Customer segmentation - Dimensionality Reduction and Clustering Techniques

Student: Emanuele Morales

The dataset refers to the clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories. The products considered for this projects are: Fresh, Milk, Grocery, Frozen, Detergents Paper, Delicatessen.
This project is divided into three main parts: Data exploration and Preprocessing, Dimensionality reduction with Multidimensional scaling and PCA, Clustering Analysis with K-means and Model-Based Techniques.

1 Data exploration and Preprocessing

The distribution of the data can be analysed from the following matrix of plots, where the distribution of the variables, bivariate scatter plots and the value of correlation are reported.

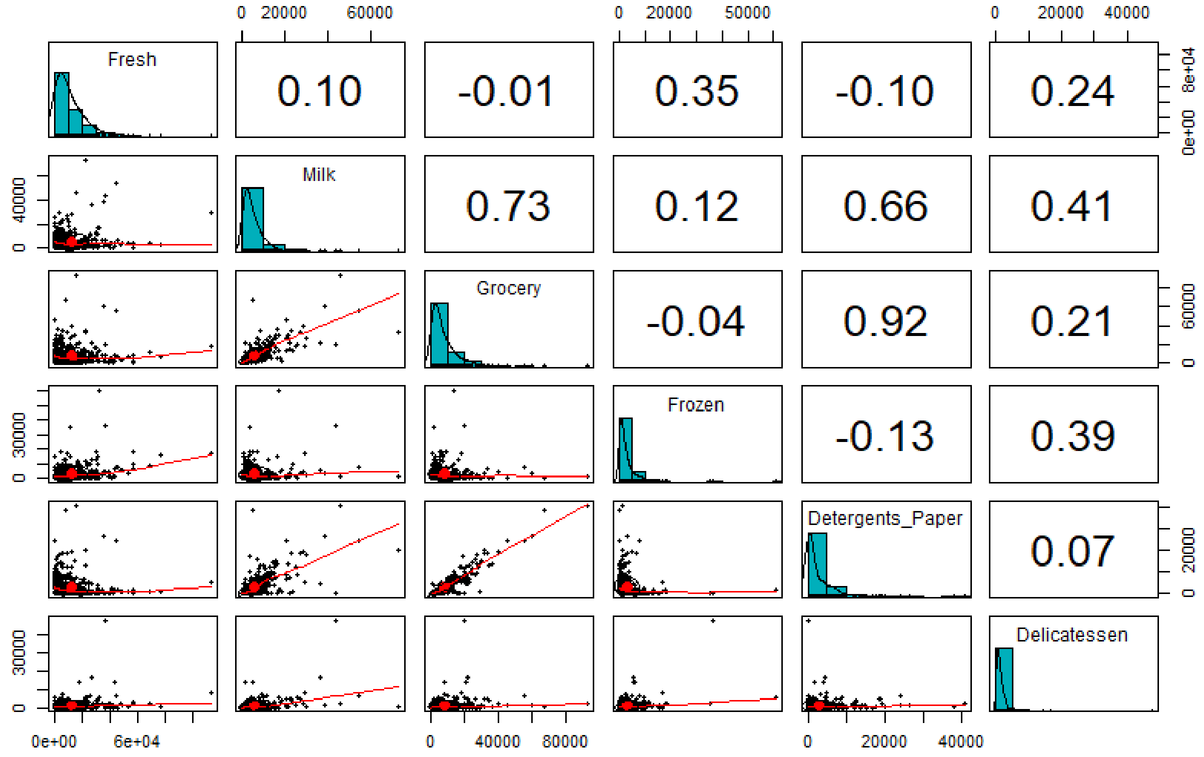


Figure 1: Data exploration matrix

It can be observed that data are skewed to the right and that there are some variables that are correlated (e.g. Grocery and Detergent Paper). Moreover, from the distribution and the analysis of the box-plots, it can be noticed the presence of several outliers that could influence the output of clustering and that must be identified.

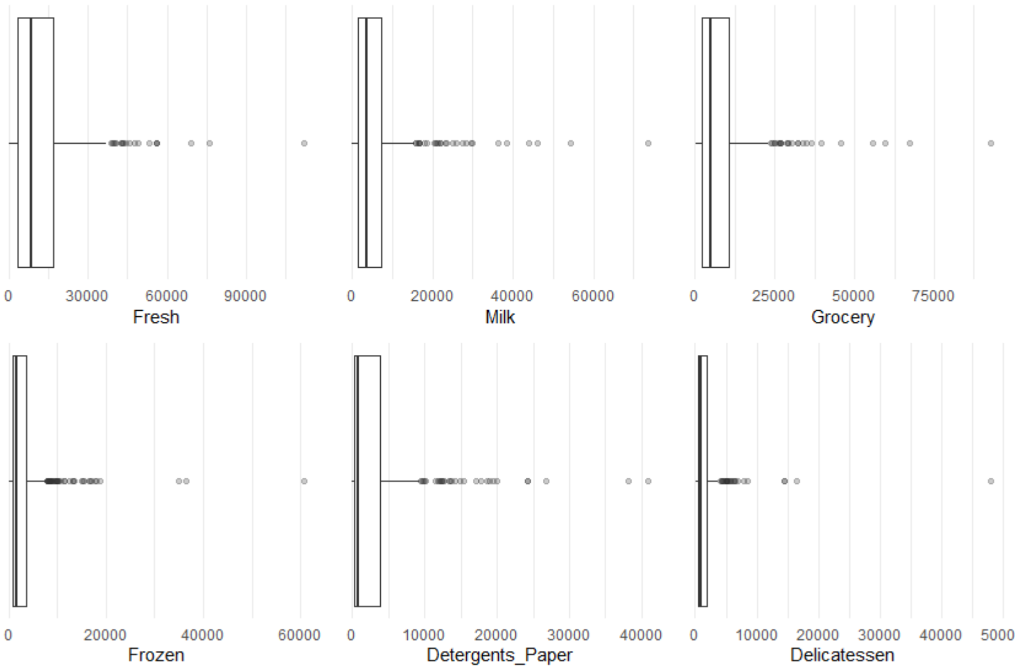


Figure 2: Box-plots

In order to overcome the skewness and make data more normal, it is applied the logarithmic transformation. After this step, outliers are identified by calculating the Mahalanobis distance, that is the distance scaled by the statistical variation in each component of the point: $MD = ((\bar{x} - \bar{y})C^{-1}(\bar{x} - \bar{y})^{\frac{1}{2}}$. Customers that have a value of the MD distribution outside the range between the quantiles of order 0.15 and 0.85 are excluded from this analysis, because they represent particular cases that could influence the clustering analysis.

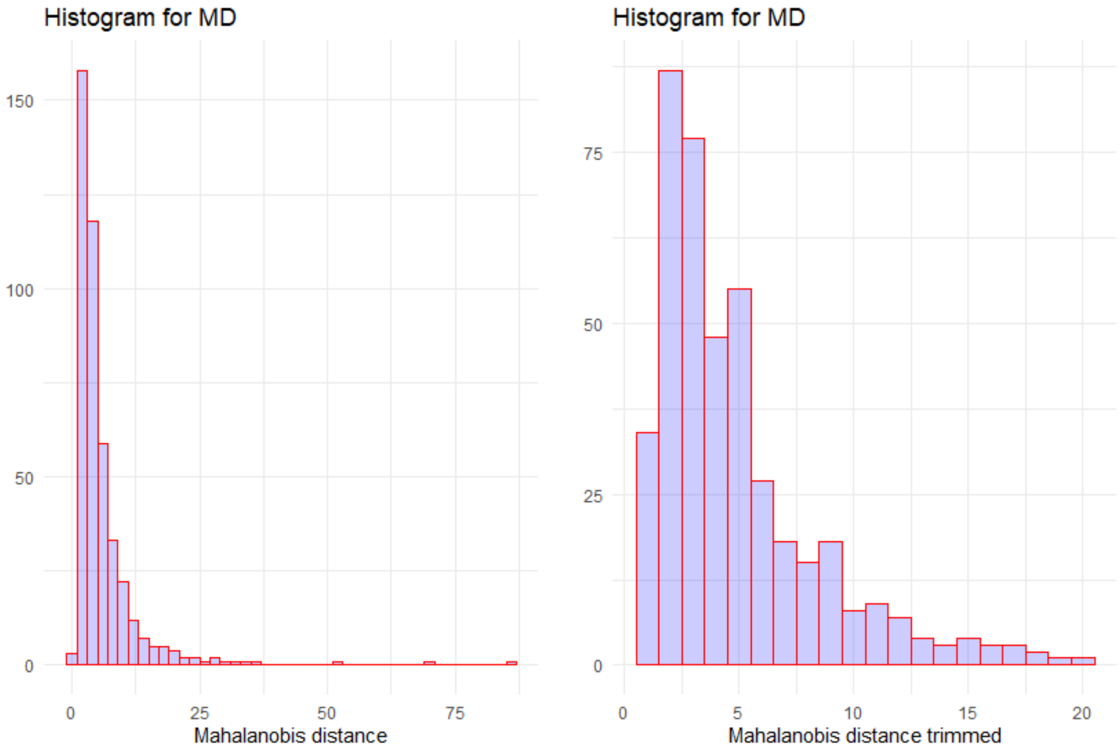


Figure 3: Original and trimmed Mahalanobis distance comparison

It follows the representation of cleaned and pre-processed data.

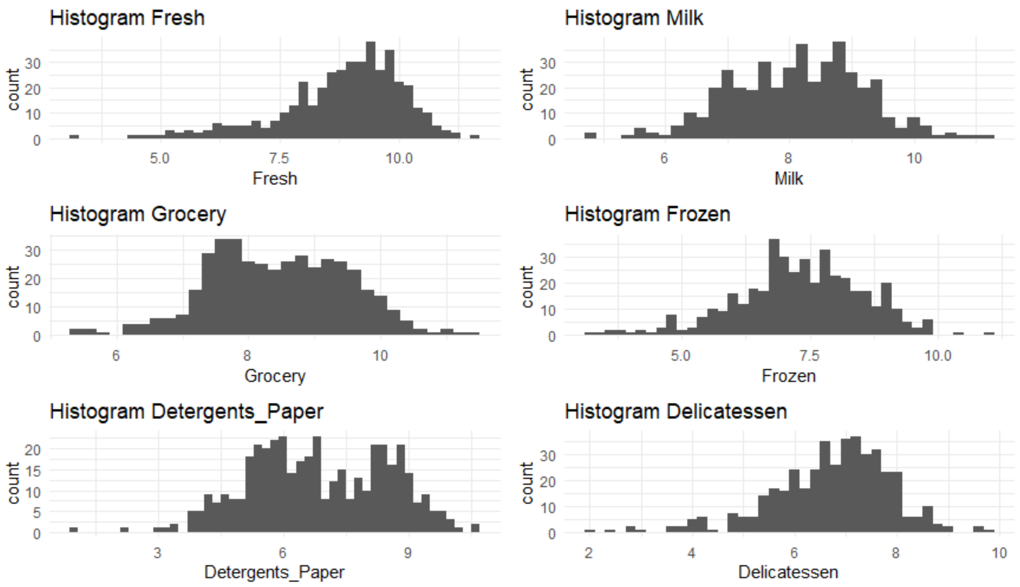
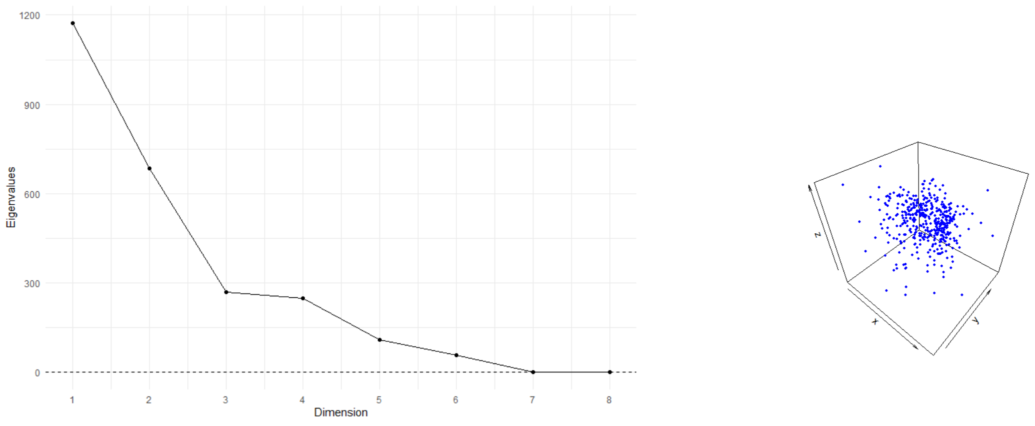


Figure 4: New data distribution after pre-processing

2 Dimensionality Reduction

In order to visualize the data in a 2d/3d plot it is necessary to reduce the dimensionality of the data. The first method applied to obtain this result is the Multidimensional scaling, a technique that works on matrices of distance. In particular here it is used the "Euclidean" distance. It can be observed from the first of the following two plots that eigenvalues stabilize at the third dimension. In the plot on the right is reported the representation of data in three dimensions.



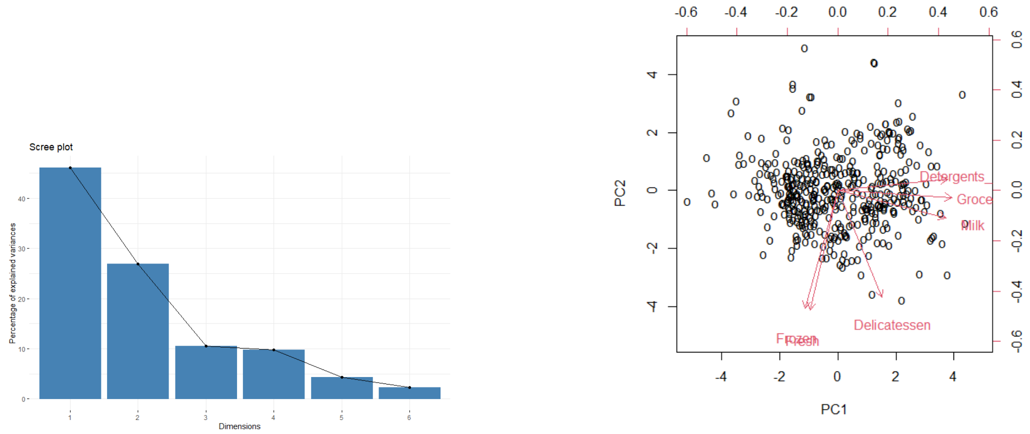
(a)

(b)

Figure 5: (a) Eigenvalues representation (b) 3d representation of data

Another technique to reduce the dimensionality of data is PCA, that consists in projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

In the following plots are reported the scree-plot, that explains the percentage of cumulative variance of the PC (in this case 2 Principal Components explains the 73% of the variance, that is a satisfactory value) and the biplot, a representation of scores on a bi-dimensional plot and the loadings given to each feature. The loadings of the variables more correlated are as expected to be close each others.

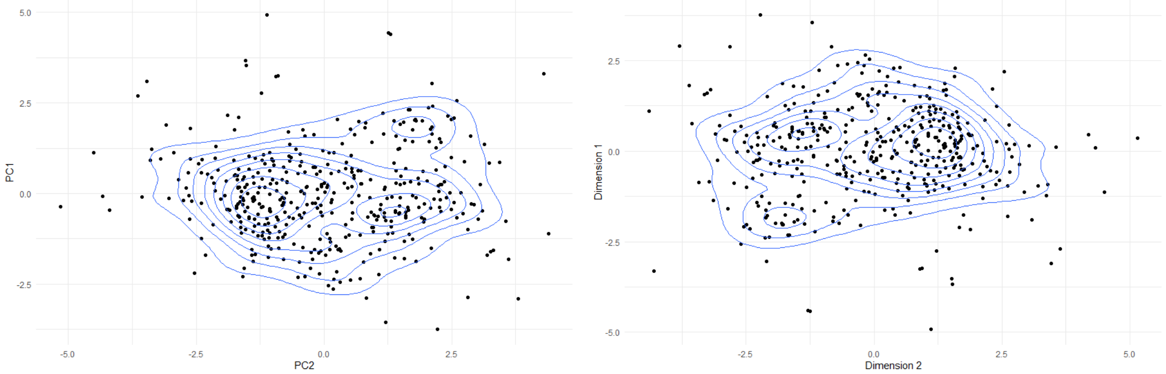


(a)

(b)

Figure 6: (a) Scree Plot (b) Biplot

With Euclidean distances the output of PCA and Multidimensional scaling is the same. The following plots show the results of PCA and MDS on two dimensions. As expected they are the same but with a different orientation.



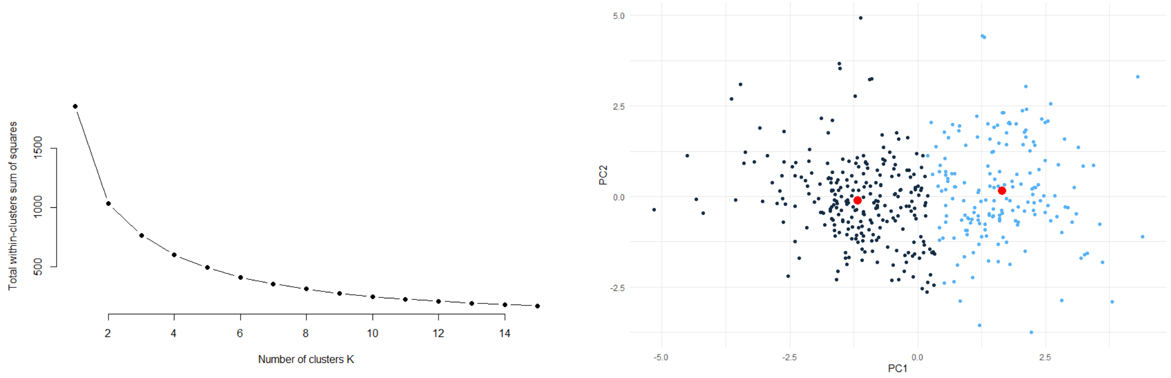
(a)

(b)

Figure 7: (a) Density plot of PCA (b) Density plot of MDS

3 Clustering

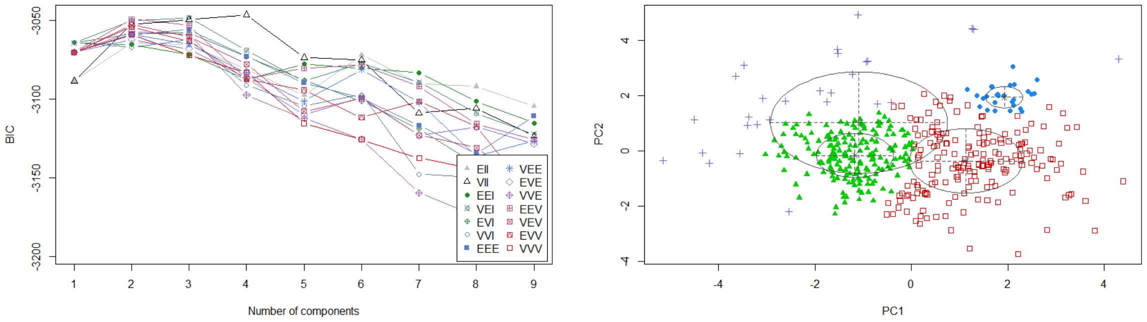
The first technique applied, is the k-means clustering, that attempts to form the cluster by minimizing the within-cluster variation and exploiting the concept of centroids. The following plots represent the output of the elbow method and the output of k-mean clustering considering k = 2.



(a) (b)

Figure 8: (a) Density plot of PCA (b) Density plot of MDS

Since the shape of the elbow is not very defined, it could be necessary to look for another method of clustering and observe what is the optimal number of cluster suggested. Particularly it is interesting to apply a model-based approach, that is based on an inferential idea instead of the distance-based approach of the k-means. In fact, the number of clusters coincides with the identification of the number of multivariate distributions in the mixture of the ones that compose data. The first of the following two plots represents the model selection criterion (BIC) varying with the number of the clusters and the type of model used. In this case, it is suggested to use a 4-components VII model, where VII stands for "varying volume, round shape (spherical covariance)". The second plot represents the application of this model on the data.



(a) (b)

Figure 9: (a) Model Selection (b) 4-clusters VII model

It can be concluded that the wholesale distributor can divide his customers in 4 segments and combining with the information given by the biplot it can be said:

- Blu segment with positive values of PC1 and PC2: it could be a cluster composed by grocery shops.
- Red segment with positive PC1 and negative PC2: it could be a cluster composed by supermarkets.
- Green segment with negative PC1 and PC2: it could be a cluster composed by restaurants.
- Purple segment with positive PC2 and PC1: it could be a cluster composed by small generalist shops.