***Report on Titanic Dataset: EDA***

**1) Introduction:**

The Dataset contains information about passengers who were onboard the Titanic.

This report delves into EDA of the Titanic dataset seeking insights into passenger traits and the factors influencing survival rates.

**2) Data Overview**

The dataset consists of the following columns:

* PassengerId: Unique identifier for each passenger
* Survived: Survival status (0 = No, 1 = Yes)
* Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
* Name: Passenger's name
* Sex: Passenger's gender
* Age: Passenger's age
* SibSp: Number of siblings/spouses aboard
* Parch: Number of parents/children aboard
* Ticket: Ticket number
* Fare: Passenger fare
* Cabin: Cabin number
* Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

**3) Data Cleaning and Preprocessing**

Before conducting analysis, the following preprocessing steps were performed:

* Handling missing values

Total number of missing values in ‘Age’ column were 177.

For females whose ages were missing, their titles were used to estimate their ages. For those with Miss in their titles, the average age was calculated and used that value to estimate the age for females with missing ages. The same process was followed for females with the title "Mrs.

Calculated the average age for males in each class and used those averages to fill the ages for males with missing ages in each class.

* Removing irrelevant columns

The column containing the information about the Cabin number was removed.

* Data type conversion

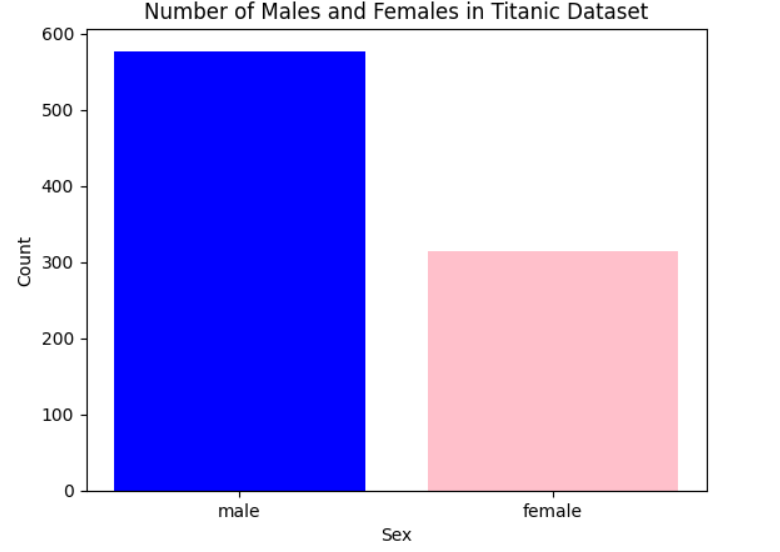
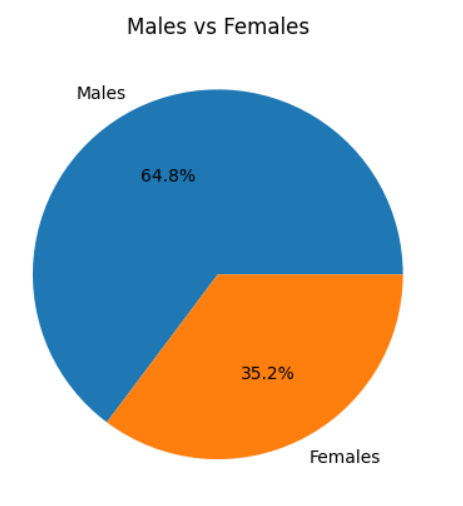
(Will encounter Data converasion in the later stages of this report)

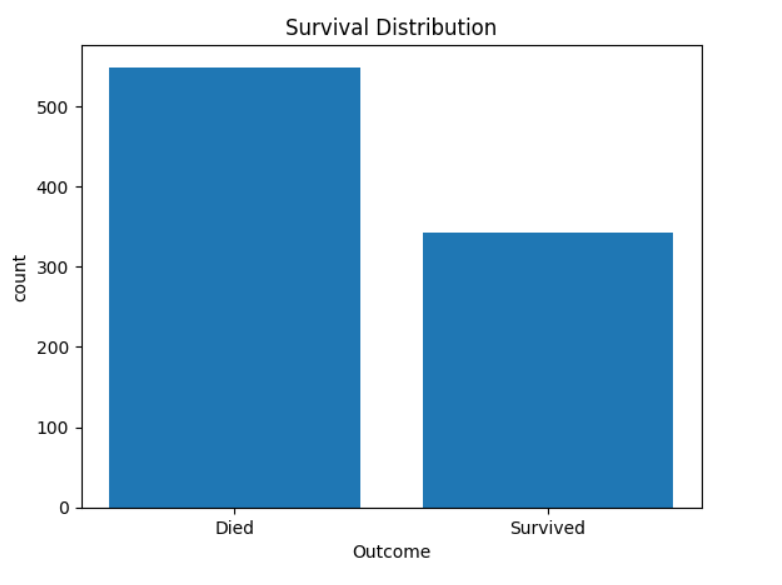
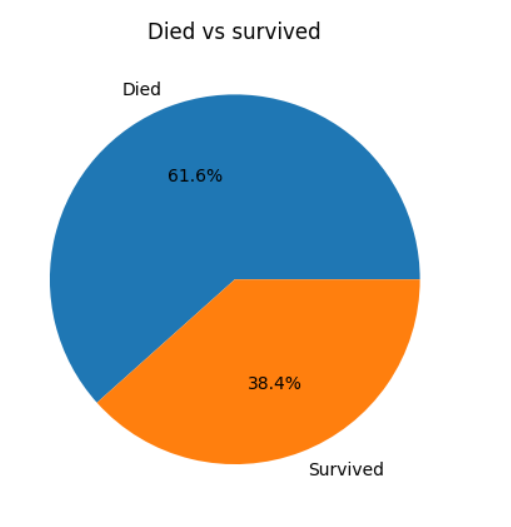
**4) Exploratory Data Analysis**

The oldest person on the ship was 80 years old.

The youngest person on the ship was not even half a year old.

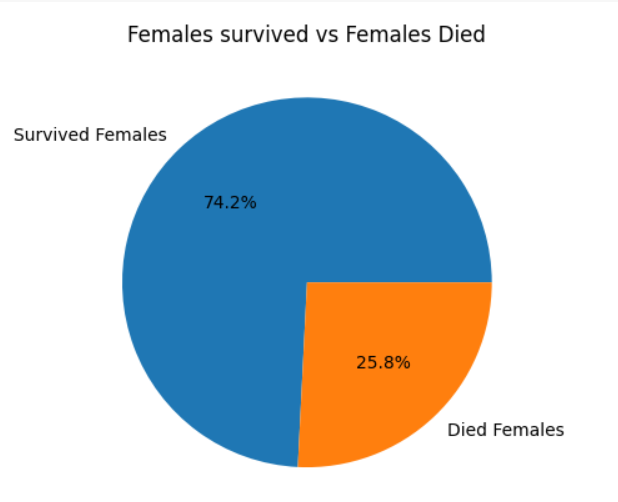
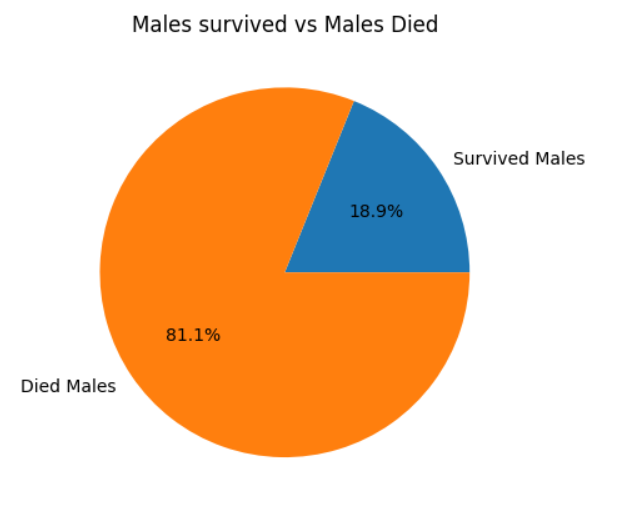
On the ship, there were a total of 891 passengers. Among them, 577 were male and 314 were female. This indicates that males comprised approximately 64.8% of the passengers, while females made up about 35.2%.

   
Out of the total 891 passengers onboard, 342 survived while 549 did not. This means the survival rate was approximately 38.4% and the mortality rate was around 61.6%.

Out of the total passengers who survived, 233 were females and 109 were males. This indicates that approximately 68.1% of the survivors were females, while about 31.9% were males.

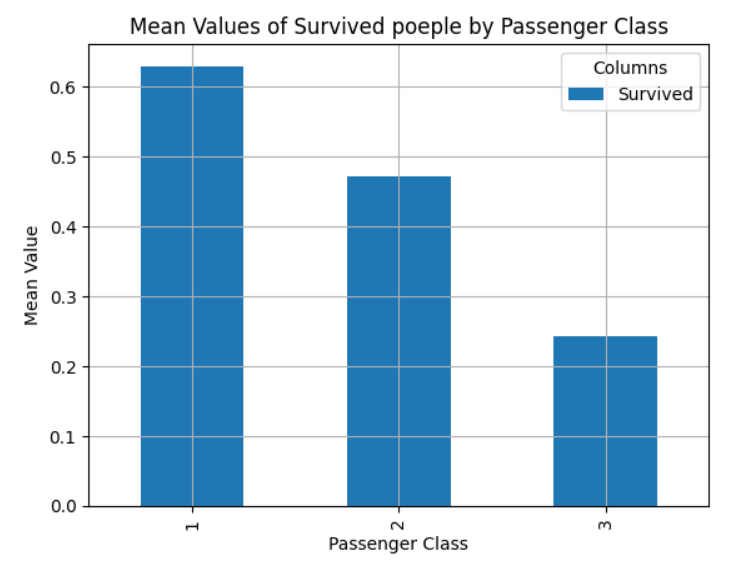
The number of men on the ship is lot more than the number of women. Still the number of women saved is almost twice the number of males saved. The survival rates for a**women on the ship is around 75% while that for men in around 18-19%.**

** **

The mean survival rate for passengers in each ticket class (Pclass) is as follows:

* For passengers in first class (Pclass 1), the mean survival rate is approximately 63%.
* For passengers in second class (Pclass 2), the mean survival rate is approximately 47%.
* For passengers in third class (Pclass 3), the mean survival rate is approximately 24%.

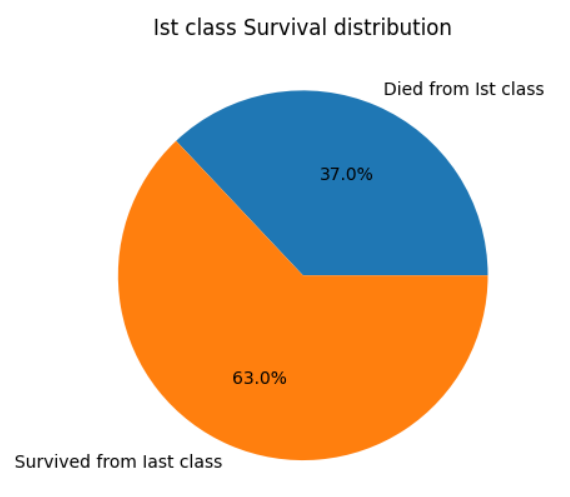
This suggests that passengers in higher classes had higher chances of survival compared to those in lower classes.

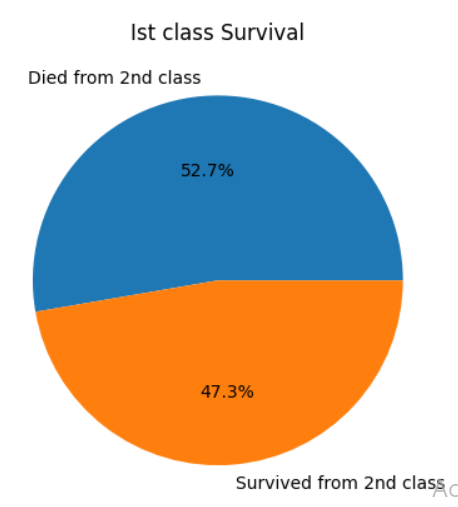


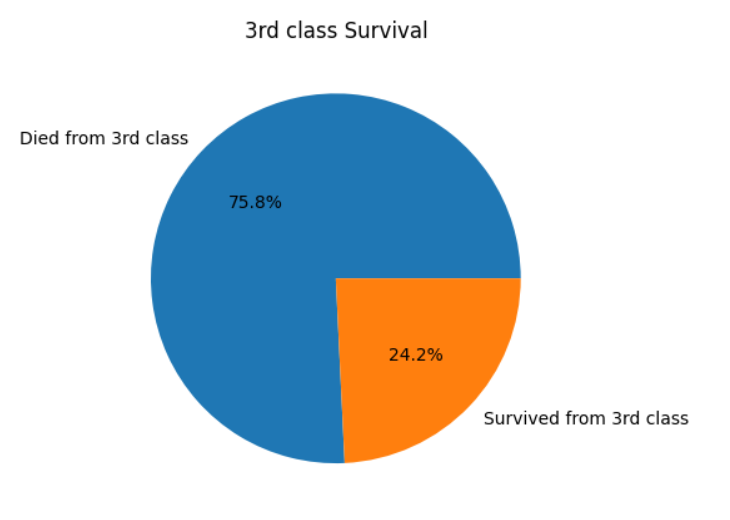
Number of people survived in class 1: 136

Number of people survived in class 2: 87

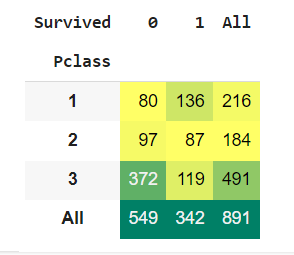
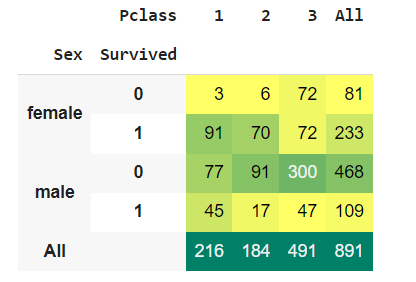
Number of people survived in class 3: 119







**Top of Form**

****

The table provides a breakdown of survival outcomes based on passenger gender and ticket class on the Titanic.

For females in the first class, 91 survived while 3 did not. In the second class, 70 females survived and 6 did not. Among females in the third class, 72 survived, and 72 did not. In total, out of 314 females onboard, 233 survived, and 81 did not.

For males in the first class, 45 survived, and 77 did not. In the second class, 17 males survived, and 91 did not. Among males in the third class, 47 survived, and 300 did not. In total, out of 577 males onboard, 109 survived, and 468 did not.

Overall, among the 891 passengers onboard, 342 survived, and 549 did not.

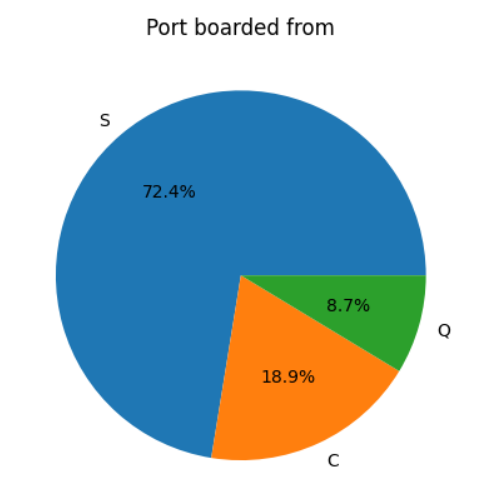
To put it in perspective:

* Approximately 68.1% of the survivors were females, while about 31.9% were males.
* The majority of survivors were from the first-class (216 out of 342 survivors), while the majority of casualties were from the third-class (491 out of 549 casualties).

The count of passengers embarking from different areas is as follows:

* 644 passengers embarked from Southampton (S).
* 168 passengers embarked from Cherbourg (C).
* 77 passengers embarked from Queenstown (Q).

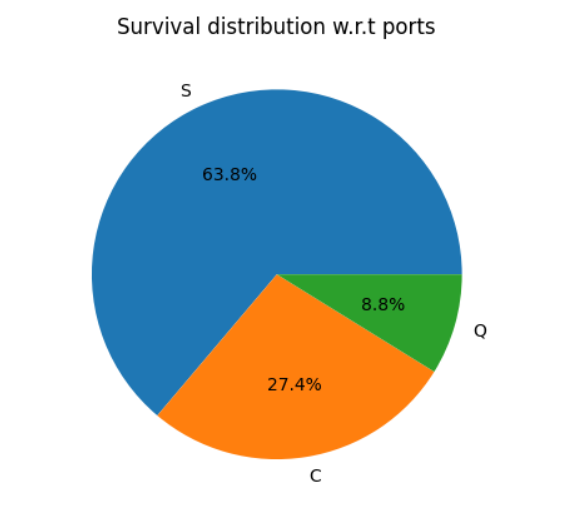
This indicates that Southampton had the highest number of embarked passengers, followed by Cherbourg and Queenstown.

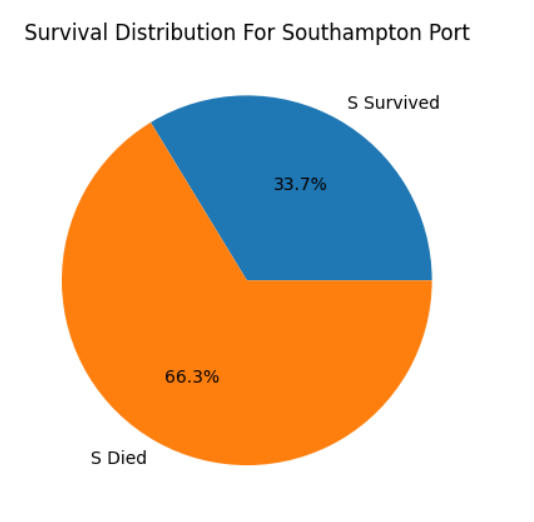


The number of passengers who embarked from Southampton (S) and survived is 217.

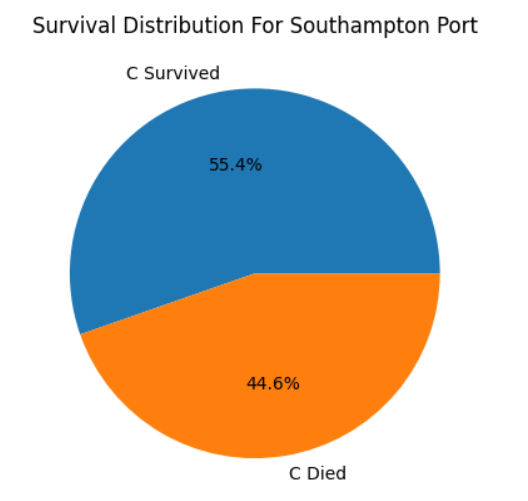
The number of passengers who embarked from Cherbourg(C) and survived is 93.

The number of passengers who embarked from Queenstown (Q) and survived is 30.





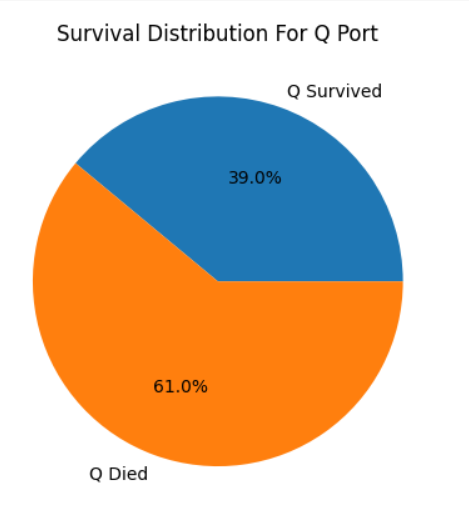
Out of 644 people who embarked from S port, only 33.7% people survived.

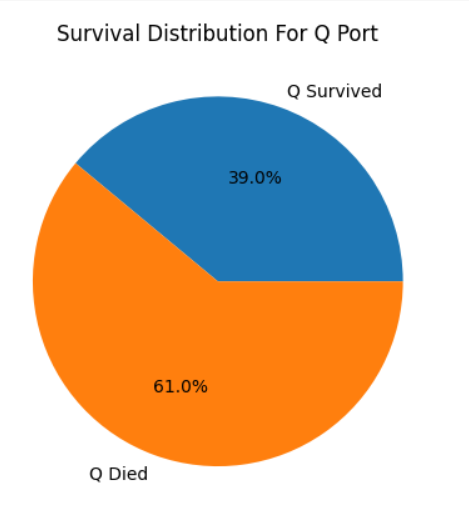


Out of 168 people who embarked from C port, only 55.4% people survived.

Top of Form

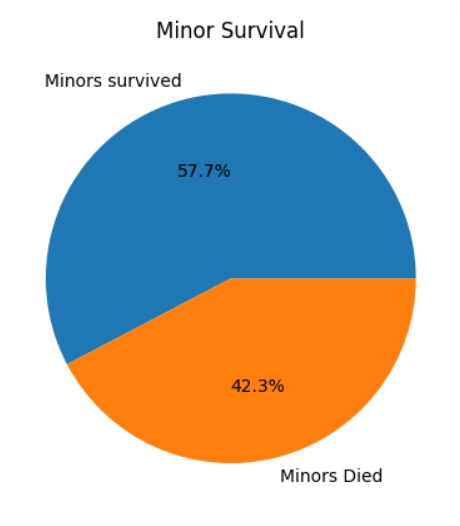
Top of Form

Top of Form

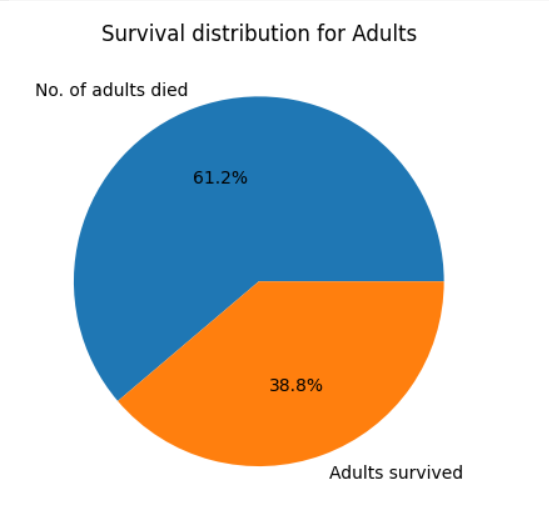
****

Out of 77 people who embarked from Q port, only 39% people survived.

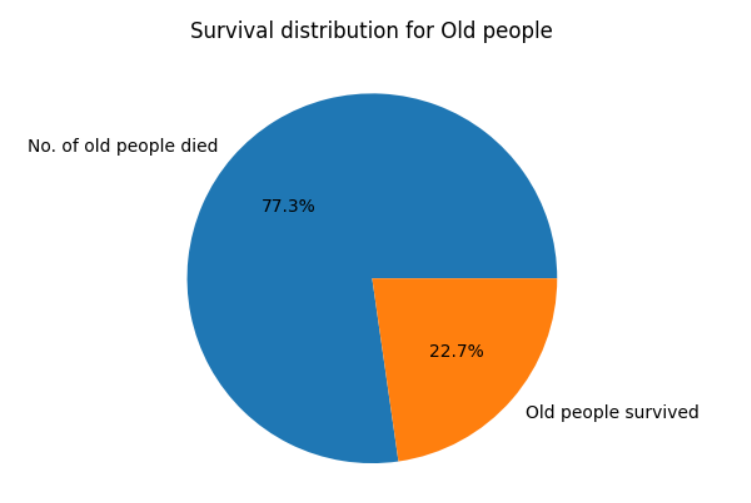
There were 78 Minors on the ship. Out of which only 45 survived.



There were 726 Adults on the ship, out of which only 265 survived.

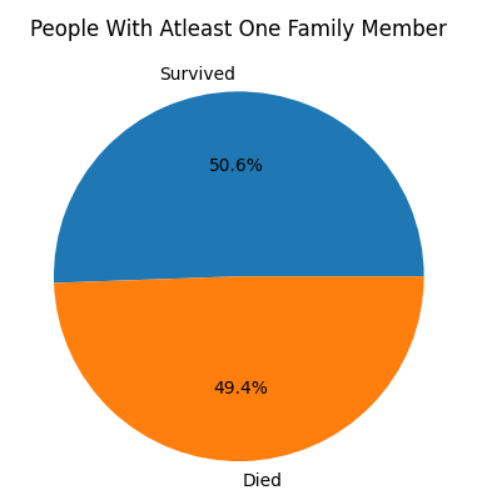


There were 22 Senior Citizens on the ship, out of which only 5 survived.



Number of people Who were on the ship with at least one family member: 354

Out of which only 175 survived.



**Correlation Analysis:**

Correlation analysis is a statistical technique used to measure the strength and direction of the linear relationship between two variables. Correlation coefficients range from -1 to 1, where:

* 1 indicates a perfect positive correlation (as one variable increases, the other also increases).
* -1 indicates a perfect negative correlation (as one variable increases, the other decreases).
* 0 indicates no correlation between the variables.
* A correlation heatmap provides a quick and intuitive way to identify patterns and relationships between numerical variables in a dataset. Variables with high positive correlation are represented by lighter colors, while variables with high negative correlation are represented by darker colors. Variables with low or no correlation are represented by neutral colors.
* It's important to note that correlation does not imply causation. While a strong correlation between two variables indicates a relationship, it does not necessarily mean that changes in one variable cause changes in the other. Correlation analysis should be used in conjunction with other analytical techniques and domain knowledge to draw meaningful conclusions about the data.

The correlation coefficient between "Survived" and "Gender" is 0.5434, indicating a moderate positive correlation between these two variables.

**Conclusion:**

* The majority of passengers were in the third class.
* The dataset contains more male passengers than female passengers.
* Passengers in the first class had a higher survival rate compared to those in the second and third classes.
* Children had higher survival rates compared to other age groups.
* Females exhibited a higher survival rate compared to males.
* Top of Form