BUSINESS CASES WITH DATA SCIENCE

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

Business Case 3 - Instacart Market Basket Analysis

Group W

Ana Luísa Mestre, number: 20200599

Beatriz Pereira, number: 20200674

Mariana Domingues, number: 20201040

Nadine Aldesouky, number: 20200568

April, 2020

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

INDEX

1. INTRODUCTION

Instacart wants to analyze consumer behavior using market basket analysis in order to understand the relationship between its products (i.e. which products are complementary, substitutes and/or require more variety of options).
Our team of data scientists has taken on this project to help Instacart better understand its customers' tendencies in order to capitalize on them in terms of product offerings.

2. BUSINESS UNDERSTANDING

2.1. Determine Business Objectives

1. **Background**: Instacart is an American company which delivers groceries around the US and Canada. The company offers delivery and pick-up service using a website and a mobile app. It currently has a large amount of transactional data waiting to be analyzed for insights and potential strategic benefits. This unused data problem affects the Sales, Accounts, Inventory, Supply Chain and Partnerships, Product, Marketing, and Customer Service departments. As a result, Jane Doe, the District Manager has hired Group W to help understand the business better using the transactional data. Indeed, Jane aims to implement proactive marketing strategies based on the market basket analysis to provide extended amounts of offerings for popular products, and cross-promotional programs for complementary products etc. Currently, the lack of data use and data-driven decisions has had an implicit negative effect on Instacart as the company continues to miss on opportunities.
2. **Business Objective/Problem**: To analyse consumer behaviour from existing transactional data to uncover purchasing patterns across products in order to optimize product offerings, marketing programs and webpage or app layout.
3. **Business Success Criteria**: Successful analysis will result in accurate identification of relationships between products to understand which are complements and substitutes and provide a holistic view of Instacart's portfolio of products.

2.2. Assess the Situation

- **Inventory of Resources**: Currently we have a database of 200000 orders which will be analyzed by a team of four business analysts known as Group W. To manage our problem we will use Python, Microsoft Word, Power Point, and Excel.
- **Requirements, Assumptions, and Concerns**: This report is for managers' use which means that it is for private use in Instacart only. Additionally, this is not a technical audience meaning that all aspects must be comprehensible in business terms without technical jargon. For this project, we will be working with 4 relational datasets containing information on a sample of 200000 transactions made by more than 100000 users.
- **Risks and Constraints**: The main constraint is the lack of computing power and memory space available to host the datasets. As a result, for each user, only a few of their orders are listed in these datasets. Also, since the original dataset was too large, the products were grouped by type and generalized into 134 products.

Risks	Contingencies
Missing relevant information	Ask for more detailed explanation in meta data e.g. order_number
Losing the data	Create a copy of all aspects of the project
General level of detail	Ask for the original dataset

Table 2.1.– Risks and Contingencies

- **Terminology**:

1. *Add to cart order:* This is like a virtual shopping cart so when you put groceries in your trolley in the supermarket, you are adding orders to your cart in the app or web-shop.
2. **Costs and Benefits**:
   As this is an academic project, we have no information about the costs of the data collection neither the ones regarding the development and the implementation of our solution for the business problem.
   The benefits include increasing sales, increasing repeated orders, increasing ROI, and improving customer satisfaction and targeting, leading to higher revenues and lower costs.

2.3. Determine Data Mining Goals

- **Data Mining Goals**: To complete a market basket analysis on existing transactional data to uncover association rules and insights on relationships between products.
- **Data Mining Success Criteria**: Successful market basket analysis will result in the identification of different association rules for different products with a good confidence factor and a high level of support.

2.4. Produce Project Plan

Steps

- General exploratory analysis, identifying clear problems in the data such as missing values, outliers, or incoherencies; and revealing insights and patterns.
- Describe the insights obtained from the data exploration by producing a set of visualizations. This is to understand the absolute frequencies of the variables, identify the presence of outliers and extract the maximum amount of information from the data.
- Apply the *Apriori* algorithm[1][2] to build the association rules between the different products and consequently distinguish between complementary and substitute products.
- Define different thresholds for the lift, confidence, and support metrics, and then compare and interpret the different resulting outputs.
- Review the project, identifying the main limitations that were experienced.
- Provide a deployment plan.

Initial Assessment of Tools and Techniques

Technique	Pros	Cons
***Apriori* Algorithm**	- Simple and easy to implement. - Works well with large datasets (includes pruning steps).	- Computationally expensive based on the thresholds and diversity of items. - Very complex in terms of time and space.

		- Performance reduced by the numerous times the algorithm scans the dataset.

Table 2.2.– Tools and Techniques

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data Understanding

The dataset was provided as four separate csv files by the District Manager of Instacart, Jane Doe, directly from the company. It contains a sample of 200000 orders, 105273 users, and 134 generalized product groupings.

The four provided datasets were the following:

- Orders with 200000 rows and 6 columns, from which 5 are non-metric variables and 1 is a metric-variable.
- Order_products with 2019501 rows and 4 columns, all of which are non-metric variables, including 1 binary variable – "reordered".
- Departments with 21 rows and 2 columns, all of which are non-metric variables.
- Products with 134 rows and 3 columns, all of which are non-metric variables.

Next, we present the different aspects of the dataset which are divided into three parts: general analysis, departments analysis, and products analysis.

General Analysis

- In the following heatmap we can see the distribution of the orders by day of week and hour. The hours with more orders are between 10am and 4pm.
- The most common (in terms of highest number of orders) days of the week to do grocery shopping on Instacart are Sunday and Monday. The period in which the largest number of orders was registered was Sunday at 1pm.

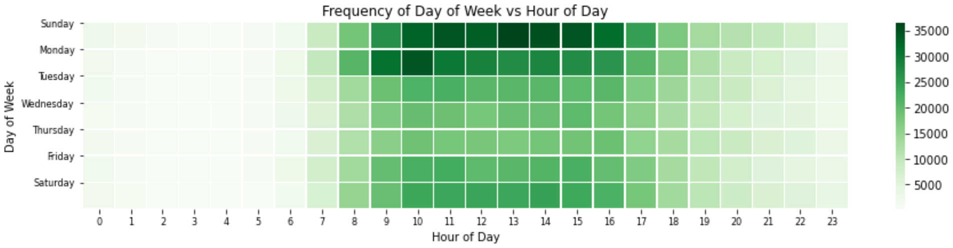


Figure 3.1.– Heatmap of Orders by Hour and Day of Week

- The average number of products per order is equal to 10 with a large standard deviation of 7.5.

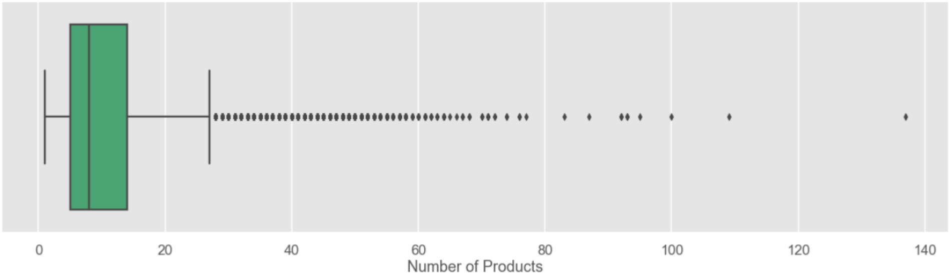


Figure 3.2.– Boxplot of Number of Products by Order

- 75% of the orders have less than 14 products (third quartile). The most frequent number of products bought in each order is between 4 and 7.
- Regarding the number of orders by user, given that the data does not have information about the specific date of the orders, and it does not seem to follow a continuous timeline, there is an incoherence between the count of orders by user and their maximum “order_number”. If the “order_number” variable follows the number of orders by user, it is expected to have its max as the quantity of orders in the data but, when counting the orders, we get fewer numbers. This is because, as mentioned earlier, only a few orders per user were provided in this dataset for lack of computational resources. To demonstrate, it is possible to have a user with three observations in the data being the 3º, 5º and 10º order of that user, meaning that we already received 10 orders from the user, but this dataset only contains 3 of those orders. The figure below illustrates the distribution of these two variables, the count of the orders (on top) and the maximum “order_number” by user (on the bottom).

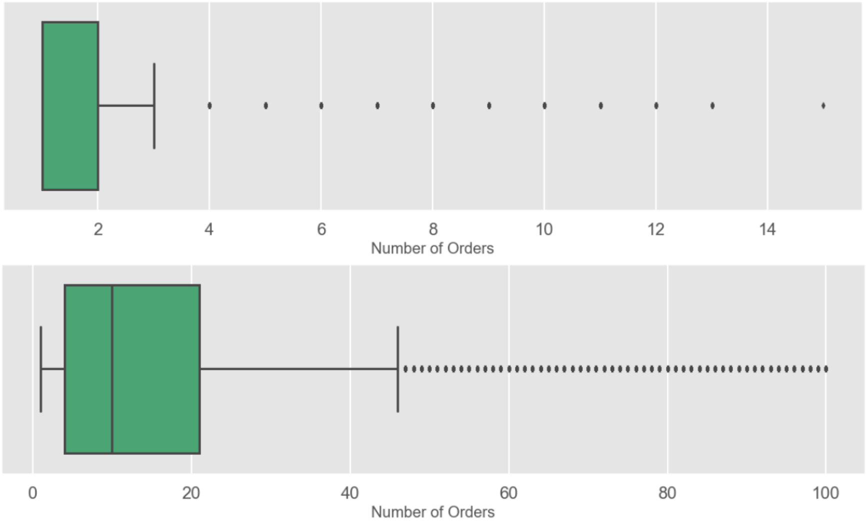


Figure 3.3.– Boxplot of Number of Orders by User

- Interpreting only the observations (orders) presented in the provided dataset, approximately 55% of the users have only 1 order and 75% of the users have 1 or 2 orders (third quartile).
- Taking into consideration all the orders except the first user orders, that surely have a reorder rate of zero, in the next graphics we can see the distribution of the reorder rate by user and by order:
    - The average number of the reorder rate by user is around 0.58.
    - It is most common to have users with a reorder rate between 0.6 and 0.8.
    - There are 11235 users that have a reorder rate of 1, i.e., users that always buy the same products as they ordered in the past, and 5547 with a rate of 0, i.e., users that always buy new products.
    - The average number of the reorder rate by order is around 0.64.
    - It is most common to have orders with a reorder rate equal to 1.
    - There are 43037 orders that have a reorder rate of 1, i.e., orders with all the products reordered by a specific user, and 11402 with a rate of 0, i.e., that are only composed of new products ordered by a specific user.

- In general, it is more common to have reordered products than orders with new products. There are 1190986 reordered products (around 63% of the data).

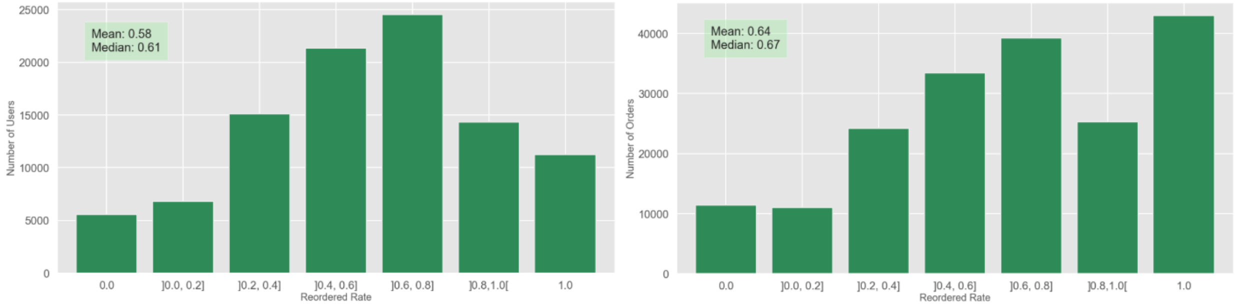


Figure 3.4.– Distribution of the Reorder Rate by user and by order

- Counting the reordered products by the variable order_number from each user we can see a clear relationship, remarkably similar to a logarithmic function. The reorder rate increases a lot in the first 20 orders of a user and increases more slowly after that. We can also conclude that the higher the number of the order, the higher the reorder rate fluctuation, so the biggest uncertainty of this relationship.

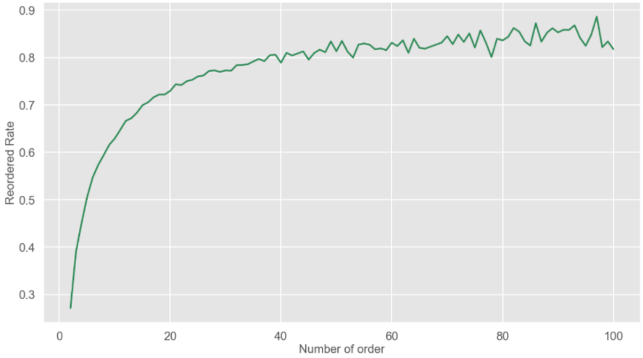


Figure 3.5.– Order Number (from user) vs Reorder Rate

Departments Analysis

- There are 21 different departments of different sizes forming a total of 134 distinct products. (A table listing the products per department can be found in the appendix 1)
- The department with the highest number of orders is Produce (with 588996 orders) while the department with the lowest number of orders is Bulk (with 2133 orders).
- Customers' shopping behaviour in departments such as household, beverages, pets, and snacks is spread out throughout the week (distribution on the left of figure 3.6).
- Customers shopping behaviour in the rest of the departments is mostly focused during weekends and Mondays (distribution on the middle of figure 3.6).
- For the department of alcohol, there is a clear pattern in the shopping behaviour where it increases until Friday and then drops. The reason the number of alcohol orders drops in the weekend is probably because delivery fees at Instacart are higher in the weekend and so customers will prefer to buy their alcohol beforehand (distribution on the right of figure 3.6).

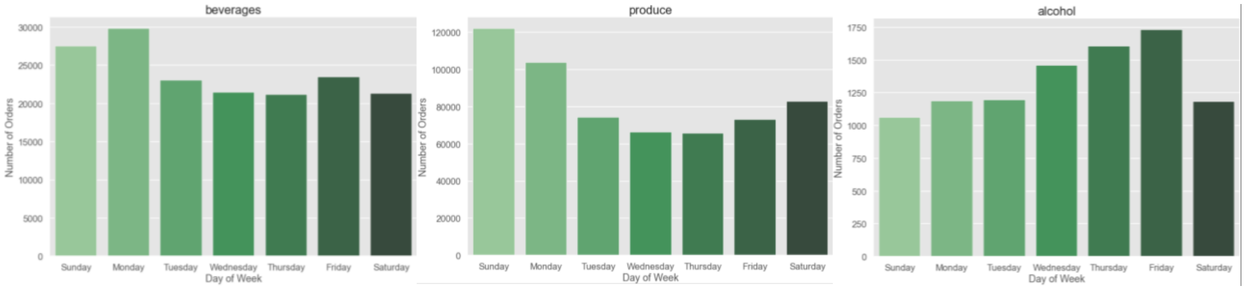


Figure 3.6.– Distribution of Orders by Day of Week and Department

Products Analysis

- The top 4 products sold are fresh fruits, fresh vegetables, packaged vegetables fruits and yogurt.
- The top 4 products reordered are the same as above. However, taking the ratio between the reorders and the product orders into consideration, the top 4 reordered products are milk, water seltzer sparkling water, fresh fruits, and eggs. The graphic below was created excluding the first orders, given that in the first orders it is impossible to have reordered products.

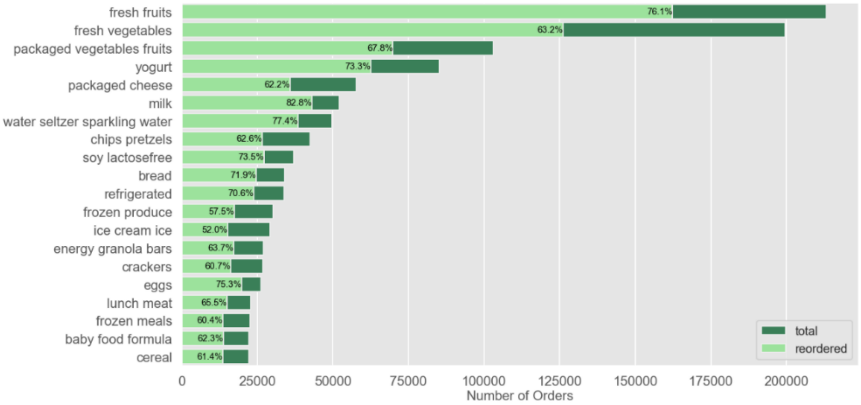


Figure 3.7.– Total Orders and Reorders from Most Popular Products

- Fresh Fruits is the product that appears in the highest number of orders as the first one being added to the cart.

3.2. Data Preparation

As already mentioned above, the dataset was provided in a form of four separated csv files - *departments.csv*, *order_products.csv*, *orders.csv*, and *products.csv*. Thus, in order to start analyzing and preparing the data, the first step was to merge the different datasets, according to the analysis to be performed.

This data preparation phase was quite short and simple since the data had already gone through a selection process by the District Manager of the company.

The dataset contained missing values only in the variable “days_since_prior_order” which, in this case, are associated to the first order of a user and, naturally, there is no prior order.

As for the outliers’ analysis, we explored the only metric variable in the data - “days_since_prior_order”. Using the IQR method and setting the multiplier to 1.5, the variable had no datapoints beyond the boundaries, so we did not remove any observations.

Additionally, we created the variable nproducts_byorder which represents the number of products by order. This variable was created by getting the maximum number of the variable “add_to_cart_order” by order. However, this variable was only used for preliminary analysis and was not added to the final dataframe. As previously shown in figure 3.2, we detected some outliers in this variable. There were 6589 orders outside of the upper bound which is the 27 products limit. These orders represented a high percentage of the total data which we did not want to lose. As a result, we decided to define a higher limit and only delete the orders that had more than 80 products. This corresponded to only 10 orders which were composed of a total of 981 products (or rows). Thus, the dropped observations represented only 0.05% of the data so we believe we kept the integrity of the original data. To the next step the input data has 2018520 rows.

3.3. Modeling

3.3.1 Association Technique

For this project, we used the *Apriori* algorithm which assumes that all subsets of the frequent item sets must also be frequent while all supersets of the infrequent items will also be infrequent and so exculded from the generation of association rules. We created the frequent itemset supported by at least 3% of transactions; this is to show as many rules as possible which we will adjust later by setting higher thresholds. Moreover, it is because when we filtered out the top products from the frequent itemset, higher thresholds did not generate enough rules at all. The resulting frequent itemset only included 62 of the 134 Instacart products which created a total of 397 items with sets ranging from a single product to a maximum of 5.

Indeed, the model was able to point out many correlations between the products after these low thresholds. However, most of the association rules generated included one of the top 4 products bought (i.e. Fresh Fruits, Fresh Vegetables, Packaged Vegetables Fruits, Yogurt). This makes these rules **trivial** because either way the consumers will buy these top products and they are not actually linked to anything.

The association rules for both the complementary products (filtered by confidence higher than 0.75 and lift higher than 1.8) and the substitute products (filtered by confidence higher than 0.20 and lift lower than 1.04) are listed in the appendix 2.

Since this list of rules proved to be trivial, we decided to filter out any rules which included any of the top 4 products. This is to focus on the rules generated by the *less* popular products.

Now, the resulting frequent itemset only included 58 of the 134 Instacart products which created a decreased total of 102 items with combinations ranging from a single product to a maximum of 2. This means that the other generated combinations of 3, 4, or 5 products included one or more of the top 4 products.

Filtering the product associations with a confidence higher than 0.10 and lift higher than 1.5 the resulting rules for the **complementary** products are listed in the appendix 3. The following **actionable rules** were noticed:

- Packaged cheese is usually complemented by lunch meat, hot dogs bacon sausage, other cream cheeses, crackers and bread.
- Chips pretzels tend to be complemented by fresh dips tapenades and crackers.
- Bread is usually complemented by lunch meat, eggs, and packaged cheese.
- Milk is usually complemented by cereal.
- Frozen produce is usually accompanied by soy lactose free.

Filtering the product associations with a confidence higher than 0.15 and lift lower than 1.05, the resulting rules for the **substitute** products are listed in the appendix 4. We noticed that Milk is usually substituted by soy lactose free and sparkling water.

Although these rules have quite low confidence and support, we decided to trust them as logically they make sense and are explainable. Additionally, the reason why their support and confidence levels are low is because of the disproportionate difference between the popularity of the top 4 products and the rest of the Instacart products. The following NetworkX diagram [3] illustrates the relationships and connections between the different products:

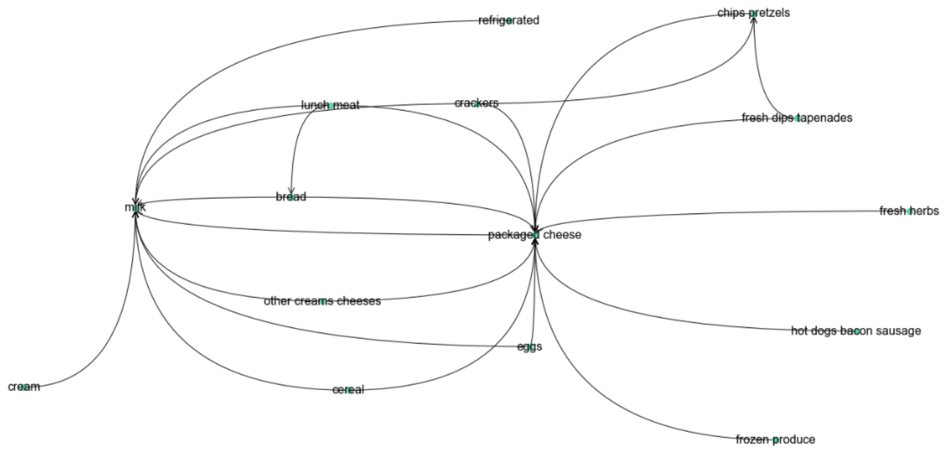


Figure 3.8.– NetworkX Directed Graph

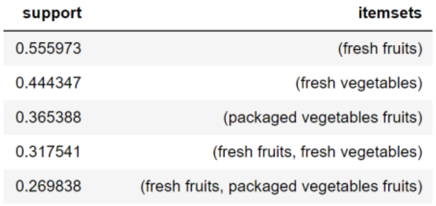
We also applied the algorithm in two different perspectives – weekdays vs. weekends, and hours of the day - to try to create different rules for each group. However, the groups gave very similar results, so we discarded this option, assuming that there is no significant different between these groups.

3.4 Evaluation

3.4.1 Evaluate Association Model's Results

"A credible rule should have a high confidence factor, a high level of support and a lift higher than 1 (for complementary products) and lower than 1 (for substitute products)."

Unfortunately, in this case the rules with the highest support and confidence were the ones that included the top 4 products. This is just because these products are generally bought more often and so they will surely occur together with other products; this does not prove correlation or association. The figure below demonstrates how support increases with the popularity of the product or product set.



support	itemsets
0.555973	(fresh fruits)
0.444347	(fresh vegetables)
0.365388	(packaged vegetables fruits)
0.317541	(fresh fruits, fresh vegetables)
0.269838	(fresh fruits, packaged vegetables fruits)

Figure 3.9.– Support of Popular Products

Furthermore, in this abnormal case, the consequent happens more often than the antecedent (because the consequents are the top 4 products). This results in a higher confidence for the rules which include the top 4 products. Indeed, looking at the confidence formula (A∩C/A), a small denominator will result in a higher value. Given the explanations above, we do not consider these rules credible. This can be illustrated in appendix 5, a Venn diagram representing the relationship with the highest confidence, of 91%.

On the other hand, the rules generated from the combinations excluding the top 4 products are the ones we consider credible. This is because they are logical rules that can be explained and transformed into actionable rules. We believe that their lower confidence and support is mostly because of the excessive difference between the popularity of the products. In appendix 6, we illustrate the relationships with the highest confidence levels by two Venn diagrams for complementary and substitute products.

3.4.2 Review Process

During this project, our biggest limitation was the missing information. To make this process more accurate we should have more information about the products and their categories. The products that the data contains are already grouped. It would be beneficial to have more detailed knowledge on how they were grouped. Another important component which was missing are the orders' specific dates that could be used to get information about the month, year, weather, season or holiday in which the orders took place. All these parameters can influence the consumer behavior and the delivery taxes of the orders. Another drawback was the explanation or reasoning behind the selection of the included orders per user. The number of orders made by user in the data does not correspond to the maximum number of the variable "order_number", meaning that we do not have the orders in a continuous timeline.
Additionally, the data collection process was quite ambiguous which restricted our conclusions. In other words, we did not have a timeline to base our insights on. Was this data collected from the last year? The last quarter? This context information is important to better understand the dataset and find justifications for the insights it reveals.

3.4.3 Determine Next Steps

- Proceed to the deployment step.
- Set the appropriate level of detail (specify the products instead of the groups of products).
- Improve the thresholds for the support, confidence, and lift metrics.
- Improve the "tree" with the association rules - tunning the parameters of the plot.

4. DEPLOYMENT

4.1. Deployment Plan

After having carried out the analysis, we think that it would be advantageous to have a greater variety of packaged vegetables fruits and yogurt as they are two types of products that have high popularity but possibly did not reach their full potential. In this dataset, there are 109596 orders of packaged vegetable fruits and 90751 orders of yogurt. Additionally, they have a high reorder rate of 67.8% and 73.3% respectively. Thus, we think that the company would be able to sell more of these types of products if it had a more diversified portfolio, reaching quantities as high as the fresh fruits and fresh vegetables sold for example. An extended variety of yogurt can include different brands, flavours, fat percentages, vegan bases etc. An extended variety of packaged vegetables fruits could include different brands, mixes, sizes, themes, purposes etc.

Furthermore, concerning the complementary products, when a customer adds a product to the cart, you should suggest one or more of its complementary items (e.g. chips pretzels and crackers). Additionally, the layout of the landing page can be changed to place the complementary products closer together. Indeed, offers can be created where bundles of complementary products are sold together for discounted prices.

Regarding products that are substitutes of each other, our suggestion would be, when a product is out of stock, suggest to the customer its substitute (e.g. milk and soy lactose free). Additionally, the products added to the cart and the scrolling behaviour can be monitored so that if the customer passed milk and did not add it to their cart for example, you should immediately suggest soy lactose free products. Then mark or tag this customer to later notify them with any offers on soy lactose free products and move these products up on their page when they shop.

Moreover, the lower selling products can be paired with the popular products to create the free ride effect where customers begin to see and buy these non-popular items more often until they get used to them and they become popular. For example, the following products can be paired with **packaged vegetables fruits** to increase their popularity:
Oils vinegars, pickled goods olives, preserved dips spreads, salad dressing toppings, spices seasonings, spreads, condiments, prepared meals, prepared soups salads, tofu meat alternatives, soup broth bouillon, canned meals beans.
These products would go well together to make salads, soups and they are a fit for the lazy consumer who wants ready, easy and quick food.

4.2. Plan Monitoring and Maintenance

- Review the model when a new product becomes available at Instacart.
- Evaluate the model, after a certain number of orders (to be determined by the company), to find out if there are new rules for the antecedents vs. consequents products or even if those previously considered are still valid.
- Evaluate the model at least twice a year (every six months) to keep on having confidence in the association rules to apply and to monitor their credibility (whether it increases or decreases).
- Monitor the model depending on the seasons (it is necessary to start collecting this data, regarding the date or season), since there are products that can be complementary / substitutes in certain seasons but not in others - an alternative to this approach is to reach out to our team whenever a season changes.

4.3. Review Project

The market basket analysis requires a large amount of data in order to be successful. Indeed, the data we received was large, but it did not have the detail we expected. The product groupings were too general meaning that the product list was not granular enough. We believe that having more subcategories of the presented products would have given a more accurate and detailed analysis.
Our biggest limitation was the missing information, essentially about the products families, the orders' specific dates and the app layout. To manage the goal of defining different types of consumer behavior, it would be essential to have more information about the products and the orders. This could entail the products' prices, whether they were bought in a promotion or not, the orders' dates and delivery tax.
To provide ideas on how to optimize the web or app layout, it would also be crucial to understand how the app is managed and how the customers interact with it. Particularly, what the customer sees when opening the app and how that is decided. Plus, if there are any filters to show ongoing promotions, previous ordered products, or a simple menu with the products in an alphabetic order. It is important to clarify how the app/web is operated and if it has any targeting mechanisms in place that adjust and customize the layout based on the individual shopper using it.
Another aspect that could have influenced our results, related with the orders dates, is not having the orders in a continuous timeline.

It is fundamental to give all the necessary knowledge about the client's business to a data science team in order to reach the optimal results. Otherwise, the project gets burdensome, and the final results do not fulfill the client's requests.

5. APPENDIX

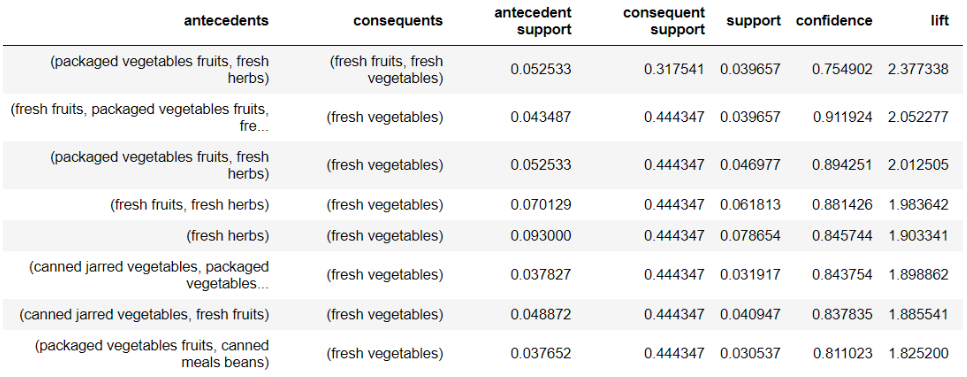
Appendix 1

Other (1 product)	**Missing** (1 product)	**Bulk** (2 products)
Other	Missing	Bulk dried fruits vegetables Bulk grains rice dried goods
Pets (2 products)	**Alcohol** (4 products)	**Babies** (4 products)
Cat food care Dog food care	beers coolers red wines specialty wines champagnes spirits	Baby accessories Baby bath body care Baby food formula Diapers wipes
Breakfast (4 products)	**International** (4 products)	**Dry goods pasta** (5 products)
Breakfast bars pastries Cereal Granola Hot cereal pancake mixes	Asian foods Indian foods Kosher foods Latino foods	Dry pasta Fresh pasta Grains rice dried goods Instant foods Pasta sauce
Canned goods (5 products)	**Deli** (5 products)	**Produce** (5 products)
Canned fruit applesauce Canned jarred vegetables Canned meals beans Canned meat seafood Soup broth bouillon	Fresh dips tapenades Lunch meat Prepared meals Prepared soups salads Tofu meat alternatives	Fresh fruits Fresh herbs Fresh vegetables Packaged produce Packaged vegetables fruits
Bakery (5 products)	**Meat Seafood** (7 products)	**Beverages** (8 products)
Bakery desserts Bread Breakfast bakery Buns rolls Tortillas flat bread	Hot dogs bacon sausage Meat counter Packaged meat Packaged poultry Packaged seafood Poultry counter Seafood counter	Cocoa drink mixes Coffee Energy sports drinks Juice nectars Refrigerated Soft drinks Tea Water seltzer sparkling water
Dairy eggs (10 products)	**Household** (10 products)	**Frozen** (11 products)
Butter Cream Eggs Milk Other creams cheeses Packaged cheese Refrigerated pudding desserts Soy lactosefree Specialty cheeses Yogurt	Air fresheners candles Cleaning products Dish detergents Food storage Kitchen supplies Laundry More household Paper goods Plates bowls cups flatware Trash bags liners	Frozen appetizers sides Frozen breads doughs Frozen breakfast Frozen dessert Frozen juice Frozen meals Frozen meat seafood Frozen pizza Frozen produce Frozen vegan vegetarian Ice cream ice
Snacks (11 products)	**Pantry** (12 products)	**Personal care** (17 products)

Candy chocolate Chips pretzels Cookies cakes Crackers Energy granola bars Fruit vegetable snacks Ice cream toppings Mint gum Nuts seeds dried fruit Popcorn jerky Trail mix snack mix	Baking ingredients Baking supplies décor Condiments Doughs gelatins bake mixes Honeys syrups nectars Marinades meat preparation Oils vinegars Pickled goods olives Preserved dips spreads Salad dressing toppings Spices seasonings Spreads	Beauty Body lotions soap Cold flu allergy Deodorants Digestion Eye ear care Facial care Feminine care First aid Hair care Muscles joints pain relief Oral hygiene Protein meal replacements Shave needs Skin care Soap Vitamins supplements

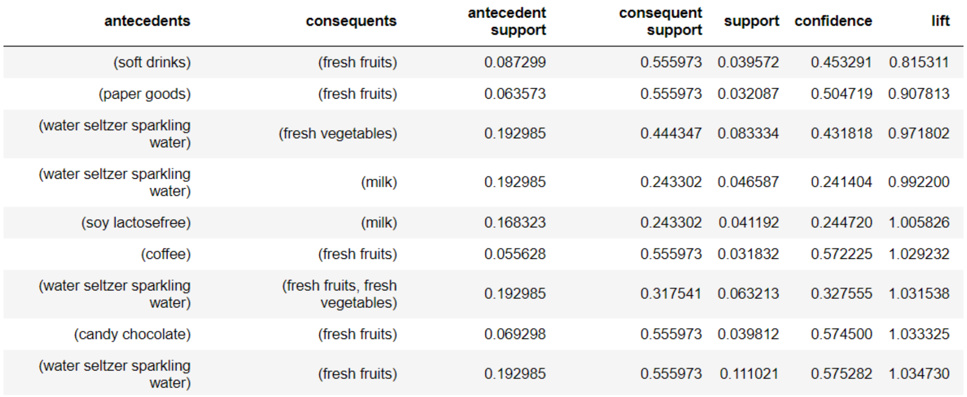
Table 5.1.– Products by Department

Appendix 2



antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(packaged vegetables fruits, fresh herbs)	(fresh fruits, fresh vegetables)	0.052533	0.317541	0.039657	0.754902	2.377338
(fresh fruits, packaged vegetables fruits, fre...	(fresh vegetables)	0.043487	0.444347	0.039657	0.911924	2.052277
(packaged vegetables fruits, fresh herbs)	(fresh vegetables)	0.052533	0.444347	0.046977	0.894251	2.012505
(fresh fruits, fresh herbs)	(fresh vegetables)	0.070129	0.444347	0.061813	0.881426	1.983642
(fresh herbs)	(fresh vegetables)	0.093000	0.444347	0.078654	0.845744	1.903341
(canned jarred vegetables, packaged vegetables...	(fresh vegetables)	0.037827	0.444347	0.031917	0.843754	1.898862
(canned jarred vegetables, fresh fruits)	(fresh vegetables)	0.048872	0.444347	0.040947	0.837835	1.885541
(packaged vegetables fruits, canned meals beans)	(fresh vegetables)	0.037652	0.444347	0.030537	0.811023	1.825200

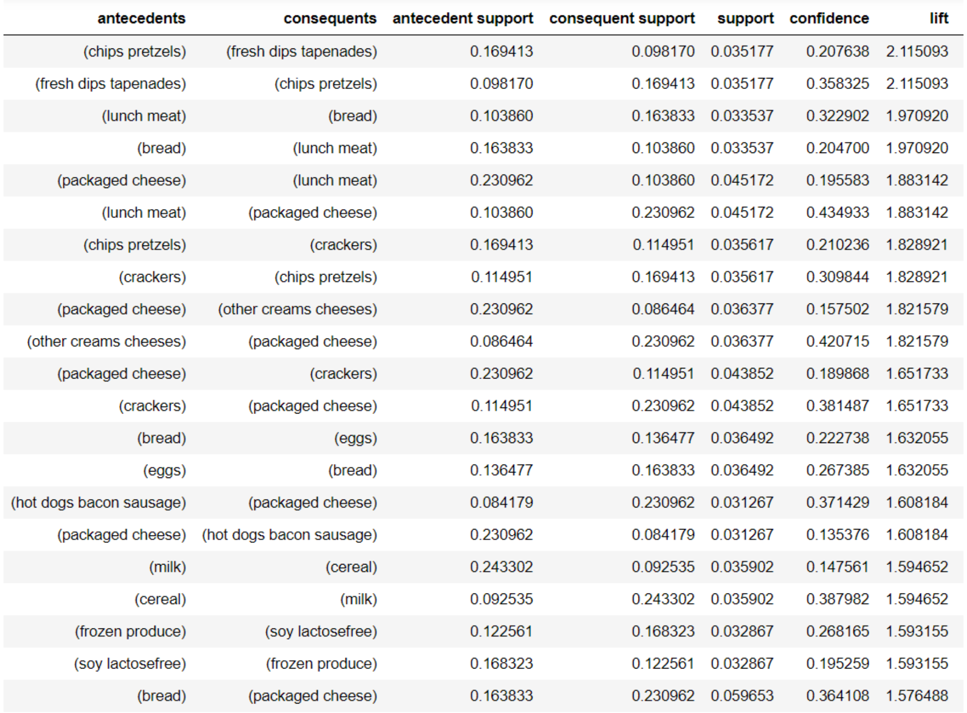
Figure 5.1.– Complementary Products (descendent order by lift)



antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(soft drinks)	(fresh fruits)	0.087299	0.555973	0.039572	0.453291	0.815311
(paper goods)	(fresh fruits)	0.063573	0.555973	0.032087	0.504719	0.907813
(water seltzer sparkling water)	(fresh vegetables)	0.192985	0.444347	0.083334	0.431818	0.971802
(water seltzer sparkling water)	(milk)	0.192985	0.243302	0.046587	0.241404	0.992200
(soy lactosefree)	(milk)	0.168323	0.243302	0.041192	0.244720	1.005826
(coffee)	(fresh fruits)	0.055628	0.555973	0.031832	0.572225	1.029232
(water seltzer sparkling water)	(fresh fruits, fresh vegetables)	0.192985	0.317541	0.063213	0.327555	1.031538
(candy chocolate)	(fresh fruits)	0.069298	0.555973	0.039812	0.574500	1.033325
(water seltzer sparkling water)	(fresh fruits)	0.192985	0.555973	0.111021	0.575282	1.034730

Figure 5.2.– Substitute Products (ascendent order by lift)

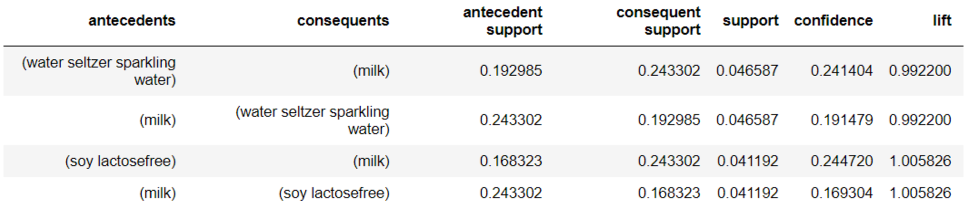
Appendix 3



antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(chips pretzels)	(fresh dips tapenades)	0.169413	0.098170	0.035177	0.207638	2.115093
(fresh dips tapenades)	(chips pretzels)	0.098170	0.169413	0.035177	0.358325	2.115093
(lunch meat)	(bread)	0.103860	0.163833	0.033537	0.322902	1.970920
(bread)	(lunch meat)	0.163833	0.103860	0.033537	0.204700	1.970920
(packaged cheese)	(lunch meat)	0.230962	0.103860	0.045172	0.195583	1.883142
(lunch meat)	(packaged cheese)	0.103860	0.230962	0.045172	0.434933	1.883142
(chips pretzels)	(crackers)	0.169413	0.114951	0.035617	0.210236	1.828921
(crackers)	(chips pretzels)	0.114951	0.169413	0.035617	0.309844	1.828921
(packaged cheese)	(other creams cheeses)	0.230962	0.086464	0.036377	0.157502	1.821579
(other creams cheeses)	(packaged cheese)	0.086464	0.230962	0.036377	0.420715	1.821579
(packaged cheese)	(crackers)	0.230962	0.114951	0.043852	0.189868	1.651733
(crackers)	(packaged cheese)	0.114951	0.230962	0.043852	0.381487	1.651733
(bread)	(eggs)	0.163833	0.136477	0.036492	0.222738	1.632055
(eggs)	(bread)	0.136477	0.163833	0.036492	0.267385	1.632055
(hot dogs bacon sausage)	(packaged cheese)	0.084179	0.230962	0.031267	0.371429	1.608184
(packaged cheese)	(hot dogs bacon sausage)	0.230962	0.084179	0.031267	0.135376	1.608184
(milk)	(cereal)	0.243302	0.092535	0.035902	0.147561	1.594652
(cereal)	(milk)	0.092535	0.243302	0.035902	0.387982	1.594652
(frozen produce)	(soy lactosefree)	0.122561	0.168323	0.032867	0.268165	1.593155
(soy lactosefree)	(frozen produce)	0.168323	0.122561	0.032867	0.195259	1.593155
(bread)	(packaged cheese)	0.163833	0.230962	0.059653	0.364108	1.576488

Figure 5.3.– Complementary Products (descendent order by lift)

Appendix 4



antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(water seltzer sparkling water)	(milk)	0.192985	0.243302	0.046587	0.241404	0.992200
(milk)	(water seltzer sparkling water)	0.243302	0.192985	0.046587	0.191479	0.992200
(soy lactosefree)	(milk)	0.168323	0.243302	0.041192	0.244720	1.005826
(milk)	(soy lactosefree)	0.243302	0.168323	0.041192	0.169304	1.005826

Figure 5.4.– Substitute Products (ascendent order by lift)

Appendix 5

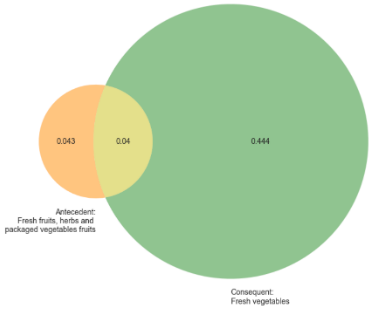


Figure 5.5.– Venn Diagram Phenomena for Popular Products

Appendix 6

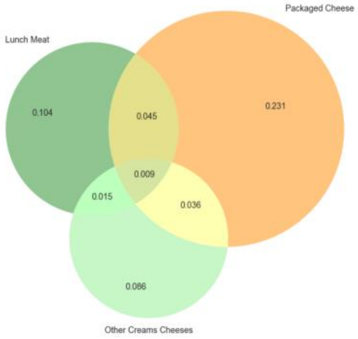


Figure 5.6.– Venn Diagram for Complementary Products

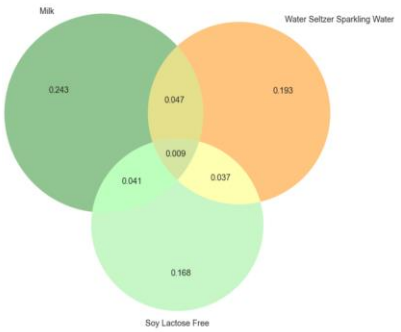


Figure 5.7.– Venn Diagram for Substitute Products

6. REFERENCES

[1] Great Learning Team, “What is Apriori Algorithm in Data Mining Implementation and Examples?,” 2020. https://www.mygreatlearning.com/blog/apriori-algorithm-explained/.

[2] Sebastian Raschka, “Frequent Itemsets via Apriori Algorithm,” 2020. http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/.

[3] NetworkX Developers, “NetworkX 2.5 documentation,” 2020. https://networkx.org/documentation/stable//tutorial.html.