Task 2 Interview Questions - Exploratory Data Analysis (EDA)

1. What is the purpose of EDA?

EDA (Exploratory Data Analysis) helps understand the dataset before applying ML. It reveals patterns, missing values, outliers, and relationships.

Example: In Iris dataset, petal length helps separate species.

2. How do boxplots help in understanding a dataset?

Boxplots show the distribution, median, IQR, and outliers.

Example: Boxplot of sepal_width shows unusual values easily.

3. What is correlation and why is it useful?

Correlation measures the strength of the relationship between features.

Useful for feature selection.

Example: petal_length and petal_width are highly correlated in Iris.

4. How do you detect skewness in data?

Use .skew() method or check histogram shape.

Example: Right-skewed = more small values, tail to the right.

5. What is multicollinearity?

When two or more features are highly correlated, it causes redundancy.

Can mislead models like linear regression.

Example: petal_length and petal_width might cause multicollinearity.

6. What tools do you use for EDA?

- Pandas for stats
- Matplotlib/Seaborn for visualizations
- Plotly for interactive graphs

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

In the Titanic dataset, I saw 'Cabin' had many missing values using .isnull().sum().

I dropped it early, saving time during modeling.

8. What is the role of visualization in ML?

It helps understand data, detect trends/outliers, and communicate insights.

Example: Pairplot showed Iris species are separable based on petal size.