

**BUSINESS CASES WITH DATA SCIENCE**

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

**Business Case #1 - Wine store**



Data4Business Consulting

Débora Santos, number: 20200748

Diana Furtado, number: 20200590

Pedro Henrique Medeiros, number: 20200742

Rebeca Pinheiro, number: 20201096

February, 2021

INDEX

[1. INTRODUCTION 1](#_Toc65425090)

[2. BUSINESS UNDERSTANDING 1](#_Toc65425091)

[2.1. Background 1](#_Toc65425092)

[2.2. Business Objectives 2](#_Toc65425093)

[2.3. Business Success criteria 2](#_Toc65425094)

[2.4. Situation assessment 2](#_Toc65425095)

[2.4.1. Inventory of resources 2](#_Toc65425096)

[2.4.2. Requirements, assumptions and constraints 3](#_Toc65425097)

[2.4.3. Risks and contingencies 3](#_Toc65425098)

[2.5. Determine Data Mining goals 4](#_Toc65425099)

[2.6. Project Plan 4](#_Toc65425100)

[3. CLUSTERING ANALYSIS 5](#_Toc65425101)

[3.1. Data understanding 5](#_Toc65425102)

[3.2. Data preparation 5](#_Toc65425103)

[3.3. Clustering (K-means) 7](#_Toc65425104)

[3.4. Evaluation 9](#_Toc65425105)

[4. RESULTS EVALUATION 9](#_Toc65425106)

[5. DEPLOYMENT AND MAINTENANCE PLANS 9](#_Toc65425107)

[6. CONCLUSIONS 9](#_Toc65425108)

[6.1. Considerations for model improvement 9](#_Toc65425109)

[7. REFERENCES 9](#_Toc65425110)

[8. APPENDIX (OPTIONAL) 10](#_Toc65425111)

# 

# INTRODUCTION

Thank you for choosing **Data4Business Consulting(D4B)** to help you in the challenge of understand better your customers’ characteristics and segmentation. Our main objective is helping your business to improve its set-up, increase your gains, bring new customers, and retain the current ones.

The world is experiencing a great technological and digital revolution where understanding business data, customers segmentation and their necessities are essential for the business success. The exponential technological advances, such as data mining techniques, artificial intelligence, internet of things and more can help the business to have great advance.

Through innovative technological programs, well-referenced data mining methods and insights of marketing digital, the present report intends to give an overview of the process behind the analysis, presents the results and provides insights you need to be successful in new era.

We are excited to be a part of this challenge.

# BUSINESS UNDERSTANDING

## Background

**Wonderful Wines of the World (WWW)** has been present in the wine market for 7 years. The company aims to provide customers with a premium selection of wine and wine accessories.

The key persons in this business are the owner (Fernando Bação) and the managers (João Fonseca and David Silva). Fernando is interested on increasing wine and accessories selling. João and David are looking on the actions needed to get the outcome the owner expects.

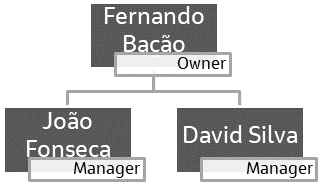


Figure 2.1 - Organizational chart.

During the 7 years of existence, this company has been using a marketing strategy based on previous experience on how to sell more: send a catalog (which is renewed every 6 weeks) to the 350,000 customers on the list (from the past 4 years) and expect the customer to approach the company to buy wine and accessories. Also, from previous analysis, the company already knows that most of its customers are wine lovers who have no financial constrains to get good quality wine.

The customers currently have three different ways of purchasing wine and accessories from WWW: in person (through one of the ten stores WWW has in major cities around the USA), by telephone (through the catalog) or online (on WWW’s web site).

The problem area this business is now trying to achieve is how can they improve the wine and accessories selling (not only to existing, but also new customers) using the knowledge on the current customers and this is why they have reached Data4Business Consulting.

## Business Objectives

At this point, WWW there is no specific knowledge about the customers. All marketing actions are based on market reports, information from sales team and intuition.

The customer’s primary objective is to increase wine and accessories selling by understanding the following:

* Identify key characteristics that best distinguish the customers.
* Which and how many customer segments there are in the provided database.
* Understand how the business can reach new and existing customers from each segment and which ones should be prioritized.
* Improve the interaction with the customers by creating new marketing strategies.

## Business Success criteria

The expected outcome will be well defined customers’ segments which can make possible to build a customized marketing strategy and maximize the return of investment.

Another expected outcome of this report is suggestion of marketing strategies and business applications for the findings.

The success of the proposed task will be evaluated by WWW’s owner and managers and, if needed, we will go back to the model until we get an outcome that matches with the board’s expectation.

## Situation assessment

### Inventory of resources

This project was made following the CRISP-DM reference model (Cross Industry Standard Process for Data Mining). CRISP-DM is a standard process built in the end of 90’s and it was built by more than 200 members lead by a consortium of big companies. *CRISP-DM succeeds because it is soundly based on the practical, real-world experience of how people conduct data mining projects. And in that respect, we are overwhelmingly indebted to the many practitioners who contributed their efforts and their ideas throughout the project.[5]*

From WWW side, the project has the support of Management as well as the IT team.

From D4B Consulting, this project will be conducted by the following team:

* Débora Santos (Executive sponsor)
* Diana Furtado (Project leader)
* Pedro Medeiros (Data miner)
* Rebeca Pinheiro (Data expert)

We have been provided by the WWW’s IT team with a database of the customers who purchased in the last 18 months, composed by 10,000 customers and 29 attributes of them. It was also provided a metadata file of this dataset.

The main technology used to achieve the objectives of this report was Python. Python is one of most important and commonly used program languages in data science projects.

We also used Prezi application to produce the final presentation to the board.

### Requirements, assumptions and constraints

The completion date of this phase of this project is March 1st, 2021. But we expect to continue giving support and helping WWW to achieve the next goals for the growth of the business.

Even though the sale of alcoholic beverages to people under 21 is prohibited in the USA, on this project we considered all customers with age 18 or more that bought in the past 18 months.

The dataset provided has only 10.000 customers even the total database has 350.000 customers. One of the assumptions is that 10.000 customer will well represent the entire data.

### Risks and contingencies

Table 2.1 identifies a list of risks and contingency proposed.

|  |  |
| --- | --- |
| **Risk** | **Contingency** |
| Insufficient number of features | Work with remaining features or ask for different variables |
| Insufficient observations | Ask for more observations (customers) |
| Model overfitting[[1]](#footnote-1) | Ask for observations |

Table 2.1 - Risks and contingency.

#### Terminology

***Business glossary***

* Small wine rack, large wine rack, wine cellar humidifier, plated cork extractor and silver wine bucket are all wine accessories:

1. Wine cellar humidifier is designed to increase humidity levels in any size commercial or residential wine room.
2. Wine rack is some furniture to keep the wines.
3. Plated cork extractor is used to open the wine bottle.
4. Silver wine bucket is a premium accessory to keep the wine temperature during its consumption.

* Dry red wines, sweet or semi-dry reds wines, dry white wines, sweet or semi-dry white wines, dessert wines (port, sherry, etc.) and Exotic wines are types of wines according to its ingredients, flavours, and other characteristics.

***Data mining glossary***

* Clustering: It is a data mining technique. The technique consists in apply some algorithms that will classify the observations (customers) into groups according to the similarity of their attributes.
* Normalization: The major algorithms of clustering need the data be scaled to a standard range. The process of applying some transformations in the data to have it in the same range is called normalization.

## Determine Data Mining goals

The data mining goals states project objectives in technical terms:

1. Segment customers according to their willingness to purchase wine and accessories, considering their demographic and social information (age, years of education, presence or absence of children, income, etc.), their 18 months’ records of commercial information (purchases, complaints, websites visits, etc.).

*Success criteria*: Visualize the cluster solution. Each cluster must have similar characteristics between the customers into it and distinct characteristics from the other clusters. We should be able to describe the characteristics best distinguish the clusters.

1. Evaluate the cluster quality.

*Success criteria*: The TNSE (visualization technique to visualize the distribution of the clusters) must have a good distribution of the clusters; Also compare the R2 metric of some cluster techniques applied and chose the one with the highest R2.

1. Suggest marketing strategies and business applications.

*Success criteria*: Present marketing strategies for each cluster and show business applications to help implement these strategies.

## Project Plan

Figure 2.2 - Project’s timeline.

Resources wise, for the business understanding we plan to use all the information provided in the kickoff meeting’s presentation. For the core stages of the project, we plan to use Python to work the data provided by WWW’s IT team. To present the results, we expect to use Word for the report, Prezi for the presentation and flask app system coded in Python to provide a user-friendly visualization of the results.

We consider the Modelling dependent of the Data preparation state as the quality of the clustering process will be directly connected with the quality of the input data. For this reason, we identify the Modelling stage dependent of the Data preparation stage. During the project, we must go and back between Data preparation and Modelling many times, repeat this iteratively until we get the desired outcome.

For the Modelling stage we aim to build an unsupervised model (clustering) using K-means algorithm. Due to the timescales we opted for using this algorithm as it is fast and efficient in terms of computational cost, simple to implement and the interpretation of clustering results is straightforward. The clustering quality evaluation will be made using R squared and some visualizations to check the good distribution between clusters.

# CLUSTERING ANALYSIS

In this section we go through the process of understanding and preparing the data for modelling, the modelling itself and different algorithms used and, finally, the results evaluation.

## Data understanding

At this stage we analysed the dataset to understand its potential and limitations. We got a first insight of the features, what they mean, how are they distributed, if there is noise, missing and/or duplicated values we should process, which of these features are relevant for the final goal and which features are redundant.

We have used the Pandas profiling to have an overview of the dataset: what variables are in the dataset, number of variables (29 features, from which 11 categorical and 18 numerical as shown on Table 3.1) and observations (10.000 customers), duplicated rows (none), missing values (none), correlation between variables.

We have also looked at the metadata file provided to understand the meaning of each feature to understand their relevancy in the project – in here we flagged feature LTV (Lifetime value of the customer) for posterior analysis since there is no clear information on how it was calculated and what does it exactly mean.

|  |  |
| --- | --- |
| **Numeric** | **Categorical** |
| *Dayswus*, *Age*, *Edu*, *Income*, *Freq*, *Recency*, *Monetary*, *LTV*, *Perdeal*, *Dryred*, *Sweetred*, *Drywh*, *Sweetwh*, *Dessert*, *Exotic*, *WebPurchase*, *WebVisit, Access* | *Custid, SMRack*, *LGRack*, *Humid*, *Spcork*, *Bucket*, *Complain*, *Mailfriend*, *Emailfriend*, *Kidhome*, *Teenhome* |

Table 3.1 - Numerical and categorical features.

## Data preparation

This is the stage when the input data for clustering is prepared, so we have ensured the data meets the requirements for this purpose [3]: Numerical variables only, data has no noise or outliers, data has symmetric distribution of variables, variables are on the same scale, there is no collinearity, few numbers of dimensions.

On the first step of data preparation, we set the variable *CustId* as index. The next step on the data preparation was feature engineering. We built a feature representing the average spent per purchase (*Avg\_ticket*) which was calculated dividing the feature Monetary by Freq.

As states on the first paragraph of this section (3.2), we have focuses on cleaning only the numeric variables sub-dataset, since those are the ones contributing to the model. A Pearson correlation matrix was prepared to look for these correlations. From the analysis of this matrix, we verified there were two groups of strongly correlated variables, as we can see in Figure 3.1:

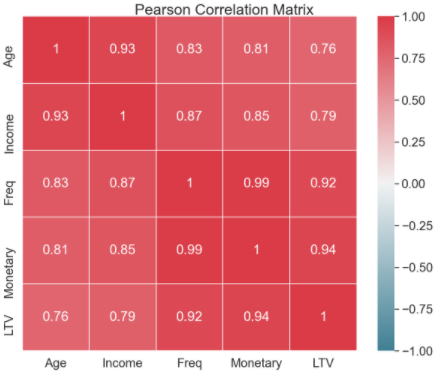


Figure 3.1 – Pearson correlation sub-matrices for numeric variables strongly correlated.

1. *Age*, *Income*, *Freq*, *Monetary* and *LTV* – to avoid redundancy and overfitting we dropped *Age*, *Income*, *Monetary* and *LTV* features as we have considered *Freq* to be more relevant for the analysis.
2. *WebPurchase* and *WebVisit* – for the same reason as stated in the previous point, we decided to drop *WebPurchase*.

After dropping these five features, we have checked once again the Person correlation matrix, which can be found in the notebook provided.

We have then looked for missing values, duplicated values and features which variance is lower than 10%:

* Missing and/or duplicated values but we concluded there were none in this dataset.
* Features which variance is lower than 10% and the outcome was the following eight variables: *Emailfriend*, *SMRack*, *Mailfriend*, *Complain*, *LGRack*, *Spcork*, *Humid* and *Bucket*. All these variables are categorical, so we did not drop them.

To check for the presence of outliers on the numeric variables we looked at the box and whiskers plots for each numeric feature and concluded that features *Sweetwh, Dessert, Sweetred, Freq* and *Drywh*, in Figure 3.2, seem to have outliers that need to be removed.

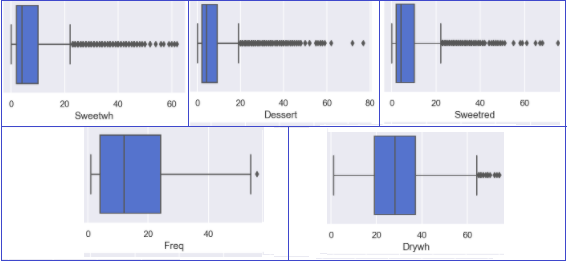


Figure 3.2 - Box and whiskers plot for features *Sweetwh*, *Dessert*, *Sweetred*, *Freq* and *Drywh*.

To remove these outliers, we tested some automatic approaches by applying six different methods for the entire dataset and using these methods individually or combined between them to remove observations. The number of observations removed with each of these approaches is summarized in Table 3.2.

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual methods** | | **Combined methods** | |
| Z-score | 1355 (14%) | 1 method | 1865 (19%) |
| Inter Quartile Range (IQR) | 2848 (28%) | 2 methods | 628 (6%) |
| Local Outlier Factor (LOF) | 634 (6%) | 3 methods | 506 (5%) |
| Isolation Forest | 1663 (17%) | 4 methods | 412 (4%) |
| Support Vector Machines (SVM) | 632 (6%) | 5 methods | 200 (2%) |
| Density based spatial clustering of applications with noise (DBSCAN) | 293 (3%) | 6 methods | 23 (0,23%) |

Table 3.2 - Number of outliers excluded with different approaches.

After testing removing based on 4 combined methods and realising that did not seem to make a big difference on the box and whiskers plot in terms of outliers, we have dropped this approach and decided to apply Inter quartile range (IQR) only on the five features mentioned above (*Sweetwh, Dessert, Sweetred, Freq* and *Drywh*). This way we have removed 354 observations, representing 3.5% of the dataset. The resulting, clean, box and whiskers plot for each feature can be found in the notebook.

In order to prepare the data for clustering, we have one-hot-encoded categorical variables (to transform in binary values) and applied the StandardScaler so that the dataset will have a mean value of 0 and a standard deviation of 1.

We have finished the data preparation with the following 14 numeric variables for clustering: *Dayswus, Edu, Freq, Recency, Perdeal, Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic, WebVisit, Access and Avg\_ticket*.

## Clustering

As previously stated in section 2.6 - Project Plan we have opted for using K-means clustering as it is the most reliable and efficient method. In this process we have created a set of functions that can be consulted in the notebook. From these we highlight the following outcomes from the functions:

* Inertia plot, average silhouette plot and Davies-Bouldin plot – showing the dispersion of the points within the cluster, how well each object lies within the cluster and a ratio between distances within the cluster and distances between clusters, for the different numbers of clusters to facilitate the decision on the optimum number of clusters for the present clustering challenge.

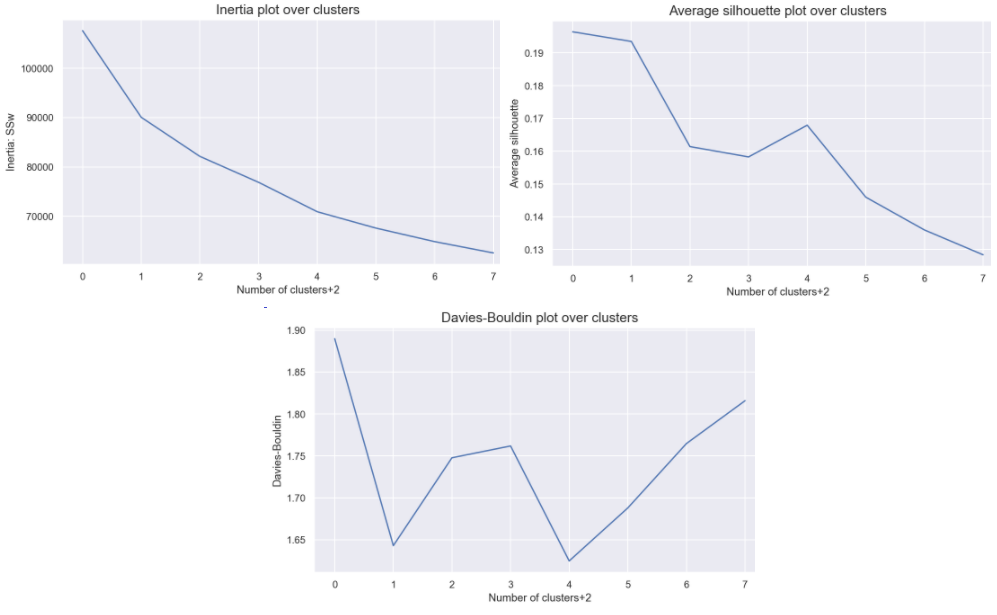


Figure 3.3 – Inertia, Average silhouette and David-Bouldin plots.

* Silhouette plot – shows a coefficient for the clusters’ quality depending on the number of clusters chosen.

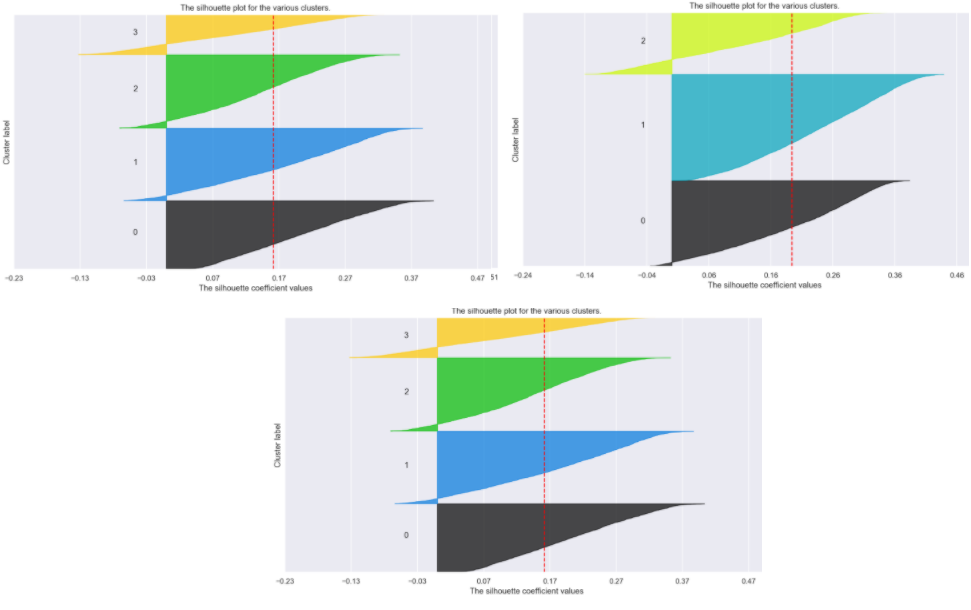


Figure 3.4 - Silhouette plot.

From the analysis of the plots above, we have chosen to proceed with 3 clusters and produced their profiles, showing the cluster’s means for each feature and the clusters’ absolute frequency for the three clusters, which enabled us to confirm the clusters are well balanced.

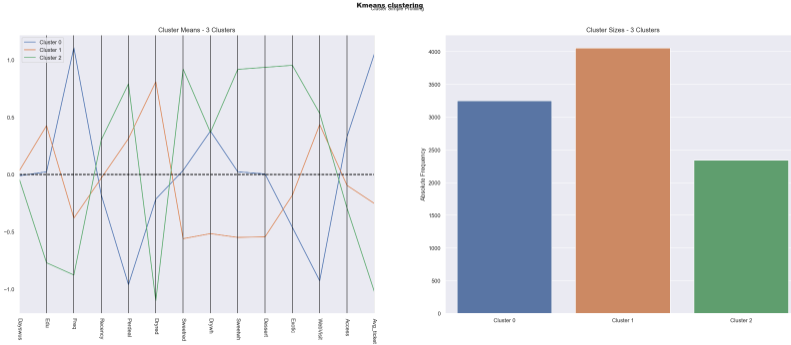


Figure 3.5 - Clusters profile.

It is worth to point that we also tested Hierarchical clustering whose outcome was reasonable, however, because K-means clustering showed a higher R2, we have opted for proceeding the analysis with the latter method.

## Evaluation

To evaluate the quality of our clusters, we have confirmed visually, in two dimensions, that the clusters are well defined (Figure 3.6, on the left) and the position of its centroids (Figure 3.6, on the right) through the T-Distributed Stochastic Neighbor Embedding (t-SNE) plots.

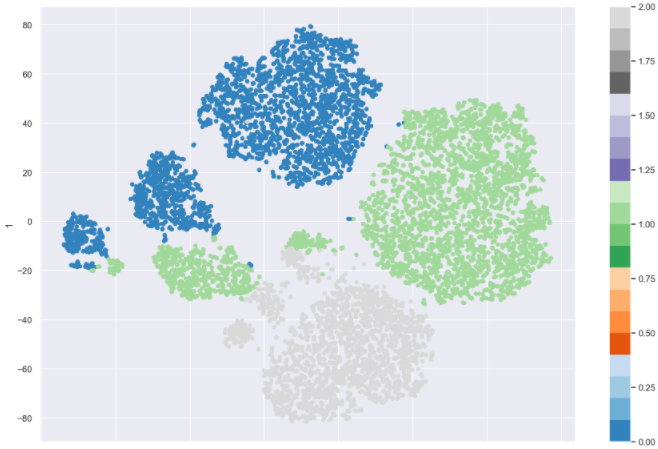
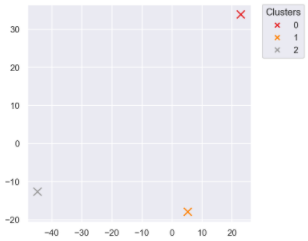
 

Figure 3.6 - t-sne plots.

In addition, we calculated the R2 metric for K-means that presents the result of 0,36, higher than hierarchical cluster that presents a result of 0,31.

We also applied a decision tree classifier, after splitting the data into  
training and test, to test our solution. It was able to predict 93,94% of the customer correctly  
We also got the feature importance of each variable in predicting the cluster. In terms of feature importance, *Avg\_ticket* and *Dryred* stand out as the most relevant features to the target. The full ranking can be found in the notebook provided.

# RESULTS EVALUATION

With this model we were able to identify three different segments of customers who have different characteristics and purchase behaviour.

Cluster 0 buys more frequently, has a higher income and tend to acquire more accessories when compared to the other groups however they do not purchase online or visit the website as often as the other groups. For this cluster we decided to create small events at the stores such as tasting days (fortnightly) and / or small workshops (monthly). To contact them, the best approach is via SMS or phone call, since they do not catch up the news about the products online, for those clients the contact should be closer and personal.

The idea with the events is stimulate the working capital and mix the products with worst selling wine (*exotic*/*sweetred*/*dessert*). Promotions and discounts were discarded in stores, since cluster 0 does not make many purchases on discounted items, leading us to a different approach exploring more the customer experience as wine lovers in addition to expanding knowledge about the product consumed since this is the cluster with the second largest education years.

This group prioritizes quality over price and is willing to pay more. Therefore, the probability of these events attracting interest is high. As a way of attracting new customers, the sale of tickets for the events would be processed in the following ways: a) the first customers to confirm would pay a lower price on the ticket or b) when inviting a second person, the ticket price would be 10% cheaper compared to individual ticket value.

Another factor highlighted is related to accessories, since this cluster have the most items. To stimulate the frequency beyond the days when events don’t occur, strategies such as buying a best seller (*drywh*) and worst seller (*exotic*) get an accessory would be adopted. Personalized accessories for the days of the events could also be made and distributed, such as glass of wine or cork stoppers.

Cluster 1 is characterized for the highest number of years of education and a high number of accesses to the website and online purchases. The strategy adopted for this cluster would be contacting the customers via e-mail as they use the internet as a way of communication. Promotional emails would also include information about the product consumed (e.g. curiosities, composition, benefits, ideal consumption rate, recipes, etc.), the production chain and distribution. In that way, customers would have knowledge of the company purpose in addition to its concern regard quality and commitment to the consumer.

It was also noted that cluster 1 buys products on discount, so considering that the *dryred* product has good sales frequency in this cluster, there is no need to promote it. The promotions are made with the second largest sales product (*drywh*) combined with the product with lower sales alternately, i.e. *drywh* + *sweetred*, *drywh* + *sweetwh* or *drywh* + *dessert*.

Finally, cluster 2 is the one that purchases less frequently, has the highest percentage of online purchases, the highest volume of website visits and it is also the youngest group, with the lowest income and money spent per purchase. Since this is the cluster with the highest percentage of purchases of products on discount, we propose two different approaches:

* Send sporadic promotional links to the client's account on the websites and app, there would also be gift vouchers to be given on holidays and also promotions for the products with the second largest sale, added to the one with the lowest sale.
* Give the customers the opportunity to indicate new customers. For each new consumer indicated, the current customer would accumulate points that to be converted into discounts on future purchases.

These strategies aim to increase the frequency on the websites and to attract new customers through referral.

For all clusters, three categories could be created to classify customers according to the value of accumulated purchases (regular, gold and premium customers).

For gold customers, a loyalty program could be offered so they would receive 2 types of wine monthly based on their preference, while a premium customer, in addition to having the same advantage as gold customers, would also have access to pre-sales and exclusive accessories. The transition from one category to another would occur as the value of purchases increased.

# DEPLOYMENT AND MAINTENANCE PLANS

The next steps would be:

* If it does not exist yet, create a customer account.
* Use the telephone (either call or SMS) as the main way of contacting customers on cluster 0 and create in-store events for these customers as well.
* Create promotions for cluster 1 when the customer buys the second largest sales product combined with the product with lower sales.
* Send sporadic promotional links for cluster 2 customers, to their client account and app and give gift vouchers on holidays and create promotions for the products with the second largest sale when sold together with the product with the lowest sale.
* Create a points program for cluster 2 where the customer indicating new customers would accumulate points that to be converted into discounts on future purchases.
* To create the three customers categories proposed in the section above (regular, gold and premium customers).
* To start the loyalty program proposed for gold and premium customers.
* Monitoring the model performance for the current and new customers.
* Start using digital marketing to reach new customers (e.g. Tweet Sentiment Visualization: https://www.csc2.ncsu.edu/faculty/healey/tweet\_viz/tweet\_app/).

We propose WWW prioritises customers that fall on cluster 0 since those are the ones that purchase more frequently and higher amounts.

App

# CONCLUSIONS

Final remarks on the project.

## Considerations for model improvement

The model has improvement margin if we get additional datasets to test its performance.

# REFERENCES

[1] Chapman, P, Clinton, J, Kerber, R., Khabaza, T., Reinartz, T, Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0*, CRISP-DM consortium

[3] Ryzhkov, E 2020, 5 Stages of Data Preprocessing for K-means clustering, viewed 27 February 2021, <<https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932>>

1. Model overfitting may happen as we are dealing only with 10.000 customers while the entire data has 350.000 customers (less than 3%). This problem will be identified when the WWW starts applying the solutions proposed to the other customers. If it happens, our consultancy is up to work to improve the modelling many times will be necessary. [↑](#footnote-ref-1)