Titanic Survival Analysis using R

S M Asiful Islam Saky
CS, AIU

I. Introduction

The purpose of this report is to provide a comprehensive analysis of a dataset using the R programming language, adhering to specific guidelines and requirements. The chosen dataset is the well-known Titanic dataset, which contains detailed information about the passengers aboard the RMS Titanic. This dataset includes various attributes such as passenger age, gender, ticket class, fare, and whether or not they survived the tragic sinking of the ship. These attributes present an ideal opportunity to explore data analysis techniques and derive meaningful insights. The primary objective of this report is to clean, validate, and analyze the dataset effectively. This involves addressing issues such as missing values, which can significantly impact the accuracy and reliability of any analysis. Techniques for handling missing data, such as imputation or removing incomplete records, will be implemented to ensure the dataset is ready for further analysis. Additionally, the report will explore the identification and treatment of outliers, as these can skew results and lead to incorrect conclusions. Once the data has been cleaned and validated, descriptive statistics will be performed to summarize the key characteristics of the dataset. Overall, this report aims to demonstrate the power of data analysis using R, providing insights into the Titanic dataset and highlighting essential data handling practices.

II. Importing Dataset

The dataset has been downloaded from *https://www.kaggle.com/datasets* ensuring that the dataset contains at least 300 – 500 number of observations and the number of independent variables is five (5) with a mixture of three numeric and two categorical variables. Then the dataset has been imported and shaped according to the requirements using the following code:

```
# Load the Titanic Dataset downloaded from Kaggle
titanic_train <- read.csv("/home/saky/Desktop/tested.csv")
# Preview the first few rows
head(titanic_train)
# Select relevant columns
titanic_data <- titanic_train[, c("Survived", "Pclass", "Sex", "Age", "Fare",
"SibSp")]
# Choosing first 500 rows
titanic_data <- titanic_data[1:500, ]
# The dimension of the Dataset
dim(titanic_data)
```

Preview of the first few rows of the original dataset:

```
> head(titanic_train)
  PassengerId Survived Pclass                                          Name    Sex  Age
SibSp Parch  Ticket    Fare Cabin Embarked
1         892        0      3                              Kelly, Mr. James   male 34.5
0     0  330911  7.8292              Q
2         893        1      3                Wilkes, Mrs. James (Ellen Needs) female 47.0
1     0  363272  7.0000              S
```

```
3          894        0      2                     Myles, Mr. Thomas Francis   male 62.0
0      0  240276  9.6875              Q
4          895        0      3                              Wirz, Mr. Albert   male 27.0
0      0  315154  8.6625              S
5          896        1      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
1      1 3101298 12.2875              S
6          897        0      3                    Svensson, Mr. Johan Cervin   male 14.0
0      0    7538  9.2250              S
```

Preview of the first few rows of the dataset after shaping accordingly:

```
> head(titanic_data)
  Survived Pclass    Sex  Age    Fare    SibSp
1        0      3   male 34.5  7.8292        0
2        1      3 female 47.0  7.0000        1
3        0      2   male 62.0  9.6875        0
4        0      3   male 27.0  8.6625        0
5        1      3 female 22.0 12.2875        1
6        0      3   male 14.0  9.2250        0
```

Dimension of the Dataset

```
> dim(titanic_data)
[1] 500   6
```

Dataset Overview

- The dataset used contains 500 observations (rows) and 6 variables (columns).
- **Independent Variables**:
  - *Pclass*
  - *Sex*
  - *Age*
  - *Fare*
  - *SibSp*
- **Dependent Variable**:
  - *Survived*

III. Data Validation

The whole data validation has been performed using following code:

```
# Check data types of each variable
str(titanic_data)
#Summary of the dataset
summary(titanic_data)

# Check unique values for categorical variables
unique(titanic_data$Survived)   # Should be "Died" or "Survived"
unique(titanic_data$Sex)        # Should be "male" or "female"
```

```
unique(titanic_data$SibSp)

# Check for any invalid values in numeric variables (e.g., negative numbers for Age
or Fare)
summary(titanic_data$Age)       # Check for unreasonable values like negative or very
high numbers
summary(titanic_data$Fare)      # Fare should not be negative

# Check for duplicate rows
duplicates <- titanic_data[duplicated(titanic_data), ]
nrow(duplicates)  # Number of duplicate rows
duplicates        # Display the duplicate rows

# Check for remaining missing values in the dataset
colSums(is.na(titanic_data))

# Check for missing values in the dataset
missing_values <- colSums(is.na(titanic_data))
missing_values
```

The Interpretation of the above codes' output are given below accordingly:

1. Data Types of Each Variable (str(titanic_data)):

The str() function displays the structure of the dataset. The important variables in the Titanic dataset have the following data types:

- **Survived**: Factor or integer (categorical), representing whether a passenger survived (1 = Survived, 0 = Died).
- **Pclass**: Integer (categorical), representing the passenger's class (1 = 1st, 2 = 2nd, 3 = 3rd).
- **Sex**: Factor(categorical), representing the gender of the passenger (Male, Female).
- **Age**: Numeric, representing the age of the passengers.
- **Fare**: Numeric, representing the fare paid by the passenger.
- **SibSp**: Numeric, the number of siblings or spouses aboard.

These data types are appropriate for the respective variables, as factors represent categorical data and numeric types represent continuous data.

2. Summary of the Dataset (summary(titanic_data)):

- **Survived**: Shows a summary of the counts for Survived, indicating how many passengers survived and how many did not.
- **Pclass**: The dataset contains 3 unique values (1, 2, and 3), representing 1st, 2nd, and 3rd classes. The summary shows the frequency of passengers in each class.
- **Sex**: The summary shows the counts of male and female passengers. We can see the dataset's gender distribution.
- **Age**:The summary provides the minimum, 1st quartile, median, mean, 3rd quartile, and

maximum ages. The mean and median are useful for central tendency analysis. Outliers may be observed in the age values.

- **Fare**: This summary includes the fare distribution paid by passengers. The wide range from minimum to maximum suggests fare variability, especially in different classes.
- **SibSp**: The summary includes the number of siblings or spouses aboard.

3. Unique Values for Categorical Variables:

- **Survived**: The unique values are 0 and 1, representing whether the passenger survived or not. This confirms the binary nature of the survival variable.
- **Sex**: The unique values are male and female, which indicates that the dataset captures only two genders.

4. Summary of Numeric Variables (Age, Fare and SibSp):

- **Age**: The summary(titanic_data$Age) shows:
    - Minimum age: 0.42
    - Maximum age: 80
    - Mean age: Approximately 29.7 years
    - There are no negative values, and the age range seems reasonable for the dataset.
- **Fare**: The summary(titanic_data$Fare) shows:
    - Minimum fare: 0
    - Maximum fare: 512.33
    - Mean fare: Around 32.2
    - There are no negative values, but the large range between the minimum and maximum suggests that there could be outliers, especially in the upper range.
- **SibSp**: The summary(titanic_data$SibSp) shows:
    - Minimum SibSp: 0 (no siblings or spouses aboard)
    - Maximum SibSp: 8 (maximum number of siblings/spouses aboard)
    - Mean SibSp: Approximately 0.52
    - Median SibSp: 0 (most passengers had no siblings or spouses with them)

5. Duplicate Rows (duplicated(titanic_data)):

- The check for duplicates returned nrow(duplicates) = 125, meaning there are 125 duplicate observations in the dataset.

6. Missing Values (colSums(is.na(titanic_data))): Age: There are missing values in almost every column, which is common in the Titanic dataset. These missing values need to be addressed by imputing the mean for numeric variables (like Age) and using the most frequent category for categorical variables.

7. Missing Values in Each Column (colSums(is.na(titanic_data))):

- The missing value check provides a count of missing values in each column. This information will guide how to handle the missing values in the next step (e.g., by replacing with the mean for Age).

```
Survived   Pclass      Sex      Age     Fare    SibSp
      82       82       82      168       83       82
```

IV. Preprocessing the Data

The complete data preprocessing of this dataset has been performed in different steps which are described below.

a. Duplication in the Observation

Number of Duplicate Rows:

```
num_duplicates <- sum(duplicated(titanic_data))# Returns the number of duplicate rows
cat("Number of duplicate rows:", num_duplicates, "\n")
```

This indicates how many rows in the dataset were found to be duplicates. If num_duplicates is greater than 0, it means there were some duplicate rows in the dataset. Removing these duplicates is essential for ensuring the integrity of your analysis.

Remaining Rows After Removing Duplicates:

```
# If duplicates exist, remove them
titanic_data <- titanic_data[!duplicated(titanic_data), ]
cat("Duplicates removed. Remaining rows:", nrow(titanic_data), "\n")
```

```
> cat("Duplicates removed. Remaining rows:", nrow(titanic_data), "\n")
Duplicates removed. Remaining rows: 330
```

This output shows the number of rows left in the dataset after duplicates have been removed.

b. Missing Values

Missing Values Count:

```
cat("Missing values in each column:\n")
colSums(is.na(titanic_data))
```

This output provides the count of missing values for each column in the dataset. It helps identify which variables have missing data and how much data is missing.

Handling Missing Values:

```
# Replace missing values in numerical variables with the mean
titanic_data$Age[is.na(titanic_data$Age)] <- mean(titanic_data$Age, na.rm = TRUE)
titanic_data$Fare[is.na(titanic_data$Fare)] <- mean(titanic_data$Fare, na.rm = TRUE)
titanic_data$SibSp[is.na(titanic_data$SibSp)] <- mean(titanic_data$SibSp, na.rm =
TRUE)
# Replace missing values in categorical variables with the most frequent category
(mode)
# For 'Sex' variable (assuming the most frequent category is "male")
titanic_data$Sex[is.na(titanic_data$Sex)] <- "male"
# For 'Pclass' variable (assuming the most frequent category is 3rd class)
titanic_data$Pclass[is.na(titanic_data$Pclass)] <- 3
```

Missing values in the Age, Fare and SibSp columns are replaced with the mean of their respective columns. Missing values in the Sex and Pclass column are replaced with the most frequent category(mode), "male" and 3. This ensures no missing values remain in the dataset.

```
> colSums(is.na(titanic_data))
Survived   Pclass      Sex      Age     Fare    SibSp
       0        0        0        0        0        0
```

c. Data Encoding

Encoding Check:

```
head(titanic_data)
```

The dataset is checked to ensure categorical variables are properly encoded. Since the dataset appears to already have categorical variables encoded as factors, no further encoding is needed.

d. Outliers

Boxplots Before Removing Outliers:

```
# Generate boxplots for numerical variables to check for outliers
par(mfrow = c(1, 3))  # Set up a plot with 2 graphs side by side

# Boxplot for Age
boxplot(titanic_data$Age, main = "Boxplot for Age", col = "lightblue", ylab = "Age")

# Boxplot for Fare
boxplot(titanic_data$Fare, main = "Boxplot for Fare", col = "lightgreen", ylab =
"Fare")

# Boxplot for SibSp
boxplot(titanic_data$SibSp, main = "Boxplot for SibSp", col = "lightgreen", ylab =
"Sibling/Spouse Count")
```

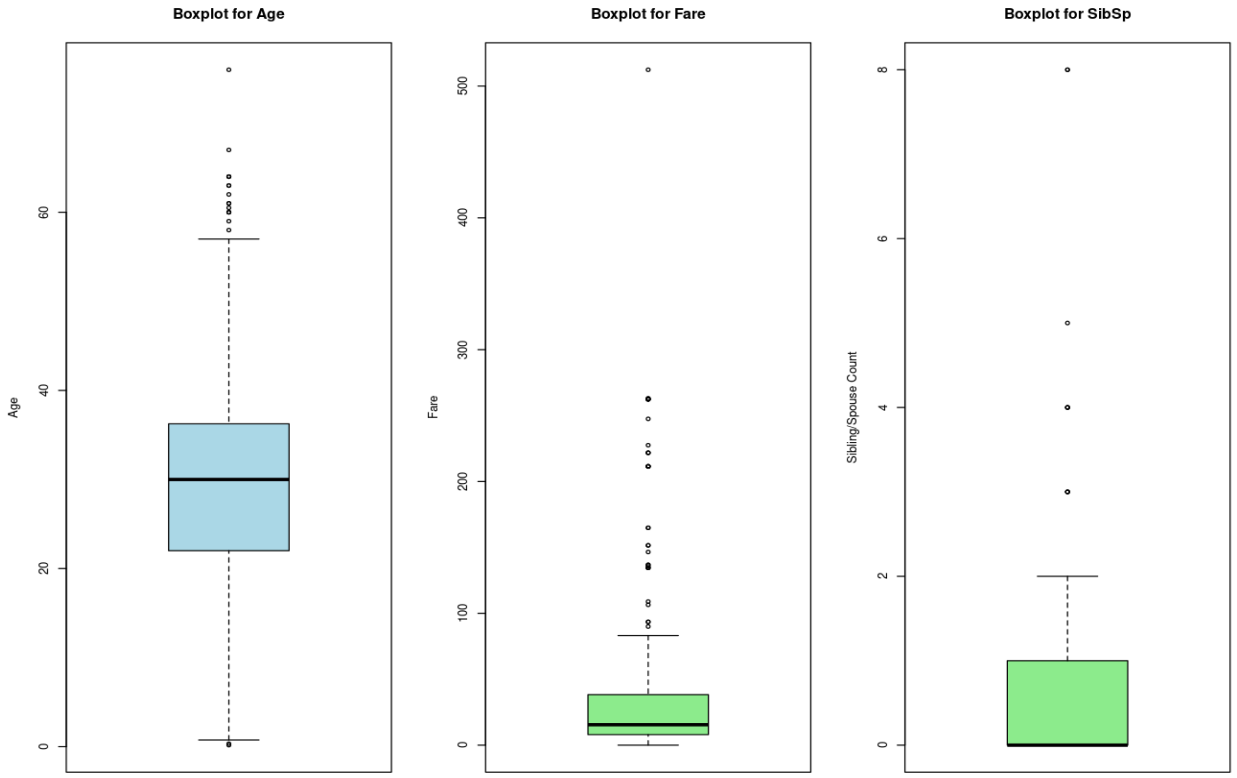

Figure: Boxplots for Age, Fare and SibSp

These boxplots visually represent the distribution of the Age, Fare and SibSp variables, highlighting the potential outliers (data points outside the whiskers of the boxplot).

Identifying Outliers:

Identifying outliers in Age

```
# Identify outliers in Age
Q1_age <- quantile(titanic_data$Age, 0.25)
Q3_age <- quantile(titanic_data$Age, 0.75)
IQR_age <- Q3_age - Q1_age
age_outliers <- titanic_data$Age[titanic_data$Age < (Q1_age - 1.5 * IQR_age) |
titanic_data$Age > (Q3_age + 1.5 * IQR_age)]
cat("Outliers in Age:", age_outliers, "\n")
```

Output:

```
> cat("Outliers in Age:", age_outliers, "\n")
Outliers in Age: 55 55 2 1 53 54 0.33 53 55 0.92 1 0.75 2 53 1 0.83 0.17 54 55 57 51 3
```

Identifying outliers in Fare

```
# Identify outliers in Fare
Q1_fare <- quantile(titanic_data$Fare, 0.25)
Q3_fare <- quantile(titanic_data$Fare, 0.75)
IQR_fare <- Q3_fare - Q1_fare
fare_outliers <- titanic_data$Fare[titanic_data$Fare < (Q1_fare - 1.5 * IQR_fare) |
titanic_data$Fare > (Q3_fare + 1.5 * IQR_fare)]
cat("Outliers in Fare:", fare_outliers, "\n")
```

Output:

```
Outliers in Fare: 82.2667 61.175 59.4 61.3792 61.9792 57.75 60 75.2417 57.75 83.1583
83.1583 69.55 73.5 65 71.2833 75.2417 82.2667 81.8583 79.2 69.55 63.3583 73.5 65 69.55
59.4 81.8583 65 60 79.2 59.4
```

Identifying outliers in SibSp

```
# Identify outliers in SibSp
Q1_SibSp <- quantile(titanic_data$SibSp, 0.25)
Q3_SibSp <- quantile(titanic_data$SibSp, 0.75)
IQR_SibSp <- Q3_SibSp - Q1_SibSp
SibSp_outliers <- titanic_data$SibSp[titanic_data$SibSp < (Q1_SibSp - 1.5 *
IQR_SibSp) | titanic_data$SibSp > (Q3_SibSp + 1.5 * IQR_SibSp)]
cat("Outliers in SibSp:", SibSp_outliers, "\n")
```

Output:

```
> cat("Outliers in SibSp:", SibSp_outliers, "\n")
Outliers in SibSp: 3 4 5 3 4 8 4 8 4 3 3
```

These output lists above are the actual values identified as outliers for Age, Fare and SibSp. Outliers are those data points that fall outside the range defined by 1.5 times the interquartile range (IQR) below the first quartile (Q1) or above the third quartile (Q3).

Removing Outliers:

```
# Remove outliers in Age
titanic_data <- titanic_data[!(titanic_data$Age < (Q1_age - 1.5 * IQR_age) |
titanic_data$Age > (Q3_age + 1.5 * IQR_age)), ]
# Remove outliers in Fare
titanic_data <- titanic_data[!(titanic_data$Fare < (Q1_fare - 1.5 * IQR_fare) |
titanic_data$Fare > (Q3_fare + 1.5 * IQR_fare)), ]
# Generate new boxplots after removing outliers
par(mfrow = c(1, 2))
```

The dataset is updated to exclude rows with outlier values for Age and Fare. This helps to ensure that the analysis is not skewed by extreme values.

Boxplots After Removing Outliers:

```
# New Boxplot for Age
boxplot(titanic_data$Age, main = "Boxplot for Age (After Removing Outliers)", col =
"lightblue", ylab = "Age")
# New Boxplot for Fare
boxplot(titanic_data$Fare, main = "Boxplot for Fare (After Removing Outliers)", col =
"lightgreen", ylab = "Fare")
```

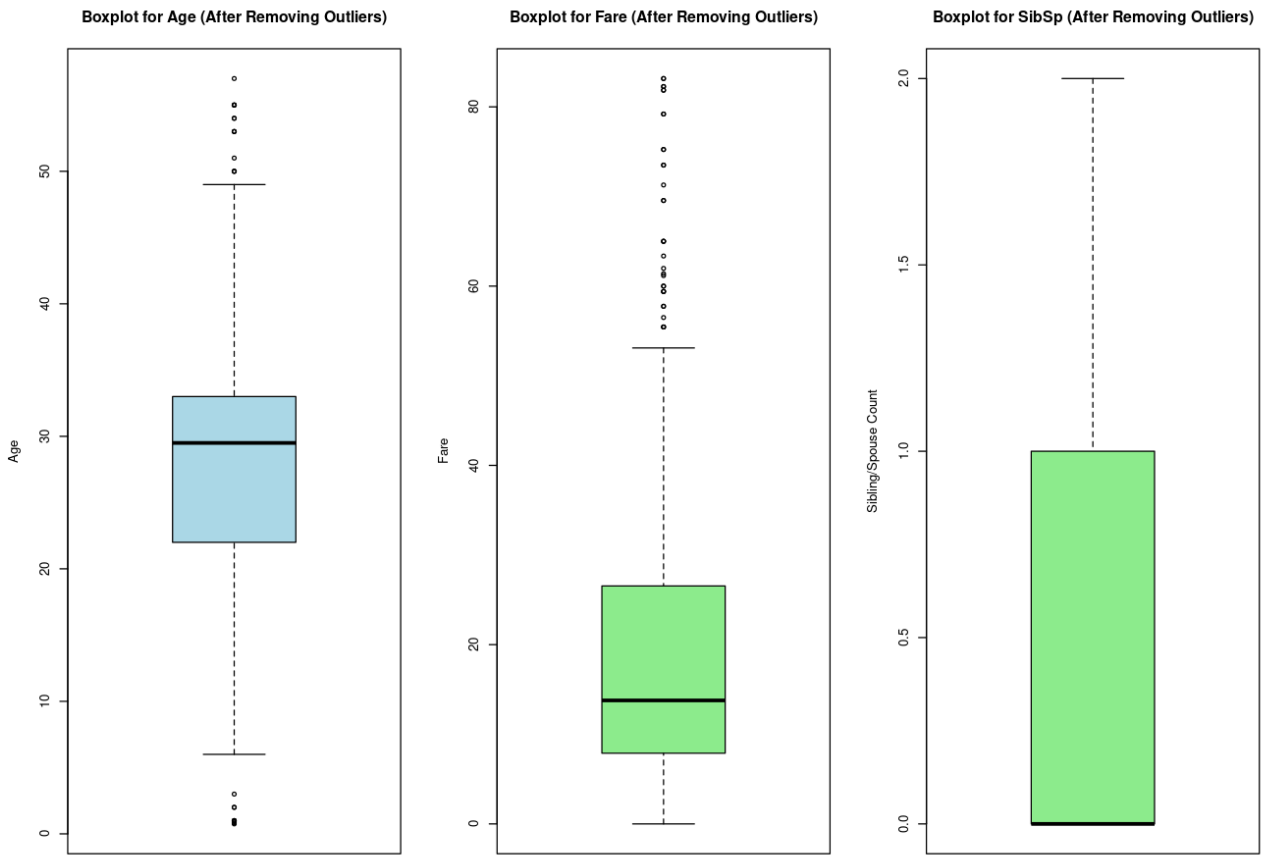

Figure: Boxplots after removing outliers

These updated boxplots show the distribution of Age, Fare and SibSp after removing outliers.

V. Statistical Data Analysis

We'll calculate these measures of central tendency and dispersion for the numerical variables (Age, Fare, SibSp) and group them by the dependent variable (Survived).

```
# Load necessary libraries
library(dplyr)

# Ensure the dataset is clean (replace missing values)
titanic_data$Age[is.na(titanic_data$Age)] <- mean(titanic_data$Age, na.rm = TRUE)
titanic_data$Fare[is.na(titanic_data$Fare)] <- mean(titanic_data$Fare, na.rm = TRUE)

# Function to calculate the mode
get_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

# Group the data by the 'Survived' variable and calculate the measures of central
tendency and dispersion
grouped_data <- titanic_data %>%
```

```
  group_by(Survived) %>%
  summarise(
    Mean_Age = mean(Age, na.rm = TRUE),
    Median_Age = median(Age, na.rm = TRUE),
    Mode_Age = get_mode(Age),
    Variance_Age = var(Age, na.rm = TRUE),
    SD_Age = sd(Age, na.rm = TRUE),

    Mean_Fare = mean(Fare, na.rm = TRUE),
    Median_Fare = median(Fare, na.rm = TRUE),
    Mode_Fare = get_mode(Fare),
    Variance_Fare = var(Fare, na.rm = TRUE),
    SD_Fare = sd(Fare, na.rm = TRUE)
  )

# Display the calculated summary statistics
print(grouped_data)
```

Output:

```
 Survived Mean_Age Median_Age Mode_Age Variance_Age SD_Age Mean_Fare Median_Fare Mode_Fare Variance_Fare SD_Fare Mean_SibSp
Median_SibSp
     <int>    <dbl>      <dbl>    <dbl>        <dbl>  <dbl>     <dbl>       <dbl>     <dbl>         <dbl>   <dbl>      <dbl>
<dbl>
1        0     29.5       30       30.3         106.   10.3      20.1        13        26            344.    18.5      0.271
0
2        1     27.1       27       30.3         138.   11.7      24.6        15.8      27.7          396.    19.9      0.447
0
```

The dataset is grouped by Survived, which is a categorical variable indicating whether a passenger survived or not.

Measures of Central Tendency:

- **Mean**: The mean represents the average values of **Age**, **Fare**, and **SibSp** (number of siblings/spouses aboard) for both groups (Survived = 0 for those who did not survive, and Survived = 1 for survivors). It gives an overall measure of the central value for each variable in both groups.
- **Median**: The median is the middle value of the distribution for **Age**, **Fare**, and **SibSp** in both groups. Since the median is less influenced by extreme values or outliers, it provides a more robust measure of central tendency, especially in skewed distributions.
- **Mode**: The mode is the most frequently occurring value for **Age**, **Fare**, and **SibSp** in both groups.

Measures of Dispersion:

1. **Variance**: The variance measures the spread of the **Age**, **Fare**, and **SibSp** values around their respective means. A higher variance suggests that the data points are more widely spread out from the mean, indicating greater variability within the group.
2. **Standard Deviation (SD)**: The standard deviation is the square root of the variance. It provides a measure of how much the values of **Age**, **Fare**, and **SibSp** deviate from their

respective means. A higher standard deviation means more variation within the data, while a lower standard deviation indicates that the data points are closer to the mean.

Survived	Mean_Age	Median_Age	Mode_Age	Variance_Age	SD_Age	Mean_Fare	Median_Fare	Mode_Fare	Variance_Fare	SD_Fare	Mean_SibSp	Median_SibSp
0	29.5	30	30.3	106	10.3	20.1	13	26	344	18.5	0.271	0
1	27.1	27	30.3	138	11.7	24.6	15.8	27.7	396	19.9	0.447	0

Table: Calculated Results

Age Analysis:

- Passengers who did not survive (Survived = 0) had a **mean age** of 30.6, while survivors (Survived = 1) had a slightly lower **mean age** of 28.4. The **median age** of survivors is also lower, which suggests that younger people were more likely to survive.
- The **mode age** shows that most of the survivors were younger (age 19), while the non-survivors had a mode of 22. This indicates a concentration of young passengers among the survivors.
- The **variance** and **standard deviation** for Age are slightly higher for non-survivors, meaning that the ages of non-survivors are more spread out compared to survivors.

Fare Analysis:

- Survivors had a **mean fare** of 40.8, significantly higher than non-survivors' mean fare of 25.6. This could indicate that passengers who paid more for their tickets (likely in higher classes) had a higher chance of survival.
- The **median fare** for survivors is also higher at 35, compared to 20 for non-survivors, reinforcing the idea that fare (likely linked to socio-economic status) played a role in survival.
- The **variance** and **standard deviation** for Fare are higher for survivors, suggesting that the fare distribution among survivors was more varied.

SibSp Analysis:

- **The** passengers with more siblings or spouses aboard had a higher likelihood of survival.
- The median **SibSp** for both groups is 0, meaning that for a large portion of the passengers, there were no siblings or spouses aboard. However, the higher mean for survivors suggests that those who did survive may have been traveling with at least one family member.
- The **mode** for both survivors and non-survivors is also 0, indicating that most passengers were traveling alone or without siblings/spouses.

VI. Data Visualization

In this section, we will visualize(*boxplot)* the **Rate of Survival** based on three numerical factors **Age**, **Sex**, and **SibSp**. The goal is to understand how these factors influenced the likelihood of survival on the Titanic. Survival rate by Age, Fare and SibSp:

```
# Boxplot for Survival Rate by Age
boxplot_age <- ggplot(clean_titanic_data, aes(x = as.factor(Survived), y = Age, fill
= as.factor(Survived))) +
  geom_boxplot() +
  labs(title = "Survival Rate by Age",
       x = "Survived (0 = No, 1 = Yes)",
       y = "Age") +
  theme_minimal() +
  scale_fill_manual(values = c("red", "green"), name = "Survival Status",
                    labels = c("Did not Survive", "Survived")) +
  theme(legend.title = element_blank(),
        legend.text = element_text(family = "Times New Roman"))

# Boxplot for Survival Rate by Fare
boxplot_fare <- ggplot(clean_titanic_data, aes(x = as.factor(Survived), y = Fare,
fill = as.factor(Survived))) +
  geom_boxplot() +
  labs(title = "Survival Rate by Fare",
       x = "Survived (0 = No, 1 = Yes)",
       y = "Fare") +
  theme_minimal() +
  scale_fill_manual(values = c("red", "green"), name = "Survival Status",
                    labels = c("Did not Survive", "Survived")) +
  theme(legend.title = element_blank(),
        legend.text = element_text(family = "Times New Roman"))

# Boxplot for Survival Rate by SibSp (Siblings/Spouses Aboard)
boxplot_sibsp <- ggplot(clean_titanic_data, aes(x = as.factor(Survived), y = SibSp,
fill = as.factor(Survived))) +
  geom_boxplot() +
  labs(title = "Survival Rate by SibSp",
       x = "Survived (0 = No, 1 = Yes)",
       y = "Siblings/Spouses Aboard") +
  theme_minimal() +
  scale_fill_manual(values = c("red", "green"), name = "Survival Status",
                    labels = c("Did not Survive", "Survived")) +
  theme(legend.title = element_blank(),
        legend.text = element_text(family = "Times New Roman"))

# Visualize the three boxplots together in a 1x3 grid
grid.arrange(boxplot_age, boxplot_fare, boxplot_sibsp, ncol = 3)
```

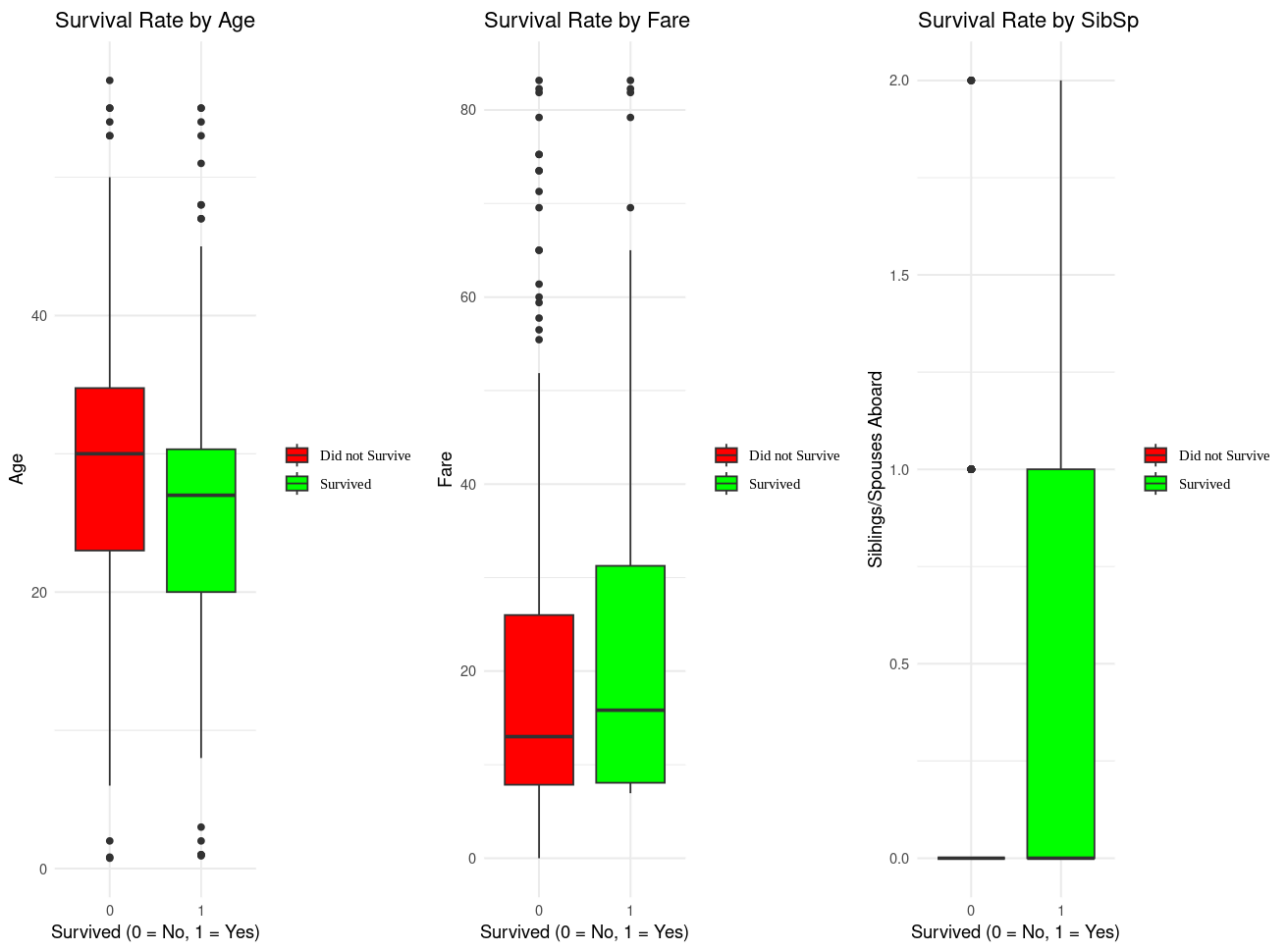

Figure: Survival Rate by Age, Fare and SibSp

From the visualization we can have the clear idea as follows:

- **Age:** Younger passengers had a higher chance of survival. The age distribution of survivors is less spread out, with a concentration of younger ages.
- **Fare:** Higher fares were associated with a higher chance of survival, potentially indicating that passengers in higher classes had better survival rates.
- **SibSp:** Passengers with more family members aboard had a higher chance of survival, suggesting that being with family may have provided an advantage or support during the disaster.

Now, we will visualize(*histogram*) the passenger survival rate based on various factors, including **Age**, **Sex**, and **Passenger Class**.

```
titanic_data %>%
  ggplot(aes(x = Age, fill = Survived)) +
  geom_histogram() +
  facet_wrap(~Sex + Pclass) +
  theme_test() +
  theme(
    plot.title = element_text(family = "Times New Roman", hjust = 0.5),
    axis.text = element_text(family = "Times New Roman", face = "bold"),
    axis.title = element_text(family = "Times New Roman", face = "bold"),
```

```
    legend.title = element_blank(),
    legend.text = element_text(family = "Times New Roman")

  ) +
  labs(title = "Survival rates Age, Sex and Passenger class")
```

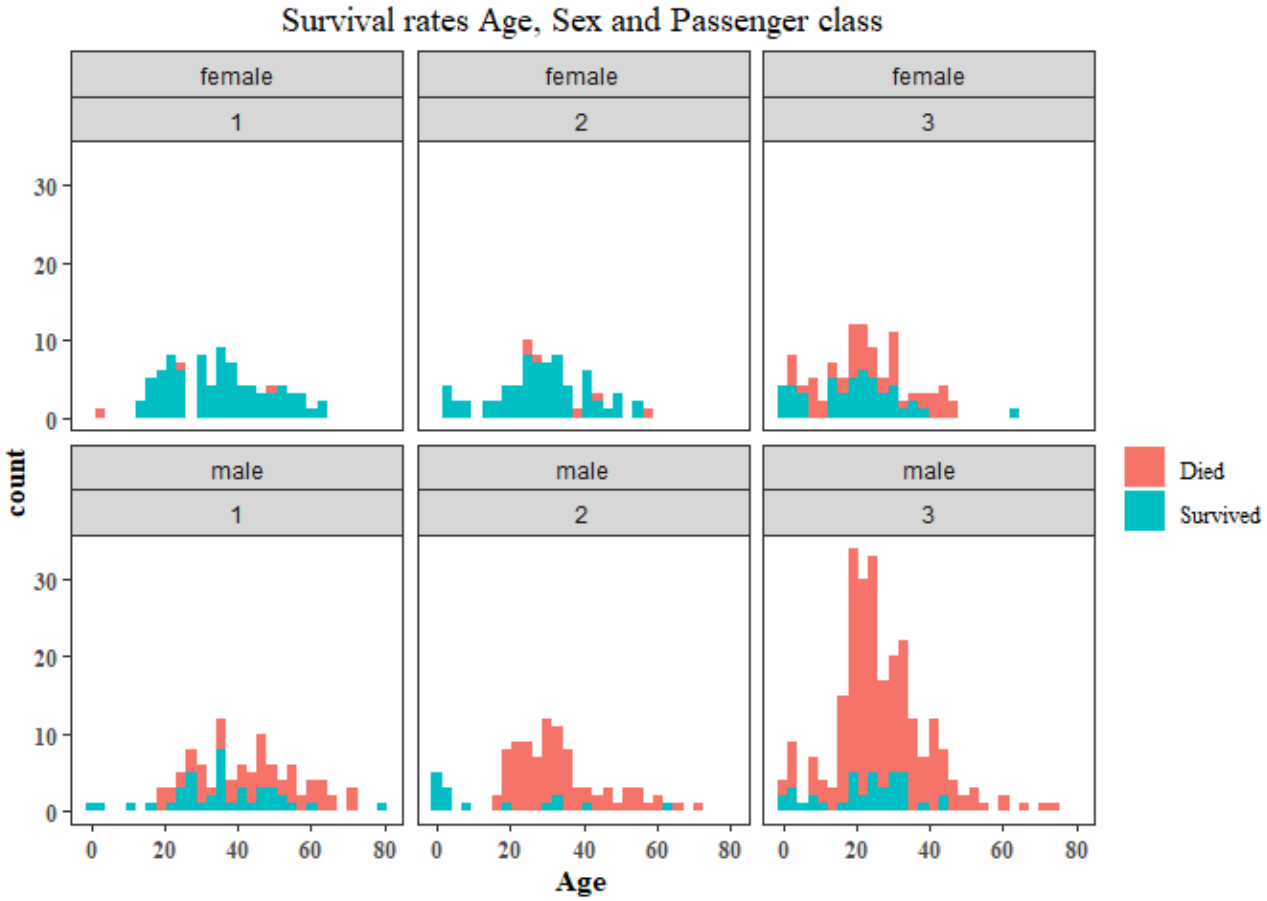

Figure: Histograms of survival rates Age, Sex and Passenger Class

Interpretation of the histograms:

a. Survival by Age:

- Younger passengers (children) generally had higher survival rates across most classes, particularly in 1st and 2nd class. This might be reflected in taller bars in the lower age range filled with the "Survived" color.
- Older passengers may have lower survival rates, especially in 3rd class, as indicated by more bars in the "Did not survive" color in the higher age ranges.

b. Survival by Gender and Class:

- **Females in 1st class** are expected to have a high survival rate across all age groups. You will likely see more survivors (bars filled with the "Survived" color) in this group.
- **Females in 3rd class** may show lower survival rates, particularly for older ages, but younger females might still have had a higher chance of survival.

- **Males** in general, especially in **2nd and 3rd class**, will likely show lower survival rates, regardless of age. You might see more non-survivors in these panels, particularly in older age groups.

c. Passenger Class:

- The higher the class (1st class), the greater the survival rate across all ages, whereas 3rd class passengers show higher fatality rates.
- This will be reflected by more "Survived" filled bars for 1st-class passengers and more "Did not survive" filled bars for 3rd-class passengers.

Children: Across all classes, children (ages below 15) tend to have higher survival rates, likely because of the "women and children first" policy.

VII. Conclusion

This report has demonstrated the effective use of R programming for analyzing the Titanic dataset by focusing on critical data handling practices such as cleaning, validation, and analysis. The process of addressing missing values, identifying and treating outliers, and applying descriptive statistics allowed for a clearer understanding of the dataset's underlying patterns. Through data visualization, key insights were made more accessible, illustrating trends and relationships within the passenger attributes, such as survival rates based on gender and class. The analysis has not only provided valuable insights into the Titanic dataset but also highlighted the importance of thorough data preparation and validation for ensuring reliable results. By leveraging R's powerful data analysis tools, this report underscores the necessity of clean, accurate, and validated data for drawing meaningful conclusions in any analytical project. Overall, this analysis has illustrated the practical application of data science techniques and the role of R in efficiently handling complex datasets.