Business Case 1: WWW

Master Degree Program in Data Science and Advanced Analytics

Academic Year 2021/2022

Business Cases with Data Science

Github repository:

https://github.com/andremforte/BC1_GroupV.git

Group V:

Anis Tmar (m20211157)

André Forte (m20210590)

Opeyemi Mary Akande (m20211320)

Rafael Nunes (m20210832)

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

Index

1 Business Understanding

Wonderful Wines of the World (WWW) is a 7-year-old enterprise. Its mission is to delight its customers with well-made and unique wines. Sales are realized essentially in the USA via digital and physical channels (Website/Stores).

The company has started using their database to improve sales. But so far it has no segment market. WWW aims to meet customer's needs by assessing them. By doing so, marketing campaigns will improve, focused programs will be developed and hopefully profit will increase. Our goal is to develop a two-tiered data segmentation: first one based on the value and engagement of the customers and the second on their buying behavior (types of wines they prefer or buy most). Finally, we will merge the two perspectives and segment our final clusters.

2. Data Understanding

Data Understanding involves taking a closer look at the data available for mining. This step is critical in avoiding unexpected problems during next phases. The dataset provided is a subset of Wonderful Wines of the World (WWW)'s customer database. Our data "WonderfulWinesoftheWorld.xlsx" contains 10000 rows and 18 columns (see Annex 1) which were all metric variables.

We used statistical techniques and visualizations to be able to explore and understand the different characteristics of our data. After we implemented the Pearson correlation, we found some redundant variables: "Monetary", "Freq", "LTV" and "Age", "Income" and "WebVisit", "WebPurchase" (see Annex 2). Also, from the Boxplots' analysis, the variables "Recency", "LTV", "Sweetred", "Drywh", "Sweetwh", "Dessert" and "Exotic" have few outliers.

We proceeded to check if any of the variables had null values and verified that there were none. We also noticed that there were no duplicated values in the dataset. Then, we checked the data types of each variable and found some features that were not assigned the appropriate data type. Regarding coherence check, we inspected the variables both individually, as well as in relation to each other. This analysis checks were done to shed more light on information that could be hidden in the dataset.

3. Data Preparation

3.1. Outliers' Removal

We started by visualizing the boxplots and came to the conclusion that our models would perform better after removing some outliers. We tried three different approaches of outliers'

treatment: manual filtering, IQR method and Local Outlier factor. Then, we concluded on furthering our analysis using the manual approach since it's the method that allowed us to select the critical assessed points to drop. The points dropped were values that stands out as extreme from other data points as seen on the boxplots and histograms. The conditions chosen to drop outliers are in Annexes (see Annex 3). After outliers' removal, we kept 97.8% of the data.

3.2 Feature Engineering

In this step, we wanted to study the creation of new features that would bring value to the models. Some highly correlated variables like "Monetary" and "Frequency", due to redundancy in such relationship, were used to create another variable ("Average spending Amount per visit") which would be more useful for the modelling process. After that, we dropped the old variables. The new feature was further discretized and was used for categorization purpose. Also due to the high correlation between "Age" and "Income" we decided to discretize "Age" and used it as a categorical variable. In like manner, "WebVisit" and "Perdeal", which were highly correlated with "WebPurchase", were transformed into discrete variables. After that, the previous continuous variables were dropped.

This idea of creating discrete variables were employed so as to reduce redundant information as well as to provide clearer and additional interpretations of each final cluster. Summary of the feature engineered variables are provided in Annexes (Annex 4).

After feature engineering, we had 17 variables which were divided into metric and non-metric variables. Metric variables are: "LTV", "Dayswitus", "Educ", "Income", "Recency", "Dryred", "Sweetred", "Drywh", "Sweetwh", "Dessert", "Exotic", "WebPruchase". Non-metric variables are: "AgeGroup", "DiscountGroup", "AvgAmountPerVisit" and "PerdealGroup". After this step, the heat plot showed that the high correlations had been avoided.

3.3 Data Normalization

The variables are in different units and scales. To address these issues, we transformed the variables into a more normal distribution. We performed these transformations because, in general, learning algorithms benefit from standardization of the data set.

The goal of standardization or normalization is to change the values to a common scale, without distorting differences in the ranges of values. The metric variables were scaled using StandardScaler (mean = 0, standard deviation = 1) and OneHotEncoder was used to scale the non-metric variables. The histograms, boxplot and heatmap of the variables after transformation were then visualized and we noticed that the data were cleaned and prepared to modeling.

3.4. Second Outliers' Check

Before starting the Modeling process, we wanted to assure that our dataset was clean again, so we decided to apply DBSCAN to filter any noise that could be in the dataset at this point. With this algorithm, we removed 105 observations, temporarily, which are regarded as local outliers, and this represented 1.1% of our original dataset.

4. Modeling

In the modeling part, it was decided to analyze the clusters by perspectives in order to get better insights from the data: one approach based on customer value and another focused on buying behavior.

For clustering, it was used three main algorithms. The first two were K-means and K-Prototypes. Since we transformed numerical variables into categorical, K-Prototypes may be implemented. This algorithm calculates the distance between numerical features using Euclidean distance and between categorical features using the number of matching categories. Therefore, we can combine more different variables in order to get a better customers' characterization, so that's one of the reasons we continued with K-Prototypes solutions.

To merge the clustering perspectives, we used the third algorithm, the Hierarchical Clustering, and, after that, T-SNE was applied to visualize the clusters.

Perspective 1 - Engagement (Customer Value)

Regarding this perspective, it was defined a list with the related variables. Our analysis was divided into the types of customers that provided more value in terms of profit for the company. Using K-Prototypes, three clusters were obtained: "BronzeCustomers", "SilverCustomers" and "GoldCustomers". Individuals from the first cluster, "BronzeCustomers", had the lowest income and LTV (0.85 standard deviation below the mean and 0.71 standard deviation below the mean, respectively). Individuals from the last cluster, "GoldCustomers", had the highest income and LTV (1.24 standard deviation above the mean and 1.52 standard deviation above the mean, respectively). Finally, individuals from "SilverCustomers" registered values near the average. The parallel coordinate plot and the centroids' values of this perspective can be seen in Annexes (Annexes 5 and 6).

Perspective 2 – Buying Behavior

In the second perspective, the selected variables corresponded to the types of wine they could buy. In addition, we included the "Age" variable in order to start understanding some profiles.

The K-Prototypes Solution provided us three clusters: “DryredWine”, “DrywhWine” and “SweetWineLovers”. The first cluster includes customers that prefer Dryred Wine (0.98 standard deviation above the mean). Customers from “DrywhWine” tend to prefer “Drywh Wine (0.78 standard deviation above the mean). Finally, “SweetWineLovers” corresponds to the individuals who really like all types of sweet wine (SweetRed, SweetWh, Exotic and Dessert). The parallel coordinate plot and the centroids’ values of this perspective can also be seen on Annexes (Annex 7 and 8).

Merged Perspective – Final Solution

We merged the two K-Prototypes perspectives using Hierarchical Clustering, Using visual analysis and interpretation of the centroids to understand which was the best number of clusters to extract, we came to a conclusion of 4 clusters. The clusters’ characterization was based on partial analysis, with different features, and on previous knowledge from the individual perspectives.

After Clustering process, we imputed the removed outliers, from DBSCAN, and classified them using DecisionTreeClassifier. With this algorithm, we could assign to them a specific cluster for each perspective.

4.1. Clusters’ Characterization

Our final segmentation had 4 different groups of customers:

“DryredSilverCustomers” – customers with high preferences for the Dryred Wine (1 standard deviation above the mean). They registered values for income and LTV near the average. In this cluster, 70% of the individuals have ages between 38-57 and 63% of them visited our website more than 5 times per month.

“DrywhGoldCustomers” – in this group, we see the customers that prefer Drywh Wine (0.49 standard deviation above the mean). The individuals from this cluster registered the highest difference from the mean regarding Income and LTV (1.01 standard deviation above the mean and 1.07 standard deviation above the mean). In this cluster, 90% of the individuals have ages between 58-78. According to the “WebPurchase” variable, we can understand that these individuals don’t tend to buy wine online (1.05 standard deviation below the mean).

“SweetWineGoldCustomers” – 66% of these individuals have ages between 58-78. This cluster includes customers with high income and high LTV, compared to the mean (0.67 standard deviation and 0.45 above the mean). They prefer Sweet Wine – Sweetred, SweetWh, Exotic, Dessert. Like the “DrywhGoldPCustomers”, they don’t tend to buy wine on our website (0.86 standard deviation below the mean).

“SweetWineBronzeCustomers” – this is cluster is constituted by 88% of individuals with ages between 18-37. These customers have the lowest income and LTV (1.09 standard deviation and 0.72 standard deviation below the mean). It’s important to notice that 84% of these individuals had more than 32% of their purchases on discount. Also, they tend to buy more online (0.83 standard deviation above the mean).

All of the important visualizations used to establish these patterns and our final clusters’ profiles can be found on Annexes (Annexes 9 to 16).

5. Business Applications

With customers’ segmentation, it’s possible to have a better understanding of their profiles, which is helpful to define business approaches.

We noticed that there was one cluster (“SweetWineBronzeCustomers”) that registered more webpurchases and most of them were on discount. So, our suggestion is to create a loyalty program based on previous online purchases to increase LTV and reduce the number of days between purchases (Recency).

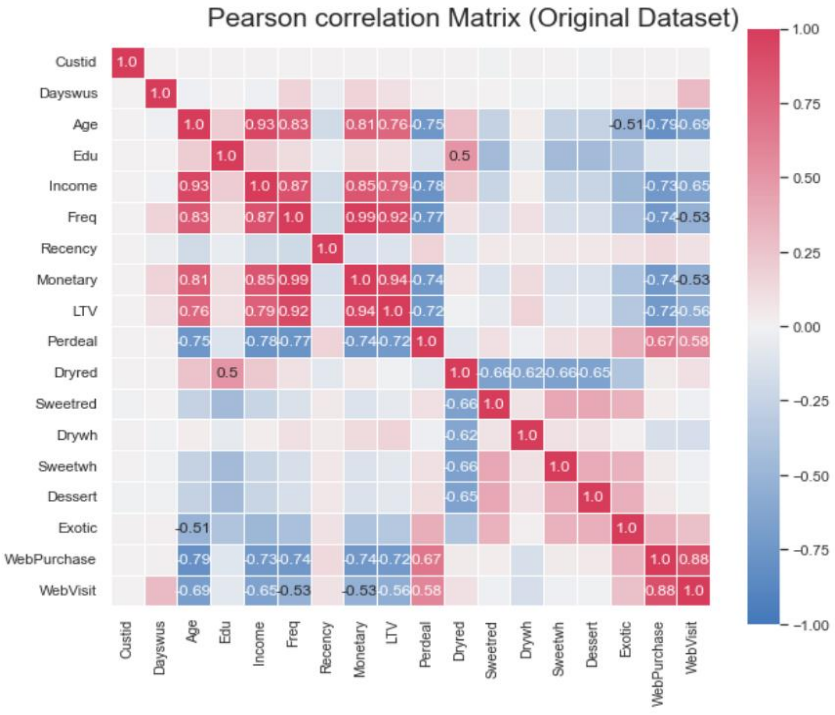
For clusters “DryredSilverCustomers”, “DrywhGoldPCustomers” and “SweetWineGoldCustomers”, we could create a discount campaign where we offer an extra bottle of different wine when they buy three bottles of their favorite type of wine. It could be one time per month. This campaign could be an opportunity to our customers try different types of wine they don’t prefer and, maybe, increase consumption by buying more products, since they had high incomes.

For the “SweetWineGoldCustomers” and “DrywhGoldPCustomers” customers, we could offer an experience per year (trip to wine’estate, for instance) to award our three best customers.

Annexes

Variables	Description
CUSTID	Unique Identification of the customer
DAYSWUS	Number of days as customer
AGE	Customer's age
EDUC	Years of education
INCOME	Household income
FREQ	Number of purchases in last 18 mo
RECENCY	Number of days since last purchase
MONETARY	Total sales of customer in last 18 mo
LTV	Lifetime value customer
PERDEAL	% Purchases bought on discount
DRYRED	% Of dry red wines
SWEETRED	% Of sweet red wines
DRYWH	% Of dry white wines
SWEETWH	% Of sweet white wines
DESSERT	% Of dessert wines
EXOTIC	% Of unusual wines
WEBPURCH	% Of purchases made on website
WEBVISIT	Average visits to the website per months

Annex1: Description of the variables (initial dataset)



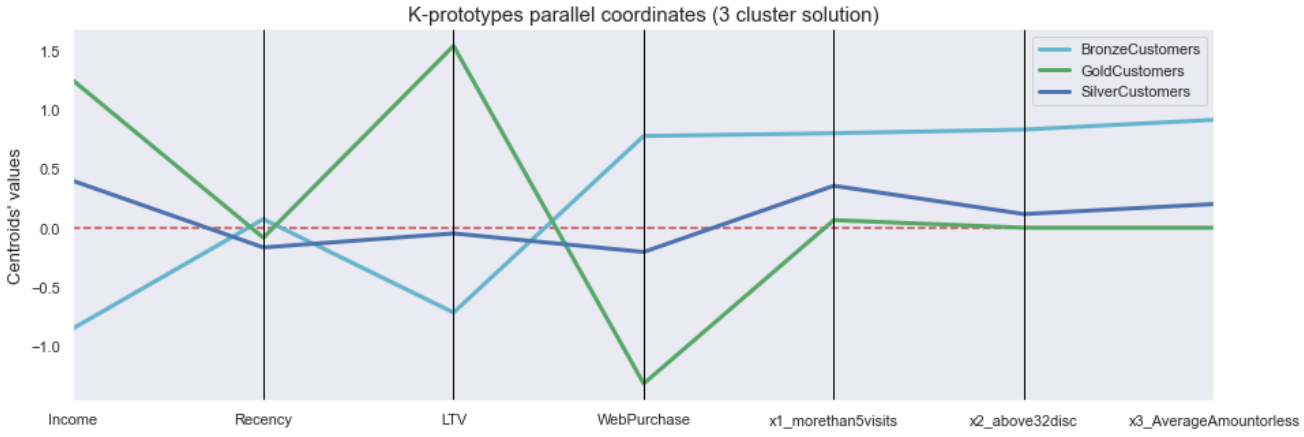
Annex2: Pearson Correlation Matrix (Initial Dataset)

Feature	Filter
Recency	<=400
LTV	<=1500
Sweetred	<=50
Drywh	<=80
Sweetwh	50
Dessert	<=50

Annex 1: Outliers' manual filtering

Name	Information Provided	Formula
AgeGroup	Divided "Age" into 3 groups	18-37, 38-57, 58-78
NumberWebVisit	If the number of visit is less or more than the average value =5.	WebVisit>5 Or WebVisit<=5
DiscountGroup	The % of customer purchases that were bought on discount, if it's more or less than 32% (the mean value)	Perdeal<=32 Or perdeal>32
AvgAmountPerVisit	Average amount spent per visit in the last 18 months (if it's less or more than the average= 32)	AvgAmountperVisit<=32 Or AvgAmountperVisit> 32

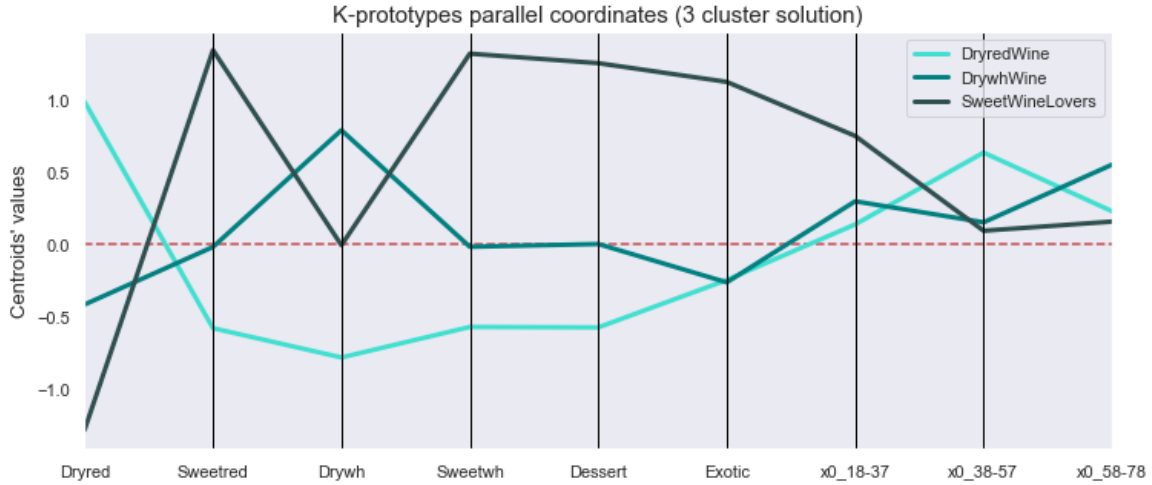
Annex 4: Feature Engineering



Annex 5: First customer Perspective (parallel coordinates)

labels	Income	Recency	LTV	Web Purchase	x1_more-than5visits	x2_above 32disc	x3_Average-Amount orless
Gold Customers	1.247463	-0.08477	1.541064	-1.321119	0.064458	0	0
Silver Customers	0.396725	-0.167316	-0.048271	-0.205591	0.355631	0.1157	0.200683
Bronze Customers	-0.853099	0.073166	-0.719322	0.777523	0.801198	0.832039	0.915021

Annex 6: First Perspective: customer Value (centroids' values)

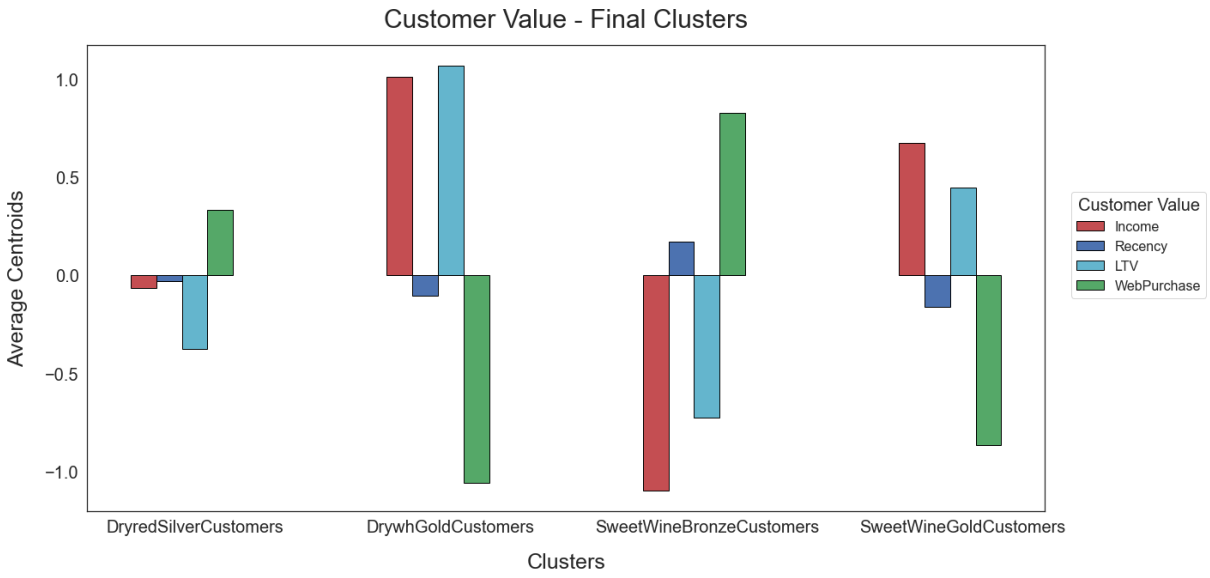


Annex 7: Second Perspective: buying behavior (centroids' values)

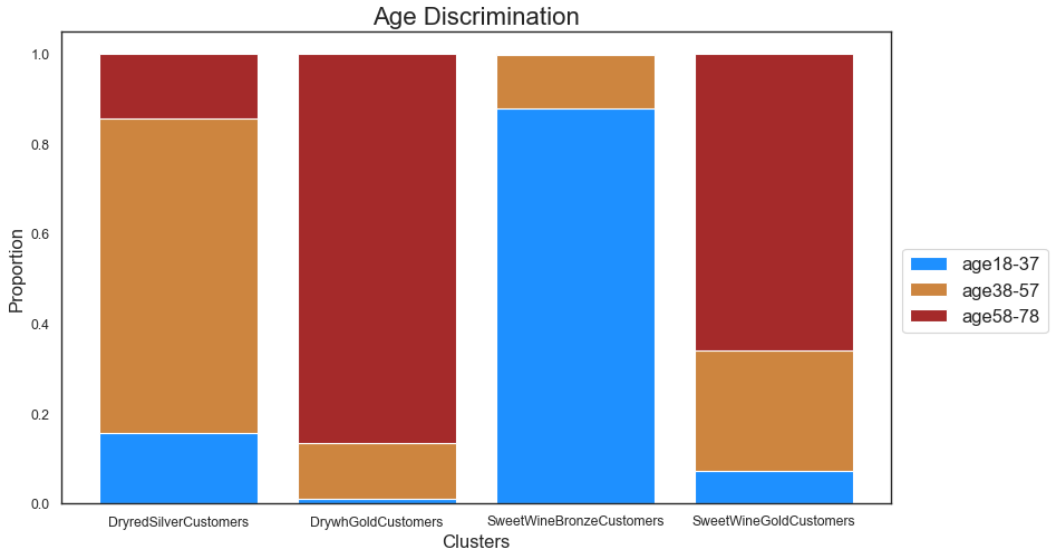
labels	Dryred	Sweet red	Drywh	Sweetwh	Dessert	Exotic	x0_18-37	x0_38-57	x0_58-78
DryredWine	0.984325	-0.577352	-0.779305	-0.568846	-0.572089	-0.245334	0.138008	0.633142	0.228850
DrywhWine	-0.414541	-0.017337	0.787339	-0.016350	0.002945	-0.261524	0.297392	0.152457	0.550150
SweetWine-Lovers	-1.279086	1.339771	-0.005113	1.315475	1.249752	1.120808	0.748210	0.094272	0.157518

Annex 8: Second Perspective: buying behavior (centroids' values)

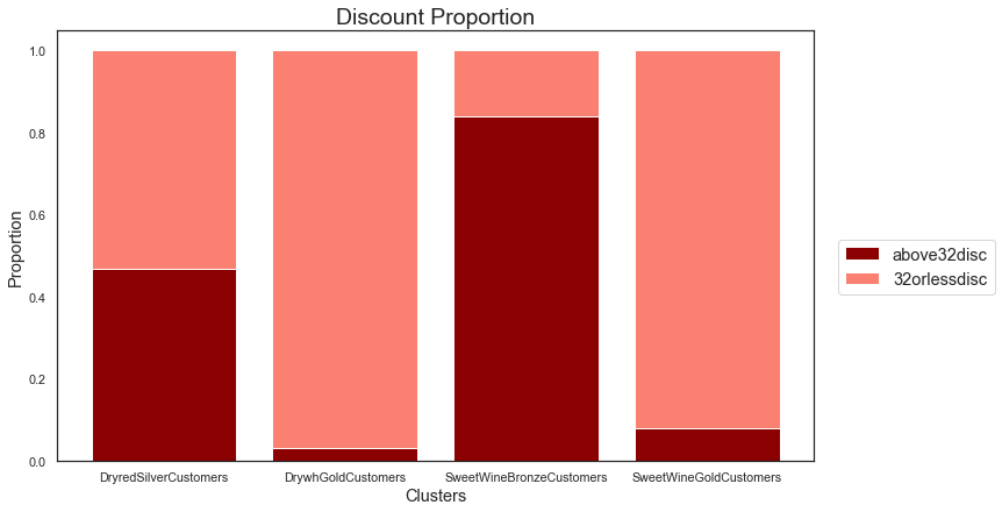
Annex 9: Final Clusters: wine types



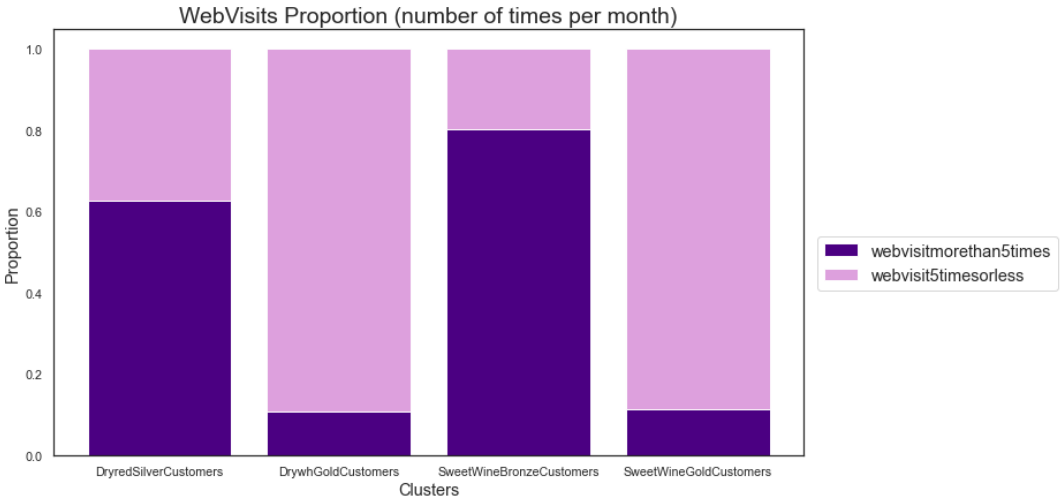
Annex 10: Final Clusters: customer value



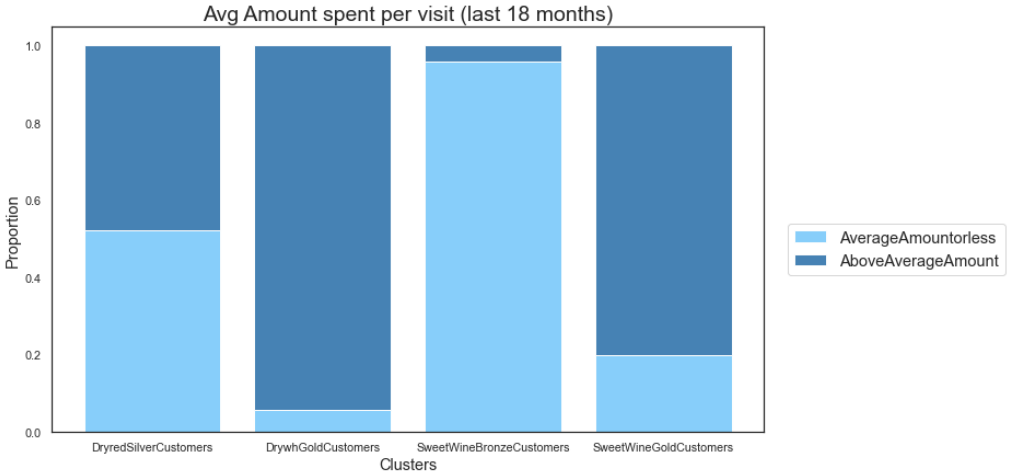
Annex 11: Final Clusters: Age groups



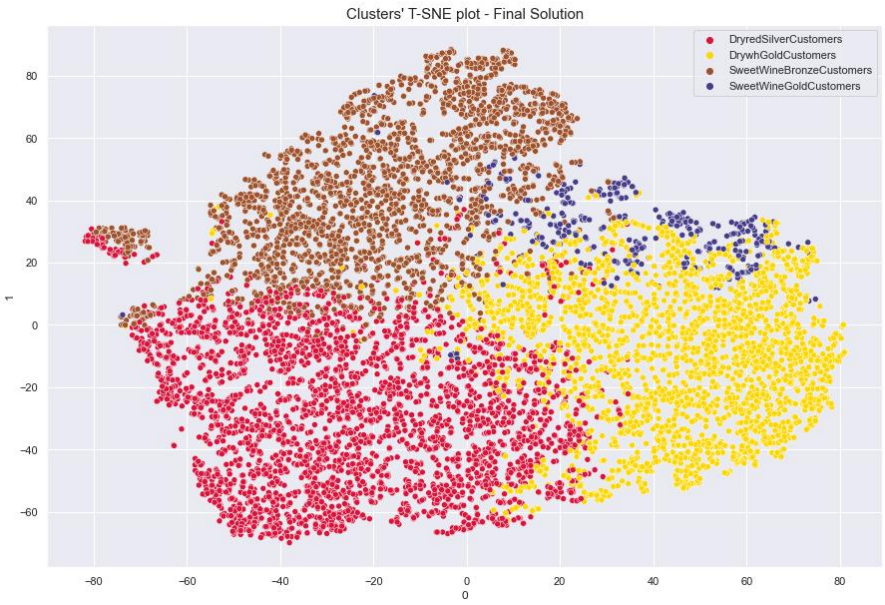
Annex 12: Final Clusters: discount groups



Annex 13: Final Clusters: WebVisit groups



Annex 14: Final Clusters: Avg Amount per visit proportion



Annex 15: t-SNE plot: final clusters

Labels	Income	Recency	LTV	Dryred	Sweet red	Drywh	Sweetwh	Dessert	Exotic	Web Puchase
DryredSilver Customers	-0.064593	-0.029273	-0.376775	1.000861	-0.580874	-0.803307	-0.572919	-0.575812	-0.195283	0.332364
DrywhGold Customers	1.012566	-0.104712	1.067991	-0.225701	-0.055221	0.499921	-0.046401	-0.031956	-0.499161	-1.058194
SweetWine-Bronze Customers	-1.095379	0.172028	-0.725409	-0.885730	0.605576	0.518602	0.618497	0.622102	0.737355	0.830722
SweetWine-Gold Customers	0.675908	-0.162023	0.446210	-1.017424	1.346266	-0.226353	1.117840	1.009732	0.308327	-0.864138

Labels	age18-37	age38-57	age58-78	Webvisitmore than5times	above32disc	AverageAmout orless
DryredSilverCustomers	0.154989	0.700673	0.144339	0.627803	0.468610	0.521861
DrywhGoldCustomers	0.009333	0.123667	0.867000	0.108000	0.030000	0.056667
SweetWineBronzeCusto-mers	0.878248	0.119260	0.002492	0.801353	0.840869	0.959416
SweetWineGoldCustomers	0.070175	0.268170	0.661654	0.112782	0.080201	0.197995

Annex 16: Final Merged Perspective (centroids' values)