Exploring - IRIS Data Set

Someshwar Rao Sattiraju

March 18, 2017

Exploring IRIS Data Set- Performing descriptive statistics

We are intrested in knowing the essential descriptive statistics of the IRIS Data set to understand how to differentiate one species from the other.

So we are firstly computing the dimentions of the data set and we are doing the summary statistics which will give us the mean and quartiles data for Length & Width of Petals and Sepal

```
dataset<-read.csv("S:/R/Iris/Iris.csv")
print(dim(summary))
```

```
## NULL
```

```
print(summary(dataset))
```

```
##        Id         SepalLengthCm    SepalWidthCm   PetalLengthCm  
##  Min.   :  1.00   Min.   :4.300   Min.   :2.000   Min.   :1.000  
##  1st Qu.: 38.25   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600  
##  Median : 75.50   Median :5.800   Median :3.000   Median :4.350  
##  Mean   : 75.50   Mean   :5.843   Mean   :3.054   Mean   :3.759  
##  3rd Qu.:112.75   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100  
##  Max.   :150.00   Max.   :7.900   Max.   :4.400   Max.   :6.900  
##   PetalWidthCm              Species  
##  Min.   :0.100   Iris-setosa    :50  
##  1st Qu.:0.300   Iris-versicolor:50  
##  Median :1.300   Iris-virginica :50  
##  Mean   :1.199                       
##  3rd Qu.:1.800                       
##  Max.   :2.500                       
```

```
print(names(dataset))
```

```
## [1] "Id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm"
## [5] "PetalWidthCm"  "Species"
```

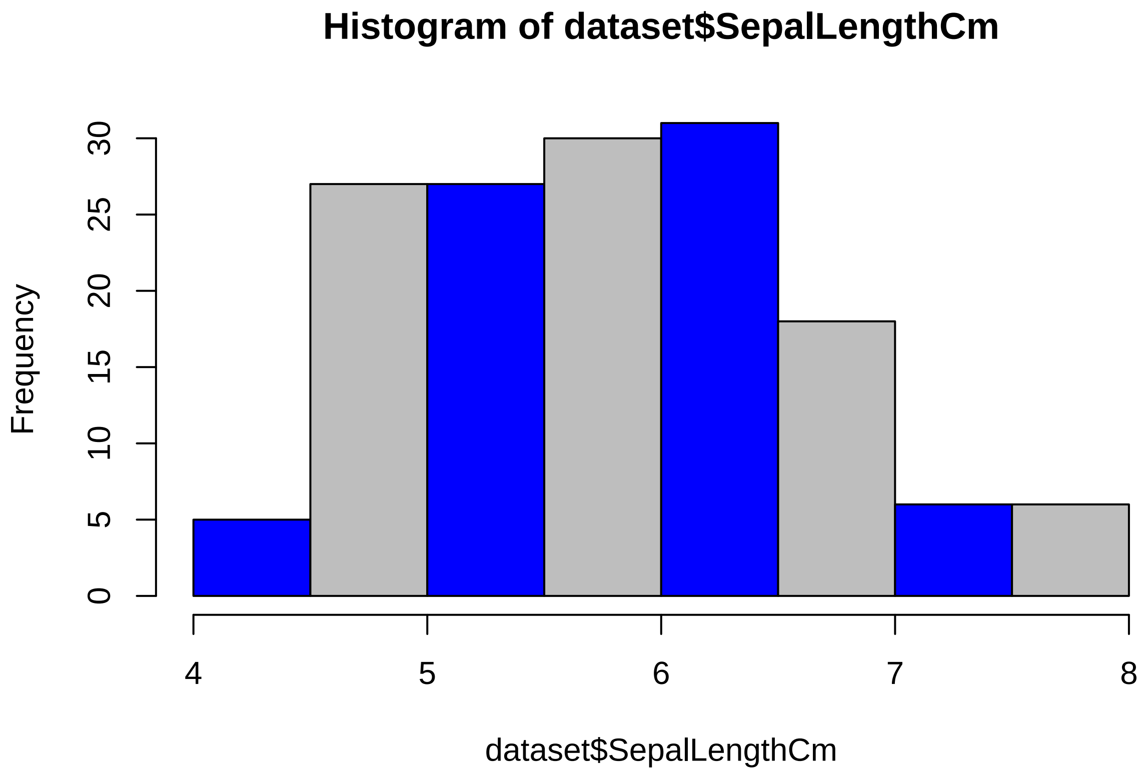
Here we are understanding the attriutes or column names in the data set.

```
print(attributes(dataset))
```

```
## $names
## [1] "Id"            "SepalLengthCm" "SepalWidthCm"  "PetalLengthCm"
## [5] "PetalWidthCm"  "Species"      
## 
## $class
## [1] "data.frame"
##
## $row.names
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [18]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35]  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52]  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69]  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86]  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102
## [103] 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
## [120] 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136
## [137] 137 138 139 140 141 142 143 144 145 146 147 148 149 150
```

A plot of histogram to know the Sepal Length

```
colors = c("blue","grey","blue","grey","blue","grey","blue","grey")
hist(dataset$SepalLengthCm,col=colors)
```

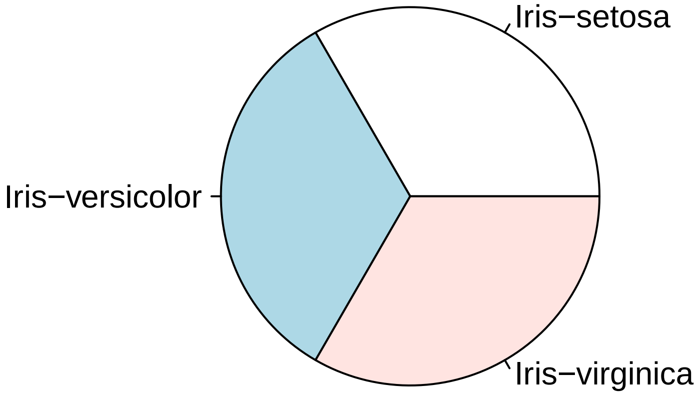


Different Spicies in the dataset

```
print(table(dataset$Species))
```

```
##
##     Iris-setosa Iris-versicolor  Iris-virginica
##              50              50              50
```

```
pie(table(dataset$Species))
```



Now in order to understand how to differentiate these species, we need to understand the length and width measures of Sepal and Petal for all the species, we want to understand those factors that differentiate one species from the other so we are need of finding out the correlation between each the availabe attributes.

```
print(cov(dataset[,2:5]))
```

```
##               SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
## SepalLengthCm    0.68569351  -0.03926846     1.2736823    0.5169038
## SepalWidthCm    -0.03926846   0.18800403    -0.3217128   -0.1179812
## PetalLengthCm    1.27368233  -0.32171275     3.1131794    1.2963875
## PetalWidthCm     0.51690380  -0.11798121     1.2963875    0.5824143
```

```
print(cor(dataset[,2:5]))
```

```
##               SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
## SepalLengthCm     1.0000000   -0.1093692     0.8717542    0.8179536
## SepalWidthCm     -0.1093692    1.0000000    -0.4205161   -0.3565441
## PetalLengthCm     0.8717542   -0.4205161     1.0000000    0.9627571
## PetalWidthCm      0.8179536   -0.3565441     0.9627571    1.0000000
```

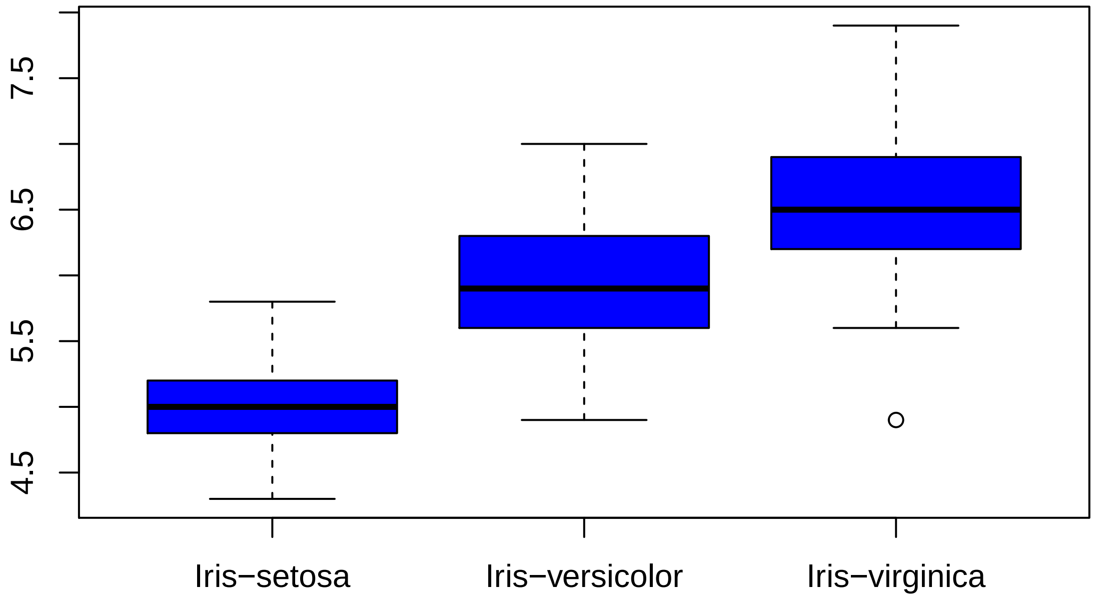
Futher we need to investigate if the summary statistics of each species seprately to know the measures of distinctive species.

```
print(aggregate(dataset$SepalLengthCm~dataset$Species,summary,data=dataset))
```

```
##   dataset$Species dataset$SepalLengthCm.Min. dataset$SepalLengthCm.1st Qu.
## 1     Iris-setosa                      4.300                         4.800
## 2 Iris-versicolor                      4.900                         5.600
## 3  Iris-virginica                      4.900                         6.225
##   dataset$SepalLengthCm.Median dataset$SepalLengthCm.Mean
## 1                        5.000                      5.006
## 2                        5.900                      5.936
## 3                        6.500                      6.588
##   dataset$SepalLengthCm.3rd Qu. dataset$SepalLengthCm.Max.
## 1                         5.200                      5.800
## 2                         6.300                      7.000
## 3                         6.900                      7.900
```

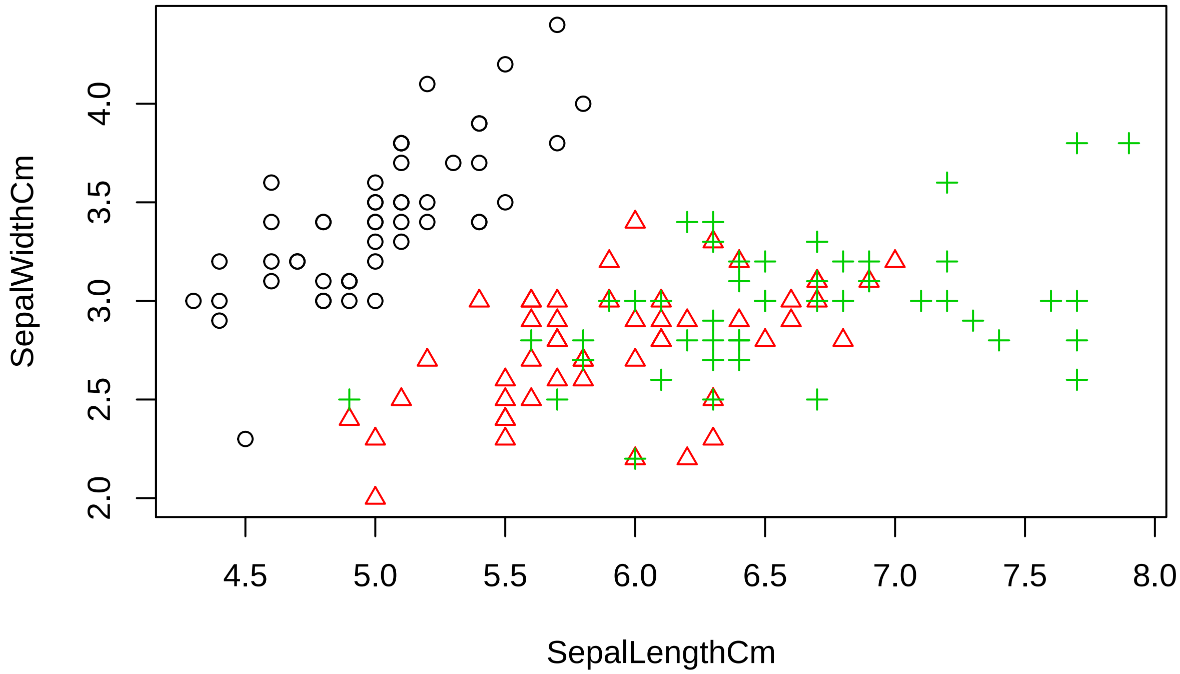
Box Plot to See the variability of Sepal Lengths between the different species

```
boxplot(dataset$SepalLengthCm~dataset$Species,data = dataset,col= "blue")
```



Now we plot the scatter plot with Species Sepal lengths and Sepal Widths

```
with(dataset,plot(SepalLengthCm,SepalWidthCm,col=Species,pch=as.numeric(Species)))
```



Now to try clustering we remove the attribute Species from the data set

```
newdata=dataset[,c(2,3,4,5,6)]
newdata$Species = NULL
result = kmeans(newdata,3)
print(result)
```

```
## K-means clustering with 3 clusters of sizes 38, 50, 62
##
## Cluster means:
##   SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
## 1      6.850000     3.073684      5.742105     2.071053
## 2      5.006000     3.418000      1.464000     0.244000
## 3      5.901613     2.748387      4.393548     1.433871
##
## Clustering vector:
##   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [36] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [71] 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1
## [106] 1 3 1 1 1 1 1 1 3 3 1 1 1 1 3 1 3 1 3 1 1 3 3 1 1 1 1 1 3 1 1 1 1 3 1
## [141] 1 1 3 1 1 1 3 1 1 3
##
## Within cluster sum of squares by cluster:
## [1] 23.87947 15.24040 39.82097
##  (between_SS / total_SS =  88.4 %)
##
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Now we shall check if the clustering is good by checking if the grouping produce is as per the Species acutually present in the data set.

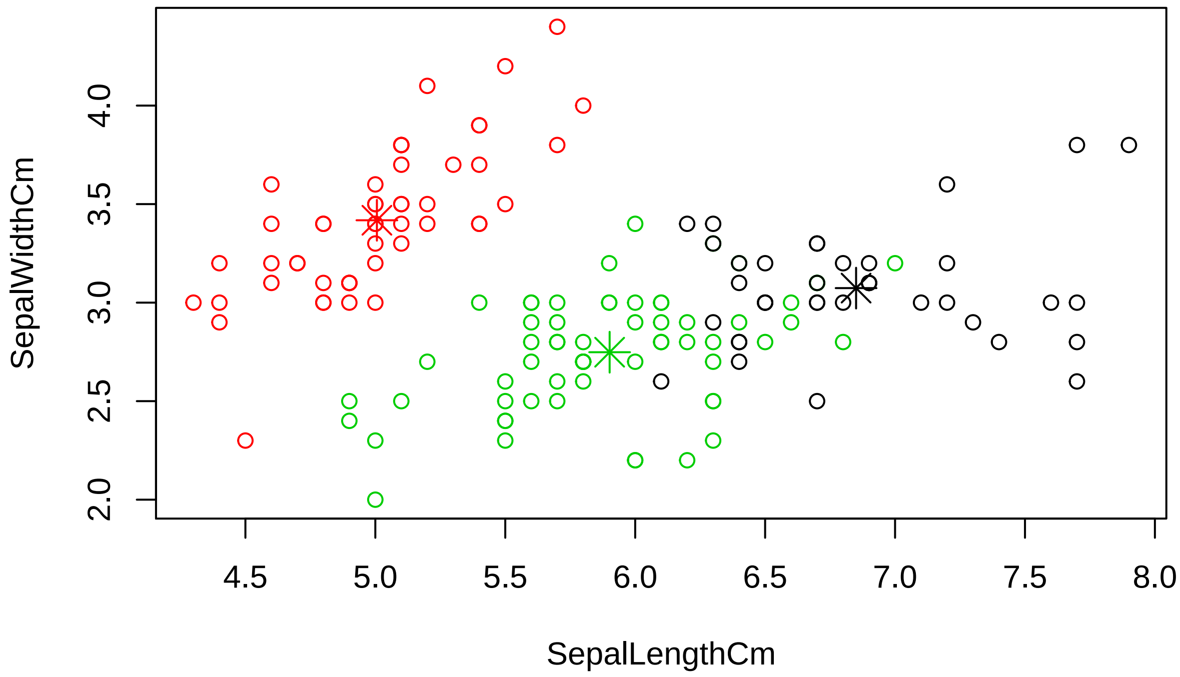
```
table(dataset$Species, result$cluster)
```

```
##
##                    1  2  3
##   Iris-setosa      0 50  0
##   Iris-versicolor  2  0 48
##   Iris-virginica  36  0 14
```

```
plot(newdata[,c("SepalLengthCm","SepalWidthCm")],col=result$cluster)

#Plotting Cluster Centers

points(result$centers[,c("SepalLengthCm","SepalWidthCm")],col=1:3,pch=8,cex=2)
```


SepalWidthCm
SepalLengthCm