EDA (Exploratory Data Analysis) বা এক্সপ্লোরেটরি ডেটা অ্যানালাইসিসের একটি অত্যন্ত গুরুত্বপূর্ণ অংশ ইউনিভ্যারিয়েট অ্যানালাইসিস **(Univariate Analysis)** সম্পর্কে বিস্তারিত আলোচনা . ইউনিভ্যারিয়েট অ্যানালাইসিস মানে হলো একটি ডেটাসেটের প্রতিটি কলাম বা ভেরিয়েবলকে অন্যদের ওপর নির্ভর না করে স্বাধীনভাবে এবং আলাদা আলাদাভাবে বিশ্লেষণ করা।

১. ইউনিভ্যারিয়েট অ্যানালাইসিস কী?

'ইউনি' (Uni) মানে হলো এক এবং 'ভ্যারিয়েট' (Variate) মানে হলো ভেরিয়েবল বা কলাম। সুতরাং, যখন আপনি একটি ডেটাসেটের প্রতিটি কলামকে আলাদাভাবে ধরে সেটির প্রকৃতি বোঝার চেষ্টা করেন, তখন তাকে ইউনিভ্যারিয়েট অ্যানালাইসিস বলা হয়। এটি করার মূল উদ্দেশ্য হলো ডেটার ভেতরে লুকিয়ে থাকা তথ্যগুলো বের করে আনা এবং প্রতিটি কলামকে ভেতর-বাহির থেকে চেনা।

২. ডেটার প্রকৃতি বোঝা

যেকোনো বিশ্লেষণ শুরু করার আগে ডেটার ধরন জানতে হয়। ডেটা মূলত দুই প্রকারের হয়:

- ক্যাটেগরিকাল ডেটা **(Categorical Data):** যে ডেটাগুলো বিভিন্ন বিভাগ বা শ্রেণিতে বিভক্ত থাকে। যেমন- জেন্ডার (পুরুষ/মহিলা), টাইটানিকের প্যাসেঞ্জার ক্লাস (১ম, ২য় বা ৩য় শ্রেণি), বা কোন শহর থেকে যাত্রী উঠেছিল।
- নিউমেরিক্যাল ডেটা **(Numerical Data):** যে ডেটাগুলো সংখ্যায় প্রকাশ করা যায় এবং যার ওপর গাণিতিক হিসাব করা সম্ভব। যেমন- বয়স, টিকিটের দাম (Fare), বা উচ্চতা।

৩. ক্যাটেগরিকাল ডেটা বিশ্লেষণের পদ্ধতি

ক্যাটেগরিকাল কলামগুলোর তথ্য দেখার জন্য দুটি প্রধান উপায় দেখানো হয়েছে:

- কাউন্ট প্লট **(Count Plot):** এটি একটি কলামের প্রতিটি ক্যাটেগরি কতবার এসেছে তা বার গ্রাফের মাধ্যমে দেখায়। যেমন- টাইটানিকে কতজন বেঁচেছিলেন (১) এবং কতজন মারা গিয়েছিলেন (০), তা কাউন্ট প্লট দিয়ে সহজে বোঝা যায়। এটি মূলত ডেটার ফ্রিকোয়েন্সি বা সংখ্যা গণনার কাজ করে।
- পাই চার্ট **(Pie Chart):** যদি আপনি জানতে চান প্রতিটি ক্যাটেগরি পুরো ডেটার কত শতাংশ (Percentage) জায়গা জুড়ে আছে, তবে পাই চার্ট ব্যবহার করা হয়। যেমন- মোট যাত্রীর কত শতাংশ পুরুষ ছিল বা কত শতাংশ যাত্রী ৩য় শ্রেণিতে ভ্রমণ করছিল।

৪. নিউমেরিক্যাল ডেটা বিশ্লেষণের পদ্ধতি

সংখ্যাবাচক তথ্যের বিন্যাস এবং প্যাটার্ন বোঝার জন্য নিচের প্রযুক্তিগুলো ব্যবহার করা হয়:

- হিস্টোগ্রাম **(Histogram):** এটি পুরো ডেটাকে ছোট ছোট রেঞ্জ বা 'বিন' (Bin)-এ ভাগ করে দেখায় যে কোন রেঞ্জে কত বেশি তথ্য আছে। যেমন- ২০ থেকে ৩০ বছর বয়সী যাত্রীর সংখ্যা কেমন ছিল, তা হিস্টোগ্রাম দিয়ে এক পলকে বোঝা যায়।
- ডিস্টপ্লট বা পিডিএফ **(Distplot/PDF):** এটি একটি কার্ভ বা বক্ররেখার মাধ্যমে ডেটার বিন্যাস দেখায়। এর মাধ্যমে বোঝা যায় কোনো নির্দিষ্ট মানের হওয়ার সম্ভাবনা কতটুকু। এখান থেকেই জানা যায় ডেটাটি কি সিমেট্রিকাল (মাঝখানে বেশি এবং দুই পাশে সমান) নাকি স্কিউড (একদিকে বেশি ঝুঁকে আছে)।

- বক্সপ্লট **(Boxplot):** এটি ডেটার ৫-নম্বর সামারি প্রদান করে (সর্বনিম্ন, সর্বোচ্চ, মিডিয়ান এবং প্রথম ও তৃতীয় কোয়ার্টাইল)। এর সবচেয়ে বড় কাজ হলো আউটলায়ার **(Outliers)** খুঁজে বের করা। আউটলায়ার হলো এমন অস্বাভাবিক মান যা সাধারণ ডেটার প্যাটার্ন থেকে অনেক দূরে থাকে (যেমন- ১০০ বছর বয়সী যাত্রী বা অস্বাভাবিক বেশি দামের টিকিট)।

৫. গাণিতিক ও পরিসংখ্যানগত বিশ্লেষণ

গ্রাফের পাশাপাশি ডেটাকে গাণিতিকভাবেও বিশ্লেষণ করা হয়:

- **Mean, Median, Max, Min:** ডেটার গড়, মধ্যক এবং সর্বনিম্ন ও সর্বোচ্চ সীমা বের করা।
- স্কিউনেস **(Skewness):** ডেটা কি সুষমভাবে সাজানো নাকি একদিকে বেশি ঝুঁকে আছে তা গাণিতিকভাবে বের করা। যদি স্কিউনেস '০' হয় তবে সেটি একদম সিমেট্রিকাল, পজিটিভ হলে সেটি ডান দিকে ঝুলে আছে এবং নেগেটিভ হলে সেটি বাম দিকে ঝুলে আছে।