Titanic Dataset Analysis Report

1. Introduction

The Titanic dataset provides details about passengers aboard the RMS Titanic, including their demographics, ticket class, and survival status. This report explores the dataset by performing data cleaning, exploratory data analysis (EDA), and visualizing key relationships among variables.

2. Data Cleaning

2.1 Handling Missing Values

- The **Age** column had missing values, which were filled using the median age of passengers.
- The **Embarked** column had missing values, which were filled with the most frequently occurring port.
- The **Deck** column had too many missing values and was dropped from the dataset.

2.2 Removing Duplicates

- Duplicate entries were checked and removed to ensure data integrity.

2.3 Detecting and Treating Outliers

- The **Fare** column contained extreme values. The Interquartile Range (IQR) method was used to remove outliers that exceeded 1.5 times the IQR.

2.4 Standardizing Categorical Variables

- The **Sex** column values were standardized to lowercase for consistency.
- The **Embark Town** values were capitalized properly.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- **Summary Statistics**: The dataset was summarized using measures such as mean, median, and standard deviation.
- **Age Distribution**: A histogram was plotted to visualize the age distribution of passengers.

3.2 Bivariate Analysis

- **Fare vs. Pclass**: A box plot showed that passengers in higher classes paid significantly more for their tickets.
- **Age vs. Sex**: A box plot was created to compare age distributions across genders.

3.3 Multivariate Analysis

- **Correlation Matrix**: A heatmap visualized correlations among numerical variables, highlighting relationships such as the negative correlation between **Pclass** and **Fare**.

4. Key Findings

- Most missing values were in the **Deck** column, making it unreliable for analysis.
- Older passengers and those in higher classes were more likely to have paid higher fares.
- Survival rates varied significantly based on **Pclass** and **Fare**.
- Correlations among variables helped identify potential patterns in passenger data.

5. Conclusion

This analysis provided valuable insights into passenger demographics and survival trends on the Titanic. Data cleaning ensured data quality, and EDA helped uncover meaningful relationships between variables. Future work can involve predictive modeling to estimate survival probabilities.

6. Recommendations

- Further analysis using machine learning models can improve understanding of survival factors.
- Additional feature engineering (e.g., family size) may enhance predictive power.
- Deeper investigation into ticket fares and class distributions can provide richer insights into passenger experiences.

End of Report