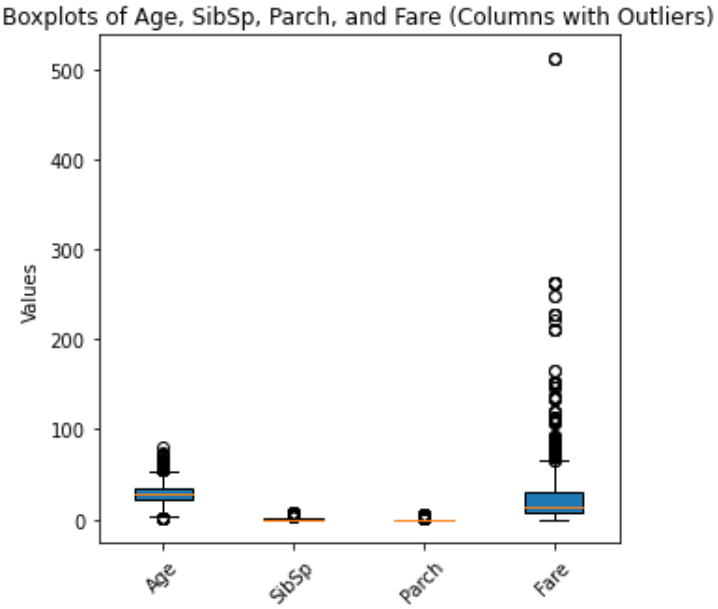
Titanic Dataset EDA Results

For the EDA on Titanic Dataset we did few visualizations. Such as:

1. **Outliers Detection:**
   Boxplots for the numeric variables containing outliers:



Interpretation:

Here for Age:

1. The median is around 25-40 range.
2. The dots below and above the whiskers mean that there are outliers below 0 and above 60. So, we can say the passengers that were boarded - some of them were with very young age and some of them were very old age.
3. The distribution is moderate but due to outliers it's right-skewed.

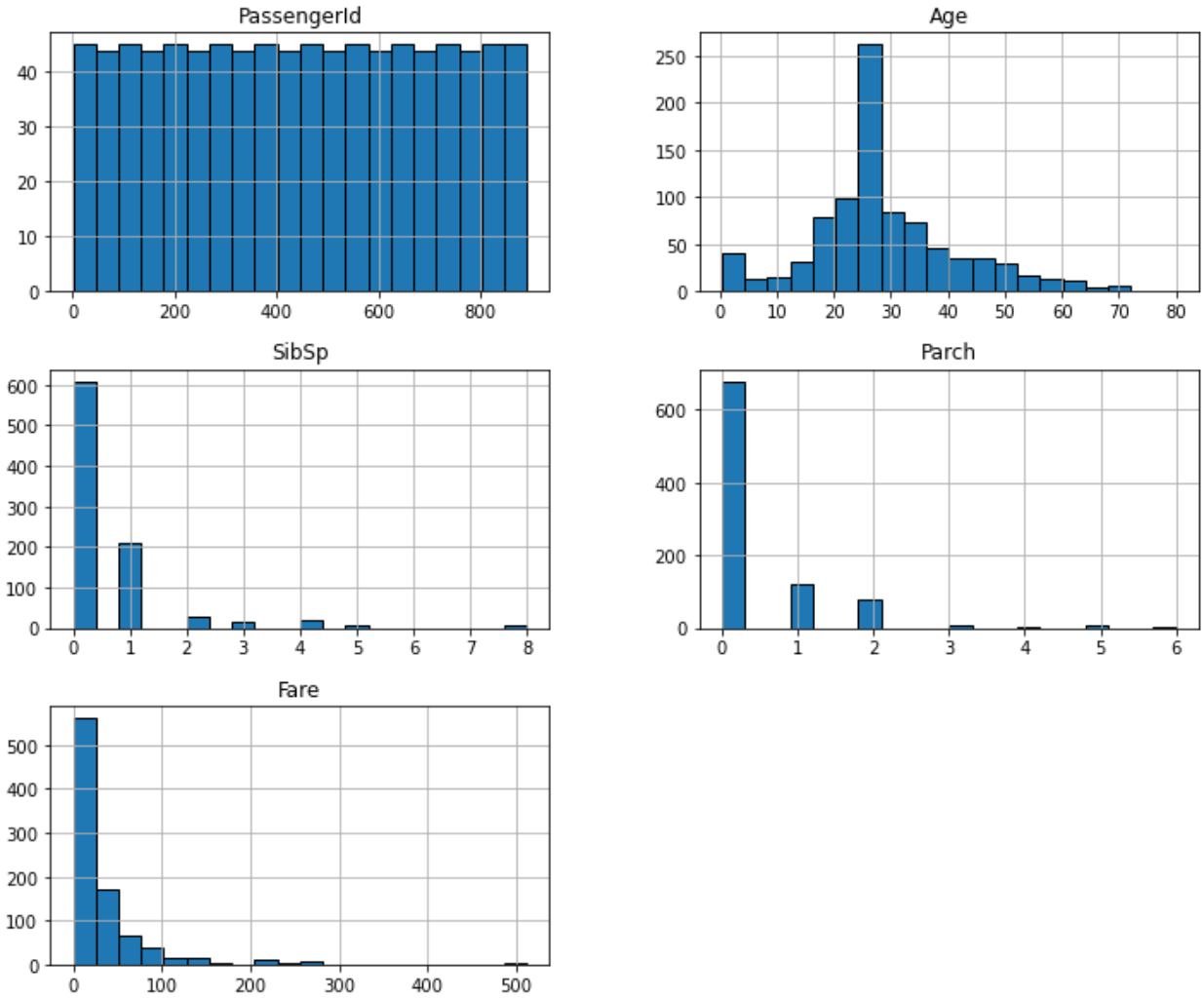
Here for SibSp and Parch:

1. Here values are mostly around 0-1 range.
2. So, we can say most passengers have no or maybe few siblings and children as well.
3. There are outliers showing few families with higher numbers of siblings, children on board.

Here for Fare:

1. The median is around 15-25 range.
2. There are outliers showing that some values are in extreme higher range.
3. The distribution is rightly skewed due to the presence of outliers. 4. So, it means most of the passengers are travelling with lower fares but few passengers are travelling with 1st class ticket which is showing extreme outliers.

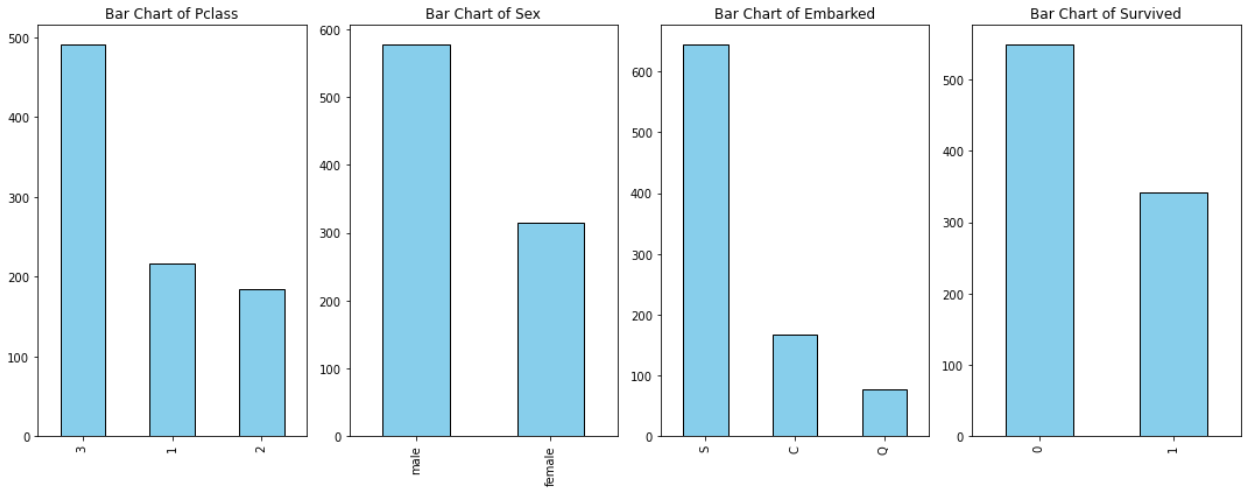
2. Univariate Analysis:
Histograms for the Numerical variables:



Interpretation:

- For Age histogram, Here's the data is rightly skewed. The most passengers are from the age of 10 - 40 means most of the passengers were from children and young adults group. There's also highest peak at the age of 25 -30 indicating the above statement also. The count decreases when the age increases.
- For SibSp and Parch histogram, the data is showing exponential distribution. Most passengers on board had fewer family members as most of the values are in 0 - 1 range. Though few passengers were travelled with higher family members which is evident by the presence of outliers.
- For Fare histogram, the distribution of the data is bimodal as there are two distinct peaks around 0-1 range. Most of the passengers were in 3rd class while boarding and few also paid extreme for the luxurious first class is visible through the outliers.

Bar charts for Categorical variables:

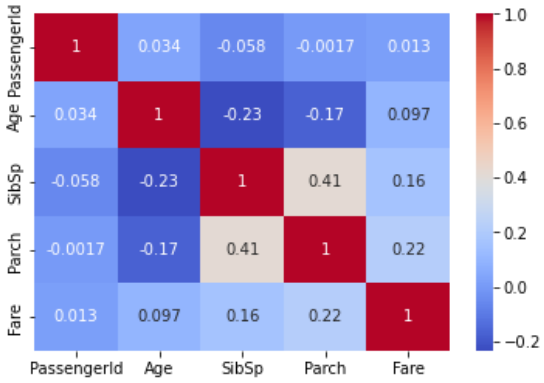


Interpretation:

- More passengers were in 3rd class on board.
- Male passengers outnumbers the female passengers by 60%.Remaining 40% were female passengers.
- Among the three ports, most passengers boarded at Southampton, with smaller groups embarking at Cherbourg and Queenstown.
- The non-survivors number is greater than the survivors for the passengers who boarded Titanic.

3. **Correlation Matrix :**

Between all numerical columns:



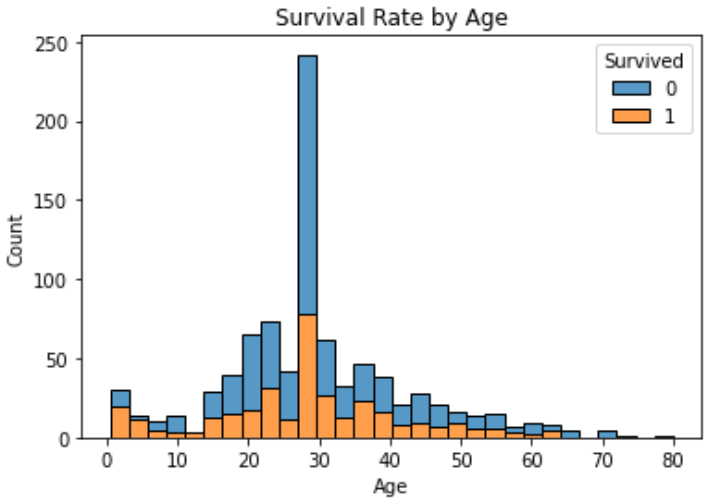
Interpretation:

- Age shows weak positive correlations with PassengerId (0.034) and Fare (0.097), and weak negative correlations with SibSp (-0.23) and Parch (-0.17).
- SibSp has a weak negative correlation with PassengerId (-0.058) and Age (-0.23), a moderate positive correlation with Parch (0.41), and a weak positive correlation with Fare (0.16).

- Parch has very weak negative correlations with PassengerId (-0.0017) and Age (-0.17), a moderate positive correlation with SibSp (0.41), and a weak positive correlation with Fare (0.22).
- Fare shows very weak positive correlations with PassengerId (0.013), Age (0.097), SibSp (0.16), and Parch (0.22)

4. Hypothesis testing:

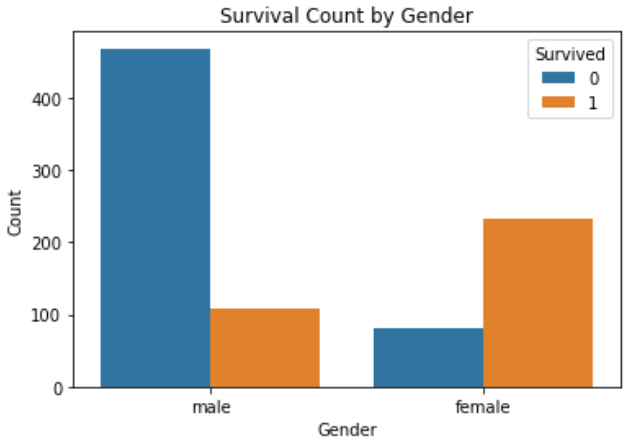
- **Survival Rate for Age by stacked histogram**



Interpretation:

- A significant number of passengers who did not survive were around the age of 30.
- There were generally more non-survivors than survivors across most age groups.
- The age distribution is spread out, with passengers of all ages affected by the tragedy.

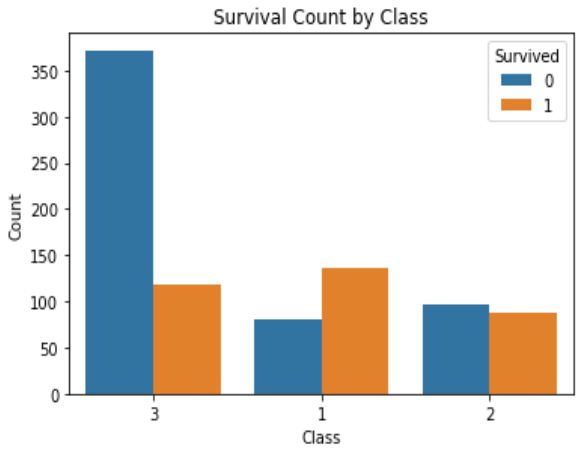
- **Survival Rate by Gender by count plot**



Interpretation:

- There are significantly more non-survivors than survivors among male passengers.
- There are more survivors than non-survivors among female passengers.
- This chart highlights that a larger proportion of female passengers survived compared to male passengers.

- **Survival Rate by Passenger class by count plot**



Interpretation:

- There are more survivors than non-survivors in 1st class. So the survival rate is higher here.
- The survival rate between survivors and non-survivors are balanced relatively in 2nd class. Showing the number of survivors with a slight higher number.
- The survival rate of 3rd class passengers are significantly higher for non-survivors. So the survival rate is lower for this class.