Market Basket Analysis

Team members: Shiva Sai Amaravadi
M Manikanta Venkata Pasumarthi
Sai Teja Uppu
Hariprasad Yedluri

ABSTRACT:-

Market Basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. Helps in analysing large-scale datasets. Whereas now data cleaning and a clear understanding have been done by using statistical analysis using apriori algorithm to find the product relation with the customer which helps in improving the business for the organisation.

Keywords—apriori algorithm.

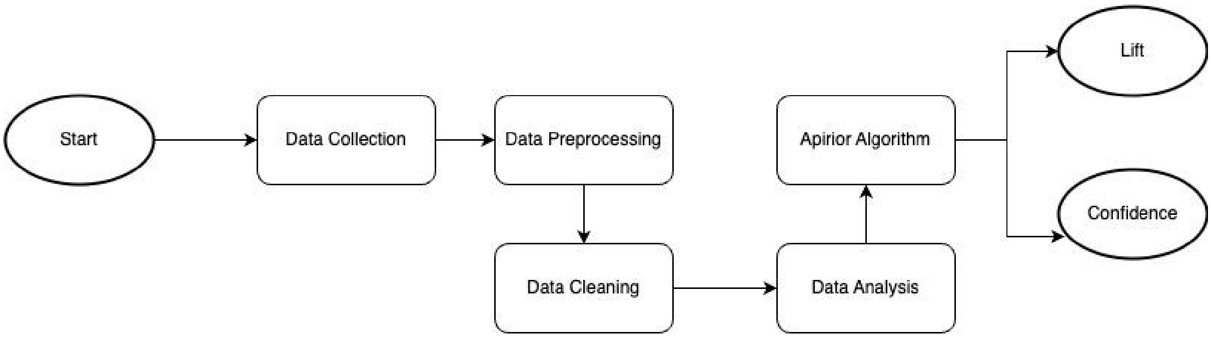
MOTIVATION AND SIGNIFICANCE:

The project is motivated by the need to address and get the analysis of products which are going to sale fast and first and the product which are associated with each other like if a person likes to buy milk he is likely to buy bread and eggs also.It is based on Association rule mining.

OBJECTIVES:

To help retailers increase sales through better Market Basket Analysis using statistical analysis using apirior algorithm.

WORKFLOW DIAGRAM WITH EXPLANATION



The Kaggle serving as a wonderful repository for datasets where we collected it from kaggle. The acquired statistics then undergoes a transformative preprocessing degree, where it's far cleansed and organized to render it amenable to evaluation. The concurrent facts cleansing phase makes a speciality of making sure the pristine quality of the dataset via rectifying errors. Once the data is prepared, a complete statistics analysis is carried out, unraveling styles and insights via exploratory data analysis. Transitioning into affiliation rule mining, the venerable Apriori algorithm takes the level, sifting through the Kaggle-acquired dataset to unveil frequent itemsets and articulate association regulations based on assist and confidence metrics. The subsequent assessment entails scrutinizing raise and self assurance, in which raise elucidates the chance of co-occurrence of related objects, whilst self belief gauges the reliability of such institutions

FEATURES:

The dataset has 3 columns Member number, Date and Item Description. The member number is a unique transaction number where each number represents the translation id. whereas the data is in day month and year format. And item description has product details with space separated by next.

I. DATASET

Market Basket Analysis is one of the important techniques used by retailers. It works by understanding the relationship between frequently brought products and the transaction. It helps business people understand which product is important and which people buy. The data set consists of 38765 rows of purchases of customers from supermarket stores. By history, we can analyse the data and association rules can be generated using the Market Basket Algorithm for example Aprioor algorithm. A few terms like support, confidence and lift are explored from this dataset.

In total there are 38765 rows of translation data is there in the dataset. The dataset has 3 columns Member number, Date and Item Description. The member number is a unique transaction number where each number represents the translation id. whereas the data is in day month and year format. And item description has product details with space separated by next. When data was collected from the famous Kaggle website which is very reliable and authentic source to collect data for the data science community. The columns are represented in int 64 and object format. Where as date also represented in object format. In a sequence of generating the features of our model date was converted into date format using famous python library called pandas. And members id are converted into string format becuase there are uninque in nature.

II. ANALYSIS OF DATA

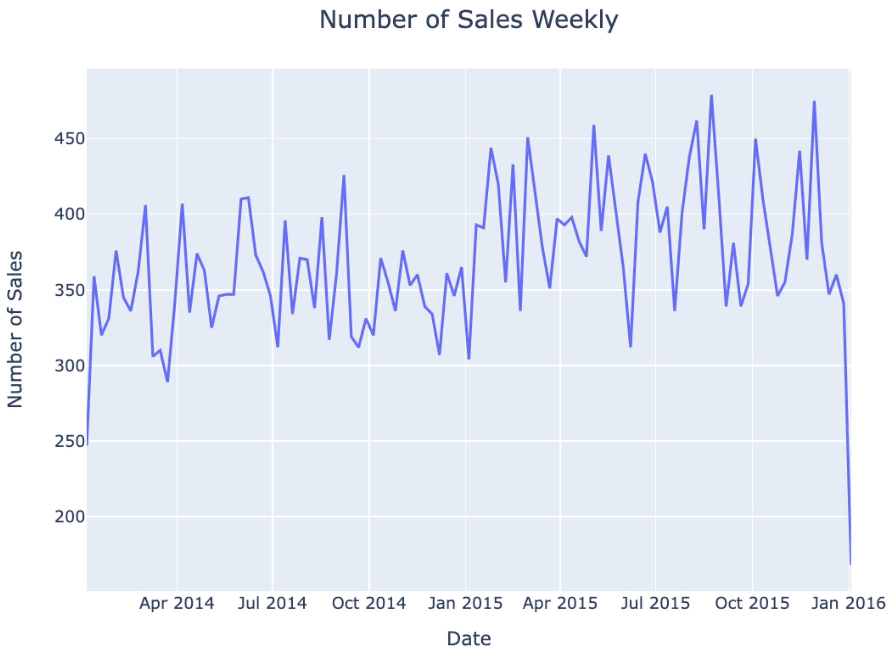


Figure 1 Number of sales weekly

This graph Figure 1 the number of sales per week as depicted in the graph below. There has been an incline in the sales recorded on a weekly basis over the last few months. From August 2014 to July 2015, there was a small growth in the sale while there were more considerable gains in sale from July 2015 to January 2016. In the given graph, January 2016 registers the highest number of sales among all the other months. However, some probable reasons for the sales increase can be identified. The firm may have introduced a new product that is in demand among its clients. On the other hand, a company could have spread its wings in new marketplaces. Moreover, it could have been due to the fact that the company might have enhanced its marketing and advertising campaigns.If there are no details on why sales have increased, it would be hard to tell for certain why this has happened. Nevertheless, the graph depicts an unquestionable upward movement in sales. This denotes a good omen for the company as it indicates the right direction forward.

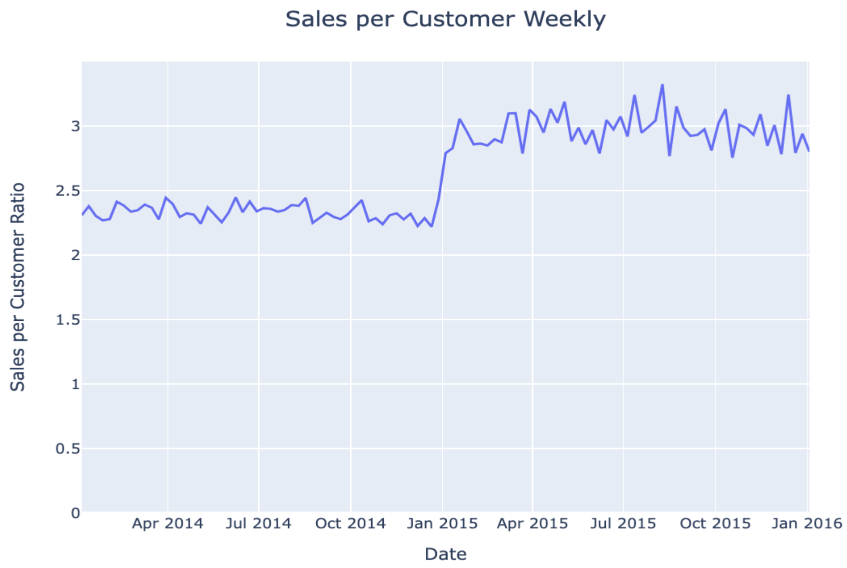


Figure 2 Sales per Customer weekly

The graph that you have forwarded indicates customer's per week. As seen on the graph, customer numbers increase with time albeit sporadic changes occur in between. Then in march 2014, it reached the worst point whereby there were just eighty visitors to the site. Nonetheless, customers started growing with momentum thereafter, and in December 2014, it totaled to 160. The highest number of the customers was recorded at 180 in January 2015. Nonetheless, the number of customers started to fall slightly from mid-January 2015. By February 2015, there were a total of 170 customers while March showed that customers had reduced to a total of 160. In April 2015, the number of customers stood at 150. However, there was an upsurge for the number of customers in May 2015, resulting in a rise from 270 to 160. In August, after another 170 people joined, the customer base was increasing again. By August 2015, there were 180 customers, the same peak shown in the graph. Nonetheless, customer number gradually reduced since August 2015. The amount of customers for the month of Sept 2015 was 170 while for Oct 2015 was 160. In November 2015, the drop continued and reached a total of 150 customers for that month. Nevertheless, the number of customers started increasing again in December 2015, while by January 2016, there were 160 customers.

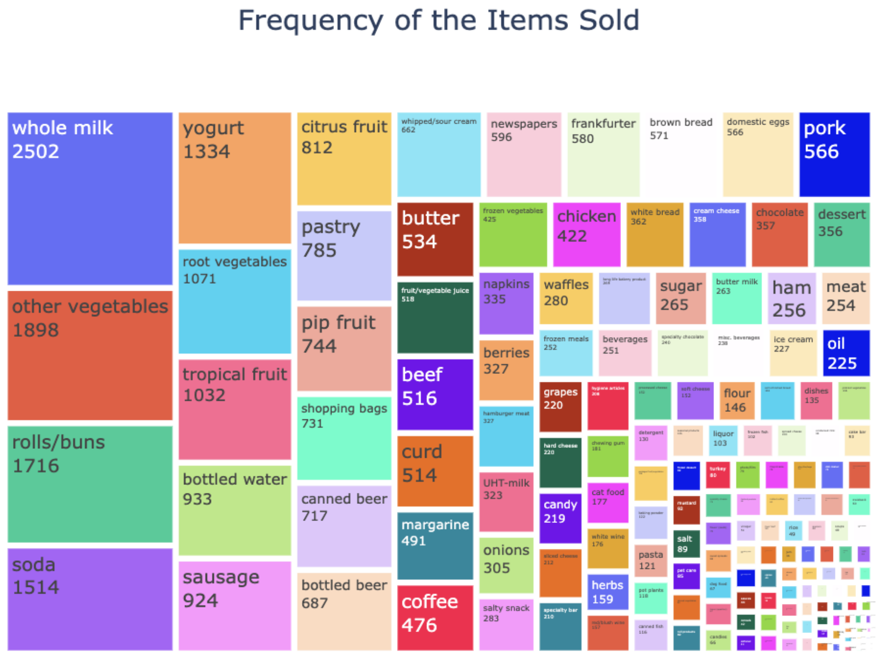


Figure 2 Frequency of the Items Sold

Frequency of the items sold. They usually sell bottled beer, water, and whole milk. They are cheap and highly demanded. For instance, other regularly purchased items are brown bread, coffee, grapes, candy, ham, eggs, beef, and pork, shopping bags, newspapers, and white bread.These include cream cheese, pasta, and berries that happen to be the least frequent sales. They are priced expensively, or less popular. Some other occasional items are citrus fruit, canned beer, frankfurters, butter, margarine, and napkins.It is evident in the graph as well, that some items are sold in greater quantities than others. Another scenario is that whole milk is more popular than bottled water, whereas bottled water is more popular than bottled beer. Whole milk forms part of a normal diet but other drinks like bottled water and bottled beers are optional.In a similar way, some items are sold at different quantities as depicted well on the graph. As an illustration, the amount of whole milk may be up to about 1716, and bottled water amounting to approximately 785. This is probably because of the larger amounts sold of whole milk compared to bottled water.In general terms, the graph reveals a range in which the frequency at which given items are sold depends on each item. Some products are sold on a daily basis, while others are bought at different times, as well as, at different volumes of purchase.

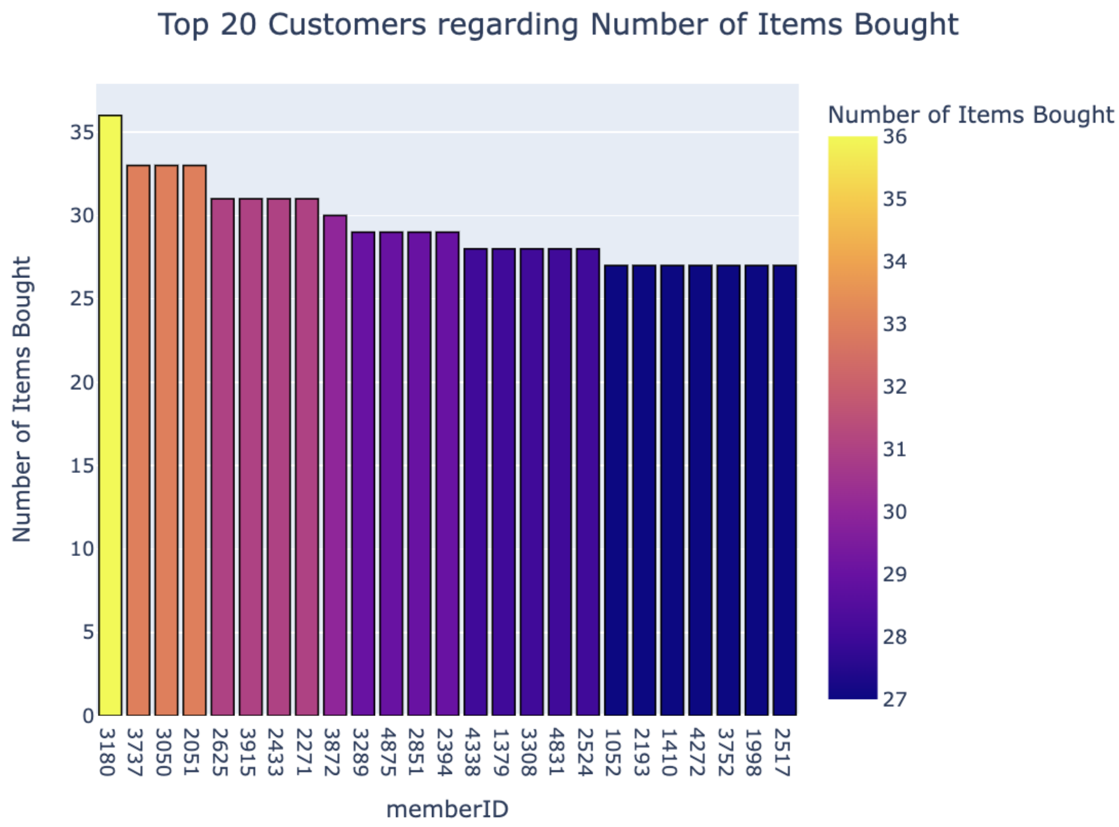


Figure 3 Customer regarding NUmber of items brough

As per the graph 3180 are the best customers who brought frequently, as per the dataspurce the graph is drarm by the number of items brough by the members Id. The graph sorted the best customers to worst customers. Top right are the best which shows in the yellow colous and blue colours are not good customers are per the graph.

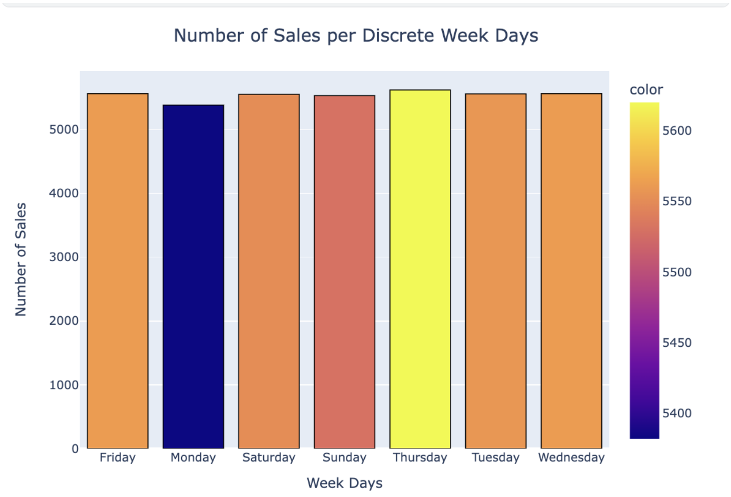


Figure 4 Number of Sales per Discrete Week Day

As per the above graph Figure 4 it is drawn between the number of sales and the week. To identify the mood of customers during the day of the week. We can understand that during Thursday people are likely to shop more than the rest of the day . next comes on Sunday and saturday cause its a weekend. To interpret the graph the more yellow the more customers purchased on the day of the week.

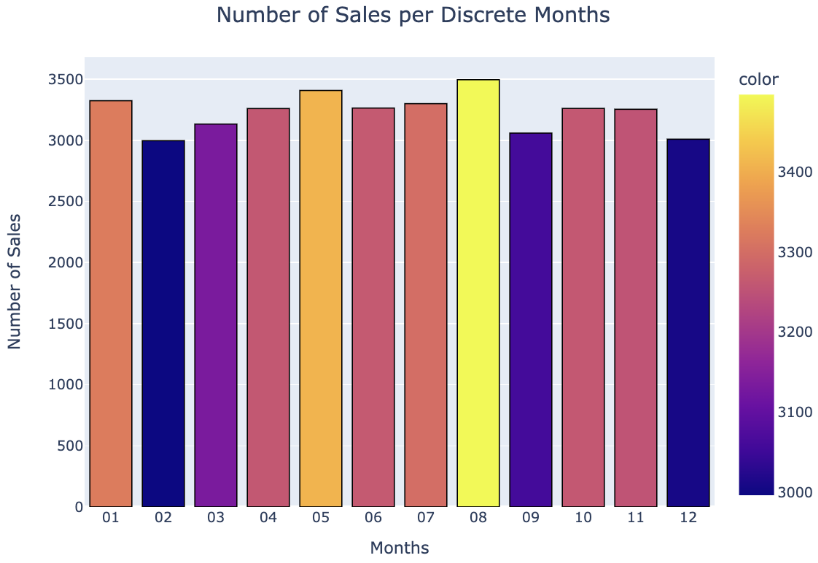


Figure 5 Number of sales per discrete month

The above graph Figure 5 states the number of sales vs the month. To understand which month had the most sales during the year. So the next year's market can be analysed. As we can see during 8th month sales are high and in 1st month are in 2nd place.

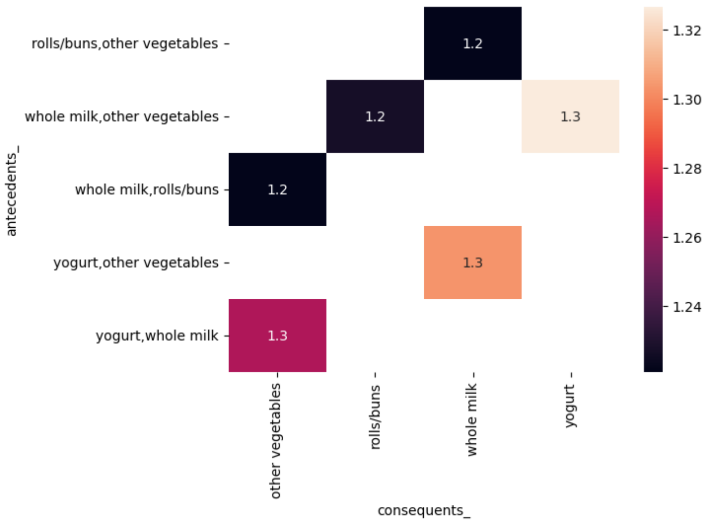


Figure 6 Heat Map antecedents vs consequents

The above heat graph (Figure 6) shows the correlation between better consequents vs antecedents where we can see the clear correlation between [other vegetables and yoghurt & other vegetables], [other vegetables vs whole milk vegetables], [other vegetables vs rolls/buns]

Whereas yoghurt has a very high correlation between milk, rolls, and other vegetables.

Market Basket Analysis

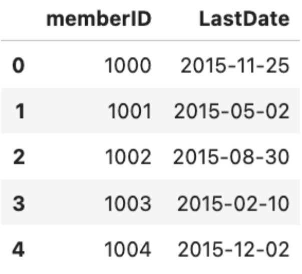
Stores use market basket analysis to figure out what people like to buy together. They look at a lot of receipts to see what products are often bought together. Then, they use this information to make decisions about how to stock their shelves and what discounts to offer. This can help them sell more products and make more money. The customers-items matrix is a table that shows what products each customer has bought. Each row in the table represents a different customer, and each column represents a different product. The numbers in the table show how many times each customer has bought each product. The most items that appeared together are butter and shopping bags, and spread cheese.

Created association rules for indicating antecedent and consequent items add Code Add Markdown, sausage = yogurt, rolls/buns, root vegetables, whole milk = shopping bags, rolls/buns, soda = sausage, butter, whole milk = yogurt, and etc. have strong relationships.

RFM Analysis:

Calculate the Recency

To last purchase date of each customer

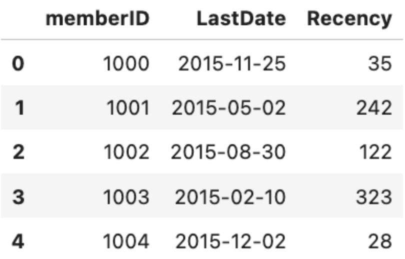


	memberID	LastDate
0	1000	2015-11-25
1	1001	2015-05-02
2	1002	2015-08-30
3	1003	2015-02-10
4	1004	2015-12-02

To last date for our dataset

Timestamp('2015-12-30 00:00:00')

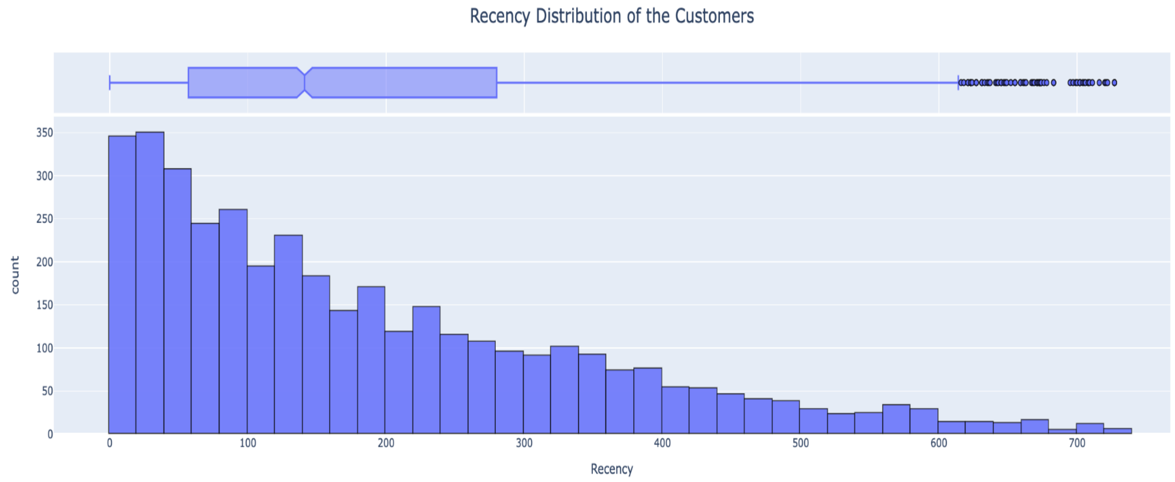
Calculating Recency by subtracting (last transaction date of dataset) and (last purchase date of each customer)



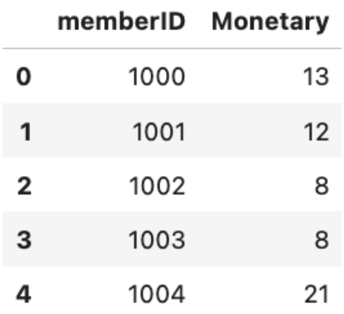
	memberID	LastDate	Recency
0	1000	2015-11-25	35
1	1001	2015-05-02	242
2	1002	2015-08-30	122
3	1003	2015-02-10	323
4	1004	2015-12-02	28

Recency Distribution of the Customers

* Visit Frequency

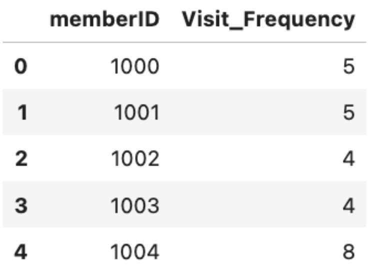


* Visit Frequency Distribution of the Customers



	memberID	Monetary
0	1000	13
1	1001	12
2	1002	8
3	1003	8
4	1004	21

Monetary



	memberID	Visit_Frequency
0	1000	5
1	1001	5
2	1002	4
3	1003	4
4	1004	8

III.Implementations

Apriori algorithm refers to an algorithm for computing association rules between objects. This shows how two or more things are related. In other words, we can say that the apriori algorithm is an inverse association rule that checks if people who bought product A also bought product B .

The main objective of the apriori algorithm is to establish association rules between objects. The association law describes how two or more entities are related to each other. Apriori algorithms are also known as frequent pattern mining. Typically, you apply the Apriori algorithm to a database with many jobs. Let us understand the apriori algorithm with the help of an example; Imagine going to Big Bazaar to buy different things. It helps consumers shop faster and increases the size of retail markets. In this tutorial, we will discuss the apriori algorithm with an example.

Apriori algorithm refers to the algorithm used in mining frequent product sets and related association rules. Generally, the apriori algorithm works on a database with a large number of transactions. For example, merchandise customers but in a Big Bazaar. The Apriori algorithm helps customers shop faster and increases the sales performance of a particular store.

The given three components comprise the apriori algorithm.

Support, Confidence and Lift

Support refers back to the default reputation of any product. You find the aid as a quotient of the department of the quantity of transactions comprising that product by the total range of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)

= four hundred/4000 = 10 percent.

Confidence refers to the opportunity that the clients offered each biscuits and candies collectively. So, you need to divide the variety of transactions that incorporate both biscuits and sweets with the aid of the entire wide variety of transactions to get the self belief.

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits)= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

Consider the above example; raise refers back to the boom inside the ratio of the sale of sweets while you promote biscuits. The mathematical equations of lift are given below.

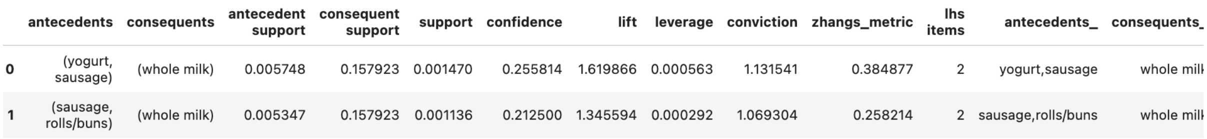
Lift = (Confidence (Biscuits - chocolates)/ (Support (Biscuits)

= 50/10

= 5

IV. Result

Confidence:



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	lhs items	antecedents_	consequents_
0	(yogurt, sausage)	(whole milk)	0.005748	0.157923	0.001470	0.255814	1.619866	0.000563	1.131541	0.384877	2	yogurt,sausage	whole mil
1	(sausage, rolls/buns)	(whole milk)	0.005347	0.157923	0.001136	0.212500	1.345594	0.000292	1.069304	0.258214	2	sausage,rolls/buns	whole mil

The table shows the following:

The maximum frequent itemset is (complete milk), which has a assist of zero.384877.

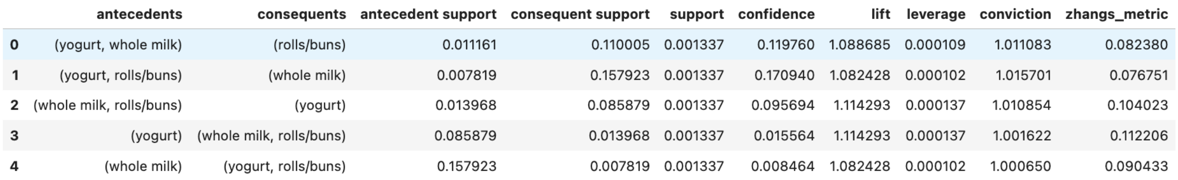
The strongest affiliation is among the itemsets (yogurt, sausage) and (rolls/buns), with a lift of 1.619866 and a conviction of 1.131541. This way that clients who purchase yogurt and sausage are 1.619 times much more likely to additionally buy rolls/buns than clients who do now not purchase yogurt and sausage.

The desk may be used by shops to pick out product combos which might be regularly bought collectively. This records can then be used to create promotions or to location products close together in the store.

Here is an instance of ways the table will be used:

A retailer may want to see from the table that clients who buy yogurt and sausage also are likely to buy rolls/buns. The retailer should then create a advertising that gives a discount on rolls/buns to clients who additionally buy yogurt and sausage. This might encourage clients to buy more products and boom the retailer's sales.

Lift:

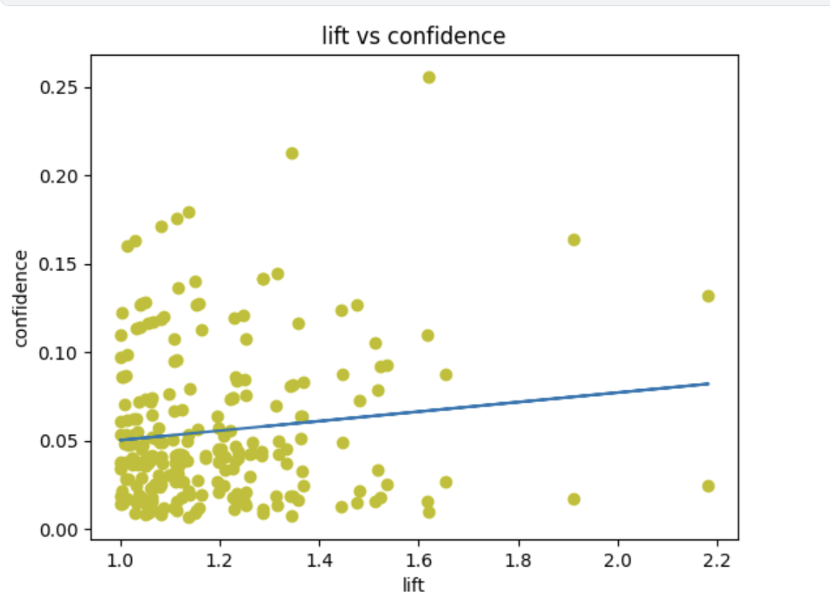


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(yogurt, whole milk)	(rolls/buns)	0.011161	0.110005	0.001337	0.119760	1.088685	0.000109	1.011083	0.082380
1	(yogurt, rolls/buns)	(whole milk)	0.007819	0.157923	0.001337	0.170940	1.082428	0.000102	1.015701	0.076751
2	(whole milk, rolls/buns)	(yogurt)	0.013968	0.085879	0.001337	0.095694	1.114293	0.000137	1.010854	0.104023
3	(yogurt)	(whole milk, rolls/buns)	0.085879	0.013968	0.001337	0.015564	1.114293	0.000137	1.001622	0.112206
4	(whole milk)	(yogurt, rolls/buns)	0.157923	0.007819	0.001337	0.008464	1.082428	0.000102	1.000650	0.090433

The carry table can be used to pick out the strongest institutions among itemsets. This information can then be used to create promotions or to region products close together in the shop.

For example, the elevate table suggests that the affiliation rule (yogurt, sausage) -> (rolls/buns) has a boost of one.619866. This means that customers who buy yogurt and sausage are 1.619 instances much more likely to additionally buy rolls/buns than clients who do not buy yogurt and sausage. This is a sturdy affiliation, and the store may want to remember developing a promotion that gives a reduction on rolls/buns to customers who also buy yogurt and sausage.

The carry table can also be used to discover poor institutions between itemsets. For instance, the raise table suggests that the association rule (entire milk, rolls/buns) -> (yogurt) has a boost of zero.834941. This method that clients who buy complete milk and rolls/buns are much less probable to additionally purchase yogurt than clients who do not purchase whole milk and rolls/buns. This is a poor association, and the retailer may additionally want to avoid placing yogurt subsequent to complete milk and rolls/buns in the shop.



The figure is a scatter plot showing the relationship among raise and self assurance. The blue line represents the self belief degree, and the green dots constitute the raise level.

The plot shows that the elevate degree is generally lower than the self assurance degree. This is due to the fact lift is a extra stringent degree of association than self belief. Lift takes under consideration the guide of the antecedent and consequent itemsets, whilst self belief does now not.

The plot additionally indicates that there are some outliers, wherein the elevate level is a good deal better than the self assurance level. These outliers constitute robust associations between itemsets.

Overall, the plot suggests that the Apriori algorithm is effective at figuring out robust institutions between itemsets. However, it is essential to observe that the lift degree is generally lower than the confidence degree.

Here are a few unique insights that can be gained from the plot:

- The most powerful associations have elevate values of greater than 1.Five.
- The weakest associations have carry values of much less than zero.Five.
- There are a few outliers with raise values which might be a good deal better than 1.Five. These outliers represent the strongest institutions within the dataset.
- The universal trend is that the carry stage is lower than the self belief degree. This is because raise is a greater stringent measure of affiliation than self belief.

Retailers can use this records to perceive product mixtures which might be frequently purchased collectively. This statistics can then be used to create promotions or to location merchandise close together in the shop.

Implementation status report:

- Work completed
    - Data cleaning:-
    - ○ Removed numbers from our dataset in a required column that is question column
    - ○ Removed punctuations for the same column.
    - ○ Replaced negations with antonyms as cited in the code
    - ○ And also dealt with capitalised phrases
    - ○ Removed forestall words that are in large part took place in our dataset.
    - ○ Replaced repetition of punctuations with respective sting as required for our dataset.
    - ○ Replaced contractions for the query column
    - ○ Lowered the case for the complete dataset

- Analysis
    - Number of sales weekly
    - Number of customers weekly
    - Sales per customer weekly
    - Fequency of the Items Sold
    - Top 20 Customers regarding number of items bought
    - Number of Sales per discrete week days
    - Number of sales per discrete Month

Task	Person	Status	Contribution
Data Cleaning	Sai Teja	completed	25%
Analysis	Shiva sai	completed	25%
Analysis	Hariprasad	completed	25%
Model Training	Manikanta	completed	25%

I. References

1. Apriori-algorithm https://www.javatpoint.com/apriori-algorithm
2. *Groceries-dataset https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset*
3. Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining https://www.sciencedirect.com/science/article/pii/S1877050916305208
4. Market Basket Analysis for a Super Market https://www.researchgate.net/publication/365489098
5. MARKET BASKET ANALYSIS: TREND ANALYSIS OF ASSOCIATION RULES IN DIFFERENT TIME PERIODS https://run.unl.pt/bitstream/10362/80955/1/TEGI0458.pdf
6. A Study on Market Basket Analysis Using a Data Mining Algorithm https://www.academia.edu/73783800/A_Study_on_Market_Basket_Analysis_Using_a_Data_Mining_Algorithm