MD2201: Data Science

Name of the student: Rajeev Tapadia **Roll No.: 67**

Div: C **Batch: 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)?

```
      # 1. Mean of Temp when Month is 6
mean(pollutant$Temp[pollutant$Month == 6])
# 2. Number of observations
nrow(pollutant)
# 3. Print last two rows
tail(pollutant, 2)
# 4. Ozone value in 47th row
pollutant$Ozone[47]
# 5. Missing values in Ozone
sum(is.na(pollutant$Ozone))
# 6. Mean of Ozone excluding missing values
mean(pollutant$Ozone[!is.na(pollutant$Ozone)])
# 7. Mean of Solar.R for Ozone > 31 and Temp > 90
mean(pollutant$Solar.R[pollutant$Ozone > 31 & pollutant$Temp > 90])
```

```r
# 8. Maximum Ozone value in May (Month = 5)
max(pollutant$Ozone[pollutant$Month == 5])
```

```
> # 1. Mean of Temp when Month is 6
> mean(pollutant$Temp[pollutant$Month == 6])
[1] 79.1
> # 2. Number of observations
> nrow(pollutant)
[1] 153
> # 3. Print last two rows
> tail(pollutant, 2)
    Ozone Solar.R Wind Temp Month Day
152    18     131  8.0   76     9  29
153    20     223 11.5   68     9  30
> # 4. Ozone value in 47th row
> pollutant$Ozone[47]
[1] 21
> # 5. Missing values in Ozone
> sum(is.na(pollutant$Ozone))
[1] 37
> # 6. Mean of Ozone excluding missing values
> mean(pollutant$Ozone[!is.na(pollutant$Ozone)])
[1] 42.12931
```

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?

```r
# 1.    How many people have brown eye color?
brown_eye <- subset(dataset, dataset$Eye.Color == "Brown")
nrow(brown_eye)

# 2.    How many people have Blonde hair?
sum(dataset$Hair.Color == "Blonde")

# 3.    How many Brown haired people have Black eyes?
sum(dataset$Hair.Color == "Brown" & dataset$Eye.Color == "Black")

# 4.    What is the percentage of people with Green eyes?
green_eye_percent <- sum(dataset$Eye.Color == "Green") * 100 / nrow(dataset)
cat(green_eye_percent, "%")

# 5.    What percentage of people have red hair and Blue eyes?
sum(dataset$Hair.Color == "Red" & dataset$Eye.Color == "Blue") * 100 /
nrow(dataset)
```

```
> # 1.  How many people have brown eye color?
> brown_eye <- subset(dataset, dataset$Eye.Color == "Brown")
> nrow(brown_eye)
[1] 10
> # 2.  How many people have Blonde hair?
> sum(dataset$Hair.Color == "Blonde")
[1] 6
> # 3.  How many Brown haired people have Black eyes?
> sum(dataset$Hair.Color == "Brown" & dataset$Eye.Color == "Black")
[1] 2
> # 4.  What is the percentage of people with Green eyes?
> green_eye_percent <- sum(dataset$Eye.Color == "Green") * 100 / nrow(dataset)
> cat(green_eye_percent, "%")
10 %
> # 5.  What percentage of people have red hair and Blue eyes?
> sum(dataset$Hair.Color == "Red" & dataset$Eye.Color == "Blue") * 100 / nrow(da
[1] 5
>
```

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?

```
# 1. avg no of seeds germinated for uncovered boxes with watering 6?
subset1 <- subset(dataset, dataset$Box == "Uncovered" & dataset$water_amt
== 4)
mean(subset1$germinated)

# 2.    What is the median value for the data covered boxes?
covered_boxes <- subset(dataset, dataset$Box == "Covered")
median(covered_boxes$germinated)

```

```
> # 1. avg no of seeds germinated for uncovered boxes with watering 6?
> subset1 <- subset(dataset, dataset$Box == "Uncovered" & dataset$water_
> mean(subset1$germinated)
[1] 78
> # 2.  What is the median value for the data covered boxes?
> covered_boxes <- subset(dataset, dataset$Box == "Covered")
> median(covered_boxes$germinated)
[1] 45
>
```

4. Write a single R code :

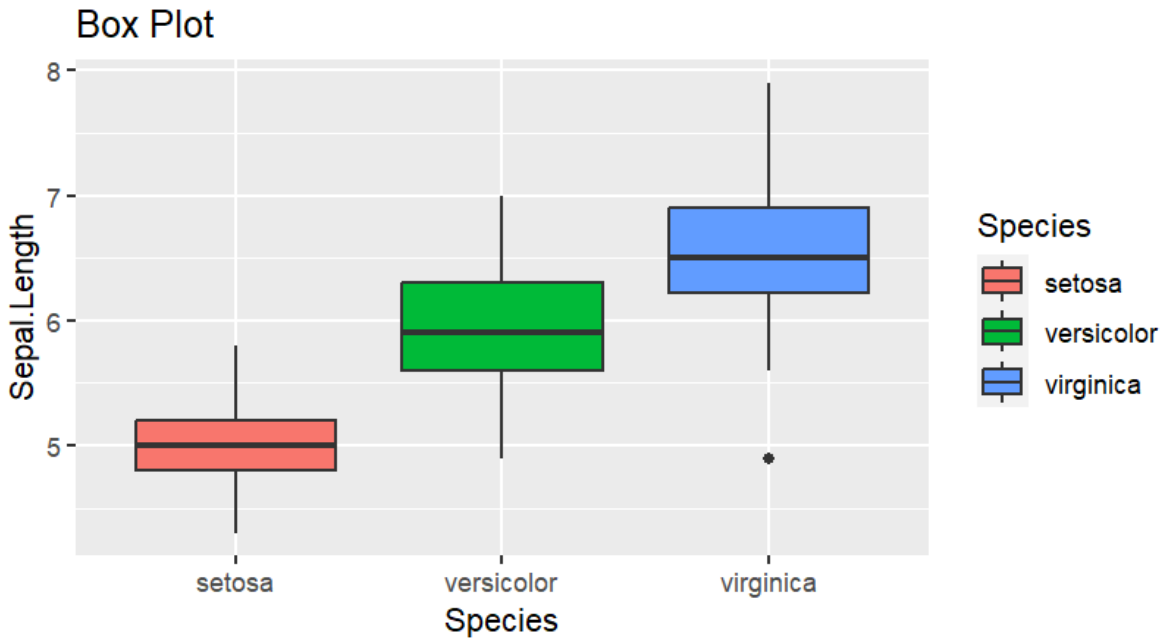
i. To display the Boxplot for sepal length of iris data set as shown below

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

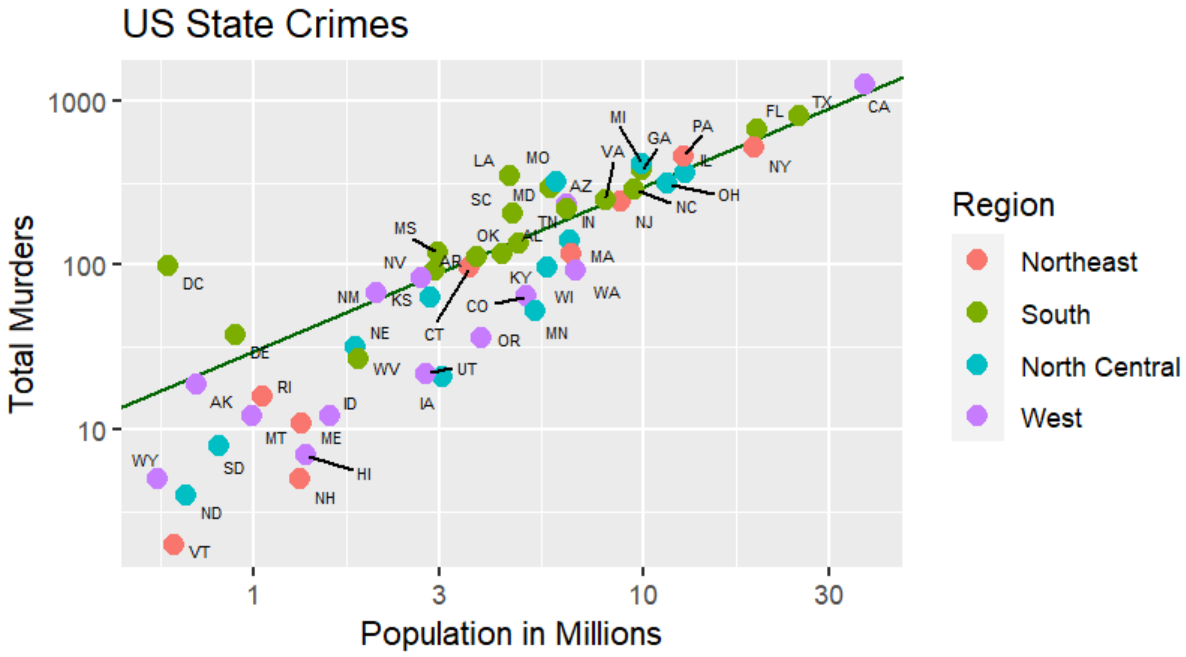
Give proper title, x,y axis label etc. to each plot.

```
# plot using ggplot2
ggplot(data = iris, aes(x = Species, y = Sepal.Length, )) +
  geom_boxplot(aes(fill = Species)) +
  ggtitle("Box Plot")
```

Expected Boxplot:



Expected Scatter Plot:




```
ggplot(murderdf, aes(x = population/ 10^6, y = total)) +
  geom_abline(intercept = log10(r), col = "darkgreen") +
  geom_point(aes(color = region), size = 3) +
  geom_text_repel(nudge_x = 0.005, size = 2, aes(label = abb)) +
  scale_x_log10("Population in Millions") +
  scale_y_log10("Total Murders") +
  ggtitle("US State Crimes") +
  scale_color_discrete(name = "Region")
```