A STUDY ON MARKET BASKET ANALYSIS AND IT'S APPLICATION IN SALES DATA

PROJECT REPORT

submitted to the University of Calicut in a partial fulfillment of the requirement for the award of the degree of

MASTER OF SCIENCE IN **STATISTICS**

submitted by

THAHSEEN.T P

Reg. No. FKAVMST012

PG & RESEARCH DEPARTMENT OF STATISTICS

FAROOK COLLEGE (AUTONOMOUS)

KOZHIKODE

JULY 2023

PG AND RESEARCH DEPARTMENT OF STATISTICS
FAROOK COLLEGE (AUTONOMOUS)
FAROOK COLLEGE P.O, KOZHIKODE – 673632

CERTIFICATE

This is to certify that the project entitled **“A STUDY ON MARKET BASKET ANALYSIS AND ITS APPLICATION IN SALES DATA”** is bona-fide work done by **Ms. THAHSEEN.T P** during the year **2022-2023.** The project is submitted to the Department of Statistics, Farook College, for the partial fulfillment of the requirement for the award of the degree of Master of Science in Statistics.

Dr. MOHAMMED HISHAM M
(Project Guide)

DR. ABDUL RASHEED K V
(Head of the Department)

Name & Signature of Examiners:

1.

2.

Place: FAROOK COLLEGE

Date:

DECLARATION

I hereby declare that this project entitled **"A STUDY ON MARKET BASKET ANALYSIS AND ITS APPLICATION IN SALES DATA"** submitted to FAROOK COLLEGE (Autonomous), affiliated with the University of Calicut, in partial fulfillment of requirements for the award of the degree of Master in Statistics is a record of original work carried out by me under the supervision and guidance of **Dr. MOHAMMED HISHAM M**, Assistant Professor, Department of Statistics, Farook College (Autonomous).

Place:

Date:

Ms. THAHSEEN.T P
(Reg. No. FKAVMST012)

ACKNOWLEDGMENT

At the outside I record my profound thanks to Dr. K.A.Ayisha Swapna, Principal, Farook College and the management for giving me an opportunity to pursue M.Sc. degree in this prestigious institution and to undertake this project work.

I extend my sincere thanks to, Dr. Abdul Rasheed K V, HOD of Statistics Department and Dr. R. M. Juvairiyya, former HOD of Statistics Department who provided me with all facilities and necessary encouragement during the project.

I feel great pleasure in acknowledging my heartfelt thanks to my project coordinator Dr. Mohammed Hisham M for providing me with guidance and facilities for the project.

I also extend my sincere thanks to all other faculty members of Statistics Department, Farook College and my friends for their support and encouragement. Above all I bow my head before God Almighty whose blessings and Grace guided me throughout my life.

Ms. THAHSEEN. T P
(Reg. No. FKAVMST012)

CONTENTS

Chapter 1

INTRODUCTION

Market Basket Analysis (MBA) is the fundamental technique adopted by prominent marketers or retailers to understand the association between products in markets and thereby increase the cross-selling of products. It helps retailers to figure out the relationship between items more frequently bought together in markets. Market basket analysis is one of the best examples of unsupervised machine learning technique helping the retail industry largely. Thus, market basket analysis enables in identifying customer's interests and purchasing patterns, which results in increased sales performance. It works by analyzing customers' past purchasing behavior and finding which product combinations are frequently bought together by customers.

The process of market basket analysis identifies the association between different items that the customers pick into their shopping baskets. After finding the association between items, it would be helpful for marketers to adopt better marketing strategies. Nowadays market basket analysis is not only applied in grocery stores but also online retail. This analysis works based on the theory that if a customer buys a particular item, the customer is are more (or less) likely to buy another item. For example, if a customer buys bread, it is checked how probably he buys jam also from the supermarket. Typically, the relationship will be in the format of 'A' then 'B'.

The set of items a customer buys is called the itemset. The following rules can represent the relationship between two items A and B:

- The probability that a customer will buy product A is referred to as 'support of A'. The conditional probability of buying products A and B is called 'confidence'. The ratio of confidence to the expected confidence is called the 'lift'. The lift value shows how much association is between products A and B. The larger the lift, the more significant is the association.

- Implementing these rules in the analysis is termed 'association rule mining'. It is widely used to understand the customer's purchasing behaviour, thereby uplifting the firm's profit. Generally, under the association rule mining technique, we can find a set of rules like "if this, then that", which can control the firm's overall sales. The equation below represents the association mining rule

$$\{IF\}\ A\ \ =>\ \{THEN\}\ B$$

This is found as an if-then relationship. If the customer purchases item A, the chance of item B being bought together is calculated. Here A is called the antecedent and B the consequent. Antecedents are items primarily found in the basket, and consequents are those found in the basket after the antecedent.

$$\textbf{Support of A} = \frac{\mathbf{freq(A)}}{\mathbf{N}}$$

$$\textbf{Support of A \& B} = \frac{\mathbf{freq(A,B)}}{\mathbf{N}}$$

Thus, based on support values filtering out less frequently occurring item sets from the data is possible.

$$\textbf{Confidence} = \frac{\mathbf{freq(A,B)}}{\mathbf{freq(A)}}$$

It shows how frequently items A and B are bought together.

$$\textbf{Lift} = \frac{\mathbf{support(A,B)}}{\mathbf{support(A)\times support(B)}}$$

It indicates the strength of any association rule. The more the lift more is the association.

1.1 ADVANTAGES OF MARKET BASKET ANALYSIS

The following are some of the advantages of Market Basket Analysis.

- **Helps in setting prices:** Market basket analysis helps the retailers to identify which SKUs (stock keeping units) are more preferred among customers. For example, tea powder and milk are frequently bought together, so analysts assign a high probability of association for this combo. Without market basket analysis, retailers will decrease milk and tea powder stock intake. However, MBA can point out that whenever a customer buys tea powder, they tend to purchase milk as well. So, whenever the sale of milk and tea powder is expected to rise, retailers can increase their stock intake and customize their prices by understanding that this combination is inevitable for customers even if the price is increased.

- **Arranging SKU display:** A standard display format adopted by supermarket chains is the department system, where products are categorized and sorted by department. For instance, dairy products, body care products, hygienic products, snacks, cosmetics, cool drinks, etc., are correctly classified and displayed in different sections. Market basket analysis identifies those items that have a close affinity to each other, even if they are in different categories. With the help of this knowledge, retailers will place products with greater affinity close to each other in order to boost sales. For instance, if chips are placed near the beer bottle, customers may always buy both. In contrast, if they were placed in two extreme places, the customer would walk into the store to buy beer and leave the store without buying chips which would cause the least sales of chips.
- **Customizing promotions:** Marketers can study the purchase behaviour of individual customers to estimate with relative certainty what items they are more likely to purchase next. Today, many retailers use this idea to analyse the purchase behaviour of each individual accessing their websites. Such retailers can estimate with certainty what items the individual may purchase next or at a specific time. For instance, a customer who bought pizza would likely purchase pizza sauce on some weekends. So, retailers can customize offers to create 2 pizzas with one pack of pizza sauce at a discounted price in every weekend to boost their sales.
- **Identifying sales influencers:** A retail store contains a number of items that have some relationship with each other. In most cases, the sale of an item is driven by the increase or decrease in sales of other products. For example, consumers of tea in a particular area can impact the sale of sugar. Another example is if the production of French fries is lowered, the sale of mayonnaise will decrease, even though there is no relation between the products. In this way, retailers can understand the influence of such activities on customer behaviour and sales.
- **Recommendation engines:** Market basket analysis is an important principle that empowers recommendation engines. Online retailers mainly use this idea. Even you would have come across a recommendation while you are shopping online. A recommendation is applied by understanding the consumer behaviour and interests of each customer who access the websites, which may, in turn, increase sales.
- **Content placement:** It is crucial in e-commerce business to display or arrange the products in the correct order in order to increase the conversion rates. Online retailers

use market basket analysis to display the content that customers will likely read next. It will put the customers to be engaged in the website. Market basket analysis helps increase website traffic and get better conversion rates.

1.2 APPLICATIONS OF MARKET BASKET ANALYSIS

The following are some of the applications of Market Basket Analysis.

- **Cross-selling:** Cross-selling can be defined differently depending on the factors involved. In simpler words, cross-selling is when a marketer suggests complementary goods to customers with the primary product already purchased. MBA helps retailers to understand customer purchasing patterns and then enhance cross-selling of products.
- **Customer behaviour:** Customer behaviour, in general, is a study of how an individual or a group of customers purchase items to satisfy their wants and needs. So, market basket analysis helps to understand customer behaviours. It provides an insight into customers' purchasing mode and helps retailers identify the association between two products a customer tends to purchase.
- **Fraud detection:** Market basket analysis is also used to detect fraudulent activities. It is possible to identify fraudulent transactions by understanding customer purchasing patterns under different conditions. For instance, data that contains the usage of the credit card can help in detecting fraud which is associated with the purchase behaviour.
- **Affinity promotion:** Affinity promotion is the method of discovering the co-occurrence relationships among the items purchased by specific individuals or groups using data analytics. By understanding customer purchasing patterns, retailers can decide on cross-selling, loyalty programs, and store layouts and prepare discount plans.
- **Bioinformatics / Pharmaceutical:** Market basket analysis helps understand the co-occurrence among the pharmaceutically active ingredients and the diagnosis prescribed to different patients.
- **Telecom:** The increasing attention to customer service of telecom companies is eased with the help of market basket analysis. For instance, telecom companies have started to deliver television and internet packages together, apart from other discounted online services, to eliminate churn.
- **Medicine:** Market basket analysis was used to determine symptom analysis and comorbid conditions in the medical field. It also helps determine the hereditary traits and genes almost associated with the local environmental effects.

1.3 OBJECTIVE OF THE STUDY

The objectives of the study are as follows:

1. To study the basic concepts of Market Basket Analysis by going through its methodologies and applied tools.
2. To apply the concept of Market Basket analysis on a real-life dataset which is the sales data of a bakery located in Edinburg with the help of Python software.
3. To understand the most frequently sold item among all the items in the bakery and obtaining the percentage of its sales from the dataset.
4. To know most frequent combinations of items which are more likely to be purchased by customers and applies appropriate business strategies for promoting them.
5. To see which months and days in a week are more productive with higher sales and also checks that, in which part of the day the bakery seems to be more occupied.

Chapter 2

METHEDOLOGY

Market basket analysis is a business intelligence technique used to predict future purchasing behaviour of customers. It observes and analyses the purchases that commonly happens together and thereby makes decisions and strategies which increases sales and profit. MBA is one of the main applications of machine learning. By knowing which products customers frequently buy together enables the merchants to organise their stores and websites in a beneficial manner. This method of MBA is conducted by looking into prior purchases of customers. Marketers use this as a cross-selling tool for enhancing the profit. MBA is not only employed in retail industry but also in detecting false credit card transaction and insurance claims.

2.1 BASIC CONCEPTS OF ASSOCIATION RULE MINING

Association rule mining deals with the use of machine learning models to analyse the data for patterns or co-occurrences, in a database. It involves finding if-then associations which themselves are the association rules. An association has two parts:

- Antecedent (IF)
- Consequent (THEN)

Applications of association mining is widely seen in marketing, market basket analysis in retailing, clustering and classification. In market basket analysis association rule counts the frequency of items that occur together, seeking to find association that occur far more often than expected. It can tell you what products do customers frequently bought together by generating a set of rules called association rules. Retailers can use those rules for various marketing strategies:

- Customer behaviour analysis
- Catalogue design

- Changing the store layout according to the trends
- Identifies the trending products among customers
- Cross marketing on online store
- Customised emails with add on sales

Association mining rules are widely used to analyse retail basket or transaction data, and are trying to identify strong rules found in transaction data using its methodologies and metrics. There are 3 key metrics to consider in association rule mining.

The main 3 components in A-PRIORI algorithm are:

- Support
- Confidence
- Lift

Support:

This is the percentage of the item contained in the dataset. The minimum support threshold required by apriori can be set based on knowing the domain. Support is calculated by dividing number of transactions of particular item by total number of transactions.

Confidence:

If two items A and B are given, confidence measures the percentage of items that item B is purchased, given that item A is already purchased. This is expressed as:

Confidence[A->B] =support [A, B]/support [A]

Confidence values from range 0 to 1. Where 0 indicates that B is never purchased when B is bought, and 1 indicates that B is purchased whenever A is purchased. Note that the measuring confidence is directional. That is, we can also compute the percentage of times that item A is purchased, given that item Bis already purchased.

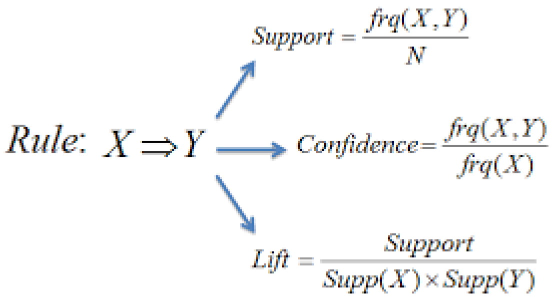
Confidence[A->B] =support [A, B]/support [B]

Lift:

Lift is calculated as the ratio of confidence and support. Unlike the confidence matric lift has no direction. This means that lift{A->B} is equal to the lift{B->A}.

lift{A->B} = lift{B->A} = support {A, B} / (support {A} *support {B})

Lift value that is equal to one implies that there is no relationship between A and B. That is A and B occur together only by chance. Lift value greater than one implies that there is a positive relationship between A and B. That is A and B occur together more often than random. Lift less than one implies that there is a negative relationship between A and B.


$Rule: X \Rightarrow Y$
$Support = \frac{frq(X,Y)}{N}$
$Confidence = \frac{frq(X,Y)}{frq(X)}$
$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

Support is an indication of how frequently the product appears in dataset and Confidence is the indication of how often the rule has been found to be true. Lift measures how likely a product is purchased when another product was purchased, while controlling for how popular both items are.

Rule generation is the primary and foremost task in the mining of frequent patterns in retail industry. An association rule is an implication expression of the form X→Y, where X and Y are disjoint items. The main task of the rule is to evaluate the "interest" of such an association rule. Suppose item X is being picked by the customer, then the chance of item Y also being picked by the customer under same transaction ID is found out. For this purpose, we use the matrices support, confidence and lift. The ranges of these matrices are given below,

Support→ [0,1]

Confidence→ [0,1]

Lift→ [0, ∞)

An association rule is considered of having the required "interest", if it satisfies both minimum support threshold and minimum confidence threshold. So, both these matrices measure how interesting the rule is. The thresholds set by the client help to compare the rule strength according to the client's will. The closer to the threshold the more is the rule powerful.

Applications of Association rule mining beside MBA

- **Medical Diagnosis:**

Association rule mining in medical diagnosis can be useful for assisting physicians for curing patients. Diagnosing diseases is not an easy process and has a scope of errors which may result in creating unreliable results. Using relational association rule mining, physicians can identify the probability of occurrence of disease concerning various factors and symptoms. Using these learning techniques, inferences can be extended by adding new symptoms and defining relationships between the signs of the new diseases.

- **Census Data:**

  Every government has a huge volume of census data. This data can be used to plan efficient public services (food, education, health, transportation, etc) as well as help public businesses (for building new factories, shopping malls, and even marketing some particular products). This application of association rule mining and data mining has great potential in supporting sound public policy and bringing an efficient functioning of democratic society.

- **Protein Sequences:**

  Protein sequences are made up of 20 types of amino acids. Each protein present in the protein sequence bears a unique 3D structure which depends upon the sequence of these amino acids. A slight change in the protein sequence can cause a change in structure which might change the functioning of the protein. This dependency of the protein functioning on amino acid sequence has become a subject of great researches. Earlier it was believed that these sequences are random, but now found that they aren’t. Scientists deciphered the nature of associations between different amino acids present in a protein. Knowledge and understanding these association rule became extremely powerful during the synthesis of artificial proteins.

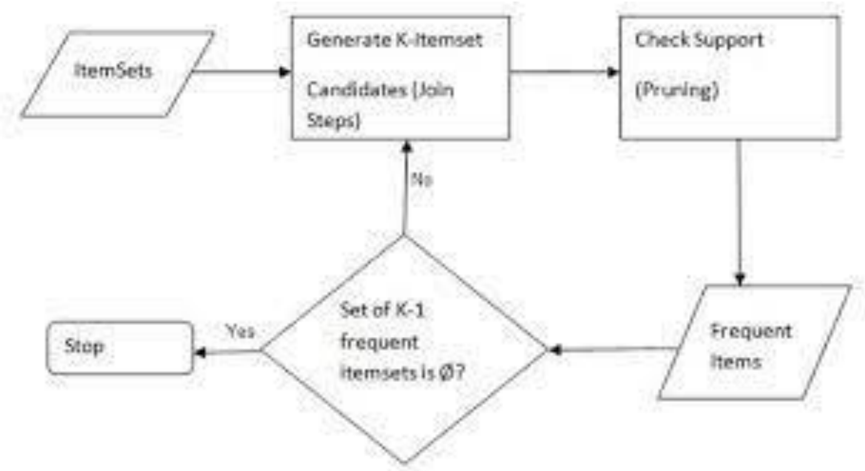
2.2 A-PRIORI ALGORITHM AND ITS USAGE IN MBA

The apriori algorithm is one of the most popular algorithms used in association rule mining over relational database. It identifies the products in a dataset and further extends them to larger and larger itemset. The apriori algorithm introduced by R Agarwal and R Srikanth in 1994 has a great significance in data mining. Apriori algorithms are used to generate association rules using frequent dataset. It is based on the concept that a subset of a frequent itemset will also be a frequent itemset, where frequent itemset is an itemset whose support value is greater than threshold value.

The apriori algorithm proposes that the probability of an itemset is not frequent if the probability of the itemset is less than minimum support threshold and if any subset within the itemset has value less than minimum support threshold. One of the characteristics of apriori algorithm is the “anti-monotone property”. That is, for an instance, if the probability of buying sandwich is less than the minimum support threshold, then probability of buying both sandwich and coffee will definitely be less than minimum support threshold as well.

The apriori algorithm uses mainly two steps called “join” and “prune” and are used to reduce the search space. It is an iterative approach to discover the most frequent itemset. Steps for apriori algorithm in data mining are following:

1. In join step a (K+1) itemset is generated from K-itemset by joining each item with itself.
2. In prune step count of each item in the database is scanned. Any candidate itemset that does not meet the minimum support requirement is regarded as infrequent and removed. This step is used for reducing the size of candidate itemset. It iterates both of the steps until there is no further possible extensions left.


ItemSets
Generate K-Itemset Candidates (Join Steps)
Check Support (Pruning)
No
Stop
Yes
Set of K-1 frequent itemsets is Ø?
Frequent Items

Steps in Apriori

Apriori algorithm is a sequence of steps to be executed in order to find out the most frequent itemset in the given database. This data mining method follows both the join and prune step iteratively until the most frequent itemset is found. A minimum support threshold is given in the problem or it is decided by the analyst. The key steps in performing apriori algorithm are the following:

1) In the first iteration of the apriori algorithm each item is taken as a one itemset candidate. Then the algorithm counts the occurrences of each item.
2) Let there be some fixed minimum support. The set of one itemset whose occurrence is satisfying the minimum support threshold are determined. Only those candidates which have support value greater than or equal to minimum support threshold are chosen for the next iteration and others are pruned.
3) In this step, the two itemset with minimum support threshold are found. For this step, in the join step the two itemset is generated by forming a group of two items by combining items with each one.
4) The two itemset candidates are pruned using the minimum support threshold value. Now the table will have the two itemset with minimum support only. Then move on to the next iteration.
5) In this iteration, the three itemset are formed using join and prune step. Then look for their support value by examining the support of all its possible subsets, that is two itemset. If all the two itemset subsets are meet by the minimum support threshold value then the corresponding superset is regarded as frequent and it is selected and otherwise pruned.
6) In this step, 4 itemset is formed by joining 3 itemset with itself and pruning if its subset does not meet the minimum support criteria. The algorithm is stopped when the most frequent itemset are achieved.

2.3 THE SALES DATA EXPLORATION

In this dissertation, we are discussing about the influence of market basket analysis among huge retailers and also applying this technique for analysing and interpreting the sales of a ‘Bakery sales data’ which is extracted from Kaggle. The dataset belongs to “The bread basket”, a bakery located in Edinburg. The data provides the transaction details of customers who ordered different items from this bakery throughout during a 2 year of time period. The dataset has 20507 entries over 9000 transactions and 4 columns.

The only numeric variable used in the data is transaction number, where 9465 distinct values are present. And all other variables are categorical. They are ‘items’ in 94 distinct values, ‘datetime’ in 9182 distinct values, ‘daypart’ in 4 distinct values and ‘day type’ in 2 distinct values. ‘Transaction number’ is a unique identifier for each transaction, ‘items’ are the items purchased, ‘datetime’ is the date and time for each transaction, ‘daypart’ is the part of the day when a transaction is done and ‘day type’ classifies whether a transaction is made in weekend or weekdays.

The dataset is ideal for anyone trying to practice association rule mining and market basket analysis, which helps to understand the business context of data mining for understanding the buying patterns of customers. The method of market basket analysis is executed on this data using **Python** software and the required inferences are noted in the third chapter.

Chapter 3

ANALYSIS AND FINDINGS

In this chapter we will see the analysis of the dataset mentioned previously. All the analysis are conducted for reaching in a result which can be used for enhancing the daily sales of the bakery. The technique of association rule mining is implemented by finding the values of its three matrices such as support, confidence and lift. The support value is used for finding the most frequent item from the entire itemset. Confidence is used for understanding the strongest associations and lift is used as a tool for measuring the strength of each association. The associations are extracted from the most frequent associations by setting the minimum lift threshold equal to 1. Also, finding the most productive months and peak hours of sales by observing the number of transactions during different time periods.

3.1 DATA FEATURES

In this chapter, we are applying the ideas and methodologies of market basket analysis for analyzing and studying the sales characteristics of a ‘Bakery sales data’ which is extracted from Kaggle. The dataset belongs to “The bread basket”, a bakery located in Edinburg. The market basket analysis is executed with the help of Python software. The data provides the transaction details of all customers who purchased different items from this bakery over 2 years in between 2016 and 2017.

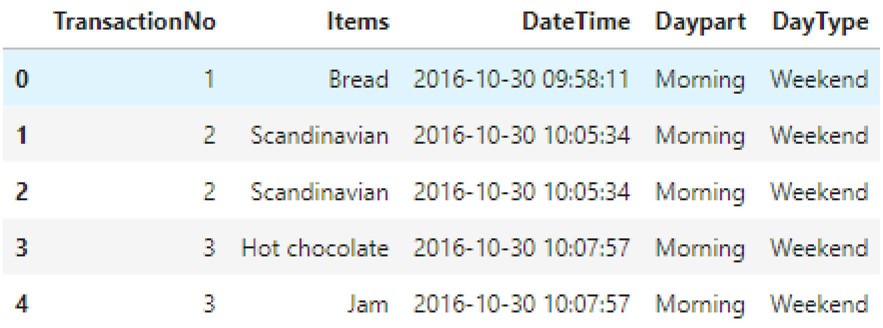
The dataset contains 20507 entries over 9000 transactions and the variables are located in 5 columns. The variables included are transaction number, item, datetime, daypart and day type. Among them, transaction number is the only numeric variable and all others are categorical. There are 94 distinct items in the data. Daypart is classified into 4, morning, afternoon, evening and night. Also, there are 2 types of day type, weekend or weekdays. Exactly, there is 9465 total transactions.

The dataset is ideal and useful for those who looking for practicing market basket analysis and also association rule mining. This data and the analysis on it is highly helpful for understanding the business context of data mining and association rule generation for predicting the purchasing patterns of the customers.

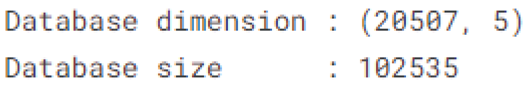
3.2 THE ANALYSIS AND OUTPUTS

The market basket analysis on the ‘bakery sales data’ is performed using the python software. The following represents the executed steps and corresponding outputs required for reaching at a crucial conclusion about the sales of the mentioned bakery.

The data from the excel file is Imported to the python using the necessary code. Also, the dimension and size of the database is found.

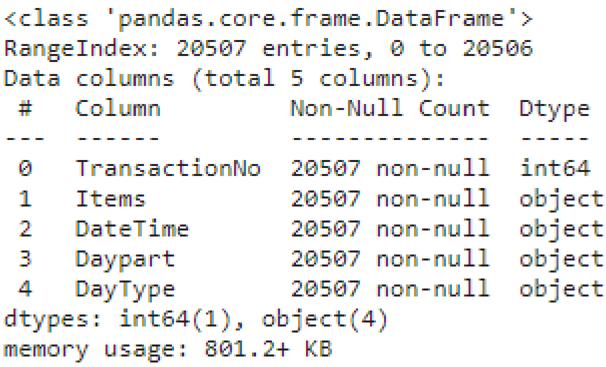


	TransactionNo	Items	DateTime	Daypart	DayType
0	1	Bread	2016-10-30 09:58:11	Morning	Weekend
1	2	Scandinavian	2016-10-30 10:05:34	Morning	Weekend
2	2	Scandinavian	2016-10-30 10:05:34	Morning	Weekend
3	3	Hot chocolate	2016-10-30 10:07:57	Morning	Weekend
4	3	Jam	2016-10-30 10:07:57	Morning	Weekend



```
Database dimension : (20507, 5)
Database size      : 102535
```

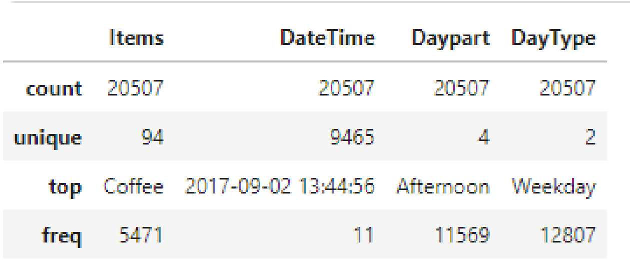
The overall data summary is found using appropriate codes. Here all the information about the variables is checked. The table below gives the details of the variables present in the database.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20507 entries, 0 to 20506
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   TransactionNo  20507 non-null  int64 
 1   Items          20507 non-null  object
 2   DateTime       20507 non-null  object
 3   Daypart        20507 non-null  object
 4   DayType        20507 non-null  object
dtypes: int64(1), object(4)
memory usage: 801.2+ KB
```

In next step, the number of transactions is found. The output for the code gives there are total 9465 transactions occurred during the time period

Finally, the information regarding all the variables, their count and frequency are found and output includes a brief table. The table also contains the one item which has the highest frequency, that is ‘coffee’. Also checks which date &time, daypart, and day type has the highest frequency and they are included in the same table.

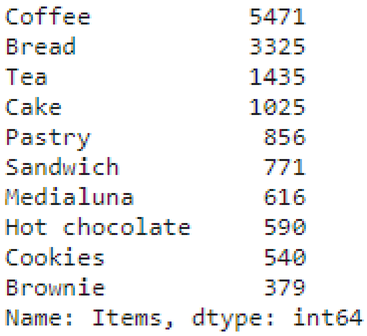


	Items	DateTime	Daypart	DayType
count	20507	20507	20507	20507
unique	94	9465	4	2
top	Coffee	2017-09-02 13:44:56	Afternoon	Weekday
freq	5471	11	11569	12807

The dataset provides the transaction details of all the items purchased by customers from the bakery in between 2016 and 2017. The dataset has 20507 entries over 9465 transactions and 5 columns. There is total 5 variables present in the data, which are ‘transaction number, item, date time, daypart and day type’. Among them transaction number only is the numeric variable and all others are categorical variables. There are no missing cells present in the dataset. 94 different items are present in the dataset. Daypart is classified into 4 and day type into 2, weekend and weekdays.

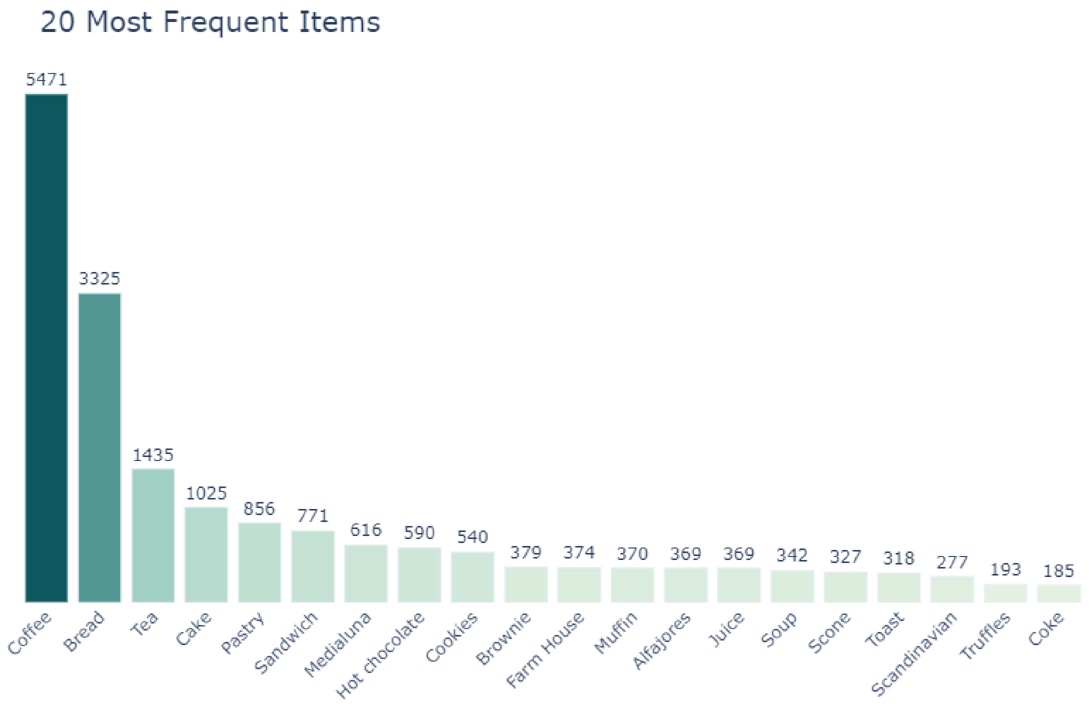
Data Exploration and Visualization

In this step we understood that coffee is the most frequent item among all items in the dataset. So, we need to find the further items having the high frequencies after coffee. The output for the code needed for checking this provides first 10 items with higher frequencies.



```
Coffee           5471
Bread            3325
Tea              1435
Cake             1025
Pastry            856
Sandwich          771
Medialuna         616
Hot chocolate     590
Cookies           540
Brownie           379
Name: Items, dtype: int64
```

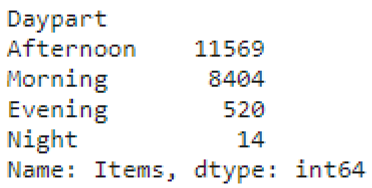
A bar graph is plotted below showing 20 items with larger frequencies along with their frequencies in a descending order of magnitude.



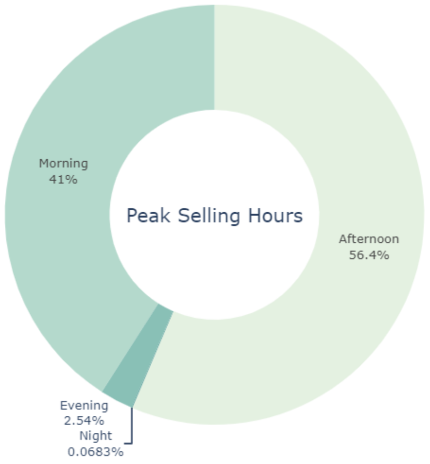
From this graph also we can conclude that coffee is the best-selling product, followed by bread and tea.

Looking for peak daypart of sales

In this step, the daypart (afternoon, morning, evening, night) in which higher sales occurred is analysed. Also, a graph is plotted using the results.



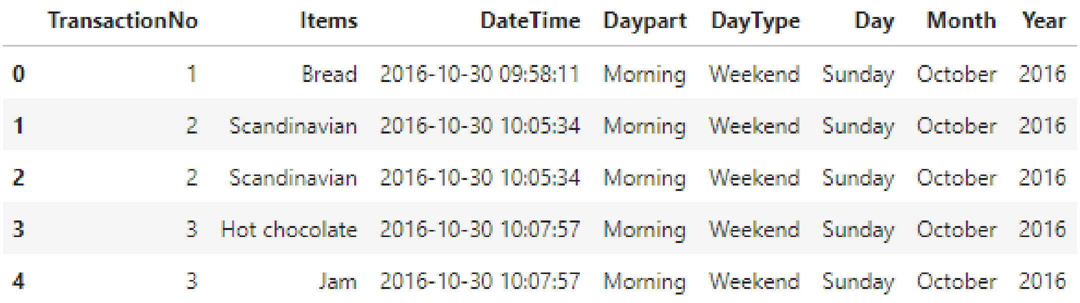
```
Daypart
Afternoon    11569
Morning       8404
Evening        520
Night           14
Name: Items, dtype: int64
```



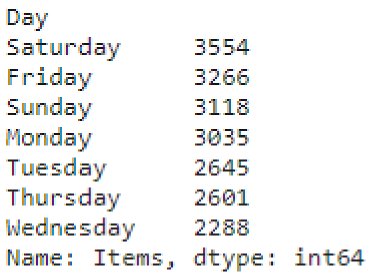
From this result, it is understood that the bakery is making a large part of its sales in the afternoon everyday with over 56% of the total sales. After that in evening and night sales decreases. However, the bakery makes a decent part of its sales in morning over 41%.

Looking for monthly and weekly sales

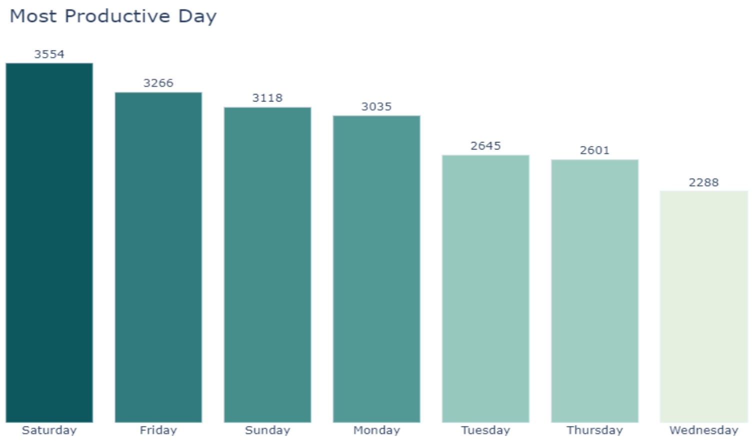
For this analysis, the extraction of months and days from the dataset is done. The number of transactions happened in each day in a week is discovered and a bar graph is plotted with the 7 days in a week and corresponding number of transactions. The results are shown below.



	TransactionNo	Items	DateTime	Daypart	DayType	Day	Month	Year
0	1	Bread	2016-10-30 09:58:11	Morning	Weekend	Sunday	October	2016
1	2	Scandinavian	2016-10-30 10:05:34	Morning	Weekend	Sunday	October	2016
2	2	Scandinavian	2016-10-30 10:05:34	Morning	Weekend	Sunday	October	2016
3	3	Hot chocolate	2016-10-30 10:07:57	Morning	Weekend	Sunday	October	2016
4	3	Jam	2016-10-30 10:07:57	Morning	Weekend	Sunday	October	2016

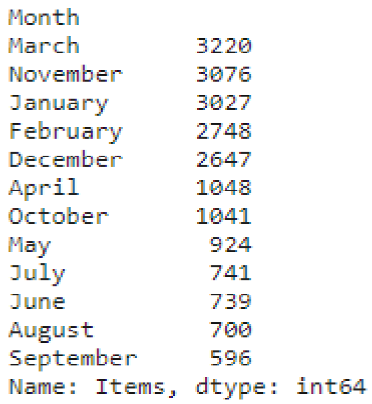


```
Day
Saturday     3554
Friday       3266
Sunday       3118
Monday       3035
Tuesday      2645
Thursday     2601
Wednesday    2288
Name: Items, dtype: int64
```



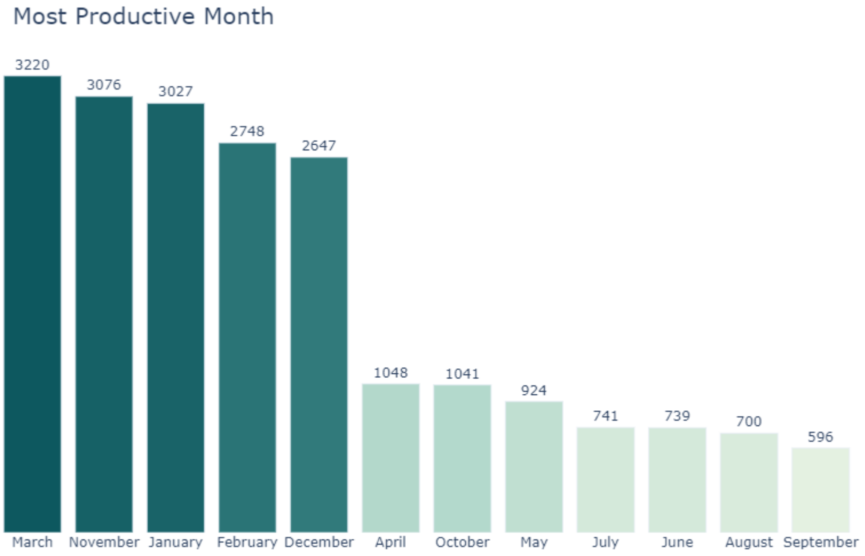
From this graph, it is clear that the sales are huge during the weekends, that is in Friday and Saturday. Also, the sales seem to be uniform in other days of the week.

The sales count is checked in each month of the year and a bar graph is plotted showing the results.



```
Month
March        3220
November     3076
January      3027
February     2748
December     2647
April        1048
October      1041
May           924
July          741
June          739
August        700
September     596
Name: Items, dtype: int64
```

The sales count in each month of the year

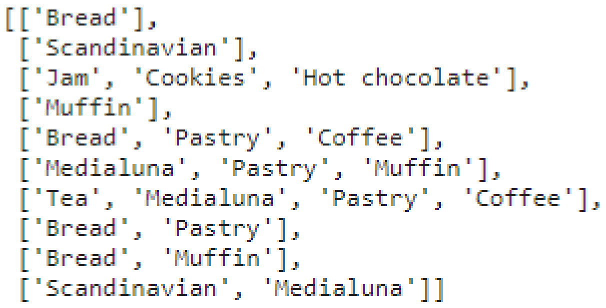


For some obvious reasons the bakery seems to have higher sales from November to March. The rest of the months shows a lowered and uniform sales.

Coffee is the best seller product, followed by bread and tea respectively. The bakery is making most of its sales in afternoon everyday with over a 56% of the total sales. Sales decreases after that in evening and night. But there is a decent volume of sales in morning. For some reasons, bakery has higher sales during weekends. The bakery seems to have most of its sales during November to December, and makes most of their business during these months.

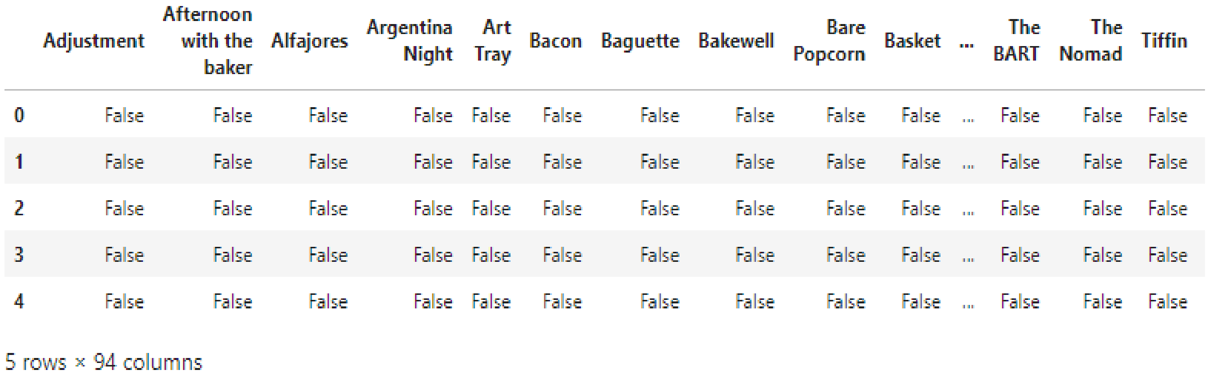
Association rules generation

Apriori algorithm requires a data frame in such a way that all the transactions are one hot encoded for all the items. For this, list of all transactions is made and it is given below.



```
[['Bread'],
 ['Scandinavian'],
 ['Jam', 'Cookies', 'Hot chocolate'],
 ['Muffin'],
 ['Bread', 'Pastry', 'Coffee'],
 ['Medialuna', 'Pastry', 'Muffin'],
 ['Tea', 'Medialuna', 'Pastry', 'Coffee'],
 ['Bread', 'Pastry'],
 ['Bread', 'Muffin'],
 ['Scandinavian', 'Medialuna']]
```

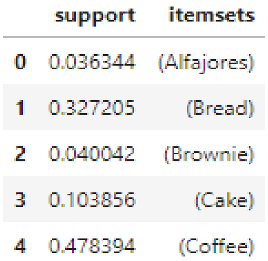
One hot encoding:



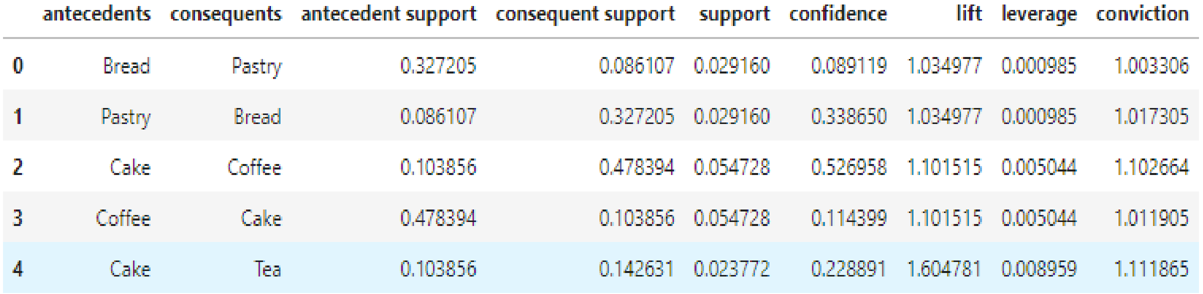
	Adjustment	Afternoon with the baker	Alfajores	Argentina Night	Art Tray	Bacon	Baguette	Bakewell	Bare Popcorn	Basket	...	The BART	The Nomad	Tiffin
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False

5 rows × 94 columns

Support value of each item is found and the association rules are generated using the necessary codes. Here, the association rules are generated by extracting the combinations with minimum lift threshold equal to 1.



	support	itemsets
0	0.036344	(Alfajores)
1	0.327205	(Bread)
2	0.040042	(Brownie)
3	0.103856	(Cake)
4	0.478394	(Coffee)

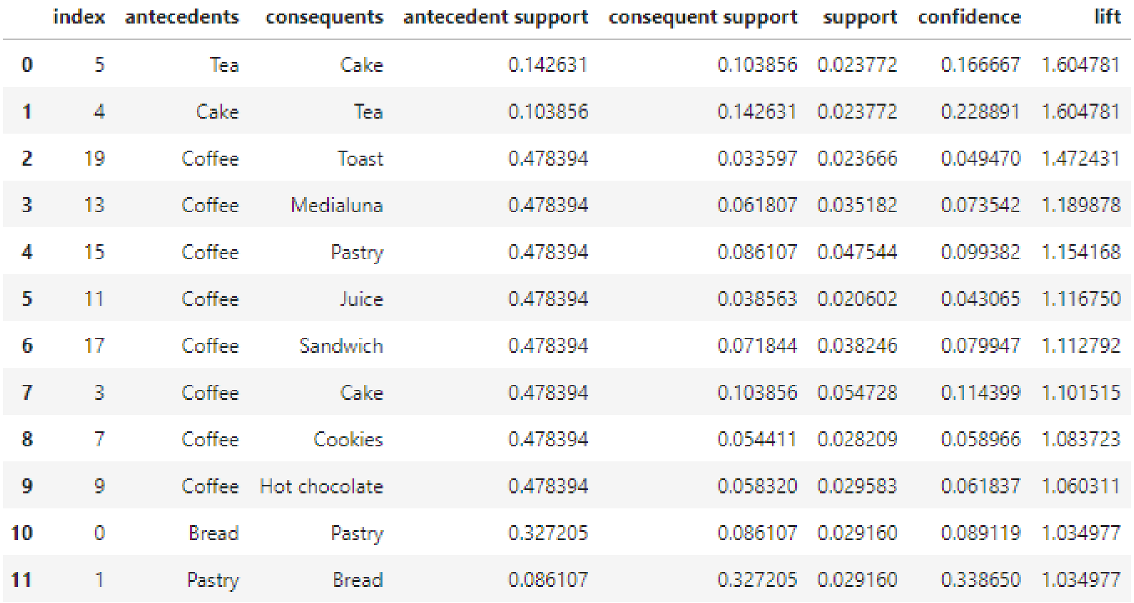


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	Bread	Pastry	0.327205	0.086107	0.029160	0.089119	1.034977	0.000985	1.003306
1	Pastry	Bread	0.086107	0.327205	0.029160	0.338650	1.034977	0.000985	1.017305
2	Cake	Coffee	0.103856	0.478394	0.054728	0.526958	1.101515	0.005044	1.102664
3	Coffee	Cake	0.478394	0.103856	0.054728	0.114399	1.101515	0.005044	1.011905
4	Cake	Tea	0.103856	0.142631	0.023772	0.228891	1.604781	0.008959	1.111865

The support, confidence and lift were found for each most frequent antecedent-consequent pairs. The mostly occurred combinations with lift threshold values equals to 1 are bread-pastry, cake-coffee and cake-tea. By observing the lift value for each pair, we can say that the cake-tea bond is stronger than others.

Refining Rules

The confidence for the most frequent consequent is always very high even if there is a weak association. Here, coffee is the most frequent item and the best seller item. Therefore it can be recommented with every other item. So, we can generate rules by setting coffee in the consequent part inorder to get a clearer unknown rules generated from the data.



	index	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	5	Tea	Cake	0.142631	0.103856	0.023772	0.166667	1.604781
1	4	Cake	Tea	0.103856	0.142631	0.023772	0.228891	1.604781
2	19	Coffee	Toast	0.478394	0.033597	0.023666	0.049470	1.472431
3	13	Coffee	Medialuna	0.478394	0.061807	0.035182	0.073542	1.189878
4	15	Coffee	Pastry	0.478394	0.086107	0.047544	0.099382	1.154168
5	11	Coffee	Juice	0.478394	0.038563	0.020602	0.043065	1.116750
6	17	Coffee	Sandwich	0.478394	0.071844	0.038246	0.079947	1.112792
7	3	Coffee	Cake	0.478394	0.103856	0.054728	0.114399	1.101515
8	7	Coffee	Cookies	0.478394	0.054411	0.028209	0.058966	1.083723
9	9	Coffee	Hot chocolate	0.478394	0.058320	0.029583	0.061837	1.060311
10	0	Bread	Pastry	0.327205	0.086107	0.029160	0.089119	1.034977
11	1	Pastry	Bread	0.086107	0.327205	0.029160	0.338650	1.034977

Since, coffee is the best seller of this bakery, it is showing associations with 8 other items. By analyzing the confidence values for each pair, we can interpret the following statements. Among the tea consumers, 16% of them buys cake also. Also 22% of cake buyers will buy tea also. Over 11% of coffee lovers also buy cake. Among the pastry lovers, over 33% of them buys bread. We can conclude that all the above associations are stronger, since their lift values are greater than 1.

3.3 RESULTS & INTERPRETATIONS

Looking into the insights of above outputs, there are some key information about the sales of the bakery over 2 years. From those results we can adopt some crucial business strategies for boosting the sales of the firm.

Coffee is the frequent item and the best-seller in the bakery why so it shows a strong or moderately strong association with 8 other frequent items. The most powerful association is given by the tea-cake pair. Over 16% of the total sales, the customers who bought tea also buys cake and in reverse order it is 22%. Over 11% of coffee lovers also bought cake with it.

There is a number of business strategies that the bakery can adopt if it is practical, in order to boost the daily sales by considering all above results. We have seen a strong association of 8 items with coffee. So, this information can be used for setting promotional discounts and combo offers for coffee and any another item. Also, tea-cake and pastry-bread associations also can be promoted. Arranging placements of items close to coffee in ordering counters and menu chart can be adopted as a good strategy for boosting the sales.

Coffee is the best seller product, followed by bread and tea respectively. The bakery is making most of its sales in afternoon everyday with over a 56% of the total sales. Sales decreases after that in evening and night. But there is a decent volume of sales in morning. For some reasons, bakery has higher sales during weekends. The bakery seems to have most of its sales during November to December, and makes most of their business during these months.

Chapter 4

SUMMARY AND CONCLUSIONS

In this work, we have explained the basic concepts involved in Market Basket Analysis.

A retailer is highly interested in understanding what product is the best seller and the frequent product combinations by which they can adopt strategies for enhancing cross-selling activities. Market basket analysis is a machine learning tool implemented for this purpose, boosting the firm's sales by applying appropriate strategies.

In this dissertation, a detailed analysis of sales data of a bakery located in Edinburg is done using the methodologies of market basket analysis. The analysis is executed in Python software. The important methodology used is association rule mining. The three matrices used in association rule mining are support, confidence, and lift. All these are used for measuring associations. The main algorithm used in market basket analysis is the apriori algorithm. The important results obtained from the analysis of the dataset are the following:

- Coffee is the best seller item, and it shows associations with eight other items.
- The tea-cake combination is the most powerful association of all. Offers and promotional discounts can be applied to these items to boost sales. Rearranging the ordering counter or display shelves based on the most frequent combination is also a good strategy for enhancing sales and, thereby, the bakery's profit.
- In analyzing on time of occurrence of sales, the months from November to March are the most productive. We obtained that the weekends are more occupied with the customers when compared to weekdays. Also, over 56% of total sales is happened in the afternoons and 41% in the morning. Then sales gradually lowered after that in the evenings and nights due to some unknown reasons.

In essence, we got an intellectual summary of the bakery data's sales characteristics using the market basket analysis technique, and the obtained results can be used for planning necessary business strategies.

4.1 FUTURE SCOPE AND LIMITATIONS OF THE STUDY

Some of the future scopes of the study are –

- The study could be extended to explore and incorporate advanced techniques such as sequential pattern mining, association rule mining with constraints, or graph-based methods to uncover more complex and nuanced patterns in customer purchasing behavior.
- The future scope of the study could involve incorporating additional variables such as customer demographics, loyalty program data, or product attributes to gain deeper insights into the factors influencing customer preferences and purchasing decisions.
- Conducting a comparative analysis of different Market Basket Analysis algorithms or approaches can help evaluate their performance and identify the most effective method for extracting valuable insights from transactional data.
- Building predictive models based on Market Basket Analysis results could be explored to anticipate customer behavior, recommend personalized product offerings, and optimize inventory management.

The following are some of the limitations of the study -

- The findings of the study may be specific to the analyzed dataset, which might limit the generalizability of the results to other contexts or industries. Replicating the study with diverse datasets and domains could provide more robust and widely applicable insights.
- Market Basket Analysis primarily focuses on transactional data without considering contextual information such as customer motivations, external factors, or specific events. Incorporating contextual data could provide a more comprehensive understanding of customer behavior and purchasing patterns.
- While Market Basket Analysis identifies associations between items, it does not necessarily establish causality. Further research may be required to validate and understand the underlying reasons behind observed associations.

REFERENCES

1. Gupta, S., & Mamtora, R. (2014). A survey on association rule mining in market basket analysis. *International Journal of Information and Computation Technology*, *4*(4), 409-414.
2. Kurniawan, F., Umayah, B., Hammad, J., Nugroho, S. M. S., & Hariadi, M. (2018). Market Basket Analysis to identify customer behaviours by way of transaction data. *Knowledge Engineering and Data Science*, *1*(1), 20.
3. Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. *Procedia computer science*, *85*, 78-85.
4. Idris, A. I., Sampetoding, E. A., Ardhana, V. Y. P., Maritsa, I., Sakri, A., Ruslan, H., & Manapa, E. S. (2022). Comparison of Apriori, Apriori-TID and FP-Growth Algorithms in Market Basket Analysis at Grocery Stores. *The IJICS (International Journal of Informatics and Computer Science)*, *6*(2), 107-112.
5. Yabing, J. (2013). Research of an improved apriori algorithm in data mining association rules. *International Journal of Computer and Communication Engineering*, *2*(1), 25.