APPLIED STATISTICAL ANALYSIS LAB

NAME: ADITI KULKARNI

ROLL NO: 55

YEAR: SY

DIVISION: E (E3)

SRN NO: 202201893

ASSIGNMENT 3

STATEMENT: *Obtain summary statistics for variables using pre-existing file.*

THEORY:

1. **Minimum, Maximum, and Range:**
   - *The minimum and maximum values of the 'SepalLengthCm' column are calculated and displayed.*
   - *The range of the 'SepalLengthCm' column is calculated and displayed.*
   - *A custom function range2 is defined to calculate the range of a* ***given variable.***
2. **Mean, Median, Quartiles, and Interquartile Range:**
   - *The mean and median of the 'SepalLengthCm' column are calculated and displayed.*
   - *The 50th percentile (median), 25th percentile (first quartile), and 75th percentile (third quartile) are calculated and displayed.*

- *The interquartile range (IQR) of the 'SepalLengthCm' column is calculated and displayed.*

3. **Standard Deviation and Variance:**
    - *The standard deviation and variance of the 'SepalLengthCm' column are calculated and displayed.*
4. **Summary and Grouped Summary:**
    - *A summary of the dataset is displayed using summary.*
    - *The by function is used to display summaries grouped by the 'Species' column.*
5. **Mode:**
    - *The mode (most frequent value) of the 'SepalLengthCm' column is calculated and displayed.*
    - *The code then sorts and displays occurrences of unique values in descending order.*
    - *Similar sorting and display are done for 'Species' as well.*
6. **Bar Plot and Relative Frequency:**
    - *A bar plot of the 'Species' column is created using the barplot function.*
    - *A bar plot of the relative frequencies of 'Species' is created using prop.table and barplot.*
7. **Histograms:**
    - *A histogram of the 'SepalLengthCm' column is created using the hist function.*
    - *An alternative histogram is created using the ggplot2 library.*
8. **Box Plot:**
    - *A box plot of 'SepalLengthCm' is created using the boxplot function.*
    - *A box plot of 'SepalLengthCm' grouped by 'Species' is created.*
9. **Dot Plot:**

- *A dot plot of 'SepalLengthCm' grouped by 'Species' is created using* ***the lattice library.***

10. Scatter Plot:

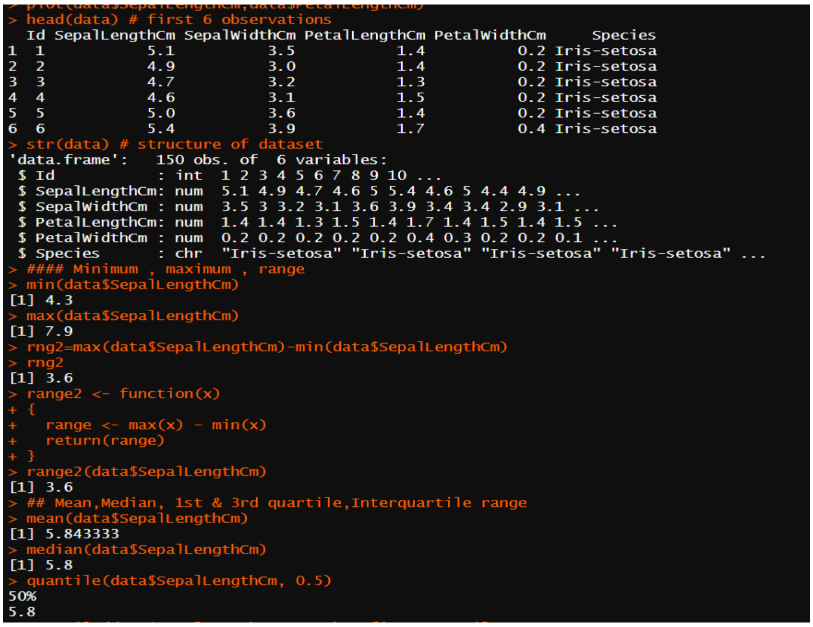
- *A scatter plot of 'SepalLengthCm' against 'PetalLengthCm' is created using the plot function.*

SOURCE CODE:

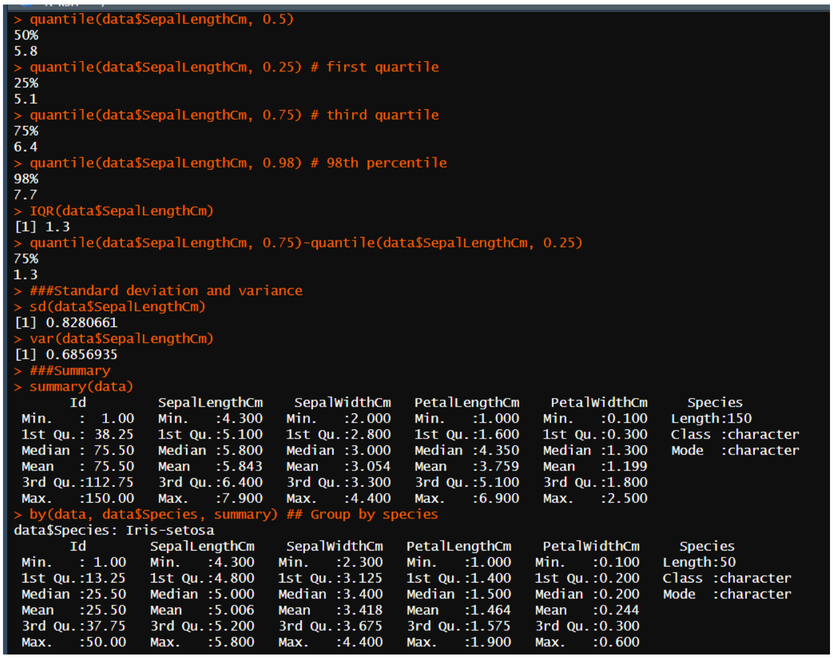
```
data=read.csv(file.choose()) # read dataset
########## Viewing Data ########
View(data)
head(data) # first 6 observations
str(data) # structure of dataset
#### Minimum , maximum , range
min(data$SepalLengthCm)
max(data$SepalLengthCm)
rng2=max(data$SepalLengthCm)-min(data$SepalLengthCm)
rng2
range2 <- function(x)
{
  range <- max(x) - min(x)
  return(range)
}
range2(data$SepalLengthCm)
## Mean,Median, 1st & 3rd quartile,Interquartile range
mean(data$SepalLengthCm)
median(data$SepalLengthCm)
quantile(data$SepalLengthCm, 0.5)
quantile(data$SepalLengthCm, 0.25) # first quartile
quantile(data$SepalLengthCm, 0.75) # third quartile
quantile(data$SepalLengthCm, 0.98) # 98th percentile
IQR(data$SepalLengthCm)
quantile(data$SepalLengthCm, 0.75)-quantile(data$SepalLengthCm, 0.25)
###Standard deviation and variance
sd(data$SepalLengthCm)
var(data$SepalLengthCm)
###Summary
summary(data)
by(data, data$Species, summary) ## Group by species
### Mode
tab <- table(data$SepalLengthCm) #gives the number of occurrences for each unique value
tab
sort(tab, decreasing = TRUE)
sort(table(data$SepalLengthCm), decreasing = TRUE)
sort(table(data$Species), decreasing = TRUE)
summary(data$Species)
##Barplot
barplot(table(data$Species))
barplot(prop.table(table(data$Species)),main="Bar Graph for Iris Species",xlab="Iris Species"
,ylab="Length",col="lightsteelblue1")
## Relative frequency
## Histogram
```

```
hist(data$SepalLengthCm, main= "Histogram for Sepal Length",xlab = "Iris Species", ylab= "Length", col="cadetblue3")

## installed.packages - ggplot2 if not installed

#install.packages("ggplot2")

library(ggplot2)

ggplot(data) + aes(x = SepalLengthCm) + geom_histogram(colour="pink4", fill="rosybrown3")

## Boxplot

boxplot(data$SepalLengthCm)

boxplot(data$SepalLengthCm ~ data$Species, col = "slategray2" ,medcol = "slategrey")

## species wise

## installed.packages - lattice

#install.packages("lattice")

library("lattice")

dotplot(data$SepalLengthCm ~ data$Species)

## Scatterplot

plot(data$SepalLengthCm,data$PetalLengthCm)
```

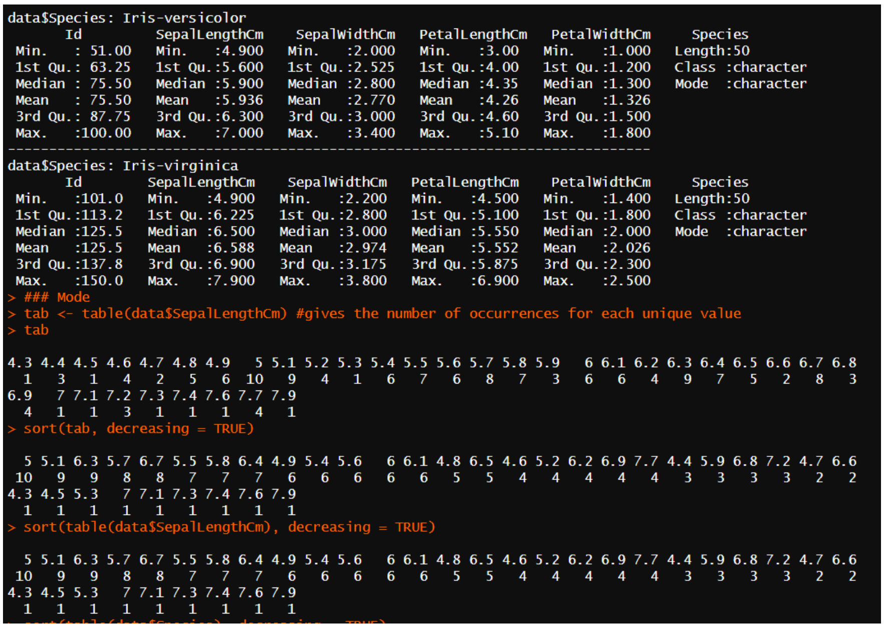
OUTPUT:



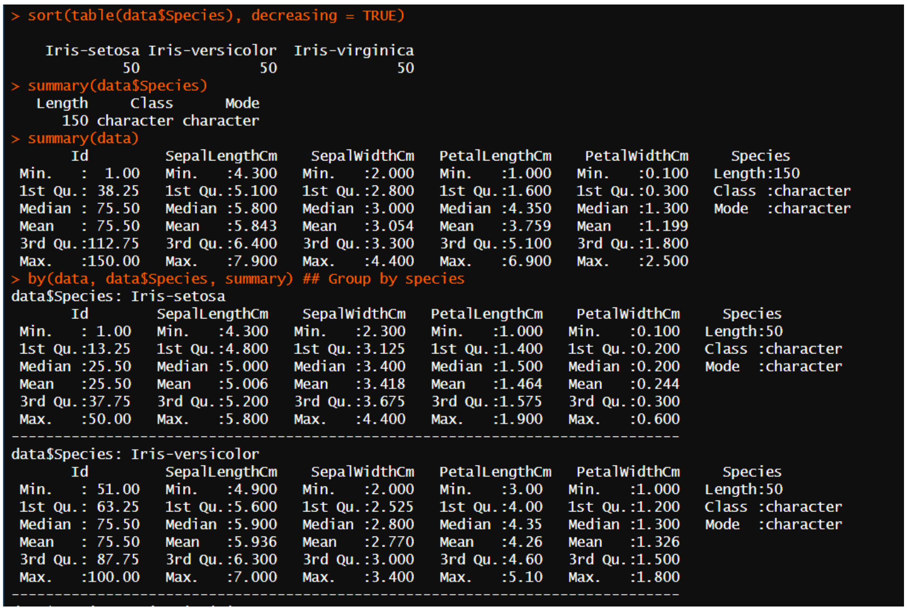
```
> plot(data$SepalLengthCm,data$PetalLengthCm)
> head(data) # first 6 observations
  Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
1  1           5.1          3.5           1.4          0.2 Iris-setosa
2  2           4.9          3.0           1.4          0.2 Iris-setosa
3  3           4.7          3.2           1.3          0.2 Iris-setosa
4  4           4.6          3.1           1.5          0.2 Iris-setosa
5  5           5.0          3.6           1.4          0.2 Iris-setosa
6  6           5.4          3.9           1.7          0.4 Iris-setosa
> str(data) # structure of dataset
'data.frame':   150 obs. of  6 variables:
 $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ SepalLengthCm: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ SepalWidthCm : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ PetalLengthCm: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ PetalWidthCm : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : chr  "Iris-setosa" "Iris-setosa" "Iris-setosa" "Iris-setosa" ...
> #### Minimum , maximum , range
> min(data$SepalLengthCm)
[1] 4.3
> max(data$SepalLengthCm)
[1] 7.9
> rng2=max(data$SepalLengthCm)-min(data$SepalLengthCm)
> rng2
[1] 3.6
> range2 <- function(x)
+ {
+   range <- max(x) - min(x)
+   return(range)
+ }
> range2(data$SepalLengthCm)
[1] 3.6
> ## Mean,Median, 1st & 3rd quartile,Interquartile range
> mean(data$SepalLengthCm)
[1] 5.843333
> median(data$SepalLengthCm)
[1] 5.8
> quantile(data$SepalLengthCm, 0.5)
50%
5.8
```



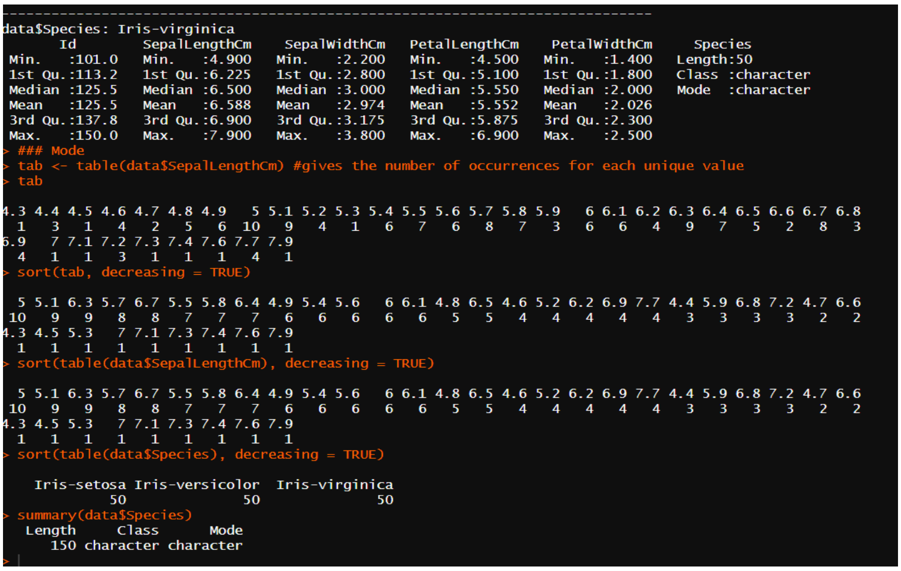
```
> quantile(data$SepalLengthCm, 0.5)
50%
5.8
> quantile(data$SepalLengthCm, 0.25) # first quartile
25%
5.1
> quantile(data$SepalLengthCm, 0.75) # third quartile
75%
6.4
> quantile(data$SepalLengthCm, 0.98) # 98th percentile
98%
7.7
> IQR(data$SepalLengthCm)
[1] 1.3
> quantile(data$SepalLengthCm, 0.75)-quantile(data$SepalLengthCm, 0.25)
75%
1.3
> ###Standard deviation and variance
> sd(data$SepalLengthCm)
[1] 0.8280661
> var(data$SepalLengthCm)
[1] 0.6856935
> ###Summary
> summary(data)
       Id         SepalLengthCm    SepalWidthCm   PetalLengthCm    PetalWidthCm     Species
 Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Length:150
 1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Class :character
 Median : 75.50   Median :5.800   Median :3.000   Median :4.350   Median :1.300   Mode  :character
 Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
 3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> by(data, data$Species, summary) ## Group by species
data$Species: Iris-setosa
       Id        SepalLengthCm    SepalWidthCm   PetalLengthCm    PetalWidthCm     Species
 Min.   : 1.00   Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100   Length:50
 1st Qu.:13.25   1st Qu.:4.800   1st Qu.:3.125   1st Qu.:1.400   1st Qu.:0.200   Class :character
 Median :25.50   Median :5.000   Median :3.400   Median :1.500   Median :0.200   Mode  :character
 Mean   :25.50   Mean   :5.006   Mean   :3.418   Mean   :1.464   Mean   :0.244
 3rd Qu.:37.75   3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
 Max.   :50.00   Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
```



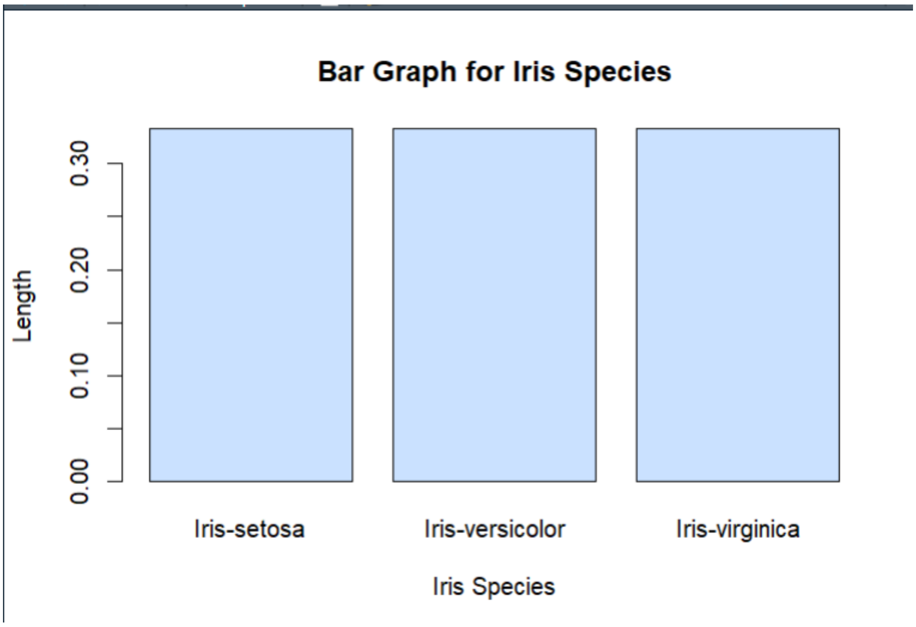
```
data$Species: Iris-versicolor
       Id          SepalLengthCm    SepalWidthCm    PetalLengthCm    PetalWidthCm      Species
 Min.   : 51.00   Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   Length:50
 1st Qu.: 63.25   1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   Class :character
 Median : 75.50   Median :5.900   Median :2.800   Median :4.35   Median :1.300   Mode  :character
 Mean   : 75.50   Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326
 3rd Qu.: 87.75   3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500
 Max.   :100.00   Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800
------------------------------------------------------------------------------------------
data$Species: Iris-virginica
       Id         SepalLengthCm    SepalWidthCm    PetalLengthCm    PetalWidthCm      Species
 Min.   :101.0   Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400   Length:50
 1st Qu.:113.2   1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800   Class :character
 Median :125.5   Median :6.500   Median :3.000   Median :5.550   Median :2.000   Mode  :character
 Mean   :125.5   Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026
 3rd Qu.:137.8   3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300
 Max.   :150.0   Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
> ### Mode
> tab <- table(data$SepalLengthCm) #gives the number of occurrences for each unique value
> tab

4.3 4.4 4.5 4.6 4.7 4.8 4.9   5 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9   6 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8
  1   3   1   4   2   5   6  10   9   4   1   6   7   6   8   7   3   6   6   4   9   7   5   2   8   3
6.9   7 7.1 7.2 7.3 7.4 7.6 7.7 7.9
  4   1   1   3   1   1   1   4   1
> sort(tab, decreasing = TRUE)

  5 5.1 6.3 5.7 6.7 5.5 5.8 6.4 4.9 5.4 5.6   6 6.1 4.8 6.5 4.6 5.2 6.2 6.9 7.7 4.4 5.9 6.8 7.2 4.7 6.6
 10   9   9   8   8   7   7   7   6   6   6   6   6   5   5   4   4   4   4   4   3   3   3   3   2   2
4.3 4.5 5.3   7 7.1 7.3 7.4 7.6 7.9
  1   1   1   1   1   1   1   1   1
> sort(table(data$SepalLengthCm), decreasing = TRUE)

  5 5.1 6.3 5.7 6.7 5.5 5.8 6.4 4.9 5.4 5.6   6 6.1 4.8 6.5 4.6 5.2 6.2 6.9 7.7 4.4 5.9 6.8 7.2 4.7 6.6
 10   9   9   8   8   7   7   7   6   6   6   6   6   5   5   4   4   4   4   4   3   3   3   3   2   2
4.3 4.5 5.3   7 7.1 7.3 7.4 7.6 7.9
  1   1   1   1   1   1   1   1   1
```

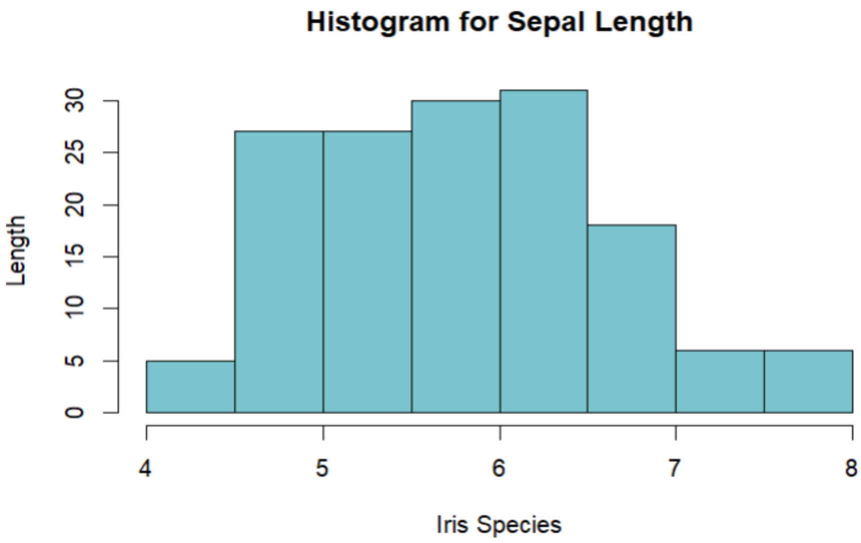


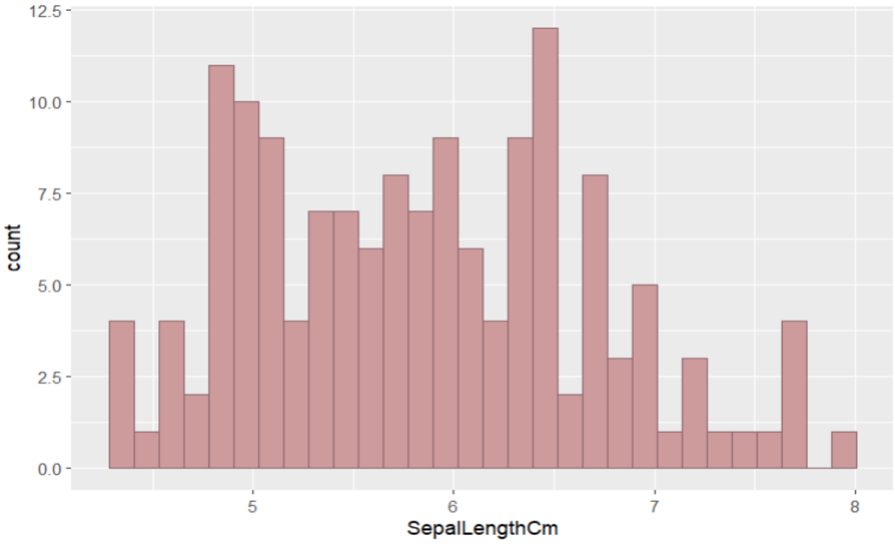
```
> sort(table(data$Species), decreasing = TRUE)

    Iris-setosa Iris-versicolor  Iris-virginica
             50              50              50
> summary(data$Species)
   Length     Class      Mode
      150 character character
> summary(data)
       Id          SepalLengthCm    SepalWidthCm    PetalLengthCm    PetalWidthCm      Species
 Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   Length:150
 1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   Class :character
 Median : 75.50   Median :5.800   Median :3.000   Median :4.350   Median :1.300   Mode  :character
 Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759   Mean   :1.199
 3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> by(data, data$Species, summary) ## Group by species
data$Species: Iris-setosa
       Id         SepalLengthCm    SepalWidthCm    PetalLengthCm    PetalWidthCm      Species
 Min.   : 1.00   Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100   Length:50
 1st Qu.:13.25   1st Qu.:4.800   1st Qu.:3.125   1st Qu.:1.400   1st Qu.:0.200   Class :character
 Median :25.50   Median :5.000   Median :3.400   Median :1.500   Median :0.200   Mode  :character
 Mean   :25.50   Mean   :5.006   Mean   :3.418   Mean   :1.464   Mean   :0.244
 3rd Qu.:37.75   3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
 Max.   :50.00   Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
------------------------------------------------------------------------------------------
data$Species: Iris-versicolor
       Id          SepalLengthCm    SepalWidthCm    PetalLengthCm    PetalWidthCm      Species
 Min.   : 51.00   Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   Length:50
 1st Qu.: 63.25   1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   Class :character
 Median : 75.50   Median :5.900   Median :2.800   Median :4.35   Median :1.300   Mode  :character
 Mean   : 75.50   Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326
 3rd Qu.: 87.75   3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500
 Max.   :100.00   Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800
------------------------------------------------------------------------------------------
```

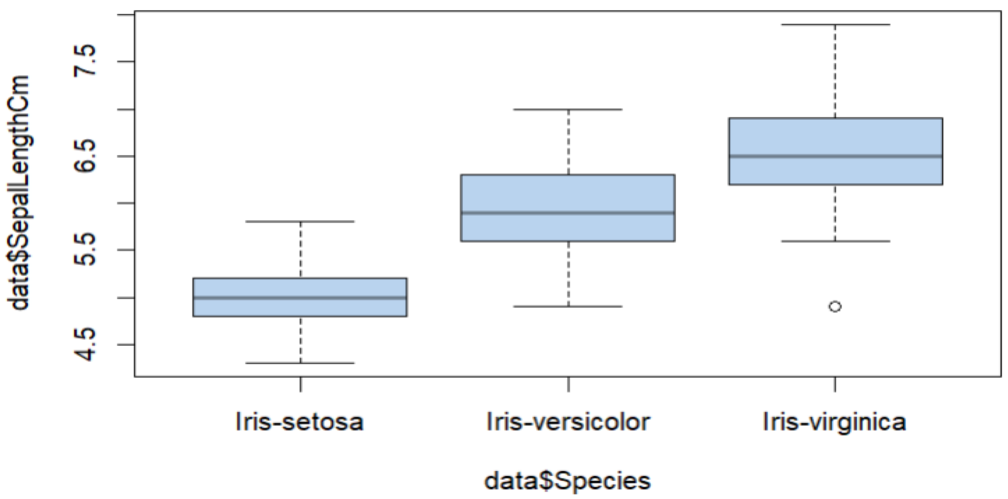


```
------------------------------------------------------------------------------------------
data$Species: Iris-virginica
       Id         SepalLengthCm    SepalWidthCm    PetalLengthCm    PetalWidthCm      Species
 Min.   :101.0   Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400   Length:50
 1st Qu.:113.2   1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800   Class :character
 Median :125.5   Median :6.500   Median :3.000   Median :5.550   Median :2.000   Mode  :character
 Mean   :125.5   Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026
 3rd Qu.:137.8   3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300
 Max.   :150.0   Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
> ### Mode
> tab <- table(data$SepalLengthCm) #gives the number of occurrences for each unique value
> tab

4.3 4.4 4.5 4.6 4.7 4.8 4.9   5 5.1 5.2 5.3 5.4 5.5 5.6 5.7 5.8 5.9   6 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8
  1   3   1   4   2   5   6  10   9   4   1   6   7   6   8   7   3   6   6   4   9   7   5   2   8   3
6.9   7 7.1 7.2 7.3 7.4 7.6 7.7 7.9
  4   1   1   3   1   1   1   4   1
> sort(tab, decreasing = TRUE)

  5 5.1 6.3 5.7 6.7 5.5 5.8 6.4 4.9 5.4 5.6   6 6.1 4.8 6.5 4.6 5.2 6.2 6.9 7.7 4.4 5.9 6.8 7.2 4.7 6.6
 10   9   9   8   8   7   7   7   6   6   6   6   6   5   5   4   4   4   4   4   3   3   3   3   2   2
4.3 4.5 5.3   7 7.1 7.3 7.4 7.6 7.9
  1   1   1   1   1   1   1   1   1
> sort(table(data$SepalLengthCm), decreasing = TRUE)

  5 5.1 6.3 5.7 6.7 5.5 5.8 6.4 4.9 5.4 5.6   6 6.1 4.8 6.5 4.6 5.2 6.2 6.9 7.7 4.4 5.9 6.8 7.2 4.7 6.6
 10   9   9   8   8   7   7   7   6   6   6   6   6   5   5   4   4   4   4   4   3   3   3   3   2   2
4.3 4.5 5.3   7 7.1 7.3 7.4 7.6 7.9
  1   1   1   1   1   1   1   1   1
> sort(table(data$Species), decreasing = TRUE)

    Iris-setosa Iris-versicolor  Iris-virginica
             50              50              50
> summary(data$Species)
   Length     Class      Mode
      150 character character
>
```

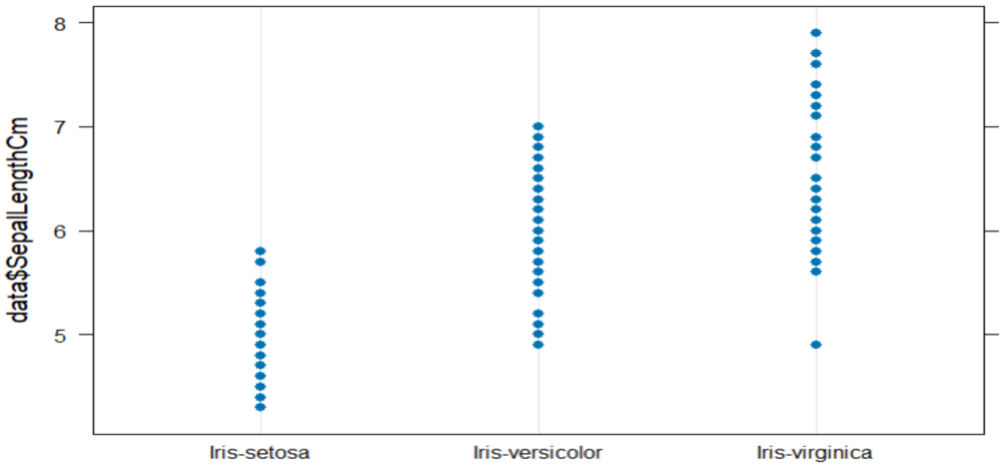

Bar Graph for Iris Species
Length
0.00
0.10
0.20
0.30
Iris-setosa
Iris-versicolor
Iris-virginica
Iris Species


Histogram for Sepal Length
Length
0
5
10
15
20
25
30
4
5
6
7
8
Iris Species

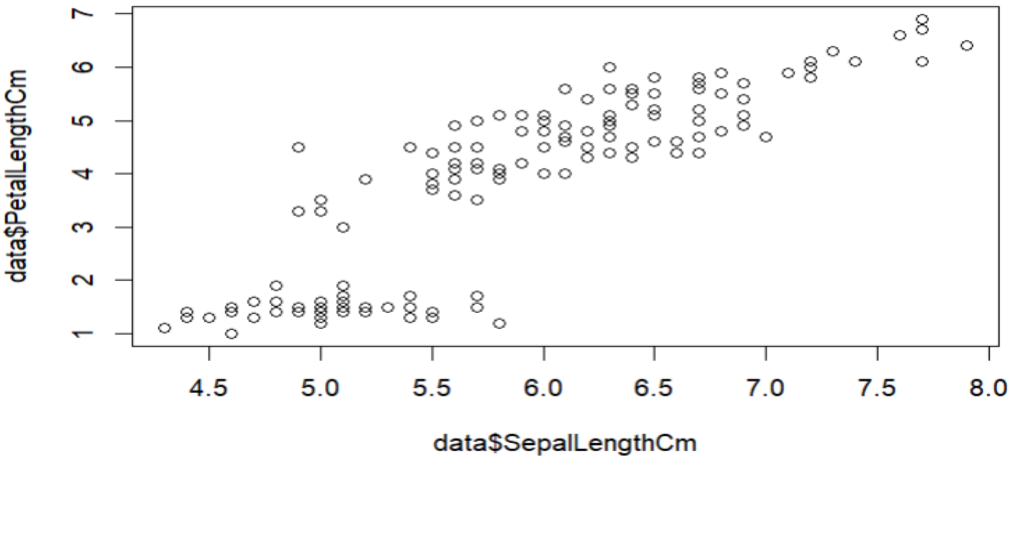

count
0.0
2.5
5.0
7.5
10.0
12.5
5
6
7
8
SepalLengthCm



data$SepalLengthCm
4.5
5.5
6.5
7.5
Iris-setosa
Iris-versicolor
Iris-virginica
data$Species



data$SepalLengthCm
5
6
7
8
Iris-setosa
Iris-versicolor
Iris-virginica



data$PetalLengthCm
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
data$SepalLengthCm

CONCLUSION: *This R code reads a dataset, likely the Iris dataset, and performs various analyses and visualizations to explore its characteristics. It begins by viewing the data's structure and displaying the first few observations. The code then calculates and presents the minimum, maximum, and range of the 'SepalLengthCm' column. It defines a custom function for range calculation as well. Next, it computes and displays the mean, median, quartiles, and interquartile range of the 'SepalLengthCm'. Standard deviation and variance are also calculated and shown. A summary of the dataset is provided, and summary statistics are computed for each species group using the 'by' function. The code determines the mode of 'SepalLengthCm' and generates sorted frequency tables. It creates bar plots representing species occurrences and their relative frequencies. Additionally, histograms and a histogram-like plot using the 'ggplot2' library are generated for 'SepalLengthCm'. Box plots illustrate the distribution of 'SepalLengthCm', both overall and grouped by species. Finally, a scatter plot showcases the relationship between 'SepalLengthCm' and 'PetalLengthCm'.*

In conclusion, This R code efficiently explores and analyzes the Iris dataset using a variety of statistical measures and visualization techniques. It provides insights into the dataset's descriptive statistics, distribution, and relationships, allowing for a comprehensive understanding of the data's characteristics and patterns.