হাতেকলমে ডেটা অ্যানালাইটিক্স ও ভিজ্যুয়ালাইজেশন

ডেটা অ্যানালাইটিক্স এবং ডেটা সায়েন্সের রোডম্যাপ

রকিবুল হাসান

হাতেকলমে ডেটা অ্যানালাইটিক্স ও ভিজ্যুয়ালাইজেশন

ডেটা অ্যানালাইটিক্স এবং ডেটা সায়েন্সের রোডম্যাপ

The world's most valuable resource is no longer oil, but data.

— The Economist, 2017

হাতেকলমে ডেটা অ্যানালাইটিক্স ও ভিজ্যুয়ালাইজেশন

ডেটা অ্যানালাইটিক্স এবং ডেটা সায়েন্সের রোডম্যাপ

রকিবুল হাসান

তাম্রলিপি

হাতেকলমে ডেটা অ্যানালাইটিক্স ও ভিজ্যুয়ালাইজেশন

রকিবুল হাসান

গ্রন্থস্বত্ব : লেখক

প্রথম প্রকাশ : ফেব্রুয়ারি ২০২৩

তাম্রলিপি : ৭২৮

প্রকাশক
এ কে এম তারিকুল ইসলাম রনি
তাম্রলিপি
৩৮/৪ বাংলাবাজার, ঢাকা-১১০০

প্রচ্ছদ
আমান হাসান

বর্ণবিন্যাস
তাম্রলিপি কম্পিউটার

মুদ্রণ
একতা প্রিন্টিং প্রেস

মূল্য : ৬৬৭.০০

Hatekolome Data Analytics O Visualization
By : Rakibul Hassan
First Published : February 2023 by A K M Tariqul Islam Roni
Tamralipi, 38/4 Banglabazar, Dhaka-1100

Price : 667.00 $20

ISBN : 978-984-97605-8-0

উৎসর্গ

আরশিয়া, নাবিল এবং নতুন
প্রজন্ম— যাদের হাত ধরে
ঘুরে দাঁড়াবে বাংলাদেশ

সূচি

ভূমিকা ১৩

ডেটা এবং একুশ শতকের দক্ষতা সিরিজ ১৫

ডেটা ইজ দ্য নিউ অয়েল ১৫

কেন ডেটা অ্যানালাইসিস নিয়ে বইটা লিখতে চাইলাম? ১৭

ডেটাকে বুঝতে পারা ১৭

এ যুগের স্কিলসেট, ডেটা ড্রিভেন সমাজের জন্য ১৯

একটা সত্যিকারের হাইব্রিড প্ল্যাটফর্ম (বই+ভিডিও+কোড) ১৯

ঠিক প্রশ্নটা করে সেটার সমাধান বের করা ১৯

ভালো প্রশ্ন করতে পারা একটা বিশাল স্কিল ২০

বাড়তি সাহায্য এবং কৃতজ্ঞতা ২৩

সাহায্যের লিস্ট ২৩

ডেটা স্টোরিটেলিং ২৪

এগিয়ে থাকতে ডেটাকে ঠিকমতো ‘কমিউনিকেট’ করতে পারা ২৪

পুরোনো ডেটা থেকে ধারণা নিয়ে ভবিষ্যৎ দেখার উপলব্ধি তৈরি ২৫

ক্যারিয়ারের বিভিন্ন ধাপে কমিউনিকেশনের ব্যবহার ২৬

কমিউনিকেশনে কনটেক্সটের প্রযোজ্যতা ২৭

কমিউনিকেশন স্কিল তৈরিতে নিজেকে ইনভেস্ট করা ২৭

ডেটা এক্সপ্লোর করার সফট-স্কিল ২৯

কেন শুরুতে ডেটা অ্যানালাইটিক্স স্কিল প্রয়োজন? ৩১

কেন শুরুতে ডেটা অ্যানালিস্ট? ৩১

‘ডিভাইস সেন্ট্রিক’ সিস্টেমকে ‘ইন্টারপ্রেট’ করার মানুষটা কই? ৩১

ডেটা লিটারেসি এবং ‘কেন’ প্রশ্ন? ৩২

ডেটা লিটারেসি কেন দরকার? ৩৩

তাহলে কোথায় শুরু করতে হবে? ৩৪

ঠিক প্রশ্নটা করতে পারা ৩৫
‘কেন’ প্রশ্নের উত্তর খোঁজার একটা উদাহরণ ৩৬
ম্যানেজারদের জন্য ডেটা অ্যানালাইটিক্স ৪৪
ডেটা = মানুষের অভিজ্ঞতা ৪৪
ক্রিটিক্যাল থিঙ্কিং এবং প্রবলেম সলভিং স্কিল ৪৬
ক্রিটিক্যাল থিঙ্কিং এবং আমাদের চিন্তাধারা চিন্তা করতে পারা ৪৬
শেখার ধারণা ৪৭
কেন ‘প্রজেক্ট টাইটানিক’ দিয়ে ডেটা অ্যানালাইটিক্স? ৪৮
কেন ‘প্রজেক্ট টাইটানিক’ একটা বিজনেস ‘প্রক্সি’ ডেটা? ৪৮
কেন আমাদের প্রশ্ন করতে জানতে হবে? ৪৯
ডেটাকে ঠিকমতো ‘নেভিগেট’ করার ধারণা ৫৩
ডেটা ভিজ্যুয়ালাইজেশনের কিছু ধারণা ৫৪
স্বাতীর গল্প ৫৪
কিছু বেসিক ডেটা ভিজ্যুয়ালাইজেশন ৫৬
ভার্টিক্যাল বার চার্ট ৬২
ওয়াটার ফল চার্ট ৬৩
হরাইজন্টাল বার চার্ট ৬৪
‘স্টেকড’ হরাইজন্টাল বার চার্ট ৬৫
সেকেন্ডারি y-axis রিপ্রেজেন্টেশন ৬৭
হাতেকলমে ডেটা অ্যানালাইটিক্সের শুরুটা কোথায়? ৬৯
ডেটা অ্যানালাইসিস/ভিজ্যুয়ালাইজেশনের কিছু উদাহরণ ৬৯
ক্রিমিয়ার যুদ্ধ ৬৯
একনজরে মার্কিন যুক্তরাষ্ট্রের বাজেট ৭১
হাতেকলমের কিছু টুল — গিটহাব ৭৩
আমার পছন্দের কিছু ওপেনসোর্স টুল ৭৫
ফুল তুমি কার? ৭৬
প্রশ্ন: কেন ঘটনাগুলো ঘটছে? ৭৭

দরকার একটা 'অ্যাডভান্সড টুল'	৭৭
ডাউনলোড করুন আইরিস ডেটাসেট	৭৮
ট্যাবলো পাবলিক 'ওয়েব ভার্সন' এবং আইরিস প্রজাতি	**৭৯**
বিজনেস কোশ্চেন	৭৯
নতুন একটা ফুলের ডেটা কী বলছে?	৮২
সাহায্যকারী ভিডিও	৮৩
ট্যাবলো দিয়ে আইরিস প্রজাতির ক্লাস্টার অ্যানালাইসিস	**৮৪**
বিজনেস অ্যানালাইটিক্স টুলের নিজের বুদ্ধি কেমন?	৮৪
ট্যাবলোয়ের মাথা অর্থাৎ অ্যানালাইটিক্স	৮৫
তৈরি করি 'ক্যালকুলেটেড ফিল্ড', তুলনা করার জন্য	৮৮
ক্লাস্টারকে ভেতর থেকে দেখা, 'ডেস্ক্রাইব' করা	৮৯
ট্যাবলো—প্রজাতির ক্লাস্টার অ্যানালাইসিস এবং মিসম্যাচ	**৯১**
আগের তৈরি ক্লাস্টার এবং রেকর্ড আইডি দিয়ে টেবিল	৯২
মিস ক্লাফিকেশন — রং দিয়ে	৯৩
মাইক্রোসফট পাওয়ার বিআই এবং ট্রাভেল ইনস্যুরেন্স	**৯৫**
ট্রাভেল ইনস্যুরেন্স ডেটাসেট, 'কে' এবং 'কেন' প্রশ্নের উত্তর	৯৫
ডেটা ডিকশনারি, 'ট্রাভেল ইনস্যুরেন্স' ডেটাসেট	৯৬
দরকার পাওয়ার বিআই-এর মতো একটা 'অ্যাডভান্সড টুল'	৯৭
পাওয়ার বিআইতে ট্রাভেল ইনস্যুরেন্স ডেটা লোডিং	৯৮
ট্রাভেল ইনস্যুরেন্স ভিজ্যুয়ালাইজেশন	৯৯
ডেটা নিয়ে বেসিক ধারণা	**১০২**
ডেটা কী এবং ডেটার সোর্স	১০২
ভ্যারিয়েবল	১০৩
ডেটা সোর্স	১০৩
কাঁচা মানে 'র' ডেটার কিছু উদাহরণ	১০৪
তথ্যসূত্র অর্থাৎ ডেটা সোর্স কোথা থেকে হতে পারে?	১০৫

ডেটার বৈশিষ্ট্য এবং ভেতরের ধারণা	**১০৬**
মিনিংফুল অর্থাৎ অর্থপূর্ণ ডেটার কিছু বৈশিষ্ট্য	১০৬
ডেটার কনসিসটেন্সি অর্থাৎ সংগতি	১০৭
ডেটার সেগমেন্টেশন, অর্থাৎ বিভিন্ন ভাগে ভাগ করতে পারা	১০৯
ডেটাকে ‘অরগানাইজ’ করার কিছু প্র্যাকটিস	**১১০**
কলাম এবং সারি	১১০
পরিবারের ৪ জন সদস্যের একটা উদাহরণ	১১১
ভালো স্ট্রাকচারাল উদাহরণ	১১২
পুরোনো অসংগতিপূর্ণ টেবিলকে নতুনভাবে স্ট্রাকচার করা	১১৪
‘পিভোট’, কলাম স্প্লিটিং ব্যাপারগুলো কী?	**১১৬**
মানসিক ধারণা থেকে টেবিল	১১৬
কলাম স্প্লিটিং	১১৮
বিভিন্ন ভ্যারিয়েবল, ফিল্ড টাইপ নিয়ে কিছু ধারণা	**১২০**
কোয়ালিটেটিভ এবং কোয়ান্টেটিভ ভ্যারিয়েবল	১২০
অর্ডিনাল কোয়ালিটেটিভ ভ্যারিয়েবল	১২২
ভিজ্যুয়ালাইজেশনে ভ্যারিয়েবল	১২৩
ডিসক্রিট, কন্টিনিউয়াস ভ্যারিয়েবল এবং এগ্রিগেশন	১২৫
ডেটার গ্র্যানুলারিটি	**১২৮**
ডিসক্রিট ভ্যারিয়েবলের ডিস্ট্রিবিউশন	১২৯
প্রবাবিলিটি ডিস্ট্রিবিউশন গ্রাফ	১৩১
কন্টিনিউয়াস ভ্যারিয়েবলের ডিস্ট্রিবিউশন	১৩২
ডিস্ট্রিবিউশনের বিভিন্ন শেপ এবং হিস্টোগ্রাম	**১৩৫**
ডিস্ট্রিবিউশনের বিভিন্ন শেপ	১৩৫
হিস্টোগ্রাম	১৩৫
সিমেট্রিক্যাল ডিস্ট্রিবিউশন	১৩৬
‘পজিটিভ স্কিউ’ ডিস্ট্রিবিউশন	১৩৭
‘নেগেটিভ স্কিউ’ ডিস্ট্রিবিউশন	১৩৮

দুটো জটিল উদাহরণ: বক্স প্লট এবং হিস্টোগ্রাম ১৪০

শুরুর উদাহরণ: বক্স প্লট ১৪০

কন্টিনিউয়াস ভ্যারিয়েবলের ডিস্ট্রিবিউশন: বক্স প্লট ১৪০

তৈরি করি একটা বক্স প্লট ১৪২

পার্সেন্টাইল ক্যালকুলেশন ১৪৩

বক্সকে প্লট করি পার্সেন্টাইল হিসেবে ১৪৪

দলছুট অর্থাৎ আউটলাইয়ার ভ্যালু নিয়ে ধারণা ১৪৯

বক্স প্লট এবং হিস্টোগ্রাম ১৫৩

বক্স প্লট এবং হিস্টোগ্রাম পাশাপাশি দেখা ১৫৩

বক্সপ্লটের উপযোগিতা ১৫৪

স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ ১৫৫

ডেটা অ্যানালাইটিক্সে ডেটাবেইসের ধারণা ১৫৫

রিলেশনাল ডেটাবেইস ১৫৬

আমাদের মনের কথা বলবে, স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ— ১৫৮

লং টার্ম ‘এসকিউএল’ শেখার রোডম্যাপ ১৫৮

BigQueryতে কীভাবে কাজ করব? ১৫৯

এসো নিজে করি: গুগল ডেটা স্টুডিও ১৬৯

টেক্সটবুক এবং ভিডিও সিরিজের সংযোগ ১৭০

ওপেনসোর্সে হাজারো ডেটা অ্যানালাইটিক্স টুল ১৭৯

প্লটলি এক্সপ্রেস ১৭৯

ওপেনসোর্সে আইরিস ডেটাসেট ১৮০

গ্যাপমাইন্ডার ডেটাসেট ১৮২

জিও লোকেশন ডেটাসেট ১৮৯

রিয়েল ওয়ার্ল্ড ওপেনসোর্স প্রজেক্টের জন্য টুল ১৯১

গ্রাফানা: সব কাজের কাজি ১৯১

নাঈম— ‘KNIME’: এন্ড টু এন্ড ডেটা সায়েন্স ১৯২

অ্যাপাচি স্পার্ক: লার্জ স্কেল ডেটা প্রসেসিং টুল ১৯৩

কিবানা: ইলাস্টিক সার্চ ডেটাকে দেখতে ১৯৪
ডিথ্রি: অসাধারণ ভিজ্যুয়ালাইজেশন জাভায় ১৯৪

অনলাইনে বিভিন্ন সার্টিফিকেশন ১৯৫

এ যুগের প্রশিক্ষণ, ডেটা ড্রিভেন সমাজের জন্য ১৯৫
আমার গল্প: গ্লোবাল সার্টিফিকেশন ১৯৭
গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন: ব্যাকস্টোরি ১৯৮

ইন্ডাস্ট্রি কেন শেখাচ্ছে ডেটা অ্যানালাইটিক্স? ২০০

গ্রো উইথ গুগল সিরিজ ২০০
গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন: সামারি ২০২

ডেটা সায়েন্স/অ্যানালাইটিক্স রোডম্যাপ ২০৪

লার্নিং রোডম্যাপ: ব্যাপারটা কী? ২০৫
ডেটা সায়েন্স সামারি, (*) মার্ক = অ্যানালাইটিক্স ২০৬
বেসিক ‘এন্ড টু এন্ড’ পাইপলাইনের ধারণা (ঐচ্ছিক) ২০৮
ডেটা সায়েন্সের ‘হায়ারার্কি অফ নিডস’ * ২০৯
ডিপ ডাইভ: ‘হায়ারার্কি অব ডেটা সায়েন্স নিডস’ (সবার জন্য প্রযোজ্য নয়) ২১০
(পাইথন) ডেটা সায়েন্স ওয়ার্কস্পেস, অ্যানাকন্ডা ২১৬
বেসিক অঙ্ক, লিনিয়ার অ্যালজেব্রা, ক্যালকুলাস ২১৬
পরিসংখ্যান, ডেসক্রিপটিভ, সঙ্গে ডিস্ট্রিবিউশন * ২১৭
প্রোগ্রামিং স্কিল, পাইথন/‘আর’ এনভায়রনমেন্ট * ২১৮
(পাইথন) দিয়ে ডেটার চারটি কাজ ২১৯
এক্সেল, গুগলশিট, ট্যাবলো, পাওয়ার বিআই * ২২০
(পাইথন) মডেল তৈরি, মেশিন লার্নিং ২২১
এসকিউএল/সিক্যুয়েল স্কিল, SQL * ২২২
নো সিক্যুয়েল (ঐচ্ছিক) ২২২
কাজের পোর্টফোলিও ‘শো-কেসিং’ * ২২৩
ডেটা সায়েন্স কমিউনিটি তৈরি ২২৪

ভূমিকা

এই বইটা লিখেছি, অনেকটাই গল্পের ছলে। আমার অভিজ্ঞতা নিয়ে। নিজের দর্শন এবং অন্যান্য ডেটা এক্সপার্টদের ধারণা নিয়ে। আমার অন্য বইগুলোর মধ্যে এখানেও গল্প (দর্শন) ভর্তি। যেকোনো জিনিস ভালোভাবে শিখতে গেলে 'স্টোরিটেলিং'-এর বিকল্প নেই। পিথাগোরাসের সূত্র না বলতে পারলেও কচ্ছপ এবং খরগোশের গল্প মনে আছে আমাদের। আর সে কারণেই আমি গল্পচ্ছলে বই লিখি। এটা আমার দুর্বলতা।

অন্যান্য টেকনোলজির বইয়ের মতো এখানে খুব বেশি কোড খুঁজবেন না। 'স্ট্যাক ওভারফ্লো' থেকে শুরু করে হাজারো সাইট আছে, যেখানে ডেটা অ্যানালাইসিস নিয়ে হাজারো কোড পাবেন, তবে সেই কোডগুলো শুরুতে কেন লাগছে, সেটা নিয়েই বেশি ধারণা থাকছে এখানে। ডেটা দিয়ে কী করা যাবে আর কী করা যাবে না, তার পাশাপাশি ডেটা আপনার কাজে আসছে কি না সেটা নিয়েই বেশি ব্যস্ত থাকব আমরা।

এই বইটাতে আমি প্রচুর কমার্শিয়াল অ্যাপ্লিকেশন নিয়ে কথা বলেছি, কারণ এগুলো নিয়ে দিয়ে সহজেই যেকোনো বিজনেসের মানুষজন সরাসরি কাজ করতে পারেন। উদাহরণ হিসেবে, এখানে এক্সেলের কথা বলতে পারি, যা পুরো পৃথিবীকে পাল্টে দিয়েছে। স্বাতী, আমার স্ত্রী যে ব্যাংকে (ফিন্যান্সিয়াল সার্ভিস কোম্পানি) ২২ বছর ধরে আছেন, সেই ব্যাংকের প্রতিটি কাজের ক্ষেত্র থেকে শুরু করে গ্লোবালি সব প্রতিষ্ঠানের বিজনেসের লোকজনই চালাচ্ছে ডেটা অ্যানালাইসিসের মতো কমপ্লেক্স জিনিস। এর অর্থ হচ্ছে, প্রোগ্রামার নয়; বরং বিজনেসের মানুষজন ডেটা নিয়ে কাজ করতে পারছেন, নতুন প্রজন্মের ডেটা অ্যানালাইটিক্স টুল দিয়ে। এই কমার্শিয়াল অ্যাপ্লিকেশনগুলোতে প্রচুর বুদ্ধিমত্তা ঢুকিয়ে দেয়া হয়েছে মানুষের সাহায্যে, যাতে যে কেউ কাজ করতে পারে ডেটা নিয়ে। তবে আমার হৃদয় পড়ে আছে ওপেনসোর্সে।

আমি অনেক বড় বড় কোম্পানির ইন্টার্নাল কমার্শিয়াল অ্যাপ্লিকেশন নিয়ে কথা বলব, অর্থাৎ যে কোম্পানিগুলোর একটা বড় রেভিনিউ আসে কিছুটা ক্যাপটিভ মার্কেট থেকে, সেখানে তারা বড় বড় কমার্শিয়াল অ্যাপ্লিকেশনে ইনভেস্ট করেন। কারণ তাদের জন্য 'টাইম ইজ মানি'। তবে, আমার দেখা মতে ছোট ছোট স্টার্টআপ, যারা সামাজিক বিভিন্ন সমস্যার সমাধানে এগিয়ে আসছে, তাদের জন্য শুরুতে এ ধরনের এক্সপেন্সিভ কমার্শিয়াল ভেঞ্চারে যাওয়াটা অনেক কঠিন। সে কারণেই তাদের মূল স্ট্রেন্থ আসে ওপেনসোর্স ডেটা অ্যানালাইটিক্স প্ল্যাটফর্ম থেকে।

মজার কথা হচ্ছে, আমি নিজেও কয়েক শ মিলিয়ন ডলার প্রজেক্টে প্রচুর ওপেনসোর্স টুল (কিছুটা 'মোডিফাই' করে) ব্যবহার করেছি, যা বর্তমান যুগের একটা ট্রেন্ড। আজকে গুগল, ফেসবুক, অ্যামাজন, নেটফ্লিক্স থেকে শুরু করে সবাই এ ধরনের ওপেনসোর্স টুল ব্যবহার করে বানিয়ে নিয়েছে তাদের কিছুটা কাস্টমাইজড অ্যাপ্লিকেশন। সে কারণেই, ওপেনসোর্স আমার অসম্ভব প্রিয়। পরের দিকে এই ওপেনসোর্স টুল নিয়ে একটা আলাদা চ্যাপ্টার থাকবে যেটা আপনাকে পথ দেখাবে নতুন ভেঞ্চারে।

আমি নিজে 'ফার্স্ট প্রিন্সিপাল থিঙ্কিং' প্রসেসে বিশ্বাসী বলে সবকিছু একদম গোড়া থেকে শুরু করতে চাই। একদম নতুন শিশুর মতো— কিছুই জানি না। তবে প্রতিটা জিনিস ধরে ধরে প্রশ্ন করে নিজের মতো করে 'জানা'— আমরা যেটাকে মনে করি যে আমরা জানি (আসলে জানি না) সেটাকে প্রশ্ন করে নতুন কিছু শেখা। আমরা যেটাকে জানি— অর্থাৎ আমাদের নিজেদের 'অ্যাজাম্পশন'কে প্রশ্ন করলে অনেক অজানা জিনিস বের হয়ে আসবে। সেই মেথোডলজিতে এগিয়েছে বইটা। কিছু কিছু জিনিস একদম বেসিক মনে হতে পারে— তবে সেই ধারণাটা আমাদের নিয়ে যাবে ডেটা সায়েন্সের রাস্তায়।

ডেটা এবং একুশ শতকের দক্ষতা সিরিজ

ডেটা ইজ দ্য নিউ অয়েল

কয়েক দশক আগে পৃথিবীর সব বিলিয়ন ডলার কোম্পানি হিসেবে চলে আসত এনার্জি/পাওয়ার কোম্পানিগুলোর নাম। ব্রিটিশ পেট্রোলিয়াম, জেনারেল ইলেকট্রিক, শেল, সৌদি আরামকো— কারণ, তখন তেল চালাত বিশ্ব। এখন বিশ্ব চালাচ্ছে ডেটা। যার কাছে বেশি ডেটা, সে-ই চালাচ্ছে পৃথিবী। আমরাও নেচে যাচ্ছি তাদের অ্যালগরিদমের পাল্লায় পড়ে। ডেটা থেকে প্রজ্ঞা নিয়ে দুনিয়া দাপাচ্ছে গুগল, ফেসবুক, আমাজন, আলিবাবা, দারাজ, নেটফ্লিক্সের মতো ডেটা ড্রিভেন কোম্পানি। এই কোভিড ও কোভিডপরবর্তী সময়েও ডেটার দাপটের কমতি নেই, যখন অন্য বিজনেসগুলো ক্রমশ বিপদে পড়েছে।

এবার আসি বাংলাদেশে। অনেক কোম্পানির ম্যানেজমেন্ট কথা বলেন আমার সাথে বিশেষ করে ডেটা অ্যানালাইসিস নিয়ে। তারা চাচ্ছেন, কীভাবে তাদের বছরের পর বছরের 'পাইল করা' ডেটা থেকে প্রজ্ঞাকে ব্যবহার করা যায় কোম্পানির স্বার্থে। আমি সরকারে থাকার সময় একটা বড় সময় কাটিয়েছি ডেটার ব্যবহার নিয়ে। এর অর্থ আমাদের বাজারও তৈরি হচ্ছে। একটু সময় লাগছে, এই যা। সবাই ডেটা থেকে জ্ঞান নিতে চাইছেন।

হিসেব করে দেখা গেছে যে, পৃথিবীর প্রত্যেক মানুষের জন্য তৈরি হচ্ছে ১.৭ মেগাবাইট ডেটা, প্রতি সেকেন্ডে। এর সরলীকরণ হচ্ছে, সেই ১.৭ মেগাবাইট দিয়ে প্রায় ৮৫০ পাতার বই বানানো যায় এক সেকেন্ডে। ধারণা করুন, কত কত ডেটা তৈরি হচ্ছে আমাদের অজান্তে। প্রতিদিন ৩ হাজার কোটি ইমেইল চালাচালি হয় পৃথিবীজুড়ে। গুগল প্রায় ৬০০০ কোটি সার্চ

রিকোয়েস্ট প্রসেস করে একই দিনে। ইউটিউবে প্রতি মিনিটে ৪ মিলিয়ন ভিডিও দেখে পৃথিবীর মানুষ। এই মুহূর্তে বলা যেতে পারে, প্রতিদিন প্রায় ৩ 'এক্সাবাইট' ডেটা জেনারেট হয় আমাদের কাজে।

মনে আছে ২০১৩ সালের আইবিএমের সেই রিপোর্টের কথা? সেখানে বলা হয়েছিল ওই সময়ে ৯০% পুরো পৃথিবীর ডেটা জেনারেট হয়েছিল ২০১০ থেকে ২০১২ সালের মধ্যে। চিন্তা করা যায়? অর্থাৎ পৃথিবীর জন্ম থেকে ২০০৯ সাল পর্যন্ত পৃথিবী জেনারেট করতে পেরেছিল মাত্র ১০% ডেটা। অনেকটাই অবিশ্বাস্য। সামনে এর হার আরও বাড়বে।

এখন বলুন, এই ডেটা প্রসেস করবে কে?

মানুষ? না। যন্ত্র অবশ্যই। তবে, সেই যন্ত্রের হাতে সেই সিদ্ধান্তের চাবিকাঠি দেয়ার আগে মানুষকে বুঝতে হবে ডেটা কী বলতে চাইছে? কিছুটা 'এক্সপ্লেইনেবল এআই' অর্থাৎ 'ব্ল্যাকবক্স' কী কাজ করছে, সেটা জানা জরুরি মানুষের। সে কারণে ডেটার ভেতরে কী ঘটছে, সেটা জানলে এর ভেতরের ধারণা পরিষ্কার হবে। মানুষকে জানতে হবে ডেটা থেকে এমন কী 'ইনসাইট' পাওয়া যায়, যা যন্ত্র এখনো ধরতে পারছে না। যন্ত্র অবশ্যই ভালো, বিশেষ করে স্পেশালাইজড কাজে, তবে যন্ত্রের ওপর 'সুপারভিশন' এবং ডেটার হাই লেভেল ধারণা পেতে মানুষকে জানতে হবে 'ডেটা অ্যানালাইটিক্স'। ডেটা নিয়ে অ্যানালাইসিস করতে করতে ডেটা বোঝার জন্য শুরুতে প্রোগ্রামিং দরকার নেই। এখন এমন টুল চলে এসেছে— যা প্রশ্নাতীতভাবে সহজ। প্রচুর 'এআই' টুল আমি চালিয়েছি— যা হাজারো লাইন কোডিং থেকে অনেক সহজ করে দিয়েছে। তবে, গোড়াতে ঢুকতে কিছু প্রোগ্রামিং দরকার। এ ব্যাপারে শেষ চ্যাপ্টারে ডেটা অ্যানালাইসিস/ডেটা সায়েন্স রোডম্যাপ চ্যাপ্টারটা আমাদের একটা বড় পার্সপেক্টিভ দেবে। এই বইগুলো সে কারণে একটা সিরিজের অংশ।

কেন ডেটা অ্যানালাইসিস নিয়ে বইটা লিখতে চাইলাম?

ডেটাকে বুঝতে পারা

Your relevance as a data custodian is your ability to analyze and interpret it. If you can't, your replacement is due.

— Kwashie Mensah

মনে আছে প্রমিথিউসের কথা— মানবসভ্যতার জন্য জিউসের কাছ থেকে আগুন চুরি করার জন্য শাস্তি পেতে হয়েছিল? অথচ, এই আগুনই মানবসভ্যতার জন্য একটা বিশাল মাইলস্টোন। ব্যাপারটা গল্প হলেও এর পেছনের ধারণা মানবসভ্যতার উৎকর্ষকে ঘিরে। আজ সেই আগুনকে জ্ঞানের সমার্থক ধারণা হিসেবে বলা যেতে পারে।

ছাপাখানা মানুষের জ্ঞানকে ছড়িয়ে দিয়েছিল বিশ্বজুড়ে। বাষ্পীয় ইঞ্জিন মানবসভ্যতাকে সাহায্য করেছিল দূর-দূরান্তে যেতে। এরপর বিদ্যুৎ পাল্টে দিল পুরো সমাজব্যবস্থা। আজ ইন্টারনেটের বদৌলতে পুরো পৃথিবীর তথ্য ছড়িয়ে পড়েছে গ্রহের আনাচেকানাচে। তথ্য, অর্থাৎ ডেটা হচ্ছে পরবর্তী গেম চেঞ্জার। এই ডেটা আসছে বিদ্যুতের মতো ‘গ্রিড’ থেকে। মানুষের প্রতিটা কাজের পেছনের সিদ্ধান্ত যদি ডেটানির্ভর হয়, তাহলে সেই কাজ নির্ভুল এবং সহজেই করা সম্ভব।

ডেটা নিয়ে পাঁচটা বই, বিভিন্ন ফোরামে বক্তৃতা দেয়া ও শোনা থেকে শুরু করে শতাধিক ভিডিও এবং ব্লগ পোস্ট নিয়ে কাজ করার সময় কথা হয়েছে নতুন প্রজন্ম এবং ম্যানেজমেন্ট পর্যায়ের অনেক মানুষের সাথে। আমার মনে হয়েছে— সবাই ডেটার সুবিধাগুলোর ব্যাপারে জানেন, তবে

কোথা থেকে শুরু করলে একদম বেসিক তথ্যপ্রযুক্তি জ্ঞান নিয়ে এগোনো যাবে, সেটা নিয়ে দ্বিধা-দ্বন্দ্ব রয়েছে। আমি তাদের দোষ দিতে পারি না, কারণ এই ডেটা নিয়ে প্রচুর ক্যারিয়ার ট্র্যাক তৈরি হয়েছে এরই মধ্যে।

আপনি যেহেতু ডেটা নিয়ে ক্যারিয়ার শুরু করতে চাচ্ছেন, সেখানে ডেটার এই দরজা দিয়ে ঢুকতে গেলে প্রথমেই দরকার ডেটাকে ঠিকমতো বুঝতে পারা। ডেটাকে বুঝতে প্রয়োজন ডেটা অ্যানালাইসিস, ডেটা আসলে কী বলতে চাচ্ছে আমাদের। ডেটার 'স্টোরিটেলিং' তৈরি করতে পারেন একজন ডেটা অ্যানালিস্ট। ডেটা থেকে বিভিন্ন ধরনের বিজনেস ইন্টেলিজেন্স বের করতে প্রয়োজন বিভিন্ন টুলস। তবে সেই টুলস ব্যবহার করতে জানার আগে প্রয়োজন সেটার পেছনে দর্শনকে বোঝা। 'ডেটা অ্যানালাইসিস' ঠিকমতো বুঝলে ডেটা সায়েন্সের রাস্তায় নামা সোজা হয়ে যাবে। শুরুতে ডেটা অ্যানালাইসিস, পরে ডেটা সায়েন্স।

তিন বছর আগে গুগল একটা অসাধারণ কাজ শুরু করে। ডেটা অ্যানালাইটিক্স নিয়ে সব ধরনের গুঞ্জন মিটিয়ে নিয়ে আসে আটটা সাব কোর্সের একটা 'এন্ড টু এন্ড' কোর্সের পাইপলাইন। গুগল ক্যারিয়ার সার্টিফিকেট প্রোগ্রামের অংশ হিসেবে ডেটা অ্যানালাইটিক্স কোর্সটা অডিটিং করতে গিয়ে দেখলাম— পুরো জিনিসটা মিলে যাচ্ছে আমার পার্সপেক্টিভের সাথে। ধরুন, আমি যদি একটা অনলাইন/অফলাইন অথবা হাইব্রিড কোর্স চালু করতাম, তাহলে হয়তোবা এই একই মেথডোলজির স্ট্রাকচারে কোর্সটা চালু করতাম। তবে, অনেক ছাত্র যারা বাংলাদেশ থেকে 'কোর্সেরা'র স্কলারশিপ পেয়েছেন— তারা কিছু কিছু জায়গায় সাহায্য চেয়েছেন। কেন তারা সাহায্য চাচ্ছেন এ ধরনের কোর্সে? আমি সাহায্য চাইলাম নিজের বাসায়।

আর্শিয়া (একাদশ) এবং নাবিল (নবম) আমার দুই সন্তান, যারা ডেটার ব্যাপার ছাড়া বাকি সব ব্যাপারেই অতি উৎসাহী। তাদের অনুরোধ করে অনেক কষ্টে গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন কোর্সে এনরোল করিয়েছি। মজার ব্যাপার হচ্ছে, তারা আমার সাহায্য ছাড়াই জিনিসটা ধরতে পারছে। তবে, আমাদের শিক্ষাক্রমে ডেটা নিয়ে সে ধরনের ধারণা না থাকাতে হোঁচট খাচ্ছেন অনেকে। ডেটা জিনিসটা কী, এটা কীভাবে দেখতে হয়, এর পেছনের আন্ডারস্ট্যান্ডিংগুলো আসছে সামনে।

এ যুগের স্কিলসেট, ডেটা ড্রিভেন সমাজের জন্য

সামনে বেশির ভাগ কাজ ডেটার। ইদানীং গুগল বলছে ডেটানির্ভর চার বছরের ডিগ্রির জায়গায় ৬-১২ মাসের নতুন সার্টিফিকেশন তৈরি করে দেবে নতুন জনবল— একুশ শতকের জন্য। গুগল ট্রেন্ড বলছে— মার্কেটে একটা বিশাল গ্যাপ তৈরি হয়েছে এই মহামারির কারণে। ডেটা বোঝা এবং ডেটা থেকে সিদ্ধান্ত নিতে পারা জনবল নিয়ে।

তিন বছর ধরে কোর্সগুলোকে ফুল স্কেলে অডিটিং করে যা বোঝা গেল, বিশাল ইনভেস্টমেন্ট করেছে গুগল, ফেসবুক, নেটফ্লিক্স এবং আইবিএমের মতো কোম্পানিগুলো— এই গ্যাপ পূরণে। আমি তাদের কোর্সওয়ার্ক ম্যাটেরিয়াল দেখে সত্যিকারভাবে অভিভূত। এই কোর্সগুলো করতে বিশেষ কোনো প্রাতিষ্ঠানিক পড়াশোনার প্রয়োজন নেই। এর পাশাপাশি, ওই কোর্সওয়ার্কগুলো এবং আমার অভিজ্ঞতা মিলিয়ে নন-প্রোগ্রামারদের জন্য বাংলায় আমার ডেটা অ্যানালিস্ট বই।

একটা সত্যিকারের হাইব্রিড প্ল্যাটফর্ম (বই+ভিডিও+কোড)

অনেকগুলো প্রযুক্তিগত বই লিখে এই প্রথম একটি পূর্ণ হাইব্রিড বই তৈরি করছি। যেকোনো বিষয়ের জন্য তথ্যগত জ্ঞানের পাশাপাশি ভিডিও টিউটোরিয়াল অনেক সাহায্য করে। ভিডিও টিউটোরিয়ালের পাশাপাশি বেজলাইন তথ্যগত টেক্সট ছাড়া যেকোনো বিষয়ে 'মাস্টারিং' করা অসম্ভব হয়ে দাঁড়ায়। আর এ কারণেই এই বইটার প্রথম অংশ তাত্ত্বিক এবং এর পেছনের দর্শনকে প্রায়োরিটি দেওয়া হয়েছে। ডেটা অ্যানালাইটিক্সের ওপরে সঠিক এবং প্রয়োজনীয় জ্ঞানের জন্য প্রাথমিক তাত্ত্বিক এবং এর পেছনের 'আন্ডারলায়িং' দার্শনিক ধারণাগুলোর সাথে শতাধিক ভিডিও যোগ করা হচ্ছে। দুটোই প্রয়োজনীয়, এই স্কিলসেট অর্জনের জন্য।

ঠিক প্রশ্নটা করে সেটার সমাধান বের করা

১০ বছর আগেও যে ঘটনা দেখেছি, সেদিক থেকে— বর্তমানে সরকারি/ বেসরকারি প্রতিষ্ঠানগুলোতে এত বেশি ডেটা আসছে এবং এত অল্প সময়ের মধ্যে, যা আমি নিজেও বিশ্বাস করতে পারছি না। এটা ঠিক যে আগের

ডেটাগুলো অনেক ছোট ছিল— যেটাকে সহজে ম্যানেজ করা যেত। তবে এই বিগ ডেটার যুগে সেটা মাথা খারাপ করে দিচ্ছে। যেমন সামান্য একটা প্রডাক্ট বিক্রি করার এত চ্যানেল তৈরি হয়েছে, যা মানুষের পক্ষে প্রতিটা চ্যানেল ট্র্যাক করা অসম্ভব হয়ে পড়েছে। সেটাকে ছাড়তে হচ্ছে যন্ত্রের ওপরে। আর সেই যন্ত্র জেনারেট করে চলছে হাজারো ডেটা। কী যন্ত্রণা! এ কারণে ডেটা থেকে প্যাটার্ন বোঝার জন্য আসবে মেশিন লার্নিং, যা ডেটা সায়েন্সের মধ্যে চলে যাবে। এ ব্যাপারে বিস্তারিত পাবেন শেষের রোডম্যাপে।

আমরা দেখছি যে ডেটা শুধু ভালো ভবিষ্যৎ প্রেডিকশন করছে না; বরং একটা ফ্লাইওভারের লেন কোন দিক দিয়ে নামালে সামনের ১০ বছরে— পুরো শহরের হাজার লক্ষ ঘণ্টা বেঁচে যাবে সেটাও হিসেব করে দিতে পারছে এখনই, ফ্লাইওভার বানানোর আগে। মেট্রোরেলের মতো একটা পাবলিক ট্রানজিট সিস্টেম প্রতিদিন শহরের কার্বন কমানো এবং মানুষের সময় কত কমাচ্ছে, সেটা জানা যাবে মেট্রোরেলে সিস্টেম বসানোর আগে। তখন যেকোনো ইনভেস্টমেন্ট সহনীয় হয়ে আসে। সেদিক থেকে আপনার কোম্পানির নতুন রিটেইল লোকেশন কোথায় হলে সবচেয়ে বেশি বিক্রি হবে, সেখানেও সাহায্য করছে ডেটা অ্যানালাইটিক্স। এর ব্যাকগ্রাউন্ডে, ডেটাকে প্রশ্ন করতে পারছি বলেই সেগুলো জানতে পারছি। কারণ, ডেটার ভেতরে লুকিয়ে আছে সব উত্তর।

ভালো প্রশ্ন করতে পারা একটা বিশাল স্কিল

If I had an hour to solve a problem and my life depended on the solution, I would use the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes.

— Albert Einstein (1879-1955)

এর অর্থ হচ্ছে আপনি ডেটাকে ঠিকমতো প্রশ্ন করতে পারছেন বলে সেটার উত্তর পাচ্ছেন। আমাদের সমস্যার একটা বড় অংশের সমাধান পাই না ঠিক প্রশ্ন করতে না পারার কারণে। আবার ঠিকমতো প্রশ্ন

করতে পারলে একটা প্রডাক্টের রিটার্নের হার আপনাকে বলে দেবে কোম্পানির ডিস্ট্রিবিউশন চ্যানেলে কোথায় সমস্যা হচ্ছে। এদিক-সেদিক কথা বলে ম্যানেজমেন্টকে 'বুঝ' দেওয়ার সুযোগ নেই আর। আগের ডেটা, বিশেষ করে একটা 'ফ্লপ' অ্যাডভার্টাইজমেন্ট ক্যাম্পেইন, অর্থনৈতিক মন্দার ইফেক্ট, কোম্পানির জনবলের খারাপ পারফরম্যান্সের ভেতরের ডেটা আমাদের সাহায্য করছে কীভাবে সামনের 'লাইন অব ওয়ার্ক' ঠিকমতো বুঝে 'অ্যাডজাস্ট' করে কোম্পানির প্রবৃদ্ধি বাড়ানো যায়। দিন শেষে কোম্পানির প্রবৃদ্ধি নির্ভর করছে ডেটাকে কতটা দক্ষভাবে ব্যবহার করা যায় তার ওপর।

আমাদের দেশে অনেকগুলো ডেটা ড্রিভেন সরকারি এবং বেসরকারি অ্যাপ্লিকেশন দেখছি, তবে এই মুহূর্তে বাংলাদেশের জনগণের জন্য ভ্যাকসিনেশন অ্যাপ্লিকেশন, সুরক্ষা অ্যাপ— পুরোটাই ডেটা ড্রিভেন। কবে, কোথায়, কখন, পার্টিকুলার কোন ভ্যাকসিনেশন, কার মাধ্যমে, পরবর্তী ভ্যাকসিনের জন্য কবে ডাকা হতে পারে ইত্যাদি ইত্যাদি ডিটেইলিং থেকে শুরু করে, অনেক কিছুর পেছনে আছে এর ডেটা অ্যানালাইসিস। কখন ৪০ বছরের সীমা থাকবে ওপরে এবং কখন সেটা নামাতে হবে— কাদের কোন মুহূর্তে অগ্রাধিকার দিতে হবে, কাদের কোন ভ্যাকসিন দিতে হবে এবং এর সাথে ভ্যাকসিনের সরবরাহকে, অর্থাৎ এর সাপ্লাই চেইনের প্রচুর ডেটা দিয়ে প্রতিদিনের হিসেব কীভাবে মেলানো হবে— সবকিছুই যুক্ত হচ্ছে এর পেছনের ডেটা মডেলে। ধীরে ধীরে এর সঙ্গে যুক্ত হচ্ছে অন্যান্য সরকারি অ্যাপ্লিকেশন। যেমন— এ মুহূর্তে যুক্ত হয়েছে প্রবাসীকল্যাণ ও বৈদেশিক কর্মসংস্থান মন্ত্রণালয় এবং আস্তে আস্তে আরও অনেক সংস্থা যুক্ত হচ্ছে এর সাথে। এর পেছনে চমৎকারভাবে কাজ করছে একটা ডেটা অ্যানালাইসিস টিম।

সামনে আমরা দেখব, কীভাবে একটা ডেটাসেট থেকে কিছু নির্দিষ্ট ভ্যারিয়েবলগুলোকে আলাদা করলে কোম্পানির প্রবৃদ্ধি কী কী জিনিসে আটকে যাচ্ছে, সেটাকে ধরে সমস্যা 'আইসোলেট' করে ফেলব। এই কাজ করব, যাতে আমাদের ফলাফল ঋণাত্মক না হয়। আমরা এটাও দেখার চেষ্টা করব, কীভাবে ইন্ডাস্ট্রি এবং বিভিন্ন মার্কেটপ্লেসে এই ছোটখাটো ভ্যারিয়েবল আমাদের সিদ্ধান্তগুলোকে ঠিকমতো বাস্তবায়নে বাধা দিচ্ছে।

কেন আমরা সনাতন ‘ইনসটিঙ্কট’কে অগ্রাধিকার দিচ্ছি ডেটা থেকেও? ডেটা হচ্ছে ফ্যাক্ট, সত্যিকারের তথ্য। সেই ডেটা আমাদের নেগেটিভ ইফেক্ট দিচ্ছে, তাদের আলাদা করে ‘ট্রিটমেন্ট’ দিতে পারতে জানতে হবে। এই ‘ঋণাত্মক’ ডেটা ফেলে না দিয়ে সেটার ‘ফিগার ইঞ্জিনিয়ারিং’ নিয়ে কথা বলেছি, ‘হাতেকলমে মেশিন লার্নিং’ বইটাতে।

বাড়তি সাহায্য এবং কৃতজ্ঞতা

সাহায্যের লিস্ট

ডেভিড ল্যাংগার, পরিচয় হয় শিকাগোতে, একটা বুটক্যাম্পে। তিনি হাতে ধরে দেখিয়েছিলেন কীভাবে 'আর' এর অসাধারণ লাইব্রেরি দিয়ে 'লাইফ লাইক' ভিজ্যুয়ালাইশেন তৈরি করা যায়।

ফ্রান্সেসকো মস্কোনি, যার সাথে আমার পরিচয় হয় সানফ্রান্সিসকোতে। যার হাত দিয়ে হাতেখড়ি হয় ডিপ লার্নিং প্ল্যাটফর্মগুলোর অসাধারণ ডেটা ম্যানিপুলেশন স্কিলে।

আলেকজান্ডারের ভিজিবল ডিফারেন্স, https://vizdiff.blogspot.com/— ট্যাবলো দিয়ে তৈরি একটা চমৎকার সাইট।

'লন্ডন বিজনেস অ্যানালাইটিক্স গ্রুপের ইনপুট হিসেবে প্রগতি জৈন বেশ কিছু অনলাইন ক্লাস আমাকে 'পাওয়ার বিআই'তে আসতে বাধ্য করেছেন।

ট্যাবলো ওয়েবসাইটের বেসিক ডেটা অ্যানালাইটিক্স এবং ভিজ্যুয়ালাইজেশন টুলের ওপরে অনলাইন ট্রেনিংগুলো অনেকটাই 'আউট অব দ্য ওয়ার্ল্ড'। এই ট্রেনিংগুলো বেসিক ধারণাকে পোক্ত করার জন্য এক কথায় অসাধারণ।

মাইক্রোসফটের ডেটা অ্যানালিস্ট এর ট্রেনিং সাইট 'গেট স্টার্টেড উইথ মাইক্রোসফট ডেটা অ্যানালাইটিক্স'।

ফাইনালি বিটিআরসি এবং এনটিএমসির মতো প্রতিষ্ঠানে ডেটা নিয়ে কাজ করার অভিজ্ঞতা আমাকে একটা বিশাল 'মাইলেজ' দিয়েছে। এই তেরো বছরের কাজের অভিজ্ঞতা আমাকে 'ইন্সপায়ার' করেছে ডেটা নিয়ে পড়ে থাকার ব্যাপারে।

ডেটা স্টোরিটেলিং

এগিয়ে থাকতে ডেটাকে ঠিকমতো ‘কমিউনিকেট’ করতে পারা

ডেটাকে ঠিকমতো ‘কমিউনিকেট’ করার স্কিল + স্টোরিটেলিং

একটা পরিসংখ্যানে দেখা গেছে, প্রতিটি ‘জব’ অর্থাৎ কাজে ডেটার ব্যবহার বেড়েছে বলে সেই ডেটাকে কীভাবে কমিউনিকেট করা যায়, সেটা নিয়ে অনেক রিসার্চ হচ্ছে। যে জিনিসটা নিয়ে সবচেয়ে বেশি আলাপ হচ্ছে, সেটা হলো কোথাও এগিয়ে থাকতে হলে ডেটাকে আমাদের স্টেকহোল্ডারদের (আমাদের ম্যানেজমেন্ট অথবা ক্লায়েন্ট) মধ্যে ঠিকমতো কমিউনিকেট করতে পারতে হবে। এদিকে আমরা যখন ডেটাকে ঠিকমতো ‘কমিউনিকেট’ করব, তখন সেটার জন্য প্রয়োজনীয় ‘রিকমেন্ডেশন’ যেমন আগের প্রপোজাল, অ্যাকশন প্ল্যান অথবা রিপোর্টগুলোকে প্রযোজ্য মাপে কাটছাঁট করতে হবে যাতে যেখানে যেটা প্রযোজ্য, সেখানে সে ধরনের ডেটা থেকে ‘ইনসাইট’ দিতে পারা যায়। যিনি সবচেয়ে ভালো কমিউনিকেটর, তিনি ডেটাকে ঠিকমতো স্ট্রাকচারে ফেলে একটা বিশ্বাসযোগ্য (যেহেতু পেছনে ডেটা আছে) এবং মনে রাখার মতো স্টোরিটেলিং ন্যারেটিভে প্লট করেন বলেই উনি বাকিদের থেকে এগিয়ে থাকেন।

আমাদের এই স্টোরিটেলিং এর সুবিধা হচ্ছে, তাদের স্টেকহোল্ডারকে (আমাদের ম্যানেজমেন্ট অথবা ক্লায়েন্ট হতে পারেন) ঠিক কী মেসেজ কমিউনিকেট করা হচ্ছে, সেটা বোঝানো গেলে পুরো ব্যাপারটাই একটা ভালো ‘রেশ’ ফেলে তাদের মনে। ফলে ব্যাপারটা মনে থাকে তাদের। এ কারণে ঠিক ধরনের ভিজ্যুয়াল এবং মুখের গল্পে যখন একটা ‘রিকমেন্ডেশন’ কমিউনিকেট করা হয়, তখন সেখানে অসাধারণ কিছু স্টোরিটেলিং এবং ডেটা প্লটের ব্যাপার চলে আসে। এ ছাড়া, শুধু মুখের কথায় ‘বুঝ’ দেওয়ার

ধারণাগুলো চলে যাচ্ছে আস্তে আস্তে। যেকোনো ধারণাকে বিশ্বাসযোগ্য করতে হলে গল্পের সাথে যোগ করতে হবে ডেটা। তবে হাতে শুধু গল্প এবং ডেটা থাকলেই হবে না, সেটাকে ইফেক্টিভলি কমিউনিকেট করতে পারতে হবে।

সেটা কীভাবে সম্ভব?

আমার ব্যক্তিগত অভিজ্ঞতা বলে, ধরুন আপনি যদি কোনো ধারণাকে আপনার অরগানাইজেশনকে ঠিকমতো পিচ করতে পারেন এবং অরগানাইজেশন যদি সেই ধারণাকে নিজের মনে করে, অর্থাৎ 'ওনারশিপ' নেয়— তাহলে আপনার কমিউনিকেশন সাকসেসফুল। আপনার কমিউনিকেশন তখনই সবার গ্রহণযোগ্যতা পাবে, যখন সেটার পেছনে ডেটা এবং অসাধারণ স্টোরিটেলিং থাকবে। শুধু পরিসংখ্যান কেউ মনে রাখে না বলে ডেটাকে পরিণত করতে হয় ন্যারেটিভে। যারা ডেটা নিয়ে কাজ করেন, তারা স্বভাবগতভাবে ভালো স্টোরিটেলার নন। তবে গল্পের স্ট্যাটাস, কনফ্লিক্ট এবং রেজল্যুশনের ভিত্তিতে দরকারি ডেটাগুলোকে সামনে আনতে হবে। মনে করতে হবে, সব ডেটা নয়; বরং ইন্টারেস্টিং ডেটাগুলো।

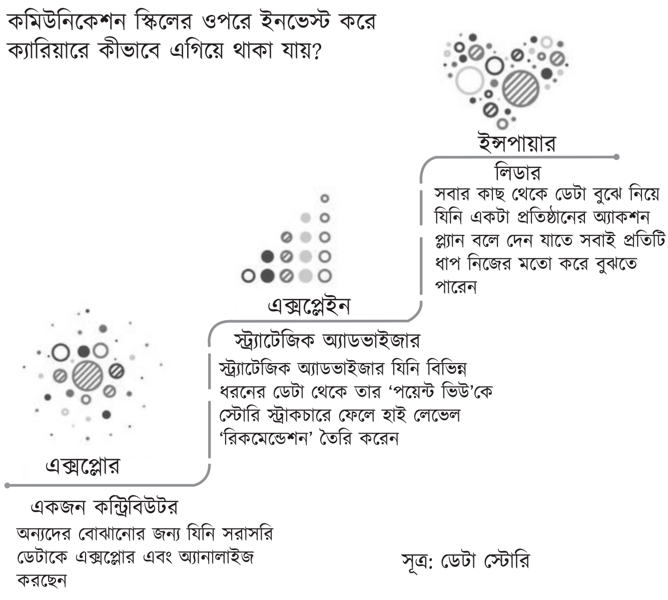
যেকোনো জিনিসের ভাইটাল স্ট্যাটিসটিকস আপনার মনে না থাকলেও সেই ছোটবেলায় পড়া খরগোশ এবং কচ্ছপের গল্প ঠিকই মনে আছে। কেন? স্টোরিটেলিং এবং একটা দর্শন। যখন ডেটা অর্থাৎ পরিসংখ্যানকে একটা গল্পের ন্যারেটিভে ট্রান্সফর্ম করা হয়, তখন সেই কমিউনিকেশন অধিকমাত্রায় গ্রহণযোগ্যতা পায়। সে কারণে আমার বইগুলোতে নিজের এবং আশপাশের গল্পের ছড়াছড়ি।

পুরোনো ডেটা থেকে ধারণা নিয়ে ভবিষ্যৎ দেখার উপলব্ধি তৈরি

ডেটার একটা বড় কাজ হচ্ছে পুরোনো ঘটনাগুলোকে ঠিকমতো ক্যাটালগিং করতে পারা। একটা ভালো সিদ্ধান্ত নেবার জন্য এ ধরনের পুরোনো তথ্য বিশেষ করে হিস্টোরিক্যাল ডেটা খুবই প্রয়োজনীয়। আর যারা ডেটা নিয়ে কাজ করেন তারা সত্য ঘটনা উদঘাটনে বিশ্বাসী। যখন আমরা আস্তে আস্তে একটা লিডারশিপ পজিশনে যাব, তখন এই সত্য ঘটনাকে অবলম্বন করে আমাদের ভবিষ্যৎ দেখতে পারতে হবে। সে কারণে পুরোনো

ডেটা থেকে ধারণা নিয়ে ভবিষ্যৎ দেখার উপলব্ধি তৈরি করে সেটাকে সবার মধ্যে কমিউনেকেট, অর্থাৎ ছড়িয়ে দিতে পারতে হবে। তাহলেই একজন মানুষ লিডারশিপ পজিশনে এগিয়ে যেতে পারবেন। আগের ঘটনা থেকে ঠিকমতো ‘ইনসাইট’ নিয়ে সামনের ডাইরেকশন ঠিকমতো বুঝে সে ধরনের কর্মপন্থা তৈরি করার জন্য আপনাকে কমিউনিকেশন স্কিলে এক্সপার্ট হতে হবে।

ক্যারিয়ারের বিভিন্ন ধাপে কমিউনিকেশনের ব্যবহার



চিত্র ১: ডেটাকে কমিউনিকেট করার কয়েকটা ধাপ

আমি যেটা বুঝি, এই একুশ শতকে— ডেটা দিয়ে কমিউনিকেশন স্কিল ডেভেলপ করার জন্য, ক্যারিয়ারের শুরু থেকেই ডেটা এবং কমিউনিকেশন স্কিলের ওপর মনোযোগ দেওয়া উচিত। শুরুতে শিক্ষানবিস হিসেবে আমরা যখন ডেটা নিয়ে ‘এক্সপ্লোরেশন’ এবং অ্যানালাইজ করতে শুরু করি (যাতে অন্যরা বুঝতে পারেন), তখন আমাদের এই এক্সপ্লোরেশন যাত্রা শুরু হয়। এ ধরনের ডেটা থেকে টিমের অন্যদের আইডিয়া এবং

নিজের 'পয়েন্ট অব ভিউ' থেকে যখন একটা সম্মিলিত 'রিকমেন্ডেশন' তৈরি হয়— তখন সেটাকে 'এক্সপ্লেইন' করার দায়িত্ব 'মিড লেভেল' ম্যানেজমেন্টের ওপরে পড়ে। প্রতিষ্ঠানের সবচেয়ে ওপরের ম্যানেজমেন্ট লিডারদের কাজ হচ্ছে এ ধরনের 'ফিলোসফিক্যাল' আইডিয়া এবং 'পয়েন্ট অব ভিউ'গুলোকে নিজের মতো করে অ্যানালাইজ করে সবার জন্য আলাদা আলাদা অ্যাকশন প্ল্যান ডেভেলপ করে সবার মধ্যে 'কমিউনিকেট' অর্থাৎ ছড়িয়ে দেওয়া। এ ধরনের 'ইন্সপায়ার' করার কাজের জন্য আগের ডেটা থেকে প্রচুর 'ইনসাইট' নেওয়া যেতে পারে। ডেটা থেকে 'ইনসাইট' নিতে ডেটার 'কনটেক্সট' বুঝতে পারতে হবে। ডেটা কোন ভিত্তির ওপর দাঁড়িয়ে কথা বলছে, সেটা বুঝলে এর কনটেক্সট বোঝা যাবে।

কমিউনিকেশনে কনটেক্সটের প্রযোজ্যতা

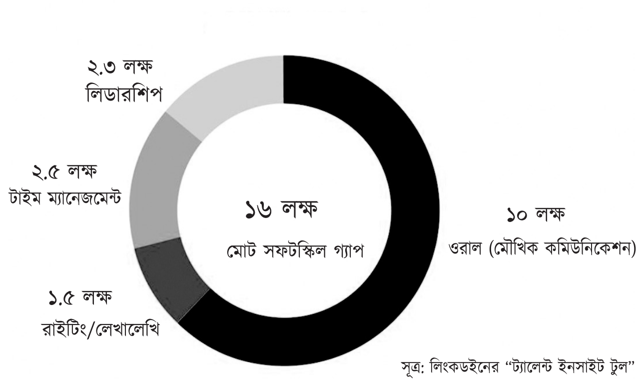
আমি যেটা দেখেছি, আপনার একটা রিকমেন্ডেশন যদি ম্যানেজমেন্টকে ঠিকমতো বোঝাতে হয় তাহলে ঠিক সময়ে ঠিক ডেটা এবং তার জন্য প্রযোজ্য ন্যারেটিভ আপনাকে কয়েক ধাপ এগিয়ে নিতে পারে। ভিজ্যুয়াল এবং মৌখিক অর্থাৎ ওরাল এক্সপ্রেশনে কোন ধরনের কিওয়ার্ড ব্যবহার করতে হবে এবং সেখানে সবার সাথে সহমর্মিতার লেয়ার (অন্যকে বোঝার ক্ষমতা) ব্যাপারটাকে ভালো মাইলেজ দেবে। ডেটাকে ঠিকমতো কমিউনিকেশন মানে এই নয় যে সুন্দর সুন্দর চার্ট এবং মিষ্টি মিষ্টি কথাবার্তা; বরং কনটেক্সটের প্রযোজ্যতাকে ঠিকমতো জানা, কতটুকু তথ্য শেয়ার না করলেই নয়, কার সাথে এবং কোন কনটেক্সটে। এই ব্যালেন্সটা আসে ডেটার সঠিক ব্যবহারের অভিজ্ঞতা থেকে। এই ব্যালেন্সটা তৈরি হবে যখন আমরা পুরো বইটা পড়ব।

কমিউনিকেশন স্কিল তৈরিতে নিজেকে ইনভেস্ট করা

লিংকডইনের 'ট্যালেন্ট ইনসাইট টুল'-এর 'ফাইন্ডিংস'

২০১৮ সালের শেষদিকের ঘটনা। জেফ ওয়াইনার, লিংকডইনের 'সিইও' কয়েকটা স্টাডির 'ফাইন্ডিংস' দেখিয়েছিলেন, কীভাবে কোম্পানিগুলো কিছু নির্দিষ্ট সফট-স্কিলের পেছনে ছুটছে। তাদের (লিংকডইনের) নিজস্ব

‘ট্যালেন্ট ইনসাইট টুল’ ব্যবহার করে দেখিয়েছিলেন— যে ধরনের জব ‘ওপেনিংস’ অনলাইনে পোস্ট করা হয়, সেগুলোর চাহিদাকৃত কোন কোন স্কিলগুলো তাদের চাহিদা ঠিকমতো মেটাতে পারছে না। সেগুলোর মধ্যে সবচেয়ে বড় স্কিল গ্যাপ হচ্ছে সফট স্কিল। পরের পৃষ্ঠার ছবিটি দেখুন এই ১.৬ মিলিয়ন সফট-স্কিল গ্যাপের মধ্যে প্রায় ৯৯৩ হাজার প্রার্থীদের মৌখিক মানে ‘ওরাল কমিউনিকেশন’ স্কিল পায়নি কোম্পানিগুলো। ভয়ংকর, তাই না? এর পাশাপাশি, এই সফট-স্কিল গ্যাপে ১৪০ হাজার পোস্ট যেখানে ‘রাইটিং’ অর্থাৎ লেখালেখির স্কিলের মানুষ পায়নি তারা। এর পাশাপাশি, ‘লিডারশিপ’ এবং ‘টাইম ম্যানেজমেন্ট’ নিয়েও একটা বড় মার্কেট গ্যাপ আছে এখনো।

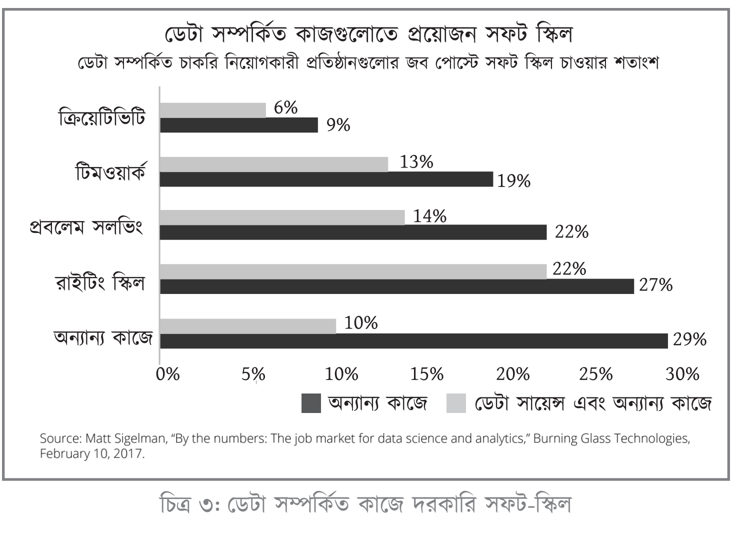


চিত্র ২: লিংকডইনের ‘ট্যালেন্ট ইনসাইট টুল’-এর ‘ফাইন্ডিংস’

তবে মজার কথা হচ্ছে, এই সবকিছুর সাথে ডেটা অ্যানালাইসিসের একটা ভালো সম্পর্ক আছে। কোম্পানিগুলো থেকে হায়ারিং ম্যানেজাররা আশা করেন, প্রার্থীরা যে কাজেই আসুক না কেন, তার ডেটা অ্যানালাইসিস এবং অন্যান্য কমিউনিকেশন স্কিল যাতে সঙ্গে থাকে। ওয়াইনারের বক্তব্য হচ্ছে, যতই নতুন ধরনের একুশ শতকের স্কিলের ডিমান্ড আসুক না কেন, ভালো ধরনের কমিউনিকেশন স্কিল লাগবে সব সময়ে। এই কমিউনিকেশন স্কিল ততটাই ভালো লেভেলে যাবে যখন এর সাথে ডেটাকে ব্যবহার করা হবে। এর অর্থ হচ্ছে ডেটা আপনার অন্যান্য স্কিলকে শক্ত করে দেবে। আগের মতো ‘বুঝ’ দেয়া যাবে না ডেটা ছাড়া।

ডেটা এক্সপ্লোর করার সফট-স্কিল

‘ডেটা এক্সপ্লোরেশন’ একটা বেসিক স্কিলে দাঁড়িয়ে গেছে ইদানীং। ওপরের ঘটনার পাশাপাশি, আইবিএমের রিসার্চের জন্য ‘বার্নিং গ্লাস টেকনোলজি’ বলে একটা কোম্পানি বড় ধরনের স্টাডি করেছিল। সেখানে দেখা গেছে, কোম্পানিগুলো প্রার্থীদের কাছ থেকে ডেটাকে ঠিকমতো ব্যবহার করার ব্যাপারে বেশকিছু সফট-স্কিল আশা করছে। কোভিডের পরে এই ব্যাপারটা অনেকটাই ‘মাস্ট (must)’ হয়ে গেছে। কোম্পানিগুলো ধারণা করছে, একজন প্রার্থীকে যেই কাজের জন্য নেয়া হোক না কেন— তাদের যাতে ‘ডেটা এক্সপ্লোর’ করার ক্ষমতা থাকে। ডেটা থেকে ধারণা নেবার ক্ষমতা থাকে। কারণ, এখন আশপাশে সবই ডেটা। এই ডেটা এক্সপ্লোর করার ক্ষমতার সাথে ভালো ‘প্রবলেম সলভিং’ ‘ক্রিটিক্যাল থিঙ্কিং’ এবং লেখালেখির স্কিল চলে আসে। ইদানীং এগুলোকেই বেসিক স্কিল হিসেবে ভাবছে সবাই।



চিত্র ৩: ডেটা সম্পর্কিত কাজে দরকারি সফট-স্কিল

আপনি ওপরের ডায়াগ্রামটা দেখুন, শুরুতেই চলে এসেছে ‘ক্রিয়েটিভিটি’, যার অর্থ হচ্ছে ডেটা নিয়ে কাজ করার সময় যত ধরনের সমস্যা আসবে, সেই সমস্যাগুলোকে ‘সলভ’ করতে গিয়ে তাদের অনেকটাই ‘ক্রিয়েটিভ’ হতে হবে। এর সাথে সংযোগ আছে ‘ক্রিটিক্যাল থিঙ্কিং’ এর। যারা

ভালো ‘প্রবলেম সলভার’, তারা জানেন কীভাবে ডেটা থেকে একটা সমস্যার সমাধানের ভেতরে ঢুকতে হয়। তারা জানেন কীভাবে ‘রুট কজ অ্যানালিইসিস’ তাদের জীবনকে সহজ করে তুলবে। সবচেয়ে মজার কথা হচ্ছে, এ ধরনের সফট-স্কিল আমাদের সায়েন্স কোর্সওয়ার্ক, অর্থাৎ যেগুলো আমরা সচরাচর পড়ছি, সেই সনাতন সায়েন্স, ফিন্যান্স এবং পরিসংখ্যান পড়ার সময় সময় শেখানো হয় না; বরং এ ধরনের আইডিয়াগুলো আসে, আশপাশের সমস্যাগুলো কীভাবে সমাধান (সনাতনভাবে নয়) করা হচ্ছে তার ভেতর থেকে। এ ধরনের সফট-স্কিল ডেভেলপ করার জন্য ডেটার ব্যবহার জানা একুশ শতকে জরুরি হয়ে পড়েছে। আমি বলব, ‘এ মাস্ট স্কিল’।

কেন শুরুতে ডেটা অ্যানালাইটিক্স স্কিল প্রয়োজন?

কেন শুরুতে ডেটা অ্যানালিস্ট?

যে জিনিসটা আমি সব সময় বলে থাকি, আমাদের সবাইকে ডেটা সায়েন্টিস্ট হবার দরকার নেই। বরং ডেটা কীভাবে কাজ করে সেটার (বিশেষ করে, বোঝার জন্য) সম্পর্কে কোর-স্কিল থাকলে বর্তমান যুগে ডেটা থেকে সব ধরনের সুবিধা নেয়া সম্ভব। এক্সেলের কথাই ধরুন, বিজনেসের মানুষ প্রোগ্রামিং না করেও কাঁপিয়ে দিচ্ছেন ডেটাকে ব্যবহার করে। আইটি সেক্টর ছাড়া প্রতিটা জায়গায় ডেটার ব্যবহার এতটাই বেড়েছে যে, ছোট একটা মুদির দোকান চালাতে গেলেও ইদানীং ছোটখাটো ফিন্যান্সিয়াল অ্যানালাইসিস ছাড়া এই ব্যবসায় টিকে থাকা অসম্ভব। সবচেয়ে বড় কথা হচ্ছে, আমাদের আশপাশে প্রতিনিয়ত যে ধরনের ডেটা তৈরি হচ্ছে সেটা বোঝার জন্য বিশাল কোনো সফটওয়্যার প্রয়োজন নেই। বেসিক লেভেল থেকে অ্যাডভান্স লেভেল পর্যন্ত মাইক্রোসফট এক্সেল এবং গুগল শিট এখনো অপরিহার্য। আপনি বলুন— আমি কি ভুল বলছি?

'ডিভাইস সেন্ট্রিক' সিস্টেমকে 'ইন্টারপ্রেট' করার মানুষটা কই?

সরকারি এবং বেসরকারি প্রতিটা সংস্থার 'অপারেশন' এবং ব্যবসায়িক সিদ্ধান্তগুলো পুরোটাই নির্ভর করে সেই সংস্থা এবং তার সাথে সম্পর্কিত ডেটা বোঝার ওপরে। বর্তমান কাজের পরিধিতে, ধীরে ধীরে মানুষের সময় কমে যাওয়াতে, বিশ্বে মানুষের ভুল সিদ্ধান্তের ফলে প্রতিষ্ঠানের সময় এবং টাকার অপচয় রোধ করতে ডেটা থেকে জ্ঞান নিয়েই সবাই সিদ্ধান্ত নিচ্ছেন। কারণ, মানুষ আবেগের বশে ভুল সিদ্ধান্ত নেন, সেখানে ডেটাকে ব্যবহার করলে সেটার সম্ভাবনা অনেক কমে আসে। আমি

নিজেই যেহেতু প্রাতিষ্ঠানিকভাবে ভিন্ন ভিন্ন কিছু প্রতিষ্ঠানের নিরাপত্তা, অফিস চালানোর ব্যবস্থাপনা, সেই অফিস সম্পর্কিত বিজনেস ইন্টেলিজেন্স, নিজের সায়েন্টিফিক রিসার্চ নিয়ে কাজ করছি, সেখানে আমরা ডেটা ছাড়া বড় ধরনের সিদ্ধান্ত নিতে পারি না। বর্তমানে সেটা ক্রাইমের পর্যায়ে পড়ে।

আমাদের আশপাশে এত বেশি 'ডিভাইস সেন্ট্রিক' সিস্টেম চলে এসেছে, এখানে প্রতিটা ডিভাইস থেকে যত ধরনের 'লগ' জেনারেট হচ্ছে— সেখানে মানুষের মস্তিষ্ক কুলিয়ে পারছে না। সেগুলো নিয়ে কাজ করার জন্য প্রয়োজন পড়ছে বিজনেস ইন্টেলিজেন্স সফটওয়্যার। যার পেছনে মানুষের মাথার হাই লেভেল আইডিয়ার কাজ প্রয়োজন। ডিভাইসের কাজ ডিভাইস করছে, তবে এই ডিভাইসের ডেটা বোঝার জন্য দরকার মানুষের। এখানে মানুষ যিনি অ্যানালিস্ট হিসেবে কাজ করছেন, তার সবচেয়ে বড় কাজ হচ্ছে এই ডিভাইসের ক্যাপচারকৃত ডেটাকে ঠিকমতো 'ইন্টারপ্রেট' করতে পারা। এবং যখন এটাকে কেউ ঠিকমতো 'ইন্টারপ্রেট' করতে পারছেন না, সেখানে ডেটাকে প্রয়োজনীয় 'ম্যানিপুলেশন' (খারাপ অর্থে নয়; বরং যতটুকু 'মডিফাই' করলে ডেটাটাকে ঠিকমতো বোঝা যাবে) করে তার ভেতর থেকে প্রজ্ঞাটা বের করার প্রয়োজন। এতে বোঝা যাবে সেই ডেটাটা আমাদের কী ধরনের সিদ্ধান্ত নিতে বলছে।

ডেটা লিটারেসি এবং 'কেন' প্রশ্ন?

আমরা যখন ডেটা অ্যানালাইসিস নিয়ে কথা বলতে চাই, তখন এর সমার্থক হিসেবে দুটো শব্দ হিসেবে চলে আসে— 'ডেটা লিটারেসি'। এতে ভয় পাবার কিছু নেই। কারণ আপনি আমি এই ব্যাপারটা সব সময় ব্যবহার করছি প্রতিদিনের কাজে। মনে আছে, আমরা কোভিড ভ্যাকসিন নিয়ে কম আলাপ করিনি, সেটার কিছু প্রশ্ন ফেলছি এখানে। ডেটা অ্যানালাইসিসের প্রথম ধাপ।

১. তিন-চারটা ভ্যাকসিন নিয়ে কোনটার কার্য-দক্ষতা কোথায়, কার 'এফিকেসি' কেমন?

২. ১ ডোজ ভালো নাকি ২ ডোজ ভালো? কারা বাজারে আসছে আগে? সরকার কোনটা কিনছে?

৩. সবচেয়ে কাছের কোন ভ্যাকসিন সেন্টারে গেলে আগে ভ্যাকসিন পাওয়া যাবে?

অথবা,

১. এই মহামারির সময় কী ধরনের ফলমূল, ভেষজ এবং শাকসবজি খেলে শরীরের রোগ প্রতিরোধক্ষমতা বাড়াবে?

২. পরিবারের কোন সদস্যের কী কী লক্ষণ দেখা দিলে, হাসপাতালে নিতে হতে পারে— অথবা বাড়িতে চিকিৎসা নেওয়া যেতে পারে?

এ রকম হাজারো প্রশ্নের মধ্য দিয়ে যেতে হয় আমাদের, প্রতিনিয়ত। বলুন তো, কী করেছি আমরা এ সমস্ত বিষয়ে?

এ সব ক্ষেত্রে আমরা শুরুতেই একটা প্রশ্ন নিয়ে শুরু করি। সেই প্রশ্নের উত্তর পাওয়ার জন্য অনেক তথ্য জোগাড় করি, যাতে আমাদের প্রশ্নের উত্তর তৈরি করা যায়। সবচেয়ে বড় কথা হচ্ছে একটা প্রশ্নের উত্তর খুঁজতে গিয়ে তার সাথে সম্পূরক প্রশ্ন, অর্থাৎ আরও কিছু প্রশ্ন বানাতে হয়েছে— যাতে আমাদের আসল প্রশ্নের উত্তর পাওয়া যায়। সবার সাহায্যে আমরা যখন এই প্রশ্নগুলোর উত্তর নিয়ে সেটাকে সমাধান করেছি, তখন সেই ব্যাপারটিকে জানিয়েছি বাকি সবাইকে, যাতে অন্যদের সেই একই সমস্যায় পড়তে না হয়। এটাই ডেটা লিটারেসি। এ কাজটাই আমরা করছি প্রতিনিয়ত।

তাহলে এর ডেফিনেশন কী দাঁড়াচ্ছে?

Data literacy is the ability to explore, understand, and communicate with data.

ডেটা লিটারেসি কেন দরকার?

প্রতিনিয়ত আমাদের আশপাশের মানুষের জীবনকে পাল্টে দিচ্ছি ডেটার ব্যবহার করে। আগে একেকটা ভ্যাকসিনের ভ্যারিয়েন্ট, বিশেষ করে ড্রাগ ডিসকভারিতে যত সময় লাগত, এখন এক মাসের কম সময়ে একেকটা নিয়ে কাজ করতে পারছে ফার্মাসিউটিক্যাল কোম্পানিগুলো। এই মহামারিতে কীভাবে নতুন দক্ষতায় একটা ব্যবসা টিকিয়ে রাখা

যায়, সেটা নিয়ে চলছে বিস্তর গবেষণা— বিশেষ করে ডেটা নিয়ে। ইন্ডাস্ট্রিয়াল এফিশিয়েন্সি বেড়েছে অনেক গুণ, যাতে কম মানুষকে যুক্ত করে অটোমেশন দিয়ে অনেক কাজ ডেলিভারি করা যায়।

সরকারগুলো ডেটা ব্যবহার করে, কীভাবে তাদের খরচকে সীমিত করে আনছেন— সেটার পেছনে কাজ করছে ডেটা। অন্যান্য জায়গায় খরচ কমিয়ে কীভাবে মানুষের জীবনকে (জনগণের) উন্নত করা যায়, সেটা নিয়ে ভাবছে অনেক দেশের সরকারগুলো। মূল কথা হচ্ছে, ডেটা খুলে দিচ্ছে অনেক সুবিধাজনক পরিস্থিতি। ডেটা আমাদের নতুন করে চিন্তার সুযোগ করে দিচ্ছে।

তাহলে কোথায় শুরু করতে হবে?

To begin making better use of data, consider an important aspect of data literacy, exploration. The key to successful data exploration is the formulation of good questions.

ডেটাকে ঠিকমতো ব্যবহার করতে হলে— ডেটা লিটারেসির যে অংশটা আমাদের হাতে ধরে এগিয়ে নিয়ে যাবে, সেটাকে আমরা বলছি 'এক্সপ্লোরেশন'। তত্ত্বানুসন্ধান। এ ছাড়া এর একটা টেকনিক্যাল নাম আমরা প্রায় ব্যবহার করি, 'এক্সপ্লোরাটরি ডেটা অ্যানালাইসিস', ইডিএ! ধারণাটা পরিষ্কার করতে 'হাতেকলমে মেশিন লার্নিং' বইটা দেখতে পারেন। আমাদের বের হতে হবে প্রশ্ন এবং উত্তরের অন্বেষণে। এর সঙ্গে চলবে বিশ্লেষণ এবং পুঙ্খানুপুঙ্খভাবে অনুসন্ধান।

ধরুন, স্বাতীর মতো আপনি বাগান পরিচর্যা পছন্দ করেন। এর পাশাপাশি, আপনি স্বাস্থ্যসচেতন ও বাজার থেকে প্রচুর শাকসবজি কেনেন। এখন একটা প্রশ্ন হতে পারে, বাগানে শাকসবজি হলে নিজের খাবারে পুষ্টি যোগ করা যায় সরাসরি, তবে সেই সময়টা ইনভেস্ট করার পেছনে সময় দিতে চাইবেন কী? এখন এখানে একটা সম্পূরক প্রশ্ন আসতে পারে, নিজের ছাদে বাগান পরিচর্যা করলে যেভাবে নিজের মানসিক স্বাস্থ্য পরিচর্যা করা যায় (যারা বাগান পরিচর্যা করেন, বাগান পরিচর্যার সেই নিবিড় সময়টাতে পৃথিবীর সব সমস্যা ভুলে যেতে পারেন), সেভাবে সেই বাগানের শাকসবজি শরীরের জন্য ভালো ফলাফল বয়ে নিয়ে আসতে পারে। প্রথম

প্রশ্নটা একটু ‘ওপেন এন্ডেড’ থাকলেও পরের প্রশ্নটা সেটাকে প্রায় গুছিয়ে এনেছে। শারীরিক স্বাস্থ্য এবং নিজের বাগানের শাকসবজি— দুটো ব্যাপারই আপনাকে ভাবিয়েছে। ব্যাপারটা কিছুটা কান টানলে মাথা আসার মতো।

ঠিক প্রশ্নটা করতে পারা

We get wise by asking questions, and even if these are not answered, we get wise, for a well-packed question carries its answer on its back as a snail carries its shell.

— James Stephens

আপনি যখন একটা প্রশ্ন শুরু করবেন, তখন সেই প্রশ্নের লেজ ধরে আরেকটা প্রশ্ন আসতে বাধ্য— যতক্ষণ পর্যন্ত আপনি সেই উত্তরে পরিতুষ্ট হন। আমাদের প্রথম প্রশ্নের ফলোআপ হিসেবে নিজের মানসিক স্বাস্থ্য নিয়ে যে প্রশ্নটা আসবে, সেটার উত্তর দেয়ার জন্য আমরা প্রস্তুত কি না? মানে, আমরা মানসিক স্বাস্থ্য নিয়ে চিন্তা করব কেন? কেন নয়? এ রকম প্রতিনিয়ত আমাদের আশপাশের হাজার প্রশ্ন— কেন বিক্রি কমে যাচ্ছে? কেন অপারেটিং কস্ট বাড়ছে? কেন এই প্রডাক্টটা রিটার্ন হচ্ছে বেশি একটা রিজিওন থেকে? কেন গত দুই মাসে আমাদের প্রফিট কমে গেছে —এ রকম প্রচুর ‘কেন’ ঘুরছে আমাদের আশপাশে। এই ‘কেন’-এর শুরুটা কোথায়?

এই ‘কেন’ প্রশ্নের ‘ফাইভ হোয়াই’ পদ্ধতিটা এসেছে টয়োটা মোটরসের প্রতিষ্ঠাতা সাকিচি টয়োডার কাছ থেকে। বাচ্চাদের বারবার ‘কেন’ প্রশ্নের মতো। ওনার প্রস্তাবনা বেশ গভীরের। ব্যাপারটা এরকম যে— আমাদের যে সমস্যাটাকে (‘হোয়াই’ অফ আ প্রবলেম) আইডেন্টিফাই করছি, যাকে বেশ কয়েকটা ‘কেন’ প্রশ্ন জিজ্ঞাসা করব, তখন তার ফলোআপ প্রতিটা প্রশ্ন এবং তার উত্তর অথবা ব্যাখ্যা দেবার জন্য প্রশ্নোত্তরের একটা পাইপলাইন তৈরি করব। আমাদের শুরুর কৌশলটি যদিও প্রথম ‘কেন’ প্রশ্নোত্তরের জন্য নির্ধারণ হলেও (এবং এর মধ্যে উত্তর পেয়ে গেলেও) সেটার আরোও ভেতরে যাবার সুযোগ নেয়া প্রয়োজন এবং নিতে চাই আমরা।

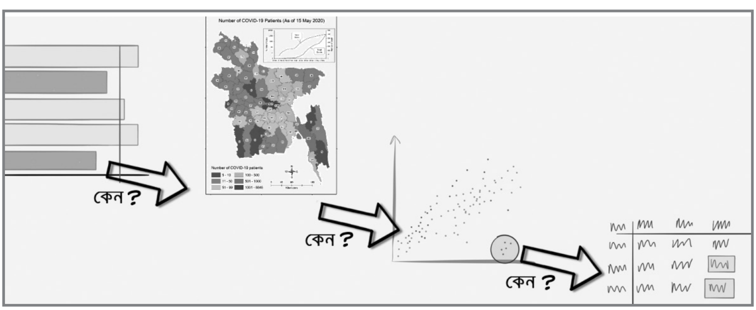
এতে শুধু প্রাথমিক সমস্যা নয়, বরং ভবিষ্যতে যে ধরনের সমস্যাগুলো আসতে পারে, তার ফলাফল এই কয়েকটা 'কেন'তে সমাধান করা সম্ভব। এতে 'রুট কজ অ্যানালাইসিস' ধরে আগানো আরও সহজ। যেমন, 'কেন' এত আগুন ধরছে ফ্যাক্টরিগুলোতে? সেটার সাথে শত শত 'কেন' প্রশ্নের আউটকাম এবং সেটার 'রুট কজ অ্যানালাইসিস' আমাদের ড্যাশবোর্ডের বাইরে তাকাতে সাহায্য করবে। কারণ আমরা জানি, সনাতন ড্যাশবোর্ড বেশি ব্যস্ত থাকে 'হোয়াট'-এর উত্তর খুঁজতে।

শুরুর কাজ, তবে শেষ নয়

Put another way, dashboards should be the start, not the end.

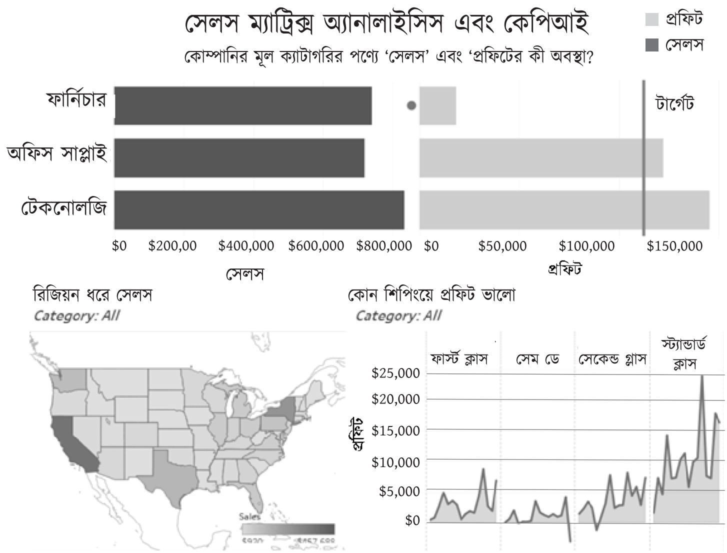
আমার অভিজ্ঞতা বলে, ডেটাকে ঠিকমতো প্রশ্ন করার একমাত্র উপায় হচ্ছে ডেটা ধরে ধরে 'ড্রিল ডাউন' করা, একেকটা করে প্রশ্ন ধরে ধরে। এর মধ্যে, অনেক প্রশ্নের সরাসরি উত্তর না পেলেও সেগুলোর উত্তর আসল প্রশ্নের উত্তরের আশপাশে পৌঁছে বলে, কোনো প্রশ্নই খারাপ নয়। একেকটা 'কেন' প্রশ্নের উত্তরের সাথে আরও অনেক 'কেন' প্রশ্নকে আসতে দিতে হবে। এটা না হলে সমস্যার ভেতরে ঢোকা দুষ্কর। একটা-দুটো 'কেন' প্রশ্ন দিয়ে, হয়তোবা একটা সমস্যার সমাধান ওই মুহূর্তে দেয়া যেতে পারে তবে, 'রুট কজ অ্যানালাইসিস' না করতে পারলে সমস্যাটার চিরস্থায়ী সমাধান সম্ভব নয়। রুট কজ অ্যানালাইসিস নিয়ে আমার একটা ভিডিও দেখুন প্লেলিস্টে, এটাও ঠিক যে অনেক ব্যবসা বেশি 'কেন' শুনতে চান না।

'কেন' প্রশ্নের উত্তর খোঁজার একটা উদাহরণ



চিত্র ৪: কীভাবে অনেকগুলো ভালো প্রশ্নের উত্তর সাহায্য করে ডেটা অ্যানালাইসিসে

চলুন, আপনাকে একটা ‘গাইডেড ট্যুর’ দিই পৃথিবীর সবচেয়ে বেশি ব্যবহৃত ডেটা অ্যানালাইসিস ‘ডেটাসেট’ থেকে। এই ডেটাসেটটা পেয়েছি বহুল ব্যবহৃত ডেটা ভিজ্যুয়ালাইজেশন এবং অ্যানালাইসিস সফট্ওয়্যার ‘ট্যাবলো’ থেকে। মাইক্রোসফট ‘পাওয়ার বিআই’-এর বাংলাদেশে বাজার ভালো, তবে ‘ট্যাবলো’ দিয়ে শেখা সহজ। গ্লোবাল সুপারস্টোর ডেটাসেট, নামে সবাই চেনে, ডেটা অ্যানালাইসিস পৃথিবীতে। এই সাধারণ জিনিসটা বোঝার জন্য আপনাকে ডেটা অ্যানালিস্ট হবার প্রয়োজন নেই। এ ধরনের চার্ট নিয়ে আমরা কাজ করেছি স্কুলে। বার চার্ট। শুরুতেই একটা ড্যাশবোর্ড। কী পারফরম্যান্স ইনডিকেটর (কেপিআই) দেখতে। কেপিআই— যেকোনো জিনিসের ভেতরে ঢুকতে পিলার হিসেবে কাজ করবে। আচ্ছা বলুন তো, যেকোনো কোম্পানির ম্যানেজমেন্ট, শুরুতেই কী জানতে চান? আপনার কোম্পানি হলে কী করতেন?



চিত্র ৫: কোম্পানির মূল ক্যাটাগরি ভিত্তিক ‘সেলস’ এবং ‘প্রফিট’

কত বিক্রি হলো? লাভ হয়েছে কী?

ব্যবসা চলেই— বিশেষ করে, ‘সেলস’ এবং ‘প্রফিট’ নিয়ে।

ছবিটি দেখুন। এই সাধারণ বার চার্টে কোম্পানির মূল ক্যাটাগরিভিত্তিক ‘সেলস’ এবং ‘প্রফিট’ দেখাচ্ছে। যেকোনো কোম্পানির একটা প্রাথমিক

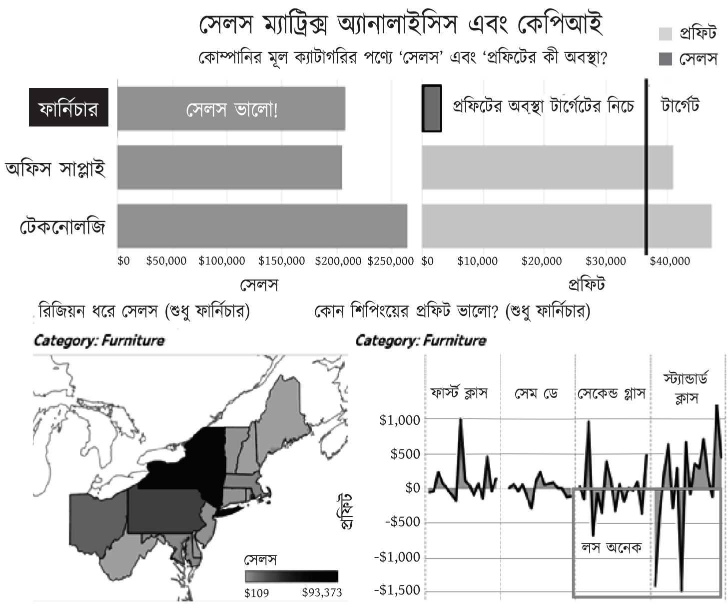
প্রশ্নের উত্তর হচ্ছে এই বার চার্ট। আপনাকে অনুরোধ করব ছবিটি খুঁটিয়ে খুঁটিয়ে দেখতে।

১. একটু এক্সপ্লোর করি, রিজিয়ন ধরে সেলসের কী অবস্থা?

২. কোন শিপিং মোডে 'প্রফিট' বেশি?

তবে আমাদের প্রশ্নের উত্তরে, এই বার চার্টে একটা সমস্যা দেখাচ্ছে। কী হতে পারে সমস্যাটা? আবার ছবিটা দেখুন। আমাদের ফার্নিচার ক্যাটাগরির 'প্রফিট' তার 'টার্গেট' থেকে অনেক নিচে। সেটার পাশে একটা 'ডট' দিয়ে রাখা আছে। এটা একটা বড় সমস্যা। এখন আমাদের কী করা উচিত?

আমরা প্রশ্ন করি— কেন? মানে 'কেন' নিচে?



চিত্র ৬: শুধু ফার্নিচারের 'সেলস' এবং 'প্রফিট'

'কেন' প্রশ্নের উত্তরে আমরা ডেটাসেটের বিভিন্ন 'রিজিয়ন' ধরে 'সেলস' দেখেছি আগে। সফটওয়্যারে ডেটাকে বিভিন্ন ফিল্টারে ফেলে সেলসের একটা ধারণা নেওয়া যায়। কোথায় বেশি হচ্ছে, কোথায় কম হচ্ছে, কোন শিপিং মোডে বেশি প্রফিট হচ্ছে ইত্যাদি ইত্যাদি। নিচের ছবিটা দেখুন। এই ডেটাকে

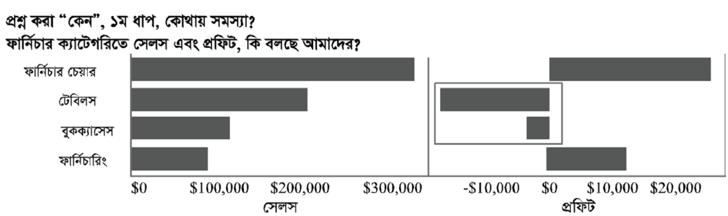
একটা ম্যাপের ওপর প্লট করে দেখতে পারি রিজিওন ধরে কোথায় ‘সেলস’ কমবেশি। লিজেন্ড হিসেবে ‘গাঢ়’ রং মানে বেশি সেলস, ফ্যাকাশে রং মানে সেলস কম। এদিকে ডেটা বলছে, শিপিং মোডে ‘স্ট্যান্ডার্ড ক্লাস’ বেশি প্রফিট করছে।

যতই ডেটা দেখি না কেন, আমাদের মনের মধ্যে খচখচ করছে কেন ফার্নিচারের লাভ এত কম? বিশেষ করে টার্গেট থেকে সেটা অনেক কম। এটা হলে তো কোম্পানি ডুবে যাবে। এ কারণেই আমরা ‘ড্রিল ডাউন’ করছি— কোথায় সমস্যা? বিভিন্ন ধরনের ডেটাকে আমরা যখন প্লটিং করছি, তখন একই প্রশ্নের বিভিন্ন অ্যাঙ্গেল দেখতে পাচ্ছি।

১. ফার্নিচারে সেলস খারাপ নয়। বেশ ভালো।
২. শিপিংয়ে সেকেন্ড ক্লাস এবং স্ট্যান্ডার্ড শিপিংয়ে লস দেখছি। এটা একটা সমস্যার ইন্ডিকেশন।

এ কারণে আমাদের এই ‘রুট কজ অ্যানালাইসিসে’ আমরা আরও একটার পর একটা ডেটাকে প্লট করব। যাতে সমস্যাটা ভালোভাবে বোঝা যায়। এরপর ঢুকব, ফার্নিচার ক্যাটাগরিতে। আরো বুঝতে।

এখন কী করব? আবার প্রশ্ন করব— ‘কেন?’

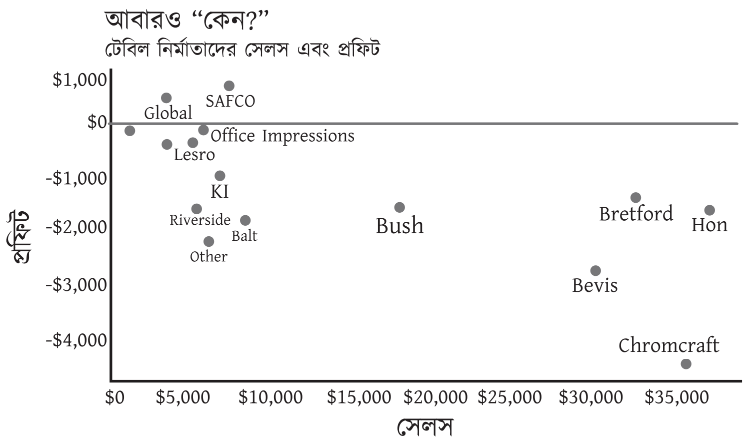


চিত্র ৭: টেবিল এবং বুককেসের ‘সেলস’ ভালো হলেও ‘প্রফিট’ একদম মাটিতে

যেহেতু ফার্নিচারে একটা সমস্যার ‘ইন্ডিকেশন’ পেয়েছি, সে কারণে আমরা ফার্নিচারের ভেতরে ঢুকব। একে ‘ড্রিলিং ডাউন’ বলা যায়। ফার্নিচারের ভেতরের আইটেমগুলোর সেলস এবং প্রফিট পাশাপাশি প্লট করে দেখি। শুধু সেলস দেখলে আমরা হয়তোবা সমস্যাটা বুঝতাম না, কিন্তু প্রফিটের চার্টে তাকাতেই মাথা গরম। ফার্নিচার আইটেম হিসেবে ‘টেবিল’ এবং ‘বুককেস’-এ আমরা টাকা হারাচ্ছি। টেবিলের অবস্থা বেশি খারাপ।

ম্যানেজমেন্ট তো গরম চোখে তাকিয়ে আছেন আমাদের দিকে। আমরা ভয়ে ভয়ে ওনাদের নজর ফার্নিচার চেয়ারের দিকে সরাতে চাইলাম। ওনারা তো সরছেন না।

আবারও প্রশ্ন করি ‘কেন?’



চিত্র ৮: মাত্র দুটো টেবিল নির্মাতা কোম্পানি ছাড়া সবাই লসে

টেবিল নির্মাতাদের ‘সেলস’ এবং ‘প্রফিট’ ইনভেস্টিগেট করতে গিয়ে দেখা গেল টেবিল নির্মাতাদের মধ্যে মাত্র দুটো কোম্পানি লাভ করছে। ছবিটা দেখুন। ‘০’ ডলার প্রফিট লাইনের ওপরে মাত্র দুটি নাম। তাহলে এত টেবিল নির্মাতারা ‘লস’ করছে কেন? এটা তো হবার কথা না।

বোকার মতো প্রশ্ন করি— ‘কেন?’

আমরা টেবিল নির্মাতাদের টেবিলের সেলস ধরে আরও ড্রিল ডাউন করতে গিয়ে দেখা গেল বিভিন্ন শিপিং মোড। সে রকম কিছু পেলাম না। কাস্টমার সেগমেন্টের দিকে তাকালাম, অর্থাৎ তাদের ডেটাকে প্লট করলাম। একই ঘটনা। তবে সেখানে কিছু সমস্যা থাকলেও সেটাকে ‘মাইনর’ হিসেবে ধরে নিচ্ছি। তবে হাল ছেড়ে দিলে চলবে না। আমরা ‘কেন’-এর উত্তর চাই।

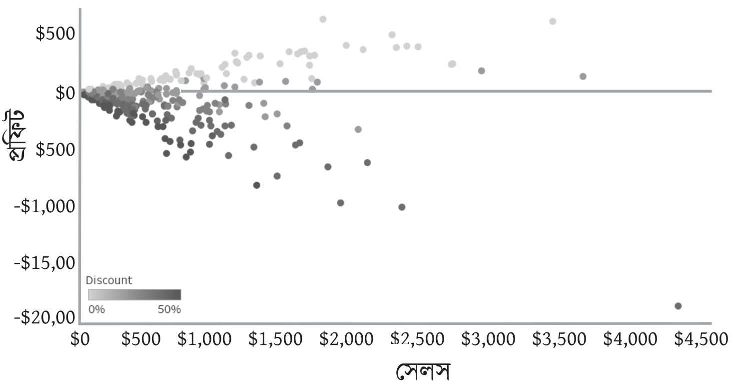
আমরা যখন ‘ডিসকাউন্ট’-এর দিকে তাকালাম, তখন দেখা গেল— আমাদের সেলসের লোকজন এই টেবিলে বেশি ‘সেল’ দেখানোর জন্য

অতিরিক্ত ডিসকাউন্ট দিচ্ছে— যা কোম্পানির প্রফিট মার্জিনকে নামিয়ে আনছে মাটিতে। এখানে কাস্টমার মেসমেন্টও জড়িত।

‘কেন?’ মুখ ফসকে বেরিয়ে গেল

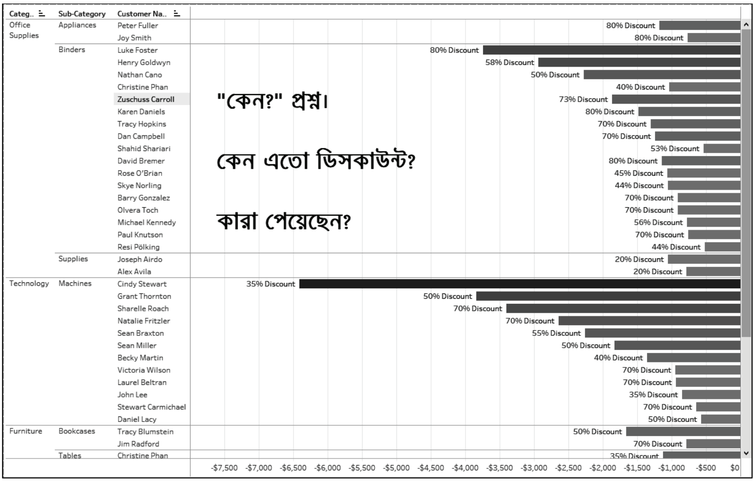
কেন, কেন এবং কেন? ভালো প্রশ্ন করতে জানা

টেবিলের অর্ডার ধরে সেলস এবং প্রফিট অ্যানালাইসিস; গাঢ় কালো রং মানে বেশি ডিসকাউন্ট



চিত্র ৯: ‘সেলস’ এবং ‘প্রফিট’ এর মধ্যে অতিরিক্ত ডিসকাউন্ট একটা বড় সমস্যা

ছবিতে ভালোভাবে দেখলে বুঝতে পারব, যখনই কোনো প্রডাক্টে ডিসকাউন্ট দেওয়া হয়, সেখানে প্রফিট ঋণাত্মক, অর্থাৎ নেগেটিভ চলে যায়। এখানে গাঢ় রং মানে বেশি ডিসকাউন্ট, সেলস এবং প্রফিটের নেগেটিভ দিকে কাজ করছে। কোম্পানির সেলস বাড়ানোর জন্য, অথবা নতুন বাজারে প্রবেশ করার জন্য এ ধরনের ‘ডিসকাউন্টিং’ মডেল শুরুতে কাজ করলেও এর ‘রুট কজ অ্যানালাইসিসে’-এর ভেতরের সমস্যা বের হয়ে এসেছে। এর আরও ভেতরে ঢুকলে আমরা দেখব— কিছু নির্দিষ্ট কাস্টমারও এর সাথে জড়িত। আরও ভেতরে ঢুকতে চাইলে ঢোকা যায়, তবে আমার ধারণা আপনারা বেসিক আন্ডারস্ট্যান্ডিংটা পেয়ে গেছেন এর মধ্যে। বইয়ের পৃষ্ঠা কমাতে বেশি ছবি এখানে আমি আনিনি। তবে ট্যাবলো ব্যবহার করে এই একই ডেটাসেট নিয়ে অনেকগুলো ভিডিও দেখতে পারেন প্লেলিস্টে। সব করতে হবে হাতেকলমে। শুধু দেখলে বোঝা যাবে না।



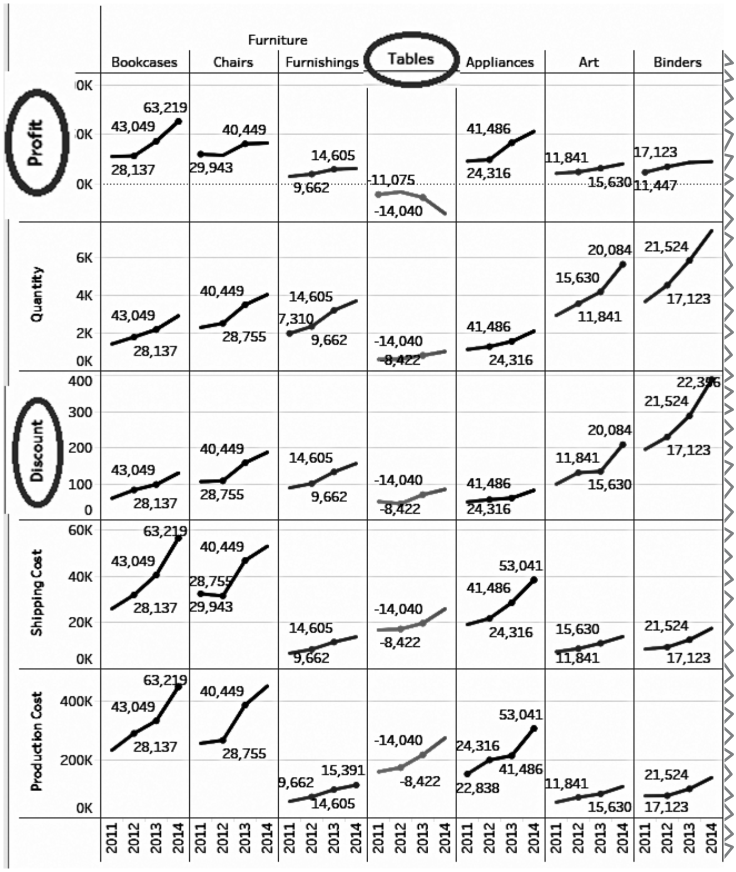
চিত্র ১০: কিছু কাস্টমারের জন্য অতিরিক্ত ডিসকাউন্ট

একটা হোমওয়ার্ক রইল। আপনাকে বলতে হবে নিচের ছবিটা কী বলতে চাইছে? কেন রংটা আলাদা হয়ে গেছে? হিন্টস দিচ্ছি, কালো রঙের ফিচারগুলো দেখুন। আমার ধারণা— বুঝতে পেরেছেন। না পারলে একটু মাথা খাটান। না পারলেও সমস্যা নেই, এর ব্যাপারে একটা বড় ভিডিও আছে। প্লেলিস্টে ট্যাবলোতে নিজে হাতে ফেলতে পারলে আপনার শেখা হয়ে যাবে। ডাউনলোড করতে পারেন পুরো এক্সারসাইজটা নিচের লিংক থেকে, ভবিষ্যৎ কাজে।

এক্সারসাইজ লিংক

https://public.tableau.com/app/profile/rakibul.hassan2768/viz/WhywhyandwhyRootCauseAnalysis/sheet5

সমস্যা ডিসকাউন্ট ক্যাটাগরিতে: একটা হিন্ট দেখুন এখানে



চিত্র ১১: কোথায় সমস্যা? টেবিলে এবং ডিসকাউন্টে

আমি আপনাকে এটুকু বলতে পারি, ডেটা অ্যানালাইসিস এবং গ্লোবাল সুপারস্টোর ডেটাসেট সম্পর্কে ন্যূনতম ধারণা না থাকলেও শুধু এই কয়েকটা ছবি দেখেই ডেটা অ্যানালাইসিসের ব্যাপারটা ধরতে পারবেন। নিজে হাতে করতে চলুন সামনে। আমরা দেখব, কীভাবে 'ট্যাবলো' চালাতে হয়। মূল মেসেজ একটাই, ডেটা অ্যানালাইসিসের জন্য কোনো অনন্য স্কিলসেট প্রয়োজন নেই, ডেটা নিয়ে লেগে থাকলে এই জিনিসটা আয়ত্ত করা সম্ভব।

ম্যানেজারদের জন্য ডেটা অ্যানালাইটিক্স

ডেটাকে ঠিকমতো বোঝা, অর্থাৎ ডেটা অ্যানালাইসিস এখন শুধু ডেটা এক্সপার্টদের হাতে নেই; বরং প্রতিষ্ঠানগুলো আশা করছে তাদের প্রতিটা ম্যানেজার ডেটা অ্যানালাইটিক্স ব্যাপারটা নিজে থেকে বুঝুন। অ্যাটলিস্ট এর বেসিক জিনিসটা। ইদানীং মহামারির সময়— প্রতিষ্ঠানগুলো তাদের দক্ষতা বাড়াতে দুই-তিন জনের কাজ চাপিয়ে দিচ্ছে একজনের কাছে।

আপনি একটা প্রতিষ্ঠানে যেখানেই (যেকোনো লেভেলে) কাজ করুন না কেন, ডেটার ব্যাপারটা ঠিকমতো না বুঝলে ওপরে ওঠার সিঁড়িতে আটকে যাওয়ার আশঙ্কা আছে। ডেটা কী বলছে অথবা আমাদের অ্যানালিস্ট ডেটার কী ধরনের 'এভিডেন্স' আপনাকে দেখাচ্ছে, সেটাকে ঠিকমতো বুঝে কোম্পানি ম্যানেজমেন্টকে ঠিকমতো 'ইন্টারপ্রেট' করানো এখনকার নিত্যপ্রয়োজনীয় স্কিল। প্রতিষ্ঠানের উচ্চপদস্থ কর্মকর্তারা এই স্কিলসেটকে অত্যাবশ্যকীয় বলে ধারণা করছেন ইদানীং। আমরা সারা জীবন যেকোনো জিনিস নিয়ে অ্যানালাইসিস করেছি, তবে এবার সত্যিকারের ডেটার সাহায্যে।

ডেটা = মানুষের অভিজ্ঞতা

আগেও ডেটা ছিল, সেটাকে আমরা বলতাম অভিজ্ঞতা। বয়সের সাথে সাথে আসত সেটা। এখন ডেটাই দিচ্ছে অভিজ্ঞতা, আমাদের এই অল্প বয়সে। এর পাশাপাশি, ওই তথ্যটা কোথা থেকে এল এবং কীভাবে সেটা 'জেনারেট' হয়েছে, সেটার ব্যাকগ্রাউন্ড না জানলে এই কাজে ঠিকমতো অগ্রসর হওয়া যাবে না।

আমরা যেটা দেখেছি, সিদ্ধান্ত নেবার জন্য এ ধরনের অ্যানালাইসিস শুধু নতুন ধারণা দেয় না; বরং এই ব্যাপারগুলো আপনার 'বিকল্প' পছন্দ

ঠিকমতো নির্বাচনে অনেক সাহায্য করে। এই জিনিসগুলো নিয়ে কাজ করতে করতে একসময় আপনার ভেতরের আত্মবিশ্বাস বাড়িয়ে তোলে। কারণ ডেটা থেকে সিদ্ধান্ত নিলে ভুল হয় কম। ঠিকমতো অ্যানালাইসিস করতে পারলে ভুল হবার আশঙ্কা প্রায় শূন্যে ঠেকে। এই ব্যাপারগুলো আমার ব্যক্তিগত অভিজ্ঞতা থেকে বলছি।

দেখুন, প্রতিদিন এত ডেটা আসছে— আমরা কি ধরে নিতে পারি যে, প্রতিটা ডেটাকে আমরা বুঝতে পারছি? অবশ্যই না। এটা সম্ভবও নয়। এর পাশাপাশি যে ডেটাগুলো আসছে— সেগুলোকে আমাদের ডেটা অ্যানালিস্ট এর হাতে ধরিয়ে দিলেই কি সব ফলাফল পাওয়া যাবে? না। কারণ তাদের বিজনেস প্রসেস সম্পর্কে ধারণা কম। এদিকে, আমরা বলতে পারি, আমি ম্যানেজার— আমার কাজ ভিন্ন। ডেটা অ্যানালিস্ট আমাকে যেটা বুঝিয়ে দেবে— আমি সেটাকেই প্রতিষ্ঠানের নীতিনির্ধারণীতে বসে থাকা মানুষের কাছে পৌঁছে দেব, তাহলে ম্যানেজার হিসেবে আমরা কী ভ্যালু যোগ করছি? আমাদের অভিজ্ঞাতার 'কনটেক্সট' কোথায় যোগ হলো?

ধরে নিলাম, আমাদের প্রতিষ্ঠানে সবচেয়ে 'স্মার্ট' ডেটা অ্যানালিস্ট আছেন, তবে সেই জ্ঞান প্রতিষ্ঠানের কাজে লাগবে না যতক্ষণ পর্যন্ত ম্যানেজার হিসেবে আমি নিজে না জানব 'আমাদের ডেটা অ্যানালিস্ট ডেটা নিয়ে কী কাজ করছেন এবং সেটার 'আউটকাম' হিসেবে ওনারা কী বোঝাতে চাচ্ছেন? সবচেয়ে বড় কথা হচ্ছে— ম্যানেজার হিসেবে আমার জানার প্রয়োজন, কীভাবে ডেটা ব্যবহার করতে হয় একটা সিদ্ধান্ত নিতে, কারণ ডেটা তো আর কিছু নয় সংখ্যা ছাড়া। বাকি কনটেক্সট কে বুঝবে? সেই ডেটা অ্যানালিস্ট? আমার মনে হয় না। ডেটা অ্যানালিস্টের কাজ উপাত্ত দিয়ে সেটার সাথে আমাদের বিজনেস প্রসেসের অভিজ্ঞতা যোগ করে একটা 'ডেডলি কম্বিনেশন' করা, যাতে কোম্পানির সিদ্ধান্তে কোনো ভুল না হয়।

ক্রিটিক্যাল থিঙ্কিং এবং প্রবলেম সলভিং স্কিল

আমাদের আশপাশে এত ডেটা জমা হচ্ছে প্রতিনিয়ত, বিশেষ করে সরকারে এবং অন্যান্য প্রতিষ্ঠানে— যা পাল্টে দিচ্ছে আমাদের প্রতিদিনের কাজের ধারণা। আমরা আগে যেভাবে সিদ্ধান্ত নিতাম— এবং এখন যখন ডেটা ব্যবহার করে সিদ্ধান্ত নিচ্ছি, তার মধ্যে প্রচুর ব্যবধান দেখছি প্রতিনিয়ত। কাজের আউটকামে সরকার অথবা ব্যবসাপ্রতিষ্ঠানে একটা ভুল সিদ্ধান্তের ফলে যেভাবে খরচ, মানুষের ভোগান্তি বাড়ে, সময় নষ্ট হয়— সেখানে ডেটা ব্যবহার না করাটাই এখন অপরাধের পর্যায়ে পড়ে যায়।

ক্রিটিক্যাল থিঙ্কিং এবং আমাদের চিন্তাধারা চিন্তা করতে পারা

The essence of the independent mind lies not in what it thinks, but in how it thinks.

— Christopher Hitchens

ডেটা সায়েন্টিস্ট হিসেবে যে কয়েকটা স্কিলসেট দরকার তার মধ্যে—

ক. কিছু অবজেক্টের মধ্যে রিলেশন
খ. প্যাটার্ন বের করতে পারা
গ. ডেটার মধ্য থেকে প্রশ্ন খুঁজে বের করতে পারা
ঘ. ক্রিটিক্যাল থিঙ্কিং
ঙ. প্রবলেম সলভিং স্কিল

সত্যি বলতে, এই স্কিলগুলো আছে সব মানুষের। শিক্ষার ব্যাকগ্রাউন্ড নির্বিশেষে। আমার কাজ সেটাকে ‘ধার’ অর্থাৎ ধারালো করে দেয়া, যাতে আপনি একজন অ্যানালিস্টের চোখ দিয়ে সমস্যাগুলো দেখতে পারেন। সে জন্য এই বইটা। বেসিক গাইডলাইন হিসেবে। বাকি কিছু পাবেন প্লেলিস্টে।

শেখার ধারণা

The greatest experts in any field are those who never forget how much they have to learn.

— Stephen Few, Now You See It

একটা বই পড়ছিলাম কয়েক দিন ধরে। নাউ ইউ সি ইট, বেশ মজার বই। বইটার লেখক তথ্য প্রযুক্তিবিদ স্টিফেন ফিউ মানুষের এমন সেই সব একটি বৈশিষ্ট্যের লিস্ট নিয়ে কথা বলছিলেন, যা ডেটাকে কার্যকরভাবে ব্যবহার করতে সাহায্য করে। একটা মানুষের 'অ্যাপটিটিউডস' অর্থাৎ স্বাভাবিক ক্ষমতা এবং 'অ্যাটিটিউড' অর্থাৎ মনোভাব ঠিক থাকলে এই কাজ করা সম্ভব। এই বৈশিষ্ট্যগুলো আমাদের আরও ভালো প্রশ্ন করতে সাহায্য করবে।

ডেটার বৈশিষ্ট্য নিয়ে বিশাল আলাপ করেছি 'ডেটা নিয়ে বেসিক ধারণা' ও 'ডেটার বৈশিষ্ট্য এবং ভেতরের ধারণা' নামের দুটো চ্যাপ্টারে।

কেন ‘প্রজেক্ট টাইটানিক’ দিয়ে ডেটা অ্যানালাইটিক্স?

কেন ‘প্রজেক্ট টাইটানিক’ একটা বিজনেস ‘প্রক্সি’ ডেটা?

ডেটা সায়েন্স শেখার শুরুতে সবাই প্রজেক্ট টাইটানিক দিয়ে তাদের যাত্রা শুরু করেন। প্রজেক্ট টাইটানিক হচ্ছে ডেটা সায়েন্সের গেটওয়ে প্রজেক্ট। টাইটানিক জাহাজডুবিতে কারা বেঁচে অথবা মারা যাবেন তাদের ভাগ্যকে প্রেডিক্ট করাটাই ছিল এই প্রজেক্টের লক্ষ্য। সেই যাত্রীদের মধ্যে তাদের ডেমোগ্রাফিক ইনফরমেশন, কে কোন শহর থেকে উঠেছেন, তাদের বয়স, তাদের সাথে সন্তান অথবা বাবা-মা উঠেছেন কি না, তারা কোন ক্লাসে ট্রাভেল করছেন, এর পাশাপাশি তাদের কেবিন থাকলে কেবিন নম্বর, যেটা থেকে সবচেয়ে কাছের লাইফবোট এক্সেস কত দূরে হতে পারে, তাদের জেন্ডার ইত্যাদি ইত্যাদি যেকোনো ব্যবসায়িক কাস্টমার ডেটার মতো মনে হবে। আর এই কারণে অনেকে বিজনেস প্রবলেমের সাথে টাইটানিক প্রজেক্টকে সংযোগ করান। এতে একজন যাত্রী, উনি জাহাজডুবিতে বেঁচে যাবেন না মারা যাবেন, তার সাথে বিজনেসের কাস্টমার ডেটার সম্পর্ক রয়েছে।

টাইটানিক প্রজেক্ট নিয়ে পুরো ধারণা নেওয়ার জন্য আপনারা ‘হাতেকলমে মেশিন লার্নিং’ বইটা অনুগ্রহপূর্বক পড়বেন। এতে ডেটা সায়েন্স রাস্তায় হাঁটা সহজ হবে।

ডেটা সায়েন্সের রাস্তায় হাঁটতে গেলে ‘প্রজেক্ট টাইটানিক’ নিয়ে কাজ করা জরুরি। সে কারণে এখানে একটা বেসিক ধারণা দিয়ে রাখছি, যাতে সামনে প্রজেক্টে ঢোকা সহজ হয়। যেমন যেকোনো বিজনেসে কাস্টমার ডেটার মধ্যে থেকে কোন কাস্টমার নতুন একটা প্রডাক্ট অর্ডার করতে পারেন অথবা কোন প্রডাক্টগুলো ফেরত আসতে পারে, এর পাশাপাশি

ভবিষ্যতে কোন প্রডাক্টগুলোর সাথে কোন প্রডাক্টগুলোর সংযোগ একটা কোম্পানিকে আরও নতুন নতুন প্রডাক্ট বিক্রি করতে সহায়তা করতে পারে সেটার একটা ধারণা আসবে এই কাস্টমার ডেটা থেকে। টাইটানিক ডেটাসেটের মধ্যে এ ধরনের ইনফরমেশন অনেকটাই কাস্টমার ডেটা স্পেসিফিক ধারণায় পড়ে যায় বলে টাইটানিক ডেটাসেটটা একটা ভালো বিজনেস প্রক্সি। সুতরাং এটা একটা ভালো স্টার্ট।

কেন আমাদের প্রশ্ন করতে জানতে হবে?

যেকোনো 'বিজনেস কোশ্চেন'-এর উত্তর বের করতে। (আচ্ছা, 'বিজনেস কোশ্চেন'-এর সুন্দর বাংলা কী হতে পারে?)।

অবাক হচ্ছেন? আমরা নিজেদের জন্য, কিংবা ধরুন— সরকার অথবা কোনো কোম্পানির হয়ে যখন কাজ করি, তখন আমাদের প্রতিনিয়ত বিভিন্ন 'বিজনেস কোশ্চেন', অর্থাৎ বিভিন্ন সমস্যার উত্তর খুঁজে বের করতে হয়। যেমন, কী বিষয় নিয়ে পড়াশোনা করলে জীবন অর্থবহ হবে, কোন ক্যারিয়ার আমার জন্য ভালো হবে, কাকে ধরলে একটা বিজনেসের লাইন পাওয়া যাবে— সারা জীবন ধরে এ ধরনের 'বিজনেস কোশ্চেন'-এর উত্তর খুঁজে বেড়াই, নিজের অভিজ্ঞতার কাছে। তবে সেই উত্তর সঠিক হবার সম্ভাবনা থাকে যখন তার উত্তর আসে সরাসরি ডেটা থেকে।

এই কোভিডের সময় সরকারগুলোর একটা বড় 'বিজনেস কোশ্চেন' ছিল প্রকৃত দুস্থদের খুঁজে বের করা, যাতে তাদের সামাজিক নিরাপত্তা তহবিল, অর্থাৎ 'স্টিমুলাস' প্যাকেজগুলোর সঠিক ব্যবহার করা যায়। কাদের আসলেই সহায়তা দরকার সেটা বের করতে দরকার ডেটা। আমাদের দেশেও এই কাজটা হয়েছে ধাপে ধাপে— সামাজিক নিরাপত্তা বেষ্টনীর সুবিধা বণ্টন ব্যবস্থাপনায় বিভিন্ন ভাতা বণ্টন করতে গিয়ে। এত এক্সারসাইজ, তার ফাইনাল আউটকাম একটাই, প্রকৃত দুস্থদের খুঁজে বের করা, যাতে রাষ্ট্রীয় কোষাগারের টাকার উপযুক্ত ব্যবহার হয়। এই কাজের একটা বড় অংশ হয়েছিল আমার প্রতিষ্ঠানে।

একটু পেছনে ফিরে যাই। পৃথিবীতে প্রযুক্তির উৎকর্ষ নিয়ে যে ইন্ডাস্ট্রিতে সবচেয়ে বেশি কাজ হয়েছে, সেটা হচ্ছে 'অ্যাড' ইন্ডাস্ট্রি। অর্থাৎ সেই প্রকৃত ব্যক্তিকে খুঁজে বের করা, যিনি আপনার/আমার প্রডাক্টটা

কিনবেন। মনে আছে সেই এসএমএসগুলোর কথা? আমি পুরুষ/মহিলা, সেটার বাছবিচার না করে বিভিন্ন দোকান থেকে অথবা হাজারো মহিলা প্রডাক্টের এসএমএস দিয়ে ভর্তি হয়ে যেত মোবাইল। এখন সেখানে, এই অ্যাডগুলো এতটাই টার্গেটেড, সেখানে ডেটা বলতে পারে কারা কারা একটা 'স্পেসিফিক' প্রডাক্ট কিনবেন।

আপনি যদি একটা অ্যাড ক্যাম্পেইন চালান, সেখানে ১০০ জন মানুষকে টার্গেট করলে, যদি ৯০ জনকে ওই প্রডাক্টটা বিক্রি করতে পারেন, তাহলে আপনার 'অ্যানালাইসিস' সার্থক। ই-কমার্স কোম্পানিগুলো বহু আগে থেকে আপনি ভবিষ্যতে তাদের কাছ থেকে কী কী কিনতে পারেন সেই প্রডাক্টগুলোকে মজুধ করছে আপনার বাসার কাছের ওয়্যারহাউসে। যেদিন কেনার জন্য ক্লিক করবেন, তার আগের রাতেই ডেলিভারি গাড়ির প্যালেটে উঠে গেছে সেই প্রডাক্ট! বিশ্বাস হচ্ছে না, আমার কথা? গুগল করে দেখতে পারেন, 'অ্যান্টিসিপেটরি শিপিং', যার পেটেন্ট পেয়েছে আমাজন ২০১২ সালে।

আগে ডিজিটাল স্পেসে এ ধরনের অ্যাড ক্যাম্পেইন হতো প্রতি 'ক্লিক' অথবা ইম্প্রেশনের ওপর ভিত্তি করে। এখন ব্যাপারটা এমন হয়ে যাচ্ছে যে আমার একটা নির্দিষ্ট টাকার প্রডাক্ট বিক্রি করে দেবেন, সেটার জন্য কিছু স্পেসিফিক কিওয়ার্ড 'ডেটা অ্যানালাইসিস' থেকে আসবে। অর্থাৎ এর ভেতরে একটা 'বিডিং' বা 'অকশন' হবে যাতে প্ল্যাটফর্ম প্রোভাইডারের সাথে সামান্য কিছু 'রেভিনিউ শেয়ারিং মডেলে' আপনি সেই স্পেসিফিক ক্রেতার কাছে পৌঁছাতে পারেন।

ধরুন, যারা অ্যাড ক্যাম্পেইন অর্থাৎ প্ল্যাটফর্ম চালাচ্ছেন, তারা বলতে পারেন এক লাখ টাকার ৪% ওনাদের দিলে ওনারা ওই ১ লাখ টাকার প্রডাক্ট সেল করে দিতে পারবেন। এখন ওনাদের ডেটা অ্যানালাইসিস করে খুঁজে বের করতে হবে সেই ক্রেতাদের, যারা নিশ্চিতভাবে প্রডাক্টটা কিনবেন। সেটাই আমাদের বিজনেস কোশ্চেনের উত্তর। এর পেছনে কাজ করবে ডেটা অ্যানালাইটিক্স, মেশিন লার্নিং ইত্যাদি। শুরুতে এটা একটা 'কমপ্লেক্স মেট্রিক' মনে হলেও এর ভেতরের গল্প আরও মজাদার, সামনে কথা বলা যাবে এগুলো নিয়ে।

এই ব্যাপারগুলো ‘অ্যাড ক্যাম্পেইন’ দিয়ে শুরু হলেও এই প্রযুক্তি ছড়িয়ে পড়েছে সব জায়গায়। একটা ব্যাংক লোন অ্যাপপ্রুভ করার আগে জানতে চাইতে পারেন সেই লোনটা ‘রিকভার’ করা যাবে কি না? যিনি লোন দেবেন, তিনি চাইবেন সেই লোনটাকে ‘রিকভার’ করে আরেকজনকে দেবার জন্য, এতে তার পোর্টফোলিও বাড়বে। এতে এ ছাড়া ইনভেস্টরদের কাছে ব্যাংকের পারফরম্যান্স ভালোভাবে দেখানো যাবে।

প্রতিটা বড় বড় ছুটি, যেমন— দুই ঈদে যেখানে একটা বিশাল জনসংখ্যা অনেকটাই ‘অরক্ষিত’ অবস্থায় রাস্তায় থাকেন তাদের বাড়িতে পৌঁছানোর জন্য। এই রাস্তায় থাকা সময়টুকুতে দুর্ঘটনা এবং বিভিন্ন অনাকাঙ্ক্ষিত ঘটনা থেকে বাঁচার জন্য ডেটা অ্যানালাইটিক্সের সাহায্য নেয়া যেতে পারে আগেভাগেই। কোথায় কোথায় দুর্ঘটনা এবং অপ্রীতিকর ঘটনাগুলো ঘটে, কী কারণে ঘটে, কী করলে সেই দুর্ঘটনাগুলোকে এড়ানো সম্ভব, সেগুলোর ‘অসাধারণ’ প্রেডিকশন নিয়ে আসা যায় আগের ডেটা অ্যানালাইসিস থেকে। কোন রাস্তায় (বিশেষ করে কোন কোন বাঁকে বা টার্নিংয়ে) দুর্ঘটনা হওয়ার সম্ভাবনা বেশি, কোথায় কোথায় ১ মাইলের বেশি লম্বা জ্যাম হতে পারে, কোন রুটে লঞ্চের চাপ বেশি হতে পারে, কীভাবে টিকিট বিক্রি শুরু করলে এই চাপ এড়ানো সম্ভব, সবকিছুর পেছনে রয়েছে সেই ব্যাপারগুলোকে (‘বিজনেস কোশ্চেন’) ঠিকমতো শনাক্ত করতে পারা। এই প্রেডিকশন দিয়ে প্রচুর সড়ক, নৌপথে দুর্ঘটনাতে প্রাণহানি এড়ানো সম্ভব। কখন রওনা দিয়ে কখন বাড়ি পৌঁছানো সম্ভব, সেটাও আসবে প্রেডিকশনে।

আর ঠিক এই কারণে, পুরো ডেটা সায়েন্স পৃথিবী খুঁজে বের করেছে ১০০ কিলোবাইটের ‘অসাধারণ’ প্রজ্ঞাসহ একটা ডেটাসেট, টাইটানিকের সত্যিকারের ঘটনা নিয়ে। এটার ‘বিজনেস কোশ্চেন’ একটাই, যা মিলে যায় পৃথিবীর অন্যান্য ‘বিজনেস কোশ্চেন’-এর সাথে। এখানে আমরা শনাক্ত করছি টাইটানিক জাহাজডুবি থেকে কে কে বেঁচে যেতে পারেন? (উত্তর না দেখেই)। কোন কোন পারিপার্শ্বিক জিনিসগুলো (বয়স, লিঙ্গ, নাম, কোথা থেকে উঠেছেন, কোন ক্লাসের যাত্রী, টিকিট নম্বর ইত্যাদি ইত্যাদি ফিচার থেকে) তার বেঁচে যাওয়াকে প্রভাবিত করছে?

সরকার জানতে চাইতে পারেন কোন ব্যক্তি/প্রতিষ্ঠানগুলো আগেভাগে ট্যাক্স রিটার্ন জমা দেন, সেই ব্যক্তিগুলোকে ঠিকমতো শনাক্ত করে

প্রণোদনা দিলে সরকারের আয়কর ভালোভাবে আদায় করা সম্ভব। এর পাশাপাশি, কোন ব্যক্তি অথবা প্রতিষ্ঠান এই আয়কর নিয়ে সব সময় সমস্যা করেন তাদের আগেভাগে শনাক্ত করে সম্ভাব্য জালিয়াতি থেকে উদ্ধার পাওয়া যায়। ডেটা অ্যানালাইটিক্স দিয়ে এ ধরনের ট্রেন্ড বোঝা সহজ হয়ে গেছে ইদানীং। বছর ধরে এই বছর কত টাকা ট্যাক্স আয় হবে সেটা বের করা যায় কয়েক বছর আগেই।

পুলিশ জানতে চাইতেই পারেন, কোন এলাকাগুলোতে সন্ত্রাসী কর্মকাণ্ড, অর্থাৎ খুন, রাহাজানি, চাঁদাবাজি, বেআইনি কার্যকলাপ বাড়ছে, যাতে আগে থেকেই সেই এলাকাগুলোতে টহল জোরদার করে পরিস্থিতি নিয়ন্ত্রণে আনা যায় ঘটনা ঘটার আগেই। এর পেছনে কাজ করবে আগের ক্রাইম প্রেডিকশন ডেটা, ডেটা অ্যানালাইটিক্স এবং ডেটা থেকে প্যাটার্ন বের করার ধারণা। আরও জানতে পড়ুন, আমার আগের বই ‘কৃত্রিম বুদ্ধিমত্তায় মানবিক রাষ্ট্র’।

কী বুঝলাম আমরা? এই একই কাজটা আমরা করতে পারছি আমাদের টাইটানিক ডেটাসেট থেকে। এই ডেটাসেটের মাত্র ১২টা ফিল্ড থেকে আমাদের বের করতে হবে এই টাইটানিক জাহাজডুবিতে শেষে কারা বেঁচে গিয়েছিলেন। এটাও একটা ‘প্রক্সি’ ধারণা, যার মাধ্যমে এর আগে আলাপ করা অ্যাড ক্যাম্পেইন, রাজস্ব আদায়, ব্যাংকের লোন রিকভারি, ইলেকশন ক্যাম্পেইন, কার্যকরী ভ্যাকসিন তৈরি, হাসপাতালে ভর্তির সম্ভাবনা ইত্যাদি বের করা সম্ভব। এই ডেটা অ্যানালাইটিক্স দিয়ে কীভাবে নিজের দেশকে আরও ভালোবাসা যায়, সেটা পাবেন কৃত্রিম বুদ্ধিমত্তায় মানবিক রাষ্ট্র বইটাতে। কিনতে হবে না, পুরোটাই আছে অনলাইনে।

ডেটাকে ঠিকমতো ‘নেভিগেট’ করার ধারণা

আমি প্রচুর ভিডিও বানিয়েছি এর মধ্যে। নন প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স শেখার জন্য QR কোড স্ক্যান করে এই প্লে লিস্টটি আপনারা দেখতে পারেন। এই বইয়ের সবগুলো এক্সারসাইজ পাবেন এই প্লেলিস্টে। শতাধিক ভিডিও পাবেন এই প্লেলিস্ট এবং নতুন ভিডিওগুলোতে। তবে বই না পড়ে ভিডিও না দেখার অনুরোধ রইল।

https://www.youtube.com/watch?v=K_XEnWAJO-U&list=PL5NPSn95n_L4Z5bpCOqHW_QU8nxgywGM2

আমি বিশ্বাস করি, মানুষ মাত্রই ডেটা অ্যানালাইসিস করতে পারেন, কারণ আমরা প্রতিনিয়ত সিদ্ধান্ত নিচ্ছি আমাদের আশপাশের ডেটার ইনপুটের ওপর ভিত্তি করে।

ডেটা ভিজ্যুয়ালাইজেশনের কিছু ধারণা

স্বাতীর গল্প

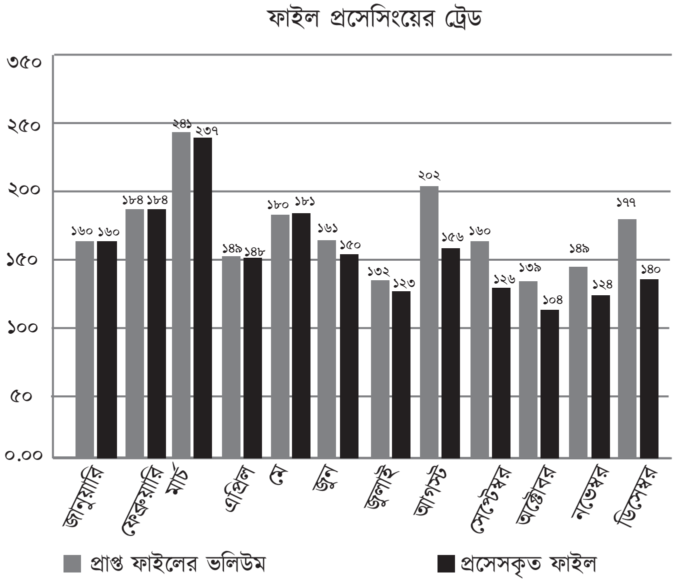
আমার স্ত্রী একজন ব্যাংকার। ২০ বছর ধরে একটা সেক্টরে থাকলে তাকে কি 'জাঁদরেল' ব্যাংকার বলা যেতে পারে? হয়তোবা। আমি সামরিক বাহিনীর অফিসার[১] হলেও দক্ষ প্রাইভেট সেক্টরে কী হয়, এর পাশাপাশি কোন কোন জিনিস আমার স্ত্রী স্বাতীকে চাপে রাখে, সেটার চাক্ষুষ সাক্ষী আমি।

এটা ঠিক যে প্রাইভেট সেক্টরে সুযোগ সুবিধা বেশি থাকার কারণে 'ইন্টার ব্যাংক সুইচিং' একটা খুব সাধারণ ঘটনা। এর বিপরীত ঘটনাও আছে। স্বাতীর টিম থেকে বেশ কয়েকজন অন্য কোথাও চলে গেলে সেখানে নতুন করে রিসোর্স পেতে যে কী সমস্যা সেটা দেখেছি একটা লম্বা সময় ধরে। ভালো সুবিধা পেলে অনেকে চলে যাবেন এটাই স্বাভাবিক। তবে সেই জায়গায় নতুন করে জনবল নিয়োগ দেওয়া একটা সময়সাপেক্ষ ব্যাপার। সেখানে স্বাতীর নাভিশ্বাস উঠে গেলেও ওপরের ম্যানেজমেন্ট এ ব্যাপারে সেভাবে 'রেস্পন্সিভ' থাকেন না। আমিও হতাম না। এমনিতেই কাজ হয়ে গেলে নতুন মানুষ নেয়া ঝামেলার বটে। এখন প্রাইভেট সেক্টরে এসে স্বাতীর বিপদ হাড়ে হাড়ে বুঝছি।

তবে, লম্বা সময় ধরে রিসোর্স না থাকলে সেটা একটা বড় সমস্যা। এটা প্রাইভেট সেক্টর বলে নয়— এটা সব জায়গায় প্রযোজ্য। একটা টিমের মেম্বার কমে গেলে সেখানে নতুন টিম মেম্বার না দিয়ে যতটুকু চালানো যায়। সেটা ম্যানেজমেন্ট ভালো করেই জানেন। কোভিড আমাদের অনেক

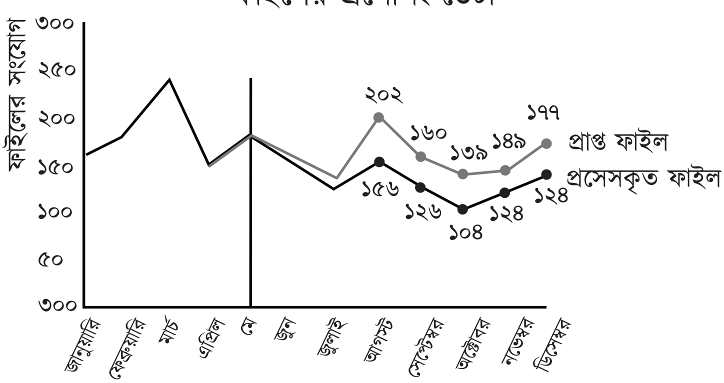
১. বর্তমানে অবসরে চলে এসেছি।

কিছু শিখিয়েছে। (নিচের ছবি দুটো কাল্পনিক, তবে যেকোনো ইন্ডাস্ট্রির জন্য প্রযোজ্য) এই ছবিতে— একটা ব্যাংকের সমস্যা শুরু হয় মে মাস থেকে, যখন দু জন এমপ্লয়ি চলে যান সেই ব্যাংক ছেড়ে।



ডেটার প্লটিং

চিত্র ১২: প্রাপ্ত ফাইলের সাথে প্রসেসকৃত ফাইলের সংখ্যা



চিত্র ১৩: মে মাসের পর থকে প্রসেসকৃত ফাইল কমেছে

এখন নতুন ম্যানেজমেন্টকে কীভাবে নতুন দুজনকে নিয়োগ দেওয়ার ব্যাপারে ভিজ্যুয়ালি সাহায্য করা যেতে পারে? এখানে দুটো চার্ট ব্যবহার করা হয়েছে, যেখানে মে মাস থেকে দুজন টিম মেম্বার ছাড়াই ফাইল প্রসেসিং হচ্ছে। শুরুর দিকে সবাই মিলে চালিয়ে নিলেও জুলাই মাস থেকে সেই প্রেসার নিতে পারেননি বাকি টিম মেম্বাররা। সে কারণেই একটা বড় 'ডাউন-ফল'। ওপরের ম্যানেজমেন্ট ভাবছেন— সব তো চলছে আগের মতো করে।

এখন আসি আমাদের কাজে। ম্যানেজমেন্টের জন্য প্রথম ছবিতে ভিজ্যুয়ালাইজেশন যতটা সফলতা দেবে, সেখানে দ্বিতীয় ছবিটা— নতুন দুজন টিম মেম্বারকে 'এক নিমেষেই' নিয়োগ দেওয়ার জন্য সহায়তা করবে। আপনিও দিতেন। কারণ, মে মাস থেকে ফাইল প্রসেসিং পড়ে থাকছে। সার্ভিস 'ইমপ্যাক্টেড' হচ্ছে। ডেটা ভিজ্যুয়ালাইজেশনের ক্ষমতা এখানেই। এর পরেও কথা আছে।

১. সব ধরনের গ্রাফ, চার্ট সব ক্ষেত্রে প্রযোজ্য নয়।
২. কোন ছবি বা গ্রাফ কোথায় প্রযোজ্য হবে, সে ধরনের ধারণা বোঝার জন্য এই ডেটা অ্যানালাইসিস।

চলুন দেখে আসি, কখন কোন ধরনের ডেটা ভিজ্যুয়ালাইজেশন প্রযোজ্য হবে। ডেটার ভিজ্যুয়ালাইজেশন অনেকটাই আর্ট, তবে এর পেছনে বিশাল 'সায়েন্স' লুকিয়ে আছে। আমরা প্রতিনিয়ত: অনেক ধরনের গ্রাফ, চার্ট, টেবিল, হিট-ম্যাপ, স্ক্যাটার প্লট, বার চার্ট ইত্যাদি ইত্যাদি ব্যবহার করি যাতে ডেটা থেকে প্রাপ্ত 'ইনসাইট' (প্রজ্ঞা) ঠিকমতো আমাদের 'স্টেকহোল্ডার' এবং সিনিয়র ম্যানেজমেন্টে ঠিকমতো পৌঁছানো যায়। যেমনটা আমি দেখেছিলাম স্বাতীর বেলায়।

কিছু বেসিক ডেটা ভিজ্যুয়ালাইজেশন

টেবিল

টেবিল ঠিক একটা ভিজ্যুয়ালাইজেশন টুল নয়, তবে এই টুল থেকে আমরা বেশ কিছু 'কমিউনিকেশন' বুঝতে পারি। এটা কিছুটা আমাদের কথা পড়ে যাবার মতো, যেখানে একটা টেবিলকে আমরা সাধারণত বাংলা স্টাইলে পড়ে যাই। আমরা টেবিল বুঝি না, বরং টেবিলটাকে বাম থেকে

ডানে পড়ে থাকি। বিশেষ করে, যেই সারি এবং কলাম— যার জন্য যেটা প্রযোজ্য। যে লাইনটা যাকে যে ধরনের 'ইনসাইট' দিতে পারে। আমাদের 'ইনডেক্স' আঙুল দিয়ে একটা টেবিলের কলাম অথবা সারি ধরে আমরা এগিয়ে যাই, যাতে সেই কলাম এবং সারিতে দেওয়া তথ্যগুলোকে আমরা বুঝে 'হৃদয়ঙ্গম' করতে পারি। আমরা যাদের একটা টেবিল দেখাব তাদের মধ্যে একেকজন একেকটা সারি অথবা কলাম নিয়ে 'ইন্টারেস্টেড' হবেন বলে একটা টেবিল দিয়ে অনেক সময় কয়েক ধরনের 'অডিয়েন্স'কে বোঝানো যেতে পারে।

মোটা বর্ডার

গ্রুপ ১	মেট্রিক ক	মেট্রিক খ	মেট্রিক গ
গ্রুপ ২	টাকা ১.১	ক %	১০
গ্রুপ ৩	টাকা ১.২	ক %	২০
গ্রুপ ৪	টাকা ১.৩	ক %	৩০
গ্রুপ ৫	টাকা ১.৪	ক %	৪০
গ্রুপ ৫	টাকা ১.৫	ক %	৫০

হালকা বর্ডার

গ্রুপ ১	মেট্রিক ক	মেট্রিক খ	মেট্রিক গ
গ্রুপ ২	টাকা ১.১	ক %	১০
গ্রুপ ৩	টাকা ১.২	ক %	২০
গ্রুপ ৪	টাকা ১.৩	ক %	৩০
গ্রুপ ৫	টাকা ১.৪	ক %	৪০
গ্রুপ ৫	টাকা ১.৫	ক %	৫০

একদম মিনিমালিস্ট বর্ডার

গ্রুপ ১	মেট্রিক ক	মেট্রিক খ	মেট্রিক গ
গ্রুপ ২	টাকা ১.১	ক %	১০
গ্রুপ ৩	টাকা ১.২	ক %	২০
গ্রুপ ৪	টাকা ১.৩	ক %	৩০
গ্রুপ ৫	টাকা ১.৪	ক %	৪০
গ্রুপ ৫	টাকা ১.৫	ক %	৫০

চিত্র ১৪: টেবিল এবং টেবিল বর্ডার

আমরা সামনে অনেক ধরনের ভিজ্যুয়ালাইজেশন দেখব— যেখানে বিভিন্ন গ্রিড ব্যবহার করে ডিজাইনের ব্যাকগ্রাউন্ডে একটা টেবিল থাকতে পারে, তবে সত্যিকারের টেবিল নয়। একটা টেবিলের তথ্য দিয়ে আমরা বিভিন্ন ধরনের ডেটা ভিজ্যুয়ালাইজেশন তৈরি করতে পারি যা সেই একই তথ্য আরও চমৎকারভাবে অন্য মিডিয়াম দিয়ে ঠিকমতো 'কমিউনিকেটেড' হতে পারে। তবে টেবিল যদি ব্যবহার করতেই হয়— সেখানে মোটা সনাতন বর্ডার থেকে হালকা বর্ডার ব্যবহার করা বর্তমানের স্টাইল। মিনিমাল বর্ডার চলছে ইদানীং, তবে এর সাথে আশপাশের প্রেজেন্টেশনও মিনিমালিস্ট হতে হবে। তবে, আমি কখনোই বর্ডার দিতে চাই না, বরং শেড দিয়ে 'কাভার' করে ফেলি।

হিট ম্যাপ

টেবিলের কথা বলতে বলতেই আরেকটা আইডিয়া চলে এল— যা দেখতে অনেকটাই টেবিলের মতো। টেবিল যেমন কিছুটা 'লিখিত লিখিত'ভাবে চলে, সেখানে টেবিলের ভেতরের কনটেন্টের ওপর ভিত্তি করে এমন কিছু শেডের ব্যবহার করা যায় যাতে এর ভেতরের বিভিন্ন ক্যাটাগরিগুলোর গুরুত্ব ক্রমান্বয়ে বেশি বা কম বোঝানো যায়। একটা টেবিলের মধ্যে যেই ডিটেইলসটা মিসিং, সেটাকে ভিজ্যুয়ালি আলাদা করতে পারে এই হিট ম্যাপ।

টেবিলের ভেতরের সংখ্যার পাশাপাশি একটা সেলের রঙের গাঢ়ত্ব দিয়ে প্রতিটা সেলের সাথে আপেক্ষিকতা বোঝানো যেতে পারে। এখানে নিচে পাশাপাশি দুটো টেবিল দেখুন। প্রথম টেবিলে আমাদের তিনটা কলাম আলাদা আলাদা করে পড়ে— এর সাথে সারিকে মিলিয়ে বুঝতে হবে কোথায় এর শতাংশ বেশি। অথচ, পাশের টেবিলে সরাসরি চোখ চলে যাবে সেই কলাম এবং সারির 'ক্রসিং' পয়েন্টে— যেখানে রঙের ঘনত্ব কোথায় কম বা বেশি আছে। মানুষের ব্রেইন রঙের ঘনত্বের ওপর ভিত্তি করে বুঝতে পারে কোথায় গুরুত্ব দিতে হবে। এখানে কালার হলে আরও ভালো বোঝা যায়।

যেকোনো হিট ম্যাপে যেখানে যত বেশি ঘনত্ব— তার আপেক্ষিকতা পারিপার্শ্বিক সেল থেকে অনেক বেশি। ডানের ছবিতে সবচেয়ে ছোট সংখ্যা ১১% এবং বড় সংখ্যা ৫৮%-এর মধ্যে রঙের ঘনত্ব একটা ভালো ধারণা দিচ্ছে ভিজ্যুয়ালি। এক্সেল বা গুগল শিটের মতো এ ধরনের অ্যাপ্লিকেশন 'কন্ডিশনাল ফরম্যাটিং' দিয়ে এ ধরনের ম্যাপ তৈরি করে দেয় এখনই। এর পাশাপাশি 'ট্যাবলো' অথবা 'পাওয়ার বিআই' এ ব্যাপারে একদম সিদ্ধহস্ত।

টেবিল

	ক	খ	গ
ক্যাটাগরি ১	১৫%	২২%	৪২%
ক্যাটাগরি ২	৪০%	৩৬%	২০%
ক্যাটাগরি ৩	৩৫%	১৭%	৩৪%
ক্যাটাগরি ৪	৩০%	২৯%	২৬%
ক্যাটাগরি ৫	৫৫%	৩০%	৫৮%
ক্যাটাগরি ৬	১১ %	২৫%	৪৯%

হিটম্যাপ

কম-বেশি

	ক	খ	গ
ক্যাটাগরি ১	১৫%	২২%	৪২%
ক্যাটাগরি ২	৪০%	৩৬%	২০%
ক্যাটাগরি ৩	৩৫%	১৭%	৩৪%
ক্যাটাগরি ৪	৩০%	২৯%	২৬%
ক্যাটাগরি ৫	৫৫%	৩০%	৫৮%
ক্যাটাগরি ৬	১১ %	২৫%	৪৯%

চিত্র ১৫: শেডের কাজ

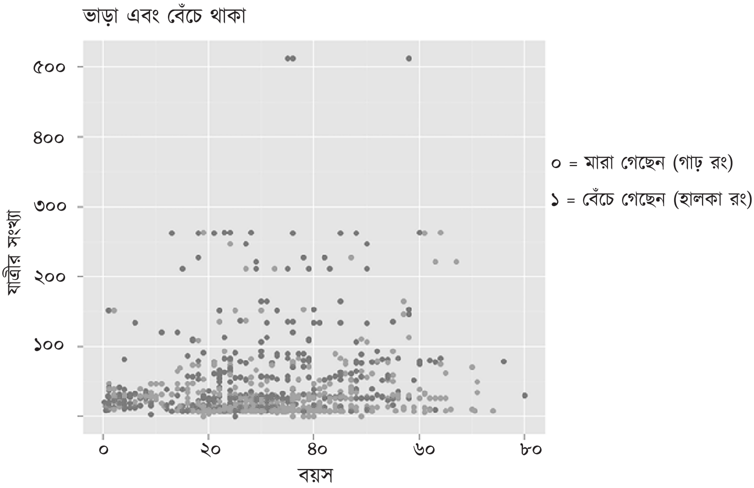
গ্রাফ

টেবিল অথবা হিট ম্যাপ যতটাই আমাদের লেখা থেকে পড়ায় (এখন যেভাবে পড়ছি), সেখানে গ্রাফগুলো সরাসরি আমাদের মাথার ভিজ্যুয়াল সিস্টেমের সাথে কাজ করে। মানুষের মাথার 'প্রি-ফ্রন্টাল' কর্টেক্স একদম শুরুতেই বুঝিয়ে দেয় গ্রাফ সিস্টেম কতটা 'ইফেক্টিভ'। মজার ব্যাপার হচ্ছে, একটা ভালো ডিজাইনের গ্রাফ সাধারণ টেবিল থেকে অনেক তাড়াতাড়ি

তথ্য বের করে দিতে পারে। আমি নিজেই সাধারণত টেবিল থেকে গ্রাফ ব্যবহার করি বেশি। তবে পৃথিবীতে এত এত ধরনের ভালোমন্দ গ্রাফ আছে, সেখানে তলিয়ে যাওয়া কোনো ব্যাপার নয়। তবে সত্যিকার অর্থে দরকারি গ্রাফ হাতে গোনা। বিশ্বাস হচ্ছে না?

স্ক্যাটার প্লট

বিশ্বাস করুন আর নাই করুন, স্ক্যাটার প্লট দিয়ে দুটো ভ্যারিয়েবলের মধ্যে সম্পর্ক বের করা খুবই সহজ। এটা পৃথিবীর সবচেয়ে বেশি ব্যবহৃত প্লট। দুটো ভ্যারিয়েবল থেকে তাদের সম্পর্ক বের করা যে কত বিশাল জিনিস— সেটা না করলেই অনেক কিছু 'মিস'। আবারও বলছি, এটা যেহেতু লম্বালম্বিভাবে এক্স-অ্যাক্সিস (x-axis) এবং ভূমির সমান্তরালে y-axis— দুটো এনকোডেড ডেটার মধ্যে অসাধারণ সম্পর্ক বের করতে পারে, সে কারণে সায়েন্টিফিক জগতের ব্যবহার সবচেয়ে বেশি। তবে, ইদানীং ব্যবসায়িক ঘটনাতেও এই স্ক্যাটার প্লট ব্যবহার হচ্ছে বেশ ভালোই। মনে আছে, আমার প্রথম বই 'হাতেকলমে মেশিন লার্নিং' অংশে টাইটানিক ডেটাসেটের কথা? সেখানে প্যাসেঞ্জার ক্লাসের সাথে মানুষের বেঁচে থাকার সম্ভাবনার একটা ভালো সম্পর্ক পেয়েছিলাম এই স্ক্যাটার প্লট থেকে।

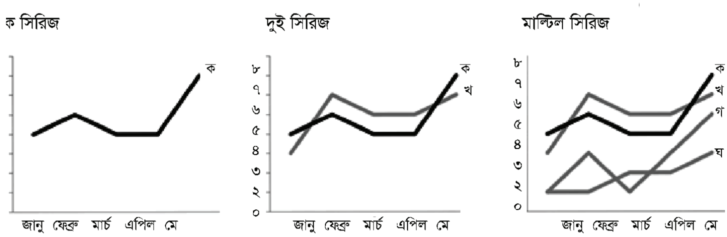


চিত্র ১৬: বয়স, ভাড়া এবং বেঁচে থাকার একটা সম্পর্ক

যেকোনো সম্পর্কের জন্য ‘এক্স এবং ওয়াই অ্যাক্সিসে’ দুটি ভ্যারিয়েবলকে পাশাপাশি রাখলে সেটার ‘ক্রস সেকশন’-এ যে ধরনের ধারণা আসে, সেটা কখনোই অন্য কোনো গ্রাফে পাওয়া যাবে না। যেমন বয়স এবং টিকিটের ভাড়া কীভাবে টাইটানিক যাত্রীদের বেঁচে থাকার ব্যাপারে প্রভাবিত করতে পারে, সেটার একটা ধারণা পাচ্ছি ওপরের ছবিটা থেকে।

লাইন গ্রাফ

আমার মতে এবং আপনারাও দেখেছেন যে লাইন গ্রাফ সবচেয়ে বেশি ব্যবহার হয় যখন আমরা ‘কন্টিনিউয়াস’ ডেটাকে প্লট করতে চাই। যখন পয়েন্টগুলোকে সরাসরি লাইন দিয়ে যুক্ত করি— তখন এর মধ্যে এক পয়েন্ট থেকে আরেক পয়েন্টে যাবার জন্য এর অগ্রগতি বা নিম্নগতি আমাদেরকে ভাবতে শেখায়। মজার ব্যাপার না? এ ব্যাপারগুলো ক্যাটাগরিক্যাল ডেটার জন্য প্রযোজ্য হবে না, তবে যেখানে কন্টিনিউয়াস ডেটা বিভিন্ন সময়ের সাথে চলে— যেমন দিন, মাস, ত্রৈমাসিক বা কোয়ার্টার অথবা বছর, তখন এ ব্যাপারগুলো অন্যভাবে ‘রিপ্রেজেন্ট’ হয়। চলুন, কিছু উদাহরণ দেখে আসি।

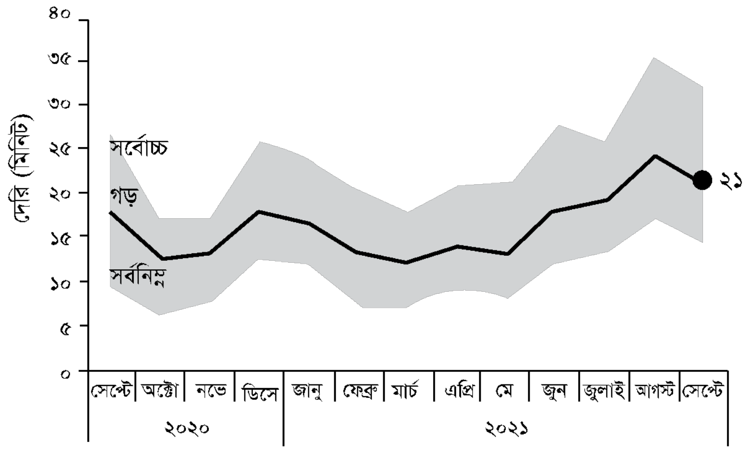


চিত্র ১৭: সময়ের সাথে কিছু জিনিসের পরিবর্তন

যেকোনো লাইন গ্রাফে একটি সিরিজ ডেটা থেকে শুরু করে দুটো অথবা বেশ কিছু সিরিজের ডেটা প্লট করা যায়। তবে এখানে একই টাইম লাইনে কয়েক ধরনের মাত্রা যোগ করে প্লটিং-এ, সেটাই আমাদের কাজে দেয়। ধরুন, আপনি এক বছরের ‘সেলস’-এর ডেটা একটা লাইনে ফেললেও গত ৪৫ বছরের ডেটাও ফেলা যাবে একটি গ্রাফে, অর্থাৎ ‘মাল্টিপল সিরিজ’ ব্যবহার করে। এখানে একটা জিনিস মনে রাখতে হবে, যাতে অনেকগুলো বছর যখন আমরা আপলোড করব, তখন এর

মধ্যে বছরগুলোর সময়ের মধ্যে যে গ্যাপ থাকবে, সেগুলো প্রতিটা বছরের জন্য যাতে একই রকম টাইমলাইনে থাকে। প্রতিবছরের রেফারেন্স যাতে একই লেভেল থেকে শুরু হয়— সেটা দেখবার বিষয়। এখানে বেশ কয়েকটা সিরিজের ডেটার উদাহরণ দেখা দিতে পারে।

পাসপোর্ট কন্ট্রোল অপেক্ষার সময় (ওয়েটিং টাইম)



লাইন গ্রাফে একটা রেঞ্জের মধ্যে গড় (মোট কালো রং)

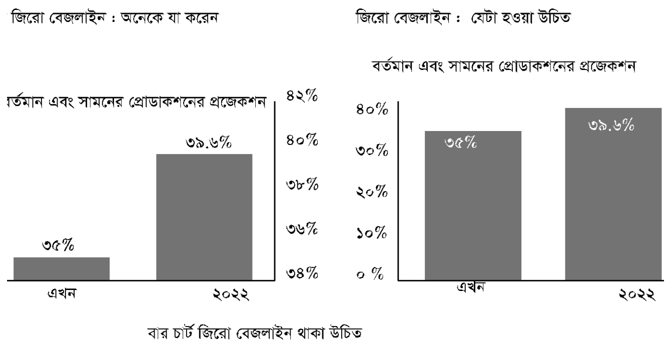
চিত্র ১৮: অপেক্ষার সময়ের সর্বোচ্চ এবং সর্বনিম্ন সময়— গড়

এয়ারপোর্ট দিয়ে আমরা যখন দেশ থেকে বাইরে অথবা দেশের বাইরে থেকে বাংলাদেশে ফিরে আসি, তখন পাসপোর্ট কন্ট্রোলের ওয়েটিং টাইম নিয়ে কিছু ডেটা দেখা যেতে পারে। এই একই ডেটা সিরিজে পাসপোর্ট কন্ট্রোলে সর্বোচ্চ, সর্বনিম্ন এবং সেটার গড় একসাথে লোড করলে একটা ভালো ধারণা পাওয়া যায়। আমরা এখানে ১৩ মাসের একটা ধারণা পেতে পারি। সেপ্টেম্বরে কি বেশি মানুষ আসে? নভেম্বর-ডিসেম্বর একটা ভালো ধারণা দেয়।

বার চার্ট

পৃথিবীর সবচেয়ে সহজতম গ্রাফিং, অর্থাৎ চার্টের মধ্যে ‘বার চার্ট’ খুব জনপ্রিয়। খেয়াল করে দেখুন, যেকোনো পত্রপত্রিকায় এটাই ব্যবহার করা হয়। কারণ একনজরে এই ডেটা থেকে সহজে ‘ইনসাইট’ পাওয়া যায়। এটা খুব সহজেই একটা জিনিস থেকে আর একটা জিনিসের

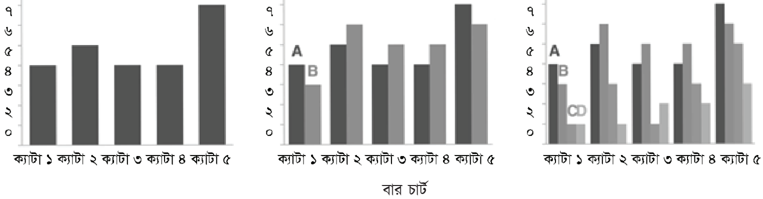
‘আপেক্ষিকতা’, অর্থাৎ কোনটা বড়, কোনটা ছোট বা কোনটার মধ্যে ইনক্রেমেন্টাল ডিফারেন্স, অর্থাৎ বিভিন্ন ক্যাটাগরির মধ্যে কোনটা সবচেয়ে ভালো করছে বা খারাপ করছে— সেটা সহজেই বোঝা যায়। এখানে একটা জিনিস লক্ষণীয় যে সঠিক ডেটা দেখানোর জন্য যেকোনো বার চার্ট এ শূন্যের বেইজলাইন যাতে ঠিকমতো ব্যবহার করা হয়। তবে দৃষ্টিনন্দন বার চার্টগুলোর জন্য নিচের ছবির মতো কিছু গাইডলাইন দেখা করা যেতে পারে। নিচে দেখুন ডানের ছবিটা ঠিক আছে।



চিত্র ১৯: বার চার্ট দিয়ে কোনো একটি জিনিসের ধারণা পাওয়া

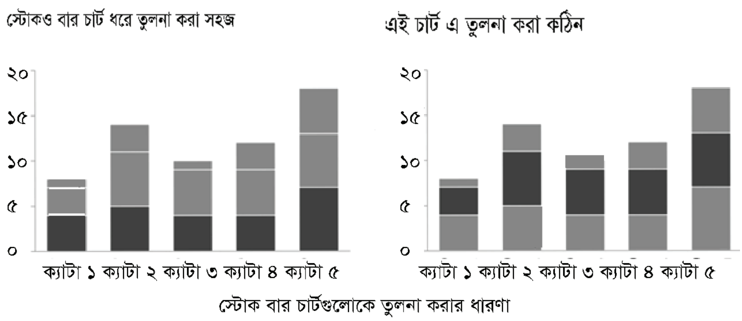
ভার্টিক্যাল বার চার্ট

বার চার্টগুলোর মধ্যে ভার্টিক্যাল বার চার্ট সবচেয়ে বেশি জনপ্রিয়। এতে অনেকগুলো সিরিজ যোগ করা যায়— তবে বেশি সিরিজ যোগ করলে এর ভেতরের কম্প্লেক্সিটি বাড়তে পারে। সে কারণে মোটামুটি মাল্টিপল সিরিজ হলেও তিনের বেশি সিরিজ যোগ করানোটা যুক্তিযুক্ত হবে না। নিচের ছবি দেখুন।



চিত্র ২০: একটার ওপর আরেকটা ‘স্টেক’ ভার্টিক্যাল বার চার্ট

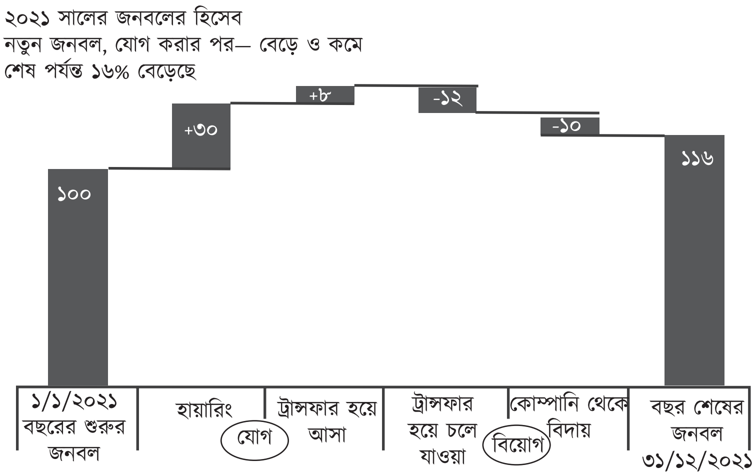
আরেকটা উদাহরণ দেখুন। এটাকে অনেক সময় ‘স্টেক’ ভার্টিক্যাল বার চার্ট বলা হয়, যেখানে একটা লাইনের ওপরে বিভিন্ন ভাগগুলোকে আলাদা করে দেখানো যেতে পারে। সমস্যা একটাই। এটা বোঝার জন্য একটু বেশি সময় প্রয়োজন। প্রথম ধাক্কায়, অনেকে ব্যাপারটা নাও বুঝতে পারেন। সে কারণে এ ধরনের একটার ওপর আরেকটা ভার্টিক্যাল বার চার্ট ব্যবহার করার আগে আপনার স্টেকহোল্ডাররা ঠিকমতো বুঝবেন কি না সেটাও যাচাই করে নেবার প্রয়োজন আছে। নিচের দুটো ছবি দেখুন।



চিত্র ২১: ডানের চার্ট তুলনা করা কঠিন

ওয়াটার ফল চার্ট

এ ধরনের চার্ট অনেকগুলো ধারণাকে এক জায়গায় নিয়ে এসে পাশাপাশি ভালো ‘তুলনামূলক’ জ্ঞান দিতে পারে। তবে সেটা গল্পের মতো করে বললে আরও সুন্দর হয়। বিশেষ করে টাইমলাইনে। একটা জায়গা থেকে শুরু থেকে শেষ পর্যন্ত সিরিজ হিসেবে কয়েকটা ধারণা নিয়ে স্টোরি লাইন তৈরি করা যায়। এটার একটা ভালো উদাহরণ দেওয়া যেতে পারে, একটা ‘হিউম্যান রিসোর্স’ ম্যানেজমেন্ট ঝামেলাকে কেন্দ্র করে। একটা অফিসের শুরুতে যতজন থাকে, তার সঙ্গে নতুন নিয়োগপ্রাপ্তরা কীভাবে যুক্ত হচ্ছে, অন্য ডিপার্টমেন্ট থেকে কতজন আসছে বা অন্য ডিপার্টমেন্টে কতজন চলে যাচ্ছে, কতজন ওই কোম্পানি থেকেই একেবারে চলে যাচ্ছেন এবং একটা সময় পরে সেই কোম্পানিতে কতজন জনবল থাকছেন সেটার একটা হিসেব পাওয়া যেতে পারে এই ‘হেডকাউন্ট’ স্টোরিলাইন দিয়ে। মাথা খারাপ করার কিছু নেই— ছবিটা দেখলেই পুরোপুরি পরিষ্কার হয়ে যাবে।

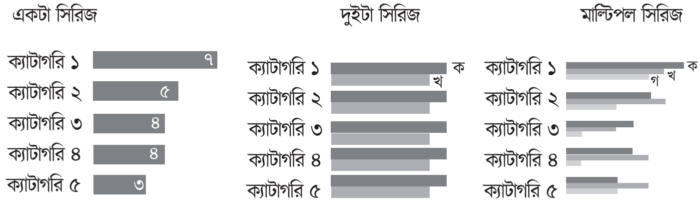


চিত্র ২২: টাইমলাইনের ভিত্তিতে জনবলের হিসেব

হরাইজন্টাল বার চার্ট

ভার্টিক্যাল বার চার্ট থেকে হরাইজন্টাল বার চার্ট পড়তে সোজা। বিশ্বাস হচ্ছে না? এটা আমার কথা নয়; বরং যারা ইন্ডাস্ট্রিতে ডেটা ভিজ্যুয়ালাইজেশন নিয়ে কাজ করেন তাদের সবার এটার ব্যাপারে মতামত পজেটিভ। যেহেতু হরাইজন্টাল বার চার্টগুলো ভূমির সমান্তরালে থাকে— সে কারণে বিভিন্ন ক্যাটাগরির নাম পূর্ণাঙ্গভাবে লেখা যায়। ভার্টিক্যাল বার চার্টে ক্যাটাগরির নামগুলো ছোট জায়গায় থাকে বলে ঠিকমতো লেখা যায় না। তখন এই ধরনের চার্ট খুব ভালো কাজ করে।

মানুষ যেহেতু সাধারণত অক্ষর বাম থেকে ডানদিকে লেখে এবং পরে সেই ধারণায় ব্যাপারটা ভালো সুবিধা দেয়। Y-axis এর সাথে এটার ভালো কানেকশন আছে বলে ওপর থেকে নিচ পর্যন্ত এটাকে ‘সর্টিং’ করা যায় ভালোভাবে। ভার্টিক্যাল বার চার্টের মতো হরাইজন্টাল বার চার্টে একটা, দুইটা অথবা মাল্টিপল সিরিজে বিভিন্ন ক্যাটাগরিতে রিপ্রেজেন্ট করা যায়। এখানে বেশিতেও খুব একটা সমস্যা হয় না।

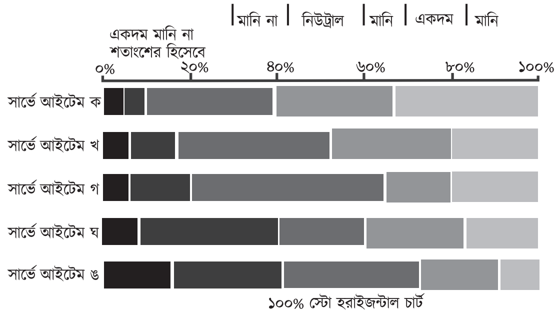


চিত্র ২৩: কয়েকটা সিরিজে ডেটা রিপ্রেজেন্টেশন

‘স্টেকড’ হরাইজন্টাল বার চার্ট

আমরা যখন বিভিন্ন ক্যাটাগরিতে একটা জিনিসকে ১০০ শতাংশের মধ্যে ফেলতে চাইব, তখন এ ধরনের বার চার্ট অসাধারণ কাজ করে। পৃথিবীতে যেটাই করি না কেন সবকিছুই সংখ্যা ‘এক’ অর্থাৎ ১০০% ধরে কোনো কিছু করতে গেলে এ ধরনের ‘স্ট্যাক’ করা ভিজ্যুয়ালাইজেশন অসাধারণ কাজ করে। একটা জিনিসের মধ্যে যত ধরনের সাব-কম্পোনেন্ট থাকবে— সেগুলোর যোগফল ১০০ শতাংশ হিসেবে রিপ্রেজেন্ট করলে ভূমির সমান্তরালে এটাকে দেখতে বেশ সুবিধা হয়।

আমরা সার্ভে কী কারণে করি? একটা বিষয়ে মানুষের মতামত (কেউ কি একমত, অথবা একমত না) নেওয়ার জন্য। নিচে, আমরা একটা সার্ভের রেজাল্টকে প্লটিং করেছি যেখানে যেকোনো একটা বিষয় ১০০%-এর মধ্যে কত শতাংশ ব্যাপারটাকে একদমই মানেননি, মানেননি, নিউট্রাল অবস্থায় ছিলেন, মেনেছেন এবং সম্পূর্ণভাবে মেনেছেন এই কতগুলো ক্যাটাগরিতে আলাদা করে দেখানো হয়েছে। এখানে গ্রে রঙের মাধ্যমে এই পাঁচটা ক্যাটাগরিকে ভাগ করা হয়েছে বলে এগুলোকে সহজেই আলাদা করা যাচ্ছে।

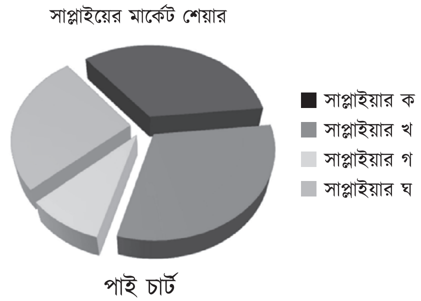


চিত্র ২৪: হরাইজন্টালে সব ধারণা এক লাইনে

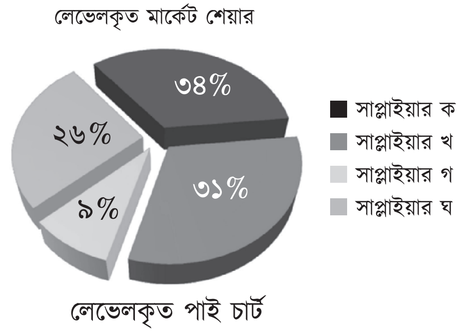
পাই চার্ট

বিশ্বাস করুন, আমি খুব একটা পছন্দ করি না এই চার্টটাকে। পাই চার্টের ভিজ্যুয়ালাইজেশন অনেক সময়— দরকারি মতো কাজের রিফ্লেকশন আনতে পারে না বলে এ ধরনের চার্টগুলোকে ব্যবহার না করলেই ভালো হয়। ডেটাকে ঠিকমতো 'রিপ্রেজেন্ট' করতে গেলে মনস্তাত্ত্বিকভাবে নিচের মতো এ ধরনের ভাগ ঠিকমতো বুঝতে পারে না আমাদের মাথা।

একটা উদাহরণ দেওয়া যাক। নিচের ছবিতে পাঁচটা সাপ্লাইয়ারের মার্কেট শেয়ার যদি আমরা ঠিকমতো 'রিপ্রেজেন্ট' করতে যাই— তাহলে কোন সাপ্লাইয়ারের মার্কেট শেয়ার বেশি মনে হচ্ছে? উত্তরটাকে আপনার মধ্যে রেখে, আমরা এর পরের ছবিতে যাই— যেখানে সাপ্লাইয়ারের মার্কেট শেয়ারগুলোকে আমরা লেবেল করে দিয়েছি। এখন আপনি বুঝতে পারবেন কেন এ ধরনের পাই চার্ট অনেকটাই 'ডিসেপটিভ'। ঠিকমতো মেসেজ দেয় না অনেক সময়ে। খালি চোখে ৩১% বেশি মনে হচ্ছে।

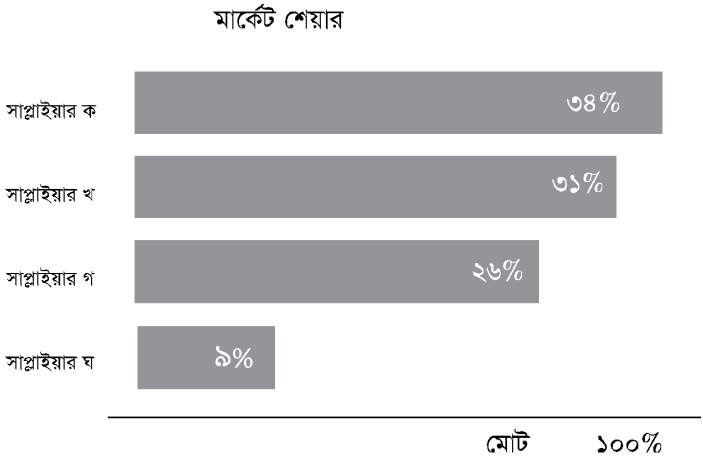


চিত্র ২৫: থ্রিডি, স্টাইলিশ



চিত্র ২৬: কিছুটা 'ডিসেপটিভ'

এই জায়গায় আমরা যদি বার চার্ট ব্যবহার করি, তাহলে এটার রিপ্রেজেন্টেশন খুবই সহজভাবে বোঝা যায়। এর অর্থ হচ্ছে পাই চার্ট থেকে বার চার্ট অনেকটাই 'ভিজ্যুয়ালি' বুঝতে সোজা। আপনি পাই চার্ট ব্যবহার করবেন, না ডোনাট চার্ট ব্যবহার করবেন— সেটা অনেকটাই 'কেইস টু কেইস' অর্থাৎ বিষয়ের ওপরে নির্ভর করে। তবে এ দুটো চার্টকে আপনারা বাদ দিতে পারেন সাধারণ ভিজ্যুয়াল রিপ্রেজেন্টেশন থেকে। আগের ছবি থেকে এই ছবি ভালো বোঝা যায়।



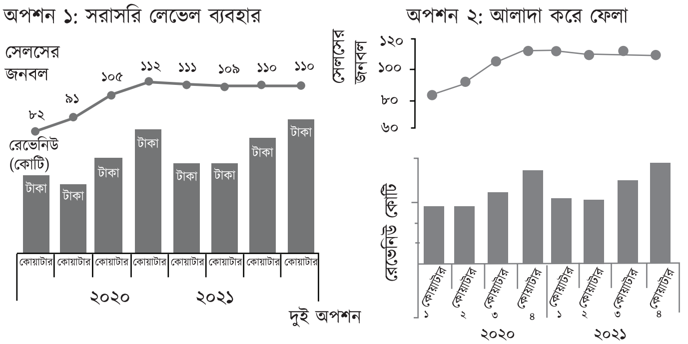
চিত্র ২৭: পাই চার্ট থেকে হরাইজন্টাল বার চার্ট ভালো নয় কী?

যতটুকু পারা যাবে, আমরা 'থ্রি ডাইমেনশন ভিজ্যুয়ালাইজেশন' বাদ দেবার চেষ্টা করব। দেখতে সুন্দর হলেও এর থেকে কাজের আউটপুট কম। বরং মাথায় কম্প্লেক্সিটি বাড়ায় বেশি।

সেকেন্ডারি y-axis রিপ্রেজেন্টেশন

যখন দুটো বা তিনটে ডেটাকে একই অ্যাক্সিসে প্লট করতে হয়— তখন এ ধরনের সেকেন্ডারি y-axis ডেটা ভিজ্যুয়ালাইজেশন ব্যবহার করার প্রয়োজন পড়ে। নিচের ছবি দেখুন। যখন কোনো ধরনের ডেটা ভিজ্যুয়ালাইজেশনে এ জায়গার, অর্থাৎ স্পেসের সংকট থাকে, তখন হয়তোবা এ ধরনের মাল্টিপল সেকেন্ডারি y-axis ডেটা ভিজ্যুয়ালাইজেশনে ব্যবহার করলেও এর ব্যবহারে নতুন কিছু জিনিস মনে রাখা যেতে পারে।

ভালো করে দেখুন। মাল্টিপল সিরিজে ডেটা ‘লোড’ করতে গেলে দুটো বা তিনটে ডেটা সিরিজের জন্য আলাদা আলাদা করে লেবেলিং করলে ব্যাপারটা অনেক সহজ হয়ে আসে। আর, আমরা যদি লেবেলিং না করতে চাই— তাহলে দুটো প্লটকে আলাদা করে আলাদা সেটিং-এ প্লট করা যেতে পারে, যেখানে এই দুটো ছবিকে আলাদা আলাদা চার্ট হিসেবে গণ্য করা যায়।



চিত্র ২৮: অপশন ১: একের মধ্যে দুই

হাতেকলমে ডেটা অ্যানালাইটিক্সের শুরুটা কোথায়?

ডেটা অ্যানালাইসিস/ভিজ্যুয়ালাইজেশনের কিছু উদাহরণ

শুরুতেই ডেফিনেশন। ডেটা এবং তথ্যের সঠিক গ্রাফিক্যাল রিপ্রেজেন্টেশনকে অনেক সময় আমরা 'ডেটা ভিজ্যুয়ালাইজেশন' বলতে পারি। আমাদের দেখা মতে, যেকোনো ভিজ্যুয়াল এলিমেন্ট যেমন চার্ট, গ্রাফ, ম্যাপ ইত্যাদি আমাদের সাহায্য করে একটা স্পেসিফিক ডেটা দেখতে। বিশেষ করে, ট্রেন্ড অ্যানালাইসিস, প্যাটার্ন থেকে ডেটা কী বলতে চাইছে, কোন ধরনের ডেটা আউটলাইয়ার, অর্থাৎ বাকিদের সাথে মিলছে না, অ্যানামলি ডিটেকশন (কোথায় ঝামেলা)– ইত্যাদি আমাদের সাহায্য করে যেকোনো তথ্য থেকে ডেটা ড্রিভেন সিদ্ধান্ত নিতে।

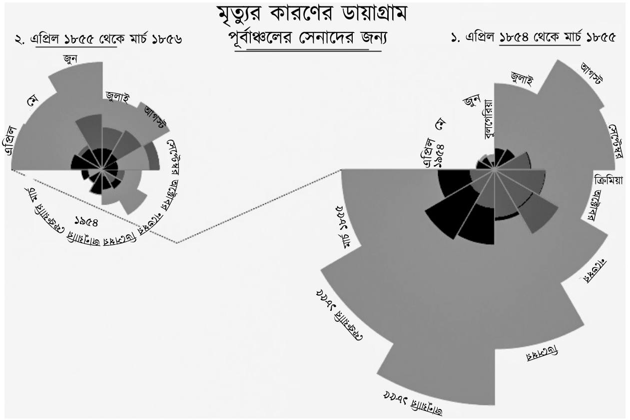
আমরা এখানে কয়েকটা হিস্টোরিক্যাল খুবই সহজ ভিজ্যুয়ালাইজেশন নিয়ে কথা বলব, যা বুঝিয়ে দেবে ডেটা আমাদের বহু বছর আগে থেকেই কীভাবে সাহায্য করছে। শুরুতে ফ্লোরেন্স নাইটিঙ্গেল।

ক্রিমিয়ার যুদ্ধ

১৮৫৩ সালে ক্রিমিয়ার যুদ্ধ শুরু হলেও সেটা শেষ হতে হতে ৫৬ সাল পর্যন্ত লেগেছিল। ইতিহাস ঘেঁটে বলা যায়, যে যুদ্ধে রেলওয়ে, টেলিগ্রাফ, ফটোগ্রাফি এবং হাই এক্সপ্লোসিভ শেল ব্যবহার হয়েছে– সেটাকে বর্তমান কালের যুদ্ধের মধ্যে ফেলা যেতে পারে। এই যুদ্ধের সবচেয়ে বড় সমস্যা ছিল, প্রচুর সৈনিক মারা যাচ্ছিলেন প্রতিদিন, যা পরিসংখ্যানগতভাবে যুক্তরাজ্যের রাজনীতিবিদ, সমরবিদদের বিপদে ফেলছিল। সবাই ভাবছিল, যুদ্ধেই বুঝি মারা যাচ্ছিল এত সৈন্য। তবে সেটার ভেতরের খবর বের করেন ফ্লোরেন্স নাইটিঙ্গেল, তার ডেটার সাহায্য নিয়ে। সেই ডেটা প্রকাশ করতে বিপদে পড়েছিলেন উনি।

ফ্লোরেন্স নাইটিঙ্গেল তার ৩৮ জন সহকর্মী নিয়ে যুদ্ধক্ষেত্রে সৈন্যদলের সেবা-শুশ্রূষার পাশাপাশি ডেটা কালেক্ট করছিলেন— বিশেষ করে, হাসপাতাল সিস্টেমের রিফর্ম নিয়ে নিয়ে কাজ করার জন্য। পরবর্তী সময়ে, সেই দুর্দান্ত 'পোলার এরিয়া ডায়াগ্রাম' ব্যবহার করে নীতিনির্ধারণী পর্যায়ে জানানোর চেষ্টা করলেন— যে যুদ্ধ নয়, বরং যুদ্ধ পরবর্তী সময়গুলোতে 'ফ্রস্ট-বাইট' অর্থাৎ বরফ সংক্রান্ত রোগ, আমাশয়, কলেরা এবং টাইফাস জ্বরে মারা যাচ্ছিল বেশি। উনার পয়েন্ট ছিল— যদিও এটা যুদ্ধক্ষেত্র, তবে মানুষ মারা যাচ্ছে সেই সব রোগে— যে রোগগুলো পুরোপুরি প্রতিরোধযোগ্য। মানে তাদের বাঁচানো যেত। সৈন্যরা মারা যাচ্ছিল কারণ বিশেষ করে, কম্বলের সংখ্যা অপ্রতুল ছিল ওখানে। এর পাশাপাশি, হাসপাতালে হাইজেনিক সমস্যা যেমন ওই জায়গাতে থাকা প্রচুর ইঁদুর এবং মাছি এই প্রতিরোধযোগ্য রোগগুলোকে ছড়াচ্ছিল বেশি।

এ অবস্থায় হাসপাতাল এবং সামরিক বাহিনীর প্রশাসনের সাথে কথা বলার পাশাপাশি ফ্লোরেন্স নাইটিঙ্গেল ক্রিমিয়ার যুদ্ধেই বসে সৈন্যদের দেখভাল এবং নিজস্ব পরিসংখ্যানের জন্য ডেটা কালেকশন শুরু করলেন। তার সঙ্গে ৩২ জন নার্স এবং সামরিক বাহিনীর সাহায্যে বেশ কিছু ডেটা কালেক্ট করে নিজেই চমকে গেলেন। যুদ্ধ শেষে ডেটা থেকে 'ফাইন্ডিংস' যখন পাবলিশ করতে গেলেন, তখন বড় ধরনের একটা প্রতিরোধের সম্মুখীন হয়েছিলেন। তখন তার অ্যানালাইটিক্যাল মাইন্ডসেট থেকে তৈরি করলেন 'পোলার এরিয়া ডায়াগ্রাম' তিনি যা একনজরে যে কাউকে বোঝানোর জন্য অসাধারণ ছিল।



চিত্র ২৯: প্রাণঘাতী রোগ না যুদ্ধ?

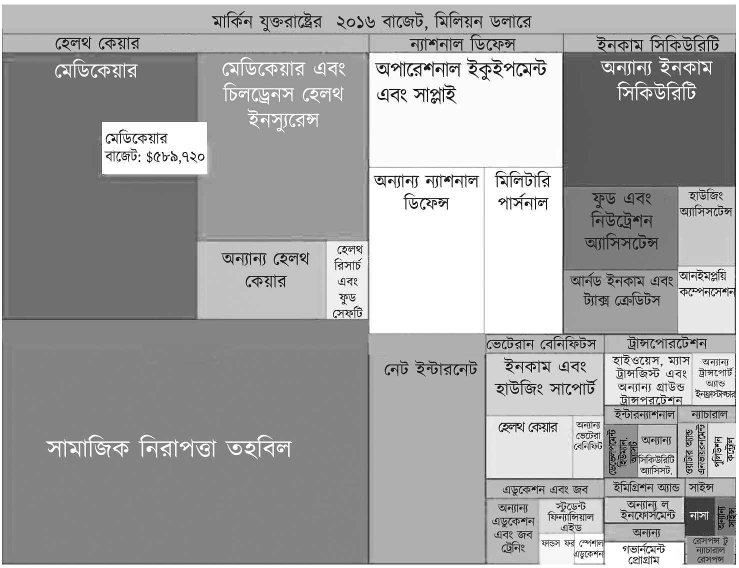
ছবিটা ভালোভাবে দেখুন। প্রথম বছর অর্থাৎ ছবিটার সর্বডানের ডায়াগ্রামের সময়কাল ছিল ১৮৫৪-৫৫ পর্যন্ত। ওই বছরেই ফ্লোরেন্স নাইটিঙ্গেল কাজ শুরু করেন ক্রিমিয়ায়। ডায়াগ্রামে প্রতিটা ত্রিভুজ অংশ হচ্ছে একেকটা মাস। ওই একেকটা ত্রিভুজের ভেতরে অংশগুলোকে বিভিন্ন রং দিয়ে ভাগ করে ফেলা হয়েছে কতজন সৈন্য মারা গেছেন ওই নির্দিষ্ট মাসটাতে। এর মধ্যে সবচেয়ে বড় অংশ, মানে গ্রে রঙের অংশটাতে সেই সব সৈন্যের সংখ্যা মারা গেছে, যারা প্রতিরোধযোগ্য রোগে আক্রান্ত হয়ে মারা গেছেন। যুদ্ধ নয়; বরং ক্রিমিয়া অস্বাস্থ্যকর পরিবেশের মধ্যে সংক্রামক রোগের কারণে তারা মারা গেছেন। এদিকে অল্প, মানে হালকা কালো অংশটা হচ্ছে যারা যুদ্ধক্ষেত্র থেকে প্রাপ্ত ক্ষতের কারণে মারা গেছেন। কালো অংশ হচ্ছে সেই সৈন্যদের সংখ্যা, যারা অন্যান্য কারণে মারা গেছেন।

বামদিকের চার্টটা পরের বছরের। ফ্লোরেন্স নাইটিঙ্গেল আগের বছরের ডেটা ব্যবহার করে যুদ্ধক্ষেত্রে মৃত্যুর সংখ্যা কমিয়ে নিয়ে এসেছিলেন অনেকখানি। দুটো চার্ট একই স্কেলে বলে বোঝা যাচ্ছে কত কম মৃত্যু হয়েছে বামের চার্ট থেকে। শুধু ঠিকমতো ডেটা ব্যবহার করে তিনি কমিয়ে নিয়ে এসেছিলেন প্রতিরোধযোগ্য রোগের মৃত্যুর হার। এই পোলার চার্ট ডায়াগ্রাম দেখে সবাই দুটো জিনিস বুঝতে পারবে। ১. আগের বছরের ডেটা ব্যবহার করে কমিয়ে নিয়ে আসা হয়েছিল অনেক মৃত্যু। ২. যুদ্ধক্ষেত্রের ক্ষত থেকে অধিকাংশ ক্ষেত্রে সংক্রামক প্রতিরোধযোগ্য রোগই মেরে ফেলছিল সৈন্যদের। এই পোলার চার্ট ডায়াগ্রাম শেষ পর্যন্ত ‘পাবলিশ’ হয়েছিল ‘ইংল্যান্ড অ্যান্ড হার সোলজারস’ বইটাতে। তার জন্য ভুগতে হয়েছিল তাকে। শেষমেশ ওই বইটার ধারণা নীতিনির্ধারণী পর্যায়ের সবাইকে বাধ্য করেছিল পরবর্তী যুদ্ধের নীতিমালাগুলো সংস্করণে।

একনজরে মার্কিন যুক্তরাষ্ট্রের বাজেট

এই ম্যাপের ব্যাপারে আমার নিজস্ব কিছু মতামত রয়েছে। নেশা অথবা পেশা হোক, ডেটার ভেতরে ‘কোরিলেশন’ বুঝতে সারাক্ষণ মাথা ব্যাকগ্রাউন্ডে ‘কম্পিউট’ করতে থাকে। আপনারও করে। মানুষের মাথার কাজ একটাই। একটার সাথে আরেকটার তুলনা করা। বিশ্বব্যাপী সরকারি ডেটাকে ঠিকমতো রিপ্রেজেন্ট করতে সংবাদপত্রগুলো বেশ

ভালো কাজ করে। নিউইয়র্ক টাইমসের ২০১৩ সালের বাজেটের চার্ট দেখে চমকে গিয়েছিলাম। বেশ ইন্টারঅ্যাক্টিভ ছিল সেটা। ক্লিক করলে এক্সপ্যান্ড করত বাজেটের হেড ধরে ধরে। বইয়ের কাগজে তো আর সেই ইন্টারঅ্যাকটিভিটি দেখানো যাবে না, তাই একটা সাধারণ ট্রি ম্যাপ।



চিত্র ৩০: 'ট্রি ম্যাপ': একটা বাজেটকে একনজরে আনা

বিশ্বব্যাপী যেকোনো সরকারের বাজেট বোঝা বেশ কষ্টকর, তবে এ ধরনের একটা 'ট্রি ম্যাপ' আমাদের বুঝিয়ে দেয় একনজরে যে কোন কোন খাত সবচেয়ে বেশি টাকা পাচ্ছে। এটা বানানো হয়েছিল— হোয়াইট হাউস থেকে, বারাক ওবামার সময়ে। আমার দেখামতে, একটা অসাধারণ ডেটা ভিজ্যুয়ালাইজেশন, বিশেষ করে বাজেটের মতো কমপ্লেক্স জিনিস দিয়ে। তাও সেটা বোঝা যাচ্ছে একনজরে। এতেই বোঝা যায় যে, একটা কমপ্লেক্স জিনিসকে কত সহজে ভিজ্যুয়ালাইজেশনের মাধ্যমে জনসাধারণকে 'কমিউনিকেট' করা যায়। পৃথিবীর সবচেয়ে কঠিন কাজ হচ্ছে আপনার মেসেজটাকে সবার কাছে ঠিকমতো 'কমিউনিকেট' করতে পারা।

হাতেকলমের কিছু টুল — গিটহাব

গিটহাব ডেটা অ্যানালাইটিক্স টুলকিট, ফর্ক অথবা ডাউনলোড করে নিন আপনার কাজে। এ কারণে বইটার নাম ‘হাতেকলমে’।

আপনারা দেখেছেন, আমি সব সময়ই হাতেকলমে কাজ করতে বিশ্বাসী। আর সে কারণে, অনেকগুলো ডেটাসেট, এক্সারসাইজ করে দেওয়া মাইক্রোসফট এক্সেল, পাওয়ার বিআই করে দেয়া ডকুমেন্টসহ অন্যান্য প্রয়োজনীয় টুল ডাউনলোড করুন নিচের লিংক থেকে। তবে আমার অনুরোধ থাকবে, পুরো লিংকটাকে ফর্ক করে নিন আপনার গিটহাব অ্যাকাউন্টে। ফলে সেখানে ডকুমেন্ট ‘মডিফাই’ করতে বাধা থাকবে না আর।

https://github.com/raqueeb/data_analyst

মাইক্রোসফট এক্সেল, গুগল-শিট, মাইক্রোসফট পাওয়ার বিআই, ট্যাবলো, মাইক্রোসফট এক্সেল অথবা গুগল-শিট নিয়ে সামনে কাজ দেখাব। তবে আমার আইডিয়া হচ্ছে আপনাদের বেশ কিছু অ্যাডভান্সড বিজনেস অ্যানালাইটিক্স টুলের সাথে পরিচয় করে দেয়া— যাতে আমরা পৃথিবীর সাথে টেক্কা দিতে পারি। আমি এখনো মাইক্রোসফট এক্সেল এবং গুগল শিট নিয়ে কাজ করি, তবে যখন আমি ধারণা করি ডেটা আমাদের সারপ্রাইজ দেবে, (কিছুটা, সারপ্রাইজ মি স্টাইলে)— তখন মাইক্রোসফট পাওয়ার বিআই, অথবা ট্যাবলো চালু করি ভবিষ্যৎ দেখার জন্য। আমাদের একটা জিনিস মনে রাখতে হবে, টুল ডিপেন্ডেন্ট হওয়া যাবে না। সব টুলের বেসিক ‘আন্ডার লায়িং প্রিন্সিপাল’ একই। সেটা বুঝলে সব জায়গায় কাজ করা সম্ভব। তবে আমাদের কাছে মাইক্রোসফটের প্রডাক্ট সহজলভ্য বলে আমরা মাইক্রোসফট দিয়ে শুরু করতে পারি। তবে ভবিষ্যৎ অন্য কথা বলবে।

ইদানীং মাইক্রোসফটের সব অ্যাপ্লিকেশন পাওয়া যায় মাইক্রোসফট স্টোরে। অথবা, নিচের এই লিংক নিয়ে যাবে মাইক্রোসফট স্টোরে। মাইক্রোসফট স্টোর দিয়ে পাওয়ার বিআই অ্যাপ্লিকেশন সরাসরি ডাউনলোড এবং ইনস্টল হয়ে যায়। যেহেতু অ্যাপ্লিকেশনটা শেখার জন্য ফ্রি, সে কারণে এই অ্যাপ্লিকেশনটা আপনার পিসিতে রাখার জন্য এ মুহূর্তে একটা মাস্ট লিস্ট।

https://powerbi.microsoft.com/en-us/downloads/

এদিকে বিজনেস অ্যানালাইটিক্স ‘এন্ড টু এন্ড’ সিস্টেমের জন্য ট্যাবলোয়ের নাম বিশ্বখ্যাত। এর ট্যাবলো পাবলিক (ফ্রি) শেখার জন্য দুর্দান্ত জিনিস। শিখতে চাইলে পৃথিবীর সব বড় বড় বিজনেস অ্যানালাইটিক্স সল্যুশন খুলে রেখেছে আমাদের জন্য। শিখতে চাইলে বাধা নেই। সবচেয়ে মজার ব্যাপার হচ্ছে, শিক্ষার্থীদের জন্য এদের লাইসেন্স সহনীয়, চাইলেই পাওয়া যায়। তবে ট্যাবলো পাবলিক দিয়ে শুরুর দিকে চমৎকারভাবে শেখা সম্ভব। নিজের ইমেইল অ্যাড্রেস দিয়ে ডাউনলোড করে ইনস্টল করে নিন এই অসাধারণ টুল। বলতে পারেন, আমার ডেটা অ্যানালাইটিক্সের ভালোবাসা শুরু হয়েছে এই ট্যাবলো দিয়ে।

```
https://public.tableau.com/
```

আরেকটা কথা। যারা আমার মতো ফাঁকিবাজ, তাদের জন্য ট্যাবলো পাবলিক নিয়ে এসেছে ওয়েব ভার্সন। সহনীয় স্পিডের ইন্টারনেট থাকলে, ওয়েব ভার্সন ট্যাবলো পাবলিক দিয়েই চমৎকার কাজ চলে। তবে এই সুবিধা নেবার জন্য আপনাকে একটা প্রোফাইল তৈরি করে নিতে হবে ট্যাবলো পাবলিক সাইটে। নিজের প্রোফাইল হতে পারে—

```
https://public.tableau.com/profile/yourname/
```

আপনার প্রোফাইলে গেলে সবচেয়ে বাম পাশে পাবেন একটা নীল রঙের বাটন। লেখা, [+ Create a Viz]

নিচের কাজগুলো না করে সামনে যাওয়া যাবে না

১. ফর্ক অথবা ডাউনলোড করুন ডেটা অ্যানালাইটিক্স টুলকিট লিংক (আগে করা হয়েছে)

২. চালু করুন ট্যাবলো পাবলিক ডেস্কটপ, অথবা চাপ দিন [+ Create a Viz] বাটনের ওপর, ট্যাবলো পাবলিক ওয়েব ভার্সন (নিজের প্রোফাইল তৈরি করে নিতে হবে আগে) অথবা,

৩. চালু করুন মাইক্রোসফট পাওয়ার বিআই ডেস্কটপ অথবা ওয়েব (ছোট ডেটাসেটের জন্য) ভার্সন— (সহজলভ্য হওয়ায় করপোরেটে জনপ্রিয়)

১, ২ অথবা ৩ — তৈরি তো?

এর পাশাপাশি, যেকোনো স্টার্টআপদের জন্য (যারা নিজের হাতে সব ক্ষমতা চান, কাজ করতে চান একদম গ্রাউন্ড-আপ থেকে) — আমার ডেটায় হাতেখড়ি হয় ওপেনসোর্স টুল দিয়ে।

আমার পছন্দের কিছু ওপেনসোর্স টুল

পাইথনভিত্তিক

১. ডেটাপ্রেপ
২. সুইটভিজ
৩. এক্সপ্লেইনার ড্যাশবোর্ড
৪. প্লটলি এক্সপ্রেস *
৫. ডিথ্রি
৬. আরো পাবেন https://github.com/raqueeb/data_analyst লিঙ্কে।

ওপরের ওপেনসোর্স টুলগুলো নিয়ে প্রচুর ভিডিও বানিয়েছি আমি। কারণ, আমি চাই ডেটার এই ক্ষমতা ছড়িয়ে পড়ুক দেশজুড়ে। নিজেদের ছোট ছোট স্টার্টআপের ডেটা অ্যানালাইসিস আজই শুরু করা যায় এই টুল দিয়ে।

ফুল তুমি কার?

ডেটা অ্যানালাইসিসের প্রথম শর্ত হচ্ছে প্রশ্নের উত্তর খুঁজে বের করা। তবে এর সাথে আর একটা শর্ত হচ্ছে, সেই প্রশ্নের উত্তর খুঁজতে সঠিক প্রশ্ন করতে পারতে হবে। তবে এটাও ঠিক যে সঠিক প্রশ্ন করতে পারার আগে আমাদের অনেক ভুল প্রশ্ন করতে হতে পারে। আর এখানেই আসল মজা। আমি চাই, আপনারা ভুল অথবা শুদ্ধ যা-ই হোক না কেন, আপনাদের প্রশ্ন করতে পারতে হবে। কারণ বলা যায় না, একটা ভুল প্রশ্ন করতে গিয়ে আমরা বেশ কিছু ইনসাইট পেতে পারি। আর এ কারণেই আমরা কখনোই বলব না যে প্রশ্নটা ভুল অথবা শুদ্ধ।

আমরা যে প্রতিষ্ঠানের হয়ে কাজ করি না কেন, প্রতিনিয়ত আমাদের অনেক প্রশ্নের উত্তর খুঁজতে হয়। যেমন আমাদের একটা স্পেসিফিক প্রডাক্ট কেন রিটার্ন হচ্ছে? অথবা, কেন একটা স্পেসিফিক লোকেশন থেকে প্রডাক্টের রিটার্ন বেশি? পুরো বছর ধরে কেন গত দুইমাসে সেলস কমে গেল? কারা আমাদের প্রডাক্টের রেটিং খারাপ দিচ্ছে? এর পেছনের কারণটাই বা কী? কেন এই রাস্তায় দুর্ঘটনা বেশি হচ্ছে? এর পেছনে কী কারণ থাকতে পারে? কেন এই 'স্পেসিফিক' ইনস্যুরেন্স' বেশি ক্লেইম হচ্ছে?

আমাদের প্রশ্নের উত্তর খুঁজতে, শুরুতেই আমরা চেষ্টা করব আমাদের হাতের ডেটাকে প্লট করতে। মজার ব্যাপার হচ্ছে আমরা যখনই ডেটাকে প্লট করি, তখন চমৎকার কিছু ধারণা পাই আমরা। মানুষের মস্তিষ্ক যেহেতু নিউরাল নেটওয়ার্ক দিয়ে তৈরি, সে কারণে যেকোনো ডেটা ঠিকমতো ভিজ্যুয়ালাইজেশন করতে পারলে— এর ভেতর থেকে প্যাটার্ন বের করা সোজা। আর এই প্যাটার্ন বুঝতে পারলে, আগের সব প্রশ্নের উত্তর বের করা সহজ। আর এ কারণেই বিজনেস অ্যানালাইটিক্সের একটা পূর্বশর্ত হচ্ছে ডেটা ঠিকমতো প্লট করতে পারা।

মনে আছে আমাদের কালজয়ী আইরিশ ডেটাসেটের কথা? এই ডেটাসেটকে ব্যবচ্ছেদ করেছিলাম 'শূন্য থেকে পাইথন মেশিন লার্নিং' বইটাতে। আইরিশ ডেটাসেটকে যখন আমরা খালি চোখে দেখেছিলাম, তখন তিনটা ফুলের প্রজাতি কীভাবে একে অপরকে আলাদা করতে পারে— সেটার কূলকিনারা করতে পারছিলাম না। তবে যখন সেটাকে আমরা প্লটিংয়ে ফেলে দিলাম, অর্থাৎ সবগুলো ডেটাকে এক্স এবং ওয়াই অ্যাক্সিসে ভিজ্যুয়ালাইজেশন লেভেলে ফেলে দিলাম, তখনই সব পরিষ্কার হয়ে গেল।

প্রশ্ন: কেন ঘটনাগুলো ঘটছে?

একটা জিনিস চিন্তা করুন, কোন প্রডাক্ট রিটার্ন হচ্ছে, কেন একটা রাস্তায় বেশি দুর্ঘটনা ঘটছে, কেন গত দুই মাস সেলস কমে গিয়েছে, সমাজের কোন পর্যায়ের ক্রেতারা আমাদের প্রডাক্ট খারাপ রেটিং দিচ্ছে, সেগুলোর ডেটা কে ঠিকমতো প্লট করলে সমস্যার কূলকিনারা পাওয়া যাবে। আমি আগেই বলেছি, মানুষের মাথার নিউরাল নেটওয়ার্ক চমৎকারভাবে যেকোনো ডেটার ভিজ্যুয়ালাইজেশন থেকে অসংগতি বুঝতে পারে। আমরা যদি আইরিশ প্রজাতির ফুলের ডেটার প্লট থেকে তিনটা প্রজাতিকে খালি চোখে আলাদা করতে পারি, তাহলে আমরা ডেটা অ্যানালাইসিসের মর্ম বুঝতে পারছি। কেন এটা বা ওটা হচ্ছে, সেটার পেছনে কারণ খুঁজতে গেলে প্রয়োজন ডেটার প্লটিং।

আমরা এর আগের বইগুলোতে 'আর' এবং পাইথন প্রোগ্রামিং এনভায়রনমেন্টে বিভিন্ন লাইব্রেরি ব্যবহার করে এই ভিজ্যুয়ালাইজেশন করেছিলাম। আমি যেহেতু নন প্রোগ্রামারদের এই ডেটা অ্যানালাইসিস ফিল্ডে নিয়ে আসতে চাচ্ছি, সে কারণে নতুন ব্যবস্থা। সময়ের সাথে সাথে মানুষের হাতের টুলগুলো বুদ্ধিমান হয়ে যাচ্ছে। সেই একই কাজ সম্ভব, 'ড্র্যাগ অ্যান্ড ড্রপ' করে— বিশেষ করে, স্মার্ট টুল দিয়ে। মাইক্রোসফট এক্সেল ভালো, তবে পাওয়ার বিআই এবং ট্যাবলোর মতো একটা স্মার্ট টুল কীভাবে কাজ করে সেটা দেখলে ডেটার ব্যাপারে ভক্তি চলে আসবে। এরপর সেটা প্রোগ্রামিং দিয়ে করব কি না— সেটা আপনি বিচার করবেন।

দরকার একটা 'অ্যাডভান্সড টুল'

আমি অনুরোধ করব, এই সুযোগে একটা ভালো 'অ্যাডভান্সড টুল' আমার সাথে সাথে বুঝে ফেলতে। এতে আপনাদের ডেটা অ্যানালাইসিস করার

ভয়টা কেটে যাবে, সব ধরনের টুল নিয়ে; সেটা যতই অ্যাডভান্সড টুল হোক না কেন। আমি শুরুতে একটা 'অ্যাডভান্সড টুল নিয়ে কথা বলতে চাচ্ছি, যাতে আপনারা দেখতে পারেন— ডেটা অর্থাৎ সংখ্যা দিয়ে বানানো একটা কমপ্লেক্স জিনিসকে— কীভাবে ভেঙে ছোট ছোট সমস্যায় রূপান্তর করা যায়। ডেটা আমাদের কী বলতে চাইছে, সেটা বোঝা যাবে এই 'স্টেপ বাই স্টেপ' প্রসেসে।

চলুন, ডেটাকে ব্যবচ্ছেদ করি আমাদের নতুন টুল ট্যাবলো পাবলিক ওয়েব ভার্সন অথবা 'মাইক্রোসফট পাওয়ার বিআই' ওয়েব/ডেস্কটপ ভার্সন দিয়ে। পালাবেন না অনুগ্রহ করে! এই পর্যায়ে যারা দাঁত কামড়ে আমার সাথে লেগে থাকবেন, তাদের জয় নিশ্চিত।

ডাউনলোড করুন আইরিস ডেটাসেট

পাব কোথায়? কোথায় আবার? ক্যাগলে। ডাউনলোড করতে লাগবে গুগল অ্যাকাউন্ট। মাত্র ৫ কিলোবাইট। ফাইল নাম হচ্ছে iris.csv, যা এক্সেল অথবা গুগলশিটে খোলা যায়। তবে আমি বলব ট্যাবলো পাবলিক ওয়েব ভার্সন অথবা 'মাইক্রোসফট পাওয়ার বিআই' ডেস্কটপ দিয়ে খুলতে।

https://www.kaggle.com/uciml/iris

ডেটাসেট দেখতে কেমন?

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

চিত্র ৩১: আইরিস প্রজাতির ডেটা, শূন্য থেকে পাইথন মেশিন লার্নিং বই থেকে

ট্যাবলো পাবলিক ‘ওয়েব ভার্সন’ এবং আইরিস প্রজাতি

আইরিস প্রজাতির ডেটা নিয়ে আমাদের বিজনেস ‘কোশ্চেন’ কি ছিল?

বিজনেস কোশ্চেন

১. আমরা কি খালি চোখে আইরিস প্রজাতির ফুলগুলোকে আলাদা করতে পারব কি না?

২. ধরা যাক, আপনাকে একটা নতুন ফুলের ডেটা দেয়া হলো। সেই ডেটাকে প্লট করলে সেটা কোন প্রজাতির, বের করা সম্ভব কি না?

(ব্যাপারটা এ রকম হতে পারে, আমাদের ভবিষ্যৎ কোন প্রডাক্ট রিটার্ন হবে অথবা হবে না অথবা আমাদের ক্রেতাদের মধ্যে কারা কারা ‘রিটার্নিং কাস্টমার’ অর্থাৎ ফিরতি ক্রেতা হতে পারেন, সেটা বের করা যাবে এই পদ্ধতিতে)।

এর পাশাপাশি, লেখা বুঝতে অসুবিধা হলে বইয়ের প্লেলিস্ট দেখুন। ভিডিওটা পাবেন।

১. চালু করুন, ট্যাবলো পাবলিক ‘ওয়েব ভার্সন’ https://public.tableau.com/s/, একদম ডানে দেখুন, মাই প্রোফাইল, ক্লিক করুন— নিজের প্রোফাইলে।

২. ডানের বাটন, চাপ দিন [+ Create a Viz], চালু হয়ে গেল— নতুন ওয়ার্কশিট, সঙ্গে ‘কানেক্ট টু ডেটা’

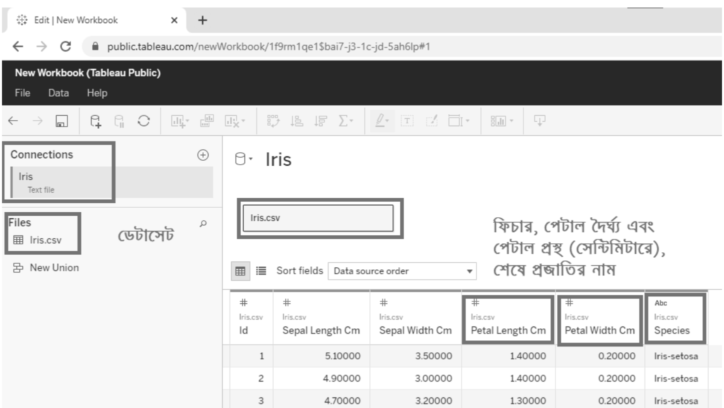
৩. ড্র্যাগ এন্ড ড্রপ অথবা ব্রাউজ করে দেখিয়ে দিন আগের ক্যাগল থেকে ডাউনলোড করা iris.csv ফাইল

৪. নিচের ‘ডেটা’ অংশে iris ক্লিক করে দেখুন, ডেটার স্ট্রাকচার, টেবিলের কলাম হেডিং, ৪টা ফিচার;

১. পেটাল দৈর্ঘ্য
২. পেটাল প্রস্থ
৩. সিপাল দৈর্ঘ্য
৪. সিপাল প্রস্থ এবং
৫. ফুলের প্রজাতির নাম

ওয়েববেসড অ্যাপের টাইমআউট সমস্যা

ট্যাবলো পাবলিক ‘ওয়েব ভার্সন’ ক্লাউডে কাজ করে বলে ডেটাসেট নিয়ে একটা নির্দিষ্ট সময় পর টাইমআউট হতে পারে। ভয় পাবেন না, আবার আপলোড করে নেবেন। সেদিক থেকে ট্যাবলো পাবলিক ‘ডেস্কটপ ভার্সন’ খারাপ নয়। একবার ইনস্টল করে নিলেই হলো।

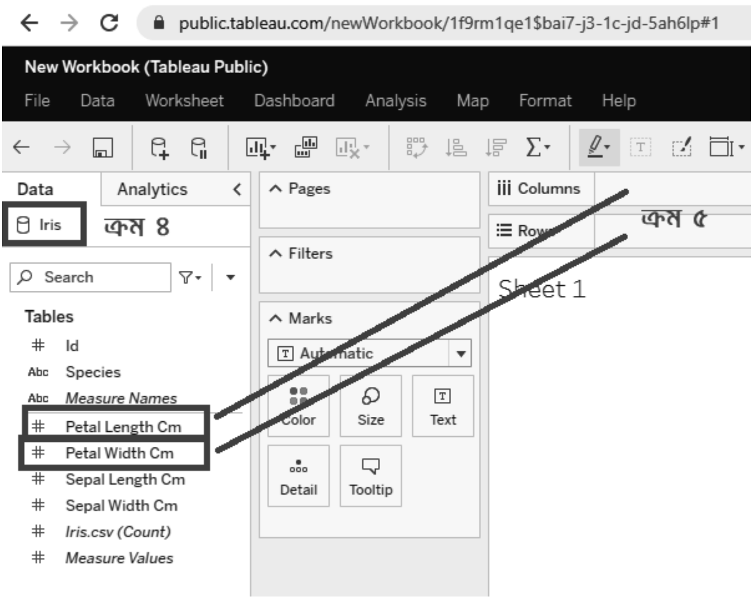


চিত্র ৩২: iris ডেটাসেট দেখুন, দুটো ফিচার ক্লিক করে কলাম এবং সারিতে বসান

ওপরের ছবিটা ভালোভাবে লক্ষ করুন। এখানে ফাইনাল বিজনেস কোশ্চেন হচ্ছে প্রজাতির নাম বের করা। ফলে দুটো ফিচার ১. পেটাল দৈর্ঘ্য ২. পেটাল প্রস্থর সম্পর্ক ধরে বের করব সেটা কোন প্রজাতি হতে পারে। দুটো ফিচার ১. পেটাল দৈর্ঘ্য ২. পেটাল প্রস্থর কো-রিলেশন আমাদের বের করে দেবে ফুলটা কোন প্রজাতির। সেই কো-রিলেশনটা দেখতে পারব ভিজ্যুয়ালাইজেশনে।

৫. আইরিস ডেটাসেটের কানেকশন, ডেটাসেট, টেবিল, আমাদের কাজ যেহেতু দুই ডাইমেনশনে, চারটার জায়গায় প্লট করব দুটো ফিচার।

কারণ, আমাদের কার্যক্ষেত্র দুটিই ডাইমেনশন। এখানে দুটো ফিচারের নাম— ১. পেটাল দৈর্ঘ্য ২. পেটাল প্রস্থ, প্রতিটা প্রজাতির আলাদা করে মাপ দেয়া আছে। আমরা অন্য ফিচারগুলো নিয়েও প্লট করতে পারি। তবে একেকবার দুটো ফিচার করে। আমি অনুরোধ করব— সব ধরনের টেস্ট করার জন্য। ডেটা অ্যানালিস্টদের কাজই তো এটা। ডেটাকে খোঁচাখুঁচি করা।



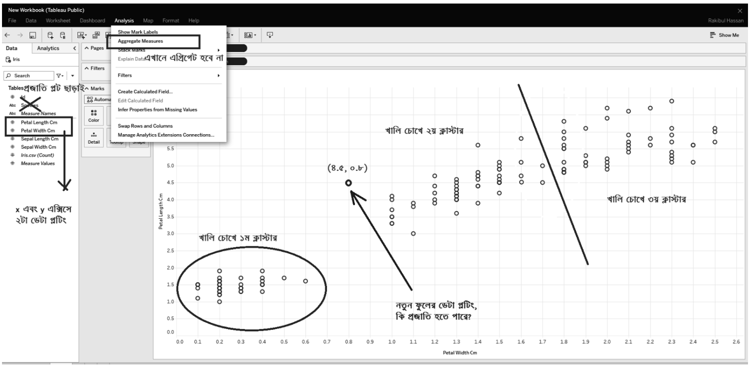
চিত্র ৩৩: ডেটাসেট কানেকশন, দুটো ফিচারের মাপ এবং প্রজাতির নাম

৬. এগুলোকে যদি এক্স এবং ওয়াই অ্যাক্সিসে প্লট করি, তাহলে কী হবে? ডিফল্ট হিসেবে সব ডেটা মেজারকে এগ্রিগেট করে বলে— ‘অ্যানালাইসিস’ মেন্যু থেকে একে ‘আন-চেক’ করে দিতে হবে। এখন খালি চোখে দেখি, আমাদের মাথা কী বলে? পরের পৃষ্ঠার ছবি দেখুন (পৃষ্ঠা: ৮২)। আমাদের মস্তিষ্ক কি এই ডেটা প্লটিংকে ঠিকমতো আলাদা করতে পারছে কি না? আমার ধারণা, ছবি দেখে আপনারা সবাই বলবেন, আমরা খালি চোখে এখানে তিনটা প্রজাতিকে আলাদা করতে পারছি। ডেটা অ্যানালাইসিসের এটাই সার্থকতা। সংখ্যাগুলোকে আমরা ঠিকমতো বুঝতে না পারলেও, এগুলোকে দুটো অ্যাক্সিসে, এক্স এবং ওয়াই অ্যাক্সিসে,

অর্থাৎ দুই ডাইমেনশনে প্লট করলে— ডেটাই বলে দিচ্ছে, আমাদের সমস্যার সমাধান কী হতে পারে। ঠিক একইভাবে, কেন একটা প্রডাক্টের রিটার্ন খুব বেশি, অথবা গত দুই মাসে প্রডাক্টের সেলস কেন নেমে গেল— সেই ডেটাগুলোকে ঠিকমতো প্লট করতে পারলে সমস্যার সমাধান, অর্থাৎ বিজনেস কোশ্চেনের উত্তর পাওয়া সম্ভব।

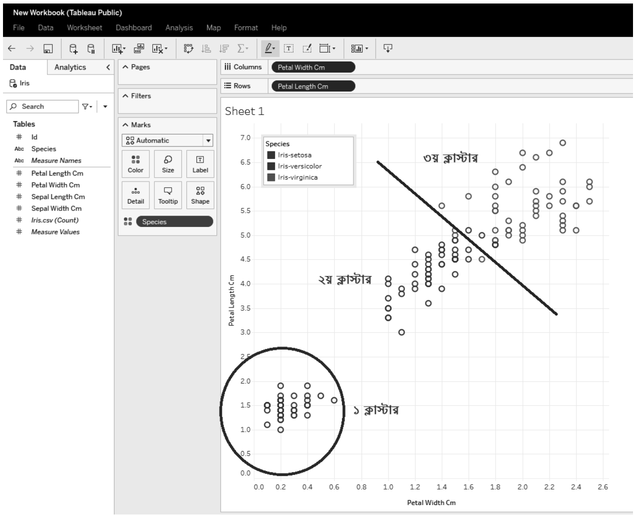
নতুন একটা ফুলের ডেটা কী বলছে?

আমাদের নতুন একটা ফুলের ডেটা দেওয়া হলো। সেটাকে প্লটিং করে আমাদের বলতে হবে সেটা কোন প্রজাতির হতে পারে? এটার উত্তর দেওয়া সম্ভব, কোনো ধরনের বাড়তি ক্যালকুলেশন না করেই। দ্বিতীয় ক্লাস্টারের কাছাকাছি বলে এটা অবশ্যই দ্বিতীয় ক্লাস্টারের প্রজাতির হবে।



চিত্র ৩৪: ডেটাকে প্লট করতে পারলে মানুষের মাথা সহজেই বুঝতে পারে

৭. এখানে 'স্পিসিস' হিসেবে রঙের ভাগ দেখি। অর্থাৎ আসল ডেটা প্লট করি। মানে আমাদের খালি চোখে যেভাবে ক্লাস্টার আলাদা করতে পেরেছিলাম, আসল ডেটার প্রজাতি কি একই জিনিস বলে? এটাই ডেটা অ্যানালাইসিসের সার্থকতা। বিভিন্ন গ্রেস্কেলে (কম্পিউটারে বিভিন্ন রঙে পাবেন) ভাগ করার আগে আসল ডেটা এবং আমাদের খালি চোখের ডেটার বিভাজন ঠিক আছে কি না, সেটাই আমাদের বিজনেস কোশ্চেন ছিল। আমাদের নতুন ফুলের ডেটা নিয়ে মেশিন লার্নিং মডেল করে দেখতে পারেন। শূন্য থেকে হাতেকলমে পাইথন মেশিন লার্নিং বইটা দেখুন।



চিত্র ৩৫: ডেটাকে প্লট করতে পারলে মানুষের মাথা সহজেই বুঝতে পারে

সাহায্যকারী ভিডিও

কেন একজন মানুষ মাত্রই ডেটা অ্যানালিস্ট, সেটা দেখানোর জন্য ওপরের (ডেটাকে ঠিকমতো ‘নেভিগেট’ করার ধারণা)-এ অধ্যায়ের প্লে লিস্টটিতে। প্রোগ্রামার অথবা নন-প্রোগ্রামার বলে নয়, প্যাটার্ন দেখে ডেটা অ্যানালাইসিস করার স্টেপিং স্টোন হতে পারে এই ভিডিওগুলো।

১৩	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১১: অ্যানালাইটিক্স টুল ট্যাবলো (আইরিস ডেটাসেট)	https://youtu.be/PWiaXY8Q35c
১৪	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১২: ট্যাবলো দিয়ে ক্লাস্টার (আইরিস ডেটাসেট)	https://youtu.be/_rFq8vVXt7A
১৫	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১৩: ট্যাবলো দিয়ে ক্লাস্টার অ্যাকুরেসি -আইরিস	https://youtu.be/_7WnMDGjpLk
১৬	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১৪: ট্যাবলো দিয়ে ক্লাস্টার, স্কয়ার-সাম	https://youtu.be/tKWssvSCe_w

তবে, শুরু থেকে বুঝতে সাথে লাগবে আরেকটা বই। ‘শূন্য থেকে পাইথন মেশিন লার্নিং’ বইটা।

ট্যাবলো দিয়ে আইরিস প্রজাতির ক্লাস্টার অ্যানালাইসিস

বিজনেস অ্যানালাইটিক্স টুলের নিজের বুদ্ধি কেমন?

অবশ্যই ভালো; বরং আমি দেখছি— সেটা বাড়ছে প্রতিদিন। বুঝব কীভাবে? আচ্ছা, টেস্ট করা যায় কীভাবে?

একটা পুরোনো ধারণা দিয়ে। বিশেষ করে, ক্লাস্টার অ্যালগরিদম দিয়ে। এখন আমাদের কাজ হচ্ছে, আমাদের মাথা (মানে চোখ) যেখানে আইরিস প্রজাতির ফুলকে ঠিকমতো ক্লাস্টার করতে পারছে, সেখানে আমাদের দেখা উচিত এই ডেটা অ্যানালাইটিক্স টুল নিজে থেকে ক্লাস্টার করতে পারে কি না? এর পাশাপাশি সত্যিকারের ডেটার সাথে টুলের অ্যানালাইসিসের মাধ্যমে প্রাপ্ত নতুন ক্লাস্টারের কতটুকু পার্থক্য সেটাও দেখব এখানে। এই পুরো জিনিসটা, হাতেকলমে করলে ব্যাপারটা পরিষ্কার হবে।

শুরুতে বুঝতে না পারলেও, কয়েকবার প্র্যাকটিস করলেই ধারণাটা বোঝা যাবে। ব্যাপারটা এ রকম যে আসল ডেটাসেটের প্রতিটা ফুলের মাপের সঙ্গে প্রজাতিকে যেভাবে ট্যাগ (মানে লেবেল) করা আছে, সেখানে প্রতিটা ফুলের মাপকে অ্যানালাইটিক্সের ওপর ছেড়ে দিলে— সে নিজে থেকে ক্লাস্টার তৈরি করবে। তখন আসল ডেটার সাথে তার অ্যাকুরেসি কতটুকু ওঠানামা করে, সেটা বুঝলেই আমাদের কাজ চলে যাবে। প্লেলিস্টে ক্লাস্টার অ্যানালাইসিস ভিডিওটা দেখুন।

আমাদের প্রথম ওয়ার্কশিটের কনটেন্টকে কপি করে নিয়ে আসি পরের ওয়ার্কশিটে। এ কারণে আমরা প্রথম ওয়ার্কশিটের ওপরে রাইট ক্লিক করে 'ডুপলিকেট ওয়ার্কশিট' ক্লিক করব যাতে একই জিনিস পরের ওয়ার্কশিটে কপি হয়ে আসে। আমরা এখানে যেহেতু ট্যাবলোকে ক্লাস্টার

অ্যানালাইসিস করতে দেব— ট্যাবলোর নিজস্ব ক্লাস্টার অ্যানালাইটিক্স, সে কারণে আগের প্রজাতি অর্থাৎ 'স্পিসিস' অংশকে (যাকে আমরা 'লেবেল' বলি) ড্রপ করে দেব। এতে ডেটার প্লটিং থাকবে তবে, আমরা 'ভিজ্যুয়ালি' রঙের পার্থক্য দেখে আলাদা করব না। আপনারা চাইলে আসল ডেটাসেট থেকে স্পিসিস কলাম বাদ দিয়েও দেখতে পারি। সত্যিকার অর্থে, এই ক্লাস্টার ব্যবহার করার জন্য আমাদের প্রজাতি কলামটা প্রয়োজন নেই।

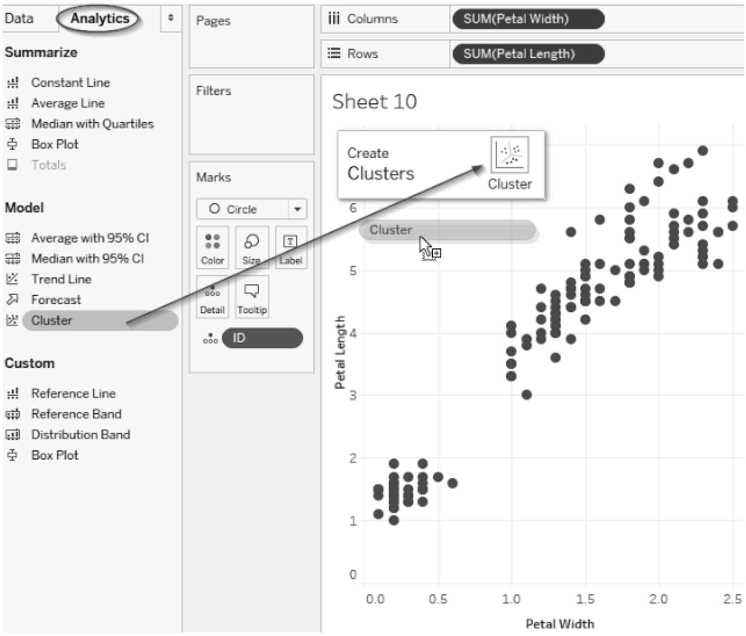
বরং এখানে ট্যাবলোকে বলছি, তার নিজস্ব অ্যানালাইটিক্স দিয়ে ক্লাস্টারিং করে দেখাতে। এটা খুবই সহজ একটা অ্যালগরিদম, এখানে ব্যবহার করছি শুধু দেখানোর জন্য— কীভাবে মানুষের মতো এ ধরনের আন-সুপারভাইজড অ্যালগরিদম ডেটার মধ্যে প্যাটার্ন বুঝতে পারে। ডেটার মধ্যে প্যাটার্ন বুঝতে পারলেই আমাদের অ্যানালাইসিস ধরনের টুল সাহায্য করতে পারে। ক্লাস্টারিং নিয়ে বিশদ লেখা হয়েছে 'হাতেকলমে মেশিন লার্নিং' বইটাতে।

এই প্রক্রিয়া করার জন্য আমরা চলে যাব ট্যাবলোয়ের অ্যানালাইটিক্স অংশে। এই অ্যানালাইটিক্স অংশটা আছে ওপরের ডেটা অংশের পাশে। পৃষ্ঠা ৮৬-এর ছবিটা দেখুন।

ট্যাবলোয়ের মাথা অর্থাৎ অ্যানালাইটিক্স

১. ট্যাবলো অ্যানালাইটিক্স অংশ থেকে আমরা বেছে নেব মডেল সেগমেন্টটাকে। সেখান থেকে ক্লাস্টার অংশটা টেনে নিয়ে ছেড়ে দেব আমাদের ওয়ার্কশিটের ওপরে। ওয়ার্কশিটের ওপরে ক্লাস্টারকে টেনে আনা মাত্রই সেখানে একটা 'ক্রিয়েট ক্লাস্টার' বলে ডায়ালগ বক্স আসবে। ক্লাস্টারকে সেই ডায়ালগ বক্সের ওপরে ছেড়ে দেব। কতগুলো ক্লাস্টার হতে পারে সেটা বলে না দিয়ে তার ওপর ছেড়ে দিতে পারি। এটাই তো পরীক্ষা! ক্লাস্টারের সংখ্যাকে আমরা এখানে অটোমেটিক হিসেবে রেখে দিই। সঙ্গে সঙ্গে স্বয়ংক্রিয়ভাবে এর পেছনের অ্যালগরিদম আইরিস ফুলের প্রজাতিকে তিন ভাগে ভাগ করে দেখাবে। মাঝের দুটো রং কিছুটা মিলে গেলেও আমরা ডানদিকে ক্লাস্টারের রঙে পাল্টে নিতে পারি। মাঝখানের ক্লাস্টারটাকে আমরা আলাদা একটা রং দিতে পারি।

শুরুতে না বুঝলেও, আপনি এই ওয়ার্কশিটগুলো অনুগ্রহপূর্বক কয়েকবার প্র্যাকটিস করবেন। এতে বোঝা সহজ হবে পরের ওয়ার্কশিটে।



চিত্র ৩৬: ক্লাস্টারকে টেনে নিয়ে আসি নতুন ওয়ার্কশিটের ওপর

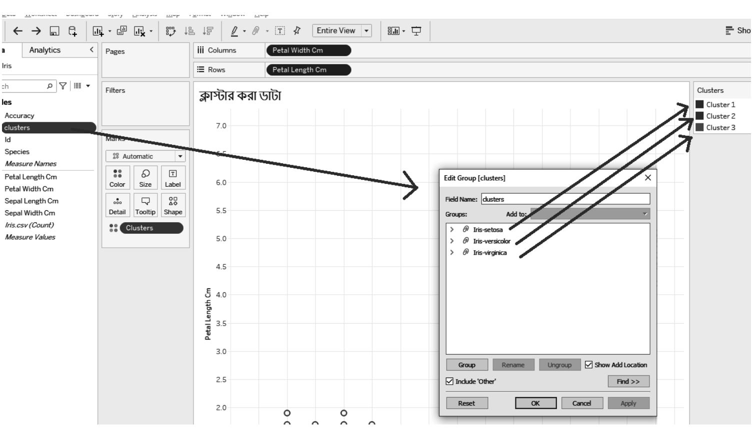
২. আমাদের এই তিন রঙের ক্লাস্টার, (ভিডিওতে 'কালার' দেখা যাবে) এবং সত্যিকারের ডেটার ক্লাস্টারকে আমরা মিলিয়ে দেখি, কতটুকু ঠিকমতো করতে পেরেছে? প্রথম ক্লাস্টার নিয়ে আমাদের কোনো সমস্যা নেই, তবে দ্বিতীয় এবং তৃতীয় ক্লাস্টারে কিছু ওভারল্যাপিং থাকাতে সেটা অঙ্কের মাধ্যমে বের করতে চাই। আমাদের ডেটাতে যদি 'স্পিসিস' অর্থাৎ প্রজাতি কলাম না থাকত তাহলে আমাদের এই রাস্তার মাধ্যমেই বের করতে হতো।

৩. আমাদের নতুন ক্লাস্টারের নাম ইংরেজিতেও ক্লাস্টার হিসেবে রাখি। এই মেজারকে আমরা টেনে ডাইমেনশনের অংশে নিয়ে যেতে পারি। মেজার হচ্ছে যেখানে কোয়ান্টেটিভ ভ্যালু অর্থাৎ সংখ্যা থাকে এগ্রিগেট, যেমন— যোগ, গড় করা যায়। যেমন বয়স, দাম, লাভ ইত্যাদি। এই ভ্যালু

কন্টিনিউয়াস হতে পারে। ডাইমেনশন হচ্ছে, কোয়ালিটিভ ভ্যালু, যেমন, দেশের নাম, আমাদের নাম, পণ্যের নাম যাদের বিভিন্ন ক্যাটেগরিতে ফেলা যায়। ডেটাকে বিভিন্ন গ্রুপে সেগমেন্ট করা যায়। সেখানে নিয়ে এটাকে ক্লাস্টার হিসেবে নাম পাল্টে ফেলি। যেহেতু আমাদের অ্যানালাইটিক্স ‘স্পিসিস’ সম্বন্ধে জানে না, সে কারণে সে তিনটা ক্লাস্টার—

ক্লাস্টার-১, ক্লাস্টার-২ এবং ক্লাস্টার-৩ নাম স্বয়ংক্রিয়ভাবে রেখেছে। চিত্র: ৩৬ দেখুন। এ জন্য আমাদের নাম পাল্টাতে হবে মূল ‘স্পিসিস’-এর সাথে তুলনা করার জন্য। একদম নাম ধরে ধরে। যাতে অঙ্কের ‘==’ স্টাইলে দুটোকে তুলনা করা যায়।

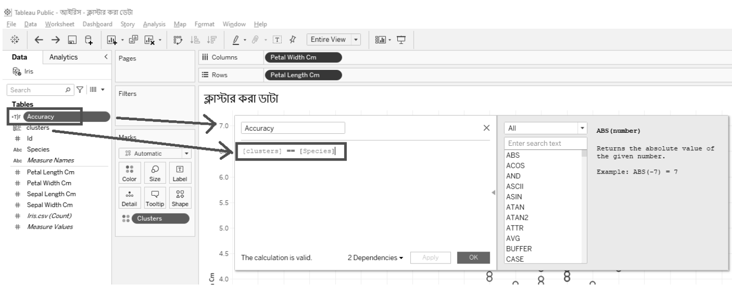
৪. এডিট গ্রুপ থেকে ‘রিনেম’ করে আমাদের ডেটা সোর্স থেকে কপি করে নিয়ে আসি ‘Iris-setosa’ যাতে ক্লাস্টার-১-এর সাথে তুলনা করা যায়। এভাবে, ক্লাস্টারগুলোকে নাম পাল্টে ৩টা প্রজাতির নামে পাল্টে ডেটাসেটের মতো করে নিয়ে আসি। ছবি দেখুন। পাশাপাশি প্লেলিস্টের ভিডিও দেখতে পারেন।



চিত্র ৩৭: ক্লাস্টারগুলোকে নাম পাল্টে ৩টা প্রজাতির নামেই এডিট করছি

৫. এই মুহূর্তে আমাদের প্রজাতী অর্থাৎ স্পিসিসের সাথে ক্লাস্টারকে তুলনা করতে হবে। সেটা কীভাবে করা যায়? অবশ্যই নতুন একটা ‘ক্যালকুলেটেড ফিল্ড তৈরির মাধ্যমে। অ্যানালাইসিস মেন্যু থেকে আমরা সিলেক্ট করে নিচ্ছি ‘ক্যালকুলেটেড ফিল্ড’। ক্যালকুলেটেড ফিল্ড হচ্ছে—

একটা নতুন ডাইমেনশন তৈরি করা, যা এখানে আগে থেকে তৈরি করা নেই। এর কাজ হচ্ছে নতুন ক্যাটাগরি বা সেগমেন্ট তৈরি করা। এর অনেক পাওয়ারফুল ফিচার পাবেন কাজ করতে গিয়ে। নাম দিচ্ছি, 'অ্যাকুরেসি'— আপনারা নাম দিতে পারেন ইচ্ছেমতো। এখানে আমাদের প্রজাতির সাথে '==' অপারেটর দিয়ে তুলনা করতে হবে ক্লাস্টারের। ছবি দেখুন। আমাদের, (লক্ষ করুন) যদি ক্লাস্টারের নাম এবং প্রজাতির এই নাম হয়, তাহলে লজিক কী হতে পারে? তুলনা করতে আমাদের বলতে হবে [clusters] == [Species], আপনার নাম ভিন্ন হলে আলাদা কথা। কপি করবেন না, লজিকটা বোঝার চেষ্টা করুন।



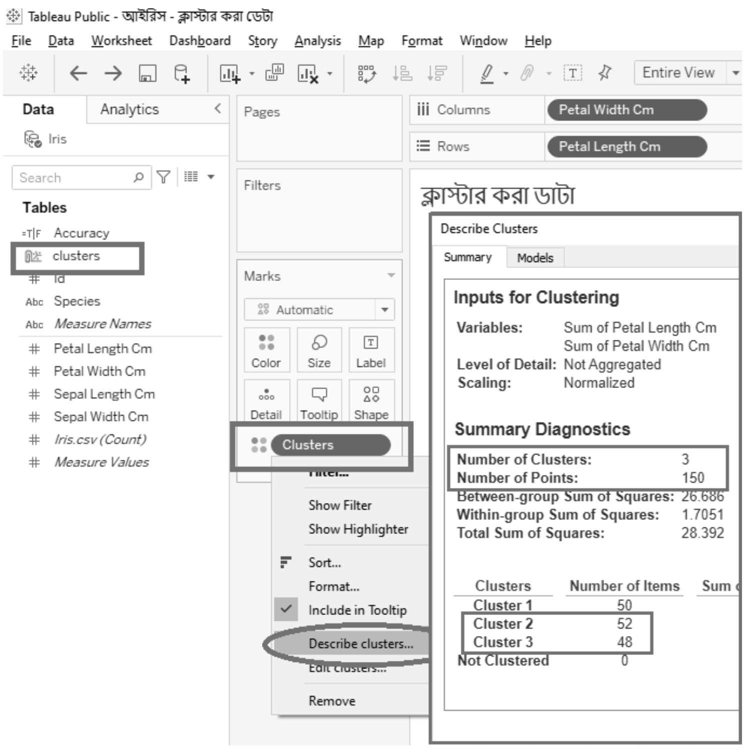
চিত্র ৩৮: ক্যালকুলেটেড ফিল্ড তৈরি

তৈরি করি 'ক্যালকুলেটেড ফিল্ড', তুলনা করার জন্য

৬. নতুন ওয়ার্কশিটে যাই। 'অ্যাকুরেসি' ক্যালকুলেটেড ফিল্ডকে নিয়ে আসি ওপরের সারিতে। এর পাশাপাশি, আইরিস প্রজাতির ডেটাসেটকে 'ড্র্যাগ অ্যান্ড ড্রপ' করে ফেলে দিই অ্যাকুরেসির সংখ্যার অংশে। দেখা যাচ্ছে, অ্যানালাইটিক্স তার প্রেডিকশনে ১৫০টা রেকর্ডের মধ্যে মাত্র ৬টা ভুল করেছে। এর অর্থ হচ্ছে, মেশিন আস্তে আস্তে 'স্মার্ট' হচ্ছে। তারা ডেটার সাহায্য নিয়ে সিদ্ধান্ত নিতে পারছে। ফলে আমরা প্রতিনিয়ত ছোট ছোট সিদ্ধান্তের কাজ ছেড়ে দিতে পারি যন্ত্রের ওপর।

৭. ফিরে আসি ক্লাস্টারে। ক্লাস্টারকে যখন 'কালার'-এর ওপরে নিয়ে এসে ছেড়ে দিয়েছিলাম, তখন অ্যানালাইটিক্স ক্লাস্টারকে তিন ভাগে তিন রং দিয়ে দেখিয়েছে। প্লেলিস্টের ভিডিও দেখতে পারেন কালারের জন্য। এর মধ্যে আমরা যদি ক্লাস্টারের ওপরে রাইট ক্লিক করে 'ডেস্ক্রাইব

ক্লাস্টার’ বলে দিই, তাহলে সে দুই এবং তিন নম্বর ক্লাস্টারে পর্যায়ক্রমে ৫২ এবং ৪৮ সংখ্যা দেখিয়েছে। চিত্র: ৩৮ দেখুন। এর অর্থ হচ্ছে, দ্বিতীয় ক্লাস্টারে দুটো বেশি, মানে মিস-ক্লাসিফিকেশন এবং তৃতীয় ক্লাস্টারে দুটো কম ক্লাসিফিকেশন ঠিক করেছে। এদিকে সত্যিকারের ডেটার সাথে তাদের ফারাক কত? ৬ এবং ১৪৪ কী হতে পারে? এর সম্পর্ক কি বের করতে পারবেন?

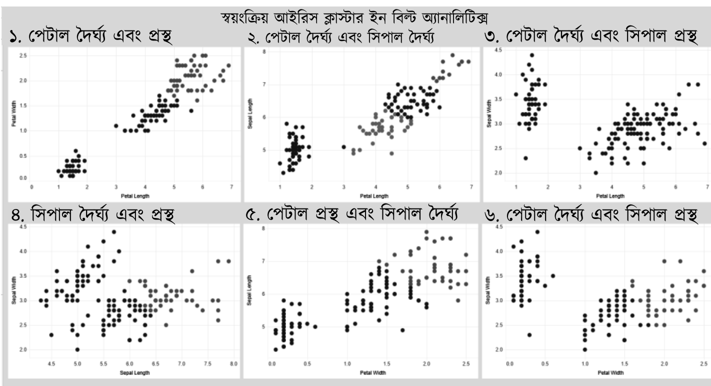


চিত্র ৩৯: ডেস্ক্রাইব ক্লাস্টার করা, যন্ত্রের তৈরি ক্লাস্টার

ক্লাস্টারকে ভেতর থেকে দেখা, ‘ডেস্ক্রাইব’ করা

৮. আমরা এ পর্যন্ত পেটাল দৈর্ঘ্য এবং প্রস্থের মধ্যে আসল সেগমেন্ট এবং ক্লাস্টারিং পদ্ধতিতে পুরো জিনিসটা দেখিয়েছি। এখানে আপনাদের কাজ হবে বাকিগুলো, যেমন সিপাল দৈর্ঘ্য, প্রস্থ অথবা পেটাল দৈর্ঘ্য

এবং প্রস্থের পার্মুটেশন এবং কম্বিনেশন করে মোট ৬ ধরনের আসল ডেটা এবং ক্লাস্টারিং করে মেশিনের অ্যাকুরেসি দেখা যেতে পারে। আমি অনুরোধ করব— এগুলো নিজের হাতে করতে, যাতে এই পুরো ব্যাপারটাতে আপনার ভালো ধারণা চলে আসে। হাতে-কলমে না করলে ডেটা অ্যানালাইটিক্স শেখা কষ্টকর। (আমার সাইটে করে দেয়া থাকলেও, নিজে করা জরুরি।) প্লেলিস্টের ভিডিওগুলো দেখলে হবে না, নিজে হাতে করতে হবে।



চিত্র ৪০: প্রতিটা কম্বিনেশন দেখে অ্যাকুরেসি চেক

আমি আবারও বলব, সিপাল দৈর্ঘ্য, প্রস্থ অথবা পেটাল দৈর্ঘ্য এবং প্রস্থের পার্মুটেশন এবং কম্বিনেশন করে দেখতে হবে নিজ হাতে।

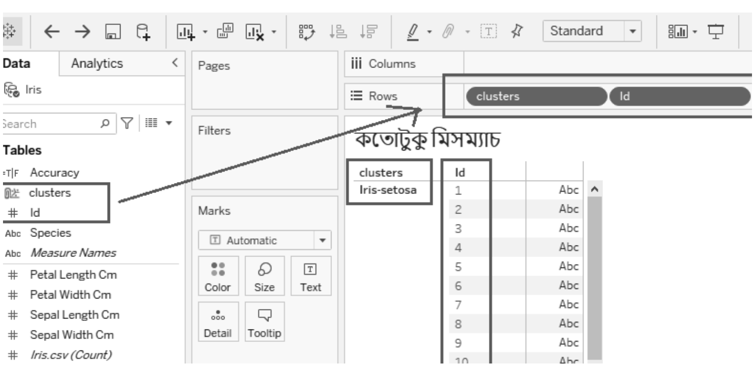
একবারে না বুঝলে, আবার দেখুন। এর পাশাপাশি ভিডিওগুলো দেখতে পারেন। বই থেকে ভিডিওগুলোতে কিছুটা ‘আলাদাভাবে’ করে দেয়া আছে, একই জিনিস যে অনেকভাবে করা যায়, সেটা নিশ্চিত করে বোঝাতে।

১. https://youtu.be/PWiaXY8Q35c
২. https://youtu.be/_rFq8vVXt7A
৩. https://youtu.be/_7WnMDGjpLk
৪. https://youtu.be/tKWssvSCe_w

ট্যাবলো—প্রজাতির ক্লাস্টার অ্যানালাইসিস এবং মিসম্যাচ

শুরুতে এক্সারসাইজটা না বুঝলেও আপনারা করে ফেলুন। শেষে পার্সপেকটিভ পাবেন পুরো ঘটনার।

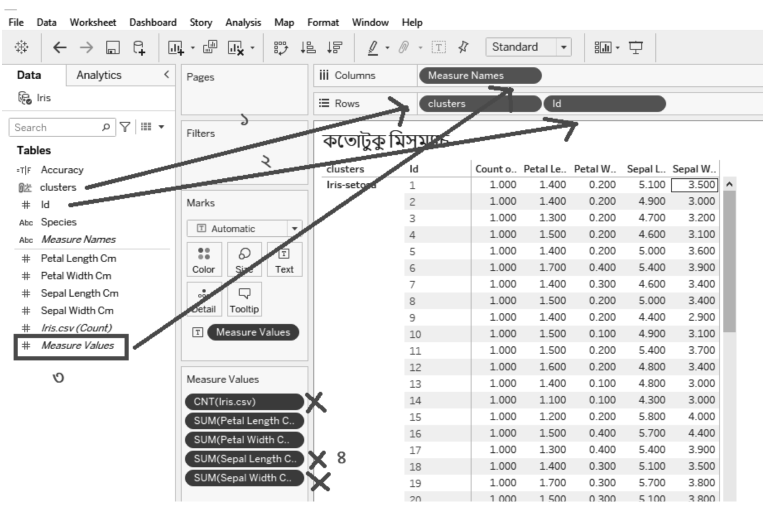
ভালো কথা— আগের এক্সারসাইজে দেখলাম, ৬টা ডেটা 'মিস-ম্যাচ'। এখন কী হবে? এ মুহূর্তে, আমরা জানতে চাইব— প্রজাতির আসল ডেটা এবং ট্যাবলো তার অ্যানালাইটিক্স দিয়ে যেই ক্লাস্টার করেছে— তার মধ্যে কতটুকু মিস-ম্যাচ? বিশেষ করে, কোন কোন রেকর্ডে মিস-ম্যাচ, সেটা বের করতে চাইব। এর পরিপ্রেক্ষিতে আমরা শুরুতে ক্লাস্টার ডেটা এনে ছেড়ে দেব ওয়ার্কশিটের 'সারির' ওপরে। এতে আমাদের তিনটা ক্লাস্টারের একটা টেবিল তৈরি করে দেবে— যেখানে এই মুহূর্তে কোনো সংখ্যা থাকছে না। আমাদের এখানে যেহেতু ১ থেকে ১৫০টা রেকর্ড, সে কারণে শুরুতেই আইডিকে টেনে নিলাম সারিতে। আমরা জানতে চাইছি, প্রতিটা রেকর্ড ধরে কোন কোন রেকর্ড আসল ডেটাতে কী প্রজাতি বলেছে এবং এর প্রেক্ষিতে ক্লাস্টার অ্যানালাইসিসে কোথায় কোথায় ভুল ক্লাসিফাই করেছে? এখানে তিনটা ক্লাস্টার ধরে আলাদা আলাদা করে আইডি দেখা যাচ্ছে শুরুতে।



চিত্র ৪১: ক্লাস্টার ডেটা এনে দিলাম 'সারি'তে

আগের তৈরি ক্লাস্টার এবং রেকর্ড আইডি দিয়ে টেবিল

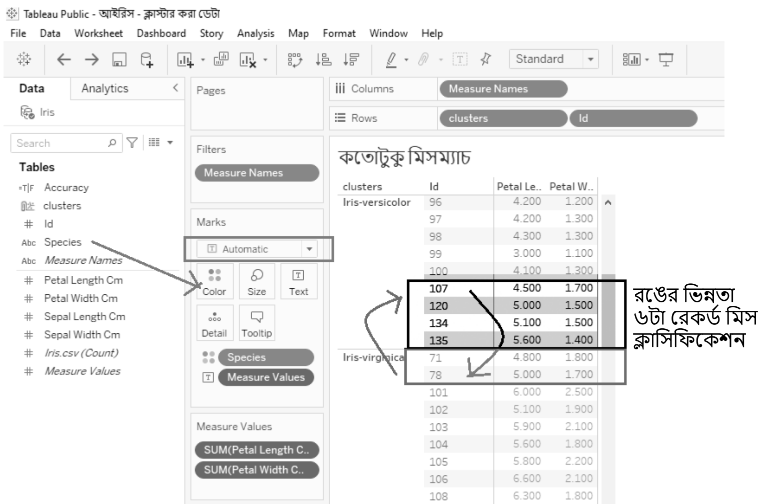
ট্যাবলোতে আমরা যখন একটা ডেটাসেট ইমপোর্ট করে নিয়ে আসি, তখন প্রতিটা ফিচার অর্থাৎ কলামের ভ্যালুগুলো আনার সময় তিনটা অটো জেনারেটেড 'মেজার' তৈরি করে আমাদের কাজের সুবিধার জন্য। এর মধ্যে মেজার নেম, মেজার ভ্যালু এবং পুরো ডেটাসেটের রেকর্ডের একটা 'কাউন্ট মেজার' নিয়ে আসে। সে কারণে, যেকোনো ডেটাসেটের একটা পরিপূর্ণ ধারণা পেতে চাইলে 'মেজার ভ্যালু' ডাবল ক্লিক করলে সব ধরনের পার্সপেক্টিভ পাওয়া যায়। সেই ধারণা থেকে আমরা 'মেজার ভ্যালু' ডাবল ক্লিক করছি, যেখানে সিপাল এবং পেটাল দৈর্ঘ্য এবং প্রস্থ (মোট চারটি) এবং রেকর্ডের কাউন্ট (১ করে মোট ১৫০টি) আইডি টেবিলে পাব।



চিত্র ৪২: মেজার ভ্যালু দিয়ে পুরো ধারণা নিতে চাওয়া, এক টেবিলেই

আমাদের এখানে যেহেতু 'পেটাল দৈর্ঘ্য' এবং প্রস্থ ছাড়া বাকি মেজারের প্রয়োজন নেই, সে কারণে আমরা রেকর্ড কাউন্ট এবং বাকি দুটো মেজার মুছে দিচ্ছি। চিত্র: ৪১ দেখুন। এতে আমাদের প্রয়োজনীয় দুটো মেজার থাকছে। এখন এই ক্লাস্টারকে তুলনা করতে হবে ডেটাসেটের আসল প্রজাতির ভ্যালুর সাথে, যেটা এখানে 'স্পিসিস' হিসেবে আছে। আমরা যেহেতু ভিজ্যুয়ালি এ দুটোর পার্থক্য বুঝতে চাই, সে কারণে স্পিসিস

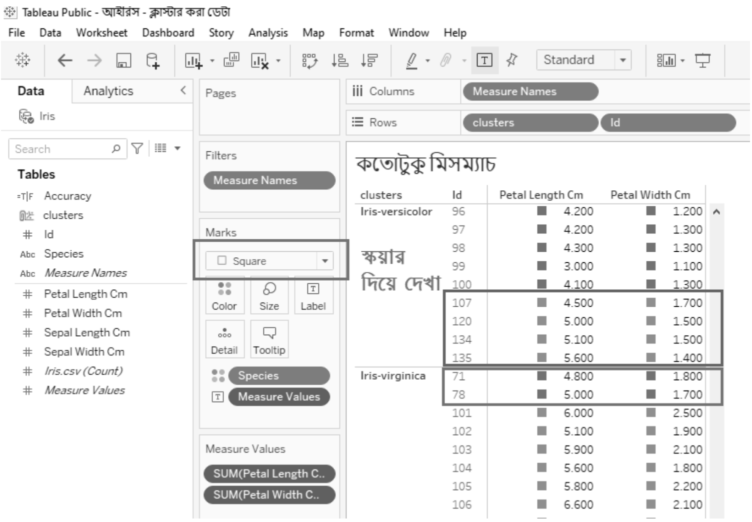
ডাইমেনশনকে টেনে আনছি কালারের ওপরে। চিত্র: ৪২ দেখুন। আমি আপনাদের অনুরোধ করব, ইচ্ছেমতো বিভিন্ন ডাইমেনশন এবং মেজারগুলো ওয়ার্কশিটে টেনে আনতে, যাতে আপনার ডেটা অ্যানালাইটিক্স সম্পর্কে ভয় ভেঙে যায়। আপনি কিছু নষ্ট করতে পারবেন না। বলছি! আপনি যা-ই করুন, ডেটা অ্যানালাইসিস সফটওয়্যার নষ্ট করতে পারবেন না! নিজের মতো করে বিভিন্ন জিনিস টেনে ফেলে দেখুন কী হয়!



চিত্র ৪৩: রঙের পার্থক্যই বলে দিচ্ছে মিসম্যাচ, রেকর্ড ধরে

মিস ক্লাফিকেশন — রং দিয়ে

দেখুন রেকর্ড নম্বর, ১০৭, ১২০, ১৩৪, ১৩৫, ৭১ এবং ৭৮? শেডের অমিল পাচ্ছেন? প্রজাতি আইরিস ভার্সিকালারের মধ্যে চারটা এবং প্রজাতি ভার্জিনিকার মধ্যে দুটো ভিন্ন রঙের মোট ৬টি মিস-ক্লাসিফিকেশন দেখতে পাচ্ছি। তবে এই মিস-ক্লাসিফিকেশন আমরা 'মার্কস' থেকে গোলাকার অথবা বর্গাকার চিহ্ন দিলে আমাদের দেখার সুবিধা হবে। আপনাদের সুবিধার জন্য পুরো পাঁচটা ওয়ার্কশিট এবং একটা ড্যাশবোর্ড তৈরি করে দিয়েছি নিচের লিংকে। ডাউনলোড করে নিন আপনার ট্যাবলো ডেস্কটপ অথবা কপি করে নিন ওয়েব ভার্সনের জন্য।



চিত্র ৪৪: মানুষ হিসেবে বর্গাকার ভিজ্যুয়ালাইজেশন, আসল মিস-ম্যাচ, রেকর্ড ধরে

ওয়ার্কশিট ডাউনলোড করুন এই লিংকে—

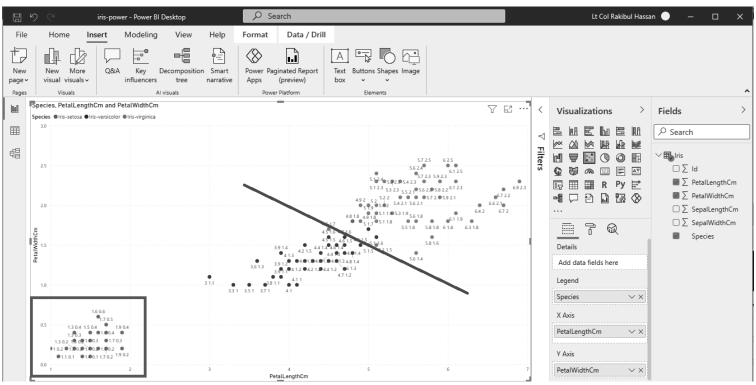
https://public.tableau.com/views/-_16259206654450/sheet5

— কথা দিতে হবে, পরেরগুলো নিজেই করবেন সব!

আপনি হাতেকলমে নিজে করতে পারলে আমাকে আর প্রয়োজন পড়বে না।

এই এক্সারসাইজটা প্রমাণ করে, আমাদের আসল ডেটা এবং যন্ত্রের ক্লাস্টার করার ডেটার মধ্যে খুব একটা ফারাক নেই। এর অর্থ হচ্ছে, ডেটা থেকে সিদ্ধান্ত নিলে আমাদের ভুল হবার আশঙ্কা কম।

মাইক্রোসফট পাওয়ার বিআই এবং ট্রাভেল ইনস্যুরেন্স



চিত্র ৪৫: ডেটাকে প্লট করতে পারলে মানুষের মাথা সহজেই বুঝতে পারে

ট্রাভেল ইনস্যুরেন্স ডেটাসেট, ‘কে’ এবং ‘কেন’ প্রশ্নের উত্তর

ডেটা অ্যানালাইটিক্স নিয়ে বই লেখার আগে একটা বড় সময় দিয়েছিলাম ডেটাসেট সিলেকশনের ব্যাপারে। প্রতিটা বই লেখার আগে এ ধরনের রিসার্চ বেশ সময় নিলেও এতে ফলাফল ভালো আসে। আমি সেই ধরনের ডেটাসেটগুলোকে রিলিজ ‘ক্যান্ডিডেট’ হিসেবে পছন্দ করি,

একটা ডেটাসেট পছন্দ করার পন্থা

১. যেগুলোর সাথে বর্তমান ইন্ডাস্ট্রির একটা ভালো সংযোগ আছে, বিশেষ করে ‘রেলেভেন্সি’র দিক থেকে

২. যেই ডেটাসেটগুলো নিয়ে পুরো পৃথিবীর একটা বিশাল অংশ আলাপ করছে— প্রচুর রিসার্চ হয়েছে যার ওপর

৩. ছোট ডেটাসেট এবং বৈচিত্র্য বেশি— যেই ডেটাসেটগুলো অল্প ফুটপ্রিন্টে আমাদের ভেতরের প্রশ্ন তৈরি করার সুযোগ করে দেয়— পথ দেখায় কীভাবে প্রশ্ন করলে সেটার সমাধান পাওয়া যাবে।

টাইটানিক ডেটাসেটের (হাতেকলমে মেশিন লার্নিং বই থেকে) পাশাপাশি আমাদের হাতেকলমের এই অংশ শুরু হচ্ছে 'ট্রাভেল ইনস্যুরেন্স' নিয়ে। এই ডেটাসেটের ভিত্তি হচ্ছে একটা সিঙ্গাপুরভিত্তিক থার্ড-পার্টি ইনস্যুরেন্স সার্ভিসিং কোম্পানি, যারা ভ্রমণকারীদের কাছে একটা সময়ের জন্য ট্রাভেল ইনস্যুরেন্স বিক্রি করে থাকেন। স্বভাবতই— এর সবচেয়ে বড় প্রশ্ন হচ্ছে, একজন ব্যবহারকারী ইনস্যুরেন্স কেনার পরে সেটা ক্লেইম করবেন কি না? বলতে গেলে— এটা অবশ্যই একটা বড় বিজনেস কোয়েশ্চেন। বিশেষ করে, যিনি বিক্রি করছেন। ইনস্যুরেন্স' কোম্পানির একটা বড় চিন্তা থাকে যে, তাদের ইনস্যুরেন্স বিক্রি করার পর সেটাকে কারা কারা 'ক্লেইম' করতে পারেন? কারণ, ক্লেইম বেশি হলেই সমস্যা। সেটা যদি আগে থেকে বোঝা যায়, তাহলে একটা ভালো 'বিজনেস কোয়েশ্চেন'-এর সমাধান করা যায়।

আমাদের এই ডেটাসেটটা ক্যাগলে থাকলেও এর খোঁজ পাই 'লন্ডন বিজনেস অ্যানালাইটিক্স গ্রুপের মাধ্যমে। ছোট গ্রুপ হলেও তাদের ইন্ডাস্ট্রি স্পেসিফিক কিছু কাজ চোখে লাগার মতো। সেখানে এই ডেটাসেট নিয়ে আলাপ করেছিলেন অ্যানালিস্ট প্রগতি জৈন। ৪৫০ কিলোবাইটের এই ছোট্ট তবে বৈচিত্র্যময় ডেটাসেটটা পাওয়া যাবে নিচের এই লিংকে।

https://www.kaggle.com/mhdzahier/travel-insurance

ডেটা ডিকশনারি, 'ট্রাভেল ইনস্যুরেন্স' ডেটাসেট

ক্রম	ফিচার/কলামের নাম	মানে কী হতে পারে?
১.	Agency	যেই এজেন্সির মাধ্যমে ইনস্যুরেন্স বিক্রি হচ্ছে, তার কোড
২.	Agency Type	কী ধরনের এজেন্সি, ট্রাভেল এজেন্ট নাকি এয়ারলাইনস
৩.	Distribution Channel	ডিস্ট্রিবিউশন চ্যানেল, যেটা অনলাইন অথবা অফলাইন হতে পারে

ক্রম	ফিচার/কলামের নাম	মানে কী হতে পারে?
৪.	Product Name	ইনস্যুরেন্স প্রডাক্টের নাম
৫.	Claim	ইনস্যুরেন্সটা ক্লেইম করা হয়েছে কি না সেটার স্ট্যাটাস
৬.	Duration	কত দিনের জন্য এই ট্রাভেল ইনস্যুরেন্স কাভার করবে
৭.	Destination	যেই গন্তব্যের জন্য ট্রাভেল ইনস্যুরেন্স কেনা হয়েছে
৮.	Net Sales	ইনস্যুরেন্সটার নেট বিক্রির হিসেব
৯.	Commision (in value)	ইনস্যুরেন্সটার জন্য কত কমিশন পাওয়া গেছে
১০.	Gender	ভ্রমণকারী মহিলা নাকি পুরুষ?
১১.	Age	ইনস্যুরেন্স ক্রেতার বয়স

টেবিল ১: ডেটা ডিকশনারি

আমরা যখন ডেটা নিয়ে কাজ করি, তখন বিশ্বাস করতে চাই ডেটার মধ্যে একটা 'কোহেরেন্স' আছে, অর্থাৎ ডেটার বিভিন্ন ফিচারের মধ্যে একটা ভালো সম্পর্ক আছে। এখানে ফিচার হিসেবে প্রতিটি কলামের যেই এট্রিবিউটগুলো আছে তাদের ভেতরের ইন্টার্নাল সম্পর্কের কথা বলছি। মনে আছে টাইটানিক ডেটাসেটের কথা? টাইটানিক ডেটাসেটের মতো এই ডেটাসেটেও একটা বিজনেস প্রবলেম আছে, অর্থাৎ ট্রাভেল ইনস্যুরেন্স কেনার পাশাপাশি তাদের 'ক্লেইম' ফ্রিকোয়েন্সি কিছুটা বেশি। কেন? সেটা জানতেই এই অ্যানালাইসিস ইনস্যুরেন্স কোম্পানিগুলো চাইবে যাতে তাদের প্রডাক্টের কম 'ক্লেইম' হয়। সেটাই স্বাভাবিক নয় কি?

দরকার পাওয়ার বিআই-এর মতো একটা 'অ্যাডভান্সড টুল'

এটা কিছুটা ডেটা সায়েন্স সমস্যা হলেও এখানে ডেটা অ্যানালাইসিস করার ভালো স্কোপ রয়েছে। সমস্যা হচ্ছে, মানুষ যখন ডেটা, অর্থাৎ সংখ্যাগুলো নিয়ে কাজ করে তখন সেই ডেটা থেকে ধারণা পাবার জন্য 'সায়েন্টিফিক টুল' না ব্যবহার করলে এ ব্যাপারে ভেতরের সম্পর্ককে বের করা কষ্টকর হয়ে পড়ে। ছোট ডেটা হলে কথা ছিল, এখন তো বিগ-ডেটার যুগ। আর সে কারণেই, যেকোনো কিছু করার আগে আমরা এই ১১টা ফিল্ড ভালো করে দেখব। তবে শুরুর কথা হচ্ছে, এর ভেতরের রিলেশন কীভাবে বোঝা যায়?

ভিডিও লিংক—

পাওয়ার 'বি-আই' ইনস্যুরেন্স ক্লেইম	https://youtu.be/iGuq_vqdygY

পাওয়ার বিআইতে ট্রাভেল ইনস্যুরেন্স ডেটা লোডিং

১. মাইক্রোসফট পাওয়ার বিআইতে শুরুতেই আপনার কাছে ডেটা সোর্স চাইবে। সেটা ক্যানসেল করে দিলেও সেটাকে নতুন করে যোগ করে নিতে পারেন 'Get Data' মেন্যু থেকে। সেখান থেকে ট্রাভেল ইনস্যুরেন্স 'CSV' ফাইল travel_insurance.csv ফাইলটাকে 'csv' হিসেবে লোড করলেই একটা হাই লেভেল ডেটা ডায়াগ্রাম চলে আসবে, যেখানে '২০০' সারি ডেটা দেখাবে।

২. এটাতে 'লোড' এবং 'ট্রান্সফর্ম' বলে দুটো বাটন চলে আসবে। এ মুহূর্তে আমরা 'লোড' না করে 'ট্রান্সফর্ম' নিয়ে কাজ করতে চাইব ডেটা দেখার জন্য। ভিডিও দেখুন বিস্তারিত ধারণার জন্য পাওয়ার কোয়েরি এডিটর নিয়ে।

৩. লোড অথবা পাওয়ার কোয়ারি এডিটর থেকে 'ক্লোজ অ্যান্ড অ্যাপ্লাই' শেষে তাকান সবচেয়ে ডানে। দেখুন 'Fields'-এর নিচে 'Travel-insurance' টেবিল। সেটাকে 'এক্সপান্ড' করলে দেখুন সব ফিল্ডের নাম। এর পাশাপাশি, দুটো ডেটাসেটের মধ্যে সম্পর্ক দেখার জন্য লোড করব 'Accidents' ডেটা।

৪. নতুন ডেটাসেট যোগ করার অর্থ হচ্ছে, আপনাদের দেখাল, কীভাবে আরও আলাদা ডেটাসেট যোগ করা যায়, যেখানে তাদের মধ্যে রিলেশনশিপ তৈরি করা যায়। আমাদের 'accidents.csv' নিয়ে আসব ওপরের 'Get data' থেকে 'csv/txt' বেছে নিয়ে। এই ফাইলটা পাওয়া যাবে Githab লিংক থেকে। এখানে দুটো কলাম দেখছি country এবং accidents।

৫. ডানে দেখুন। পাশাপাশি ভিডিওটা দেখুন। ফিল্ডের নিচে দুটো ডেটাসেট। accident এবং travel insurence। এখন সবচেয়ে বামে দেখুন, 'রিলেশনশিপ' বাটন ক্লিক করুন। দুটো ডেটাসেটের মধ্যে কমন ফিল্ড 'Country'কে 'destination' এর সাথে যোগসূত্র করে দেব।

যেকোনো ডেটাসেটের ওপর রাইট ক্লিক করে 'ম্যানেজ রিলেশনশিপ' থেকে 'Country' তে দেশের নাম একবার এবং 'Travel insurance' টেবিলে 'destination' এ একটা দেশের নাম কয়েকবার থাকাতে আমরা 'কার্ডিনালিটিতে' 'One to many' বেছে নিলাম। এতে দুটো ডেটাসেটের মধ্যে কমন ফিল্ডের মধ্যে সম্পর্ক তৈর হলো। আমরা এই সম্পর্ক থেকে অনেক ভিজ্যুয়ালাইজেশন তৈরি করতে পারি বা আমি আপনাদের করতে বলব। এই অংশটুক আমি আপনাদের ওপরে ছাড়তে চাই।

ট্রাভেল ইনস্যুরেন্স ভিজ্যুয়ালাইজেশন

৬. চলুন, কিছু ভিজ্যুয়ালাইজেশন— অর্থাৎ বিজনেস প্রশ্ন তৈরি করি উত্তর দেখার জন্য। ভিডিওটা দেখুন। প্রথম প্রশ্ন কী হতে পারে? যেমন, এজেন্সিভিত্তিক কারা বেশি সেল করছে? এয়ারলাইনস না ট্রাভেল এজেন্সি?

৭. সবচেয়ে বেশি ভিজ্যুয়ালাইজেশনের মধ্যে 'ক্লাস্টার্ড বার চার্ট' খুবই সোজা এবং সবারই প্রিয়। ভিজ্যুয়ালাইজেশন থেকে সেটাকে টেনে 'Page 1' এর ওপর ছেড়ে দিলাম। এরপর, একটা ব্ল্যাংক ম্যাট্রিক্স চলে আসবে। এর পাশাপাশি, ভিজ্যুয়ালাইজেশনের নিচে অনেকগুলো ফিল্ড চলে আসবে। এখন, আমাদের 'Agency Type' কিন্তু টেনে এনে ফেলে দিতে হবে 'Axis'-এ এরপর, 'Product Name' ফেলে দিতে হবে 'value' তে।

৮. দেখুন মজার একটা জিনিস। 'Values' এ পুরো 'প্রডাক্ট নেমে' কাউন্ট অর্থাৎ যোগ হয়ে স্বয়ংক্রিয়ভাবে বসে যাবে। এখন ফরম্যাটিংয়ে চাপ দিয়ে আপনার মতো করে ডেটা লেবেল, ডেটা কালার, বিভিন্ন অ্যাক্সিস ধরে মেন্যু সিলেকশন, সব সম্ভব। এই ভিজ্যুয়াল কী দেখাচ্ছে? ট্রাভেল এজেন্সি আসলে বেশি টিকিট বিক্রি করছে। কেন? এটার উত্তর কয়েকভাবে দেওয়া যেতে পারে। সহজ অথবা 'ইন-ডেপথ' স্টাইলে। তবে এটা কমফোর্ট জোনের বাইরে। আপনাদের শেখাতে।

৯. ওপরে চাপ দিন, 'New Measure', একটা ফর্মুলা বার চলে আসবে। মেজার ছাড়াও এই জিনিস করা সম্ভব, তবে প্রোগ্রামেটিক্যালি এগুলো সোজা। শুরুতে একটা 'মেজার' নাম দিতে হবে। ডিফল্ট নেয়া যায়। তবে একটা মিনিংফুল নাম দিই। Product Volume = count (travel

insurance [Product Name]) এই পুরো জিনিসটাই, খানিকটা 'অটো কমপ্লিট' হিসেবে চলে আসবে। এটা ট্যাবলো-এর ক্যালকুলেটেড ফিল্ডের মতো। দেখুন, ফিল্ডের নিচে 'Product Volume' নামে একটা ক্যালকুলেটর স্টাইলে একটা মেজার আসবে। এই নতুন মেজারকে 'Value' হতে এনে ছেড়ে দিই। কী হলো? একই জিনিস তবে এখানে ফরম্যাটিংয়ে অপশন বেশি। 'Count of Product' এবং 'Product Vloume'-এর আউটকাম একই— তবে 'Product Vloume'-এ ফরম্যাটিংয়ে যাই। নতুন মেজার ক্লিক করি, উপরে দেখুন $v%, ইত্যাদি দেখাচ্ছে। থাউসেন্ড সেপারেটর '৴' ক্লিক করি। এখানে ৪৫৮৬৯ এর জায়গায় ৪৫,৮৬৯ দেখাচ্ছে। এখন 'Count of Product Name' অংশটি 'Value' থেকে মুছে দিলাম। থাউসেন্ড সেপারেটর দিয়ে দেখতে ভালোই লাগছে।

১০. এখন আগের প্রশ্নে ফিরে যাই, কেন ট্রাভেল এজেন্সি বেশি বিক্রি করতে পারছে? চলুন, আবার ডেটা দেখি। Commission (in value) দেখি। এজেন্সিগুলো এয়ারলাইন টিকিট বিক্রি করার সময় কমিশন পায়। এবার ভিজ্যুয়ালাইজেশন থেকে ক্লাস্টার্ড কলাম চার্ট নিয়ে 'Page 1'-এ ফেলি। 'Axis'-এ 'Agency Type' টেনে ফেলে দিই। এখন এজেন্সি ধরে কে কত করে কমিশন পাচ্ছে, সেটা দেখার জন্য 'Commission' টেনে 'Values'-এ ফেলে দিই। নতুন চার্টটা দেখুন। কোথাও নিউমেরিক ভ্যালু পাওয়ার বিআইতে ফেললে সেটা সাধারণত 'Sum' করে। ভিডিও দেখুন। তবে আমরা সেটাকে 'Values' থেকে ড্রপ ডাউন দিয়ে 'Average' সিলেক্ট করে দিই। এরপর ফরম্যাটিং থেকে 'ডেটা লেবেল' চালু করে দিই।

এখানে এয়ারলাইন এবং ট্রাভেল এজেন্সির গড় পজিশন দেখতে পাচ্ছি। ভয়লা! এটাই আমাদের উত্তর। কেন ট্রাভেল এজেন্সি বেশি টিকিট বিক্রি করতে পারছে? এয়ারলাইনসগুলো প্রতিটা ট্রাভেল ইনস্যুরেন্স বিক্রি করতে ১৬ পাউন্ড চার্জ করছে, সেখানে ট্রাভেল এজেন্সি মাত্র ৮ পাউন্ড— যা ৫০% কম। আমি নিজেই ট্রাভেল এজেন্সি থেকে কিনব— আর সে কারণে ট্রাভেল এজেন্সি বেশি প্রডাক্ট বিক্রি করছে।

১১. ধরুন আমরা এই ইনস্যুরেন্স বিক্রি জেন্ডারভিত্তিক চাই। কতজন মহিলা/পুরুষ এই প্রডাক্টগুলো কিনেছেন? ভিজ্যুয়ালাইজেশন থেকে 'স্লাইসার' টেনে 'Page 1'-এ ফেলি। এখানে ফিল্ড 'Gender' ফেলি জেন্ডার স্লাইসের চেকবক্সে 'M', 'F', এবং 'U'-(জানা নেই) চলে আসবে। এখানে M/F-এর যেকোনো চেকবক্সে ক্লিক করলেই আবার দুটো ক্লাস্টার বার চার্ট পাল্টে যাবে। চমৎকার না?

এভাবে প্রায় ১০টির মতো ভিজ্যুয়ালাইজেশন পাবেন ভিডিওতে। বইয়ের পৃষ্ঠা কমাতে এই ব্যবস্থা। তবে ভিডিও দেখে, নিজের হাতেকলমে চেষ্টা করুন। তাহলে ভালো ধারণা আসবে।

ডেটা নিয়ে বেসিক ধারণা

ডেটা কী এবং ডেটার সোর্স

আমরা জেনেছি যে ডেটা লিটারেসি, অর্থাৎ সাক্ষরতা হলো— (১) ডেটাকে ঠিকমতো অন্বেষণ করতে পারা, (২) সেটাকে বোঝা এবং (৩) সেই ডেটা থেকে ইনসাইট বের করে সেটাকে অন্যদের মাঝে ছড়িয়ে দেবার ক্ষমতা তৈরি করা।

কিন্তু 'ডেটা' ঠিক কী জিনিস?

ডেটার কালেকশন হচ্ছে অনেকগুলো তথ্যের সন্নিবেশনে 'ফ্যাক্ট'। আরও বিশেষভাবে বলতে গেলে, এই 'এক্সপান্ডেড' সংজ্ঞাটি দেখুন। জনস হপকিন্স ব্লুমবার্গের স্কুল অব পাবলিক হেলথের অধ্যাপক হিসেবে কর্মরত ডেটা বিজ্ঞানী জেফরি লিক উইকিপিডিয়া থেকে এই 'এক্সপান্ডেড' সংজ্ঞাটি নিয়ে বলছেন: 'ডেটা অনেকগুলো গুণগত (কোয়ালিটেটিভ) বা পরিমাণগত (কোয়ান্টিটেটিভ) ভ্যারিয়েবলের মান নিয়ে তৈরি, যা ওই আইটেমের একটি সেটের অন্তর্গত।'

ঝামেলা মনে হচ্ছে? টেক্সট বইয়ের মতো? চলুন, আমরা এই পুরো জিনিসটাকে ব্যবচ্ছেদ করে নিজেরা বুঝি।

আইটেমের সেট

আমরা যে জিনিসটা নিয়ে কাজ করতে আগ্রহী, সেই আইটেমের অংশকে আমরা বেশির ভাগ সময়ে 'পপুলেশন', অর্থাৎ জনসংখ্যা বলি। আইরিস প্রজাতির ডেটা একটা সেট হতে পারে। সে কারণে এর নাম আইরিস ডেটাসেট।

ভ্যারিয়েবল

এটা এমন একটা পরিমাপের বৈশিষ্ট্য যা কিছু জিনিসের ইনপুটের উপরে পরিবর্তন হতে পারে। এর বিপরীত অর্থ হচ্ছে কনস্ট্যান্ট যা কখনোই পাল্টায় না— যেমন ‘পাই’-এর মান। এদিকে তাপমাত্রা একটা ভ্যারিয়েবেল হতে পারে যা দিনের সময়ের সাথে সাথে পাল্টায়।

কোয়ালিটেটিভ ভ্যারিয়েবল

এটা গুণগত মানের একটা পরিমাপক, যেমন একজন মহিলা না পুরুষ, মানুষের অথবা দেশের নাম, তবে এটা পরিবর্তনশীল যা রেকর্ডের উপরে নির্ভর করে পাল্টানো যায়। একটা ক্লাসে ৫০ জন ছাত্র-ছাত্রী থাকলে প্রতিটা সারিতে ছাত্র-ছাত্রীর নাম পাল্টাবে অর্থাৎ একটা সারিতে একেকজনের নাম হবে। এটাকে আমরা যদি একটা বাক্স ধরি, যেমন একটা দেশের নাম, তার ভেতরে তার জনগণের নাম ও কোয়ালিটেটিভ ভ্যারিয়েবল হবে। একটা সুপারশপে তার বিক্রি আইটেমগুলোকে বিভিন্ন ক্যাটাগরি এবং সাব-ক্যাটাগরিতে ফেলা যেতে পারে। বাক্সের ভেতরে বাক্স।

কোয়ান্টিটেটিভ ভ্যারিয়েবল

এটাকে আমরা পরিমাণগত একটা সূচক বলতে পারি, যা সংখ্যা দ্বারা নির্ধারিত। যেমন মানুষের উচ্চতা, ওজন এবং তাপমাত্রা পর্যায়ক্রমে বয়স এবং দিনের সময়ের সাথে পাল্টায়। একটা দোকানের সেলসের পরিমাণ, লাভ, ক্ষতি ইত্যাদি হচ্ছে কোয়ান্টিটেটিভ ভ্যারিয়েবলের উদাহরণ।

এখন আমরা আলাপ করি ডেটাকে কীভাবে জোগাড় করতে হয়?

ডেটা সোর্স

আমাদের দরকারমতো ডেটাকে কালেকশন (কালেক্ট) করতে বিভিন্ন ধরনের পদ্ধতি ব্যবহার করি। যেমন—

১. বিভিন্ন ধরনের প্রশ্ন তৈরি করে প্রশ্ন স্পেসিফিক ডেটা কালেকশন।

২. অনেকের সাথে সাক্ষাৎকার (কোয়েশ্চেনিয়ার) থেকে সেই ডেটাকে একটা জায়গায় নিয়ে গিয়ে সেটাকে বিশ্লেষণ।

৩. অনেকের কাছ থেকে কাগজপত্র, নথিকে একটা জায়গায় এনে সেটা থেকে বিশ্লেষণ।
৪. মেশিন থেকে লগ জেনারেট হলে সেটাকে প্রি-প্রসেসিং করে আমাদের নিজস্ব বিশ্লেষণ এবং
৫. ইদানীং আমরা ওয়েব স্ক্রাপিংয়ের মাধ্যমে বিভিন্ন ডেটা সংগ্রহ ইত্যাদি ইত্যাদি।

তবে এই অটোমেশনের যুগে সবচেয়ে বেশি ডেটা পাচ্ছি বিভিন্ন মেশিনের লগ থেকে। কারণ ইদানীং প্রচুর মেশিন আমাদের ডেটা দিচ্ছে কাজকে সহজ করার জন্য। আমাদের ডেটা আসছে এই যন্ত্রের ভেতর থেকে, বিশেষ করে বিভিন্ন ডেটা স্টোর হতে প্রাপ্ত ডেটা আমাদের অনেক ধরনের 'ইনসাইট' দিচ্ছে।

তবে আমরা যেভাবে ডেটা জোগাড় করি না কেন, এই প্রাপ্ত বা সংগৃহীত তথ্যকে আমরা বলছি 'কাঁচা' ডেটা অর্থাৎ 'র' ডেটা। যখন এ ধরনের ডেটা আমরা সরাসরি একটা প্রাথমিক সোর্স থেকে পাই, যাকে কোনোভাবেই 'ম্যানিপুলেশন' বা আলাদা প্রসেসিংয়ের মধ্যে যেতে হয়নি। আমরা যেটাকে বলতে পারি একটা সত্যিকারের ডেটা সোর্স থেকে ডেটা পাওয়ার পরে সেটাকে কোনোভাবেই সামারাইজেশন, অর্থাৎ সংক্ষিপ্তকরণ, আলাদা প্রসেসিং এবং কোনো ডেটা মুছে ফেলা, অর্থাৎ সরানো হয়নি তাহলে সেটাকে কাঁচা অর্থাৎ 'র' ডেটা বলতে পারি। আমরা সাধারণত, শুরু থেকে কাঁচা ডেটা (যেটা টেম্পারিং করা হয়নি) নিয়ে কাজ করব, যাতে আসল 'ইনসাইট' পাওয়া যায়।

কাঁচা মানে 'র' ডেটার কিছু উদাহরণ

সরাসরি ডেটা

১. বিভিন্ন মেশিন থেকে তৈরি বাইনারি ফাইল।
২. একদম ফরম্যাট না করা এক্সেল ফাইল, কমা সেপারেটেড ভ্যালু, ইত্যাদি যেখানে কোনো ডেটা ম্যানিপুলেশন করা হয়নি।
৩. বিভিন্ন এপিআই থেকে স্ক্র্যাপ করা 'জেসন' ডেটা।
৪. (একটা উদাহরণ হিসেবে), অণুবীক্ষণ যন্ত্রের মাধ্যমে দেখার সময় হাতে লেখা সংখ্যাগুলো সংগ্রহ করা ডেটা। ল্যাব থেকে প্রাপ্ত ডেটাও হতে পারে।

তথ্যসূত্র অর্থাৎ ডেটা সোর্স কোথা থেকে হতে পারে?

আমরা ডেটা জোগাড় করি যাতে সেই ডেটা থেকে সেটার 'এক্সপ্লোরেশন' এবং ঠিকমতো বুঝে অন্য সবার মধ্যে যাতে সেটাকে কমিউনিকেট করতে পারি। আপনারা দেখবেন, আমরা 'পাওয়ার বিআই' অথবা ট্যাবলো যেটাই ব্যবহার করি না কেন, শুরুতে আমরা ডেটা সোর্সের সাথে সেটাকে যুক্ত করে নিই। এভাবে আমরা বিভিন্ন ধরনের ডেটা সোর্সের সাথে যুক্ত করি, যাতে সেটাকে ঠিকমতো অ্যানালাইসিস করা যায়। এ ধরনের ডেটা সোর্স হিসেবে স্প্রেডশিট অর্থাৎ মাইক্রোসফট এক্সেল অথবা গুগল ওয়ার্কশিট, রিলেশনাল ডেটাবেইস যেমন, মাইএসকিউএল, ওরাকল, পোস্টগ্রেসকিউএল, ম্যাংগো ডিবি ইত্যাদি যোগ করা যায়। ক্লাউডের যোগ করার সুযোগ থাকার কারণে যেকোনো বিজনেস অ্যানালাইটিক্স সফটওয়্যার সরাসরি ক্লাউডের সাথে যুক্ত হয়ে অ্যানালাইসিস করতে পারে। এর ফলে আমাদের ডেটা অ্যানালাইসিস টুলগুলো ডেটাকে ঠিকমতো নিয়ে এসে সেটাকে ফ্ল্যাট স্ট্রাকচার অর্থাৎ রেকর্ডগুলোকে বিভিন্ন সারিতে ফেলতে পারে।

আমরা অনেক সময় 'র' ডেটা নিয়ে কাজ করতে চাই, যাতে ডেটা থেকে গুরুত্বপূর্ণ জিনিস কোনোভাবে মিসিং না হয়। ডেটাকে 'ম্যানিপুলেট' করলেও এর 'এসেন্স' ঠিক করার চেষ্টা করতে হবে।

ডেটার বৈশিষ্ট্য এবং ভেতরের ধারণা

ডেটার বৈশিষ্ট্য কী হতে পারে? ডেটার কী কী বৈশিষ্ট্য ভালো কাজ দিতে পারে?

প্রযুক্তির বই বাংলা লিখতে— এই একটা ঝামেলা। না করা যায় পূর্ণ বাংলা, না করা যায় পুরো ইংরেজি। সে কারণে, এই চ্যাপ্টারটা একটু খটমটে লাগতে পারে। তবে এ 'মাস্ট রিড' চ্যাপ্টার।

ডেটা লিটারেসি নিয়ে তথ্যপ্রযুক্তিবিদ স্টিফেন ফিউ এমন কিছু বৈশিষ্ট্য নিয়ে আলোচনা করেছেন, যা আমাদের ডেটা নিয়ে কার্যকরভাবে কাজ করতে সাহায্য করবে। ডেটা নিয়ে ভালোভাবে কাজ করতে গেলে আমাদের যেভাবে বিভিন্ন বৈশিষ্ট্যের দিকে তাকাতে হয়, সেভাবে ভালো ডেটার আরও কিছু বৈশিষ্ট্য রয়েছে। স্টিফেন ফিউর মতে, সে রকম উচ্চমানের ডেটার কিছু কিছু বৈশিষ্ট্যের প্রয়োজন, যা ডেটা বিশ্লেষণের একটি অপরিহার্য অংশ এবং জিনিসটি ডেটার গুরুত্বপূর্ণ অন্তর্দৃষ্টি দেয়। চলুন দেখে আসি ডেটার সে ধরনের কিছু বৈশিষ্ট্য।

মিনিংফুল অর্থাৎ অর্থপূর্ণ ডেটার কিছু বৈশিষ্ট্য

স্টিফেন ফিউ এমন কিছু বৈশিষ্ট্যের একটি তালিকা যোগ করেছেন, যা অর্থপূর্ণ ডেটার মধ্যে পাওয়া যাবে। আমরা যখন এই বৈশিষ্ট্যগুলো পড়তে থাকব, তখন আমাদের নিজের অভিজ্ঞতা আনতে চেষ্টা করব, বোঝার পাশাপাশি। আপনার কী মনে হয়? আমাদের ব্যক্তিগত বা কর্মজীবনে এ ধরনের উপযোগী ডেটার সম্মুখীন হয়েছিলাম কখনো? এই ডেটাগুলোর মধ্যে কোন ধরনের বৈশিষ্ট্য ছিল? আমরা কি এমন ডেটার মুখোমুখি হয়েছি, যা আমাদের ব্যক্তিগত বা অফিসের কাজে উপযোগী ছিল না? এই বৈশিষ্ট্যের মধ্যে কোনটিতে সেই ডেটার অভাব ছিল? আপনি নিজেকে প্রশ্ন করুন এই নিচের বৈশিষ্ট্যগুলোর মধ্যে থেকে।

ডেটার ভলিউম

প্রচুর পরিমাণে প্রাসঙ্গিক, অ্যাভেলেবেল ডেটা থাকতে হবে। (এবং যেই ডেটা আমার কাছে আছে) ডেটার অর্থ হলো আমাদের প্রশ্নের উত্তর দেওয়ার জন্য আমাদের যত ডেটা প্রয়োজন, সেখানে প্রাসঙ্গিক যত ডেটা হবে, ততই ভালো কাজ হবে। যত গুড় তত মিষ্টির মতো ব্যাপার কিছুটা। তবে সেই ডেটাটাকে প্রাসঙ্গিক হতে হবে আমাদের প্রশ্নের সাথে। আমরা অনেক সময় প্রচুর ডেটা কালেক্ট করি, সেখানে তার প্রাসঙ্গিকতা না থাকলে পুরো জিনিসটাই ভেস্তে যায়।

ডেটার ইতিহাস

আপনার মনে আছে— সময়ের সাথে সাথে চলা ডেটা আমাদের কীভাবে দেখতে দেয়? বিশেষ করে, সময়ের সাথে সাথে যেসব প্যাটার্ন তৈরি হয়েছে যার কারণে বর্তমান পরিস্থিতি কীভাবে তৈরি হয়েছে? যেমন আমাদের ডেটা ইতিহাস প্রয়োজন— গত ১০ বছরে বিক্রয়ের প্রবণতা বৃদ্ধি বা হ্রাস দেখার জন্য। ডেটার ইতিহাস দিয়ে আমরা কিছু প্যাটার্ন বের করতে পারি যাতে ভবিষ্যতে কী হতে পারে বা তার ফোরকাস্টিং কেমন হবে, সেটা বোঝা যায়।

ডেটার কনসিসটেন্সি অর্থাৎ সংগতি

সময়ের সাথে অনেক ভ্যারিয়েবল পাল্টাবে, তবে সেই পরিবর্তনের সাথে সাথে, ডেটার ভেতরে কনসিসটেন্সি অর্থাৎ সামঞ্জস্য না থাকলে সেই ডেটার মধ্যে অনেক ধরনের সমস্যা তৈরি হতে পারে। যেমন মুদ্রাস্ফীতির সাথে তখনকার বেতনের সামঞ্জস্য বজায় রাখা এর একটি ভালো উদাহরণ। আপনার কাছে ১০ বছর আগের এবং পরের ডেটাতে বেতনের কোনো পরিবর্তন না থাকলে, এর অর্থ হচ্ছে সেই দেশে কোনো মুদ্রাস্ফীতি ঘটেনি, যা একটা অবিশ্বাস্য ঘটনা। এর অর্থ হচ্ছে সময়ের সাথে সাথে বেতন বাড়বে সেটাই আমরা ধারণা করি। ফলে ডেটার কনসিসটেন্সি একটা বড় ইস্যু।

একাধিক ভ্যারিয়েবল হচ্ছে ডেটার বড় শক্তি

যেকোনো মিনিংফুল ডেটাতে পরিমাণগত কোয়ান্টেটিভি (সংখ্যাসূচক পরিমাপযোগ্য) এবং গুণগত কোয়ালিটেটিভ (বৈশিষ্ট্যগত, সংখ্যাসূচক

পরিমাপযোগ্য নয়) ভ্যারিয়েবল থাকতে হবে। ডেটাতে যত বেশি ভ্যারিয়েবল, তত বেশি আপনি এটি থেকে নতুন নতুন ধারণা আবিষ্কার করতে পারবেন। টাইটানিক ডেটাসেটের কথাই ধরুন। ছোট ডেটাসেট, তবে এর মধ্যে পরিমাণগত এবং গুণগত, কোয়ালিটেটিভ এবং কোয়ান্টিটেটিভ ডেটা থাকাতে এর থেকে ধারণা পাওয়া যায় বেশি। যত বেশি ভ্যারিয়েবল তত বেশি সেখান থেকে জ্ঞান পাওয়া যাবে। আবার বেশি ডেটা মানেই যে ভালো ডেটা, তা নয়।

লোয়েস্ট লেভেলের বিস্তারিত ডেটা

ডেটা যত সূক্ষ্মভাবে ডিটেইলিং লেভেল ব্যবচ্ছেদ অর্থাৎ বিশদ, আমরা সেখানে বিস্তারিতভাবে বিভিন্ন স্তরে এটি পরীক্ষা করতে পারব। উদাহরণস্বরূপ, আমরা যদি পুরো দেশে ভ্যাকসিনেশন প্রোগ্রাম ঠিকমতো চলছে কি না, সেটা জানার প্রবণতা বুঝতে চাই তাহলে এই একই কাজ শুধু শহর না, এর পাশাপাশি পুরো বাংলাদেশব্যাপী এই কাজটা কীভাবে চলছে তার জন্য বিভাগীয় জেলা এবং উপজেলা লেভেলে কীভাবে কাজ চলছে, সেটার প্রবণতাগুলোকে বোঝার জন্য সর্বনিম্ন লেভেলের বিস্তারিত ডেটা থাকা প্রয়োজন।

ডেটার ভেতরে 'ক্লিনলিনেস' অর্থাৎ শুচিতা ঠিকমতো বোঝা

আমাদের ধরে নিতে হবে, ডেটা অর্থপূর্ণ হওয়ার জন্য, পুরো ডেটাকে সম্পূর্ণ হতে হবে। এর মধ্যে ডেটা ভুল বা অসম্পূর্ণ হওয়া উচিত নয়। সর্বোপরি, ডেটার ভেতর ত্রুটি থাকা উচিত নয়। ডেটাকে ম্যানিপুলেট করা যাবে না শুরুতে।

ডেটার ক্ল্যারিটি, অর্থাৎ নির্মলতা

ডেটার ফিচারের নাম এমনভাবে লেখা উচিত, যা সহজে বোঝা যায়, কোডে লিখব না। উদাহরণ হিসেবে একজন মানুষের বিভিন্ন ফিচার নিয়ে কথা বললে, বয়সের জায়গায় ইংরেজিতে এইজ, নামের জায়গায় নাম, ঠিকানার জায়গায় ঠিকানা ইত্যাদি ইত্যাদি লেখা যেতে পারে। কিন্তু বয়সের জায়গায় বার ঠিকানার জায়গায় যদি শর্ট করে কিছু লেখা হয় তাহলে অন্য আরেকজন এই ফিচারগুলো দেখে সহসাই বুঝবেন না। তখন তাদের জন্য আলাদা করে ডিকশনারি বানিয়ে দিতে হবে, যা আসলে সময়ের অপচয়।

ডাইমেনশনের বিভিন্ন স্ট্রাকচার

ডেটাকে কাজ করাতে হলে সেটাকে স্ট্রাকচারাল সিস্টেমে ঢোকাতে হবে। ডেটাকে ঠিকমতো অ্যাকসেসযোগ্য করতে হলে এটিকে দুইভাবে স্ট্রাকচার করা যায়। ডাইমেনশন অর্থাৎ (গুণগত মান) এবং মেজার পরিমাপ (পরিমাণগত মান)। ডেটাকে ঠিকমতো ব্যাখ্যা করতে হলে একটা অর্গানাইজেশনাল টেবিল স্ট্রাকচার ব্যবহার করতে হবে, যাতে ডেটাকে ঠিকমতো বোঝা যায়।

ডেটার সেগমেন্টেশন, অর্থাৎ বিভিন্ন ভাগে ভাগ করতে পারা

একই ধরনের বৈশিষ্ট্যের ওপর ভিত্তি করে বিভিন্ন ডেটাকে গ্রুপে ভাগ করা যায়। এই কাজটি করতে কিছু কিছু বৈশিষ্ট্যকে গ্রুপিং করে সেটাকে সহজ বিশ্লেষণের জন্য এক জায়গায় রাখা যেতে পারে। আমাদের কাছে একই ধরনের ডেটাকে গ্রুপিং করে নিজস্ব বাস্কেট ভিত্তিতে ফেলা যেতে পারে। উদাহরণ হিসেবে আমরা যখন সিনেমার ডেটাগুলোকে এর জন্য বিভিন্ন জনরাকে ধরে তাকে একটা শ্রেণিবদ্ধ অর্থাৎ গ্রুপিং করা যেতে পারে (অ্যাকশন, সায়েন্স ফিকশন, রোমান্স, কমেডি ইত্যাদি)।

ডেটার উৎপত্তির ব্যাপারে স্বচ্ছ ধারণা, ট্রান্সপারেন্ট অরিজিন নিয়ে চিন্তা করা

ডেটা বিশ্বাস করার জন্য আমাদের জানতে হবে যে এটি একটি নির্ভরযোগ্য উৎস থেকে এসেছে এবং এটি বিশ্বাসযোগ্য উপায়ে প্রসেসিং করা হয়েছে। যে ডেটাকে আমরা বিশ্বাস করছি, তার সঠিক সোর্স এবং এর মাঝখানে কোনো ম্যানিপুলেশন হয়েছে কি না সেটা সম্বন্ধে নিশ্চিত থাকতে হবে।

এত বৈশিষ্ট্যের কারণে সমস্যা মনে হচ্ছে? চলুন, ডেটার অরগানাইজ করার কিছু ধারণা বোঝার চেষ্টা করি।

ডেটাকে ‘অরগানাইজ’ করার কিছু প্র্যাকটিস

এই চ্যাপ্টারটা ‘বোরিং’ মনে হলেও অনুরোধ করছি পুরোটা পড়ার জন্য। এটা ডেটার ব্যাপারে ‘বেজ লাইন’ তৈরি করে দেবে।

মনে আছে তো— আমরা ভালো মানের ডেটার কিছু বৈশিষ্ট্য নিয়ে আলাপ করেছিলাম? এই পর্যায়ে ডেটাকে কীভাবে ‘অরগানাইজ’ করতে হয়, সেটার একটা ভালো ধারণা থাকা উচিত। আমার দেখা মতে, এই ডেটাকে ঠিকমতো ‘অরগানাইজ’ অর্থাৎ কোন ধরনের ডেটাকে ‘ভার্টিক্যাল’ অথবা ‘হরাইজন্টাল’ অ্যাক্সিস লেভেলে ফেললে ব্যাপারটা কাজ করবে, সেটা নিয়ে আলাপ করতে চাই এখানে।

একটা টেবিলের কথা চিন্তা করুন। যেকোনো ডেটাকে শুরুতে ‘অরগানাইজ’ করা যায় তার সারি এবং কলামের ওপর নির্ভর করে। সব ডেটাকে ইচ্ছেমতো কলাম বা সারিতে ফেলা যায় না।

কলাম এবং সারি

ক. কলাম হচ্ছে যেখানে আমরা ডেটাকে লম্বালম্বিভাবে রাখি, অর্থাৎ যেকোনো ডেটাসেটের ভেতরে লম্বালম্বি কলামের মধ্যে ডেটার ফিচার অর্থাৎ ভ্যারিয়েবলগুলোকে রাখা হয়। এই ভ্যারিয়েবল অর্থ হচ্ছে যেকোনো ডেটাসেটের বিভিন্ন ফিচার— যেমন টাইটানিক ডেটাসেটের ‘ট্রেইন’ ডেটাফ্রেমে বারোটা ফিচার অর্থাৎ বারোটা কলাম নিয়ে কাজ করে। সাধারণত কলামে এই ফিচারগুলো রাখার ফলে বিভিন্ন ভ্যারিয়েবলের নামগুলো ফিচারের নাম ধরে বোঝা যায় জিনিসগুলো কী এবং এর ভেতরে কী আছে? আবারও বলছি— মনে আছে ভালো মানের ডেটার কিছু বৈশিষ্ট্যের কথা? টাইটানিক ডেটাসেটে নাম, বয়স, কোথা থেকে উঠেছেন এ ধরনের ফিচারগুলোর কাজ সেই ভ্যারিয়েবলের নাম থেকেই বোঝা যায়।

খ. এখানে ‘সারি’ মানে হচ্ছে ভূমির সাথে সমান্তরাল, অর্থাৎ একটা টেবিলের ওপরের কলামের বাম পাশ ধরে যেই সারিগুলোতে বিভিন্ন রেকর্ড রাখা হয় এখানে সেটার একেকটা ইউনিটকে সারি/রেকর্ড হিসেবে আমরা ব্যবহার করব অ্যানালাইসিসে। একটা কোম্পানির কাস্টমার ডেটাবেইসে কাস্টমারের নাম, বয়স, ঠিকানা ইত্যাদি কলামে গেলেও এর সারিতে চলে আসছে সেই কোম্পানির কাস্টমারদের রেকর্ড, অর্থাৎ একেক কাস্টমারের রেকর্ড একেকটা সারিতে। ধরা যাক, আমাদের কোম্পানিতে যদি ১০০০ কাস্টমার থাকেন, তাহলে সেই ডেটাসেটে ১০০০ সারি থাকবে। একেকটা সারি হচ্ছে একেকটা কাস্টমারের রেকর্ড। এদিকে কলাম হেডার হিসেবে যদি ফিচারের নাম চলে আসে, তাহলে মোট সারির সংখ্যা হবে ১০০১। কারণ, ওপরেরটা হেডার কলাম ট্যাবলো এবং মাইক্রোসফট এক্সেল এবং ‘পাওয়ার বিআই’ ধরনের বিজনেস ইন্টেলিজেন্স প্ল্যাটফর্মে এগুলোকে অনেক সময় ‘ভ্যালু’ বলা হয়।

পরিবারের ৪ জন সদস্যের একটা উদাহরণ

ধরুন— একটা পরিবার, সেই পরিবারের সদস্যসংখ্যা, তাদের বয়স, উচ্চতা এবং তাদের পছন্দের খাবার একটা টেবিলে ফেলতে চাই। সেটা কীভাবে ‘অরগানাইজ’ করা যায়? মানে, আমি বলছি সবচেয়ে সহজে কীভাবে আমরা ৪ জন সদস্যের একটা পরিবারের সবাইকে একটা টেবিলে ফেলতে পারি?

নাম	বয়স	উচ্চতা	প্রিয় খাবার
শামিমা	৮	৫০”	আইসক্রিম
জামাল	১২	৬৩”	পিৎজা
অন্তরা	৪২	৬৭”	দই
কামাল	৩৯	৭০”	কেক

চিত্র ৪৬: টেবিলে সবার নাম রেকর্ডে

এখানে ডেটাসেটের ফিচার হিসেবে নাম, বয়স, উচ্চতা এবং তার পছন্দের খাবার যদি ফিচার হিসেবে নিয়ে আসি, তাহলে এই ফিচারগুলো কোথায় যেতে পারে? অবশ্যই কলামে। ৪ জন সদস্যের নাম এবং আরও তথ্য অর্থাৎ রেকর্ডগুলো তাহলে কোথায় যেতে পারে? অবশ্যই 'সারি' হিসেবে। সেটাই আমরা এখন দেখছি ভিজ্যুয়ালি, একটা টেবিলে। সায়েন্টিফিক্যালি এভাবেই ডেটা অরগানাইজ করতে হয়।

আপনারা কী দেখলেন এখানে? এখানে প্রতিটা ভ্যারিয়েবল যাকে আমরা 'ফিল্ড' হিসেবে বলছি, সেগুলোকে 'অরগানাইজ' করা হয়েছে কলামে। এখানে একেকটা কলাম হচ্ছে নাম, বয়স, উচ্চতা এবং তাদের পছন্দের খাবার। এখানে প্রতিটা সারি যাকে আমরা একেকটা অ্যানালাইসিস ইউনিট হিসেবে ধরে নিচ্ছি। এই সারিগুলো একেকটা একেকজন মানুষের অর্থাৎ প্রতিটা সারি একজন মানুষকে 'রিপ্রেজেন্ট' করছে তাদের নাম, বয়স, উচ্চতা এবং পছন্দের খাবার নিয়ে। আমরা এখানে দেখছি শামিমা, অন্তরা, যার বয়স ৪২ এবং ইনি ৬৭ ইঞ্চি লম্বা। সবচেয়ে বড় কথা হচ্ছে, উনি দই খেতে খুব পছন্দ করেন।

ভালো স্ট্রাকচারাল উদাহরণ

ঠিক ধরেছেন। একটা ভালো ডেটা স্ট্রাকচার তৈরি করতে কলাম এবং সারিতে 'কী' 'কী' যেতে পারে, সেটা জানা খুবই জরুরি। কোনটা সারিতে গেলে ভালো— আর কোনটা কলামে গেলে ভালো ইনসাইট পাওয়া যেতে পারে— সেটা না জানলে পরে সমস্যা হয় বেশি। আমার জীবনে প্রচুর টেবিল দেখেছি বা ভুলভাবে ডিজাইন করার ফলে সেখান থেকে প্রজ্ঞা পাওয়া যায় না।

আরও বলেছি, নো ডেটাসেটের প্রতিটা ভ্যারিয়েবল, অর্থাৎ 'ফিচার' যাকে অনেক সময় আমরা 'ফিল্ড' হিসেবে ধরে থাকি, সেগুলো যাবে একেকটা কলামে। একটা ফিচার— একটা কলাম। সবচেয়ে ওপরের সারিটাকে আমরা সেই ভ্যারিয়েবল দিয়ে কলাম হেডার বানাতে পারি। আমাদের আগের টেবিলে নাম, বয়স, উচ্চতা এবং পছন্দের খাবার হচ্ছে আমাদের ফিল্ড এবং সবগুলো মিলিয়ে একটা কলাম হেডার।

প্রতিটা সারিতে আলাদা আলাদা করে যে রেকর্ড সেটা একেকজনের ‘অবজারভেশন।’ আমরা এই রেকর্ডগুলোকে আলাদা আলাদা ইউনিট হিসেবে ধরে নিই। যেমন এখানে প্রতিটা মানুষের কিছু বৈশিষ্ট্য নিয়ে একেকটা অবজারভেশন যা প্রতিটা মানুষকে আলাদাভাবে বুঝতে সাহায্য করছে। যেমন অন্তরা একজন মাঝারি গড়নের সবেমাত্র তরুণী, যিনি দই পছন্দ করেন। এটার কিছু আলাদা ইমপ্লিকেশন আছে। যারা উঠতি বয়সের তরুণী, দই পছন্দ করেন, তারা সবার সাথে মানিয়ে চলতে ভালোবাসেন। এটা আসছে আলাদা ডেটা থেকে। (সেটা হতে পারে— সাইকোমেট্রিক প্রোফাইলিং ডেটাসেট)।

সতর্কতা: আমার ৩০ বছরের অভিজ্ঞতা বলে কোন ‘স্পেসিফিক’ তথ্য কলাম এবং কোনটা সারিতে যাবে, এটা শুরুতেই ঠিক না করা হলে (মানে না বুঝলে) পরবর্তী সময়ে বড় সমস্যায় পড়তে হয়। আমি দেখেছি, ঠিকমতো ডেটা স্ট্রাকচার অর্থাৎ আসল ফরম্যাটে ডেটাকে ঠিকমতো শুরুতেই স্ট্রাকচার না করতে পারলে পরবর্তী সময়ে বছর বছর ধরে এই অসংগতি সম্পন্ন টেবিল এবং ডেটা ফরম্যাটকে টানতে হয়। আপনারা দেখবেন— প্রচুর সরকারি এবং বেসরকারি অফিসে টেবিল স্ট্রাকচার ঠিকমতো কলাম এবং সারি ফরম্যাট না করার ফলে পরবর্তী সময়ে সেটাকে সফটওয়্যার নিয়ে অ্যানালাইসিস করা সম্ভব হয় না। এটা সিস্টেমের বড় ক্ষতি।

আমি অনেক সময় দেখেছি, যে ধরনের ফিল্ডগুলোকে কলামে নেবার কথা— সেগুলোকে সারিতে নেবার ফলে প্রচুর ঝামেলা হয়। টাইটেলে কী কী যাবে সেটাকে সারিতে ফেললে তার পাশাপাশি কলাম হেডারে অর্ধেক অংশ এবং সারিতে বাকি অংশ নেবার ফলে পুরো প্রক্রিয়াটা আরো জটিল পর্যায়ে দাঁড়ায়। এক টেবিলে কোন কোন কলাম প্রযোজ্য, তার বাইরে অতিরিক্ত কলাম ডেটা অ্যানালাইসিসকে সমস্যায় ফেলে। সব সময় ‘লেস ইজ মোর’। বেশি ডেটা মানেই ভালো না। অনেক সময় কলাম হেডারগুলোকে সাবটাইটেল হিসেবে ফরম্যাট করলে সেটাকে ঠিকমতো কোথাও ফেলা যায় না।

আমি এখানে শুরুতে কিছু অসংগতি সম্পন্ন টেবিল দেখাতে চাচ্ছিলাম, তবে বইয়ের পৃষ্ঠার সীমাবদ্ধতা লক্ষ করে সেটাকে বাদ দিয়েছি। গুগল করতে পারেন, ‘ডেটাবেইস ডিজাইন ব্যাড প্র্যাকটিসেস’। তবে মজার

কথা হচ্ছে, পুরোনো অসংগতি সম্পন্ন টেবিলগুলোকে বর্তমানে ডেটা অ্যানালাইসিস টুল যেমন 'আর' অথবা 'পাইথন' এবং বিভিন্ন বিজনেস অ্যানালাইটিক্স টুল যেমন ট্যাবলো, পাওয়ার বিআই, মাইক্রোসফট এক্সেল, পাওয়ার কোয়েরি এডিটরের 'ই টি এল' অর্থাৎ 'এক্সট্রাক্ট ট্রান্সফর্ম এবং লোড' পাইপলাইনে একটা ভালো স্ট্রাকচারের টেবিলে রূপান্তর করা যায়।

পুরোনো অসংগতিপূর্ণ টেবিলকে নতুনভাবে স্ট্রাকচার করা

আমাদের পুরোনো টেবিল যদি অসংগতিপূর্ণ হয় তাহলে সেটাকে নতুনভাবে স্ট্রাকচারে ফেলার জন্য আমাদের বিভিন্ন টুল ব্যবহার করার আগে সেটার আন্ডারলায়িং সমস্যাটা বুঝতে হবে। ডেটাবেইসকে নতুন করে স্ট্রাকচার করার জন্য সেটাকে কাগজে ফেলে কোথায় কোন ধরনের তথ্য গেলে ডেটা অ্যানালাইসিস করতে সুবিধা হবে, সেটা বুঝতে পারলেই কাজটা সহজ হবে। অর্থাৎ আগে কাগজে ফেলতে হবে। আমি এই কাজে কম্পিউটারের কাছ থেকে দূরে থাকি হাজার মাইল। ডিজাইন সব সময় কাগজে ভালো। আমার বন্ধুদের কল্যাণে গুগল, ফেসবুক ইত্যাদি অফিসে দেখে এসেছি সনাতন কাগজের ড্রয়িংবোর্ডের দৌরাত্ম্য।

এর পরের অংশে আসে টুলের ব্যবহার। এ পর্যায়ে আমরা সেই ডেটা টেবিলে আসল ডেটাকে মডিফাই করার আগে ডেটাকে কাস্টমাইজ করা যেতে পারে তার 'মেটা ডেটা' দিয়ে। এগুলো এখন ঠিকমত না বুঝলেও আমি বলছি। কারণ, আমাদের প্লেলিস্ট ভিডিওগুলোতে এসব কথাকে কানেক্ট করতে সুবিধা হবে। আমাদের এই 'মেটা ডেটা' মানে এটার বিভিন্ন ধরনের নতুন ক্যালকুলেশন ফিল্ডগুলোকে পাল্টে নতুন নাম দেওয়া এবং এর পাশাপাশি কোন ডেটাকে কোথায় ফেলতে হবে, সেটার ধারণাকে একটা পর্যায়ে আবার কাগজে ফেলা।

আবার কাগজ কেন?

আমি দেখেছি যে কোনো ডেটা স্ট্রাকচারের শুরুতে— সেটাকে যদি আমরা কাগজ-কলমে ফেলি তাহলে মস্তিষ্ক সেটাকে ঠিকমতো অ্যানালাইসিস করতে পারে। যেকোনো ধরনের ডেটা স্ট্রাকচার তৈরির শুরুতেই আমরা কাগজ-কলম নিয়ে বসব। সত্যি বলছি, এখানেও কম্পিউটার আসবে অনেক

পরে। শুরুতে কাগজ এবং পেনসিল দিয়ে কোথায় কোন ডেটা ফেললে অর্থাৎ কোনটা কোনদিকে ফেললে আমাদের জন্য ডেটাকে অ্যানালাইসিস এবং ‘আউটকাম’ পেতে সুবিধা হবে, সেটা বের করতে পারলেই আমাদের কাজ সহজ হয়ে যাবে। দক্ষতার জন্য এ ধরনের কাস্টমাইজেশন সবাই করে থাকে, তাদের ডেটাবেইস যতই পুরোনো হোক না কেন। এ ধরনের কাজ আমরা অহরহ দেখব বিজনেস অ্যানালাইটিক্স টুলে। কম্পিউটার যে সবকিছুতে ভালো সেটা কিন্তু নয়।

একদম বেসিক লেভেলে ডেটাকে ৩ ভাগে স্ট্রাকচার করা যায় এ মুহূর্তে—

১. কলাম থেকে সারিতে অথবা সারি থেকে কলামে ‘পিভটিং’ (সামনে আসছি)

২. ফিল্ডকে আরও সাব লেভেলে ভাগ না করে ওখানেই ‘স্প্লিট’ করা

৩. সংখ্যা অথবা ক্যাটাগরিক্যাল ভ্যারিয়েবল হলে ডেটার চাহিদা অনুযায়ী নির্দিষ্ট ফিল্ডে ডেটা ফেলা

চলুন, সামনের চ্যাপ্টারে।

বইটা না পড়ে ভিডিও প্লেলিস্টে আগেই যাবেন না। ভিডিও বুঝতে ভালো, তবে শুরুতে টেক্সট না পড়লে বেসিক ধারণা সেই তিমিরেই থেকে যায়।

‘পিভোট’, কলাম স্প্লিটিং ব্যাপারগুলো কী?

শুরুতেই প্রশ্ন আসতে পারে ‘পিভোট’ ব্যাপারটা কী?

বেশি ভেতরে না গিয়ে বলতে হয় যে— কোনো টেবিলের কলামগুলোকে সারি এবং সারিগুলোকে কলামে ‘ট্রান্সফার’ করার পদ্ধতিকে আমরা ‘পিভোট’ পদ্ধতি বলতে পারি। আমরা এভাবে বলতে পারি— ‘পিভোট’ পদ্ধতি ব্যবহার করার জন্য, আমাদের কাছে যদি এ ধরনের কোনো ডেটাসেট থাকে যার একপাশে লম্বালম্বিভাবে কলামের সংখ্যা বেশি এবং কলামগুলোতে বেশির ভাগই একই ধরনের তথ্য থাকে; বিশেষ করে এটা কিছুটা কলামনির্ভর ডেটাসেট, যেখানে কাছাকাছি সব ডেটা কলামেই রাখা হয়। আমাদের হাতের লেখা খাতা-কলমে টেবিলগুলোর মতো। এ ধরনের টেবিলগুলোকে সাধারণত আমাদের নিজেদের সুবিধার জন্য তৈরি করে থাকি— যেগুলোকে বিভিন্ন ইউজার রিপোর্ট হিসেবে ব্যবহার হয়। আমরা যখন এ ধরনের স্ট্রাকচার নিয়ে কাজ করি, তখন অ্যানালাইসিসের ক্ষেত্রে এগুলো বেশ কিছু সমস্যা তৈরি করে। যেহেতু এটা ‘মানবিক’, অর্থাৎ ‘হিউম্যান সেন্টারড মাইন্ডসেট’ থেকে তৈরি করা হয় বলে এর ভেতর থেকে অ্যানালাইসিস করে ডেটা বের করা দুষ্কর।

মানসিক ধারণা থেকে টেবিল

একটা উদাহরণ নিয়ে কথা বললে ব্যাপারটা আরও পরিষ্কার হবে। নিচের টেবিলটা দেখুন, যেখানে কয়েকজন ব্যক্তি পার্কিং ফি দিয়েছেন বিভিন্ন সময়ে। টিকিট কাউন্টারে একজন মানুষ যখন তার খাতাকলমে প্রতিটা ব্যক্তির জন্য আলাদা আলাদা করে প্রতিদিন ধরে পার্কিং টিকিটের হিসেব করেন, তখন সেটা এ ধরনের একটা টেবিলে দাঁড়ায়।

নাম	২/৫/২০২১	২/৬/২০২১	২/৭/২০২১	২/৮/২০২১	২/৯/২০২১
জামিল	১০	১০	১০	১০	১০
রফিক	১০				
কমল	১০				১০
কামাল	১০		১০		

টেবিল ২: তারিখ ধরে তৈরি টেবিল, দক্ষ নয়

এ ধরনের ডেটা স্ট্রাকচার আমরা যদি 'মাইক্রোসফট এক্সেল' / 'পাওয়ার বিআই' অথবা 'ট্যাবলো'তে ব্যবহার করি, তাহলে প্রতিটা কলামের জন্য একটা করে ফিল্ড পাব। ভালো করে লক্ষ করলে দেখা যাবে যে এখানে পাঁচজন মানুষের জন্য একেক দিন করে কত টাকা পার্কিং ফি হচ্ছে, সেটা দেখা যাচ্ছে। তবে এ থেকে 'সরাসরি' কোনো ধরনের অ্যানালাইসিস করা যাচ্ছে না— যেহেতু সময়ের সাথে ডেটাগুলো আলাদা আলাদা ফিল্ডে চলে গেছে। এটা ডেটা অ্যানালাইসিসের জন্য একটা বড় অন্তরায়। ব্যাপারটা বোঝার চেষ্টা করুন।

আর সে কারণেই এমপ্লয়ির নাম, তারিখ এবং পার্কিং ফি এগুলো কলামে যাওয়া উচিত কলামের হেডারসহ। এর পাশাপাশি, প্রতিটা মানুষের নাম আমরা নিতে পারি সারিতে যা 'অবজারভেশন', অর্থাৎ 'ভ্যারিয়েবলের অবজারভেশন ভ্যালু' হিসেবে ব্যবহার হবে। ঠিক তো? এখানে সারি হিসেবে সে কারণে ব্যক্তির নাম তার সাথে সেই দিনের তারিখ এবং সেই দিনের তারিখে কত টাকা পার্কিং ফি দেওয়া হয়েছে সেটার হিসেব থাকলে আমাদের জন্য অ্যানালাইসিস করতে সুবিধা হবে। এই পিভোট পদ্ধতির পরে আমরা পুরোনো টেবিলটাকে নতুন টেবিলে দেখলে ব্যাপারটা আরও পরিষ্কার হবে। আগের লম্বালম্বি টেবিলটা এখন হরাইজন্টাল লম্বা হিসেবে চলে এসেছে, যাতে ক্যালকুলেশন সহজ হয়। এখানেই মানুষের সাথে যন্ত্রের হিসাব কিতাব আলাদা।

নাম	তারিখ	পার্কিং ফি
জামিল	২/৫/২০২১	১০
জামিল	২/৬/২০২১	১০
জামিল	২/৭/২০২১	১০
জামিল	২/৮/২০২১	১০
জামিল	২/৯/২০২১	১০
রফিক	২/৫/২০২১	১০
কমল	২/৯/২০২১	১০
কমল	২/৫/২০২১	১০
কামাল	২/৫/২০২১	১০
কামাল	২/৭/২০২১	১০

টেবিল ৩: ডেটা অ্যানালাইসিসের জন্য টেবিল

কলাম স্প্লিটিং

মনে আছে ওই ধরনের টেবিলের কথা— যেখানে একটা কলামের মধ্যে অনেকগুলো তথ্য থাকে, যা মানুষের চোখের জন্য সহনীয় হলেও অ্যানালাইসিসের জন্য সেটা সমস্যায় দাঁড়ায়। সে কারণে অনেক সময় আমরা একটা কলামের তথ্যগুলোকে ভাগাভাগি করে ফেলি বিভিন্ন কলামে— যাতে কলাম ধরে সেগুলোকে 'সর্টিং' এবং অন্যান্য অ্যানালাইসিস করতে সুবিধা হয়। এই ধরনের স্প্লিটিং সাধারণত একটা 'স্ট্রিং' ভ্যালুর মধ্যে যেমন 'কমা' (,) অথবা কোলন (-) হাইফেন হিসেবে থাকে, যেখানে এই জিনিসগুলোকে আলাদা আলাদা করে বিভিন্ন কলামে ফেললে সেগুলোকে অ্যানালাইসিস করতে সুবিধা হয়।

এখানে একটা উদাহরণ দেখি। এটা একটা খেলনা উদাহরণ, তবে এ ধরনের অনেক বড় কাজ দেখবেন ভিডিওর প্লেলিস্টে।

এয়ারপোর্টের নাম
ঢাকা: DAC
চট্টগ্রাম: CGP
সিলেট: ZYL
কক্সবাজার: CXB

এয়ারপোর্ট	কোড
ঢাকা	DAC
চট্টগ্রাম	CGP
সিলেট	ZYL
কক্সবাজার	CXB

টেবিল ৪: নাম দিয়ে কোডকে আলাদা করা

নাম
আরিফুর রহমান : ৩৪
আমিনুল করিম : ৪৭
রফিকুল বারী : ২৫
কানিজ ফাতেমা : ২৭

নাম	বয়স
আরিফুর রহমান	৩৪
আমিনুল করিম	৪৭
রফিকুল বারী	২৫
কানিজ ফাতেমা	২৭

টেবিল ৫: নাম এবং বয়স আলাদা করে কলাম

এ ধরনের স্প্লিটিং— একটা ফিল্ড কম্পোনেন্টের মধ্যে আলাদা আলাদা ধরনের 'মিনিংফুল' ইনফরমেশনগুলোকে বিভিন্ন কলামে ফেলে— যাতে সেগুলোকে 'ডেটা অ্যানালাইসিস' টুল ঠিকমতো বুঝতে পারে। আমাদের ওপরের উদাহরণে সে রকম একটা ধারণা দিয়েছে, যেখানে একটা এয়ারপোর্টের নাম এবং সেই এয়ারপোর্টের কোড মাঝখানে 'কোলন' দিয়ে ভাগ করা থাকলেও সেটাকে আলাদা কলামে না ফেললে ডেটা অ্যানালাইসিস টুল এটা নিয়ে সমস্যায় পড়তে পারে। এই স্প্লিটিংয়ের কারণে এখন এয়ারপোর্টের নাম এবং এয়ারপোর্টের কোড এখন আলাদা আলাদা কলামে চলে গিয়েছে। পরের উদাহরণ দেখুন। অনেক সময় আমি একই কলামে নামের সাথে বয়স দেখি [রহমান খান ৩৭], সেখানের নাম একটা কলাম এবং বয়স আরেকটা কলামে ফেললে ডেটা অ্যানালাইসিসে অনেক সুবিধা হয়।

বিভিন্ন ভ্যারিয়েবল, ফিল্ড টাইপ নিয়ে কিছু ধারণা

কোয়ালিটেটিভ এবং কোয়ান্টেটিভ ভ্যারিয়েবল

Data is comprised of values of qualitative or quantitative variables belonging to a set of items.

আগের অনেকগুলো আলাপে আমরা দেখেছিলাম কীভাবে 'কোয়ালিটেটিভ' (গুণগত) ভ্যারিয়েবল এবং 'কোয়ান্টিটেটিভ' (সংখ্যায়, যেটাকে যোগ, বিয়োগ, গুণ, ভাগ ইত্যাদি করা যায়) ভ্যারিয়েবলকে আলাদা করা যায়। এখানে 'কোয়ালিটেটিভ' ভ্যারিয়েবল হচ্ছে সেগুলো— যা সংখ্যা দিয়ে মাপা যায় না; বিশেষ করে কিছু কিছু ক্যাটাগরি, অথবা নাম, জায়গার নাম ইত্যাদি। এর পাশাপাশি সংখ্যায় যেটাকে মাপা যায়, সেটাকে 'কোয়ান্টিটেটিভ' ভ্যারিয়েবল বলছি। আমরা যখন এই আইটেমগুলোকে একটা ডেটাসেটে যোগ করি— সেখানে এ ধরনের কোয়ালিটেটিভ ভ্যারিয়েবল একই বৈশিষ্ট্যের কোয়ালিটেটিভ ফিল্ড (অথবা কলাম বলতে পারি) এবং কোয়ান্টিটেটিভ ভ্যারিয়েবলগুলো সেই একই কোয়ান্টিটেটিভ ফিল্ড অথবা কলামে পৌঁছায়। বোঝার জন্য আমরা কি এর আগের টেবিলটা দেখতে পারি?

নাম	বয়স	উচ্চতা	প্রিয় খাবার	ডামি ভ্যারিয়েবল
শামিমা	৮	৫০'	আইসক্রিম	০০
জামাল	১২	৬৩'	পিৎজা	০১
অন্তরা	৪২	৬৭'	দই	১০
কামাল	৩৯	৭০'	কেক	১১

টেবিল ৬: কোয়ালিটেটিভ এবং কোয়ান্টিটিভ ভ্যারিয়েবল

একটা আইটেমের সেটে কোয়ালিটেটিভ এবং কোয়ান্টিটেটিভ নিয়েই একটা ডেটাসেট বানানো যায়। আগেই বলেছি— একটা হচ্ছে গুণগত মান আরেকটা হচ্ছে সংখ্যার মান। বিভিন্ন ক্যাটাগরিক্যাল ভ্যারিয়েবলগুলোকে সাধারণত গুণগত মান দিয়ে মাপা করা যায়। আপনার-আমার নাম, আমরা কোন এলাকায় থাকি, আমাদের প্রিয় খাবার কী, ইত্যাদি সেই গুণগত বৈশিষ্ট্যের মধ্যে পড়ে যায়। কম্পিউটার যেহেতু ভেতরে সংখ্যা ছাড়া কিছু বোঝে না, সেখানে আমরা এই ক্যাটাগরিক্যাল ভ্যারিয়েবলকে বিভিন্নভাবে সংখ্যায় কনভার্ট করে নিয়ে আসি (গুগল 'ডামি ভ্যারিয়েবল') যাতে অ্যানালাইসিসের সমস্যা না হয়। তবে সেটা এখানে বিবেচ্য নয়। যেহেতু গুণগত মানসম্পন্ন বৈশিষ্ট্যের ভ্যারিয়েবেলগুলো কোনো একটা জিনিসের মানের ওপর নির্ভর করে, সে কারণে এ ধরনের কোয়ালিটেটিভ ভ্যারিয়েবলগুলোকে আরও দুই ধরনের ক্লাসিফিকেশন ফেলা যায়। এর একটা হচ্ছে নমিনাল এবং আরেকটা হচ্ছে অর্ডিনাল। না বুঝলে আমরা ছবি দেখে বুঝতে পারি।

আমাদের আগের টেবিলে যেমন 'মানুষের নাম' এবং 'পছন্দের খাবার' হচ্ছে কোয়ালিটেটিভ ফিল্ড, যেগুলোকে সংখ্যায় মাপা যায় না। এর পাশাপাশি 'বয়স' এবং উচ্চতা হচ্ছে কোয়ান্টিটেটিভ যাকে সংখ্যায় মাপা যায়। আমরা চিন্তা করছি যে কীভাবে এই কোয়ান্টিটেটিভ এবং কোয়ালিটেটিভ ভ্যারিয়েবেলগুলো নিয়ে আমাদের বোঝাপড়া শুরু হতে পারে। এখন আমাদের দেখতে হবে, এই ভ্যারিয়েবলগুলোকে আরও কী কী ভাবে ক্ল্যাসিফাই করা যায়?

এর পাশাপাশি কোয়ান্টিটেটিভ ভ্যারিয়েবল— যেগুলোকে সংখ্যায় ভাগ করা যায়— যেটাকে আমরা টাকা, মাসের সংখ্যা, উচ্চতা, ওজন ইত্যাদিতে ফেলা যেতে পারে। একটা ডেটাসেটে গুণগত মানের ভ্যারিয়েবলগুলো সাধারণত সেই ধরনের বৈশিষ্ট্যের কলাম হিসেবে দাঁড়ায়— যা মানুষ সরাসরি ভ্যারিয়েবলগুলোর নাম দেখেই বুঝতে পারে। এর পাশাপাশি, সংখ্যার মান নিয়ে যেমন বয়স, ওজন, উচ্চতা সেগুলো ভিন্ন ধরনের কলাম হয়ে থাকে এবং এগুলো সংখ্যায় থাকে। কেন একই কথা বারবার বলছি, বলুন তো? এই বেসিকগুলো ঠিকমতো বুঝলে— সামনে আমাদের কাজ সহজ হয়ে যাবে, বিশেষ করে ভিডিও লেকচারে।

এখন ফিরে আসি, নমিনাল কোয়ালিটেটিভ ভ্যারিয়েবল নিয়ে আলাপে। এটা এমন ধরনের ভ্যারিয়েবল অর্থাৎ ক্যাটাগরি— যেগুলোকে আলাদাভাবে একটার সাথে আরেকটার র‍্যাঙ্কিং অর্থাৎ কোনটা ওপরে বা কোনটা নিচে সেটাকে ভাগ করা যায় না। উদাহরণ হিসেবে আমরা যদি বেশ কয়েকটা ফলের নাম বলি— যেমন কলা, আনারস, আম, কাঁঠাল— এগুলোকে আমরা নমিনাল ভ্যালু বলি, কারণ এগুলোর মধ্যে কোনো ধরনের র‍্যাঙ্কিং স্ট্রাকচার করা যায় না। কোনটা উপরে বা কোনটা নিচে এ ধরনের কোনো ব্যাপার নেই এই নমিনাল কোয়ালিটেটিভ ভ্যারিয়েবেলগুলোতে। আমরা যদি সাধারণভাবে এটাকে মনে রাখতে চাই তাহলে এ ধরনের নমিনাল ভ্যারিয়েবলগুলোকে আমরা নাম হিসেবে উল্লেখ করতে পারি— যেখানে কেউ কারও থেকে ওপরে না।

অর্ডিনাল কোয়ালিটেটিভ ভ্যারিয়েবল

নমিনাল কোয়ালিটেটিভ ভ্যারিয়েবল এবং অর্ডিনাল ভ্যারিয়েবলের মধ্যে একটা বড় পার্থক্য আছে। কী হতে পারে সেটা? এগুলোকে র‍্যাঙ্কিং অর্থাৎ কোনটা একটার ওপরে বা কোনটা আরেকটার নিচে, সেটা বলে দেওয়া যায়। তারা কোয়ালিটেটিভ, তবে তাদের সংখ্যায় প্রকাশ করা যায় না। উদাহরণ হিসেবে আমরা বিভিন্ন সার্ভেতে কোনো সার্ভিস নিয়ে আলাপ করতে গেলে 'মোটেই খুশি না', 'অল্প খুশি', 'মোটামুটি', 'বেশ ভালো' অথবা 'অসাধারণ সার্ভিস' এ রকম দেওয়া হলে তখন সেগুলোকে অর্ডিনাল কোয়ালিটেটিভ ভ্যালু হিসেবে ধরা যেতে পারে। আগের ছবি দেখুন এটাকে মনে রাখার সবচেয়ে ভালো উপায় হচ্ছে অর্ডিনাল, মানে অর্ডারলি ফ্যাশন অর্থাৎ এটাকে র‍্যাঙ্কিং করা যায়। তবে বিভিন্ন মেশিন লার্নিং মডেলে আমরা এ ধরনের র‍্যাঙ্কিংকে নিউমেরিক ভ্যালু, অর্থাৎ সংখ্যায় কনভার্ট করে নিই যাতে খুবই খারাপ অবস্থাকে '০' এবং 'অত্যন্ত খুশি'কে ১০ নম্বর দিয়ে কনভার্ট করে দেওয়া যায়।

আচ্ছা একটা বাসায় বাচ্চাদের সংখ্যা কী হতে পারে? কোয়ান্টিটেটিভ ভ্যারিয়েবল।

ভিজ্যুয়ালাইজেশনে ভ্যারিয়েবল

ভিজ্যুয়ালাইজেশন ভ্যারিয়েবল কীভাবে ফেলা যেতে পারে?

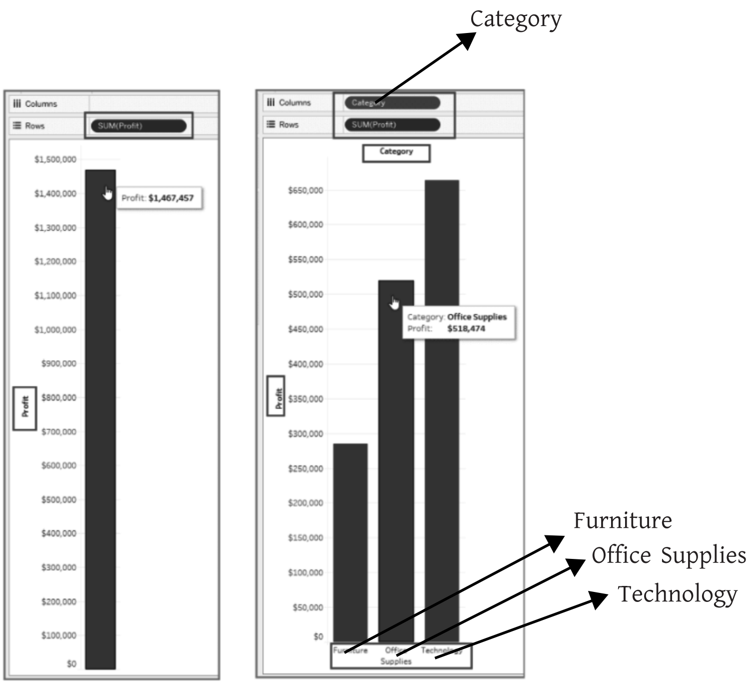
আমরা যখন এই ভ্যারিয়েবলগুলোকে একটা ডেটাসেটে ব্যবহার করি, তখন সেগুলোকে সাধারণত 'ফিল্ড' হিসেবে অর্থাৎ কলামে ফেলতে চাই। এখানে 'ভেরি গুড' কোয়ালিটেটিভ অথবা কোয়ান্টিটেটিভ– দুটোই হতে পারে। চলুন, আমরা কিছু ভিজ্যুয়ালাইজেশন দেখি যেখানে এ ধরনের দুটো ভ্যারিয়েবলকে প্লট করতে পারি।

কোয়ান্টিটেটিভ এবং কোয়ালিটেটিভ ভ্যারিয়েবলগুলোর আলাদা ধরনের ব্যবহার আছে ভিজ্যুয়ালাইজেশনে:

ক. কোয়ান্টিটেটিভ ভ্যারিয়েবল হচ্ছে সেই ধরনের ডেটা এলিমেন্ট, যাকে আমরা 'ক্যালকুলেট' করতে পারি। এর অর্থ হচ্ছে– তাদের আমরা 'এগ্রিগেশন', যোগ (সাম), গড় (এভারেজ) করতে পারা হচ্ছে 'এগ্রিগেশন' এর দুটো উদাহরণ।

খ. কোয়ালিটেটিভ ভ্যারিয়েবল ভিজ্যুয়ালাইজেশনের বিভিন্ন লেভেলে ডিটেইলিং তৈরি করে। এটা করার কারণ হচ্ছে– আমরা যখন ডেটার মধ্যে বিভিন্ন ক্যাটাগরির ভাগ, আলাদা করে সেগমেন্ট করে সেটাকে বুঝতে চাই, যা আমাদের সাহায্য করে ডেটার ভেতরের ডিটেইলস ঠিকমতো ধারণা করতে। আমরা যখন এক্সেল অথবা পাওয়ার বিআই নিয়ে কাজ করব, তখন শুরুতে এ ধরনের ধারণা কাজে না লাগলেও, এই ব্যাপারগুলোতে 'ট্যাবলো' অসাধারণ মানের প্রজ্ঞা দেয়। বেসিক জিনিস নিয়ে কাজ করার জন্য 'ট্যাবলো' এখনো অসাধারণ। এ কারণে শুরুতে আমি 'ট্যাবলো' নিয়ে কাজ করার জন্য অনুরোধ করব।

নিচে ট্যাবলোর দুটো উদাহরণ দেখুন।



শুধু কোয়ালিটেটিভ ভ্যারিয়েবল

কোয়ান্টিটেটিভ এবং কোয়ালিটিভ ভ্যারিয়েবল দিয়ে তৈরি

চিত্র ৪৭: দুটো ভ্যারিয়েবল এক সাথে দেখা

আমরা ওপরে বাম পাশে (ভিজ্যুয়ালাইজেশন) বিশ্লেষণটা তৈরি করেছি—সেটা একটা কোয়ান্টিটেটিভ ভ্যারিয়েবল, যেখানে প্রফিটকে দেখিয়েছি। এখানে প্রফিট হচ্ছে সংখ্যাগত একটা মান যাকে একটা বার-চার্ট দিয়ে আমরা রিপ্রেজেন্ট করছি। বার চার্ট দিয়ে প্রফিট কতটুকু সেটা এক টানেই আমরা বুঝতে পারছি। এখানে প্রফিটের কথায় সংখ্যাই হচ্ছে মূল। এবং সংখ্যা দিয়েই একটা ভিজ্যুয়ালাইজেশনের ভিত্তি তৈরি হচ্ছে।

ওপরের ডান পাশের ছবিটি দেখুন। এখানে আমরা কোয়ান্টিটেটিভ ভ্যারিয়েবলের পাশাপাশি কোয়ালিটেটিভ ভ্যারিয়েবলকে এনেছি। এই ভিজ্যুয়ালাইজেশনে একটা ব্যবসাপ্রতিষ্ঠানের প্রফিটের পাশাপাশি কোন কোন ব্যবসায়িক ক্যাটাগরিগুলো কী পরিমাণ প্রফিট করছে, সেগুলো দেখাতে পারছি। এখানে কোয়ান্টিটেটিভ ভ্যারিয়েবল হচ্ছে পুরো ব্যবসার প্রফিট, যা এগ্রিগেটেট, অর্থাৎ ‘সাম’।

তবে এর পাশাপাশি, যখন তিনটা কোয়ালিটেটিভ ভ্যারিয়েবল যেটাকে আমরা ক্যাটাগরি হিসেবে দেখাচ্ছি, সেখানে সেই ক্যাটাগরি অনুযায়ী বিভিন্ন প্রফিট সেগমেন্ট হিসেবে আমরা দেখছি। এর মধ্যে ফার্নিচার, অফিস সাপ্লাই এবং টেকনোলজি আলাদা করে দেখতে পাচ্ছি। একই ভিজ্যুয়ালাইজেশন– তবে দুই ধরনের ভ্যারিয়েবল দুটো আলাদা ডাইমেনশন এনে দিয়েছে।

ডিসক্রিট, কন্টিনিউয়াস ভ্যারিয়েবল এবং এগ্রিগেশন

যেহেতু ডেটার মধ্যে সংখ্যার খেলা, সে কারণে সংখ্যা নিয়ে কাজ করার সময় আমরা এর দুটো ভাগ নিয়ে কথা বলতে পারি।

ডিসক্রিট ভ্যারিয়েবল: ডিসক্রিট– মানে পূর্ণাঙ্গ, যেখানে সংখ্যাগুলো আলাদাভাবে 'ডিস্টিনক্ট' অর্থাৎ এগুলো একটা পূর্ণ সংখ্যা হিসেবে বলা যেতে পারে। ধরুন, একটা বাসার সদস্যসংখ্যা আমরা যখন হিসেব করি তখন সেখানে একটা বাচ্চা অথবা তিনটা বাচ্চা হিসেবে বলতে পারি, কিন্তু কখনোই আমরা বলতে পারি না যে ৩.৪৯ বাচ্চা। এদিকে, আমরা কোনো একটা জিনিসের দাম হিসেব করলে সেখানে এ ধরনের ভগ্নাংশ অবশ্যই যোগ করা যায়। আমার কথা হচ্ছে, আমরা যখন ডিসক্রিট ভ্যারিয়েবল নিয়ে কথা বলব, তখন কোনো ভগ্নাংশ ব্যবহার করব না, যাতে সেটা ডেটা ঠিকমতো রিপ্রেজেন্ট করতে পারে। কয়টা পাখি সংখ্যায় বললে সেটা ৩/৫/৭ হতে পারে, তবে বাচ্চাসহ ৩.৫ নয়।

আবার ধরা যাক, আপনাকে জিজ্ঞাসা করা হলো আপনার কাছে কয়টা কলম আছে? তার উত্তরে অবশ্যই আপনি একটা ডিসক্রিট ভ্যারিয়েবল সংখ্যা বলবেন, যা হয়তোবা ৫ অথবা ৯ হতে পারে। ৫.৬ নয়।

কন্টিনিউয়াস ভ্যারিয়েবল: সংখ্যায় ব্যবহৃত এ ধরনের ভ্যারিয়েবলগুলো ভগ্নাংশ হয় না, অর্থাৎ একটা স্টেপ থেকে আরেকটা স্টেপে যেতে সেটার মধ্যে কোনো 'ব্রেক' থাকে না। ব্যাপারটা এ রকম যে– যখন আমরা কোনো কিছু মাপতে যাই, তখন দুটো সংখ্যার মধ্যে 'ইনফিনিটি'

লেভেলের স্টেপ থাকতে পারে। যেমন, আপনি যখন সময় নিয়ে আলাপ করবেন, তখন সংখ্যার একেকটা ঘণ্টার মধ্যে, মিনিট এবং সেকেন্ডের দুই লেভেলের ধাপ আছে। এখন আপনি এক ঘণ্টা ত্রিশ মিনিট না বলে সেটাকে ১.৫ ঘণ্টাও বলা যেতে পারে। ঘণ্টা ব্যাপারটা ডিসক্রিট ভ্যারিয়েবল হলেও আমরা যখন সেই ঘণ্টাকে একটা ভগ্নাংশের মধ্যে অর্থাৎ সেই ঘণ্টার একটা কন্টিনিউয়াস মোড হিসেব করতে যাব, তখন সেটাকে কন্টিনিউয়াস ভ্যারিয়েবল বলা যেতে পারে। যেটা ৩ জন মানুষ এবং ৪ জন মানুষের মধ্যে আলাদা স্টেপ করে সম্ভব না। কারণ, আমরা মানুষকে ভাঙতে পারব না।

এই একই গল্প প্রযোজ্য যখন আমরা তাপমাত্রা, দূরত্ব বা ওজন নিয়ে কথা বলব। দূরত্ব এক মাইল অথবা ১.৫ মাইল হতে পারে। এদিকে আমরা যখন ওজন নিয়ে কথা বলব, তখন যেকোনো জিনিসের ওজন ২ কেজি অথবা ২.৭৫ কেজি হতে পারে। এটাকেই কন্টিনিউয়াস ভ্যারিয়েবল হিসেবে আমরা দেখছি।

সংখ্যার আরেকটা ডাইমেনশন: এগ্রিগেশন

ডেটার এগ্রিগেশনের একটা ফর্ম হচ্ছে কীভাবে অনেকগুলো ডেটাকে এক জায়গায় এনে একটা বিশেষায়িত অ্যাঙ্গেল থেকে ঠিকমতো বোঝা যায়। ধরুন, আপনি অনেকগুলো মানুষের স্যালারি নিয়ে কাজ করতে গিয়ে তাদের গড় স্যালারি একটা শহরের সব মানুষের আয়ের ব্যাপারে একটা ভালো ধারণা দেয়। এটাই এগ্রিগেশন। একটা সংখ্যা দিয়ে একটা বড় ডেটাসেটের ধারণা পাওয়া।

একটা ডেটার ভেতরে যদি ট্রেন্ড অ্যানালাইসিস করতে হয় তাহলে এগ্রিগেশন বিশাল কাজ করে। কোয়ান্টিটেটিভ ফিল্ডগুলোকে চমৎকারভাবে এগ্রিগেট করা যায়। ট্যাবলো এবং মাইক্রোসফট পাওয়ার বিআইতে ডিফল্ট হিসেবে যেকোনো ডেটাকে শুরুতে ‘এগ্রিগেট’ করে। কারণ এতে শুরুতেই বড় একটা ডেটার ট্রেন্ড বোঝা যায়।

নিচের টেবিল থেকে আমরা কিছু ধারণা নেই:

এগ্রিগেশন ধারণা

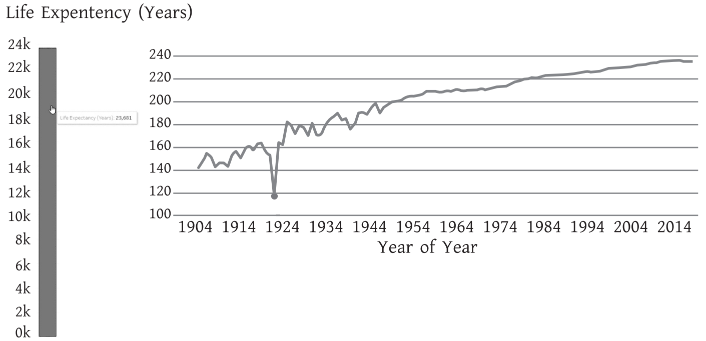
এগ্রিগেশন পদ্ধতি	বর্ণনা
সাম	এতে ডেটার সবগুলো ডেটা পয়েন্টের যোগফল বেরিয়ে আসে
এভারেজ	সব ডেটা পয়েন্টের গড়
মিডিয়ান	ডেটার ভেতরে বিভিন্ন 'স্কিউনেসের' ওপর ভিত্তি করে ডেটা কোন দিকে বেশি টানছে, সেটার একটা মধ্যম ভ্যালু জানিয়ে দেয় সে
মিনিমাম	আমাদের ডেটা পয়েন্টের সবচেয়ে কম ভ্যালুটা দেখায়
ম্যাক্সিমাম	ডেটা পয়েন্টের সর্বোচ্চ ভ্যালু
কাউন্ট	ডেটার কতগুলো রেকর্ড, অর্থাৎ সারি সংখ্যা

টেবিল ৭: এগ্রিগেশন পদ্ধতি

ডেটার গ্র্যানুলারিটি

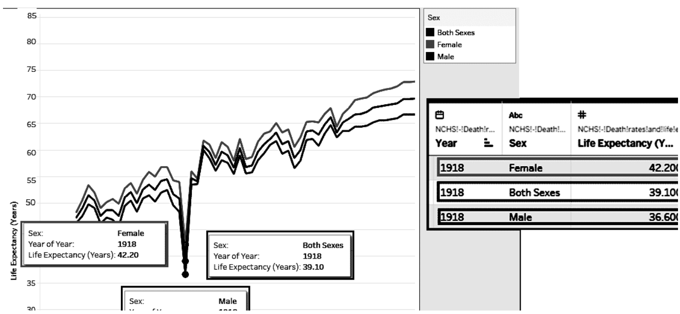
ধরা যাক, আমাদের কোম্পানিতে প্রতিবছর প্রফিট মার্জিন নিয়ে আলাপ হয়। এখানে গত বছর এবং তার আগের বছরে কোম্পানি প্রফিট করেছে তার ধারণা আলাদা আলাদাভাবে বছরভিত্তিক পেলেও এর ভেতরে কোন মাসে বা ১২ মাসের মধ্যে কোন কোন মাসে প্রফিট ভালো থাকে সেটার ভেতরে গেলে এর বছরভিত্তিক প্রফিট লেভেল থেকে মাসভিত্তিক প্রফিট লেভেলে একটা ভালো গ্র্যানুলারিটি পাওয়া যায়। এখানে ডেটাটা আর এগ্রিগেটেড থাকছে না। ফলে, ভিজ্যুয়ালাইজেশনে ডেটাকে আরও ডিটেইল লেভেলে দেখাবে।

এখানে একটা উদাহরণ নিয়ে আলাপ করি। মনে আছে, লাইফ এক্সপেক্টান্সি ডেটাসেটের কথা? বিশ্বব্যাংকের এই ডেটাসেট নিয়ে অনেক কাজ হয়েছে। এর পাশাপাশি বিবিসিতে একটা লাইভ ভিডিও দেখতে পারেন। তবে এই ডেটাসেট সম্বন্ধে না জানলেও নিচের ছবিটা দেখলে পরিষ্কার হয়ে যাবে। এই ছবিতে আপনারা দেখছেন মানুষের লাইফ এক্সপেক্টান্সি ফিল্ডটা একটা 'এগ্রিগেটেড সাম'। এখানে তথ্যগুলো সেভাবে ডিটেইল করা নেই, ফলে এর ভেতরের তথ্যগুলো সম্বন্ধে সেভাবে জানা যাচ্ছে না।



চিত্র ৪৮: প্লটিং সে রকম ডিটেইল নয়

সেই একই ছবি যখন আমরা কয়েক বছর ধরে প্লট করতে চাচ্ছি, তখন এর গ্র্যানুলারিটি বেড়ে গেছে। এর অর্থ হচ্ছে আমরা যত বেশি ডেটাকে বিভিন্ন মাত্রায় অর্থাৎ ডাইমেনশনে ফেলব, তখন সেই ডেটার গ্র্যানুলারিটি সম্বন্ধে আরও ভালো ধারণা আসবে। এর পাশাপাশি লিঙ্গ অর্থাৎ জেন্ডারকেও যদি আমরা এখানে প্লট করতে চাই, তাহলে এখানে নতুন মাত্রা যোগ হবে।



চিত্র ৪৯: ডেটাকে বিভিন্ন ডাইমেনশনে ফেলে গ্র্যানুলারিটি বোঝা

ডিসক্রিট ভ্যারিয়েবলের ডিস্ট্রিবিউশন

ডিসক্রিট ভ্যারিয়েবলের ডিস্ট্রিবিউশন কেমন হতে পারে? তার আগে একটু ব্যাকগ্রাউন্ড নিয়ে কথা বলি।

বই পড়া একটা নেশা। ডেটা অ্যানালাইটিক্স বলি অথবা পরিসংখ্যান বলি— সবকিছুই মার্জ হয়ে আছে একদম বেসিক লেভেলে। সামনের দুটো চ্যাপ্টার আমি নিয়েছি বিশ্বখ্যাত বই 'ইন্ট্রোডাকশন টু স্ট্যাটিসটিকস' থেকে। পরিসংখ্যানের মতো খটমটে জিনিসটাকে তারা চমৎকারভাবে বুঝিয়ে লিখেছেন। পড়া যায় অনলাইনে। বইটা একটা অবশ্যপাঠ্য। ধারণাগুলো তাদের কাছ থেকে নেওয়া। এতে আপনার সুবিধা হবে পরবর্তীতে এই বইটা নিয়ে কাজ করতে গিয়ে।

মনে আছে এমঅ্যান্ডএম (m&m) ক্যান্ডির কথা? আমার বাচ্চারা খুব পছন্দ করে। সেই এমঅ্যান্ডএমের একটা ব্যাগ কিনেছিলেন আমাদের লেখক। একটা প্যাকেটের ভেতর ৬ রঙের ক্যান্ডি পেয়েছিলেন তিনি। পুরো একটা ব্যাগ খুলে একটা গোনা দিতেই দেখা গেল যে, সেখানে মোট ৫৫টা ক্যান্ডি রয়েছে। তার মধ্যে ১৭টা হচ্ছে বাদামি রঙের, ১৮টা

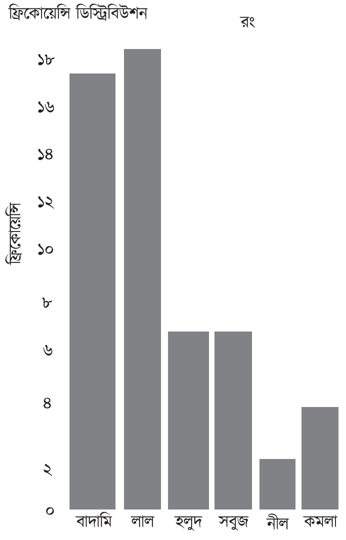
লাল, ৭টা হলুদ এবং সবুজ এবং ৪টা কমলা রঙের এবং শেষে দুটো নীল রঙের। এসব ডেটাকে একটা টেবিলে ফেলি।

রং	ফ্রিকোয়েন্সি
বাদামি	১৭
লাল	১৮
হলুদ	৭
সবুজ	৭
নীল	২
কমলা	৪

টেবিল ৮: এমঅ্যান্ডএম ব্যাগের ক্যান্ডির পরিমাণের ফ্রিকোয়েন্সি

ফ্রিকোয়েন্সি টেবিল

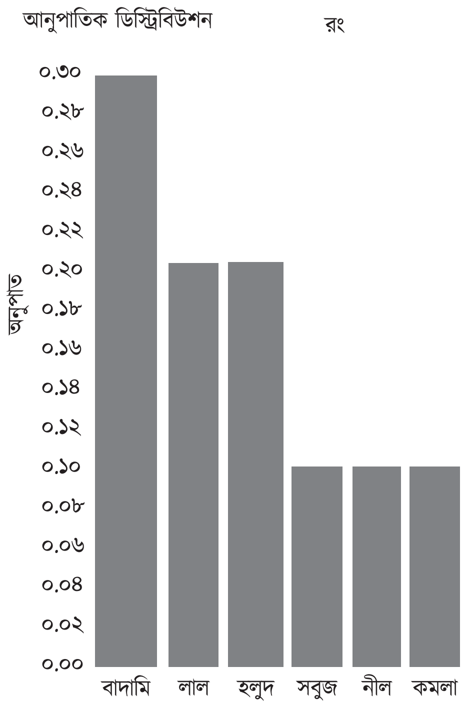
এই টেবিলটাকে আমরা ফ্রিকোয়েন্সি টেবিল বলতে পারি, যেখানে এমঅ্যান্ডএম ক্যান্ডিগুলোর রঙের ফ্রিকোয়েন্সি নিয়ে আলাপ করা হয়েছে। মজার কথা হচ্ছে, এ ধরনের ডিস্ট্রিবিউশনকে ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন বলা হয়। এই টেবিলটাকে আমরা যদি গ্রাফে ফেলি তাহলে নিচের মতো হতে পারে।



চিত্র ৫০: রঙের ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন

প্রবাবিলিটি ডিস্ট্রিবিউশন গ্রাফ

আমরা সাধারণত যখন এ ধরনের ছোট ছোট ক্যান্ডির প্যাকেট কিনি, সেখানে কোম্পানিগুলো কালার ধরে ধরে সমসংখ্যক ক্যান্ডি ফেলে না। কারণ এতে তাদের অপারেশন খরচ বেড়ে যাবে। বরং তারা সবগুলো রংকে একসাথে মিশিয়ে একবারই প্যাকেজিং করে যাতে খরচ কমে যায়। সে কারণে আমরা যখন ক্যান্ডিগুলোর ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন দেখছি আগের ছবিতে, সেটা পরের ছবিতে পাল্টে যাচ্ছে। পরের ছবিতে আমরা যদি যেকোনো একটা রং ধরে তার প্রবাবিলিটি ডিস্ট্রিবিউশন বের করতে চাই, সেখানে আমরা একমাত্র একটা রংকে পিক করব দৈবচয়নের ভিত্তিতে— তাহলে বাদামি (ধরে নিচ্ছি) রঙের ক্যান্ডি পাবার অনুপাত হচ্ছে ০.৩০। সে কারণে আমরা নতুন একটা ডিস্ট্রিবিউশন গ্রাফ ফেলছি সেই অনুপাতের ওপর ভিত্তি করে।



চিত্র ৫১: প্রবাবিলিটি ডিস্ট্রিবিউশন

৫৫টা ক্যান্ডির প্রবাবিলিটি ডিস্ট্রিবিউশন

আগের ছবিতে আমরা দেখেছি ওই ব্যাগে ৫৫টা ক্যান্ডির ডিস্ট্রিবিউশন। পরের ছবিতে আমরা পুরোপুরি এমঅ্যান্ডএম কোম্পানির তৈরি ৬ ধরনের রঙের প্রপোরশন টেবিল, অর্থাৎ অনুপাত বের করার চেষ্টা করেছি। এই একই জিনিস প্রযোজ্য হবে যখন এমঅ্যান্ডএম কোম্পানি দৈবচয়নের ভিত্তিতে আলাদা আলাদা ব্যাগ তৈরি করবে। কোন কোন ব্যাগের মধ্যে এই গ্রাফের মতো রঙের ডিস্ট্রিবিউশন হবে। আবার কোনো কোনো ব্যাগে সেটা আগের ডিস্ট্রিবিউশন থেকে কিছুটা ব্যতিক্রম হবে। তবে যেটাই হোক না কেন সবগুলোর যোগফল ১ হবে। এটার ব্যাপারে বিস্তারিত জানতে ‘হাতেকলমে পাইথন ডিপ লার্নিং’-এর ১১৭-১১৯ পৃষ্ঠা দেখতে পারেন।

কন্টিনিউয়াস ভ্যারিয়েবলের ডিস্ট্রিবিউশন

কন্টিনিউয়াস ভ্যারিয়েবলের ডিস্ট্রিবিউশনে এটা কীভাবে কাজ করবে?

এর আগের উদাহরণে আমরা ডিসক্রিট ভ্যারিয়েবল নিয়ে আলাপ করলেও এখন চলে আসছি একই ধারণাতে— যেখানে কন্টিনিউয়াস ভ্যারিয়েবল কীভাবে একটা ডিস্ট্রিবিউশন দেখাবে। আমরা নিচের টেবিলে একটা উদাহরণ টানছি, যেখানে ২০টি ডেটা পয়েন্ট আছে। এখানে একেকটা রেসপন্স সময় রেকর্ড করা আছে, যখন একটা কার্সারকে নড়ানো হচ্ছে একটা বিন্দুর কাছাকাছি। এখানে যেহেতু ভ্যারিয়েবল অর্থাৎ রেসপন্স সময় হচ্ছে কন্টিনিউয়াস, যার ইউনিট হচ্ছে মিলি সেকেন্ড। সে কারণে দুটো রেসপন্স সময় কখনোই এক হবে না।

কার্সার নাড়ানোর ট্রায়াল	রেসপন্স সময় (মিলি সেকেন্ড)	কার্সার নাড়ানোর ট্রায়াল	রেসপন্স সময় (মিলি সেকেন্ড)
১	৫৬৮	১১	১
২	৫৭৭	১২	৭২৮
৩	৫৮১	১৩	৭২৯
৪	৬৪০	১৪	৭৭৭
৫	৬৪১	১৫	৮০৮

কার্সার নাড়ানোর ট্রায়াল	রেসপন্স সময় (মিলি সেকেন্ড)	কার্সার নাড়ানোর ট্রায়াল	রেসপন্স সময় (মিলি সেকেন্ড)
৬	৬৪৫	১৬	৮২৪
৭	৬৫৭	১৭	৮২৫
৮	৬৭৩	১৮	৮৬৫
৯	৬৯৬	১৯	৮৭৫
১০	৭০৩	২০	১০০৭

টেবিল ৯: কার্সার নাড়ানোর ট্রায়ালের সময় (মিলি সেকেন্ডে)

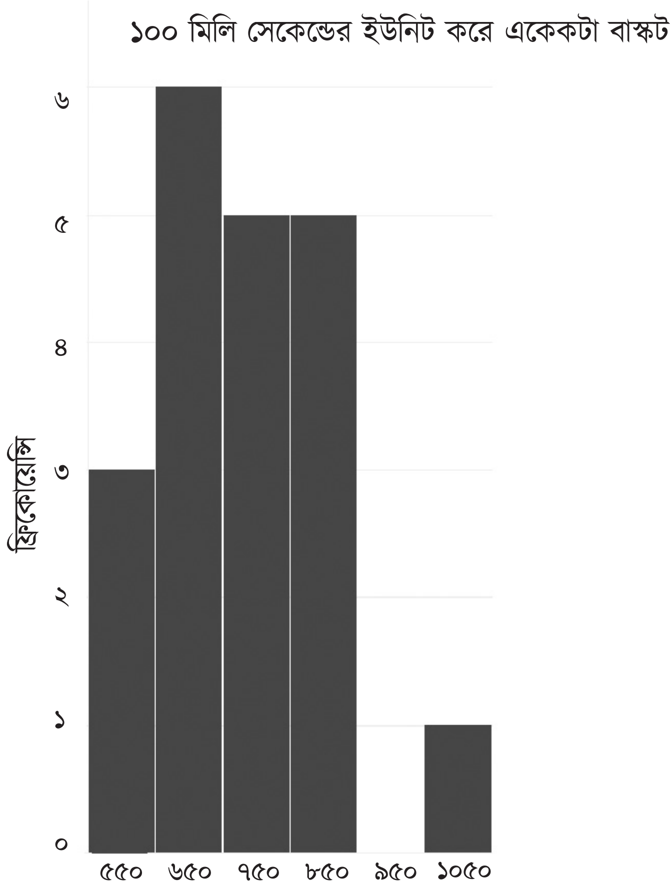
ভালো করে লক্ষ করুন, এখানে আগের মতো ফ্রিকোয়েন্সি ডিস্ট্রিবিউশন ব্যবহার করলে আমাদের কোনো কাজে লাগবে না। একেকটা ভ্যালুর ফ্রিকোয়েন্সি এখানে ১। তাহলে উপায় কী? এই সমস্যার সমাধানে আমাদের এই মিলি সেকেন্ডের রেসপন্স টাইমগুলোকে একটা রেঞ্জের বাস্কেটের মধ্যে ফেলতে হবে, যাতে তার ফ্রিকোয়েন্সি ধরা যায়। এটাকে আমরা অনেক সময় 'বিন' বলি। যেমন, ১০০ মিলি সেকেন্ডের একটা বিন। একেকটা বিনের মধ্যে একেকটা রেঞ্জ থাকে যার মধ্যে মিলি সেকেন্ডগুলোকে আমরা ধরে রাখতে পারি।

বিন (মিলি সেকেন্ড)	ফ্রিকোয়েন্সি
৫০০-৬০০	৩
৬০০-৭০০	৬
৭০০-৮০০	৫
৮০০-৯০০	৫
৯০০-১০০০	০
১০০০-১১০০	১

টেবিল ১০: মিলি সেকেন্ডের বিন

এই পুরো জিনিসটাকে আমরা গ্রাফিক্যালি দেখতে পারি। এই দেখার পদ্ধতির একটা নাম আছে, যেটাকে সাধারণত হিস্টোগ্রাম বলা হয়। 'এক্স' অ্যাক্সিসে বাস্কেট অথবা বিন ব্যবহার করা হয়েছে, সেটার মধ্যে ভ্যালু থেকে পুরো ব্যাপারটাকে ধরা হয়। আপনারা নিচের হিস্টোগ্রাম দেখতে পারেন।

গ্রাফ: হিস্টোগ্রাম



চিত্র ৫২: ফ্রিকোয়েন্সির হিস্টোগ্রাম

ডিস্ট্রিবিউশনের বিভিন্ন শেপ এবং হিস্টোগ্রাম

ডিস্ট্রিবিউশনের বিভিন্ন শেপ

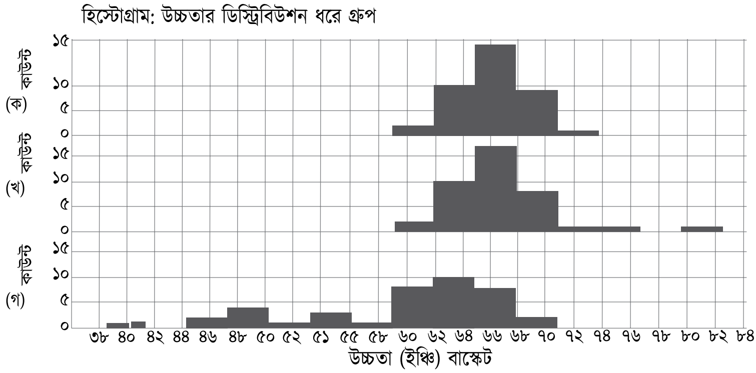
আমরা যখন ডেটাকে প্লট করি, তখন সেই ডেটার পয়েন্টগুলোর ডিস্ট্রিবিউশনগুলো বিভিন্ন আকারে আসতে পারে। কিছু কিছু ডিস্ট্রিবিউশন অনেকটাই সিমেট্রিক্যাল যেখানে তার ভ্যালুগুলো পুরো জায়গা ধরে সমানভাবে ডিস্ট্রিবিউটেড থাকে তার মধ্যকে কেন্দ্র করে। এদিকে আবার অনেক সময় এই ডিস্ট্রিবিউশনের মধ্যে একটা 'পজিটিভ স্কিউ', অর্থাৎ একদিকে ডেটার ক্লাস্টার ঘেঁষা থাকে যেটাকে পজিটিভের ক্ষেত্রে 'ডান দিকে' আমরা ধরে নিই। সেদিক থেকে যখন এই ডিস্ট্রিবিউশনের ডেটাগুলো ক্লাস্টার থেকে কিছুটা বামে থাকে, তখন সেটাকে আমরা 'নেগেটিভ স্কিউ' বলি। ঠিকই ধরেছেন। ডেটারও ডান বা বাম ঘেঁষা বৈশিষ্ট্য হয়।

নিচে একটা উদাহরণ নিয়ে আলাপ করি। ধরে নিই, আমাদের কাছে তিনটা আলাদা আলাদা গ্রুপের মানুষের উচ্চতার হিসেব আছে যার মাধ্যমে নিচের একটা হিস্টোগ্রাম তৈরি করেছি। ৩ গ্রুপের জন্য আলাদা আলাদা হিস্টোগ্রাম আমরা দেখতে পাচ্ছি নিচে।

হিস্টোগ্রাম

এই উদাহরণটা বোঝার জন্য চিত্র: ৫৩ ভালোভাবে লক্ষ করুন। তিনটি গ্রুপের স্টুডেন্টদের নিয়ে এই উচ্চতার আলাদা আলাদা ভাগ। চিত্র: ৫৩তে তিনটি গ্রুপকে একসাথে প্লট করা হয়েছে ডেটার গ্রুপকে একসাথে প্লট করা হয়েছে ডেটার 'ডিসপারশন' বুঝতে।

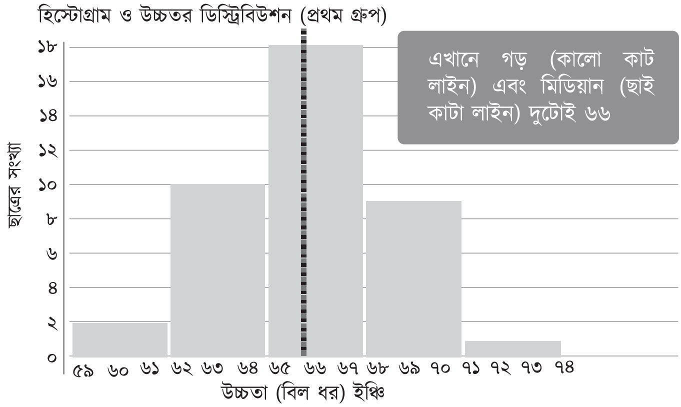
আমরা দেখতে পাচ্ছি এখানকার ডিস্ট্রিবিউশনের বাস্কেট, অর্থাৎ বিন সাইজ হচ্ছে ২.৯৫ ইঞ্চি। ফলে এখানে তিনটি গ্রুপের স্টুডেন্টদের উচ্চতা ভ্যালুগুলোকে 'বিনিং' অর্থাৎ বাক্সবন্দী করা হয়েছে ৫৯-৬১.৯৫, ৬২-৬৪.৯৫ ... এভাবে চলে আসবে সবার জন্য। চলুন, আমরা এই হিস্টোগ্রামকে একটু কাছ থেকে বুঝে আসি।



চিত্র ৫৩: উচ্চতার ডিস্ট্রিবিউশন

সিমেট্রিক্যাল ডিস্ট্রিবিউশন

আমাদের এই মানুষের উচ্চতার গ্রুপে যদি ডেটাগুলো দেখি, তাহলে একটা গ্রুপ অনেকটাই সিমেট্রিক্যাল, কারণ আমরা যখন এই পুরো হিস্টোগ্রামকে দুই পাশে কাল্পনিকভাবে ভাঁজ করব, তখন এটা দুটো অংশ পাশাপাশি সমান জায়গা নিয়ে 'ম্যাচ' করবে সুন্দরভাবে। এ ধরনের সিমেট্রিক্যাল ডিস্ট্রিবিউশনে ডেটার ক্লাস্টারে ঠিক মাঝখানটাতে এর 'মিন' অর্থাৎ গড়টা থাকে। এর পাশাপাশি তার 'মিডিয়ান' অংশ— যেখানে সম্পূর্ণ ভ্যালুগুলোর মধ্যের আসল মধ্যম পয়েন্ট, কাছাকাছি থাকে কারণ এই দুটো ভ্যালুর 'আউটকাম' প্রায় সমান সমান। সে কারণে এই ডেটার, বিশেষ করে দুই ধরনের ভ্যালুর মধ্যম জায়গা এবং তার 'ডিসপারশন' অর্থাৎ ডেটাটা কীভাবে ছড়িয়ে আছে দুই পাশে, সেটাও প্রায় সমান সমান হয়। নিচের ছবিটা দেখুন।



চিত্র ৫৪: সিমেট্রিক্যাল ডিস্ট্রিবিউশনের ছবি

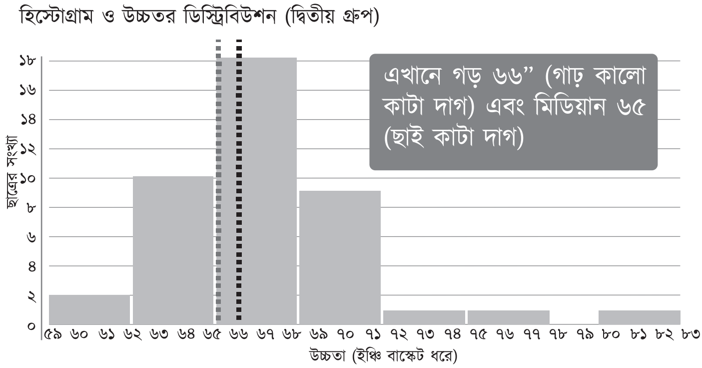
‘পজিটিভ স্কিউ’ ডিস্ট্রিবিউশন

পৃথিবীর বেশির ভাগ ডেটায় ঠিকমতো সিমেট্রি তৈরি হয় না বলে ডেটার ঠিকমতো ডিস্ট্রিবিউশন নিয়ে কাজ করতে গেলে এর ভেতরে দু ধরনের অর্থাৎ ডান এবং বাম ঘেঁষা ডেটা পাওয়া যায়। নিচের ছবি দেখুন। ডান দিকে ঘেঁষা, বিশেষ করে যদি ডেটার ডিস্ট্রিবিউশন পজিটিভ দিকে যায় তাহলে সেটাকে ‘পজিটিভ স্কিউ’ বলি আমরা। এটা পজিটিভ বলছি এ কারণে যেখানে ডেটাকে কিছুটা টেনে লম্বা দেখানো হচ্ছে ডান দিকে। ডান দিকে এর ‘ডিসপারশন’ বেশি। যদি ডেটার ‘ডিস্ট্রিবিউশন স্কিউ’ পজিটিভ দিকে হয়, তখন মিডিয়ান গড় অথবা ‘মিন’ থেকে ছোট হয়।

তবে এখানেও কিছু সমস্যা আছে। একটা শহরের সবার গড় আয় মাঝামাঝি থাকলেও সেখানে যদি কয়েকজন কোটিপতি থাকে তাহলে সেই কোটিপতিদের কারণে পুরো শহরের গড় ইনকাম একদিকে মানে কিছুটা ওপরে চলে যাবে। এ জন্যই এ ধরনের কাজে কোটিপতিদের হিসেব একদিকে টানলেও পুরো শহরের জনগণের আয়ের আসল হিসেব চাইলে গড়ের পরিবর্তে ‘মিডিয়ান’ আয় ভালো কাজ করবে। কোটিপতিরা সেখানে কিছুটা ‘আউটলাইয়ার।’

আমরা যখন আমাদের দ্বিতীয় গ্রুপের উচ্চতার ডেটা দেখছি, সেখানে তিনজন ব্যক্তির কারণে বাকি সবার জন্য ডেটাটা পজিটিভ স্কিউ, অর্থাৎ ডান

দিকে ঘেঁষে গেছে। তিনজনের উচ্চতা আমরা এখানে দেখতে পাচ্ছি প্রায় ৬ ফুট অথবা তার কাছাকাছি। ছবিগুলো ভালোভাবে দেখুন এই তিনজনের উচ্চতা বাকি সব ডেটার গড়কে আরও উপরে নিয়ে গেছে। এ জন্য এখানে বিশেষ করে, এই গ্রুপের জন্য মিডিয়ান ভ্যালু নেওয়াটাই সবার জন্য সঠিক রিপ্রেজেন্টেশন হবে।



চিত্র ৫৫: ডেটার ‘পজেটিভ স্কিউ’

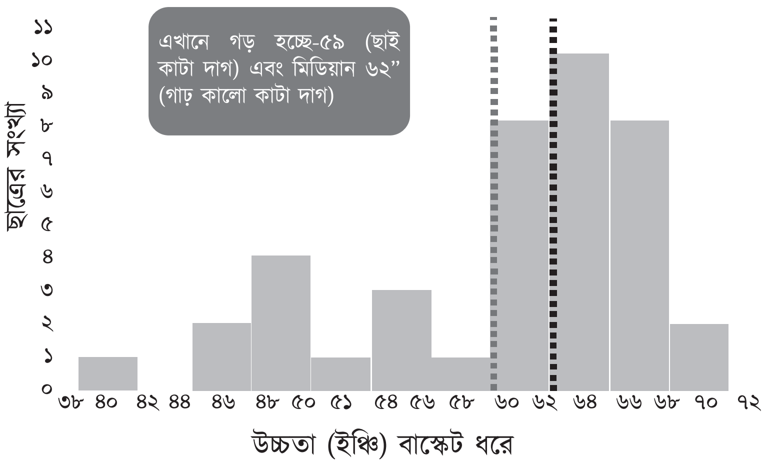
‘নেগেটিভ স্কিউ’ ডিস্ট্রিবিউশন

অ্যাসিমেট্রিক্যাল ডিস্ট্রিবিউশনের আরেকটা ভাগ হচ্ছে ‘নেগেটিভ স্কিউ’ ডিস্ট্রিবিউশন— যেখানে ডেটার ক্লাস্টারে ডেটার একটা বড় অংশ নেগেটিভ দিকে চলে গেছে। এই ধরনের ডেটা ‘ডিসপারশন’ যখন বাম দিকে বেশি চলে যায়, তখন এটাকে ‘নেগেটিভ স্কিউ’ বলা হয়। অর্থাৎ এর বাম দিকের ডেটার অংশটা বেশি এবং লম্বা। যখন আমাদের ডিস্ট্রিবিউশন নেগেটিভ দিকে ‘স্কিউ’ হয়, তখন ডেটার মিডিয়ান ডেটার গড় থেকে বড় হয়।

একটা উদাহরণ দেয়া যাক। ২০ জনের একটা ক্লাসে যখন দুজন ছাত্র কখনোই যদি ক্লাসে না আসে অথবা অ্যাসাইনমেন্ট জমা না দেয়, তাহলে কী হতে পারে? তাদের ফাইনাল গ্রেড অবশ্যই শূন্য হবে। আর এই দুজন ছাত্রের কারণে পুরো ২০ জনের ছাত্রের ক্লাসে সবার ফাইনাল গ্রেডের গড় আসল গড় থেকে বেশ কিছুটা নিচে নেমে যাবে। আর সে কারণে ২০ জন ছাত্রের ফাইনাল গ্রেডের সত্যিকারের রিপ্রেজেন্টেশন চাইলে এখানে গ্রেডের ‘মিডিয়ান’ ভ্যালু ভালো ‘আউটকাম’ দেবে।

আমাদের আগের ডেটা অনুযায়ী যখন তিনটা গ্রুপের ছাত্রদের উচ্চতার ডেটা হিসেব করছি, সেখানে তৃতীয় গ্রুপটা কিছুটা নেগেটিভ স্কিউ, কারণ এখানে একজন ব্যক্তি আছেন যার উচ্চতা ৫ ফুটের কম। সে কারণে তাদের আসল গড় উচ্চতা একটু কমে গেছে, সে কারণে এর 'ডিসপারশন' একদিকে অর্থাৎ বাম দিকে বেশি ঘেঁষার কারণে হয়েছে। ছবি দেখুন।

হিস্টোগ্রাম: উচ্চতর ডিস্ট্রিবিউশন (তৃতীয় গ্রুপ)



চিত্র ৫৬: 'নেগেটিভ স্কিউ'র মিডিয়ান

দুটো জটিল উদাহরণ: বক্স প্লট এবং হিস্টোগ্রাম

শুরুর উদাহরণ: বক্স প্লট

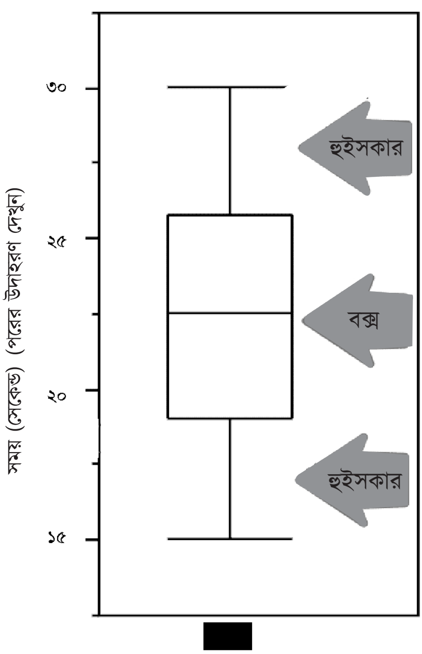
(কিছু তাত্ত্বিক জিনিস বইয়ের লেখায় ভালো, অন্যদিকে ছবিগুলোকে আমরা যখন প্লট করব— তখন ভিডিও প্লেলিস্ট ভালো। আর সে কারণেই শুধু বক্স প্লট নিয়ে কাগজ-কলমে বোঝানোর চেষ্টা করছি। বক্স প্লেট বুঝলে অন্য ভিজ্যুয়ালাইজেশন বোঝা সহজ হবে। কারণ, বক্স প্লট একটা কঠিন প্লট বটে!

কন্টিনিউয়াস ভ্যারিয়েবলের ডিস্ট্রিবিউশন: বক্স প্লট

এই অংশে আমরা চেষ্টা করব কীভাবে কিছু ভ্যারিয়েবলের ডিস্ট্রিবিউশন বিভিন্ন অ্যাঙ্গেল থেকে দেখা যায়? আর এ কারণে আমরা এই মুহূর্তে একটা খুব প্রয়োজনীয় গ্রাফ নিয়ে আলোচনা করব, যাকে আমরা সাধারণত 'বক্স প্লট' বলে থাকি। বক্স প্লটে একটা সুবিধা হচ্ছে কীভাবে একটা ডেটাসেটের ডিস্ট্রিবিউশন সহজে এবং কম জায়গায় দেখা যায়। এখানে সবচেয়ে বড় ব্যাপার হচ্ছে, এর ভেতরে ডেটার বিভিন্ন বৈশিষ্ট্য খুব সহজে বোঝা যায়। শুরুতেই একটা বক্স প্লটে যেই বাক্সটা প্লটিংয়ের ভেতরে থাকে সেটা আসলে ৫০% মধ্যম, অর্থাৎ ঠিক মাঝখানে থাকে। এই বক্সটা শুরু হয় ২৫% পার্সেন্টাইল থেকে ৭৫% পার্সেন্টাইল পর্যন্ত। এর ঠিক মাঝখানে মিডিয়ান, অর্থাৎ সেটা ঘিরে থাকে ৫০ পার্সেন্টাইল লাইন ঘিরে। ছবি দেখলেই পরিষ্কার হয়ে যাবে।

'পার্সেন্টাইল'-এর কাজ হচ্ছে একটা ডেটাসেটের ভেতর কীভাবে একটা জিনিসকে আরেকটার সাথে 'কমপেয়ার' করা যায়। এখন যদি আপনার

‘লাজুকতা’ অর্থাৎ ইন্ট্রোভার্টনেস নিয়ে যদি হিসেব করতে চাই তাহলে সেটার স্কোরিং সরাসরি বের করা কিছুটা সমস্যাজনক হবে বটে। তার বদলে আপনার ইন্ট্রোভার্টনেস যদি অন্য আরেকজনের সাথে আপেক্ষিকভাবে বের করা যায় তাহলে হয়তোবা বোঝা যাবে তার ইন্ট্রোভার্টনেসের কাছে আপনি বেশি না কম। এখানে সেটাকে, মানে সেটার রিলেটিভনেসকে শতাংশ আকারে ধরে নিতে পারি। ধরা যাক, একটা কুইজ টেস্ট নেওয়া হলো যেখানে বিভিন্ন প্রশ্নের মাধ্যমে একজন মানুষ কতটুকু লাজুক, সেটা বের করতে হবে। সেখানে যদি ৬৫% বাকি কুইজ টেস্ট অংশগ্রহণকারীর স্কোর আপনার থেকে কম হয়, তাহলে আপনার স্কোর হবে ৬৫তম পার্সেন্টাইল।



চিত্র ৫৭: একটা বক্স প্লট

বক্স প্লটের ভূমির সমান্তরালে যে লাইনগুলো আছে, সেগুলোকে ‘হুইসকার’ বলি আমরা। বক্স প্লটের সাথে এগুলো জোড়া দেওয়া হয় যাতে ডেটার ডিস্ট্রিবিউশন আরও ভালোভাবে বোঝা যায়। সে কারণে হুইসকার, বক্সের

বাইরে প্লট করা হয়। দেখুন, হুইসকারের জিনিসগুলোর ডিস্ট্রিবিউশনের মধ্যে সাধারণত ৫০ শতাংশের মধ্যে থাকে না— বরং বাইরের তথ্যগুলোকে ঠিকমতো বোঝা যায়। ডেটার ডিস্ট্রিবিউশনে কোনটা কতদূর গিয়েছে এবং কোনটার মধ্যে 'আউটলাইয়ার' অর্থাৎ দলছুট ভ্যালু আর আছে, সেটাও এ ধরনের ডিস্ট্রিবিউশনে বোঝা যায়। আমাদের দেখা মতে, এই আউটলাইয়ারগুলো সাধারণত নরমাল রেকর্ডের মতো হয় না; বরং এটা ডিস্ট্রিবিউশনের মধ্যম পয়েন্ট থেকে এক্সট্রিম পর্যায়ে যেতে পারে। মজার কথা হচ্ছে, আজকের এই জনপ্রিয় বক্স প্লট তৈরি করা হয়েছিল ১৯৭০ সালের দিকে। মার্কিন অঙ্কবিদ 'জন টাকি' এটা বানিয়েছিলেন। আমার প্রিয় একটা প্লটিং মেকানিজম। ইন্ডাস্ট্রির সবাই ব্যবহার করেন।

তৈরি করি একটা বক্স প্লট

একটা বই লিখতে গেলে তার ব্যাকগ্রাউন্ডে প্রচুর বই পড়তে হয়। অনেক বেছে ডেভিড এম লেইনের 'বক্স প্লট চ্যাপ্টার' থেকে উদাহরণটা নিয়েছি আমি। ওনার মূল বইয়ের নাম হচ্ছে 'ইন্ট্রোডাকশন টু স্ট্যাটিসটিকস', যা পাবলিক ডোমেইনে পাওয়া যায়। বিনামূল্যে, মানে এই নয় যে— এটা ভালো বই নয়। আমার দেখা মতে, এটা একটা অসাধারণ বই। ওখানে তিনি একটা ক্লাসের ৩১ জন ছাত্রকে নিয়ে একটা এক্সপেরিমেন্ট চালিয়েছিলেন। ছাত্রদের ৩০টি রঙিন আয়তক্ষেত্র দিয়েছিলেন, যেখানে তাদের কাজ ছিল কে কত তাড়াতাড়ি সেই রঙের নামটা বলতে পারে। নিচের টেবিলটা থেকে বোঝা যাচ্ছে তাদের সময় (আমরা সেকেন্ডে হিসেব করেছি) কীভাবে রেকর্ড করা হয়েছিল। সংখ্যাগুলোর অর্থ হচ্ছে কত সেকেন্ডে তারা রংগুলোর নাম বলতে পেরেছিল।

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

টেবিল ১১: সেকেন্ডের হিসেব

এই ডেটাসেট থেকে আমরা চেষ্টা করব একটা বক্স প্লট তৈরি করতে। সেই বক্স প্লট তৈরি করতে এই জিনিসগুলো লাগবে সামনে:

১. ডেটার পার্সেন্টাইল ক্যালকুলেট করা
২. সেই পার্সেন্টাইল ধরে বক্সকে প্লট করা
৩. এর মাঝে স্টেপ সাইজগুলোকে নির্ধারণ করা
৪. ‘হুইসকার’ যোগ করা
৫. আউটলাইয়ার মানে বাইরের দলছুট ভ্যালু প্লটিংয়ে নিয়ে আসা

পার্সেন্টাইল ক্যালকুলেশন

বক্স প্লটের ভেতরের বক্স আসলে ডেটার ২৫তম পার্সেন্টাইল থেকে ৭৫তম পার্সেন্ট পর্যন্ত ডেটাকে রিপ্রেজেন্ট করে। বুঝতে পারছেন মাঝখানের তারা আমাদের ৫০তম পার্সেন্টাইলকে বোঝায়। সবচেয়ে নিচের, বিশেষ করে বক্সের সবচেয়ে নিচের এলাকায় যেটাকে অনেকে লোয়ার হিঞ্জ বলেন, সেটা ২৫তম পার্সেন্টাইল থেকে শুরু হয়। এর পাশাপাশি— বক্সের ওপরের লাইন অর্থাৎ আপার হিঞ্জ হচ্ছে ৭৫তম পার্সেন্টাইল। না বুঝলে সমস্যা নেই, এখন আমরা ছবি দিয়ে উদাহরণে চলে আসছি।

আমাদের আগের ডেটাসেটের সবগুলো সংখ্যাকে যদি পার্সেন্টাইল মোডে ফেলি তাহলে শুরুতে সেই ৩১টা সংখ্যাকে ছোট থেকে বড় হিসেবে সাজাই। প্যারেডের স্টাইলে ছোট থেকে বড় হিসেবে সাজানো।

ছোট থেকে বড় হিসাবে সাজানোর হিসেব

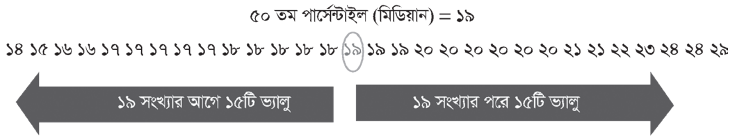
৩১টি সংখ্যাকে ছোট থেকে বড় হিসেবে সাজানো হয়েছে

১৪ ১৫ ১৬ ১৬ ১৭ ১৭ ১৭ ১৭ ১৭ ১৮ ১৮ ১৮ ১৮ ১৮ ১৯ ১৯ ১৯ ২০ ২০ ২০ ২০ ২০ ২০ ২১ ২১ ২২ ২৩ ২৪ ২৪ ২৯

চিত্র ৫৮: ছোট থেকে বড় হিসেবে সাজানো

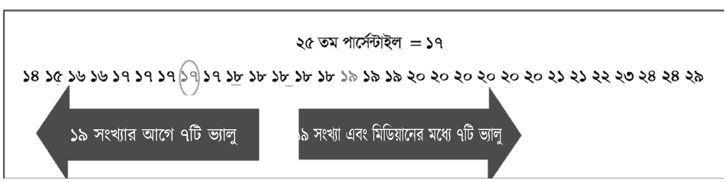
ক. এই ছবি ধরে আমরা তার মিডিয়ান ভ্যালু, অর্থাৎ সেন্ট্রাল সংখ্যাকে খুঁজে বের করি, যা শুরু এবং শেষ সংখ্যার সিকোয়েন্সের ঠিক মাঝামাঝি অংশে থাকবে। যেহেতু ৩১ সংখ্যাটা বিজোড় এ কারণে এর মধ্যম ঘরটা

হচ্ছে শুরুর ১৫ এবং শেষের ১৫-এর ঠিক মাঝখানের সংখ্যা। সে কারণে আমাদের মিডিয়ান ভ্যালু হচ্ছে ১৯।



চিত্র ৫৯: ৫০তম পার্সেন্টাইল

এখন আমাদের কাজ হচ্ছে ২৫তম পার্সেন্টাইল বের করা। এটা বেশ সহজ, কারণ আমাদের শুরু থেকে মিডিয়াম ভ্যালু পর্যন্ত যে সংখ্যা আছে তার মাঝামাঝি ভ্যালুটা হচ্ছে ২৫তম পার্সেন্টাইল। আমাদের এই ৩১টা সংখ্যার মধ্যে মিডিয়ান এবং সংখ্যাগুলোর মধ্যে প্রথম সাতটা সংখ্যার পরে এবং মিডিয়ান ভ্যালু থেকে সাতটা সংখ্যা আগে হচ্ছে এই ২৫তম পার্সেন্টাইল। এর ভ্যালু হচ্ছে ১৭।

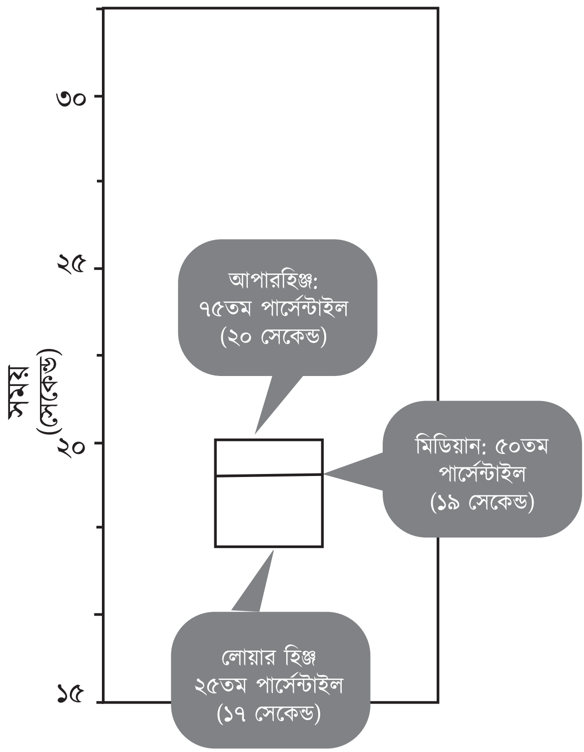


চিত্র ৬০: ২৫তম পার্সেন্টাইল

খ. আগের মতো ৭৫তম পার্সেন্টাইল নির্ধারণ করি। এই ৭৫তম পার্সেন্টাইল বসে আছে মিডিয়ান এবং সবচেয়ে শেষের সংখ্যার ঠিক মাঝখানে। ৩১ সংখ্যার মধ্যে মিডিয়ান ভ্যালু থেকে ৭ সংখ্যা এবং শেষ থেকে ৭ সংখ্যার মাঝে যে সংখ্যা সেটাই হচ্ছে ৭৫তম পার্সেন্টাইল। হিসেব মতে, সেটা আসে সংখ্যা ২০। এটার আর ছবি দিলাম না।

বক্সকে প্লট করি পার্সেন্টাইল হিসেবে

আমাদের বক্সের নিচের যে লাইনটা আছে (লোয়ার হিঞ্জ) সেটাকে আমরা বলি ২৫তম পার্সেন্টাইল এবং ওপরের লাইন অথবা আপার হিঞ্জ হচ্ছে ৭৫তম পার্সেন্টাইল। ৫০তম পার্সেন্টাইল আঁকা হয়েছে এই বক্সের মাঝখানে। এটাই আমাদের মিডিয়ান।



চিত্র ৬১: বক্স প্লট পার্সেন্টাইল হিসেবে

৩১টা স্কোরের মধ্যে আমরা যেটা বুঝতে পেরেছি;

১. ২৫তম পার্সেন্টাইল হচ্ছে ১৭
২. ৫০তম পার্সেন্টাইল অথবা মিডিয়ান হচ্ছে ১৯
৩. ৭৫তম পার্সেন্টেজ হচ্ছে ২০
৪. আপারহিঞ্জ, মিডিয়ান, লোয়ার হিঞ্জ

বক্স প্লট আঁকার হিসেব

এখন বক্সটা আঁকি,

১. ২৫তম পার্সেন্টাইল অর্থাৎ লোয়ার হিঞ্জে আমরা এ লাইন তৈরি করছি y-axis-এর সাথে ১৭কে কানেক্ট করে

২. পঞ্চাশতম পার্সেন্টাইল অর্থাৎ মিডিয়ানকে ধরে এ লাইন তৈরি করছি y-axis-এর সাথে ১৯ নম্বরে

৩. ৭৫তম পার্সেন্টাইল, অর্থাৎ আপার হিঞ্জে লাইন হচ্ছে সংখ্যা ২০-এর সাথে এই একই y-axis-এ।

আমাদের ২৫তম পার্সেন্টাইল এবং মিডিয়ান সংখ্যার মধ্যে যেই ভ্যালুটা আছে সেটাকে অনেকে ‘দ্বিতীয় কোয়ারটাইল’ বলে থাকেন। আবার ৫১তম পার্সেন্টাইল, অর্থাৎ মিডিয়ানের পরের ভ্যালু থেকে শুরু করে ৭৫তম পার্সেন্টাইল পর্যন্ত সংখ্যাগুলোকে ‘তৃতীয় কোয়ারটাইল’ বলা হয়ে থাকে। এবং ৭৫তম পার্সেন্টাইলে ওপরের সব সংখ্যাকে ‘চতুর্থ কোয়ারটাইল’ বলতে পারি আমরা। এখানে ‘হুইসকার’ এবং অন্যান্য লাইনকে ব্যবহার করি প্রথম এবং চতুর্থ কোয়ারটাইল ভ্যালুগুলোকে দেখানোর জন্য।

স্টেপ সাইজ নির্ধারণ করা

আমরা চেষ্টা করব— এখন ‘হুইসকার’ এর ওপর এবং নিচে কীভাবে বাড়তি ইনফরমেশন দেওয়া যায় যাতে ডেটার ডিস্ট্রিবিউশন ঠিকমতো বোঝা যায়। হুইসকারের প্লেসমেন্ট নির্ভর করে তার স্টেপের ওপরে, যেখানে তার স্টেপগুলো আমরা ডিফাইন করতে পারি ১.৫ x IQR (ইন্টার কোয়ারটাইল রেঞ্জ) দিয়ে। এটা কিছুটা কমপ্লিকেটেড মনে হতে পারে, তবে আমাদের এই ‘ইন্টার কোয়ারটাইল রেঞ্জ’ নির্ভর করছে ওপরের যেই আপার হিঞ্জ (আমরা যাকে বলছি ৭৫তম পার্সেন্টাইল) ভ্যালু থেকে নিচের লোয়ার হিঞ্জ (যাকে বলি ২৫তম পার্সেন্টাইল) এর মধ্যে। আমাদের মনে রাখতে হবে যে— মধ্যে ৫০% ডেটা ভ্যালু এই বক্সের মধ্যে থাকছে। আমাদের এই সেটের মধ্যে আপার হিঞ্জ হচ্ছে ২০ এবং নিচের লোয়ার হিঞ্জ হচ্ছে ১৭। এর অর্থ হচ্ছে ‘ইন্টার কোয়ারটাইল রেঞ্জ’ হবে ২০ — ১৭ = ৩।

এখন আমাদের স্টেপ সাইজ নির্ধারণ করতে চাইলে আমাদের ১.৫কে গুণ করতে হবে ৩ দিয়ে। এর ফলাফল হচ্ছে ৪.৫, যা আমাদের স্টেপ সাইজ।

হুইসকার যোগ করা

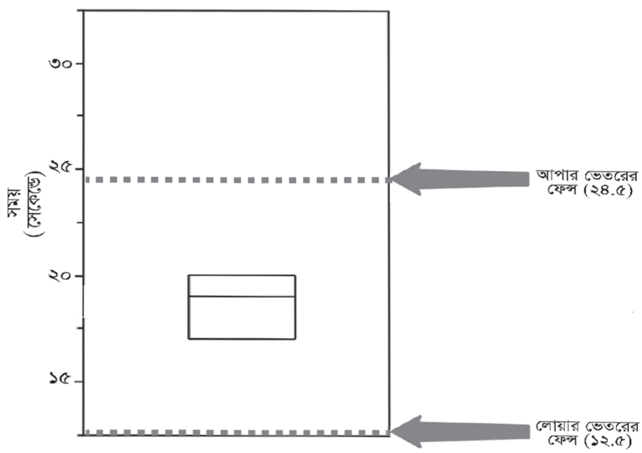
হুইসকার যোগ করার আগে বুঝতে হবে কীভাবে এই জিনিসটাকে প্লট করা যেতে পারে। এর পাশাপাশি, এর কিছু টার্মস এবং তার ভ্যালুগুলোকে কীভাবে যোগ করতে হবে বিভিন্ন পজিশনে। সেটা দেখুন সামনে।

ভেতরের ফেন্স (উপরের এবং নিচের ফেন্স)

ওপরের ভেতরের ফেন্স: আপার হিঞ্জ থেকে এক স্টেপ ওপরে বসবে এটা। আমাদের ডেটা অনুযায়ী এর ভ্যালু হবে ২০ + ৪.৫ = ২৪.৫ যা বসবে 'ওয়াই অ্যাক্সিসের' ওপরে।

নিচের ভেতরের ফেন্স: আগের মতোই লোয়ার হিঞ্জ থেকে একটা স্টেপ নিচে বসবে এই জিনিস। এবং আমাদের ডেটা অনুযায়ী, এর ভ্যালু হবে ১৭ - ৪.৫ = ১২.৫ যা বসবে 'ওয়াই অ্যাক্সিসের' ওপরে।

এখানে উল্লেখ্য, আমাদের ফাইনাল গ্রাফে এই ট্রেন্ডসগুলো দেখা যাবে না, কারণ এগুলো তৈরি করা হয় সেই ধারণার ওপর নির্ভর করে কোথায় 'হুইসকার' বসবে।



চিত্র ৬২: আপার এবং লোয়ার ফেন্স

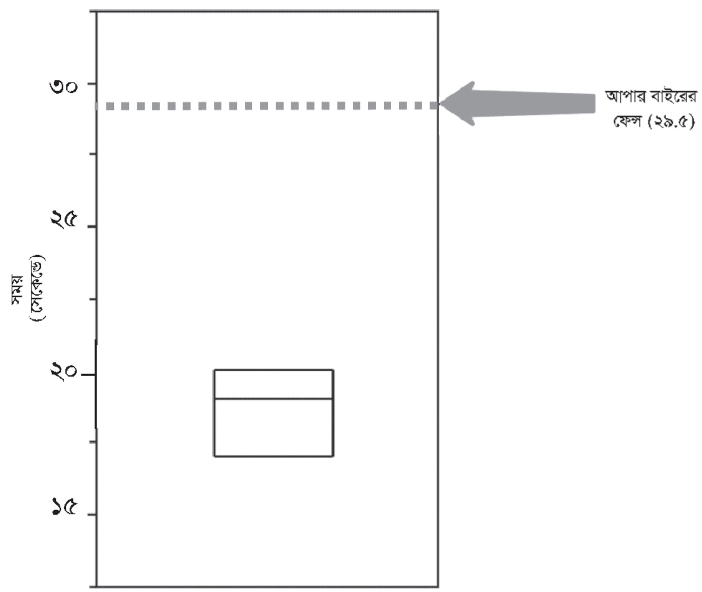
বাইরের ফেন্স (ওপরের এবং নিচের ফেন্স)

উপরের বাইরের ফেন্স: আপার হিঞ্জ থেকে দুই স্টেপ ওপরে বসবে এটা। আমাদের ডেটা অনুযায়ী এর ভ্যালু হবে ২০ + ৪.৫ + ৪.৫ = ২৯, যা বসবে 'ওয়াই অ্যাক্সিসে'র ওপরে।

নিচের বাইরের ফেন্স: আগের মতোই লোয়ার হিঞ্জ থেকে দুই স্টেপ নিচে বসবে এটা। এবং আমাদের ডেটা অনুযায়ী, এর ভ্যালু হবে

১৭ - ৪.৫ - ৪.৫ = ৮, যা বসবে 'ওয়াই অ্যাক্সিসে'র উপরে। (ক্লাটার হবার ভয়ে ছবিতে আর দেখাচ্ছি না এখানে)

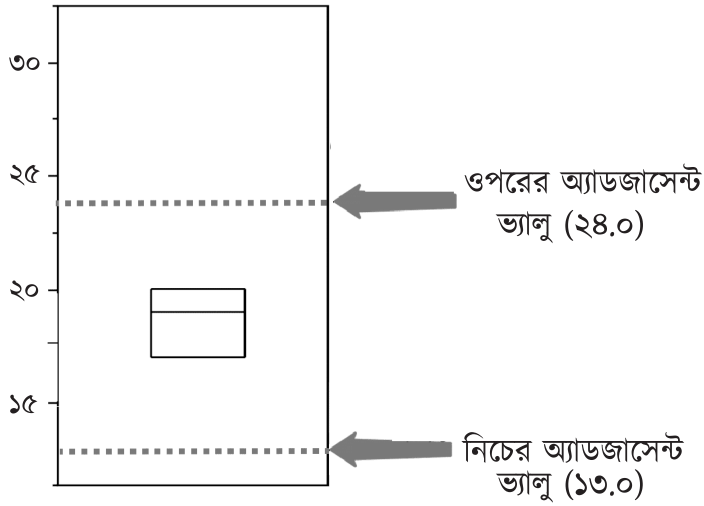
এখানে উল্লেখ্য, আমাদের ফাইনাল গ্রাফে এই ট্রেন্ডসগুলো দেখা যাবে না, কারণ এগুলো তৈরি করা হয় সেই ধারণার ওপর নির্ভর করে কোথায় হুইসকার বসবে।



চিত্র ৬৩: বাইরের ফেন্স (ওপরের)

আশপাশের সংখ্যা (অ্যাডজাসেন্ট ভ্যালু)

ওপরের অ্যাডজাসেন্ট সংখ্যা: মনে আছে ওপরের ভেতরের ফেন্সের কথা? সেই ফেন্সের নিচের বড় সংখ্যাটাই আমাদের ওপরের অ্যাডজাসেন্ট ভ্যালু। আমাদের উদাহরণে যেহেতু এই সংখ্যাটা ২৪, সে কারণে এর ভ্যালু হচ্ছে ২৪ এবং এটা বসবে 'ওয়াই অ্যাক্সিসে'র ওপরে।



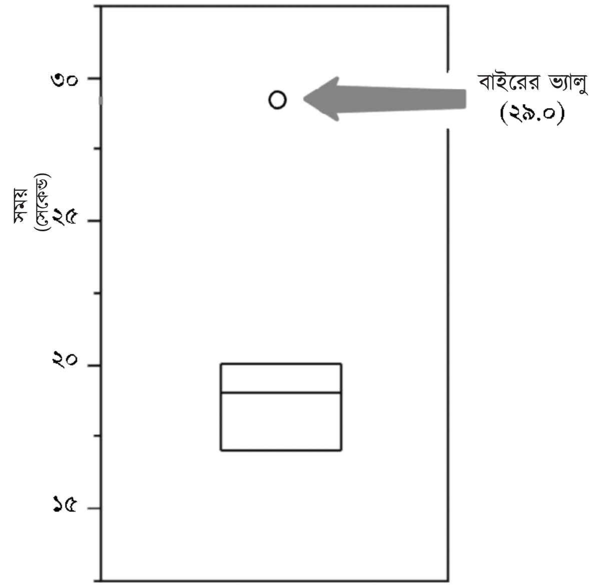
চিত্র ৬৪: অ্যাডজাসেন্ট ভ্যালু

নিচের অ্যাডজাসেন্ট সংখ্যা: নিচের ভেতরের ফেন্সের নিচে সবচেয়ে ছোট সংখ্যাটি হচ্ছে নিচের 'অ্যাডজাসেন্ট' সংখ্যা। আমাদের উদাহরণে এই সংখ্যা ছিল ১৩, তাই এর ভ্যালু হচ্ছে ১৩ এবং সেটা বসবে 'ওয়াই অ্যাক্সিসে'র উপরে।

দলছুট অর্থাৎ আউটলাইয়ার ভ্যালু নিয়ে ধারণা

আউটলাইয়ারের জন্য বাইরের ভ্যালু: এটা এমন ধরনের একটা সংখ্যা হতে পারে, যা ভেতরের ফেন্সের বাইরে থাকতে পারে, তবে সেটা বাইরের ফেন্সের ভেতরে থাকবে। আমাদের ডেটাসেটে সে রকম একটা ভ্যালু আমরা পাচ্ছি যা হচ্ছে সংখ্যা ২৯ এবং এই সংখ্যাটা বসবে 'ওয়াই অ্যাক্সিসে'র ওপরে।

একদম দূরের ভ্যালু: এ রকম একটা ভ্যালু যা বাইরের ফেন্সের একদম বাইরে থাকতে পারে। আমাদের উদাহরণে এ ধরনের ভ্যালু নেই বলে আমরা এখানে সেটা দেখাচ্ছি না।

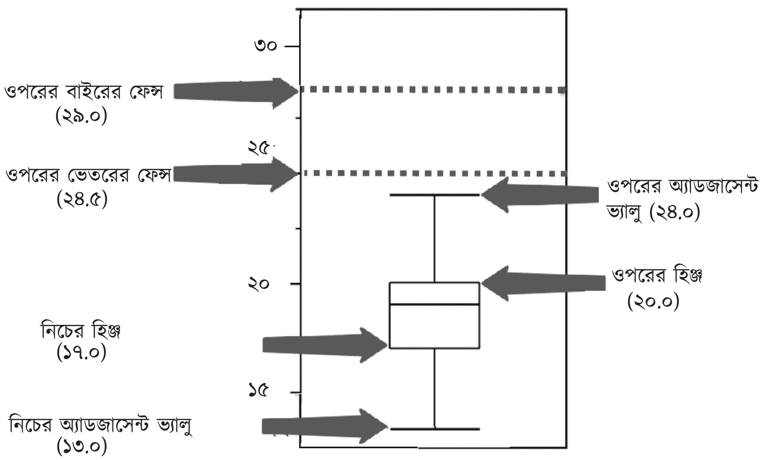


চিত্র ৬৫: আউটলাইয়ারকে প্লট

তাহলে আমাদের 'হুইসকার' কোথায় বসতে পারে?

হুইসকারগুলো সাধারণত আঁকা হয় উপরের হিঞ্জ থেকে ওপরের এডজাসেন্ট সংখ্যা পর্যন্ত এবং নিচের হিঞ্জ থেকে নিচের অ্যাডজাসেন্ট সংখ্যা পর্যন্ত। ছবি ৬৬ দেখুন। একটা কথা মনে রাখবেন, আমরা সাধারণত 'হুইসকার' আঁকি না, যেগুলোর ভ্যালু একটা রেঞ্জের বাইরে পড়ে। তবে আমরা সেগুলোকে বক্স প্লটের মধ্যে রিপ্রেজেন্ট করতে চাই। এই ধরনের বাইরের ভ্যালুগুলোকে আমরা সাধারণত ছোট ছোট 'শূন্য' দিয়ে রিপ্রেজেন্ট করি। ছবি ৬৫ দেখুন। এর পাশাপাশি যে ভ্যালুগুলো একদম দূরের, তাদের '* তারকা' দিয়ে দেখিয়ে দিই।

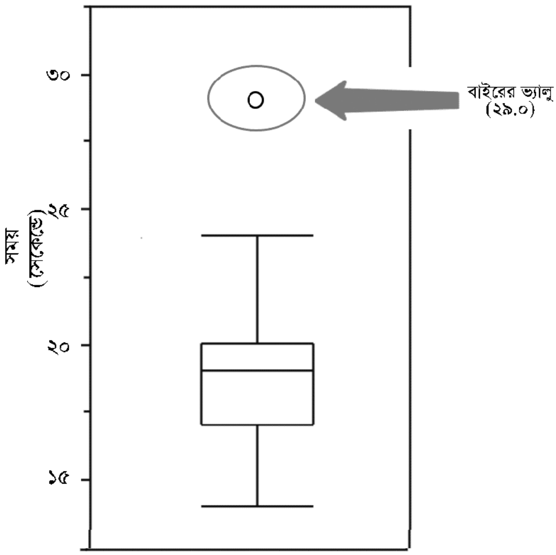
আমাদের ডেটাসেটের ইনপুট নিয়ে যে কথাটা বলা যায়, 'হুইসকার' শুরু হয় উপরের হিঞ্জ ভ্যালু (২০) থেকে, যা চলে যায় একদম ওপরের অ্যাডজাসেন্ট সংখ্যা (২৪) পর্যন্ত। এই একই জিনিস শুরু হয় নিচের হিঞ্জ ভ্যালু (১৭) থেকে, যা চলে যায় একদম নিচের অ্যাডজাসেন্ট সংখ্যা (১৩) পর্যন্ত।



চিত্র ৬৬: একটা পূর্ণাঙ্গ বক্সপ্লট হতে পারে

বাইরের ভ্যালুর কী হবে?

আমাদের বাইরের ভ্যালু হচ্ছে যেটা ভেতরের ফেন্সের বাইরে তবে বাইরের ফেন্সের ভেতরে। বলুন তো— সে ধরনের সংখ্যা কতগুলো আছে? আগে দেখিয়েছি চিত্র ৬৫তে আমরা ঠিকমতো হিসেব করলে দেখা যাবে যে একটামাত্র ভ্যালু যেটা বাইরের ফেন্সের সাথে যাচ্ছে তবে এর বাইরে নয়। সেই সংখ্যাটা হচ্ছে ২৯। এই সংখ্যাকে ঠিকমতো দেখানোর জন্য আমরা একটা ছোট ‘0’ আকৃতির ছবি এঁকে দিতে পারি। আমাদের ডেটা অ্যানালাইসিস সফটওয়্যার নিজে থেকেই করে দেবে। এভাবেই আমাদের বক্স প্লট তৈরি শেষ। শুরুতে একটু মাথা খারাপের মতো হতে পারে, তবে ছবির সাথে নিজে একবার প্র্যাকটিস করলে সব সহজ হয়ে যাবে।



চিত্র ৬৭: আউটলাইয়ারসহ প্লট

বক্স প্লট এবং হিস্টোগ্রাম

বক্স প্লট এবং হিস্টোগ্রাম পাশাপাশি দেখা

এখন আমাদের মনে হতে পারে যে কীভাবে একটা বক্স প্লট, হিস্টোগ্রাম থেকে আলাদা হবে যখন ডিস্ট্রিবিউশন নিয়ে কথা বলছি?

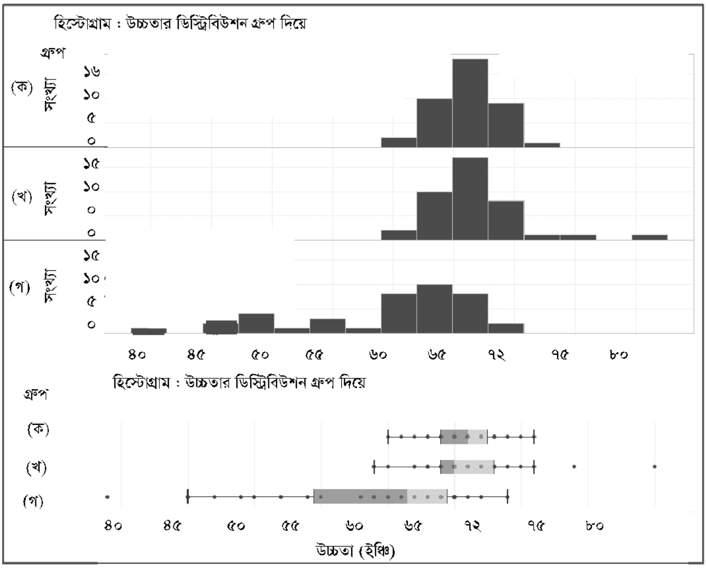
প্রযোজ্যতা কোথায়?

১. হিস্টোগ্রাম সাধারণত 'বিণ' অর্থাৎ একেকটা বাক্স ব্যবহার করে যার মাধ্যমে বিভিন্ন ভ্যালুর ফ্রিকোয়েন্সি বোঝা যায়।

২. বক্স প্লটে ডেটা ডিস্ট্রিবিউশনে মধ্যের ৫০%-এর ডেটা রিপ্রেজেন্টেশন বক্সের ভেতরে থাকে। এর পাশাপাশি 'আউটলাইয়ার' বক্সের বাইরে এবং হুইসকারের বাইরে থাকে।

এই দুটো জিনিসকে একেবারে পাশাপাশি প্লট করলে নিচের ছবির মতো আসতে পারে। আমরা যদি আগের ডেটাসেটের প্রতিটা মানুষের উচ্চতার ডিস্ট্রিবিউশন দেখতে চাই তাহলে এ ধরনের দুটো ছবি দেখা যেতে পারে। ওপরেরটা হিস্টোগ্রাম এবং নিচেরটা বক্স প্লট। মজার কথা হচ্ছে, বক্স প্লট কম জায়গায় হয়ে যায়। এবং এগুলোকে পাশাপাশি খুব সহজে তুলনা করা যায়। সে দিক থেকে তিনটা আলাদা আলাদা হিস্টোগ্রাম অনেক বেশি জায়গা নিয়ে নেয়। তবে সত্যি কথা বলতে, দুটোর উপযোগিতা দুইধরনের। তবে, বক্স প্লট দিয়ে ডেটার ডিস্ট্রিবিউশন দেয়া সোজা।

বক্সপ্লটের উপযোগিতা



চিত্র ৬৮: বক্স প্লট দিয়ে ডিস্ট্রিবিউশন দেখা সোজা

এ ধরনের প্লট আপনারা চেষ্টা করবেন পাওয়ার বিআই এবং ট্যাবলোতে। যত চেষ্টা করবেন, তত এর ভেতরে ঢোকা সহজ হবে।

স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ

ডেটা অ্যানালাইটিক্সে ডেটাবেইসের ধারণা

স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ কেন প্রয়োজন পড়ল?

ভিডিও লিংক—

ডেটা অ্যানালাইটিক্স: সিকুয়েল/SQL নিয়ে ক্লাস	https://youtu.be/HPu7l_lqaLU

ডেটা অ্যানালাইসিস করার একটা পর্যায় প্রশ্ন আসতে পারে, আচ্ছা ডেটা থাকে কোথায়? আমাদের কম্পিউটারে ডেটা কোথায় স্টোর করা হয়? এর সহজ উত্তর হতে পারে ডেটাবেইস, যেটা আমাদের কম্পিউটার সিস্টেমে রাখে। ডেটাবেইস কী হতে পারে? একটা ডেটাবেইস হচ্ছে যেখানে ডেটাকে কালেক্ট করে একটা লোকেশনে স্টোর করে রাখে, বিশেষ করে কম্পিউটার সিস্টেমে এই ডেটাগুলো রাখার স্টোরেজ নিয়ে ধারণা দেখুন ওপরের ভিডিওতে। আমরা কাজ করতে করতে বুঝে যাব কীভাবে একটা ডেটাবেইস আমাদের সাহায্য করবে দরকারি ডেটাগুলোকে কোয়েরি করে বের করে আনা যায়। 'কোয়েরি'টা বোঝাই আসল।

আমাদের কাজের সময় যখন ডেটাগুলোকে শর্ট করে নিয়ে আসব ভেতরে ঢুকতে, যাতে আমাদের রিপোর্টে গ্র্যানুলার ডেটা পাওয়া যায়। এই গ্র্যানুলার ডেটা হচ্ছে একদম ভেতরের ডেটা যেটাকে আমরা বলছি মেটাডেটা। উদাহরণ হিসেবে বলা যেতে পারে, একটা গল্পের বইয়ের ক্যারেক্টার যেটাকে ভবিষ্যতে রেফারেন্সিং করা যায়, সেটাও একটা 'মেটা'। মোদ্দা কথা হচ্ছে ভেতরের ডেটার সাথে একটা রেফারেন্সিং পয়েন্ট যার মাধ্যমে ভেতরের ডেটা খুঁজে পাওয়া যায়, সেটাকেই 'মেটা' বলা যায়।

আমরা যখন ডেটাবেইসের সাথে কাজ করব, তখন মেটাডেটা খুবই গুরুত্বপূর্ণ একটা কনসেপ্ট যখন ডেটাবেইসের ভেতর থেকে রেফারেন্স

করে বিভিন্ন তথ্য বের করে নিয়ে আসব। এটাকে যদি আমরা 'রেফারেন্স গাইড'-এর সাথে তুলনা করি, তাহলে এই গাইড দিয়ে ভেতরের বিভিন্ন ডেটাকে কানেক্ট করতে পারা। অর্থাৎ 'কন্টেক্সটুয়ালিজেশন' করতে পারলে আমাদের কাজ হয়ে যাবে। মেটাডেটা আমাদেরকে বলে কোথা থেকে ডেটাটা আসছে, কীভাবে এবং কোথায় এই ডেটাটা এসেছে, কখন তৈরি হয়েছে, এর সঙ্গে ফাইনাল আউটকাম কীভাবে আসবে তার মধ্যে রেফারেন্সিং হলেই হলো। এটাই বুঝতে হবে। ডেটা অ্যানালাইসিস ডেটাবেইস, অ্যানালিস্টকে সাহায্য করে ডেটাকে ঠিকমতো স্টোর, ম্যানিপুলেট এবং প্রসেস করতে। এ ব্যাপারটা সাহায্য করবে কীভাবে দক্ষতার সাথে ভালো ইনসাইট নিয়ে আসা যায় একটা বা অনেকগুলো ডেটাবেইস থেকে।

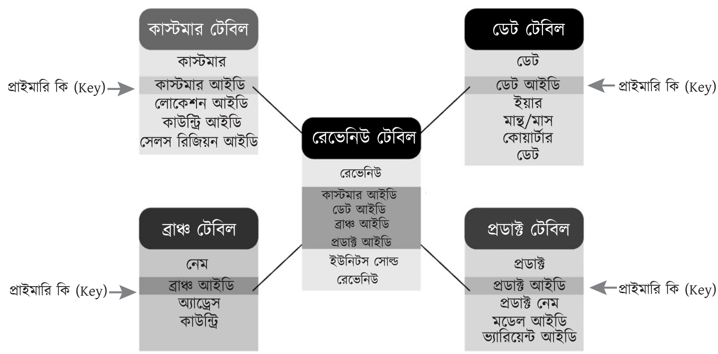
সবচেয়ে মজার কথা হচ্ছে, আমরা যখন বিভিন্ন ডেটাবেইস থেকে ডেটাগুলোকে নিয়ে আসতে পারব বহুল ব্যবহৃত স্প্রেডশিট প্রোগ্রামে, তাহলে আমাদের কাজের পসিবিলিটি অনেক বেড়ে যাবে। 'সিকিউর কমান্ড' দিয়ে ডেটাবেইস থেকে যখন ডেটাগুলোকে স্প্রেডশিট প্রোগ্রামে আনা যাবে, তখন সেটাকে অ্যানালাইসিস করা সহজ হবে। এর মধ্যে 'এসকিউএল' একটা গুরুত্বপূর্ণ অংশ বিশেষ করে এ ধরনের স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ ডেটাবেইস থেকে সরাসরি প্রয়োজনীয় ডেটা নিয়ে আসতে পারে।

রিলেশনাল ডেটাবেইস

যেকোনো ডেটাবেইসের ভেতরে সিরিজ আকারে অনেকগুলো টেবিল থাকতে পারে, তবে টেবিলগুলোর মধ্যে 'রিলেশনশিপ' বা সম্পর্ক না থাকলে সেই টেবিলগুলোর মধ্যে ডেটা ঠিকমতো 'রেফারেন্সিং' করা সম্ভব নয়। আর সে কারণে রিলেশনাল ডেটাবেইসের ভেতর টেবিলগুলোর মধ্যে রেফারেন্সিং করতে হয়, বিশেষ করে যখন টেবিলগুলোর মধ্যে একটা কলামের ডেটা একই রকম থাকে। রিলেশনাল ডেটাবেইস টেবিলগুলোর মধ্যে কমন ফিল্ডগুলোর মধ্যে একটা টেবিলে প্রাইমারি এবং আরেকটা টেবিলে ফরেন কি দিয়ে আমরা নিচের কিছু উদাহরণ দেখতে পারি। প্রাইমারি কি হচ্ছে একটা আইডেন্টিফায়ার, যার কলামের প্রতিটা ভ্যালু ইউনিক, যাতে ওই কলামের মধ্যে প্রতিটা রেকর্ডকে আলাদা আলাদা

করে চেনা যায়। টেবিলের ওই পার্টিকুলার কলামে যতগুলো রেকর্ড থাকবে সেটা পুরো টেবিলের মধ্যে আলাদা আলাদা হবে। কারণ সেই সারিগুলোতে যতগুলো রেকর্ড থাকবে, সেগুলোকে অন্য টেবিলের সাথে সম্পর্ক করতে গেলে সেগুলোকে ‘ইউনিক’ হতে হবে। যেমন, দুটো টেবিলের মধ্যে স্কুলের ছাত্রদের রোল নম্বর থাকে তাহলে কোনো সারিতে একই রোল নম্বর অন্য কোনো সারিতে থাকতে পারবে না। দুটো টেবিলের মধ্যে ছাত্রের রোল নম্বর ‘ইউনিক’ থাকবে, যাতে দুটোর মধ্যে রেফারেন্সিং করা যায়।

কলামের মধ্যে ‘ফরেন কি’ হচ্ছে একটা টেবিলের মধ্যে সংযোগসূত্র যেটা ‘প্রাইমারি কি’ হতে পারে আর একটা টেবিলে। একটা টেবিলে যেভাবে একটাই ‘প্রাইমারি কি’ হতে পারে, সেখানে অন্য টেবিলে অনেকগুলো ‘ফরেন কি’ হতে পারে, ফলে এই দুটোর মধ্যে সম্পর্ক তৈরি হবে যখন প্রাইমারি কি এবং ‘ফরেন কি’ আলাদা ভ্যালু নিয়ে বিভিন্ন টেবিলে থাকলেও সেটা নিজেদের মধ্যে একই ভ্যালু দিয়ে রেফারেন্সিং করবে।



চিত্র ৬৯: প্রাইমারি কি দিয়ে সম্পর্ক

ওপরের উদাহরণ দেখুন। অনেক সময় কিছু কিছু টেবিলে প্রাইমারি কি’র প্রয়োজন নেই। এখানে ‘রেভিনিউ টেবিলের’ মধ্যে বেশ কিছু ‘ফরেন কি’ আছে, তবে কোনো ‘প্রাইমারি কি’ নেই। কখনো কখনো একটা প্রাইমারি কি টেবিলের মাল্টিপল অংশে মানে বেশ কয়েকটা কলাম নিয়ে তৈরি হতে পারে যাদের বলা হয় কম্পোজিট কি। এখানে রেভিনিউ টেবিলে দেখুন। বিভিন্ন টেবিলের ‘প্রাইমারি কি’গুলো ‘ফরেন কি’ হিসেবে রেভিনিউ টেবিলে চলে এসেছে।

আমাদের মনের কথা বলবে, স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ—

আগেই বলেছি ডেটাবেইসগুলো একটা স্পেশাল ল্যাঙ্গুয়েজ ব্যবহার করে বিভিন্ন ধরনের কোয়েরি অর্থাৎ নিজেদের মধ্যে কমিউনিকেশন করে। এর মধ্যে স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ যেটাকে আমরা বলি 'এসকিউএল' অথবা 'সিক্যুয়েল', যার মাধ্যমে ডেটা অ্যানালিস্ট কোয়েরি করতে পারে বিভিন্ন ডেটাবেইসকে। এর ফলে একজন ডেটা অ্যানালিস্ট 'এসকিউএল' ব্যবহার করে তার পছন্দমতো ডেটাকে আনার জন্য কোয়েরি করে একটা বড় ডেটাসেট থেকে। রিলেশনাল ডেটাবেইসে একজন ডেটা অ্যানালিস্ট কোয়েরি করে ডেটা আনতে পারে তার সম্পর্কযুক্ত টেবিলগুলো থেকে। আপনারা সামনে দেখবেন এই 'এসকিউএল' কত বড় একটা ক্ষমতাবান টুল, বিশেষ করে যখন বড় বড় ডেটাসেট থেকে 'গ্র্যানুলার' ডেটা নিয়ে আসতে চায়।

লং টার্ম 'এসকিউএল' শেখার রোডম্যাপ

'এসকিউএল' নিয়ে কাজ করার জন্য সবচেয়ে সহজ উপায় কী?

এসকিউএল নিয়ে কাজ করার জন্য সবচেয়ে ভালো উপায় হচ্ছে দুটো কমিউনিটি প্রডাক্ট নিজের পিসিতে ডাউনলোড করা।

১. মাইএসকিউএল কমিউনিটি সার্ভার এডিশন।

২. মাইএসকিউএল ওয়ার্কবেঞ্চ। দুটো প্রডাক্ট কমিউনিটি এডিশন বলে ফ্রি হিসেবে ব্যবহার করা যায়। এতে খুব সহজে 'এসকিউএল' কমান্ড প্র্যাকটিস করা যায়। তবে যারা কোনো ধরনের অ্যাপ্লিকেশন ডাউনলোড করতে চান না এবং শুধু ব্রাউজার দিয়ে 'এসকিউএল' শিখতে চান, তাদের জন্য 'গুগল বিগ কোয়েরি' খুব ভালো একটা অপশন।

ডেটাবেইসের সাথে কীভাবে কানেক্ট করতে হয়?

ডেটাকে ম্যানিপুলেট করার জন্য আমরা আপডেট, ডিলিট, অ্যাড করতে পারি। আমাদের দেখামতে, ডেটা সব সময় ক্লিন থাকে না। আমরা যদি ক্লিন করতে চাই ডেটাকে, স্প্রেডশিট ব্যবহার করতে পারি একদম

শুরুতে। স্প্রেডশিটের পাশাপাশি এসকিউএল ব্যবহার করতে পারি ডেটা ম্যানিপুলেশনের শুরুতে। ‘প্রসেস ডেটা ফর্ম ডার্টি টু ক্লিন’ মানে আমরা শুরুতে ধরে নিই যে আমাদের ডেটাকে প্রসেস করে এর ভেতর থেকে ‘অদরকারি’ ডেটা আমরা ‘ক্লিন’ করে ব্যবহার উপযোগী করব। আমাদের আশপাশে যত ডেটা আছে, সেই ডেটাগুলো বিশেষ করে ক্লিন থাকে না, কারণ আমরা বিভিন্ন রকম সেন্সর থেকে ডেটা নিই, আমাদের ‘আইওটি’ সেগমেন্ট ডেটাবেইস থেকে ডেটা নিই। এ ধরনের ডেটার মধ্যে অনেক সমস্যা থাকে। এই জায়গাটা নিয়ে আলাপ করতে আমি গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন কোর্স থেকে ধারণা নিয়েছি। গুগল ডেটা অ্যানালাইসিস সার্টিফিকেশনের পারসপেক্টিভ থেকে যদি বলি তারাও কিন্তু ডেটা ম্যানিপুলেশনে সিকিউএল নিয়ে কাজ করেছে। সুতরাং, আমাদের এখানে যে কাজগুলো হচ্ছে, সেই কাজগুলোর মধ্যে একটা বড় পারসপেকটিভ হচ্ছে আমরা কীভাবে ডেটাকে ক্লিন করতে পারি? এখানে আমরা হাতেকলমে দেখব কীভাবে আমাদের ডেটাগুলোকে ম্যানিপুলেট করতে পারি বা ডেটা দেখতে পারি।

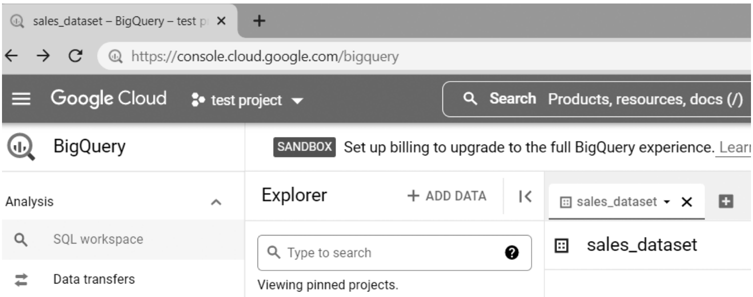
BigQueryতে কীভাবে কাজ করব?

আমরা কোনো কিছু ইন্সটলেশন না করতে চাইলে ওয়েববেসড গুগল ‘বিগ কোয়েরি’ ব্যবহার করতে পারি। এতে একটা ওয়েব ব্রাউজারই যথেষ্ট। গুগল ডেটা অ্যানালাইটিক্স কোর্সে এ ব্যাপারটি চমৎকারভাবে উপস্থাপন করেছে।

গুগল ক্লাউট প্ল্যাটফর্ম থেকে ফ্রিতে শেখা যায়

আমরা জানি যে BigQuery একটি গুলের প্রডাক্ট। আমরা গুগলে BigQuery লিখে সার্চ দিলে আমাদের প্রথমে চলে আসবে, Google BigQuery.

তবে সেরা উপায় হচ্ছে নিচের লিংকটি ব্রাউজারে লিখে সেখানে চলে যাওয়া। https://console.cloud.google.com/bigquery



চিত্র ৭০: গুগল বিগ কোয়েরি লিংক

শুরুতে আমরা 'স্যান্ডবক্স' ব্যবহার করব। এটা ফ্রি, রেজিস্ট্রেশনের পরে কোনো ধরনের ঝামেলার মধ্যে যেতে হবে না। গুগল ক্লাউডে যাওয়ার পর আমরা BigQuery platfrom-এর নিচে SQL Workspace চলে যাব। BigQuery-এর SQL Workspace সিকুয়েল শেখার জন্য একটি অসাধারণ জায়গা।

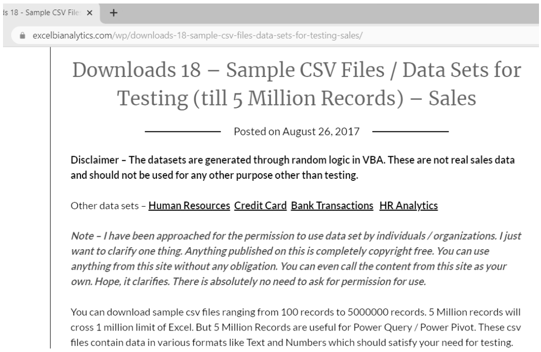
নতুন ডেটাসেট

যেকোনো জিনিস শিখতে শুরুতে দরকার একটা ভালো ডেটাসেট, যা আমাদেরকে 'সিকুয়েল' শিখতে সাহায্য করবে। আমি একটু কষ্ট করেছি, একটা ভালো ডেটাসেট বের করতে। সে কারণে, এই ডেটাসেট এখান থেকে অনেক নতুন 'ইনসাইট' পাওয়া যাবে। আমি অনুরোধ করব, এই এক্সারসাইজের বাইরে নিজের প্র্যাকটিস চালু রাখবেন। চলুন, নিচের ডেটাসেটে।

ডেটাসেট পাওয়া যাবে—

```
https://excelbianalytics.com/wp/wp-content/
uploads/2017/07/10000-Sales-Records.zip
```

ডেটাসেটের কিছু ধারণা পাবেন এই সাইটে। ডাউনলোড করে ডেটাসেট খুলে দেখুন মাইক্রোসফট এক্সেল অথবা গুগল ডেটাশিটে।



ds 18 - Sample CSV File ×

excelbianalytics.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/

Downloads 18 – Sample CSV Files / Data Sets for Testing (till 5 Million Records) – Sales

Posted on August 26, 2017

Disclaimer – The datasets are generated through random logic in VBA. These are not real sales data and should not be used for any other purpose other than testing.

Other data sets – Human Resources Credit Card Bank Transactions HR Analytics

Note – I have been approached for the permission to use data set by individuals / organizations. I just want to clarify one thing. Anything published on this is completely copyright free. You can use anything from this site without any obligation. You can even call the content from this site as your own. Hope, it clarifies. There is absolutely no need to ask for permission for use.

You can download sample csv files ranging from 100 records to 5000000 records. 5 Million records will cross 1 million limit of Excel. But 5 Million Records are useful for Power Query / Power Pivot. These csv files contain data in various formats like Text and Numbers which should satisfy your need for testing.

চিত্র ৭১: ডেটাসেট পাবার লিংক

এর পাশাপাশি, যদি আগের লিংক কখনো কাজ না করে, সেখানে এই ডেটাসেটটি পাবেন এই বইয়ের গিটহাব পেজে।

```
https://github.com/raqueeb/data_analyst/raw/main/10000-Sales-Records.zip
```

ডাউনলোড করে নিন। কী দেখছেন?

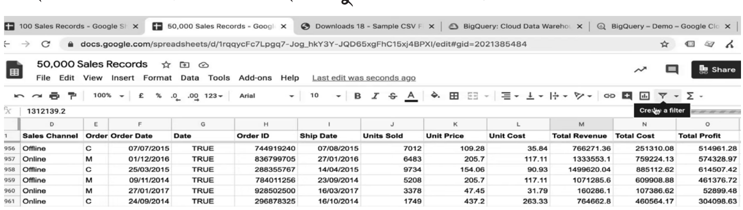
এই ডেটাসেটকে যেকোনো জায়গায় সেলস ক্যাটাগরিতে ফেলা যাবে। আমরা দিন শেষে সবাই সেলসম্যান!

এই 'সিএসভি' ফাইলের কলামের নামগুলো দেখলে আপনাদের এই সেলস ডেটাসেট সম্পর্কে ভালো ধারণা চলে আসবে।

Region	**Country**	**Item Type**	**Sales Channel**	**Order Priority**	**Order Date**	**Order ID**
Ship Date	**Units Sold**	**Unit Price**	**Unit Cost**	**Total Revenue**	**Total Cost**	**Total Profit**

চিত্র ৭২: ডেটাসেটের কলামের নাম

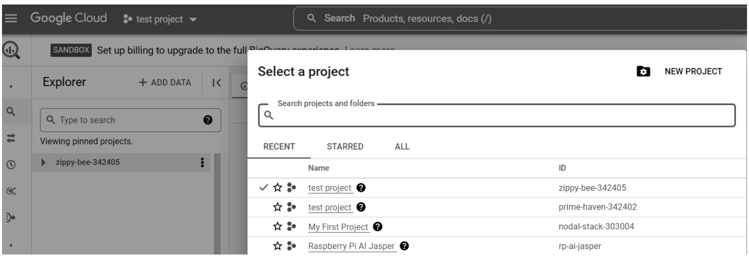
গুগলশিটে খুলে দেখুন। এটা ৫০,০০০ রেকর্ডের ফাইল। বাকি ফাইলগুলোর কনটেন্ট একই। শুধু রেকর্ড সংখ্যা বেশি।



	D	E	F	G	H	I	J	K	L	M	N	O
1	Sales Channel	Order	Order Date	Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
956	Offline	C	07/07/2015	TRUE	744919240	07/08/2015	7012	109.28	35.84	766271.36	251310.08	514961.28
957	Online	M	01/12/2016	TRUE	836799705	27/01/2016	6483	205.7	117.11	1333553.1	759224.13	574328.97
958	Offline	C	25/03/2015	TRUE	288355767	14/04/2015	9734	154.06	90.93	1499620.04	885112.62	614507.42
959	Offline	M	09/11/2014	TRUE	784011256	23/09/2014	5208	205.7	117.11	1071285.6	609908.88	461376.72
960	Online	M	27/01/2017	TRUE	928502500	16/03/2017	3378	47.45	31.79	160286.1	107386.62	52899.48
961	Online	C	24/09/2014	TRUE	296878325	16/10/2014	1749	437.2	263.33	764662.8	460564.17	304098.63

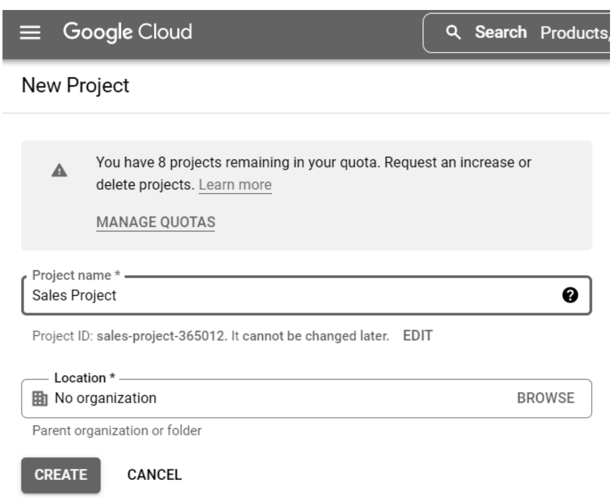
চিত্র ৭৩: সেলস রেকর্ড দেখুন গুগলশিটে

১. চলে যাই, https://console.cloud.google.com/bigquery — গুগল অ্যাকাউন্ট ব্যবহার করে স্যান্ডবক্স চালু করুন। শুরুতেই প্রজেক্টের নাম। ধরা যাক— ‘টেস্ট প্রজেক্ট’ অথবা ‘Sales Project’, যেটা আপনার সুবিধা হয়। ক্লাউড প্ল্যাটফর্ম লেখার পাশে প্রজেক্টে ক্লিক করে নতুন প্রজেক্ট তৈরি করে নিই। এখানে কিছু দৈবচয়নের ভিত্তিতে প্রজেক্টের নাম আসতে পারে। আমরা নিউ প্রজেক্ট তৈরি করি আমাদের পছন্দমতো নাম দিয়ে। নাম দিলাম সেলস প্রজেক্ট, তবে সেটা দৈবচয়নের ভিত্তিতে এরপর কিছু সংখ্যা যোগ হয়ে গেল।



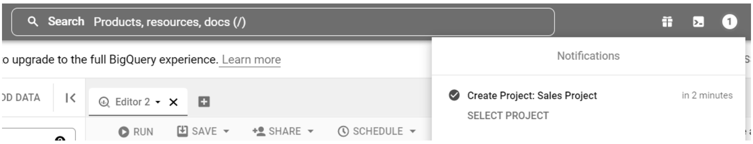
চিত্র ৭৪: প্রজেক্টের নাম দিই এখানে

নিচের ছবিটা দেখুন, এখানে নতুন প্রজেক্টর নাম এবং অন্যান্য অংশ স্বয়ংক্রিয়ভাবে যোগ হয়ে গেল।



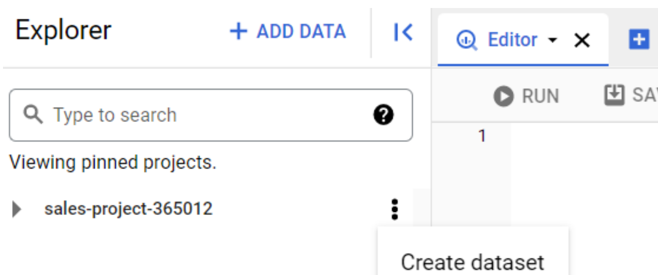
চিত্র ৭৫: সেলস প্রজেক্টের নাম

সেলস প্রজেক্ট হিসেবে প্রজেক্ট তৈরি হয়ে গেল, এখন এটাকে সিলেক্ট করার পালা।

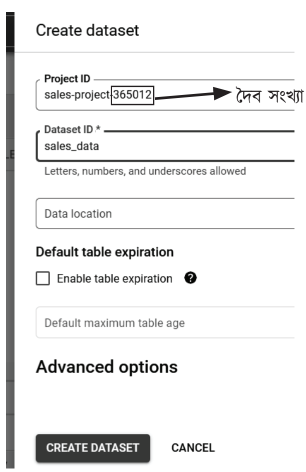


চিত্র ৭৬: তৈরি হয়ে গেল প্রজেক্ট

প্রজেক্টের ওপর ‘রাইট ক্লিক’ করে আমরা নতুন ডেটাসেট তৈরি করি। ডেটাসেটের নাম দিলাম সেলস ডেটা। এটাকে অনেক সময় ‘ডেটাসেট আইডি’ বলা হয়ে থাকে।

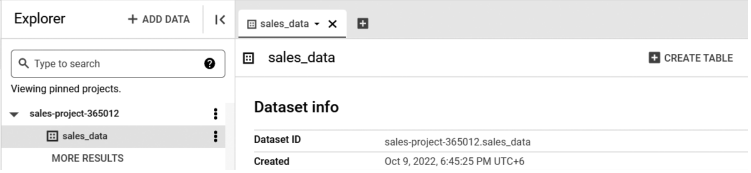


চিত্র ৭৭: প্রজেক্টের সাথে দৈব সংখ্যা যুক্ত হয়ে গেছে



চিত্র ৭৮: ডেটাসেট তৈরি

ডেটাসেট তৈরি হয়ে গেলে সেই ডেটাসেটের মধ্যে টেবিল তৈরি করার পালা। 'ক্রিয়েট টেবিল' বাটন চাপ দিয়ে নিচের ছবিটাকে ফলো করি। আমরা নতুন করে একটা ব্ল্যাংক টেবিল তৈরি করতে পারতাম, তবে যেহেতু আমাদের 100000 রেকর্ডের সেলসের একটি টেবিল ডাউনলোড করা আছে, সেটাকেই বরং আপলোড করে দিই এই ডেটাসেটের ভেতরে।

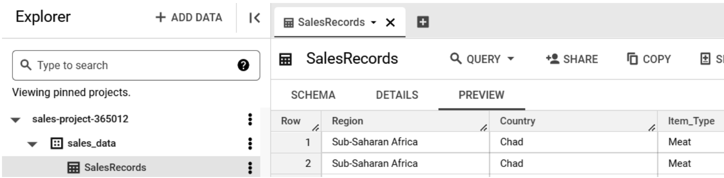


চিত্র ৭৯: এখান থেকে টেবিল তৈরি করা যায়

এখানে সোর্স হিসেবে 'আপলোড' সিলেক্ট করে সিলেক্ট ফাইল থেকে আমাদের ডাউনলোড করা 100000 সেলস রেকর্ডের ফাইলটাকে দেখিয়ে দিই। এখানে ফাইল ফরম্যাট স্বয়ংক্রিয়ভাবে 'সিএসভি' চলে এসেছে। এর পাশাপাশি আমরা 'বিগ কোয়েরি'কে বলছি 'স্কিমা' জেনারেট করতে, যাতে 'সিএসভি' ফাইল থেকে সঠিক 'স্কিমা' তৈরি করতে পারে।

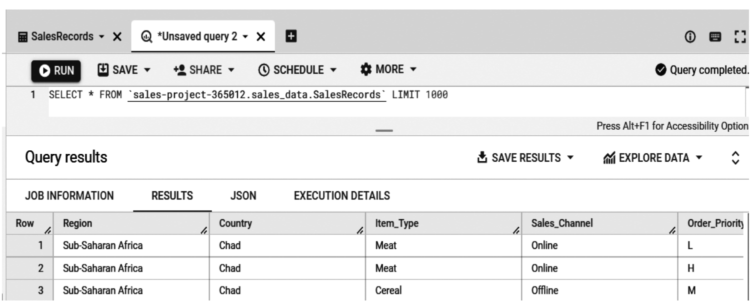
চিত্র ৮০: ডেটাসেট ফাইল আপলোড করুন এখানে

টেবিল তৈরি হয়ে গেল। এখানে আমরা সেলস রেকর্ড থেকে স্কিমা, ডিটেইলস এবং প্রিভিউ থেকে ডেটাগুলোকে গুগলশিটের মতো করে দেখতে পারি। আমাদের এখানে কাজ হবে, কোয়েরি বাটনে চাপ দিয়ে 'অটো জেনারেটেড কোয়েরি স্ট্রিং' জেনারেট করা।



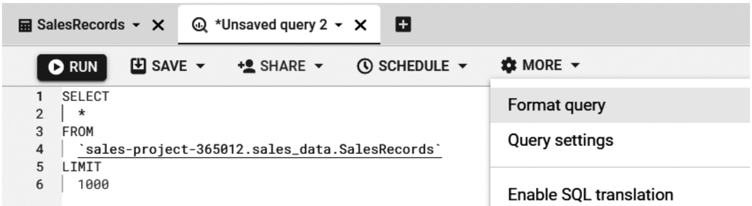
চিত্র ৮১: টেবিল 'প্রিভিউ' করি

এখানে, 'সিলেক্ট' একটা এসকিউএল স্টেটমেন্ট। অটো জেনারেটেড কমান্ডের মধ্যে আমরা * (ওয়াইল্ড কার্ড) মানে 'ক্যাচ অল' চিহ্ন দেব যাতে পুরো টেবিলের মধ্যে সবগুলো কলাম নিয়ে আসে। এতে একটা বেসিক লিমিট দেওয়া হচ্ছে যেটা 1000, যা আসলে দেখার জন্য অনেক বটে। এখানে সেলস প্রজেক্টের নামের পরে ডট, এরপর ডেটাসেটের নামের পরে ডট এবং শেষে টেবিলের নাম। তবে এসকিউএল কমান্ডের ফরম্যাট অনুযায়ী জিনিসটা সাজাতে গেলে আমরা 'মোর' চাপ দিয়ে ফরম্যাট কোয়েরি সিলেক্ট করে নেব।



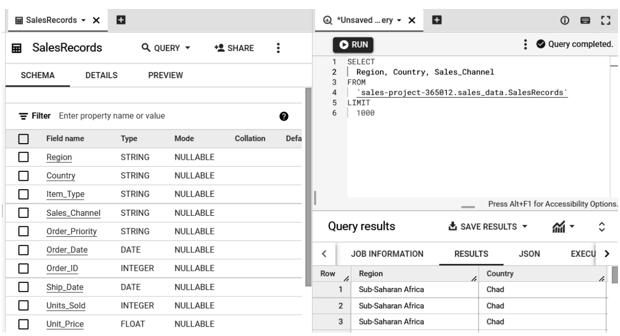
চিত্র ৮২: 'সিলেক্ট' স্টেটমেন্ট

নিচের ছবি দেখুন, এসকিউএল কমান্ড ফরম্যাটে জিনিসটা নিজে থেকে সাজিয়ে নিয়েছে। এতে আমাদের প্রতিটা ঘরে বিভিন্ন কমান্ড যোগ করে 'রান' বাটন চালানো যাবে। এখানে * অর্থাৎ সবগুলো কলাম না সিলেক্ট করে, বরং স্কিমা থেকে বিভিন্ন কলাম সিলেক্ট করে এখানে আনা যায়। ফলে 'স্টার' চিহ্ন অর্থ সব কলাম বাদ দিয়ে আমাদের পছন্দমতো যেগুলো প্রয়োজন সেই কলামগুলোকে স্কিমা থেকে সিলেক্ট করলে স্বয়ংক্রিয়ভাবে সেগুলো আমাদের কোয়েরি অংশে ঢুকে পড়বে। ভিডিও দেখুন।



চিত্র ৮৩: সুন্দর করে সাজিয়ে নিয়েছে নিজেই

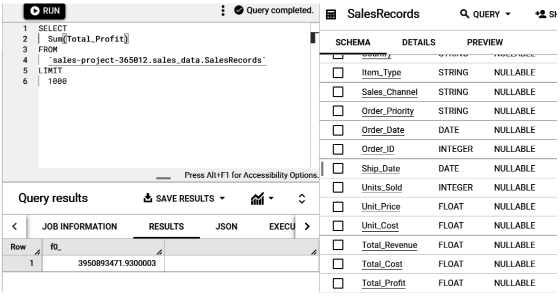
এখানে আমরা স্টার চিহ্ন ফেলে দিয়ে পাশের স্কিমা থেকে রিজিওন, কান্ট্রি এবং সেলস চ্যানেল সিলেক্ট করা মাত্রই সিলেক্টের নিচে সবগুলো কলাম স্বয়ংক্রিয়ভাবে চলে এসেছে। এর ফলে আমরা আমাদের টেবিল থেকে শুধু এই তিনটা কলামের ডেটা দেখার সুযোগ করে নিচ্ছি। এভাবেই আমরা ফিল্টার করে গ্র্যানুলার ভিত্তিতে ডেটা দেখতে পারব। এ কারণেই এসকিউএল এত গুরুত্বপূর্ণ এবং অনেক ক্ষমতাশালী একটি টুল। মিলিয়ন বিলিয়ন ডেটার রেকর্ড থেকে গ্র্যানুলার ডেটা আনার সহজ উপায়।



চিত্র ৮৪: বায়ের কলাম নাম টেনে আনুন সিলেক্টের ওপর

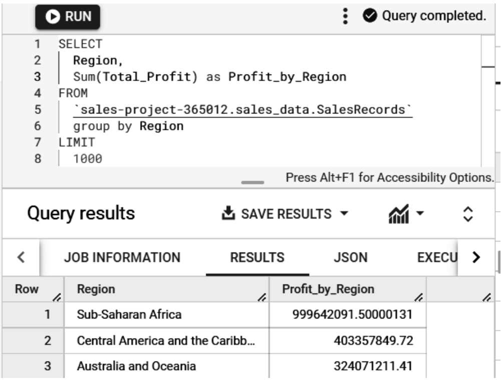
এবার একটা যোগফল দেখে আসি। আমাদের স্কিমা এবং টেবিলের শেষে 'টোটাল প্রফিট' বলে একটা কলাম আছে। এই 'টোটাল প্রফিট'কে আমরা যোগ করে একটা কলামে দেখানোর চেষ্টা করব। আমরা স্কিমা থেকে টোটাল প্রফিট কলামের নামটা সিলেক্ট করে আনি, এরপর সেটার সামনে 'সাম' যোগ করে দিই। টোটাল প্রফিটকে ব্র্যাকেটের মধ্যে নিয়ে এনে রান বাটন চালাই। f0_ নামে একটা টেম্পোরারি কলাম যোগ হয়েছে, যার মধ্যে 'টোটাল প্রফিট'-এর যোগফল দেখানো হয়েছে। ব্যাপারটা খুবই সোজা, টোটাল প্রফিট কলাম থেকে সবগুলো রেকর্ডের যোগফলকে সিলেক্ট করে

ফলাফল দেখানো হয়েছে। এখানে ফ্রম, প্রজেক্ট নাম. ডেটাসেটের নাম. টেবিলের নাম, দেখানো হয়েছে, যেখানে লিমিট করা হয়েছে 1000, শুধু দেখানোর জন্য।



চিত্র ৮৫: স্কিমা ধরে 'সিলেক্ট' স্টেটমেন্ট

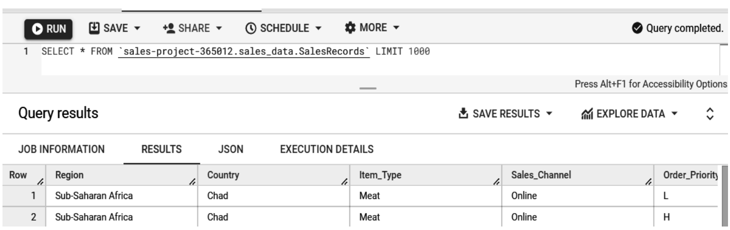
আমরা একই জিনিস যদি 'রিজিওনভিত্তিক' দেখতে চাই তাহলে কী করা উচিত? আমি আপনাকে প্রশ্ন করছি। ঠিক ধরেছেন, আপনার উত্তর ছিল সিলেক্টের পরে রিজিওন ঢুকিয়ে দিতে, এখানে রিজিওন হাতে লেখার দরকার নেই। কারণ এই রিজিওন আমরা ক্লিক নিয়ে আসতে পারি টেবিলের স্কিমা থেকে, রিজিওন কলাম ধরে। তবে আমরা যেহেতু রিজিওনভিত্তিক গ্রুপ করে দেখতে চাচ্ছি, সে কারণে এখানে ফ্রমের পর 'গ্রুপ বাই' রিজিওন বলতে হবে। নিচের রেজাল্ট টেবিলে দেখুন, তার পুরো ফলাফল নিয়ে এসেছে। এখন প্রশ্ন হচ্ছে এই টেম্পোরারি কলামের নাম কি নিজের মতো করে দেওয়া যায় না? অবশ্যই যায়। এক্ষেত্রে এখানে 'সাম অফ টোটাল প্রফিট'কে অ্যাজ হিসেবে যে কলামের নাম দিব, সেটাই এখানে দেখাবে। নিচের ছবিটা দেখুন।



চিত্র ৮৬: কলামের নাম নিজের মতো করে

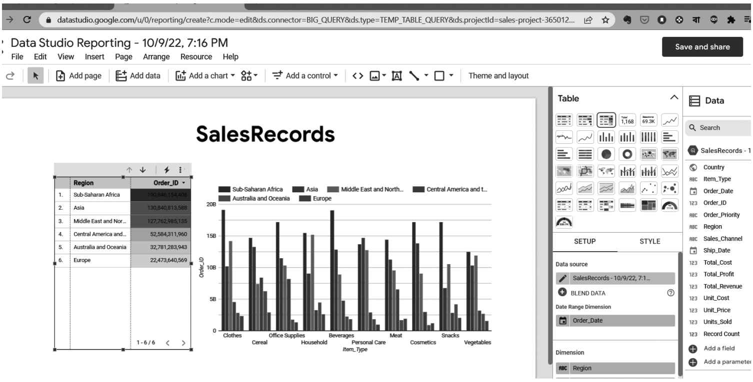
ঠিকই f0_ টেম্পোরারি কলামের নাম পাল্টে আমাদের পছন্দের কলামের নাম দিয়েছে। সত্যি কথা বলতে, এসকিউএল, একটা অসাধারণ কোয়েরি ল্যাঙ্গুয়েজ, যার মাধ্যমে যেকোনো ধরনের ডেটাবেইসকে অ্যাকসেস করে পছন্দমতো 'কোয়েরি' করা যায়। গুগলের বিগকোয়েরি ব্যবহার করার পেছনে আরেকটা উদ্দেশ্য ছিল আমার। আপনাদের গুগল ডেটা স্টুডিওর মতো একটা অসাধারণ ডেটা ভিজ্যুয়ালাইজেশন এবং রিপোর্টের সাথে পরিচয় করে দেওয়া, যেটা কয়েক ক্লিকেই অসাধারণ ড্যাশবোর্ড বানাতে পারে। মাইক্রোসফট পাওয়ার বিআই অথবা ট্যাবলো-এর সমকক্ষ এই ভিজ্যুয়ালাইজেশন এবং রিপোর্টিং টুল এখনই ব্যবহার করতে পারবেন বিগ কোয়েরির মধ্যে বসে।

চলুন, আমরা ফিরে আসি একদম শুরুর এসকিউএল সিনট্যাক্সে। কারণ এতে ভিজ্যুয়ালাইজেশনটা আরও চমৎকারভাবে বোঝা যাবে। বিগ কোয়েরির মধ্যে নিচের দিকে 'এক্সপ্লোর ডেটা' বলে একটা অপশন পাবেন, যা আমাদের নিয়ে যাবে গুগল ডেটা স্টুডিওতে।



চিত্র ৮৭: এক্সপ্লোর ডেটা দিয়ে চলে যান গুগল ডেটা স্টুডিওতে

'এক্সপ্লোর ডেটা' চাপ দেওয়ার সঙ্গে সঙ্গে খুলে যাবে নিচের মতো করে গুগল ডেটা স্টুডিওর রিপোর্টিং টুল। এই রিপোর্ট তৈরি করার জন্য ডানদিকে ডেটাসহ বিভিন্ন ধরনের ভিজ্যুয়ালাইজেশন টেমপ্লেট এবং কোন ডেটা কোন ধরনের ডাইমেনশনে যাবে, সেটা ভালোভাবেই বোঝা যায়। সত্যি কথা বলতে, 'গুগল ডেটা স্টুডিও' যেকোনো ডেটা অ্যানালিস্টের জন্য একটা গোল্ড মাইন।



চিত্র ৮৮: গুগল ডেটা স্টুডিওর রিপোর্টিং

এসো নিজে করি: গুগল ডেটা স্টুডিও

গুগল ডেটা স্টুডিওর কারুকাজ কিছুটা দেখানো হয়েছে প্লেলিস্টের ভিডিওতে। এখানে আপনার অপছন্দমতো কলাম বা মেজারকে টানতে ভুলবেন না। আপনি জিনিসটাকে নষ্ট করতে পারবেন না। ডানের টেবিল এবং ডেটা থেকে ইচ্ছেমতো ডেটা এবং টেমপ্লেট টানুন ‘ভিজুয়্যালাইজেশন’ অংশে। ভালো ইনসাইট আসবে এতে। গুগল ডেটা স্টুডিওর ব্যবহার অনেকটাই ট্যাবলো এবং পাওয়ার বিআইর মতো।

টেক্সটবুক এবং ভিডিও সিরিজের সংযোগ

ডেটা অ্যানালাইটিক্স শিখতে যেভাবে টেক্সটবুক (যেমন, এই বইটা) প্রয়োজন— সেভাবে টেক্সটবুকের সাহায্যকারী ভিডিও অনেকাংশেই সাহায্য করে। আর সে কারণে এই ভিডিও সিরিজ। তবে টেক্সটবুক হিসেবে এই বইটি না পড়ে সরাসরি ভিডিওতে চলে গেলে গোড়ায় গলদ থাকবে। আমাদের প্রয়োজন বেসিক ধারণা, যেখানে ফার্স্ট প্রিন্সিপাল থিঙ্কিং অনুযায়ী সাহায্য করবে এই টেক্সট বই। এরপরে বইয়ে কোনো কিছু না বুঝলে আমরা সরাসরি চলে যাব ভিডিওতে।

লিংক:
https://www.youtube.com/playlist?list=PL5NPSn95n_L4Z5bpCOqHW_QU8nxgywGM2

ভিডিও প্লেলিস্ট

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
১	ডেটা অ্যানালাইটিক্স মেন্টরশিপ: গুগল এবং আইবিএম ডেটা অ্যানালাইটিক্স সার্টিফিকেশন	https://youtu.be/K_XEnWAJO-U
২	ডেটা অ্যানালাইটিক্স মেন্টরশিপ: ওপেনসোর্স ডেটা টুল, ডেটা ট্রান্সফরমেশন/স্টোরেজ	https://youtu.be/fSmkkisToh0
৩	ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১: সিদ্ধান্ত কীভাবে নিই আমরা?	https://youtu.be/WP39lGIYzvk

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
৪	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ২: ডেটা স্টোরি-টেলিং	https://youtu.be/VYRS3IECSWM
৫	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ৩: ডেটা স্টোরি-টেলিং	https://youtu.be/oU9x7qXU7BE
৬	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ৪: ডেটা স্টোরি-টেলিং	https://youtu.be/QK7UPaevBuw
৭	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ৫: ডেটা অ্যানালাইসিসের ৬ ধাপ	https://youtu.be/gBo5MLVHlZw
৮	নন-প্রোগ্রামারদের ডেটা অ্যানালাইটিক্স মেন্টরশিপ	https://youtu.be/12ETNqeBotQ
৯	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ৭: ডেটা ইকোসিস্টেম কী?	https://youtu.be/bZc8VF9XYrU
১০	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ৮: কোডিং ছাড়া ডেটা ভিজ্যুয়ালাইজেশন	https://youtu.be/MWEfjAoyyPA
১১	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ৯: অ্যানালাইটিক্যাল থিঙ্কিং — ভিজ্যুয়ালাইজেশন	https://youtu.be/10RqvuX7SaE
১২	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স টুল ১০: ট্যাবলো/ এক্সেল/গুগলশিট (আইরিস/ টাইটানিক ডেটাসেট)	https://youtu.be/TKM8zUvRG3A
১৩	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১১: অ্যানালাইটিক্স টুল ট্যাবলো (আইরিস ডেটাসেট)	https://youtu.be/PWiaXY8Q35c
১৪	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১২: ট্যাবলো দিয়ে ক্লাস্টার (আইরিস ডেটাসেট)	https://youtu.be/_rFq8vVXt7A

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
১৫	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১৩: ট্যাবলো দিয়ে ক্লাস্টার অ্যাকুরেসি -আইরিস	https://youtu.be/_7WnMDGjpLk
১৬	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইটিক্স মেন্টরশিপ ১৪: ট্যাবলো দিয়ে ক্লাস্টার, স্কয়ার-সাম	https://youtu.be/tKWssvSCe_w
১৭	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ১৫: ট্যাবলো এক্সপ্লোরেটোরি অ্যানালাইসিস - টাইটানিক ডেটাসেট	https://youtu.be/LnfprC9GzyM
১৮	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ১৬: ডেটা অ্যানালিস্ট নাকি ডেটা সায়েন্টিস্ট?	https://youtu.be/QEszjMwkcZQ
১৯	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ১৬: ডেটা অ্যানালিস্ট নাকি ডেটা সায়েন্টিস্ট? (Updated)	https://youtu.be/Qol1b9E2Et8
২০	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ১৭: টাইটানিক ডেটা ভিজ্যুয়ালাইজেশন - ট্যাবলো দিয়ে	https://youtu.be/XWgi8wJJdvU
২১	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ১৮: মাইক্রোসফট এক্সেল অ্যানালাইসিস টুলপ্যাক -আইরিস ডেটাসেট	https://youtu.be/-onXgxSFsWo
২২	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ১৯: এক্সেল অ্যানালাইসিস টুলপ্যাক - কোরিলেশন, আইরিস ডেটাসেট	https://youtu.be/Ryz9N092ixU
২৩	নন-প্রোগ্রামারদের জন্য ডেটা মেন্টরশিপ ২০: টাইটানিক ডেটাসেট থেকে ডেটা ভিজ্যুয়ালাইজেশন, ট্যাবলো	https://youtu.be/nsfpvfogwJU
২৪	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২১: গ্লোবাল সুপারসেট ডেটাসেট থেকে অ্যানালাইসিস, ট্যাবলো	https://youtu.be/SMlSbP1kJhY
২৫	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২২: ট্যাবলো, গ্লোবাল সুপারসেট ডেটাসেট, পিভট টেবিল	https://youtu.be/Full0_fX2v8

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
২৬	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৩: ট্যাবলো, টাইটানিক ডেটাসেট, ডাইমেনশন এবং মেজার	https://youtu.be/Xktl7aAqJNA
২৭	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৪: এক্সেল, টাইটানিক ডেটা, ডেটা কীভাবে ফাইনাল আউটকাম নির্ধারণ করে?	https://youtu.be/GzhzP9w21Ow
২৮	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৫: টাইটানিক, কীভাবে যেকোনো 'বিজনেস কোশ্চেন'-এর উত্তর বের করা যায়?	https://youtu.be/3ULpTdKa3oE
২৯	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৬: টাইটানিক, কীভাবে মিসিং ডেটাকে নতুন করে তৈরি করব?	https://youtu.be/AQpPC3N-wkY
৩০	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৭: টাইটানিক ডেটাসেট, বিজনেস ইন্টেলিজেন্স, পিভট টেবিল, পিভট চার্ট	https://youtu.be/X1iT0Qo_sPM
৩১	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৮: পাওয়ার বিআই থেকে টাইটানিক ডেটাসেটের 'কি ইনফ্লুয়েন্সার'	https://youtu.be/TobshfkmrNE
৩২	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ২৯: পাওয়ার বিআই থেকে টাইটানিক ডেটাসেটের 'কি সেগমেন্টেশন'	https://youtu.be/bdez7go5KLU
৩৩	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ৩০: ব্যবসায়িক ড্যাশবোর্ড, কীভাবে 'বিজনেস ডিসিশন' নেন কোম্পানিগুলো?	https://youtu.be/kRNkZ8Yx5h0
৩৪	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ৩১: ডেটা অ্যানালাইটিক্স / বিজনেস অ্যানালাইটিক্সের পার্থক্য কী?	https://youtu.be/76Nc989r2Ck
৩৫	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ৩২: ডেটার ট্রান্সফরমেশনের আগেই ডেটার ব্যাপারে ধারণা পাওয়া	https://youtu.be/fKZ-RIK7INE
৩৬	নন-প্রোগ্রামারদের জন্য মেন্টরশিপ ৩৩: ডেটা থেকে 'বিজনেস প্রবলেম' অর্থাৎ চিকিৎসায় ক্যানসার শনাক্তকরণ	https://youtu.be/fpFFSYsDee8

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
৩৭	মেন্টরশিপ ৩৪: কেন টাইটানিক জাহাজডুবিতে তৃতীয় শ্রেণির পুরুষ যাত্রীরা বেশি মারা গিয়েছিলেন?	https://youtu.be/O9ThTJRw1JU
৩৮	মেন্টরশিপ ৩৫: কেন টাইটানিক জাহাজডুবিতে তৃতীয় শ্রেণির মহিলা এবং পুরুষ যাত্রীরা বেশি মারা গিয়েছিলেন? প্লট	https://youtu.be/A_mK0xHDf8k
৩৯	মেন্টরশিপ ৩৬: টাইটানিক জাহাজডুবিতে কোন শ্রেণির মহিলা এবং পুরুষ যাত্রীরা বেশি মারা গিয়েছিলেন? প্লটিং	https://youtu.be/IYp3vFtvPoQ
৪০	মেন্টরশিপ ৩৭: সুপারস্টোর ডেটাসেট, বিভিন্ন ডাইমেনশনে ‘রুট কজ অ্যানালাইসিস’ আসলে ভেতরে কী হয়েছিল?	https://youtu.be/Uka9mQQhCHs
৪১	মেন্টরশিপ ৩৮: নন-প্রোগ্রামারদের ৭ লাইনের মেশিন লার্নিং: কে কে বেঁচে গিয়েছিলেন? (আর প্রোগ্রামিং)	https://youtu.be/GcM3W3gjVjc
৪২	মেন্টরশিপ ৩৯: দুটো ডোমেইনের মধ্যে 'মেনি-টু মেনি' রিলেশন ম্যাপিং ভিজ্যুয়ালাইজেশন, স্যানকে চার্ট	https://youtu.be/PpeJWvg4qrU
৪৩	মেন্টরশিপ ৪০: ডেটা থেকে ‘কেন’ প্রশ্ন তৈরি করতে পারা: কারা এবং কেন বিক্রি করতে পারছে বেশি?	https://youtu.be/q4yAmB1NxQQ
৪৪	মেন্টরশিপ ৪১: ৫ সেকেন্ড রুল: ডেটা অ্যানালাইসিস থেকে ভিজ্যুয়ালাইজেশন বোঝার সময়সীমা — ট্যাবলো (ব্রাউজার)	https://youtu.be/w6DKIll5iG4
৪৫	প্রোগ্রামারদের জন্য ডেটা—অ্যানালাইটিক্স ৪২: শুরু করব কোথায়? পাইথন দিয়ে প্রজেক্ট টাইটানিক	https://youtu.be/bRVjpEENeJI
৪৬	নন-প্রোগ্রামারদের জন্য ডেটা-অ্যানালাইটিক্স ৪৩: সুপারস্টোর: সাব-ক্যাটাগরিতে ৪ বছরের 'গ্রোথ' কেমন?	https://youtu.be/wvuvqfXkeN4

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
৪৭	নন-প্রোগ্রামারদের জন্য ডেটা-অ্যানালাইটিক্স ৪৪: ডেটা অ্যানালাইটিক্স নিয়ে বাংলায় বই	https://youtu.be/hSZHGm82tEk
৪৮	নন-প্রোগ্রামারদের জন্য ডেটা-অ্যানালাইটিক্স ৪৫: খালি চোখে প্যাটার্ন দেখে ডেটা অ্যানালাইসিস	https://youtu.be/ynCo03nSd8A
৪৯	নন-প্রোগ্রামারদের জন্য ডেটা-অ্যানালাইটিক্স ৪৬: খালি চোখে প্যাটার্ন দেখে ডেটা অ্যানালাইসিস, পাওয়ার-বিআই	https://youtu.be/YI2EumlQEEs
৫০	অ্যানালাইটিক্স ৪৭: গ্লোবাল সুপারস্টোর ডেটা অ্যানালাইসিস, সমস্যা কোথায়? কেন এর উত্তর এবং কেন? ট্যাবলো	https://youtu.be/1u4ggeoH10A
৫১	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস (ট্যাবলো): ডেটা অ্যানালাইসিসের পেছনে কৃত্রিম বুদ্ধিমত্তা	https://youtu.be/ABVFcHJvEps
৫২	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস (ট্যাবলো): ডেটা থেকে ট্রেন্ড অ্যানালাইসিস/ফোরকাস্টিং	https://youtu.be/mbMajdygrXw
৫৩	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস (ট্যাবলো): দ্বিতীয় ক্লাস - বিজনেস সমস্যার ভিজ্যুয়ালাইজেশন	https://youtu.be/zhk_w9WUxPY
৫৪	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস (ট্যাবলো): ডেটা অ্যানালাইসিসের ডাইমেনশন এবং মেজার	https://youtu.be/xcooaCbZV_c
৫৫	নন-প্রোগ্রামারদের অ্যানালাইসিস (ট্যাবলো): ডেটা থেকে ট্রেন্ড অ্যানালাইসিস/ফোরকাস্টিং (তৃতীয় অনলাইন ক্লাস)	https://youtu.be/NlRNgHipHkI
৫৬	ডেটা অ্যানালাইসিস - চতুর্থ অনলাইন ক্লাস: ডিসিশন ট্রি / মাইক্রোসফট পাওয়ার 'বি-আই' দিয়ে ডিকম্পোজিশন ট্রি	https://youtu.be/WZcoV8ZfJjw

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
৫৭	ডেটা অ্যানালাইসিস - পঞ্চম অনলাইন ক্লাস: ডেটার প্রোফাইল, ডিস্ট্রিবিউশন/ পাওয়ার 'বি-আই' ইনস্যুরেন্স ক্লেইম	https://youtu.be/iGuq_vqdygY
৫৮	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস: ওপেন সোর্স টুল দিয়ে বিজনেস অ্যানালাইটিক্স	https://youtu.be/3fHOlKNHmgg
৫৯	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস: ওপেন সোর্স টুল দিয়ে বিজনেস অ্যানালাইটিক্স (আরও কিছু টুল)	https://youtu.be/8BAOLlVTg7o
৬০	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস: কোভিড-১৯-এর সত্যিকারের ডেটাসেট নিয়ে কিছু কাজ (ট্যাবলো)	https://youtu.be/i2C4YEiaFz0
৬১	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস: হাউস প্রাইস প্রেডিকশন ডেটাসেট	https://youtu.be/fiAsC--qc1M
৬২	ডেটা অ্যানালাইটিক্স: ডেটা অ্যানালাইটিক্স এবং ফোরকাস্টিং: Tableau Public	https://youtu.be/CkFYWllnP1g
৬৩	ডেটা অ্যানালাইটিক্স: ট্যাবলো পাবলিক (ফ্রি) দিয়ে ডেটা অ্যানালাইসিস এবং ফোরকাস্টিং (২) Tableau Public	https://youtu.be/QdGlag2k9LQ
৬৪	ডেটা লিটারেসি জিনিসটা কী: এক্সপ্লোর, আন্ডারস্ট্যান্ড এবং কমিউনিকেট করতে পারা	https://youtu.be/i25dwPY38PU
৬৫	ডেটা নিয়ে কাজ করার আগ্রহ/ অ্যাটিচ্যুড - শেখার আগ্রহ আছে তো?	https://youtu.be/DJdvk0kkJss
৬৬	ডিজাইন থিঙ্কিং: ডেটা থেকে 'ইনসাইট'কে কীভাবে কমিউনিকেট করব?	https://youtu.be/uOR9Zwn8qhg
৬৭	বেসিক ডেটা ভিজ্যুয়ালাইজেশন: গুগল শিট/এক্সেল	https://youtu.be/7rj_6IRQd8I

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
৬৮	মাইক্রোসফট এক্সেল দিয়ে ডেটা অ্যানালাইসিস: পিভট টেবিল (টাইটানিক ডেটাসেট)	https://youtu.be/35KJJCpOrrw
৬৯	‘নন-টেকনিক্যাল’ মেশিন লার্নিং এবং লিনিয়ার ‘অ্যালজেব্রা’ লাইভ	https://www.youtube.com/live/2pyWKSSFa2E?feature=share
৭০	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস: ওপেন সোর্স টুল দিয়ে বিজনেস অ্যানালাইটিক্স	https://www.youtube.com/live/Hls2XAWAo-A?feature=share
৭১	নন-প্রোগ্রামারদের জন্য ডেটা অ্যানালাইসিস: ট্যাবলো, একদম বেসিক Tableau — কোথা থেকে শুরু করা যায়?	https://www.youtube.com/live/guMOoW2ER5c?feature=share
৭২	ট্যাবলো দিয়ে আইরিস প্রজাতির ক্লাস্টার অ্যানালাইসিস	https://www.youtube.com/live/TxKNd9H18w8?feature=share
৭৩	ট্যাবলো কেন? কেন ট্যাবলো শেখা সহজ? Tableau - Why is it easier than you think?	https://youtu.be/hH9ZxC3FVQI
৭৪	ডেটা অ্যানালাইটিক্স: ট্যাবলো দিয়ে গ্লোবাল সুপারস্টোরের ধারণা Tableau Global Superstore Dataset	https://youtu.be/Ko90p_ORvdE
৭৫	ডেটা অ্যানালাইটিক্স: মাইক্রোসফট পাওয়ার বিআই (টাইটানিক ডেটাসেট) Power BI	https://youtu.be/JoO6v_LF9Hw
৭৬	ডেটা অ্যানালাইটিক্স: ৩ সার্টিফিকেশন নিয়ে আলাপ, গুগল, আইবিএম, মাইক্রোসফট: ওয়েবিনার	https://youtu.be/UY-zgwR6K6A
৭৭	ডেটা অ্যানালিস্টের কাজ কী? (মাইক্রোসফট সার্টিফাইড ডেটা অ্যানালাইটিক্স অ্যাসোসিয়েট)	https://youtu.be/j3X5qPDEsZA
৭৮	‘পাওয়ার বিআই’-এর বিল্ডিং ব্লক: (মাইক্রোসফট সার্টিফাইড ডেটা অ্যানালাইটিক্স অ্যাসোসিয়েট) Power BI	https://youtu.be/CLX1BS-WxZw
৭৯	ডেটা সায়েন্সের ‘হায়ারার্কি অব নিডস’:: একদম শুরুর গল্প!	https://youtu.be/kTYcFCnyq3c

ক্রম	ভিডিও টাইটেল	ভিডিও লিংক
৮০	ডেটা সায়েন্স/অ্যানালাইটিক্স কীভাবে শুরু করবেন? ডেটা থেকে প্রজ্ঞা নেবার উপায় কী?	https://youtu.be/BlzHgnVe1mE
৮১	বেসিক ধারণা নিয়ে আলাপ: ডেটা-অ্যানালাইসিস জিনিসটা কী? Data Analytics	https://youtu.be/dLB0hIGy7bA
৮২	ডেটা অ্যানালাইটিক্স: সিকুয়েল/SQL নিয়ে ক্লাস	https://youtu.be/HPu7l_lqaLU
৮৩	ডেটা নিয়ে কাজ করার স্টার্টিং পয়েন্ট কোথায়? | How to Start with Data?	https://youtu.be/BrSvNlaefuw
৮৪	ফার্স্ট প্রিন্সিপালস থিঙ্কিং | কেন 'বেসিক' জানা এত এত প্রয়োজনীয়? | Data Science, back to basics?	https://youtu.be/o2yOPqu_FbU
৮৫	প্রিয় ওপেনসোর্স অল্টারনেটিভ: ডেটা সায়েন্স/ডেটা অ্যানালাইটিক্স টুল	https://youtu.be/8HyQLeFV8R0
৮৬	নন-প্রোগ্রামারদের জন্য কম্পিউটেশনাল থিঙ্কিং ধারণা | Computational Thinking for Problem Solving	https://youtu.be/uZhxfGgCVkE
৮৭	টেকনোলজি এবং সার্ভিস ইন্ডাস্ট্রি: রুট কজ অ্যানালাইসিস, সমস্যার ভেতরে ঢোকা | Root Cause Analysis RCA	https://youtu.be/0jElnqwWXGs

ভিডিও যোগ হচ্ছে প্রতিনিয়ত।

ওপেনসোর্সে হাজারো ডেটা অ্যানালাইটিক্স টুল

ডেটা অ্যানালাইসিসের হাতেখড়ি হয় আমার ওপেনসোর্স টুল দিয়ে। এর শুরুটা 'আর স্টুডিও'র সাথে 'জি জি প্লট' ব্যবহার করে অসাধারণ কিছু গ্রাফ তৈরি করলেও সেটার একটা বড় অংশ চলে আসে পাইথনে। পাইথন লাইব্রেরি হিসেবে 'ম্যাট প্লট লিব' শুরুতে কাজ করলেও পরে সিবর্ণ অসাধারণ কিছু কাজ দেখায়। সেই রাস্তা ধরে প্রায় ১০ বছর পর হাতে পাই 'প্লটলি এক্সপ্রেস', যা দিয়ে একসময় প্রচুর কাজ করেছি ছোটখাটো প্রজেক্টে। এখন সেটার কাজ দেখছি সরকারি, বেসরকারি বড় বড় প্রজেক্টে।

প্লটলি এক্সপ্রেস

আমরা ধারণা করছি, আমরা 'শূন্য থেকে পাইথন মেশিন লার্নিং' বইটা পড়েছি। সেই বইটা থেকে বেসিক পাইথন ধারণা থাকলে এই অংশটা আমরা ঠিকমতো বুঝতে পারব। বেসিক পাইথন অন্য সোর্স থেকেও শেখা যেতে পারে।

এখানে ছোট ছোট করে কিছু উদাহরণ দেখাব, যা আমাদের পুরো ব্যাপারটাকে সহজ করে দেবে। এই পুরো জিনিসটাই আপনি পাবেন গিটহাব লিংকে, তবে সেটাকে গুগল কোলাবে খুললে ছবিসহ সরাসরি দেখতে পারবেন। এতে আপনি প্যারামিটারগুলোকে পাল্টে নতুন করে রান করে দেখার সময় ডেটাসেট পাল্টে নিতে পারেন। কীভাবে পাইথন লাইব্রেরি ডেটা সায়েন্স এবং অ্যানালাইটিক্সের ব্যবহার হয়, সেটার একটা ভালো ধারণা পাবেন 'শূন্য থেকে পাইথন মেশিন লার্নিং' বইটা থেকে। পাইথনে বিভিন্ন লাইব্রেরি কীভাবে ইমপোর্ট করতে হয় সেটাসহ আরও কিছু একটা ভালো ধারণা পাবেন সেই বইটা থেকে।

শুরুতেই প্লটলি এক্সপ্রেসকে ইমপোর্ট করে নিয়ে আসি 'পিএক্স' হিসেবে। 'প্লটলি'-এর একটা বড় শক্তি হচ্ছে, 'ইন্টারঅ্যাকটিভ' গ্রাফ তৈরিতে। 'সিবর্ন' দিয়ে চমৎকার গ্রাফ করা গেলেও যেগুলো অনেকটাই স্ট্যাটিক এবং ইন্টারঅ্যাকটিভ নয়। শুরুতে এটার কিছু স্ট্যাটিক গ্রাফ তৈরির পর এটাকে আমরা ব্যবহার করব স্ক্যাটার অ্যানিমেশন তৈরি করার জন্য। আমরা যখন বিভিন্ন ভ্যালু একটার সাথে আরেকটা সংযোগ করব, তখন সেটার স্ক্যাটার প্লট এখানে তৈরি হবে। নিজের হাতে প্র্যাকটিস করতে লিংকটি খুলুন গুগল কোলাবে।

```
https://colab.research.google.com/github/raqueeb/data_analyst/blob/main/data_analyst_plotly_express-update.ipynb
```

শুরুতে লোড করে নিই প্লটলি এক্সপ্রেস লাইব্রেরি।

```
import plotly.express as px
```

ওপেনসোর্সে আইরিস ডেটাসেট

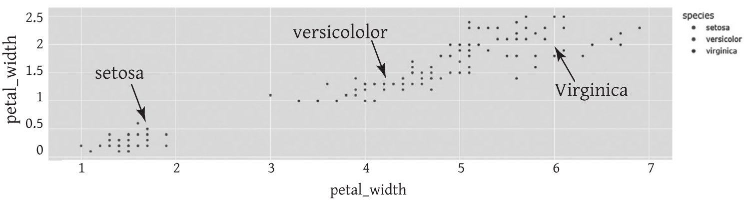
প্রথম উদাহরণটা দেখি, আইরিস ডেটাসেট নিয়ে। আমাদের প্লটলি এক্সপ্রেসের মধ্যে 'ডেটা.আইরিস'-এর মধ্যে আইরিস ডেটাসেট আছে। ফলে, আগে থেকে ডাউনলোড করতে হচ্ছে না। আমরা জানি যে প্রতিটা লাইব্রেরির ভেতরে বেসিক ডেটাসেটগুলো থাকে যাতে আমরা সহজেই শিখতে পারি। প্লটলিকে বেছে নেওয়ার সবচেয়ে বড় উদ্দেশ্য হচ্ছে এটা একটা ইন্টারঅ্যাকটিভ ভিজ্যুয়ালাইজেশন, অর্থাৎ স্ট্যাটিক ভিজ্যুয়ালাইজেশন গ্রাফের ওপর ক্লিক করলে অতিরিক্ত তথ্য পাওয়া যায় না, সেখানে ইন্টারঅ্যাক্টিভ ভিজ্যুয়ালাইজেশন মধ্যে দরকারি জায়গায় ক্লিক করলে সেটার ওপরে আরও অন্যান্য নতুন ব্লক অথবা গ্রাফ এবং বিভিন্ন তথ্যের ওপর মাউস 'হোভার' করলে সেখানে বাড়তি তথ্য চলে আসে। এখানে অনেক সময় 'ড্যাস' ওয়েব অ্যাপ্লিকেশন ফ্রেমওয়ার্ক ব্যবহার করে এর দক্ষতা আরও বাড়ানো যায়।

আমাদের এখানে ডেটাফ্রেম হচ্ছে 'ডিএফ', (আমরা অন্য নাম দিতে পারতাম) যেখানে ডেটাফ্রেম লোড হয়ে স্ক্যাটার প্লটের জন্য তৈরি হচ্ছে।

দুই ডাইমেনশনে কোন কোন ফিচার ব্যবহার হবে সেটা বলে দিচ্ছি এখানে, অর্থাৎ এক্স এবং ওয়াই অ্যাক্সিসে এবং সেটাকে ৩ রঙে ভাগ করে দেবে 'স্পিসিস' অংশে। এখানে বিভিন্ন অংশে কোথায় কোন আর্গুমেন্ট বসছে, সেটা ঠিকমতো বুঝতে পারলে আপনি নিজে থেকেই নতুন নতুন ফিচার নাম বসিয়ে অন্যান্য ডেটা সম্পর্কে তথ্য পাওয়া যাবে।

```
df = px.data.iris()

[ ]
fig = px.scatter(df,
                 x="petal_length",
                 y="petal_width",
                 color="species")
fig.show()
```

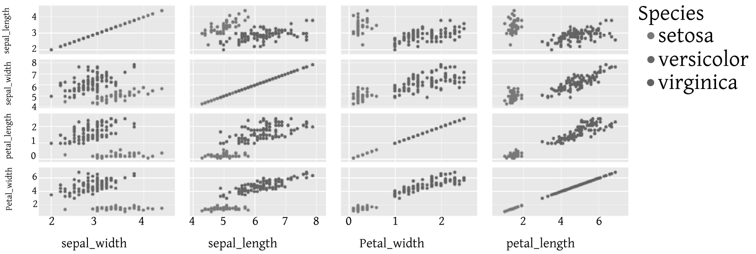


চিত্র ৮৯: আইরিস ডেটার প্লটিং

এখানে স্ক্যাটার প্লট ম্যাট্রিক্স ডেটা ফ্রেমকে চারটা ডাইমেনশনে ফেলে বিভিন্ন রং দিয়ে প্রজাতিকে আলাদা করেছে। px.scatter_matrix প্লটলি এক্সপ্রেসের একটা ফাংশন। এর কাজ হচ্ছে ডেটা ফিল্ডের কলামগুলোকে স্ক্যাটার ম্যাট্রিক্সে ফেলা। 'বাই ডিফল্ট' সে সবগুলো কলামকে কনসিডার করবে।

```
fig = px.scatter_matrix(df,

                        dimensions=[‘sepal_width", ‘sepal_length",
‘petal_width", ‘petal_length"],
                        color="species")
fig.show()
```



চিত্র ৯০: তিনটি প্রজাতির আলাদা প্লটিং

নতুন উদাহরণ।

গ্যাপমাইন্ডার ডেটাসেট

এবার আরেকটা ডেটাসেট, বিশ্বখ্যাত গ্যাপমাইন্ডার ডেটাসেট। পাঠিয়ে দিলাম নতুন ডেটা ফ্রেমে। এখানে 'ডিএফ১', দেয়া যেত, তবে আমরা আগের ডেটাগুলোকে রিপ্লেস করে ফেলছি একই ডেটাফ্রেম দিয়ে।

```
df = px.data.gapminder()
```

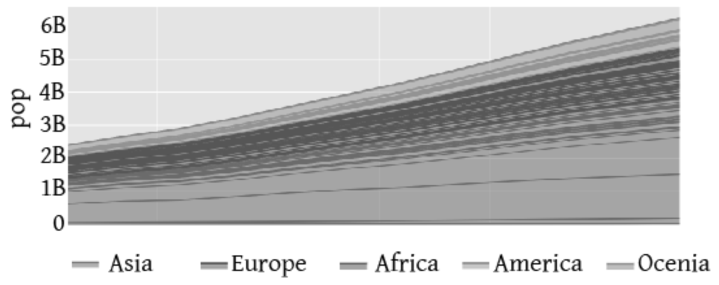
সবশেষের কয়েকটা রেকর্ড দেখি, ধরুন– শেষের পাঁচটা!

```
df.tail()
```

	country	continent	year	lifeExp	pop	gdpPercap	iso_alpha	iso_num
1699	Zimbabwe	Africa	1987	62.351	9216418	706.157306	ZWE	716
1700	Zimbabwe	Africa	1992	60.377	10704340	693.420786	ZWE	716
1701	Zimbabwe	Africa	1997	46.809	11404948	792.449960	ZWE	716
1702	Zimbabwe	Africa	2002	39.989	11926563	672.038623	ZWE	716
1703	Zimbabwe	Africa	2007	43.487	12311143	469.709298	ZWE	716

এখানে px.area একটা স্ট্যাকেবল এরিয়া চার্ট তৈরি করে, line_group প্যারামিটার দিয়ে বাকিগুলোকে গ্রুপিং করে কলাম ধরে দেখিয়ে দিচ্ছে। এখানে প্রতিটা কালারের ওপর মাউস ধরলেই সেটার কনটেক্সট ডেটা পেয়ে যাবেন।

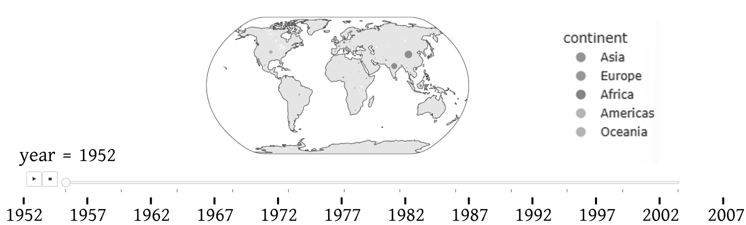
```
fig = px.area(df,
              x="year",
              y="pop",
              color="continent",
              line_group="country")
fig.show()
```



চিত্র ৯১: মহাদেশ ধরে জনসংখ্যা

এই ছবিটা আপনাকে কম্পিউটারে দেখতে হবে, অ্যানিমেশন ধরে ১৯৫২ সাল থেকে ২০০৭ সাল পর্যন্ত প্রতিটা দেশ ধরে ধরে পপুলেশনের অ্যানিমেশন দেখাবে পৃথিবীর প্রজেকশনের ওপর ভিত্তি করে। আগের লিংক থেকে ব্রিজের দেখুন এই অ্যানিমেশন। নিচে, 'প্লে' বাটনটা চাপ দিন বছর ধরে অ্যানিমেশন দেখার জন্য।

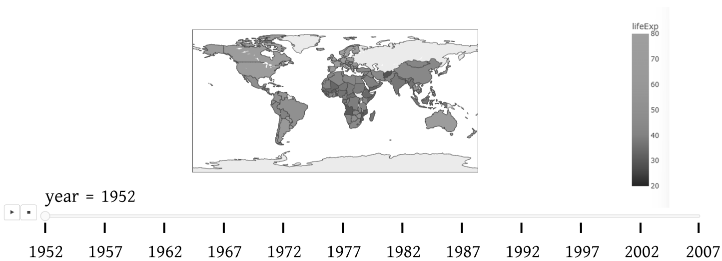
```
fig = px.scatter_geo(df,
                     locations="iso_alpha",
                     color="continent",
                     hover_name="country",
                     size="pop",
                     animation_frame="year",
                     projection="natural earth")
fig.show()
```



চিত্র ৯২: এখানে অ্যানিমেশন ফ্রেম বছর ধরে

আরেকটা অ্যানিমেটেড গ্রাফ, এভাবে দেখতে থাকুন আরো আরো গ্রাফ। এখানে কালার দিয়ে লাইফ এক্সপেক্টেন্সি দেখানো হয়েছে। নিচে ‘প্লে’ বাটন চাপ দিয়ে দেখুন কীভাবে ‘কালার’ পাল্টায়। পাইথন এ ব্যাপারে অসাধারণ।

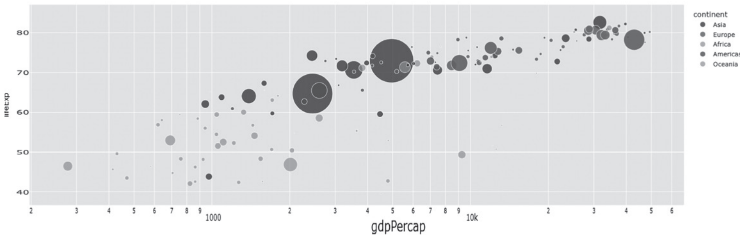
```
fig = px.choropleth(df,
                    locations="iso_alpha",
                    color="lifeExp",
                    hover_name="country",
                    animation_frame="year",
                    range_color=[20,80])
fig.show()
```



চিত্র ৯৩: লাইফ এক্সপেক্টেন্সি, কালারে

এখানে ২০০৭ সালকে কোয়েরি হিসেবে ধরে আমরা একটা ফিগার জেনারেট করছি যেটার কন্টিনেন্ট বিভিন্ন কালারে দেখাবে। এখানে মাউস হোভার করলে কান্ট্রি নাম হিসেবে দেখাবে।

```
fig = px.scatter(df.query("year==2007"),
                 x="gdpPercap",
                 y="lifeExp",
                 size="pop",
                 color="continent",
                 hover_name="country",
                 log_x=True,
                 size_max=60)
fig.show()
```



চিত্র ৯৪: জিডিপি পার ক্যাপিটা হিসেবে একটা অ্যানিমেশন

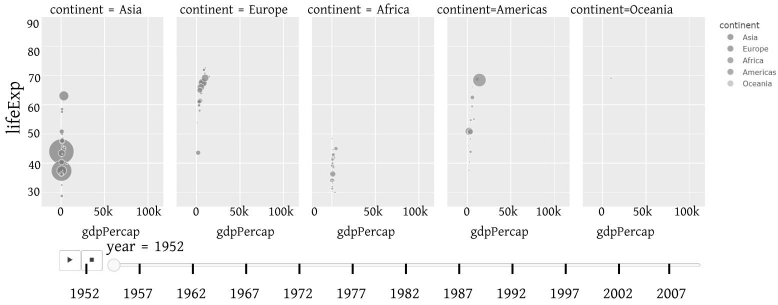
নিচের ফিগারে অ্যানিমেশন ফ্রেমে কান্ট্রি দেখাবে, যেটা অ্যানিমেশন গ্রুপ হিসেবে কান্ট্রিকে ব্লক করা হয়েছে। এখানেও মাউস হোভার কান্ট্রি হিসেবে কাজ করবে।

```
fig = px.scatter(df,
                 x="gdpPercap",
                 y="lifeExp",
                 size="pop",
                 color="continent",
                 animation_frame ="year",
                 animation_group ="country",
                 hover_name="country",
                 log_x=True,
                 size_max=60)
fig.show()
```

চিত্র ৯৫: ১৯৫২ থেকে ২০০৭ সাল পর্যন্ত একটানা অ্যানিমেশন

নিচের ছবিতে কন্টিনেন্টগুলোকে আলাদা আলাদা গ্রুপ করে বছরভিত্তিক অ্যানিমেশনের জন্য আলাদা স্ক্যাটার প্লট ব্যবহার করা হয়েছে।

```
fig = px.scatter(df,
     x ="gdpPercap",
     y ="lifeExp",
     animation_frame ="year",
     animation_group ="country",
     size ="pop",
     color ="continent",
     hover_name ="country",
     facet_col ="continent",
     size_max = 45,
     range_y =[25, 90]
)
fig.show()
```

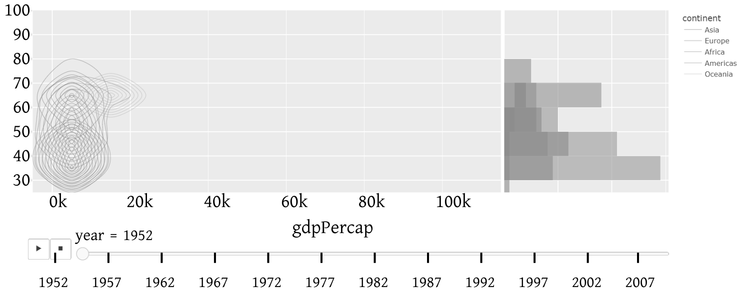


চিত্র ৯৬: বিভিন্ন কন্টিনেন্ট ধরে আলাদা আলাদা গ্রুপ করে বছরভিত্তিক অ্যানিমেশন

এ রকম আরও অনেকগুলো উদাহরণ নিচে দেয়া হলো— যেগুলো ব্যবহার করে আপনারা বিভিন্ন প্যারামিটার নিজেদের মতো করে পাল্টে রান করাতে পারবেন। কাগজের সীমাবদ্ধতার কথা লক্ষ করে বেশ কিছু প্লটিং এখান থেকে আমরা কমিয়ে দিয়েছি, যা ওপরের লিংক থেকে পুরোপুরি পাবেন।

```
fig = px.density_contour(df,
                         x ="gdpPercap",
                         y ="lifeExp",
```

```
                                   color ="continent",
                                   marginal_y ="histogram",
                                   animation_frame ="year",
                                   animation_group ="country",
                                   range_y =[25, 100])
fig.show()
```

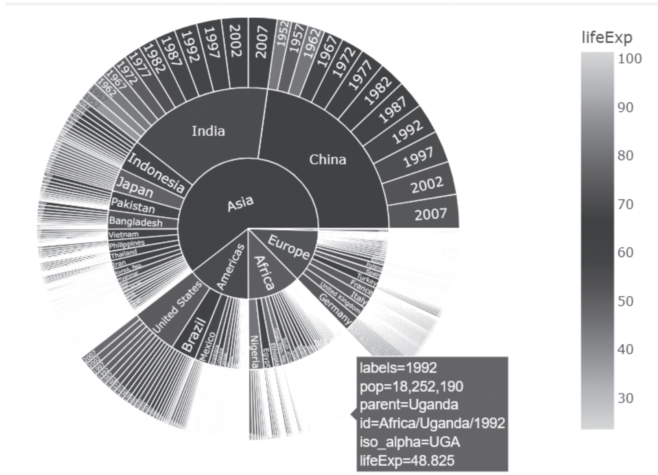


চিত্র ৯৭: প্লটের পাশাপাশি একটা হিস্টোগ্রাম জেনারেট করা হয়েছে, যা অ্যানিমেশনের সাথে পরিবর্তন হবে

আপনার ব্রাউজারে যদি ঠিকমতো প্লটলি ছবি জেনারেট না করে তাহলে প্লটলিকে আপগ্রেড করে নিতে পারেন।

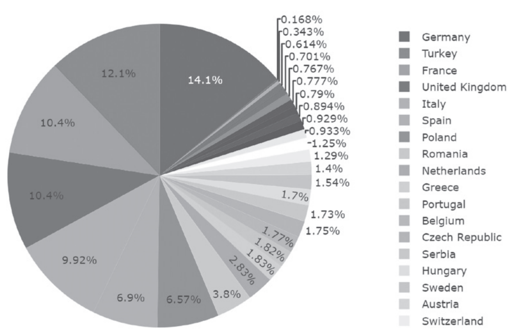
```
# !pip install --upgrade plotly
```

```
fig = px.sunburst(df, path=['continent', 'country', 'year'],
values='pop',
                          color='lifeExp', hover_data=['iso_alpha'],
                          color_continuous_scale='twilight',
                          color_continuous_midpoint=np.
average(df['lifeExp'], weights=df['pop']))
fig.show()
```



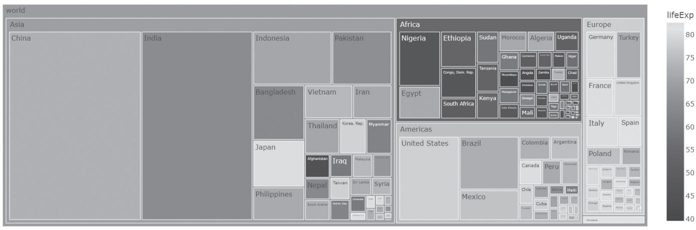
চিত্র ৯৮: এখানে একটা সানবার্স্ট প্লট করা হয়েছে যেখানে কন্টিনেন্ট, কান্ট্রি হায়ারার্কি হিসেবে এসেছে

```
df = px.data.gapminder().query("year == 2007").query("continent
== 'Europe'")
df.loc[df['pop'] < 2.e6, 'country'] = 'Other countries' # Represent
only large countries
fig = px.pie(df, values='pop', names='country', title='Population
of European continent')
fig.show()
```



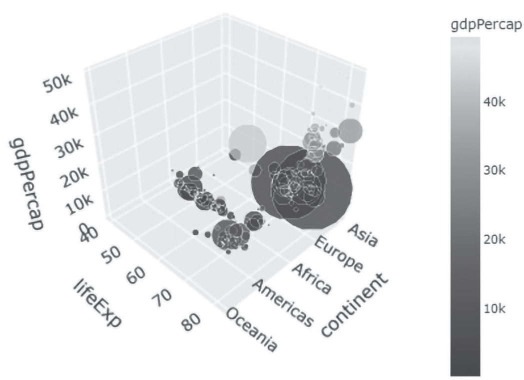
চিত্র ৯৯: এটা একটা পাই চার্ট যেখানে ইউরোপিয়ান কন্টিনেন্টকে দেখানো হয়েছে

```
fig = px.treemap(df, path=[px.Constant('world'), 'continent', 'country'], values='pop', color='lifeExp', hover_data=['iso_alpha'])
fig.show()
```



চিত্র ১০০: ওপরের ট্রি ম্যাপে পৃথিবী, কন্টিনেন্ট, কান্ট্রি স্টেপ বাই স্টেপ এসেছে (মাউস হোভারে কন্টেক্সসহ)

```
fig = px.scatter_3d(df, z="gdpPercap", y="lifeExp", x='continent', color="gdpPercap", size='pop', hover_name="country", size_max=100)
fig.show()
```

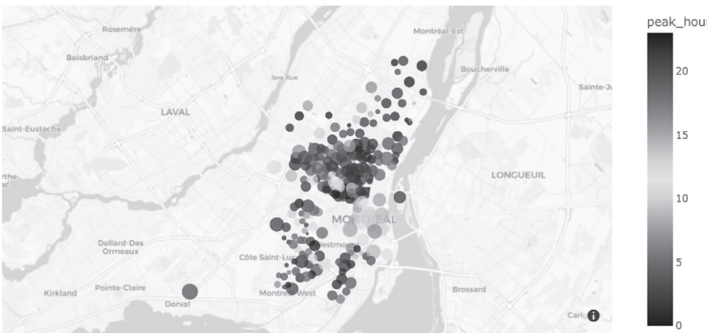


চিত্র ১০১: থ্রিডি ম্যাপে ‘এক্স’ এবং ‘ওয়াই’ অ্যাক্সিসে লাইফ এক্সপেক্টেন্সিসহ স্ক্যাটার প্লট

জিও লোকেশন ডেটাসেট

এতক্ষণ আমরা আলাপ করেছি গ্যাপমাইন্ডার ডেটাসেট নিয়ে। জিও লোকেশন ডেটা নিয়ে কার শেয়ার ডেটাসেট থেকে একটা ম্যাপ বক্স প্লট।

```
df = px.data.carshare()
fig = px.scatter_mapbox(df, lat="centroid_lat", lon="centroid_
lon", color="peak_hour", size="car_hours",
                  color_continuous_scale=px.colors.cyclical.IceFire,
size_max=15, zoom=10,
                  mapbox_style="carto-positron")
fig.show()
```



চিত্র ১০২: কার শেয়ার ডেটাসেট থেকে জিও লোকেশন ডেটাসেট প্লটিং

রিয়েল ওয়ার্ল্ড ওপেনসোর্স প্রজেক্টের জন্য টুল

রিয়েল ওয়ার্ল্ড প্রজেক্টে আমার প্রিয় ডেটা অ্যানালাইসিস টুলসগুলো নিয়ে একটু আলাপ করা যাক। এগুলো দুর্দান্তভাবে চলছে পৃথিবীর সব হাইটেক কোম্পানিগুলোতে। আমিও ব্যবহার করেছি বিশাল বিশাল প্রজেক্টে। কারণ ওপেনসোর্স ব্যবহার করে যেকোনো প্রজেক্টে খুব সহজেই দরকারি কম্পোনেন্টগুলোকে যোগ করে পুরো 'এন্ড টু এন্ড' পাইপলাইন তৈরি করা যায় সহজে। এর শুরুতেই থাকবে 'গ্রাফানা', যেটাকে আমি ব্যবহার করছি অনেক বছর ধরে। ডেটা সেন্টার থেকে শুরু করে নেটওয়ার্কের সব ভিউ আমি দেখতাম এই টুলে বসে।

গ্রাফানা: সব কাজের কাজি

গ্রাফ আনা হচ্ছে সে ধরনের একটা ওপেনসোর্স ডেটা অ্যানালাইটিক্স প্ল্যাটফর্ম যেটা আমাদের একটা জায়গায় বসে বিভিন্ন ম্যাট্রিক্স (শুধু মনের শান্তির জন্য নয়) তার থ্রেশহোল্ড লেভেলে কোথায় কী সমস্যা হচ্ছে, সেটাও জানাতে পারে খুব সহজে। এর পাশাপাশি, আমরা যেভাবে অ্যালার্ম সেট করব, সেভাবে বিভিন্ন ইভেন্ট ধরে real-time ইনসাইট দিতে পারবে দরকারি ভিজ্যুয়ালাইজেশনের পাশাপাশি। আমার দেখা মতে, এ ধরনের সফটওয়্যার সাধারণত 'ডেভঅপস' ইঞ্জিনিয়াররা ব্যবহার করেন তাদের সিস্টেমগুলোকে ঠিকমতো মনিটর করতে, যাতে সেই অ্যানালাইটিক্স থেকে বিভিন্ন ধরনের ম্যাট্রিক্স অর্থাৎ কোন কোন ডেটাগুলোকে পুল করলে একটা কাস্টমাইজেবল ডেটা ড্যাশবোর্ড বানানো যায়, সেটার ভালো ধারণা দেয়।

গ্রাফানা দিয়ে আমরা ডেটার 'জিও ম্যাপ' হিটম্যাপ হিস্টোগ্রাম থেকে শুরু করে সব ধরনের ধারণা দেয় যে, কীভাবে ডেটাকে বোঝা যায়। এখানে

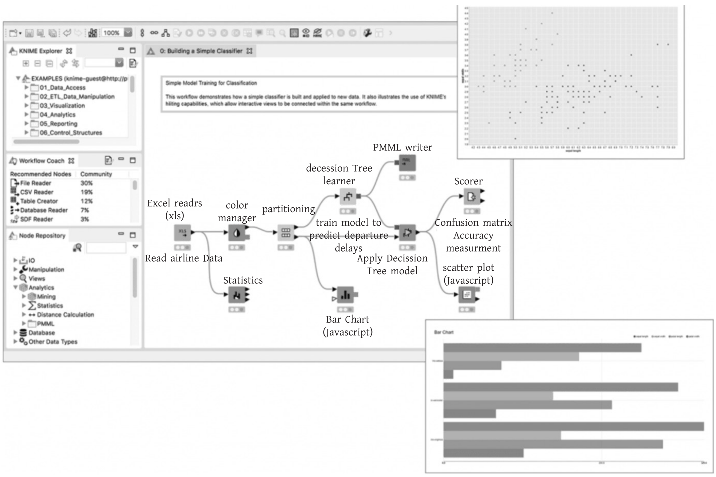
কোন ডেটার সাথে কোন ডেটা গেলে ভালো ‘কন্টেক্সট’ আসবে এবং কোন ডেটা কার সাথে রিলেটেড না সে ধরনের ধারণা আমি দেখেছি কাজ করতে গিয়ে। এই প্ল্যাটফর্ম নিয়ে কাজ করতে গিয়ে শত শত ‘প্লাগিন’ এবং ড্যাশবোর্ড পেয়েছি যেগুলো আছে তাদের অফিশিয়াল লাইব্রেরিতে। এর পাশাপাশি ‘ট্রান্সফার’ না ৫০টির বেশি ওপেনসোর্স এবং কমার্শিয়াল ডেটা সোর্স থেকে ডেটা ‘কানেক্ট’ করতে পারে যেখানেই আমরা সেটাকে রাখি (‘হোস্ট’) না কেন। এর পাশাপাশি এর একটা বিল্ট-ইন গ্রাফানা ‘কোয়েরি পার্সার’ আছে, যেটা যেকোনো ধরনের ডেটাকে সহজে রিড এবং ‘এডিট’ করতে পারে আমাদের কাজের স্বার্থে।

চিত্র ১০৩: গ্রাফানা দিয়ে ভিজ্যুয়ালাইজেশন, গ্রাফানা আমার ব্যবহৃত সবচেয়ে বড় প্ল্যাটফর্ম

নাঈম— ‘KNIME’: এন্ড টু এন্ড ডেটা সায়েন্স

‘এন্ড টু এন্ড’ ডেটা সায়েন্স প্যাকেজ হচ্ছে ‘নাঈম’।

তবে আমার নতুন ভালোবাসা হচ্ছে ‘নাঈম’, যেখানে হাতেকলমে, বুঝে বোঝে নন-টেকনিক্যাল মান ডেটা সায়েন্স প্ল্যাটফর্ম হিসেবে দুর্দান্ত কাজ করছে।



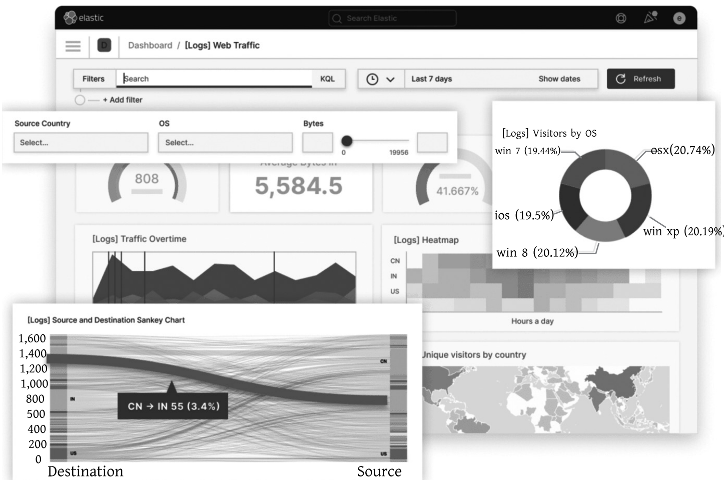
চিত্র ১০৪: নাঈমের প্রসেস ফ্লো

অ্যাপাচি স্পার্ক: লার্জ স্কেল ডেটা প্রসেসিং টুল

আমার দেখামতে, অ্যাপাচি স্পার্ক ডিস্ট্রিবিউটেড large-scale ডেটা প্রসেসিংয়ের জন্য অসাধারণ প্ল্যাটফর্ম। আমি নিজে এটা নিয়ে বেশ কাজ করেছি, যেখানে 'ওয়ার্কফ্লো' হিসেবে এর মাঝে অন্যান্য সল্যুশন রেখেছিলাম কাজের স্বার্থে। বড় বড় ক্লাস্টার ম্যানেজার দিয়ে ডেটার প্রসেসকে অটোমেটিক করা যায় এ ধরনের সল্যুশন দিয়ে— যেখানে ফল্ট টলারেন্স এবং ডেটা হারানোর কোনো ভয় থাকে না। স্পার্ককে ব্যবহার করা যায় আর, জাভা, পাইথন স্ক্যালা এসকিউএল ইত্যাদি দিয়ে যেটাকে ইন্টিগ্রেট করা যায় যেকোনো ধরনের মেইনস্ট্রিম বিগ ডেটা ওয়ার্কফ্লোর মধ্যে। তবে এর সিকিউরিটি দেবার জন্য আলাদা সল্যুশন ব্যবহার করতাম আমি, কারণ এর ব্যাকওয়ার্ড কম্পটিবিলিটি থেকে শুরু করে অন্যান্য সমস্যা, বিভিন্ন 'অ্যাটাক' ইত্যাদি এর মধ্যে ঝামেলায় ফেলে দিতে পারে এই সল্যুশন। এর সাপোর্টের জন্য ওপেনসোর্স কমিউনিটি এবং ডকুমেন্টেশন বেশ ভালো। যদিও ইন মেমোরি প্রসেসিংয়ের জন্য হার্ডওয়্যারের মেমোরি দরকার হয় প্রচুর।

কিবানা: ইলাস্টিক সার্চ ডেটাকে দেখতে

ইলাস্টিক সার্চ ডেটাকে ঠিকমতো নেভিগেট এবং ভিজ্যুয়ালাইজ করতে কিবানা একটা অসাধারণ ওপেনসোর্স ইন্টারফেস দিয়ে থাকে। এর মাধ্যমে আমাদের অ্যাপ্লিকেশনের মধ্যে বিভিন্ন কোয়েরি এবং লোডকে ঠিকমতো বুঝতে কিবানা অসাধারণ কাজ করে। মেশিন জেনারেটেড হাজারো লগ ফাইল থেকে হিস্টোগ্রাম, লাইন গ্রাফ, পাই চার্ট, সানবার্স্ট ইত্যাদি ইত্যাদি ভিজ্যুয়ালাইজেশন আনতে কিবানার জুড়ি নেই। এককথায় আমাদের ডেটা ল্যান্ডস্কেপকে সার্ভে করার জন্য এটা একটা ভয়ংকর টুল বটে।



চিত্র ১০৫: কিবানা, ইলাস্টিক সার্চে ডেটা দেখার জন্য

ডিথ্রি: অসাধারণ ভিজ্যুয়ালাইজেশন জাভায়

আমার প্রিয় ভিজ্যুয়ালাইজেশন লাইব্রেরি জাভাভিত্তিক। ডিথ্রি.জেএস হচ্ছে একটা জাভাস্ক্রিপ্ট লাইব্রেরি যেটা ডেটার ওপর ভিত্তি করে বিভিন্ন ডকুমেন্টকে ম্যানিপুলেট করতে পারে। আমরা দেখেছি ডিথ্রি যেকোনো ডেটাকে এইচটিএমএল, এসভিজি, সিএসএস ফরম্যাটে খুব সহজেই আনতে পারে। ডিথ্রির সবচেয়ে বড় শক্তি হচ্ছে এটা মডার্ন ব্রাউজারের ক্যাপাবিলিটিকে মাথায় রেখে ওয়েব স্ট্যান্ডার্ডকে ঠিকমতো ফলো করে অসাধারণ গ্রাফিকস জেনারেট করে দিতে পারে।

অনলাইনে বিভিন্ন সার্টিফিকেশন

এ যুগের প্রশিক্ষণ, ডেটা ড্রিভেন সমাজের জন্য

Recipes tell you nothing. Learning techniques is the key.

— Tom Colicchio

পেশাগত কারণে প্রচুর প্রশিক্ষণের (৭৫+) মুখোমুখি হতে হয়েছে আমাকে। প্রযুক্তিগত উৎকর্ষের একটা অংশে অনলাইন প্ল্যাটফর্মগুলো (লার্নিং ম্যানেজমেন্ট সিস্টেম) নিয়ে প্রচুর রিসার্চ করতে হয়েছে এক দশক ধরে। এবারের ঘটনা একদম মাথা খারাপ করার মতো।

সামনে বেশির ভাগ কাজ ডেটার। গুগল বলছে ডেটানির্ভর চার বছরের ডিগ্রির জায়গায় ৬-১২ মাসের নতুন সার্টিফিকেশন তৈরি করে দেবে নতুন জনবল— একুশ শতকের জন্য। গুগল ট্রেন্ড বলছে— মার্কেটে একটা বিশাল গ্যাপ তৈরি হয়েছে এই মহামারির কারণে। ডেটা বোঝা এবং ডেটা থেকে সিদ্ধান্ত নিতে পারা জনবল নিয়ে।

৩ বছর ধরে নিচের কোর্সগুলোকে ফুল স্কেলে অডিটিং করে যা বোঝা গেল, বিশাল (ঠিক ধরেছেন) মানে, বিশাল ইনভেস্টমেন্ট করেছে গুগল, মাইক্রোসফট এবং আইবিএমের মতো কোম্পানিগুলো— এই 'নলেজ' গ্যাপ পূরণে। আমি তাদের কোর্সওয়ার্ক + ভেতরের ম্যাটেরিয়াল দেখে সত্যিকারভাবে অভিভূত। এলাহি কাণ্ড করেছে তারা! একটা জিনিস বুঝতে পারলাম— এই কোর্সগুলো করতে বিশেষ কোনো প্রাতিষ্ঠানিক পড়াশোনার প্রয়োজন নেই শুরুতে। কারণ, আমাদের আশপাশের দেশ ভারত, ফিলিপাইন, ইস্ট ইউরোপের দেশগুলোর ছেলেমেয়েদের মাথায় এমন কী আছে যেটা আমরা জানি না? আমাদের সমস্যা একটাই— ভালো ডেটা কমিউনিটি নেই

আমাদের। অনেকে জানতেও পারেন না, কত মণিমুক্তা ছড়িয়ে আছে এই নেটে। ফলে কেউ কেউ বিষণ্ন হয়ে পড়ে থাকেন টিকটক, ফেসবুকে।

কোর্স	লিংক
গুগল ডেটা অ্যানালাইটিক্স প্রফেশনাল সার্টিফিকেট	https://www.coursera.org/professional-certificates/google-data-analytics
আইবিএম ডেটা অ্যানালাইটিক্স সার্টিফিকেশন	https://www.coursera.org/professional-certificates/ibm-data-analyst
মাইক্রোসফট সার্টিফাইড ডেটা অ্যানালাইটিক্স অ্যাসোসিয়েট	https://docs.microsoft.com/en-us/learn/certifications/data-analyst-associate/

টেবিল ১২: অনলাইনে তিনটা অসাধারণ ডেটা অ্যানালাইটিক্স কোর্স

In a world deluged by irrelevant information, clarity is power. Today having power means knowing what to ignore.

— Yuval Noah Harari

এই ইনফরমেশন ওভারলোডের যুগে আজকের জীবনের বড় কাজ হচ্ছে— যেহেতু সব খাওয়া যাবে না, সেহেতু 'প্রায়োরিটি' সেট করে জীবন যতটুকু 'ব্যালান্স' করা যায়। আর বাকিটা ফেলে দেব ফিল্টারে। এর পাশাপাশি, প্রয়োজনীয় স্কিলসেট না থাকলে আবার বাজারে 'বাতিল' বলে গণ্য হব। আজকের কিওয়ার্ড, মার্কেটে 'রেলেভেন্ট' থাকা। তাহলে, কোনটা রাখব আর কোনটা ফেলব— এই টানাপোড়োনের গল্প নিয়ে এই মেন্টরশিপ।

বেশ কয়েকটা 'অসাধারণ' কাজ হয়েছে গত তিন বছরে। গুগল ডেটা অ্যানালাইটিক্স প্রফেশনাল সার্টিফিকেট নিয়ে বেশ নড়াচড়া পড়েছে বাংলাদেশে। আমি এটাই চেয়েছিলাম, যাতে অনেকে নিজে থেকে শুরু করতে পারেন। সেটা শুরু হয়েছে, তবে অনেকে মাঝপথে এসে সাহায্য চাচ্ছিলেন। এই একই জিনিস আমি দেখেছিলাম আমার সন্তানদের কোর্সটা

করার সময়। কিছু কিছু জিনিস মুখস্থ করে না বুঝেই পরীক্ষায় পাস করে যাচ্ছিল, যেটা আসলে আমাদের 'গোল' হবার কথা নয়।

গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন প্রোগ্রাম 'পাস' করে সার্টিফিকেট পাওয়া খুব সোজা, তবে ভেতরের আসল 'আন্ডারস্ট্যান্ডিং' পেতে বাইরের সাহায্য লাগে। এর পেছনে যা বোঝা গেছে, কোর্সটিকে ছোট করিয়ে আনার জন্য গুগল কিছু কিছু চমৎকার লিংক এবং কিওয়ার্ড ব্যবহার করেছে, যা আলাদাভাবে না জানলে বোঝা কিছুটা দুষ্কর। এর মোদ্দা কথা হচ্ছে, গুগল সার্টিফিকেশন কোর্সওয়ার্ক ভালো; তবে সেটা ভালোভাবে বুঝতে কোর্সওয়ার্কের পাশাপাশি বাইরের নলেজ কিছুটা প্রয়োজন। সেটার জন্য এই বইটা আপনাদের কাজে দেবে।

আমার গল্প: গ্লোবাল সার্টিফিকেশন

আমার প্রথম গ্লোবাল সার্টিফিকেশন আসে ১৯৯৯ সালে। বাংলাদেশে গ্লোবাল সার্টিফিকেশন এক্সাম সেন্টারগুলো না আসায় সিঙ্গাপুর/ভারতে গিয়ে পরীক্ষা দেয়া শুরু করল কেউ কেউ। তক্কে তক্কে বসে রইলাম, কবে আসে বাংলাদেশে। বসে থাকিনি, বরং মক টেস্ট দিচ্ছিলাম নিজের মতো করে। একটু ফুরসত পেলেই প্লেব্যাক করতাম, নিজের মাথায়। ফাইনালি, সেটা এল ১৯৯৯তে, ফুরুৎ করে পরীক্ষা দিলাম ৮টা! নিজের চাকরি থেকে টাকা বাঁচিয়ে। প্লাস, সব তো ডাউনলোডেড মাথায়!

লুকিং ব্যাক, গ্লোবাল সার্টিফিকেশন কেন নিয়েছিলাম আমি? চাকরির জন্য কি? না। বরং নিজেকে যাচাই করতে— গ্লোবাল স্কেলে কোথায় আছি— সেটা বুঝতে। সিম্পলি, একটা ভ্যালিডেশন। ওই গ্লোবাল সার্টিফিকেশন আমাকে যে কনফিডেন্স দিয়েছিল সেই ১৯৯৯ সালে— যা দিতে পারেনি অন্য কিছুই। বুঝলাম কীভাবে বিদেশের চাকরি পায়ে ঠেলে থাকতে হয় দেশে, একটা বড় 'কলিং' নিয়ে। বিদেশে একটা কিউবিকলে থাকার চেয়ে 'টাচিং পিপলস লাইভ' কীভাবে একটা বড় কলিং হতে পারে, সেটা বুঝে গেলাম একটা পর্যায়ে। সেই 'ইনসাইট' এবং কনফিডেন্স আমাকে নিয়ে গেছে জাতীয় পর্যায়ের ডেটা ড্রিভেন সব প্রতিষ্ঠানতে। আমাকে পেছনে তাকাতে হয়নি আর।

আমাদের নতুন প্রজন্মের মধ্যে অনেকে সেই ভ্যালিডেশন চান নিজের কাছে। তবে শুধু চাকরির জন্য স্কিল শেখা, সেটা কখনো টেকেনি, আর টিকবেও না কোনো দিন। আমাদের বর্তমান চাকরির পাশাপাশি যে কাজটা ভালো লাগে, সেটার জন্য স্কিল ডেভেলপমেন্ট অন্যায় কিছু নয়। স্কাই ইজ দ্য লিমিট।

আমার কষ্ট অন্য জায়গায়। আমরা একটা ইউনিভার্সিটি ডিগ্রির জন্য ৫-১০ লাখ টাকা (যারা প্লেসমেন্টের দায়িত্ব নেন না) খরচ করলেও স্কিল ডেভেলপমেন্টে পয়সা খরচ করতে চাই না। ফলে, সেটা যায় আমাদের নিজেদের জীবনের ওপর দিয়ে। ডিগ্রির পরে চাকরির আসল যুদ্ধে নামলে নিজের স্কিল বাজারের সাথে অসামঞ্জস্য হওয়ায় হতাশ হয়ে পড়ি।

গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন: ব্যাকস্টোরি

৬টা ডেটা সায়েন্সের বই লিখতে লিখতে মনে হলো নন-প্রোগ্রামারদের ডেটা বোঝার সহজ উপায় কী? সেই হিসেবে ডেটা অ্যানালাইটিক্সের জন্য একটা এক্সক্লুসিভ বই, এর পাশাপাশি প্রায় এক শ ভিডিও তৈরি করছিলাম ব্যাপারটিকে সহজ করে আনার জন্য। এই উন্মুক্ত, অর্থাৎ ফ্রি জিনিসগুলো নিয়ে কাজ করতে গিয়ে মনে হয়েছে, বিশেষায়িত জিনিসপত্রের আবেদন হারায় যখন এটাকে প্রিমিয়াম মডেলে চালানো হয়। তখন থেকেই আমি অন্য চিন্তাভাবনা শুরু করি। বই এবং ভিডিও সিরিজ।

আমি কোর্সেরার ব্যবহারকারী অনেক অনেক দিন থেকে। এক বছর আগে নিজের ডেটা অ্যানালাইটিক্সের বই লিখতে লিখতে গুগলের অ্যালার্ট পেলাম এই কোর্সের। সেই পার্সপেক্টিভ থেকে আমি নন-প্রোগ্রামারদের জন্য চার মাসের একটি ডেটা অ্যানালাইটিক্স মেন্টরশিপ প্রোগ্রাম ডিজাইন করি যেখানে গ্রাউন্ড-আপ থেকে একজন শিক্ষার্থী কীভাবে ডেটা থেকে ধারণা নিতে হয়, তা শিখতে পারবেন। কারণ, আমি চাই ডেটা সায়েন্স ঠিকমতো বোঝার আগে, বিশেষ করে এই বিশেষায়িত জায়গাটা কেন শুধু পৃথিবীর কয়েক হাজার মানুষের মধ্যে সীমাবদ্ধ থাকবে এবং সেই পাইপলাইনটিকে সহজ করার জন্য শুরুতে ডেটা অ্যানালাইসিস অংশটি হাতেকলমে শুরু করেছিলাম।

আমি শেখাচ্ছি নিজের মতো করে, অনলাইনে— তবে এর পাশাপাশি যদি একটা গ্লোবাল সার্টিফিকেশন কারিকুলাম ঠিকমতো ব্যবহার করা যায় তাহলে একজন শিক্ষার্থী গ্লোবাল কারিকুলামের পাশাপাশি একটি আন্তর্জাতিক মানের সার্টিফিকেট পেতে পারে। এই সার্টিফিকেট নিয়ে এসে মার্কিন যুক্তরাষ্ট্রের কাজ পাবে কি না, সেটাকে বিবেচনায় না এনে তারা আসলে গ্লোবাল লেভেলে নিজেদের যাচাই করতে পারছে কি না, সেটাই মুখ্য উদ্দেশ্য।

আমার রিসার্চ বলে গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন সাধারণত ১৭৪ ঘণ্টা সময় নেয় কমপ্লিট করতে। পরের চ্যাপ্টারে বিস্তারিত লিখেছি ডেটাসহ। এটা কমিয়ে আনা সম্ভব, তবে 'এইড অ্যাপ্লিকেশন' প্রসেস শুরু থেকে শেষ পর্যন্ত বাড়তি সময়ের কারণে ৩ মাসের নিচে আনা কষ্টকর।

এই সার্টিফিকেশনে ৮টি আলাদা আলাদা কোর্স (একটি ক্যাপস্টোন প্রজেক্টসহ) একেকটা করতে (সার্টিফিকেশন পাওয়ার জন্য নয়— আসল শেখা) গড়ে প্রায় ২০ ঘণ্টা লাগবে। পুরো সার্টিফিকেশনে গুগলের নিজেরই ২১ ঘণ্টার ভিডিও আছে— যাকে একটা কোর্স হিসেবে দেখলে প্রতি কোর্স ধরে ১৭৫ মিনিটের ভিডিও দেখতে হবে (মাস্ট)।

এই পুরো জিনিসটা আপনাকে ডাউনলোড করতে হবে আপনার ব্রেইনে ৩২ ঘণ্টায়। তবে একটা সিরিয়াস 'কমিটমেন্ট' ছাড়া নিজেকে 'সার্টিফাই' কষ্টকর। সিরিয়াস লার্নার এবং 'ভবিষ্যৎ দেখার জন্য প্রস্তুত' হবার মানুষ চাই আমি। ডেটা নিয়ে ব্যাকগ্রাউন্ড না থাকলেও সমস্যা নেই, 'অ্যাটিচুড' ঠিক থাকলে সব সম্ভব। এই সার্টিফিকেশন নিয়ে বিস্তারিত হিসেব দেখুন পরের চ্যাপ্টারে।

ইন্ডাস্ট্রি কেন শেখাচ্ছে ডেটা অ্যানালাইটিক্স?

আমরা চাই বা না চাই, আস্তে আস্তে সনাতন ডিগ্রিকে 'রিপ্লেস' করে ফেলবে ইন্ডাস্ট্রি স্কিলসেট। ট্রিলিয়ন ডলারের এডুকেশন ইন্ডাস্ট্রি ভার্সেস ইন্ডাস্ট্রির (যারা চাকরি দেয়) মধ্যে টানাপোড়েন শুরু হয়ে গেছে অনেক আগেই। ইন্ডাস্ট্রিকে যদি একজন ডিগ্রিধারীকে আলাদাভাবে ইন্টার্নশিপ করিয়ে সিস্টেমে ঢোকায়, তাহলে সে কেন নিজে এডুকেশনে ঢুকবে না? গুগল, আইবিএম কেন তৈরি করছে নিজেদের কারিকুলাম? ইনফোসিস কেন তৈরি করেছে নিজের একাডেমি?

আবার, গুগল/আইবিএম/ইনফোসিস যদি প্রডাক্টই তৈরি করে, তাহলে সে কেন ১০ বছরের পুরোনো কারিকুলাম পড়াবে? সে তো চাইবে, সামনে কী আসবে সেটার বেজলাইন ধরে কী পড়ালে তাকে সে সাহায্য করতে পারবে সামনের বছরে। ইন্ডাস্ট্রি তো তার কী দরকার— এই নলেজ ইকোনমিতে।

সেই পার্সপেক্টিভ থেকে গুগল কোর্সেরার সাথে মিলিয়ে একটা চমৎকার এন্ট্রি লেভেলের ডেটা অ্যানালাইটিক্স কোর্স এনেছে, যা দিয়ে গুগল থেকে শুরু করে মার্কিন যুক্তরাষ্ট্রের অনেকগুলো (১৫০+) কোম্পানিতে প্লেসমেন্ট নিয়ে কথা বলছে। আমার কথা জব প্লেসমেন্ট নয়— বরং এই কোর্সটা হচ্ছে যেকোনো ডেটা সায়েন্স রোডম্যাপের শুরুর ক্রসরোড। এই রাস্তায় ওঠানোটা জরুরি।

গ্রো উইথ গুগল সিরিজ

আমার দুটো সন্তানকে (৯ম এবং কলেজে পড়ছিল) দেখিয়েছিলাম কীভাবে গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন শুরু করতে হয়। ফিন্যান্সিয়াল অ্যাসিস্ট্যান্স থেকে শুরু করে তাদের জার্নিটা আমি দেখেছি— অসাধারণ

কনসেপ্ট। এটা ঠিক যে গুগল এবং কোর্সেরা মিলে অসাধারণ কোর্স কনটেন্ট (হাজারো স্টাডি ম্যাটেরিয়াল) নিয়ে এসেছে— তবে সেগুলো বেশ 'কনসাইজ' এবং আমাদের এডুকেশন সিস্টেমের সাথে একটা বেজলাইন গ্যাপ থাকাতে, শুরুতে এই জিনিসটা বুঝতে হয়তোবা অনেককে কিছু কাঠখড় পোড়াতে হবে।

আবারও বলছি, সমস্যা হচ্ছে এটা মোট ৮টা কোর্সের বিশাল আলাপ। ডেটাকে কীভাবে একদম শুরু থেকে বুঝতে হয়, হ্যান্ডেল করতে হয় এবং ডেটা থেকে 'লব্ধ জ্ঞান' নিয়ে সেটা কীভাবে ব্যবসায় ব্যবহার করা যায়, সেটা না দেখলে বোঝার উপায় নেই। গুগল এবং কোর্সেরা তাদের এত দিনের জ্ঞান দিয়ে তৈরি করেছে এই চমৎকার সার্টিফিকেশন। আমি মুগ্ধ।

আমার রিসার্চ বলে গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন সাধারণত ১৭৪ ঘণ্টা সময় নেয় কমপ্লিট করতে। নিচের টেবিল দুটো দেখুন। এটা কমিয়ে আনা সম্ভব, তবে 'এইড অ্যাপ্লিকেশন' প্রসেস শুরু থেকে শেষ পর্যন্ত বাড়তি সময়ের কারণে ৩ মাসের নিচে আনা কষ্টকর। তবে কিওয়ার্ড, সার্টিফিকেশন ঠিক আছে, আমরা শিখব নিজে থেকে।

এই সার্টিফিকেশনে ৮টি আলাদা আলাদা কোর্স (একটি ক্যাপস্টোন প্রজেক্টসহ) একেকটা করতে (সার্টিফিকেশন পাওয়ার জন্য নয়— আসল শেখা) গড়ে প্রায় ২০ ঘণ্টা লাগবে। পুরো সার্টিফিকেশনে গুগলের নিজেরই ২১ ঘণ্টার ভিডিও আছে— যাকে একটা কোর্স হিসেবে দেখলে প্রতি কোর্স ধরে ১৭৫ মিনিটের ভিডিও দেখতে হবে (মাস্ট)।

'গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন' হয়ে গেলে আপনারা দেখে নেবেন ডেটা সায়েন্টিস্ট হবার ক্রসরোড। এটাই ফাইনাল ডেস্টিনেশন! রোডম্যাপ চ্যাপ্টারটা দেখুন।

কোর্সে এনরোল হবার আগে এই ছবিগুলো বোঝা মাস্ট। গুগল ডেটা সার্টিফাইড হলে বাড়তি সুবিধা হিসেবে গুগল থেকে গ্লোবাল কমিউনিটি, ইউএস মার্কেট ধারণা/এক্সেস এবং ১ বছরের 'বিগ-ইন্টারভিউ'র সাবস্ক্রিপশন আপনাকে নিয়ে যাবে অন্য লেভেলে।

গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন: সামারি

	Course Name	Approx. Total Time to complete	Video Lecture Time
1	Foundations: Data, Data, Everywhere	20 hours	156 minutes
2	Ask Questions to Make Data-Driven Decisions	17 hours	149 minutes
3	Prepare Data For Exploration	22 hours	140 minutes
4	Process Data from Dirty to Clean	22 hours	213 minutes
5	Analyze Data to Answer Questions	23 hours	193 minutes
6	Share Data Through the Art of Visualization	23 hours	179 minutes
7	Data Analysis with R Programming	37 hours	213 minutes
8	Google Data Analytics Capstone: Complete a Case Study	10 hours	51 minutes
	Total	**174 hours**	**21 hours, 33 Minutes**

চিত্র ১০৬: গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন: একনজরে

Google Data Analytics Professional Cert Content Breakdown: Google Data Analytics					
	Course Name	Videos	Readings	Practice Quizzes	Graded Quizzes
1	Foundations: Data, Data, Everywhere	39	33	18	6
2	Ask Questions to Make Data-Driven Decisions	38	29	18	5

Google Data Analytics Professional Cert Content Breakdown: Google Data Analytics					
3	Prepare Data for Exploration	43	29	23	5
4.	Process Data From Dirty to Clean	43	27	19	5
5.	Analyze Data to Answer Questions	39	28	22	5
6	Share Data Through the Art of Visualization	39	31	22	5
7	Data Analysis with R Programming	31	36	35	6
8	Google Data Analytics Capstone: Complete a Case Study	14	23	7	1

টেবিল ১৩: গুগল ডেটা অ্যানালাইটিক্স সার্টিফিকেশন: কোথায় কী আছে?

ডেটা সায়েন্স/অ্যানালাইটিক্স রোডম্যাপ

ডেটা সায়েন্স দিয়ে একটা রোডম্যাপ— একদম শুরু থেকে। ফর অ্যাবসলিউট বিগিনার্স! পিএইচডি দরকার নেই এই ডেটা সায়েন্স করতে!

A good Plan is like a road map : it shows the final destination and usually the best way to get there.

— H. Stanley Judd

ডেটা সায়েন্স/অ্যানালাইটিক্স নিয়ে অনেক ধরনের 'আলাপ' চালু আছে ইন্ডাস্ট্রিতে, বিশেষ করে কীভাবে এবং কত সময়ের মধ্যে একজন ভালো ডেটা সায়েন্টিস্ট/অ্যানালাইটিক্স হওয়া যায়। এটার উত্তর অনেকটাই আপেক্ষিক, কারণ ডেটার ব্যাপারে কে কীভাবে ভাবছে, তার ওপর নির্ভর করছে কে ডেটা নিয়ে কোন লেভেল পর্যন্ত কাজ করতে চাইবেন। ডেটা নিয়ে কাজ করতে হলে শুরুতে ডেটা অ্যানালাইটিক্স (ডেটা থেকে জ্ঞান) দিয়ে শুরু করা যেতে পারে। তবে ডেটা অ্যানালাইটিক্স (ক্ষেত্রবিশেষে ৫-৬ মাস) দিয়ে শুরু করে ডেটা সায়েন্স (১২-১৬ মাস)-এর মধ্যে করা সম্ভব।

ডেটা সায়েন্স একটা মহিরুহ জিনিস (বটগাছ), তবে আমার মতে ডেটা সায়েন্সের মতো গভীর বিষয়ে যাবার আগে ডেটা থেকে কীভাবে 'ইনসাইট' নিতে হয় (ডেটা অ্যানালাইটিক্স) সেটা বুঝতে পারলে আস্তে আস্তে 'ডেটা সায়েন্স' নিয়ে কাজ করা সুবিধাজনক হবে। ডেটাকে কীভাবে 'হ্যান্ডেল' করা হচ্ছে, এর পাশাপাশি সেই ডেটা কী ধরনের আউটকাম দিতে পারে— সেগুলো বোঝা গেলে ডেটা সম্পর্কিত এই সায়েন্স বোঝা অনেকটাই সোজা হয়ে যায়। যারা হাজারো ডেটা সায়েন্স নিয়ে বিভিন্ন টিউটোরিয়াল দেখছেন, তাদের বেসিক ধারণা ঠিক না থাকার ফলে মাঝপথ থেকেই আগ্রহ হারিয়ে ফেলছেন। সে কারণেই আমার এই গাইডলাইন।

আবারও বলছি,

এখানে সোজা এবং সিলেক্টিভ হচ্ছে, ডেটা অ্যানালাইটিক্স রোডম্যাপ, যার শুধু * অংশগুলো নেয়া হয়েছে পুরো ডেটা সায়েন্স রোডম্যাপ থেকে। আবারও বলছি, ডেটা সায়েন্স একটা মহিরুহ জিনিস (বটগাছ), শুরুতে ডেটা অ্যানালাইটিক্স শেখা সহজ।

সময়সীমা	ডেটা সায়েন্স/অ্যানালাইটিক্স
ডেটা অ্যানালাইটিক্স	৫-৬ মাস
ডেটা সায়েন্স	১০-১৬ মাস

টেবিল ১৪: ডেটা অ্যানালাইটিক্স এবং ডেটা সায়েন্স শেখার সময়সীমা

যেকোনো জিনিস শেখার জন্য শুরুতে একটা লার্নিং ফ্রেমওয়ার্ক থাকা ভালো। এ ধরনের লার্নিং ফ্রেমওয়ার্কে কী কী জিনিস যোগ করতে হবে এবং তার জন্য কী ধরনের রিসোর্স এই মুহূর্তে পাওয়া যায়— প্লাস, তার সাথে কোন ধরনের টুল এই মুহূর্তে ব্যবহার হচ্ছে, সেগুলো নিয়ে ধারণা দেবার জন্যই এই রোডম্যাপ। লার্নিং ফ্রেমওয়ার্কটা একটা প্রসেস, প্রসেসগুলো ঠিকমতো বুঝে গেলে কোন ধরনের টুল সামনে ব্যবহার হবে সেটাও বোঝা যাবে— যেটা সময়ের সাথে পাল্টাবে।

এ কারণে আমি সাধারণত 'টুল ডিপেন্ডেন্ট' হবার কথা বলি না। আমার কথা একটাই— প্রসেস ঠিকমতো বুঝে গেলে যেকোনো টুল শেখা সোজা। আমরা কী শিখব, কেন শিখব— কেন দরকার— সেটার সাথে 'রেলেভেন্স' আছে কি না, সেটা জানাও দরকার ব্রডলাইনে।

লার্নিং রোডম্যাপ: ব্যাপারটা কী?

এটা অনেকটা মাইন্ড ম্যাপের মতো, যেখানে 'মাল্টি-লেভেল' স্কিল তৈরি করা খুব একটা কষ্টকর কিছু নয়। এখানে বলব, কী কী ধরনের স্কিল আমরা শিখতে চাইব এবং সেই প্রসেসগুলোর আউটকাম কীভাবে 'মেজার' করতে হবে, সেটা বুঝে গেলে আমাদের কাজ কমে আসবে। প্রতিটা লেভেলের প্রসেস এবং তার সম্পর্কিত টুলের ব্যবহার— সেগুলোও দেখব এখানে। এর পাশাপাশি আমরা দেখব, কী কী ধরনের টেকনিক কাজে

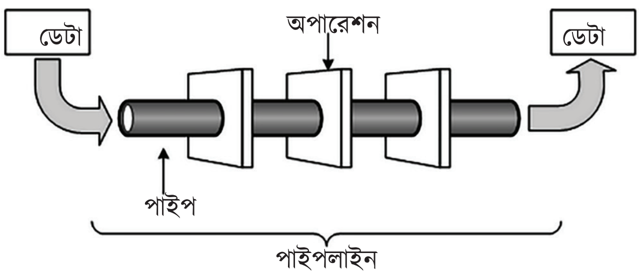
দেবে প্রতিটা স্কিলকে মাস্টারিং করতে। সেখানে কোথায় গেলে এই জিনিসগুলো সহজে শেখা যাবে, অথবা কোন কোন জায়গায় 'কম্প্লেক্সিটি' বাড়তে পারে, সেটা দেখা যাবে সামনের ধাপগুলোতে।

ডেটা সায়েন্স সামারি, (*) মার্ক = অ্যানালাইটিক্স

ডেটা সায়েন্স	অ্যানালাইটিক্স
'এন্ড টু এন্ড' পাইপলাইন	ঐচ্ছিক
'হায়ারার্কি অব নিডস'	*
ডেটা সায়েন্স ওয়ার্কস্পেস	
বেসিক অঙ্ক, লিনিয়ার অ্যালজেব্রা	
পরিসংখ্যান, ডিস্ট্রিবিউশন	*
প্রোগ্রামিং স্কিল, পাইথন/'আর'	*
এক্সপ্লোরাটরি ডেটা অ্যানালাইসিস	*
ডেটা ভিজ্যুয়ালাইজেশন	*
প্রফেশনাল ভিজ্যুয়ালাইজেশন টুল	*
বিজনেস অ্যানালাইটিক্স/ইন্টেলিজেন্স টুল	*
মডেল তৈরি, মেশিন লার্নিং	
সিক্যুয়েল স্কিল, SQL	*
নো সিক্যুয়েল (ঐচ্ছিক)	
ডেটা অ্যানালাইটিক্স ক্যাপস্টোন প্রজেক্ট	*
ডেটা সায়েন্স ক্যাপস্টোন প্রজেক্ট	
ডেটা ইঞ্জিনিয়ারিং, এপিআই	

টেবিল ১৫: ডেটা অ্যানালাইটিক্স এবং ডেটা সায়েন্সের চেকলিস্ট, * হচ্ছে বাধ্যতামূলক

আমরা জানতে চাইব— কীভাবে একজন সফল ডেটা সায়েন্টিস্ট/অ্যানালিস্ট হওয়া যায়? একদম শুরুতে, (কয়েক লাইনের) ছোট্ট ডেটা সম্পর্কিত একটা 'এন্ড টু এন্ড' প্রজেক্ট নিয়ে একটা সাধারণ ধারণা পাওয়া গেলে বাকি কাজগুলো খুব সহজ হয়ে যায়। শুরুতে ডেটা সায়েন্স/অ্যানালিস্ট পাইপলাইন বোঝা জরুরি। ডেটা কীভাবে পাইপ দিয়ে একেকটা অপারেশনের মধ্য দিয়ে ট্রান্সফর্ম হচ্ছে, সেটা না জানলে এই জিনিস বোঝা কষ্টকর। এ জন্য দরকার দুটো প্রজেক্ট।



চিত্র ১০৭: ডেটার পাইপলাইন, স্টেপ বাই স্টেপ

এরপরে থিউরি হিসেবে কিছু অঙ্ক এবং পরিসংখ্যানের ধারণা পেলে সামনে আগানো অনেকটাই সহজ হয়ে পড়বে। আমি বলছি না, একদম হার্ডকোর অঙ্ক অথবা পরিসংখ্যান শিখতে হবে— তবে আমাদের কাজে যতটুকু লাগে সেই ধারণাটা পাবার জন্য এই কাজটা জরুরি। পরবর্তী সময়ে এই পরিসংখ্যান এবং অঙ্কগুলো দিয়ে বিভিন্ন মডেল তৈরি অর্থাৎ মেশিন লার্নিং অংশে এই অঙ্ক এবং পরিসংখ্যান কীভাবে কাজ করে সেটা বোঝা সহজ হবে।

তত্ত্বকথার পাশাপাশি কিছু প্রোগ্রামিং এনভায়রনমেন্ট জানা প্রয়োজন এই সেগমেন্টে। যাদের কোনো ধরনের প্রোগ্রামিং অভিজ্ঞতা নেই তাদের জন্য 'আর' প্রোগ্রামিং এনভায়রনমেন্ট অনেকটাই গডসেন্ড। যাদের প্রোগ্রামিং সম্বন্ধে ধারণা আছে তারা সরাসরি পাইথন দিয়ে শুরু করতে পারেন। তবে মেশিন লার্নিং অর্থাৎ ডেটা অ্যানালাইসিস বুঝতে 'আর' প্রোগ্রামিং এনভায়রনমেন্ট এখনো উচ্চমানের 'লার্নিং কার্ভ' দিয়ে থাকে।

ডেটা থেকে প্রাপ্ত জ্ঞান থেকেই আমাদের সিদ্ধান্ত তৈরি করতে হয়। সেই পার্সপেক্টিভ থেকে ডেটা 'ভিজ্যুয়ালাইজেশন' ব্যাপারটা বোঝা জরুরি। এই ডেটা ভিজ্যুয়ালাইজেশনের জন্য 'আর' এবং পাইথন প্রোগ্রামিং এনভায়রনমেন্টে অসাধারণ কিছু টুল তৈরি আছে অনেক আগে থেকে। ফাইনালি, এত জিনিস শেখার পর আমাদের সরাসরি 'টেস্ট' করতে হবে রিয়েল লাইফ এনভায়রনমেন্টে। অর্থাৎ আমরা যা শিখলাম সেগুলো বাস্তব জীবনে ব্যবহার করতে পারছি কি না, সেটাকে বোঝার জন্য করতে হবে বেশ কয়েকটা ক্যাপস্টোন, অর্থাৎ ডেটা সায়েন্স প্রজেক্ট। সামারি শেষ, চলুন বিস্তারিত গাইডলাইনে।

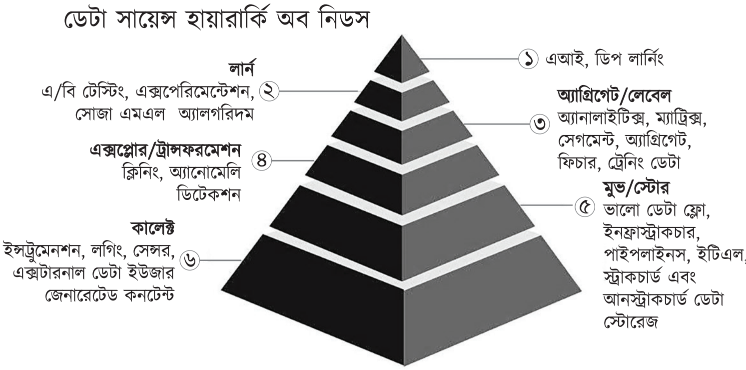
বেসিক 'এন্ড টু এন্ড' পাইপলাইনের ধারণা (ঐচ্ছিক)

ডেটা সায়েন্সের বিগ পিকচার পাবার জন্য 'মনিকা রোগাতি'র ডেটা সায়েন্সের 'হায়ারার্কি অব নিডস' একটা আবশ্যিক জিনিস। এটা ঠিকমতো বুঝতে পারলে পুরো রোডম্যাপটি মাথায় চলে আসবে। নিচের ছবিটা দেখুন। এর কয়েকটা 'ভ্যারিয়েন্ট' দেখা উচিত ইন্টারনেটে, অন্য লেভেলের পার্সপেক্টিভ বোঝার জন্য। ছবিগুলো ইংরেজিতেই রাখছি— ভালো বোঝার জন্য। তবে মাসলোর তত্ত্ব অর্থাৎ বেসিক 'হায়ারার্কি অব নিডস' গুগল করে দেখা উচিত সবারই।

চিত্র ১০৮: 'মনিকা রোগাতি'র সোজা ভার্সন

ডেটা সায়েন্সের ‘হায়ারার্কি অফ নিডস’ *

কয়েকটি ডায়াগ্রাম দেখুন। একটার সাথে আরেকটা কানেক্টেড। নিচের ছবিতে বিস্তারিত বলেছি।



চিত্র ১০৯: ‘মনিকা রোগাতি’র আরেকটা ভার্সন

কাজের ধারণায়, ওপরের ছবিটা অনেকটাই ইন্টারকানেক্টেড, তবে এটা বেসিক থেকে একটু বেশি ধারণা দেয়। কোথায় কাজ করতে চান আপনি, সেটার পার্সপেক্টিভ পাবেন এখানে। মাসলোর তত্ত্বের মধ্যে ডেটা সায়েন্সের হায়ারার্কি অব নিডসের শুরুটা হচ্ছে ডেটা কালেকশন মেকানিজম দিয়ে। দিন শেষে ডেটা সায়েন্সের কাঁচামাল হিসেবে ডেটাকে কালেক্ট করার জন্য এর প্রয়োজনীয় ইন্সট্রুমেন্টেশন, বিভিন্ন ডেটা লগ, বিভিন্ন ধরনের ডেটাসেট অর্থাৎ ডেটাবেইস, সেন্সর থেকে প্রাপ্ত ডেটা, ইত্যাদি ঠিকমতো কালেকশন করতে না পারলে এর পরবর্তী স্টেজে যাওয়া দুষ্কর। এই ছয় নম্বর সেগমেন্টে কী ধরনের টুল ব্যবহার প্রয়োজন— সেটার বিস্তারিত আলাপ থাকছে সামনে।

ডেটা কালেকশন হয়ে গেলে, এর জন্য স্টোরেজ এবং ইনফ্রাস্ট্রাকচার নিয়ে পঞ্চম ধাপ। এই ধাপে ডেটার রিলায়েবল ফ্লো, বিভিন্ন ধরনের পাইপলাইন, এর সাথে ‘ইটিএল’ অর্থাৎ ‘এক্সট্রাক্ট ট্রান্সফর্ম এবং লোড’ ভিত্তিক পাইপলাইনের ইনফ্রাস্ট্রাকচার তৈরি হয়ে গেলে চতুর্থ ধাপে যাওয়া সহজ হয়। চতুর্থ ধাপে ডেটার কিছু এক্সপ্লোরেশন, সঙ্গে ট্রান্সফর্মেশন— যার মধ্যে ডেটা ক্লিনিং, অথবা ‘এনামলি ডিটেকশন’ ইত্যাদি ইত্যাদি করা

সম্ভব। ডেটাকে ঠিকমতো ক্লিন করা গেলে সেটাকে ট্রান্সফর্ম করে পাঠানো যেতে পারে তিন নম্বর ধাপে।

এই তিন নম্বর ধাপটি খুবই গুরুত্বপূর্ণ— কারণ এখানে ডেটার এগ্রিগেশন, সম্ভাব্য লেবেলিং করা সম্ভব। ডেটার ভেতর থেকে জ্ঞান পাবার জন্য অ্যানালাইটিক্স (যেটা এই বইয়ের মূল লক্ষ্য), এর সঙ্গে বিভিন্ন ধরনের 'ম্যাট্রিক্স'-এর ধারণা পাওয়া এবং তার সম্পর্কিত 'ফিচার সেগমেন্টেশন' এই ধাপে করা সম্ভব। এ কারণেই তৃতীয় ধাপ অনেকের কাছে খুবই আকর্ষণীয়। তবে একদম শুরুর ষষ্ঠ ধাপ থেকে শুরু করলে ডেটা সায়েন্সের আসল কষ্টটা বোঝা যায়। বলতে গেলে, ডেটা কালেকশন এবং ডেটা ক্লিনিং সবচেয়ে কষ্টকর অংশ, যেখানে ডেটা সায়েন্সের প্রায় ৮০% সময় চলে যায়।

তৃতীয় ধাপে ডেটা অ্যানালাইটিক্স অর্থাৎ ডেটা থেকে ঠিকমতো জ্ঞান পাওয়া গেলে দ্বিতীয় ধাপ অর্থাৎ মেশিন লার্নিং (বেসিক লেভেলের) অ্যালগরিদম ব্যবহার থেকে শুরু করে বিভিন্ন ধরনের 'এক্সপেরিমেন্টেশন' যেমন এ/বি টেস্টিং, (গুগল করে দেখুন) ধরনের বেসিক কাজগুলো এখানেই সম্ভব। সত্যি কথা বলতে, দ্বিতীয় ধাপের কাজ দিয়েই আমাদের অধিকাংশ প্রডাক্ট ডেভেলপমেন্ট করা যায়। তবে আরও আধুনিক ডিপ লার্নিং মডেল এবং কৃত্রিম বুদ্ধিমত্তার কাজগুলো চলে যাবে শেষের প্রথম ধাপে, যেখানে এই মুহূর্তে আমাদের মাথা না ঘামালেও চলবে। প্রতিটা ধাপের টুলগুলো ওপেনসোর্স ফ্রেমওয়ার্ক হিসেবে ব্যবহার হচ্ছে পৃথিবীজুড়ে। আমি যেটা বলতে চাইছি, ডেটা সায়েন্সের 'হায়ারার্কি অব নিডস'-এর সব টুল ওপেনসোর্স ফ্রেমওয়ার্কে পাওয়া যায়।

ডিপ ডাইভ: 'হায়ারার্কি অব ডেটা সায়েন্স নিডস' (সবার জন্য প্রযোজ্য নয়)

এ ব্যাপারটা গুরুত্বপূর্ণ বলেই নতুন করে লিখছি— যারা ডেটা সায়েন্সে ক্যারিয়ার গড়তে চান। এখানে বেশ কিছু জিনিস 'রিপিট' করছি, যাতে একজন সাধারণ শিক্ষার্থী আরও ভেতরে ঢুকতে পারেন।

বড় বড় গল্প দেবার আগে ডেটা অ্যানালিস্ট, ডেটা ইঞ্জিনিয়ার এবং ডেটা সায়েন্টিস্টের মধ্যে পার্থক্য বুঝতে পারলে কাজগুলো খুব সহজ হয়ে যাবে। অনেকেই বুঝতে পারেন না তাদের জন্য কোন কাজটা নিজের স্কিলের সাথে যাবে, অর্থাৎ তার কাজের প্রোফাইলের সাথে কোন

স্কিলসেটটা যেতে পারে? এর পাশাপাশি আমাদের জীবনটাকে আরও কমপ্লেক্স (!) করার জন্য বিভিন্ন কোম্পানি তাদের জব পোস্টিং বিভিন্ন কথা বলে থাকেন— যা দেখে মনে হয় এই তিনটা ডেফিনেশনের মধ্যে নিজেরাই তালগোল পাকিয়ে ফেলছেন। এই ঝামেলা থেকে বের হওয়ার সবচেয়ে সহজ উপায় হচ্ছে জব পোস্টিংগুলোতে জব রেস্পন্সিবিলিটি (জব ডেসক্রিপশন)গুলোকে খুঁটিয়ে খুঁটিয়ে পড়া। এতে হেডিংয়ে কী বলা থাকছে অথবা থাকছে না— তার থেকে ভালো ধারণা দেবে ভেতরের ডেসক্রিপশন পড়লেই। সবচেয়ে মজার কথা হচ্ছে, এই ডেসক্রিপশন পড়লেই ওই কোম্পানিতে আমাদের কী কাজ করতে হবে সেটা বোঝা যাবে সহজে। আর এর ফলে ডেটা অ্যানালিস্ট, ডেটা ইঞ্জিনিয়ার অথবা ডেটা সায়েন্টিস্টের 'ট্যাগলাইন' অর্থাৎ 'জব হেডিং' বিবেচ্য নয়।

এই ঝামেলা থেকে বের হওয়ার জন্য আমরা আবারও আলাপ করছি 'মনিকা রোগাতির' 'হায়ারার্কি অব ডেটা সায়েন্স নিড' নিয়ে, তবে এবার আলোচনা করছি একদম ভেতরে গিয়ে। এই বিভিন্ন পজিশন নিয়ে আলাপ করার আগে আমাদের বোঝা উচিত— কীভাবে, কোথায় এবং কখন (অর্থাৎ কোন স্টেজে) এই ডেটাকে ব্যবহার করতে হয়। এই ছবিটা দেখে 'ডেটা কালেকশন মেকানিজম' সম্বন্ধে আমরা যদি না জানি, তাহলে এর ওপরের ডেটা সায়েন্স/মেশিন লার্নিং অথবা কৃত্রিম বুদ্ধিমত্তা নিয়ে কাজ করার কোনো ফলাফল ঠিকমতো আসবে না। কৃত্রিম বুদ্ধিমত্তা পৃথিবীর সব সমস্যার সমাধান অলৌকিকভাবে করে দেবে না এবং এর পাশাপাশি আমাদের বর্তমান কাজে সব জায়গায় 'কৃত্রিম বুদ্ধিমত্তা' অথবা 'মেশিন লার্নিং' স্পেসিফিকভাবে লাগবেও না। 'কৃত্রিম বুদ্ধিমত্তা' অথবা 'মেশিন লার্নিং' নিয়ে কাজ করার আগে প্রচুর ছোট ছোট কাজ (লো হ্যাঙ্গিং ফ্রুট) আছে, যেগুলো করতে এত ঝামেলা পোহাতে হয় না। তবে, আমাদের বিজনেসে ডেটা ব্যবহার করতে হলে— সেই ডেটা কীভাবে স্টোর (সংগ্রহ, এবং সংরক্ষণ) করা আছে সেটা জানতে হবে আগে। বিভিন্ন জায়গা থেকে যখন ডেটা পাই, তখন সেই ডেটাকে ঠিকমতো স্টোর করার আগ পর্যন্ত সেটাকে কীভাবে 'ম্যানিপুলেট' করা হবে— সেটার ধারণা আসে না।

আমাদের ডেটাকে ‘স্টোর’ করার জন্য কোনো রিলেশনাল ডেটাবেইস অথবা সাধারণ ‘কমা সেপারেটেড ভ্যালু’, অর্থাৎ ‘সিএসভি’ অর্থাৎ খানিকটা এক্সেল ফরম্যাটে থাকুক না কেন, সেগুলোকে একটা জায়গা থেকে আরেকটা জায়গায় নেবার জন্য কিছু ডেটা পাইপলাইন তৈরি করতে হবে— যাতে একই ডেটা এক ফরম্যাট থেকে আরেক ফরম্যাটে ‘ট্রান্সফর্ম’ করা যায়। শুরুতে এ ধরনের ডেটা পাইপলাইন ঠিকমতো বোঝা না গেলেও বড় বড় কম্প্লেক্স সমস্যা সমাধানে এ ধরনের ডেটা পাইপলাইন যেকোনো সমস্যাকে সহজ করে দেয়। এতে ডেটা এক জায়গা থেকে সরিয়ে আরেক জায়গায় তার দরকারি ফরম্যাটে স্টোর করা যায়। এ ধরনের কমপ্লেক্স তবে ডিস্ট্রিবিউটেড ডেটাবেইস সিস্টেমের জন্য— বিশেষ করে ডেটাকে এক জায়গা থেকে সরিয়ে আরেক জায়গায় নিজস্ব ফরম্যাটে রাখতে আমাদের প্রয়োজন হবে ডেটা ইঞ্জিনিয়ারদের।

এখানে এই পাইপলাইন লেখা মানে তৈরি করতে ডেটা ইঞ্জিনিয়ারদের ডেটা স্ট্রাকচার সম্বন্ধে ধারণা থাকা জরুরি। ডেটা ইঞ্জিনিয়ারদের একটা বড় কাজ হচ্ছে এই ডেটাগুলোকে ঠিকমতো ‘ক্লিন’ এবং ট্রান্সফর্ম করা, যাতে ডেটাগুলোকে ব্যবহার করা যায়। ডেটাকে ঠিকমতো কোয়েরি করা না গেলে সেই ডেটার প্রয়োজন নেই। উদাহরণ হিসেবে, ডেটা প্রিপারেশনের সময় যদি ডেটা ‘ডুপ্লিকেট’ হয়ে গেলে সেটাকে ‘ডি-ডুপ্লিকেট’ করার ধারণা থাকতে হবে আমাদের। এখানে ‘সিক্যুয়েল’ (অথবা, অনেকের ভাষায় এসকিউএল) দিয়ে ঠিকমতো ডেটাকে ‘কোয়েরি’ করা গেলে আমাদের অনেক কাজই কমে আসবে। ডেটাকে ঠিকমতো কোয়েরি করে সাধারণ কিছু ‘রেজ্যুলেশন’ ডেটার ভেতরের ধারণা সম্বন্ধে বেসিক ধারণা দিতে পারে এই পর্যায়ে।

এর পরে আসছে ডেটা অ্যানালিস্ট, বিজনেস অ্যানালিস্ট, প্রডাক্ট ম্যানেজার, সফটওয়্যার ইঞ্জিনিয়ার— তারা সবাই ডেটাকে ‘কোয়েরি’ করে তাদের প্রশ্নের উত্তর বের করতে পারলে সমস্যা অনেকটাই সমাধানের পথে হাঁটতে পারি। এখানেই বিভিন্ন ধরনের ‘কোয়েরি’ চালিয়ে তাদের বিজনেস প্রবলেমগুলো সম্বন্ধে একটা ভালো ধারণা পাওয়া যায়— তার অ্যানালাইটিক্যাল অংশে। আমাদের প্রশ্নের অনেক উত্তর পাওয়া যায় ঠিকমতো প্রশ্ন করতে পারলে এবং সেটা হবে এই স্টেজে। আমাদের

অনেক প্রডাক্ট তৈরি এবং বিজনেস ডিসিশন আসবে এই অংশ থেকে। ‘ডেটা অ্যানালিস্ট’-এর ডোমেইন এই অংশটুকু। যেহেতু আমরা এর মধ্যে দুটো বড় বড় স্টেজ পার হয়ে এসেছি— ফলে এই ‘ডেটা অ্যানালাইসিস’ অংশে ডেটা ঠিকমতো চলে এলে সঠিক কোয়েরি চালিয়ে বিভিন্ন প্রশ্নের উত্তর পাওয়া যাবে এই স্টেজে।

এর মধ্যে আমাদের চারটা স্টেজে ডেটাগুলোকে এমনভাবে ট্রান্সফর্ম করা হয়েছে, যাতে ডেটাগুলোকে দিয়ে আরও ‘গ্র্যানুলার’ কাজ করা যায়। এই স্টেজে দেখা যায়, বিভিন্ন ধরনের ‘এ/বি টেস্টিং ফ্রেমওয়ার্ক’ যেখানে বিজনেসের কোথায় কোথায় ‘ফাইন টিউনিং’ করলে আলাদা আলাদা ধরনের প্রডাক্ট বিভিন্ন মানুষের কাছে পৌঁছানো সম্ভব। একটা প্রডাক্টের ‘ইউ-এক্স’ ডিজাইনের নীল নাকি লাল বাটনটা শেষমেশ কাজ করবে, সেটা বোঝা যাবে এই ধারণা থেকে। এখানে বোঝা যাবে, কোন প্রডাক্টের কোন কোন ফিচার আসলেই দরকারি এবং এর সাথে কোন ধরনের ‘ইনক্রিমেন্টাল চেঞ্জ’ প্রডাক্টটাকে আরও আকর্ষণীয় করে তুলবে।

একই প্রডাক্টের ভিন্ন ভিন্ন ফিচার দিয়ে গ্রাহকের বিভিন্ন সেগমেন্টের এ ধরনের ‘এ/বি টেস্টিং’ করতে পারলে সেটা একটা মজার জিনিস হবে। এখানে সাধারণ ‘লিনিয়ার রিগ্রেশন’ অ্যালগরিদম চালিয়ে দেখতে পারি যাতে একজন ব্যবহারকারীর ভবিষ্যৎ বিহেভিয়ার ধরতে পারি— ভবিষ্যতে তিনি কোন ফিচারগুলো আরও পছন্দ করবেন। এখানে আমাদের প্রয়োজনমতো সাধারণ কিছু অ্যালগরিদম দিয়ে এ ধরনের টেস্ট প্রেডিক্ট করা যায়— যাতে ভবিষ্যৎ ব্যবহারকারীদের আচরণ আমরা আগে থেকেই ধরতে পারি।

সবচেয়ে ওপরের স্টেজে কৃত্রিম বুদ্ধিমত্তা অথবা ডিপ লার্নিং নিয়ে কাজ করতে গেলে সবচেয়ে প্রথমেই প্রয়োজন ক্লিন ডেটা। আর সে কারণেই আগের সবগুলো স্টেজে যে পাইপলাইনগুলো তৈরি করা হয়েছিল, সেগুলোকে ঠিকভাবে কাজ করতে হবে। আমাদের ট্রেনিং ডেটাগুলোকে ঠিকমতো ‘লেবেল’ করতে পারতে হবে, যাতে আমাদের ভুলগুলো কমে আসে। এখানে আমাদের ‘প্রি-রিকুইজিট’গুলোকে ঠিকমতো ‘আইডেন্টিফাই’ করতে পারলে ওপরের স্টেজগুলোতে কাজ সহজ হয়ে যাবে। আমাদের আগের স্টেজে সাধারণ মেশিন লার্নিং অ্যালগরিদম যেমন, ‘লিনিয়ার ইকুয়েশন’ ইত্যাদি ইত্যাদি আশানুরূপ ফলাফল না দিলে এরপরের স্টেজে ‘ডিপ লার্নিং’ নিয়ে কাজ করা যেতে পারে।

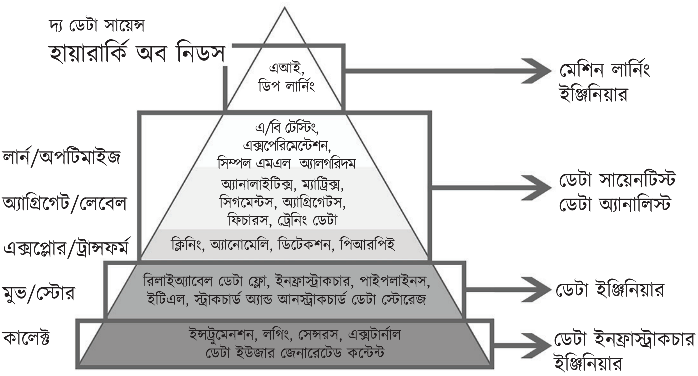
জব ডেসক্রিপশন, কোথায় ডেটা অ্যানালিস্ট, ডেটা ইঞ্জিনিয়ার এবং ডেটা সায়েন্টিস্ট কাজ করবেন?

এখন আমাদের প্রশ্ন হতে পারে— এই পিরামিডের কোথায় কোথায় ডেটা অ্যানালিস্ট, ডেটা ইঞ্জিনিয়ার এবং ডেটা সায়েন্টিস্ট নিজেদের 'ফিট' করাতে পারেন? আমাদের অভিজ্ঞতা বলে, ডেটা ইঞ্জিনিয়ার সাধারণত কালেকশন, এক্সপ্লোর এবং ট্রান্সফর্মেশন ইত্যাদি নিয়ে কাজ করেন, সবচেয়ে নিচের দুটো ব্লকে। সফটওয়্যার ইঞ্জিনিয়ার সাধারণত কালেকশন পার্টটা নিয়ে কাজ করেন, যাতে 'ফ্রন্ট-এন্ড' এবং 'ব্যাক-এন্ডে' ঠিকমতো যোগসূত্র করা যায়। কারণ, এখান থেকেই বিভিন্ন ধরনের 'ইউজার ডেটা' সংগ্রহ করা হয়। ডেটা অ্যানালিস্ট সাধারণত কাজ করেন 'এগ্রিগেট' অথবা 'লেবেলিং' লেভেলে। এখানে ডেটাকে 'এগ্রিগেট' করে, যাতে একটা ইনফর্মড ডিসিশন নেওয়া যায়, এই স্টেজে আমাদের প্রশ্নের সচরাচর উত্তর আসে। আমরা সাধারণত সিদ্ধান্ত নিই এই ডেটার উত্তর থেকে।

একজন ভালো অ্যানালিস্ট এই ডেটা থেকে কোম্পানির ভবিষ্যৎ 'ডাইরেকশন' এবং কী কী জায়গায় কাজ (সেটা ফিচার অথবা প্রডাক্ট হতে পারে) প্রয়োজন সেটা বের করতে পারেন। তারা টেকনিক্যাল এবং প্রডাক্টের 'ইনটুইশন' নিয়ে বেশ ভালো ধারণা নিয়ে থাকেন, এর পাশাপাশি— তাদের 'কমিউনিকেশন স্কিল' অসাধারণ লেভেলের হতে হয়। এই কমিউনিকেশন স্কিল তাদের কোম্পানির সবাইকে এক ধারণায়/লেভেলে রাখতে সাহায্য করে। অনেক কোম্পানিতে, তাদের জব ডেসক্রিপশন অনুযায়ী ডেটা অ্যানালিস্টকে ডেটা সায়েন্টিস্ট হিসেবে দেখায়, যদিও ডেটা সায়েন্টিস্টদের আরও বেশি টেকনিক্যাল স্কিলসেট থাকে। সেদিক থেকে একজন ডেটা সায়েন্টিস্ট সাধারণ মেশিন লার্নিং মডেল থেকে শুরু করে কৃত্রিম বুদ্ধিমত্তার কিছু কিছু অংশ নিজে সলভ করে থাকেন।

এখানে অনেক সময় রিসার্চ সায়েন্টিস্ট, এমএল ইঞ্জিনিয়ারের সহায়তায় কমপ্লেক্স ওপরের সমস্যাগুলোকে সলভ করার চেষ্টা করেন। এই জব ডেসক্রিপশনে অনেকেই কাজ করেন— যারা 'পিএইচডি ক্যান্ডিডেট' হিসেবে থাকেন। তবে কোম্পানিভেদে এই জিনিসগুলো বিভিন্নভাবে ডেলিভার হতে পারে। যেমন, এই অংশটুকু লেখার জন্য আমি সাহায্য নিয়েছি একজন 'এক্স গুগলার' থেকে, তার নাম 'জোমা', যাকে গুগলে

থাকার সময় সবকিছুই করতে হয়েছে। আবার, যারা ছোট স্টার্টআপ, তাদের একজন ডেটা সায়েন্টিস্টকে সব কাজই করতে হয়। মোদ্দা কথা হচ্ছে, এখন থেকে যত ডেটা সায়েন্টিস্ট এবং ডেটা অ্যানালিস্টের জব পোস্টিং দেখবেন, সব সময় জব ডেসক্রিপশন পড়ার চেষ্টা করবেন। এটাই আপনাকে সাহায্য করবে আসলেই সেই পজিশনটা আপনার জন্য ফিট করে কি না।



চিত্র ১১০: কার কোথায় কাজ, বেসিকভাবে

এই পুরো জিনিসটি করতে আপনাকে সাহায্য করবে আমার প্রথম দুটো বই। ‘হাতেকলমে মেশিন লার্নিং’ বইয়ের অর্ধেক অংশ আলাপ করা হয়েছে আমার ব্যক্তিগত অভিজ্ঞতা এবং গল্প দিয়ে। দ্বিতীয় বই ‘শূন্য থেকে পাইথন মেশিন লার্নিং’ বইটা যেকোনো নন-প্রোগ্রামিং ব্যাকগ্রাউন্ডের ছাত্রছাত্রী শুরু করতে পারবেন কোনো সমস্যা ছাড়াই। হাতেকলমে পাইথন এবং ‘আর’ প্রোগ্রামিং এনভায়রনমেন্ট (প্রোগ্রামিং এনভায়রনমেন্ট জানার দরকার নেই) দিয়ে হাঁটি হাঁটি পা পা করে পুরো জিনিসের পাইপলাইনগুলো বোঝানো হয়েছে খুব সহজ করে।

পাইপলাইন বুঝলে বেসিকভাবে একটা/দুটো ছোট (আসলেই ছোট) কয়েক লাইনের প্রজেক্ট করলেই একটা ধারণা চলে আসবে। বই কেনার দরকার নেই শুরুতে। প্রজেক্ট দুটো হচ্ছে—

প্রজেক্ট	বইয়ের লিংক	সাহায্যকারী বইয়ের নাম
আইরিস প্রজাতি প্রজেক্ট	t.ly/I54H	শূন্য থেকে পাইথন মেশিন লার্নিং
টাইটানিক প্রজেক্ট	t.ly/fsBP	হাতেকলমে মেশিন লার্নিং

এখন আমরা যেহেতু খুবই ভিজুয়্যাল (ভিডিও দেখতে ভালোবাসি), সে কারণে এই ভিডিওটা একটা ভালো ধারণা দিতে পারে। এটা হাতেকলমে করলেই একটা বড় ধারণা চলে আসবে। শুরুতে না বুঝে করলেও আস্তে আস্তে বুঝে যাবেন করার সময়।

(পাইথন) ডেটা সায়েন্স ওয়ার্কস্পেস, অ্যানাকন্ডা

কী নেই এখানে। মনে আছে, ডেটা সায়েন্সের 'হায়ারার্কি অব নিডস'-এর প্রতিটা ধাপের টুলগুলো ওপেনসোর্স ফ্রেমওয়ার্ক পাচ্ছেন এই অ্যানাকন্ডাতে। এতে আছে জুপিটারের নোটবুকসহ পুরো এনভায়রনমেন্ট। এর পাশাপাশি, কিছু না ইনস্টল করতে চাইলে আছে— গুগল কোলাবোরেটরি টুল।

বেসিক অঙ্ক, লিনিয়ার অ্যালজেব্রা, ক্যালকুলাস

ডেটা সায়েন্সের ভেতরের খুঁটিনাটি জানতে শুরুতে লিনিয়ার অ্যালজেব্রা'' একটা চমৎকার জিনিস। ডেটাকে প্রসেস করার আগে ওই ডেটাকে ঠিকমতো কনটেইনারে রাখার জন্য স্কেলার, ভেক্টর এবং ম্যাট্রিক্স জানা জরুরি। বিশেষ করে লিনিয়ার অ্যালজেব্রার ভেতরে ভেক্টর এবং ম্যাট্রিক্স কীভাবে কাজ করে, বিশেষ করে ট্রান্সফর্মেশন, ইনভার্স, ট্রান্সপোজ এবং ম্যাট্রিক্স মাল্টিপ্লিকেশন নিয়ে কিছু ধারণা নেয়া ভালো। ভেক্টরের মধ্যে ডট এবং ক্রস প্রডাক্ট কীভাবে করা যায়, সেটা খুব সহজেই শেখা যায় 'খান একাডেমি' থেকে। আমি নিজেও শিখেছি। এখানে লিনিয়ার ডিপেন্ডেন্সি, ইন্ডিপেন্ডেন্সি, এদের মধ্যে কম্বিনেশন ইত্যাদি শেখা যায় 'মিডিয়াম'-এর মতো আরও অনেক সাইটে। তবে সবকিছুর জন্য ভালো 'খান একাডেমি'। অথবা বিভিন্ন ম্যাথ সাইট।

ক্যালকুলাসের জন্য কীভাবে 'ডিরাইভেটিভ' বের করতে হয়, ডিফারেন্সিয়াল ইকুয়েশন,— 'ইন্টিগ্রালস'-এর সাথেই এমনি চলে আসে। সত্যি বলতে, এগুলো খুব সহজেই কয়েক লাইন দিয়ে তৈরি করা যায় পাইথন অথবা

‘আর’ প্রোগ্রামিং এনভায়রনমেন্ট দিয়ে। তবে ভেতরের অঙ্কটা জানা থাকলে কাজ করার সময় কীভাবে ভেতরের জিনিসগুলো শেষমেশ আউটকাম হিসেবে বের করেছে, সেটা বোঝা সহজ হয়। অঙ্ক জানা ভালো, কারণ এতে ডেটা ইঞ্জিনিয়ারিং-এর অংশ চমৎকারভাবে বোঝা যায়।

পরিসংখ্যান, ডেসক্রিপটিভ, সঙ্গে ডিস্ট্রিবিউশন *

ডেটা সায়েন্সের পেছনে যে বিষয়টি সবচেয়ে বড় ভূমিকা রাখছে সেটা হচ্ছে পরিসংখ্যান। পরিসংখ্যানের সব ধারণাই মেশিন লার্নিংয়ে, অর্থাৎ ডেটা সায়েন্সে ঢুকে গেছে যখন ডেটা প্রসেসিং চলে এসেছে কম্পিউটারের ভেতরে। ডেসক্রিপটিভ স্ট্যাটিসটিকস এবং ডেটার ডিস্ট্রিবিউশন ঠিকমতো বোঝা গেলে ডেটা সায়েন্স বোঝা অনেকটাই সহজ হয়ে যায়। ডেটার মধ্যে সর্বোচ্চ, সর্বনিম্ন, মিডিয়ান, মোড, গড়, সেন্ট্রাল টেনডেন্সি ইত্যাদি ডেটার ভেতরের ইতিহাস বের করে দেয়। ডেটার ‘নর্মাল ডিস্ট্রিবিউশন’ বোঝার জন্য সাধারণ কয়েকটা লেসন ইন্টারনেটে পাওয়া যায় খুব সহজে।

এর পাশাপাশি হিস্টোগ্রাম, স্ক্যাটার প্লট এবং বক্স প্লট করার ধারণা চলে এসেছে সবার ঘরে ঘরে। ডেটার ভ্যারিয়েন্স এবং স্ট্যান্ডার্ড ডেভিয়েশন জিনিসগুলো সহজেই বোঝা যাচ্ছে খান একাডেমির কল্যাণে। এগুলো এখন সাধারণ অঙ্কের মধ্যেই পড়ে। যেকোনো জিনিসের ‘কনফিডেন্স ইন্টারভ্যাল’ এবং ‘হাইপোথিসিস টেস্টিং’ চলে এসেছে রিসার্চের সাথে সাথে। এগুলো সব চলে আসছে পপুলেশন প্রপোরশন, ‘পি ভ্যালু’ এবং ‘সিগনিফিকেন্স টেস্ট’ করতে গিয়ে। দুষ্টলোকেরা এটার সাথে ‘এসপিএসএস’ প্যাকেজ মেলাবেন। প্রথম প্রথম এগুলো কঠিন মনে হলেও আসলে ব্যাপারগুলো অনেক সোজা। একটু গুগল করেই দেখুন না!

প্রোবাবিলিটির কোর কনসেপ্টগুলো খুব একটা সমস্যা না করলেও এর ভেতরের এক্সপেরিমেন্টাল প্রোবাবিলিটি, ইন্ডিপেন্ডেন্ট এবং ডিপেন্ডেন্ট ইভেন্ট, কন্ডিশনাল প্রোবাবিলিটি জানা ভালো। এ ব্যাপারে গুগল ‘কিওয়ার্ড’ এবং খান একাডেমি আমাদের হাতের পাঁচ।

প্রোগ্রামিং স্কিল, পাইথন/'আর' এনভায়রনমেন্ট *

আগেই বলেছি— ডেটা সায়েন্স অথবা ডেটা অ্যানালাইটিক্স ঠিকমতো কাজে লাগাতে গেলে একটা প্রোগ্রামিং এনভায়রনমেন্ট জানা প্রয়োজন। এখানে বিশ্বখ্যাত দুটো প্রোগ্রামিং এনভায়রনমেন্ট নিয়ে কথা বলছি— ডেটা সায়েন্স এবং ডেটা অ্যানালাইটিক্স ঠিকমতো শুরু করার জন্য। প্রোগ্রামিং ('আর' অথবা পাইথন প্রোগ্রামিং এনভায়রনমেন্ট) নিয়ে আমাদের অনেকের মধ্যে কনফিউশন তৈরি হয়েছে, বিশেষ করে এর ডেটা সায়েন্সের ব্যবহার নিয়ে। শুরুতে বুঝতে হবে, আমরা কেন প্রোগ্রামিং এনভায়রনমেন্ট ব্যবহার করি ডেটা সায়েন্সে? যেকোনো সমস্যার শুরুতে ধারণা পেতে আমাদের ডেটা 'পুল' করতে হয় বিভিন্ন ডেটাসেট থেকে। ডেটা সায়েন্সে সেই ডেটা টেনে আনার জন্য আমরা ব্যবহার করি 'এপিআই' (অ্যাপ্লিকেশন প্রোগ্রামিং ইন্টারফেস), (ডেটাসেট থেকে আনা), অথবা 'স্ক্র্যাপিং'' (ওয়েবসাইটগুলো থেকে বিভিন্ন ডেটা টেনে আনা)-এর মাধ্যমে। পাশাপাশি, ডেটা ম্যানিপুলেশন এবং ট্রান্সফর্মেশনে (ইটিএল) ব্যবহার করি প্রোগ্রামিং। এই জায়গাগুলোতেই প্রোগ্রামিং এনভায়রনমেন্ট ব্যবহার হয়, তার আগে নয়।

'আর' প্রোগ্রামিং এনভায়রনমেন্ট

যারা নন-প্রোগ্রামিং ব্যাকগ্রাউন্ড থেকে আসেন, তাদের জন্য 'আর' প্রোগ্রামিং এনভায়রনমেন্ট অনেকটাই 'গডসেন্ড'। কারণ, পরিসংখ্যানের জন্য স্পেসিফিক্যালি তৈরি করা হয়েছিল এই 'আর' প্রোগ্রামিং টুল, যা ডেটা সায়েন্স/ডেটা অ্যানালাইটিক্সের শুরুটা ধরতে অসাধারণভাবে সাহায্য করে। যে কাজটা হয়তোবা পাইথনের করতে পাঁচ লাইনের কোড প্রয়োজন পড়ত, সেটা হয়তোবা এক লাইনেই করে ফেলতে পারে 'আর'-এর হাজারো লাইব্রেরি। এর পাশাপাশি, ডেটা ভিজ্যুয়ালাইজেশনের জন্য 'আর' এর অসাধারণ কিছু টুল এখনো বিশ্বসেরা। এর ভালো ধারণা পাওয়ার জন্য আমার প্রথম বই 'হাতেকলমে মেশিন লার্নিং' দিয়ে শুরু করা যায়। ক্লাস টেনের শিক্ষার্থীও বুঝবে ব্যাপারগুলো।

পাইথন প্রোগ্রামিং এনভায়রনমেন্ট

যাদের আগে থেকে প্রোগ্রামিংয়ের ধারণা আছে, তারা সাধারণত পাইথন দিয়ে ডেটা সায়েন্স এবং ডেটা অ্যানালাইটিক্স শুরু করেন। 'আর'-এর

মতো পাইথনের বহু ডেটা সায়েন্স লাইব্রেরি সহজ করে দিচ্ছে এই ডেটা সায়েন্স জার্নিকে। পাইথন দিয়ে ডেটা সায়েন্স শেখার জন্য আমার 'শূন্য থেকে পাইথন মেশিন লার্নিং' বইটা দেখা যেতে পারে। এর পরের বইটা 'হাতেকলমে পাইথন ডিপ লার্নিং' অ্যাডভান্সড লেভেলকে নিয়ে এসেছে সহজ করে।

(পাইথন) দিয়ে ডেটার চারটি কাজ

পাইথন দিয়ে ডেটা সায়েন্স অথবা ডেটা অ্যানালাইটিক্স বোঝার জন্য পুরো ব্যাপারটাকে চার ভাগে আলাদা করে ফেলি। ডেটা নিয়ে কাজ করতে গেলে প্রথমে আমাদের হাত দিতে হবে 'ডেটা কালেকশন'-এ। এর মানে, ডেটাকে ঠিকমতো আমাদের সিস্টেমে লোড না করতে পারলে— কোনো কাজই ঠিকমতো শুরু করা যাবে না।

(পাইথন) ডেটা কালেকশন

ডেটা কালেকশন করতে গেলে বেসিক লেভেলে পাইথনের বেশ কিছু লাইব্রেরি ব্যবহার করব যার মধ্যে পান্ডা খুব নামকরা। সত্যি কথা বলতে, এই 'পান্ডাজ' দিয়ে হেন কাজ নেই যা করা যায় না। এর একটু অ্যাডভান্স লেভেলে গেলে 'সেলেনিয়াম' এবং 'স্ক্র্যাপি' ব্যবহার করা যায়। গুগল করুন এই দুটো কিওয়ার্ড।

এক্সপ্লোরাটরি ডেটা অ্যানালাইসিস, ইডিএ *

এর পরের ধাপ হচ্ছে ডেটা অ্যানালাইসিস। যেকোনো ডেটা কালেকশন করার পর সেটাকে ঠিকমতো ক্লিন এবং ট্রান্সফর্মেশন করতে আমাদের প্রয়োজন বেশ কিছু পাইথন লাইব্রেরি। ডেটা থেকে ঠিকমতো ইনসাইট পেতে ডেটা কালেকশনের মতো একই ধরনের লাইব্রেরি ব্যবহার করলেও এখানে বেশ কিছু নতুন ধারণা চলে আসছে। এ ক্ষেত্রে ডেটা কীভাবে স্টোর হয় এবং সেই ডেটাকে কীভাবে ম্যানিপুলেশন করা যায়, সেটার জন্য আমরা ব্যবহার করব (numpy) নামপাই এবং পান্ডাজ লাইব্রেরি। এটাকে আমরা 'এক্সপ্লোরাটরি ডেটা অ্যানালাইসিস' বলি, যা আমাদেরকে ডেটাকে ঠিকমতো এক্সপ্লোর করতে দেয়, এর ভেতরের প্যাটার্ন বোঝার জন্য।

ডেটা ভিজ্যুয়ালাইজেশন (ইডিএ, এক্সপ্লোর) *

ডেটার ভেতরে ঠিক প্যাটার্নটা বুঝতে আমাদের প্রয়োজন ডেটা ভিজ্যুয়ালাইজেশন। মানুষের মাথা অসাধারণভাবে প্যাটার্ন বুঝতে পারে বলে— সেই ডেটাকে যখন বিভিন্ন অ্যাক্সিসে প্লট করা হয়, তখন সেগুলোর মধ্যে প্যাটার্ন বা সম্পর্কগুলো উঠে আসে। শুরুতে বার, বক্স প্লট এবং পাই চার্ট একটা ভালো ধারণা দেয়— কীভাবে ডেটাগুলো নিজেদের বিভিন্ন ফিচারের মধ্যে 'ইন্টার কানেক্টেড'। এটা বুঝতে শেখায় পৃথিবীতে সবকিছুই একটা প্যাটার্নে চলে— যার অনেকটাই আমাদের অজানা। এই কাজটি করতে পাইথনের 'ম্যাট-প্লট-লিব' এবং 'সিবর্ন' লাইব্রেরিগুলো চমৎকার কাজ করে।

বিজনেস অ্যানালাইটিক্স এবং ইন্টেলিজেন্স টুল *

94% of people using data in their current job role agree that data helps them do their job better.

ডেটা থেকে ঠিকমতো ধারণা নিয়ে সেটাকে স্টেকহোল্ডারকে (যারা আমাদের সিনিয়র ম্যানেজমেন্ট অথবা কাস্টমার)কে ঠিকমতো কমিউনিকেট করতে হলে ডেটার ভেতরে কী ধরনের অ্যানামলি/সম্পর্ক আছে, সেটার জন্য স্ক্যাটার প্লট, বক্স প্লট ইত্যাদি ইত্যাদি তৈরি করতে কিছু প্রফেশনাল বিজনেস অ্যানালাইটিক্স এবং ইন্টেলিজেন্স টুল ব্যবহার করতে হয়। ওপেনসোর্স টুল হিসেবে সিবর্ন ভালো, তবে অল্প সময়ের মধ্যে ভালো অ্যানালাইটিক্স এবং বিজনেস ইন্টেলিজেন্স টুল হিসেবে মাইক্রোসফট এক্সেল, গুগলশিট, ট্যাবলো এবং মাইক্রোসফট পাওয়ার বিআই অন্য লেভেলের জিনিস। ডেটা থেকে অ্যাডভান্সড অ্যানালাইটিক্স বোঝার জন্য এই কয়েকটা টুল অসাধারণ কাজ করে। মাইক্রোসফট এক্সেল, ট্যাবলো এবং মাইক্রোসফট পাওয়ার বিআই নিয়ে আমার অনেকগুলো ভিডিও আছে।

এক্সেল, গুগলশিট, ট্যাবলো, পাওয়ার বিআই *

আগেই বলেছি, ওপেনসোর্স টুল হিসেবে পাইথন চমৎকারভাবে কাজ করলেও বড় বড় ব্যবসাপ্রতিষ্ঠান ডেটা নিয়ে কাজ করার জন্য অনেক ধরনের 'প্রপাইটোরি' টুল ব্যবহার করে থাকেন। এ ধরনের 'প্রপাইটোরি'

টুলগুলো শুধু স্মার্ট নয়, একজন ব্যবহারকারীকে কত সহজে নানা ধরনের রিপোর্ট অল্প সময়ের মধ্যে বের করে দিতে পারে, সেটা নিয়েই চলছে হাজারো গবেষণা। যেহেতু সবকিছুকে তৈরি করা হচ্ছে ব্যবহারকারীর চাহিদাকে মাথায় রেখে, সে কারণে এ ধরনের টুলের দাম নির্ধারণ হয় তাদের কাজের কমপ্লেক্সিটির ওপর। সবার কম্পিউটারে মাইক্রোসফট এক্সেল সেটারই প্রমাণ। এর পাশাপাশি গুগলশিট, বিনা মূল্যে সার্ভিস দিলেও করপোরেট ডেটার ব্যাপারে কিছু সীমাবদ্ধতা রয়েছে। বিজনেস অ্যানালাইটিক্স নিয়ে ট্যাবলো, পাওয়ার বিআই অন্য লেভেলের জিনিস। তবে এর খরচ কম নয়।

(পাইথন) মডেল তৈরি, মেশিন লার্নিং

ডেটা সায়েন্সের শেষ পর্যায়ে থাকছে কীভাবে ডেটা থেকে ধারণা নিয়ে মডেল তৈরি করতে হয়। ডেটা অ্যানালাইটিক্স এবং এক্সপ্লোরাটরি ডেটা অ্যানালাইসিস থেকে যে জ্ঞানটা পাচ্ছি, সেটাকে মডেলে নিয়ে আসতে আমাদের প্রয়োজন মেশিন লার্নিং। এই অংশটুকু ডেটা অ্যানালাইটিক্সে থাকছে না। এই মডেল তৈরি এবং মডেল থেকে প্রাপ্ত জ্ঞানগুলোকে কীভাবে ব্যবহার করতে হয় সেটার পেছনে স্ট্যাটিসটিকস, অর্থাৎ পরিসংখ্যান এবং অঙ্কের সাহায্য লাগবে। আমি আবারো বলছি, অঙ্ক এবং পরিসংখ্যানের সব কাজে বিভিন্ন টুল ব্যবহার করে করলেও এর পেছনের 'আন্ডারলায়িং' আন্ডারস্ট্যান্ডিং আসবে অঙ্ক এবং পরিসংখ্যান থেকে; বিশেষ করে মডেলগুলো কীভাবে কাজ করছে সেটা বুঝতে।

এই জায়গায় বেসিক লেভেলে আমরা ব্যবহার করব পাইথনের scikit-learn এবং ডিপ লার্নিংয়ের জন্য 'টেনসরফ্লো' ফ্রেমওয়ার্ক। যারা রিসার্চে কাজ করতে চান তাদের জন্য পাইটর্চ লাইব্রেরিটা খারাপ নয়। পাইথনের scikit-learn লাইব্রেরি বিশ্বের অন্যতম ভালো ডকুমেন্টেড লাইব্রেরি, যা শেখা খুবই সহজ। আমার দ্বিতীয় বই 'শূন্য থেকে পাইথন মেশিন লার্নিং' এ ব্যাপারে চমৎকার ধারণা দিয়েছে। 'টেনসরফ্লো' ফ্রেমওয়ার্ক অর্থাৎ 'ডিপ লার্নিং' নিয়ে কাজ করতে চাইলে আমার কিছুটা অ্যাডভান্সড বই 'হাতেকলমে পাইথন ডিপ লার্নিং' ব্যবহার করতে পারেন।

এসকিউএল/সিক্যুয়েল স্কিল, SQL *

ডেটা কালেকশন করার জন্য বিভিন্ন টুল ব্যবহার করলেও সেই ডেটা কোথায় আছে সেটা নিয়েও অনেক কাজ করার স্কোপ রয়েছে। সে কারণে রিলেশনাল ডেটাবেইস কীভাবে কাজ করে তার কনসেপ্ট কী, কীভাবে সিক্যুয়েল কুয়েরি দিয়ে একটা ডেটাবেইস থেকে তার আউটকামস বের করে আনা যায়, সেটা জানা জরুরি। ডেটা সায়েন্স এবং ডেটা অ্যানালাইটিক্সের জন্য সিক্যুয়েল একটা কার্যকরী মেকানিজম।

ডেটাবেইসে কোয়েরি করার জন্য সিক্যুয়েলের বিভিন্ন স্টেটমেন্ট, যেমন 'ডিডিএল' হিসেবে create, alter, drop, truncate, rename, ডিএমএল স্টেটমেন্ট হিসেবে, select, insert, update, delete, এবং শেষে বিসিএল স্টেটমেন্ট হিসেবে grant, revoke, টিসিএল স্টেটমেন্ট হিসেবে commit, rollback ইত্যাদি জানা প্রয়োজন। এর পাশাপাশি বিভিন্ন টেবিলকে কীভাবে জয়েন, ইউনিয়ন করতে হয় তা জানা জরুরি। এখানে শুরুতে ওপেনসোর্স ডেটাবেইস ফ্রেমওয়ার্ক 'মাইএসকিউএল' নিয়ে কাজ করা যায়।

নো সিক্যুয়েল (ঐচ্ছিক)

সিক্যুয়েলের পাশাপাশি আন-স্ট্রাকচার্ড ডেটার জন্য নো সিক্যুয়েলের জনপ্রিয়তা বাড়ছে দিনে দিনে। নন-টেবুলার এবং রিলেশনাল ডেটাবেইস নয়— এ ধরনের ক্ষেত্রে এর কাজ অসাধারণ। এই পার্সপেক্টিভ থেকে 'মঙ্গোডিবি' নিয়ে কাজ করা যেতে পারে। এই ধারণাগুলো ভবিষ্যতে কাজে লাগবে।

ডেটা সায়েন্স/অ্যানালাইটিক্স ক্যাপস্টোন প্রজেক্ট

একটা জিনিস পুরোপুরি বুঝতে গেলে কাজের পাইপলাইন ধরে শুরু থেকে শেষ পর্যন্ত একটা সত্যিকারের প্রজেক্ট না করলে পুরো শেখাটাই বিফলে যেতে পারে। সে কারণে ডেটা সায়েন্স এবং ডেটা অ্যানালাইটিক্সের জন্য কিছু ইন্ডাস্ট্রি সম্পর্কিত ক্যাপস্টোন প্রজেক্ট করতে হবে। এতে সবগুলো বিষয় যা বলেছি, সেগুলোর ফাউন্ডেশন ধারণাগুলোকে রিয়েল ওয়ার্ল্ড সিনারিওর সাথে যুক্ত করা যাবে।

গ্লোবাল সুপারস্টোর প্রজেক্ট (অ্যানালাইটিক্স) *

এখানে আমরা কিছু কাজ দেখাব, সুপারস্টোর ম্যানেজমেন্ট নিয়ে। মাথা খারাপ করা একটা রিয়ালিস্টিক প্রজেক্ট।

মুভি রিকমেন্ডেশন প্রজেক্ট

নেটফ্লিক্স এবং ইউটিউব যেভাবে আমাদের রিকমেন্ড করে সেই একই পদ্ধতি ব্যবহার করে আজ দারাজ, আলীএক্সপ্রেস, আমাজন তাদের প্রডাক্ট রিকমেন্ড করে— যেটা ব্যবহারকারী হিসেবে আমরা দেখে অথবা কিনে থাকি। এ ধরনের মুভি এবং প্রডাক্ট রিকমেন্ডেশন প্রজেক্ট নিয়ে কাজ করলে রিয়েল ওয়ার্ল্ড সিনারিওর সাথে আমাদের শেখাকে সংযুক্ত করা যায়। এখানে একটা সিম্প্লিফাইড মুভি রিকমেন্ডেশন সিস্টেম বানাতে পারলে পৃথিবীর ই-কমার্সের বিশাল একটা অংশ মাথায় নিয়ে আসা সম্ভব।

কাস্টমার চার্ন ম্যানেজমেন্ট প্রজেক্ট

যেকোনো কোম্পানির জন্য তার কাস্টমার বড় সম্পদ। কোনো কাস্টমার তার সাথে থাকবেন অথবা তার প্রডাক্ট ব্যবহার না করে অন্য প্রডাক্টে সুইচ করবেন, সেটা আগে থেকে বুঝতে পারলে ব্যবসা সহজ হয়ে যায় অনেক জায়গায়। একটা মোবাইল ফোন কোম্পানি তাদের পুরোনো ব্যবহারকারীকে কীভাবে ফিরিয়ে আনবেন অথবা কোন কোন কাস্টমার কেন এবং কখন সুইচ করে অন্য কোম্পানিতে চলে যাচ্ছেন, অথবা একটা কোম্পানি থেকে একজন এমপ্লয়ি কবে অন্য কোম্পানিতে সুইচ করবেন— সেটা আগে থেকে জানা গেলে মানবসম্পদ, অর্থাৎ এইচআর ডিপার্টমেন্ট অনেক অপ্রীতিকর ঘটনা কমিয়ে আনতে পারেন।

কাজের পোর্টফোলিও ‘শো-কেসিং’ *

প্রজেক্টে কাজ করার পাশাপাশি সেই কাজগুলো অন্যদের দেখানো (‘শো-কেসিং’) জরুরি। এতে ভালো ফিডব্যাক পাওয়া যায়। নিজের আত্মবিশ্বাস বাড়ে। একটা জিনিস কীভাবে শিখলাম এবং সেটাকে কীভাবে ডেলিভারি দিতে পারছি— সেটার একটা ভালো যোগসূত্র হচ্ছে— আমাদের সব কাজের ‘পোর্টফোলিও’ সাইট তৈরি করা। যেকোনো ডেটা সায়েন্স এবং ডেটা অ্যানালিস্ট কাজ পেতে এ ধরনের ‘শো-কেসিং’ জরুরি। শুরুর

দিকে ‘শো-কেসিং’-এর জন্য— বিশেষ করে এ ধরনের সাইট তৈরিতে গিটহাব পেজ অসাধারণ কার্যকারিতা দেয়। এই সাইটটাও সেই গিটহাব পেজ দিয়ে তৈরি করা। এর পাশাপাশি, নেটওয়ার্কিং ইভেন্টগুলোকে কানেক্ট করতে হবে সেখানে।

ঐচ্ছিক: ডেটা ইঞ্জিনিয়ারিং, এপিআই

ঐচ্ছিক বিষয় হিসেবে ডেটা ইঞ্জিনিয়ারিং এবং তার ডেপ্লয়মেন্ট, এর পাশাপাশি এদের মধ্যে এপিআইভিত্তিক যোগাযোগ, বিশেষ করে জেসন ফাইল, রেস্ট, ওপেন এপিআই ইত্যাদি নিয়ে ধারণা থাকলে যেকোনো ইকোসিস্টেমে এই কাজ সহজে ফিট করতে পারবে। এর সাথে, গ্রাফ থিওরি, নেটওয়ার্ক অ্যানালাইসিস, টাইম সিরিজ অ্যানালাইসিস, অপটিমাইজেশন, কম্বিনেটরিক্স মাঝেমধ্যে ভালো কাজে লাগে।

ডেটা সায়েন্স কমিউনিটি তৈরি

আমাদেরকেই তৈরি করতে হবে সেটা বাংলাদেশে। ভার্চুয়ালি বসার পাশাপাশি একটা ভেন্যুতে বসলে নিজেদের মধ্যে হৃদ্যতা বাড়বে।

ডেটা সায়েন্স/অ্যানালাইটিক্স শুরুতে প্রথম চারটা বাংলা বই

চিত্র ১১১: ডেটা নিয়ে শুরুর প্রথম চারটা বাংলা বই

‘ডেটা অ্যানালাইসিস’ নিয়ে কাজ করার একটা মজার ধাপ হচ্ছে ডেটা ভিজ্যুয়ালাইজেশন। আমাদের আশপাশের ডেটাকে কীভাবে ‘ভিজ্যুয়াল’ কনটেক্সটে ট্রান্সলেট করা যায় যাতে আমাদের মাথা এর ভেতরের প্যাটার্ন বুঝতে পারে। আর, আমাদের মাথা বুঝতে পারার অর্থ হচ্ছে, সেই ডেটা থেকে ‘অ্যাকশনেবল ইনসাইট’ ধরে সিদ্ধান্ত নিতে পারা। এতেই আমাদের কাজ হাসিল।

তবে, ডেটা ভিজ্যুয়ালাইজ করে সিদ্ধান্ত নিতে যাওয়ার আগে আমাদের দরকার ডেটা অ্যানালাইসিস, যার মধ্যে ডেটা কালেকশন, প্রসেসিং, মডেলিং ইত্যাদি দরকার। সব মিলিয়ে একটা বেসিক ধারণা নিয়ে এসেছে এ বইটি। অন্য ছয়টি বইয়ের মতো এ বইটিও (লেখা+ভিডিও) লেখকের ডেটার সাথে লম্বা জার্নির একটা ফলাফল।

● প্রবলেম সলভিং ● ডেটা স্টোরিটেলিং ● এক্সপ্লোরেটরি ডেটা অ্যানালাইসিস (ইডিএ) ● ডেটা ভিজ্যুয়ালাইজেশন ● ট্যাবলো ● মাইক্রোসফট পাওয়ার বিআই ● ডেটার বৈশিষ্ট্য ● ডেটাকে অর্গানাইজ করার ধারণা ● ডেটার গ্র্যানুলারিটি এবং ডিস্ট্রিবিউশন ● স্ট্রাকচার্ড কোয়েরি ল্যাঙ্গুয়েজ ● ওপেনসোর্স টুল ● গ্লোবাল সার্টিফিকেশন ● ডেটা সায়েন্স/অ্যানালাইটিক্স রোডম্যাপ

9 789849 760580

হাতেকলমে ডেটা অ্যানালাইটিক্স ও ভিজ্যুয়ালাইজেশন
রকিবুল হাসান
মূল্য : ৬৬৭.০০ টাকা
www.tamralipi.com