# STATISTICAL TECHNIQUES USING R

**Subject Code:**24CAP-612

PROJECT REPORT

ON

# Train(Titanic) Data set analysis


***Submitted by: -***

***Aayush Khatiyan (24MCI10109)***

***Yash Aggarwal(24MCI10107)***

***Heena Sharma(24MCI10111)***

***Submitted to: -***

***Ms. Mausam kumari***


***in partially fulfillment for the award of the degree of***

***Master of Computer Application***

***in***

***Artificial Intelligence & Machine Learning***


CHANDIGARH UNIVERSITY

NOVEMBER 2024

## Acknowledgement

We want to express our sincere gratitude to our project supervisor, **Ms. Mausam Kumari**, from the Department of University Institute of Computing, for her invaluablee guidance and support throughout the development of this project. Their expertise and insights were instrumental in shaping the direction and quality of our work. Together, we tackled the challenges of that comes during this project, learning from one another and pushing each other to achieve our best. Our combined efforts in brainstorming, coding, and testing have greatly enriched the project.

Additionally, we would like to acknowledge the resources available through the Department of UIC, including access to libraries and software that facilitated us development process. The encouragement and support from faculty and staff have also been crucial in helping us navigate challenges throughout the project.

GitHub link:-https://github.com/Yash17aggarwal/Rproject/upload

# Introduction: -

The dataset Train.csv is taken form Kaggle repository, the tain dataset is consist the data of titanic ship in which several information such as name, age, gender, fare and many more things on which we have to perform the function and plot the different graphs for data analytics.

# Data set: -

- Firstly, we have to load the dataset into the R-studio. For that follow the code:-
  As our file is in .csv, here we use read.csv to read and load the file.

```
20  # Load the dataset
21  titanic_data <- read.csv("Desktop/train.csv")
22
```

- After loading the dataset, took a glance on the data set by the following commands.

a) Head(): - It's gives the first 6 row of the dataset.
b) Tail(): - It's gives the last 6 row of the dataset.
c) Summary(): - It basically provides a detailed statistical summary of each variable in a dataset.
d) Glimpse(): - It basically shows a transposed snapshot of the data structure and the first few values.
e) ncol(): - It tells how many column present in the dataset.
f) nrow(): - It tells how many row present in the dataset.

```
> head(titanic_data)
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch           Ticket    Fare Cabin Embarked
1           1        0      3                              Braund, Mr. Owen Harris   male  22     1     0        A/5 21171  7.2500              S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0         PC 17599 71.2833   C85        C
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282  7.9250              S
4           4        1      1       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0           113803 53.1000  C123        S
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0           373450  8.0500              S
6           6        0      3                                     Moran, Mr. James   male  28     0     0           330877  8.4583              Q
>
```

```
> tail(titanic_data)
    PassengerId Survived Pclass                                      Name    Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
886         886        0      3     Rice, Mrs. William (Margaret Norton) female  39     0     5     382652 29.125              Q
887         887        0      2                   Montvila, Rev. Juozas   male  27     0     0     211536 13.000              S
888         888        1      1            Graham, Miss. Margaret Edith female  19     0     0     112053 30.000   B42        S
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  28     1     2 W./C. 6607 23.450              S
890         890        1      1                    Behr, Mr. Karl Howell   male  26     0     0     111369 30.000  C148        C
891         891        0      3                      Dooley, Mr. Patrick   male  32     0     0     370376  7.750              Q
>
```

```
> glimpse(titanic_data)
Rows: 891
Columns: 12
$ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36,…
$ Survived    <fct> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1,…
$ Pclass      <fct> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3, 2, 2, 3, 1, 3, 3, 3, 1, 3, 3, 1, 1, 3, 2, 1, 1, 3, 3, 3, 3, 3, 2, 3, 2, 3,…
$ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Florence Briggs Thayer)", "Heikkinen, Miss. Laina", "Futrelle, Mrs. Jacques He…
$ Sex         <fct> male, female, female, female, male, male, male, male, female, female, female, female, male, male, female, female, male, male, female, …
$ Age         <dbl> 22, 38, 26, 35, 35, 28, 54, 2, 27, 14, 4, 58, 20, 39, 14, 55, 2, 28, 31, 28, 35, 34, 15, 28, 8, 38, 28, 19, 28, 28, 40, 28, 28, 66, 28…
$ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0, 0, 0, 0, 0, 3, 1, 0, 3, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 2, 1, 1, 1, 0, 1, 0,…
$ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 5, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0,…
$ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "373450", "330877", "17463", "349909", "347742", "237736", "PP 9549", "113783",…
$ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625, 21.0750, 11.1333, 30.0708, 16.7000, 26.5500, 8.0500, 31.2750, 7.8542, 16.00…
$ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", "", "", "", "G6", "C103", "", "", "", "", "", "", "", "", "D56", "", "A6", "", "", "", "C23 …
$ Embarked    <fct> S, C, S, S, S, Q, S, S, S, C, S, S, S, S, S, S, Q, S, S, C, S, S, Q, S, S, S, C, S, Q, S, C, C, Q, S, C, S, C, S, S, C, S, S, C, C, Q,…
>
```

```
> ncol(titanic_data)
[1] 12
```

```
> nrow(titanic_data)
[1] 891
```

```
> summary(titanic_data)
  PassengerId     Survived Pclass      Name              Sex           Age            SibSp           Parch           Ticket              Fare
 Min.   :  1.0   0:549    1:216   Length:891         female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00
 1st Qu.:223.5   1:342    2:184   Class :character   male  :577   1st Qu.:22.00   1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91
 Median :446.0            3:491   Mode  :character              Median :28.00   Median :0.000   Median :0.0000   Mode  :character   Median : 14.45
 Mean   :446.0                                                  Mean   :29.36   Mean   :0.523   Mean   :0.3816                      Mean   : 32.20
 3rd Qu.:668.5                                                  3rd Qu.:35.00   3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00
 Max.   :891.0                                                  Max.   :80.00   Max.   :8.000   Max.   :6.0000                      Max.   :512.33
    Cabin             Embarked
 Length:891          :  2
 Class :character    C:168
 Mode  :character    Q: 77
                     S:644
```

Till here we get the overview of the data. Now check if contain any null value: -

- ✧ To check missing value.

```
> # Check for missing values
> colSums(is.na(titanic_data))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0           0           0           0           0           0           0           0           0
>
```

- ✧ To fill the missing values and missing Embarked values.

```
> # Fill missing Age values with the median
> titanic_data$Age[is.na(titanic_data$Age)] <- median(titanic_data$Age, na.rm = TRUE)
>
```

```
> # Fill missing Embarked values with the most frequent value ('S')
> titanic_data$Embarked[is.na(titanic_data$Embarked)] <- "S"
>
```

- ✧ To convert the categorical variable to factors.

```
> # Convert categorical variables to factors
> titanic_data$Survived <- as.factor(titanic_data$Survived)
> titanic_data$Pclass <- as.factor(titanic_data$Pclass)
> titanic_data$Sex <- as.factor(titanic_data$Sex)
> titanic_data$Embarked <- as.factor(titanic_data$Embarked)
>
```

Till here our analysis of data part and filling the null/missing values are done.

Let's move to our next part---- that is data visualization.
Before moving to visualization load the necessary libraries. And check if they installed by using library() function.

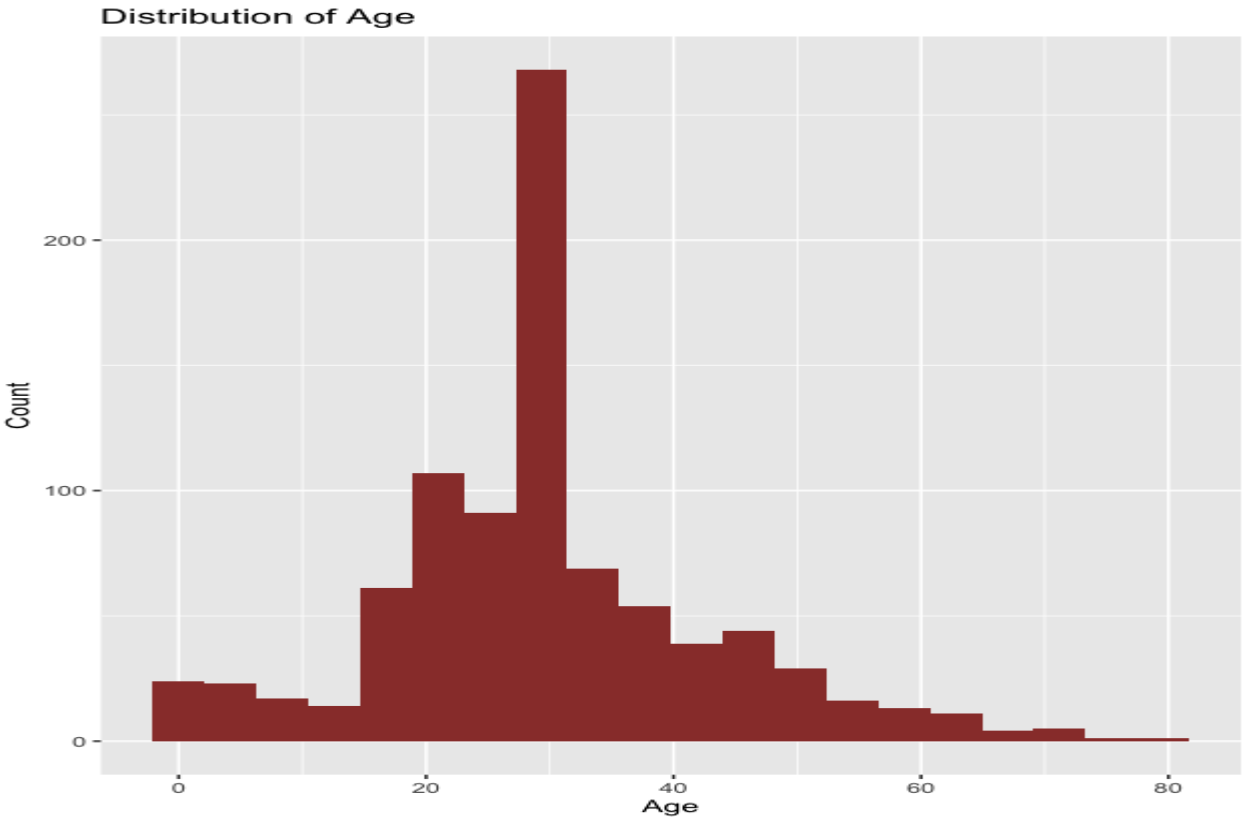
```
# Install necessary packages (if not installed)
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
```

```
# Load the libraries
library(ggplot2)
library(dplyr)
library(tidyr)
```

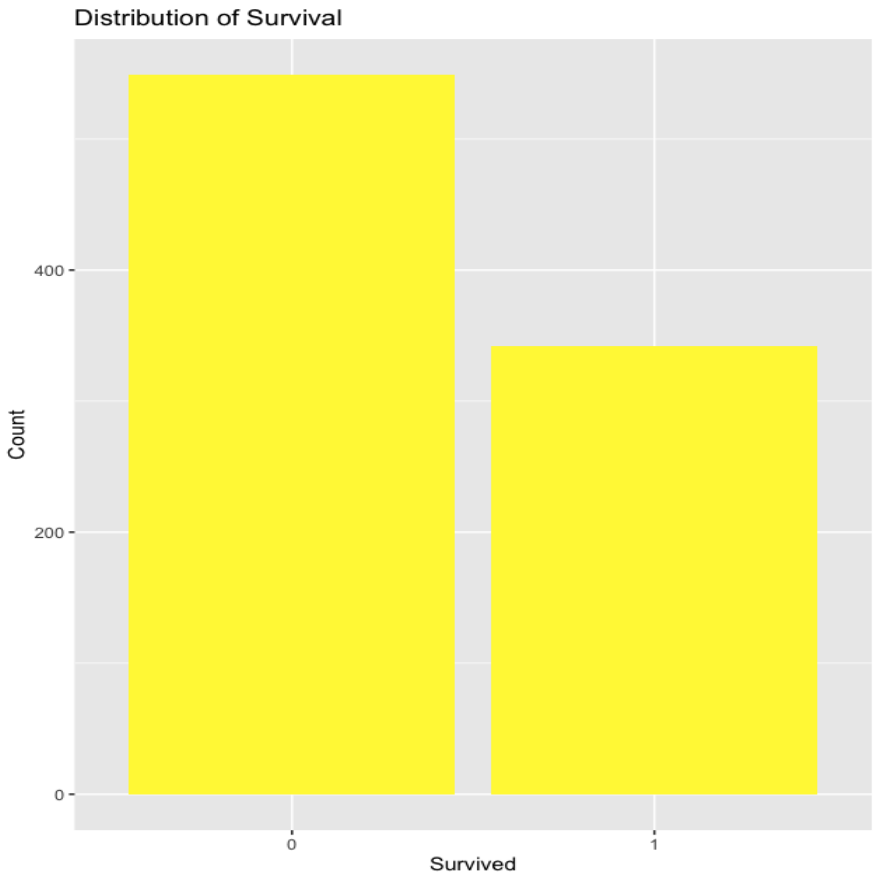
The first graph we used is Histogram, which is used to distribute the frequency of the data.

```
# Histogram of Age distribution
ggplot(titanic_data, aes(x = Age)) +
  geom_histogram(fill = "brown", bins = 20) +
  labs(title = "Distribution of Age", x = "Age", y = "Count")
```

Output: -



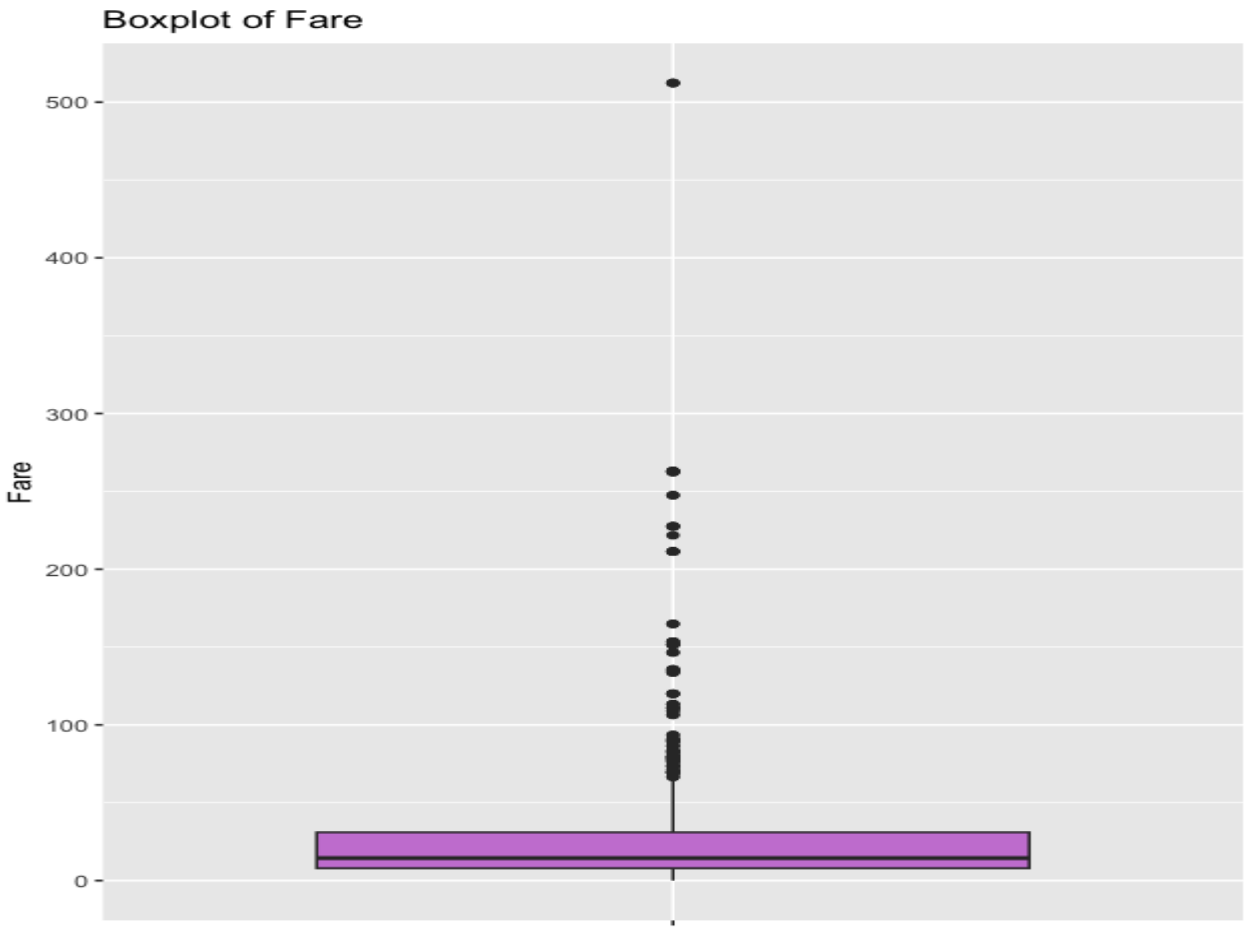

```r
# Plot the distribution of survival (Survived column)
ggplot(titanic_data, aes(x = Survived)) +
  geom_bar(fill = "yellow") +
  labs(title = "Distribution of Survival", x = "Survived", y = "Count")
```




In dataset their are some outliers that are present in dataset. To detect them we use the boxplot.

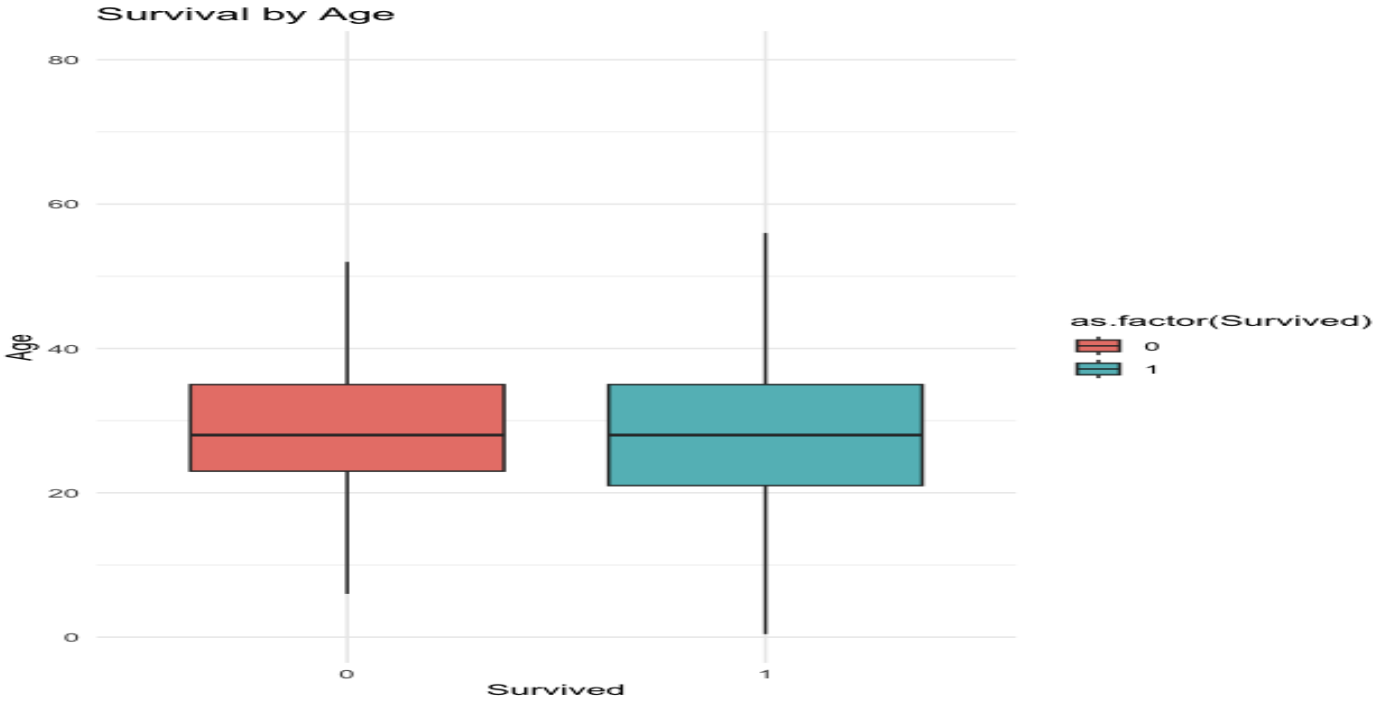
The boxplot basically contain 5 things Minimum, 1 quartile (Q1), Median, 3 quartile(Q3), Maximum.
The range of Q1 and Q3 should be below 1.5, if the range is increases from 1.5 it is consider as outliers.

```
73  # Boxplot for Age
74  ggplot(titanic_data, aes(x = "", y = Age)) +
75    geom_boxplot(fill = "coral") +
76    labs(title = "Boxplot of Age", x = "", y = "Age")
77
```



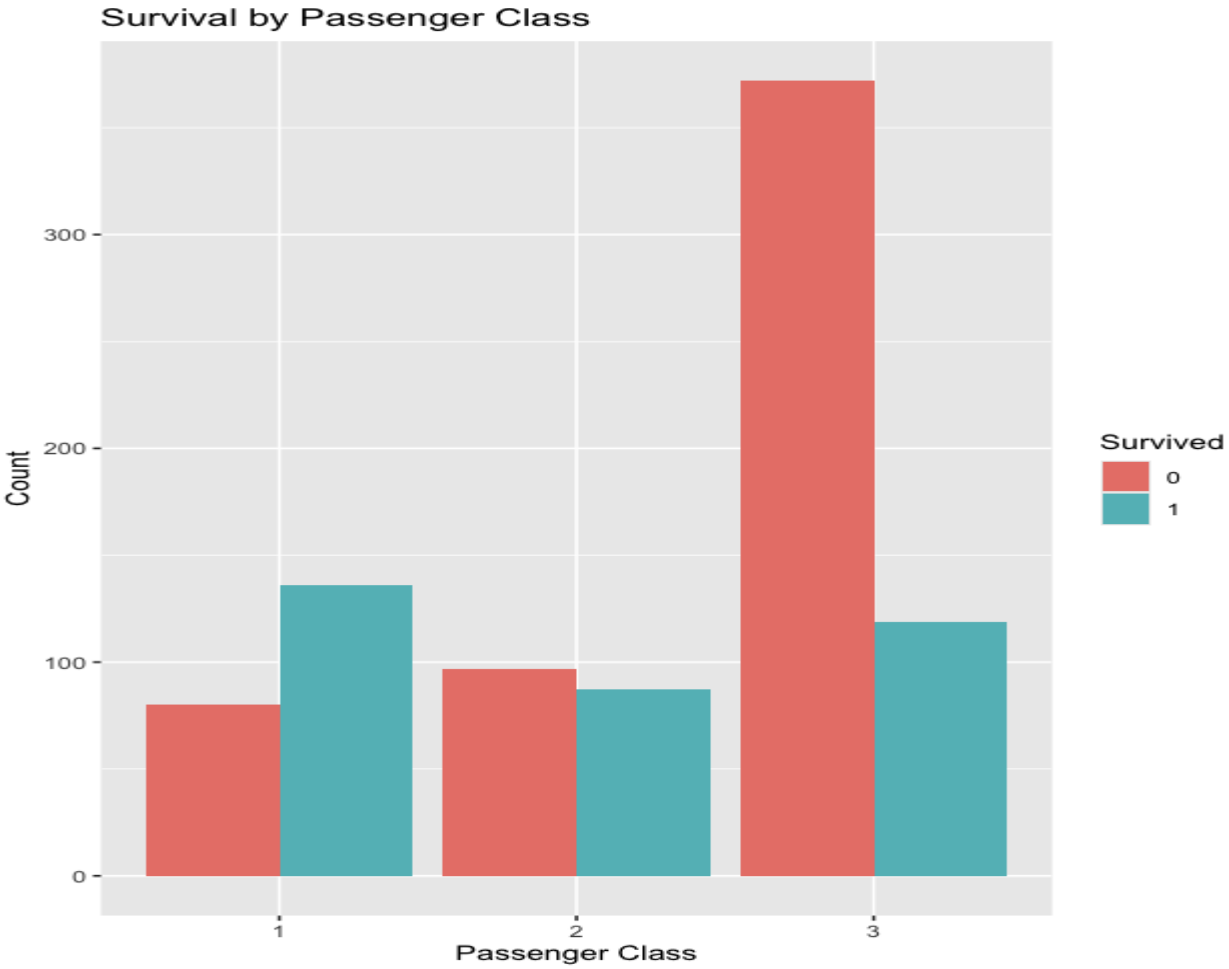

Here, in the figure dots represent the outliers. But it can be ignored in graph by giving the outlier.shape=NA in the code. In next figure we will look on it.

```
90  # Boxplot of Survival vs. Age
91  ggplot(titanic_data, aes(x = as.factor(Survived), y = Age, fill = as.factor(Survived))) +
92    geom_boxplot(outlier.shape = NA) +   # This ignores the outliers
93    labs(title = "Survival by Age", x = "Survived", y = "Age") +
94    theme_minimal()
```




Let's took a glance on bar chart which is basically shows the distribution of the categorical data.

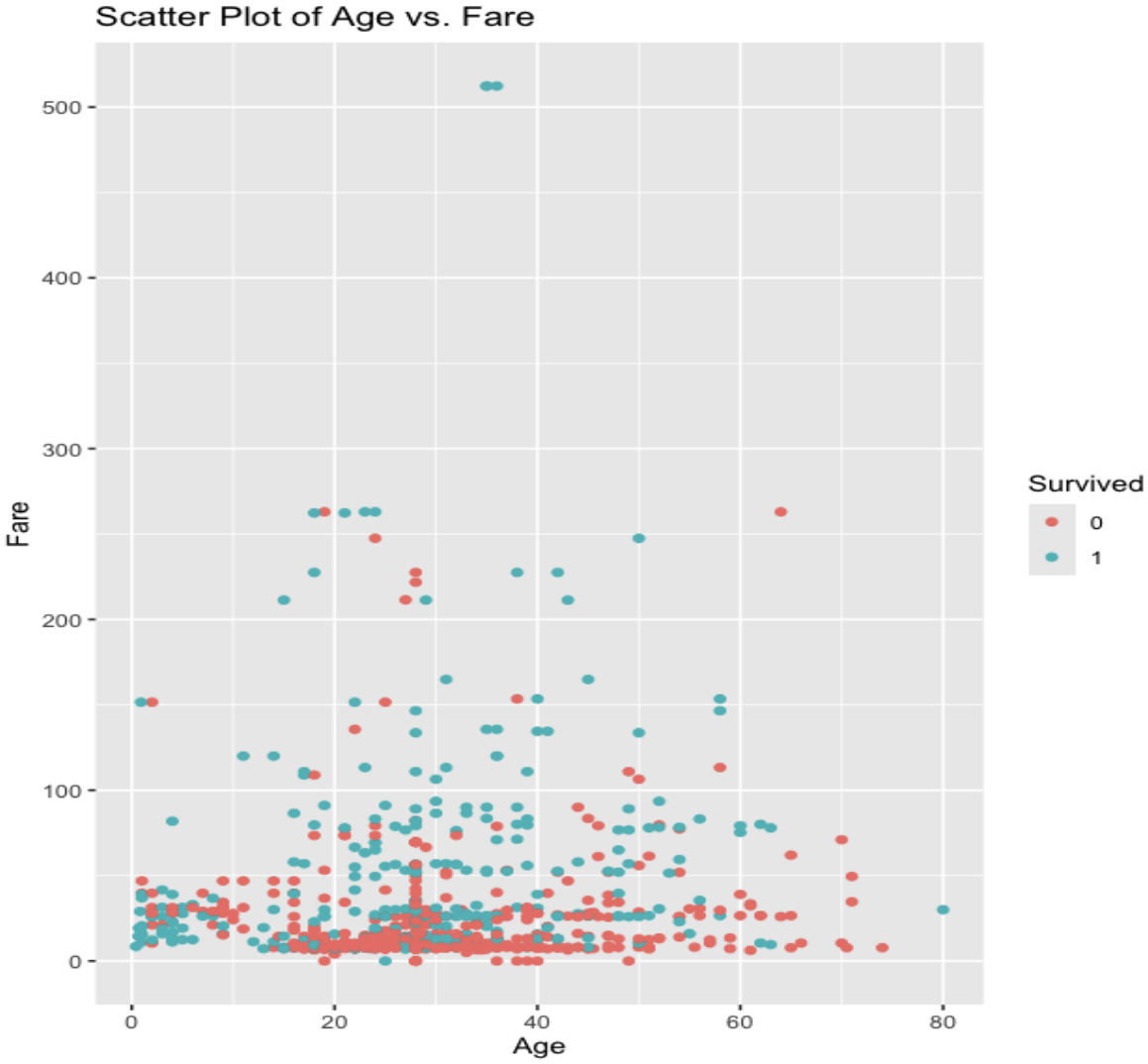
```
# Bar chart for Survival by Pclass
ggplot(titanic_data, aes(x = Pclass, fill = Survived)) +
  geom_bar(position = "dodge") +
  labs(title = "Survival by Passenger Class", x = "Passenger Class", y = "Count")

```




After bar chart, move to scatter chart which is an important chart for displaying the multiple values.

Scatter plot: - uses dots to represent values for two different numeric variables

```
109  # Scatter plot of Age vs. Fare, colored by Survival
110  ggplot(titanic_data, aes(x = Age, y = Fare, color = Survived)) +
111    geom_point() +
112    labs(title = "Scatter Plot of Age vs. Fare", x = "Age", y = "Fare")
113
```


Scatter Plot of Age vs. Fare
Fare
500
400
300
200
100
0
0
20
40
60
80
Age
Survived
0
1

**The implementation of the code**

```
115  # Load necessary libraries
116  library(shiny)
```

GUI: - In the GUI section we use a R package called Shiny for better understanding and visualization. Below down there is a brief explanation of Shiny package.

✧ **Shiny** is a free, open-source R package that allows users to create interactive web applications

Components of the Shiny Application:

1. User Interface (UI)
   - The UI of the app includes several input fields where users can enter values for the independent variables (e.g., BMI, glucose level, physical activity status).
   - A "Predict Age" button triggers the model to predict the user's age based on the inputs.
2. Server Logic
   - When the "Predict Age" button is clicked, the server-side logic takes the user inputs, feeds them into the trained linear regression model, and displays the predicted age.
   - The residual plots are also displayed in real-time for additional diagnostic analysis.
3. Interactivity
   - Users can experiment with different inputs to see how changes in health metrics affect the predicted age.

Code snippet: -

```r
# Define UI
ui <- fluidPage(
  titlePanel("Titanic Dataset Exploratory Data Analysis"),

  sidebarLayout(
    sidebarPanel(
      selectInput("plotType", "Choose Plot Type:",
                  choices = c("Histogram", "Boxplot", "Scatter Plot")),

      # Histogram options
      conditionalPanel(
        condition = "input.plotType == 'Histogram'",
        selectInput("histVar", "Choose Variable for Histogram:",
                    choices = c("Age", "Fare", "SibSp", "Parch"))
      ),

      # Boxplot options
      conditionalPanel(
        condition = "input.plotType == 'Boxplot'",
        selectInput("boxVar", "Choose Variable for Boxplot:",
                    choices = c("Age", "Fare", "SibSp", "Parch"))
      ),

      # Scatter plot options
      conditionalPanel(
        condition = "input.plotType == 'Scatter Plot'",
        selectInput("scatterX", "Choose X Variable for Scatter Plot:",
                    choices = c("Age", "Fare", "SibSp", "Parch")),
        selectInput("scatterY", "Choose Y Variable for Scatter Plot:",
                    choices = c("Age", "Fare", "SibSp", "Parch"))
      ),

      actionButton("update", "Update Plot")
    ),

    mainPanel(
      plotOutput("mainPlot")
    )
  )
)
```

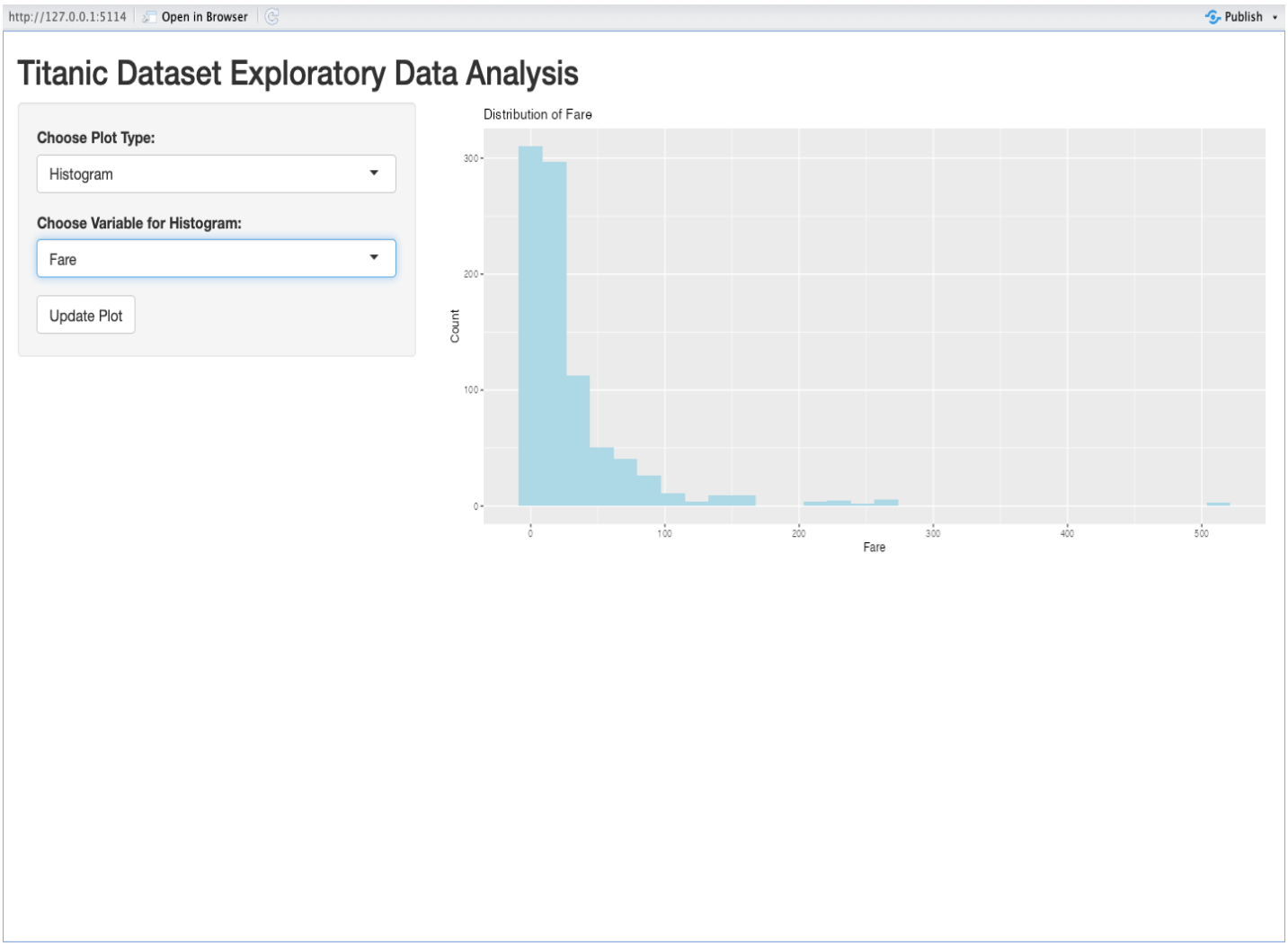
272:35 (Top Level)

Console

```
241
242 # Define server logic
243 server <- function(input, output) {
244
245   # Render main plot based on selected type
246   output$mainPlot <- renderPlot({
247     req(input$update)  # Ensure the plot updates only when the button is clicked
248
249     if (input$plotType == "Histogram") {
250       ggplot(titanic_data, aes_string(x = input$histVar)) +
251         geom_histogram(fill = "lightblue", bins = 30) +
252         labs(title = paste("Distribution of", input$histVar),
253              x = input$histVar, y = "Count")
254
255     } else if (input$plotType == "Boxplot") {
256       # Create a dummy variable for x-axis
257       ggplot(titanic_data, aes_string(x = "1", y = input$boxVar)) +
258         geom_boxplot(fill = "coral") +
259         labs(title = paste("Boxplot of", input$boxVar),
260              x = "", y = input$boxVar)
261
262     } else if (input$plotType == "Scatter Plot") {
263       ggplot(titanic_data, aes_string(x = input$scatterX, y = input$scatterY)) +
264         geom_point(color = "blue") +
265         labs(title = paste("Scatter Plot of", input$scatterY, "vs", input$scatterX),
266              x = input$scatterX, y = input$scatterY)
267     }
268   })
269 }
270
271 # Run the application
272 shinyApp(ui = ui, server = server)
```

```
Listening on http://127.0.0.1:5114
```

GUI of the project:-



http://127.0.0.1:5114 | Open in Browser | Publish

# Titanic Dataset Exploratory Data Analysis

**Choose Plot Type:**

Scatter Plot

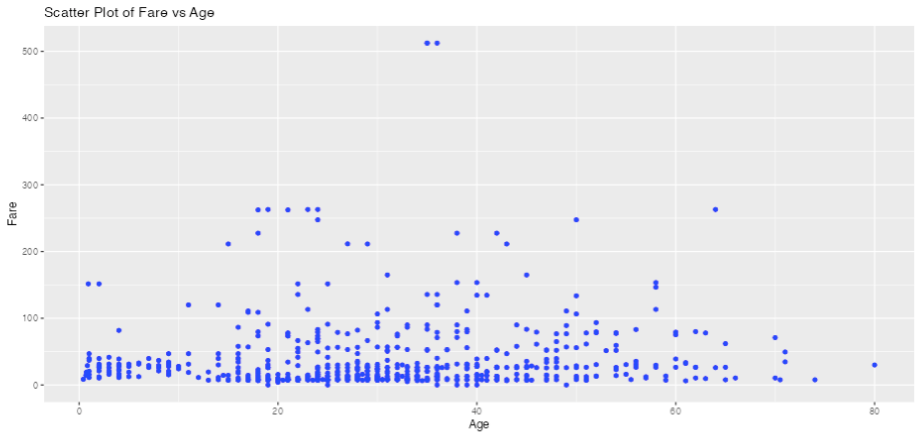
**Choose X Variable for Scatter Plot:**

Age

**Choose Y Variable for Scatter Plot:**

Fare

Update Plot




http://127.0.0.1:5114 | Open in Browser | Publish

# Titanic Dataset Exploratory Data Analysis

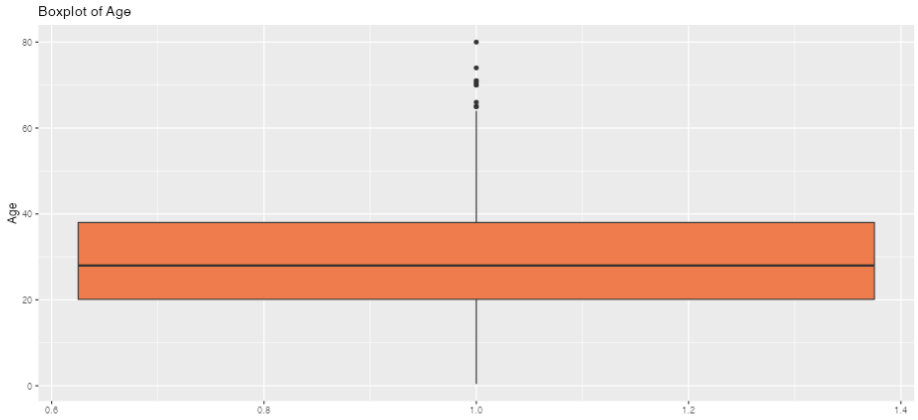
**Choose Plot Type:**

Boxplot

**Choose Variable for Boxplot:**

Age

Update Plot



## Conclusion: -

The Exploratory Data Analysis project on the titanic dataset provide a valuable insights during the execution of this project, helps to categories b/w different data as performed during the project. Apart from this it gives various challenges regarding cleaning and plotting the various graphs.

**Key points: -**

1. **Data distribution: -** During the performance of this project we face the distribution of different dataset such as fare vs age or age vs parch.
2. **Data cleaning: -** In this dataset, it contain some value which is not suitable with the format or didn't have the value. To fill null value it uses a method(mean,mode,medain) to fill these values.
3. **Outlier detection: -** In titanic dataset, there are few data that contain outlier value(value > 1.5). For the detection of these outlier done with the help of boxplot.