BUSINESS CASES WITH DATA SCIENCE

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

Wonderful World of Wines:
Customers segmentation.
How well do you know your customers?

Group O

Eleonora Sbrissa, M20200628

Luis Reis, M20200636

Pedro Godeiro, M20200396

Sara Michetti, M20200626

March, 2021

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

INDEX

INTRODUCTION

The dataset was provided by Wonderful Wines of the World (WWW), which is a 7-year-old wine company that delights their customers with multiple wine types. The company mostly makes their selling using three main channels: catalogs (telephone), a web site, and ten small stores in major cities around the USA. This team of data scientists was hired by the CEO of the company to identify the differences that exist among their clients, the different characteristics and the different segments to which each client belongs. The company would like to better their customers' knowledge, to make new marketing strategies in the future that are more specific to different types of customers, to both reach new and existing buyers.

1. BUSINESS UNDERSTANDING

1.1. BACKGROUND

What do we know about Wonderful Wines of the World (WWW)?

- **Company Age**: 7-year-old enterprise
- **Product**: Wine Sales. Several hundred selections
- **Mission**: delight customers with well-made, unique and interesting wines
- **Sales channels**: Catalogs (telephone), website, 10 small stores in major cities around USA
- **Customers:** around 350k customers in the database
- **Past customer acquisition**: aggressive promotions in wine and food magazine
- No loyalty programs

1.2. BUSINESS OBJECTIVES

The company would like to understand the characteristics of their customers, in order to identify target markets for cross-selling opportunities. The goal is to make new strategic marketing choices through the identification of different groups of customers and behaviors.

1.3. BUSINESS SUCCESS CRITERIA

The business success criteria consist in clearly differentiating the segments of customers. For example, who are the customers that prefer discounts or web purchases; or if there is any group that favor red wines against the exotic ones; if customer's age is a relevant variable, how to use it to communicate with the segments.

1.4. SITUATION ASSESSMENT

The *Wonderful Wines of the World* team gave access to a sample of 10.000 observations out of a database of 350.000 customers. The dataset, contained in an excel file, is in a table format in which each row is a different customer. The figures reflect the last 18 months behavior of the buyers.

The Team we are in contact with is composed by the CEO, Fernando, and two analysts, David and Joao. They got the data from the IT department but there is no access to this team to ask further questions.

For example, we are not aware about what “the last 18 months” mean, if it was from February 2021 or from before the month we are working in.

Some variables are not known by the team either, like *Rand* or *LTV*. It seems that *Rand* was merely a random number and *LTV* represents the lifetime value of the customer but calculated by an analyst that doesn’t work for the company anymore.

An assumption that was made is that all customers that bought from the website are eligible to drink.

Right now, the marketing campaign is not based on customer knowledge, hence it is not targeted. All customers get the catalog, and there are no loyalty programs nor attempts to identify target markets for cross-selling opportunities. This may produce a high cost on marketing, with low return on investment. With this clustering and identification of different profiles of customer, the company may focus on specific targeted markets and improve the ROI and without increasing marketing budget.

To analyze the data the programs used are Python in the Google Colab environment and for data visualization purposes Power BI will be used.

1.5. DETERMINE DATA MINING GOALS

The data mining goal is to identify, via unsupervised learning techniques, different clusters based on the variables given in the dataset.

To do so a new dataset will be delivered, with new engineered features and cleaned data, that will be used to determine the different characteristics that best distinguish your customers and how to reach new ones from each group.

The success of this project depends on the results of the clustering algorithm and on the goodness of fit of the clusters.

2. PREDICTIVE ANALYTICS PROCESS

2.1. DATA UNDERSTANDING

The initial dataset contains 10.001 rows and 30 variables related to Demographics (like age, income, number of children and education level), Behavior of the customers (like the percentage of wines bought per category, the percentage of purchases made with discount, the number of accessories bought) and technographics (web visits and web purchases).

From the 30 variables, 19 were numerical, 10 boolean and 1 categorical.

For a first analysis we used Pandas Profiling in order to check data distribution, data consistency and to detect missing data: there were no missing values; there were no duplicated observations either.

2.1.1. What do we know of our customers?

Customers are on average 47 years old and range from 18 to 78. Income is normally distributed among customers.
They bought at least once online and have been registered on the website since at least 18 months.

The average monthly visit is 5.21. The average value spent on the website for the last 18 months is 622$ and we know that online sales represented 42% of the total sales.

Below we can see the online sales related to the last 18 months, divided by type of wine.

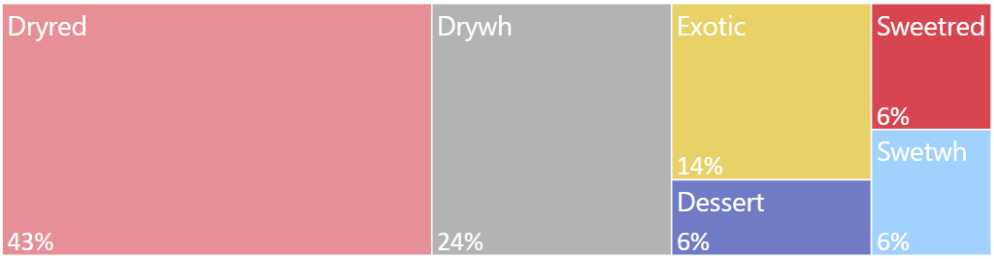

Figure 1 - Sales distribution per type of wine (last 18 months)

It seems that customer prefer in general dry red wine, followed by dry white wine.

Here below in figure 2 an analysis on web purchases average based on different age groups. We can see that the older the people, the less is the average of online purchases.

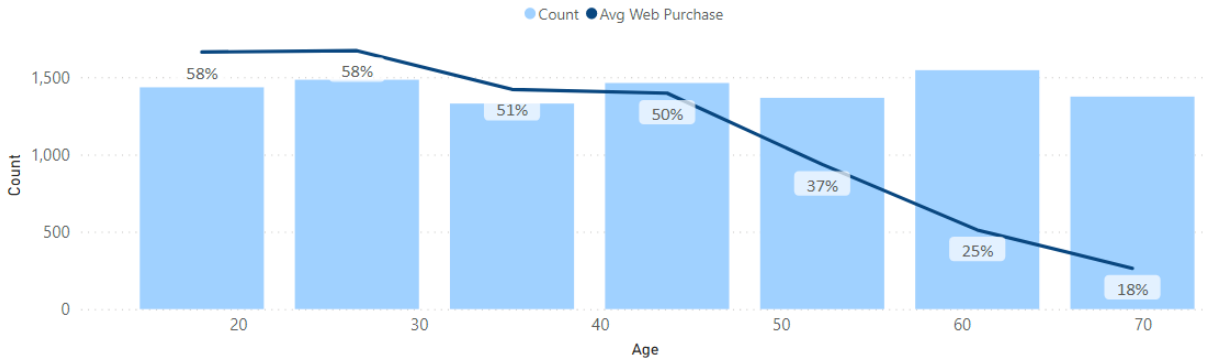

Figure 2 - Influence of age on web purchases

We also checked that only 3142 accessories have been bought by the customers in these last 18 months.

To have a clearer understanding of the relationship between features, Pearson correlation matrix for the numerical variables has been analyzed. In the code you may also find the Spearman correlation matrix for the Boolean variables.

2.2. DATA PREPARATION

First, the last column “Rand” (there was no information about it) has been dropped and the last row has been deleted too as it is the mean of the values and it does not represent a customer. The LTV variable has been dropped too as it is highly correlated with Monetary and Frequency variables, and the company is not sure of its meaning.

2.2.1. Outliers

Outliers have been detected. Based on the boxplots only Recency was considered as critical, being the number of days that passed since the last purchase. It has been assumed that customers who bought

more than 200 days ago (outside the Inter Quartile Range) are lapsed customers, hence they have been saved in a different table and taken out from the main dataset. It was also highlighted that these customers just bought once on the website.

2.2.2. Feature Engineering

Feature engineering has been performed in order to add some meaningful information to the dataset. Table 1 shows the variables created:

Engineered Feature	*Definition*	*Preprocess*
AVG_PURCH	Average purchase of the last 18 months	division between Monetary and Frequency
ACC_SPCORK	Number of accessories bought by the customers in total, taking SPCORK in consideration as well	Sum of all the accessories variables in the dataset
NUMB_WEB_PURCH_PER_MONTH	web purchases per month per client	Product of Web_purchases and Frequency normalised by 18
CONVERSION_RATE	% web purchases on the number of visits of the website	Number of purchases online divided by WebVisit

Table 1 – Feature Engineering

2.2.3. Correlation and Feature Selection

Feature selection is important to identify the variables that will be used for the modeling. It's key to delete redundancy in the data and to reduce dimensionality. Both correlation matrix and PCA have been calculated to perform this task.

In figure 3 the correlation matrix for numerical variables. The percentages of types of wines sold are not shown in the correlogram as it was seen that they are not correlated with any other variable, and they will all be used in the clustering algorithm.

From the matrix Income and Age have the same behavior towards the other variables, hence it was decided to drop **Income** and keep Age. Also, **Frequency** has been dropped as it has a linear correlation of 0.99 with Monetary.

PCA was used to understand the variables to use for the clustering algorithm. Both **Recency, Days with us** and **Accessories** did not have a significant beta on the first three principal components (that account for 62% of the variance) hence they have been deleted. For reference, please check the PCA section on the code.

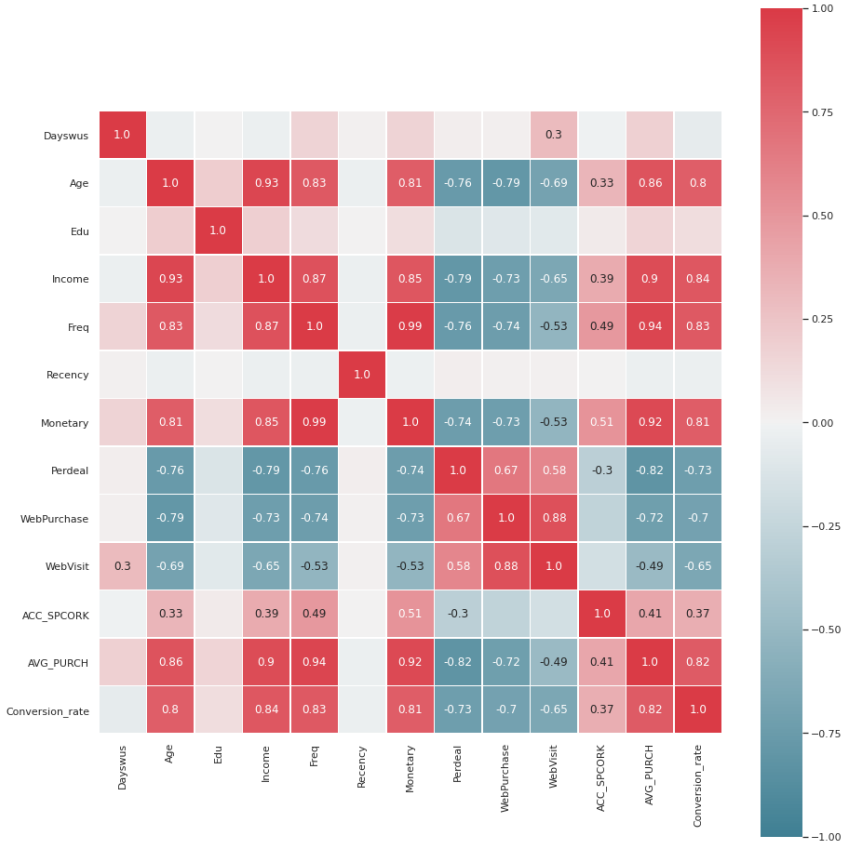

Figure 3 – Correlogram

Below follows a detailed description of the variables that were selected for the modeling:

Variable Name	***Range***	***Description***
Edu	12-20	Years of education
Monetary	$6-$3052	Total Sales in the past 18 months
Age	18-78	Customer's age (we considered 18 years old as a legal age to drink)
Perdeal	0-97%	% purchases with discount
Dryred	1-99%	% of dry red wines bought
Sweetred	0-75%	% of sweet or semi-dry red wines bought
Drywh	1-74%	% of dry white wines bought
Sweetwh	0-62%	% of sweet or semi-dry white wines bought
Dessert	0-77%	% of dessert wines bought
Exotic	0-96%	% of very unusual wines bought
WebPurchase	4-88%	% purchases made on the website in the past 18 months
WebVisit	0-10	Average number of visits to website per month
AVG_PURCH	6-54$	Average purchase related to the past 18 months
Conversion_rate	0-33%	% of web purchases on the number of visits of the website

Table 2 – Features selected for clustering

2.2.4. Scaling

Lastly, all the variables that will be included in the model have been scaled using Robust Scaler.

2.3. MODELING

To model our data and find the meaningful clusters the process used was the following:

- Hierarchical clustering over 100 cluster found via k-means
- Find the right number of clusters using Ward's Dendogram
- Use k-means algorithm with the number of clusters identified

To proceed with the clustering, only numeric variables have been used (Table 2).

To use the right distance calculation between clusters for the hierarchical algorithm, the R^2 for every cluster solution has been calculated (the plot can be seen in the code, under the section "Hierarchical Clustering"). Based on these results, the ward distance was the most appropriate one. The dendogram for 100 clusters is shown in figure 4.

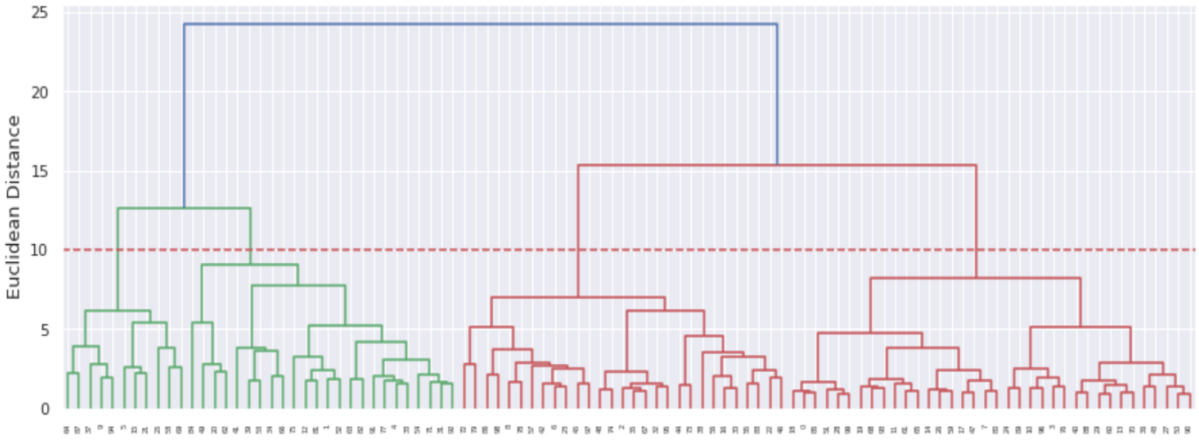

Figure 4 – Hierarchical clustering: Ward's Dendogram

It seems wise to select an amount of 4 clusters to proceed with K-means to have enough differentiation in the characteristics of the customers without getting too detailed. Speaking also from a marketing point of view it's important not to have too many different customers segmentation.

The k-means algorithm resulted in the following groups:

labels	Edu	Monetary	Perdeal	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	WebPurchase	WebVisit	Age	AVG_PURCH	Conversion_rate
1	15.0	116.0	52.0	16.0	19.0	26.0	19.0	20.0	40.0	55.0	6.0	26.0	18.0	2.0
3	17.0	136.0	55.0	44.0	7.0	36.0	7.0	6.0	23.0	55.0	7.0	34.0	21.0	2.0
0	18.0	534.0	32.0	76.0	2.0	18.0	2.0	2.0	10.0	47.0	6.0	49.0	33.0	6.0
2	17.0	1361.0	5.0	43.0	8.0	34.0	7.0	7.0	8.0	21.0	3.0	67.0	47.0	12.0

Table 3 – k-means clusters

2.4. EVALUATION

To assess the model a R^2 metric was used, resulting in 0.6225 goodness of fit.

The R^2 was then decomposed into the R^2 for each variable, to get an idea of the importance of each variable in the clustering. In table 5 it is possible to see them.

Age	0.745214	WebVisit	0.493485
Dryred	0.686808	Dessert	0.448999
AVG_PURCH	0.660792	Sweetwh	0.417993
Monetary	0.623570	Drywh	0.398522
WebPurchase	0.614507	Sweetred	0.394641
Conversion_rate	0.601417	Exotic	0.378108
Perdeal	0.520333	Edu	0.231271

Table 5 - R^2 Evaluation for each variable

3. RESULTS EVALUATION

Based on the data shown in table 3, some visualizations have been created to understand better the different groups of customers. Taking the most relevant variables, in figure 5 it is possible to notice how customers in every group behave differently. For a better understanding of the results a description for every cluster has been identified:

Cluster 0: the average customer, likes deals
Cluster 1/3: the youngest, deal lover, low spender
Cluster 2: the eldest, prefer catalog or store purchase, goes to the website just to buy, high spender

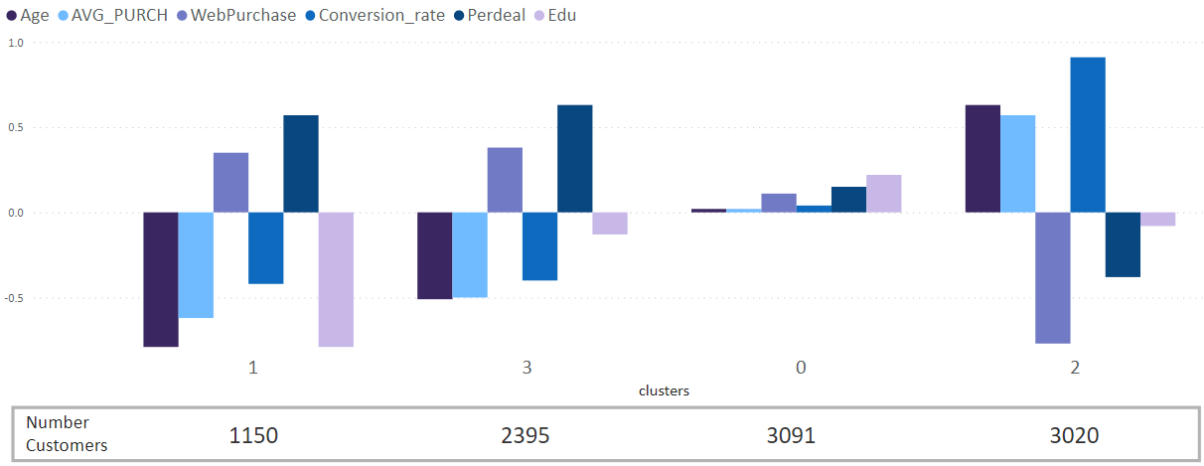

Figure 5 - Clusters characteristics

The reason why there is still four clusters and not three lies on the wine preferences for every cluster. Our young customers, based on where they sit, have different wine taste. In figure 6 it is possible to see how cluster 3 prefers dry red wines (36%) while cluster 1 prefers the exotic ones (29%). Interesting to notice that cluster 0, the average customer, consumes 70% of dry red wine, 16% of dry white, without largely buying the other categories.

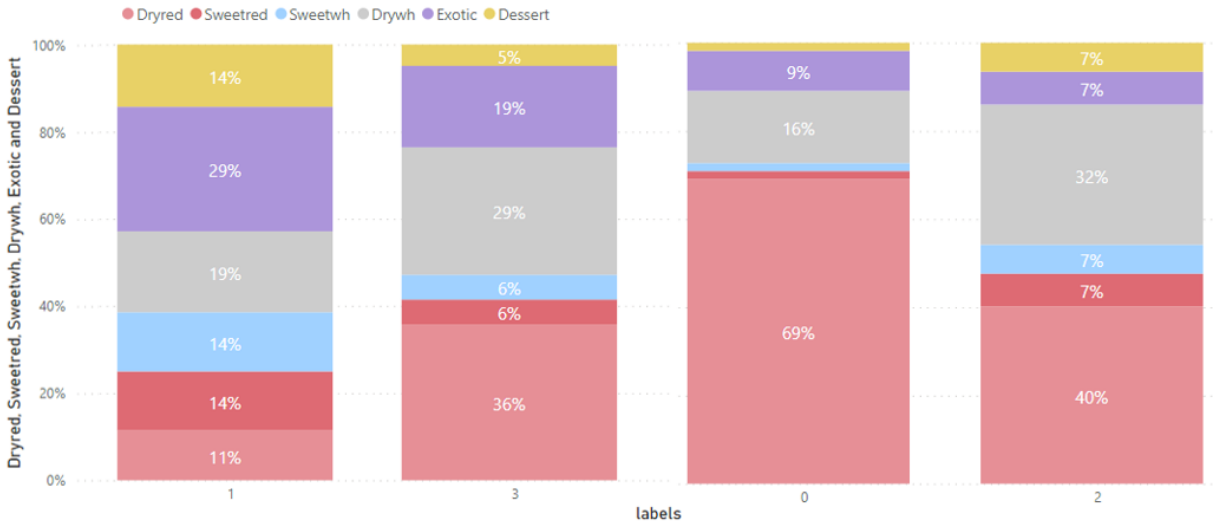

Figure 6 - WebSales of different wines category per cluster

It was also seen that the elder people (Cluster 2) bought on average more accessories than the rest.

It is possible, based on the previous analysis, to have different marketing campaigns for different customers both looking at it from a taste and purchase behavior perspective. This meets both the business goal and the data mining goal set at the beginning.

4. DEPLOYMENT AND MAINTENANCE PLANS

4.1. DEPLOYMENT

The teams involved for the deployment are IT, marketing and sales.

IT needs to make sure that the data is collected as before, and additionally, calculate the new features, including the cluster to which the customer belongs to.

The marketing team can then use those clusters and do new targeted marketing campaigns. Here some suggestions:

- Cluster 0: Average Customer, they tend to buy more on promotions
  As the percentage of dry red wines is really high, sales promotion should be focused more on discounts on the other categories, in order to increase interest and sales on these. The campaign can be done online as the web purchases are slightly higher than average
- Cluster 2: Elder people, they don't buy online most of the times, but the conversion rate is really high as when they visit the website they go there to purchase.
  On average they tend to buy accessories, so a marketing strategy suggestion can be based on product bundling, to match wines with accessories. This promotion should be done offline through mails and phone.
- Cluster 1/3: Low spenders, have a tendency to buy more on deals
  For these clusters you should try to increase the average purchase also giving more deals.
- For all the clusters, try to stick with the different wine tastes

The sales team can then understand where to invest and divest regarding the wine categories.

The customers that were considered as lapsed buyers (the ones who bought more than 200 days ago) were also analyzed. It was seen that they bought just once on the website. A different marketing campaign could be done for them, for example offering a promotion on the same type of wine bought the first time.

4.2. MAINTENANCE

As for the maintenance, after 2 weeks, 1 month, 3 months and 6 months it is needed to check if the cluster approach is improving the business overall. Are sales improving? By how much? Are the wines tastes reflecting the sales trends? Did the conversion online increased?

The suggestion is to recalculate them via k-means every season. When the company reaches enough seasons, the advice is to make a comparison among them and check the overall behaviors across clusters.

5. CONCLUSIONS

Starting from a dataset containing 30 variables and 10k rows, the team was able to analyze the data, identify the most important features, to create new valuable ones and to group customers into four different clusters, achieving a 70% on R^2. For each cluster, the team was also able to suggest different marketing campaigns. We hope the company can take some valuable insights from this analysis in order to improve the strategy and increase sales. Some considerations regarding the model have been done.

5.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

To improve the model few things are needed. The data analyzed for this business case is in a way too general to be more specific around which deals to make and to who. More detailed information would be ideal. For example: did the customer always buy the same wine? Was it always on promotion or not? If they purchase in store, do they buy the same products? Hence data from physical stores and catalogs are needed too. It would be also ideal to get the reviews for the wines from the website.

Another aspect to take in consideration is: what happens when a new customer registers but has not done any purchase yet? Which kind of marketing campaign can you give him? We are suggesting here the use of a decision tree based on what found on in this project.

6. REFERENCES

https://fourweekmba.com/customer-segmentation/

https://www.wordstream.com/conversion-rate

https://github.com/davidsilva98/DMDSAA