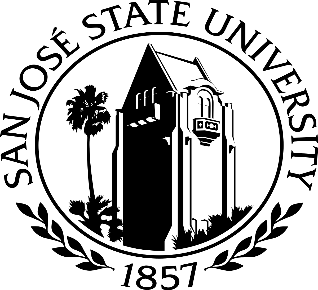
Project Report

MARKET BASKET ANALYSIS

DATA-294 - DATA ANALYTICS SEM

San Jose State University  
Master of Science in Data Analytics



Aarathi Srinivas Nadathur – 013704405

**Table of Contents**

[**Abstract** 3](#_Toc8589766)

[**Introduction** 4](#_Toc8589767)

[Problem Statement 5](#_Toc8589768)

[Literature Review 5](#_Toc8589769)

[Background and significance 6](#_Toc8589770)

[Association rule mining 6](#_Toc8589771)

[Affinity analysis 6](#_Toc8589772)

[Dataset 7](#_Toc8589773)

[Algorithms used 7](#_Toc8589774)

[K-means clustering 7](#_Toc8589775)

[PCA 8](#_Toc8589776)

[Implementation 8](#_Toc8589777)

[Basic Analysis 8](#_Toc8589778)

[Data wrangling 8](#_Toc8589779)

[PCA analysis 11](#_Toc8589780)

[K-means clustering 12](#_Toc8589781)

[Conclusion 13](#_Toc8589782)

[Future works 13](#_Toc8589783)

[**Bibliography** 14](#_Toc8589784)

**Abstract**

Market Basket Analysis (MBA) helps to find out what products clients tend to buy together. It makes it easy for understanding the buying conduct of customers by data mining. A solution to this in distributed data mining is that the massive dataset can be collected and stored in a warehouse by reducing the dimension. In this paper I have used K means clustering and Principal Component Analysis (PCA) algorithm to reduce the dimensions and with clustering, finding patterns which will be easier to find possible explanations for the identified clusters.

Keywords: PCA, K-means.

**Introduction**

Market Basket Analysis is a modeling technique which gives an indication of what a client would have bought if the idea originated. Market Basket Analysis involves transactional data listing of all goods bought by a customer in an individual purchase. The technique determines the plans for the purchase of other products.

If{A} then{B}. The If part of the rule ({A}) is known as the antecedent and the THEN part of the rule is known as the consequent ({B}).

The antecedent is the situation and the result are the outcome. The Association Regulations have three steps which convey trust in the rule: support, confidence and lift.

Support: the proportion of purchases containing all goods in a package.

Confidence: the possibility that a transaction containing the products on the left side of the law will also contain the object on the right.

Lift: the probability of all items that occur together (support) in a rule, as if there was no connection between them, by the product of the probabilities of the items on the left and right.

The product assistance demonstrates the popularity of the item in the order set. "The item is more widely distributed, more supportive. "These steps can assist identify the reason behind the traffic. Confidence and increased profitability can be used for item positioning approach. In the presence of Item B, lifting values above 1.0 are indicative of the transactions containing item B, which tend to contain item A more frequently than transactions not containing item B. Lift provides information on the probability change of item A.

[1]

# Problem Statement

The advent of electronic commerce expanded the usage and application of transactional statistics. So, using the records of client orders of the past, we can determine what will he/ she be purchasing in the future based on some pattern of purchase. Data mining tools have turn out to be the surest weapon for studying a big amount of information and breakthrough in making accurate selections. The goal of this paper is to analyze many statistics thereby exploiting the customer behavior and make a suitable choice. A sample is generated from the clients preceding purchases and analyzed.

# Literature Review

In retailing, most purchases are bought on impulse. MBA gives clues as to what a customer might have bought if the idea had occurred to them. Although MBA conjures up pictures of shopping carts and supermarket shoppers, it is important to realize that there are many other areas in which it can be applied. These include: Credit card analysis, Analysis of telephone calling patterns, etc.

[2]In this paper the author describes about MBA using association mining techniques. The author has worked on dynamic data and performed periodic mining. The analysis is performed using ARM algorithm, found the outliers and examined the customer behavior that will assist in increasing the sales.

[3] In this paper the author has used apriori algorithm and association mining techniques to perform market basket analysis. The approach examines the buying habits of the customers and identifies the associations among the items purchased by the customers in their baskets. The author has improved the efficiency of the algorithm by grouping items into higher conceptual groups and Reducing the number of scans of the entire database.

[4] In this paper the author describes about how different predictive analysis techniques can be used for market basket analysis. Using supervised learning, finding patterns within the data which can then be used to predict a label or value, given some set of parameters. Unsupervised learning were historical data is not used to learn the data. The author provides an overview of what is market basket analysis and how the to make it efficient by using different predictive models.

[5]The paper discusses the challenges in implementing market basket analysis to leverage near-instant results. This in turn facilitates “train- of- thought” or interactive analysis, enabling retailers to drill down into customer buying patterns over a period to precisely understand and target specific combinations of products, brands, categories and even time of a day. The author has used Apriori algorithm using XLMiner and created different Rules to suggest Confidence for the products.

[6] The paper presents a survey about the existing data mining algorithm for market basket analysis. The author has discussed about how apriori algorithm can be used to for data mining and how the efficiency of the algorithm can be improved by combining fuzzy logic technique with apriori algorithm which will improve its time complexity, accuracy and also provide better selection of association rules for market basket analysis.

# Background and significance

## Association rule mining

Association rules are sincere If /then statements that help to find links between reputedly autonomous relational databases or unique databases. Association rule mining is a way geared closer to watching commonly going on dispositions, correlations, or connections from data sets discovered by a variety of databases which includes relational databases, transactional databases, and different repository kinds.

**Example:**

Bread -> Jam-The above rule shows that there is a strong relationship between bread and jam. It shows that many customers buy bread and jam together. These rules can be helpful for retailers to understand buying nature of customers. One of the most common methods of data mining is to discover regular items from a data set and obtain association laws.

## Affinity analysis

Affinity analysis is a method which helps in revealing the connection or relation between the different process or transactions performed by people or computer. In general, this can be used for any process in which agents can be identified uniquely and information can be stored about their activities. Affinity analysis can be used in market basket analysis for the retailers to get an idea about the customer’s purchasing pattern.

# Dataset

The data set has been taken from Kaggle website. The dataset belongs to Instacart. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, we provide between 4 and 100 of their orders, with the sequence of products purchased in each order.

Customers are recognized from user\_id. every row of the orders table represents an order made by using a user. Order is recognized by order\_id; each order of a consumer is characterized by using an order\_number which specifies while it has been made with respect to others of the equal user; every order includes a set of merchandise each characterized by using an add\_to\_cart\_order characteristic representing the series wherein they were introduced to the cart in that order; for every user we may additionally have n-1 prior orders and 1 teach order OR n-1 earlier orders and 1 check order wherein we ought to state what products have been reordered.

# Algorithms used

Finding a possible customer segmentation enabling to classify customers according to their different purchases.

## K-means clustering

The objective of K-means is to split results into K groups. The group with the nearest sum is equivalent to each sample. The method of clustering is as follows: (1) Split the items into the initial K cluster by combining each sample with the closest median, (2) Add an item to the cluster and (3) Assign the centroid for the cluster. The K-means cluster algorithm is used to identify unclear objects. This can be used to check what types of entities exist or to find unknown entities in complex data sets. When the algorithm is present, new information can be assigned to the correct unit. [7]

## PCA

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce variables to a small set with no data loss. PCA decreases the attribute range to the lower number and is a non-dependent procedure. The aim is to reduce sizes, and the sizes are not to be interpreted. PCA decreases room of the attributes to a few numbers of factors from a greater number of variables. [8]

# Implementation

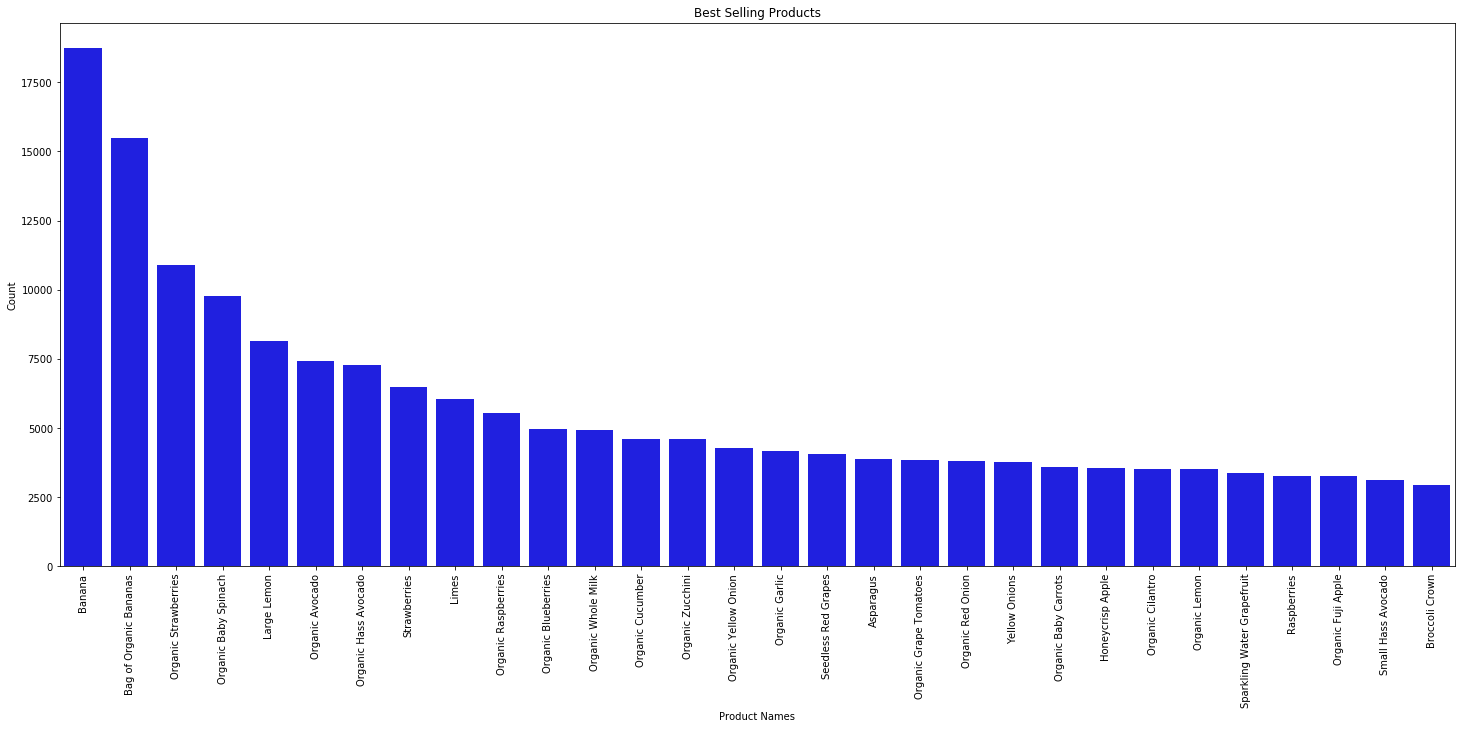
## Basic Analysis

### Data wrangling

The insta-cart data set has 5 different data sets (tables). So, first the data was cleaned by checking for null values and replacing them. Then the first 30000 records were selected to analyze. The necessary tables were merged, and the first 30000 records were selected.

The following graphs were plotted to get an idea of the shopping pattern of the customers:

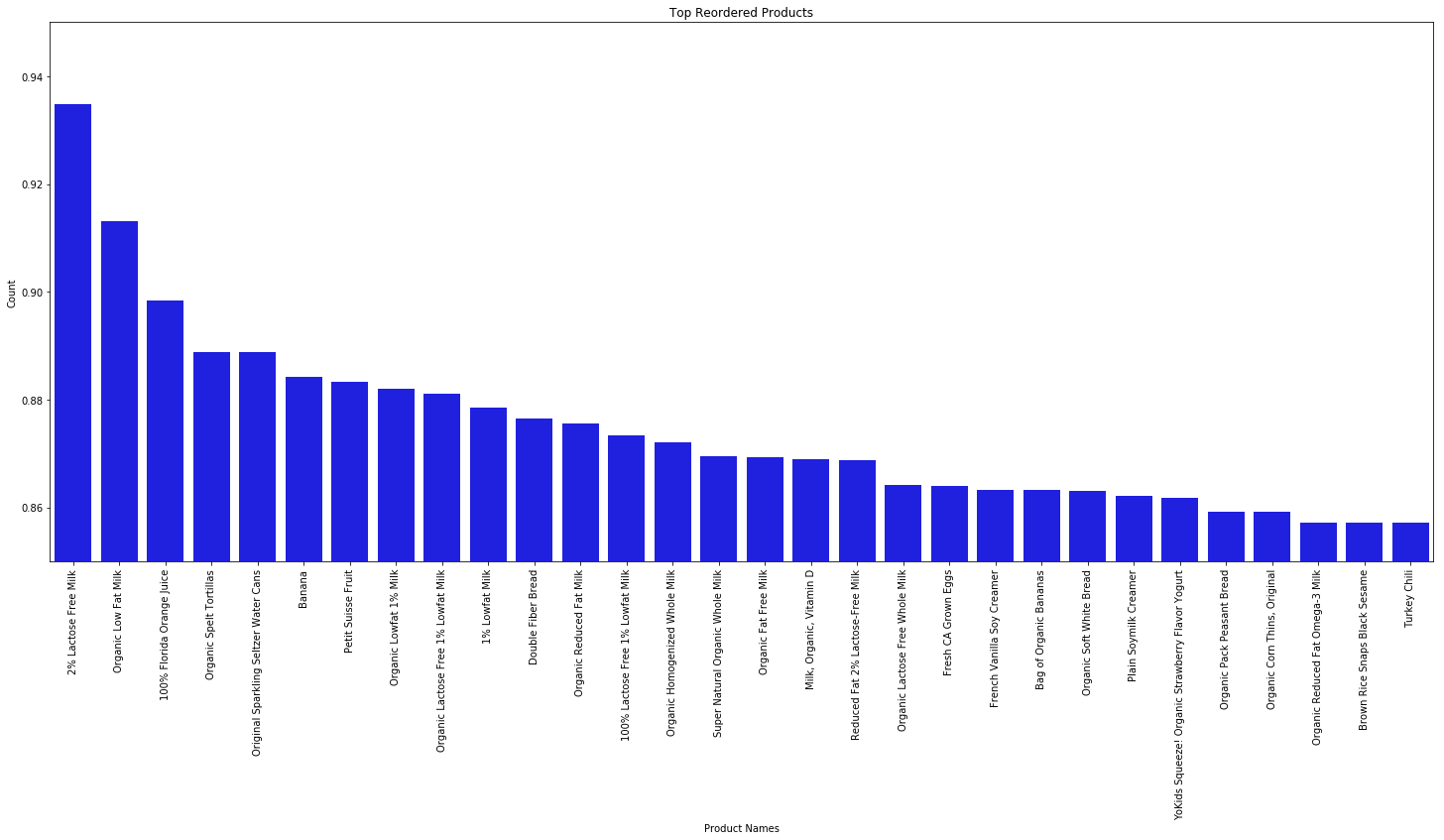
**Best-selling item**



*Figure 1*

The best-selling item is banana and the least selling is broccoli crown. So, the shopkeepers can stock up the products accordingly.

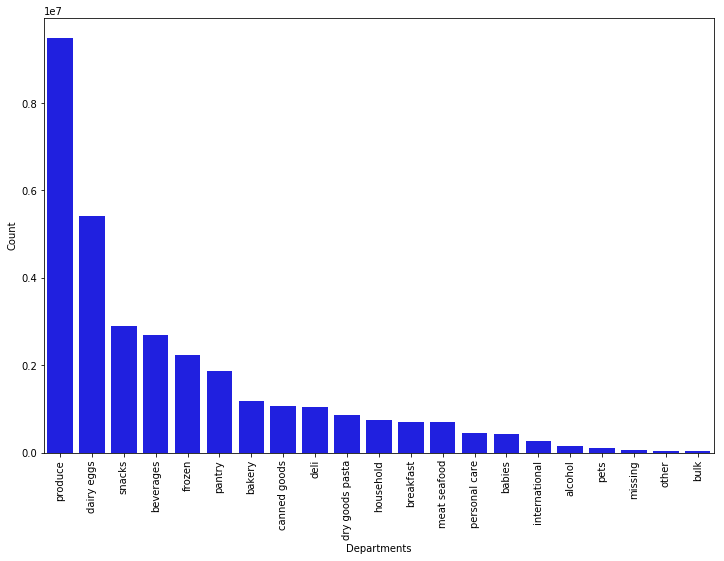
**Top reordered items**



*Figure 2*

The top reordered items will give the shopkeeper an idea about the demand of each product. So, from the graph we can say that 2% Lactose free milk is most reordered. Top reordered items are most important as the customers who buy these are regular customers. So, the shopkeeper must make sure that the products which are reordered the most much be always on stock.

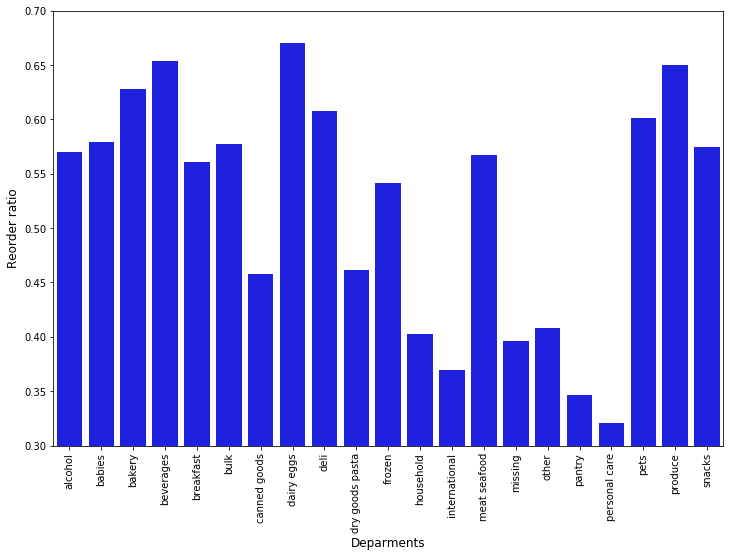
**The count of products in each department**



*Figure 3*

From the above graph we can say that most items that are bought are from produce department.

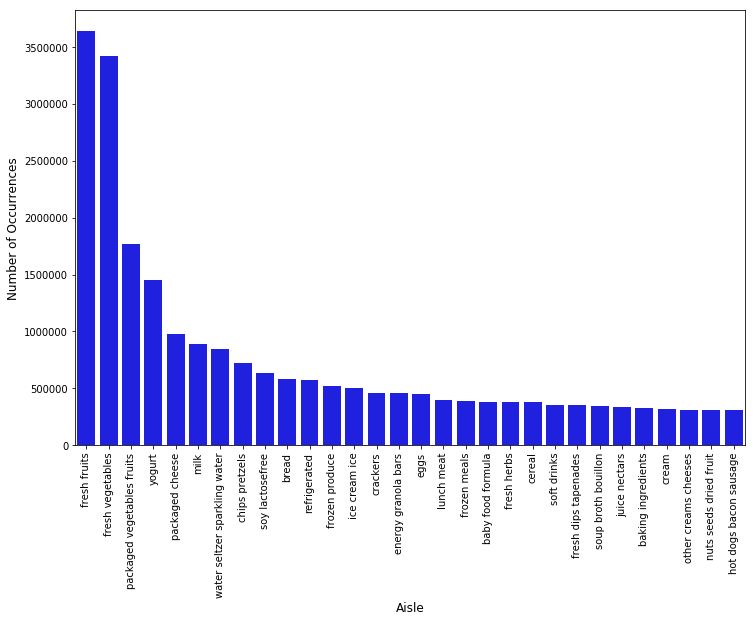
**Which department products are most reordered?**

****

*Figure 4*

The highest reorder ratio is for dairy eggs.

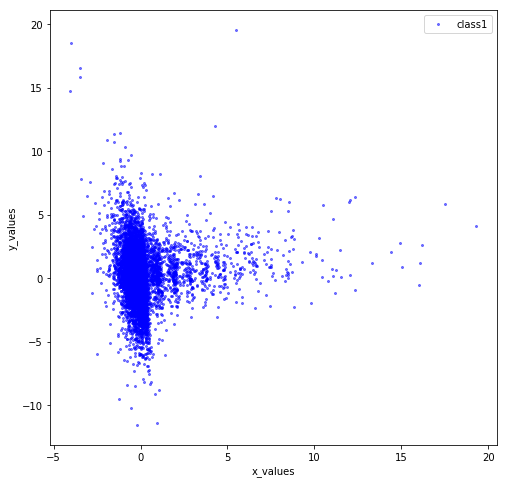
**Most Popular aisles**



*Figure 5*

From the graph we can say that fresh fruits are most popular followed by fresh vegetables. From graph 4 and 5, we can come to the point that most people buy fresh produce.

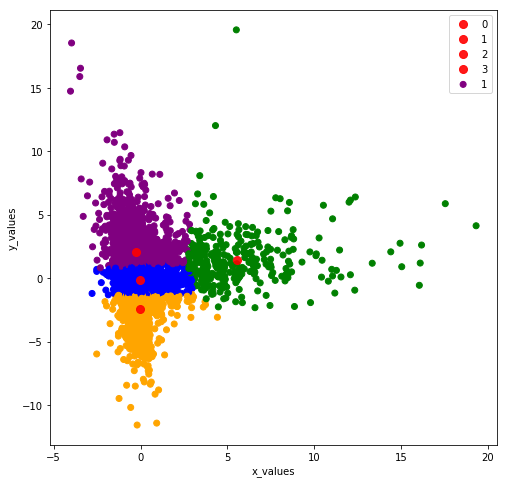
## **PCA analysis**



*Figure 6*

Using PCA for dimensionality reduction involves zeroing out one or more of the smallest principal components, resulting in a lower-dimensional projection of the data that preserves the maximal data variance. For analysis we are mainly considering aisles column. Using PCA, it will reduce the number of features from the number of aisles to 6.

## K-means clustering



*Figure 7*

After the principal component points have been achieved, pick 2 PC scores and pursue the clustering method of the k means where the information is grouped into four different groups depending on the range from the data point from the cluster center. Clustering provides a pattern common in customers and similar customers are clustered together.

After Clustering the different customers and substitute single user\_id with the cluster to which they are assumed to belong, I found the most frequent bought items in each cluster, and the results are as follows:

Cluster 0: Fresh vegetables

Cluster 1: Fresh fruits

Cluster 2: Fresh fruits

Cluster 3: Baby food formula

# Conclusion

Market Basket Analysis is a standout amongst the most well-known and valuable kinds of information investigation for advertising and retailing. Recognizing what items individuals buy as a gathering can be extremely useful to a retailer or to some other organization. Advertisers could utilize the market basket analysis results to figure out what new items to offer their earlier clients. Data mining methods have turned out to be surest weapon for dissecting tremendous measure of information and leap forward in settling on right choices. In this paper I have used Principal Component Analysis for dimension reduction so that the data can be easily analyzed. Then using K means clustering, clustered the similar data. The main objective was to find a pattern in the data so that it will be easy to know the customer buying habits and it will also be useful for the seller/ shopkeeper to know the demand of products and also help them increase the profits.

# Future works

Other methods for data mining can be tested to create association laws to render them functionally more efficient, which can be implemented in the existing recommendation system Better and efficient rule mining methods can be used to improve the recommendation system's efficiency. In order to improve the precision of the recommendation scheme, effective neural network methods using feed forward layers can be included with Association Rule Mining.

# Bibliography

|  |  |
| --- | --- |
| [1] | [Online]. Available: https://www.researchgate.net/publication/315049886\_Building\_Prediction\_Model\_using\_Market\_Basket\_Analysis. |
| [2] | S. K. Manpreet Kaur, "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining," *International Conference on Computational Modeling and Security (CMS 2016),* 2016. |
| [3] | M. M. Phani Prasad J1, "A Study on Market Basket Analysis Using a Data Mining Algorithm," *International Journal of Emerging Technology and Advanced Engineering,* 2013. |
| [4] | D. B. K. 2. ,. D. S. D. G. Roshan Gangurde 1, "Building Prediction Model using Market Basket Analysis," *International Journal of Innovative Research in Computer and Communication Engineering,* 2017. |
| [5] | D. R. S. K. R. K. Sinha2, "AFFINITY ANALYSIS: A COMPETITIVE TOOL FOR UNORGANISED RETAIL SECTOR," *PEZZOTTAITE JOURNALS.* |
| [6] | D. M. P. Dr. M. Dhanabhakyam, "A Survey on Data Mining Algorithm for Market Basket Analysis," *Global Journal of Computer Science and Technology,* 2011. |
| [7] | [Online]. Available: https://www.datascience.com/blog/k-means-clustering. |
| [8] | [Online]. Available: https://gerardnico.com/data\_mining/pca . |
| [9] | F. M. Last Name, "Article Title," *Journal Title,* pp. Pages From - To, Year. |
| [10 | F. M. Last Name, Book Title, City Name: Publisher Name, Year. |
| [11] | [Online]. Available: https://www.researchgate.net/publication/315049886\_Building\_Prediction\_Model\_using\_Market\_Basket\_Analysis. |