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

Helen O'Shea

Galway Mayo Institute of Technology

Table of Contents

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 consisted 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 and clear differences between sepals and petals. A third cluster was seen when the data categorised.

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 It was collected by the American botanist, Edgar Anderson in 1935 yet remained unpublished till Fisher published it in his 1936 paper (ibid). Anderson's data consisted of fifty samples each, from 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 while the Iris Versicolor was sourced elsewhere.

The data was 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 five measures, the species type and its corresponding length and width of its petals and sepals. Measures were taken in centimetres (cm). The petals are the inner flower while the sepals protect the emerging flower on the outer part of the plant structures as seen in figure 1. Thus plant sepals are present for a greater duration of the Iris's lifecycle than its petals.

Figure 1: Sepal and Petal of the Iris Versicolor

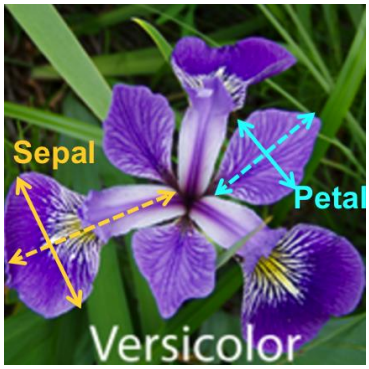


Figure 2, figure 3 and figure 4 are pictures of the flowering Iris Setosa, Iris Virginica and Iris Versicolor respectively.

Figure 2: Iris Setosa

Figure 3: Iris Virginica

Figure 4: Iris Versicolor

Thus the data Anderson collected consisted of 150 samples of Iris spanning 5 dimensions; namely 4 one dimensional measures in cm, sepal length (sepalL), sepal width (sepalW), petal length (petalL), petal width (petalW); And one categorical dimension of Iris species (name). The words in brackets are the names given to these dimensions in the code attached to this project.

This project used the version of the Iris data hosted at UCI machine learning repository (Iris Data set, Iris Data, donated 1988). The UCI version contains two error rows (3 errors values in total) from Anderson's original data which were not amended in the analysis (Bezdek, Keller et al., 1999). Both errors rows were the Iris Setosa samples (Iris Data, donated 1988).

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 many aspects of computer science. Reasons the Iris data set is frequently used in computer science include the following; 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. Fisher's Iris data set is often used as a student learning tool in data analytics, and artificial neural networks. The data is small enough to be manageable for beginners yet sufficiently challenging in what it can reveal. There is more to investigate with the Iris data, with its 4 non categorical dimensions, than can be explored with 2 dimensional data. 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 (Fisher, 1936). Iris Versicolor was intermediate between the other two species with some overlap of measures with Iris Virginica (ibid). Once the samples were categorised by species then three clusters emerged. This project sought to replicate these findings and explore the data further with the more modern tools of computer science.

This project explored the Iris Data using the programming and scripting language Python Version 3.6. In addition GitHub was used as a development platform. 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 allow statistical data to be visualised while Scipy has useful statistical tests.

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 to explore areas that might need further investigations. Finally introductory inferential statistics were used to explore if there was significant differences between the Iris species.

Method

This report produced descriptive, graphical visualisations and inferential tests 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 (Appendix 1). The column headings are as follows; sepalL represents the sepal lengths, sepalW represents sepal widths, petalL represents petal lengths, petalW represents petal

widths and name represents the iris species, Iris-setosa, Iris-versicolor and Iris-virginica. All measures are in cm.

Table 1a

Iris	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. A computer and internet connection was required for conducting research. In addition the python modules, NumPy, Scipy, Pandas, Matplotlib, Scipy and Seaborn were used. GitHub in conjunction with Visual Studio Code was used for project 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 /project /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. Then descriptive statistics were produced for each species of Iris (length/width ratios, mean, maximum and minimum values of the measures).

Then the data was explored using visualisation techniques including scatter plots, histograms, box plots and various density plots. The data was visualised using the NumPy extension modules Matplotlib and Seaborn. Scipy was also used for q-q plots.

Inferential statistics were computed using one way independent t-test for comparing two means from a normal distribution with equal variances. Welch's adaption of this test was used to compare means from a normal distribution with unequal variances.

Results

Descriptive statistics

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 Readme.md contains a list of all graphs and the dimensions used to product them.

Table 2: Iris Dataset descriptive statistics

Iris cm	sepalL	sepalW	petalL	petalW
count	150	150	150	150
mean	5.843	3.054	3.759	1.199
std	0.828	0.434	1.764	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

Table 2 shows that mean lengths are greater than widths for both sepals and petals. Sepal widths are slightly smaller than petal lengths, which had the largest spread of all the measures. The maximum values approximately double the minimum values except in with petal lengths. The data was grouped by species name and analysed further.

Table 3: Iris Setosa descriptive statistics

Setosa cm	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 Setosa 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 (0.244cm). The sepal length had the highest mean (5.006cm) and was over three times larger than the petals length (1.464cm). Although the sepal mean width (3.418cm) was smaller than its mean length (5.006cm), it was over 14 times bigger than the corresponding mean width for the petals (0.244cm). The sepal length had the largest maximum (5.8cm) and minimum value (4.3cm) of all the measures while the petal width showed the smallest maximum (0.6cm) and minimum (0.1cm) value. The standard deviation (std) showed that the sepal values (length 0.352 and width 0.381) were more spread out than the petal values (length 0.174 and width 0.107). The table also shows the percent quartiles.

Table 4 Iris Virginica descriptive statistics

Virginica cm	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

This table shows that for the 50 Iris Virginica samples, the mean was smallest for the petal width (2.026cm) and largest for the sepal length (6.588cm) as with the Iris Setosa. The maximum (highest at 7.9cm for sepal length and lowest at 2.5cm for petal width) and minimum (highest at 4.9 for sepal length and lowest at 1.4cm for petal width) were higher than the Iris Setosa'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 widths which were smaller (2.2cm minimum and 3.8cm maximum). All quartiles except the sepal width were greater than the Iris Setosa quartiles.

Table 5 Iris Versicolor descriptive statistics

Versicolor cm	**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

For the 50 Iris Versicolor samples, the means for the petal dimensions (4.26cm length, 1.326cm width) were smaller compared to the corresponding means for the sepals (5.936cm length, 2.77cm width). The largest mean was for the sepal length (5.936cm) and smallest for the petal width (1.326cm) with the values falling between the lower Setosa measures and higher Virginica ones. The maximum value was the sepal length (7cm) which was just under the maximum for the Virginica species (7.9cm). The minimum measure was the petal width (1cm) similar to the other Iris species. The measure of spread for the Iris Versicolor (0.516 for sepal length, 0.314 for sepal width, 0.470 for petal length, 0.198 for petal width) was slightly lower for the corresponding values for the Iris Virginica (0.636, 0.322, 0.522, 0.275 respectively) but higher than the Iris Setosa (0.352 , 0.381 , 0.174, 0.107, respectively) The quartiles lay between the lower Setosa and higher Virginica except for the Sepal width which had the lowest quartiles (25% at 2.525, 50% at 2.8 and 75% at 3) out of all the samples.

The ratio of sepal length to width and petal length to width were also calculated as seen in 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

This shows that the Setosa petal ratio had the largest length to width mean ratio at 7.078cm whist the Setosa sepal ratio had the smallest length to width ration 1.475cm. The Versicolor and Virginia had comparable ratios ranging between lengths being twice as big as widths to lengths being under 4 times as big as widths. Overall these figures suggested that the Setosa petals were an area of interest.

Visualisations

Scatter plots

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. Graphs produced by visu.py code were saved in the folder project/graphs/. For a full index of graphs produced see /project/readme.md.

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

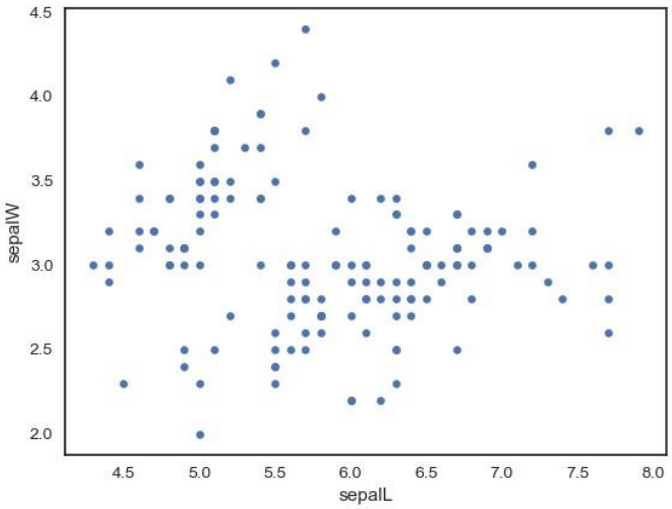


Figure 5 shows the scatter plot of all the data without labelling it by species. This shows that sepal lengths tend to be greater than sepal widths. There appears to be a cluster in the top left of the plot and a larger dispersed cluster in the mid to right section of the plot. Figure 6 shows the corresponding plot of the petal lengths.

Figure 6 Scatter plot of the Iris's petal length versus widths in cm

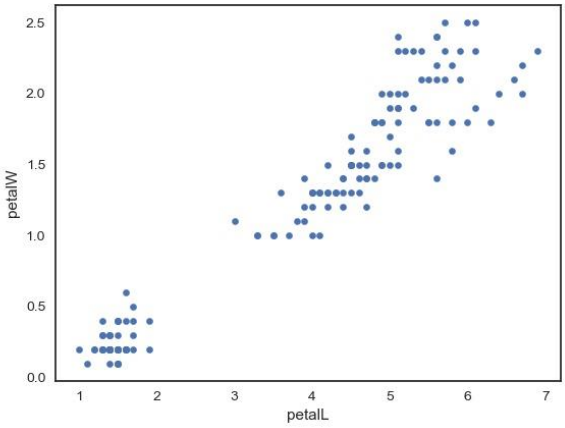


Figure 6 shows a similar pattern of lengths being greater than widths. However here, larger lengths tend to have larger widths. Two distinct clusters appear to be present, a small tight cluster in the lower left of the scatter plot and a larger more dispersed cluster in the middle to upper right of the scatter plot. The data appears to have a very strong liner relationship and a high positive correlation between petal lengths and widths.

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

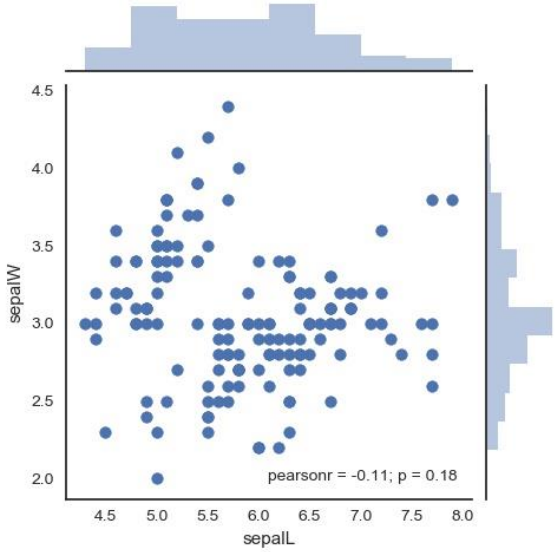


Figure 7 and 8 were produced using the seaborn module which produces a scatter plot that also calculates Pearsons correlation coefficient r. Figure 7 shows the scatter plot for sepeal lengths and widths. This shows Pearson's r as -0.11 indicating a very slight negative corralation between sepal lengths and widths. However the association is not signifigiant as shown by the p vlaue of 0.18 > 0.05;

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

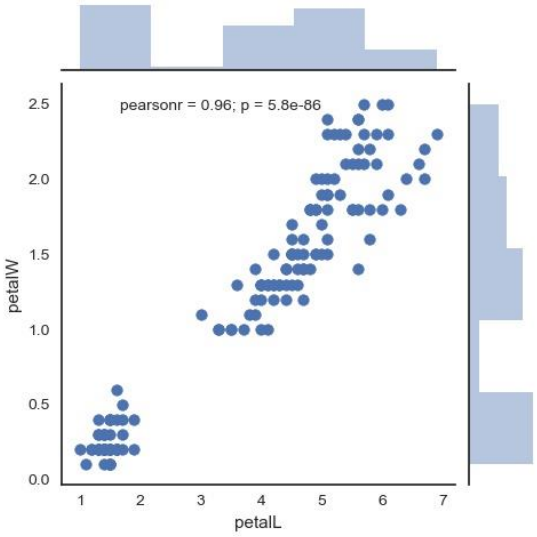
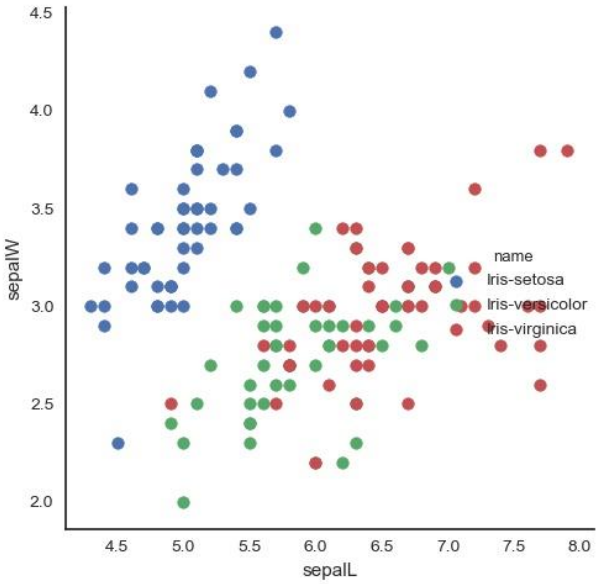


Figure 8 shows Pearson's r at 0.96 indicating an almost perfectly positive linear relationship between petal lengths and widths. This association is significant as the p value is less than 0.05.

The seaborn module was also used to explore this difference between sepals and petals by identifying each Iris species within the scatter plot as shown in Figures 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) with its sepals as seen in the mid to upper left of the graph. The Iris Setosa appears to show a strong positive correlation in sepal length to width dimensions that is absent in the other two species.

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. The lengths and widths also appear to be slightly positively correlated for Virginica and Versicolor but to a far lesser degree than with the Setosa dimensions.

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

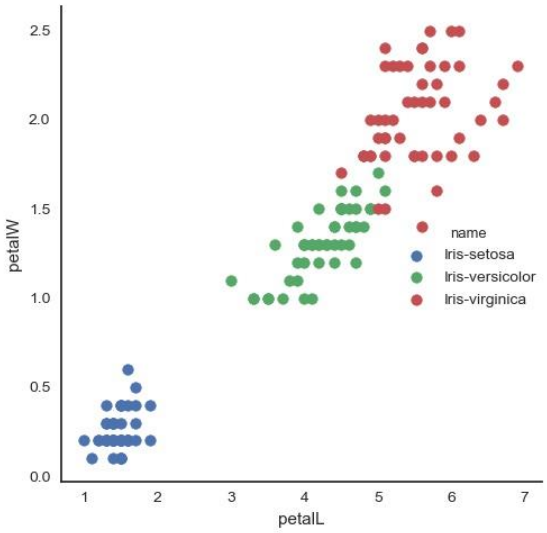


Figure 10 clearly shows three clusters with on the petal dimension. The Iris Setosa has clearly smaller petal dimensions than the other two, with the Iris Versicolor falling linearly between the Iris Setosa and Iris Virginica with some overlap between larger Versicolor petals and smaller Virginica petals. There is no overlap in petal measures between the Setosa and the other two Iris species.

Histograms

Figures 11– 14 show histograms of the Iris data. These graphs are saved in /project/graphs/as ISW.jpg, ISL.jpg, IPW.jpg and IPW.jpg.

Figure 11 Histogram of Iris sepal widths

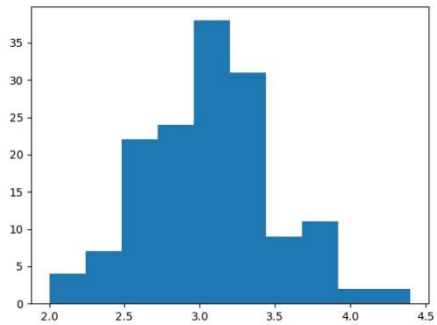


Figure 12 Histogram of Iris sepal lengths

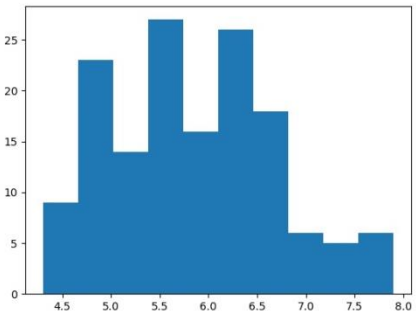


Figure 13 Histogram of Iris petal widths

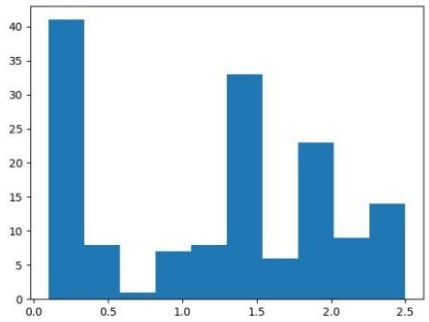
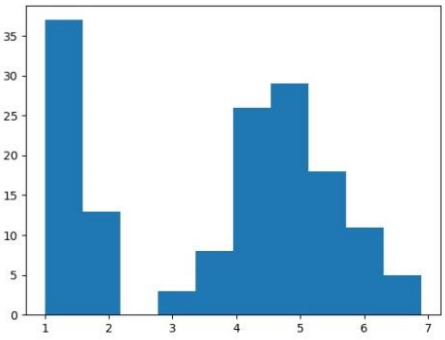


Figure 14 Histogram of Iris petal lengths



Figures 11-14 suggest that the sepal dimensions may be normally distributed, however the petal dimensions appear to be from a multimodal distribution. The data was then plotted by name.

Figure 15 Histogram of Iris Setosa sepal widths

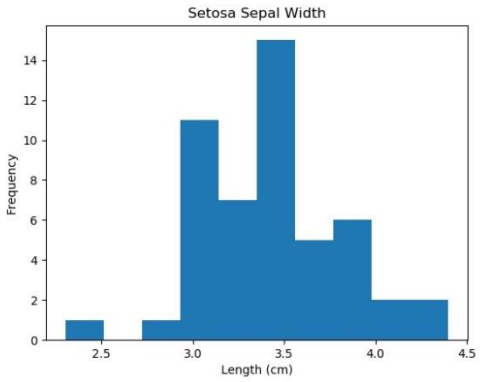


Figure 16 Histogram of Iris Setosa sepal lengths

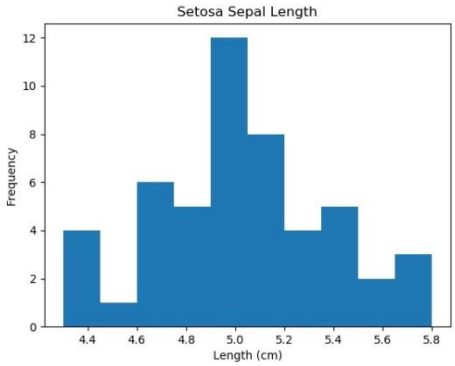


Figure 17 Histogram of Iris Setosa petal widths

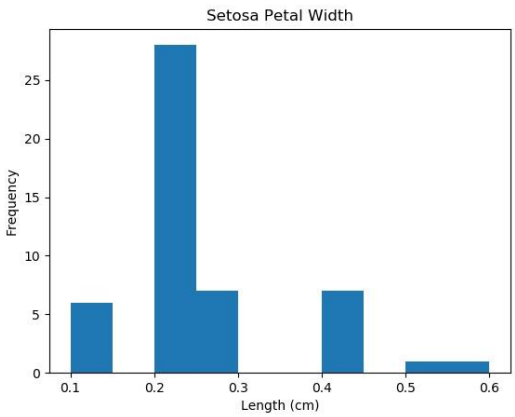
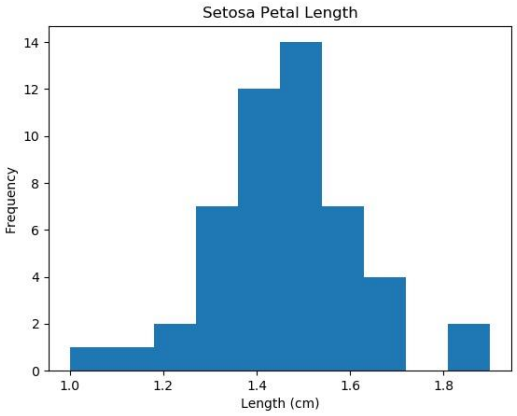


Figure 18 Histogram of Iris Setosa petal lengths



The Iris Setosa sepal histograms show signs of symmetry around the centre and could be from a normal distribution although sepal widths may be right skewed. The Setosa's petal lengths also might be normally distributed however the petal widths do not look normally distributed with skewness to the right.

Figure 19 Histogram of Iris Versicolor sepal widths

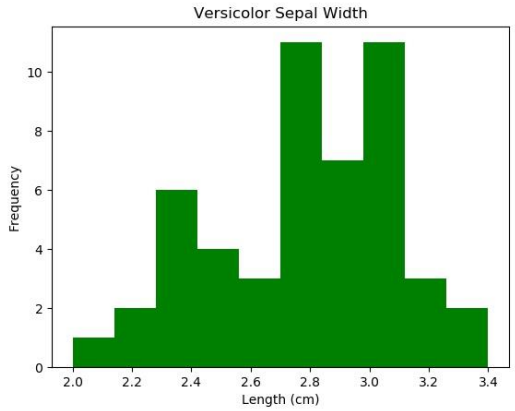


Figure 20 Histogram of Iris Versicolor sepal lengths

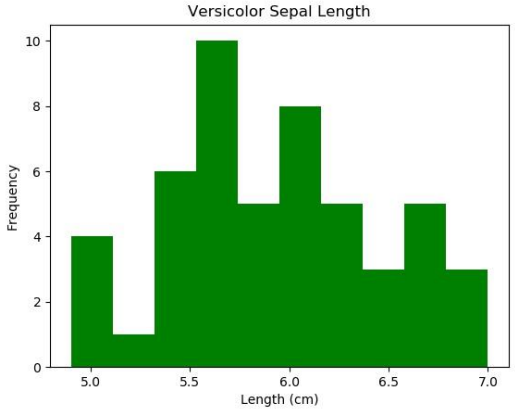


Figure 21 Histogram of Iris Versicolor petal widths

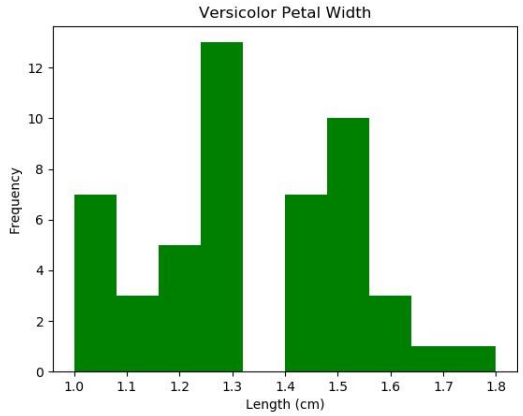
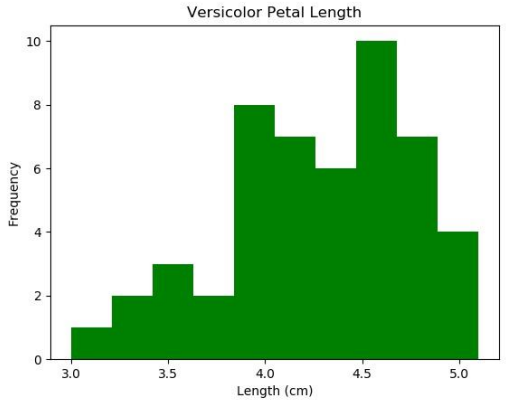


Figure 22 Histogram of Iris Versicolor petal lengths



Figures 19 – 22 show the histograms of the Iris Versicolor. The data does not look obviously normally distributed. Spal widths appear multimodal or left skewed, while the lengths could possibly follow a normal distribution but seem skewed to the right. The petal widths suggest a multimodal distribution with right skew while the petal lengths look skewed to the left.

Figure 23 Histogram of Iris Virginica sepal widths

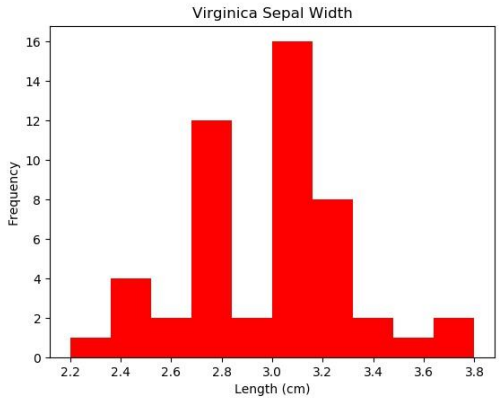


Figure 24 Histogram of Iris Virginica sepal lengths

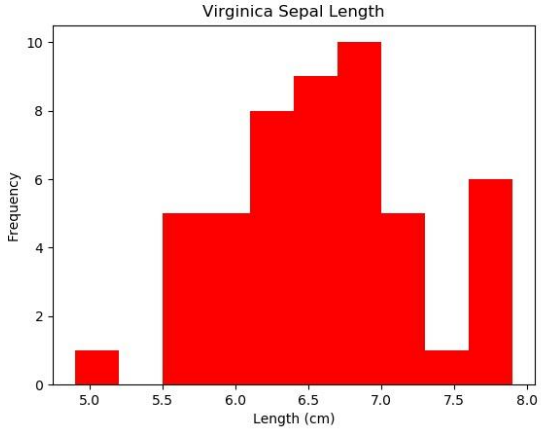


Figure 25 Histogram of Iris Virginica petal widths

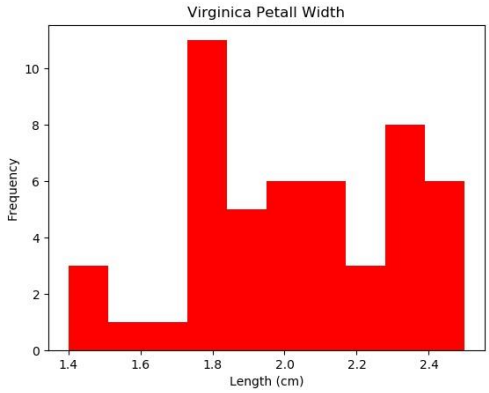
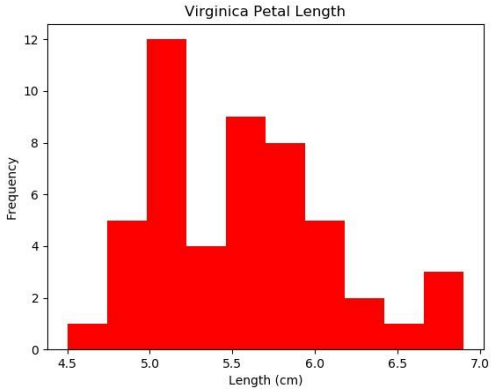


Figure 26 Histogram of Iris Virginica petal lengths



The 3 of the histograms for the Iris Virginica do not show clear normal distribution, while the sepal widths looks to have a central axis. The petal lengths appear skewed to the right and the petal widths appear multimodal.

Box plots

Figures 27-30 show the box plots of the sepal and petal lengths and widths sorted by Iris name.

Figure 27 Boxplot of sepal lengths cm

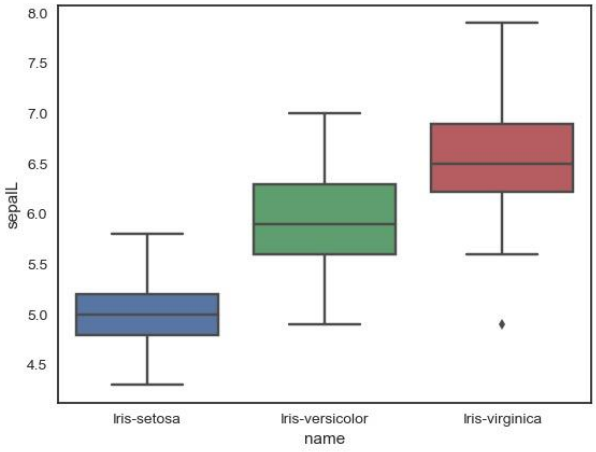


Figure 27 shows that sepal length quartiles are smallest for the Iris Setosa and largest for the Iris Virginica with the Iris Versicolor falling between them. The quartile spread is similar for the Iris Virginica and Iris Versicolor while narrower and more symmetrical for the Iris Setosa. The maximum 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 28 Boxplot of sepal widths cm

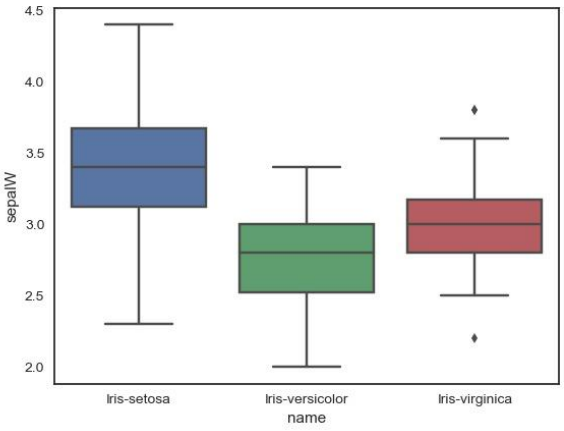


Figure 28 showed that the Iris Setosa width has the larger quartiles compared to the other two with a greater spread between its maximum and minimum. The Iris Veriscolor has the lowest quartiles while the Iris Virginica lies between the Iris setosa and Iris versicolor. The Iris Virginica also shows outliers in its highest and lowest sepal width. The range of sepal widths for the Iris Setosa encompasses all the Iris Virginica measures and the maximum

and quartile span of the Iris Versicolor, with only Versicolor values close to the minimum falling outside of the Setosa's range.

Figure 29 Boxplot of petal lengths cm

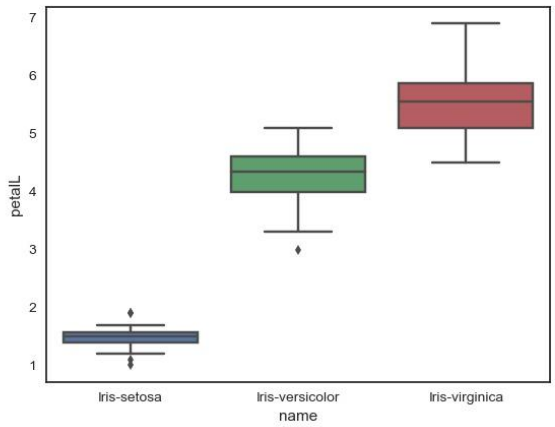


Figure 29 shows a great deal of separation between the Iris Setosa petal lengths and the other two species, with some outliers in its highest and lowest lengths. Iris Versicolour petal lengths fall between the smaller Iris Setosa and larger Iris Virginica, with no overlap with the former and some overlap between the upper Versicolor values and the mid to lower Virginica values. Some separation between the Iris Versicolor and Iris Virginica was present as the Versicolor quartiles were lower than the Virginicas' first quartile. In addition, 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 30 Boxplot of petal widths cm

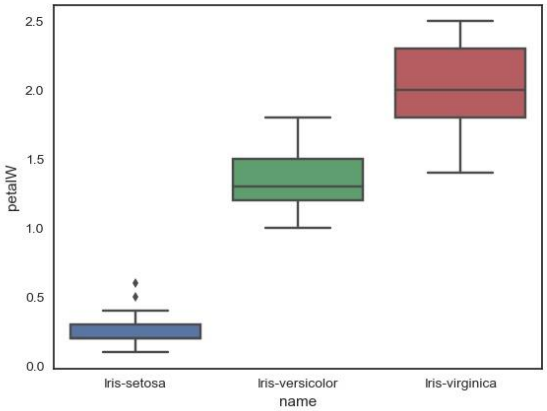


Figure 30 shows a similar pattern in petal widths as found in petal lengths. There is stronger separation between petal widths of the Iris Versicolor and Iris Virginica as the maximum value of the former is just over the first quartile of the later. There is no overlap between these two and the Iris Setosa. The Setosa also showed the presence of two outliers at the upper petal widths.

Figures 31-34 show the boxplot overlaid with the split plot to show the distribution of data over its boxplot measures.

Figure 31Boxplot spilt plot overlay of sepal lengths in cm

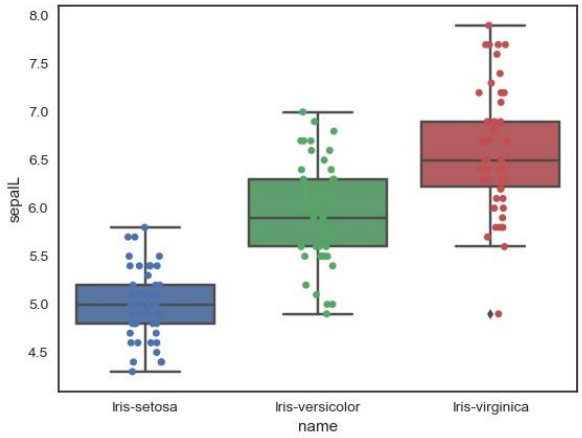


Figure 31 appears to show distribution density along within the quartiles for the Setosa sepal lengths, with less distribution within this band for the Iris versicolor. The Virginica appears to show comparable distribution outside the quartiles as between them.

Figure 32 Boxplot spilt plot overlay of sepal widths in cm

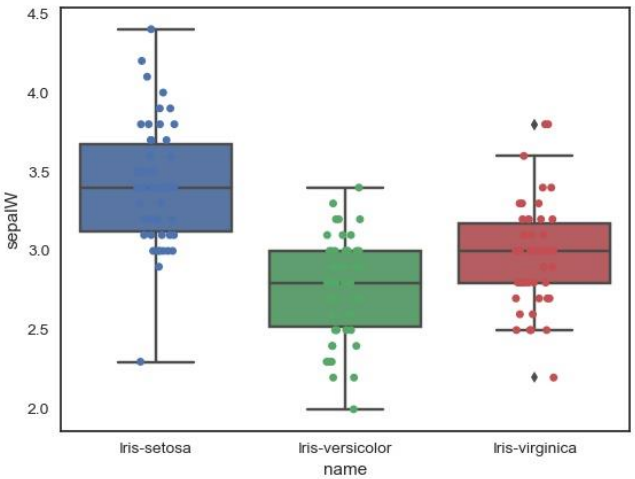


Figure 32 shows the distribution of the Iris Setosa sepal widths are clustered between the upper and lower quartile whereas the dispersion appears to be more uniform between the minimum and maximum values for the other two species. The Setosa spread shows a longer tail at the upper width values.

Figure 33 Boxplot spilt plot overlay of petal lengths in cm

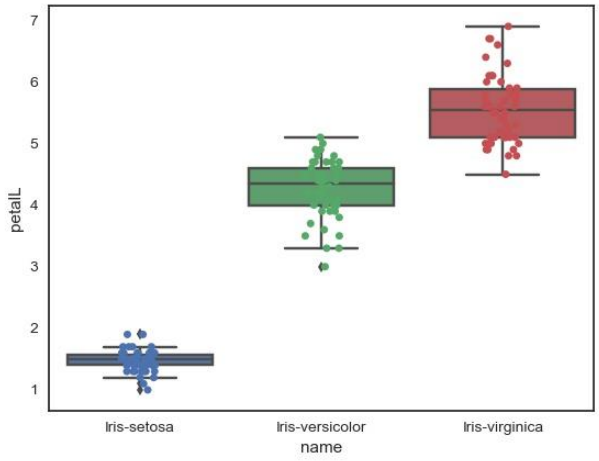


Figure 33 shows the distribution of Iris Setosa petal lengths are focused between the narrow range of its maximum and minimum length as are the other two albeit within a larger range. The Iris Versicolor petal lengths are dispersed between the lower quartile and maximum whereas the Iris Virginica shows clustering within the upper and to just under the lower quartile.

Figure 34 Boxplot spilt plot overlay of petal widths in cm

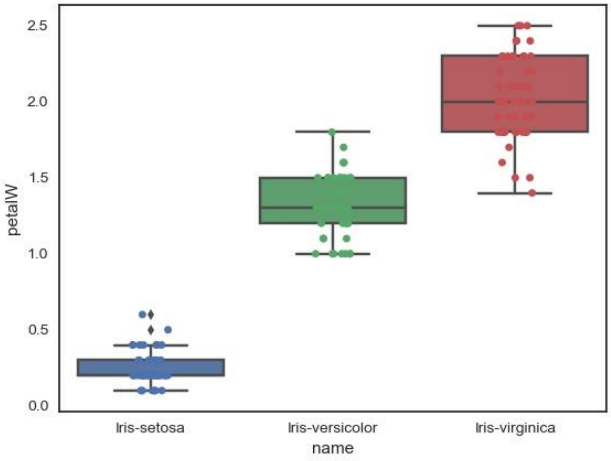


Figure 34 shows clustering of the Iris Versicolor within its upper and lower quartile but with several samples at the minimum petal width. The Iris Setosa petal widths are also focused between the narrow range of quartiles and then at the maximum and minimum. The Iris Virginia has the greatest spread and highest values of petal width.

Density plots

Figures 35 to 38 show the violin plots of the data which incorporates box plot details with the data's density distribution.

Figure 35Violin plot of sepal lengths in cm

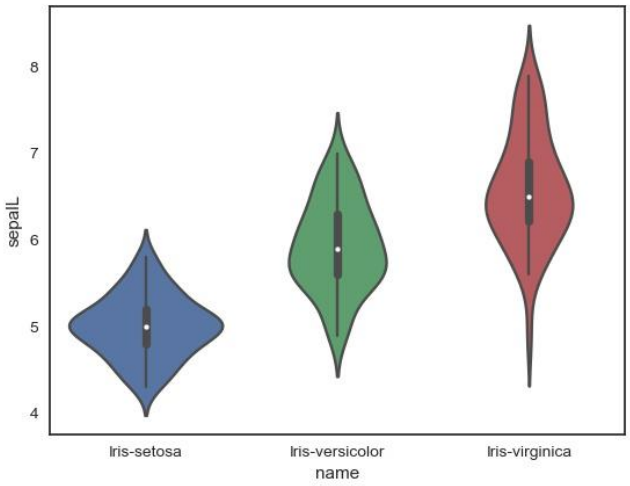


Figure 35 shows the distribution in sepal lengths for all the species is symmetrical with most values focusing round the mean (the central bulge) as in a normal distribution. The Iris Setosa shows a wide area around the mean indicating a higher probability that more samples from the Iris Setosa will 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, suggesting the distribution is wider than for the Setosa. The sepal lengths of the Iris Virginica show a long narrow tail for the smaller sepal lengths and slightly thicker tail for the upper sepal lengths. This suggests the density is quite dispersed. The Versicolor is not as spread out as the Virgnica while the Setosa is quite compact.

Figure 36 Violin plot of sepal widths in cm

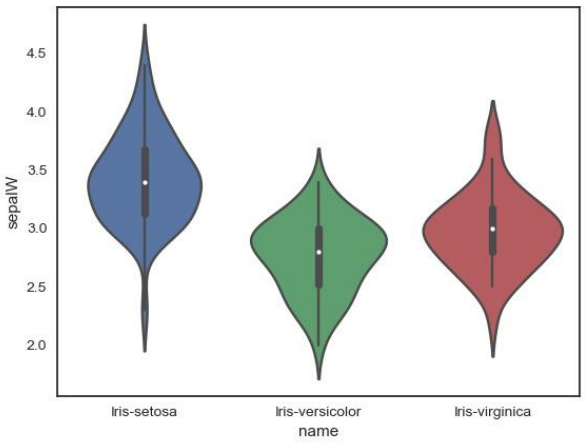


Figure 36 shows the violin plot of sepal widths. This shows a similar pattern in terms of symmetrical density around the mean; however the Setosa widths show a long tail in the lower sepal widths. The density around the mean is less pronounced in the Setosa sepal widths than it was in their lengths. Whereas the pattern is reversed with the other two as the area around their mean is more pronounced than it was with lengths. The Setosa has the highest values, then the Virginica and the Versicolor has the lowest measures in sepal widths.

Figure 37 Violin plot of petal lengths in cm

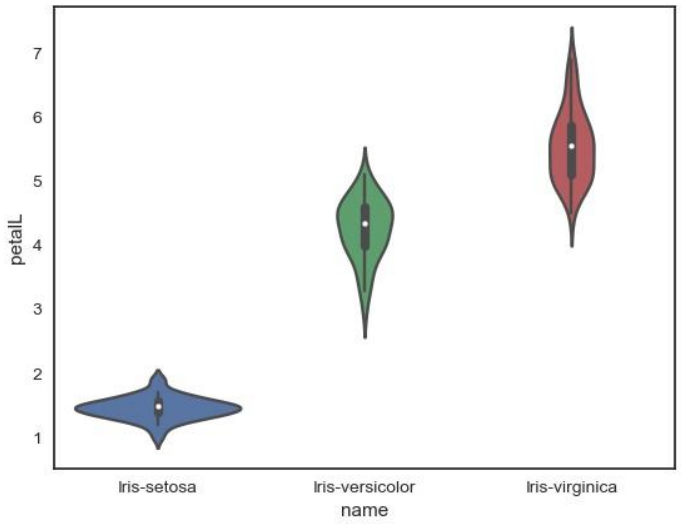


Figure 37 shows the Violin plot of petal lengths. The Iris Setosa central density is quite pronounced. The Iris Versicolor has a long tail at the lower petal lengths and the opposite pattern with a long tail in the upper petal lengths was found in the Iris Virginica.

Both the Iris Versicolor and Iris Virginica are of broadly similar shape whereas the Iris Setosa appears to be from a different distribution.

Figure 38 Violin plot of petal widths in cm

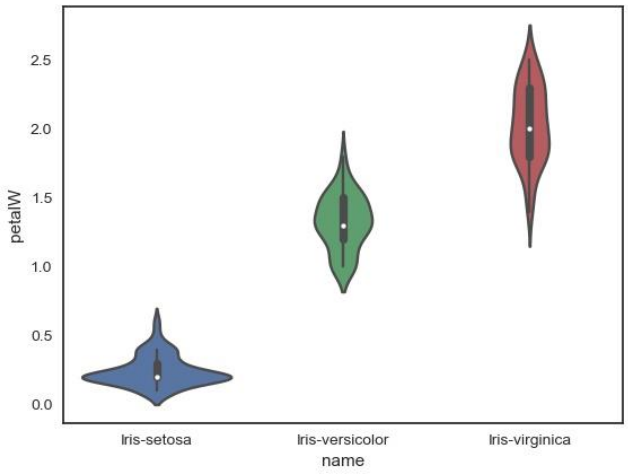
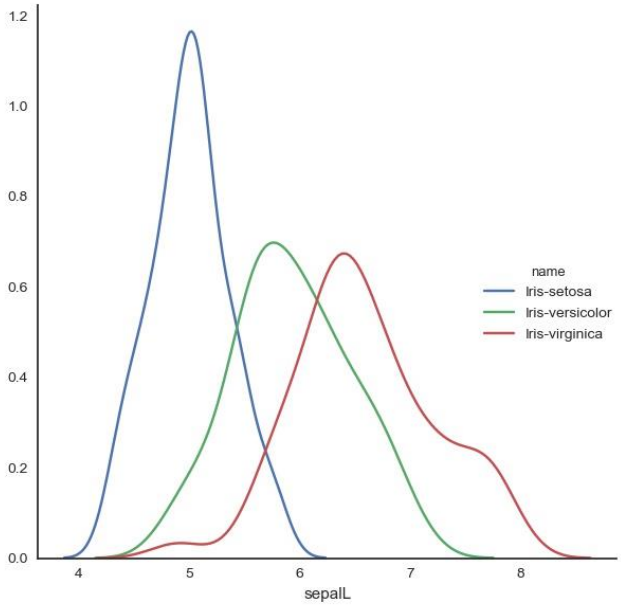


Figure 38 shows the violin plots for the petal widths. There appears to be a bimodal distribution in the Iris Setosa with a large density around the mean but another density developing in the upper petal widths. The Iris Versicolor shows a much milder version of this bimodality at the lower width values.

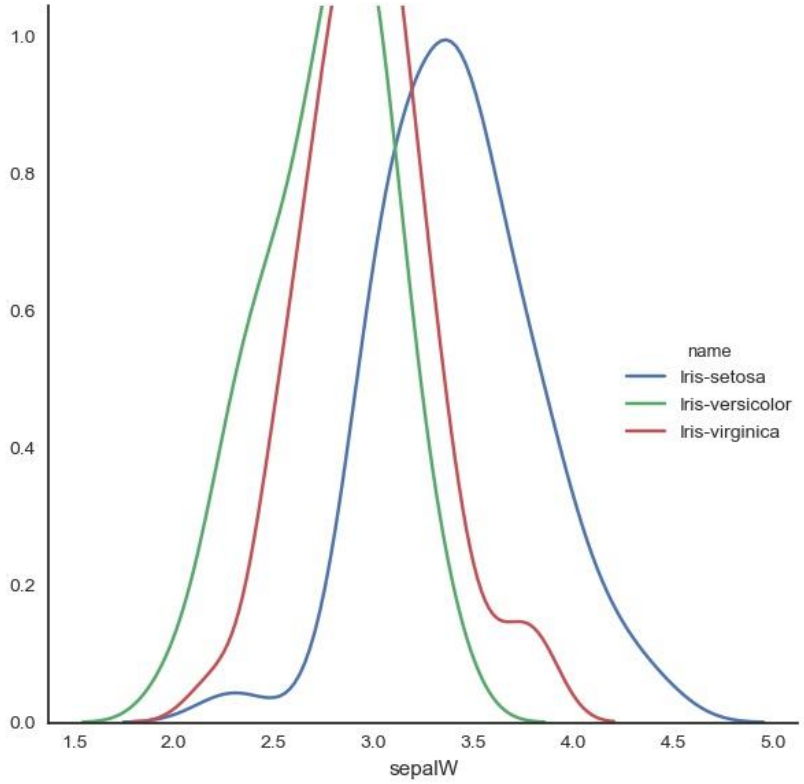
Figures 39 to 42 show the kde plots of the Iris data.

Figure 39 kde plot of Iris sepal lengths



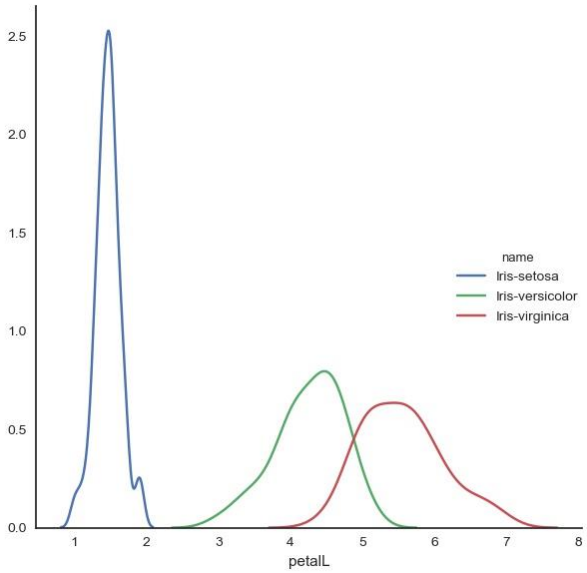
This shows that the iris Setosa differs dramatically from the Versicolor and Virginica. The Iris Setosa kde shows symmetrical distribution around the central axis so a normal distribution may apply here. It may also apply with the Versicolor sepal lengths, however the kde pattern shows several humps with Virginica suggesting bi – or higher modal distribution. There is some overlap between the higher lengths of the Setosa and lower sepal lengths of the other two but far more overlap between the Virginica and Versicolor. The Setosa's peak is higher than the highest peaks of the others.

Figure 40 kde plot of Iris sepal widths



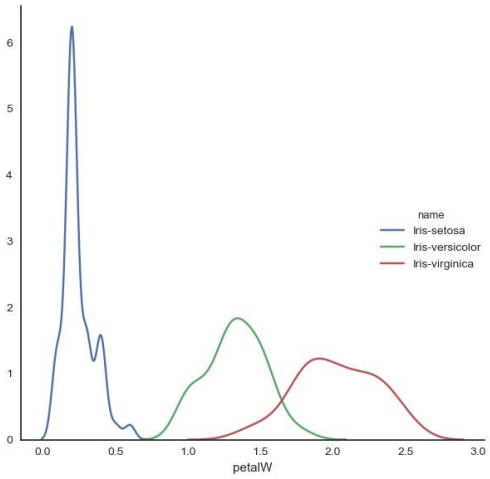
This shows Setosa sepal widths are longer than the others. There is also a suggestion of bimodal distribution with the hump on the lower left. The Virginica lies between the sepal widths of the other two and its KDE suggests bi modality with a small hump on the right. The graphs all look to be of comparable shape around their central values. However the graph is cut at the top so difficult to interpret. The Virginica shows a bump at the upper sepal widths while the Setosa shows a bump at its lower sepal widths.

Figure 41 kde plot of Iris petal lengths



This shows that the kernel density estimation is more similar for the Versicolor and Virginica, with overlapping populations than with the Setosa petal lengths which do not overlap. Whereas the Setosa has a different shape and a small peak at its upper lengths and small a hump a lower lengths. The Versicolor and Virginica show some aspects of being normally distributed; However, there appears to be some skewness in the right for the Virginica and left for the Versicolor.

Figure 42 kde plot of Iris petal widths



This shows the kde plot of Iris petal widths. The data does not appear to be normally disturbed. There appears to be 3 peaks in the Setosa plot and skewness and a mini hump on the other two species. The Versicolor has a hump on the left and is left skewed and the Virginica is right skewed with a hump on the right.

Figure 43 and 44 show the pair plots of the bivariate relationship between each pair of features. Figure 43 shows histograms along the diagonal and figure 44 shows kde plots along the diagonal.

Figure 43 pairplot of bivariate relations of the Iris data with histogram along the diagonal

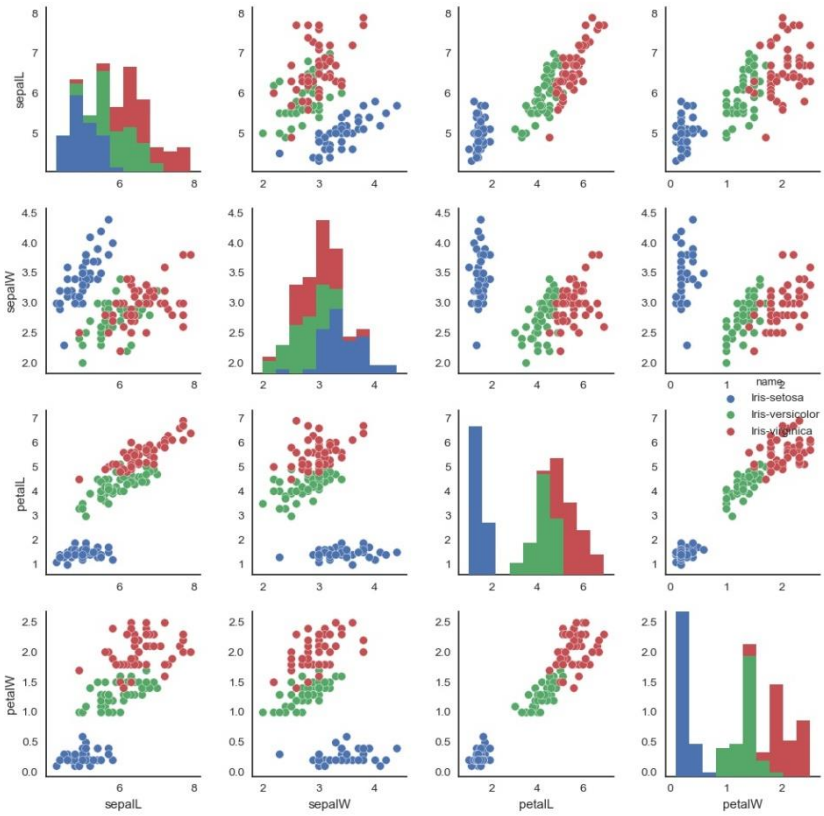


Figure 43 with histograms along the diagonal, shows each measure plotted against the other measures. In each case the Setosa cluster remains clear with the Versicolor and Virginica cluster emerging strongest with petal lengths (3rd column and 3rd row). The Versicolor and Virginica cluster appears weakest when comparing sepal widths.

Figure 44 pairplot of bivariate relations of the Iris data with kde plot along the diagonal

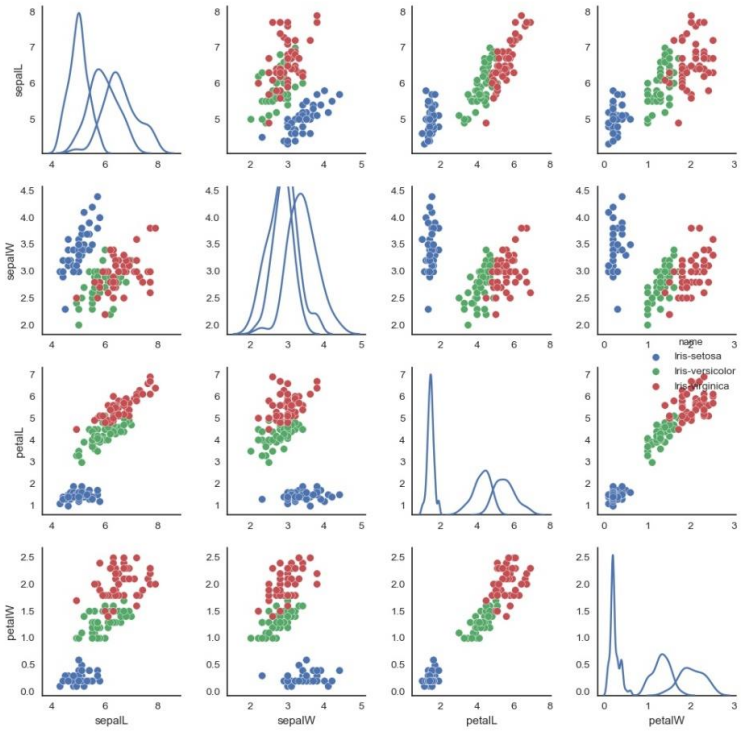


Figure 44 shows the same scatter plots as in figure 26 however here, the kernel density estimations are plotted along the diagonal for comparison.

High dimensional plots

Figure 45 Andrew Curves plot of Iris data

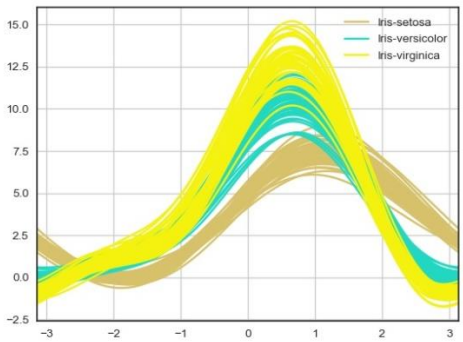


Figure 45 plots the Andrew Curves for the Iris data. This shows that the Iris Setosa (in tan) has tightly bound curves and cross the other two species at two narrow ranges at the lower and upper sections. The Setosa curves are clearly separate from the other two. The Setosa

pattern is down up down whereas the pattern for the others is up down up. There is a great deal of overlap with the Iris Versicolor and Iris Virginica, particularly in the lower ranges, with hardly any in the upper range. The Iris Versicolor has smaller peaked Andrew curves in mid ranges but is larger than the Iris Setosa curves.

Figure 46 Parallel Coordinates plot of Iris data

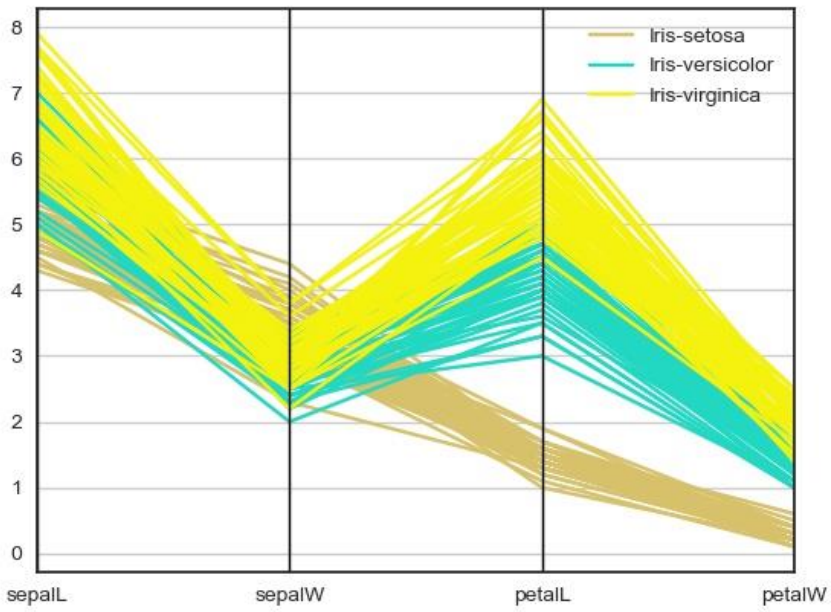


Figure 46 plots the parallel coordinates of the Iris data. This shows sepal lengths are high (between 4 and 8) for all species of Irises sampled. Sepal widths are lower for the Versicolor and Virginica and higher in the case of the Setosa. The difference between sepal lengths and widths is more extreme for the Versicolor and Virginica as seen by the low vertices at sepal widths and the less severe of the Setosa. The Setosa petal lengths and petal widths continue their downward trajectory whereas the Versicolor and Virginia move in the upward direction. The petal widths for the Setosa is its lowest values while the petal widths of the other two show some overlap with each other [1, ~2.8] and are only slightly lower than their sepal widths [2 ~4]. The Setosa petal lengths and widths show no overlap with the other petal lengths and widths whereas some overlap is present with its sepal dimensions. The data shows that the lows at sepal widths are tighter than the more disperse peak at the petal lengths and widths in the case of Versicolor and Virginia.

Figure 47 Radviz plot of Iris data

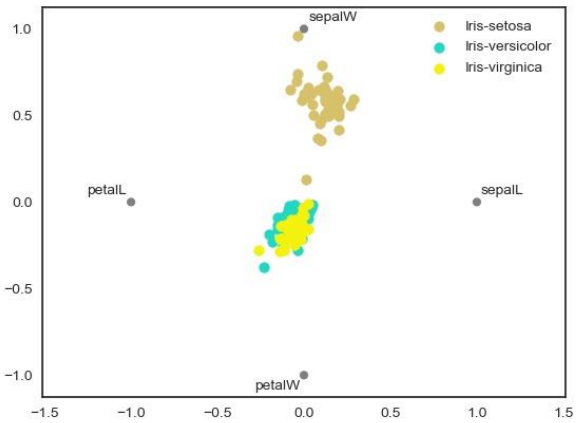


Figure 47 shows the Radviz plot of the Iris data. This shows that the Virginia and Versicolor samples are close to the centre of the plot. This indicates that the measures for these species are approximately equal across their dimensions or they have opposing pulls with one dimension going in the opposite direction to the other dimension. The Setosa data sample points (bar one near the centre) are pulled towards the sepal widths. This indicates that sepal widths are much greater than its other measures.

Inferential statistics

The data as seen by the histograms, figures 15 to 26, did not show clear evidence for normal distribution. However this could be due to the low sample size. It would be useful to test the measures for normal distribution and apply the t-test if applicable. A one way sample t – test was conducted. In all cases the null hypothesis was

H0: u1=u2 i.e. the means of the Iris species (Setosa, Versicolor or Virginica) measure (sepal or petal, length or width) were the same. The alternative hypothesis was

HA: u1 ≠ u2 that is, that the means of the Iris species measure in question were not the same.

The location for this sections code is /project/pyscripts/infer.py. This code also checked the assumptions for t-test before applying it. The results are noted below.

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) was 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, $p=<0.05$).

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) was 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, $p<0.05$).

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 with slightly less 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, $p<0.05$).

Comparing sepal lengths

A one way t-test investigated if there was 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 slightly 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, $p<0.05$).

A one way t-test investigated if there was 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, $p<0.05$).

A one way t-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, $p<0.05$).

Comparing petal widths

A one way t-test investigated if there was 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) was 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 Sestosa and versicolor failed the Shapiro Wilk test for normal distribution. Thus the t test was not conducted.

A one way t-test investigated if there was 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.

A one way t-test investigated if there was 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) so variances were unequal; The Iris Versicolor failed the Shapiro Wilks test for normal distribution hence no t-test was performed.

Comparing petal lengths

A one way t-test investigated if there was 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) was smaller and has less variation than Iris-versicolor (M=4.26, SD=0.465). Levene's test for homogeneity of variances was not 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, $p<0.05$).

A one way t-test investigated if there was 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) was 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 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, $p<0.05$).

A one way t-test investigated 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 $p<0.05$).

The above findings are summarised in table 7 below.

***Table 7*: Summary of hypotheses testing**

Iris			Sepals						Petals					
			Setosa		Versicolor		Virginica		Setosa		Versicolor		Virginia	
			Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length
Sepal	Setosa	Width			Sig t		Sig t							
		Length				Sig w-t		Sig w-t						
	Versicolor	Width	Sig t				Sig t							
		Length		Sig w-t				Sig t						
	Virginica	Width	Sig t		Sig t									
		Length		Sig w-t		Sig t								
Petals	Setosa	Width									NA		NA	
		Length										Sig t		Sig w-t
	Versicolor	Width							NA				NA	
		Length								Sig t				Sig t
	Virginica	Width							NA		NA			
		Length								Sig w-t		Sig t		

A simplified version of table 7 can be seen in table 8.

Table 8 Inferential statistics test results for Iris Data

		Setosa		Versicolor		Setosa		Versicolor	
		Sepals				Petals			
		Width	Length	Width	Length	Width	Length	Width	Length
Versicolor	Width	Sig t				NA			
	Length		Sig w-t				Sig t		
Virginica	Width	Sig t		Sig t		NA		NA	
	Length		Sig w-t		Sig t		Sig w-t		Sig t

In table 7 and 8, Sig t represents independent t test significance with equal variance, Sig w-t represents Welch's test with unequal variance and NA represents not tested due to failure of test assumptions (not normally distributed). The table showed that where the tests were significant in an independent t test for Setosa sepals with the other species they were not applicable with the Setosa petals and the other species. The variances were not equal in the Setosa sepal lengths tested against the other two species sepal lengths yet still significant. The versicolor and Virginica sepal lengths and widths were significant in an independent t-test. Equal variance was also not assumed in Setosa and Virginica t-test of petal measures but still significant.

Discussion

This project explored the Iris data set with a number of approaches. Initially descriptive analysis of the Iris data set was taken. Here, the sepal length measures had the greatest mean, minimum, maximum and inter quartile range with the petal widths having the smallest (table 2). Grouping the data showed that the sepal length data was largely influenced by the Iris Setosa sepal length. This was mostly due to the large sepal lengths of the Iris Setosa. The opposite pattern was seen for the Setosa's petal width which bore little relation to the full iris samples petal widths.

Scatter plots of the unlabelled data plotted against their respective lengths and widths showed a clear positively correlated linear relationship with petals and two strong clusters,

with one cluster being more compact than the other figure 8. The pattern in the sepal scatter plot was not as pronounced though two weak clusters could also be identified, see figure 7.

Identifying the type of Iris confirmed two clusters as the Iris Setosa was the clear cluster in the petals and the week cluster in the sepals. It was also possible to split the 2nd uncategorised cluster into two as the scatter plot for the petals showed Iris Virginica samples mostly rested along the top right of the plot with the Iris Versicolor upper values overlapping lower Iris Virginica values (figure 10). This pattern was also present with sepals, however overlapping between Virginica and Versicolor was more pronounced and covered a greater range.

Histogram plots did not show a clear normal distribution but there was signs of symmetrical density around the mean in the Setosa petal and sepal lengths (figures 18 and 16). The remaining histograms suggested bi or multimodal distribution.

Box plots of the sepal lengths compared to sepal widths showed different patterns with Setosa having the lowest box plot followed by Versicolor, with Virginica having the largest. However the largest was Setosa followed by Virginica and Versicolor being the lowest with sepal widths. The petal lengths and widths followed the same pattern as the sepal lengths with the Setosa at the lower end of the scale, Versicolor in mid-range and Versicolor at the upper end. The box plots also showed several outliers. Iris Virginia has a very low sepal length, upper and lower outliers in sepal widths. The Iris Versicolor had one low outlier in petal lengths and the Iris Setosa had several upper and lower outliers in petal lengths and 2 upper outliers in petal widths.

The violin plots (figure 35-38) showed Versicolor and Virginica data resembled each other more than they resembled Setosa. The Setosa showed most similarly with the other two species with sepal widths. All plots were broadly symmetrical and were denser close to or at their mean. The Setosa violin plot was most different from the other two for petals, particularly for petal widths figure 38.

The kernel density estimation plots (figures 39-42) showed some elements of a normal distribution, with a central peak and symmetry axis however there were signs skewness in opposite directions with the Versicolor and Virginica data and also signs of bi or

multimodal distribution. This was most pronounced with the kde plot of petal widths (figure 42).

High dimensional plots showed distribution similarity around central values between species as seen in the Andrew Curves figure 45. Parallel coordinate plot also showed Versicolor and Virginica follow the same trajectory along its petal and sepal measures whereas the Setosa differs along petal lengths (figure 46). The Setosa is also marked out in the Radviz plot as its values are heavily influenced by its sepal width (figure 47).

Inferential analysis (one way t-test) showed the Iris Setosa sepal widths were more than likely from a different normal distribution than the Versicolor and Virginica. In addition the Setosa sepal lengths also appeared to be from a different distribution from the other two but with less degree of freedom. The Versicolor and Virginia sepals also appeared to be from different populations. However the histograms showed evidence of skewness with Versicolor and Virginica sepals. This would affect the reliability of these results. It would be useful to retest using a larger sample size. While the sepals, particularly the Setosa, showed signs of normal distribution and significant differences in means, the petals were a different story.

The Setosa and Versicolor petal widths did not pass tests for normal distribution. The Setosa and Versicolor petal lengths did pass tests for normal distribution for an independent t test with significant mean differences in petal lengths between Setosa tested with Versicolor, and Versicolor tested with Virginica. Setosa and Virginica petal lengths were also significant with Welch's t-test. However skewness was present in the histograms of petal lengths of the Versicolor and Virginica so less confidence should be put in this statistic compared to the more normal looking Setosa petal lengths. Again, it would be useful to retest with a larger sample size.

In conclusion, this analysis showed several areas of interest and lines for further inquiries. Two clear clusters, with a third emerging once data are categorised were present. The Iris Setosa petals appear to not be normally distributed and heavily influenced by sepal widths (figure 47). It may be that the sepal width is influencing the distribution of petal, possibly pulling the distribution in the direction of bimodality. Sepal widths also show a different pattern than the rest of the data as seen with box plots. However, the violin plots of the Setosa sepal widths resembled the shape of the Virginia and Versicolor sepal widths, if

not its direction, more so than its other measures did. Some points to consider in this analysis include the number of samples. 50 samples from each species may not be sufficient to test for normal distribution. The presence of outliers may also indicate poor sample representation or signify an area requiring further analysis. The data used in this project contains 3 errors in 2 data points both of which come from the Iris Setosa and include 2 errors with sepal width (Iris Data, 1988). It would be useful to re analyse this data with adjustments for outliers and errors. Fisher investigated if the Versicolor was a hybrid of the Setosa and Virginica. If this conjecture has merit it may explain the opposing skewness shown by the Versicolor and Virginica. The Versicolor may be an young hybrid branch off Virginia. This would explain Versicolors' similarity to Virginica but also reflect the strong influence by the genes from the Setosa pulling sepal widths in the opposite direction from Virginica.

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

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

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