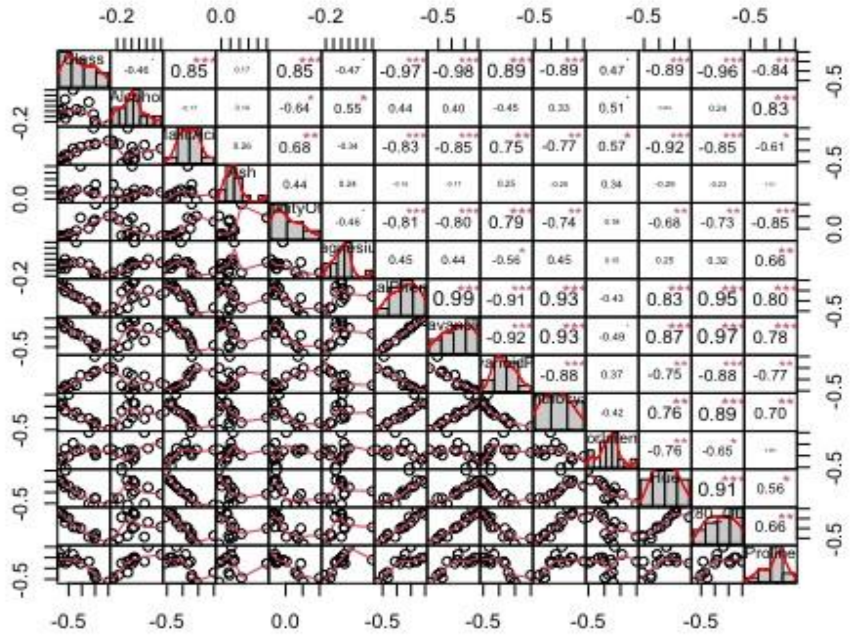
Cluster Analysis on
Wine Data

The data which I used is the Wine Data set. The dataset consists of various measurements related to different chemical properties which are found in a variety types of wine.The data set comprises of 178 observations, each observation in the data set corresponds to an each wine sample. The data set has 13 features which are used to represent the chemical components wine. There are 13 numeric variables which like alcohol, milacid, phenols, and more. There is a categorical variable “class” which indicates the wine class.

The data is taken from the UCI Machine Learning Repository.

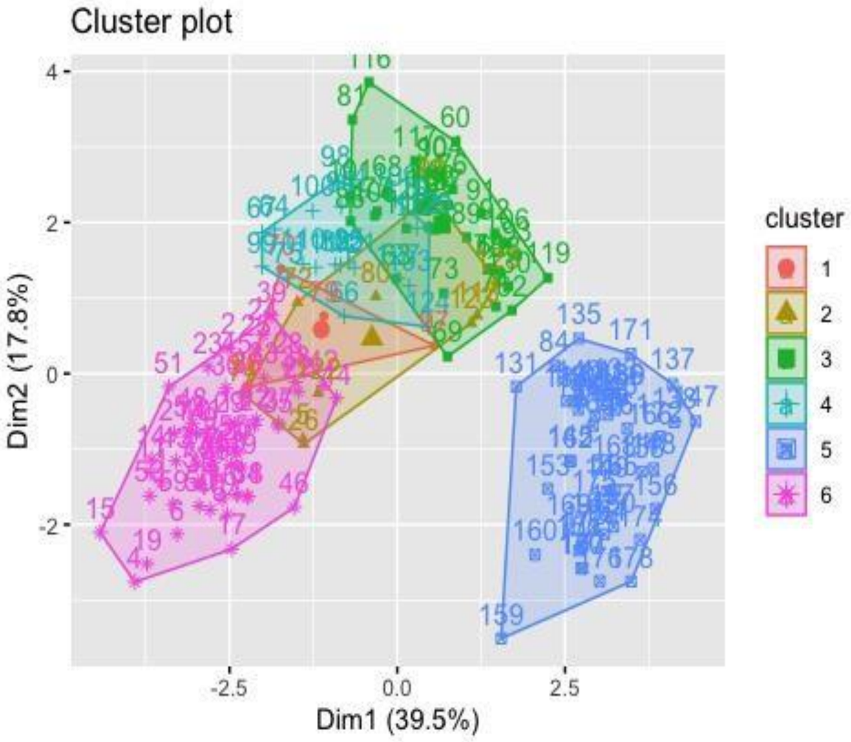
URL - https://archive.ics.uci.edu/dataset/109/wine

Correlation Plot Of Data -



K – means

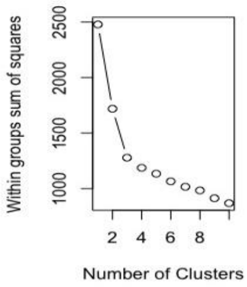
- For the K- means method, we used a Partitional clustering method . First we have assumed and taken the number of clusters (K value) as “6” The cluster plot for the K- means model is shown below in the figure.

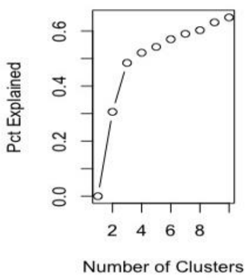


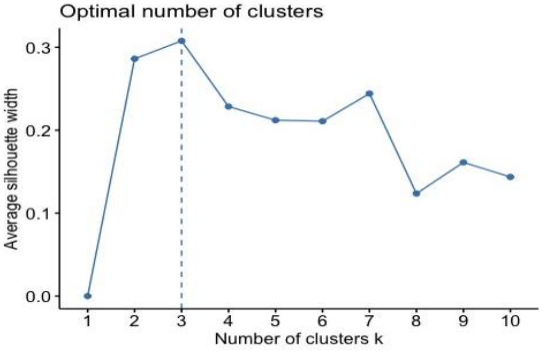
- As we can see from the above plot, we can say that for the number of clusters (K value) as “6” , the clusters are colliding in all directions which does not make k equal to “6” as an ideal value.

- Hence we have opted to find Optimal – k using the Elbow Method.

Determining the optimal k using the elbow method



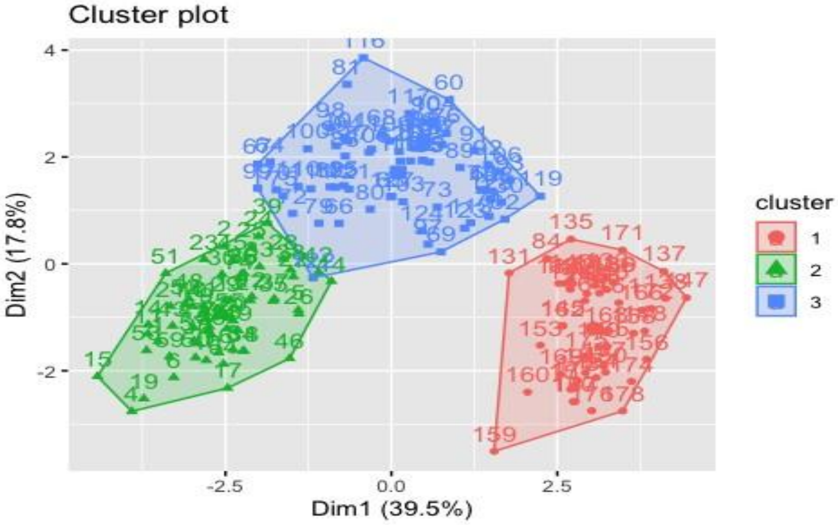




Determining the optimal {k}:

- Using the Elbow plot method and Silhouette method for each k.
- Plotting the silhouette scores and choose the k with the highest average silhouette score.

- Based on the Elbow Plot applied above, we can observe from the plot that there is a clear elbow at k=3, which indicates that there is an optimal number of clusters.
- Based on the Silhouette plot, the highest average silhouette score is observed at k = 3.
- Hence we can conclude that the optimal value of K is “3”

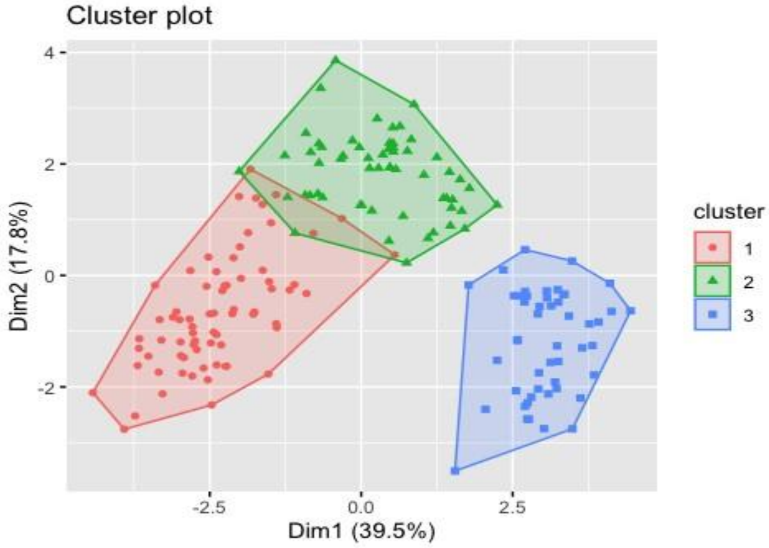


```
table(wine_data$Class, wine_data$cluster)


     1  2  3

1  0 59  0

2  1  2 68

3  48  0  0
```

K- Medoid –

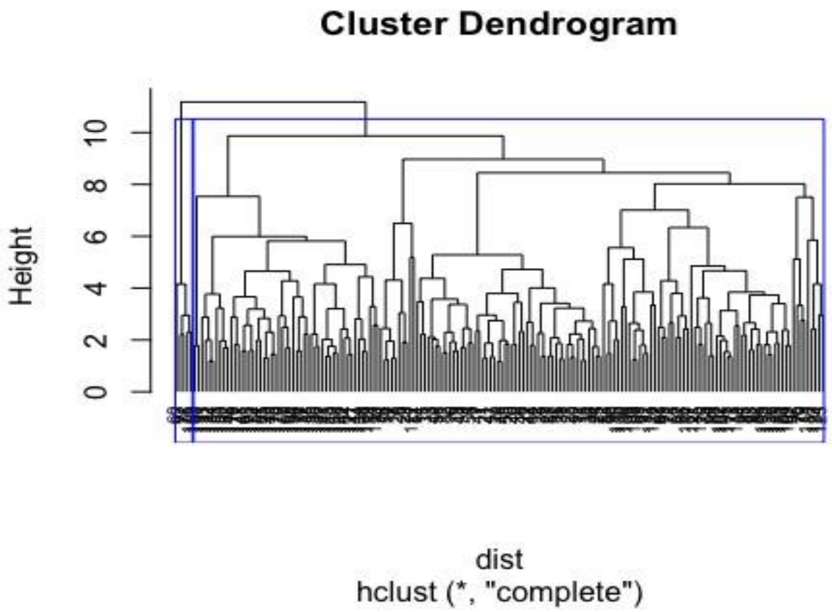
- For k -medoid, we have taken the optimal value of “K” from the k – means method, which is k = 3.
- Used the pam model with distance metric as Euclidean distance.

- Plotted the cluster plot which is shown below.



Hierarchial Clustering -

- First we have taken the distance metric as Euclidean distance
- The best K value is chosen as “3” which is the cuttree at height k=3 as shown in the figure.
- For this, we have used complete linkage method. This method is a bit difficult to analyse and understand.
- Plotted the cluster Dendogram to show the visualization of clustering.



```
table(wineScaled$Class, wineScaled$highclust)

##
##                        1   2   3
##   -1.21052889135896   59   0   0
##   0.0797354359577757 63   5   3 ##
1.36999976327451    0   0 48
```

{Q} This criterion is linked to a Learning OutcomeChoose one set of cluster results (from one method) and provide some interpretation as to the meaning of the clustering.

- For Wine data set, partional clustering k- means is best suited methodology for clustering.
- From the cluster plot, by using k -means method the data is divided based on class into three categories. Since the data is Scaled and Class levels are also in negative
- Here, we can observe that three levels for class, the K- means clustering categorized the 178 observations into cluster 1 , cluster 2, and cluster 3.
- Each cluster has each level of class except three values, which means the levels are clustered into three which is the main aim of clustering.

```
table(wine_data$Class, wine_data$cluster)
```

```
     1  2  3
1  0 59  0
2  1  2 68
3  48  0  0
```