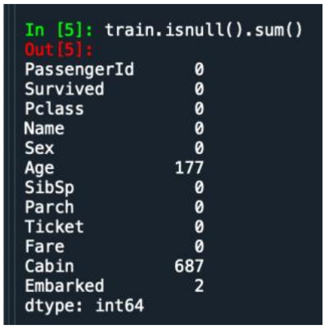
# Titanic Data Set Report

## Importing the Data:

First, I started with all the imports that I thought I would need during this assignment. I added matplotlib and seaborn because I thought it would be beneficial when doing analysis with certain graphs and charts. Then I loaded in the training and test dataset in, and looked at the first 5 rows of data, along with the column values. Furthermore, I looked into the different data types of the variables, which would help me later in deciding what variables I needed to change or manipulate.

## Data Cleaning:

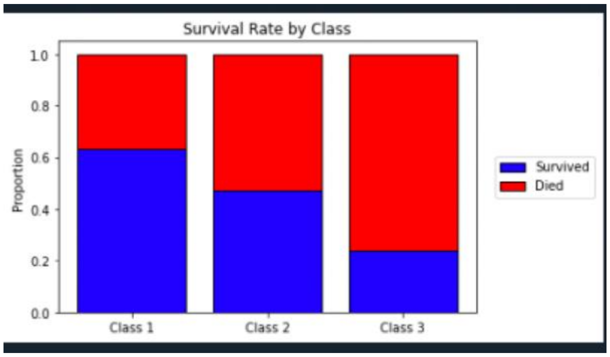
Next, I moved onto the data cleaning process. I looked at which variables had null values. The training data set had null values within the age, cabin, and embarked variables. The testing data set had null values within the age, cabin, and fare variables. To fix this issue I filled in the age and fare gaps with medians and the embarked variable with the mode of the already present letters. I then changed the floats to integers to have normality across the grid. This meant changing age and fare into integers.

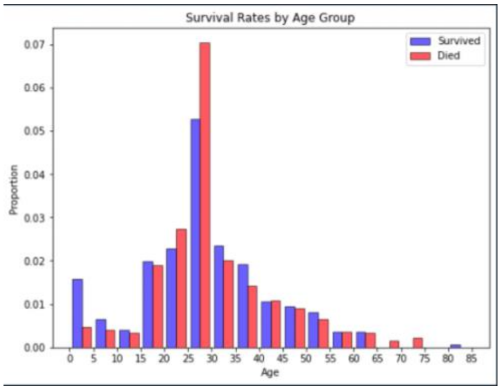


```
In [5]: train.isnull().sum()
Out[5]:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

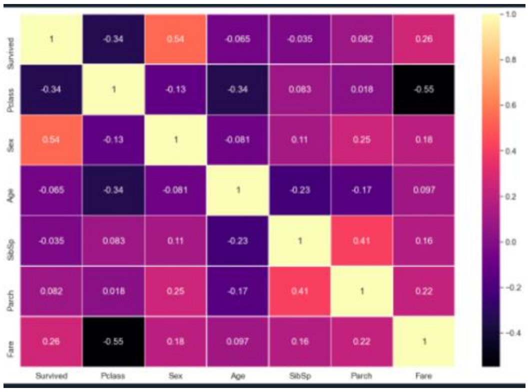
## Data Analyzing:

Moving onto some exploratory data analysis within the training data set. I first looked at the relationship between survivors and non survivors. It appears more people died than survived. Diving more into the survival rate, I investigated who was more likely to survive, men or women. It turned out that more women survived than men. Next, I created a graph that proved that the higher your class, the more likely you were to survive, meaning the wealthier you were, the more chance you had at surviving. The last bar chart I investigated was survival vs. age. It appears the younger you were, the more chance you had at surviving.





I ran a correlation matrix to see what variables were most correlated with surviving. Pclass, Age, and SibSp all had a negative correlation, while Sex, Parch, and Fare all had a positive correlation.

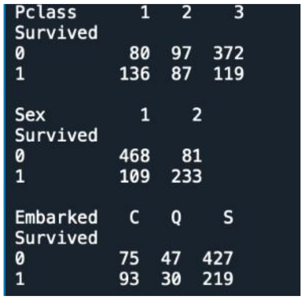


| | Survived | Pclass | Sex | Age | SibSp | Parch | Fare |
|---|---|---|---|---|---|---|---|
| Survived | 1 | -0.34 | 0.54 | -0.065 | -0.035 | 0.082 | 0.26 |
| Pclass | -0.34 | 1 | -0.13 | -0.34 | 0.083 | 0.018 | -0.55 |
| Sex | 0.54 | -0.13 | 1 | -0.081 | 0.11 | 0.25 | 0.18 |
| Age | -0.065 | -0.34 | -0.081 | 1 | -0.23 | -0.17 | 0.097 |
| SibSp | -0.035 | 0.083 | 0.11 | -0.23 | 1 | 0.41 | 0.16 |
| Parch | 0.082 | 0.018 | 0.25 | -0.17 | 0.41 | 1 | 0.22 |
| Fare | 0.26 | -0.55 | 0.18 | 0.097 | 0.16 | 0.22 | 1 |

This matrix helped to determine what variables were most important when determining survival, which was Pclass, Sex, and Age.

I did not really feel as though ticket number, cabin number, or fare really had any large impact on the predictions of whether someone survived. Although they may have some pull in the prediction, it is hard to justify without further analysis, especially with machine learning algorithms.

I figured it was best to create 6 pivot tables for all the variables I thought were important. When looking at class, even though there were significantly less people in first class than there were in third, the people in first class were still more likely to survive. Even though there were more men on board, more women survived. If you embarked from Cherbourg, you were more likely to survive than not, could possibly be a city or area with more wealthy people, young children, or women. I ran a pivot table on SibSp and Parch, but I did not really draw any conclusions from them both because nothing significant stood out between the variables and the survived.



```
Pclass      1   2    3
Survived
0          80  97  372
1         136  87  119

Sex         1    2
Survived
0         468   81
1         109  233

Embarked   C    Q    S
Survived
0         75   47  427
1         93   30  219
```

I then narrowed down my predictions by age, sex, and class. A child (under the age of 18) who is first class has a 91% survival rate, which were the best odds out of all the groups I calculated. Groups of people with the worst chance of surviving were men over the age of 18. Thirds class men had a 76% death rate and first-class men had a 64% death rate. On the other hand, if you were a first-class woman, only 12% did not survive.

**Conclusions:**

Overall, if you are a child or a woman, you had the greatest chance of survival. The saying woman and children first does seem fitting in this instance. I would like to investigate this data set more. I think that tree-based methods would be an interesting technique to use for this problem. Especially since there are so many factors that go into deciding if someone survived or not.