**Exploratory Data Analysis (EDA) Report on Titanic Dataset**

**Name**: Kandarp Singh Mahobiya

**Enrollment number**: 23114045

**Introduction:**

The Titanic dataset is a well-known dataset that contains information about the passengers aboard the RMS Titanic, which sank in 1912. The dataset includes various features such as passenger class, age, gender, survival status, and more. This report aims to provide a detailed analysis of the dataset through exploratory data analysis (EDA), univariate analysis, bivariate analysis, and multivariate analysis.

**Data Cleaning:**

Before diving into the analysis, it is crucial to clean the data to ensure accurate results. The following steps were taken to clean the dataset:

1. Removing Unnecessary Columns: Columns such as 'who', 'deck', and 'alive' were dropped as they were deemed unnecessary for the analysis.

2. Handling Missing Values: Missing values were handled using forward fill (ffill) to ensure continuity in the data.

3. Removing Duplicates: Duplicate rows were removed to avoid redundancy.

4. Standardizing Categorical Values: Categorical values in the 'sex' and 'embarked' columns were standardized to lowercase and uppercase, respectively.

5. Correcting Outliers: Outliers in numeric columns were corrected using the Interquartile Range (IQR) method.

**Univariate Analysis**

Univariate analysis focuses on analyzing each variable individually to understand its distribution and characteristics.

**Summary Statistics**

The summary statistics for numeric columns provide insights into the central tendency, dispersion, and shape of the data distribution.

* Survived: The mean survival rate is approximately 40%, indicating that less than half of the passengers survived.
* Pclass: The mean passenger class is around 2.28, suggesting that most passengers were in the second or third class.
* Age: The mean age of passengers is approximately 29.72 years, with a standard deviation of 14.71 years, indicating a wide range of ages.
* SibSp: The mean number of siblings/spouses aboard is 0.45, with a maximum of 2.5.
* Parch: The mean number of parents/children aboard is 0.38, with a maximum of 2.5.
* Fare: The mean fare is approximately 25.34, with a standard deviation of 21.64, indicating a wide range of fares.

**Frequency Distributions for Categorical Data**

The frequency distributions for categorical columns provide insights into the distribution of categories within each variable.

* Sex: There are 577 males and 314 females in the dataset.
* Embarked: The majority of passengers embarked from Southampton (644), followed by Cherbourg (168) and Queenstown (77).
* Class: The majority of passengers were in the third class (491), followed by the first class (216) and second class (184).
* Adult Male: There are 537 adult males and 354 non-adult males.
* Embark Town: The majority of passengers embarked from Southampton (644), followed by Cherbourg (168) and Queenstown (77).
* Alone: There are 537 passengers who were alone and 354 who were not alone.

**Bivariate Analysis**

Bivariate analysis explores the relationship between two variables to understand if there is any correlation or dependency between them.

**Correlation Analysis**

Correlation analysis helps in understanding the relationship between numeric variables. A correlation coefficient close to 1 or -1 indicates a strong relationship, while a coefficient close to 0 indicates a weak or no relationship.

* Survived vs. Pclass: There is a negative correlation, indicating that passengers in lower classes (higher class numbers) had a lower survival rate.
* Survived vs. Age: There is a weak negative correlation, suggesting that younger passengers had a slightly higher survival rate.
* Survived vs. Fare: There is a positive correlation, indicating that passengers who paid higher fares had a higher survival rate.
* Pclass vs. Fare: There is a negative correlation, indicating that passengers in lower classes paid lower fares.

**Survival Rate by Gender**

* + Female Survival Rate: The survival rate for females is significantly higher than for males, with 74.2% of females surviving compared to 18.9% of males.

**Survival Rate by Passenger Class**

* + First Class: The survival rate for first-class passengers is the highest, at 62.96%.
  + Second Class: The survival rate for second-class passengers is 47.28%.
  + Third Class: The survival rate for third-class passengers is the lowest, at 24.24%.

**Multivariate Analysis**

Multivariate analysis involves examining the relationship between three or more variables to understand complex interactions.

**Survival Rate by Gender and Passenger Class**

* + First Class Females: The survival rate for first-class females is the highest, at 96.81%.
  + Third Class Males: The survival rate for third-class males is the lowest, at 12.53%.

**Age Distribution by Survival Status**

* + Survived: The age distribution for survivors is slightly skewed towards younger passengers, with a mean age of 28.34 years.
  + Not Survived: The age distribution for non-survivors is slightly skewed towards older passengers, with a mean age of 30.63 years.

**Conclusion**

The analysis of the Titanic dataset provides several insights into the factors that influenced survival rates:

1. Passenger Class: Passengers in higher classes (first and second) had a significantly higher survival rate compared to those in the third class.

2. Gender: Females had a much higher survival rate than males, indicating that gender played a crucial role in survival.

3. Age: Younger passengers had a slightly higher survival rate, although the difference is not as pronounced as with class and gender.

4. Fare: Passengers who paid higher fares had a higher survival rate, likely due to their higher class status.

The correlation analysis revealed that certain variables, such as passenger class and fare, are strongly correlated with survival, while others, like age, have a weaker correlation. This suggests that while class and gender were significant factors in determining survival, age played a less critical role.

Overall, the analysis highlights the importance of socioeconomic status and gender in determining the likelihood of survival on the Titanic. These insights can be valuable for understanding historical data and making informed decisions in similar contexts.