



DESCRIPTIVE DATA MINING COURSE'S PROJECT REPORT

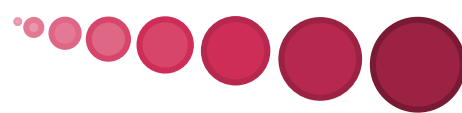
Professor Fernando Bação

Group 049

Raimundo Mujica Costa M20221342






Simão Pedro Acciaioli Gouveia Dos Santos M20220561

Tomás Caldeira Cardoso Soares Esteves M20220377



Abstract

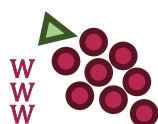
In this report, the group describes all the steps and explains the processes to realize a segmentation for both Wonderful Wines of the World Client's Profile and their Buying Behaviour, namely:

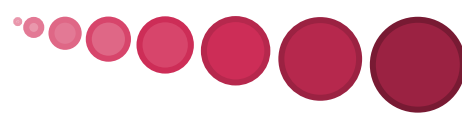
-  a check on the dataset's coherence and its visualization,
-  the process of selecting the crucial variables to enter the segmentation process,
-  the preparation of the data where we unveiled some of the data's problems and try to minimize their impact on the following clustering process,
-  the segmentations assembly itself,
-  and a segmentation polishing, so the outcomes of the segmentation are more evident to a practical marketing approach.

In the end we'll have reached a segmentation of 10 groups that, on a later stage were combined to be the target groups of marketing campaigns.

Table of Contents

Abstract	1
Introduction	3
The Company	3
Introduction to Wonderful Wines of the World (WWW) company	3
Business Model	3
Company's Problem Scenario	3
Methodology	4
Variable Interpretation	5
Variables Discernment	5
New Variables	6
Variables' Group Division	6
Data Integrity and Requisites Confirmation	6
Resolution on Data Missing Values and Duplicate Rows	6
Resolution on Data Incoherent Records	6
On clients' age	6
On clients' purchase recency	6
On clients' wine taste	7
Data's First Statistics and Data Visualization	7
Clients' Profile dataset	7
Buying Behaviour dataset	8
Data Pre-Processing	11
Resolution on Atypical Non-Standard Values (Outliers)	11
Data Normalization	13
Clustering Analysis	15





Client Profile Segmentation	16
Buying Behaviour Segmentation	17
Client Segmentation Explanation and Marketing Approach Suggestions	18
Appendix	20

Table of Illustrations

FIGURE I VARIABLES' GROUPS	6
FIGURE II CLIENTS' PROFILES_ PEARSON MATRIX FOR VARIABLES' CORRELATION	8
FIGURE III CLIENTS' PROFILE_ TEENHOME BOXPLOTS	8
FIGURE IV BUYING BEHAVIOUR_ PEARSON MATRIX FOR VARIABLES' CORRELATION	9
FIGURE V BUYING BEHAVIOUR_ KIDHOME BOXPLOTS	10
FIGURE VI BOXPLOT OF LTV VARIABLE	11
FIGURE VII BOXPLOT OF SWEETRED VARIABLE	12
FIGURE VIII BOXPLOT FOR SWEETWH VARIABLE	12
FIGURE IX BOXPLOT FOR DESSERT VARIABLE	12
FIGURE X CLIENTS' PROFILE_ DENDOGRAM FOR AVERAGE LINKAGE	16
FIGURE XI CLIENTS' PROFILE_ HEATMAP	16
FIGURE XII CLIENTS' PROFILE_ RADAR PLOT	16
FIGURE XIII BUYING BEHAVIOUR_ INERTIA PER NUMBER OF CLUSTERS	17
FIGURE XIV BUYING BEHAVIOUR_ RADAR PLOT	17
FIGURE XV BUYING BEHAVIOUR_ HEATMAP	17
FIGURE XVI PURCHASE COMPOSITION VARIABLES' HISTOGRAMS	20
FIGURE XVII CLIENTS' PROFILE_ VARIABLES' PAIRWISE RELATIONS	20
FIGURE XVIII CLIENTS' PROFILE_ KIDHOME VARIABLE IMPACT	21
FIGURE XIX BUYING BEHAVIOUR_ VARIABLES' PAIRWISE RELATIONS	21
FIGURE XX BUYING BEHAVIOUR_ TEENHOME VARIABLE IMPACT	22
FIGURE XXI CLIENTS' PROFILE_ VARIABLES' HISTOGRAMS	22
FIGURE XXII CLIENTS' PROFILE_ VARIABLES' BOXPLOTS	22
FIGURE XXIII BUYING BEHAVIOUR_ VARIABLES' HISTOGRAMS	23
FIGURE XXIV BUYING BEHAVIOUR_ VARIABLES' BOXPLOTS	23
FIGURE XXV CLIENTS' PROFILE_ HISTPLOTS FOR OUTLIERS DETECTION	23
FIGURE XXVI CLIENTS' PROFILE_ LTV VARIABLE HISTOGRAM	23
FIGURE XXVII BUYING BEHAVIOUR_ HISTPLOTS FOR OUTLIERS DETECTION	24
FIGURES XXVIII BUYING BEHAVIOUR_ HISTOGRAMS FOR WYNE TYPES	24
FIGURE XXIX CLIENTS' PROFILE_ HISTOGRAMS FOR VARIABLES AFTER MINMAX NORMALIZATION	24
FIGURE XXX CLIENTS' PROFILE_ HISTOGRAMS FOR VARIABLES AFTER STANDARD NORMALIZATION	25
FIGURE XXXI CLIENTS' PROFILE_ HISTOGRAMS FOR VARIABLES AFTER Z-SCORE NORMALIZATION	25
FIGURE XXXII BUYING BEHAVIOUR_ HISTOGRAMS FOR VARIABLES AFTER MINMAX NORMALIZATION	25
FIGURE XXXIII BUYING BEHAVIOUR_ HISTOGRAMS FOR VARIABLES AFTER STANDARD NORMALIZATION	26
FIGURE XXXIV BUYING BEHAVIOUR_ HISTOGRAMS FOR VARIABLES AFTER Z-SCORE NORMALIZATION	26
FIGURE XXXV CLIENTS' PROFILE_ VARIABLES' DISTRIBUTION PER CLUSTER	26
FIGURE XXXVIII CLIENTS' PROFILE_ CLUSTERS' 3D VIEW	27
FIGURE XXXIX BUYING BEHAVIOUR_ CLUSTERS' 3D VIEW	28



Introduction

Data mining is a critical component of modern business, enabling organizations to extract valuable insights and patterns from vast datasets. These insights can be leveraged to inform decision making, optimize processes, and ultimately enhance performance. The present report outlines the data mining process utilized to achieve two distinct segmentations.

Segmentation is a vital technique in data mining, allowing businesses to partition their customer or market base into smaller, more homogeneous groups based on shared characteristics. This can be accomplished through either supervised or unsupervised methods. Supervised segmentation relies on pre-labelled data and seeks to classify individuals into predefined groups. Unsupervised segmentation, on the other hand, involves grouping individuals based on common characteristics identified through the data mining process. By implementing segmentation techniques, businesses can customize their strategies and more effectively serve their target audiences, ultimately leading to improved performance.

The Company

Introduction to Wonderful Wines of the World (WWW) company

Wonderful Wines of the world is a ten-year-old enterprise that seeks out small, unique wineries worldwide and brings its wines to its customers. Very recently they have announced that they are expanding their range to over one hundred different premium spirits from around the world. WWW's mission is to provide its clients with all kinds of wines from all over the world. The wine selling company has amassed a total amount of three hundred and fifty thousand different entries to its membership program.

Business Model

WWW's business model is based on a mass market approach. Namely, the marketing department's approach is not different.

Wonderful Wines of the World is selling its products through various platforms: website, mobile app, catalogues and even ten physical stores around the United States of America.

Company's Problem Scenario

As earlier introduced, WWW's marketing campaigns have zero differentiation on its customer's profile and preferences.

The problem here is that all customers get the same catalogue which gives no opportunity to cross-sell any products or to even identify its main target type clients. Therefore, the company's operations are not as effective as they could be.

WWW intends to get a smarter and more profitable approach on its customer's data by starting to differentiate customers and developing more focused programs like loyalty programs and more focused offerings.

WWW needs to do a client segmentation based on vital business criteria. The organization supplies a random sample of one hundred thousand of its client's personal profile and their acquisition historic. Wonderful Wines of the World aspires to get its hands on an evaluation on the value of the different type of clients they have, besides understanding their customers' buying behaviours. This simple "Get-to-Know Your Customer" initiative can make a stark difference on Wonderful Wines of the World sales.





Methodology

“Cogito, ergo sum (I think, therefore I am)”

(René Descartes, 1637)




Being certain of the problem's questions we need to address; we developed a working task procedure to ensure we reach our goals.

Firstly, we realised that our eyesight should prior the dataset. It is crucial to understand the different variables and data types within the data sample that Wonderful Wines of the World supplies to our Data Analysis team. It is important to consider that by combining the variables that currently exist in our database, new useful variables to our analysis might appear.

We will also take some time to make sure the data is structured correctly: we have the best interest to work with data coherent to the project's goals. Also, we need to clarify which are the variables truthfully important to answer WWW's needs.

Only after understanding “what we have in hands”, sort of speaking, the first approach on a true analysis can take place.

As previously mentioned, Wonderful Wines of the World intends to segment its client's portfolio. For that reason, we will do a cluster analysis. To achieve a more synthetized and faithful customer's breakdown, it is mandatory that we prepare the earlier identified data. This preparation consists of the following steps:

-  Removing atypical clients' record that show easily distinguishable, non-standardized behaviours (outliers). By doing this the risk of skewing the analysis drops.
-  Checking if there is any variable that might be discretized, so the cluster analysis runs more smoothly.
-  Normalizing the variables so drastic range differences do not interfere with the clustering processes' performance. Not doing this could promote the less populated groups of values within variables with huge asymmetric displays to merge with more relevant ones, decreasing the segmentation quality.

We can finally attain the line-up for the cluster analysis: this process might use several techniques. Consequently, a decision on the algorithm(s) that better suits both the dataset and the project objectives so its/their implementation starts to take place.

Consummating the client's segmentation, our team will be ready to offer an integrate description on Wonderful Wines of the World customers' briefcase.

Variable Interpretation

"It is a capital mistake to theorize before one has data."

(Sir Arthur Conan Doyle, 1891)

The data sample provided by Wonderful Wines of the World grants notions on multiple attributes that characterize a client's profile. However, not all these variables refer to the same common field of study. The two different segmentations ordered by WWW will surely require different data as they cover various aspects of the client's identification.

Variables Discernment

We start by producing a disclosure on each variable's true meaning.

CUSTID

The customer ID number serves only for a specific client identification as it is the only attribute unique to each client.

DAYSWUS

The variable *Dayswus* indicates the number of days since the individual started to be a customer. This value does not refer to the client's first purchase. Instead, it refers to the instant the client entered WWW membership program.

AGE

The variable *Age* represents the consumer's age.

EDUCATION

The variable *Education* represents the amount of client's years of education.

INCOME

The variable *Income* represents the amount of the household's income.

KIDHOME

The variable *Kidhome* is a *boolean* that distinguishes the presence or not of an under thirteen years old child at the client's home.

TEENHOME

The variable *Teenhome* is a *boolean* that distinguishes the presence or not of a teenager with thirteen to nineteen years of age at the client's home.

FREQ

The variable *Freq* displays the number of the client's purchases during the past eighteen months. This variable has a minimal value of one, as the clients will be eliminated from the active membership dataset after a year and half without buying any product from WWW.

REGENCY

The variable *Recency* represents the number of days since the last time the client acquired a product from WWW.

MONETARY

The variable *Monetary* represents the total monetary value of the client's past eighteen months.

LTV

The variable *LTV* represents the total income WWW can expect to bring in from the customer as long as he remains a client.

PERDEAL

The variable *Perdeal* represents the percentage of products the client acquired while they were on discount

DRYRED

The variable *Dryred* represents the percentage of dry red wines the client acquired.

SWEETRED

The variable *Sweetred* represents the percentage of sweet red wines the client acquired.

DRYWH

The variable *Drywh* represents the percentage of dry white wines the client acquired.

SWEETWH

The variable *Sweetwh* represents the percentage of sweet white wines the client acquired.

DESSERT

The variable *Dessert* represents the percentage of dessert wines the client acquired.

EXOTIC

The variable *Exotic* represents the percentage of exotic wines the client acquired.

WEBPURCHASE

The variable *WebPurchase* represents the percentage of the number of purchases the client does via WWW's website.

WEBVISIT

The variable *WebVisit* represents the average number of times the clients navigate through WWW's website per month.

ACCESS

The variable *Access* represents the number of accessories the client bought in past eighteen months.

NOTE: DRYRED + SWEETRED + DRYWH + SWEETWH + DESSERT = 100%



New Variables

By combining some of the just introduced variables, new ways of evaluating Wonderful Wines of the World's clientele might be conceived. The group is of the opinion the following criteria could be relevant to the explanation of both segmentations taking place.

AVRPUR = MONETARY / FREQ , average value of the client's purchases in the last 18 months

Variables' Group Division

Going through all the data variables explanation aids to take notice on the possible different data fields. As it serves only for client identification, we already decided not to use CUSTID. We divide the data attributes into three groups.

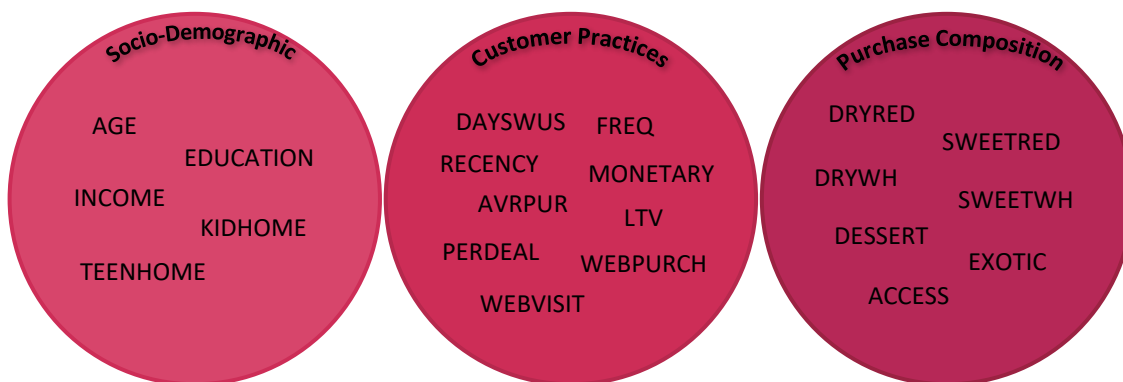


FIGURE 1 VARIABLES' GROUPS

These are the established initial datasets to the later clustering process.

Client Profile dataset: Socio-Demographic and Customer Practices variables

Buying Behaviour dataset: Socio-Demographic and Product Composition variables

Data Integrity and Requisites Confirmation

"Integrity is the seed for achievement. It is the principle that never fails."

(Earl Nightingale, 1957)

First, for being able to guarantee a truthful, unbiased look into Wonderful Wines of the World clientele, the team decided to check the quality of the dataset and confirm if the records have details that need to be changed for a better future analysis.

Resolution on Data Missing Values and Duplicate Rows

No missing values or duplicate rows within both datasets.

Resolution on Data Incoherent Records

On clients' age

For the customer's age, we checked if the records' variable *Age* values are all above eighteen years old. Therefore, we conclude that there is no record of having minors in the dataset.

On clients' purchase recency

We check if the customers' purchases were made in the last eighteen months (variable *Recency*), since customers who have not purchased within this time range should not be considered. As the amount of customers' records not complying with these requirements was negligible, we replaced their variable values to the limit value, equal to saying these client's last purchase was precisely eighteen months ago.





On clients' wine taste

Also, we check if the sum of the variables *Dryred*, *Sweetred*, *Drywh*, *Sweetwh* and *Dessert* is equal to 100. As we find some values with the sum 99 and 101, we analyse the Purchased Composition Variables (Figure XVIII from the appendix) and we notice that the variable *Dryred* is the one closer of "being a uniform" distribution, more "well-behaved". So, for the records whose sum is equal to 99, we are adding 1 value on variable *Dryred* and, for the record which "wine" variables that sum 101, we are reducing 1 value on this same variable.

Data's First Statistics and Data Visualization

"The beginning is the most important part of the work."

(Plato, 372 BC)

After completely checking the dataset, we can move forward on the Data Analysis. Besides the visual tools that will be presented, the team also assembled histograms and boxplots for both datasets' variables (Figure XIX, XX, XXI, XXII from the appendix).

NOTE: We apply the Sturges' Rule [$\log_2 10000 + 1 \sim 15$] to achieve an "ideal number" on the histograms' classes.

Clients' Profile dataset

Here are some brief statistics of Clients' Profile dataset variables:

	count	mean	std	min	25%	50%	75%	max
Age	10000.0	47.927300	17.302721	18.0	33.00	48.00	63.00	78.0
Educ	10000.0	16.739100	1.876375	12.0	15.00	17.00	18.00	20.0
Income	10000.0	69904.358000	27612.233311	10000.0	47642.00	70012.00	92147.00	140628.0
Kidhome	10000.0	0.418800	0.493387	0.0	0.00	0.00	1.00	1.0
Teenhome	10000.0	0.469800	0.499112	0.0	0.00	0.00	1.00	1.0
Dayswus	10000.0	898.102000	202.492789	550.0	723.75	894.00	1074.00	1250.0
Freq	10000.0	14.628100	11.969073	1.0	4.00	12.00	24.00	56.0
Recency	10000.0	62.406800	69.874255	0.0	26.00	52.00	78.25	549.0
Monetary	10000.0	622.555200	647.135323	6.0	63.00	383.00	1077.00	3052.0
AVRPUR	10000.0	31.942299	13.711039	6.0	20.00	31.92	44.88	54.5
LTV	10000.0	209.071200	291.986040	-178.0	-2.00	57.00	364.00	1791.0
Perdeal	10000.0	32.397200	27.897094	0.0	6.00	25.00	56.00	97.0
WebPurchase	10000.0	42.376200	18.522062	4.0	28.00	45.00	57.00	88.0
WebVisit	10000.0	5.216600	2.330457	0.0	3.00	6.00	7.00	10.0

TABLE 1 CLIENT'S PROFILE STATISTICS

Having the different statistics for each one of Clients' Profile dataset variables, we decide to measure the correlation between them with a Pearson's correlation matrix. We care to check the variables' interaction between each other.



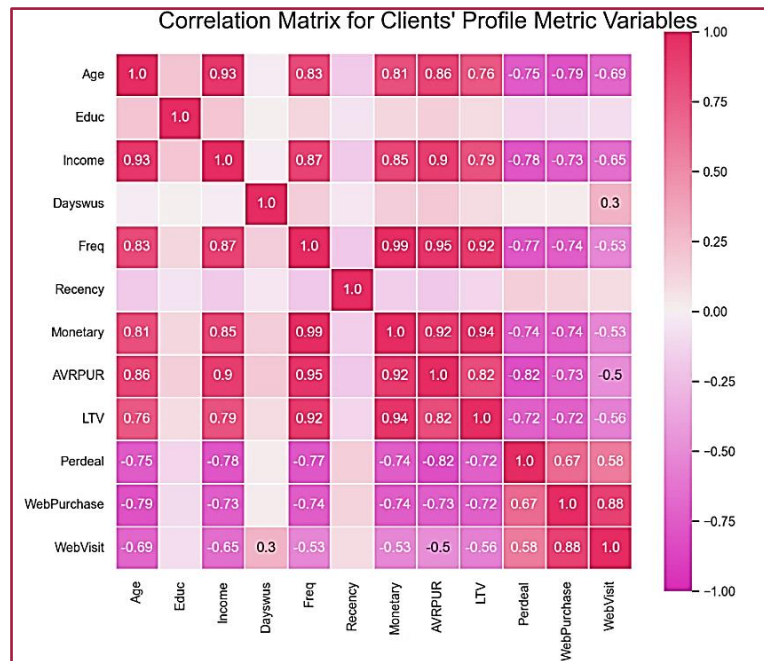


FIGURE II CLIENTS' PROFILES_PEARSON MATRIX FOR VARIABLES' CORRELATION

We also build Boxplots to measure the effect of both variables *Kidhome* and *Teenhome*.

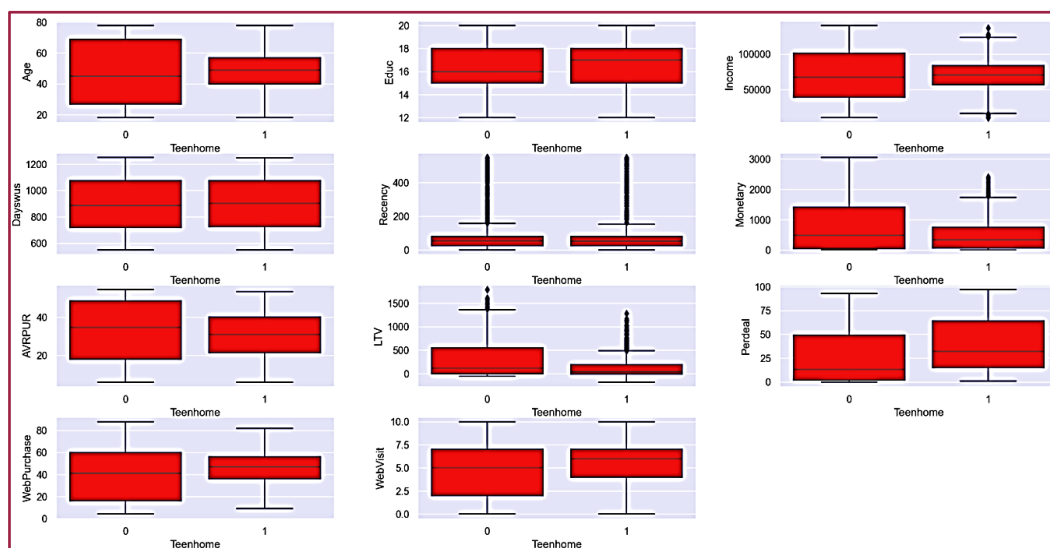


FIGURE III CLIENTS' PROFILE_TEENHOME BOXPLOTS

Through the observations of these visual tools, we attain that variables *Educ*, *Teenhome*, *Dayswus* and *Recency* do not have a strong correlation with the other variables. Also, we discover that variable *AVRPUR* variable has a good correlation with the other variables.

So, we decide to drop variables *Educ*, *Teenhome*, *Dayswus*, *Recency*, *Monetary* and *Freq*. The last two variables notwithstanding being strongly correlated with some of the other variables, are redundant if variable *AVRPUR* is taken into consideration, so we decide to drop these variables for the upcoming segmentation.

Client Profile dataset variables = [*Age*, *Income*, *Kidhome*, *AVRPUR*, *LTV*, *Perdeal*, *WebPurchase*, *WebVisit*]

Buying Behaviour dataset

Finished analysing the Client Profile dataset, now it is the time to look at the Buying Behaviour dataset variables:



	count	mean	std	min	25%	50%	75%	max
Age	10000.0	47.9273	17.302721	18.0	33.0	48.0	63.0	78.0
Educ	10000.0	16.7391	1.876375	12.0	15.0	17.0	18.0	20.0
Income	10000.0	69904.3580	27612.233311	10000.0	47642.0	70012.0	92147.0	140628.0
Kidhome	10000.0	0.4188	0.493387	0.0	0.0	0.0	1.0	1.0
Teenhome	10000.0	0.4698	0.499112	0.0	0.0	0.0	1.0	1.0
Dryred	10000.0	50.4070	23.480445	1.0	32.0	51.0	69.0	99.0
Sweetred	10000.0	7.0545	7.866544	0.0	2.0	4.0	10.0	75.0
Drywh	10000.0	28.5213	12.583957	1.0	19.0	28.0	37.0	74.0
Sweetwh	10000.0	7.0698	8.015083	0.0	2.0	4.0	10.0	62.0
Dessert	10000.0	6.9474	7.879546	0.0	2.0	4.0	9.0	77.0
Exotic	10000.0	16.5466	17.247672	0.0	4.0	10.0	23.0	96.0
Access	10000.0	0.2460	0.539178	0.0	0.0	0.0	0.0	3.0

TABLE 2 BUYING BEHAVIOUR STATISTICS

Having assembled these statistics, we decide to measure the correlation between the Buying Behaviour variables with a Pearson's correlation matrix. As carried before, we also plot the variable's pairwise relations.

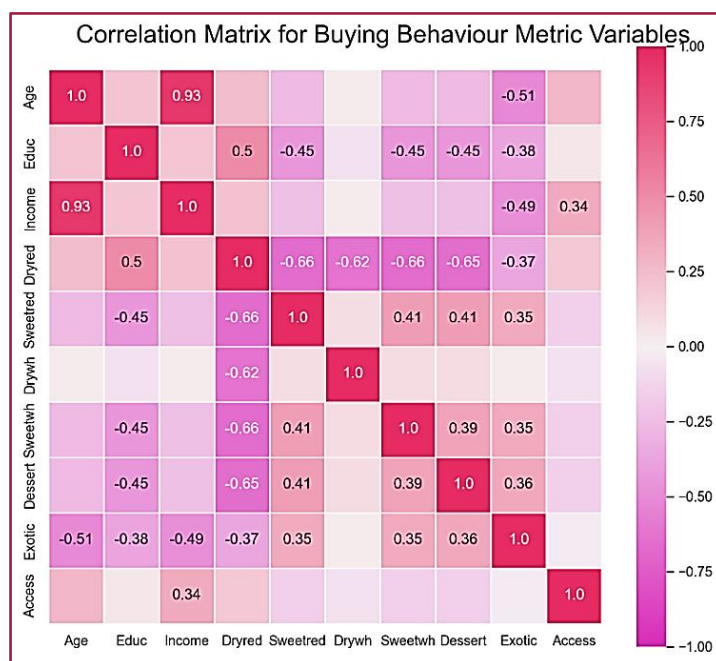


FIGURE IV BUYING BEHAVIOUR_PEARSON MATRIX FOR VARIABLES' CORRELATION

We also build Boxplots to measure the effect of both variables *Kidhome* and *Teenhome*.

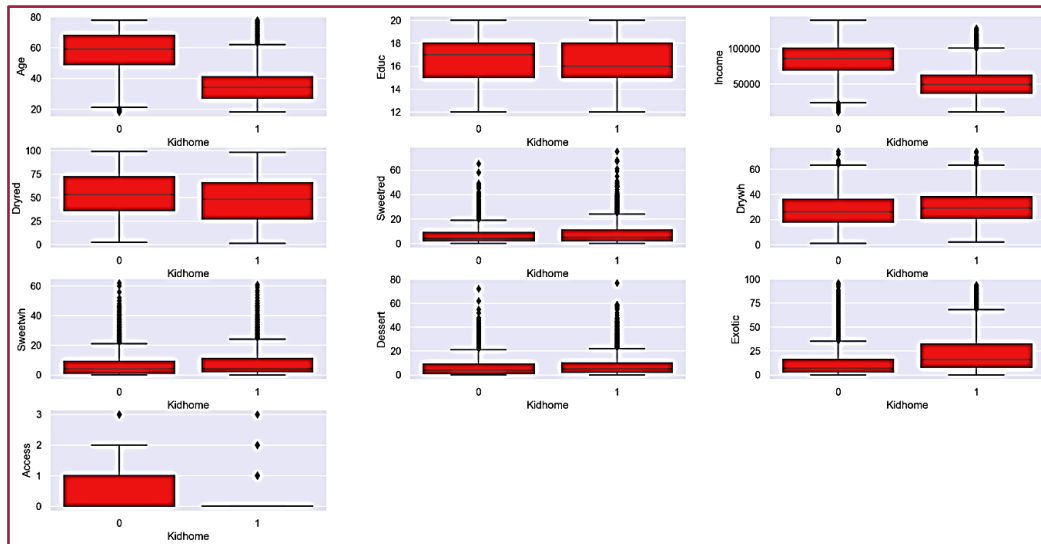


FIGURE V BUYING BEHAVIOUR_KIDHOME BOXPLOTS

Considering that the client's buying behaviour is heavily dependent on their taste (and we must bear in mind taste is a very subjective matter) we will have to look very carefully at the variables' importance within the dataset's tendencies. Through the observations of these visual tools, we attain that variables *Kidhome*, and *Access* have a relatively weak correlation with the other variables. Also, they are not as important as the others for logical reason: we do not expect children to be drinking and accessories are majorly bought in rare occasions, with the evident exception of "rich clients" who prior their "wine tasting experience" and can afford all the types of exquisite accessories. Therefore, we decide to drop variables *Kidhome* and *Access* for the segmentation process.

Buying Behaviour dataset variables = [*Age*, *Educ*, *Income*, *Dryred*, *Sweetred*, *Drywh*, *Sweetwh*, *Dessert*, *Exotic*]

Data Pre-Processing

“Without clean data, or clean enough data, your data science is worthless.”

(Michael Stonebraker, 2019)

The project team may now start preparing both datasets to the segmentation process. The group will drop the binary variables, KIDHOME and TEENHOME, for this pre-processing step.

Resolution on Atypical Non-Standard Values (Outliers)

For this part, we are looking for values that could create some problems in the clustering process. We glimpse these odd values with the help of histograms and boxplots (Figure XXVII and XXIX from the appendix).

(!) Although we look at each of the datasets individually on the code approach, for integrity purposes the outliers' designation will be the same for the entire data collection. This fact enables us to show the process full disclose at once, which is more convenient for the reader's understanding.

We go across each variable individually and decide on the outlier's removal.

Socio-Demographic Variables

The team did not remove any records based on any Socio-Demographic variable, namely: *Age*, *Educ* and *Income*.

Customer Practices Variables

Within these group of variables, we focus on *LTV* variable.

As we can observe clearer on the boxplot graphic, *LTV* variable is different to the other Customer Practices variables, because there are records outside the barriers.

At the first time we use the IQR method, but if we follow this method, we realize that we would be left with 96,77% of the data after eliminating the outliers. We would be eliminating a very large percentage of records in a very early stage. Also, this group of records are representing the most valuable clients, which are a legit important clientele group.

Therefore, as a team, we believe it is best to delete records manually.

We check the possibility to drop clients with an *LTV* superior to one thousand (1000). This way, we are keeping 97,99% of the data after removing the outliers. So, we agree to take this option instead of the IQR method.

Nothing to report on the other variables, namely: *AVRPUR*, *Perdeal*, *WebPurchase* and *WebVisit*.

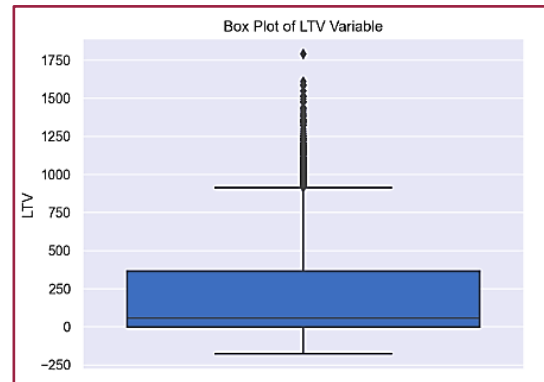


FIGURE VI BOXPLOT OF LTV VARIABLE

Purchase Composition Variables

We have “problematic” records when concerning all this group variables’ values.

Following the same steps as before, we use the IQR method on these variables. We realize that if we follow this method, we are keeping with the 74.46% of data after removing the outliers: that is an unthinkable number of data being removed. Also, individually, we notice that the IQR process displays problems for all variables except for *Drywh* variable. So, we will have to follow another path for dealing with these variables’ outlier treatment.

Beginning with *Sweetred* variable, we can observe a substantial number of customers that prefer sweet red wine, for this reason we cannot afford to disregard this important group of clients.

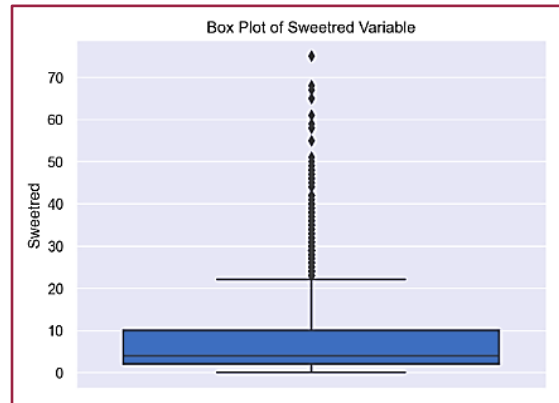


FIGURE VII BOXPLOT OF SWEETRED VARIABLE

We check the possibility of letting drop records with a value superior to forty-five (45) on this variable.

With this process we are keeping 99.73% of the data. So, we agree to take this option instead of the IQR method.

For *Sweetwh* variable, we can observe a considerable number of clients with an extraordinary preference for sweet white wines.

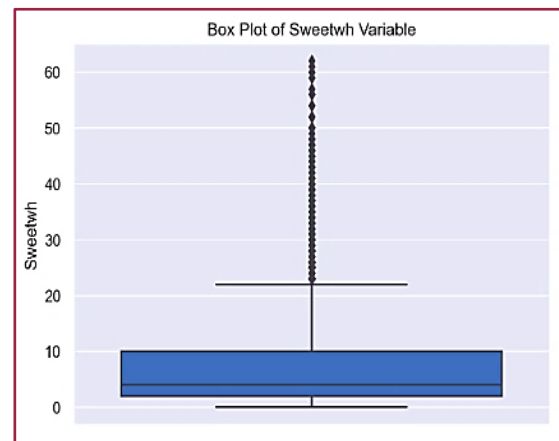


FIGURE VIII BOXPLOT FOR SWEETWH VARIABLE

As we think the group of values outside the barrier is important, we decide to check the possibility of letting drop records with a value superior to forty-five (45) on this variable.

The result was great, keeping 99.63% of the data after removing outliers. So, we decided to take this option instead of the IQR method.

For *Dessert* variable, we also can observe a big number of records outside the barrier, where all the outliers are representing the clients that have a huge preference on dessert kind of wines.

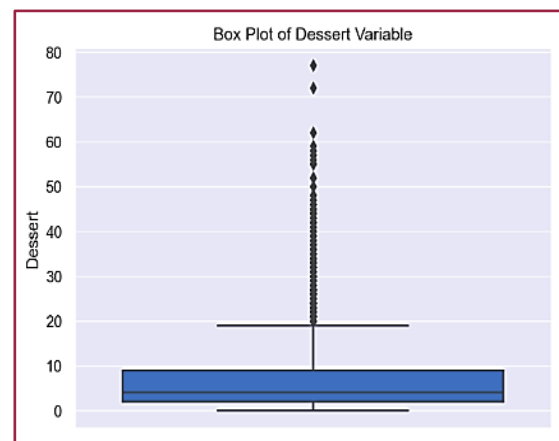


FIGURE IX BOXPLOT FOR DESSERT VARIABLE

As we consider this group of records important for our analysis, we decide to check the possibility of letting drop records with a value superior to forty (40) on the variable.

After checking this possibility, we would have a great percentage of kept data, 99,36%. So, we are considering this option instead of the IQR method.

Summing up...

After finishing the outliers’ removal process, we acknowledge the percentage of data kept on both datasets: 94,95%.





Data Normalization

We start the normalization process, which is a crucial step to get a good clustering process afterwards.

Since there is more than one method for normalization, we decide to evaluate three types of normalization procedures and then evaluate which one of these three methods will have a better impact on our database. After each one of the methods' performance, we draw some visual tools to help us deciding what technique(s) will be choose (Figure XXXI, XXXII, XXXIII, XXXIV, XXXV, XXXVI from the appendix)

Before this, we need to remove the categorical variables, which are *Kidhome* in Client's profile variables and *Teenhome* in Buying Behaviour variables.

The Normalization process is based on assigning a new value to the records of the variables, which are designated within a specific range. As it stands, all the variables will be analysed and compared with the same unit of measurement and the data can be analysed in a more precise way.

The first normalization method that we decide to analyse is Minmax normalization, in which a scale from 0 to 1 is applied to all variables. It is a common technique used to pre-process data before applying machine learning algorithms. The normalization formula is as follows:

$$normalized_value = \frac{value - min}{(max - min)}$$

Standard normalization is another method used to transform the values of a variable to a specific scale. Instead of using a fixed range like min-max normalization, standard normalization uses the standard deviation of the variable's values to determine the scale. Standard normalization is useful because it allows us to compare variables that may have different units of measurement or different scales. It is also useful for identifying outliers in the data, as values that are significantly larger or smaller than the mean will have a higher standardized value.

$$normalized_value = \frac{value - mean}{standard_deviation}$$

Finally, we decide to assess the Z-Score normalization method, which allows us to create a scale where we can compare values with different mean and different standard deviations based on the statistical formula, $New_value = \frac{x - \mu}{\sigma}$.

After implementing these three different methods for Client's Profile and Buying Behaviour variables, we realized that Minmax normalization is the type of normalization where the difference between the different records is better appreciated, which makes our work in the cluster process easier and clearer. We also consider this normalization technique for its' ability to diminish the remaining outliers' impact. This is because the outliers are scaled along with the rest of the data, so they are not as extreme relative to the rest of the data after normalization.

Imagine you are trying to predict the price of a house based on its size. You have collected data on 100 houses, with sizes ranging from 800 to 4,000 square feet. The majority of the houses are between 1,000 and 2,000 square feet, but a few are much larger, such as a mansion at 10,000 square feet. This mansion would be considered an outlier in the dataset.

To scale the size variable, you can apply min-max normalization. This involves subtracting the minimum value (800) from all the values and then dividing the result by the range (4,000-800), resulting in a new range from 0 to 1. As a result, the large mansion is scaled down along with the rest of the data and is not as extreme relative to the other values after normalization. However, it is



important to note that the mansion's size is still significantly larger than most of the other houses in the dataset and may have a considerable influence on the model's predictions.



Clustering Analysis

“Divide et Impera (Divide and Rule)”

(Philip II of Macedon, 340 BC)

We are on the verge of beginning our clustering processes and achieving the desired segmentations to Wonderful Wines of the World corporation. With these segmentations, the company will have a much clear notion on positive future possibilities for increasing its sales and top the performance ratings. Not only we wish to guarantee the best possible results, but also to ensure that if WWW decides to conduct future segmentations, the algorithm's efficiency holds on to the minimal time employment. For these reasons, we will use five clustering techniques on both datasets and, on a later stage, decide which one delivers the best clusters to define marketing approaches.

We use the following techniques: **K-Means, Mean Shift, Agglomerative Hierarchical with Average Linkage, Agglomerative Hierarchical with Ward Linkage and DBSCAN.**

We describe the Mean Shift, Agglomerative Hierarchical with Ward Linkage and DBSCAN techniques. Both K-Means and Agglomerative Hierarchical with Average Linkage were chosen as processes for data segmentation and will be presented on a later stage.

Mean-shift clustering is a machine learning technique used to identify clusters within a dataset. The algorithm defines a kernel function, which assigns a weight to each data point based on its relative influence or importance. The kernel function is then used to estimate the probability density function of the data, allowing the algorithm to identify the densest regions of the dataset. The algorithm iteratively shifts the kernel to these dense regions and repeats the process until convergence is reached, resulting in the identification of clusters within the data. It is important to carefully consider the choice of kernel function, as it can impact the results of the mean-shift algorithm and may be more suitable for certain types of data.

Agglomerative Hierarchical Clustering with Ward Linkage is a robust method for identifying clusters within a dataset. The algorithm begins by treating each data point as a separate cluster and progressively merges the closest clusters together until all points are contained in a single cluster. Ward Linkage is used to measure the distance between clusters and determine which clusters should be merged at each step of the process. The resulting dendrogram (tree-like diagram) allows the user to easily visualize the hierarchy of clusters and determine the optimal number of clusters for the data. This clustering method is highly versatile and effective for a wide range of applications.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a state-of-the-art clustering algorithm that leverages density to identify clusters within a dataset. The algorithm defines a minimum number of points (minPts) that must be within a certain distance (epsilon) in order to form a cluster, while points that do not meet these criteria are labelled as "noise" and excluded from the clusters. This powerful method is effective at identifying clusters of any shape and is resistant to the presence of noise or outliers in the data. It is no surprise that DBSCAN is widely utilized and highly regarded in a variety of fields.

Client Profile Segmentation

After segmenting the client profile dataset with the already mentioned five techniques, we chose the Agglomerative Hierarchical Clustering with Average Linkage. For size related issues and to keep the report as clean and clear as we can, we decide to leave the graphics for the not chosen methods only in the python notebook.

This type of methodology bases itself on merging data points on a certain measure level, that in its growing process will merge all records together on a same group. Choosing the Average Linkage as the “connecting measure” means that the distance between two clusters is defined as the average of distances between all pairs of objects.

We define a max distance between records within the dataset and, consequently, identify three distinct clusters that we now will describe. We display some of the visual tools created for the clusters’ interpretation that we consider more interesting.

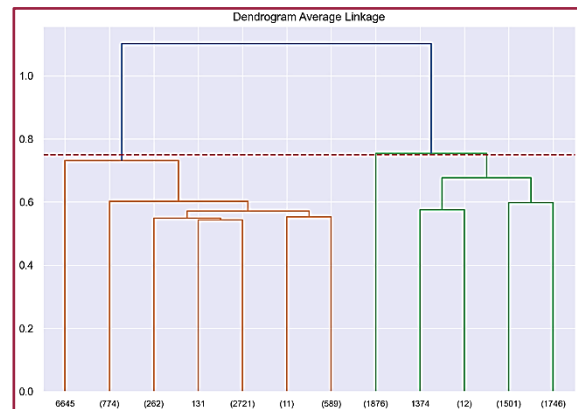


FIGURE X CLIENTS' PROFILE_DENDROGRAM FOR AVERAGE LINKAGE

	Age	Income	AVRPUR	LTV	Perdeal	WebPurchase	WebVisit	Kidhome
CP								
0	33.197752	46717.830695	19.884992	-1.395962	56.621932	55.510209	6.573985	0.788254
1	55.497853	81245.981902	38.115402	202.197239	17.044172	38.671779	4.874233	0.140491
2	69.974947	104761.735075	49.395064	637.592217	2.640725	18.982409	2.715885	0.044776

TABLE 3 CLIENTS' PROFILE_CLUSTERS VARIABLES' MEANS

The first cluster consists of individuals who are younger and have lower incomes. They tend to make more purchases on the website and are more likely to take advantage of discounts on products.

🍇 The individuals comprising the second cluster are adults with an average age of fifty-five (55) and higher incomes compared to those in the first cluster. They occasionally make use of discounts when making purchases on the website, but they also make purchases at full price. This results in a higher lifetime value for this group. They make fewer purchases on the website compared to the first cluster.

🍇 The third cluster consists of the oldest individuals with the highest income. This cluster is the opposite of cluster one, as in addition to being older and having a higher income, they also prefer to purchase products without discounts and prefer to make their purchases in person rather than online.

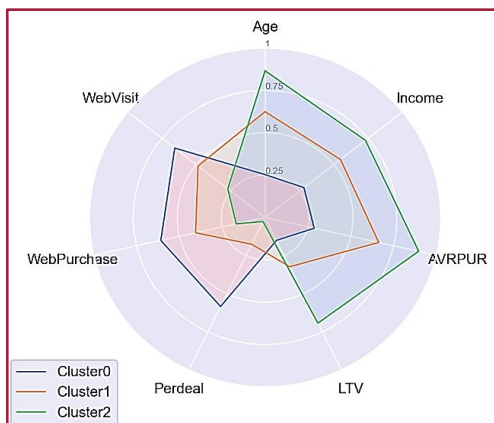


FIGURE XII CLIENTS' PROFILE_RADAR PLOT

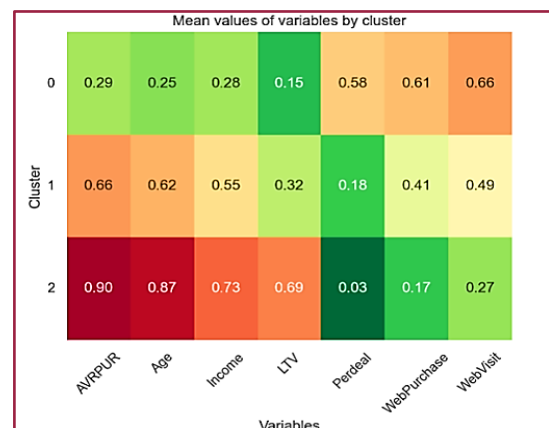


FIGURE XI CLIENTS' PROFILE_HEATMAP



Buying Behaviour Segmentation

After segmenting the Buying Behaviour dataset with the already mentioned five techniques, we chose the K-Means process. For size related issues and to keep the report as clean and clear as we can, we decide to leave the graphics for the not chosen methods only in the python notebook.

This method assigns the data points to a pre-defined number of clusters in such a basis that the sum of the squared distance between the data points and the cluster's centroid (the mean reference point within the cluster) is at the minimum.

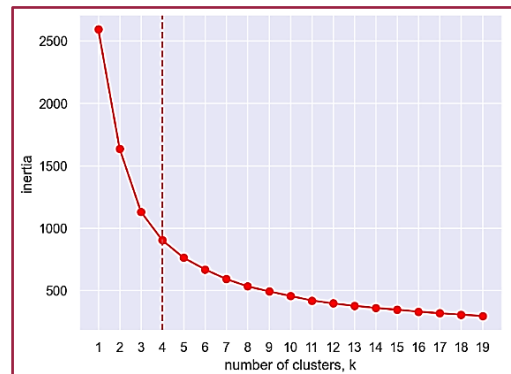


FIGURE XIII BUYING BEHAVIOUR_ INERTIA PER NUMBER OF CLUSTERS

We initially measure the inertia per number of clusters to decide the ideal number of clusters. By applying the “Elbow Law” we reach the ideal number of four clusters.

We begin to describe them. To help us achieving this, we assemble some visual tools.

	Age	Educ	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	Teenhome
BB									
0	51.063778	17.808972	76.299969	2.183399	17.280439	2.150137	2.086054	9.877327	0.774489
1	34.071209	16.830769	46.962637	5.916923	35.652308	5.744615	5.723516	20.524835	0.460659
2	26.058494	14.863782	19.072115	16.983173	29.980769	17.241186	16.722756	34.269231	0.179487
3	66.618182	16.467532	40.534694	8.220779	35.134694	8.261224	7.848609	8.862338	0.305380

TABLE 4 BUYING BEHAVIOUR_ CLUSTERS VARIABLES' MEANS

The first cluster consists of adults with an average age of fifty-one (51) years. This group tends to favour dry wines over sweet ones, with a particular preference for red wine. They do not often purchase exotic wines, suggesting a preference for more traditional varieties.

The second cluster consists of adults with an average age of thirty-four (34) years. Like the first cluster, they tend to prefer dry wines over sweet wines, but have a more balanced preference between red and white wines. This group also tends to be more curious about trying exotic wines.

The third cluster is the youngest group, with an average age of twenty-six (26) years. They tend to have a balanced preference for all types of wine but have a slight preference for dry white wines. This cluster is also the most interested in trying exotic wines.

The fourth cluster consists of individuals with an average age of sixty-seven (67) years, making them the oldest group. Their preference for wine leans towards dry varieties, as opposed to sweet ones, and they do not seem to have a particular interest in trying exotic wines.

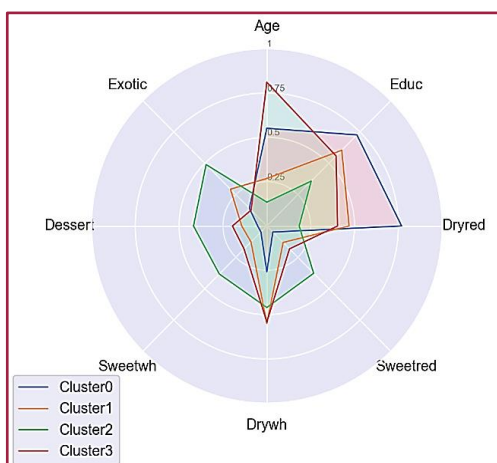


FIGURE XIV BUYING BEHAVIOUR_ RADAR PLOT

Mean values of variables by cluster								
Cluster	Age	Dessert	Dryred	Drywh	Educ	Exotic	Sweetred	Sweetwh
0	0.55	0.05	0.77	0.26	0.73	0.14	0.05	0.05
1	0.27	0.14	0.47	0.55	0.60	0.29	0.13	0.13
2	0.13	0.42	0.18	0.46	0.36	0.49	0.38	0.38
3	0.81	0.20	0.40	0.54	0.56	0.13	0.18	0.18

FIGURE XV BUYING BEHAVIOUR_ HEATMAP



Client Segmentation Explanation and Marketing Approach

Suggestions

We merge both segmentations and assemble eleven final clusters

$[(3 \text{ clients' profiles clusters} * 4 \text{ buying behaviour clusters}) - 1 \text{ empty join} = 11]$

	AVRPUR	Age	Dessert	Dryred	Drywh	Educ	Exotic	Income	LTV	Perdeal	Sweetred	Sweetwh	WebPurchase	WebVisit	Kidhome	Teenhome
Final_Cluster																
0	22.773223	42.091160	1.815838	76.658379	17.803867	17.862799	12.003683	57733.682320	-6.785451	57.713628	1.832413	1.889503	52.778085	6.241252	0.725599	0.848066
1	20.028020	32.938537	5.703415	46.445854	36.177561	16.891220	20.192683	46953.462927	-0.147317	57.034146	5.923902	5.749268	56.355610	6.698049	0.852195	0.448293
2	16.715177	24.765203	16.697635	18.499155	30.474662	14.867399	34.257601	35363.269426	1.802365	55.030405	16.974662	17.353885	56.918919	6.696791	0.737331	0.163007
3	28.173077	55.179487	8.179487	42.307692	29.179487	15.794872	8.615385	72295.769231	-14.051282	52.871795	8.230769	12.102564	44.384615	5.589744	0.717949	0.846154
4	37.807925	53.413441	2.176344	76.886559	16.351075	17.677957	9.503226	79197.737634	189.496237	18.920968	2.339785	2.246237	43.651075	5.615054	0.139247	0.818280
5	35.657689	44.391111	5.906667	51.671111	30.866667	16.280000	23.551111	69560.995556	145.760000	21.537778	5.853333	5.702222	44.186667	5.871111	0.302222	0.573333
6	35.672167	48.916667	16.916667	30.550000	20.666667	14.850000	34.366667	75478.916667	160.933333	17.916667	17.166667	14.700000	37.083333	4.600000	0.150000	0.516667
7	39.255749	61.570404	8.565022	42.046637	31.502242	16.178475	10.451121	87331.076233	236.993722	12.959641	8.893274	8.992825	29.338117	3.452018	0.109417	0.609865
8	49.448882	67.299094	2.465257	71.827795	20.785498	18.368580	5.003021	102209.903323	645.758308	3.873112	2.456193	2.465257	26.709970	3.996979	0.108761	0.287009
9	51.305000	66.000000	21.250000	16.500000	23.500000	14.000000	36.250000	107613.750000	742.500000	0.750000	16.750000	22.000000	16.500000	2.750000	0.000000	0.000000
10	49.378546	70.560026	7.321869	39.395847	37.913692	16.693705	7.719014	105302.454250	635.565866	2.380921	7.733939	7.634653	17.329007	2.440623	0.031149	0.071382

TABLE 5 FINAL CLUSTER_VARIABLES' MEANS

We may now begin to identify the marketing campaigns' target groups, by combining the eleven clusters in a way that is doable to Wonderful Wines of the World's pockets and manpower. We make use of a Decision Tree to relocate the small niches' clients, namely clusters three (3), five (5), six (6), eight (8) and nine (9).

Decision trees are a widely used method for performing supervised learning tasks, including classification and regression. These algorithms construct a tree-like model of decisions based on certain features, dividing the data into subsets according to the most relevant features as determined by a measure such as information gain. Predictions can be made by following the branches of the tree to its leaf nodes, which contain the predicted class or value. To train a decision tree, the relevant features must be selected, and the optimal split points determined, while also pruning the tree to prevent overfitting. We train our algorithm with the clusters we want to maintain.

After we achieve the final clusters, the target groups, we reintroduce the outliers into the data, applying the same trained algorithm used in the last explained process.

	AVRPUR	Age	Dessert	Dryred	Drywh	Educ	Exotic	Income	LTV	Perdeal	Sweetred	Sweetwh	WebPurchase	WebVisit	Kidhome	Teenhome
Final_Cluster																
Discount_Lovers	22.822452	42.162261	1.841386	76.540565	17.853236	17.844120	12.154057	57843.636281	-6.621696	57.541477	1.855971	1.908842	52.761167	6.236098	0.723792	0.847767
Everything_Dry	38.990017	60.963485	8.843154	41.826556	31.080498	16.140249	11.712863	86667.667220	230.428216	13.705394	9.090456	9.159336	29.767635	3.524481	0.119502	0.611618
Life_is_Sweet	16.815547	24.717583	17.734555	17.763747	29.268839	14.752885	39.200950	35255.143924	4.618466	54.653768	17.368635	17.864223	56.602172	6.647658	0.726409	0.162254
Sugar_Daddy	38.476156	54.202489	2.294878	76.021063	16.902824	17.466721	9.916707	80599.343705	213.331259	18.012925	2.434179	2.347056	42.859263	5.532791	0.140258	0.783150
Sugar_GrandDaddy	49.823463	70.465633	6.842894	42.697674	36.378811	16.884238	7.337468	106082.589664	699.492506	2.335401	7.088372	6.992248	17.888889	2.623256	0.034625	0.084238
Young_and_Broke	20.831168	33.426170	5.742844	46.392549	36.056338	16.856429	20.609723	47969.970468	6.515220	55.284416	5.983644	5.824625	55.651976	6.656974	0.826897	0.451159







TABLE 6 TARGET GROUPS_VARIABLES' MEANS

We know during Christmas people tend to flock to road markets. Also, during the pandemic, many of the sellers that usually have a stand in the already referred markets invested in online platforms. Now things are running back in a more "normal way", the costumers' online sales are already a established trend. With Christmas fast approaching, we propose a strategic alliance and a joint marketing plan with some of the most renowned Christmas Markets' organizations to boost sales from December 20th to January 5th. This is an excellent opportunity to capitalize on the holiday shopping season and drive more sales for both of our businesses.





Moving forward we will specify the marketing campaigns for each target group:

-  **Discount Lovers:** “Are you a fan of Dry Red wines and looking for a deal for your Christmas dinner? Look no further! For the holiday season, purchase two Dry Red wines online or at the Christmas markets and receive a 30% discount on your meal. This offer is also valid on the website for all Christmas delicacies, so you can stock up on everything you need for the perfect festive feast.”
-  **Everything Dry:** “If you cannot decide between Dry Red or Dry White wines, this deal is for you! Get a third Dry Red or Dry White wine for free when you purchase two at the Christmas markets or online. Whether you prefer a bold and full-bodied red or a crisp and refreshing white, you can enjoy the best of both worlds with this offer.”
-  **Life is Sweet:** “Don't let the name fool you – this campaign is for sweet wine lovers and bargain hunters alike! Purchase one sweet wine online or at the Christmas markets and receive a 25% discount on your Christmas dinner. This offer is also valid on the website for all Christmas delicacies, so you can indulge in your favourite sweet wine without breaking the bank.”
-  **Sugar Daddy:** “Looking to stock up on Dry Red wines for the holiday season? When you buy two, either online or at the Christmas markets, you will receive a third for free. With this generous offer, you can enjoy a variety of red wines at a fraction of the cost.”
-  **Sugar Grand Daddy:** “Why choose between Dry Red and Dry White wines when you can have both? Purchase two wines, either online or at the Christmas markets, and receive a third for free on all Dry wines at the Christmas markets or online. Additionally, receive a 25% discount on your Christmas dinner, which is also valid on the website for all Christmas delicacies. Treat yourself to a festive feast without breaking the bank with this amazing offer.”
-  **Young and Broke:** “Are you a wine lover on a budget? This campaign is for you! For the holiday season, purchase one wine online or at the Christmas markets and receive a 30% discount on your Christmas dinner. This offer is also valid on the website for all Christmas delicacies, so you can indulge in your favourite wine and treat yourself to a festive feast without breaking the bank. Whether you prefer a rich and fruity red, a crisp and refreshing white, or a sweet and aromatic dessert wine, you'll find something to suit your taste and budget with this amazing deal. Don't miss out on this opportunity to save on your holiday wine and food purchases.”

Appendix

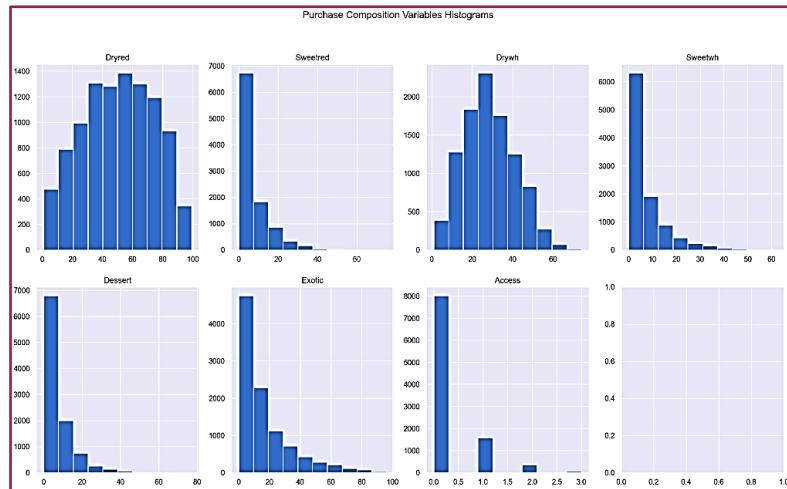


FIGURE XVI PURCHASE COMPOSITION VARIABLES' HISTOGRAMS

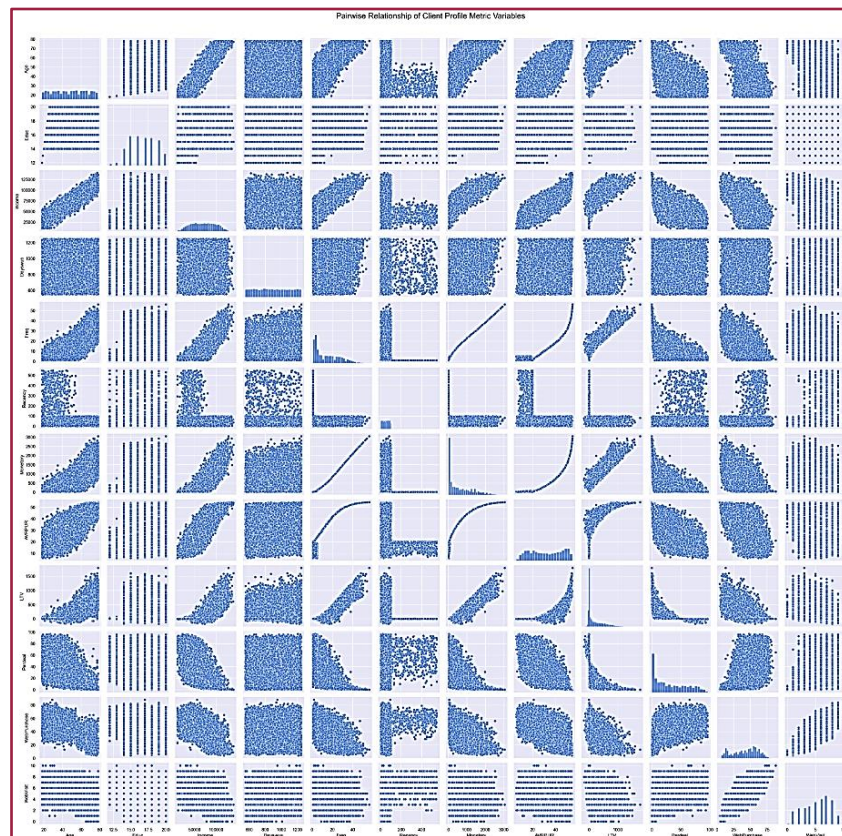


FIGURE XVII CLIENTS' PROFILE VARIABLES' PAIRWISE RELATIONS

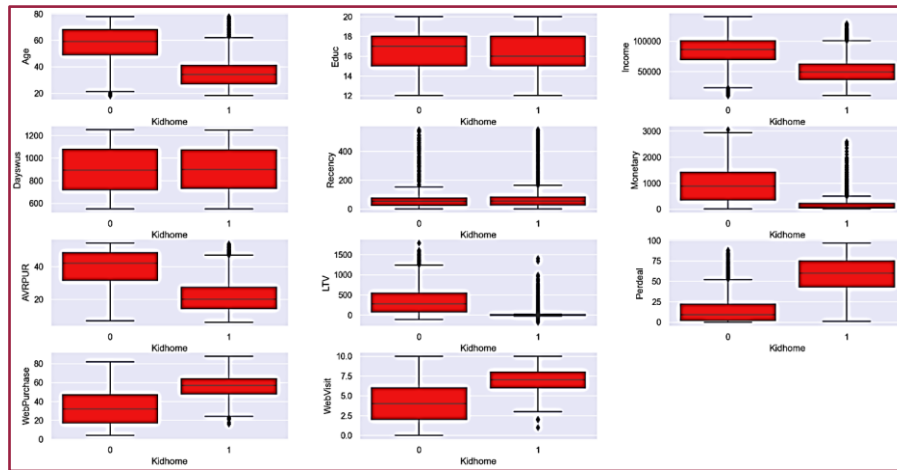


FIGURE XVIII CLIENTS' PROFILE_KIDHOME VARIABLE IMPACT

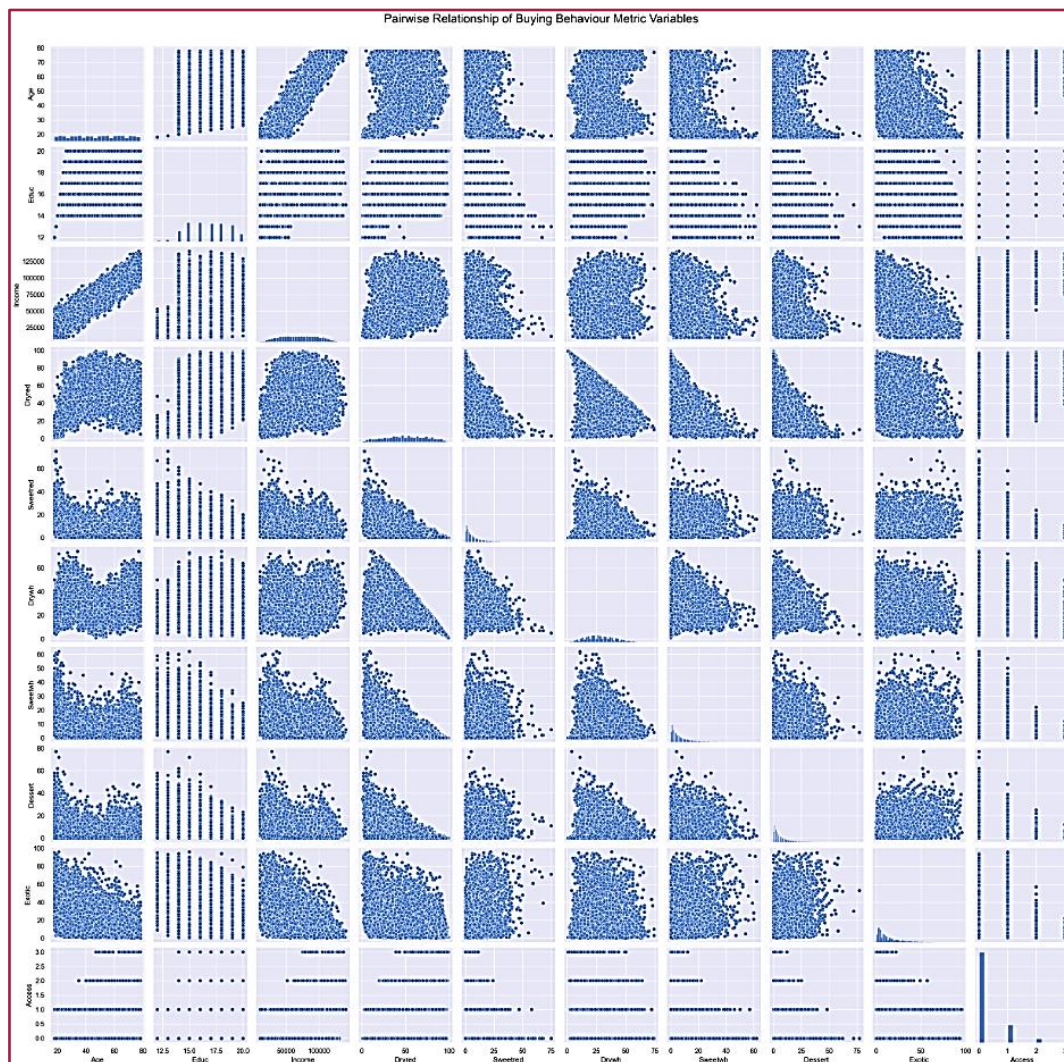


FIGURE XIX BUYING BEHAVIOUR_VARIABLES' PAIRWISE RELATIONS

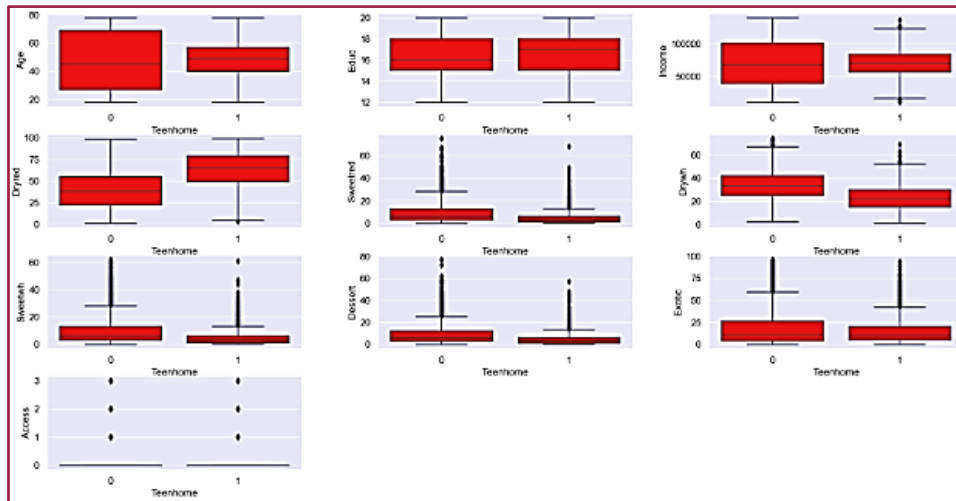


FIGURE XX BUYING BEHAVIOUR_ TEENHOME VARIABLE IMPACT

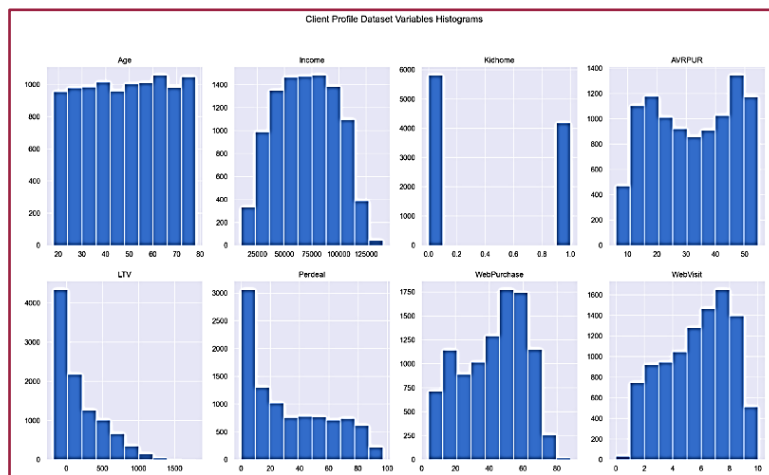


FIGURE XXI CLIENTS' PROFILE_ VARIABLES' HISTOGRAMS

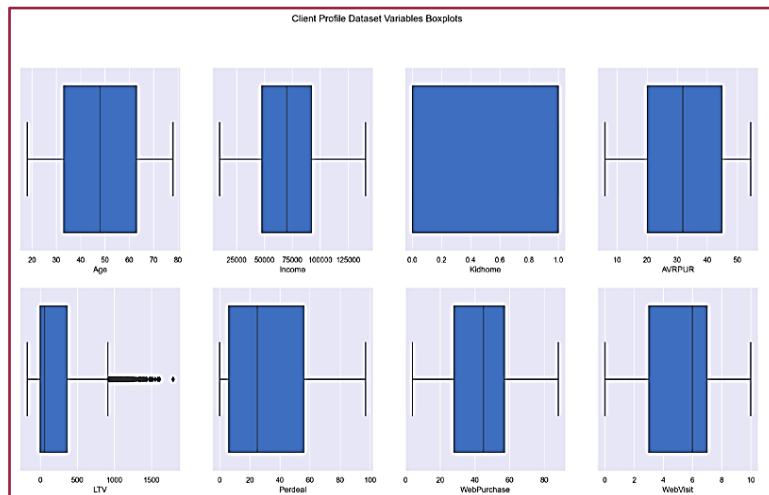


FIGURE XXII CLIENTS' PROFILE_ VARIABLES' BOXPLOTS

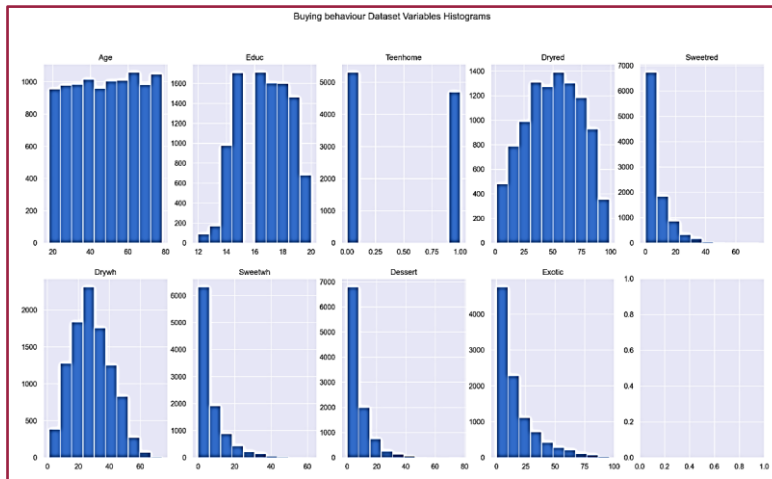


FIGURE XXIII BUYING BEHAVIOUR_VARIABLES' HISTOGRAMS

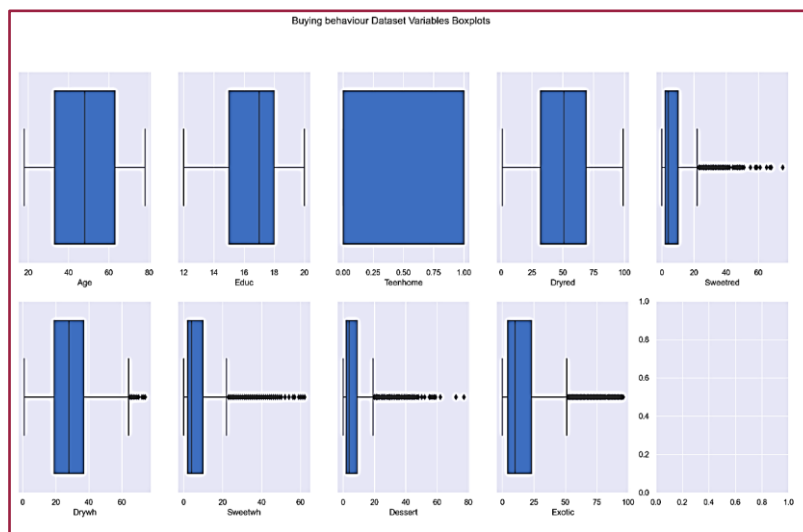


FIGURE XXIV BUYING BEHAVIOUR_VARIABLES' BOXPLOTS

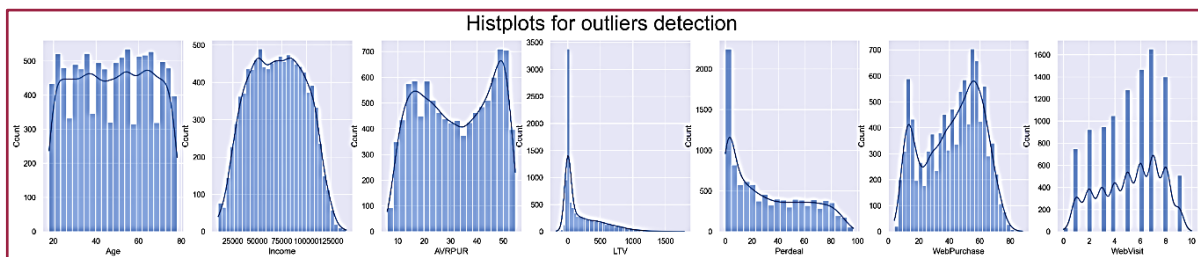


FIGURE XXV CLIENTS' PROFILE_HISTPLOTS FOR OUTLIERS DETECTION

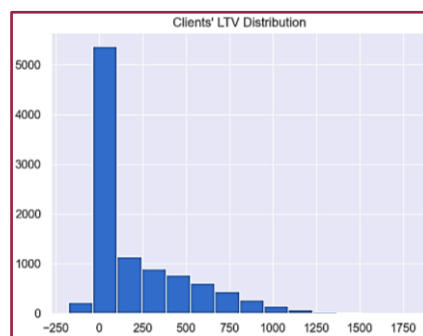


FIGURE XXVI CLIENTS' PROFILE_LTV VARIABLE HISTOGRAM

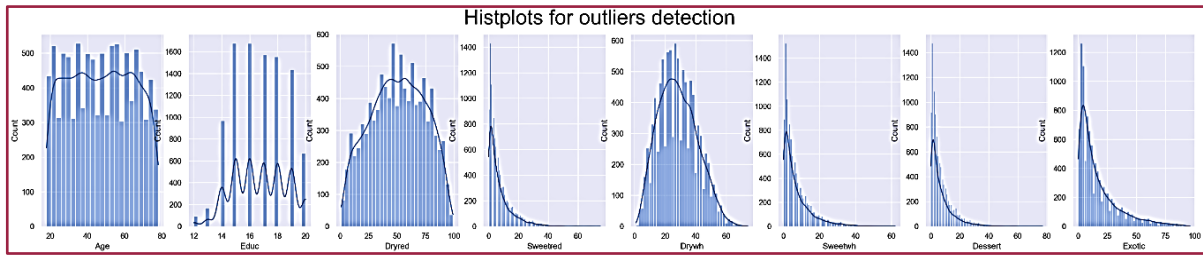
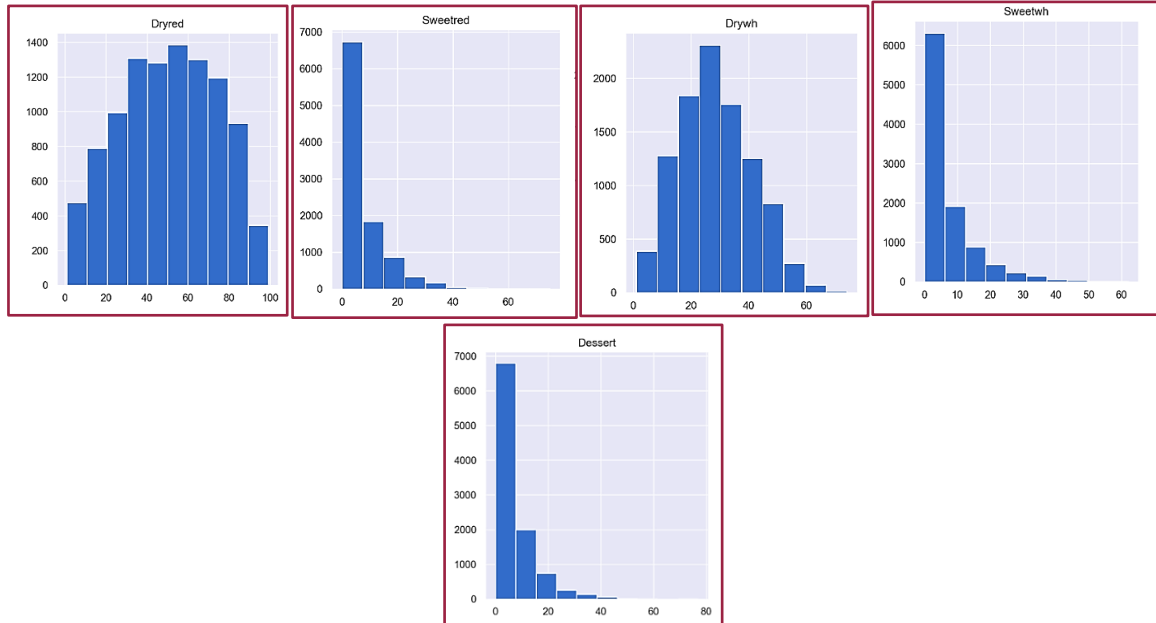


FIGURE XXVII BUYING BEHAVIOUR_ HISTPLOTS FOR OUTLIERS DETECTION



FIGURES XXVIII BUYING BEHAVIOUR_ HISTOGRAMS FOR WYNE TYPES

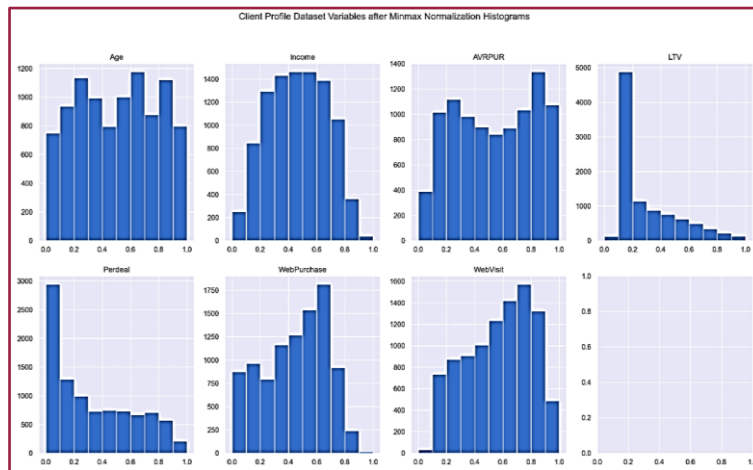


FIGURE XXIX CLIENTS' PROFILE_ HISTOGRAMS FOR VARIABLES AFTER MINMAX NORMALIZATION

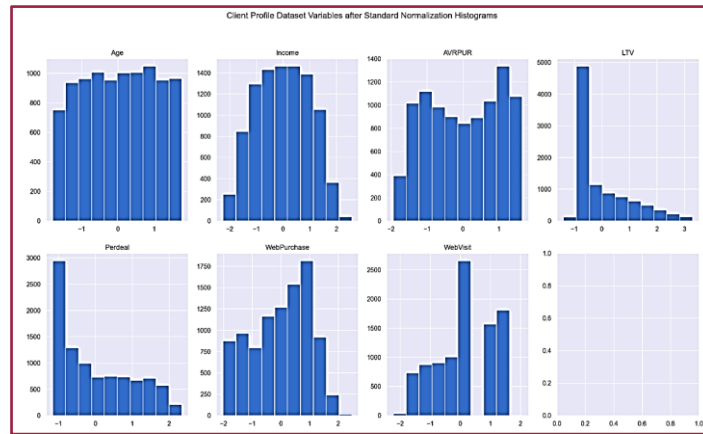


FIGURE XXX CLIENTS' PROFILE_ HISTOGRAMS FOR VARIABLES AFTER STANDARD NORMALIZATION

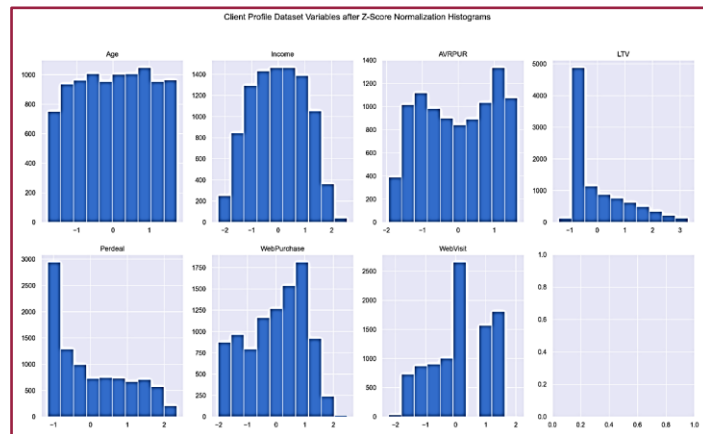


FIGURE XXXI CLIENTS' PROFILE_ HISTOGRAMS FOR VARIABLES AFTER Z-SCORE NORMALIZATION

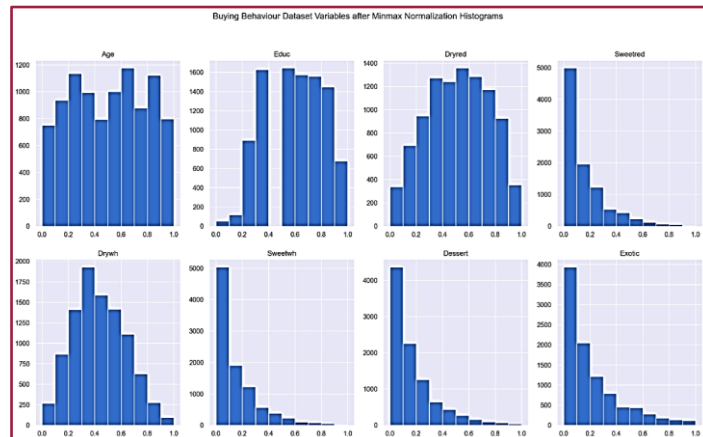


FIGURE XXXII BUYING BEHAVIOUR_ HISTOGRAMS FOR VARIABLES AFTER MINMAX NORMALIZATION



FIGURE XXXIII BUYING BEHAVIOUR_ HISTOGRAMS FOR VARIABLES AFTER STANDARD NORMALIZATION

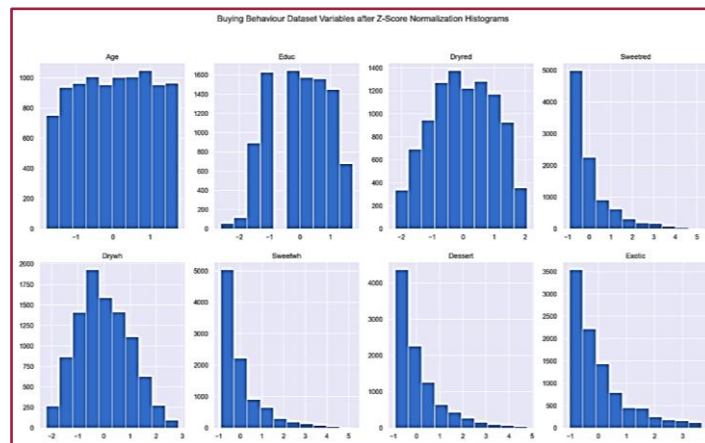


FIGURE XXXIV BUYING BEHAVIOUR_ HISTOGRAMS FOR VARIABLES AFTER Z-SCORE NORMALIZATION

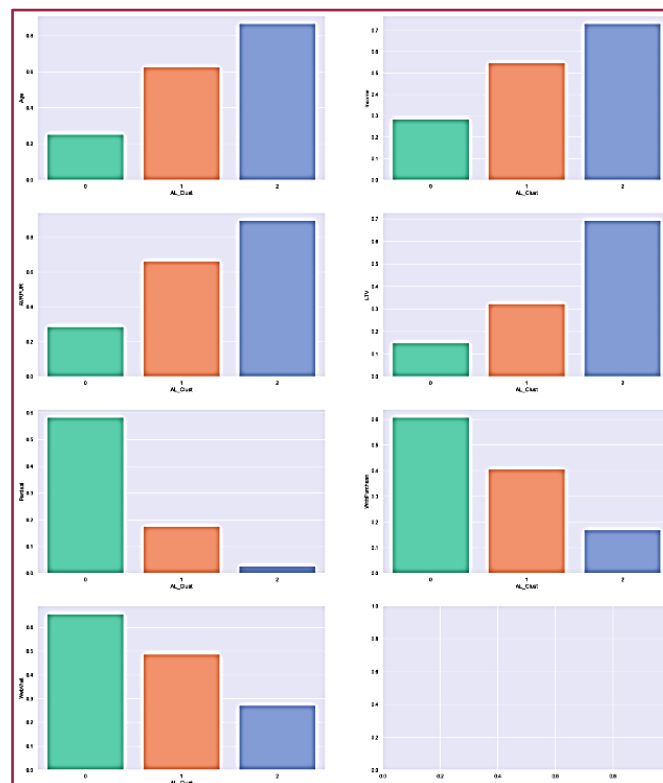


FIGURE XXXV CLIENTS' PROFILE_ VARIABLES' DISTRIBUTION PER CLUSTER

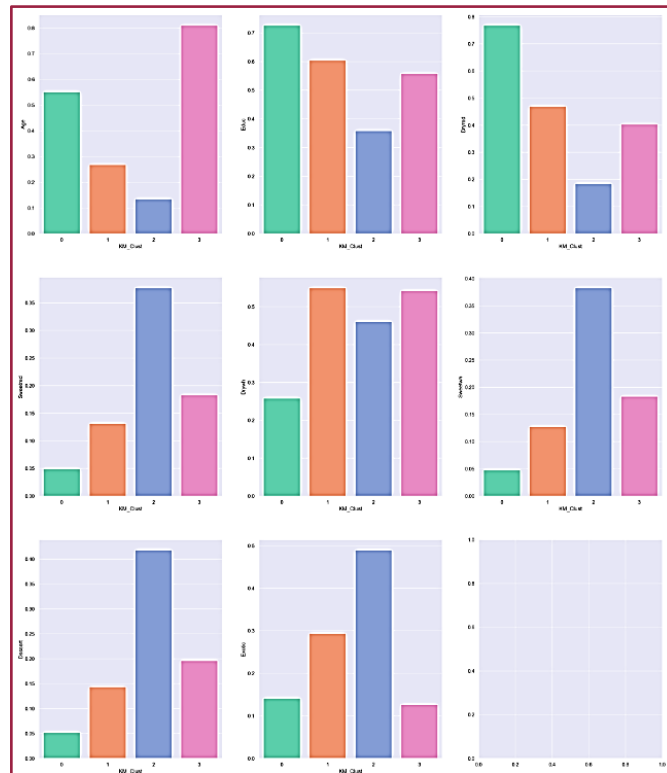


FIGURE XXXVI 'BUYING BEHAVIOUR_VARIABLES' DISTRIBUTION PER CLUSTER

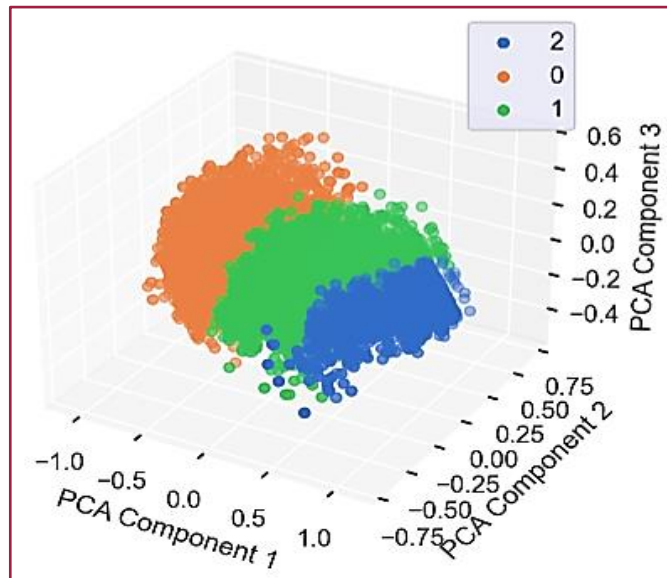


FIGURE XXXVI CLIENTS' PROFILE_CLUSTERS' 3D VIEW

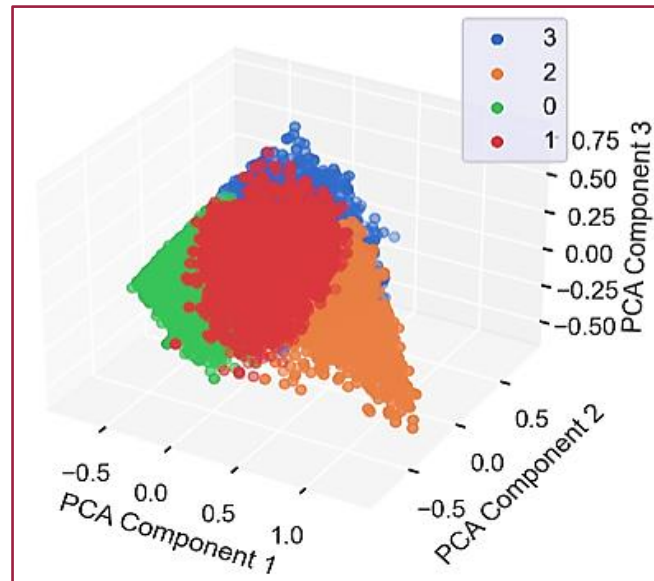


FIGURE XXXVII BUYING BEHAVIOUR_ CLUSTERS' 3D VIEW