Titanic Survival

William Duquette

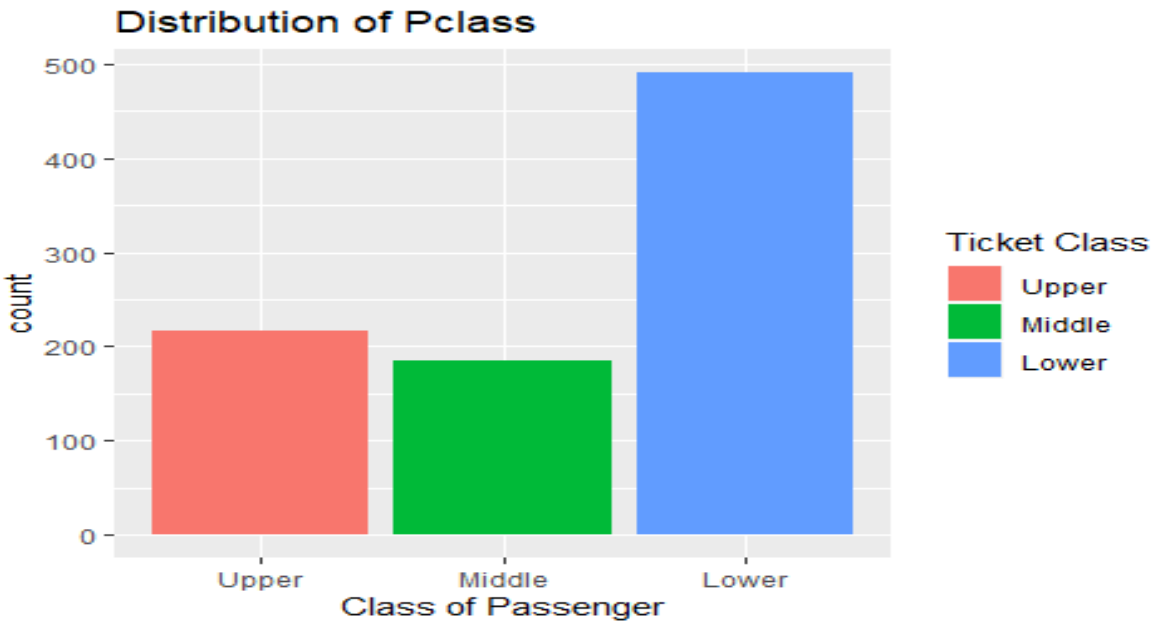
5/7/2020

Project Description: In 1912, the Titanic attempted to sail across the Atlantic Ocean from England to the New York. It was hailed as a wonder of engineering and was thought to be unsinkable. Before it could reach New York, it struck an iceberg, and thousands of people died. It remains one of the most infamous sinking's of all time, a sinking in which many people perished. The data being used in this study was gathered from the manifest of the Titanic, and other contemporaneous records. I will build a model that will predict which factors increase and decrease the odds of a passenger surviving.

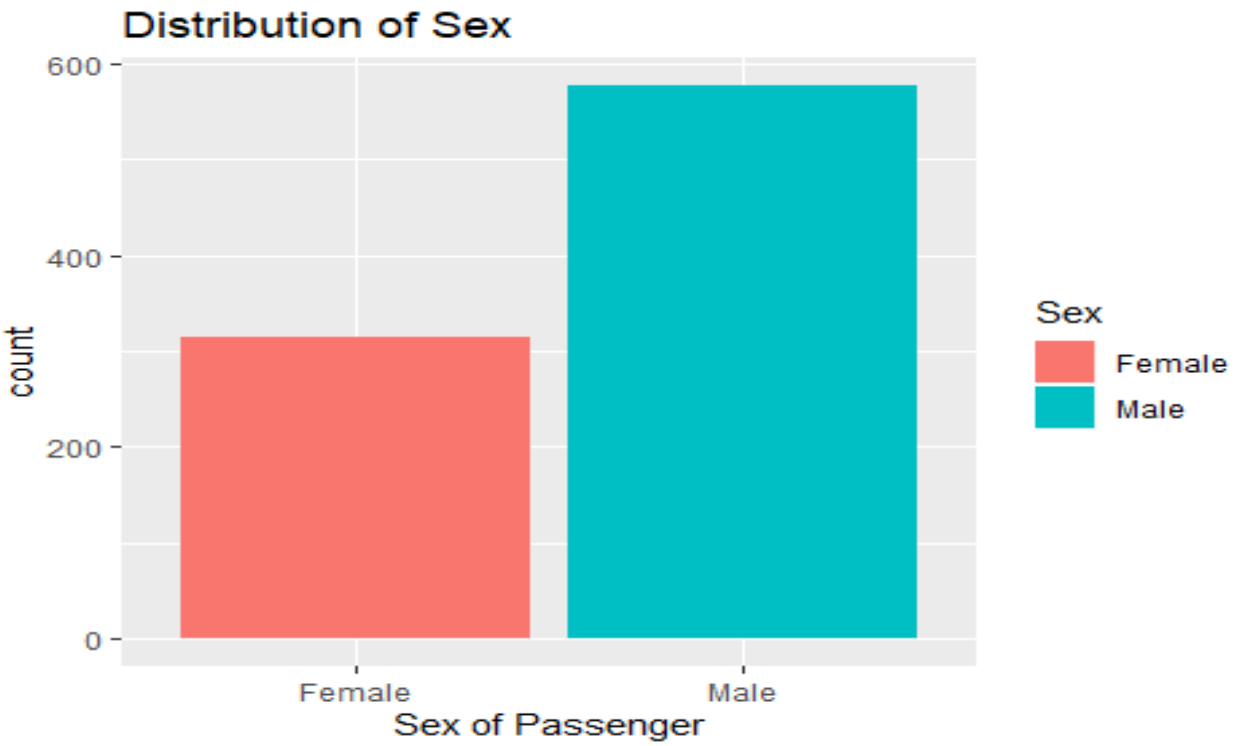
Relevant Variables: 9 variables are inside this data set. For this question, survived will be the response variable. This variable, Survived, is an indication of whether the passenger survived or not (see Exploratory Data Analysis). The explanatory variables (before possible transformations) that could be included in the model are listed below:

1. PassengerID - This identification variable was a unique identification number given to each passenger.

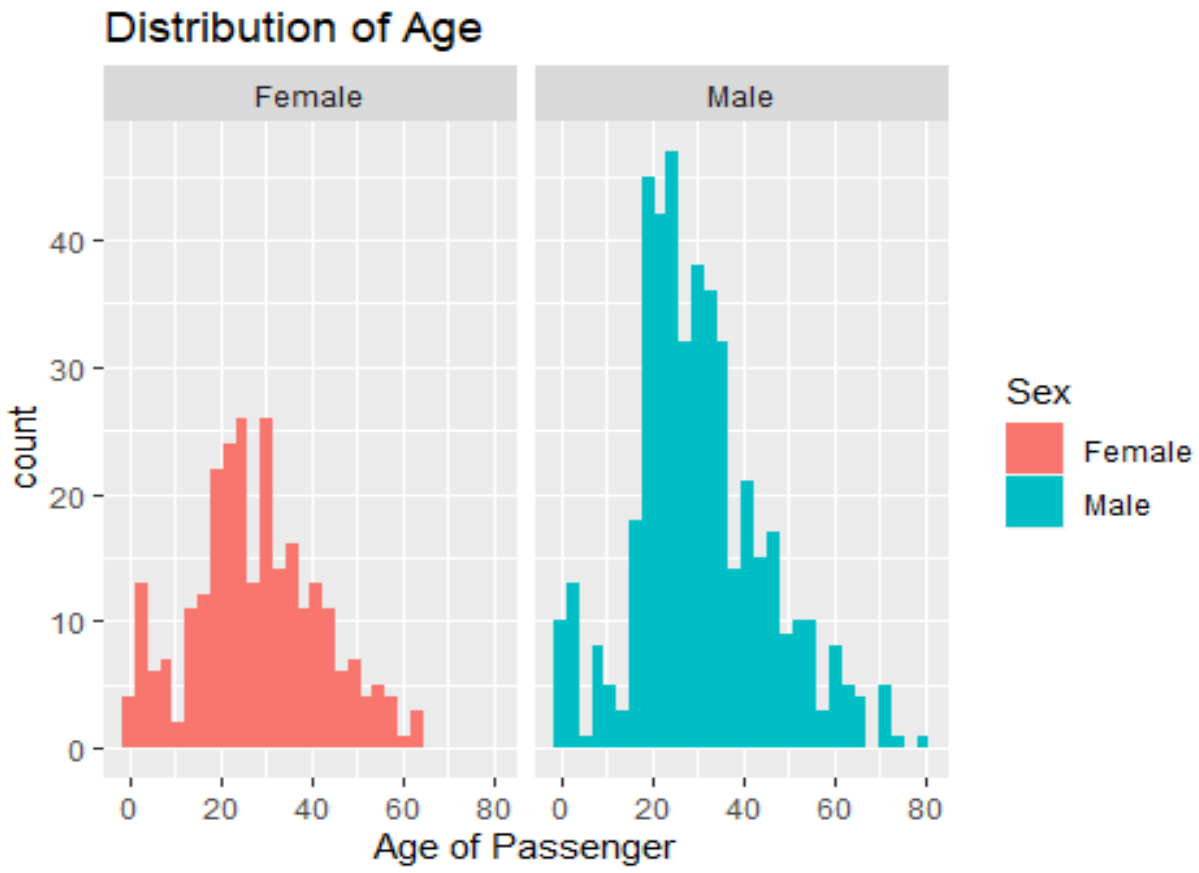
2. Pclass - This categorical variable denotes the class of the passenger. This could be used as a stand in for socio-economic standing, due to the price difference of tickets. The three possible classes are: 1 (upper class), 2 (middle class), 3 (lower class). Below is a graph that shows the distribution of the explanatory variable. As you can see there are more passengers in third class than any other class, while there is only a difference of 32 passengers between first and second class.



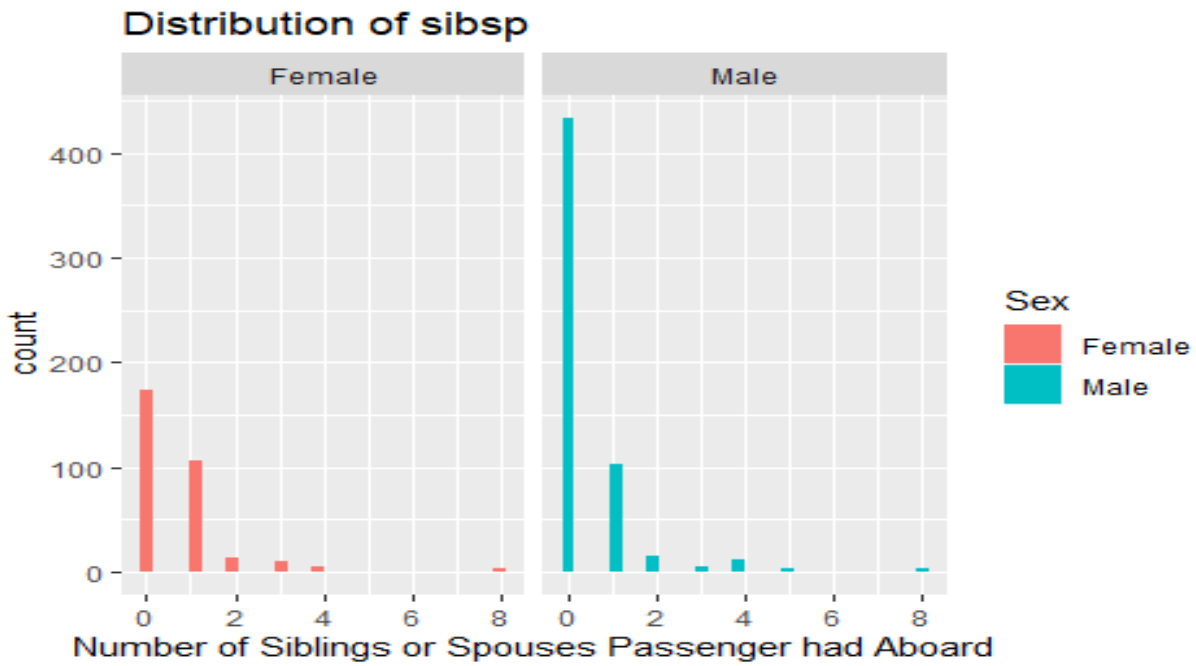
3. Sex - This categorical variable denotes the sex of the passenger. Here, the possible genders are male and female. Below is a graph that shows the distribution of the explanatory variable. As you can see there are significantly more male passengers than female passengers.

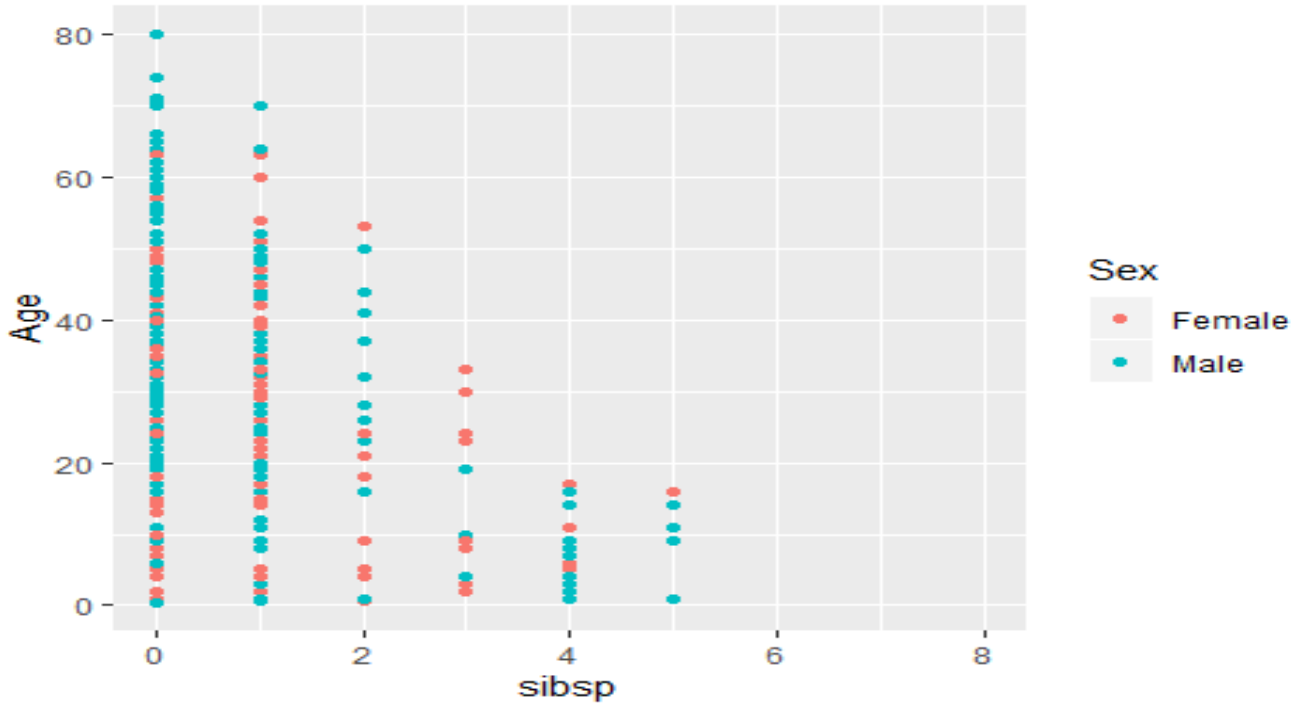


4. Age - This variable measures the age of the passenger. Below is a graph that shows the distribution of the explanatory variable. As you can see the distribution of age is roughly normal.

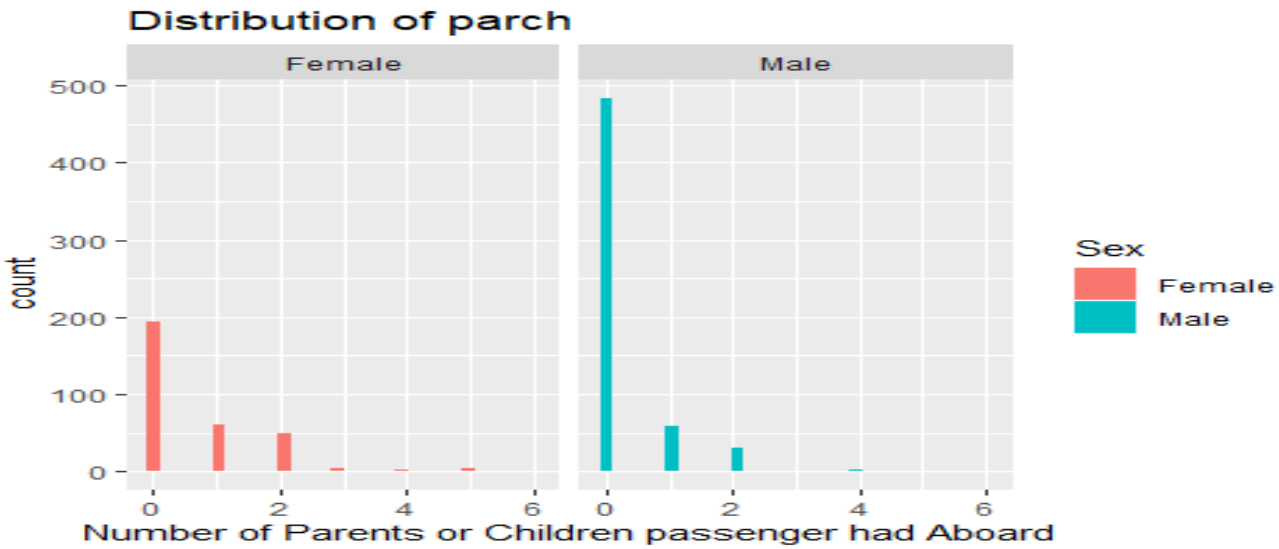


5. sibsp - This variable measures the number of siblings or spouses a passenger had aboard. Below is a graph that shows the distribution of the explanatory. As you can see this variable is skewed right, but this distribution makes sense given that a large amount of the passengers on board were young men less likely to have spouses or their siblings traveling with them.

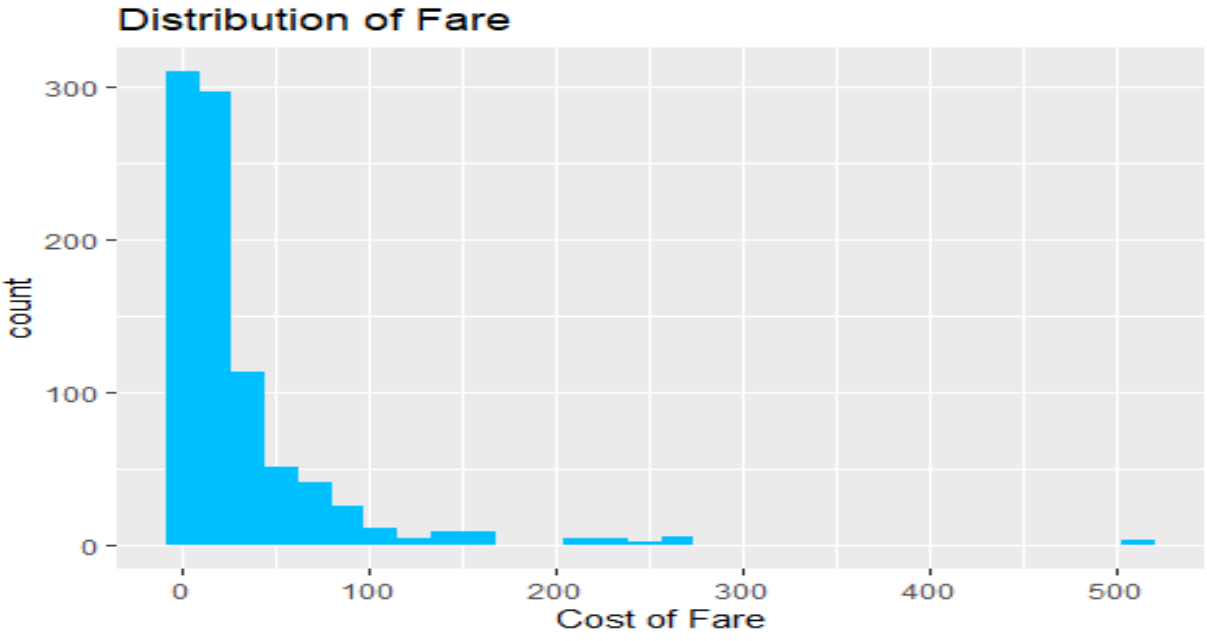




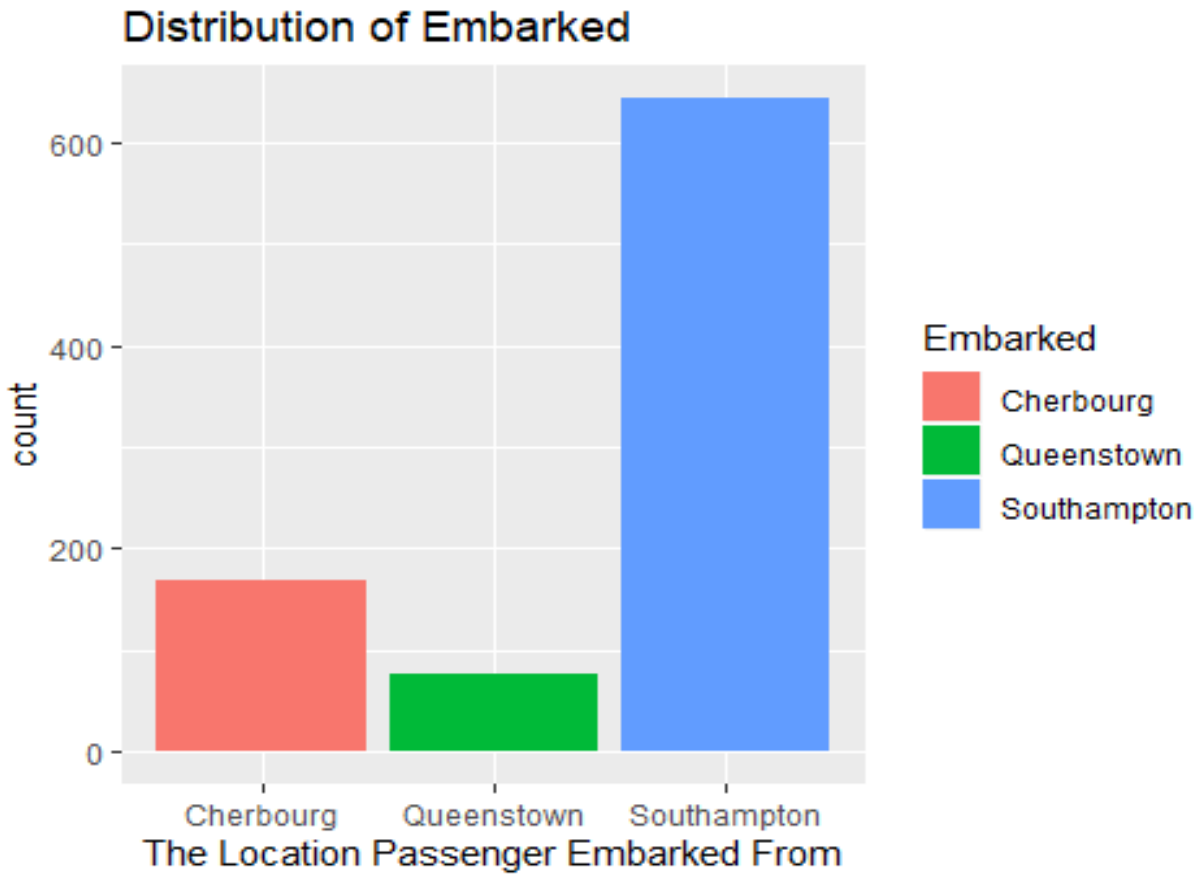
6. parch - This variable measures the number of parents or children the passenger had aboard. Below is a graph that shows the distribution of the explanatory variable. This distribution of this variable is skewed right, but again this distribution makes sense because much of the ship's population was young men who are less likely to have kids or their parents traveling with them.



7. Fare - This variable measures the cost of the fare. Below is a graph that shows the distribution of the explanatory variable. I have applied the log() transformation to this variable (see Exploratory Data Analysis).

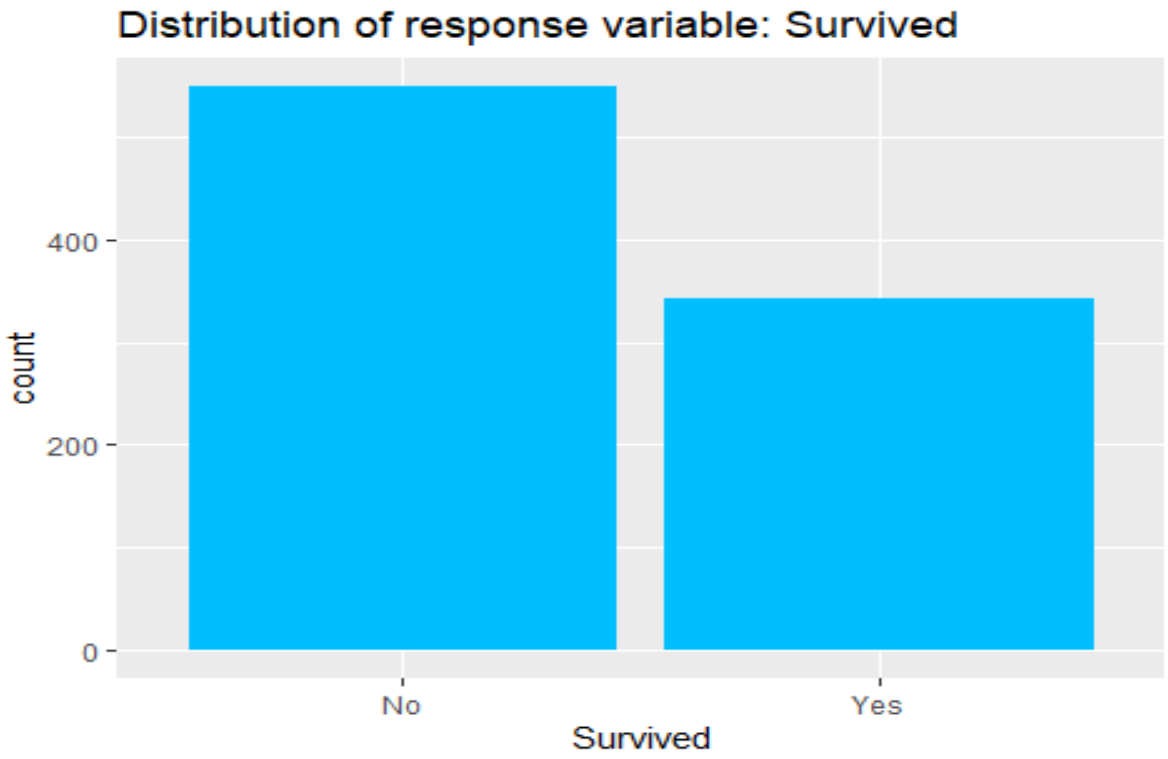


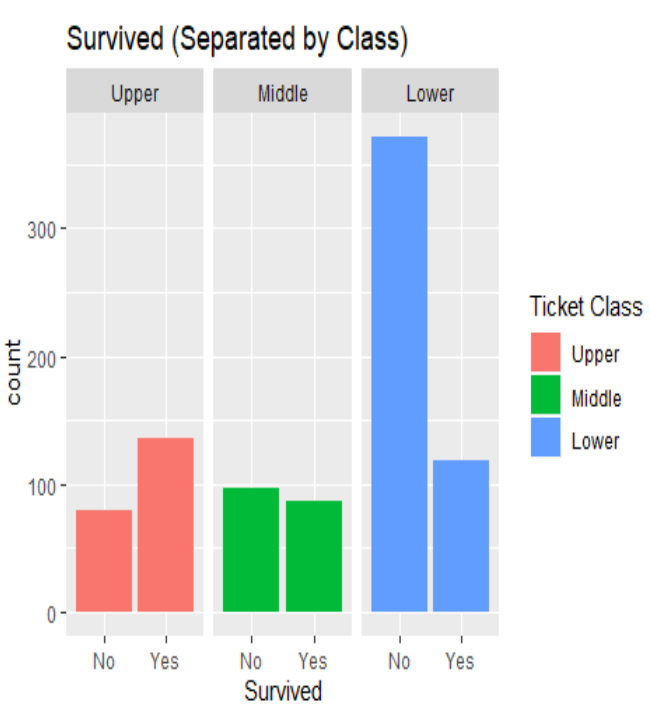
8. Embarked - This categorical variable denotes where the passenger embarked from. The three possible locations to embark are Cherbourg, Queenstown, and Southampton. Below is a graph that shows the distribution of the explanatory variable. As you can see the most passengers embarked the Titanic at Southampton.

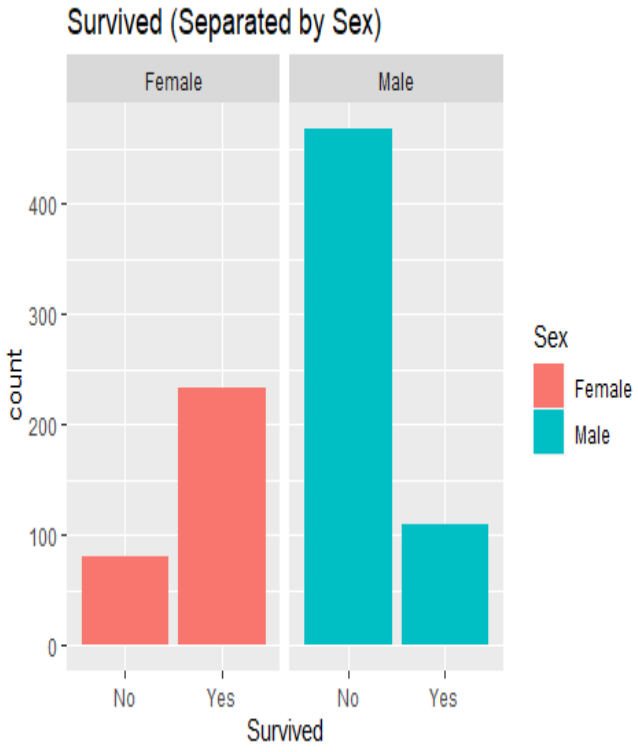


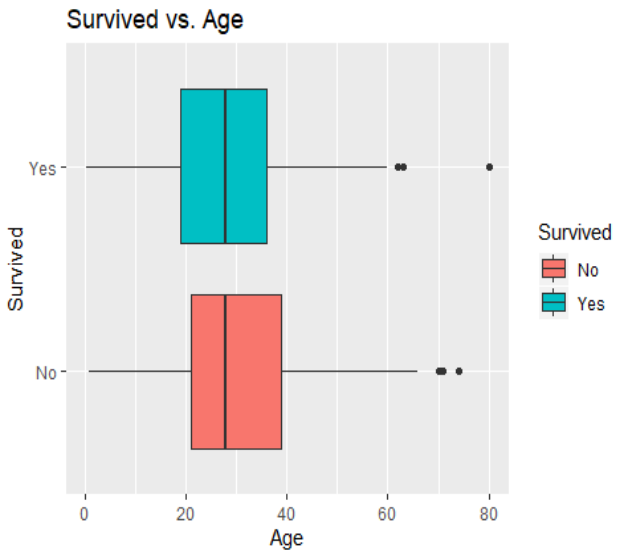
Exploratory Data Analysis: This first graph shown is a bar graph that shows the distribution of the response variable: Survived. Survived is a categorical variable that indicates whether a passenger lived or died. The second set of graphs, which uses the original data set, are the graphs that show each possible explanatory variable's relationship to the response variable (Survived). As you can see *Fare* needs to be transformed (see below). A few things that are worth noting is the graph of "Survived (Separated by Class)" and "Survived (Separated by Sex)." Something to note on the "Survived (Separated by Class)" graph is that significantly more lower-class people died rather than survived, compared to higher class passengers where more survived rather than died. With the "Survived (Separated by Sex)" graph it should be noted that more men

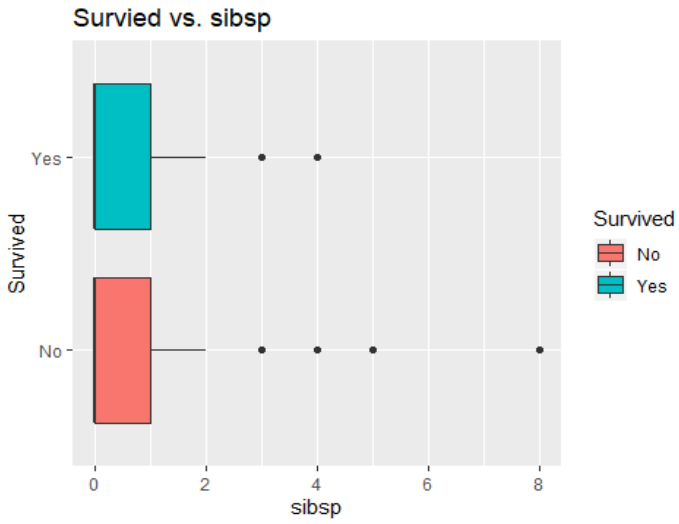
died than survived. However, more women survived than died. This points to societal views on class and gender that will be addressed in the conclusion section.

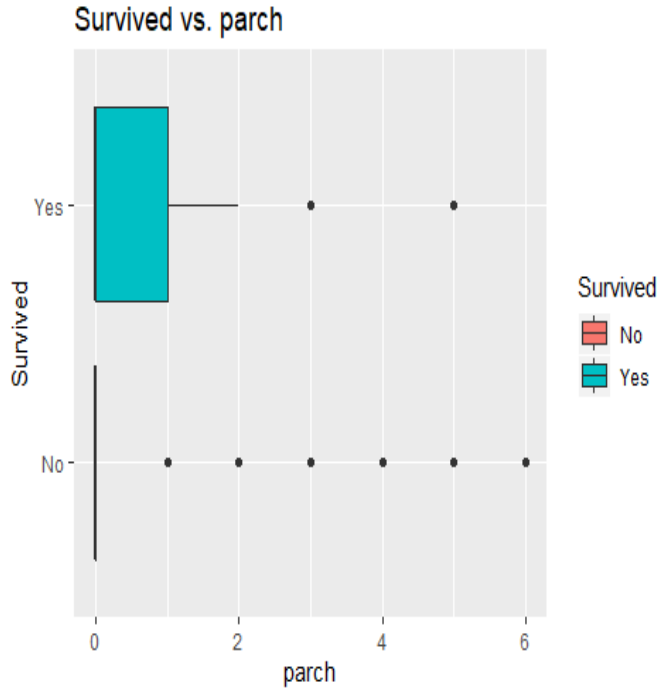


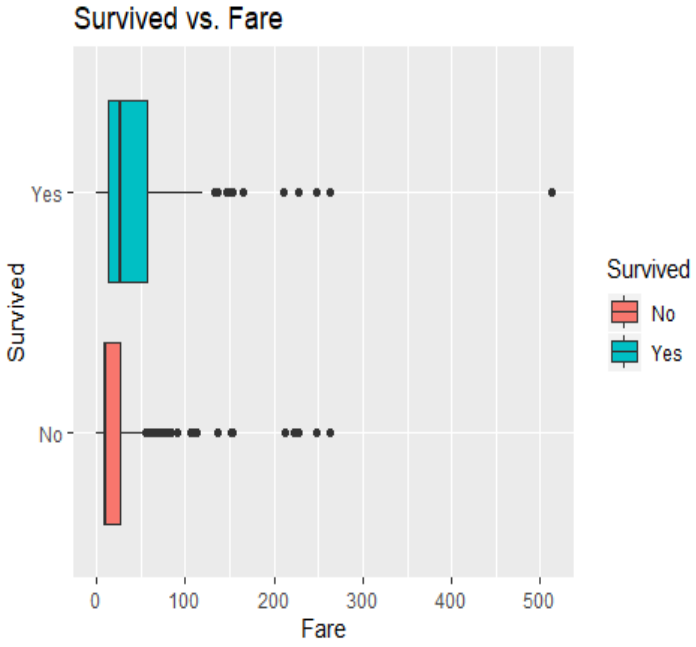


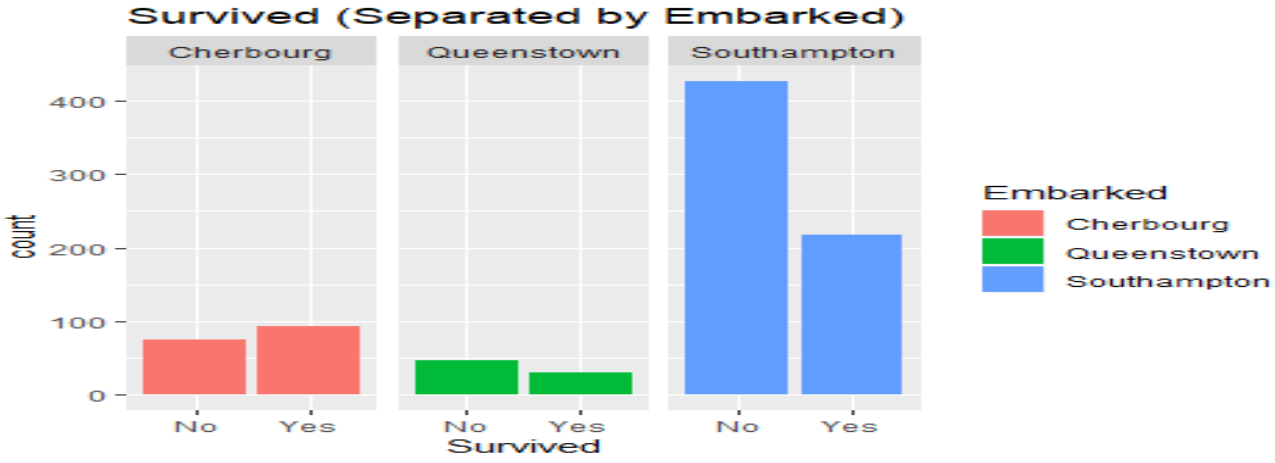



Survived vs. Age
Survived
Yes
No
0
20
40
60
80
Age
Survived
No
Yes

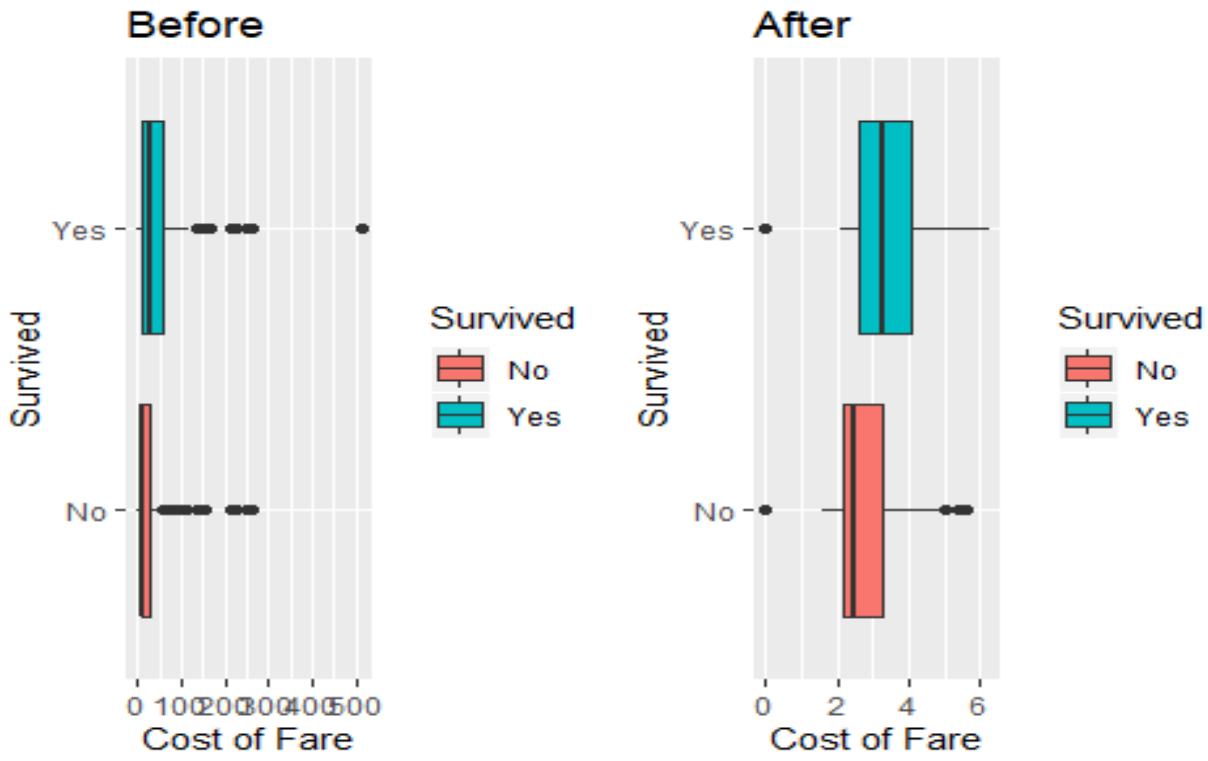

Survied vs. sibsp
Survived
Yes
No
0
2
4
6
8
sibsp
Survived
No
Yes


Survived vs. parch
Survived
Yes
No
0
2
4
6
parch
Survived
No
Yes


Survived vs. Fare
Survived
Yes
No
0
100
200
300
400
500
Fare
Survived
No
Yes


Survived (Separated by Embarked)
Cherbourg
Queenstown
Southampton
count
400
300
200
100
0
No
Yes
No
Yes
No
Yes
Survived
Embarked
Cherbourg
Queenstown
Southampton

Below you can see the analysis done on *Fare* to see what transformation is necessary. The transformation used was log(). This transformation was used because the distribution of fare covered several orders of magnitude. As you can see in the before graph, most people paid close to zero for their fare, while a few passengers paid over 500 for their fare. In the graph titled after, you can see that after the transformation fare is more evenly distributed.



Results: I have found two interaction terms that are statistically significant. An interaction term is a variable that is a function of the current explanatory variables. An interaction term is appropriate inside a model if one explanatory variable is dependent on another explanatory variable.

```
##
## Call:
## glm(formula = Survived ~ Pclass * Sex, family = binomial, data = Titanic2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6248  -0.5853  -0.5395   0.4056   1.9996
```

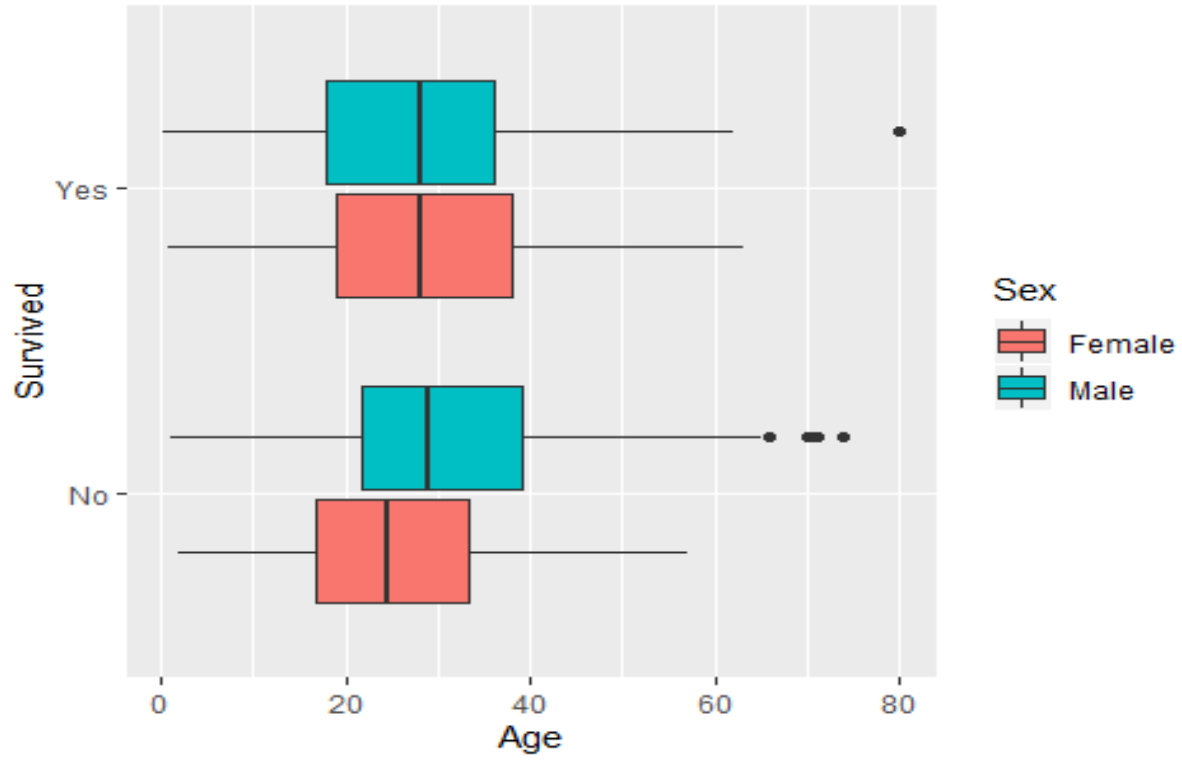
```
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)          3.4122     0.5868   5.815 6.06e-09 ***
## PclassMiddle        -0.9555     0.7248  -1.318  0.18737
## PclassLower         -3.4122     0.6100  -5.594 2.22e-08 ***
## SexMale             -3.9494     0.6161  -6.411 1.45e-10 ***
## PclassMiddle:SexMale  -0.1850     0.7939  -0.233  0.81575
## PclassLower:SexMale    2.0958     0.6572   3.189  0.00143 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  798.1  on 885  degrees of freedom
## AIC: 810.1
##
## Number of Fisher Scoring iterations: 6
```

The above summary shows the coefficients of the model that shows the interaction term between Pclass and Sex. Given that both a Pclass and Sex are factor variables, I will use a table to best illustrate the significance of the interaction variable. The table below shows the decreasing odds that a passenger survived. This makes sense because it is well documented that higher class passengers, and upper-class women in particular, were evacuated first. The horizontal columns show the passenger's class, and the vertical columns show the passengers sex. As you can see from this table, the odds that a passenger survived decreases as the ticket class lowers. Also, the effect of ticket class on the two genders is different.

Table 1

	Upper	Middle	Lower
Male	98.07337%	99.38414%	99.48347%
Female		61.53802%	96.70314%

For my second interaction term I found that there was an increased effect on a whether a passenger survived or not when both their age and sex was taken into account. For example, the older a male passenger was the higher chance he had of dying. You can see that in the graph of the male passengers that the males that died were older than the males who lived.

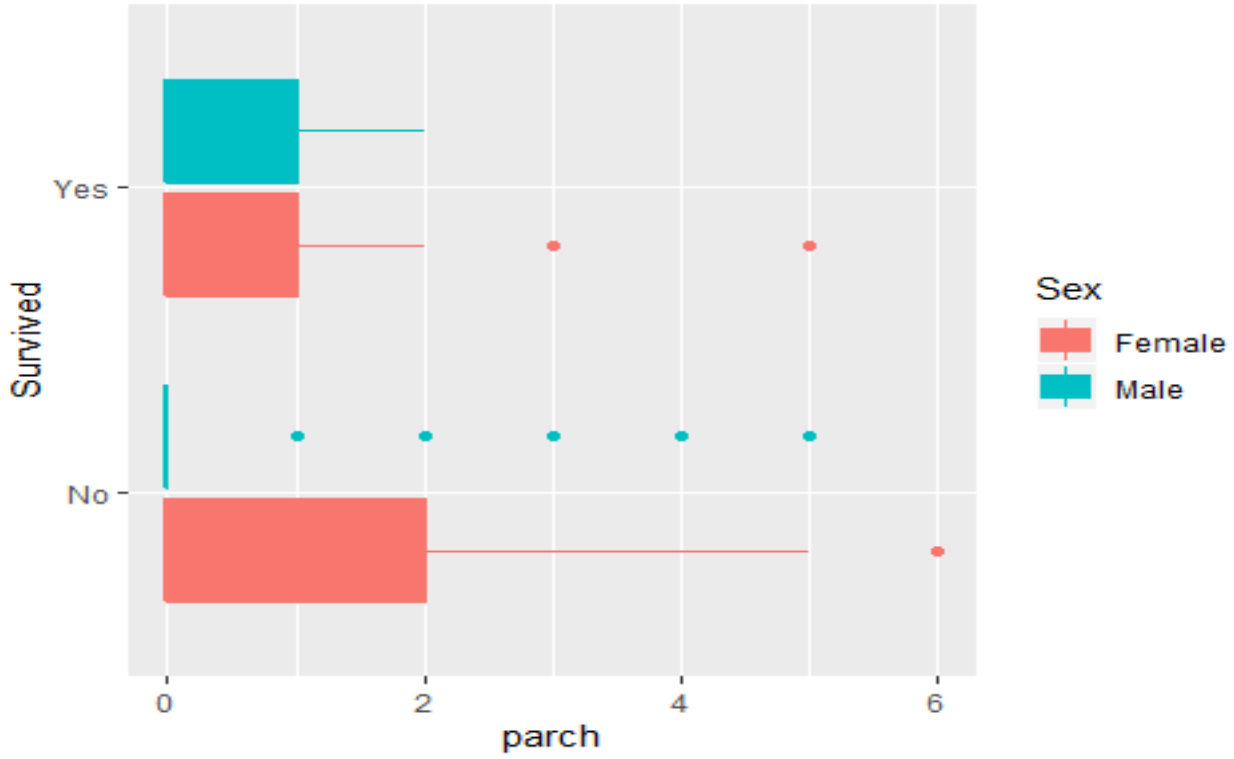



```
##
## Call:
## glm(formula = Survived ~ Sex * Age, family = binomial, data = Titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9401  -0.7136  -0.5883   0.7626   2.2455
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.59380    0.31032   1.913  0.05569 .
## SexMale     -1.31775    0.40842  -3.226  0.00125 **
## Age          0.01970    0.01057   1.863  0.06240 .
## SexMale:Age -0.04112    0.01355  -3.034  0.00241 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 740.40  on 710  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 748.4
##
## Number of Fisher Scoring iterations: 4
```

The summary above shows the coefficients of the model used to analyze the significance of the interaction term between Sex and Age. When the passenger is male an increase in age of 1 year will result in their odds of surviving decreasing by 2.119222%. This is compared to a female passenger where every 1 year increase in age results in their odds of surviving increasing by 1.989533%.

For my third interaction term I found that there was increased effect on whether a passenger survived or not if the passenger was male and had kids or parents aboard. You can see from the graph that the males that survived had more kids or parents on board than the males that died.



```
##
## Call:
## glm(formula = Survived ~ parch * Sex, family = binomial, data = Titanic2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7972  -0.6204  -0.6204   0.6660   1.8668
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.3931     0.1646   8.462  < 2e-16 ***
## parch          -0.4558     0.1234  -3.694 0.000221 ***
## SexMale        -2.9432     0.2017 -14.589  < 2e-16 ***
## parch:SexMale   0.7949     0.1949   4.078 4.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  898.68  on 887  degrees of freedom
## AIC: 906.68
##
## Number of Fisher Scoring iterations: 4
```

The summary above shows the coefficients of the model made to analyze the significance of the interaction term between Sex and parch. When the passenger is male an increase of 1 in the number of parents or children brought aboard increased the odds of survival by 40.36837%. This is compared to a female passenger where every increase of 1 in the number of parents or children brought aboard resulted in the odds of survival decreasing by 36.60594%.

I used to backwards stepwise model selection to build my model, which is an algorithm that finds the best model for predicting the average evaluation score. Having the step function be in the backwards direction means that the model will start with all the variables and will remove variables as the algorithm sees fit. It compares the models it produces using the AIC scores from each model; the lower the score the better.

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + sibsp + parch +
```

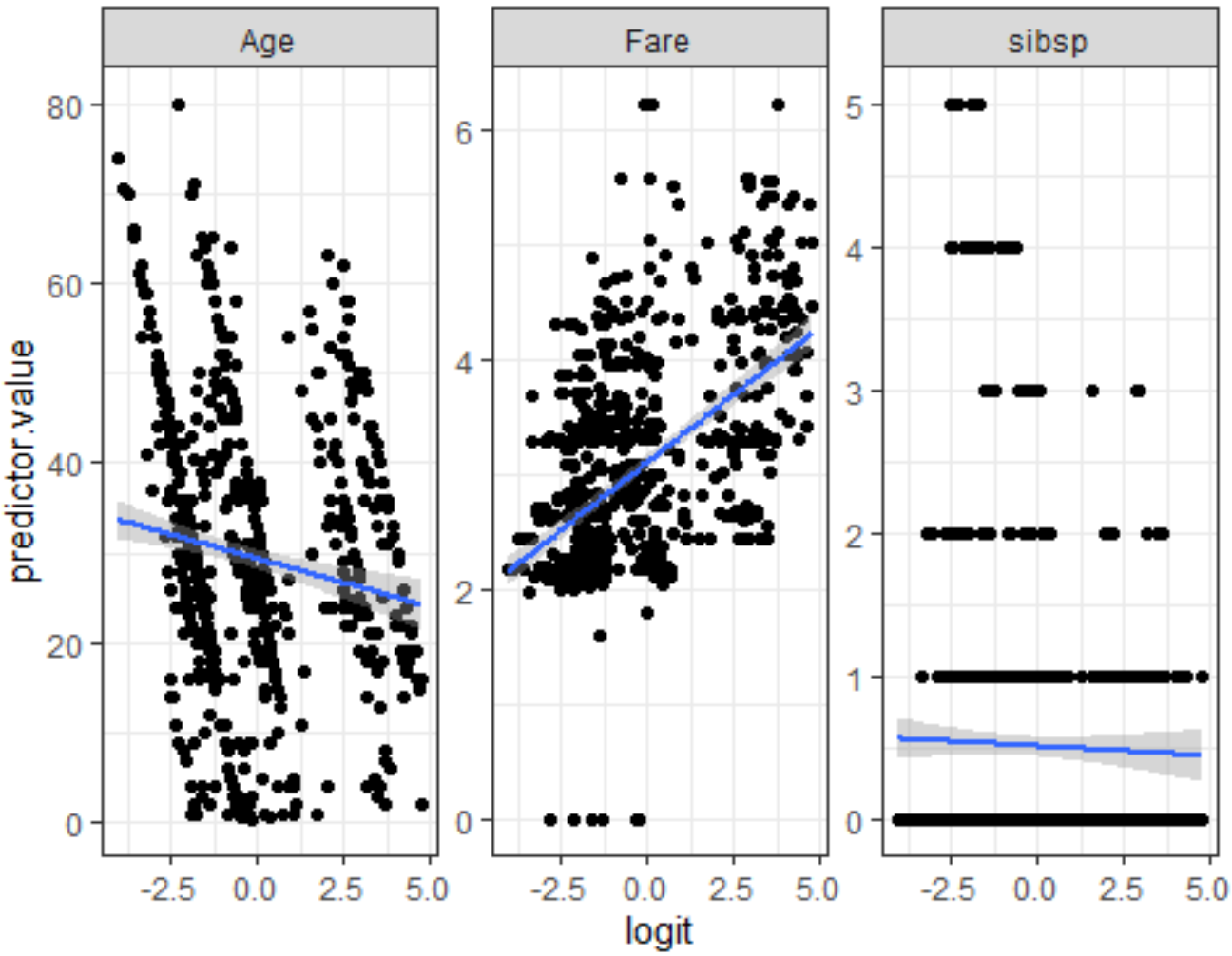
```
##     Pclass:Sex + Sex:parch, family = binomial, data = na.omit(Titanic2))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0833  -0.6518  -0.4402   0.3918   2.6093
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          5.474605   0.723755   7.564 3.90e-14 ***
## PclassMiddle        -1.195058   0.739575  -1.616  0.10612
## PclassLower         -4.153615   0.660557  -6.288 3.21e-10 ***
## SexMale             -3.859745   0.641277  -6.019 1.76e-09 ***
## Age                 -0.049187   0.009069  -5.423 5.85e-08 ***
## sibsp               -0.401816   0.126557  -3.175  0.00150 **
## parch               -0.114795   0.162753  -0.705  0.48060
## PclassMiddle:SexMale -0.740724   0.823229  -0.900  0.36824
## PclassLower:SexMale   2.130643   0.693499   3.072  0.00212 **
## SexMale:parch         0.410359   0.243206   1.687  0.09155 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 599.47  on 702  degrees of freedom
## AIC: 619.47
##
## Number of Fisher Scoring iterations: 6

##               GVIF Df GVIF^(1/(2*Df))
## Pclass     23.999084  2       2.213343
## Sex         8.577077  1       2.928665
## Age         1.600841  1       1.265243
## sibsp       1.318573  1       1.148291
## parch       2.003627  1       1.415495
## Pclass:Sex 33.579755  2       2.407240
## Sex:parch   1.986181  1       1.409319
```

This is the model created from the stepwise algorithm. As you can see there are 5 explanatory variables. I have conducted a multicollinearity test, and as you can see there is no multicollinearity in this model given that the values from that test are less than 5, which is the generally accepted threshold. The estimate column shows how much the log(odds) would increase or decrease if the other variables are held constant. I have changed the estimates from

log(odds) to odds to make them easier to interpret. Below is an explanation of each of the coefficient:

Variable	Effect
PclassMiddle & PclassLower	Refer to table 1
Sexmale	Refer to table 1,2 and 3
Age	Refer to table 2
sibsp	For every increase in 1 sibsp the odds of surviving decreased by 33.08962%.
parch	Refer to table 3
PclassMiddle: SexMale	Refer to table 1
PclassLower: SexMale	Refer to table 1
SexMale: parch	Refer to table 3

When you make a logistic regression model there are certain assumptions that have to be made. The first assumption is that the response variable is binary; as you can see, survived is binary given that you can either live or die. The second assumption is that the variables are independent, which in this data set they are. Third, there can be no multicollinearity, which as you can see from the above summary there is not. The fourth assumption is that there is a linear relationship between the explanatory variables and the log of the odds, also known as logit; as you can you can see from the three graphs below there is a clear linear relationship between the log(odds) and the explanatory variables. The sample size for logistic regression needs to be sufficiently large. The equation used to test sample size is $n = 50 \cdot v/p$. In order for this model to be suffiently large enough. In this case the value you get from that equation is 619, which is less than the 891 observations in the data set making this model sufficiently large.



Discussion: Many of the problems that plague statistical studies do apply in this study. For example, the data was clean, except for a few minor fixes, and only one variable needed to be transformed. One problem I would point out is that there is no variable that measures where the passenger's room was located on the ship. Given that the ship hit the iceberg at close to midnight most people would have been in bed. A variable that measures where someone's room is located could estimate how long it would take a passenger to get to the life boats. Given that some lower-class passengers survived it would be useful to know if they survived only because they were closer to the life boats. What is alarming about this study is that gender played a large part in if the passenger would survive. For example, if the passenger was male the odds he survived decreased by 73.22633%. Every 1 year older a man got the odds he survived decreased by 2.119222%, where as a woman's odds increased by 1.989533% for every 1 year increase in

her age. The reason that a passenger survives should have nothing to do with that passenger's gender. Another alarming trend this study found was that there was a significant difference between the ticket classes regarding the odds that a passenger survived (This can be seen in table 1). This is very troubling because how wealthy you are should not determine whether you live or die. Everyone, no matter how much money a passenger has, deserves an equal chance of survival.

Conclusion: What happened to the Titanic was tragic, with hundreds of people dying. It is well documented that sinking of the Titanic resulted in wide changes to how large cruise ships operate. The biggest change that could have saved more people was the new order that all boats must have enough lifeboats for all passengers. The events on the Titanic and the order in which passengers were evacuated in points to several issues present in the past. How human lives were valued in the past was unfair and unjust. People would favor the lives of rich people over poor people. Along with that the lives of females were considered more important to save. These beliefs are beliefs that as a society we need to work on. Although, they are not as present today as they were back then, there is still evidence that some of these views carry over to today. It is important for everyone's lives, no matter the gender or class, to be valued the same.