NOVA IMS Information Management School

Business Case #3

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

Instacart Market Basket Analysis

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

Index

INTRODUCTION

In the retail industry, especially in online services, is crucial to understand the purchasing patterns/behaviors of consumers and identify different relationships between products, providing a holistic view about its customers and the portfolio of products it has to offer. Considering this kind of information and the expected outcomes required by the Instacart district manager, the best approach adopted was the implementation market basket analysis. Market basket analysis is a data mining technique that is used to inform a retailer about products that are frequently bought together, which of these are seen as substitutes to one another, and customer segmentation.
After the application of this analysis, it would be easier for the company to understand the purchasing patterns and construct relationships between products.

BUSINESS UNDERSTANDING

BACKGROUND

Instacart is an American company that provides a grocery delivery and pick-up service via a website or mobile app in the United States and Canada. This company uses transactional data to get some insights into their customers. Our consultant group was contracted to help in order to take full advantage of the data that they have available.
The dataset available is a relational set of files describing customers' orders over time. This dataset contains 200, 000 grocery orders from more than 100, 000 users. For each user, is provided a few of their orders, with the sequence of products purchased in each order. The products are grouped by their types, containing 134 generalized products. We also have information about the week and hour of the day the order was placed and a relative measure of time between orders.

BUSINESS OBJECTIVES

The objective of this business passes by giving important insights to the company, providing purchasing patterns of the costumers with the Market Basket Analysis (MBA). The data science team should aim to provide an overview of Instacart's business as complete as possible.
With the provided dataset, the team could come up with data analysis for the below-mentioned business problems:

- Optimize the store's layout
- Develop cross-promotional programs and customized marketing strategies for its customers with a deep understanding of the consumer behavior

Answering the topics below:

1. What are the main types of consumer behavior in the business?
2. Which types of products should have an extended amount of product offerings?
3. Which types of products can be seen as substitutes?
4. Which items are complementary?

BUSINESS SUCCESS CRITERIA

There is a natural need to understand and especially to find out what would be the substitute products and the complementary products. Thus, for the complimentary products, it was defined as a rule, the presentation of all products (associated with each other) that had a confidence greater than or equal to 0.8 and a lift greater than or equal to 1.8. For substitute products, the lift parameter was considered the most relevant, indicating a list of products in which the lower the value of this parameter, the higher the substitute product relationship was between them.

DETERMINE DATA MINING GOALS

One of the key techniques used by large retailers is called Market Basket Analysis (MBA), which uncovers associations between products by looking for combinations of products that frequently co-occur in transactions. In other words, it allows supermarkets to identify relationships between the products that people buy. The discovery of these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. To do this, the data science team have done the following steps:

- Linking of the provided datasets, to cross-check relevant information
- Analyze the anonymized data of 3 million grocery orders from more than 200,000 Instacart users open-sourced by Instacart
- Find out the hidden association between products for better cross-selling and upselling
- Perform customer segmentation for targeted marketing and anticipate customer behavior

The data mining goals defined to answer the business objectives were:

- Customer's data behavior analysis
- Customer segmentation with K-Means
- Market Basket Analysis (MBA) with Apriori algorithm

ANALYTICS PROCESS

DATA UNDERSTANDING

All the information presented consists of the analysis of four datasets that were provided.

First, the dataset *"order_products"* is made up of the variables: *"order_id"* which indicates the number of the order placed by the customer; *"product_id"* which indicates

which products were purchased by the customer; *"add_to_cart_order"* which indicates the sequence in which the products were placed in the cart; *"reordered"* which indicates whether the products are reordered or not (0 isn't reordered and 1 is reordered).
Next, the "*products"* dataset is made up of the variables: *"product_id"*; *"department_id"* which corresponds to the numeric identification per department; *"product_name"* which corresponds to the name of each product.
The *"departments"* dataset is made up of the following variables: *"department_id"*; *"department"* which corresponds to the name of the department.
Finally, the *"orders"* dataset has the following variables: *"order_id"*; *"user_id"* which corresponds to the direct identification of each customer who has made at least one purchase at the store; *"order_number"* which corresponds to the number of the order made by each customer; *"order_dow"* which corresponds to the day of the week when the order was made; *"order_hour_of_day"* which corresponds to the time of day when the order was made; *"days_since_prior_order"* which corresponds to the number of days between making an order and placing a new order.
Regarding *order_products* data, we realized that includes a total of 134 products and 20000 orders.

DATA PREPARATION

To prepare the analysis, it was necessary to perform a merge between the different datasets, where their link consisted of the corresponding ids of each dataset. For the count of products by the user was necessary to proceed with a merged between *"orders"* and *"order_products"* and execute a group by to get the sum of order size of each user. This merged happened also to discover the number of products by of the week where we made also a group by between *"order_dow"* and *"order_size"* with sum. The exact same thing occurred to obtain the number of products purchased per hour. To extract the most important departments by the number of products was crucial to merge *"products"* and *"departments"* datasets, creating an *"items"* dataset, in order to obtain the total number of products per department through sum group by. After this, to get the best-selling departments by number or orders was decided to merge *"orders"* and *"order_products"* and then merge this result with the *"items"* dataset.

In "Which types of products should have an extended amount of product offerings?" the topic was necessary to proceed with also a merged in-between *"products"* and *"order_products"* to obtain the top 10 best-selling product and the most ordered products. This procedure was made for most reordered products and most popular products too.

Regarding data cleaning and manipulation, there were no missing values for the datasets like *"order products"*, *"departments"* and *"products"* datasets. *"Orders"* dataset has some null values in the *"days_since_prior_order"* variable and only 6.5% of the values were found to be missing and this has been rejected since the count is very low to be a significant issue.

DATA ANALYSIS

Following the data preparation, a brief analysis of the dataset is handled in order to understand the customer behavior, to have an overview of Instacart's business as complete as possible, and answer the case study topics.

The datasets gave us important insights about the customer to take into account in the previous sections. They are the following:

- According to Figure 1 and Figure 2, the distribution of the number of orders and products by the day of the week is similar – the majority of orders and products are held on the weekend, with a higher incidence on Saturday. Although, whereas the number of orders has a linear decrease on workdays, the purchase of the products tends to increase as of Tuesday to Saturday. The hour with higher incidence is between 10 and 15.

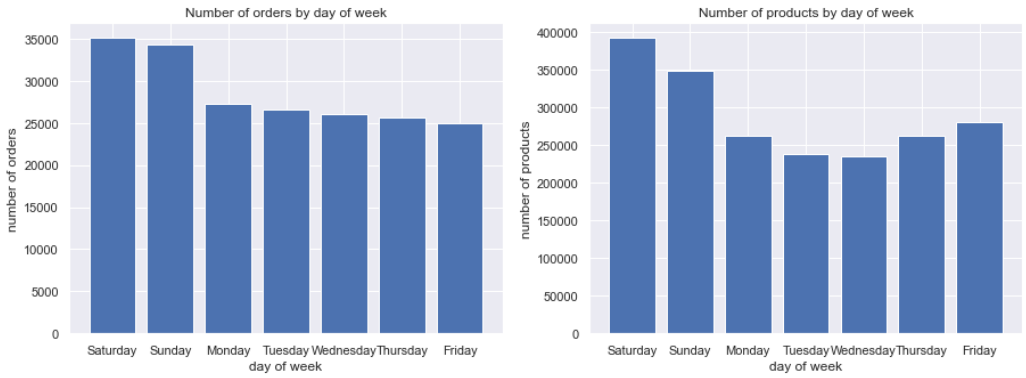

Figure 1 - Number of orders and products by day of the week.

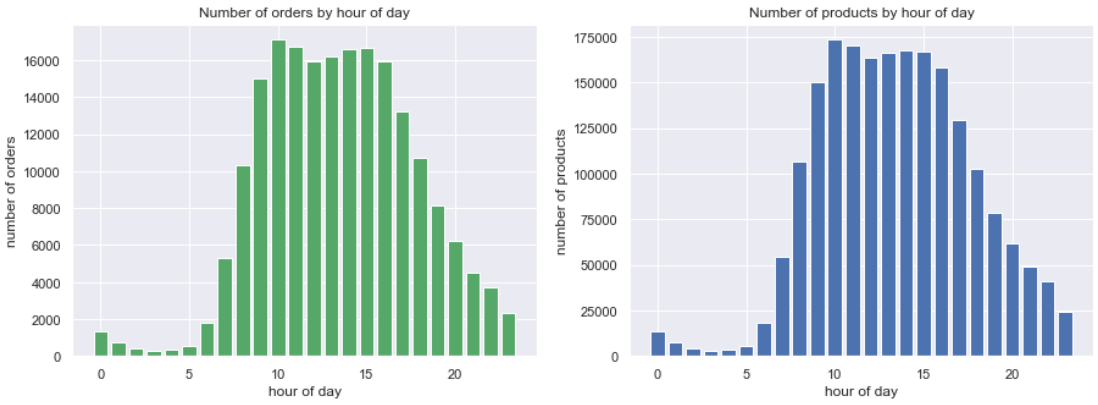

Figure 2 – Number of orders and products by the hour of the day.

- According to Figure 6, the majority of the reorders are made during the daytime with peaks on the 7th and 30th day of the month.

- As can be seen in Figure 3, the most popular products are basic needs in food, such as fruits, vegetables, yogurt, and milk. In general, the most reorder products match the most purchased products in store.

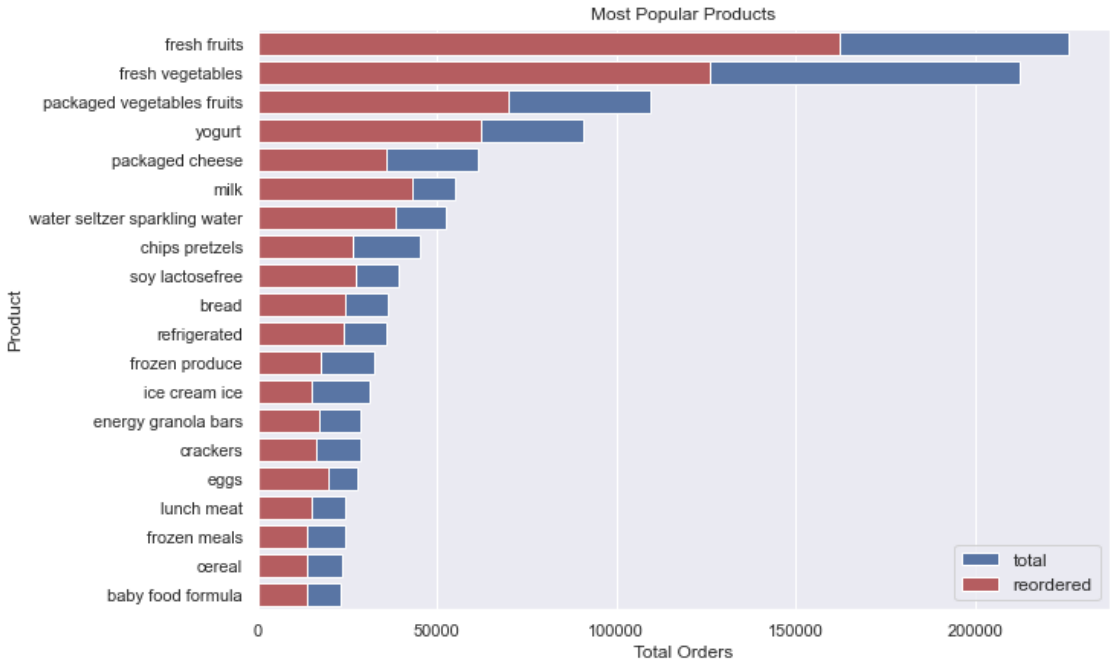

Figure 3 - Most popular products by reorders and the total.

MARKET BASKET ANALYSIS WITH APRIORI ALGORITHM

Market Basket Analysis (MBA) is a modeling technique based upon the theory that if you buy a certain group of items, you are more or less likely to buy another group of items. Market basket analysis may provide the retailer with information to understand the purchase behavior of a buyer.

The analytical approaches in the association rules of MBA are:

- Support
- Confidence
- Lift

The Apriori algorithm has been designed to operate on databases containing transactions, such as purchases by customers of a store. An item set is considered as "frequent" if it meets a user-specified support threshold. This analysis will be used in Topic 3 and 4 with the use of the association's rules.

CASE STUDY TOPICS:

1) What are the main types of consumer behavior in the business?

For an understanding of the consumer types that are presented in the datasets presented, it was necessary to perform a cluster analysis (displayed in Figure 4), with the focus of understanding what type of products are purchased by the consumer. Thus, there is a relationship between each customer individually and the products they purchase.

Four customer types were therefore generated, from a record of 200,000 orders for a total of 134 products. The first group, consisting of 24,940 customers, can be identified as having a fresh product consumption profile, with a clear preference for fresh vegetables followed by fresh fruits. The second group, consisting of 15,323 customers, has a profile that places a lot of importance on fruit in general, buying mostly fresh fruit. The third group, consisting of 62,389 customers, is not a particularly product-oriented group, buying a variety of products that the store sells, starting with fresh fruit, yogurt, or packaged vegetable fruits. Finally, regarding the fourth and last group, consisting of 2,621 customers, you have a fairly clear preference in purchasing seltzer sparkling water.

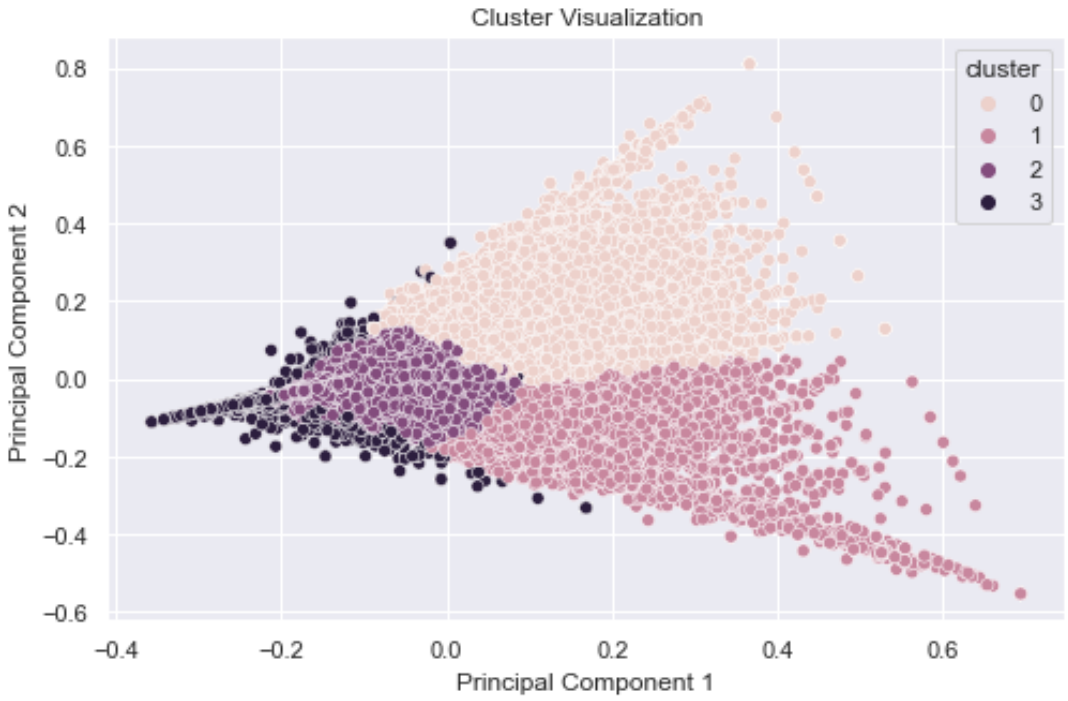

Figure 4 - Cluster visualization with PCA

2) Which types of products should have an extended amount of product offerings?

There are several factors to take into consideration to have an understanding of the products that are sold and especially a more efficient stock management. Thus, it is essential to calculate the products that are most wanted by the customer and to establish a direct relationship, in which these products are ordered again by the same

customer after the first purchase. This direct relation can be verified in Figure 3, presenting for example as products of extreme importance fresh fruits, fresh vegetables, or the packaged vegetable fruits.

Another relevant factor to take into account, and which will directly influence the most important products, is the number of items purchased for each purchase that is made. Once again it is clearly evident that there is a general preference for fresh products, such as fresh fruit or vegetables (Figure 11).

An analysis by the department was also performed in Figure 9, which shows which departments of the store are the most important considering the number of products that department contains, as well as Figure 8, which shows the departments with the best sales results considering the number of orders placed.

3) Which types of products can be seen as substitutes?

For a better interpretation of the results obtained regarding substitute products, it is necessary to understand the concept of the association rule, lift.

This association rule can be seen as the relationship between two types of products (A and B), it indicates the probability that sales of a consequent product (B) will increase knowing that the antecedent product was sold. Thus, carrying the concept to the discussion of this topic, there is a particular interest in understanding which products, when related to each other, have a minor lift as much as possible. For the effective analysis of the problem, consider products with a lift of less than 1, because this is the only way to indicate that there is no great relationship between the two products, leading even to their substitution.

This data can be seen in Figure 16 and from this visualization, it is possible to conclude that the product "fresh vegetables" relates in both directions to the product "water seltzer sparkling water", giving an indication that these products substitute each other.

4) Which items are complementary?

For the analysis of complementary products, it is not only extremely important to have a lift with a high value, shown that there is a direct relationship between the products, but it is also crucial that there is a reference to the concept of confidence. The concept of trust shows us directly the relationship between the product types in evidence, showing the strength of their association, that is, the percentage of a consequent product being sold knowing that the antecedent product has been sold.

Thus, all products with a confidence greater than 0.8 and a lift greater than 1.8 were selected (Figure 15). Only in 2 cases, this happens, however, it can be concluded that when there is a purchase of "fresh fruits" and "fresh herbs" there is a natural tendency to purchase the "fresh vegetables" product. This is also true when the product "fresh herbs" is purchased, leading to a natural tendency to purchase the product "fresh

vegetables". It can thus be said, that in these two cases, they are complementary products.

DEPLOYMENT AND MAINTENANCE PLANS

Based on the associative rule methodology and graphical results, some of the recommendations have been made:

- Personalized communications can be very lucrative by reminding the customers to reorder the products or can be added to the cart automatically based on the customer preferences.
- We would recommend Instacart to add the products directly to the customer's cart or to provide a suggestion list when they make their purchase in order to enhance the customer experience.

CONCLUSIONS

Comprehending the main types of consumer behavior, the products that should have an extended number of offerings and the relationships between the products are important to the company for being a good competitor in the market or even a leader market. Explaining these outcomes for the company managers will bring other views and knowledge in order to understand the customers and their behavior concerning certain types.
Regarding the questions answered, we can conclude the followed:

- Exist 4 types of customers that purchase products;
- Mostly fresh products, fresh vegetables, and fruits are the ones who should have an extended amount of offerings;
- Products "fresh vegetables" and "water seltzer sparkling water" are substitutes;
- "Fresh vegetables" is a complementary product of "fresh fruits" and "fresh herbs" and is also a complementary of "fresh herbs" only.

REFERENCES

(s.d.). Obtido de Frequent Itemsets via Apriori Algorithm: http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/

(s.d.). Obtido de Instacart Market Basket Analysis: https://github.com/archd3sai/Instacart-Market-Basket-Analysis

APPENDIX

Figure 5 - Number of Orders by order number.

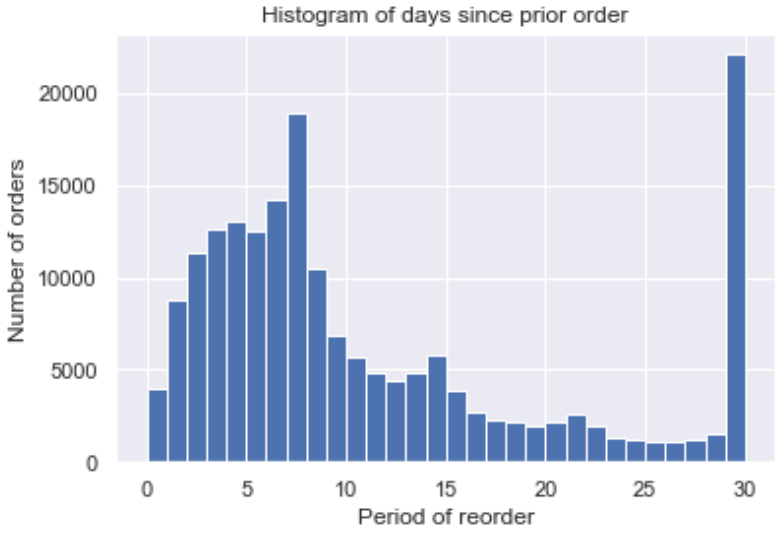

Figure 6 - Number of Orders with x days since the last purchase.

Figure 7 - Count of Order Size.

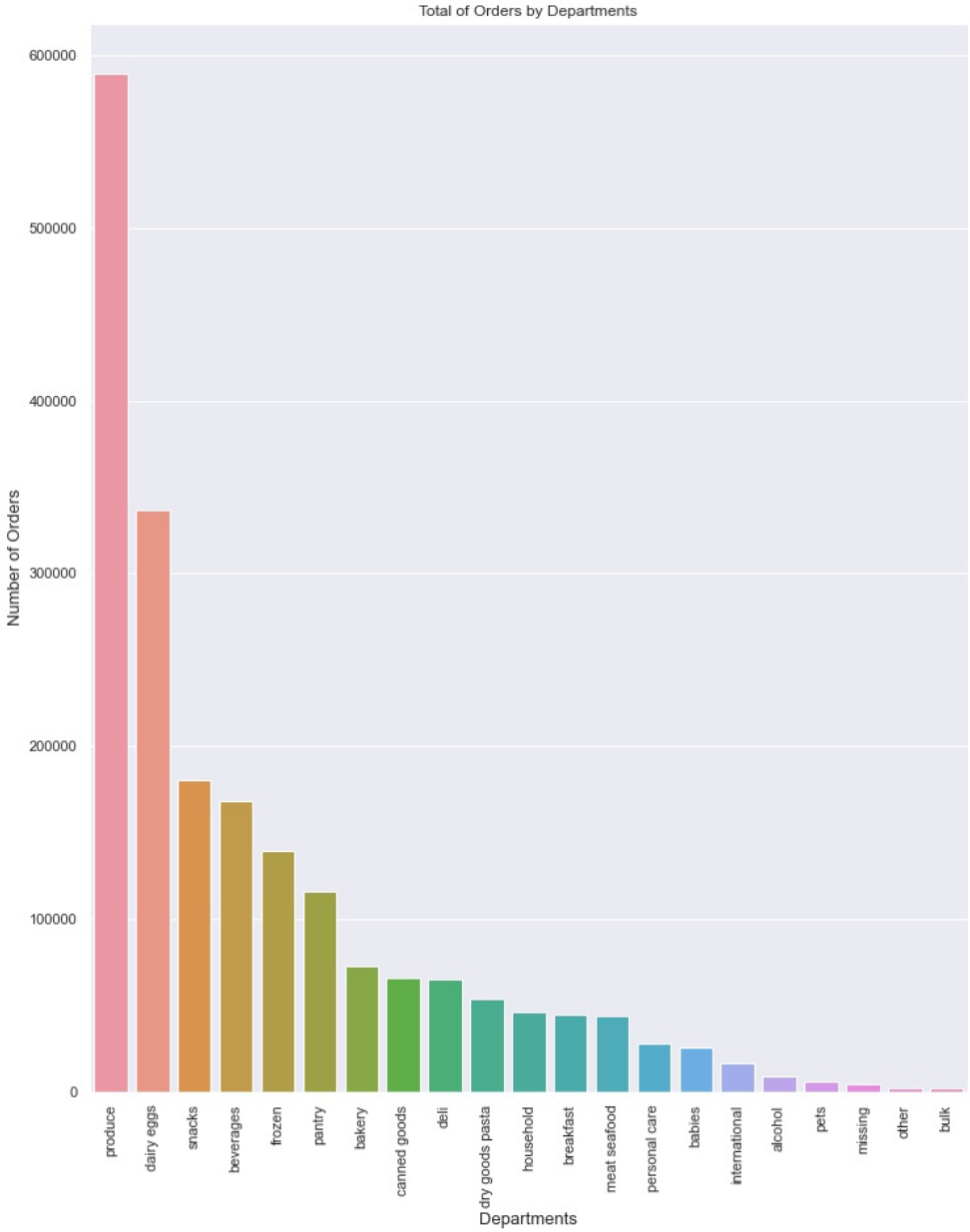

Figure 8 - Most important Departments by the number of orders.

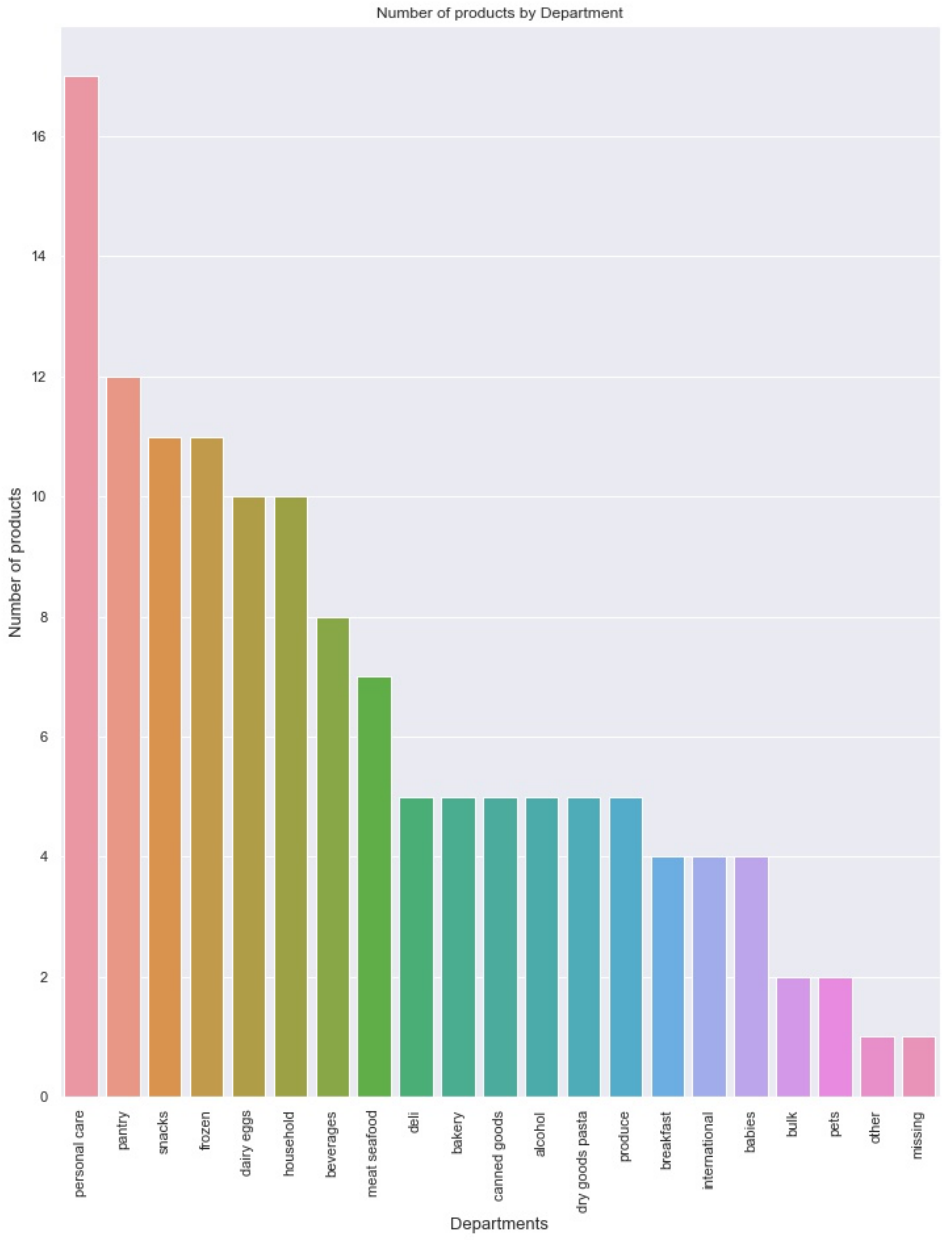

Figure 9 - Number of products by Departments.

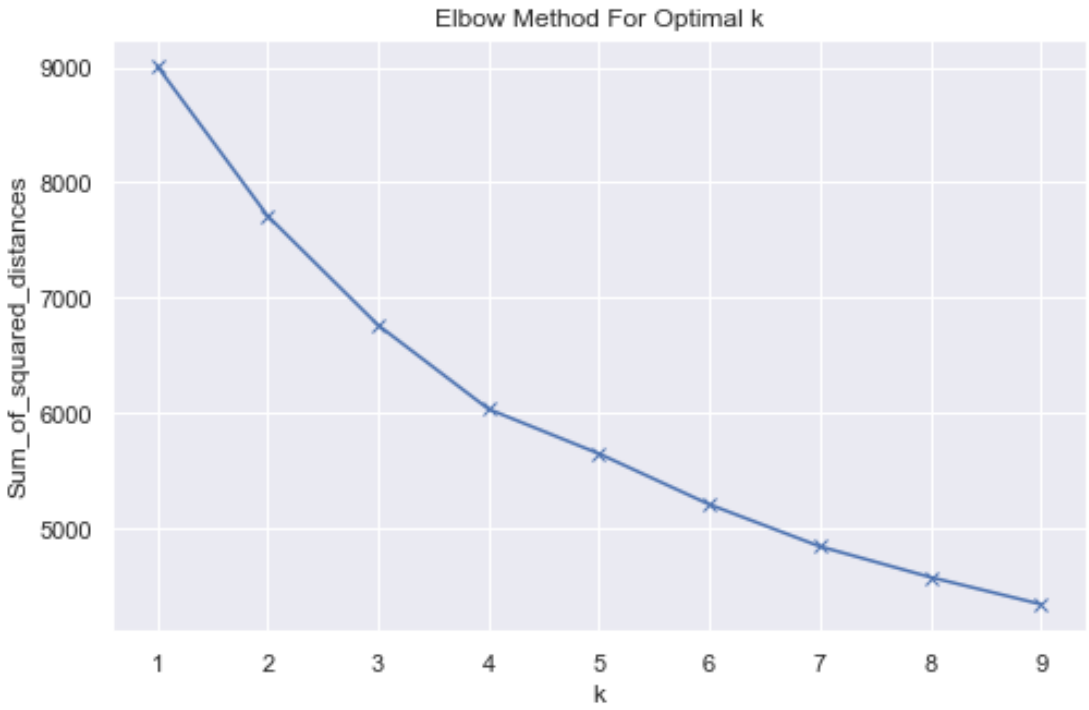

Figure 10 - Elbow Method for Optimal K.

	product_id	total_reorders	product_name
23	24	226039	fresh fruits
82	83	212611	fresh vegetables
122	123	109596	packaged vegetables fruits
119	120	90751	yogurt
20	21	61502	packaged cheese
83	84	55150	milk
114	115	52564	water seltzer sparkling water
106	107	45306	chips pretzels
90	91	39389	soy lactosefree
111	112	36381	bread

Figure 11 - Top 10 best-selling products.

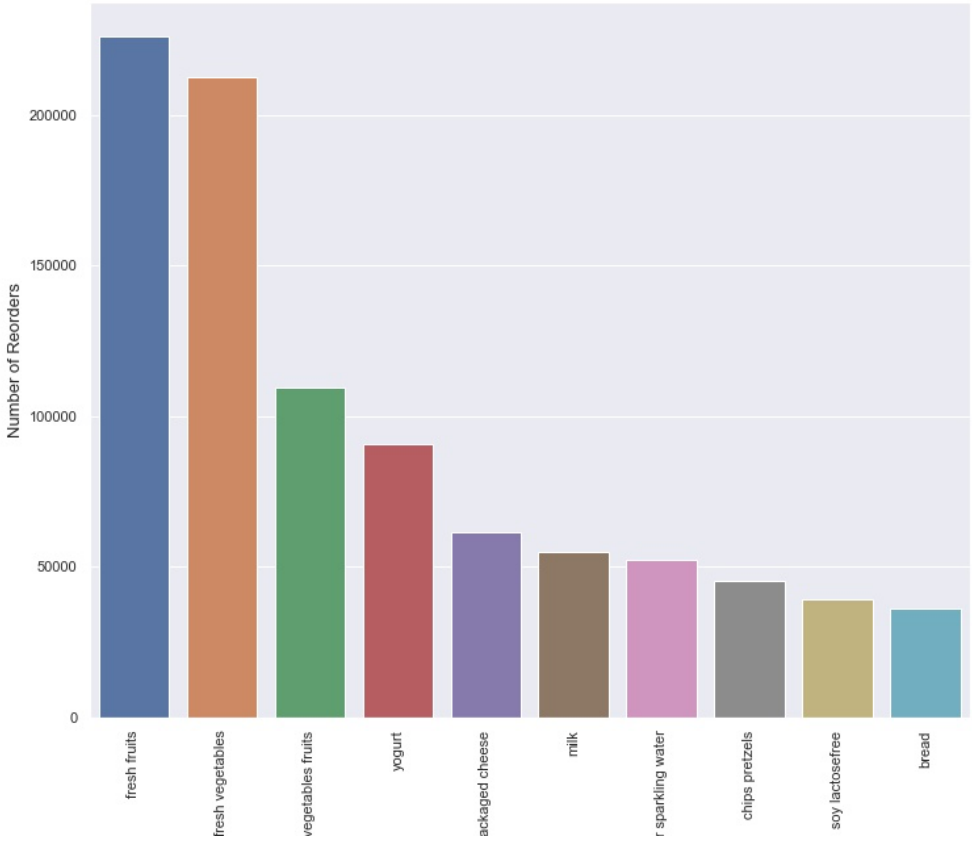

Figure 12 - Top 10 best-selling products according to reorders.

	reordered	total_products	Ratios
0	0	828515	0.410257
1	1	1190986	0.589743

Figure 13 - Reorder Frequency

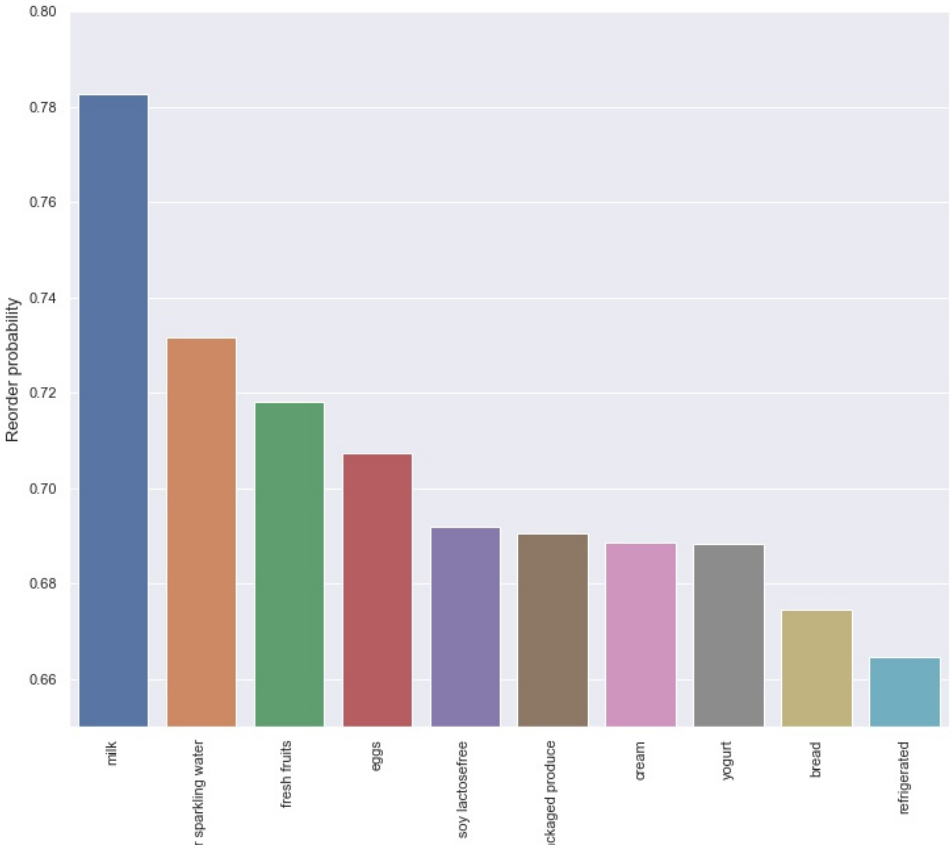

Figure 14 - Top 10 best-selling products according to reorders (in percentage).

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
64	(fresh fruits, fresh herbs)	(fresh vegetables)	0.070135	0.44436	0.061815	0.881372	1.983463	0.030650	4.683872
34	(fresh herbs)	(fresh vegetables)	0.093005	0.44436	0.078655	0.845707	1.903203	0.037327	3.601205

Figure 15 - Complementary products.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
110	(water seltzer sparkling water)	(fresh vegetables)	0.193005	0.444360	0.083355	0.431880	0.971915	-0.002409	0.978033
111	(fresh vegetables)	(water seltzer sparkling water)	0.444360	0.193005	0.083355	0.187584	0.971915	-0.002409	0.993328
232	(fresh fruits, fresh vegetables)	(water seltzer sparkling water)	0.317560	0.193005	0.063235	0.199128	1.031723	0.001944	1.007645
233	(water seltzer sparkling water)	(fresh fruits, fresh vegetables)	0.193005	0.317560	0.063235	0.327634	1.031723	0.001944	1.014983
82	(water seltzer sparkling water)	(fresh fruits)	0.193005	0.555995	0.111045	0.575348	1.034807	0.003735	1.045573
83	(fresh fruits)	(water seltzer sparkling water)	0.555995	0.193005	0.111045	0.199723	1.034807	0.003735	1.008395
133	(packaged vegetables fruits)	(water seltzer sparkling water)	0.365415	0.193005	0.073715	0.201730	1.045204	0.003188	1.010929
132	(water seltzer sparkling water)	(packaged vegetables fruits)	0.193005	0.365415	0.073715	0.381933	1.045204	0.003188	1.026725
59	(ice cream ice)	(fresh fruits)	0.110510	0.555995	0.064485	0.583522	1.049509	0.003042	1.066094
58	(fresh fruits)	(ice cream ice)	0.555995	0.110510	0.064485	0.115981	1.049509	0.003042	1.006189

Figure 16 - Substitutes products.