|  |  |  |
| --- | --- | --- |
|  | **SAVEETHA SCHOOL OF ENGINEERING**  **SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES**  **INSTITUTE OF COMPUTER SCIENCE AND ENGINEERING**  **ITA 04 - STATISTICS WITH R PROGRAMMING MARCH 2023**  **ASSIGNMENT 5** |  |

**Reg No :192124140**

**Name :S.DIVYA BHASKAR**

1. i) Describe how histogram charts are created in R. Create a histogram chart for the below given age attribute.

Age : 5,45,23,30,33,32,34,35,42,41,28,29

# A.

SOURCE CODE:

#

age <- c(5,45,23,30,33,32,34,35,42,41,28,29)

# Create histogram chart

hist(age, main="Age Distribution", xlab="Age", ylab="Frequency", col="blue")

ii) Create a 3D Pie Chart for the dataset “political Knowledge” with suitable labels and colour.

SOURCE CODE:

# Load plotrix package

library(plotrix)

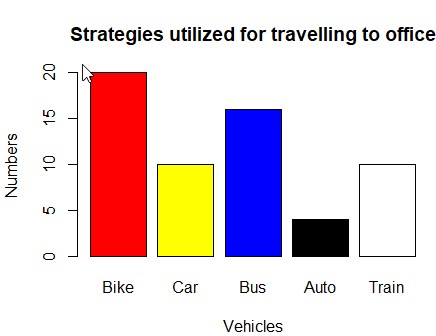
# Political Knowledge data

political\_knowledge <- c(25, 30, 10, 35)

# Create 3D pie chart

pie3D(political\_knowledge, labels=c("Very Low", "Low", "Medium", "High"), col=c("red", "yellow", "green", "blue"), explode=0.1)

1. Write R code for the below output Figure 1 shows Bike is assigned red , car is assigned yellow , bus is assigned blue , auto is assigned black , and train is assigned white. Mention the parameters used in the below barchart.



# A.

SOURCE CODE:

x <- c("bike","car","bus","auto","train")

> y <- c(0,5,10,15,20)

> colurs <- c("red","yellow","blue","black","white")

> barplot(y,names.arg=x,xlab="Vehicle",ylab="Numbers",col=colurs)

1. Create a Boxplot graph for the relation between "mpg"(miles per galloon) and "cyl(number of Cylinders) for the dataset "mtcars" available in R Environment. v)Write R program to find the sum of Natural Numbers using Recursion

# A.

SOURCE CODE:

# Load the mtcars dataset

data(mtcars)

# Create boxplot graph for mpg and cyl

boxplot(mpg ~ cyl, data = mtcars, main = "Miles per Gallon vs Number of Cylinders", xlab = "Number of Cylinders", ylab = "Miles per Gallon")

# Function to calculate the sum of natural numbers

sum\_natural\_numbers <- function(n) {

if(n == 1) {

return(1)

} else {

return(n + sum\_natural\_numbers(n-1))

}

}

# Test the function

sum\_natural\_numbers(5)

1. a. Melt ‘airquality’ data set which inbuild dataset in ‘R’ and display as a long – format data?

# A.

# Load the "reshape2" package

library(reshape2)

# Melt the "airquality" dataset

airquality\_melted <- melt(airquality)

# Print the melted dataset

print(airquality\_melted)

* 1. Melt air quality data and specify month and day to be “ID variables”?

# A.

# Melt the "airquality" dataset and specify "month" and "day" as ID variables

airquality\_melted <- melt(airquality, id.vars = c("Month", "Day"))

# Print the melted dataset

print(airquality\_melted)

* 1. Cast the molten ‘airquality’ data set.

# A.

# Cast the molten "airquality" dataset

airquality\_casted <- dcast(airquality\_melted, Month + Day ~ variable)

# Print the casted dataset

print(airquality\_casted)

* 1. Use cast function appropriately and compute the average of Ozone, Solar, Wind and temperature per month?

# A.

# Load the "reshape2" package

library(reshape2)

# Cast the "airquality" dataset

airquality\_casted <- cast(airquality, Month ~ variable, mean)

# Select only the relevant columns

airquality\_casted <- airquality\_casted[, c("Month", "Ozone", "Solar.R", "Wind", "Temp")]

# Print the casted dataset

print(airquality\_casted)

* 1. Create a boxplot for ozone reading of ‘airquality’ dataset. Add title, label and color.

# A.

# Create a boxplot for the "Ozone" reading of the "airquality" dataset

boxplot(airquality$Ozone, main = "Ozone Reading of Air Quality Dataset", xlab = "Ozone Reading", col = "blue")

1. a. Write a program for creating a pie-chart in R using the input vector (21,62,10,53). Provide labels for the chart as ‘London’, ‘New York’, ‘Singapore’, ‘Mumbai’. Add a title to the chart as ‘city pie-chart’ and add a legend at the top right corner of the chart.
   1. Using linear regression analysis establish a relationship between height and weight of a person using the input vector given below.

# Values of height

151, 174, 138, 186, 128, 136, 179, 163, 152, 131 # Values of weight.

63, 81, 56, 91, 47, 57, 76, 72, 62, 48 Predict the weight of a person with height 170.

c. Visualize the regression graphically.

d. Call ‘mtcars’ which is built in dataset in ‘R’ and plot distribution of mpg feature.

Make x axis range from 10 to 35 and plot title as “More trends in 70’s Vehicles”. e. Find statistical summary of the ‘mtcars’ dataset.

# A.

# Input vector

> # Input vector

> input\_vector <- c(21, 62, 10, 53)

> # Labels for chart

> labels <- c("London", "New York", "Singapore", "Mumbai")

> # Create pie chart

> pie(input\_vector, labels = labels, main = "City Pie Chart")

> # Add legend

> legend("topright", legend = labels, fill = rainbow(length(labels)))

> # Input vectors

> height <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)

> weight <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)

> # Perform linear regression

> fit <- lm(weight ~ height)

> # Predict weight for height 170

> predict(fit, data.frame(height = 170))

1

76.22869

> # Create scatter plot

> plot(height, weight, main = "Height vs Weight")

> # Add regression line

> abline(fit, col = "red")

> # Set x-axis range

> x\_range <- c(10, 35)

> # Create histogram

> hist(mtcars$mpg, xlim = x\_range, main = "More Trends in 70's Vehicles", xlab = "MPG")

> e)summary(mtcars)

1. There is a popular built-in data set in R called "mtcars" (Motor Trend Car Road Tests), which is retrieved from the 1974 Motor Trend US Magazine.

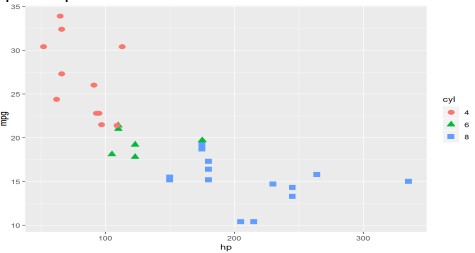
(i)Find the dimension of the data set & Give the statistical summary of the features. (ii)Create correlation matrix between mpg vs all other features and print the high 3 correlated Features(both +ve and -ve)

(iii)Plot the Box plot for “mpg” group by “cyl” feature.

(iv)Create a scatter plot graph for the relation between "mpg"(miles per gallon) and

"hp"(horse power) group by cyl(number ofcylinder)

**Sample Output:**



(v) Generate a multiple regression model to establish the relationship between "mpg" as a response variable with "disp","hp" and "wt" as predictor variables. Plot the regression line. Find the MSE of the model.

# A.

data(mtcars)

> # Dimensions of the data set

> dim(mtcars)

[1] 32 11

> # Statistical summary of the features

> summary(mtcars)

mpg cyl disp hp drat

Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0 Min. :2.760

1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5 1st Qu.:3.080

Median :19.20 Median :6.000 Median :196.3 Median :123.0 Median :3.695

Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7 Mean :3.597

3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0 3rd Qu.:3.920

Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0 Max. :4.930

wt qsec vs am gear

Min. :1.513 Min. :14.50 Min. :0.0000 Min. :0.0000 Min. :3.000

1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:3.000

Median :3.325 Median :17.71 Median :0.0000 Median :0.0000 Median :4.000

Mean :3.217 Mean :17.85 Mean :0.4375 Mean :0.4062 Mean :3.688

3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:4.000

Max. :5.424 Max. :22.90 Max. :1.0000 Max. :1.0000 Max. :5.000

carb

Min. :1.000

1st Qu.:2.000

Median :2.000

Mean :2.812

3rd Qu.:4.000

Max. :8.000

> # Correlation matrix between mpg and other features

> correlations <- cor(mtcars)

> # Sort the correlations by the absolute values

> correlations\_sorted <- sort(abs(correlations[,'mpg']), decreasing = TRUE)

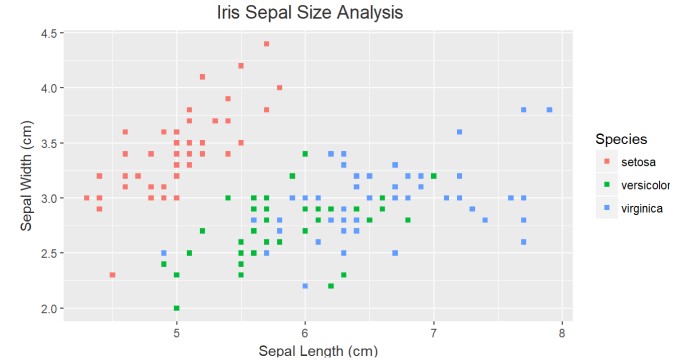
> # Print the top three highest correlated features

> correlations\_sorted[1:3]

mpg wt cyl

1. 0.8676594 0.8521620
2. (i) Use melt and cast function to find mean of numeric data in dataset based on Species group.
   1. Generate a suitable plot which summaries statistical parameter of Sepal.Width based on Species group
   2. Generate scatter plot between Sepal.Length vs Sepal.Width grouped by Specias.

**Sample Output:**



# A.

a)# Load the reshape2 package

library(reshape2)

# Melt the iris data set

melted\_iris <- melt(iris, id.vars = 'Species')

# Cast the molten data set to find the mean of the numeric variables based on the species group

mean\_iris <- dcast(melted\_iris, Species ~ variable, mean)

# View the mean\_iris data frame

mean\_iris

b)# Load the ggplot2 package

library(ggplot2)

# Create a box plot of Sepal.Width grouped by Species

ggplot(iris, aes(x = Species, y = Sepal.Width)) +

geom\_boxplot() +

ggtitle("Box plot of Sepal.Width by Species")

c)# Create a scatter plot of Sepal.Length vs Sepal.Width grouped by Species

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +

geom\_point() +

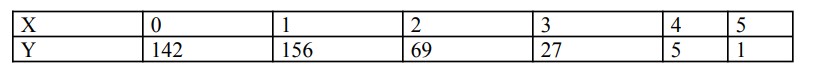
ggtitle("Scatter plot of Sepal.Length vs Sepal.Width by Species")

1. A) Heights(in cm) of father and son are given as follows

Father(X): 150 152 155 157 160 161 164 165 Son (Y) : 154 156 158 159 160 162 161 164

Fit a regression line parameters to predict the height of son given the height of father. Write R code for same.

B) Fit a regression line parameters distribution with the following data



# A.

A)

# Heights of father and son

father\_height <- c(150, 152, 155, 157, 160, 161, 164, 165)

son\_height <- c(154, 156, 158, 159, 160, 162, 161, 164)

# Fit a regression line to predict the height of son given the height of father

regression\_model <- lm(son\_height ~ father\_height)

# Print the summary of the regression model

summary(regression\_model)

B)

# Fit a regression line to the data

regression\_model <- lm(y ~ x, data = data\_df)

# Print the summary of the regression model

summary(regression\_model)