Name- Harshit Kumar **EnrollmentNO. -23114036**

Analysis Report

Introduction

This report presents an in-depth exploratory data analysis (EDA) of the Titanic dataset, which contains information about the passengers aboard the Titanic, including their survival status, age, gender, class, and other attributes. The goal of this analysis is to understand the relationships between different variables and uncover patterns that may explain the survival outcomes of the passengers.

Dataset Overview

The dataset contains 891 entries and 15 columns, including:
survived: Survival status (0 = No, 1 = Yes)
pclass : Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
sex: Gender of the passenger
age: Age of the passenger
sibsp : Number of siblings/spouses aboard
parch: Number of parents/children aboard
fare: Fare paid by the passenger
embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
class: Passenger class (categorical)
who: Gender classification (man, woman, child)
adult_male: Whether the passenger is an adult male
deck: Cabin deck
embark_town: Port of embarkation (categorical)

alive: Survival status (alive or not)
alone: Whether the passenger was alone

Data Cleaning and Preprocessing

1. Handling Missing Values:
   The age column had 177 missing values, which were filled with the mean age.
   -The deck column had 688 missing values, and the embark_town and embarked columns had 2 missing values each. These rows were dropped.

2. Data Types:
   The class and deck columns were converted to categorical data types.
   - The adult_male and alone columns were converted to boolean types.

3. Outlier Removal:
   - Outliers in numerical columns (age, sibsp , parch, fare) were identified and removed using the Interquartile Range (IQR) method.

4. Feature Engineering:
   The class column was converted to lowercase for consistency.

5. Dropping Unnecessary Columns:
   The 'who' column was dropped as it was redundant with the sex column.

Univariate Analysis

Numerical Columns

survived: Mean = 0.67, Mode = 1, Median = 1, Variance = 0.22, Skewness = -0.71
pclass : Mean = 1.20, Mode = 1, Median = 1, Variance = 0.28, Skewness = 2.62
age : Mean = 35.06, Mode = 29.70, Median = 33.00, Variance = 224.17, Skewness = 0.13
sibsp : Mean = 0.45, Mode = 0, Median = 0, Variance = 0.40, Skewness = 1.46
parch : Mean = 0.44, Mode = 0, Median = 0, Variance = 0.54, Skewness = 1.62
fare : Mean = 76.31, Mode = 26.55, Median = 55.00, Variance = 5580.81, Skewness = 2.73

Categorical Columns
sex : Male (106), Female (95)
embarked : S (128), C (69), Q (4)
class : First (173), Second (16), Third (12)
embark_town : Southampton (128), Cherbourg (69), Queenstown (4)
alive : Yes (134), No (67)

Observations
The majority of passengers were in the first class.
Most passengers embarked from Southampton.
The survival rate was higher for females compared to males.

Bivariate Analysis

Correlation Analysis
survived vs pclass : There is a negative correlation, indicating that higher-class passengers had a higher survival rate.

survived vs age : The correlation is low, suggesting that age had a minimal impact on survival.
survived vs fare: There is a positive correlation, indicating that passengers who paid higher fares had a better chance of survival.

Survival Rate by Gender

Female : Higher survival rate (95 out of 201)
Male : Lower survival rate (106 out of 201)

Survival Rate by Passenger Class

First Class : Highest survival rate
Second Class : Moderate survival rate
Third Class : Lowest survival rate

Multivariate Analysis

Survival Rate by Gender and Class

First Class Females : Highest survival rate
Third Class Males : Lowest survival rate

Age Distribution by Survival Status

Survived : Slightly younger age distribution compared to non-survivors.
Not Survived : Slightly older age distribution.

Fare Distribution by Survival Status

Survived : Higher fare distribution.
Not Survived : Lower fare distribution.

Key Findings

1. Survival Rate:

- The overall survival rate was approximately 67%.
- Females had a significantly higher survival rate compared to males.
- Passengers in the first class had the highest survival rate, while those in the third class had the lowest.

2. Age and Survival:

- Age had a minimal impact on survival, with a low correlation between age and survival status.

3. Fare and Survival:

- There was a positive correlation between fare and survival, indicating that passengers who paid higher fares had a better chance of survival.

4. Embarkation Point:

- Most passengers embarked from Southampton, and the survival rate was highest among those who embarked from Cherbourg.

5. Family Size:

- Passengers traveling alone had a lower survival rate compared to those traveling with family members.

Conclusion

The analysis reveals that survival on the Titanic was significantly influenced by factors such as gender, passenger class, and fare paid. Females and higher-class passengers had a better chance of survival, while age had a minimal impact. The findings suggest that socio-economic status and gender played crucial roles in determining survival outcomes during the Titanic disaster.