DAI – 101
Assignment - 1
Data Cleaning and EDA (Exploratory Data Analysis)

Harshit Kumar Meena
Enrollment Number: 23114037
Dataset Used: Titanic

Contents

1 Introduction

The **Titanic dataset** provides details about passengers on the Titanic, including demographic information, ticket details, and survival status. This analysis aims to clean the data, explore its structure, and derive meaningful insights about survival factors.

2 Data Review

The dataset consists of **891 rows and 12 columns**, describing various attributes of Titanic passengers. Below is an overview of each column:

- **Survived**: Indicates whether the passenger survived (1) or not (0).
- **Pclass**: Passenger class (1 = First, 2 = Second, 3 = Third).
- **Name**: Passenger's full name.
- **Sex**: Gender of the passenger (male/female).
- **Age**: Age of the passenger.
- **SibSp**: Number of siblings/spouses aboard the Titanic.
- **Parch**: Number of parents/children aboard the Titanic.
- **Ticket**: Ticket number.
- **Fare**: Fare paid for the ticket.
- **Cabin**: Cabin number (many missing values).
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

3 Data Cleaning

3.1 Handling Missing Values

- The **Cabin** column contains too many missing values and is therefore dropped.
- The **PassengerId** column does not contribute to the analysis and is also dropped.
- The **Age** column has 177 missing values, which are imputed using the median of available values.
- The **Embarked** column, a categorical variable, has missing values filled with its mode.

3.2 Handling Duplicate Values

The dataset was checked for duplicates, and no duplicate records were found.

3.3 Outlier Detection and Removal

- Boxplot analysis revealed outliers in the dataset.
- Outliers were identified in numerical columns such as **Fare** and **Age**.
- Extreme outliers were handled to prevent them from skewing the analysis.

4 Exploratory Data Analysis (EDA)

4.1 Univariate Analysis

- **Survival Rate**: Around 38% of the passengers survived, while 62% did not.
- **Age Distribution**: The majority of passengers were between 20-40 years old.
- **Gender Distribution**: There were more male passengers than female passengers.
- **Passenger Class**: Most passengers belonged to the third class.
- **Embarkation Ports**: Most passengers embarked from 'S' (Southampton).

4.2 Bivariate Analysis

- **Survival Rate by Gender**: Female passengers had a significantly higher survival rate (74%) compared to male passengers (19%), aligning with the "women and children first" policy.
- **Survival Rate by Passenger Class**: First-class passengers had the highest survival rate (63%), followed by second-class (48%), and third-class (25%).
- **Effect of Embarkation Port**: Passengers who embarked from 'C' (Cherbourg) had the highest survival rate (55%), followed by 'Q' (Queenstown) and 'S' (Southampton).
- **Fare vs. Survival**: Higher ticket prices were associated with higher survival rates, as first-class passengers generally paid more and had a better chance of survival.

4.3 Multivariate Analysis

- **Gender & Class Influence on Survival**: First-class female passengers had the highest survival rate, while third-class male passengers had the lowest.
- **Age & Survival Across Classes**: Children (age ¡ 10) in first and second class had a much higher survival rate compared to adults in third class.
- **Embarkation, Class, and Survival**: First-class passengers from Cherbourg had the highest survival rate, whereas third-class passengers from Southampton had the lowest.

5 Conclusions

- **Gender, class, and fare price** were strong determinants of survival.
- **Women and children** had higher survival rates.
- **First-class passengers** had better survival chances than lower-class passengers.
- **The embarkation point** had some impact on survival, potentially due to economic status differences.