BUSINESS CASES WITH DATA SCIENCE

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

Business Case 5 – Demand Forecasting

Group W

Ana Luísa Mestre, number: 20200599

Beatriz Pereira, number: 20200674

Mariana Domingues, number: 20201040

Nadine Aldesouky, number: 20200568

June
2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

INDEX

1. INTRODUCTION

Mind Over Data (MOD) is an Australian company founded in 2016 which focuses on technology services such as machine learning, business intelligence and data management Consulting. It offers global services but is currently housed in 2 offices in Lisbon, Portugal and Sydney, Australia. The company tackles different industries including retail, transport, Insurance, Healthcare, among others. Fortunately, MOD, is offering a challenge for Nova IMS' data scientists to provide a business intelligence report for their PoS Appliance's Retail project. Our team of data scientists has taken on this challenge to prove to MOD their substantial skillset and help provide them with a very insightful report.

2. BUSINESS UNDERSTANDING

2.1. Determine Business Objectives

- **Background**: Mind Over Data is an Australian data and analytics consulting company. The company's services are split in 3 groups: advisory, analytics, cloud services and delivery. The company operates across numerous industries including retail, insurance, automotive, transport, telecommunications, transport, healthcare, higher education and lotteries. Currently, MOD resides in two offices in Lisbon and Sydney, but its projects extend across the globe. Since the company is already an expert in the field of data, it is hosting a challenge at Nova IMS to help students grow and learn by offering them real-life datasets and business problems to solve and work with. As a result, Vasco Jesus, the Head of Analytics at MOD, has challenged the Nova IMS' data scientists to provide him with a full analysis of their PoS Appliances Retail project.
- **Business Objective/Problem**: To complete a full analysis on PoS Appliance's Retail by deriving insights from the data provided.
- **Business Success Criteria**: A successful analysis will result in insightful information such as PoS segmentation and demand forecasting that can be used by MOD to improve its services.

2.2. Assess the Situation

- **Inventory of Resources**: The company provided a database in a CSV file which will be analyzed by a team of four business analysts known as Group W. To manage the problem, Python, Microsoft Word, Power BI and PowerPoint will be used.
- **Requirements, Assumptions, and Concerns**: This report is for managers' use which means that it is for private use at MOD only. For this project, we will be working with the *NOVAIMS_MAA_2020e21_BusinessCasesDataScience_MindOverData_RetailChallenge.csv* dataset composed of a sample of 182 342 304 records and 9 variables.
- **Risks and Constraints**: The dataset contains all the transactions occurring between 01/01/2016 and 01/11/2019. Additionally, this report must be completed and presented to management within three weeks' time.

Risks	Contingencies
Useless features for the task	Ask for different variables or derive new features
Missing relevant information	Ask for more details about each variable, and additional variables
Lack of funding or available time	Request an extended deadline
Losing the data	Create a copy of all aspects of the project

Lack of computing power	Request better machinery or laptops

Table 2.1.– Risks and Contingencies

- **Terminology**:

SKU: Stock Keeping Unit, a scannable barcode printed on the labels of products in retail stores, this code helps sellers to keep up with the movement of their inventory.
PoS: Point of Sale, the location at which a retail transaction takes place

- **Costs and Benefits**:

As this is an academic project, we have no information about the costs of the data collection neither the ones regarding the development and the implementation of our solution for the business problem.
The benefits include increased better understanding of sales across the PoSs to increase sales, improve customer satisfaction, improve supplier and partner relationships, decrease costs, and increase revenues in the future.

2.3. Determine Data Mining Goals

- **Data Mining Goals**: To complete a quarterly analysis of the top products, a market basket analysis to understand product co-occurrences, a PoS clustering, and a product forecasting 6 weeks ahead.
- **Data Mining Success Criteria**: A successful analysis will result in acceptable performance metrics per model and produce very insightful and actionable information regarding the performance of different products and PoSs.

2.4. Produce Project Plan

- **Steps**

1. General exploratory analysis, identifying clear problems in the data such as missing values, outliers, or incoherencies.
2. Deeper exploratory analysis.
3. Describe the insights obtained from the data exploration - explore; produce and interpret visualizations to understand the absolute frequencies of the variables, or the presence of outliers.
4. Complete the quarterly analysis of the top 5 products sold and the top 5 product families and categories with regards to market share.
5. Prepare the data to the model (market basket, clustering, time series forecasting) input – select the required features, apply features transformations, create new features, detect outliers, and/or correcting incoherencies.
6. Review the data again, after cleaning it (e.g.: see if the insights are the same or if they changed slightly).
7. Define and apply the clustering as well as market-basket-analysis technique to create unsupervised models.
8. Split the data into train and test, then apply the forecasting technique to create a supervised predictive model.
9. Fine Tuning (making small adjustments to parameters to achieve the desired output of performance) of the models, and selection of the one that provides the best solution.
10. Assess and interpret the results from the chosen models.
11. Provide a deployment plan.

Initial Assessment of Tools and Techniques

Technique	Pros	Cons

Kmeans[1]	Good for metric features Good for spherical-shaped clusters. Efficient and fast implementation.	Sensitive to outliers and initialization method Need to set # of clusters beforehand.
Hierarchical Clustering[2]	Easy to implement.	Difficult to decide the appropriate number of clusters from the dendrogram
***Apriori* Algorithm**[3]	-Simple and easy to implement. -Works well with large datasets (includes pruning steps).	- Computationally expensive based on the thresholds and diversity of items. - Very complex in terms of time and space. - Performance is reduced by the numerous times the algorithm scans the dataset.
SARIMA[4]	Suitable for univariate time series with seasonal components.	Does not consider exogenous variables
SARIMAX[4]	Adds exogenous variables to SARIMA. These variables can play a significant role in accounting for extraordinary variations.	Costly computation
Prophet[5]	It is robust to missing data and shifts in the trend, and typically handles outliers well.	Works better with time series that have strong seasonal effects and several seasons of historical data.

Table 2.2.– Tools and Techniques

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data Understanding

The dataset was provided to us in the form of a csv file, which contains 9 variables and around 19000000 observations. The data was taken from the period between 1st January 2016 and 1st November 2019. Inside the dataset, there is a set of 5 variables regarding the hierarchy of products – ProductFamily_ID, ProductCategory_ID, ProductBrand_ID, ProductName_ID, ProductPakSKU_ID, being the last variable the most detailed one. Following this, there are 21 families of products, 178 categories, 1535 brands, and 8660 SKU's. The remaining 4 variables – Point-of-Sale_ID, Date, Measures, and Value - are regarding the different points of sales (410), the date in which the products were sold and the number of units and total price of those products.

When exploring the dataset, we expected to have pairs of rows that would be equal for the same product SKU, point of sale, and date, meaning that a certain product was bought in a certain date and point of sale with an X quantity and Y price in total. However, we noticed that this not always happens - for instance, there are 10682254 rows grouped by a set of 4 duplicated values for these variables instead of 2.

Additionally, the prices for the same product may vary from point of view and throughout the time. Lastly, there are 13584 duplicated observations and no missing values.

Initial exploration of the dataset showed that 2018 was the best performing year with a revenue of 84677 million. Additionally, at first sight it seems like 2019 is the worst performing year since it has the lowest revenue (73456 million). However, this is most likely because we only have data until Nov 1st of 2019 which means we are missing 2 months from its last quarter. When comparing to the performance of previous years, it looks like 2019 is performing quite well and on target. To clarify, in general the company generates around 20 to 22 million in Q4. In the first month of 2019 Q4, the company has already generated 8093 million so most probably in the next 2 months it would exceed last year's revenue. As a result, the lowest performing year is actually 2016. This is because it generated a total of 75956 million while the ten months of 2019 generated a total of 73456 million.

Moreover, we noticed the following insights:

- The top performing Quarters in terms of total generated revenue over the last four years (since 2016) across all PoS are:
  - **2018 Q3 with 22.627 million** (see the appendix, Figure 6.1)
  - 2017 Q4 with 22.158 million
  - 2019 Q1 with 22.067 million
- The top performing Point of Sale IDs in terms of total units sold over the last four years (since 2016) across all PoS are:
  - **2019 Q1 with 13679k units**
  - 2018 Q4 with 13559k units
  - 2018 Q1 with 13243k units

3.1.1 Quarterly Analysis by each Point-of-Sale

After exploring the dataset, we were able to decipher the following insights:

- The top performing Point of Sale IDs in terms of total generated revenue over the last four years (since 2016) are:
  - **PoS ID 282 with 1552 million**
  - PoS ID 78 with 1363 million
  - PoS ID 283 with 1287 million
- The top performing Point of Sale IDs in terms of total units sold over the last four years (since 2016) are:
  - **PoS ID 282 with 904k units**
  - PoS ID 78 with 886k units
  - PoS ID 280 with 861k units

3.1.1.1 Top products sold

When diving deeper into this insight, we were able to complete a quarterly analysis in which we identified the highest performing quarter of the top PoS ID 282. This was 2018 Q3 both in terms of generated revenue (117 million) and units sold (63k units). Additionally, you can see the top selling products (in terms of units sold) during this quarter at this PoS, in the appendix figure 6.2.

3.1.1.2 Market share by Family and Category

Moreover, the tree maps in the appendix (Figure 6.3) indicate the top 5 product families and categories (in terms of market share) during this quarter at this PoS.

You can dive deeper into this PoS quarterly analysis which focuses on the product using our dashboard with its interactive filters in the following link:

https://app.powerbi.com/view?r=eyJrIjoiZDk2NWY3ZDItOWVjOS00ZGYwLWE4NzMtNTQzYzE0NzNiNjFjIiwidCI6ImU0YmQ2OWZmLWU2ZjctNGMyZS1iMjQ3LTQxYjU0YmEyNDkwZSIsImMiOjh9&pageName=ReportSection891b6da36ae0ecabe78e.

3.2. Data Preparation

To effectively meet the challenge, we needed to do a general preprocessing before applying the required analysis. Since we are talking about a big data the first step was to change some variables type to occupy less space in order to simplify the data and be able to work with it in a more efficient way. Most importantly, we converted the 5 product hierarchy variables into integers, taking out their definition from their values; and transformed the variable Measures into a binary one, where the value 0 represents the sell-out units and 1 the sell-out values. Additionally, while analyzing the data we noticed that there was an observation with a negative value that was deleted from the data.

Since the manager required a product name level analysis, the next step was to group the data set by product name, point of sale, date, and measures, excluding the product SKU from the data. In order to have the information reg
arding the number of units and the total price of the same sale in a single row, we created two variables from the variable Measures – Unit and Price.

In the end, the data was left with 75980067 rows and 9 columns. Only losing the product SKU variable, we could pass from a file with 18.6 GB to 2.95 GB with this simple initial treatment.

In the next sub-sections, we will present some data cleaning, do several feature engineering and prepare the data to the input of the different models being applied.

3.2.1 Data Treatment for Clustering

The data frame used in the clustering by value contains 410 observations and a set of 8 variables that were created:

- total_units, which is the total number of units sold for each PoS
- total_price, which represents the total revenue with the units sold for each PoS
- avg_units, which is the value in average of units sold for each PoS
- avg_price, which is the value in average of revenue for each PoS
- std_units, which is the standard deviation of the units for each PoS
- std_price, which is the standard deviation of the revenue for each PoS
- nr_days, which represents the total number of days with sales for each PoS
- difference, which is the difference, in days, between the most recent and the oldest dates

On the other hand, the data frame used in the clustering by preference contains 410 observations and a set of 21 variables which correspond to the 21 different families of products that exist in the dataset, meaning that each row corresponds to a point of sale and each column contains the total number of units sold for that product family in the 410 points of sales.

Regarding the outliers, we decided not to drop them since we considered them to be important for the analysis.

3.2.2 Data Treatment to Forecast

Given that we want to perform a forecast 6 weeks ahead we started to exclude the products that don't have sales for more than 6 months, i.e., with no sales after May of 2019. From an initial number of 2820, we were left with 2188 products to forecast.

To perform the weekly forecast on the variable Unit we needed to resample the data to a week frequency, summing the values of the days by week. According to the dimension of the history of each product a different predictive analysis is presented in the next section.

3.3. MODELING

3.3.1 Product Co-ocorrences – Apriori Algorithm

To be able to understand what were the co-occurrences that happened in the clients' stores, we used the *Apriori* algorithm. This algorithm assumes that all subsets of the frequent item sets must also be frequent while all supersets of the infrequent items will also be infrequent and so excluded from the generation of association rules. We created the frequent itemset for each quarter supported by at least 60% of transactions; this will help us understand what the most required products are too. When we say that a product, or even a subset of products, needs to have, at least, 60% of support, this means that this product or subset of products needs to appear in 60% of our transactions.

We considered transaction the combination of Date + Point-of-Sale_ID, so, our "baskets" were basically the purchases made each date and at each point of sales; and we analyse them for each quarter in each year.

Since with this algorithm we want to know which are the co-occurrences, we do not need to have a lift less than 1, which means that the products are substitutes, and that if we buy one the probability of buying the other is lower. That is why we considered a threshold of 1.1 for the lift, that is, only combinations of products with a lift greater than 1.1 would be considered. With this we were able to build 2 data frames with the association rules we wanted: one for the co-occurrences between product families and the other for the co-occurrences of the product category.

In both of the data frames we can see that the confidence and support of each rule are high so, we can trust in these rules.

3.3.2 Point-of-Sales Clustering - by Value and by Preference

In order to cluster the 410 points of sale by value and preferences perspectives, we started by normalizing the two different data frames with MinMax scaler, since the algorithms used require it. Then, we applied two different clustering techniques for each perspective of analysis: one with the KMeans, and the other with the KMeans (with a big number of clusters, 60) followed by the Hierarchical Clustering, where, from the dendrogram, we selected a final number of clusters, much smaller than the previous one.

To select the best approach, we changed the parameters regarding the number of clusters in each clustering method to analyse the different solutions that they gave us in terms of distributions, metrics, and profiles of the different clusters. In the following table are presented the variables and models used for each perspective, as well as the values obtained for each evaluation metric.

	KMeans	**KMeans+HC**
Preference	**R2**= 0.482 **Calinski**= 125.881 **Davies**= 1.687	**R2**=0.211 **Calinsk**i=36.140 **Davies**=2.742
Features used	Product_FamilyID (1-21)	Product_FamilyID (1-21)
Clusters	4	4
Value	**R2**=0.484 **Calinski**=190.937 **Davies**=1.110	**R2**=0.145 **Calinski**=34.414 **Davies**=2.897
Features used	total_units, total_price, avg_units, avg_price, std_units, std_price,	total_units, total_price, avg_units, avg_price, std_units, std_price,

	last_date, nr_days, difference	last_date, nr_days, difference
Clusters	3	3

Table 3.1.– Clustering Results

Considering the values and the criteria mentioned above, we chose to use the KMeans for both the preference and the value perspectives with 4 and 3 clusters, respectively.

In the preferences' perspective, we also try a different approach by applying the KPrototypes algorithm. To do it, we used a data frame grouped by point of sale and by product family, corresponding to 7250 rows and 6 columns. Here, we started by creating 2 new variables – total_units, total_prod_fam – so that we could have, for each product family of each point of sale, the total of units sold and the number of times each product family appears. To be able to apply the KPrototypes, we created two binary variables from the previous ones – total_units_bin, total_prod_fam_bin – that, from a certain threshold, would assume the value 1 if the considered product family would be a preference and 0 if not. To do this, we had to delete the outliers of these variables so that we could have more homogeneous distributions. In the end, since the clusters with this approach were not so good, we discarded this option.

3.3.3 Sales Forecast

After considering some forecasting models we decided to apply Prophet, that is an open-source software released by Facebook's Core Data Science team. It stands out because of its reasonable forecast on messy data with no manual effort and its performance getting forecasts in just a few seconds, using Stan[6] to fit models. They use a decomposable time series model with three main model components: trend, seasonality, and holidays.[5]

For this data, we use the default linear model from Prophet with the automatic changepoint detection, given the possibility to change the trend over time. The frequency of these changes can be adjusted with the argument *changepoint_prior_scale* (0 to 1) to control a possible overfitting (too much flexibility) or underfitting (not enough flexibility). For the seasonality component, estimated using a partial Fourier sum, we tried to fit monthly, quarterly, and yearly seasonality. The Fourier order used for each seasonality is the one recommended by Prophet documentation. For the last component, the holidays, it was left empty giving lack of information about this topic in the data. Since we did not have the PoS regions we could not try to implement the holidays factor in our model. Other recurring events, like promotion days/weeks, were also not given to us.

Given the variation of the history that we have by each product we needed to adjust the model fitness by some thresholds. There are products with only 2 weeks of history, or with a big proportion of zeros (no sales) in middle weeks.

As mentioned in[7], there is no justification for the magic number of 30 observations often given as a minimum for ARIMA modelling. The only theoretical limit is that we need more observations than there are parameters in our forecasting model. Therefore, given the parameters needed to Prophet, we deleted all products with less than 5 weeks of history.

To apply the grid search of the parameters mentioned above, the following thresholds were applied:

- Products with less than 3 months:
    - Trend flexibility (*changepoint_prior_scale*) = 0.05 (default)
    - Seasonality - None
- Products with less than 6 months, a proportion of zeros greater than 70% or a history with less than 3 months without zeros (no sales):
    - Trend flexibility (*changepoint_prior_scale*) = [0.05, 0.1]
    - Seasonality – None or monthly
- Otherwise:

- Trend flexibility (*changepoint_prior_scale*) = [0.05, 0.1, 0.2]
- Seasonality – [None, monthly, quarterly, monthly and quarterly] + default yearly

3.4 EVALUATION

3.4.1 Product Co-ocorrences – Apriori Algorithm

As stated earlier, in the data frames developed for co-occurrences we can see that all confidence levels are higher than 0.6 and that support also checks the previous condition. With this, we can conclude that the rules found by us, are credible.

We can also conclude that the co-occurrences that happened more often and that are more likely to happen again are the combinations whose lift is higher. However, according to the data provided and treated by us, the lift with the highest value does not exceed 1.16.

3.4.2 Point-of-Sales Clustering

Regarding the evaluation of the different clustering algorithms and consequently selection of the best one for each perspective, as already mentioned above, we used three different metrics: R2, Calinski Harabasz, and Davies Bouldin scores. Additionally, we plotted the clusters so that we could better visualize the distributions of the points of sale through them, as well as their main differences in terms of variables used. In the end, we got the plots in the appendix (see figures 6.4 and 6.5):

By Preferences

Cluster 1: This is the largest cluster, containing 160 points of sale, however, it is the one where the preferences are not so clear. Here, we can say that the less preferred categories are numbers 20 and 11, and the most preferred are 2 and 18, for example.

Cluster 2: This is the smallest cluster which is composed by 57 points of sale. It represents the group which buys the highest number of products from the families number 4, and from 7 to 20. Regarding the families less frequent, these are mainly families number 5 and 6.

Cluster 3: This cluster comprises 120 points of sale. In general, this is the group which sells the smallest number of units for most of the products' families, for example families number 5 and 19. On the other hand, these points of sale prefer the families 2, 8, and 18.

Cluster 4: This cluster is composed by a total of 73 points of sale. They are the ones which sell the highest number of units from the products' families 1, 2 6, 9, and 12. Additionally, this is the group which sells less units of the families 11 and 20.

By Value

Cluster 1: This cluster comprises 125 points of sale. It is the one with the highest total number of units sold, leading to the highest revenue (total_price) and, consequently, the highest average number of units and revenue generated. Also, in this group is where the price and the units vary the most. Finally, it is also the cluster with the bigger number of days with sales.

Cluster 2: This is the smallest cluster, composed by 70 points of sale. The points of sale belonging to this group have the lowest total number of units sold and revenue. They also represent the group with the lowest number of days with sales. However, they occupy the intermediate position in terms of average and standard deviation of units and price.

Cluster 3: This cluster represents the largest group, containing 215 points of sale. This group has the shops with the lowest average number of units sold and revenue amount, and it is also the one where the number of units and the revenue vary the less. On the other hand, it occupies the intermediate position in terms of total number of units sold, total revenue and total number of days with sales.

In terms of the difference between the most recent date and the oldest one, it is not possible to differentiate the clusters because they have an equal behavior in this case.

3.4.3 Sales Forecast

To evaluate the model performance, we used the Prophet diagnostics that provides a time series cross validation. This is done by selecting cutoff points in the history, and for each of them fitting the model using data only up to that cutoff point. In our case, we initialized it with 60% as training data and made folds with a length equal to 20% of the training data. The metric used to evaluate each fold predictions was WAPE (weighted absolute percentage error), that weights the forecast error adding sales volumes to the equation. In this article [8], it is mentioned that MAPE calculates a fake prediction measure when predicting a product with a very low volume of sales. This measure can also fail when calculating a week with no sales given that its denominator will be zero. Additionally, this author considers WAPE the best method to measure the forecast accuracy with this type of data.

To define the final forecast accuracy, in order to give more weight to the most recent folds, a weighted mean of the folds' WAPE is calculated.

Unfortunately, given the restricted time to develop this project and computational power we couldn't predict the sales for all products and PoS. However, from the PowerBI dashboard we can have an idea of the accuracy of this procedure. Of course, that, given the history variation from product to product, this accuracy will fluctuate a lot.

In the appendix, Figure 6.6, we present an example of the forecast of the product 1147 sales in the PoS 49. In the graphic, the confidence interval is represented by the dashed line, the predictions in the gray line and the real sales the points in red.

Despite MAPE and WAPE being most commonly used, these measures have one important drawback. They calculate the percentage of error but don't distinguish between overstocks and out-of-stocks, because they are symmetric.

However, from the retailer's perspective, overstocks and out-of-stocks affect business differently. In other words, the cost of forecasting error equal to 10 wholesale packages of milk is not the same for stockouts and overstocks. While the former are potential losses that include loyalty losses and a decrease in store traffic, the latter consists of storage costs, cost of capital, and write-offs.

These statistical metrics should be aligned with the business ones. Depending on the retailer's priorities, these additional metrics may include the number of write-offs, costs of markdowns, the number of out-of-stocks, lost sales, or food wastage. The combination of statistical and business metrics gives an opportunity to make better decisions on the stock management. After all, it is not the percentage of error but the cost of error that makes a difference.[9]

3.4.4 Review Process

To improve and detailed our analysis and results, we think that we could have tried more clustering algorithms and maybe apply outliers' analysis. However, the dimension of the data and the discovery of the better way to deal with it was a time barrier in the implementation of this project, which lead us to the optimization of our decisions in order to reach all the objectives required.

3.4.5 Determine Next Steps

- Move to the deployment step.
- Improve the data preparation step, such as the outliers' analysis, for example.
- Improve the clustering by PoS – tunning the parameters or trying different clustering algorithms, such as SOM, Mean Shift, or DBSCAN, for example.
- In the clustering by preference perspective, build the clusters with the total units sold by product category, increasing the level of detail of the analysis.
- Improve the association rules analysis by trying different values for the support, lift and confidence metrics.

- Add exogenous variables to the forecast, like promotion dates or holidays.
- Extend the forecast analysis: making a bigger grid search in the Prophet Model, try more advanced forecasting methods like Neural Networks or bootstrapping time series [10].
- Work with big data to be able to process this huge data faster. Using other tools like Py spark, work with non-relational databases, using multi-threading in Python, etc.

4. DEPLOYMENT

4.1. Deployment Plan - Recommendations and Challenges

After analyzing the different stores of the client and, in each quarter, the co-occurrences, having added the point of sales in different groups, as well as the forecast for the intended 6 weeks, the next step is to arrange the necessary stock to the weeks ahead.

We advise that this forecast is monitored and that the values which were verified are reported to us, so that we can improve it and help the business to grow even more.

Additionally, the company can and should use the dashboard designed to better understand its customers and the needs, as well as the needs of each of its points of sale.

The dashboard is a very useful tool that can be updated from time to time according to the needs of the company Mind Over Data, however, this continuous monitoring should be done by us, and therefore we should perhaps have an element of our team that weekly access the data and make a check of it and what needs to be solved or not.

We also advise you to also pay attention to the products' co-occurrences and forecast. This is because if we know that two products tend to be bought together, then, in the future, when they expect to have in stock a certain quantity of a product, also have the other product with which the first is usually bought. However, you should pay attention to whether these co-occurrences happen or happened because some type of promotion was developed that encouraged the purchase of the products together

4.2. Plan Monitoring and Maintenance

- If current conditions do not change, the analysis of the clusters must be reviewed within one year's time at most to ensure its viability and relevance.
- Review the co-occurrences when a certain number of products becomes available at Mind Over Data. And evaluate them, after a certain number of orders (to be determined by the company), to find out if there are new rules for the antecedents vs. consequents products or even if those previously considered are still valid. And if the current conditions do not change, the analysis of the co-occurrences must be reviewed twice a year at most to ensure its viability and relevance
- Regarding to the forecast, we will have an alert that informs us when the predictions foreseen by us err a certain number of times for a reasonable amount of error; these parameters must be discussed with the Mind Over Data manager
- Have an element of our team that weekly access the data and make a check of it and what needs to be solved or not

4.3. Review Project

During this project, we felt some difficulties mainly related with the size of the dataset, the set of objectives required, and the time that we had to achieve them. Regarding the size of the dataset, we lost much time figuring out the best way to load all the dataset to Python so that we could start doing the project. Additionally, our computers were not prepared for working with these huge datasets, which lead to frequent break outs and wastes of time during the development of our work. Besides this, the required outputs were a lot (almost like a compilation of 4 different projects in only 1), so it would have been helpful if we had additional time to run more code and improve our analysis, or if we had less required outputs, in order to focus more on the forecasting analysis, for example, since it was the new one.

5. REFERENCES

[1] "Data Clustering Algorithms - k-means clustering algorithm." https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm?authuser=0 (accessed Jun. 01, 2021).

[2] "Data Clustering Algorithms - Hierarchical clustering algorithm." https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm (accessed Jun. 01, 2021).

[3] G. L. Team, "What is Apriori Algorithm in Data Mining Implementation and Examples?," 2020. https://www.mygreatlearning.com/blog/apriori-algorithm-explained/.

[4] "11 Classical Time Series Forecasting Methods in Python (Cheat Sheet)." https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/ (accessed Jun. 01, 2021).

[5] "Prophet | Forecasting at scale." https://facebook.github.io/prophet/ (accessed Jun. 01, 2021).

[6] "Stan - Stan." https://mc-stan.org/ (accessed Jun. 01, 2021).

[7] "12.7 Very long and very short time series | Forecasting: Principles and Practice (2nd ed)." https://otexts.com/fpp2/long-short-ts.html (accessed Jun. 01, 2021).

[8] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons," *Int. J. Forecast.*, vol. 8, no. 1, pp. 69–80, Jun. 1992, doi: 10.1016/0169-2070(92)90008-W.

[9] "Measuring forecast accuracy. Is most accurate forecast always the best?" https://www.dslab.ai/measuring-forecast-accuracy (accessed Jun. 01, 2021).

[10] "Chapter 11 Advanced forecasting methods | Forecasting: Principles and Practice (2nd ed)." https://otexts.com/fpp2/advanced.html (accessed Jun. 01, 2021).

6. APPENDIX

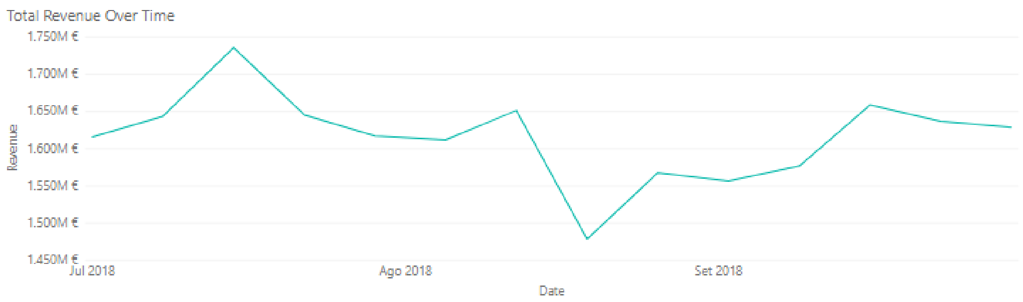


Figure 6.1 - Total Generated Revenue in 2018 Q3

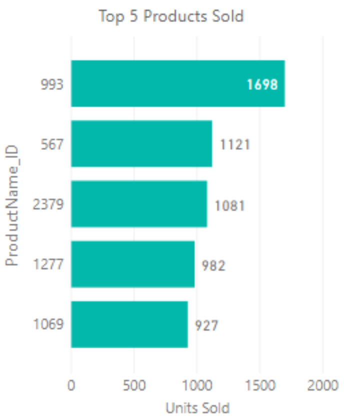


Figure 6.2 - Top 5 Products Sold in PoS ID 282 in 2018 Q3

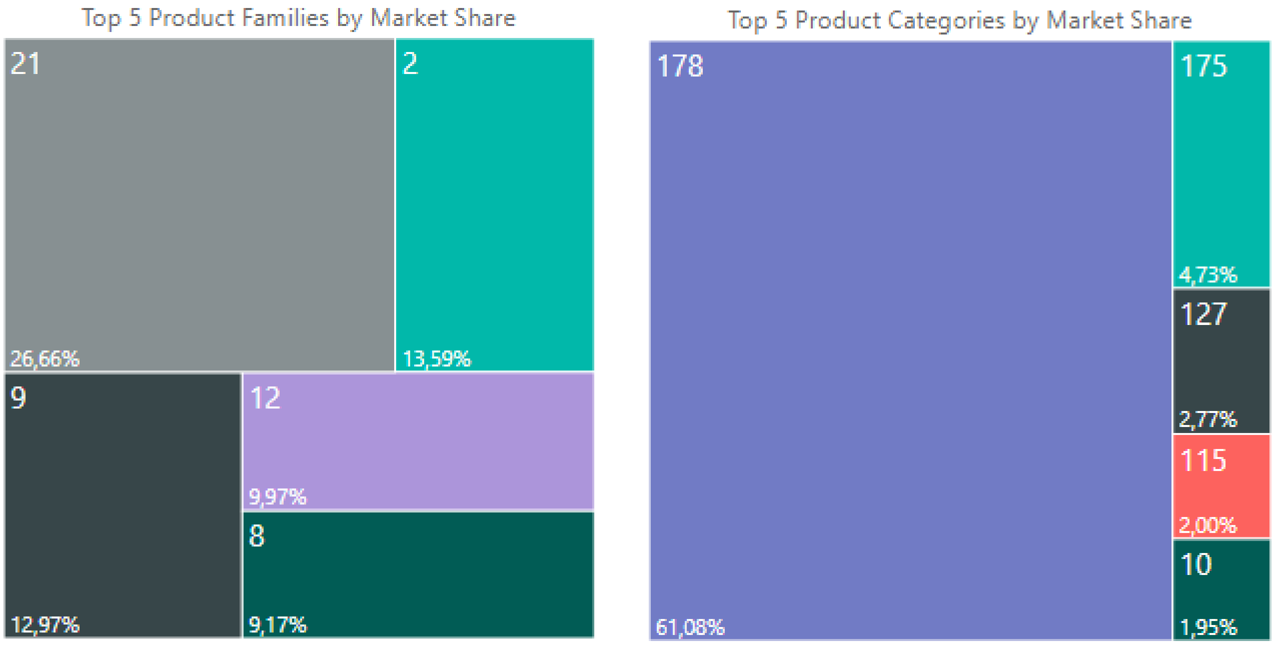


Figure 6.3 - Top 5 Product Families and Categories (in terms of market share) in PoS ID 282 in 2018 Q3

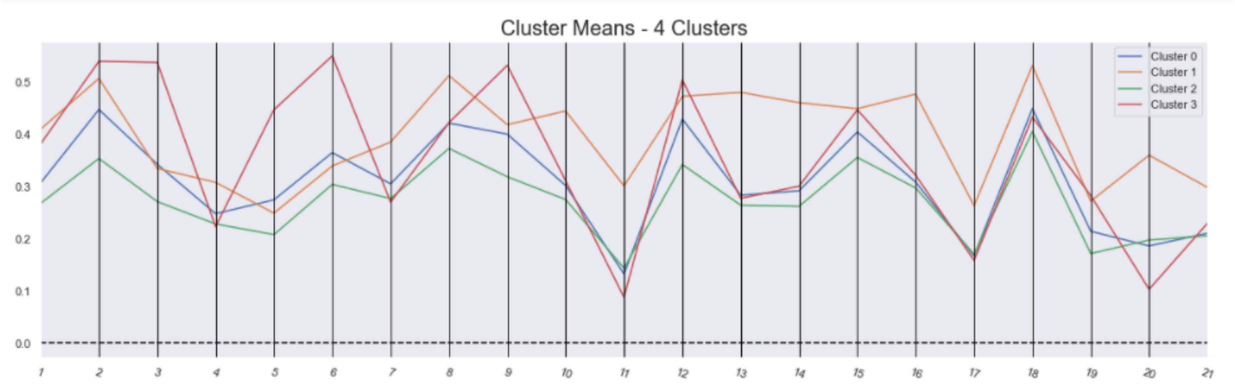


Figure 6.4 – Clusters by Family Preference

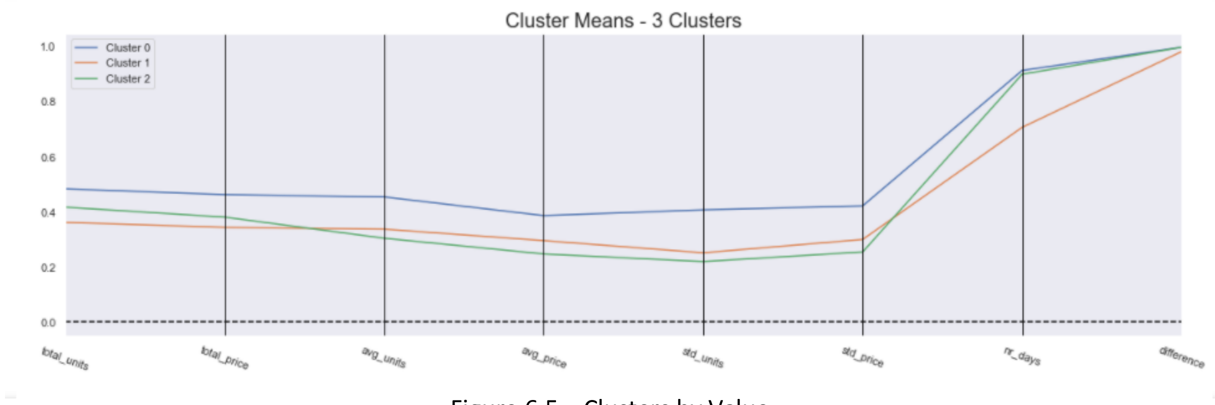
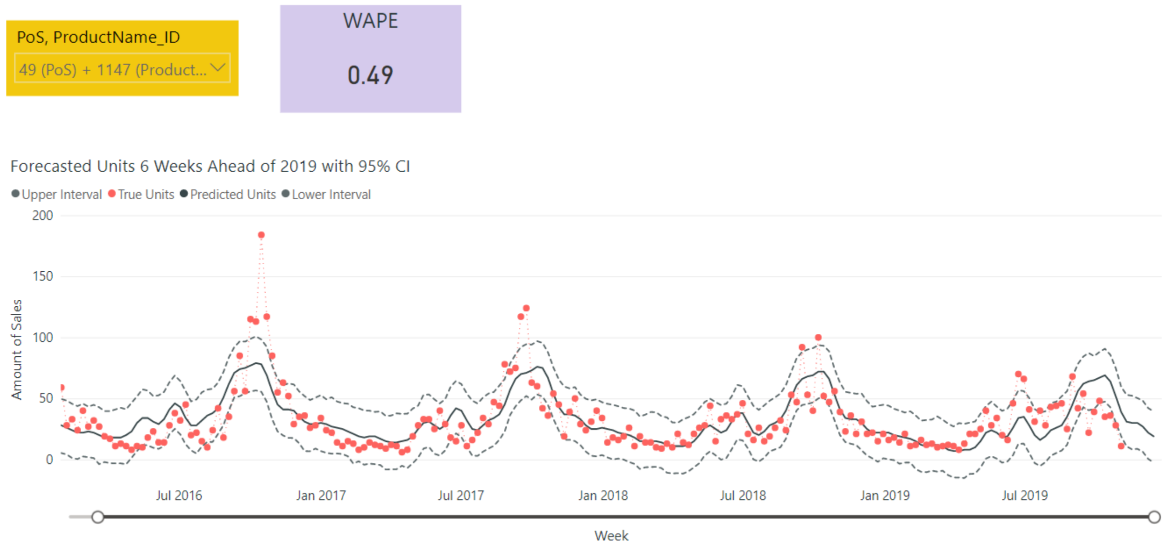


Figure 6.5 – Clusters by Value



PoS, ProductName_ID

49 (PoS) + 1147 (Product...

WAPE

0.49

Figure 6.6 - Forecasting of the Product 1147 Sales in the PoS 49