Hyperiondev

TASK

Exploratory Data Analysis on the Titanic Dataset

Visit our website

Introduction

The sinking of the Titanic is a renowned maritime disaster that occurred on April 15, 1912, during its inaugural voyage. Despite being hailed as "unsinkable," the RMS Titanic collided with an iceberg, leading to its tragic downfall. Regrettably, the limited availability of lifeboats resulted in the loss of 1,502 out of 2,224 passengers and crew members.
Although chance played a role in survival, certain demographics appeared to have a higher likelihood of surviving than others.

The datasets titled Titanic includes passenger information like name, age, gender, socio-economic class, etc. It contains the details of a subset of the passengers on board (891 to be exact) and whether they survived or not.

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	

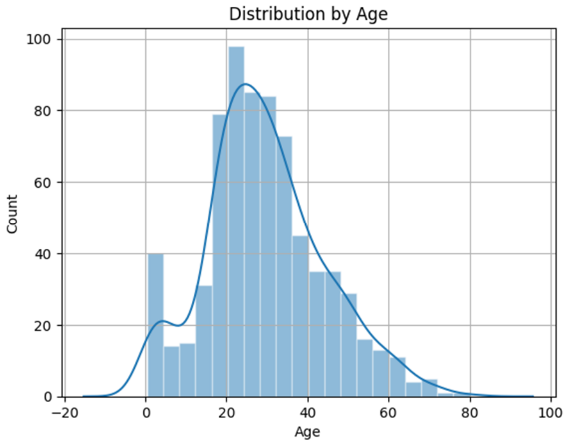
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

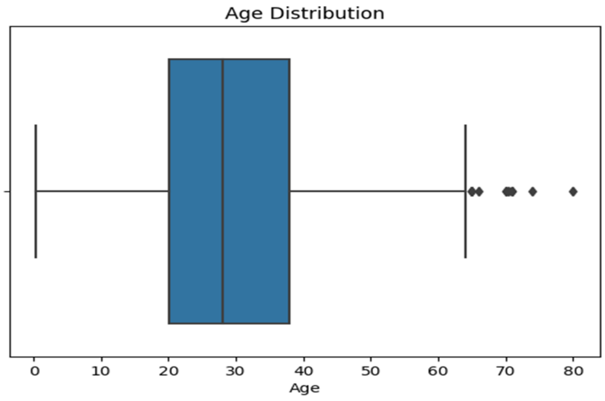
DATA CLEANING

During the process of data cleaning, irrelevant information, duplicated data, null data or missing values were explored and identified. All the unique values, null data, and duplicated data were revealed. Bar chart, box plot, scatter plot and missingno matrix were used to observe the outliers and null data or missing values.

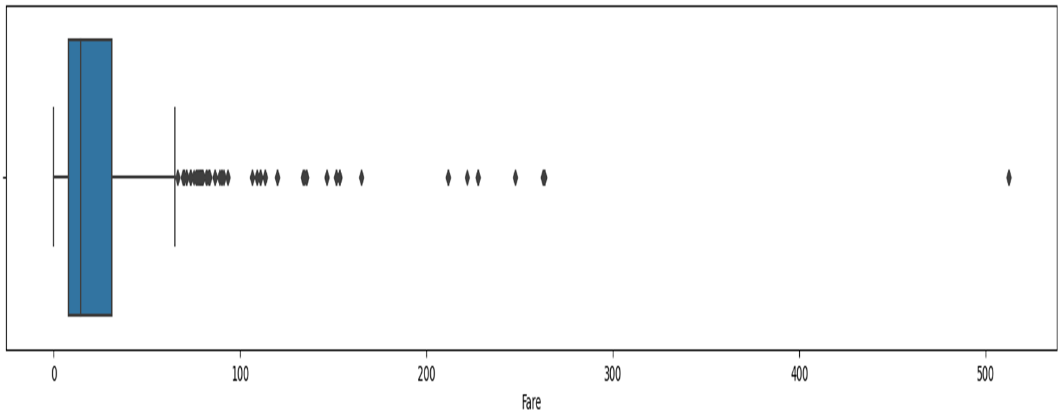
In the exploratory data analysis (EDA) on the titanic dataset, survival is the primary interest. Passenger Id, Name and Ticket are considered to be irrelevant information after exploring the data types and nature of the data. No duplicated data was found. Null data/missing values were found.

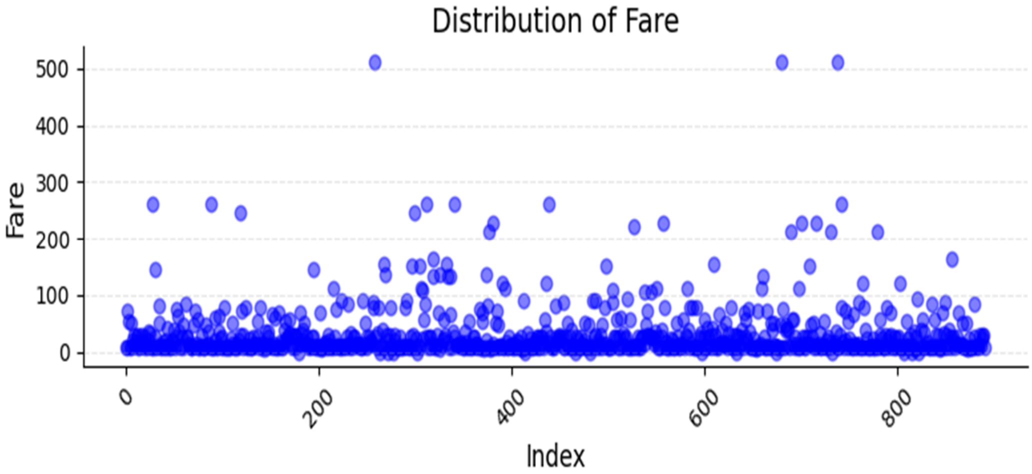
Age:



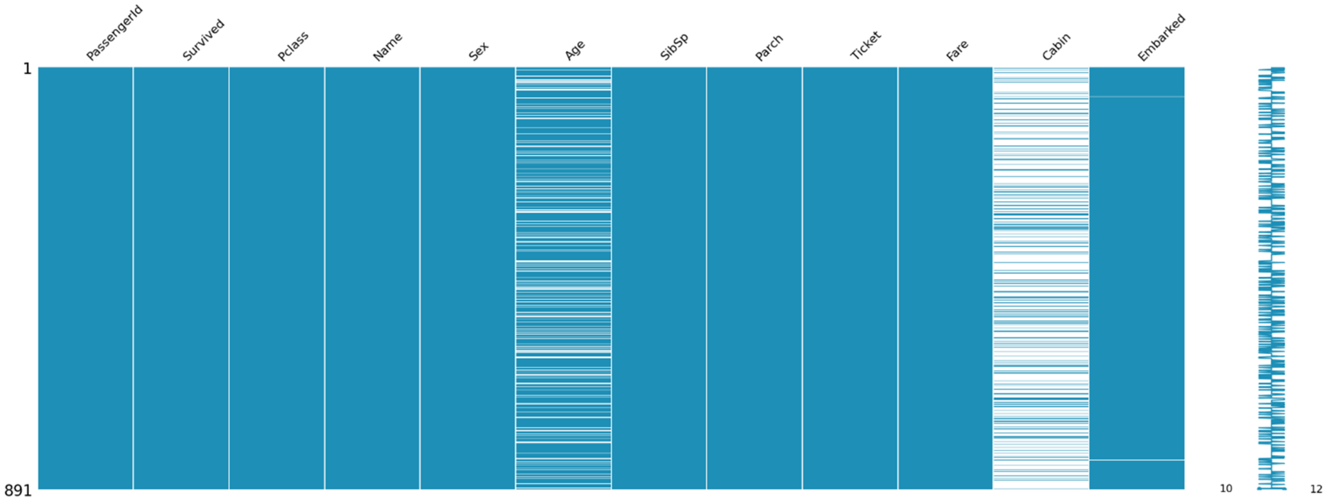


Fare:





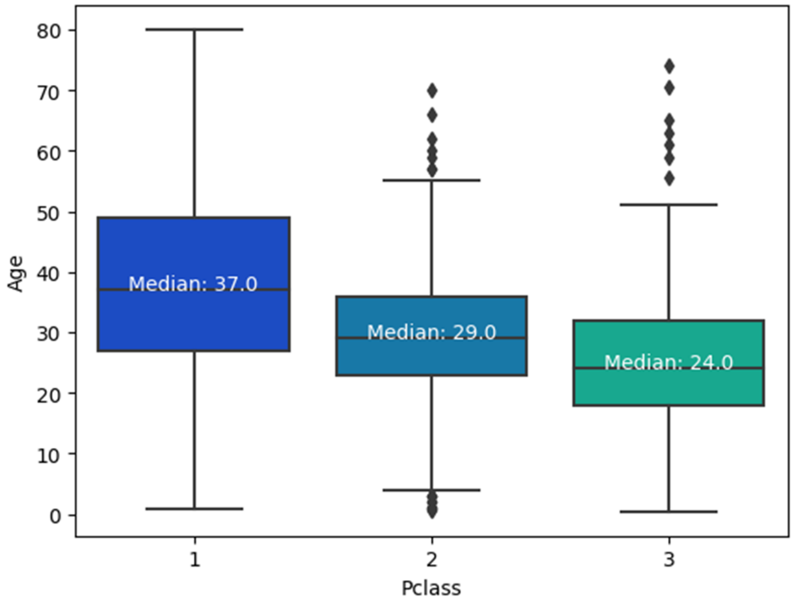
The data outliers were considered to have significant values in EDA, none of them were removed. However, the variables of Passenger Id, Name and Ticket were dropped (irrelevant information).



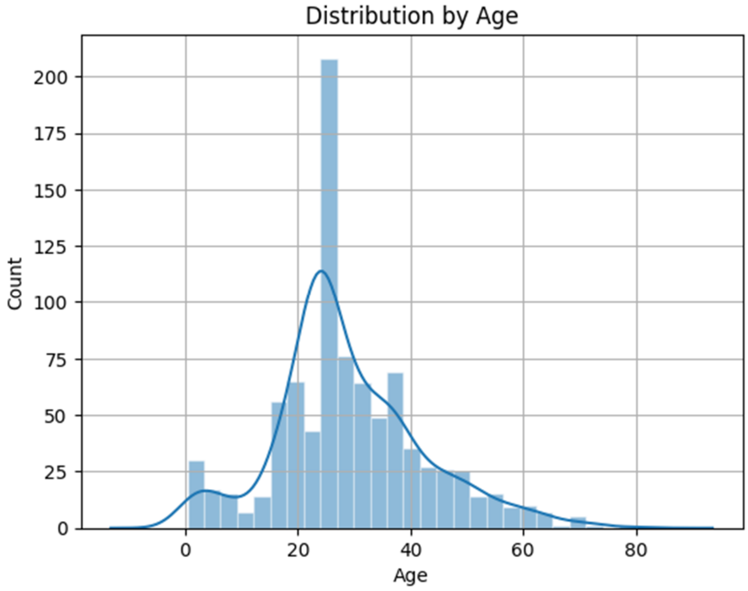
Null values illustrated in missingno matrix above consist 77.1% (687/891), 19.9%(177/891) and 0.2%(2/891) in the variables of Cabin, Age and Embarked respectively. Theoretically, 25 to 30% is the maximum missing values are allowed, beyond which we might want to drop the variable from analysis. Variables of Cabin was dropped from the dataset.

MISSING DATA

687, 177 and 2 missing values were identified in Cabin, Age and Embarked respectively.
Cabin was dropped from the dataset due to the high proportion (77.1%) of missing values to the whole dataset. Considering that we cannot easily replace the missing values with a reasonable value, the 2 missing values in Embarked was also dropped.
Regarding the 177 missing values in Age, there were replaced with the median values of age according to the Pclass the individual data belongs to (ie. 37, 29 and 24 respectively), as showed the following box plot.



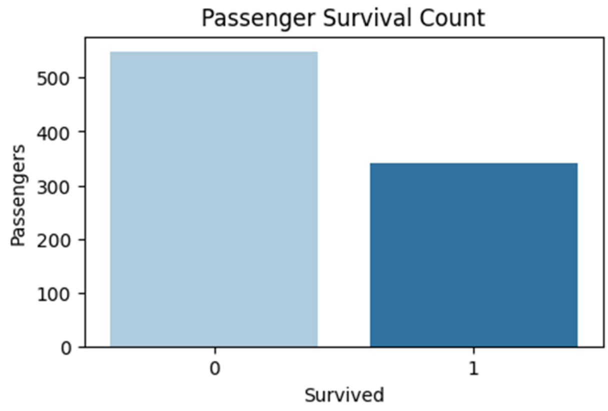
After handling the missing values, the age distribution changed as follows:

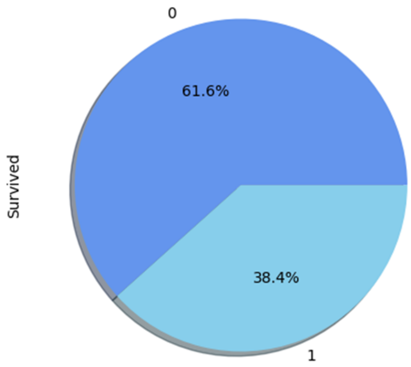


DATA STORIES AND VISUALISATIONS

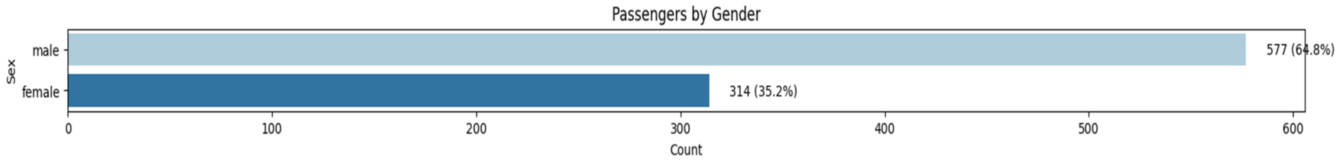
General statistics of the titanic dataset:

Survival:

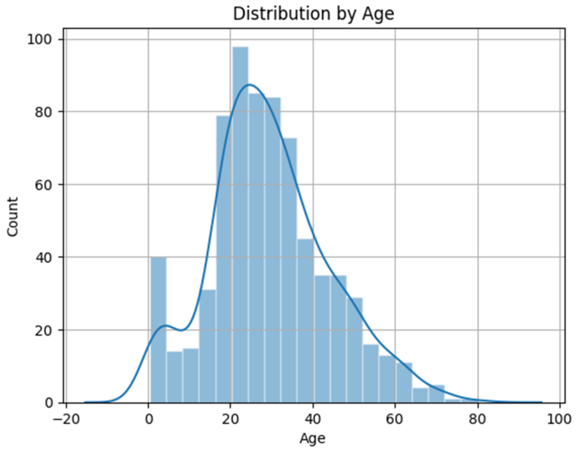


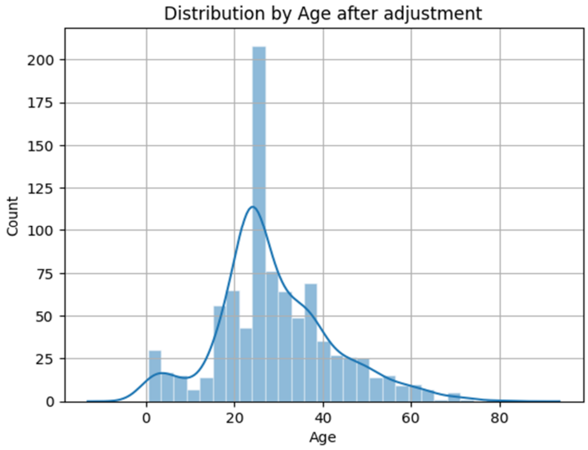


Sex:

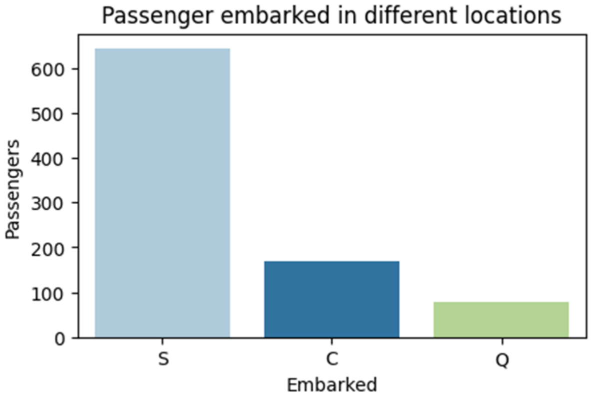


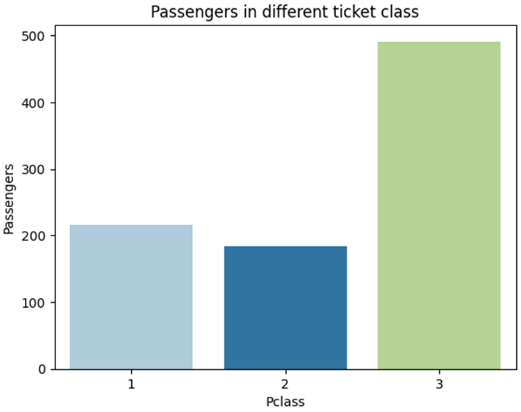
Age (before and after replacement of missing values):





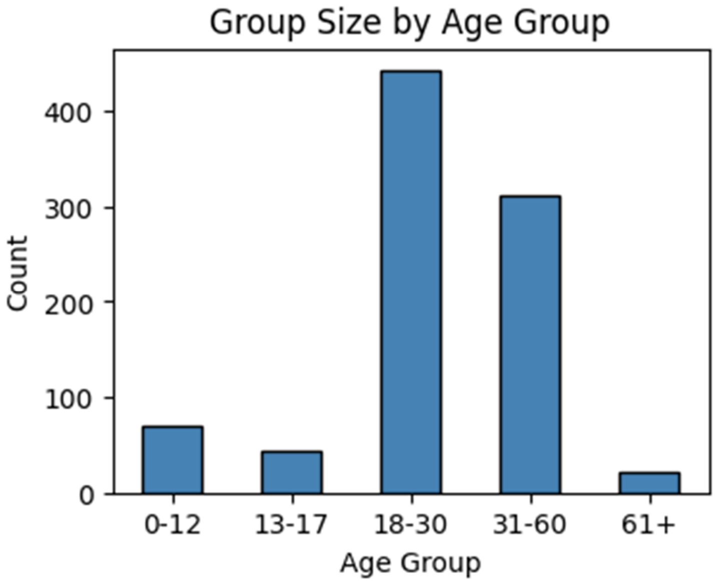
Embarked & Pclass:



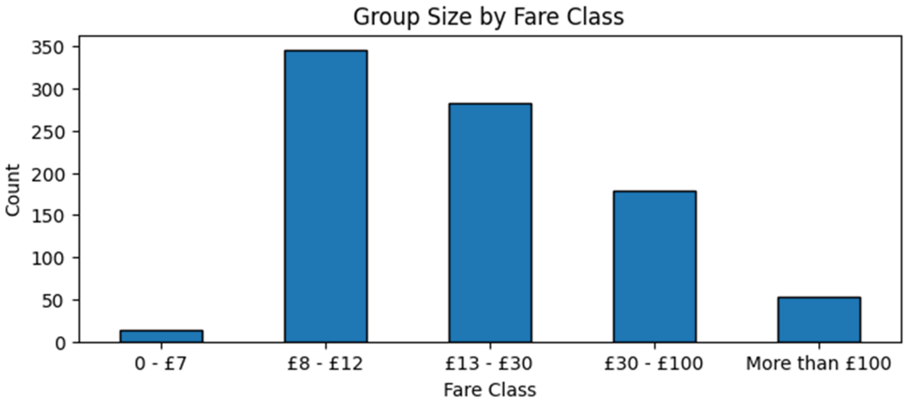


From the charts above, we have an overview of the general statistics of the dataset:
The survival rate is around 38%, which is lower than the death: around 62%.
The male gender is dominant (65%) compared with female (35%).
People in middle age were the major passengers in dataset.
Cherbourg is the most popular port for the passengers to embark.
Majority of passengers were paying class 3 tickets.

Age group and fare class:
To understand the contributing factor of age, age groups were divided as following (based on the fact that 18 is the age in the UK considered as adult and the above age distribution): Group 1: from 0 - 12 (Children), Group 2: from 13 - 17 (Adolescents), Group 3: from 18 - 30 (Adults 1), Group 4: from 31 - 60 (Adults 2), Group 5: Ages 61+ (Seniors).

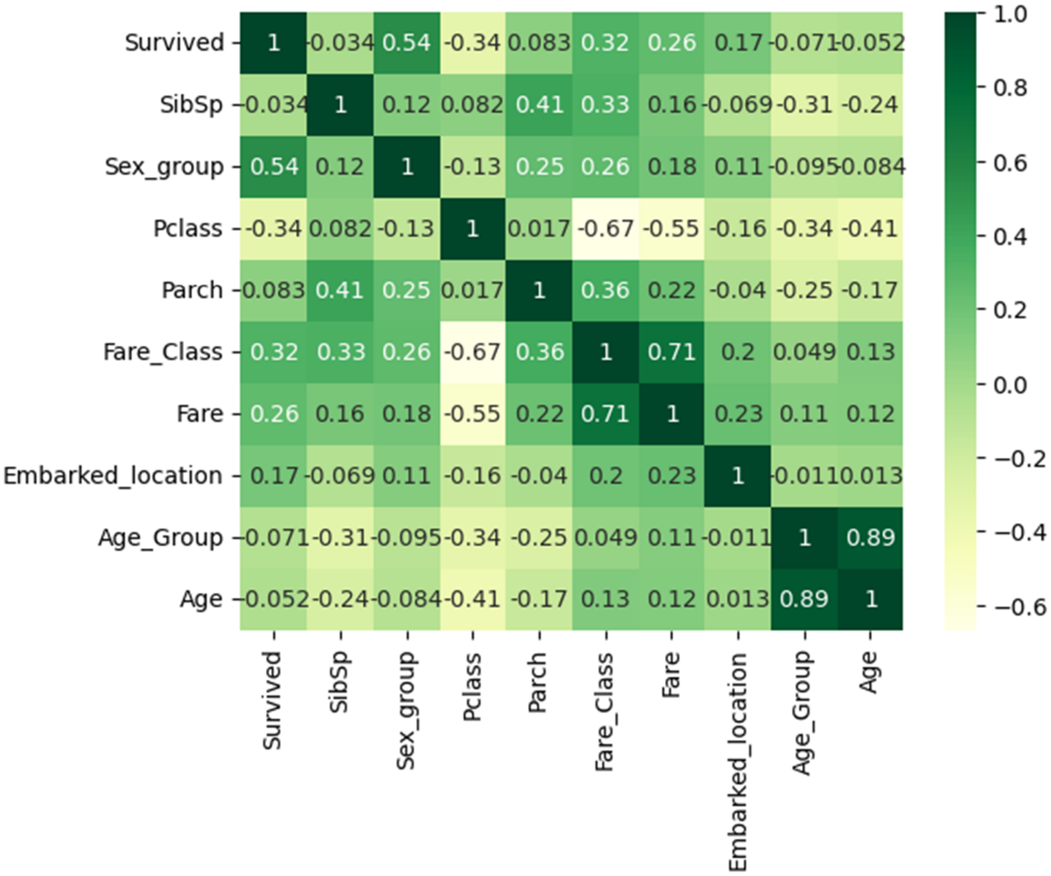


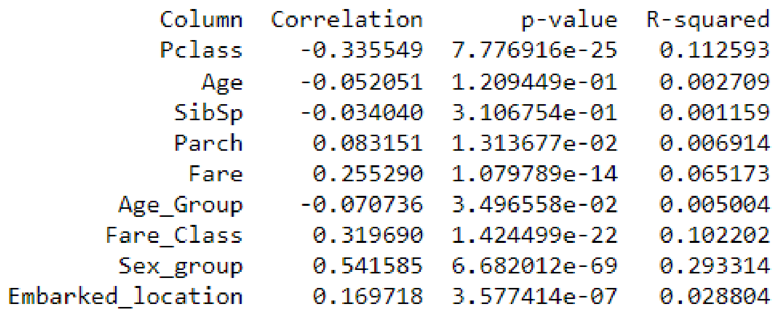
In 1912, the fare was divided in different class: third Class: £7, second Class: £12, first class berth: £30, first class suite: £870. Fare classes were divided as following: group 1: from 0 - £7, group 2: from £8 - £12, group 3: from £13- £30, group 4: from £30- £100, group 5: more than £100.



Exploratory Data Analysis

Heatmap:

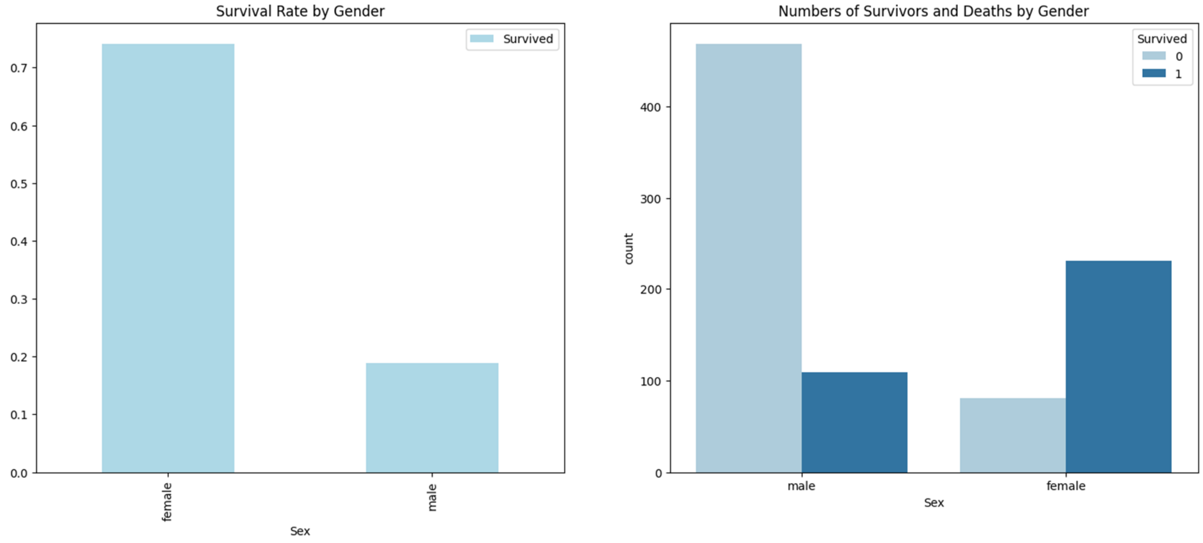




```
          Column  Correlation       p-value  R-squared
          Pclass    -0.335549  7.776916e-25   0.112593
             Age    -0.052051  1.209449e-01   0.002709
           SibSp    -0.034040  3.106754e-01   0.001159
           Parch     0.083151  1.313677e-02   0.006914
            Fare     0.255290  1.079789e-14   0.065173
       Age_Group    -0.070736  3.496558e-02   0.005004
      Fare_Class     0.319690  1.424499e-22   0.102202
       Sex_group     0.541585  6.682012e-69   0.293314
Embarked_location    0.169718  3.577414e-07   0.028804
```

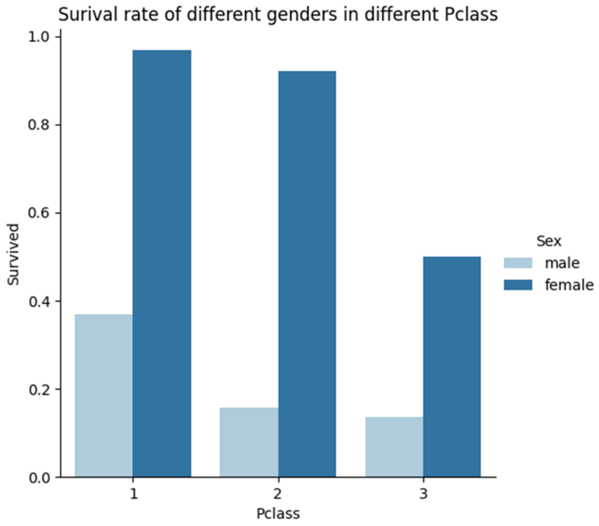
From the heatmap, strong correlations found between: gender and survival: 0.54 (positive correlation), Pclass and survival: -0.34 (negative correlation), fare/fare class and survival: 0.26/0.32 (positive correlation).
Overall, the variables with significant relationships (based on low p-values(< 0.05)) with the survived are: Pclass, parch, fare, age group, fare class, gender, and embarked location. These variables have correlations that suggest a meaningful association with the survival outcome. However, it's important to consider other factors such as effect size, domain knowledge, and potential confounding variables when interpreting the significance of relationships.

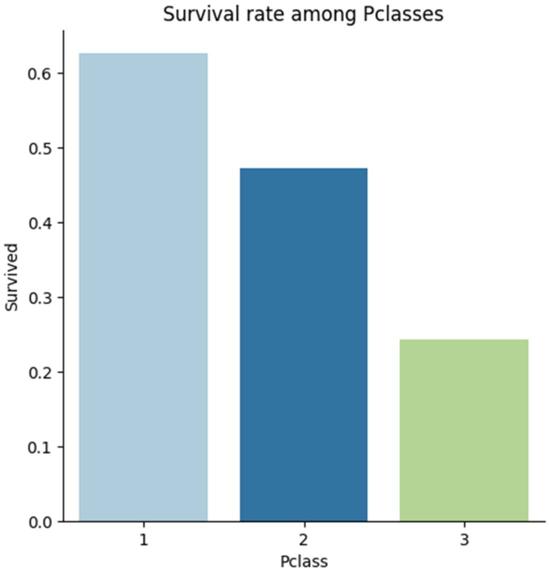
Gender:
Gender is the most important factor in determining survival of the Titanic incident. Female was likely to survive compared with male. Sex has the highest correlation coefficient (0.541585) among all the other variables, and its p-values is 6.682012 × 10^{-69} which is <0.05. The following graphs will show the survival in different genders:



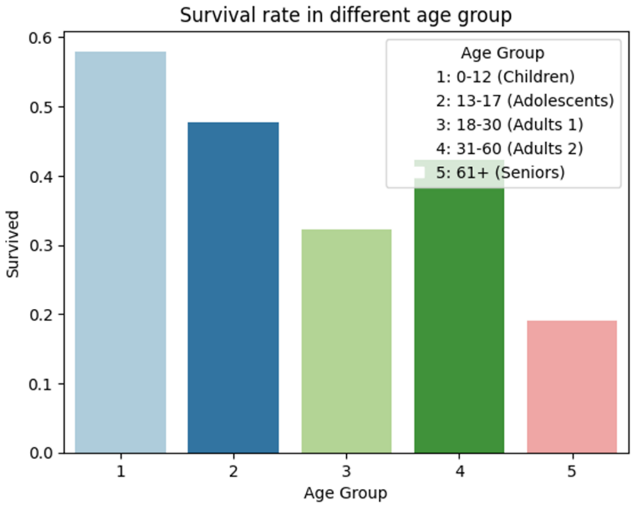
The dataset showed us this could be the important factor affect the survival. However, we knew that the dataset contains the data with 577 male and 314 female, which may be not accurate enough to decide this factor is significantly affecting the result due to the significant sample size difference.

Pclass:
The second important factor in determining survival of the Titanic incident is Pclass, the upper-class the passengers were in, the higher chance to survive. Pclass has the correlation coefficient: -0.338481 (negative correlation), and its p-values is 7.776916 × 10 $^{-25}$ which is <0.05. ‘The upper-class passengers were given preference on lifeboats’ in the movie does make sense. The following bar chart also illustrates this statement makes sense to both genders. We have a significant and strong correlation between the Pclass/gender and survival, females at class 1 and 2 are almost certain to survive! While males have much lower survival chance across all classes.

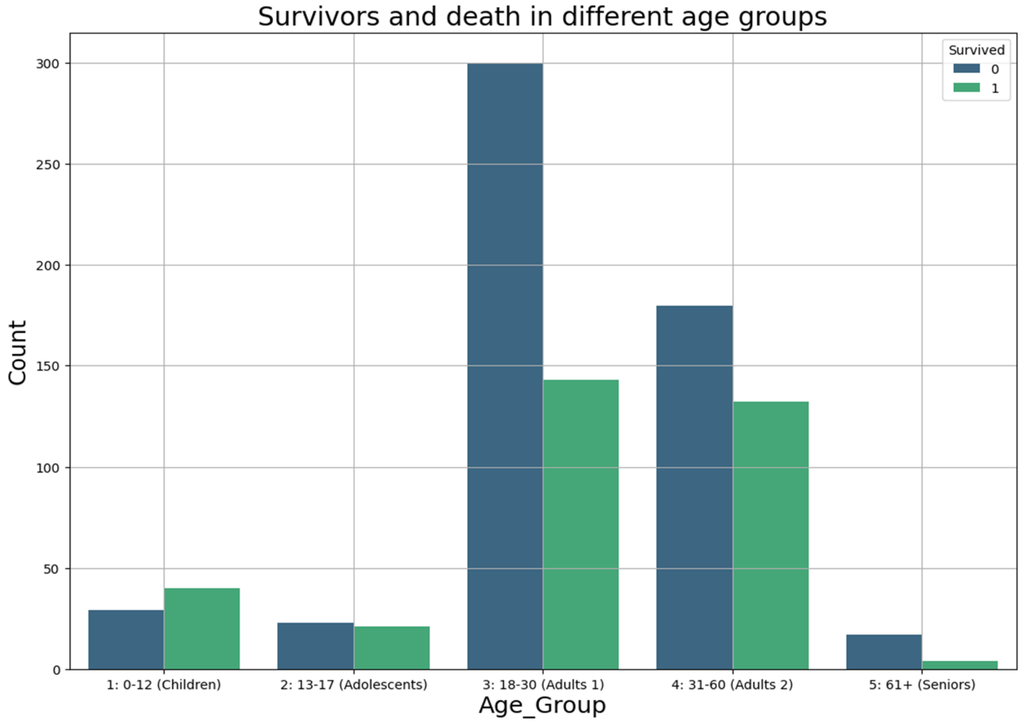




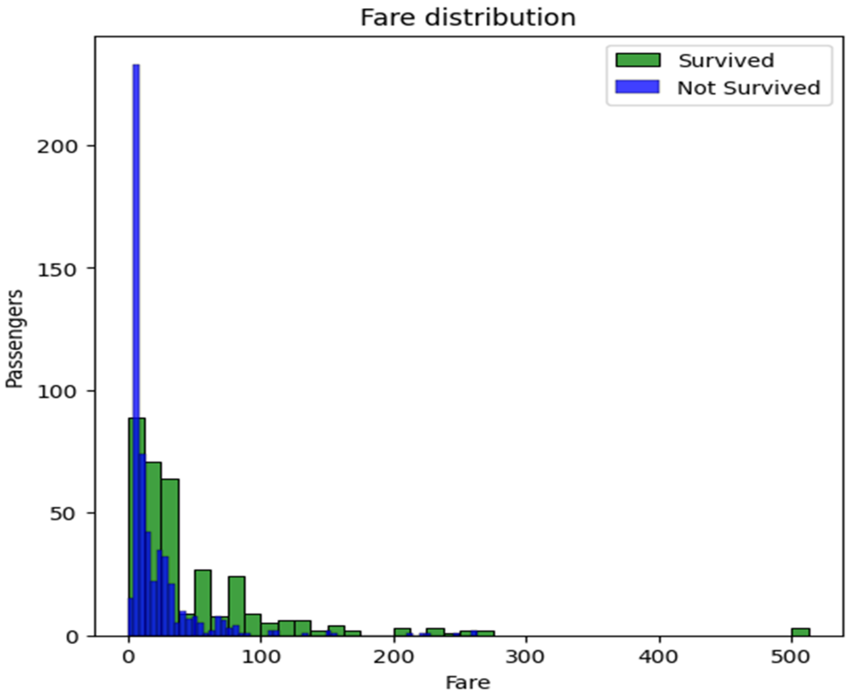
Age:
The correlation coefficient between age and survival is -0.052051, and the p-value is 1.209449 x 10-1, which is weak and insignificant negative correlation.
However, age group has correlation coefficient: -0.070736 with p-value turns to 3.496558 x 10^{-2}, which indicates a weak and significant negative correlation. It illustrates certain age groups have a higher likelihood of survival. “Women and children first” is likely correct.

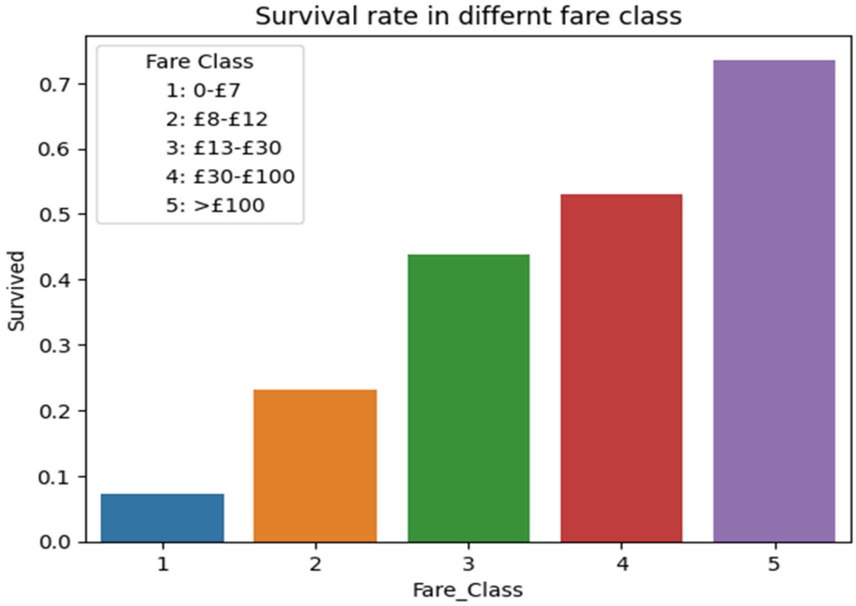


Except children group, the death number is higher than the survival number in all other age group:

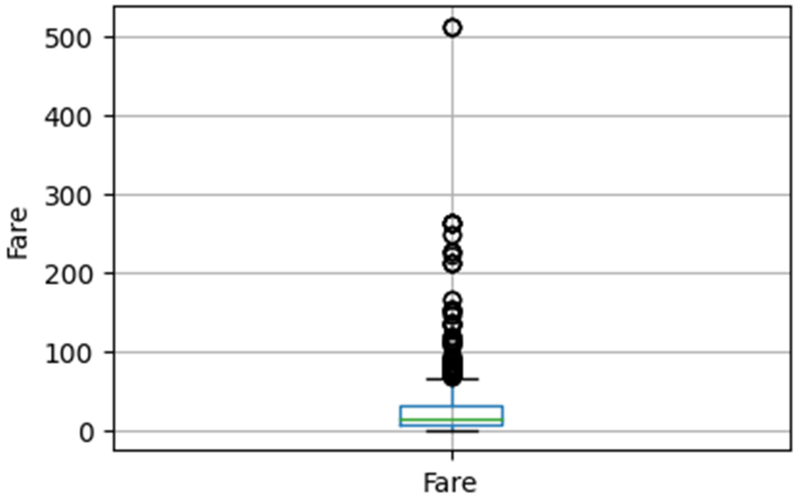


Fare:
The people who paid more expensive fares had more chances of survival:

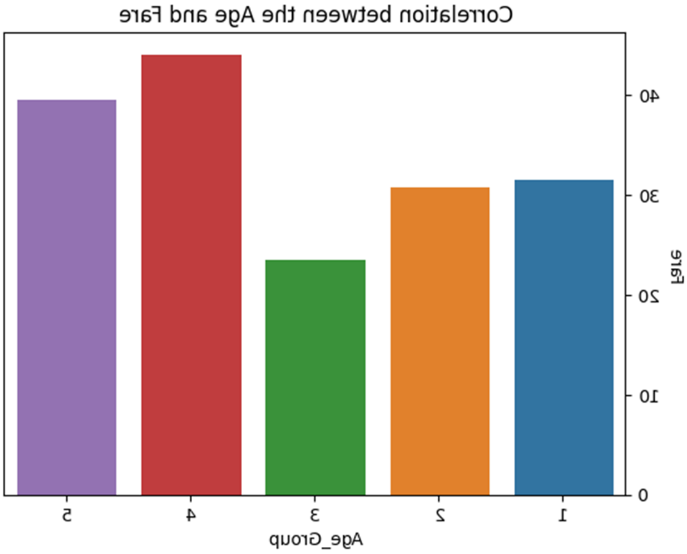




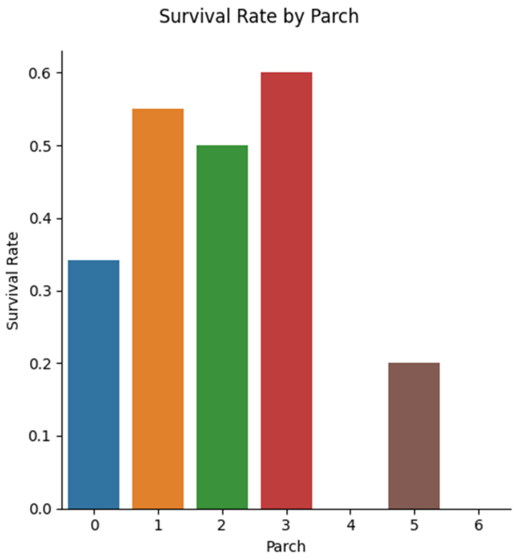
Many people may pay more than it needs for the tickets:

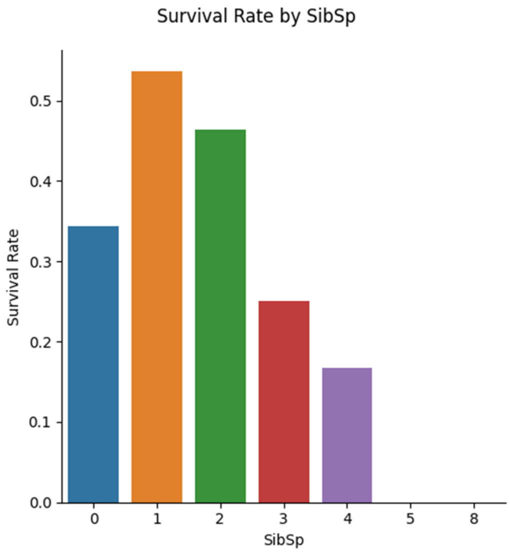


Older people generally paid the higher fares:

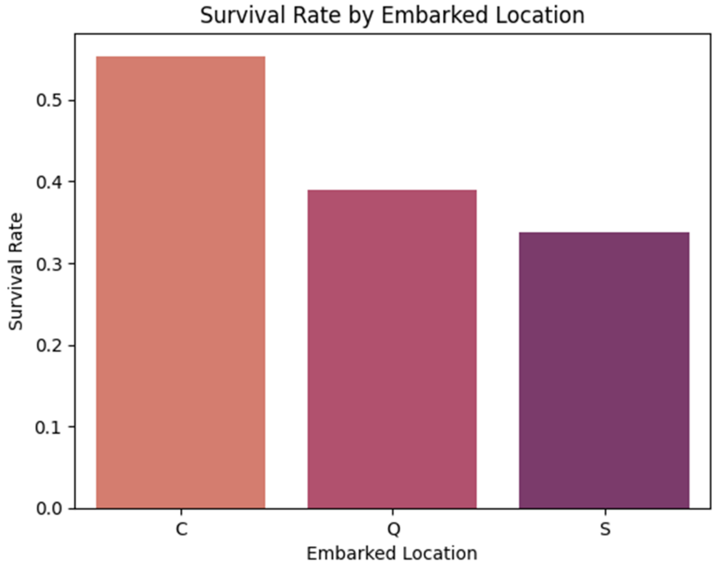


SibSp & Parch:
Passengers who were alone had a lower chance to get rescued:





Embarked location:
People embarked in Cherbourg had higher chance to survive:



In conclusion, the correlations of different variables with survival were explored. The strong and significant correlations include gender, Pclass and fare/fare class. There are some significant but weak correlations with parch, age group and embarked location.

THIS REPORT WAS WRITTEN BY : Chu , Alex Wai Leung