BUSINESS CASES WITH DATA SCIENCE

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

Market Basket Analysis

Group O

Eleonora Sbrissa, M20200628

Luis Reis, M20200636

Pedro Godeiro, M20200396

Sara Michetti, M20200626

April, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

INTRODUCTION

1. BUSINESS UNDERSTANDING

Most retail companies nowadays lack of using their data to create value and new business opportunities.

Understanding the purchasing patterns/behaviors of consumer is a key task for any retail company. Identifying the different relationships between each type of product gives the retailer the chance to optimize their store's layout, develop promotional campaigns and develop new marketing strategies for their customers.

1.1. BACKGROUND

The dataset was provided by Instacart, an American company that provides a grocery delivery and pick-up service via website or mobile app in USA and Canada. The dataset contains a sample of 200k grocery orders from more than 100k customers. As it's a sample, there is information related to only some of the orders made by the customers, with the sequence of products purchased in each order.

1.2. BUSINESS OBJECTIVES

The company would like to have an overview of its business as complete as possible. They would like to know which type of customers they have, which products are more likely to be bought again and which items have buying patterns to identify important relationships.

1.3. BUSINESS SUCCESS CRITERIA

Instacart would like to know better what the main types of consumer behavior are, and which important relationships exist between the different kind of products bought by the customers, meaning which kind of products can be considered substitutes and which ones complementary. They would also like to understand which types of products should have an extended amount of product offerings.

1.4. SITUATION ASSESSMENT

Jane Doe, one of Instacart district managers gave access to 4 datasets containing different kind of information:

- Products: list of 134 different products that the retailer sells, with its product id and the department id each product belongs to.
- Departments: list of 21 departments each product belongs to, with its department id.
- Orders: records of 200k orders made by customers with user id, day of the week in which the order was made, hour of the day and days since prior order.
- Order_products: combination of order id and product id.

1.5. Determine Data Mining goals

To make a good analysis on Instacart customers buying pattern/behavior, the first thing to do is to explore and understand the datasets provided. K means algorithm will be used to find customers characteristics. Another type of analysis will be conducted in order to find product characteristics, using the apriori algorithm and association rules.

2. PREDICTIVE ANALYTICS PROCESS

2.1. DATA UNDERSTANDING

The four datasets provided contain information related to the products, the departments, and the orders customers did.

The following analysis and the one related to the different clusters were made available on a Power BI dashboard, in order for the company to have a clearer overview of their data. The dashboard is interactive, and you can see more specific insights related to specific departments/products/clusters. The dashboard is available at this link.

Products are grouped into 21 departments. Next figure shows an insight of how many products there are in each department, and how many orders there are for products related to each department:

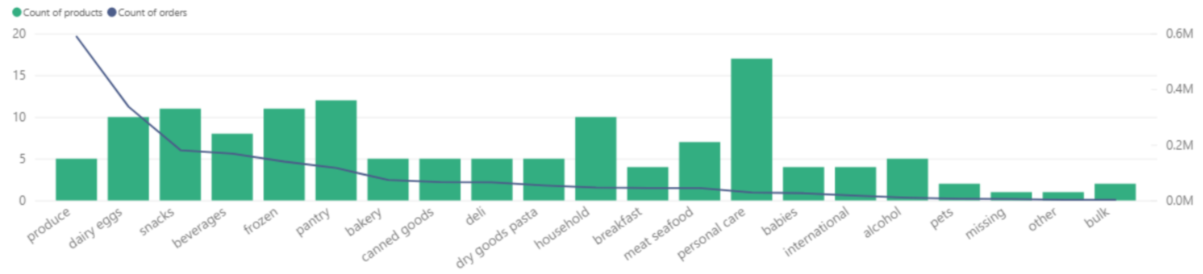

Figure 1 – Number of products per department, number of orders per department

The highest number of products belong to the department personal care, even though the number of orders related to this type of products is really low.

It is interesting to see that the biggest number of orders is related to products belonging to the department "produce", which includes fresh fruits and vegetables. This accounts for 75% of the total number of orders.

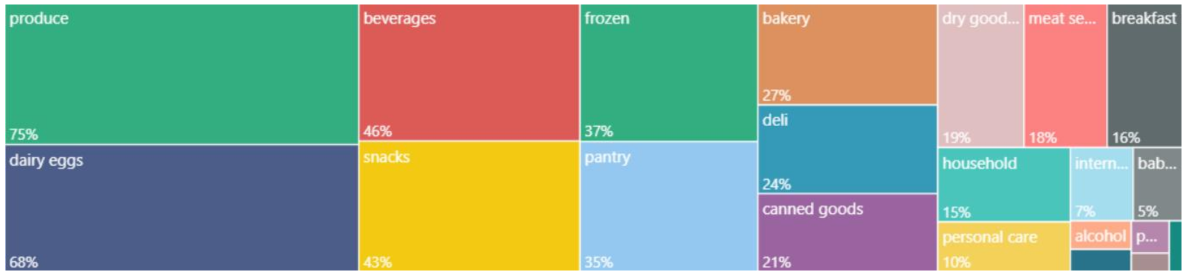

Figure 2 – Most bought products divided by Department

Getting a closer look to the top 3 departments of Instacart, it is possible to see the split in products for each department and see which products are preferred by the customers. For example fresh fruits account for 38% of the total produce orders, yogurt 27% of the total dairy eggs and seltzer sparkling water can be found in 31% of the orders when a beverage was bought. More insights can be seen in the Power BI dashboard.

Looking at customers behavior, most of the orders are made between 10:00 and 15:00, but the distribution of the orders follows something like a normal distribution after 5:00. After 19:00 the orders decrease by 51% compared to 15.00.

In order to have a better overview of the customers behavior, the team decided to group the exact hour of the purchase (available in the column 'order_hour_of_day') into time slots as follows: night from 0 to 6, morning from 7 to 10, noon from 11 to 14, afternoon from 15 to 18, evening from 19 to 23.

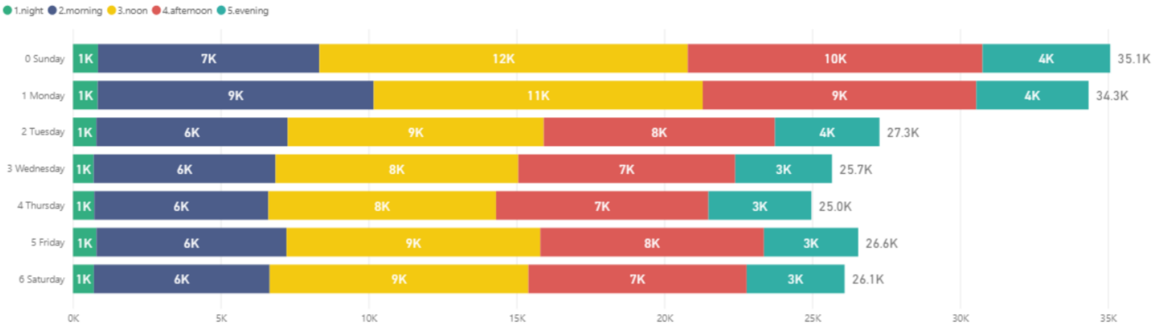

Figure 3 – Amount of orders per day of the week, divided by time slots

People tend to buy more on Sundays and Mondays, in the hours between noon and afternoon.

There is also another question to be answered: how many times a week customers buy on the website? The analysis below shows that customers tend to buy more within a week from the previous order, even though the average number of products bought is not high. There is a peak after one week since the previous order, peak that is also reflected in the average number of products bought. The other peak happens after 1 month since the previous order: It could be related to people that buy online once a month looking for discounts. It can be seen that the average number of products is not higher than average.

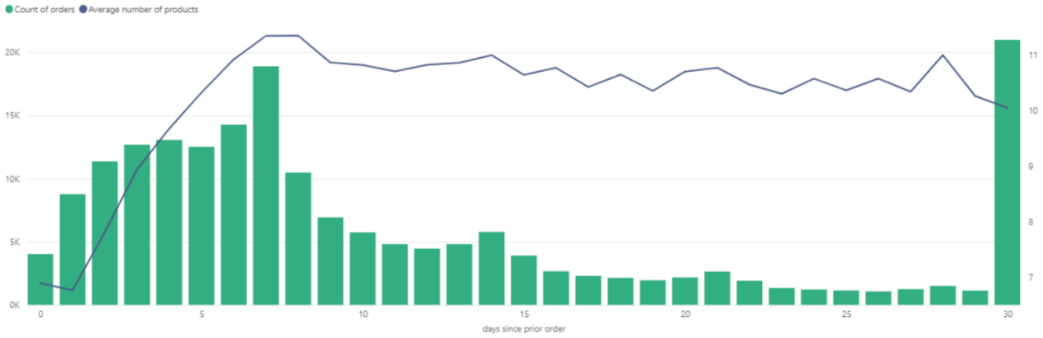

Figure 4 - Count of orders based on the previous order made, average number of products bought

2.2. DATA PREPARATION

The order_product dataset contained 28% of duplicated rows considering the combination of order id and product id. This means the same person bought two or more units/boxes of the same product in the same order. As an assumption, this analysis will not consider how many times the same product has been bought in one order, but only if that item was bought or not, so duplicates have been removed.

The report will now be divided in two parts as there are two different approaches to be taken considering data preparation, modeling and results evaluation: the first analysis will focus on understanding the different customer behaviors (through clustering), the second one will be more focused on finding substitutes and complementary products (through association rules).

2.3. K-MEANS CLUSTERING

To understand how the customers differ and in order to find different consumers behaviors it was decided to run the unsupervised learning method of clustering using the following features: order id, day of the week, hour of the day in which the order was made, days since the previous order and amount of products per basket.

This is done for the company Instacart to have a better understanding of the various groups that populate their database and understand how they could tackle different customers in different situations, but also for marketing purposes.

2.3.1. Data Preparation

To proceed with the analysis, the order_product dataset was merged with orders dataset. The analysis will be done based on the different orders, not on the different customers. As this is a sample of the dataset and it gathers only some orders made by the customers, it would be biased to focus the clustering on the customers as there is no total view of the costumer behaviors. Instead, running the algorithm based on the orders can detect some patterns.

The first thing done was to impute the missing values related to the feature 'days since previous order' using the mean. Data were then scaled using Robust Scaler.

2.3.2. Modeling

To model the data the process used was the following:

- Hierarchical clustering over 50 clusters found via k-means
- R^2 was calculated for every cluster solution in order to find the right distance between clusters for the hierarchical algorithm
- the right number of clusters was found using Ward's Dendogram
- k-means algorithm was used with the number of clusters identified

The ward distance was used in the hierarchical algorithm in order to find the final number of clusters, which were 4.

K-means algorithm resulted in the following clusters:

Cluster 0	Cluster 1	Cluster 2	Cluster 3
37857	64491	63027	34625

2.3.3. Evaluation and Results

To assess the model R^2 metric was used, resulting in 0.645 goodness of fit.

To understand the results of the clustering an analysis has been done in Power BI to show the different orders behavior.

In terms of number of orders, Cluster 1 and 2 appear to be the most numerous clusters. Even though their behavior seems to be on the average, they differ in the moment of the day in which the orders are made, as shown in the next Figure. Even if the clustering was done using the different times of the day, for visualization purposes they were grouped in different time slots.

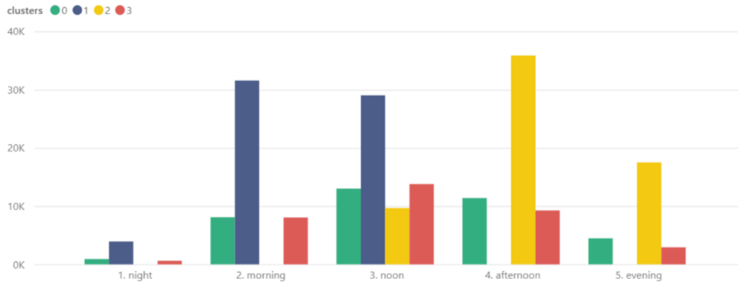

Figure 5 – Number of orders per cluster, divided by time slots

While orders belonging to cluster 1 are done at night and in the first part of the day (till noon), cluster 2 prefers to buy from noon till evening. Cluster 0 orders focus more between noon and afternoon, as well as cluster 3.

In terms of how many products are bought per order, there is a clear difference in behavior for cluster 3: orders belonging to this cluster are mainly big orders, with a minimum of 10 products.

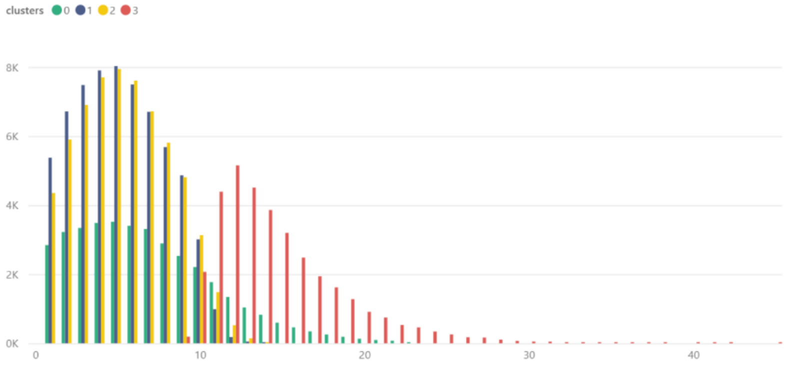

Figure 6 – Count of products per order by cluster

Orders belonging to cluster 0 are more normally distributed, with a maximum number of products of 23. Cluster 1 and 2 focus more on orders with less products, reaching the maximum at 14.

Cluster 0 is also related to orders made once a month, while all the other clusters tend to buy within a week from the previous order. More insights in the next figure.

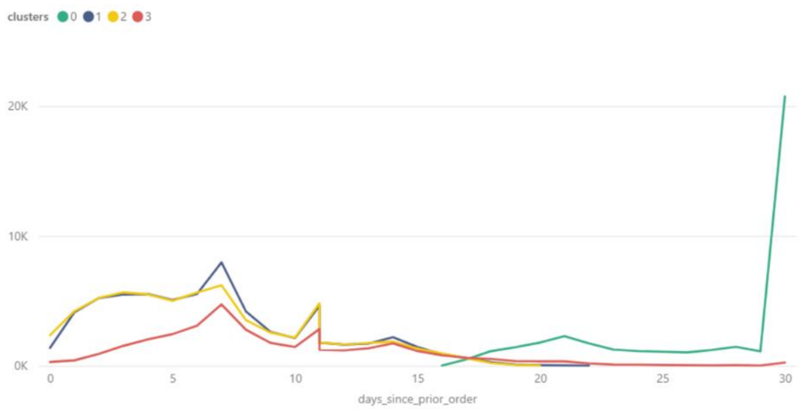

Figure 7 – Count of orders per day since prior order by cluster

To conclude, the 4 clusters have different behaviors:

- Cluster 0 – the monthly orders: orders made mostly between noon and the afternoon, once a month, with an average of 7 products per order; most likely promotions seekers.
- Cluster 1 – morning orders: orders made mostly in the morning till noon, once a week and with low quantity of products (average is 5).
- Cluster 2 – evening orders: orders with small number of products made in the afternoon mostly, once a week.
- Cluster 3 – big baskets: orders made between noon and afternoon mostly, every 9 days on average and with a bigger number of products per order.

Regarding the day in which the order is placed there was no significant difference between the clusters. You can see an insight of this analysis in the Power BI dashboard.

2.4. ASSOCIATION RULES

To provide valuable information to the business about the products sold, association rules were used. There are 3 types of rules: Actionable Rules - contains high-quality, actionable information; Trivial Rules - information already well-known by those familiar with the business; Inexplicable Rules - no explanation and do not suggest action. Trivial and Inexplicable Rules occur most often but the team is most interested in the actionable rules to supply value to the business.

To create association rules, a list of frequent itemsets is needed. This was created using the Apriori algorithm. Association rules provide information about the association between antecedent and consequent products.

Itemset: X,Y is a representation of the list of all items that form the association rule. Association Rule: X → Y is a representation of finding Y on the basket which has X on it.

To evaluate these rules there are some metrics that need to be taken into consideration:

- Support: The percentage of transactions that contain all of the items in an itemset;
- Confidence: probability of occurrence of Y given X is present;

- Lift: degree to which the products are dependent on one another; when lift is less than 1, products are substitutes; if lift is equal to one, products have no association; the bigger the lift is, the more likely the products are going to be bought together.

2.4.1. Data Preparation

To proceed with the analysis, the order_product dataset was merged with products dataset and orders dataset. A new table was then created with each order made by the customers as index and the products bought as one hot encoded.

2.4.2. Modeling

To create the association rules and check the association between products, the Apriori algorithm was used with a 5% support threshold to get the list of items in at least 5% of the transactions. The frequent itemset list was then used to create the rules by setting confidence and lift thresholds. The team also explored frequent itemsets with different lengths, and found out which products were complementary, and which were substitutes by tweaking the threshold of the different metrics. The same analysis was also done with the support of 2.5%.

2.4.3. Results Evaluation

First, the team created a first set of association rules with 60% confidence as a threshold to find complementary products. As a result, 566 rules were created. Some interesting findings can be found in the next table.

antecedents	consequents	confidence	lift
fresh fruits, fresh herbs	fresh vegetables	**0.88**	1.98
yogurt, packaged vegetables fruits, fresh vegetables	fresh fruits	**0.87**	1.56
milk, fresh vegetables, packaged vegetables fruits	fresh fruits	**0.86**	1.54
fresh herbs	fresh vegetables	**0.85**	1.90
fresh vegetables, packaged cheese, packaged vegetables fruits	fresh fruits	**0.83**	1.50
yogurt, packaged vegetables fruits	fresh fruits	**0.83**	1.49
soy lactosefree, packaged vegetables fruits	fresh fruits	**0.82**	1.48
yogurt, fresh vegetables	fresh fruits	**0.82**	1.47
milk, packaged vegetables fruits	fresh fruits	**0.81**	1.46
packaged vegetables fruits, frozen produce	fresh fruits	**0.81**	1.46

Table 1 - Complementary products with support higher than 5%

On these 10 rules the consequent is always either *fresh vegetables* or *fresh fruits* and there are different antecedents. All those transactions appear at least in more than 5% of the dataset, and every antecedent has a positive association with *fresh vegetables*. In other words, when a customer has in the cart any of those antecedents, it is highly probable from them to add *fresh vegetables*, becoming then complementary products. Because of the high number of fresh fruits and vegetables, it is expected that they are, in the end, the most *consequent* items in the dataset.

Considering all the information on Table 1, it is better to focus on the most important antecedents and consequents so that it is possible to extend the amount of offers for those products. For example, if someone puts fresh fruits or fresh herbs in his online basket, it is highly likely that the same person is also interested in buying fresh vegetables. Since Instacart is an online store, the suggestion is for the website to automatically recommend fresh vegetables to the customer. This logic repeats to every other pair of antecedents and consequences highlighted on Table 1.

An analysis has been conducted to find which types of products could be seen as substitutes. Substitute products are the ones that have a lift of less than 1. With 5% support threshold that was previously used, there were not many rules obtained, therefore the threshold was lowered to 2.5%. The association rules for the substitute products can be seen in Table 2.

Antecedent	Consequent	Antecedent Support	Consequent Support	Support	Confidence	Lift
soft drinks	fresh vegetables	0.087	0.044	0.027	0.32	**0.72**
soft drinks	fresh fruits	0.087	0.56	0.04	0.45	**0.82**
paper goods	fresh fruits	0.063	0.56	0.03	0.50	**0.90**
candy chocolate	fresh fruits	0.069	0.44	0.026	0.40	**0.91**
seltzer sparkling water	milk	0.19	0.24	0.046	0.24	**0.99**

Table 2 – Substitute products

Each of the rules on the table has a lift smaller than 1, meaning that the products are not likely to be sold together. For example, a customer that adds to the cart *soft drinks* is very unlikely to buy *fresh vegetables* or *fresh fruits*. This rule does not take into consideration the roles of antecedent and consequent, so the opposite works as well. If a customer adds to the cart *fresh vegetables*, it is unlikely that he will buy *soft drinks*. Below in figure 10 another insight related to substitutes products, done with a chord graph.

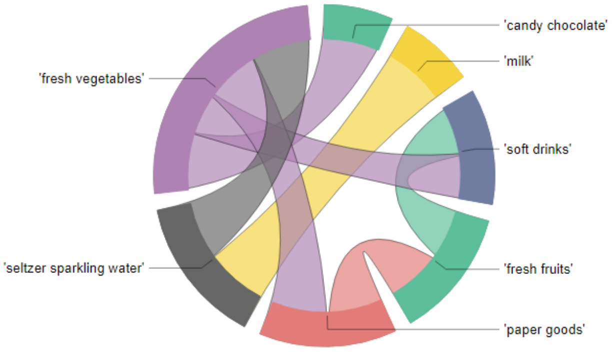

Figure 3 – Substitutes products (lift less than 1)

Next, the team searched for more complementary products with the threshold of 2.5%, as in the previous analysis in table 1 the consequents were all *fresh vegetables and fresh fruits*. Even changing the confidence bigger than 0.6 and lift bigger than 1.6, the results didn't change much related to table 1. This table is available in the Jupiter notebook in the session "Complementary products".

The team also did an analysis related to products with lift more than one: these are products that are more likely to be bought together than alone. Some interesting insights can be seen in the next table.

Antecedents	Consequents	Support	Confidence	Lift
chips pretzels	fresh dips tapenades	0.0352	0.20	**2.11**
lunch meat	bread	0.03	0.32	**1.97**
other creams cheeses	packaged cheese	0.03	0.42	**1.82**
yogurt, packaged cheese	milk	0.03	0.43	**1.78**
crackers	packaged cheese	0.04	0.38	**1.65**
hot dogs bacon sausage	packaged cheese	0.03	0.37	**1,.6**
lunch meat	chips pretzels	0.02	0.27	**1.59**
butter	packaged cheese	0.02	0.36	**1.56**

Table 3 – Products that are bought together

This analysis shows some habits related to the customers: for example, people that buy chips pretzels buy also fresh dips tapenades, or people that buy lunch meat will also buy bread. This insight is useful for the company in order to do some promotions about products bundling.

3. DEPLOYMENT AND MAINTENANCE PLANS

3.1. DEPLOYMENT

For the deployment of both the clusters, association rules and the Power BI dashboard few steps are needed.

1. Make sure that the data is readily available for the analysts to run Python and the other programs;
2. Do company-wide and leaders-wide meetings to show the results of the analysis and start understanding where it can be used next. For example here below some ideas on how to use it from now on:
    a. The rules can be used to choose the page layout and product display on the website;
    b. As a recommendation system once a customer is choosing a product;
    c. To do promotions on specific groups of products to increase both their sales and the complementary ones;
    d. The power BI dashboard can be broadly used for seasonal and end of year meetings, reporting and to have a better understanding of the business and how the orders are placed. It is an easy tool that anyone in the company could use;
    e. To start forecasting the demand of products based on their complementary or substitute.

3.2. MAINTENANCE

Maintenance wise it is important that the update of the analysis is scheduled by the business. To have a better understanding of different periods of the year the suggestion is to divide analysis based on events such has religious, national events and, as well, season wise updates. This can help to both check cross events how the behavior of customers differs and, on the other hand, checking how it changed based on new actions the business took; such as bundle promotions, different website designs and displays ect.

4. CONCLUSIONS

With this business case for Instacart it was possible to deep dive in the data and come up not only with insights, but also with information that can be used to act right away.

Starting from understanding the general behavior of customers: buying more on Sundays and on Monday; preferring both to buy before 7 days from the current order or after 30 days. Then the focus shifted on clustering the orders in different groups finding out that not only there is a clear group of orders that are made in the morning or in the evening, but also that there is a significant group of orders which prefer bigger amount of products while also the "monthly orders" cluster which contain all those orders that are made after at least 20 days, with 55% of the orders made on the 30th day.

Finally, understanding the association rules, how to use the information to improve the current way the business takes decisions on promotions, how to tackle groups of products, understanding as well that not only products can be complementary or substitutes, but can clearly define different types of customers and lifestyles.

4.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Regarding what can be done better in the future there are a few points worth mentioning:

- When calculating the clusters, some orders were the first ones ever for that specific customer, leaving them without a value for the variable "days_since_prior_order". To obviate to the problem, the mean of the population was imputed to these lines. For future improvements, the ideal strategy would be to check the mean for a certain group of users and imputing it based on that;
- Having access to the whole dataset and not just a sample of the orders. As well, having the date of each order and not only the time references as day or time of the day;
- Adding to the Power BI dashboard as well all the association rules so that the team can navigate them freely.

5. REFERENCES

https://pbpython.com/market-basket-analysis.html
https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/
https://towardsdatascience.com/association-rules-2-aa9a77241654
https://www.solver.com/xlminer/help/association-rules
https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/
http://mattimeyer.github.io/2016-12-21-Substitution-Rule-Mining/
https://towardsdatascience.com/instacart-market-basket-analysis-part-1-which-grocery-items-are-popular-61cadbb401c8
https://towardsdatascience.com/instacart-market-basket-analysis-part-2-which-groups-of-customers-are-similar-618e88b0866d
https://towardsdatascience.com/instacart-market-basket-analysis-part-3-which-sets-of-products-should-be-recommended-to-shoppers-9651751d3cd3