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

EDA helps in:

* **Understand data structure**: Identify feature types (numeric/categorical), distributions, and missing values (e.g., Age had 20% missing values).
* **Detect patterns/trends**: Found that **1st-class passengers (Pclass=1)** had a 63% survival rate vs. 24% for 3rd-class.
* **Spot anomalies**: Outliers in Fare (e.g., a $512 ticket).
* **Guide modeling**: Revealed that Sex and Pclass should be key features for predicting survival.

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

* **Visualize spread**: Show median (Q2), quartiles (Q1, Q3), and whiskers (IQR).
  + *Example*: Boxplots revealed most fares were under $50, but extreme outliers ($200+) existed.
* **Identify outliers**: Points beyond 1.5×IQR (e.g., a few passengers paid exceptionally high fares).
* **Compare groups**: Survival rates varied by Pclass; boxplots showed 1st-class passengers were typically older.

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

* Measures linear relationships (-1 to 1). In the Titanic dataset:
  + Fare and Survived had a **weak positive correlation (0.26)**: Higher fares linked to better survival odds.
  + Pclass and Fare were **negatively correlated (-0.55)**: Lower classes paid less.
* **Usefulness**:
  + Guides feature selection (e.g., prioritize correlated features like Pclass).
  + Detects multicollinearity (e.g., SibSp and Parch had mild correlation).

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

* **Visually**: Histograms with long tails (e.g., Fare was right-skewed: most values clustered under $50, but a few exceeded $500).
* **Statistically**:
  + Skewness coefficient:
    - Fare = 4.3 (highly skewed).
    - Age = 0.4 (near-symmetric).
  + Q-Q plots: Deviations from the diagonal line indicate skewness.

**5. What is multicollinearity?**

* When **independent variables are highly correlated**, causing unstable model coefficients.
  + *Example*: If SibSp (siblings/spouses) and Parch (parents/children) were combined into FamilySize, including both could introduce redundancy.
* **Detection**:
  + Correlation matrix (e.g., SibSp and Parch had 0.41 correlation).
  + Variance Inflation Factor (VIF) > 5 suggests multicollinearity.

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

* **Python Libraries**:
  + Pandas (summary stats, df.describe()).
  + Seaborn/Matplotlib (boxplots, histograms, sns.pairplot()).
  + Plotly (interactive plots, e.g., px.scatter\_3d()).
* **Automated EDA**: pandas-profiling for quick overviews.
* **Statistical Tests**: Scipy (stats.skew, stats.zscore).

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

* **Issue**: Missing Embarked values (2 rows) could skew port-based analysis.
* **EDA Action**:
  + Used df[df['Embarked'].isnull()] to locate missing rows.
  + Found both were high-fare, 1st-class passengers; imputed with ‘C’ (Cherbourg’s mode for 1st-class).
* **Impact**: Avoided bias in survival rate calculations by port.

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

* **Pre-Modeling**:
  + **Feature selection**: Histograms showed Age bins (e.g., children) had higher survival.
  + **Outlier handling**: Boxplots flagged extreme fares needing winsorization.
* **Post-Modeling**:
  + Explain predictions (SHAP plots).
  + Debug errors (residual plots for regression).