**Homework 4**

**Instructions:**

The data file for this homework is Cereals.xls, which is to be downloaded from Canvas. Create a new Word document and save it as HW4Answers\_X (where X is your team number). Where required, write your answers or paste screenshots into this Word document. Your response should not exceed 100 words for each below question. Write every member’s full name and participation on the first page of the Word document as follows. You need to submit this Word document and XLMiner solution.

|  |  |  |  |
| --- | --- | --- | --- |
| Participant | Complete the Assignment before the Meeting (Y/N) | Percentage of Contribution | Justification |
| Anchal Atlani | Y | 100 |  |
| Nikhil Jadhav | Y | 100 |  |
| Soham Dhodapkar | Y | 100 |  |
|  |  |  |  |

The dataset Cereals.xls includes nutritional information, store display, and consumer ratings for 77 breakfast cereals.

Tasks:

a) Data Preprocessing. Remove all cereals with missing values.

Ans: Preprocessing steps:

1. The total number of records having missing values are 4, we have deleted those records.
2. Created dummy variables for mfr and type categorical variables.
3. Since the data consists of many numeric values we have first normalized the data.

The final sheet we have used for creating clusters is - Rescaling\_TrainingTransform.

b) Apply hierarchical clustering to the data using Euclidean distance to the normalized measurements. Compare the dendrograms from single linkage and complete linkage, and look at cluster centroids. Comment on the structure of the clusters and on their stability. Hints: (1) To obtain cluster centroids for hierarchical clustering, apply Excel’s Pivot Table to the “Predicted Clusters” table. (2) Running hierarchical clustering in XLMiner is an iterative process – run it once with a guess at the right number of clusters, then run it again after looking at the dendrogram, adjusting the number of clusters if needed.

Answer:

**Single Linkage Hierarchical clustering:**

1. We have used Single Linkage Hierarchical clustering using 3 clusters and 4 clusters, both are giving almost same dendrogram.

For K=4 we find the clustering more stable.

Refer HC\_Output4 and HC\_Output6 (for K=4) sheet single and complete linkage.

Centroids for 4 clusters:

Variables:

Calories, protein, fat, sodium, carbo, potass, sugars, fiber, vitamins, shelf, weight, cups, rating, mfr\_A, mfr\_K, mfr\_N, mfr\_G, mfr\_P, mfr\_Q, type\_C, mfr\_R, type\_H

Cluster 1:

0.309125614,0.360557769, 0.041832669, 0.090650584, 0.555493896, 0.412703955, 0.14713715, 0.328815977, 0.100079984, 0.301980198, 0.466666667, 0.388976378,0.668885939, 0, 0, 1, 0, 0, 0, 1, 0, 0

Cluster 2:

0.402615109, 0.258110415, 0.34348321, 0.330367745,0.365229111, 0.172356404, 0.409644284, 0.082229468, 0.142928557, 0.782885431, 0.359943978, 0.476940382, 0.313849747, 0, 0, 0, 0, 0, 1, 1, 0, 0

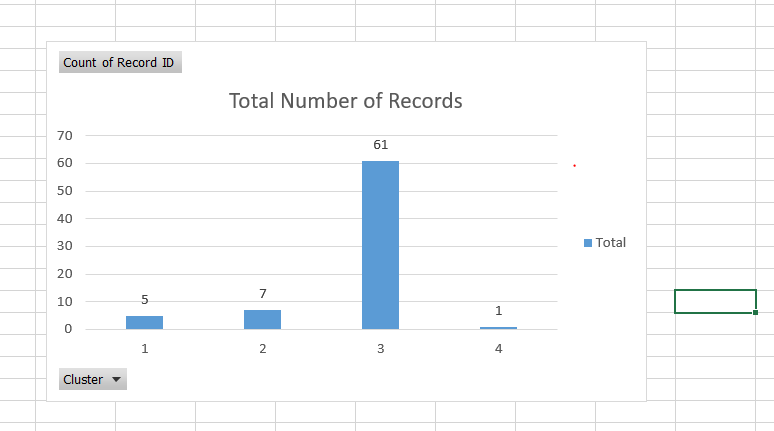
Cluster 3:

0.549916409, 0.299164, 0.197929593, 0.57018004, 0.558223104, 0.263871355,0.512551571, 0.152720937, 0.323805731, 0.613617919, 0.555448409, 0.459145476, 0.291313656, 0, 0.37704918, 0, 0.360655738, 0.147540984, 0, 1, 0.114754098, 0

Cluster 4:

0.454553718, 0.599601594, 0.201195219, 3.1248E-05, 0.610987791, 0.253983874, 0.200399467, 0.000713267, 0.25004999, 0.5, 0.5, 0.598425197, 0.486483395,1, 0, 0, 0, 0, 0, 0, 0, 1

Number of Records in each Single Linkage Clusters:



Single Linkage Distribution of records

**Complete Linkage Centroids for 4 clusters:**

Centroids for 4 clusters:

Variables:

calories, fat, protein, sodium, fiber, sugars, potass, shelf, vitamins, rating, weight, cups, mfr\_A, mfr\_K, mfr\_G, mfr\_N, mfr\_Q, mfr\_P, type\_C, type\_H, mfr\_R, carbo

Cluster 1:

0.121280979, 0.134794157, 0.599601594, 0.552080078, 0.785306705, 0.244784731, 0.936480223, 0.995049505, 0.25004999, 0.737448367, 0.5, 0.115485564, 0, .666666667, 0, 0.333333333, 0, 0, 1, 0, 0, 0.093044765

Cluster 2:

0.402615109, 0.34348321, 0.258110415, 0.330367745, 0.082229468, 0.409644284, 0.172356404, 0.782885431, 0.142928557, 0.313849747, 0.359943978, 0.476940382, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0.365229111

Cluster 3:

0.565473458, 0.201195219, 0.288979675, 0.568322001, 0.130068425, 0.524261437,

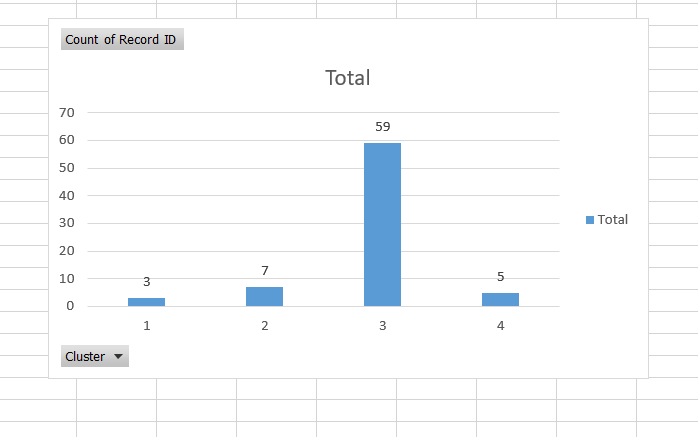
0.239456951, 0.600688035, 0.326305925, 0.274971813, 0.557328016, 0.470038703,

0, 0.355932203, 0.372881356, 0, 0, 0.152542373, 1, 0, 0.118644068, 0.572424237

Cluster 4:

0.363661153, 0.041832669, 0.360557769, 0.009405662, 0.186162625, 0.107190413, 0.295251095, 0.202970297, 0.100079984, 0.633072812, 0.466666667, 0.494488189, 0.2, 0, 0, 0.8, 0, 0, 0.8, 0.2, 0, 0.677580466

Total Number of Records Distribution



Records Distribution

**Cluster Structure**:

As visible from HC\_CLuster6 (From Complete Linkage) and HC\_Cluster4 (By Single Linkage) the low calorie and high vitamin cereals are grouped in one cluster i.e. cluster 1 and high calorie and less fiber content cereals are grouped in one cluster so we from this it can be deduce that proper clustering happened.

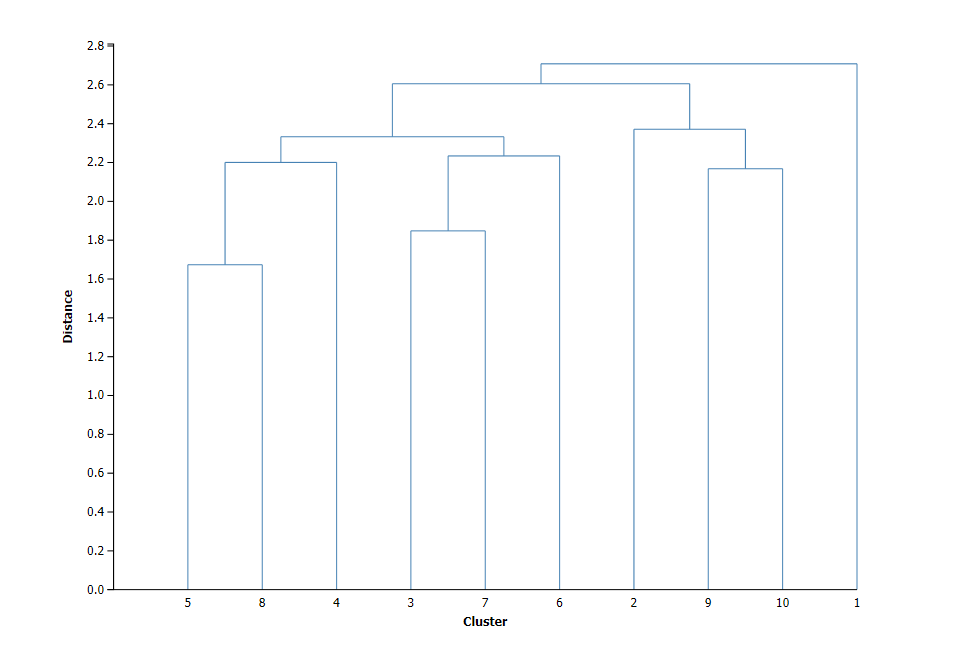
We have tried with different values of like k=3 and clustering for k=3 as well have similar kind of distribution, from this we can say that the clusters are quite stable.

For K=3 the output is present in the XL sheet.

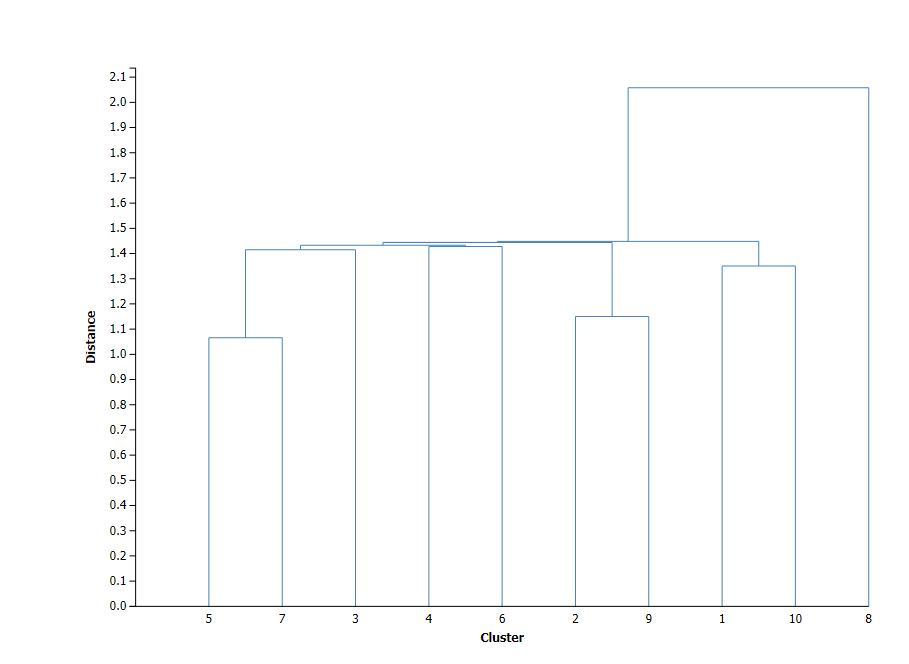
c) Which method leads to the most insightful or meaningful clusters?

Answer. Following are the dendrogram generated using Complete linkage and single linkage method:

**Complete Linkage:**



**Single Linkage**



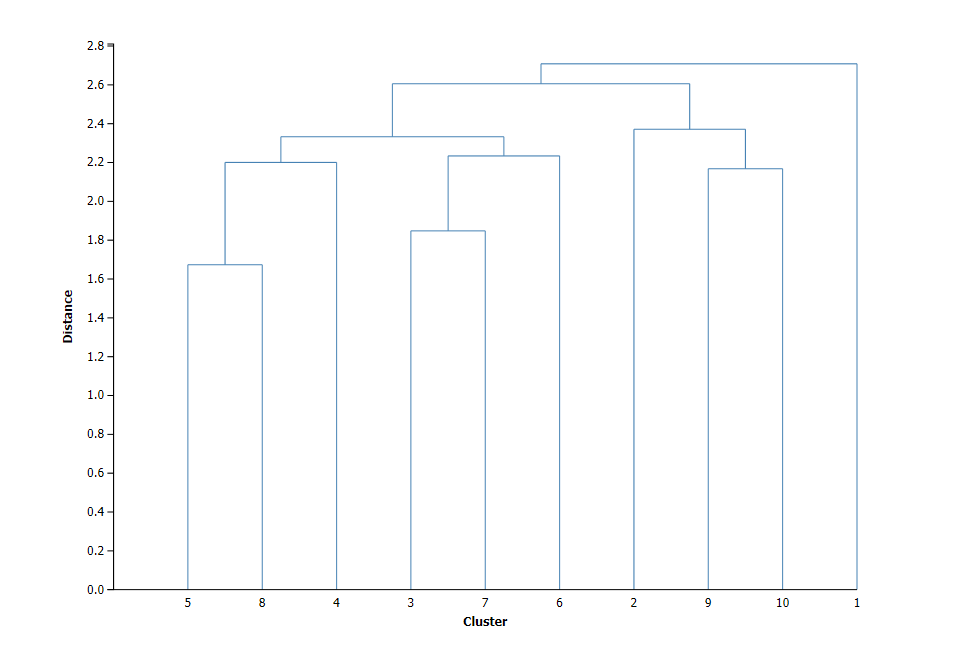
Looking at the dendrogram generated using single linkages a complete linkage.

The complete linkage dendrogram gives the even distribution of leaves when compared to that of the single linkage dendrogram. In the single dendrogram, we face the common issue of chaining, whereas in complete linkage we are not facing this issue. Also, the distance between each cluster is considerably good in complete linkage while in single linkage the distance between all clusters is very less which leads to chaining.

d) Choose one of the methods. How many clusters would you use? What distance is used for this cutoff? (Look at the dendrogram.)

Answer: Dendrogram generated using complete linkage gives a good distribution of leaves when the cluster size is set to 4.

**Complete linkage:**



However, we do not get these good results when we consider a single linkage method.

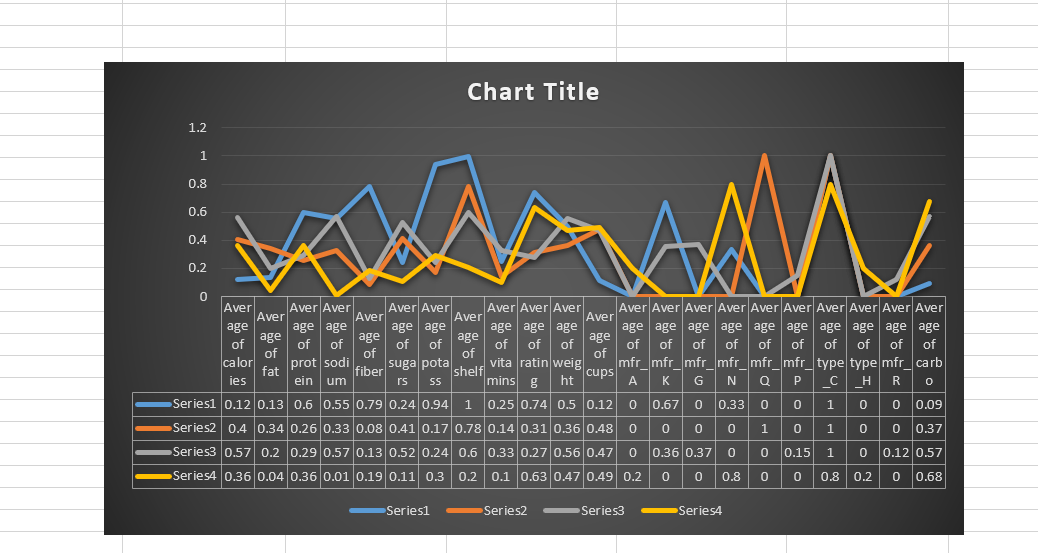
Looking at the dendrogram generated the cutoff value for a cluster size of 4 is **2.37115**.

e) 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?

Answer: We are considering that the cereals with higher protein, vitamin, fiber and lower fat could be considered as a healthy cereal for children. In this case from the cluster analyses performed in the tasks above, it is conclusive that cereals from Cluster 1 is the healthier of all the other formed clusters. Also, the data had to be normalized from the previous tasks. If the data is not normalized, it could lead to worse results. As normalization scales all the variables to a certain range, it leads to a better performance although it would not affect the clustering results as such.

In conclusion, Cluster 1 has the healthiest cereals for the elementary schools to include in their cafeteria.

If we look at the line chart of centroids of Complete Linkage we will get the cluster 1(series 1 in below line chart) having all the properties of healthy cereals.



**Important submission instructions**

Save your Word file and XLMiner solution. Use the link “Homework 4” to upload these files. **Due by 11.59 P.M. April 15, 2019.**