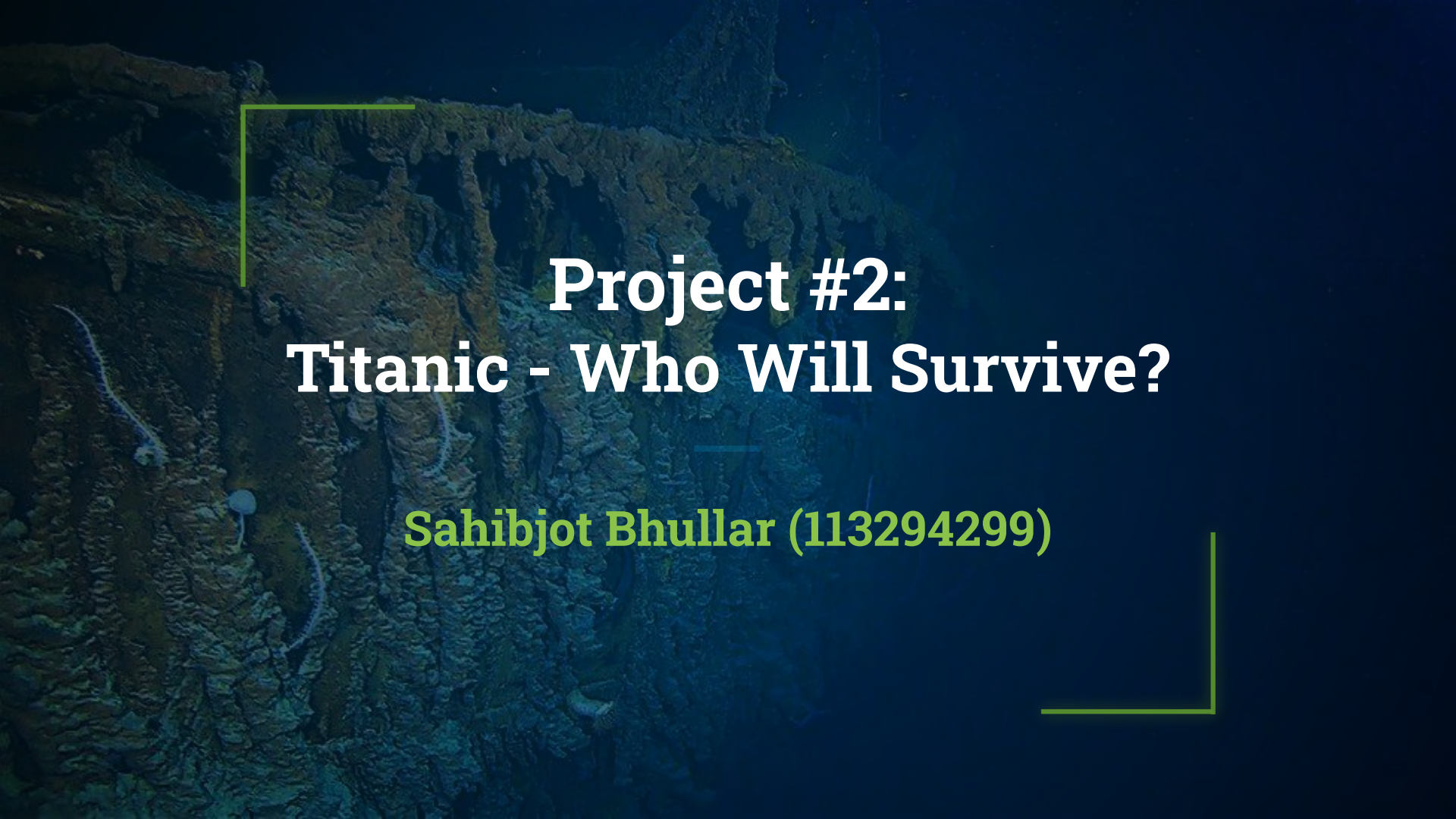


Project #2:
Titanic - Who Will Survive?

Sahibjot Bhullar (113294299)

Cleaning The Data

Dropped:

- The whole Cabin columns because it was missing 77% of information, This column had insignificant data
- Dropped missing Age value rather than filling it with average age
- Dropped the ticket column because it had 681 unique tickets , which makes it very hard to group them in categories for training the models
- Dropped the Name column as it doesn't help

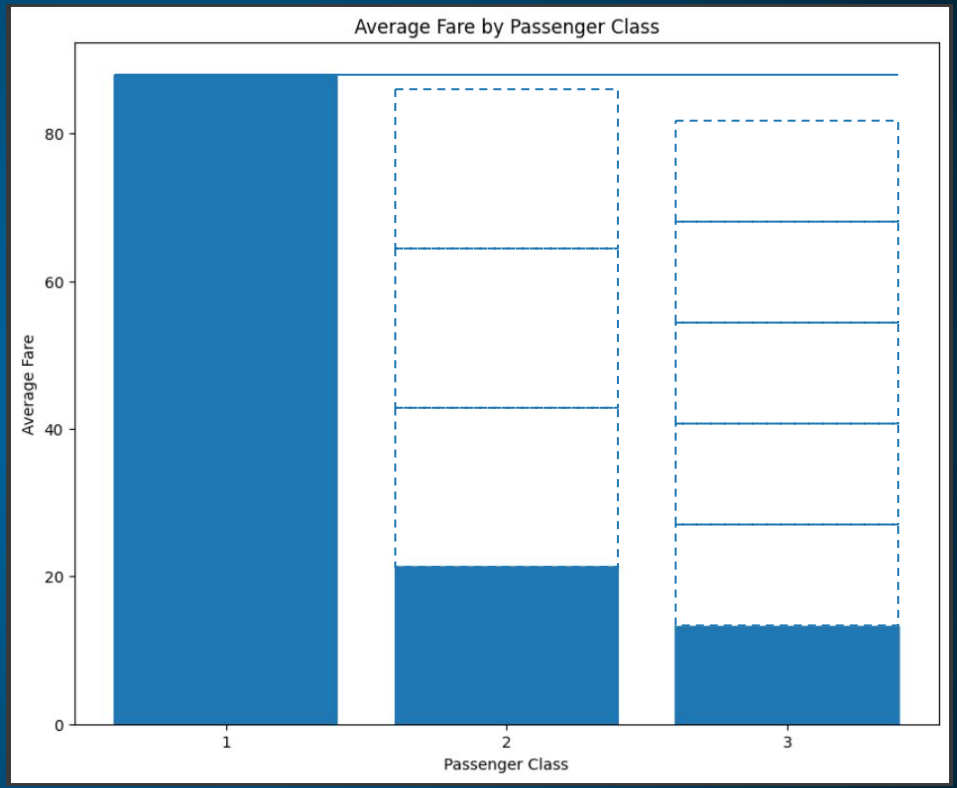
The Socioeconomic Status of the Passengers

Average Fare by Passenger Class

The *Average Fare by Passenger Class* graph shows that:

- Class 1 paid more than 4 times that of Class 2, and
- Class 1 paid more than 6 times that of Class 3

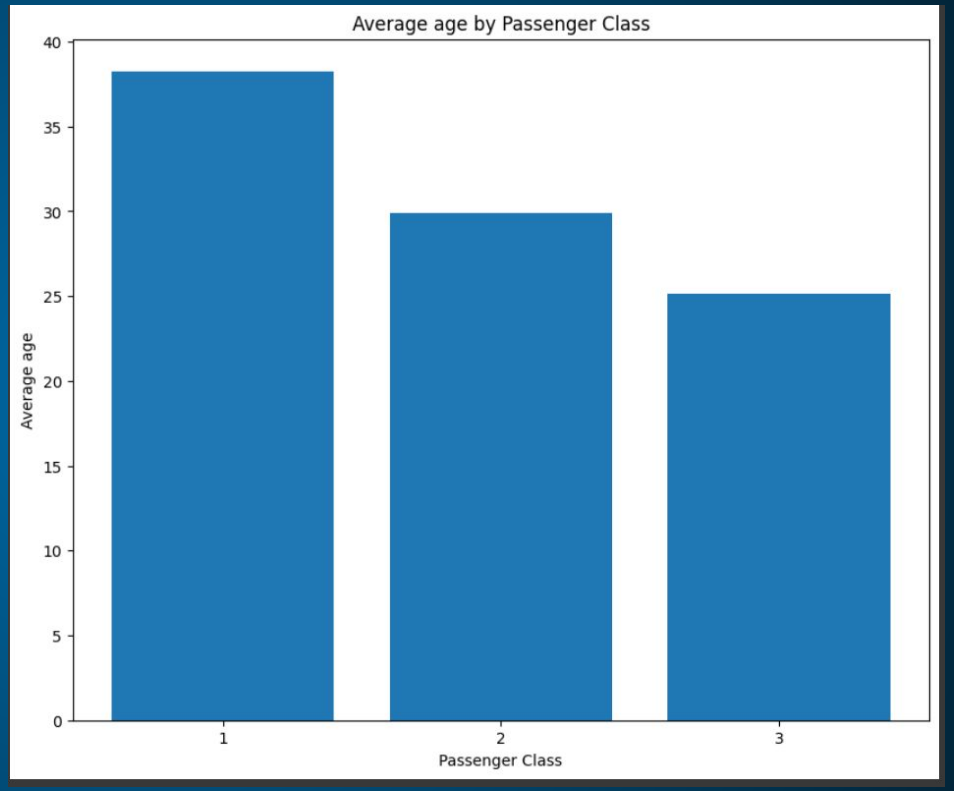
The indication of a significantly higher average fare makes it reasonable to assume that Class 1 passengers are wealthier than Class 2 and Class 3 passengers.



Average Age by Passenger Class

The *Average Age by Passenger Class* graph corroborates the assumption that Class 1 passengers are wealthier than Class 2 and Class 3 passengers.

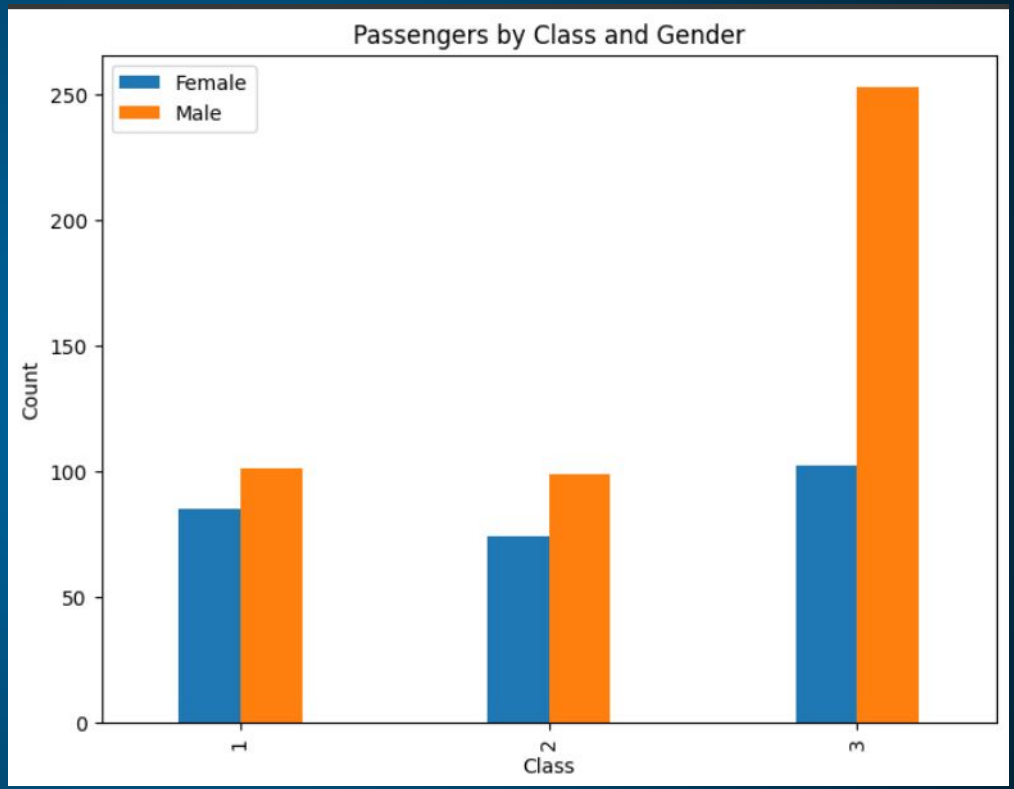
The graph asserts that class prestige and age have a direct relationship, which can be expected as wealth increases with age.



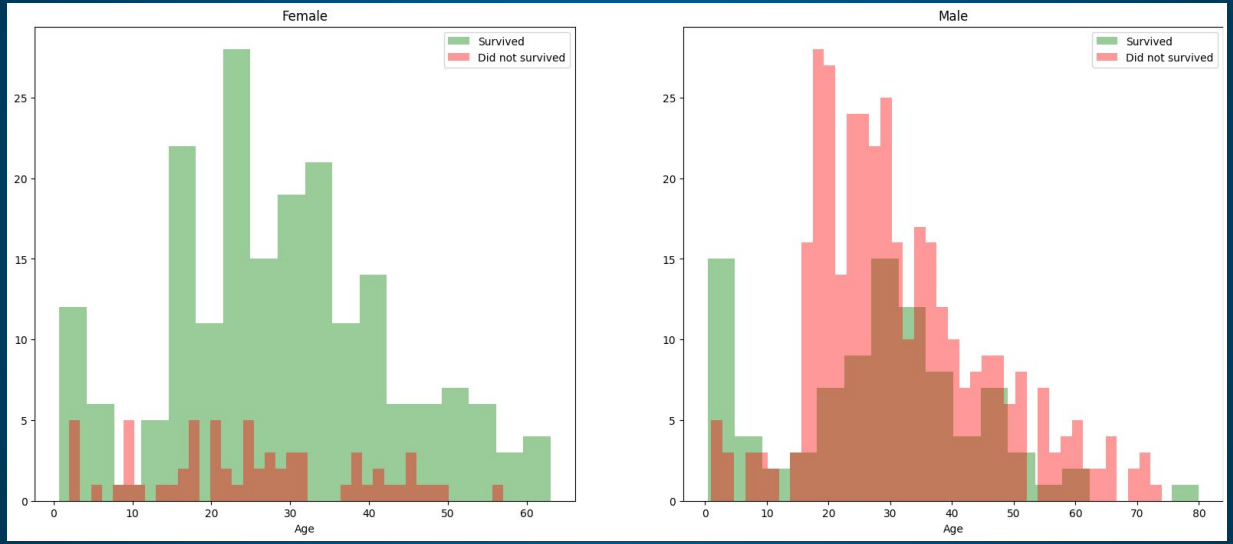
Passengers by Class and Gender

The *Passengers by Class and Gender* graph shows that Class 1 and Class 2 have almost identical distributions of females and males, making it reasonable to assume that most Class 1 and Class 2 passengers were couples as women typically did not travel alone at the time.

Whereas Class 3 has more than double the number of males than females. Combined with the *Average Age by Passenger Class* graph's assertion of Class 3 having the youngest average age, it makes it fair to assume that most of Class 3 were single and possibly traveling to seek job opportunities.

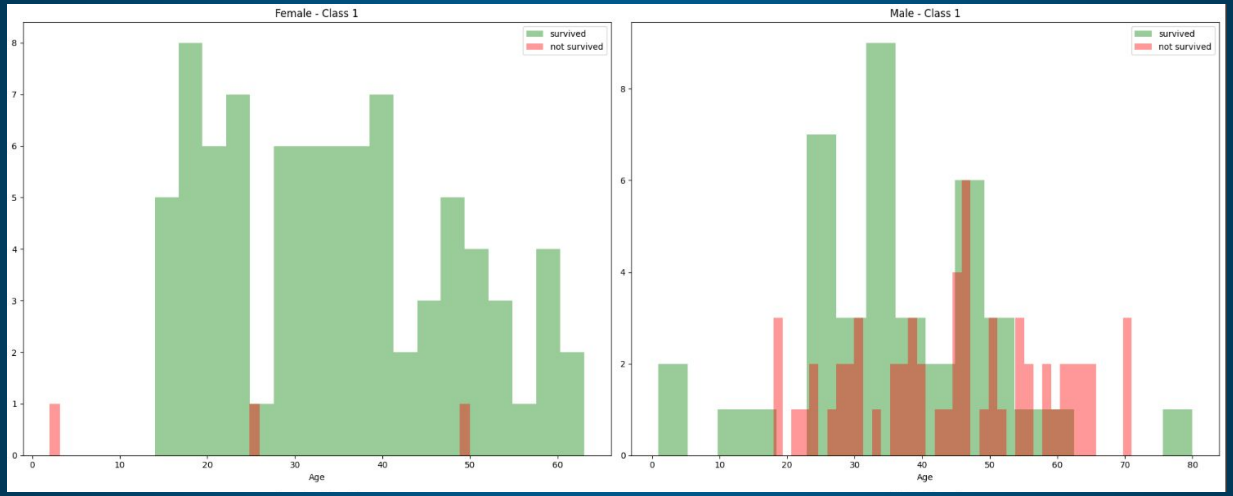


Distribution of Survivors by Age and Gender



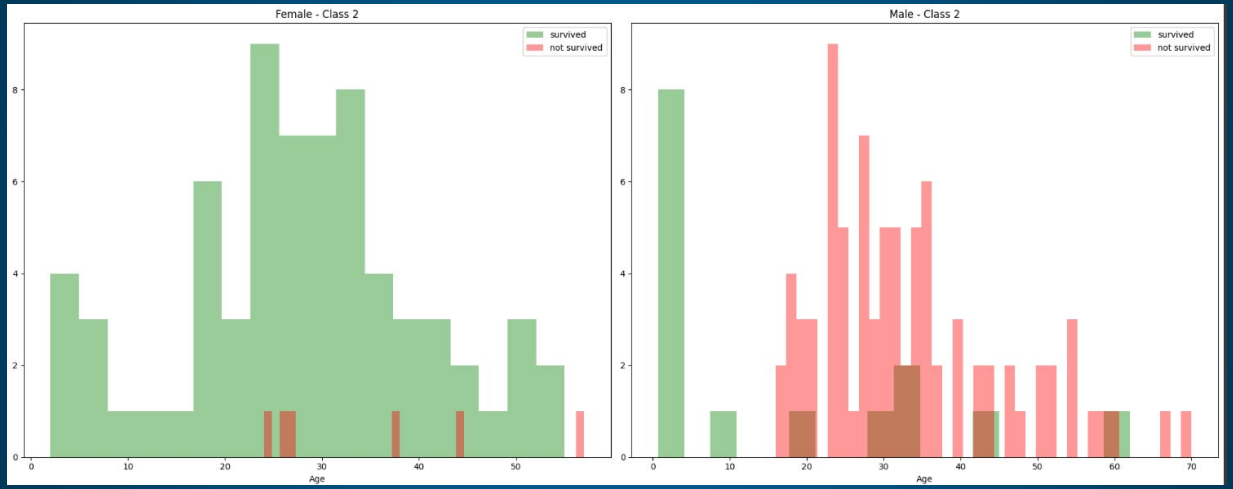
The graph above shows that overall, females had a higher chance of surviving, as well as kids between the ages of 0 and 8, regardless of gender. The highest survival rate for females was between the ages of 10 and 40 (30-year span), and for males, it was between the ages of 10 and 18 (8-year span).

Class 1



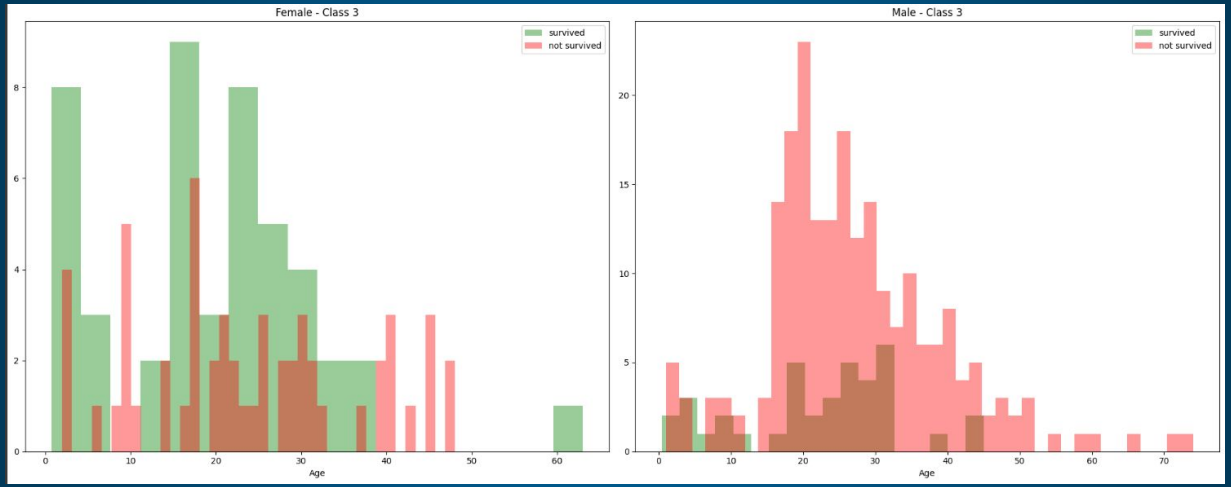
The graph above shows that Class 1 passengers across the board had a higher chance of surviving. We see the pattern continue of females having a higher opportunity to live. The age span for the highest survival rate for females was larger than for males again, but Class 1 male passengers had a much higher chance of surviving than male passengers in general.

Class 2



The graph above shows the pattern continue of females and children between the ages of 0 and 8, regardless of gender, having a higher opportunity to live. The age span for the highest survival rate for females was larger than for males once again, but Class 2 adult male passengers had a much higher mortality rate.

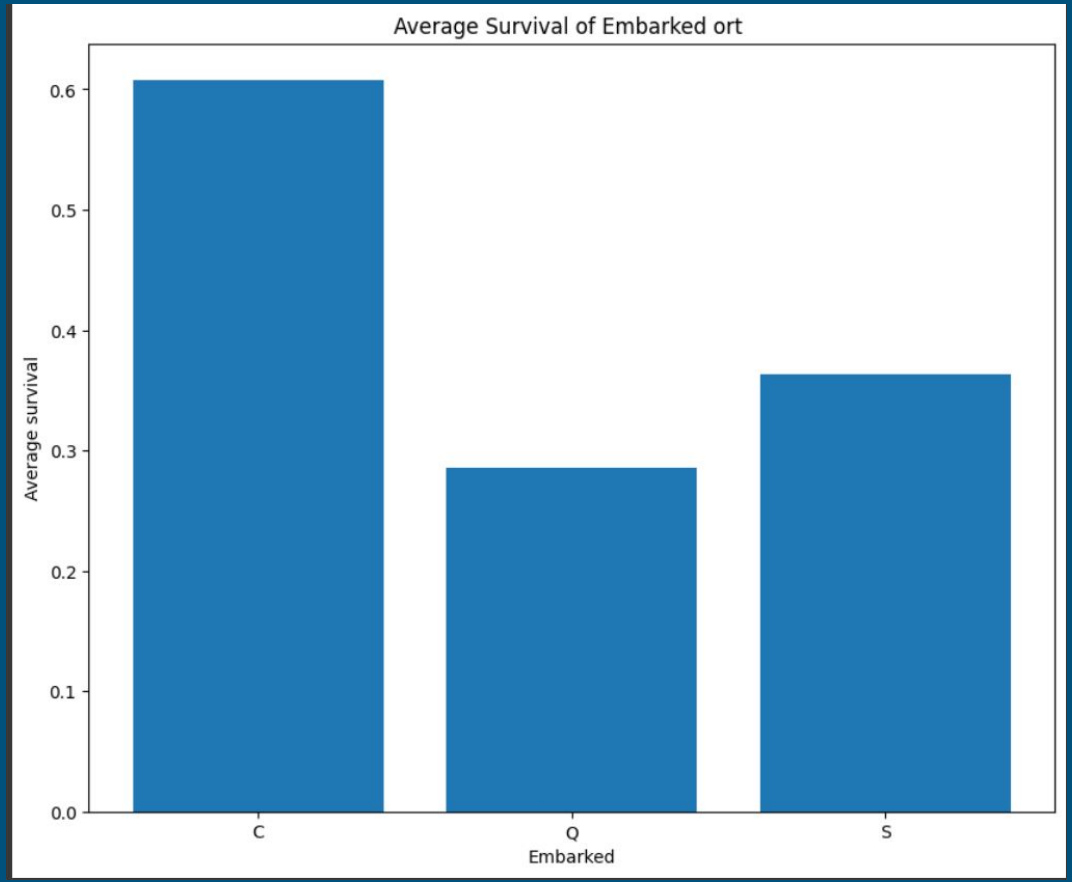
Class 3



The graph above shows the pattern continues of females having a higher opportunity to live. The age span for the highest survival rate for females was larger than for males once more, but Class 3 male passengers had a much higher mortality rate regardless of age.

Conclusion on the Socioeconomic Status of Passengers

- On the basis of gender alone, females had a significantly better chance of surviving than their male counterparts.
- In terms of female passengers, Class 1 and Class 2 female passengers had higher chances of surviving than Class 3 female passengers.
- In terms of male passengers, Class 1 male passengers had higher chances of surviving than Class 2 or Class 3 male passengers.
    - More specifically, Class 1 and Class 2 male children passengers had a much higher chance of surviving than Class 3 male children passengers.


Average Survival of Embarked ort
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Average survival
C
Q
S
Embarked

Correlation Coefficient for Survival

with Fare:	0.26818861687447876	
with Age:	-0.07722109457217768	← Lowest
with Gender:	0.5388255930146364	← Highest
with SibSp:	-0.01735836047953421	
with Parch :	0.09331700774224293	
with Embarked:	0.1896570103153622 8	

Gender seems to have the highest correlation with survival.

Logistic Regression Model

- Logistic regression finds the relation between independent variables and probability of dependent variable
- Works as binary classification by finding probability classify the data between 0 and 1 of dependent variable
- Applies the logistic function
- In my code it uses logistic regression to predict the likelihood of passenger surviving using different columns

Random Forest

- Builds multiple decision trees using random subsets of training and features
- Averages the prediction to make final predictions
- In my code I use the scikit-learn library, I specified the number of decision trees and it works by creating collection of decision trees and each is trained on randomly selected subset of the data and features, finally it averages the predictions of all the individual decision trees to make a final prediction.

Nearest Neighbor

- This model works by finding the k closest points in the training set to a new data point and predicting the label based on the majority label of those k neighbors.
- I chose the number of nearest neighbors to consider parameter to be 5
- In this case the closest neighbor are either 1 or 0.

Evaluating Models

	Accuracy	Precision	Recall	F1 score
Logistic Regression Model	0.804	0.843	0.683	0.754
Random Forest	0.762	0.754	0.683	0.717
Nearest Neighbor	0.692	0.646	0.667	0.656

Evaluating Models Table