Report - Dietary Habits using Cluster Analysis

Multivariate Data Analysis (MVDA)

M.Sc. High Integrity Systems

Guidance:

Prof. Dr. Christina Andersson

By:

Gaurav Kapadiya (1319237)

Kshitij Yelpale (1322509)

Parag Tambalkar (1322596)

Safir Mohammad Shaikh (1322554)

Shubham Girdhar (1323003)

Abstract

Multivariate Data Analysis (MVDA) is a statistical technique to analyse and visualise data in order to check how multiple variables behave in combination. Specifically, Cluster Analysis, an unsupervised learning technique, is used to explore and divide data into multiple groups based on their characteristics. In this report, we have illustrated the same with a real-world example related to Dietary Habits. The analysis is performed to identify dietary habits of people belonging to different regions of the world. People with similar dietary habits are placed into one cluster and vice versa. The analysis is conducted using hierarchical clustering and k-means clustering techniques. The difficulty of examining every possible partition creates the need of finding algorithms which look for the minimum values of the clustering measures by rearranging existing allotments and keeping the new one only if it provides an improvement. The experiment involves data collection, data pre-processing and clustering of data using different algorithms which result into the formation of clusters of similar input data points.

1. Introduction

Cluster Analysis is used to group similar objects together in groups called clusters. The task of clustering includes dividing the data points into number of groups in such a way that similar data points are placed in the same group and dissimilar data points are placed in different groups. The goal of our experiment is to obtain groupings of similar data points based on dietary habits of participants from various regions of the world. We have analysed the cluster formation for our collected data using hierarchical clustering and k-means clustering techniques.

1.1 Hierarchical Clustering Techniques [1]

As discussed in previous section, aim of cluster analysis is to form clusters of similar data points. To implement cluster analysis using hierarchical clustering, distance matrix or raw data is used. Distance between observations is needed to calculate distance matrix. The Euclidean distance is generally preferred to calculate distance between two data points. When raw data is provided to some statistical software (for example RStudio), it automatically computes a distance matrix. The distance between multiple input data points is showed by the distance matrix.

There are two types of hierarchical clustering depending on the cluster formation approaches:

1. Divisive Clustering (Top-down approach)

In this technique, initially all the observations are considered into one cluster, and then successively splitting these clusters.

2. Agglomerative Clustering (Bottom-up approach)

At the beginning of hierarchical clustering, we treat each observation as a separate cluster and iterate below steps until all the clusters are merged together.

Process:

STEP 1: Make each data point a single-point cluster => That forms N clusters

STEP 2: Take the two closest data points and make them one cluster => That forms N-1 clusters

STEP 3: Take the two closest clusters and make them one cluster => That forms N-2 clusters

STEP 4: Repeat STEP3 until there is only one cluster

The output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters.

1.2 K Means Clustering [2]

K means clustering is an iterative algorithm that groups the input datapoints into k pre-defined distinct clusters where each data point belongs to only one group. This algorithm helps to group most similar datapoints into one cluster and it also tries to keep the formed clusters separated from each other at a maximum possible distance. The clusters are formed in such a way that the arithmetic mean of distances of all the data points belonging to that cluster form centroid of the cluster. The intra cluster variation should be less in order to get more homogeneous datapoints in one cluster. The algorithm measures the sum squared distances of datapoints from all cluster centroids. Datapoints are placed into respective cluster depending on the minimum distance of datapoint from the centroid.

Process:

STEP 1: Choose the number K of clusters

STEP 2: Select at random K points, the centroids (not necessarily from your dataset)

STEP 3: Assign each data point to the closest centroid => that forms K clusters

STEP 4: Compute and place the new centroid of each other

STEP 5: Reassign each data point to the new closest centroid if any reassignment took place, go to STEP4, otherwise go to FINISH

2. Implementation and Results

2.1 Data Collection

To study cluster analysis of dietary habits, we collected data through a survey using google form. The survey form was designed in such a way that it covered general but important questions related to eating habits of individuals. We also asked participants to enter some personal and demographic details. Total 160 participants from different parts of the world provided feedback. The feedback data received from participants was then processed to conduct the cluster analysis.

2.2 Data Pre-Processing

After getting raw data from participants, first and obvious step of our implementation was to pre-process the data. We have a large data set horizontally but not vertically, therefore we face the problem of curse of dimensionality. We removed the variables such as Timestamp which were not related to our analysis. We found in the Gender variable that we received only one response as 'Prefer not to say'. We merged less frequent responses (less than 3) like this to another response containing less inputs. In this case we considered the response 'Prefer not to say' as 'Female'. Similarly, we found out and merged other frequent responses for respective variables. In the raw data variable names were appearing as questions so we renamed the variables to the relevant names. The data contained multiple categorical variables and blank inputs. We converted categorical variables such as Gender, Region, Source of Main Meal to integer by treating every category as an individual dummy variable. For example, in response to Region we received multiple responses from Europe, Asia and Other regions. We created two dummy variables Region 1 and Region 2 as below:

Original Region	**Region 1**	**Region 2**
Asia	1	0
Europe	0	1
Other	0	0

After data pre-processing we actually proceeded for cluster analysis of the data. At the very beginning we implemented cluster analysis on complete dataset using both Hierarchical and K-means method. Moreover, we tried to find relationship between significant variables and created groups of input variables for example one group consisted of sweet and weight. We will discuss both approaches in detail in upcoming sections.

2.3 Hierarchical Cluster Analysis

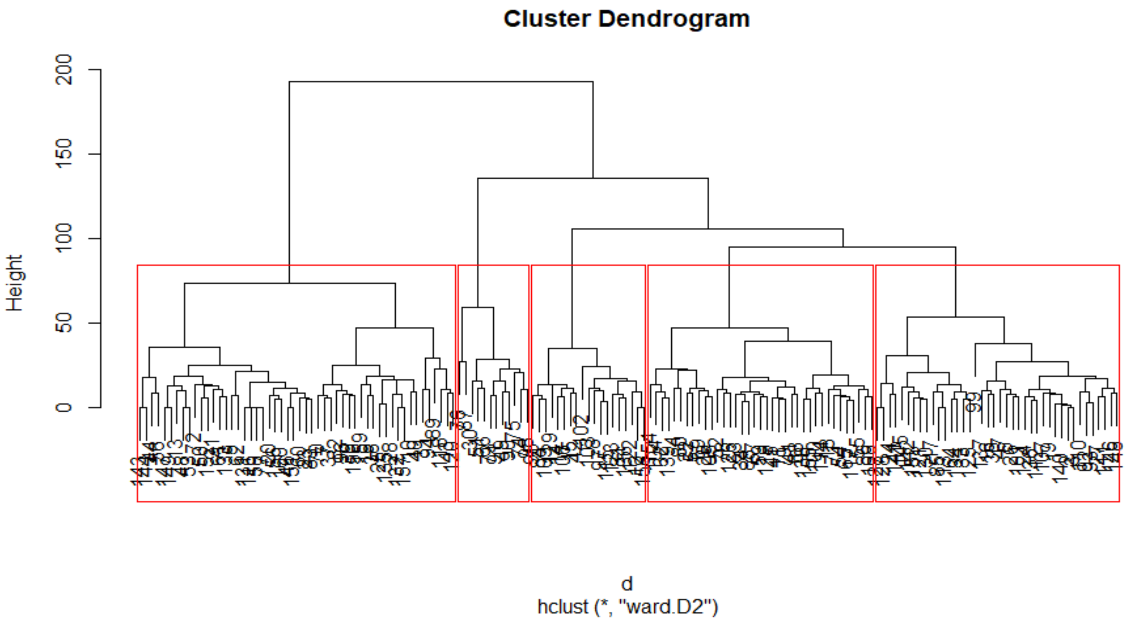
Dendrogram is used in this method to visualize the data. On X axis, the points have been plotted and distance between points is plotted on Y axis. So, points are connected by horizontal line at the height of distance. In this way, the tree is formed. The further away two points or clusters are, more dissimilar they are. To find optimal clusters using dendrograms we need to take below two steps into consideration,

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

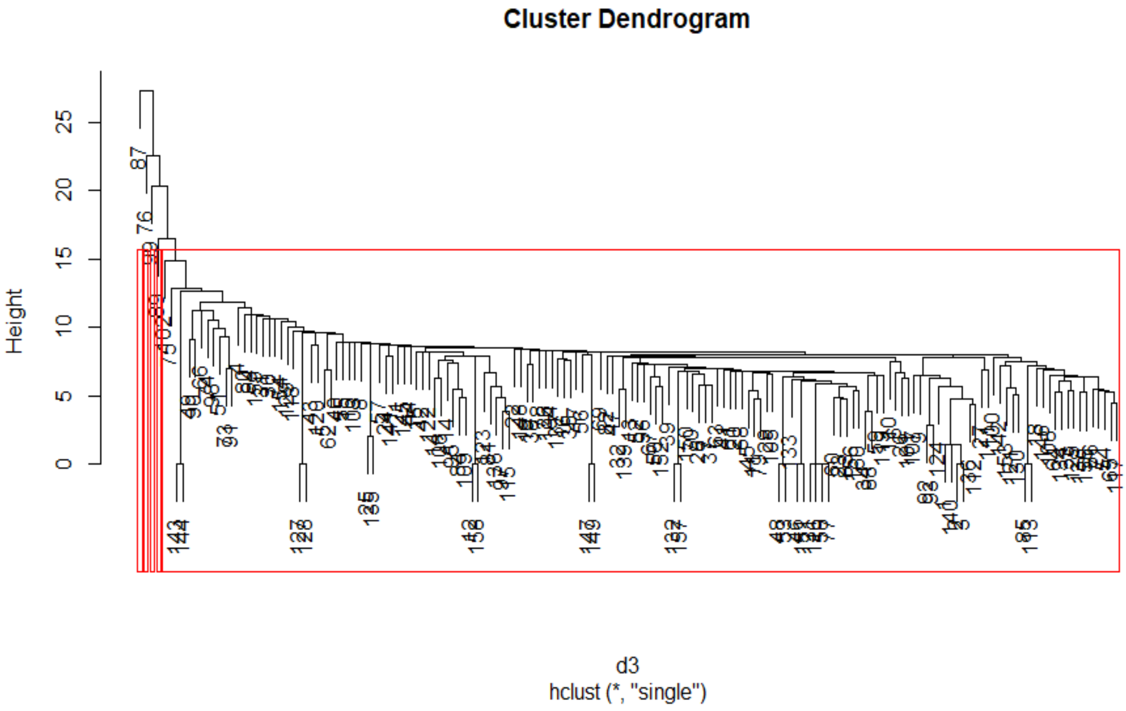
We can also call vertical line as height or distance of dissimilarity between points. We implemented several hierarchical cluster analysis algorithms in this project for which the output dendrogram for complete dataset and group one is shown in below sections:

1. For Complete Dataset
    a. Ward’s Method (ANOVA based approach)

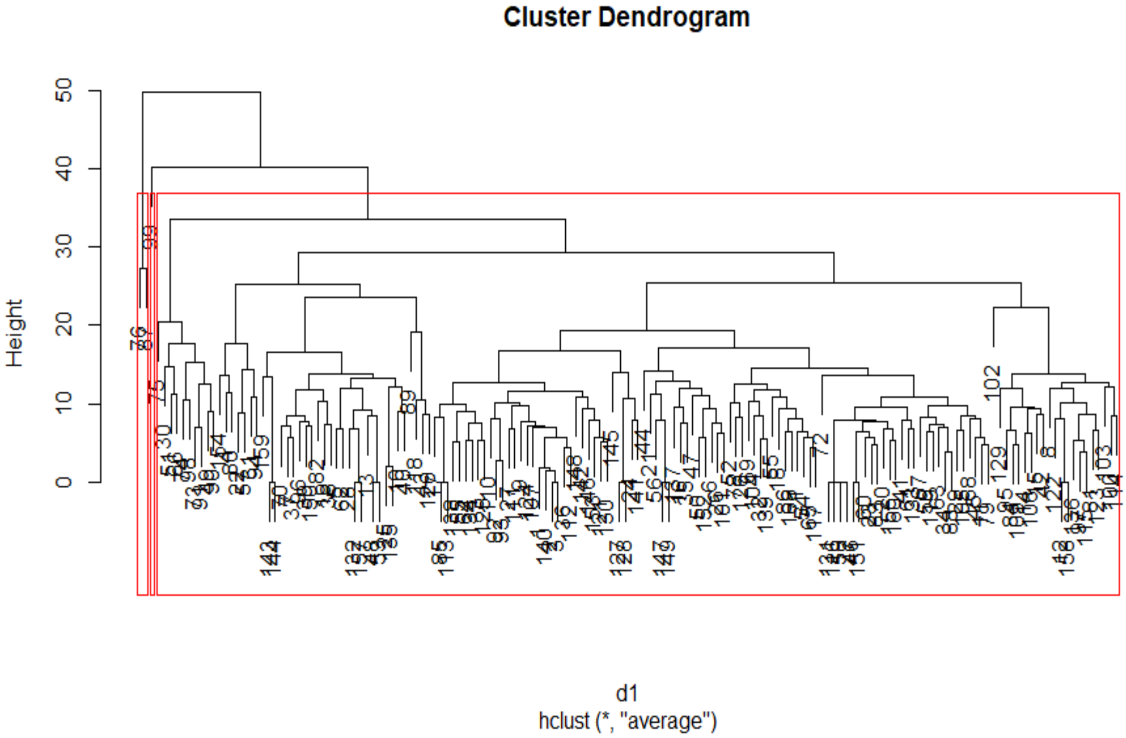
We have used the latest ‘Ward.D2’ as a clustering method which implements Ward’s clustering criterion and squares the dissimilarities before cluster updating.



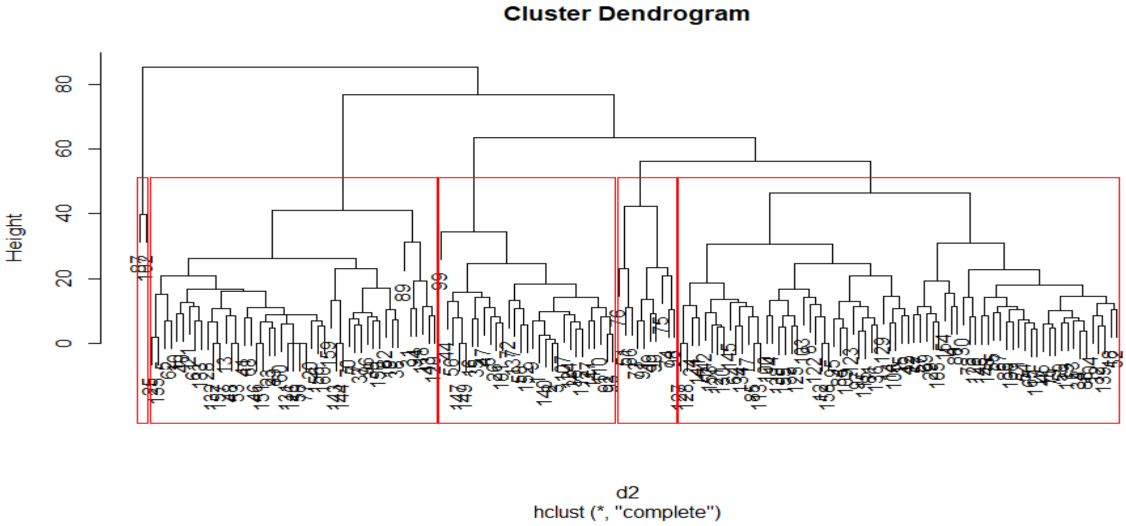
b. Single Linkage Clustering (Closest points)



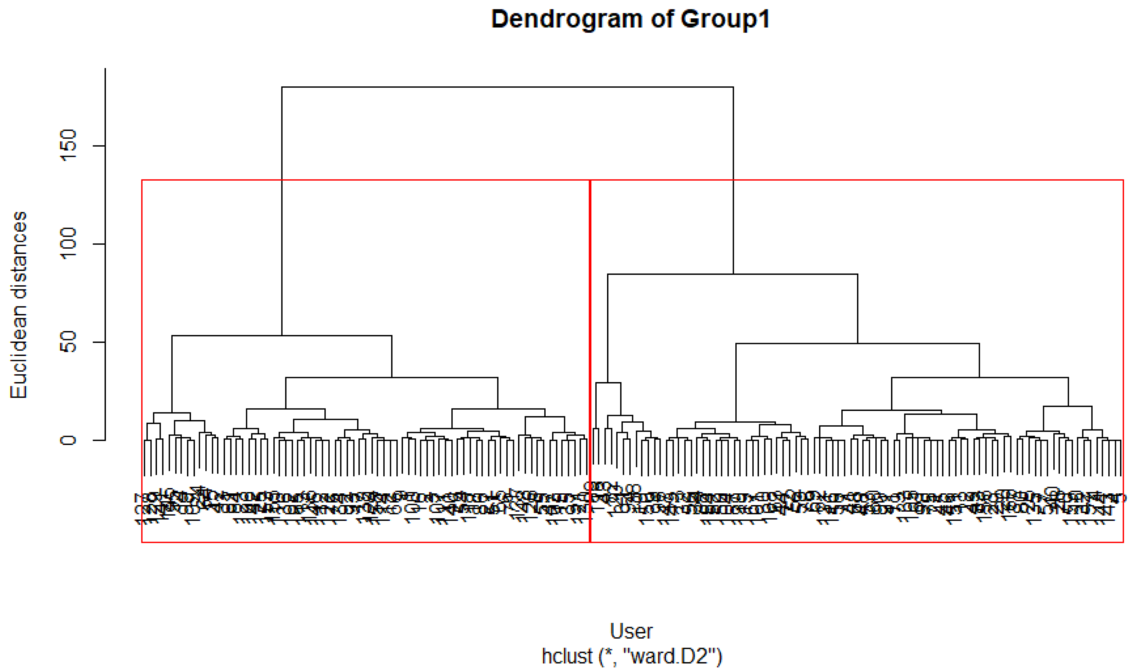
c. Average Linkage Clustering (Average distance)



d. Complete Linkage Clustering (Farthest points)



2. For Group 1
    a. Ward’s Method



2.4 K-means Clustering

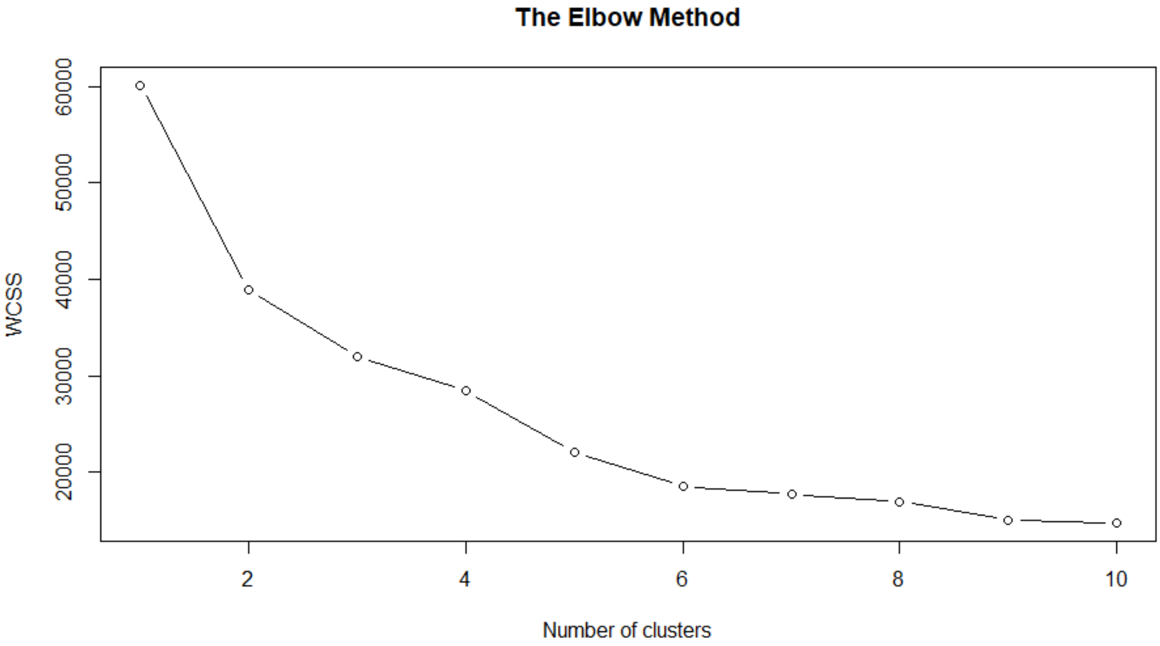
In k-means clustering, the number of clusters must be specified initially. Bad random initialization can lead to the wrong convergence of the algorithm. This can be solved by kmeans++ algo. So, instead of guessing the number of clusters, we used one of the methods called Elbow method that gives the optimal number of clusters based on the data. 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 k-means finds out k clusters and then calculates the sum of squares within the cluster. So that we need to try out for different number of clusters.

$$\text{WCSS} = \sum_{C_k}^{C_n} \Big(\sum_{d_i \, in \, C_i}^{d_m} distance(d_i, C_k)^2 \Big)$$

Where,
C is the cluster centroids and d is the data point in each Cluster.

So, at the end we can form any number of clusters (maximum 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 technique needs to plot all the WCSS$_i$ points against i clusters and find out from where the curve starts to bend like a knee or an elbow. That is why it is called an elbow method. It is usually vague since it is hard to interpret the bending point sometimes, and depends on the individual perspective and requirements.

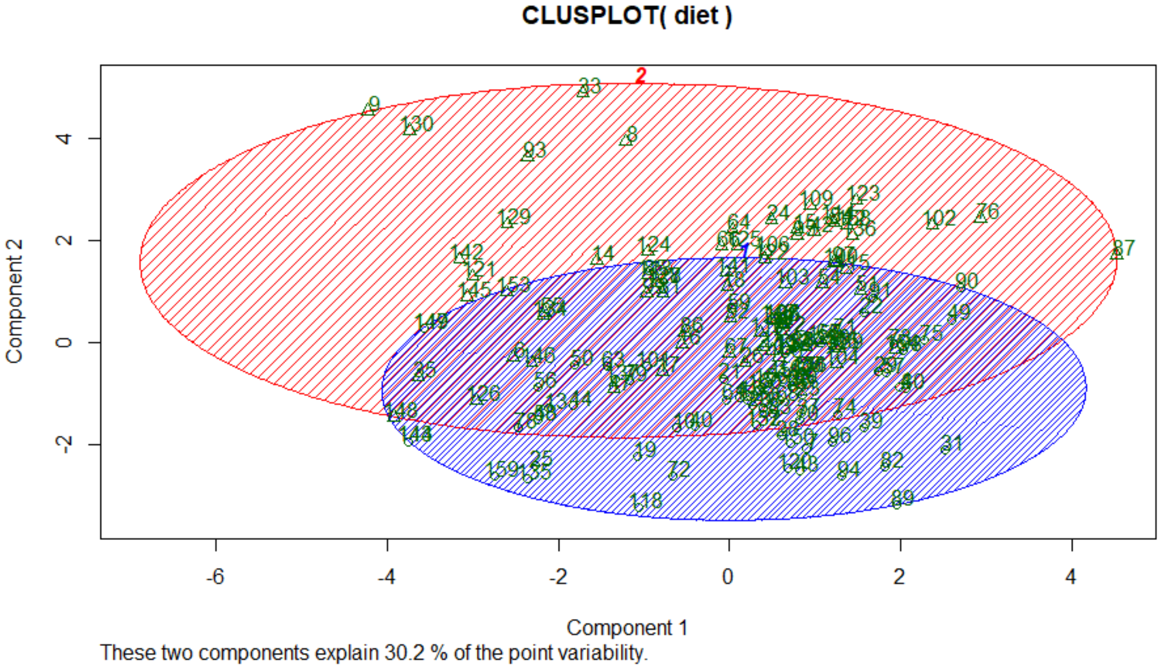
1. For Complete Dataset
    a. Elbow Method



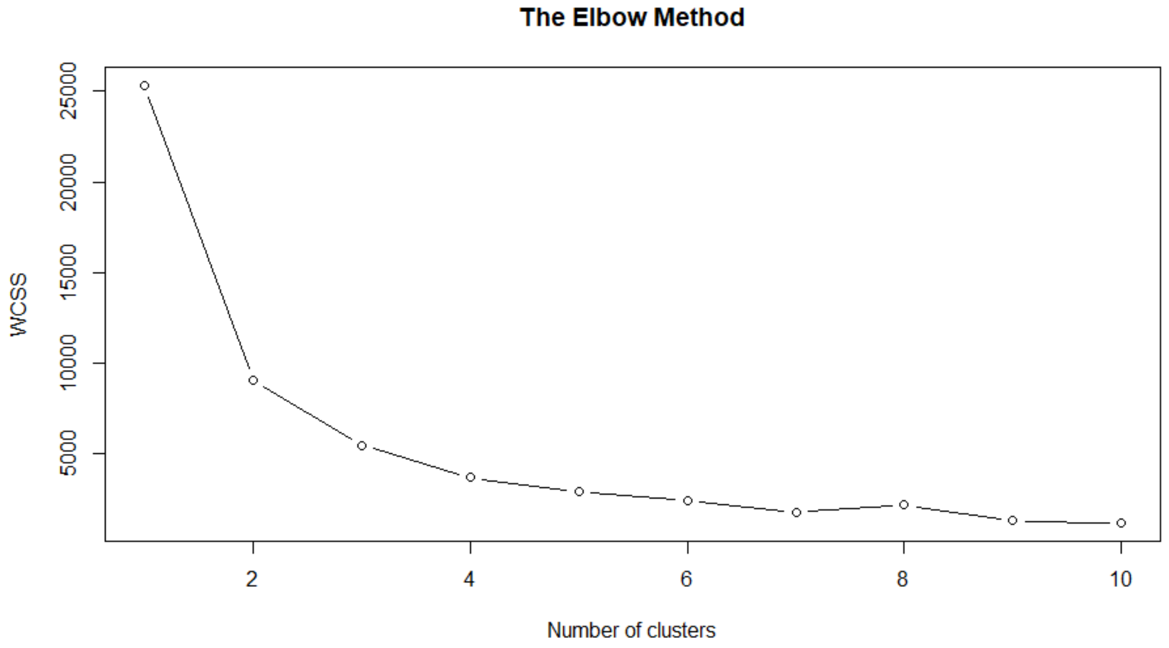
As already stated, it is very difficult to detect the point where it starts bending. All the points after 2 seem to be linear with little change, so we chose 2 as the number of clusters for the dataset.

Now let’s try to visualize these 2 clusters using k-means.

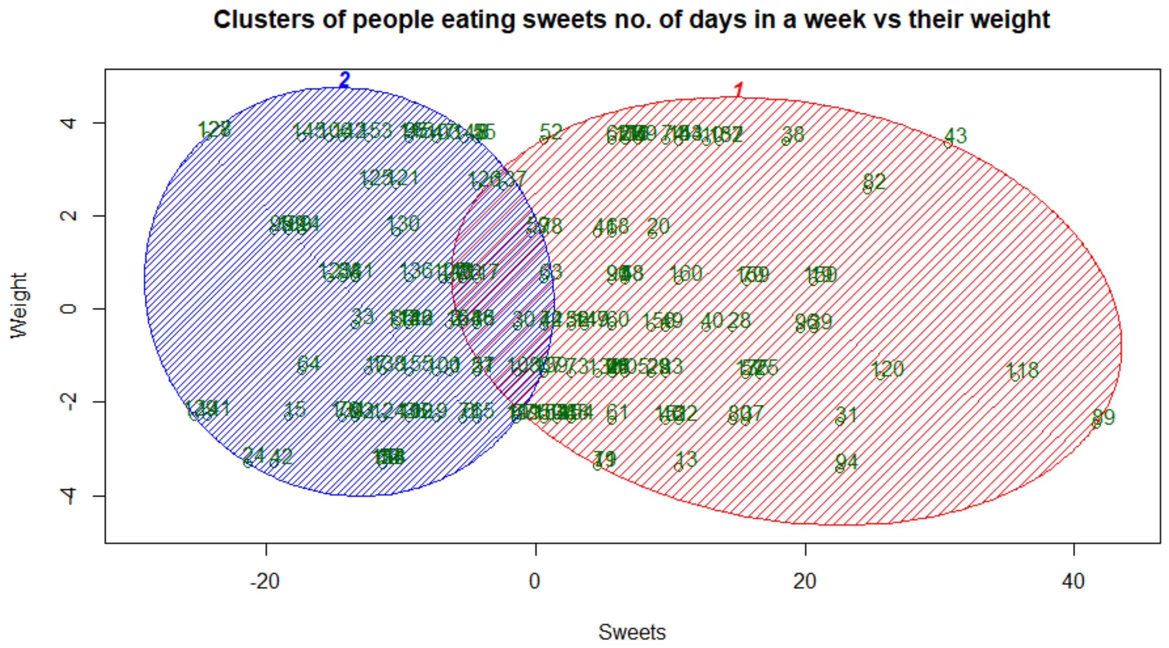
    b. K-Means Clustering



2. For Group1
   a. Elbow Method



   b. K-Means Clustering



These two components explain 100 % of the point variability.

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.

The code for this project is available at the Github link: https://github.com/Shubham-Girdhar/MVDA-Cluster-Analysis

Acknowledgement:

This project was possible under the guidance and direction of Prof. Dr. Christina Andersson. This report is the pre-requisite for the course 'Multivariate Data Analysis' taught in the Master's programme High Integrity Systems (HIS).

References:

[1] Bock, T., n.d. [Blog] *What is Hierarchical Clustering?*, Available at: <https://www.displayr.com/what-is-hierarchical-clustering/> [Accessed 9 January 2021].

[2] Dabbura, I., 2018. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. [Blog] Available at: <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a> [Accessed 9 January 2021].