**Titanic Dataset Exploratory Analysis Summary**

This project aimed to determine if the survival of the passengers from the historic Titanic is associated with their class, sex, or age. The data used in the analyses were based on Kaggle's Titanic dataset. The following table below shows the index of each variable in the Python dataframe being created from the dataset, variable name, number of null values per variable, and variable type.

```
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
```

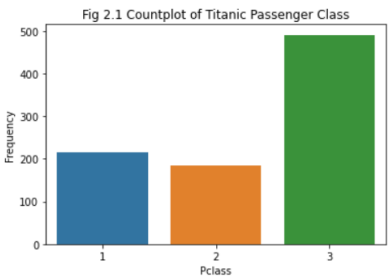
From the above table, it can be noted from the data that:

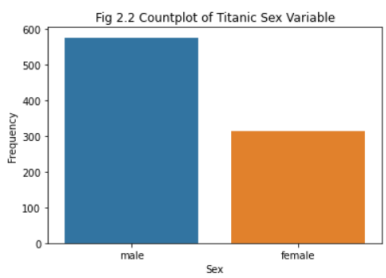
- There are 891 observations in total.
- Variables *Age, Cabin*, and *Embarked* have 177, 687, and 2 null values, respectively.
- There are a total of 12 variables: 7 of which are in number form while 5 are strings. However, although variables *PassengerId*, *Survived*, and *Pclass* are in number form, these are considered categorical variables based on what they represent. With this, there are 4 numerical and 8 categorical variables in the data:
    - **Numerical:** *Age, SibSp, Parch, Fare*
    - **Categorical:** *PassengerId, Survived, Pclass, Name, Sex, Ticket, Cabin, Embarked*

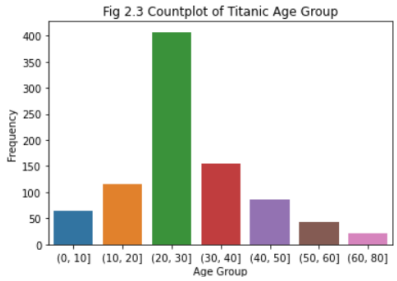
Summary statistics of the numerical variables are shown below:

| | Age | SibSp | Parch | Fare |
|---|---|---|---|---|
| **count** | 714.000000 | 891.000000 | 891.000000 | 891.000000 |
| **mean** | 29.699118 | 0.523008 | 0.381594 | 32.204208 |
| **std** | 14.526497 | 1.102743 | 0.806057 | 49.693429 |
| **min** | 0.420000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 20.125000 | 0.000000 | 0.000000 | 7.910400 |
| **50%** | 28.000000 | 0.000000 | 0.000000 | 14.454200 |
| **75%** | 38.000000 | 1.000000 | 0.000000 | 31.000000 |
| **max** | 80.000000 | 8.000000 | 6.000000 | 512.329200 |

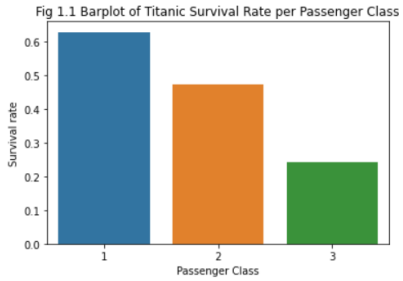
This project, however, only focused on the variables *Survival*, *Pclass*, *Sex*, and *Age* as it aimed to know if survival rate is associated with any of the last three variables mentioned. Figures below are the count plots for variables *Pclass*, *Sex*, and *Age* (grouped accordingly) which show the distribution of the values of the variables:

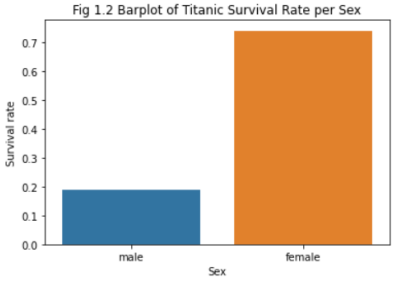


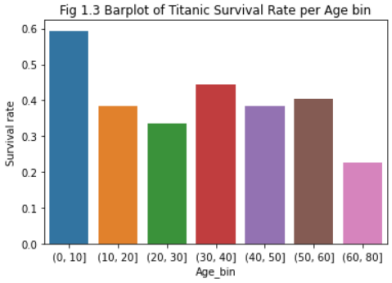




Moreover, the bar charts below show the survival rate for each value of the variables *Pclass*, *Sex*, and *Age*, respectively:




To check if survival is associated with passenger class, sex, or age, three different Chi-Square tests were conducted. Note that from numerical type, age data is grouped accordingly which creates a new age variable that is of categorical type.

The null hypotheses for the three tests are as follows:

Test 1: The passenger's survival has no association with passenger class.
Test 2: The passenger's survival has no association with passenger's gender.
Test 3: The passenger's survival has no association with passenger's age.

Below table shows the results of the Chi-Square tests:

| | Chi-Square Statistics | p-value | Degrees of Freedom |
|---|---|---|---|
| **Test 1 (vs Pclass)** | 102.888989 | 4.549252e-23 | 2 |
| **Test 2 (vs Sex)** | 260.717020 | 1.197357e-58 | 1 |
| **Test 3 (vs Age Group)** | 20.995233 | 1.838238e-03 | 6 |

At significance level of 0.05, we reject the null hypothesis of each of the three tests as the resulting p-values from all tests are less than 0.05. This means that there is a statistically significant association between the survival rate of Titanic passengers and the three factors: passenger class, sex, and age.