Market Basket Analysis

University of connecticut school of business |

Market BASKET ANALYSIS FOR INSTACART

chAkradhar rajineni

2020

**Table of Contents**

**INTRODUCTION:**

Instacart is an on-demand grocery delivery service found in 2012 that helps customers shop online from local grocery stores by providing a 'personal shopper' to fulfill and deliver the order in as little as an hour. One of the main differences between Instacart and other delivery services is that Instacart is not a grocery store; it shops and delivers the products fresh from the stores that customers choose and it does not pile up on products like other services.

**How it works:**

Customer signs up => Chooses the store of choice => adds products to the cart => selects the delivery date and time window => provides address, contact number, card details, and places the order => Personal shopper shops and delivers the order

Instacart also provides customers with additional options like:

- adding notes to the products in the cart

- preferences to select an option for 'If out of stock, replace with...'

**Instacart Process – A Brief look:**

- The store options customers are provided with depends on the customer location

- Customer can choose to either get the items picked-up or delivered depending upon the options

- Once the order is placed, an initial hold will be placed on the customer's card for an amount that is slightly higher than the order amount

- Instacart assigns a personal shopper sometime before the chosen delivery window

- Once the personal shopper is assigned, the name and contact details of the personal shopper are updated.

- The application also allows the customer to communicate with the personal shopper via the application, mobile (call), and text messaging.

- Until the personal shopper starts purchasing, Instacart allows customers to modify the order details. Once the personal shopper is assigned, the customer will no longer be able to modify the

order in the application but can choose to do so by communicating with the personal shopper and the order amount will be adjusted accordingly

- If the personal shopper is unable to find the item, the amount of that item will be refunded.

- When selecting substitutes for items that are out of stock, personal shopper acts based on customer's instructions. If the customer's choice is to substitute with back-up items, then the personal shopper will choose a different brand that's similar to the one customer wanted, and customer might end up paying more or less depending on the chosen option. On the other hand, if customer requests no substitutes, then the item will ne be purchased.

- Amount is debited from the customer’s account after the order is delivered.

- The final amount debited from the customer's account will be based upon the items that are delivered finally

**LITERATURE**

**Strategies used by e-commerce industry to recommend products:**

E-commerce industry uses several methods to recommend products and some of them are given below:

1. Displaying a list of suggested products based on the visitor’s browsing history
2. Recommendations based on “Frequently bought together”
3. Using product recommendation engines to personalize email campaigns.
4. Showing “Related to items you’ve viewed” suggestions
5. Recommendations based on “Customers who bought [this item] also bought [that item]”
6. Providing recommendations by showing items related to previous purchases
7. Featuring best-selling items across brands

**Market Basket Analysis:**

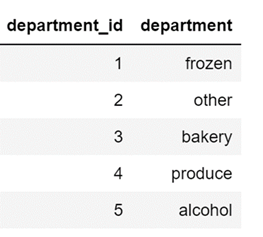
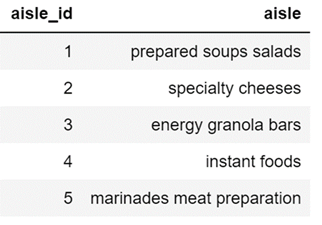
Market basket analysis is a modeling method that helps in identifying the products that will be purchased together. In simple terms, if a customer buys a product, what is the probability that he/she will buy another product along. The purpose of market basket analysis is to increase sales by identifying the products bought together by customers. Based on this prediction a set of recommendations can be displayed on the e-commerce website. This analysis helps to increase sales by recommending appropriate products to the customers who are likely to buy. It improves user experience by providing specific recommendations and reduces the time spent by the customers on product search.

**Our Work:**

We are using Market Basket Analysis to find combinations of products bought together and then associate these combinations with appropriate customers. We are also segmenting customers into relevant segments that will provide better insights and improve the associations. Combining Market Basket Analysis with Customer Segmentation will help us in providing better recommendations to the customers.

**2 Data Dictionary and Description**

**2.1 Data Dictionary**

****

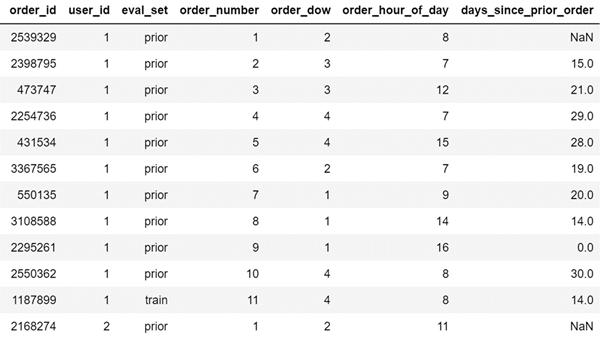
**2.1.1 2.1.2**

****

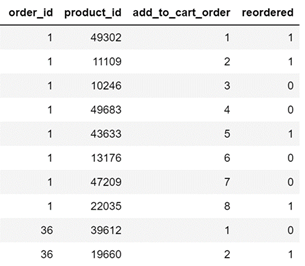
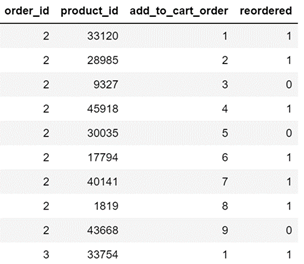
**2.1.3**

First of all, we have 3 dataset files named as aisle, department and product. In aisle and department files, they simply just contain two columns : unique ID and categorical name. In aisle, for example, it has 1 to 134 Id numbers representing each unique aisle according to their categories name in the next column. Then we have product files which contain information about all products with their unique IDs linked with them under which aisle and department.

We have 134 aisles, 21 departments and 49,688 products in total from our complete data source.

****

**2.1.4**

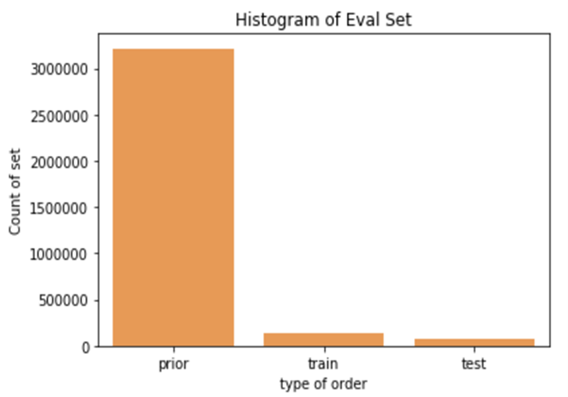
****

**2.1.5 2.1.6**

In addition, we have another 3 data files that contain all of Instacart orders’ information. In 2.1.4, the first 12 records are shown in our order datasets. To interpret the first 12 rows, for example, user 1 has placed 11 orders with 11 unique order IDs in the past on Instacart App, and it also tells us about on which day of the week and hour of the day he placed each particular order, how many days interval between these orders. In addition, it also tells these records belong to which data set: prior, train or test? In this example, 10 of the orders for user 1 are in the prior dataset, and 1 of them is in the train dataset.

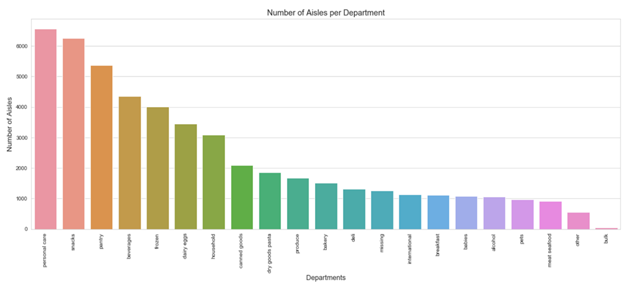
In the chart 2.1.5 and 2.1.6, these are abstracts from our prior and train data files. Unlike the previous Order data file, these two files contain 2 new information: item add to cart order and whether the product has been purchased before. In 2.1.5, for instance, there were 9 items purchased in order Id 2 with unique product Ids, and we have the order purchased for these 9 items in time sequence, and 6 of the items have been reordered by this particular customer before, and 3 of them have not. Same interpretation ideas applied in the training data file.

In sum, we have over 3 million order records of Instacart apps in our mega data source from Kaggle.

**2.1.7**

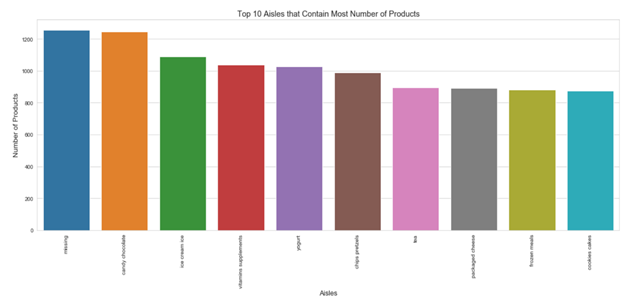
In 2.1.7, this is the data types distribution. We can see that there are about 3.4 million records in the prior dataset which is really large, and 131k, 75k in the train and test dataset.

**2.2 Data description**

****

**2.2.1**

Here is the product distribution in 21 departments, we can see the most number of items fall under personal care, snacks and pantry departments within about 6k items per each.

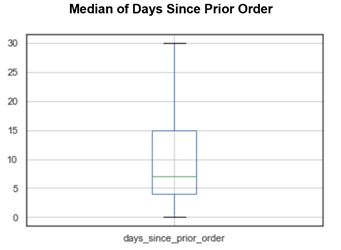
****

**2.2.2**

Then, there are 134 aisles in total. Here are the top 10 aisles that contain the most number of products in 2.2.2, and we can tell that missing, candy chocolate, ice cream are the top 3 aisles that contain the most items.

****

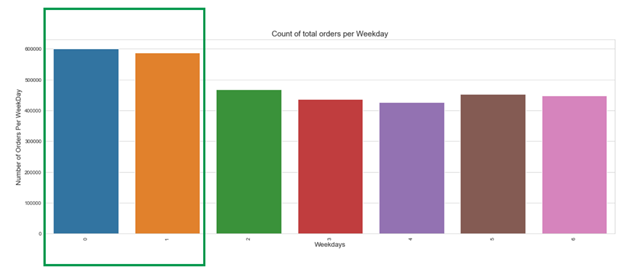
**2.2.3 2.2.4**

****

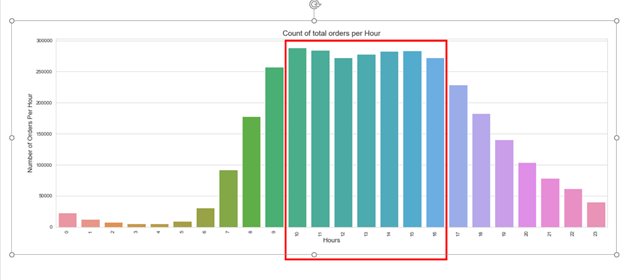
**2.2.5 2.2.6**

We count the number of reorders by aisles, departments and items. Then, we rank top 10 popular aisles, departments and items. We can tell that department produce, dairy egg and snacks are the most 3 popular departments. Moreover, under the most popular aisles fresh fruits and fresh vegetables, bananas, organic strawberries, baby spinach and avocado are the most frequently bought items.

In 2.2.6, the median day of the frequency of customers shopping in Instacart is 7 days, which means in general consumers are placing orders in the app weekly.

****

**2.2.7**

****

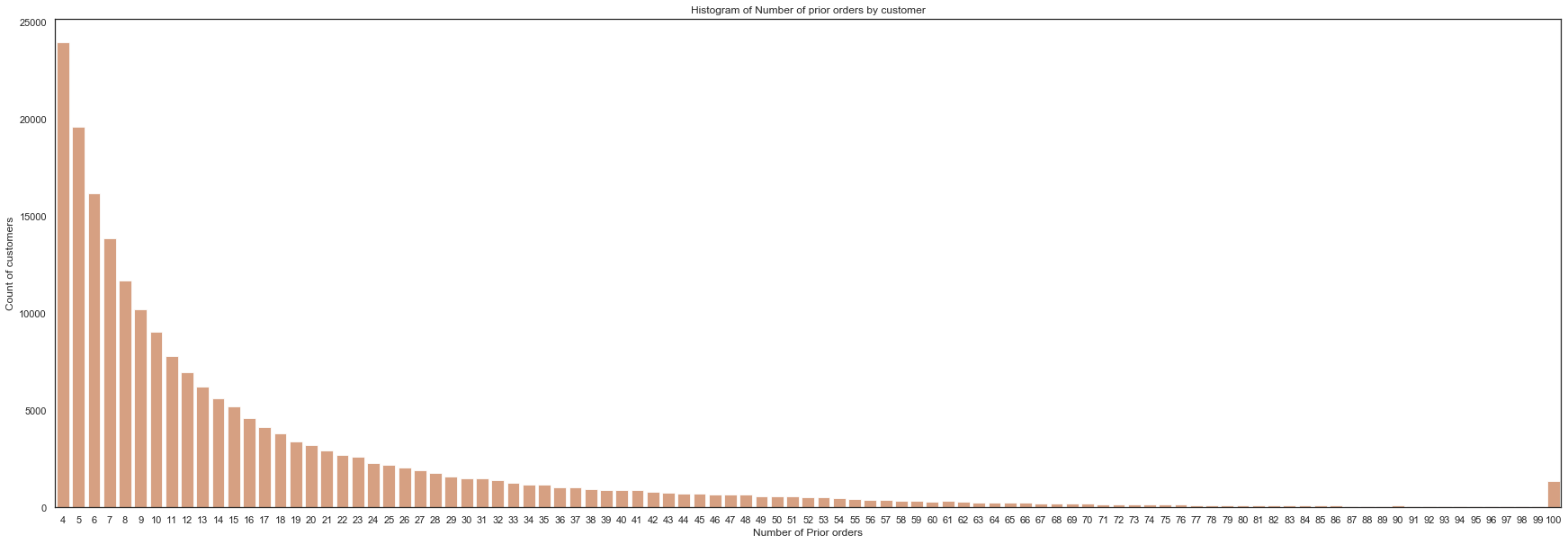
**2.2.8**

To identify most busy hours and days by counting the number of orders. There is a clear effect of the day of the week. Most orders are made on days 0 and 1. Unfortunately there is no info regarding which values represent which day, but one would assume that this is the weekend.

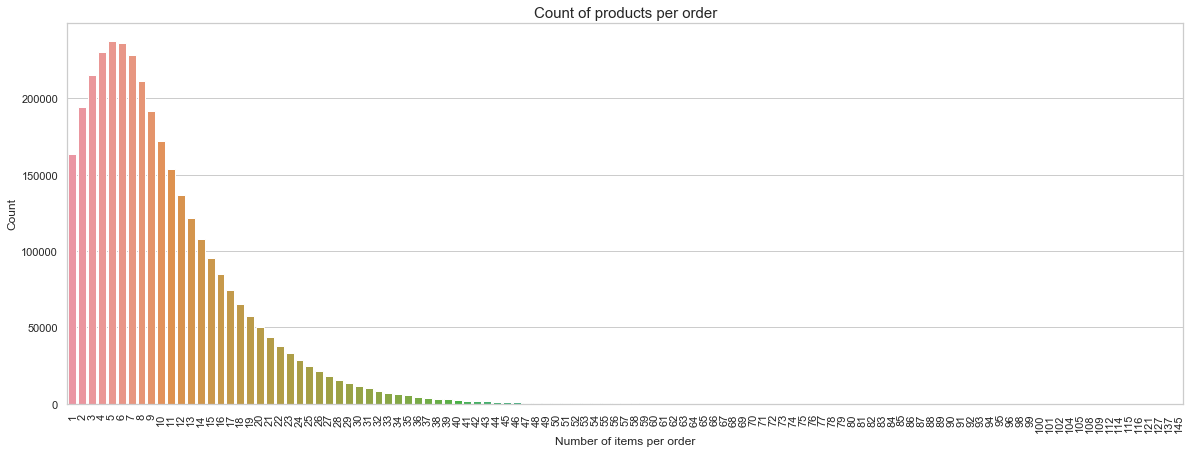
In addition, there is also a bell shaped graph in busy hours distribution, and we could see that most of the orders were placed within 10:00 to 16:00.

**3 Data Exploration**

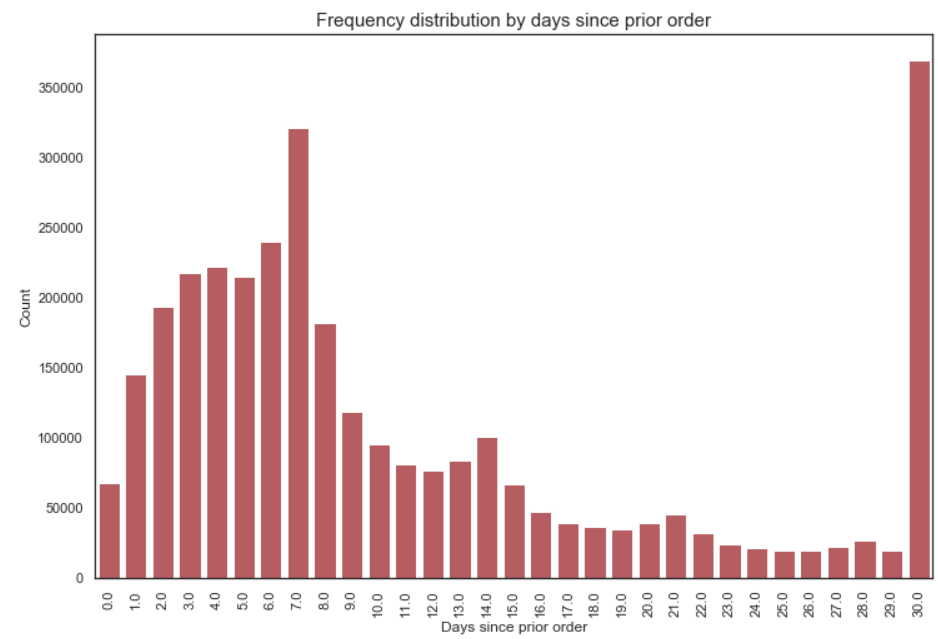
**3.1 Customer Behaviors**



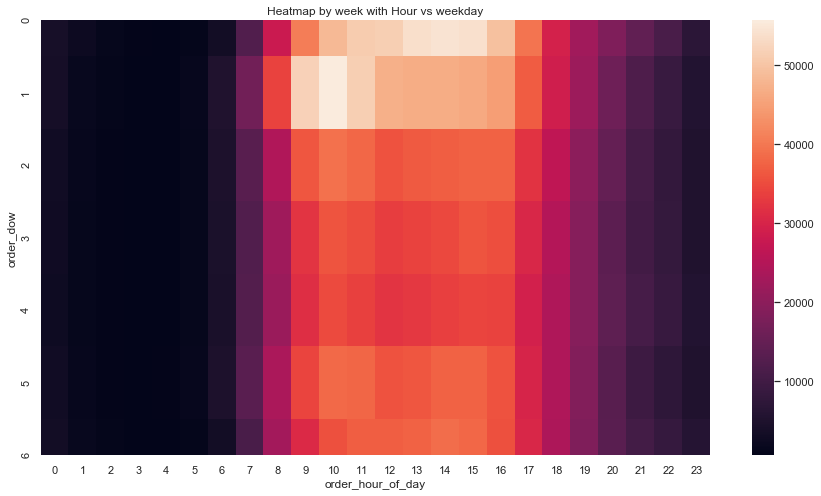
The reorders’ frequency plot follows the power-law distribution. The number of prior orders is concentrated on the range from 4 to 12. Considering Instacart is a fresh company founded in 2012, it makes sense that most customers are relatively new registers that don’t have many purchase records.



From figure2, we can see that the distribution is peaked at 5 with right skewness, with median somewhere between 8 and 10. Therefore, we may expect that usually one order will contain around 7 products.

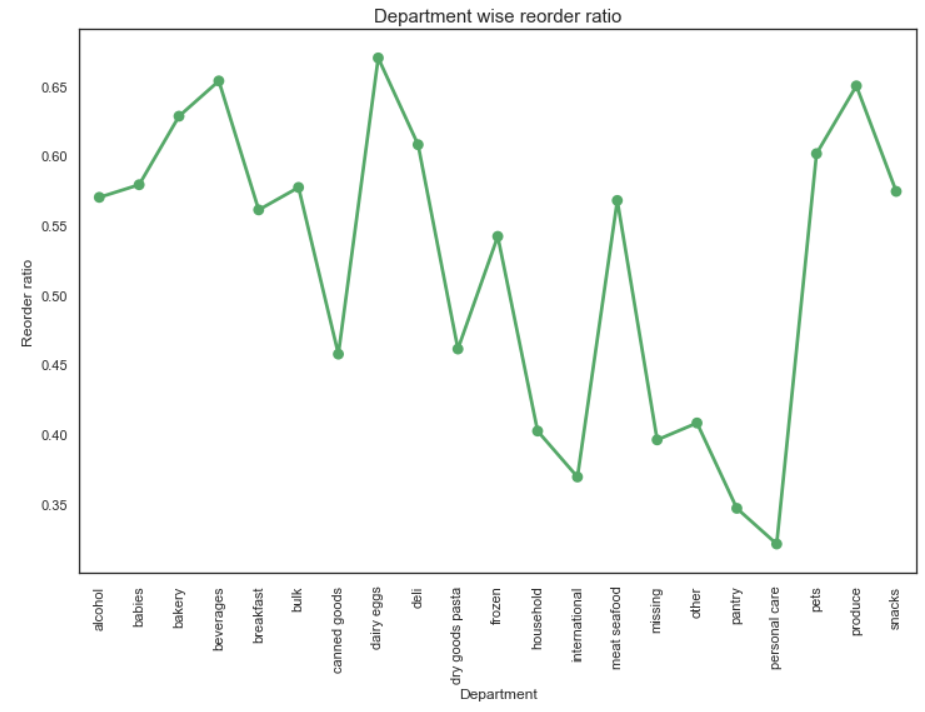


In figure3, there are 2 spikes that one is 7, the other is 30. It indicates that customers are willing to place orders again every other week or month. This is also consistent with a frequency of food purchases by households.

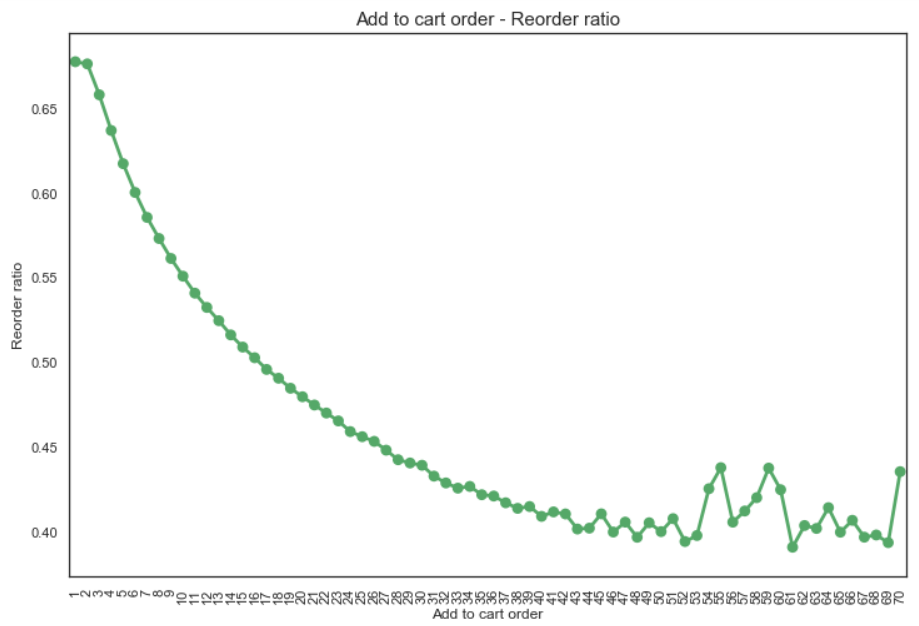


In figure4 y-axis, 0 represents Saturday, 1 represents Sunday, 2 represents Monday and so on. It’s obvious that the traffic of customers is boosted on weekends. That makes sense, because people have more leisure time on weekends to shop. Also, the “heat” hours in map is from 9am to 5pm, it’s also in line with people's work and rest laws.

**3.2 Reorder Pattern**



According to reorder ratio of departments in figure5, we can infer the popularity of different types of food on Instacart. Thus, beverages, dairy eggs and produce are the most popular products on Instacart.



The probability of products to be reordered will reduce as the increasing times of adding to cart. *I’m not clear about this plot, what does the number of add to cart order mean.*

**Customer Segmentation**

Segmentation is a widely used technique for better understanding of either the target consumer base or the existing customers to cater to their needs in a better way. This can help in better marketing, cross-sell recommendations and other customer outreach programs.

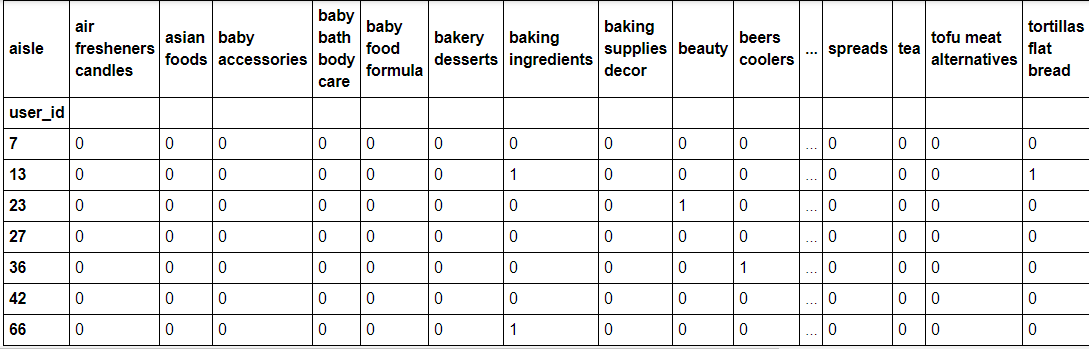
**Goal:**

For Instacart customers, before making any product recommendations, it will be helpful to understand the buying behavior of the already existing consumer base. Using that information, we can build an algorithm to segment consumers into different buckets with identical features and use that data to learn about what a particular consumer might want to buy in their next cart. This will also help Instacart to segment their new consumers (based on demographics features) and use the already known machine knowledge to make shopping recommendations to them.

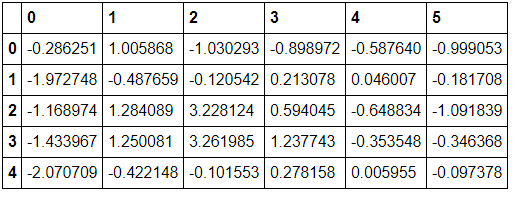
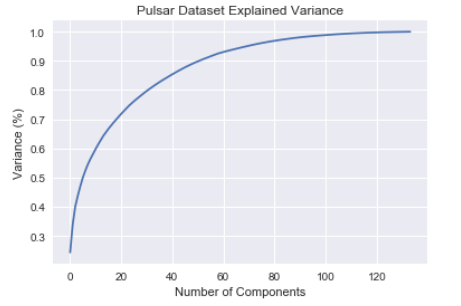
Thus, we used consumer segmentation using K-Means unsupervised learning as one of the ways to learn more about Instacart’s consumer buying preferences and make shopping recommendations to them based on that.

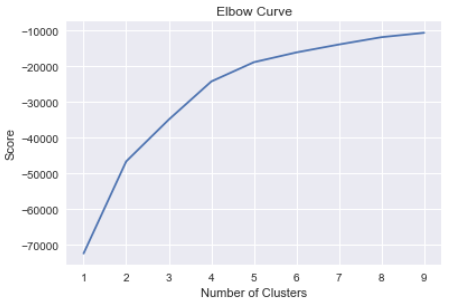
**Process:**

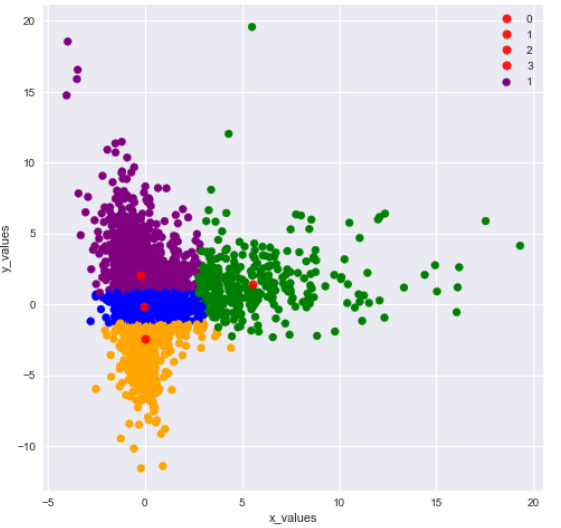
To segment the consumers, we could either do it by products, aisles or departments. Now, there are ~50K different products, 134 different aisles and only 21 departments. Now, since there are a lot of products, it may not be the most efficient way to perform clustering based on that. So, we ran clustering both on departments and aisles and observed that we were able to get a much better distribution of consumers by clustering on aisles. Therefore, going forward the documentation will elaborate on clustering based on aisles only.

Our data was clean and therefore we did need to do any data manipulation and could start with the clustering right away. We merged different tables to get the data in the following form:

The different columns in the abve are different features we need to base our clustering on. There are 134 different aisles. This makes our data very wide and it will be easier if we could reduce the number of variables using principal component analysis.

**Principal Component Analysis**: This is a technique used to condense the information stored in multiple variables in fewer number of mathematical vectors that can then be used for clustering instead of using 134 aisles for that purpose. However, how do we find the ideal number of components to reduce our features into? We plotted the graph to the right to see how many components we need to explain maximum variance of the original 134 features. The graph shows that about 80 components will be able to explain close to 98% variance of the original features. Thus it will be ideal to reduce the original 134 features into ~80 components. However, since that was still complicated our purpose, we settles with only 50% of the variance of the original features and generated 6 principal components. Please note – this was only for simplicity sake and serves the purpose of our personal learning through this project. We got the following principal components:

**Clustering:** To identify the optimal number of clusters, we used the elbow method. This method is used to identify the number of clusters after which the sum of squared distances of samples to their closest cluster center start decreasing. In the figure in the right, the score starts flattening out after ‘4’ clusters. Thus we start building our clustering model for 4 clusters. By permutation and combination, we identify that using principal component 1 & 4 give the best distribution of consumers across all segments. Thus, we built our clustering model for 4 clusters using 2 principal components – 1 & 4 and we got the following result:

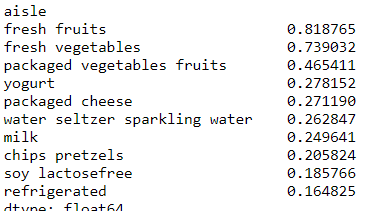
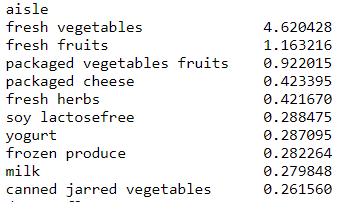


The above figure shows 4 clusters. We can now go ahead and study each cluster to a greater depth and understand the buying pattern across them.

**Insights:**

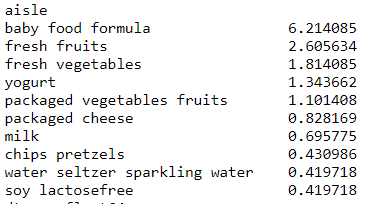
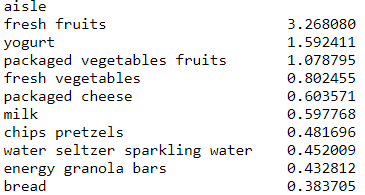
1. One of the insights we can derive from the clusters is the top 10 items items (aisles bought from) in each of these clusters. Now, even though there are a lot of common aisles, the probability of the consumer buying from each one of them gives us the real information. This also highlights some unique characteristics of each cluster that can be used to make some recommendations for consumers in that cluster. For example: baby food in cluster 4.

See below:

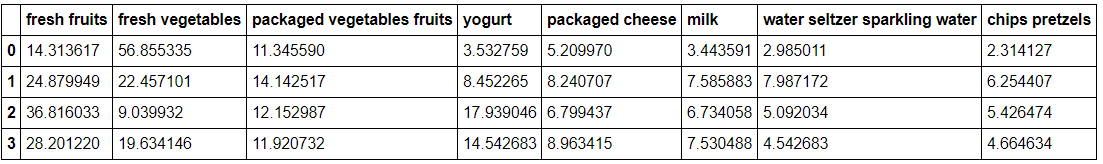


2.

2. We can also



1. The following table depicts the percentage of the popular items with respect to the other top 8 in each cluster. We can see some interesting results. For example: people in cluster 1 buy more fresh vegetables than the other clusters and cluster 2 buy more yogurt than other cluster.



**APRIORI Algorithm:**

As we investigate exploring the association analysis more, we get into exploring it through Apriori Algorithm. Apriori Algorithm is a frequent item set mining and association rule learning, proceeds by identifying the most frequently associated items by using a breadth first approach and hash tree structure. The candidate set contains all frequent items as a set and it scans the relational data base to find out the frequency of the items. The algorithm uses the bottom up approach, which extends the associations one at a time and run until no further baskets with the items occur. Therefore, the algorithm can be extended to any number of item sets to a large data base.

The affinity analysis to find out the co-occurrence relationships among the activities, in retail is called Market Basket Analysis. For this dataset, we have taken the set length as ‘two’, which means we are going to explore the association between different any two different items. Three important concepts in association rules mining are

1. Support: The support is the percentage of orders, in which the item set appears. Let’s, say A and B are our items.

Support {A, B} = Number of orders with set {A, B}/ Total number of orders

We have taken minimum support as 0.001, which allows us to explore most important products.

1. Confidence: Confidence measures the percentage of times, when one item in the set is purchased, what is the number of times both items are purchased

Confidence{A->B} = support {A, B}/support{A}

Confidence is unidirectional and always ranges between 0 and 1.

1. Lift: Lift define weather there is a relationship between two items or the items are purchased simply by chance.

Lift {A, B} = support {A, B} / (support{A}\*support{b})

We have taken minimum support as 0.001, which means the item set has to appear in every 1000 orders to filter the over searching done by APRIORI. We have a total set of 11554 items out of the total 1384617 orders with support of 0.001. Each of these items are paired against each other and then filtered. The stats are as below

Starting order\_item: 1384617

Items with support >= 0.01: 11554

Remaining order\_item: 1274989

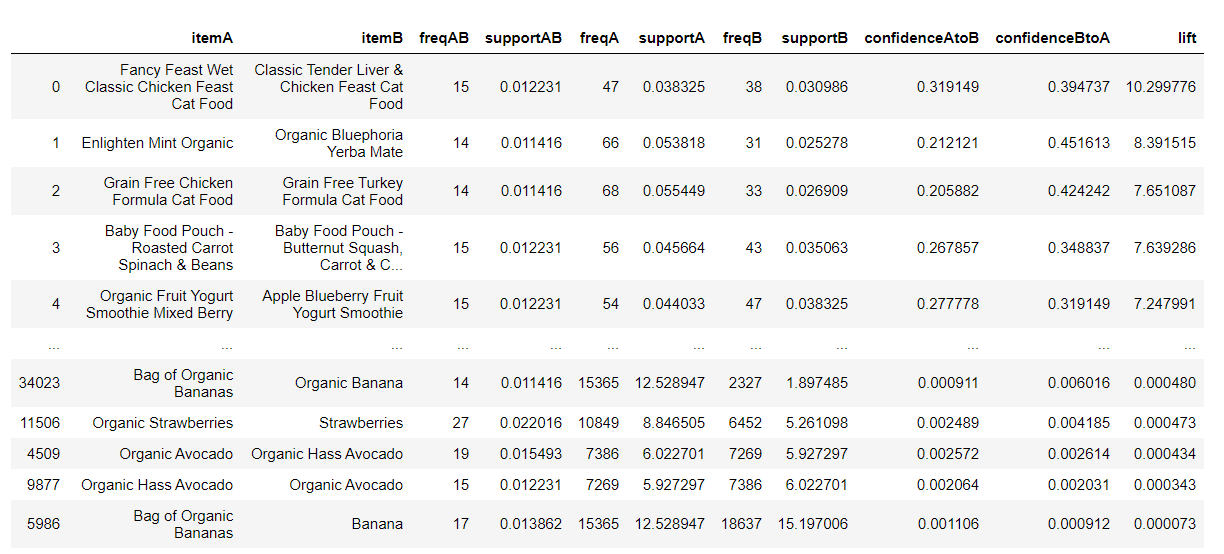
Remaining orders with 2+ items: 122636

Remaining order\_item: 1267339

Item pairs: 5004378

Item pairs with support >= 0.01: 55341

Once we ran the whole model and sort the model by lift, the top ten item pairs are as follows.



**Observations:**

**Rule:** We can see that cat food with either different flavors or different items in the cat food has highest lift, which means they are frequently bought together

Conclusion1: when selling these items, there is no point of discounting them as a pair, as people are already buying both.

Conclusion2: If a new product is released, it will be good to mix it with other item as it leads to more visibility and because of the brand value, we can have the cover for launching more.

**Rule:** We can see that customers often buy different items for same purpose (feeding cat, baby food, different types of yogurt), they have similar value but offers variety.

Conclusion: As a marketing manager, product diversification is one of the important criteria to improve the market share. If we are selling the same product but just with a variation, it is always good to have a multiple pack for the similar item but with different flavors.

**Rule:** In the case of bananas, the product is similar but varies in sizes. Hence the lift is low.

Conclusion: As a store manager, if I am designing my aisles, it would be good to place these items in the same shelf as it would offer options for the similar value items for the customer and thereby making it easy for them to explore all options before they buy.

As Instacart is an Online portal, the purpose of association rule is to build recommendation system. With these item pairs, as soon as a customer buys on item, a recommendation aisle would be shown with the different items with maximum lift associated with the product. This would add the value to the service platform, as it would give more options for customer to buy. We can further explore the algorithm into finding association rules for more than 2 item pairs but sue to the nature of larger dataset and computational power problems, we are concluding with 2 item pairs.

**5 Conclusions:**

**6 Recommendations for Business**

1) Using Apriori algorithm, we could find the correlations between items. So, if a customer added product A in cart, we may recommend product B which are highly related to product A to the customer. For example, a customer who buys Fancy Feast Wet Classic Chicken Feast Cat Food are 10 times likely to buy Classic Tender liver& Chicken Feast Cat Food than other randomly picking products. In this way, we can not only sell our products but also save the customer’s time to search that product that he/she may potentially buy.

2) For a returning customer, we can use segmentation to classify the customer into a particular cluster and recommend the relative products to that customer based on the cluster he/she has been put into. This is a sufficient way to target our customers and potentially increase the revenues.

3) For a new customer, we may recommend products that are purchased most frequently by majority of the users from the store to them.

4) We could provide a pre-add products list to the basket of a customer (i.e. before he/she starts adding products to the basket) based on the customer’s previous purchase, to make the purchase experience quick and easy.

5) For every product that we recommend, we will consider it as a success if the consumer ends up purchasing the recommended product. So, the success can be calculated by accuracy (Accuracy for an order = no. of recommended products purchased / total no. of recommended products; Final accuracy = mean of accuracy of all the orders).