Report on the Titanic Data Set

Dimensions of the data set

In total, there are ***890*** *observations* and ***12*** *variables.*

Occurrence of missing values

There are missing values in the following variables:

- Age: 177
- Cabin: 687
- Embarked: 2

Summary statistics

Numerical variables:

1) **Age**
   Age range of passengers:
   5 months - 80 years
   Average age: 29.69 years
   Median age: 28
   Mode: 24
   - Distribution skewed to the right.

2) **Fare**
   Fare range of tickets: 0 - 512.33
   Average fare: 32.2
   Median fare: 14.4542
   Mode: 8
   - Distribution skewed to the right.

3) **No. of siblings / spouses aboard**
   Range of no. of siblings/spouses aboard: 0 - 8
   Median & Mode = 0
   - Distribution skewed to the right.

   We can hereby conclude that most passengers did not have their siblings/spouses on board with them.

4) **No. of parents/children aboard**
   Range of no. of parents/children aboard: 0 - 6
   Median & Mode = 0
   - Distribution skewed to the right.

   We can hereby conclude that most passengers did not have their parents/children on board with them.

Categorical variables:

1) **Survival**
   Distributed as follows-
   342 lived
   549 passed

   General survival rate = 0.38

2) **Pclass**
   Distributed as follows-
   1st Class 216
   2nd Class 184
   3rd Class 491

Largest proportion of passengers found in 3rd class, followed by 1st class and the least found in 2nd class.

3) **Sex**
Distributed as follows-
Male 577
Female 314

There were a greater number of males on board.

4) **Embarked**
Distributed as follows-
S 644
C 168
Q 77

Most passengers embarked from the Port of Southampton followed by Cherbourg and Queenstown.

Possible Hypotheses

- Determining if the survival rate is associated to the class of passenger

  1st Class 62.96%

  2nd Class 47.28%

  3rd Class 24.24%

  - We can conclude that passengers of **higher classes** had **greater survival** rates.

- Determining if the survival rate is associated to the gender

  Female 74.2%

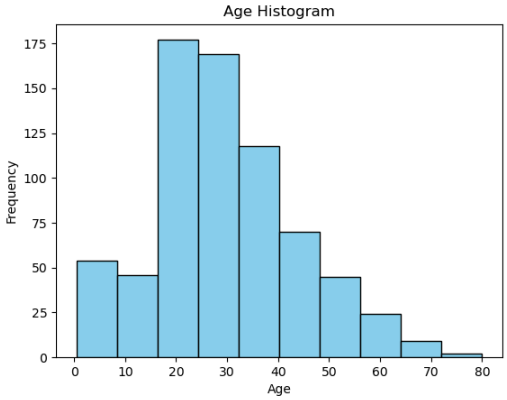
  Male 18.89%

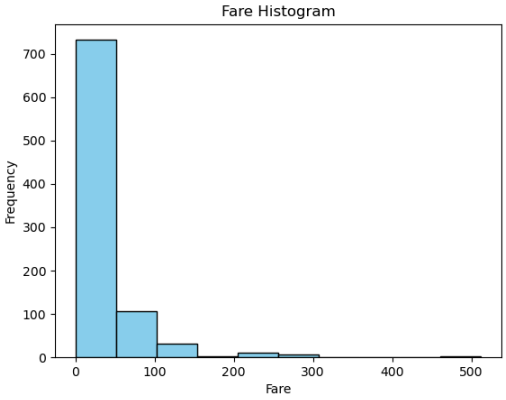
  - We can conclude that **females** had a much **greater survival** rate as compared to males.

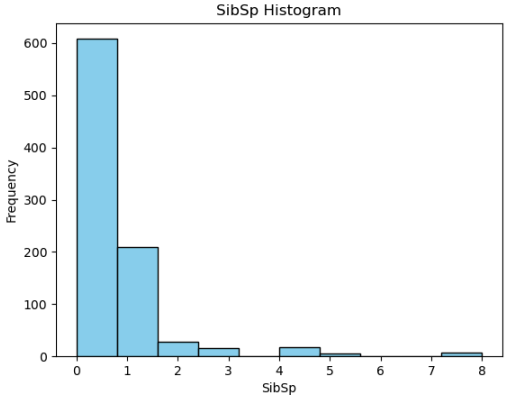
- Determining if the survival rate is associated to the age

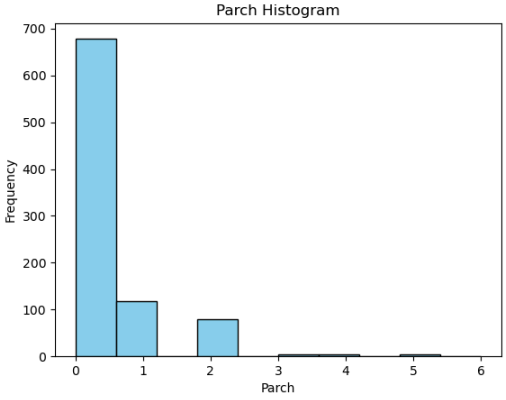
  - Age and Survival rate appear to have a correlation coefficient of *-0.25408475* implying there is only a **weak negative correlation** between the two variables. Hence, we cannot conclude that the survival rate is associated with age.
  - Similarly, the boxplot of the two variables show that the interquartile range of those who **survived**, leans slightly towards passengers of a **younger age**.

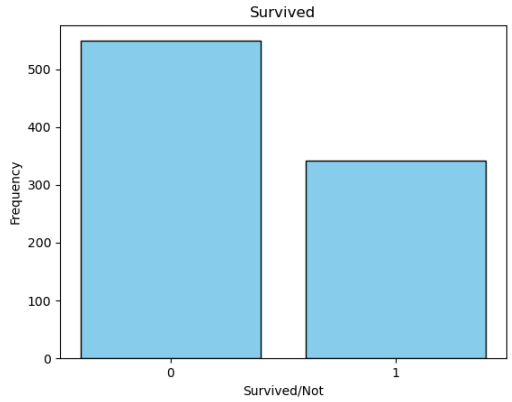
Figures

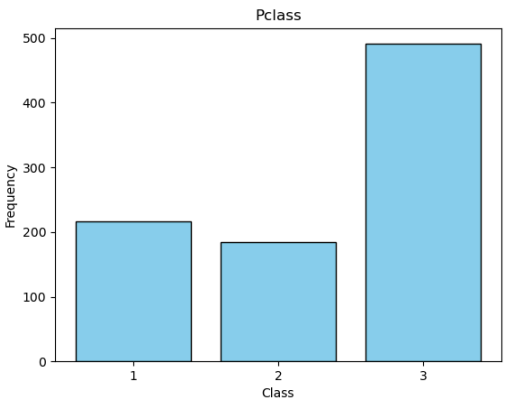

Age Histogram
Frequency
Age
0
25
50
75
100
125
150
175
0
10
20
30
40
50
60
70
80

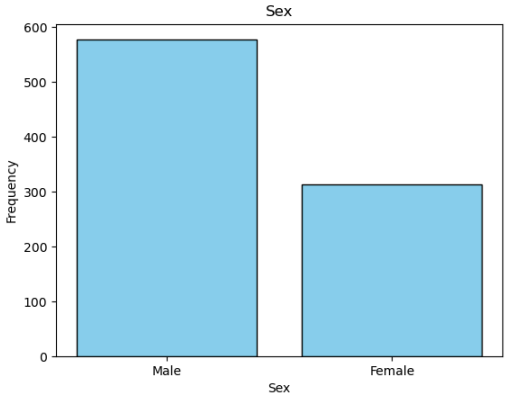

Fare Histogram
Frequency
Fare
0
100
200
300
400
500
600
700
0
100
200
300
400
500

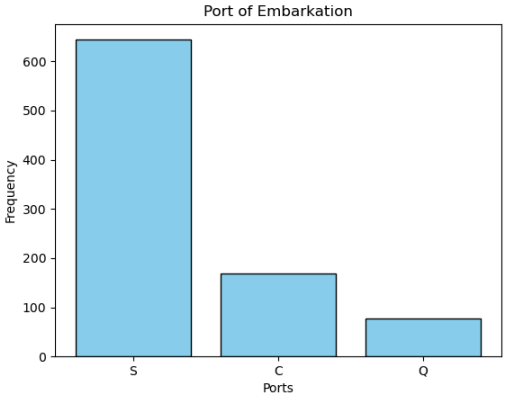

SibSp Histogram
Frequency
SibSp
0
100
200
300
400
500
600
0
1
2
3
4
5
6
7
8

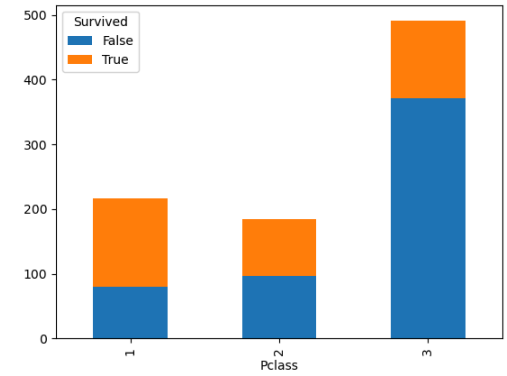

Parch Histogram
Frequency
Parch
0
100
200
300
400
500
600
700
0
1
2
3
4
5
6

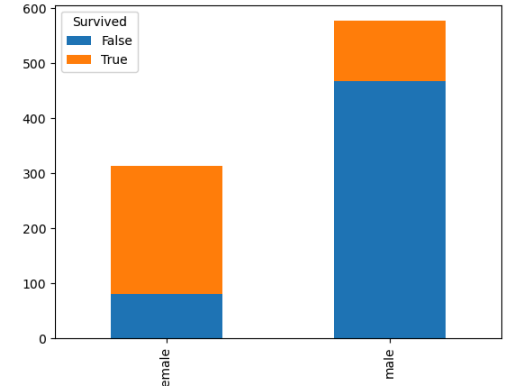

Survived
Frequency
Survived/Not
0
100
200
300
400
500
0
1

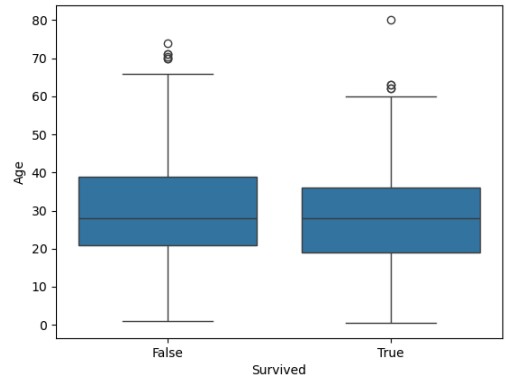

Pclass
Frequency
Class
0
100
200
300
400
500
1
2
3


Sex
Frequency
Male
Female
Sex


Port of Embarkation
Frequency
S
C
Q
Ports


Survived
False
True
Pclass


Survived
False
True
male


Age
False
True
Survived