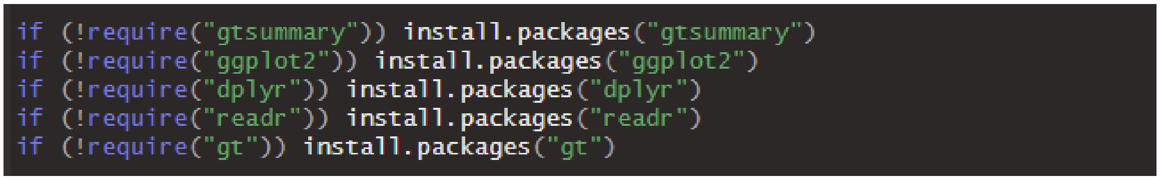
Assignment2

Niramai Arkhom

- **How to write coding**

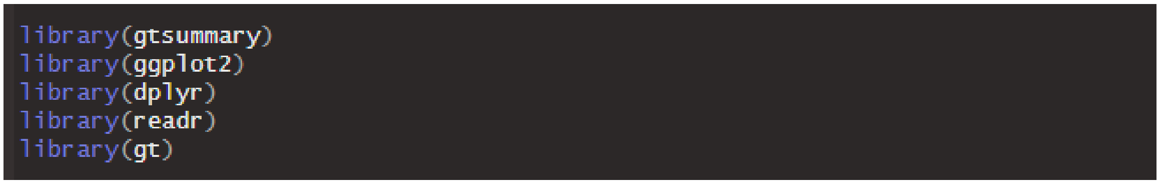
1. Loading and Installing Required Packages



```
if (!require("gtsummary")) install.packages("gtsummary")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("dplyr")) install.packages("dplyr")
if (!require("readr")) install.packages("readr")
if (!require("gt")) install.packages("gt")
```

The code checks if the required R packages are installed on your system (gtsummary, ggplot2, dplyr, readr, and gt). If they are not already installed, it will automatically install them. require() loads the package if installed, and install.packages() installs it if necessary.

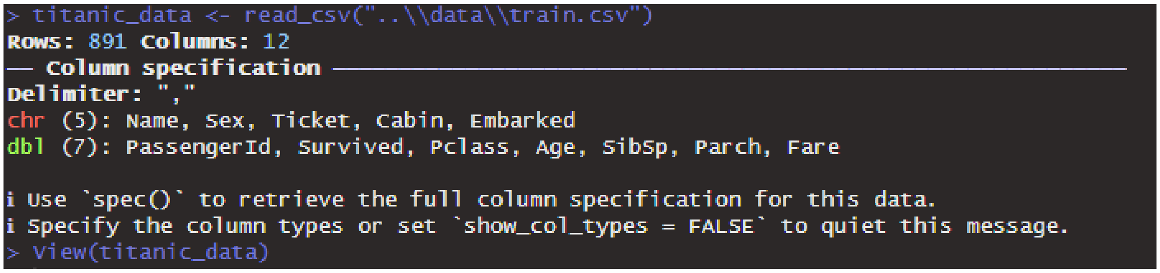
2. Loading Libraries



```
library(gtsummary)
library(ggplot2)
library(dplyr)
library(readr)
library(gt)
```

After ensuring that the necessary packages are installed, this step loads the libraries into your R environment to access their functions.

3. Import and Reading the Titanic Dataset



```
> titanic_data <- read_csv("..\\data\\train.csv")
Rows: 891 Columns: 12
— Column specification ————————————————————————————————————————
Delimiter: ","
chr (5): Name, Sex, Ticket, Cabin, Embarked
dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
> View(titanic_data)
```

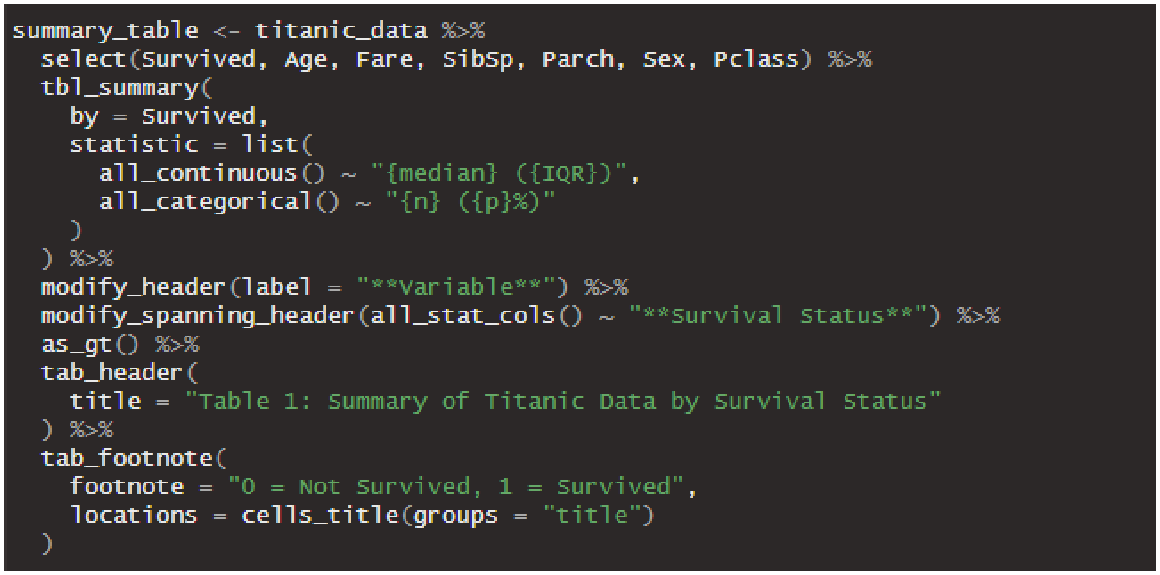
4. Creating a Summary Table Grouped by 'Survived'

This block creates a summary table of selected variables (Survived, Age, Fare, SibSp, Parch, Sex, and Pclass) grouped by the Survived column.

Step-by-step:

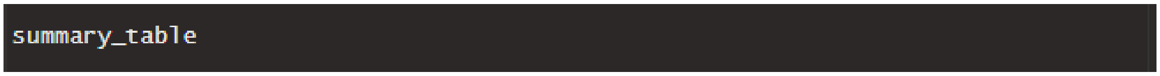
- select() filters the specified columns from the dataset.

- tbl_summary() from the gtsummary package generates summary statistics for continuous and categorical variables, grouping by Survived.
    - Continuous variables (like Age and Fare) are summarized using the median and interquartile range (IQR).
    - Categorical variables (like Sex and Pclass) are summarized using counts (n) and percentages (p).
- modify_header() renames the header for the first column.
- modify_spanning_header() groups the columns under the label "**Survival Status**".
- as_gt() converts the summary table into a format supported by the gt package for better visualization.
- tab_header() adds a title to the summary table.
- tab_footnote() adds a footnote explaining the survival codes (0 = Not Survived, 1 = Survived).



```
summary_table <- titanic_data %>%
  select(Survived, Age, Fare, SibSp, Parch, Sex, Pclass) %>%
  tbl_summary(
    by = Survived,
    statistic = list(
      all_continuous() ~ "{median} ({IQR})",
      all_categorical() ~ "{n} ({p}%)"
    )
  ) %>%
  modify_header(label = "**Variable**") %>%
  modify_spanning_header(all_stat_cols() ~ "**Survival Status**") %>%
  as_gt() %>%
  tab_header(
    title = "Table 1: Summary of Titanic Data by Survival Status"
  ) %>%
  tab_footnote(
    footnote = "0 = Not Survived, 1 = Survived",
    locations = cells_title(groups = "title")
  )
```

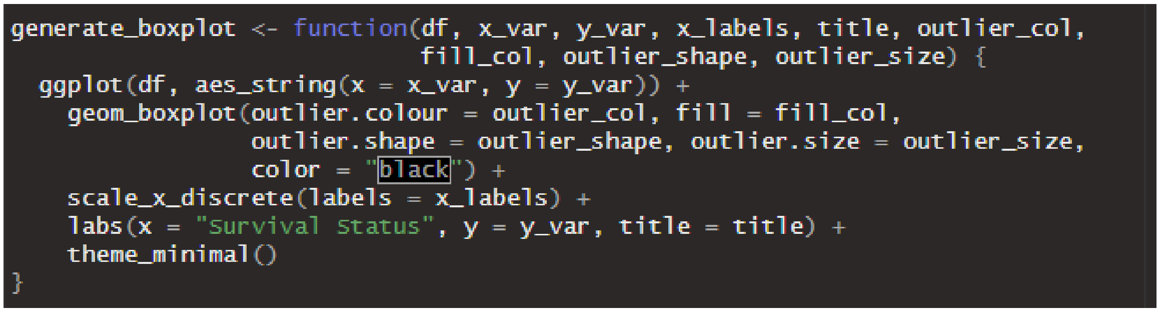
5. Displaying the Summary Table



```
summary_table
```

This command displays the summary table created in the previous step in the RStudio Viewer.

6. Creating a Function to Generate Boxplots



```
generate_boxplot <- function(df, x_var, y_var, x_labels, title, outlier_col,
                             fill_col, outlier_shape, outlier_size) {
  ggplot(df, aes_string(x = x_var, y = y_var)) +
    geom_boxplot(outlier.colour = outlier_col, fill = fill_col,
                 outlier.shape = outlier_shape, outlier.size = outlier_size,
                 color = "black") +
    scale_x_discrete(labels = x_labels) +
    labs(x = "Survival Status", y = y_var, title = title) +
    theme_minimal()
}
```

This function, generate_boxplot(), is used to generate boxplots. It takes various arguments to customize the appearance of the boxplots.

Arguments:

- df: The data frame (in this case, titanic_data).
- x_var: The x-axis variable (in this case, Survived).
- y_var: The y-axis variable (e.g., Age, Fare).
- x_labels: Custom labels for the x-axis (to convert 0 and 1 into "Not Survived" and "Survived").
- title: The title of the plot.
- outlier_col: The color of the outliers.
- fill_col: The fill color for the boxplot.
- outlier_shape: The shape of the outliers.
- outlier_size: The size of the outliers.

7. Creating and Displaying Boxplots for Age and Fare

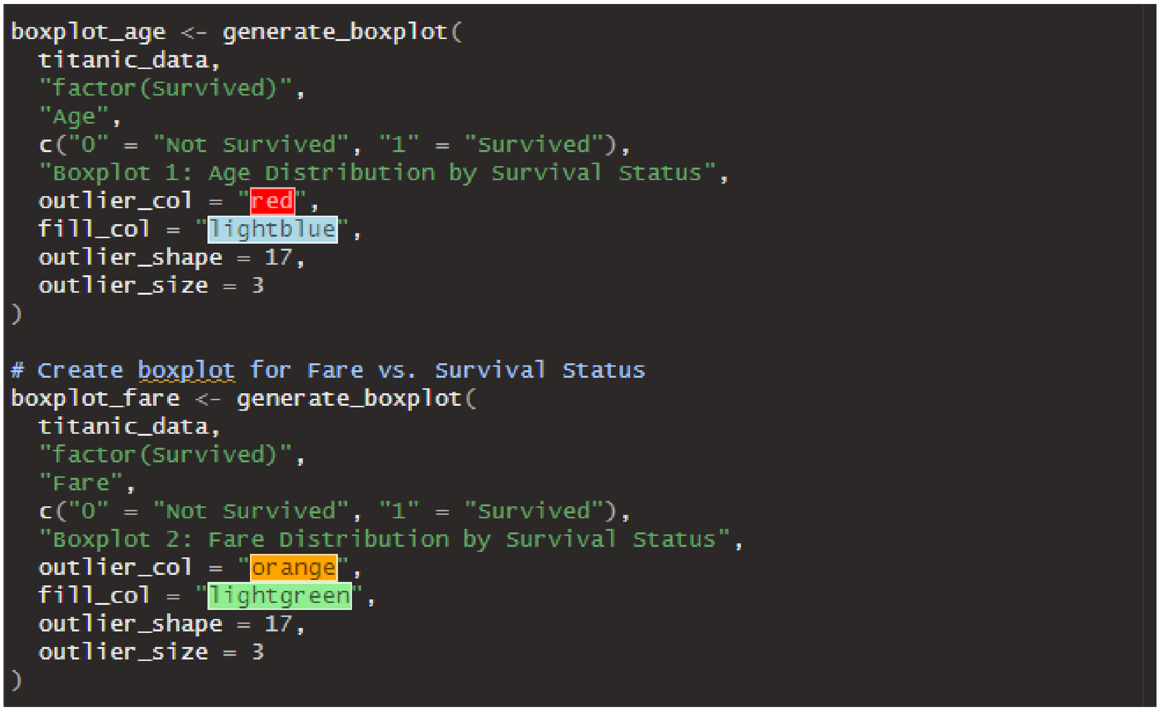
This block generates two boxplots:

- boxplot_age: A boxplot for Age vs. Survived.
- boxplot_fare: A boxplot for Fare vs. Survived.

Customization:

- factor(Survived): Ensures the Survived variable is treated as a factor (categorical variable).
- The custom labels (0 = "Not Survived", 1 = "Survived") and title are added for clarity.
- The outliers are highlighted in specific colors and shapes for easy identification.

The boxplots are displayed using boxplot_age and boxplot_fare. The output is visualized in RStudio's plot viewer.



```
boxplot_age <- generate_boxplot(
  titanic_data,
  "factor(Survived)",
  "Age",
  c("0" = "Not Survived", "1" = "Survived"),
  "Boxplot 1: Age Distribution by Survival Status",
  outlier_col = "red",
  fill_col = "lightblue",
  outlier_shape = 17,
  outlier_size = 3
)

# Create boxplot for Fare vs. Survival Status
boxplot_fare <- generate_boxplot(
  titanic_data,
  "factor(Survived)",
  "Fare",
  c("0" = "Not Survived", "1" = "Survived"),
  "Boxplot 2: Fare Distribution by Survival Status",
  outlier_col = "orange",
  fill_col = "lightgreen",
  outlier_shape = 17,
  outlier_size = 3
)
```

End of Code

- ## Interpret charts and data

Table 1: Summary of Titanic Data by Survival Status

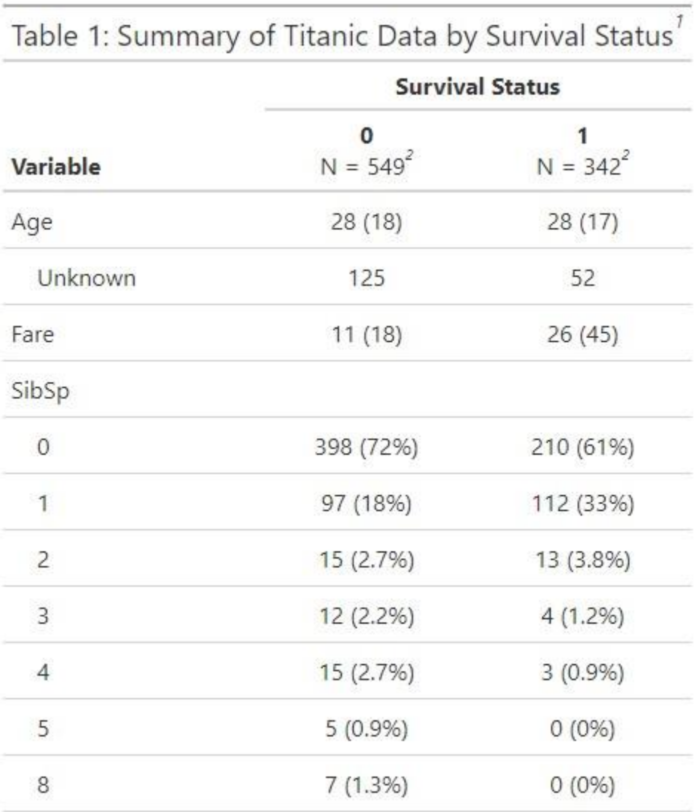
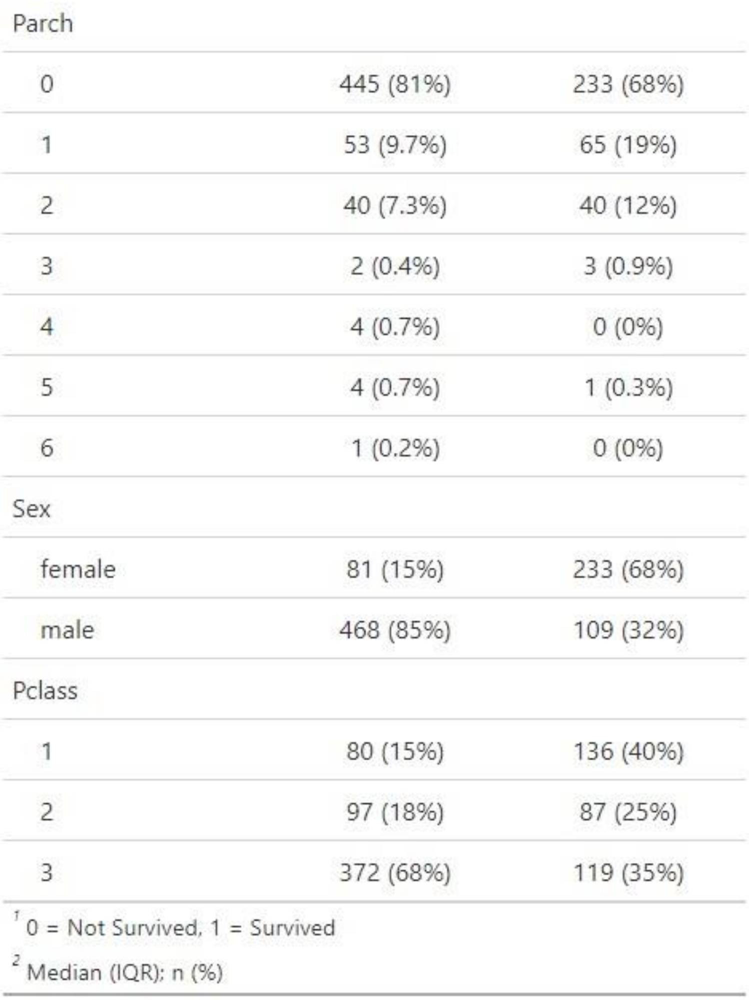


Table 1: Summary of Titanic Data by Survival Status[1]

	Survival Status	
Variable	**0** N = 549[2]	**1** N = 342[2]
Age	28 (18)	28 (17)
Unknown	125	52
Fare	11 (18)	26 (45)
SibSp		
0	398 (72%)	210 (61%)
1	97 (18%)	112 (33%)
2	15 (2.7%)	13 (3.8%)
3	12 (2.2%)	4 (1.2%)
4	15 (2.7%)	3 (0.9%)
5	5 (0.9%)	0 (0%)
8	7 (1.3%)	0 (0%)



Parch		
0	445 (81%)	233 (68%)
1	53 (9.7%)	65 (19%)
2	40 (7.3%)	40 (12%)
3	2 (0.4%)	3 (0.9%)
4	4 (0.7%)	0 (0%)
5	4 (0.7%)	1 (0.3%)
6	1 (0.2%)	0 (0%)
Sex		
female	81 (15%)	233 (68%)
male	468 (85%)	109 (32%)
Pclass		
1	80 (15%)	136 (40%)
2	97 (18%)	87 (25%)
3	372 (68%)	119 (35%)

[1] 0 = Not Survived, 1 = Survived

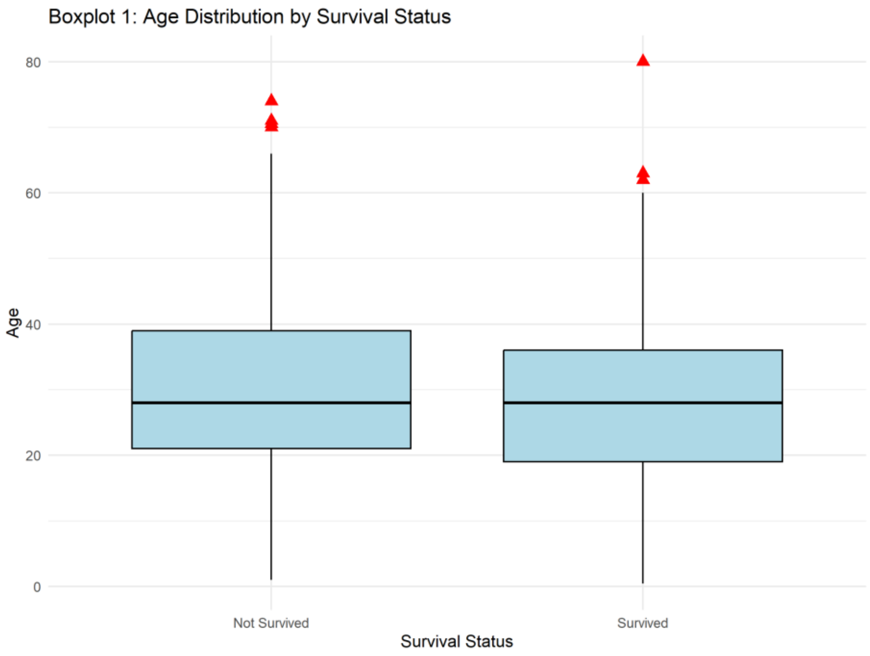
[2] Median (IQR); n (%)

- **Age**:
  - The median age of both survivors and non-survivors was around 28 years, with a slightly lower interquartile range (IQR) for survivors (28 years (17) compared to 28 (18) for non-survivors).
  - A significant portion of data on **Age** was missing: 125 entries (non-survivors) and 52 entries (survivors).
- **Fare**:
  - Survivors had paid higher median fares (26 with IQR 45), compared to non-survivors who paid a median fare of 11 (IQR 18), suggesting that passengers who paid more had a better chance of survival (likely due to better access to lifeboats in higher classes).
- **SibSp**:
  - A majority of passengers had no siblings/spouses aboard: 72% (398) among non-survivors and 61% (210) among survivors.

    - Passengers with 1 sibling/spouse had a better survival rate, with 33% surviving.
    - As the number of siblings/spouses increased, the survival rate generally decreased.
- **Parch**:
    - Most passengers had 0 parents/children onboard: 445 (81%) among those who did not survive and 233 (68%) among survivors.
    - Small percentages of passengers had 1–6 family members onboard, with the survival rate decreasing as the number of family members increased beyond 1.
- **Sex**:
    - **Females** had a much higher survival rate: 68% (233 survived) compared to 15% (81 did not survive).
    - **Males**, on the other hand, had a much lower survival rate: 32% (109 survived) compared to 85% (468 did not survive).
- **Pclass**:
    - Passengers in **1st class** had the highest survival rate: 40% (136 survived), while only 15% (80 did not survive).
    - The majority of those who didn't survive were in **3rd class**: 68% (372), while only 35% (119) in this class survived.

Boxplot 1: Age Distribution by Survival Status

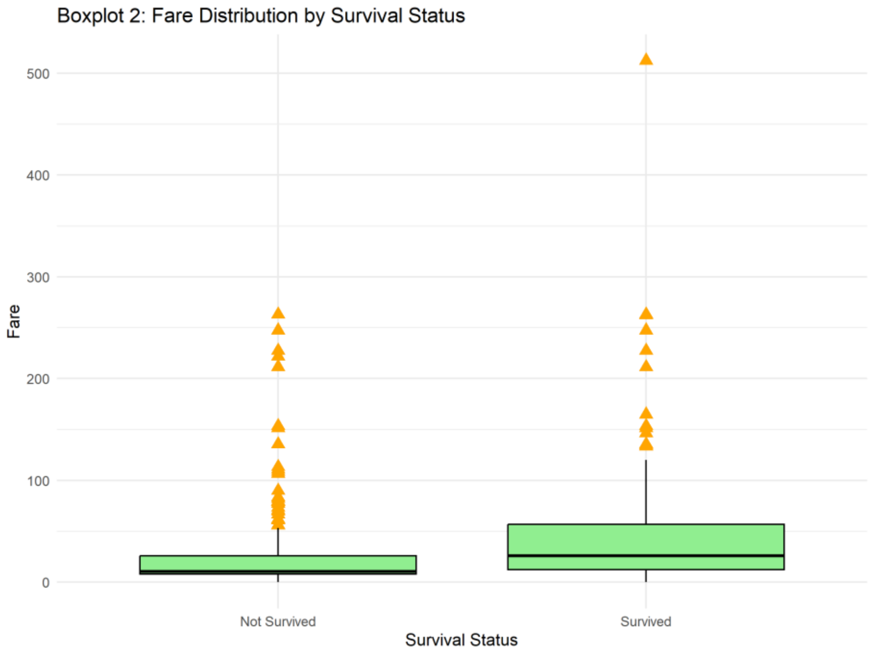
The first boxplot shows the distribution of age for passengers who survived and those who did not survive. From the plot, we observe the following:
1. Both survived and non-survived groups have similar median ages, which are centered around the mid-30s.
2. The interquartile ranges (IQR) for both groups appear to be similar, suggesting comparable age distribution between them.
3. There are outliers in both groups, indicated by the red triangles. For the 'Not Survived' group, the upper outliers extend above 80 years. In the 'Survived' group, the outliers also extend above 80 years.
4. The overall distribution is slightly right-skewed, with a wider spread on the upper end.



Boxplot 2: Fare Distribution by Survival Status

The second boxplot presents the distribution of fare paid by passengers who survived and those who did not survive. Key insights include:

1. The median fare for passengers who survived is significantly higher than that of passengers who did not survive.
2. The fare distribution for those who survived has a much larger interquartile range, indicating a higher fare variability among survivors.
3. There are many outliers in both groups, as shown by the orange triangles. These outliers represent passengers who paid substantially higher fares.
4. The majority of 'Not Survived' passengers paid relatively lower fares compared to the survivors.


Boxplot 2: Fare Distribution by Survival Status
Fare
Not Survived
Survived
Survival Status