**Interview Questions**

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

**Exploratory Data Analysis (EDA)** helps us understand the structure, patterns, and relationships in a dataset before applying any machine learning model.  
It allows us to:

* Detect missing values and outliers.
* Understand variable distributions.
* Find correlations and relationships between features.
* Make data-driven decisions for preprocessing and feature engineering.

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

A **boxplot** shows the **spread and distribution** of numerical data using five main points — *minimum, Q1, median, Q3, and maximum*.  
It helps to:

* Identify outliers easily.
* See the skewness of data (if the median is not centered).
* Compare distributions across different categories.

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

**Correlation** measures how strongly two variables are related (ranges from -1 to +1).

* +1 → Perfect positive relationship
* -1 → Perfect negative relationship
* 0 → No relationship  
  It is useful because it helps identify features that move together, which aids in feature selection and avoids redundancy.

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

**Skewness** shows whether the data is symmetrical or biased to one side.  
You can detect it by:

* Histogram or density plots (visual check).
* Skewness value using pandas: df['column'].skew().
  + Positive skew → long right tail
  + Negative skew → long left tail

**5.What is multicollinearity?**

**Multicollinearity** occurs when two or more independent variables are highly correlated with each other.  
This makes it difficult for models (like linear regression) to estimate feature importance correctly.  
It can be detected using:

* Correlation matrix (high correlation between features).
* Variance Inflation Factor (VIF) (>10 indicates high multicollinearity).

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

Common tools and libraries:

* Python: pandas, numpy, matplotlib, seaborn, plotly
* R: ggplot2, dplyr
* BI Tools: Power BI, Tableau (for visual EDA)

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

For example:  
While analysing the Titanic dataset, EDA showed that the Age column had many missing values and Fare had extreme outliers.  
By identifying and fixing these issues, the model’s accuracy improved significantly.

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

**Visualization** helps to understand data patterns, trends, and relationships at a glance.  
It allows:

* Better feature understanding and selection.
* Detection of outliers, missing values, and skewness.
* Clear communication of insights to stakeholders.