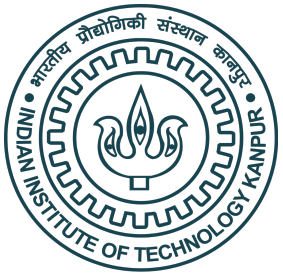
Data Mining Project

Subject Code:MTH552A

Association Rule Mining and Market Basket Analysis Using Apriori Algorithm

Submitted by
Bidhisha Ghosh-211423
Sampriti Dutta-211366

under supervision of
Dr.Amit Mitra



Department:STATISTICS
IIT KANPUR

Acknowledgement

We would like to express a deep sense of thanks and gratitude to Dr. Amit Mitra for providing us the opportunity to prepare this project and constantly motivating us with constructive advices. It has been a great learning experience building practical insights of the theoretical knowledge gathered during course lectures.

Last, but not the least, our parents provided us with continuous encouragement and extensive support throughout the session. So , with due regards we express our gratitude to them for completion of the project within the stipulated time-period.

Abstract

Product suggestions are an important part of customizing sales in any store. If a customer is buying a certain item from a store, he/she is much likely to buy some other items along with it to complement the first item bought by him/her. For example, if a customer buys a shirt, he is likely to buy a pant or jeans along with it. Depending on some previous purchases of a customer, we can figure out what items he/she is going to purchase next and thus provide the customer with suitable product suggestions for their ease of shopping. This is where Market Basket Analysis comes in. In Market Basket Analysis (MBA), we find association rules among products, based on the probability of purchasing an item, say item B, provided that the customer has already purchased another item, say item A. This analysis can be used as a way to offer products based on items that has been purchased often together. This project deals with a dataset from a groceries store, provided by Kaggle. In this project, we have grouped the data with respect to the items purchased corresponding to the customer number and the date of purchase. We have then used the apriori algorithm to find the possible basket items (or the frequent item sets) that can be purchased together and then finally provided the association rules based on the confidence of each frequent item sets.

Introduction

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction. Apriori Algorithm is a widely-used and well-known Association Rule algorithm and is a popular algorithm used in market basket analysis.

Our main objective in this project is to increase tha sale by better understanding customer purchasing pattern.To meet this purpose we have taken the data of sale in a grocery store and by analyzing the purchase history, we reveal some product groupings, as well as products that are likely to be purchased together.

Association Rule Mining (ARM)

Association rule mining finds interesting associations or relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Basket Analysis: The main objective is to find associations between different itemsets that customers place in a shopping basket. Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently. Using this technique, given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

Applications:
The main applications of Market Basket Analysis are in the following field:

1. Cross Marketing
2. Catalogue or Floor Design
3. Web log Analysis for e-commerece

Itemset:A set of items together is called an itemset. If any itemset has k-items it is called a k-itemset. An itemset consists of two or more items.

Data Structures: Let, the set of transactions be $T = t_1, t_2, ..., t_n$. Where each t_k is an itemset.

The rule is generated in the form **Antecedent** $\rightarrow$ **Consequent**.

Our aim is to find frequent patterns, i.e. associations among sets of items in T. Then represent these relationships as association rules of the form: $X \rightarrow Y$.

Support count: Support count is the number of occurances of an itemset in the database. It is denoted as S({itemset})

Support: Support is the fraction of transactions containing the itemsets. It is denoted as $\frac{S(\{itemset\})}{|T|}$

Frequent itemset: An itemset is said to be a frequent itemset if its support is greater than or equal to a minimum threshold value, say min Sup.

Confidence: Confidence is the measure of how often the consequent appears in the transactions containing the antecedant. $C(A \rightarrow B) = \frac{S(A,B)}{S(A)}$

Lift: lift refers to the increase in the ratio of the sale of item B when you sell item A. The mathematical equations of lift are given below. Lift = $C(A \rightarrow B)$/ (S(A)

Pruning:Pruning means to change the model by deleting the child nodes of a branch node. The pruned node is regarded as a leaf node.

ARM task: Given a set of transactions, the goal of ARM is to find all rules such that

i $Support \geq$ min Sup

ii $Confidence \geq$ min conf

where min Sup and min conf are fixed aprior.

Apriori Algorithm

Apriori Algorithm is one of the algorithm used for transaction data in Association Rule Learning. It allows us to mine the frequent itemset in order to generate association rule between them. It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule leaning that analyzes that people who bought product A also bought product B

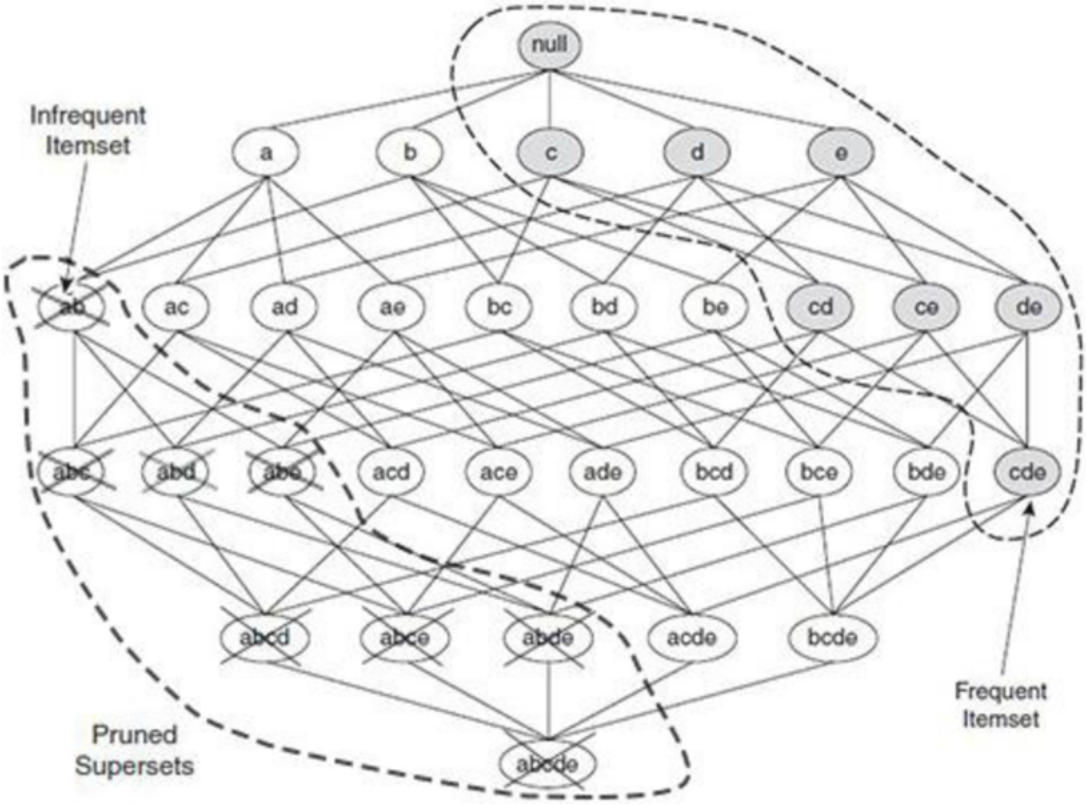
Components of Apriori Algorithm

The given three components comprise the apriori algorithm.
1. Support
2. Confidence
3. Lift

Principles in Apriori Algorithm

The key concept of Apriori algorithm is its anti-monotonicity of support measure.
1. Apriori assumes that all subsets of a frequent itemset must be frequent(Apriori property).
2. If an itemset is infrequent, all its supersets will be infrequent.

This can be explained by the following example:



In the above example **ab** is found to be infrequent (i.e. support¡ min sup). Then all itemsets which are supersets of **ab** (**abc, acd, abe, abcd, abce, abde, abcde**)are also infrequent .All such branches will be pruned.
Also **cde** is found to be frequent (i.e. support¿ min sup).Then all subsets od **cde**(**cd, ce, de, c, d, e**)are frequent.

2 Step ARM of Apriori Algorithm

Step 1 Generate all frequent itemsets with $support \geq$ min Sup

Step 2 Generate association rules using these frequent itemsets.

Pseudo Code of Step 1 of Apriori Algorithm

- Set k=1
- Generate frequent itemsets of length 1.
- Prune itemsets of higher order (i.e. Supersets), if necessary.
- Generate itemsets of length k+1 from frequent itemsets of length k. Compute the support of new itemsets with respect to min Sup.
- Prune if necessary.
- Set k= k+1.
- Repeat until no frequent itemsets are found.

Generation of itemsets for next level

Let L_k denote the frequent itemsets at level k and C_k denote the set of all candidates at level k. At first the items in L_{k-1} are listed in an order.

Step 1: Join 2 items from L_{k-1}, i.e. self joining as$L_{k-1} * L_{k-1}$.
Let two itemsets be $[p_1, p_2, ..., p_{k-1}]$ and $[q_1, q_2, ...q_{k-1}]$,under the given order.
Then insert the itemset $[p_1, p_2, ..., p_{k-1}, q_{k-1}]$ into C_k.

Step 2: For all itemsets C in C_k and for all (k-1) subsets S of C(under the given order); prune C from C_k if S is not in C_{k-1}.

Pseudo Code of Step 2 of Apriori Algorithm

- Generate association rules using frequent itemsets.
- Given any frequent itemset L; find all non empty subsets F of L'
- Output each rule $F \rightarrow \{L - F\}$ that satisfies the threshold on confidence.

Exploaratory Data Analysis

In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

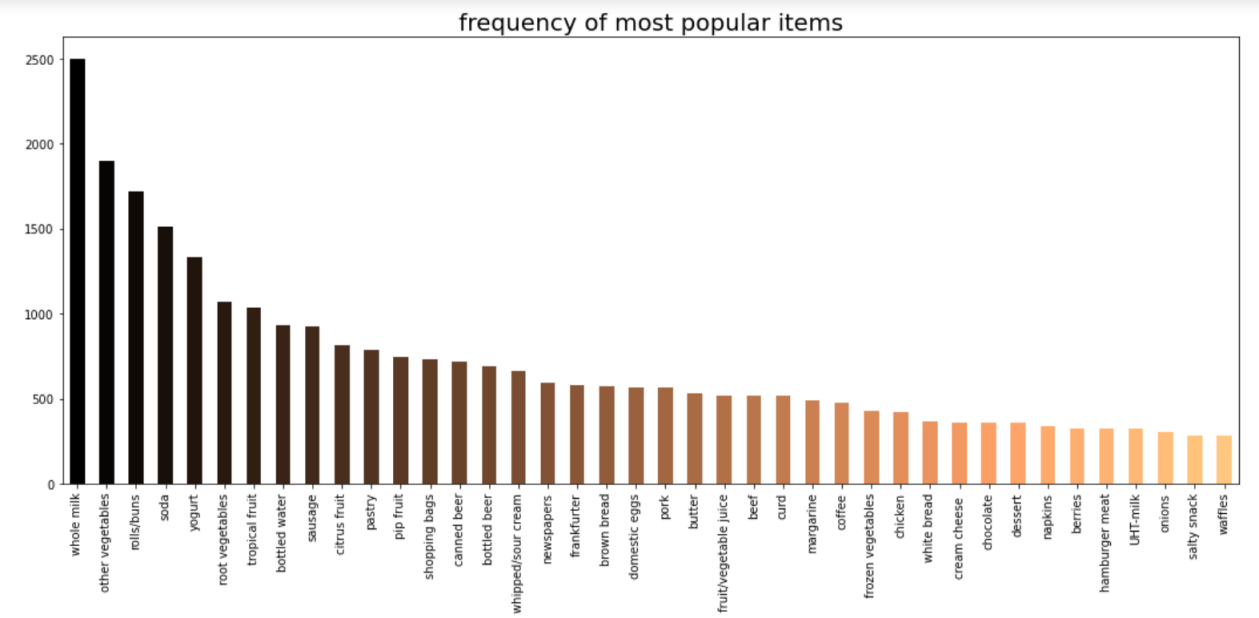


Fig.1 Bar plot of frequency of the overall sale of items

Insights:The sale of whole milk is maximum in that community and the sale of waffles is minimum among all items.

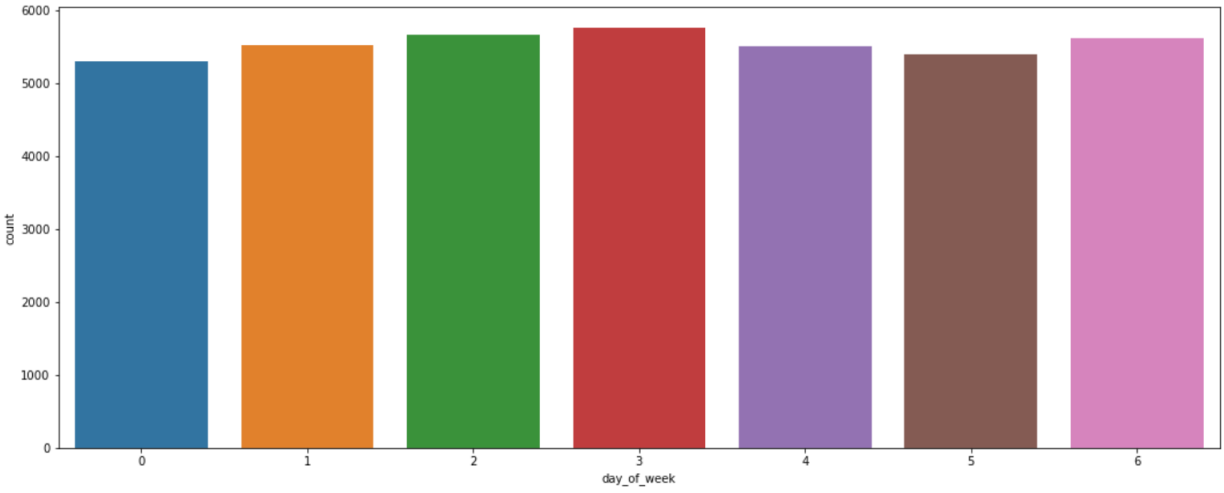


Fig.2 Sale of items on different days of a week

Insights:There is no significant difference in the sale of various items in a week.

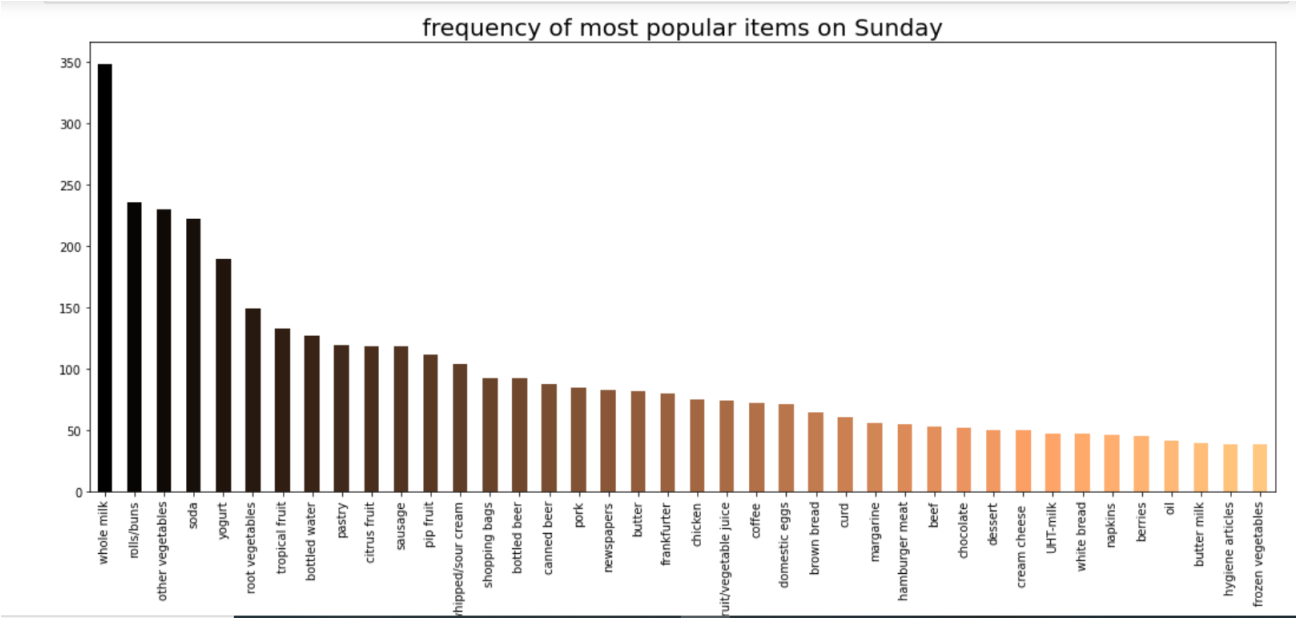


Fig.3 Barplot of frequency of sale of items on Sunday

Insights:Whole milk is sold most frequently on Sunday.

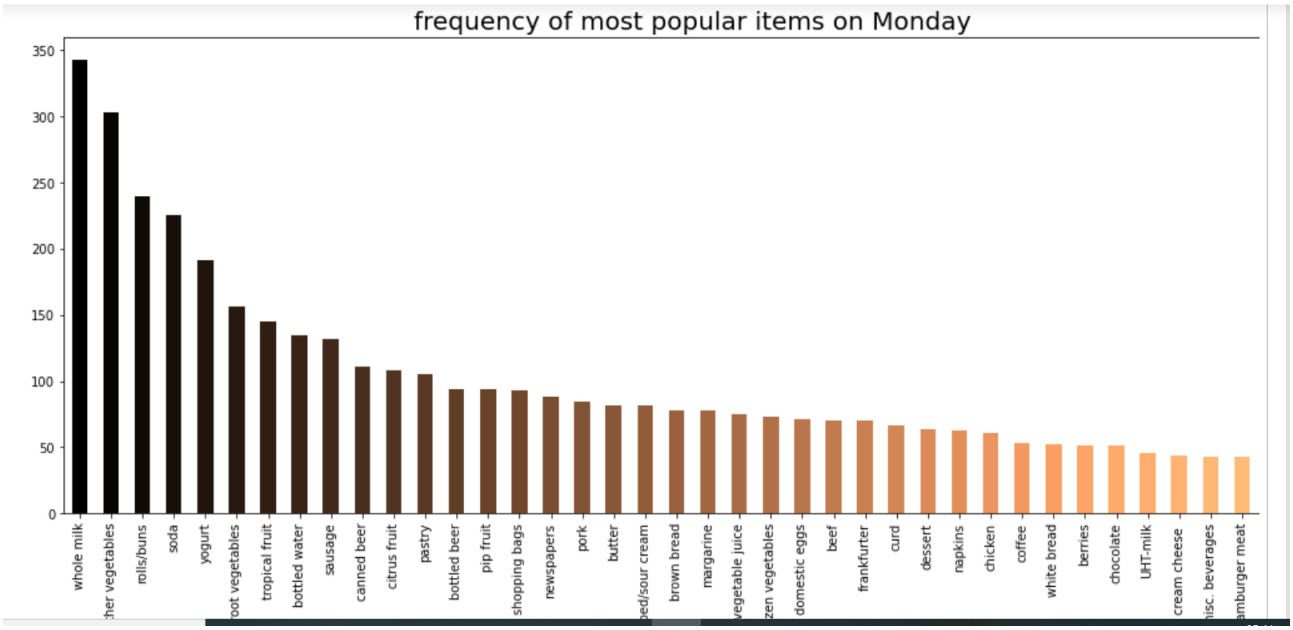


Fig.4 Barplot of frequency of sale of items on Monday

Insights:Whole milk is sold most frequently on Monday also but the sale of vegetables is also high on that day.

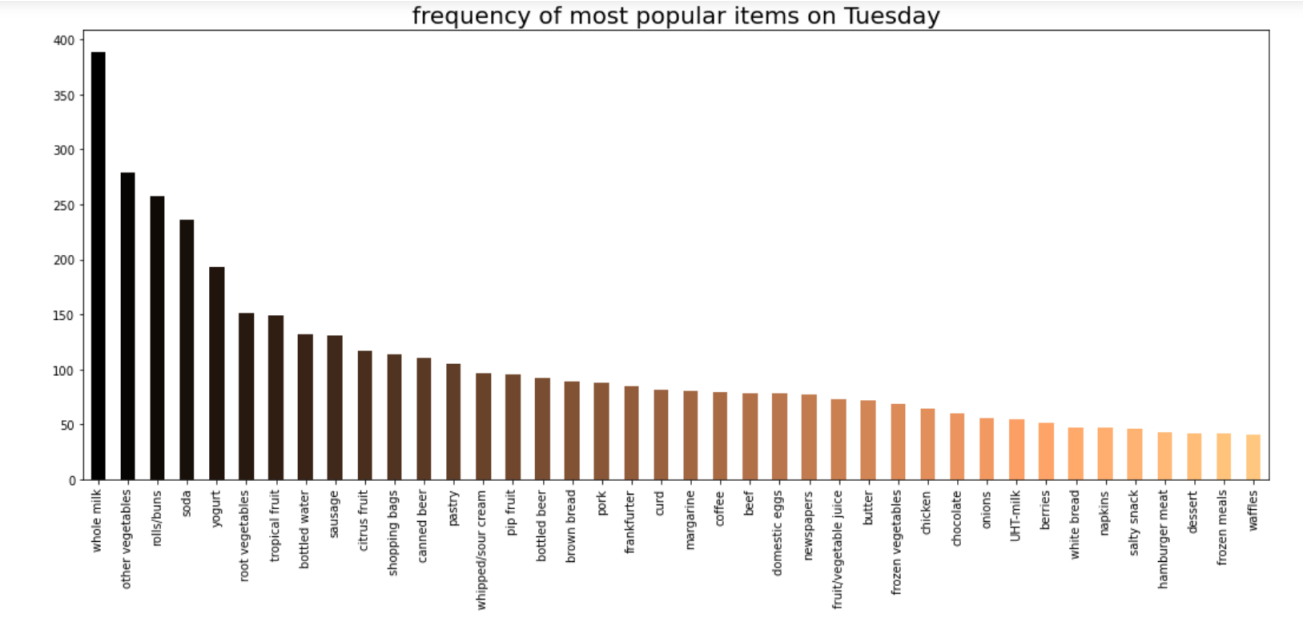


Fig.5 Barplot of frequency of sale of items on Tuesday

Insights:Whole milk is sold most frequently on Tuesday.

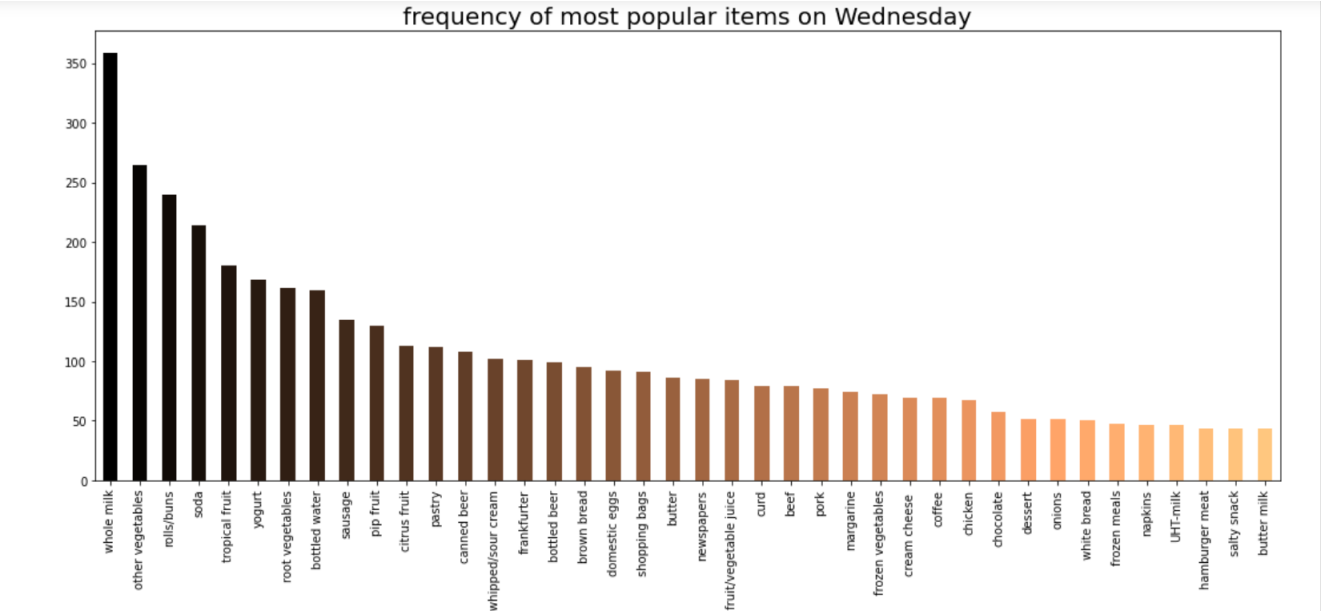


Fig.6 Barplot of frequency of sale of items on Wednesday

Insights:Whole milk is sold most frequently on Wednesday.

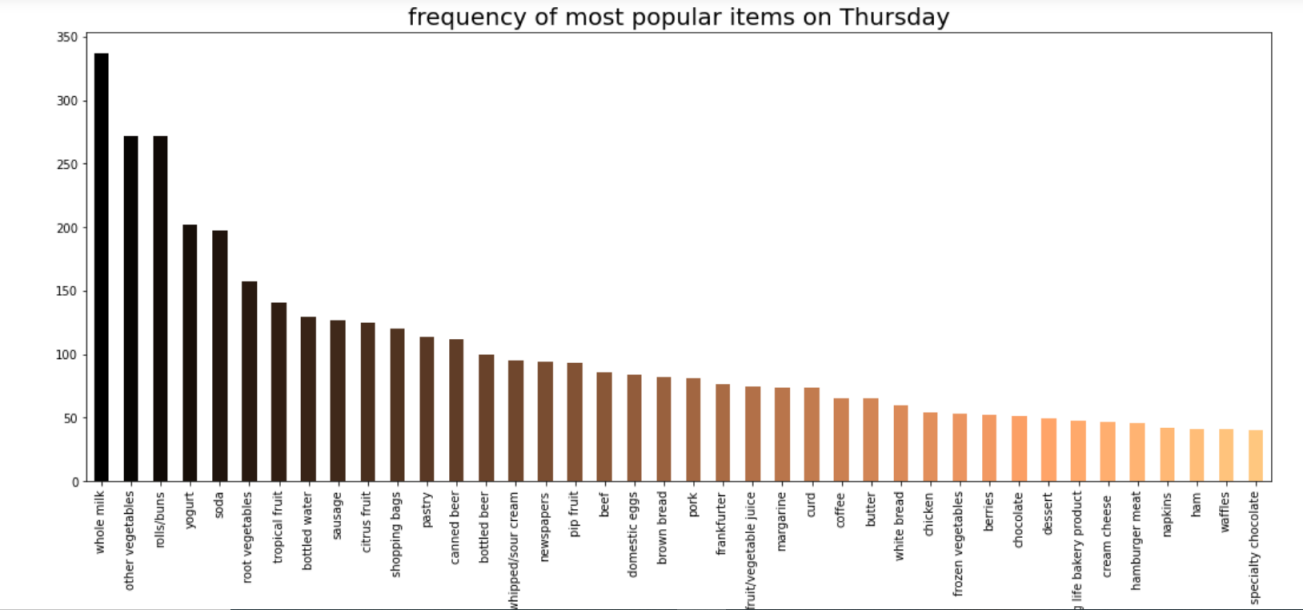


Fig.7 Barplot of frequency of sale of items on Thursday

Insights:Whole milk is sold most frequently on Thursday too and also the sale of buns and vegetables are also on same level.

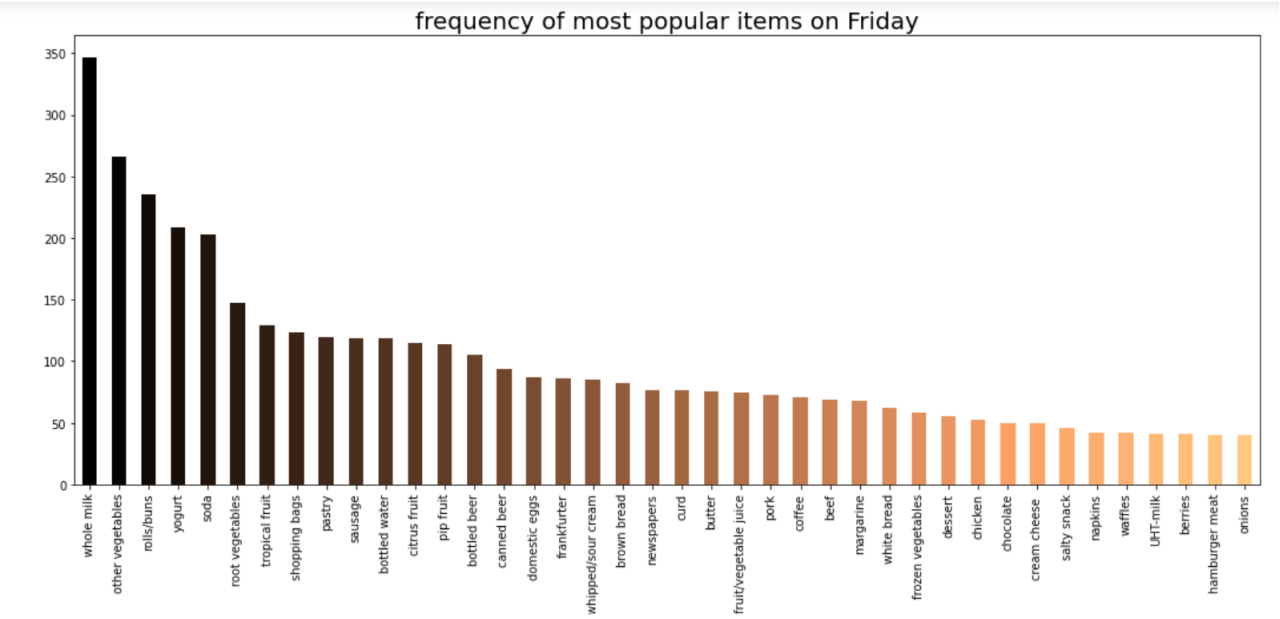


Fig.8 Barplot of frequency of sale of items on Friday

Insights:Whole milk is sold most frequently on Friday and then gradually the sale of other items decreases.

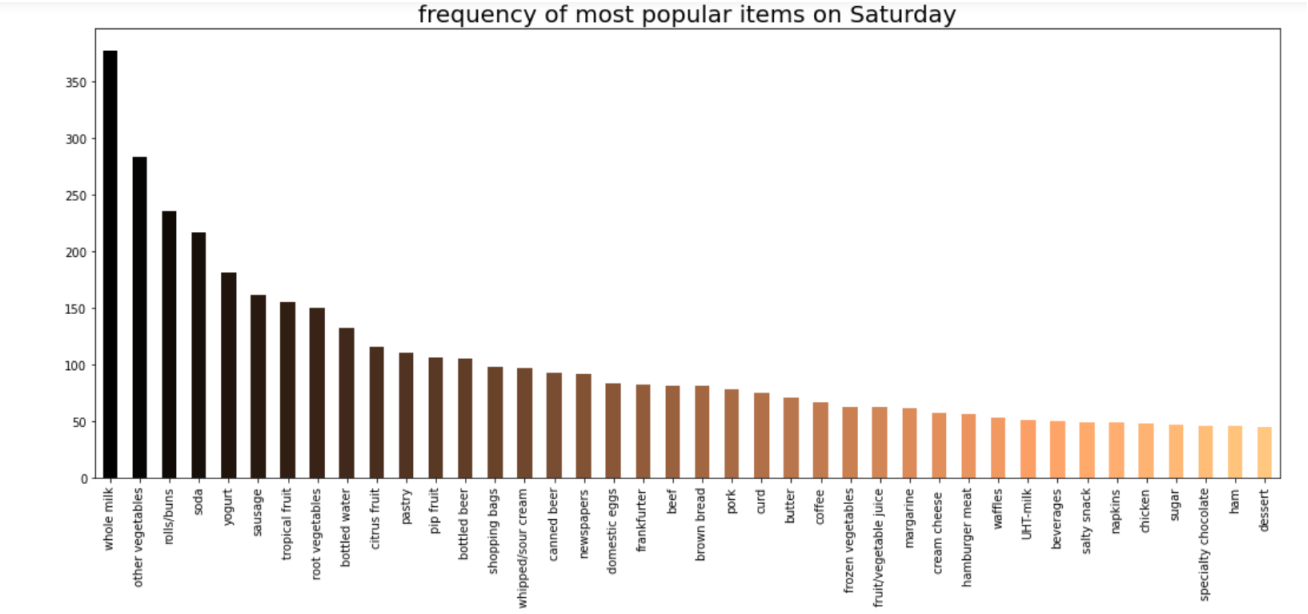


Fig.9 Barplot of frequency of sale of items on Saturday

Insights:Whole milk is sold most frequently on Saturday too and for the rest of items it is decreasing.

Market Basket Analysis for the Data of Grocery Items

1. After processing and cleaning the data, the count for the different items has been taken.

```
groceries.groupby('itemDescription').size().sort_values(ascending=False)
```

```
itemDescription
whole milk                2502
other vegetables          1898
rolls/buns                1716
soda                      1514
yogurt                    1334
                          ...
rubbing alcohol              5
bags                         4
baby cosmetics               3
preservation products        1
kitchen utensil              1
Length: 167, dtype: int64
```

2. According as the apriori algorithm, to get the frequent itemsets, minimum value of support(i.e. min Sup) was taken as 0.005 and we obtained a set of 125 itemsets .

3. Finally the minimum value of the confidence (i.e. min Con) has been taken as 0.13 and those itemsets has been kept for which confidence $\geq$ min con.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(bottled beer)	(whole milk)	0.045312	0.157923	0.007151	0.157817	0.999330	-0.000005	0.999874
8	(sausage)	(whole milk)	0.060349	0.157923	0.008955	0.148394	0.939663	-0.000575	0.988811
5	(newspapers)	(whole milk)	0.038896	0.157923	0.005614	0.144330	0.913926	-0.000529	0.984114
2	(domestic eggs)	(whole milk)	0.037091	0.157923	0.005280	0.142342	0.901341	-0.000578	0.981834
4	(frankfurter)	(whole milk)	0.037760	0.157923	0.005280	0.139823	0.885388	-0.000683	0.978958
3	(frankfurter)	(other vegetables)	0.037760	0.122101	0.005146	0.136283	1.116150	0.000536	1.016420
7	(pork)	(whole milk)	0.037091	0.157923	0.005012	0.135135	0.855703	-0.000845	0.973652
6	(pip fruit)	(whole milk)	0.049054	0.157923	0.006616	0.134877	0.854071	-0.001130	0.973362
1	(citrus fruit)	(whole milk)	0.053131	0.157923	0.007151	0.134591	0.852259	-0.001240	0.973040
9	(shopping bags)	(whole milk)	0.047584	0.157923	0.006349	0.133427	0.844887	-0.001166	0.971732

Results

Apriori algorithm was the first algorithm that was proposed for frequent itemset mining. It was later improved by R Agarwal and R Srikant and came to be known as Apriori, because it uses prior knowledge of frequent itemsets properties.
In our study, we can interpret the confidences as: 15% of the customers, who purchased bottled beer also bought whole milk.
14% of the customers, who purchased sausage, newspaper, domestic eggs, frankfurter respectively also bought whole milk.
13% of the customers who purchased frankfurter also bought other vegetables.
13% of the customers, who purchased pip fruits, citrus fruit and shopping bag also bought whole milk.
If minimum confidence is 13% , then these rules can be considered as strong association rules.

Conclusions

Apriori Algorithm can be slow. The main limitation is time required to hold a vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets i.e. it is not an efficient approach for large number of datasets. For example, if there are 10^4 from frequent 1- itemsets, it need to generate more than 10^7candidates into 2-length which in turn they will be tested and accumulate. Furthermore, to detect frequent pattern in size 100 i.e. v1, v2... v100, it have to generate 2^{100} candidate itemsets that yield on costly and wasting of time of candidate generation.
Though some limitations exist , the apriori algorithm is widely used on a database containing a large number of transactions .In our data there were 38764 members ,then only 10 rules were considered as strong association and whole milk was almost common in each pair. So it is desirable to keep any offer (combo or discount) on whole milk by the shopping mall to get the best profit.

References

1. An Introduction to Statistical Leraning- By Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.

2. https://www.javatpoint.com/apriori-algorithm.

3. https://www.geeksforgeeks.org/apriori-algorithm/

4. https://hello.iitk.ac.in/course/mth552a22.