Final Project Statistical Learning

Matthew Clarke and Lisa Olsson

Introduction:

The sinking of the RMS Titanic in 1912 is one of the most infamous disasters in maritime history. Despite being considered unsinkable, the ship collided with an iceberg and tragically claimed the lives of many passengers and crew. However, among the survivors, certain groups of people were more likely to make it out alive than others. In this machine learning project, we aim to predict which factors contributed to the survival of Titanic passengers, based on their personal information such as age, gender, socio-economic status, and cabin class. By analyzing and modeling the dataset of Titanic passengers, we can gain insights into the characteristics of those who were most likely to survive the disaster.

Data:

- PassengerID: Unique identifier for passenger
- Survival: 0 = No, 1 = Yes
- Pclass: Ticket class, 1 = 1st, 2 = 2nd, 3 = 3rd (A proxy for socio-economic status)
- Name: Full Name of Passenger
- Sex: Male, Female
- Age: Age in years
- Sibsp: # of siblings / spouses aboard the Titanic,
- Parch: # of parents / children aboard the Titanic,
- Ticket: Ticket number,
- Fare: Passenger fare paid for the ticket
- Cabin: Cabin number (the first letter indicates location onboard)
- Embarked: Port of Embarkation, C = Cherbourg, Q = Queenstown, S = Southampton

Part 1: Exploratory Analysis

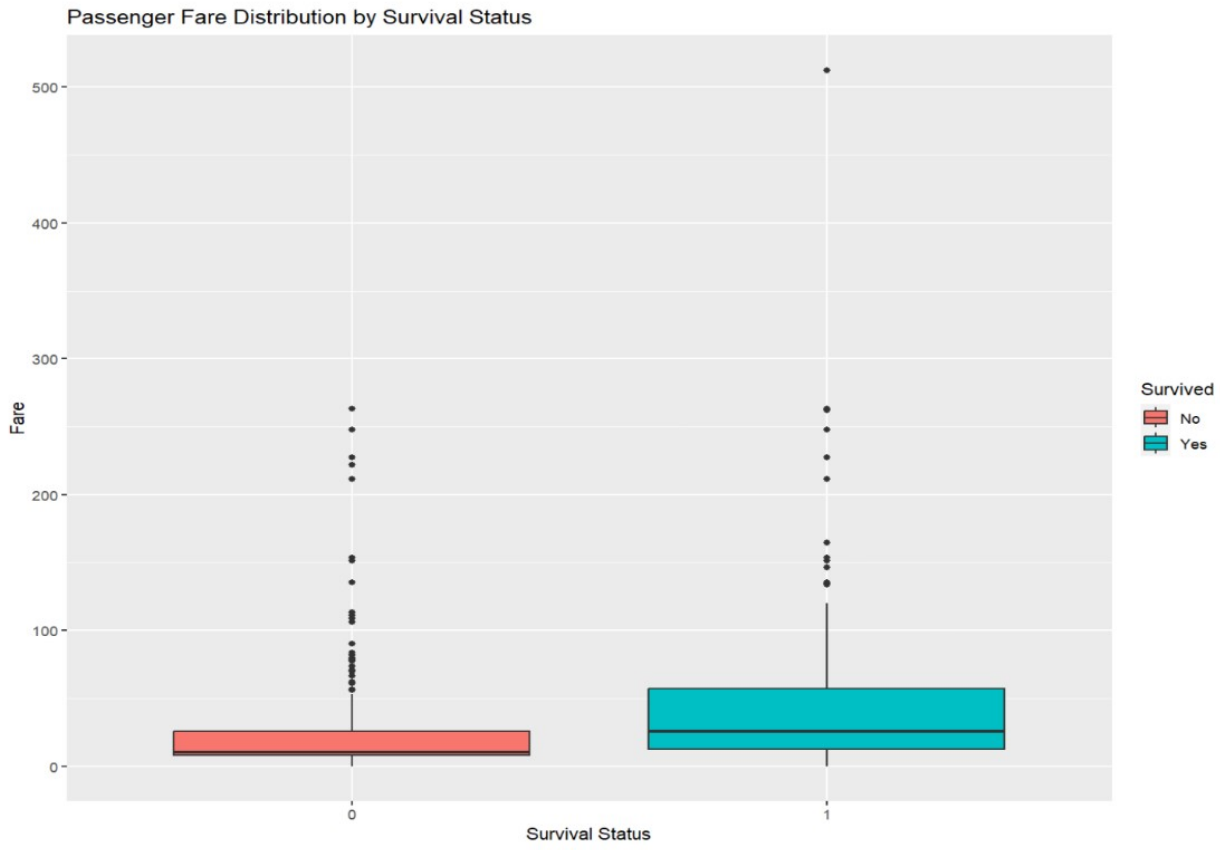
Looking at the data 177 rows had NAs in them. However, these all came from the same variable, Age.

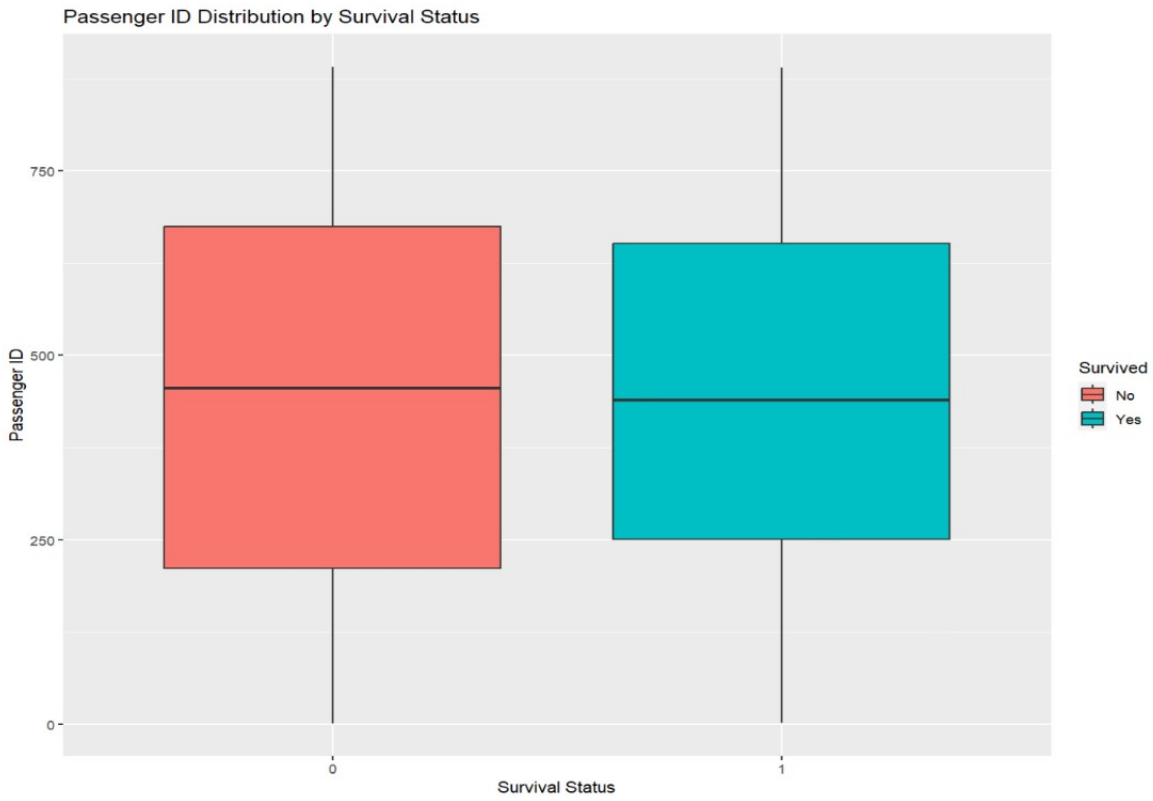
```
  PassengerId       Survived          Pclass          Name               Sex                 Age       
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42  
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12  
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                                         Mean   :29.70  
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00  
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00  
                                                                                        NA's   :177    
     SibSp           Parch           Ticket               Fare            Cabin              Embarked        
 Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00   Length:891         Length:891        
 1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character   Class :character  
 Median :0.000   Median :0.0000   Mode  :character   Median : 14.45   Mode  :character   Mode  :character  
 Mean   :0.523   Mean   :0.3816                      Mean   : 32.20                                        
 3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00                                        
 Max.   :8.000   Max.   :6.0000                      Max.   :512.33                                        
```

Searching for NAs in all variables:

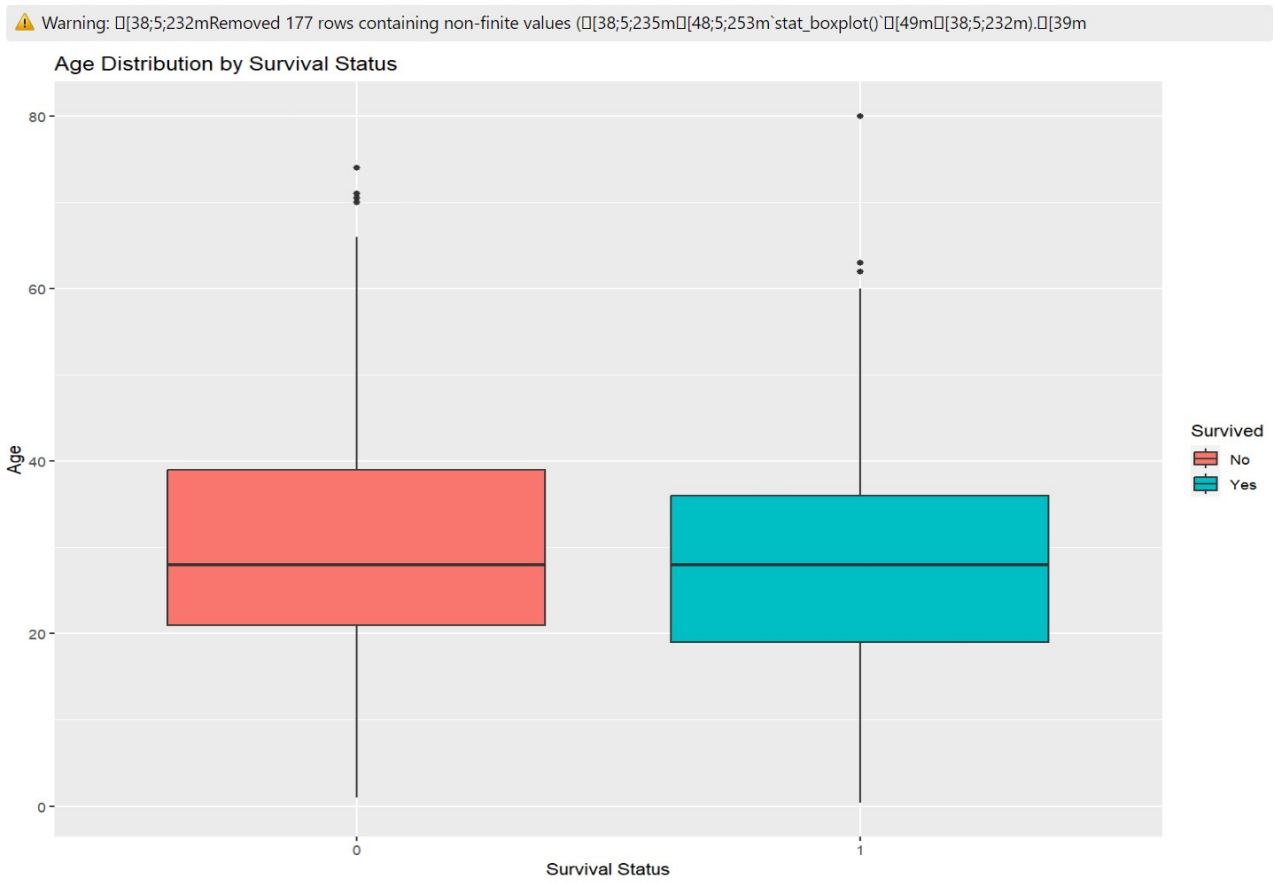
```
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket
          0           0           0           0           0         177           0           0           0
       Fare       Cabin    Embarked
          0           0           0
```

Numeric Variables and Survival Status:



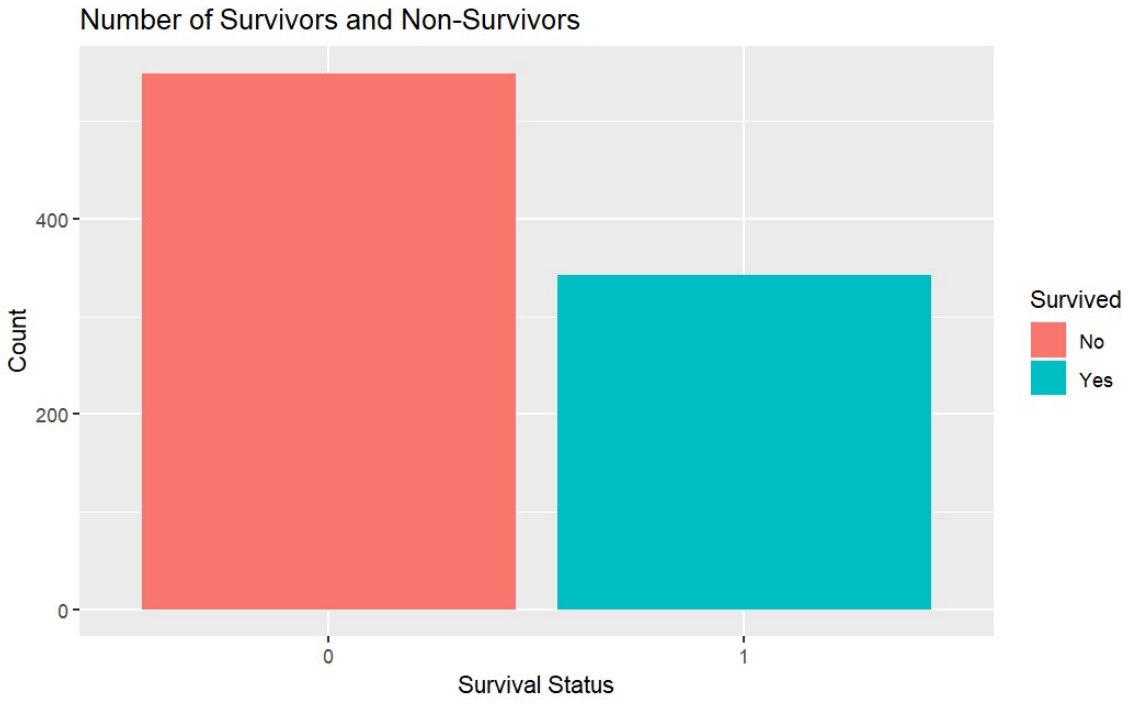


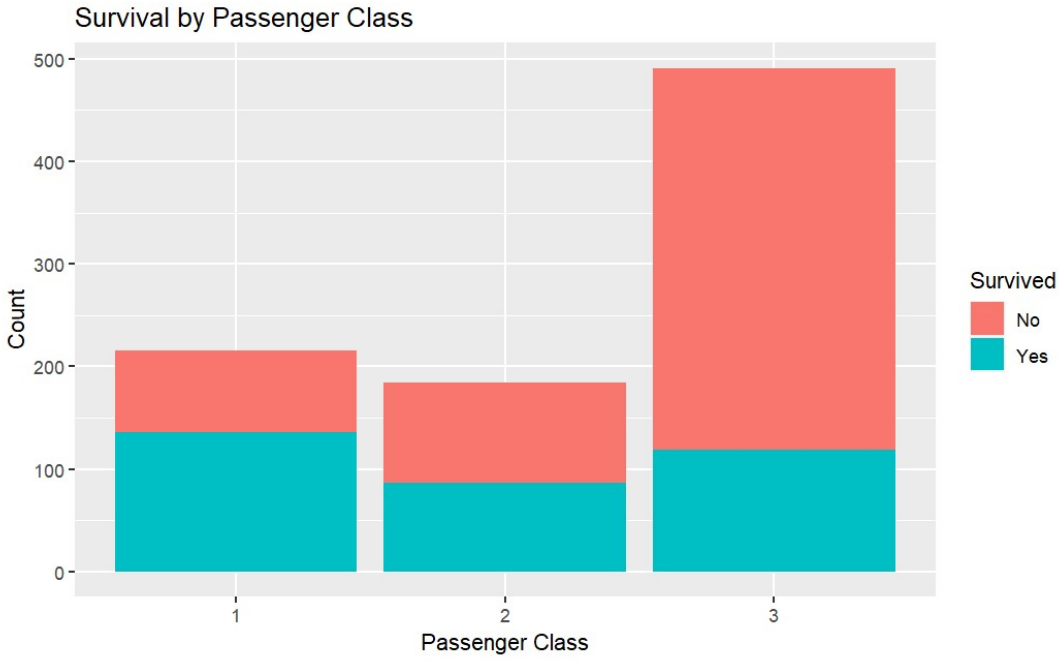
Passenger ID is not supposed to mean anything, its an arbitrary number for each passenger. We still plotted it to confirm that this theory held true and that there wasn't any information hidden in the variable.

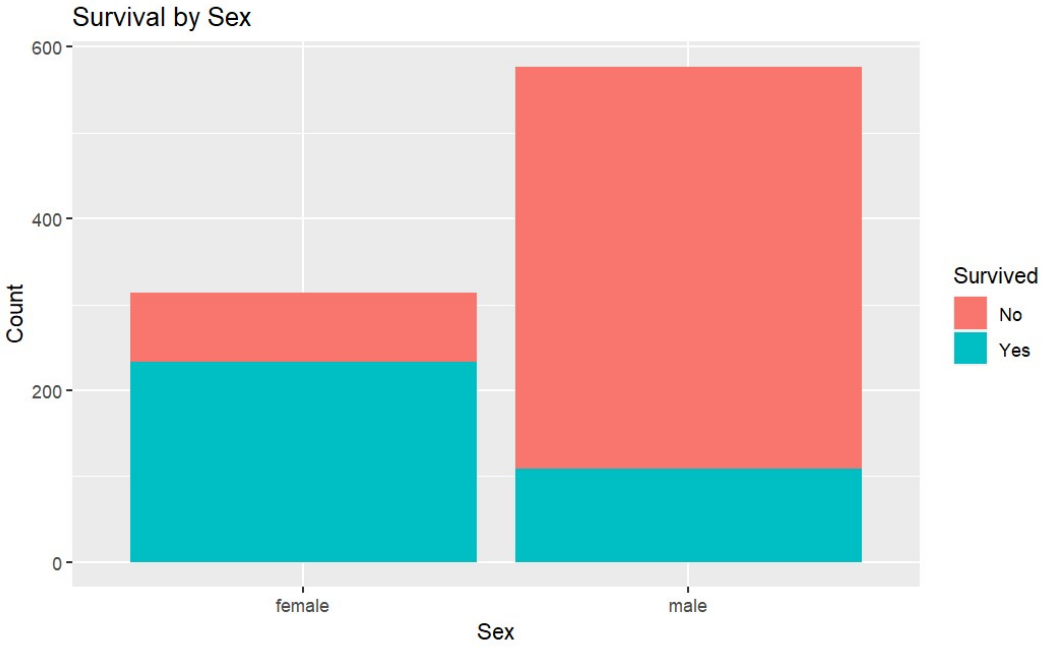


For the numeric variables it appears as if fare is the one with the strongest relationship to survival status. For Age there is a slight skew towards younger passengers surviving however the difference does not appear to be too significant.

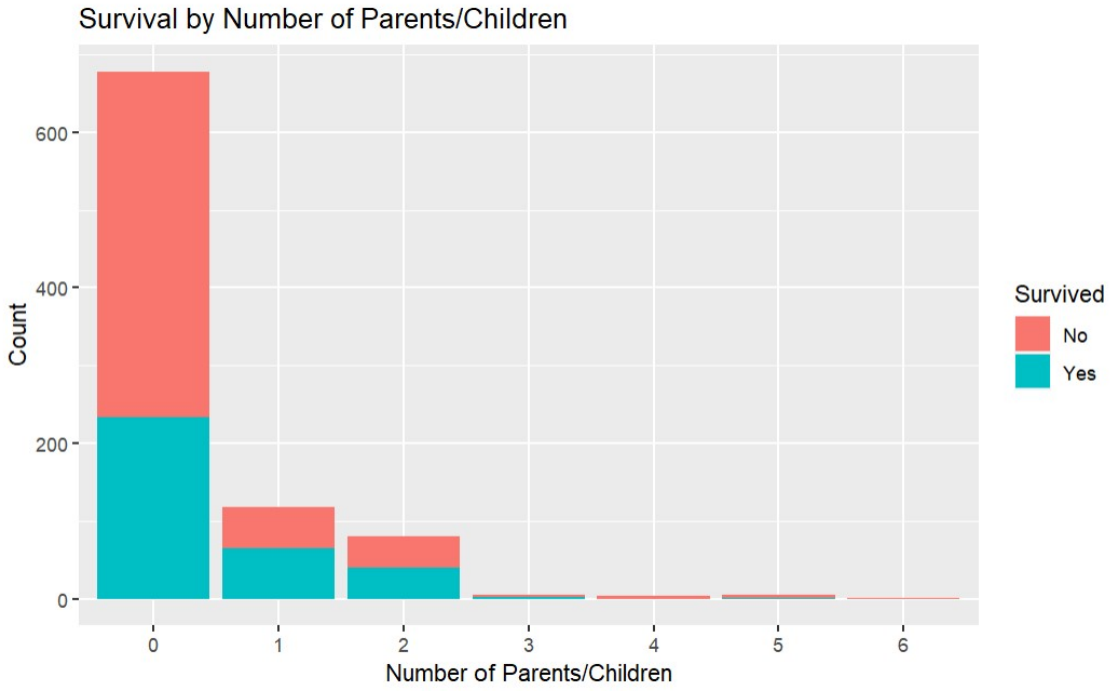
Categorical Variables:

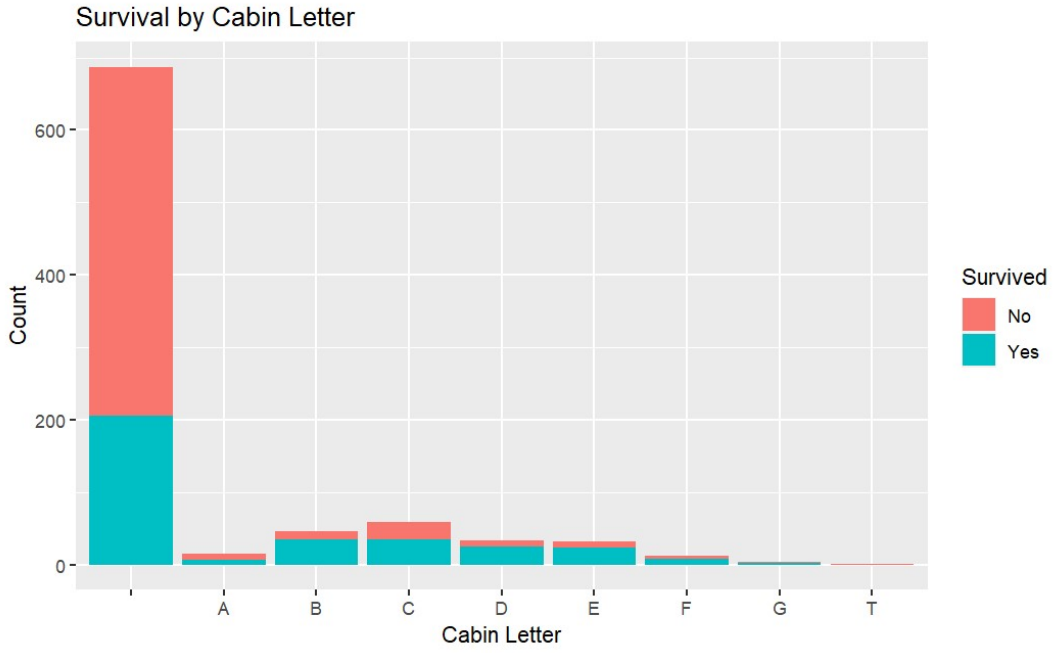

Number of Survivors and Non-Survivors
Count
400
200
0
0
1
Survival Status
Survived
No
Yes

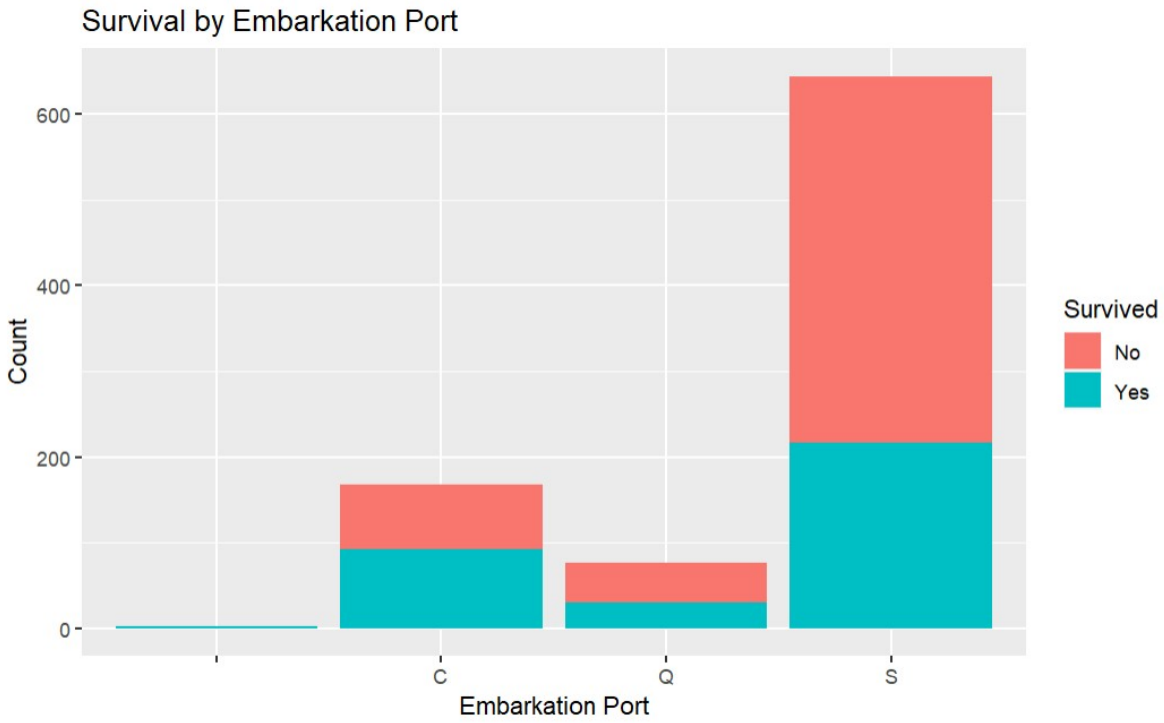

Survival by Passenger Class
Count
500
400
300
200
100
0
1
2
3
Passenger Class
Survived
No
Yes


Survival by Sex
600
400
200
0
Count
female
male
Sex
Survived
No
Yes


Survival by Number of Siblings/Spouses
600
400
200
0
Count
0
1
2
3
4
5
8
Number of Siblings/Spouses
Survived
No
Yes


Survival by Number of Parents/Children
600
400
200
0
Count
0
1
2
3
4
5
6
Number of Parents/Children
Survived
No
Yes


Survival by Cabin Letter
600
400
200
0
Count
A
B
C
D
E
F
G
T
Cabin Letter
Survived
No
Yes



In the categorical variables we can see that there were overall more people dying in our dataset than surviving.

Out of the passengers in class 1 the majority are survivors, in class 2 it looks like about a 50/50 split of survivors and non survivors while class 3 has a significant majority of non survivors. Class 3 is also by far the biggest class in terms of number of people overall.

There are far more males than females onboard the Titanic however the representation of female survivors is significantly larger than male survivors.

Looking at family members onboard is hard to make any inference of the relationship to survival or non-survival.

The cabin letter is a little interesting. Cabin letter indicates where on the ship the cabin is located. Even though the majority of this variable is missing information we could potentially see if there are any zones on the ship which favored surviving. This could however likely be correlated to the Pclass variable. From the bar chart it is hard to draw any conclusions.

Thinking logically about the project, embarkation port shouldn't be useful information, however we still plotted it to make sure we didn't miss any hidden information.

Data Manipulation:

We turned all character values plus the binary variable Survived and the Pclass variable to the data type factor.

The variable Cabin contained information about where on the ship the cabin was located through the first letter of the record. Therefor we extracted the first letter from each record and got rid of the remaining numbers that simply indicated which exact unit.

We will later discover that Age is a statistically significant factor in the logistic regression. It is also one of few numeric values for our KNN model. In order to not lose the variable or the 177 records with NA values we decided to fill in the missing values with the mean of the non-missing age values, which was 29.7.

We also decided to exclude the Variable Name and Ticket as these are text variables where each instance has a unique input. Attempting to do any analysis with these text variables falls outside the scope of the project.

Part 2: Logistic Regression

```
Call:
glm(formula = Survived ~ ., family = binomial, data = test)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3689  -0.4789  -0.1835   0.4995   2.5207

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.039e+01  2.625e+03   0.008  0.99380
PassengerId  5.303e-05  1.064e-03   0.050  0.96026
Pclass2     -5.504e-01  1.266e+00  -0.435  0.66378
Pclass3     -2.818e+00  1.393e+00  -2.024  0.04297 *
Sexmale     -2.675e+00  5.552e-01  -4.819 1.44e-06 ***
Age         -8.484e-02  2.757e-02  -3.077  0.00209 **
SibSp       -7.205e-01  4.339e-01  -1.661  0.09681 .
Parch        1.505e-01  3.175e-01   0.474  0.63559
Fare         3.401e-03  1.051e-02   0.324  0.74627
CabinA      -1.677e+01  1.887e+03  -0.009  0.99291
CabinB       1.430e+00  1.750e+00   0.817  0.41385
CabinC      -2.747e-01  1.656e+00  -0.166  0.86824
CabinD       4.166e-01  1.225e+00   0.340  0.73374
CabinE       2.821e+00  1.648e+00   1.712  0.08699 .
CabinF       1.182e+00  1.548e+00   0.763  0.44523
CabinG      -1.979e+01  3.956e+03  -0.005  0.99601
EmbarkedC   -1.542e+01  2.625e+03  -0.006  0.99531
EmbarkedQ   -1.532e+01  2.625e+03  -0.006  0.99534
EmbarkedS   -1.537e+01  2.625e+03  -0.006  0.99533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 198.03  on 146  degrees of freedom
Residual deviance: 106.59  on 128  degrees of freedom
  (32 observations deleted due to missingness)
AIC: 144.59

Number of Fisher Scoring iterations: 16
```

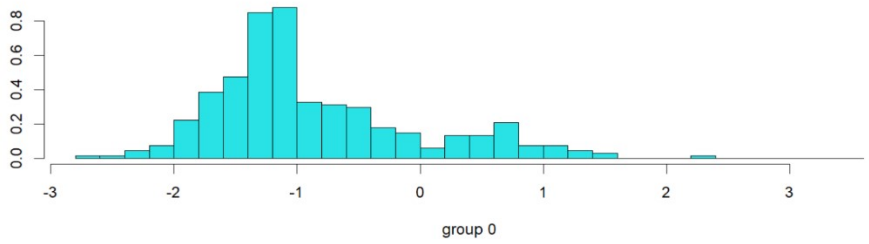
Confusion Matrix:

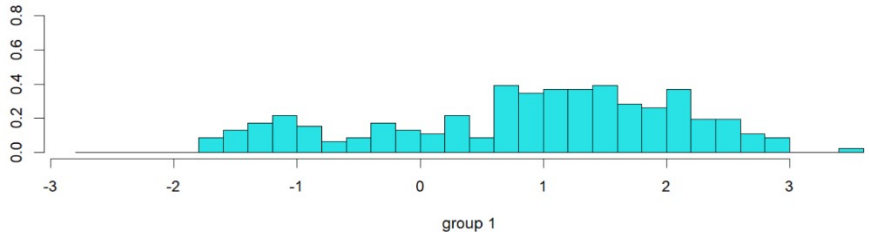
	True	
logpred	0	1
0	99	14
1	14	52

Performance Metrics:

- Accuracy: 84,4%
- Error Rate: 15,6%
- Sensitivity (TP/(TP+FN)):78.8%
- Specificity (TN/(TN+FP)):87.6%

LDA:





	0	1
0	70	18
1	15	44

Based on the logistic regression, which we found much more accurate and interpretable than LDA, the main variables that impact one's likelihood of death on the Titanic were Age, Sex being Male, and PClass 3. Our prior thoughts, especially with the knowledge derived from the Titanic film, led us to believe that Age, Sex, PClass and Fare price would have statistically significant impacts on determining survivors on the ship. PClass and Fare price, due to their ties to socioeconomic status, were expected to have similar

impacts, but this assumption was not supported by our Logisitic Regression of the data. Additionally, Logistic Regression resulted in a rather high accuracy of 84.%.

Part 3: KNN

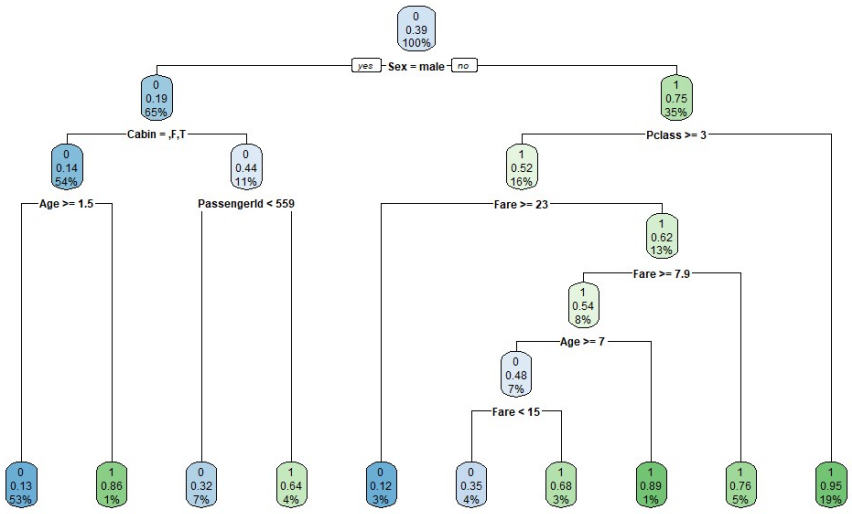
For our KNN model we picked the numeric variables, Age and Fare as well as the factor variable Pclass. Since the levels of the Pclass variable have inherent meaning and we saw the significance of the variables in the logistic regression we choose to do our best to keep it. We turned the variable back into an integer for the KNN model. The interpretation of this is that the move between first class and second class is as big as the move between the second and third. We feel comfortable with this interpretation and therefore included it in our model.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 97 18
         1 16 48

               Accuracy : 0.8101
                 95% CI : (0.7448, 0.8647)
    No Information Rate : 0.6313
    P-Value [Acc > NIR] : 1.537e-07

                  Kappa : 0.5894

 Mcnemar's Test P-Value : 0.8638

            Sensitivity : 0.8584
            Specificity : 0.7273
         Pos Pred Value : 0.8435
         Neg Pred Value : 0.7500
             Prevalence : 0.6313
         Detection Rate : 0.5419
   Detection Prevalence : 0.6425
      Balanced Accuracy : 0.7928

       'Positive' Class : 0
```

The model is giving us a similar result to the logistic regression but with a lot less variables. One big loss is that we didn't have the gender variable in the KNN model. We could potentially have added it as a binary variable. But we chose not to do so and will instead look at how including the gender variable (among others) will impact the decision tree models which are better at handling categorical variables.

Part 4: Decision Tree

Starting with a simple decision tree to visualize the output.

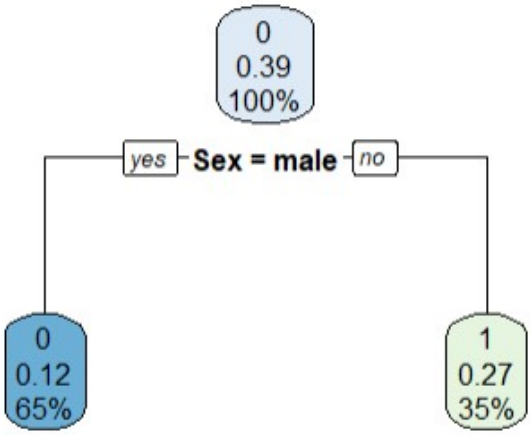


	predicted	
	0	1
0	97	16
1	20	46

Performance Metrics:

- Accuracy: 79.9 %
- Error Rate: 20.1%
- Sensitivity (TP/(TP+FN)):69.7%
- Specificity (TN/(TN+FP)):85.8%

Pruned Decision Tree



The unpruned decision tree had lower accuracy with predicting as compared to Logistic Regression, but was very interpretable when assessing and following the tree. In a sense this could have been used by on board officials to decide who was to be prioritized in the evacuation, as this produced a robust, black and white walk through with a respectable accuracy of 79.9%. The pruned tree resulted in an extremely simple breakdown that was only influenced by the gender of the passenger.

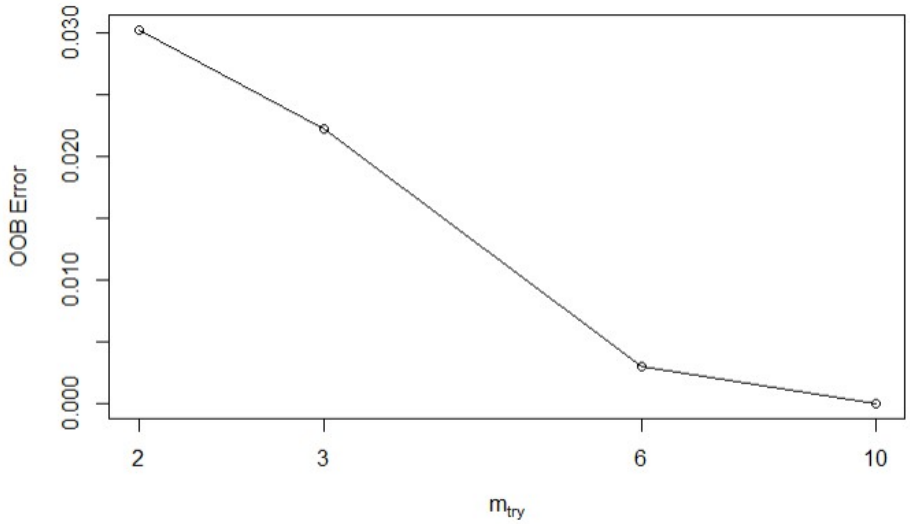
Boosting:

```
##### xgb.Booster
raw: 7.8 Kb
call:
  xgb.train(data = ddata1, nrounds = 5, watchlist = list(train2 = ddata1,
    eval = ddata1), max_depth = 2, eta = 1, nthread = 4, objective = "binary:logistic")
params (as set within xgb.train):
  max_depth = "2", eta = "1", nthread = "4", objective = "binary:logistic",
validate_parameters = "TRUE"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
  cb.evaluation.log()
# of features: 10
niter: 5
nfeatures : 10
evaluation_log:
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 106   0
         1   0  73

               Accuracy : 1
                 95% CI : (0.9796, 1)
    No Information Rate : 0.5922
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5922
         Detection Rate : 0.5922
   Detection Prevalence : 0.5922
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

With the XG Boost technique we were able to derive a prediction model that had an accuracy of 100%, which was the highest among any previously constructed model and maintained 100% sensitivity and specificity. Since there was such a jump in accuracy as compared to previous models, we were cautious by the results, but XG Boost is a powerful method that helps reduce bias, something that is prevalent in our data. Additionally, with our dataset having a high dimensionality, but low number of observations, the misclassifications must have been quickly reduced through iterations of decision trees produced. Overall, the best model in terms of accuracy, but not for interpretability.

Random Forrest:




```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 96 16
         1 17 50

               Accuracy : 0.8156
                 95% CI : (0.751, 0.8696)
    No Information Rate : 0.6313
    P-Value [Acc > NIR] : 5.992e-08

                  Kappa : 0.6052

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8496
            Specificity : 0.7576
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.7463
             Prevalence : 0.6313
         Detection Rate : 0.5363
   Detection Prevalence : 0.6257
      Balanced Accuracy : 0.8036

       'Positive' Class : 0
```

Through optimization of the m_{try}, we found the minimal Out-Of-Bag Error at $m_{try} = 10$. We decided to use 500 trees as this was just above half of our raw data size in order to determine a model for predicting the Survived variable. The model had a rather high accuracy at 0.8156 and a sensitivity of 0.8496. This

indicates a high performance for our model and based on the outputs from the decision tree, minimal false positives and false negatives were the results.

Part 5: Support Vector Machine – SVM

```
Call:
svm(formula = Survived ~ ., data = train, type = "C-classification", kernel = "linear", cost = 0.1, scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1

Number of Support Vectors:  348

 ( 171 177 )


Number of Classes:  2

Levels:
 0 1


Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 95 22
         1 18 44

               Accuracy : 0.7765
                 95% CI : (0.7084, 0.8353)
    No Information Rate : 0.6313
    P-Value [Acc > NIR] : 2.151e-05

                  Kappa : 0.5139

 Mcnemar's Test P-Value : 0.6353

            Sensitivity : 0.8407
            Specificity : 0.6667
         Pos Pred Value : 0.8120
         Neg Pred Value : 0.7097
             Prevalence : 0.6313
         Detection Rate : 0.5307
   Detection Prevalence : 0.6536
      Balanced Accuracy : 0.7537

       'Positive' Class : 0
```

Similar to KNN, Support Vector Machine is a supervised learning algorithm, but is designed to work best with linear and non-linear classification data. Due to the efficiency and strength of this model, we were expecting similar results to XG Boost, but instead it was our worst performing model (excluding LDA). We suspect that this is due to the limited observations and high number of variables that were fed through the model.

Part 6: Results and Concluding Thoughts

	Logistic Regression	Decision Tree	KNN	XG Boost	SVM
Accuracy	84.4%	79.9%	81%	100%	77.7%
Error Rate	15.6%	20.1%	19%	0%	22.3%
Sensitivity	78.8%	69.7%	85.9%	100%	84.6%
Specificity	87.7%	85.8%	72.3%	100%	66.6%

Based on our initial MLR, most of our variables did not have a measurable impact on whether an individual died while on the Titanic with only Age and Pclass3 being statistically significant. However, through our modeling techniques we were able to predict our variable of interest with an accuracy as high as 100% and a sensitivity of 100% (XG Boost). The accuracies among all our models were roughly the same, besides XG Boost, the average was 80.75% with similar Sensitivity and Specificity percentages. With such a perfect prediction result from XG Boost, we were left slightly skeptical by the vast improvement from our other models to one with 100% accuracy. We would potentially include our KNN and Logistic Regression models as one of the better models due to the interpretability and still low error rate. Adjusting the techniques used to replicate Age and Cabin for missing data could provide more clarity as to if our prediction models were produced properly. And with refining or addition of data, we would be able to tune the current models which may uncover more relationships among the variables than previously observed. Overall, we found ease with creating and understanding the models taught in class.

The learning outcomes derived from this project were very tangible for future application of statistical and prediction models, with a highlight on communicating results and incorporating unique data into the models learned in the lectures and other assignments. This allowed us to understand that some models are rather inflexible and require more caution or additional code to accommodate. Additionally, the project was unique in the fact that there wasn't necessarily a research question for our group to consider as compared to previous MSBA projects, which proved to be helpful with choosing the data set, but at first caused confusion around what was derived from the models.