Exploratory Data Analysis

Introduction

The Titanic dataset provides information about the passengers aboard the ill-fated Titanic voyage. This exploratory data analysis (EDA) aims to uncover insights, patterns, and relationships among the features to better understand factors influencing survival. Techniques like statistical summaries and visualizations such as histograms, boxplots, scatterplots, and heatmaps are employed.

Observations for Each Visuals

1. df.info() and df.describe():

- These commands provide an overview of the dataset, including column types, missing values, and basic descriptive statistics such as mean, median, standard deviation, min, and max values for numerical columns.
- Look for missing data in Age, Embarked, or any other column, and decide how to handle it (e.g., imputation or removal).

2. Value Counts for Categorical Variables:

- These give you a quick understanding of the distribution of categorical variables (e.g., whether more passengers survived, their class distribution, and gender distribution).
- For example, how many passengers survived (Survived = 1) vs. did not survive (Survived = 0), and whether gender or class (Pclass) played a role in survival.

3. Pairplot:

- The pairplot() helps to visualize pairwise relationships in the dataset for the selected numeric columns.
- You can quickly identify trends or separations between survived vs. non-survived passengers. For example, if passengers who survived tend to have higher Fare and lower Age compared to those who didn't survive.

4. Correlation Heatmap:

- The heatmap will show correlations between numeric variables like Age, Fare, and the number of family members aboard (SibSp, Parch).
- Look for strong correlations like between Age and Fare or low correlations between Pclass and SibSp, which might help in predicting survival.

5. Histograms:

- Histograms of variables like Age, Fare, and SibSp help you understand their distribution.
- For instance, Age might show a normal distribution with a peak in the young adult range, while Fare might show a skewed distribution (more low-fare passengers).

6. Boxplot of Fare by Pclass:

- Boxplots help to compare distributions. Here, you’ll see how fare varies across different passenger classes.
- First-class passengers are expected to have higher fares, and the boxplot will show that distinct separation.

7. Boxplot of Age by Survival:

- This boxplot reveals if there is a significant difference in the age distributions of survivors and non-survivors.
- Typically, younger passengers tend to have higher survival rates, so this visualization can reveal if age is a significant factor in survival.

8. Scatterplot of Age vs Fare:

- This scatterplot helps you visually understand the relationship between Age and Fare and how it interacts with survival.
- You can identify if older passengers (perhaps with higher fares) tend to have a higher survival rate.

Final Summary of Findings

- **Sex** was a major determinant of survival, with females having significantly higher chances.
- **Pclass** was crucial: first-class passengers survived at a much higher rate.
- **Age** impacted survival: younger passengers (especially children) had a better survival rate.
- **Fare** also influenced survival: passengers paying higher fares were more likely to survive.
- Missing data in 'Age' and 'Cabin' requires handling for deeper modeling.
- Overall, social status (class and gender) played a crucial role in survival odds on the Titanic.

Conclusion

The exploratory analysis highlighted clear trends in the Titanic dataset: being female, traveling first-class, and paying a higher fare were associated with better survival rates. These insights will be vital when building predictive models. Addressing missing values and engineering additional features (like family size) could further enhance data understanding and prediction accuracy.