Assignment-1

Data Cleaning and EDA (Exploratory Data Analysis)

Name: Navodit Verma

Enrollment no.: 23114071

Data Set used: Titanic

Analysis Report

Introduction

The Titanic dataset provides information about the passengers aboard the Titanic, including demographic details, ticket class, fare, and survival status. The goal of this analysis is to explore patterns in survival rates based on various features and gain insights into factors affecting survivability. This report includes data preprocessing, exploratory data analysis (EDA), and conclusions drawn from the data.

Data Overview

The dataset consists of the following columns:

1) **PassengerId**: Unique identifier for each passenger.
2) **Survived**: Survival status (0 = No, 1 = Yes).
3) **Pclass**: Ticket class (1st, 2nd, 3rd), indicating socio-economic status.
4) **Name**: Passenger's full name.
5) **Sex**: Gender (male/female).
6) **Age**: Passenger's age in years; missing values require imputation.
7) **SibSp**: Number of siblings/spouses aboard, useful in analysing family survival trends.
8) **Parch**: Number of parents/children aboard, important for family-based survival analysis.
9) **Ticket**: Ticket number, which may contain grouped passengers.
10) **Fare**: Ticket fare paid; higher fares may indicate better accommodations and higher survival chances.
11) **Cabin**: Cabin number; missing for many passengers.
12) **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton), which may correlate with social class.

Data Preprocessing

1. Handling Missing Values

- **Age**: Contains missing values, filled using the median age to maintain statistical integrity.
- **Cabin**: Contains many missing values (more than 50%), making it unreliable for analysis. It is dropped from further consideration.
- **Embarked**: Few missing values (=2 passengers), filled with the most frequent value.

2. Data Encoding and Transformation

- **Sex**: Encoded as 0 (female) and 1 (male) for numerical analysis.
- **Embarked**: Encoded as 0 (C) and 1 (Q) and 2(S) for numerical analysis.
- **Pclass**: Treated as a categorical variable to analyze its impact effectively.
- **Fare and Age Outlier removal**:

Exploratory Data Analysis (EDA)

1. Univariate Analysis

Univariate analysis examines each feature separately to understand its distribution and characteristics.

- Survival Rate: Around 34% of passengers survived.
- Gender Distribution: More males than females, but survival was higher for females.
- Age Distribution: Most passengers were between 20-40 years old, with children and elderly making up a smaller fraction.
- Fare Distribution: Right-skewed, meaning most passengers paid lower fares.
- Pclass Distribution: Majority of passengers were in 3rd class.

2. Bivariate Analysis

Bivariate analysis explores relationships between two variables to identify patterns and trends.

- **Survival rate by gender:**
  - Female: **~68%**
  - Male: **~18%**
  - Women had a significantly higher survival rate, likely due to the "women and children first" policy.

- **Survival Rate by Pclass**
  - Passengers in 1st class had the highest survival probability, followed by 2nd class, while 3rd class passengers had the lowest chances of survival.
    - 1^{st}: **~51%**
    - 2^{nd}: **~48%**
    - **3^{rd}: ~24%**
  - This suggests that wealth and social status played a crucial role in access to lifeboats.

- **Survival by Age**
  - Infants (Age < 10) had a higher survival rate**,** likely due to the "children first" evacuation policy.
  - Older passengers (Age > 50) had a lower survival probability, indicating that mobility and physical condition might have influenced survival.
  - A broader look at the age distribution suggests that survival rates were highest among younger individuals.

- **Family Size and Survival Impact**

  - Small families (1-3 members) had a higher survival rate than solo travelers or large families (>5 members).
  - Solo travelers had lower survival rates, possibly due to difficulty securing lifeboat spots.
  - Large families might have struggled to stay together, leading to lower survival rates.

- **Fare and Survival**

  - Higher fare-paying passengers had better survival chances, which aligns with the Pclass survival trend.
  - Passengers who paid higher fares were often in first class and had better access to lifeboats.
  - Fare distribution suggests a correlation between economic status and survival probability.

- **Embarkation Point and Survival**

  - Passengers who embarked from Cherbourg (C) had the highest survival rate, possibly due to a larger number of first-class passengers.
  - Passengers from Southampton (S) had the lowest survival rate, likely because most third-class passengers boarded here.
  - Queenstown (Q) passengers had moderate survival rates, though fewer passengers embarked from this port.

3. Multivariate Analysis

Multivariate analysis studies the combined effect of multiple features on survival.

- **Gender, Pclass, and Survival**:
  - Female 1st class passengers had the highest survival rates

        - Male 3rd class passengers had the lowest survival rates
    - **Age, Pclass, and Survival**:
        - Young children in 1st class had the best chances of survival
        - Elderly in lower classes had poor survival rates
    - **Family Size and Survival**:
        - Small families (1-3 members) had better survival chances
        - Large families (>5 members) had lower survival rates

Conclusion

- Gender and Class were the strongest indicators of survival: Women and first-class passengers had significantly higher survival rates.
- Younger passengers, particularly children, had a better chance of survival, aligning with the "women and children first" evacuation practice.
- Higher ticket fare and smaller family sizes improved survival chances, reflecting the advantage of wealth and family support.
- Embarkation point had a minor but noticeable impact, with Cherbourg passengers having slightly better survival odds.

This analysis provides a comprehensive look into the Titanic dataset, highlighting key factors influencing survival rates. Further analysis using machine learning models could help quantify the impact of each feature on survival more precisely.