Interview Questions:

Q1.What are the different types of missing data?
Ans. There are three main types of missing data in statistics and data analysis:

1. MCAR (Missing Completely at Random)
The missingness is completely unrelated to any data (observed or unobserved).
Example: A survey respondent skips a question by accident.
Impact: Least problematic; does not bias results.

2. MAR (Missing at Random)
The missingness is related to observed data, but not to the missing value itself.
Example: Younger people are less likely to report income, but age is known.
Impact: Can be handled well with techniques like imputation using other features.

3. MNAR (Missing Not at Random)
The missingness is related to the missing value itself.
Example: People with very high incomes choose not to disclose their income.
Impact: Most difficult to handle; may introduce bias unless modeled explicitly.

Q 2.How do you handle categorical variables?
Ans. Categorical variables in a pandas DataFrame must be converted into numeric form for analysis or machine learning. This can be done using Label Encoding, which assigns a unique number to each category, or One-Hot Encoding, which creates separate binary columns for each category. One-Hot Encoding is preferred for unordered (nominal) data, while Label Encoding suits ordered (ordinal) data. You can also use mapping for manual conversions or mean encoding for advanced modeling, though it requires caution to avoid data leakage.

Q 3.What is the difference between normalization and standardization?
Ans. Normalization and standardization are techniques used to scale numerical features in a dataset. Normalization scales values to a fixed range, typically 0 to 1, using the formula (x - min) / (max - min). It is useful when features have different units or scales. Standardization, on the other hand, transforms data to have a mean of 0 and a standard deviation of 1 using the formula (x - mean) / std. It is preferred when data follows a normal distribution or for algorithms that assume centered data, like linear regression or SVM.

Q4.How do you detect outliers?
Ans. Outliers can be detected using statistical methods or visualizations. A common approach is the IQR (Interquartile Range) method, where values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ are considered outliers. Another method is using z-scores, where values with a z-score greater than 3 or less than −3 may be outliers. Visual tools like boxplots and scatter plots also help identify unusually high or low values in the data.

Q 5.Why is preprocessing important in ML?
Ans. Preprocessing is essential in machine learning because it prepares raw data for modeling, ensuring better performance and accuracy. It involves handling missing values, encoding categorical variables, scaling features, and removing outliers. These steps help models learn patterns more effectively by reducing noise, improving data consistency, and ensuring all features contribute equally during training.

Q 6.What is one-hot encoding vs label encoding?
Ans. One-hot encoding and label encoding are techniques used to convert categorical data into numerical form. Label encoding assigns a unique integer to each category, which can imply an unintended order. It's best for ordinal data. One-hot encoding creates separate binary columns for each category, avoiding any implied ranking. It's ideal for nominal data with no inherent order, ensuring models don't misinterpret category relationships.

Q7.How do you handle data imbalance?
Ans. Handling data imbalance is important to prevent machine learning models from being biased toward the majority class. Common techniques include resampling methods like oversampling the minority class (e.g., using SMOTE) or undersampling the majority class. Alternatively, you can use class weights in models to give more importance to the minority class or choose algorithms that handle imbalance well, such as Random Forest or XGBoost.

Q8.Can preprocessing affect model accuracy?
Ans. Yes, preprocessing can significantly affect model accuracy. Proper preprocessing—such as handling missing values, encoding categorical variables, scaling features, and removing outliers—ensures that the data is clean, consistent, and suitable for the model. Poor or incomplete preprocessing can lead to misleading patterns, biased predictions, or poor generalization, ultimately reducing the model's performance.