Exploratory Data Analysis – Titanic Dataset

1)Data Exploration:

The Titanic dataset is loaded from a CSV file to start the analysis. Viewing the first and last rows, enumerating distinct column names, producing descriptive statistics for numerical columns, displaying data types and memory usage, and verifying the dataset's dimensions are all part of the initial exploration process.

2)Managing Missing Values:

The frequency of missing values in each column is determined. To lessen the effect of outliers, the median is used to fill in the missing values for the "Fare" column. Because the mode is less sensitive to

outliers in this situation, it is used for imputation for the 'Age' column. Due to its limited relevance and large number of missing values, the 'Cabin' column has been removed.

3)Data Visualization:

The following significant observations and conclusions were drawn from the data visualizations and analysis:

- **Survival Rate:** A sizable fraction of travelers perished in the catastrophe.

- **Gender and Survival:** Compared to male passengers, female passengers had a significantly higher survival rate. This implies that in lifeboats, women and children were prioritized.

- **Passenger Class and Survival:** Compared to passengers in third class, those in first and second classes had a higher chance of surviving. Because wealthier passengers had better access to safety resources, this demonstrates the social injustices of the era.

- **Age and Survival:** Younger adults made up a larger portion of the age distribution. Perhaps because they were given priority in lifeboats, infants and kids had a comparatively high survival rate. The survival rate was lower for older passengers.

- **Fare and Survival:** The study found a positive correlation between fare and survival, indicating that passengers who paid higher fares had a higher chance of surviving. Given that higher fares were linked to higher passenger classes,

this supports the observation regarding passenger class.

- **Family Size and Survival:** The survival rate was higher for passengers who were traveling with family, particularly spouses and kids. Families remaining together and helping one another during the evacuation may be the cause of this.

- **Fare and Pclass Correlation:** Higher passenger classes were linked to higher fares, according to a strong negative correlation. This demonstrates how passenger class and socioeconomic status are related.

- **Age and Pclass Correlation:** Younger passengers were more likely to be in higher classes, according to a moderately negative correlation.

This might have to do with travel habits and family wealth.

- **SibSp and Parch Correlation:** Passengers who traveled with their spouses or siblings were also more likely to travel with their parents or kids, according to this positive correlation. The typical family structures of the era are reflected in this.

4)Data Preprocessing:

To facilitate analysis, the 'Sex' column is transformed into numerical values.

5)Correlation Analysis:

Significant relationships between variables like Fare, Pclass, Age, SibSp, and Parch are revealed by formal

correlation analysis, which validates the observations and conclusions drawn from visualizations.

6)**Conclusion:**

This EDA offers insightful information about the elements that affected Titanic survival. Age, fare, gender, and passenger class were all found to be significant predictors of survival. The likelihood of surviving the disaster was higher for female passengers, passengers in higher classes, younger people, and passengers who paid higher fares. Building predictive models and comprehending the historical background of the Titanic disaster can benefit both from these insights.