Project #2: Titanic - Who will survive?

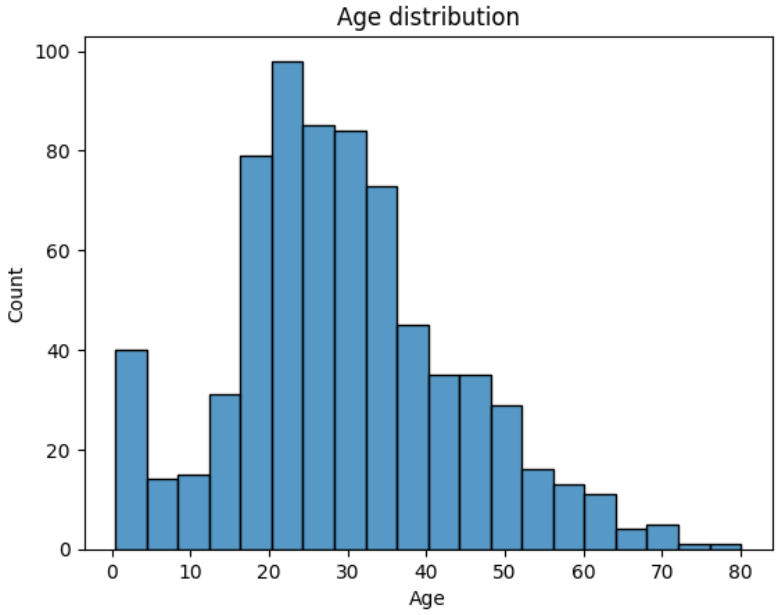
CSE 351 Jason Cheng and Jaden Wong

Language and Libraries:

- Python
- Pandas
- Seaborn
- Sklearn
- Scipy

Cleaning Data:

From the test.csv, we have a total of 891 data points for passengers. Next we checked the number of null values for the fields in the dataset. In the “Cabin” field there were 687 missing values, in the “Age” field there were 177 missing values, and in the “Embarked” field there were 2 missing values. Since about 77.28% of the data points are missing the “Cabin” field, we determined that it was not worth keeping the field for our exploratory data analysis and modeling since it makes up a large portion of our data set. From the “Embarked” field, there are only 2 values missing and we believe that embark location is unlikely to make a major difference if a passenger survives so we decided to drop the rows where the “Embarked” field is missing. Finally for “Age”, we deemed that it was important enough of a feature to include because the age of a person could be important for someone’s survival such as the elderly likely being more frail or prone to drowning. Due to 20% of the age data being missing and an important feature we decided to impute the data. We decided to use either mean or median imputation because we wanted to have a simple model. The values for the mean and median are 29.7 and 28 respectively, very similar. From looking at the age distribution, the data does not have any outliers and decided to fill the missing “Age” values with the mean



Other fields that were cleaned were “Sex”, changing it to binary where 0 = male and 1 = female to make it easier for EDA. Other columns that were dropped where “Name” and “Ticket” since they mostly likely wouldn’t affect a passengers survival rate because they are mostly identifiers

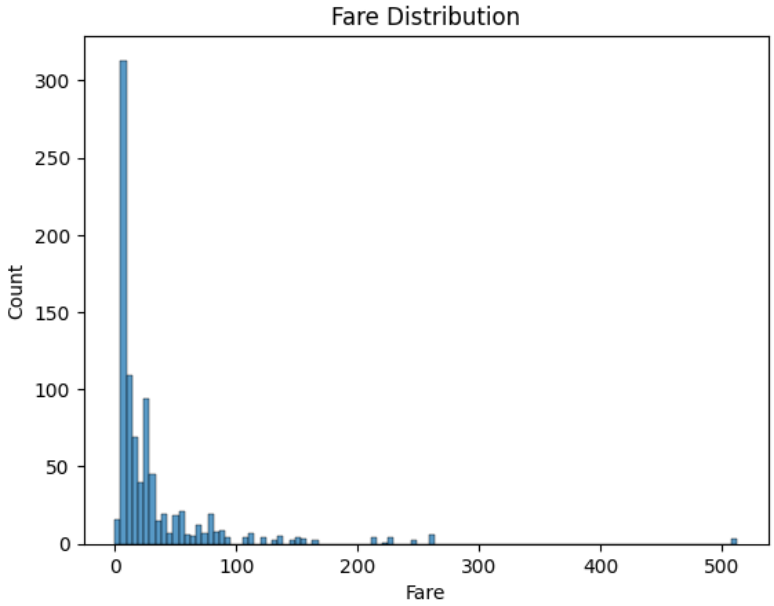
EDA:

First thing we wanted to look at was to see if there was any difference between the different Pclasses. By conducting a one way ANOVA F-Test, we can see if the means of certain attributes of different groups are the same or different. After sorting the data into groups based off their “Pclass”, we conducted the test on the numerical attributes fare,SibSp, Age, and Parch with the following result below:

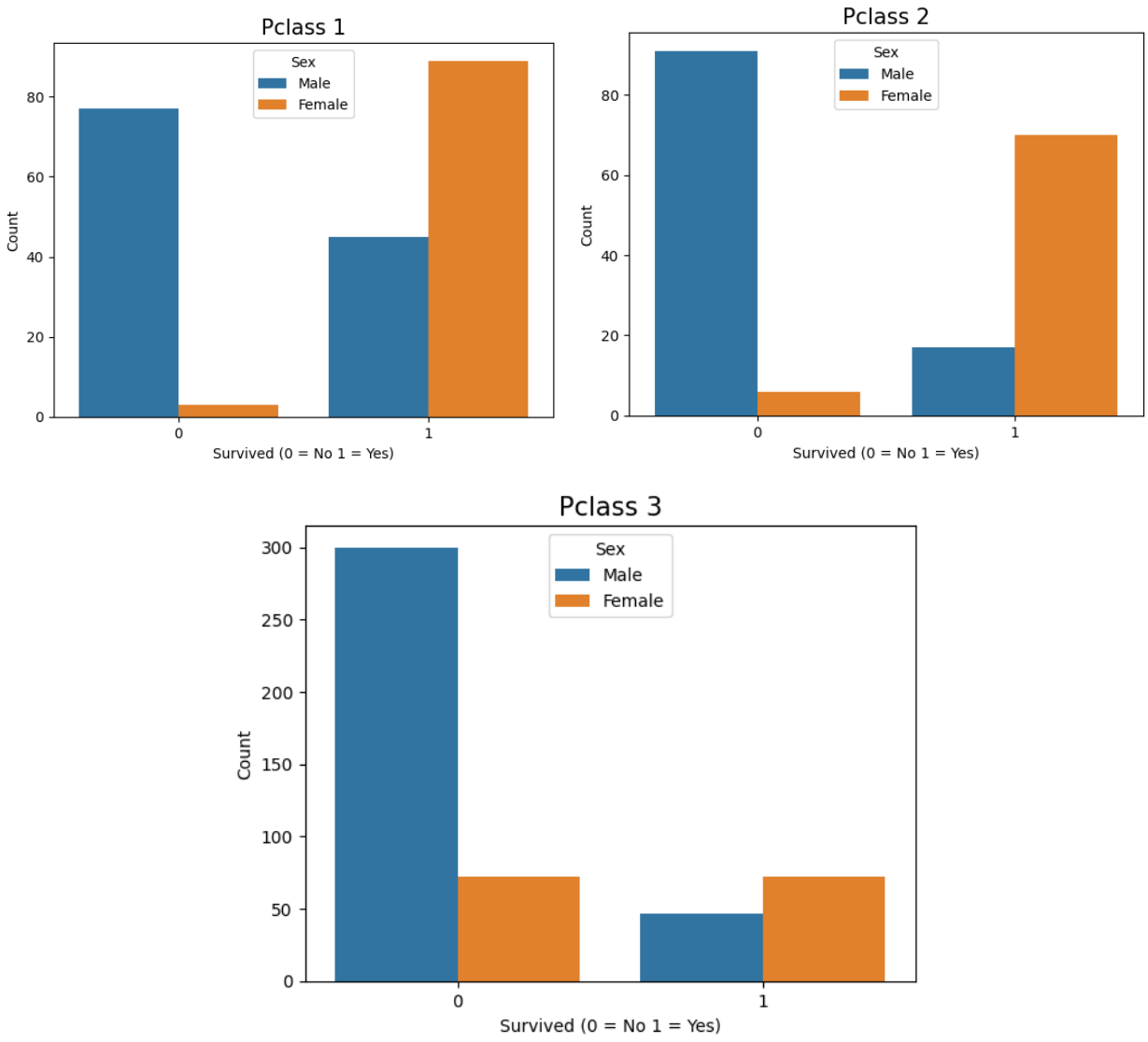
```
Fare: F_onewayResult(statistic=240.38829529293847, pvalue=3.9731247008621683e-84)
SibSp: F_onewayResult(statistic=3.7553742290688574, pvalue=0.02376488529670891)
Age: F_onewayResult(statistic=55.08746430410622, pvalue=2.8214004828185865e-23)
Parch: F_onewayResult(statistic=0.1271498529782774, pvalue=0.8806177675750773)
```

At alpha level 0.01, the p-values for SibSp and Parch would get rejected and for Age and Fare, it would get accepted. This shows that the result is statistically significant for Age and Fare and that there is a difference between the Pclasses for Fare and Age

We can verify the difference between the Fares in each Pclass by observing the centers for each Pclass. Since the “Fare” distribution is skewed, it is best to use median. There we learned Pclass = 1 is the highest standing at 58.69, Pclass = 2 at 14.25, and Pclass = 8.05. This shows that Pclass 1 is overall paid more for their fare than Pclass 3.



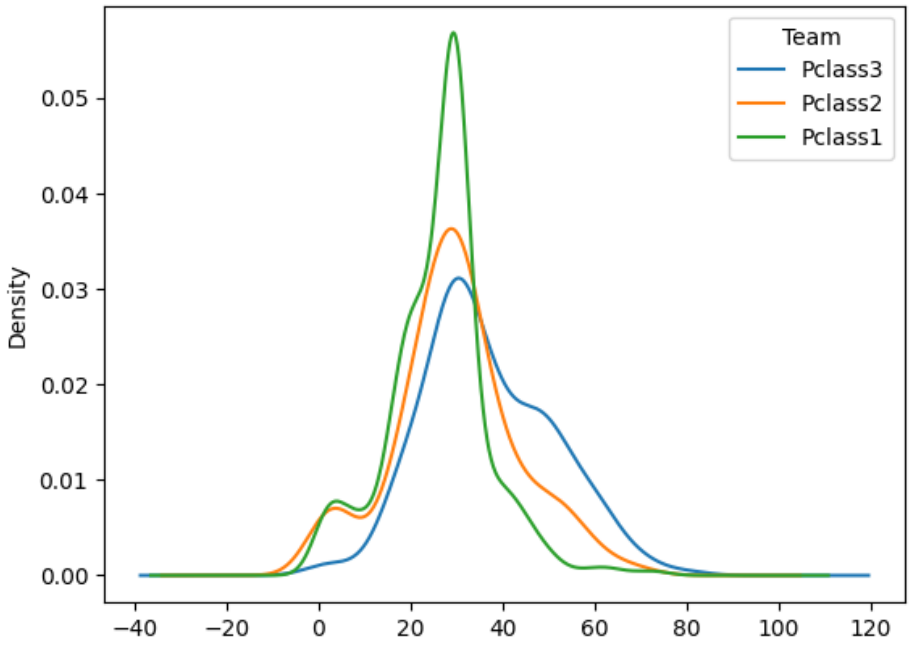
Next thing we wanted to check was how gender affected the survival rate of passengers by making a count plot between the Pclasses



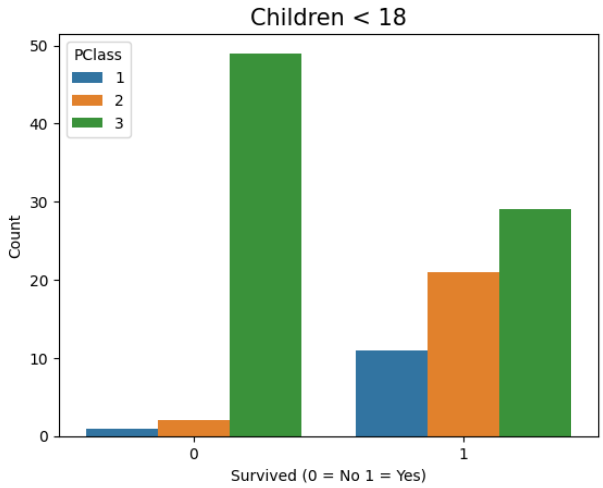
Some interesting things to note is that a greater portion of people from the upper classes such as Pclass 2 compared to Pclass 3. In the case of Pclass 1, there were more people that were able to survive than the passengers who didn't. This can indicate that people in higher classes are being prioritized for rescuing compared to lower classes. Another point of interest is for all classes, there were always more females who survived than males and more males than females that did not survive. So overall being female increases a passengers survival chance. Below are the respective survival rates for each gender in each Pclass:

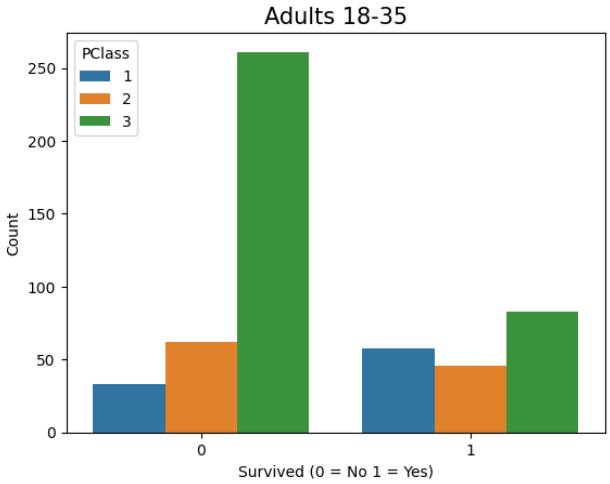
```
PClass 1 Women Survival: 96.73913043478261%
PClass 1 Men Survival :36.885245901639344%
PClass 2 Women Survival: 92.10526315789474%
PClass 2 Men Survival :15.74074074074074%
PClass 3 Women Survival: 50.0%
PClass 3 Men Survival :13.544668587896252%
```

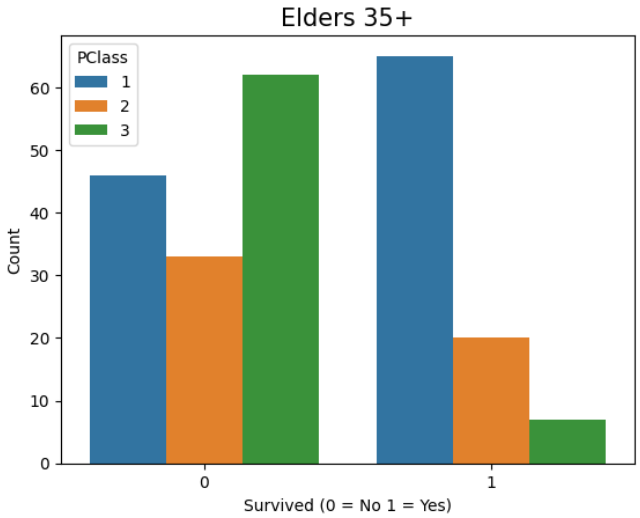
As found earlier, mean age between the Pclasses are different. We wanted to further analyze how age can affect survival groups. First we plotted the KDE graph based on a passengers age to determine the distribution of ages in our data.



From there we can determine that a significant amount of passengers are adults between the ages 20-40. Next we wanted to split the data from all passengers into age groups to see the survival rates differ between the age groups. We decided on three groups, (0,18),[18,-35), and 35+. The reason for this is the first group represents people not considered adults. From observing the percentiles of the ages, 35 years old represents the third quartile which we find to be an appropriate stopping point for a young adult with the final group representing everyone older than 35.





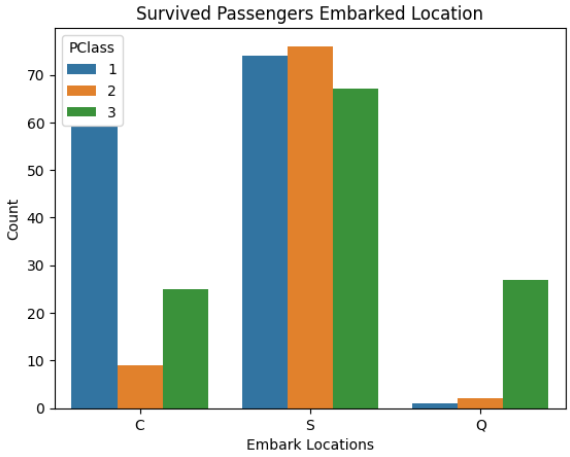


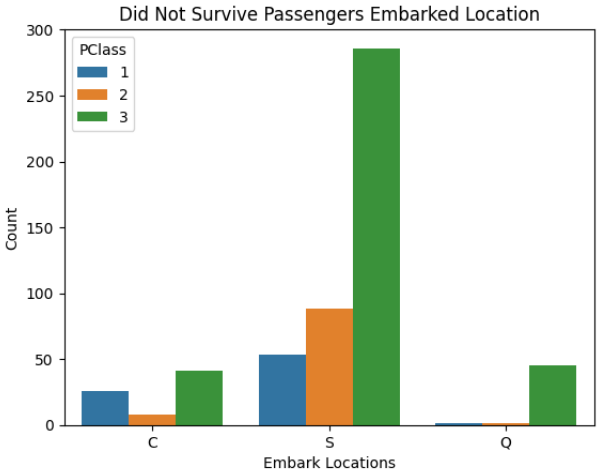
From the calculated percentages:

```
Children: 53.98230088495575 %
Adult: 34.43830570902394 %
Elder: 39.48497854077253 %
```

Children overall had the highest survival rate at about 54%. Elders had about a 5% higher chance of survival compared to adults. One thing to note is this count does not account for the Pclass of the passenger. From looking at the charts, elders had a significantly higher survival rate if they are Pclass 1 while children had the most in the other classes, once again showing Pclass seeming important for a passengers survival rate.

Finally, we wanted how the port where the passengers embarked from could play a role in a passengers survival rate. This was done by conducting another count plot, separating the passengers by their Pclass.





The embark locations key is: C = Cherbourg, Q = Queenstown, S = Southampton
We would expect that the location from where passengers boarded the Titanic wouldn't affect a passenger survival rate, however it is much more obvious that fro embarking from Southampton has overall a much higher survival rate with all the amount of passengers that survive be around the same for all Pclasses.

Modeling:

Since our goal is to predict if a passenger on the Titanic would survive, a logistic regression model would be fitting, since we are classifying a passenger ,given its data fields, if they survived, 1, or did not ,0. This works by mapping a value from a linear function onto a logistic function. The range from this logistic function is from [0,1] which we treat as probability with values above 0.5 as positive cases and below as negative cases. The library of choice to make these models were from Sklearn, in python. I splitted the data into the training and testing groups for our models with 80% of the original data being used for training and the remaining for verification.

1: To start off we wanted a more simple model to predict the passengers survival class and used the attributes of “Age” and “Fare”. Those variables were considered significant at alpha level = 0.01 because of the low p-value and are numerical data. This gave the following results for the confusion matrix and in the linear regression line used in the logistic regression:

```
Confusion Matrix
[[102  13]
 [ 42  21]]
Coefficients: [[-0.02060535  0.01551172]]
Intercept: [-0.3003101]
```

```
Accuracy 0.6910112359550562
Precision 0.6176470588235294
Recall 0.3333333333333333
F1 Score 0.4329896907216495
```

This shows that overall using only the fields of “Age” and “Fare” for a logistic regression model would be a poor choice. With a poor F1 score of 0.433, it is too low to be a reliable predictor for passenger survival. One thing the mode suffers from is having a lot of false negatives, where it predicts a passenger survives when they did not and reflects in the recall score of 0.33. On the other hand the model does well in predicting true negatives which shows that this model is too simple and needs to be more complicated to predict positive cases.

2: To improve the model we decided to include more categorical variables to increase the complexity of the model. The attributes that we added are “Sex” and “Pclass” since from our earlier analysis, there appears to be differences in the number of passengers who survive, particularly for females and Pclass 1. This gave us the following results:

```
Confusion Matrix
[[96 19]
 [17 46]]
Coefficients: [[-3.30970231e-02  1.36183381e-03 -1.00252616e+00  2.44342669e+00]]
Intercept: [1.81473253]
```

```
Accuracy 0.797752808988764
Precision 0.7076923076923077
Recall 0.7301587301587301
F1 Score 0.7187500000000001
```

Overall, the added complexity to this model helped it improve much more compared to model 1 with all the summary classifiers values being higher for every field. The F1 score of 0.71875 shows that this model performs decency well at predicting the survival rates of passengers. Most of the improvement comes with much less false negative predictions at the cost of a little true negative guesses, which is a worthwhile trade off for a better performing model in predicting positive cases.

3: For our third model, we went with a different approach to use a decision tree to predict if a passenger survives. Decision trees work by splitting the data into branches based on the criteria of the node, such as being male or female, and each node keeps on branching with its own criteria, until a decision is reached at the children nodes of the tree. This time, we went with mostly categorical variables, which a decision tree often leads itself to, which are, “Pclass”, “Sex”, and “Embarked”. These categories were all shown to be correlated with a passenger's survival. Also we included “Age” of a passenger by breaking up the age of a passenger into the same groups shown before of children [0,18), adults [18,35), and elders (35+). This gave use the following results:

```
Confusion Matrix
[[110   5]
 [ 23  40]]
```

```
Accuracy 0.8426966292134831
Precision 0.8888888888888888
Recall 0.6349206349206349
F1 Score 0.7407407407407407
```

*The decision tree can be viewed in the notebook
This model had a slightly higher F1 score than model 2 with a score of 0.7401 and still much better than model 1. The difference between model 2 and 3 in F1 score is not large enough to claim that this model will always be more performant than model 2. However, this model does a much better job at predicting true negatives and making less false positives. The performance on false negatives and true negatives is slightly worse than model 2. This is shown with every statistic improving while trading off recall.

Cross-validation:

Cross-validation is a technique to split the data into folds and train on all folds except for one of them, which is reserved for verification, and repeat this process for all folds such that each one gets to be used as a verification set. This can make sure the model is not too overfitting compared to when I split the 80% of the data for training from variance. I decided on fold lengths of 10 and computed the results for each model.

Model 1:

```
Average Accuracy 0.6591547497446373 STD: 0.04332839348130216
Average Precision 0.6439906584643427 STD: 0.13724597270007546
Average Recall 0.2352941176470588 STD: 0.08823529411764705
Average F1 score 0.3403840126775094 STD: 0.10544077403954101
```

Model 2:

```
Average Accuracy 0.7896450459652706 STD: 0.021398964314535
Average Precision 0.7371481650942698 STD: 0.02485488104750012
Average Recall 0.7029411764705883 STD: 0.08964559208310686
Average F1 score 0.7160921594985707 STD: 0.0441195740440703
```

Model 3:

```
Average Accuracy 0.8256511746680285 STD: 0.03523787807385179
Average Precision 0.8902189274026583 STD: 0.05047306451573542
Average Recall 0.6205882352941178 STD: 0.08049371872590592
Average F1 score 0.7290231738587578 STD: 0.06325461721020288
```

For each of my models, the F1 scores are all slightly lower for models 2 and 3. However, they all fall within the standard deviation from the average F1 score calculated from the cross-validation technique. This shows that the model we went for is not too overfitting of our training set. The difference between the F1 score and other statistics also do not warrant that there is a lot of difference between the cross-validate models and splitting the training set by 80% since they fall within one standard deviation. The cross-validation model if anything shows that model 1 performance is not that good since the F1 score is lower and has a much larger variance.

Contributions:
Cleaning/exploratory analysis - Jaden Wong
Modeling/ exploratory analysis - Jason Cheng
Report - Jason Cheng
Powerpoint - Jaden Wong