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

**Helen O’Shea**

**Galway Mayo Institute of Technology**

Table of Contents

[**Fishing Fisher's Flower Figures - An exploration of the Iris Dataset** 1](#_Toc511219490)

[**Helen O’Shea** 1](#_Toc511219491)

[**Galway Mayo Institute of Technology** 1](#_Toc511219492)

[**Abstract** 3](#_Toc511219493)

[**Introduction** 3](#_Toc511219494)

[**Method** 5](#_Toc511219495)

[**Results** 7](#_Toc511219496)

[**Discussion** 26](#_Toc511219497)

[**References** 29](#_Toc511219498)

[**Appendix** 29](#_Toc511219499)

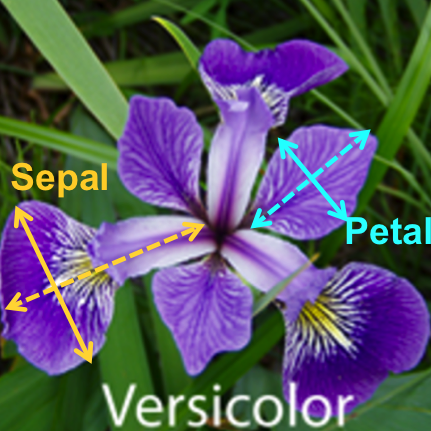
**Abstract**

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 centimetres. The dataset was analysed using three approaches, descriptive statistics, graphical visualisations and inferential statistics. The method used was statistical analysis using python and python modules. 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 It was collected by the American botanist, Edgar Anderson in 1935 yet remained unpublished till Fisher published his paper which used this data set (ibid).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 figure 1.

*Figure 1*



See figures 2, 3 and 4 for images of Iris setosa, Iris virginica and Iris versicolor respectively.

*Figure 2* Iris Setosa



*Figure 3* Iris Virginica



*Figure 4* 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 (name). 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 modules that can be used to investigate particular aspects of the data. For example the Pandas module 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 for each of the species of Iris. 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 modules, 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 module 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 for each species of Iris (mean, maximum and minimum values of the measures). The data was visualised using the NumPy extension modules Matplotlib and Seaborn. Inferential statistics were produced using the module xx and yy.

**Results**

#### *Descriptive statistics*

Add a paragraph on the ratios of length/width

The descriptive statistics were taken by running the script project/pyscripts/desc.py. Results were saved in folder project/data/ with data from desc\_se.csv shown in table 2; desc\_vi shown in table 3 and desc\_ve shown in table 4.

The descriptive statistics for the Iris Setosa is shown in table 2.

*Table 2* Iris Setosa descriptive statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sepalL | sepalW | petalL | petalW |
| count | 50 | 50 | 50 | 50 |
| mean | 5.006 | 3.418 | 1.464 | 0.244 |
| std | 0.352 | 0.381 | 0.174 | 0.107 |
| 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 |

This shows that the data consisted of 50 samples with measures taken of the Iris’s sepal and petals length and width. The mean was smallest for the petals, with the width having the smallest mean. The sepal length had the highest mean and was over three times larger than the petals length. Although the sepal width was smaller than its length, it was over 14 times bigger than the corresponding value for the petal. The sepal length had the largest maximum and minimum value of all the measures while the petal width showed the smallest maximum and minimum value. The standard deviation (std) showed that the sepal values were more spread out than the petal values. The table also shows the percent quartiles.

Table 3 below shows the corresponding results for the Iris Virginica samples

*Table 3* Iris Virginica descriptive statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sepalL | sepalW | petalL | petalW |
| count | 50 | 50 | 50 | 50 |
| mean | 6.588 | 2.974 | 5.552 | 2.026 |
| std | 0.636 | 0.322 | 0.552 | 0.275 |
| 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 |

Here the mean was smallest for the petal width and largest for the sepal length as per the Iris setosa. However the ratios were less striking with the length of the sepal being just under 20% bigger than the petal’s length and the corresponding widths being 45% bigger. The maximum and minimum values were higher than the Iris sertosa’s but showed the same pattern of the sepals being longer and wider with the exception of the sepal width. The spread of the data was greater than the Iris setosa over all measures excepting the sepal width. All quartiles except the sepal width were greater than the Iris setosa quartiles.

Table 4 below shows the results for the Iris Versicolor samples

*Table 4* Iris Versicolor descriptive statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | sepalL | sepalW | petalL | petalW |
| count | 50 | 50 | 50 | 50 |
| mean | 5.936 | 2.77 | 4.26 | 1.326 |
| std | 0.516 | 0.314 | 0.470 | 0.198 |
| 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 |

The same pattern again was found of the mean being smaller for the petal dimensions compared to the sepal corresponding dimensions. The largest mean was for the sepal length and smallest for the petal width as before with the values falling between the lower setosa measures and higher virginica ones. The maximum value was for the sepal length which was just under the maximum for the virginica species. The minimum measure was the petal width similar to the other species. The sepal width maximum and minimum values differed from the rest of the data in that the were smaller than the Iris virginica and setosa whereas the other measures lay between the setosa and virginica Iris species. The measure of spread was slightly lower than for the Iris virginica but higher than the Iris setosa. The quartiles lay between the lower setosa and higher virginica except for the sepal width which had the lowest quartiles out of all the samples.

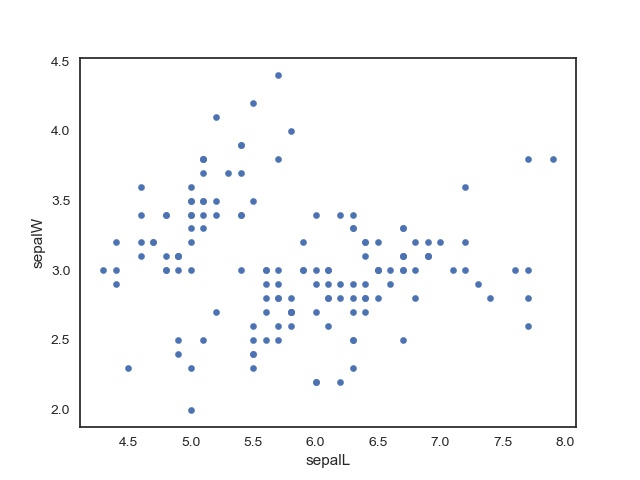
The ratio of sepal length to width and petal length to width were also calculated. See table 6

*Table 6* Ratios of Length to Widths of Iris Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ratio L/W | Setosa  Sepal | Setosa  Petal | Vericolor  Sepal | Vericolor  Petal | Virginica  Sepal | Virginica  Petal |
| count | 50 | 50 | 50 | 50 | 50 | 50 |
| mean | 1.475 | 7.078 | 2.160 | 3.243 | 2.230 | 2.781 |
| std | 0.119 | 3.124 | 0.229 | 0.312 | 0.247 | 0.407 |
| min | 1.268 | 2.667 | 1.765 | 2.667 | 1.824 | 2.125 |
| 25% | 1.395 | 4.688 | 2.034 | 3.017 | 2.032 | 2.511 |
| 50% | 1.468 | 7.000 | 2.161 | 3.240 | 2.170 | 2.667 |
| 75% | 1.548 | 7.875 | 2.232 | 3.418 | 2.343 | 3.056 |
| max | 1.957 | 15.0 | 2.818 | 4.1 | 2.962 | 4.0 |

#### *Visualisations*

Visualisations of the data were produced by running /project/pyscripts/visu.py. Results were saved in /project/graphs/. Figure 5 and 6 shows the initial scatter plot of the sepal measures and petal measures respectively via matplotlib.pyplot. Script figures are saved in the folder /graphs/.

*Figure 5* Scatter plot of the Iris’s sepal length versus widths in cm

This shows that sepal lengths tend to be greater than sepal widths. The figure also suggests that sepal widths are negatively correlated to lengths. The data also suggests smaller lengths have a steeper incline compared to larger lengths. Figure 6 shows the corresponding plot of the petal lengths.

*Figure 6* Scatter plot of the Iris’s sepal length versus widths in cm

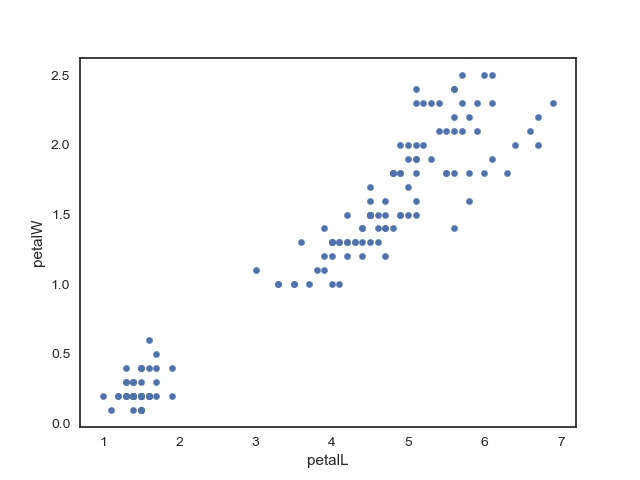


Figure 6 shows a similar pattern of larger lengths tending to have larger widths. Two distinct clusters appear to be present two clusters. The data appears to have a liner relationship and a strong positive correlation between petal lengths and widths. Figure 7 and 8 confirms this correlation using Pearson’s r.

*Figure 7* scatter plot of sepal lengths and widths in cm with correlation

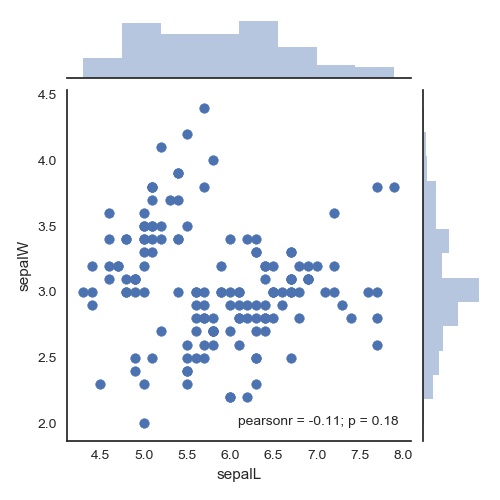


Figure 7 and 8 were produced using the seaborn module. Figure 7 shows Pearson’s r is -0.11 indicating a slight negative corralation between sepal lengths and widths.

*Figure 8* scatter plot of petal lengths and widths in cm with correlation

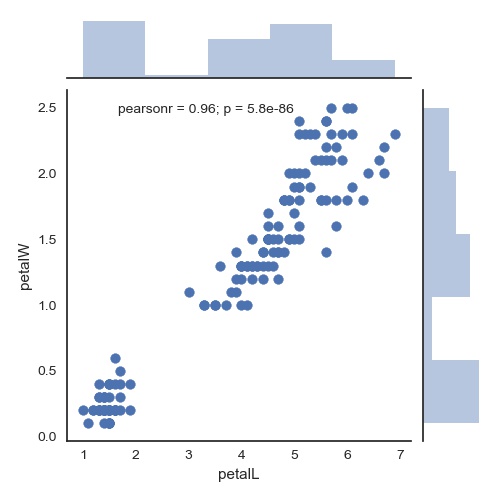
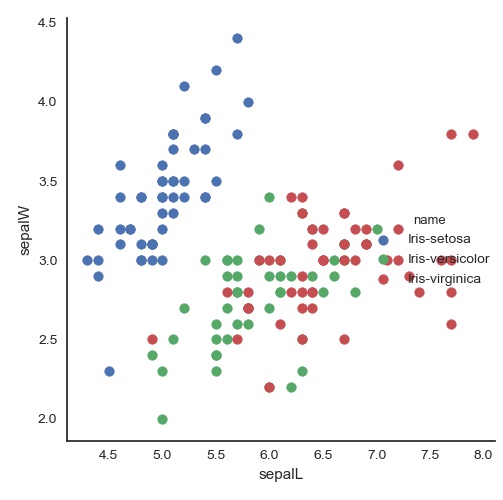


Figure 8 shows Pearson’s at 0.96 indicating an almost perfectly positive linear relationship between petal lengths and widths.

The seaborn module was used to explore this difference between sepals and petals by identifying each Iris species within the scatter plot as shown in figure 9 and 10.

*Figure 9* scatter plot of sepal lengths and widths in cm labelled by Iris name



This shows a clear cluster amongst the Iris setosa (in blue) in sepal dimensions. The Iris setosa appears to show a strong positive correlation in sepal dimensions that is absent in the other two sepcies. Clusters are less pronounced between the Iris versicolor (green) and Iris virginica (red) although it appears that the virginica have larger sepal dimensions than the versicolor with some degree of overlap around the (6.25, 2.75) position.

*Figure 10* scatter plot of petal lengths and widths in cm labelled by Iris name

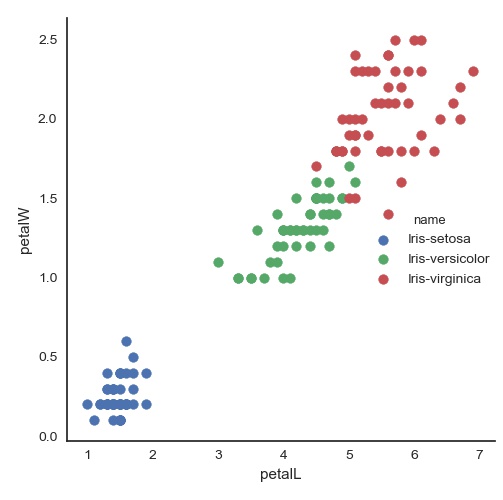


Figure 10 shows three clusters. The Iris setosa has clearly smaller petal dimensions than the other two, with the Iris versicolor falling linearly between the Iris setosa and Iris virginica on the Iris virginica side.

Figures 11-15 show the box plots of the sepal and petal lengths and widths sorted by Iris name.

*Figure 11* Boxplot of sepal lengths cm

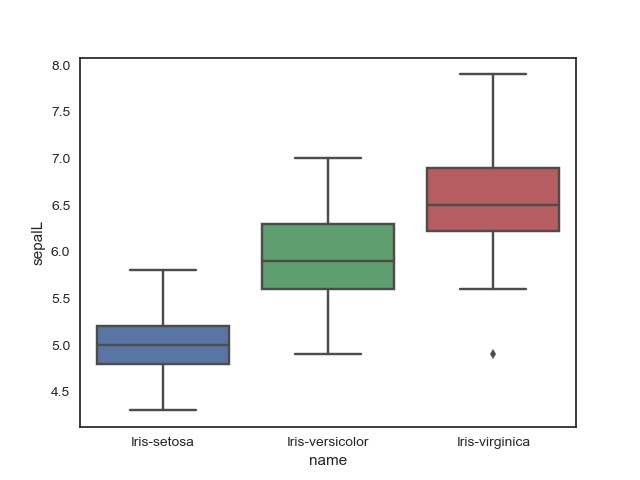


Figure 11 shows that sepal length quartiles are smallest for the Iris setosa and largest for the Iris virginica with the Iris versicolor between them. The quartile spread is similar for the Iris virginica and Iris versicolor while narrower for the Iris setosa. The maximum value of the Iris setosa and the minimum of Iris virginica barely overlap. The maximum value of the Iris virginica is further from its median compared to the other species of Iris.

*Figure 12* Boxplot of sepal widths cm

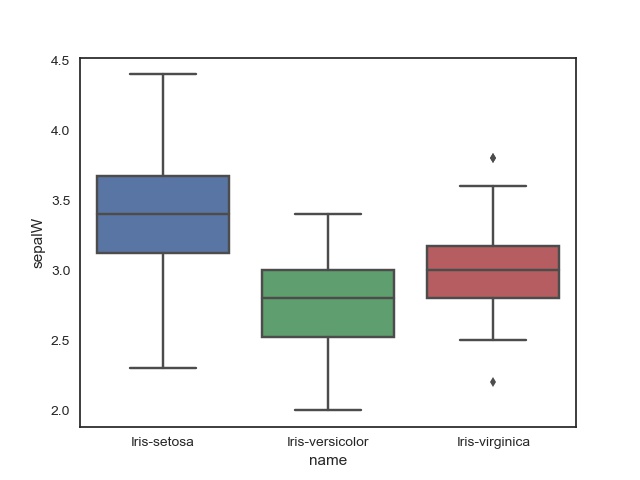


Figure 12 shows that the Iris setosa width has the larger quartiles than the other two with a greater spread between its maximum and minimum. The Iris veriscolor as the lowest quartiles while the Iris virginica lies between the Iris setosa and Iris versicolor. The Iris virginica also shows outliers in its highest and lowest sepal width.

*Figure 13* Boxplot of petal lengths cm

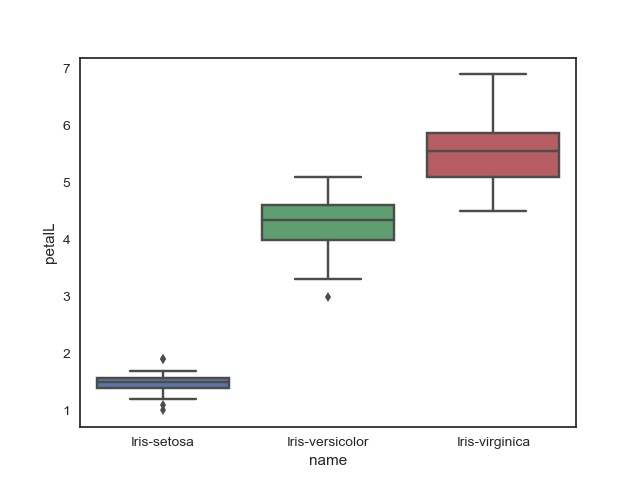


Figure 13 shows a great deal of separation between the Iris setosa petal lengths and the other two species, with some outliers in its highest and lowest lengths. Iris versicolour petal lengths fall between the smaller Iris setosa and larger Iris virginica. There is some separation between the Iris versicolor and Iris virginica as the formers quartiles are lower than the laters first quartile and the Iris versicolor petal length maximum value is below the Iris virginica’s median petal length. There is also an outlier low value in the Iris versicolor and low and high outliers in the Iris setosa petal length.

*Figure 14* Boxplot of petal widths cm

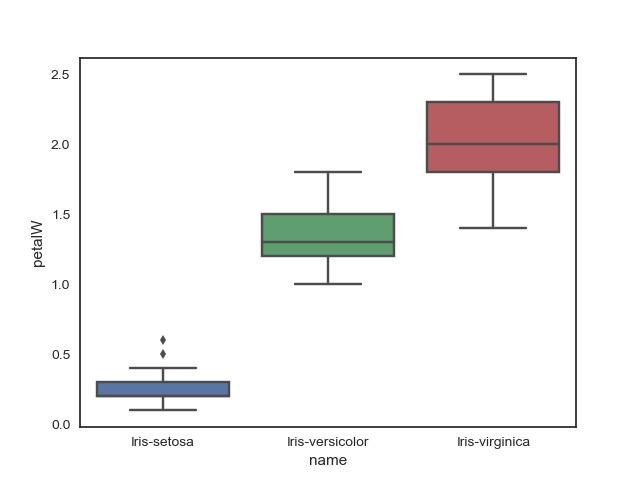


Figure 14 shows a similar pattern in petal widths as found in petal lengths. There is stronger separation between the widths of the Iris versicolor and Iris virginica as the maximum value of the former is just over the first quartile of the later. There is no overlap between these two and the Iris setosa which also shows the presence of two outliers at the upper end.

Figures 15-18 show the boxplot overlaid with the split plot to show the distribution of data over its boxplot measures.

*Figure 15* Boxplot spilt plot overlay of sepal lengths in cm

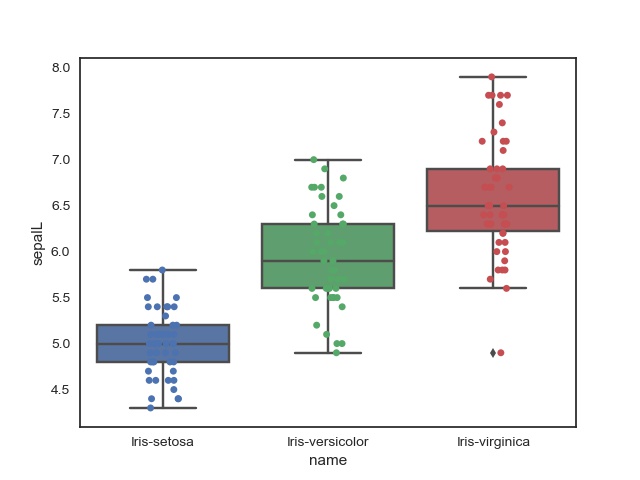


Figure 15 appears to show distribution density along within the quartiles for the Iris setosa sepal lengths with less distribution within this band for the Iris versicolor. The Iris virginica appears to show comparable distribution outside the quartiles as between them.

*Figure 16* Boxplot spilt plot overlay of sepal widths in cm

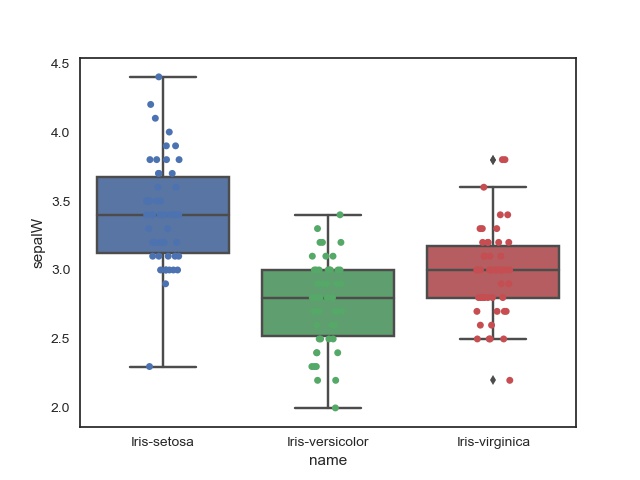


Figure 16 shows the distribution of the Iris setosa sepal widths are clustered between the upper and lower quartile whereas the dispersion is more uniform between the minimum and maximum values for the other two species.

*Figure 17* Boxplot spilt plot overlay of petal lengths in cm

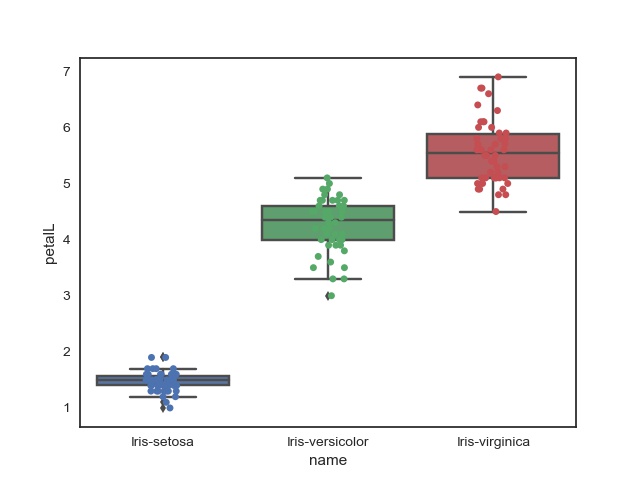


Figure 17 shows the distribution of Iris setosa petal lengths are focused between the narrow range of its maximum and minimum length. The Iris versicolor petal lengths are dispersed between the lower quartile and maximum whereas the Iris virginica show clustering within the upper and to just under the lower quartile.

*Figure 18* Boxplot spilt plot overlay of petal widths in cm

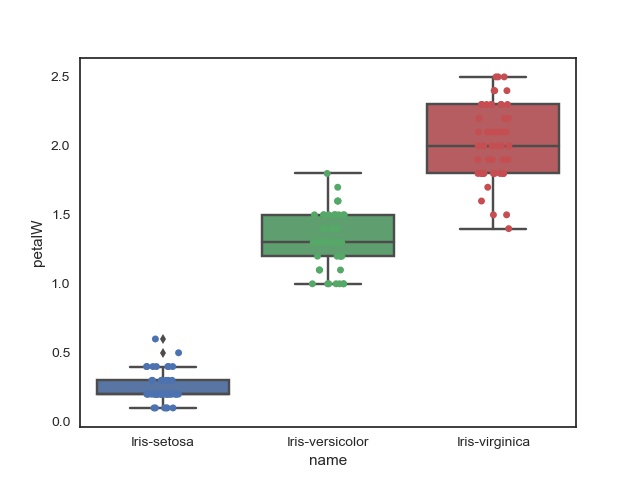
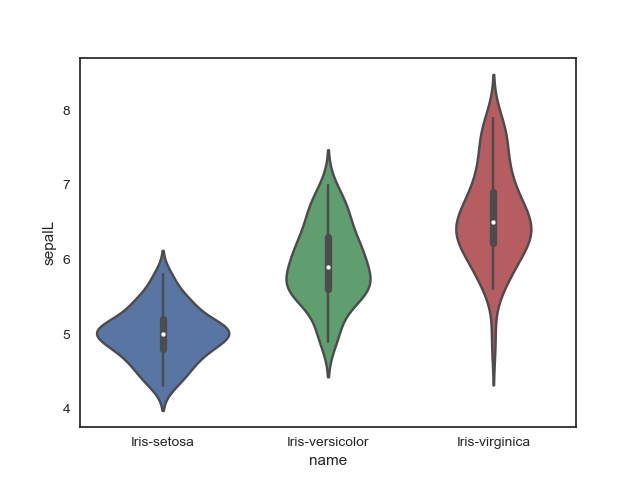


Figure 18 shows clustering of the Iris versicolor within its upper and lower quartile but with several samples at the minimum petal width. The Iris setosa petal widths are also focused between the narrow range of quartiles and then at the maximum and minimum. The Iris Virginia has the greatest spread and highest values of petal width.

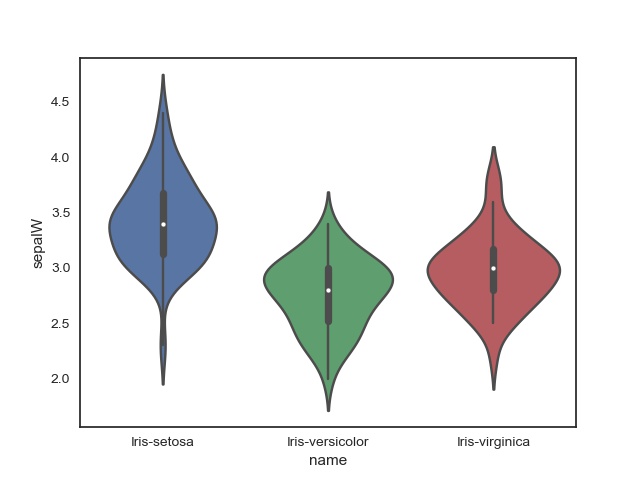
Figures 19 to 22 show the violin plots of the data which incorporates the data’s density distribution.

*Figure 19* Violin plot of sepal lengths in cm



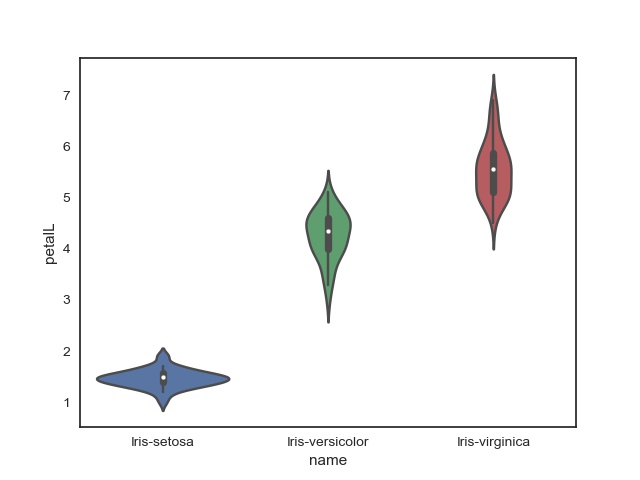
This shows the distribution for all is symmetrical. The Iris setosa shows a wide area around the mean indicating a higher probability that more samples from the Iris setosa will have a value close to this mean. The Iris versicolor plot indicates that further samples would fall just below the mean. However both the Iris versicolor and Iris virginica are narrower than the Iris setosa.

*Figure 20* Violin plot of sepal widths in cm



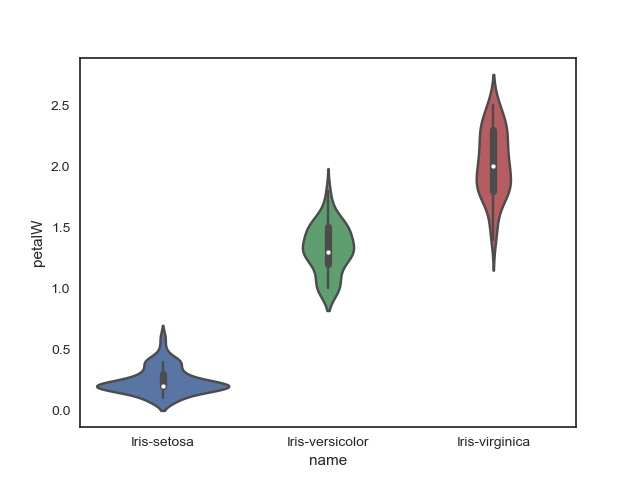
This shows…

Figure 21 Violin plot of petal lengths in cm



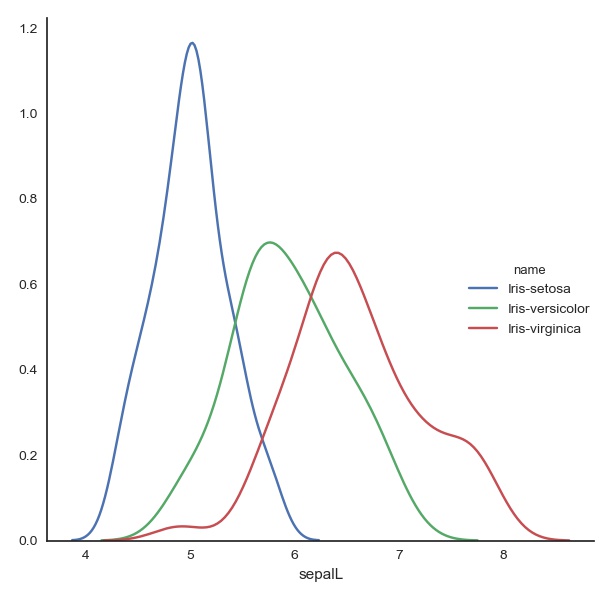
This shows….

Figure 22 Violin plot of petal widths in cm



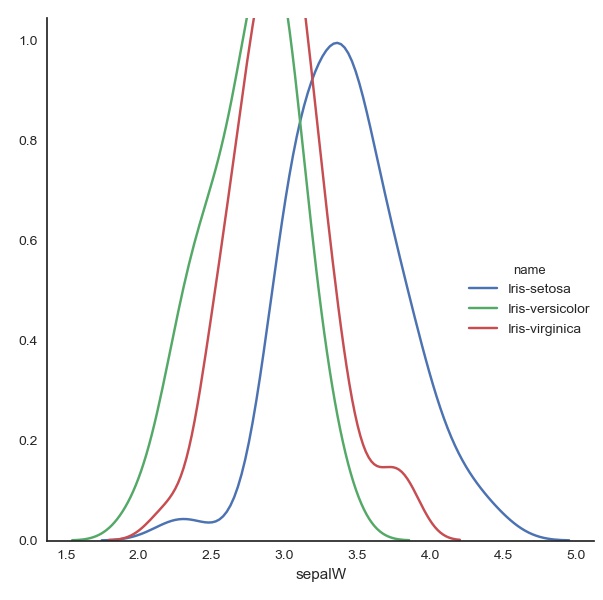
Figures 22 to 25 show the kde plot of the Iris data.

*Figure 22* kde plot of Iris sepal lengths



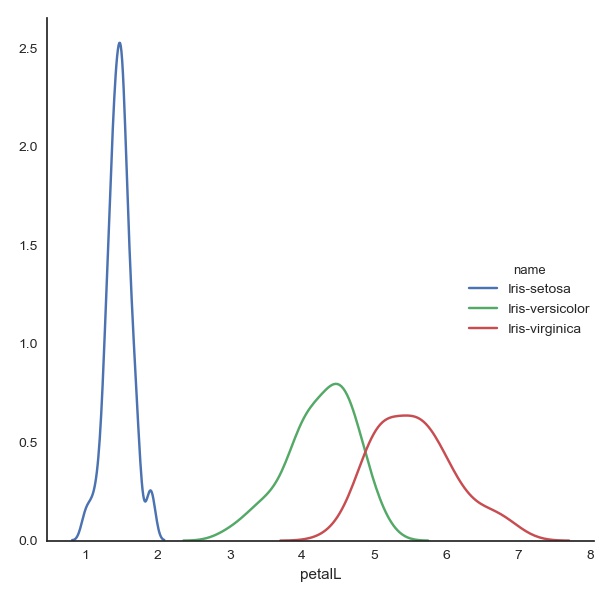
This shows…

*Figure 23* kde plot of Iris sepal widths



This shows…

*Figure 24* kde plot of Iris petal lengths



This shows…

*Figure 25* kde plot of Iris petal widths

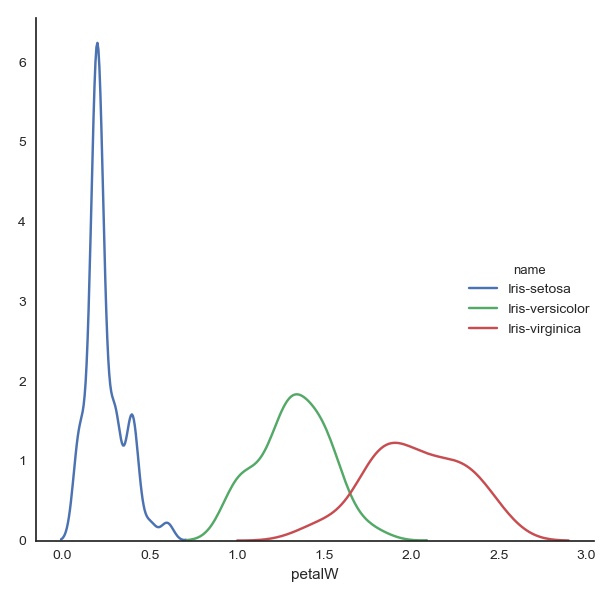
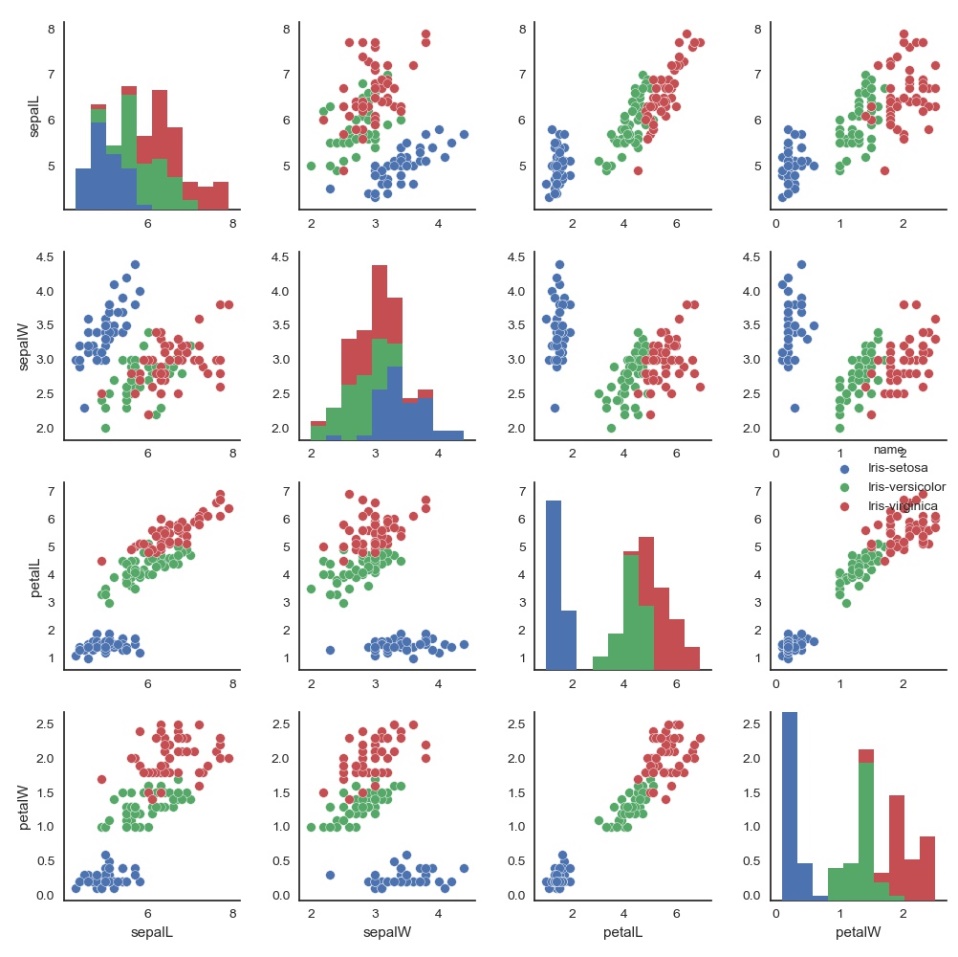


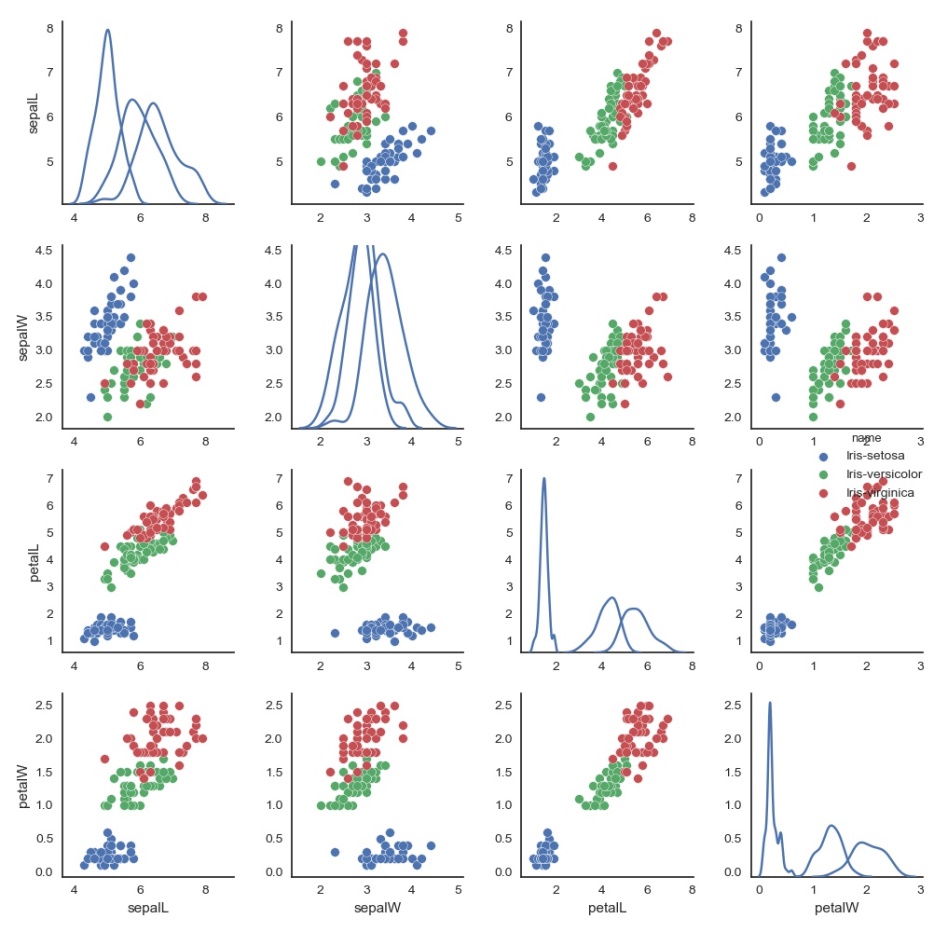
Figure 26 and 27 pairplots show the bivariate relationship between each pair of features. Figure 26 has the histogram along the diagonal and Figure 27 has the kde plot along the diagonal.

*Figure 26* pairplot of bivariate relations of the Iris data with histogram along the diagonal



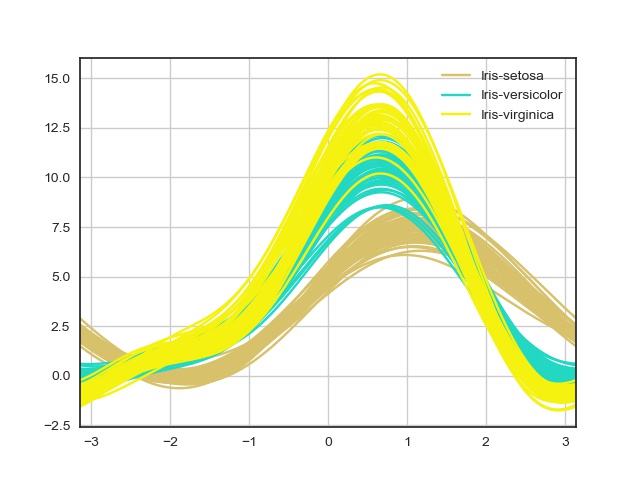
This shows….

*Figure 27* pairplot of bivariate relations of the Iris data with kde plot along the diagonal



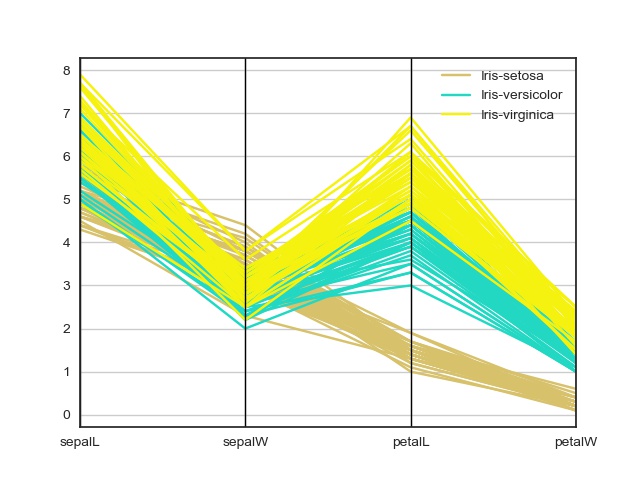
This shows…god with the graphs never end…

*Figure 30* Andrew Curves plot of Iris data



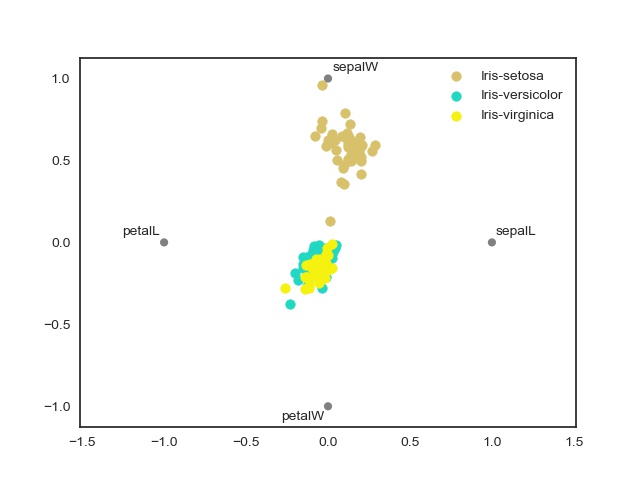
This shows something to do with fourier transformations…don’t know what yet or how to interpret graph…god I hope this is the last graph.

*Figure 3*1 Parallel Coordinates plot of Iris data



This shows…the iris set differs crossing over at the sepal width….. ahhhh.

Figure 32 Radviz plot of Iris data



This shows….

#### *Inferential* *statistics*

##### Comparing sepal widths

A one way t-test between the Iris setosa and Iris versicolor sepal widths was conducted. This tested if there is a significant difference in sepal widths between the Iris setosa and Iris versicolor. The Iris setosa’s average sepal width (M=3.418 , SD= 0.377) is wider and has greater variation than Iris versicolor (M= 2.77, SD=0.311). Levene’s test for homogeneity of variances indicated equality of variance (F=1.057 *, p=0.306*); therefore an Independent t-test was used. Results showed a significant difference in sepal widths between Iris-setosa and Iris-versicolor (t(98)=9.283, *p=4.362e-15*).

A one way t-test between the Iris setosa and Iris virginica sepal widths was conducted. This tested if there is a significant difference in sepal widths between the Iris setosa and Iris virginica. The Iris setosa’s average sepal width (M=3.418 , SD=0.377 ) is wider and has slightly greater variation than Iris-virginica (M= 2.974, SD=0.319). Levene’s test for homogeneity of variances indicated equality of variance (F= 0.967, p=0.181); therefore an Independent t-test was used. Results showed a significant difference in sepal widths between Iris-setosa and Iris-virginica (t(98)=6.289, p=8.917e-09).

A one way t-test between the Iris versicolor and Iris virginica sepal widths was conducted. This tested if there is a significant difference in sepal widths between the Iris versicolor and Iris virginica. The Iris versicolor’s average sepal width (M=2.77 , SD=0.311 ) is shorter slightly smaller variation than Iris-virginica (M=2.974 , SD=0.319). Levene’s test for homogeneity of variances indicated equality of variance (F=0.087 , p=0.768); therefore an Independent t-test was used. Results showed a significant difference in sepal widths between Iris-versicolor and Iris-virginica (t(98)=3.206, p=0.002).

##### Comparing sepal lengths

This test investigated there is a significant difference in sepal lengths between the Iris setosa and Iris versicolor. The Iris setosa’s average sepal length (M=5.006, SD=0.349) is smaller and has less variation than Iris-versicolor (M=5.936 , SD=0.511). Levene’s test for homogeneity of variances was significant (F=8.172, p<0.05); therefore Welch's t-test was used. Results showed a significant difference in sepal widths between Iris-setosa and Iris-versicolor (t(86.538)=-10.521, p<0.05).

This test investigated if there is a significant difference in sepal lengths between the Iris setosa and Iris virginica. The Iris setosa’s average sepal length (M= 5.006, SD=0.349) is smaller and has less variation than Iris-virginica (M=6.588 , SD=0.629). Levene’s test for homogeneity of variances was significant (F=11.454 , p=0.001); therefore Welsh's t-test was used. Results showed a significant difference in sepal lengths between Iris setosa and Iris-virginica (t(76.516)=-15.386, p<0.05).

This test investigated if there is a significant difference in sepal lengths between the Iris versicolor and Iris virginica. The Iris versicolor's average sepal length (M=5.936 , SD=0.511 ) is shorter and has slightly less variation than Iris-virginica (M=6.588 , SD=0.629). Levene’s test for homogeneity of variances indicated equality of variance (F=1.025, p=0.314); therefore an Independent t-test was used. Results showed a significant difference in sepal widths between Iris-versicolor and Iris-virginica (t(98)=-5.629, p=1.725e-07).

##### Comparing petal widths

This tested if there is a significant difference in petal widths between the Iris setosa and Iris versicolor. The Iris setosa’s average petal width (M= 0.244, SD=0.106) is wider and has less variation than Iris-versicolor (M=0.106 , SD=1.326 ). Levene’s test for homogeneity of variances was significant and both the setosa and versicolor failed the Shapiro Wilk test for normal distribution. Thus the t test was not conducted.

This tested if there is a significant difference in petal widths between the Iris setosa and Iris virginica. The Iris setosa’s average petal width (M=0.244, SD=0.106 ) is smaller and has less variation than Iris-virginica (M=2.026 , SD=0.272). Levene’s test for homogeneity of variances was significant (F=38.107, p=1.517e-8); The Shapiro Wilk for the Iris setosa petal widths was significant (F=0.814, p=1.853e-06) whereas the Shapiro Wilk test for the Iris Virginica was not significant (F=0.960, p=0.09) therefore no t-test was performed.

This tested if there is a significant difference in petal widths between the Iris versicolor and Iris virginica. The Iris versicolor's average petal width (M=1.326 , SD=0.196) is smaller and has less variation than Iris-virginica (M=2.026 , SD=0.272). Levene’s test for homogeneity of variances was significant (F=6.546, p=0.012); therefore Welch's t-test was used. Results showed a significant difference in petal widths between Iris-versicolor and Iris-virginica (t(89.043)=-14.625, p<0.05).

##### Comparing petal lengths

This tested if there is a significant difference in petal lengths between the Iris setosa and Iris versicolor. The Iris setosa’s average petal length (M=1.464 , SD=0.172 ) is smaller and has less variation than Iris-versicolor (M=4.26, SD=0.465 ). Levene’s test for homogeneity of variances was significant (F=30.897, p=2.348); therefore an Independent t-test was used. Results showed a significant difference in petal lengths between Iris-setosa and Iris-versicolor (t(62.118)=-39.469, p<0.05).

This tested if there is a significant difference in petal lengths between the Iris setosa and Iris virginica. The Iris setosa’s average petal length (M=1.464 , SD= 0.172) is much smaller and has less variation than Iris-virginica (M=5.552, SD=0.546). Levene’s test for homogeneity of variances was significant (F=39.977 , p=7.651e-09); therefore an Welch's t-test was used. Results showed a significant difference in petal lengths between Iris-setosa and Iris-virginica (t(58.593)=-49.9657, p<0.05).

This tested if there is a significant difference in petal lengths between the Iris versicolor and Iris virginica. The Iris versicolor's average petal length (M=4.26 , SD=0.465 ) is smaller and has less variation than Iris-virginica (M=5.552 , SD=0.546). Levene’s test for homogeneity of variances indicated equality of variance (F=1.067 , p=0.304); therefore an Independent t-test was used. Results showed a significant difference in petal lengths between Iris-versicolor and Iris-virginica (t(98)=29.023 p=6.428).

**Discussion**

This project analysed the Iris data set.

Add in a paragraph on the description of the whole data. …

Firstly the ratio’s between the length and width was explored. The Setosa sepal ratios were fairly similar with no lengths being more than double the widths. In contrast the Setosa petals lengths varied between double and 15 times larger than the widths. The sepals of both the Versicolor and Virginica also showed this modest pattern with the sepal measures with no lengths being more than 3 times the corresponding width. There was greater similarity between the sepal measures of the Virginica and Versicolor than the smaller ratio between the Setosa measures. The petal length/width ratio for the Virginica and Versicolor were more modest compared to the Setosa petal ratios. No length was greater than just over 4 times its width and averaging at around 3 times the width. Thus the sepal’s appeared more similar across the three species. The petals were comparable between the Virginia and Versicolor but in stark contrast to those of the Setosa. This pattern was illustrated in the scatter plot of the Iris petals figure 6. Figure 6 plotted unlabelled petal measures and strongly indicates the presence of two clusters. Figure 10 is the same scatter plot but with the three species of Iris displayed in differing colours. This plot suggested further clustering, with the Setosa petals tightly bunched in the lower left of the graph, the Versicolor in the middle and the Virginica in the upper left. There was some overlap between the Versicolor and Virginica measures but no overlap of the Setosa and the petals of the other species. XRandolf?X (Ref) suggested that the Versicolor was a hybrid of the Setosa and Virginica and an initial analysis supported this hypothesis as the Versicolor lay between the other two. However Anderson selected the Setosa and Virginica from the same region whereas the Versicolor was collected from a different region. It would be useful to select all the data from one region to investigate if the stratification is more or less pronounced.

The labelled data set shows cluster patterns for both the sepal and petal dimensions. The petal labelled data shows three clusters while the sepal labelled data shows a setosa cluster with the other two species forming the another cluster. The Virginia and Versicolor sepal measures show overlap around the 5.5-6.5cm sepal lengths. This becomes more stratified at the upper and lower extremities with the Virginica tending towards the upper right of the scatter plot and the versicolor along the lower left. However the stratification in sepal measures is far less pronounced than with the petal measures which had far less overlap between the Versicolor and Virginica.

The cluster’s suggested by the descriptive statistics and visualisation of the data set was tested for association. There were significant differences in sepal widths between the setosa and the other two species. There was also a significant difference in sepal widths between the versicolor and virginica. This indicates that the means of the sepal widths of all the species are from different distributions. In these cases the data satisfied the equality of variance criteria for independent t-test. The sepal lengths inferential analysis was not clearly separated as in the case of the Setosa paired with both the Versicolor and Virginica, the independence of variance criteria was not met. Yet there was still a significant difference in the means albeit with less degrees of freedom. The Versicolor and Virginia t-test for sepal length satisfied the equality of variance test and was significant suggesting these came from different normal distributions.

Inferential statistics on petal width showed that the Setosa and Versicolor did not satisfy the criteria for performing Student’s t-test or Welch’s t-test. This would require further analysis as to why the data was not normally distributed. The box plot for the petal widths shows the presence of outliers on the upper end of the Setosa measurments. Note that the data presented by UCI contains two errors in the 35th and 38th  (entries 34 and 37 in appendex 1 ) sample both of which affect the Iris Setosa dimension petal width, with another transcription error in the Setosa’s petal widths for data point 38 (37 in appendex 1). It would be useful to repeat this test with outliers removed and data corrected to see if t-test criteria would then be satisfied. It would also be useful to do an analysis of variance ANOVA on the data. Similarly no t-test was performed between the Setosa and Virginica petal widths as normal distribution criteria was not satisfied. Again this could be due to outliers and errors in the Setosa data.

Petal length tests were more conclusive. Setosa and Versicolor Welch’s t-test was significant suggesting they came from different normal distributions. Significant differences were found when the Setosa was tested against Virginica petal lengths. The Versicolor and Virginica also showed significate difference in means with greater degrees of freedom than with the setosa petal lengths.

In sum, the mean sepal widths were significantly different across all species, with the Setosa (3.42cm) being larger than the Virginica (2.97cm) which in turn is larger than the Versicolor (2.77cm). Sepal length means were also significantly different with the setosa having the smallest sepal length (5.01cm), then the Versicolor (5.94cm) and Virginica (6.59cm) with the largest measure. This suggests that the flower could be identified by the either its sepal length or width. Petal widths were not as good an indicator of Iris species since the Setosa appeared not to follow a normal distribution. However it was possible to differentiate between the Virginica (2.03cm) and Versicolor (1.33cm) species. Petal lengths were also useful in identifying plant species. The mean petal lengths for the Iris species was also useful for identification purposes as the Setosa (1.46cm) had the smallest mean, followed by the Versicolor (4.26cm), with the Virginica (5.55cm) having the largest petal length. Further research could investigate this dataset without outliers. The ratio length:width might also be a useful method for identifying the setosa from the other two species given the relatively large ratio compared to Versicolor and Virginica ratios. This project showed the usefulness and power python can bring to data analysis in performing descriptive, inferential and visual analysis of data.

**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) []() #

**Appendix**

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 |