Titanic Dataset Analysis Report

1. Introduction

The Titanic disaster of 1912 is one of the most studied maritime tragedies. This dataset, derived from passenger records, helps analyze various factors influencing survival rates. By applying data science techniques such as data cleaning, exploratory data analysis (EDA), and statistical inference, we can uncover patterns that determined who had a better chance of survival.

This report follows the given problem statement, covering:

- Data Cleaning: Handling missing values, duplicates, and inconsistencies.
- Exploratory Data Analysis (EDA): Investigating numerical and categorical variables.
- Statistical Inference: Identifying factors that affected survival rates.
- Visualizations and Data Interpretation.

The findings will help answer key questions such as:

1. Did ticket class impact survival chances?
2. Were women and children prioritized during rescue?
3. How did fare price affect survival probability?
4. Were there patterns in age distribution among survivors?

2. Data Cleaning

2.1 Inspecting the Dataset

The dataset was loaded and examined using:

- .info() – to check data types and missing values.
- .describe() – for summary statistics.
- .head() – to preview the first few rows.

Key Observations:

- Missing Values: Age (~20%), Cabin (~77%), Embarked (~2%).
- Categorical Columns: Pclass, Sex, Embarked.
- Numerical Columns: Age, Fare, SibSp, Parch.
- Duplicates: Some duplicate records were found and removed.

2.2 Handling Missing Data

Missing values were treated as follows:

- Age: Imputed using the median age (~29.7 years).
- Embarked: Imputed using the mode (most common value: 'S').
- Cabin: Dropped due to excessive missing values (~77%).

2.3 Outlier Detection and Handling

Outliers were identified in Age and Fare columns using box plots. Handling techniques used:

- Z-score method** for Age (since it has small skewness).
- Interquartile Range (IQR) method** for Fare (as it is highly skewed).

2.4 Standardizing Categorical Variables

The Sex column had inconsistent capitalization ('Male' vs. 'male').

The Embarked column contained missing values, replaced with the mode ('S').

The columns Name, Ticket, and PassengerID were dropped as they had no impact on survival.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

Statistical summaries and visualizations were used to analyze individual variables:

- Survival Rate: Only 38.4% of passengers survived.
- Age Distribution: Right-skewed, with a median of 28 years.
- Fare Distribution: Highly skewed, with a few high outliers.
- Gender Distribution: Males (~65%) outnumbered females (~35%).
- Embarked Port: Most passengers boarded from Southampton.

3.2 Bivariate Analysis

Correlation and visual analysis were used to explore relationships:

- Survival vs. Fare: Positive correlation – passengers who paid more had better survival chances.
- Survival vs. Pclass: Negative correlation – first-class passengers had a better survival rate.

- Survival vs. Sex: Females had a significantly higher survival rate than males.

Visualizations used:

- Bar Chart: Showed females had a higher survival rate (~74%) compared to males (~18%).
- Boxplot: Revealed older passengers had lower survival chances.
- Scatter Plot: Showed younger passengers paid lower fares.

3.3 Multivariate Analysis

- Multiple variables were analyzed together:
- Heatmap: Showed strong correlations between Pclass, Fare, and Survival.
- Pair Plots: Explored relationships between Fare, Age, Pclass, and Survival.
- Survival Rates by Class and Gender:
  - First-Class Females: 97% survival rate (highest).
  - Third-Class Males: Only 13% survival rate (lowest).
  - Children (Age <10): Had better survival rates compared to adults.

4. Key Findings and Inferences

The Titanic dataset revealed clear patterns in survival rates:

- Women and children had the highest survival rates.
  - 74% of females survived, while only 18% of males did.
- First-class passengers had significantly higher survival rates.
  - First-class: 63%
  - Third-class: 24%.
- Higher fares were associated with better survival.
- Older passengers had lower survival rates.
- Third-class passengers had the highest mortality.

5. Conclusion

This analysis confirmed that survival on the Titanic was influenced by gender, class, and ticket fare. Women, children, and first-class passengers had the best survival odds.