MDSAA

Master Degree Program in Data Science and Advanced Analytics
- Major in Business Analytics

Market Basket Analysis

Group C

Catarina Moreira, number: 20201034

Luísa Barral, number: 20201045

Madalena Valério, number: 20200657

Yu Song, number: 20200572

April, 2021

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

INDEX

INTRODUCTION 3

BUSINESS UNDERSTANDING 3

- Background 3
- Business Objectives 3
- Business Success criteria 3
- Situation assessment 3
- Determine Data Mining goals 4

PREDICTIVE ANALYTICS PROCESS 4

- Data understanding 4
- Data preparation 7
- Modeling 8
- Evaluation 9

RESULTS EVALUATION 11

DEPLOYMENT AND MAINTENANCE PLANS 13

CONCLUSIONS 14

- Considerations for model improvement 14

REFERENCES 14

APPENDIX 15

1. INTRODUCTION

This project is a business case of practicing market basket analysis to find out deeper value in understanding the purchasing behaviours of customers with Instacart data set. Instacart is a grocery ordering and delivery service company in America. Our group applies clustering and association analysis in Python using the Apriori algorithm and data mining techniques.

2. BUSINESS UNDERSTANDING

2.1. Background

Instacart is a U.S. company that provides door-to-door grocery delivery services in the U.S. and Canada through a website or mobile app. The order is completed and delivered by a private buyer, who will select, pack and deliver the order in advance within the time frame specified by the customer. The company wants to make full use of their transaction data to understand users' shopping habits and buying patterns to provide a more delightful shopping experience.

2.2. Business Objectives

The company wants to raise their revenue by understanding their customers' shopping habits and improving the supply of goods based on its supply chain and customer's preference.

2.3. Business Success criteria

Our group should present a detailed overview of the company's business, especially in the topics of the main types of consumer behaviour, types of goods that can be extended to provide a variety of product choices, substitution types of products and complementary items. We are also required to present useful information in non-technical business ways with visualizations of the data set.

2.4. Situation assessment

Our team was given 4 csv files that contain basic information of each order with the sequence of products purchased. There are 2 CSV files,namely products.csv and departments.csv that specify the product has 134 types grouped in 21 departments. More specifically, order_products.csv and orders.csv contains which products were purchased, 'reordered' status, day of the week, hour of the day and days since previous order in each order.

2.5. Determine Data Mining goals

In Data Mining terms, we intend to find out segmentation of consumers to classify the customers by their purchase through clustering techniques. Our group used the variables ‘department’ and ‘product name’ to cluster the consumers, our goal is to find the appropriate types of shopping behaviours by comparing the results of different methods including K-Means, Gaussian Mixture, clustering, Mini Batch K-Means, and SOM.

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data understanding

We started by reading the description of each file from the relational dataset that describes the customers’ orders over time. We have a sample of 200.000 grocery orders from 105.273 Instacart users. Information about them is divided into 4 files: departments.csv, order_products.csv, orders.csv and products.csv that we loaded to pandas Dataframes. When analysing the tables we found missing values only in the “days_since_prior_order” variable. These were only for “order_number” equal to 1, therefore they are missing, because there can’t be days since the last order for the first order the client has made.

Following this, we began the Exploratory Data Analysis to get our initial insights of the data. We found that most clients shop between 10AM and 4PM as we can see from ***Figure 1***.

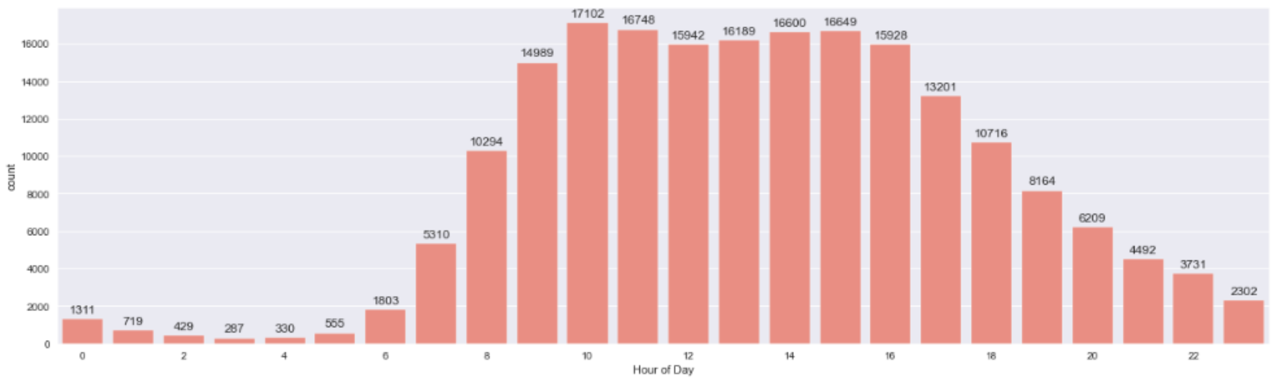


Figure 1 - Number of orders for each hour of the day

As for day of week, there isn't much difference, but there are slightly more orders on Sunday and Monday (***Figure 2***).

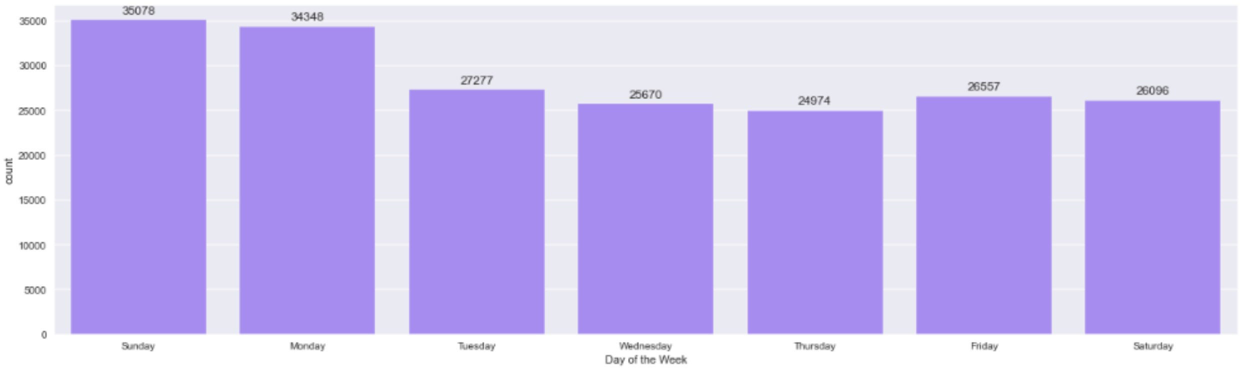


Figure 2 - Number of orders for each day of the week

For the variable "days_since_prior_order" it is clearly shown, in ***Figure 3***, that customers tend to go back to shopping after 7 or 30 days, which means Instacart customers are usually either weekly or monthly shoppers.

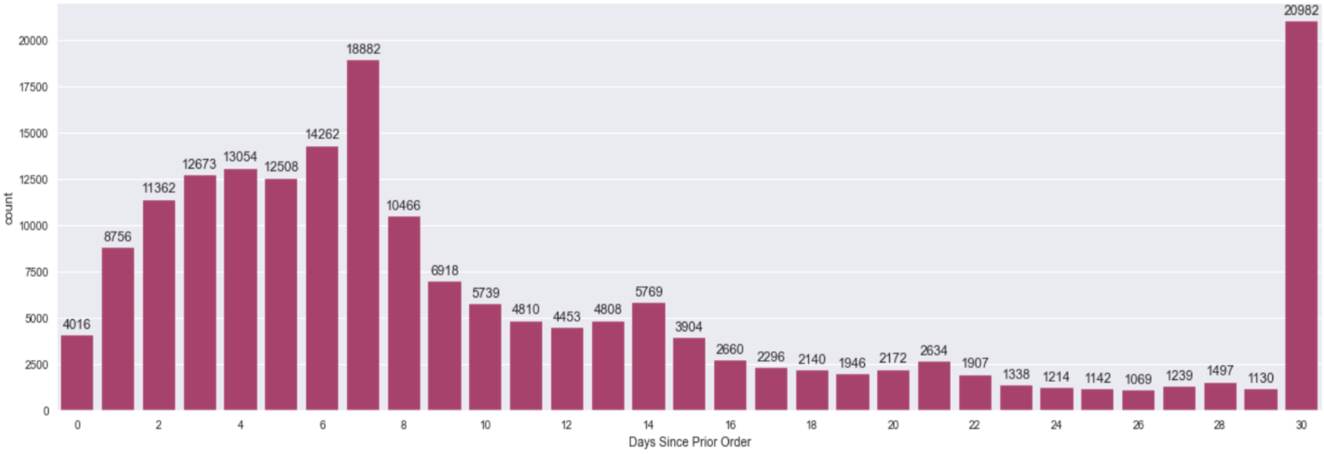


Figure 3 - Number of orders for each day since last order

We also wanted to know what were the overall most bought from departments and the most popular products, and as we can see from ***Figure 4***, Produce is by far the department that has most sold products, followed by Dairy Eggs and Snacks. ***Figure*** 5 shows us the top 30 most bought products, from which Fresh Fruits and Fresh Vegetables are the most popular.

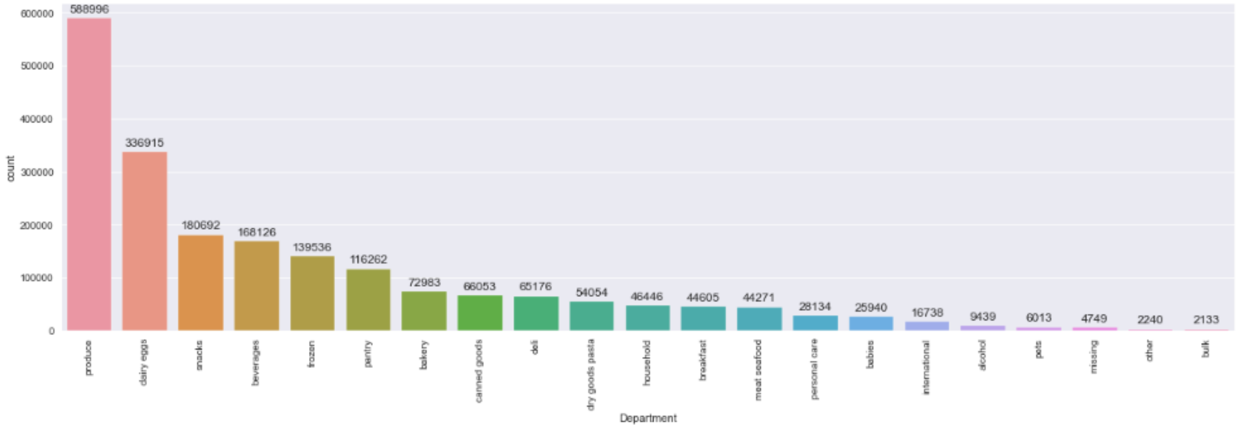


Figure 4 - Number of products each Department has sold

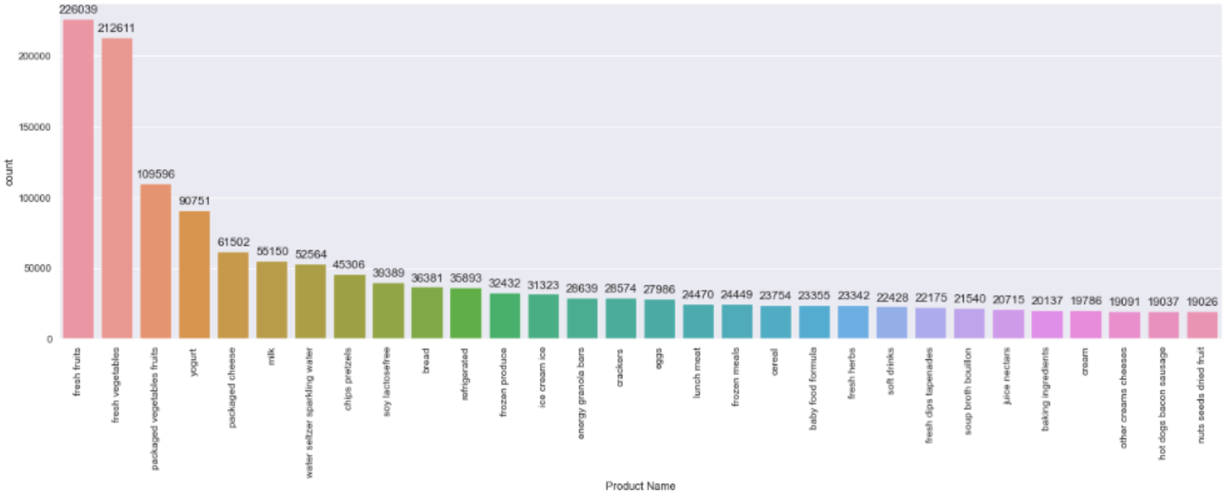


Figure 5 - Number of times each product was sold

When looking at the number of products each order contains, we discovered that most clients buy between 3 and 8 products (***Figure 6***).

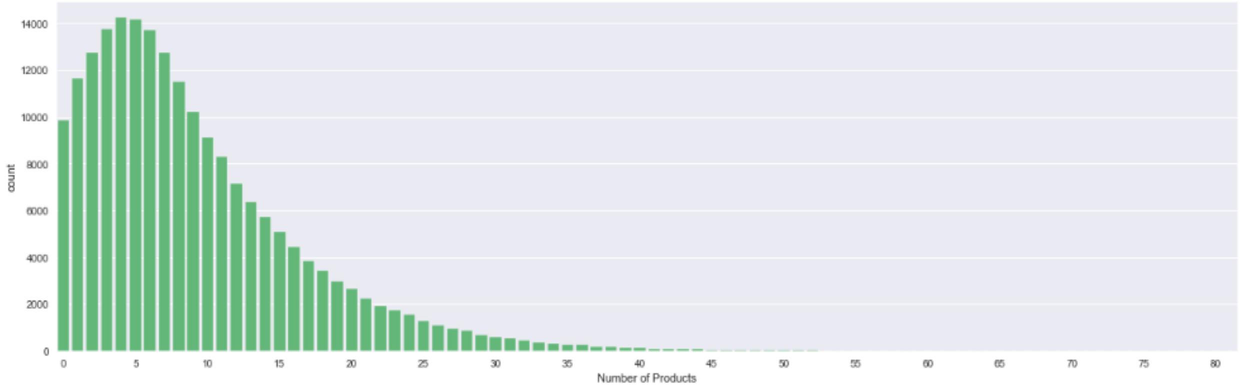


Figure 6 - Size of the orders

3.2. Data preparation

After doing some visualizations to understand the data better, we did the data preparation.

As we are dealing with a huge amount of data (observations), it’s really important to prepare correctly the data.

Initially, and after merging the four tables, we obtain a dataframe with 10 variables and 2019501 rows. During the preparation, we eliminated some variables as well as the missing values.

First of all, we started to check if there were missing values on the dataframe. As we could see, only the ‘days_since_prior_order’ had some Nan values. This column had 6% of Nan values in the dataset. Although this value is over 3%, we decided to eliminate it.

After eliminating the Nan values, we checked if the dataset had duplicate values that were not important. However, as we can see, it did not have any duplicated values. We were already expecting this result since we included the 'add_cart_order column. In this way and after analyzing the dataset, we decided to group by all columns except the ‘add_to_cart_order’ and have only the line with the max value of this last column.

After that, we started doing data exploration. Thus, we started by dividing the metric features and the non metric features and then we made a heat map with the metric features.

The purpose of making this matrix was to be able to detect the high-correlated variables.If two variables have a high correlation, this means that they are variables that give us redundant information and, in this way, we can eliminate one of them. In this category, we found that there were no redundant variables.

Lastly, we tried to solve the problems related to the quality of the data.

In this step, we started by drawing the boxplot for the metric features and we realized, then, that there were values that were significantly different from the other observations - outliers. In fact, we noticed that we had 9% of outliers but we decided not to remove them.

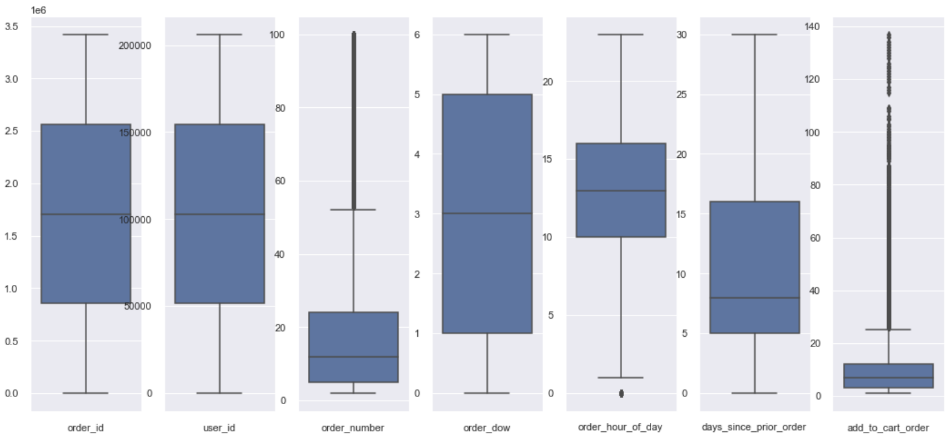


Figure 7 - Boxplot of the different metric features

Before doing the clustering, we dropped some columns that we believed that they were not important to our cluster algorithms. The columns that we eliminated were: 'days_since_prior_order', ‘order_dow’, and ‘reordered’.

3.3. Modeling

3.3.1. Clustering

For a better understanding of our customers behaviors, the group decided to apply different clustering algorithms and then analyse them to see each type of products, a certain cluster of customers , buys more.

Firstly we started by performing the One Hot Encoding, in order for us to use the categorical variables like “product_name” and “department”, and then we normalized the data.

The group applied the **K-Means + Hierarchical**, the **K-Means**, the **Gaussian Mixture**, clustering, **Mini Batch K-Means**, and **SOM.**

To group both departments (‘department’) and products (‘product_name’), we used K-Means + Hierarchical.

To group only the departments to draw some conclusions, we used the rest.

The first algorithm that we implemented was the **hierarchical cluster model with K-means**. Through the Dendrogram we took the number of clusters to use in the K-means algorithm.

After that, we implemented the **K-means** model. The **K-Means** is a clustering algorithm that tries to group in “K” groups the data points that are similar. The choice of “K” was made with the aid of the elbow plot and, therefore, we found that a good value to assign to “K” would be 4.

After **K-Means**, we used the **Gaussian Mixture** model. This model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

We also used the **Mini Batch K-Means**. Its main idea is to use small random batches of examples of a fixed size so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until convergence. As the number of iterations increases, the effect of new examples is reduced, so convergence can be detected when no changes in the clusters occur in several consecutive iterations.

Lastly, we applied the **Self-Organizing Maps(SOM).** SOM is a type of artificial neural network based on competitive learning. SOM is a technique to generate topological representations of data in reduced dimensions. With this technique we can explore some graphics and analyze and see the clusters and the idea is to iteratively adapt a connected two-dimensional matrix of vectors to the higher-dimensional topology of the input dataset. At each cycle, a node is selected and its elements (the weights) are updated, together with those of its neighbors, to approach a randomly chosen datapoint from the training set. The competitive element comes into play during the update stage, since the closest node to the extracted datapoint is selected for the weights update at each iteration.

To conclude, we use the K-means to analyze customers' behaviour.

3.3.2. Association Analysis

After the clustering we began to analyse the associations between products. For this, it was necessary to consolidate the items into 1 order per row with each product one hot encoded. With the data structured properly we generated frequent itemsets with the Apriori algorithm and defined Support to be at least 5%, this gave us 156 itemsets. The last step was to generate the Association Rules. For each rule we have the Antecedent, Consequent, Support, Confidence and Lift. Support is the relative frequency that the rule shows up, Confidence is a measure of the reliability of the rule, in other words it's the probability of finding the Consequent of the rule in transactions given that these transactions also contain the Antecedent, and Lift is a factor by which the likelihood of Consequent increases given an Antecedent.

We tested the rules with different metrics (Confidence and Lift) and minimum thresholds to understand which products have the strongest associations and to later define which are Complementary and Substitutes.

3.4. Evaluation

As we said before, we used the K-Means with the Hierarchical clustering to group both departments and products. Through this analysis, we could make a better exploration and we could analyse the different people's behaviour. We grouped into 4 groups (we used 4 clusters) and we are going to describe them better in the next point.

As we affirmed, we used the Dendrogram and we took the number of clusters to use in the K-means algorithm. In the figures below, we presented the dendrogram and the clusters.

After applying this method, we applied other clustering models only in the departments. In this way, although it is not such a detailed description and exploration, we were able to draw some conclusions and clustering analysis.

With respect to the Association analysis, after analysing the relations between products we noticed that, when generating the rules with Lift as the metric and with a minimum threshold of 0 we can see, from ***Figure 8***, that those with low Support are also strong in terms of Confidence and Lift, therefore they shouldn't be discarded.

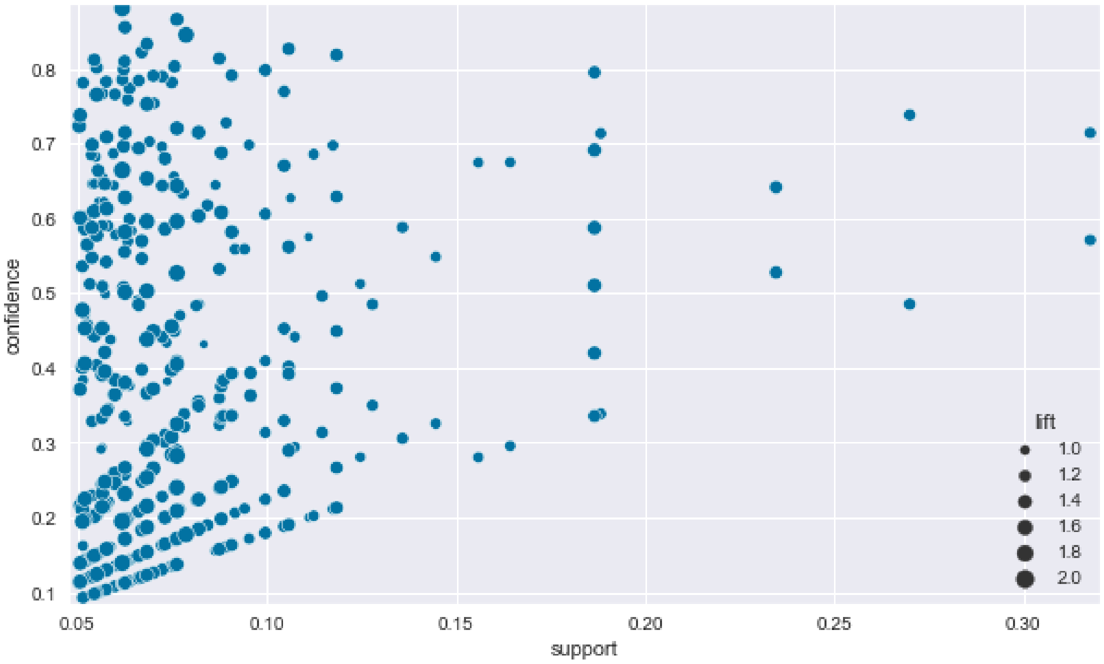
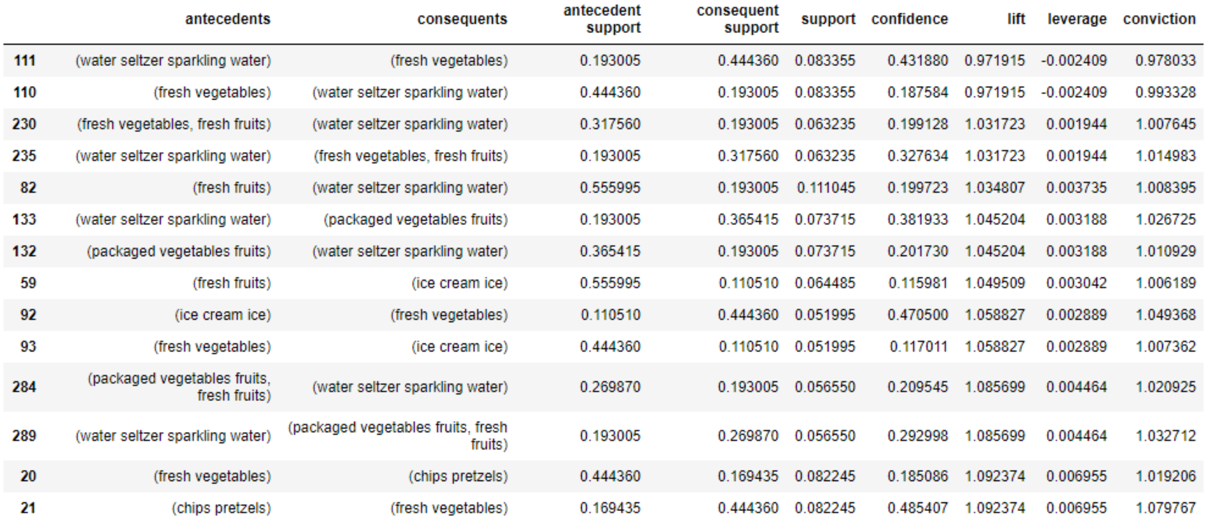


Figure 8 - Scatterplot of Association Rules

With further analysis we concluded that we mostly have rules with Lift larger than 1, which means that there is a positive correlation within the itemset, that is, the items are more likely to be bought together. Despite this, Lift has relatively low values.

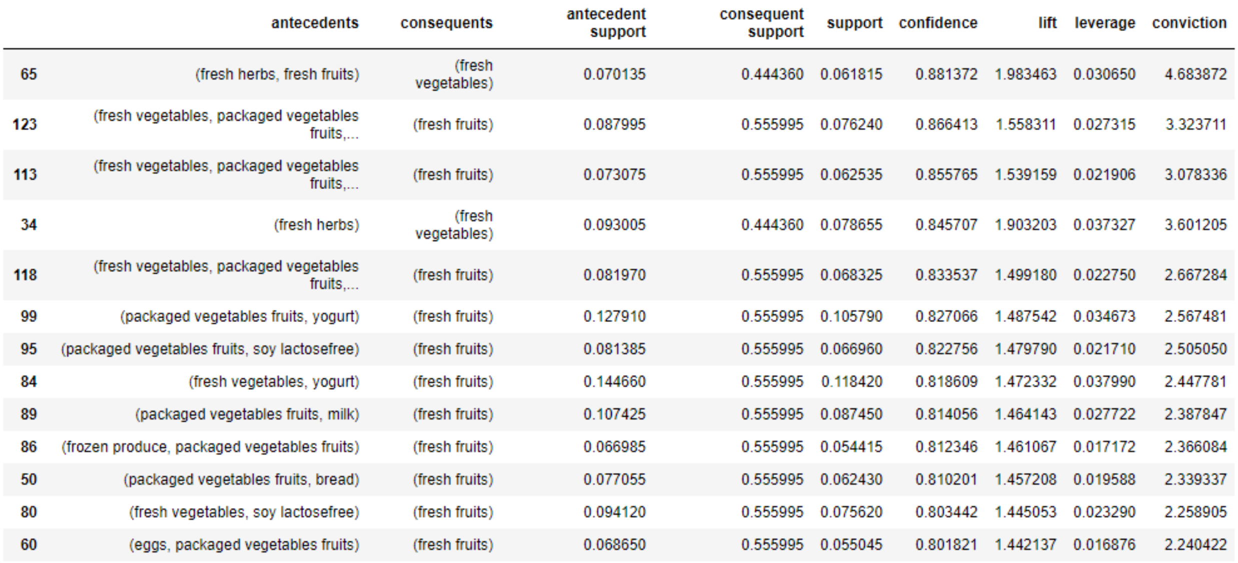
To find substitute products, we only selected the rules that have Confidence below 50% and Lift below 1.1. Usually only lift <1 can tell us if items are substitutes, however we only have 2 rules that have those values. Therefore we decided to have a bigger lift that gave us more substitute products to work with.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
111	(water seltzer sparkling water)	(fresh vegetables)	0.193005	0.444360	0.083355	0.431880	0.971915	-0.002409	0.978033
110	(fresh vegetables)	(water seltzer sparkling water)	0.444360	0.193005	0.083355	0.187584	0.971915	-0.002409	0.993328
230	(fresh vegetables, fresh fruits)	(water seltzer sparkling water)	0.317560	0.193005	0.063235	0.199128	1.031723	0.001944	1.007645
235	(water seltzer sparkling water)	(fresh vegetables, fresh fruits)	0.193005	0.317560	0.063235	0.327634	1.031723	0.001944	1.014983
82	(fresh fruits)	(water seltzer sparkling water)	0.555995	0.193005	0.111045	0.199723	1.034807	0.003735	1.008395
133	(water seltzer sparkling water)	(packaged vegetables fruits)	0.193005	0.365415	0.073715	0.381933	1.045204	0.003188	1.026725
132	(packaged vegetables fruits)	(water seltzer sparkling water)	0.365415	0.193005	0.073715	0.201730	1.045204	0.003188	1.010929
59	(fresh fruits)	(ice cream ice)	0.555995	0.110510	0.064485	0.115981	1.049509	0.003042	1.006189
92	(ice cream ice)	(fresh vegetables)	0.110510	0.444360	0.051995	0.470500	1.058827	0.002889	1.049368
93	(fresh vegetables)	(ice cream ice)	0.444360	0.110510	0.051995	0.117011	1.058827	0.002889	1.007362
284	(packaged vegetables fruits, fresh fruits)	(water seltzer sparkling water)	0.269870	0.193005	0.056550	0.209545	1.085699	0.004464	1.020925
289	(water seltzer sparkling water)	(packaged vegetables fruits, fresh fruits)	0.193005	0.269870	0.056550	0.292998	1.085699	0.004464	1.032712
20	(fresh vegetables)	(chips pretzels)	0.444360	0.169435	0.082245	0.185086	1.092374	0.006955	1.019206
21	(chips pretzels)	(fresh vegetables)	0.169435	0.444360	0.082245	0.485407	1.092374	0.006955	1.079767

Table 1 - Substitute Products

On the other hand, to find complementary products we selected the rules with confidence above 80% and Lift above 1.4, so that we can have only the very top products that are usually bought together, for a simpler analysis.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
65	(fresh herbs, fresh fruits)	(fresh vegetables)	0.070135	0.444360	0.061815	0.881372	1.983463	0.030650	4.683872
123	(fresh vegetables, packaged vegetables fruits,...	(fresh fruits)	0.087995	0.555995	0.076240	0.866413	1.558311	0.027315	3.323711
113	(fresh vegetables, packaged vegetables fruits,...	(fresh fruits)	0.073075	0.555995	0.062535	0.855765	1.539159	0.021906	3.078336
34	(fresh herbs)	(fresh vegetables)	0.093005	0.444360	0.078655	0.845707	1.903203	0.037327	3.601205
118	(fresh vegetables, packaged vegetables fruits,...	(fresh fruits)	0.081970	0.555995	0.068325	0.833537	1.499180	0.022750	2.667284
99	(packaged vegetables fruits, yogurt)	(fresh fruits)	0.127910	0.555995	0.105790	0.827066	1.487542	0.034673	2.567481
95	(packaged vegetables fruits, soy lactosefree)	(fresh fruits)	0.081385	0.555995	0.066960	0.822756	1.479790	0.021710	2.505050
84	(fresh vegetables, yogurt)	(fresh fruits)	0.144660	0.555995	0.118420	0.818609	1.472332	0.037990	2.447781
89	(packaged vegetables fruits, milk)	(fresh fruits)	0.107425	0.555995	0.087450	0.814056	1.464143	0.027722	2.387847
86	(frozen produce, packaged vegetables fruits)	(fresh fruits)	0.066985	0.555995	0.054415	0.812346	1.461067	0.017172	2.366084
50	(packaged vegetables fruits, bread)	(fresh fruits)	0.077055	0.555995	0.062430	0.810201	1.457208	0.019588	2.339337
80	(fresh vegetables, soy lactosefree)	(fresh fruits)	0.094120	0.555995	0.075620	0.803442	1.445053	0.023290	2.258905
60	(eggs, packaged vegetables fruits)	(fresh fruits)	0.068650	0.555995	0.055045	0.801821	1.442137	0.016876	2.240422

Table 2 - Complementary Products

4. RESULTS EVALUATION

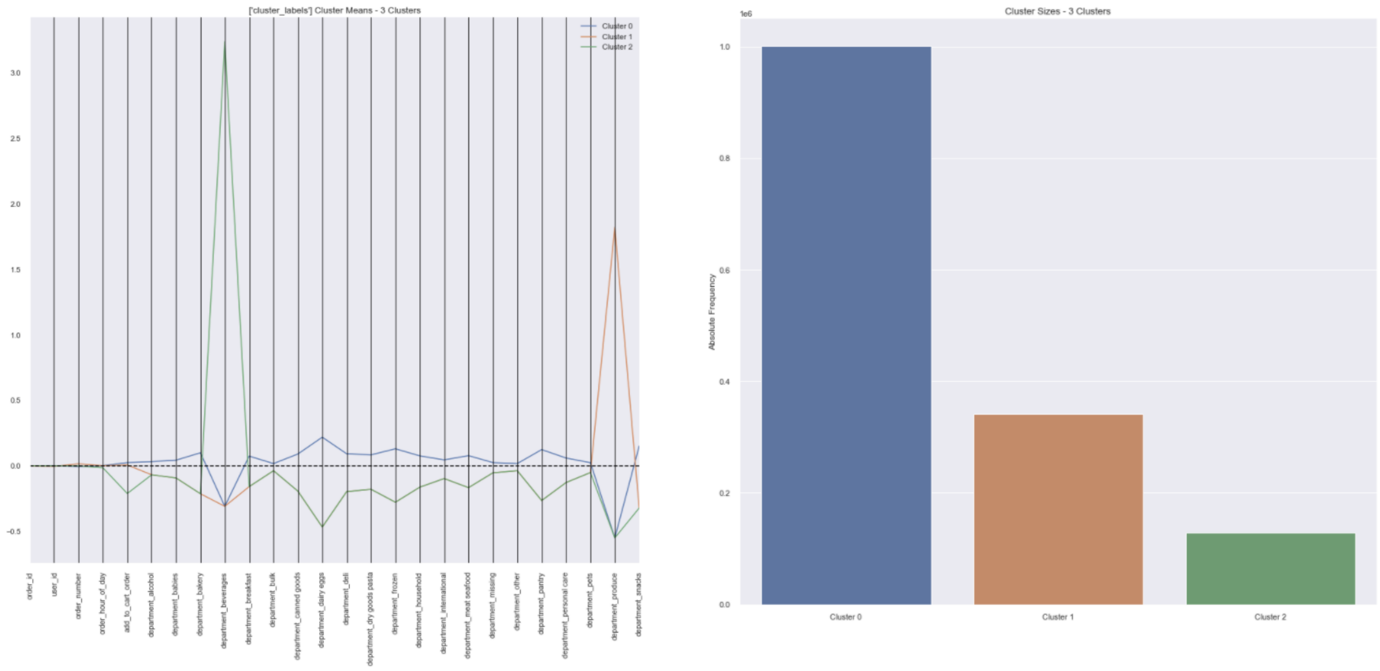


Figure 9 - Clusters from K-means algorithm

We found three clusters of shopping behaviour of the Instacart consumers.

Cluster 0: This type of customer buys most frequently frozen food such as frozen meals, dessert and ice cream. They also buy snacks, but have a low consumption of alcohol or beverages, baby products and fresh food.

We could see them as people that pay less attention to cooking with fresh food because they buy a lot of frozen food and snacks instead.

Cluster 1: The customers in this category are more likely to buy alcohol and beverages, fresh fruits and vegetables, meat, seafood, pasta and pet food. They buy fewer dairy products and eggs, snacks, frozen foods, pantry, breakfast, among few others

We could see this kind of clients as people who like to have a full dispensary so they are fond of alcohol and beverages, various fresh food products and pets products.

Cluster 2: Customers in this category are more likely to buy bakeries, breakfast, canned food, dairy and eggs, household products, personal care and baby products. They made fewer purchases of alcohol, beverages, frozen foods, meat and seafood, fresh vegetables and snacks.

We could see them as young families with babies who prepare breakfast themselves with dairy, eggs and pantry food. They pay more attention to bakery, personal care, imported food and deli.

So, after doing the clustering and the customers' behavior analysis, we began to analyze the associations between products. Thus, after implementing rules, we were able to find the substitute and complementary products.

The substitute products that we obtained are (***Figure 10***):

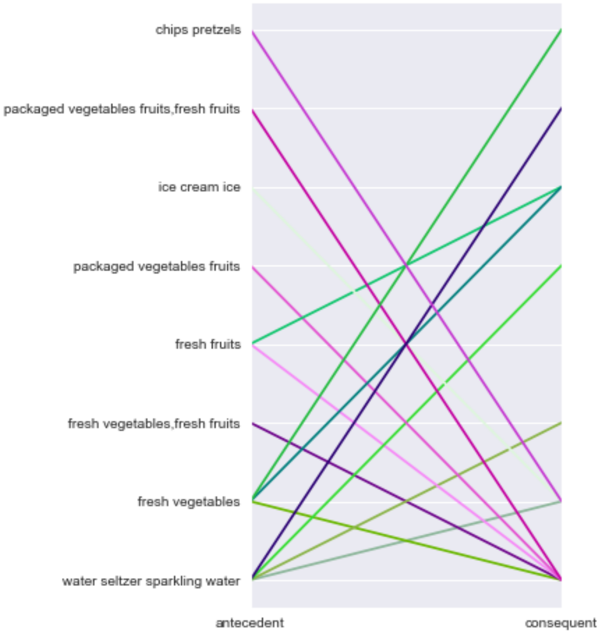


Figure 10 - Substitute products

The complementary products are (***Figure 11***):

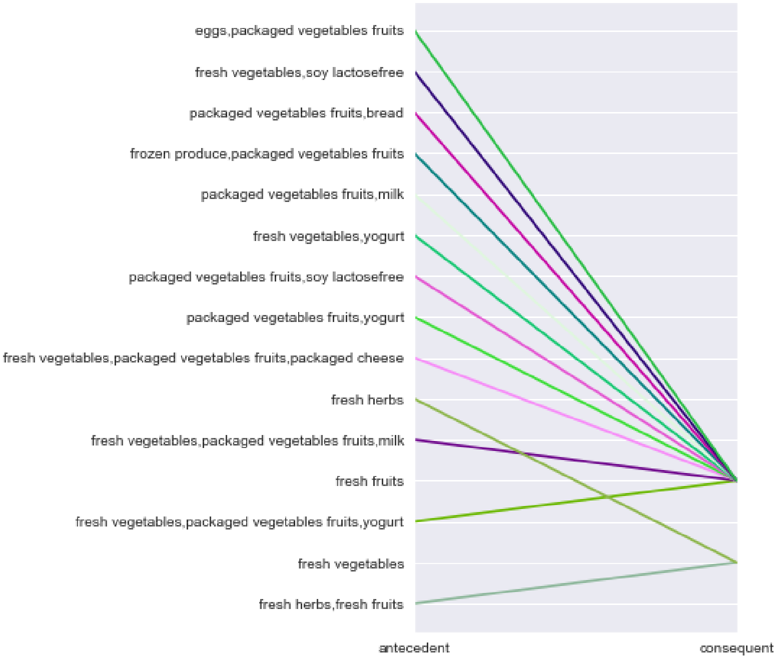


***Figure 11* -** Complementary products

After the market basket analysis, we tried to create cross-selling opportunities in order to sell items that aren't so popular for Instacart users, and that should have an extended amount of product offerings. First, we created a *Hygiene Basket* that includes Skin Care Products, Eye Ear Care Products, Facial Care Products, Hair Care Products and Body Lotion Care Products. Then we created an *Ice Cream Promotion,* with this promotion, Instacart will be able to boost their sales of ice cream topping products by selling them alongside ice creams at a lower price. After that, we created a *Salad Promotion*, that will boost the consumption of Prepared Salads and Dressing Toppings by selling them together at lower prices. Lastly, we developed the *Healthy Basket* idea, since Fruits and Vegetables are popular items, that includes the Bulk dried Fruits Vegetables, Canned Fruit Applesauce, Fruit Vegetable Snacks, Nuts Seeds Dried Fruit, where the products can all be bought as a basket or in pairs.

5. DEPLOYMENT AND MAINTENANCE PLANS

For the deployment and maintenance of the Instacart market basket analysis, the group did some investigation, in order to find the best tool for our presenting and updating our recommendations to the shoppers. We discovered Saturn Cloud that provides the data scientists with all the necessary tools for seamless collaboration, effortlessly scalable compute resources, and easy analytics, without the need of specialized DevOps. The data scientists can work within the Jupyter Notebook which is hosted on a specified server created by the system.

After data pre-processing we suggest the implementations of certain functions to assist on the main association rules function like frequency, that counts for items and item pairs, order_count, that returns og unique orders. After this the main association rule function should be implemented.

Using Jupyter Notebook in the cloud is very intuitive with Saturn Cloud, once a notebook is running, we can easily share it from within the notebook, with the public. Hosting Jupyter Notebooks with Saturn Cloud while also taking care of versioning and the ability to scale in or out as needed can tremendously simplify the life of our data scientists and shareholders because it decreases our time to market, decreases cost and the need for expert cloud skills.

As for the maintenance of our model, Saturn Cloud will also help us with that through access monitoring, intrusion detection, environmental concerns, and asset management. The Saturn Cloud application, our Jupyter Notebook, will be accessible via internet through an Internet Gateway.

6. CONCLUSIONS

In conclusion, there was a lot of information extracted from our Market Basket Analysis, such as the most bought products, the hour of the day in which people usually go to the supermarket, the 3 main behaviours displayed by our customers, the substitute and complementary products and some extended product offerings. With use of clustering algorithms, we were able to analyse, through different methods, the main 3 behaviours our customers display when going shopping. When resorting to the association rules, we hoped to find the confidence and lift levels that would tell us which products are substitutes and which ones are complementary. We then concluded by using all this information extracted and applying it in selecting which products to promote by extending product offerings, pair them or creating baskets containing similar products.

6.1. Considerations for model improvement

Firstly, the group suggests an investment on the data collection process, so we can have more details about our customers and their favorite products, like the name of the product so we can suggest more customized promotions. For model improvement, the group suggests a more deep understanding of the problem and an even better and more detailed suggestion of products. Deep Learning might be a good tool to apply in this problem.

7. REFERENCES

Moffitt, C. (2017, July 3). *Introduction to Market Basket Analysis in Python*. Retrieved from Practical Business Python: https://pbpython.com/market-basket-analysis.html

Chauhan, N. S. (n.d.). *Market Basket Analysis: A Tutorial*. Retrieved from KDnuggets: https://www.kdnuggets.com/2019/12/market-basket-analysis.html

Simmons, J. (2020, July 27). *Market Basket Analysis in Python*. Retrieved from SROSE: https://www.srose.biz/research-analysis/market-basket-analysis-in-python/

Implementing Self-Organizing Maps with Python and TensorFlow. (2018, August 27). Retrieved from Rubik's Code: https://rubikscode.net/2018/08/27/implementing-self-organizing-maps-with-python-and-tensorflow/

Otieno, D. (2019, December 31). *Overview of Self Organizing Maps (SOM) with its python implementation in determining safe airlines over time.* Retrieved from DaliCodes: https://dalicodes.medium.com/overview-of-self-organizing-maps-som-with-its-python-implementation-in-determining-safe-airlines-db8f6018a2b

sklearn.cluster.DBSCAN. (n.d.). Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Harman, M. (2020, August 27). *DBSCAN with Python*. Retrieved from Medium: https://towardsdatascience.com/dbscan-with-python-743162371dca

8. APPENDIX

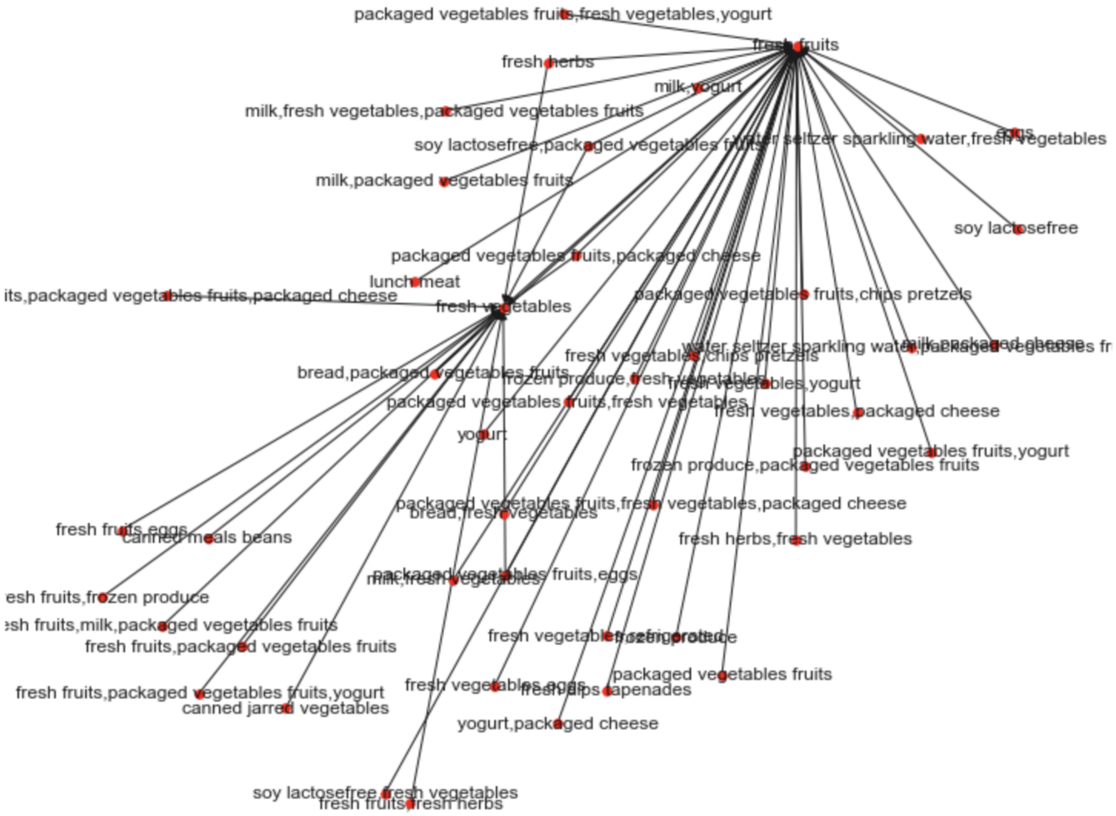

packaged vegetables fruits,fresh vegetables,yogurt
fresh fruits
fresh herbs
milk,yogurt
milk,fresh vegetables,packaged vegetables fruits
soy lactosefree,packaged vegetables fruits
water seltzer sparkling water,fresh vegetables
eggs
milk,packaged vegetables fruits
soy lactosefree
packaged vegetables fruits,packaged cheese
lunch meat
its,packaged vegetables fruits,packaged cheese
fresh vegetables
packaged vegetables fruits,chips pretzels
water seltzer sparkling water,packaged vegetables fr
milk,packaged cheese
fresh vegetables,chips pretzels
bread,packaged vegetables fruits
frozen produce,fresh vegetables
fresh vegetables,yogurt
packaged vegetables fruits,fresh vegetables
fresh vegetables,packaged cheese
yogurt
packaged vegetables fruits,yogurt
frozen produce,packaged vegetables fruits
packaged vegetables fruits,fresh vegetables,packaged cheese
bread,fresh vegetables
fresh fruits,eggs
canned meals beans
fresh herbs,fresh vegetables
packaged vegetables fruits,eggs
milk,fresh vegetables
esh fruits,frozen produce
esh fruits,milk,packaged vegetables fruits
fresh fruits,packaged vegetables fruits
fresh vegetables,refrigerated
frozen produce
packaged vegetables fruits
fresh fruits,packaged vegetables fruits,yogurt
canned jarred vegetables
fresh vegetables,eggs
fresh dips tapenades
yogurt,packaged cheese
soy lactosefree,fresh vegetables
fresh fruits,fresh herbs

***Figure 11* -** Directed Network Graph of the top 50 confidence rules