

**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](#_Toc65365259)

[2. BUSINESS UNDERSTANDING 1](#_Toc65365260)

[2.1. Background 1](#_Toc65365261)

[2.2. Business Objectives 2](#_Toc65365262)

[2.3. Business Success criteria 2](#_Toc65365263)

[2.4. Situation assessment 2](#_Toc65365264)

[2.4.1. Terminology 2](#_Toc65365265)

[2.4.2. Risks and contingencies 2](#_Toc65365266)

[2.5. Determine Data Mining goals 3](#_Toc65365267)

[2.6. Project Plan 3](#_Toc65365268)

[3. CLUSTERING ANALYSIS 3](#_Toc65365269)

[3.1. Data understanding 3](#_Toc65365270)

[3.2. Data preparation 4](#_Toc65365271)

[3.3. Modeling 6](#_Toc65365272)

[3.4. Evaluation 6](#_Toc65365273)

[4. RESULTS EVALUATION 6](#_Toc65365274)

[5. DEPLOYMENT AND MAINTENANCE PLANS 6](#_Toc65365275)

[6. CONCLUSIONS 6](#_Toc65365276)

[6.1. Considerations for model improvement 6](#_Toc65365277)

[7. REFERENCES 7](#_Toc65365278)

[8. APPENDIX (OPTIONAL) 8](#_Toc65365279)

# 

# INTRODUCTION

Thank you for choosing Data4Business Consulting for help you in the challenge of understand better your customers characteristics and segmentation. Our main objective is helping your business to improve the set-up of the business, 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.

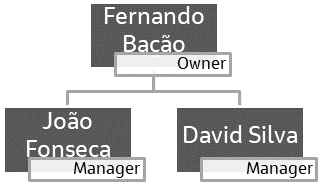


Figure 2.1 - Organizational chart.

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 Data4Business 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 while 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 with the ingredients, flavors, 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

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).

## 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, with the following libraries: *pandas*, *numpy*, *seaborn*, *scipy*, *sompy*, *matplotlib*, *itertools*, *math*, *sklearn*. 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 risk of, after the Modelling stage, to have to go back to the Data preparation and 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 (aiming for a value close to 1).

# 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), missing cells and 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 K-means is prepared, so we have ensured the data meets the requirements for this purpose [2]: 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.

As states above, because in the project plan we have decided to use K-means algorithm for clustering and, 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:

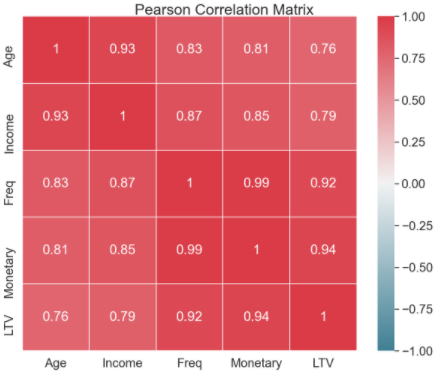


Figure 3.1 – Pearson correlation sub-matrices for numeric variables strongly correlated.

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 new Person correlation matrix, which can be found in the notebook provided.

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*.

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.

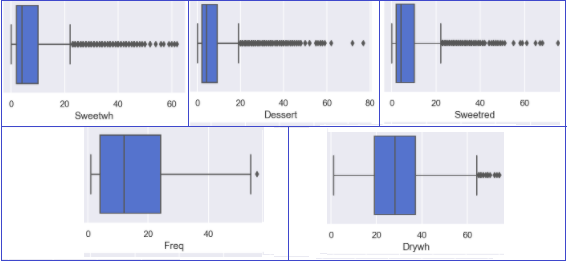


Figure 3.2 - Box and whiskers plot for features *Sweetwh*, *Dessert*, *Sweetred*, *Freq* and *Drywh*.

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 summarised 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 seemed 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.

To avoid confusing the model, we have one hot encoded the categorical variables and then we have used StandardScaler to normalize the dataset. Mas porque se so utilizamos as numricas?

We have finished the data preparation with the following 15 variables: *Cusid*, *Dayswus*, *Edu*, *Freq*, *Recency*, *Perdeal*, *Dryred*, *Sweetred*, *Drywh*, *Sweetwh*, *Dessert*, *Exotic*, *WebVisit*, *Access* and *Avg\_ticket*.

We then looked at the numeric variables distribution to ensure

What did we conclude from the pairwise relationship between variables visualization? + Not to include following graph

Aiming to use K-means we needed to ensure the numeric features follow a normal distribution. For this reason we have used histograms to verify its distribution and skewness and concluded that…..

Feature engineering (Value per purchase)

RobustScaler is based on percentiles and therefore not influenced by a few numbers of very large marginal outliers.[14].

## Modeling

Bla, bla.

## Evaluation

Results described in technical terms (e.g., reached an Accuracy of 95%).

# RESULTS EVALUATION

Describe the degree to which the model meets the business objectives. If that cannot be done without the application of the model in a real environment, describe how could that be done.

Assess the data mining results in respect to the business success criteria.

# DEPLOYMENT AND MAINTENANCE PLANS

Describe how the strategy to deploy the model into production (necessary steps, persons involved, systems that may require changes, etc.).

State how after deployment the model’s performance should be monitored and maintained.

# CONCLUSIONS

Final remarks on the project.

## Considerations for model improvement

Bla, Bla

# REFERENCES

Author, A. A., Author, B. B., & Author, C. C. (Year). Title of article. *Title of Periodical, volume number* (issue number), pages.

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

[2] https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

# APPENDIX (OPTIONAL)

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)