**1. What is the purpose of EDA?**

**Exploratory Data Analysis (EDA)** helps you:

* Understand the structure, patterns, and distribution of data.
* Detect anomalies, missing values, or outliers.
* Uncover relationships between variables.
* Prepare data for modeling by guiding preprocessing decisions.

**2. How do boxplots help in understanding a dataset?**

A **boxplot**:

* Shows the distribution of numerical data via five-number summary: min, Q1, median, Q3, and max.
* Highlights **outliers** clearly.
* Helps compare distributions across categories.
* Indicates **skewness** through the position of the median and length of whiskers.

**3. What is correlation and why is it useful?**

**Correlation** measures the strength and direction of a linear relationship between two variables (range: -1 to +1).

* Useful for:
  + Identifying **highly related features** (for feature selection).
  + Avoiding **redundant data** in models.
  + Understanding relationships (e.g., positive/negative influence on target).

**4. How do you detect skewness in data?**

* **Numerically** using skewness value:
  + 0 = symmetric, >0 = right-skewed, <0 = left-skewed.
* **Visually** using:
  + Histogram
  + Boxplot (asymmetric box/whiskers)
  + Density plot

**5. What is multicollinearity?**

**Multicollinearity** occurs when two or more independent variables in a dataset are highly correlated.

* Causes:
  + Inflated coefficients
  + Unstable model interpretation
* Detection:
  + **Correlation matrix**
  + **Variance Inflation Factor (VIF)** > 5 or 10 indicates concern

**6. What tools do you use for EDA?**

* **Python**: pandas, matplotlib, seaborn, plotly, pandas-profiling, sweetviz
* **R**: ggplot2, dplyr, DataExplorer
* **Jupyter Notebooks** for interactive exploration
* **Excel** or **Power BI/Tableau** for quick visual profiling (non-programmatic)

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

Yes. In one project, while predicting customer churn, EDA revealed:

* One feature had **80% missing values**, making it unusable.
* Another numeric feature had **extreme outliers** (age > 200) due to data entry errors.
* Correlation analysis showed two features were nearly identical, so one was dropped to avoid **multicollinearity**.  
  Fixing these improved model accuracy and interpretability.

**8. What is the role of visualization in ML?**

Visualization helps:

* Understand data distribution and structure
* Identify patterns, trends, and anomalies
* Detect outliers, missing data, or imbalance
* Interpret model results (e.g., feature importance, confusion matrix)
* Communicate findings to stakeholders clearly