**Nova Information Management School**
**Advanced Analytics Master's degree**
**Data Mining**
**Project 1: Wonderful Wines of the World**

**Group 20: Carolina Bellani (M20170098), Gonçalo Passão (M20170450) and Sofia Jerónimo (M20170070)**

## 1. Introduction

Wonderful Wines of the World (WWW) is a seven-year-old company that sells wines and wine accessories from different wineries around the world. Customers can shop at one of the ten small stores in major cities around the USA, by telephone (after checking the wine catalog) or on the website.

WWW has now 350.000 customers in it's four-year database. Each person that did not purchase a product within the last 18 months is deleted from the active database.

This company is willing to start using this database to differentiate customers and developing more focused marketing strategies. To achieve these goals, we performed basic statistics to understand digital impact on the behaviour of the customers and a SEMMA approach (Sample, Explore, Modify, Model and Assess) to profile the adequate number of clusters in a random sample of 10.000 customers from WWW's active database.

## 2. Analysis and Conclusions

### Part 1: Understanding customer behaviour related to newsletter and shareMedia

For the first part of the project, we used the SAS Enterprise Guide 7.1. Firstly, the datasets called 'ValueEngage' and 'Newsletter' were merged together into one table named 'FullTable'. The missing values for the variables 'Newsletter' (1=Signed up to receive newsletters) and 'shareMedia' (1=Shared at least one WWW post) were changed to zero. For the remaining variables, we did not have missing data and other kind of errors.

We checked the distribution of every variable (**Figure S1.1-1.3**) through histograms.

We analysed if the customers that received a newsletter and/or shared at least one post WWW have bought more products compared to those that did not receive and/or share the post. We expected that the digital approach would have a positive impact on the customer behaviour and interest in the goods of the company, more specifically in decreasing the 'Recency' (number of days since last purchase), and increasing the 'Monetary' (total sales in 18 months) as well as 'Freq' (number of purchases in the past 18 months), 'WebPurchase' (% of purchases made on website) and 'WebVisit' (average number of visits to website per month).

Although the mean and standard deviation differences between those customers that received the newsletter and/or shared at least one post compared to those that did not receive and/or share the post (**Table 1.1**) were rather minimal for 'Freq', 'WebPurchase' and 'WebVisit'. For 'Recency' and 'Monetary' there were some changes in the mean and standard deviation.

**Table 1.1.** Mean, standard deviation, minimum and maximum for the 4 different customer groups: Newsletter = 0 and shareMedia = 0, Newsletter = 0 and shareMedia = 1, Newsletter = 1 and shareMedia = 0 and Newsletter = 1 and shareMedia = 1.

| Newsletter | shareMedia | Variable | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|---|
| 0 | 0 | Freq | 5648 | 14.6285411 | 11.9467169 | 1.0000000 | 56.0000000 |
| | | Recency | 5648 | 61.5872875 | 68.3055990 | 0 | 549.0000000 |
| | | Monetary | 5648 | 622.2004249 | 646.1216632 | 7.0000000 | 3052.00 |
| | | WebPurchase | 5648 | 42.3121459 | 18.5644565 | 4.0000000 | 84.0000000 |
| | | WebVisit | 5648 | 5.1942280 | 2.3519725 | 0 | 10.0000000 |
| 0 | 1 | Freq | 1865 | 14.2573727 | 11.8787793 | 1.0000000 | 52.0000000 |
| | | Recency | 1865 | 63.3844504 | 71.0879939 | 0 | 543.0000000 |
| | | Monetary | 1865 | 602.9898123 | 640.9094961 | 7.0000000 | 2821.00 |
| | | WebPurchase | 1865 | 42.9474531 | 18.4136841 | 5.0000000 | 82.0000000 |
| | | WebVisit | 1865 | 5.2804290 | 2.2939012 | 0 | 9.0000000 |
| 1 | 0 | Freq | 1899 | 14.8335966 | 12.1307727 | 1.0000000 | 50.0000000 |
| | | Recency | 1899 | 66.0331754 | 77.9408496 | 0 | 546.0000000 |
| | | Monetary | 1899 | 634.6840442 | 656.6261029 | 6.0000000 | 2705.00 |
| | | WebPurchase | 1899 | 42.2996314 | 18.4510646 | 6.0000000 | 88.0000000 |
| | | WebVisit | 1899 | 5.2343339 | 2.2901633 | 0 | 10.0000000 |
| 1 | 1 | Freq | 588 | 15.1360544 | 11.9390214 | 1.0000000 | 45.0000000 |
| | | Recency | 588 | 55.4659864 | 49.9360804 | 0 | 479.0000000 |
| | | Monetary | 588 | 648.8486395 | 645.4496545 | 9.0000000 | 2410.00 |
| | | WebPurchase | 588 | 41.4268707 | 18.6787536 | 5.0000000 | 80.0000000 |
| | | WebVisit | 588 | 5.1717687 | 2.3680336 | 0 | 9.0000000 |

We performed a t-test (for simplicity, we opted to use T-test although some variables do not have a normal distribution and we assumed independent and identically distributed samples) to verify if the means for 'Freq', 'Recency', 'Monetary', 'WebPurchase', 'WebVisit' were statistically different from the means of the same variables considering all the combinations of the variables 'Newsletter' and 'shareMedia' (Newsletter = 0 and shareMedia = 0, Newsletter = 0 and shareMedia =1, Newsletter = 1 and shareMedia = 0, Newsletter = 1 and shareMedia = 1).
Contrary to the expectations, the mean differences between those customers that received a newsletter and/or shared at least one post compared to those that did not receive and/or share the post were not statistically significant for 'Freq', 'Monetary', 'WebPurchase' and 'WebVisit' (all p-value are >= 0.1381, **Table 1.2**). However, for 'Recency', the mean was significantly different for clients that received the newsletter and/or share a post compared to the customers (Newsletter = 1 and shareMedia =1: p-value = 0.0008, Newsletter = 1 and shareMedia = 0: p-value = 0.012, **Table 1.2**) that did not receive the newsletter and/or share a post.

**Table 1.2.** T-test results for the 4 different customer groups. We used the Satterhwaite method for the variables that had a significantly different variance and Pooled method for the variable that had not a significantly different variance (Folded F statistic, **Table S1.1**).

| Variable | Newsletter = 0 and shareMedia = 0 | | | Newsletter = 0 and shareMedia = 1 | | | Newsletter = 1 and shareMedia = 0 | | | Newsletter = 1 and shareMedia = 1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DF | t-value | Pr > \|t\| | DF | t-value | Pr > \|t\| | DF | t-value | Pr > \|t\| | DF | t-value | Pr > \|t\| |
| **Freq** | 9998 | 0.00 | 0.9967 | 9998 | 1.48 | 0.1381 | 9998 | -0.83 | 0.4059 | 9998 | -1.06 | 0.2888 |
| **Recency** | 9114.9 | -1.33 | 0.1844 | 9998 | -0.67 | 0.5029 | 2612.3 | -2.31 | 0.0212 | 743.49 | 3.37 | 0.0008 |
| **Monetary** | 9998 | -0.06 | 0.9502 | 9998 | 1.45 | 0.1477 | 9998 | -0.91 | 0.3642 | 9998 | -1.02 | 0.3099 |
| **WebPurchase** | 9998 | -0.39 | 0.6936 | 9998 | -1.48 | 0.1398 | 9998 | 0.20 | 0.8414 | 9998 | 1.28 | 0.2002 |
| **WebVisit** | 9998 | -1.09 | 0.2741 | 9998 | -1.31 | 0.1897 | 9998 | -0.37 | 0.7126 | 9998 | 0.48 | 0.6307 |

Thus, we can conclude that the digital approach has an impact on the customer behaviour. It has a positive impact when the customers receive the newsletter and share at least one post, having a lower 'Recency' average than the remaining customers. However, there is a negative impact when the customers receive a newsletter and do not share any post, these customers have on average a higher 'Recency' than the rest of the clients.

## Part 2: Customer segmentation and profile

For the second part of this project, we used both SAS Enterprise Miner 14.1 and SAS Enterprise Guide 7.1 to perform customer (value and engage) and consumption segmentations using a SEMMA approach.
The preliminary analysis described below was done in SAS Enterprise Miner.
We did not consider the variables concerning the number of accessories bought and the binary variables such as if the client did or did not buy a wine bucket, a wine cellar humidifier, a small wine rack, a large wine rack or a silver-plated cork extractor. This is because these variables do not show the wine affinity of the customer. Similarly, we did not consider 'Newsletter' and 'shareMedia' since we had already analysed them in the first part of this project.
Before performing the clustering, we did several preliminary analyses on the variables we decided to keep. We did a cross tabulation between two variables 'kidhome' (1=child under 13 lives at home) and 'teenhome' (1=child 13-19 years lives at home), to analyse and understand the relationship of these categorical data (**Table 2.1**).

**Table 2.1.** Cross tabulation between 'Kidhome' and 'Teenhome'.

| Variable1 | Variable2 | Value1 | Value2 | Frequency | Percent | RowPercent | ColPercent |
|---|---|---|---|---|---|---|---|
| Kidhome | Teenhome | 0 | 0 | 2946 | 29.46 | 50.68823 | 55.56394 |
| Kidhome | Teenhome | 0 | 1 | 2866 | 28.66 | 49.31177 | 61.00468 |
| Kidhome | Teenhome | 1 | 0 | 2356 | 23.56 | 56.25597 | 44.43606 |
| Kidhome | Teenhome | 1 | 1 | 1832 | 18.32 | 43.74403 | 38.99532 |

We did a Chi-square test of independence to determine if there is a significant relationship between 'kidhome' and 'teenhome' by testing the frequency of having a kid or not at home, for the variable 'kidhome' compared across the frequency of having or having a teenager or not at home, for the variable 'teenhome'. This test showed that there is a relationship between these variables (p-value < 0.001) which means the customers that have one kid at home, generally do not have a teenager at home (2356 customers) and the clients that have one teenager at home, tend to not have a kid at home (2866 customers). It is less common to have clients that have both a kid and a teen at home (only 1832 customers). The most frequent case is to have no kid and no teenager at home (2946 customers).
In order to avoid correlated variables that can be detrimental to the clustering analysis (multicollinearity problem), we analysed the Pearson correlation coefficients between the variables (**Figure 2.1**).

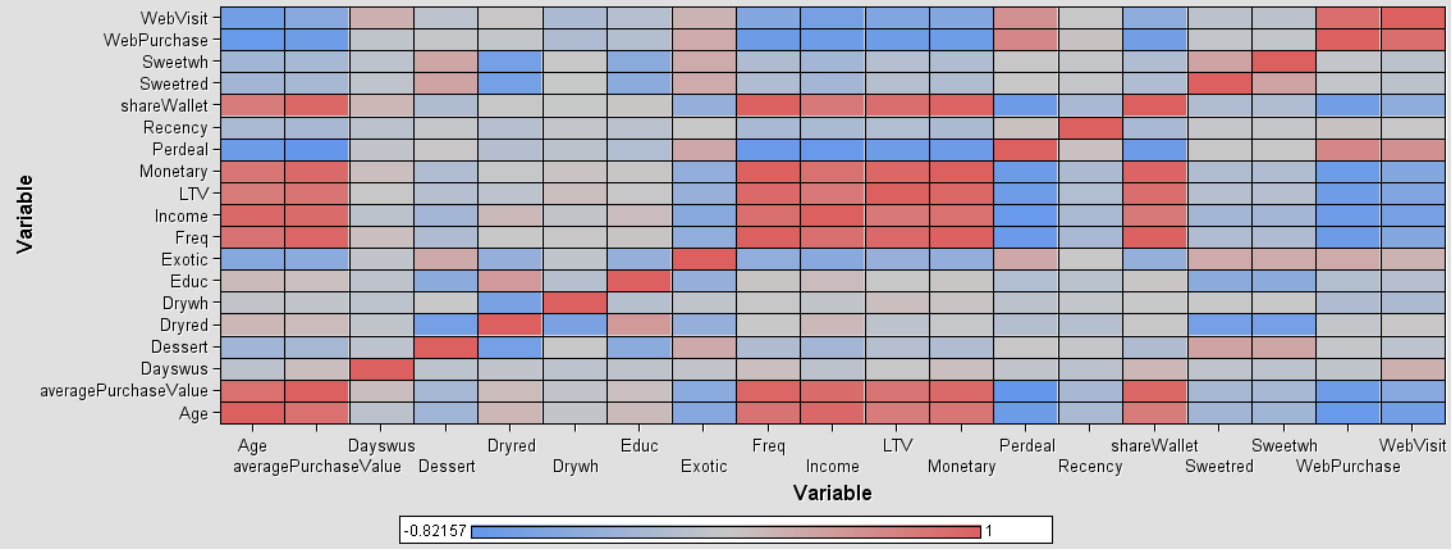


**Figure 2.1.** Matrix of Pearson correlation coefficients.

We can see that a high positive correlation exists (indicated in red, all coefficients are >= 0.788, between 'Age', 'Freq', 'Income', 'LTV' and 'Monetary' and between 'WebVisit' and 'WebPurchase'. There are some slight negative correlations (indicated in blue, all the coefficients are >= -0.786), for example between 'WebPurchase' and 'Age' and between 'Perdeal' and 'Income'.
To decrease the number of correlated variables in our clustering, we created new variables called 'shareWallet', which is the ratio between 'Monetary' and 'Income' (household income) and 'averagePurchaseValue', which is the ratio between 'Monetary' and 'Freq'. The 'shareWallet' variable was still highly correlated to 'Age', 'Freq', 'Income', 'LTV' (**Figure 2.1**), but since this ratio will be important to help us defining the customer value and engagement, it was kept for the clustering analysis. 'averagePurchaseValue' was highly correlated to 'Age', 'Income', 'LTV' and 'ShareWallet' (**Figure 2.1**) and since it does not add more information than other considered variables, it was discarded from the analysis.
We also decided to drop the variable 'WebVisit' because it does not add more information than the variable 'WebPurchase'.
To get a better understanding of our data, we explore our data by plot different variables against one another, here we present the most relevant graphs (**Figure 2.2**).

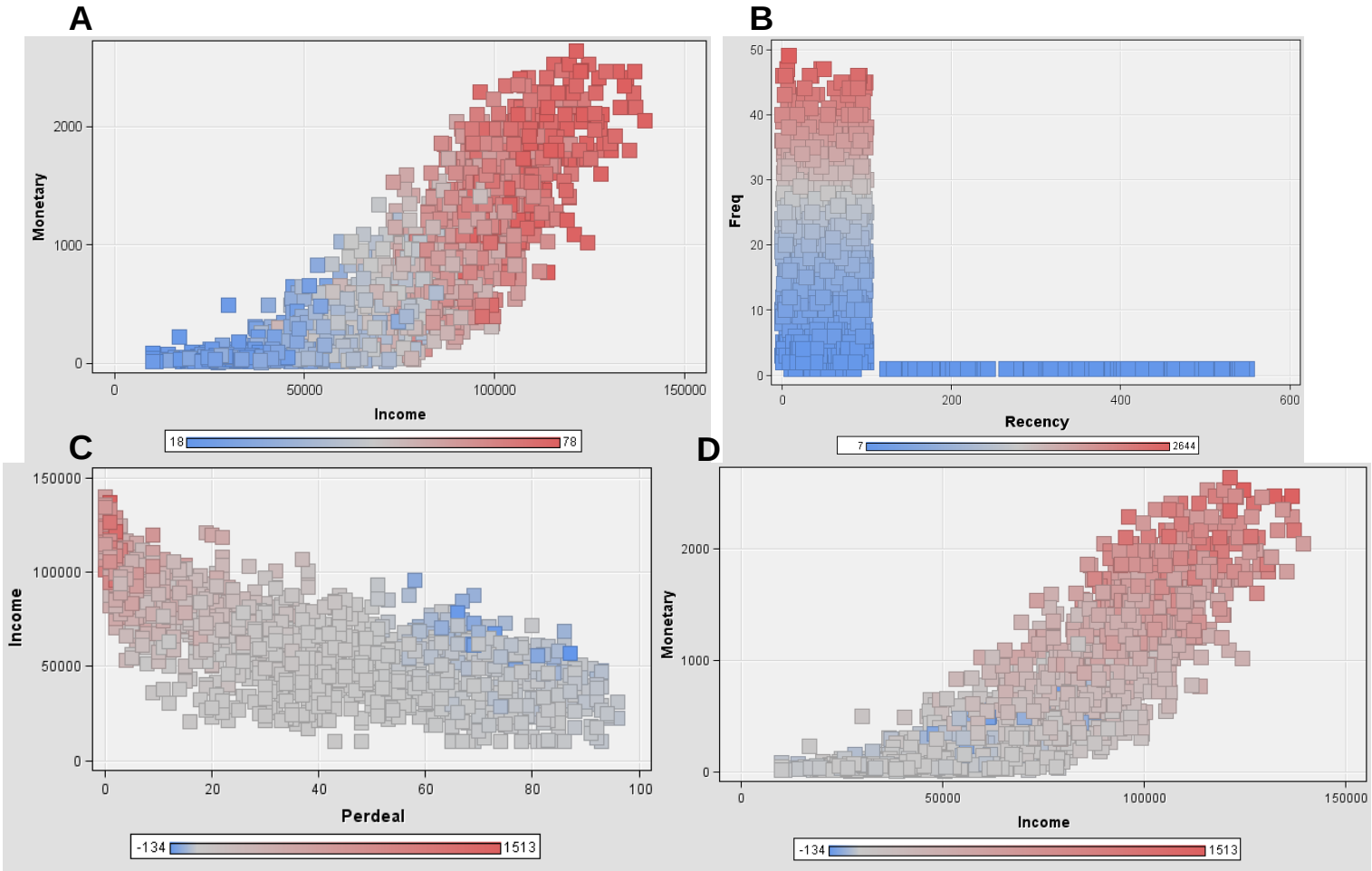


**Figure 2.2.** Scatter plot of A) 'Income' vs. 'Monetary', with colour representing 'Age'. The variables have values that are linear and positive correlated; B) Recency' vs. 'Freq', with colour representing 'Monetary'; C) 'Perdeal' vs. 'Income', with colour representing 'LTV'; D) 'Income' vs. 'Monetary', with colour representing 'LTV'.

In **Figure 2.2A**, we can see that as 'Age' increases, the 'Income' and the 'Monetary' of the customer also increases. In **Figure 2.2B**, we plotted the RFM variables, among customers that have less than 100 days since the last purchase ('Recency'), there are two profiles of customers, the ones that have high 'Freq' and 'Monetary' and other customers that have low 'Freq' and 'Monetary'. The customers that have 'Recency' higher than 100 days, have low 'Freq' and 'Monetary' values. In **Figure 2.2C**, the customers that have higher 'Income' and higher 'LTV' (lifetime value of the customer) have lower percentages of products bought on discount ('Perdeal'). In the last graph, **Figure 2.2D**, the clients that have higher incomes spent more money in the last 18 months ('Monetary') and have a higher 'LTV'.

To avoid outliers that can have a negative effect for our clustering analysis, we decided to remove the customers that did not purchase any products in the last 400 days because we consider those customers as not active and so, there will be not important for our profiling. However, we did not remove the possible outliers of 'Monetary', 'Income', 'Freq', 'WebPurchase', 'shareWallet' ('shareWallet''s histogram in **Figure S2.1**) and of the products 'Dessert', 'Drywh', 'Exotic', 'Sweetred', 'Sweetwh' because those customers have the highest value for the company and for that reason, they are important to be considered in the clustering. We also did not remove possible outliers of 'LTV' because this will allow us to segment clients into profitable and not profitable categories and consequently, have better business strategies in the future. Finally, we did not remove possible outliers of 'Perdeal' because these clients will probably respond well to future discount campaigns. In the end of this process, we excluded only 172 out of 10,000 customers.

We performed again the Pearson correlation coefficients, but this time excluding the outliers and we obtained similar results (**Figure S2.2**).

Concluding our preliminary analysis, we proceed to the clustering.

Firstly, we decided to use SAS Enterprise Guide, in particular the High Performance (HP) Clustering using ABC with Global Peak Criterion. For the consumption clusterings, we decided to focus on the % of wine affinity (dry red: 'dryred', sweet or semi-dry reds: 'sweetred', dry white: 'drywh', sweet or semi-dry white: 'sweetwh' and port, sherry wines: 'dessert') and % of unusual wines ('exotic'). For the value and engage clustering, we used the variables 'shareWallet', 'Perdeal', 'Recency', 'LTV', 'WebPurchase', 'Age'. Considering this together with the previous preliminary analysis, we decided these variables are very relevant for our clusterings and they represent a good trade-off between information and numbers of variables.

To avoid the problem of having variables with different weights in the clustering, we standardised all of them with the min/max method. The min/max standardisation decreases the standard deviations which can suppress the effect of the outliers. We decided 10 as the number of iterations since we considered this number a good compromise between complexity of our clustering and computer performance.

For the consumption clustering, the method ABC estimated the number of clusters as 2, which cluster 1 have the centroid's mean of the products 'Dessert', 'Sweetred', 'Drywh', 'Exotic', 'Sweetwh', higher than the values of the cluster 2, and cluster 2 has the centroid's average of the product 'Dryred' higher than the value of the cluster 1 (**Table 2.2**).

In **Figure 2.3**, we can see in the Elbow plot that the angular coefficient stops growing when the number of clusters is around 5, in contrast to the result obtained from the ABC.

**Table 2.2.** Results of HP clustering with ABC criteria for Consumption clustering when the estimated number of cluster is 2. A) Cluster Summary and B) Centroid average for each standardised variable and C) Centroid average for each variable.

| A) Cluster Summary | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Distance from Cluster Centroid to Observation | | | | | | |
| Cluster | Frequency | Maximum | Minimum | Average | SSE | Standard Deviation | Nearest Cluster | Distance to Nearest Cluster Centroid |
| **1** | 4713 | 1.1157 | 0.0460 | 0.3408 | 647.8 | 0.3708 | 2 | 0.5039 |
| **2** | 5115 | 0.7010 | 0.0228 | 0.2075 | 263.1 | 0.2268 | 1 | 0.5039 |

| A) Centroid average for each standardised variable | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Obs | Iter | CustID | S_Dessert | S_Dryred | S_Drywh | S_Exotic | S_Sweetred | S_Sweetwh |
| **1** | 10 | 1 | .14481 | .29466 | .48191 | .23137 | .15082 | .18500 |
| **2** | 10 | 2 | .04013 | .69452 | .28235 | .11658 | .04222 | .04892 |
| **C) Centroid average for each variable** | | | | | | | | |
| Obs | Iter | CustID | Dessert | Dryred | Drywh | Exotic | Sweetred | Sweetwh |
| 1 | 10 | 1 | 11.1501 | 29.8762 | 36.1797 | 22.2119 | 11.3112 | 11.4702 |
| 2 | 10 | 2 | 3.0896 | 69.0625 | 21.6116 | 11.1914 | 3.1662 | 3.0329 |

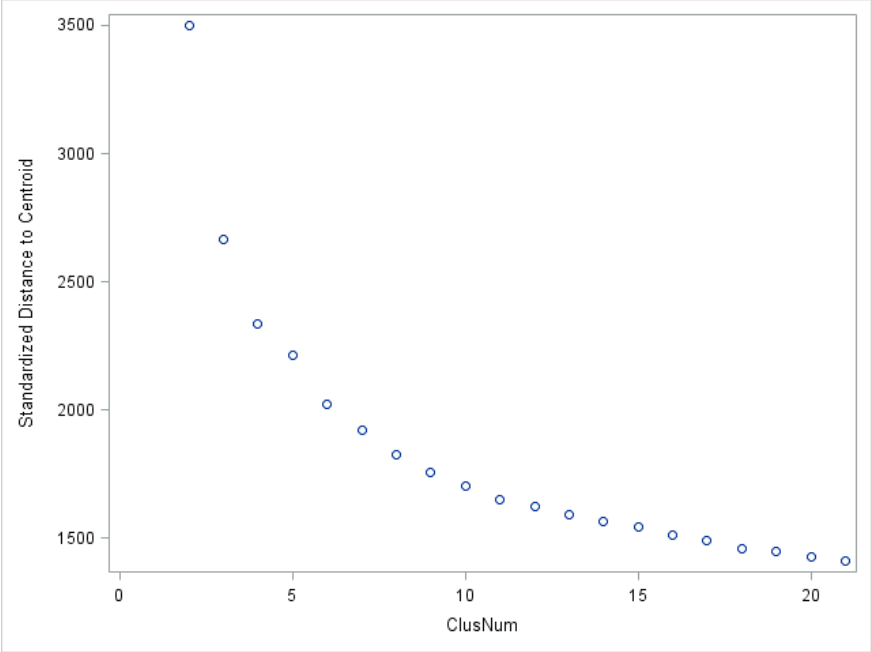


**Figure 2.3.** Elbow plot for the Consumption HP clustering.

For Value and Engage clustering, the estimated number of clusters is 3 (**Table 2.3**). From now on, we will take into consideration the non-standardised values' representation because it has the natural and logical range. The average 'shareWallet' of the centroid is higher in group 3 than in group 1 and 2 (respectively 0.015, 0.002, 0.007). 'Perdeal' average of the centroid is different between the clusters (cluster 1: 60.030, cluster 2: 20.977, cluster 3: 3.589). 'Recency' has no big differences in the average of the centroid between the group 2 (50.946) and 3 (50.454), but it is higher in group 1 (62.023). 'LTV' mean of the centroid has differences for the three groups; group 1 has -5.925, group 2 has 143.459 and group 3 has 604.737. Furthermore, 'WebPurchase' has differences in the average of the centroid for all the three groups (group 1: 56.603, group 2: 43.416, group 3: 19.762). The age characterises the three groups; group 1 has mean age of the centroid 31.508, group 2 51.663, group 3 68.588.

**Table 2.3.** Results of HP clustering with ABC criteria for Value and Engage clustering when the estimated number of cluster is 3. A) Cluster Summary and B) Centroid average for each standardised variable and C) Centroid average for each variable.

**A) Cluster Summary**

| Cluster | Frequency | Distance from Cluster Centroid to Observation | | | SSE | Standard Deviation | Nearest Cluster | Distance to Nearest Cluster Centroid |
|---|---|---|---|---|---|---|---|---|
| | | Maximum | Minimum | Average | | | | |
| **1** | 3990 | 0.9188 | 0.0384 | 0.2924 | 396.8 | 0.3153 | 2 | 0.5831 |
| **2** | 3066 | 0.8088 | 0.0491 | 0.2783 | 263.0 | 0.2929 | 3 | 0.5694 |
| **3** | 2772 | 0.7216 | 0.0481 | 0.2408 | 182.7 | 0.2568 | 2 | 0.5694 |

**B) Centroid average for each standardised variable**

| Obs | Iter | ClusID | S_shareWallet | S_Perdeal | S_Recency | S_LTV | S_WebPurchase | S_Age |
|---|---|---|---|---|---|---|---|---|
| **1** | 10 | 1 | .07505 | .61887 | .15506 | .08739 | .62623 | .22513 |
| **2** | 10 | 2 | .26004 | .21625 | .12736 | .16326 | .46923 | .56105 |
| **3** | 10 | 3 | .53959 | .03700 | .12613 | .39753 | .18764 | .84313 |

**C) Centroid average for each variable**

| Obs | Iter | ClusID | shareWallet | Perdeal | Recency | LTV | WebPurchase | Age |
|---|---|---|---|---|---|---|---|---|
| **1** | 10 | 1 | .002152 | 60.0301 | 62.0230 | -5.925 | 56.6032 | 31.5075 |
| **2** | 10 | 2 | .007085 | 20.9765 | 50.9459 | 143.459 | 43.4157 | 51.6629 |
| **3** | 10 | 3 | .014541 | 3.5894 | 50.4540 | 604.737 | 19.7616 | 68.5879 |

In this case, we can see that in the Elbow plot (**Figure 2.4**), the number of clusters reaches a *plateau* around 3.

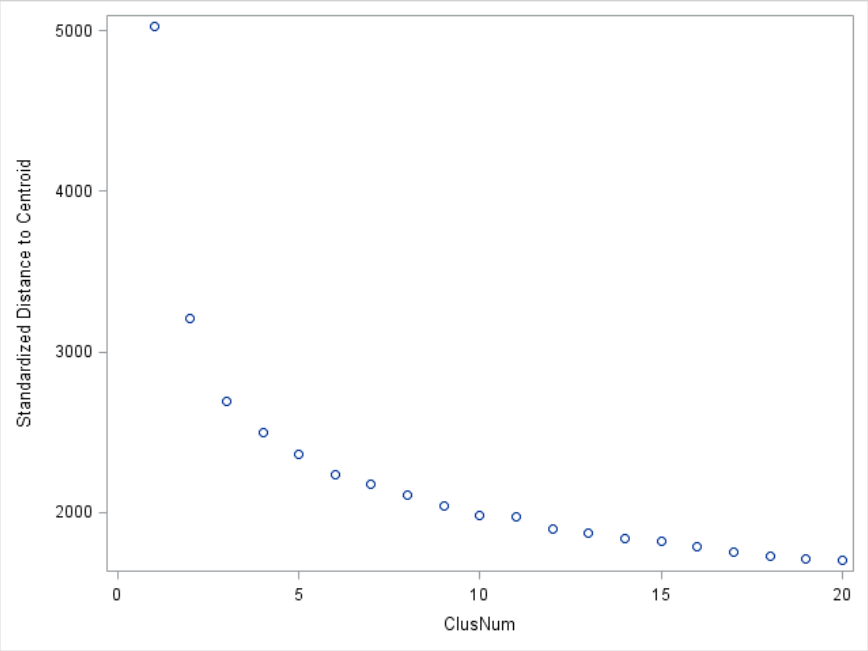


**Figure 2.4.** Elbow plot for the Value and Engage HP clustering.

To rank the variables included in the clustering, in **Figure 2.5** we calculated the maximum log 'Worth' for the three clusters. 'Dryred' has a higher value of worth, followed by 'Drywth', 'Sweetwh', 'Dessert', 'Sweetred' and 'Exotic'.

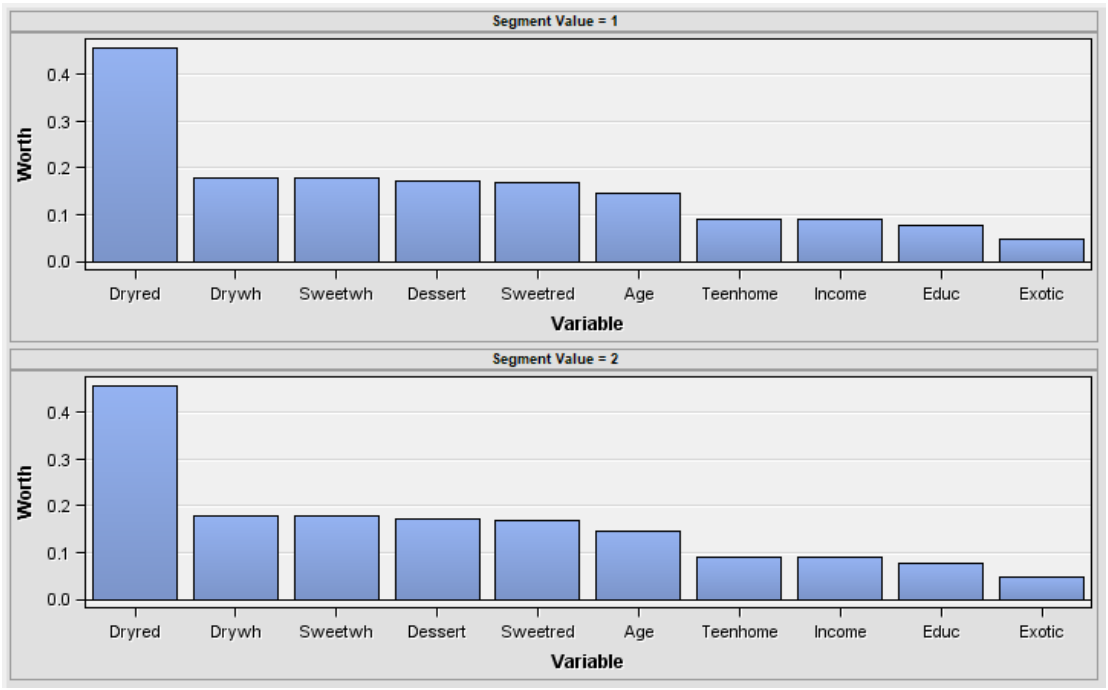


**Figure 2.5.** Worth variable for the 2 segments of the Consumption HC clustering.

After obtaining these results, we used SAS Enterprise Miner to carry out a HP Cluster and Hierarchical Clustering (HC) Cluster. The result of HP Cluster confirmed the previous analysis in SAS Enterprise Guide and since we wanted to be sure that different types of clustering arrive at the same results, we also performed HC Cluster.
For HP Cluster, as we did in SAS Enterprise Guide, we selected the min/max method to standardise the variables, the Euclidian distance as the distance measure, 10 as the numbers of iterations, 5% as the percentage of clustering change, the ABC with the Global Peak as estimation criterion and finally, principal component analysis (PCA) as the align reference dataset.
For HC Cluster, we standardised the values, once again, with min/max method and we compared the results changing the initialisation cluster seed method (random vs. PCA).
For Consumption HC clustering, we compared the clustering when we selected the number of clusters equals to 5 (**Table 2.4**) as Elbow plot indicated (**Figure 2.3**) and when we select 2 as ABC criteria previously suggested. When the number of clusters is set to 5, we can see that the first branching of the tree is related to the variable 'Dryred'.

**Table 2.4.** Consumption HC Clustering when the estimated number of clusters is 5 with random initialization method. A) Parameters per cluster, B) Importance of the variables and C) $R^2$ determination coefficient

| A) Parameters per cluster | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Segment** | **Freq** | **RMSSTD** | **Radius** | **Near** | **GAP** | **Dessert** | **Dryred** | **Drywh** | **Exotic** | **Sweetred** | **Sweetwh** |
| **1** | 767 | 0.121386 | 0.629794 | 4 | 0.41683 | 8.719687 | 37.95698 | 36.103 | 49.34029 | 8.859192 | 8.337679 |
| **2** | 3843 | 0.078491 | 0.626275 | 5 | 0.426916 | 2.527973 | 74.23784 | 18.05282 | 11.02394 | 2.566485 | 2.562581 |
| **3** | 570 | 0.158104 | 0.838339 | 4 | 0.476964 | 22.98246 | 11.16316 | 21.91228 | 51.66491 | 21.44386 | 22.46491 |
| **4** | 1477 | 0.117351 | 0.671075 | 5 | 0.31037 | 12.98375 | 25.08192 | 32.00812 | 14.50914 | 14.54909 | 15.37305 |
| **5** | 3171 | 0.084565 | 0.593988 | 4 | 0.31037 | 5.993377 | 43.99117 | 38.79376 | 9.437401 | 5.821192 | 5.398297 |

| B) Importance of the variables | | | |
|---|---|---|---|
| **Name** | **N Rules** | **N Surrogates** | **Importance** |
| **Dryred** | 13 | 2 | 1 |
| **Drywh** | 8 | 9 | 0.965987 |
| **Sweetwh** | 2 | 14 | 0.944674 |
| **Dessert** | 1 | 15 | 0.93231 |
| **Sweetred** | 0 | 14 | 0.929873 |
| **Exotic** | 6 | 7 | 0.81308 |

| C) $R^2$ determination coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Type** | **Over_All** | **Dessert** | **Dryred** | **Drywh** | **Exotic** | **Sweetred** | **Sweetwh** |
| **$R^2$** | 0.639381 | 0.46797 | 0.792272 | 0.541552 | 0.626892 | 0.475046 | 0.526448 |

In **Figure S2.2**, we can see how the data are divided into 5 groups and we can notice that the data are predominantly divided into two major groups, as ABC criteria suggested. When the prefixed number of clusters is equal to 2 (Seed initialisation method: Random **Table 2.5**, PCA **Table 2.6**), the variable 'Dryred', as we have seen before in the tree decision, is the most important variable. Comparing these results with the HP clustering (**Table 2.2**), we have the same results.

**Table 2.5**. Consumption HC Clustering when the estimated number of clusters is 2 with random initialisation method. A) Parameters per cluster, B) Importance of the variables and C) $R^2$ determination coefficient

| A) Parameters per cluster | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Segment** | **Freq** | **RMSSTD** | **Radius** | **Near** | **GAP** | **Dessert** | **Dryred** | **Drywh** | **Exotic** | **Sweetred** | **Sweetwh** |
| 1 | 5092 | 0.092315 | 0.701185 | 2 | 0.503882 | 3.048311 | 69.47977 | 21.36744 | 11.18559 | 3.098782 | 2.967793 |
| 2 | 4736 | 0.151293 | 1.116752 | 1 | 0.503882 | 11.01415 | 30.30469 | 36.11613 | 21.97149 | 11.20144 | 11.35135 |

| B) Importance of the variables | | | |
|---|---|---|---|
| **Name** | **N Rules** | **N Surrogates** | **Importance** |
| Dryred | 8 | 1 | 1 |
| Sweetwh | 0 | 15 | 0.886228 |
| Dessert | 1 | 15 | 0.884564 |
| Sweetred | 6 | 9 | 0.881868 |
| Drywh | 3 | 16 | 0.880621 |
| Exotic | 8 | 10 | 0.789288 |

| C) $R^2$ determination coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Type** | **Over_All** | **Dessert** | **Dryred** | **Drywh** | **Exotic** | **Sweetred** | **Sweetwh** |
| **$R^2$** | 0.406233 | 0.259584 | 0.700908 | 0.343017 | 0.098805 | 0.267879 | 0.279316 |

**Table 2.6**. Consumption HC Clustering when the estimated number of clusters is 2 using PCA as Seed Initialisation Method. A) Parameters per cluster, B) Importance of the variables and C) $R^2$ determination coefficient

| A) Parameters per cluster | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Segment** | **Freq** | **RMSSTD** | **Radius** | **Near** | **GAP** | **Dessert** | **Dryred** | **Drywh** | **Exotic** | **Sweetred** | **Sweetwh** |
| 1 | 5004 | 0.091397 | 0.702212 | 2 | 0.503714 | 2.993405 | 69.82694 | 21.15967 | 11.17326 | 3.04956 | 2.931455 |
| 2 | 4824 | 0.150983 | 1.120624 | 1 | 0.503714 | 10.92579 | 30.6592 | 36.0626 | 21.78752 | 11.10468 | 11.23611 |

| B) Importance of the variables | | | |
|---|---|---|---|
| **Name** | **N Rules** | **N Surrogates** | **Importance** |
| Dryred | 7 | 2 | 1 |
| Dessert | 0 | 15 | 0.888923 |
| Sweetwh | 1 | 11 | 0.88598 |
| Drywh | 6 | 9 | 0.882808 |
| Sweetred | 0 | 12 | 0.876974 |
| Exotic | 8 | 7 | 0.796068 |

| C) $R^2$ determination coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Type** | **Over_All** | **Dessert** | **Dryred** | **Drywh** | **Exotic** | **Sweetred** | **Sweetwh** |
| **$R^2$** | 0.40636 | 0.25766 | 0.70133 | 0.350572 | 0.095779 | 0.265004 | 0.274351 |

For value and engage HC clustering, we selected the number of clusters as 3 (**Table 2.7**), as Elbow plot (**Figure 2.4**) and ABC criteria suggested (**Table 2.3**).
The segment sizes are shown in **Figure S2.4**. The same results were observed before in HP clustering (**Table 2.3**).

**Table 2.7.** Value and Engage HC Clustering when the estimated number of clusters is 3 with random initialisation method. A) Parameters per cluster, B) Importance of the variables and C) $R^2$ determination coefficient

| A) Parameters per cluster | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Segment** | **Freq** | **RMSSTD** | **Radius** | **Near** | **GAP** | **Age** | **LTV** | **Perdeal** | **Recency** | **WebPurchase** | **shareWallet** |
| 1 | 3901 | 0.12776 | 0.921828 | 3 | 0.569381 | 31.32992 | -6.36965 | 60.616 | 62.1474 | 56.70879 | 0.002126 |
| 2 | 2908 | 0.107269 | 0.737153 | 3 | 0.576586 | 68.21458 | 590.9508 | 3.904402 | 50.37827 | 20.38549 | 0.014349 |
| 3 | 3019 | 0.119984 | 0.8002 | 1 | 0.569381 | 50.86519 | 131.5724 | 21.89367 | 51.21464 | 44.16496 | 0.006812 |

| B) Importance of the variables | | | |
|---|---|---|---|
| **Name** | **N Rules** | **N Surrogates** | **Importance** |
| LTV | 10 | 12 | 1 |
| Age | 9 | 12 | 0.968433 |
| shareWallet | 3 | 26 | 0.964242 |
| Perdeal | 4 | 16 | 0.947778 |
| WebPurchase | 9 | 18 | 0.945181 |
| Recency | 0 | 14 | 0.736275 |

| C) $R^2$ determination coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Type** | **Over_All** | **Age** | **LTV** | **Perdeal** | **Recency** | **WebPurchase** | **shareWallet** |
| **$R^2$** | 0.71169 | 0.785735 | 0.737656 | 0.764382 | 0.015427 | 0.657621 | 0.718347 |

The most important variable is ‘LTV’ followed by ‘Age’, ‘shareWallet’, ‘Perdeal’, ‘WebPurchase’ and ‘Recency’. We can see that LTV has also the higher value of worth (**Figure 2.6**).

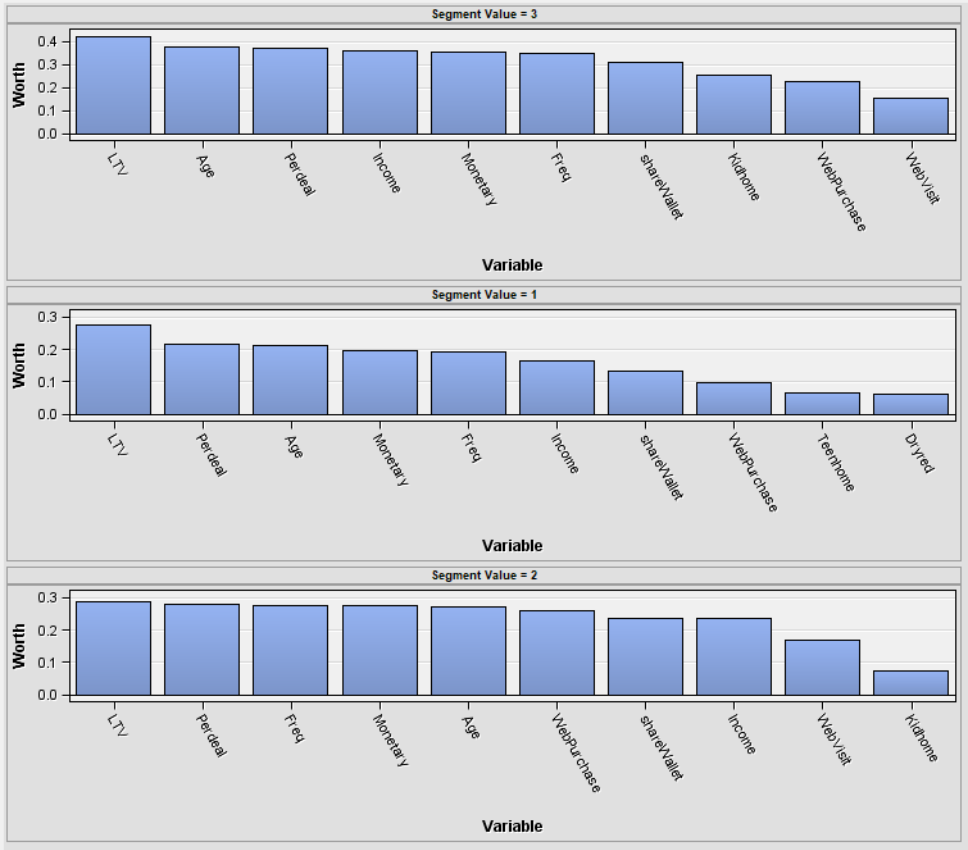


**Figure 2.6.** Worth variable for the 3 segments of the Value and Engage HP clustering.

To check the previous analysis against a new technique, we also performed the Self-Organising Map, in particular a Kohonen net. Using the Min/Max standardisation as before, Principal Component as method for the initialization of the seeds, 100 as number of neurons this algorithm has some remarkable results.
For the Consumption segmentation, we could see the following U-matrixes (**Figure 2.7-2.8**):

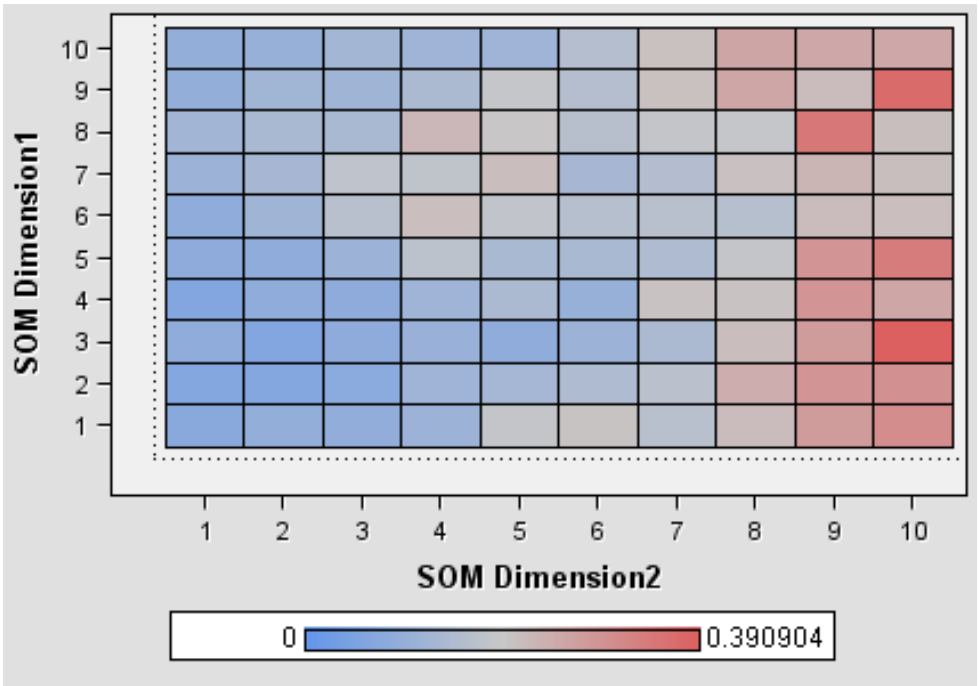


**Figure 2.7.** Distance to the nearest cluster in Consumption segmentation.

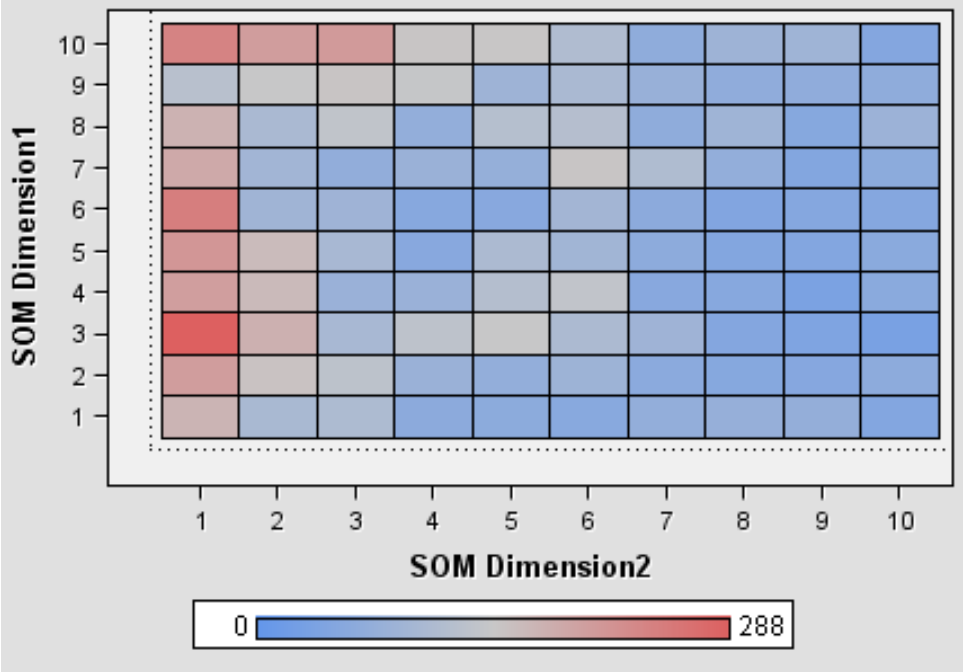


**Figure 2.8.** Frequency of Cluster in Consumption segmentation.

We can see in the west of the matrix, the neurons are distant but there is a presence of copious data points and in the centre-west, the neurons are closer to each other but the area is less populated by data points.
Instead, for the Value and Engage we had the following U-matrixes (**Figure 2.9-2.10**):

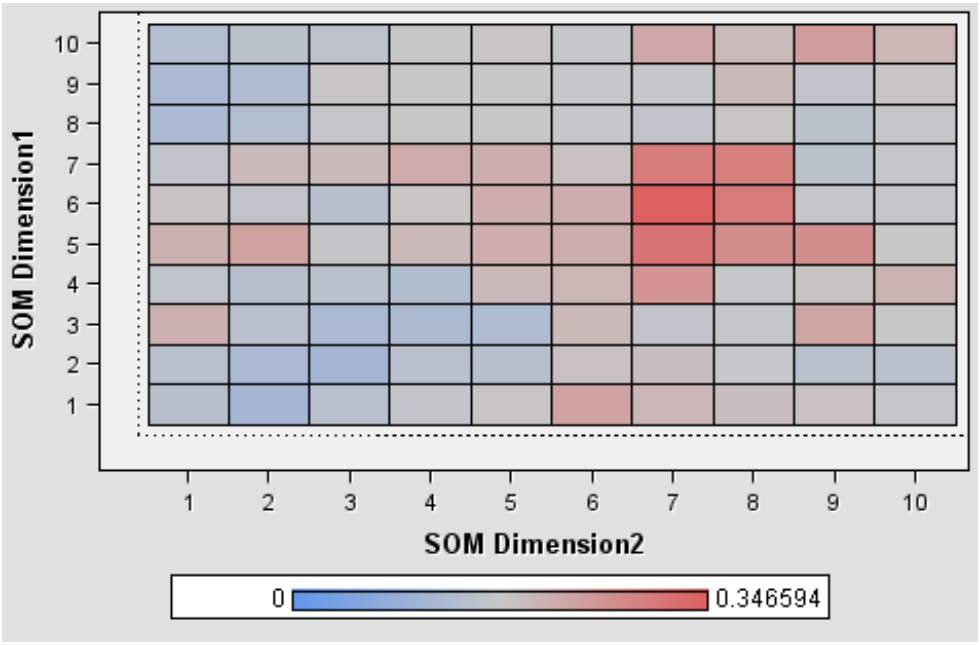


**Figure 2.9.** Distance to the Nearest Clustering in Value and Engage segmentation.

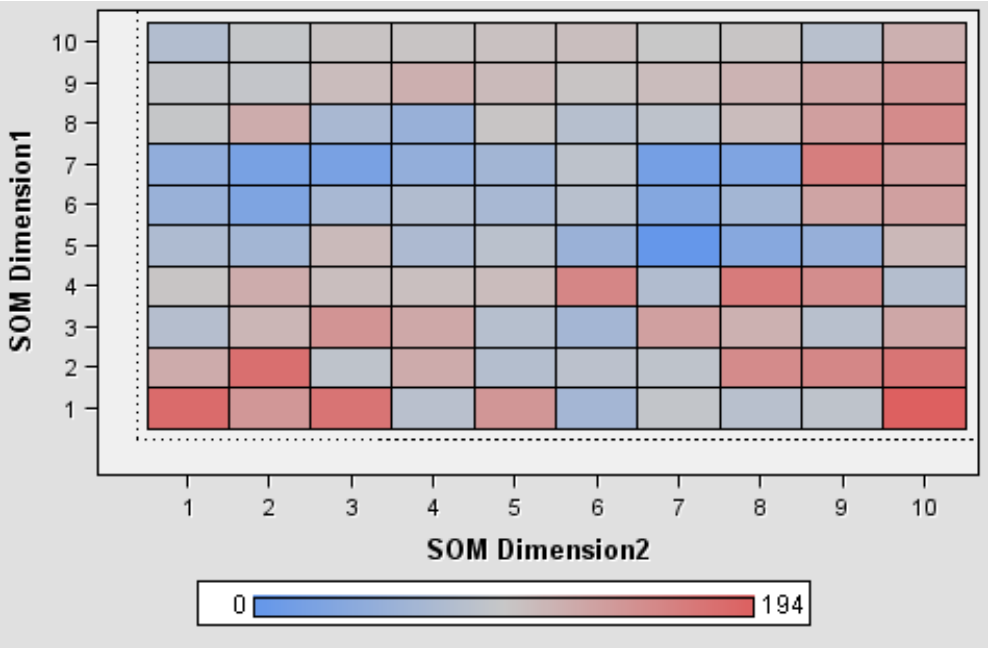


**Figure 2.10.** Frequency of Cluster in Value and Engage segmentation.

It is possible to see that in the centre-east of the matrix the distance between neurons is more pronounced and in the south-east and south-west corners it is more populated.
We decided to classify the outliers we took out at the start.
Considering HC clustering for the consumption we have the following results (**Table 2.9**):

**Table 2.9.** Classification of outliers in HC clustering for the consumption.

| Class Variable Summary Statistics | | |
|---|---|---|
| Data Role=SCORE Output Type=SEGMENT | | |
| Cluster | Frequency Count | Percent |
| SEGMENT 1 | **152** | 88.3721 |
| SEGMENT 2 | **20** | 11.6279 |

The majority (~88%) of outliers are classified in the first clustering (**Table 2.6**).
Considering HP clustering for value and engage we have the following results (**Table 2.10**):

**Table 2.10.** Classification of outliers in HP clustering for value and engage.

| Class Variable Summary Statistics | | |
|---|---|---|
| Data Role=SCORE Output Type=SEGMENT | | |
| Cluster | Frequency Count | Percent |
| CLUSTER 2 | **5** | 2.9070 |
| CLUSTER 1 | **167** | 97.0930 |

The majority (~97%) of outliers are classified in the first clustering (**Table 2.3**).

Ultimately, we decided to concatenate the results of the two segmentations: consumption and value engage (**Table 2.8**). For this, we decided to use the results of HC clustering for the consumption with the number of clusters equals to 2 and HP clustering for value and engage when the number of clusters is 3. We selected these clusterings because both methods HC and HP have a similar response.

**Table 2.8.** Concatenation between clusters from Consumption HC and Value Engage HP Clusterings.

| Clusters of Value Engage HP Clustering | Clusters of Consumption HC Clustering | | |
|---|---|---|---|
| **Frequency** | **1** | **2** | **Total** |
| **1** | 1984 | 2006 | 3990 |
| **2** | 1652 | 1414 | 3066 |
| **3** | 1368 | 1404 | 2772 |
| **Total** | 5004 | 4824 | 9828 |

Concerning the consumption clustering, we verified that 5004 customers that belong to the first group have higher percentage affinity for 'Dryred' and the remaining 4824 customers (second group) have a higher affinity for the other type of wines, particularly the 'Drywh'.

In the second clustering, the three clusters have similar customer densities. The first segment represents the youngest clients (average 32 years old) that buy more frequently online comparing to the second and third segments, however, these customers have the lowest value and engagement (low 'shareWallet' and 'LTV', high 'Recency' and 'Perdeal'). The third segment represents the best profitable customers, whom are senior (average 69 years old) and have sufficient money to indulge their passion for wine, as we can see from the high 'shareWallet' and 'LTV' and most (or all) of the sales are bought at full price. Finally, the second segment represents the intermediate level of client kind.

To conclude, we decided to establish six marketing strategies for the company development taking into account the six clusters determined before.

For the customers having lowest value and engagement, we would suggest that the company has an aggressive online marketing strategy in this case. The company should frequently send new discounts campaigns, vouchers and recommendations to improve customer's engagement. Moreover, to have a more personalised advertisement, we would filter the subject of the promotions concerning the wine affinity (dry red vs. other products).

For the third group (the best customers), we would suggest having a premium selection of product recommendations, regarding their product affinity, in the local stores to improve their recency and take the advantage of their full potential wallet,

For the intermediate category of the customers, we would follow a mixed strategy from the previous ones.

As a final remark, we would suggest to the company to invest more in improving both the quantity and quality of the dry red wines, because there is a big and committed portion of the clients that have a high affinity for these wines.

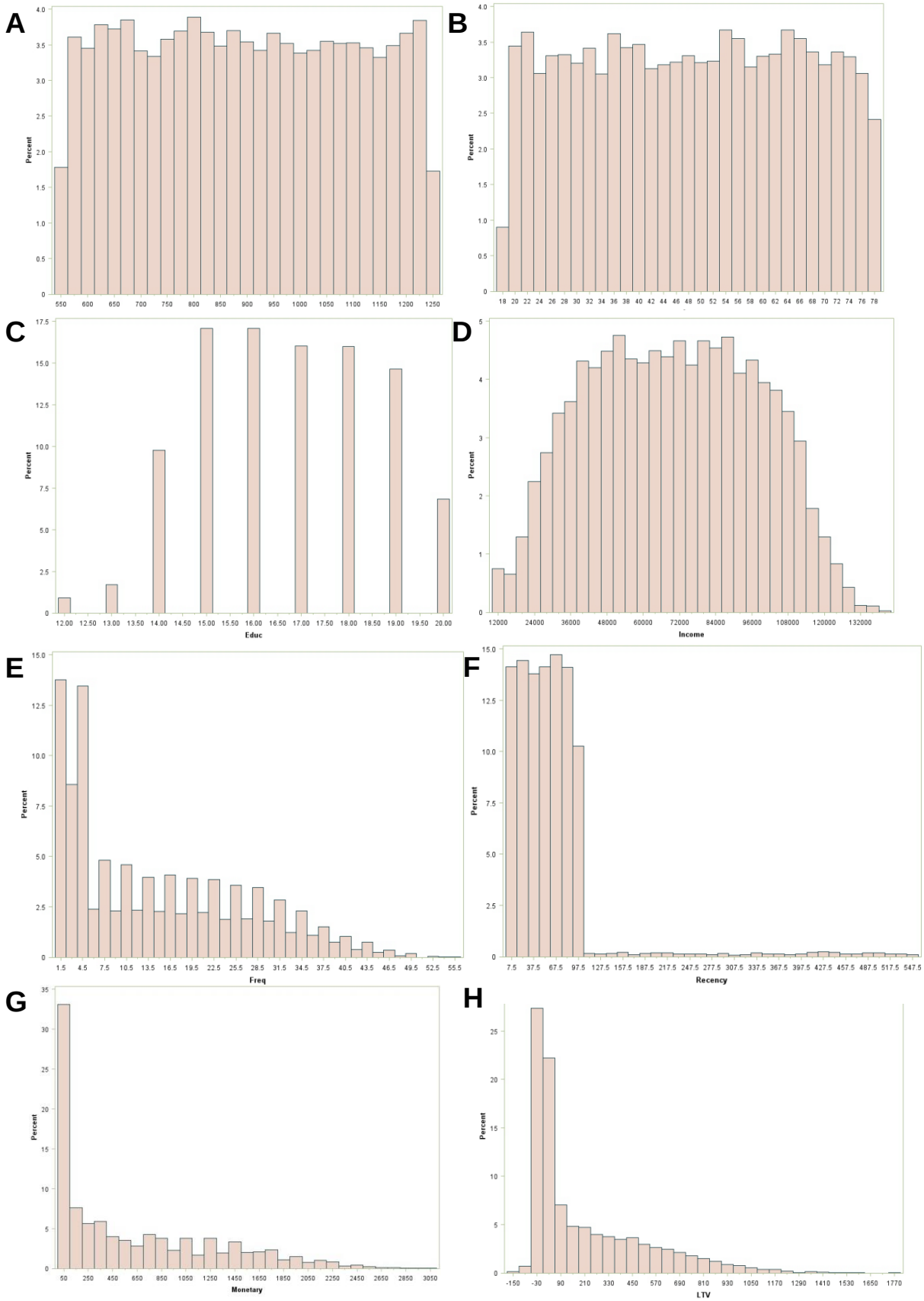


**Figure S1.1.** Histogram of the variable 'Dayswus' (A), 'Age' (B), 'Educ' (C), 'Income' (D), 'Freq' (E), 'Recency' (F), 'Monetary' (G) and 'LTV' (H).

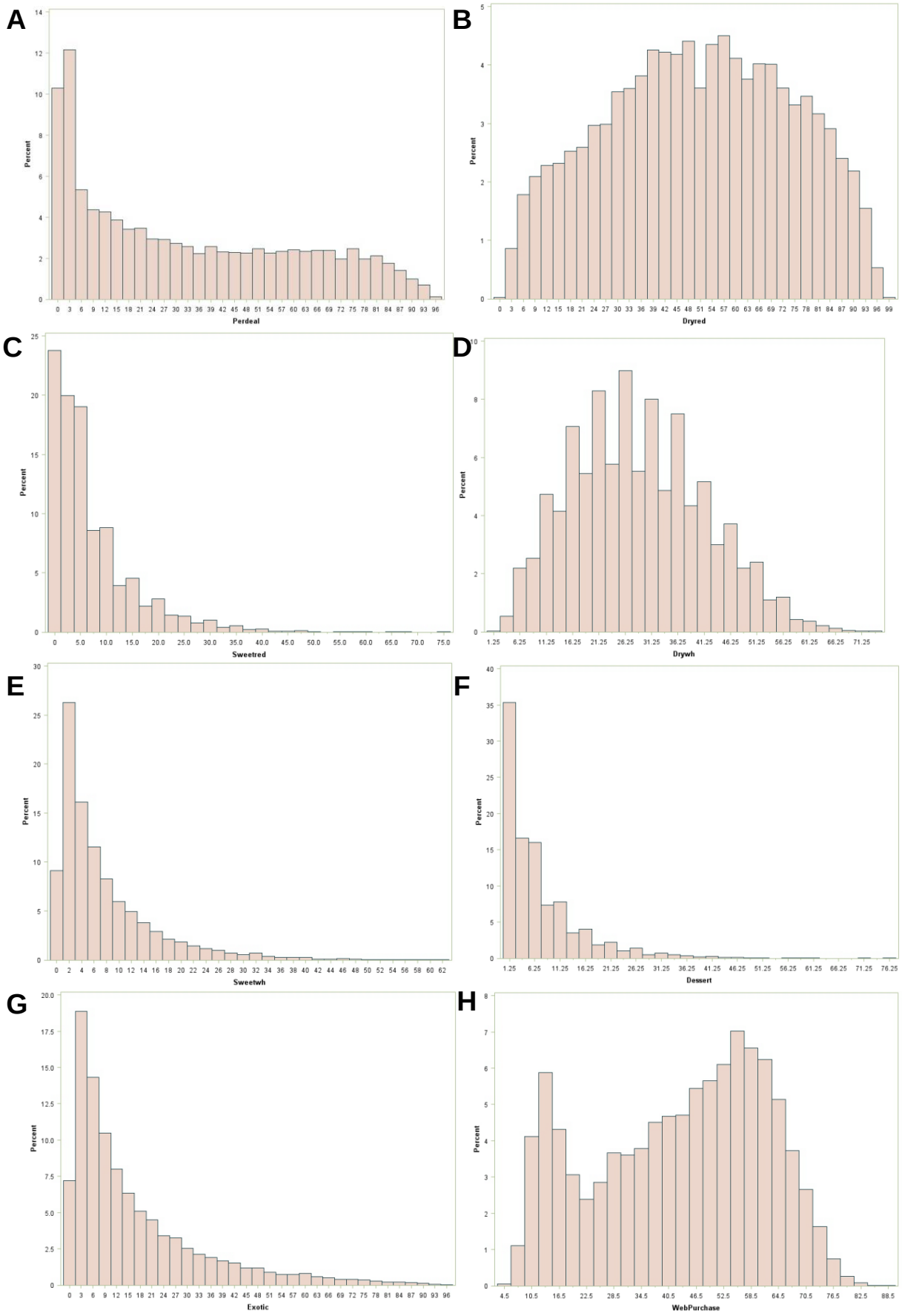


**Figure S1.2.** Histogram of the variable 'Perdeal' (A), 'Dryred' (B), 'Sweetred' (C), 'Drywh' (D), 'Sweetwh' (E), 'Dessert' (F), 'Exotic' (G) and 'WebPurchase' (H).

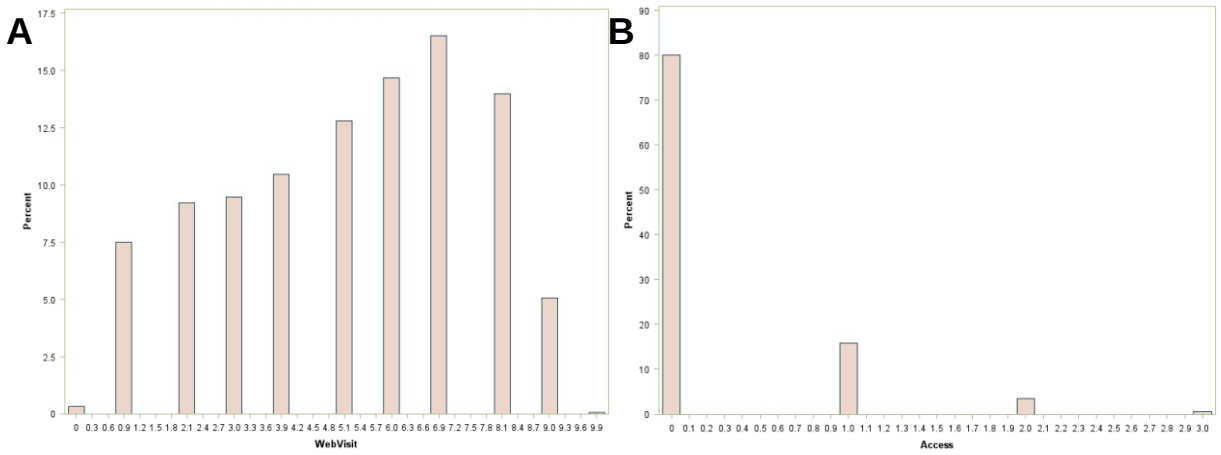


**Figure S1.3.** Histogram of the variable 'WebVisit' (A) and 'Access' (B).

**Table S1.1.** Folded F statistics.

| Newsletter | shareMedia | Variable | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|---|---|
| 0 | 0 | Freq | 4351 | 5647 | 1.01 | 0.7568 |
| | | Recency | 4351 | 5647 | 1.11 | 0.0004 |
| | | Monetary | 4351 | 5647 | 1.01 | 0.7939 |
| | | WebPurchase | 5647 | 4351 | 1.01 | 0.7179 |
| | | WebVisit | 5647 | 4351 | 1.04 | 0.1343 |
| 0 | 1 | Freq | 8134 | 1864 | 1.02 | 0.6177 |
| | | Recency | 1864 | 8134 | 1.04 | 0.2375 |
| | | Monetary | 8134 | 1864 | 1.02 | 0.5222 |
| | | WebPurchase | 8134 | 1864 | 1.01 | 0.7006 |
| | | WebVisit | 8134 | 1864 | 1.04 | 0.2928 |
| 1 | 0 | Freq | 1898 | 8100 | 1.03 | 0.3519 |
| | | Recency | 1898 | 8100 | 1.32 | <0.0001 |
| | | Monetary | 1898 | 8100 | 1.04 | 0.3122 |
| | | WebPurchase | 8100 | 1898 | 1.01 | 0.7963 |
| | | WebVisit | 8100 | 1898 | 1.04 | 0.2383 |
| 1 | 1 | Freq | 9411 | 587 | 1.01 | 0.9436 |
| | | Recency | 9411 | 587 | 2.02 | <0.0001 |
| | | Monetary | 9411 | 587 | 1.01 | 0.9408 |
| | | WebPurchase | 587 | 9411 | 1.02 | 0.7511 |
| | | WebVisit | 587 | 9411 | 1.03 | 0.5588 |

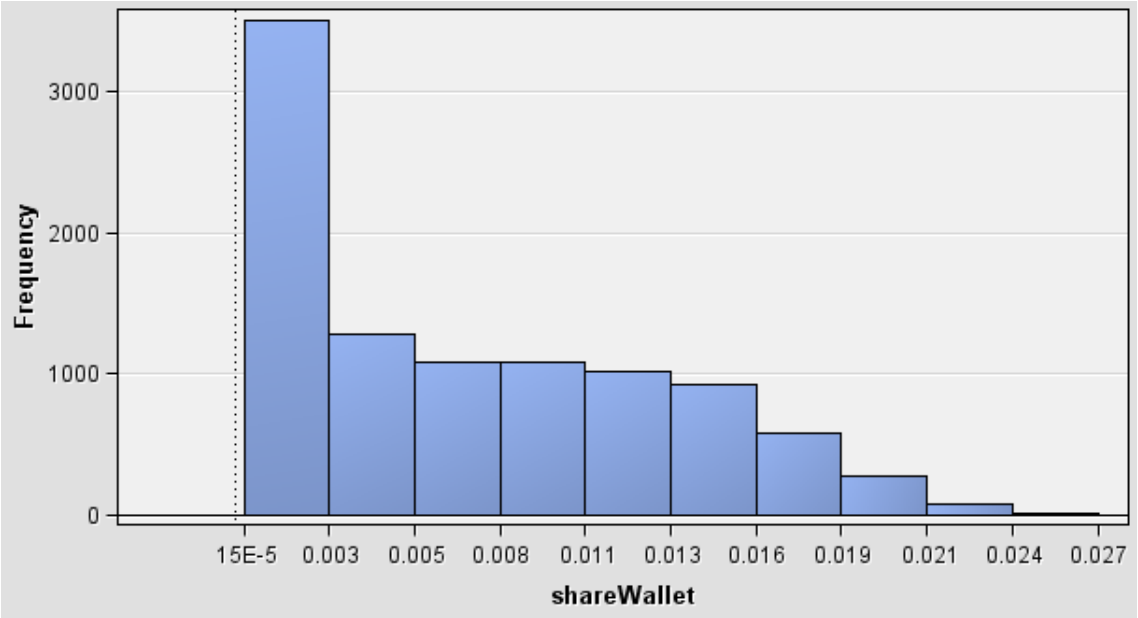


**Figure S2.1.** Histogram of the new variable 'shareWallet'.

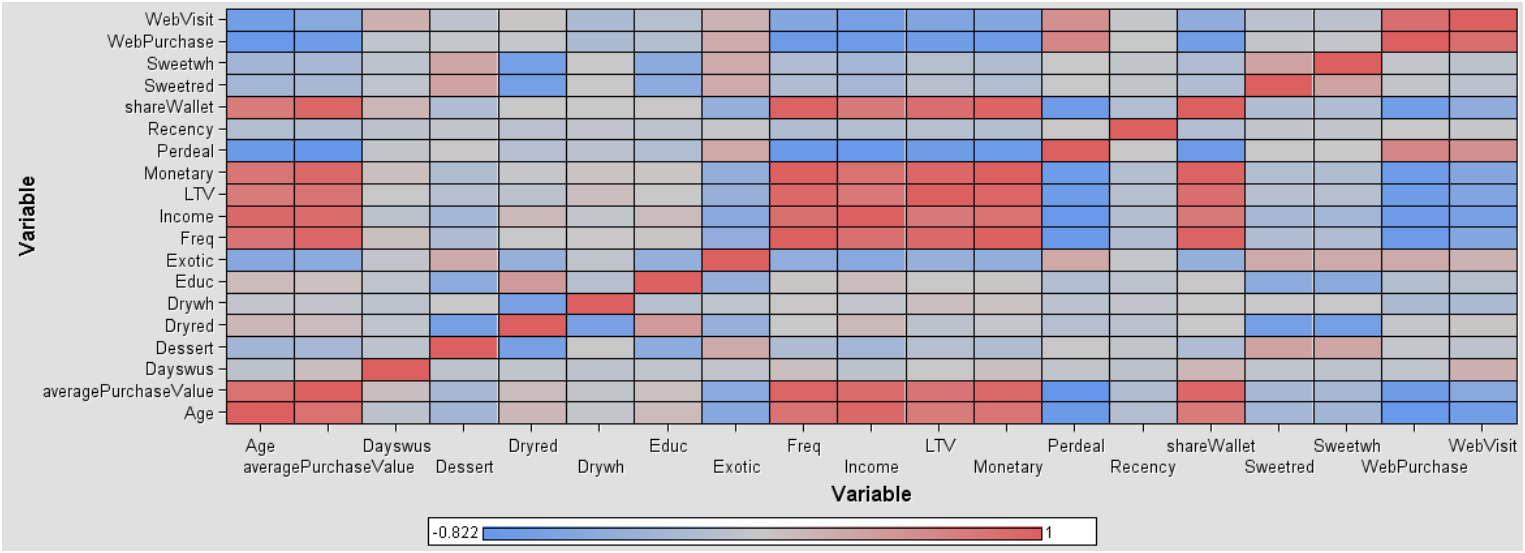


**Figure S2.2.** Matrix of Pearson correlation coefficients for the dataset after removing the outliers.

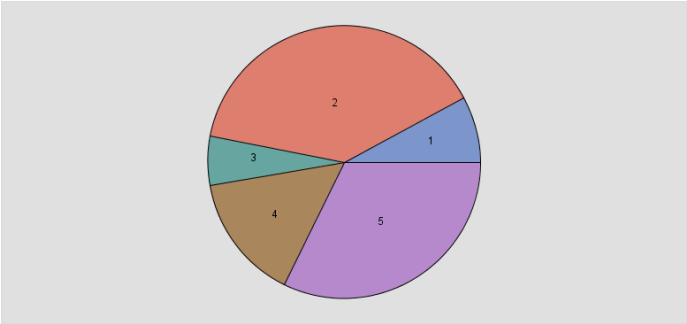


**Figure 2.3.** Segment sizes of Consumption HC Clustering**.**

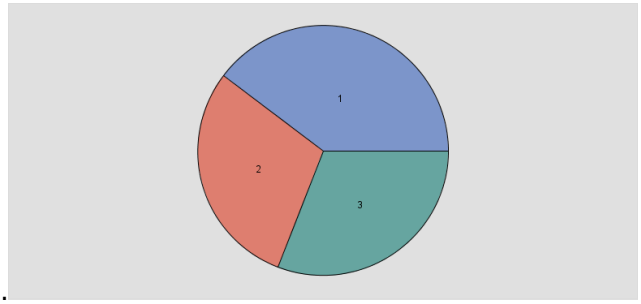


**Figure S2.4.** Segment sizes of Value and Engage HC Clustering.