BUSINESS CASES WITH DATA SCIENCE

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Instacart, Market Basket Analysis

Understanding purchasing behaviour

Group Y

João César (20200669)

João Henriques (20200670)

Pedro Sancho Vivas de Castro (20200132)

Vilmar Bussolaro (20200268)

April 19th, 2021

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

INDEX

INTRODUCTION

In this project we are asked to perform market basket analysis, a popular solution to an old business problem.

With the introduction of OLTP (Online transaction processing), companies started collecting large amounts of data, with detailed purchase orders. As it is in most cases with consumer behavior, one of the best predictors of future buying behavior is past buying behavior. Hence, the value in understanding associations in products that are commonly acquired together.

This knowledge has many implications on the marketing mix. For example, smart product placement strategy can reduce friction by making it easier for customers to acquire products. In regular supermarkets, that means placing baby massage oil and ointment next to diapers, or bread, butter and cheese close to each other. On online grocery stores like Instacart, it means having the right products in the catalog, and promoting items in the app that are more likely to convert into sale.

This work covers the business case of Instacart, and how Data Science can improve sales using Market Basket Analysis.

On a preliminary note, we think that there is a possibility of using a cluster implementation on the numeric variables of the dataset and possible feature engineered numeric variables.

Our main notebook and presentation will also be present on this GitHub repository: https://github.com/PedroSancho

BUSINESS UNDERSTANDING

2.1. Background

With COVID-19, Digital Transformation of enterprises have accelerated greatly. Lockdowns, social distancing, and travel restrictions have forced many people into new shopping habits. Grocery shopping is such a case. Reports from eMarketer[1] show a 41.9% increase in sales in just one year, with a more modest increase projected after the pandemic is over.

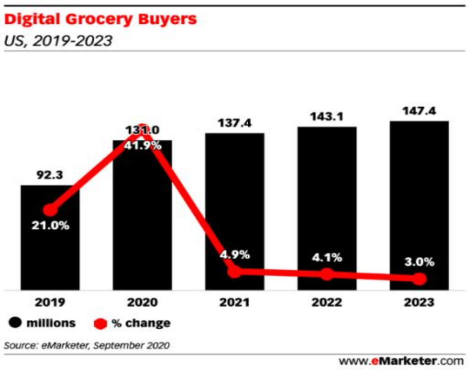


Figure 1 - Online grocery shopping sales in the United States (in billion U.S. dollars)

Data from a recent study made by Statista[2] shows that during the pandemic, its market share has increased to about a third of the total online sales of Fresh and Frozen food in the US, showing that Instacart is a big player.

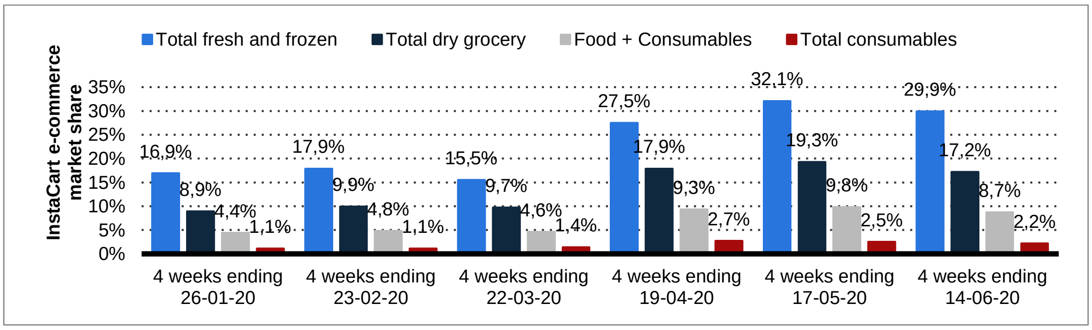


Figure 2 InstaCart e-commerce market share during the coronavirus pandemic in the US from January to April 2020

This change of habits in consumer spending has greatly boosted Instacart prospects. As with every rapid growth case, there´s opportunity alongside the risks. Instacart must keep its customers satisfied by taking care of the logistics surge, but also there is the chance to deliver such a good experience in grocery shopping through their app, that people might choose to continue shopping online for convenience, even after the pandemic is over.

2.2. Business Objectives

To further develop the customer experience, we have been asked to perform an analysis on a portion of the dataset, to try to come up with answers regarding: **(I) Which types of products should have an extended amount of product offerings? (II) Which types of products can be seen as substitutes? (III) Which items are complementary? (IV) What are the main types of consumer behavior in the business?**

[1] https://www.emarketer.com/content/online-grocery-sales-will-increase-by-nearly-53-this-year
[2] https://www.statista.com/statistics/1147639/instacart-market-share-by-category-us/

2.3. BUSINESS SUCCESS CRITERIA

Business success criteria in this project is not so easily defined. Our goal with this project is to give good and meaningful answers to Jane so she can better understand the market. But that information should be able to bring positive outcomes to the business.

With that in mind, to set a business success criterion, we should perform an A/B testing with users and see how sales react with and without the recommendation system that could be generated with insights we are going to bring through the market basket analysis and its respective association rules and the clustering solution for identifying purchase patterns. It would be considered a success, in our assessment, if sales in the test group rise by 10% or more than the placebo group.

2.4. SITUATION ASSESSMENT

Instacart already uses transactional data to understand which products a user is likely to buy again, try for the first time, or add to their cart next during a session. By analysing the dataset of 200,000 grocery orders from more than 100,000 Instacart users, we expect to bring some new information to the district manager Jane.

Our team is composed by 4 Data Scientists, which had to prepare a 5-minute presentation to the C-Suite, as well as a 10-Page Report (which you are reading) and the accompanying code for the market basket analysis and the identification of purchasing patterns at Instacart application and website.

2.5. DETERMINE DATA MINING GOALS

To have good understanding of the business, in particular consumer behavior habits, what products we need to have a wider amount of product offerings, what products are substitutes and which are complementary. Therefore, we are going to perform a market basket analysis, which is a rule-based machine learning approach that generates the relationship between variables in a dataset. Before applying the algorithm, preprocessing will have to be done to understand the data, link the different files into a *dataframe* that makes sense to the proposed business problems, perform some engineering with the features and analyze results.

The outcome of this analysis will be visualizations that will help us better understand consumer behavior as a whole, followed by lists of substitute and complementary products, achieved with the *Apriori algorithm*, as well as a cluster analysis with *kmeans* to understand different types of consumers, their buying habits and behavior. Additionally, we will test a second algorithm for a benchmark comparison of the *Aprori* implementation.

Regarding the cluster quality, we are setting as acceptable the quality of R2 at 0.5 or above. The lists of complementary and substitute products are deemed good if it is between 10 and 300 items, which seems a reasonable amount to be acted upon by the marketing and strategy teams.

2.6. PROJECT PLAN

1. Data understanding

2. Data preparation

   2.1. Pre-processing (connect different tables with scattered information into dataframes)

   2.2. Feature engineering (that will help us create better clusters)

3. Apply the *Apriori* algorithm & Assess through FP Growth Algorithm benchmark

5. Perform *kmeans clustering* on numeric variables

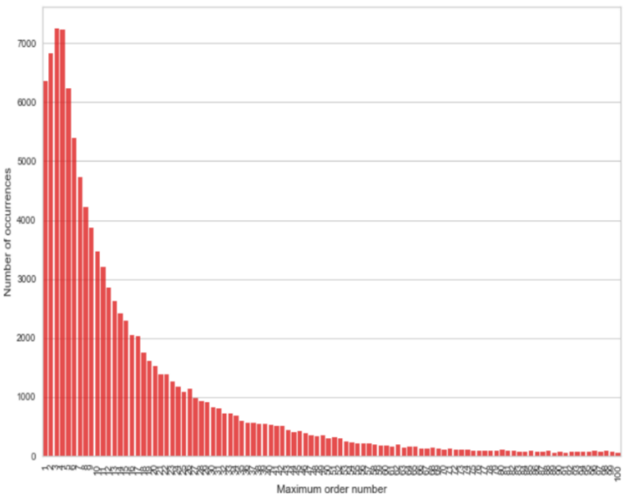
6. Extract lists and visualizations

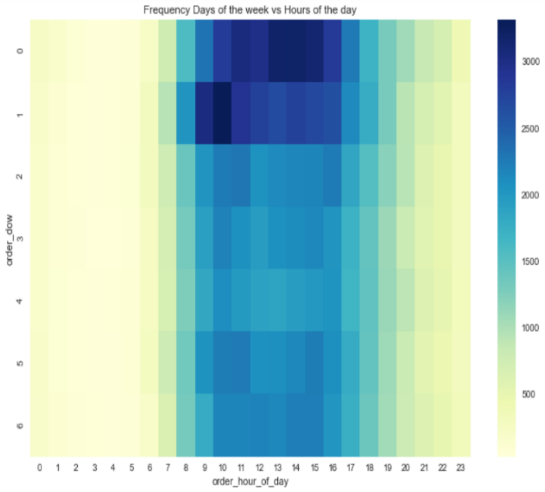
7. Deployment and maintenance

3. MARKET BASKET ANALYSIS AND CLUSTERING SOLUTION FOR PURCHASING PATTERNS

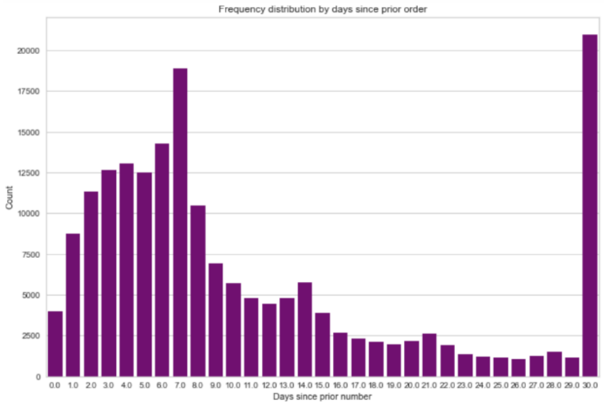
3.1 DATA AND OVERALL UNDERSTANDING OF ORDERS AND CUSTOMERS

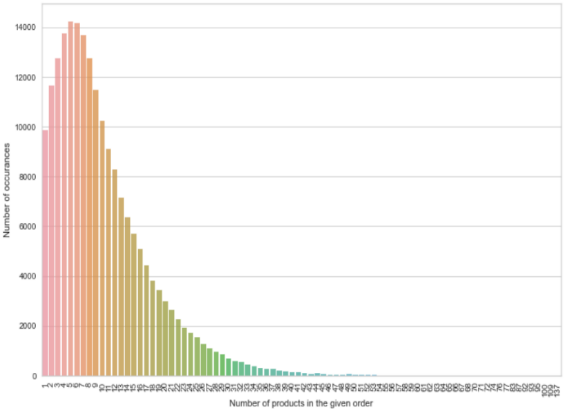
To better understand our data, we decided to do a series of simple visualizations (histograms, heatmaps, scatterplots and pie charts). We consider this to be a crucial part of this project because understanding the data will allow us to better judge our customers and their behaviors.





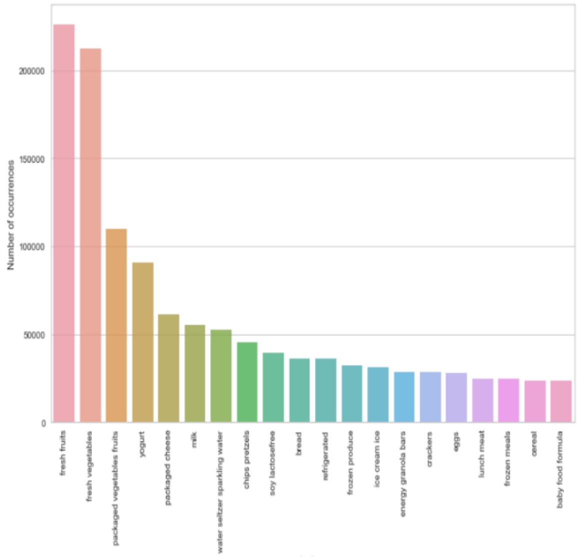
We can clearly see that most of our customers are new, after the 4th order frequency seems to fall, this may be because some type of promotion in the app ends after the 4th order. Most frequent days of the week to shop is day 0 and 1, and most common schedule is between 10 and 15. This information is re-assured on the previous heatmap we displayed above.

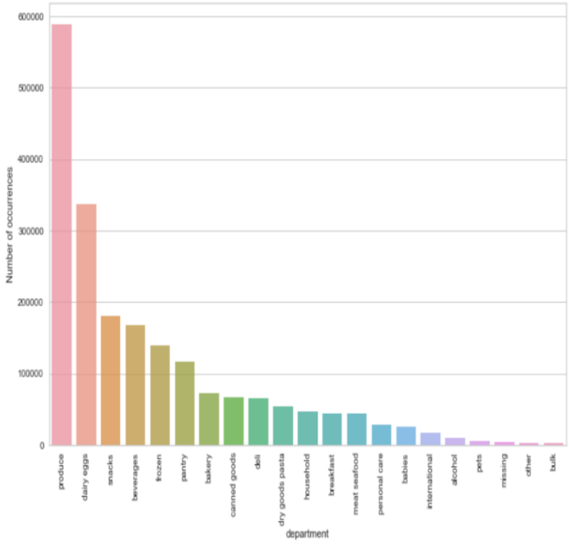


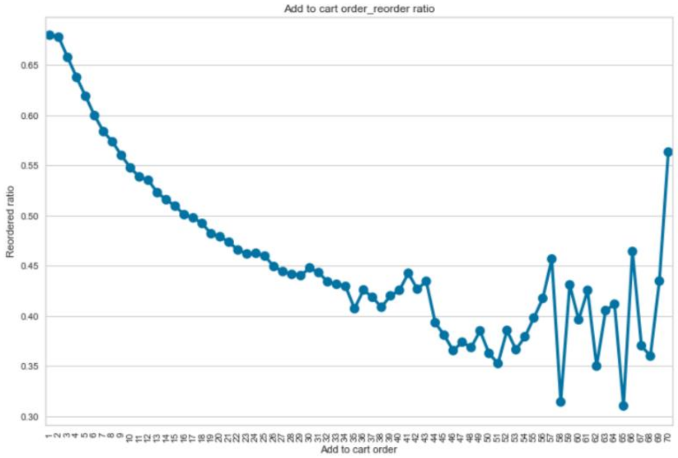


We can see that orders most likely occur before the end 1st week after the last purchase, this is important because it shows us that our costumers enjoyed both the services and the products provided by Instacart. There are spikes on days 7, 14, 21 and 30, the first 3 spikes are explained for weekly, bi-weekly and tri-weekly ordering recursion, the last spike is because it is counted as 30+ days, so we know that we have a lot of costumers that no longer use our services, probable churned customers.

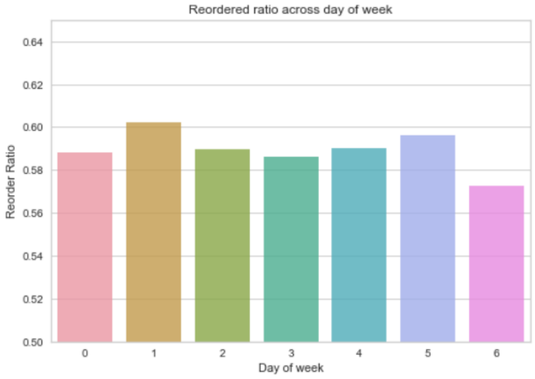
It is noticeable that most orders are small, averaging around 8 items per order, most probably with the fresh items that customers must refill. Costumers tend to buy in majority perishable goods, them being fruits, vegetables, eggs, etc. This information is going to be key for the market basket analysis.

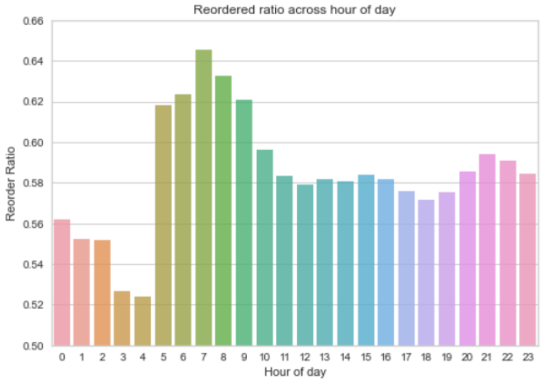






Here we can see that most orders start by adding products that costumer already ordered. This shows us some type of regularity and familiarity from costumers to our products.





We can see that in average at least half of the products in each order, regardless of the day an hour of the day, are products that the costumer already ordered before.

3.2. ALGORITHM SELECTION

We are using the *Apriori* Algorithm, mostly due to the familiarity of this approach when doing Market Basket Analysis. This algorithm will help us to better understand the clients' patterns and association between products. Besides that, we will reassess the algorithm with a benchmark algorithm, in order to evaluate different possibilities for maintenance phase.

We will first find the more frequent item-sets with a certain level of support, we chose 2%, and also some associations rules using different measures, like Confidence, Lift and Conviction.

Also, to answer the point regarding the type of purchasing patterns, we have we applied the already well-known K-means Clustering technique.

Finally, it is worth mentioning that we tested the FP Growth Algorithm. This technique is a more straight forward approach, but we did not end up using because the results did not differ enough for it to be worth using.

3.3. DATA PREPARATION

Our preprocessing steps for the *Apriori* Algorithm were pretty much merging the different data sources into certain dataframes for the purpose of our analysis.

We first merged into the dataframe "df" the variables from the excel files "order_products", "products" and "departments", dropping the id variables for the two latest files. Later we we polished "df" so we could get some information on the orders: the order that product was added to cart, a binary feature that indicated if that was or was not a reorder the product name and the consequent department.

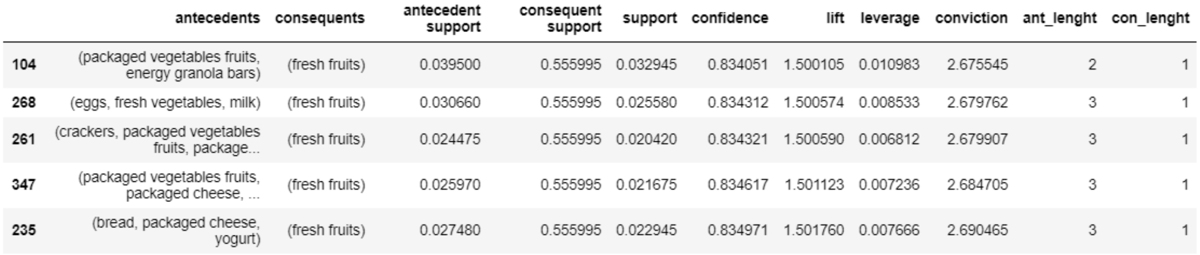
Then, we created a dataframe only for the product labels, and finally we merged the variables "order id" and "department" so we could have an idea of what departments were shopped in each order.

In relation to a K-means Clustering we did some feature engineering, creating the variables "number_of_items" and "%_of_reordered" that indicate the number of items per order and the percentage of products that were previously bought, respectively.

3.4. MODEL APPLICATION

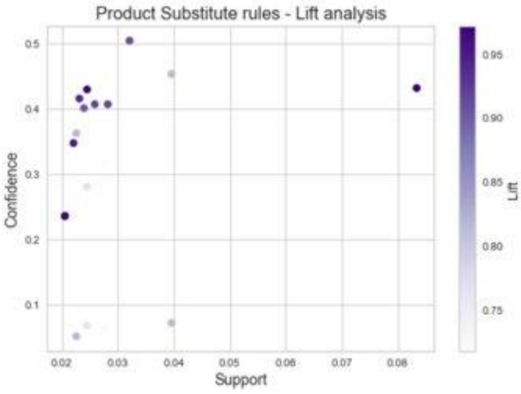
For our minimum support we decided to go with 2%, after trying with different levels of support, and backing our decision with the framework used by Sohaib Zafar Ansari, oriented by Fernando Bação on his Dissertation Thesis on Market Basket Analysis (support was 1%).

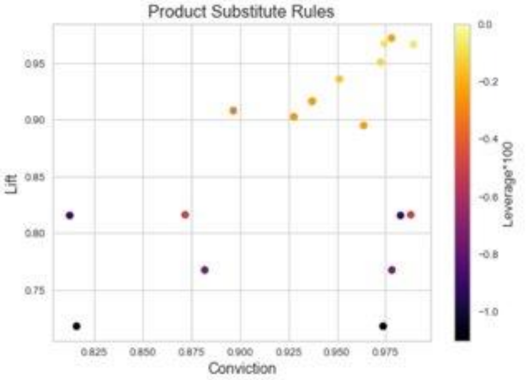
Through using association rules library and the output of the item-sets from the *Apriori Algorithm*, with lists of antecedents item-sets and consequent item-sets for those groups, applying the algorithm to get information on the support of both the antecedents and consequents alone and of them together, as well as support, confidence, lift, leverage, conviction and the number of item-sets in both the antecedents and consequents.



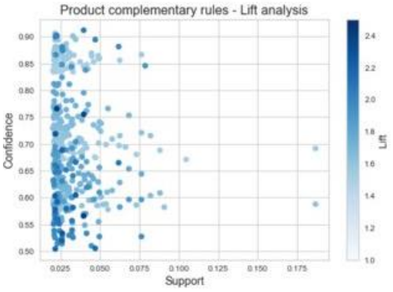
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	ant_lenght	con_lenght
104	(packaged vegetables fruits, energy granola bars)	(fresh fruits)	0.039500	0.555995	0.032945	0.834051	1.500105	0.010983	2.675545	2	1
268	(eggs, fresh vegetables, milk)	(fresh fruits)	0.030660	0.555995	0.025580	0.834312	1.500574	0.008533	2.679762	3	1
261	(crackers, packaged vegetables fruits, package...	(fresh fruits)	0.024475	0.555995	0.020420	0.834321	1.500590	0.006812	2.679907	3	1
347	(packaged vegetables fruits, packaged cheese, ...	(fresh fruits)	0.025970	0.555995	0.021675	0.834617	1.501123	0.007236	2.684705	3	1
235	(bread, packaged cheese, yogurt)	(fresh fruits)	0.027480	0.555995	0.022945	0.834971	1.501760	0.007666	2.690465	3	1

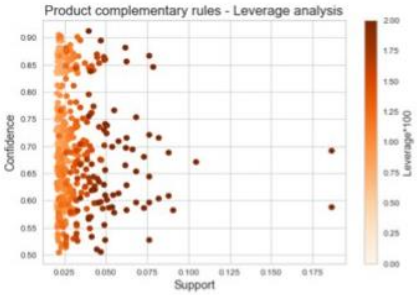
With this information we created our association rules for the relationships of the goods and consequent departments. For the substitute products we set a maximum Lift of 1.0 and a maximum conviction of 0.95, while for the substitute departments we keep our Lift limit of 1.0 but the conviction could reach a higher value of until 0.99.

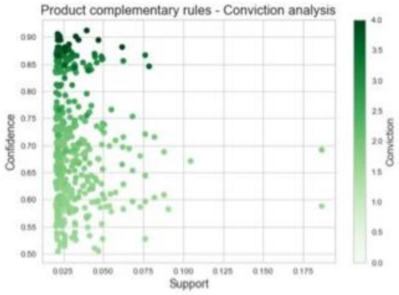




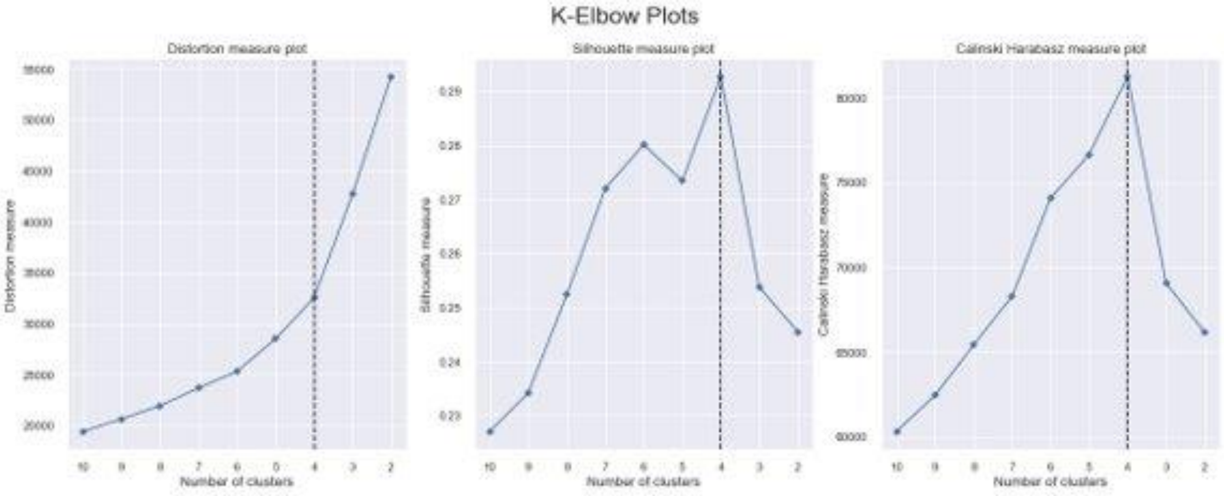
Regarding the complementary goods we considered the products with a confidence higher than 0.7 as well as a Lift and Conviction of over 1.5. For the complementary departments we kept the same levels of Lift and Conviction, but the Confidence had to be higher than 0.5 instead of 0.7.







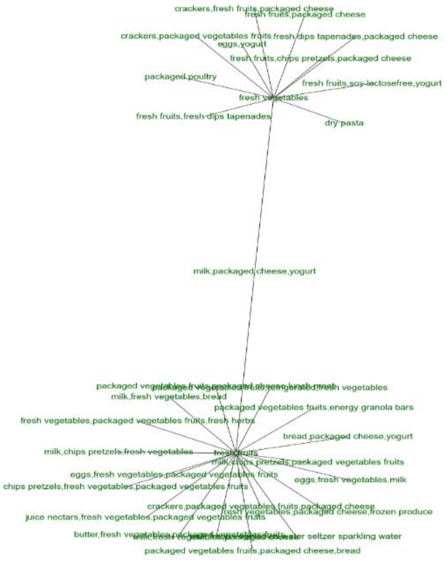
Regarding the K-means algorithm we used the variables "order_dow", "order_hour_of_day", "days_since_prior_order" and the previously created in our feature engineering "number_of_items" and "%_of_reordered". From a Distortion, a Silhouette and a Calinski Harrabasz measure plots, we concluded the optimal number of clusters is 4, with R2 of 0.53.



4. RESULTS EVALUATION

Through the usage of different data mining algorithms cited before we have successfully gathered information to answer the questions posed initially by the Account Manager, Jane Doe. The first 3 questions were answered with the use of *Apriori* algorithm, we will go through them one by one. The later one will be answered through a cluster implementation.

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


crackers,fresh fruits,packaged cheese
fresh fruits,packaged cheese
crackers,packaged vegetables fruits
fresh dips tapenades,packaged cheese
eggs,yogurt
fresh fruits,chips pretzels,packaged cheese
packaged poultry
fresh fruits,soy lactosefree,yogurt
fresh vegetables
fresh fruits,fresh dips tapenades
dry pasta
milk,packaged cheese,yogurt
packaged vegetables fruits,packaged cheese,lunch meat
packaged vegetables fruits,refrigerated,fresh vegetables
milk,fresh vegetables,bread
packaged vegetables fruits,energy granola bars
fresh vegetables,packaged vegetables fruits,fresh herbs
bread,packaged cheese,yogurt
milk,chips pretzels,fresh vegetables
fresh fruits
milk,chips pretzels,packaged vegetables fruits
eggs,fresh vegetables,packaged vegetables fruits
eggs,fresh vegetables,milk
chips pretzels,fresh vegetables,packaged vegetables fruits
crackers,packaged vegetables fruits,packaged cheese
fresh vegetables,packaged cheese,frozen produce
juice nectars,fresh vegetables,packaged vegetables fruits
butter,fresh vegetables,packaged vegetables fruits
water seltzer sparkling water
packaged vegetables fruits,packaged cheese,bread

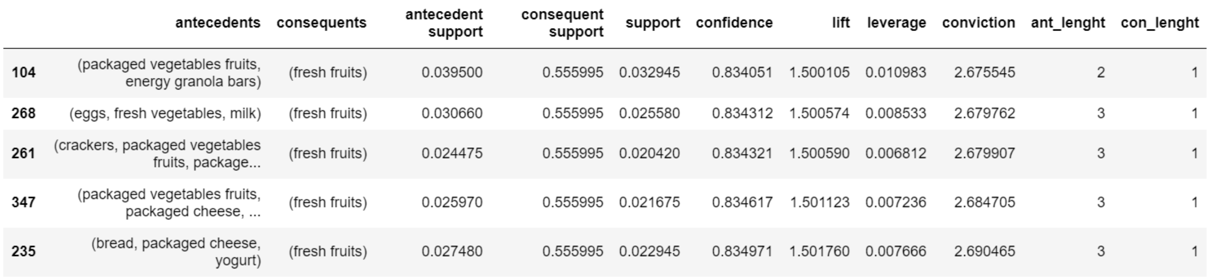
After analyzing the network (refer to notebook for better visualization), the most represented products are fresh ones. Therefore, our group considers that the best bet for Instacart is to extend product offering in exactly this area, in order to remain the top contender in this niche. Some of these categories of products are the following:

1. Fresh fruits
2. Packaged vegetables & fruits
3. Bakery
4. Eggs and milk derivatives

Which types of products can be seen as substitutes?

Substitute products can be one of two things, both products that replace one another and products that clients tend not to buy together.

Soft drinks and fresh products seem to not get well together, as expected. If a customer is buying soft drink,drinks he will not often buy fresh vegetables, fresh fruits or packaged vegetables fruits.

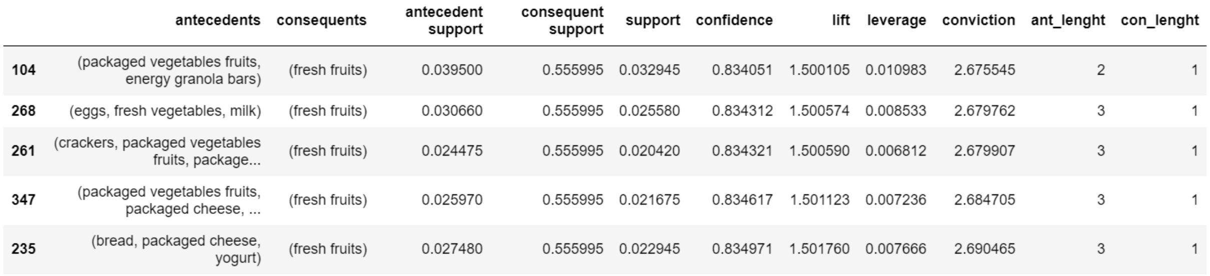


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	ant_lenght	con_lenght
104	(packaged vegetables fruits, energy granola bars)	(fresh fruits)	0.039500	0.555995	0.032945	0.834051	1.500105	0.010983	2.675545	2	1
268	(eggs, fresh vegetables, milk)	(fresh fruits)	0.030660	0.555995	0.025580	0.834312	1.500574	0.008533	2.679762	3	1
261	(crackers, packaged vegetables fruits, package...	(fresh fruits)	0.024475	0.555995	0.020420	0.834321	1.500590	0.006812	2.679907	3	1
347	(packaged vegetables fruits, packaged cheese, ...	(fresh fruits)	0.025970	0.555995	0.021675	0.834617	1.501123	0.007236	2.684705	3	1
235	(bread, packaged cheese, yogurt)	(fresh fruits)	0.027480	0.555995	0.022945	0.834971	1.501760	0.007666	2.690465	3	1

Which items are complementary?

Complementary items are the ones that are more likely to be seen together than separated.

Fresh fruits is the item-set that complements the most amount products, these other products seem to be other fresh products like packed vegetables eggs, milk, fresh vegetables, bread and yogurt.



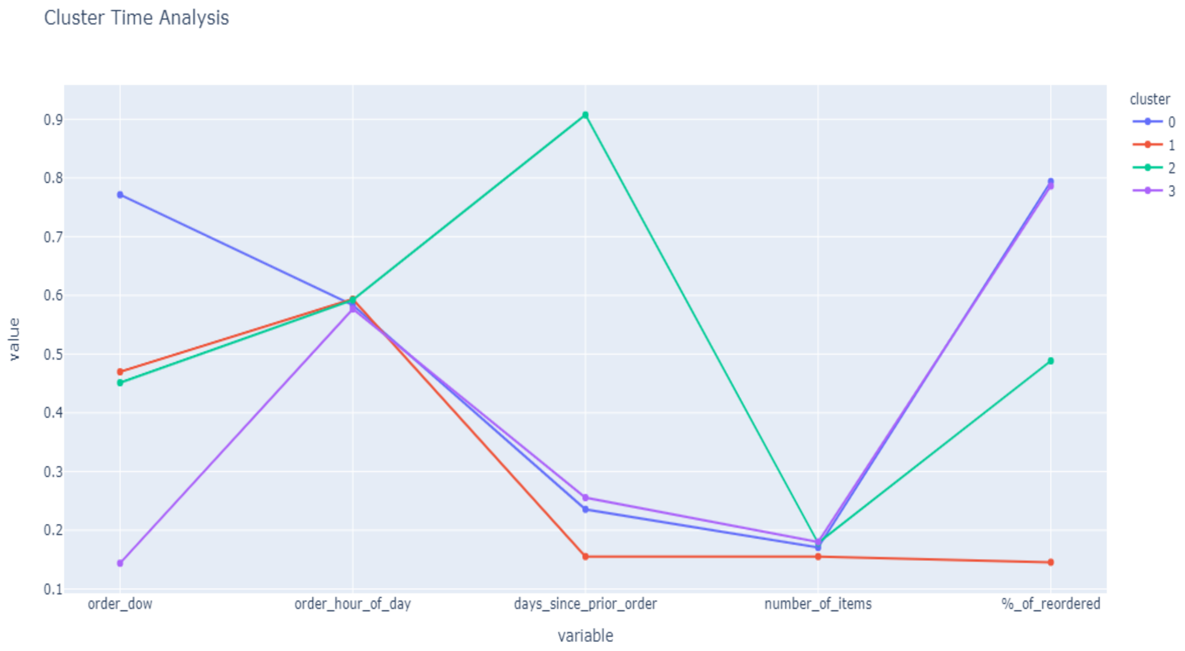
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	ant_lenght	con_lenght
104	(packaged vegetables fruits, energy granola bars)	(fresh fruits)	0.039500	0.555995	0.032945	0.834051	1.500105	0.010983	2.675545	2	1
268	(eggs, fresh vegetables, milk)	(fresh fruits)	0.030660	0.555995	0.025580	0.834312	1.500574	0.008533	2.679762	3	1
261	(crackers, packaged vegetables fruits, package...	(fresh fruits)	0.024475	0.555995	0.020420	0.834321	1.500590	0.006812	2.679907	3	1
347	(packaged vegetables fruits, packaged cheese, ...	(fresh fruits)	0.025970	0.555995	0.021675	0.834617	1.501123	0.007236	2.684705	3	1
235	(bread, packaged cheese, yogurt)	(fresh fruits)	0.027480	0.555995	0.022945	0.834971	1.501760	0.007666	2.690465	3	1

We also did association rules for substitute and complementary departments. The department that substitutes the highest number of other departments is the produce department and the ones that are most complementary are the dairy eggs and snacks.

The last question was answered with a clustering solution.

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

In general, we can say that the average costumer did 5 purchases, shops between 9 and 17, and mostly on days 0 and 1, he usually makes small purchases, around 8 items being at least half of them re-ordered. In an attempt of having a wider analysis we used a cluster solution and got the following 4 clusters:



Cluster 0 "*Weekend Usual Purchase*": these customers tend to buy medium amount of products, half of them being reordered, usually at the end of the week.

Cluster 1 "*First Try Purchase*": This are our new costumers, order at the middle of the week and since they are just trying out the app, they tend to order low number of products.

Cluster 2 "*Last Purchase*": Churned costumers that buy mostly during the middle of the week and buy in average a medium number of products.

Cluster 3 "*Monday Usual Purchase*": - Low to medium number of products, purchases are usually made at the beginning of the week and with a high reordering percentage.

5. CONCLUSION, DEPLOYMENT & MAINTENANCE

5.1. PLAN DEPLOYMENT

Overall, the scope of the Data Science team that performed the Market Basket Analysis and the Clustering solution for purchase patterns does not encapsulate the implementation within the Instacart's application.

Engineering team, that is, Front-End and Back-End development should jointly use the output of our analysis to think of how best use the insights and information provided above with the necessary framing of the usage of User Design and User Intelligence. o keep track of the customers' preferences a web application should be developed in order to track throughout time each customer's behaviour. Nowadays, the analysis performed through orders could use more information about each user and their purchase patterns more concatenated as well as socio-demographics of the customers.

5.2. MAINTENANCE

The application could have information about the products they order the most, if they vary in terms of brands, the order of each product when products to the cart, etc. Also, a recommendation system could be made so the store could better recommend products to the customers. This could be extremely helpful and time saving for the clients when they order goods that have good complementary products.

Also, we recommend that an analysis like this could be repeated throughout time. There are products that will always be consumed and are timeless, let us say, but there are also some products that have a floating interest according to trends. Trends can very much change the major preferences of the customers and call for a higher need to adapt the product line to the new needs. Having said, we believe it would be great to repeat this analysis once a year, for example.

This analysis could be made using the FP Growth Algorithm, as this method returns identical results to the *Apriori* Algorithm in a faster and easier way. Although, we cannot guarantee this would be the case every time for every dataset assessed. Despite that, we found that it was the case for the current Instacart dataset and we think there is room for improvement by using FP Growth Algorithm once the Account Manager inputs new data and reassess the MBA rules

6. Conclusion

Instacart provided us with 4 csv files that had information about orders, departments, products and how orders and products are related. With that information we were asked to provide value to the company by replying to some questions made by Jane Doe, Account Manager at Instacart. These questions were about costumer behavior, complementary products, substitute products and what products needed to more offerings.

Data understanding and preprocessing was a crucial part of this project since it gave us a lot of valuable information about our customers and their preferences. To do the market basket analysis we used an *Apriori* algorithm and to get a better understanding of purchasing patterns we used a K-means algorithm with 4 clusters.

With the usage of both algorithms, we were able to reply in a meaningful way to Jane's questions and we hope that with this valuable information, and with the implementation of cross-selling and product placement techniques, Instacart will increase their profits.

We provided recommendations on how Instacart should maintain this type of approach continuously for a better and most efficient performance, as well as a quick response to possible eventualities that may happen in the business throughout time.