EDA Visual Observations & Summary

1. Heatmap of Descriptive Statistics

The heatmap of df.describe() highlighted that 'Fare' has a high range and standard deviation, suggesting extreme variability and possible outliers. 'Age' also showed moderate variability.

2. Pairplot between Age and Fare

The pairplot revealed no strong linear correlation between 'Age' and 'Fare'. The distribution of 'Fare' is right-skewed, while 'Age' appears more normally distributed.

3. Bar Chart of Mean Fare by Age

Some age groups had significantly higher average fares, indicating a possible link between age and class of travel. Variability was large, hinting at differing passenger categories.

4. Boxplot of All Numerical Columns

Boxplots showed strong outliers in 'Fare' and moderate ones in 'Age'. This confirms the variability observed earlier and indicates that preprocessing (like outlier handling) is needed.

5. Mean Age by Embarkation Port

Mean age varies slightly across embarkation ports. This could suggest socio-economic differences among passengers from different ports.

Summary of Findings

- Missing values exist in the dataset (details visible via .isnull().sum()).
- 'Fare' is highly skewed and contains significant outliers.
- 'Age' is mostly normal but has a few outliers.
- No strong correlation between 'Age' and 'Fare'.
- Port of embarkation shows minor age differences.
- Categorical variables like 'Sex' and 'Embarked' can be encoded.
- Feature scaling and transformation may be necessary for modeling.