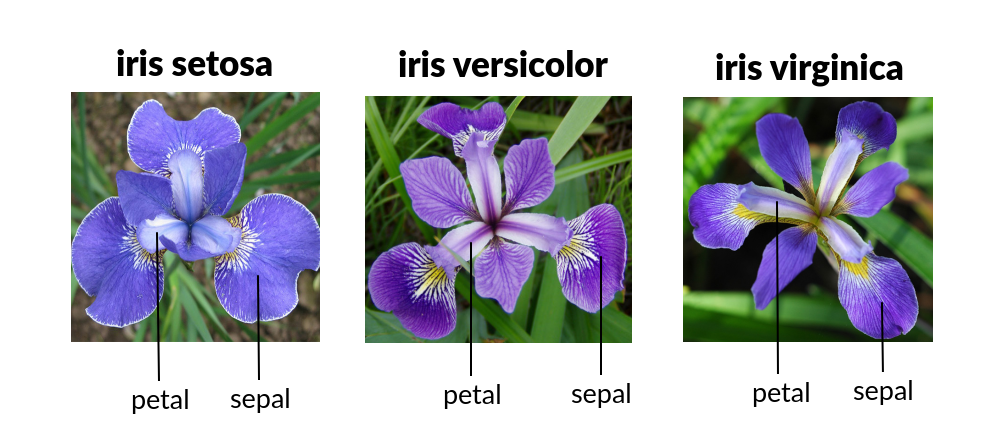
**IRIS DATASET**

**EDA, KNN, and K-MEANS – Assignment 1**

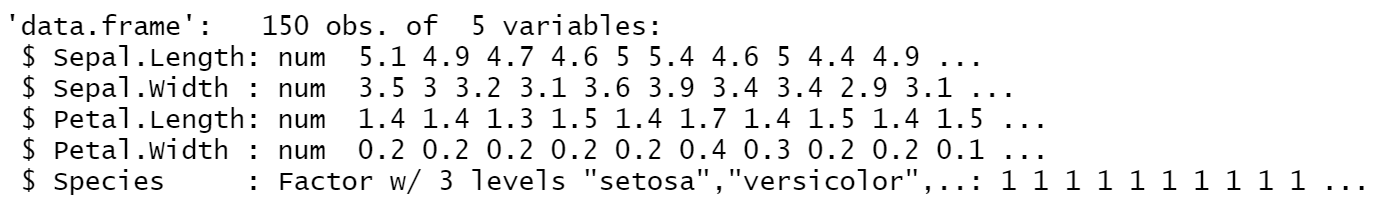
**Chris Steege and Shuki Saito**

**Exploratory Data Analysis:**

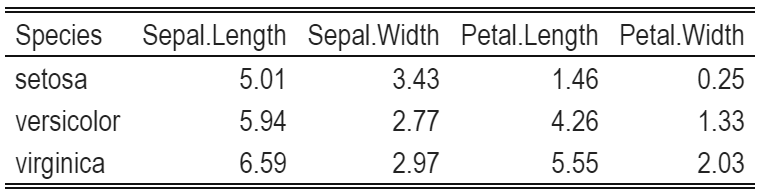
1. **Background**

The Iris Flower data set was collected and used in the 1936 paper *The use of multiple measurements in taxonomic problems* as an example of linear discriminant analysis by a British statistician, eugenicist, and biologist named Ronal Fischer. The data was originally collected by Edgar Anderson for the purpose of quantifying the morphological spectrum of sepal and petal attributes of 3 species of Iris flowers. There are 50 observations in this data set associate with each species (Iris Setosa, Iris Virginica, and Iris Vericolor). The sepal/petal lengths and widths are measured in centimeters.

1. **Summary and Statistics**

We want to understand the types of data we are dealing with. By checking the structure using the str() function built in R, it reveals that we have four numeric data types and one factor data type. The factor data type is our species. Factor data types are useful, because R can automatically create dummy variables our of them. We have 150 observations and 5 attributes associated with each attribute. The column names are Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width.

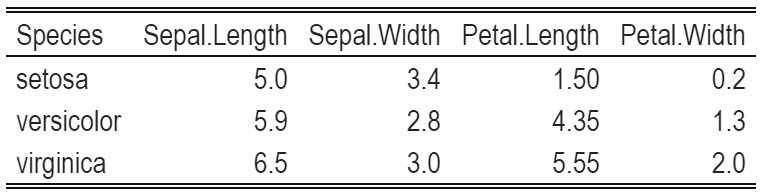
**Table 1: Mean Values of Attributes by Species**



The table indicates the following about our data set:

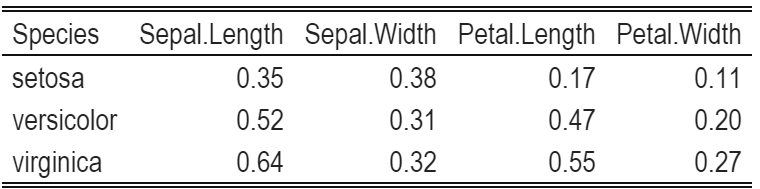
* Smallest mean Sepal Length: Iris Setosa, Largest mean Sepal Length: Iris Virginica
* Smallest mean Sepal Width: Iris Versicolor, Largest mean Sepal Width: Iris Setosa
* Smallest mean Petal Length: Iris Setosa, Largest mean Petal Length: Iris Virginica
* Smallest mean Petal Width: Iris Setosa, Largest mean Petal Width: Iris Virginica

**Table 2: Mean Values of Attributes by Species**



The table indicates the following about our data set:

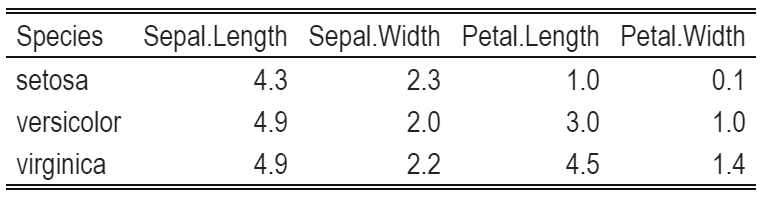
* Smallest median Sepal Length: Iris Setosa, Largest median Sepal Length: Iris Virginica
* Smallest median Sepal Width: Iris Versicolor, Largest median Sepal Width: Iris Setosa
* Smallest median Petal Length: Iris Setosa, Largest median Petal Length: Iris Virginica
* Smallest median Petal Width: Iris Setosa, Largest median Petal Width: Iris Virginica

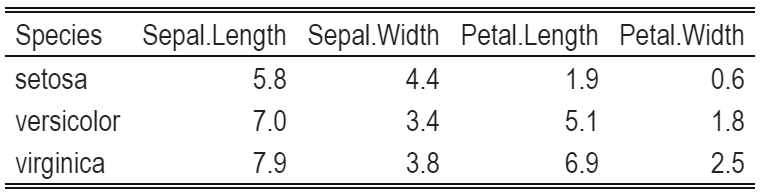
**Table 3: Standard Deviation of Attributes by Species**

This table indicates the following about our data set:

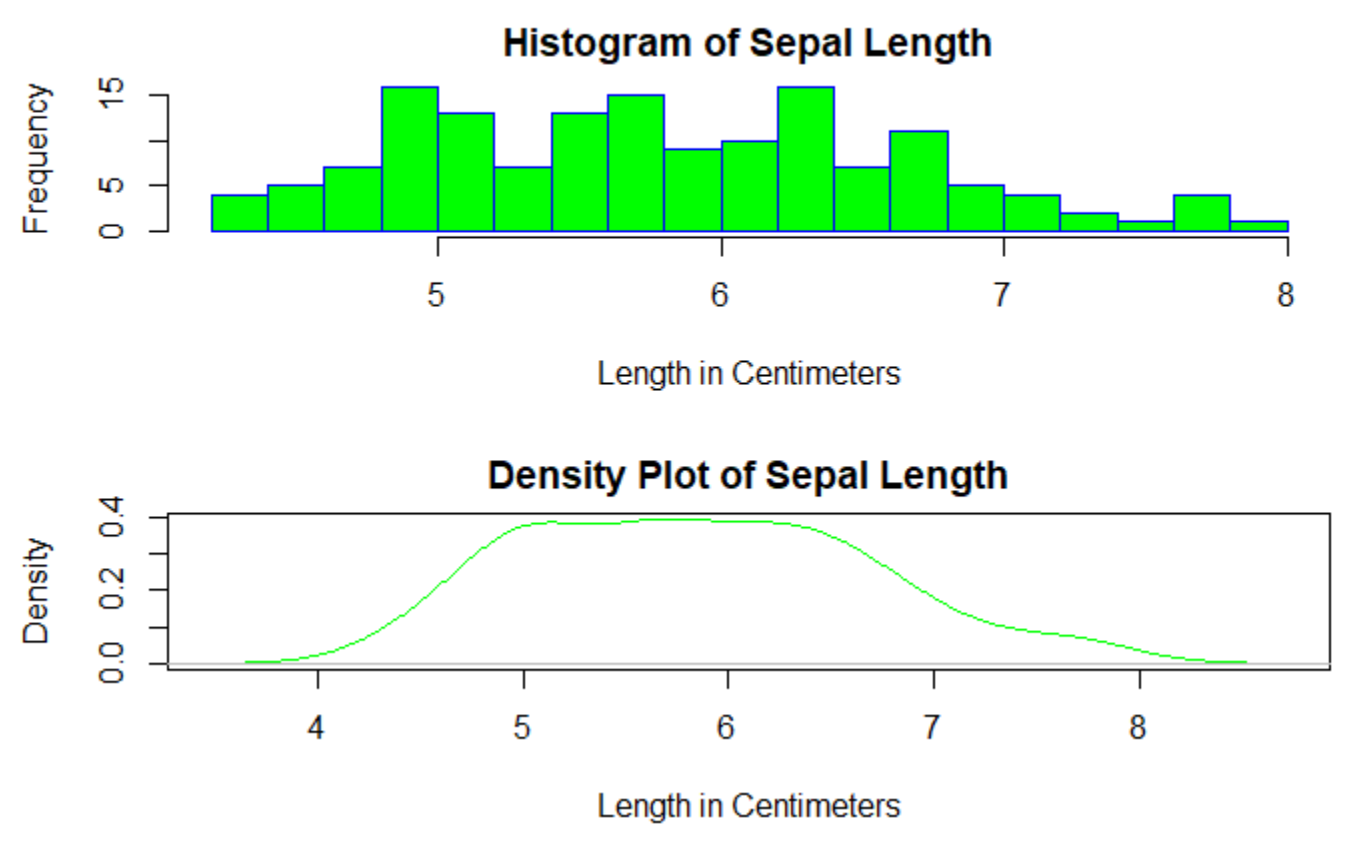
* Smallest SD of Sepal Length: Iris Setosa, Largest SD of Sepal Length: Iris Virginica
* Smallest SD of Sepal Width: Iris Versicolor, Largest SD of Sepal Width: Iris Setosa
* Smallest SD of Petal Length: Iris Setosa, Largest SD of Petal Length: Iris Virginica
* Smallest SD of Petal Width: Iris Setosa, Largest SD of Petal Length: Iris Virginica

Upon analyzing the mean and median values for our four attributes of each flower type, the data seems to be fairly normally distributed. This seems to be the case, because the mean and median tend to be close to each other. You could also conclude there is a slight positive skew in the data because it is slightly more common that the attributes have higher means than medians, though, it does not seem to be very significant. I slight positive skew makes sense because there is a lower bound (zero) on length and width. The smallest and largest mean, median, and standard deviation all correspond with each other.

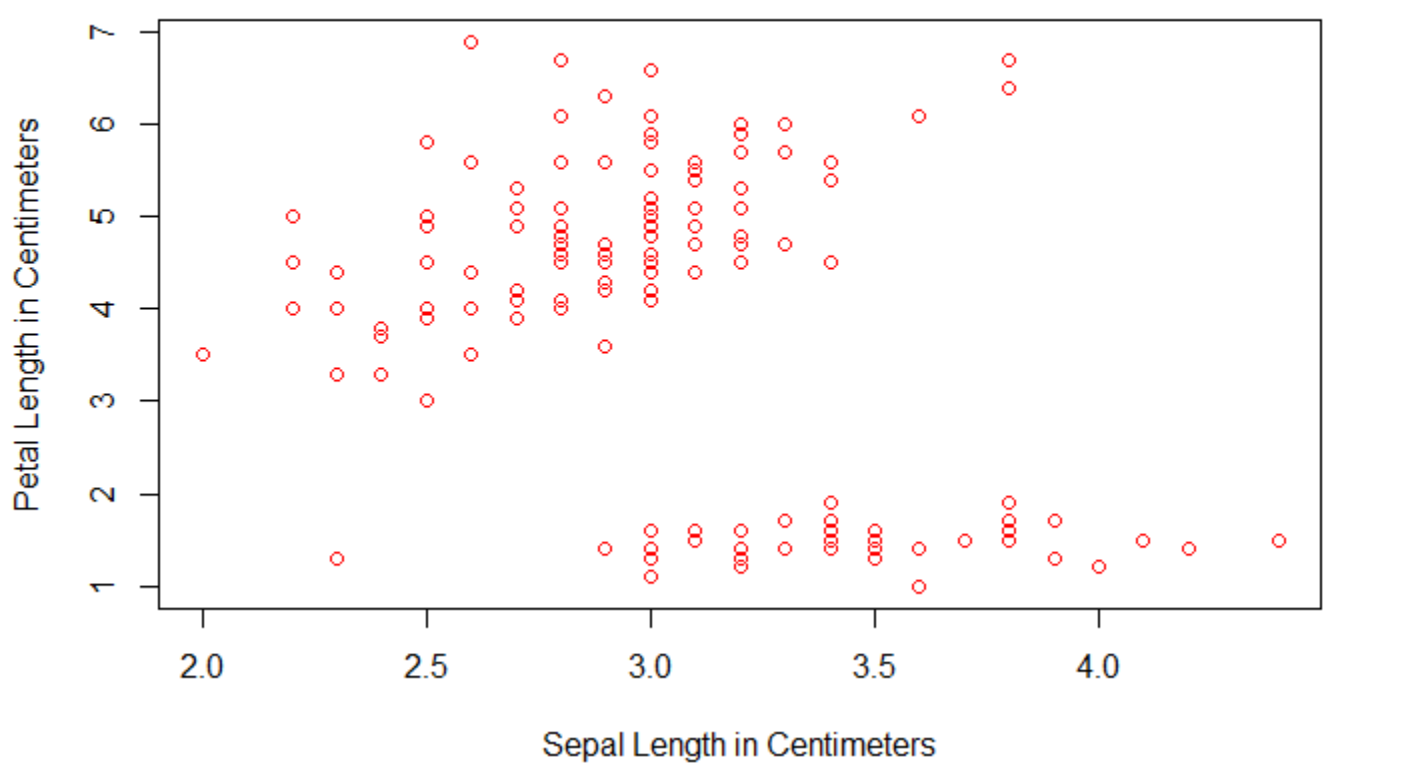
**Table 4: Minimum Values of Attributes by Species**

**Table 5: Maximum Values of Attribute by Species**

Upon inspection of our maximum and minimum values, it does not appear that there are any outrageous outliers. The maximum and minimum values for each species also correspond with the species with the largest and smallest means, medians, and standard deviations.

**Plot 1 and 2: Histogram and Density plot of Sepal Length**

The density plot and histogram reveal that the sepal length of the flowers is not very skewed. In fact, the distribution looks quiet normal – though not fully. The density is nearly even around the mean and only decreases once the length becomes significantly larger or smaller than the mean.

**Plot 3: Scatter Plot of Sepal Width vs Petal Length**

The scatter plot has interesting structure because it appears that a line would do well to separate and classify the point. It is likely that the points in the bottom right are Iris Setosa flowers, because they have a high mean sepal width, but a low mean petal length. The other two flowers do not have as drastic of a sepal width/petal length ratio as Setosa. The top left would then contain Iris Versicolor and Iris Virginica. This is reminiscent of the original purpose of this data set for Linear Discriminant Analysis.

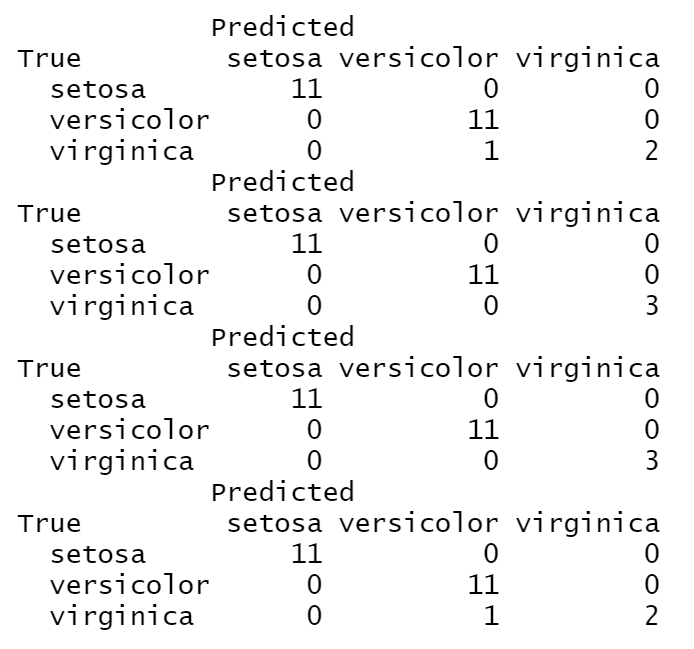
**Supervised Learning:**

The procedure in supervised learning is to separate the data into a train set and a test set. The training set is used to train the algorithm on your data set and then the test set can be used to give an evaluation of the algorithm trained by your training data set. Typically, you split the data 80/20 (train/test), but this can be adjusted to suit your needs. Supervised learning deals with data that has some response variable that you want to predict or classify.

The Iris data set is an excellent candidate to use a supervised learning classification algorithm, such as KNN, if we want to predict the species. We have some evidence that this may be an effective procedure from the scatter plot analysis we have done earlier. KNN works by classifying a data point based on the classes that are nearby. If the classes are well distinguished by their continuous variables without too significant overlap, then It is likely that we will have success.

Below I will train multiple KNN algorithms using various K values. The K values tells the algorithm how many neighbors to investigate before determining the candidate class. The algorithm will choose whichever class appears most within its neighborhood.

**Table 5: KNN Using Multiple K Values**



K=3

K=5

K=7

K=10

After training the algorithm four times using different K values, I went ahead and tested the prediction accuracy. I got slight variation in the results. Our median choices of K for KNN worked perfectly with no inaccurate predictions. Using 3 and 10 neighbors led to 1 misclassification. This is due to the shortsightedness of the algorithm in the first case and the algorithm being too open-minded in the latter case. The algorithm needs the “right” sized neighborhood to perform optimally. Somewhere between 5-7 seems like a great choice.

**Unsupervised Learning:**

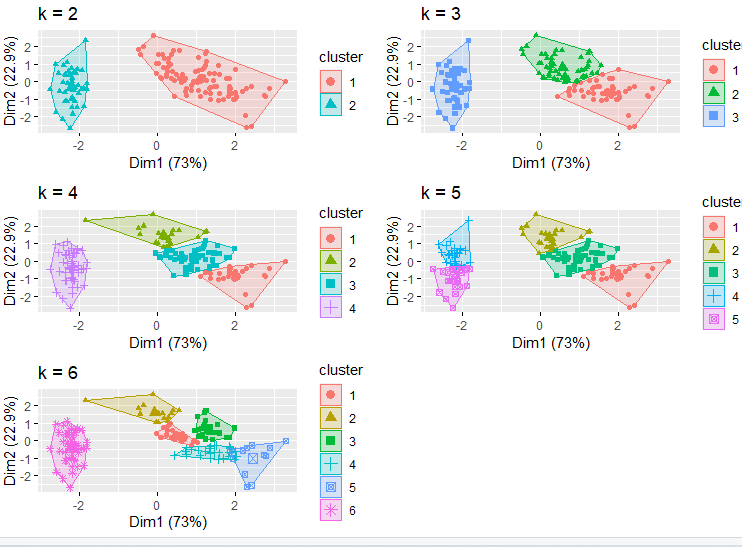
Unsupervised learning is a type of algorithm used in order for pattern recognition. In this algorithm, there is no target value for the clustering process. Clustering is the process to discover groups of similar features within a dataset.

K-mean Clustering is a common method for unsupervised learning, where a dataset will be distributed into K number of clusters based on the similarity of data. Since there is no target value in this method, the result of clustering might be varied. Therefore, it might be necessary to conduct several trials in order to recognize decent patterns in data.

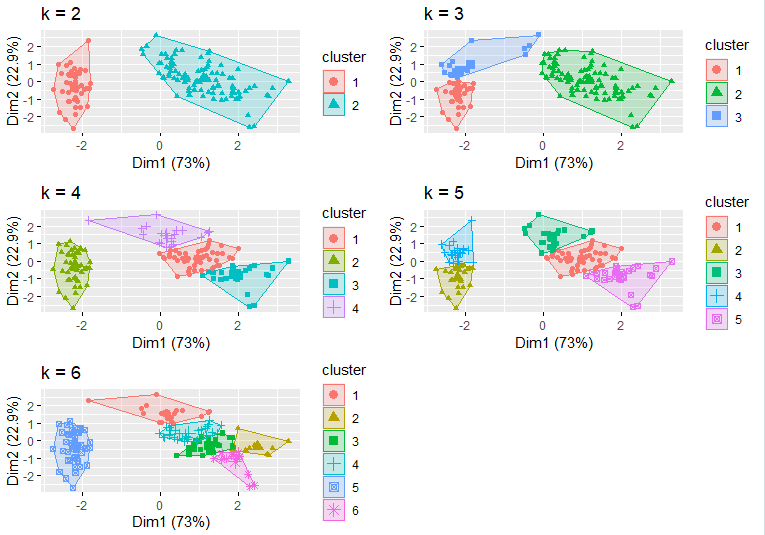
In K-means clustering, it is important that the data is normalized in order to maintain accuracy of clustering.

Below are the results of K-mean clustering for Iris dataset.

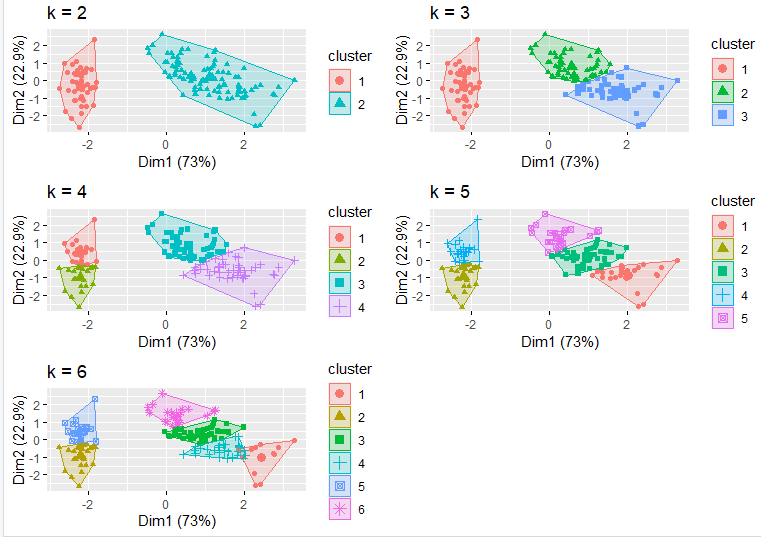
**Plot 4: 1st trial for K-Means Clustering**



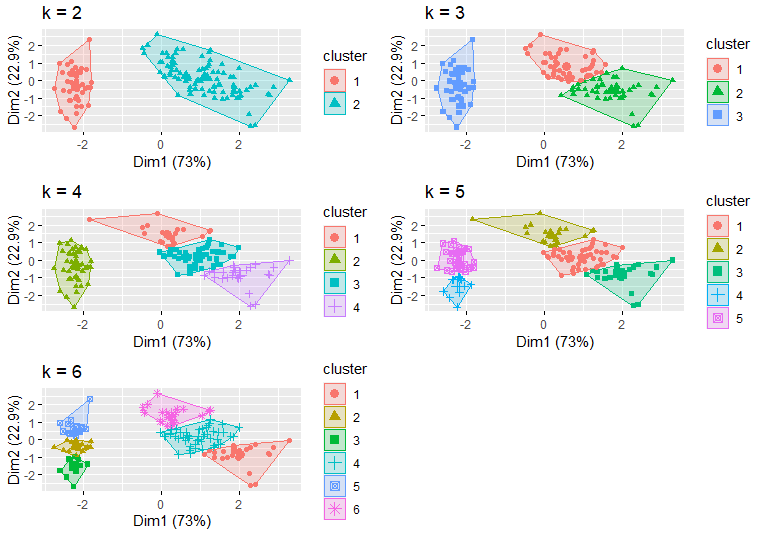
**Plot 5: 2nd trial for K-Means Clustering**



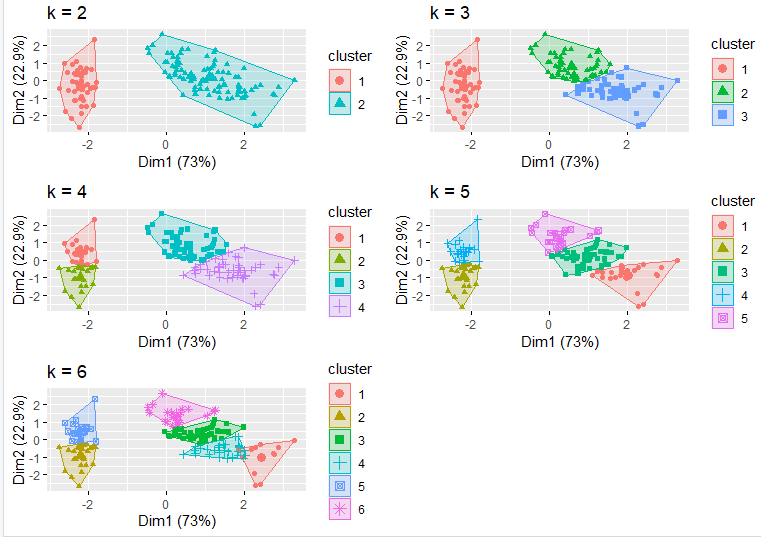
**Plot 6: 3rd trial for K-Means Clustering**



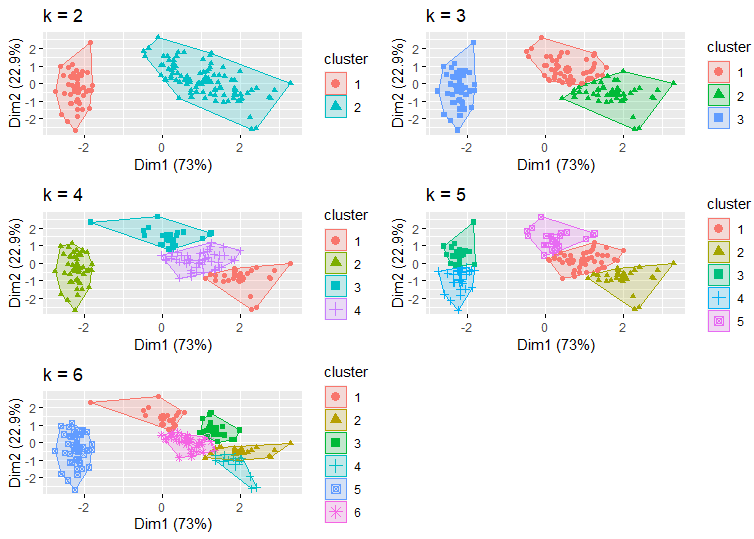
**Plot 7: 4th trial for K-Means Clustering**



**Plot 8: 5th trial for K-Means Clustering**



**Plot 9: 6th trial for K-Means Clustering**



As it is seen in the plots, the clustering result varies in each time when K = 6, while the results look stable when 2 >= k >= 5. Therefore, the Iris data is probably divided into 2 to 5 groups based on the data similarity. Based on what we know about the data set, it is likely that there are either 2 or 3 groups, because we have only 3 species of Iris flowers. It could be the case that there are more groups than this if it is the case that the flowers of the same species were not sampled under the same conditions. The group on the left of our plots appears to represent a distinct species, because it is well separable. It is not clear from the plots alone that the right-side group, which likely contains the other two species of Iris flowers, is clearly separable by these two attributes alone. From what we have understood from the EDA, we know that the different species of Iris flowers do have different distributions for their petal and sepal attributes. Knowing this and our previous results, we should be able to tease apart the different species and accurately classify the species using different combinations of attributes or including more than two.