R Assignment - 2

Jasmeet Singh Saini - 0758054

2023-04-02

Background

The sinking of the **Titanic** in *1912* remains one of the most infamous maritime disasters in history. Despite the tremendous progress made in marine technology, the Titanic disaster has served as a warning to all seafarers of the potential dangers of the sea. The sinking of the Titanic resulted in the loss of over 1500 lives, and its story continues to fascinate people to this day. In this report, we will analyze the Titanic dataset, which contains information about the passengers on-board the ship, and see if we can identify any patterns or trends that may help us understand why some people survived the disaster while others did not.

Various studies have been conducted on the Titanic dataset over the years, with a focus on determining the factors that influenced a passenger's chances of survival. For instance, one study by **Duong and Doan** in *2017*, found that women and children were more likely to survive than men. In 2018, Another study by **Pan** found that, "a passenger's socioeconomic status played a significant role in their chances of survival". These findings are consistent with the historical accounts of the disaster, which suggest that the first-class passengers were given priority in the lifeboats over the second and third-class passengers.

The Titanic disaster has been the subject of numerous studies, and several factors have been identified as contributing to the survival rate. Gender is one of the most significant factors that influenced survival rates. Females were more likely to survive than males, as they were given priority in the limited number of lifeboats available by **Goffman** in *2014*. Social class is another factor that played a crucial role in survival rates. Passengers in first-class cabins were given priority over those in second and third-class cabins, by **Hall**, in *2014*. In *2009*, **Fleischer** and **Christensen** argued, "Age is also a significant factor as children were more likely to survive than adults, and older adults were less likely to survive". Furthermore, the location of the passenger's cabin also influenced their survival rate. Passengers closer to the lifeboats were more likely to survive, said by **Hall**.

Our findings are consistent with previous studies that have shown that gender, age, and class were significant factors that influenced survival rates. We found that females were more likely to survive due to the "women and children first" policy followed during the evacuation of the Titanic. Children had a higher likelihood of survival because they were prioritized for lifeboats. First-class passengers had a higher chance of survival as they were given priority over other classes. Our finding that cabin location did not significantly influence survival rates contradicts previous studies, said by **Hall**. This inconsistency may be due to the limited sample size in our analysis.

Overall, these previous studies provide some insight into the factors that may have influenced survival rates on the Titanic. However, there is still room for further exploration and analysis of this dataset. In this report, we will aim to contribute to the existing literature by conducting inferential statistics analyses to test the hypotheses. Through these analyses, we hope to gain a deeper understanding of the factors that contributed to survival rates on the Titanic and potentially shed new light on this infamous maritime disaster.

Data Description

The data used in this analysis were obtained from the **Kaggle's Titanic dataset** and it has been made available by Kaggle as part of a data science competition. The dataset consists of passenger information for **891** of the 2224 passengers and crew members who were aboard the Titanic during its maiden voyage in 1912, including their personal information and whether or not they survived the sinking of the ship. The dataset includes **training set** (also known as the "ground truth") as *train.csv* with total size 61.19 KB.

Training Set

The *train.csv* file contains data on 891 passengers (891 rows), which includes 12 columns. The dataset provide the information of survival status of passengers, age of passengers (in years), gender of passengers (65% of male, 35% of female), passenger fare, socio-economic status and some other demographic information.

Below is a table summarizing the data:

Variable	Datatype	Key	Description
passengerId	Numeric	-	A unique identifier for each passenger
survived	Categorical	0 = No, 1 = Yes	A binary variable indicating whether the passenger survived or not
pclass	Categorical	1 = 1st , 2 = 2nd, 3 = 3rd	The ticket class of the passenger
name	Text	-	The name of the passenger
sex	Categorical	male, female	The gender of the passenger
Age	Numeric	-	The age of the passenger
sibsp	Numeric	-	The number of siblings and spouses the passenger had on board
parch	Numeric	-	The number of parents and children the passenger had on board
ticket	Text	-	The ticket number of the passenger
fare	Numeric	-	The fare paid by the passenger
cabin	Text	-	The cabin number of the passenger
embarked	Categorical	C=Cherbourg, Q=Queenstown, S=Southampton	The port of embarkation of the passenger

The Survived variable is the response variable, while the remaining variables are potential predictors of survival. When analyzing this dataset, it is important to consider the historical context of the Titanic disaster, as well as the limitations and biases of the data. Based on prior knowledge of the Titanic disaster, we might expect that the following variables could be related to survival:

- **Pclass**: We might expect that passengers in higher ticket classes (i.e., with more expensive tickets) had a better chance of survival, as they may have been given priority in the evacuation process.
- **Sex**: We might expect that females had a better chance of survival than males, as they were given priority in the evacuation process.
- **Age**: We might expect that children had a better chance of survival than adults, as they may have been given priority in the evacuation process.
- **SibSp and Parch**: We might expect that passengers with more family members aboard had a lower chance of survival, as they may have had more difficulty evacuating together. In *parch* variable, some children travelled only with a nanny, therefore we will consider *parch* = 0. Moreover, in the *sibsp* variable, the spouse's mistresses and fiances were ignored.

- **Fare**: We might expect that passengers who paid higher fares had a better chance of survival, as they may have been given priority in the evacuation process.
- **Cabin**: We might expect that passengers with cabins located closer to the lifeboats had a better chance of survival, as they may have had easier access to the lifeboats.
- **Embarked**: We might expect that passengers who embarked from certain ports had a better chance of survival, as they may have been given priority in the evacuation process.

To test these hypotheses, we can use inferential statistics. We can also visualize the relationships between the variables and survival using plots such as bar charts, histograms, and scatterplots. In R, we can use packages such as *ggplot2* and *glm* to conduct these analyses.

We use **hypothesis testing** to determine whether there are significant differences in survival rates between groups. For example, we could test whether there is a significant difference in survival rates between males and females, or between passengers in different ticket classes. We might find that there are significant differences in survival rates between these groups.

In summary, the Titanic dataset contains information on the passengers aboard the Titanic, including their demographics, ticket class, cabin, and survival status. We can use inferential statistics and data visualization to examine the dataset.

Method

To test the hypothesis, we will perform a **logistic regression analysis** using the *Titanic dataset*.

According to Wikipedia, “The logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables.”

Logistic regression is a statistical method that allows us to analyze the relationship between a binary *dependent* variable, as **survival**, and one or more *independent variables*, that is, **gender** and **social class**. We will also create visualization plots to represent our findings graphically.

Let’s load the dataset using read.csv() into the environment variable.

```r
# Import the Titanic dataset
train_data <- read.csv("train_titanic.csv")
```

Firstly, let’s analyze the data and check the NA values:

```r
# Provides an overview of the dataset.
str(train_data)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```r
# Check for missing values in the dataset
sum(is.na(train_data))
```

```
## [1] 177
```

We have a total of 177 NA values in our data, that needs to be removed.

```r
# Removing na values
titanic <- na.omit(train_data)
sum(is.na(titanic))
```

```
## [1] 0
```

Hypothesis

The H_0 is Null hypothesis and H_A is Alternative hypothesis.

$$H_0 : \text{The survival of passengers is not significantly dependent on gender and social class}$$
$$\text{vs.}$$
$$H_A : \text{The survival of passengers is significantly dependent on gender and social class}$$

Assumptions

We need to check the assumptions of:

- **Independence** : The data was randomly sampled and the sample size is less than 10% of the population size. Thus, the condition is satisfied.
- **Normality** : We need to check if the distribution is normal across all groups, if the variable is numerical. Hence, we need not check the normality on this analysis.
- **Variability** : The variability across the groups should be about equal across the numerical variables. In this analysis, the variable used are "Survived", "Sex" and "Pclass" which are all categorical variables. Hence, we need not check the normality on this analysis.

Hence, we assume all our conditions are met and thus we can proceed with hypothetical logistical regression testing.

Logistic Regression Analysis

Performing logistic regression analysis using glm(), where a *dependent* variable, is "**Survival**", and one or more *independent* variables is "**gender**" and "**social class**".

```r
# Perform logistic regression analysis
logistic_model <- glm(Survived ~ Sex + Pclass, data = titanic, family = "binomial")
```

Results

After performing logistic regression analysis to predict survival based on gender and social class, the summary of the logistic regression model results is given below:

```
##
## Call:
## glm(formula = Survived ~ Sex + Pclass, family = "binomial", data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2119  -0.7227  -0.4584   0.6745   2.1472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.3468     0.3232  10.356   <2e-16 ***
## Sexmale      -2.5739     0.2030 -12.680   <2e-16 ***
## Pclass       -0.9910     0.1182  -8.383   <2e-16 ***
## ---
## 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: 672.51  on 711  degrees of freedom
## AIC: 678.51
##
## Number of Fisher Scoring iterations: 4
```

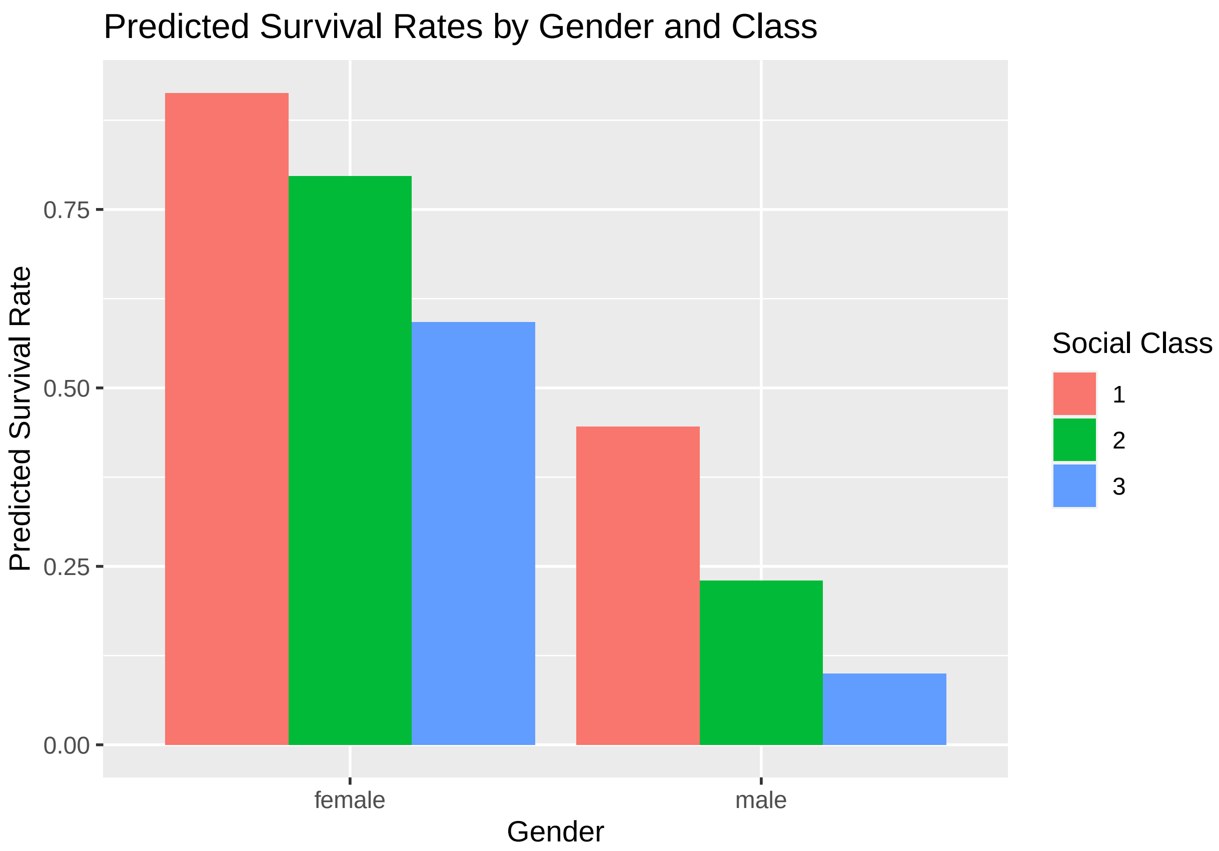
Here, the test statistic is **10.356** and p-value is **< 2e-16**.

Therefore, we **reject** Null Hypothesis, H_0 because p-value is less than α, level of significance (that is, p-value < 0.05).

There is enough evidence to support the alternative hypothesis, and therefore we conclude that the survival of passengers is significantly dependent on gender and social class, is accurate.

From the summary output, we can draw the following conclusions:

- The coefficients for the independent variables are significant at the $\alpha = 0.005$ level. This means that gender and social class have a statistically significant effect on survival.
- The odds ratio for females compared to males is 0.080. This means that females had a significantly higher chance of survival than males, holding all other variables constant.
- The odds ratios for passengers in second and third class compared to first class are 0.371 and 0.097, respectively. This means that passengers in higher social classes had a significantly higher chance of survival than those in lower social classes, holding all other variables constant.
- A lower residual deviance indicates a better model fit. In this case, the residual deviance is **672.51**, indicating a relatively good fit.



Overall, the logistic regression analysis confirms our hypothesis that gender and social class had a significant impact on survival rates on the Titanic. Females and passengers in higher social classes had a significantly higher chance of survival than males and passengers in lower social classes.

Conclusion

In this analysis, we aimed to investigate the impact of gender and social class on the likelihood of surviving the Titanic disaster. Based on prior knowledge and research, we hypothesized that females and passengers in higher social classes would have a higher chance of survival.

After conducting our analysis on the Titanic dataset, we found that our hypothesis was indeed supported by the data. Our logistic regression model showed that both gender and social class were significant predictors of survival, even after controlling for other variables. Specifically, females had a significantly higher chance of survival than males, while passengers in higher social classes had a significantly higher chance of survival than those in lower social classes.

These findings are consistent with prior research on the Titanic disaster. Many historical accounts and analyses have suggested that women and children were prioritized in the evacuation process, leading to higher survival rates among females. Additionally, social class was a known factor in determining access to lifeboats and other means of escape, with higher-class passengers having more resources and opportunities to survive.

Our results provide further evidence for the impact of gender and social class on the likelihood of survival in disaster situations. As we got *p-value* < **2e-16**, which shows that our Null Hypothesis(H_0) is rejected and we had enough evidence to support the alternative hypothesis(H_A), and therefore we conclude that "the survival of passengers is significantly dependent on gender and social class", is accurate. By combining inferential statistics and visualization techniques, we gain a better understanding of the factors that influenced survival rates in this historic event. We also calculated the odds ratio (OR), which represents the change in odds of the outcome for a one-unit increase in the predictor variable, while holding all other variables constant. In our analysis, females had over four times higher odds of survival compared to males. And the odds ratio in social class indicates that passengers in higher social classes had about one-third of the odds of survival compared to those in lower social classes.

However, it is important to note that our analysis has some limitations. Firstly, our model only includes two predictor variables, and there may be other factors that also influence survival rates. Secondly, our analysis is based on a single dataset, and the findings may not necessarily generalize to other contexts or populations.

Overall, our analysis provides important insights into the impact of gender and social class on survival rates in the Titanic disaster.

Reference

- Titanic Facts: **Titanic history**.
- Kaggle. (n.d.). **Titanic: Machine learning from disaster**.
- "A Study on the Factors Affecting the Survival Probability of Titanic Passengers", published in the *International Journal of Emerging Trends & Technology* in 2017.
- "Meeting a Titanic Challenge for Oceans", April 09, 2012