EDA Interview Questions - Answers by Bhagavi

1. What is the purpose of EDA?

The purpose of Exploratory Data Analysis (EDA) is to understand the structure, patterns, and relationships in a dataset before applying any machine learning models. It helps in detecting outliers, missing values, and discovering trends which guide better data cleaning and feature engineering.

2. How do boxplots help in understanding a dataset?

Boxplots visually summarize the distribution of a dataset. They highlight the median, quartiles, and potential outliers. With boxplots, I can quickly understand the spread of data and detect if any variable has extreme values that may affect analysis.

3. What is correlation and why is it useful?

Correlation measures the linear relationship between two variables. It's useful because it helps identify how changes in one feature may affect another. This is especially important for feature selection in ML, as highly correlated variables might be redundant.

4. How do you detect skewness in data?

Skewness can be detected using the skewness statistic (`df.skew()`) or by visualizing data with histograms. If the distribution tail is longer on one side, it's skewed. Skewed data can affect model performance and may need transformation.

5. What is multicollinearity?

Multicollinearity occurs when two or more independent variables are highly correlated. This can confuse regression models, making it difficult to determine the effect of each variable. It's often detected using a correlation matrix or VIF (Variance Inflation Factor).

6. What tools do you use for EDA?

I use Python libraries such as Pandas for data handling, Seaborn and Matplotlib for visualizations, and sometimes Plotly for interactive plots. Jupyter Notebook is my go-to environment for quick analysis.

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

Yes! During my Titanic EDA task, I noticed many missing values in the 'Age' column. I also detected outliers in 'Fare' through a boxplot. These insights helped me decide how to fill missing values and whether to cap outliers before modeling.

8. What is the role of visualization in ML?

Visualization is essential in ML because it makes patterns, distributions, and relationships between features easier to understand. As Bhagavi, I believe visuals tell stories that numbers can't. They help in feature selection, understanding data imbalance, and communicating insights effectively to others.