# INTERVIEW PREPARATION QUESTIONS

**1. What is EDA and why is it important?**

EDA (Exploratory Data Analysis) is the process of exploring and analyzing datasets to summarize their main characteristics, often using visualizations and statistical summaries.

Importance:

- Understand data distributions, patterns, and anomalies
- Identify missing values and data quality issues
- Select appropriate statistical tests and models
- Generate hypotheses and guide feature engineering

**2. Which plots do you use to check correlation?**

- Heatmap (correlation matrix visualized as colors) → Good for overall correlation view
- Scatter plot → Good for visualizing correlation between two continuous variables
- Pairplot → Multiple scatter plots with histograms/ KDE for quick multi-variable correlation view
- Correlogram (advanced) → Displays correlation coefficients and their signs visually

**3. How do you handle skewed data?**

If data distribution is highly skewed (not symmetric):

- Log Transformation (np.log1p(x)) → for right-skewed data with all positive values
- Square Root / Cube Root Transformation → reduces skew but keeps zeros
- Box-Cox Transformation (requires positive values)
- Yeo-Johnson Transformation (can handle negatives)
- In classification, sometimes binning or normalization can help instead

**4. How to detect multicollinearity?**

- Correlation Matrix: Check if correlation coefficient > 0.8 or < -0.8
- Variance Inflation Factor (VIF):
    - VIF > 5 (or 10 in some cases) indicates multicollinearity
    - Formula: VIF = 1 / (1 - $R^2$) where $R^2$ is from regressing one feature on all others

**5. What are univariate, bivariate, and multivariate analyses?**

- Univariate → Analysis of one variable at a time (e.g., histogram, boxplot)
- Bivariate → Analysis of two variables to study relationships (e.g., scatter plot, bar plot)
- Multivariate → Analysis of more than two variables together (e.g., heatmap, PCA, multiple regression)

**6. Difference between heatmap and pairplot?**

- Heatmap → Shows correlation matrix or data values as a color-coded grid (good for numeric summaries)
- Pairplot → Creates scatter plots for each pair of variables plus histograms/ KDEs (good for distribution + relationship view)

**7. How do you summarize your insights?**

- Start with data quality summary → missing values, duplicates, outliers
- Mention key trends & distributions → e.g., “Survival rate higher in females and 1st class passengers”
- Highlight correlations & patterns → e.g., “Fare is positively correlated with survival”
- Note important anomalies or unexpected findings
- Keep it concise and actionable