Interview Questions:

Q1.What is the purpose of EDA?
Ans.The purpose of Exploratory Data Analysis (EDA) is to understand the structure, patterns, and quality of a dataset before applying any modeling or decision-making. It helps identify trends, relationships, outliers, and missing values using statistical summaries and visualizations. EDA guides data cleaning, feature selection, and modeling by providing insights that shape the analytical approach.

Q 2.How do boxplots help in understanding a dataset?
Ans. Boxplots help in understanding a dataset by visually summarizing the distribution, central tendency, and variability of a numeric variable. They highlight the median, quartiles, and potential outliers, making it easy to detect skewness and compare distributions across different groups. This allows for quick identification of data spread and anomalies.

Q 3.What is correlation and why is it useful?
Ans. Correlation measures the strength and direction of a linear relationship between two numerical variables. It is useful because it helps identify which variables are related, allowing analysts to detect patterns, reduce redundancy, and select relevant features for modeling. Understanding correlation guides better decision-making and improves the effectiveness of data analysis.

Q4.How do you detect skewness in data?
Ans. Skewness in data is detected by examining the asymmetry of the distribution of values. It can be identified visually using histograms or boxplots, where a longer tail on one side indicates skewness. Statistically, skewness is measured using a skewness coefficient: a value greater than zero indicates right (positive) skew, less than zero indicates left (negative) skew, and around zero suggests a symmetric distribution.

Q 5.What is multico linearity?
Ans. Multicollinearity occurs when two or more independent variables in a dataset are highly correlated, meaning they provide redundant information. This can make it difficult to determine the individual effect of each variable in regression models, leading to unstable coefficients and reduced model interpretability.

Q 6.What tools do you use for EDA?
Ans.Common tools for EDA include Pandas for data manipulation, Matplotlib and Seaborn for visualizations, and NumPy for numerical analysis. These tools help summarize, visualize, and explore data efficiently.

Q7.Can you explain a time when EDA helped you find a problem?
Ans. During an EDA of customer transaction data, I discovered several records with negative purchase amounts through summary statistics and boxplots. This anomaly indicated a data entry or processing error. Investigating further, I found these were unintended duplicates or refunds not properly labeled. Identifying this early through EDA helped clean the data before modeling, improving accuracy and insights.

Q8.What is the role of visualization in ML?
Ans. Visualization plays a crucial role in machine learning by making complex data and model outputs more understandable. It helps identify patterns, trends, and anomalies in the data before modeling, ensuring better data quality and feature selection. During model evaluation, visual tools like confusion matrices, ROC curves, and learning curves allow for a clear understanding of performance.

Visualization also aids in communicating insights to stakeholders, making results more accessible and actionable.