নিচে মূল আলোচনা সহজভাবে বুঝিয়ে বলা হলো:

১. ডেটা সেটটি কত বড়? (How big is the data?)

প্রথমেই জানতে হবে আপনার ডেটাতে কতগুলো তথ্য বা সারি (Rows) এবং কতগুলো বৈশিষ্ট্য বা কলাম (Columns) আছে। এটি জানার জন্য আপনি `df.shape` কোডটি ব্যবহার করতে পারেন। এটি আপনাকে ডেটার আকার সম্পর্কে একটি পরিষ্কার ধারণা দেবে।

২. ডেটা দেখতে কেমন? (How does the data look?)

ডেটা সরাসরি দেখার জন্য আমরা সাধারণত `df.head()` ব্যবহার করি, যা প্রথম ৫টি সারি দেখায়। তবে একটি ভালো পরামর্শ দেওয়া হয়েছে যে, কেবল প্রথম ৫টি সারি না দেখে `df.sample(5)` ব্যবহার করা ভালো। এটি আপনাকে পুরো ডেটাসেট থেকে দৈবচয়ন বা রেন্ডমভাবে ৫টি সারি দেখাবে, যার ফলে ডেটাতে কোনো পক্ষপাত বা বায়াস (Bias) থাকলে তা সহজে ধরা পড়বে।

৩. কলামগুলোর ডেটা টাইপ কী কী? (What are the data types of columns?)

প্রতিটি কলামের ডেটা টাইপ (যেমন- সংখ্যা, শব্দ বা অবজেক্ট) জানা খুব জরুরি। এটি জানতে `df.info()` ফাংশনটি ব্যবহার করা হয়। এখান থেকে আপনি কলামগুলো নিউমেরিক্যাল নাকি ক্যাটাগরিক্যাল তা বুঝতে পারবেন। এছাড়াও এটি ডেটা কতটুকু মেমরি দখল করে আছে তাও জানায়। অনেক সময় অপ্রয়োজনীয় বড় ডেটা টাইপ (যেমন Float) কমিয়ে ছোট টাইপে (Int) রূপান্তর করলে মেমরি সাশ্রয় হয় এবং মডেল দ্রুত কাজ করে।

৪. ডেটাতে কোনো মিসিং ভ্যালু আছে কি? (Are there any missing values?)

ডেটাতে কোনো ফাঁকা ঘর বা মিসিং ভ্যালু থাকলে তা মডেলের পারফরম্যান্সে সমস্যা করতে পারে। `df.isnull().sum()` ব্যবহার করে প্রতিটি কলামে কতগুলো মিসিং ভ্যালু আছে তা বের করা যায়। যদি কোনো কলামে অনেক বেশি মিসিং ভ্যালু থাকে (যেমন টাইটানিক ডেটাসেটের 'Cabin' কলাম), তবে সেই কলামটি বাদ দেওয়া বা অন্য কোনো মান দিয়ে পূরণ করার সিদ্ধান্ত নিতে হয়।

৫. গাণিতিক দিক থেকে ডেটা কেমন? (How does the data look mathematically?)

নিউমেরিক্যাল কলামগুলোর একটি সারসংক্ষেপ পাওয়ার জন্য `df.describe()` ফাংশনটি ম্যাজিকের মতো কাজ করে। এটি আপনাকে ডেটার গড় (Mean), স্ট্যান্ডার্ড ডেভিয়েশন (Standard Deviation), সর্বনিম্ন ও সর্বোচ্চ মান এবং কোয়ার্টাইল ভ্যালুগুলো জানিয়ে দেয়। এর মাধ্যমে আপনি ডেটাতে কোনো অসংগতি বা অস্বাভাবিকতা আছে কি না তা সহজে ধরতে পারেন।

৬. কোনো ডুপ্লিকেট সারি আছে কি? (Are there duplicate values?)

একই তথ্য বারবার থাকলে মডেল ভুল শিখতে পারে। তাই `df.duplicated().sum()` ব্যবহার করে দেখা উচিত ডেটাতে কোনো ডুপ্লিকেট সারি আছে কি না। থাকলে সেগুলো ফেলে দেওয়া উচিত।

৭. কলামগুলোর মধ্যে সম্পর্ক বা কো-রিলেশন কেমন? (How is the correlation between columns?)

সবশেষে গুরুত্বপূর্ণ হলো কলামগুলোর একে অপরের সাথে সম্পর্ক বোঝা। এটি বের করার জন্য `df.corr()` ফাংশন ব্যবহার করা হয়। এটি -১ থেকে +১ এর মধ্যে মান দেয়।

- যদি মান পজিটিভ হয়, তবে একটি বাড়লে অন্যটিও বাড়বে (যেমন- টাইটানিক ডেটাসেটে 'Fare' বা ভাড়া বাড়লে বেঁচে যাওয়ার সম্ভাবনা বা 'Survived' বাড়ার সম্পর্ক আছে)।
- যদি মান নেগেটিভ হয়, তবে একটি বাড়লে অন্যটি কমবে।
- যদি মান ০ এর কাছাকাছি হয়, তবে তাদের মধ্যে কোনো সম্পর্ক নেই (যেমন- প্যাসেঞ্জার আইডি-র সাথে বেঁচে যাওয়ার কোনো সম্পর্ক নেই)।

উপসংহার: এই সাতটি প্রশ্নের উত্তর জানলে আপনার ডেটা সম্পর্কে একটি শক্তিশালী প্রাথমিক ধারণা তৈরি হবে। পরে **EDA (Exploratory Data Analysis)** সম্পর্কে আরও বিস্তারিত আলোচনা করা হবে যেখানে গ্রাফ এবং চার্টের মাধ্যমে ডেটাকে বিশ্লেষণ করা শিখবেন।