Exploratory Data Analysis of Flower Dataset

Rishi Shah, Akshay Kapoor

31/01/2020

Introduction

Flower.csv is comma separeted file containing the data of iris-setosa, iris-virginica, and iris-versicolor classes of flowers. Data has two features named petal length and petal width for the given classes of data.

petal_length : It is numerical vector giving length of the petal

petal_width : It is numerical vector giving width of the petal

We have to carry out the exploratory data analysis of the given data set in order to improve our knowledge to impress our grandpa who likes flowers a lot.

Analysis

To begin the analysis, we need to summarise the dataset first to understand the size of the dataset and features available to us.

```
## 'data.frame':    150 obs. of  3 variables:
##  $ 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 ...
##  $ class       : Factor w/ 3 levels "Iris-setosa",..: 1 1 1 1 1 1 1 1 1 1 ...
```

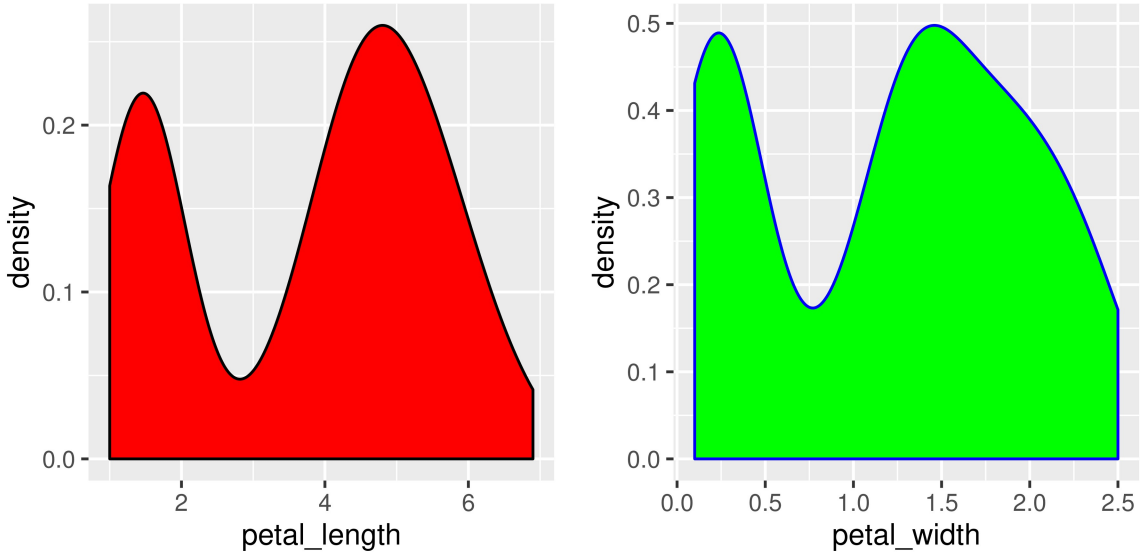
It can be shown from the summary that Total **number of observations available in the dataset is 150**. There are a total of **3 variables** named petal-length, petal-width and class of the flower. Petal length and petal width are of the numerical data type. While class is the factor data type with 3 levels namely iris-setosa, iris-virginica, and iris-versicolor.

Central tendency, Dispersion and Range of the Dataset

In order to understand each variable we need to calculate it's statistics and hence understand central tendency, dispersion and range of the same.

```
##   petal_length    petal_width                class
##  Min.   :1.000   Min.   :0.100   Iris-setosa    :50
##  1st Qu.:1.600   1st Qu.:0.300   Iris-versicolor:50
##  Median :4.350   Median :1.300   Iris-virginica :50
##  Mean   :3.759   Mean   :1.199
##  3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :6.900   Max.   :2.500
```

1. Petal_length : Petal length is the length of the petal within the range of 1 o 6.9. It can be observed from the above summary that 50% of the petal length resides within the range of 1.6 to 5.1.



From the above density plot of the petal_length, it can be clearly depicted that the data distribution is coming from two different classes having two different mean and standard deviation.Mean of the Petal_length for the complete dataset is 3.759 and standard deviation is 1.7644.

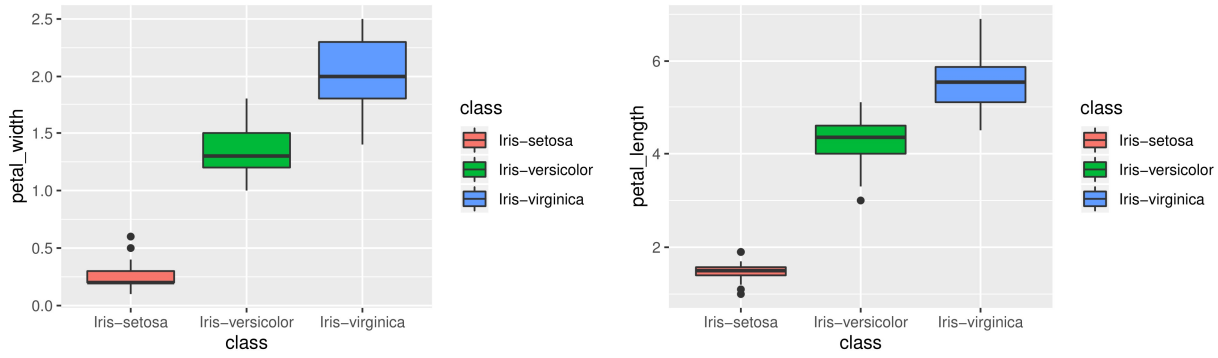
2. Petal_width : Width of petal ranges between 0.1 to 2.5.The interquartile range is 1.5. Above density plot of petal_width represents the distribution of the width. similar to length, Both the classes of flowers have different mean petal_width and distribution.

petal_width Mean of the entire data set is 1.199 and the standard deviation is 0.7631607.

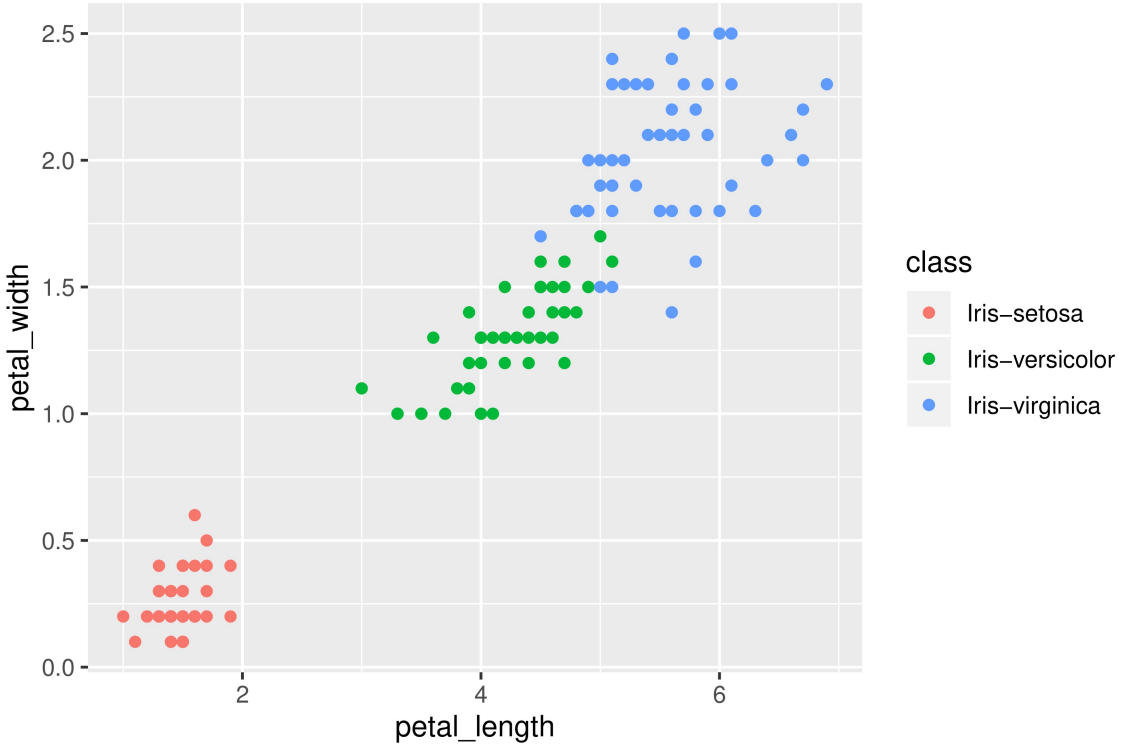
Moving on, we can explore the central tendency and dispersion of the features for each class of the iris flower. Following table represents the mean and standard deviation for each class of the data.

```
##   width_mean length_mean  width_sd length_sd           class
## 1      0.244       1.464 0.1735112 0.1735112     Iris-setosa
## 2      1.326       4.260 0.4699110 0.4699110 Iris-versicolor
## 3      2.026       5.552 0.5518947 0.5518947  Iris-virginica
```

Below plotted blox plot gives us the relative visualization of length and width for different classes and will help us to classify the different classes of the flowers.

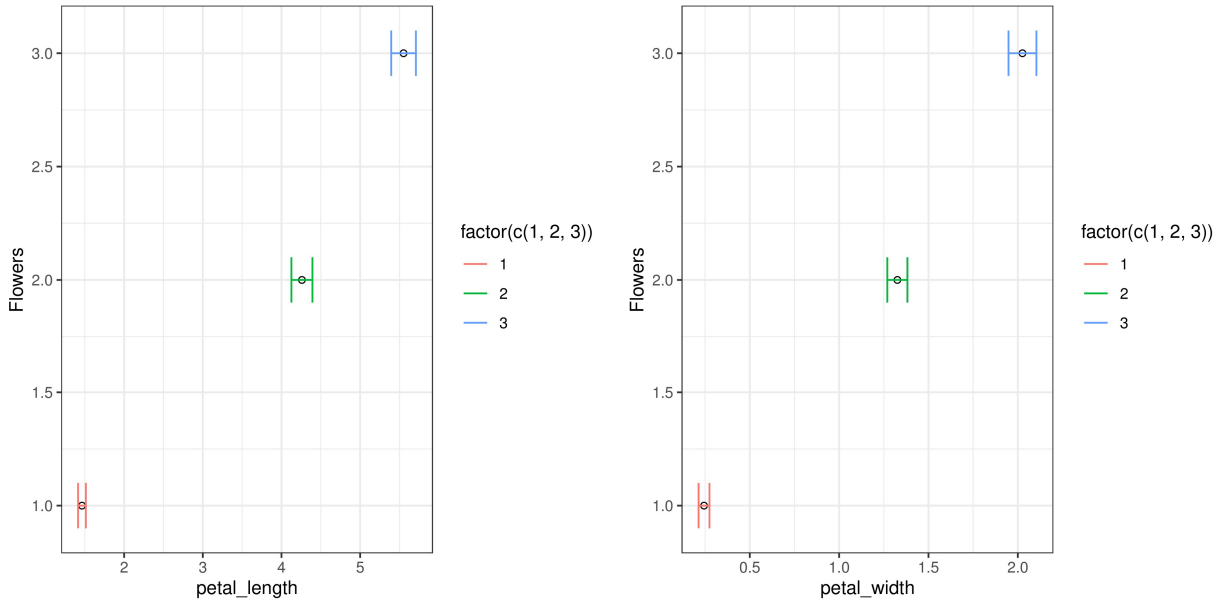


To visualise the distribution of all the data points in regard to the different flowers, We plotted the petal width over the petal length in a scatter plot.



From the above scatter plot, we infer that iris-setosa flower has the lowest petal as its petal_lengths and widths are clustered at lower value.While in contrast, iris-virginica has the largest petal from all the other classes of iris flowers.

This result can be supported by looking at the confidence interval. Following plots represent the confidence interval of the pelat_length and petal_width for every class.



Looking at the above confidence interval graph for petal_length and petal_width and calculations carried out in R for different class, we can observe following things:

1. Petal_length of the iris-setosa(class 1) has minimum mean (1.464) with a confidence interval(95%) of 1.5133 to 1.41468. On the contrary, iris - virginica(class 3) flower has the largest mean petal_length of 5.552 and can be found within 5.7088 to 5.3951, with a confidence of 95%. Mean petal length of iris-versicolor(class 2) is 4.260 which is greater than the class 1 flower but less than the class 3 flower.
2. Similar to the length, iris-setosa(class 1) has minimum mean width equals to 0.244 and can be found with in the range of 0.2744 to 0.2135, with a confidence of 95%. While iris-virginica(class 3) has the maximum mean width of 2.026 with 95% confidence margin of 0.07805 on both positive and negative side of mean. iris-versicolor (class 2) has an average mean petal width of 1.326.

With the quantitative evidence from the calculations above, we can say that **iris-setosa(class 1) flower has the smallest petal** while **iris-virginica has the largest petal**.

Conclusion

Comparing the estimate of means with 95% confidence for petal length and petal width of given dataset and carrying out the exploratory data analysis on the same, We can conclude that iris-setosa (class 1) tend to have the smallest petal and iris-virginica(class 3) tend to have the largest petal.

References

1. Field, A., Miles, J., and Field, Z. (2012).Discovering Statistics using R.https://studysites.uk.sagepub.com/dsur/
2. https://r4ds.had.co.nz/
3. https://stackoverflow.com/questions/34535155/figure-size-in-r-markdown

Team Contribution

Akshay Kapoor and Rishi Shah have worked together on this report.Rishi has contributed in data exploration and plotig the right graphs whereas Akshay has worked on knitting the report and analysing the data.