Q1.What are the main steps you take in an Exploratory Data Analysis?

My EDA process follows these key steps: 1) Data Loading and Inspection to understand the dataset's structure, data types, and basic statistics. 2) Data Cleaning , where I identify and handle missing values and correct any data type inconsistencies. 3) Univariate Analysis to understand the distribution of individual variables using plots like histograms and count plots. 4) Bivariate and Multivariate Analysis to explore relationships between variables using correlation heatmaps, scatter plots, and grouped bar charts. 5) Insight Generation , where I synthesize my findings into actionable business insights.

Q2.Which plots help detect skewness and outliers in numerical data?

To detect skewness , histograms and Kernel Density Plots (KDE) are most effective as they visually show the shape of the distribution. For detecting outliers , box plots are the standard tool because they explicitly plot points that fall outside the typical range (the whiskers).

Q3.How would you deal with missing values in a dataset during EDA?
Ans :- My approach depends on the context: 1) Removal: If the number of missing values is very small (e.g., <5%) and randomly distributed, I might remove the rows or columns. 2) Imputation: For numerical data, I would impute with the mean or median. For categorical data, I'd use the mode. 3) Advanced Imputation: In more complex scenarios, I might use model-based imputation like k-Nearest Neighbors (k-NN) to predict the missing values based on other features.

Q4. What does a correlation heatmap show, and how do you interpret it?
Ans :- A correlation heatmap is a graphical representation of the correlation matrix. It uses color to show the strength and direction of the linear relationship between pairs of numerical variables. A value near +1 (often shown in a warm color) means a strong positive correlation, a value near -1 (a cool color) means a strong negative correlation, and a value near 0 (a neutral color) means no linear relationship.

Q5.How would you analyze the relationship between two categorical variables?
Ans :- The best way is to use a contingency table (or crosstab), which shows the frequency of each combination of categories. To visualize this, I would use a grouped or stacked bar chart . For example, a countplot in Seaborn with the hue parameter set to the second categorical variable.

Q6.When would you use a boxplot vs a histogram?
Ans :- I use a histogram to understand the distribution, shape, and skewness of a single numerical variable. I use a boxplot to quickly compare the distributions of a numerical variable across several different categories and to clearly identify outliers.

Q7.Can EDA help detect data leakage? How?

Ans :- Yes, absolutely. Data leakage occurs when your training data contains information about the target that would not be available at the time of prediction. During EDA, you can detect this by looking for features that have an unrealistically high correlation (close to 1 or -1) with the target variable. This is a major red flag that the feature is likely a proxy for the target or was created using information from it.

Q8.How do you summarize key takeaways in a business-readable format?
Ans :- I focus on storytelling with data. I translate statistical findings into clear, actionable business insights. I use simple language, avoid jargon, and lead with the conclusion. For example, instead of saying "The survival rate for Pclass 3 was 24%," I would say, "Passengers in the third class had a significantly lower chance of survival, highlighting a stark difference in outcomes based on socio-economic status."