Market Basket Analysis of Instacart Data

By Shraddha Patel, Naga Surya Suresh, Nevil

Goal :

Analyze data ,find out hidden association between products for better upselling and predict which previously purchased product will be in user’s next order

Content

Introduction

Data Description

Data Analysis

Data Modeling

Model selection and Analysis

Association Analysis

Model to Predict Next Purchase

Conclusion and Future Work

References

Introduction

Instacart is online grocery delivery and pick up service provided by a American Company.

Our aim is to build a model that predict which previously purchased items will be in customer's next order which can boost sales for the company by improved customer experience, increased customer interaction and suggesting relevant products to consumers, which saves time.

Data Description

Data is extracted from the Kaggle

This is a relational group of files that tracks the orders that customers place over time.

Each user receives between 4 and 100 of their order details, including the order of the things they purchased in each order, the day and week it was placed, and the amount of time between orders.

Each entity (customer, product, order, aisle, etc.) has an associated unique id and its related data.

Data Description

aisles: This file contains different aisles and there are total 134 unique aisles

departments: This file contains different departments and there are total 21 unique departments

orders: This file contains all the orders made by different users.

products: This file contains the list of total 49688 products and their aisle as well as department

prior: This file gives information about which products were ordered and in which order they were added in the basket. It also tells us that if the product was reordered or not

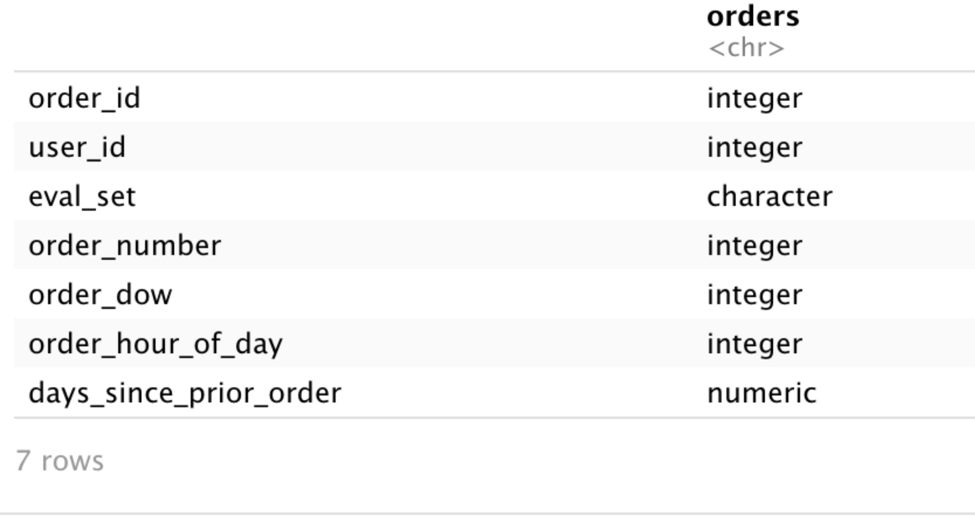
Data Processing

Data Processing is important in further Analysis

After analyzing the initial dataset, we converted various character variables to Factors and numeric values to numeric

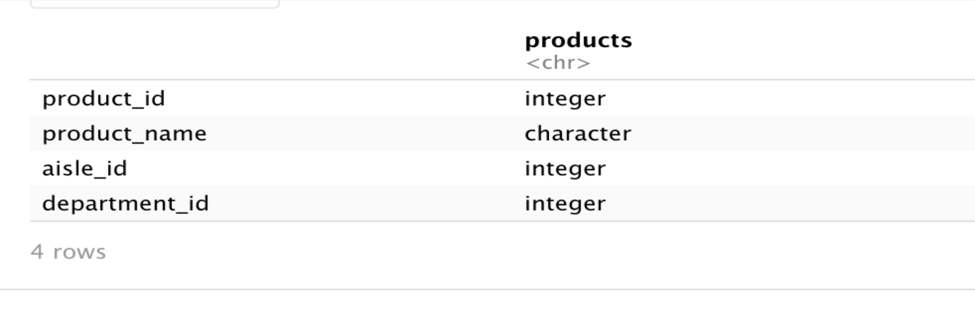
The factor conversion is done on orders, products, aisles, and department data sets

Transformed Data set



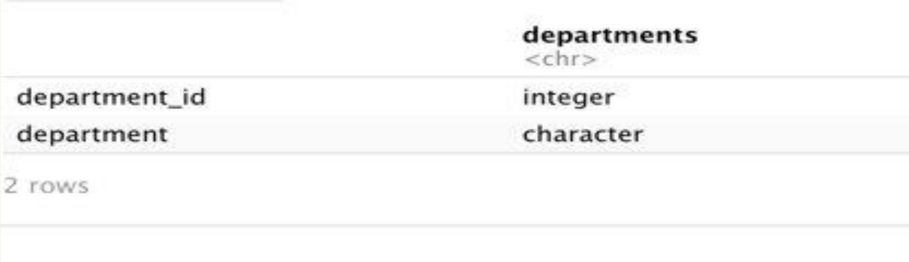
	orders <chr>
order_id	integer
user_id	integer
eval_set	character
order_number	integer
order_dow	integer
order_hour_of_day	integer
days_since_prior_order	numeric

7 rows



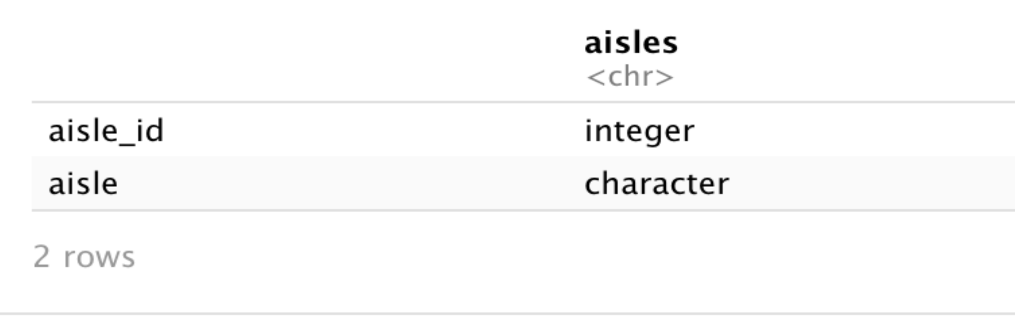
	products <chr>
product_id	integer
product_name	character
aisle_id	integer
department_id	integer

4 rows



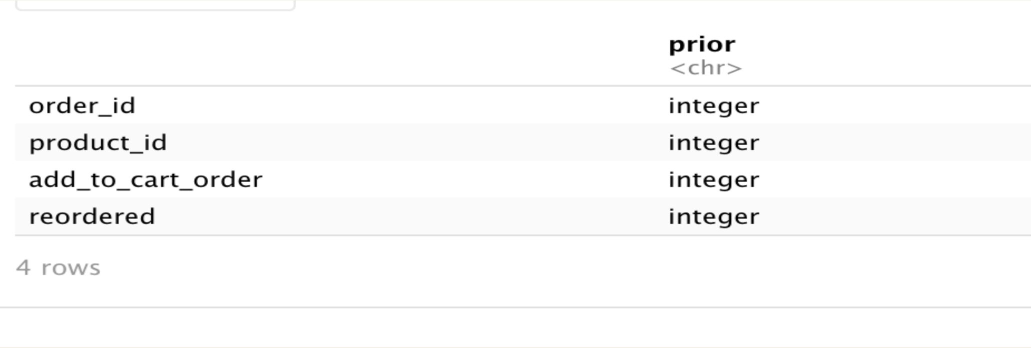
	departments <chr>
department_id	integer
department	character

2 rows



	aisles <chr>
aisle_id	integer
aisle	character

2 rows



	prior <chr>
order_id	integer
product_id	integer
add_to_cart_order	integer
reordered	integer

4 rows

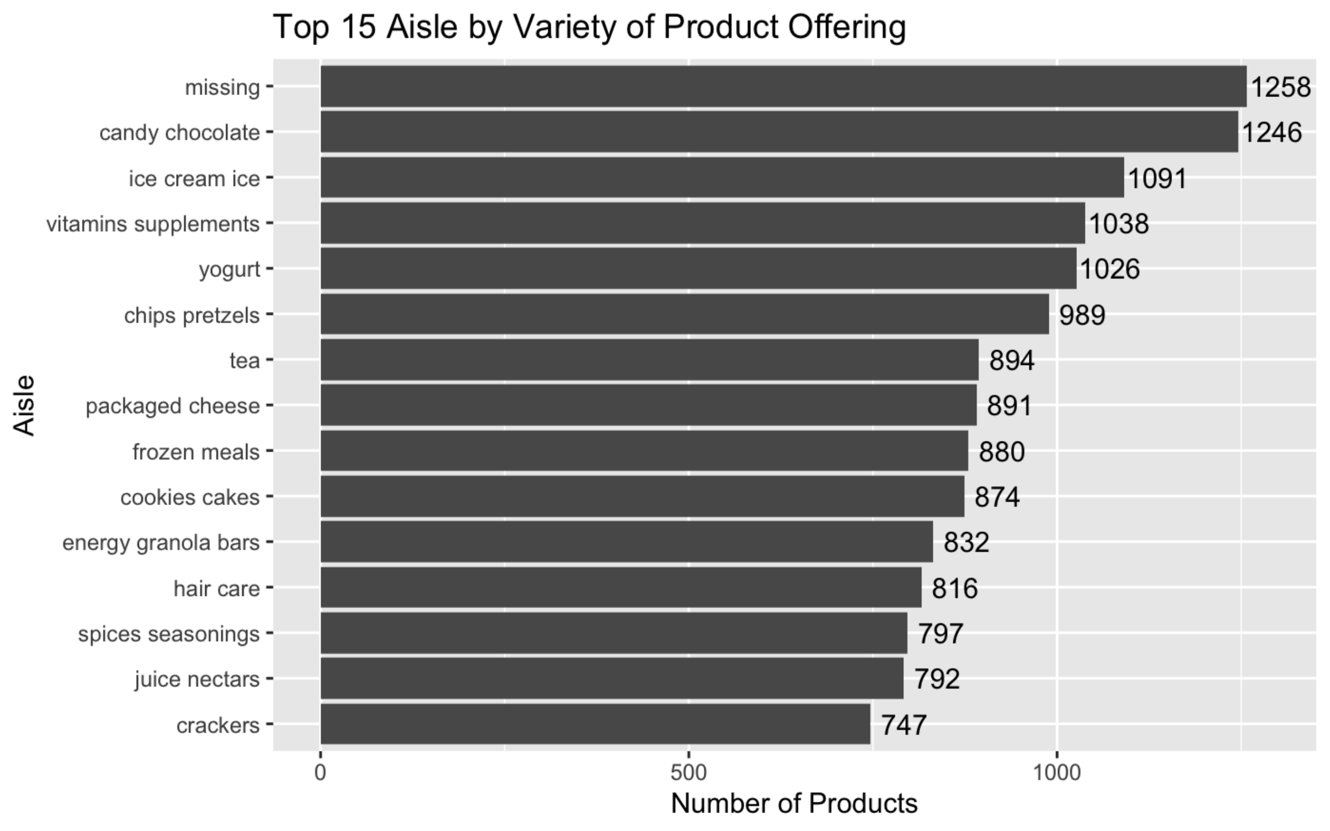
Data Analysis

We merged datasets of products, Aisles and Departments data to find the exact product offerings.

The Merged products, Aisles and Departments data has 49688 Rows and 6 Columns.

We further explore the data to find the product offerings of the Instacart data.

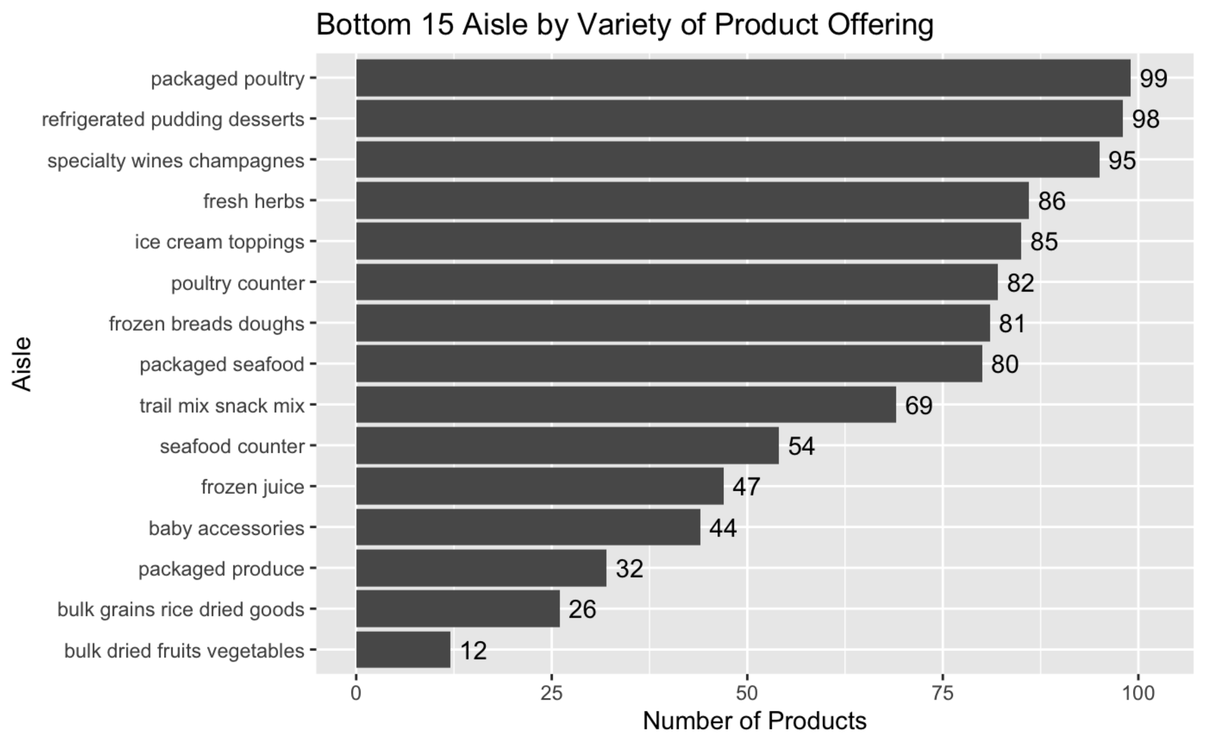
Aisle by Variety of Product Offering



Graph shows that most customers orders and purchases products like crackers, juice nectars etc., This shows the preference of the customers.

Also, there are some missing items in the aisle which means we don't know where the product is located or the product_id does not match the aisle_id

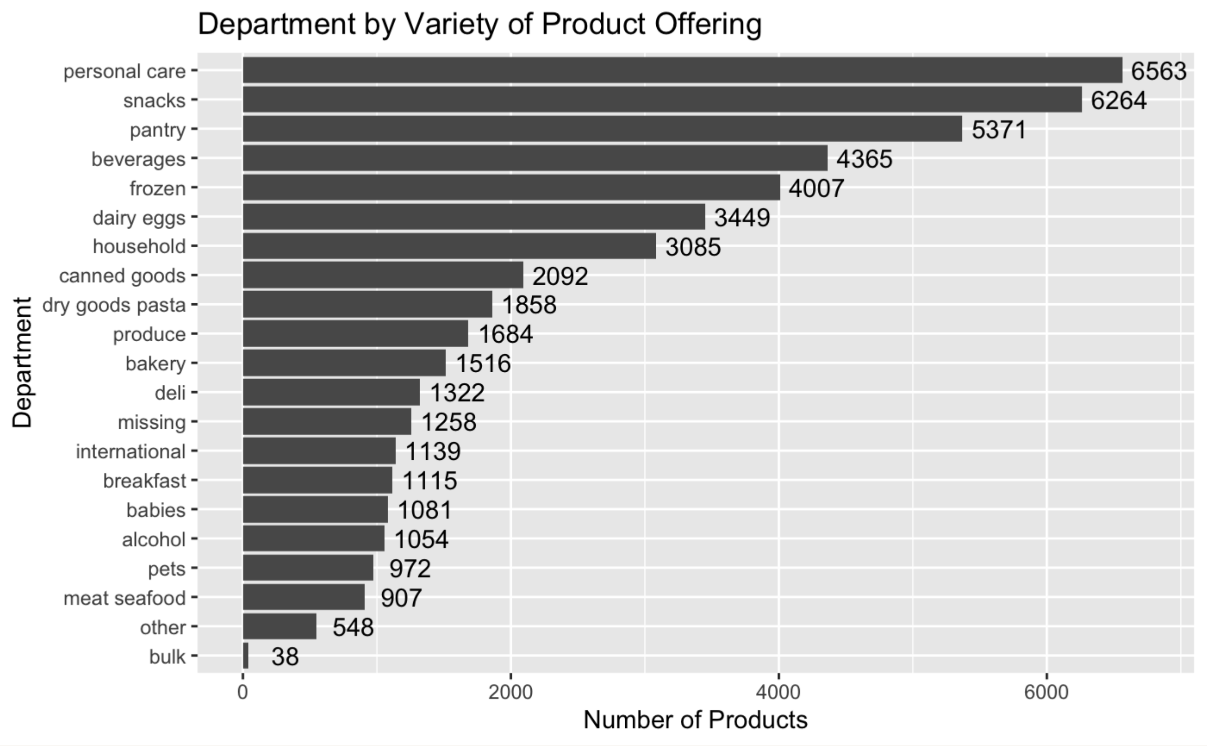
Aisle by Variety of Product Offering



Graph show that bulk products like dried fruits, vegetables and dries grain rice are most ordered and purchased than the wines and poultry products.

Also, there is no missing items in the aisle which means no item is mis-matched in this set of data .

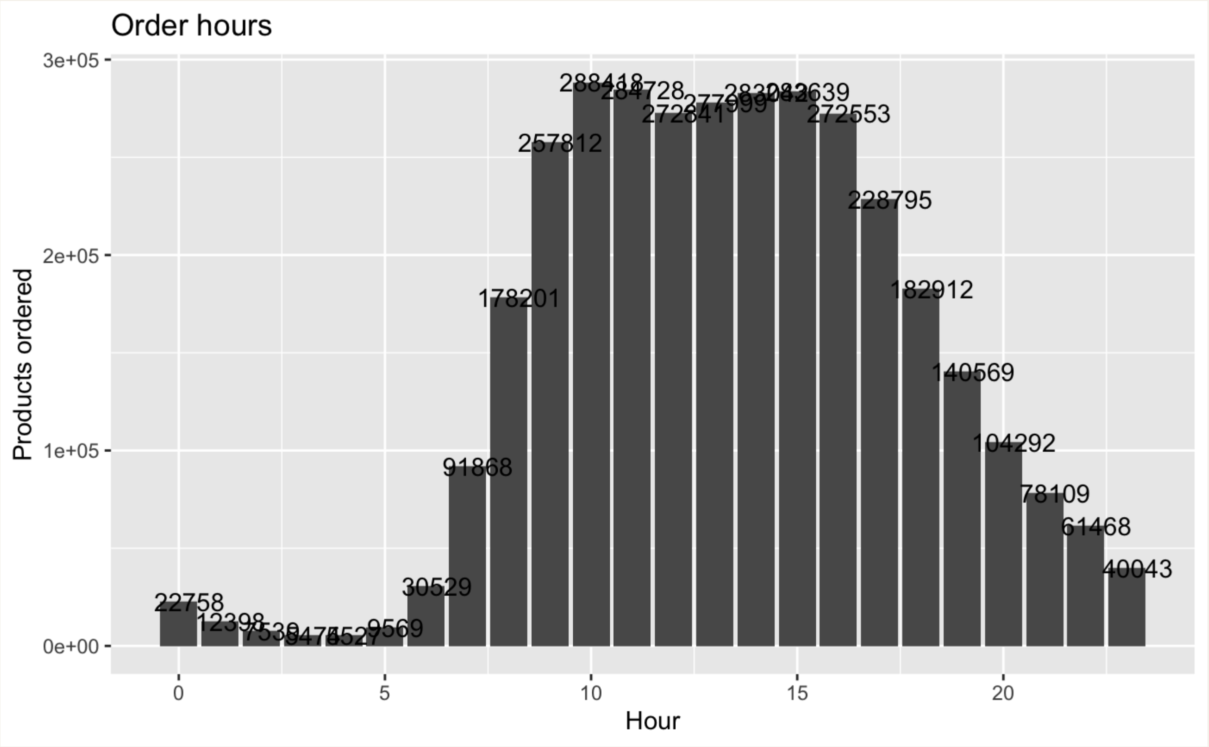
Department by Variety of Product Offering



Graph shows popular departments which are very close to each other

The departments that are in bulk are ordered and purchased most by the customers.

Orders by hour

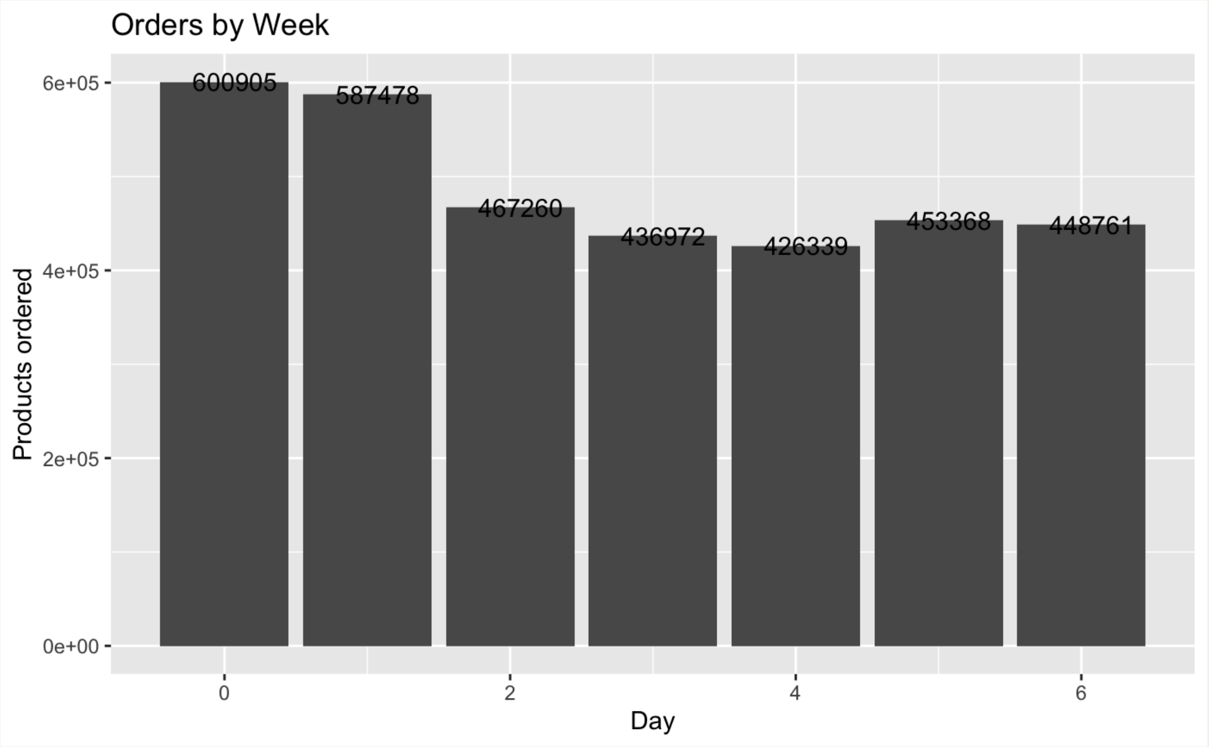


Plot shows that for the first 5 to 6 hours the number of products ordered are very less.

The products ordered between 7 to 20 hours are more and then decreases after 20 hours.

This explains that middle of the day is more active than the opening and closing hours.

Order by week

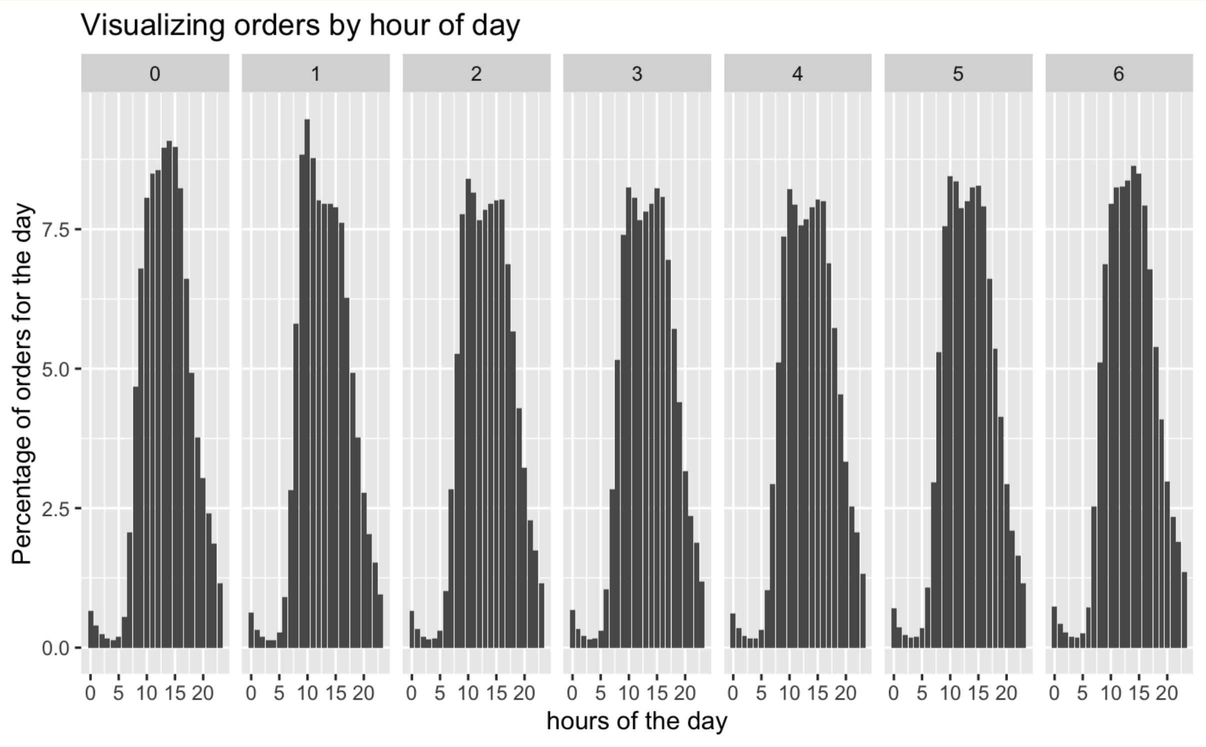


Graph clearly state that in the

Start of the week the product ordered are high and by the mid of the week the ordering decreases and gets stable by the end of the week.

Orders by Every day Every hour

The Graph shows the product order percentage vs hours in a day.



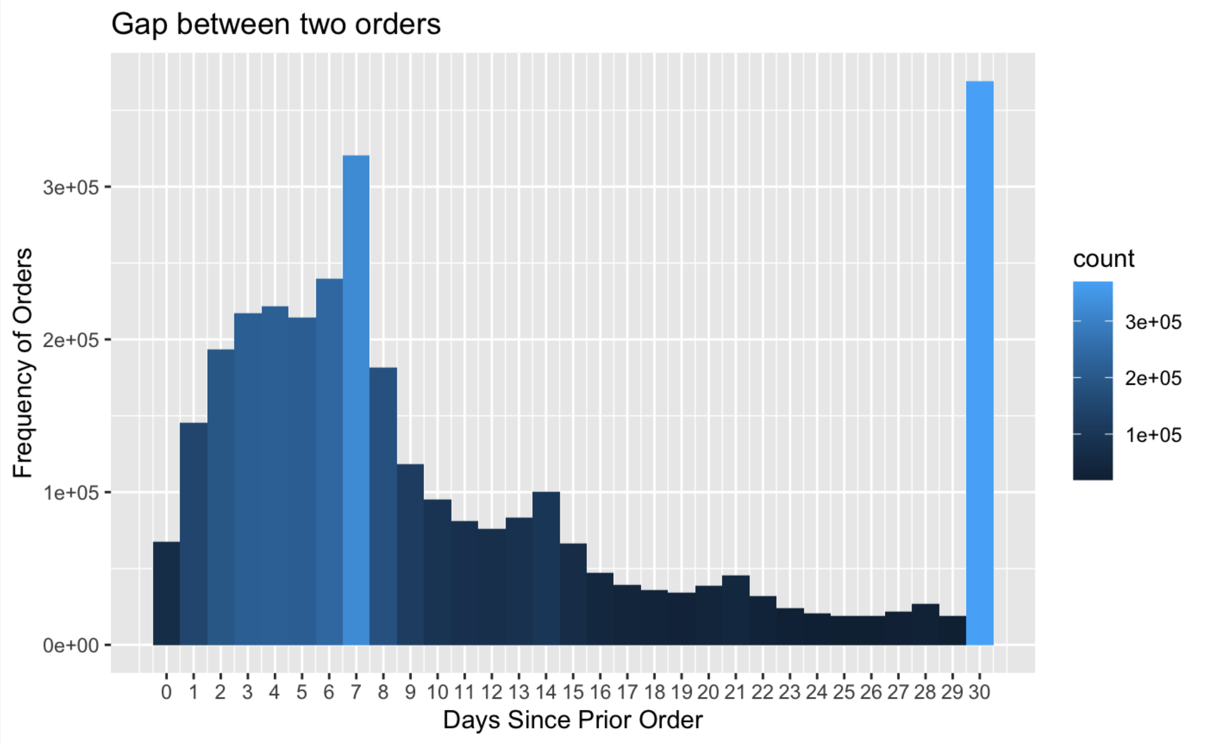
Each and every hour has a different order pattens.

The percentage of products ordered by the middle of the hour reaches maximum at each and every hour.

By this we can analyze the peak ordering time in an hour.

Days Since Prior Order Analysis

Based on 30th day and 7th day peaks. People reorder on average 11% monthly and 9% weekly.



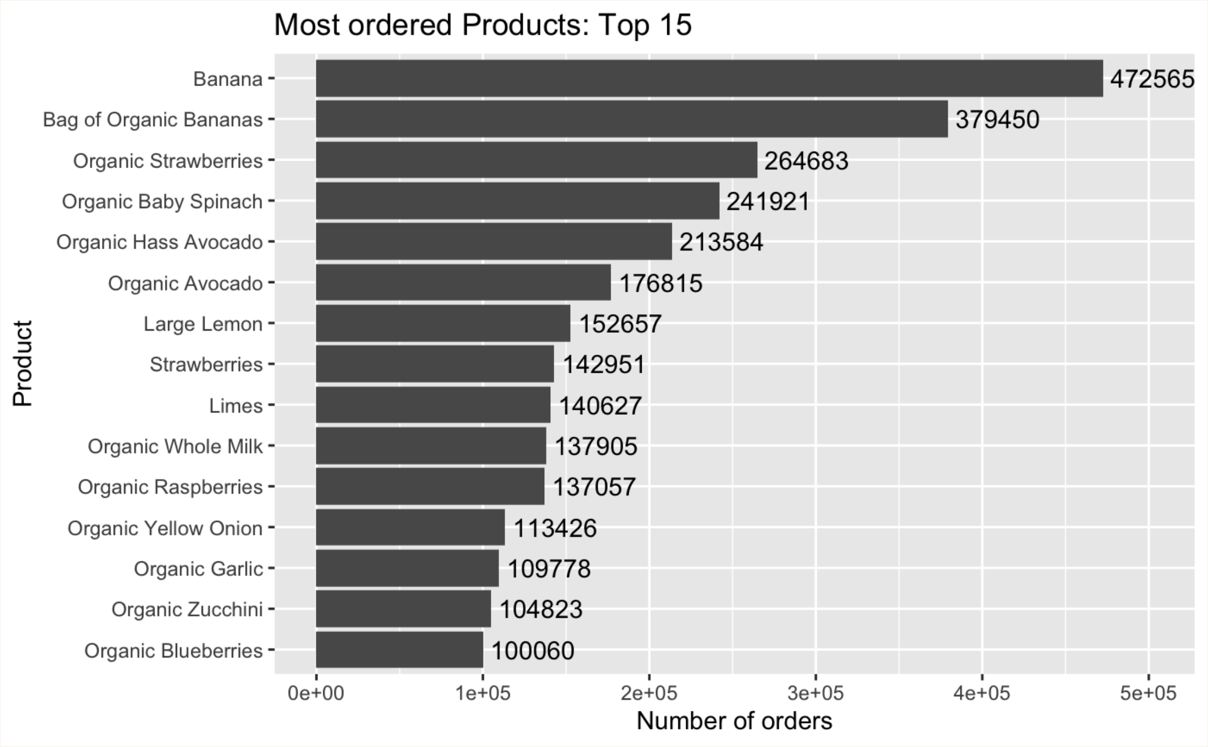
This demonstrates that there are some people who restock their food every month and others who restock them weekly.

The frequency of NA denotes the total number of distinct users and their initial order.

There is a continuous spike in orders from day 1 to day 6, shows that some people are frequent buyers with short window of restocking.

Prior Table Analysis

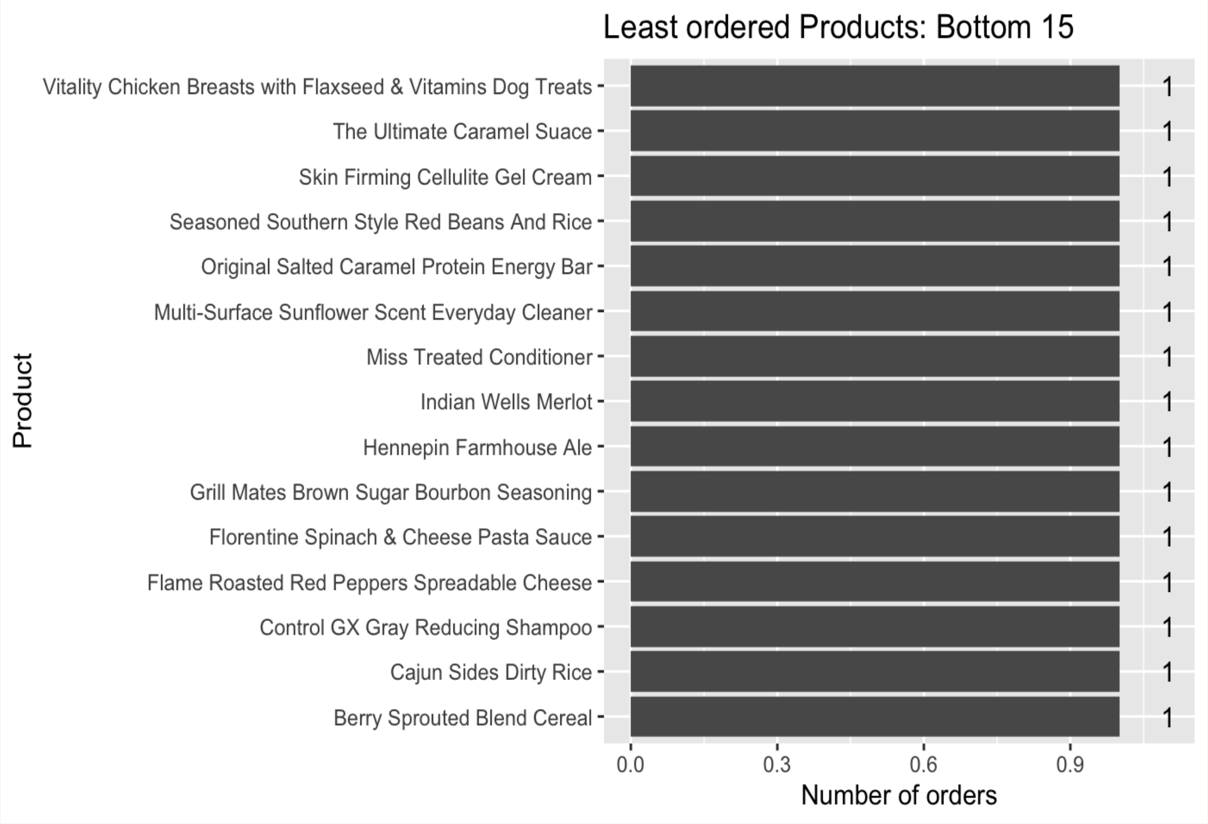
We can predict their next purchase or order by looking the history of their order using prior table.



From the graph we can see the result that the most ordered product is Banana and Organic Banana which tops in the top 15 products.

The graph represents the history of orders by all the users and their most preferred product.

Prior Table Analysis



The graph shows the top 15 least ordered products from the prior table.

It clearly shows that the frequency of the products ordered by the customers are less.

Data Modeling : Model selection and Analysis

Market Basket Analysis is a modelling technique based upon the logic that if you buy a certain set of products, you are more or less likely to buy another set of products.

This information can be used for the purpose like cross-selling, product placement, affinity marketing, fraud detection, and consumer behavior

Association Rule Mining is used when we want to find an association between different objects in a group ,find frequent patterns in a database or any other information repository.

We used the **Apriori algorithm** for mining association rules and make a comparison with the Frequent Pattern Growth Algorithm

Data Modeling : Model selection and Analysis

The user id and product id serve as the keys of a denormalized structure that is formed after the **features are created.**

We have used XGBoost (Extreme Gradient Boosting) Classifier for classification of the features created.

XGBoost is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset’s features.

Metrices used to find association rules

Market Basket Analysis using Apiroi algorithms to predict grouped items, transactions, frequency, rules and define meaning full definitions.

Apriori Algorithm: Apriori algorithm assumes that any subset of a frequent itemset must be frequent.

Association rules are created by analyzing pattern and using criteria support and confidence to identify most crucial relationship.

Metrices used to find association rules

Support : Its the default popularity of an item. In mathematical terms, the support of item A is the ratio of transactions involving A to the total number of transactions.

Confidence : Likelihood that customer who bought both A and B. It is the ratio of the number of transactions involving both A and B and the number of transactions involving B.

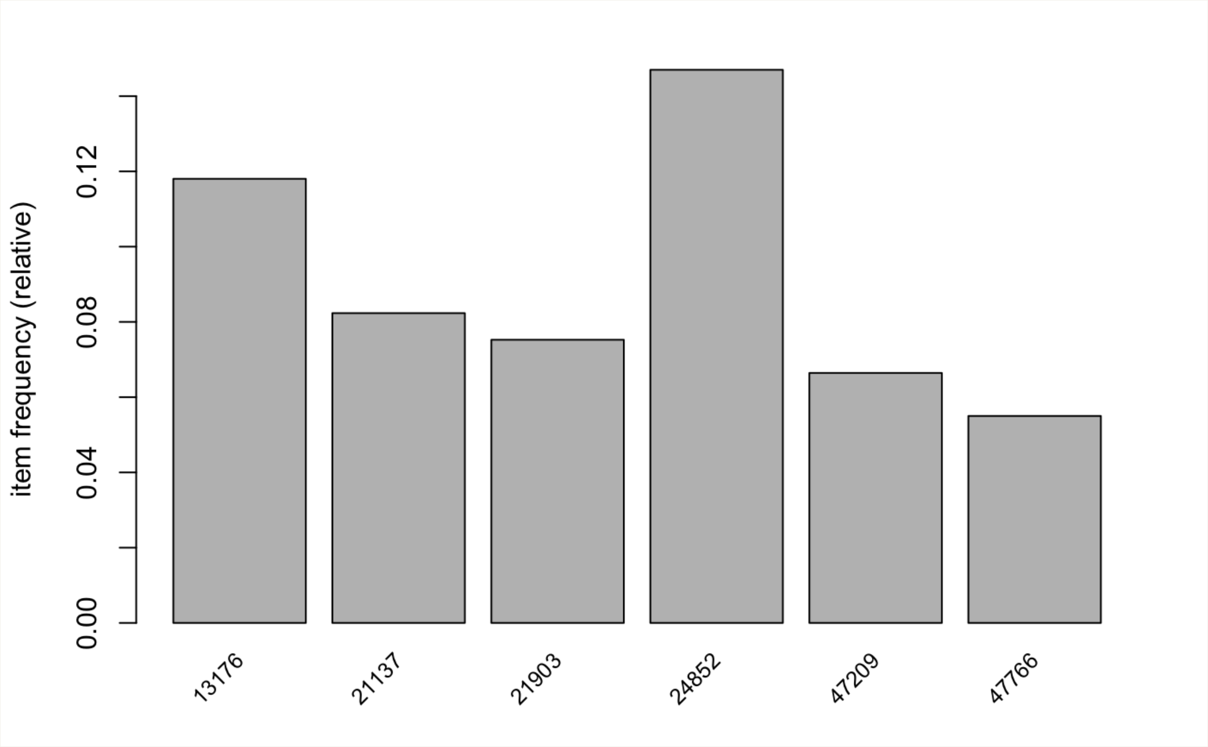
- Confidence(A => B) = Support(A, B)/Support(B)

Lift : Increase in the sale of A when you sell B.

- Lift(A => B) = Confidence(A, B)/Support(B)

Products that occurred most frequently

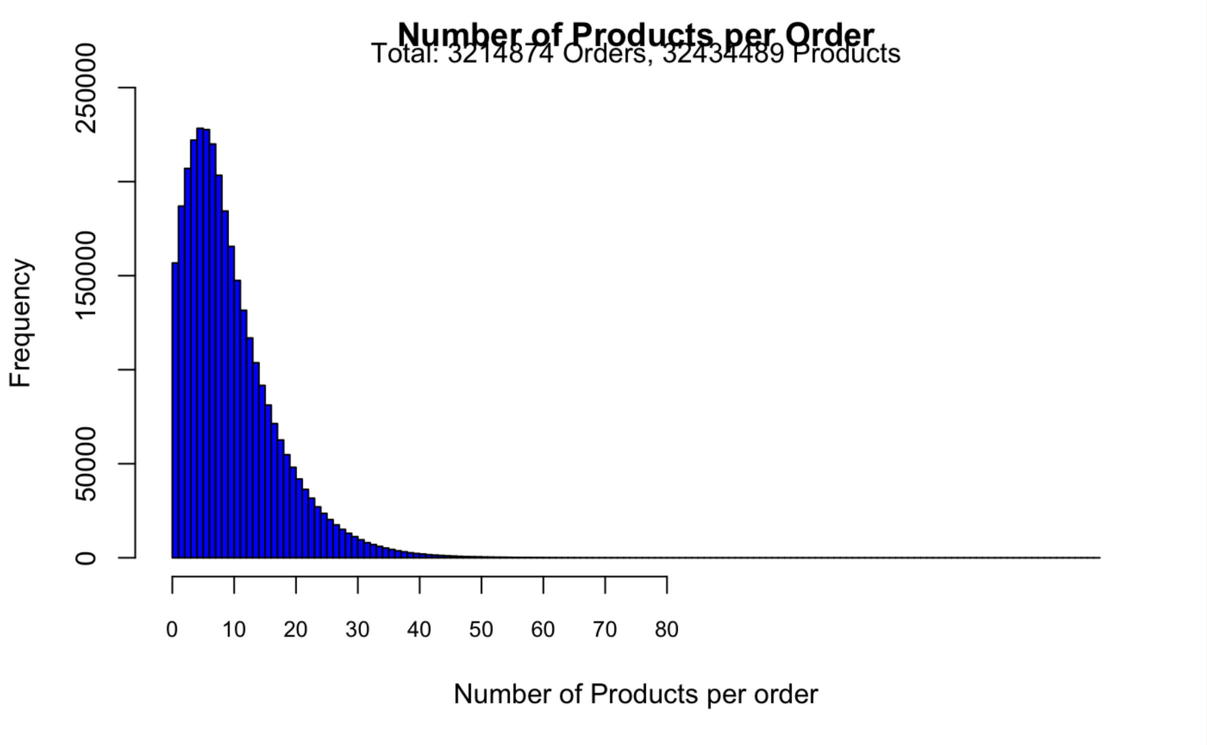
Support of at least 0.05



The productsID in the x-axis indicated the most frequently brought products.

Total Product per order

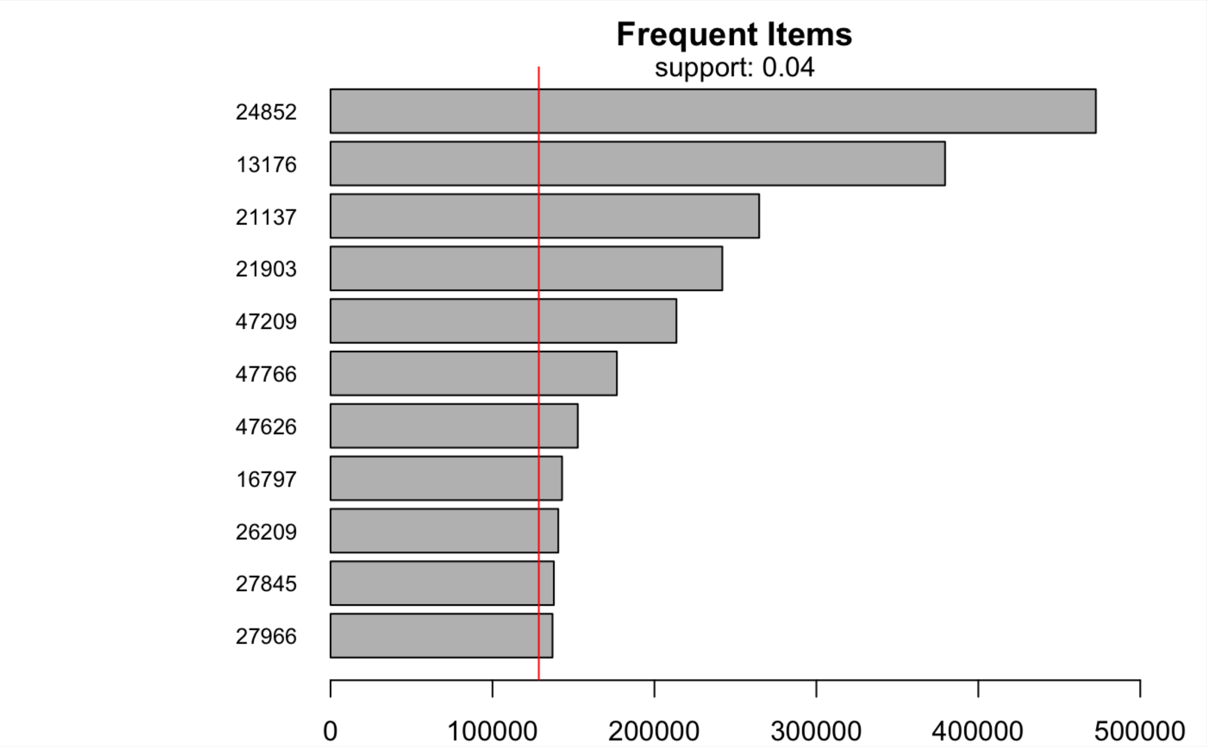
Shows number of products per order



This plot shows customer usually buy around 1 to 40 products per order.

Frequent Products

Support is 0.04 which is about 4%

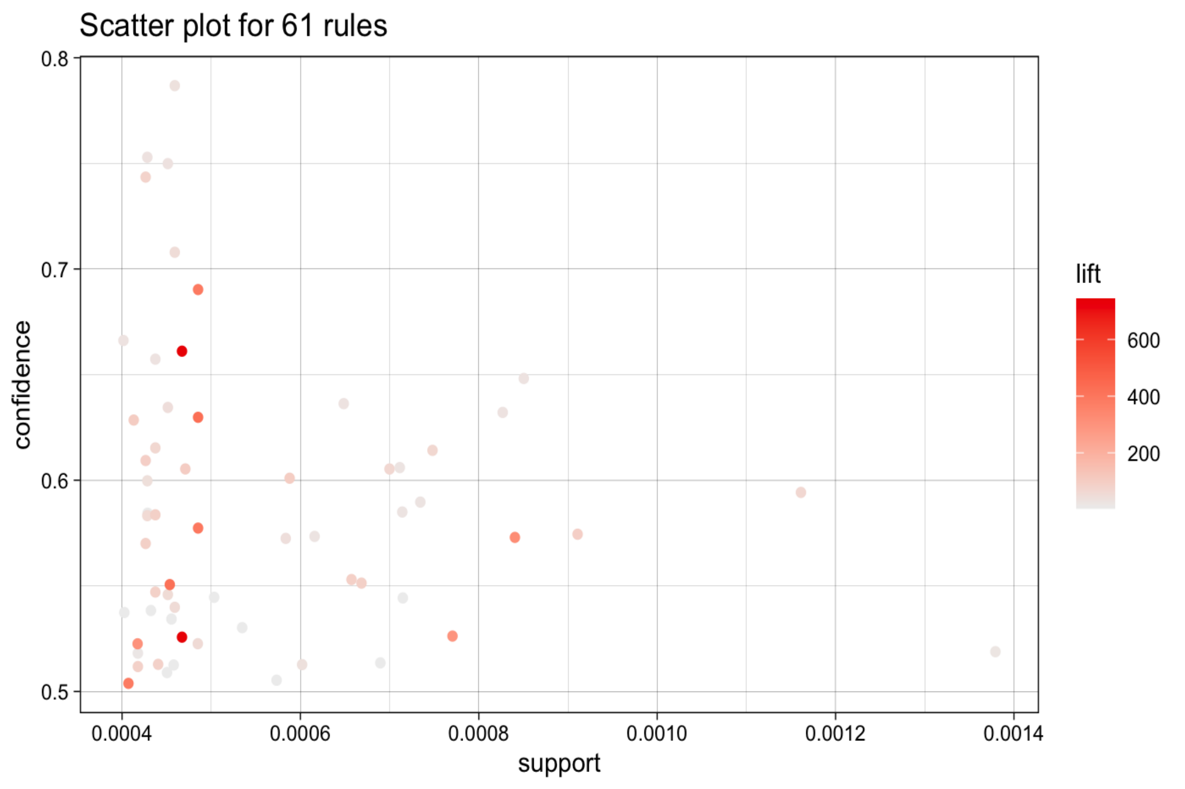


This plot shows there are 11 products occur when the support is set at 0.04.

This means these products are found in 4% of the total transactions which is approximately about 140k.

Scatter plot of rules

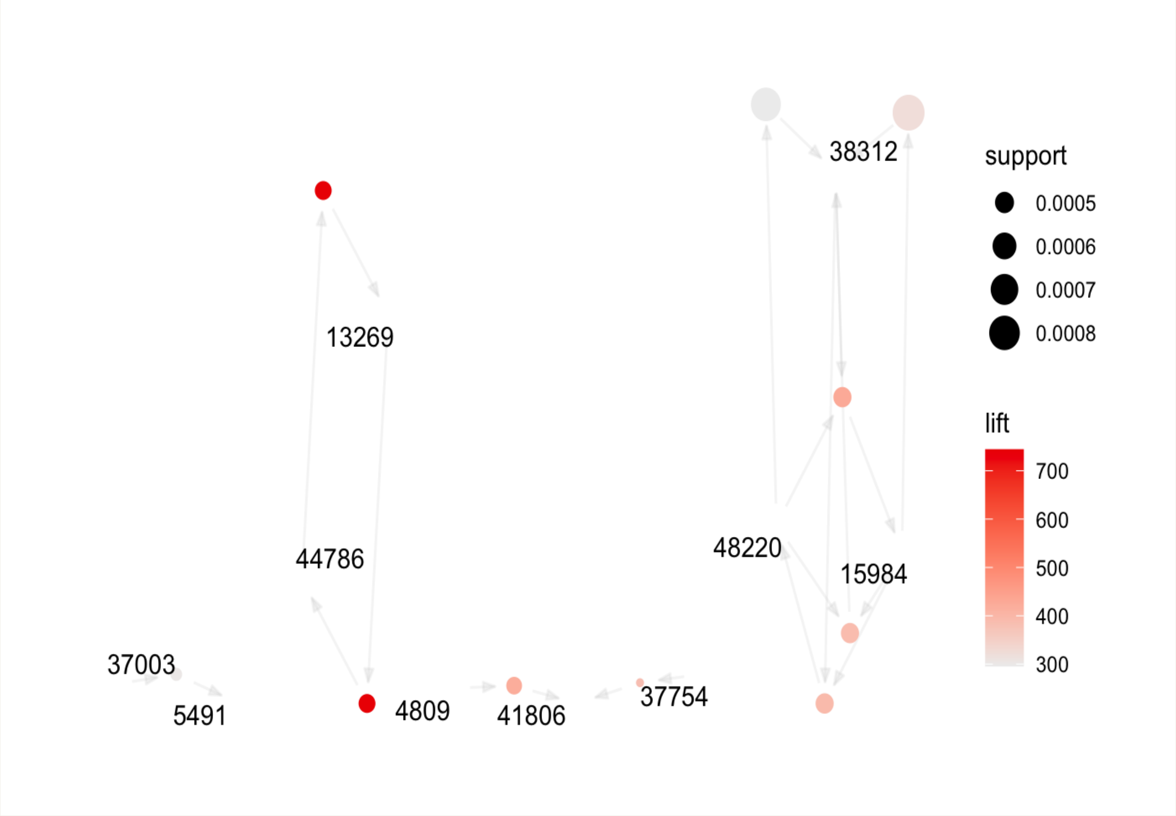
We found out associations from most frequent products which resulted in 11 product and can see good lift are between the confidence of 50-70% and a support of 0.0004 -0.0005



By lowering support value of 0.0004 since we want to produce 2 items and 3 product combinations.

This means that the product gets sold 1200 times out of 3 million transactions, or around 0.04% of all transactions.

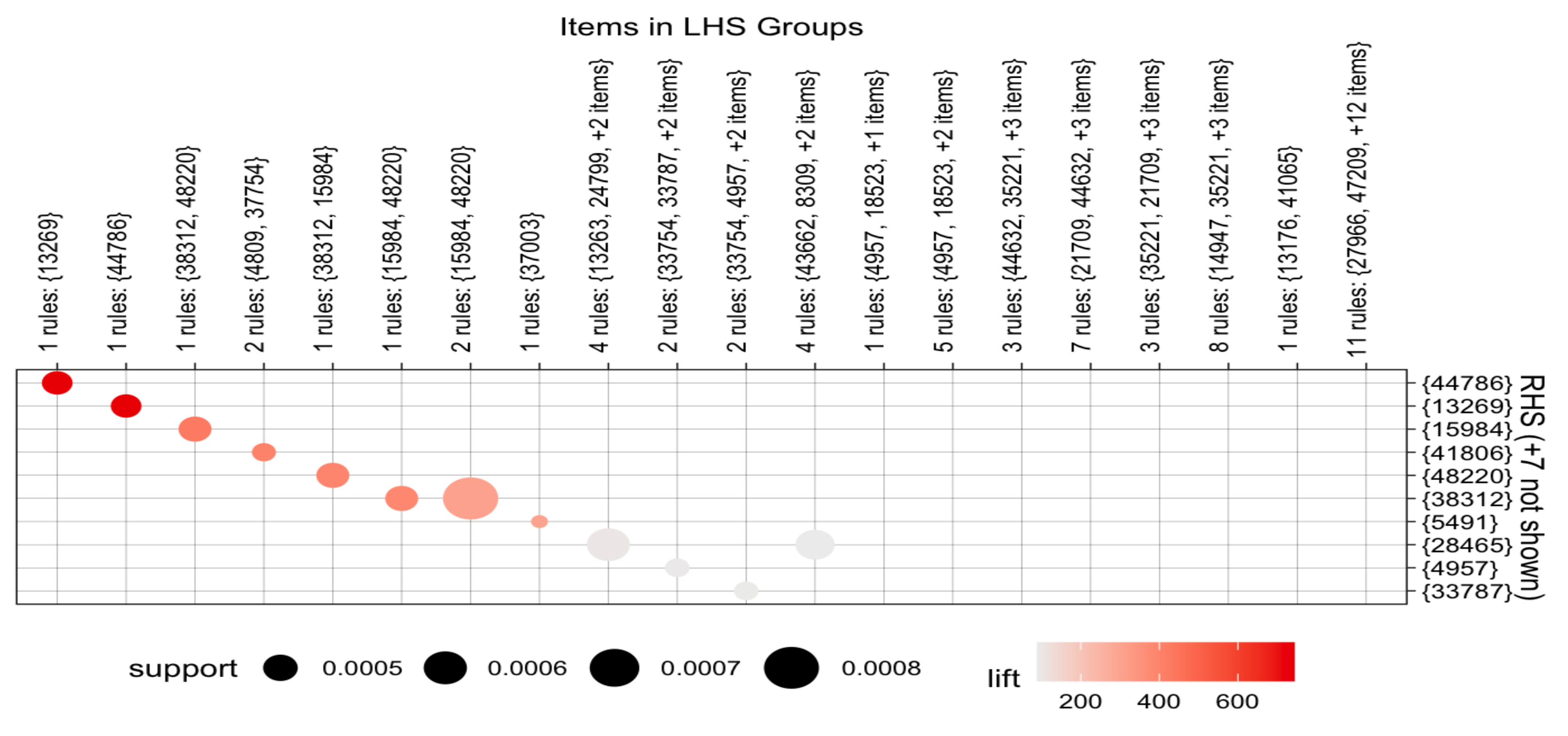
Rules and its confidence and support graph



Graph shows that ,
Product 13269 has a strong lift with product 4809

Support, Confidence and Lift of rules

Odered by lift value, shows the effectiveness of the rule, and create product combinations as a result



Confidence is set at 50%

List of 61 rules is provided.

ML Model for Prediction next order

We used XGBoost Classifier for prediction

we used Product and User metrics to make our prediction

Product metrics grouped by product ID	User metrics grouped by User ID
Total Orders	Total Orders
Total Order ratio for each product	Mean day of the week
Mean of add to cart	Mean hour of the day
Total times the product was reordered.	Mean day of the week and hour of the day
	Days since the prior order

We used combination metrics Group by user_id and product_id to perform our classification

Classification Metrices	
1	Total orders
2	Mean add to cart
3	Total reorders

Hyper Parameters Matrix

eta	0.1
max_depth	6
min_child_weight	10
gamma	0.7
subsample	0.77
colsample_bytree	0.95
alpha	0.00002
lambda	10

Combination metrics and scores are converted into matrix

we split the data to test and train and then performed the training of the model with these hyperparameters.

The model is used to predict the next orders where if the product is ordered (a probability greater than 21%) then we consider that product is reordered in a particular order.

Final Evaluation

Accuracy would not be the right measure to evaluate the model due to imbalanced data set.

So we have used F1 score and obtained score of 0.404.

Conclusion

By employing the results of the association analysis and XGBoost classifier model , we can predict the products that would be brought by the user in near future .

By using these analytics the company could perform curated results and thus increasing the business revenue.

Rules are refined based on Confidence ,support and lift metrics and combinations

Future Work

Further improve of the F1 score

Identify other methods to handle imbalanced data set, which includes hyperparameter tuning and other supervised classification models

Includes prediction based on neural nets, deep learning and using different metrics to predict the next buy.

Also, Collaborative filtering can be used to suggest products to customers