# Week 7 – Assignment 2 – Unsupervised Learning

By: Zach Adair

Regis University

MSDS 650 – Data Analytics

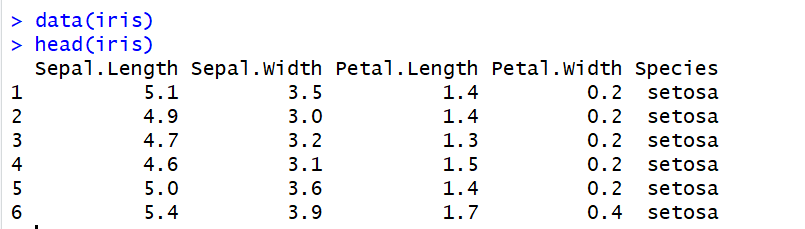
## Introduction

For this week’s second assignment I will be walking through an unsupervised learning exercise with the Iris data set as well as answering some questions that are related to unsupervised learning. I will start by going through the exercise and then at the end I will answer the questions and wrap up with my final thoughts on the assignment.

## Unsupervised Learning Exercise

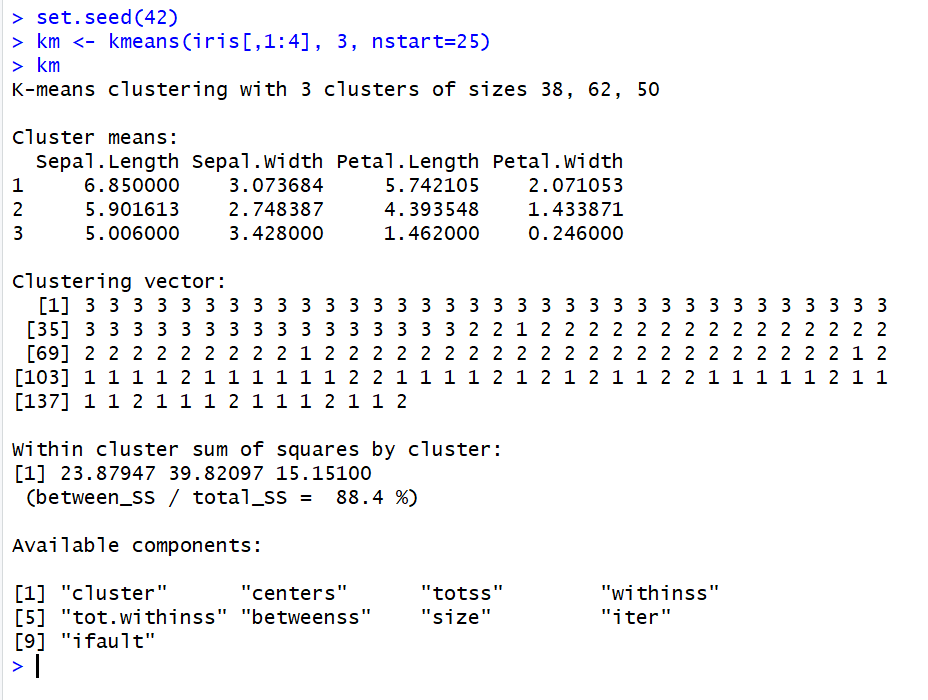
### Load the Data

Can’t do any analysis without any data, so I will start by loading a dataset, for this assignment I will be using the Iris dataset.



### Set seed and apply K means Function

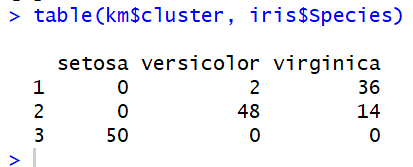
In this step we will be starting by setting a seed to the data, then I will set up and run the k means function with three clusters of the iris data being created and tested. The reason three clusters are chosen is because there are three different types of iris’s in the dataset.



Now that the k means output is here we can compare the different cluster compiled and then plot the results.

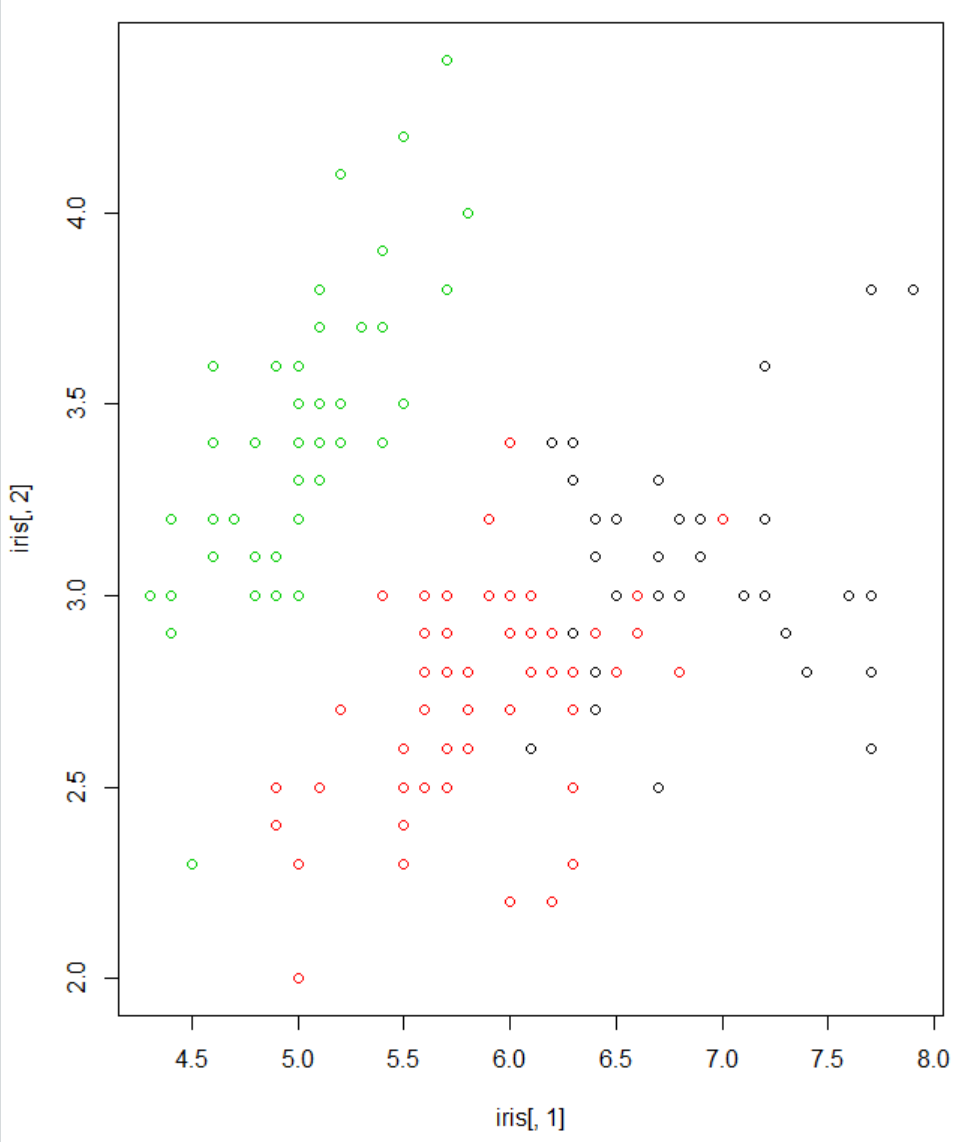
### Compare Clusters with the Species and then Plot Results

We will start by comparing the clusters to the species, and looking at which species were in which cluster.



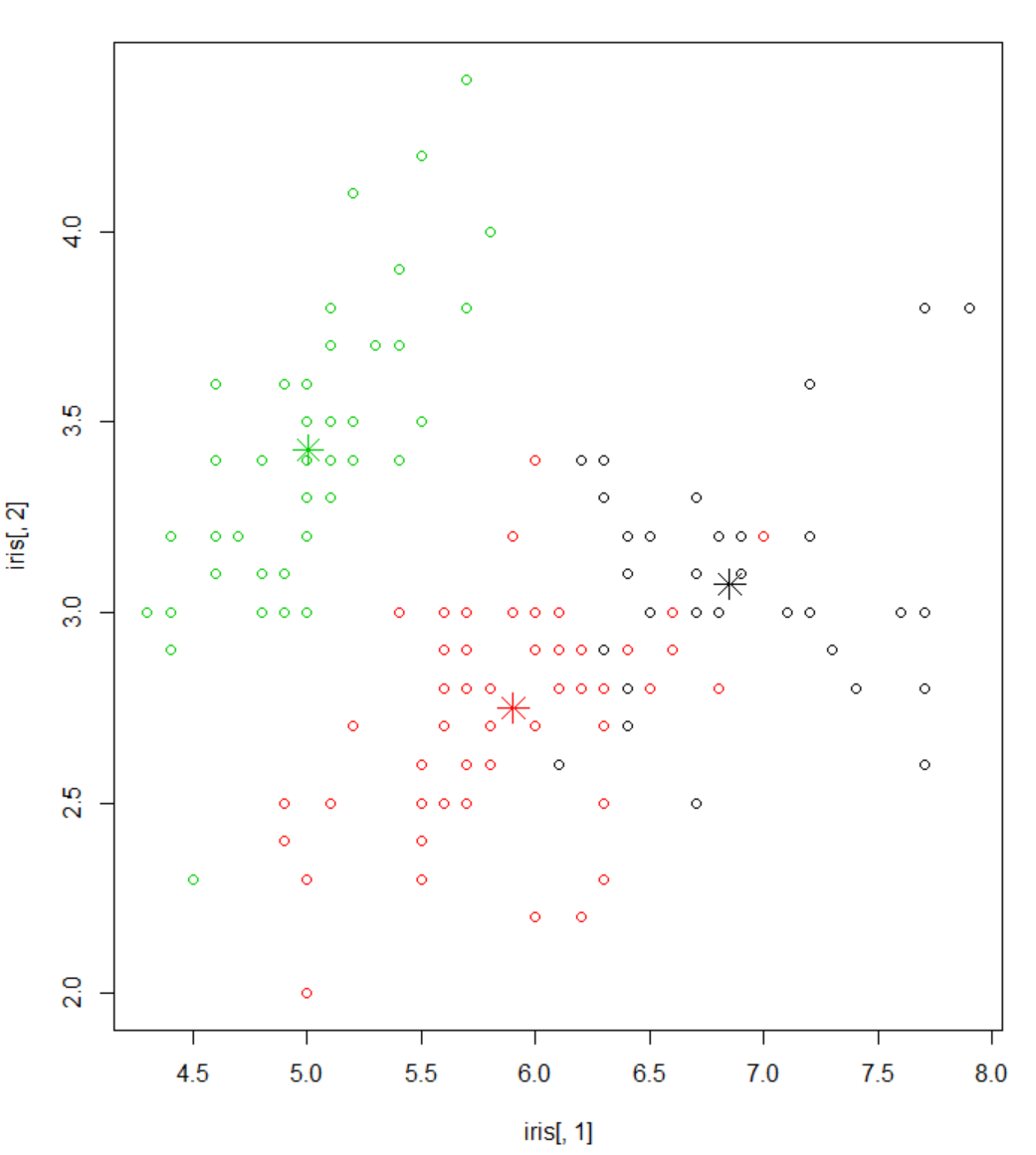
From the table created, the first cluster was compiled of 2 versicolor iris’s and 36 virginica iris’s; the second cluster was made up of 48 versicolor iris’s and 14 virginica iris’s; and finally cluster 3 was compiled of all 50 setosa iris’s. The iris’s that were the minority within the cluster they were assigned to were wrongfully put into their cluster. .So in this scenario, there are 2 versicolors which are wrongfully classified and 14 virginica iris’s that were wrongfully classified. Now that we know which iris’s are in which cluster, let’s plot the data.





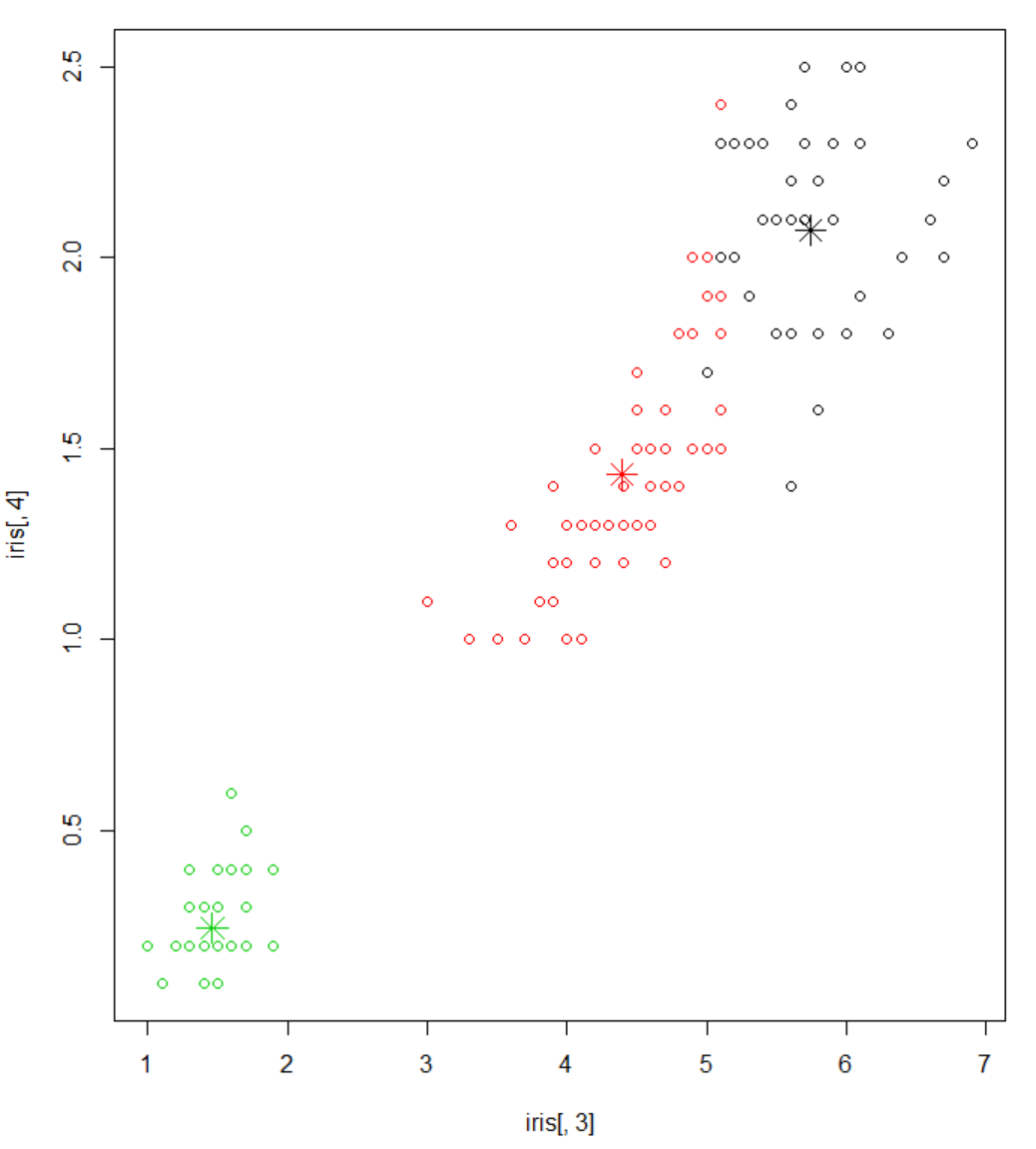
As you can see from the initial plot, the green points of the data are of cluster 3, cluster 2 appears to be the red points, and cluster 1 appears to be the black points. It would be interesting to see the center of each cluster so next we will plot the clusters with a center pointed out specifically.





With the centers showing more, it is easier to find where the cluster’s centers are located. Next, we will plot the lengths and widths of petals for the 3 species.





Within this last plot, it is much easier to determine which points within clusters 1 and 2 are they associated with and it is very easy to recognize the differences in petal length and width between the 3 species and you can conclude that there is a clear difference between the setosa and both the versicolor and the virginica iris’s, with those two being a little mixed but still fairly distinguishable by petal length and width with the last plot.

## Questions

1. **How is unsupervised learning related to the statistical clustering problem?**

Unsupervised learning and statistical clustering are related to each other because with unsupervised learning you are taking data which has no real labels and your having the algorithm put together some sort of input and you can do that by classifying data or clustering it. The statistical clustering problem is one of the most used techniques by unsupervised learning because it is a way to find relation in the data and group it together either based on whatever distinct similarities or differences it finds in the group of records.

1. **What packages in R and Python perform unsupervised learning?**

RStudio: pvclust, mclust, cluster, fpc,

Python: DBSCAN, t-SNE

1. **What measures of quality for the learning algorithm might you expect to see?**

Classification accuracy is always a good one to measure, so whenever the model is set against the actual data and the model errored in classifying the cluster or class certain data points are supposed to be assigned to it, that is negatively counted against the model. A confusion matrix can help measure the learning algorithm in the same way, so in a 2x2 matrix the algorithm will have 4 options and two of them would mean that they predicted correctly and two of the options would mean the algorithm chose incorrectly.

## Summary

Exercise two was also a very informative exercise for this week. I learned a lot more about the differences between supervised and unsupervised learning and also getting to learn about reinforced learning through my readings and research was really cool because I think my definition of each was kind of blurred together so it was nice to clean those up. Unsupervised learning actually gave me a lot of good ideas for my actual profession because of how it uses clustering and classification for data mining, I think that is useful technique when dealing with a new data set and you are unsure about any trends that are potentially within the set. The questions I thought were useful as well, over the course of my reading and research as well as going through the assignment you really get a feel for how important clustering is to unsupervised learning, so much so it appears to me that clustering is by far the most important technique used for an unsupervised learning model because it helps establish the mappings that it is left without when compared to supervised or reinforced learning strategies.