Lab Sheet 04

IT2120 - Probability and Statistics

IT24103420

Jayawardhana H.G.T.N.

Exercise

Instructions: Create a folder in your desktop with your registration number (Eg: "IT....."). You need to save the R script file and take screenshots of the command prompt with answers and save it in a word document inside the folder. Save both R script file and word document with your registration number (Eg: "IT....."). After you finish the exercise, zip the folder and upload the zip file to the submission link.

1. Import the dataset ('Exercise.txt') into R and store it in a data frame called "branch_data".

```
#Question 01
setwd("C:\\Users\\it24103420\\Desktop\\IT24103420")
branch_data<-read.table("Exercise.txt",header =TRUE, sep=",")
head(branch_data)
```

```
> #Question 01
> setwd("C:\\Users\\it24103420\\Desktop\\IT24103420")
> branch_data<-read.table("Exercise.txt",header =TRUE, sep=",")
> head(branch_data)
  Branch Sales_X1 Advertising_X2 Years_X3
1      1      3.4            120        4
2      2      4.1            150        7
3      3      2.8             90        3
4      4      5.0            200       10
5      5      3.7            110        5
6      6      4.5            175        6
```

2. Identify the variable type and scale of measurement for each variable.

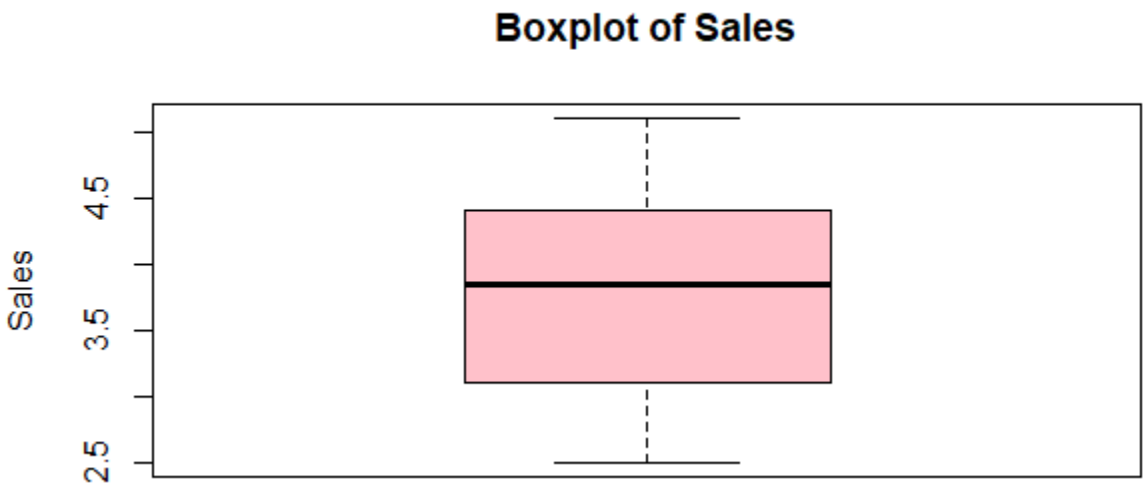
```
#Question_02
str(branch_data)
```

```
> #Question_02
> str(branch_data)
'data.frame':   30 obs. of  4 variables:
 $ Branch        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Sales_X1      : num  3.4 4.1 2.8 5 3.7 4.5 3 4.9 3.2 2.5 ...
 $ Advertising_X2: int  120 150 90 200 110 175 95 185 105 80 ...
 $ Years_X3      : int  4 7 3 10 5 6 2 9 4 1 ...
```

3. Obtain boxplot for sales and interpret the shape of the sales distribution.

```
#Question_03
boxplot(branch_data$Sales,
        main = "Boxplot of Sales",
        ylab = "Sales",
        col = "pink")
```

```
> #Question_03
> boxplot(branch_data$Sales,
+         main = "Boxplot of Sales",
+         ylab = "Sales",
+         col = "pink")
```



4. Calculate the five number summary and IQR for advertising variable.

```
#Question_04
fivenum(branch_data$Advertising)

summary(branch_data$Advertising)


IQR(branch_data$Advertising)
```

```
> #Question_04
> fivenum(branch_data$Advertising)
[1]  80.0 100.0 132.5 160.0 210.0
>
> summary(branch_data$Advertising)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   80.0   101.2   132.5   134.8   158.8   210.0
>
>
> IQR(branch_data$Advertising)
[1] 57.5
```

5. Write an R function to find the outliers in a numeric vector and check for outliers in years variables.

```
#Question_05
find_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
  outliers <- x[x < lower | x > upper]
  return(outliers)
}
find_outliers(branch_data$Years)
```

```
> #Question_05
> find_outliers <- function(x) {
+   Q1 <- quantile(x, 0.25)
+   Q3 <- quantile(x, 0.75)
+   IQR <- Q3 - Q1
+   lower <- Q1 - 1.5 * IQR
+   upper <- Q3 + 1.5 * IQR
+   outliers <- x[x < lower | x > upper]
+   return(outliers)
+ }
> find_outliers(branch_data$Years)
integer(0)
```