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

My EDA process involves: 1) Data Loading & Inspection to understand its structure, types, and basic statistics. 2) Data Cleaning, focusing on handling missing values and correcting data types. 3) Univariate Analysis to understand individual variable distributions. 4) Bivariate/Multivariate Analysis to explore relationships and correlations. 5) Insight Generation where I summarize findings to answer business questions.

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

Histograms and Kernel Density Plots (KDE) are excellent for visualizing the shape and skewness of a distribution. Box plots are the best tool for clearly identifying outliers, which appear as individual points beyond the whiskers.

Q3.How would you deal with missing values in a dataset during EDA?

First, I'd identify the extent and nature of missing data. Depending on the situation, I would: 1) Remove rows/columns if the amount of missing data is small and won't cause bias. 2) Impute values using the mean, median (for numerical), or mode (for categorical). 3) For more advanced cases, use model-based imputation (e.g., k-NN).

Q4.What does a correlation heatmap show, and how do you interpret it?

A correlation heatmap is a visual representation of the correlation matrix. It uses color to show the strength and direction of linear relationships between numerical variables. A value close to +1 indicates a strong positive correlation, -1 indicates a strong negative correlation, and 0 indicates no linear correlation.

Q5.How would you analyze the relationship between two categorical variables?

I would use a contingency table (created with pd.crosstab) to see the frequency distribution across the categories. To visualize it, I'd use a stacked or grouped bar chart (countplot with a hue parameter in Seaborn) to compare the proportions.

Q6.When would you use a boxplot vs a histogram?

I use a histogram to understand the overall shape, central tendency, and spread of a single numerical variable. I use a boxplot to quickly compare distributions across multiple categories and to clearly identify outliers.

Q7.Can EDA help detect data leakage? How?

Yes. Data leakage is when information from the future or the target variable is accidentally included in the features. During EDA, you might spot this by finding a feature with an unusually high correlation (near +1 or -1) with the target variable. This is a major red flag that warrants investigation.

Q8.How do you summarize key takeaways in a business-readable format?

I summarize takeaways by focusing on actionable insights. I use clear visualizations, avoid technical jargon, and present findings as a story. For example, instead of saying "Gross income is positively correlated with product line X," I'd say, "Product line X is our most profitable category, suggesting we should focus marketing efforts there."