Business Case #5

MASTER DEGREE PROGRAM, DATA SCINCEAND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Mind Over Data - Retail Challenge

Ana Sofia Silva, number: 20200220
José Francisco Alves, number: 20200653
Miguel Nunes, number: 20200615
Mohammadali Gharghi, number: 20200997

Index

INTRODUCTION

The present project was proposed by Vasco Jesus, Head of Analytics in Mind over Data company. Mind Over Data is a Data Science and Advanced Analytics boutique consultancy, providing tailor made Advanced Analytics solutions for companies worldwide. This study case is in the scope of retail business which is a company that sells mostly household appliances. The provided dataset for this business case was provided in a large file size, and for that, a several of challenges came along, such has data handling, transforming, and modelling. The proposed challenges have the goal of help the business management and give important insights on the customer behavior and segments.

BUSINESS UNDERSTANDING

BACKGROUND

This business case is a real-world data science project regarding a retail company of household appliances. The company sells a variety of products that are divided in Families, Categories, Brands and SKU's.

The data science team expects to use a set of data mining and machine learning methodologies to facilitate the company's management to respond a variety of problems such as the customer segmentation and a deeper business understanding according to historical data.

BUSINESS OBJECTIVES

The objective of this business passes by getting a deeper understanding of the business. Also, to understand each store characteristics that are represented as "Point-of-Sale" with the information of the Date, ID's and the value of each product.

With the provided dataset, the team could come up with the business solution for the problems mentioned below:

- Help the business management to have known of the types of customers
- Provide an understanding of the business and important insights
- Understand each Point-of-Sale characteristics
- Customer Segmentation
- Knowledge of the future product's demand

BUSINESS VALUE

The expected outcome of this study case includes implement adequate evaluation strategies for the retail business and give important insights for proper campaigns and sales predictions. With the provided visualizations tools and analytical study, the data science team can also provide suitable measures and an accurate evaluation of the business.

DETERMINE DATA MINING GOALS

The business case has different suggested data mining goals, which are:

1. Data Analysis and Transformation with feature engineering
2. Understand the monthly Point-of-Sale characteristics with computational methods and visualizations tools
3. Clustering with K-Means with Principal Analysis Components (PCA) and RFM techniques
4. Unit products forecast with ARIMA and Unit Product forecast by Point-of-Sale for 6 weeks ahead.

ANALYTICS PROCESS

DATA UNDERSTANDING

The provided dataset is an approximately 20Gb file that has information regarding the transactions of the retail store. The dataset has the customer's sales records between 1st January 2016 and 1st November 2019 which include the following information: 'ProductFamily_ID', 'ProductCategory_ID', 'ProductBrand_ID', 'ProductName_ID', 'ProductPackSKU_ID', 'Point-of-Sale_ID', 'Date', 'Measures' and 'Value'.

Due to the size of the data source, the dataset was read with the PySpark tool – an interface for Apache Spark in Python. It not only allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively analyzing your data in a distributed environment. PySpark supports most of Spark's features such as Spark SQL, DataFrame, Streaming, MLib (Machine Learning) and Spark Core. (PySpark Documentation, s.d.)
PySpark will be used in the following sections of Data Preparation and Analysis.

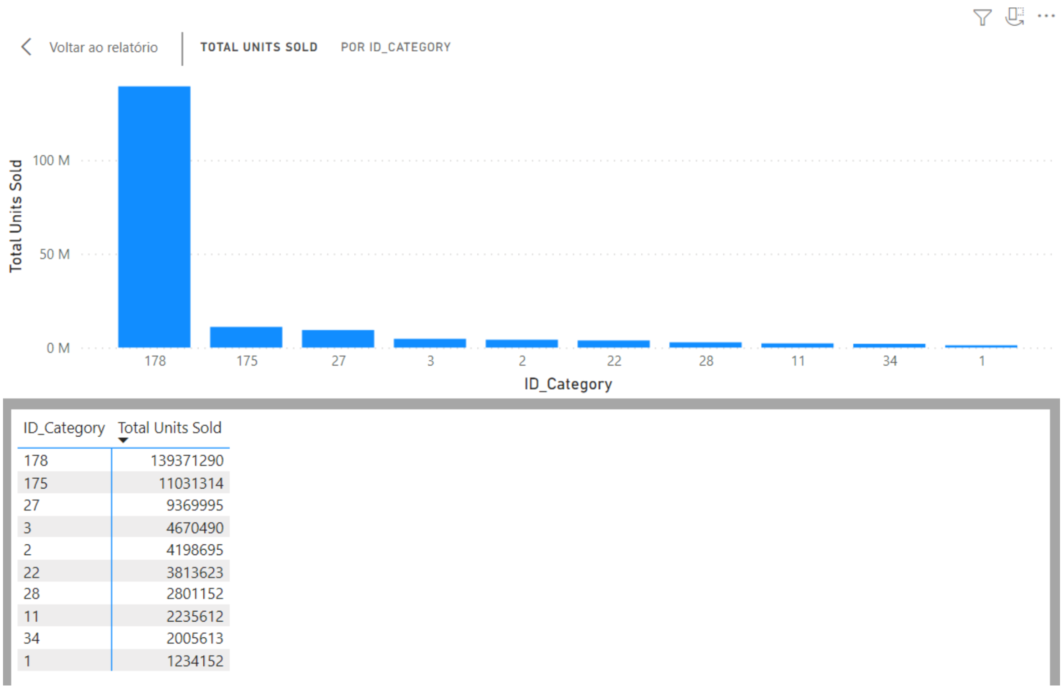
DATA PREPARATION

For better analysis and consecutive interpretation of the possible results, some data preparation was done. One of the first situations was to delete irrelevant information of the ID's columns – deleting characters and keeping the digits to reduce the file size. The two distinct entries ('Unit' and 'Value') of the original column 'Measures' were split in two columns and the duplicates were dropped. The data types of the columns were changed to 'IntegerType', excluding the 'Date' column that was changed to 'DateType'.

Another piece of data that was considered relevant to analyse was the unit price of each product. To do this, the column 'Unit Value' was created with the information of the columns 'units' and 'values'.

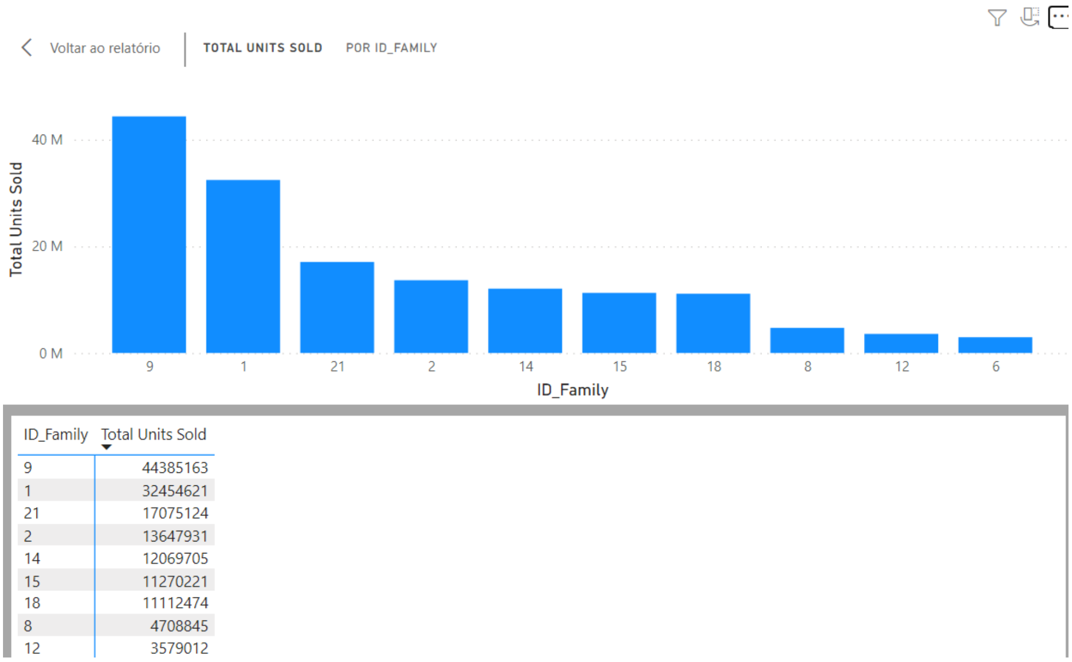
Data Analysis

For data analysis, it was done a brief overview of the units sold by Category, Family, Brand and PackSKU. The information about the top 10 sold products by ID of category, family and PackSKU was assessed, and the visualizations **Erro! A origem da referência não foi encontrada.**, **Erro! A origem da referência não foi encontrada.** and **Erro! A origem da referência não foi encontrada.** were created with PowerBI. Also, the visualizations in Figure



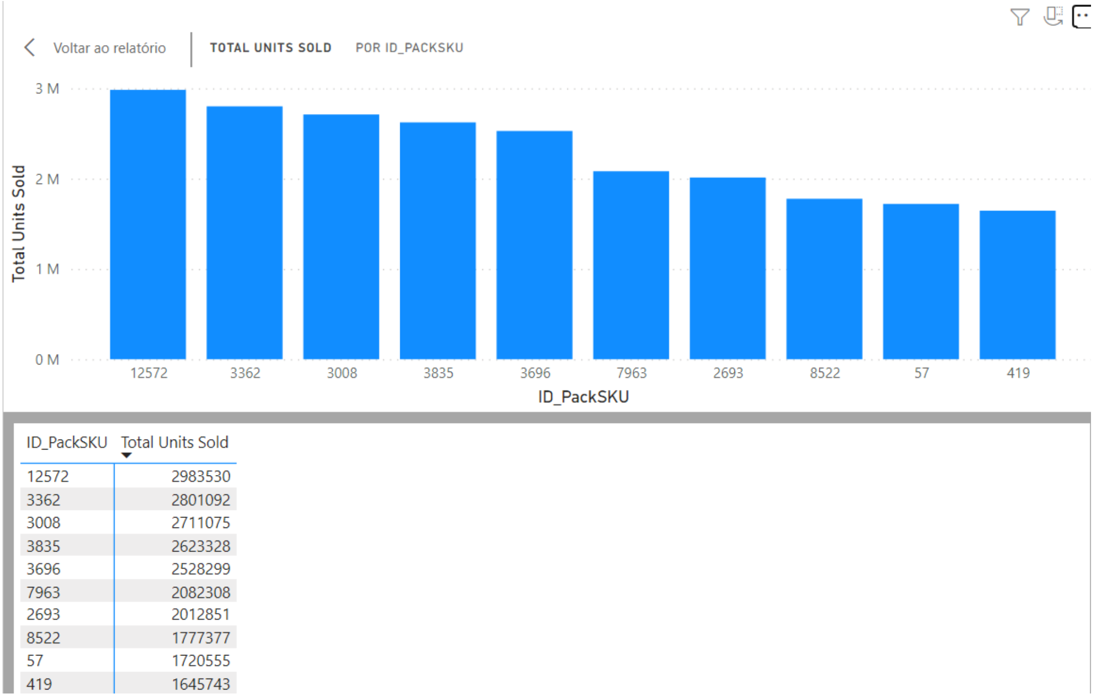
ID_Category	Total Units Sold
178	139371290
175	11031314
27	9369995
3	4670490
2	4198695
22	3813623
28	2801152
11	2235612
34	2005613
1	1234152

Figura 1 - Total units sold of ID_Category



ID_Family	Total Units Sold
9	44385163
1	32454621
21	17075124
2	13647931
14	12069705
15	11270221
18	11112474
8	4708845
12	3579012

Figura 2 - Total units sold for ID_PackSKU



ID_PackSKU	Total Units Sold
12572	2983530
3362	2801092
3008	2711075
3835	2623328
3696	2528299
7963	2082308
2693	2012851
8522	1777377
57	1720555
419	1645743

Figura 3 - Total units sold for ID_Family

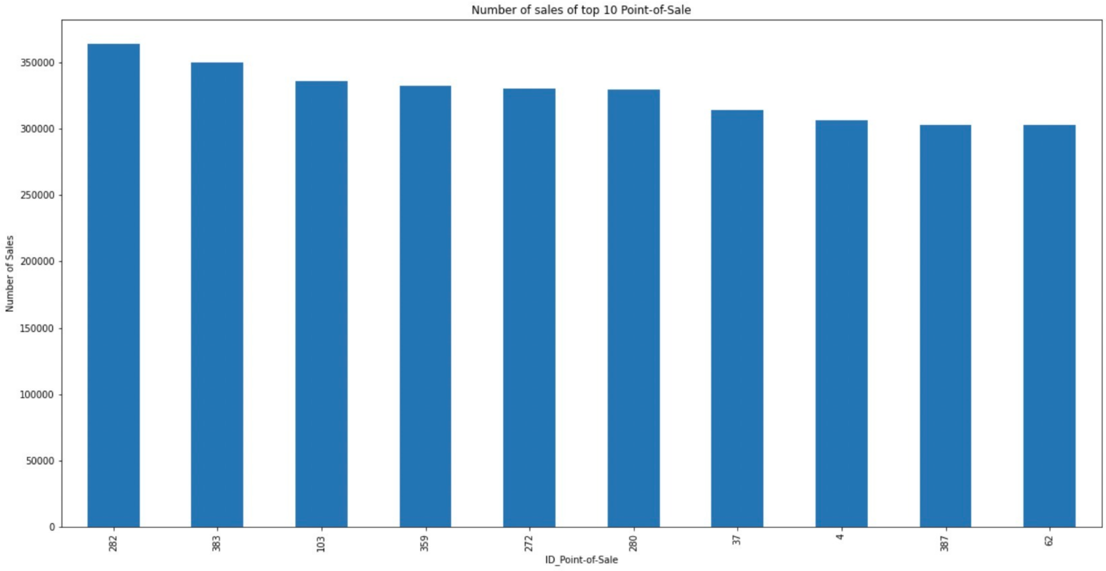


Figura 4 - Number of sales per ID Point-of-Sales

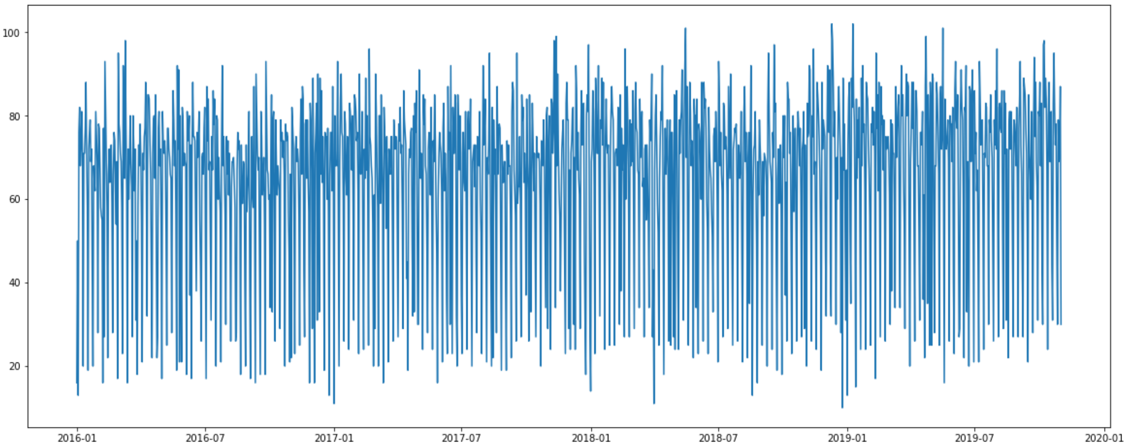


Figura 5 - The number of client's transactions during the time

CASE OF STUDY

1) Point-of-Sales clustering based on Value and Product Preference

In this challenge the clustering with K-Means was tested with multiple subsets of data according to the analysis of the groups **Value** or **Product Preference**. The following analysis were made:

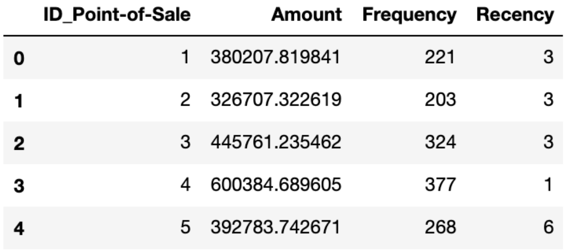
- In the first phase, we assessed the elbow method to choose an adequate number of clusters and made a visualization from it. This analysis was made for the two groups, where the group **Product** included the variables 'ID_Family', 'ID_Category', 'ID_Brand', 'ID_Name', 'ID_PackSKU' and 'ID_Point-of-Sale'. This analysis had inconclusive results due the non-numeric type of the selected variables.

The same method was tested for the **Value** analysis. The results showed a clear division of three clusters.

- The second attempt included the normalization of the entire dataset for the principal components analysis (PCA). In this section, the analysis was made for all the variables, excluding Date. The number of selected components were 4. The conclusions were that there's no meaningful relationship between Date and Point-of-Sale.

- The last attempt was based in the RFM analysis for customer segmentation. methodology. RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because frequency and monetary value affects a customer's lifetime value, and recency affects retention, a measure of engagement. (RFM analysis for Customer Segmentation, s.d.). The following table shows the dataframe after the implementation of the RFM method before the normalization.

Tabela 1 - First 5 rows of RFM analysis of the initial dataset



	ID_Point-of-Sale	Amount	Frequency	Recency
0	1	380207.819841	221	3
1	2	326707.322619	203	3
2	3	445761.235462	324	3
3	4	600384.689605	377	1
4	5	392783.742671	268	6

After the data normalization, the silhouette method was assessed, and the final number of clusters was 3.
With this methodology, it was possible to have three distinct clusters:

- Customers with Cluster ID 0 are the customers with high number of transactions as compared to other customers, spend higher amounts of money and are frequent
- Customers with Cluster ID 1 are frequent buyers.
- Customers with Cluster ID 2 are not recent buyers and hence least of importance from business point of view.

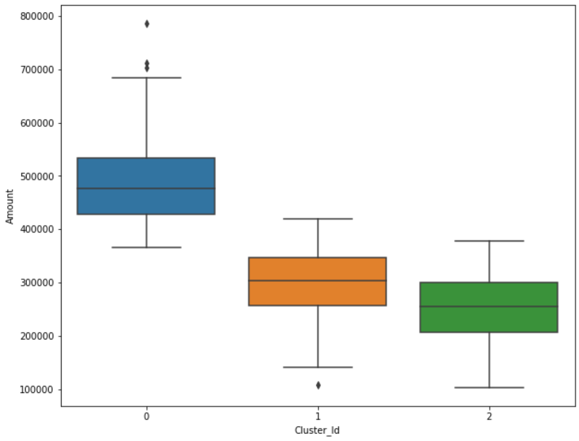


Figura 6 - Boxplot analysis between clusters ID and the variable Amount

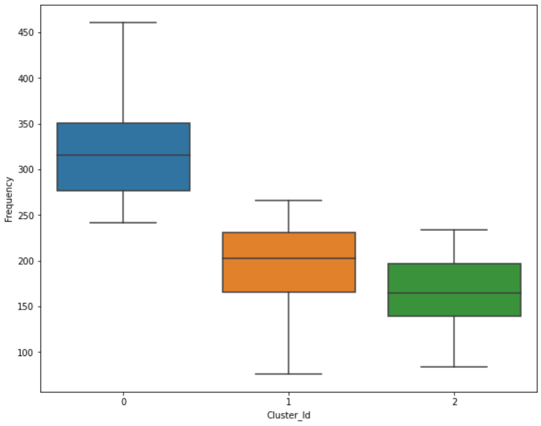


Figura 7 - Boxplot analysis between clusters ID and the variable Frequency

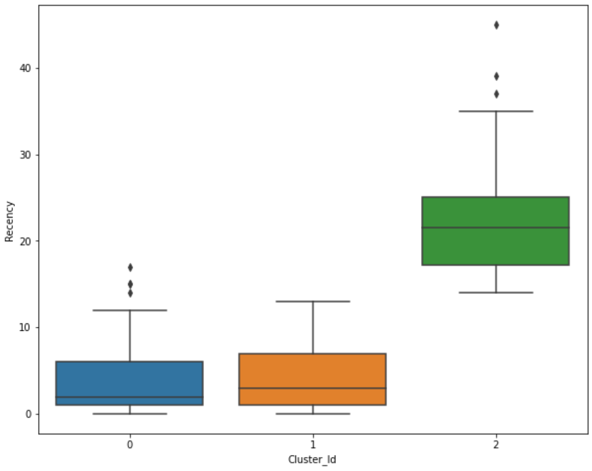


Figura 8 - Boxplot analysis between clusters ID and Recency

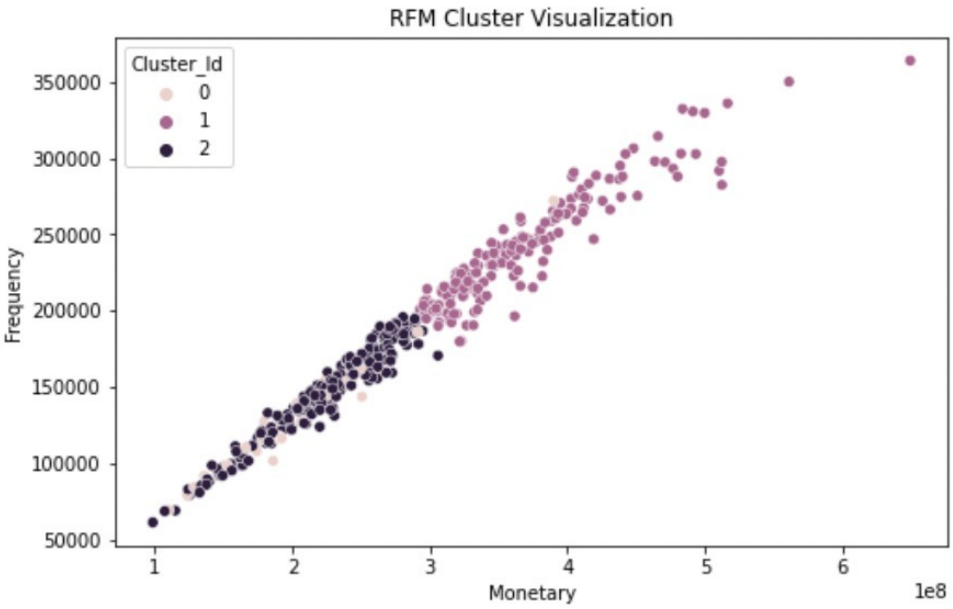


Figura 9 - Clusters visualization between Monetary and Frequency

2) Unit Product forecast (6 weeks ahead) and Unit Product forecast by Point-of-Sale (6 weeks ahead)

For the first part of the challenge, the prediction of the unit's products 6 weeks ahead, it was used ARIMA models. For that, the values from the ID's were encoded due to be categorical. The model was trained with 80% of the data and, the forecast is calculated and displayed in the orange line as we can see in Figure 7.

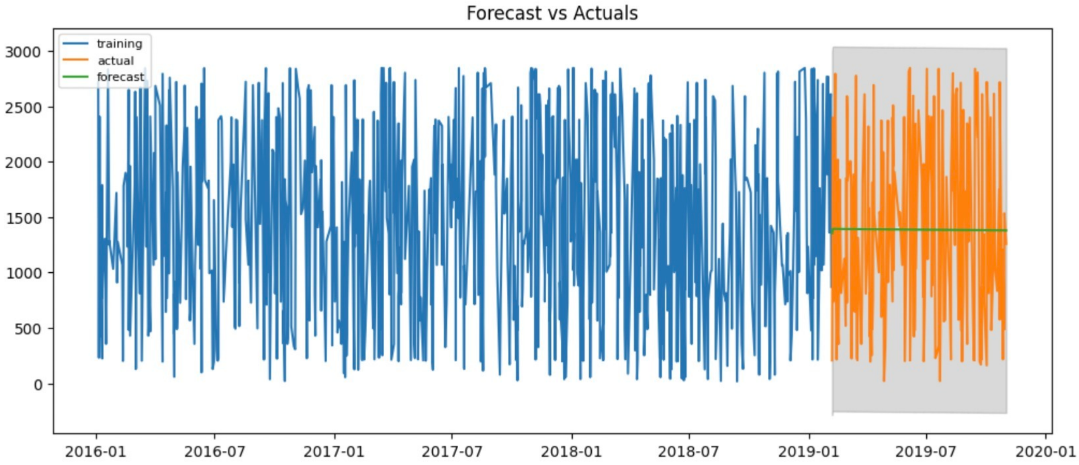
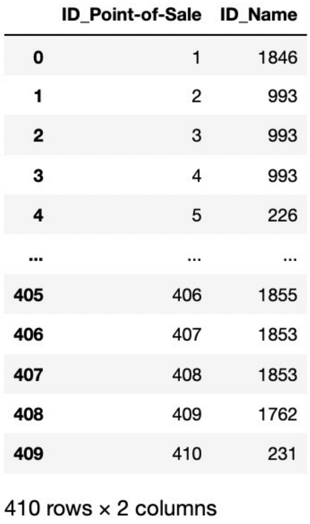


Figura 10 - Forecast for 6 weeks ahead

For prediction of products based on Point-of-Sale, because we didn't have a time series we filtered our top sales for each point of sale and suggested them, as we can see in the Figure 8. We made a filtering method to Point-of-Sale to see the predictions. As we can see in Figure 8, for example, if Point-of-Sale with ID 99 ordered product ID of 4555 in 60% of time, the chance of ordering it again is 60 percent, so our model accuracy is 0.6 too.



	ID_Point-of-Sale	ID_Name
0	1	1846
1	2	993
2	3	993
3	4	993
4	5	226
...	...	...
405	406	1855
406	407	1853
407	408	1853
408	409	1762
409	410	231

410 rows × 2 columns

Tabela 2 - Filtering Method for the unit product prediction based on Point-of-Sale

CONCLUSIONS

In this case study, the proposed challenge was analyzed, and some important insights were presented. The Point-of-sale, which is the basis of our analysis, has presented a specific number of products that have many transactions. The ID's variables were also visualized in the same way, to complement our further analysis.

To cluster the Point-of-Sales by value and product, three different methodologies were conducted, and the data science team has concluded that there are three different types of customers. In the end, with the forecast methodologies to predict the units of products, two different methods were done based on the type of forecast for the six weeks ahead.

The large dimension of the data came out as an obstacle in the preparation of the results and visualizations, due to the time spent in the script output.

REFERENCES

PySpark Documentation. (n.d.). Retrieved from https://spark.apache.org/docs/latest/api/python/

RFM analysis for Customer Segmentation. (n.d.). Retrieved from https://clevertap.com/blog/rfm-analysis/