Name: Varun Sudhir

Reg No: 21BDS0040

Exploratory Data Analysis Lab Exp 8

For the purpose of this experiment, we will utilize the popular Iris datatset , an inbuilt dataset available in R

1) Loading the dataset

Code:

```
# Varun Sudhir 21BDS0040

# Load the iris dataset
data(iris)

# View the first few rows of the dataset
head(iris)

# View the structure of the dataset
str(iris)

# View summary statistics for the dataset
summary(iris)
```

Output:

```
>
> # Varun Sudhir 21BDS0040
>
> # Load the iris dataset
> data(iris)
> # View the first few rows of the dataset
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> # View the structure of the dataset
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> # View summary statistics for the dataset
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

2) Checking for missing values

Code:

```
# Varun Sudhir 21BDS0040
# Check for missing values in the dataset
any(is.na(iris))
# Count the number of missing values per column
colSums(is.na(iris))
```

Output:

```
>
> # Varun Sudhir 21BDS0040
>
> # Check for missing values in the dataset
> any(is.na(iris))
[1] FALSE
> # Count the number of missing values per column
> colSums(is.na(iris))
Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
           0            0            0            0            0
>
```

3) Descriptive statistics

Code:

```
# Varun Sudhir 21BDS0040
# Summary statistics for numerical columns
summary(iris)
# Mean, variance, and standard deviation for Sepal.Length
mean(iris$Sepal.Length)
var(iris$Sepal.Length)
sd(iris$Sepal.Length)
```

Output:

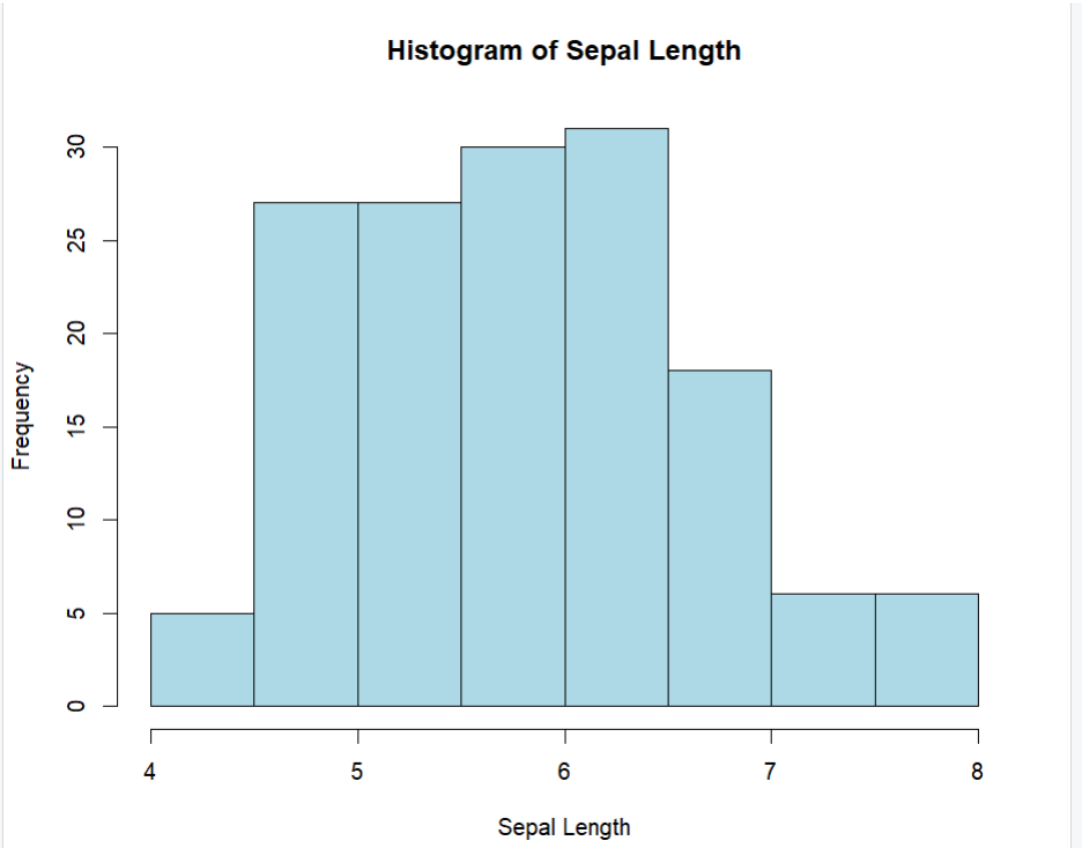
```
>
> # Varun Sudhir 21BDS0040
> # Summary statistics for numerical columns
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> # Mean, variance, and standard deviation for Sepal.Length
> mean(iris$Sepal.Length)
[1] 5.843333
> var(iris$Sepal.Length)
[1] 0.6856935
> sd(iris$Sepal.Length)
[1] 0.8280661
```

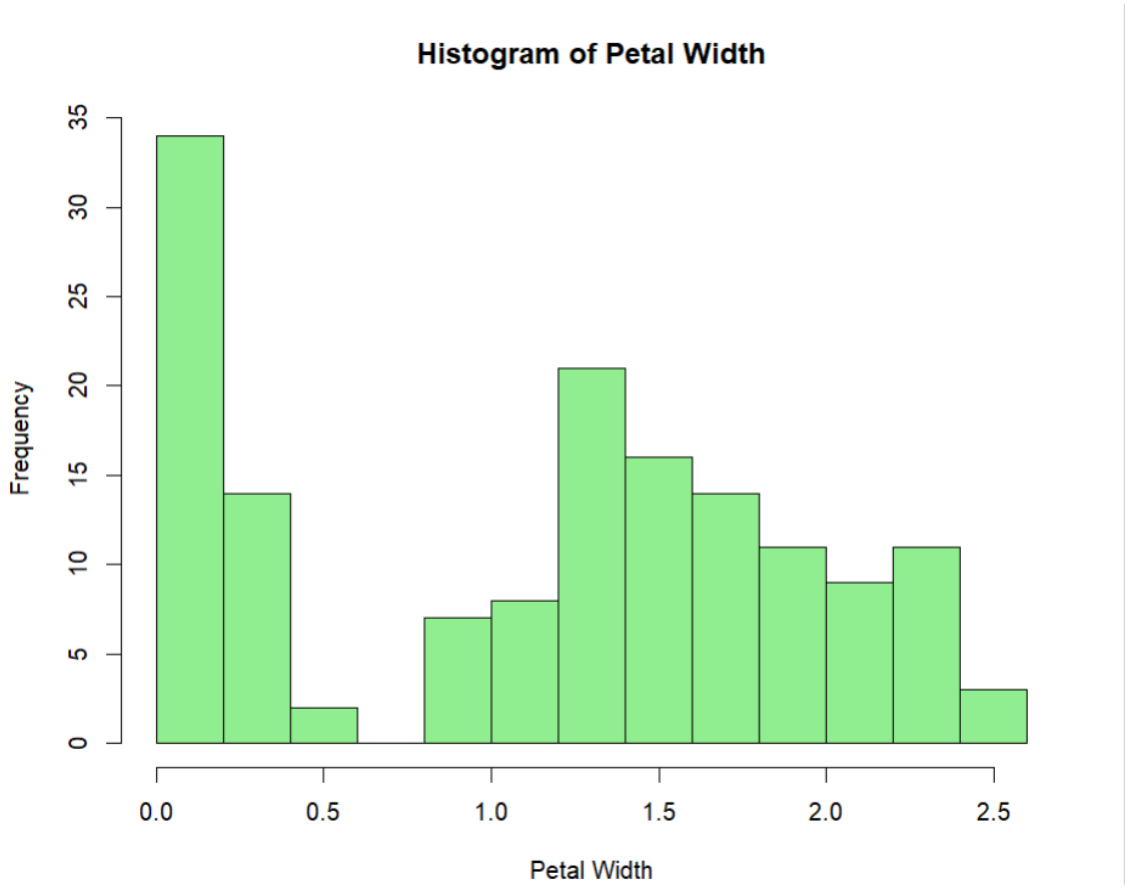
4) Visualizing Distributions

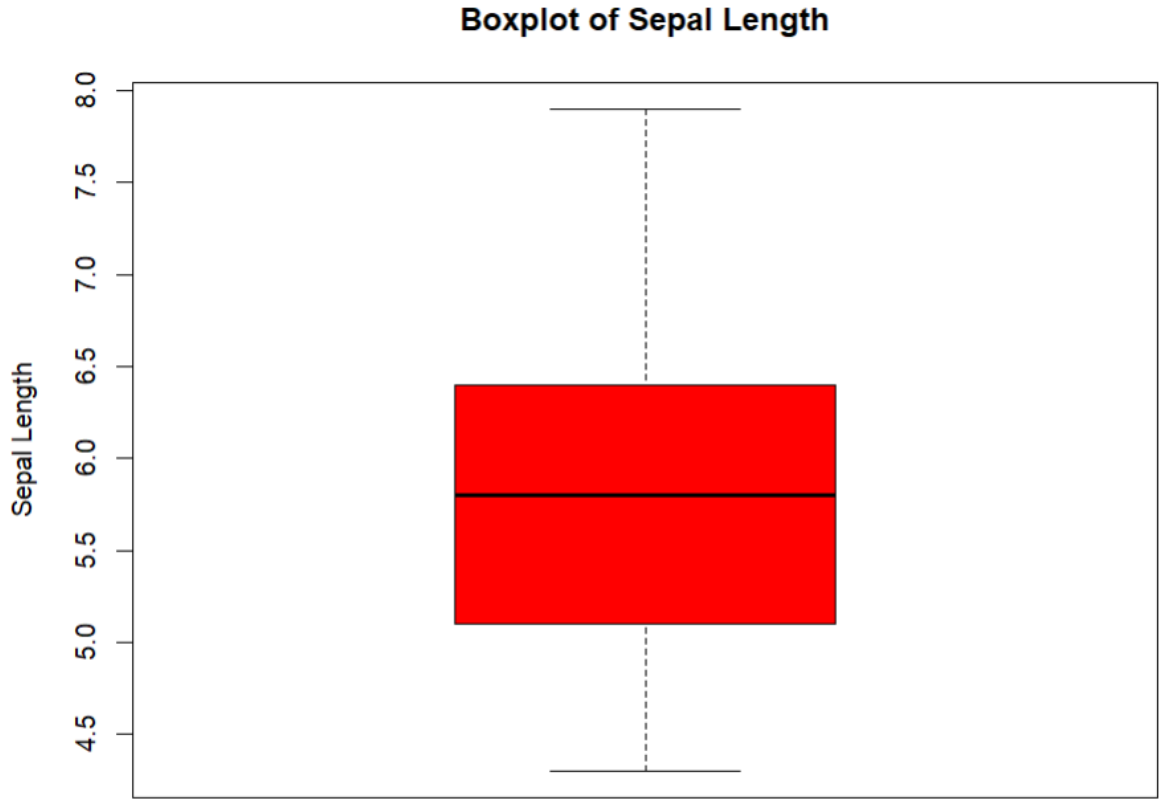
Code:

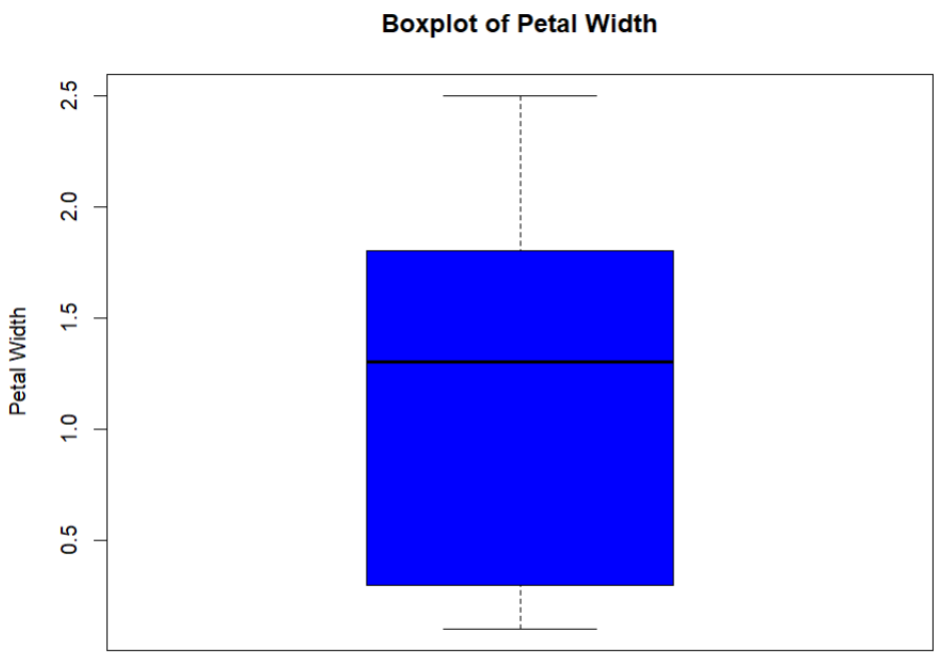
```
# Varun Sudhir 21BDS0040

# Histogram for Sepal.Length
hist(iris$Sepal.Length, main="Histogram of Sepal Length", xlab="Sepal
Length", col="lightblue", border="black")

# Histogram for Petal.Width
hist(iris$Petal.Width, main="Histogram of Petal Width", xlab="Petal Width",
col="lightgreen", border="black")


# Boxplot for Sepal Length
boxplot(iris$Sepal.Length, main="Boxplot of Sepal Length", ylab="Sepal
Length", col="red")

# Boxplot for Petal Width
boxplot(iris$Petal.Width, main="Boxplot of Petal Width", ylab="Petal
Width", col="blue")
```

Output:




Histogram of Petal Width
Frequency
0
5
10
15
20
25
30
35
0.0
0.5
1.0
1.5
2.0
2.5
Petal Width


Boxplot of Sepal Length
Sepal Length
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

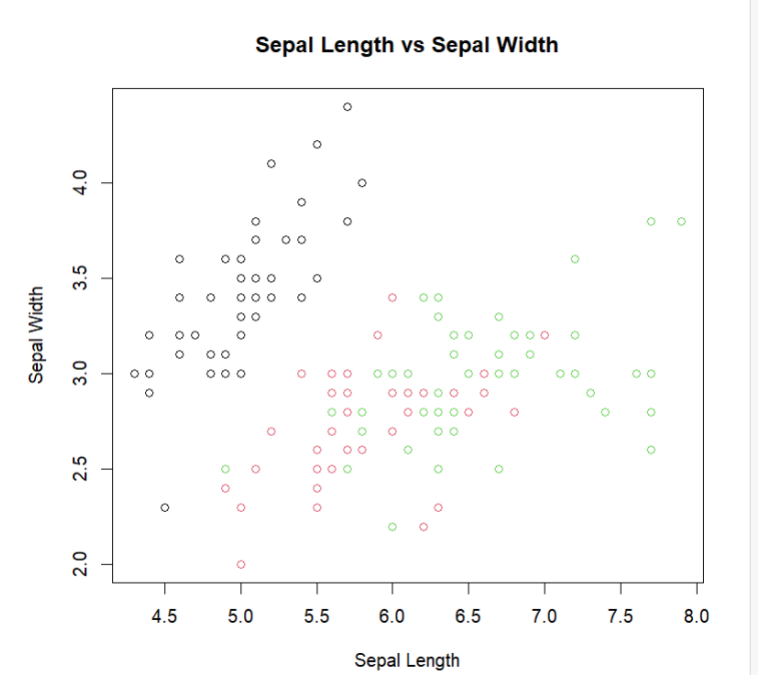


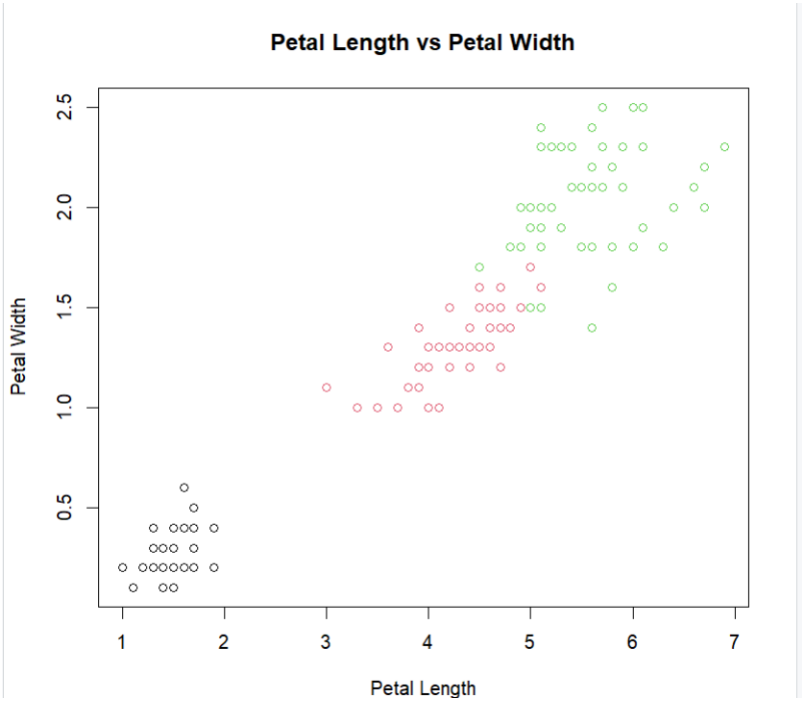
5) Relationships Between Variables

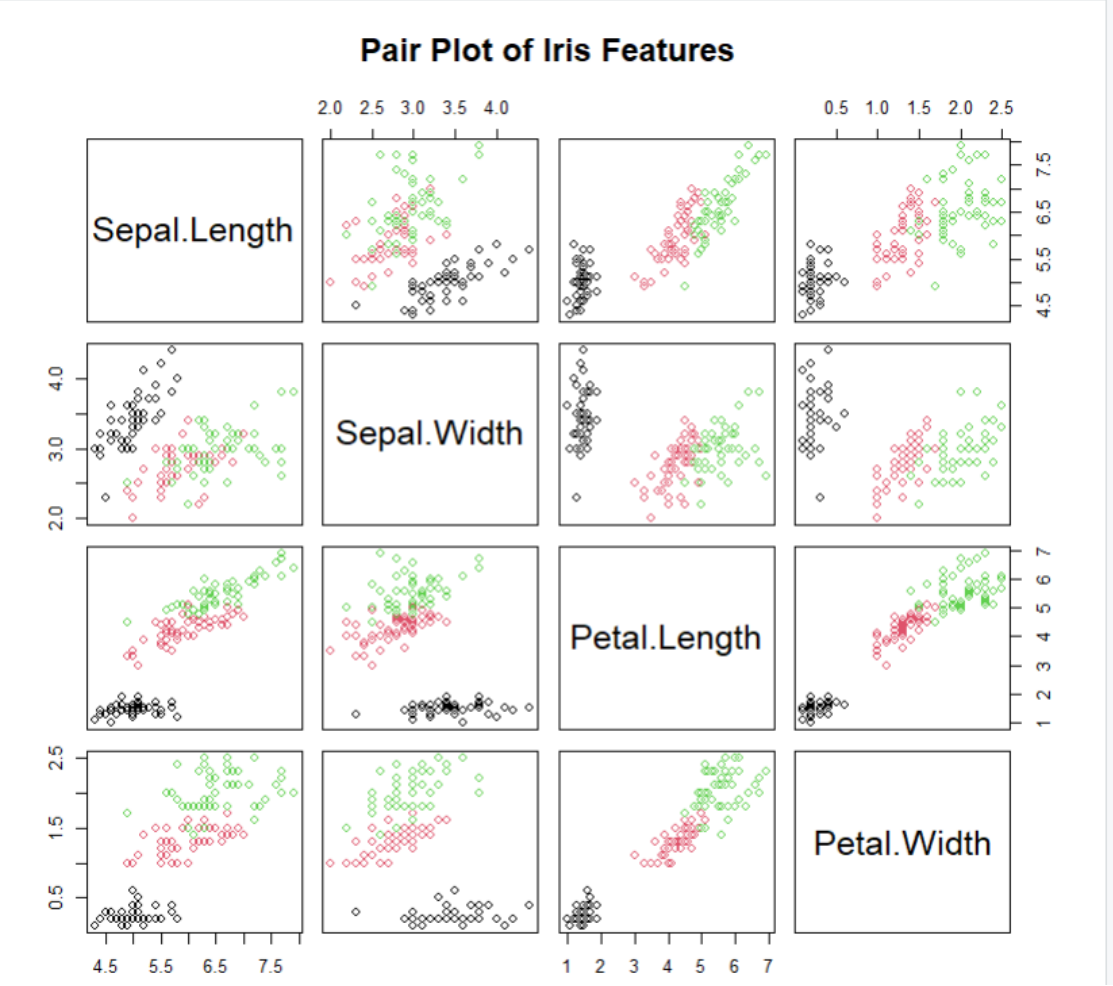
Code:

```
# Varun Sudhir 21BDS0040

# Scatter plot between Sepal Length and Sepal Width
plot(iris$Sepal.Length, iris$Sepal.Width, main="Sepal Length vs Sepal
Width",
     xlab="Sepal Length", ylab="Sepal Width", col=iris$Species)

# Scatter plot between Petal Length and Petal Width
plot(iris$Petal.Length, iris$Petal.Width, main="Petal Length vs Petal
Width",
     xlab="Petal Length", ylab="Petal Width", col=iris$Species)


# Pair plot for all variables
pairs(iris[, 1:4], main="Pair Plot of Iris Features", col=iris$Species)
```

Output:


Sepal Length vs Sepal Width
Sepal Width
Sepal Length
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


Petal Length vs Petal Width
Petal Width
Petal Length
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7



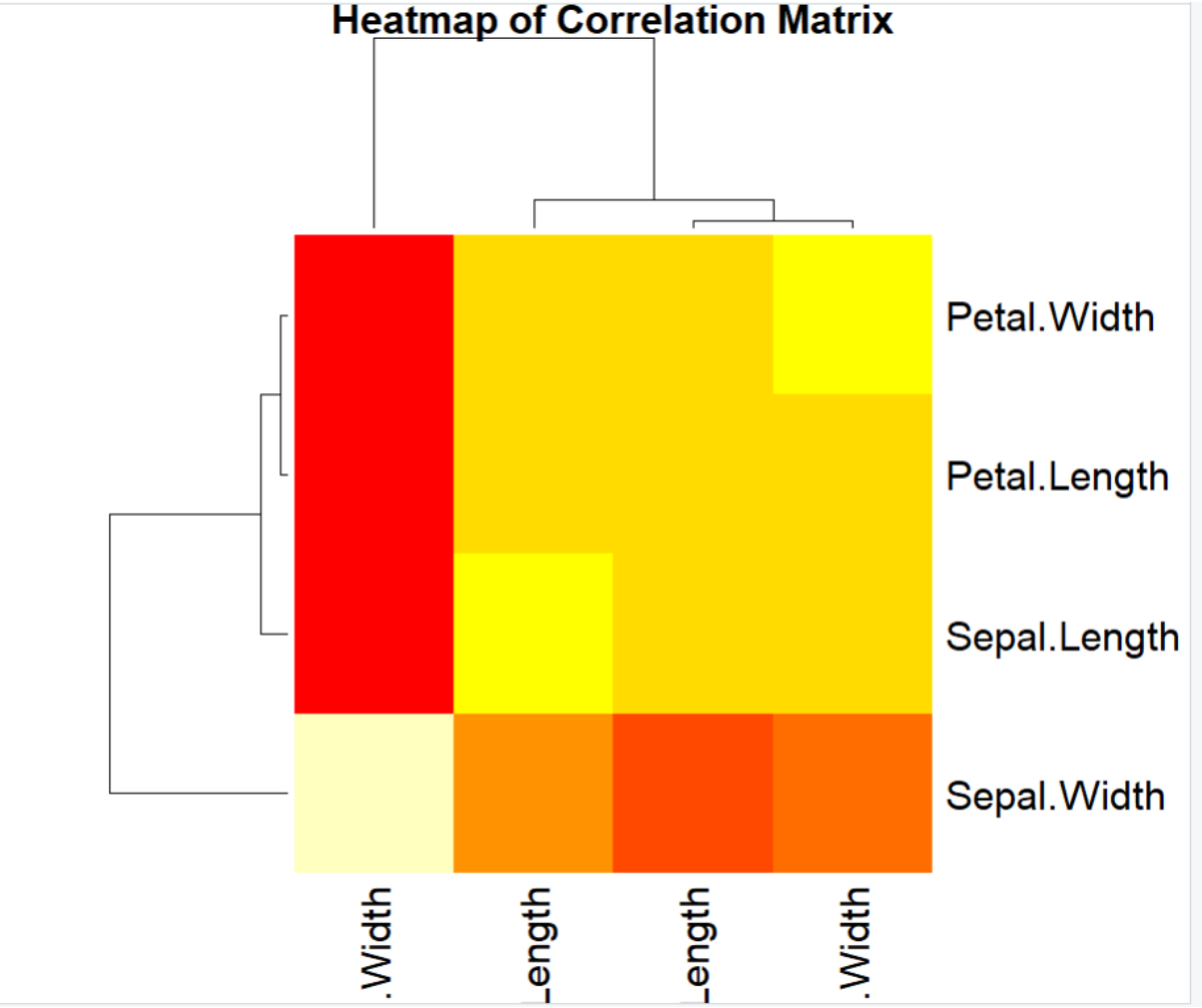
6) Correlation Matrix

Code:

```
# Varun Sudhir 21BDS0040

# Correlation matrix for numerical features
cor_matrix <- cor(iris[, 1:4])
cor_matrix

# Visualize correlation matrix using heatmap
heatmap(cor_matrix, main="Heatmap of Correlation Matrix",
col=heat.colors(10))
```

Output:

```
> # Varun Sudhir 21BDS0040
>
> # Correlation matrix for numerical features
> cor_matrix <- cor(iris[, 1:4])
> cor_matrix
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```



7) Outliers Detection

Code:

```
# Varun Sudhir 21BDS0040

# Outlier detection using IQR for Sepal Length
Q1 <- quantile(iris$Sepal.Length, 0.25)
Q3 <- quantile(iris$Sepal.Length, 0.75)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Check for outliers
outliers <- iris$Sepal.Length[iris$Sepal.Length < lower_bound |
iris$Sepal.Length > upper_bound]
outliers
```

Output:

```
> # Outlier detection using IQR for Sepal Length
> Q1 <- quantile(iris$Sepal.Length, 0.25)
> Q3 <- quantile(iris$Sepal.Length, 0.75)
> IQR <- Q3 - Q1
> lower_bound <- Q1 - 1.5 * IQR
> upper_bound <- Q3 + 1.5 * IQR
> # Check for outliers
> outliers <- iris$Sepal.Length[iris$Sepal.Length < lower_bound | iris$Sepal.Length > upper_bound]
> outliers
numeric(0)
>
```

8) Multivariate Analysis

Code:

```
# Varun Sudhir 21BDS0040

# Performing PCA
pca_model <- prcomp(iris[, 1:4], scale = TRUE)

# View summary of PCA
summary(pca_model)

# Biplot of the first two principal components
biplot(pca_model, main="PCA Biplot")
```

Output:

```
> # Performing PCA
> pca_model <- prcomp(iris[, 1:4], scale = TRUE)
> # View summary of PCA
> summary(pca_model)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```