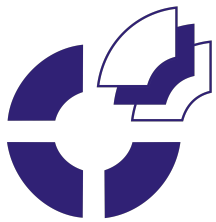

Indian Institute of Management
Indore

Multivariate Analysis

Applying Hierarchical Clustering On Wholesale Reseller Data

March 16, 2019

Aarjav Sethi	2016IPM002
Apurv Chaudhary	2016IPM022
Harshvardhan	2016IPM043
Pradeep Charan	2016IPM075
Tamanna Gupta	2016IPM112

Contents

Abstract

Hierarchical Clustering is an algorithm to group data points on the basis of their similarity. Wholesale Market Data is data from a distributor about the way they distribute various goods and the revenue generated from it. Our aim was to find similarity between such resellers based on their buying patterns. While there are various different choices to make when deciding to implement it, we did our analysis with four different distance measures - single, complete, average and centroid. We also used two different distance measures - euclidean and correlational. We finally concluded with analysis of various Dendrograms and inferences from them.

Keywords: Hierarchical Clustering, Machine Learning, Wholesale Market Dataset

1 Notes on Data

The data was obtained from University of California, Irvine's Machine Learning Repository. As per the website the data has multivariate characteristics. It has all integers as its attributes. There are 440 instances and each observation has eight different parameters. The data, fortunately, has no missing values.

The details and explanations of all the attributes are in the table below.

Attribute	**Explanation**	**Type**
FRESH	Annual Spending on Fresh Products	Continuous
MILK	Annual Spending on Milk Products	Continuous
GROCERY	Annual Spending on Grocery Products	Continuous
FROZEN	Annual Spending on Frozen Products	Continuous
DETERGENTS	Annual Spending on Detergents and Paper Products	Continuous
DELICATESSEN	Annual Spending on delicatessen products	Continuous
CHANNEL	Horeca (Hotel/Restaurant/Cafe) or Retail Channel	Nominal
REGION	Region (Lisnon, Oporto or Other)	Nominal

1.1 Descriptive Statistics

Summary Statistics of the *continuous* attributes are in the table below.

Attribute	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
FRESH	3	112151	12000.30	12647.329
MILK	55	73498	5796.27	7380.377
GROCERY	3	92780	7951.28	9503.163
FROZEN	25	60869	3071.93	4854.673
DETERGENTS	3	40827	2881.49	4767.854
DELICATESSEN	3	47943	1524.87	2820.106

Summary Statistics of *nominal* attributes are in the table below.

REGION	**Frequency**
Lisbon	77
Oporto	47
Other	316

CHANNEL	**Frequency**
Horeca	298
Retail	142

1.2 Inferences from the Data

MILK and FROZEN happen to be stable and staple diet of people. People, be it at retail shops or at at Horeca, always consume it. MILK, for instance, is consumed regularly and thus has the highest *minimum* consumption levels.

2 Problem

2.1 Problem Description

The task to be performed is to explain the concept of hierarchical clustering and use the technique to cluster the customers in the dataset. Then, we have to interpret the results and provide insights on the methods applied. After this, we have to perform clustering mechanism using different similarity-based measures.

2.2 Practical Implications

In the business, it is essential to maintain uniformity of data. Classification of data makes the process more efficient in handling the records and future sales prediction. Cluster Analysis is one of the most used classification tool. Cluster Analysis of the above data helps the seller in market segmentation, i.e., dividing the customers into the groups on the basis of some shared parameters.

Another use of analysis of this data is for seeking a better understanding of buyer behaviors by identifying homogeneous groups of buyers. This helps the seller in developing the specific sales strategies for a particular group and can make changes to current model in an efficient manner according to the demands of that set of customers.

Also, it would help in Customer Relationship Management that is Companies try to maximize its relationship with the potential and current customers, in which data analysis is used for increasing customer retention and engagement with sales and profitability growth being the ultimate target. Data Analysis by Hierarchical clustering would enable to increase customer satisfaction by customizing their offerings according to customers and discover hidden patterns and match with individual's preferences. Analysis of the following data would facilitate the future expansion of the business, predicting the demand of certain types of goods in the specific area according to the previous statistics.

It will further help the wholesalers to manage their inventory according to the customer preferences and reduce costs of inventory management and also economies of scale as buying and producing in large scale according to customer demand will yield to supply in alignment with demand in long run, and the country may run in equilibrium in long run.

3 Brief Overview of Machine Learning Algorithms

3.1 Supervised Learning

In Supervised machine learning method, each instance of a training dataset is composed of input attributes and an expected output. The input attributes of a training dataset can be any kind of data: the pixels of an image, values of a database row or even an audio frequency histogram.

Consider a mapping function, $y = f(x)$. We have input data x from which we take out the output data. The goal is to approximate this mapping function, $f()$. For example, a biometric machine takes in inputs as fingerprint, it is able to generate the student's details.

3.2 Unsupervised Learning

In the Unsupervised machine learning method, we have the input variable x but no output variable. The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. This algorithm instead of giving the output, detects the pattern based on the characteristics of the input data. In this case, it is assumed that there is some function which happens to be producing these data and instead of trying to find these underlying factors of influence, we aim to segregate them into groups with maximum similarity. Two methods for unsupervised learning are *hierarchical clustering* and *k-means clustering*.

4 Theory of Clustering

Clustering is a method of dividing a group of data into more than one subgroups such that objects belonging to the same subgroup are in some way more similar to those which belong to some other subgroup (called cluster). Clustering is the method of arranging or organizing objects in such a way so that similar objects are together in the same cluster or group and dissimilar objects in different cluster. It is particularly helpful when qualities or traits of a particular group are not known beforehand. In such cases, we have to separate the objects on the basis of how similar or dissimilar they are to others rather than going through a checklist of qualities that a subgroup should likely possess.

Clustering, in this sense, is an objective rather than a method. There is no specific method as to how to achieve these specific clusters. Different algorithms give different kinds of clusters and different algorithms may suit some situation more appropriately than others. Some common clustering algorithms are K-means clustering and Hierarchical Clustering. The algorithm should be performed until the next merge would create a cluster with low

cohesion, i.e. a bad cluster. Also when you pick a number k upfront, and stop when we have k clusters (when we know data falls into k groups).

Clustering is a case of unsupervised learning. From a causal point of view, in a supervised learning method we train the algorithm/model with a set of available data and then try to predict the rest of them and the future occurrences. In an unsupervised learning algorithm, however, we try to find the inherent structure within the data which would be otherwise be neglected. In unsupervised learning, it is assumed that the data available is a result of some "latent" variables about which there is do data available. And then, we try to group such outcomes.

Clustering is widely used in fields where there are no predefined criteria for distinguishing one object from another. Common ones are – medical industry: there are many unknown types of Cancer/Diabetes and availability of data helps us to group them is some structured fashion; market segmentation: as explained in previous examples, marketing companies have tons of data about their consumers and using clustering they group them suitably to pitch required product.

4.1 Technicalities of Clustering

Successful Clustering requires to address three basic decisions. We must need to establish a similarity measure that is the distance of how "close" two observations are. Second is choosing an apt and accurate clustering algorithm amongst the wide choices of algorithms for grouping the data. Third, we need a reliable way to measure the distance amongst different clusters.

There are two ways to begin with,

1. **Agglomerative:** This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

2. **Divisive:** This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

To define what constitutes dissimilarity between two groups, there are several options at the statistician's disposal. The exact decision of the same is always context specific and can never be generalised. For instance, say we have n different purchase bills (and thus consumption profile) of people who purchase some goods at some online store. Each customer (and there are n such customers) has p different parameters which could be, e.g. Electronics, Housing, Stationary, Grocery, etc. So now, we have to group people on the basis on similar buying pattern. We don't know beforehand what the specific groups are – and for the time being, not even how many such sub groups we want. In such cases, we follow the algorithms and try to reach a particular number of subgroups which have sufficient dissimilarity with each

other but very similarity within the subgroup. The customers are thus distributed in several subgroups on the basis of thein consumption pattern.

Distance Measures. Cluster analysis is for categorising homogeneous subgroups, as the distance between two observations reduce the similarity increases. The distance measure measures similarity, similarity can be measured between two cases using either euclidean distance or correlation(Pearson).

Euclidean Distance. It is defined as,

$$d_0 = \sqrt{\sum_{j=1}^{n}(x_i - y_i)^2},$$

where x_i and y_i refer to different observations of Variable i where there are n variables. At each step, we write the Euclidean distance in the proximity matrix. A proximity matrix is a square symmetric matrix in which cell$[i, j]$ of the matrix that is row i and column j. The diagonal entries are 0 and off-diagonal entries represent the distance between have Euclidean distances between all the two observation. In each step, the smallest values in the euclidean matrices are taken for forming clusters. Hierarchical clustering becomes prolonged because proximity matrix needs to calculated again every time to take into account another cluster. However amongst clusters, how to best link the clusters and measure distance needs to be decided, thus the linkage also needs to be decided.

Pearson Correlation Distance Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It has a value between $+1$ and -1, where 1 is a total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

Correlation-based measures take two observations similar even when the distance between them is significant, but there is a high correlation between their features. It is calculated between observation profiles' features to find the similarity.

Dendrogram A dendrogram is used in hierarchical clustering as a tree diagram that shows the similarity between observations by representing the clusters.

In the example Dendrogram 1, the data in clusters is shown in left that shows the similarity between observations. Each leaf is representing a observation and as we move up some get fused in branches. Branches at the same height are similar and the greater the vertical distance between two branches the more dissimilar. Following can be said,

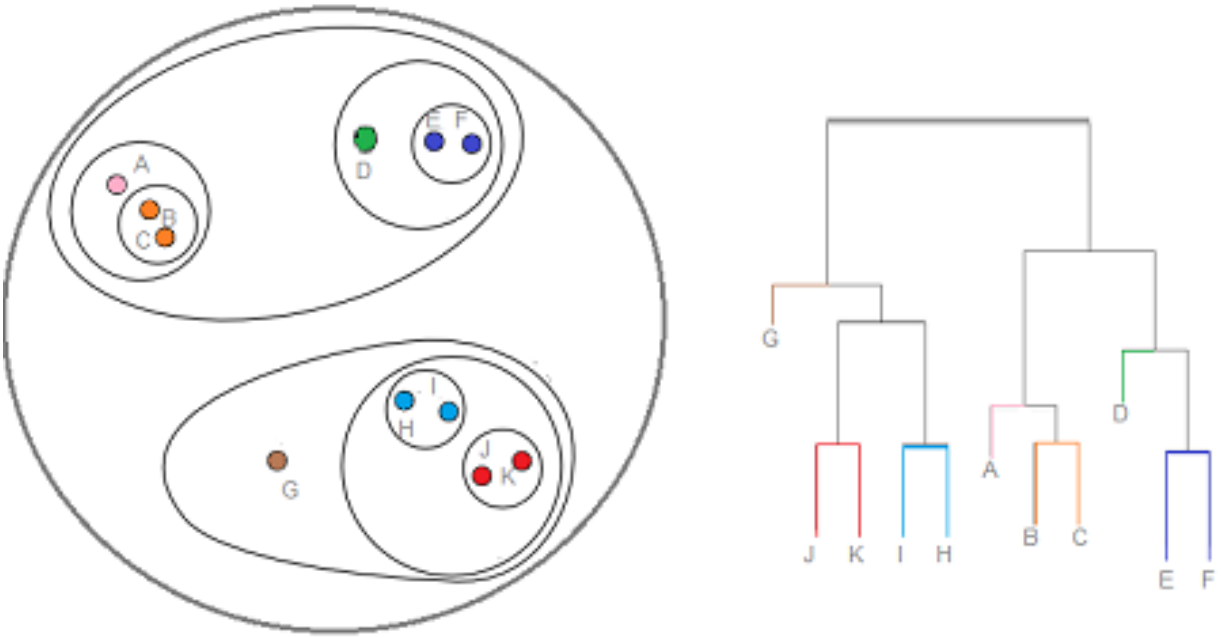

Figure 1: A dendrogram (right) representing nested clusters (left).

- J and K are similar to each other and also, I and H are similar to each other, E and F are similar to each other and also B and C are similar
- A is most similar to the cluster of B and C
- D is most similar to cluster E and F
- Cluster J and K are similar to Cluster I and H
- Point G is similar to cluster of J,K ,I and H
- Cluster of A,B,C are most similar to cluster of D,E F

4.2 Implementation of Hierarchical Clustering

The hierarchical clustering algorithm is a very simple algorithm for agglomerative clustering. The process is as follows,

1. We define a dissimilarity measure. Let it be Euclidean Distance.
2. Start from bottom of the dendrogram, each of n observations are treated as separate clusters.
3. The two most similar clusters are fused, leaving $n - 1$ clusters.

4. Again, from the remaining, the most similar clusters are fused,but there we need a dissimilarity measure for a group of observations, so we have linkages and we choose any one amongst the 4 that is single, complete, average and Centroid. Clusters that are most similar are fused leaving $n - 2$ clusters.

5. It repeats from 1 to 4 until all observations are in a single cluster.

Types of Linkages Once we are done with our choice of distance, the type of linkage to choose becomes another point of confusion. While doing hierarchical clustering as one crosses the first stage, where each object is a cluster, there forms at least one cluster of minimum two objects (as we select the objects with minimum distance between them and club them) . So the problem arises in the second stage when one has to measure the distances among the clusters with more than one object and between clusters ($objects > 1$) and clusters ($object = 1$).

There are four commonly used linkage measures used to compute distances between two clusters containing 2 or more objects. These are based on calculating distances between individual objects from either of the clusters and considering the main distance as the smallest, largest or something intermediate. The description of some of these linkages are:

1. Single Linkage

   - It is also called nearest neighbour or the minimum method.
   - As per this measure, the distance between two clusters is equal to the minimum distance of all possible distances between one object from the first cluster and the other from second cluster.
   - It is outlier-robust measure as however big the outlier may be, it only cares about the smallest distance.
   - One major drawback of this linkage measure is that it leads to high chances of chaining i.e. many clusters may join not because of true similarities but because two objects in them are very similar or close to each other.

2. Complete Linkage

   - It is also called the furthest neighbour or the maximum method.
   - Unlike the above linkage, it considers the maximum of all possible distances between one object from first cluster and the other from second cluster, as the distance between the two clusters.
   - It doesn't lead to unnecessary clustering i.e. chaining, caused due to a few close objects but it may lead to avoidance of necessary clustering also.

- It is heavily affected by the outliers as one big object can make the inter-cluster distance quite big and lead to discarding of many similarities of other smaller objects.

3. Average Linkage
   - It is also called as Unweighted Pair-Group Method using Arithmetic averages or (UPGMA).
   - This measure is a compromise between the single and the complete linkage measures as it takes the unweighted average of all the possible distances between one object from the first cluster and the other from second cluster, as the distance between the respective clusters.
   - This turns out be better as a way to evade the cons and take the pros of the above two extreme linkage measures.
4. Centroid Linkage
   - It is also a similar to average linkage but here the weighted/unweighted average is first carried out within a cluster to decide for the object or the parameters (same as of a object) which can be used to find out the distance between any two clusters.
   - It may lead to certain inversions of data which may not be desirable.

There is no "best" linkage as such and the objective, context, type of data etc. factors combined should lead to a selection of a linkage measure. We may even go for doing clustering analysis using all the linkage measures separately and later, based upon interpretation and differences adopt the most suitable of them.

5 Application and Interpretation

5.1 Application on Data

We ran various codes on R to operate on our data. The codes and explanation are presented below.

```
library(readr)
data <- read_csv("Wholesale customers data.csv",
                 col_types = cols(X10 = col_skip(), X11 = col_skip(),
                                  X12 = col_skip(), X9 = col_skip()))
#removed columns which had details of the columns
View(data)
```

```
x=matrix(c(data$Channel,data$Region,data$Fresh,data$Milk,data$Grocery,
data$Frozen,data$Detergents_Paper,data$Delicassen), ncol=8,byrow=FALSE)

#scaling variables for better implementation
x[,3:8]=scale(x[,3:8])

#new matrix storing all continuous variables
x.new=x[,3:8]

#descriptive plots of the variables
par(mfrow=c(2,4))
hist(x[,1],main="Histogram of Channel",sub="Horeca = 1, Other = 2")
hist(x[,2],main="Histogram of Region",sub="Lisbon = 1, Oporto = 2, Other = 3")
boxplot(x[,3],main="Boxplot of Fresh Products")
boxplot(x[,4],main="Boxplot of Milk Products")
boxplot(x[,5],main="Boxplot of Grocery Products")
boxplot(x[,6],main="Boxplot of Frozen Products")
boxplot(x[,7],main="Boxplot of Det_Paper Products")
boxplot(x[,8],main="Boxplot of Del_ Products")

#for parallel coordinates
library(MASS)
parcoord(x.new,var.label = TRUE)


#relabeling variables for convenience
x.orig=x
x=x.new


#performing clustering
#default distance is euclidean

hc.single=hclust(dist(x), method ="single")
hc.complete =hclust(dist(x), method="complete")
hc.average =hclust(dist(x), method ="average")
hc.centroid=hclust(dist(x), method ="centroid")


#making Dendrograms
#due to some internal error par(mfrow=c(2,2)) always crashed R sessions
```

```
plot(hc.complete,main="Complete Linkage")
plot(hc.average, main="Average Linkage")
plot(hc.single, main="Single Linkage")
plot(hc.single, main="Centroid Linkage")

#Dendrograms thus obtained didn't have any distinct clusters, so didn't use cutree()


#using correlation distance
#calculating correlation distance
hc.s=hclust(dd, method ="single")
plot(hc,hang = -1,main="Single Linkage with Correlation-Based Distance ",
xlab="", sub ="")

hc.c=hclust(dd, method ="complete")
plot(hc,hang = -1,main="Complete Linkage with Correlation-Based Distance ",
xlab="", sub ="")

hc.a=hclust(dd, method ="average")
plot(hc,hang = -1,main="Average Linkage with Correlation-Based Distance ",
xlab="", sub ="")

hc.c=hclust(dd, method ="centroid")
plot(hc,hang = -1,main="Centroid Linkage with Correlation-Based Distance ",
xlab="", sub ="")

par(mfrow=c(1,3))

c=cutree(hc,h=0.9)
x.mod=cbind(x,c)
x.mod

clust1=x[which(x.mod[,7]==1),]

clust2=x[which(x.mod[,7]==2),]

clust3=x[which(x.mod[,7]==3),]

par(mfrow=c(3,3))

boxplot(clust1[,1],main="Cluster 1",sub="FRESH")
boxplot(clust2[,1],main="Cluster 2",sub="FRESH")
```

```
boxplot(clust3[,1],main="Cluster 3",sub="FRESH")

boxplot(clust1[,2],main="Cluster 1",sub="MILK")
boxplot(clust2[,2],main="Cluster 2",sub="MILK")
boxplot(clust3[,2],main="Cluster 3",sub="MILK")


boxplot(clust1[,3],main="Cluster 1",sub="GROCERY")
boxplot(clust2[,3],main="Cluster 2",sub="GROCERY")
boxplot(clust3[,3],main="Cluster 3",sub="GROCERY")

boxplot(clust1[,4],main="Cluster 1",sub="FROZEN")
boxplot(clust2[,4],main="Cluster 2",sub="FROZEN")
boxplot(clust3[,4],main="Cluster 3",sub="FROZEN")

boxplot(clust1[,5],main="Cluster 1",sub="DETERGENTS")
boxplot(clust2[,5],main="Cluster 2",sub="DETERGENTS")
boxplot(clust3[,5],main="Cluster 3",sub="DETERGENTS")

boxplot(clust1[,6],main="Cluster 1",sub="DELICATESSEN")
boxplot(clust2[,6],main="Cluster 2",sub="DELICATESSEN")
boxplot(clust3[,6],main="Cluster 3",sub="DELICATESSEN")


x1=cbind(x.orig,c)

par(mfrow=c(1,2))
clust11=x1[which(x1[,9]==1),]


clust12=x1[which(x1[,9]==2),]

clust13=x1[which(x1[,9]==3),]

par(mfrow=c(2,3))
hist(clust11[,1],main="Channel: Cluster 1",sub=length(clust11[,1]))
hist(clust12[,1],main="Channel: Cluster 2",sub=length(clust12[,1]))
hist(clust13[,1],main="Channel: Cluster 3",sub=length(clust13[,1]))

hist(clust11[,2],main="Region: Cluster 1",sub=length(clust11[,2]))
hist(clust12[,2],main="Region: Cluster 2",sub=length(clust12[,2]))
hist(clust13[,2],main="Region: Cluster 3",sub=length(clust13[,2]))
```

```
x=x[,3:8]
sub=cutree(hc,k=3)
table(sub)

library(factoextra)
fviz_cluster(list(data = x, cluster = sub))

#developing tanglegram - comparative tanglegram between average linkages for
correlation and euclidean distances
library("dendextend")
hc1=hc.average
d1=as.dendrogram (hc)
d2=as.dendrogram (hc1)

tanglegram(d1, d2)
```

5.2 Output and Plots

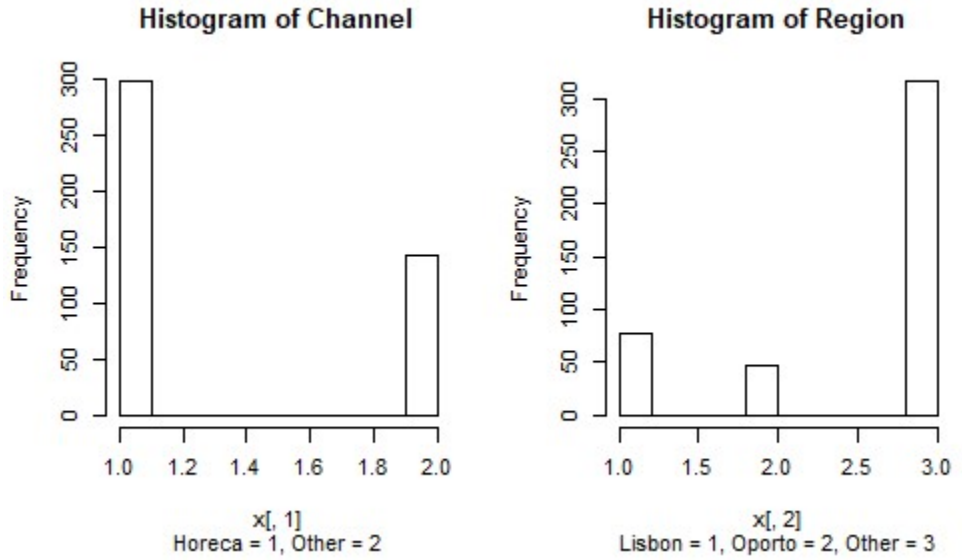

Figure 2: Histogram of Channel and Region. There are 77 Lisbon, 47 Oporto and 316 Other in Region. There are 286 in Horeca and 142 in Retail in Channel.

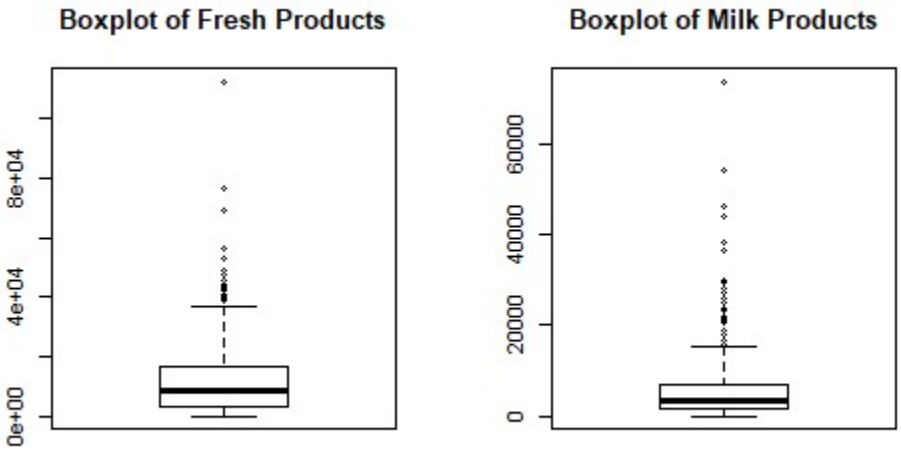

Figure 3: Boxplots of Fresh Products and Milk Products.

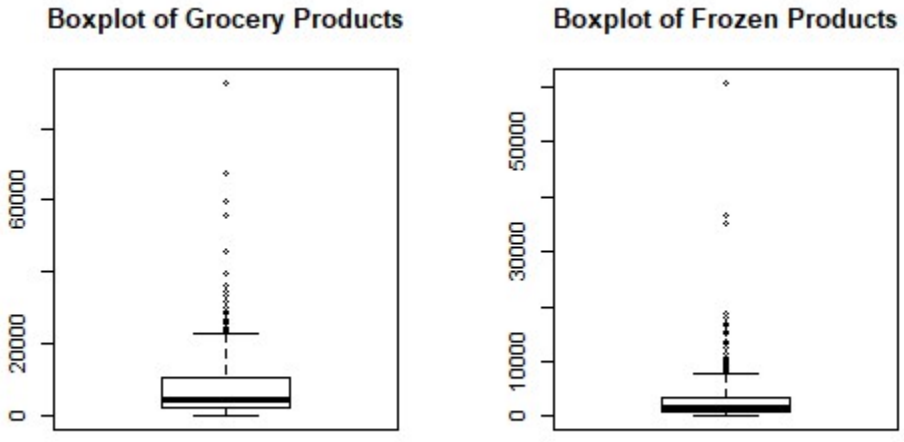

Figure 4: Boxplots of Grocery Products and Frozen Products.

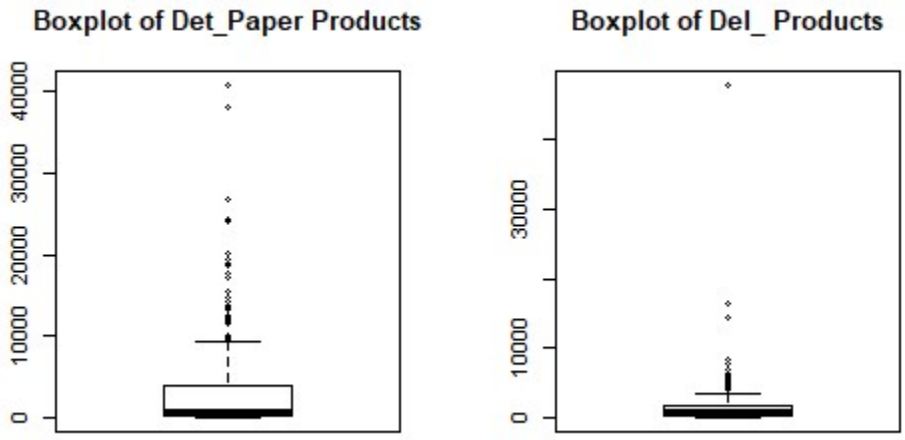

Figure 5: Boxplots of Detergent and Paper Products, and Delicatessen Products.

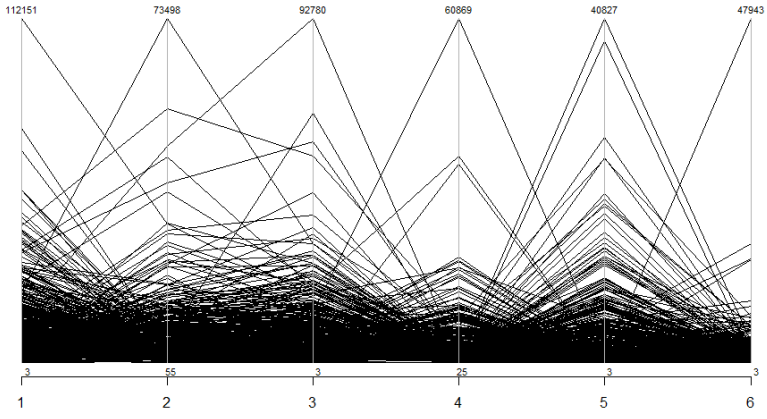

Figure 6: Parallel Coordinate Plot for all the variables.

Dendrograms

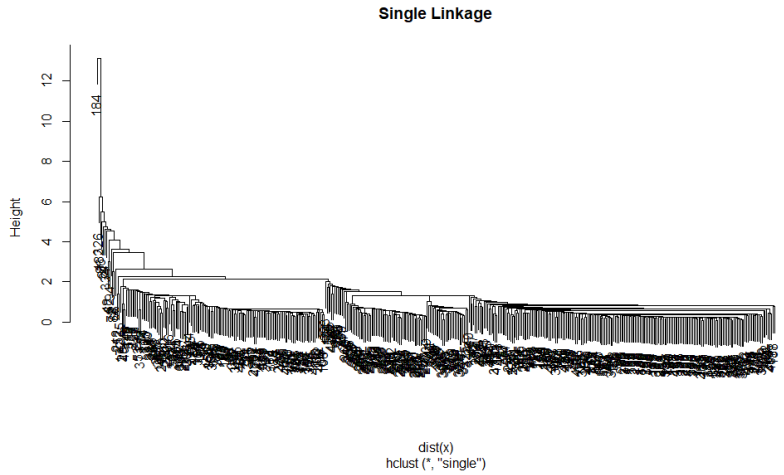

Figure 7: Dendrogram produced using *single* linkage and *euclidean* distance.

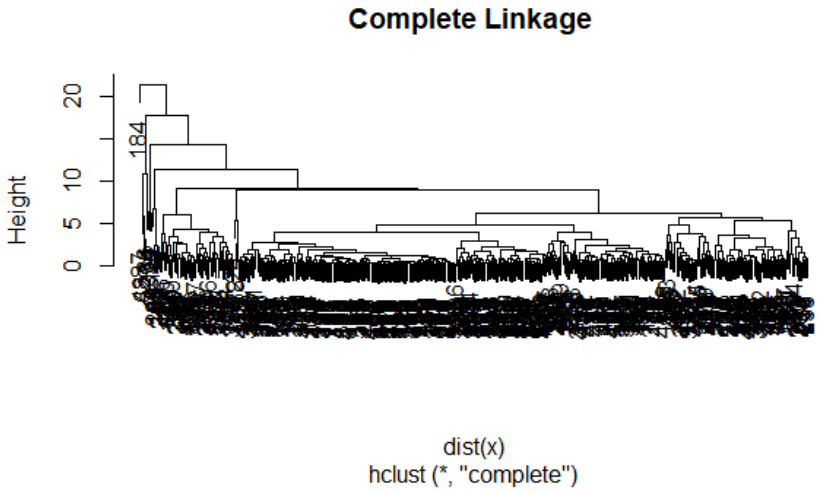

Figure 8: Dendrogram produced using *complete* linkage and *euclidean* distance.

Interpreting Clusters

As we can see, there are too many dizzy figures to begin with. To proceed, we chose *average* as our linkage method and *correlation* as our distance measure.

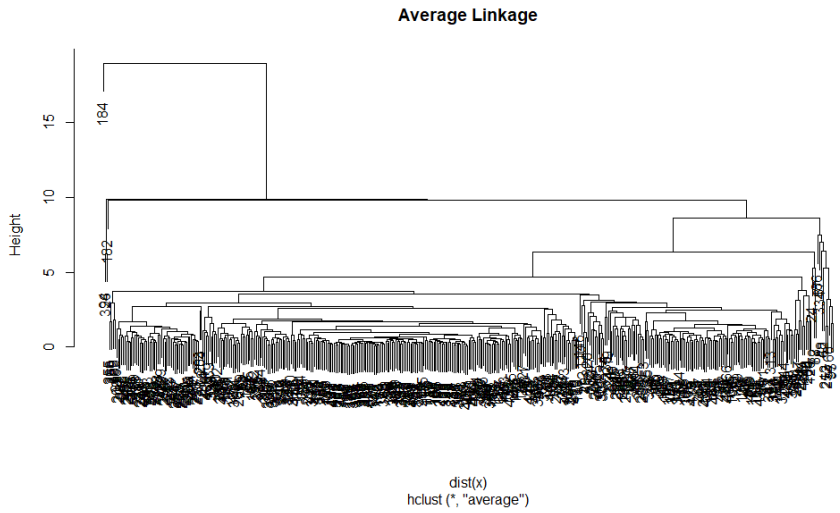

Figure 9: Dendrogram produced using *average* linkage and *euclidean* distance.

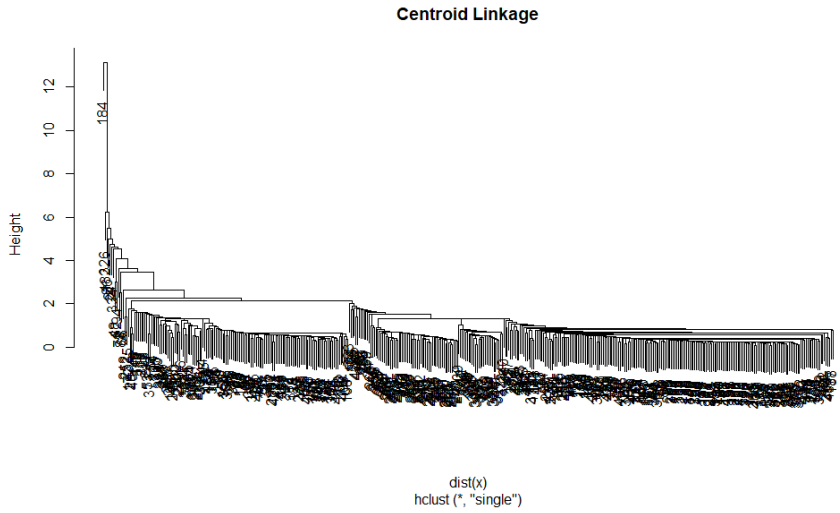

Figure 10: Dendrogram produced using *centroid* linkage and *euclidean* distance.

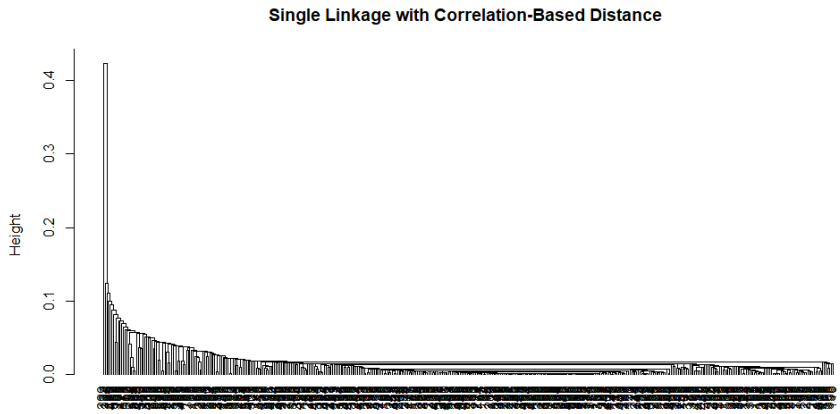

Figure 11: Dendrogram produced using *single* linkage and *correlation* distance.

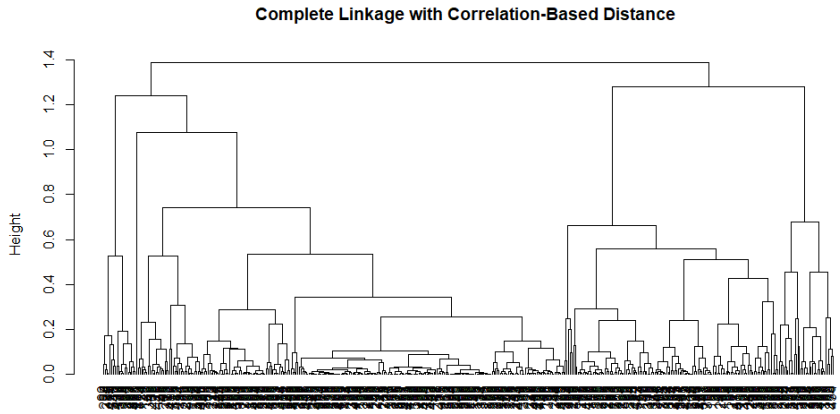

Figure 12: Dendrogram produced using *complete* linkage and *correlation* distance.

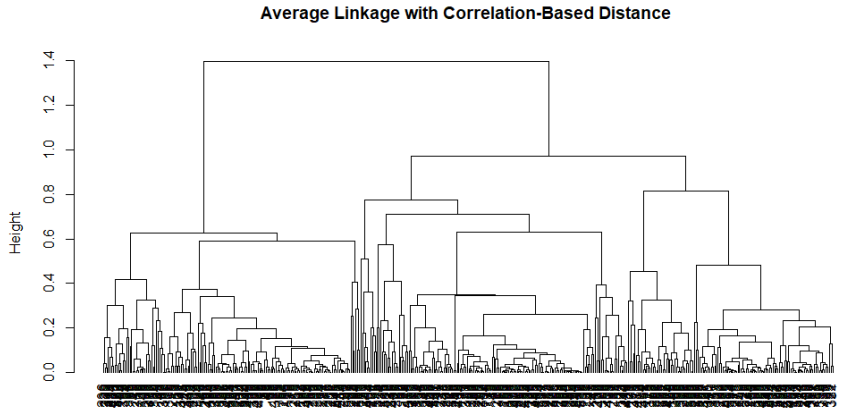

Figure 13: Dendrogram produced using *average* linkage and *correlation* distance.

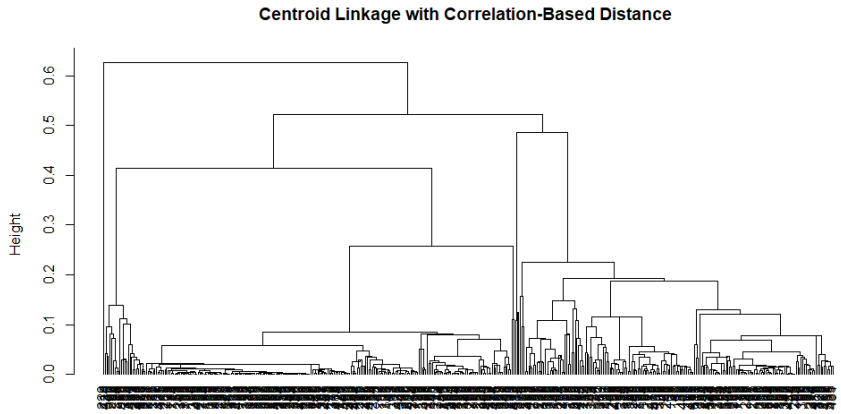

Figure 14: Dendrogram produced using *centroid* linkage and *correlation* distance.

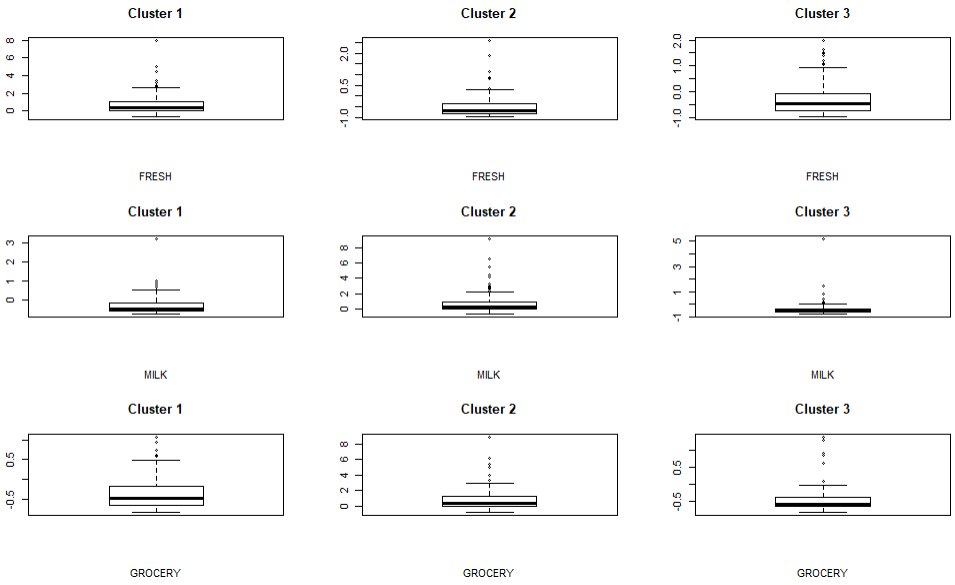

Figure 15: Comparing boxplots for FRESH, MILK and GROCERY for the chosen three clusters.

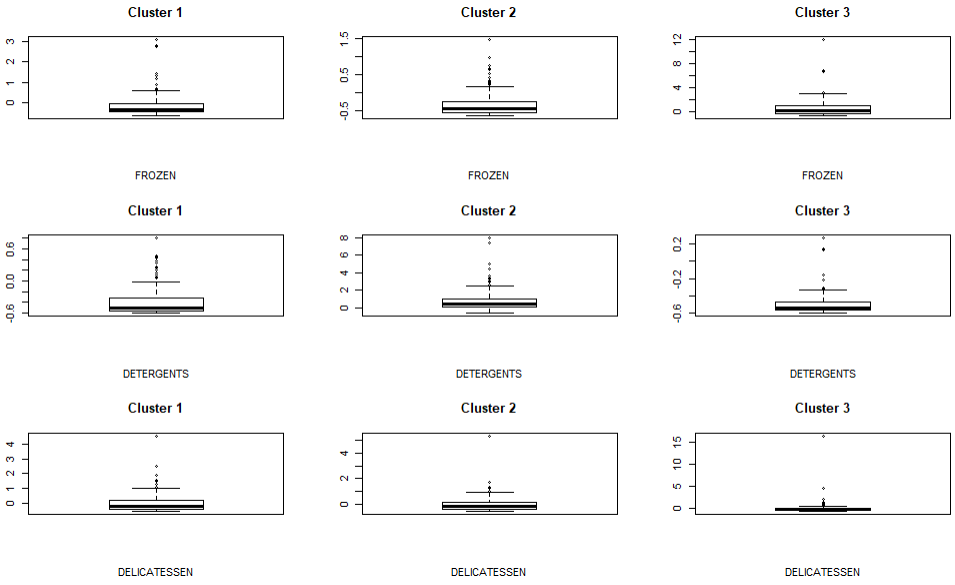

Figure 16: Comparing boxplots for FROZEN, DETERGENTS and DELICATESSEN for the chosen three clusters.

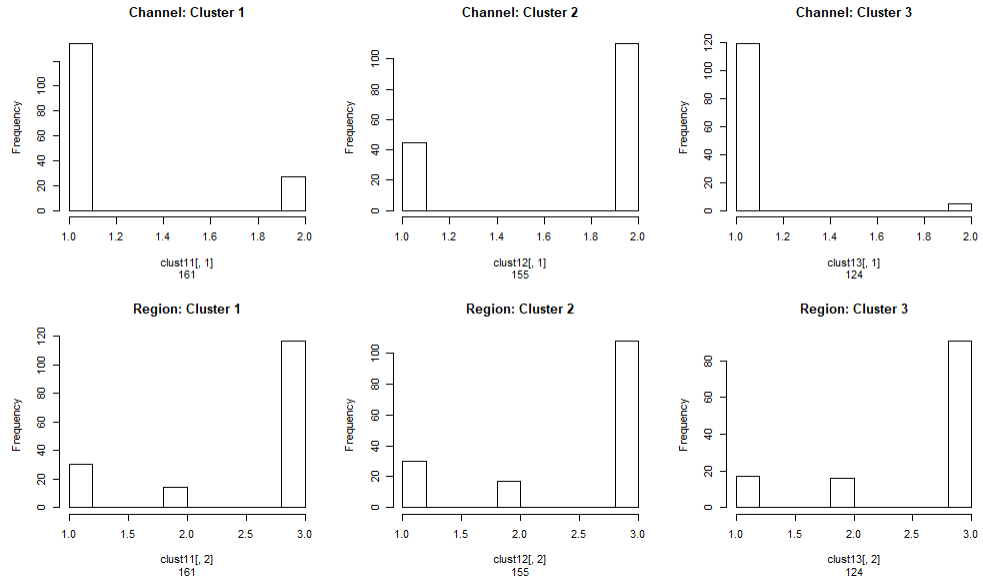

Figure 17: Comparing histograms for CHANNEL and REGION for chosen three clusters.

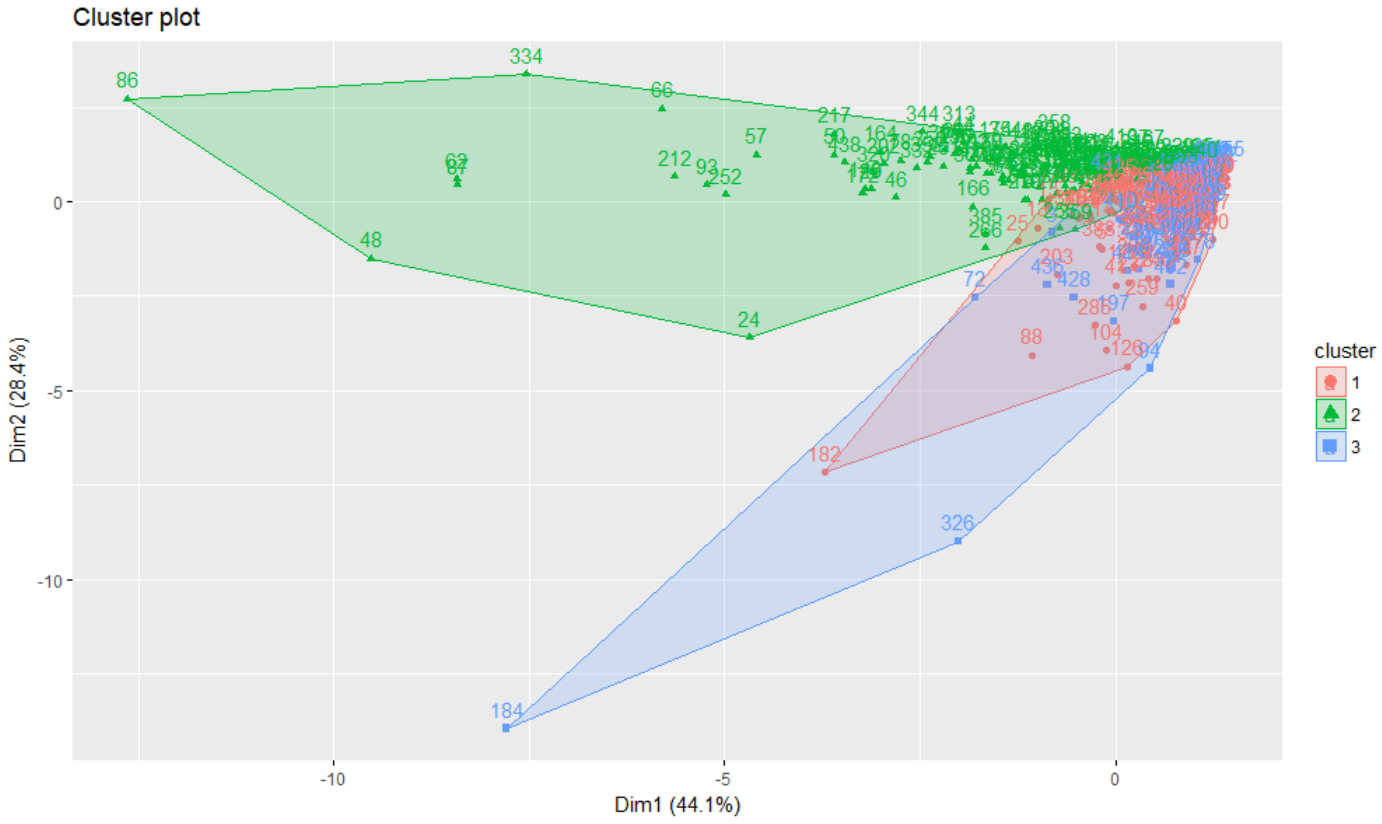

Figure 18: Cluster Plot for the clusters.

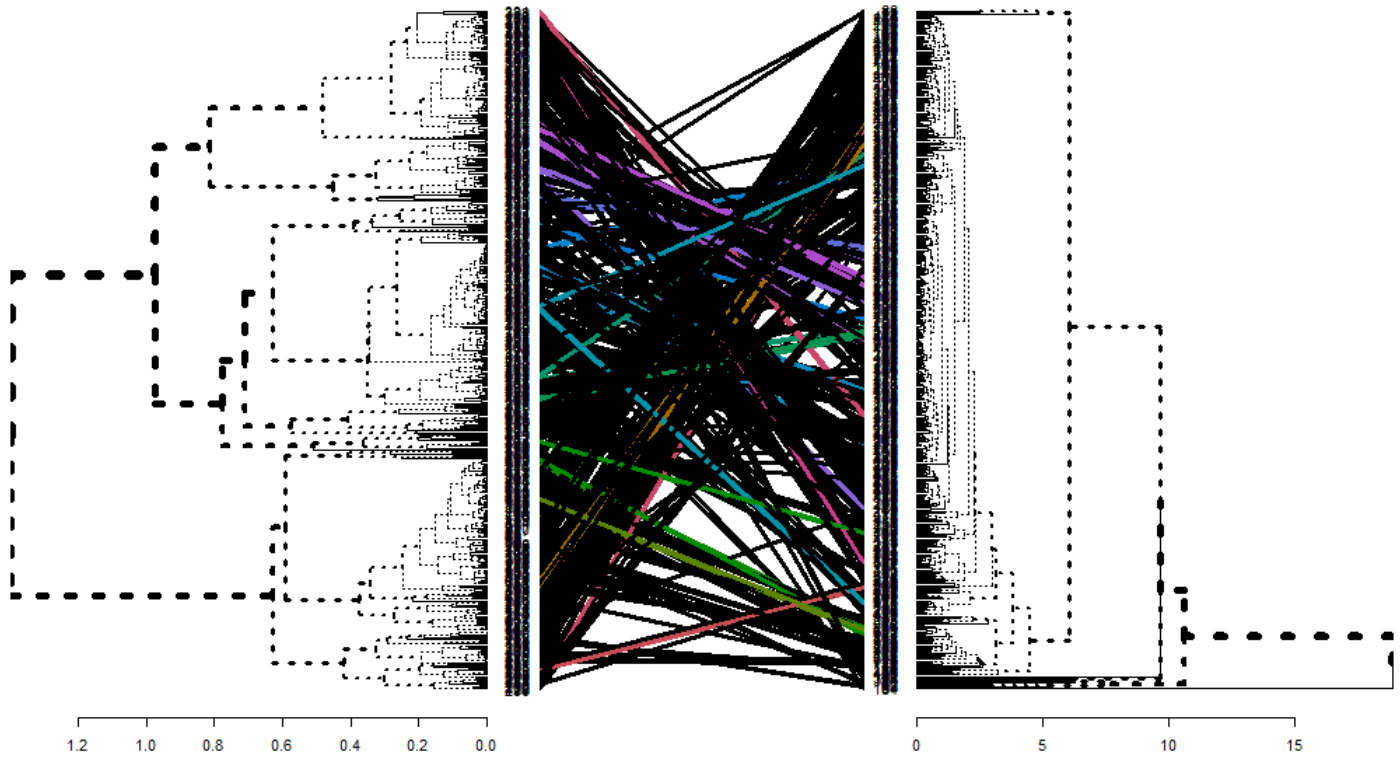

Figure 19: Tanglegram comparing Dendrograms of average linkage with correlation/euclidean distance. Correlation based Dendrogram is on left and euclidean based Dendrogram is on right.

6 Inference and Interpretations

The data shows the revenue of various enterprises who get these supplies from a wholesaler.

Supplies include,

1. FsP: Fresh products (we assume they are grown usual vegetables, fruits, etc perishable food)
2. MP: Milk Products (products from a dairy which include milk and last less)
3. GP: Grocery Products (things commonly found at malls excluding other categories)
4. FzP: Frozen products (Cold drink, ice cream having relatively longer shelf cycle)
5. DP: Delicatessen products (Unusual or foreign food items)
6. DPP: Detergents and paper products (includes soaps, cleaning sols, notebooks, diaries, etc.)

Some explicit observations,

- More number of HoReCa (298) than retail chains (142). Possible reasons,
    1. More no. of "outside consumers" than home-cookers
    2. This wholesaler's service is favourable for HoReCa enterprises
    3. Coincidental, maybe.
- More no. of enterprises in Lisbon (77), followed by Oporto (47). Possible reasons,
    1. More number of restaurants in Lisbon
    2. Coincidental, maybe.
- Fresh Products and DP almost similar over all the enterprises. There is less deviation.
- FsP uniformly spread over the range similar to unskewed distribution while others excluding DP and FsP are right skewed.
- Parallel coordinate plot shows that there is no single entity which is a complete outlier or a major player in every or most segment. There are different products and different entities as revenue leaders.

Since we haven't been given any particular context or a problem but a mere collection of data, any single Dendrogram cannot be ascertained as giving us the all-time correct clusters. As Clusters are groups of observations which can be formed on the basis of any criteria. That criteria decides the relative weight we need to give to the individual parameters.

Find below the inferred meaning of the all the possible Dendrograms obtained from the data.

Euclidean Distance

This method is biased towards the coordinates with higher variation and thus doesn't differentiate between the revenues of two product categories which may be operating at different scales. Ex. We don't buy salt as much as the wheat.

1. **Single Linkage** It is clearly visible to be affected by the chaining effect. It is showing unnecessary clustering.
2. **Complete Linkage** There are no clear clusters. All we can see is that the values go on dividing but with no emerging pattern.
3. **Average Linkage** If we cut the Dendrogram at Hight equal to 12 we get two distinguished clusters – one is of size 1 and the another contains rest of the observations. If we decrease the heights we get more and more clusters but with no significant pattern.
4. **Centroid Linkage** There are slight chaining effects. If we cut at any height between 6 to 12, we get 2 clusters which are exactly same as what we got with average! Cutting the Dendrogram at any lower height doesn't yield us any clear patterned clusters.

Correlation Distance

We try to find clusters which have high similarities and thus higher correlation with elements of the same clusters.

1. **Single Linkage** There is a high prevalence of chaining even more than what happened in the case of single linkage Euclidean.
2. **Complete Linkage** If we cut at height 1.3 we get 2 distinguished clusters. At a height of 1.15-1.2 we get 4 distinguished clusters. This process of increasing number of clusters is gradually increasing with lowering height. The Dendrogram is one of the most comfortable Dendrogram to deal with.
3. **Average Linkage** Upon cutting at a height of 0.9 approx., we get three vivid clusters. And all these three clusters are different from each other. Our hunch is that these three clusters will tell us practical differences between all these enterprises.
4. **Centroid Linkage** If we cut at 0.5 height we get 4 different clusters one of which looks like an outlier cluster. For any subsequent lower heights, there is no clear pattern.

Choose the Correct Number of Clusters

We saw that average was the most appropriate linkage that balanced the polar effects of the single and the complete linkages. Also, there are no distinct clusters for other linkages

(except for complete-correlational linkage). So, we chose average for our linkage method and correlational for our distance measure.

For our post cluster analysis, we have discarded the Euclidean distance because the coordinates with the larger variation and range makes this simple unweighted distance biased towards them and thus overshadows the effects of the smaller but important coordinates. Also, the aim of our analysis is to group buyers with similar buying patterns rather than grouping on the basis of the money they spend.

So, we finally chose three clusters for our post-clustering analysis.

1. **Fresh Product** There are significant differences in the three clusters as the median for the three are 1,0 and -0.5 for the 1st, 2nd and the 3rd cluster respectively.
2. **Milk Product** Here also there are significant differences in the three clusters with the 1st one's median is grossly negative rather the same for 2nd one is around 1 and that for 3rd is around -0.5.
3. **Grocery Products** Except for the outliers, the cluster 1 and 3 have similar median values.
4. **Frozen Products** There is a similarity in the median of the 1st and he 2nd cluster.
5. **Detergents and the paper products** There is a similarity between the median of the 1st and the 3rd clusters, ignoring the outliers.
6. **Delicatessen Product** While the median are similar there are significant differences in the variability and the spread amongst the 3 clusters.

Effects of region and channel over the cluster

- While the 1st and 3rd cluster have similar frequencies of both the channels, the 2nd cluster has a reverse trend of more retail channels than the HoReCa.
- There seem no significant differences in term of regions.

What does Tanglegram tell us?

When the average linkage of Euclidean and correlation distance methods is compared over the Tanglegram it clears comes out that the observations form part of different clusters. It signifies that the two distance methods give us drastically different clusters.

7 Conclusion

In the end, hierarchal clustering methods provided us with good clustering and groups. The algorithm can be further used in more such applications like those of marketing, consumer preference, etc. While this project followed a general approach, we can get more insights and meaningful results post knowing the problem context of the data.

Acknowledgement

We sincerely and deeply thank Prof Sayantan Banarjee for his dedicated and much helpful guidance throughout the project. We hope that he brings to us more such projects and opportunities in future.