Titanic Project **Garrett Peters**

1. Introduction

The Titanic was a renowned British luxury passenger liner that tragically sank during its inaugural journey from Southampton to New York City on April 15, 1912. The catastrophe occurred when the ship collided with an iceberg, leading to the loss of more than 1,500 passengers and crew members. Michelle, a member of a Titanic Research Group, commissioned us to investigate the factors influencing survival rates of the disaster. Our task included identifying variables associated with survival and examining potential patterns in survival rates based on combinations of passengers' class and gender. Additionally, Michelle requested an analysis of how age impacted survival, with a specific interest in visual representations of the data through graphs.Our study focuses on a population of 1,311 passengers identified by Michelle. To gather data, Michelle employed systematic sampling, selecting every fifth observation from the population, resulting in a dataset comprising information on 261 passengers. As systematic sampling was utilized, our study is classified as observational, allowing us to examine patterns and relationships within the collected dataset without intervening or manipulating variables.

1.1 - Research Questions

Question 1) Which combination of variables is best in predicting the survival of Titanic passengers?

Question 2) Which of the six combinations of class and sex variables are similar in survival probabilities practically speaking?

Question 3) Graphically, what is the relationship between age and probability of survival, given the significant variables from Question 1?

1.2 - Variables

Variables	Type	Description	Levels and Ranges
Name	Categorical	Name of the passenger	Unique to each observation
Age	Quantitative	Age of passenger	2 to 74
AdultOrChild	Categorical	If the passenger was 18 or older	Adult or Child
Class	Categorical	The class of the passenger	1st, 2nd, and 3rd

Survived	Categorical	Whether or not the person survived	1 = Yes 0 = No
Port	Categorical	The port that the passenger boarded from	Southampton, Cherbourg, and Queenstown
StaffNotes	Categorical	Whether the passengers is identified as staff	None, Interpreter, Maid, Manservant
Sex	Categorical	Gender of the passenger	Male and Female

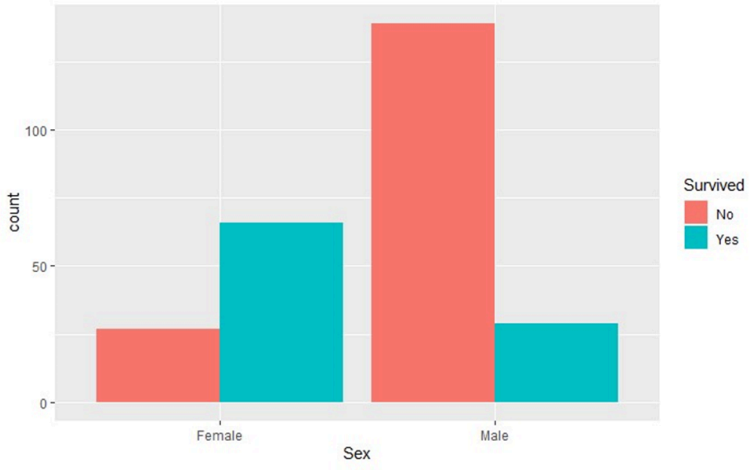
Table 1: Summary of Variables for the Titanic Study

2. Exploratory Data Analysis

	1st Class	**2nd Class**	**3rd Class**
Female - Survived	23	20	30
Female - Died	1	2	24
Male - Survived	13	4	12
Male - Died	27	29	83

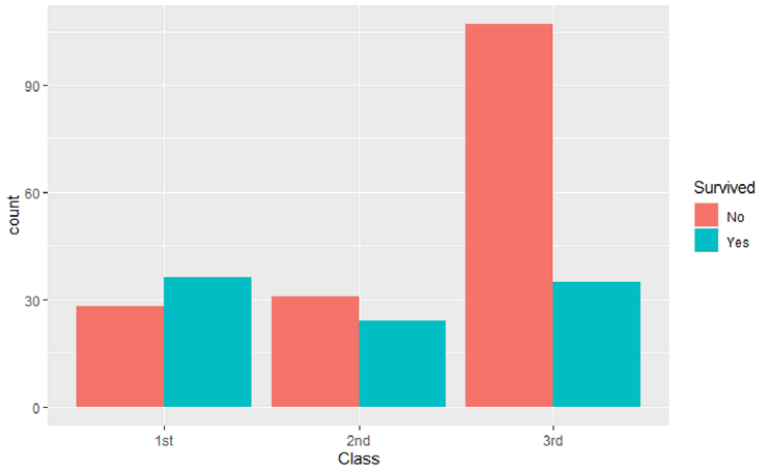
Table 2: Count of Sex and Class by Survival Status

Figure 1: Bar Chart showing frequency of survival based on Sex



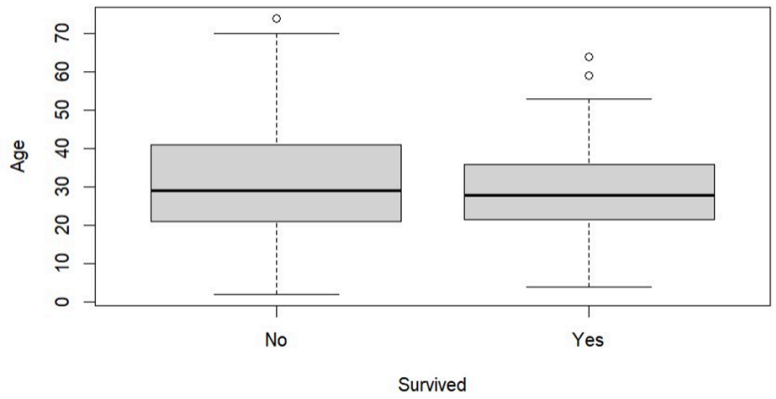
This graph is a bar chart that shows the count of males and females based on whether or not they survived. From the chart, it seems that about 65 females survived and 25 females died (72.22% survival rate. On the other hand, about 30 males survived and 165 males died (15.38% survival rate). This drastic difference is consistent with the famous "women and children first" strategy employed in boarding lifeboats. Clearly, females were much more likely to survive, in general, than males.

Figure 2: Bar Chart showing frequency of survival based on Class



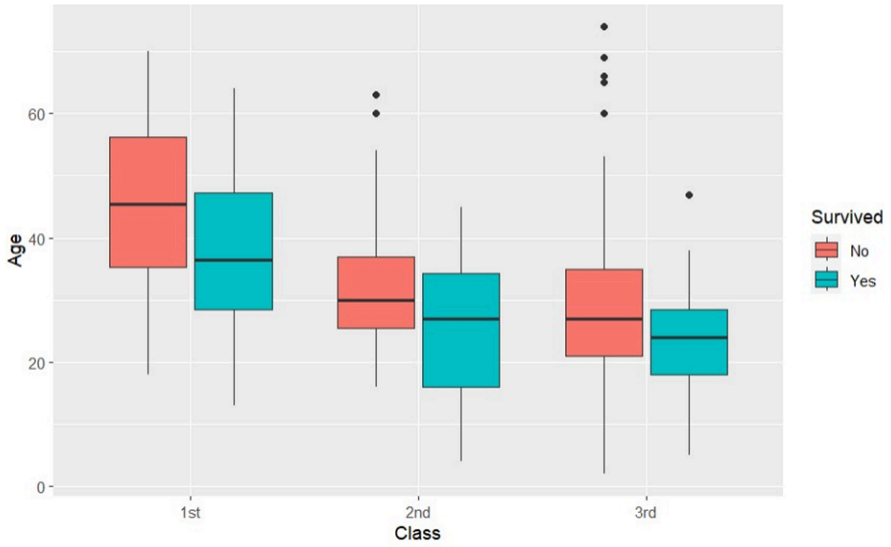
This graph is a bar chart for each of the three classes of passengers on the ship by whether or not they survived. From the graph, it seems that the survival rate for first class was slightly above 50% and the survival rate for second class was slightly below 50%. However, only about 35 third class passengers survived compared to about 110 who died (24.14% survival rate). This is also consistent with what is known about the Titanic in how the third class passengers were barred from certain areas after the ship hit the iceberg. The other two classes had relatively similar survival rates while the third class survival rate was much less.

Figure 3: Boxplot of Age by Survival Status



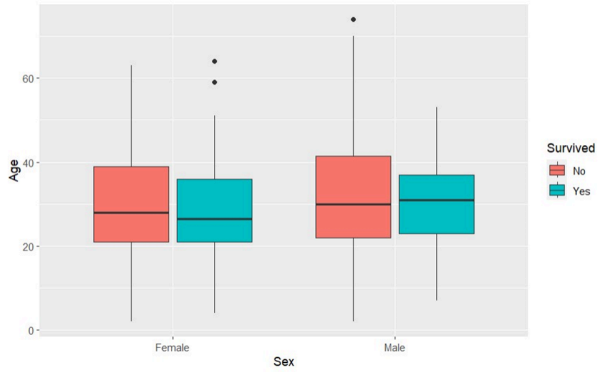
This graph shows side-by-side boxplots for age by whether or not a person survived. A boxplot is a visual representation of the five-number summary mentioned earlier where the black line in the middle is the median, the edges of the box are the first and third quartiles, and the top and bottom of the lines are the minimum and maximum. From this plot, it seems that the mean and minimum ages were about the same for those who survived and those who died. However, the maximum age of those who died is much higher than that of those who survived save a few outliers. This would suggest that above the age of 50 or so, the survival rate for passengers was much lower than the overall rate.

Figure 4: Boxplot of Age by Class and Survival Status



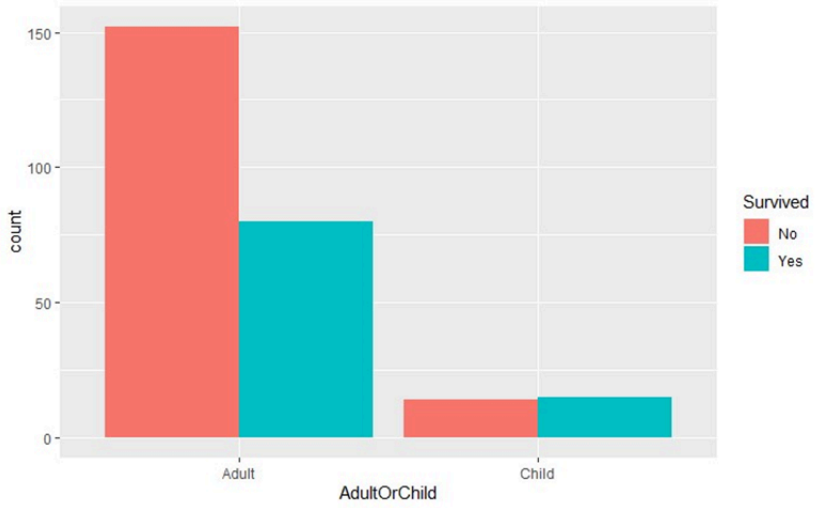
This graph shows side-by-side boxplots for age for each class by whether or not a person survived. For each class, especially first, the mean age of those who survived was much lower than those who didn't. This also follows the "women and children first strategy. There are some outliers on the upper end for second and third class. Perhaps the most interesting piece of this is the fact that there are almost no young people in first class. The mean ages for both those who survived and those who didn't in second and third class are far younger than the means in first class.

Figure 5: Boxplot of Age by Sex and Survival Status



This graph shows side-by-side boxplots for age for each sex by whether or not a person survived. The mean ages for both sexes are about the same regardless of survival status (all around 30 years old). This shows that when broken up by sex, the younger passengers are comparably a lot less likely to survive than they are in general. Also, while the male survival rate is much lower in general, their mean age is about the same.

Figure 6: Bar Chart showing frequency of survival based on being an Adult or Child



This graph is a bar chart for adults and children based on whether or not they survived. While there were far more adults sampled, there are still enough children to make inferences about survival rates. It seems that children have about a 50% survival rate while there were only about 80 adults who survived compared to about 150 who died (34.78% survival rate). This again is consistent with “women and children first”. In general, passengers under 18 years old had a far better chance of survival than passengers 18 years old and above.

3. Statistical Analysis

3.1 - Research Question 1

We opted to analyze the dataset using RStudio to address the primary inquiry: "Which combination of variables most accurately predicts the survival of Titanic passengers?" We utilized a general linear model (GLM) due to the binary nature of the response variable (1 = survived, 0 = did not survive). This GLM facilitates the identification of the most influential variables. Significance is typically established when a variable's observed p-value falls below 0.05, indicating the probability of achieving the observed outcome. Following an initial full model encompassing all variables except "Name" and their interactions, we identified Age, Sex, Class, Port, and the Class-Sex interaction as potentially significant predictors based on their respective p-values. Subsequently, our refined model, excluding the non-significant Port variable, exhibited enhanced performance. A subsequent model comprising Age, Sex, Class, and the Class-Sex interaction emerged as the most optimal, with all variables demonstrating significance. The lower AIC value in this model indicates its superior fit compared to other models, as AIC (Akaike Information Criterion) quantifies the balance between model complexity and goodness of fit, where lower values signify better model performance. Table 3 to the right shows significant variables and their corresponding p-values. It is important to note that all model assumptions were satisfied for this final model, these assumptions include normality, constant variance and independence. These plots will be listed in the Appendix below

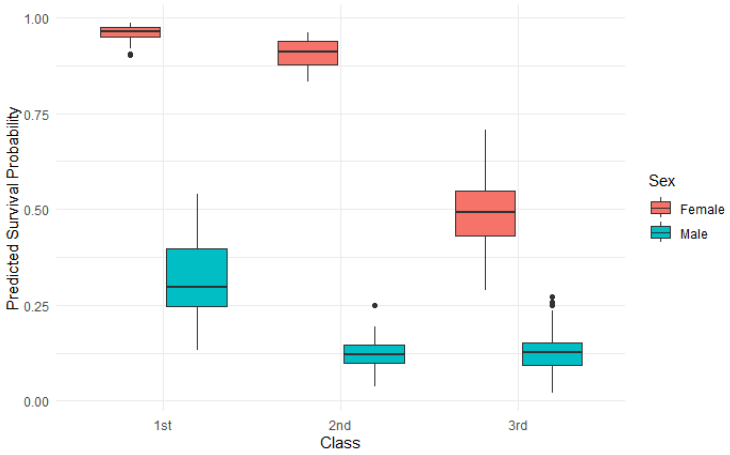
Coefficients:	P-Value
(Intercept)	0.000***
Age	0.007**
Class2nd	0.289
Class3rd	0.001***
SexMale	0.000***
Class2nd:Sex Male	0.834
Class3rd:Sex Male	0.084

Table 3: Summary of Final Model

3.2 - Research Question 2

In addressing question 2 - "Which of the six combinations of class and sex variables are similar in survival probabilities practically speaking" our approach involves determining the predicted probability of survival for each passenger, as outlined and detailed in the Appendix below. By incorporating these probabilities into the original dataset, we can visualize and compare the similarities in survival probabilities across different classes and sexes. We chose to visualize the data through the creation of boxplots, illustrating each gender's distribution within each class, as depicted in Figure 7.

Figure 7: Boxplot of gender-class survival probability



As observed, three pairs exhibit similar probabilities of survival. The first pair consists of first and second-class females, with both groups surpassing a 75% chance of survival. The second pair, comprising third-class females and first-class males, demonstrated survival rates ranging from approximately 15% to 65%. Although third-class females exhibited a slightly higher average survival rate compared to males, these groups exhibited the closest resemblance. Lastly, the second and third-class males exhibited the lowest survival rates, averaging around 12.5%.

3.3 - Research Question 3

We determined that creating a scatterplot would be the most effective approach to address research question 3: "Graphically, what is the relationship between age and the probability of survival, considering Age, Sex, and Class?" Figure 8 below illustrates this analysis.

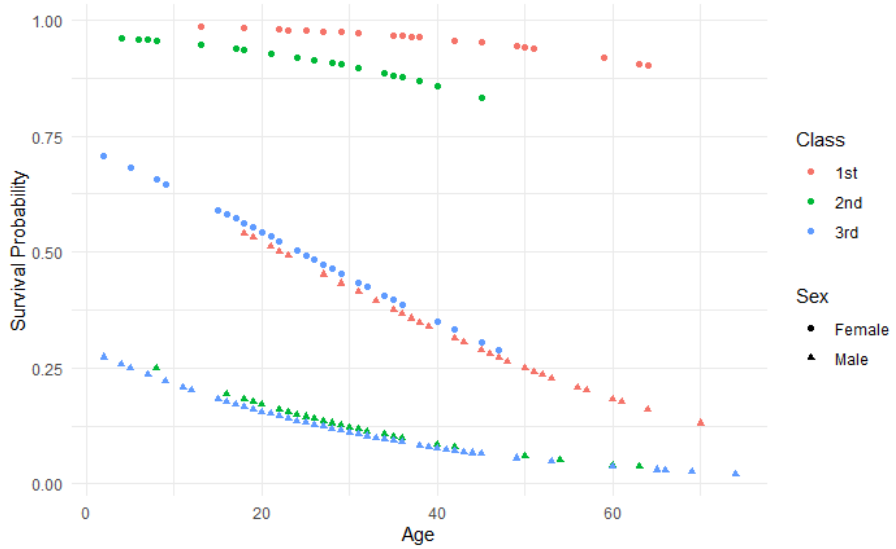


Figure 8: Scatterplot of Age and Survival Probability based on Sex and Class

In this graph, the probability of survival is plotted on the y-axis against age on the x-axis, with passengers represented as either circles or triangles (circles denote females, triangles denote males). These shapes are color-coded: red for first class, green for second class, and blue for third class. Across all six combinations of class and gender, younger individuals tend to have higher survival probabilities, yet disparities exist between these combinations. Females generally exhibit higher survival rates compared to males. Within each gender, a hierarchical pattern emerges: first-class females are more likely to survive than third-class females, along with males mirroring a similar trend. This graph further aids in visualizing research question 2, corroborating our previous analysis: first and second class females display notably similar, higher survival probabilities; third class females and first class males exhibit roughly equal probabilities; and second class and third class males demonstrate the lowest survival probabilities.

4. Recommendations

Question 1) Which combination of variables is best in predicting the survival of Titanic passengers?

- The combination of Sex, Class, Survival and the Sex-Class interaction best predict the survival of Titanic passengers. Generally, females in second and even more so in first class have a better survival rate along with young people

Question 2) Which of the six combinations of class and sex variables are similar in survival probabilities practically speaking?

- First and Second Class Females exhibit a similar probability
- Third Class Females and First Class Males exhibit a similar probability
- Second and Third Class Males exhibit a similar probability

Question 3) Graphically, what is the relationship between age and probability of survival, given the significant variables from Question 1?

- Generally the younger you are, the higher probability of survival, females also exhibit higher probability, along with being a superior class.

5. Resources

For resources related to GLM models please refer to:

https://online.stat.psu.edu/stat504/lesson/6/6.1

For resources pertaining to the creation of these graphs, please refer to:

https://ggplot2.tidyverse.org/

For resources related to R Studio, please refer to:

https://guides.libraries.psu.edu/c.php?g=894031&p=6430365

6. Additional Considerations

Assumptions and Outliers

We utilized binary logistic regression to address the first research question, examining the relative probabilities of survival. The model's assumptions were reasonably met, as detailed in the appendix. Normality assumption was assessed through a Q-Q plot, where a good plot demonstrates a linear relationship between observed and theoretical quantiles, with points clustering around the diagonal reference line. Our Q-Q plot largely adheres to this pattern, albeit with a few outliers, including observations 189 (Mr. Ali Lam), 89 (Miss Marta Hiltunen), and 57 (Mrs. Rosalie Ida Straus). Additionally, the residuals vs. fitted plot, while generally conforming to the expected pattern, exhibits slight curvature, primarily driven by observation 89 (Miss Marta Hiltunen). The Cook's plot focuses on identifying influential data points in linear regression, with observations 89 (Miss Marta Hiltunen), 57 (Mrs. Rosalie Ida Straus), and 82 (Miss Annie Clemmer Funk) standing out. Given the limited number of outliers in the dataset, their impact on the predictive variables for survivability is expected to be minimal.

Extrapolation in This Study

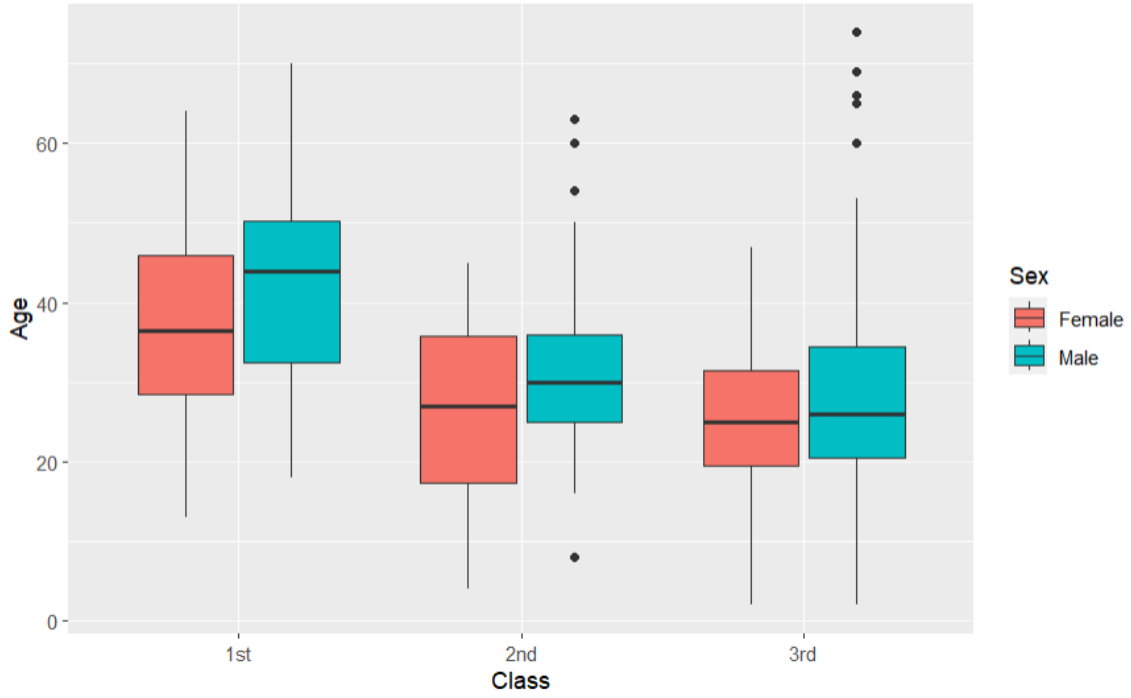
It is also important to remember that because this was an observational study, as all Titanic studies would be, it is hard to make any concrete assumptions. The Titanic sinking was a one time event over 100 years ago, and so this analysis should only be used to draw conclusions about the survival rate of that particular event.

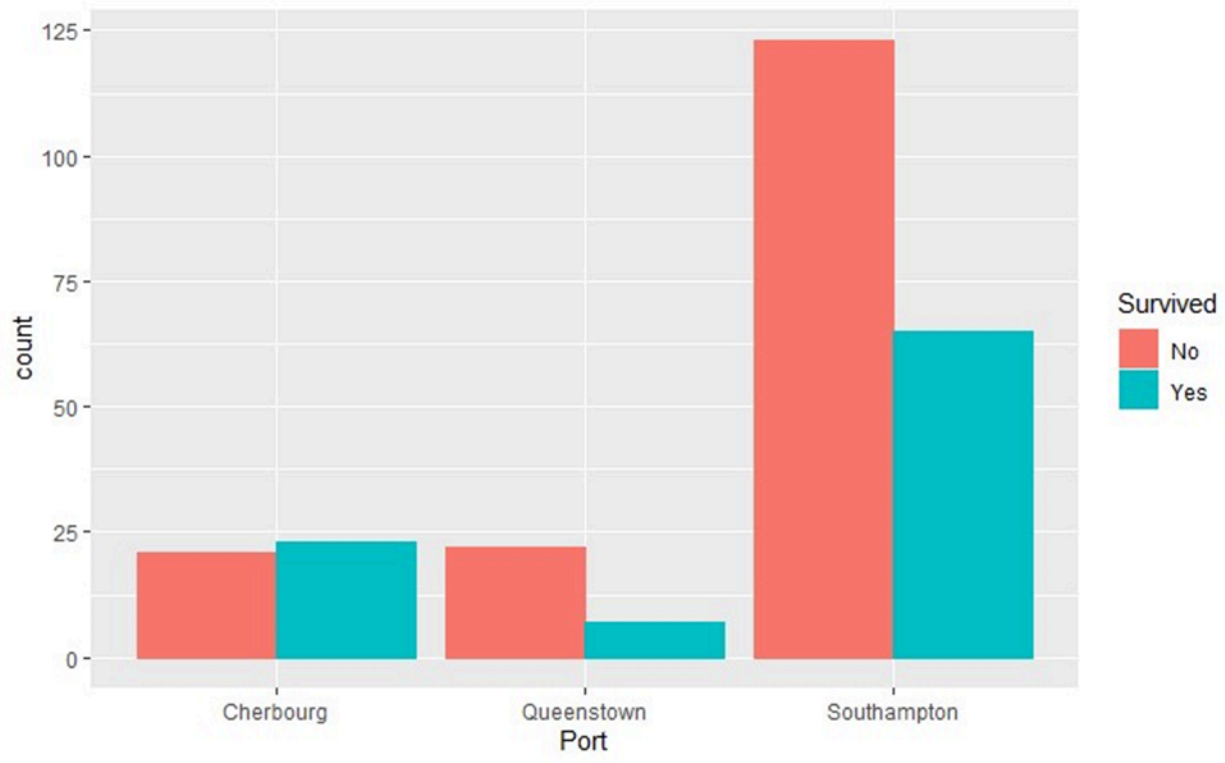
7. Appendix

Further EDA:

In the process of exploring the data we looked at the age of all passengers in a box plot categorized by class and sex. In general, male in each class tend to be older than females, and first class passengers tend to be older than second and third class passengers. Also, we looked at a bar chart of people who survived or not categorized by what port they boarded at. Although the survival rate of those who boarded at Cherbourg seem to have a higher survival rate than the other two ports, the Port variable was not significant in the binary regression analysis, so this information was not significant to our overall analysis. Below are those graphs and the corresponding code:

```{r}
library(ggplot2)
ggplot(Titanic, aes(x=Class, y=Age, fill=Sex)) + geom_boxplot()
```





R-Studio Code for in report EDA

```
#Figure 3 Boxplot of Age by Survival Status
boxplot(Titanic$Age ~ Titanic$Survived, data=Titanic, main="Boxplots",  ylab = "Age", xlab =
"Survived", title = "fmds")
```

```
#Figure 4: Boxplot of Age by Class and Survival Status
ggplot(Titanic, aes(x=Class, y=Age, fill=Survived)) + geom_boxplot()
```

```
#Figure 5: Boxplot of Age By Sex and Survival Status
ggplot(Titanic, aes(x=Sex, y=Age, fill=Survived)) + geom_boxplot()
```

Binary Regression Code for Question 1

```
```{r}
model1 = glm(Survived~ Age+Class+AdultorChild+Port+Sex+Age:Class+Age:AdultorChild+Age:Port+Age:Sex+C
lass:AdultorChild+Class:Sex+AdultorChild:Port+AdultorChild:Sex, data=Titanic, family="binomial")
  summary(model1)
```
```

```
Call:
glm(formula = Survived ~ Age + Class + AdultorChild + Port +
    Sex + Age:Class + Age:AdultorChild + Age:Port + Age:Sex +
    Class:AdultorChild + Class:Sex + AdultorChild:Port + AdultorChild:Sex,
    family = "binomial", data = Titanic)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     7.105e+00  2.758e+00   2.577  0.00998 **
Age                            -8.851e-02  5.407e-02  -1.637  0.10164
Class2nd                       -1.688e+00  2.670e+00  -0.632  0.52733
Class3rd                       -3.765e+00  2.161e+00  -1.742  0.08145 .
AdultorChildchild               2.524e+01  2.058e+03   0.012  0.99022
PortQueenstown                 -1.208e+00  2.853e+00  -0.423  0.67196
PortSouthampton                -2.261e+00  1.527e+00  -1.481  0.13867
SexMale                        -4.848e+00  2.357e+00  -2.057  0.03972 *
Age:Class2nd                   -4.751e-03  6.202e-02  -0.077  0.93894
Age:Class3rd                   -7.849e-03  4.114e-02  -0.191  0.84869
Age:AdultorChildchild           1.047e-02  1.181e-01   0.089  0.92932
Age:PortQueenstown             -5.228e-03  9.841e-02  -0.053  0.95763
Age:PortSouthampton             5.652e-02  4.490e-02   1.259  0.20815
Age:SexMale                     1.232e-02  4.539e-02   0.271  0.78604
Class2nd:AdultorChildchild     -9.286e+00  1.455e+03  -0.006  0.99491
Class3rd:AdultorChildchild     -1.173e+01  1.455e+03  -0.008  0.99357
Class2nd:SexMale                1.081e-01  1.669e+00   0.065  0.94834
Class3rd:SexMale                2.578e+00  1.511e+00   1.706  0.08797 .
AdultorChildchild:PortQueenstown -1.346e+01  1.455e+03  -0.009  0.99262
AdultorChildchild:PortSouthampton -1.424e+01  1.455e+03  -0.010  0.99219
AdultorChildchild:SexMale       1.522e-01  1.481e+00   0.103  0.91813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.26  on 260  degrees of freedom
Residual deviance: 217.67  on 240  degrees of freedom
AIC: 259.67

Number of Fisher Scoring iterations: 14
```

```
Step:  AIC=239.95
Survived ~ Age + Class + Sex + Class:Sex

            Df Deviance    AIC
<none>           225.95 239.95
- Class:Sex  2   234.19 244.19
- Age        1   233.75 245.75

Call:  glm(formula = Survived ~ Age + Class + Sex + Class:Sex, family = "binomial",
    data = Titanic)

Coefficients:
     (Intercept)               Age          Class2nd          Class3rd           SexMale
         4.75429          -0.03955          -1.36661          -3.79277          -3.88065
Class2nd:SexMale  Class3rd:SexMale
        -0.29529           2.01799

Degrees of Freedom: 260 Total (i.e. Null);  254 Residual
Null Deviance:      342.3
Residual Deviance: 225.9        AIC: 239.9
```

```r
reducedmodel <- glm(formula = Survived ~ Age + Class + Sex + Class:Sex, family = "binomial",
    data = Titanic)
summary(reducedmodel)
```

```
Call:
glm(formula = Survived ~ Age + Class + Sex + Class:Sex, family = "binomial",
    data = Titanic)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.75429    1.21936   3.899 9.66e-05 ***
Age             -0.03955    0.01482  -2.669 0.007609 **
Class2nd        -1.36661    1.28965  -1.060 0.289290
Class3rd        -3.79277    1.10619  -3.429 0.000607 ***
SexMale         -3.88065    1.08736  -3.569 0.000359 ***
Class2nd:SexMale -0.29529    1.42438  -0.207 0.835764
Class3rd:SexMale  2.01799    1.16885   1.726 0.084263 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 342.26  on 260  degrees of freedom
Residual deviance: 225.95  on 254  degrees of freedom
AIC: 239.95

Number of Fisher Scoring iterations: 5
```

.

Graphical Analysis Code For Question 3

```r
eprobs = data.frame(probs = predict(reducedmodel, type="response"))
head(eprobs)
```

Description: df [6 × 1]

	probs <dbl>
1	0.9735887
2	0.2718127
3	0.2718127
4	0.1824764
5	0.3566587
6	0.9714764

6 rows

```r
library(dplyr)
Titanic2 <- merge(Titanic, eprobs, by="row.names", all.x=TRUE)
```

```{r}
library(ggplot2)
ggplot(data = Titanic2, aes(x=Age, y=probs, color = Class, shape=Sex)) +
  geom_point()+
  theme_minimal()
```