Kmeans Clustering to Classify Various Iris Species

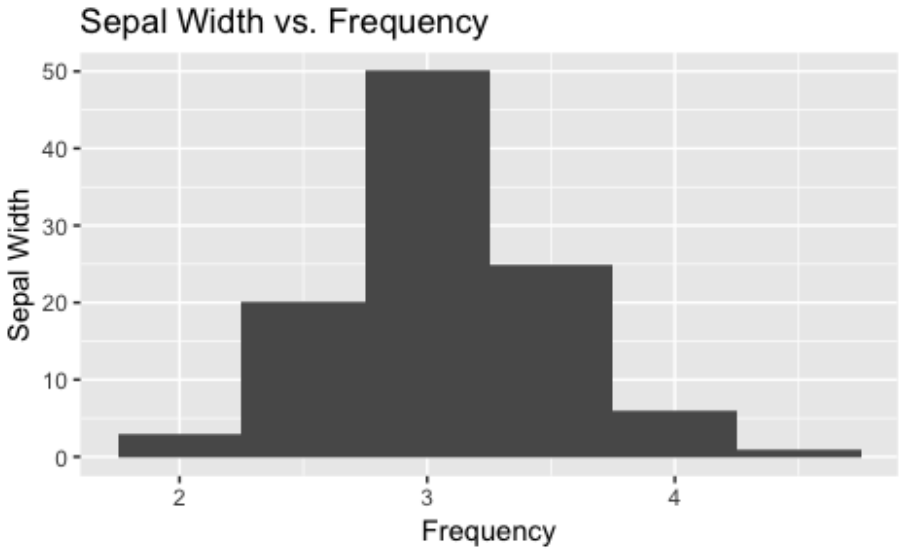
Jeffrey Cheng

06-18-19

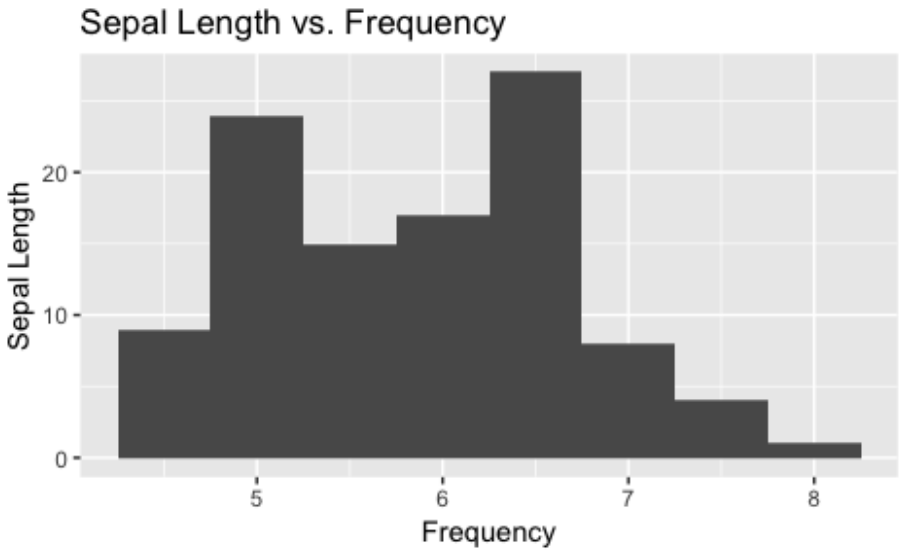
1 Introduction

This is a basic Kmeans clustering of a dataset of 3 different Iris Species from UCI. The dataset was split into a training set to create the clusters (70 percent) and a testing set to test the model with (30 percent). The goal of the project was to determine averages of sepal length, sepal width, petal length, and petal width for each species respectively.

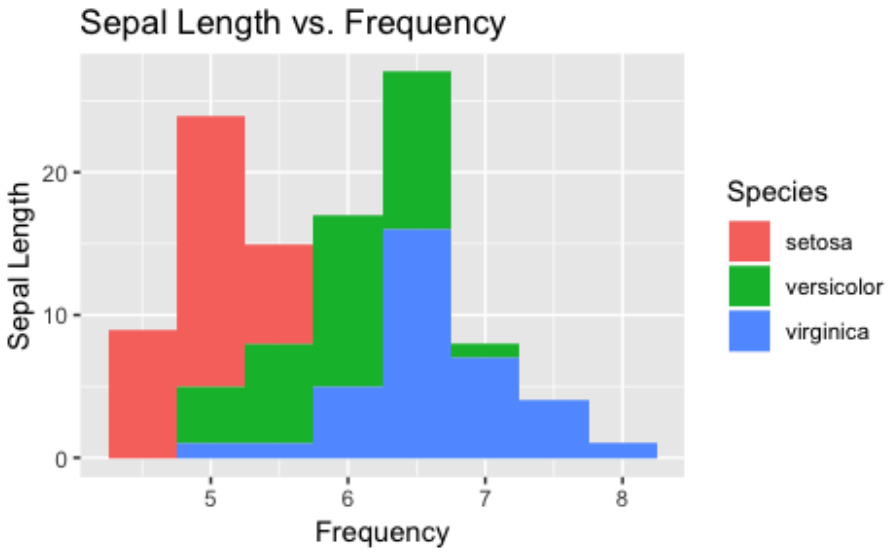
2 Examining the Data



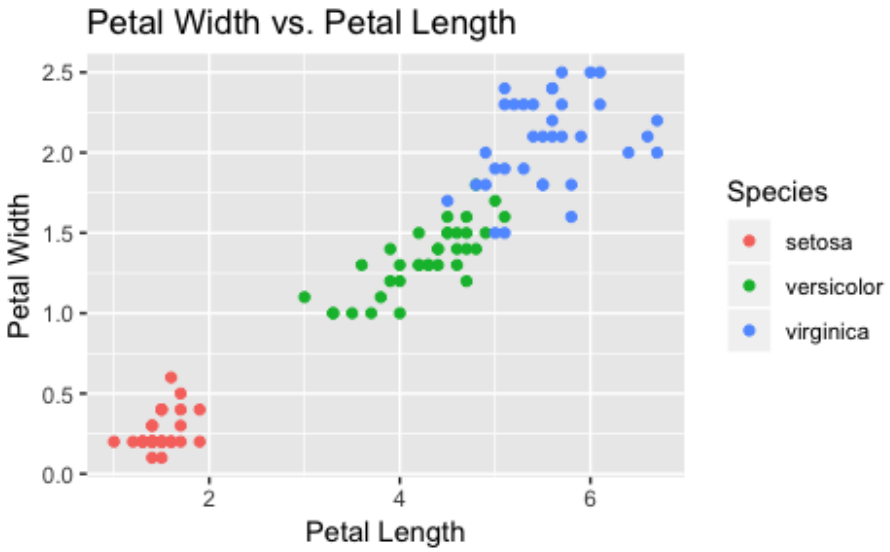
Sepal width appears to be fairly normally distributed, so it is unlikely that it has a significant relation to species.



Sepal length appears to be roughly bimodal, so it is possibly that one species has a significantly longer sepal length than another. To investigate this further, color can be added to the diagram that differentiates between species.



It's very clear from this graph that sepal length is a defining factor between different species of irises.



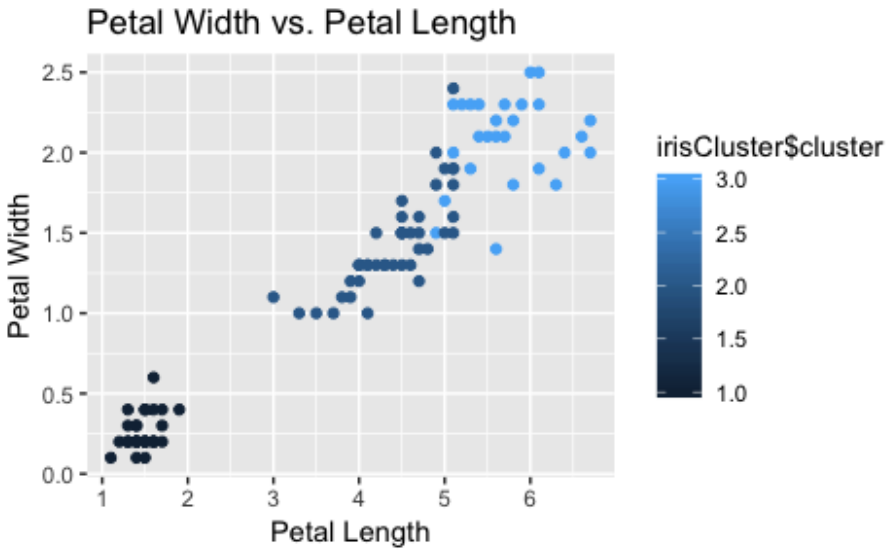
If we create a graph of petal width vs. petal length and differentiate color by species, it is very clear that petal width and petal length are defining factors between different species of irises as well.

3 Kmeans Clustering

A Kmeans clustering algorithm was run with 3 centers and 20 iterations.

Cluster	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.045714	3.448571	1.471429	0.2542857
2	5.888372	2.727907	4.406977	1.4348837
3	6.877778	3.081481	5.737037	2.0814815

Graphing petal width vs. petal length again and differentiating color by cluster shows how accurate the clusters are.



The graphs are almost identical when grouped by species and by cluster.

4 Average Dimensions for each Species

By matching respective species and clusters on the graphs above we can easily figure out approximate average dimensions for each species

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
Setosa	5.045714	3.448571	1.471429	0.2542857
Versicolor	5.888372	2.727907	4.406977	1.4348837
Virginica	6.877778	3.081481	5.737037	2.0814815

5 Conclusion

In conclusion, there are three defining factors for iris species: sepal length, petal length, and petal width. A Kmeans clustering shows this very clearly. Sepal width was shown to be not very important in classification of species.