Flower.analysis

Nada_Saiyed | Neethu_Mariya | Shadh_Shanavas

1/31/2020

R Markdown

This is an R Markdown document. The following document deals with the Exploratory Data Analysis of Flower dataset, which we have in an excel file. To perform initial analysis on the dataset, we move the excel file to our working directory of RStudio and load the data in a flower data-frame.

```
flower<-read.csv("flower.csv",header = TRUE)
```

1. Summarizing FLOWER Dataset

Now that we have our data in a better readable format, we use `dim(flower)` command and learn that there are **150 records** and **3 columns**. Further, to get a summary of our dataset, we use "summarytools" library to run `dfSummary(flower)` command and get the following output.

```
## Dimensions: 150 x 3
## Duplicates: 47
##
## ------------------------------------------------------------------------------------------------------------------
## No   Variable        Stats / Values           Freqs (% of Valid)   Graph                           Valid    Missing
## ---- --------------- ------------------------ -------------------- ------------------------------- -------- ---------
## 1    petal_length    Mean (sd) : 3.8 (1.8)    43 distinct values   :                               150      0
##      [numeric]       min < med < max:                              :         . :                   (100%)   (0%)
##                      1 < 4.3 < 6.9                                 :         : : .
##                      IQR (CV) : 3.5 (0.5)                          : :       : : : .
##                                                                    : :   . : : : : : .
##
## 2    petal_width     Mean (sd) : 1.2 (0.8)    22 distinct values   :                               150      0
##      [numeric]       min < med < max:                              :                               (100%)   (0%)
##                      0.1 < 1.3 < 2.5                               :       . .   :
##                      IQR (CV) : 1.5 (0.6)                          :       : :   :   .
##                                                                    : :   : : : . : : :
##
## 3    class           1. Iris-setosa           50 (33.3%)           IIIIII                          150      0
##      [factor]        2. Iris-versicolor       50 (33.3%)           IIIIII                          (100%)   (0%)
##                      3. Iris-virginica        50 (33.3%)           IIIIII
## ------------------------------------------------------------------------------------------------------------------
```

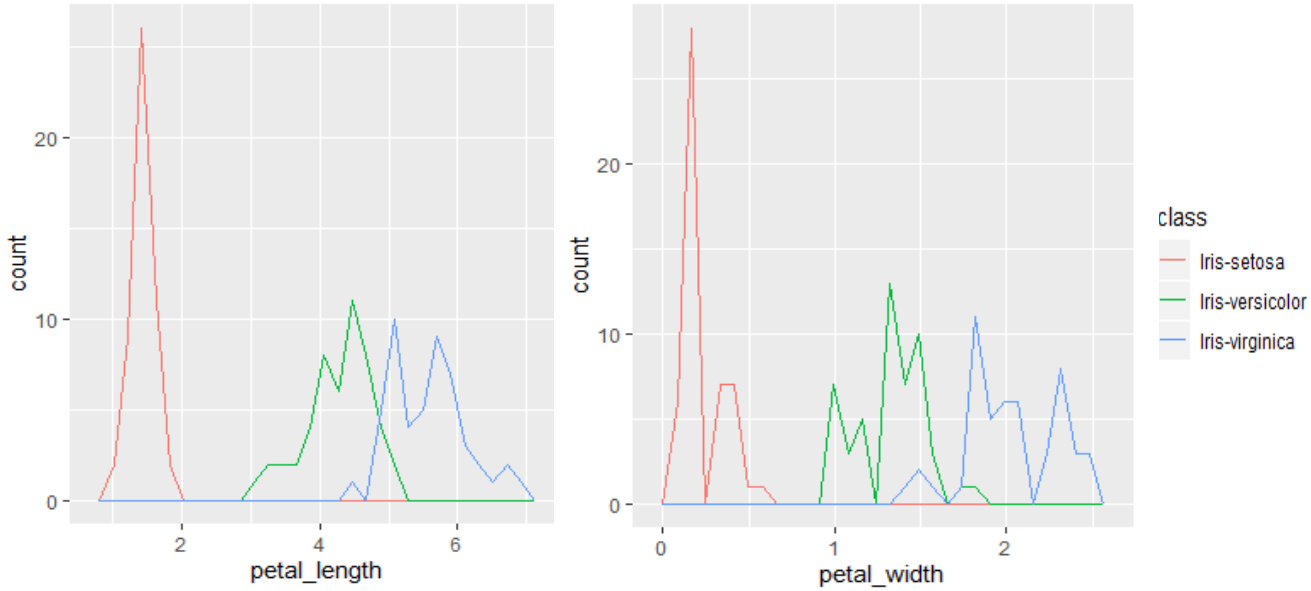
From the last two columns we can see that the data is valid and we have no missing values. Further, we can see that we have one categorical column, **class**, which has 3 unique values, namely "Iris-setosa", "Iris-versicolor" and "Iris-virginica" with equal number of occurrences i.e. 50 data points of each category. The other two columns, namely **petal_width** and **petal_length** are numeric columns with continuous data.

petal_ length ranges from 1 to 6.9 units with an overall mean of 3.8 units which is less than the median (4.3 units). From the rough histogram that we obtained, we can see that the observations are divided into two major groups, with a standard deviation of 1.8.

petal_width ranges from 0.1 to 2.5 units, with an overall mean of 1.2 units, which is less than the median (1.3 units). The rough histogram depicts that the distribution is unevenly divided into 3 groups with a standard deviation of 0.8. Comparing the s.d. of petal length and petal width we can say that petal length has more variation compared to petal widths.

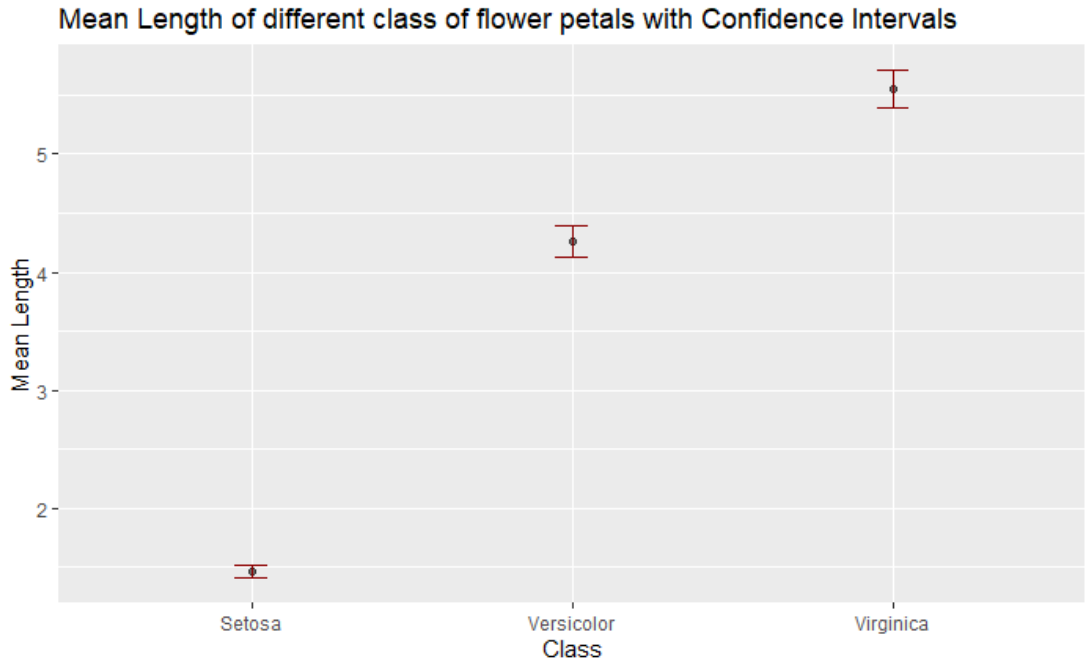
2. Comparing Petal size

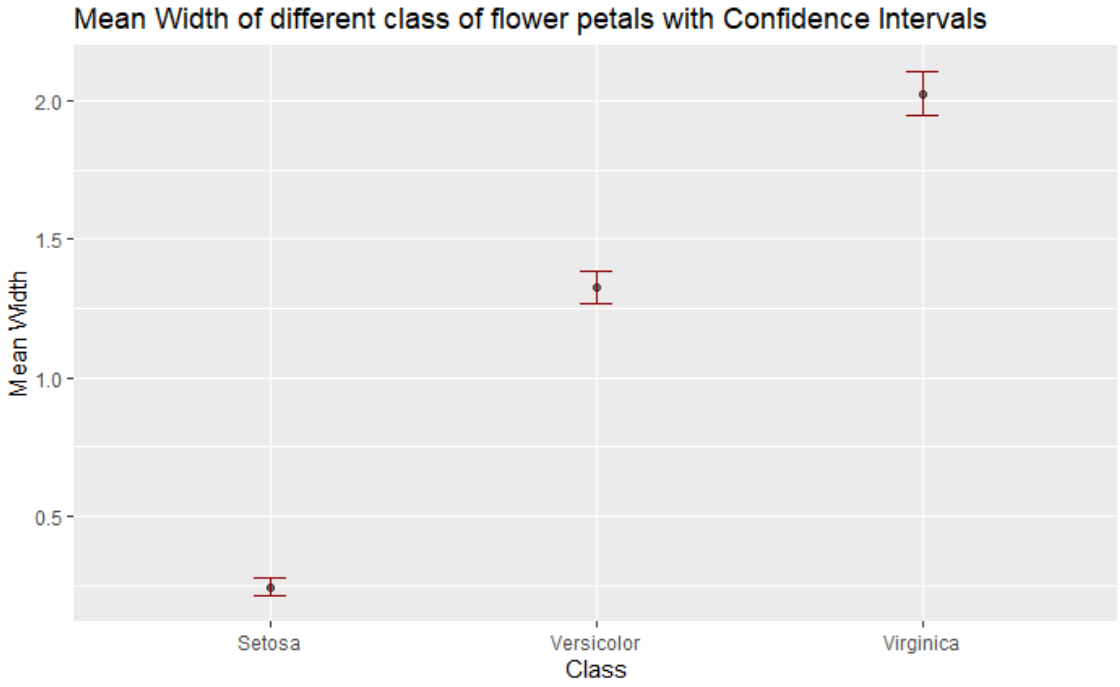
To compare the petal width and petal length for each class of flower to further analyze the distribution, we plot line graphs grouping it based on flower class.



We notice that for flower class Iris-setosa, the petal width and petal length is the smallest. From here we can conclude that **Iris-setosa is the smallest flower**. However, we cannot conclude the same for rest two classes of flower as their petal length and width overlaps a little.

Plotting a scatter plot, petal lengths against widths, we get a better analysis of flower size. From here we can definitely say that **Iris-setosa is the smallest flower and Iris-virginica is the largest**. Iris- versicolor is definitely larger than Iris-setosa, but to confirm that its smaller than Iris-virginica, we calculate and plot a 95% confidence interval for lengths and widths of the flower.

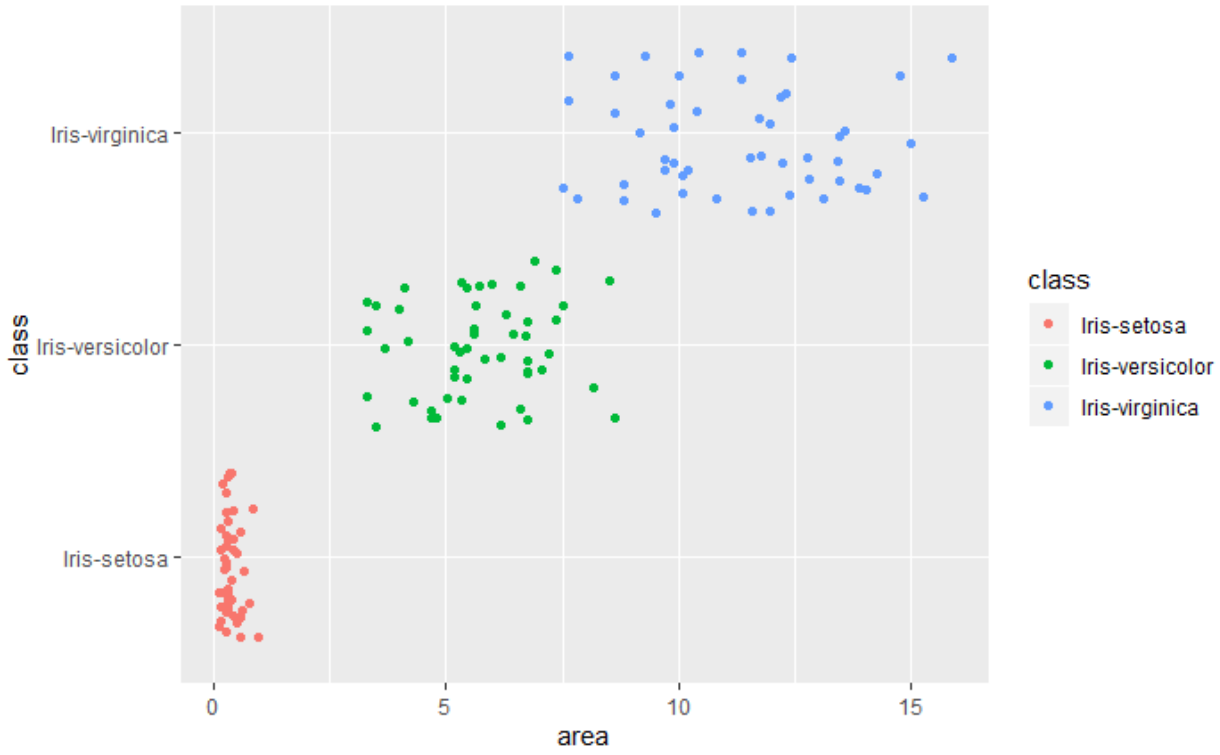




Seeing the 95% CI for each class's mean width and length we can conclude that Iris-Setosa has the smallest petal, followed by Iris-Versicolor and Iris-Virginica has the largest petal amongst all three.

3. Conclusion

From the analysis conducted on different classes of Iris flower, we have some idea about which flower is the largest and smallest. However, to support our claim, we calculate rough area of each flower petal by multiplying length and width. Then we plot the area and classify them based on classes



From the plot we can confidently say that to Ronald that Iris-setosa has the smallest petal size and Iris-virginica has the largest petal size.