Who Survived the Titanic?

Jake Derby

Bellevue University

DSC530 – Data Exploration and Analysis

Matthew Metzger

March 2, 2025

Introduction

The notorious ocean liner, Titanic, and its tragic maiden voyage in 1912 has captured the imaginations and sympathies of generations. The popular movie starring Leonardo DiCaprio has only added to the event's notoriety. Of interest to my analysis was the movie's depiction of certain classes and types of passengers that were prioritized over others for evacuation into the ship's lifeboats. My analysis attempted to get a feel for Titanic's demographic makeup and how those passenger characteristics might have impacted the ultimate outcome of who survived and who did not. Specifically, I was interested in testing the hypothesis that higher-class passengers were more likely to survive the event.

Exploration

A dataset containing records of 891 passengers onboard Titanic was obtained from Kaggle. Though mostly complete, a handful of variables contained a significant number of missing values. Before any exploration or analysis took place, NAs were either removed or filled in (for numeric variables, median was used).

The main findings of my data exploration help to confirm our assumptions about Titanic passengers. Firstly, Titanic was a luxury ocean liner that was meant to mainly carry passengers and mail. As such, a first target for exploration was to get a feel for the distribution of wealth among the passengers. Using fare paid and passenger ticket class as proxies of the respective passenger's likely wealth, I found both fare and class to be significantly skewed to cheaper and lower-class tickets, though with notable outliers to the upside in terms of fare. This suggests that, though most passengers were indeed 3rd class, a considerable proportion of passengers were likely far more affluent.

Another relevant finding includes a far greater proportion of men than women onboard (this fact is particularly crucial when analyzing the results of the logistic regression later). The age distribution of Titanic passengers does not appear to conform perfectly to a normal distribution; a disproportionate number of young people (ages 4-7 and mid 20s-30s) were onboard Titanic. The variables of age and fare were found to significantly, positively correlated (as depicted by age vs log fare and confirmed by a Pearson's r of 0.1).

Hypothesis Test

I utilized a chi-squared analysis to assess my hypothesis that passenger class was related to the likelihood of survival. Chi-squared testing is an appropriate method since that neither variable of interest is continuous nor numerical. The result of my chi-squared test found a significant relationship between class and survival ($p < 0.000$). Looking at the chi-squared table, we can easily see the nature of the relationship; passengers of higher class tended to survive at greater numbers than lower classes.

Logistic Regression

Having confirmed my hypothesis, I decided to build and fit a logistic regression to predict survival using age, sex, # of parents/children onboard, fare and port of embarkation. Since that fare and passenger class were highly correlated, I decided to use fare instead to mix it up. All variables were found to be significant predictors of survival. The model yielded a pseudo-R-squared value of 0.27 and a prediction accuracy of 79%. Given the constant variable we added to our predictors (-0.81), any given passenger did not have a very good chance of making of Titanic when all our predictors are held constant at 0. The biggest finding of this regression was that sex was by far the most powerful predictor of survival. Age barely reached significance using the $p < 0.05$ standard. As suspected, fare was the second most powerful predictor of survival.

Though this regression was successful in predicting roughly 4 out of every 5 passengers' fate, my results also strongly suggest that other factors were almost certainly involved in effecting survival. One of these factors may have been the location of a passenger's room (higher decks presumably would have had a higher survival rate since Titanic flooded from bottom-up). If we had access to information that confirmed that ticket class strongly correlated with room location (lower decks for lower-class tickets), we could presumably add this variable to our regression and simplify the model by getting rid of passenger class and fare (to avoid multicollinearity).