Bansilal RamnathAgarwal Charitable Trust's
VISHWAKARMA INSTITUTE OF TECHNOLOGY – PUNE
Department of SY Common

MD2201: Data Science

Name of the student: Bhavin Patil **Roll No. 78**

Div: D **Batch: B-3**

Date of performance:

Experiment No.1

Title: Laboratory on Data Visualization

Aim: i. To explore the dataset for different case study examples with different commands.
ii. To plot the Box plot and scatter plot.

Software used: Programming language R.

Code Statement:

1. Write a single R code to display the answers for the following questions.

Case Study: Consider the "pollutant" data set.

1. What is the mean of "Temp" when "Month" is equal to 6?
2. How many observations are there in the given data?
3. Print last two rows of the data.
4. What is the value of Ozone in 47th row?
5. How many values are missing in Ozone column?
6. What is the mean of Ozone column excluding missing values?
7. Extract the subset of rows of the data frame where Ozone values are above 31 and Temp values are above 90. What is the mean of Solar.R in this subset?
8. What was the maximum ozone value in the month of May (i.e. Month is equal to 5)?

2. Write a single R code to display the answers the following questions

Case Study: Hair Eye color Data set

1. How many people have brown eye color?
2. How many people have Blonde hair?
3. How many Brown haired people have Black eyes?
4. What is the percentage of people with Green eyes?
5. What percentage of people have red hair and Blue eyes?

3. Write a single R code to display the answers for the following questions

Case study: Germination Data Set

1. What is the average number of seeds germinated for the uncovered boxes with level of watering equal to 4?
2. What is the median value for the data covered boxes?

Establish conclusions on the basis of available data and write them in the conclusion part.

a. Association of levels of watering with the number of germinating seeds in case of covered boxes as well as uncovered boxes.

b. Association of number of germinating seeds with the fact that the boxes were covered or uncovered.

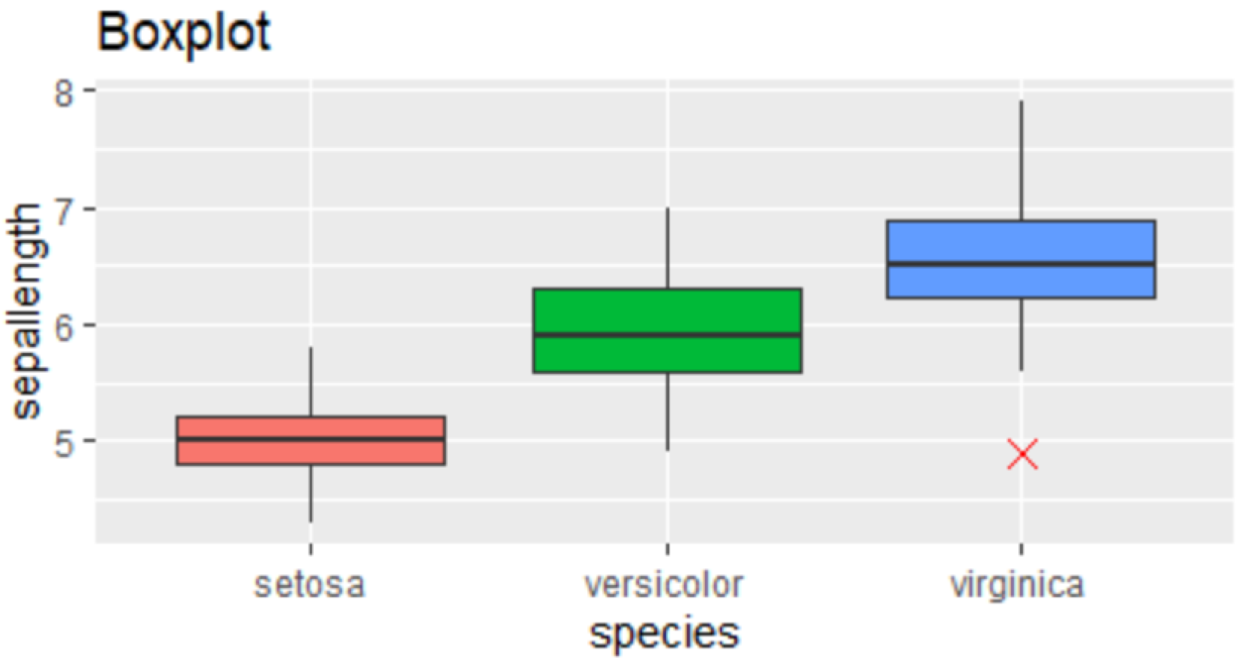
4. Write a single R code :

i. To display the Boxplot for sepal length of iris data set as shown below

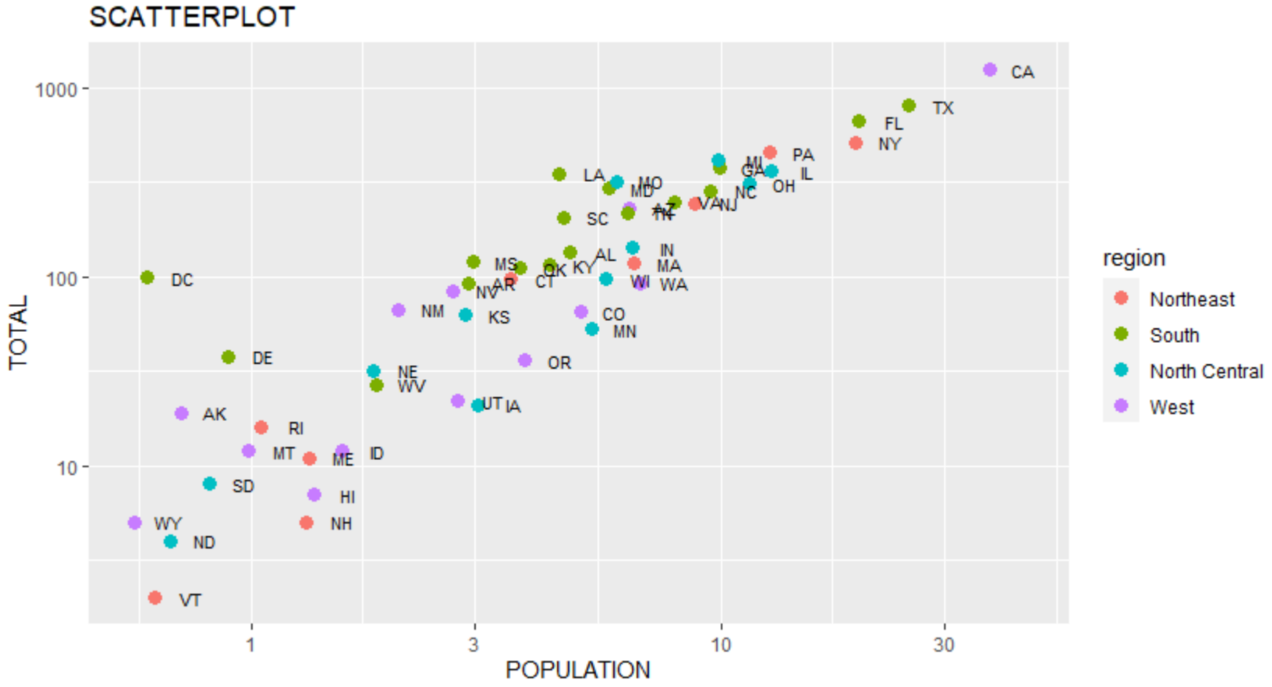
ii. To display the Scatter plot for murders data set present in "dslabs" package as shown below.

Give proper title, x,y axis label etc. to each plot.

Expected Boxplot:



Expected Scatter Plot:



Code:

```
#=====================================================================
#Case Study No. 1
#=====================================================================

cat("\n=======================Solution for Case Study 1=======================")

dataset1 <- read.csv("pollutant_csv.csv")
meanofTemp <- mean( dataset1$Temp [dataset1$Month == 6])
cat("\n\nQ. 1 Mean of Temp when Month = 6 : ", meanofTemp,"\n\n")

n <-nrow(dataset1)
cat("Q. 2 Number of observations in the given data : ", n,"\n\n")

cat("Q. 3 Last two rows: \n")
print(tail(dataset1,2))

cat("\n\nQ. 4 Value of Ozone in 47th row : ", dataset1$Ozone[47],"\n\n")

cat("Q. 5 Number of missing values in Ozone Column : ", sum(is.na(dataset1$Ozone)),"\n\n")

cat("Q. 6 Mean of Ozone column excluding missing values : ", mean(is.na(dataset1$Ozone)),"\n\n")

a <- dataset1[dataset1$Ozone > 31 & dataset1$Temp > 90,]
cat("Q. 7 Ozone above 31 and Temp above 90: ")
print(a)

a1 <- mean(dataset1$Solar.R, na.rm = T)
cat("Q. 8 Mean of Solar.R ",a1,"\n\n")
```

```
a2 <- max(dataset1$Ozone[dataset1$Month == 5], na.rm = T)
cat("Q. 9 Max Ozone Layer in the month of May: ", a2,"\n\n")

#======================================================================
#Case Study No. 2
#======================================================================

cat("\n\n\n*************Solution for Case Study 2****************")

dataset2 <- read.csv("hair_eye_color_csv.csv")

ans1 <- sum(dataset2$Eye.Color == "Brown")
cat("\n\nQ. 1) Number of people having Brown eyes: ",ans1,"\n\n")

ans2 <- sum(dataset2$Hair.Color == "Blonde")
cat("Q. 2) Number of people having Blonde Hairs: ",ans2,"\n\n")

ans3 <- sum(dataset2$Hair.Color == "Blonde" & dataset2$Eye.Color == "Black")
cat("Q. 3) Number of people having Blonde Hair and Black Eyes: ",ans3,"\n\n")

ans4 <- (sum(dataset2$Eye.Color == "Green") / length(dataset2) )* 100
cat("Q. 4) Percentage of the people with green eyes: ",ans4,"\n\n")

ans5 <- (sum(dataset2$Hair.Color == "Red" & dataset2$Eye.Color == "Blue") / length(dataset2) )* 100
cat("Q. 5) Percentage of the people with red hairs and blue eyes: ",ans5,"\n\n")

#======================================================================
#Case Study No. 3
#======================================================================

cat("\n\n\n*************Solution for Case Study 3****************")

dataset3 <- read.csv("germination_csv.csv");

answer1 <- mean(dataset3$Box == "Uncovered" & dataset3$water_amt == 4, rm.na = T) / length(dataset3)
answer1 <- mean(dataset3$germinated[dataset3$Box == "Uncovered" & dataset3$water_amt == 4])
cat("\n\nQ. 1) Average Number of seeds = ", answer1,"\n\n")

answer2 <- median(dataset3$Box == "Covered")
cat("Q. 2) Median: ", answer2)

library(dslabs)
t<-
    ggplot(iris,aes(Species,Sepal.Length,fill=Species))+geom_boxplot(outlier.color="red",outlier.shape=4,outlier.size = 4)+theme(legend.position = "none")+ggtitle("Boxplot")+xlab("species")+ylab("sepallength")
print(t)


y<-
    ggplot(murders,aes(population/10^6,total,col=region))+geom_point(size=3)+scale_x_log10()+scale_y_log10()+geom_text(aes(label=abb),size=3,nudge_x=0.050)+labs(title="SCATTERPLOT",x="Population",y="Total")
print(y)
```

Results:

```
========================Solution for Case Study 1========================


Q. 1 Mean of Temp when Month = 6 :  79.1


Q. 2 Number of observations in the given data :  153


Q. 3 Last two rows:
    Ozone Solar.R Wind Temp Month Day
152    18     131  8.0   76     9  29
153    20     223 11.5   68     9  30



Q. 4 Value of Ozone in 47th row :  21


Q. 5 Number of missing values in Ozone Column :  37


Q. 6 Mean of Ozone column excluding missing values :  0.2418301


Q. 7 Ozone above 31 and Temp above 90:      Ozone Solar.R Wind Temp Month Day
NA      NA      NA   NA   NA    NA  NA
NA.1    NA      NA   NA   NA    NA  NA
69      97     267  6.3   92     7   8
70      97     272  5.7   92     7   9
NA.2    NA      NA   NA   NA    NA  NA
NA.3    NA      NA   NA   NA    NA  NA
120     76     203  9.7   97     8  28
121    118     225  2.3   94     8  29
122     84     237  6.3   96     8  30
123     85     188  6.3   94     8  31
```

```
124     96     167  6.9   91     9   1
125     78     197  5.1   92     9   2
126     73     183  2.8   93     9   3
127     91     189  4.6   93     9   4
```

Q. 8 Mean of Solar.R 185.9315

Q. 9 Max Ozone Layer in the month of May: 115

*************Solution for Case Study 2****************

Q. 1) Number of people having Brown eyes: 10

Q. 2) Number of people having Blonde Hairs: 6

Q. 3) Number of people having Blonde Hair and Black Eyes: 1

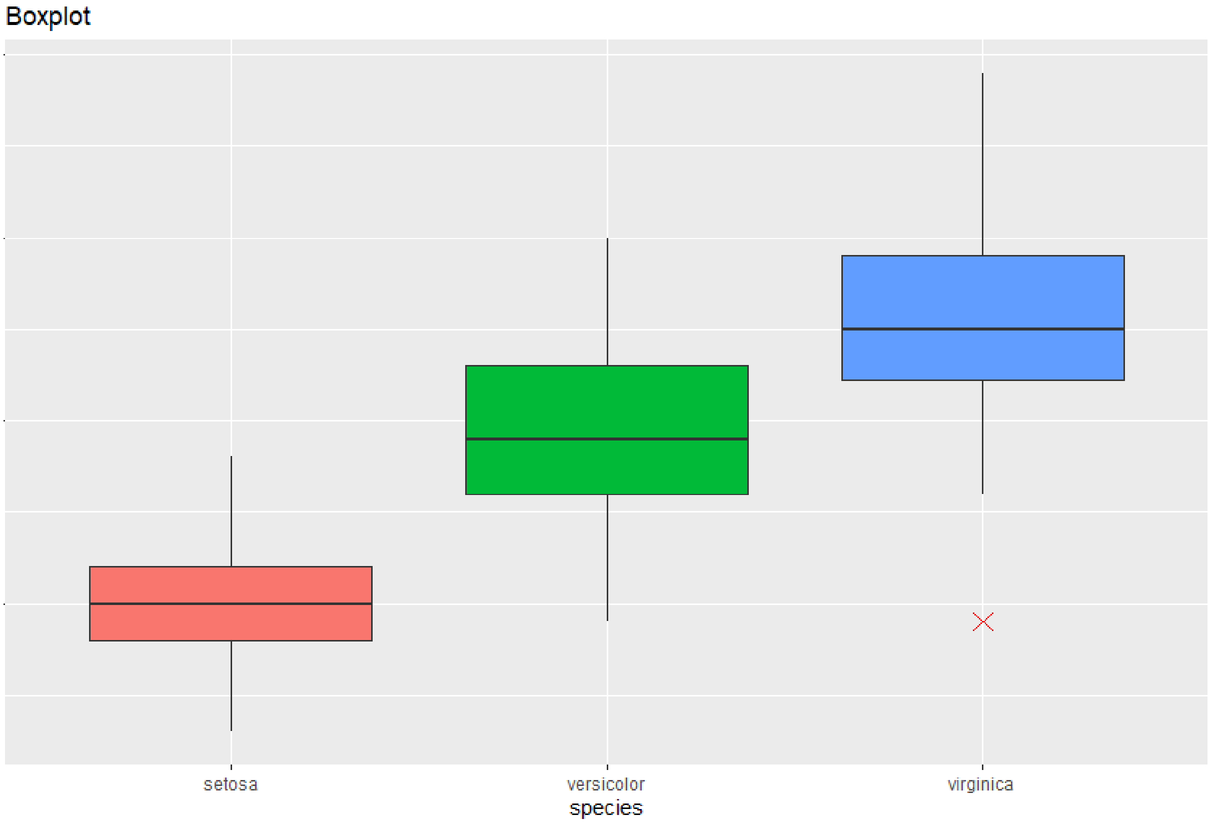
Q. 4) Percentage of the people with green eyes: 66.66667

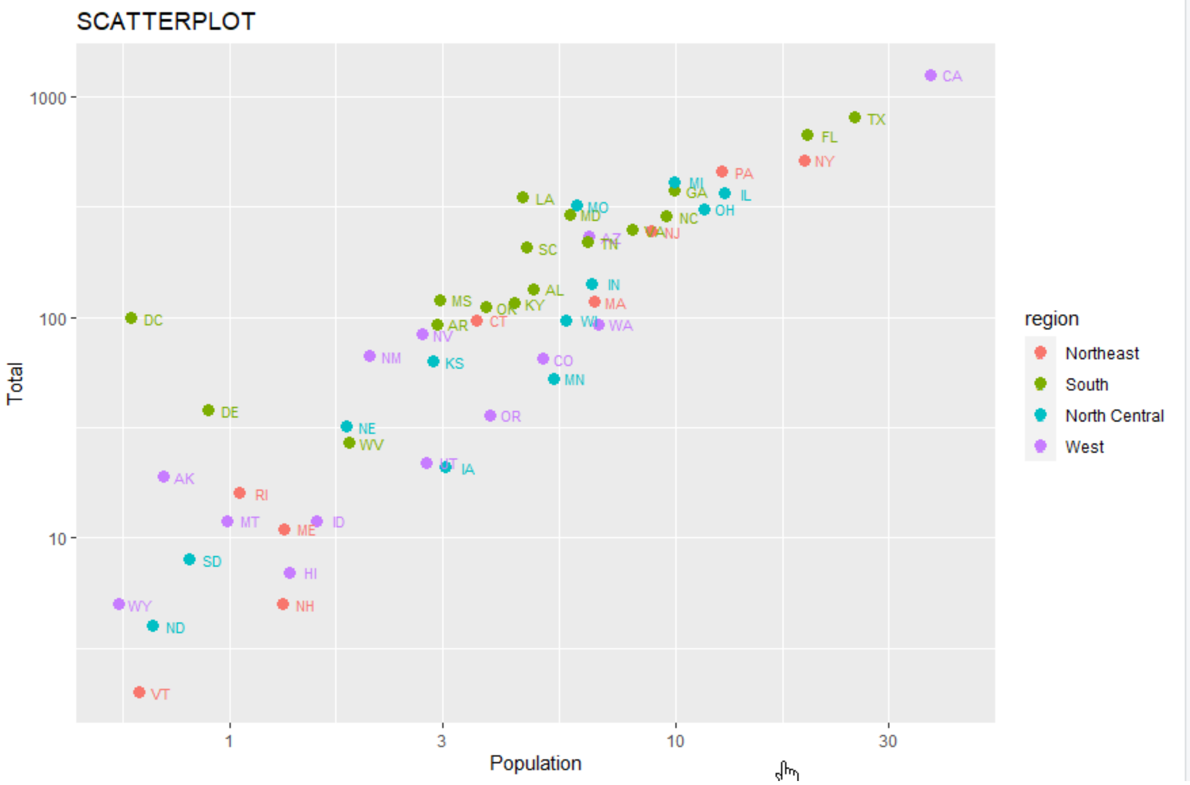
Q. 5) Percentage of the people with red hairs and blue eyes: 33.33333

*************Solution for Case Study 3****************

Q. 1) Average Number of seeds = 78

Q. 2) Median: 0.5


Boxplot
setosa
versicolor
virginica
species


SCATTERPLOT
Total
1000
100
10
1
3
10
30
Population
region
● Northeast
● South
● North Central
● West
CA
TX
FL
NY
PA
MI
GA
IL
OH
LA
MO
MD
NC
VA
NJ
AZ
TN
SC
AL
IN
MA
MS
OK
KY
AR
CT
WI
WA
NV
DC
NM
KS
CO
MN
DE
NE
WV
OR
UT
IA
AK
RI
MT
ME
ID
SD
HI
WY
NH
ND
VT

Conclusion: exploring the dataset for different case study examples using different commands for correct outputs and plotting the Box plot and Scatter plot successfully as expected.