UNIVERSITY INSTITUTE of COMPUTING
Asia's Fastest Growing University

NAAC GRADE A+
Accredited University

Titanic Dataset Analysis in R: Visualizing Data Distributions

Student Name: Prince Kumar **UID:** 24MCI10052

Branch: MCA(AIML) **Section/Group:** 1B

Semester: 1st **Date of Performance:** 25/10/2024

Subject Name: R Programming **Subject Code:** 24CAP-614

1. Aim/Overview of the Project:

Choose a dataset from a repository like Kaggle or UCI Machine Learning Repository and perform exploratory data analysis using R. Explore the distribution of variables, identify outliers, and visualize relationships between variables using plots like histograms, scatter plots, and boxplots.

2. Task to be done:

- **Histogram Summary**

Purpose: Shows the frequency distribution of the `Age` variable among passengers.

Key Components:

- Bins: Groups of ages (e.g., 0-5, 5-10, etc.) represented on the x-axis.
- Frequency: The count of passengers within each age group on the y-axis.
- Distribution Shape: Helps identify the central tendency, spread, and any potential skewness in the age distribution.

- **Boxplot Summary**

Purpose: Visualizes the distribution of the `Fare` variable across different passenger classes (Pclass).

Key Components:

- Minimum: Lowest fare in each class.

- First Quartile (Q1): 25% of fares fall below this value.
- Median (Q2): The middle value of fares, indicating the central tendency.
- Third Quartile (Q3): 75% of fares fall below this value.
- Maximum: Highest fare in each class.
- Outliers: Points outside the whiskers, indicated in red.

- **Scatter Plot Summary**

Purpose: Displays the relationship between `Age` and `Fare`, colored by survival status (`Survived`).
Key Components:

- X-Axis: Represents passenger `Age`.
- Y-Axis: Represents passenger `Fare`.
- Color Coding: Points colored based on survival status (e.g., survived or not), facilitating visual analysis of survival trends related to age and fare.
- Trends: Helps identify potential correlations or patterns between age and fare, such as whether younger or older passengers paid more.

Visual Insights

- Boxplot: Indicates fare distribution differences between classes, highlighting outliers.
- Histogram: Reveals age distribution patterns, suggesting demographic trends among passengers.
- Scatter Plot: Provides insights into how age relates to fare and survival, allowing for further analysis of potential factors influencing survival rates.

These visualizations together give a comprehensive overview of the dataset's key characteristics and relationships. If you need more detailed explanations or additional analysis, feel free to ask!
Write all the parameters needed for the boxplot.

- **Visualization of titanic data**

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

3. Steps/Commands involved to perform Project:

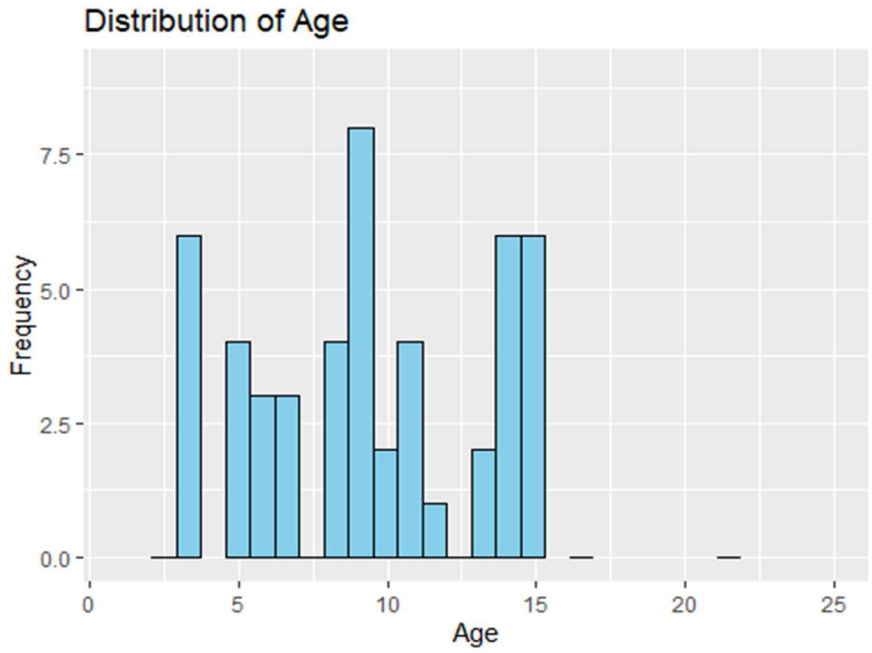
- **Install Required Libraries (if not already installed)**:

```
install.packages("ggplot2")  # For visualization
install.packages("dplyr")    # For data manipulation
install.packages("corrplot")  # For correlation matrix visualization
install.packages("reshape2")

library(ggplot2)
library(dplyr)
library(readr)# To read CSV files
library(reshape2)

# Load the Titanic dataset
titanic <- read.csv("C:/Users/Prince/Downloads/train.csv")

#  View the first few rows of the dataset
head(titanic)

#  Summary Statistics of the numerical columns
print("Summary Statistics:")
summary(titanic)
```

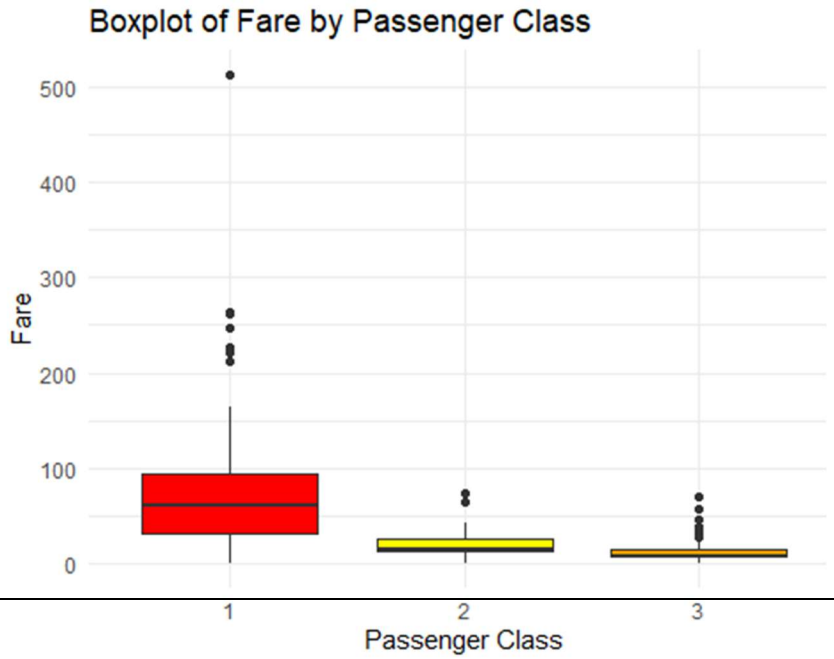
```
#  Check for missing values
print("Missing Values:")
print(colSums(is.na(titanic)))

#  Histogram: Distribution of Age

ggplot(titanic, aes(x = Age)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  ggtitle("Distribution of Age") +
  xlab("Age") +
  ylab("Frequency") +
  xlim(c(1, 25)) +
  ylim(c(0, 9))


# . Boxplot: Fare by Passenger Class (Pclass)

ggplot(titanic, aes(x = factor(Pclass), y = Fare)) +
  geom_boxplot(fill = c("red","yellow","orange")) +
  ggtitle("Boxplot of Fare by Passenger Class") +
  xlab("Passenger Class") +
  ylab("Fare")+
  theme_minimal()

#  Scatter Plot: Age vs Fare, colored by Survival status

ggplot(titanic, aes(x = Age, y = Fare, color = factor(Survived))) +
  geom_point(size = 3) +
  ggtitle("Scatter Plot: Age vs Fare by Survival Status") +
  xlab("Age") +
  ylab("Fare") +
  scale_color_manual(values = c("yellow", "blue"),
                     labels = c("Not Survived", "Survived"))

#  Correlation Matrix: Only numerical columns

numeric_data <- titanic %>%
  select(Age, Fare, SibSp, Parch)  # Select numerical columns
correlation_matrix <- cor(numeric_data, use = "complete.obs")
print("Correlation Matrix:")
print(correlation_matrix)

#  Heatmap: Correlation Matrix
```

```
melted_corr <- melt(correlation_matrix)

ggplot(data = melted_corr, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  ggtitle("Correlation Matrix Heatmap") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```

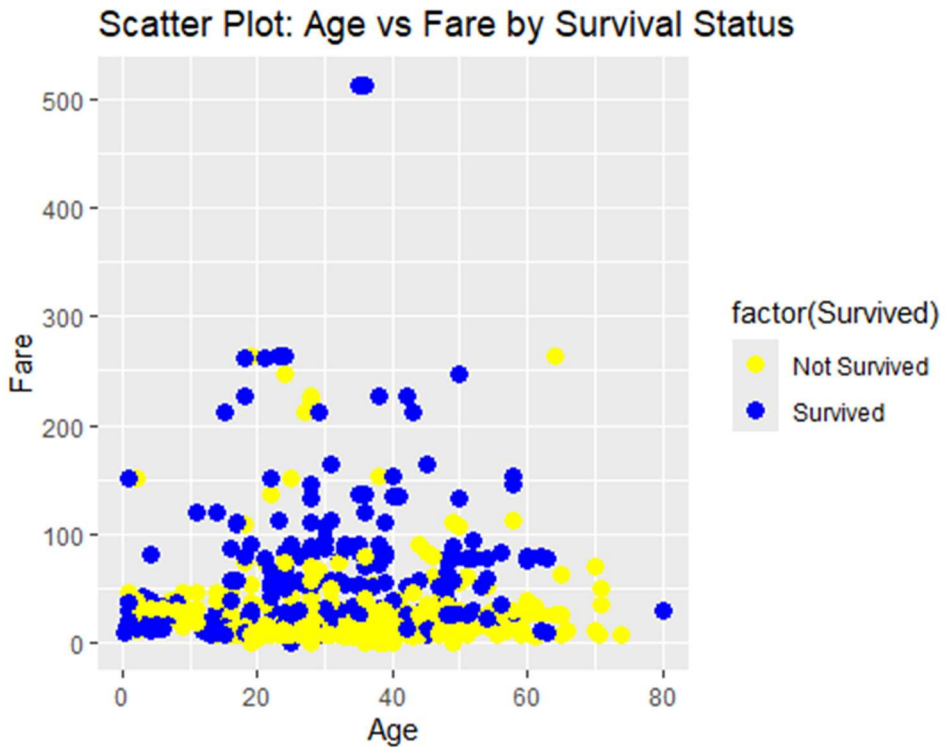
Output:-



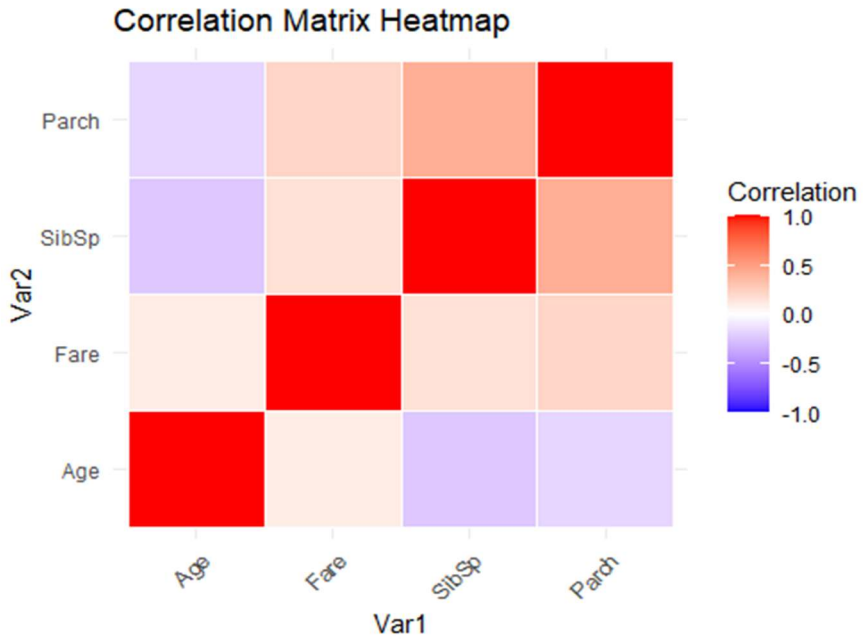
Boxplot: Sepal Length by Species



Create Scatter Plot: Sepal Length vs Sepal Width



Heatmap: Correlation Matrix



Learning Outcomes

1. Data Loading & Inspection: Learn to load CSV files and inspect datasets using functions like `head()` and `summary()`.

2. Data Cleaning: Identify and quantify missing values, understanding their impact on analysis.

3. Distribution Visualization: Create histograms to visualize distributions (e.g., Age) and identify trends.

4. Outlier Detection: Use boxplots to detect outliers in numerical data (e.g., Fare by Passenger Class).

5. Relationship Exploration: Generate scatter plots to explore relationships between variables (e.g., Age vs. Fare), using color for categorical differentiation (Survival status).

6. Correlation Analysis: Calculate and visualize correlation matrices to understand relationships among numerical variables.

7. Effective Communication: Enhance skills in using visualizations to communicate findings clearly and effectively**.**

8. Statistical Understanding: Develop an understanding of correlation coefficients and their interpretation.