**Interpretation of cluster analysis methods**

We have a large data set horizontally but not vertically, therefore we are in the problem of curse of dimensionality. So, to understand the interpretation methods let’s dive for a small group of data and understand how the cluster analysis methods help us to identify the clusters in the data. Clusters are likely related data to each other with some factors.

**Kmeans Clustering**: In this method we need to specify the clusters beforehand.

Process   
STEP1: Choose the number K of clusters

STEP2: Select at random K points, the centroids (not necessarily from your dataset)

STEP3: Assign each data point to the closest centroid => that forms K clusters

STEP4: Compute and place the new centroid of each other

STEP5: Reassign each data point to the new closest centroid if any reassignment took place, go to STEP4, otherwise go to FINISH

Bad random initialization can lead to the wrong convergence of the algorithm. This can be solved by kmeans++ algo.

But we do not know how before how many clusters will suit the data to distinguish properly. So, we use the elbow method to find out the optimal number of clusters. In this method, there is a term WCSS (Within cluster sum of square) which helps to find out the number of clusters. Initially if we randomly select cluster number then kmeans finds out that k clusters and then calculate the sum of squares within the cluster. So that we need to try out for different number of clusters.

WCSS = ∑ ∑ distance (Pi, Cj) ^ 2

Cj Pi in cluster Cj

So, at the end we can form any number of clusters (max to number of data points in dataset). If you select the number of clusters equal to number of data points then WCSS will be zero, which is a good metric but at last it will end up to zero. So, this way needs to plot all the WCSSi points against i cluster and find out from where curve starts to bend like a knee or an elbow. That is why it is called an elbow method. It is usually vague. Because sometimes it is hard to interpret the bending point, for different observer point can be different.

**Hierarchical Clustering:**

Process

STEP1: Make each data point a single-point cluster => That forms N clusters

STEP2: Take the two closest data points and make them one cluster => That forms N-1 clusters

STEP3: Take the two closest clusters and make them one cluster => That forms N-1 clusters

STEP4: Repeat STEP3 until there is only one cluster

Distance between clusters: Single linkage (closest points), Complete linkage (Furthest points), Average linkage (Average Distance), Distance between centroids, Ward distance (ANOVA based approach)

Dendrogram is used in this method to visualize the data. On X axis, the points have been plotted and distance between points is been plot on Y axis. So, points are connected by horizontal line at the height of distance. In this way, the tree has been formed up. The further away two points or clusters are, more dissimilar they are.

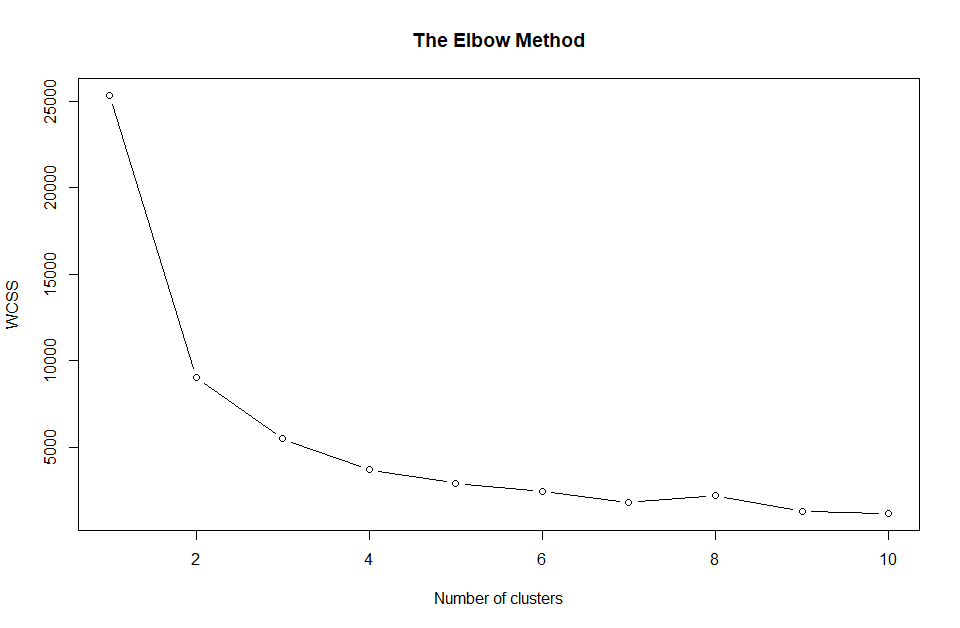
How to find max clusters using dendrograms?

1. Set the threshold and cut through the vertical lines so the lower parts will give us the clusters.
2. Find the longest vertical line which do not cross any horizontal line including hypothetical horizontal line means extend all the horizontal lines and interpret the merge point. Then find other vertical lines parallel to the selected vertical line.

We can also call vertical line as height or distance of dissimilarity between points.

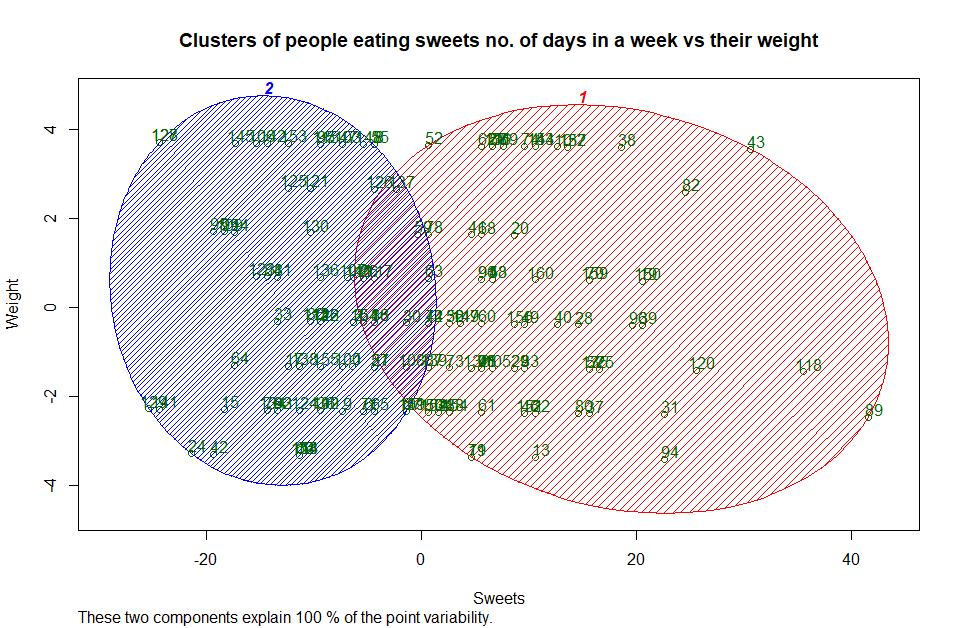
**For group 1**: Sweets intake for number of days in a week vs weight of a person

So here is the WCSS graph.



As I said initially it is very difficult to detect the point where it starts bending. Points after 2, all seems linear with little change, so I would choose 2 as a number of clusters.

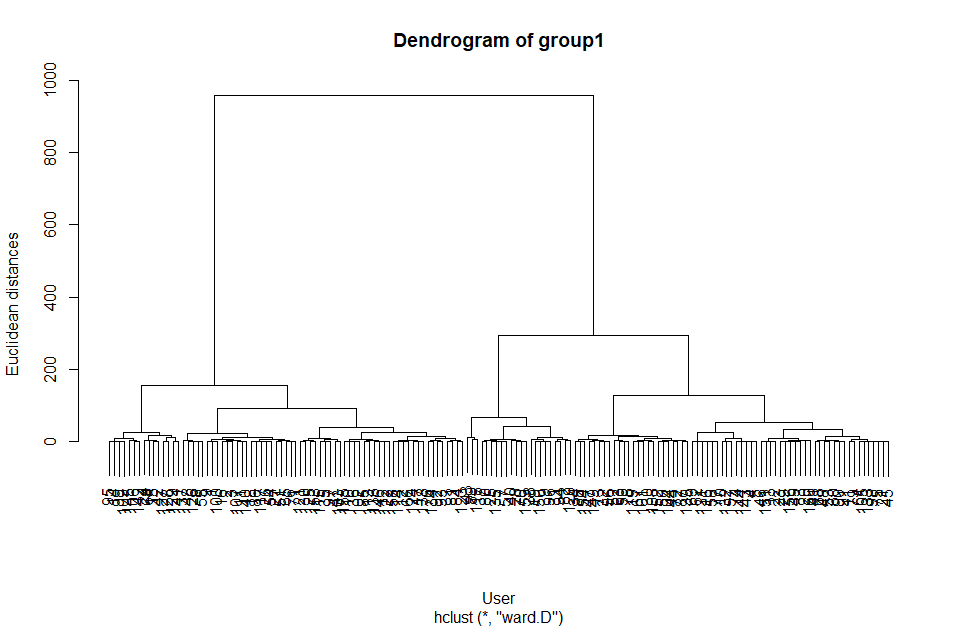
Now let’s try to visualize these 2 clusters using kmeans



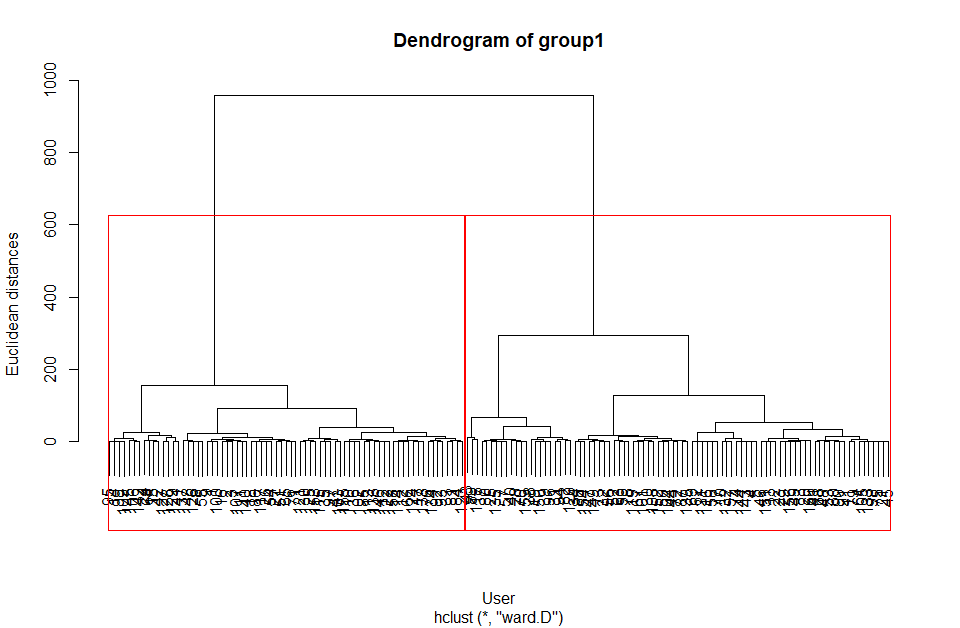
Here we can see as per weight there is not a significant difference. But clusters can be formed as per eating habits of sweets as low and high.

Hierarchical clustering for group1:

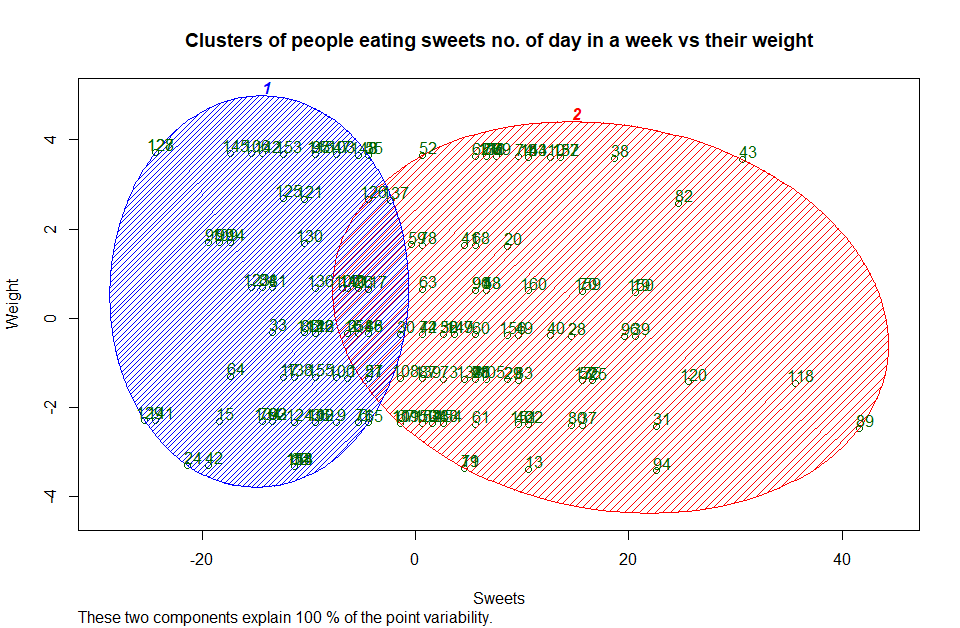
As the process explained above of dendrograms, the longest vertical line shows in the image below.



According to this let’s try to divide the dendrograms into two groups.



Now let’s visualize the clusters as we did in kmean,



In the similar way, it shows two clusters where it is clearly identified that weight does not provide significance contribution in separation of clusters, but sweets day count in a week does.