Mohamamdhossein Jafari

mohammadhossein.jafari@studenti.unimi.it

Unsupervised Learning :

Abstract:

This research tries to use clustering tasks on an agricultural dataset. The Hierarchical Clustering and k-means Clustering are used on the dataset and performance of different Linkage methods such as ward, complete and single are tested. In the next step, the research tries to find the best number of clusters by using different methods such as Gap-statistics and Elbow Method.

Goal Of The Analysis And Description Of The Data Set:

Data Set Information:

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. The data set can be used for the tasks of classification and cluster analysis. Attribute Information:

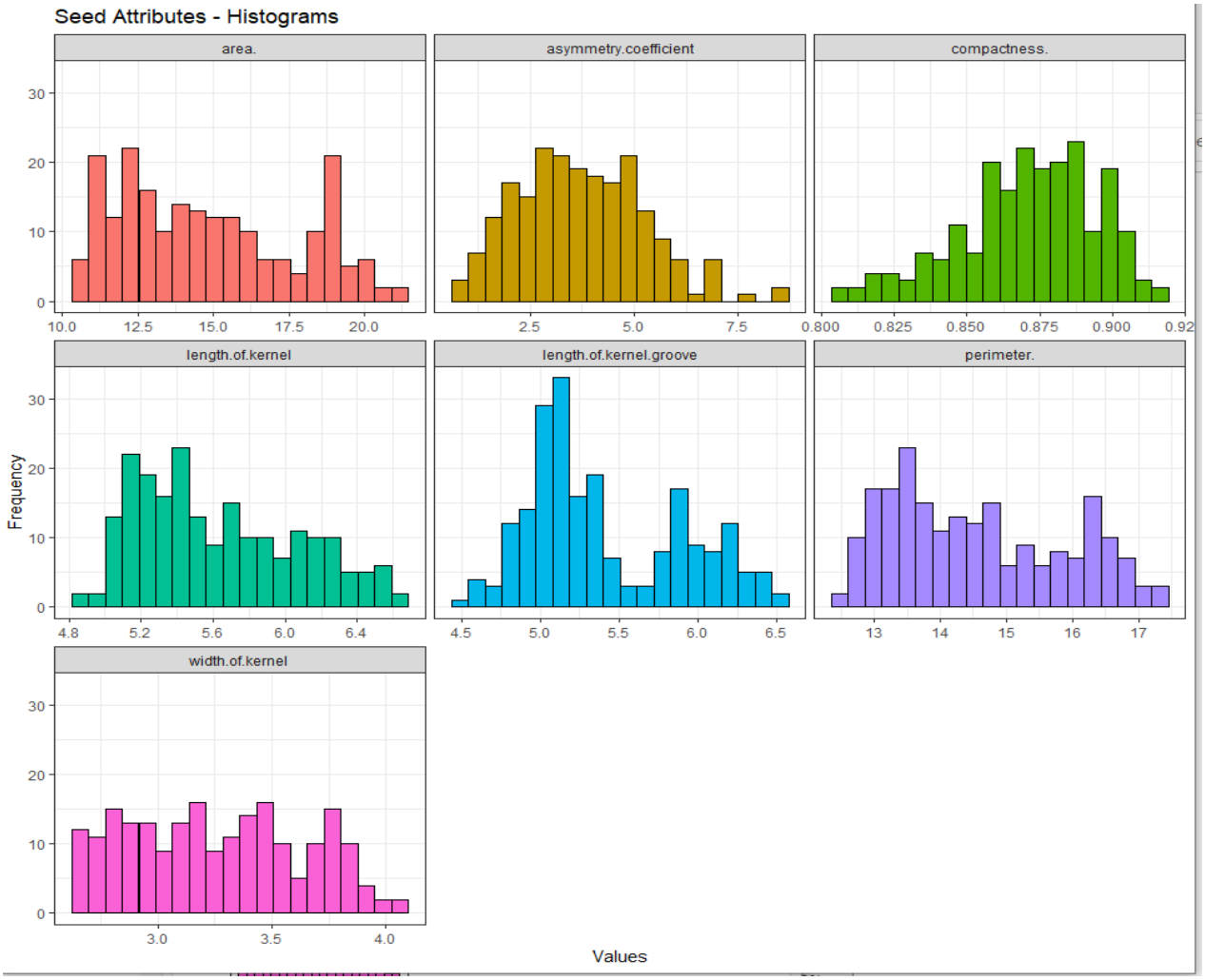
To construct the data, seven geometric parameters of wheat kernels were measured:

- area A
- perimeter P
- compactness C
- length of kernel
- width of kernel

- asymmetry coefficient
- length of kernel groove.

All these parameters were real-valued continuous.

The histograms show the frequencies of attributes at each value. It is clear that the distributions of frequencies are different in each attribute.

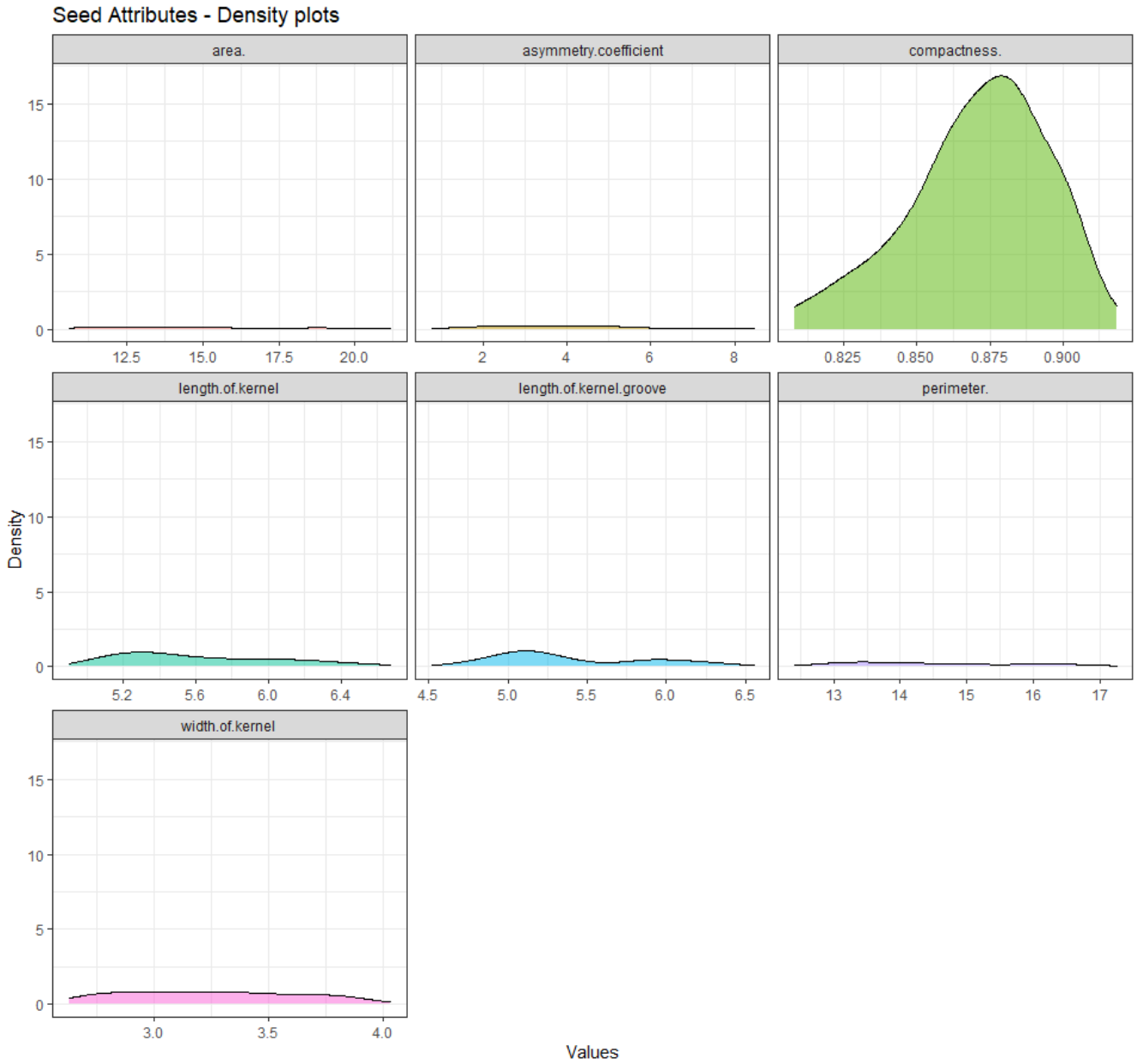


Density plots at different values has been shown in the following tables.

Scaling data:

Feature scaling through standardization can be an important preprocessing step for many machine learning algorithms. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one.

While many algorithms (such as SVM, K-nearest neighbors, and logistic regression) require features to be normalized, intuitively we can think of Principle Component Analysis (PCA) as being a prime example of when normalization is important. In PCA we are interested in the components that maximize the variance. If one component (e.g. human height) varies less than another (e.g. weight) because of their respective scales (meters vs. kilos), PCA might determine that the direction of maximal variance more closely corresponds with the 'weight' axis, if those features are not scaled. As a change in height of one meter can be considered much more important than the change in weight of one kilogram, this is clearly incorrect. A set of scaled data is generated, and we compare the results of the scaled and non-scaled data.



Findings of the Research:

- when doing the Hierarchical Clustering the best linkage method is ward.
- Doing scaling can improve the results of the research sharply
- Almost all methods to find the optimal number of Clusters is three clusters
- The Figure of the first PCA was about 82 percent

The Analysis of Models and Explanations:

Hierarchical Clustering:

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

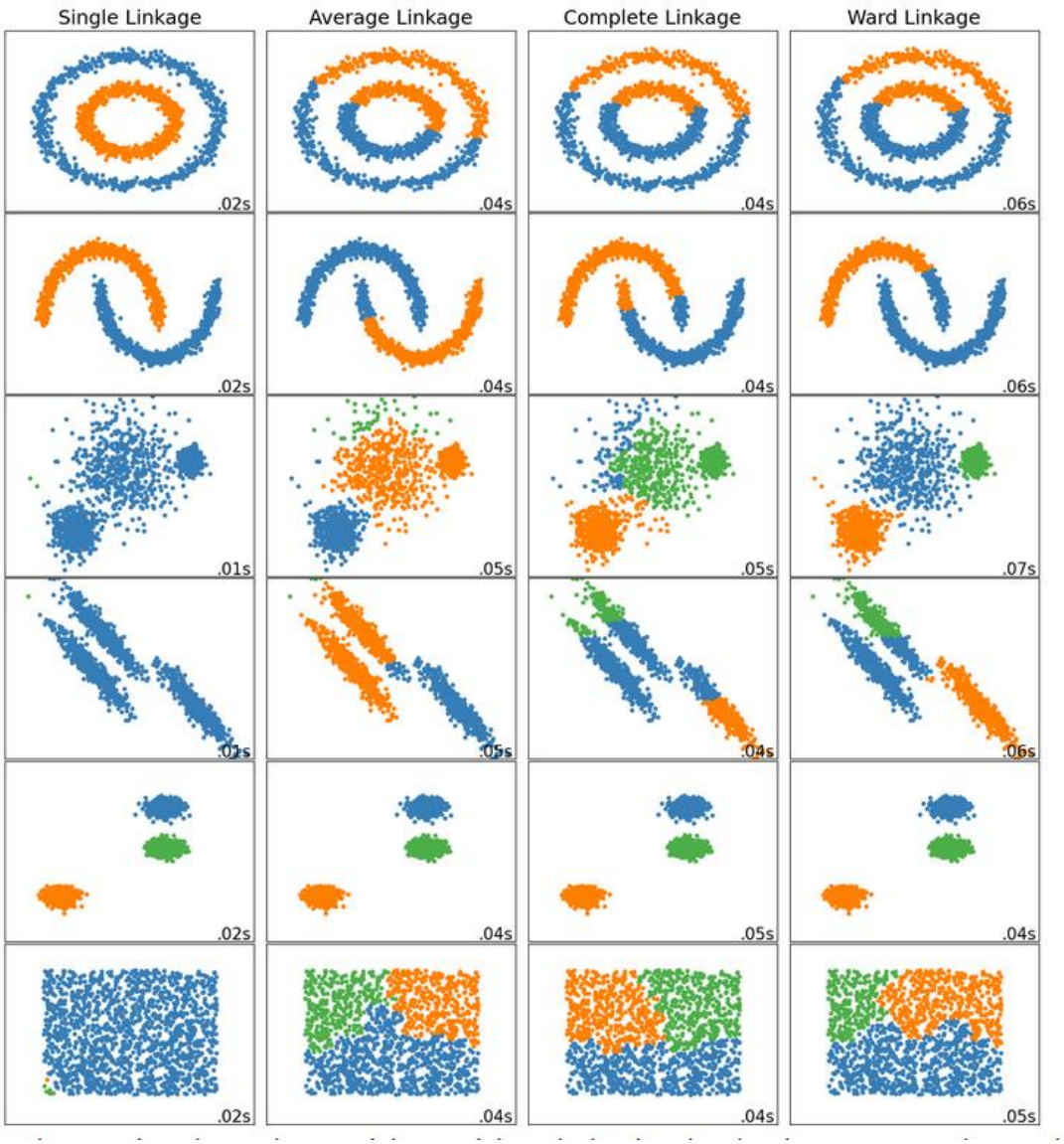
a hierarchical clustering using a bottom-up approach: each observation starts in its own cluster, and clusters are successively merged. The linkage criteria determine the metric used for the merge strategy:

- ❖ Ward minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is like the k-means objective function but tackled with an agglomerative hierarchical approach.
- ❖ Maximum or complete linkage minimizes the maximum distance between observations of pairs of clusters.
- ❖ Average linkage minimizes the average of the distances between all observations of pairs of clusters.
- ❖ Single linkage minimizes the distance between the closest observations of pairs of clusters.
- ❖

Different linkage type: Ward, complete, average, and single linkage:

Agglomerative cluster has a “rich get richer” behavior that leads to uneven cluster sizes. In this regard, single linkage is the worst strategy, and Ward gives the most regular sizes. However, the affinity (or distance used in clustering) cannot be varied with Ward, thus for non-Euclidean metrics, average linkage is a good alternative. Single linkage, while not robust to noisy data, can be computed very efficiently and can therefore be useful to provide hierarchical clustering of larger datasets. Single linkage can also perform well on non-globular data.

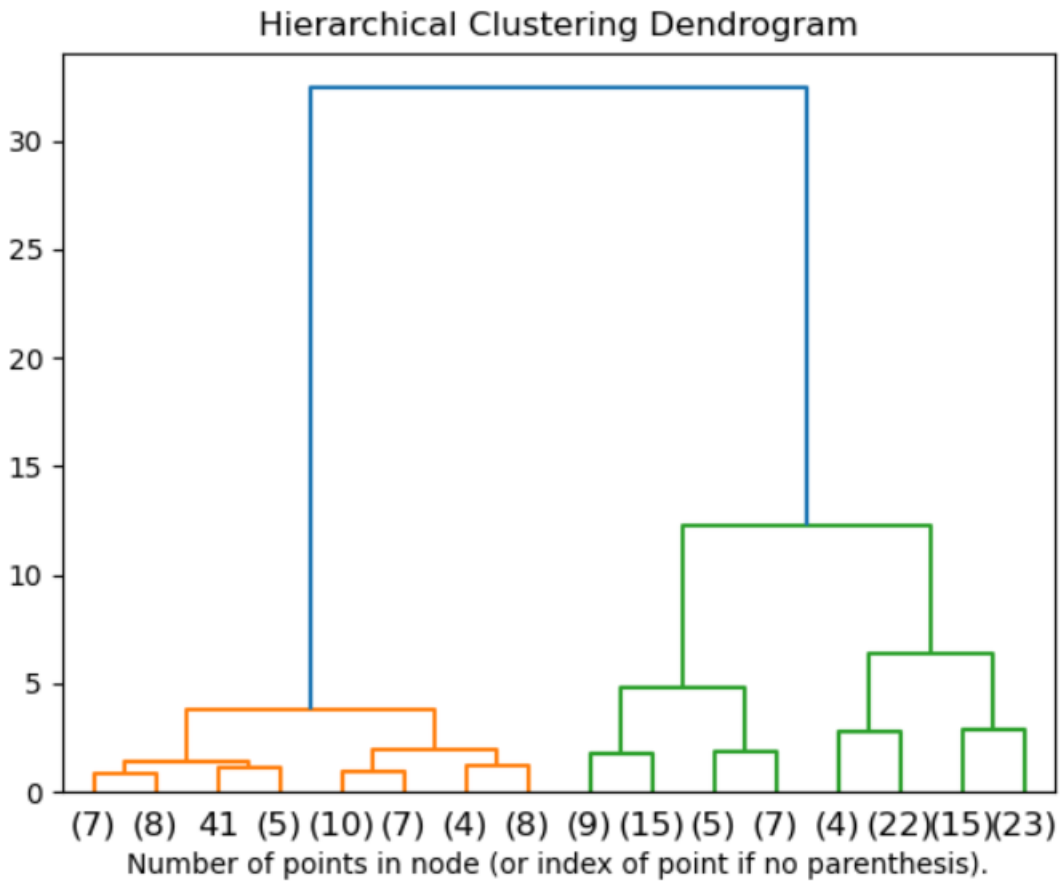
The below picture shows the usage of different linkage methods.



In the following pages we check the performance of different linkage methods on the seed dataset.

Visualization of cluster hierarchy:

It's possible to visualize the tree representing the hierarchical merging of clusters as a dendrogram. Visual inspection can often be useful for understanding the structure of the data, though more so in the case of small sample sizes.



K-means

The K means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. It scales well to large number of samples and has been used across a large range of application areas in many different fields. This algorithm requires the number of clusters to be specified.

The k-means algorithm divides a set of samples into disjoint clusters, each described by the mean of the samples in the cluster. The means are commonly called the cluster “centroids”.

The K-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum-of-squares criterion.

K-means is often referred to as Lloyd’s algorithm. In basic terms, the algorithm has three steps. The first step chooses the initial centroids, with the most basic method being to choose k samples from the dataset X. After initialization, K-means consists of looping between the two other steps. The first step assigns each sample to its

nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids are computed, and the algorithm repeats these last two steps until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly. K-means is equivalent to the expectation-maximization algorithm with a small, all-equal, diagonal covariance matrix.

The algorithm can also be understood through the concept of Voronoi diagrams. First the Voronoi diagram of the points is calculated using the current centroids. Each segment in the Voronoi diagram becomes a separate cluster. Secondly, the centroids are updated to the mean of each segment. The algorithm then repeats this until a stopping criterion is fulfilled. Usually, the algorithm stops when the relative decrease in the objective function between iterations is less than the given tolerance value. This is not the case in this implementation: iteration stops when centroids move less than the tolerance.

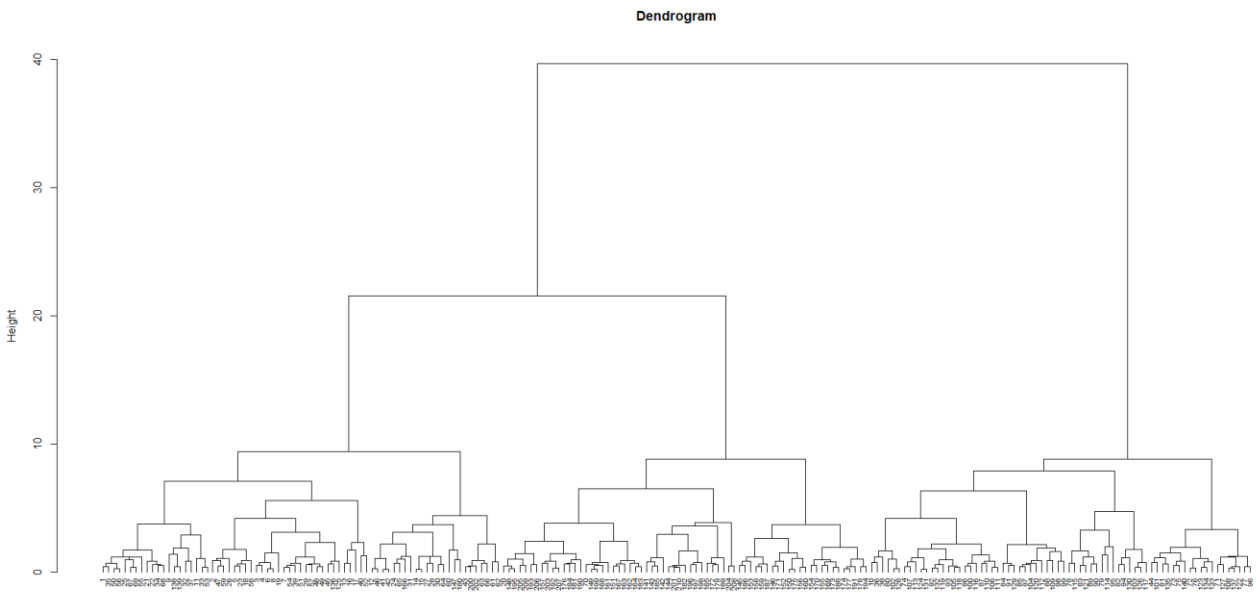
Given enough time, K-means will always converge, however this may be to a local minimum. This is highly dependent on the initialization of the centroids. As a result, the computation is often done several times, with different initializations of the centroids.

K-means can be used for vector quantization. This is achieved using the transform method of a trained model of KMeans.

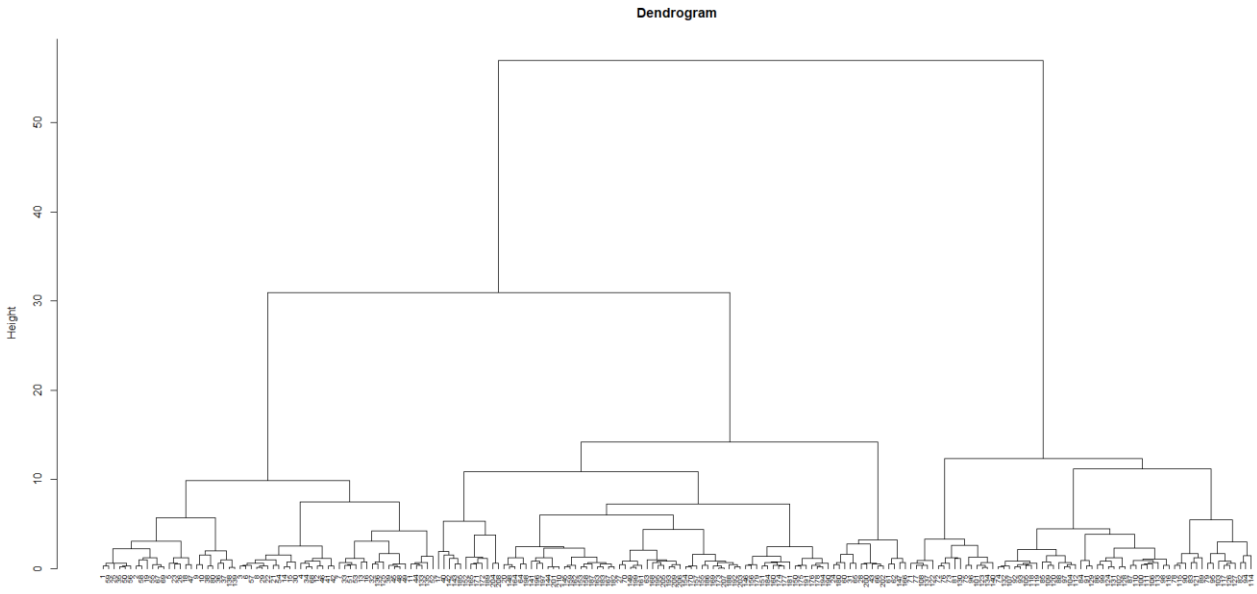
Hierarchical **Clustering on seed Dataset:**

A function is defined to check the performance of each linkage method on this data set. By using the original dataset, the ward method gets 99 percent, and the complete method gets 96 percent and the average method gets 93 percent. The worst performance is for single method which only gets approximately 71 percent. When we use the scaled dataset, the ward method again has the first position with having 98 while the worst position is for single method with only 60 percent. The interesting result here is that the feature scaling helps us to choose better the best linkage method because it increases the distance among the ward method, as the best method, and the other methods.

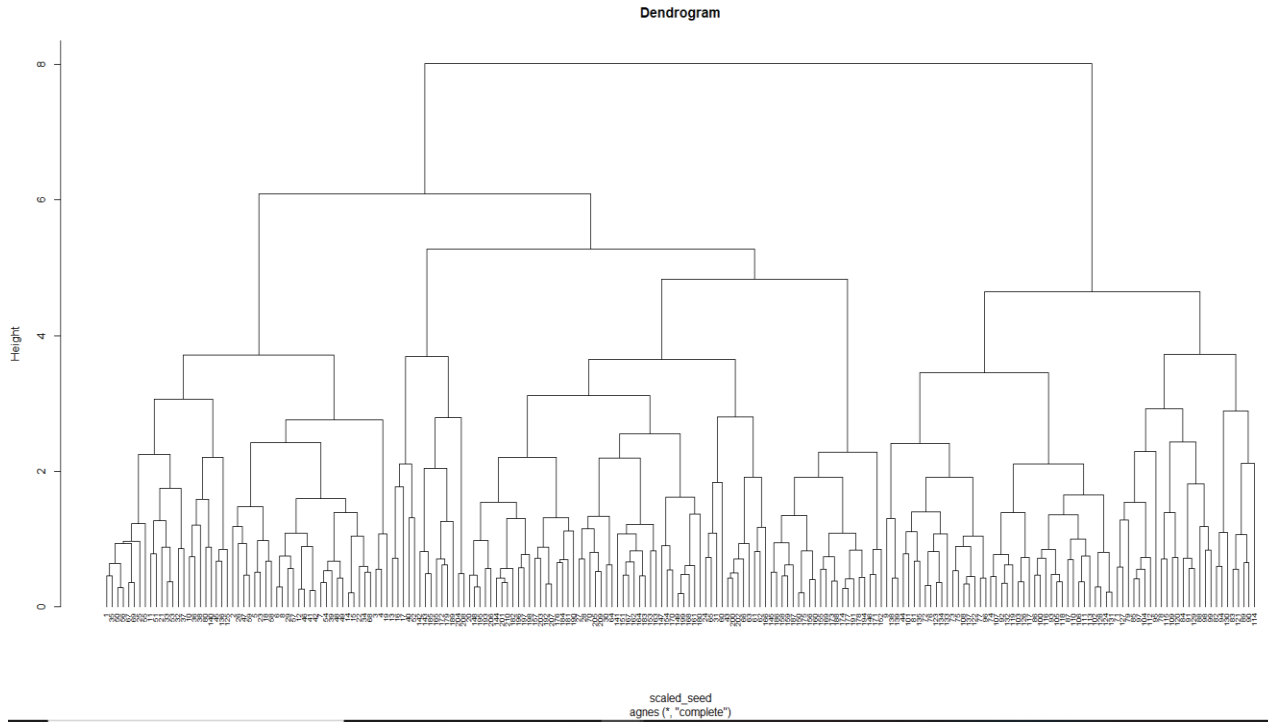
In the below picture, you can see the dendrogram of the scaled seed dataset.

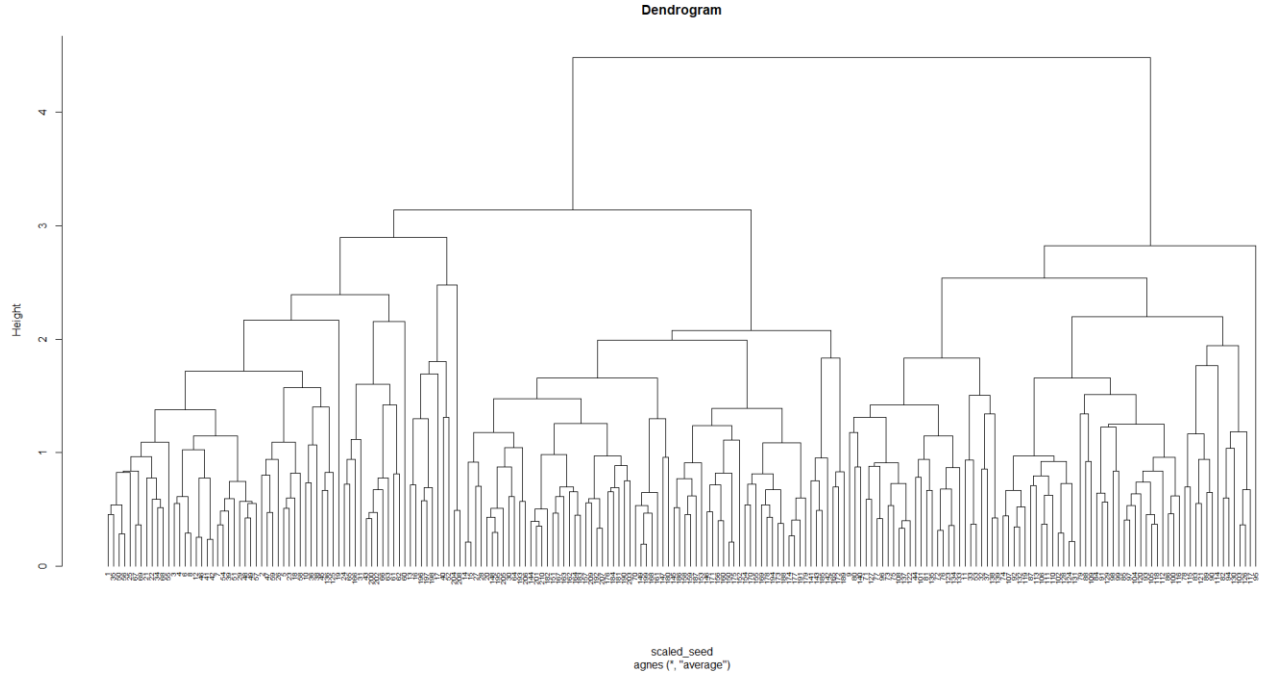


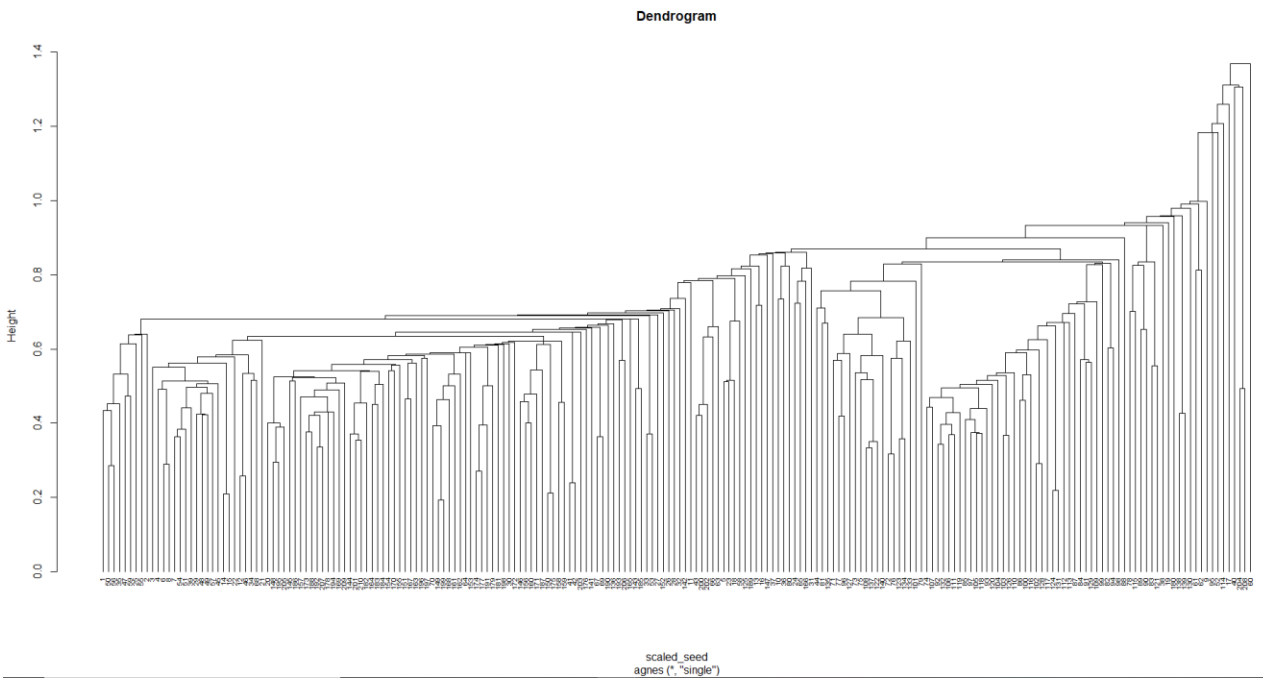
In the below picture, you can see the dendrogram of the original seed dataset.



The chosen linkage method can influence a lot the consequences of our clustering. The dendrogram for the scaled dataset by the complete linkage method has been showen in the below picture. It is obvious the height in the complete linkage method is much less than the ward method.


Dendrogram
Height
0
2
4
6
8
scaled_seed
agnes (*, "complete")

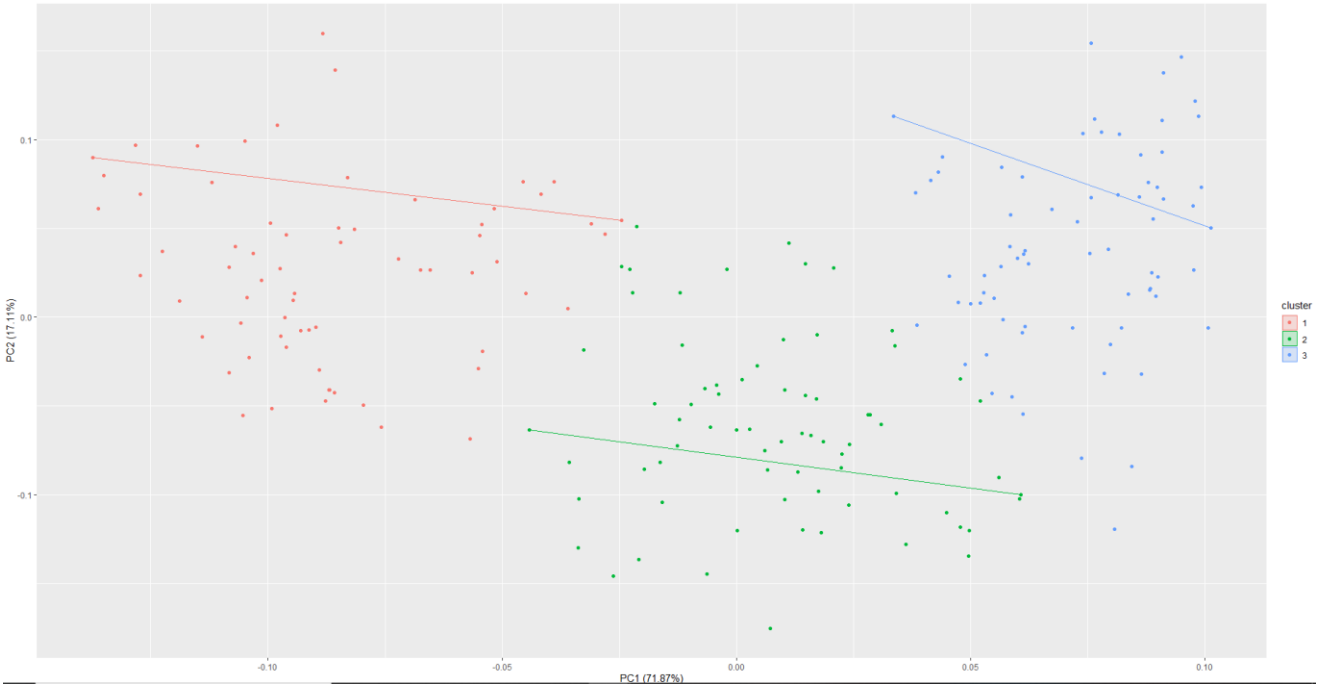

Dendrogram
Height
0
1
2
3
4
scaled_seed
agnes (*, "average")



K-Means Clustering on seed Dataset:

We choose three clusters as defaults and compares the results of this method on both scaled and original data.

the scaled ones:



Choosing the optimal number of clusters:

Hierarchical Clustering and the Optimal Number of Clusters:

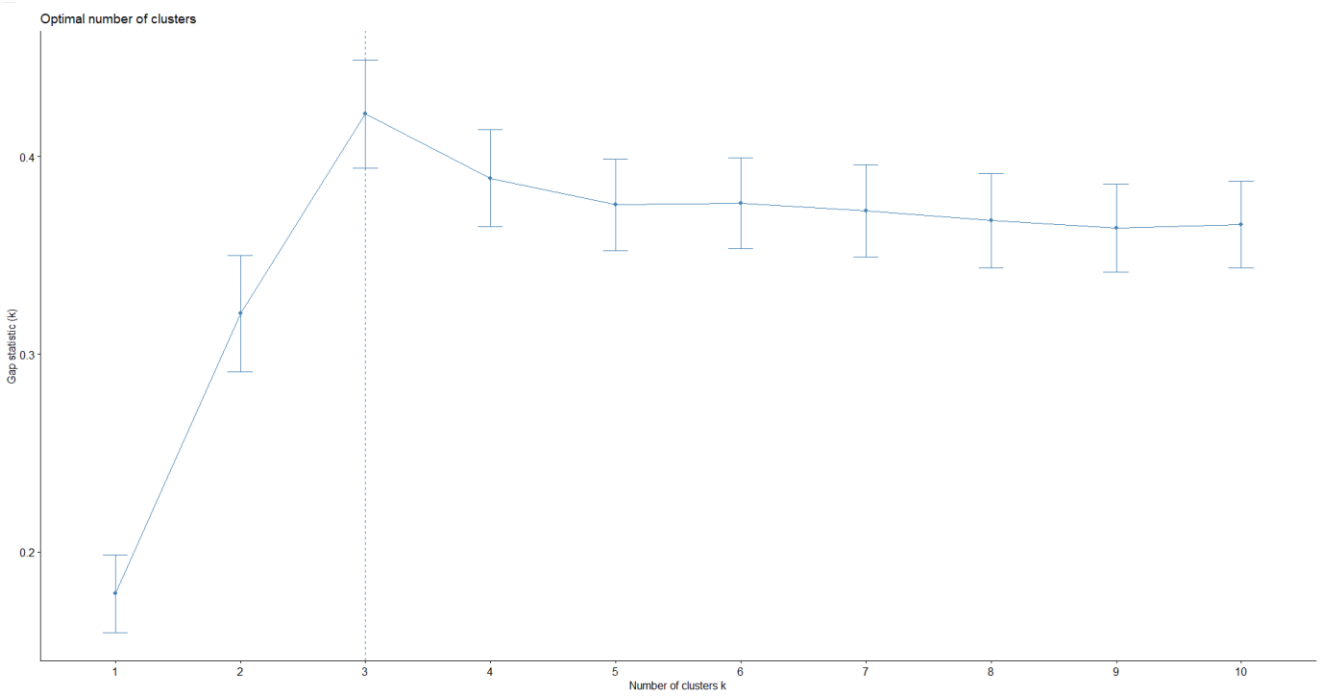
The Gap-statistics method:

The gap statistic has been published by R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001). The approach can be applied to any clustering method.

The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points.

According to Gap-statistics the larger the figure is the better is the chosen number of clusters. The Gap-Statistics chooses 3 as the best number of clusters.

If the scaled data are used there will be 73 members on cluster 1 and 70 and 67 members in cluster 2 and 3 respectively while when the original data is used, not the scaled one, we can see the first cluster has 61 members and the second and the third ones have 86 and 63 members respectively.

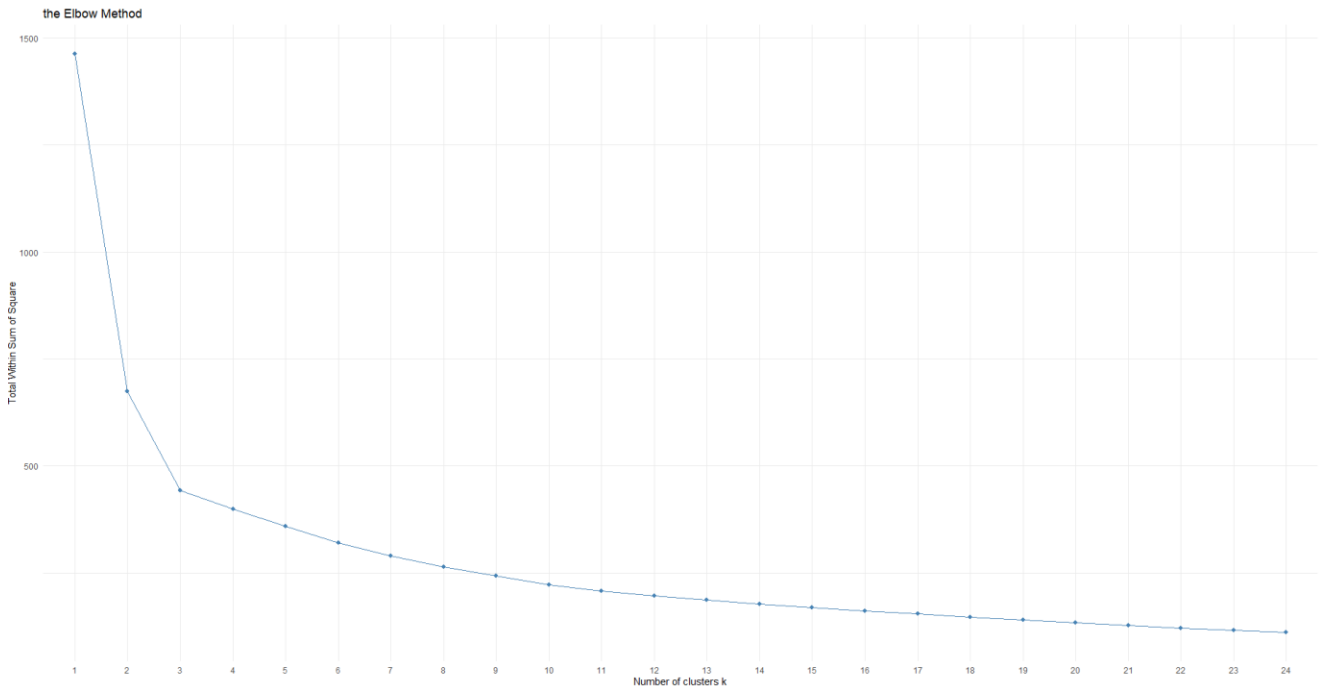


Elbow Method:

Recall that, the basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible.

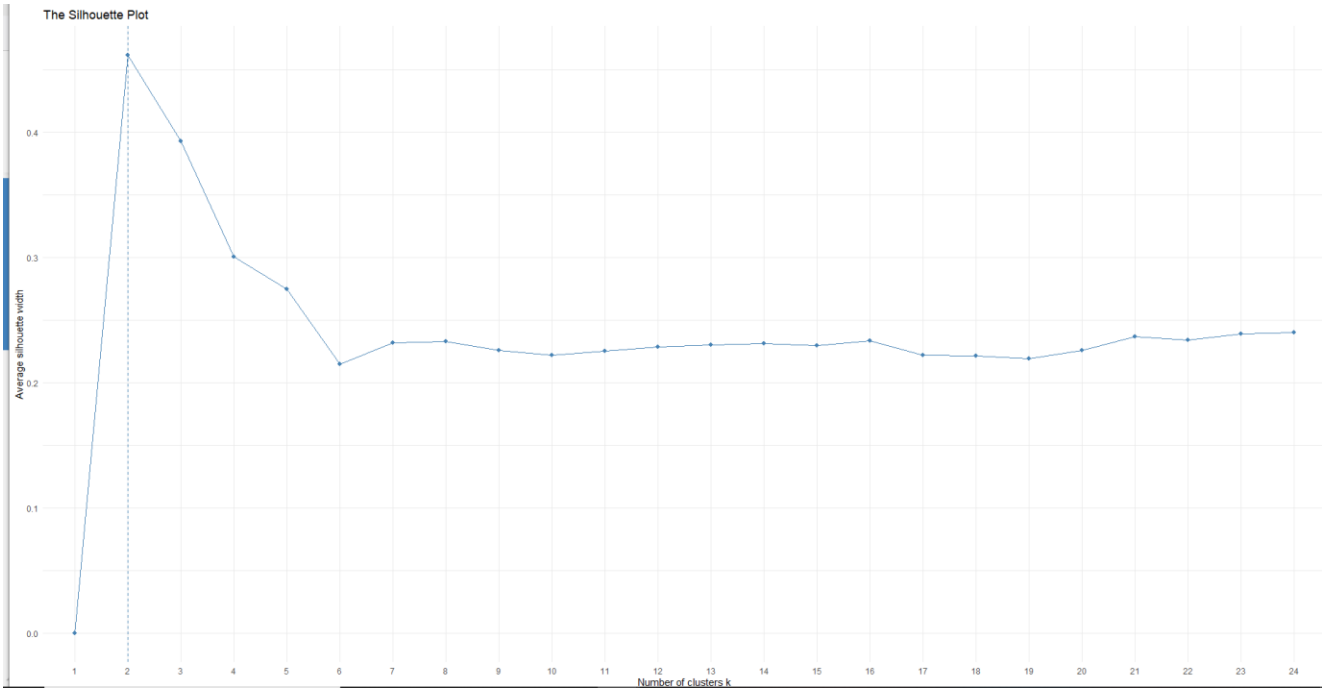
The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't improve much better the total WSS.Is it possible to use elbow method to choose the best number of clusters by considering different algorithm such as Hierarchical Clustering or K-means Clustering.

According to elbow method by considering the Hierarchical Clustering the best number of clusters are three.



Silhouette Method:

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually.

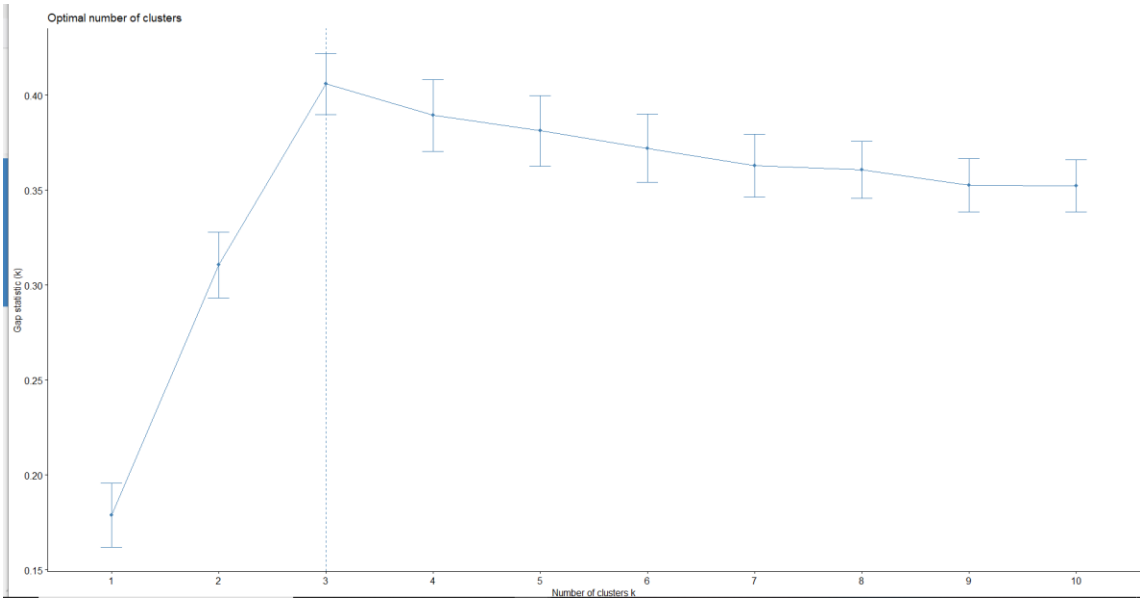


By the above method it chooses just 2 clusters.

K-MEANS and the Optimal Number of Clusters:

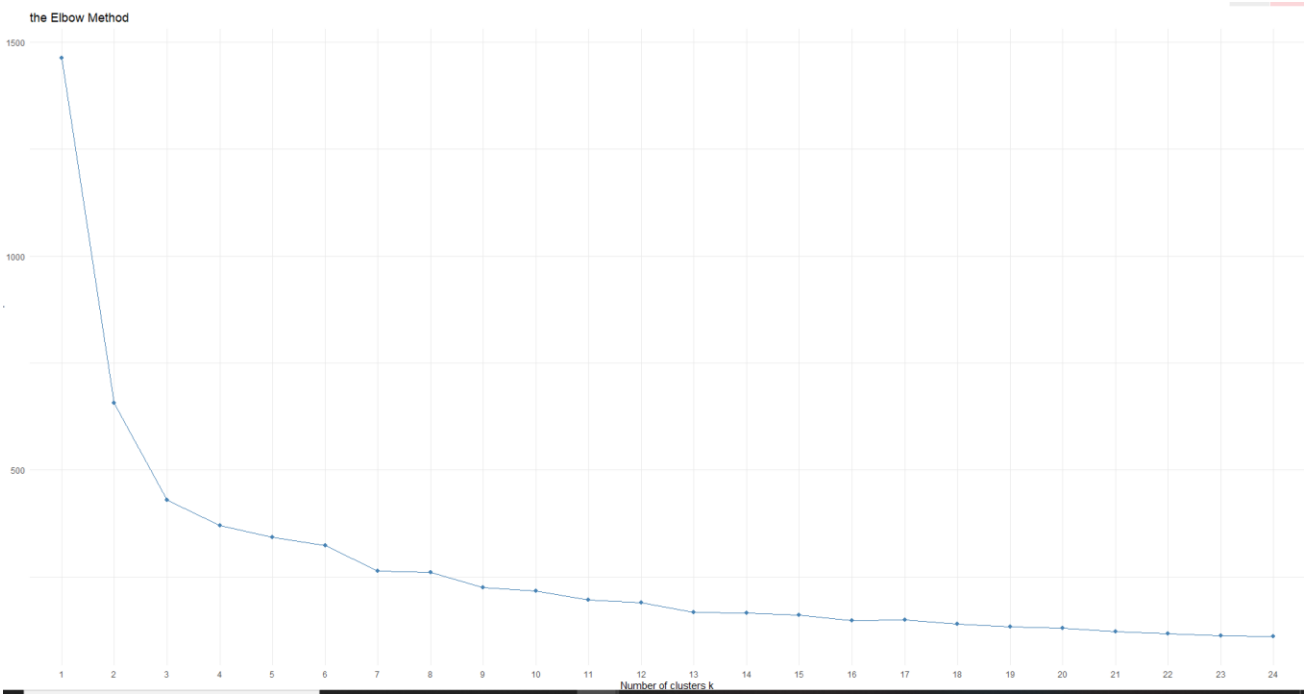
:

GAP-Statistics:



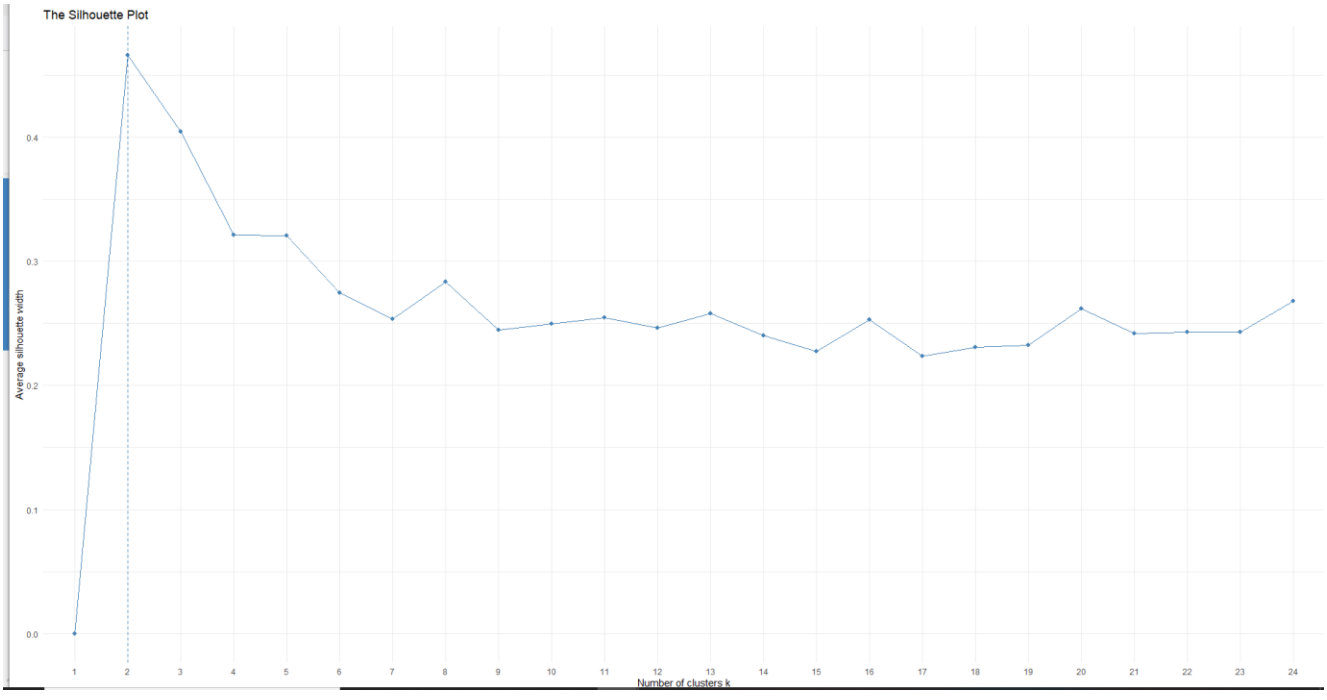
Elbow method

According to **elbow method** by considering the K-means method, the best number of clusters are again three.



Silhouette:

According to Silhouette method for k-means, the bst number of clusters are two.



Conclusion:

It is important to summarize the following findings in conclusion:

- Both clustering families agree that the best number of clusters will be 3.
- Among all linkage methods the ward method can produce more accurate results
- Feature scaling can have an important influence in producing accurate results
- The single linkage method always produces the worst result.
- The sum of the first PCA is about 82 percent and the second one is 16 percent
- The methods of choosing the best number of clusters can sometimes have difference but they almost produced the same results in this dataset.
- It is important to search about other methods of clustering on this dataset.