**NAME: SIDDHI SUBHEDAR**

**PRN: 1262240621**

**R LAB 6- CREATE A DATASET AND DO A STATISTICAL ANALYSIS ON THE DATA USING R**

install.packages("readr")

library(readr)

myData <- read\_csv("C:/Users/SoECE/Downloads/titanicsurvival (1).csv")

head(myData)

Pclass Sex Age Fare Survived

*<dbl>* *<chr>* *<dbl>* *<dbl>* *<dbl>*

1 3 male 22 7.25 0

2 1 female 38 71.3 1

3 3 female 26 7.92 1

4 1 female 35 53.1 1

5 3 male 35 8.05 0

6 3 male NA 8.46 0

summary(myData)

> summary(myData)

Pclass Sex Age Fare Survived

Min. :1.000 Length:891 Min. : 0.42 Min. : 0.00 Min. :0.0000

1st Qu.:2.000 Class :character 1st Qu.:20.12 1st Qu.: 7.91 1st Qu.:0.0000

Median :3.000 Mode :character Median :28.00 Median : 14.45 Median :0.0000

Mean :2.309 Mean :29.70 Mean : 32.20 Mean :0.3838

3rd Qu.:3.000 3rd Qu.:38.00 3rd Qu.: 31.00 3rd Qu.:1.0000

Max. :3.000 Max. :80.00 Max. :512.33 Max. :1.0000

# Calculate statistics for a specific variable (e.g., Age)

mean\_age <- mean(myData$Age, na.rm = TRUE)

median\_age <- median(myData$Age, na.rm = TRUE)

sd\_age <- sd(myData$Age, na.rm = TRUE)

# Print the results

cat("Mean Age:", mean\_age, "\n")

> cat("Mean Age:", mean\_age, "\n")

Mean Age: 29.69912

cat("Median Age:", median\_age, "\n")

> cat("Median Age:", median\_age, "\n")

Median Age: 28

cat("Standard Deviation of Age:", sd\_age, "\n")

> cat("Standard Deviation of Age:", sd\_age, "\n")

Standard Deviation of Age: 14.5265

# Frequency table for a categorical variable (e.g., Pclass)

table(myData$Pclass)

> table(myData$Pclass)

1 2 3

216 184 491

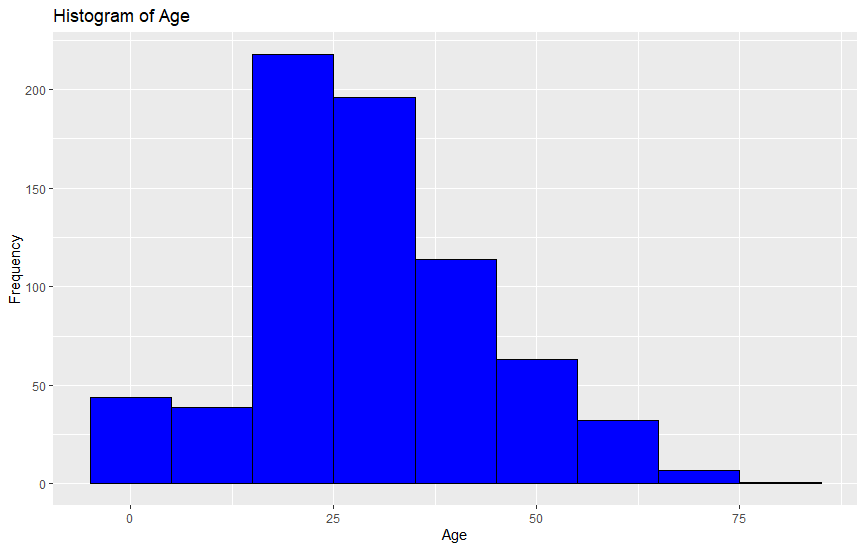
# Histogram for a numeric variable (e.g., Age)

library(ggplot2)

ggplot(myData, aes(x = Age)) +

geom\_histogram(binwidth = 10, fill = "blue", color = "black") +

labs(title = "Histogram of Age", x = "Age", y = "Frequency")



# Boxplot of a numeric variable (e.g., Age) by a categorical grouping variable (e.g., Survived)

ggplot(myData, aes(x = factor(Survived), y = Age, fill = factor(Survived))) +

geom\_boxplot() +

labs(

title = "Boxplot of Age by Survival Status",

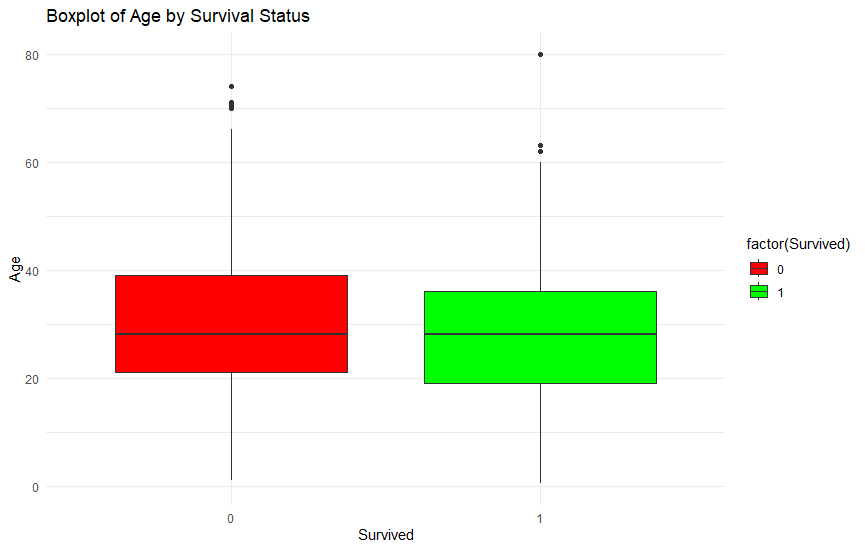
x = "Survived",

y = "Age"

) +

scale\_fill\_manual(values = c("red", "green")) + # Custom colors for each survival status

theme\_minimal() # Optional: Apply a clean theme



# t-test for a numeric variable (e.g., Age) grouped by a categorical variable (e.g., Survived)

t.test(Age ~ Survived, myData = myData)

# Chi-square test for association between two categorical variables (e.g., Pclass and Survived)

chisq.test(table(myData$Pclass, myData$Survived))

> chisq.test(table(myData$Pclass, myData$Survived))

Pearson's Chi-squared test

data: table(myData$Pclass, myData$Survived)

X-squared = 102.89, df = 2, p-value < 2.2e-16

# ANOVA to compare means of a numeric variable (e.g., Fare) across groups (e.g., Pclass)

# Perform ANOVA on Fare grouped by Pclass

anova\_result <- aov(Fare ~ factor(Pclass), data = myData)

# Display the summary of the ANOVA

summary(anova\_result)

anova\_result <- aov(fare ~ factor(class), data = myData)

summary(anova\_result)

> summary(anova\_result)

Df Sum Sq Mean Sq F value Pr(>F)

factor(Pclass) 2 776030 388015 242.3 <2e-16 \*\*\*

Residuals 888 1421769 1601

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Remove rows with missing values in Age or Fare

cleaned\_data <- na.omit(myData[, c("Age", "Fare")])

# Calculate Pearson correlation

cor(cleaned\_data$Age, cleaned\_data$Fare, method = "pearson")

> cor(cleaned\_data$Age, cleaned\_data$Fare, method = "pearson")

[1] 0.09606669

# Check if columns exist

names(myData)

> names(myData)

[1] "Pclass" "Sex" "Age" "Fare" "Survived"

# Calculate correlation matrix for Age, Fare, and SibSp

cor(myData[, c("Age", "Fare", "SibSp")], use = "complete.obs")

colnames(myData)

> colnames(myData)

[1] "Pclass" "Sex" "Age" "Fare" "Survived"

summary(myData)

> summary(myData)

Pclass Sex Age Fare Survived

Min. :1.000 Length:891 Min. : 0.42 Min. : 0.00 Min. :0.0000

1st Qu.:2.000 Class :character 1st Qu.:20.12 1st Qu.: 7.91 1st Qu.:0.0000

Median :3.000 Mode :character Median :28.00 Median : 14.45 Median :0.0000

Mean :2.309 Mean :29.70 Mean : 32.20 Mean :0.3838

3rd Qu.:3.000 3rd Qu.:38.00 3rd Qu.: 31.00 3rd Qu.:1.0000

Max. :3.000 Max. :80.00 Max. :512.33 Max. :1.0000

NA's :177

myData <- na.omit(myData)