০০১-০০৭ Statistics এর ভূমিকা (পর্ব-২)

পাইথনের বেসিক কোড শেষ করে আমরা Statistics এর ভূমিকা পর্ব-২ ধরেছি। আমরা পরবর্তীতে বারবার Statistics এর সাথে পাইথন ধরবো।

ডেটাসেট উপস্থাপনাঃ

আমরা "Statistics এর ভূমিকা পর্ব-১" এ ডেটাসেট আর কেস নিয়ে আলোচনা করেছি। এবার এর উপস্থাপনা নিয়ে আলোচনা করবো। ডেটাসেট উপস্থাপনা এর ক্ষেত্রে row আর column দুইটি অংশ থাকে। row কে অনেক সময় tuple বলা হয়; এবং column কে অনেক সময় attribute বলা হয়। row এর ক্ষেত্রে প্রত্যেকটা কেসের জন্য একই column বা attribute থাকে। যেমনঃ

	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

এই ডেটাসেট টি দেখুন, এখানে "Olivia C." কেসের জন্য Roll Number, Marks, Fee Payment আছে। বাকিদের জন্য কিন্তু সেই একই column প্রযোজ্য হয়েছে। বাকিদের জন্য ডিফারেন্ট কিছু column দেওয়া হয়নি। সুতরাং প্রত্যেকটা কেসের জন্য একই column থাকবে।

আবার column এর ক্ষেত্রে প্রত্যেকটা variable একই ধরনের হতে হবে, যেমনঃ

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 রাহুলের উচ্চতা "৫.৭ ফুট"। আবার সোহার উচ্চতা "৫.৩ ফুট"।

খেয়াল করে দেখেন, এখানে রাহুলের উচ্চতাকে ফুট আর সোহার উচ্চতা কে সেন্টিমিটারে করা হয়নি। দুইজনেরই ফুট দিয়েই হিসাব করা হয়েছে, কারণ প্রত্যেকটা column variable এ দুইধরনের ইউনিট থাকা অসম্ভব। আবার আমরা আরো বিষয়টা লক্ষ্য করি, রাহুলের উচ্চতা column এ আমরা তার উচ্চতা ব্যতীত ফোন নাম্বার, ঠিকানা দেইনি, কারণ সেটা allowed নয়। এইজন্যেই column এর ক্ষেত্রে প্রত্যেকটা variable একই ধরনের হতে হবে।

আবার মনে করিয়ে দিলাম, Column তিন ধরনের, যথাঃ case, variable, constant।

তো row আর column এর গুরুত্বপূর্ণ শর্ত হলো, প্রত্যেকটা observation এর জন্য নিজস্ব row থাকতেই হবে, আর প্রত্যেকটা ভেরিয়েবলের জন্য নিজস্ব column থাকতে হবে।

Data এর ক্লাসিফিকেশনঃ

ক্লাসিফিকেশনের ক্ষেত্রে ডেটা দুই প্রকার। যথাঃ ১) categorical ২) numerical

a) Categorical data: Categorical data কে অনেক সময় qualitative variables বলা হয়। এটা লিখিত আকারে group membership কে আইডেন্টিফাই করা হয়। অর্থাৎ string এর ব্যবহার হয় এখানে, এবং এই ডেটার সর্বোচ্চ, সর্বনিম্ন, গড়, মোট, যোগ, বিয়োগ করা যায় না। নাম, ঠিকানা, লিঙ্গ – এরা হলো Categorical data।

b) Numerical data: Numerical data কে অনেক সময় বলা হয় quantitative variables। এটাকে সংখ্যায় প্রকাশ করা হয় এবং প্রত্যেকটা কেসের numeric property দেওয়া হয়। Numerical data এর ক্ষেত্রে কখনো কখনো measurement unit থাকতেও পারে, আবার নাও থাকতে পারে। যেমনঃ মানুষের উচ্চতা প্রকাশ করা হয় numerical data অনুযায়ী, তখন ফুট, সেন্টিমিটার unit

ব্যবহার হয়; কিন্তু পরীক্ষার মার্কস এর ক্ষেত্রে numerical data ব্যবহার করা হয়, কিন্তু এর কোনো unit নেই। এই ডেটার সর্বোচ্চ, সর্বনিম্ন, গড়, মোট, যোগ, বিয়োগ করা যায়।

Categorical data আর Numerical Data এর উদাহরণ দেওয়া হলোঃ

নাম	ডিপার্টমেন্ট	উচ্চতা	মার্কস
আসিফ	CSE	5.7 ফুট	70
রাতুল	BBA	5.3 ফুট	75
মুহিত	BBA	5.4 ফুট	60
জাহিদ	EEE	5.5 ফুট	61
রাফি	CSE	5.2 ফুট	90
মজিদ	EEE	5.6 ফুট	95
শরীফ	ME	5.4 ফুট	80

এখানে “নাম”, “ডিপার্টমেন্ট” – এরা হলো categorical data। আর “উচ্চতা” হলো numerical data, যার সাথে unit আছে। আর “মার্কস” হলো numerical data, যার সাথে unit নেই। “নাম” আর “ডিপার্টমেন্ট” এ string ব্যবহার করা হয়েছে, এবং এক্ষেত্রে আমরা সেখানে সর্বোচ্চ, সর্বনিম্ন, গড়, মোট, যোগ, বিয়োগ বের করতে পারবো না। কিন্তু “উচ্চতা” আর “মার্কস” এর ক্ষেত্রে আমরা সর্বোচ্চ, সর্বনিম্ন, গড়, মোট, যোগ, বিয়োগ বের করতে পারবো সহজেই।

মনে রাখতে হবে, মোবাইল নম্বর numerical data নয়, কারণ এখানে 0 আগে আসে, এবং আমরা এর সর্বোচ্চ, সর্বনিম্ন, গড়, মোট, যোগ, বিয়োগ করি না। যেমনঃ 01745..... এখানে 0 আগে এসেছে। অর্থাৎ সকল সংখ্যা numerical data নয়, কিন্তু সকল numerical data

সংখ্যা হিসেবে থাকে। তো প্রশ্ন জাগতে পারে যে মোবাইল নম্বর কি ডেটা তাহলে? মোবাইল নম্বর হলো nominal data।

অনেক সময় unit দেওয়া numerical data এর column heading এ প্রথম বন্ধনী দিয়ে unit লেখা থাকে। যেমনঃ

নাম	ডিপার্টমেন্ট	উচ্চতা (ফুট)	মার্কস
আসিফ	CSE	5.7	70
রাতুল	BBA	5.3	75
মুহিত	BBA	5.4	60
জাহিদ	EEE	5.5	61
রাফি	CSE	5.2	90
মজিদ	EEE	5.6	95
শরীফ	ME	5.4	80

এখানে উচ্চতা column এ "(ফুট)" লিখে unit দেওয়া হয়েছে।

Cross-sectional data আর Time-sectional data:

1) Time-sectional data: Time-sectional data হলো যে ডেটা সময়ের সাথে সাথে পর্যায়ক্রমে সংগ্রহ করা হয়। এটি সময়ের পরিবর্তনের সাথে ডাটার প্রবণতা বিশ্লেষণ করতে সাহায্য করে।

   যেমনঃ নিচে Asus কোম্পানির ল্যাপটপ উৎপাদনের ডেটাসেট দেখানো হলোঃ-

তারিখ	ল্যাপটপ উৎপাদন
১ জুন	২৪০০
২ জুন	২৩০০
৩ জুন	২৯০০
৪ জুন	৫০০০

এখানে সময়ের সাথে সাথে ল্যাপটপ উৎপাদন দেখানো হয়েছে। যেমনঃ ১ জুন, ২ জুন, ৩ জুন এভাবে।

2) Cross-Sectional Data: Cross-Sectional Data হলো যে ডেটা একটি নির্দিষ্ট সময়ে সংগ্রহ করা হয়। এটি সময়ের সাথে পরিবর্তন হয় না, বরং একটি নির্দিষ্ট মুহূর্তের চিত্র তুলে ধরে। যেমনঃ ধরে নিই, ২০২৪ সালে ল্যাপটপ কোম্পানি কতগুলো ল্যাপটপ উৎপাদন করেছিলোঃ-

কোম্পানি	ল্যাপটপ উৎপাদন
Lenovo	১ বিলিয়ন
Acer	১.৩ বিলিয়ন
Toshiba	২ বিলিয়ন
Asus	৫ বিলিয়ন

এখানে ২০২৪ সাল কে স্থির রেখে প্রত্যেকটা কোম্পানির ল্যাপটপ উৎপাদন দেখানো হয়েছে।

Scale of Measurement:

ডাটার ধরণ ও তাদের পরিমাপের ভিত্তিতে শ্রেণিবিন্যাসের পদ্ধতিকে Scale of Measurement বলা হয়। এটি চার ভাগে বিভক্ত: Nominal, Ordinal, Interval, এবং Ratio।

1) Nominal Scale: এখানে ডাটাকে শুধুমাত্র আইডেন্টিফাই বা লেবেল করা হয়। এর গাণিতিক কোনো মান থাকে না। শুধুমাত্র নাম বা শ্রেণি বোঝানো হয়। এবং তুলনা করা সম্ভব নয়।

উদাহরণ:

- **লিঙ্গ:** পুরুষ, নারী, অন্যান্য
- **রক্তের গ্রুপ:** A+ve, B+ve, AB+ve, O+ve, A-ve, B-ve, AB-ve, O-ve
- **মানুষের নাম।**

নাম	লিংগ	রক্তের গ্রুপ
মুসা	পুরুষ	A+ve
প্রভা	নারী	B+ve
জাহিদ	পুরুষ	A-ve

এখানে "মুসা" হলো ক্যারেক্টার, তার নাম তাকে আইডেন্টিফাই করে। তার লিঙ্গ হলো পুরুষ, এটা তাকে লেবেল করা হয়েছে

2) Ordinal Scale: এখানে ডাটাকে একটি নির্দিষ্ট ক্রম বা sequence অনুসারে সাজানো হয়ে থাকে। এক্ষেত্রে তুলনা করা সম্ভব।

উদাহরণ:

- **পরীক্ষার ফলাফল:** প্রথম (১ম), দ্বিতীয় (২য়), তৃতীয় (৩য়)।
- **গ্রাহক সন্তুষ্টি:** খারাপ, গড়পড়তা, ভালো, খুব ভালো, অসাধারণ।
- **সামরিক পদমর্যাদা:** ক্যাপ্টেন, মেজর, কর্নেল, জেনারেল।

যেমন, কোনো ওয়েবসাইটে user এর app নিয়ে ফিডব্যাক।

ইউজার	ফিডব্যাক
ইউজার_১২৫	ভালো
ইউজার_০৪৫	ভালো
ইউজার_৯০০	মোটামুটি
ইউজার_৩৩৩	খারাপ

এখানে ফিডব্যাক হলো ordinal। কারণ আমরা এখানে ভালো, মোটামুটি, খারাপ – এগুলোর তুলনা করতে পারছি।

ভালো > মোটামুটি > খারাপ

3) Interval Scale: এখানে ডাটার ক্রমবিন্যাস ও নির্দিষ্ট পার্থক্য থাকে। এই ডেটা সংখ্যা দ্বারা প্রকাশ করা হয়। **এখানে শূন্য প্রকৃত নয়**। যেমনঃ সেলসিয়াস বা ফারেনহাইট স্কেলে তাপমাত্রা (এখানে 0°C মানে তাপমাত্রার অভাব বোঝায় না)। আরো একটা ভালো উদাহরণ দিলে বুঝতে পারবেন।

গ্রহ	তাপমাত্রা
বুধ	167°C
পৃথিবী	15°C
জুপিটার	-110°C
নেপচুন	-200°C

এখানে তাপমাত্রা ধ্বনাত্মক আছে, ঋণাত্মক আছে। প্রকৃত শূন্য না থাকার মানে হল সেই ডেটা ধ্বনাত্মক হতে পারে, ঋণাত্মক হতে পারে। প্রকৃত শূন্যের ক্ষেত্রে শূন্য মানে কোনো কিছুর অনুপস্থিতি বুঝায় না। যেমনঃ 0°C মানে তাপমাত্রার অভাব বুঝায় না, কারণ এর চেয়েও নিম্ন তাপমাত্রার অস্তিত্ব আছে, যেমনঃ জুপিটারের তাপমাত্রা -110°C

সোর্সঃ https://science.nasa.gov/resource/solar-system-temperatures/

4) Ratio Scale: এখানে ডাটার ক্রমবিন্যাস ও নির্দিষ্ট পার্থক্য, প্রকৃত শূন্য থাকে। যেমনঃ উচ্চতা *(০ ফুট মানে উচ্চতার সম্পূর্ণ অনুপস্থিতি)*, ওজন *(০ কেজি মানে ওজনের অনুপস্থিতি)*। এছাড়া আমরা বাস্তবে কখনই বলি না -৫ কেজি, -১০ কেজি, -১০ ফুট। কারণ এখানে ০ ফুট মানে উচ্চতার সম্পূর্ণ অনুপস্থিতি।

আয়, বিক্রয়, বয়স এরাও Ratio Scale এর মধ্যে পড়ে। কোনো কিছুর মূল্য, বস্তুর পরিমাণ – এরাও Ratio Scale এর মধ্যে পড়ে।

এখানে কতকগুলো পেন্সিল বাক্স আর তার মধ্যে পেন্সিল কতগুলো আছে, তা দেখানো হলঃ-

পেন্সিল বাক্স	পেন্সিলের পরিমাণ
বাক্স-1	0
বাক্স-2	10
বাক্স-3	15
বাক্স-4	11

এখানে পেন্সিলের পরিমাণ Ratio Scale এর মধ্যে পড়ে। বাক্স-1 এ পেন্সিলের পরিমাণ শূন্য, এর মানে সেখানে কোনো পেন্সিল নেই। আর পেন্সিলের পরিমাণ কখনই ঋণাত্মক হয় না। অর্থাৎ পেন্সিলের পরিমাণের ক্ষেত্রে প্রকৃত শূন্য বজায় রয়েছে।

তো আমাদের “Statistics এর ভূমিকা” সম্পূর্ণভাবে শেষ হয়েছে। পরবর্তীতে আমরা ডেটা সায়েন্সের জন্য Statistics এর আরও বিষয় জানবো।