Statistics & Machine Learning – 1st Task

Dikaiopoulos Michael - 3180050

The results of the customer grouping of a total of 440 customers using different clustering algorithms (k-means, hierarchical clustering, model based clustering) and the comparison of the results are presented.

Descriptive data analysis

Our data set contains 440 customers (comments) and 8 variables, Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen. The variables Channel and Region are categorical (2 and 3 levels respectively), while the rest are numerical.

The sets of examples for each categorical variable are as follows:

- Channel: #**1** - 298 #**2** - 142
- Region: #**1** - 77 #**2** - 47 #**3** - 316

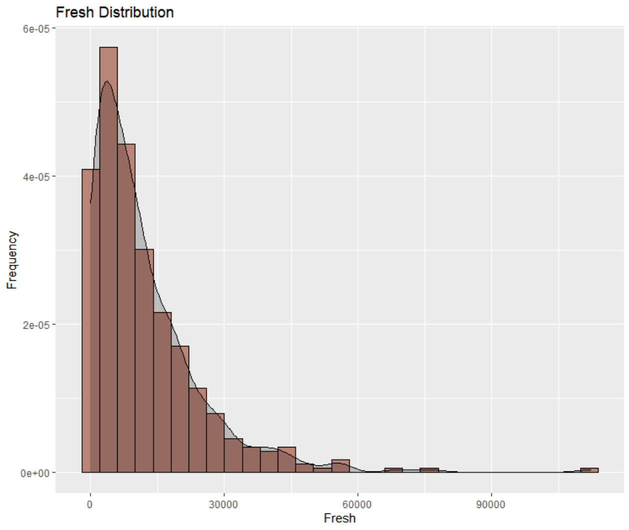
For each numerical variable, its minimum and maximum values, the sum of its values, its median and mean values, its dispersion and its standard deviation are given.

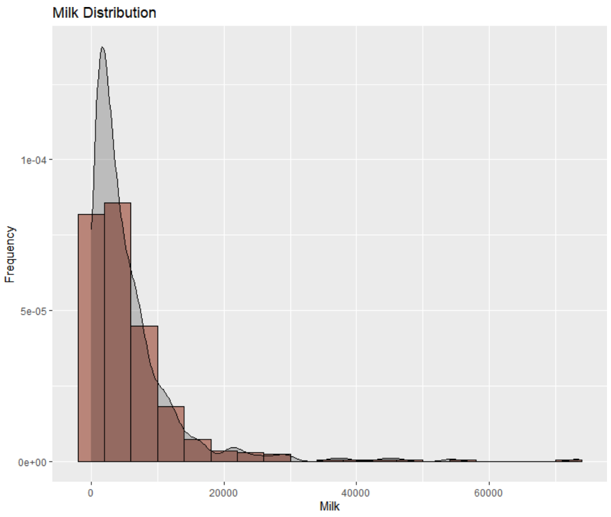
	Fresh	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicassen**
Min	3	55	3	25	3	3
Max	112151	73498	92780	60869	40827	47943
Sum	5280131	2550357	3498562	1351650	1267857	670943
Median	8504	3627	4755.5	1526	816.5	965.5
Mean	12000.3	5796.266	7951.277	3071.932	2881.493	1524.87
Variance	159954927.4	54469967.24	90310103.75	23567853.17	22732436.04	7952997.498
Standard Deviation	12647.33	7380,377	9503.163	4854,673	4767,854	2820,106

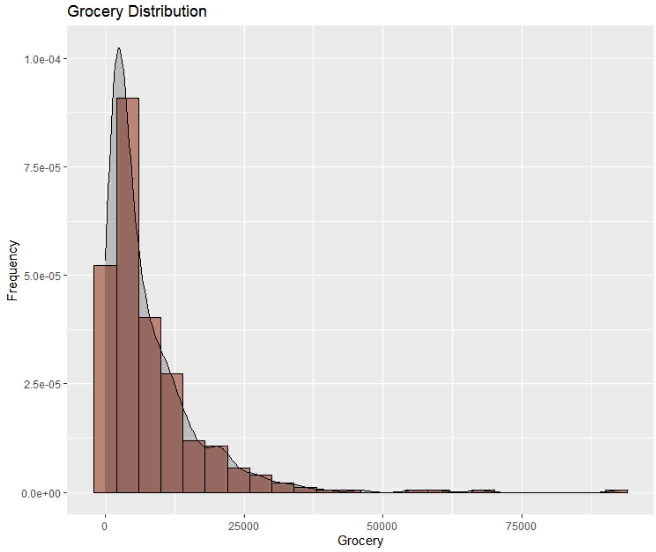
Below is the Pearson correlation table for each pair of variables. Using the practical significance value 0.7 (Chatfield, Collins - 1992), we distinguish two important correlations, the variable **Milk** withthe **Grocery** and the variable **Grocery** with the **Detergents_Paper**.

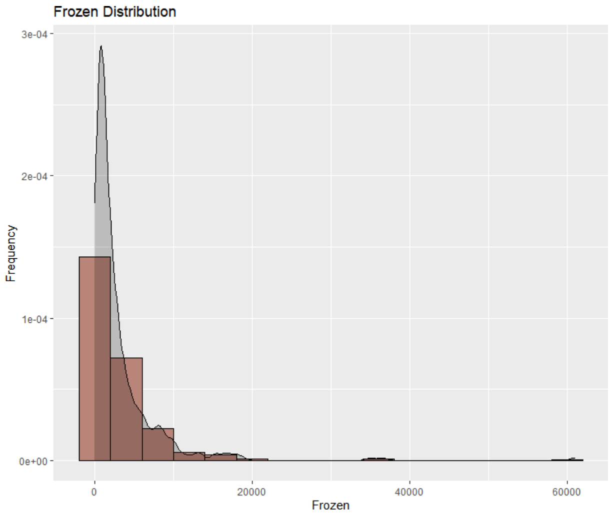
	Fresh	**Milk**	**Grocery**	**Frozen**	**Detergents_Paper**	**Delicassen**
Fresh	1	0.101	- 0.012	0.346	- 0.102	0.245
Milk	0.101	1	0.728	0.124	0.662	0.406
Grocery	- 0.012	0.728	1	- 0.04	0.925	0.205
Frozen	0.346	0.124	- 0.04	1	- 0.132	0.391
Detergents_Paper	- 0.102	0.662	0.925	- 0.132	1	0.069
Delicassen	0.245	0.406	0.205	0.391	0.069	1

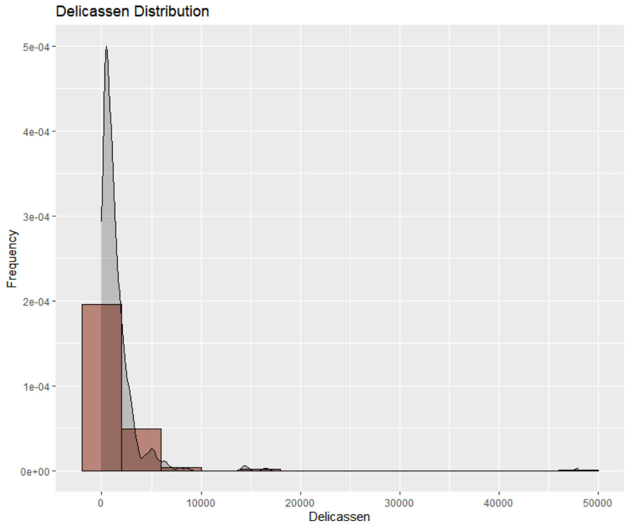
Below we show the histograms along with the distributions of the 6 numerical variables. The different value ranges (with a constant bucket length) and therefore the need to normalize the data for more reliable results are easily understood from the histograms.

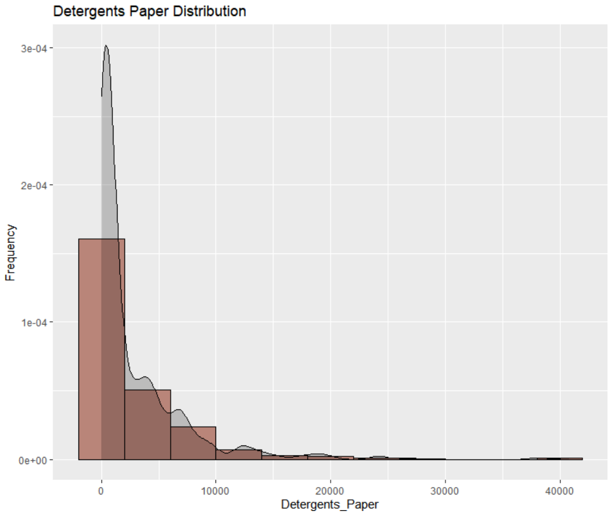












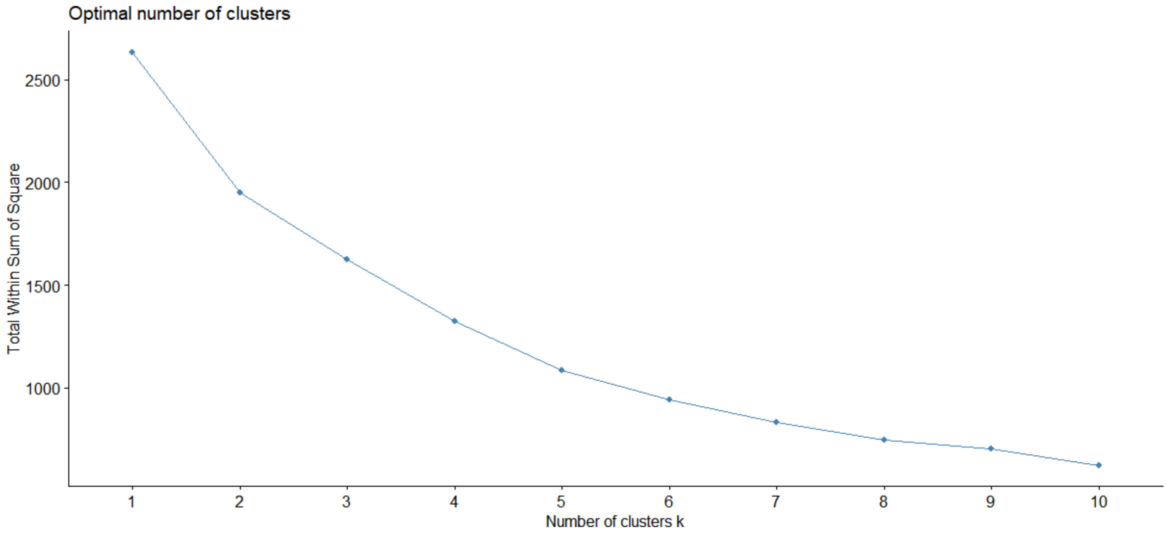
Also, for the analysis of our groupings it will be useful for us to identify the extreme outliers. Since our data are multivariate, the best distance measure is the Mahalanobis distance. We find 3 extreme outliers, lines 87, 184 and 326, with line 184 being an very extreme outlier that may greatly affect our groupings. The scatterplot with Mahalanobis distances is also shown.

Algorithm k-means

We will run the k-means algorithm, first across the dataset and then into the dataset without the extreme values we pointed out earlier. The reason is that we expect k-means to be greatly affected by these extreme values due to its nature (use of average values) and we expect better grouping quality based on the initial categorizations by subtracting the extreme values.

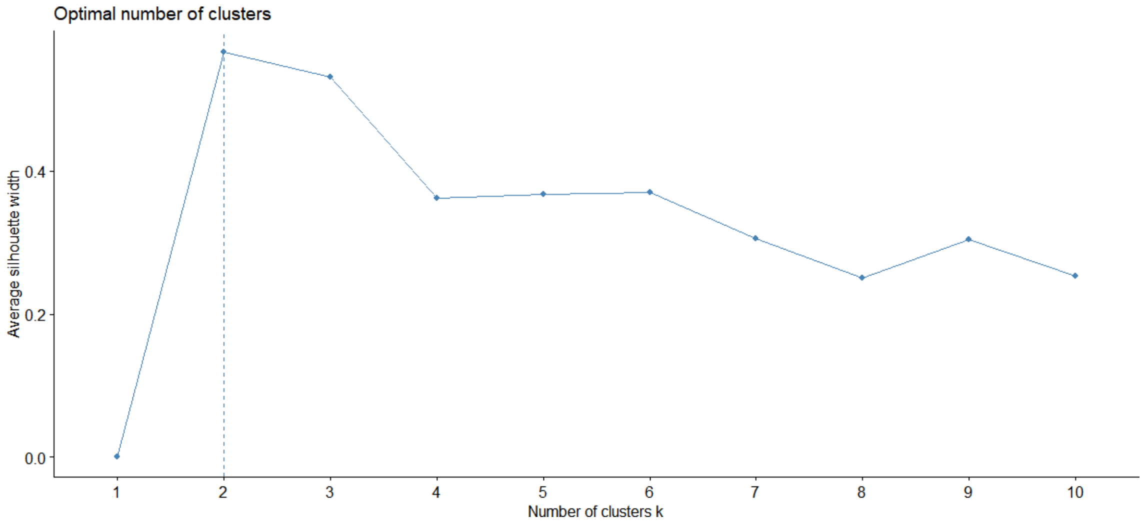
After we scale our data we run the three methods of finding the optimal k.

- 1st Method: Elbow Method



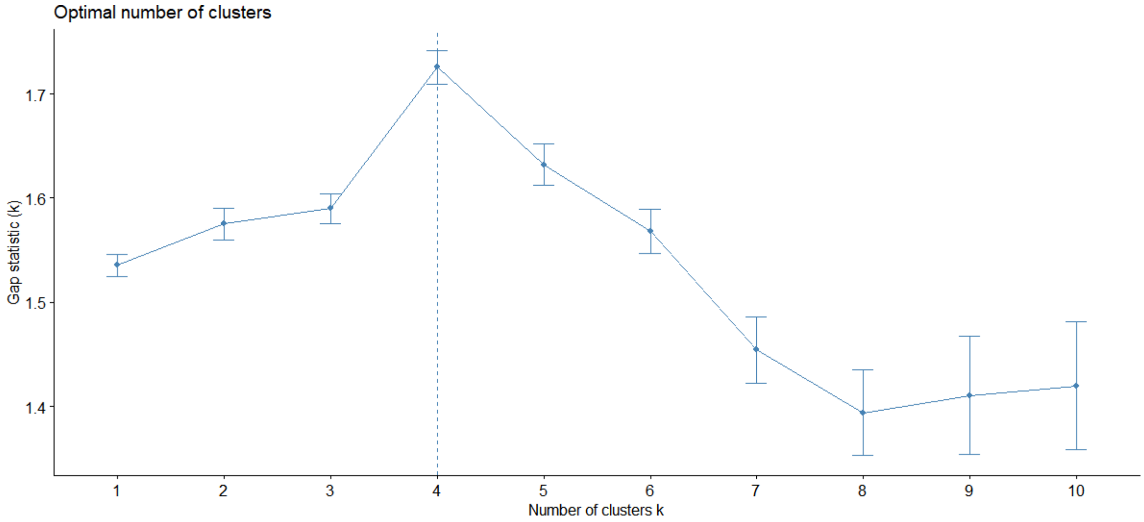
The problem here is that the graph does not look like an elbow, it has a slight curvature and we can not define k with certainty, but we consider that it is between 4 and 5 clusters.

- 2nd Method: Silhouette Method



Silhouette Method defines k = 2 as optimal k.

- 3rd Method: Gap Statistic



Here, k = 4 is defined as optimal k.

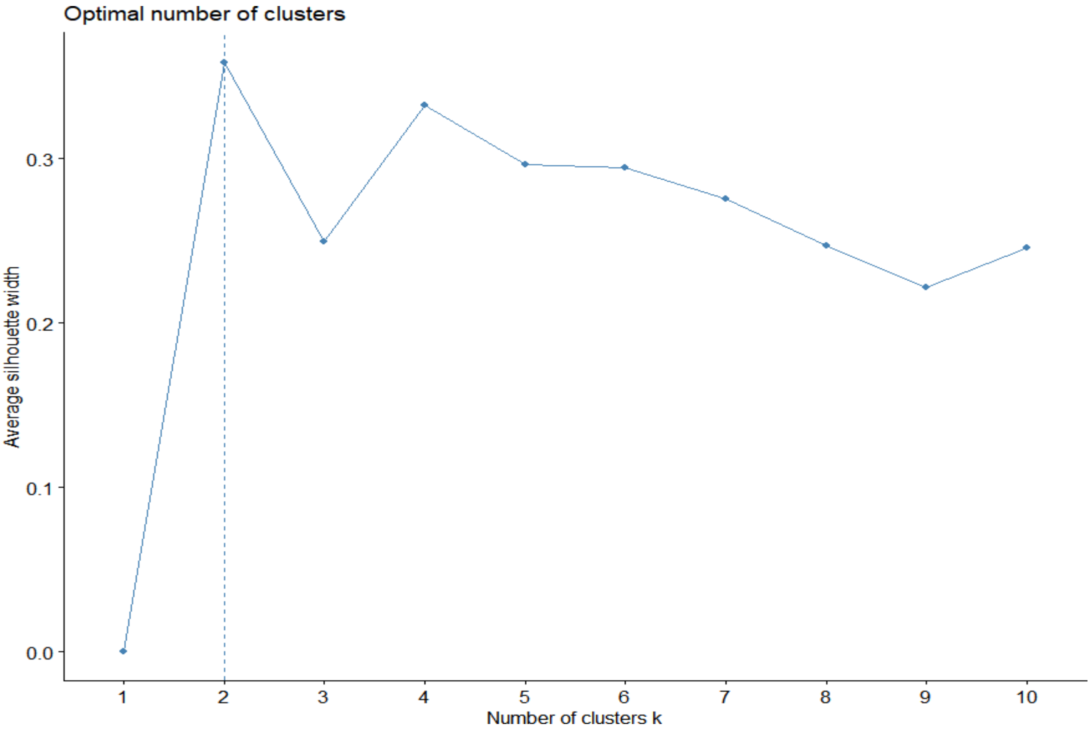
We will define as optimal k based on the results of the three methods k = 3. We already have from the data a strong estimate that they can be grouped in three clusters (based on Region) or in two clusters based on Channel. Also for k = 2 and k = 3 we have the maximum silhouette average coefficient and for k = 3 we have lower wss and higher gap statistic than for k = 2. For k = 4 we have a quite high gap statisti, but based on the average silhouette coefficient we understand that this is a grouping with a weak structure (<0.4)

- **Quality of Clustering (Unsupervised)** : For k = 3, we calculate the average silhouette coefficient as 0.54, defining the structure as marginally interesting. However because we also have 2 variables based on which we can categorize our data into two or three categories, we will also look at the quality of categorization based on these characteristics.

- **Quality of Clustering (Supervised)** : We examine whether there is a strong linear correlation between the groups found by the k-means algorithm and the existing categorization of our data. Regarding the categorization based on the Region (ie if the customers are from Lisbon, Porto or some other area), we compare with the grouping for k = 3, if we want the number of groups to be the same as the number of classes. We find a linear correlation coefficient corr = 0.034, so we do not observe any relation of our grouping with the pre-existing categorization. If we now compare the categorization of thedata based on the Channel and our grouping for k = 2, we find a linear correlation coefficient corr = 0.249,which indicates a better, but relatively weak relationship between them.

Therefore, we can not characterize the quality of grouping as good neither by the criterion of the average silhouette coefficient nor by the criterion of the pre-existing classification.

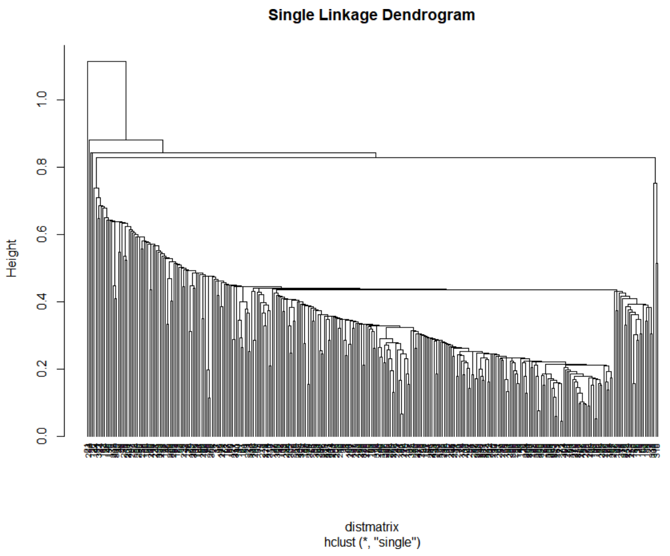
After removing the outliers, follow the same procedure as before. In the Silhouette Method graph, we notice an interesting change:

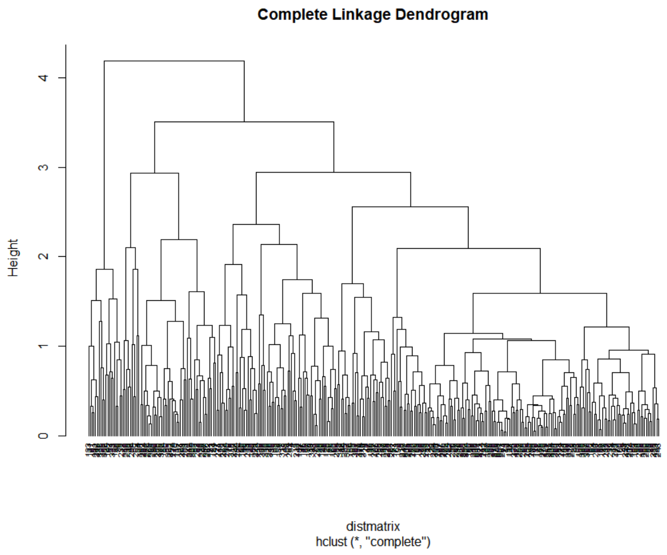


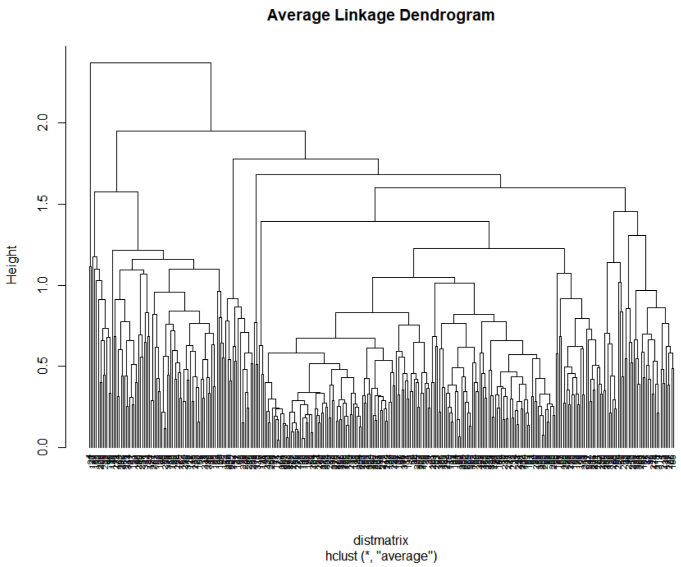
Here the average silhouette coefficient for k = 3 is greatly reduced. We think this is because one of the previous groups consisted of data that had extreme values in our data set. Indeed, if we check the data set of each group in the previous grouping, we will see that one group consisted of three samples, two of which belonged to the outliers we rejected. Having a sense of how the data are categorized based on Region into 3 groups and seeing the almost non-existent correlation we observed between the previous grouping for k = 3 and the categorization based on the Region (0.034), we conclude that the grouping into 3 groups is not ideal and it only gave a high average silhouette coefficient due to the small 3rd team. Now for k = 2, and without the outliers, we have that the average silhouette coefficient is equal to 0.366 while the linear correlation coefficient is equal to 0.613, thus observing a better approach to grouping in classification based on Channel.

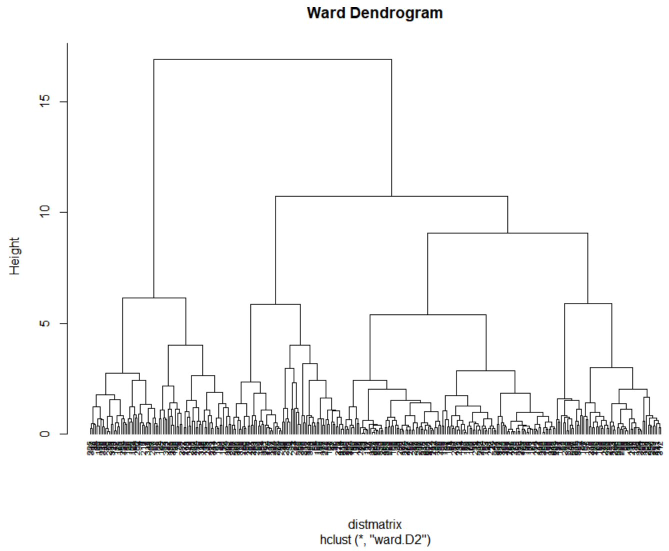
Hierarchical Clustering

We start as before normalizing our data. We perform hierarchical clustering with the methods single linkage,complete linkage, average linkage and Ward.



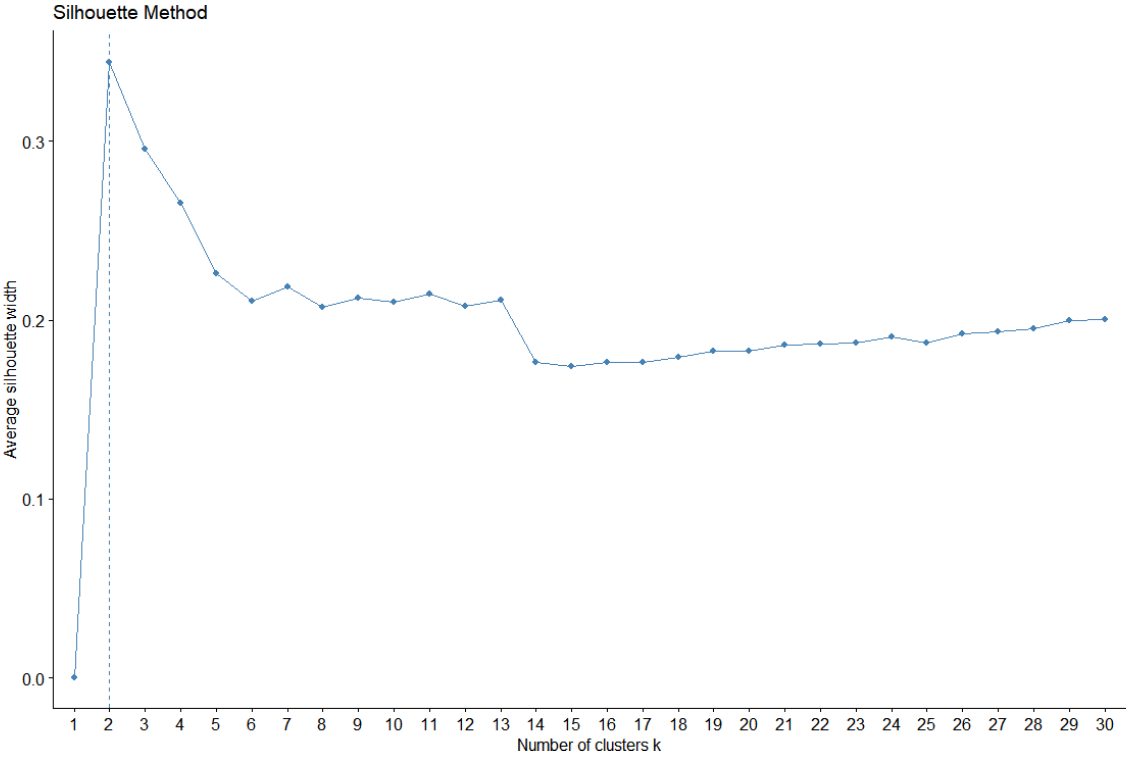






We use the agglomerative coefficient to check the strength of the clustering structure for each method, withthe Ward algorithm showing the strongest structure (coef = 0.979) and the Single Linkage method the loosest (coef = 0.712)

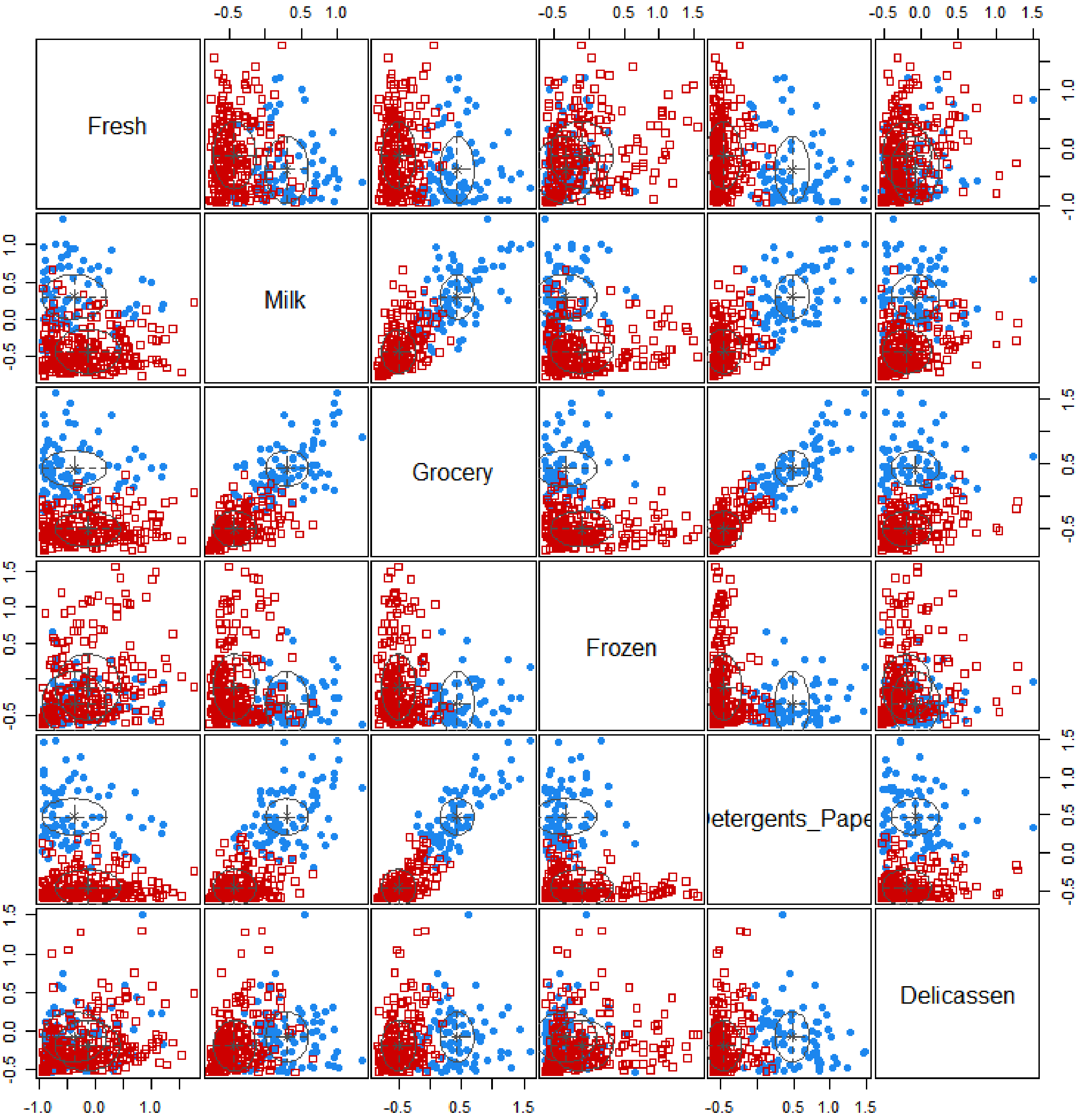
Examining the quality of clustering, initially without supervision, we observer that there is no indication of good clustering for any number of groups. As shown in the diagram below with the average silhouette coefficients, only for k = 2 the coefficient exceeds the barrier of 0.3 (specifically equal to 0.344), defining a weak structure.

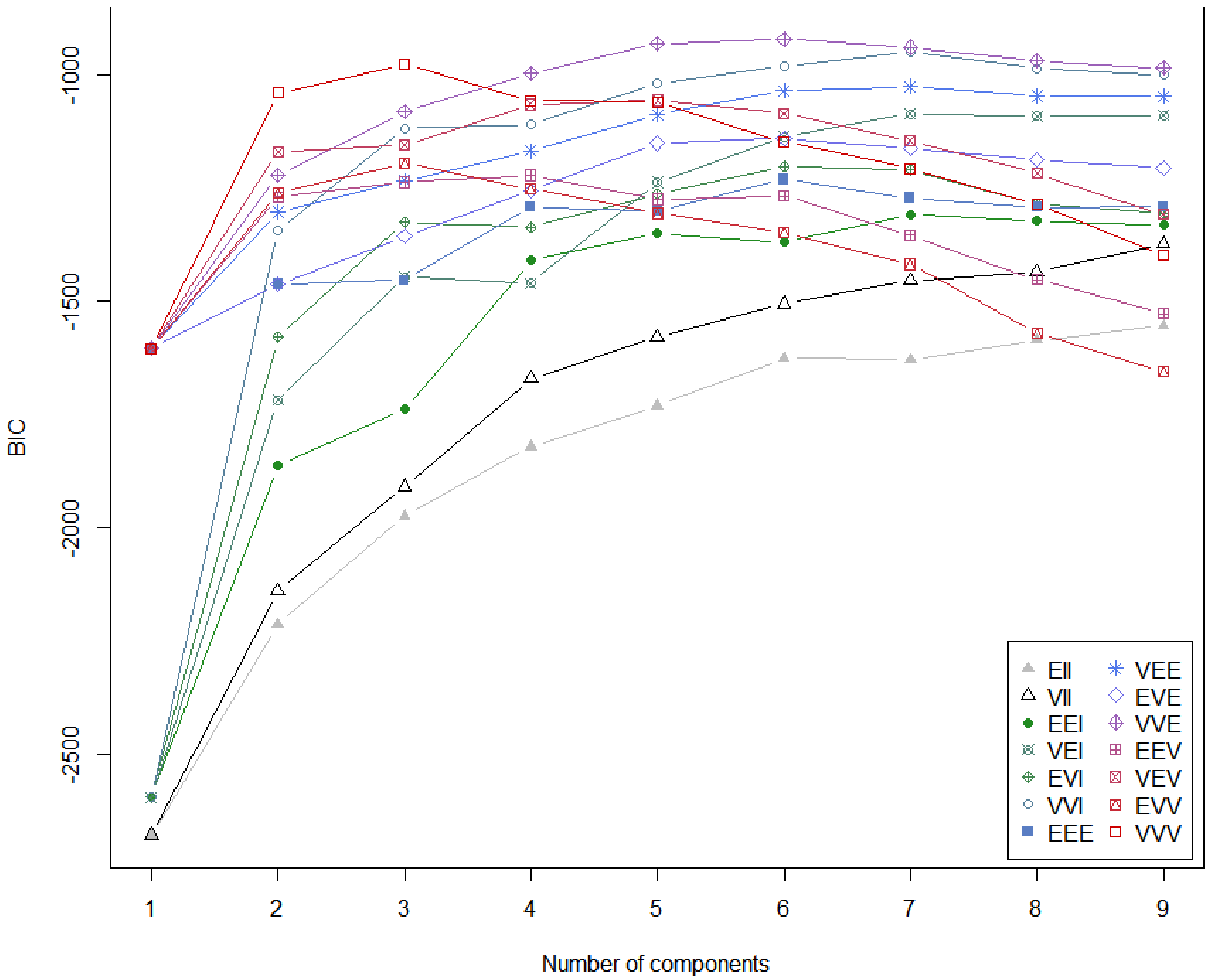


In terms of clustering quality based on Channel and Region classification, the results are again worse than those of the k-means algorithm. Using Ward's method, which achieves the best results in all hierarchical clustering methods, the linear correlation coefficient between 2 clusters and Channel categorization is equa lto 0.521, less than the 0.613 found with k-means. The corresponding coefficient between the clustering in 3 groups and the classification based on the Region is very close to 0 and therefore excludes the existence of a linear relationship between these two.

Model - Based Clustering

After normalizing our data, we perform model-based clustering, checking different models of Gaussian distributions based on the average silhouette coefficient of the model and the linear correlation coefficient of the clustering in 2 groups and the classification based on the Channel. No model shows a linear correlation between 3-group clustering and Region-based categorization. We choose the "EEI" as the optimal model, ie the 2 groups have the same volume and shape, as well as the same direction as the axes of the coordinates. The linear correlation coefficient of this model is equal to 0.652, which is the best linear relationship we have found so far between clustering and categorization based on the Region.





Final Conclusions

The main evaluation criterion we use for the algorithms we mentioned, is the categorization we already knew about our data. As we observed, all three methods, distinguished a clustering in two clusters that tries to approach the categorization based on the categorical Channel variable. In contrast, none of the three methods detects Region-based categorization. The results of k-means and model-based clustering are similar, with both methods having the same power structure and model-based clustering showing a better linear relationship with pre-existing categorization. Based on the above data, we choose Model-Based Clustering as the optimal clustering method.