**Interview Answers**

1. What is the purpose of EDA?

EDA helps researchers understand the data's characteristics and patterns before building statistical models or making assumptions about the data. It involves examining and visualizing the data to gain insights into its main features, identify outliers, and spot potential relationships between variables. This early exploration helps guide further analysis and model selection by revealing hidden patterns and potential issues in the data.

2. How do boxplots help in understanding a dataset?

Boxplots are a visual representation of data distribution, showing the median, quartiles (25th and 75th percentiles), and potential outliers. They help identify the central tendency of the data, its spread or variability, and any unusual data points that fall outside the typical range. By comparing boxplots across different groups, researchers can also visualize differences in data distribution and potential skewness.

3. What is correlation and why is it useful?

Correlation measures the strength and direction of the relationship between two variables. It helps determine if variables tend to move together (positive correlation) or in opposite directions (negative correlation). Understanding correlation is useful for identifying potential relationships between variables, which can be used for feature selection in machine learning models and for building more accurate predictive models.

4. How do you detect skewness in data?

Skewness refers to the asymmetry of a data distribution. It can be detected using various methods, including:

* **Histograms:** Histograms visually show the distribution of data, and skewed distributions will have a tail extending towards one side.
* **Boxplots:** If the median line is not in the center of the box, it indicates skewness.
* **Skewness Metrics:** Statistical measures like the Pearson coefficient or the skewness coefficient can be used to quantify the degree of skewness.

5. What is multicollinearity?

Multicollinearity occurs when independent variables in a statistical model are highly correlated with each other. This can make it difficult to determine the unique effect of each variable on the dependent variable, potentially leading to unstable model parameters and inaccurate interpretations. Detecting and addressing multicollinearity is important for building reliable and interpretable models.

6. What tools do you use for EDA?

Common tools for EDA include:

* **Python:** Libraries like Pandas (for data manipulation), Matplotlib (for plotting), and Seaborn (for statistical visualization) are widely used.
* **R:** A powerful programming language and statistical environment with extensive libraries for data analysis and visualization.
* **Statistical Software:** Programs like SPSS, SAS, and JMP also offer various tools for EDA.

7. Can you explain a time when EDA helped you find a problem?

In one project, EDA revealed a high correlation between two variables (e.g., advertising spending and sales) that were initially considered independent. This led to a re-evaluation of the assumptions about the variables and the model, and ultimately, the discovery that advertising spending had a more significant impact on sales than initially suspected. By uncovering this hidden relationship, the team was able to build a more accurate predictive model by including advertising spending as a key feature.

8.What is the role of visualization in ML?

Visualization in Machine Learning (ML) helps data scientists understand complex data, identify patterns, and improve model development. It translates raw data into visual formats like charts and graphs, making it easier to see trends, relationships, and outliers. This visual insight is crucial for informed decision-making and effective communication of results throughout the ML process.