০০১-০০৪ Statistics এর ভূমিকা (পর্ব-৩)

স্ট্রাকচারের ভিত্তিতে ডেটার প্রকারভেদ:

স্ট্রাকচারের ভিত্তিতে ডেটা দুই প্রকারঃ

১) Unstructured data

2) Structured data

Unstructured data কী? আর Structured data কী?

যেসব ডেটা Standard format এ নাই, তারা হলো Unstructured data। আর যেসব ডেটা Standard format এ আছে, তারা হলো Strucutred data।

অনেক সময় বলা হয় যে, যেসব ডেটা pre-defined format এ নাই, তারা হলো Unstructured data।

উদাহরণঃ

একটা দোকানে কাস্টমাররা কি কি দ্রব্য খরিদ করল, তার ডেটাঃ

> **কাস্টমার ১: বিস্কুট, পেপসি, পেস্ট।**
> **কাস্টমার ২: পেপসি, চিপস।**
> **কাস্টমার ৩: টি-ব্যাগ, সুগার, বিস্কুট।**

এই ডেটা হলো Unstructured data, কারণ এই ডেটা Standard format এ নাই। এই ডেটা থেকে ইনফরমেশন বের করা কঠিন (যেমনঃ “কাস্টমার ১” তো বিস্কুট, পেপসি,

আর পেস্ট খরিদ করলো, তাতে ঐ কাস্টমার কত টাকা দোকানদার কে দিল, সেই ইনফরমেশন পাওয়া যাচ্ছে না।)

কিন্তু কাস্টমাররা কি কি দ্রব্য খরিদ করল, সেই ডেটা যদি এরকম হতোঃ

কাস্টমার ১	
দ্রব্য	**মূল্য**
বিস্কুট	৭০/=
পেপসি	২৫/=
পেস্ট	৮০/=

কাস্টমার ২	
দ্রব্য	**মূল্য**
পেপসি	২৫/=
চিপস	২০/=

কাস্টমার ৩	
দ্রব্য	**মূল্য**
টি-ব্যাগ	৮০/=
সুগার	৫০/=
বিস্কুট	৭০/=

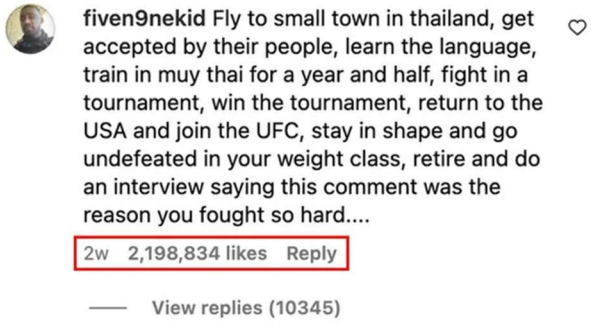
তখন এই ডেটা হবে Structured data, কারণ এই ডেটা Standard format এ আছে। এখান থেকে ইনফরমেশন বের করা যায়, যেমনঃ “কাস্টমার ১” যেসব দ্রব্য কিনলো, সেক্ষেত্রে তার কাছে দোকানদার পায় (৭০+২৫+৮০) টাকা = ১৭৫ টাকা (“/=” চিহ্ন দ্বারা টাকা বুঝানো হয়।)

Structured data এর আরেকটা উদাহরণ, এখানে ১০ জন শিক্ষার্থীর নাম, রোল আর মার্কস দেওয়া আছে।

	A	B	C
1	Name	Roll Number	Marks
2	Olivia C.	1	99
3	Alice J.	2	93
4	Frank Z.	3	91
5	Tom F.	4	91
6	Hannah X.	5	91
7	Ryan A.	6	80
8	Grace U.	7	85
9	Mia U.	8	75
10	Sophia Q.	9	83

এই ডেটা হলো Structured data কারণ এটা Standard format এ আছে। এখানে আমরা সহজে ইনফরমেশন পাচ্ছি। Structured data কে আরেকভাবে বলা যায় Tabulated data।

ইউটিউব, ফেসবুক, ইন্সটাগ্রাম – ইত্যাদির কমেন্টস হল Unstructured data। যেমনঃ



এটা হলো ইন্সটাগ্রামের সবচেয়ে জনপ্রিয় কমেন্ট। কমেন্টটা ছিলঃ

Fly to small town in thailand, get accepted by their people, learn the language, train in muy thai for a year and half, fight in a tournament, win the tournament, return to the USA and join the UFC, stay in shape and go undefeated in your weight class, retire and do an interview saying this comment was the reason you fought so hard....

এটা হল Unstructured data, কারণ এটা Structured form এ নাই। যদি আমরা এটাকে Structured form এ নিয়ে এসে Structured data করি, তাহলে এরকম হবে:

Objectives	**Where/What**
Fly	To a small town in Thailand
Get Accepted	By thai people
Learn	Thai language
Train	In Muy thai
Fight	In a tournament
Win	The tournament
Return	To USA
Join	The UFC
Stay	In shape
Go undefeated	In weight class

ঠিক এভাবে আমরা Unstructured data কে Structured data তে কনভার্ট করতে পারি।

ডেটাবেসে সংরক্ষিত তথ্য (ইনফরমেশন) তখনই কার্যকর হয় যখন আমরা সেই তথ্যের প্রাসঙ্গিকতা ও অর্থ বুঝতে পারি। যখন সংখ্যা বা টেক্সট এলোমেলোভাবে ছড়িয়ে থাকে এবং কোনো নির্দিষ্ট কাঠামো বা সংগঠন থাকে না, তখন সেই তথ্য থেকে কার্যকর সিদ্ধান্ত নেওয়া বা বিশ্লেষণ করা কঠিন হয়ে পড়ে। তাই, ডেটাবেসে তথ্যকে সুসংগঠিত ও কাঠামোবদ্ধভাবে সংরক্ষণ করা অত্যন্ত গুরুত্বপূর্ণ।

Dataset:

Dataset হলো Structured collection of data। যেমন:

Customer Name	Buying Amount	Total price (Tk)
Arif Rahman	2	200
Mojid Mollah	10	2660
Akash	6	1450
Karim Uddin	5	1330

এটা হলো ডেটাসেট। আর এটাই হলো Structured data.

Variables and Cases:

ভেরিয়েবল (Variable) কী?

স্ট্যাটিস্টিক্সে ভেরিয়েবল (Variable) হলো একটি বৈশিষ্ট্য বা গুণ যা পরিবর্তনশীল হতে পারে। এটি বিভিন্ন ব্যক্তি, বস্তু বা ঘটনাসমূহের মধ্যে পৃথক মান ধারণ করতে পারে।

উদাহরণ:

- একজন ছাত্রের বয়স, উচ্চতা, ওজন (যেগুলো ব্যক্তি ভেদে পরিবর্তিত হতে পারে)।
- কোনো পণ্যের দাম বা বিক্রির পরিমাণ (যেগুলো সময়ের সাথে পরিবর্তিত হতে পারে)।

ভেরিয়েবল দুই প্রকার:

- গুণগত বা ক্যাটাগরিকাল ভেরিয়েবল (Qualitative or Categorical Variable) – যেমন লিঙ্গ (পুরুষ/নারী), রক্তের গ্রুপ (A, B, O)।
- পরিমাণগত বা নিউমেরিক্যাল ভেরিয়েবল (Quantitative or Numerical Variable) – যেমন বয়স (২০ বছর), উচ্চতা (৫.৬ ফুট)।

কেস (Case) কী?

স্ট্যাটিস্টিক্সে কেস (Case) বলতে এমন একটি একক বা অবজেক্টকে বোঝানো হয়, যার জন্য আমরা তথ্য সংগ্রহ করি। প্রতিটি কেসের এক বা একাধিক ভেরিয়েবল থাকতে পারে।

উদাহরণ:

- একটি স্কুলের ছাত্রদের নিয়ে গবেষণা করা হলে, প্রতিটি ছাত্র একটি কেস হবে। ডেটাসেটে ছাত্রদের নাম হবে Case।
- একটি কোম্পানির বিভিন্ন পণ্যের বিক্রয় বিশ্লেষণ করলে, প্রতিটি পণ্য একটি কেস হবে। ডেটাসেটে পণ্যের নাম হবে Case

কেস সাধারণত সারি (Row) আকারে উপস্থাপন করা হয়, যেখানে প্রতিটি কেসের জন্য বিভিন্ন ভেরিয়েবলের মান লিপিবদ্ধ থাকে।

উদাহরণ:

ব্যক্তি (কেস)	বয়স (ভেরিয়েবল)	উচ্চতা (ভেরিয়েবল)	লিঙ্গ (ভেরিয়েবল)
রাহুল	২০ বছর	৫.৭ ফুট	পুরুষ
সোহা	২২ বছর	৫.৩ ফুট	নারী
হাসান	২৫ বছর	৫.৯ ফুট	পুরুষ

এখানে প্রতিটি সারি একটি "কেস", এবং প্রতিটি কলাম একটি "ভেরিয়েবল"।

এখানে “রাহুল” হোলো কেস, কারণ আমরা রাহুল কে ফোকাস করে তার বয়স, উচ্চতা আর লিঙ্গ বের করছি। একই জিনিসটা সোহা, হাসানের ক্ষেত্রেও যায়। কাজেই “সোহা”, “হাসান” হলো কেস (Case)

Constant (ধ্রুবক):

Constant হলো যা অপরিবর্তনীয়। যেমন:-

	A	B	C	D
1	Name	Roll Number	Marks	Fee Payment
2	Olivia C.	1	99	2000
3	Alice J.	2	93	2000
4	Frank Z.	3	91	2000
5	Tom F.	4	91	2000
6	Hannah X.	5	91	2000
7	Ryan A.	6	80	2000
8	Grace U.	7	85	2000
9	Mia U.	8	75	2000
10	Sophia Q.	9	83	2000

এখানে Fee Payment হলো constant, কারণ সবার Fee Payment একই।

শূণ্য বনাম খালি:

মনে করেন, ১০ মার্কের Quiz পরীক্ষার রেজাল্ট দেওয়া হয়েছে। দেখা গেছে, সেখানে আফ্রিদি পেয়েছে ৯, লাবিবা পেয়েছে ৭, রাকিব পেয়েছে ০, আর মুহিতের মার্ক আসেনি।

নাম	মার্কস (১০)
আফ্রিদি	৯
রাকিব	০
মুহিত	---
লাবিবা	৭

মুহিতের মার্ক খালি মানে মুহিত শূণ্য পায়নি। শূণ্য আর খালি এক জিনিস নয়। মুহিত পরীক্ষা দেয়নি, তাই মুহিতের মার্কস খালি। কিন্তু রাকিব তো পরীক্ষা দিয়েছে, সেই পরীক্ষায় শূণ্য পেয়েছে।

অনেক সময় খালি ডেটা কে “Not available”, “N/A” দ্বারা প্রকাশ করা হয়। খালি ডেটা কে Not available ডেটা বলে। যে ডেটা আমরা পাইনি, সেটাকে Not available ডেটা বা খালি ডেটা বলে।

[রেফারেন্সঃ]

1) “W1_L2_Introduction and types of Data - Understanding data” from IIT Madras in YouTube: https://youtu.be/5e5nHjXJtUg?si=bamKKYvlSvIdIhPC
2) ChatGPT help