BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Increasing Wine sales with Targeted Marketing Campaigns

Ana Marta Silva, number: 20200971
Natalia Cristina Castañeda, number: 20200575
María Luisa Noguera number: 20201005
Gustavo Tourinho, number: 20180846

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1. BUSINESS UNDERSTANDING

1.1 Introduction and Background

Consultant team – IMS Consulting Bouldin Group has been hired by WWW to leverage their database and produce insights regarding their business.

Although they have been on market for over seven years, they have kept a small business mindset and a one-fits-all marketing strategy that, although keeps their finances afloat, it is starting to prove inefficient and ineffective for a large percentage of their current customers.

Once the data was received and their business models and goals understood, the team moved forward with the analysis of the dataset, and strategize about how to **achieve the customer segmentation for the definition of new targeted marketing strategies.**

WWW's sales are made mainly through their stores, and around 40% of their business is conducted online.

During the initial diagnosis, it became evident that there were some pieces of information that this company was not collecting nor had a clear standard for (e.g. LTV calculation, contact form preference). Therefore, some of the variables were transformed or eliminated fer better accuracy of the clustering model.

Because the wine and accessories sales were tied to the customers, it was given in percentage of how much (or many) they purchased, and not as absolute values of sales per category. We also noticed that there was an overlap of percentages for the wine sales information, as per the "exotic" category was also included as any other (Dry or sweet Red and White, and dessert).

Although WWW has a database of 350 thousand customers, only a 10.000 data set was provided to our team for analysis. The team followed the CRISP-DM process with care and achieved both of the project objectives.

1.2 Business Objectives

Since the initial contact with WWW it was very clear that the objective of the project would be to segment their customers to better tend to their interests and optimize their marketing efforts. Therefore, the business objectives were defined as follows:

a. Using the sample dataset, identify the main customer segments.
b. Once the segments are identified, define a marketing strategy according to the profile of each one.

1.3 Business Success criteria

a. Defined and effective marketing campaigns per customer segment
b. Increase the frequency of sales
c. Increase average amount per purchase by a 10%
d. Increase the conversion of purchases through web traffic (#of visits to the website/# of purchases done through this channel)
e. Increase in sales of accessories (median 0) by offers with bundles.

1.4 Situation Assessment

a. General comments and Resources

For the project in question, we have a 10 thousand instances dataset with 30 attributes, out of which 7 are demographic, and 22 that describe purchasing behavior and preferences.

Even though the dataset includes customers that have been part of the WWW database for up to 40 months, there is an unusual trend in Recency (most purchases were made within the last 100 days). This was flagged as a potential bias source for the analysis.

All 4 members of the consulting team and the CEO will dedicate time to the project.

b. Time Constraint

The entire Bouldin Group team dedicated time to promptly run the analysis and finish the project on time for meeting the CEO's deadline for presenting to their board (Monday, 1st of March, 2021).

c. Terminology

Name	**Meaning**
CUSTID	Customer's permanent number
DAYSWUS	number of days as a customer
AGE	customer's age or imputed age
EDUC	years of education (may be imputed)
INCOME	household income (may be imputed)
KIDHOME	1=child under 13 lives at home
TEENHOME	1=child 13-19 years lives at home
FREQ	number of purchases in past 18 mo.
RECENCY	number of days since last purchase
MONETARY	total sales to this person in 18 mo.
LTV	Lifetime value of the customer
PERDEAL	% purchases bought on discount
DRYRED	% of wines that were dry red wines
SWEETRED	% sweet or semi-dry reds
DRYWH	% dry white wines
SWEETWH	% sweet or semi-dry white wines
DESSERT	% dessert wines (port, sherry, etc.)
EXOTIC	% very unusual wines
WEBPURCH	% of purchases made on website
WEBVISIT	average # visits to website per month
SMRACK	1=bought the small wine rack $50
LGRACK	1=bought the large wine rack $100
HUMID	1=bought wine cellar humidifier $75
SPCORK	1=silver-plated cork extractor $60
BUCKET	1=bought silver wine bucket $150
ACCESS	number of accessories (not SPCORK)
COMPLAIN	1=made a complaint in last 18 mo.
MAILFRND	1=appears on a purchased list of "mail friendly" customers
EMAILFRD	1=appears on a purchased list of "e-mail friendly" customers

d. Costs and Benefits

Because one of the objectives is to determine targeted marketing campaigns, the company will incur in implementation, merchandizing and promotion costs.

e. Risks and Contingencies

Risk	**Contingency**
Data in percent format	Transform into integers
Skewed purchasing data	Consider not including in clustering analysis
Clustering analysis does not clearly define the customer segments	Run the data through a different algorithm or perform further feature engineering

Historical data loss	Disclose that the analysis will be limited to the behavior of the last 18 months. Recommend to the CEO not rely on recency as a criterion for delete customers from the database.
Binary variables	Exclude from clustering analysis (input data).

1.5 Data Mining Goals

a. Submit the dataset to preprocessing to make sure we have the appropriate and sufficient information for accomplishing the Business Objectives
b. Using clustering methods, identify the number of customer segments and their characteristics in terms of demographics and purchasing behavior.
c. Use R2, davies-bouldin index, and Calinski- Harabasz index, to compare model accuracies and determine the most appropriate.

1.6 Data Mining Success Criteria

2. DATA UNDERSTANDING

In order to identify opportunities for increasing revenue and sales, the purchasing history provided by WWW was initially scrutinized and described, keeping in mind that it contained demographic and purchasing information for **the last 18 months only.**

Name	Meaning	Min. Value	Max. Value	Mean	Std	Type of Variable
CUSTID	Customer ID	1001	11.000	-	-	Numerical
DAYSWUS	Number of days as a customer	550	1.250	898,10	202,49	Numerical
AGE	Customer's age or imputed age	18	78	47,93	17,30	Numerical
EDUC	Years of education	12	20	16,74	1,88	Numerical
INCOME	Household income	10000	140.628	69.904,36	27.612,23	Numerical
KIDHOME	1 = child under 13 lives at home	0	1	0,42	0,49	Categorical (Binary)
TEENHOME	1 = child 13-19 years lives at home	0	1	0,47	0,50	Categorical (Binary)
FREQ	Number of purchases	1	56	14,63	11,97	Numerical
RECENCY	Number of days since last purchase	0	549	62,41	69,87	Numerical
MONETARY	Total sales to this person	6	3.052	622,56	647,14	Numerical
LTV	Lifetime value of the customer	-178	1.791	209,07	291,99	Numerical
PERDEAL	% purchases bought on discount	0	97	32,40	27,90	Numerical
DRYRED	% of wines that were dry red wines	1	99	50,38	23,45	Numerical
SWEETRED	% sweet or semi-dry reds	0	75	7,05	7,87	Numerical
DRYWH	% dry white wines	1	74	28,52	12,58	Numerical
SWEETWH	% sweet or semi-dry white wines	0	62	7,07	8,02	Numerical
DESSERT	% dessert wines (port, sherry, etc.)	0	77	6,95	7,88	Numerical
EXOTIC	% very unusual wines	0	96	16,55	17,25	Numerical
WEBPURCH	% of purchases made on website	4	88	42,38	18,52	Numerical
WEBVISIT	Average # visits to website per month	0	10	5,22	2,33	Numerical
SMRACK	1 = bought the small wine rack $50	0	1	0,08	0,27	Categorical
LGRACK	1 = bought the large wine rack $100	0	1	0,07	0,25	Categorical

HUMID	1 = bought wine cellar humidifier $75	0	1	0,08	0,27	Categorical
SPCORK	1 = silver-plated cork extractor $60	0	1	0,07	0,25	Categorical
BUCKET	1 = bought silver wine bucket $150	0	1	0,01	0,11	Categorical
ACCESS	Number of accessories (not SPCORK)	0	3	0,25	0,54	Categorical
COMPLAIN	1 = made a complaint in last 18 mo.	0	1	0,01	0,11	Categorical (Binary)
MAILFRND	1 = appears on a purchased list of “mail friendly” customers	0	1	0,10	0,30	Categorical (Binary)
EMAILFRD	1 = appears on a purchased list of “e-mail friendly” customers	0	1	0,05	0,22	Categorical (Binary)

While analyzing the data, we observed the following:

a. The dataset did not have any missing values,
b. The dataset had 18 (62%) numerical variables and 11 (38%) categorical variables.

In the following analysis, we will understand the high standard deviation and the not well-distributed values according to the numerical variable's mean.

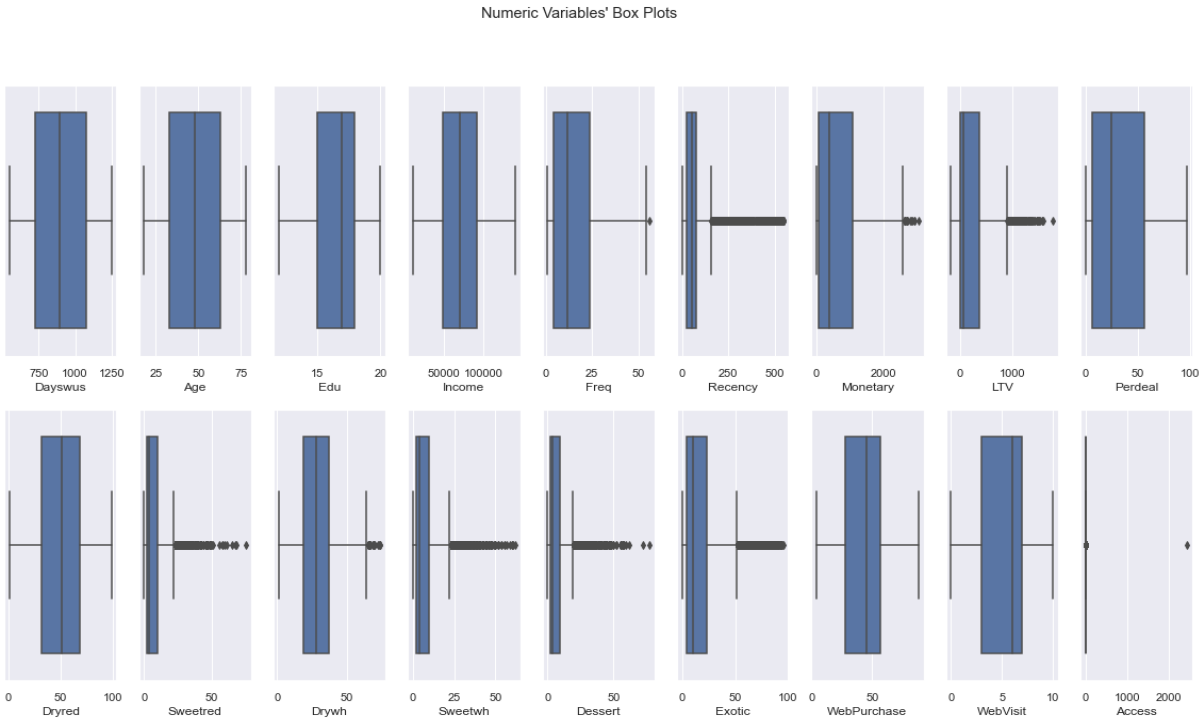


Figure 1. – Box Plots (Numerical Variables)

Further exploring the dataset, It seemed that few of the variables had outliers to be removed. However, taking a closer look, outliers for "FREQ", "LTV", "SWEETRED", and "DESSERT" were removed. (Figure 1).

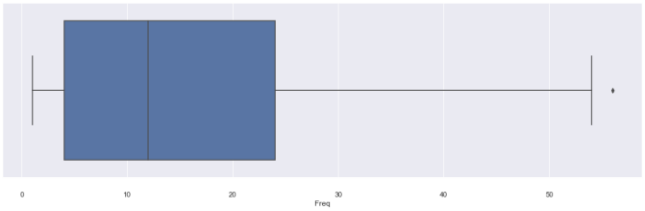


Figure 2. – Box Plot (Frequency – Number of the purchases)

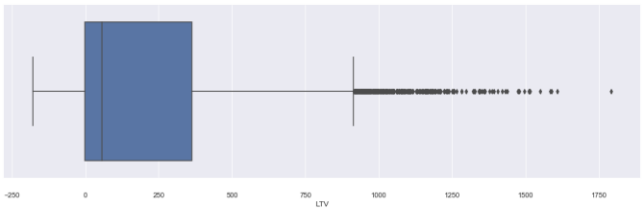


Figure 3. – Box Plot (LTV – Lifetime Value of the customer)

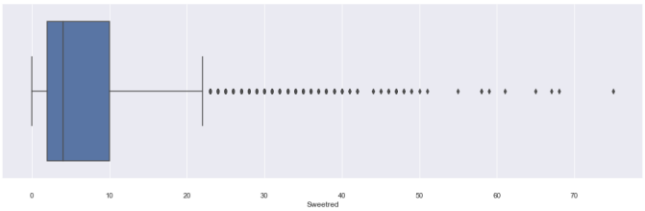


Figure 4. – Box Plot (Sweet or Semi-Dry Red Wines)

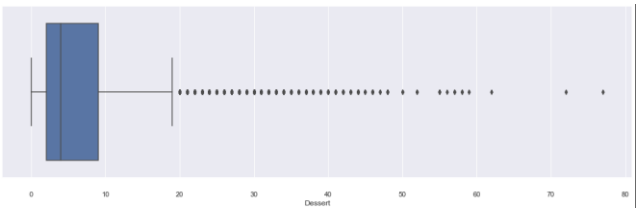


Figure 5. – Box Plot (Dessert Wines)

After isolating the aforementioned variables, we decided to remove outliers using the 25th and 75th percentiles for the following variables: "FREQ" > 84 (Figure 2), "LTV" > 1.462 (Figure 3), "SWEETRED" > 34 (Figure 4), and "DESSERT" > 30 (Figure 5).

Continuing the exploration of the dataset, we could identify that most of the variables were not normally distributed. The variables that were normally distributed were the "INCOME", "DRYRED", and "DRYWH" (Figure 6).

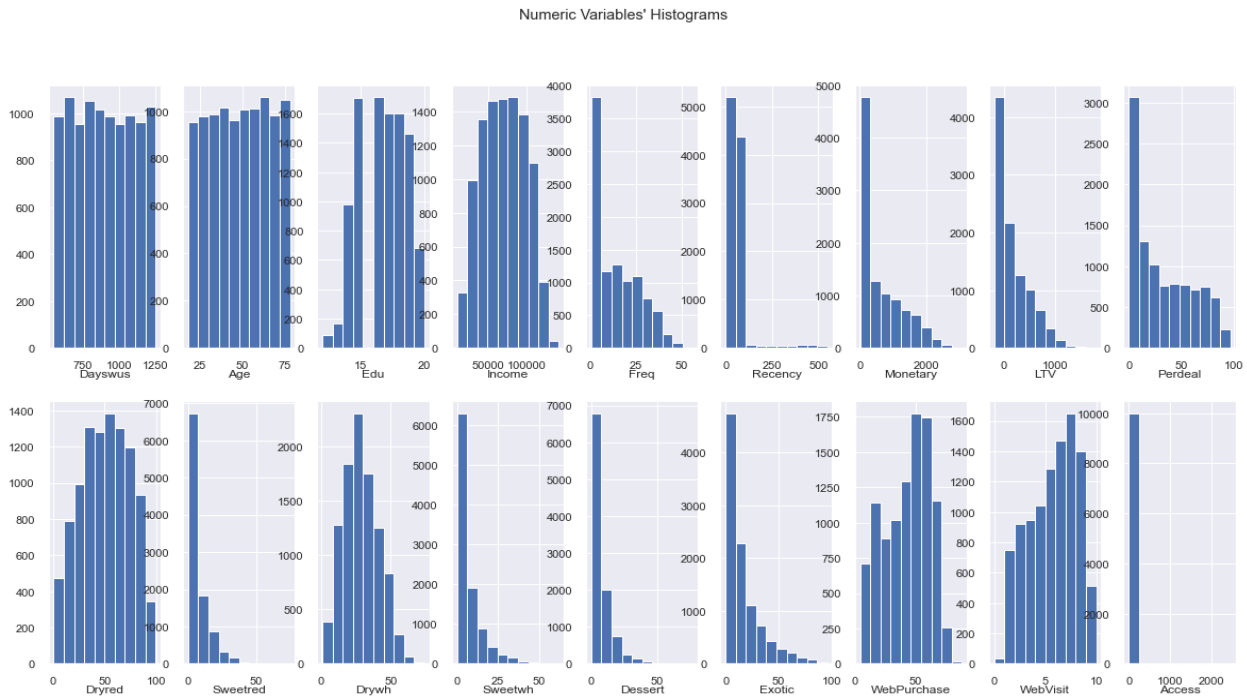


Figure 6. – Histograms (Numerical Variables)

As mentioned in "Business Understanding", we will not use all the information variables provided for the model since some of them are binary and are not appropriate for the algorithms. Variables discarded include "SMRACK", "LGRACK", "HUMID", "SPCORK", "BUCKET".

3. DATA PREPARATION

3.1 Data Quality

Given the initial outlier analysis, we decided to use the Interquartile Range Method with the usual 25th and 75th percentiles, thus removing only a 3,36% of the data from the variables who's distribution in the boxplot seemed to have defined outliers: Frequency, LTV, Sweetred, Desserts.

3.2 Feature Engineering

a. Conversion rate (conversion): number of online purchases ((WebPurchase/100)*Freq) divided by the total number of visits (WebVisit*18)
b. Average Purchase (Avg_purchase): average amount spent per purchase (Monetary/Freq)
c. Average Time Between Transactions (Days_between): Number of days between purchases (18*30)/Freq.
d. Relative Spent on each Product (Spent_dryred, Spent, Sweetred, Spent_drywh, Spent_sweetwh, Spent_dessert, Spent_exotic): average amount spent in each type of wine. Monetary*(wine/100)

3.3 Feature Exclusion

a. The variables with the percentage of purchases per wine and replace them with the relative amount spent in each type of wine.
b. The binary variables were also excluded from the clustering analysis.

3.4 Feature Significance and Correlation (Pearson, Phik, Principal Component Analysis)

Based on the two correlation analysis ran (Pearson and Phik method) we decided to eliminate the variable LTV since it's highly correlated with four variables (Income, KidHome, Freq and Monetary). Likewise, after the Principal Component Analysis we decided to exclude the variables Recency and Spent_exotic because they don't explain much of the variance of the four principal components.

3.5 Data Subsets

Four data subsets were selected for the modelling:

Subset 1 All variables (After first correlation analysis and feature engineering)	**Subset 2** After correlation analysis with new features	**Subset 3** Demographic variable	**Subset 4** Customer behavior
Dayswus	×		×
Age	×	×	
Edu	×	×	
Income		×	
Freq			×
Monetary			×
Perdeal	×		×
WebPurchase	×		×
WebVisit	×		×
Spent_dryred	×		×
Spent_sweetred	×		×
Spent_drywh	×		×
Spent_sweetwh	×		×
Spent_dessert	×		×
Days_between	×		×
Avg_purchase	×		×
conversion	×		×

4. MODELING

4.1 Modeling Technique

Before defining which clustering technique to use, we visualized the dataset in a 2-dimensional scatter plot with the 2 first principal components. Thanks to this visualization we concluded we should use only partitioning and agglomerative techniques (not density clustering techniques).

Consequently, the clustering algorithm to apply in the data subsets was K-means.

Additionally, after obtaining the labels corresponding to each perspective analysis, we merged social-demographic and costumer behavior by applying the hierarchical clustering technique.

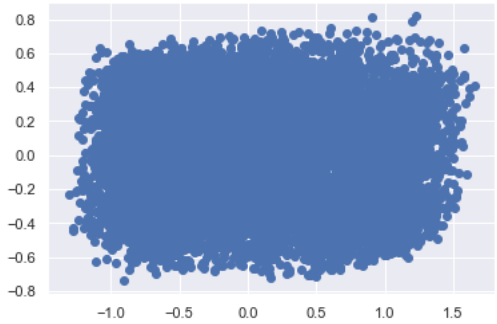


Figure 7. – Scatter plot with first 2 principal components

4.2 Modelling Assumptions

The assumptions behind each technique are: K-means assumes that the clusters are spherical shaped and that an individual can only be part of one clustering. Besides the number of clusters have to be decided a priori. The hierarchical clustering has no assumptions in terms of number of clusters or shape.

4.3 Build the Model

When applying the K-means technique we used the initialization parameter as K-means ++ because it chooses the initial seeds which are the furthest apart, thus increasing the probability that the initial centroids belong to different clusters. We also utilized the elbow method to define the number of clusters to input in the algorithm.
In applying the hierarchical clustering, the parameter linkage was set to Ward because it minimizes the variance of the clusters being merged.

4.4 Assess the model

To assess the various approaches applied we computed the following assessment metrics: R2 score, the Davies-Bouldin index (which represents the average similarity between clusters, so the lower the score the better) and the Calinski- Harabasz index (which corresponds to the ratio between the sum of inter-cluster dispersion and the intra-cluster dispersion for all clusters, so the higher the score the better). The results are shown in the tables below.
For R2 results:

	all variables	less variables	socio	behav	merged
kmeans	0.559499	0.516719	0.73377	0.669201	0.546402

For Davies-Bouldin index:

	all variables	less_variables	socio	behav	merged
kmeans	1.339752	1.446228	0.985941	1.334567	1.360106

For Calinski - Harabasz index:

	all variables	less variables	socio	behav	merged
kmeans	6135.419876	5164.710934	8874.796905	4885.007016	5818.802922

Since the kmeans technique, applied to all the variables, presented the best scores (R2 score 55.9, Davies-Bouldin index 1.33, Calinski- Harabasz index 6135), we concluded that this was the best

model to produce very well separated and diverse segmentation results. This technique awarded us with 3 segments of very similar size: cluster 1 - 3755, cluster 2 –3132, cluster 3- 2777 costumers.

The segmentation allowed us to define 3 clear segments of customers, which are:
Cluster 1 – This cluster is constituted by the **older costumers (68)** and with the **highest income ($102k)**. **Most frequent** costumers (bought 29 times/ 18 months). **Highest value per purchase** ($48). Only **4% of the times use discounts**. **Lowest percentage of web purchases** (19%), lowest web visits (3 times/ month), which amounts for the higher conversion rate. Have lowest number of days between purchases (20). Prefer dry red wine, then dry white wine, being for this type of wine their best costumer. Have the **lowest preference for exotic wines** (8% of their choices)

Cluster 2 – These costumers are the **youngest (32)** with the **lowest income ($44k). Less frequent** (bought 4 times/18 month). **Lowest average value per purchase ($18)**. Highest number of days between purchases (212). **Highest usage of promotions** (60%) , Highest number of visits to the website (7 times/month). **Highest percentage of web purchases** (56%). **Best costumers of exotic wines** (24% of choices). Relatively highest preference of sweet red , sweet white and dessert wines.

Cluster 3 – This cluster is constituted by our **second oldest (51)** and **second richest(75k**). Bought 15 times /18 months, the second highest. Consequently, have **average purchase value of $36**. Use discounts 23% of the times. **45% of purchases are online**. Days between their purchases are on average 44. Have the **highest preferences for dry red**. Exotic wines are chosen 13% of the times.

Finally, another insight obtained was that education , number of days as clients and recency are features with no segmentation power because all the segments presented have very similar values for these variables. Costumers bought with WWW in the last 2.5 years and have an average of 16 to 17 years of education, which means a Bachelor and sometimes a Master degree. They have bought in WWW in the last 50 to 80 days.

Regarding having children, we concluded that the cluster 2 is the one who have more kids with age below 13 years old which is understandable since they are the younger ones.

5. EVALUATION

5.1 Evaluation results

The results of the clustering techniques allowed us to identify three very distinct types of costumers which emphasizes the importance of this segmentation effort. The finding of costumers with so different social characteristics and buying behaviors is a clear indication that the marketing initiatives of WWW shouldn't be mass targeted.

From a population (the available dataset) with on average sales of $622/18 months per costumer, a frequency of 15, a web consumption of 42% and an average value per purchase of $32, this segmentation broadens the possibilities of improvement by showing a range of different behaviors: $75-$1431, 4-29, 19%-56% and $18-$48 for the variables mentioned,, respectively.

Another input obtained was that the accessories are very seldomly bought by the costumers, so the company have to rethink the accessories variety and prices in order to make them more attractive.

5.2 Review process

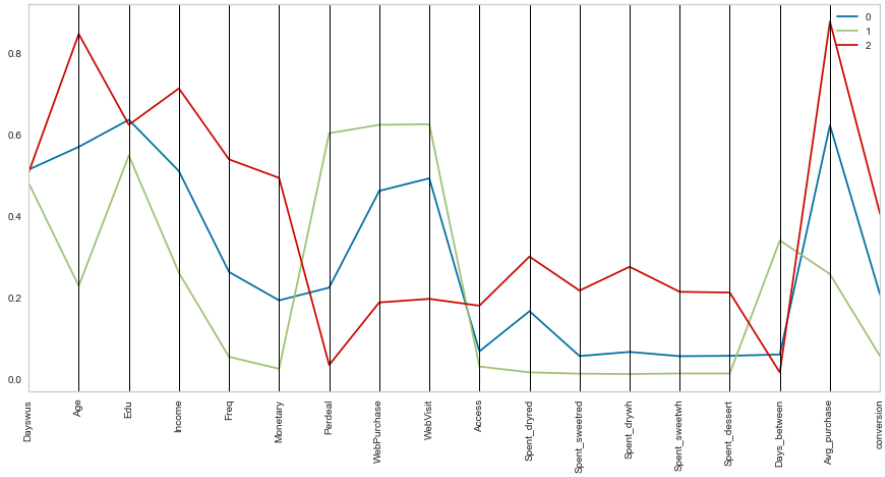
Technically we didn't identify any revisions to apply to the project. However, we consider that the sample available for the project could have been bigger as it only corresponded to 2.8% of the all dataset of the costumers of the WWW company.

5.3 Next steps

Following this segmentation endeavor, the company should proceed with the marketing campaigns that leverage these results and allow to increase the average purchase value. Moreover, the company should apply the model of kmeans one time each year to assess if the pool of costumers have changed to justify adjustments in the segments defined.

6. DEPLOYMENT

6.1 Marketing initiatives based on segmentation



Segment Description	Marketing Strategy
Cluster 0 – Silver Foxes	Mid-spenders, these customers have high potential to be migrated to high-spenders and thus increase future purchases. Create a wine rewards program where if the member makes certain number of purchases and average ticket per quarter, he/she is invited to wine events and special discounts.
Cluster 1 – Eldering millennials	Low-spenders, like to drink wine, but don't know what kind to buy and may select by the price. Personalize social media marketing communications, creating educational, interactive and dynamic content about wine. Create a wine club where if the member is subscribed, he/she gets special online discounts.
Cluster 2 - Connoisseurs	High-spenders. Traditionalist, wine connaisseurs, they enjoy wines from established wineries. They should be introduced to a wider variety of well-known brands, exclusive and unique collections. Create a wine membership where if the member makes certain number of purchases in one year, he/she has special deals in well-known and high-ticket price wine brands and is invited to exclusive wine events every month.