BUSINESS CASES WITH DATA SCIENCE

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

Business Case 3 – Instacart's Business

Group H

Hiromi Nakashima, m20201025

Manuel A. F. Carreiras, m20200500

Luis F. R. Agottani, m20200621

Venâncio Munhangane, m20200579

April, 2021

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

TABLE OF CONTENTS

1. Introduction

In this project we were designated to analyze the Instacart's Business through market basket analysis.

Instacart is an online grocery delivery and pick-up service in the United States and Canada. The aim difference in that business is with the customer experience, Instacart intend to facilitate the "do your groceries" through an app to create a personal list of shopping followed by the delivery in the customer place.

This report will describe the steps of the Cross-Industry Standard Process for Data Mining (CRISP-DM) and the marketing basket analysis of the mentioned business.

The GitHub repository where all the present analysis is saved can be accessed through the following link: https://github.com/hnakashima96/Business-Case-.

2. Business Understanding

At this stage we defined the essential business guidelines to grant a good result of the project and to develop the best solution for Instacart's Business. Currently, the business has a basic knowledge of the user and products correlation based in actual numbers from the cart. In order to improve the results of the customer, we will describe the business objectives and how those objectives will be converting to data mining steps on the following topics.

2.1. Business Objectives

The goals of Instacart's Business are:

- Create customers profiles of shopping.
- Identify correlation/patterns of shopping between the products.
- Identify substitutes/complementary products.

2.2. Business Success criteria

Based on the business objectives description, a main result was defined to guarantee the success of this project: provide a understanding of the basket based on the products.

2.3. Determine Data Mining view

The business goals guide the Data Mining goals as shown in the table below (Table 1). Also, we defined the success criteria for these goals to support the confidence of our results.

Table 1. Data Mining Goal s and Success Criteria.

Business Goals	Data Mining Goals
Create customers profiles of shopping	Clustering the carts to identify shopping patterns.
Identify correlation/patterns of shopping between the products.	Develop a market basket analysis
Identify substitutes/complementary products	Measure association rules metrics

3. Data undertanding

The Instacart provided the business information through four different datasets: orders, order_products, products and departments.

- Orders: contains details about orders and time history (day of the week, hour of the day, number of days since last order) for each user/id. This dataset contains in total 200.000 orders done for 105.273 customers.
- Order_products: contains information about the products and its quantities bought and reordered for each customer.
- Products: Products table contains 133 different products of the grocery store. This dataset connects products to a certain a department id.
- Department: aggregates department's names and id of the grocery store. The store contains 20 different departments.

Now we are going to describe the principal activities in order to obtain the final dataset. All the datasets were merged using the primary key and was found out that there are 4749 observation labelled with missing on products and on department. These cases represented 0.23% of all the cases and was considered missing cases and excluded.

In order to improve the data understanding we developed some data visualizations. After this process we were able to highlight three main schemas:

a) Number of orders by time Period

Understand the customer behavior, specifically in a time point of view can collaborate to find real opportunities for the business.

The figure 1 illustrates the number of orders through the day of the week and by time period (Dawn, morning, afternoon and night). Most orders to happen on Sundays and Mondays. On Thursdays and Wednesdays, the grocery has the lowest number of orders. In terms of period there is a tendency for most orders to happen during the afternoon and morning (reaching almost 40k). On the opposite and as expected there are less orders during night and dawns.

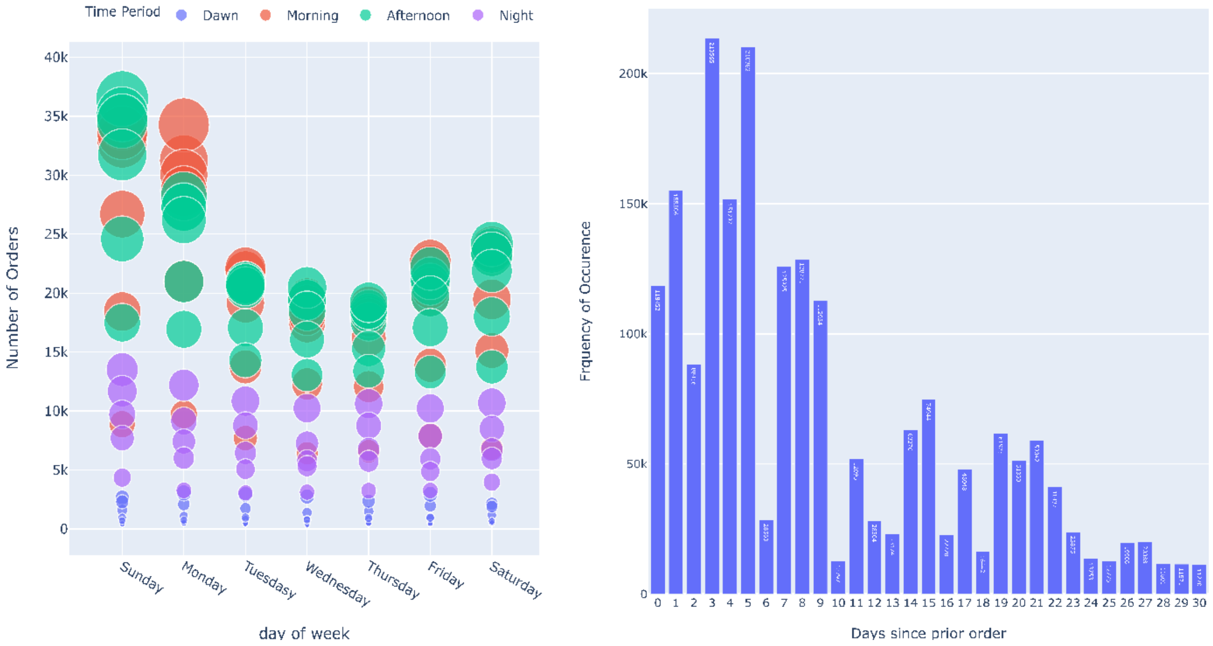

Figure 1: Numbers of orders throughout Period of Time and Frequency of days since Prior Order.

b) Quantity of sales on each department by time Period Information

After understanding the orders it is also important to analyse the quantity of products sold by each department throughout time. This information can be analysed on the graphics bellow, the only difference remains on the time discriminated. On the left it concerns the days of the week and on the right, period within a full day of work.

When it comes to evaluate the left bubble chart it is possible to conclude that for **all departments the highest quantity of products is sold on Sunday except for beverages which tend occur on Monday**. On opposite, **Thursday is verified the lowest quantity of products sold for all departments**. **Produce department has more quantity of products** sold. Sunday this department reaches it's peak and sells more than 120k products, almost doubling the second-best department (dairy goods), which follows the same tendencies.

In terms sales throughout the period of the day, according to the right bubble chart it is possible to verify that in general most departments sell **more quantity of products during the mornings and afternoons.**

As expected, these charts follow the same tendencies of the previous one presented on point a).

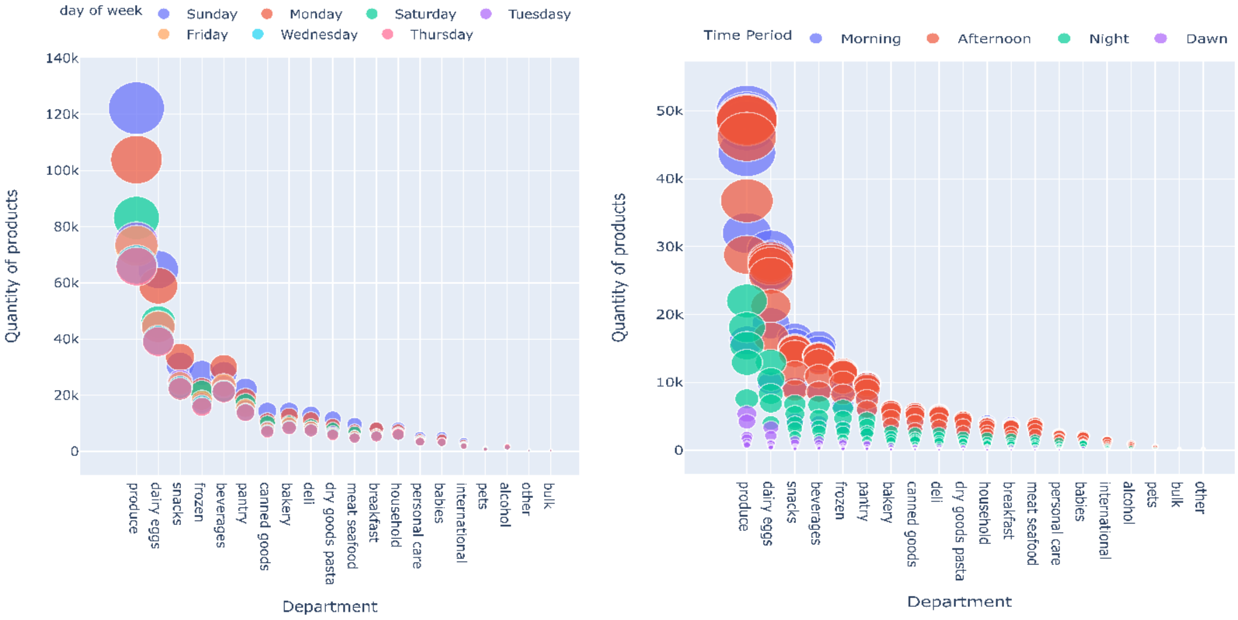

Figure 2: Quantity of sale for each department throughout time period

c) Reorder ratio and Quantity of sales on products

Subsequently it is presented a few insights about product reorders and products shopping trends considered important for a good understanding of the grocery Business. Firstly, at figure 6 it is observed the frequency of the number of products bought on each order. It is possible to view that any client **shops just one product per order and there is an inclination for clients to shop 5 products, being this the most frequent number of combinations of products per order**, followed by 6 and 4 respectively.

At figure 7 it is illustrated the reorder ratio by each department. The information provided shows that most of the clients tends to shop the same products in many of the departments. More than 50% of sales made on 13 departments are referent to products bought before. There are only 7 departments with a reorder ratio below of 50%. **This indicates clients tend to always buy the same products, revealing a behavior to do so.**

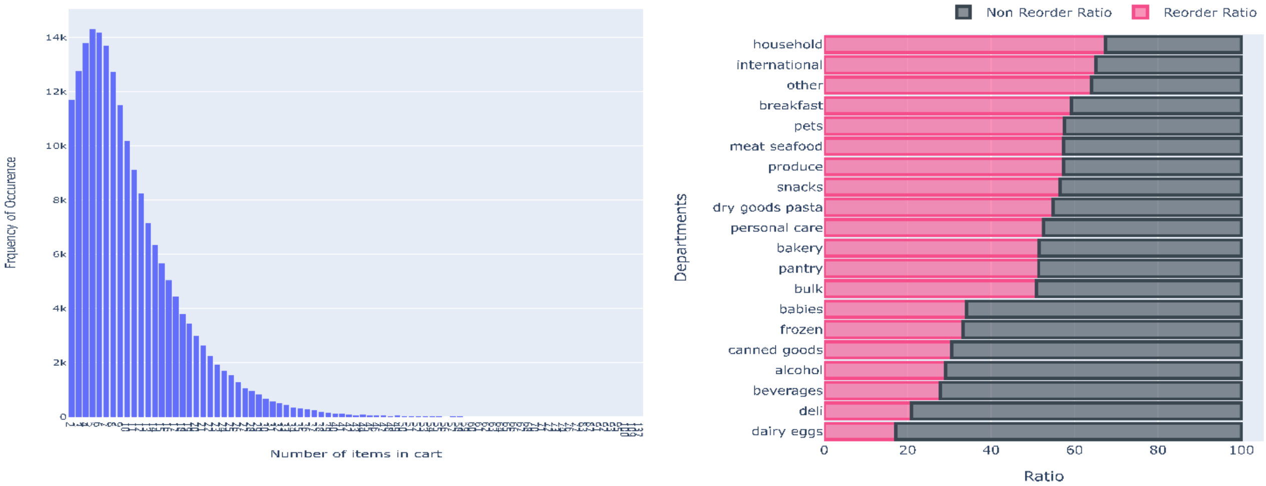

Figure 3: Frequency of nº of items in cart and Reorder Ratio vs. Non Reorder Ratio.

To finish it is presented a scatterplot regarding the number of shops by products and the reorder ratio by each product. Most products have **less than 50k shops**, so we considered this value as an important threshold for this dimension. On top of that it was also measured that most products have a **reorder ratio above 0,5**, being this the other threshold. With this evidence this chart has a main insight: a division, allowing the **distinguishment of 7 products** highlighted on the left graphic. Among this 7 products Fresh Fruits has the largest number of shops (above 226k) and most clients bought milk frequentl y (near 0.8 reorder ratio).

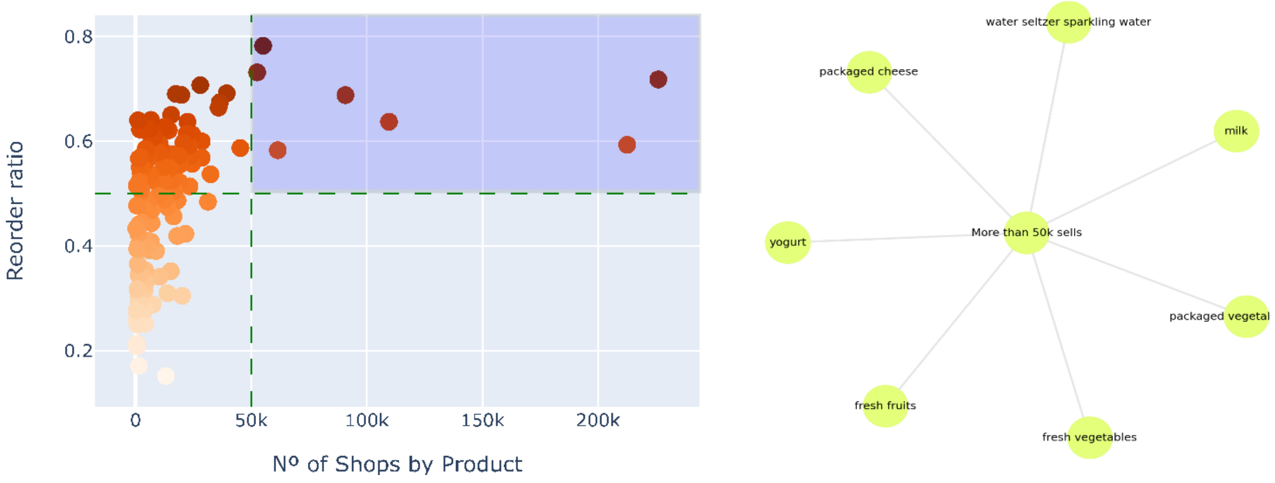

Figure 4: Reorder Ratio Vs. Number of shops by product and products with more than 50k sells

4. Clustering Process

4.1. Data Preparation

In order to cluster the data, we had to filter the user_id from the orders table and departments from the products table. A cross table was created to identify the quantity of each product that each user bought.

4.2. Methodology

It is a clustering algorithm used in scientific and industrial applications. The algorithm begins with the choose of all initial centres uniformly and randomly, from the data points. In this report was used for customer segmentation. Which the aim was to divide customers into groups based on which products they buy. Because there are 133 of products and thousands of customers and K-Means does not produce good results on higher dimensions, was utilized the departments which represents categories of products and all features were rescaled on a range in [0,1]. To choose the optimal number of clusters was used *Silhouette score* and was considered two numbers 4 and 5, but a solution with 5 cluster was considered better and was presented how the cluster are distributed on a 2-dimensional space using t-Distributed Stochastic Neighbour Embedding (t-SNE).

4.3. Clustering Result

As mentioned, the optimal result is a solution with 5 clusters. Cluster 1 results into 72832 consumers, cluster 2 results into 4500, Cluster 3 results into 23384, Cluster 4 results into 4548 and Cluster 5 results into 3615. The table 2 presents the top 10 average products purchased in each cluster.

In all the clusters, customers show interest on the top seven products and specially for fruits and vegetables. The first three can be discriminated by specific products and the last two by quantity of purchase. The cluster 1 has a high interest in yogurt and sparkling water. While the cluster 2, is characterized by high interest on dairy eggs products. The cluster 3 is differentiated by the interest in more industrialized products like hot dog and chips.

To conclude the cluster 4 was identified with low quantity of purchase. And the cluster 5, with high quantity of purchase.

Table 2. Top 10 average products purchased in each cluster.

Cluster 1	**Avg. Pur.**	**Cluster 2**	**Avg. Pur.**
fresh fruits	2,31	fresh vegetables	3,77
fresh vegetables	1,82	fresh fruits	3,69
yogurt	1,23	packaged vegetables fruits	1,82
packaged vegetables fruits	1,21	yogurt	1,51
water seltzer sparkling water	1,01	packaged cheese	1,01
paper goods	0,87	milk	0,85
packaged cheese	0,74	chips pretzels	0,68
chips pretzels	0,74	water seltzer sparkling water	0,65
refrigerated	0,68	soy lactosefree	0,63
milk	0,66	frozen produce	0,59
Cluster 3	**Avg. Pur.**	**Cluster 4**	**Avg. Pur.**
fresh vegetables	4,01	fresh fruits	1,21
fresh fruits	3,56	fresh vegetables	0,97
packaged vegetables fruits	1,87	packaged vegetables fruits	0,55
yogurt	1,28	yogurt	0,47
hot dogs bacon sausage	1,13	water seltzer sparkling water	0,37
packaged cheese	1,09	milk	0,32
milk	0,84	packaged cheese	0,29
water seltzer sparkling water	0,64	chips pretzels	0,26
chips pretzels	0,64	soy lactosefree	0,22
bread	0,56	refrigerated	0,21

Cluster 5	Avg. Pur.
fresh vegetables	9,18
fresh fruits	8,90
packaged vegetables fruits	4,42
yogurt	3,63
packaged cheese	2,82
milk	2,02
chips pretzels	1,72
bread	1,55
frozen produce	1,52
soy lactosefree	1,45

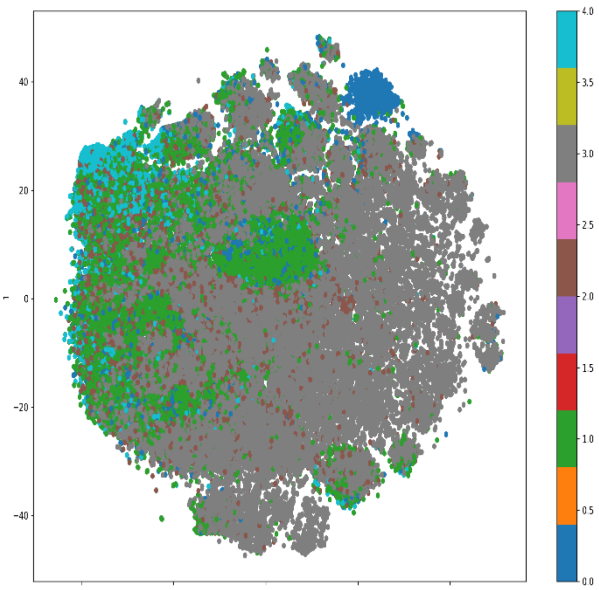

Figure 5: t-SNE representation of the clusters.

5. Marketing basket Analysis

According with [1] Market basket analysis has the objective of identifying products or group of products which tend to occur together (are associated) in buying transactions (Baskets). The insights extracted from this method can be useful. [1] Says that these insights can be employed by a supermarket to reorganize its layout, taking products frequently sold together and locating them in close proximity. But it can also be used to improve the efficiency of a promotional campaign:

Products that are associated should no be put on promotion at the same time. By promoting just one of the associated products, should help to increase the sales of the other and get accompanying sales increases for the associated products[1].

5.1. Data Preparation

In order to model the theory "*if one client buys a product or a certain group of items, he is more or less likely to buy another product or group of items*" and get the insights explained above was used the association rule mining based on the following metrics:

i. Support: Shows how frequently the combination occurs in the database;

ii. Confidence: It is the ratio of the number of transactions involving both A and B and the number of transactions involving B;

iii. Lift: Equal to the confidence factor divided by the expected confidence. Lift is a factor by which the likelihood of consequent increases given an antecedent.

With this metrics was possible to evaluate the combinations of the products orders and the number of times these combinations are repeated. The possible combination that is repeated many times and has higher confidence to occur is the potential one to be a rule for the theory "*if the condition, then the result*".

In this Business case was applied the basket analysis following the steps bellow:

1. Was applied a priori algorithm and the goal was to find all the possible subsets of items with popularity or probability to occur higher or equal to 1% (minim support 0,001. This threshold resulted on 178277 combinations.

2. On all the 1782277 combinations, association rule was applied and resulted in 4787804 possible rules. These rules were dived in two groups:
    a. First group to identify complementary products (was selected all the rules with lift higher than 1) with 4781132 rules;
    b. The second group to identify substitutes products (was selected all the rules with lift low than 1) with 6672 rules.

5.2. Approach to complementary products

Taking on account the number of rules first was decided to keep rules with confidence higher than the median (threshold: median of confidence higher than 0.0716) reducing the number of rules to 2390566. Using these rules was found out that there are more chances to have as antecedents a combination of 3 products and as a consequent a combination of two products (see the fig 66). Due to these results we decided to focus our analysis on the rules that have as antecedents a combination of 3 products.

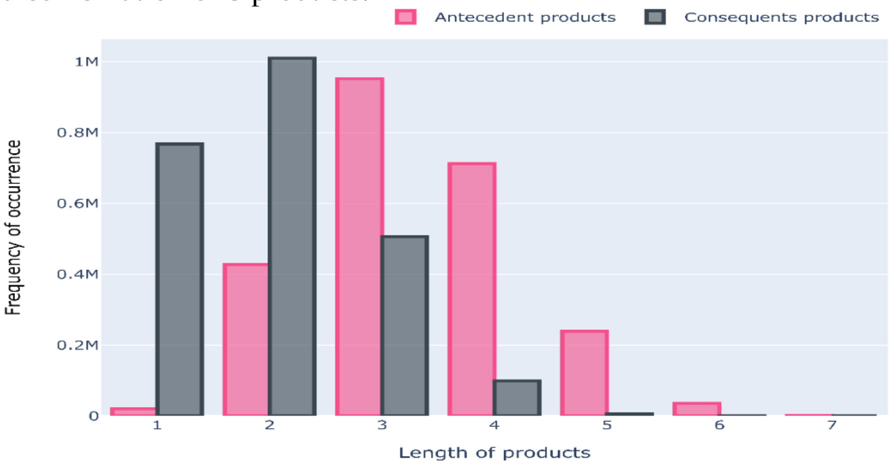

Figure 6: *Frequency of possible combinations by consequents and antecedents*

The step above resulted on 952548 rules and among these rules were selected those with lift higher than 1.8 and confidence higher or equal than 0.5. These resulted on 47380 rules.

5.3. Approach to substitute products

For the substitute products was followed the same approach:

i. Select the rules with confidence higher than the median (median confidence 0.047)
ii. Was verified that is better to analyse substitute products with a combination of 1 product as antecedent and consequent These resulted on 668 rules.
iii. Among the rules resulted above was selected those with lift low than 0.9 and confidence higher or equal than 0.2. These resulted on 76 rules.

The analysis of approaches described above are shown on the chapter X but is going to be presented only the top 10 rules for complementary products and the top 20 for substitutes products.

5.4. Results

After applying all the steps described on the points 5.1, 5.2, 5.3 it is going to be presented the main results in the obtain on the association with the analyse basket analyses for complementary and substitute products.

5.4.1. Complementary products

After applying all the thresholds was found a set of 6672 rules for complementary products when the antecedents of the products is a combination of three products. The figure above shows the top ten rules. It is possible to verify that the group of products contains fresh herbs, packaged cheese and baby food formula are complementary with the group of fresh fruits, fresh vegetables and yogurt are. The result of lift was 4.45 and the confidence was 0.52. This indicates that every time that the customer goes to buy the first group is four (4) times more likely to add on the cart the second group.

The best rule found on the complementary products was for the antecedent of spices seasonings, canned meals beans and soup broth bouillon with canned jarred vegetables. The lift was 7.1 and the confidence 0.51 meaning that every time that the customer goes to buy the first group is 7 times more likely to add on the cart canned jarred vegetables.

Other results of the complementary can be analysed on the notebook on the line 110.

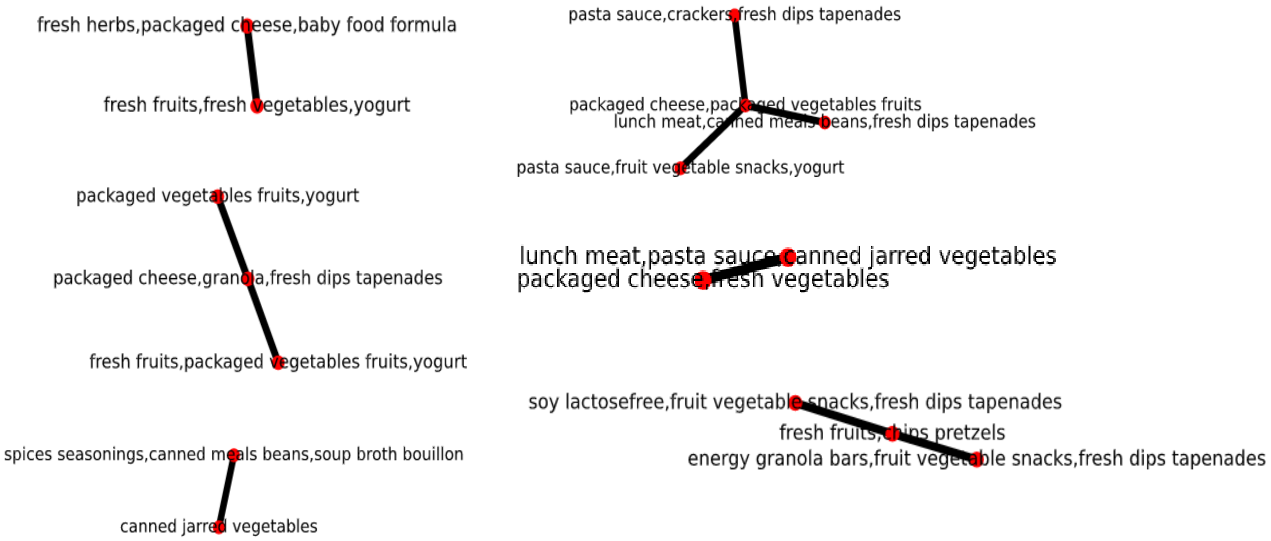
fresh herbs,packaged cheese,baby food formula
fresh fruits,fresh vegetables,yogurt
packaged vegetables fruits,yogurt
packaged cheese,granola,fresh dips tapenades
fresh fruits,packaged vegetables fruits,yogurt
spices seasonings,canned meals beans,soup broth bouillon
canned jarred vegetables
pasta sauce,crackers,fresh dips tapenades
packaged cheese,packaged vegetables fruits
lunch meat,canned meals beans,fresh dips tapenades
pasta sauce,fruit vegetable snacks,yogurt
lunch meat,pasta sauce,canned jarred vegetables
packaged cheese,fresh vegetables
soy lactosefree,fruit vegetable snacks,fresh dips tapenades
fresh fruits,chips pretzels
energy granola bars,fruit vegetable snacks,fresh dips tapenades

Figure 7: Top 10 rules for complementary products.

5.4.2. Substitute products

After applying all the thresholds was found a set of 79 rules for substitute products. With the substitute products is possible to conclude that a client that tends to buy a product won't by it's consequent. The figure above shows the top twenty rules. It is possible to verify fresh vegetables is a substitute products of energy sport drinks, cold flu allergy, trail mix snack mix, spirits, body lotion soup and mint gum. Meaning that if a client buys fresh vegetables, he is more likely to not buy one of the products mentioned. Other case is fresh fruits which has seven substitutes' products namely: Eye ear care, facial care, beers coolers, spirits, mint gum, red wines, white wines and body loations soup.

Other results of the substitute products can be analysed on the notebook on the line 132.

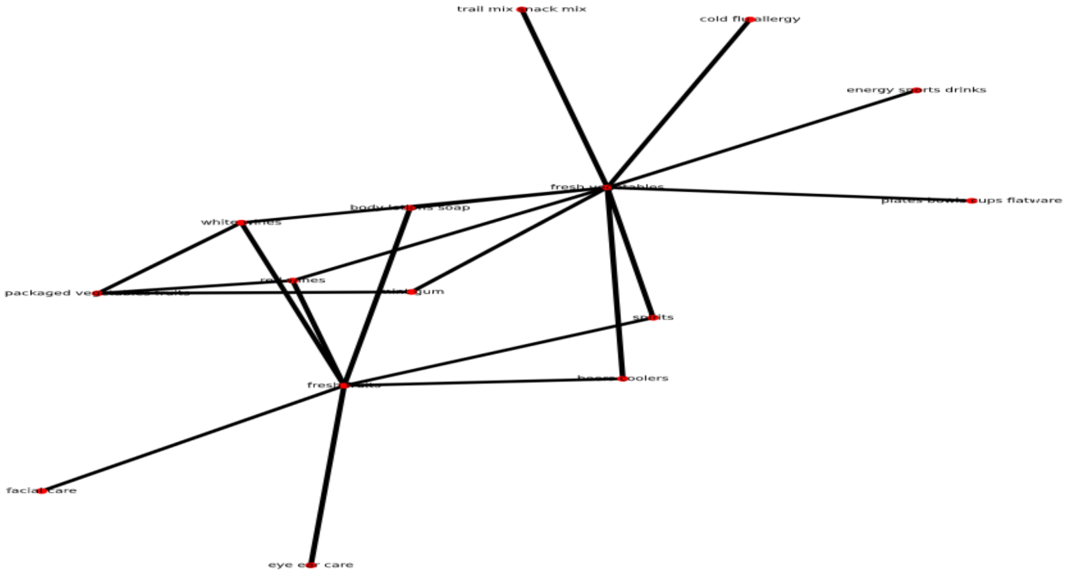

Figure 8: Top 20 Rules for substitute products.

5.4.3. Frequency of the top 7 products on the rules

The other insight that is going to be provided is the frequency of the occurrence of the seven products on the final set of the rules. With the graphic above is possible to verify that these seven products have frequency of occurrence in most of the rules.

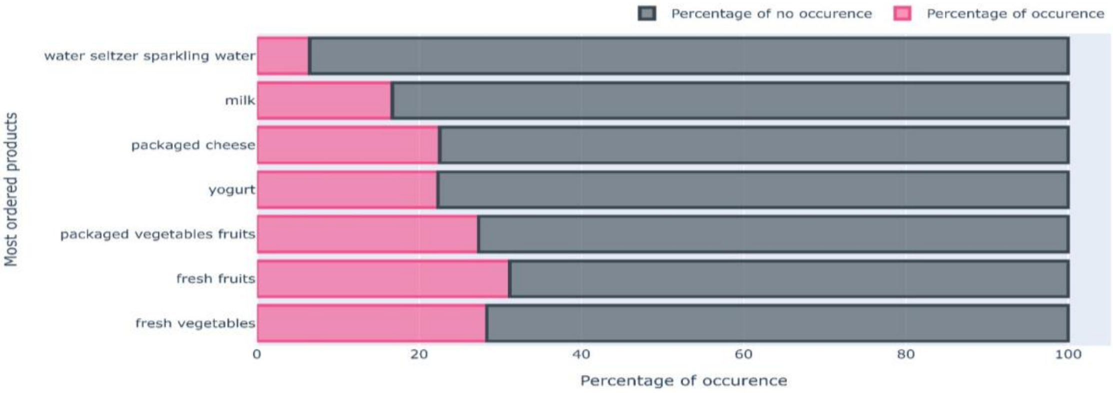

Figure 9: Percentage of occurrence of the top 7 products on the rules

6. Conclusions

Based on the business objectives description was possible to conclude the following:

- All the clusters show a high interest in fruits and vegetables.
- Based on the Market Basket Analysis was concluded that the customers look for Instacart orders with 3 products (antecedents) and 2 products as a consequent.
- Most of the customer prefer to buy 5 products per order;
- The 7 most buying products should be the focus on sales to provide the increase sales of the associated products.
- With the threshold defined was possible to find 79 rules for substitute products;
- With the threshold defined was possible to find 6672 rules for complementary products.

7. Deployment

At this stage we are going to describe some points of how the company could use the results to increase the sells:

- Products that are complementary should not be promoted at the same time. By promoting just one of the associated products, should help to increase the sales of the other. In this case it is not advisable that for instance eggs, specialty cheeses, fresh herbs (antecedents) should be promoted at the same time with fresh vegetables (consequent).
- As for the substitutes according to our rules for the top20 substitutes it is possible to conclude that a client that tends to buy a product won't by it's consequent, for example if a client that buys mint gum(antecedent) will not buy fresh vegetables (consequent) with a confidence of 20%.

8. Future work

On the basket analyses is possible to have a huge number of possible rules. This report was focused on the combination of tree products as antecedents and was selected rules with lift higher than 1.8 and confidence higher or equal than 0.5 for complementary products. One other and for substitute products was focused on 1 product as antecedent and complementary, lift low than 0.9 and confidence higher or equal than 0.5 for substitute products.

For future work can be analysed rules for other number of products as antecedents also other rules.

REFERENCE

Pete Chapman, J. C. (1999). CRISP-DM 1.0. *Step-by-step data mining guide.*

[1] Trnka, A. (2010, June). Market basket analysis with data mining methods. In *2010 International Conference on Networking and Information Technology* (pp. 446-450). IEEE.