COMENIUS UNIVERSITY BRATISLAVA

FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

Client Segmentation in Wholesale Markets using Multivariable Data Analysis Methods

Final Project

Thu Nguyen

Bratislava 2020

1. Introduction

1.1. The aim of the project

In today's competitive world, an organization's success largely depends on how much it understands customer behavior. Understanding each customer to better tailor the organization's efforts to their individual needs is a very costly task (1). The main aim of this project is to find if we can meaningfully segment the clients of a business based exclusively on the annual spending habits.

1.2. About the data set

The data set refers to clients of a wholesale distributor. It includes the annual spending habits of 440 clients in monetary units (m.u.) on diverse product categories. This data set has been sourced from the Machine Learning Repository of University of California, Irvine Wholesale Customers Data Set (UC Irvine). The UCI page mentions the original source of the data set is found in: “*Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon”.*

Here are more details on this data set:

- CHANNEL - customers’ Channel: 1 - Horeca (Hotel/Restaurant/Caffe) or 2 - Retail channel (Nominal);
- REGION - customers’ Region: 1 - Lisnon, 2 - Oporto or 3 - Other (Nominal);
- FRESH - annual spending (m.u.) on fresh products (Continuous);
- MILK - annual spending (m.u.) on milk products (Continuous);
- GROCERY - annual spending (m.u.) on grocery products (Continuous);
- FROZEN - annual spending (m.u.) on frozen products (Continuous);
- DET-PAP - annual spending (m.u.) on detergents and paper products (Continuous);
- DELICATESSEN - annual spending (m.u.) on and delicatessen products (Continuous).

2. Data exploration

2.1. The visual Data Analysis

Before delving into the projects, we will begin exploring the data through visualization and how each feature is related to others to get a better understanding of the data. We construct a summary for data (given by figure 1) and scatter matrix (given by figure 2).

```
> wholesale  <- read.table("data.txt", header=TRUE)
> head(wholesale)
  Channel Region Fresh Milk Grocery Frozen Det_Pap Delicatessen
1       2      3 12669 9656    7561    214    2674         1338
2       2      3  7057 9810    9568   1762    3293         1776
3       2      3  6353 8808    7684   2405    3516         7844
4       1      3 13265 1196    4221   6404     507         1788
5       2      3 22615 5410    7198   3915    1777         5185
6       2      3  9413 8259    5126    666    1795         1451
> attach(wholesale)
> wholesale$Channel = as.factor(wholesale$Channel)
> wholesale$Region = as.factor(wholesale$Region)
> summary(wholesale)
 Channel Region     Fresh             Milk           Grocery          Frozen           Det_Pap          Delicatessen
 1:298   1: 77   Min.   :     3   Min.   :   55   Min.   :    3   Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
 2:142   2: 47   1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153   1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
         3:316   Median :  8504   Median : 3627   Median : 4756   Median : 1526.0   Median :  816.5   Median :  965.5
                 Mean   : 12000   Mean   : 5796   Mean   : 7951   Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
                 3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656   3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
                 Max.   :112151   Max.   :73498   Max.   :92780   Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
> |
```

Figure 1: A summary about wholesale data

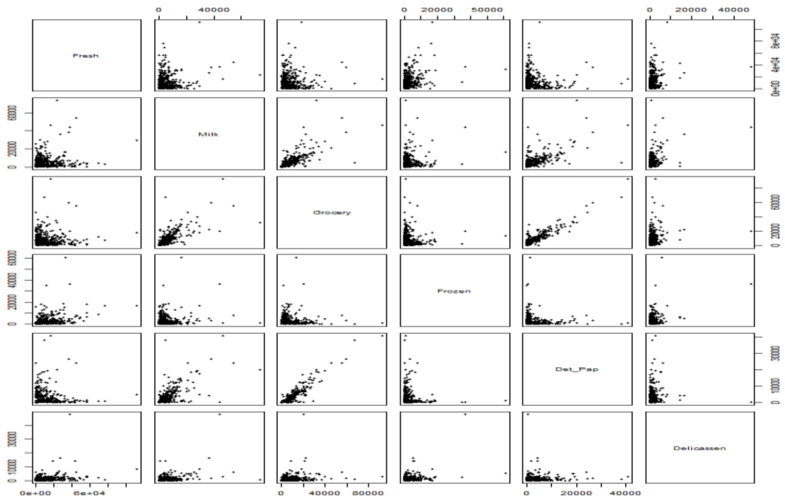

Figure 2: Scatter matrix

Some preliminary assessments of the data:

1. The dataset has 440 observations and there are 8 columns, 6 continuous variables (the last six variables) and two descrete variables (Channel and Region). While 6 continuous record goods that were brought by distributions from the wholesaler, the two descrete variables are factors representing the location and channel of purchase.

For the purpose of this project, the features “Channel” and “Region” will be excluded in the analysis – with a focus instead of on the six product categories recorded for customers.

2. There is a very strong linear correlation between the ‘Grocery’ & ‘Det_Pap’ features; and a correlation between ‘Milk’ & ‘Grocery’, ‘Milk’ & ‘Det_Pap’ as well. We wish our data are independent to implement clustering model.
3. We can also see that the data is not normally distributed due to: There are many outliers, and most of the data points lie on the left, the features for this data are heavily skewed. We need the data features to be normally distributed as clustering algorithms required.
4. To get a deeper understanding of how much normalized of our data, we quickly use some simple tests for normality such as: Kolmogorov-Smirnov test, Shapiro, and Q-Q plot. Tests reinforce the conclusion that our data is not normality.

2.2. Data Preprocessing

We will preprocess the data to ensure that our obtained results are significant and meaningful.

1. We will check the missing value of observations in our dataset. If there are missing values, then we should be removed from those instances. From the result (see figure 3), we see that our data does not have any missing values.
2. Remove two features Channel and Region.
3. We need our data is normality, so the simplest way which can work in most cases would be applying the natural logarithm. After using the natural logarithm scale to the data, the distribution of each function should appear much more normality (see figure 4).
4. Standardize the variables by using the function scale().

```
> #Data Preprocessing###
> print(apply(wholesale, 2, function (x) sum(is.na(x)))) #Check missing values
     Channel       Region        Fresh         Milk      Grocery       Frozen  Det_Pap Delicatessen
           0            0            0            0            0            0        0            0
```

Figure 3: Check missing values

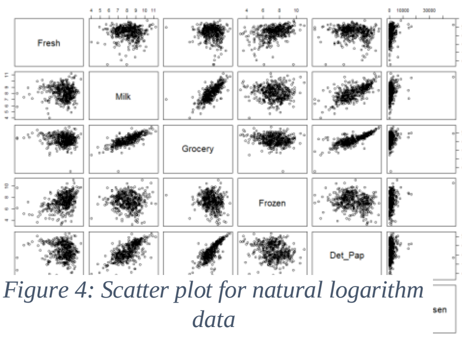

Figure 4: Scatter plot for natural logarithm data

3. Dimensionality reduction using PCA

To understand the data more deeply, I employed the PAC analysis. Principal component analysis (PCA) is one of the simplest multivariable data analysis methods. The main goal of PCA is to reduce the dimensionality of the data - in effect, reducing the computational cost on the original data set. By using PCA, we can create a new space with less dimension but can represent data as well as the old space. Although some theoretical results about principal components assume normality, it is routinely applied also to non-normal data (2).
After applying PCA for the preprocessed model (good_data), we obtained some results shown as the following figures.

```
> PCA <- prcomp(good_data, scale = TRUE)
> print(PCA)
Standard deviations (1, .., p=6):
[1] 1.6246398 1.2757977 0.8033712 0.7801817 0.5425900 0.4294099

Rotation (n x k) = (6 x 6):
                  PC1          PC2         PC3         PC4         PC5          PC6
Fresh      -0.1046266  0.590473852 -0.63189359  0.48852525 -0.04115995 -0.027446847
Milk        0.5422741  0.133145373 -0.07607607 -0.06138591  0.76163365  0.313957708
Grocery     0.5716940 -0.006282363 -0.13344981 -0.09567158 -0.09810121 -0.797835027
Frozen     -0.1383505  0.589534934 -0.03363014 -0.79160890 -0.07414434  0.005888702
Det_Pap     0.5513378 -0.068624237 -0.19725842 -0.07734734 -0.61832599  0.513903505
Delicassen  0.2122351  0.530389241  0.73285210  0.34028597 -0.14412416  0.002237943
> summary(PCA)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.6246 1.2758 0.8034 0.7802 0.54259 0.42941
Proportion of Variance 0.4399 0.2713 0.1076 0.1014 0.04907 0.03073
Cumulative Proportion  0.4399 0.7112 0.8187 0.9202 0.96927 1.00000
> |
```

Figure 6: A summary for PCA

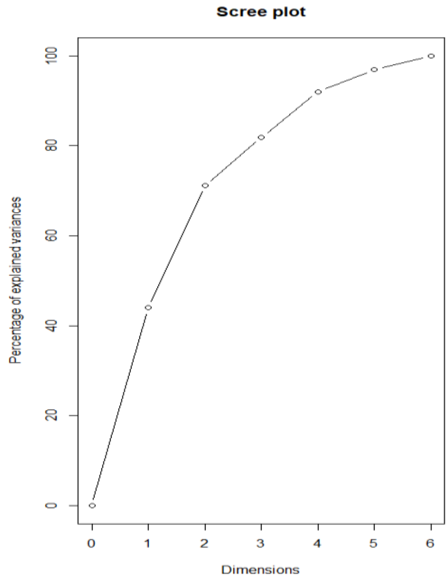

Figure 5: Scree plot

Scree plot (figure 5) used to determine principal components to keep in a principal component analysis (PCA). The components need to describe at least 80% of the cumulative percentage of variance (2). In this case, the three components 1, 2, and 3 account up to 81,87% of the cumulative variance, and 92,02%, is explained by the first four principal components.
From the results obtained on figure 6, three components PC1, PC2, and PC3 are built based on interaction with six goods and shown in equations 1, 2, and 3

$$PC1 = -0.1\,Fresh + 0.54Milk + 0.57Grocery - 0.14Frozen + 0.55Det\_Pap + 0.21Delicassen \text{ (equation 1)}$$

$$PC2 = 0.59\,Fresh + 0.13Milk - 0.06Grocery + 0.58Frozen - 0.07Det\_Pap + 0.53Delicassen \text{ (equation 2)}$$

$$PC3 = -0.63\,Fresh - 0.07Milk - 0.13Grocery - 0.03Frozen - 0.2Det\_Pap + 0.73Delicassen \text{ (equation 3)}$$

We can see that Milk, Grocery, Dete_Pap have similar coefficients; also, Fresh and Frozen have similar coefficients. This table below discussed what kind of establishment could each of the first four principal components represent and which features primarily effect to each component

	Primarily	Establishment
PC1	(+) Det_Pap, Milk, Grocery	Retailers
PC2	(+) Fresh, Frozen, Delicatessen	Restaurant or Cafee
PC3	(+) Delicassen (-) Fresh	US convenience store
PC4	(+) Fresh, Delicassen (-) Frozen	Gas station shop

The 2D-plot and 3D-plot are shown in figure 7, show the projection of the original features along with the components with the axes are the principal components.

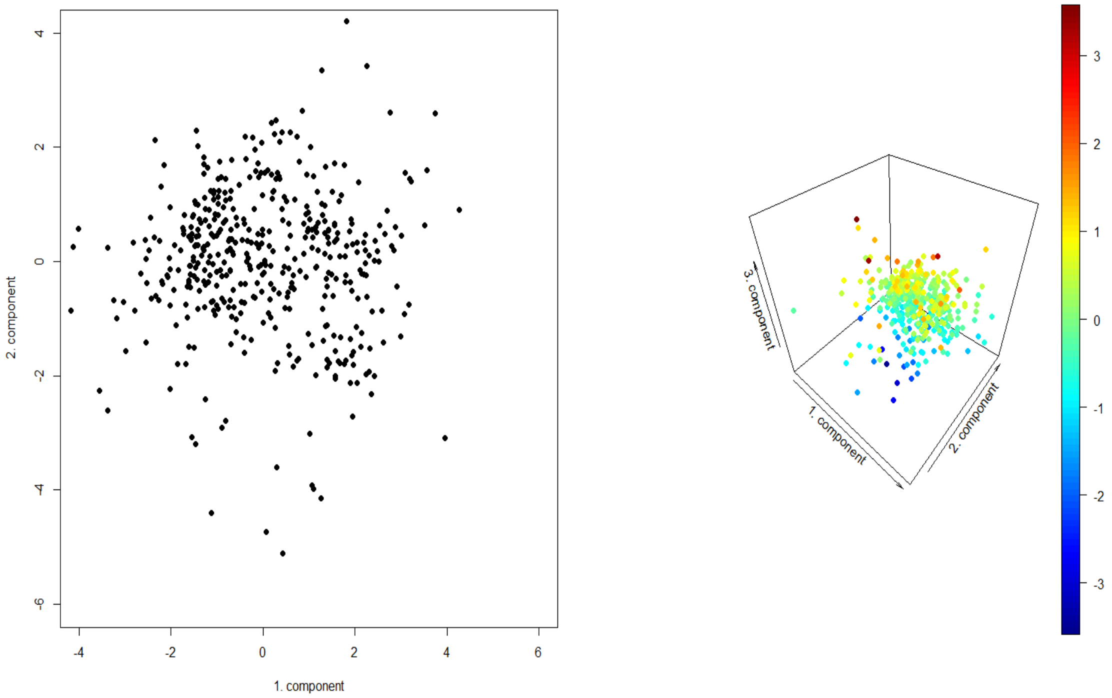

Figure 7: 2D- and 3D- plot

4. Clustering

4.1. K-means clustering algorithm

The k-means algorithm is probably the most popular and commonly used method of the partitioning methods of cluster analysis. In this project, I will choose to use the k-means clustering algorithm because it works well in practice and easy to understand and implement (3).

We do not know ahead of time how many clusters to be made with our data. To answer the value of k-clusters, sometimes, we can go to a business and ask them how many clusters they would expect in the data (1). In addition, there are several methods available to decide on this k-value. In R, the Nbclust function, found in the Nbclust package, is powerful to determine the optimal clusters. After implementing Nbclust, we come up with the best number of clusters is **k = 2.** Figure 8 shows the plot of the D index. The D index is a graphical method of determining the number of clusters. In the plot of D index, we seek a significant knee (the significant peak in D-index second differences plot) that corresponds to a significant increase in the value of the measure (4).

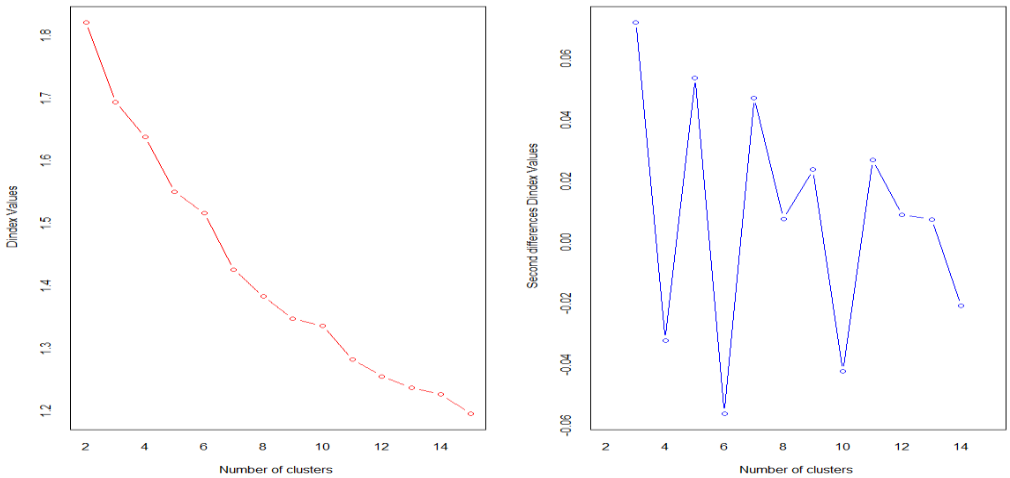

Figure 8: Total within-cluster sum of squares with euclidean distance

```
> kmeans_plot <- data.frame(good_data,  wholesale.kmeans$cluster)
> clusplot(kmeans_plot, wholesale.kmeans$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
> wholesale.kmeans
K-means clustering with 2 clusters of sizes 252, 188

Cluster means:
       Fresh       Milk    Grocery     Frozen    Det_Pap Delicatessen
1  0.2235330 -0.5960391 -0.6348117  0.2678315 -0.6632111   -0.1698832
2 -0.2996294  0.7989460  0.8509178 -0.3590082  0.8889851    0.2277158

Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  2
  2   2   2   1   2   2   2   2   2   2   2   1   2   2   2   1   2   1   2   2   2   1   1   2   2   2
 62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  8
  2   2   2   1   2   2   2   2   1   1   2   1   1   2   1   1   2   1   2   1   2   2   1   2   2   2
123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 14
  1   2   1   1   1   2   2   1   1   1   1   1   1   1   2   2   1   1   2   1   1   1   1   2   1   1
184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 21
  2   1   1   1   2   2   2   1   1   1   2   1   1   1   2   1   1   2   2   2   1   1   2   1   2   2
245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 27
  2   2   1   1   1   1   1   2   1   2   2   1   1   1   1   1   1   1   1   1   2   2   2   1   2   1
306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 33
  2   2   1   1   2   1   1   2   1   1   2   1   2   1   2   1   1   1   1   1   2   1   1   1   1   1
367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 39
  1   1   1   1   1   1   1   2   1   1   2   1   1   2   1   1   2   1   2   1   1   1   1   1   1   1
428 429 430 431 432 433 434 435 436 437 438 439 440
  1   2   1   2   1   1   1   2   1   1   2   1   1

Within cluster sum of squares by cluster:
[1]  826.1059 1012.4726
 (between_SS / total_SS =  30.2 %)
```

Figure 9: k-means clusstering with 2 clusters of sizes

```
> Channel
  [1] 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 1 2 1 2 1 2 1 1 2 2 2 1 1 2 1 1 1 1 1 1 2 1 2 2 1 1 1 2 2 2 2 2 2 2 2 1 1 2 2 1 1 2 2 1 1 2 2 2 2 1 2 1 2 1
 [70] 1 1 1 1 2 2 1 1 2 1 1 1 2 2 1 2 2 2 1 1 1 1 1 2 1 2 1 2 1 1 1 2 2 2 1 1 1 2 2 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1
[139] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 2 1 2 2 2 1 1 2 2 2 2 1 1 1 2 2 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 1 1 1 2 1 1 2 2 1 1 1 2 1
[208] 2 1 2 1 2 1 1 2 1 2 1 2 1 1 1 1 2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 1 1 1 1 1 1
[277] 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 2 1 2 2 2 2 2 2 2 1 1 2 1 1 2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 1 1 1 1 2 2 1 2 1
[346] 1 2 2 1 2 1 2 1 2 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1
[415] 1 2 2 1 2 1 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1
> 
```

Figure 10: Feature Channel (1 ~ Horeca; 2 ~ Retail)

To have a better view of the clustering results, we can base on the first two principal components visualize observations (see figure 11).
Interpretation: The analysis explains 71.84% of the multivariate data.

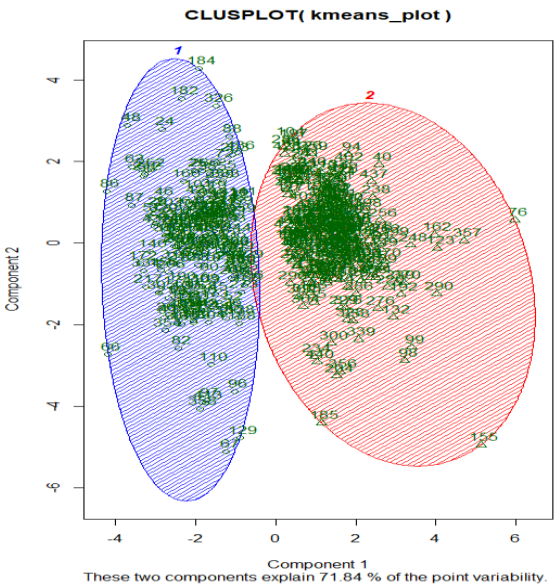

Figure 11: 2D representation of the Cluster Solution

Comment 1: From Cluster means table, we can see

- 1. Cluster spends at an above-average rate on items like Milk, Grocery, and Det_Pap, indicating that it might represent grocery stores. When compared with feature Channel, 1. cluster has similar observations as Retail (with the value indicated by 1)
- 2. Cluster spends majority in the Fresh and Frozen category, very lowly spends at Grocery, Det_Pap, and Milk, indicating that it might represent restaurants or seafood or meat market. When compared with feature Channel, 2. cluster has similar observations as Horeca (Hotel/Restaurant/Café/ with the value indicated by 2).

The purpose of this project was to find if we could meaningfully segment the clients of a business based exclusively on the annual spending habits. The 'Channel' and 'Region' features did not have any related to the spending habits, so we excluded from the dataset. Our project discovers that base on the spending habits; clients could be divided into meaningful groups (customer segments). Those groups would be purely classified as "Retailers" or "Hotels, Restaurants, Cafes". The resulting clusters can be believable and have practical significance. Based on these results, businesses can devise business strategies that target specific customer segments.

Comment 2: If the data had not been preprocessed, by a brief analysis, the result would have 3 clusters: the first cluster has 330 observations in it, and the second and third clusters are small with just 50 and 60 observations. The results are more meaningful with the data preprocessed.

4.2. Hierarchical clustering algorithms – Divisive clustering

Divisive hierarchical clustering, also known as Divisive ANAlysis (DIANA) clustering algorithm that follows a top-down approach to identify clusters in a given dataset (1). After applying the function diana(), we have a visualization of results as shown in figure 13.
In the dendrogram output, the higher level of the fusion indicates the similarity between the two observations are. When compared to k-means, we want to cut the dendrogram at 2. level, which means the number of clusters obtained is 2. The result after cutting can visualize as figure 13. We can see the result received is similar to the k-means algorithm (figure 11).

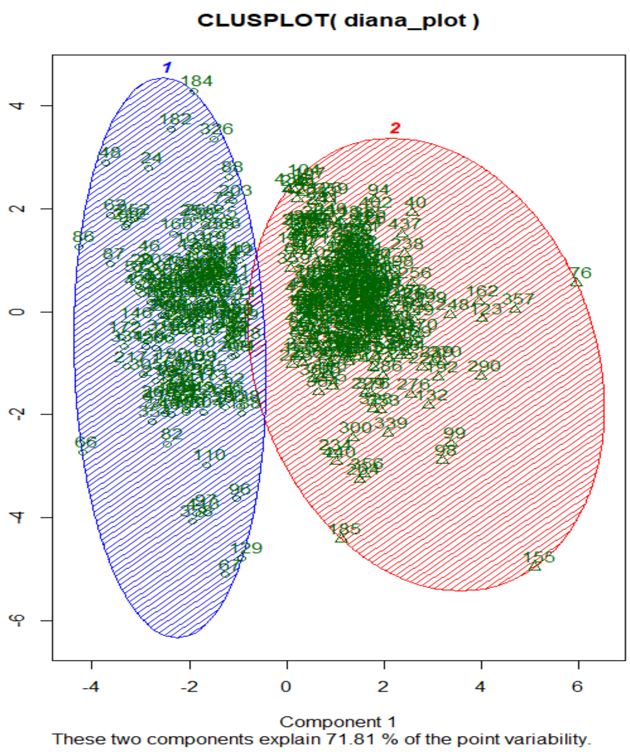

Figure 13: 2D Cluster plot

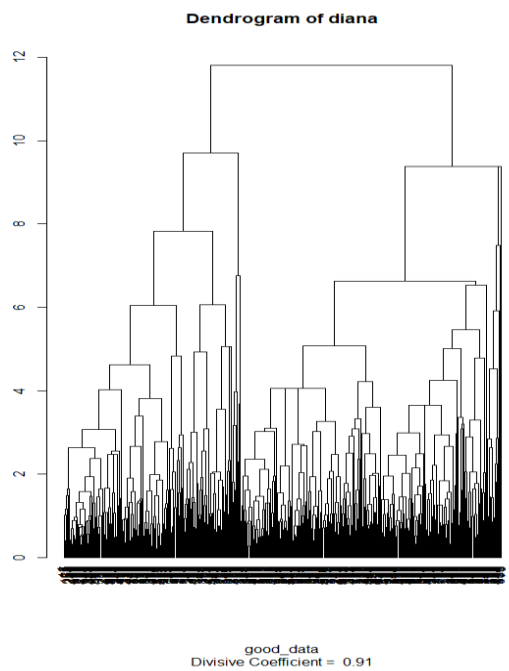

Figure 12: Cluster Dendrogram

5. Classification trees

Classification trees are non-parametric methods to recursively partition the data into more "pure" nodes, based on splitting rules. The purpose of this dataset is to predict which customers are Horeca or Retails.

$$Channel \sim Fresh + Milk + Grocery + Frozen + Det\_Pap + Delicassen$$

The dataset is ordered by the variable Channel. Before training the model, we have to split the dataset into the training (wholesale.train) and testing (wholesale.test) dataset; each dataset has 220 rows. To create our classification tree, we'll be using the rpart() function. We get the following tree:

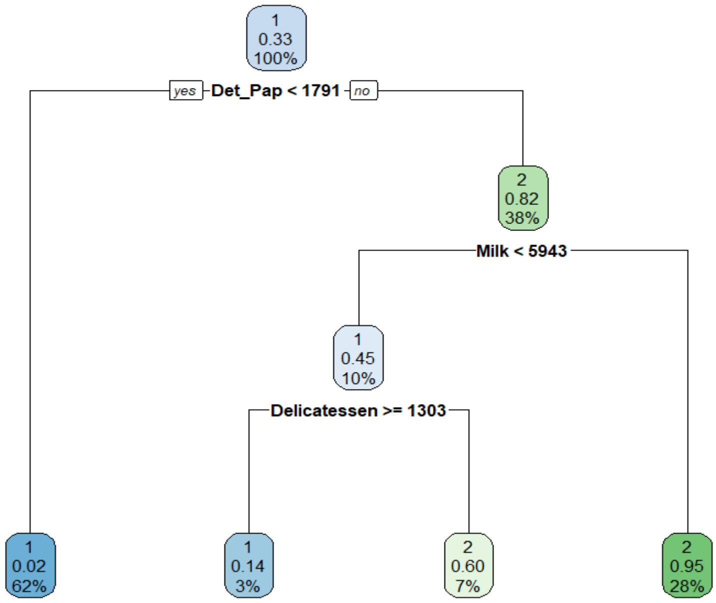

Figure 14: Classification trees.

The prediction and accuracy of the model is then estimated. The accuracy of the model is now 86.81%.

```
> table(test, wholesale.test[, 1])

test   1   2
   1 138  14
   2  12  56
> 1 - mean(test == wholesale.test[, 1])
[1] 0.1181818
```

6. Conclusion

In this project, we used a wholesale dataset available from the UCI repository and implemented clustering using the Divisive ANAlysis (DIANA) and k-means method. During this project, we also studied various aspects related to clustering, such as Principal Component Analysis (PCA) and methods for identifying the correct number of clusters. We also explored the use of classification trees to predict customer's channels. The aim of the project was to examine whether the clients of a business could be divided into meaningful groups based exclusively on spending habits. It has been discovered that clients' spending habits can indeed be used to make such discoveries and that it does so quite well.

Bibliography

1. **Cory, Lesmeister and Sunil, Kumar Chinnamgari.** *Advanced Machine Learning with R.* Birmingham : Packt Publishing Ltd., 2019. ISBN 978-1-83864-177-1.
2. **Radoslav, Harman.** *Multivariate Statistical Analysis.* Bratislava : http://www.iam.fmph.uniba.sk/ospm/Harman/VSAp.pdf, 2020.
3. *Identifying Customer Segments (Unsupervised Learning).* **Ng, Ritchie.** s.l. : https://www.ritchieng.com/machine-learning-project-customer-segments/, 2017.
4. *K-means Clustering Method.* **Rodda, Ouma.** s.l. : https://rpubs.com/oumarodd/411716, August 4, 2018.
5. *Creating Customer Segments from Spending Habits.* **Ronny, Restrepo.** s.l. : http://ronny.rest/portfolio/mlnd_cust_seg.
6. *WHOLESALE CUSTOMERS DATASET.* **Sriraag, Av.** s.l. : https://medium.com/@sa2253/wholesale-customers-dataset-65c58c692c6a.