UNIVERSITY INSTITUTE of COMPUTING
Asia's Fastest Growing University

NAAC GRADE A+
Accredited University

R LAB PROJECT

Performed by –
Anshita Goel (UID – 24MCI10012)
Ashi Mittal (UID – 24MCI10015)

Branch: MCA(AIML)
Section/Group: 24MAM 1-A
Semester: 1
Date of Performance: 26-10-24
Subject Name: R LAB
Subject Code: 24CAP-614
Github: https://github.com/anshita2003/R-Project.git

1. **Aim/Overview of the practical:** Choose a dataset from a repository like Kaggle or UCI Machine Learning Repository and perform exploratory data analysis using R. Explore the distribution of variables, identify outliers, and visualize relationships between variables using plots like histograms, scatter plots, and boxplots.

2. **Task to be done:**

Task 1: Load the Dataset
- Step 1.1: Load the Iris Dataset
- Step 1.2: View the First Few Rows
- Step 1.3: Check the Structure of the Data

Task 2: Summarize the Dataset
- Step 2.1: Generate Summary Statistics
- Step 2.2: Generate Structure (STR) Statistics

Task 3: Visualizing the Distribution of Variables
- Step 3.1: Create Histograms

Task 4: Identifying Outliers
- Step 4.1: Create Boxplot

Task 5: Analyzing Relationships Between Variables
- Step 5.1: Scatter Plots

3. Steps/Commands involved to perform project:

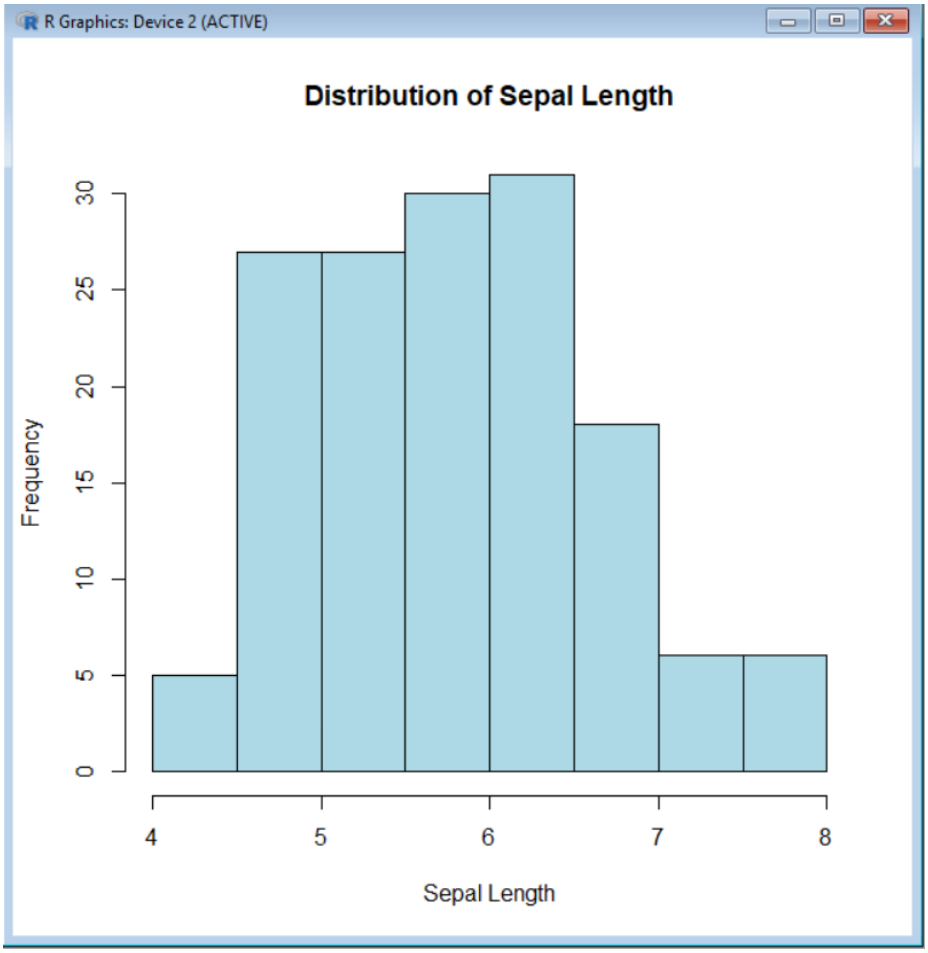
Loading the dataset IRIS and performing EDA

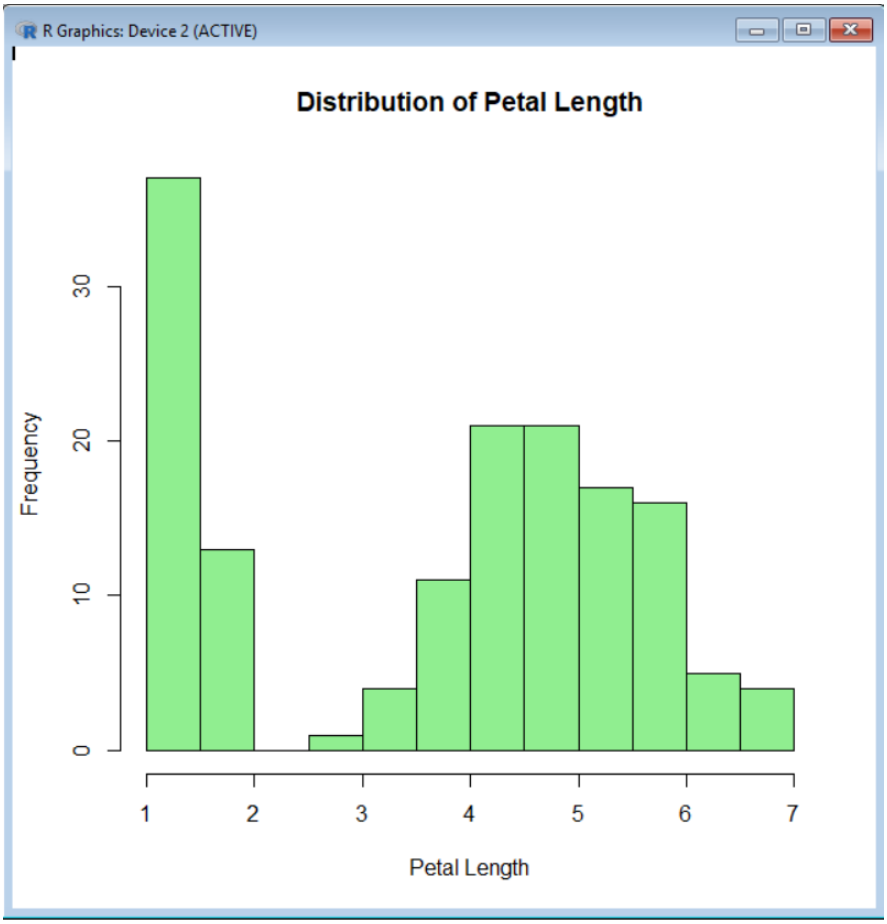
```
> # EDA(exploratory data analysis)
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
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
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 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
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

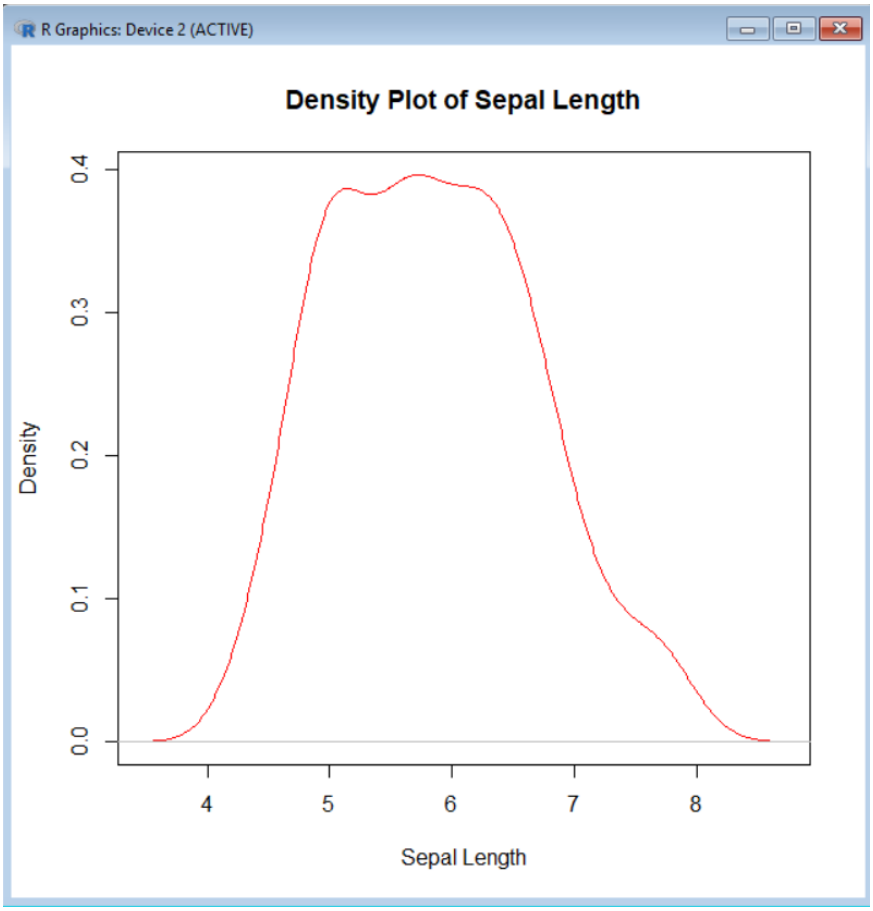
```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Visualizing the distribution of variables using HISTOGRAM:

```
> #distribution of dataset using histogram,scatterplot,boxplot
> hist(iris$Sepal.Length,
+ main = "Distribution of Sepal Length",
+ xlab = "Sepal Length",
+ col = "lightblue"
+ )
> hist(iris$Petal.Length,
+ main = "Distribution of Petal Length",
+ xlab = "Petal Length",
+ col = "lightgreen"
+ )
> #Density plot
> plot(density(iris$Sepal.Length),
+ main = "Density Plot of Sepal Length",
+ xlab = "Sepal Length",
+ col = "red"
+ )
```

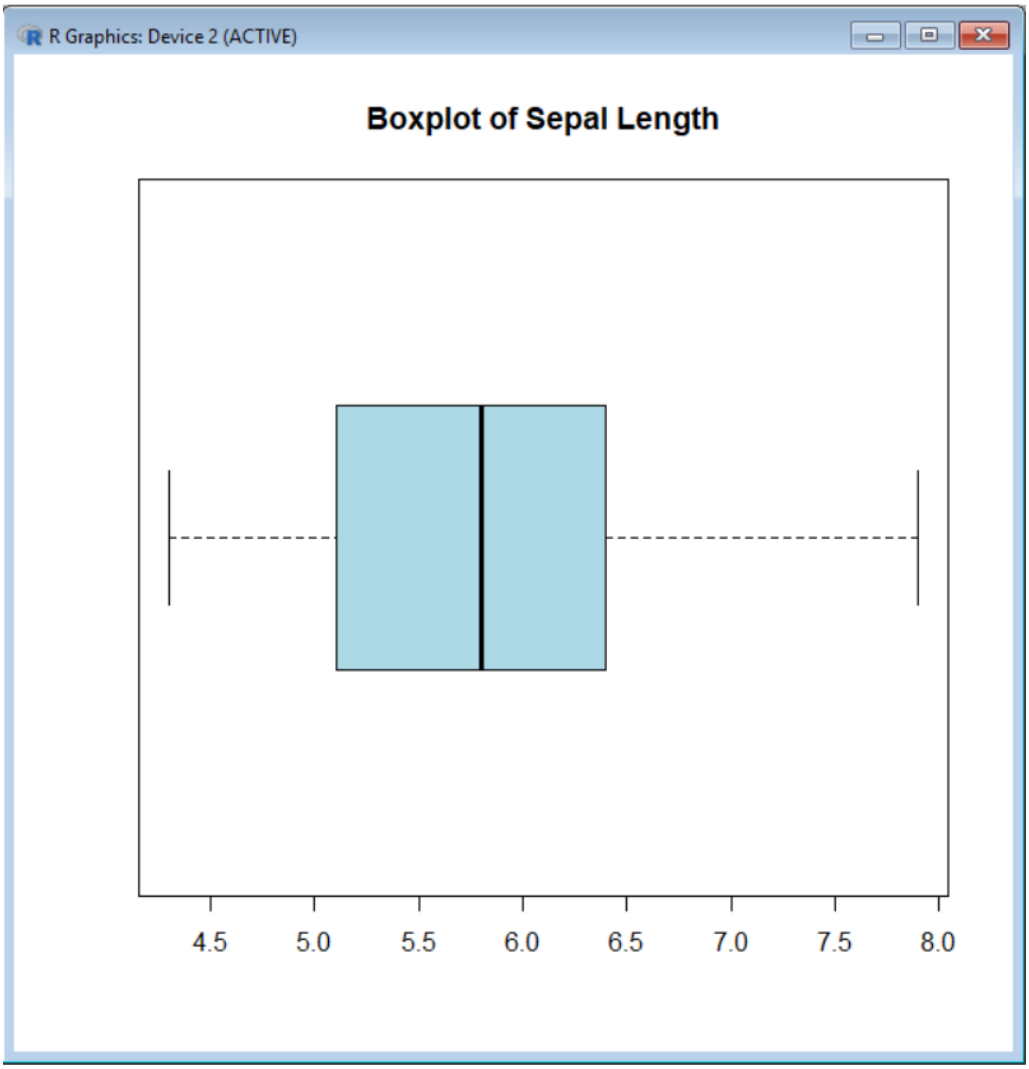


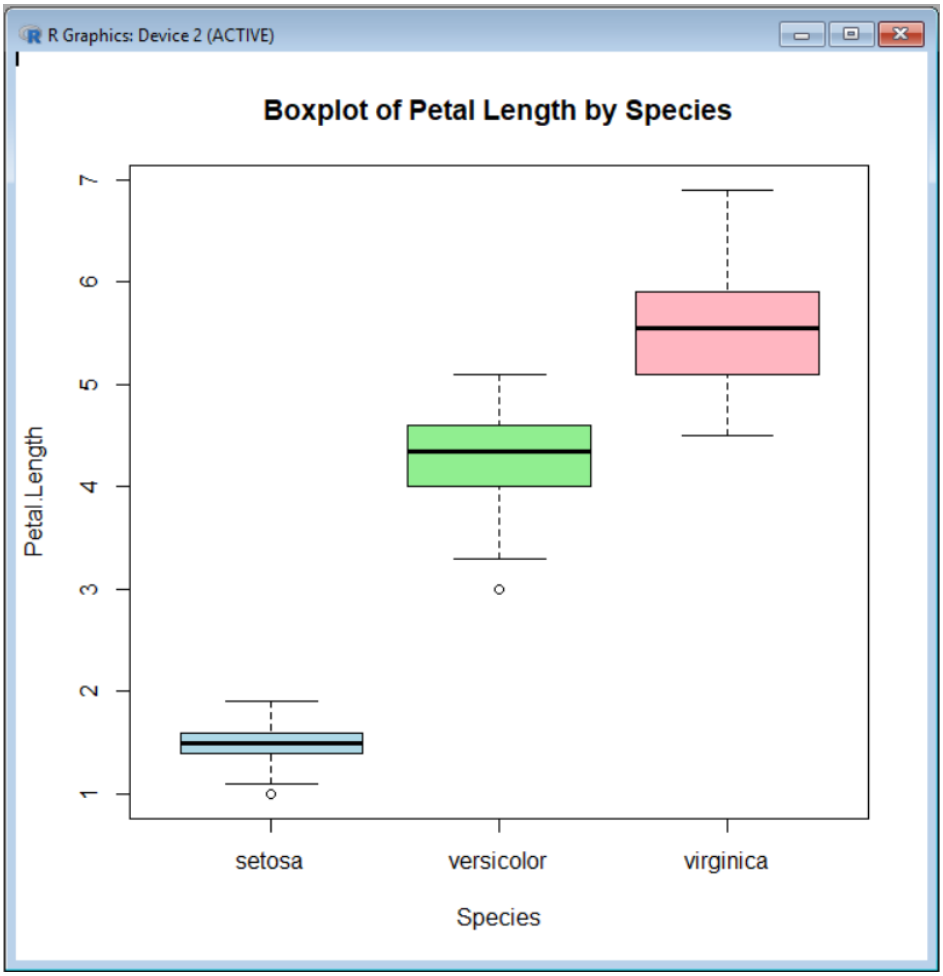

R Graphics: Device 2 (ACTIVE)
Distribution of Petal Length
Frequency
0
10
20
30
1
2
3
4
5
6
7
Petal Length


R Graphics: Device 2 (ACTIVE)
Density Plot of Sepal Length
Density
0.0
0.1
0.2
0.3
0.4
4
5
6
7
8
Sepal Length

Identify the Outliers using BOXPLOT:

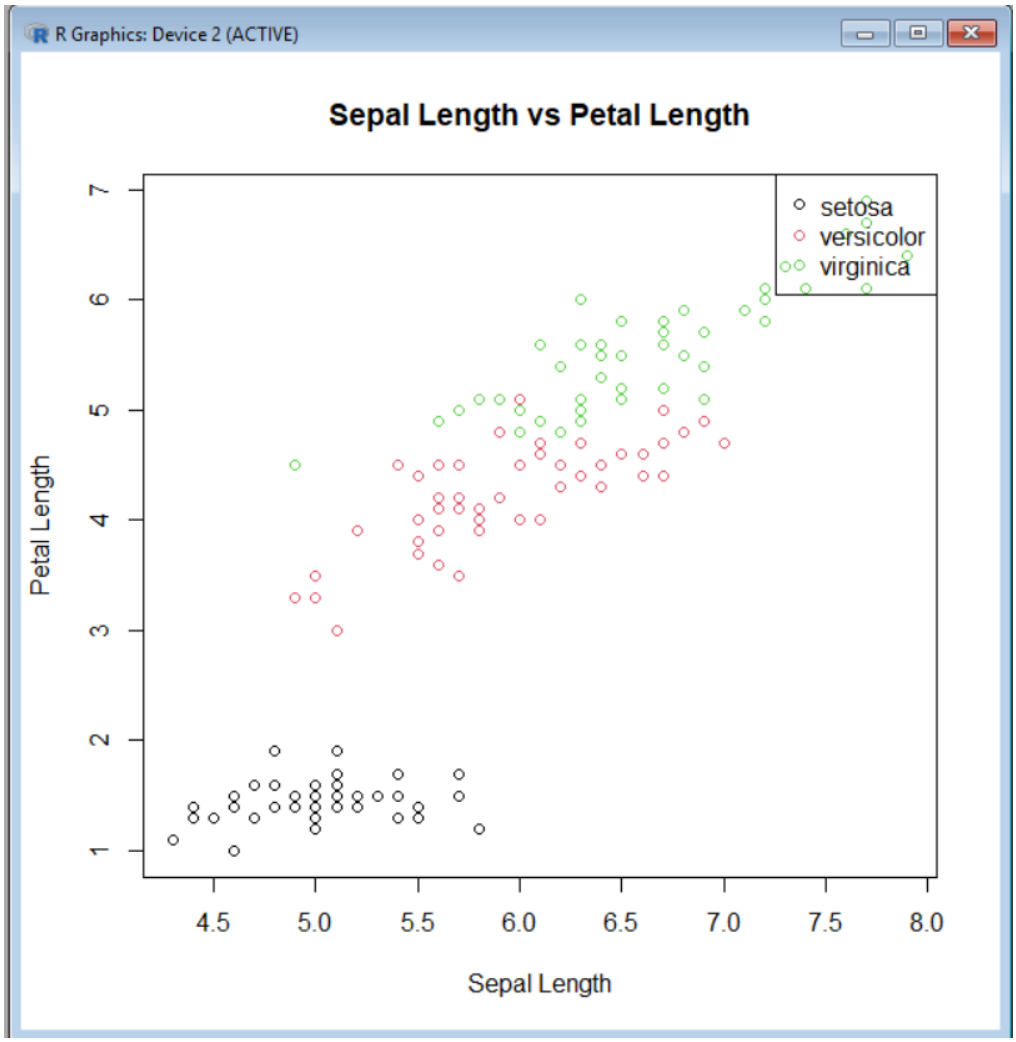
```
> #Identify Outliers
> boxplot(iris$Sepal.Length,
+ main = "Boxplot of Sepal Length",
+ col = "lightblue",
+ horizontal = TRUE
+ )
> boxplot(Petal.Length ~ Species,
+ data = iris,
+ main = "Boxplot of Petal Length by Species",
+ col = c("lightblue", "lightgreen", "lightpink")
+ )
```





Relationship between variables using SCATTER PLOT:

```
> #relationship between variables
> #scatter plot
> plot(iris$Sepal.Length,
+ iris$Petal.Length,
+ main = "Sepal Length vs Petal Length",
+ xlab = "Sepal Length",
+ ylab = "Petal Length",
+ col = iris$Species
+ )
> legend("topright", legend = levels(iris$Species), col = 1:3, pch = 1)
>
```



4. Learning Outcomes (What I have learnt) :

- Performed exploratory data analysis (EDA) on the Iris dataset or alternatively downloaded a dataset from an online source if you prefer not to use the pre-installed Iris dataset.
- Visualizing the distribution of numerical variables using histograms and density plots.
- Identifying outliers with boxplots.
- Exploring relationships between variables using scatter plots.