# INST 627 - Data Analytics For Information Professionals

# Team 9

# Final Paper

# Predicting Survival on the Titanic

Fall 2022

Ting-Hsuan Wang (tingw@umd.edu), Subject matter expert

Shashank Ramprasad (shashram@umd.edu), Technical expert

Sharvil Shastri (sharvil@umd.edu), Technical expert

Marie-Helene Trepy (mtrepy@umd.edu), Project manager

**Introduction**

This project is about the RMS Titanic's early morning sinking on April 15, 1912. This tragedy killed over 1,500 people. We want to know what attributes helped passengers survive the sinking, given the socioeconomic and gender inequality in the past. Our research question is whether ticket class, gender, or port of embarkation are statistically significant predictors of surviving the Titanic. In addition, we want to know how these attributes affect the probability of survival on the Titanic. These attributes were chosen because they give us additional information about the passengers. Ticket class symbolizes their socioeconomic level, and examining gender analyzes whether it is a vital role in a passenger's survival. Finally, the port of embarkation represents the passenger's origin.

## Methods and Data

We obtained our dataset from Kaggle: Titanic - Machine Learning from Disaster. It includes 891 actual Titanic passengers. Each row represents a passenger, and the columns describe their attributes, such as whether they survived, ticket class, age, gender, number of siblings, number of parents, etc. We selected four attributes to help us answer our research questions.

1. Whether or not they survived (0=No; 1=Yes)
2. Class of ticket purchased (1=first; 2=second; 3=third)
3. Gender (Male or Female)
4. Port of embarkation (C=Cherbourg; Q= Queenstown; S=Southampton)

**Dependent variable:** Survival

**Independent variables:** Ticket class, gender and port of embarkation

To find out if there is a relationship between our dependent and independent variables, we will initially perform three chi square tests of independence for each DV/IV pair since all our variables are categorical in nature. If a significant relationship is found, then we will perform a logistic regression to find the exact relationship between our independent variables and the dependent variable. Logistic regression is chosen because the predicted outcome is binary (survived or not survived).

## Results

### *Chi Square test 1*

Dependent variable: Survival. Independent variable : Ticket class

**Descriptive statistics**

1) For Dependent variable:

| **Status** | **Number of people** | **% of population** |
|---|---|---|
| Survived | 342 | 38.38 |
| Did not Survive | 549 | 61.61 |

*Table 1: Number of people that survived vs did not survive on the Titanic*

2) For Independent variable:

| **Type of Class** | **Number of Tickets purchased** | **% of Population** |
|---|---|---|
| | | |
| First Class | 216 | 24.24 |
| Second class | 184 | 20.65 |
| Third class | 491 | 55.1 |

*Table 2: Number of tickets purchased for First, second and third class respectively*

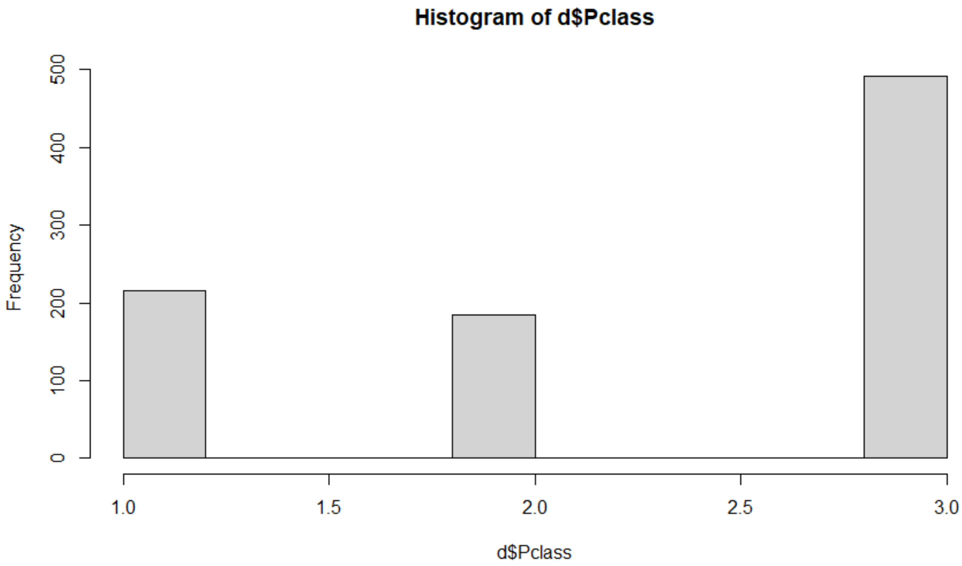
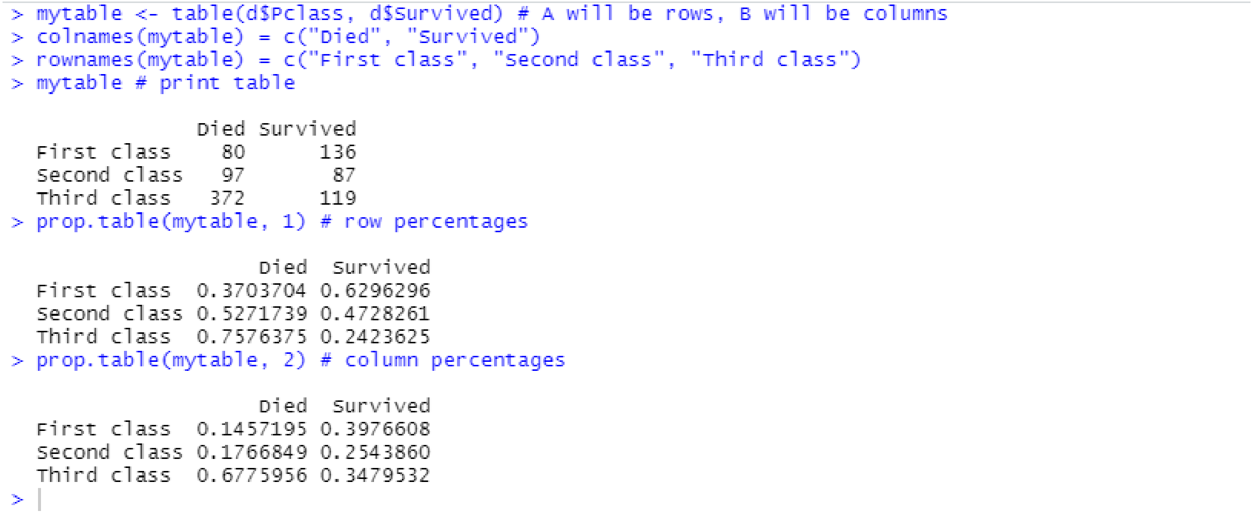


*Fig 1: Histogram plot of different classes and their frequency on the RMS Titanic*

The cross table result is attached below. The prop table calculates the value in each cell as a proportion to all the other values in the table. For example, the first cell in prop.table(mytable, 1) indicates the percentage of first-class passengers who died on the Titanic. We can see that approximately 75% of the passengers with a third class ticket perished as compared to the 52% second-class passengers  and 37% first class tickets respectively.



```
> mytable <- table(d$Pclass, d$Survived) # A will be rows, B will be columns
> colnames(mytable) = c("Died", "Survived")
> rownames(mytable) = c("First class", "Second class", "Third class")
> mytable # print table

             Died Survived
First class    80      136
Second class   97       87
Third class   372      119
> prop.table(mytable, 1) # row percentages

                  Died  Survived
First class  0.3703704 0.6296296
Second class 0.5271739 0.4728261
Third class  0.7576375 0.2423625
> prop.table(mytable, 2) # column percentages

                  Died  Survived
First class  0.1457195 0.3976608
Second class 0.1766849 0.2543860
Third class  0.6775956 0.3479532
> |
```

*Fig. 2: Class* x *Survival cross table results*

**State hypothesis**

$H_0$: There is no relationship between the passenger’s ticket class and their chances   of survival.

$H_A$: There is a statistically significant relationship between the passenger’s ticket class and their chances of survival.

**Observed Values**

| **Class** | **Survived** | **Not Survived** | **Total** |
|---|---|---|---|
| **1** | 136 | 80 | 216 |
| **2** | 87 | 97 | 184 |
| **3** | 119 | 372 | 491 |
| **Total** | 342 | 549 | 891 |

*Table 3: Observed values of Passengers that survived and didn't survive among different classes*

**Expected Values**

| **Class** | **Survived** | **Not Survived** | **Total** |
|---|---|---|---|
| **1** | 83 | 133 | 216 |
| **2** | 71 | 113 | 184 |
| **3** | 188 | 303 | 491 |
| **Total** | 342 | 549 | **891** |

*Table 4: Expected values of Passengers that survived and didn't survive among different classes*

Using the formula,

$$\chi 2 = \Sigma \frac{(Observed\ values - Expected\ values)^2}{(Expected\ Values)}$$

$$\chi 2 = 101.8721341$$

Using RStudio, we have

pchisqr(101.8721341,2,lower.tail=FALSE) <- 7.563924e-23

Since our p value is lower than the significance level of 0.05, we can **reject** the null hypothesis and conclude that there is a relation between the ticket class of a passenger and their likelihood of surviving the Titanic.

## *Chi Square test 2*

Dependent variable : Survival. Independent variable : Gender

**Descriptive statistics**

For independent variable:

| **Gender** | **Number** | **% of population** |
|---|---|---|
| Male | 577 | 64.75 |
| Female | 314 | 35.24 |

*Table 5: Number of Males as compared to Females on the RMS Titanic*

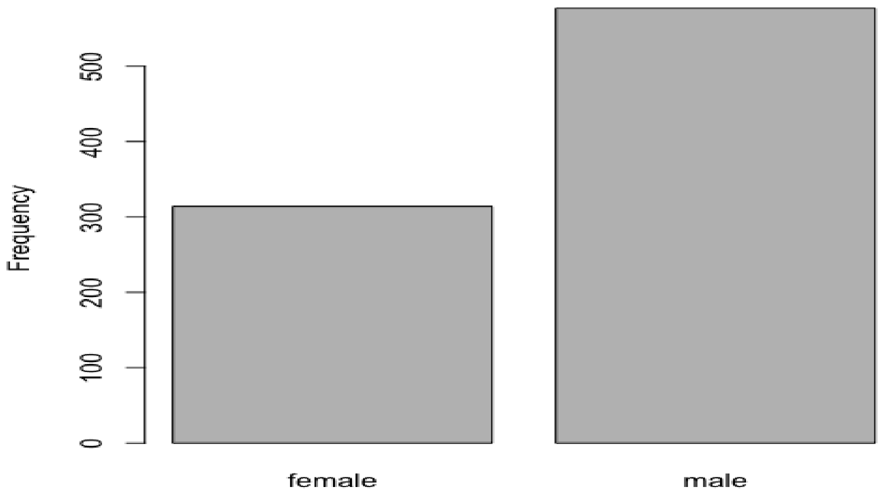
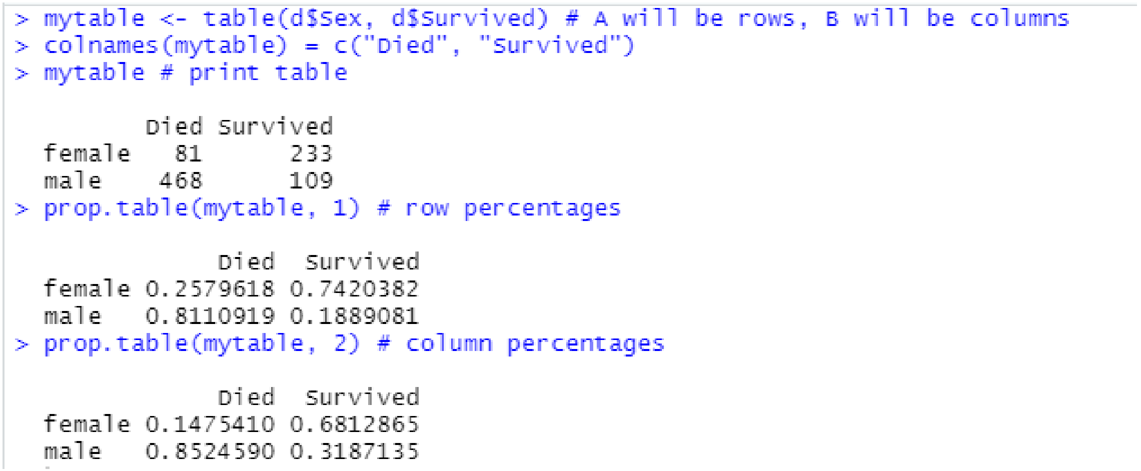


*Fig 3: Plot of frequency distribution of gender*



```
> mytable <- table(d$Sex, d$Survived) # A will be rows, B will be columns
> colnames(mytable) = c("Died", "Survived")
> mytable # print table

         Died Survived
  female   81      233
  male    468      109
> prop.table(mytable, 1) # row percentages

              Died  Survived
  female 0.2579618 0.7420382
  male   0.8110919 0.1889081
> prop.table(mytable, 2) # column percentages

              Died  Survived
  female 0.1475410 0.6812865
  male   0.8524590 0.3187135
```

*Fig. 4: Gender* x *Survival cross table results*

Through the cross tables we can see that almost 75% of the female passengers survived as opposed to 19% of the male passengers.

**State hypothesis**

$H_0$: There is no relationship between the passenger's gender and their survivability.

$H_A$: There is indeed a relationship between the gender of the passenger and their survivability.

**Observed values**

| **Gender** | **Survived** | **Not Survived** | **Total** |
|---|---|---|---|
| **Male** | 109 | 468 | 577 |
| **Female** | 233 | 81 | 314 |
| **Total** | 342 | 549 | 891 |

*Table 6: Observed values of Males and Females that survived and didn't survive*

**Expected values**

| Gender | Survived | Not Survived | Total |
|---|---|---|---|
| **Male** | 221 | 356 | 577 |
| **Female** | 121 | 193 | 314 |
| **Total** | 342 | 549 | 891 |

*Table 7: Expected values of Males and Females that survived and didn't survive*

$$\chi 2 \ = \ \Sigma \frac{(Observed\ values - Expected\ values)^2}{(Expected\ Values)}$$

$$\chi 2 \ = \ 260.660372$$

pchisqr(260.660372,1,lower.tail=FALSE) <- 1.231886e-58

Since our p value is lower than the significance level of 0.05, we can **reject** the null hypothesis and conclude that there is a relation between the passenger's gender and whether or not they survived.

## *Chi Square test 3*

Dependent variable : Survival. Independent variable : Port of Embarkation

**Descriptive statistics:**

[1]For Independent Variable:

| **Port of Embarkation** | **Number of Individuals** | **% of Population** |
|---|---|---|
| Cherbourg | 168 | 18.89 |
| Southampton | 644 | 72.44 |
| Queenstown | 77 | 8.66 |

*Table 8: Number of Individuals with different port of embarkation respectively*

Below is a frequency distribution of the port of embarkment (C = Cherbourg, Q = Queenstown, S = Southampton and the blank indicates the two null values)

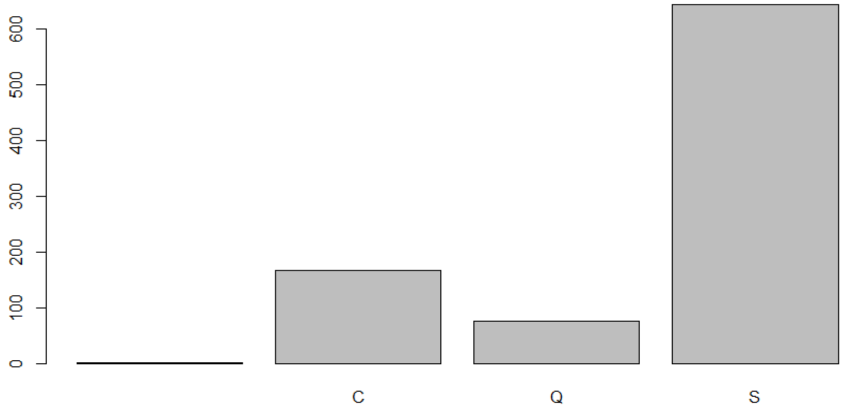
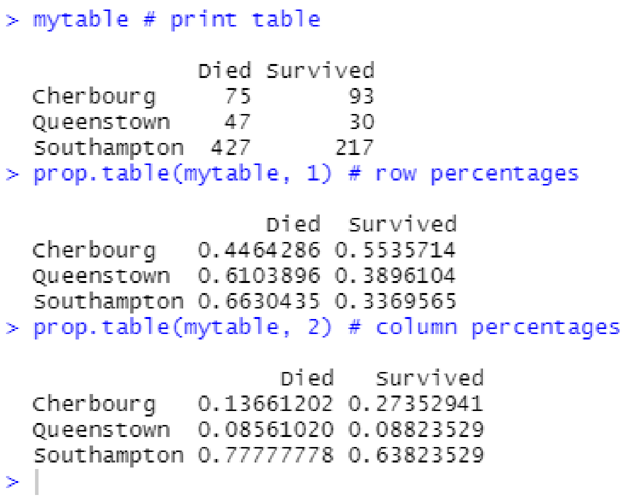


*Fig 5: Plot of Frequency distribution of the port of embarkment*

[1] In this dataset, port of embarkment for two of the passengers was missing, therefore we have ignored them in our analysis and our total sample size comes to 891-2 = 889.

Following are the cross table results:



```
> mytable # print table

             Died Survived
  Cherbourg    75       93
  Queenstown   47       30
  Southampton 427      217
> prop.table(mytable, 1) # row percentages

                  Died  Survived
  Cherbourg   0.4464286 0.5535714
  Queenstown  0.6103896 0.3896104
  Southampton 0.6630435 0.3369565
> prop.table(mytable, 2) # column percentages

                   Died   Survived
  Cherbourg   0.13661202 0.27352941
  Queenstown  0.08561020 0.08823529
  Southampton 0.77777778 0.63823529
> |
```

*Fig. 6: Embarkment port* x *Survival cross table results*

Around 55% of passengers who embarked from Cherbourg survived as opposed to 38 and 33 percent for Queenstown and Southampton respectively.

**State hypothesis**

$H_0$: There is no relationship between where the passenger embarked and their survivability.

$H_A$: There is a relationship between where the passenger embarked and their survivability.

**Observed values**

| **Town** | **Survived** | **Not Survived** | **Total** |
|---|---|---|---|
| Cherbourg | 93 | 75 | 168 |
| Queenstown | 30 | 47 | 77 |
| Southampton | 217 | 427 | 644 |
| **Total** | 340 | 549 | 889 |

*Table 9: Observed values of passengers that survived and didn't survive based on their port of embarkment*

**Expected values**

| **Town** | Survived | Not Survived | Total |
|---|---|---|---|
| Cherbourg | 64.26 | 103.7 | 168 |
| Queenstown | 29.4 | 47.55 | 77 |
| Southampton | 246.3 | 397.7 | 644 |
| Total | 340 | 549 | 889 |

*Table 10: Expected values of passengers that survived and didn't survive based on their port of embarkment*

Using the formula,

$$\chi 2 = \Sigma \frac{(Observed\ values - Expected\ values)^2}{(Expected\ Values)}$$

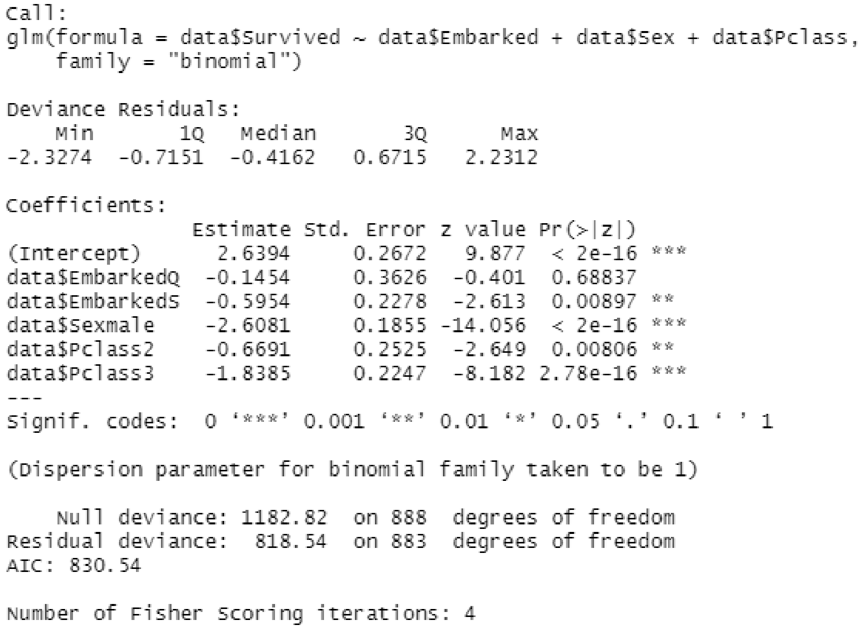
We get the value of chi square χ2 as 26.45

pchisq(26.45,2,lower.tail=FALSE) <- 1.80491e-06

Since our p value is lower than the significance level of 0.05, we can **reject** the null hypothesis and conclude that there is a relation between a passenger's port of embarkation and whether or not they survived the Titanic.

### Logistic Regression

Having found a statistically significant relation between each of our dependent and independent variables, we can now perform a logistic regression test to model the probability of surviving the Titanic based on our aforementioned predictor variables. We used the glm() function in R and found the following results:



```
Call:
glm(formula = data$Survived ~ data$Embarked + data$Sex + data$Pclass,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3274  -0.7151  -0.4162   0.6715   2.2312

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.6394     0.2672   9.877  < 2e-16 ***
data$EmbarkedQ  -0.1454     0.3626  -0.401  0.68837
data$EmbarkedS  -0.5954     0.2278  -2.613  0.00897 **
data$Sexmale    -2.6081     0.1855 -14.056  < 2e-16 ***
data$Pclass2    -0.6691     0.2525  -2.649  0.00806 **
data$Pclass3    -1.8385     0.2247  -8.182 2.78e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.82  on 888  degrees of freedom
Residual deviance:  818.54  on 883  degrees of freedom
AIC: 830.54

Number of Fisher Scoring iterations: 4
```

*Fig. 7: glm() function output in R*

We can see that for a particular variable, one of its categories is considered as the reference and the coefficients of each of the other categories are a comparison against this reference. For example, $Sexfemale is considered as the reference category (since it is not mentioned in the output) and the coefficient of $Sexmale indicates that changing the gender category from Female to Male would reduce the log odds of surviving the Titanic by 2.6.

Similarly changing the ticket class from first to second would reduce the log of the odds of surviving the Titanic by 0.66 and changing the category from first to third would decrease it by 1.83.

We can now form our logit equation based on the coefficients of the predictor variables mentioned in our output:

$$ln\left(\frac{prob_{surviving}}{1-prob_{surviving}}\right) = 2.64 - \left(0.59\ X_{EmbarkedS}\right) - \left(2.6\,X_{Gender}\right) - \left(0.67\,X_{Class2}\right) - \left(1.83X_{Class3}\right)$$

## Discussion and Conclusion

The chi square tests showed that all three factors have a significant relationship with the passenger survival. The logit equation also showed that gender had the greatest impact on survival (it had the largest regression coefficient). The Titanic's third-class passengers had the lowest survival rate.

More women tend to survive disasters because of the “women and children first” code of conduct which states that the lives of women and children are to be saved first in a life-threatening situation. Similar orders were followed by the officers on the Titanic.

An emerging question in this study is why were second and third class passengers disproportionately affected in this disaster? This was mainly due to the layout of the ship. The top deck, which held all the lifeboats, was closest to the first class cabins (as shown in the image below). Also, the number of passageways and stairways that had to be used to get to the top deck progressively increased for each of the subsequent classes as they were situated farther away making it harder for these passengers to reach the top deck in time.

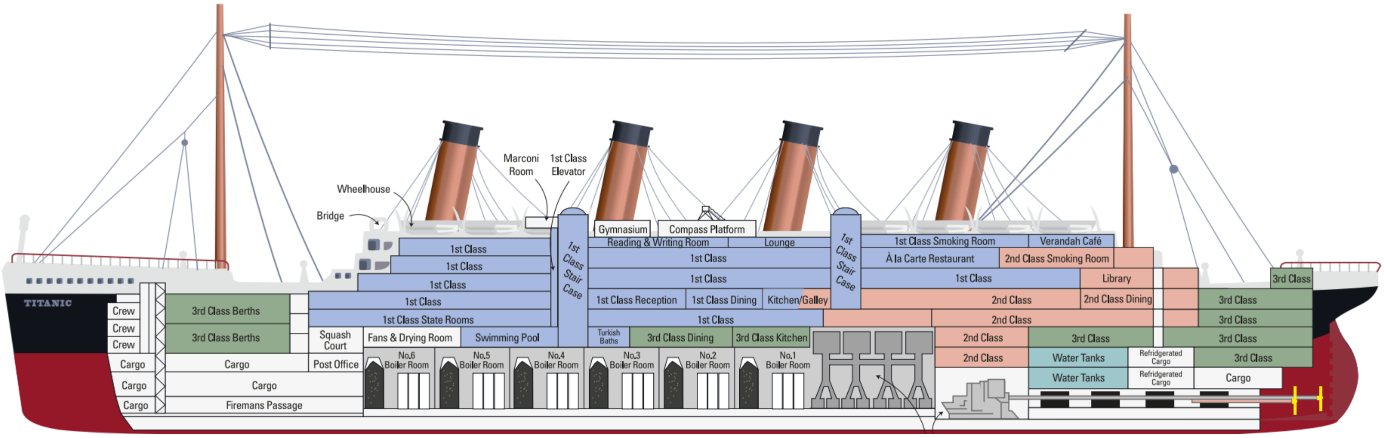


Fig. (*Titanic Deck Plan for Titanic Resource Pack (Pdf)*, n.d.)

## Limitations

**1.** Around 1,500 people died on the Titanic, but our dataset has details of only 891 passengers. The lack of complete data limits our capability to draw insights on the population.

**2.** Since age, number of siblings, and other independent variables were outside our scope, we did not include them in the logistic regression. Hence, future research could consider investigating all attributes and adjusting the logit equation accordingly.