**Project for programming and scripting data analytics GMIT 2018**

**Fishing Fisher's Flower Figures - An exploration of the Iris Dataset**

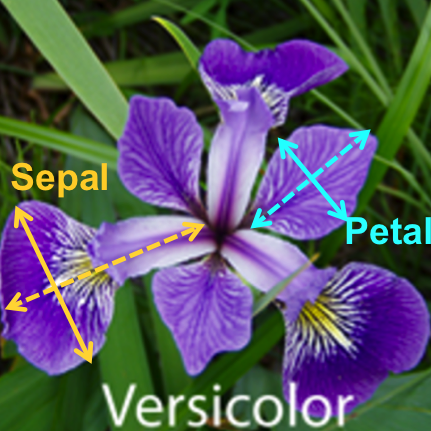
**Abstract**

This report explores the iris dataset using data analytics techniques. The iris data consists of 50 samples from three species of iris namely iris setosa, iris virgincia and iris versicolor. It consists of measures of the flowers sepal and petals’ length and width in centimeters. The dataset was analysed using three approaches, descriptive statistics, graphical visualisations and inferential statistics. The method used was statistical analysis using python and python packages. Results showed clear differences between one species and the other two when the data is unlabelled and further separation of data when all species are labelled.

**Introduction**

The iris data set, also known as Fisher or Anderson’s iris flower data set, was popularised by the statistician Ronald Fisher in his 1936 paper “The use of multiple measurements in taxonomic problems” (Fisher 1936). It was collected by the American botanist, Edgar Anderson in 1935 yet remained unpublished till Fisher published his paper which used this data set (ibid). Anderson’s data consists of fifty samples each, from the three species of iris, iris setosa, iris virginica and iris versicolor. Two of the species, iris virginica and iris setosa, were sampled from the same region and used by Fisher to illustrate discriminant functions (ibid). In addition Fisher extended this method to investigate Randolph’s hypothesis that the third species, iris versicolor, was a hybrid of the iris virginica and setosa species (ibid). Anderson recorded species type and measured the length and width of their petals and sepals in centimetres. The petals are the inner flower while the sepals are the outer structures as seen in image 1.

Image 1



See Image 2, 3 and 4 for images of iris setosa, iris virginica and iris versicolor respectively.

Image 2 Iris Setosa



Image 3 Iris Virginica



Image 3 Iris Versicolor

.

This gives the data set five dimensions, namely 4 one dimensional measures in centimetres, sepal length (sepalL), sepal width (sepalW), petal length (petalL), petal width (petalW); And one categorical dimension of iris species (species). This project used the version of the iris data hosted at UCI machine learning repository (Iris Data set, Iris Data, donated 1988). This version contains two errors from Anderson’s original data which were not amended in the analysis (Bezdek, Keller et al., 1999).

Fisher’s analysis investigated if petal/sepal measurements alone could predict which species of iris the sample came from (Fisher, 1936). This makes the data useful in exploring machine learning methods, statistical techniques and data visualisation. The iris data set is well known, and often cited. Iris Data Set (Donated, 1988) notes 99 citations with citations from as recently as 2005. The set has historical significance, as does Fisher, and is widely recognised in computer science. Fisher’s iris data set is often used as a learning tool in data analytics. The data is small enough to be manageable for beginners yet sufficiently challenging in what it can reveal. Its historical use means that there is a body of work and continuity based on it, which can be used as a benchmark to test program results and explore data analytic methods.

Fisher's analysis showed two main clusters, with the iris setosa petal and sepal measures being smaller than the iris virginica. Iris versicolor was intermediate between the other two species with some overlap of measures with iris virginica. This separation into two groups is an example of unsupervised clustering in Machine Learning and artificial Neural Networks. Once the samples are labelled by species then three clusters emerge illustrating supervised clustering (reference).

This report explored the Iris Data using the programming and scripting language Python Version 3.6. In addition GitHub was as the development platform for this project. Python is fast becoming the standard tool for data analysis. It is free and open source, unlike SPSS or SAS for example. It also has a more intuitive syntax than R for example. In addition Python has a wide selection of library packages that can be used to investigate particular aspects of the data. For example the Pandas package offers data frame manipulation and table reading tools; NumPy offers numerical and scientific computing; Matplotlib extends NumPy to include plotting ability and Seaborn extends Matplotlib to allows statistical data to be visualised.

This project aimed to explore the iris data as hosted at UCI. Firstly descriptive data was calculated using Python. This included the maximum, minimum and mean of the petal and sepal measures. Then various visualisations were produced. Finally introductory inferential statistics were investigated.

**Method**

This report produced descriptive, inferential and graphical visualisations of the Iris data set as hosted at UCI (Iris Data, Donated 1988).

**Data**

This project used a copy of the Iris dataset available from UCI (Data Set, Donated 1988). A sample of which can be viewed below in Table 1a. The full table is available in [Table 1](#table1) ([Appendix 1](#appendix1)).

Table 1a

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | sepalL | sepalW | petalL | petalW | name |
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 5 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 6 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 7 | 5.0 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 8 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 9 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |

**Apparatus and Materials**

A computer with python 3.6 software installed was used. An internet connection was required for conducting research. In addition the python packages, NumPy, Pandas, Matplotlib and Seaborn were used. GitHub in conjunction with the IDE Visual Studio Code was used for program development.

**Procedure**

The data was downloaded from UCI iris.csv (Data Set, Donated 1989) as a cvs file and stored in a local folder as /data/iris.csv. The Python package Pandas was used to import the data as a dataframe and to perform initial descriptive analysis. Firstly the shape of the data was looked at and the first 10 entries. Then descriptive statistics were produced (mean, maximum and minimum values of the measures). The data was visualised using the NumPy addon packages Matplotlib and Seaborn. Inferential statistics were produced using the package xx and yy.

**Results**

The descriptive statistics table 2 shows that the mean sepal length was ….

Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sepalL | sepalW | petalL | petalW |
| count | 150 | 150 | 150 | 150 |
| mean | 5.843 | 3.054 | 3.759 | 1.199 |
| std | 0.828 | 0.434 | 1.76 | 0.763 |
| min | 4.3 | 2 | 1 | 0.1 |
| 25% | 5.1 | 2.8 | 1.6 | 0.3 |
| 50% | 5.8 | 3 | 4.35 | 1.3 |
| 75% | 6.4 | 3.3 | 5.1 | 1.8 |
| max | 7.9 | 4.4 | 6.9 | 2.5 |

put the findings here printouts graphs scatter plots etc

**Discussion**

discuss the findings maybe a discussion of cluster analysis data mining unstructured data and structured data mabye do colourful plots like maps?

from the

**References**

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7, pt. II, 197-188

“Iris Data Set”, (donated 1988). Retrieved from <http://archive.ics.uci.edu/ml/datasets/Iris> on 1/4/2018

“Iris Data” (donated 1988) Retrieved from [http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data on 22/3/2018](http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data%20on%2022/3/2018)

J. C. Bezdek, J. M. Keller, R. Krishnapuram, L. I. Kuncheva and N. R. Pal, (1999) "Will the real iris data please stand up?," in *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 3, pp. 368-369. doi: 10.1109/91.771092

[wiki](https://en.wikipedia.org/wiki/Iris\_flower\_data\_set) [stack exchange](https://stats.stackexchange.com/questions/30788/whats-a-good-way-to-use-r-to-make-a-scatterplot-that-separates-the-data-by-trea/30789#30789) [link](https://stats.stackexchange.com/questions/74776/what-aspects-of-the-iris-data-set-make-it-so-successful-as-an-example-teaching) [link](https://archive.ics.uci.edu/ml/datasets/iris) [](https://www.kaggle.com/sridharcr/data-analysis-iris-dataset) [](https://www.kaggle.com/benhamner/python-data-visualizations) [](http://scikit-learn.org/stable/tutorial/basic/tutorial.html) []() #

**Appendices**

**Appendix 1**

Table 1 The Iris Data Set (measures in cm)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | sepalL | sepalW | petalL | petalW | name |
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 1 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 4 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 5 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 6 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 7 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 8 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 9 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 10 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| 11 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| 12 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa |
| 13 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa |
| 14 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa |
| 15 | 5.7 | 4.4 | 1.5 | 0.4 | Iris-setosa |
| 16 | 5.4 | 3.9 | 1.3 | 0.4 | Iris-setosa |
| 17 | 5.1 | 3.5 | 1.4 | 0.3 | Iris-setosa |
| 18 | 5.7 | 3.8 | 1.7 | 0.3 | Iris-setosa |
| 19 | 5.1 | 3.8 | 1.5 | 0.3 | Iris-setosa |
| 20 | 5.4 | 3.4 | 1.7 | 0.2 | Iris-setosa |
| 21 | 5.1 | 3.7 | 1.5 | 0.4 | Iris-setosa |
| 22 | 4.6 | 3.6 | 1 | 0.2 | Iris-setosa |
| 23 | 5.1 | 3.3 | 1.7 | 0.5 | Iris-setosa |
| 24 | 4.8 | 3.4 | 1.9 | 0.2 | Iris-setosa |
| 25 | 5 | 3 | 1.6 | 0.2 | Iris-setosa |
| 26 | 5 | 3.4 | 1.6 | 0.4 | Iris-setosa |
| 27 | 5.2 | 3.5 | 1.5 | 0.2 | Iris-setosa |
| 28 | 5.2 | 3.4 | 1.4 | 0.2 | Iris-setosa |
| 29 | 4.7 | 3.2 | 1.6 | 0.2 | Iris-setosa |
| 30 | 4.8 | 3.1 | 1.6 | 0.2 | Iris-setosa |
| 31 | 5.4 | 3.4 | 1.5 | 0.4 | Iris-setosa |
| 32 | 5.2 | 4.1 | 1.5 | 0.1 | Iris-setosa |
| 33 | 5.5 | 4.2 | 1.4 | 0.2 | Iris-setosa |
| 34 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 35 | 5 | 3.2 | 1.2 | 0.2 | Iris-setosa |
| 36 | 5.5 | 3.5 | 1.3 | 0.2 | Iris-setosa |
| 37 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 38 | 4.4 | 3 | 1.3 | 0.2 | Iris-setosa |
| 39 | 5.1 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 40 | 5 | 3.5 | 1.3 | 0.3 | Iris-setosa |
| 41 | 4.5 | 2.3 | 1.3 | 0.3 | Iris-setosa |
| 42 | 4.4 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 43 | 5 | 3.5 | 1.6 | 0.6 | Iris-setosa |
| 44 | 5.1 | 3.8 | 1.9 | 0.4 | Iris-setosa |
| 45 | 4.8 | 3 | 1.4 | 0.3 | Iris-setosa |
| 46 | 5.1 | 3.8 | 1.6 | 0.2 | Iris-setosa |
| 47 | 4.6 | 3.2 | 1.4 | 0.2 | Iris-setosa |
| 48 | 5.3 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| 49 | 5 | 3.3 | 1.4 | 0.2 | Iris-setosa |
| 50 | 7 | 3.2 | 4.7 | 1.4 | Iris-versicolor |
| 51 | 6.4 | 3.2 | 4.5 | 1.5 | Iris-versicolor |
| 52 | 6.9 | 3.1 | 4.9 | 1.5 | Iris-versicolor |
| 53 | 5.5 | 2.3 | 4 | 1.3 | Iris-versicolor |
| 54 | 6.5 | 2.8 | 4.6 | 1.5 | Iris-versicolor |
| 55 | 5.7 | 2.8 | 4.5 | 1.3 | Iris-versicolor |
| 56 | 6.3 | 3.3 | 4.7 | 1.6 | Iris-versicolor |
| 57 | 4.9 | 2.4 | 3.3 | 1 | Iris-versicolor |
| 58 | 6.6 | 2.9 | 4.6 | 1.3 | Iris-versicolor |
| 59 | 5.2 | 2.7 | 3.9 | 1.4 | Iris-versicolor |
| 60 | 5 | 2 | 3.5 | 1 | Iris-versicolor |
| 61 | 5.9 | 3 | 4.2 | 1.5 | Iris-versicolor |
| 62 | 6 | 2.2 | 4 | 1 | Iris-versicolor |
| 63 | 6.1 | 2.9 | 4.7 | 1.4 | Iris-versicolor |
| 64 | 5.6 | 2.9 | 3.6 | 1.3 | Iris-versicolor |
| 65 | 6.7 | 3.1 | 4.4 | 1.4 | Iris-versicolor |
| 66 | 5.6 | 3 | 4.5 | 1.5 | Iris-versicolor |
| 67 | 5.8 | 2.7 | 4.1 | 1 | Iris-versicolor |
| 68 | 6.2 | 2.2 | 4.5 | 1.5 | Iris-versicolor |
| 69 | 5.6 | 2.5 | 3.9 | 1.1 | Iris-versicolor |
| 70 | 5.9 | 3.2 | 4.8 | 1.8 | Iris-versicolor |
| 71 | 6.1 | 2.8 | 4 | 1.3 | Iris-versicolor |
| 72 | 6.3 | 2.5 | 4.9 | 1.5 | Iris-versicolor |
| 73 | 6.1 | 2.8 | 4.7 | 1.2 | Iris-versicolor |
| 74 | 6.4 | 2.9 | 4.3 | 1.3 | Iris-versicolor |
| 75 | 6.6 | 3 | 4.4 | 1.4 | Iris-versicolor |
| 76 | 6.8 | 2.8 | 4.8 | 1.4 | Iris-versicolor |
| 77 | 6.7 | 3 | 5 | 1.7 | Iris-versicolor |
| 78 | 6 | 2.9 | 4.5 | 1.5 | Iris-versicolor |
| 79 | 5.7 | 2.6 | 3.5 | 1 | Iris-versicolor |
| 80 | 5.5 | 2.4 | 3.8 | 1.1 | Iris-versicolor |
| 81 | 5.5 | 2.4 | 3.7 | 1 | Iris-versicolor |
| 82 | 5.8 | 2.7 | 3.9 | 1.2 | Iris-versicolor |
| 83 | 6 | 2.7 | 5.1 | 1.6 | Iris-versicolor |
| 84 | 5.4 | 3 | 4.5 | 1.5 | Iris-versicolor |
| 85 | 6 | 3.4 | 4.5 | 1.6 | Iris-versicolor |
| 86 | 6.7 | 3.1 | 4.7 | 1.5 | Iris-versicolor |
| 87 | 6.3 | 2.3 | 4.4 | 1.3 | Iris-versicolor |
| 88 | 5.6 | 3 | 4.1 | 1.3 | Iris-versicolor |
| 89 | 5.5 | 2.5 | 4 | 1.3 | Iris-versicolor |
| 90 | 5.5 | 2.6 | 4.4 | 1.2 | Iris-versicolor |
| 91 | 6.1 | 3 | 4.6 | 1.4 | Iris-versicolor |
| 92 | 5.8 | 2.6 | 4 | 1.2 | Iris-versicolor |
| 93 | 5 | 2.3 | 3.3 | 1 | Iris-versicolor |
| 94 | 5.6 | 2.7 | 4.2 | 1.3 | Iris-versicolor |
| 95 | 5.7 | 3 | 4.2 | 1.2 | Iris-versicolor |
| 96 | 5.7 | 2.9 | 4.2 | 1.3 | Iris-versicolor |
| 97 | 6.2 | 2.9 | 4.3 | 1.3 | Iris-versicolor |
| 98 | 5.1 | 2.5 | 3 | 1.1 | Iris-versicolor |
| 99 | 5.7 | 2.8 | 4.1 | 1.3 | Iris-versicolor |
| 100 | 6.3 | 3.3 | 6 | 2.5 | Iris-virginica |
| 101 | 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| 102 | 7.1 | 3 | 5.9 | 2.1 | Iris-virginica |
| 103 | 6.3 | 2.9 | 5.6 | 1.8 | Iris-virginica |
| 104 | 6.5 | 3 | 5.8 | 2.2 | Iris-virginica |
| 105 | 7.6 | 3 | 6.6 | 2.1 | Iris-virginica |
| 106 | 4.9 | 2.5 | 4.5 | 1.7 | Iris-virginica |
| 107 | 7.3 | 2.9 | 6.3 | 1.8 | Iris-virginica |
| 108 | 6.7 | 2.5 | 5.8 | 1.8 | Iris-virginica |
| 109 | 7.2 | 3.6 | 6.1 | 2.5 | Iris-virginica |
| 110 | 6.5 | 3.2 | 5.1 | 2 | Iris-virginica |
| 111 | 6.4 | 2.7 | 5.3 | 1.9 | Iris-virginica |
| 112 | 6.8 | 3 | 5.5 | 2.1 | Iris-virginica |
| 113 | 5.7 | 2.5 | 5 | 2 | Iris-virginica |
| 114 | 5.8 | 2.8 | 5.1 | 2.4 | Iris-virginica |
| 115 | 6.4 | 3.2 | 5.3 | 2.3 | Iris-virginica |
| 116 | 6.5 | 3 | 5.5 | 1.8 | Iris-virginica |
| 117 | 7.7 | 3.8 | 6.7 | 2.2 | Iris-virginica |
| 118 | 7.7 | 2.6 | 6.9 | 2.3 | Iris-virginica |
| 119 | 6 | 2.2 | 5 | 1.5 | Iris-virginica |
| 120 | 6.9 | 3.2 | 5.7 | 2.3 | Iris-virginica |
| 121 | 5.6 | 2.8 | 4.9 | 2 | Iris-virginica |
| 122 | 7.7 | 2.8 | 6.7 | 2 | Iris-virginica |
| 123 | 6.3 | 2.7 | 4.9 | 1.8 | Iris-virginica |
| 124 | 6.7 | 3.3 | 5.7 | 2.1 | Iris-virginica |
| 125 | 7.2 | 3.2 | 6 | 1.8 | Iris-virginica |
| 126 | 6.2 | 2.8 | 4.8 | 1.8 | Iris-virginica |
| 127 | 6.1 | 3 | 4.9 | 1.8 | Iris-virginica |
| 128 | 6.4 | 2.8 | 5.6 | 2.1 | Iris-virginica |
| 129 | 7.2 | 3 | 5.8 | 1.6 | Iris-virginica |
| 130 | 7.4 | 2.8 | 6.1 | 1.9 | Iris-virginica |
| 131 | 7.9 | 3.8 | 6.4 | 2 | Iris-virginica |
| 132 | 6.4 | 2.8 | 5.6 | 2.2 | Iris-virginica |
| 133 | 6.3 | 2.8 | 5.1 | 1.5 | Iris-virginica |
| 134 | 6.1 | 2.6 | 5.6 | 1.4 | Iris-virginica |
| 135 | 7.7 | 3 | 6.1 | 2.3 | Iris-virginica |
| 136 | 6.3 | 3.4 | 5.6 | 2.4 | Iris-virginica |
| 137 | 6.4 | 3.1 | 5.5 | 1.8 | Iris-virginica |
| 138 | 6 | 3 | 4.8 | 1.8 | Iris-virginica |
| 139 | 6.9 | 3.1 | 5.4 | 2.1 | Iris-virginica |
| 140 | 6.7 | 3.1 | 5.6 | 2.4 | Iris-virginica |
| 141 | 6.9 | 3.1 | 5.1 | 2.3 | Iris-virginica |
| 142 | 5.8 | 2.7 | 5.1 | 1.9 | Iris-virginica |
| 143 | 6.8 | 3.2 | 5.9 | 2.3 | Iris-virginica |
| 144 | 6.7 | 3.3 | 5.7 | 2.5 | Iris-virginica |
| 145 | 6.7 | 3 | 5.2 | 2.3 | Iris-virginica |
| 146 | 6.3 | 2.5 | 5 | 1.9 | Iris-virginica |
| 147 | 6.5 | 3 | 5.2 | 2 | Iris-virginica |
| 148 | 6.2 | 3.4 | 5.4 | 2.3 | Iris-virginica |
| 149 | 5.9 | 3 | 5.1 | 1.8 | Iris-virginica |