ITA0448- STATISTICS WITH R PROGRAMMING VECTORIZED EXPRESSION

NAME:P.BARATH
REG NO:19212147

DAY_5_ ASSIGNMENT

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

```
1)i)age <- c(5, 45, 23, 30, 33, 32, 34, 35, 42, 41, 28, 29)
hist(age, main = "Age Distribution", xlab = "Age", ylab = "Frequency")
ii)library(plotrix)
political_knowledge <- c("Low", "Low", "Medium", "Medium", "High",
"High", "High")
pie3D(table(political_knowledge), main = "Political Knowledge",
col = c("red", "orange", "green"), explode = 0.1)
```

2. 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.
PROGRAM:

```
data(mtcars)
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders", ylab =
"Miles per Gallon",
main = "Relationship between MPG and Cylinders")
sum_natural <- function(n) {
if (n == 0) { # Base case: sum of 0 numbers is 0
return(0)
```

```
} else { # Recursive case: sum of n numbers is n + sum of (n-1) numbers
return(n + sum_natural(n-1))
```

}
sum_natural(10)
Output:
55
3. 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
PROGRAM:
data(mtcars)
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders", ylab =
"Miles per Gallon",
main = "Relationship between MPG and Cylinders")
sum_natural <- function(n) {
if (n == 0) { # Base case: sum of 0 numbers is 0
return(0)
} else { # Recursive case: sum of n numbers is n + sum of (n-1) numbers
return(n + sum_natural(n-1))
}
}
sum_natural(10) # Output: 55
4. a. Melt 'airquality' data set which inbuild dataset in 'R' and display as a
long – format
data?
b. Melt air quality data and specify month and day to be "ID variables"?
c. Cast the molten 'airquality' data set.
d. Use cast function appropriately and compute the average of Ozone,
Solar, Wind
and temperature per month?
e. Create a boxplot for ozone reading of 'airquality' dataset. Add title,
label and color.
PROGRAM:
4)a)library(reshape2)
data(mtcars)

```
airquality_melt <- melt(airquality)
b)airquality_melt <- melt(airquality, id.vars = c("Month", "Day"))
head(airquality_melt)
c)airquality_cast <- dcast(airquality_melt, Month ~ variable)
head(airquality_cast)
d)library(plyr)
airquality_avg <- cast(airquality_melt, Month ~ variable, mean)
head(airquality_avg)
e)library(ggplot2)
ggplot(data = airquality, aes(x = "", y = Ozone)) +
geom_boxplot(fill = "lightblue", color = "blue") +
labs(title = "Boxplot of Ozone Readings", y = "Ozone Reading")
head(airquality_melt)
```

5. 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.
b. 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.
PROGRAM:

```
> input_vector <- c(21, 62, 10, 53)
> labels <- c("London", "New York", "Singapore", "Mumbai")
> pie(input_vector, labels = labels, main = "City Pie Chart")
> legend("topright", legend = labels, fill = rainbow(length(labels)))
> height <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
> weight <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)
> fit <- lm(weight ~ height)
> predict(fit, data.frame(height = 170)) 1
76.22869
plot(height, weight, main = "Height vs Weight")
> abline(fit, col = "red")
> hist(mtcars$mpg, xlim = x_range, main = "More Trends in 70's
Vehicles", xlab = "MPG")
> e)summary(mtcars)
```

6. 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.

PROGRAM:
data(mtcars)
> dim(mtcars)
[1] 32 11
> 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
> correlations <- cor(mtcars)
> correlations_sorted <- sort(abs(correlations[,'mpg']), decreasing =
TRUE)
> correlations_sorted[1:3] mpg wt cyl
1.0000000 0.8676594 0.8521620

7. (i) Use melt and cast function to find mean of numeric data in dataset based on
Species group.
(ii) Generate a suitable plot which summaries statistical parameter of Sepal.Width

based on Species group
(iii) Generate scatter plot between Sepal.Length vs Sepal.Width grouped by Specias.
Sample Output:

PROGRAM: 7)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")

8. 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
PROGRAM:

```
father_height <- c(150, 152, 155, 157, 160, 161, 164, 165)
son_height <- c(154, 156, 158, 159, 160, 162, 161, 164)
regression_model <- lm(son_height ~ father_height)
summary(regression_model)
B)# Fit a regression line to the data
regression_model <- lm(y ~ x, data = data_df)
summary(regression_model)
```