BUSINESS CASES WITH DATA SCIENCE

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

WWW CUSTOMER SEGMENTATION

Group X

Beatriz Chumbinho, number: R20170867

Inês Costa number: R20170775

Mª Leonor Morgado, number: R20170871

Rodrigo Matias, number: R20170880

March, 2021

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

INDEX:

1. INTRODUCTION

This project aims to develop a customer segmentation for "Wonderful Wines of the World" (WWW), it is a wine company, whose objective is to delight its customers with well-made, unique, and interesting wines.With the information collected by our team of data scientists the "Wonderful Wines of the World" can segment their customers based on their profile, with the engagement of the customers with the company and with their preferences about the types of wines.

Consequently, adapting different and specific marketing approaches, which will allow the company to benefit from reducing the costs and to have more satisfied customers as they will receive more targeted offers according to their interests and consumption patterns

2. BUSINESS UNDERSTANDING

2.1. Background

"Wonderful Wines of the World", most known by WWW, is an American company with seven years of history. Through catalog, a website or their ten small stores around the US, WWW brings the most amazing wines around the whole world to its customers. Until now their strategy to gather customers is to promote their products in wines and food magazines and so reach the wine lovers.The business is conducted in a very intuitive way, having feedback from customers and staff and maintaining a close relationship with their clients.

2.2. WWW Mission

Ensure quality;Deliver unique experiences;Delight with wines from all around the world that you can only find in WWW stores.

2.3. Business Objectives

WWW intends to improve their business and so increase their business value. To accomplish this, the company needs to deploy new and strategic marketing measures, not only to reach new customers but also to make the current ones buy more, increasing the WWW efficiency and profit.To ensure the appropriate strategy it is necessary to identify, classify and segment the customers, as well as to know how many customer segments there are in the database. This way is possible to develop more focused and targeted marketing measures reaching more and the right customers and so increasing the company sales.

2.4. Business Success criteria

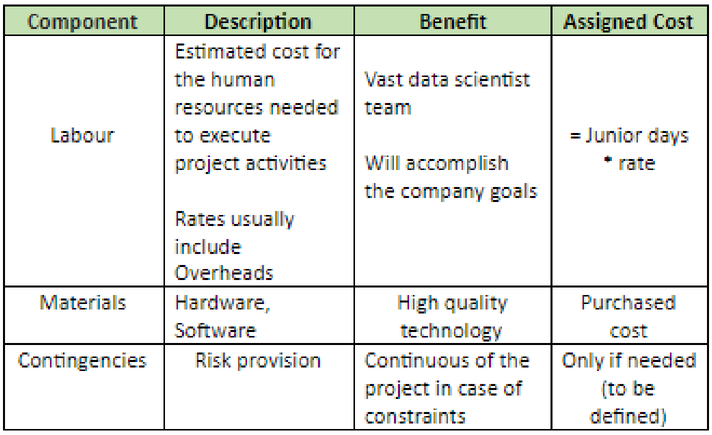
Our team's goal in the present project is to provide to WWW the following, shaped as a report and a presentation:

- Provide to the board members the distinct characteristics of WWW customers
- Produce a well documented list of the customer segments conferred in the database
- Assemble a well planned marketing mix for each segment and ways to reach them
- Deliver it by the 1st of March.

2.5. Situation assessment

To develop this project the "Wonderful Wines of the World" hired four data scientists for the time of seven days. To the data scientists, a sample dataset of the company clients was made available to them in order to perform exploration and create a model which is capable of distinguishing different groups of WWW customers. The customers in the sample data frame are only ones who have purchased in the past 18 months. The platform available to perform the data analysis task in order to accomplish the objective is Python Jupyter Notebook.

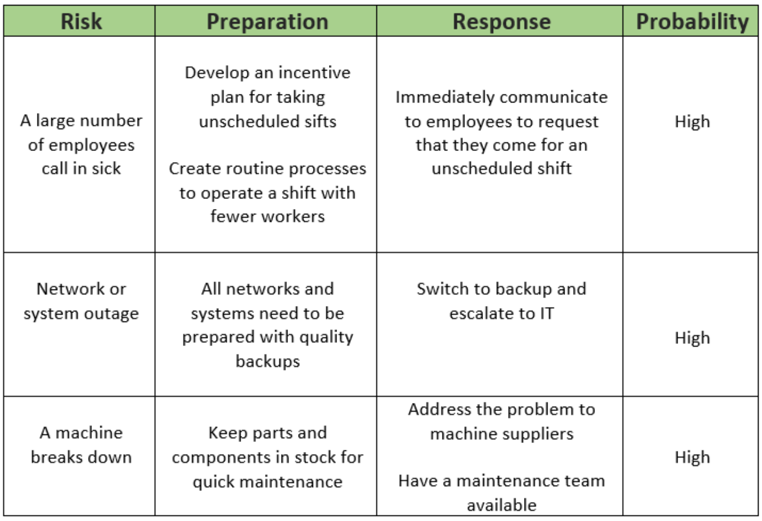
2.6. Costs and Benefits:



Component	Description	Benefit	Assigned Cost
Labour	Estimated cost for the human resources needed to execute project activities Rates usually include Overheads	Vast data scientist team Will accomplish the company goals	= Junior days * rate
Materials	Hardware, Software	High quality technology	Purchased cost
Contingencies	Risk provision	Continuous of the project in case of constraints	Only if needed (to be defined)

table 1

2.7. Risk & Contigency:



Risk	Preparation	Response	Probability
A large number of employees call in sick	Develop an incentive plan for taking unscheduled sifts Create routine processes to operate a shift with fewer workers	Immediately communicate to employees to request that they come for an unscheduled shift	High
Network or system outage	All networks and systems need to be prepared with quality backups	Switch to backup and escalate to IT	High
A machine breaks down	Keep parts and components in stock for quick maintenance	Address the problem to machine suppliers Have a maintenance team available	High

table 2

2.8. Determine Data Mining goals

The identification and classification of the WWW customers, through the dataset of the company clients, are the technical goals of this project.
The data mining methodologies will enable the identification of patterns in the data as well as the partitioning of the data into classes. Hence, we will find customers with similar behaviours and group them into classes accordingly. In a nutshell:

- Segment the customers into clusters, based on: value as a client, demographic conditions and product preferences;
- Identification of patterns in the data- Exploratory Data Analysis (EDA).

3. PREDICTIVE ANALYTICS PROCESS

In order to develop the present analysis of customer segmentation, several steps were performed.

Our team started by understanding the data we had in hands, takin into consideration the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology - a standard process model that describes common approaches used to conduct data mining studies. CRISP-DM methodology can be understanded by looking a the image on the right.
This way, we started by understanding your business, the project objectives, the requirements and the data itself that WWW provided us in a circular, iterative and interactive perspective.
The data was prepared, in the way presented lately in the present section. Then, the data was the input for our model - using SOM (Self Organizing Map) and Hierarchical clustering techniques.

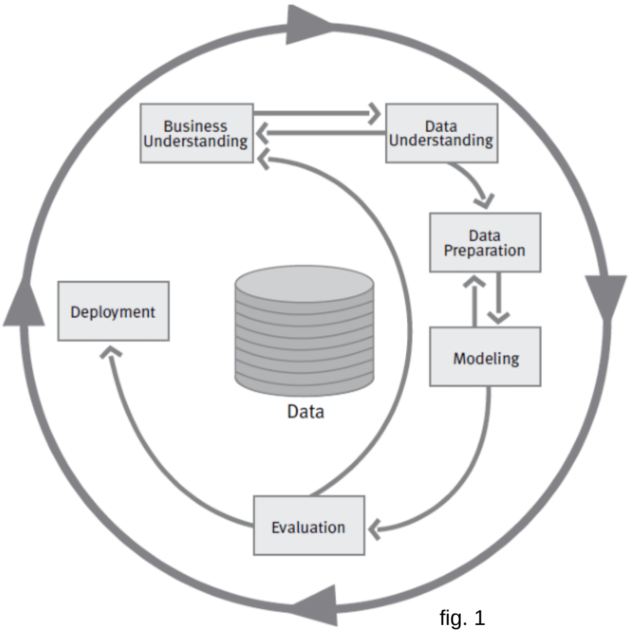


fig. 1

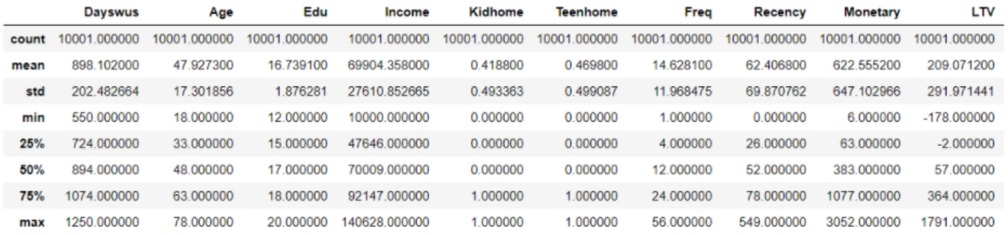
This model enabled us to create and define different and meaningful clusters representing WWW groups of clients. Having done all these steps, our team evaluated the achieved results comparing them with your business needs. Finally, we provide you the current solution.

3.1. Data understanding (in Jupyter Notebook, lines 3-7)

3.1.1.Variables:

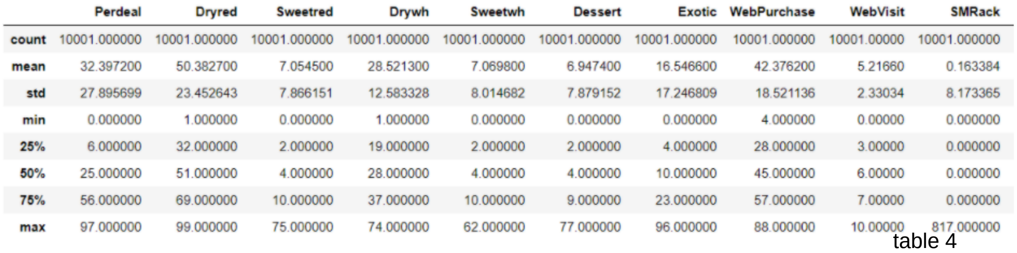
Our team worked in understanding the meaning of each variable taking into account their meaning in the process of reaching the business goal of this data mining analysis. We noticed there were variables related to demographic factors, sales patterns and sales platforms.

3.1.2.Descriptive Statistics:



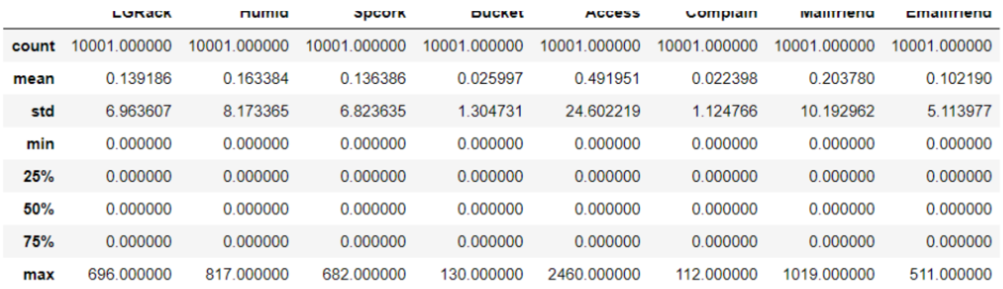
	Dayswus	Age	Edu	Income	Kidhome	Teenhome	Freq	Recency	Monetary	LTV
count	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000
mean	898.102000	47.927300	16.739100	69904.358000	0.418800	0.469800	14.628100	62.406800	622.555200	209.071200
std	202.482664	17.301856	1.876281	27610.852665	0.493363	0.499087	11.968475	69.870762	647.102966	291.971441
min	550.000000	18.000000	12.000000	10000.000000	0.000000	0.000000	1.000000	0.000000	6.000000	-178.000000
25%	724.000000	33.000000	15.000000	47646.000000	0.000000	0.000000	4.000000	26.000000	63.000000	-2.000000
50%	894.000000	48.000000	17.000000	70009.000000	0.000000	0.000000	12.000000	52.000000	383.000000	57.000000
75%	1074.000000	63.000000	18.000000	92147.000000	1.000000	1.000000	24.000000	78.000000	1077.000000	364.000000
max	1250.000000	78.000000	20.000000	140628.000000	1.000000	1.000000	56.000000	549.000000	3052.000000	1791.000000

table 3



	Perdeal	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	WebPurchase	WebVisit	SMRack
count	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.00000	10001.000000
mean	32.397200	50.382700	7.054500	28.521300	7.069800	6.947400	16.546600	42.376200	5.21660	0.163384
std	27.895699	23.452643	7.866151	12.583328	8.014682	7.879152	17.246809	18.521136	2.33034	8.173365
min	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	4.000000	0.00000	0.000000
25%	6.000000	32.000000	2.000000	19.000000	2.000000	2.000000	4.000000	28.000000	3.00000	0.000000
50%	25.000000	51.000000	4.000000	28.000000	4.000000	4.000000	10.000000	45.000000	6.00000	0.000000
75%	56.000000	69.000000	10.000000	37.000000	10.000000	9.000000	23.000000	57.000000	7.00000	0.000000
max	97.000000	99.000000	75.000000	74.000000	62.000000	77.000000	96.000000	88.000000	10.00000	817.000000

table 4



	LGRack	Humid	Spcork	Bucket	Access	Complain	Mailfriend	Emailfriend
count	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000	10001.000000
mean	0.139186	0.163384	0.136386	0.025997	0.491951	0.022398	0.203780	0.102190
std	6.963607	8.173365	6.823635	1.304731	24.602219	1.124766	10.192962	5.113977
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	696.000000	817.000000	682.000000	130.000000	2460.000000	112.000000	1019.000000	511.000000

table 5

3.1.3. Clean Noise:

In order to have quality data to apply the clustering methods and obtain accurate results we had to clean the noise. Our team noticed that the dataset did not contain missing values nor duplicated rows.

3.1.4. number of customers:

Despite WWW now having 350,000 customers in its database, our team based the study in a representative sample of 10, 000 clients provided by WWW.

3.2.Data preparation (in Jupyter Notebook, lines 8-26)

3.2.1.Outliers: Regarding the outliers, since the data was not hard to observe manually, we decided to analyze each one of the variables through the respective histograms and thus remove the noisy data. It is possible to visualize an example of this task in the figure at right, from the 'Sweetred' variable. (in Jupyter Notebook, lines 8-9)

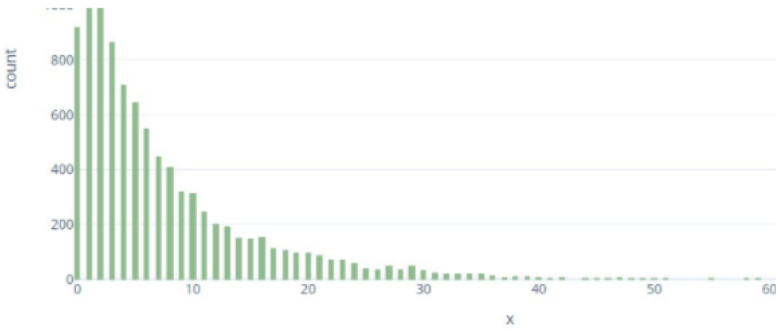


fig. 2

3.2.2. Feature engineering: Our goal was to produce tools to retrieve the maximum information from the dataset. To reach our objective we created new variables that are transformations of variables already existing in the dataset. The table at the right explains the new variables.(in Jupyter Notebook, lines 10-14)

Variable Name	Description	Creation
Dependents (new variable)	Indicates whether the customer has children or not.	We used *Kidhome* and *Teenhome* variables. We mapped each character to 0 or 1 and then we sum and then we mapped each number to 0 or 1 according if the customer has or not children.
Accessories (new variable)	Amount spent on accessories.	We used the variables *SMRack*, *LGRack*, *Humid*, *Spcork*, *Bucket*. We mapped each character to 0 or to the price of each accessories and then we sum all of prices of different accessories.

table 6

3.2.3. Feature selection: Taking always into consideration the perspectives our team found interesting to explore, a feature selection was performed.

The most appropriated path is to measure the correlation between the different variables in order to avoid redundant information entering in the clusters. We used a standard gauge in which we choose only 1 of 2 variables with a correlation equal or bigger than 0.85 between themselves.

Our criteria to choose between those variables was to compare again their correlation with the remaining variables, and the one with lower values was maintained, the other one was discarded.

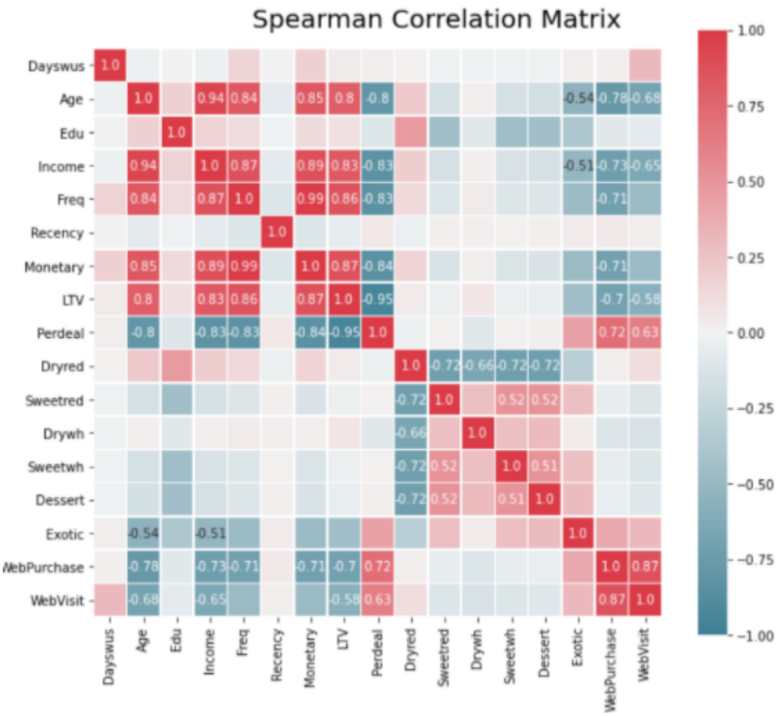


fig. 3

NOTE: We used a spearman correlation since it does not assume a linear correlation between variables.

Therefore, the variables elected to enter in the clusters were the following: 'Dayswus', 'Age', 'Edu', 'Recency', 'Monetary', 'Dryred', 'Sweetred', 'Drywh', 'Sweetwh', 'Dessert', 'Exotic', 'WebVisit', 'SMRack', 'LGRack','Humid', 'Spcork', 'Bucket', 'Access'.

The remaining numerical variables were deleted due to high correlations and the categorical variables were also deleted once they can not be used as clusters´ input, nevertheless they will be used for interpretation purposes.

Having reach this point, the image at right shows the correlations heat-map. (in Jupyter Notebook, lines 15-20)

3.2.4. Standardization: The goal of this section is to standardize the range of the continuous variables in order that each of them contributes equally to our analysis. Our team chose the MinMax scaler since this method will preserve the shape of our dataset and does not assume a normal distribution in the variables. (in Jupyter Notebook, lines 21-26)

3.3.Modeling (in Jupyter Notebook, lines 36-101)

3.3.1. Definition of perspectives: Firstly, to better conduct this study, we decided to approach the modeling techniques by different perspectives about the customer. There, we divided the variables into three groups, demographic, regarding customers age and education, client value, which includes variables such as Recency, Monetary, Dayswus, Webvisit and products containing products bought information, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic and Accessories. For each perspective the same clustering techniques were performed. Clustering:

- **SOM - Self Organizing Map:** Taking into consideration the characteristics of our dataset we believe the best approach is to use SOM as a clustering algorithm. In SOM, each neuron is a vector in the input space and, during the training, their position is adjusted and with them also came their neighbors. The input patterns are compared to each neuron and are finally assigned to one of them. Thus, the 'winner' neuron is updated and, with him, their neighbors.By applying this methodologie we were able to analyse the representation of each variable in it through the component planes visualization. Furthermore, to support the number of clusters to choose the U-matrix was used.We trained the SOM with a 50x50 grid in order to produce a clear U-Matrix so that it would be easier to understand the data structure.
  (in Jupyter Notebook, lines 38-41 / lines 58-61 / lines 73-76)

Example of a component plane:

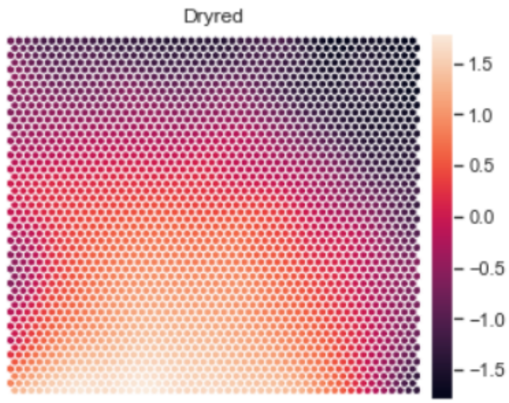


fig. 4

Example of a U-matrix:

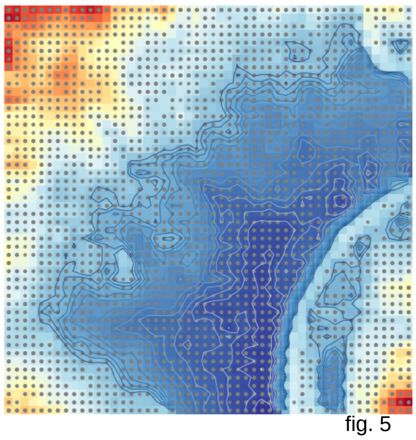


fig. 5

- **Hierarchical clustering:** On top of the SOM algorithm we applied a hierarchical technique as the good practices recommend. Therefore, we used the AgglomerativeClustering algorithm before assigning labels to each observation. Using the hierarchical clustering we can also verify the most appropriate number of clusters chosen for each perspective.After applying both techniques to each approach we joined all the three perspectives together in order to have a final clustering solution.
  (in Jupyter Notebook, lines 42-45 / lines 62-64 / lines 77-79)

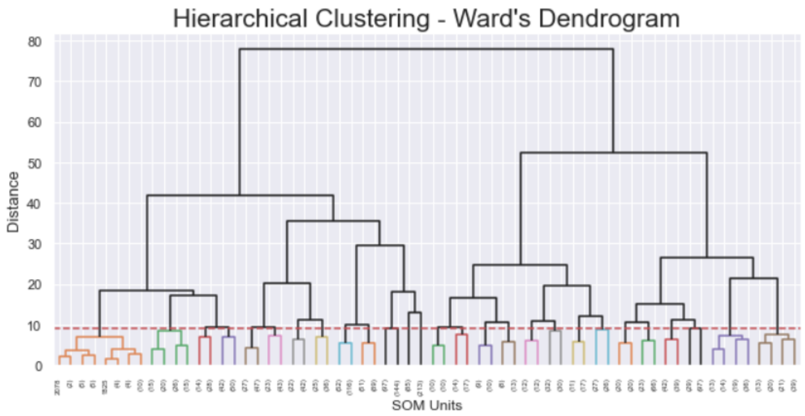


fig. 6

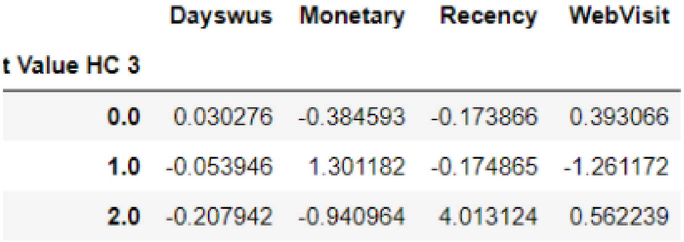
3.4. Evaluation (in Jupyter Notebook, lines 169-178)

With the purpose of assessing whether the techniques applied were the most appropriate ones, we computed the r-squared which represents the accuracy. This value is scaled between zero and one, zero meaning a very bad solution and one representing a really good solution.

Cluster Assessment By Perspective:

- Client Value: (in Jupyter Notebook, lines 47-56)

The obtained r-squared had a value of 0.39. This perspective was divided into 3 clusters and observing the figure below it is possible to affirm that the cluster 0.0 is composed by customers that have been in the WWW database for the longest period, spend moderately - not the bigger nor the lower spenders. They are also people with a low recency and sometimes they visit the website.



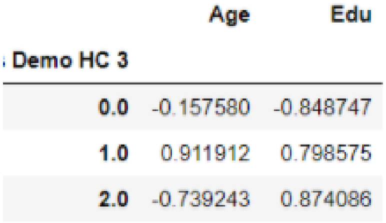
t Value HC 3	Dayswus	Monetary	Recency	WebVisit
0.0	0.030276	-0.384593	-0.173866	0.393066
1.0	-0.053946	1.301182	-0.174865	-1.261172
2.0	-0.207942	-0.940964	4.013124	0.562239

table 7

The cluster 1.0 hosts customers who are with the WWW since a short period of time but are the ones who have been spending the most, with a low recency. They do not prefer to check the website.
The cluster 2.0 are the one with the most recent customers, the ones who had spent the least and the ones with higher recency. They are the ones who like the most to visit the website.Taking this in consideration the ranking of the most valuable customers is: Cluster 1, Cluster 0, Cluster 2.

- Demographic: (in Jupyter Notebook, lines 66-71)

The obtained r-squared had a value of 0.75. This perspective was divided into 3 clusters and observing the figure below it is possible to affirm that the cluster 0.0 is composed by people in middle age that do not have a high level of education.
The cluster 1.0 has older customers with a considerable level of education.
The 2.0 cluster is the one with the youngest people and also the ones higher educated.

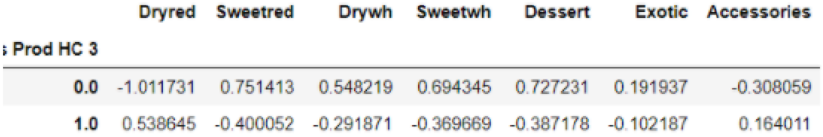


Demo HC 3	Age	Edu
0.0	-0.157580	-0.848747
1.0	0.911912	0.798575
2.0	-0.739243	0.874086

table 8

- Products: (in Jupyter Notebook, lines 81-86)

The obtained r-squared had a value of 0.59. This perspective was divided into 2 clusters and observing the figure below it is possible to affirm that the cluster 0.0 is composed by the customers who prefer all the other products instead of the Dry Red Wine and the Accessories.
The clusters 1.0 is the one with people who have preference for the Dry Red Wine and Accessories.



Prod HC 3	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	Accessories
0.0	-1.011731	0.751413	0.548219	0.694345	0.727231	0.191937	-0.308059
1.0	0.538645	-0.400052	-0.291871	-0.369669	-0.387178	-0.102187	0.164011

table 9

- Final Cluster Assessment: (in Jupyter Notebook, lines 92-101)

The previous 3 perspectives were merged in order to join their individual information into a global perspective of WWW customers. The obtained r-squared had a value of 0.41. This solution is composed by 4 different groups of customers and observing the figure below it is possible to affirm that:

0

the cluster 0.0 customers: are not the newest nor the oldest clients, spend a significant amount and have been buying recently. In general, do not visit the website and are people in, at the minimum, the middle age. These are the higher educated clients and also the biggest fans of Dry Red Wine but they also appreciate the Accessories. A lot of people in this cluster have dependents (kids or teens).

1

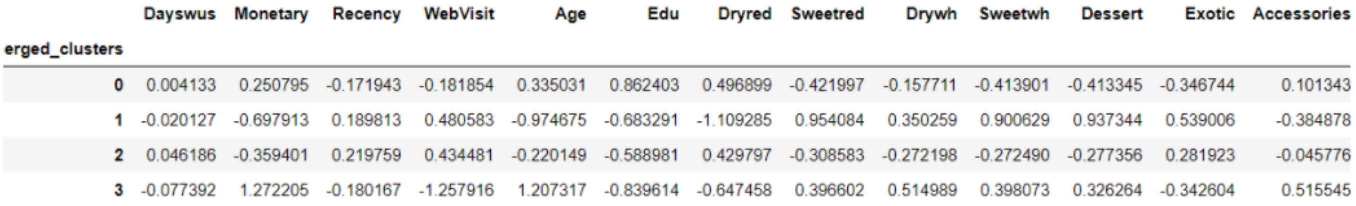
the cluster 1.0 customers: are with WWW for a small period of time but are the ones who spend the least. It Has been a while since they have not bought any product and they are the biggest users of the website. This group has the youngest customers with a medium level of education. Their preference really goes to Sweet Wines and Desserts but they also like Dry White wine and Exotic products.

2

the cluster 2.0 customers: are the oldest clients, and do not spend a lot in WWW. Are the ones who have been most away from WWW but visit the website a lot. This cluster has young people with a medium/low level of education. Their preference goes to Dry Red Wine and Exotic products. These groups appear a lot in the purchased list of “email friendly” and “mail friendly” customers. These clients tend to have a high value of dependents (kids or teens) and are the ones who complained the most in the last 18 months.

3

the cluster 3.0 customers: Are the newest clients and also the ones with the higher monetary value. They have been shopping recently and are the ones who use the less the website. These are WWW´s oldest clients in age and the ones with a lower education level. This cluster has the specificity that people here do not buy Dry Red Wine neither Exotic products and their favourite products are Dry White Wine and the Accessories. This group of customers are the one who appears less in the purchased list of “email friendly” / “mail friendly”customers.



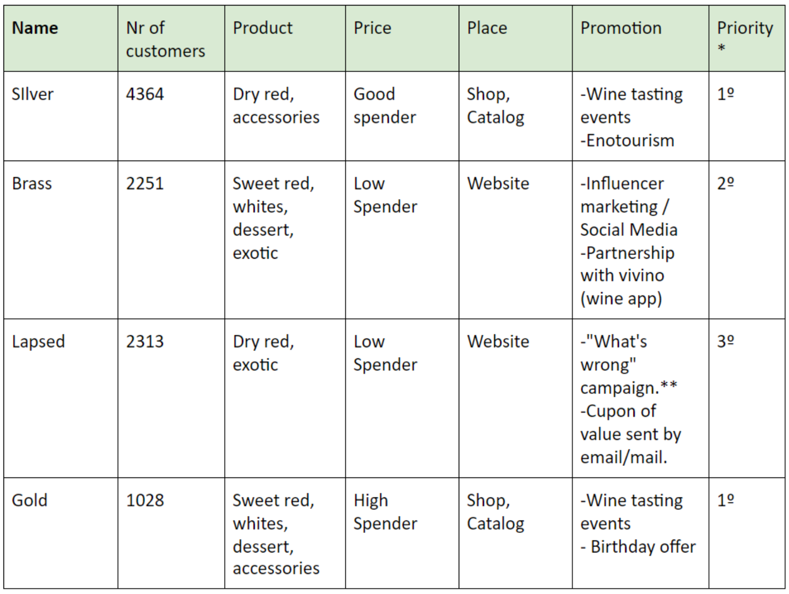
erged_clusters	Dayswus	Monetary	Recency	WebVisit	Age	Edu	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	Accessories
0	0.004133	0.250795	-0.171943	-0.181854	0.335031	0.862403	0.496899	-0.421997	-0.157711	-0.413901	-0.413345	-0.346744	0.101343
1	-0.020127	-0.697913	0.189813	0.480583	-0.974675	-0.683291	-1.109285	0.954084	0.350259	0.900629	0.937344	0.539006	-0.384878
2	0.046186	-0.359401	0.219759	0.434481	-0.220149	-0.588981	0.429797	-0.308583	-0.272198	-0.272490	-0.277356	0.281923	-0.045776
3	-0.077392	1.272205	-0.180167	-1.257916	1.207317	-0.839614	-0.647458	0.396602	0.514989	0.398073	0.326264	-0.342604	0.515545

table 10

3.5. segmentation

Marketing approaches to achieve new clients:

- Promotion in wine blogs
- Partnerships with the enotourism industry
- Social media marketing
- Influencer marketing
- Website only promotions (less 30% in online shopping, for example)
- Sponsorship of local events



Name	Nr of customers	Product	Price	Place	Promotion	Priority *
Sllver	4364	Dry red, accessories	Good spender	Shop, Catalog	-Wine tasting events -Enotourism	1º
Brass	2251	Sweet red, whites, dessert, exotic	Low Spender	Website	-Influencer marketing / Social Media -Partnership with vivino (wine app)	2º
Lapsed	2313	Dry red, exotic	Low Spender	Website	-"What's wrong" campaign.** -Cupon of value sent by email/mail.	3º
Gold	1028	Sweet red, whites, dessert, accessories	High Spender	Shop, Catalog	-Wine tasting events - Birthday offer	1º

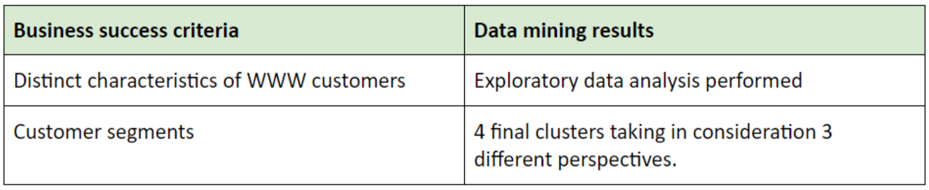
table 11

* Gold and Silver clients are the most valuable so we suggest to prioritize them. Brass are new clients and need to be motivated in order to be more loyal and frequent. Finally, Lapsed are old clients that are losing the relationship with WWW so the objective is to bring them back.

**"What's wrong campaign" will be targeted to the lapsed customers, it will consist of a questionnaire with the objective to understand what led to the churn observed. For all customers a loyalty program with specific promotions should be created, this strategy would increase the frequency of purchases and reinforce the relationship between the customer and the company.

4. RESULTS EVALUATION

- To meet the business objectives the marketing mix, concerning the product, price, place and promotion for each group of clients, was completed as well as the delivery deadline of the project.



Business success criteria	Data mining results
Distinct characteristics of WWW customers	Exploratory data analysis performed
Customer segments	4 final clusters taking in consideration 3 different perspectives.

table 12

5. DEPLOYMENT AND MAINTENANCE PLANS

Plan Deployment

There were some issues regarding the data provided (such as one row with absurd values, probably badly introduced in the database) that could be resolved with a better data collection process. There are some features that could benefit from a re-work, it would be more valuable to see how many times a customer bought a small wine rack instead we only know if he bought it or not, and it would be useful to know the difference between the store and catalog preferences.The scope of this project was to generate some business insights which will be done by presentation to the board of directors and using the conclusion of this report.

Maintenance Plans

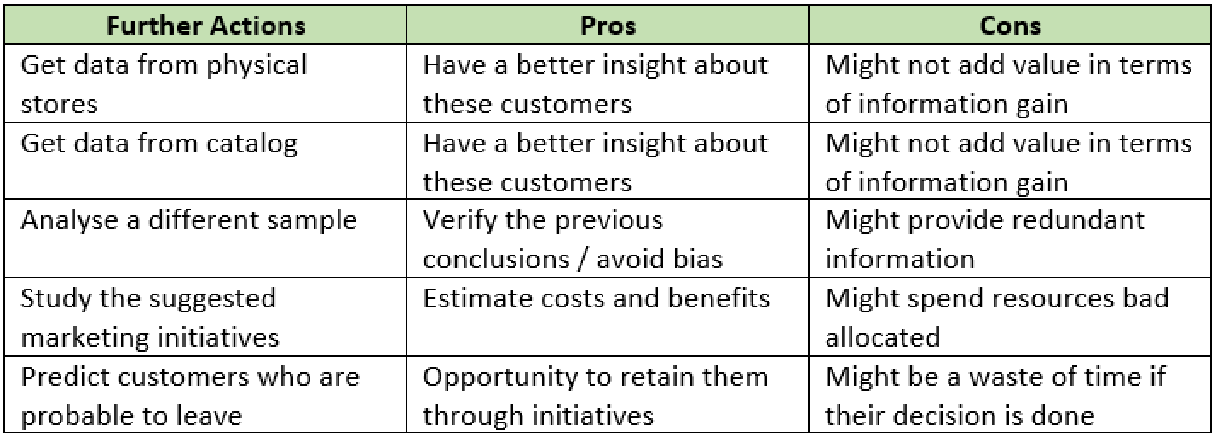
The algorithms used for this analysis might degrade in the future so if the goal of WWW is to keep track of how their customers behave this should be taken in consideration.

6. CONCLUSIONS

To sum up this project, after the cluster analysis with three different approaches, we were able to identify customers' groups for "Wonderful Wines of the World" and specified the marketing campaigns to each group.

With the implementation of these recommendation plan, WWW could see customer satisfaction increase by customizing their relation.

8. FURTHER ACTIONS



Further Actions	Pros	Cons
Get data from physical stores	Have a better insight about these customers	Might not add value in terms of information gain
Get data from catalog	Have a better insight about these customers	Might not add value in terms of information gain
Analyse a different sample	Verify the previous conclusions / avoid bias	Might provide redundant information
Study the suggested marketing initiatives	Estimate costs and benefits	Might spend resources bad allocated
Predict customers who are probable to leave	Opportunity to retain them through initiatives	Might be a waste of time if their decision is done

table 13

8. REFERENCES

- Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) (2000)."Crips-DM 1.0".Step-by-step data mining guide.
- Spacey, John-Simplicable- January 28,2016 - https://simplicable.com/new/contingency-plan - February 27, 2021.
- Smartsheet-An Expert Guide to Cost Benefit Analysis - https://www.smartsheet.com/expert-guide-cost-benefit-analysis - February 27, 2021.
- Lucas Bação, Fernando - Business Cases with Data Science- February 8, 2021.