Fisher’s Iris Data set

SOMANATHAN SUBRAMANIYAN Student Id: G00364742

Programming and Scripting

Project 2018

2018

Table of Contents

[Problem statement 2](#_Toc511580734)

[R.A Fisher and Irish Data Set 2](#_Toc511580735)

[Irish Data Set: 2](#_Toc511580736)

[Data Set - Analysis and Investigation 3](#_Toc511580737)

[Iris Setosa: 3](#_Toc511580738)

[Iris Versicolor: 3](#_Toc511580739)

[Iris Virginica: 3](#_Toc511580740)

[Petal Graph’s: 4](#_Toc511580741)

[Sepal Graph’s: 4](#_Toc511580742)

[References 5](#_Toc511580743)

# 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 information and programming 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 in total for 3 species of iris flower
* 50 samples from each of Iris flower namely Iris setosa, Iris virginica and Iris versicolor.

Four features were measured from each sample

* Petal length
* Petal width
* Sepal length
* Sepal width

The length and the width of the sepals and petals, 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.

# Data Set - Analysis and Investigation

Please find the summary of initial data analysis for each flower

* Count – Number of records in the Iris data file
* Mean – Average of the data set
* 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 |

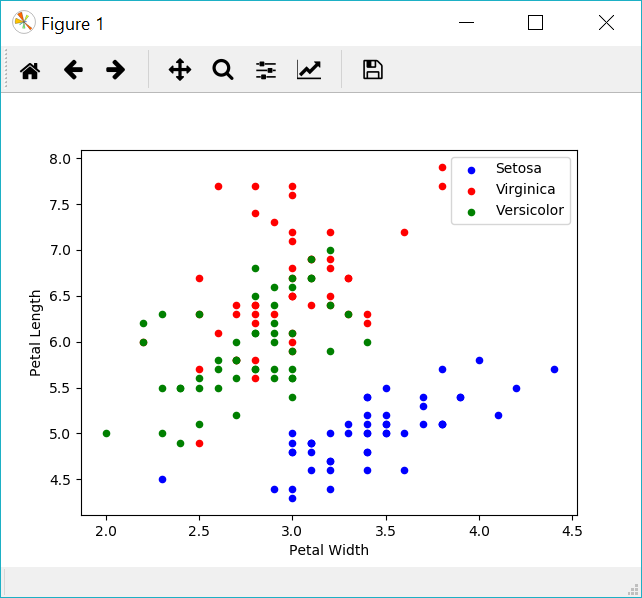
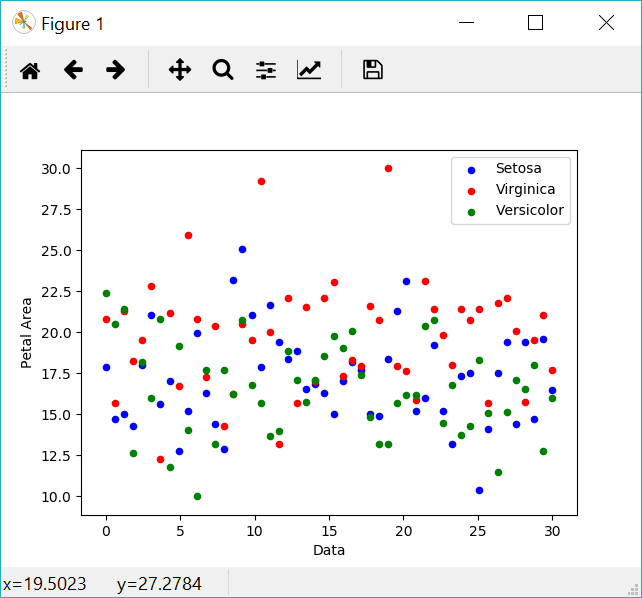
### Iris Versicolor:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

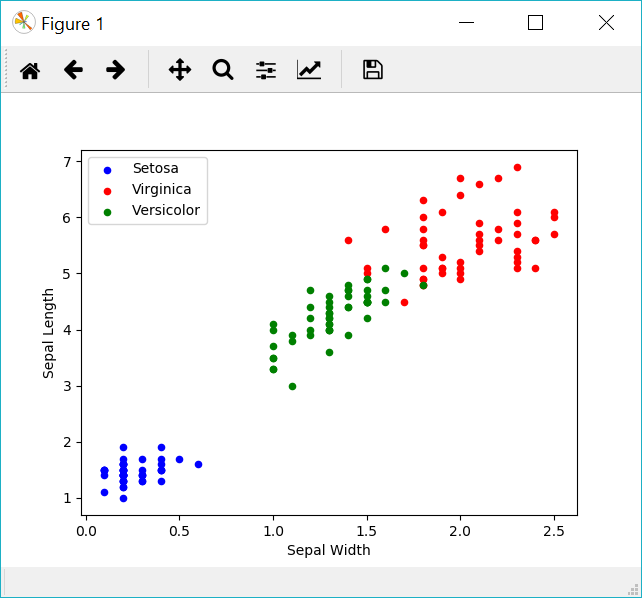
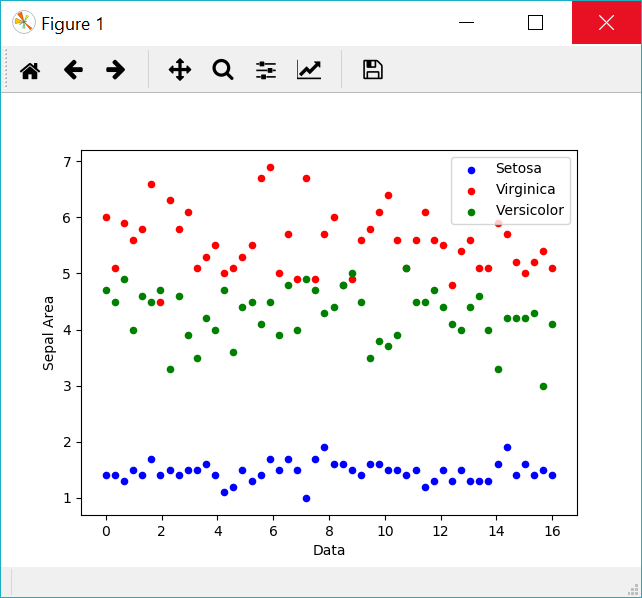
**Petal Length Vs Petal Width Petal Area Graph**

### Petal Graph’s:

* Clearly distinguish the Setosa flower against Virginica and Versicolor
* Setosa: The Sepal width and length are distinct from Virginica and Versicolor
  + Minimum length and Width: 4.3 cm and 2.3 cm
  + Maximum length and Width: 5.8 cm and 4.4 cm
* Overlap between Virginica and Versicolor

**Sepal Length Vs Sepal Width Sepal Area Graph**

### Sepal Graph’s:

* Clearly distinguish the Setosa flower against Virginica and Versicolor
* Setosa: The Sepal width and length are distinct from Virginica and Versicolor
  + Minimum length and Width: 1 cm and 0.1 cm
  + Maximum length and Width: 1.9 cm and 0.6 cm
* Small overlap between Virginica and Versicolor

# References

Technical Reference:

* 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/
* 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
* 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