

**BUSINESS CASES WITH DATA SCIENCE**

**MASTER’S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

**Mind Over Data - Retail Challenge**

Insights on the business, Point of Sales clusters and product forecast

Group Y

João César (20200669)

João Henriques (20200670)

Pedro Sancho Vivas de Castro (20200132)

Vilmar Bussolaro (20200268)

June 2nd, 2021

INDEX

[1. INTRODUCTION 1](#_Toc73552393)

[2. BUSINESS UNDERSTANDING 2](#_Toc73552394)

[2.1. Background 2](#_Toc73552395)

[2.2. Business Objectives 2](#_Toc73552396)

[2.3. Business Success criteria 2](#_Toc73552397)

[2.4. Situation Assessment 2](#_Toc73552398)

[2.5. Determine Data Mining Goals 3](#_Toc73552399)

[2.6. Project Plan 3](#_Toc73552400)

[3. MULTI-PURPOSE DATA SCIENCE SOLUTION 4](#_Toc73552401)

[3.1. Data Understanding 4](#_Toc73552402)

[3.2. Data Preparation 4](#_Toc73552403)

[3.3. Data Mining and Machine Learning Solutions implementation 4](#_Toc73552404)

[3.4. DM & ML Solutions Evaluation 6](#_Toc73552405)

[4. RESULTS EVALUATION 7](#_Toc73552406)

[4.1. Point-of-Sale Level Quarterly Analysis of the Business 7](#_Toc73552407)

[4.2. Point-of-Sales Clustering 7](#_Toc73552408)

[4.3. Product Forecast for Different Point of Sales 9](#_Toc73552409)

[5. DEPLOYMENT AND MAINTENANCE PLANS 10](#_Toc73552410)

[5.1. Plan deployment 10](#_Toc73552411)

[5.2. Maintenance 10](#_Toc73552412)

[6. CONCLUSION 11](#_Toc73552413)

# 

# INTRODUCTION

In this project we are challenged by *Mind Over Data* to develop a data mining project for an appliances store from Australia. This project forced us to face 5 different challenges, to analyse the data and transform it, to create insights from the data specially to better understand each Point-of-Sale characteristics, to create clusters for each Point-of-Sale, and to forecast sales for the next 6 weeks in general and by Point-of-Sale.

*Mind Over Data* is a company that markets itself as a smart problem solver, developing tailored software to provide solutions, processes, and protocols to other businesses.

The database provided consists on the daily sales of each of the 1535 products for each of the 410 Points-Of-Sale since January 2016 to November 2019.

We are following the CRISP-DM methodology to achieve the final product, which is both a recommender system and this report.

All information provided in this report, the presentation, and the respective notebooks with the recommendations for the company will be made available at: https://github.com/PedroSancho

# BUSINESS UNDERSTANDING

## Background

*Mind Over Data* is a software company that provides tailored solutions to its customers that can vary from big corporate institutions to small businesses. They shared with us data from an appliances store in Australia from which we know very little about, especially because the data provided only contains the ID’s excluding the names from the dataset.

The appliances chain has 410 Point-of-Sales, all of which we are assuming to be physical stores. From the high volume of sales and the number of stores it is tempting to assume that this company is one of the biggest from Australia in the appliances area of business.

This chain of stores contracted *Mind Over Data* to provide the best information possible about their business, however the most important business dimension for them seems to be the Point-of-Sale, since it is the one that is emphasize the most.

## Business Objectives

This analysis has as objective understanding the characteristics of each POS. We want to know how the top products sold, market shares and product co-occurrences change throughout each quarter, how many different POS segments there are and the respective behavior splitted between value and product preference and in the end, we want to be able to forecast the following 6 weeks sales, from both the entire company and for each specific POS.

## Business Success Criteria

The Business Success Criteria can be broken down into three main goals:

* Provide a quick to access exhaustive Business Intelligence platform to help senior management diagnose at Point-of-Sales level what are the top products, sales performance, categories market-share etc.
* Provide an intelligible clustering solution for Point-of-Sales using both lifetime value of their operations to the whole chain of retail stores as well as to capture their different preference segmentation.
* Provide a data pipeline that would help the retailer brand to predict at a Point-of-Sales level products’ sales, e.g., data ingestion, data preprocessing, model training and forecast.

That means meeting the intended client’s goals by providing solutions that will help bring new and assertive insights on the operations on fast-paced fashion, therefore, it would be safe to say that achieving this defined business success criteria will help on the decision-making of client’s senior management.

## Situation Assessment

The database had 27GB (when loaded in Pandas) of data containing information starting from the 1st of January of 2016 until the 1st of November of 2019, each row containing information about the daily sales from one POS (point-of-sale) for a specific product, daily number of sold items and the amount received from said sales. For each row we also have information about the product category as well as its brand, family and SKU (stock-keeping unit). We have a total of 21 Families of Products, 178 Categories of Products, 1.535 Brands Products, 8.660 SKUs and 410 Points-of-Sales.

Our team is composed of 4 Data Scientists, which had 2 weeks to prepare a 5-minute presentation to the C-Suite, as well as a 10-Page Report (which you are reading) and the accompanying code.

## Determine Data Mining Goals

Regarding data understanding requested by the client, to achieve actionable insights about POS characteristics, we recommend building a dashboard on PowerBI, where the user will be able to perform quarterly analysis of top products sold, as well as the product families and product categories, that way showing market-share preferences.

To segment POS based on their value and product preferences, the best approach is to use Unsupervised Learning. By using a clustering method, we can select an appropriate algorithm to identify which sets of POS are more similar than others, therefore forming groups: Those groups (or clusters) are easier to comprehend and act upon.

To forecast product sales for the next 6 weeks, on all stores and on a specific POS, this analysis will have to remove deprecated products from historical data and using time series analysis project sales patters to help stock allocation more accurately.

But before bringing information to the surface, we need to do the basics: preprocess data. We will use Data Mining knowledge to determine the optimal way in order to meet expectations.

All insights should be only based on the data, the number of clusters need to be assessed by elbow methods, validated by the results as a viable number when achieving clear differentiation between groups of customers and the forecast needs to be done with at least a silhouette of 0.5.

## Project Plan

1. Data understanding  
2. Data preparation

2.1. Regex preprocessing

2.2. Feature *downcasting*

2.3. Data trimming

2.4. Data aggregation

2.5. Feature engineering

4. Business Intelligence creation on top of Point-of-Sales level quarterly pre-processed data   
5. Cluster analysis and selection for Point-of-Sales  
6. Cluster visualization

7. Pipeline creation for forecasting products at Point-of-Sales level on a Weekly basis  
8. Deployment

9. Maintenance

# MULTI-PURPOSE DATA SCIENCE SOLUTION

## Data Understanding

After taking a first look to our data, we found out the given dataset has 9 different variables. All of them being relevant, however the column value as double meaning since depending on “Measure” value it can either mean the amount that item that was purchased during that day or it can mean the amount of revenue that came from those purchases. Therefore, some data preparation will be needed in order to change this.

It’s also already known that we have 21 Families of Products, 178 Categories of Products, 1 535 Brands Products, 8 660 SKUs, 410 Point-of-Sales and data from 1 Jan 2016 until 1 Nov 2019.

## Data Preparation

Preprocessing was done using mainly *Pandas* and *Vaex* libraries. The following steps were performed: (i) *Regex* preprocessing; (ii) Feature *downcasting*; (iii) Data trimming; (iv) Data aggregation; (v) Feature engineering;

The dataset provided was the first big challenge of this project. We had two lines for each product, for each of the 410 points of sale, for each day of this period of three years, so our first step was to reduce the size of the dataset which was crucial to our whole project.

First, we *downcasted* all the variables with IDs to integers and the variable “Value” to a float so we could save some space on this dataset.

Then we faced a problem, we had two lines for each product day sales, one for the units sold and another for the value. What we did was creating a single ID for every pair of rows, merging every ID (“ProductCategory\_ID”, “ProductFamily\_ID” and so on) as well as the date, so we can distinguish every pair of rows. Then, we added two different columns, one for the value and one for the units sold and grouped by the ID we had created. This way reduced our number of rows to half of what we had before, saving a considered amount of space and having a more practical *dataframe*, with every variable in a different column.

Our final step of preprocessing was to prepare the dataset for our different analysis, we created two different *dataframes*, one grouped by every quarter, so we could do our quarterly analysis, and another grouped by every week, so we could do the weekly forecast.

## Data Mining and Machine Learning Solutions implementation

**3.3.1. Business Intelligence Solution for Point-of-Sale Level Quarterly Analysis**

For our Point-of-Sale Quarterly Analysis we had to analyze both the top products sold and the Market Share of them top product family and product category.

Having this objective, we decided to approach this with a PowerBI Report with two pages. Both of them having two slicers, one where we can navigate through the different points of sale and another where we can select the quarter we want to analyze.

On the first page we focused more on the five products that were most sold. The visualization we chose to present this information was a tree map and we also provided a table with the name of the products and a more detailed view on the quantity sold.

The second page had two different variables analyzed separately: the Product Family and the Product Category. The goal was to provide information regarding the share each family or category had in a specific point of sale on a specific quarter. This time we decided to focus on the Value and not on the Quantity, since we are talking about market share. Apart from the slicers above mentioned we can find a pie chart for each analysis that only show the distribution around the top 5 families or categories, and also a table where we can find the value of each family and category with more detail.

**3.3.2. Point-of-Sale Clustering Solution**

In order to cluster the points-of-sale we were asked to take have two approaches first look at the value (sales) provided by each POS and then look at the product preferences.

With this directories our team considered that the best solution would be to do 2 clusters solutions, one for each point of view, and then merge them, taking full information from them and been able to potentially correlate both of them.

To do the clustering we used in both solutions a K-means clustering algorithm and to get an appropriate K for the clustering solution, the elbow method was used. The following measures were used: Distortion, Silhouette and Calinski-Harabasz. The appropriate number of clusters for both implementations turn out to be 3.

Focusing now on the value clustering, it was rather easy, just group by POS and select the columns we were using to cluster (total value, total\_units, amount\_of\_purchases), scale everything using MinMax scaller and them apply the k-means algorithm with K equal to 3.

However, due to the number of different products that existed on the dataset clustering using product preferences was way harder than original expected. Our team decided to cluster using just the families of products since they were rather less, only 21 different. One hot encoder technique was used in this ‘ProductFamily\_ID’ and then group by POS aggregating by the sum. This made it so that we ended up with a data frame composed of 21 columns (one for each family) and 410 rows (one for each POS) with the information about the number of times each family of item was purchase in each POS. Then we followed the previously explained method of scalling and then applying the K-means algorithm.

**3.3.3. Product Forecast at Point-of-Sale Level**

In order to do so, we implemented a *Pycaret* pipeline to generate these predictions and *Pandas* for the final preprocessing necessary before making doing the prediction phase. The solution provided a pipeline that would take less than 10 days (since originally we had 14 days for the solution of the business case it sounded plausible) for the 355 thousand time series to predict.

Parallelization could be performed to improve the efficiency of the pipeline, therefore, there is a clear understanding that this is a necessary improvement in order to implement this pipeline long-term and generate predictions in a quicker fashion.

## DM & ML Solutions Evaluation

**3.4.1. Business Intelligence Solution for Point-of-Sale Level Quarterly Analysis**

There is not much to say regarding this solution. PowerBI is a useful tool to transform data into information with easiness regarding the readability. Since the goal was to provide an exploratory analysis of data PowerBI is a reasonable implementation choice.

We were able to get the insights we intended in the beginning and with very good visualizations that will clearly show in a simple and objective way the information asked by Mind Over Data regarding the different points-of-sales’ top 5 products sold, product categories and product families throughout the years 2016, 2017, 2018 and 2019 quarterly.

**3.4.2. Point-of-Sale Clustering Solution**

In order to create meaningful clusters, we used the K-Prototypes, which uses K-means for numerical variables and K-modes for categorical variables. Besides that, we did some feature engineering to create a dataframe with the rows being each POS and all product families as columns and the percentage of the purchases in which that product family appears on the basket on its respective column for each POS as values.

Both implementations were assessed using R2 and how clusters’ centroids differ. Additionally, for the K-Prototypes solution, it was also used T-SNE 3D visualization and intercluster distance mapping to assess using principal components how much fuzzy some POS could be on this last K-Prototypes implementation.

**3.4.3. Product Forecast at Point-of-Sale Level**

By using the preprocessed and weekly aggregated data, the group must proceed on with a few more specific processing measures in order to trim the amount of data that will go inside the forecast pipeline and therefore accelerate the process of giving predictions. Furthermore, some rationale like products that were discontinued were trimmed in order to compute only the necessary time series for all POS.

Lastly, we implemented a pipeline which used a cross-validation method and assessed for all time series (binary of POS and the products sold) different algorithms and selected the best algorithms with respects to MAE (Mean Absolute Error). Additionally, the trained models were saved in *pickle* files for reproducibility of the results later on.

# RESULTS EVALUATION

## Point-of-Sale Level Quarterly Analysis of the Business

It is hard to translate information for every one of the 410 points of sale for every quarter in this report, but we can take a random point of sale and analyze its information for the variables we analyzed. Independently from this, the most obvious insight we can take is regarding the product category market share. We can see that the product category 178 is the dominant by a huge difference, in terms of value. Overall, in every point of sale combined through all this period, this category’s sales represented 87,22% of the whole sales.

Let’s take, for example, POS 21. On the first quarter of 2016 we can see that the top five products most sold are the products 2376, with 906 units sold, 993, with 573 units sold, 481, with 515 units sold, 1234, with 509 units sold and 567, with 496 units sold. Regarding the product categories, the five that provided more value were the 178, with more than €27 million, 127, with almost €2 million in sales, 109 with €1,7 million and then 27 and 34 with €744.457 and €633.568 respectively. Finally, regarding the product families, families 9 and 2 were the ones with the highest market share, both with almost €7 million, followed by 21 and 12, with a value rounding €6 Million and 1 with €3,6 Million.

We can find this report where all the information can be found with more detail [here.](https://app.powerbi.com/links/0aauxeH5HW?ctid=e4bd69ff-e6f7-4c2e-b247-41b54ba2490e&pbi_source=linkShare)[[1]](#footnote-2)

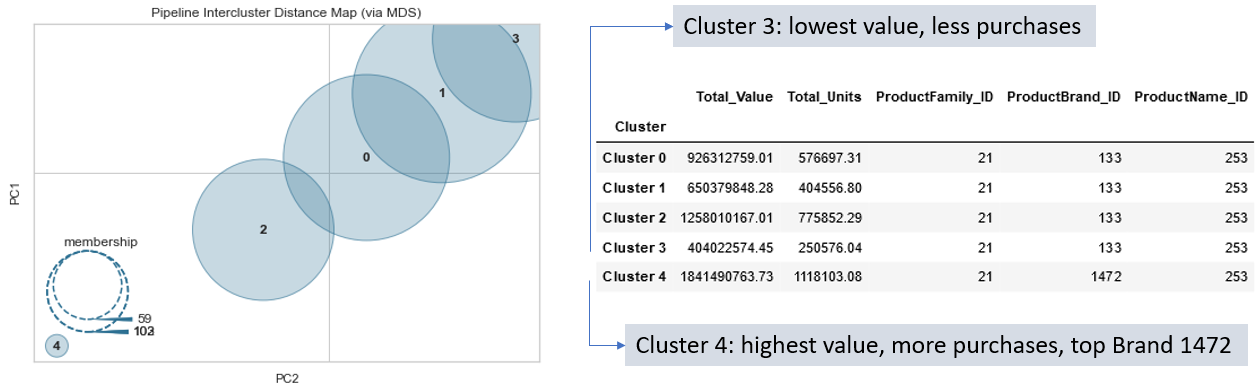
## Point-of-Sales Clustering

We approached this problem with 2 different solutions: (i) K-Prototypes of the final pre-processed dataset, using *Pycaret* library; (ii) K-means on Family Products, using feature engineered columns based on presence of family product in the baskets ordered on each POS.

**4.2.1. *K-Prototypes* implementation**

The first one was using *K-Prototypes* (on a *Pycaret* implementation). With POS\_ID as guide (groupby), our Metric Features as Total\_Value, Total\_Units, and Categorical Features: ProductFamily\_ID, ProductName\_ID, ProductBrand\_ID.

The elbow method showed an ideal value of K=5, which we implemented and achieved a Silhouette of 0.54 and Calinski-Harabasz of 1417.59.



Our clusters 1, 3 and 0 had a lot of members overlapping, but Cluster 3 is clearly the one with lowest sales and lowest number of purchases. On contrary, Cluster 4 is distinctively far away from the others, with the highest value of sales, the highest number of orders and a different ProductBrand\_ID. Cluster 4 of POS should be always in the mind of management team, because those are prime locations that bring a lot of value to the company.

**4.2.2. *K-means* on feature-engineered product preference feature set**

First, it is needed to use the *GetDummies* method for the column regarding the ProductFamily\_ID and later proceed to group by POS\_IDs using the mean, in order to retrieve per POS\_ID the amount of days that certain families were sold. Thus, the final engineered and preprocessed dataframe for this implementation will have 410 rows, accounting for all POS, and 21 columns, accounting for all product families ids. Additionally, since all values will account for percentages, ranging from 0 to 1, confirming that the implementation will be done in a numerical dataset.

Then, after performing a MinMaxScaler method on the dataframe we can proceed to implement to find the elbow point in three most used graphs to do so, which are Distortion, Silhouette and Calinski-Harabasz. Lastly, we implement the optimal solution, being k = 3, and we present the following results:

Chart, line chart

Description automatically generated

Interesting point could be extracted from the difference encountered on the centroids for the 3 clusters that depict the product preference on the POSs: (i) little differentiation on product families 19 and 17 for all three clusters; (ii) clusters 0 and 2 are pretty alike regarding preference for product family 1 and 6; (iii) clusters 0 and 1 depict very close preferences on product family 20; (iv) all product families not mentioned previously and product family 21 seems to depict the same characteristics, which are big difference between all three clusters preference for these product families, and an orderly consistent preference for product families namely cluster 2 being the highest, cluster 0 being the lowest and 0 being on the middle of the two other clusters; (v) product family 21 preference seems to inversely correlated with all other product families described on the previous point.

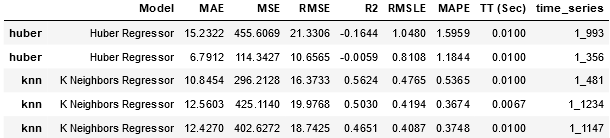
## Product Forecast for Different Point of Sales

For prediction we also used the *Pycaret* library. But the large amount of the dataset forced us to approach this problem as an MVP (Minimum Viable Product). Because the regression algorithm needs to create a different model for each *time\_series* (which is comprised of *POS* + *ProductName\_ID*), and that is being cross folded 3 times, running the whole dataset would take over 11 days in our computers. So instead, we will produce prediction results for POS 1, and the same can be repeated by the company with better computing resources.

To start the process, we did some data wrangling specifically for this purpose, with the file “weekly\_final\_for\_forecast.csv”. Here data is already grouped into time\_series, one for each product name and for each POS, and values are grouped by week.

To deal with products that were no longer on sale, we created a filter that removes products that were sold less than 10 weeks in the past years or that had zero sales in the year 2019.

After feeding the data to Pycaret in an iterative fashion (automated pipeline), the library tests 25 different regression algorithms and chooses the best one based on MAE (Mean Absolute Error). This metric is calculated using a train-test of 0.95, and it´s folded 3 times without shuffling to take into consideration that it is a time series. The best model is then fine tuned to perform even better. The hyper parameters are then exported into a pickle file for use later in the process.



In the image above, we can see 5 time\_series (POS + ProductName\_ID), each with their own correspondent Model.

Chart, histogram

Description automatically generated

With the models created in pickle files, then it´s possible to retrieve that one by one and perform the forecast. In the example above we have an example, made for POS1, Product 993.

# DEPLOYMENT AND MAINTENANCE PLANS

## Plan Deployment

Deployment could be in a POS fashion, giving the forecasting pipeline for each POS to forecast only their time series (POS+Product). By using this solution, implementation issues will be diminished since parallel workstations, e.g., from different POS, would focus on subset of the whole pool of time series, only their respective time series, for the whole chain of retail stores.

This could generate a forecast pipeline that would take as much as two hours to forecast all time series from its POS, using a 3-fold cross-validation on a rolling basis schema for training and assessing 25 different algorithms to select the best implementation for each time series using the *PyCaret* forecasting pipeline the group designed.

## Maintenance

Maintenance will be addressed in the solution since each time the forecasting pipeline is performed, each time series for every POS is retrained. This comes to the expense of spending 2h of one workstation at each POS, which accounts to 820 workstation-hours at the aggregate. Despite that, it seems to be a good implantation seems it is a distributed solution, that decreases bottlenecks on resource utilization, avoids altogether the necessity of using cloud computing and fosters data science familiarity within the retail store’s operations.

Attention points must focus on the data ingestion phase, the preprocessing notebook can receive new data and perform all the preprocessing steps necessary for data ingestion on the forecasting pipeline. Furthermore, it is important to feed and concatenate new data that was preprocessed on to the forecasting pipeline in an orderly fashion. Although not difficult, this is something that must be done in a centralized fashion and then dispersed throughout the POS just for the execution of the forecasting phase, since they will be the ones that must act upon it, with regards to supply management, customer acquisition campaigns etc.

# CONCLUSION

*Mind Over Data* trusted us to deliver a full project on an appliances store chain from Australia. The project had many phases but like any other business cases the first step was data understanding, this phase was, as mentioned above, rather not complicated, however it was harder than what it could have been if we actually had names of everything instead of just the ID’s. Data preprocessing was by far the hardest part of the project since we had to find a way to decrease the size of our csv file that originally occupied 27GB of memory when loaded into Pandas and was impossible to work with, to help us on this task we used the chunk method to make changes on portions of the dataset (chunks) at each time, as well as a number of different data preprocessing described during the report.

We were asked to do four tasks inside this project, being the first, data/feature engineering. It was done throughout the project but more specifically on the preprocessing stage, and again on the forecast task. The next task was to do a full point-of-sale quarterly analysis with had to take into consideration top products sold, market share (Family, Category), preferences and Product co-occurrences, this task was done with the help of a PowerBI dashboard, a really important tool on the business world.

For the clustering task we did two implementations: one was a *K-Prototypes* implementation with the help of the *Pycaret* library, this allowed us to segment our POS into 5 clusters where we could clearly see with ones provided the most value to the company, the other cluster implementationwas a *K-means* algorithm on feature engineered product preference feature set that after implemented made it so that we ended up with 3 cluster were we could see which POS prefer which family items.

Lastly, the prediction was done with the usage of the *Pycaret* library, due to large amount of the dataset, we had to do some more data engineering and only after that we were able to test different algorithms, 25 different algorithms for each time series (pair of POS+Product). It is important to not forget that predictions were made for the year of 2020 that was highly affected by the covid pandemic so the results more likely than not were not as predicted.

On that note, the group adds a suggestion of the implementation of a *VAR* (Vector of Autoregressive) model that could have an equation on mobility index on its specification, which could help on diagnosing the sales impact on in-store purchases due to decrease in mobility from the advent of COVID-19. Thus, by assessing through an impulse-response function (IRF), an exogenous shock could be generated in mobility and the analysis of the corresponding effect on units sold could be analyzed through the IRF graph by utilizing the *VAR* model.

Overall, our team thinks this project was a success and hope to have been able to meet expectations and as well as to somehow surprise both Mind Over Data Head of Analytics, Vasco Jesus, and the teachers on our final business case.

1. https://app.powerbi.com/links/0aauxeH5HW?ctid=e4bd69ff-e6f7-4c2e-b247-41b54ba2490e&pbi\_source=linkShare [↑](#footnote-ref-2)