**PROJECT REPORT**

**Grocery Kart-Market Basket Analysis**

*Submitted towards the partial fulfillment of the criteria for award of Data Science and Machine Learning Pro-degree by Imarticus*

*Submitted By:*

*Bharanitharan.B*

*Course and Batch: DSP20*

*2019-2020*



# **Abstract**

# Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items. For example, if you are in an English pub and you buy a pint of beer and don't buy a bar meal, you are more likely to buy crisps (US. chips) at the same time than somebody who didn't buy beer. In this analysis, a forecasting model is developed using machine learning algorithms to improve the accurately forecasts product sales.

# **Acknowledgements**

We are using this opportunity to express our gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have **Arul Francis** as our mentor. He has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: August, 2020

Place: Chennai

**Certificate of Completion**

I hereby certify that the project titled “Grocery Kart- Market Basket Analysis” was undertaken and completed under my supervision by Bharanitharan.B from the batch DSP20

Mentor: Arul Francis

Date: August, 2020

Place: Chennai

**Table of Contents**

Abstract 2

Acknowledgements 2

Certificate of Completion 3

CHAPTER 1: INTRODUCTION 6

1.1 Objective of the Study 6

1.2 Need of the Study 6

1.3 Data Source 6

1.4 Tools & Approach 7

CHAPTER 2: DATA PREPARATION AND UNDERSTANDING 7

2.1 Data Extraction and loading 8

2.2 Exploratory Data Analysis 8

2.2.1 Count of data in each Data Set 8

2.2.2 Minimum and Maximum number of Orders 9

2.2.3 Most orders by a customer in Day of a Week 10

2.2.4 Most Orders by a Customer in Hour of a Day 11

2.2.5 Frequency of Day of Week Vs Hour of Day 12

2.2.6 Top 10 Products that are Ordered in prior, Train, Test 13

2.2.7 Top 20 Aisles that are Ordered in prior, Train, Test 16

2.2.8 Top 20 Departments that are Ordered in prior, Train, Test 19

2.2.7 Department Wise Re-Order ratio 21

CHAPTER 3: MACHINE LEARNING MODELS 24

3.1 Apriori Algorithm 24

3.1 Logistic Regression 25

3.2 Random Forest Classifier 26

3.3 XGBoost 27

CHAPTER 4: RECOMMONDATION AND CONCLUSION 29

**CHAPTER 1: INTRODUCTION**

## **Objective of the study**

In the present world, more and more people are shifting from buying groceries offline (i.e. by visiting a neighborhood store) to online (i.e. ordering through an app). GroceryKart is one such platform which lets users order groceries online. After selecting products through the GroceryKart app, personal shoppers review the order and do the in-store shopping and delivery for the customers. The objective of this capstone is to Predict whether a product will be reordered or not in the future by the customer

## **Need of the Study**

A very classic problem faced by such e-commerce websites is understanding the purchase behavior of a customer. If they can correctly predict which products the customers going to buy before the customer orders them, it can give them a huge advantage in terms of warehouse stocking, delivery times, marketing strategy, improving customer experience on their app etc.

## **Data Source**

Data source for this project is given by 7 datasets. (CSV file format)

1. **Aisle**-The aisle from which the product was ordered, examples: energy granola bars, prepared soups salads etc.
2. **Product**-Contains details on the products.
3. **Department**-The department of the store from which a product was ordered, examples: frozen, bakery etc.
4. **Orders**-This file tells to which set (prior, train, test) an order belongs. We are predicting reordered items only for the test set orders.
5. **Order\_Product\_prior**-This file contains previous order contents for all customers. 'reordered' indicates that the customer has a previous order that contains the product.
6. **Order\_Product\_train**- Similar to the prior file in structure, difference being, that the above contains all prior orders before a particular order, which is present in this file.
7. **Order\_Product\_test**- Similar to the above in structure, difference being, this file contains present data which needs to be tested based on the model we build

## **Tools & Approach**

## Since the dataset provided is huge, we cannot perform the EDA as well as the model building in our PC. So, I decided to use the Google Colab for the same. Also used Python as a Programming language to get the desired outcome

In the EDA, used Pandas, NumPy along with visualization libraries like seaborn and matplotlib to find various relationships as well as the answers to some questions.

# **CHAPTER 2: DATA PREPARATION AND UNDERSTANDING**

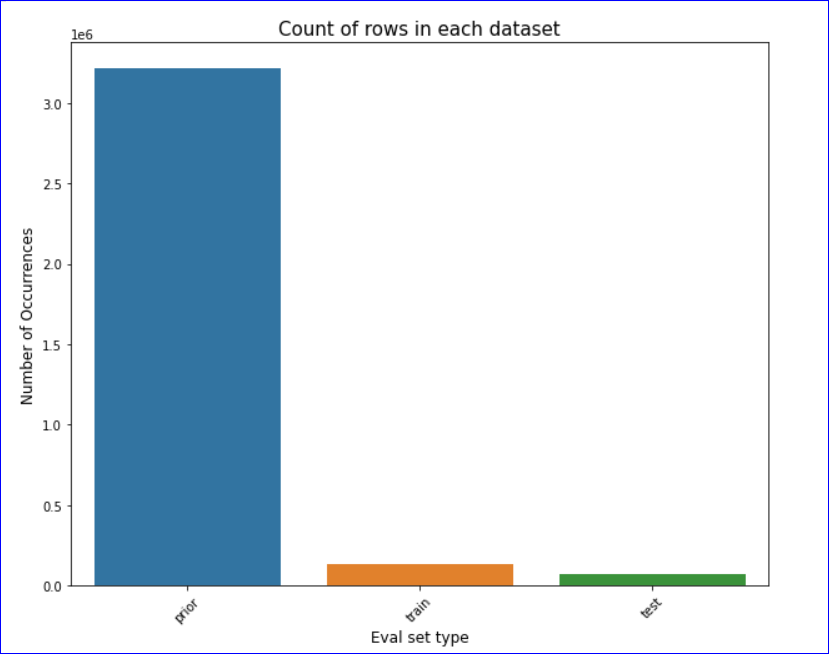
## **2.1 Data Extraction and loading**

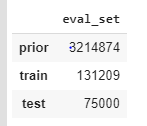
All the Datasets are imported into Google Colab with a zipped folded and unzipped using Zipfile library and Exacted with the help of Pandas and checked the shape due to dataset provided is huge used Colab for all the EDA and modelling purpose.

2.2 Exploratory Data Analysis:

Explore all the variables and check whether the data has null value. Except Orders dataset other dataset doesn’t have any null values. In order dataset also, only variable” Days\_Since\_Prior\_Order” alone has null values. But that also means that the user is doing the first order for that particular product so it was filled with Zeros.

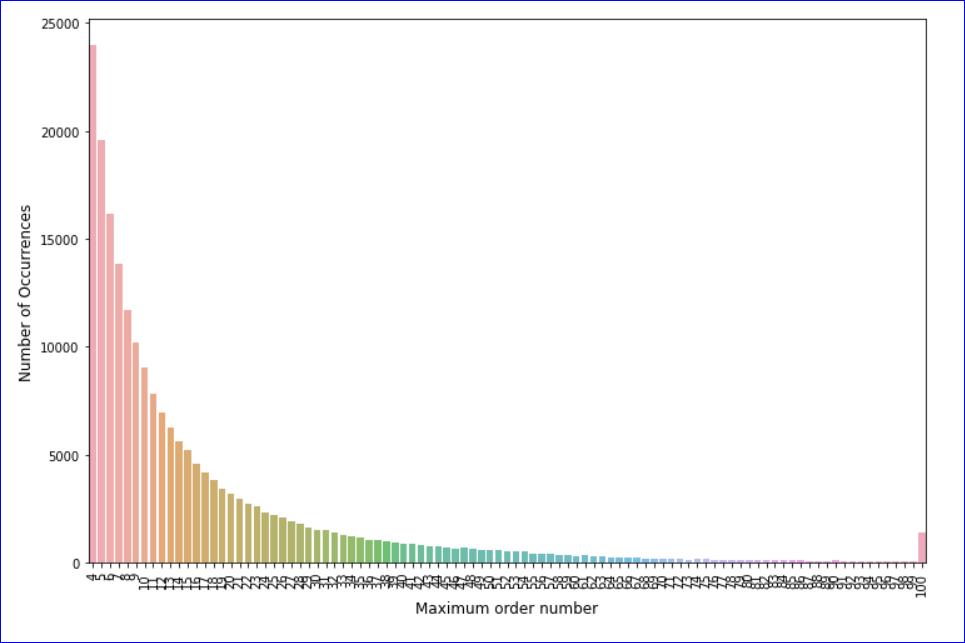
**2.2.1Count of Data in each Dataset**





First all orders are classsified as prior orders and last order. the last order is further divided into training set and test set. So prior order provides the history of a users ordering habit. The model for prediction will need evaluation based on training set. And the final prediction model will be evaluated for accuracy on the test set.

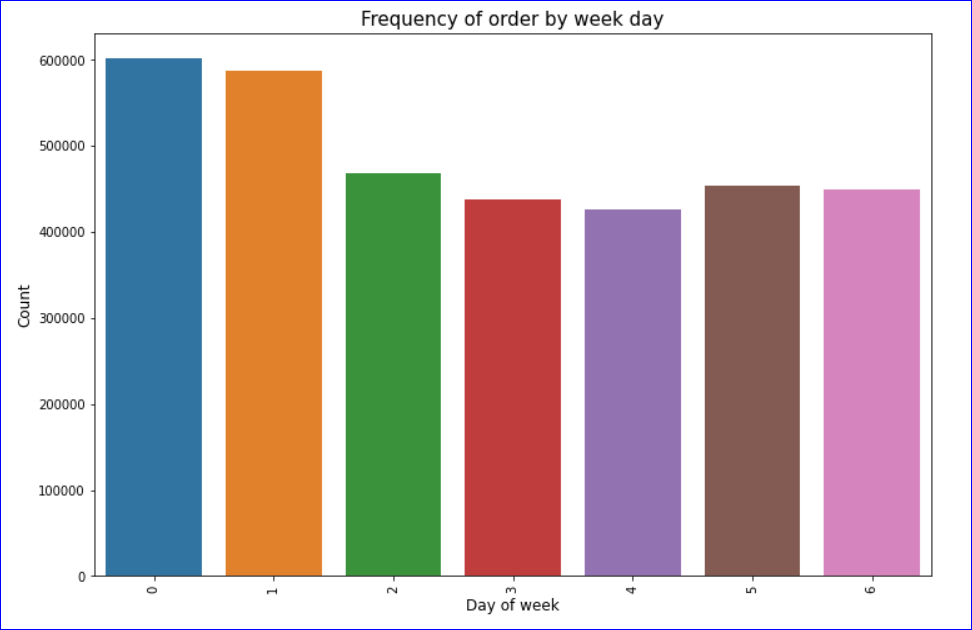
**2.2.2 Maximum and Minimum numbers of Orders by a User**



From the above graph it shows clearly that

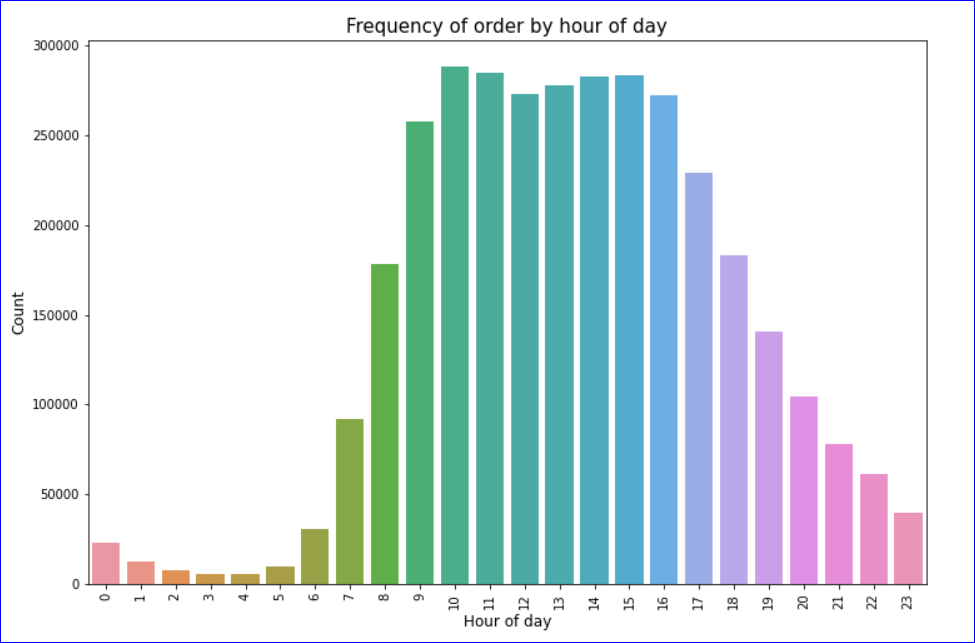
1. Minimum number of Orders placed by a user is 4
2. Maximum number of orders placed by a User is 100

**2.2.3 Count of Orders in a Day of Week**



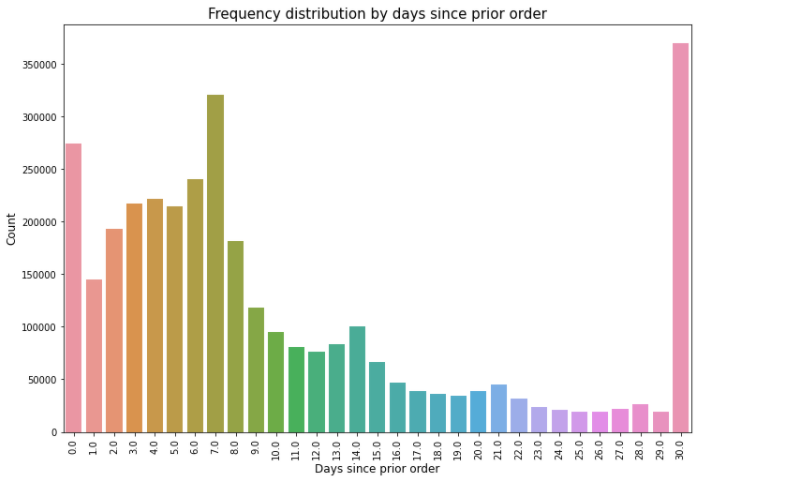
From the above graph, it’s very clear that maximum number of orders are placed in first two days of a week. i.e Sunday and Monday

**2.2.4 Frequency of Orders by Hour of a Day**



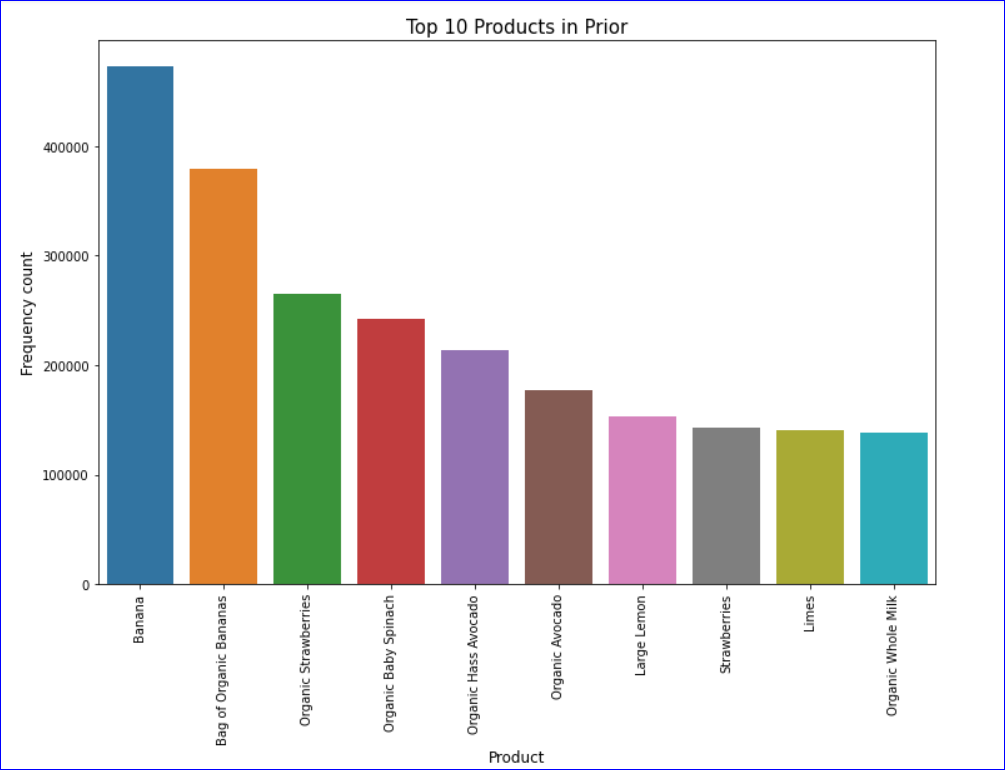
From the above, it’s clear that maximum number of Orders are placed in-between 9AM to 5PM

**2.2.5 Frequency distribution by days since prior order**



Maximum number of orders are placed in a Gap of 6 to 7 days

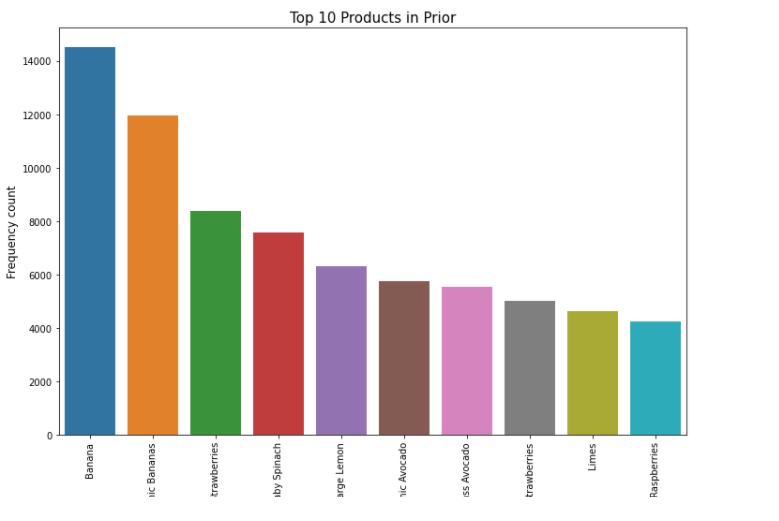
**2.2.6 Top 10 Products that are Ordered in prior**



Top 5 Products

1. Bananas
2. Bag of Organic Bananas
3. Organic Strawberries
4. Organic Baby Spinach
5. Organic Hass Avocado

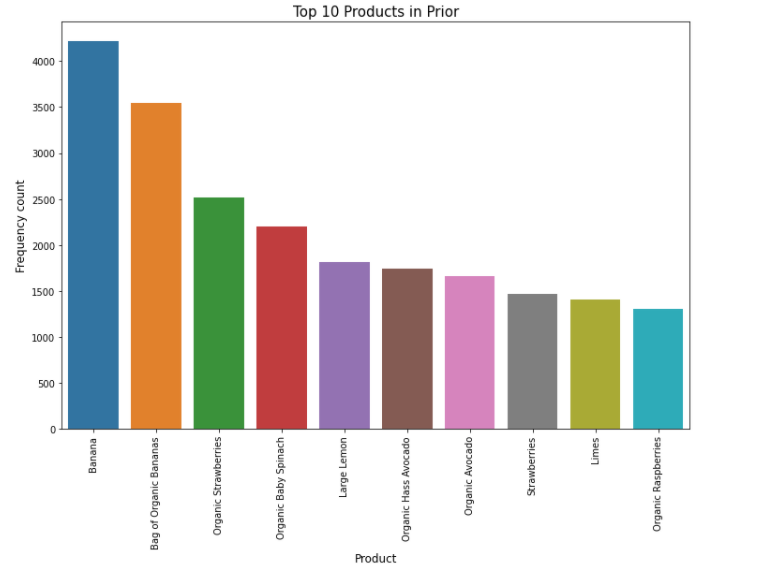
**Top 10 Products that are ordered in Train**



Top 5 Products

1. Bananas
2. Bag of Organic Bananas
3. Organic Strawberries
4. Organic Baby Spinach
5. Large lemon

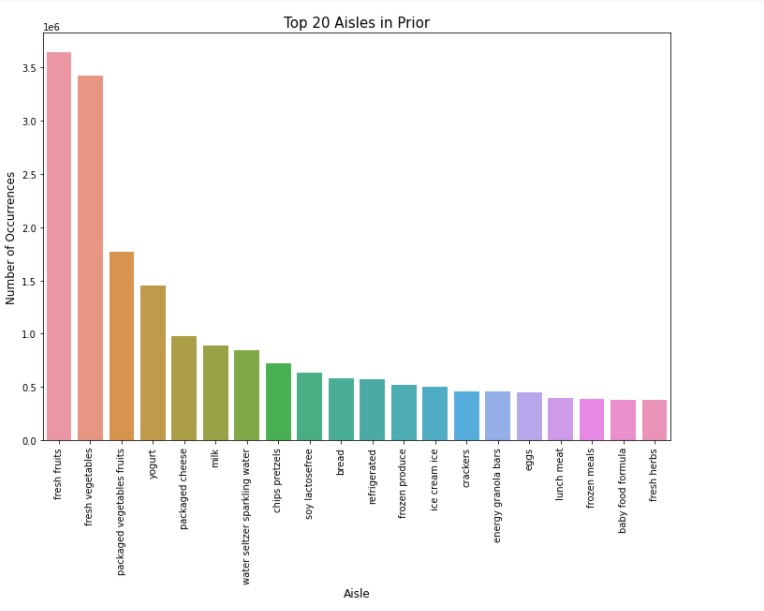
**Top 10 Products that are Ordered in TEST**



Top 5 Products

1. Bananas
2. Bag of Organic Bananas
3. Organic Strawberries
4. Organic Baby Spinach
5. Large lemon

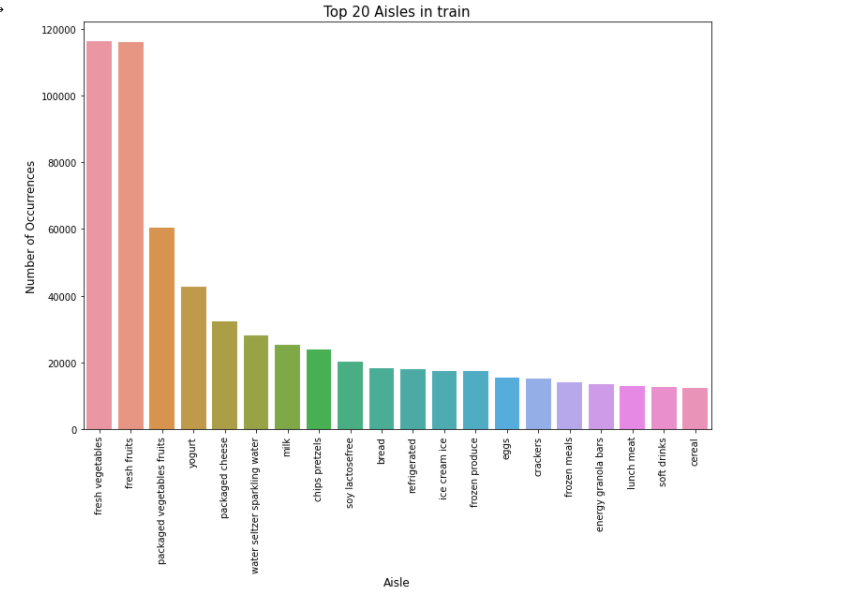
**2.2.6 Top 20 Aisles that are Ordered in Prior**



Top 5 Aisles

1. fresh vegetables
2. fresh fruits
3. packaged vegetables fruits
4. yogurt
5. packaged cheese

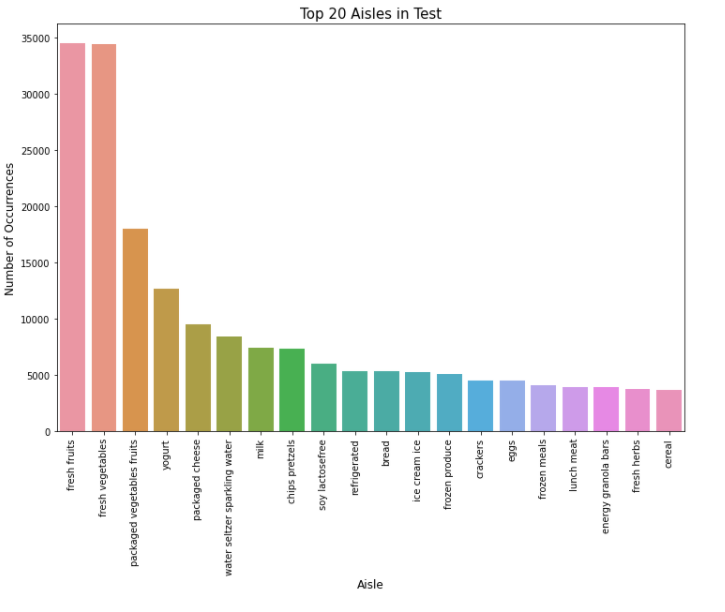
**Top 20 Aisles that are Ordered in Train**



Top 5 Aisles

1. fresh vegetables
2. fresh fruits
3. packaged vegetables fruits
4. yogurt
5. packaged cheese

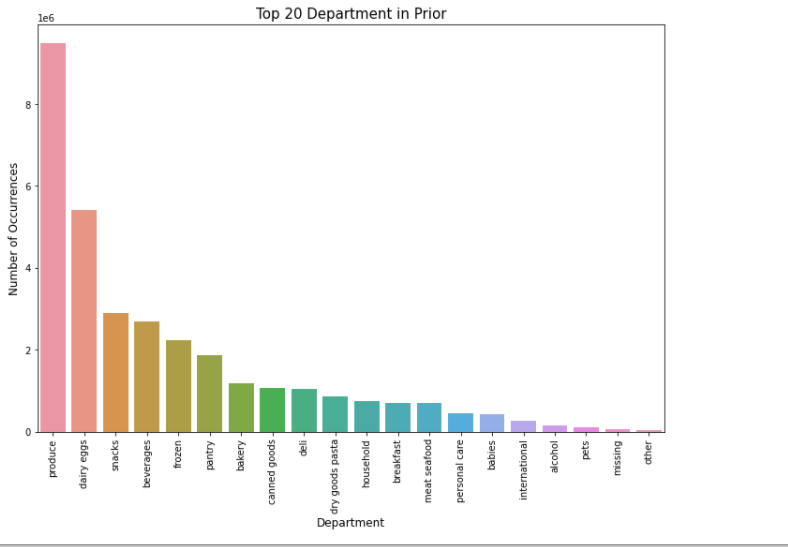
**Top 20 Aisles that are Ordered in Test**



Top 5 Aisles

1. fresh fruits
2. fresh vegetables
3. packaged vegetables fruits
4. yogurt
5. packaged cheese

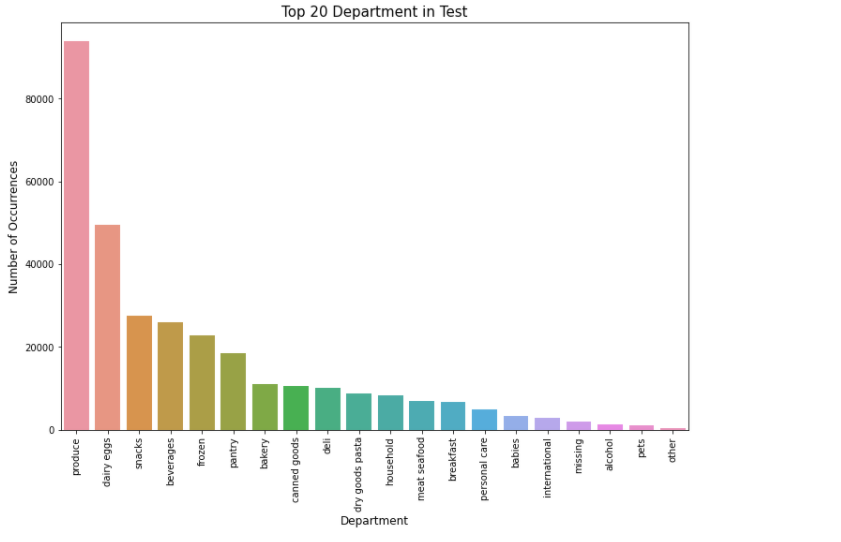
**2.2.8 Top 20 Department receive Orders in prior**



Top 5 Departments

1. produce
2. dairy eggs
3. Snacks
4. beverages
5. frozen

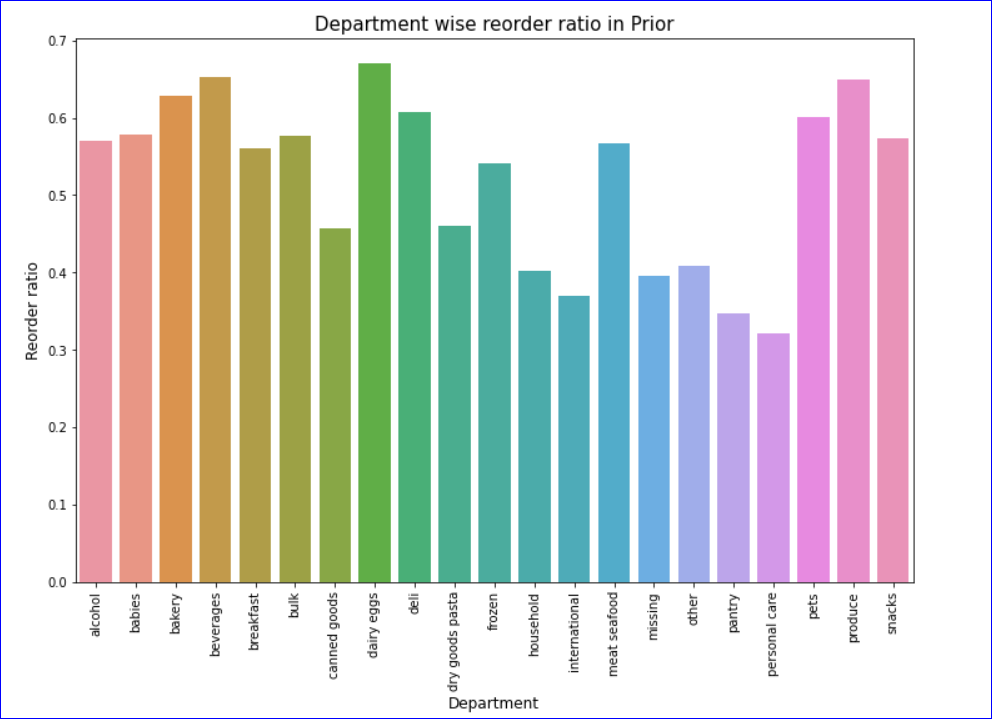
**Top 20 Department receive Orders in Test**



Top 5 Departments

1. produce
2. dairy eggs
3. Snacks
4. beverages
5. frozen

**2.2.7 Department Wise Re-Order ratio**



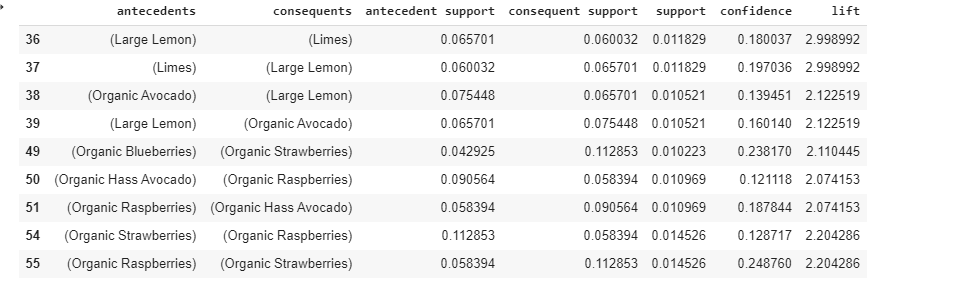
# 

# 

# **CHAPTER 3: MACHINE LEARNING MODELS**

**3.1 Apriori Algorithm**

Apriori Algorithm was used to find the antecedents, consequents, antecedent support, consequent support, support, confidence, lift



Above are the top combination of products which users brought together.

So, based on this, we can recommend the user for the products.

# **3.2 Logistic Regression to predict reorder**

# 

Build a model using X-train and Y-train.

Tested on X-test and got around 90% accuracy.(0.902)

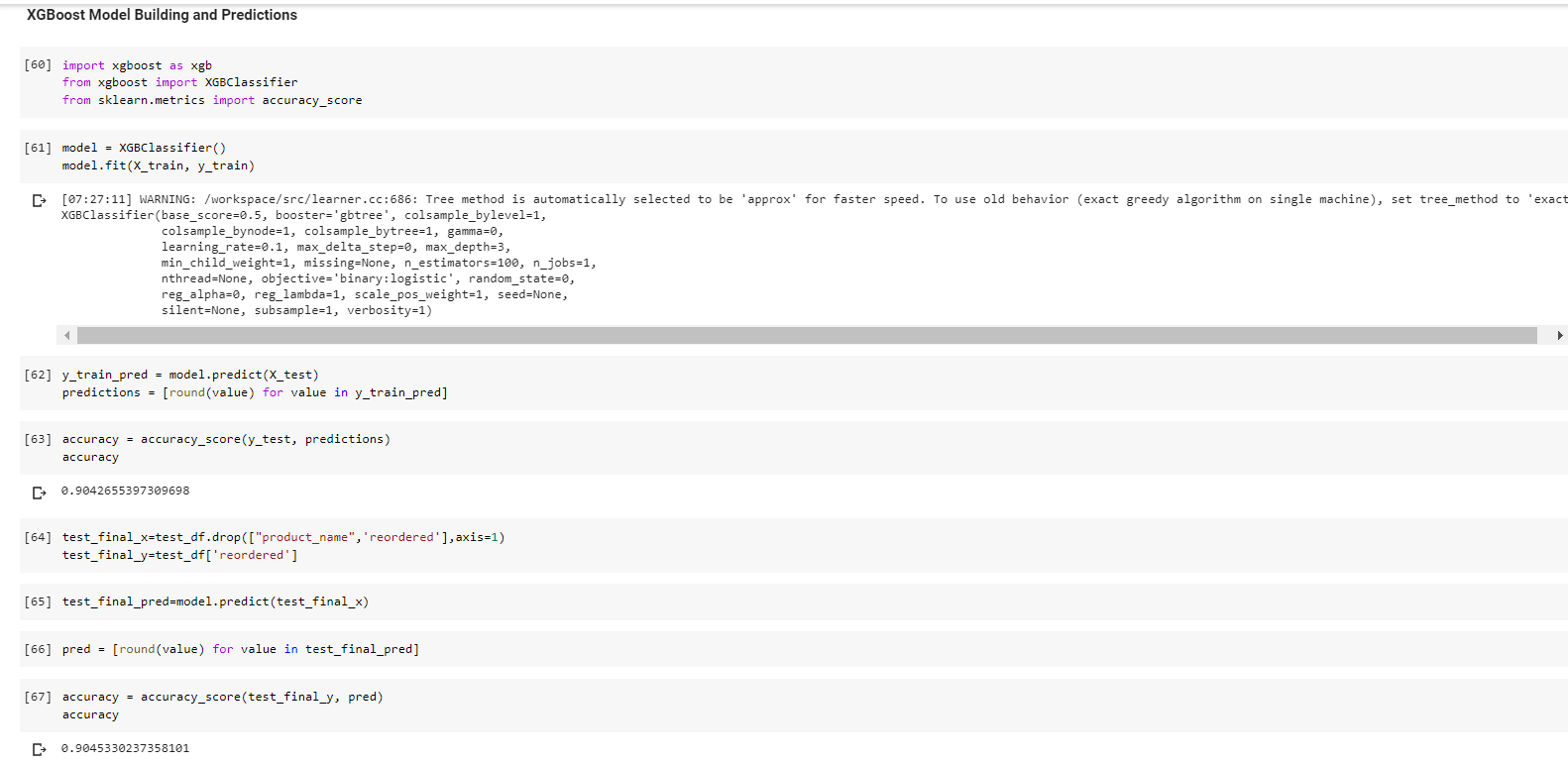
**3.3 Random Forest Classifier**



Build a model using X-train and Y-train.

Tested on X-test and got accuracy around 90% (0.905)

**3.4 XGBoost**



Build a model using X-train and Y-train

Tested on X-test and got around 90% accuracy (0.904)

**CHAPTER 4: RECOMMENDATIONS AND CONCLUSION:**

Even though the accuracy of X-test is high on Random Forest Classifier, the final accuracy on the test data set is low when compared to other machine learning models like Linear Regression and XGBoost.

So, it’s better to use XGBoosting Technique for this kind of transactional data/ Market basket analysis. If we were asked for recommendation models, then Apriori algorithm would have helped us to give a very good recommendation.