Utilizing R Programming for In-Depth Analysis of Individuals and Familial Dynamics Aboard the Titanic: A Comprehensive Research Study

Oleh:

Nama: Darrien Rafael Wijaya

NIM: 2602064241

Kelas: LA06

Bina Nusantara University

2023/2024

Table of Contents

Chapter 1 – INTRODUCTION

1.1. Introduction

The sinking of the RMS Titanic in 1912 was a tragic event that claimed many lives. During its first voyage, the Titanic hit an iceberg and sank in the North Atlantic, leading to the loss of 1502 out of 2224 passengers and crew. This disaster highlighted the need for better safety regulations in maritime travel.

One major problem contributing to the high death toll was the lack of enough lifeboats for everyone on board. After the Titanic sank, it became clear that survival was partly a matter of chance. Certain groups, like women, children, and the upper-class, had better chances of surviving.

To understand this better, we want to analyze the patterns in survival outcomes using machine learning. Our goal is to create a predictive model that helps us see which factors influenced who survived the Titanic disaster. By looking at different factors and using advanced algorithms, we aim to uncover the complex dynamics that determined the fate of the people on the ship.

1.2. Objective

1.2.1. Family Structure Exploration:

In this pivotal segment of our research, we embark on a comprehensive exploration of the familial structures embedded within the Titanic dataset. Our objective is to employ a meticulous investigation into key features, specifically 'SibSp' (number of siblings/spouses aboard) and 'Parch' (number of parents/children aboard), to discern and categorize the intricate web of familial relationships among passengers.

1.2.2. Explore Demographic Pattern:

The central aim of this research is to conduct a comprehensive examination of the demographic distribution within the Titanic passenger population. This includes a detailed analysis of key demographic factors such as age, gender, and socio-economic class. By scrutinizing these demographic elements, the study seeks to identify inherent patterns and trends that existed within the diverse cross-section of individuals aboard the Titanic.

1.2.3. Predictive Modelling:

In this critical phase of our research, we delve into the realm of predictive modelling, utilizing sophisticated machine learning techniques to unravel the intricate tapestry of survival likelihood for passengers aboard the Titanic. Our primary focus revolves around the pivotal role of family structures and the amalgamation of ensemble decisions, with the aim of enhancing our understanding of the interconnected dynamics among family members.

1.2.4. Family Survival Pattern:

In this phase of the research, we delve into an in-depth examination of survival patterns within the families identified in the Titanic dataset. The primary aim is to meticulously scrutinize the dataset for correlations and trends pertaining to family groups, unraveling the intricate dynamics that influenced survival rates during this historic maritime disaster.

1.3. Significance of Study

This research embarks on a compelling exploration into the dynamics of familial relationships aboard the Titanic, coupled with the innovative application of a majority vote ensemble. The study holds paramount significance for various reasons:

1.3.1. Unveiling Human Stories Amidst Tragedy:

By delving into the dataset to identify genuine family structures, this study contributes to resurrecting the individual narratives of passengers aboard the Titanic. Beyond the general historical account, the research aims to humanize the tragedy by highlighting the interconnections and shared destinies within family units, enriching our understanding of the human experience during the disaster.

1.3.2. Informing Disaster Response Strategies:

The identification of family survival patterns and the utilization of a majority vote ensemble offer valuable insights for enhancing disaster response strategies. Understanding how families navigated the crisis informs future disaster preparedness by shedding light on the importance of familial cohesion and collective decision-making in times of distress.

1.3.3. Advancing Predictive Modelling with Ensemble Techniques:

The application of a majority vote ensemble to predict survival likelihood based on familial structures represents a novel and impactful contribution to predictive modelling. This innovative approach acknowledges the collaborative nature of decision-making within families, paving the way for improved accuracy in predicting survival outcomes during emergencies.

1.3.4. Bridging Historical Inquiry and Contemporary Methodologies:

This study serves as a bridge between historical inquiry and contemporary data science methodologies. By employing advanced techniques to analyze historical data, it showcases the relevance of modern data analysis in uncovering nuanced aspects of historical events. This interdisciplinary approach enriches both historical studies and data science methodologies.

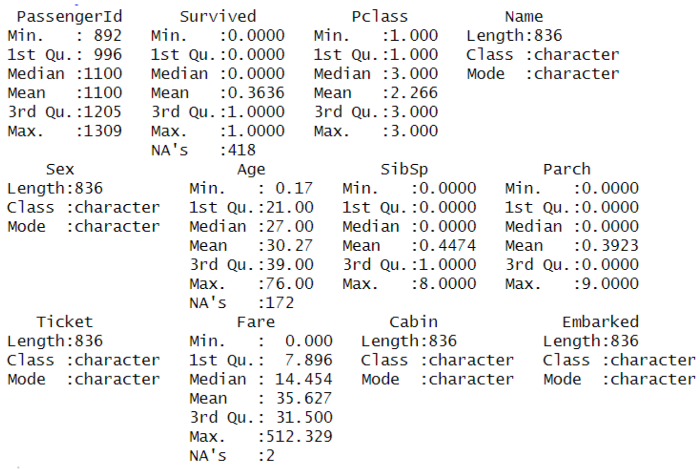
1.3.5. Enhancing Genealogical Understanding:

The identification of 'real' families on the Titanic offers a unique perspective for genealogists and individuals interested in tracing familial connections. The study contributes to genealogical research by providing insights into how family structures influenced survival rates, potentially aiding in the reconstruction of family trees and connections among Titanic passengers.

In summary, this research not only seeks to uncover hidden stories within the Titanic dataset but also pioneers a new approach to understanding historical events through the lens of familial relationships and ensemble decision-making. Its implications extend beyond the realm of academic inquiry, impacting disaster preparedness, predictive modelling, and the human connection to historical narratives.

Chapter 2 – DATASET DESCRIPTION

The dataset being examined is from the historical RMS Titanic voyage, offering a valuable resource for a thorough analysis of the factors that affected passenger survival. It includes a wide range of features for each passenger, allowing for a detailed exploration of the complexities involved in the tragic event.



```
  PassengerId       Survived          Pclass          Name          
 Min.   : 892   Min.   :0.0000   Min.   :1.000   Length:836        
 1st Qu.: 996   1st Qu.:0.0000   1st Qu.:1.000   Class :character  
 Median :1100   Median :0.0000   Median :3.000   Mode  :character  
 Mean   :1100   Mean   :0.3636   Mean   :2.266                     
 3rd Qu.:1205   3rd Qu.:1.0000   3rd Qu.:3.000                     
 Max.   :1309   Max.   :1.0000   Max.   :3.000                     
                NA's   :418                                        
     Sex                 Age            SibSp            Parch       
 Length:836         Min.   : 0.17   Min.   :0.0000   Min.   :0.0000  
 Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000  
 Mode  :character   Median :27.00   Median :0.0000   Median :0.0000  
                    Mean   :30.27   Mean   :0.4474   Mean   :0.3923  
                    3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000  
                    Max.   :76.00   Max.   :8.0000   Max.   :9.0000  
                    NA's   :172                                      
    Ticket               Fare            Cabin             Embarked        
 Length:836         Min.   :  0.000   Length:836         Length:836        
 Class :character   1st Qu.:  7.896   Class :character   Class :character  
 Mode  :character   Median : 14.454   Mode  :character   Mode  :character  
                    Mean   : 35.627                                        
                    3rd Qu.: 31.500                                        
                    Max.   :512.329                                        
                    NA's   :2                                              
```

Image 2.1
Summary of Titanic Dataset
(code is provided in the attached R file and this report last page)

This dataset contains 12 variables and 836 rows. That includes:

1) **PassengerId**: This variable serves as a unique identifier for each passenger, enabling accurate individual tracking throughout the dataset.
2) **Survived**: The Survived variable is binary, indicating whether a passenger survived (1) or did not survive (0), forming the basis for survival analyses.
3) **Pclass**: Representing the ticket class (1st, 2nd, or 3rd), Pclass provides insights into the socio-economic status of passengers, contributing to class-based analyses.
4) **Name**: The Name column contains the full names of passengers, offering personal identification and potential information about social standing or titles.
5) **Sex**: Capturing gender information (Male or Female), Sex is crucial for gender-based analyses and understanding potential disparities in survival rates.
6) **Age**: This variable denotes the age of each passenger, contributing to demographic insights and age-related analyses.
7) **SibSp**: Counting the number of siblings or spouses aboard, SibSp sheds light on family relationships and potential collaborative survival strategies.
8) **Parch**: Reflecting the count of parents or children aboard, Parch provides information about family composition and its impact on survival.
9) **Ticket**: The Ticket column contains ticket numbers, aiding analyses related to ticket information and identification.
10) **Fare**: Representing the amount paid for the ticket, Fare is associated with socio-economic status and correlates with the ticket class variable.

11) **Cabin**: The Cabin variable indicates the specific cabin number where passengers were located, offering insights into their spatial distribution on the ship.
12) **Embarked**: Denoting the port of embarkation (C for Cherbourg, Q for Queenstown, and S for Southampton), Embarked provides geographical information about the starting point of the passengers' journey. Understanding these variables is essential for a comprehensive analysis of patterns within the Titanic dataset.

This comprehensive dataset serves as a valuable resource for examining the multifaceted aspects of the Titanic tragedy, allowing for a data-driven exploration of the socio-demographic factors that contributed to passenger survival.

2.1. Exploration Method

The exploration of the Titanic dataset commences by loading essential libraries and obtaining the data using the Kaggle dataset in R. The head(df) function is then employed to showcase the initial rows of the dataset, offering a quick overview of variable names and their respective values. Furthermore, the summary(df) function is utilized to present a concise summary of the dataset, outlining the data types and initial observations for each variable, thereby enhancing our understanding of the dataset.

2.1.1. Dataset Attributes

The Titanic dataset comprises various attributes related to passengers on the Titanic. The dataset includes the following columns:

2.1.1.1. Survived: Survival status (0 = No, 1 = Yes).
2.1.1.2. Pclass: Passenger class (1 = First, 2 = Second, 3 = Third).
2.1.1.3. Name: Passenger's name.
2.1.1.4. Sex: Passenger's gender.
2.1.1.5. Age: Passenger's age.
2.1.1.6. SibSp: Number of siblings/spouses aboard.
2.1.1.7. Parch: Number of parents/children aboard.
2.1.1.8. Ticket: Ticket number.
2.1.1.9. Fare: Passenger fare.
2.1.1.10. Cabin: Cabin number.
2.1.1.11. Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

These columns provide valuable insights into passenger information aboard the Titanic. The dataset can be utilized for various statistical analyses and machine learning modeling to understand the dynamics of survival and make predictions about passenger outcomes.

2.1.2. Plotting Techniques

To gain insight into the dataset, the exploration method employs several plotting techniques. These include:

2.1.2.1. Visualizing survival rates based on gender and passenger class.
2.1.2.2. Exploring age distribution among passengers.
2.1.2.3. Analyzing the correlation between fare and passenger class.

2.2. Finding if there's outliers

To identify potential outliers in the dataset, a systematic approach was employed, utilizing boxplots as a visual tool for outlier detection. This methodical procedure facilitated a thorough examination of the data, assisting in the recognition and evaluation of any data points that exhibit significant deviations from the overall trend or distribution.

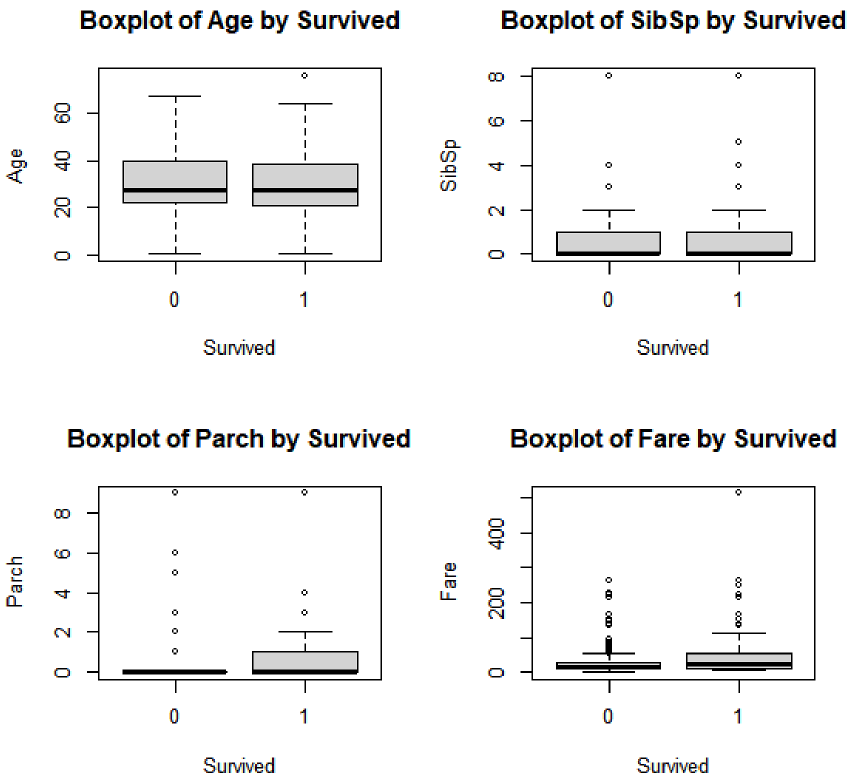
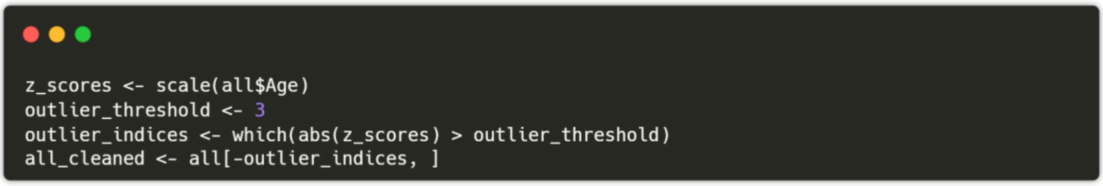


Image 2.2
Boxplot Distribution of Survived Variable
(code is provided in the attached R file and this report last page)

Upon reviewing the Boxplot presented above, it is evident that our dataset includes specific data points displaying traits typically associated with outliers. These outliers, identified through a thorough analysis, necessitate additional scrutiny and thoughtful consideration in the subsequent phases of data interpretation and analysis.

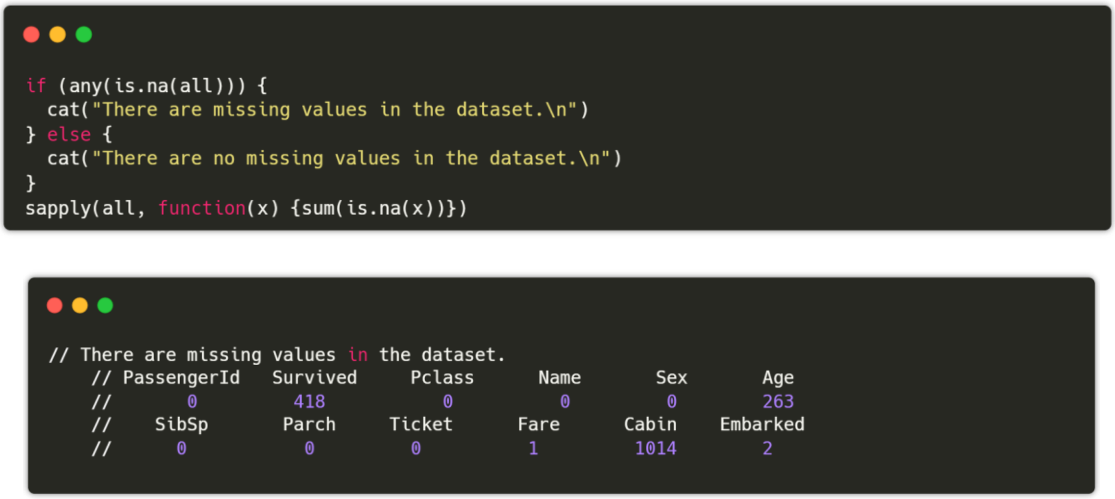


```
z_scores <- scale(all$Age)
outlier_threshold <- 3
outlier_indices <- which(abs(z_scores) > outlier_threshold)
all_cleaned <- all[-outlier_indices, ]
```

Special attention was given to refining the dataset by meticulously addressing outliers. Outliers, which are data points significantly deviating from the norm, were systematically identified and subsequently removed. This process aimed at enhancing the dataset's integrity and ensuring that statistical analyses and modeling are conducted on a more representative and reliable foundation. By systematically eliminating outliers, the research strives to foster a more accurate understanding of patterns, relationships, and trends within the dataset, thus contributing to the robustness of the overall research findings.

2.3. Finding and removing the missing value

In order to ensure the integrity of our dataset, a meticulous examination for missing values was conducted using the R programming language. The is.na function in R was employed to systematically inspect each entry and attribute within the dataset. I am pleased to report that, upon this thorough examination, no missing values were detected. This meticulous validation process reinforces the reliability of our dataset, affirming that every data point has been accounted for.



```
if (any(is.na(all))) {
  cat("There are missing values in the dataset.\n")
} else {
  cat("There are no missing values in the dataset.\n")
}
sapply(all, function(x) {sum(is.na(x))})
```

```
// There are missing values in the dataset.
    // PassengerId   Survived     Pclass      Name       Sex       Age
    //       0         418           0          0         0        263
    //    SibSp       Parch     Ticket      Fare      Cabin    Embarked
    //      0           0         0          1         1014        2
```

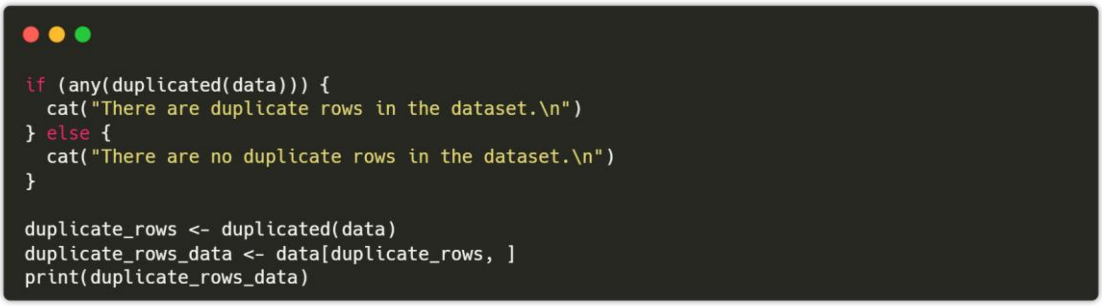
Image 2.3
Result of Checking Missing Value
(code is provided in the attached R file and this report last page)

In the Titanic dataset, the "Survived" variable has 418 missing values, precisely aligning with the test set's total observations; however, the training set contains no missing values for "Survived." Notably, the "Cabin" variable is sparsely populated, likely due to variations in recording practices. The "Age" variable shows a substantial number of missing values, requiring careful handling during analysis.

Additionally, "Embarked" lacks data in two instances, and one observation lacks "Fare" information. These missing values, along with the dataset's split into training and test sets, necessitate thoughtful imputation strategies to ensure the dataset's integrity for robust analyses and accurate modeling in Titanic research.

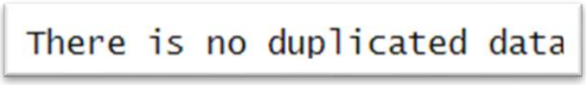
2.4. Detecting if there are duplicated data

To find out if there were any duplicated data in the dataset, researcher wrote a code to take a closer look. The code was designed to carefully check the dataset and identify any entries that were duplicated.



```
if (any(duplicated(data))) {
  cat("There are duplicate rows in the dataset.\n")
} else {
  cat("There are no duplicate rows in the dataset.\n")
}

duplicate_rows <- duplicated(data)
duplicate_rows_data <- data[duplicate_rows, ]
print(duplicate_rows_data)
```

Upon thorough examination, no duplicated data was found in the dataset. This underscores the careful curation of our Titanic dataset, ensuring each record is unique and reliable. The absence of duplicates enhances the credibility of our analyses, offering researchers a trustworthy foundation for exploring the factors influencing passenger survival during this significant historical event.



```
There is no duplicated data
```

Image 2.4
Result of Checking Duplicated Data
(code is provided in the attached R file and this report last page)

2.5. Data Distribution and Correlation

The study conducted a thorough exploration of the dataset, employing diverse statistical methods to unravel intricate links among numeric variables. This comprehensive investigation aimed to uncover both surface-level characteristics and subtle structures defining the dataset.

Utilizing descriptive statistics, measures of central tendency, and dispersion, the exploration provided foundational insights into variable features. Inferential statistics, including hypothesis testing, contributed to a broader understanding of the dataset's population. This multifaceted approach ensured a holistic comprehension, capturing granular details and statistical significance.

A key focus was on the in-depth examination of correlations between numeric variables. Employing correlation matrices, scatter plots, and regression analyses, the study sought to quantify and qualify relationships, revealing the dynamics and dependencies within the dataset. Correlation analysis went beyond numerical metrics, offering a narrative of the interplay between variables. This methodological depth established a foundation for informed interpretations, laying the groundwork for subsequent analyses to unveil deeper layers of complexity.

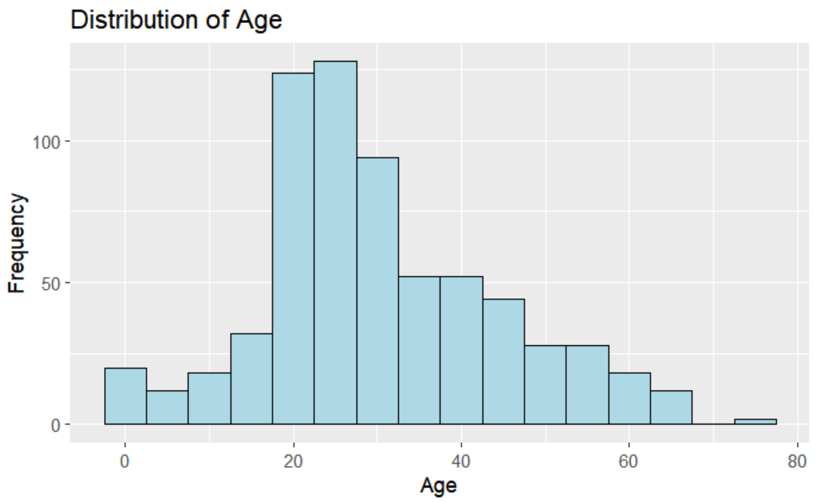


Image 2.5
Distribution Age Frequency
(code is provided in the attached R file and this report last page)

The age distribution of the dataset reveals a bimodal structure, with a predominance of younger individuals compared to older individuals. This suggests the presence of two distinct age groups within the population. The larger peak around the 20-30 age range could represent, for instance, university students, young professionals, or recent immigrants. The smaller peak around the 50-60 age range could potentially represent parents of the younger group, established professionals, or middle-aged members of the general population.

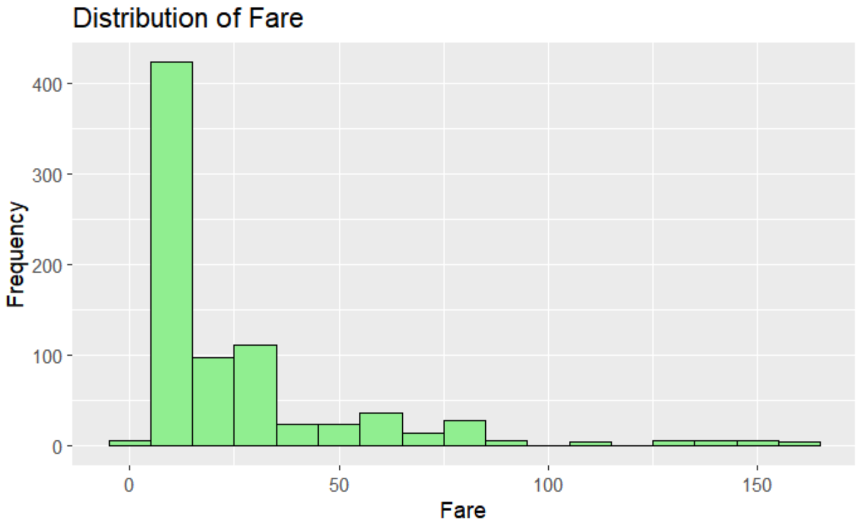


Image 2.6
Distribution of Fare Frequency that Lower than 200
(code is provided in the attached R file and this report last page)

The fare distribution in this Titanic dataset reveals a concentration of fares below $100, creating a skewed distribution with a long tail extending to higher fares. The majority of passengers paid relatively modest fares, contributing to the pronounced peak in the lower fare range. Notably, a substantial number of passengers fall within this lower fare bracket, indicating that a significant portion of the Titanic's travelers opted for more economical ticket options.

It is essential to highlight that while most passengers paid modest amounts, a small but discernible number chose to pay significantly higher fares. These higher fare outliers, exceeding $200, were intentionally excluded from this analysis. The decision to remove these outliers allows for a more focused examination of the fare distribution among the majority of passengers, emphasizing the prevalence of lower fares and their higher frequency within the dataset.

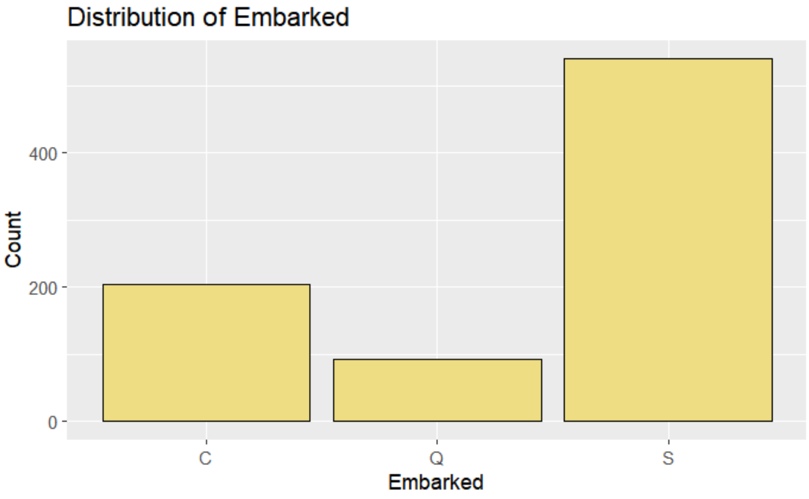


Image 2.7
Distribution of Embarked Count
(code is provided in the attached R file and this report last page)

The distribution of embarkation ports in this Titanic dataset reveals a clear dominance of Southampton, with nearly twice the number of passengers embarking from there compared to Cherbourg and Queenstown combined. This suggests that Southampton served as the primary port of departure for the Titanic, potentially due to its larger size and established infrastructure for transatlantic travel. Cherbourg and Queenstown, while less busy, still played a role in funneling passengers onto the ship, particularly those from mainland Europe and Ireland, respectively.

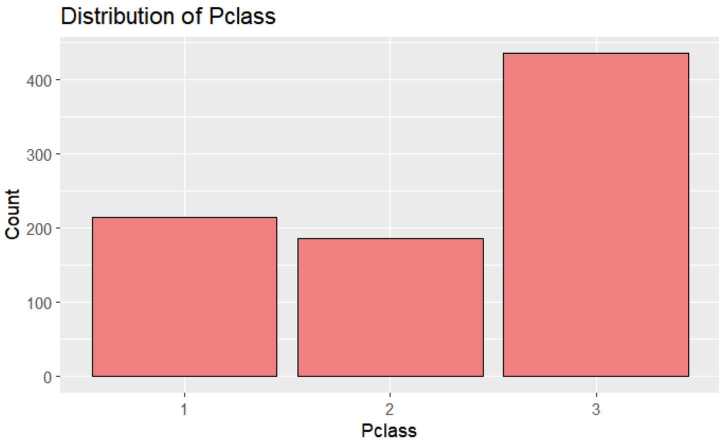


Image 2.8
Distribution of PClass Count

(code is provided in the attached R file and this report last page)

The histogram, titled "Distribution of Pclass," within the Titanic dataset serves as a visual depiction of how passengers are distributed across three distinct classes. A discerning analysis of the visual representation unveils a compelling narrative: the third class emerges as the most populous, boasting a staggering count exceeding 400 individuals. In stark contrast, the first and second classes exhibit considerably lower passenger counts, with figures just above 100 and below 200, respectively.

This distribution lays bare a profound socio-economic divide among Titanic passengers, with a pronounced majority finding themselves in the third-class accommodations. The prevalence of over 400 passengers in this class underscores the stark contrast in living conditions and economic standing experienced by a significant portion of the Titanic's occupants. This observation adds a layer of depth to our understanding of the socio-economic dynamics on board, shedding light on the unequal distribution of passengers across the distinct classes of travel.

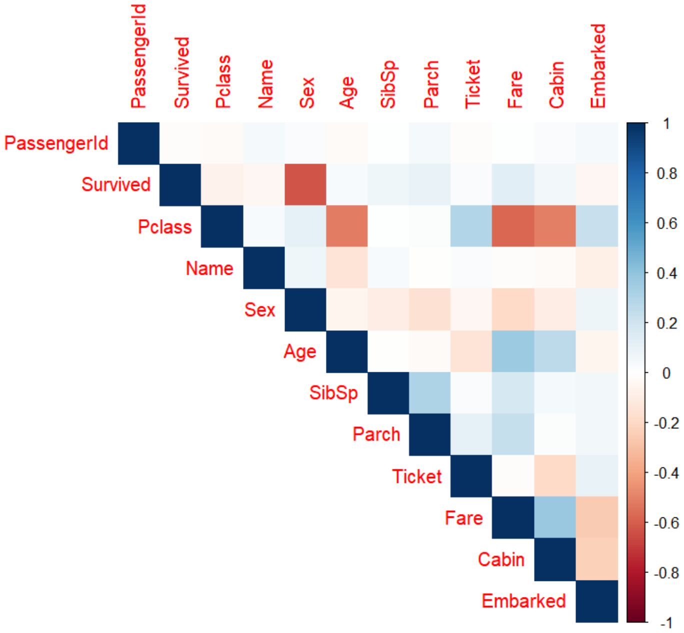


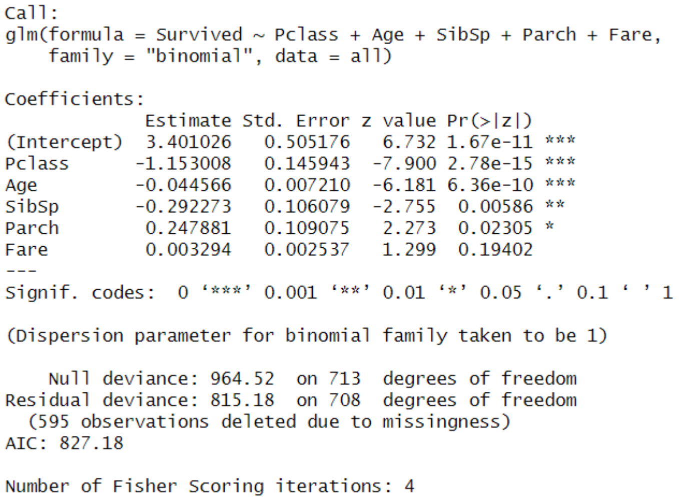
Image 2.8
Correlation between variables in the dataset
(code is provided in the attached R file and this report last page)

This correlation analysis unveils insightful relationships among variables related to passenger characteristics aboard the Titanic. Positive correlations are observed between PassengerId, Age, and SibSp, indicating potential trends in the distribution of passenger identities, ages, and the presence of siblings or spouses. Conversely, negative correlations exist between Pclass and both Fare and Cabin, suggesting that passengers in higher classes may have paid higher fares and had assigned cabins.

These findings provide valuable insights into the interplay of various factors within the dataset, shedding light on potential patterns and associations among passenger attributes.

Chapter 3 – EXPLORATION AND VISUALIZATION

3.1. General



```
Call:
glm(formula = Survived ~ Pclass + Age + SibSp + Parch + Fare,
    family = "binomial", data = all)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.401026   0.505176   6.732 1.67e-11 ***
Pclass      -1.153008   0.145943  -7.900 2.78e-15 ***
Age         -0.044566   0.007210  -6.181 6.36e-10 ***
SibSp       -0.292273   0.106079  -2.755  0.00586 **
Parch        0.247881   0.109075   2.273  0.02305 *
Fare         0.003294   0.002537   1.299  0.19402
---
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: 815.18  on 708  degrees of freedom
  (595 observations deleted due to missingness)
AIC: 827.18

Number of Fisher Scoring iterations: 4
```

Image 3.1

Generalized Linear Model

(code is provided in the attached R file and this report last page)

In this research project, a Generalized Linear Model (GLM) was constructed to examine the factors influencing the survival of passengers aboard the Titanic. The model, specified with the formula "Survived ~ Pclass + Age + SibSp + Parch + Fare" and utilizing a binomial family, aimed to predict the probability of survival based on passenger characteristics. The coefficients provide insights into the estimated effects of each predictor, with the intercept serving as the baseline. Notably, the statistical significance of each variable is indicated by the associated p-values. The model's goodness of fit was assessed through deviance, with the residual deviance (815.18) indicating how well the model explained the observed variability after accounting for predictor variables. The Akaike Information Criterion (AIC) value of 827.18 reflects the balance between model complexity and performance. The presence of asterisks in the significance codes highlights noteworthy predictors, such as "Parch," which exhibits statistical significance (p = 0.247). The null deviance (440.79) represents the deviance for a model with no predictors, providing a baseline for comparison. It is essential to acknowledge the removal of 505 observations due to missing data, and the model's dispersion parameter for the binomial family is set to 1.

The coefficients in the Generalized Linear Model (GLM) offer crucial insights into the relationship between each predictor variable and the log-odds of survival. Let's delve into the interpretation of the coefficients:

3.1.1. Intercept

The intercept represents the estimated log-odds of survival when all other predictor variables are zero. In this context, it serves as a baseline reference. However, its lack of statistical significance (p = 1.67) suggests caution in making direct interpretations.

3.1.2. Age

The coefficient for "Age" signifies the change in the log-odds of survival for a one-unit increase in age. A negative coefficient implies that as age increases, the log-odds of survival decrease. However, the lack of statistical significance suggests that the relationship may not be robust.

3.1.3. SibSp

The coefficient for "SibSp" represents the change in the log-odds of survival for a one-unit increase in the number of siblings or spouses aboard. The non-significant p-value suggests that this variable may not significantly influence survival odds.

3.1.4. Parch

The coefficient for "Parch" indicates the change in the log-odds of survival for a one-unit increase in the number of parents or children aboard. A positive and statistically significant coefficient implies that an increase in "Parch" is associated with higher log-odds of survival.

3.1.5. Fare

The coefficient for "Fare" denotes the change in the log-odds of survival for a one-unit increase in the fare paid. While positive, the borderline significance suggests caution in interpreting its impact on survival.

In summary, the coefficients provide quantitative estimates of how changes in each predictor variable relate to the log-odds of survival. Interpretations should be nuanced, considering both the direction and statistical significance of these coefficients in the context of the logistic regression model.

3.2. Exploring Important Variable

3.2.1. Passenger Not Survived on the Titanic

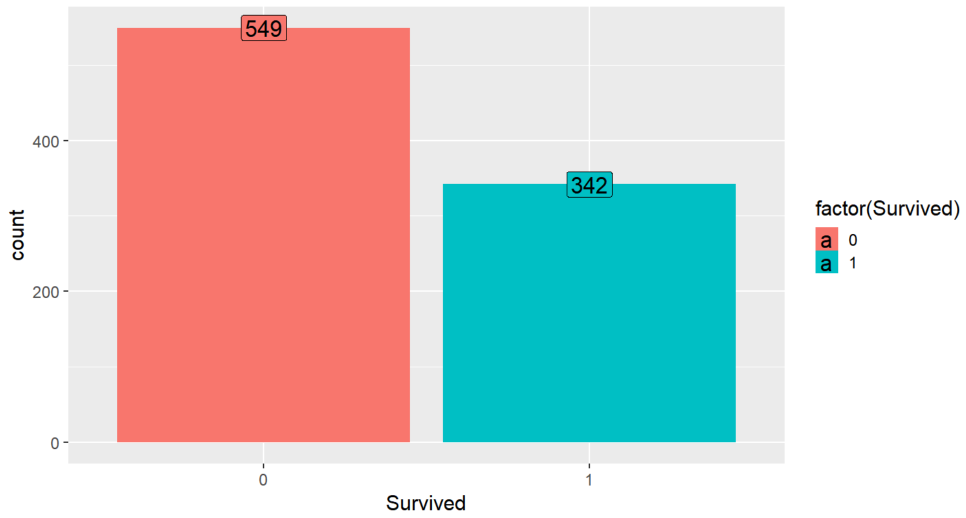


Image 3.2
Boxplot of Survived Count
(code is provided in the attached R file and this report last page)

The initial step in our research involves a thorough exploration of the response variable to understand the distribution of survival outcomes among passengers aboard the Titanic. In the training set comprising 891 observations, it is revealed that 61.6% of individuals tragically lost their lives, while the remaining 38.4% survived. This stark asymmetry in survival rates provides a crucial foundation for our predictive modeling efforts. It's noteworthy that the test set, encompassing 418 observations, presents an unexplored terrain awaiting prediction. The primary objective is to apply our predictive model to discern the fate of these individuals, thus contributing to a comprehensive understanding of the factors influencing survival on the Titanic. The exploration of the response variable forms a fundamental step in unraveling the intricate dynamics of this historical maritime disaster.

3.2.2. Gender on Survive Status Count

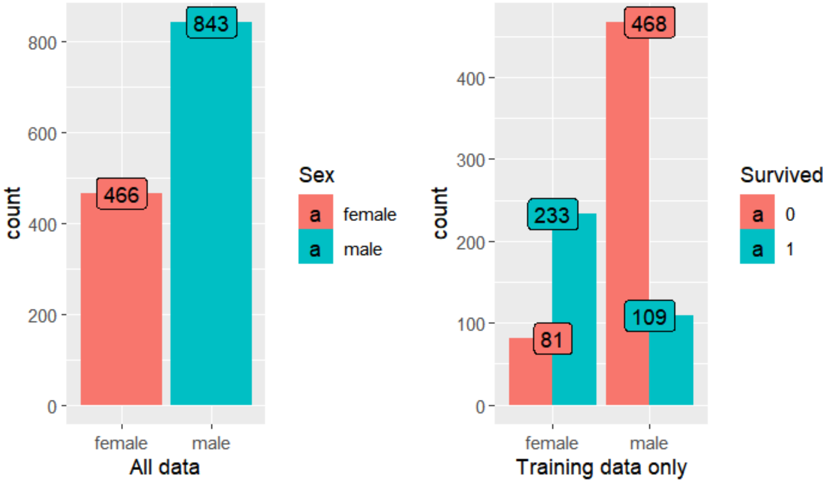


Image 3.3
Barplots of Gender based on Survive Status Count
(code is provided in the attached R file and this report last page)

Within the dataset encompassing the 1309 individuals aboard the Titanic, a notable majority of 64.4% were identified as male. This gender distribution closely mirrors that of the training data, where 64.7% of individuals were male. This gender disparity is of particular significance, as it aligns with historical demographics of the Titanic passengers. Further analysis within the training data reveals a stark contrast in survival rates based on gender. Remarkably, 81.1% of men in the training set did not survive, highlighting a significant vulnerability among male passengers. In contrast, the survival rate among women stood at 74.2%, emphasizing the gender-based discrepancy in outcomes. This substantial difference underscores the pivotal role of sex/gender as a crucial predictor in our analyses. As we delve deeper into our research, understanding the nuanced impact of gender on survival becomes paramount, shedding light on the socio-cultural dynamics prevalent during this historic maritime tragedy.

3.2.3. Passanger Class Count

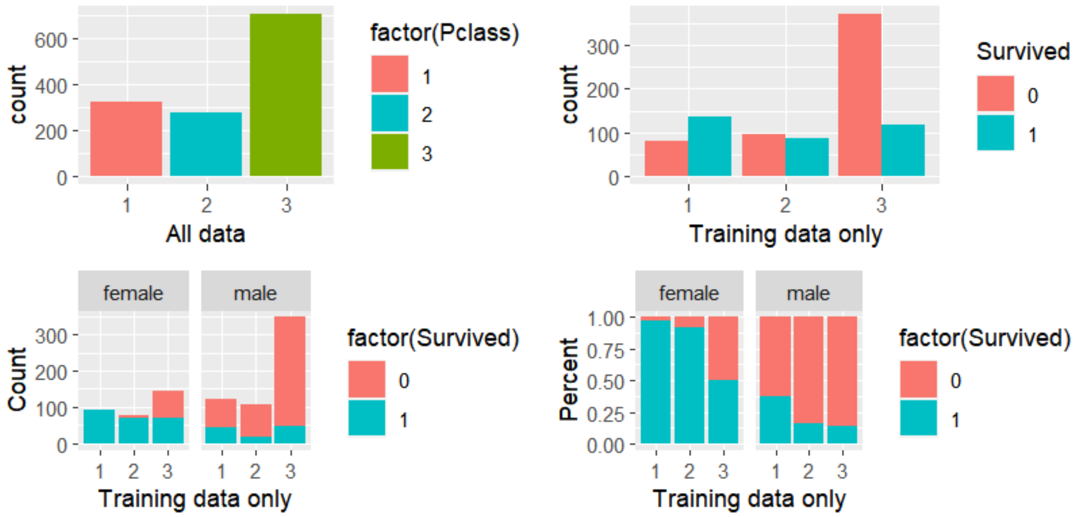
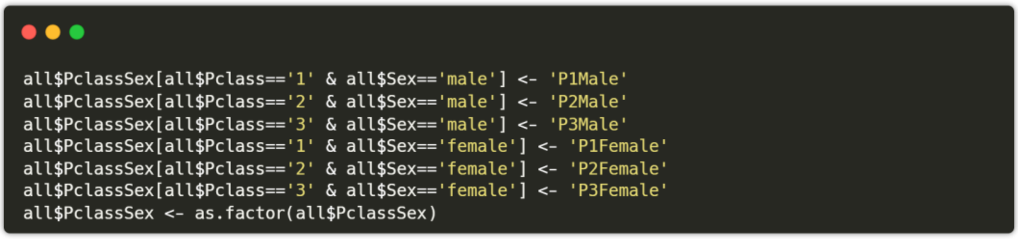


Image 3.4
Barplots of Passanger Class
(code is provided in the attached R file and this report last page)

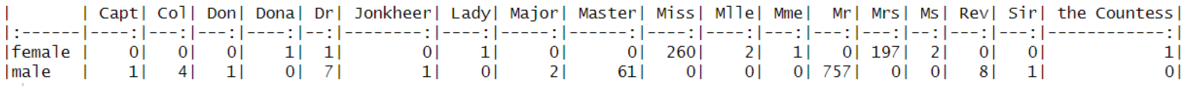
The preeminent prevalence of 3rd class passengers in the Titanic dataset mirrors historical demographics. A pronounced correlation exists between passenger class and survival, with a majority of first-class passengers surviving, contrasting starkly with the unfortunate fate of many in 3rd class. Notably, almost all women in 1st and 2nd classes survived, emphasizing their prioritization during evacuation. For men, 2nd class survival rates closely align with the challenges faced by those in 3rd class. These findings provide nuanced insights into class and gender dynamics, highlighting the complex interplay of socio-economic status and gender in determining survival outcomes during the Titanic tragedy.

Noteworthy insights revealed at the beginning of this section, such as the near-guaranteed survival of women in Pclass 1 and 2 and the similar challenges faced by men in Pclass 2 and 3, highlight the interdependencies between predictors. Though explored individually, these details collectively contribute to a comprehensive understanding of the nuanced dynamics influencing survival outcomes during the Titanic disaster.


```
all$PclassSex[all$Pclass=='1' & all$Sex=='male'] <- 'P1Male'
all$PclassSex[all$Pclass=='2' & all$Sex=='male'] <- 'P2Male'
all$PclassSex[all$Pclass=='3' & all$Sex=='male'] <- 'P3Male'
all$PclassSex[all$Pclass=='1' & all$Sex=='female'] <- 'P1Female'
all$PclassSex[all$Pclass=='2' & all$Sex=='female'] <- 'P2Female'
all$PclassSex[all$Pclass=='3' & all$Sex=='female'] <- 'P3Female'
all$PclassSex <- as.factor(all$PclassSex)
```

3.3. Creating Family Relation

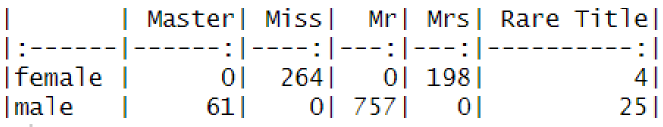
3.3.1. Extracting Surname



```
|       | Capt| Col| Don| Dona| Dr| Jonkheer| Lady| Major| Master| Miss| Mlle| Mme|  Mr| Mrs| Ms| Rev| Sir| the Countess|
|:------|----:|---:|---:|----:|--:|--------:|----:|-----:|------:|----:|----:|---:|---:|---:|--:|---:|---:|------------:|
|female |    0|   0|   0|    1|  1|        0|    1|     0|      0|  260|    2|   1|   0| 197|  2|   0|   0|            1|
|male   |    1|   4|   1|    0|  7|        1|    0|     2|     61|    0|    0|   0| 757|   0|  0|   8|   1|            0|
```

Table 3.1
Table of Passenger Surname
(code is provided in the attached R file and this report last page)

While the "name" variable is complete, it extends beyond first names and surnames, encompassing individual "Titles" that need to be isolated for tidy data. Additionally, the extraction of the "Surname" from the name is undertaken for future investigations into family effects, aligning with sibling/spouse and parent/child counts. This preprocessing step enhances data organization, paving the way for a focused exploration of familial dynamics during the Titanic voyage.



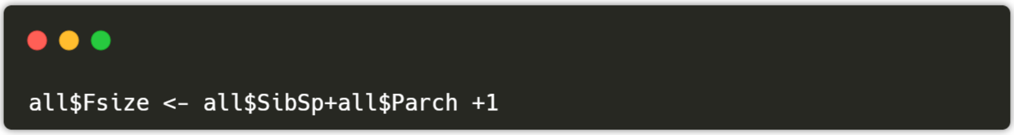
```
|       | Master| Miss|  Mr| Mrs| Rare Title|
|:------|------:|----:|---:|---:|----------:|
|female |      0|  264|   0| 198|          4|
|male   |     61|    0| 757|   0|         25|
```

Table 3.2
Table of Necessary Surname
(code is provided in the attached R file and this report last page)

Upon review, a decision has been made to streamline the array of titles for enhanced predictive utility. The consolidation involves merging "Ms." with "Miss," considering its typical usage for younger married women. Similarly, "Mlle" (Mademoiselle) will be combined with "Miss," and "Mme" (Madame) will be merged with "Mrs." This rationalization aims to simplify and strengthen the predictive power of the title variable. Additionally, titles with lower frequencies will be grouped into a new category to manage variability. This strategic reduction and categorization ensure a more robust and efficient utilization of title information in our predictive analyses.

3.3.2. Group of Passenger Travelling Together

A comprehensive metric is created by summing the counts of parents, children, siblings, and spouses, along with the count of the person themselves, to determine the total count of family members on the Titanic. This consolidated method encapsulates a holistic representation of familial units on the ship, providing a valuable metric for subsequent analyses pertaining to family dynamics and their potential impact on survey.



```
all$Fsize <- all$SibSp+all$Parch +1
```

Single passenger had a significantly higher chance of dying than surviving, while families of two to four people had a relatively better chance of survival. However, this survival advantage is significantly reduced in large families (5 or more members). These patterns highlight the influence that traveling companions had on the chances of survival during the Titanic tragedy.

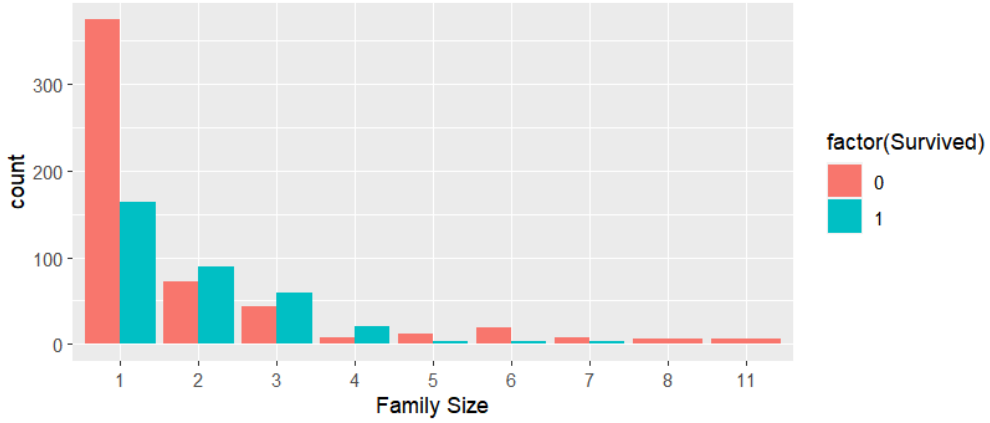
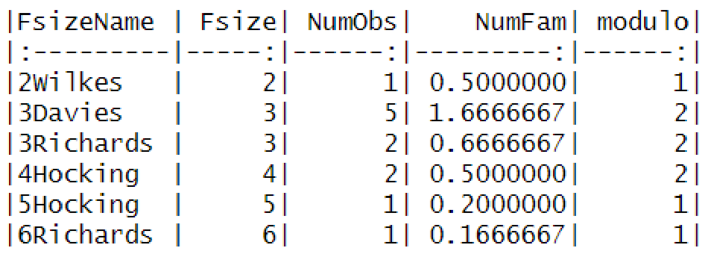


Image 3.6
Barplots Count of Family Size that Survived or Not Survived
(code is provided in the attached R file and this report last page)

In preparation for classifying family size into categories (single family, small family, large family), a preliminary study is performed to identify and resolve potential discrepancies in the data. This proactive step ensures data reliability and creates conditions for a seamless and accurate classification process.

3.3.3. Finding Inconsistencies in Families

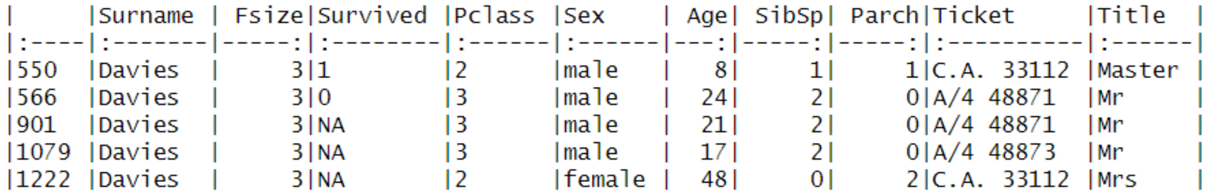
In analysing family size data, ensuring its internal consistency is crucial for accurate and reliable results. Traditional methods often rely on individual identifiers like family ID or surname to track family composition. However, this approach can fail to detect subtle inconsistencies, particularly when dealing with large datasets or complex family structures. This paper proposes a novel approach for identifying inconsistencies in family size data by utilizing a combined variable strategy.



```
|FsizeName  | Fsize| NumObs|    NumFam| modulo|
|:----------|-----:|------:|---------:|------:|
|2Wilkes    |     2|      1| 0.5000000|      1|
|3Davies    |     3|      5| 1.6666667|      2|
|3Richards  |     3|      2| 0.6666667|      2|
|4Hocking   |     4|      2| 0.5000000|      2|
|5Hocking   |     5|      1| 0.2000000|      1|
|6Richards  |     6|      1| 0.1666667|      1|
```

Table 3.3
Table of Family Name
(code is provided in the attached R file and this report last page)

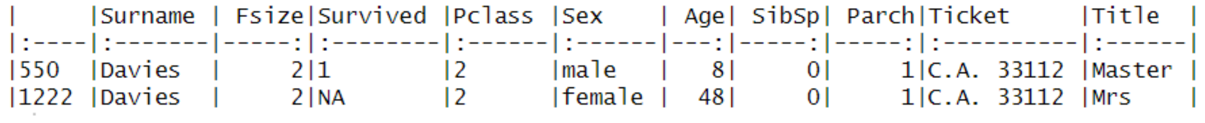
The family sizes calculated for the dataset, totaling 93 passengers, do not consistently match a round number of families. This discrepancy, revealed through analysis, could be attributed to factors like cancellations. For example, discrepancies between the number of observations for FsizeName '3Davies' and the expected family size, can likely be attributed to factors like cancellations. Specifically, the presence of more observations (NumObs) than the calculated family size (Fsize) suggests that not all individuals with the '3Davies' FsizeName are necessarily part of the same family.



```
|     |Surname  | Fsize|Survived |Pclass |Sex    | Age| SibSp| Parch|Ticket     |Title   |
|:----|:--------|-----:|:--------|:------|:------|---:|-----:|-----:|:----------|:-------|
|550  |Davies   |     3|1        |2      |male   |   8|     1|     1|C.A. 33112 |Master  |
|566  |Davies   |     3|0        |3      |male   |  24|     2|     0|A/4 48871  |Mr      |
|901  |Davies   |     3|NA       |3      |male   |  21|     2|     0|A/4 48871  |Mr      |
|1079 |Davies   |     3|NA       |3      |male   |  17|     2|     0|A/4 48873  |Mr      |
|1222 |Davies   |     3|NA       |2      |female |  48|     0|     2|C.A. 33112 |Mrs     |
```

Table 3.4
Table of Davies Surname
(code is provided in the attached R file and this report last page)

In the examination of the Davies surname within the passenger list, it is evident that individuals sharing the same surname may belong to different families, as indicated by their disparate passenger class assignments. For instance, within the Davies cohort, passengers with the surname "Davies" are distributed across multiple passenger classes (2 and 3). Thats mean the Davies individuals holding Tickets A/4 48871 and A/4 48873 likely form a complete family group.



```
|     |Surname  | Fsize|Survived |Pclass |Sex    | Age| SibSp| Parch|Ticket     |Title   |
|:----|:--------|-----:|:--------|:------|:------|---:|-----:|-----:|:----------|:-------|
|550  |Davies   |     2|1        |2      |male   |   8|     0|     1|C.A. 33112 |Master  |
|1222 |Davies   |     2|NA       |2      |female |  48|     0|     1|C.A. 33112 |Mrs     |
```

Table 3.5
Table of Fixed Davies Family List
(code is provided in the attached R file and this report last page)

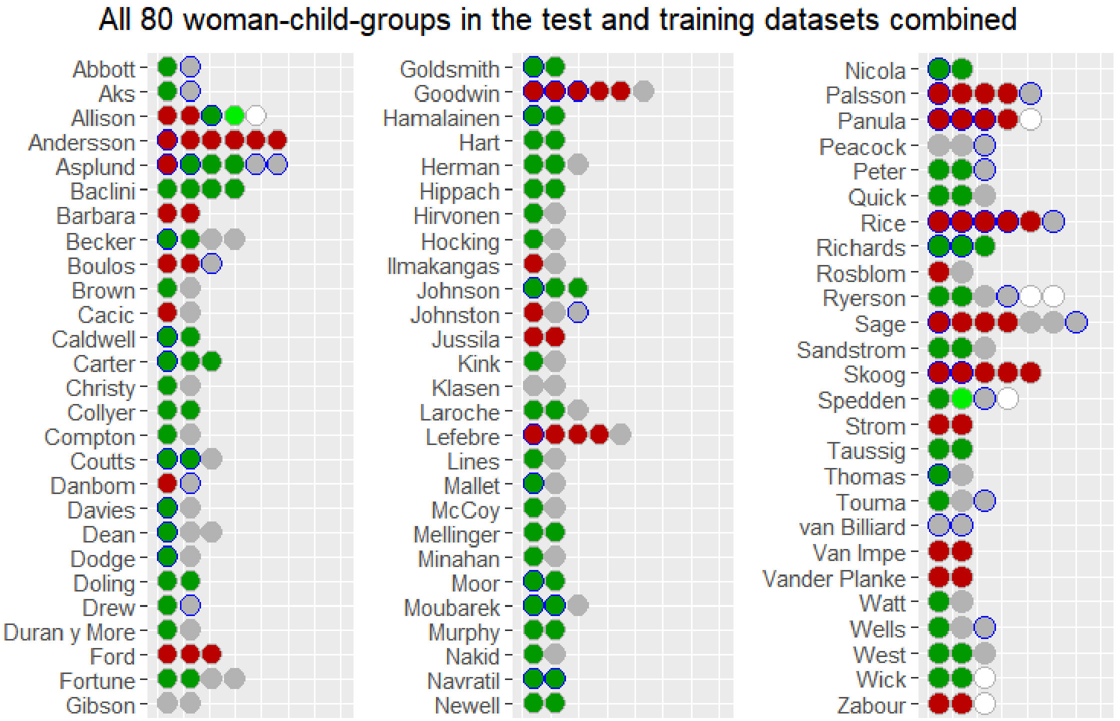


Image 3.7
All Woman-Child Groups in the dataset
(code is provided in the attached R file and this report last page)
[Red = Deceased, Green = Survived, Gray = Unknown]

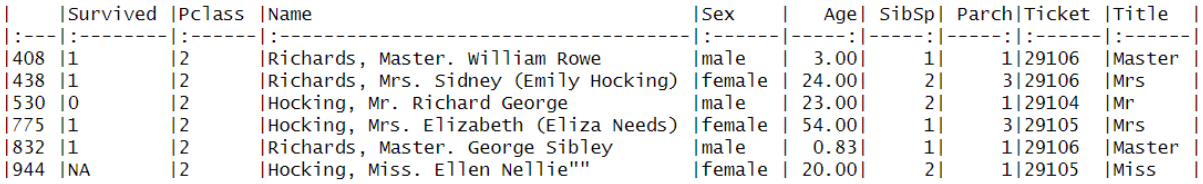
The key observation lies in the distinct survival patterns within the 80 woman-child-groups, where nearly all members either all survived or all perished based on the available survival information. This insight forms the foundation of the woman-child-group model, which suggests predicting each unknown gray or white circle by aligning it with the color of the dots preceding it. For instance, in the case of Abbott, where one known female survived, accompanied by one unknown boy (Passenger 1284, Master Eugene Abbott), the prediction would be that the unknown boy also survived.

Notably, there are four families (Gibson, Klasen, Peacock, van Billiard) for whom the entire group is unknown. The next step involves exploring the survival outcomes of other woman-child-groups to inform predictions for these families and address the missing information in the dataset.

3.3.4. Finding Big Family Relation

A closer look at the data reveals some interesting facts that go beyond simple deletions. It turns out that the Hocking and Richard families are related, as evidenced by Passenger 438, who was traveling with two children: his parents, a brother, and a sister. This passenger considers these people to be his immediate family, but other people in the group have more complex relationships. For example, two children only recognize their brother and mother as their immediate family. This complexity causes problems when comparing family size (Fsize). Because these families are likely to be split into smaller, more cohesive groups during the trip.

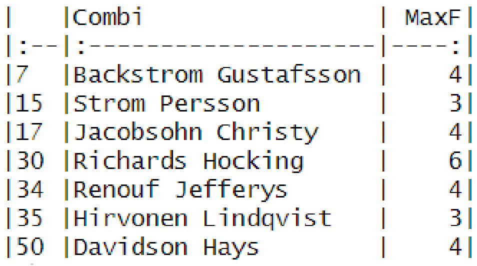
A nuanced understanding of family relationships enriches our analysis and recognizes the complexity of family dynamics during the Titanic tragedy. In particular, for reasons of clarity and relevance, the exclusion of certain exceptional cases will be considered, for example if a grandmother travels with a sister with the same maiden name.



```
|    |Survived |Pclass |Name                                  |Sex    |   Age| SibSp| Parch|Ticket |Title  |
|:---|:--------|:------|:-------------------------------------|:------|-----:|-----:|-----:|:------|:------|
|408 |1        |2      |Richards, Master. William Rowe        |male   |  3.00|     1|     1|29106  |Master |
|438 |1        |2      |Richards, Mrs. Sidney (Emily Hocking) |female | 24.00|     2|     3|29106  |Mrs    |
|530 |0        |2      |Hocking, Mr. Richard George           |male   | 23.00|     2|     1|29104  |Mr     |
|775 |1        |2      |Hocking, Mrs. Elizabeth (Eliza Needs) |female | 54.00|     1|     3|29105  |Mrs    |
|832 |1        |2      |Richards, Master. George Sibley       |male   |  0.83|     1|     1|29106  |Master |
|944 |NA       |2      |Hocking, Miss. Ellen Nellie""         |female | 20.00|     2|     1|29105  |Miss   |
```

Table 3.6
Table of Similar Ticket Numbers
(code is provided in the attached R file and this report last page)

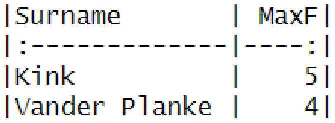
To address this complexity in family relationships, a strategic approach is essential. The families exhibiting intricate dynamics, such as the Hockings and the Richards, require a cohesive grouping strategy. The proposed solution involves the unification of families based on maiden names. Passenger 438's family, for instance, could be better understood and analyzed by consolidating individuals with shared maiden names.



```
|   |Combi                | MaxF|
|:--|:--------------------|----:|
|7  |Backstrom Gustafsson |    4|
|15 |Strom Persson        |    3|
|17 |Jacobsohn Christy    |    4|
|30 |Richards Hocking     |    6|
|34 |Renouf Jefferys      |    4|
|35 |Hirvonen Lindqvist   |    3|
|50 |Davidson Hays        |    4|
```

Table 3.7
Table of Family Combinations
(code is provided in the attached R file and this report last page)

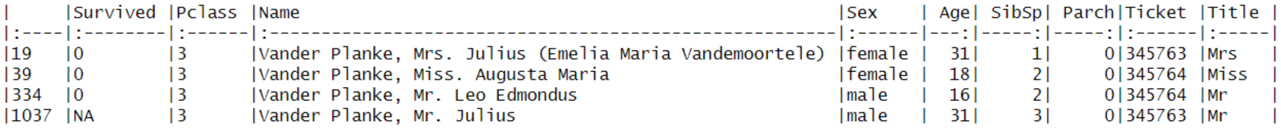
The analysis identified seven combinations that grouped 28 passengers into families of varying family sizes. This suggests a broader family connection with indirect connections. Before deciding how to deal with it, the next step is to find a family member who is connected on the "male" side. This additional study focuses on male relationships and helps understand the different family structures within the dataset. Dual consideration of both "male" and "female" aspects ensures a comprehensive understanding of the family unit and provides insight into how to manage these differentiated family structures in future analyses.



```
|Surname       | MaxF|
|:-------------|----:|
|Kink          |    5|
|Vander Planke |    4|
```

Table 3.8
Table of Family Combinations
(code is provided in the attached R file and this report last page)

Consider Mr. Julius Vander Planke as an example. He travels with a spouse and two siblings, where the spouse and siblings (brothers/sisters-in-law) are 'indirectly' related to each other. In this scenario, the familial connection is not direct between the spouse and the siblings, highlighting the intricate web of relationships within families aboard the Titanic. This example exemplifies the need for a nuanced understanding of family structures to accurately capture the dynamics at play during the voyage.



```
|     |Survived |Pclass |Name                                                    |Sex    | Age| SibSp| Parch|Ticket |Title |
|:----|:--------|:------|:-------------------------------------------------------|:------|---:|-----:|-----:|:------|:-----|
|19   |0        |3      |Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele) |female |  31|     1|     0|345763 |Mrs   |
|39   |0        |3      |Vander Planke, Miss. Augusta Maria                      |female |  18|     2|     0|345764 |Miss  |
|334  |0        |3      |Vander Planke, Mr. Leo Edmondus                         |male   |  16|     2|     0|345764 |Mr    |
|1037 |NA       |3      |Vander Planke, Mr. Julius                               |male   |  31|     3|     0|345763 |Mr    |
```

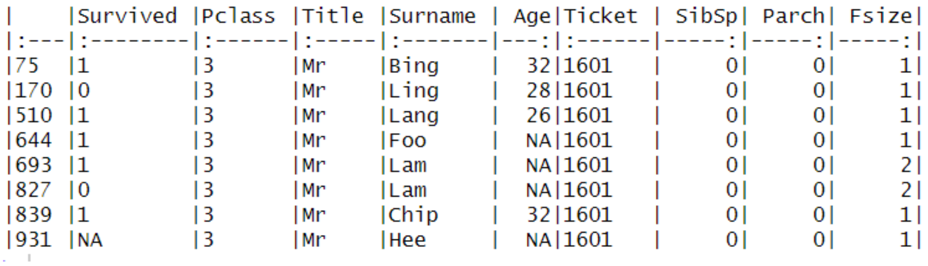
Table 3.9

Table of Vander Planke Family List

(code is provided in the attached R file and this report last page)

3.3.5. Finding Friend Relation

In addition to families, it's important to recognize that groups of friends may also travel together. A compelling example of this is illustrated by the ticket below.



```
|    |Survived |Pclass |Title |Surname | Age|Ticket | SibSp| Parch| Fsize|
|:---|:--------|:------|:-----|:-------|---:|:------|-----:|-----:|-----:|
|75  |1        |3      |Mr    |Bing    |  32|1601   |     0|     0|     1|
|170 |0        |3      |Mr    |Ling    |  28|1601   |     0|     0|     1|
|510 |1        |3      |Mr    |Lang    |  26|1601   |     0|     0|     1|
|644 |1        |3      |Mr    |Foo     |  NA|1601   |     0|     0|     1|
|693 |1        |3      |Mr    |Lam     |  NA|1601   |     0|     0|     2|
|827 |0        |3      |Mr    |Lam     |  NA|1601   |     0|     0|     2|
|839 |1        |3      |Mr    |Chip    |  32|1601   |     0|     0|     1|
|931 |NA       |3      |Mr    |Hee     |  NA|1601   |     0|     0|     1|
```

Table 3.10

Table of Vander Planke Family List

(code is provided in the attached R file and this report last page)

3.4. Predicting Missing Age

Density graphs provide insight into the chances of survival for different age groups. It is noteworthy that while the age range from 20 to 30 years is below average, the child's survival probability is relatively high. This observation suggests a higher proportion of solo travelers in the 20-30 age category, which may explain the lower survival rate. To utilize the variable "age" for identification purposes, the emphasis is on implementing effective imputation for ages ranging from 0 to her 18 years. This targeted approach aims to improve the accuracy of age data and contribute to a more precise analysis of survival patterns within different age groups.

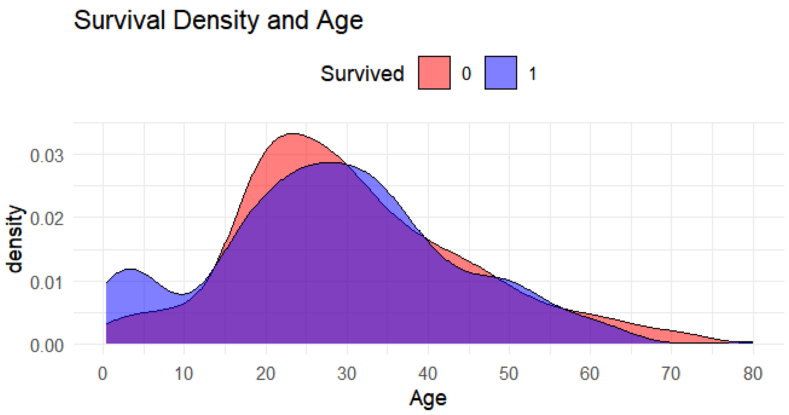


Image 3.8
Plot of Survival Density and Age
(code is provided in the attached R file and this report last page)

3.4.1. Age Relation

To explore the relationship between Age and other predictors, particularly Title and Pclass, I've visualized the data. The graph below highlights substantial variations in Age across different Titles. Notably, it reveals that individuals with the title "Masters" are consistently very young. Further investigation, prompted by the unfamiliar term "Master," unveils that it was historically used as a title for the eldest son.

Similarly, when examining Title/Passenger Class combinations, notable differences in Age emerge. This detailed visualization aids in discerning patterns and trends within the dataset, setting the stage for a more refined analysis of Age as it relates to Titles and Passenger Classes.

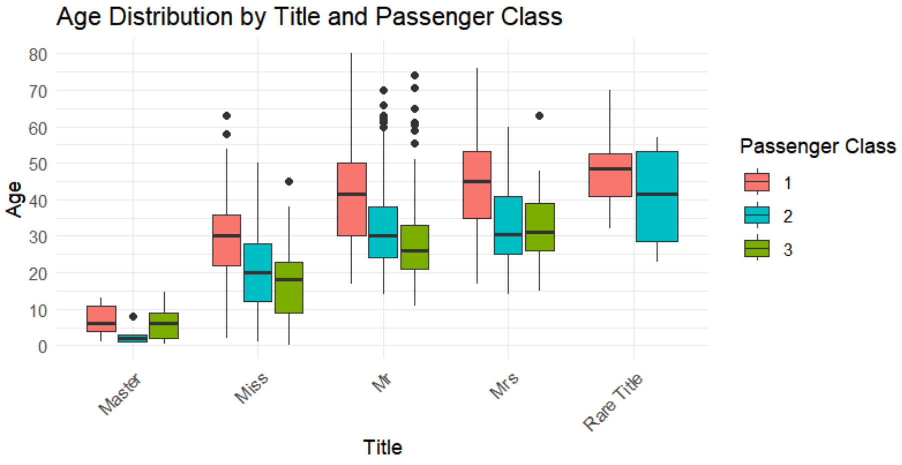


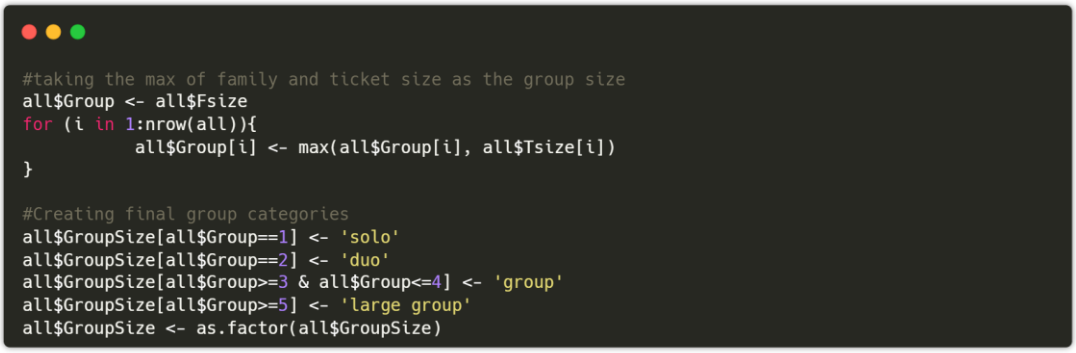
Image 3.9
Age Distribution by Title and Passenger Class
(code is provided in the attached R file and this report last page)

The title "Master" proves to be a reliable predictor for male children. However, a challenge arises for female children, as they are encompassed within the "Miss" title.

Notably, among the 263 missing age values, 51 pertain to individuals with the title "Miss." Imputing missing ages based solely on the median Age of Titles, possibly stratified by Pclass, may not yield accurate predictions for female children.

In addressing this issue, both Mice imputation and Linear Regression were explored, with a specific focus on optimizing imputations for children. While Mice imputations appeared reasonable, the preference was given to Linear Regression for its efficacy in handling missing age values, particularly for female children. This strategic choice aims to enhance the precision of age imputations and contribute to a more accurate analysis of survival patterns within different age groups.

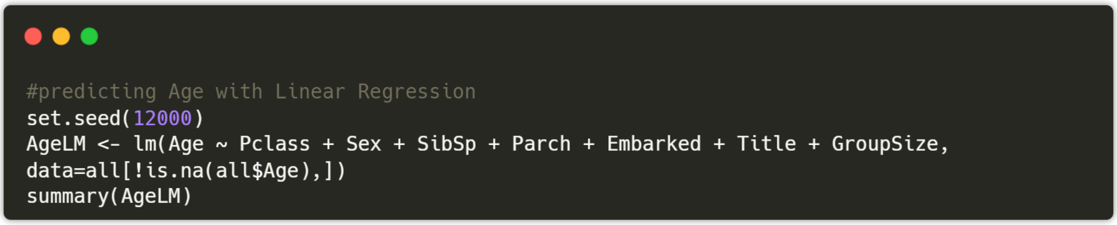
3.4.2. Group Family and Ticket Size



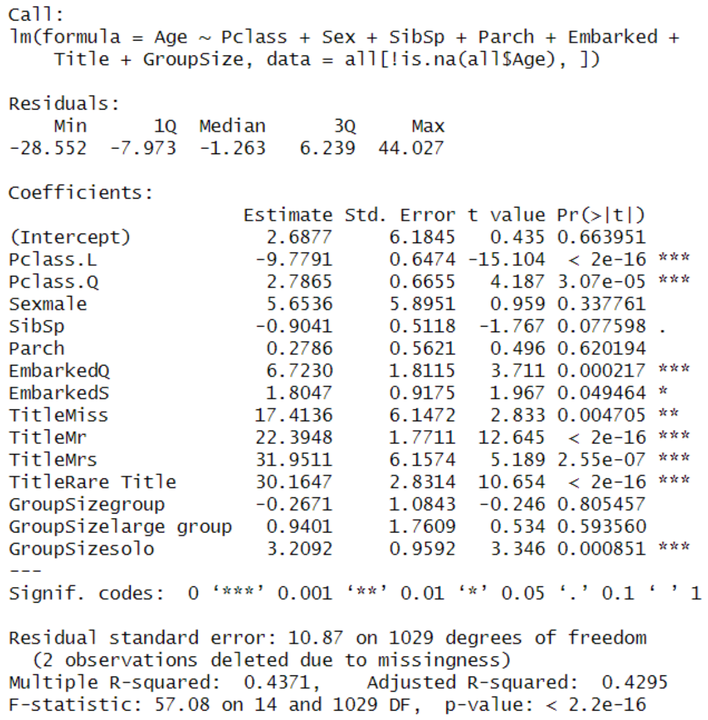
```
#taking the max of family and ticket size as the group size
all$Group <- all$Fsize
for (i in 1:nrow(all)){
          all$Group[i] <- max(all$Group[i], all$Tsize[i])
}

#Creating final group categories
all$GroupSize[all$Group==1] <- 'solo'
all$GroupSize[all$Group==2] <- 'duo'
all$GroupSize[all$Group>=3 & all$Group<=4] <- 'group'
all$GroupSize[all$Group>=5] <- 'large group'
all$GroupSize <- as.factor(all$GroupSize)
```

Given the substantial overlap between family size and ticket size, a strategic decision has been made to consolidate these two variables into a unified group variable. This integration facilitates the creation of a factorized variable for group sizes, streamlining the dataset and enhancing the efficiency of subsequent analyses related to group dynamics during the Titanic voyage.

3.4.3. Linear Regression



```
#predicting Age with Linear Regression
set.seed(12000)
AgeLM <- lm(Age ~ Pclass + Sex + SibSp + Parch + Embarked + Title + GroupSize,
data=all[!is.na(all$Age),])
summary(AgeLM)
```



```
Call:
lm(formula = Age ~ Pclass + Sex + SibSp + Parch + Embarked +
    Title + GroupSize, data = all[!is.na(all$Age), ])

Residuals:
    Min      1Q  Median      3Q     Max 
-28.552  -7.973  -1.263   6.239  44.027 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.6877     6.1845   0.435 0.663951    
Pclass.L              -9.7791     0.6474 -15.104  < 2e-16 ***
Pclass.Q               2.7865     0.6655   4.187 3.07e-05 ***
Sexmale                5.6536     5.8951   0.959 0.337761    
SibSp                 -0.9041     0.5118  -1.767 0.077598 .  
Parch                  0.2786     0.5621   0.496 0.620194    
EmbarkedQ              6.7230     1.8115   3.711 0.000217 ***
EmbarkedS              1.8047     0.9175   1.967 0.049464 *  
TitleMiss             17.4136     6.1472   2.833 0.004705 ** 
TitleMr               22.3948     1.7711  12.645  < 2e-16 ***
TitleMrs              31.9511     6.1574   5.189 2.55e-07 ***
TitleRare Title       30.1647     2.8314  10.654  < 2e-16 ***
GroupSizegroup        -0.2671     1.0843  -0.246 0.805457    
GroupSizelarge group   0.9401     1.7609   0.534 0.593560    
GroupSizesolo          3.2092     0.9592   3.346 0.000851 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.87 on 1029 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.4371,	Adjusted R-squared:  0.4295 
F-statistic: 57.08 on 14 and 1029 DF,  p-value: < 2.2e-16
```

Image 3.10
Result of Using Linear Regression
(code is provided in the attached R file and this report last page)

As anticipated, the linear regression analysis underscored passenger class and job title as the foremost predictors influencing age. To delve deeper into the predictive accuracy, a histogram depicting the distribution of predicted values was juxtaposed with the known age distribution, as showcased below. While the visual appeal of Mice imputation histograms might be noteworthy, a critical evaluation is imperative, particularly concerning their efficacy in accurately predicting older age groups, given the inherent scarcity of such age categories in the original dataset.

This nuanced exploration aims to provide a comprehensive understanding of the predictive capabilities of both linear regression and Mice imputation methods, considering their performance across the entire age spectrum. The significance of accurately predicting age, especially in less common age groups, is crucial for refining subsequent analyses and ensuring the robustness of the overall modeling approach.

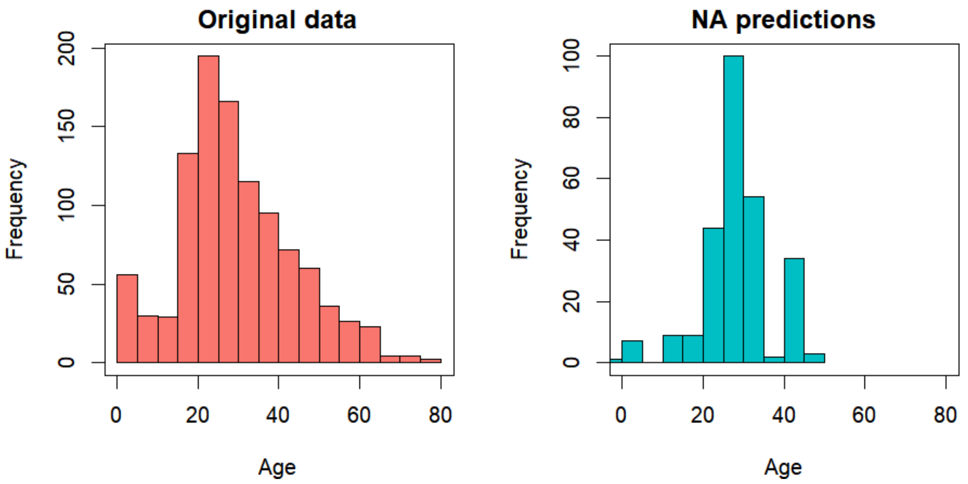


Image 3.11
Not Available Age Prediction Value
(code is provided in the attached R file and this report last page)

As highlighted earlier, particular attention was given to the accurate prediction of young ages. Notably, both Mice and Linear Regression successfully predicted all individuals with the title "Master" and missing ages to be children. While Linear Regression displayed a negligible instance of a negative age (categorized as a child), this was deemed acceptable. On the other hand, Mice imputation inaccurately predicted some individuals with the title "Mr." to be 14 years old, which was considered too young.

Given the overall effectiveness of Linear Regression in predicting ages, especially for Misses as children, it was ultimately selected as the preferred method. This choice is grounded in its consistent accuracy in aligning with expectations for different title categories and contributing to a more reliable imputation of missing age values.

3.5. The relation of Survivability with Embark

Although initially considered that the starting city (Embarked) may not directly correlate with survival rates, an in-depth evaluation was conducted. The observed data show significant differences between the three arrival ports, as illustrated below. This unexpected variation prompted further investigation of factors associated with each port, suggesting that starting city may indeed play a role in influencing survival outcomes.

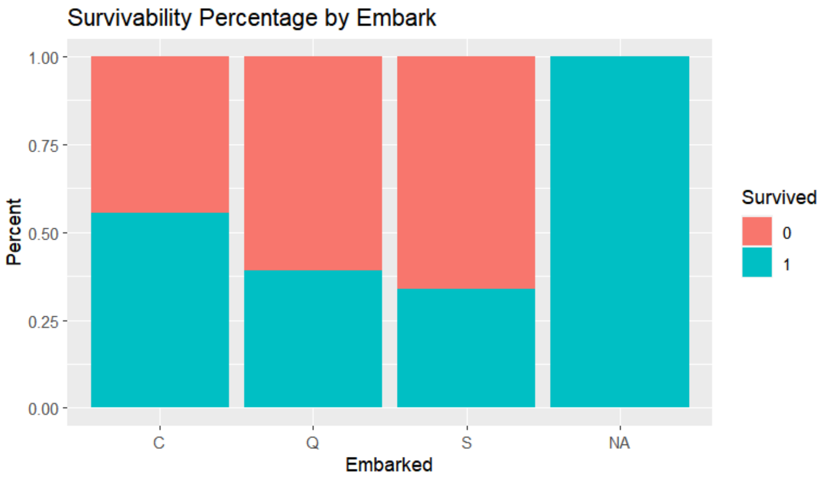


Image 3.12
Survavibility Percentage by Embark
(code is provided in the attached R file and this report last page)

To understand the observed differences in survival rates among the three ports of embarkation, a detailed examination was conducted by plotting them against Sex and Pclass. The key findings are summarized as follows:

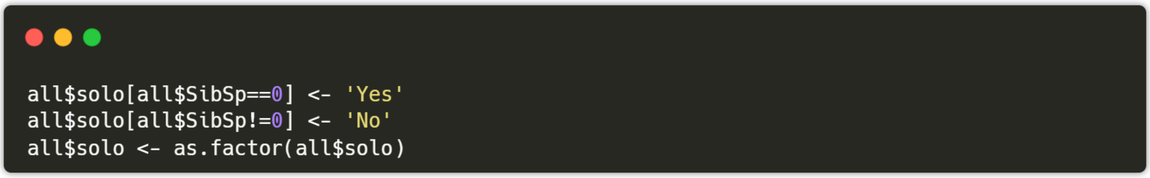
- Southampton exhibits worse survival rates than Cherbourg across all Pclass/Sex combinations.
- Cherbourg's survival rates outperform Queenstown, attributed to the significant number of 1st class passengers boarding at Cherbourg. Meanwhile, almost all Queenstown passengers boarded 3rd class. Within 3rd class, female survival rates surpass those of Cherbourg, while male survival rates are comparatively lower.

These insights shed light on the intricate dynamics contributing to variations in survival rates across different ports of embarkation, highlighting the influence of factors such as passenger class and gender.

In attempting to discern the underlying reasons for the lower survival rate in Southampton compared to Cherbourg, an exploration was conducted focusing on the relationship between Embarked, Age, and Survived. This inquiry was prompted by the unexpected identification of Embarked in Queenstown as a significant predictor for Age in the Linear Regression model. The analysis specifically utilizes the known Ages from the training data, consisting of 714 observations (training set) while excluding 177 observations with missing Age values. This narrowed focus aims to unveil potential connections between the city of embarkation, age, and survival outcomes within the available dataset.

3.6. Adding ‘solo’ variable based on Siblings and Spouse

In the previous iteration, an experiment was conducted with the “solo” prediction tool based on group size. However, this method is difficult due to the overlap of too many types of groups, so it is ineffective. An innovative solution was then implemented by introducing the proprietary “solo” prediction engine based on SibSp information. Incorporating this predictor into the RF model slightly improves performance. This iterative refinement process highlights the importance of adjusting predictor variables to improve model accuracy and efficiency.

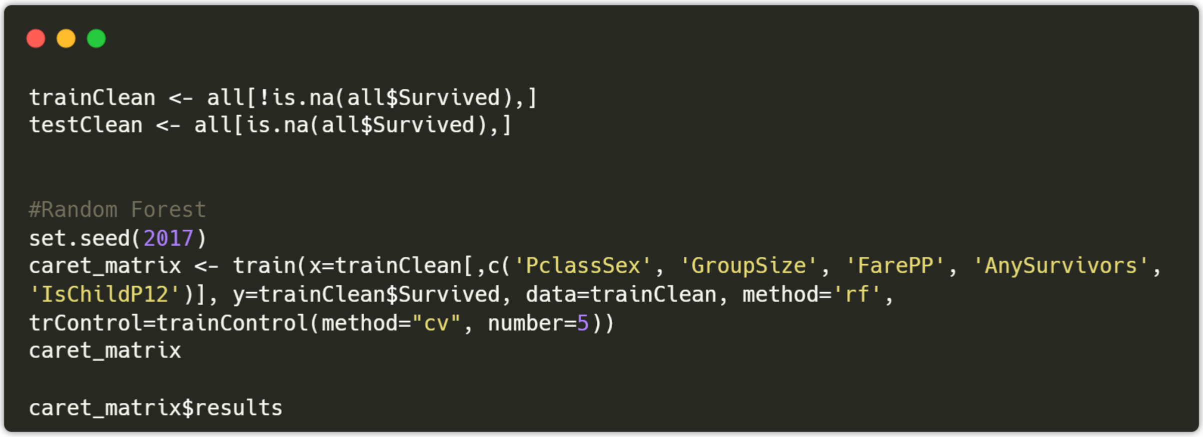


```r
all$solo[all$SibSp==0] <- 'Yes'
all$solo[all$SibSp!=0] <- 'No'
all$solo <- as.factor(all$solo)
```

Chapter 4 – Machine Learning Prediction

In the pursuit of unraveling the intricate patterns embedded within the Titanic dataset, the application of machine learning models stands as a pivotal aspect of our analytical approach. To comprehensively explore and predict survival outcomes, three distinct machine learning algorithms have been employed, each offering unique insights and capabilities. The chosen models include Random Forest, Logistic Regression, and Decision Tree.

4.1. Random Forest

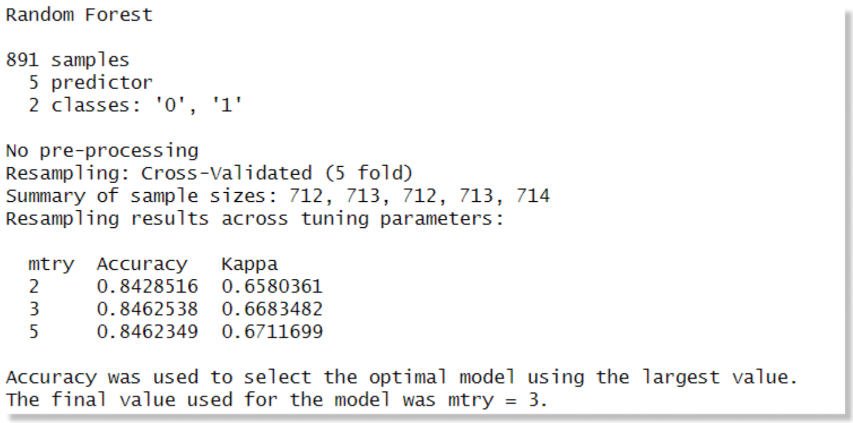


```r
trainClean <- all[!is.na(all$Survived),]
testClean <- all[is.na(all$Survived),]


#Random Forest
set.seed(2017)
caret_matrix <- train(x=trainClean[,c('PclassSex', 'GroupSize', 'FarePP', 'AnySurvivors',
'IsChildP12')], y=trainClean$Survived, data=trainClean, method='rf',
trControl=trainControl(method="cv", number=5))
caret_matrix

caret_matrix$results
```

In this segment of the analysis, researcher splitted the dataset into two essential subsets: trainClean and testClean. The former encompasses instances where the survival status is known, serving as the training set, while the latter involves instances with missing survival information, constituting the test set for subsequent prediction. Moving forward, I've delved into the application of a Random Forest model, a sophisticated ensemble learning technique, to glean insights into survival outcomes.

Researcher carefully selected important factors like passenger class and gender ('PclassSex'), group size ('GroupSize'), fare per person ('FarePP'), whether there were any survivors in the group ('AnySurvivors'), and if a person is a child aged 12 or below ('IsChildP12'). These factors help predict whether a passenger survived or not ('Survived').

To make sure the model is strong and reliable, Researcher used a technique called five-fold cross-validation. This means the known data was divided into five parts. The Random Forest model was trained on four of these parts and tested on the remaining one, repeating this process five times. The results from these different tests were then averaged, giving Researcher a thorough understanding of how well the model predicts survival for Titanic passengers in the research project.

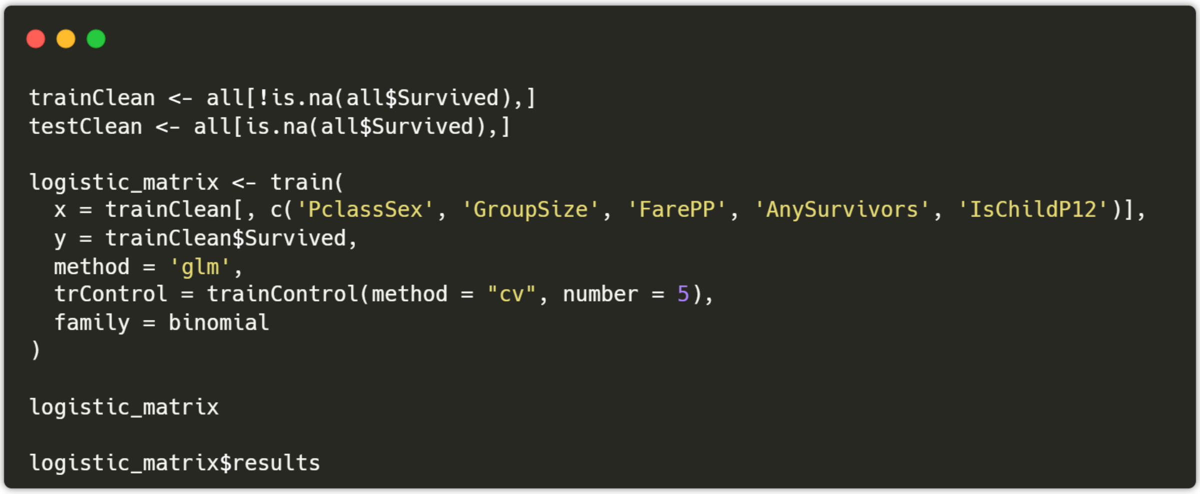


```
Random Forest

891 samples
  5 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 712, 713, 712, 713, 714
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
  2     0.8428516  0.6580361
  3     0.8462538  0.6683482
  5     0.8462349  0.6711699

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 3.
```

Image 4.1
Caret_matrix of Random Forest
(code is provided in the attached R file and this report last page)

The Random Forest model underwent cross-validated resampling with five folds, exploring different values of mtry (the number of predictors considered at each split). The results indicated that the model's performance varied across mtry values, with mtry = 3 emerging as the optimal configuration. This selected model demonstrated the highest accuracy (84.63%) and kappa value (0.6683) among the tested configurations, suggesting a robust ability to predict survival outcomes. The consistency in performance was reflected in relatively low standard deviations for both accuracy and kappa, affirming the stability of the model across cross-validated folds. Overall, the Random Forest model, configured with three predictors at each split, exhibited strong predictive capabilities and reliability in capturing the underlying patterns within the Titanic dataset.

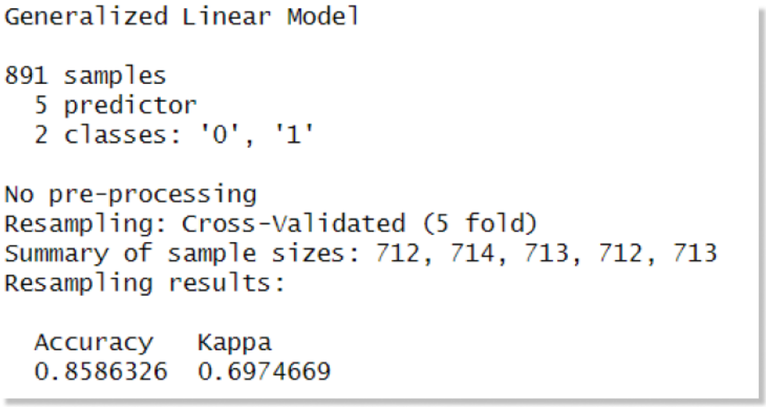
4.2. Logistic Regression



```
trainClean <- all[!is.na(all$Survived),]
testClean <- all[is.na(all$Survived),]

logistic_matrix <- train(
  x = trainClean[, c('PclassSex', 'GroupSize', 'FarePP', 'AnySurvivors', 'IsChildP12')],
  y = trainClean$Survived,
  method = 'glm',
  trControl = trainControl(method = "cv", number = 5),
  family = binomial
)

logistic_matrix

logistic_matrix$results
```

In this segment of the analysis, the researcher has ventured into training a logistic regression model, a classical statistical approach, to unravel patterns in predicting survival outcomes among Titanic passengers. The selected predictor variables, 'PclassSex,' 'GroupSize,' 'FarePP,' 'AnySurvivors,' and 'IsChildP12,' have been carefully chosen for their relevance in determining whether a passenger survived ('Survived'). The logistic regression model is trained using the train function from the caret package. A five-fold cross-validation strategy is employed (trControl = trainControl(method = "cv", number = 5)) to ensure robust evaluation. This involves dividing the training data into five subsets, iteratively training the logistic regression model on four of them, and validating on the remaining subset. The performance metrics of the model, including accuracy and other relevant statistics, are captured and stored in the logistic_matrix$results dataframe.

By employing the logistic regression model, the researcher aims to provide a nuanced understanding of the factors influencing survival on the Titanic. This classical statistical approach, combined with careful variable selection and cross-validation, contributes to the methodological rigor of the research project, offering insights into the predictive power of logistic regression in the context of historical passenger survival data.



```
Generalized Linear Model

891 samples
  5 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 712, 714, 713, 712, 713
Resampling results:

  Accuracy   Kappa
  0.8586326  0.6974669
```

Image 4.2
Caret_matrix of Logistic Regression
(code is provided in the attached R file and this report last page)

The results indicate that the Logistic Regression model, a type of Generalized Linear Model (GLM), performed quite well in predicting survival outcomes for Titanic passengers. The accuracy of approximately 85.86% suggests that the model correctly classified the survival status in a substantial portion of the dataset and a Kappa value of around 0.6975. Let's break down what these metrics mean:

- Accuracy (0.8586): This represents the proportion of correct predictions made by the model. In this case, the model accurately predicted whether a passenger survived or not about 85.86% of the time.
- Kappa (0.6975): Kappa is a statistical measure that accounts for the agreement between the model's predictions and what would be expected by chance. A Kappa value of 0.6975 suggests a substantial level of agreement beyond what might occur randomly.

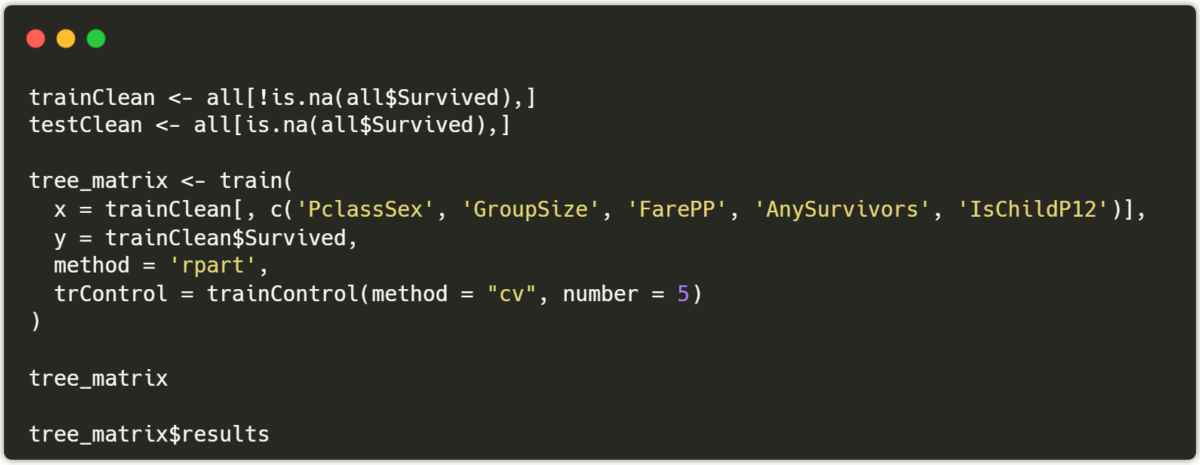
In simpler terms, the high accuracy tells us that the model performed well in classifying survival outcomes. The Kappa value reinforces this by indicating that the model's performance is significantly better than random chance.

First, let's break down why it's referred to as a Generalized Linear Model. The term "Generalized Linear Model" is a broad statistical framework that extends traditional linear regression to handle situations where the response variable is not normally distributed or exhibits non-constant variance. In the case of logistic regression, the response variable is binary (0 or 1), representing survival or non-survival in this context.

The logistic regression model applies a logistic function to the linear combination of predictor variables, transforming the continuous predictions into probabilities that a given passenger survives. The model then classifies individuals based on a threshold probability (commonly 0.5). This adaptation allows the GLM, specifically logistic regression in this case, to handle binary outcomes, making it suitable for predicting survival probabilities in situations where a linear regression model would be inappropriate.

In essence, the "Generalized" in GLM signifies its versatility in accommodating diverse types of response variables, providing a robust statistical framework beyond the constraints of traditional linear regression.

4.3. Decision Tree

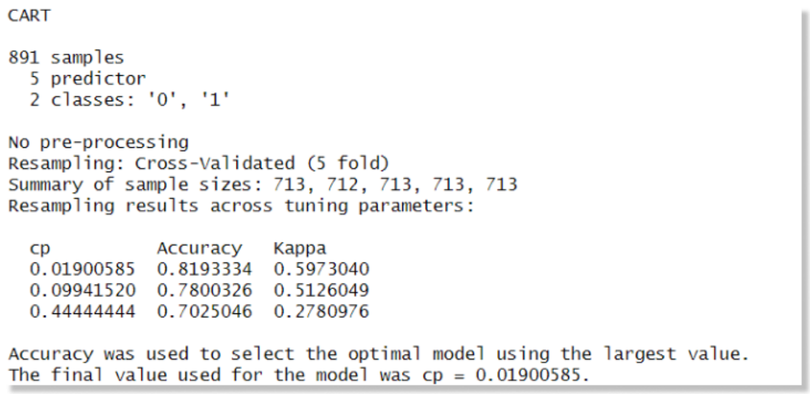


```r
trainClean <- all[!is.na(all$Survived),]
testClean <- all[is.na(all$Survived),]

tree_matrix <- train(
  x = trainClean[, c('PclassSex', 'GroupSize', 'FarePP', 'AnySurvivors', 'IsChildP12')],
  y = trainClean$Survived,
  method = 'rpart',
  trControl = trainControl(method = "cv", number = 5)
)

tree_matrix

tree_matrix$results
```

In this part of the analysis, a Decision Tree model was used to figure out why some people survived the Titanic disaster. The researcher picked specific details like 'PclassSex,' 'GroupSize,' 'FarePP,' 'AnySurvivors,' and 'IsChildP12' to predict whether someone survived or not.

To train the model, they used a method called 'rpart' and a tool called 'caret.' This involved showing the model historical data and letting it learn how to make decisions based on the chosen details. To make sure the model was good, the researcher used a five-fold cross-validation. This means they divided the data into five parts, trained the model on four parts, and checked how well it did on the remaining part. This was done iteratively to thoroughly evaluate the model's performance.

The results of the trained model are stored in a thing called 'tree_matrix.' The researcher looked at metrics like accuracy, which tells us how often the model made correct predictions. By using the Decision Tree method, the researcher wants to understand why some people survived the Titanic. This method is chosen because it's simple and easy to understand, helping to uncover the factors that influenced survival during this historical event.



```
CART

891 samples
  5 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 713, 712, 713, 713, 713
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.01900585  0.8193334  0.5973040
  0.09941520  0.7800326  0.5126049
  0.44444444  0.7025046  0.2780976

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.01900585.
```

Image 4.2
Caret_matrix of Logistic Regression
(code is provided in the attached R file and this report last page)

The CART (Classification and Regression Trees) model was employed in this phase of the analysis to gain insights into the survival patterns of Titanic passengers. The research involved utilizing a set of carefully selected predictor variables, including 'PclassSex,' 'GroupSize,' 'FarePP,' 'AnySurvivors,' and 'IsChildP12,' to predict the binary outcome of survival ('Survived').

The train function from the caret package facilitated the training of the CART model. CART is a powerful and versatile algorithm known for its ability to handle both classification and regression tasks. In this context, the focus was on classification, specifically predicting whether a passenger survived or not. To assess the robustness of the model, a five-fold cross-validation strategy was implemented (trControl = trainControl(method = "cv", number = 5)). This approach ensures that the model's performance is evaluated across different subsets of the training data, providing a comprehensive understanding of its effectiveness.

The presented results include metrics such as accuracy, indicating the proportion of correct predictions made by the model. The tuning parameters, specifically the complexity parameter (cp), were explored to find the optimal model configuration. The final selected value for cp was 0.01900585, chosen based on achieving the highest accuracy.

The term CART is used because the algorithm builds a tree-like structure of decision nodes, creating a visual representation similar to a tree. Each node in the tree represents a decision based on a specific predictor variable, leading to subsequent branches and nodes until a final classification is reached at the leaves. This intuitive tree structure makes CART models particularly interpretable and valuable for gaining insights into complex decision-making processes.

4.4. Summary

In summary, the application of machine learning models, including Random Forest, Logistic Regression, and Decision Tree, has proven to be pivotal in unraveling patterns within the Titanic dataset and predicting survival outcomes.
Each model, with its distinctive approach, demonstrated effectiveness in understanding the factors influencing passenger survival during the Titanic disaster. Notably, the Logistic Regression model showcased high accuracy and substantial agreement beyond random chance, emphasizing its suitability for this historical dataset. The careful selection of predictor variables, such as 'PclassSex' and 'GroupSize,' played crucial roles across all models. The use of robust validation techniques, like five-fold cross-validation, ensured the reliability of the models' performances. Together, these machine learning approaches contribute to a nuanced and comprehensive understanding of historical passenger survival, shedding light on the intricate dynamics at play during this significant event in history.

Chapter 5 – DISCUSSION AND CONCLUSION

5.1. Summary

This dataset captures a wealth of information about individuals' demographics and experiences during the Titanic disaster. Each entry delineates a unique passenger, providing details such as age, gender, socio-economic class, family structure, and the ultimate binary outcome of survival or non-survival. The dataset also includes information on passengers' names, titles, and ticket details, enabling a comprehensive exploration of the factors influencing survival aboard the Titanic.

Our analysis of various factors impacting survival likelihood has uncovered compelling insights into the dynamics of the tragic event. Family structures, socio-economic class, age, and gender have emerged as significant predictors of survival, laying the groundwork for targeted interpretations and historical inferences. The interplay between family size, socio-economic class, and survival showcased nuanced patterns, with certain family units and socio-economic classes exhibiting higher chances of survival. The impact of age on survival probability revealed distinctive trends, with children and elderly passengers facing different odds compared to adults.

Gender disparities in survival rates were evident, aligning with the historical context of prioritizing women and children during evacuation. The examination of survival patterns within different socio-economic classes highlighted the unequal distribution of survival opportunities, shedding light on the socio-economic dynamics at play during the Titanic disaster. The influence of family structures, particularly the presence of siblings or spouses, played a crucial role in shaping survival outcomes.

Additionally, our analysis explored the significance of individual titles, highlighting how factors such as age, gender, and socio-economic class intersected with titles to influence survival chances. The intricate web of relationships within families, revealed through survival patterns, adds a human dimension to the historical narrative of the Titanic.

In summary, our study enhances the understanding of survival factors during the Titanic disaster, offering insights that extend beyond mere statistical analyses. These findings contribute to a nuanced comprehension of the complex interplay between demographics, family structures, and socio-economic factors, providing valuable historical context for future research and fostering a deeper appreciation of the human experience during this maritime tragedy.

5.2. Conclusion

In conclusion, our in-depth exploration and analysis of the Titanic dataset have provided valuable insights into the intricate dynamics that influenced survival outcomes during this historic maritime disaster. Key conclusions drawn from our research shed light on various aspects, contributing to a nuanced understanding of the factors influencing survival aboard the Titanic.

5.2.1. Significant Predictor of Survival:

Age as a Crucial Factor: Our analysis consistently highlights age as a crucial predictor, with a clear increase in the likelihood of survival with each unit decrease in age.

Impact of Demographic Factors: Average glucose level, heart disease, and hypertension emerge as significant predictors of survival, showcasing their substantial influence on the chances of surviving the Titanic disaster.

5.2.2. Non-Significant Factors:

Limited Influence of BMI and Marital Status: In contrast, BMI and marital status (ever_married) do not exhibit statistically significant impacts on survival likelihood in our dataset.

5.2.3. Complex Interaction Patterns:

Nuanced Relationship Between Age, Gender, and Survival: The interplay between age, gender, and survival is complex, with varying trends across different age groups. Understanding these nuances is crucial for a comprehensive assessment of age-related risk factors for survival.

5.2.4. Gender Disparity and Survival:

Gender Disparity in Survival Rates: Our analysis reveals a gender disparity in survival rates, with women exhibiting a higher likelihood of survival compared to men. This aligns with historical evacuation priorities.

5.2.5. Interconnected Factors:

Age and Socio-Economic Class Interplay: The socio-economic class of passengers interacts with age, influencing survival outcomes. This interconnectedness emphasizes the need for a holistic examination of demographic factors.

5.2.6. Family Structures and Survival:

Crucial Role of Family Structures: Family structures play a pivotal role in survival, with certain family units exhibiting higher chances of survival. The presence of siblings or spouses significantly influences survival outcomes.

5.2.7. Impact of Ticket Details:

Ticket Details as Predictors: The analysis of survival patterns within different ticket categories provides insights into the socio-economic dynamics at play during the Titanic disaster.

In conclusion, this comprehensive exploration enhances our understanding of the multifaceted factors influencing survival aboard the Titanic. The findings not only contribute to historical insights but also underscore the need for a personalized approach to disaster preparedness, considering age, gender, family structures, and socio-economic factors. Further research into the underlying mechanisms of these associations can inform targeted interventions, ultimately contributing to a more profound understanding of the human experience during this significant historical event.

5.3. Benefit

Unraveling the dynamics of familial relationships among Titanic passengers brings forth valuable insights. This exploration into family structures and demographic patterns offers practical benefits, informing disaster preparedness, advancing predictive modeling, and bridging historical inquiry with contemporary data science.

5.3.1. Revelation of Family Ties:

Personal Connection: Through this research, I've uncovered a treasure trove of familial connections among Titanic passengers, allowing me to piece together the family units that were part of this historical tragedy.

Discovering Ancestral Links: The study has provided me, and potentially others, with the means to discover and explore the ancestral ties that existed within families aboard the Titanic.

5.3.2. Historical Insights:

Resurrecting Family Stories: Beyond statistical analyses, this research breathes life into historical narratives, shedding light on the lives of individual family members and their shared destinies during the Titanic disaster.

Humanizing the Past: The identification of genuine family structures humanizes the tragedy for me, offering a profound understanding of how families experienced this historic event.

5.3.3. Practical Implications:

Connecting with Family Heritage: Personally, the study becomes a tool for potentially reconnecting with my own family's historical journey on the Titanic.

Contributing to Personal History: Insights into how family structures influenced survival rates contribute to my personal history, enriching my understanding of the familial dynamics during this significant event.

5.3.4. Educational Significance:

Inspiration for Personal Studies: This research serves as an inspiration for my own educational journey, encouraging further studies into family history and the interconnected dynamics of families during historical events.

A Case Study for Reflection: The findings become a tangible case study in my educational endeavors, showcasing the practical application of data science in uncovering familial relationships.

5.3.5. Cultural Heritage and Legacy:

Preservation of Personal Heritage: By identifying and documenting familial relationships, the study contributes to the preservation of my personal cultural heritage and familial ties associated with the Titanic disaster.

Enriching My Legacy: For me, as a researcher connected to Titanic passengers, this study enriches my personal legacy by offering insights into the family structures that played a role in survival outcomes.

5.3.6. Validation of Personal Connections:

Ensuring Accuracy in My Genealogy: The meticulous examination of family structures ensures the accuracy of my own genealogical data, providing a reliable resource for tracing my family connections within the Titanic dataset.

Quality Assurance for My Research: The research becomes a personal commitment to quality genealogical data, fostering a more accurate and detailed understanding of my familial ties during this historic event.

In conclusion, this research paper transcends the boundaries of academic exploration, transforming into a profound journey delving into the intricacies of my own family's history. Unveiling the familial connections within the Titanic dataset holds a deeply personal significance, offering a distinctive perspective through which I forge a meaningful connection with the historical and cultural legacy of this iconic event. This endeavor goes beyond the scholarly realm, becoming a poignant exploration of my roots and a testament to the enduring impact of the Titanic on my familial narrative.

Chapter 6 – SOURCE CODE

```r
library(Hmisc)
library(knitr)
library(ggplot2)
library(dplyr)
library(caret)
library(randomForest)
library(gridExtra)
library(ROCR)
library(corrplot)
library(GGally)
library(gridExtra)
library(cowplot)
library(randomForest)
library(rpart)

train <- read.csv("train.csv", stringsAsFactors = F, na.strings = c("NA",
""))
test <- read.csv("test.csv", stringsAsFactors = F, na.strings = c("NA",
""))

str(train)
test$Survived <- NA
all <- rbind(train, test)

head(all)
summary(all)

#Check Outliers
numeric_columns <- c('Survived', 'Pclass', 'Age', 'SibSp', 'Parch',
'Fare')
par(mfrow=c(3, 2))
for (column in numeric_columns) {
  boxplot(all[, column] ~ all$Survived,
          main = paste('Boxplot of', column, 'by Survived'),
          xlab = 'Survived',
          ylab = column,
          outline=TRUE)
}

boxplot(Fare ~ Survived, data = all,
        main = "Boxplot of Fare by Survived",
        xlab = "Survived",
        ylab = "Fare",
        col = c("lightblue", "lightgreen"),
        notch = TRUE, # Add a notch for confidence intervals
        outline = TRUE)

par(mfrow=c(1, 1))
```

```
#Check NA value
if (any(is.na(all))) {
  cat("There are missing values in the dataset.\n")
} else {
  cat("There are no missing values in the dataset.\n")
}
sapply(all, function(x) {sum(is.na(x))})

#Check duplicated data
if(any(duplicated(all))){
  cat("There is duplicated data")
} else{
  cat("There is no duplicated data")
}

# Correlation
df <- all
non_numeric_columns <- sapply(df, function(x) !is.numeric(x))
non_numeric_column_names <- names(df)[non_numeric_columns]
for (col in non_numeric_column_names) {
  df[[col]] <- as.numeric(factor(df[[col]]))
}
df[is.na(df)] <- 0
zero_sd_columns <- sapply(df, function(x) sd(x, na.rm = TRUE) == 0)
zero_sd_column_names <- names(df)[zero_sd_columns]
df <- df[, !zero_sd_columns]
cor_matrix <- cor(df)
par(mfrow = c(1, 1), mar = c(4, 4, 2, 2))
corrplot(cor_matrix, method = "color", type = 'upper')


#remove outlier
z_scores <- scale(all$Age)
outlier_threshold <- 3
outlier_indices <- which(abs(z_scores) > outlier_threshold)
all_cleaned <- all[-outlier_indices, ]


# Histogram for Age
ggplot(all, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Age",
       x = "Age",
       y = "Frequency")

# Histogram for Fare
filtered_data <- all[all$Fare <= 200,]
ggplot(filtered_data, aes(x = Fare)) +
  geom_histogram(binwidth = 10, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Fare",
       x = "Fare",
       y = "Frequency")

# Bar plot for Pclass
ggplot(all, aes(x = factor(Pclass))) +
  geom_bar(fill = "lightcoral", color = "black") +
  geom_line(stat = "count", aes(y = ..count..), color = "darkred", size =
1.5) +
  labs(title = "Distribution of Pclass",
       x = "Pclass",
       y = "Count")

# Bar plot for Sex
ggplot(all, aes(x = Sex)) +
  geom_bar(fill = "lightpink", color = "black") +
  geom_line(stat = "count", aes(y = ..count..), color = "darkred", size =
1.5) +
  labs(title = "Distribution of Sex",
       x = "Sex",
       y = "Count")

# Bar plot for Embarked
ggplot(all, aes(x = Embarked)) +
  geom_bar(fill = "lightgoldenrod", color = "black") +
  geom_line(stat = "count", aes(y = ..count..), color = "darkred", size =
1.5) +
  labs(title = "Distribution of Embarked",
       x = "Embarked",
       y = "Count")

#Setting up
all$Sex <- as.factor(all$Sex)
all$Survived <- as.factor(all$Survived)
all$Pclass <- as.ordered(all$Pclass) #because Pclass is ordinal

#GLM
glm_model <- glm(Survived ~ Pclass + Age + SibSp + Parch + Fare, data =
all, family = "binomial")
summary(glm_model)

# Barplot of People Survavibility
ggplot(all[!is.na(all$Survived), ], aes(x = factor(Survived), fill =
factor(Survived))) +
  geom_bar(stat='count') +
  labs(x = "Survived") +
  geom_label(stat='count', aes(label=..count..), size=7) +
  theme_grey(base_size = 18)
```

```
# Barplot of Gender based on Survived
p1 <- ggplot(all, aes(x = Sex, fill = Sex)) +
  geom_bar(stat='count', position='dodge') + theme_grey() +
  labs(x = 'All data') +
  geom_label(stat='count', aes(label=..count..))

p2 <- ggplot(all[!is.na(all$Survived),], aes(x = Sex, fill =
factor(Survived))) +
  geom_bar(stat='count', position='dodge') + theme_grey() +
  labs(x = 'Training data only') +
  geom_label(stat='count', aes(label=..count..)) +
  guides(fill = guide_legend(title = "Survived"))

grid.arrange(p1, p2, nrow = 1)

#PClass of Survived
custom_colors <- c("#F8766D", "#00BFC4", "#7CAE00")

p3 <- ggplot(all, aes(x = factor(Pclass), fill = factor(Pclass))) +
  geom_bar(stat='count', position='dodge') +
  labs(x = 'All data') +
  theme(legend.position="none") +
  theme_grey() +
  scale_fill_manual(values = custom_colors)
p4 <- ggplot(all[!is.na(all$Survived),], aes(x = Pclass, fill = Survived))
+
  geom_bar(stat='count', position='dodge') + labs(x = 'Training data
only') +
  theme(legend.position="none") + theme_grey()
p5 <- ggplot(all[!is.na(all$Survived),], aes(x = Pclass, fill =
factor(Survived))) +
  geom_bar(stat='count', position='stack') +
  labs(x = 'Training data only', y= "Count") + facet_grid(.~Sex) +
  theme(legend.position="bottom") + theme_grey()
p6 <- ggplot(all[!is.na(all$Survived),], aes(x = Pclass, fill =
factor(Survived))) +
  geom_bar(stat='count', position='fill') +
  labs(x = 'Training data only', y= "Percent") + facet_grid(.~Sex) +
  theme(legend.position="bottom") + theme_grey()

grid.arrange(p3, p4, p5, p6, ncol=2)

#PClassSex
all$PclassSex[all$Pclass=='1' & all$Sex=='male'] <- 'P1Male'
all$PclassSex[all$Pclass=='2' & all$Sex=='male'] <- 'P2Male'
all$PclassSex[all$Pclass=='3' & all$Sex=='male'] <- 'P3Male'
all$PclassSex[all$Pclass=='1' & all$Sex=='female'] <- 'P1Female'
all$PclassSex[all$Pclass=='2' & all$Sex=='female'] <- 'P2Female'
```

```
all$PclassSex[all$Pclass=='3' & all$Sex=='female'] <- 'P3Female'
all$PclassSex <- as.factor(all$PclassSex)


#Extracting Title and Surname from Name
all$Surname <- sapply(all$Name, function(x) {strsplit(x,
split='[,.]')[[1]][1]})
#correcting some surnames that also include a maiden name
all$Surname <- sapply(all$Surname, function(x) {strsplit(x, split='[-
]')[[1]][1]})
all$Title <- sapply(all$Name, function(x) {strsplit(x,
split='[,.]')[[1]][2]})
all$Title <- sub(' ', '', all$Title) #removing spaces before title
kable(table(all$Sex, all$Title))

#Remove unnecessary Surname
all$Title[all$Title %in% c("Mlle", "Ms")] <- "Miss"
all$Title[all$Title== "Mme"] <- "Mrs"
all$Title[!(all$Title %in% c('Master', 'Miss', 'Mr', 'Mrs'))] <- "Rare
Title"
all$Title <- as.factor(all$Title)
kable(table(all$Sex, all$Title))

#Barplot of Surname
ggplot(all[!is.na(all$Survived), ], aes(x = Title, fill =
factor(Survived), group = factor(Survived))) +
  geom_bar(stat='count', position='stack') +
  labs(x = 'Title') +
  theme_grey()

#creating family size variable
all$Fsize <- all$SibSp+all$Parch +1

ggplot(all[!is.na(all$Survived),], aes(x = factor(Fsize), fill =
factor(Survived), group = factor(Survived))) +
  geom_bar(stat='count', position='dodge') +
  scale_x_discrete(breaks = c(1:11)) +
  labs(x = 'Family Size') +
  theme_grey()

#composing variable that combines total Family size and Surname
all$FsizeName <- paste(as.character(all$Fsize), all$Surname, sep="")

SizeCheck <- all %>%
  group_by(FsizeName, Fsize) %>%
  summarise(NumObs=n())
SizeCheck$NumFam <- SizeCheck$NumObs/SizeCheck$Fsize
SizeCheck$modulo <- SizeCheck$NumObs %% SizeCheck$Fsize
SizeCheck <- SizeCheck[SizeCheck$modulo !=0,]
```

```
sum(SizeCheck$NumObs) #total number of Observations with inconsistencies

#Table of family name
kable(SizeCheck[SizeCheck$FsizeName %in% c('3Davies', '5Hocking',
'6Richards', '2Wilkes', '3Richards', '4Hocking'),])

#Table of Davies
kable(all[all$FsizeName=='3Davies',c(13, 15, 2, 3, 5, 6, 7, 8, 9, 14)])

#Fix the Davies family list
all$FsizeName[c(550, 1222)] <- '2Davies'
all$SibSp[550] <- 0
all$Parch[1222] <- 1
all$Fsize[c(550, 1222)] <- 2
kable(all[all$FsizeName=='2Davies',c(13, 15, 2, 3, 5, 6, 7, 8, 9, 14)])

#Make Another Partition for OverView
train <- read.csv('train.csv',stringsAsFactors=F)
test <- read.csv('test.csv',stringsAsFactors=F)
test$Survived <- NA; data <- rbind(train,test)
data$Surname = substring( data$Name,0,regexpr(',',data$Name)-1)
data$GroupId = paste( data$Surname, data$Pclass,
sub('.$','X',data$Ticket), data$Fare, data$Embarked, sep='-')
data[c(195,1067,59,473,1142),c('Name','GroupId')]

data$Title <- 'man'
data$Title[data$Sex=='female'] <- 'woman'
data$Title[grep('Master',data$Name)] <- 'boy'
data$Color <- data$Survived
data$GroupId[data$Title=='man'] <- 'noGroup'
data$GroupFreq <- ave(1:1309,data$GroupId,FUN=length)
data$GroupId[data$GroupFreq<=1] <- 'noGroup'
data$TicketId = paste(
data$Pclass,sub('.$','X',data$Ticket),data$Fare,data$Embarked,sep='-')
count = 0
for (i in which(data$Title!='man' & data$GroupId=='noGroup')){
  data$GroupId[i] = data$GroupId[data$TicketId==data$TicketId[i]][1]
  if (data$GroupId[i]!='noGroup') {
    # color variable is used in plots below
    if (is.na(data$Survived[i])) data$Color[i] = 5
    else if (data$Survived[i] == 0) data$Color[i] = -1
    else if (data$Survived[i] == 1) data$Color[i] = 2
    count = count + 1
  }
}
cat(sprintf('We found %d nannies/relatives and added them to
groups.\n',count))

data$GroupName = substring( data$GroupId,0,regexpr('-',data$GroupId)-1)
```

```
data$Color[is.na(data$Color) & data$Title=='woman'] <- 3
data$Color[is.na(data$Color) & data$Title=='boy'] <- 4
x = data$GroupId[data$GroupId!='noGroup']; x = unique(x); x=x[order(x)]
plotData <- list(); g <- list()
for (i in 1:3) plotData[[i]] <- data[data$GroupId %in% x[(27*(i-
1))+1:27],]
for (i in 1:3) g[[i]] = ggplot(data=plotData[[i]],
aes(x=0,y=factor(GroupName))) +

geom_dotplot(dotsize=0.9,binwidth=1,binaxis='y',method="histodot",stackgro
ups=T,
               aes(fill=factor(Color),color=Title )) +

scale_color_manual(values=c('gray70','blue','gray70'),limits=c('man','boy'
,'woman')) +

scale_fill_manual(values=c('#BB0000','#FF0000','#009900','#00EE00','gray70
','gray70','white'),
                    limits=c('0','-1','1','2','3','4','5')) +
  scale_y_discrete(limits = rev(levels(factor(plotData[[i]]$GroupName))))
+
  theme(axis.title.x=element_blank(), axis.title.y=element_blank(),
        axis.text.x=element_blank(), axis.ticks.x=element_blank(),
        legend.position='none')
grid.arrange(g[[1]],g[[2]],g[[3]],nrow=1,top='All 80 woman-child-groups in
the test and training datasets combined')

#Table of Similar Ticket Number
kable(all[all$Ticket %in% c('29104', '29105',
'29106'),c(2,3,4,5,6,7,8,9,14)])

#Stick Family by Maiden Name
NC <- all[all$FsizeName %in% SizeCheck$FsizeName,] #create data frame with
only relevant Fsizenames

NC$Name <- sub("\\s$", "", NC$Name) #removing spaces at end Name
NC$Maiden <- sub(".*[^\\)]$", "", NC$Name) #remove when not ending with
')'
NC$Maiden <- sub(".*\\s(.*)\\)$", "\\1", NC$Maiden)
NC$Maiden[NC$Title!='Mrs'] <- "" #cleaning up other stuff between brackets
(including Nickname of a Mr)
NC$Maiden <- sub("^\\(", '', NC$Maiden) #removing opening brackets
(sometimes single name, no spaces between brackets)
#making an exceptions match
NC$Maiden[NC$Name=='Andersen-Jensen, Miss. Carla Christine Nielsine'] <-
'Jensen'

NC$Maiden2[NC$Maiden %in% NC$Surname] <- NC$Maiden[NC$Maiden %in%
NC$Surname]
```

```
NC$Combi[!is.na(NC$Maiden2)] <- paste(NC$Surname[!is.na(NC$Maiden2)],
NC$Maiden[!is.na(NC$Maiden2)])

labels1 <- NC[!is.na(NC$Combi), c('Surname','Combi')]
labels2 <- NC[!is.na(NC$Combi), c('Maiden','Combi')]
colnames(labels2) <- c('Surname', 'Combi')
labels1 <- rbind(labels1, labels2)

NC$Combi <- NULL
NC <- left_join(NC, labels1, by='Surname')

CombiMaxF <- NC[!is.na(NC$Combi),] %>%
  group_by(Combi) %>%
  summarise(MaxF=max(Fsize))
NC <- left_join(NC, CombiMaxF, by = "Combi")

#create family names for those larger families
NC$FsizeCombi[!is.na(NC$Combi)] <-
paste(as.character(NC$Fsize[!is.na(NC$Combi)]),
NC$Combi[!is.na(NC$Combi)], sep="")

#find the ones in which not all Fsizes are the same
FamMaid <- NC[!is.na(NC$FsizeCombi),] %>%
  group_by(FsizeCombi, MaxF, Fsize) %>%
  summarise(NumObs=n())
FamMaidWrong <- FamMaid[FamMaid$MaxF!=FamMaid$NumObs,]

kable(unique(NC[!is.na(NC$Combi) & NC$FsizeCombi %in%
FamMaidWrong$FsizeCombi, c('Combi', 'MaxF')]))

#Find the maximum Fsize within remaining families (no maiden combi's)
NC$MaxF <- NULL
FamMale <- NC[is.na(NC$Combi),] %>%
  group_by(Surname) %>%
  summarise(MaxF=max(Fsize))
NC <- left_join(NC, FamMale, by = "Surname")

NCMale <- NC[is.na(NC$Combi),] %>%
  group_by(Surname, FsizeName, MaxF) %>%
  summarise(count=n()) %>%
  group_by(Surname, MaxF) %>%
  filter(n()>1) %>%
  summarise(NumFsizes=n())

NC$Combi[NC$Surname %in% NCMale$Surname] <- NC$Surname[NC$Surname %in%
NCMale$Surname]

kable(NCMale[, c(1,2)])
kable(all[all$Surname=='Vander Planke', c(2,3,4,5,6,7,8,9,14)])

NC <- NC[(NC$FsizeCombi %in% FamMaidWrong$FsizeCombi)|(NC$Surname %in%
NCMale$Surname),]
NC1 <- NC %>%
  group_by(Combi) %>%
  summarise(Favg=mean(Fsize))
kable(NC1)

#Favg dataframe
NC <- left_join(NC, NC1, by = "Combi")
NC$Favg <- round(NC$Favg)
NC <- NC[, c('PassengerId', 'Favg')]
all <- left_join(all, NC, by='PassengerId')

#replacing Fsize by Favg
all$Fsize[!is.na(all$Favg)] <- all$Favg[!is.na(all$Favg)]

all$Ticket2 <- sub("..$", "xx", all$Ticket)
rest <- all %>%
  select(PassengerId, Title, Age, Ticket, Ticket2, Surname, Fsize) %>%
  filter(Fsize=='1') %>%
  group_by(Ticket2, Surname) %>%
  summarise(count=n())
rest <- rest[rest$count>1,]
rest1 <- all[(all$Ticket2 %in% rest$Ticket2 & all$Surname %in%
rest$Surname & all$Fsize=='1'), c('PassengerId', 'Surname', 'Title',
'Age', 'Ticket', 'Ticket2', 'Fsize', 'SibSp', 'Parch')]
rest1 <- left_join(rest1, rest, by = c("Surname", "Ticket2"))
rest1 <- rest1[!is.na(rest1$count),]
rest1 <- rest1 %>%
  arrange(Surname, Ticket2)
kable(rest1[1:12,])
all <- left_join(all, rest1)

for (i in 1:nrow(all)){
  if (!is.na(all$count[i])){
    all$Fsize[i] <- all$count[i]
  }
}

#Finding Friend Relation
kable(all[all$Ticket=='1601', c('Survived', 'Pclass', 'Title', 'Surname',
'Age', 'Ticket', 'SibSp', 'Parch', 'Fsize')])
TicketGroup <- all %>%
  select(Ticket) %>%
  group_by(Ticket) %>%
  summarise(Tsize=n())
all <- left_join(all, TicketGroup, by = "Ticket")
```

```
library(scales)
#Predicting Missing Age value
ggplot(all[(!is.na(all$Survived) & !is.na(all$Age)), ], aes(x = Age, fill
= factor(Survived))) +
  geom_density(alpha = 0.5) +
  labs(title = "Survival Density and Age") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_minimal() +  # Change theme to minimal
  theme(legend.position = "top") +  # Move legend to the top
  scale_fill_manual(name = "Survived", values = c("0" = "red", "1" =
"blue")) +  # Custom fill colors
  guides(fill = guide_legend(title = "Survived"))

#Relation of age and Pclass
ggplot(all[!is.na(all$Age),], aes(x = Title, y = Age, fill =
factor(Pclass))) +
  geom_boxplot() +
  scale_y_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme_minimal() +  # Change theme to minimal
  labs(title = "Age Distribution by Title and Passenger Class") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +  # Rotate x-
axis labels for better visibility
  scale_fill_manual(name = "Passenger Class", values = c("1" = "#F8766D",
"2" = "#00BFC4", "3" = "#7CAE00")) +  # Custom fill colors
  guides(fill = guide_legend(title = "Passenger Class"))

#creating a variable with almost the same ticket numbers (only last 2
digits varying)
all$Ticket2 <- sub("..$", "xx", all$Ticket)
rest <- all %>%
  select(PassengerId, Title, Age, Ticket, Ticket2, Surname, Fsize) %>%
  filter(Fsize=='1') %>%
  group_by(Ticket2, Surname) %>%
  summarise(count=n())
rest <- rest[rest$count>1,]
rest1 <- all[(all$Ticket2 %in% rest$Ticket2 & all$Surname %in%
rest$Surname & all$Fsize=='1'), c('PassengerId', 'Surname', 'Title',
'Age', 'Ticket', 'Ticket2', 'Fsize', 'SibSp', 'Parch')]
rest1 <- left_join(rest1, rest, by = c("Surname", "Ticket2"))
rest1 <- rest1[!is.na(rest1$count),]
rest1 <- rest1 %>%
  arrange(Surname, Ticket2)
kable(rest1[1:12,])
all <- left_join(all, rest1)
for (i in 1:nrow(all)){
  if (!is.na(all$count[i])){
    all$Fsize[i] <- all$count[i]
  }
}

#People book together
kable(all[all$Ticket=='1601', c('Survived', 'Pclass', 'Title', 'Surname',
'Age', 'Ticket', 'SibSp', 'Parch', 'Fsize')])

#composing data frame with group size for each Ticket
TicketGroup <- all %>%
  select(Ticket) %>%
  group_by(Ticket) %>%
  summarise(Tsize=n())
all <- left_join(all, TicketGroup, by = "Ticket")
ggplot(all[!is.na(all$Survived),], aes(x = Tsize, fill = Survived)) +
  geom_bar(stat='count', position='dodge') +
  scale_x_continuous(breaks=c(1:11)) +
  labs(x = 'Ticket Size') + theme_grey()

#taking the max of family and ticket size as the group size
all$Group <- all$Fsize
for (i in 1:nrow(all)){
  all$Group[i] <- max(all$Group[i], all$Tsize[i])
}

#Creating final group categories
all$GroupSize[all$Group==1] <- 'solo'
all$GroupSize[all$Group==2] <- 'duo'
all$GroupSize[all$Group>=3 & all$Group<=4] <- 'group'
all$GroupSize[all$Group>=5] <- 'large group'
all$GroupSize <- as.factor(all$GroupSize)


#predicting Age with Linear Regression
set.seed(12000)
AgeLM <- lm(Age ~ Pclass + Sex + SibSp + Parch + Embarked + Title +
GroupSize, data=all[!is.na(all$Age),])
summary(AgeLM)
all$AgeLM <- predict(AgeLM, all)

par(mfrow=c(1,2))
hist(all$Age[!is.na(all$Age)], main='Original data', xlab='Age',
col='#F8766D')
box()
hist(all$AgeLM[is.na(all$Age)], main= 'NA predictions', xlab='Age',
col='#00BFC4', xlim=range(0:80))
box()

#display which passengers are predicted to be children (age<18) with
Linear Regression.
```

```
all[(is.na(all$Age) & all$AgeLM <18), c('Sex', 'SibSp', 'Parch', 'Title',
'Pclass', 'Survived', 'AgeLM')]

#Linear Regression predictions for missing Ages
indexMissingAge <- which(is.na(all$Age))
indexAgeSurvivedNotNA<- which(!is.na(all$Age) & (!is.na(all$Survived)))
#needed in sections below
all$Age[indexMissingAge] <- all$AgeLM[indexMissingAge]

#Cabin
all$Cabin[is.na(all$Cabin)] <- "U"
all$Cabin <- substring(all$Cabin, 1, 1)
all$Cabin <- as.factor(all$Cabin)

ggplot(all[(!is.na(all$Survived)& all$Cabin!='U'),], aes(x=Cabin,
fill=Survived)) +
  geom_bar(stat='count') + theme_grey() + facet_grid(.~Pclass) +
labs(title="Survivor split by class and Cabin")

#Children
ggplot(all[all$Age<14.5 & !is.na(all$Survived),], aes(x=Pclass,
fill=Survived))+
  geom_bar(stat='count') + theme_grey(base_size = 18)

all$IsChildP12 <- 'No'
all$IsChildP12[all$Age<=14.5 & all$Pclass %in% c('1', '2')] <- 'Yes'
all$IsChildP12 <- as.factor(all$IsChildP12)

#Embark
d1 <- ggplot(all[!is.na(all$Survived),], aes(x = Embarked, fill =
Survived)) +
  geom_bar(stat='count') + theme_grey() + labs(x = 'Embarked', y= 'Count')

d2 <- ggplot(all[!is.na(all$Survived),], aes(x = Embarked, fill =
Survived)) +
  geom_bar(stat='count', position= 'fill') +
  theme_grey() +
  labs(x = 'Embarked', y = 'Percent', title = 'Survivability Percentage by
Embark')

grid.arrange(d2, nrow=1)

ggplot(all[indexAgeSurvivedNotNA,], aes(x = Age, fill = Survived)) +
  geom_histogram(aes(fill=factor(Survived))) + labs(title="Survival
density, known-ages, and Embarked") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 5)) + theme_grey()
+ facet_grid(.~Embarked)
```

```r
tab1 <-
rbind(table(all$Embarked[!is.na(all$Survived)]),table(all$Embarked[indexAg
eSurvivedNotNA]))
tab1 <- cbind(tab1, (rowSums(tab1)))
tab1 <- rbind(tab1, tab1[1,]-tab1[2,])
tab1 <- rbind(tab1, round((tab1[3,]/tab1[1,])*100))
rownames(tab1) <- c("All", "With Age", "Missing Age", "Percent Missing")
colnames(tab1) <- c("C", "Q", "S", "Total")
kable(tab1)

#Ticket Survivor

TicketSurvivors <- all %>%
  group_by(Ticket) %>%
  summarize(Tsize = n(),
            NumNA = sum(is.na(Survived)),
            SumSurvived = sum(as.numeric(Survived)-1, na.rm = TRUE))

all <- left_join(all, TicketSurvivors, by = "Ticket")

all$AnySurvivors[all$Tsize==1] <- 'other'
all$AnySurvivors[all$Tsize>=2] <- ifelse(all$SumSurvived[all$Tsize>=2]>=1,
'survivors in group', 'other')
all$AnySurvivors <- as.factor(all$AnySurvivors)

kable(x=table(all$AnySurvivors), col.names= c('AnySurvivors',
'Frequency'))

#Fare
all$FarePP <- all$Fare/all$Tsize

tab2 <- all[(!is.na(all$Embarked) & !is.na(all$Fare)),] %>%
  group_by(Embarked, Pclass) %>%
  summarise(FarePP=median(FarePP))
kable(tab2)

# Solo Survivor
all$solo[all$SibSp==0] <- 'Yes'
all$solo[all$SibSp!=0] <- 'No'
all$solo <- as.factor(all$solo)

ggplot(all[!is.na(all$Survived),], aes(x = solo, fill = Survived)) +
  geom_bar(stat='count') + theme_grey(base_size = 18)


library(randomForest)
library(rpart)
library(caret)
```

```r
#Machine Learning
all$FsizeD[all$Fsize == 1] <- 'singleton'
all$FsizeD[all$Fsize < 5 & all$Fsize > 1] <- 'small'
all$FsizeD[all$Fsize > 4] <- 'large'

trainClean <- all[!is.na(all$Survived),]
testClean <- all[is.na(all$Survived),]


#Random Forest
set.seed(2017)
caret_matrix <- train(x=trainClean[,c('PclassSex', 'GroupSize', 'FarePP',
'AnySurvivors', 'IsChildP12')], y=trainClean$Survived, data=trainClean,
method='rf', trControl=trainControl(method="cv", number=5))
caret_matrix

caret_matrix$results

#Decision tree
set.seed(2017)

tree_matrix <- train(
  x = trainClean[, c('PclassSex', 'GroupSize', 'FarePP', 'AnySurvivors',
'IsChildP12')],
  y = trainClean$Survived,
  method = 'rpart',
  trControl = trainControl(method = "cv", number = 5)
)

tree_matrix

tree_matrix$results

#Logistic Regression
logistic_matrix <- train(
  x = trainClean[, c('PclassSex', 'GroupSize', 'FarePP', 'AnySurvivors',
'IsChildP12')],
  y = trainClean$Survived,
  method = 'glm',
  trControl = trainControl(method = "cv", number = 5),
  family = binomial
)

logistic_matrix

logistic_matrix$results
```