**Kmeans Clustering on Iris DataSet:**

**How To Run:**

1. First the kmeans.py have to be run, to get the means and the cluster sets which will be used to plot the graphs in kmeansplot.py
2. Run the kmeansplot.py to see various sets of graphs plotted between two features , three features and four features.
3. Also, kmean.py can be run to be see the The Elbow curve, used for determining which k has to be used for the Iris dataset.

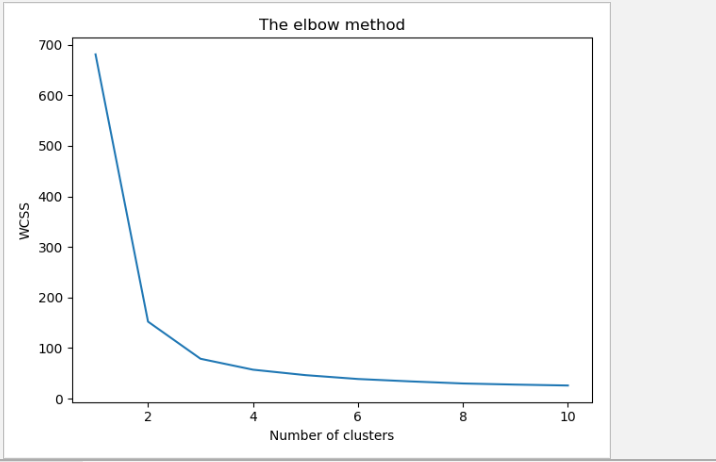
**Implementation:**

1. In the kmeans.py, we have function to find the Euclidean distance and the OriginalMeans to find the means between the datapoints.
2. Then updatingMEans fucntion can be used to update the first set of means.
3. The findingClusters can be used to see where the sets of data can be joined according to the Euclidean distance calculated and the means generated.
4. After the clusters are formed, the algorithm keeps on iterating the same code until it reaches a point where it reaches constant clustering mechanism.
5. We can see this in the code, where a large number of clusters are generated and in the end, it gets stopped at a particular dataset point.
6. Also minmaxOfColumns fucntion is used to find the minimum and the maximum of the feature set. According to this feature set, we will generate a new first random mean set.

**Beyond the implementation,**

**The Elbow Curve** is one of the methods to do the evaluation of the Iris data set to see which “k” value will be efficient enough for the dataset.

The Elbow curve below can be seen,



As we can see in this image, the curve keeps on decreasing in the terms of WCSS i.e. within cluster sum of squares.

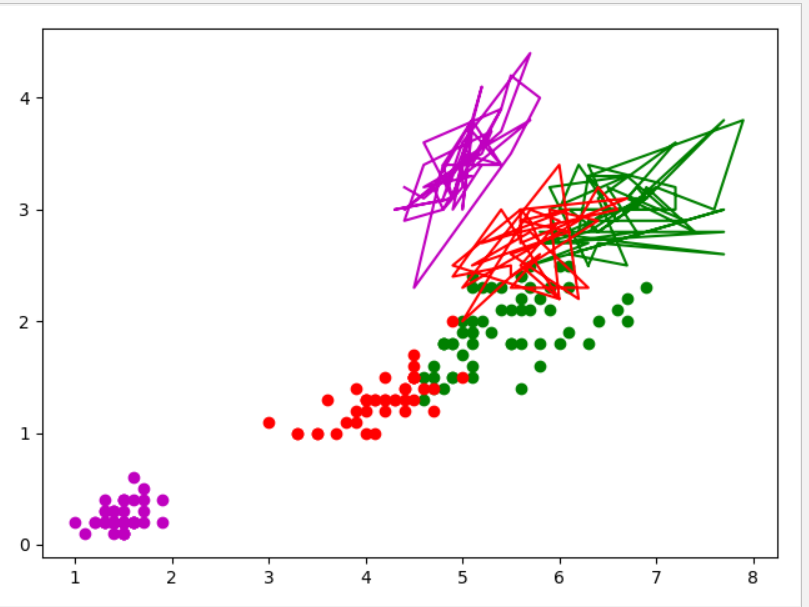
Now, after k = 3, these sum of squares keep on getting constant. This is an indication that the distances between the data point are not being decreased and forming better clusters anymore. So we can safely, assume that k = 3 will give us the best “k” value.

After we get the means and cluster points from the kmeans.py file, we will import this file to the kmeansplot.py and then there we will, plot these points on a matplotlib graph.

**Graphs:**

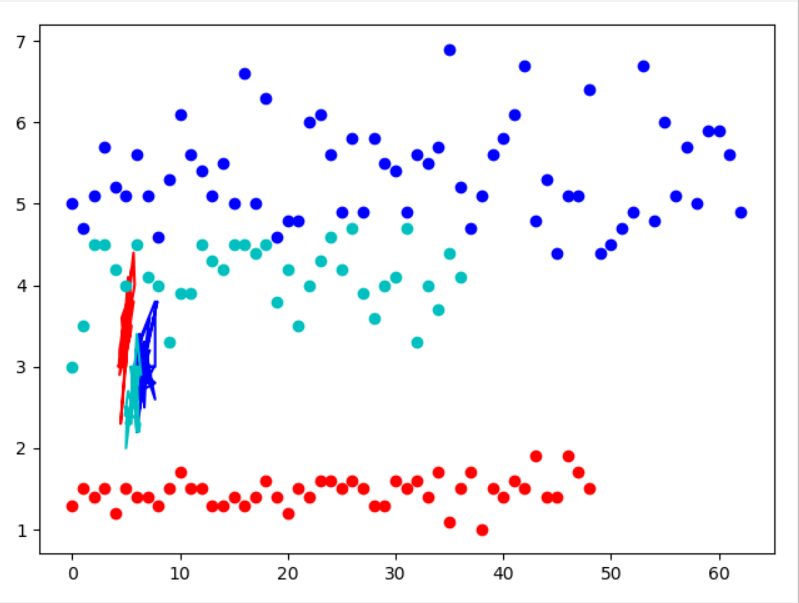
Some of these variations of these graphs are:

When 4 features are being used,



Also, we can see few other variations of the curves we may get when the feature values change but the “k” value remain constant.

When feature value = 3



When feature value = 2

