Fisher’s Iris Data set

SOMANATHAN SUBRAMANIYAN Student Id: G00364742

Programming and Scripting

Project 2018

2018

Table of Contents

[Problem statement 2](#_Toc512546868)

[R.A Fisher and Irish Data Set 2](#_Toc512546869)

[Irish Data Set: 2](#_Toc512546870)

[How to run the python code and technical summary 2](#_Toc512546871)

[Data Set - Analysis and Investigation 3](#_Toc512546872)

[Iris Setosa: 3](#_Toc512546873)

[Iris Versicolor: 3](#_Toc512546874)

[Iris Virginica: 4](#_Toc512546875)

[Petal Length Vs Petal Width Graph: 4](#_Toc512546876)

[Sepal Width Vs Sepal Length Graph’s: 5](#_Toc512546877)

[References 6](#_Toc512546878)

# Problem statement

This project requires researching the Fisher’s Iris data set, and then writing documentation and code in the Python programming language based on that research. The below are the tasks to be performed as part of the project.

* Background information and summary of the Iris Data set
* Download the data set and write Python code to investigate it.
* Summarise the data set and document the investigations.
* Document the supporting tables and graphics.
* Document the references used

# R.A Fisher and Irish Data Set

**Sir Ronald Aylmer Fisher** (R.A Fisher) was a British statistician and geneticist. His work in statistics created the foundations for modern statistical science and considered as most important person in 20th century statistics.

* The key contributions of R.A Fisher are listed below
* One of the key founders of population genetics
* Fisher’s principle
* Fisherian Runaway or runaway selection
* Sexy son hypothesis
* Analysis of Variance (AVOVA)

### Irish Data Set:

The Iris flower is a multivariate data set introduced by the R.A Fisher in his 1936 paper and it is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.

Two of the three species were collected in the Gaspé Peninsula and picked on the same day and measured at the same time by the same person with the same apparatus. The dataset contains

* 150 records/samples in total for 3 species of Iris flower
* 50 samples were taken from each of Iris flower namely Iris setosa, Iris virginica and Iris versicolor.
  + Four features (Petal length, Petal width, Sepal length and Sepal width) were measured from each sample
  + The length and the width of the sepals and petals are in centimetres.

Based on the combination of these four features, R.A Fisher developed a linear discriminant model to distinguish the species from each other.

# How to run the python code and technical summary

* Download Anaconda
* Install Anaconda
* Create a "data" folder in your work directory. Copy iris data file in the "data" folder and name the file as "iris"
* Execute Iris-flowerpy in the Terminal using command "python Iris-flower.py"

The program is divided into 3 sections

* Section 1 : Using Iris Data, create the python-pandas data frame
  + Read the data from iris.csv file and Create python list for each flower with attributes
  + Create Dataframes for each flower using the List
* Section 2 : Describe the Iris flower data stored in the dataframes. Using the describe function, display the below for each flower
  + Count of the records/data for each flower
  + Mean/Average of the flower attributes Sepal width and legth; Petal width and length
  + -Standard deviations of the each attribute
  + Minimum and maximum value of the each attribute
* Section 3 : Plot the scatter, bar and histogram charts/graphs using the matplot library
  + Create a dataframe called dflower, which holds the data of all the 3 flowers
  + Create columns in the dataframe (dflower) for each flower attribute

# Data Set - Analysis and Investigation

Please find the summary of data analysis for each flower based on the given features/attributes

* Count – Number of records in Iris data file for each flower
* Mean – Average of the data
* Standard deviation (std) – How much the data differs from the mean of the data set
* Minimum (Min) – Minimum value of the data in the data set
* Maximum(Max) – Maximum value of the date in the data set

### Iris Setosa:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Petal Length | Petal Width | Sepal Length | Sepal Width |
| count | 50 | 50 | 50 | 50 |
| mean | 5.006 | 3.418 | 1.464 | 0.244 |
| std | 0.35249 | 0.381024 | 0.173511 | 0.10721 |
| min | 4.3 | 2.3 | 1 | 0.1 |
| 25% | 4.8 | 3.125 | 1.4 | 0.2 |
| 50% | 5 | 3.4 | 1.5 | 0.2 |
| 75% | 5.2 | 3.675 | 1.575 | 0.3 |
| max | 5.8 | 4.4 | 1.9 | 0.6 |

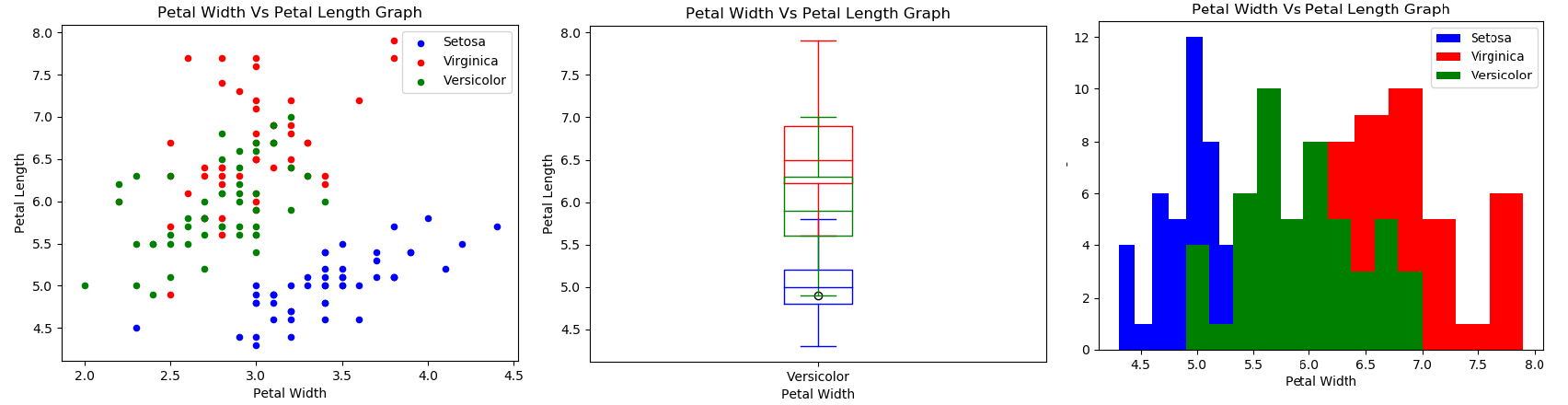
### Iris Versicolor:

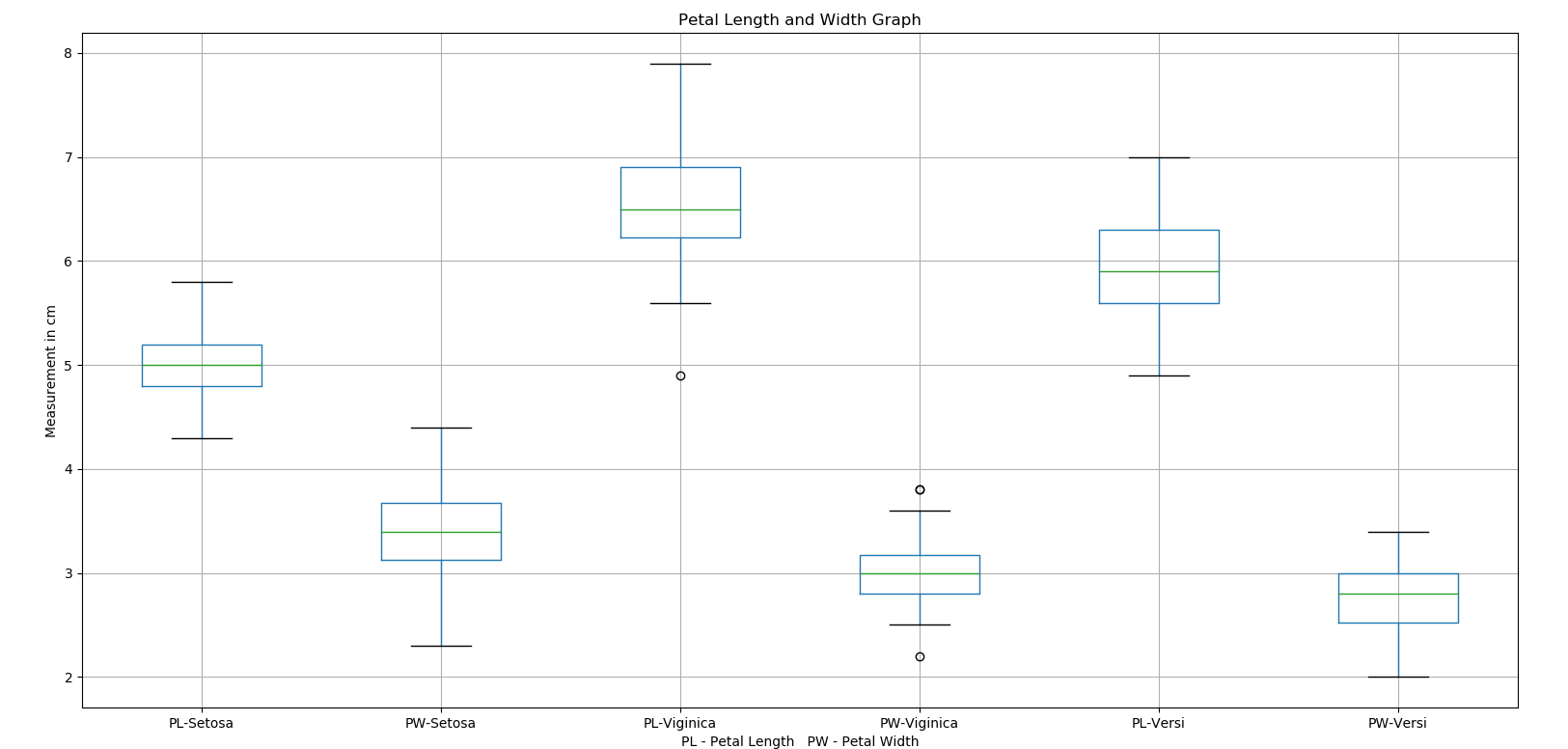
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Petal Length | Petal Width | Sepal Length | Sepal Width |
| count | 50 | 50 | 50 | 50 |
| mean | 6.588 | 2.974 | 5.552 | 2.026 |
| std | 0.63588 | 0.322497 | 0.551895 | 0.27465 |
| min | 4.9 | 2.2 | 4.5 | 1.4 |
| 25% | 6.225 | 2.8 | 5.1 | 1.8 |
| 50% | 6.5 | 3 | 5.55 | 2 |
| 75% | 6.9 | 3.175 | 5.875 | 2.3 |
| max | 7.9 | 3.8 | 6.9 | 2.5 |

### Iris Virginica:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Petal Length | Petal Width | Sepal Length | Sepal Width |
| count | 50 | 50 | 50 | 50 |
| mean | 5.936 | 2.77 | 4.26 | 1.326 |
| std | 0.516171 | 0.313798 | 0.469911 | 0.197753 |
| min | 4.9 | 2 | 3 | 1 |
| 25% | 5.6 | 2.525 | 4 | 1.2 |
| 50% | 5.9 | 2.8 | 4.35 | 1.3 |
| 75% | 6.3 | 3 | 4.6 | 1.5 |
| max | 7 | 3.4 | 5.1 | 1.8 |

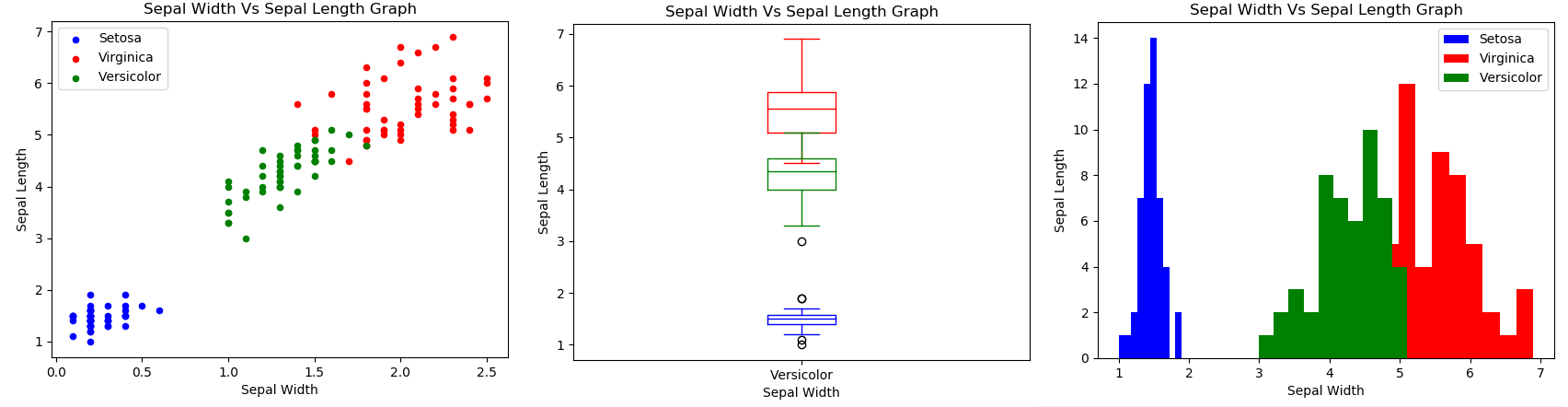
### Petal Length Vs Petal Width Graph:

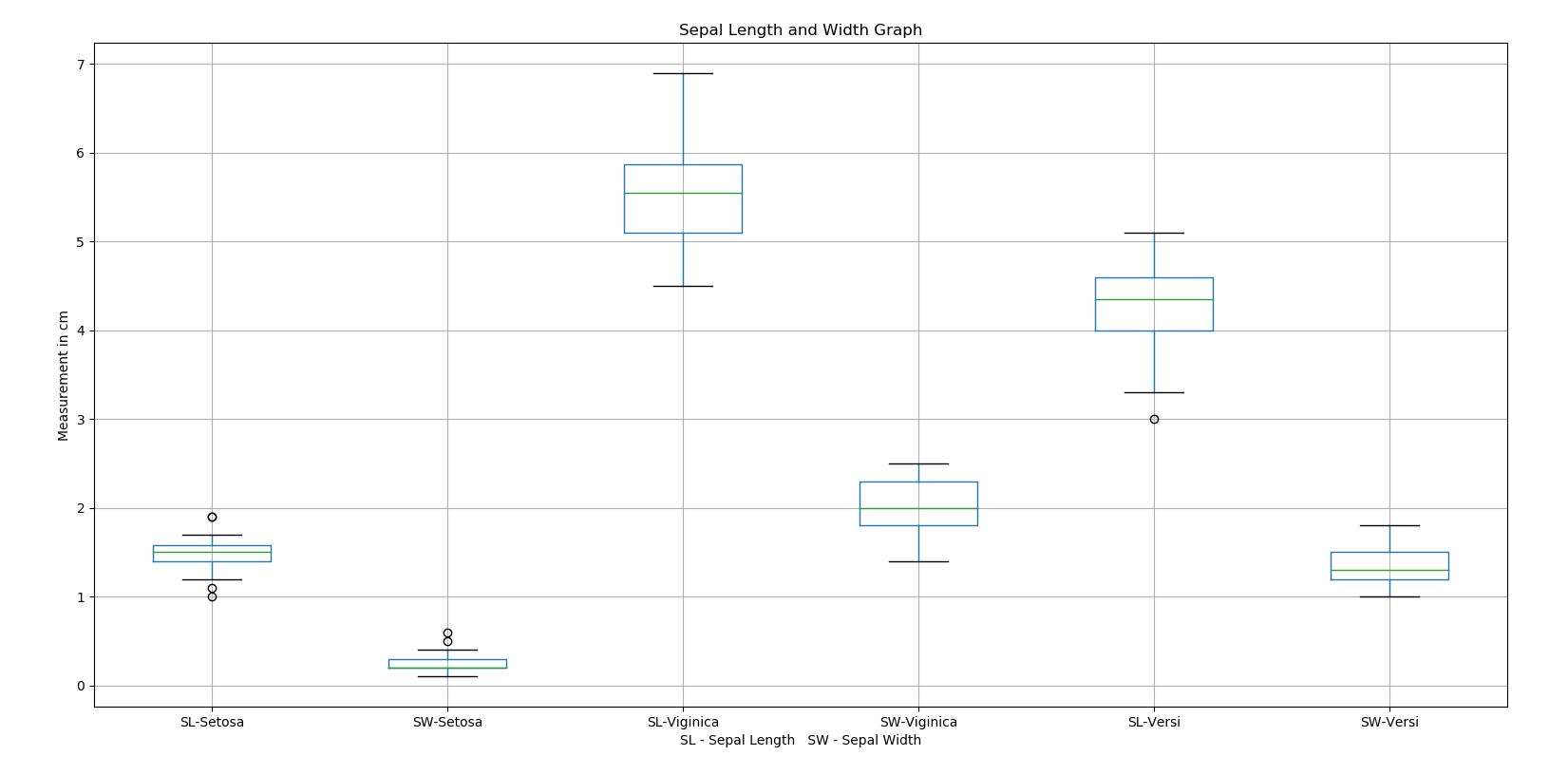




* Blue colour in the graph represents Setosa flower
* Red colour in the graph represents Viriginca flower
* Green colour in the graph represents versicolour flower
* Based on the given data,
  + Setosa attributes/features are distinct from other species of Iris flower (Viriginca and Versicolor)
  + Setosa:
    - Minimum length and Width: 4.3 cm and 2.3 cm
    - Maximum length and Width: 5.8 cm and 4.4 cm
    - No other flowers have petal length falls below 4.5
    - No other flowers have petal width falls above 4
    - Straight line/plane can be drawn to demarcate the setosa flower
  + Virginica:
    - Minimum length and Width: 4.9 cm and 2 cm
    - Maximum length and Width: 7 cm and 3.4 cm
    - No other flowers have petal length lower than 2.2
  + Versicolor:
    - Minimum length and Width: 4.9 cm and 2.2 cm
    - Maximum length and Width: 7.9 cm and 3.8 cm
    - No other flowers have petal length greater than 7 cm
* Overlap between Virginica and Versicolor flower attributes/features

### Sepal Width Vs Sepal Length Graph’s:





* Blue colour in the graph represents Setosa flower
* Red colour in the graph represents Viriginca flower
* Green colour in the graph represents versicolour flower
* Based on the given data,
  + Setosa attributes/features are distinct from other species of Iris flower (Viriginca and Versicolor)
  + Setosa:
    - Minimum length and Width: 1 cm and 0.1 cm
    - Maximum length and Width: 1.9 cm and 0.6 cm
    - No other flowers have Sepal length between 1 cm and 1.9 cm
    - No other flowers have Sepal width between 0.1 cm and 0.6 cm
    - Straight line/plane can be drawn to demarcate the setosa flower
  + Virginica:
    - Minimum length and Width: 4.5 cm and 1.4 cm
    - Maximum length and Width:6.9 cm and 2.5 cm
    - No other flowers have Sepal length more than 5.1 cm
    - No other flowers have sepal width more than 1.8 cm
  + Versicolor:
    - Minimum length and Width: 3 cm and 1 cm
    - Maximum length and Width: 5.1 cm and 1.8 cm
    - No other flowers have Sepal width between 1 cm and 1.4 cm
    - No other flowers have Sepal length between 3 cm and 4.5 cm
* Overlap between Virginica and Versicolor flower attributes/features

# References

Iris Data - Subject References:

* https://en.wikipedia.org/wiki/Iris\_flower\_data\_set
* https://en.wikipedia.org/wiki/Ronald\_Fisher
* https://en.wikipedia.org/wiki/Statistical\_classification
* https://en.wikipedia.org/wiki/Multivariate\_statistics
* https://en.wikipedia.org/wiki/Edgar\_Anderson

Technical References:

* https://www.techopedia.com/definition/32880/iris-flower-data-set
* https://www.youtube.com/watch?v=azXCzI57Yfc
* https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
* https://pythonprogramming.net/support-vector-machine-intro-machine-learning-tutorial/
* http://scikit-learn.org/stable/auto\_examples/datasets/plot\_iris\_dataset.html
* https://stackoverflow.com/
* http://pandas.pydata.org/pandas-docs/version
* https://www.tutorialspoint.com/python\_pandas/python\_pandas\_series.htm
* https://matplotlib.org/api/pyplot\_api.html
* https://matplotlib.org/examples/index.html
* https://stackoverflow.com/questions/4270301/
* https://www.datascience.com/learn-data-science/tutorials/creating-data-visualizations-matplotlib-data-science-python