Assignment_myusuf2_5_Summary

Fundamentals of Machine Learning-Assignment-5

Hierarchical Clustering

Problem Statement

The dataset Cereals.csv includes nutritional information, store display, and consumer ratings for 77 breakfast bowls of cereal.

Data Preparation

As a matter of priority, we would load all of the requisite packages that would be needed for this problem. Specifically, "ISLR", "caret", "dplyr", "tidyverse", "factoextra", "ggplot2", "proxy", "NbClust", "ppclust", "dendextend", and "cluster" would be loaded for the purpose of solving all the problems in question.

Review of Data Structure:

I used the command “head(Assignment5)” to review the first few rows of the data set, I was able to investigate the structure of the data set using str(Assignment5), and investigated the summary of the data set with the application of “summary(Assignment5):.

Data Preprocessing:

Remove all cereals with missing values.

We created a duplicate of the data set for preprocessing, scaled the dataset before placing it into a clustering algorithm, thereafter we removed all the missing values “NA” from the data set, and finally removed all the Nas by reviewing the scaled data set.

We tidy the data set by application of pivot longer on columns 4 to 14.

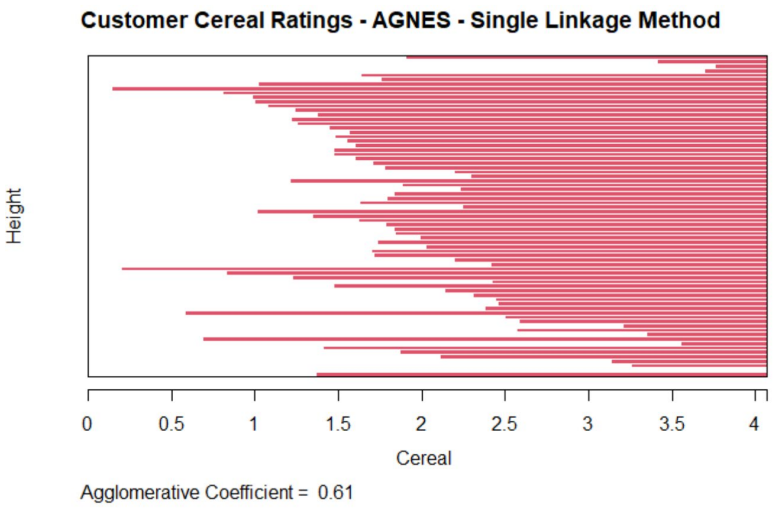
Subsequently, following the pre-processing and scaling of the data, the total number of observations were reduced from 77 to 74. This means that there were only 3 records with "NA" values in the data set.

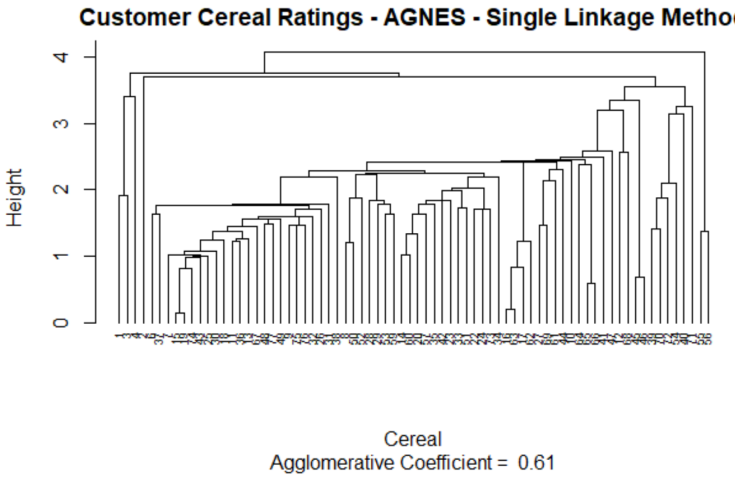
QA. Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements. Use Agnes to compare the clustering from single linkage, complete linkage, average linkage, and Ward. Choose the best method.

Solution:

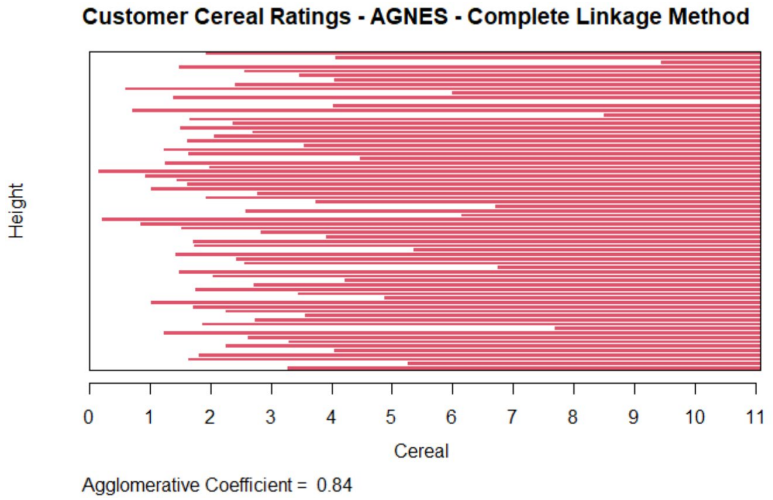
We created the dissimilarity matrix for the numeric values in the data set via Euclidean distance measurements, then Performed hierarchical clusterings via the single linkage, complete, average, and ward linkage methods before Plotting the results of the different methods as follows:

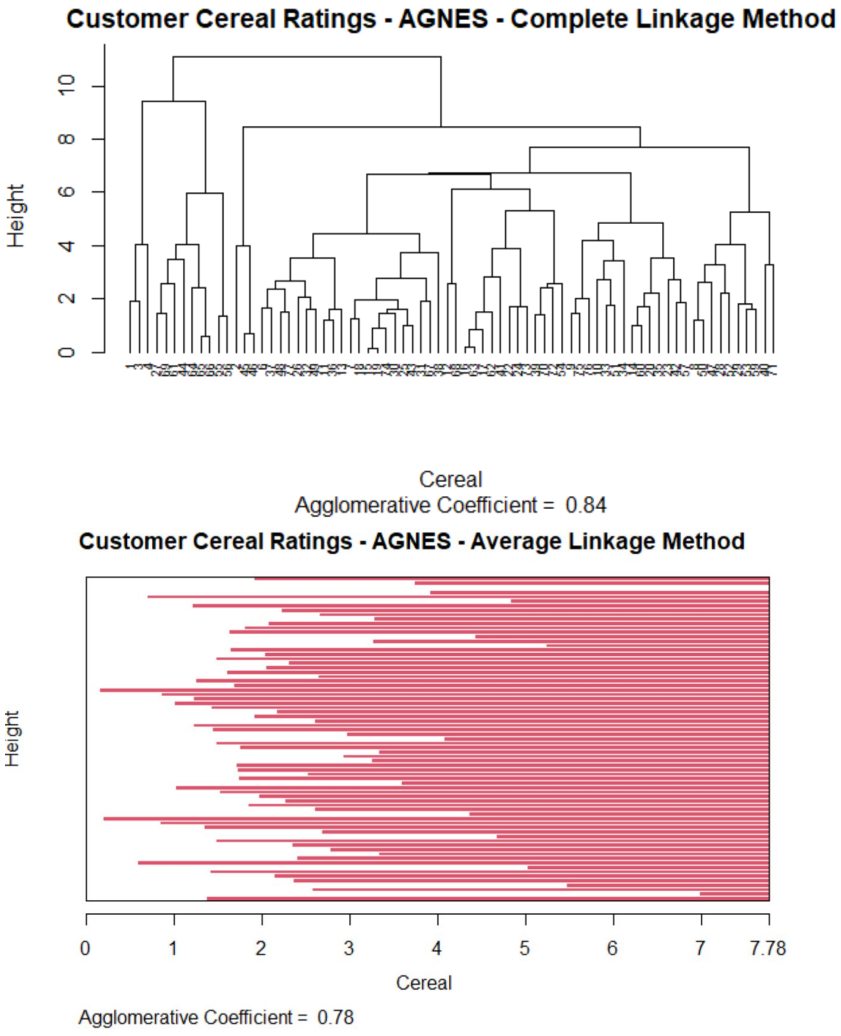
Single Linkage:



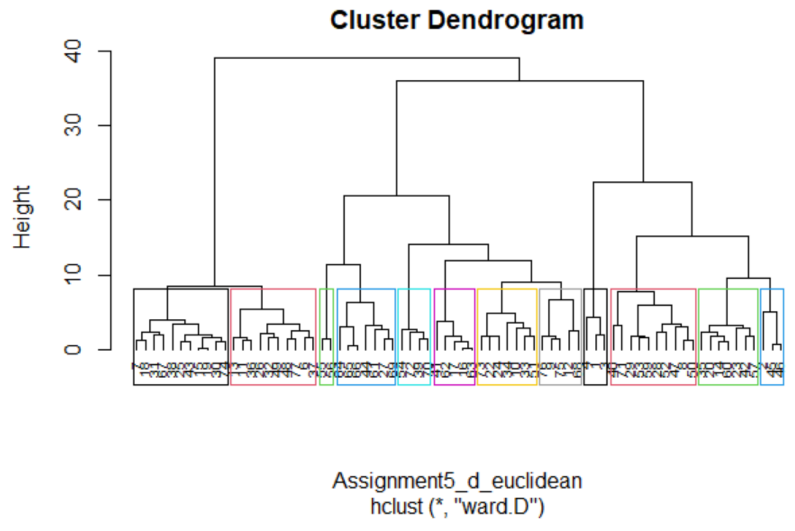


Complete Linkage:





Ward Linkage:



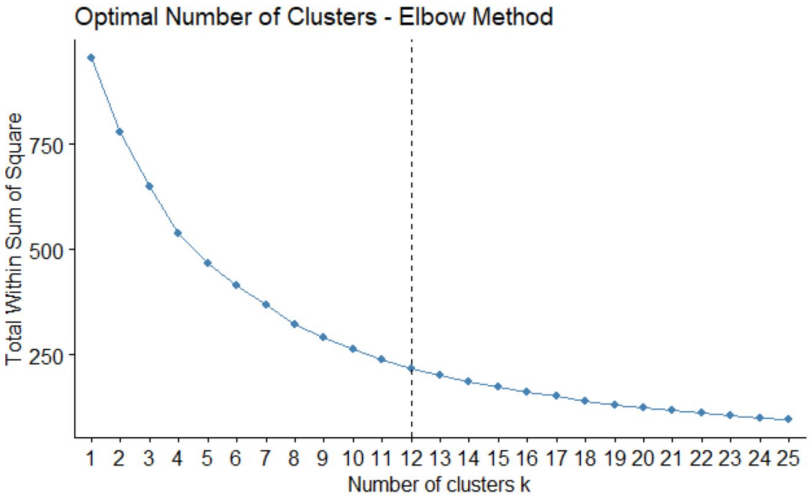
In the context of the concept of clustering, the ideal and best clustering method would be based on the agglomerative coefficient that is returned from each method. The closer the clustering structure to the value of 1.0, the better the cluster would be. This means the method with the value closest to 1.0 which is 0.90 is chosen.

Single Linkage: 0.61
Complete Linkage: 0.84
Average Linkage: 0.78
Ward Method: 0.90

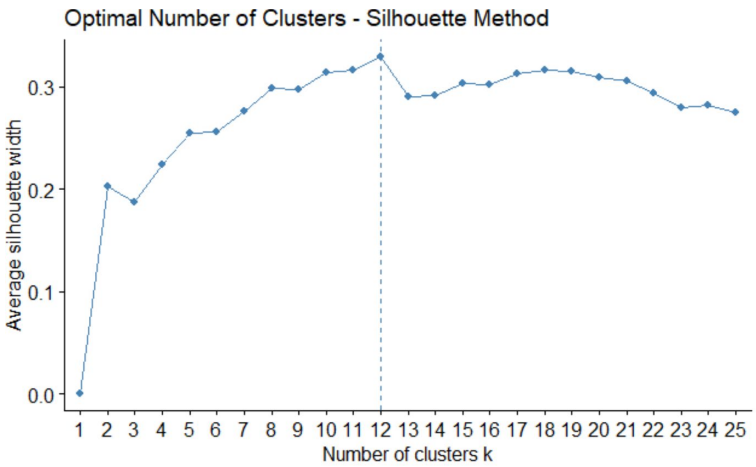
The above indicates that Ward AGNES is the method with the highest number of 0.90. However it should be noted that Single is sensitive to noise and outliers, Complete is less sensitive to noise and outliers but tends to break large clusters and is biased. So, Average linkage hierarchical is the ideal clustering as it compromises between Single and Complete but we should maintain our choice of Ward AGNES method.

QB. How many clusters would you choose?
In order to ascertain the appropriate number of clusters to choose from, we have chosen to use the elbow and silhouette methods. There are other methods, anyway.

Elbow method:

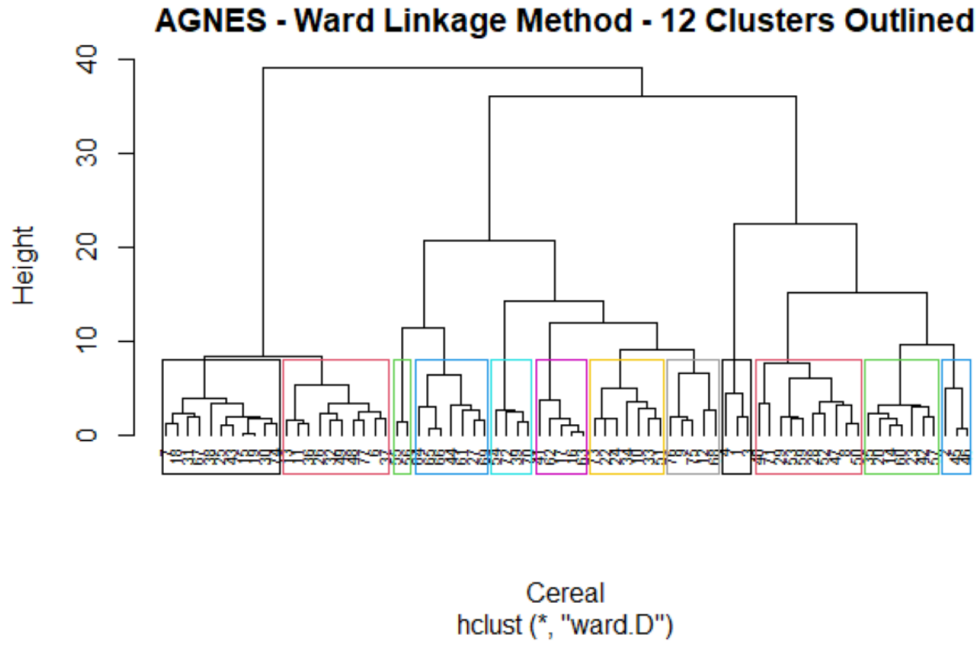


Silhouette Method:



Grounded in the analysis of the silhouette and elbow methods, we now know that the most appropriate number of clusters would be 12 in tune with the rules of Hierarchical Clustering.

I now proceeded to outline the 12 clusters on the hierarchical tree on the Ward Linkage method as depicted below:



Solution:

QC. Comment on the structure of the clusters and their stability. Hint: To check stability, partition the data and see how well clusters formed based on one part apply to the other part. To do this:
1. Cluster partition A
2. Use the cluster centroids from A to assign each record in partition B (each record is assigned to the cluster with the closest centroid).
3. Assess how consistent the cluster assignments are compared to the assignments based on all the data.

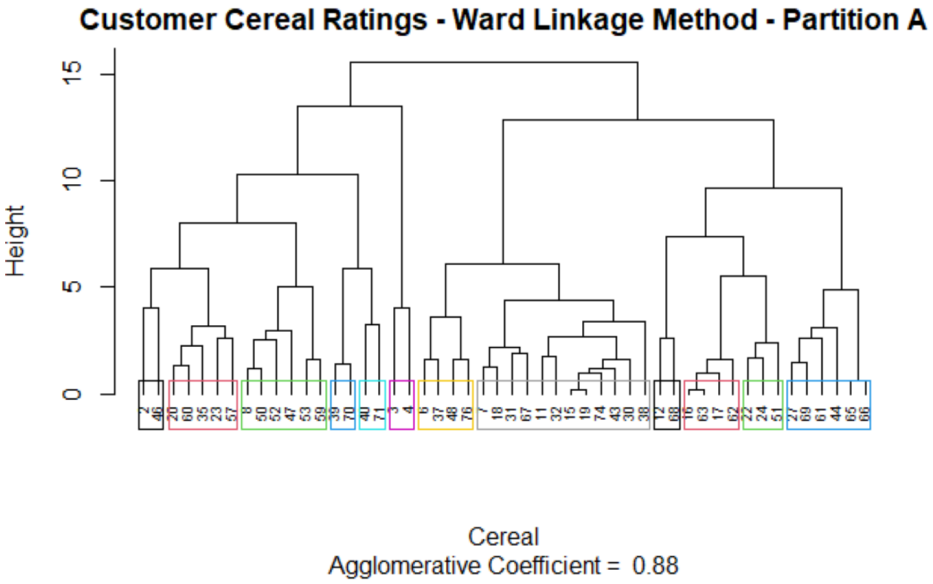
Solution:

At first, I began with all Data Assigned Clusters by establishing that the assigned clusters for all data sets would be in the "Assignment5_preprocessed_1".

I now partitioned the data set. In my quest to check the stability of clusters, the data set was split and listed into a 70:30 partition. The 70% was used to create cluster assignments again, and then the remaining 30% was assigned based on their closest centroid among the variables.

I set the seed for randomized functions to “123”, and Split the data into 70% partition A and 30% partition B, respectively.

The partitioned data is used to re-run the clustering and for the assignment, we maintained the same K value of 12 and ward the clustering method to determine the stability of the clusters. We then assigned the clusters to the nearest points in the Partitioned B data set for clusters 1 to 12 as identified. We have the plotted results of the different methods as depicted herein:



To achieve the required outcome, the centroids for each of the clusters would need to be calculated individually, so we could find the closest centroid for the data points in partition B. The summarized centroids are as follows:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.250040 -0.141338  0.008143  0.145452  0.396052  0.935190
```

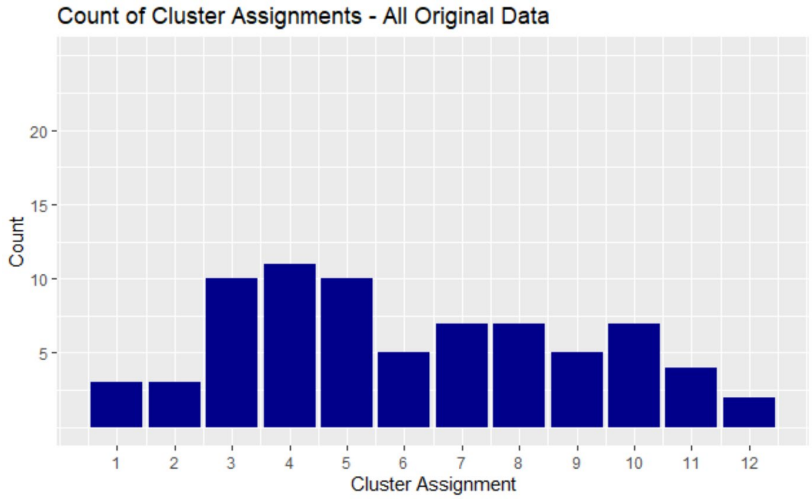
The next step is for us to calculate the Centers of the partitioned B data set, then calculate the distance between the centers of partitioned A and the values of partitioned B.

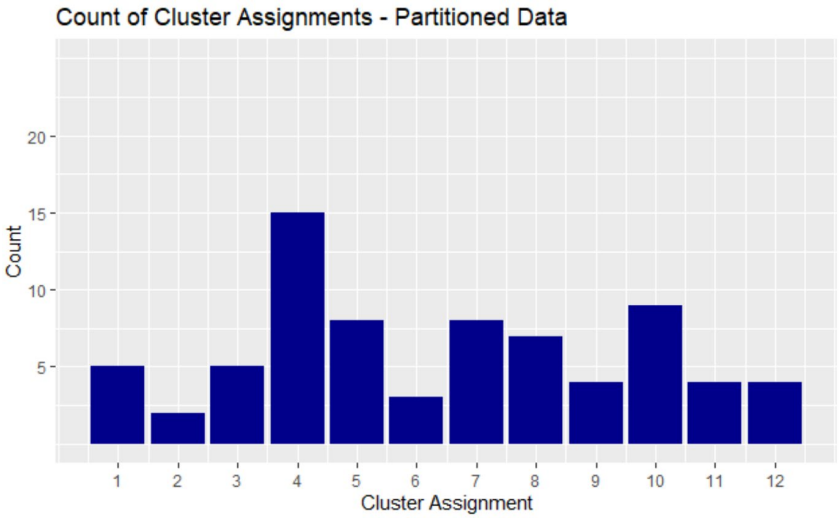
As it is, currently, the data has been assigned by both methods (whole data and partitioned data), we could then compare the number of matching assignments to see the stability of the clusters.

Given the outcome of the analysis, it could be established that the clusters are unstable. Vividly, it could be seen that with 70% of the data available, the resulting assignments were only

identical for 38 out of the 74 observations made. This results in a % repeatability of the assignment under review.

The next phase is; we visualized the cluster assignments to see any difference arising between the two clusters. We did that by plotting the original hierarchical clustering algorithm, and then the partitioned clustering algorithm as depicted below;





From the output, we could vividly assess the visualized data, we could equally see that Cluster 3 significantly shrunk when using the partitioned data. As a result, several of the other clusters became relatively larger. In addition to that, we could see that from the chart, it seems that the clusters are more evenly distributed across the 12 clusters when the data is partitioned, unlike the original data.

QD. The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of “healthy cereals.” Should the data be normalized? If not, how should they be used in the cluster analysis?

Solution:

Although this is a bit complex, in this situation, normalizing the data would not be the best idea. It would not be proper, because the scaling and normalizing of the cereal's nutritional information are basically based on the sample of cereal in question being analyzed. Therefore, the collected dataset could include only cereals with very high sugar content and very low fiber, iron, and other nutritional information as the case may be. This means, the moment it is scaled and normalized across the sample set, it is would be challenging to state how much nutrition the cereal would give a child. To an unacquainted viewer, he/she might assume a cereal with 0.999 for iron would mean it has almost all of the nutritional iron a child needs; however, it may just be the best of the worst in the sample set having nearly no nutritional value at all.

in the end, a more suitable means for preprocessing the data would be to make it a proportion to the daily suggested calories, fiber, carbohydrates, to mention but a few for a child. This would enable specialists to make a better-informed decision about the types of clusters when evaluating, but this is not to permit a few larger variables to surpass the distance estimates. It is suggested that in assessing the clusters, an expert could examine the average for the cluster to ascertain what proportion of students' daily recommended nutrition would come from certain components of the cereals. This would obviously allow the team to make informed decisions on what the ideal healthy cereal clusters to select from, based on the established inferences.