# Titanic Dataset Analysis Report

## 1. Data Cleaning Insights

Missing Values:

- Missing values in Age were replaced with the median value, which ensures that extreme values do not distort the data.
- Embarked was filled with the mode value, as it represents the most common point of embarkation.

Duplicate Records:

- Duplicate records were removed, ensuring that the analysis is based on unique observations.

Outliers:

- Outliers in Fare were treated using the IQR method to remove extreme values while retaining the overall distribution.

Standardization:

- Categorical values were standardized for consistency (e.g., Sex converted to lowercase and Embarked values mapped to meaningful names).

## 2. Univariate Analysis Insights

Age Distribution:

- The distribution of Age shows a slight right skew, indicating that more passengers were younger, with a few older individuals.

Passenger Class Distribution:

- The majority of passengers belonged to 3rd class, followed by 1st and 2nd class.

Fare Distribution:

- Fare showed a right-skewed distribution with a few high-value outliers, which were addressed through outlier treatment.

## 3. Bivariate Analysis Insights

Correlation Matrix:

- Weak correlation between Age and Fare (correlation ~ 0.1) - indicating that age and fare are largely independent.
- Strong negative correlation between Pclass and Fare (correlation ~ -0.55) - suggesting that higher-class tickets cost more.

Age vs. Fare (Scatter Plot):

- No significant pattern between Age and Fare, confirming the low correlation.

Survival Rate by Class and Sex (Bar Plot):

- Higher survival rate for 1st-class passengers compared to 2nd and 3rd classes.
- Women had a higher survival rate than men in all passenger classes - consistent with the "women and children first" policy.

Fare Distribution by Class and Sex (Violin Plot):

- Higher fare for 1st-class passengers with noticeable variation between male and female passengers.

## 4. Multivariate Analysis Insights

Pair Plot:

- Survival shows a pattern influenced by both Pclass and Fare.

- Passengers in higher classes with higher fares had a greater likelihood of survival.

Heatmap:

- Low correlation between most numerical features indicates that the Titanic dataset has mostly independent variables.

Grouped Bar Plot:

- Clear pattern: 1st-class passengers, especially women, had the highest survival rates.

## 5. Overall Inferences

- Higher-class passengers and women had a higher survival rate.
- The low correlation between Age and Fare indicates that age was not a major determinant of ticket price.
- Strong correlation between Pclass and Fare shows that higher-class tickets were more expensive.
- The survival rate differences between sexes and classes suggest that social factors played a major role in survival outcomes.
- Outlier treatment and standardization helped in producing clean and consistent data for better insights.

Conclusion:

The analysis confirms that survival on the Titanic was not random - it was influenced by factors like passenger class, gender, and fare. Higher-class passengers and women had a significant survival advantage, highlighting the societal norms and emergency protocols followed during the disaster.