TASK

Exploratory Data Analysis on the Titanic Dataset

Visit our website

Introduction

SUMMARY OF THE DATASET

The Titanic dataset contains information about passengers aboard the RMS Titanic, including their demographic data, cabin class, fare, ticket information, survival status, etc. The goal is to predict whether a passenger survived or not based on various features.

Titanic Data Description:

Survival: Survival (0 = No; 1 = Yes)

Pclass: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

Name: Name

Sex: Sex

Age: Age

Sibsp: Number of Siblings/Spouses Aboard

Parch: Number of Parents/Children Aboard

Ticket: Ticket Number

Fare: Passenger Fare

Cabin: Cabin

Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

DATA CLEANING

Before performing EDA, it's essential to clean the dataset. Here are some common cleaning steps for the Titanic dataset:

- Handling missing values: Many columns in the dataset have missing values, which need to be addressed.
- Removing irrelevant columns: Some columns may not contribute significantly to the analysis or modelling, so they can be dropped.
- Converting data types: Convert columns to appropriate data types for analysis and visualization.

- Handling outliers: Identify and handle any outliers in the dataset.

MISSING DATA

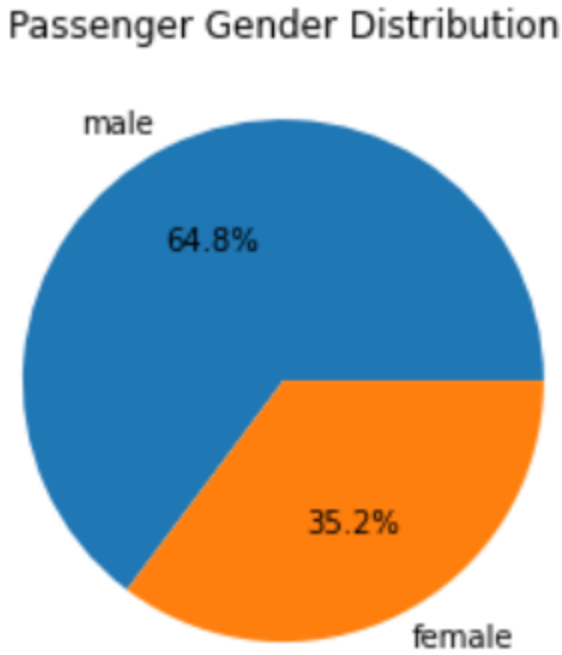
Missing data is a common issue in datasets. In the Titanic dataset, missing values can be found in columns like "Age," "Cabin," and "Embarked." Different strategies can be used to handle missing data:
- For numerical features like "Age," missing values can be imputed with the mean, median, or a custom value based on the data distribution.
- For categorical features like "Embarked," missing values can be imputed with the mode (most frequent value) or a custom value.
- The "Cabin" column has many missing values, so it can be dropped from the analysis.

DATA STORIES AND VISUALISATIONS

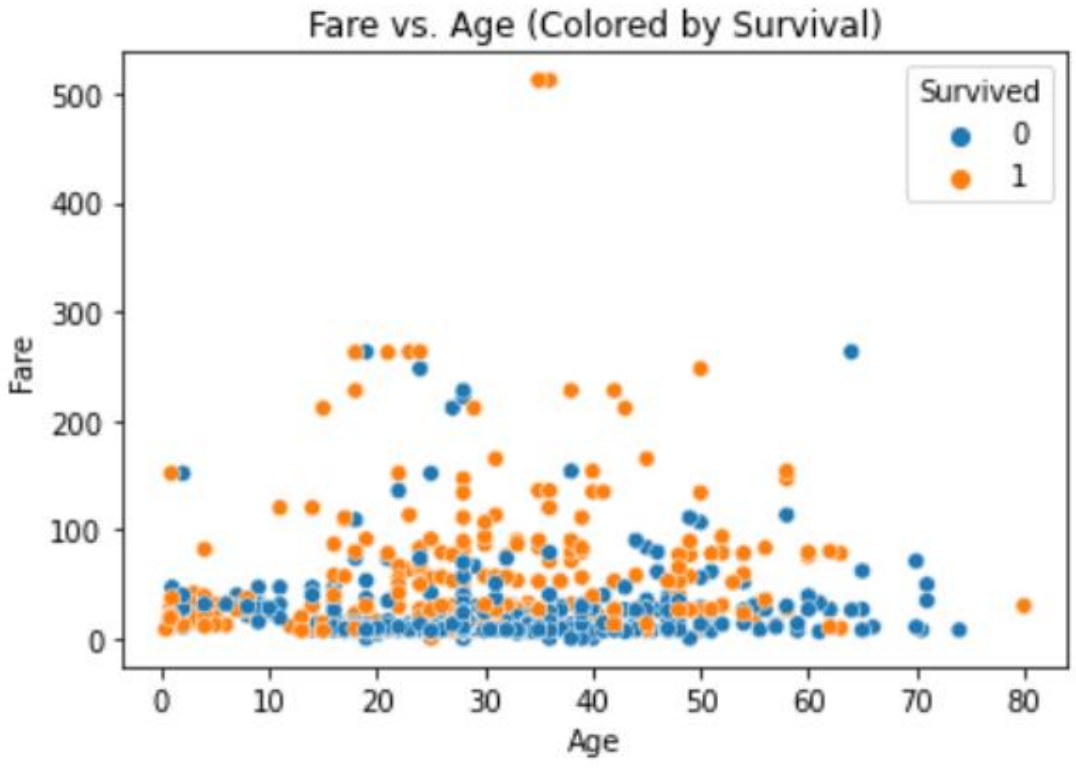
Passenger Gender Distribution

The pie chart of passenger gender distribution, which shows the percent age of male and female passengers on board. We can see that more than h

alf of the passengers were male (64.8%), while less than half were fema
le (35.2%). This might reflect the gender norms and roles at that time,
as men were more likely to travel for work or leisure, while women were
more likely to stay at home or accompany their husbands.

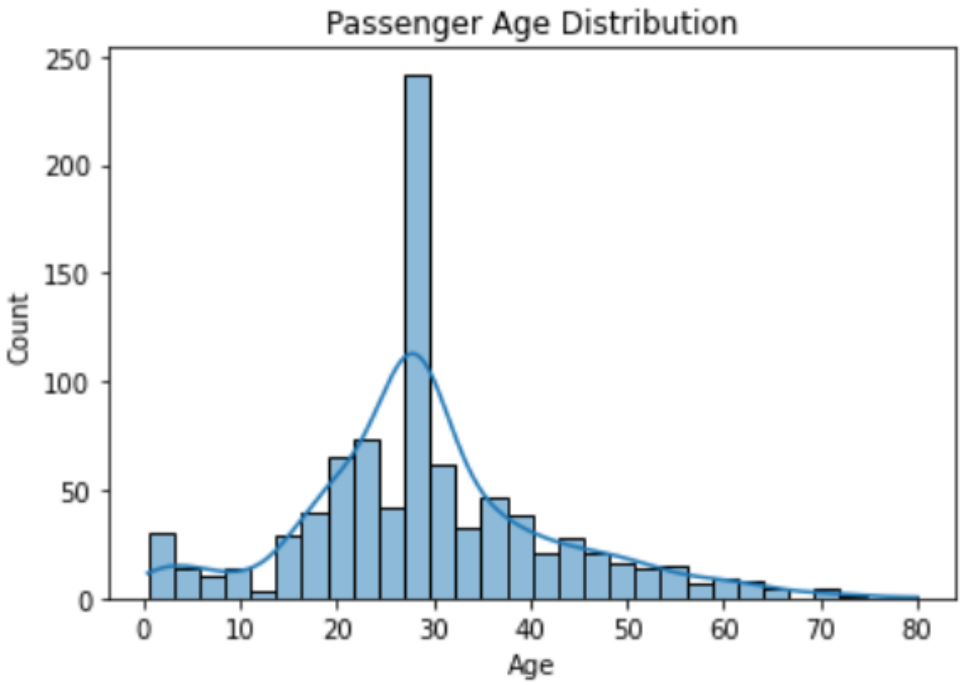


Fare against the Age



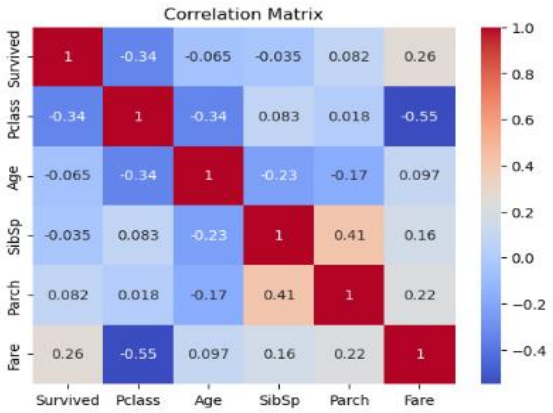
Passenger Age Distribution

We can see that most of the passengers were in their 20s or 30s, maybe they were looking for some adventure or romance on board. We can also see that there were some children and elderly people on board, who might have been traveling with their families or friends.

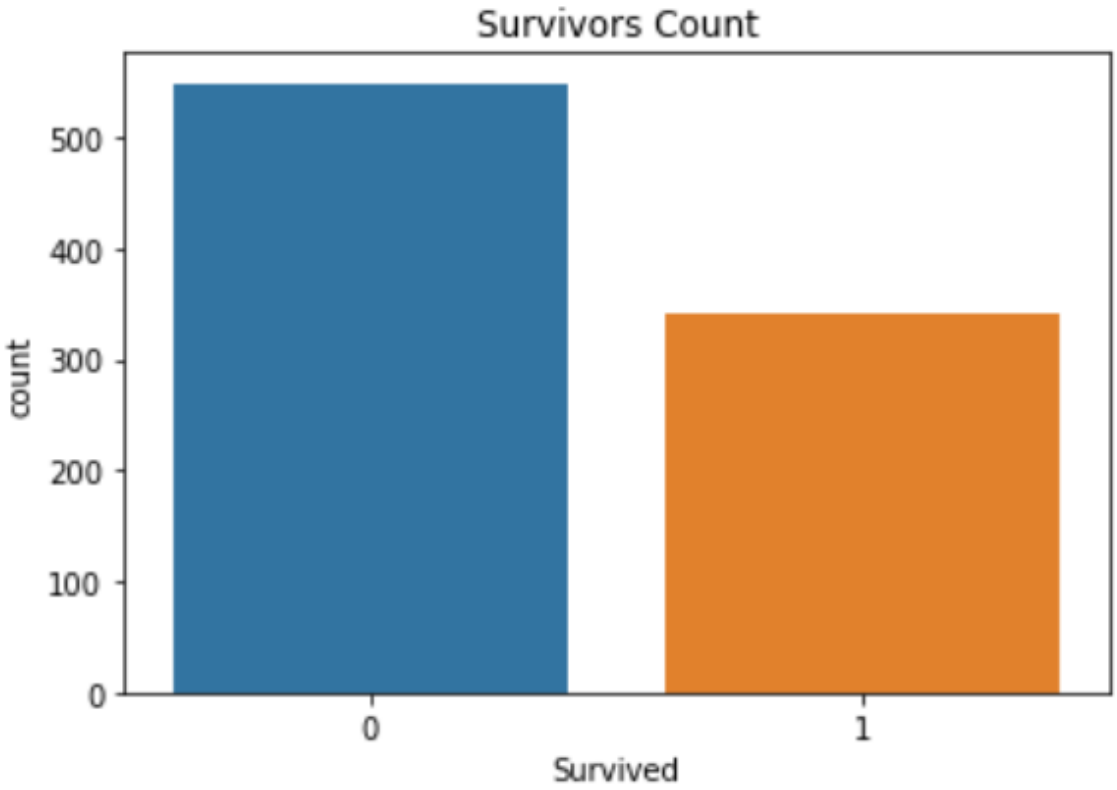


Correlation Matrix

A positive correlation means that as one variable increases, the other variable also increases. A negative correlation means that as one variable increases, the other variable decreases. A zero correlation means that there is no relationship between the two variables. We can see that some of the variables have strong positive or negative correlations, such as Pclass and Fare (-0.55), Pclass and Survived (-0.34), etc. These correlations indicate that these variables have a significant impact on each other and on the survival outcome



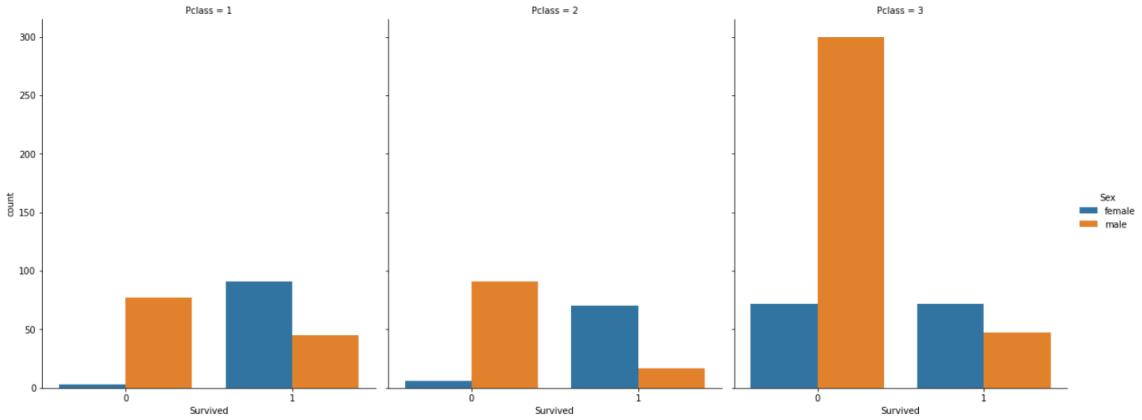
Distribution of passengers who survived during wreck of the Titanic



How class contributed to the survival of passengers in the Titanic

- The survival rate was higher for passengers in higher classes than lower classes. This is consistent with the historical fact that passengers in higher classes had better access to the lifeboats and were more likely to be rescued.
- The survival rate was lowest for male passengers in the 3rd class, as they had the least priority and access to the lifeboats. Only a few of them survived the disaster.

- The survival rate was highest for female passengers in the 1st class, as they had the most priority and access to the lifeboats. Most of them survived the disaster.

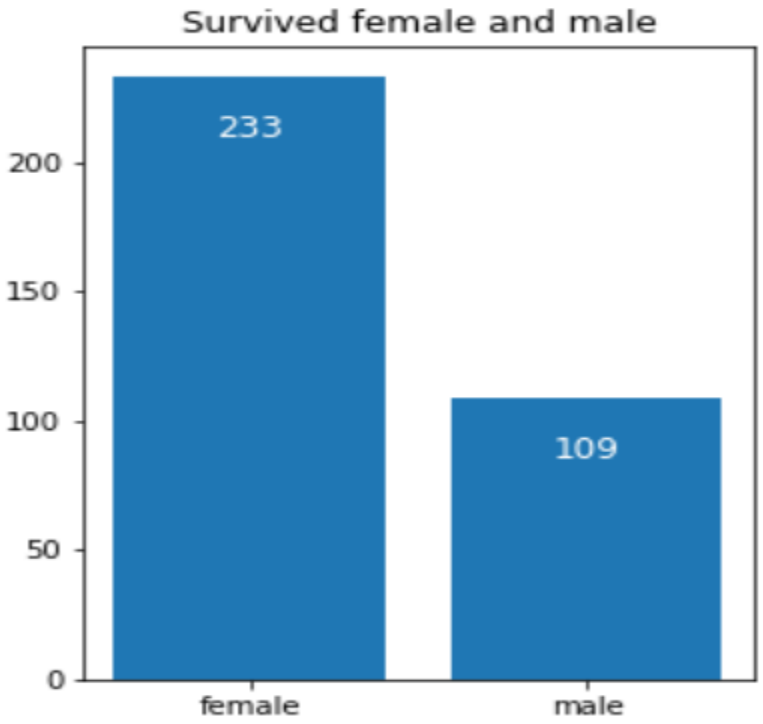


How gender affected survival

- The number of female survivors is much higher than the number of male survivors. There are 233 female survivors and only 109 male survivors, which means that females have more than twice the chance of surviving than males.
- This is consistent with the historical fact that women were given priority to board the lifeboats, as part of the "women and children first" policy. This policy was followed by some of the crew members and

passengers, who believed that women and children should be saved before men.

- The chart also reflects the gender inequality and social norms of the time, which favoured women over men in terms of protection and safety. Men were expected to be brave and selfless, and to sacrifice themselves for women and children.



THIS REPORT WAS WRITTEN BY : Ndatadzeyi B Chiota