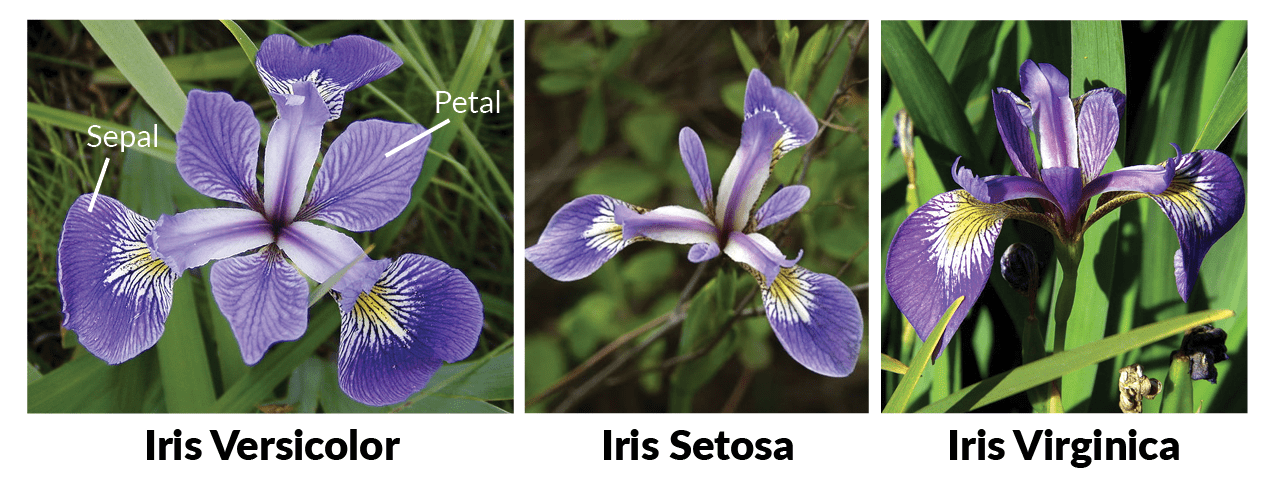
**A Statistical Analysis of Fisher’s Iris Dataset**

by Gerry Callaghan

**The Iris Flower**

The iris is a flowering plant genus of 310 accepted species with the term “iris” being widely used as a common name for all Iris species. They are perennial plants, either growing from creeping rhizomes (rhizomatous irises) or, in drier climates, from bulbs (bulbous irises). They have long, erect flowering stems which may be simple or branched, solid or hollow, and flattened or have a circular cross-section. The inflorescences are in the shape of a fan and contain one or more symmetrical six-lobed flowers (three petals and three sepals). (Source: Wikipedia1)

Source: <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Machine+Learning+R/iris-machinelearning.png>



**History of the Iris Dataset**

Almost a century ago now, in 1934, Edgar Anderson – an American botanist and geneticist collected 50 samples from two species of iris, the Setosa and the Versicolor, from the Gaspé Peninsula in Canada. Anderson measured the width and length of every petal and also the width and length of every sepal.

He set out to investigate if the similarities could show one species evolved from the other. It should be noted that the current iris dataset contains a third species of iris, the Virginica which differs from the other two samples and was taken from a different colony. It is not clear if indeed Anderson collected this third sample.

Around the time Anderson collected this dataset, Sir Ronald Aylmer Fisher, a famous mathematician and statistician, was investigating linear discrimination analysis (LDA)2, that is, a method of finding a linear combination of features that characterises two or more classes of objects or events. The resulting combination could be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

Anderson’s dataset had characteristics (which we will soon talk about) that made it suitable for Fisher’s study and was therefore used by Fisher in his 1936 classic paper, “The Use of Multiple Measurements in Taxonomic Problems”.

**Appeal of the Iris Dataset**

For Fisher, he needed a reasonably-sized dataset, and one that could be decomposed into two or more groups, and for there to be a few differentiating features. The subject matter of the data was irrelevant for this study. The iris dataset met many criteria, such as, it comprised:

- Three species or classes of iris

- Four differentiating features for each observation

- Fifty samples of each species were recorded

Therefore, each observation had five attributes, four of them being measurements (in centimetres) of the sepal’s width and length, and the petal’s width and length, while the fifth attribute was the species or class of iris to which that observation belonged.

More importantly for Fisher, based on the combination of the four measurement features, one species was said to be linearly separable from the other two, while the other two were not linearly separable from each other.

This made it ideal for Fisher to use in his paper, because he wanted to find the best linear combination of independent variables that would discriminate between the categories of the dependent variables, and determine if significant differences existed among the groups of predictor variables.

**Conclusion of Fisher’s Study**

Fisher showed that the difference between the two species, Setosa and Versicolor, was substantially greater than the standard deviations of the compound measurements. Furthermore, the difference between the species, Virginica and Versicolor, was less than four times their standard deviations. Therefore, he concluded that unlike with Setosa, the distributions of the Virginica and the Versicolor were not as easily distinguished from one another based solely on the four measurements.

**Objective of My Analysis**

I wanted to use readily available mathematical and statistical functions in python to import the dataset as a dataframe, manipulate the data into arrays for each of the three classes, which I could then use to:

* Calculate the central mean tendencies of each of the four features
* Understand the distributions of those features on histograms
* Plot scatter plots of two features I believed were interdependent
* Determine a regression line for each pair of features
* Deduce the interdependency/correlation of all four features
* Visualize the pairwise relationships using pairplot

I felt that this allowed me to confidently state how each of the three species/classes of iris could be distinguished, if at all.

**Conclusion of My Study**

According to my statistical analysis, the petal length was the most effective feature (variable) in determining the species of iris, followed closely by the petal width.

I will quickly run through each of the exercises I undertook on this dataset to arrive at this conclusion.

1. Summary Statistics

* Sepal Widths -> The whole dataset had a mean width of 3.05 centimetres (cms) and a standard deviation of 0.43 cms. The mean Setosa sepal width of 3.42 cms was well above the whole sample’s mean while the average (mean) sepal widths of the Versicolor and Virginica were much lower at 2.77 and 2.97 cms, respectively. The median and the 25%/75% percentiles showed a similar pattern in terms of the Setosa being greater than the average while the two other species were smaller.
* Sepal Lengths -> The opposite was the case when it came to the sepal length. With a mean length of 5.84 centimetres (cms) for the whole dataset and a standard deviation of 0.83 cms, the mean Setosa sepal length of 5.01 cms and standard deviation of 0.35 cms were well below those of the whole sample. In this case, the sepal length of the Virginica at 6.59 cms was much higher than the overall mean, while the Versicolor with a mean of 5.94 cms and standard deviation of 0.51 cms was close to the overall mean. The median and the 25%/75% percentiles showed a similar pattern.
* Petal Widths -> In contrast to the sepal width measurements, the width of the Setosa petal was the smallest mean at 0.24 cms, well below that of the overall mean of 1.2 cms. The Versicolor mean of 1.33 and standard deviation of 0.2 cms was closer to the overall mean while the Virginica at 2.03 cms was well above it. Again, this was mirrored in the median and 25%/75% percentiles.
* Petal Lengths -> Similar to the sepal length, the shortest petal was the Setosa with a mean of 1.46 cms compared to an overall of 3.76 cms. Both the Versicolor at 4.26 cms and the Virginica at 5.55 cms were well above the overall mean. This was mirrored in the median and 25%/75% percentiles.

Therefore, it could be seen that the Setosa was either well above or below that of the other two, never in the middle when it came to measurements. In the table below, it can be seen that for every feature, Setosa showed the greatest deviation from the mean. The gap was especially big for the petal width. Therefore, I concluded that Setosa was very different from the other two species.

| Difference from Overall Mean (%) | | | |
| --- | --- | --- | --- |
|  | Setosa | Versicolor | Virginica |
| Sepal Width | 12.13% | -9.18% | -2.62% |
| Sepal Length | -14.21% | 1.71% | 12.84% |
| Petal Width | -80.00% | 10.83% | 69.17% |
| Petal Length | -61.17% | 13.30% | 47.61% |

2. Histograms of Distributions

* Sepal widths -> The histogram of the sepal widths showed a lot of overlap between the three classes. However, the distributions for the three classes differed greatly. For example, for Setosa there were 2-3 talls bars in the middle suggesting most of its observations had a width between 3-3.5cms, while the bars above that width were much lower suggesting fewer had such large widths.

In contrast, the Versicolor had much lower bars and all to the left of 3cms approximately, suggesting move variations in the widths of its observations. Finally, Virginica, it had some very high bars between 2.75-3.25cms and then much lower bars either side suggesting most of its observations had widths of between 2.75-3.25cms. Therefore, I concluded the average Setosa sepal width was higher than the other two classes.

* Sepal lengths -> The histogram of sepal lengths also showed a lot of overlap. For Setosa, its bars were narrower and taller at the lower end of the scale. This suggested more groups of width variations between 4.25-5.25 cms approximately, and each grouping having a lot of observations.   
  Versicolor while not having bars as narrow as those for Setosa, many bars were reasonably high, especially between 5.5-6.15cms. Virginica showed higher and wider bars than many Setosa and Versicolor, suggesting its data was more concentrated. The Virginica bars were always to the right of the Setosa bars and many were to the right of Versicolor. I concluded that Setosa had the lowest mean length white Virginica had the highest mean length.
* Petal widths -> This histogram of petal widths was very different to the two previous histograms. This histogram showed little or no overlap. This immediately told me the best way to distinguish each class would be by the petal width. The Setosa distribution is concentrated in only a handful of really narrow bars at the lower end of the scale and with one of its bars dwarfing the rest of the histogram. Clearly, many of the Setosa observations had the same petal width. The bars for the Versicolor and Virginica were some distance higher up the width scale, but with the exception of 2-3 bars, all the Virginica bars were to the right of those of Versicolor. Both had bars of similar heights, suggesting that their standard deviations might be comparable. From this distribution, I concluded that the Setosa had the lowest mean petal width by some distance, while the Virginica had the highest mean width.
* Petal lengths -> This histogram, similar to the petal width, again showed little to no overlap. Furthermore, the bars for the Setosa were incredibly narrow and its distribution very much reflected that of a normal distribution. Its standard deviation around its mean looks exceptionally low, and the distance between the Setosa bars and those of Versicolor, suggest its mean is well below that of Versicolor. The distributions of Versicolor and Virginica were relatively similar in terms of shape and heights etc, but almost all the bars for Virginica were to the right of those for Versicolor. Therefore, I concluded that Setosa had the lowest mean length white Virginica had the highest mean petal length.

3. Scatter plots

* Sepal lengths versus sepal widths -> This scatter plot showed a good bit of overlap for the Versicolor and Virginica, but at no stage did any observation of the Setosa overlap with that of either of the other two classes. Its lengths were on average lower than the other two species, while its widths were higher on average. I could conclude from this that the Setosa species was very different to that of the other two Versicolor and Virginica.
* Petal lengths versus petal widths -> This scatter plot showed less overlap between the Versicolor and the Virginica, but there was a very small amount between Versicolor and Virginica. In contrast, the Setosa observations were very far away, because its petal widths and lengths were much smaller than those of the other two species, Versicolor and Virginica. This suggested that the Setosa species differed greatly from the other two classes.

4. Regression (fitted) line

* Sepal length versus sepal width -> The regression line for the Setosa may have had a relatively similar slope to that of the other two classes, but its intercept was well below that of the other two classes.
* Petal lengths versus petal widths -> The regression line for the Setosa differed greatly in terms of slope and especially intercept relative to the other two classes.

5.Boxplots

A boxplot is a chart that conveys a lot of information in a visually uncluttered and simple form.

* Each variable will have a box with an orange line. This orange line shows the median or the 50% percentile. ,
* The bottom of the box is the 25% percentile, indicating that 25% of the values are less than that. ,
* The top of the box is the 75% percentile, indicating that 25% of the values are above that. (50% of the values are in the box, 50% are outside, be it above or below),
* The whiskers either side of the box are calculated by subtracting the value at the 25% percentile from the 75% percentile (the interquartile range), and multiplying it by 1.5 either side of the box, and then letting it retreat back towards the box until it hits the nearest observation. The whisker length tells you how disperse the data is.
* Observations outside the length of 1.5 x interquartile range, are referred to as outliers - they are indicated by circles above the top whisker or below the bottom whisker. If they exist, they are the maximums and minimums, otherwise the end of the whiskers when they hit the nearest observations are the maximums and minimums.

Petal Lengths

* Medians:   
  Looking at the orange lines, the Setosa petal median was about 1.5cms, Versicolor was about 4.5cms, and Virginica was about 5.5cms. In all three cases, the orange line is in the *middle* of the respective boxes, so I concluded there was little skewness in the data for each of the classes.
* Interquartile Ranges:  
  The interquartile range for the Setosa was very tight, while it was relatively consistent between the Versicolor and Virginica, although Virginica was still bigger. Given that the interquartile range shows where 50% of the values lie, then I could say that the variance of the length of Setosa petal was much smaller than the other two. This reflected what I found when calculating the standard deviations in the summary statistics earlier (Setosa 0.17, Versicolor 0.47, and Virginica 0.55).
* Whiskers:  
  The whiskers for the Setosa were very short, while the whiskers for the Virginica were the longest suggesting the greatest variance in its data. Also, the lower whisker for the Versicolor was longer than its upper whisker suggesting it had observations at a further distance below the median than it had above the median, ignoring outliers. For Virginica, the opposite is the case, the upper whisker was longer than its lower whisker suggesting it had observations at a further distance above the median than it had below the median, ignoring outliers.
* Outliers:  
  There were some outliers above and below the whiskers for the Setosa class, not too far above and below it must be said. Given how tight the data was for Setosa, this was expected. For the Versicolor class, I already said how the lower whisker was longer than its upper whisker, well there was an outlier even below the bottom whisker.
* Overlap:  
  There was very little overlap between the three classes of iris, absolutely no overlap between Setosa and the other two. There was some overlap between high values of Versicolor against low values for Virginica. However, it was only the lower whisker of the Virginica, (that is 25%) that overlapped with the upper whisker of Versicolor and the top of the box (about 50% of Versicolor).
* I concluded from this boxplot, that the Setosa was very different class to the other two.

Petal Widths

* Medians:  
  Looking at the orange lines, the Setosa petal median was around 0.25cms, while Versicolor was about 1.4cms, and Virginica was about 2.0cms.(From my summary statistics above, the Setosa mean was 0.24, Versicolor was 1.33, and Virginica was 2.03). For Setosa, the orange line was at the bottom of the box, indicating that its data was skewed – a quarter of its observations were compressed in the very bottom of the box, while a quarter filled the rest of the box. For the Versicolor, while the data was also skewed with the medican closer to the bottom of the box, the skewness was much lower. The Virginica, likewise, had only a small amount of skewness.
* Interquartile Ranges:  
  The interquartile range for the Setosa was again very tight. This time, the interquartile range for Versicolor was much smaller than that of Virginica. Given that the interquartile range tells me where 50% of the values lie, then I could say that the variance of the width of Setosa petal was much smaller than the other two. The variance of the width of the Virginica petal was the greatest. This tallies with my standard deviations calculated in my summary statistics previously (Setosa 0.11, Versicolor 0.2, and Virginica 0.27).
* Whiskers:  
   The whiskers for Setosa were very short, while the whiskers for the Virginica were the longest suggesting the greatest variance in its data. Also, the upper whisker for the Versicolor was longer than its lower whisker suggesting it had observations at a further distance above the median than it had above the median, ignoring outliers. For Virginica, the opposite was the case, the lower whisker was longer than its upper whisker suggesting it had observations at a further distance below the median than it had above the median, ignoring outliers.
* Outliers:  
  There were some outliers above the whiskers for the Setosa class (I saw earlier in my summary statistics, the maximum value was 0.6). For the Versicolor and Virginica classes, there are no outliers.
* Overlap:  
  There was no overlap between Setosa and the other two classes of iris. There was some overlap between high values of Versicolor against low values for Virginica. However, it was only the lower whisker of the Virginica, (that is 25%) that overlaps with the upper part of the Versicolor box, while the upper whisker of Versicolor only barely reaches the lower part of the box of Virginica.

I deduced from this boxplot, that again Setosa was very different to the other two classes.

Sepal Lengths

* Medians:  
  Looking at the orange lines, the Setosa sepal median was around 5cms, while Versicolor was about 6cms, and Virginica about 6.5cms.(From my summary statistics above, the Setosa mean was 5.01, Versicolor was 5.94, and Virginica was 6.59). For Setosa, the orange line was in the centre of the box, indicating that its data had little skewness. For Versicolor, while the data was also skewed with the median closer to the bottom of the box, the skewness looked quite small. Virginica, likewise, had only a small amount of skewness to the lower part of its box.
* Interquartile Ranges:  
  The interquartile range for the Setosa was again very tight. This time, the interquartile range for the Versicolor was very similar to that of Virginica. Given that the interquartile range tells me where 50% of the values lie, then I could say that the variance of the width of Setosa petal was much smaller than the other two. Again this tallied with the standard deviations in my summary statistics calculated earlier (Setosa 0.35, Versicolor 0.51, and Virginica 0.63).
* Whiskers:  
   The whiskers for Setosa were much longer than they were for its petals, in fact, they were similar to the whiskers for Versicolor and Virginica, suggesting more variance in the sepal data. While the whiskers for Setosa and Versicolor were relatively consistent above and below their respective boxes, for Virginica the upper whisker was longer than its lower whisker suggesting it had observations at a further distance above the median than it had below the median, ignoring outliers.
* Outliers:  
  There were outliers well below the lower whisker for the Virginica class (I saw earlier in my summary statistics the minimum was 4.9). For the other two classes, there were no outliers.
* Overlap:  
  There was more overlap between Setosa and the other two classes of iris in this boxplot. There was some overlap between upper whiskers of Setosa against low values of Versicolor (box) and even Virginica (whiskers). Furthermore, upper box of the Versicolor, overlaps with the lower part of the Versicolor box.

I was unable to deduce from this boxplot, that when it comes to their sepals lengths, one of the classes could be easily distinguished from the other two.

Sepal Widths

* Medians:  
   Looking at the orange lines, the Setosa sepal median was around 3.4cms, Versicolor was about 2.75cms, and Virginica about 3.0cms.(From my summary statistics above, the Setosa mean was 3.42cms, Versicolor was 2.77cms, and Virginica was 2.97cms). For all three boxes, the orange line was moreorless in the centre of the box, indicating the data had little skewness.
* Interquartile Ranges:  
  The interquartile range for the Setosa was very similar to that of Versicolor and Virginica. In fact, in this case, Virginica looked to have the tightest box. Given that the interquartile range tells me where 50% of the values lie, then I could say that the variance of the width of the Virginica sepal was as small as the other two. The standard deviations in my summary statistics calculated earlier back this up (Setosa was 0.38cms, Versicolor 0.31cms, and Virginica was 0.32cms),
* Whiskers:  
  The whiskers for the Setosa were much longer than they were for its petals, in fact, they were even longer than the whiskers for the Versicolor and Virginica, suggesting more variance in its sepal data. The lower whiskers for the Setosa and Versicolor were longer than their upper whiskers suggesting it had observations at a further distance below the median than it had above the median, ignoring outliers.
* Outliers:  
  There were outliers well below the lower whisker and well above the upper whisker for the Virginica class (I saw earlier in my summary statistics, the minimum was 2.2cms and maximum 3.8cms). For the other two classes, there were no outliers.
* Overlap:  
  There is more overlap between Setosa and the other two classes of iris in this boxplot. In fact, this is the first boxplot where the box for Setosa is far higher than the other two. There is some overlap between the box for Versicolor and the box for Virginica.

Overall, I would deduce from this boxplot, that when it comes to their sepals widths, the classes are relatively closer to one another.

6. Correlation Coefficients

Given that I was trying to prove that one of the species (Setosa) was very different to that of the other species, I looked at the correlations between the features in each class of iris.

Setosa  
Relative to the overall iris classification, there was a massive difference in correlation between the sepal width and length for Setosa (-0.10 vs. 0.75). Likewise, there was a large difference for petal length and sepal length (0.87 vs. 0.26), and petal width and sepal length (0.82 vs. 0.28). The relationship between the sepal width and petal length went from a negative 0.42 to a positive 0.17, while that of petal width and sepal width went from a negative relationship of 0.36 to a positive 0.28. Finally, it can be seen that the relationship between petal length and petal width went from a massive 0.96 to a much lower 0.31. So, I concluded the Setosa class differed greatly to the overall iris dataset.

Versicolor  
Relative to the overall iris classification, a massive difference in correlation was seen between the sepal width and length for Versicolor (from -0.10 to 0.52). Likewise, there was a large difference for petal length and sepal length (0.87 to 0.75), and petal width and sepal length (0.82 to 0.55). The relationship between the sepal width and petal length went from a negative 0.42 to a positive 0.56, while that of petal width and sepal width went from a negative relationship of 0.36 to a large positive of 0.66. Finally, it can be seen that the relationship between petal length and petal width went from a massive 0.96 to a more realistic 0.79. Therefore, while the correlations for Versicolor differed to the overall iris dataset, the difference was less marked than that of Setosa.

Virginica  
Relative to the overall iris classification, there was a massive difference in correlation between the sepal width and length for Virginica (from -0.10 to 0.45). However, there was only a small difference for petal length and sepal length (0.87 to 0.86). There was a large difference in the relationship between petal width and sepal length (0.82 to 0.28). Finally, the relationship between the petal length and petal width was much weaker. The relationship between the sepal width and petal length went from a negative 0.42 to a positive 0.40, while that of petal width and sepal width went from a negative relationship of 0.36 to a positive of 0.53. Finally, it can be seen that the relationship between petal length and petal width went from a massive 0.96 to a more modest 0.32. Therefore, similar to Versicolor, while Virginica differed to the overall iris dataset, but not as marked as Setosa.

7. Pairplots

According to Analytics Vidhya a pair plot, also known as a scatterplot matrix, is a:

“...matrix of graphs that enables the visualization of the relationship between each pair of variables in a dataset. It combines both histogram and scatter plots, providing a unique overview of the dataset’s distributions and correlations. The primary purpose of a pair plot is to simplify the initial stages of data analysis by offering a comprehensive snapshot of potential relationships within the data.

A pair plot consists of:

* **Histograms**: Diagonal plots showing the distribution of a single variable.
* **Scatter plots**: Off-diagonal plots showing the relationship between two variables. These can reveal patterns, trends, and correlations.

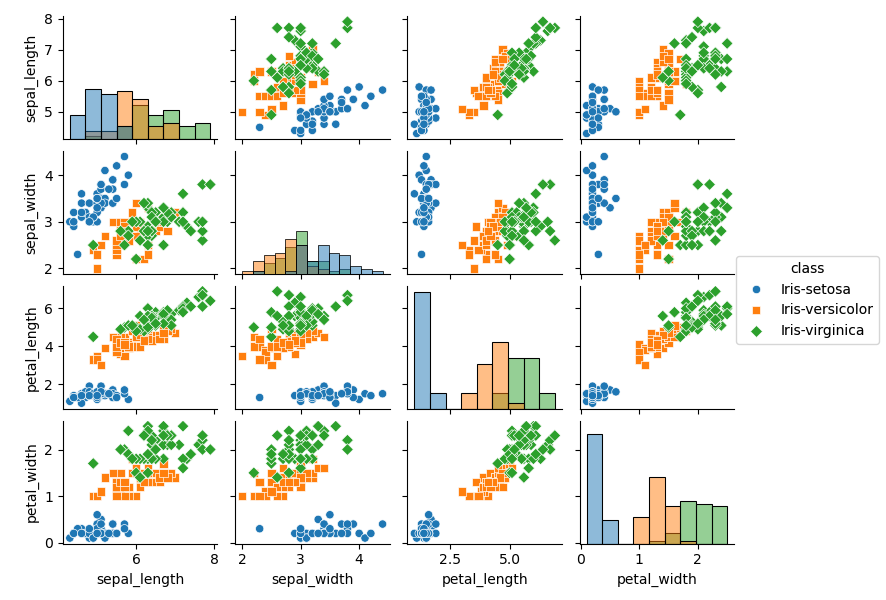
They enable data scientists to:

* **Visualize distributions**: Understand the distribution of single variables.
* **Identify relationships**: Observe linear or nonlinear relationships between variables.
* **Detect anomalies**: Spot outliers that may indicate errors or unique insights.”

The seaborn package, which is a high-level visualization package based on matplotlib and integrated with pandas’ data structures, allowed me to create a pairplot of the iris dataset.

Along the diagonal, the histograms showed the distributions of that particular feature, be it petal width, petal lengths, sepal widths, or sepal lengths. I’ve already talked about these histograms above and there was no additional information in these histograms along the diagonal.

What was very important from this exercise was that it gave a great visualization of the data in the one overall chart. When I looked at both the petal length and the petal width columns, I could clearly see that there was some overlap in the scatter plots for each pair of features, for Versicolor and Virginica. But, there was absolutely no overlap with the Setosa scatter plot, so I knew its correlations were very different to that of the two other classes.

****

**Summary**  
According to the description of the data on the UC Irvine website where the original dataset was sourced (via a personal github account), they stated

“one class is linearly separable from the other two; the latter are not linearly separable from each other.”

My analysis would lead me to agree with the dataset’s description, that one class (Setosa) can be linearly separable from the other two.

**Bibliography**

1. <https://en.wikipedia.org/wiki/Iris_(plant)>

2 <https://en.wikipedia.org/wiki/Ronald_Fisher>

**References**

https://www.angela1c.com/projects/iris\_project/the-iris-dataset/

https://www.kaggle.com/datasets/uciml/iris

https://g.co/gemini/share/14f512935c90

https://www.analyticsvidhya.com/blog/2024/02/pair-plots-in-machine-learning/

http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.html

https://zion-oladiran.medium.com/exploratory-data-analysis-iris-dataset-68897497b120

https://www.kaggle.com/code/abhishek0032/learn-classify-the-iris-dataset-adventure

https://www.ashokcharan.com/Marketing-Analytics/~sma-analysis-and-visualization-of-the-iris-dataset.php