Survival Analysis of Titanic Passengers

Table of Contents

Introduction

For this project, we chose the Titanic passengers dataset to analyze patterns related to passenger, travel classes, and survival rates. The dataset provides information about passengers, including their age, gender, ticket class, fare, and whether they survived the tragedy. We use basic statistics like averages, ranges, and graphs like boxplots to better understand the data. The project also includes simple and multiple regression analysis to see how some factors, like age and fare, are related. Each step is explained along with interpretation

Dataset Variables

Link to dataset: https://github.com/datasciencedojo/datasets/blob/master/titanic.csv

Variable	Type	Description
Survived	Qualitative	Survival status (0 = No, 1 = Yes).
Pclass	Qualitative	Passenger class (1 = First, 2 = Second, 3 = Third).
Sex	Qualitative	Gender of the passenger (male, female).
Age	Quantitative	Passenger's age (some missing values).
SibSp	Quantitative	Number of siblings/spouses on board.
Parch	Quantitative	Number of parents/children on board.
Embarked	Categorical	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

Variables Analysis and Exploration

Data Preparation

First, we read the Titanic dataset and prepare it for analysis by creating age bins and calculating the number of family members for each passenger.

```r
df <- read.csv("D:\\R-statistics\\titanic.csv")

ages <- df[['Age']]
age_breaks <- c(0, 13, 19, 25, 36, 51, 66, 100)
```

```r
df$age_bin <- cut(df$Age, breaks = age_breaks, right=FALSE)
df$family_members <- as.numeric(df$Parch + df$SibSp)
```

Function to Create Pie Charts

```r
create_pie_chart <- function (freq_table, title){
  labels <- rownames(freq_table)
  pie(freq_table, labels = labels, main = title,
    col = c("lightblue", "lightgreen", "lightcoral", "lightpink"),
    radius = 0.8, border = "white", init.angle = 90, cex = 0.8,
    clockwise = TRUE)
}
```

Passenger Gender Distribution

```r
print(table(df$Sex))
```

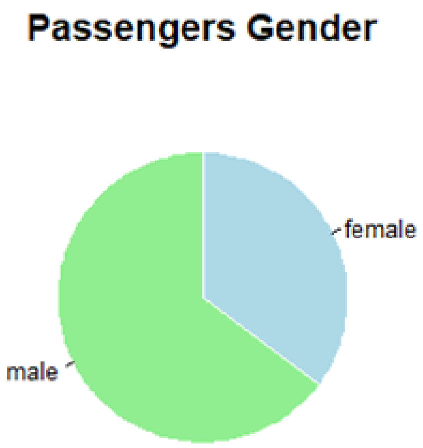
```
female   male 
   314    577 
```

```r
print(prop.table(table(df$Sex)))
```

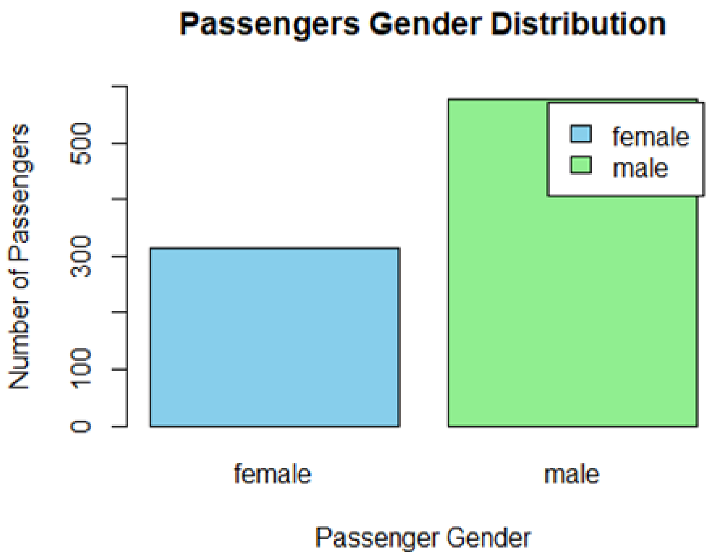
```
  female     male 
0.352413 0.647587 
```

```r
create_pie_chart(table(df$Sex), "Passengers Gender")
```

```r
barplot(
  table(df$Sex),
  main = "Passengers Gender Distribution",
  xlab = "Passenger Gender",
  ylab = "Number of Passengers",
  col = c("skyblue", "lightgreen", "salmon"),
  legend = rownames(table(df$Sex)),
  ylim = c(0, 600)
)
```



Passengers Gender



Passengers Gender Distribution

Interpretation

Frequency Distribution (Table):

- **Female**: 314 passengers.
- **Male**: 577 passengers.
- **Interpretation**: The dataset contains more **male passengers** (577) than **female passengers** (314). There are almost double the number of **male passengers** compared to **female passengers**.

Relative Frequency:

- **Female**: 35.24% of the total passengers.
- **Male**: 64.76% of the total passengers.
- **Interpretation**: **64.76%** of the passengers were **male**, while **35.24%** were **female**, indicating a greater proportion of male passengers in the Titanic dataset.

Pie Chart:

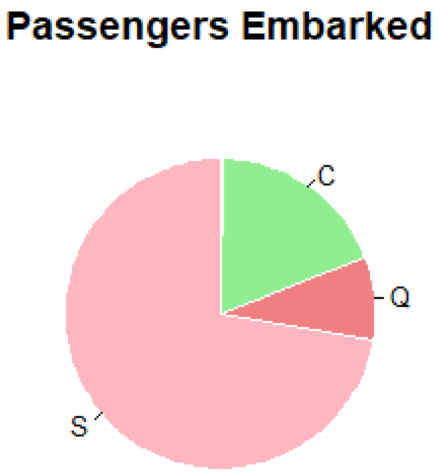
- The pie chart visually confirms that **male passengers** make up a larger share (around 64.76%) of the total passengers, while **female passengers** constitute a smaller proportion (around 35.24%).

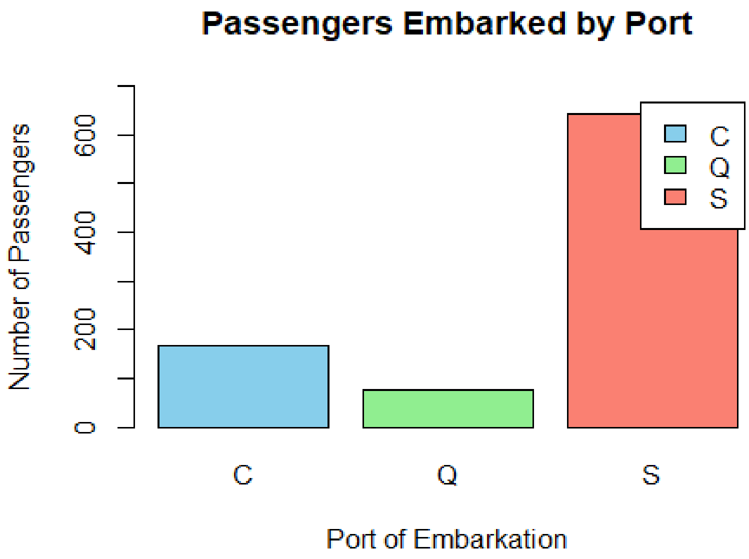
Bar Plot:

- The bar plot shows that there are **more male passengers** than female passengers. The number of **female passengers** is noticeably smaller than the number of **male passengers**, with a greater height for the male bar compared to the female bar.

Passenger Embarked Distribution

```
# Passengers embarking on specific port
print(table(df$Embarked))

      C   Q   S 
  2 168  77 644 

print(prop.table(table(df$Embarked)))
                      C           Q           S 
0.002244669 0.188552189 0.086419753 0.722783389 

create_pie_chart(table(df$Embarked), "Passengers Embarked")
barplot(tbl, main = "Passengers Embarked by Port",
  xlab = "Port of Embarkation", ylab = "Number of Passengers",
  col = c("skyblue", "lightgreen", "salmon"),
  legend = rownames(tbl), ylim = c(0, 700))
```





Interpretation

Frequency Distribution (Table):

- **Southampton (S)**: 644 passengers (largest group).
- **Cherbourg (C)**: 168 passengers.
- **Queenstown (Q)**: 77 passengers (smallest group).
- **Unknown**: 2 passengers.
- **Interpretation**: Most passengers embarked from **Southampton**, followed by **Cherbourg**, with very few passengers boarding from **Queenstown**.

Relative Frequency:

- **Southampton (S)**: 72.28% of passengers.
- **Cherbourg (C)**: 18.86% of passengers.
- **Queenstown (Q):** 8.64% of passengers.
- **Unknown**: 0.22% of passengers.
- **Interpretation**: Most passengers (72.28%) embarked from **Southampton**, a smaller proportion from **Cherbourg (18.86%)**, and very few from **Queenstown (0.22%)**.

Pie Chart:

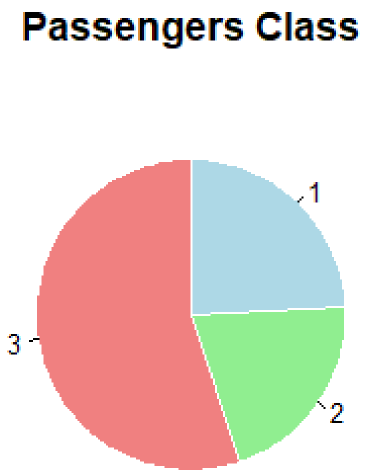
- The **pie chart** visually confirms that the largest share of passengers boarded at **Southampton**, with **Cherbourg** and **Queenstown** having a much smaller share.

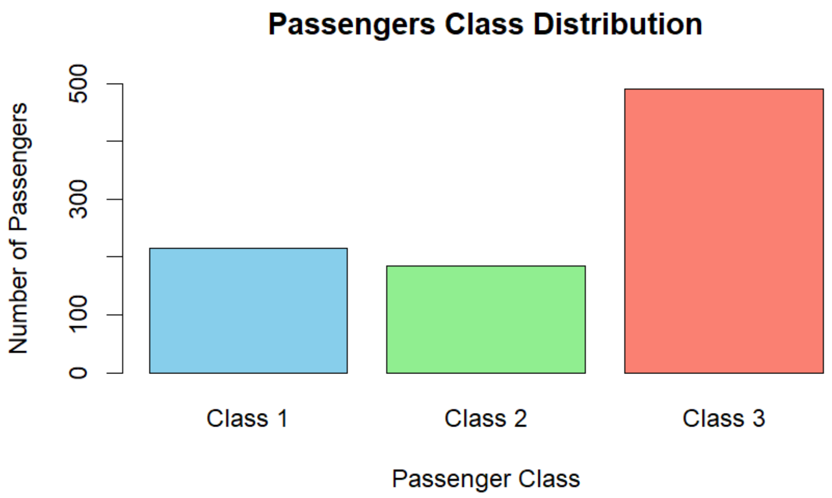
Bar Plot:

- The **bar plot** shows the most passengers from **Southampton (S)**, followed by **Cherbourg (C)**, **Queenstown (Q)** and with a very small number.

Passenger Class Distribution

```
print(table(df$Pclass))

  1   2   3 
216 184 491 

print(prop.table(table(df$Pclass)))

        1         2         3 
0.2424242 0.2065095 0.5510662 

create_pie_chart(table(df$Pclass), "Passengers Class")

barplot(
  table(df$Pclass),
  main = "Passengers Class Distribution",
  xlab = "Passenger Class",
  ylab = "Number of Passengers",
  col = c("skyblue", "lightgreen", "salmon"),
  names.arg = c("Class 1", "Class 2", "Class 3"),
  ylim = c(0, 500)
)
```





Interpretation

Frequency Distribution (Table):

- **Class 1**: 216 passengers.
- **Class 2**: 184 passengers.
- **Class 3**: 491 passengers (largest group).
- **Interpretation**: Most passengers were in **Third Class (Pclass 3)**, with a smaller number in **First Class (Pclass 1)** and **Second Class (Pclass 2)**.

Relative Frequency:

- **Class 1**: 24.24% of passengers.
- **Class 2**: 20.65% of passengers.
- **Class 3**: 55.11% of passengers.
- **Interpretation**: Over half of the passengers (55.11%) were in **Third Class**, while **First Class** accounted for 24.24% and **Second Class** 20.65%.

Pie Chart:

- The **pie chart** visually illustrates that the majority of passengers were in **Third Class (Pclass 3)**, followed by **First Class (Pclass 1)** and **Second Class (Pclass 2)**.

Bar Plot:

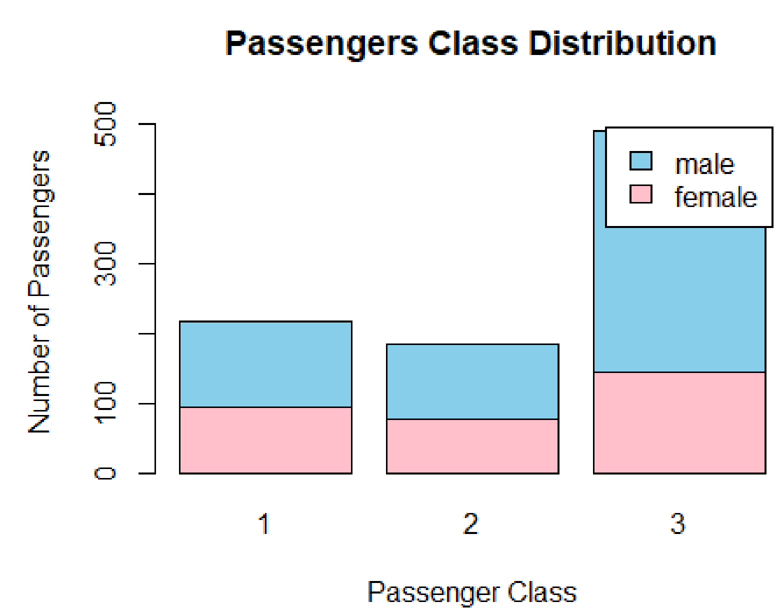
- The **bar plot** will show the largest number of passengers in **Third Class**, with **First Class** and **Second Class** having smaller bars.

Passenger Class Distribution By Gender

```
gender_class_table <- table(df$Sex, df$Pclass)
print(gender_class_table)
```

```
##
##             1   2   3
##   female   94  76 144
##   male    122 108 347
```

```
barplot(
  gender_class_table,
  main = "Passengers Class Distribution
",
  xlab = "Passenger Class",
  ylab = "Number of Passengers",
  col = c("pink", "skyblue"),
  legend = rownames(gender_class_table)
,
  ylim = c(0, 520),
)
```



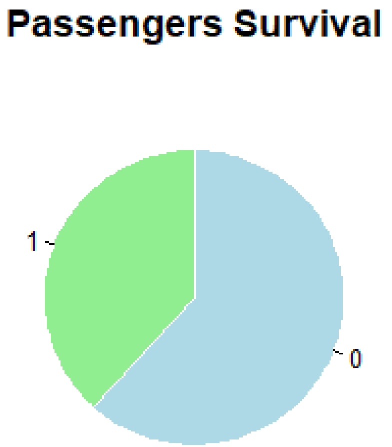
Interpretation

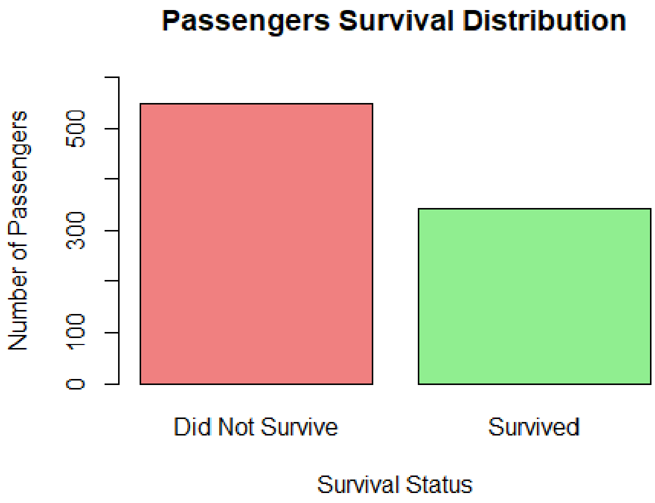
- **Class 1**: 94 females and 122 males.
- **Class 2**: 76 females and 108 males.
- **Class 3**: 144 females and 347 males.
- Males dominate across all classes, especially in Class 3, where the count of males (347) is significantly higher than females (144).
- The gender distribution in Class 1 and Class 2 is more balanced, but males still outnumber females in both classes.

Survival Distribution

```
print(table(df$Survived))

  0   1 
549 342 

print(prop.table(table(df$Survived)))

        0         1 
0.6161616 0.3838384 

create_pie_chart(table(df$Survived), "Passengers Survival")

barplot(table(df$Survived), main = "Passengers Survival Distribution",
  xlab = "Survival Status", ylab = "Number of Passengers",
  col = c("lightcoral", "lightgreen"),
  names.arg = c("Did Not Survive", "Survived"), ylim = c(0, 600))

# Create a table for survival counts based on gender
survival_gender_table <- table(df$Sex, df$Survived)
```

```r
# Create a stacked bar plot
barplot(survival_gender_table, main = "Survival Count by Gender",
    xlab = "Survival Status", ylab = "Number of Passengers",
    col = c("lightcoral", "lightgreen"),
    legend = rownames(survival_gender_table), beside = FALSE,
    ylim = c(0, max(survival_gender_table) + 200),
    names.arg = c("Did Not Survive", "Survived"))
```





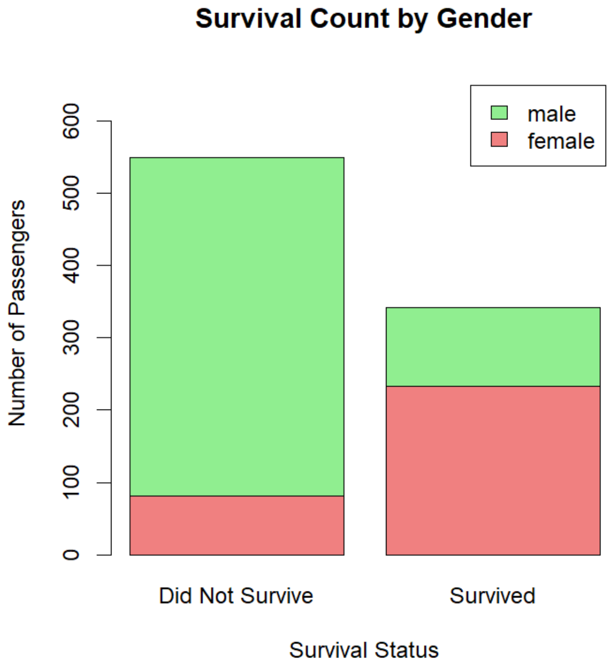
Interpretation

Frequency Distribution (Table):

- **Survived (1)**: 342 passengers.
- **Did Not Survive (0)**: 549 passengers (majority).
- **Interpretation**: More passengers **did not survive** (549) than survived (342), reflecting the tragic nature of the Titanic disaster.

Relative Frequency:

- **Survived (1)**: 38.38% of passengers.
- **Did Not Survive (0)**: 61.62% of passengers.



- **Interpretation**: A higher proportion of passengers did not survive (61.62%), while 38.38% survived, which indicates that survival was less likely.

Pie Chart:

- The **pie chart** visually shows that most passengers **did not survive**, with a smaller portion having survived.

Bar Plot:

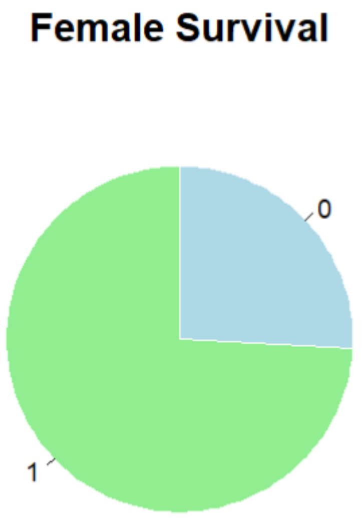
- The **bar plot** shows a higher bar for **Did Not Survive** compared to **Survived**, emphasizing the higher number of non-survivors.

Stacked Bar Chart

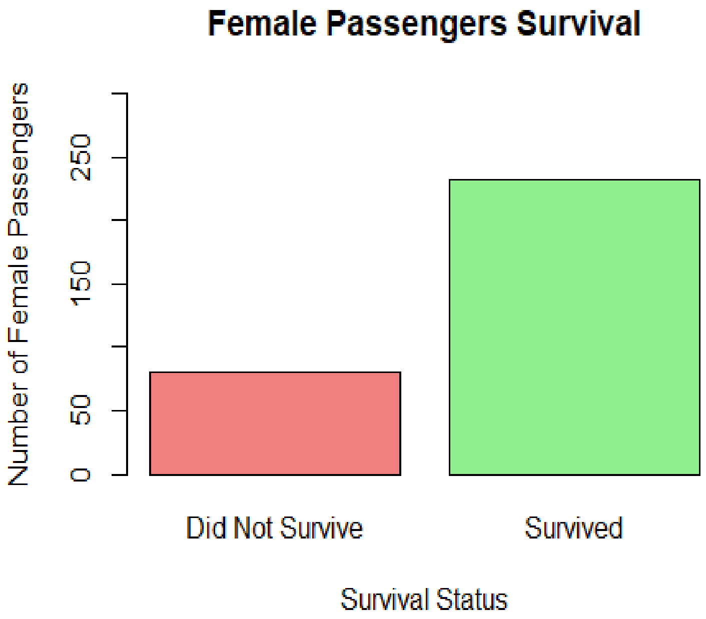
- The bar plot shows that there is greater number of men who did not survived as compared to women who comparatively had better survival chance.

Survival Distribution By Gender (Female)

```
female_survival_table = table(df[df$Sex == "female",]['Survived'])
print(female_survival_table)

  0   1 
 81 233 

# relative freq
print(prop.table(female_survival_table))

        0         1 
0.2579618 0.7420382 

create_pie_chart(female_survival_table, "Female Survival")

barplot(female_survival_table, main = "Female Passengers Survival",
  xlab = "Survival Status", ylab = "Number of Female Passengers",
  col = c("lightcoral", "lightgreen"),
  names.arg = c("Did Not Survive", "Survived"), ylim = c(0, 300))
```



Female Survival



Female Passengers Survival

Interpretation

Frequency Distribution (Table):

- **Survived (1)**: 233 females.
- **Did Not Survive (0)**: 81 females.
- **Interpretation**: A higher number of **female passengers survived** compared to those who did not, suggesting that women had a better chance of survival.

Relative Frequency:

- **Survived (1)**: 74.20% of females.
- **Did Not Survive (0)**: 25.79% of females.
- **Interpretation**: The survival rate for **female passengers** was **74.20%**, much higher than the survival rate for the general population.

Pie Chart:

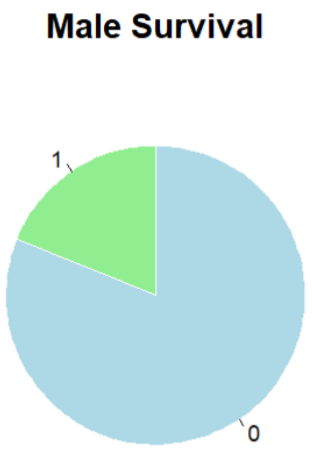
- The **pie chart** shows that the majority of females **survived**, with a small portion not surviving.

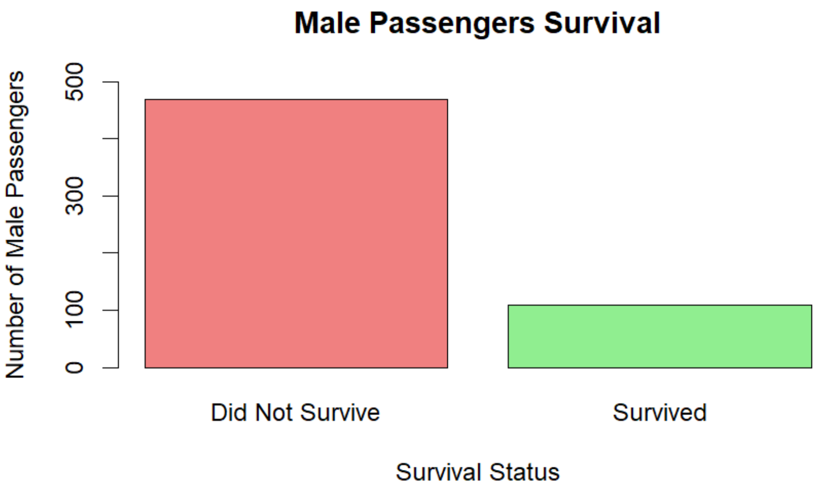
Bar Plot:

- The **bar plot** will show a much higher bar for **Survived (1)** compared to **Did Not Survive (0)** for females, with survival being the dominant outcome.

Survival Distribution By Gender (Male)

```
male_survival_table = table(df[df$Sex == "male",]['Survived'])
print(male_survival_table)

  0   1 
468 109 

# relative freq
print(prop.table(male_survival_table))

        0         1 
0.8110919 0.1889081 

create_pie_chart(male_survival_table, "Male Survival")

barplot(male_survival_table, main = "Male Passengers Survival", xlab = "Survival Status", ylab = "Number of Male Passengers", col = c("lightcoral", "lightgreen"), names.arg = c("Did Not Survive", "Survived"), ylim = c(0, 400))
```





Interpretation

Frequency Distribution (Table):

- **Survived (1)**: 109 males.
- **Did Not Survive (0)**: 468 males (majority).
- **Interpretation**: A much larger number of **male passengers did not survive** compared to those who did, indicating a lower survival rate for males.

Relative Frequency:

- **Survived (1)**: 18.89% of males.
- **Did Not Survive (0)**: 81.11% of males.

- **Interpretation**: The survival rate for **male passengers** was very low (**18.89%**), emphasizing the disparity in survival between males and females.

Pie Chart:

- The **pie chart** shows a significant portion of **males did not survive**, with a much smaller proportion of males surviving.

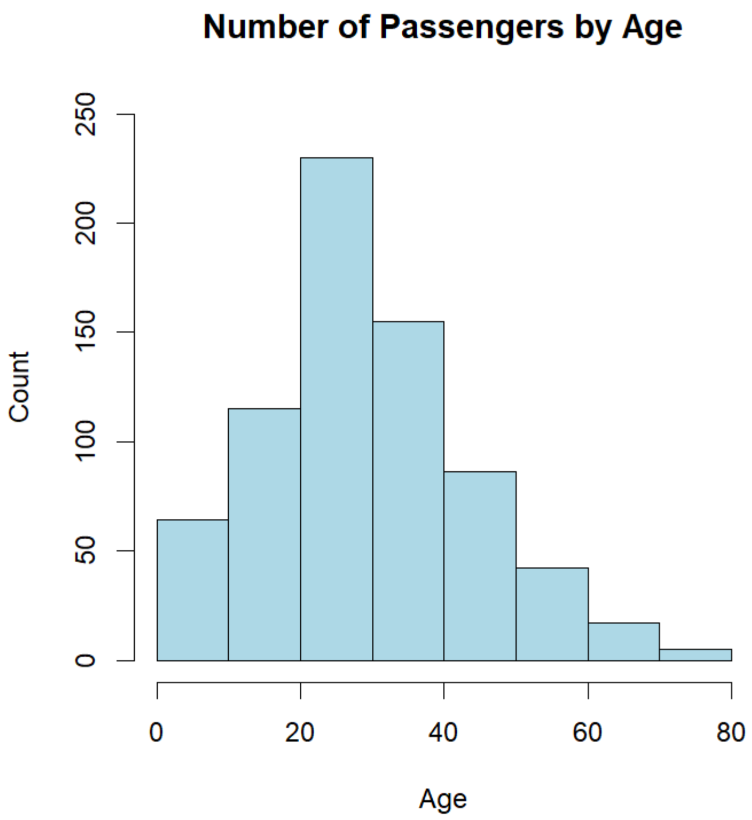
Bar Plot:

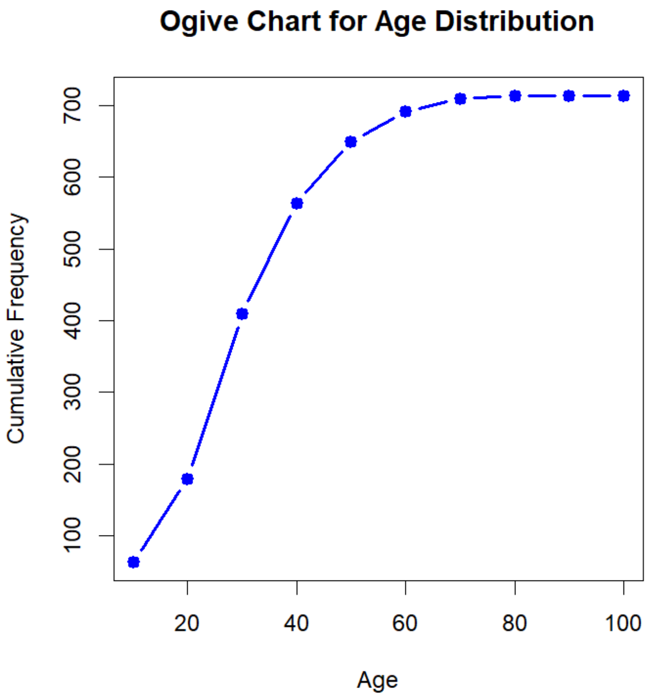
- The **bar plot** for **male passengers** will show a higher bar for **Did Not Survive** compared to **Survived**, reflecting the low survival rate for men.

Age Distribution

We create a histogram to understand the age distribution of passengers.

```r
hist(df$Age, breaks=10, main = "Number of Passengers by Age", xlab = "Age", y
lab = "Count", border='black', col = 'lightblue')

age_bins <- seq(0, 100, by = 10)  # Define age bins
age_hist <- hist(df$Age, breaks = age_bins, plot = FALSE)

# Calculate cumulative frequency
cumulative_freq <- cumsum(age_hist$counts)

# Create an ogive plot with straight lines connecting the points
plot(age_bins[-1], cumulative_freq,
    type = "b", col = "blue", lwd = 2, xlab = "Age",
    ylab = "Cumulative Frequency", main = "Ogive Chart for Age Distribution")

# Add points for better visualization
points(age_bins[-1], cumulative_freq, col = "blue", pch = 16)
```





Interpretation

- The **age distribution** of Titanic passengers, as shown in the histogram, reveals how passengers are spread across different age ranges. The histogram shows that most passengers were in their **20s and 30s**, with a sharp decline in passengers as the age increases.
- The **young adults** (20-30 years) are the most heavily represented, while the number of passengers drops steadily for those over 50.
- The **elderly** passengers (over 60) were the least represented, reflecting a small number of older individuals aboard the Titanic.

Ogive Chart

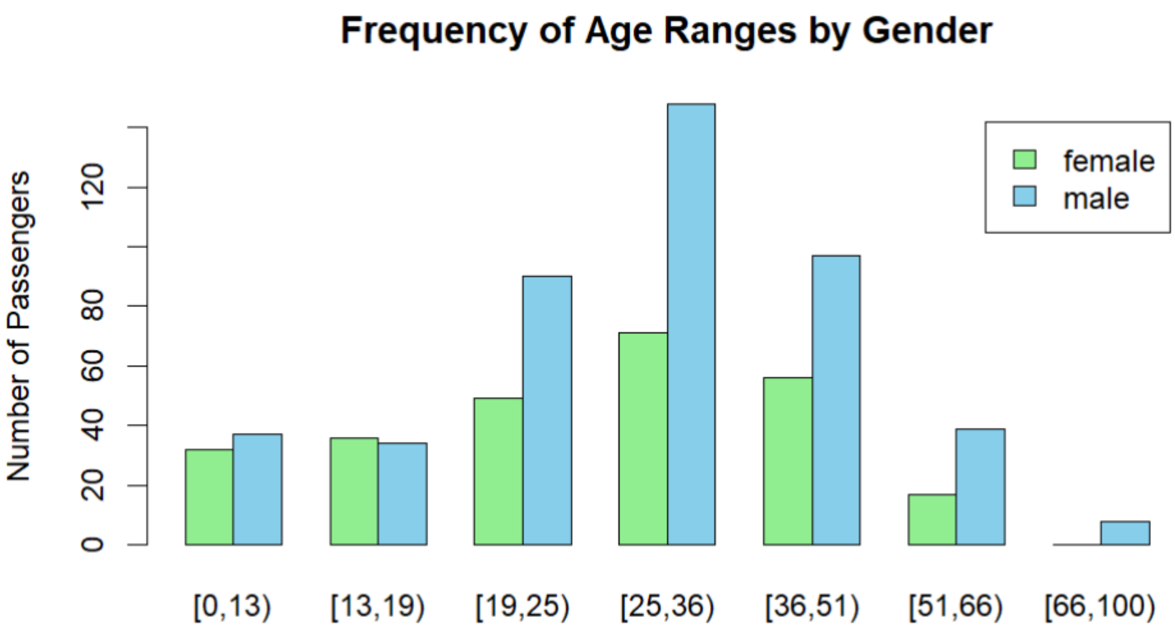
- The steep slope between 20-40 years indicates that there were a significant number of passengers in this age range.
- A flattening of the curve after 60 years suggests that the number of passengers in the higher age ranges (60+) was relatively low.

Age and Gender Distribution

```
age_gender_table <- table(df$Sex, df$age_bin)
print(age_gender_table)
```

```
       [0,13) [13,19) [19,25) [25,36) [36,51) [51,66) [66,100)
female     32      36      49      71      56      17        0
male       37      34      90     148      97      39        8
```

```
barplot(age_gender_table, main = "Frequency of Age Ranges by Gender",
  xlab = "Age Range", ylab = "Number of Passengers",
  col = c("lightgreen", "skyblue"),
  legend = rownames(age_gender_table), beside = TRUE)
```



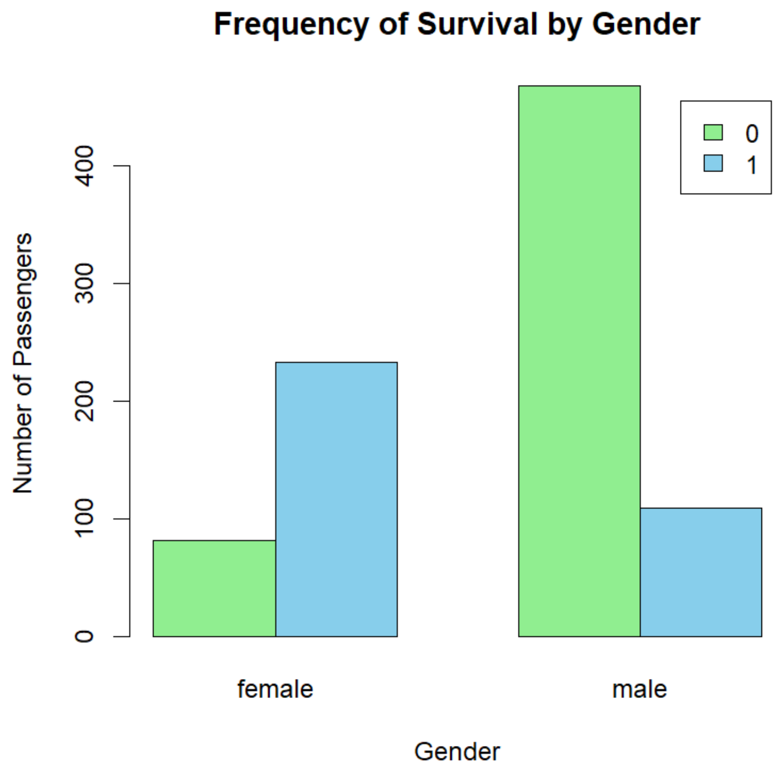
Interpretation

- The majority of passengers, both **male and female**, fall within the 19 – 36 years of age, with the largest group being **male passengers in the [25,36)** age range.
- Both **female and male passengers** show a decline in numbers in the older age groups, particularly in the **[66,100)** group.

Survival Frequency w.r.t Genders

```
survival_by_gender_table <- table(df$Survived, df$Sex)
print(survival_by_gender_table)

  female male
0     81  468
1    233  109
```

```
barplot(survival_by_gender_table, main = "Frequency of Survival by Gender",
xlab = "Gender", ylab = "Number of Passengers", col = c("lightgreen", "skyblu
e"), legend = rownames(survival_by_gender_table), beside = TRUE)
```



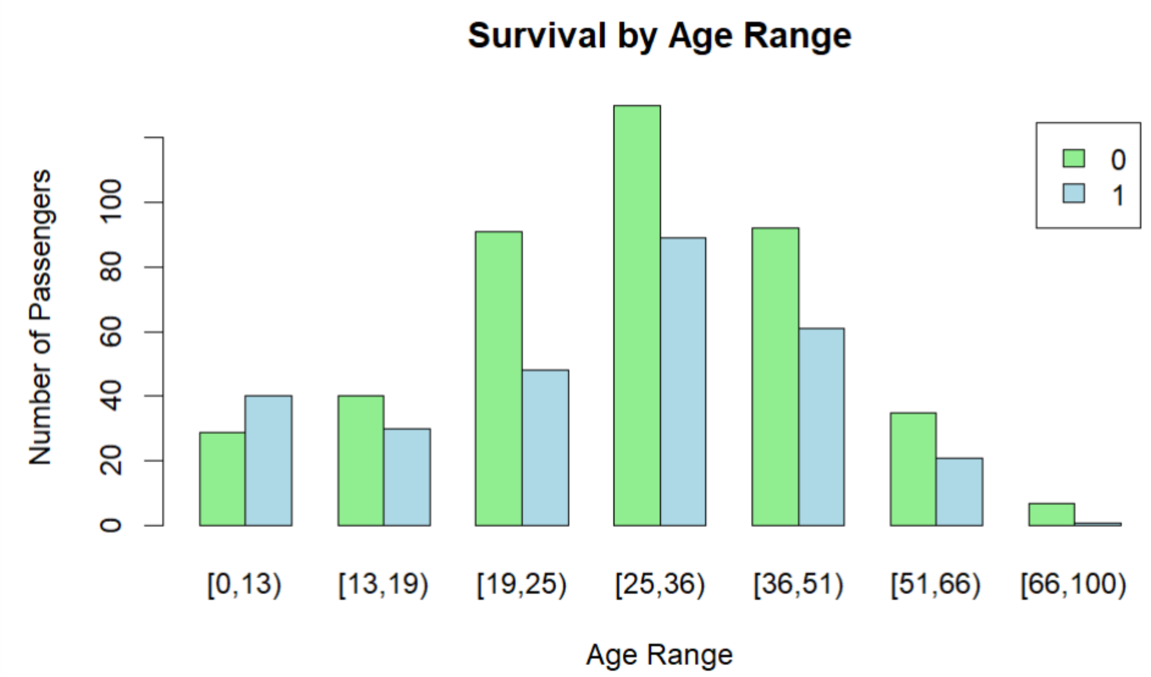
Interpretation

- **Females**: 81 did not survive & 233 survived.
- **Males**: 468 did not survive & 109 survived.
- A significantly higher proportion of **females survived** compared to **males**. In total, **233 females** survived, which is almost three times the number of **109 males** who survived.
- The **number of males who did not survive** is much higher (468) than the number of females who did not survive (81).

Survival rates by age

```
survival_by_age_freq = table(df$Survived, df$age_bin)
print(survival_by_age_freq)

    [0,13) [13,19) [19,25) [25,36) [36,51) [51,66) [66,100)
  0     29      40      91     130      92      35        7
  1     40      30      48      89      61      21        1
```

```r
barplot(survival_by_age_freq, main = "Survival by Age Range",
  xlab = "Age Range", ylab = "Number of Passengers",
  col = c("lightgreen", "lightblue"),
  legend = rownames(survival_by_age_freq), beside = TRUE)
```



Interpretation

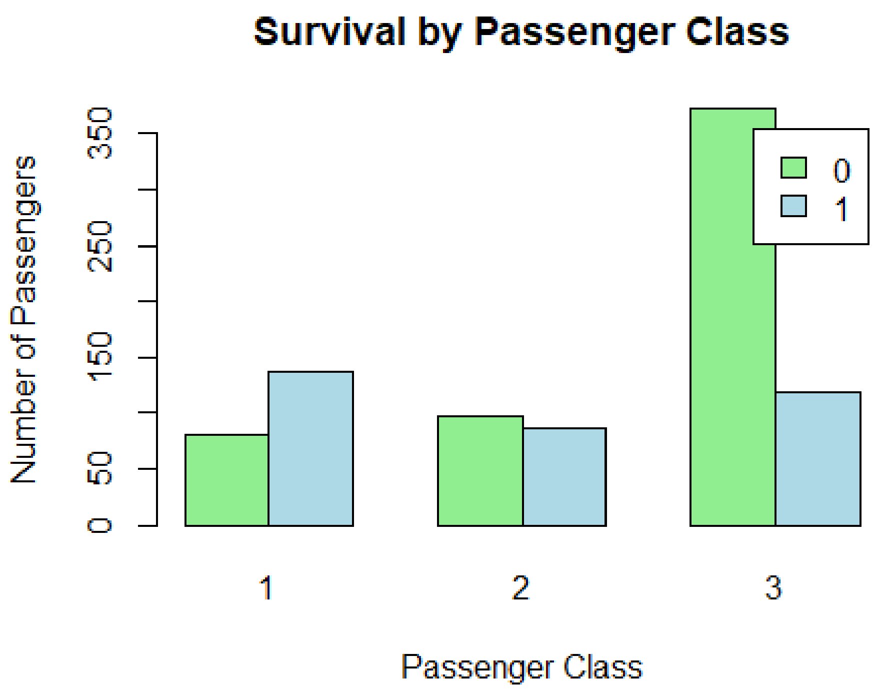
- The **multiple bar chart** clearly shows that survival rates were higher among children (0-13) compared to other age groups.
- **Children (0-13)** had the highest survival rate relative to their non-survival count, with **40 survived** and **29 not survived**. This suggests priority given to children during evacuation.
- Young adults and middle-aged passengers (19-51) had the highest non-survival counts.
- Elderly passengers (66+) faced the poorest survival outcomes, reflecting potential physical and logistical challenges during evacuation.

Survival rates by passenger class

```r
survival_by_pc_freq = table(df$Survived, df$Pclass)
print(survival_by_pc_freq)
```

```
     1   2   3
  0  80  97 372
  1 136  87 119
```

```
barplot(survival_by_pc_freq, main = "Survival by Passenger Class",
  xlab = "Passenger Class", ylab = "Number of Passengers",
  col = c("lightgreen", "lightblue"), legend = rownames(survival_by_pc_freq),
  beside = TRUE)
```

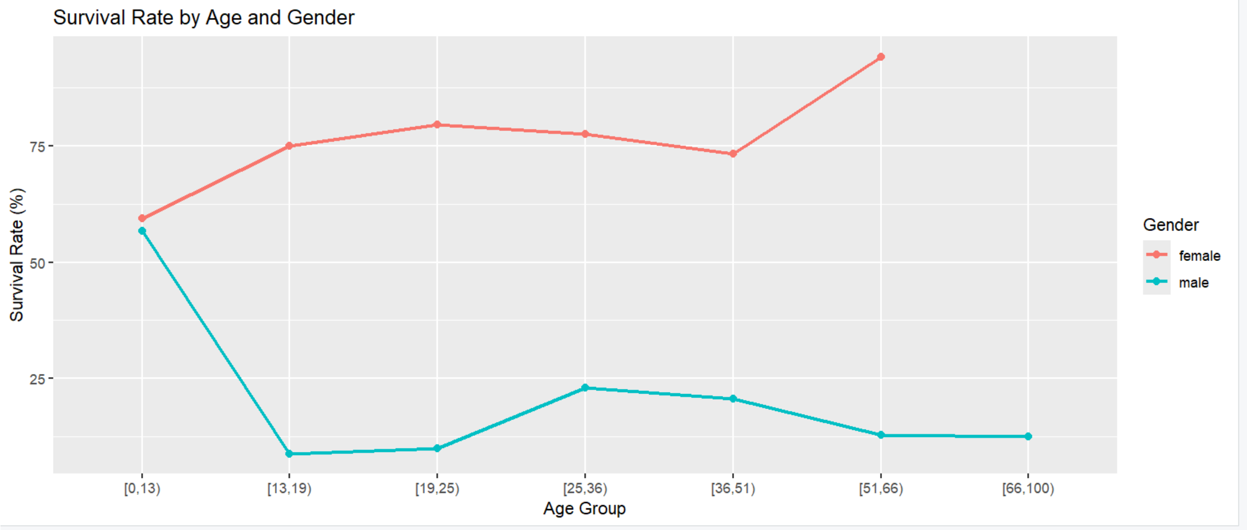


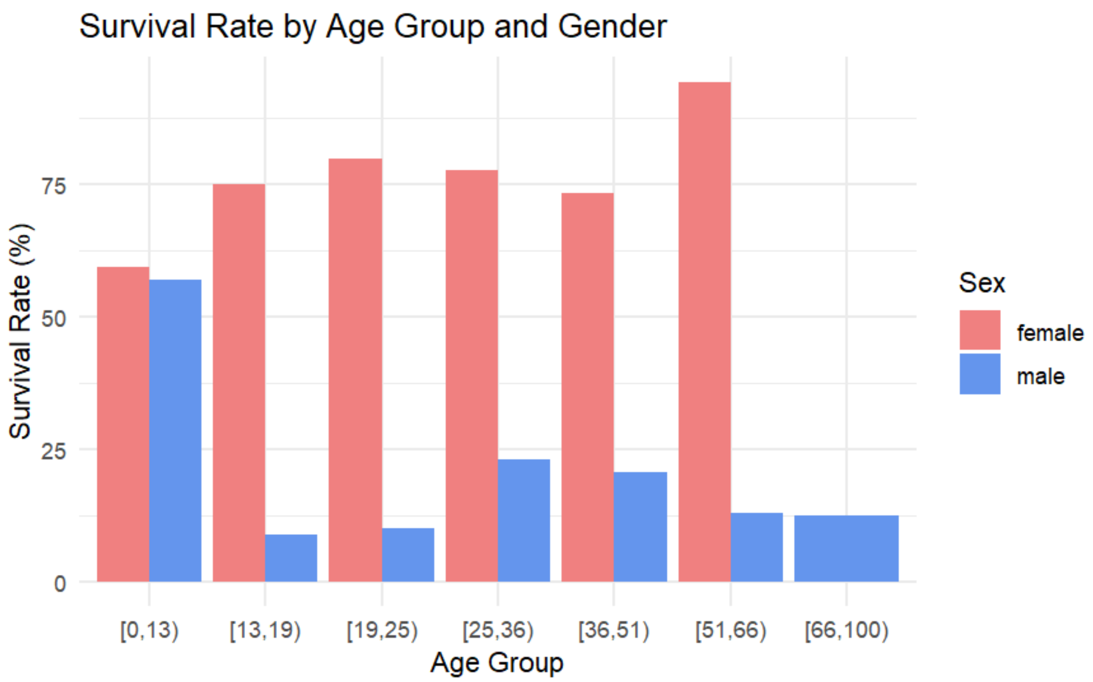
Interpretation

- **1st Class passengers** had the best chances of survival (**136 survived**) compared to non-survivors (**80 not survived**), likely due to their proximity to lifeboats and prioritization during evacuation.
- **2nd Class passengers** faced mixed outcomes (**87 survived** and **97 not survived)**, with survival and non-survival counts relatively close.
- **3rd Class passengers** had the poorest survival outcomes (**372 not survived** and only **119 survived)**, reflecting social and logistical disadvantages such as lower deck placement and delayed access to lifeboats.

Survival Rate by Age and Gender

```
age_gender_survival_rate <- df %>%
  group_by(age_bin, Sex) %>%
  summarise(
    Total_Passengers = n(),
    Survived = sum(Survived == 1),
    Survival_Rate = (Survived / Total_Passengers) * 100
  )
print(age_gender_survival_rate)

   age_bin  Sex    Total_Passengers Survived Survival_Rate(%)
 1 [0,13)   female               32       19          59.4
 2 [0,13)   male                 37       21          56.8
 3 [13,19)  female               36       27          75
 4 [13,19)  male                 34        3          8.82
 5 [19,25)  female               49       39          79.6
 6 [19,25)  male                 90        9          10
 7 [25,36)  female               71       55          77.5
 8 [25,36)  male                148       34          23.0
 9 [36,51)  female               56       41          73.2
10 [36,51)  male                 97       20          20.6
11 [51,66)  female               17       16          94.1
12 [51,66)  male                 39        5          12.8
13 [66,100) male                  8        1          12.5

ggplot(age_gender_survival_rate, aes(x = age_bin, y = Survival_Rate, color =
Sex, group = Sex)) + geom_line(size = 1) + geom_point(size = 2) +
  labs(title = "Survival Rate by Age and Gender",
       x = "Age Group", y = "Survival Rate (%)", color = "Gender")

ggplot(age_gender_survival_rate, aes(x = age_bin, y = Survival_Rate, fill = S
ex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Survival Rate by Age Group and Gender", x = "Age Group",
    y = "Survival Rate (%)") + theme_minimal() +
  scale_fill_manual(values = c("lightcoral", "cornflowerblue"))
```





Interpretation

As indicated by line and bar chart the females had consistently higher survival rate as compared to men. This indicate the preference given to women during evacuation.

- **Females consistently have higher survival rates** across all age groups, particularly in the younger and older age groups.
- **Males show a marked decline in survival rates** as they age, with significantly lower survival chances in older age groups.
- **Younger females** (in the [0,13), [13,19), and [19,25) groups) have the highest survival rates, while **males** show a sharp decline in survival rates with age.

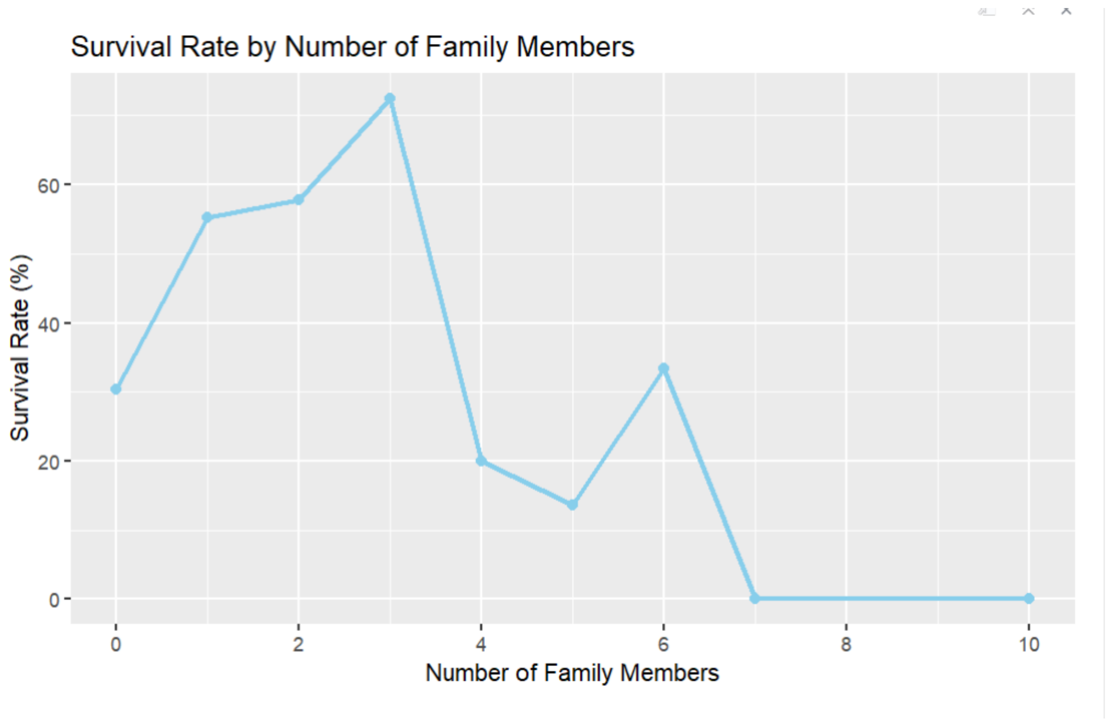
Survival Rate by Number of Family Members

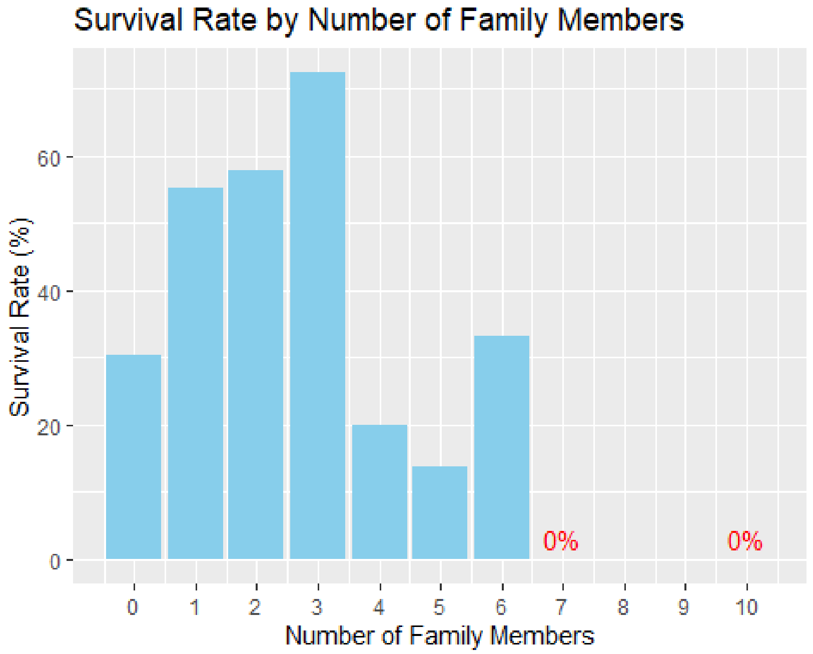
We also analyze how survival rates vary with the number of family members.

```
survival_rate_wrt_family_members <- df %>%
  group_by(family_members) %>%
  summarize(
    Total_Passengers = n(),
    Survived = sum(Survived == 1),
    Survival_Rate = (Survived / Total_Passengers) * 100
  )

print(survival_rate_wrt_family_members)
```

```
##   family_members Total_Passengers Survived Survival_Rate(%)
## 1              0              537      163          30.4
## 2              1              161       89          55.3
## 3              2              102       59          57.8
## 4              3               29       21          72.4
## 5              4               15        3          20
## 6              5               22        3          13.6
## 7              6               12        4          33.3
## 8              7                6        0           0
## 9             10                7        0           0
```

```
ggplot(survival_rate_wrt_family_members, aes(x = family_members, y = Survival
_Rate)) +
  geom_line(size = 1, color = "skyblue") +
  geom_point(size = 2, color = "skyblue") +
  labs(title = "Survival Rate by Number of Family Members",
       x = "Number of Family Members", y = "Survival Rate (%)",
       color = "Family Members") +
  scale_x_continuous(breaks = seq(
    min(survival_rate_wrt_family_members$family_members),
    max(survival_rate_wrt_family_members$family_members),
    by = 2))

ggplot(survival_rate_wrt_family_members, aes(x = family_members, y = Survival
_Rate)) +
    geom_bar(stat = "identity", position = "dodge", fill = "skyblue") +
    geom_text(data = subset(survival_rate_wrt_family_members, Survival_Rate =
= 0),
              aes(label = "0%"), vjust = -0.5, color = "red") +
    labs(title = "Survival Rate by Number of Family Members",
         x = "Number of Family Members",
         y = "Survival Rate (%)") +
    scale_x_continuous(breaks = seq(
        min(survival_rate_wrt_family_members$family_members),
        max(survival_rate_wrt_family_members$family_members),
        by = 1))
```





Interpretation

- **Survival Rate Increases with Fewer Family Members**: Passengers with no or fewer family members had higher survival rates compared to larger family groups.
- **Peak Survival with 3 Family Members**: The highest survival rate (72.4%) was observed for passengers traveling with 3 family members.

- **Sharp Decline for Larger Family Groups**: Survival rates dropped significantly for families with 4 or more members, with some groups experiencing 0% survival.
- **No Survival for Large Family Groups**: Passengers with 7 or 10 family members aboard had no survivors, suggesting that larger families faced more challenges during the disaster.

Measure of Central Tendencies and Dispersion

Load Necessary Libraries

```
install.packages("moments")  # For skewness and kurtosis

library(moments)
```

Central Tendency and Dispersion for Age

```
mean_age <- mean(titanic_data$Age, na.rm = TRUE)
median_age <- median(titanic_data$Age, na.rm = TRUE)
mode_age <- as.character(names(which.max(table(titanic_data$Age))))
variance_age <- var(titanic_data$Age, na.rm = TRUE)
sd_age <- sd(titanic_data$Age, na.rm = TRUE)

print("Mean:", mean_age, "\n")                   Mean: 29.69912
print("Median:", median_age, "\n")               Median: 28
print("Mode:", mode_age, "\n")                   Mode: 24
print("Variance:", variance_age, "\n")           Variance: 211.0191
print("Standard Deviation:", sd_age, "\n\n")     Standard Deviation: 14.5265
```

Interpretation

- The average age of passengers on the Titanic is approximately 29 years.
- The median age, which indicates the middle value, is 28 years. This means that 50% of passengers were below the age of 28 years and 50% of passengers are over the age of 28 years.
- The mode represents the most frequent age, which is `24 years`.
- The variance of the age variable is `211`, showing how much ages vary from the mean.
- The standard deviation is `14.5 years`, indicating the average distance of each age from the mean.

Central Tendency and Dispersion for SibSp

```
mean_sibsp <- mean(titanic_data$SibSp, na.rm = TRUE)
median_sibsp <- median(titanic_data$SibSp, na.rm = TRUE)
mode_sibsp <- as.character(names(which.max(table(titanic_data$SibSp))))
```

```
variance_sibsp <- var(titanic_data$SibSp, na.rm = TRUE)
sd_sibsp <- sd(titanic_data$SibSp, na.rm = TRUE)

print("Mean:", mean_sibsp)                Mean: 0.5230079
print("Median:", median_sibsp)            Median: 0
print("Mode:", mode_sibsp)                Mode: 0
print("Variance:", variance_sibsp)        Variance: 1.216043
print("Standard Deviation:", sd_sibsp)    Standard Deviation: 1.102743
```

Interpretation

- The **average number of siblings/spouses** aboard the Titanic is approximately **0.5**. This suggests that most passengers had less than one sibling or spouse on board.
- The **median number of siblings/spouses** is **0**. This indicates that 50% of passengers had no siblings or spouses traveling with them.
- The **mode** indicates that the most frequently occurring number of siblings/spouses aboard is **0**, meaning that it was common for passengers to travel alone.
- The **variance** of the SibSp variable is **1.21**, reflecting a low variability in the number of siblings/spouses among passengers. This suggests that most passengers had a similar number of siblings/spouses aboard.
- The **standard deviation** is approximately **1.1**. This indicates that the average distance from the mean (0.5) in the number of siblings/spouses is small, highlighting the consistency in this count among passengers.

Central Tendency and Dispersion for Parch

```
mean_parch <- mean(titanic_data$Parch, na.rm = TRUE)
median_parch <- median(titanic_data$Parch, na.rm = TRUE)
mode_parch <- as.character(names(which.max(table(titanic_data$Parch))))
variance_parch <- var(titanic_data$Parch, na.rm = TRUE)
sd_parch <- sd(titanic_data$Parch, na.rm = TRUE)

print("Mean:", mean_parch)                Mean: 0.3815937
print("Median:", median_parch)            Median: 0
print("Mode:", mode_parch)                Mode: 0
print("Variance:", variance_parch)        Variance: 0.6497282
print("Standard Deviation:", sd_parch)    Standard Deviation: 0.8060572
```

Interpretation

- The **average number of parents/children** aboard the Titanic is approximately **0.3**. This indicates that most passengers traveled without parents or children.
- The **median number of parents/children** is **0**, signifying that 50% of the passengers had no parents or children traveling with them.

- The **mode** for the number of parents/children is **0**, meaning that it was common for passengers to travel without any parents or children.
- The **variance** of the Parch variable is **0.6**, which suggests a low level of variability in the number of parents/children among passengers.
- The **standard deviation** is approximately **0.8**, indicating that the average distance from the mean (0.3) is small, reinforcing the notion that most passengers had a similar number of parents or children aboard.

Central Tendency and Dispersion for Fare

```
mean_fare <- mean(titanic_data$Fare, na.rm = TRUE)
median_fare <- median(titanic_data$Fare, na.rm = TRUE)
mode_fare <- as.character(names(which.max(table(titanic_data$Fare))))
variance_fare <- var(titanic_data$Fare, na.rm = TRUE)
sd_fare <- sd(titanic_data$Fare, na.rm = TRUE)

print("Mean:", mean_fare)                    Mean: 32.20421
print("Median:", median_fare)                Median: 14.4542
print("Mode:", mode_fare)                    Mode: 8.05
print("Variance:", variance_fare)            Variance: 2469.437
print("Standard Deviation:", sd_fare)        Standard Deviation: 49.69343
```

Interpretation

- The **average fare** paid by passengers on the Titanic is approximately **$32.20**.
- The **median fare** is **$14.45**, indicating that 50% of the passengers paid less than this amount, while the other half paid more.
- The **mode** of the fares is **$8.05**, meaning that this fare was the most commonly paid by passengers on board.
- The **variance** of the fare variable is **251.18**, indicating a substantial amount of variability in the fare prices. This suggests that there were both very low and very high fares among passengers.
- The **standard deviation** is approximately **$49.91**, highlighting that the average fare differs from the mean fare by this amount, indicating significant differences in fare prices among passengers.

Central Tendency for Qualitative Variables

Mode for Sex

```
mode_sex <- as.character(names(which.max(table(titanic_data$Sex))))

print("Mode:", mode_sex)                     Mode: male
```

Interpretation

The **mode** for the sex variable indicates that the most frequently occurring gender among passengers is **male**. This suggests that there were more male passengers compared to female passengers on board.

Mode for Embarked

```
mode_embarked <- as.character(names(which.max(table(titanic_data$Embarked))))

print("Mode:", mode_embarked)            Mode: S
```

Interpretation

The **mode** for the embarked variable indicates that the most frequently occurring embarkation point is **Southampton**. This suggests that a significant number of passengers boarded the Titanic from this port.

Mode for Passenger Class

```
mode_pclass <- as.character(names(which.max(table(titanic_data$Pclass))))

print("Mode:", mode_pclass)              Mode: 3
```

Interpretation

Most of the passengers had class 3 tickets

Mode for Survival

```
mode_survived <- as.character(names(which.max(table(titanic_data$Survived))))

print("Mode:", mode_survived)            Mode: 0
```

Interpretation

- Most of the passengers didn’t survive.

Measures of Location

Quantiles, Percentiles, and Deciles for Age

```
quantiles_age <- quantile(titanic_data$Age, na.rm = TRUE)
percentiles_age <- quantile(titanic_data$Age, probs = seq(0, 0.1, 0.01), na.rm = TRUE)
deciles_age <- quantile(titanic_data$Age, probs = seq(0, 1, 0.1), na.rm = TRUE)
```

```
print("Quantiles for Age: ")

print(quantiles_age)

     0%     25%     50%     75%    100% 
  0.420 20.125 28.000 38.000 80.000 

print("\nDeciles for Age: ")

print(deciles_age)

    0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100% 
  0.42 14.00 19.00 22.00 25.00 28.00 31.80 36.00 41.00 50.00 80.00 

print("\nPercentiles for Age: ")

print(percentiles_age)

    0%     1%     2%     3%     4%     5%     6%     7%     8%     9%    10% 
  0.42   1.00   2.00   2.00   3.00   4.00   5.00   7.91   9.00 11.00 14.00 
…
   70%     71%     72%     73%     74%     75%     76%     77%     78%     79%     80% 
36.000 36.000 36.000 37.000 38.000 38.000 39.000 39.000 40.000 40.135 41.000 
…
   95%     96%     97%     98%     99%    100% 
56.000 58.000 60.610 62.740 65.870 80.000 
```

Interpretation for Age Quantiles, Percentiles, and Deciles

- The **quantiles** show the distribution of ages. For instance, the 25th percentile (Q1) is approximately **20 years**, meaning that 25% of passengers are aged 20 years or younger, while the 75th percentile is around **38 years**, indicating that 75% of passengers are aged 38 years or younger.
- The **deciles** indicate that the first decile (D1) is around **20 years**, meaning that 10% of passengers are 20 years old or younger. The fifth decile (D5) or median confirms that 50% of passengers are aged 28 years or younger, while the ninth decile (D9) at **40 years** shows that 90% of passengers are aged 40 years or younger.
- The **percentiles** provide a detailed breakdown of age distribution, with the 90th percentile indicating that 80% of passengers were of age **41** or less indicating that most of passengers were not old aged.

Quantiles, Percentiles, and Deciles for Fare

```
quantiles_fare <- quantile(titanic_data$Fare, na.rm = TRUE)
percentiles_fare <- quantile(titanic_data$Fare, probs = seq(0, 1, 0.01), na.r
m = TRUE)
deciles_fare <- quantile(titanic_data$Fare, probs = seq(0, 1, 0.1), na.rm = T
RUE)
```

```
print("Quantiles for Fare:\n")

print(quantiles_fare)

       0%      25%      50%      75%     100%
   0.0000   7.9104  14.4542  31.0000 512.3292

print("\nDeciles for Fare:\n")

print(deciles_fare)

       0%      10%      20%      30%      40%      50%      60%      70%
   0.0000   7.5500   7.8542   8.0500  10.5000  14.4542  21.6792  27.0000
      80%      90%     100%
  39.6875  77.9583 512.3292

print("\nPercentiles for Fare:\n")

print(percentiles_fare)

        0%         1%         2%         3%         4%         5%         6%         7%
   0.00000    0.00000    6.39750    6.97500    7.05252    7.22500    7.22500    7.22920
        8%         9%        10%
   7.25000    7.25000    7.55000
...
       48%        49%        50%
  13.08334   14.01083   14.45420
```

Interpretation for Fare Quantiles, Percentiles, and Deciles

- The **quantiles** for fare show that the 25th percentile (Q1) is approximately **$7.9**, indicating that 25% of passengers paid $7.9 or less, while the 75th percentile (Q3) is around **$31.00**, suggesting that 75% of passengers paid $31.00 or less.

- The **deciles** reveal that the first decile (D1) is around **$7.55**, indicating that 10% of passengers paid this amount or less, while the fifth decile (D5) (or median) confirms that half of the passengers paid **$14.45** or less. The ninth decile (D9) at approximately **$77.95** shows that 90% of passengers paid this amount or less.

- The **percentiles** provide a detailed breakdown of fare distribution, with the 90th percentile indicating that 90% of passengers paid **$77.95** or less, highlighting the presence of a few passengers who paid significantly higher fares.

Skewness and Kurtosis

Skewness and Kurtosis for Age

```
skewness_age <- skewness(titanic_data$Age, na.rm = TRUE)
kurtosis_age <- kurtosis(titanic_data$Age, na.rm = TRUE)
```

```r
print("Skewness for Age:", skewness_age, "\n")
```

```
 Skewness for Age: 0.3882899
```

```r
print("Kurtosis for Age:", kurtosis_age, "\n")
```

```
 Kurtosis for Age: 3.168637
```

Interpretation

- **Skewness for Age**: The skewness of the age distribution is approximately `0.388`. A positive skewness value indicates that the distribution of ages is slightly **right-skewed**, meaning that there are more younger passengers, but a few older passengers pull the tail to the right.
- **Kurtosis for Age**: The kurtosis of the age distribution is approximately `3.16`. Since this value is close to 3, it indicates that the age distribution is **mesokurtic**, meaning it has a similar peak to a normal distribution, neither too flat nor too peaked.

Skewness and Kurtosis for Fare

```r
skewness_fare <- skewness(titanic_data$Fare, na.rm = TRUE)
kurtosis_fare <- kurtosis(titanic_data$Fare, na.rm = TRUE)

print("Skewness for Fare:", skewness_fare, "\n")
```

```
 Skewness for Fare: 4.779253
```

```r
print("Kurtosis for Fare:", kurtosis_fare, "\n")
```

```
 Kurtosis for Fare: 36.20429
```

Interpretation

- **Skewness for Fare**: The skewness of the fare distribution is approximately `4.79`. This high positive skewness indicates that the **fare distribution is highly right-skewed**, meaning that while most passengers paid lower fares, a few passengers paid very high fares, creating a long tail to the right.
- **Kurtosis for Fare**: The kurtosis for fare is approximately `36.2`. This high value indicates there are a few extreme outliers (passengers who paid very high fares).

Skewness and Kurtosis for SibSp

```r
skewness_sibsp <- skewness(titanic_data$SibSp, na.rm = TRUE)
kurtosis_sibsp <- kurtosis(titanic_data$SibSp, na.rm = TRUE)

print("Skewness for SibSp:", skewness_sibsp, "\n")
```

```
 Skewness for SibSp: 3.689128
```

```
print("Kurtosis for SibSp:", kurtosis_sibsp, "\n")

 Kurtosis for SibSp: 20.77351
```

Interpretation

- **Skewness for SibSp**: The skewness for SibSp is `3.6`, which suggests that the number of siblings/spouses aboard the Titanic is **positively skewed**, with most passengers having few or no siblings/spouses, but a few passengers had multiple.
- **Kurtosis for SibSp**: The kurtosis for SibSp is approximately `20.77`. This indicates that most passengers have similar numbers of siblings/spouses (often zero), with fewer passengers having much larger numbers of family members aboard.

Skewness and Kurtosis for Parch

```
skewness_parch <- skewness(titanic_data$Parch, na.rm = TRUE)
kurtosis_parch <- kurtosis(titanic_data$Parch, na.rm = TRUE)

print("Skewness for Parch:", skewness_parch, "\n")

 Skewness for Parch: 2.744487

print("Kurtosis for Parch:", kurtosis_parch, "\n")

 Kurtosis for Parch: 12.71661
```

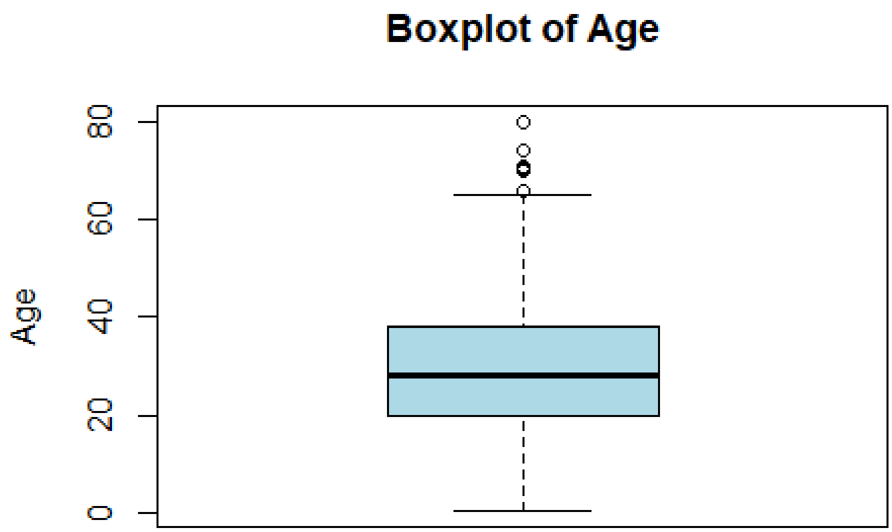
Interpretation

- **Skewness for Parch**: The skewness for Parch is `2.75`, which shows that the number of parents/children aboard the Titanic is **highly right-skewed**. Most passengers traveled without parents/children, while a few had multiple.
- **Kurtosis for Parch**: The kurtosis for Parch is approximately `12.7`. This suggests that the majority of passengers had no parents or children with them, but a small number had many, resulting in extreme outliers.

BoxPlots

Boxplot of Age

```
boxplot(titanic_data$Age, main = "Boxplot of Age", ylab = "Age", col = "light
blue", outline = TRUE)
```

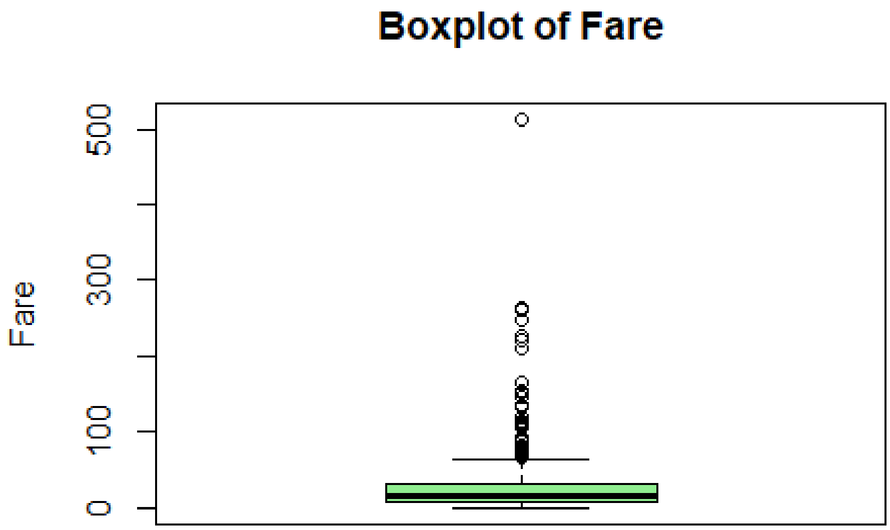


Interpretation:

- **Median**: The median age of passengers is around **28 years**, indicating that half of the passengers were younger than 28 and half were older.
- **Whiskers**: The whiskers extend to roughly **1 and 80 years**, suggesting the typical range of ages for passengers.
- **Outliers**: There are several **outliers** representing younger children and older individuals. These outliers are marked beyond the whiskers and show the presence of a few very young children (less than 10) and some passengers aged above 60.
- **Distribution Shape**: The longer whisker on the upper side suggests a slight **right skew**, indicating that there are more older passengers in the upper range than younger passengers in the lower range.

Boxplot of Fare

```
boxplot(titanic_data$Fare, main = "Boxplot of Fare", ylab = "Fare", col = "li
ghtgreen", outline = TRUE)
```

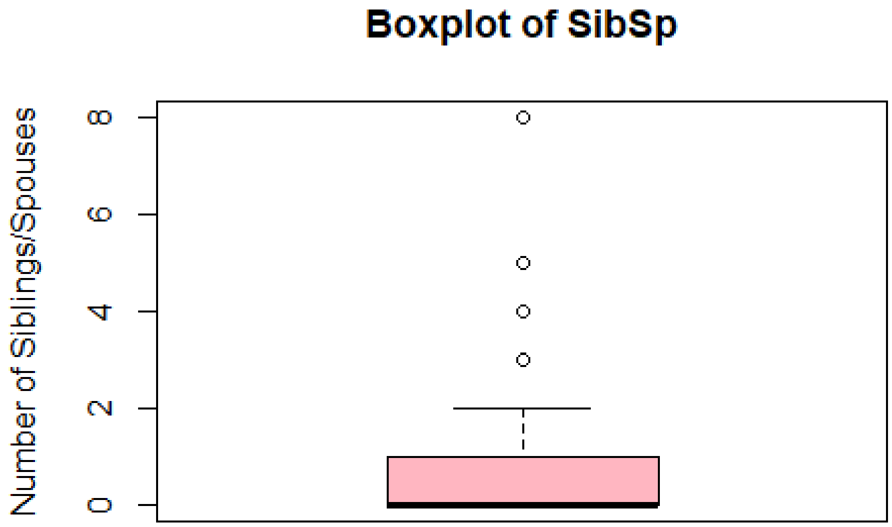


Interpretation:

- **Median**: The median fare is approximately **$14.5**, meaning half of the passengers paid less than $14.5, while the other half paid more.
- **Whiskers**: The lower whisker extends to about **$4**, while the upper whisker reaches around **$65**, showing that most passengers paid fares within this range.
- **Outliers**: There are many **outliers**, indicating that some passengers paid significantly higher fares. These outliers reflect first-class passengers who paid very high fares, well above the typical range (above $100).
- **Distribution Shape**: The **right-skewed** distribution is apparent from the many outliers and long upper whisker, reflecting that while most passengers paid lower fares, a few passengers paid extremely high amounts.

Boxplot of SibSp (Number of Siblings/Spouses Aboard)

```
boxplot(titanic_data$SibSp, main = "Boxplot of SibSp", ylab = "Number of Sibl
ings/Spouses", col = "lightpink", outline = TRUE)
```

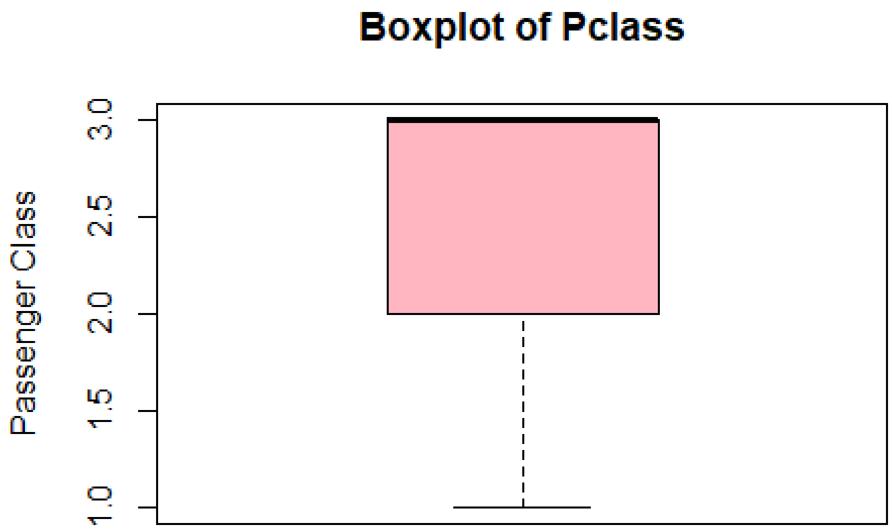


Interpretation:

- **Median**: The median number of siblings or spouses aboard is **0**, indicating that about half of the passengers traveled without a sibling or spouse.
- **Whiskers**: The whiskers indicate a range from **0 to 2**, showing that most passengers traveled with either none or a few siblings/spouses.
- **Outliers**: There are some **outliers**, representing families who had 4 or more siblings/spouses aboard. These larger families were relatively uncommon.
- **Distribution Shape**: The boxplot reveals a **right-skewed** distribution, as indicated by the outliers and the longer whisker on the upper side. This suggests that while most passengers traveled with no or few siblings/spouses, a small number traveled with several.

Boxplot for Pclass

```
boxplot(titanic_data$Pclass, main = "Boxplot of Pclass", ylab = "Passenger Cl
ass", col = "lightpink", outline = TRUE)
```



Interpretation

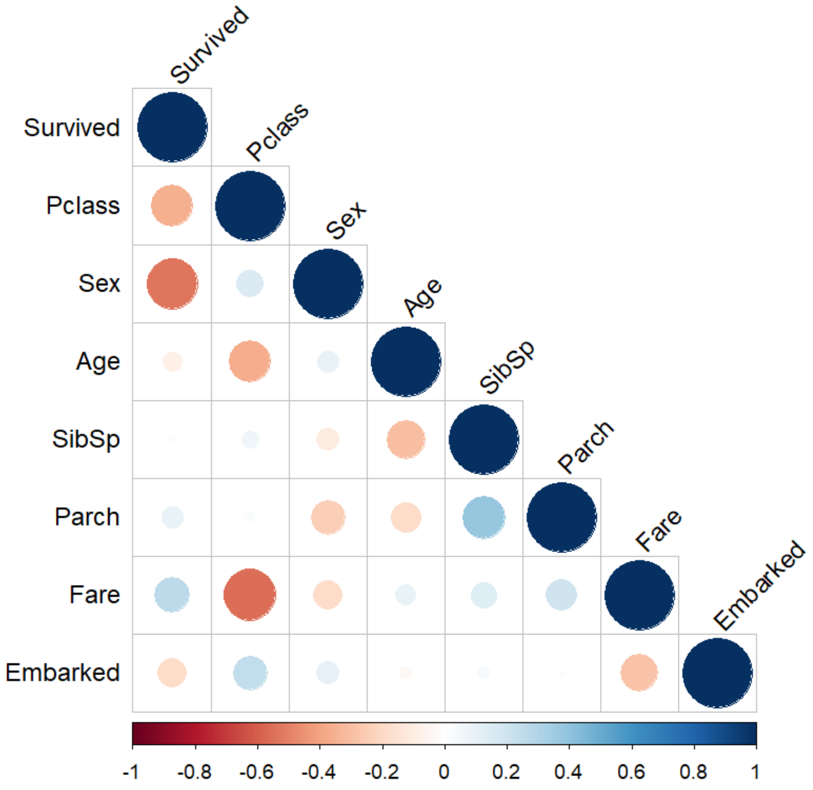
- **Median**: The median Passenger Class is 3, indicating that half of the passengers belonged to the 3rd class. This suggests that the majority of passengers were traveling in a lower class.
- **Interquartile Range (IQR)**: The box represents the middle 50% of the data. The IQR ranges between Passenger Classes 2 and 3. This implies that most passengers were concentrated in the 2nd and 3rd classes.
- **Whiskers**: The whiskers extend down to class 1 (indicating the minimum value) and up to class 3 (indicating the maximum value). This shows that the Passenger Class variable has three discrete values (1st, 2nd, and 3rd class), with no outliers outside this range.
- **Distribution Shape**: The boxplot is not continuous, and since Passenger Class is a discrete variable with only three levels, the boxplot visually represents this structure. The larger height of the box in the upper range (closer to 3) suggests that a greater proportion of passengers belonged to the 3rd class.

Regression Analysis

```r
library(corrplot)
library(ggplot2)
library(dplyr)

df <- read.csv("D:\\R-statistics\\titanic.csv")
```

Simple Linear Correlation Coefficient

```
df_mod <- df %>% select(-PassengerId) # drop PassengerId col
df_mod$Embarked <- as.numeric(factor(df_mod$Embarked, levels = c("C", "Q", "S
")))
# convert gender col to int
df_mod$Sex <- ifelse(df$Sex == "male", 1, 0)

numeric_columns <- df_mod %>% select_if(is.numeric)

# Calculate correlation matrix for all numerical columns including 'Survived'
cor_matrix_with_survival <- cor(numeric_columns, use = "complete.obs")
print(round(cor_matrix_with_survival, 2))

##          Survived Pclass   Sex   Age SibSp Parch  Fare Embarked
## Survived     1.00  -0.36 -0.54 -0.08 -0.02  0.10  0.27    -0.18
## Pclass      -0.36   1.00  0.15 -0.37  0.07  0.02 -0.55     0.24
## Sex         -0.54   0.15  1.00  0.10 -0.11 -0.25 -0.18     0.11
## Age         -0.08  -0.37  0.10  1.00 -0.31 -0.19  0.09    -0.03
## SibSp       -0.02   0.07 -0.11 -0.31  1.00  0.38  0.14     0.03
## Parch        0.10   0.02 -0.25 -0.19  0.38  1.00  0.21     0.01
## Fare         0.27  -0.55 -0.18  0.09  0.14  0.21  1.00    -0.28
## Embarked    -0.18   0.24  0.11 -0.03  0.03  0.01 -0.28     1.00

corrplot(cor_matrix_with_survival, method = "circle", type = "lower",
tl.col = "black", tl.srt = 45)
```



Interpretation

- The given figure shows nature and level of relationship b/w any 2 variables.

1. Survived and Other Variables:

- **Survived and Pclass**: **-0.36** (Negative): Passengers in higher classes (Pclass 1) had a better chance of survival, while those in lower classes (Pclass 3) had a lower survival rate.
- **Survived and Sex**: **-0.54** (Strong negative): Females (coded as 0) had a significantly higher chance of survival compared to males (coded as 1).
- **Survived and Age**: **-0.08** (Very weak negative): Age had a minimal effect on survival, showing no significant relationship.
- **Survived and SibSp**: **-0.02** (Very weak negative): Number of siblings/spouses aboard had almost no effect on survival.
- **Survived and Parch**: **0.10** (Weak positive): Passengers with more parents/children aboard had a slightly higher chance of survival.
- **Survived and Fare**: **0.27** (Moderate positive): Passengers who paid higher fares had a slightly better chance of survival.
- **Survived and Embarked**: **-0.18** (Negative): Passengers who embarked from **S** (Southampton) had a slightly lower chance of survival compared to those who embarked from **C** (Cherbourg) or **Q** (Queenstown).

2. Pclass and Other Variables:

- **Pclass and Sex**: **0.15** (Weak positive): Females were more likely to be in higher classes, while males were more likely to be in lower classes.
- **Pclass and Age**: **-0.37** (Negative): Older passengers were more likely to be in lower classes.
- **Pclass and SibSp**: **0.07** (Very weak positive): Very little relationship between the number of siblings/spouses aboard and the class.
- **Pclass and Parch**: **0.02** (Very weak positive): No significant relationship between the number of parents/children aboard and class.
- **Pclass and Fare**: **-0.55** (Strong negative): Passengers in higher classes (Pclass 1) paid significantly higher fares compared to passengers in lower classes.
- **Pclass and Embarked**: **0.24** (Positive): Passengers embarking from **C** (Cherbourg) or **Q** (Queenstown) were more likely to be in higher classes compared to those from **S** (Southampton).

3. Sex and Other Variables:

- **Sex and Age**: **0.10** (Weak positive): A slight tendency for older passengers to be male.
- **Sex and SibSp**: **-0.11** (Weak negative): Males tended to have fewer siblings/spouses aboard than females.
- **Sex and Parch**: **-0.25** (Moderate negative): Males had fewer parents/children aboard compared to females.
- **Sex and Fare**: **-0.18** (Moderate negative): Males generally paid lower fares than females.
- **Sex and Embarked**: **0.11** (Positive): Females were slightly more likely to embark from **C** (Cherbourg) or **Q** (Queenstown) compared to **S** (Southampton).

4. Age and Other Variables:

- **Age and SibSp**: **-0.31** (Negative): Younger passengers were more likely to travel with siblings/spouses.
- **Age and Parch**: **-0.19** (Weak negative): Younger passengers tended to have fewer parents/children aboard than older passengers.
- **Age and Fare**: **0.09** (Weak positive): Older passengers tend to pay slightly higher fares.
- **Age and Embarked**: **-0.03** (Very weak negative): No significant relationship between age and embarkation port.

5. SibSp and Other Variables:

- **SibSp and Survived**: **-0.02** (Very weak negative): The number of siblings or spouses aboard had almost no effect on survival.
- **SibSp and Pclass**: **0.07** (Very weak positive): Passengers with more siblings or spouses tended to be slightly more likely to travel in higher classes, but the relationship is very weak.
- **SibSp and Sex**: **-0.11** (Weak negative): Males tended to have fewer siblings or spouses aboard compared to females.
- **SibSp and Age**: **-0.31** (Moderate negative): Younger passengers were more likely to travel with siblings or spouses, while older passengers had fewer siblings or spouses.
- **SibSp and Parch**: **0.38** (Moderate positive): Passengers who had more siblings or spouses aboard tended to also have more parents or children with them.
- **SibSp and Fare**: **0.14** (Weak positive): Passengers with more siblings or spouses tended to pay slightly higher fares.

- **SibSp and Embarked**: **0.03** (Very weak positive): There is almost no relationship between the number of siblings/spouses aboard and the embarkation port.

6. Parch and Other Variables:

- **Parch and Survived**: **0.10** (Weak positive): Passengers with more parents or children aboard tended to have a slightly higher chance of survival.
- **Parch and Pclass**: **0.02** (Very weak positive): Very little relationship between the number of parents/children aboard and the class of travel.
- **Parch and Sex**: **-0.25** (Moderate negative): Males tended to have fewer parents or children aboard compared to females.
- **Parch and Age**: **-0.19** (Weak negative): Younger passengers tended to have fewer parents or children aboard compared to older passengers.
- **Parch and SibSp**: **0.38** (Moderate positive): Passengers with more parents or children aboard were more likely to have more siblings or spouses aboard.
- **Parch and Fare**: **0.21** (Weak positive): Passengers with more parents or children aboard tended to pay slightly higher fares.
- **Parch and Embarked**: **0.01** (Very weak positive): No significant relationship between the number of parents/children aboard and embarkation port.

7. Fare and Other Variables:

- **Fare and Pclass**: **-0.55** (Strong negative): Passengers in higher classes (Pclass 1) paid significantly higher fares than those in lower classes (Pclass 3).
- **Fare and Sex**: **-0.18** (Moderate negative): Males paid lower fares on average compared to females.
- **Fare and Age**: **0.09** (Weak positive): Older passengers tended to pay slightly higher fares.
- **Fare and SibSp**: **0.14** (Weak positive): Passengers with more siblings/spouses aboard tended to pay slightly higher fares.
- **Fare and Parch**: **0.21** (Weak positive): Passengers with more parents/children aboard tended to pay slightly higher fares.
- **Fare and Embarked**: **-0.28** (Moderate negative): Passengers embarking from **S** (Southampton) paid lower fares compared to those embarking from **C** (Cherbourg) or **Q** (Queenstown).

8. Embarked and Other Variables:

- **Embarked and Pclass**: **0.24** (Positive): Passengers embarking from **C** (Cherbourg) or **Q** (Queenstown) were more likely to be in higher classes (Pclass 1 or 2) compared to passengers from **S** (Southampton).

- **Embarked and Sex**: **0.11** (Positive): Females were slightly more likely to embark from **C** (Cherbourg) or **Q** (Queenstown) than males.
- **Embarked and Age**: **-0.03** (Very weak negative): No significant relationship between age and embarkation port.
- **Embarked and SibSp**: **0.03** (Very weak positive): No significant relationship between the number of siblings/spouses and embarkation port.
- **Embarked and Parch**: **0.01** (Very weak positive): No significant relationship between the number of parents/children and embarkation port.

Gender and Survival: The most significant finding is that **females had a higher likelihood of survival** (strong negative correlation with **Survived**).

Fare and Embarkation: The **embarkation port** also showed some correlation with fares, with those embarking from **S (Southampton)** paying lower fares on

Passenger Class and Fare: **Lower-class passengers** (Pclass) paid significantly lower fares compared to those in higher classes, and this is strongly reflected in the fare distribution.

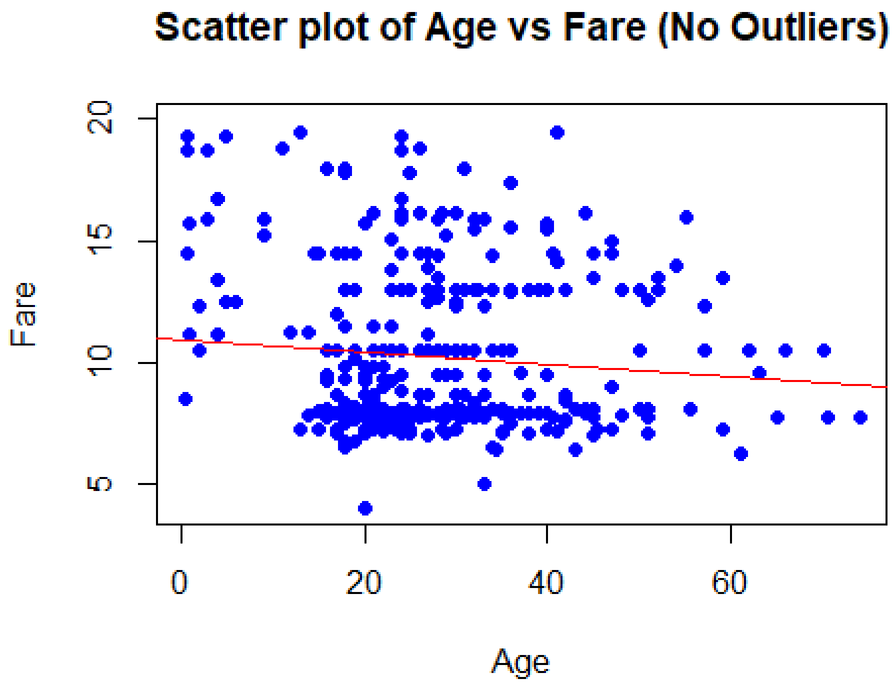
Regression Analysis

For our analysis, we will focus on Fare and Age, as these are the only continuous variables available in the dataset even though there is no strong relationship between these 2, while the rest are categorical.

Initially, we will visualize the relationship between these two variables and perform a detailed analysis. Subsequently, we will expand our approach by incorporating multiple variables into the model, aiming to achieve higher R^2 value.

Fare & Age

```r
# here we will apply fare limit of (0, 20] to reduce outliers
df_mod_no_outliers =  df_mod[df_mod$Fare > 0 & df_mod$Fare <= 20, ]
# scatter plot after removing outliers
plot(df_mod_no_outliers$Age, df_mod_no_outliers$Fare, main = "Scatter plot of Age vs Fare (No Outliers)", xlab = "Age", ylab = "Fare", pch = 19, col = "blue")

# Fit a linear regression model
linear_model_no_outliers <- lm(Fare ~ Age, data = df_mod_no_outliers)

# Add regression line to the plot
abline(linear_model_no_outliers, col = "red")
```



```
summary(linear_model_no_outliers)


## Residuals:
##    Min     1Q Median     3Q    Max
## -6.415 -2.480 -1.589  2.712  9.606
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.93616    0.41436  26.393   <2e-16 ***
## Age         -0.02541    0.01361  -1.867   0.0626 .
## ---
##
## Residual standard error: 3.263 on 382 degrees of freedom
##   (116 observations deleted due to missingness)
## Multiple R-squared:  0.009046,   Adjusted R-squared:  0.006451
## F-statistic: 3.487 on 1 and 382 DF,  p-value: 0.06262

# Coefficient of Determination (R²)
r_squared_no_outliers <- summary(linear_model_no_outliers)$r.squared
cat("R²: ", r_squared_no_outliers, "\n")

R²:  0.009045579
```

Interpretation

`lm()` by default uses Least Squared Method to fit the linear regression model as this method minimizes the error to find the best-fitting line.

- **Intercept (Estimate: 10.94)**:
  - When `Age` is 0, the expected value of `Fare` is approximately 10.94. Although, this may not be practically meaningful since `Age` cannot be zero in this dataset.
- **Age Coefficient (Estimate: -0.02541)**:
  - For each one-unit increase in `Age`, the `Fare` is expected to decrease by approximately 0.025 units. This indicates a negative relationship between `Age` and `Fare`.
- **R-squared (0.009)**:
  - The R-squared value of 0.009 means that only about **0.9%** of the variation in `Fare` is explained by `Age`. This is very low, suggesting that `Age` alone is not a strong predictor of `Fare`.
- **Residuals**:
  - The residuals represent the difference between the observed values and the predicted values of `Fare`. The distribution of residuals shows:
    - **Minimum residual**: -6.415
    - **1st Quartile (25%)**: -2.480
    - **Median (50%)**: -1.589
    - **3rd Quartile (75%)**: 2.712
    - **Maximum residual**: 9.606

  This range indicates that the model’s predictions are off by up to 9.606 units in either direction.
- **S.D. of Random Errors (Residual Standard Error)**:
  - The **residual standard error** (S.E.) is 3.263. This is the standard deviation of the residuals, and it represents the typical size of the error in predicting `Fare` based on `Age`. A lower value indicates a better fit, but here we see that there is room for improvement.
- **Degrees of Freedom**:
  - In a simple linear regression model, the **degrees of freedom** (df) for the residuals is the total number of observations minus 2 (one for the intercept and one for the slope). Here, we have 382 degrees of freedom, which means that there are 382 data points used to estimate the model’s parameters.
- **Hypothesis Testing Result:**

  The key output here is the t-statistic and the p-value for the `Age` variable.

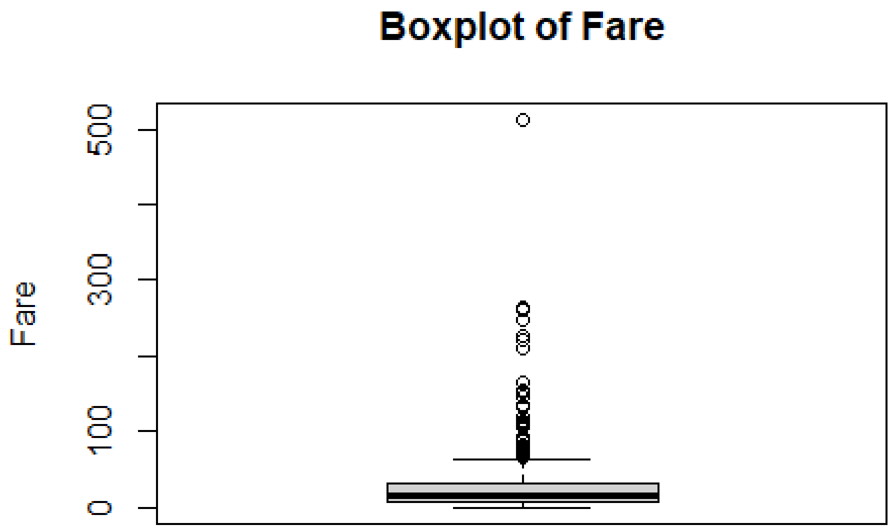
- **Null Hypothesis (H_0)**: $\beta Age=0$ (Age has no effect on Fare)
- **Alternative Hypothesis (H_1)**: $\beta Age\neq0$ (Age affects Fare)
- **p-value for Age (0.0626)**: Since the p-value is greater than the typical significance level of **0.05**, we **fail to reject** the null hypothesis. This means there is not enough evidence to conclude that `Age` significantly affects `Fare` at the 5% significance level. However, the result is still marginally significant at the 10% significance level.
- **Conclusion**: Based on this hypothesis test, we conclude that there is insufficient statistical evidence to suggest that `Age` is a significant predictor of `Fare`.

Multiple variables Regression Analysis

```
linear_model <- lm(Fare ~ Pclass + Sex + Age + SibSp + Parch + Embarked, data
= df_mod)

# Display the summary of the new linear regression model
summary(linear_model)

## Residuals:
##    Min     1Q Median     3Q    Max 
## -64.65 -18.66  -3.31  11.05 428.07 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 137.4414     8.1890  16.784  < 2e-16 ***
## Pclass      -33.9735     2.1004 -16.175  < 2e-16 ***
## Sex          -3.0279     3.4186  -0.886  0.37608    
## Age          -0.1594     0.1228  -1.298  0.19472    
## SibSp         5.6466     1.8772   3.008  0.00272 ** 
## Parch        10.4151     2.0306   5.129 3.77e-07 ***
## Embarked    -10.6048     2.0711  -5.120 3.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 41.5 on 705 degrees of freedom
##   (179 observations deleted due to missingness)
## Multiple R-squared:  0.3907, Adjusted R-squared:  0.3856 
## F-statistic: 75.36 on 6 and 705 DF,  p-value: < 2.2e-16

# Coefficient of Determination (R²)
r_squared <- summary(linear_model)$r.squared
print(paste("R² after removing outliers: ", r_squared))

R² after removing outliers:  0.390744832824502
```

Since R^2 has a lower value we check for existance of outliers that may be effecting the calculation.

```
# Visualize Fare distribution using a boxplot
boxplot(df_mod$Fare, main = "Boxplot of Fare", ylab = "Fare")
```



Since we have outliers we will drop them and refit the model.

```
# Calculate Q1 (25th percentile) and Q3 (75th percentile)
Q1 <- quantile(df_mod$Fare, 0.25)
Q3 <- quantile(df_mod$Fare, 0.75)

IQR_value <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR_value
upper_bound <- Q3 + 1.5 * IQR_value

# Remove rows with Fare outside the bounds (outliers)
df_mod_no_outliers <- df_mod[df_mod$Fare >= lower_bound & df_mod$Fare <= uppe
r_bound, ]

print(paste("Rows before removing outliers: ", nrow(df_mod)))
```

```
Rows before removing outliers:  891
```

```
print(paste("Rows after removing outliers: ", nrow(df_mod_no_outliers)))
```

```
Rows after removing outliers:  775
```

```
# Fit the linear regression model again after removing outliers
linear_model_no_outliers <- lm(Fare ~ Pclass + Sex + Age + SibSp + Parch + Em
barked, data = df_mod_no_outliers)
```

```
# Display the summary of the new linear regression model
summary(linear_model_no_outliers)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.786  -4.896   0.213   1.971  49.603
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.24174    2.05782  22.957  < 2e-16 ***
## Pclass      -12.24817    0.50535 -24.237  < 2e-16 ***
## Sex           0.34218    0.76032   0.450  0.65284
## Age           0.02832    0.02741   1.033  0.30180
## SibSp         5.26747    0.40539  12.994  < 2e-16 ***
## Parch         4.55016    0.45221  10.062  < 2e-16 ***
## Embarked     -1.56087    0.48659  -3.208  0.00141 **
## ---
##
## Residual standard error: 8.333 on 606 degrees of freedom
##   (162 observations deleted due to missingness)
## Multiple R-squared:  0.6373, Adjusted R-squared:  0.6337
## F-statistic: 177.5 on 6 and 606 DF,  p-value: < 2.2e-16

# Coefficient of Determination (R²)
r_squared_no_outliers <- summary(linear_model_no_outliers)$r.squared
print(paste("R² after removing outliers: ", r_squared_no_outliers))

R² after removing outliers:  0.637279008493999
```

Before Outlier Removal:

- **Model Fit**: The R^2 value was 0.39, indicating that the model explained only 39% of the variance in the Fare variable.
- **Residuals**: Residual standard error was 41.5, suggesting large variations between observed and predicted values, likely due to the influence of outliers.

After Outlier Removal:

- **Improved Model Fit**: The R^2 value increased to 0.637, meaning the model now explains approximately 63.7% of the variance in Fare, a significant improvement.
- **Reduced Residuals**: The residual standard error dropped to 8.333, reflecting better predictive performance with smaller errors.

Interpretation

- The removal of outliers significantly improved the model's performance, as evidenced by the increase in R^2 and decrease in residual error.

- The final model explains **63.7% of the variation in Fare**, with the remaining **36.3%** attributable to factors not included in the model or random error.

- The strongest predictor of Fare is Pclass, with a negative coefficient indicating that higher-class passengers pay significantly more.
- SibSp and Parch positively influence Fare, showing that families or groups tend to pay higher overall fares.
- Embarked had a modest negative impact, potentially reflecting differences in fare structures across embarkation points.

Hypothesis Testing

- **Null Hypothesis (H_0)**: βi=0 (The predictor does not have an effect on the outcome)
- **Alternative Hypothesis (H_1)**: βi≠0 (The predictor has an effect on the outcome)

➢ *Pclass:*

- **Coefficient**: −12.24817
- **t-value**: −24.237
- **p-value**: <2e−16

Interpretation: The p-value is extremely small, which means that `Pclass` is highly statistically significant in predicting `Fare`. We reject the null hypothesis, meaning `Pclass` has a significant effect on `Fare`.

➢ ***Sex:***

- **Coefficient**: 0.34218
- **t-value**: 0.450
- **p-value**: 0.65284

Interpretation: The p-value for `Sex` is 0.65284, which is greater than the typical significance level of 0.05. Therefore, we **fail to reject** the null hypothesis, meaning `Sex` does not have a statistically significant effect on `Fare` in this model.

➢ ***Age:***

- **Coefficient**: 0.02832
- **t-value**: 1.033
- **p-value**: 0.3018

Interpretation: The p-value for `Age` is 0.3018, which is greater than 0.05. This means we **fail to reject** the null hypothesis for `Age`, suggesting that `Age` does not have a statistically significant effect on `Fare` in this model.

➢ ***SibSp:***

- **Coefficient**: 5.26747
- **t-value**: 12.994
- **p-value**: <2e−16

Interpretation: The p-value for `SibSp` is extremely small, indicating that `SibSp` is highly statistically significant. We reject the null hypothesis and conclude that `SibSp` has a significant effect on `Fare`.

- ***Parch:***
  - **Coefficient**: 4.55016
  - **t-value**: 10.062
  - **p-value**: <2e−16

  Interpretation: The p-value for `Parch` is extremely small, which suggests that `Parch` is statistically significant in predicting `Fare`. We reject the null hypothesis for `Parch`, meaning `Parch` has a significant effect on `Fare`.
- ***Embarked:***
  - **Coefficient**: −1.56087
  - **t-value**: −3.208
  - **p-value**: 0.00141
  - **Interpretation**: The p-value for `Embarked` is 0.00141, which is less than 0.05. This suggests that `Embarked` is statistically significant in predicting `Fare`. We reject the null hypothesis for `Embarked`, meaning `Embarked` has a significant effect on `Fare`.

Statistical Inference

Age

```
# Set seed for reproducibility
set.seed(123)

# Randomly sample 100 observations from Age
sample_age <- sample(df$Age, size = 100, replace = FALSE)

# Estimate the population standard deviation using the entire dataset
population_sd <- sd(df$Age, na.rm = TRUE)

# Calculate the sample mean
sample_mean <- mean(sample_age, na.rm = TRUE)
n <- length(sample_age) # Sample size

margin_of_error = 1.96 * (population_sd / sqrt(n))
print(paste("Estimated Population Standard Deviation (σ):", round(population_
sd, 2)))
```

```
## [1] "Estimated Population Standard Deviation (σ): 14.53"
```

```
print(paste("Point Estimate (Mean):", round(sample_mean, 2)))
```

```
## [1] "Point Estimate (Mean): 29.99"

print(paste("Margin of Error:", round(margin_of_error, 2)))

## [1] "Margin of Error: 2.85"

# Calculate the Z-value for a 90% confidence level
alpha <- 0.10
z_value <- qnorm(1 - alpha / 2)

# Calculate the margin of error for CI
margin_of_error_ci <- z_value * (population_sd / sqrt(n))

# Calculate the confidence interval
lower_bound <- sample_mean - margin_of_error_ci
upper_bound <- sample_mean + margin_of_error_ci

print(paste("90% Confidence Interval: (", round(lower_bound, 2), ",", round(u
pper_bound, 2), ")"))

## [1] "90% Confidence Interval: ( 27.6 , 32.38 )"
```

Interpretation

- **Point Estimate = 29.99** is the **sample mean** age of the passengers. This is the best estimate of the **true average age** of the entire population based on the sample with margin of error of **2.85 years** means that true population mean could differ from the sample mean by up to **2.85 years** due to sampling variability.

- The **90% confidence interval** of **(27.6, 32.38)** means that, based on sample, we are **90% confident** that the true average age of all passengers in the population falls between **27.6 years** and **32.38 years**.

Fare

```
# Set seed for reproducibility
set.seed(123)

sample_fare <- sample(df$Fare, size = 100, replace = FALSE)
population_sd <- sd(df$Fare, na.rm = TRUE)
sample_mean_fare <- mean(sample_fare, na.rm = TRUE)
n <- length(sample_fare) # Sample size
margin_of_error = 1.96 * (population_sd / sqrt(n))
print(paste("Estimated Population Standard Deviation (σ):", round(population_
sd, 2)))

## [1] "Estimated Population Standard Deviation (σ): 49.69"

print(paste("Point Estimate (Mean):", round(sample_mean_fare, 2)))

## [1] "Point Estimate (Mean): 28.8"
```

```
print(paste("Margin of Error:", round(margin_of_error, 2)))

## [1] "Margin of Error: 9.74"


alpha <- 1 - 0.98
z_value <- qnorm(1 - alpha / 2)
margin_of_error_ci <- z_value * (population_sd / sqrt(n))
lower_bound <- sample_mean_fare - margin_of_error_ci
upper_bound <- sample_mean_fare + margin_of_error_ci

print(paste("98% Confidence Interval: (", round(lower_bound, 2), ",", round(u
pper_bound, 2), ")"))

## [1] "98% Confidence Interval: ( 17.24 , 40.36 )"
```

Interpretation

- **Point Estimate = 28.8** is the **sample mean** fare that passengers paid. This is the best estimate of the **true average fare** of the entire population based on the sample with margin of error of **$9.74** means that true population mean could differ from the sample mean by up to **9.74** due to sampling variability.

- The **98% confidence interval** of **(17.24, 40.36)** means that, based on sample, we are **98% confident** that the true average fare of all passengers in the population falls between $17.24 - **$40.36**.