**Customer Behavior Analytics and Recommendation Algorithms**

*Jehan Joshi*

Table of Contents

[**Executive Summary** 6](#_Toc500361316)

[**Introduction** 7](#_Toc500361317)

[**Problem Statement** 8](#_Toc500361318)

[**Purpose** 9](#_Toc500361319)

[**Research Questions** 9](#_Toc500361320)

[**Research Methodology** 9](#_Toc500361321)

[**Scope, Assumptions and Limitations** 10](#_Toc500361322)

[**Literature Review** 11](#_Toc500361323)

[**Systems Development and General Methodology** 18](#_Toc500361324)

[**Description and Presentation of the System** 20](#_Toc500361325)

[**Research Question 1: Identifying the key factors used in analysis of the customer basket** 22](#_Toc500361326)

[**Research Question 2: Role of Predictive Analytics in Customer Behavior Analysis and Product Recommendations for Retail Industry** 37](#_Toc500361327)

[**Conclusion** 53](#_Toc500361328)

[**References** 55](#_Toc500361329)

[**Appendix A: Code** 58](#_Toc500361330)

[Step 0: Load the data in RStudio 58](#_Toc500361331)

[Step 1: Glimpse of the data 59](#_Toc500361332)

[Step 2: Checking for any missing values in each table column wise 62](#_Toc500361333)

[Step 3: Data Preparation 63](#_Toc500361334)

[Step 4: Performing Exploratory Analysis on the Data Sets 63](#_Toc500361335)

[Step 5: Association Rule mining 67](#_Toc500361336)

[**Appendix B: Data Dictionary** 70](#_Toc500361337)

Table of Figures

[*Figure 1 Data Preparation* 19](#_Toc499569161)

[*Figure 2 Hours of maximum orders* 21](#_Toc499569162)

[*Figure 3 Order distribution across week* 22](#_Toc499569163)

[*Figure 4 Best Business Hours* 23](#_Toc499569164)

[*Figure 5 Reorder frequency of customers* 24](#_Toc499569165)

[*Figure 6 Prior Orders* 24](#_Toc499569166)

[*Figure 7 Order/Reorder Proportion* 25](#_Toc499569167)

[*Figure 8 Count of items per order* 26](#_Toc499569168)

[*Figure 9 Best sellers* 26](#_Toc499569169)

[*Figure 10 The front runners* 27](#_Toc499569170)

[*Figure 11 Relation between last ordered and reorder* 28](#_Toc499569171)

[*Figure 12 Product Reorder Probability* 28](#_Toc499569172)

[*Figure 13 Organic vs Not-Organic* 29](#_Toc499569173)

[*Figure 14 Reorder frequency* 30](#_Toc499569174)

[*Figure 15 Organization of aisles within the departments* 31](#_Toc499569175)

[*Figure 16 Ranking aisles based on number of unique products* 32](#_Toc499569176)

[*Figure 17 Ranking departments based on number of unique products* 33](#_Toc499569177)

[*Figure 18 Ranking aisles based on number of unique products in the maximum orders* 34](#_Toc499569178)

[*Figure 19 Ranking departments based on number of unique products in the maximum orders* 34](#_Toc499569179)

[*Figure 20 Predictive Analytics* 36](#_Toc499569180)

[*Figure 21 Items per basket* 40](#_Toc499569181)

[*Figure 22 Items brought frequently* 41](#_Toc499569182)

[*Figure 23 Rule Plot 1* 42](#_Toc499569183)

[*Figure 24 Rule plot 2* 44](#_Toc499569184)

List of Tables

[*Table 1 Rule 1 – Sorted on Lift* 43](#_Toc499815907)

[*Table 2 Rule 1 - Sorted on confidence* 43](#_Toc499815908)

[*Table 3 Rule 2 - Sorted on lift* 45](#_Toc499815909)

[*Table 4 Rule 2 - Sorted on confidence* 45](#_Toc499815910)

[*Table 5 Product Recommendation* 46](#_Toc499815911)

[*Table 6 Item based probability for the users (Sample Set)* 48](#_Toc499815912)

[*Table 7 Item Based Probability for Rel\_Freq > 0.8 (Sample Set)* 50](#_Toc499815913)

# ***Executive Summary***

*With the advent of data analytics, organizations across the world are heavily relying on the beneficial insights obtained from data to develop key operational strategies. Organizations have realized that in order to be a market leader and to sustain the competitive advantage, embracement of the power of data is indispensable. The operational and transactional data garnered from the day to day operations when subjected to optimum analysis can provide the organizations with an efficient acumen for understanding their operations and to identify the needed enhancements. The paper revolves around the focal point of how data analytics can bring in new wave of changes in the way retailers conduct the business. Identifying the customer behavior patterns and understanding the hidden details from the customer purchase history can help retailers to boost their sales in a considerable manner by utilizing techniques like cross-selling strategy. The paper discusses these concepts in detail using the dataset from Instacart, a renowned groceries-on-the go application. Instacart is valued for nearly $3 billion in the startup space. As the competition is fiercer and to win their fight over the giants like Amazon, Instacart roots its operational strategies on data analytics.*

*As a part of this paper, we performed “Market Basket Analysis” on the transactional dataset from Instacart to understand the relationship between customer purchases and to predict the nature of purchase of customers. Market Basket Analysis and Recommendation Analytics are two areas where retail industry put all its bets on. Market Basket analysis is a marketing technique that facilitates the study of customer purchase trends along with promotions and pricing of products. The set of items customer buys is called as an item set and market basket analysis focuses on finding relationships between those products. Market Basket Analysis also gives insights into the impulse purchasing which contributes to a larger percentage of retail shopping. The paper answers two main research questions, the key factors involved in identifying a customer’s basket and understanding the role of recommendation algorithms in the retail market. The analysis including but not limited to understanding the customer-reordering patterns, purchasing patterns on weekdays, weekend and on specific time slots, price impact on purchasing decisions. A detailed study of the dataset has opened the Pandora box. Understanding the relationship between various products is very crucial for retailing. Knowing the probability of a certain product to be sold more can give the retailers heads up and hence a chance to proactively manage their inventory. We have used the unsupervised technique to understand the purchasing patterns of each customer and predicted the items which can be bought by them in future. In a nutshell, the project sheds detailed insights into the customer behavior analytics which retailers can rely upon to roll out their operational strategies aimed at overall organizational success.*

# **Introduction**

Market basket analysis is a marketing analytic technique which facilitates the study of customer purchase trends coupled with promotions. The intent of the analysis is to help retailers understand the pattern changes in the shopping behavior of customers and thereby to come up with promotions on various products based on the study. One major fact in retail which is often noticed or looked after by the retailers is how the products are connected. For example, if a customer buys sugar and coffee powder, the chances of the same customer buying a bottle of milk is higher. By understanding this basic thumb rule, promotions can be placed accordingly. In the marketing terms, this is known as association and the association rules obtained as the outcome can be used to provide customized recommendations to customers by the retailers.

In this project, we use a dataset from ‘Instacart’, a smartphone application which is available on both Android and iOS platforms. Instacart enables shoppers to place their order with the products from the nearby retail shops like Walmart, Target and later the products will be delivered to their doorstep by independent contractors working for Instacart. We can call it as the ‘Uber for Groceries’. Once the order is placed, the delivery will happen within 2-3 hours which actually provides a better-enhanced experience that the online shopping. Instacart’s core workflow model depends on various entities such as customer’s familiarity with the local shop and its reputation. For example, a customer will have a good knowledge on the tomato pack in the Walmart near their home and will be confident to buy that packet rather than one from Amazon or so. Since we are talking about perishable goods and ‘specific’ groceries and produce, the factor of ‘returns’ is not coming into the picture. Selecting data from such an application will provide us with a more sophisticated data points which is tightly coupled with new generation smartphone application users. The data points can be more prevalent in nature compared to traditional retail datasets.

By analyzing this dataset, the intent is to espy the underlying connections among the customer purchases. R programming language was used for detailed analysis on the dataset and hence to predict the products that would be present in a customer’s order. R provides an excellent suite of algorithms for MBA in the various library packages.

# **Problem Statement**

The intention of this study is to analyze the effectiveness of “groceries on the go” shopping applications and how they enhance the shopping experience for the customers. Cross selling is one among the most preferred way of revenue generation by marketers. With our study, we intend to increase the efficiency of the cross-selling strategy by making it more user specific. For the analysis, we have taken an anonymized dataset of a unicorn startup “Instacart”. By analyzing the dataset, we aim to study the shopping behavior of the customers and thereby understanding the data patterns. Mobile applications reinvented the computing a decade back. They literally changed the way human beings do their daily activities. Now coupled with the powerful data analytics tools, they are restructuring the human behavior. Instacart offers a shopping experience, which has a nature of online shopping and yet has the flavor of brick and mortar shopping. With this study, our focus is to emphasize how an application influences the shopping habit of people and how it can predict the shopping nature of repeated customers.

# **Purpose**

The purpose of this paper is to identify the role of market basket analysis in optimizing the shopping experience of users by analyzing a dataset belonging to an online grocery delivery store. We intend to explore the effective strategy to be used by the grocery on the go applications to predict the possible products that would be reordered by a customer. Along with this exploration, the paper intends to conduct a thorough study of the market basket analysis technique and analyze the various methods used to conduct the market basket analysis. Also, the paper would discuss about the different recommendation systems and algorithms that can be used in building a model for the predicting a customer’s basket

# **Research Questions**

1. What are the key factors analyzed in predicting a customer's basket?
2. How predictive analytics can bring profuse changes in customer behavior analysis and product recommendations for retail industry?

# **Research Methodology**

Data Analysis, is all about trying different approaches and selecting which approach is the closest fit to answering your problems.

Firstly, studying similar problems to understand the spectrum of solutions that the paper can focus on, which would involve conducting an exhaustive literature review. After which the discusses about the data at hand identifying underlying relationships and trends in the data that would help a retailer understand the problem much better and further point them in the right direction.

This could be done by data visualization by using various tools such as R

Once the intricacies of the data are well understood, the process would start by developing a model that would solve the problem. After the development of the model the next process would be to implement this model using the R interface and test its performance on a small section of the data. If the model meets the required evaluating parameters it would be replicated throughout the data.

To summarize these steps, they can be divided into different phases as the paper progresses with the project

* Analyzing the data/Exploratory Analysis
* Model design
* Model testing using an interface such as R
* Model Implementation

# **Scope, Assumptions and Limitations**

The scope of the project is to apply principles of data analytics for studying the behavior of the “grocery on the go” customers. The study aims to understand the shopping nature of the customers by analyzing the underlying data patterns. In the study, the paper focuses on finding the association among the products bought by the customers and thereby helping the retailers to optimize the cross-selling feature.

**Assumption**

The underlying assumption in market basket analysis is that joint occurrence of two or more products in most baskets imply that these products are complements in purchase, therefore, purchase of one will lead to purchase of others. In the present study, the paper discusses into two item rules.

**Limitation**

* The analysis is based on the anonymized dataset provided by the corresponding company and the paper is not checking the validity of the data.
* The paper is not performing any kind of survey among the users to have the results verified.

# **Literature Review**

**Market Basket Analysis (MBA)**

Market basket analysis has been a buzz word in retail space for a very long time and recently its relevance has increasingly become popular in online retail space. Market basket analysis is a data mining technique commonly used in marketing and retailing to help the retailers understand the consumers purchase behavior. The term market basket analysis is named after the whole idea of a customer putting the purchased items into a shopping cart for the duration of the shopping (Vats, 2015). Understanding the purchase behavior helps organizations to make better business decisions. Organizations use these information for attracting more customers and increasing the overall sales by performing cross selling, upselling, designing store layout and in creating promotional offers. Analyzing a customer’s shopping behavior and deriving rules for creating predictive systems, which can help retailers in determining the cross-selling strategies and thereby increasing the overall sales is the main idea behind the MBA technique (Decker & Monien, 2003). The technique uses an assumption that the joint occurrence of two or more products in majority of the baskets implies the correlation between the items and hence gives a high probability of being bought together (Kamakura, 2012). Market basket analysis can be considered as an application of data mining technique**.** Usage of Market basket analysis originated in the field of marketing but it’s been effectively used in other areas such as bioinformatics, education, nuclear science, immunology and geophysics. Market basket technique can be grouped as an application of data mining technique. Data mining also known as knowledge discovery is the process of gathering useful information from large amounts of data following a step wise process of data selection, data cleaning, applying data mining techniques and interpretation of results. (Kaur & Kang, 2016). Market basket analysis can be performed by two alternative methods such as Association Rule mining and Time series Clustering.

**Association Rule Mining**

Association analysis is an unsupervised data mining technique which can be used in identifying relationship between the entities in a large data set. The output of the analysis will be rules that can be used to identify hidden relationship between the items in the data set. From the association rules the set of frequent items in a transaction can be identified. The technique is widely used for performing market basket analysis due to its ability in determining the relationship between items in the transaction data.

The techniques of association rule mining start by identification of most frequent item sets from the transaction dataset. Items are the base objects on which the analysis process is carried out. Transactions indicate the co-occurrence of a group of items together (Yali, 2012). Using the frequent item sets as the starting point, association rules are derived. The output of the market basket analysis will be a set of rules which can be used to predict the output of a customer purchase. Apriori algorithm is the most commonly used algorithm to perform the association analysis. Since the paper explains the basics of association rule, its mandatory to introduce the terms Antecedent and Resultant. For example, in an association rule A🡪B, A is the antecedent and B is the resultant. Few other terms that are commonly used in the explaining the efficiency of association rules are Confidence, Support and Lift ratio. Support value for a rule indicates the impact of a rule in the entire data set. For example, if we say the support for a rule “If Milk and Sugar then Butter” is 60%. It means that this rule affects 60% of the total data set. If the support value of the rule is less, then it indicates that the effect of the rule on the data set is considerably less. Support is a key measure in determining the rule as a low value of support means that the rule occurred just by chance. Considering the business perspective, it is not profitable to promote items that are rarely bought by the customers. Hence, the support measure is often used as a filter in removing irrelevant rules. Confidence of a rule can be defined as the likelihood of the occurrence of two items together. For example, confidence of the rule A🡪B can be defined as the probability of occurrence of A&B together to the probability of occurrence of A in the transaction. Hence confidence can be considered as a measure which determines the relevance of a rule. The lift ratio indicates the likelihood of an item Y to be purchased with the purchase of item X considering the popularity of item Y. Lift provides inference as what is the chance of an item, consider Y, to be purchased with another item (X) with respect to general buying rate of Y. Rules with value of lift ratio greater than one is considered as relevant (Gupta, Kumar, & Shaikh, 2015). The results of the association rule mining are sorted and selected with much caution using the measures of support, confidence and lift to ascertain the reliability of the rule in predicting the customer behavior. The process of association rule mining can be split into two major steps or subtasks as determining the frequently bought items and rule generation. From the first step, as the name infers, the frequently bought item sets are retrieved. In the second step using these frequent item sets high confidence rules known as strong rules are extracted.

**Time Series Clustering:**

Another method which can be utilized in performing analysis on market-basket data is Time series Clustering. There is a prevalent argument that the efficiency of the time series clustering method in analyzing large volumes of transaction data is better than association rule mining. In case of association rule mining, the data matrix created in the process will turn to be very large and sparse which require longer processing time. Whereas in the case of time series clustering the transaction data is summarized as time series there by resulting in a smaller data set. Also, the rules generated from the method will provide fewer insights when compared to the Time series clustering methods which can create insightful details even from a much smaller data set that will be only a fraction of the size of the data matrix which is required for the association analysis. Association rule mining does not consider the time factor while analyzing the transactions, but as the name implies since time series clustering groups the data on the temporal basis it is possible for the analyst to easily identify the products which are commonly purchased across a certain time period (Tan&Lau,2013). This analysis will help the retailer to plan accordingly and to increase the sales which serve the whole purpose of Market Basket Analysis. In a nutshell, based on the literature research it can be conveyed that Time series clustering is considered as a superior alternative to association rule mining. The technique that is used in this paper is Association rule mining.

**Applications of Market Basket Analysis (MBA)**

**Movie Business**

Market Basket Analysis has been used to predict the movie trend by finding association between different genres. By considering the Support and Confidence, which are two measures of any rule, usefulness and certainty of discovered rules are reflected. Based on associations of genres a viewer’s trend or movie business’s trend can be discovered (Chowdhury, 2013). Based on this trend distributors can pick movies which showed similar trend in the past during the same time period and launch genre kind of movies during the same time to maximize the profit. Considering only successful movies, which made a positive profit, associated rules have been generated for genres and the genres associated with profit with highest support have been tagged as a profitable trend(s). Though the profit or loss of new movie to be released is difficult to be predicted but predicting the possible movie trend based on genres is possible. This trend can be used by production houses to plan their release dates of their movies based on the trends predicted for genres to maximize their profitability (Chowdhury, 2013). The article with respect to this study had achieved a success rate of 86.36%, which is pretty good considering the high volatility that revolves around the movie business.

**Casino Business**

A particular casino owner was facing a challenge of understanding whether they were creating incremental win with their slot moves or were they just moving money around the floor. Market Basket Analysis was brought in to understand the player’s selection of machine behavior based on the relationship between slot machines. The insight gained from this was used for title changes, placements, machines swaps and bank swaps. In a process of 8 weeks, this resulted in $41.50 Average Daily Win increase realized for titles impacted by swaps, players who swapped titles increased their wallet by 5% in the post vs pre-Market Basket Analysis (Casino Floor Optimization, 2016).

**Traffic Accidents**

The increasing amount car crashes led into an investigation to find the associated factors that led to them. Though data mining methodologies were used earlier to analyze the car crash reports, the use of association rule mining was never put into much thoughts to analyze the car crash reports. Most significant factor noted in this analysis was that most of the crashes in a rainy weather were single vehicle ROR (Run-Off-Road) crashes and drivers with age of 55 and above, tend to involve in a crash more than others during a rainy weather in daylight (Das, 2014). On a straight level aligned roadways, most of the crashes during rainy weather was associated with the poor illumination of the roadways and young drivers in ages of 15-24 are the ones to be involved in a crash when there is no illumination on the roadways and when they are curve-aligned (Das, 2014). Another study determines similar rules for traffic accidents by presenting few interesting insights. The study provides rules that probability of adults involved in a crash was about 14% on a general basis (Donepudi, 2013) and was least if the day of the week was Friday and the crashes would involve two vehicles. The probability of crashes involving youth is about 1% where the lighting condition is poor and these crashes do not involve any other vehicle apart from the one driven by the driver and about 0.2% of the crashes caused by the adults are due to the improper condition of the vehicle (Donepudi, 2013).

**Other Applications**

Medical. In Medical, applying the rules of association in diagnosis of a patient can be used by a physician. But the problem over here is the amount of reliability because no process can be noted for 100% guarantee. But still association rules can be used over here to identify likely symptoms appearing together and based on this appropriate diagnosis can be suggested (Kaluza, 2016).

Fraud Detection. In Fraud Detection, isolation of specific part of a pattern, such as false claims, the system would search through many factors related to the person and try to establish a relationship or find a pattern that the current action would fit into, otherwise would be considered as a fraudulent activity (Aksar, 2016).

Census Data. In Census Data, a huge variety of statistical data is collected related to the society, which is available to researchers and general public. This can be used to form associations to forecast in planning public services such as education, health, transport, setting up factories, banks, shopping malls, marketing appropriate products (Kaluza, 2016).

In Retail, Market Basket Analysis would help with associating products with each other and determining which product goes with what. This helps an organization to stock the products appropriately and place them closer to each other. Product associations can also be used to understand what products are sold at different seasons. Market Basket Analysis can also be used for targeted promotions and marketing (Explanation of the Market Basket Model, n.d.).

Customer Relationship Management. In CRM, Market Basket Analysis can be used to derive associated rules to understand customers to provide better quality services. Customer behavior can be detected using the association rules to discover changes and generate rules according to the appropriate rule matching (Kaluza, 2016).

Telecommunications. In Telecommunications, with the ever-growing use of internet the services related to it has been on constant increase. Market Basket Analysis can be used to find the services being utilized and purchases made by customers and open this to direct marketing efforts that can be made on the customers (Explanation of the Market Basket Model, n.d.).

IT Operation Analytics. In IT Operation Analytics, a lot of data is collected with regards to the day-to-day IT operations that can be used to identify patterns and critical changes. If all the issues originate from the same source this frequent pattern mining can help to quickly cut through a number of alerts, allowing the IT operators to focus on truly critical changes (Kaluza, 2016).

Banks. In Financial, Market Basket Analysis can be used for fraud detection to detect irregular transactions and also can be used to cross sell offers and promotions present at a bank. The transactions anomaly can be detected for every purchase irrespective of any season and generated rules can be used by the staff of the financial organization to provide services appropriately to a customer (Explanation of the Market Basket Model, n.d.).

Protein Synthesis. Proteins are considered the most important component of any organism and are made up of several amino acids, using association rule mining the associations between different amino acids present in the protein, which could provide hints in understanding regarding the interactions between particular sets of these in the protein.

CRM of Financial Institutions. Customer Relationship Management (CRM) helps the financial institutions to tailor their products and services according to the customer groups, now an association rule technique would help the institutions to understand and manage their customer by providing marketing teams with required insights to understand their customers well.

# **Systems Development and General Methodology**

With the importance of developing a data mining project by researching the digital libraries and online data set repositories. The data was obtained from Kaggle.com one of the known data set repository and data mining competition sites. The Instacart data set was chosen for the demand in the Online grocery services booming.

The process begins with obtaining the data and this data was in xlsx format. Since it had many rows of data, about 30 million rows in Order\_Products\_Prior table, viewing the data in xlsx was difficult hence the data sets were imported in R. The structure of the data was analyzed. With the research question in mind the project is structured on deciding which table is most important in the data set. Plotting technique was used for exploring the data. It gave insights about the spread of the data. The research question to be answered in this question was to build a model that can predict a customer’s basket. At first the frequency of the products was to be analyzed. To get this it was decided to choose two data tables ‘Products and Order\_Products\_Prior’. These tables contain the details pertaining to the products and the order placed by a customer. The project identifies a primary key from the table and used it for the joining of the table. Using the key concept, an inner join was performed using “Product\_ID” and grouped the result using “Order\_ID”. Once this was done the number of observations in the table product was 49,688 and the table of Order\_Products\_Prior which earlier had 32 million record which came down to 3.2 million records. That indicates that the Order\_Id’s were repeated almost 10 times for each user. Hence each basket was divided based on order id and once this was done, the Association algorithm was performed. Firstly, to perform the association algorithm the data set which was grouped based on Order\_Id was converted to transaction Id’s. This transformed data is in the form of a sparse matrix, the rows will contain transaction id’s and column’s as items of the products. Once this format is obtained the item frequency was to be seen and plotted. This indicated which item was frequently brought with every other item. Then top 20 most frequently brought items were obtained and plotted using the basic histogram.

The basic histogram plotted showed the most frequently brought items in each basket. Then the association rule was applied based on support, confidence and lift value. The support is nothing but an item bought for the number of transactions. The confidence is an item purchased when another item is purchased in particular. Lift is nothing but measure of how many times more likely the product is brought.

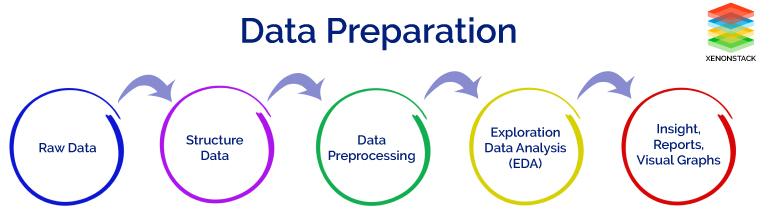
In general, the support value is set for 20% and confidence is set to 80% and the apriori algorithm is applied, which is an algorithm in R for association rule. Also, length of the rule is given at 2. That is the number of rules the items are associated with on LHS and RHS. Then looking at the summary statistics again the support is reduced and the confidence is increased that is the support is decreased to 2% and the confidence is set to 90% then the length is set to minimum 3.

By doing the above procedure the association rules were obtained, which helped understand items that are bought with the most frequently bought items. This answered the question that which items are bought along with the most frequently bought items.

# **Description and Presentation of the System**

Data Analysis is the most important aspect in the current business world but just saying this would-be understatement. Before any analysis is performed it is important to take a step back and understand what is it that one is looking for, in general need to understand what would be the purpose of conducting data analysis and in specific would be to help humans in better decision making (Inoue, 2015). The process of data analysis comprises of inspecting, cleaning, transforming and lastly modeling the data to generate a meaningful insight, which can be used to arrive at a conclusion for any given situation. Now the analysis that goes into the data would depend on the type of the data, whether one should use quantitative and qualitative analysis. When the data is of non-numerical type like textual analysis, we would perform quantitative analysis and similarly when the data is of numerical type that can measured using statistics then qualitative analysis would be performed (Kalpesh, 2013).

After having a clear understanding of the purpose of the data analysis, there is need to check for the accuracy of the data. If the data is not accurate enough it would present biased results and the conclusion that would be determined using the results would not be accurate enough. In order to perform data analysis, R language has been used, which is in an open source language for statistical computing and graphics. R has a wide variety of statistical and graphical techniques, which are highly flexible (CRAN, 1993).



*Figure 1 Data Preparation*

The figure 1 above describes the steps involved in data preparation and cleaning, which are described in detail below. The data cleaning and preparing in this project begins with Step 0: Load the data in RStudio, which loads the raw data into R using the fread function, of which the query is present in the appendix A. There are several ways to load the data into R but using the fread function is one of the best ways. The reason being R processes entire data in the temporary memory and as file size increases the space would be held up by a query the ran previously and at some point, the application would crash due to unavailability of memory. The fread function addresses this issue by freeing up the space periodically as a query executes, which would save processing time and resource. The data been loaded needs to be verified for its accuracy and once determined should begin the phase of preparing the data by exploring through it.

The exploration of the data begins at Step 1: Glimpse of the data, using few basic and very informative functions. This step also helps in structuring the data according to the needs to the desired analysis. The dim is used on data frames (used to store data tables), which would provide information such as the number of variables and observations in each data frame. This would provide with the information understanding the dataset and using the Names function the names of the variables would be determined so that those variables could be precisely called and statistical analysis be performed on them. Once an understanding is achieved about the number of observations and variables, next would be to look into small samples of the data. The Glimpse function provides the data type of the variables and also a small sample of the values of each variable for better understanding. The Summary function is used on each of the data frames to provide the descriptive statistics of the data frames that can be used to understand the range, the mean, the median of the values within them. All of this has been performed to get in detail understanding of the dataset loaded and the complete query with respect to this entire operation can be seen in the appendix A. An in detailed description of the individual data frames are presented in the data dictionary part of the appendix A as well.

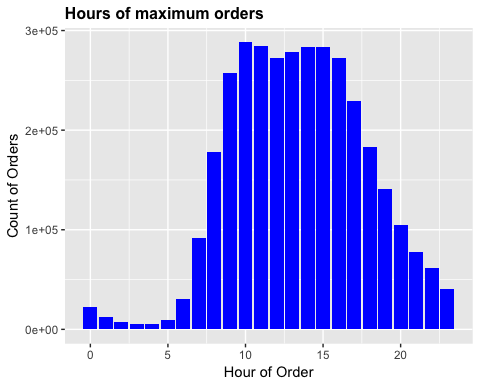
The preprocessing of the data begins by identifying if there is any necessity to clean the data. The current dataset has been accessed using a well-known reputable open data repository, Kaggle, which is also known for its clean datasets. But despite of its reputation, a protocol always needs to be followed to check the data at least for missing values. By using the sapply and is.na functions we search for ‘NA’ values in the data frame for each of the variables. Though there were zero missing values across all the data frames except ‘Orders’ data frame. The variable ‘days\_since\_prior\_order’ had 206209 missing values, this could be inferred as for all order\_number = 1 orders (the first order placed by a user) the days since prior order would obviously be zero, this is an interesting find as in general sense it would be confusing as to why there are missing values in the variable. The other action taken in data preprocessing with respect to data preparation is to convert continuous variable to categorical variable for better interpretation. The variables such as product name, aisle name, department name and eval\_set were converted to categorical variables in the data frames loaded. This can be seen in the code present in the appendix A as Step 3: Data Preparation.

The next course of action was to perform exploration data analysis and generate visual insights and graphs.

# **Research Question 1: Identifying the key factors used in analysis of the customer basket**

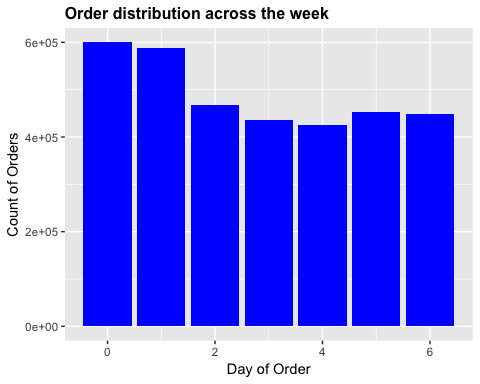
One of the research question intended to be addressed with the study is identification of the key factors involved in deciding or predicting a customer’s basket. With this aim, an exploratory analysis was conducted on the data set analyzing the different independent variables and identifying the impact of these variables on the customer’s purchase behavior.

The first analysis was conducted to study whether there is any specific day in a week or hour in a day in which there is a surge in the number of customer orders. In order to conduct this analysis, the independent variables denoting the hour and day of a particular order was considered. After aggregating the number of orders on the basis of hour and day of the week, two separate histograms were plotted to obtain a vivid picture of the scenario. The visualization in Figure 2 displays the hours for which the maximum sale is recorded. From the visualization, it could be inferred that the maximum sale occurs between 8:00 and 18:00 hours of the day.



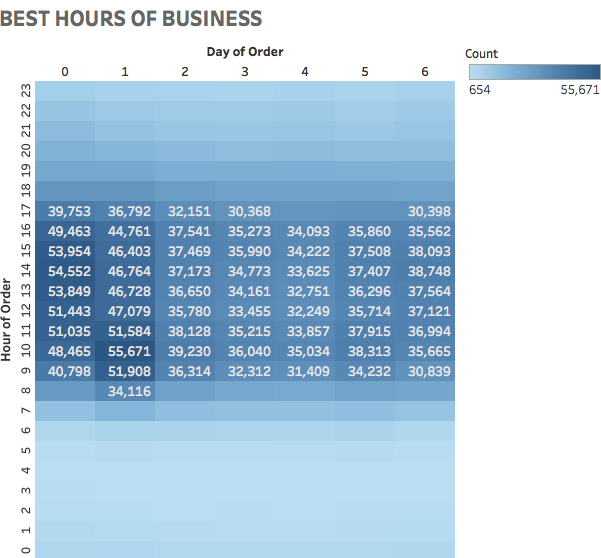
*Figure 2 Hours of maximum orders*

Figure 3 displays the order distribution across the days of the week in which 0 & 1 have the highest sales, indication of weekend days. From the visualization, it can be inferred that the maximum sales occur on the weekend and can be determined that 0 & 1 are weekend as that makes it logically.



*Figure 3 Order distribution across week*

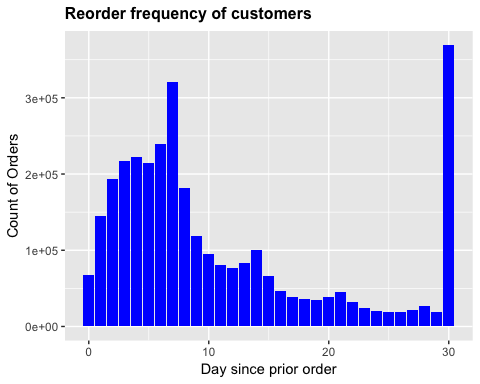
From the visualizations in Figure 2 and Figure 3, it can be clearly inferred that the day and time has an impact on the customer orders. In order to portray this concept in a better manner, an aggregated chart depicting the effect of both the day and hour is plotted in figure 4. The chart clearly shows the maximum sales are occurring on weekends when compared to the other days of the week. Also, among all the days between hours of 8:00 and 18:00 the total number of sales is high.



*Figure 4 Best Business Hours*

The next factor considered for the study of analyzing the key factors was the reorder frequency of the customers. The data set used for analysis had an indicator to inform whether a customer is placing an order for the first time or whether the customers is a returning customer placing another order. Using this data an analysis was performed to identify whether there is a particular pattern in which a returning customer places the order, more specifically the aim was to identify the reorder frequency of the customer.

The visualization in Figure 5 portrays the reorder frequency of the customers. From the plot, it can be noticed that in majority of the cases the there is a spike in the number of reorders exactly after a week span.



*Figure 5 Reorder frequency of customers*

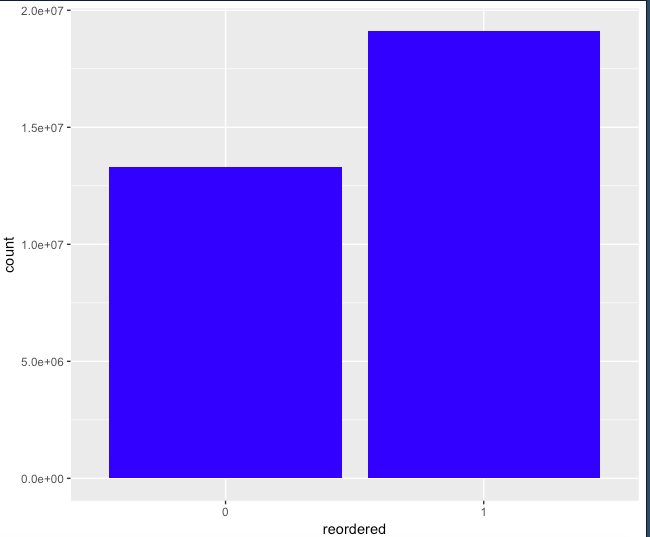
Along with considering the reorder frequency, an analysis was also conducted to understand the minimum number of orders placed by the returning customers. This study will help to draw a clear picture in understanding the customer retention rate. The more the orders placed by the returning customers, the more the customer retention.

Figure 6 visualizes the number of orders placed by returning customers. From the plot, an inference can be made that customer retention rate is good as there are about 200000 returning customers who made a minimum of 3 orders.



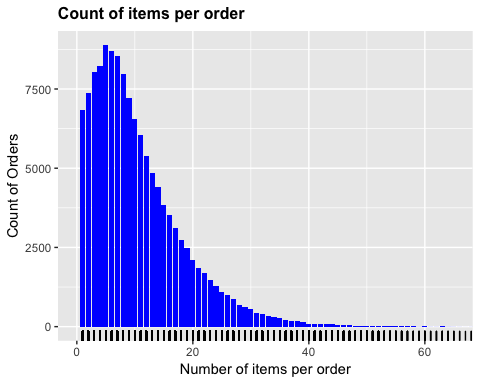
*Figure 6 Prior Orders*

The next analysis is conducted to evaluate what percentage of the orders are reorders. In the Figure 6 it was clearly depicted that the organization has a fairly good number of returning customers. So now the obvious question to be answered is regarding the percentage of orders and reorders in the overall purchases. The below visualization in Figure 7 answers this question. It depicts the proportion of orders placed by the first-time customers and orders placed by the returning customers. It can be inferred that the number of reorders are more when compared to the new orders. But the proportion is good considering the fact that there are a fairly good number of new customers placing the orders along with many of the returning customers. This shows that the firm is able to achieve the customer expectation.



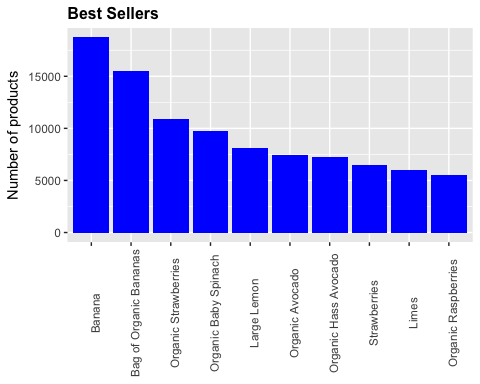
*Figure 7 Order/Reorder Proportion*

Now that the number of returning and new customers are evaluated, next an analysis is conducted to review the minimum number of items included in each of the orders placed by the customers. Identifying this number will help to gather a picture about the overall sales trend. From Figure 8 it can be noticed that in majority of the orders a minimum of five items are purchased by a customer.



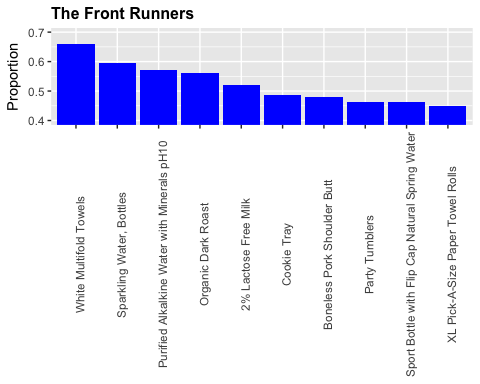
*Figure 8 Count of items per order*

As an overall picture is gathered about the generic trend of sales, the next analysis is conducted to identify the most popular items purchased by the customers. Knowing the best seller products will help the retailers in providing promotional offers in connection with the popular items. To identify the popular items the individual customer orders are analyzed in detail and the result is plotted in a visualization. The Figure 9 exhibits the best seller products among the orders and it can be noticed that the best seller item among the lot is Bananas.



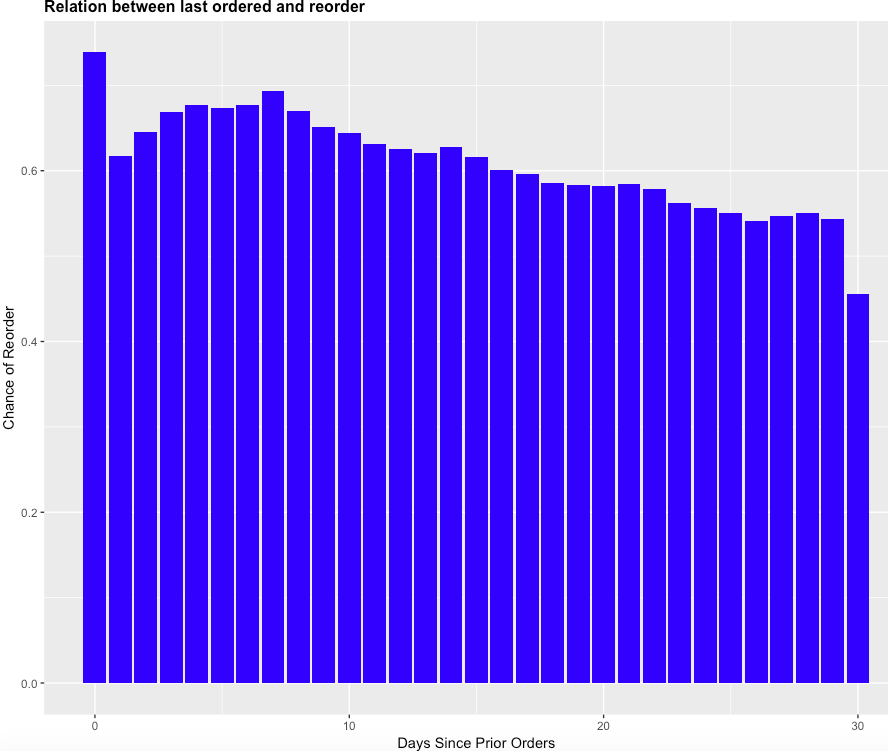
*Figure 9 Best sellers*

The dataset taken for the study had a variable which was used to denote the sequence number in which an item was added to the cart by the customer. This entry was available for all the items in the different orders. Since this was an exploratory study, an analysis was conducted to identify whether the order in which a product or item is added to the cart has any influence on the overall purchasing behavior of the customer. The study gave the result that certain items are always added first to the cart. The below visualization in Figure 10 displays the proportion of number of times the frequently bought items are added first to the cart in comparison with the total number of times these items are brought by the customer.



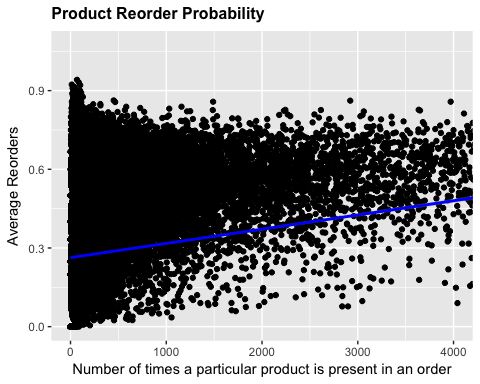
*Figure 10 The front runners*

Another key factor which was identified from the analysis was that there is an interesting relation between customer purchase behavior and the time when a customer is placing the order. By plotting a relation between time of last ordered and the next order (reorder) it was observed that when a customer places an order on the same day as the prior order, they tend to order the same product again whereas, if an order is placed after 30 days, newer products are added to the order. Figure 11 clearly depicts this relationship. As the number of days from the first order increases the chance of same item to be reordered decreases.



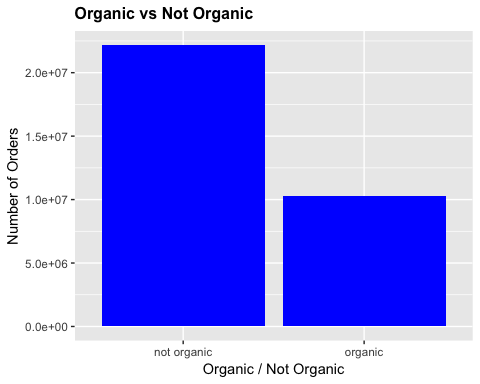
*Figure 11 Relation between last ordered and reorder*

Next analysis was conducted to analyze whether there is a chance for the popular items bought by the customers to be present in the future orders. From the analysis, it can be inferred that there is a high probability for items that are present in majority of the orders to be ordered again. Figure 12 plots this and it can be observed that the reorder chance of popular items is very high.



*Figure 12 Product Reorder Probability*

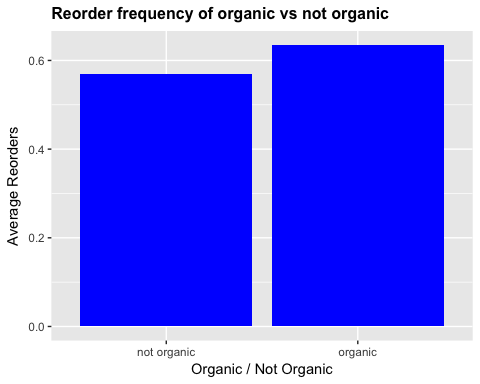
Retailers provide different pricing and promotional offers for the organic and non-organic products. Organic products are priced more than the non-organic products. But organic products are advertised as good for the overall health and well-being. The next analysis identifies whether despite the increased price tempted by the advertisements does customers buy organic products over non-organic products. The visualization in Figure 13 plots the proportion of organic and not organic products in the orders placed by the customers. From the plot, it can be clearly inferred that price is still an important factor considered by customers in decisions regarding purchases.



*Figure 13 Organic vs Not-Organic*

From the analysis done in Figure 13 it was observed that proportion of non-organic products purchased is more than the organic products. Still there is a fairly good number of organic products purchased by the customers. This resulted in the next analysis where it was analyzed whether there is a particular category of items like milk, produce etc. that are ordered from the organic product section. This analysis was done by evaluating the reorder frequency of organic products in customer orders. If the products bought from the organic section includes popular products like Milk with high reorder frequency, then the overall reorder frequency of organic products will be more.

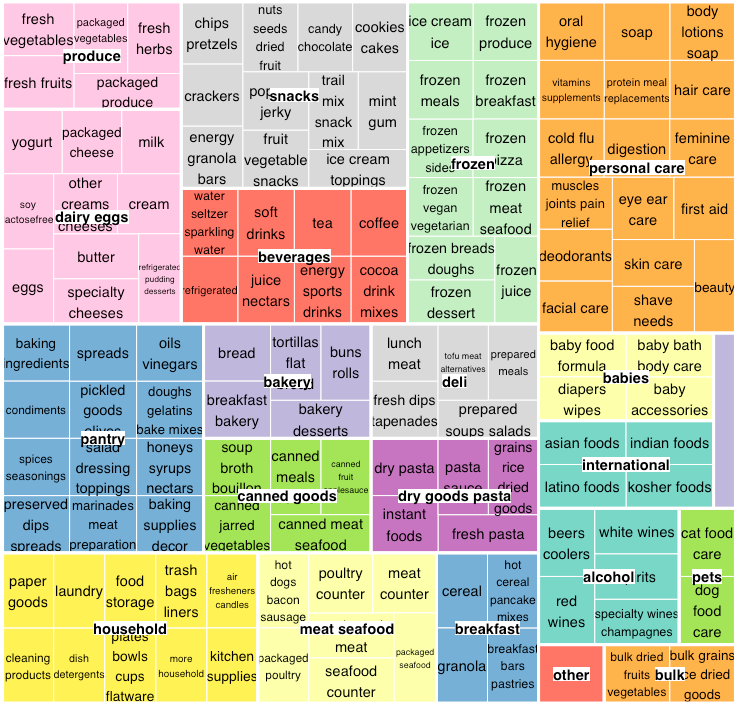
The plot in Figure 14 depicts the reorder frequency of organic and non-organic products. From the visualization, we can see that the probability of reordering organic products is higher than the chance of reordering non-organic products. From the analysis, it can also be inferred that although the overall purchase rate of organic products is less when compared to non-organic products, those items which are bought from organic aisles are items with high reorder frequency.



*Figure 14 Reorder frequency*

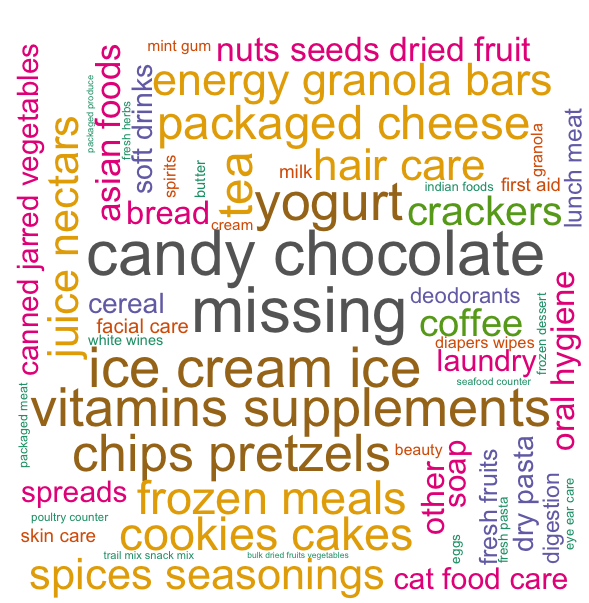
In the data set used for this study the products present in each order are organized in virtual aisles and departments. There are 134 aisles and 21 departments records in the data set. The tree map in Figure 15 maps the aisles to their respective departments and this visualization provides a better interpretation of how the aisles are distributed within the departments. The different departments are produce, snacks, frozen, personal care, dairy eggs, beverages, pantry, bakery, deli, babies, canned goods, dry goods pasta, international, household, meat seafood, breakfast, alcohol, pets, bulk and other. The aisles within each of these departments are mapped in the figure 15.

**Organization of aisles within the departments**



*Figure 15 Organization of aisles within the departments*

Now that the mapping of aisles and departments are clearly identified, next an analysis was conducted to identify the popularity of aisles and departments based on the count of unique products. An aisle or department is ranked based on the number of unique products available for purchase to the customers from each department and aisle. The aisle and department with the most number of unique products are highlighted with the largest size in the figure 16 and figure 17 respectively. From the figure 16 it can be observed that an aisle named “missing” is one among the top 5 popular aisles. From the analysis, it was inferred that this aisle includes the products which are sold at a reduced-price due to several factors.



*Figure 16 Ranking aisles based on number of unique products*



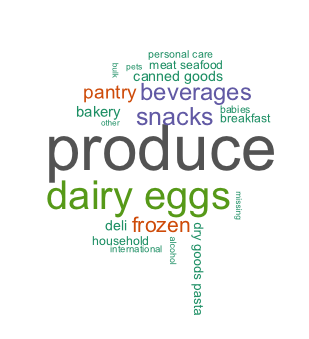
*Figure 17 Ranking departments based on number of unique products*

In the next analysis, aisles and departments are ranked based on the number of unique products which are present in majority of customer orders. The analysis helps to identify the aisles and departments which had the maximum sales with respect to number of products. The plot is obtained using basic text mining functions. The figure 18 represents the ranking of aisles based on the unique products present in majority of orders and it can be observed that the most popular aisle is “fresh vegetables” and “fresh fruits”. Packaged vegetable and fruits and yoghurt are the second most popular aisles. So, based on this inference, it can be guessed that the most popular department is produce and this guess is confirmed in the figure 19.



*Figure 18 Ranking aisles based on number of unique products in the maximum orders*

The figure 18 represents the ranking of departments based on the unique products present in majority of orders and it can be observed that the most popular department is “produce”.



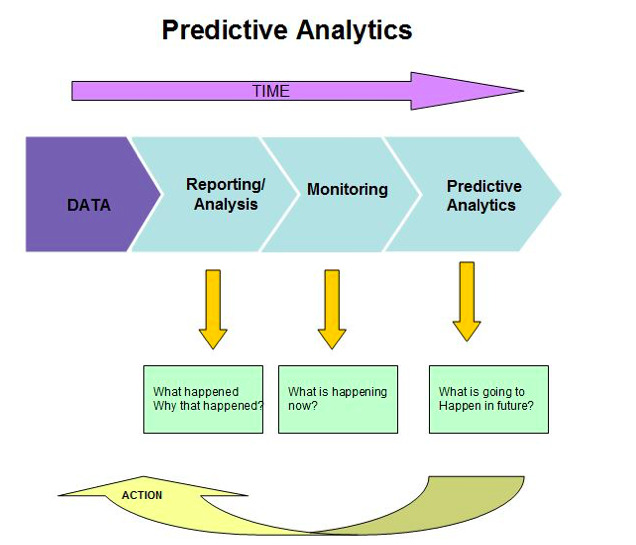
*Figure 19 Ranking departments based on number of unique products in the maximum orders*

From the exploratory analysis conducted in the anonymous data set, it was identified that the distribution of products in aisles and departments, the reorder frequency of the customers, the time and day when a customer places the order, the price of the products are among the key independent variables defining unique behavioral patterns among the customers. Next, to take this analysis a step ahead the paper discusses the role of predictive analytics in identifying customer behavior patterns and will give a snapshot of the different prediction methods that can be employed in developing recommendation systems. Along with that an in-depth analysis of the association rule mining technique is conducted on the data set for identifying strong rules among the frequently bought item sets which can further facilitate the prediction process.

# **Research Question 2: Role of Predictive Analytics in Customer Behavior Analysis and Product Recommendations for Retail Industry**

Predictive analytics is part of an advanced analytics that is used to make predictions about some unknown event, uncover new information, discover patterns and future outcomes (Exastax, 2017). Predictive analytics is the use of statistical algorithms and techniques of machine learning to analyze the historical data and determine the possible future predictions (Predictive Analytics Today, n.d.). The future predictions are performed using various data mining, predictive modeling and analytical techniques to get together the various management, information technology and business modeling processes to make all the predictions (SAS, n.d.).

The following figure 20, visually presents the activities that occur over the time during predictive analytics and following these activities appropriate actions would be taken to address the future consequences. Performing Predictive Analytics would equip an organization to become forward looking by anticipating the behaviors and outcomes using the data rather than just based on assumptions and hunches.



*Figure 20 Predictive Analytics*

In order to perform predictive analytics there are several prediction methods and this paper discusses few of them.

The main types of prediction methods we are discussing in this paper are recommender lab and black box algorithms. Along with that in the upcoming sections, the paper illustrates how association rule mining method can be used to derive strong rules among the frequent item sets and hence make appropriate predictions of an item to be paired with another item. Also, by employing the method of collaborative filtering in the data set used in the study, an item set matrix has been created which gives the absolute and relative frequencies of the already purchased products from the purchase history of each user. This matrix can be used to identify the probability rate of the item to be purchased again by the same customer.

**Recommender Lab**

R has a package of methods which can be used to develop recommendation systems called “Recommender lab”. The Recommender lab is a framework for development and testing recommendation system algorithm in R. The collaborative filtering is a recommendation algorithm in which the users rating data are taken for the items the users have rated on a liqueur scale like 1 to 5, using this we can predict a user’s rating for an item that is unknown to him/her. Also, create a TOP-N lists of recommended items (Hahsler,2016). This ‘recommenderlab’ provides an infrastructure in R to recommender systems. There are several ways to approach these recommendation systems and few of these are discussed below,

The Collaborative Filtering (CF), where the recommendation is based on how similar users liked an item, is mainly divided into two main types as:

* Memory based CF: The memory based CF means it uses the whole user database to create recommendations. This consumes a lot of time and computational space as it uses the whole user database for recommendations.
* Model based CF: This type of CF uses only the user’s data collected database. This is used for computation by clustering.

In turn, these main types of collaborative filtering methods can be further subdivided into subtypes as below:

* User Based Collaborative Filtering

Here this is basically a memory based algorithm which follows the logic that users with similar item preferences will also rate items similarly. In the present paper, one can extrapolate this logic to develop a hypothesis that users with similar item preferences will also buy similar items. Thus, items for a user can be predicted by first finding a neighborhood of similar users and then analyze the items bought by these similar users to form a prediction. Over here effort would be made to calculate the average times each product has been ordered by the similar users and only products which have been ordered more than a certain threshold value are suggested to the active user. Similar users are found by using k nearest neighbor’s method. Where the use of Pearson’s correlation coefficient to find the similar users. The user selects threshold values of the correlation and how many nearest neighbors should be analyzed i.e. the value of K is selected by the user. Example if K = 3, 3 nearest neighbors will be considered (Hahsler, 2017).

* Item Based Collaborative Filtering

Item Based collaborative filtering is a different version of the user based collaborative filtering. Here the recommendations are based on the similarity and relationships between the items in the data. The items similarity is calculated by statistical methods such as cosine similarity and Pearson’s coefficient. For each item, a list of K most similar items is stored. Now to create effective recommendations the project will calculate the weighted sum of the number of times the user has ordered the related items and the similar items with the highest weighted sums are suggested to the active user. Here again a decision would be made, which similar items to consider for recommendation and computation. For example, if k = 3 is considered, only the top 3 items with the highest reorder value are considered for computation and possible recommendation (Hahsler, 2017).

* User and Item-Based CF using 0-1 data

This model is used when there is no direct rating data available. This happens when a user will not rate the items he/she bus or likes. For this the solution is taken based on the usage behavior of the user. The 0 denotes the user doesn’t like the item and 1 denotes the user likes the item. Based on this the R matrix is computed (Hahsler ,2017).

* Recommendations for 0-1 data based on Association rules

Using the association analysis, the recommendations are built. Here the user matrix R is created in the form of the sparse matrix. This matrix is converted into transaction type to perform in the computational process. Based on the Item brought and the user transaction the support and confidence are set to find the user liking item. Then the items in the RHS are shown that it is brought by the user based on the items in the LHS (Hahsler, 2017).

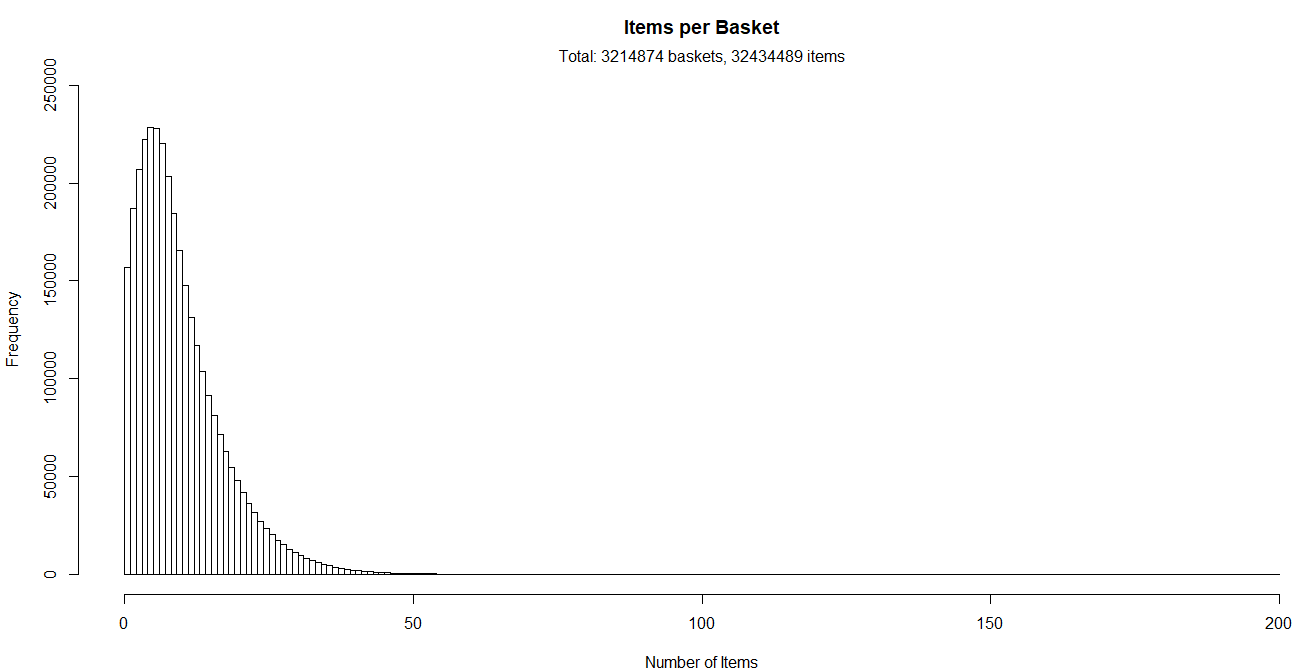
**Black Box Algorithms**

Black Box algorithms are extremely popular for their predictive power and parallel computation benefits and are widely used when one is dealing with large datasets. Algorithms such as XGBOOST are extremely popular in this domain as they can solve classification as well as regression problems both (Chen, 2016). They have an extremely high accuracy; such algorithms are termed as black box algorithms as they can only be viewed in terms of their inputs and outputs and their internal working cannot be studied in detail (Rouse, 2008). Such algorithms are popular when the main objective is prediction efficiency, in cases such as recommendation systems. These algorithms aren’t popular in cases where the need is to develop robust solutions and stable state systems based on understandable logic. For example, doctors wouldn’t trust such algorithms to predict a patient’s disease outcome as they cannot understand the internal processes that achieve such a result and risking a patient’s life based on this wouldn’t be acceptable.

**Association Rule Mining**

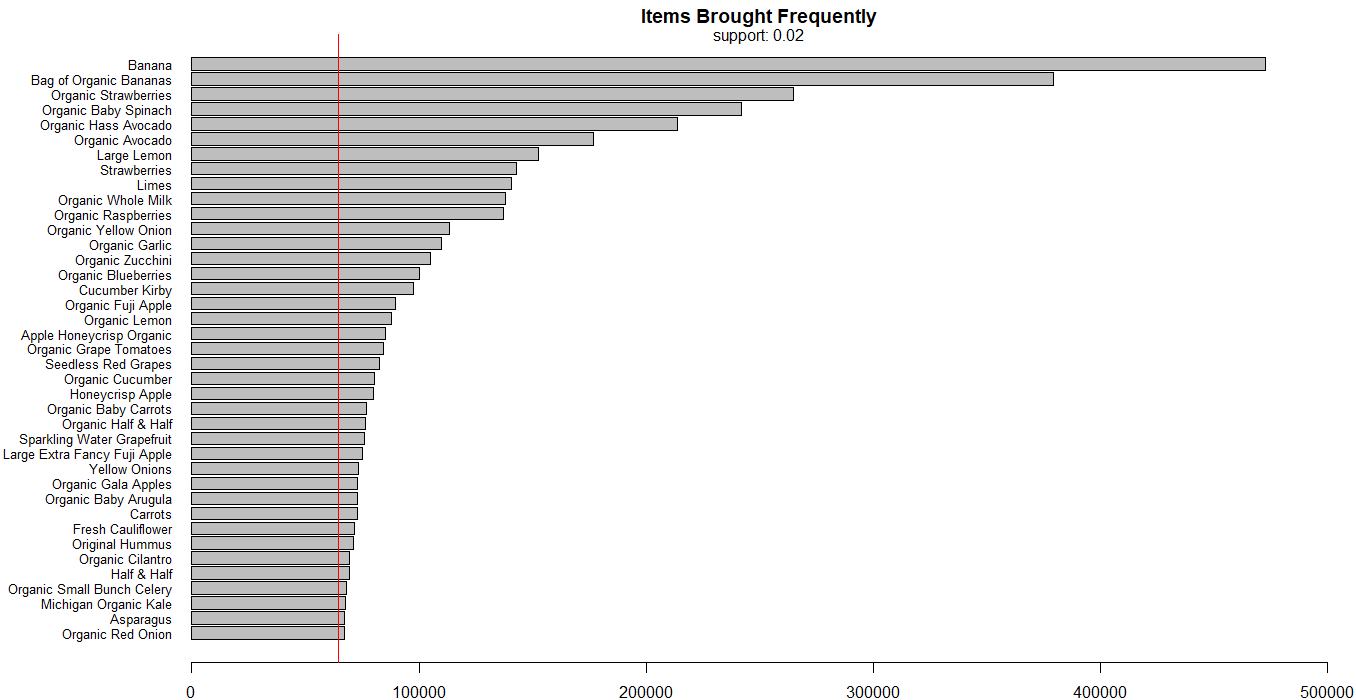
Association rule analysis is using a set of if/then statements to find frequent patterns, correlations and associations in a dataset (Technopedia, n.d.). It is a technique to understand how the items are associated to each other (Ng, 2016). A rule generated from this analysis comprises of an antecedent (if), a product or an item found in the data, and a consequent (then), a product or an item found in the data in combination with the antecedent (Rouse, n.d.). Support, lift and Confidence are the three ways to measure an association. Support describes how popular an itemset is, by measuring the transactions and determining the itemsets in which the items consist, Confidence provides the likelihood of consequent being purchased when an antecedent is purchased but there is a possibility that it might misrepresent the importance of an association and finally Lift also provides the likelihood of consequent being purchased when an antecedent is purchased while controlling the how popular consequent is. Before the association rules are applied the data is further explored.

The data set include 3.2 million baskets and 32 million items. Before proceeding with the creation of rules an analysis of the item distribution is conducted. The histogram in figure 20 depicts the distribution of number of items per basket, which has been generated using the code present in appendix A.



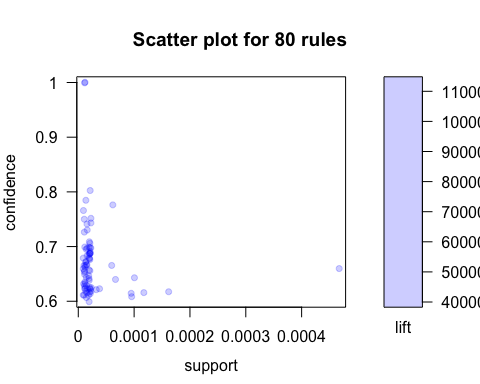
*Figure 21 Items per basket*

Next the frequent items sold are determined by analyzing the support level of the items. An item is considered to be frequently brought based on the presence of the particular item in the total percentage of transactions. The below histogram in Figure 21 shows the frequently bought items with a minimum support value of 2%.



*Figure 22 Items brought frequently*

After analyzing the item distribution in the basket, rules are generated. The first set of rules are generated with a low threshold value (0.00001) and high confidence value (0.6) so that even the less frequent items are accommodated. A scatter plot is generated for the first 80 rules generated using this setting. The code used for generating the table and the scatter plot in Figure 23 are present in the appendix A.



*Figure 23 Rule Plot 1*

Higher lift indicates a stronger relationship between the associated items, which can be inferred from the above plot in figure 23. This can be investigated further by sorting first ten rules based on lift (Table 1) and confidence (Table 2).

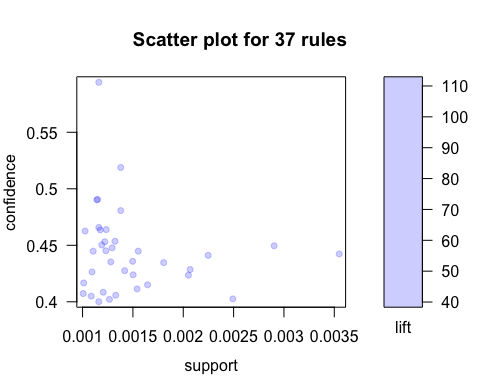
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
| **S** | **LHS** |  | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| [1] | {Moisturizing Facial Wash} | => | {Moisturizing Non-Drying Facial Wash} | 0.00001306428 | 1.0000000 | 76544.619 | 42 |
| [2] | {Moisturizing Non-Drying Facial Wash} | => | {Moisturizing Facial Wash} | 0.00001306428 | 1.0000000 | 76544.619 | 42 |
| [3] | {Raspberry Vinaigrette Salad Snax} | => | {Thousand Island Salad Snax} | 0.00002146274 | 0.8023256 | 23030.140 | 69 |
| [4] | {Extra Virgin Olive Oil Spray} | => | {All-Purpose Unbleached Flour} | 0.00001244217 | 0.7843137 | 8055.814 | 40 |
| [5] | {2nd Foods Turkey Meat} | => | {2nd Foods Chicken & Gravy} | 0.00006283294 | 0.7769231 | 4887.886 | 202 |
| [6] | {Apple Strawberry Banana Squeezable Fruit} | => | {Graduates Grabbers Fruit & Yogurt Strawberry Banana} | 0.00001026479 | 0.7674419 | 9908.550 | 33 |
| [7] | {Chocolate Love Bar} | => | {Ultra-Purified Water} | 0.00002364012 | 0.7524752 | 1624.656 | 76 |
| [8] | {Ocean Whitefish} | => | {Premium Classic Chicken Recipe Cat Food} | 0.00001026479 | 0.7500000 | 32148.740 | 33 |
| [9] | {Chocolate Love Bar} | => | {Lite Energy Drink} | 0.00002332906 | 0.7425743 | 5953.323 | 75 |
| [10] | {Ancient Grains Apricot Blended Low-Fat Greek Yogurt} | => | {Oats Ancient Grain Blend with Mixed Berry Low-Fat Greek Yogurt} | 0.00001430849 | 0.7419355 | 29088.160 | 46 |

*Table 1 Rule 1 – Sorted on Lift*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S** | **LHS** |  | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| [1] | {Moisturizing Facial Wash} | => | {Moisturizing Non-Drying Facial Wash} | 0.0000130643 | 1.00000000 | 76544.6200 | 42 |
| [2] | {Moisturizing Non-Drying Facial Wash} | => | {Moisturizing Facial Wash} | 0.0000130643 | 1.00000000 | 76544.6200 | 42 |
| [3] | {Prepared Meals Simmered Beef Entree Dog Food} | => | {Prepared Meals Beef & Chicken Medley Dog Food} | 0.0000127532 | 0.62121210 | 32211.5900 | 41 |
| [4] | {Prepared Meals Beef & Chicken Medley Dog Food} | => | {Prepared Meals Simmered Beef Entree Dog Food} | 0.0000127532 | 0.66129030 | 32211.5900 | 41 |
| [5] | {Ocean Whitefish} | => | {Premium Classic Chicken Recipe Cat Food} | 0.0000102648 | 0.75000000 | 32148.7400 | 33 |
| [6] | {Ancient Grains Apricot Blended Low-Fat Greek Yogurt} | => | {Oats Ancient Grain Blend with Mixed Berry Low-Fat Greek Yogurt} | 0.0000143085 | 0.74193550 | 29088.1600 | 46 |
| [7] | {Thousand Island Salad Snax} | => | {Raspberry Vinaigrette Salad Snax} | 0.0000214627 | 0.61607140 | 23030.1400 | 69 |
| [8] | {Raspberry Vinaigrette Salad Snax} | => | {Thousand Island Salad Snax} | 0.0000214627 | 0.80232560 | 23030.1400 | 69 |
| [9] | {Mighty Veggie Carrot Pear Pomegranate & Oats Vegetable & Fruit Smoothie} | => | {Organic Yogurt Baby Food} | 0.0000111980 | 0.66666670 | 22095.3500 | 36 |
| [10] | {Organic Baby Food Fruit Mashup Strawberry Patch 9+ Mo} | => | {Organic Baby Food Fruit Mashup Mama Bear Blueberry 7+Mo} | 0.0000118201 | 0.62295080 | 21081.1400 | 38 |

*Table 2 Rule 1 - Sorted on confidence*

The next set rules are generated with lower confidence value (0.4) and higher support value (0.001) when compared to the previous compilation. By doing so, the aim is to fetch more frequent items with lesser confidence value. The scatter plot in Figure 24 is displays the rules generated using this setting of support and confidence values.



*Figure 24 Rule plot 2*

Inspecting the Rule 2 further based on lift and confidence value to draw further insights. From the list of rules generated the top 10 rules are selected by sorting the rules using he lift and confidence value. The results of these process are listed in table 3 and table 4 respectively.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S** | **LHS** |  | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| [1] | {Non Fat Acai & Mixed Berries Yogurt} | => | {Icelandic Style Skyr Blueberry Non-fat Yogurt} | 0.001220888 | 0.4531286 | 75.648 | 3925 |
| [2] | {Non Fat Raspberry Yogurt} | => | {Icelandic Style Skyr Blueberry Non-fat Yogurt} | 0.002247055 | 0.4411064 | 73.641 | 7224 |
| [3] | {Total 2% Lowfat Greek Strained Yogurt with Peach, Total 2% with Strawberry Lowfat Greek Strained Yogurt} | => | {Total 2% Lowfat Greek Strained Yogurt With Blueberry} | 0.001161165 | 0.4658098 | 72.142 | 3733 |
| [4] | {Nonfat Icelandic Style Strawberry Yogurt} | => | {Icelandic Style Skyr Blueberry Non-fat Yogurt} | 0.001418096 | 0.4275532 | 71.378 | 4559 |
| [5] | {Total 2% Lowfat Greek Strained Yogurt With Blueberry, Total 2% with Strawberry Lowfat Greek Strained Yogurt} | => | {Total 2% Lowfat Greek Strained Yogurt with Peach} | 0.001161165 | 0.4000643 | 64.605 | 3733 |
| [6] | {Total 2% Lowfat Greek Strained Yogurt With Blueberry, Total 2% Lowfat Greek Strained Yogurt with Peach} | => | {Total 2% with Strawberry Lowfat Greek Strained Yogurt} | 0.001161165 | 0.5942375 | 63.908 | 3733 |
| [7] | {Total 2% Lowfat Greek Strained Yogurt With Blueberry} | => | {Total 2% with Strawberry Lowfat Greek Strained Yogurt} | 0.002902447 | 0.4495134 | 48.343 | 9331 |
| [8] | {Total 2% Lowfat Greek Strained Yogurt with Peach} | => | {Total 2% with Strawberry Lowfat Greek Strained Yogurt} | 0.002492788 | 0.4025517 | 43.293 | 8014 |
| [9] | {Zero Calorie Cola} | => | {Soda} | 0.001234885 | 0.4638934 | 41.669 | 3970 |
| [10] | {Sparkling Lemon Water, Sparkling Water Grapefruit} | => | {Lime Sparkling Water} | 0.001379214 | 0.4807025 | 33.202 | 4434 |

*Table 3 Rule 2 - Sorted on lift*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S** | **LHS** |  | **RHS** | **Support** | **Confidence** | **Lift** | **Count** |
| [1] | {Total 2% Lowfat Greek Strained Yogurt With Blueberry, Total 2% Lowfat Greek Strained Yogurt with Peach} | => | {Total 2% with Strawberry Lowfat Greek Strained Yogurt} | 0.001161165 | 0.5942375 | 63.907895 | 3733 |
| [2] | {Lime Sparkling Water, Sparkling Lemon Water} | => | {Sparkling Water Grapefruit} | 0.001379214 | 0.5188392 | 21.980375 | 4434 |
| [3] | {Honeycrisp Apple, Strawberries} | => | {Banana} | 0.001149967 | 0.4905786 | 3.337421 | 3697 |
| [4] | {Organic Fuji Apple, Strawberries} | => | {Banana} | 0.001141258 | 0.4901804 | 3.334712 | 3669 |
| [5] | {Sparkling Lemon Water, Sparkling Water Grapefruit} | => | {Lime Sparkling Water} | 0.001379214 | 0.4807025 | 33.201521 | 4434 |
| [6] | {Total 2% Lowfat Greek Strained Yogurt with Peach, Total 2% with Strawberry Lowfat Greek Strained Yogurt} | => | {Total 2% Lowfat Greek Strained Yogurt With Blueberry} | 0.001161165 | 0.4658098 | 72.141821 | 3733 |
| [7] | {Zero Calorie Cola} | => | {Soda} | 0.001234885 | 0.4638934 | 41.668546 | 3970 |
| [8] | {Organic Hass Avocado, Organic Navel Orange} | => | {Bag of Organic Bananas} | 0.001175474 | 0.4635672 | 3.927554 | 3779 |
| [9] | {Cucumber Kirby, Organic Fuji Apple} | => | {Banana} | 0.001024924 | 0.462586 | 3.146986 | 3295 |
| [10] | {Organic Avocado, Organic Fuji Apple} | => | {Banana} | 0.001322291 | 0.4535851 | 3.085753 | 4251 |

*Table 4 Rule 2 - Sorted on confidence*

In the exploratory analysis conducted on the data set it was identified that bananas are the best seller products (Figure 9). So now combining this information with the rules generated using the association rule mining the study was extrapolated to identify the items that are strongly tied with the item banana. The below table (Table 5) lists the products thus identified.

|  |  |
| --- | --- |
| Serial No. | Suggested Products |
| 1 | Pineapple Super Greens Kombucha |
| 2 | Fruit Squish'ems! Squeezable Fruit Pouch Apple |
| 3 | Oikos Toasted Coconut Vanilla Greek Yogurt |
| 4 | Uncured Applewood Smoked Ham |
| 5 | Ancient Grains Apricot Blended Low-Fat Greek Yogurt |
| 6 | Broccoli Crown |
| 7 | Organic Avocado |
| 8 | Organic Hass Avocado |
| 9 | Organic Navel Orange |
| 10 | Organic D'Anjou Pears |
| 11 | Organic Kiwi |
| 12 | Organic Large Extra Fancy Fuji Apple |
| 13 | Organic Raspberries |
| 14 | Strawberries |
| 15 | Large Lemon |
| 16 | Organic Avocado |
| 17 | Seedless Red Grapes |
| 18 | Cucumber Kirby |
| 19 | Organic Fuji Apple |
| 20 | Organic Raspberries |
| 21 | Organic Cucumber |
| 22 | Organic Lemon |
| 23 | Organic Strawberries |
| 24 | Blueberries |
| 25 | Organic Gala Apples |
| 26 | Original Hummus |
| 27 | Yellow Onions |
| 28 | Organic Large Extra Fancy Fuji Apple |
| 29 | Honeycrisp Apple |
| 30 | Seedless Red Grapes |
| 31 | Organic Fuji Apple |
| 32 | Apple Honeycrisp Organic |
| 34 | Organic Grape Tomatoes |
| 35 | Organic Blueberries |
| 36 | Organic Zucchini |
| 37 | Organic Garlic |
| 38 | Organic Yellow Onion |
| 39 | Bag of Organic Bananas |
| 40 | Organic Baby Spinach |

*Table 5 Product Recommendation*

This means that the products listed in the table can be recommended by the retailer to their customers when they purchase banana. The list does not imply that all the above-mentioned items are to be present in the customer’s basket, rather it is just a recommendation based on the analysis conducted in the study. Using the customer’s purchase history and the above suggested products the retailer can make better decisions regarding targeted promotional offers to returning customers. Also, based on the seasonality the company can give offers for those products. For example, there are products in the above recommendation table such as strawberries and Organic strawberries. When the customer orders strawberries and if it’s not available or is out of stock in the store then the retailer can deliver them the organic strawberry for the price of strawberry. This will make the customer happy that the premium quality product was delivered at the price of normal product and if the customer feels the organic strawberry is a better product then in the next purchase they might opt for organic strawberries which will also help the retailer in gaining more profit.

This recommendation was strongly tied only to one product in the lot, which is banana. **Next in the study, an effort was made to provide customer based product recommendation using the method of collaborative filtering.** Based on the purchase history of a customer, the relative frequency of an item purchased by every customer was computed. Based on this computation, the study can suggest the probability of a certain product to be purchased by the customer in his next purchase. The Table 6 shows the purchase history of a customer and the probability of the customer to buy the same products again in his next purchase.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **User\_Id** | **Product\_Id** | **Product\_Name** | **Abs\_Freq** | **Rel\_Freq** |
| 1 | 196 | Soda | 11 | 1.0000 |
| 1 | 10258 | Pistachios | 10 | 0.9091 |
| 1 | 10326 | Organic Fuji Apples | 1 | 0.0909 |
| 1 | 12427 | Original Beef Jerky | 10 | 0.9091 |
| 1 | 13032 | Cinnamon Toast Crunch | 4 | 0.3636 |
| 1 | 13176 | Bag of Organic Bananas | 2 | 0.1818 |
| 1 | 14084 | Organic Unsweetened Vanilla Almond Milk | 1 | 0.0909 |
| 1 | 17122 | Honeycrisp Apples | 1 | 0.0909 |
| 1 | 25133 | Organic String Cheese | 9 | 0.8182 |
| 1 | 26088 | Aged White Cheddar Popcorn | 3 | 0.2727 |
| 1 | 26405 | XL Pick-A-Size Paper Towel Rolls | 3 | 0.2727 |
| 1 | 27845 | Organic Whole Milk | 1 | 0.0909 |
| 1 | 30450 | Creamy Almond Butter | 1 | 0.0909 |
| 1 | 35951 | Organic Unsweetened Almond Milk | 1 | 0.0909 |
| 1 | 38928 | 0% Greek Strained Yogurt | 2 | 0.1818 |
| 1 | 39657 | Milk Chocolate Almonds | 2 | 0.1818 |
| 1 | 41787 | Bartlett Pears | 1 | 0.0909 |
| 1 | 46149 | Zero Calorie Cola | 4 | 0.3636 |
| 1 | 49235 | Organic Half & Half | 3 | 0.2727 |
| 2 | 23 | Organic Turkey Burgers | 1 | 0.0667 |
| 2 | 79 | Wild Albacore Tuna No Salt Added | 1 | 0.0667 |
| 2 | 1559 | Cherry Pomegranate Greek Yogurt | 6 | 0.4000 |
| 2 | 1757 | Organic Cashew Carrot Ginger Soup | 1 | 0.0667 |
| 2 | 2002 | The \""World's Best\"" Veggie Burger | 4 | 0.2667 |
| 2 | 2361 | Mint Chip | 1 | 0.0667 |
| 2 | 2573 | Garlic Pepper Ramen | 2 | 0.1333 |
| 2 | 3151 | Super Tea Power Greens | 1 | 0.0667 |
| 2 | 4071 | Organic Lemongrass Ginger Ramen | 1 | 0.0667 |
| 2 | 4957 | Total 2% Lowfat Greek Strained Yogurt With Blueberry | 1 | 0.0667 |
| 2 | 5212 | Watermelon Chunks | 1 | 0.0667 |
| 2 | 5322 | Gluten Free Dark Chocolate Chunk Chewy with a Crunch Granola Bars | 1 | 0.0667 |
| 2 | 5450 | Small Hass Avocado | 2 | 0.1333 |
| 2 | 5699 | Gluten Free Mushroom Risotto Bowl | 1 | 0.0667 |
| 2 | 5869 | Unsweetened Carob Chips | 1 | 0.0667 |
| 2 | 5907 | Apple Cinnamon Fig Bar | 1 | 0.0667 |
| 2 | 7781 | Organic Sticks Low Moisture Part Skim Mozzarella String Cheese | 3 | 0.2000 |
| 2 | 7963 | Gluten Free Whole Grain Bread | 2 | 0.1333 |
| 2 | 8138 | Traditional Hummus | 1 | 0.0667 |
| 2 | 8296 | OG Sesame Tamari Rice Cake Organic Rice Cakes | 1 | 0.0667 |
| 2 | 8479 | Original Black Box Tablewater Cracker | 1 | 0.0667 |

*Table 6 Item based probability for the users (Sample Set)*

The results of this computation can be used as an input in providing personalized recommendation to the customers when they visit the store the next time. The above table is a sample of the records of the complete table, which has relative frequency column. The value of the relative frequency determines the probability that a user would order a particular product again. For instance, the probability of user 1 ordering **pistachios** is 90.91 %, which is calculated based on the relative frequency. The table 6 over here provides an understanding of the probability of a product being ordered by the existing users again. Retailers can use this information to make better informed decisions with respect to inventory management. Higher the relative frequency higher are the chances that the product is going to reordered by the user. Similarly, when a product has a lower relative frequency it has a lower chance of being ordered again. This table has been generated using the code provided in Appendix A. As told earlier, the recommendation is just a possible option for the customer and there is no guarantee that the recommendation will turn into a purchase. As mentioned in table 6, computation is done for all the users and corresponding probabilities are generated. The below table 7, lists highest probability items that might be ordered by the returning customers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **User\_Id** | **Product\_Id** | **Product\_Name** | **Abs\_Freq** | **Rel\_Freq** |
| 1 | 25133 | Organic String Cheese | 9 | 0.8182 |
| 1 | 10258 | Pistachios | 10 | 0.9091 |
| 1 | 12427 | Original Beef Jerky | 10 | 0.9091 |
| 1 | 196 | Soda | 11 | 1.0000 |
| 3 | 39190 | Vanilla Unsweetened Almond Milk | 10 | 0.8333 |
| 5 | 11777 | Red Raspberries | 4 | 0.8000 |
| 5 | 26604 | Organic Blackberries | 4 | 0.8000 |
| 8 | 23165 | Organic Leek | 4 | 1.0000 |
| 9 | 27973 | Almond Non-Dairy Yogurt Made from Real Almonds Plain Low Fat | 4 | 1.0000 |
| 13 | 4210 | Whole Milk | 12 | 0.9231 |
| 13 | 27086 | Half & Half | 13 | 1.0000 |
| 14 | 23803 | Jalapeno Pepper | 12 | 0.8571 |
| 14 | 29509 | 80 Vodka Holiday Edition | 14 | 1.0000 |
| 16 | 5134 | Organic Thompson Seedless Raisins | 5 | 0.8333 |
| 16 | 21903 | Organic Baby Spinach | 5 | 0.8333 |
| 17 | 7350 | Natural Lime Flavor Sparkling Mineral Water | 33 | 0.8049 |
| 18 | 36216 | Lime Italian Sparkling Mineral Water | 6 | 0.8571 |
| 19 | 17008 | Shredded Sharp Cheddar Cheese | 8 | 0.8889 |
| 20 | 9387 | Granny Smith Apples | 4 | 1.0000 |
| 20 | 13575 | Apples | 4 | 1.0000 |
| 29 | 39170 | Oatmeal CrÌ¬me Pies | 18 | 0.9474 |
| 29 | 49615 | Cran-Apple Juice Drink | 18 | 0.9474 |
| 32 | 9637 | Black Forest Berry Caffeine-Free Tea | 4 | 0.8000 |
| 32 | 24852 | Banana | 4 | 0.8000 |
| 32 | 49215 | Kids Sensible Foods Broccoli Littles | 4 | 0.8000 |
| 33 | 7969 | Lime | 3 | 1.0000 |
| 33 | 8501 | Breakfast Blend Medium Keurig Brewed K-Cups Ground Coffee | 3 | 1.0000 |
| 33 | 15718 | Ultra Thin Mild Cheddar Slices | 3 | 1.0000 |
| 33 | 32441 | Fresh Goat Cheese Log | 3 | 1.0000 |
| 33 | 32912 | Baked Whole Grain Wheat Hint of Salt Crackers | 3 | 1.0000 |
| 33 | 33198 | Sparkling Natural Mineral Water | 3 | 1.0000 |
| 33 | 37131 | Tomato Yellow Cherry | 3 | 1.0000 |
| 35 | 4942 | Vanilla Almond Breeze | 8 | 0.8889 |
| 38 | 27509 | Organic Seasoned Yukon Select Potatoes Hashed Browns | 11 | 0.8462 |
| 39 | 25890 | Boneless Skinless Chicken Breasts | 6 | 0.8571 |
| 39 | 30962 | Vanilla Greek Yogurt 0% Fat | 6 | 0.8571 |
| 39 | 48679 | Organic Garnet Sweet Potato (Yam) | 6 | 0.8571 |
| 40 | 24799 | Vanilla Skyr Nonfat Yogurt | 9 | 1.0000 |
| 41 | 28985 | Michigan Organic Kale | 6 | 1.0000 |
| 55 | 4658 | Imported Mineral Water | 8 | 1.0000 |

*Table 7 Item Based Probability for Rel\_Freq > 0.8 (Sample Set)*

In reference to table 7, it can be inferred that the probability of user 1 purchasing the products, Cheese, Pistachios, Beef, Soda is more than 80%. Similarly, the relative frequency column determines the probability of a product to be present in a user’s next order. A retailer can use this information for generating targeted personalized recommendations for their customers.

# **Conclusion**

The gaining acceptance of the internet as a medium for doing business has led to a lot of momentum in online delivery models and hence the revenues for companies involved in this have significantly increased. Amongst the leading models witnessing a wide acceptance, ‘Groceries on the Go’ business segment is on the top. The Forrester Research reports ‘Online Retail Forecast’ to reach about $45 billion in 2021 (Sehgal, Kumar, & Meena, 2017). The competition is neck tight among online grocery delivery with companies such as AmazonFresh Grocery, Google Express, Instacart and Hyvee.

The exploratory data analysis performed on the dataset of Instacart gave an understanding on how a retailer should stock their inventory, by identifying the purchasing patterns of its customers. The paper also provided insights into the advantages of using a recommendation system to provide targeted promotional offers and product recommendations to customers. Unsupervised learning approach was used in explaining and implementing the concepts.

The future scope of this project would be to perform supervised learning and run regression algorithms to predict what product any particular user would reorder. Also, by automating the user item matrix and exploring the full capabilities of collaborative filtering will be an efficient way to proceed further. This has been limited in the current research due to the large size and unavailability of computational power to perform the analysis. Cloud computational services like Amazon web services and Big data technologies like Hadoop or Spark can be used in attaining this goal.

# **References**

Aksar, P. (2016, January 11). *Beyond Market Basket Analysis: Extending Association Rules*. Retrieved from Ironside: https://www.ironsidegroup.com/2016/01/11/beyond-market-basket-analysis-extending-association-rules/

Boztug, Y., & Reutterer, T. (2008). A combined approach for segment-specific market. *European Journal of Operational Research*, 294–312.

Casino Floor Optimization (June 16, 2016).

Cavique, L. (2007). A scalable algorithm for the market basket analysis. *Journal of Retailing and Consumer Services*, 400–407.

Chowdhury, N. (2013, July). Movie Business Trend Prediction using Market Basket Analysis. *International Journal of Computer Applications , 74*(9).

CRAN. (1993). *What is R?* Retrieved from https://www.r-project.org/about.html

Exastax. (2017, Jan 13). *THE IMPORTANCE OF PREDICTIVE ANALYTICS FOR YOUR BUSINESS BIG DATA*. Retrieved from Exastax: https://www.exastax.com/big-data/the-importance-of-predictive-analytics/

*Explanation of the Market Basket Model*. (n.d.). Retrieved from informationbuilders: https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL\_rstat/source/marketbasket49.htm

Gupta, G., Kumar, S., & Shaikh, F. (2015, April 16). *Effective Cross Selling Using Market Basket Analysis*. Retrieved from Analyticsvidhya: https://www.analyticsvidhya.com/blog/2014/08/effective-cross-selling-market-basket-analysis/

Inoue, T. (2015, April 3). *The 4 Important Things About Analyzing Data Part 2: Understand the Purpose of the Analysis and Who Needs the Results*. Retrieved from Treasure Data: https://blog.treasuredata.com/blog/2015/04/03/4-important-things-about-analyzing-data-understand-the-purpose-and-results/

Kalpesh. (2013, June 5). *Importance of Data Analysis in Research*. Retrieved from Dissertation India: http://www.dissertationindia.com/blog/importance-of-data-analysis-in-research

Kaluza, B. (2016, April). *Market Basket Analysis*. Retrieved from Packt: https://www.packtpub.com/books/content/market-basket-analysis

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data. *Procedia Computer Science*, 78-85.

Ng, A. (2016). *Association Rules and the Apriori Algorithm: A Tutorial*. Retrieved from KDnuggets: https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html

Predictive Analytics Today. (n.d.). *What is Predictive Analytics?* Retrieved from Predictive Analytics Today: https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/

Rouse, M. (n.d.). *association rules (in data mining)*. Retrieved from TechTarget: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

SAS. (n.d.). *Predictive Analytics*. Retrieved from SAS: https://www.sas.com/en\_us/insights/analytics/predictive-analytics.html

Technopedia. (n.d.). *Association Rule Mining*. Retrieved from Technopedia: https://www.techopedia.com/definition/30306/association-rule-mining

Vats, T. (2015, August). Market Basket Analysis by Using Apriori Algorithm in Terms of Their Effectiveness Against Various Food Product. *Indian Journal of Applied Research, 5*(8), 633-634.

Yali. (2012, April). *Market Basket Analysis: identifying products and content that go well together*. Retrieved from Snowplowanalytics: https://discourse.snowplowanalytics.com/t/market-basket-analysis-identifying-products-and-content-that-go-well-together/1132

# **Appendix A: Code**

Data Cleaning and Preparation

library(readr)  
library(dplyr)  
library(tidyr)  
library(arules)  
library(arulesViz)  
library(methods)  
library(data.table)  
library(magrittr)  
library(knitr)  
library(ggplot2)   
library(treemap)  
library(wordcloud)  
library(stringr)

# Step 0: Load the data in RStudio

Aisles <- fread('aisles.csv')  
Departments <- fread('departments.csv')  
Products <- fread("products.csv")  
Orders <- fread('orders.csv')  
Order\_Products\_Prior <- fread("order\_products\_\_prior.csv")  
Order\_Products\_Train <- fread('order\_products\_\_train.csv')

# Step 1: Glimpse of the data

**The Dim function provides the number of obervations(rows) and variables(columns) for each table.**

dim(Aisles)

## [1] 134 2

dim(Departments)

## [1] 21 2

dim(Products)

## [1] 49688 4

dim(Orders)

## [1] 3421083 7

dim(Order\_Products\_Prior)

## [1] 32434489 4

**The Name function provides the column names of each table.**

names(Aisles)

## [1] "aisle\_id" "aisle"

names(Departments)

## [1] "department\_id" "department"

names(Products)

## [1] "product\_id" "product\_name" "aisle\_id" "department\_id"

names(Orders)

## [1] "order\_id" "user\_id"   
## [3] "eval\_set" "order\_number"   
## [5] "order\_dow" "order\_hour\_of\_day"   
## [7] "days\_since\_prior\_order"

names(Order\_Products\_Prior)

## [1] "order\_id" "product\_id" "add\_to\_cart\_order"  
## [4] "reordered"

**The Glimpse function provides the structure of each table.**

glimpse(Aisles)

## Observations: 134  
## Variables: 2  
## $ aisle\_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16...  
## $ aisle <chr> "prepared soups salads", "specialty cheeses", "energy...

glimpse(Departments)

## Observations: 21  
## Variables: 2  
## $ department\_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...  
## $ department <chr> "frozen", "other", "bakery", "produce", "alcohol...

glimpse(Products)

## Observations: 49,688  
## Variables: 4  
## $ product\_id <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...  
## $ product\_name <chr> "Chocolate Sandwich Cookies", "All-Seasons Salt"...  
## $ aisle\_id <int> 61, 104, 94, 38, 5, 11, 98, 116, 120, 115, 31, 1...  
## $ department\_id <int> 19, 13, 7, 1, 13, 11, 7, 1, 16, 7, 7, 1, 11, 17,...

glimpse(Orders)

## Observations: 3,421,083  
## Variables: 7  
## $ order\_id <int> 2539329, 2398795, 473747, 2254736, 4315...  
## $ user\_id <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, ...  
## $ eval\_set <chr> "prior", "prior", "prior", "prior", "pr...  
## $ order\_number <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 1, 2...  
## $ order\_dow <int> 2, 3, 3, 4, 4, 2, 1, 1, 1, 4, 4, 2, 5, ...  
## $ order\_hour\_of\_day <int> 8, 7, 12, 7, 15, 7, 9, 14, 16, 8, 8, 11...  
## $ days\_since\_prior\_order <dbl> NA, 15, 21, 29, 28, 19, 20, 14, 0, 30, ...

glimpse(Order\_Products\_Prior)

## Observations: 32,434,489  
## Variables: 4  
## $ order\_id <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3,...  
## $ product\_id <int> 33120, 28985, 9327, 45918, 30035, 17794, 401...  
## $ add\_to\_cart\_order <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 2, 3, 4, 5, 6,...  
## $ reordered <int> 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,...

**The Summary function provides the descriptive statistics of each table.**

summary(Aisles)

## aisle\_id aisle   
## Min. : 1.00 Length:134   
## 1st Qu.: 34.25 Class :character   
## Median : 67.50 Mode :character   
## Mean : 67.50   
## 3rd Qu.:100.75   
## Max. :134.00

summary(Departments)

## department\_id department   
## Min. : 1 Length:21   
## 1st Qu.: 6 Class :character   
## Median :11 Mode :character   
## Mean :11   
## 3rd Qu.:16   
## Max. :21

summary(Products)

## product\_id product\_name aisle\_id department\_id   
## Min. : 1 Length:49688 Min. : 1.00 Min. : 1.00   
## 1st Qu.:12423 Class :character 1st Qu.: 35.00 1st Qu.: 7.00   
## Median :24844 Mode :character Median : 69.00 Median :13.00   
## Mean :24844 Mean : 67.77 Mean :11.73   
## 3rd Qu.:37266 3rd Qu.:100.00 3rd Qu.:17.00   
## Max. :49688 Max. :134.00 Max. :21.00

summary(Orders)

## order\_id user\_id eval\_set order\_number   
## Min. : 1 Min. : 1 Length:3421083 Min. : 1.00   
## 1st Qu.: 855272 1st Qu.: 51394 Class :character 1st Qu.: 5.00   
## Median :1710542 Median :102689 Mode :character Median : 11.00   
## Mean :1710542 Mean :102978 Mean : 17.15   
## 3rd Qu.:2565812 3rd Qu.:154385 3rd Qu.: 23.00   
## Max. :3421083 Max. :206209 Max. :100.00   
##   
## order\_dow order\_hour\_of\_day days\_since\_prior\_order  
## Min. :0.000 Min. : 0.00 Min. : 0.00   
## 1st Qu.:1.000 1st Qu.:10.00 1st Qu.: 4.00   
## Median :3.000 Median :13.00 Median : 7.00   
## Mean :2.776 Mean :13.45 Mean :11.11   
## 3rd Qu.:5.000 3rd Qu.:16.00 3rd Qu.:15.00   
## Max. :6.000 Max. :23.00 Max. :30.00   
## NA's :206209

summary(Order\_Products\_Prior)

## order\_id product\_id add\_to\_cart\_order reordered   
## Min. : 2 Min. : 1 Min. : 1.000 Min. :0.0000   
## 1st Qu.: 855943 1st Qu.:13530 1st Qu.: 3.000 1st Qu.:0.0000   
## Median :1711048 Median :25256 Median : 6.000 Median :1.0000   
## Mean :1710749 Mean :25576 Mean : 8.351 Mean :0.5897   
## 3rd Qu.:2565514 3rd Qu.:37935 3rd Qu.: 11.000 3rd Qu.:1.0000   
## Max. :3421083 Max. :49688 Max. :145.000 Max. :1.0000

# Step 2: Checking for any missing values in each table column wise

**Aisles Dataset**

data.frame(missing.values = sapply(Aisles, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## aisle\_id 0  
## aisle 0

**Departments Dataset**

data.frame(missing.values = sapply(Departments, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## department\_id 0  
## department 0

**Products Dataset**

data.frame(missing.values = sapply(Products, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## product\_id 0  
## product\_name 0  
## aisle\_id 0  
## department\_id 0

\*\* Orders Dataset\*\*

data.frame(missing.values = sapply(Orders, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## order\_id 0  
## user\_id 0  
## eval\_set 0  
## order\_number 0  
## order\_dow 0  
## order\_hour\_of\_day 0  
## days\_since\_prior\_order 206209

**Order\_Products\_Prior Dataset**

data.frame(missing.values = sapply(Order\_Products\_Prior, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## order\_id 0  
## product\_id 0  
## add\_to\_cart\_order 0  
## reordered 0

**Order\_Products\_Train Dataset**

data.frame(missing.values = sapply(Order\_Products\_Train, function(x) {  
 sum(length(which(is.na(x))))  
}))

## missing.values  
## order\_id 0  
## product\_id 0  
## add\_to\_cart\_order 0  
## reordered 0

# Step 3: Data Preparation

**Converting Continuous variable to Categorical for better interpretation**

Orders<-Orders %>% mutate(eval\_set=as.factor(eval\_set))  
Products<-Products %>% mutate(product\_name=as.factor(product\_name))  
Aisles <- Aisles %>% mutate(aisle=as.factor(aisle))  
Departments <- Departments %>% mutate(department=as.factor(department))

# Step 4: Performing Exploratory Analysis on the Data Sets

**The below visualization displays the hours for which the maximum sale is recorded. The visualization depicts that the maximum sale occurs between 8:00 and 18:00 hours of the day**

hours<-Orders %>% ggplot(aes(x=order\_hour\_of\_day)) + geom\_histogram(stat = "count", fill="blue")  
hours+labs(x="Hour of Order",y="Count of Orders",title = "Hours of maximum orders")+  
 theme(plot.title =element\_text(face="bold",size=12))

**The below visualization displays the order distribution across the days of the week. From the visualization we can notice that the maximum sales occurs on the weekend**

dayOfWeek<-Orders %>% ggplot(aes(x=order\_dow)) + geom\_histogram(stat = "count", fill="blue")  
dayOfWeek+labs(x="Day of Order",y="Count of Orders",title = "Order distribution across the week")+  
 theme(plot.title =element\_text(face="bold",size=12))

**The below visualization presents the reorder frequency of the customers. From the plot, we can see that in majority of the cases the reorder occurs exactly after a week span.**

daySinceOrder<-Orders %>% ggplot(aes(x=days\_since\_prior\_order)) +   
 geom\_histogram(stat = "count", fill="blue")  
daySinceOrder+labs(x="Day since prior order",y="Count of Orders",title = "Reorder frequency of customers")+  
 theme(plot.title =element\_text(face="bold",size=12))

**The below plot visualizes the number of prior orders**

priorOrderFreq<-Orders %>% filter(eval\_set == 'prior') %>% count(order\_number) %>% ggplot(aes(order\_number,n)) + geom\_line(color = 'blue', size=1)+geom\_point(size=2,color='blue')  
priorOrderFreq+labs(x="Number of prior orders",y="Count of Orders",title = "Prior Orders")+  
 theme(plot.title =element\_text(face="bold",size=12))

**The Next plot shows the number of items purchased in each of the orders placed by the customers. We could find that in majority of the orders a minimum of five items are purchased per order by a customer**

numOfItemsPerOrder<-Order\_Products\_Train %>% group\_by(order\_id) %>% summarise(n\_items = last(add\_to\_cart\_order)) %>%   
ggplot(aes(x=n\_items)) + geom\_histogram(stat='count', fill='blue') + geom\_rug() + coord\_cartesian(xlim=c(0,65))  
numOfItemsPerOrder+labs(x="Number of items per order",y="Count of Orders",title = "Count of items per order")+  
 theme(plot.title =element\_text(face="bold",size=12))

**Next visualization is quite interesting. The plot exhibits the best seller products among the orders.**

tmp <- Order\_Products\_Train %>%   
 group\_by(product\_id) %>%   
 summarise(count = n()) %>%   
 top\_n(10, wt = count) %>%   
 left\_join(select(Products,product\_id,product\_name), by='product\_id') %>%   
 arrange(desc(count))  
  
tmp %>% ggplot(aes(x=reorder(product\_name, -count), y=count)) + geom\_bar(stat='identity', fill='blue') +   
 theme(axis.text.x = element\_text(angle = 90), axis.title.x = element\_blank())+  
 labs(y="Number of products",title = "Best Sellers")+  
 theme(plot.title =element\_text(face="bold",size=12))

**What percentage of the orders are reorders?** **The below visualization answers this question.It depicts the proportion of orders and reorders placed by the customers**

tmp <- Order\_Products\_Prior %>% group\_by(reordered) %>% summarise(count = n()) %>%   
mutate(reordered = as.factor(reordered)) %>% mutate(proportion = count / sum(count))  
  
tmp %>% ggplot(aes(x=reordered, y= count, fill=reordered)) + geom\_bar(stat = "identity")+  
 labs(x="Ordered/Reordered",y="Proportion of Order/Reorder",title = "Proportion of Order/Reorder")+  
 theme(plot.title =element\_text(face="bold",size=12))

**Certain items are always added first to the cart.The below visualization displays the number of times these frequently bought items are added first to the cart**

tmp <- Order\_Products\_Train %>% group\_by(product\_id, add\_to\_cart\_order) %>% summarize(count = n()) %>%   
 mutate(pct = count / sum(count)) %>% filter(add\_to\_cart\_order==1, count>10) %>% arrange(desc(pct)) %>%   
 left\_join(Products, by='product\_id') %>% select(product\_name, pct, count) %>% ungroup() %>% top\_n(10, wt=pct)

## Adding missing grouping variables: `product\_id`

tmp %>% ggplot(aes(x=reorder(product\_name, -pct), y=pct)) + geom\_bar(stat = 'identity', fill= 'blue') +  
 theme(axis.text.x = element\_text(angle=90, hjust =0.5), axis.title.x = element\_blank()) +coord\_cartesian(ylim= c(0.4,0.7))+  
 labs(y="Proportion",title = "The Front Runners")+  
 theme(plot.title =element\_text(face="bold",size=12))

**By plotting a relation between time of last ordered and the next order (reorder) we are able to find interesting customer behavior, when a customer tends to order the same product if the order is placed on the same day whereas if an order is placed after 30 days, newer products are added to the order**

Order\_Products\_Prior %>% left\_join(Orders, by='order\_id') %>% group\_by(days\_since\_prior\_order) %>%   
 summarise(mean\_reorder = mean(reordered)) %>% ggplot(aes(x=days\_since\_prior\_order, y=mean\_reorder)) +  
 geom\_bar(stat='identity', fill = 'blue') +   
 labs(x="Days Since Prior Orders",y="Average Reorders",title = "Relation between last ordered and reorder")+ theme(plot.title =element\_text(face="bold",size=12))

**An interesting find over here is that the products that are present in most orders have a likely chance to be ordered again, which can be inferred from the below plot.**

Order\_Products\_Prior %>% group\_by(product\_id) %>% summarize(proportion\_reordered = mean(reordered), n= n()) %>% ggplot(aes(x=n, y=proportion\_reordered)) + geom\_point() +geom\_smooth(color='blue') + coord\_cartesian(xlim = c(0,4000)) + labs(x="Number of times a particular product is present in an order",y="Average Reorders",title = "Product Reorder Probability")+ theme(plot.title =element\_text(face="bold",size=12))

**The visualization plots the proportion of organic and not organic products in the orders placed by the customers**

Products <- Products %>% mutate(organic = ifelse(str\_detect(str\_to\_lower(Products$product\_name),'organic'),'organic','not organic'), organic = as.factor(organic))  
  
tmp <- Order\_Products\_Prior %>% left\_join(Products, by = 'product\_id') %>% group\_by(organic) %>%   
summarise(count = n()) %>% mutate(proportion = count / sum(count))  
  
tmp %>% ggplot(aes(x=organic, y=count)) + geom\_bar(stat='identity', fill='blue') +   
 labs(x="Organic / Not Organic",y="Number of Orders",title = "Organic vs Not Organic")+  
 theme(plot.title =element\_text(face="bold",size=12))

**The following plot depicts the reorder frequency of organic and not organic products. From the visualization we can see that the probability of reordering organic products is higher than the chance of reordering not organic products**

tmp <- Order\_Products\_Prior %>% left\_join(Products, by='product\_id') %>% group\_by(organic) %>%   
 summarise(mean\_reordered = mean(reordered))   
  
tmp %>% ggplot(aes(x=organic, y=mean\_reordered)) + geom\_bar(stat = 'identity', fill = 'blue') +  
 labs(x="Organic / Not Organic",y="Average Reorders",title = "Reorder frequency of organic vs not organic")+  
 theme(plot.title =element\_text(face="bold",size=12))

**The treemap maps the aisles to their respective departments and using this visualization we are able to better interpret how the aisles are distributed within the departments**

Products\_Distribution <-Products %>% group\_by(department\_id,aisle\_id) %>% summarise(n=n())  
Products\_Distribution<-Products\_Distribution %>% left\_join(Departments,by="department\_id")  
Products\_Distribution<-Products\_Distribution %>% left\_join(Aisles,by="aisle\_id")  
  
tmp2<-Order\_Products\_Prior %>% group\_by(product\_id) %>%   
 summarise(count=n()) %>%   
 left\_join(Products,by="product\_id") %>%   
 ungroup() %>% group\_by(department\_id,aisle\_id) %>%  
 summarise(sumcount=sum(count)) %>% left\_join(Products\_Distribution,by=c("department\_id","aisle\_id")) %>%   
 mutate(onesize=1)

**Organization of aisles within the departments**

treemap(tmp2,index=c("department","aisle"),vSize="onesize",vColor="department",palette="Set3",title="Organization of aisles within the departments",sortID="-sumcount", border.col="#FFFFFF",type="categorical", fontsize.legend = 0,bg.labels = "#FFFFFF")

**The plot displays the popular aisle and department based on the count of unique products**

wordcloud(words = Products\_Distribution$aisle, freq = Products\_Distribution$n, min.freq = 12, max.words = 134, random.order = FALSE, rot.per = 0.35, colors = brewer.pal(8,"Dark2"))

wordcount <- Products\_Distribution %>% group\_by(department\_id) %>% summarise(sumcount = sum(n)) %>% left\_join(Departments, by="department\_id")  
  
wordcloud(words = wordcount$department, freq = wordcount$sumcount, min.freq = 1, max.words = 21,   
 random.order = FALSE, rot.per = 0.85, colors = brewer.pal(8,"Dark2"))

**The visualization depicts the most popular aisle and department based on the products present in maximum orders. The plot is obtained using basic text mining functions**

wordFreq<-tmp2 %>% group\_by(department\_id) %>% summarise(sumcount = sum(sumcount)) %>% left\_join(Departments,by="department\_id")  
  
wordcloud(words = wordFreq$department, freq = wordFreq$sumcount, min.freq = 1, max.words = 21,   
 random.order = FALSE, rot.per = 0.85, colors = brewer.pal(8,"Dark2"))

aisleFreq<-tmp2 %>%group\_by(aisle\_id) %>% summarise(sumcount = (sumcount)) %>% left\_join(Aisles,by="aisle\_id")  
  
wordcloud(words = aisleFreq$aisle, freq = aisleFreq$sumcount, min.freq = 1, max.words = 135,   
 random.order = FALSE, rot.per = 0.85, colors = brewer.pal(8,"Dark2"))

# Step 5: Association Rule mining

**There are 3.2 million baskets and 32 million items. The histogram depicts the distribution of number of items per basket**  # Step 0: Load the data in RStudio

Aisles <- fread('aisles.csv')  
Departments <- fread('departments.csv')  
Products <- fread("products.csv")  
Orders <- fread('orders.csv')  
Order\_Products\_Prior <- fread("order\_products\_\_prior.csv")  
Order\_Products\_Train <- fread('order\_products\_\_train.csv')

baskets <- Order\_Products\_Prior %>%   
 inner\_join(Products, by="product\_id") %>%   
 group\_by(order\_id) %>%  
 summarise(basket = as.vector(list(product\_name)))  
  
transactions <- as(baskets$basket, "transactions")  
  
hist(size(transactions), breaks = 0:300, xlim=c(0,300),ylim=c(0,250000), xlab="Number of Items",   
 ylab="Frequency", main="Items per Basket" )   
mtext(paste("Total:", length(transactions), "baskets,", sum(size(transactions)), "items"))

**The frequent items sold are determined. An item is considered to be frequently brought when it is present in at least 2% of all the baskets. The below histogram shows the frequently bought items, which has a support value of 2%**

itemfrequencies <- itemFrequency(transactions,type="a")  
support<-0.02  
frequentitems <- sort(itemfrequencies, decreasing = F)  
frequentitems<- frequentitems[frequentitems>support\*length(transactions)]  
  
par(mar=c(2,10,2,2)); options(scipen=5)  
barplot(frequentitems, horiz=T, las=1, main="Items Brought Frequently", cex.names=.8, xlim=c(0,500000))  
mtext(paste("support:",support), padj = .8)  
abline(v=support\*length(transactions), col="blue")

**Rules are generated with a low threshold value and high confidence so that even the less frequent items are accommodated**

Rule1<- apriori(transactions, parameter = list(supp = 0.00001, conf = 0.6, maxlen=3), control = list(verbose = FALSE))  
summary(quality(Rule1))

plot(Rule1, col=rgb(0,0,1, 0.2))

**Higher lift indicates a stronger relationship between the associated items, which can be inferred from the above plot. This can be investigated furtherby sorting first ten rules based on lift and confidence**

inspect(sort(Rule1, by="lift")[1:10])

inspect(sort(Rule1, by="confidence")[1:10])

**In Rule 2 the confidence value is decreased and support is increased to fetch more frequent items with lesser confidence value**

Rule2<- apriori(transactions, parameter = list(supp = 0.001, conf = 0.4, maxlen=3), control = list(verbose = FALSE))

summary(quality(Rule2))

plot(Rule2, col=rgb(0,0,1, 0.2))

**Inspecting the Rule 2 further based on lift and confidence value to draw further insights**

inspect(sort(Rule2, by="lift")[1:10])

inspect(sort(Rule2, by="confidence")[1:10])

**Using the purchase history of every user, a user-item matrix is created. The matrix provides information about the relative frequency of each product which the user has purchased in his previous purchases. The relative frequency can be used to define the probability value for the same product to be repurchased by the customer in his next purchase**

**User Item matrix**

ordered\_products\_per\_user<- rbind(Order\_Products\_Prior,Order\_Products\_Train ) %>%   
 inner\_join(Products,by="product\_id") %>%   
 inner\_join(Orders,by="order\_id") %>% group\_by(user\_id) %>%   
 summarise(ordered\_items=as.vector(list(product\_id)),n\_orders=length(unique(order\_id)))  
  
ordered\_products\_per\_user$item\_frequencies<-Map(function(x){as.data.frame(table(x,dnn="product\_id"),responseName="abs\_freq",stringsAsFactors=FALSE)},ordered\_products\_per\_user$ordered\_items)  
  
ordered\_products\_per\_user$item\_frequencies<-Map(function(df,n){df$rel\_freq<-df$abs\_freq/n;df},  
 ordered\_products\_per\_user$item\_frequencies,  
 ordered\_products\_per\_user$n\_orders)  
  
user\_item\_df<-ordered\_products\_per\_user %>% select(user\_id,item\_frequencies) %>% unnest(item\_frequencies)  
  
write\_csv(user\_item\_df, "user\_item\_frequencies.csv")

# **Appendix B: Data Dictionary**

The present dataset is an open source anonymized dataset, which has been taken out from a relational database and provided by Instacart (organization).

The source of the dataset is from a Kaggle competition,

<https://www.kaggle.com/c/instacart-market-basket-analysis>

1. **Aisle**

An Aisle is a virtual representation of the physical aisles present in a store. These are where all the products are grouped based on their similarity, just as how they are placed in physical stores. There are about 134 rows (observations) and 2 variables (columns) in this table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Variable Description** | **Primary Key** | **Foreign Key** |
| (for example, Numerical, Text, Categorical etc.) |
| Aisle\_id | Numerical | Identification number of the aisle | Yes | No |
| Aisle | Text | Name of the Aisle | No | No |

1. **Department**

A department generally consists a group of similar aisles placed under it. This grouping of aisles is done based on the similarity of the aisles and products within those aisles. There are 21 rows (observations) and 2 variables (columns) in the present table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Variable Description** | **Primary Key** | **Foreign Key** |
| (for example, Numerical, Text, Categorical etc.) |
| Department\_id | Numerical | Identification number of the department | Yes | No |
| Department | Categorical | Name of the department. | No | Yes |

1. **Order\_products\_prior (Orders placed previously by the customer)**

This table tells us which products were bought in an order by a customer which were ordered previously by them. The ‘reordered’ variable indicates whether a customer has previously ordered or not based on the flag values and for orders, which have no previous values there would be no values. There are 32 million rows (32,434,489 observations) and 4 variables in the present table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Variable Description** | **Primary Key** | **Foreign Key** |
| (for example, Numerical, Text, Categorical etc.) |
| Order\_id | Numerical | Identification number of the order | No | Yes |
| Product\_id | Numerical | Identification number of the product | No | Yes |
| Add\_to\_cart\_order | Numerical | How many times the product was added to the cart? | No | No |
| Reordered | Numerical | Flag, which indicates whether a customer who has a previous order has the particular product or not. | No | No |

1. **Orders**

This table contains the information of the orders that were placed and to which customer they belong to. This table also provides details like which day of the week the product was ordered, which hour of the day the product was ordered, how many has it been since the previous order. There are 3.4 million rows (observations)

7 variables (columns) in the present table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Variable Description** | **Primary Key** | **Foreign Key** |
| (for example, Numerical, Text, Categorical etc.) |
| Order\_id | Numerical | Identification number of the order | Yes | No |
| User\_id | Numerical | Identification number of the user |  |  |
| Eval\_set | Categorical | To which dataset this order belongs to? | No | No |
| Order\_number | Numerical | Number of the order | No | No |
| Order\_dow | Numerical | Which day of the week was the product ordered? | No | No |
| Order\_hour | Numerical | Which hour of the day the order was placed? | No | No |
| \_of\_day |
| Days\_since\_ | Numerical | Number of days since the previous order. | No | No |
| prior\_order |

1. **Products**

This table contains the details of the various products and their descriptions that are being sold by the grocery store and also shows the relationship to the aisle and department, which the products would belong to. There are 49,688 rows (observations) and 4 variables (columns) present in this table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Variable Description** | **Primary Key** | **Foreign Key** |
| (for example, Numerical, Text, Categorical etc.) |
| Product\_Id | Numerical | Identification number of the product | Yes | No |
| Product\_Name | Text | Name of the product | No | No |
| Aisle\_id | Numerical | Identification number of the aisle where the product is placed | No | Yes |
| Department\_id | Numerical | Identification number of the department where the product is placed | No | Yes |

**Train dataset**

1. **Orders\_product\_train**

This table contains a sample of the actual dataset upon, which analysis could be done. The model developed would be trained on the train dataset to provide insights. Generally, for larger datasets, it is a common practice to take a good sample of the data upon which the model could be trained and test. There are 1.3 million rows (observations) and 4 variables (columns) in this table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Variable Type** | **Variable Description** | **Primary Key** | **Foreign Key** |
| (for example, Numerical, Text, Categorical etc.) |
| Order\_id | Numerical | Identification number of the order | No | Yes |
| Product\_id | Numerical | Identification number of the product | No | Yes |
| Add\_to\_cart\_order | Numerical | How many times the product was added to the cart? | No | No |
| Reordered | Numerical | Flag, which indicates whether a customer who has a previous order has the particular product or not. | No | No |