Analysis Of Iris Dataset

Nivedita Pandey

2024-02-09

Objective

To perform data analysis on iris data set for identifying the effective classification characteristic for different flower species

Analysis

Iris data set consists of 50 samples from each of the three sub-species (iris setosa, iris virginica, and iris versicolor). Four features were measured in centimeters (cm): the lengths and the widths of both sepals and petals.

```
View(iris) # show as a spreadsheet
```

```
class(iris) # show the data type
```

```
## [1] "data.frame"
```

```
head(iris) # first few rows
```

```
##   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
```

```
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 ...
```

The imported data is stored in a data frame called iris, which contains information of 5 variables for 150 observations (flowers). While the first 4 variables/columns, Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width, contain numeric values. The 5th variable/column, Species, contains characters indicating which sub-species a sample belongs to.

First we will use summary function so that we can quickly summarize the data set and it will provide us information on the distribution of each variables.

```
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
##
##
##
```

This illustrates how well-balanced the species are. Each species (Iris virginica, setosa, and versicolor) has a count of 50. The mean of the Sepal length is greater than the mean of the other three measurements and petal width has the lowest average measurements.The shortest petal in the data set is 1 cm while the longest petal is 6.9 cm and the shortest sepal in the data set is 4.3 cm while the longest sepal is 7.9 cm. The widths of the petals and sepals vary from 0.1 cm to 2.5 cm and 2 cm to 4.4 cm respectively.

Now we will plot box plot as a box plot is extremely effective mechanism to get a one shot view and understand the nature of data. It displays a summary of a large amount of data in five numbers — minimum, lower quartile(25th percentile), median(50th percentile), upper quartile(75th percentile) and maximum data values.

First we will plot for all quantitative variables, to get picture of all Quantitative variables

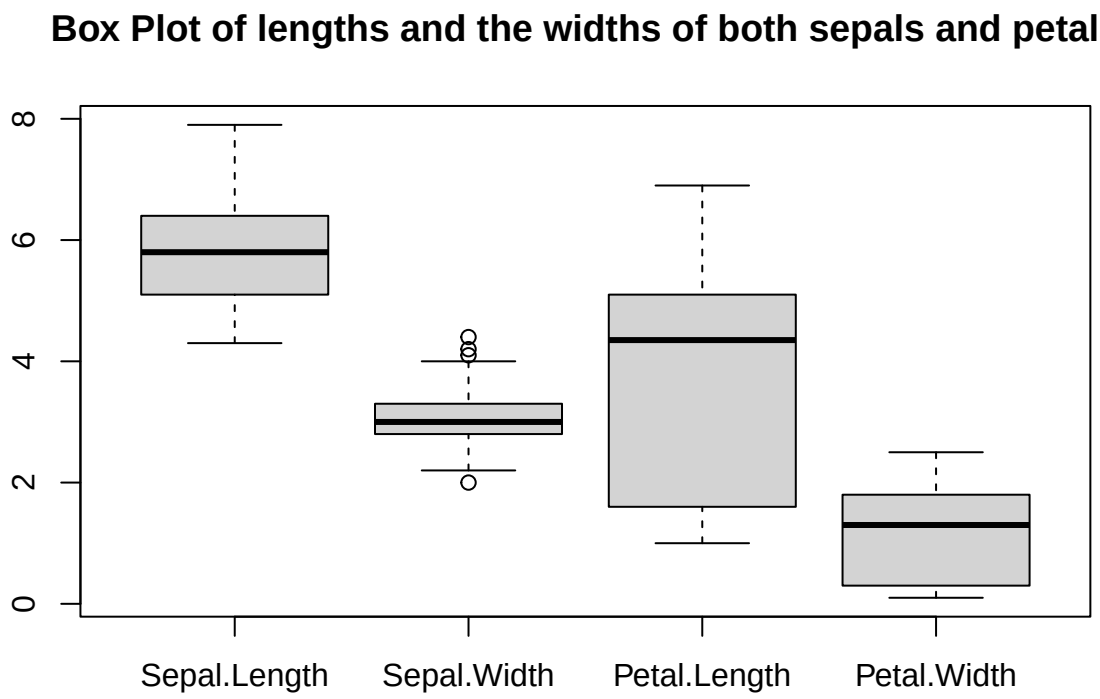

Fig 1.1 Box plot of lengths and the widths of both sepals and petals

From Fig 1.1 we can interprete the boxplot of lengths and the widths of both sepals and petals as follows

In sepal length the central rectangle or the box spans from frist quartile, 5.1 to third quartile 6.4, then the Inther Quartile Range (IQR) will be 1.3 and median is given by the line or band within the box which is 5.8.

In sepal width first and third quartiles are 2.8 and 3.3 respectively that means 25% of the data falls below 2.8 and 75% of the data falls below 3.3. Also IOR is 0.5 and median is 3. Also many outliers are present in sepal width.

In petal length first and third quartiles are 1.6 and 5.1 respectively that means 25% of the data falls below 1.6 and 75% of the data falls below 5.1 Also IOR is 3.5 and median is 4.3.

In petal width first and third quartiles are 0.3 and 1.8 respectively that means 25% of the data falls below 0.3 and 75% of the data falls below 1.8 Also IOR is 1.5 and median is 1.3.

Also distribution of sepal length and sepal width are normally distributed while petal length and petal width are negatively skewed.

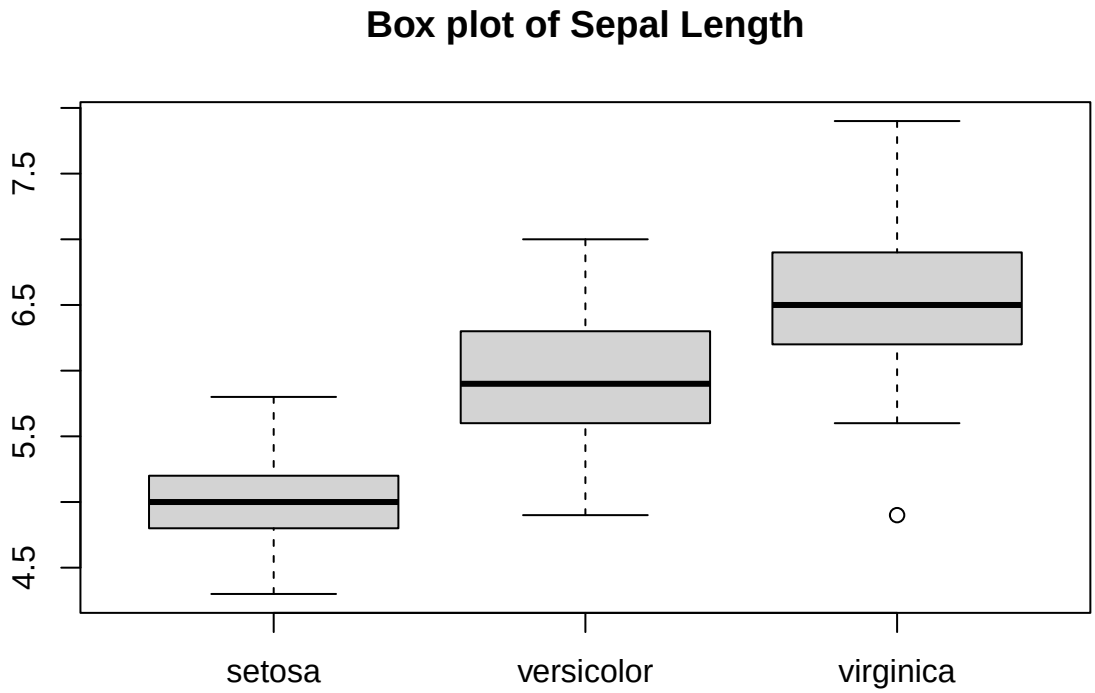

Fig 1.2 Box Plot of Sepal length

From Fig 1.2 For setosa we can observe The sepal length of setosa, versicolor, virginica approximately lies between 4.3 to 5.8, 4.9 to 7, 5.6 to 7.9 respectively. In box plot of virginica one observation is coming beyond the lower whisker so one value is unusually low value so one outlier is present in the data set of sepal length of virginica.

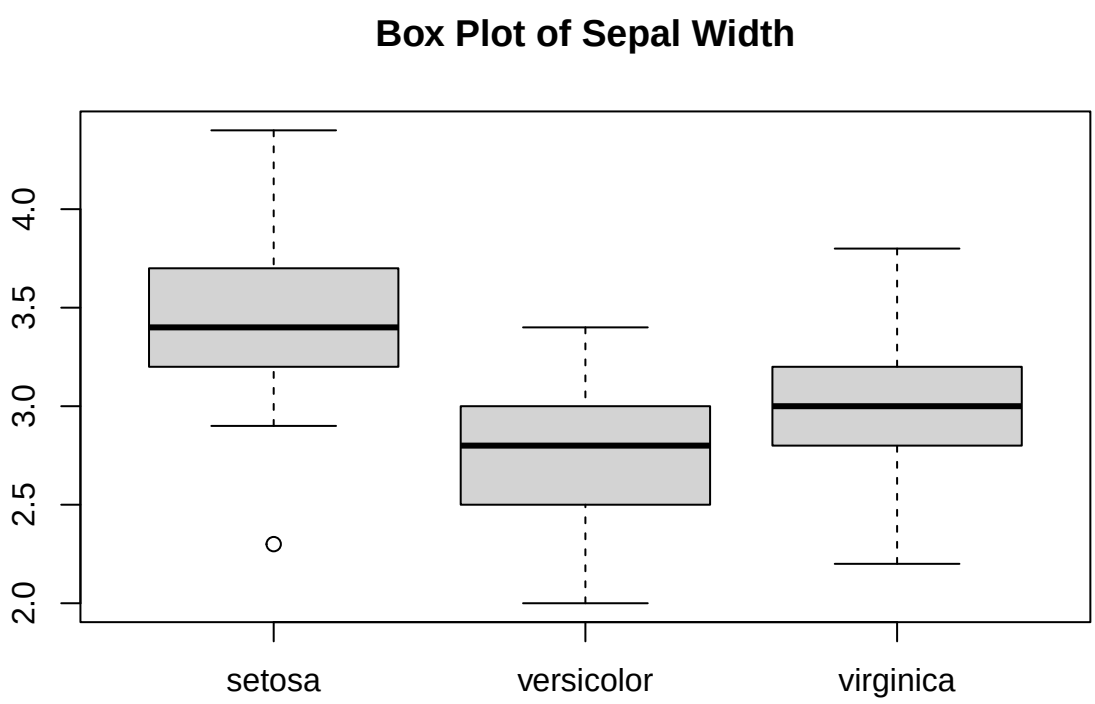

Fig 1.3 Box Plot of Sepal Width

From Fig 1.3 we can observe that the sepal width of setosa, versicolor, virginica approximately lies between 2.9 to 4.4, 2 to 3.4, 2.3 to 3.7 respectively. In box plot of setosa one observation is coming beyond the lower whisker so one value is unusually low value so one outlier is present in the data set of sepal width of setosa.

Fig 1.4 Box Plot of Petal Width

From Fig 1.4 we can observe that the sepal width of setosa, versicolor, virginica approximately lies between 0.1 to 0.4, 1 to 1.8, 1.4 to 2.5 respectively. In box plot of setosa two observations are coming beyond the upper whisker so two values are unusually high value so two outliers are present in the data set of petal width of setosa.

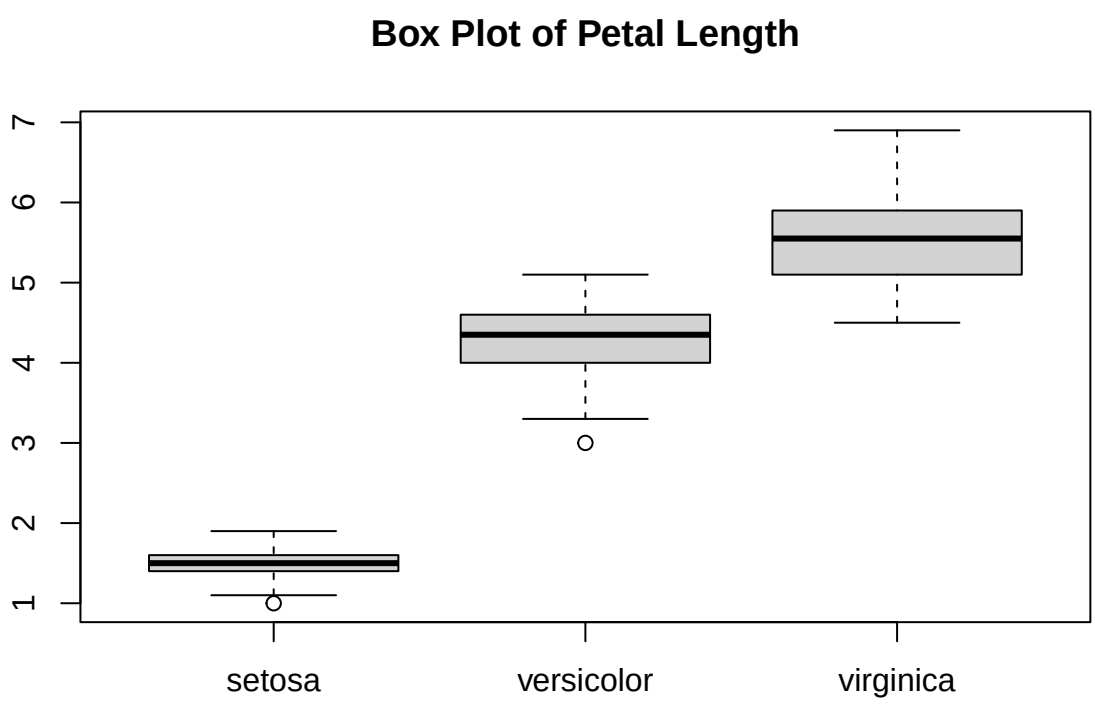

Fig 1.5 Box Plot of Petal Length

From Fig 1.5 we can observe that the sepal width of setosa, versicolor, virginica approximately lies between 1 to 2, 3.2 to 5, 4.5 to 6.9 respectively. In box plot of setosa and versicolor one observation is coming beyond the lower whisker so one value is unusually low value so one outlier is present in the data set of petal width of setosa and versicolor.

Now we will make bivariate table

```
##
##              4-4.5 4.5-5 5-5.5 5.5-6 6-6.5 6.5-7 7-7.5 7.5-8
##   setosa         5    23    19     3     0     0     0     0
##   versicolor     0     3     8    19    12     8     0     0
##   virginica      0     1     0     8    19    10     6     6
```

Table 1.1 Table for Sepal length of three species in iris dataset

```
##
##              1.5-2 2-2.5 2.5-3 3-3.5 3.5-4 4-4.5
##   setosa         0     1     7    26    13     3
##   versicolor     1    12    29     8     0     0
##   virginica      0     5    28    14     3     0
```

Table 1.2 Table for Sepal width of three species in iris dataset

```
##
##              0-0.5 0.5-1 1-1.5 1.5-2 2-2.5
##   setosa        49     1     0     0     0
##   versicolor     0     7    38     5     0
##   virginica      0     0     3    24    23
```

Table 1.3 Table for Petal width of three species in iris dataset

```
##
##              0.5-1 1-1.5 1.5-2 2-2.5 2.5-3 3-3.5 3.5-4 4-4.5 4.5-5 5-5.5 5.5-6
##   setosa         1    36    13     0     0     0     0     0     0     0     0
##   versicolor     0     0     0     0     1     4    11    20    13     1     0
##   virginica      0     0     0     0     0     0     0     1     8    16    16
##
##              6-6.5 6.5-7 7-7.5
##   setosa         0     0     0
##   versicolor     0     0     0
##   virginica      5     4     0
```

Table 1.4 Table for Petal length of three species in iris dataset

From Table 1.1 and 1.2 we can see that Sepal length and width of all species are overlapping to each other.

From Table 1.3 to Table 1.4 we can see that petal length & petal width of setosa is smaller than 2 centimeters and 0.5 centimeters respectively. Petal length and width of verginica is from 5.5 to 7 centimeters and from 1 to 2.5 centimeters respectively whereas petal length and width of versicolor lies between setosa and verginica. So we can say that setosa have smaller petal length, petal width and sepal length as compared to verginica.

We will use density plot for observing the distribution of length and width of sepal and petal of all three species.

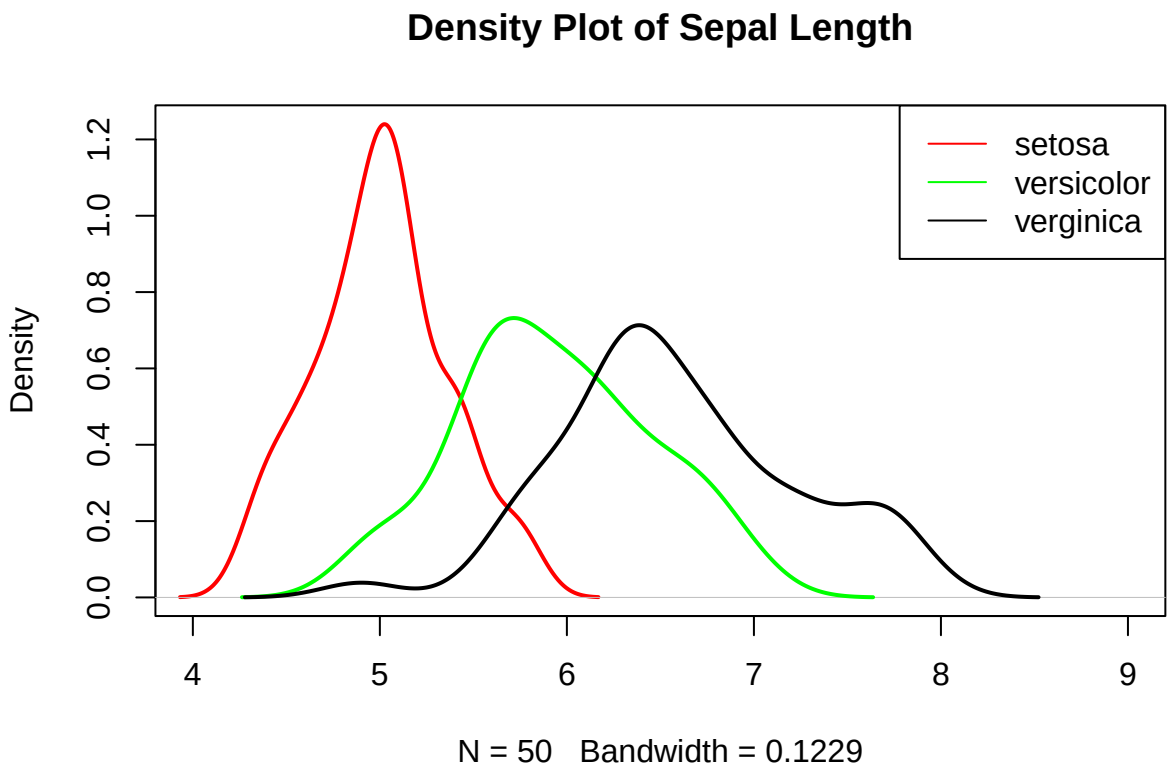

Fig 1.6 Density Plot of Sepal Length

From Fig 1.6 which is density plot of sepal length of three species we can observe that it demonstrates that the density plot overlap between the species in terms of sepal length, indicating that it is ineffective as a classification characteristic.

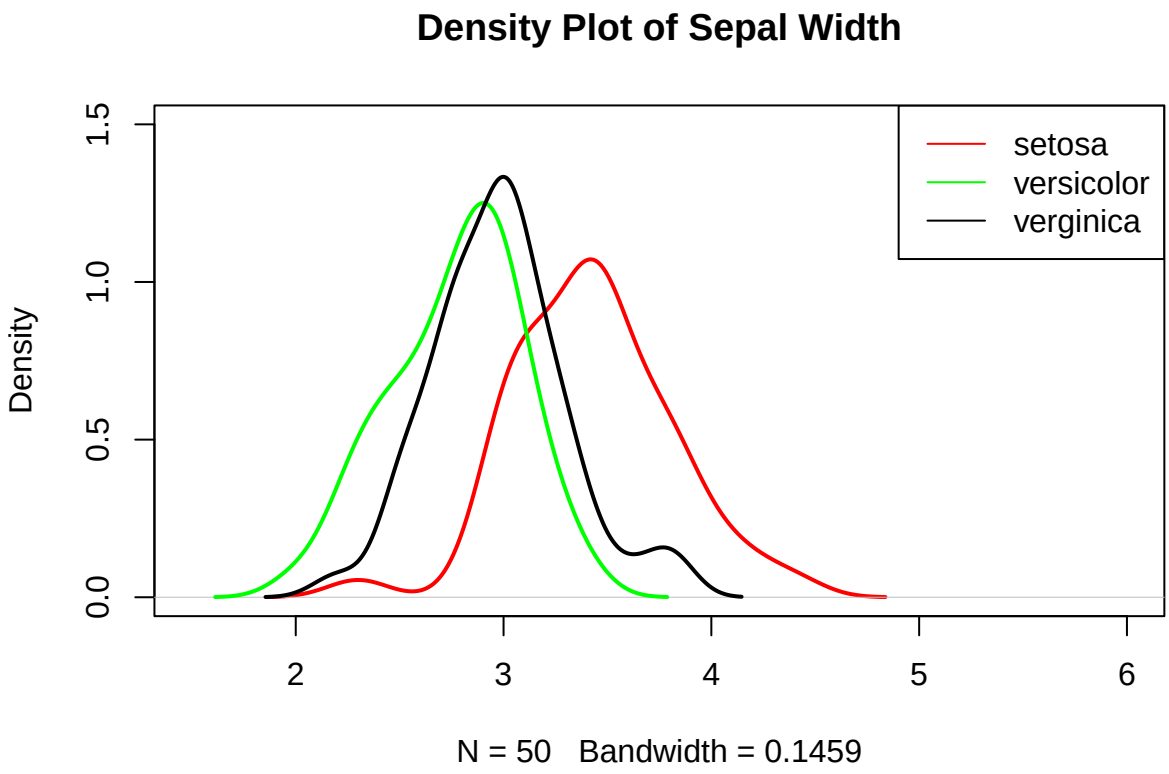

Fig 1.6 Density Plot of Sepal Width

From Fig 1.6 which is density plot of sepal width we can observe that sepal width can not be used as diffrentiator for any of the three species as density plot of sepal width of all three species setosa, versicolor and verginica more overlap between the species that mean sepal length for all three species lies in approximately same range.

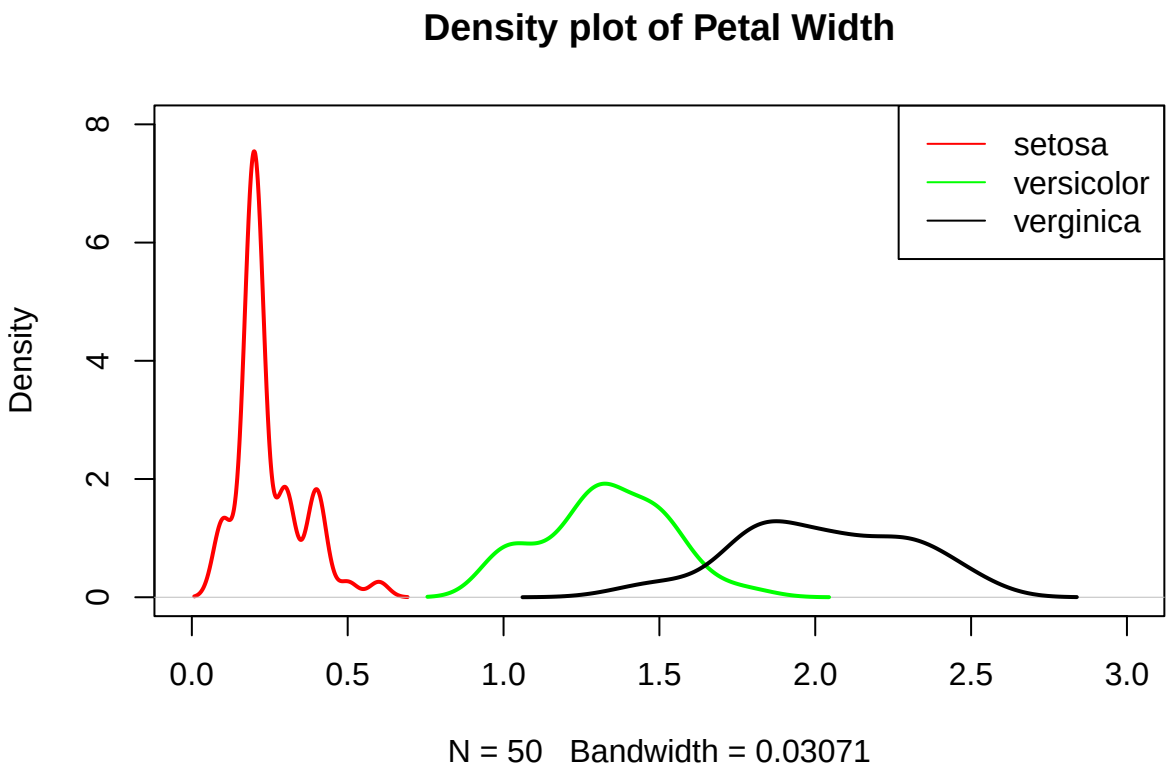

Fig 1.7 Density Plot of Petal Width

From the density plot of Petal width for all species in Fig 1.7 we can observe that petal width of setosa is very small as compared to versicolor and verginica. Also sepal length of versicolor is smaller than verginica. Petal width can be used as a diffrentiator for all species as the overlap is little (between Versicolor and Virginica), while Setosa is well separated from the other two.

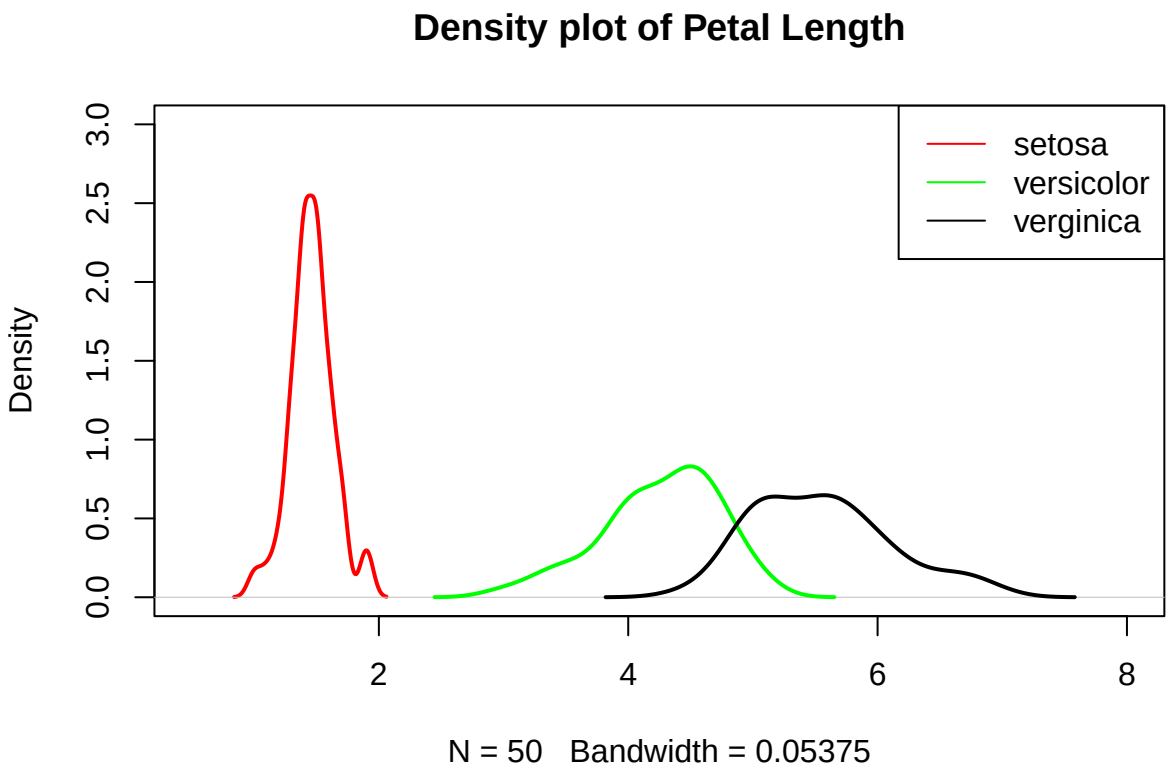

Fig 1.8 Density Plot of Petal Length

From the density plot of petal length in Fig 1.8 we can observe similar result as petal width that petal length of setosa is smaller than versicolor and verginica and petal length of versicolor is smaller than verginica. Also the overlap is little (between Versicolor and Virginica), while Setosa is well separated from the other two. So petal width and petal length can be used as a diffrentiator for all species.

Now we will use scatter plot between sepal length, sepal width and between petal length petal width for three species to get more clear picture.

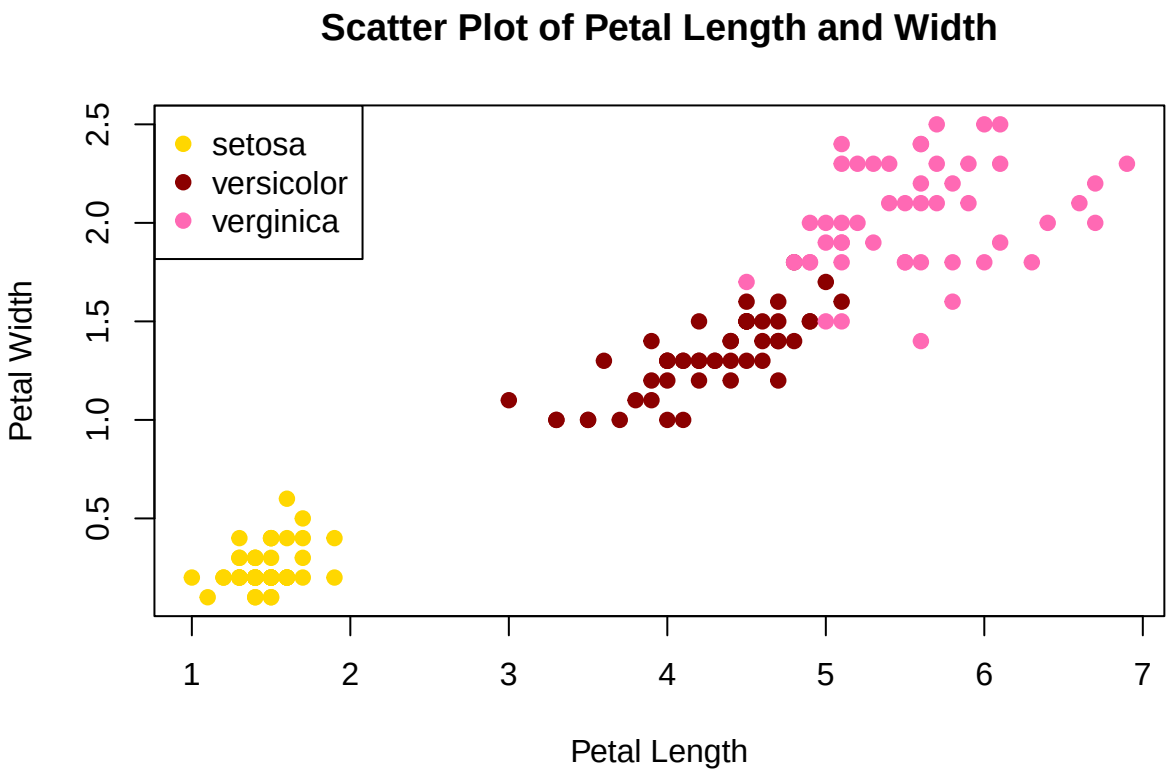

Fig 1.8 Scatter Plot of Petal Length and Width

From Fig 1.8 we observe that Setosa has the shortest petal length and breadth. Petal length and width are normal for the Versicolor species. Virginica species have the maximum petal length and width Petal length and width are diffrentiatinng the three species clearly from each other. Petal length and width of setosa is smallest and verginica is largest.

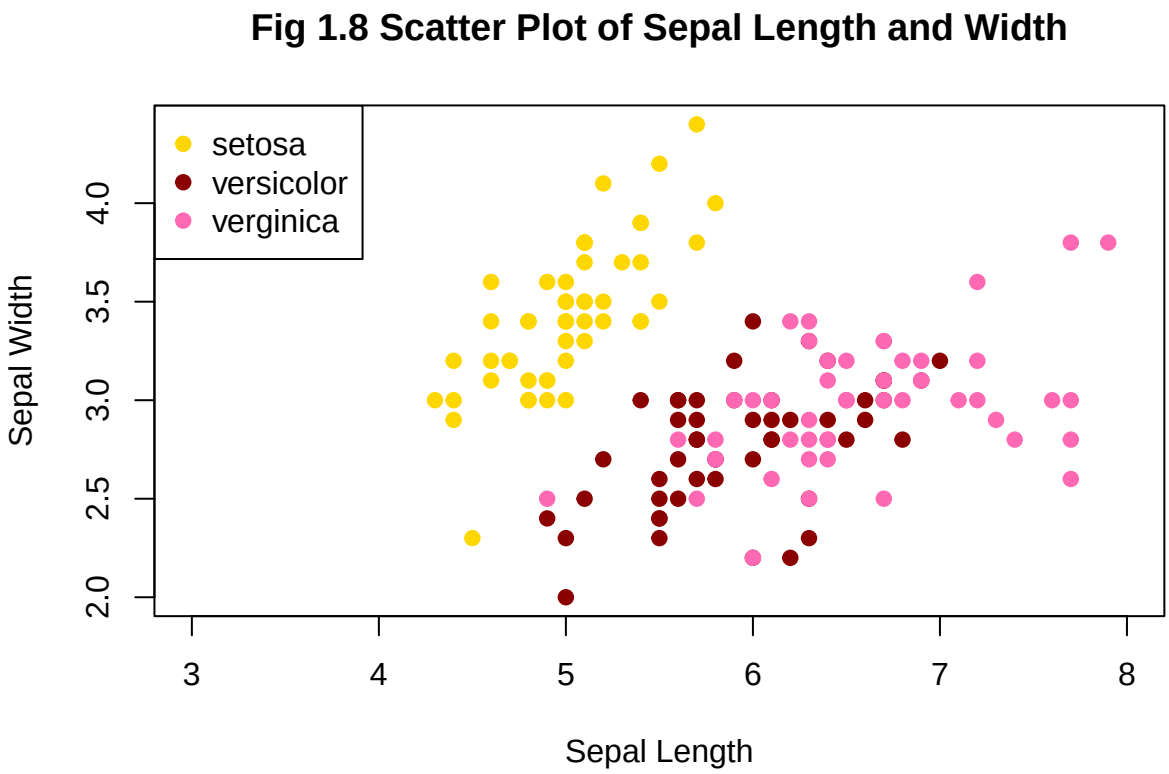

Fig 1.8 Scatter Plot of Sepal Length and Width

From Fig 1.8 we can see that Setosa has shorter sepals but wider petals. Versicolor is almost in the center in both length and width. Virginica has longer sepals and narrower sepals. Only setosa can be differentiated from other species but versicolor and verginica can not be diffrentiated.

Conclusion

From all the above analysis we can conclude that petal length and petal width can be used as effective classification characteristic for all three species of iris which are setosa versicolor and verginica. We can distinguish the setosa species effortlessly using length and width of sepal and petals but the Versicolor and Virginica species are frequently combined and can be difficult to distinguish.