Exploratory Data Analysis Report – Titanic Dataset

The Titanic dataset is mainly for the purpose of understanding for survival patterns among passengers based on demographic and socio-economic factors. The analysis began with an overview using .info() and .describe(), this showed that the dataset contains several key features such as Survived, Pclass, Sex, Age, Fare, SibSp, Parch, and Embarked. There were missing values that were observed in Age, Cabin, and Embarked, that may have required imputation or exclusion moving forward.

Only count charts were plotted for Survived, Pclass, Sex, and Embarked in order to comprehend the distribution of categorical variables. Only roughly 38% of passengers survived, according to the survival count plot, suggesting a class imbalance. Third class had the lowest survival rate and the majority of passengers, according to the Pclass chart. With more males on board but noticeably higher survival rates for females, the Sex chart showed a gender gap. Although survival rates varied slightly between ports, Southampton was the most popular port of embarkation, according to the Embarked chart.

For quantitative features, Age and Fare histograms were used to highlight distributions. The histogram for Age revealed that a right-skewed distribution has the majority of the passengers between the ages of 20 and 30 years . The Fare histogram indicated a long-tailed distribution with many high-value outliers, this implies that fare could be used as a proxy for socio-economic status.

Bivariate analysis with boxplots and count plots further illuminated survival patterns. The Age vs Survived boxplot revealed that the chances of survival were better for younger passengers, particularly children. Likewise, the Fare vs Survived boxplot revealed that survivors paid higher fares. Count plots of Pclass and Sex against Survived as hue ensured that first-class travelers and females had much higher rates of survival.

Multivariate relationships between features like Pclass, Age, Fare, SibSp, and Parch were examined using seaborn pairplots, which were colored by survival status. This visualization showed a distinct difference in survival outcomes by fare and class, with survivors clustering in higher fare and lower class values. The correlation heatmap, which showed a negative correlation between Pclass and Fare and a positive correlation between Survived and Fare, supported these findings. Interestingly, there was a positive correlation between SibSp and Parch, suggesting that family size could be a valuable trait to investigate further.

In conclusion, the EDA showed that fare, passenger class, and gender all had a significant impact on survival on the Titanic. Additionally, age was a factor, especially for younger travelers. Future preprocessing steps should address the missing values in Age, Cabin, and Embarked. With the potential to generate new variables like family size or socioeconomic status, these insights provide a strong basis for feature engineering and predictive modeling.