# **The Titanic: Survival outcomes and its variance**

**Contents**

[**The Titanic: Survival outcomes and its variance** 1](#_Toc151567156)

[**Introduction and context to the Titanic catastrophe and dataset** 2](#_Toc151567157)

[**Descriptive and confounding variables breakdown** 3](#_Toc151567158)

[**Descriptive statistics** 3](#_Toc151567159)

[**Potential confounding variables** 7](#_Toc151567160)

[**Roundup of Descriptive statistics** 8](#_Toc151567161)

[**Logistic Regression Analyses** 9](#_Toc151567162)

[**Model 1** 9](#_Toc151567163)

[**Model 2** 9](#_Toc151567164)

[**Diagnostics** 11](#_Toc151567165)

[**Limitations** 12](#_Toc151567166)

## **Introduction and context to the Titanic catastrophe and dataset**

This presentation uses data from the Titanic dataset, which was provided in Excel format, and I manipulated using R. The dataset and presentation relate to the catastrophe of the Titanic, which sank in 1912, taking the lives of many who boarded the ship.

To give some historical context, this catastrophe, mainly occurred due to some design faults regarding a lack of lifeboats, and that many of the available lifeboats sailed off under full capacity. Leaving many passengers to jump into the water where they would subsequently freeze to death within minutes.

Within this sample, 549 people did not survive, and 342 people did survive - 61% of Passengers on the Titanic died. Overall, the dataset looks at cohorts, such as Age of passengers, Fare Price paid, Number of Spouses, Children, Siblings onboard, Class of ticket and our outcome variable here – Survival rate. Survival rate is measured by whether an individual survived or not, rendering this a Binary variable, which informs the subsequent selection of a Logistic regression.



## **Descriptive and confounding variables breakdown**

### **Descriptive statistics**

A screenshot of a graph

Description automatically generatedWe move on to exploring some of the relationships between Cohorts and their Survival outcomes:

As shown here, those with 1st class tickets were overwhelming more likely to survive than those with 3rd and 2nd class tickets. With around 65% of those with 1st class tickets surviving, compared to around 25% of those with 3rd class tickets surviving.

Chart, bar chart

Description automatically generatedGiven the context that those with lower class tickets were often within Cabins that were more cramped, and lower down within the boat – this gives some context as to why Survival rate was then lower for lower Class ticket passengers.

Again, the visualization here seems to be suggesting an undercurrent of social status and chances of survival, with those onboarding from Cherbourg, France, seeming to have higher survival, which makes sense given the context that many famous French and highly prestigious people boarded here.

Chart, bar chart

Description automatically generated

By considering an individual’s Sex, in relation to Class of ticket and subsequent Chances of Survival it is shown that Ticket Class, introduced earlier, is still suggesting that higher class tickets improve chances of Survival. But, for the Male populous with 2nd and 3rd class tickets, there is a survival rate of around 18%, but for the Female populous, there is 45% of those Females with 3rd class tickets surviving, and those with 2nd class – a 90% rate. This could be pointing to the ‘Women and Children first policy’, prioritising the lives of Women and Children first and foremost over males in the recovery process.

Chart, bar chart

Description automatically generatedHere, I added Age, and in relation to the Children and Women first policy we do seem to see, regardless of Class of ticket, that lower Age groups do not necessarily have lower fatality rates. This could be pointing toward the fact that those with more children, often had more 3rd class tickets, which relates to the relationship explored earlier, of lower-class ticket and lower survival rate.

### **Potential confounding variables**

Chart, bar chart

Description automatically generatedMoving on to some potential confounding variables, which is when a variable that is added to a regression, changes the relationship between the original independent variable and the outcome variable.

When exploring the relationship opened up earlier, of Embarkation location and Class of Ticket, we can see the population Embarking from Chersbourg, France, that almost 60% of them had 1st class tickets. Providing some context to earlier insights, where those from Chersbourg had a higher survival rate to those embarking from Queenstown, Ireland, and Southampton.

**Chart, bar chart

Description automatically generated**Again, touching upon some themes established already, there is a potential confounding relationship between Sex and Class of Ticket, and its relationship to Survival. Demonstrating that the Female population was made up of more 1st class ticket holders, and less 3rd class ticket holders, than Men.

### **Roundup of Descriptive statistics**

Overall, there seems to be some strong links between people’sSex and Survival rate, but not so much Age surprisingly. Furthermore, there seems to be an undercurrent of class/ social status heavily informing the analysis of the Titanic catastrophe.

As already explored, there was a Titanic policy of ‘women and children first’, greater access to safety equipment and less cramped conditions, which is important context here. Additionally, with Fare Price and Ticket class link, it is important to understand their inherent relationship in the coefficient outcome of the subsequent regression and as confounding variables.

As the number of Parents, spouses, children, and siblings increased people’s survival rate would diminish , these variables are potential confounders to one another when analysing survival rate.

## **Logistic Regression Analyses**

### **Model 1**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

The odds of an individual on the Titanic with a 2nd class ticket Surviving on the Titanic decreases by 70% over an individual with a 1st class ticket, with statistical significance of 99.9% in the Titanic passenger population

In terms of Age, a one-unit increase, decreases survival odds by 4%, with statistical significance amongst the Titanic passenger population of 99.9%. Essentially, as age increases, survival odds decrease – replicating the small influence as explored in the visualizations, which is surprising given Women and Children first policy.

We can see, where the visualizations pointed toward some importance for the Port of Embarkation , when controlled for all other variable in the model – it does not have statistical significance as an explainer of Survival outcomes in the Census population of Titanic passengers

### **Model 2**

In Model 2, the variables that were not showing statistical significance in Model 1 were removed.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The removal of variables resulted in profounder results, where 2nd class ticket holders had 76% decreases of survival odds over a 1st class ticket holder, compared to Model 1 of a 70% decrease, but demonstrating the same statistical significance amongst the Titanic passenger population In addition, to showing 93% decrease in Survival odds for Males over Females, with 99.9% statistical significance, remaining the same across both Models. Furthermore, 3rd class ticket holders showed 93% decrease in survival odds over those with 1st class tickets, with a 99.9% statistical significance

Overall, the McFadden Psuedo R Squared tells us that the Model is explaining 30% of the variance of Survival Outcomes amongst Passengers of the Titanic – which is relatively good in general terms, and we are capturing some good variance within Survival outcomes, through the use of the selected cohorts.

### **Diagnostics**

Chart, histogram

Description automatically generated

The Cook’s Distance is showing that some influential outliers could be skewing the results of our Regression and be extreme outliers. 0.10 on the Cooks Distance is used by some as a threshold for removal, and we have a few data points above this mark.

Chart, scatter chart

Description automatically generated

But, the standardized Residuals scatterplot, also must be inspected, and this shows that not absolute standardized residual point falls above 3, or below -3. Thus, showing that these potential influential outliers highlighted in the Cooks distance, are not skewing our results too much, and do not require further attention.

In conclusion, our model indicates that the most important variables are Ticket Class and Sex, as initially expected, and that our model is relatively reliable, with small levels of confidence for our coefficients, and good variance of the survival outcomes of the titanic shown.

## **Limitations**

In terms of Limitations, we lost some variability and nuance in explaining Survival Outcomes, with around 100 Missing Values for Age, and over 500 missing values for Cabin Position. Therefore, our results regarding Age, could be impacted by the lack of data for some individuals, and with greater data available, there could be a stronger impact on Survivability amongst younger people – in line with the Women and Children priority. Additionally, as touched upon, a lack of data points for Cabin position means we cannot inspect the actual position of individuals on the boat, and the explanation that those lower down were more susceptible to death. Although Ticket Class does infer that lower quality Ticket Class would put you within worse conditions and so you would be likely to be in a lower cabin position, but we cannot be sure cabin position would be important in explaining survival outcomes on the Titanic.