Laboratory Practice I

Data Analytics

Shiva Saran
BE-B 20
2020-2021

Practical 1

Download the Iris flower dataset or any other dataset into a DataFrame. (eg https://archive.ics.uci.edu/ml/datasets/Iris)

Use Python/R and Perform following –

How many features are there and what are their types (e.g., numeric, nominal)?

```
> View(iris)
> dim(iris)
[1] 150   5

> #.............1. data set details...............
> #internal structure

> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"

> 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 ...
```

Compute and display summary statistics for each feature available in the dataset (eg. minimum value, maximum value, mean, range, standard deviation, variance and percentiles)

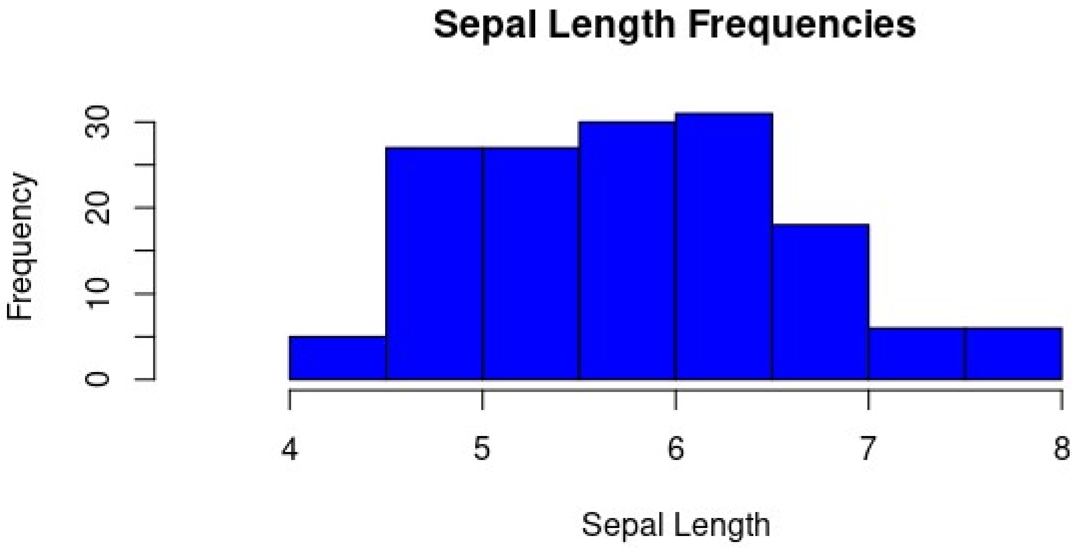
```
> #.............2.statistics..................
> #min value
> min(iris$Sepal.Length)
[1] 4.3
>
> #max value
> max(iris$Sepal.Length)
[1] 7.9

> #range
> range(iris$Sepal.Length)
[1] 4.3 7.9
```

```
>
> #standard deviation
> sd(iris$Sepal.Length)
[1] 0.8280661
>
> #variance
> var(iris$Sepal.Length)
[1] 0.6856935

> #percentile
> quantile(iris$Sepal.Length)
   0%   25%   50%   75% 100%
  4.3   5.1   5.8   6.4   7.9
>
> #For specific
> quantile(iris$Sepal.Length, c(0.35,0.75))
35% 75%
5.5 6.4
```

Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions. Plot each histogram.

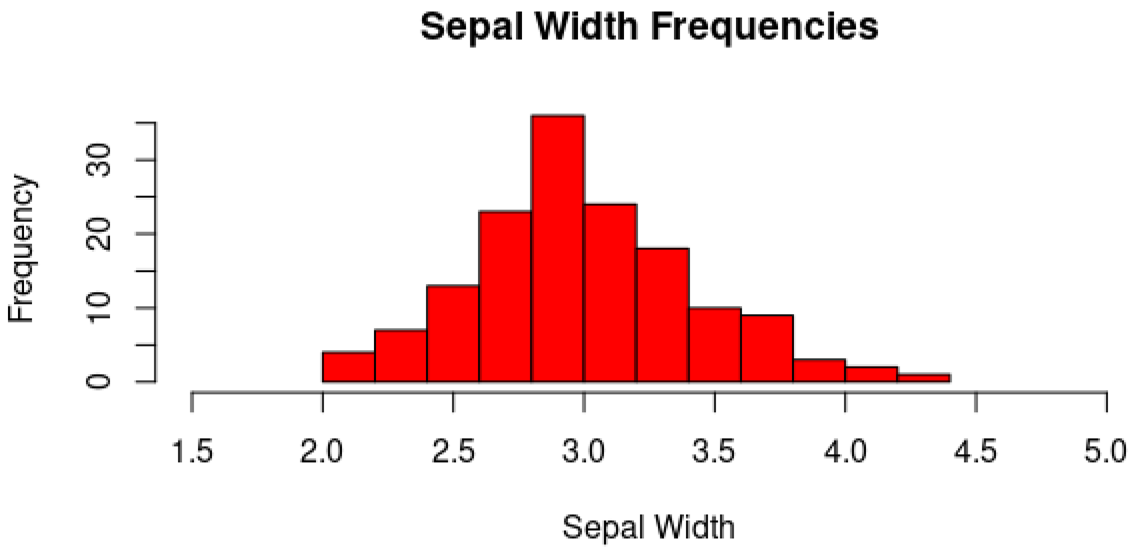
1. Sepal Length

```
> hist(iris$Sepal.Length, main = "Sepal Length Frequencies", xlab ="Sepal Length", xlim = c(3.5,8.5),
col="blue")
```



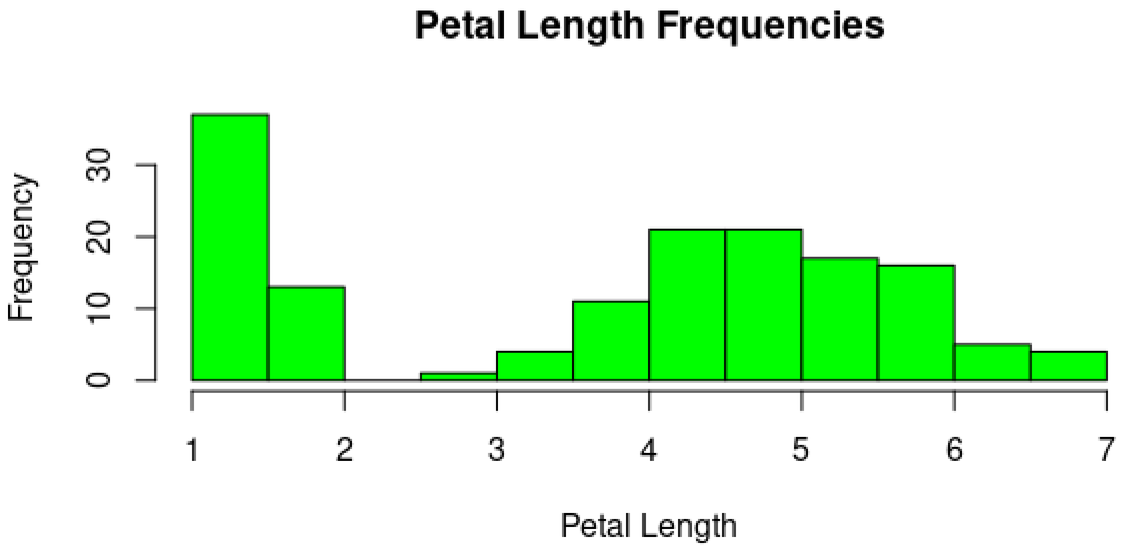
2. Sepal Width

```
> hist(iris$Sepal.Width, main = "Sepal Width Frequencies", xlab ="Sepal Width", xlim = c(1.5,5),
col="Red")
```



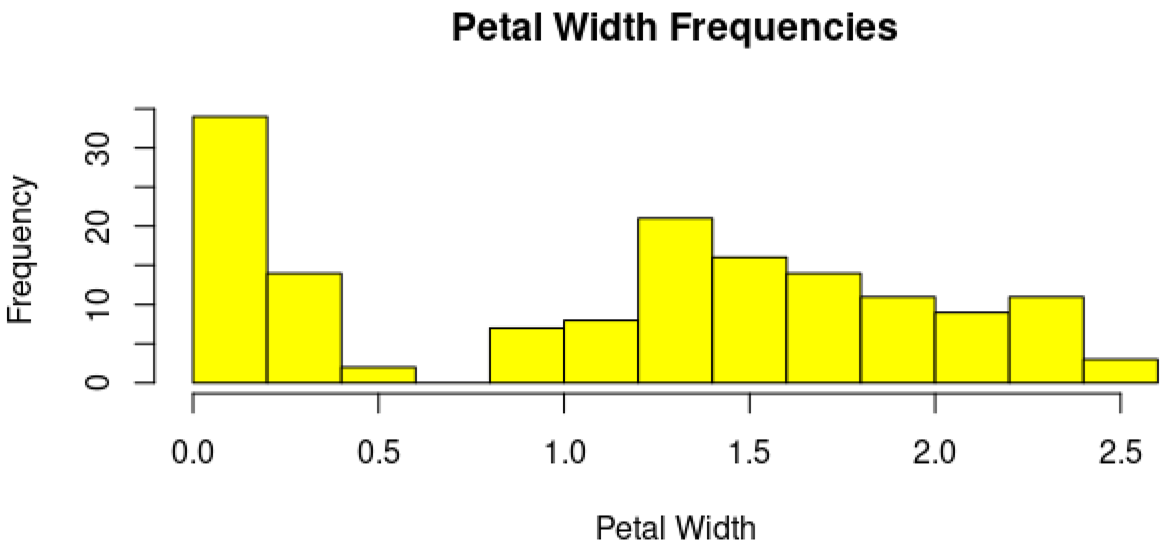
3. Petal Length

```
> hist(iris$Petal.Length, main = "Petal Length Frequencies", xlab ="Petal Length", col="Green")
```



4. Petal Width

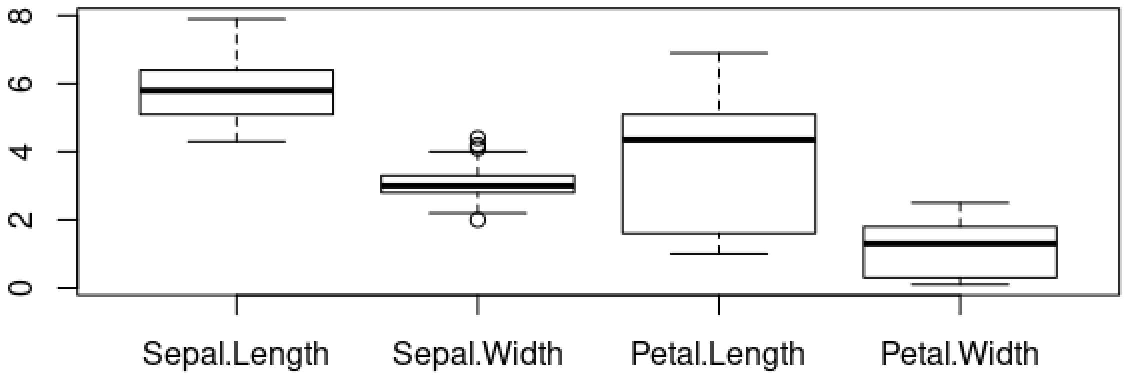
```
> hist(iris$Petal.Width, main = "Petal Width Frequencies", xlab ="Petal Width", col="Yellow")
```



Create a boxplot for each feature in the dataset. All of the boxplots should be combined into a single plot. Compare distributions and identify outliers.

```
> myboxplot<-boxplot(iris[,-5])
> myboxplot
$stats
      [,1] [,2] [,3] [,4]
[1,]  4.3  2.2 1.00  0.1
[2,]  5.1  2.8 1.60  0.3
[3,]  5.8  3.0 4.35  1.3
[4,]  6.4  3.3 5.10  1.8
[5,]  7.9  4.0 6.90  2.5

$n
[1] 150 150 150 150

$conf
           [,1]     [,2]     [,3]    [,4]
[1,] 5.632292 2.935497 3.898477 1.10649
[2,] 5.967708 3.064503 4.801523 1.49351

$out
[1] 4.4 4.1 4.2 2.0

$group
[1] 2 2 2 2
```

```
$names
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"

> myboxplot$out
[1] 4.4 4.1 4.2 2.0
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width