১। কোন ডেটাফ্রেমে আমরা Regression (Linear & Multiple) ব্যবহার করব?

Regression সাধারণত সংখ্যাসূচক (numerical) ডেটার উপর প্রয়োগ করা হয়, যেখানে এক বা একাধিক স্বাধীন চলক (independent variables) এবং একটি নির্ভরশীল চলক (dependent variable) থাকে। ডেটাফ্রেমে নিম্নলিখিত বৈশিষ্ট্য থাকতে হবে:

- **সংখ্যাসূচক ফিচার**: Regression মডেল সাধারণত সংখ্যাসূচক ডেটার সাথে কাজ করে, তবে ক্যাটাগরিক্যাল ডেটা থাকলে One-Hot Encoding বা Label Encoding করে ব্যবহার করা যেতে পারে।
- **সম্পর্ক থাকা**: নির্ভরশীল চলক ও স্বাধীন চলকের মধ্যে সম্পর্ক থাকতে হবে (যেমন: লিনিয়ার বা নন-লিনিয়ার)।
- **সঠিক ডেটা**: Outlier, missing values বা অত্যাধিক noisy ডেটা থাকলে প্রি-প্রসেসিং প্রয়োজন।

২। Simple Linear Regression প্রয়োগের জন্য ডেটাফ্রেমের কী বৈশিষ্ট্য থাকতে হবে?

Simple Linear Regression ব্যবহার করতে হলে ডেটাফ্রেমে থাকতে হবে:

- **একটি স্বাধীন চলক (X) ও একটি নির্ভরশীল চলক (Y)**
- **স্বাধীন চলক ও নির্ভরশীল চলকের মধ্যে লিনিয়ার সম্পর্ক থাকতে হবে**
- **স্বাধীন চলকের মান পরিবর্তনের সাথে নির্ভরশীল চলকের সরলরেখার মতো পরিবর্তন হবে (Linear Trend)**
- **ডেটা স্বাভাবিকভাবে বিতরণকৃত (Normally Distributed) হওয়া উচিত, তবে স্কেলিং বা ট্রান্সফরমেশন প্রয়োগ করা যেতে পারে**

৩। Multiple Linear Regression কখন ব্যবহার করা হয়, এবং এর জন্য ডেটাফ্রেম কেমন হওয়া উচিত?

- **কখন ব্যবহার করা হয়?**
  যখন নির্ভরশীল চলকের উপর একাধিক স্বাধীন চলকের প্রভাব থাকে, তখন Multiple Linear Regression ব্যবহার করা হয়।
  উদাহরণ: একটি বাড়ির দাম ভবনের আকার, লোকেশন, বয়স, সুবিধাসমূহের উপর নির্ভরশীল হতে পারে।
- **ডেটাফ্রেমের বৈশিষ্ট্য:**
  - একাধিক স্বাধীন চলক (X1, X2, X3, ...) এবং একটি নির্ভরশীল চলক (Y) থাকতে হবে।
  - স্বাধীন চলকগুলোর মধ্যে Collinearity কম থাকা উচিত।
  - Outliers ও Missing Values প্রি-প্রসেস করা উচিত।
  - Multicollinearity চেক করার জন্য Variance Inflation Factor (VIF) ব্যবহার করা যেতে পারে।

৪। Regression মডেলের জন্য Feature Selection কেন গুরুত্বপূর্ণ?

Feature Selection গুরুত্বপূর্ণ কারণ:

- **Overfitting কমায়**: অপ্রয়োজনীয় বা সম্পর্কহীন ফিচার থাকলে মডেল অপ্রাসঙ্গিক প্যাটার্ন শিখতে পারে, যা জেনারালাইজেশন কমিয়ে দেয়।

- **পারফরম্যান্স বাড়ায়**: কম এবং প্রাসঙ্গিক ফিচার ব্যবহার করলে মডেল দ্রুত ও কার্যকরভাবে প্রশিক্ষিত হয়।
- **Collinearity সমস্যা কমায়**: Highly correlated ফিচার থাকলে মডেলের অনুমান ক্ষমতা দুর্বল হতে পারে।

Feature Selection এর জন্য **Correlation Matrix, Recursive Feature Elimination (RFE), Lasso Regression** ইত্যাদি পদ্ধতি ব্যবহার করা যেতে পারে।

৫। Collinearity কীভাবে Multiple Linear Regression মডেলের পারফরম্যান্সে প্রভাব ফেলে?

Collinearity ঘটে যখন স্বাধীন চলকগুলোর মধ্যে উচ্চ মাত্রার পারস্পরিক সম্পর্ক থাকে, যা নিম্নলিখিত সমস্যাগুলো তৈরি করতে পারে:

- **অস্থির কোফিসিয়েন্ট (Unstable Coefficients)**: মডেল ফিচারগুলোর ওজন ঠিকমতো নির্ধারণ করতে পারে না, ফলে কোফিসিয়েন্টের মান পরিবর্তনশীল হয়।
- **বিশ্বাসযোগ্যতা কমে যায়**: মডেল নতুন ডেটার উপর দুর্বলভাবে পারফর্ম করে (Poor Generalization)।
- **Feature Importance বোঝা কঠিন হয়**: কোন ফিচার কতটা গুরুত্বপূর্ণ তা নির্ধারণ করা কঠিন হয়ে যায়।

Collinearity চেক করার জন্য **Variance Inflation Factor (VIF) এবং Pearson Correlation Matrix** ব্যবহার করা হয়।

৬। কোন ধরনের ডেটার উপর Regression মডেল প্রয়োগ করা উচিত?

Regression মডেল সাধারণত নিম্নলিখিত ডেটার উপর প্রয়োগ করা উচিত:

- **সংখ্যাসূচক (Numerical) ডেটা**: যেমন বিক্রয় মূল্য, তাপমাত্রা, উচ্চতা, ওজন ইত্যাদি।
- **সুনির্দিষ্ট সম্পর্কযুক্ত ডেটা**: যেখানে স্বাধীন চলক ও নির্ভরশীল চলকের মধ্যে কোনো সম্পর্ক রয়েছে।
- **সমান্তরাল বিতরণকৃত ডেটা**: Extreme Outliers বা Skewness কম থাকা ভালো।
- **ধারাবাহিক ডেটা (Continuous Data)**: যেমন সময়ের উপর ভিত্তি করে বিক্রয় পরিমাণ বা তাপমাত্রার পরিবর্তন।

যদি ক্যাটাগরিক্যাল ডেটা থাকে, তবে One-Hot Encoding বা Label Encoding করে সংখ্যা হিসেবে উপস্থাপন করতে হয়