SEMINAR REPORT

ON

Market basket analysis in Amazon

BY

Payal Mehta
T. Y. B.Tech. Computer Engineering,
Roll No -3936 , Division - C

Year:- 2020-2021

SEMINAR GUIDE
Priya Deokar

MKSSS's
Cummins College of Engineering for Women, Pune
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

CERTIFICATE

This is to certify that Ms Payal Mehta

has satisfactorily completed the seminar on

" Market Basket Analysis in Amazon"

in the partial fulfillment of her term-work (Seminar) as a part of syllabus for T.Y.B.Tech. Computer Engineering for the Academic Year 2020-2021 as prescribed by MKSSS's Cummins College of Engineering for Women, Pune (An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Prof. Priya Deokar
Seminar Guide

Dr. Mrs.S. M. Kelkar
HOD Computer

Dr. Mrs. M. B. Khambete
Principal

MKSSS's
Cummins College of Engineering for Women, Pune
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

PLAGIARISM CERTIFICATE

This is to certify that Payal Mangesh Mehta
has satisfactorily completed the seminar report on

" Market basket Analysis in Amazon"

and the data present in this report is found to be 97 % plagiarism free.

This seminar report is towards the partial fulfillment of her term-work (Seminar) as a part of the syllabus for T.Y.B.Tech. Computer Engineering for the Academic Year 2020-2021 as prescribed by MKSSS's Cummins College of Engineering for Women, Pune (An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Name of the Student and sign **Prof. Priya Deokar**
Seminar Guide

ACKNOWLEDGMENT

I wish to express my sincere gratitude to the respected Principal, Dr. Mrs. Madhuri Khambete and Dr. Mrs. Supriya Kelkar, Head of Department of Computer Engineering, Cummins College of Engineering for Women for providing me an opportunity to present a seminar on “ Market basket analysis in Amazon ”.

I thank my project guide, Priya Deokar from the Computer Department of Cummins College, for her support and encouragement at every step for carrying out this seminar.

Last but not the least I wish to avail myself of this opportunity to express a sense of gratitude to my friends and my parents for their support and strength.

Payal Mehta

Roll No. : 3936

TE Computer C

Dept.

ABSTRACT

Market basket Analysis is a data mining technique which analyzes customer purchasing habits by finding associations between the products purchased. This technique is commonly used in all e-commerce websites such as Amazon, Flipkart, Myntra, etc. Using Apriori Algorithm and association rules, one can find the relationship between one item purchased with another.

For example, when we buy any smartphone on Amazon , it gives a recommendation of screen guard or phone covers in the frequently bought section.

Here, Market basket Analysis comes into picture. Market basket analysis works on if…..then condition i.e if an item A is bought , then item B is likely to be purchased.

The question arises what is the difference between recommendation and association. When we purchase any product on Amazon , there are two categories on the product catalog page . One is "frequently bought together" and "you might also like". The first one is association and the latter is recommendation.

Amazon is using this technique to analyse customer's interests according to age, gender and location. It is the key factor to increase the productivity of the business and engage customers to explore more products on Amazon.

TABLE OF CONTENTS

CHAPTER 0
PREREQUISITES : DATA MINING

0.1 Introduction

Data mining is a technique in machine learning to extract and find different trends in large datasets. In short, it is the way of identifying hidden patterns of information to categorize it into useful data that is collected and used in particular areas such as efficient analysis, data mining algorithms, data warehouses that help in decision making and other data requirements by organizations.

It aims to cost-cutting and increase the revenue in production. It uses algorithms for datasets and finds the probability of future events. Data Mining, in other words, is also called Knowledge Discovery of Data (KDD). It is used by large business organizations to extract information from complex datasets to solve business problems.

Data mining processes includes :

1. text mining,
2. web mining,
3. audio and video mining,
4. pictorial data mining,
5. and social media mining.

It is done through highly specialized software that is simple. By using data mining, all business problems are solved faster with low operation costs. There are various organizations that use new technologies to collect data that makes it impossible to find trends in it manually. When there is a large amount of information available on various platforms and very little knowledge is accessible, the biggest challenge is to analyze the datasets and extract necessary information from it that can be used to solve a business problem for firm growth.

Various powerful instruments and techniques are available to mine data and find the trends from it.

0.2 Techniques

There are various techniques involved in Data Mining :

1] Classification

Classification is a technique to predict the class of a new item from a given dataset. It is referred to as a supervised learning technique because the dataset, which is divided into training and testing data, is used to learn the structure of the classes to which it belongs.

It includes data analysis, categorization of target data, document categorization, etc.

2] Association Rule Learning

Association rule learning is an unsupervised data mining technique which contains various datasets. In short, it is used to find the relationship between 2 or more items. It is based on the "if….. then" rule , the first one is called antecedent and the latter is consequent.

The well known association rule learning algorithm is the APRIORI algorithm, it is generally used to solve business problems in an organization.

3] Clustering Analysis

Clustering analysis is a technique to cluster or group the datasets that are similar . The datasets are grouped to understand the hidden structure and trends in data, knowledge discovery of data, etc.

The popular clustering algorithms include Expectation-Maximisation(EM), k-means clustering, fuzzy C-means, etc.

4] Correlation Analysis

Researchers can effectively detect collinear relationships among different attributes of datasets using correlation analysis, which is based on statistical measures.

5] Outlier Detection

Outlier detection is considered as a primary step in data-mining applications. An outlier is a data point that contains useful information on the anomalous behavior of the variables in the dataset. The outlier detection methods can be divided into the univariate measurement and the multivariate method-actings It is used in credit card fraud detection, financial applications, network detection , etc.

6] Regression analysis

Regression analysis is an important technique in data mining to predict the future value of data based on linear regression.More particularly, regression's primary center is to assist you reveal the exact relationship between 2 or more variables in a given information set.

7] Prediction

Prediction is one of the foremost profitable information mining strategies, since it's utilized to extend the sorts of data you'll see within the future. In numerous cases, fair recognition and understanding chronicled patterns is sufficient to chart a somewhat exact forecast of what will happen within the future. For case, you might survey consumers' credit histories and past buys to anticipate whether they'll be a credit chance within the future.

0.3 Applications

Data mining empowers a retailer to utilize point-of-sale records of client buys to create items and advancements that offer assistance to the organization to draw in the customer.

Market basket analysis in Amazon

These are the following ranges where information mining is broadly used:

1] Data mining in health care

Data mining in healthcare has amazing potential to make an efficient health care system. It uses information and analytics for superior bits of knowledge and to recognize best practices that will improve health care administrations and diminish costs. Examiners utilize information mining approaches such as Machine learning, Multidimensional database, Information visualization, Delicate computing, and measurements. Information Mining can be utilized to estimate patients in each category.

2] Data Mining in Market basket analysis

Market basket analysis may be a modeling strategy based on speculation. In the event that you buy a particular gathering of items, at that point you're more likely to purchase another bunch of items. This procedure may empower the retailer to get it the buy behavior of a buyer. This information may help the retailer in understanding the prerequisites of the buyer and modifying the store's format accordingly.

3] Data mining in education

Education data mining could be a recently developing field, concerned with creating methods that investigate information from the information produced from instructive Situations. EDM targets are recognized as certifying student's future learning behavior, considering the effect of instructive bolster, and advancing learning science. An organization can use data mining to form exact choices additionally to foresee the coming of the understudy. With this comes about, the institution can concentrate on what to educate and how to teach.

4] Data mining in CRM

Customer Relationship Management (CRM) is all about getting and holding Clients, moreover upgrading client devotion and executing customer-oriented procedures. To induce a conventional relationship with the customer, a trade organization has to collect information and analyze the information. With information mining innovations, the collected information can be utilized for analytics.

CHAPTER 1

INTRODUCTION TO MARKET BASKET ANALYSIS

1.1 Introduction

Market Basket Analysis is a strategy used by retailers to get its customer's shopping behavior from their stores. The result of the viable investigation may progress the supplier's benefit, quality of service, and client fulfillment.Nowadays, Amazon is using this strategy to attract customers to increase their revenue.

The reason for this venture is to utilize anonymized information on customers' value-based orders to center on graphic investigation on the client buy designs, things which are bought together, and units that are profoundly obtained from the store to encourage reordering and keeping up satisfactory item stock. It can be done by analyzing the accessible information in such a way that frequent itemsets can be found and can be analyzed to characterize an affiliation rule.
One of the calculations which offer assistance in finding affiliation runs the show for a frequent itemset and to recognize the relationship is the Apriori calculation.

1.2 Overview

Business organisations can now amass large amounts of data thanks to the current technological era in which we live. Currently, database technology innovation has progressed far enough to maintain these information stacks stable; yet, it is critical to examine the information in order to maximise the organization's worth.

In today's customer-centric markets, businesses must develop appropriate and low-cost advertising tactics that can respond to shifts in client perceptions and product demands. It may also aid businesses in identifying a completely new market approach that may be efficiently targeted. All together for critical market strategy decisions, stable, as much as possible, and secure proof-based data is needed.

1.3 Types

Let us delve deeper now that we have a good understanding of Market Basket Analysis and some of the major words involved with it. Market Basket Analysis approaches are classified based on how they use the available data:

1] ***Descriptive Market basket analysis***

The most common approach is descriptive market basket analysis, which only extracts insights from previous data. The study does not make any predictions; rather, it uses statistical approaches to score the relationship between products. Unsupervised learning is the name given to this form of modelling by individuals who are knowledgeable with the fundamentals of data analysis.

2] ***Predictive Market basket analysis***

This method employs supervised learning techniques such as classification and regression. Its main goal is to imitate the market in order to figure out what causes things to happen. Purchasing an extended warranty, for example, is more likely to occur after purchasing an iPhone. To calculate cross-selling, it takes into account things purchased in a specific order.

3] ***Differential Market basket analysis***

This form of analysis is useful for analysing competitors. It examines purchase histories across stores, seasons, time periods, days of the week, and other variables to uncover fascinating patterns in consumer behaviour. It can, for example, assist in determining why some consumers choose to buy the same product for the same price on Amazon vs. Flipkart — the explanation could be as simple as the Amazon reseller having more warehouses and being able to deliver faster, or it could be something more fundamental like the user experience.

1.4 Goals

The project's principal goal is to help Amazon businesses better analyse current customer behaviour and predict future purchases.Customer transaction data can be used to better analyse customers' purchasing habits, offer appropriate bundles and specials, design assortments, and manage inventory to retain consumers, increase revenues, and extend customer relationships.

The project's precise aims are given below.

1. To comprehend the purchase habits of the products in the customer's cart.
2. To research a variety of things that customers frequently purchase.
3. Research the most likely products that customers will buy, as well as a certain product category.
4. Make product recommendations and suggestions to specific customers.

1.5 Benefits

Customer happiness and sales can both benefit from market basket analysis. Retailers can optimise product placement, provide special promotions, and create new product bundles by using data to understand which products are frequently purchased together. This will drive more sales of these combinations.
These enhancements can increase revenue for the store while also making the customer's shopping experience more productive and beneficial. Customers may have a stronger attitude or brand loyalty towards Amazon if market basket analysis is used.

CHAPTER 2
BACKGROUND

2.1 Market Basket Analysis

From the typical corner businesses of the 1900s to the modern e-commerce that has rattled the retail world to its core, retail has transformed. This evolving process has ushered in a new era of limitless opportunities for trade and consumers.

Nowadays, consumers have a diverse selection of choices in practically every subject. When a customer needed to buy something in the past, he could only choose from the store's catalogue. The variety of alternatives, however, has grown dramatically in the modern era of technology and globalisation. Consumers can now choose from a wide range of items and variations.

2.2 Association Rule Mining

Ramakrishnan Srikant and Rakesh Agrawal (Agrawal, n.d.) invented the Apriori method, which is one of the most conventional algorithms for finding frequent patterns of Boolean rules. The authors expand on the concept of quantitative rules in mining in huge relational tables.

By forming such alliances, one can quickly learn about the most popular products and their market share. It is critical to determine which products are the most popular, as a large number of customers come into contact with these products every day. Given the high volume of traffic generated by departments with top items, it's critical to use this information to position other similar products.As a result, the results have examined the process of establishing association rules for the products.

Berry and Linoff (2004) set out to find patterns in transactional data by identifying associations or co-occurrences (Berry and Linoff, 2004). Customers that purchase bread frequently also purchase other bread-related items such as milk, butter, or jam.
The findings indicate why it makes sense to position these area units of groups next to each other in a retail centre so that shoppers may easily access them. Furthermore, such connected product groups should be put next to each other to remind buyers of related products and to direct them rationally around the centre.

Erpolat (2012) presented details on the Apriori and FP-Growth algorithms, which have been widely used in the study of association rules. He then used these algorithms to consumer buying data from an approved automotive service and contrasted the results.
Because of the computing rates, the researcher concluded that the FP-Growth method is better suited to analysing a data set than the Apriori algorithm.

2.3 Recommender System

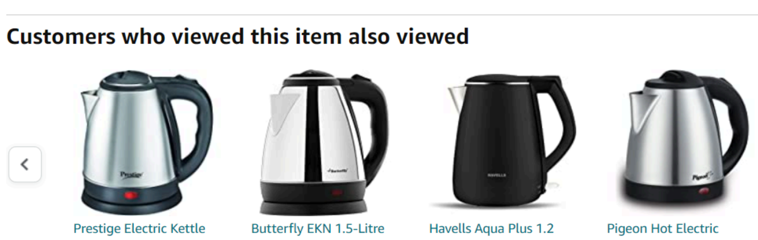
Recommendation systems are information filtering systems that use past user behavior data and attempt to forecast the user's preference for goods. They are not a differentiated category of clients and focus primarily on persons without personal experience.Since the 1990s, many recommender systems have been investigated in a variety of fields, including movies, music, books, articles, social media, and products in general. Burke divides recommendations into five categories: content-based, collaborative filtering (CF), demographical, knowledge-based, and hybrid.

One of the most essential Data Mining techniques used in Market Basket Analysis is the association rule. In a supermarket, all fruits are arranged in the same aisle, and all dairy goods are grouped under another aisle.As a result, taking the time and investing resources to organise the most important things in a logical manner not only cuts down on customer shopping time, but also aids consumers in purchasing the most appropriate items for their Market Basket. The sentence "what goes with what" is connected to the association rule.

Customers' purchases of products at a supermarket are referred to as 'Transactions.' The existence of three parameters, namely support, confidence, and lift, can be used to calculate the magnitude of an associative rule.

2.4 Association VS Recommendation

The Association's rules are not based on personal preferences. It always discovers the relationship between some sets of transaction elements.
Example :
On each product's information page, you'll notice two headings: "Frequently Bought Together" and "Customers who bought this item also bought."



Frequently bought together ➝ Association
Customers who viewed this item also viewed ➝ Recommendation

<u>CHAPTER 3</u>

Market basket analysis in Amazon

ASSOCIATION RULE MINING

3.1 Introduction

In Market Basket Analysis, the association rule is one of the most important Data Mining techniques. In a supermarket, all fruits are sorted in the same aisle, and all dairy products are grouped in another.

As a result, taking the time and investing resources to organise the most important things in a logical manner not only cuts down on customer shopping time, but also aids consumers in purchasing the most appropriate items for their Market Basket. The sentence "what goes with what" is connected to the association rule.

Association rule mining is a technique for detecting common patterns, correlations, and links in datasets stored in a variety of databases, including relational databases, transactional databases, and other repositories.

There are two part to an association rule:

- an antecedent (if) and
- a consequent (if) (then)

An antecedent is something that appears in data, while a consequent is something that appears in accordance with the antecedent. Take a note of the following rule:

“If a customer buys bread, he's likely to buy butter also”

Bread is the antecedent and butter is the consequent in the preceding association rule.
If the above rule is the result of a thorough examination of specific data sets, it can be used to increase both customer service and revenue for a corporation.
Data is rigorously analysed to find frequent if/then patterns, which are then used to develop association rules.

3.2 Effectiveness Measures

Single cardinality refers to relationships in which we can discover an association or relationship between two objects. It's all about making rules, and as the number of items grows, so does cardinality. There are numerous metrics for measuring the relationships between thousands of data pieces. These figures are as follows:

1] <u>Support:</u>
The frequency of A, or how frequently an item appears in the dataset, is called support. It's the percentage of the transaction T that has the itemset X in it. If there are X datasets, the following can be written per transaction T:

Supp(x) = freq(x)/T

2] <u>Confidence</u>
The level of confidence represents how often the rule has been proven correct. Or, since the incidence of X is already known, how frequently the elements X and Y appear together in the dataset. It's the ratio of the number of records that contain X to the number of transactions that contain X.

Confidence = freq(x,y)/freq(x)

3] Lift
It is the power of any rule, which can be expressed using the following formula:

Lift = Support(x,y)/[Support(x) * Support(y)]

If X and Y are independent of one another, it is the ratio of observed support to expected support. It can take one of three forms:

1. Lift = 1 : the chances of the antecedent and consequent occurring are independent of one another.
2. Lift > 1 : This specifies the degree to which the two itemsets are interdependent.
3. Lift < 1 : It indicates that one object can be used in place of another, implying that one item has a negative impact on another.

3.3 Implementation of Effective Measures

Consider the given dataset from Amazon for clear understanding of association rule mining measures:

Transaction	Item 1	Item 2	Item 3	Item 4
1	Phone	Phone Cover	Tempered Glass	Earphones
2	Phone	Phone Cover		
3	Phone	Tempered glass	Earphones	
4	Phone Cover	Tempered Glass		
5	Phone	Tempered Glass		
6	Earphones	Phone		
7	Phone Cover	Earphones		
8	Phone	Tempered glass	Phone Cover	

No of transactions, n = 8

1] ***Support(x) = freq(x)/n***

Support (Phone) = 6/8
= 75%

Support (Tempered Glass) = 5/8
= 62.5%

Support (Phone Cover) = 5/ 8
= 62.5%

Support (Earphones) = 4/8
= 50%

2] ***Confidence = freq(x,y)/freq(x)***

Confidence(Phone, Phone cover) = 3/ 6
= 50%
Confidence(Phone, Tempered glass) = 4/ 6
= 66.67%
Confidence(Phone cover, Tempered glass) = 3/ 5
= 60%
Confidence(Tempered glass, earphones) = 2/5
= 40%
Confidence(Phone, Earphones) = 3/ 6
= 50%
Confidence(Phone cover, Earphones) = 2/ 5
= 40%

3] ***Lift = Support(x,y)/[Support(x) * Support(y)]***

lift(phone,phone cover) = 3/8 [(6/8) * (5/8)]
= 0.8
lift(phone,tempered glass) = 4/8 [(6/8) * (5/8)]
= 1.066
lift(phone cover, tempered glass) = 3/8 [(5/8) * (5/8)]
= 0.96
lift(phone cover ,earphones) = 2/8 [(5/8) * (4/8)]
= 0.8

So, from above calculations , it can be predicted that
if lift > 1, it is likely that customer who buys phone also buys tempered glass

---

3.3 Types

There are three algorithms for learning association rules:

1] Apriori Algorithm
To build association rules, this technique employs a large number of datasets. It's made to deal with databases that have transactions in them. To calculate the itemset efficiently, this approach uses a

breadth-first search and a Hash Tree.
It is mostly used for market basket analysis and assists in determining which products can be purchased together.

2] Eclat Algorithm
Equivalence Class Transformation is the name of the Eclat algorithm. This approach finds frequent itemsets in a transaction database by using a depth-first search technique. It executes faster than the Apriori Algorithm.

3] F-P Growth Algorithm
The F-P growth algorithm is an upgraded variant of the Apriori Algorithm. It stands for Frequent Pattern. It represents the database as a frequent pattern or tree, which is a type of tree structure. This frequent tree's goal is to extract the most common patterns.

CHAPTER 4
APRIORI ALGORITHM

4.1 Introduction

The Apriori method generates association rules by using frequent itemsets, and it is designed to work with transaction databases. It determines how strongly or weakly two things are associated using these association rules. To efficiently calculate the itemset relationships, this approach uses a breadth-first search and a Hash Tree.
It is an iterative method for locating common itemsets in a large dataset.
In the year 1994, R. Agrawal and Srikant presented this algorithm. It is mostly used for market basket analysis, which aids in the discovery of products that can be purchased together.

Because it relies on prior knowledge about common itemset qualities, the algorithm is called Apriori. We employ an iterative strategy, often known as level-wise search, in which we use k-frequent itemsets to find k+1 itemsets.
An important attribute called the Apriori property is utilised to improve the efficiency of level-wise generation of frequent itemsets by minimising the search space.
To decrease the search space, this method uses two steps: "join" and "prune."

The Apriori Algorithm for data mining includes the following steps:

1] Join: By connecting each item with itself, this process yields (K+1) itemset from K-itemsets.

2] Prune : This phase counts all of the items in the database. If a candidate item fails to fulfil the minimum support requirements, it is classified as infrequent and hence deleted. The purpose of this phase is to lower the size of the candidate itemsets.

4.2 Steps

The apriori algorithm is a set of procedures that must be followed in order to determine the most frequent itemset in a database. This data mining technique repeats the join and prune procedures until the most frequently occurring itemset is found. The problem specifies a minimum support threshold, or the user assumes it.

1. Determine the transactional database's support for itemsets and choose the lowest level of support and confidence.
2. Take all supports in the transaction with higher support value than the minimum or selected support value.
3. Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
4. Sort the rules as the decreasing order of lift.

4.3 Working

Using an example and a mathematical calculation, we will learn about the apriori algorithm:

Assume we have the following dataset with multiple transactions, and we need to locate the frequent itemsets and construct the association rules using the Apriori technique from this dataset:

Transaction_Id	Item_Sets
1	Bread, Butter
2	Milk, Butter
3	Butter, Cheese
4	Bread, Milk, Butter
5	Bread, Cheese
6	Butter, Cheese
7	Bread, Butter, Cheese, Cookie
8	Bread, Butter, Cheese

ASSUME : Min. Confidence = 50%
Min. Support = 2

STEP 1: Calculation of Confidence(C1) & Limit(L1)

In the first step, we will create a table that contains the support count (The frequency of each itemset individually in the dataset) of each itemset in the given dataset. This table is called the Candidate set(C1).

Itemset	Support_Count
Bread	6
Butter	7
Cheese	5
Milk	2
Cookie	1

We'll now remove all the items with a higher support count than the Minimum Support (2). It will provide us with the table for the L1 frequently used itemset.

With the exception of Cookie, all of the itemsets have larger or equal support counts than the minimal support, hence Cookie will be deleted.

Itemset	Support_Count
Bread	6
Butter	7
Cheese	5
Milk	2

STEP 2: Calculation of C2 & L2

With the help of L1, we will generate C2 in this stage. In C2, we'll make the pair of L1 itemsets in the form of subsets.
We'll retrieve the support count from the main transaction table of datasets after we've created the subsets, which is how many times these pairs have occurred together in the current dataset. As a result, we'll receive the following table for C2:

Itemsets	Support_Count
Bread, Butter	4
Bread, Cheese	4
Bread, Milk	1
Butter, Cheese	4
Milk, Butter	2
Cheese, Cookie	0

The C2 Support count must be compared to the minimum support count once more, and the itemset with the lowest support count will be removed from the table C2. It will provide us with the table below for L2.

Itemsets	Support_Count
Bread, Butter	4
Bread, Cheese	4
Butter, Cheese	4
Milk, Butter	2

STEP 3: Calculation of C3 & L3

For C3, we'll repeat the first two steps, but this time we'll create the C3 table by combining subsets of three itemsets and calculating the support count from the dataset. It will produce the following table:

Itemsets	Support_Count
Bread, Butter, Cheese	2
Milk, Cheese, Butter	1
Bread, Cheese, Milk	0
Bread, Milk, Butter	0

We'll now make the L3 table. As shown in the C3 table above, there is only one itemset combination with a support count equal to the minimum support count. As a result, the L3 will only have one combination, namely A, B, and C.

STEP 4: Generation of Association rules:

To construct the association rules, we'll start by creating a new table with all of the potential rules from the A, B.C combination. We'll use the formula sup(A B)/A to calculate the Confidence for all of the rules. We will exclude the rules that have less confidence than the minimum threshold after calculating the confidence value for all rules (50 percent).

Take a look at this table:

Association Rules	Support	Confidence
Bread ^ Butter → Cheese	2	Sup{(Bread ^Butter) ^Cheese}/sup(Bread ^Butter)= 2/4 = 0.5 = 50%
Bread ^ Cheese → Butter	2	Sup{(Bread ^Cheese) ^Butter}/sup(Bread ^Cheese)= 2/4=0.5=50%
Cheese → Bread ^ Butter	2	Sup{(Cheese^(Bread ^Butter)}/sup(Cheese)= 2/5=0.4=40%
Bread → Butter ^ Cheese	2	Sup{(Bread^(Butter ^Cheese)}/sup(Bread)= 2/6=0.33=33.33%

The first two rules :

1. Bread ^ Butter → Cheese
2. Bread ^ Cheese → Butter

can be regarded as strong association rules for the given situation because the specified threshold or minimal confidence is 50%.

So, it can be concluded that :
1. If customers buy bread and butter, they are likely to buy Cheese.
2. If customers buy bread and cheese, they are likely to buy Butter.

4.4 Advantages & Disadvantages

The Benefits of the Apriori Algorithm:
1. This algorithm is simple to comprehend.
2. On big datasets, the algorithm's join and prune steps are simple to implement.

Apriori Algorithm Disadvantages
1. In comparison to other algorithms, the apriori algorithm is slow.
2. Because it checks the database many times, the overall performance may suffer.
3. The apriori algorithm has a time and space complexity of O(2D), which is extremely high.
4. The horizontal width of the database is represented by D.

4.5 Improvement Techniques

There are numerous approaches for increasing the algorithm's efficiency.
1] Hash-Based Technique:
For creating the k-itemsets and their accompanying counts, this method uses a hash-based structure known as a hash table. The table is generated using a hash function.

2] Transaction Reduction:
In iterations, this strategy reduces the number of transactions scanned. Transactions with infrequent items are either flagged or eliminated.

3]Partitioning:
To mine the frequent itemsets, this approach simply takes two database scans. It specifies that each itemset that is potentially frequent in the database must be frequent in at least one of the database's partitions.

4] Sampling:
This approach selects a random sample S from Database D and searches for frequently occurring itemsets in S. It's possible that a global frequently used itemset will be lost. This can be mitigated by reducing the min sup parameter.

5] Dynamic Itemset Counting:
During the scanning of the database, this approach can add new candidate itemsets at any designated start point.

CHAPTER 5
METHODOLOGY

A recommendation system's goal is to generate meaningful suggestions for items or products that are of interest to a group of users. The underlying algorithm, such as Apriori and Fp Growth, gathers information about people's tastes and understands that when people buy spaghetti and wine, they are often interested in gravies in general. In designing a recommendation system, the association rule is crucial.

After running on a data collection with details from previous shopping baskets, the Association Rule generates a set of rules. Each rule has an antecedent, a consequence, and a few class metrics including antecedent support, consequent support, support, confidence, and lift.

Machine learning is a field of study that focuses on assisting computers in learning without being explicitly programmed. They are commonly utilised to overcome a variety of life issues. In comparison to the past, Python libraries, frameworks, and modules are now incredibly easy and powerful.Python has largely supplanted many of the industry's languages, thanks to its extensive library, and is now one of the most popular programming languages for this task. The following Python libraries were used in the project:

1] <u>Pandas:</u>
Pandas is an open source, Python-licensed library that provides the Python programming language with high-performance, easy-to-use data structures and data analysis tools. The Data Frame is the foundation of all data structures. In observation rows and variable columns, a data frame allows tabular data to be stored and modified.

2] <u>Numpy:</u>
NumPy is a general-purpose array processing application. 'Numerical Python' is what it stands for. It consists of a collection of multidimensional array objects as well as a set of array processing routines. NumPy includes functions for linear algebra and random number creation.

3] <u>Matplotlib:</u>
It is the art of presenting data using charts, iconography, and presentations, among other things. It is most typical to convert complex data into understandable insights for a non-technical audience. Matplotlib is one of the most used Python tools for data visualisation. This is a cross-platform framework for creating two-dimensional graphs from arrays of records.

4] <u>MLxtend:</u>
MLxtend is a library that implements a variety of machine learning and data mining techniques as well as utilities. MLxtend's main goal is to develop widely used solutions that focus entirely on consistency with existing machine learning libraries via user-friendly and intuitive APIs. Sequential selection feature algorithms, stacked generalisation implementations for classification and regression, and frequent pattern mining algorithms are just a few of the features enforced by MLxtend. MLxtend is a collection of utilities that extend the capabilities of Python's scientific computing stack.

5.1 Data Collection

Data : A two-year sample data set of online shopping transactions.

Information about the data set:
This Online Retail data collection covers all transactions made by a UK-based and registered non-store online retailer between December 1, 2009 and December 9, 2011.
The company primarily sells one-of-a-kind all-occasion gifts. Wholesalers make up a large portion of the company's clientele.

Attributes:
1] InvoiceNo
2] StockCode
3] Description
4] Quantity
5] InvoiceDate
6] UnitPrice
7] CustomerID
8] Country

Dataset consists of 541909 rows and 8 Columns. The dataset is derived from Kaggle.

Sample from the given dataset:

```
In [68]: df.head()
Out[68]:
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

---

5.2 Understanding dataset

Dataset consists of items purchased along with their Invoice_No and stock_code, in which country it is highly and least sold, quantity of the product purchased along with their cost per quantity

1] <u>InvoiceNo:</u>
It is the number of the invoice. Each transaction is assigned a 6-digit integral number that is unique. This code denotes a cancellation if it begins with the letter 'c.'
DATA : Nominal

Sample of Invoice No:

```
In [70]: Invoice_No = (df['InvoiceNo'].value_counts().keys()).tolist()
         Invoice_No

Out[70]: ['573585',
          '581219',
          '581492',
          '580729',
          '558475',
          '579777',
          '581217',
          '537434',
          '580730',
          '538071',
          '580367',
          '580115',
          '581439',
          '580983',
```

2] <u>StockCode:</u>

It is the code for the product (item). Each different product is allocated a 5-digit integral number.
DATA : Nominal

Sample of Stock Code:

```
In [72]: (df['StockCode'].value_counts().keys()).tolist()

Out[72]: ['85123A',
          '22423',
          '85099B',
          '47566',
          '20725',
          '84879',
          '22720',
          '22197',
          '21212',
          '20727',
          '22383',
          '22457',
          '23203',
          'POST',
          '22386',
          '22469',
          '22960',
```

3]<u>Description:</u>

The name of the product (item) purchased.
DATA : Nominal.

Sample of Products:

```
In [69]: items = (df['Description'].value_counts().keys()).tolist()
         items

Out[69]: ['WHITE HANGING HEART T-LIGHT HOLDER',
          'REGENCY CAKESTAND 3 TIER',
          'JUMBO BAG RED RETROSPOT',
          'PARTY BUNTING',
          'LUNCH BAG RED RETROSPOT',
          'ASSORTED COLOUR BIRD ORNAMENT',
          'SET OF 3 CAKE TINS PANTRY DESIGN ',
          'PACK OF 72 RETROSPOT CAKE CASES',
          'LUNCH BAG  BLACK SKULL.',
          'NATURAL SLATE HEART CHALKBOARD ',
          'POSTAGE',
          'JUMBO BAG PINK POLKADOT',
          'HEART OF WICKER SMALL',
          'JAM MAKING SET WITH JARS',
          'JUMBO STORAGE BAG SUKI',
          "PAPER CHAIN KIT 50'S CHRISTMAS ",
          'JUMBO SHOPPER VINTAGE RED PAISLEY',
          'LUNCH BAG CARS BLUE',
          'LUNCH BAG SPACEBOY DESIGN ',
```

4] <u>Quantity:</u>
Per transaction, the quantity of each product (item).
DATA: Numeric

Sample of no of quantities:

```
In [73]: (df['Quantity'].value_counts().keys()).tolist()

Out[73]: [1,
          2,
          12,
          6,
          4,
          3,
          24,
          10,
          8,
          5,
          48,
          25,
          20,
```

5] <u>InvoiceDate:</u>
Invoice date and time. The day and time when a transaction was generated.
DATA : Numeric

Sample of invoices dates:

```
In [74]: (df['InvoiceDate'].value_counts().keys()).tolist()

Out[74]: ['10/31/2011 14:41',
          '12/8/2011 9:28',
          '12/9/2011 10:03',
          '12/5/2011 17:24',
          '6/29/2011 15:58',
          '11/30/2011 15:13',
          '12/8/2011 9:20',
          '12/6/2010 16:57',
          '12/5/2011 17:28',
          '12/9/2010 14:09',
          '12/2/2011 16:39',
          '12/1/2011 16:22',
          '12/8/2011 16:30',
          '12/6/2011 16:26',
          '11/24/2011 9:21',
```

6] <u>UnitPrice:</u> Price per unit. The price per unit of the product is in sterling (£)
DATA : Numeric

Sample of prices of product:

```
In [75]: (df['UnitPrice'].value_counts().keys()).tolist()

Out[75]: [1.25,
          1.65,
          0.85,
          2.95,
          0.42,
          4.95,
          3.75,
          2.1,
          2.46,
          2.08,
          0.83,
          4.13,
          1.95,
          1.45,
```

7] <u>CustomerID:</u>
CustomerID is a unique identifier for each customer. Each customer is granted a 5-digit integral number.
DATA : Nominal

Sample of IDs of Customers :

```
In [77]: (df['CustomerID'].value_counts().keys()).tolist()

Out[77]: [17841.0,
          14911.0,
          14096.0,
          12748.0,
          14606.0,
          15311.0,
          14646.0,
          13089.0,
          13263.0,
          14298.0,
          15039.0,
          14156.0,
          18118.0,
          14159.0,
```

8] <u>Country:</u>
Name of the country. The name of the country in which a consumer is based.
DATA : . Nominal

Sample of countries purchasing the product:

```
In [78]: (df['Country'].value_counts().keys()).tolist()

Out[78]: ['United Kingdom',
          'Germany',
          'France',
          'EIRE',
          'Spain',
          'Netherlands',
          'Belgium',
          'Switzerland',
          'Portugal',
          'Australia',
          'Norway',
          'Italy',
          'Channel Islands',
          'Finland',
          'Cyprus',
          'Sweden',
          'Unspecified',
          'Austria',
          'Denmark',
```

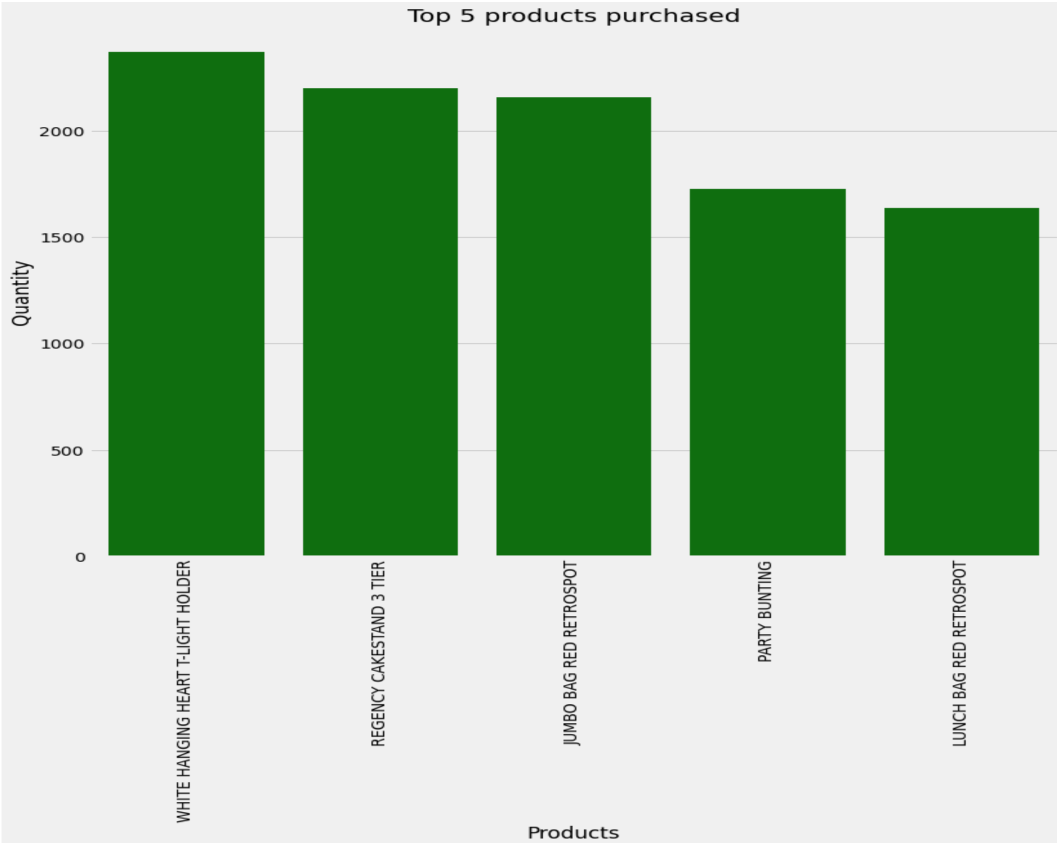
5.3Data Preproceesing

We'll have to clean up a little bit. First, several of the descriptions need to have spaces deleted. We'll also get rid of the rows with no invoice numbers and the credit transactions (those with invoice numbers containing C).

Following the cleanup, we must combine the items into a single transaction per row, with each product being hot encoded once. I'm only looking at sales in France for the sake of keeping the data set small. However, I will compare these statistics to German sales in the supplementary code following. It would be fascinating to look into more nation comparisons.

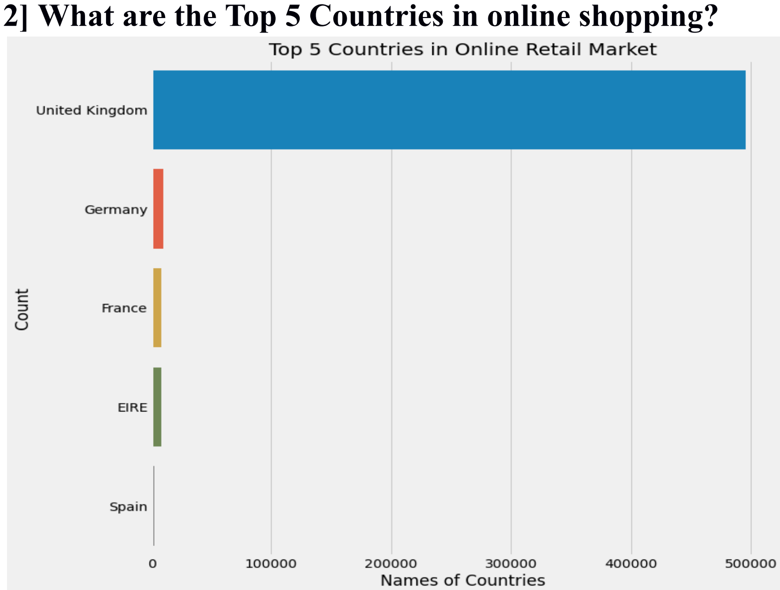
5.4 Data Analysis

1] Which are the top 10 best-selling products?



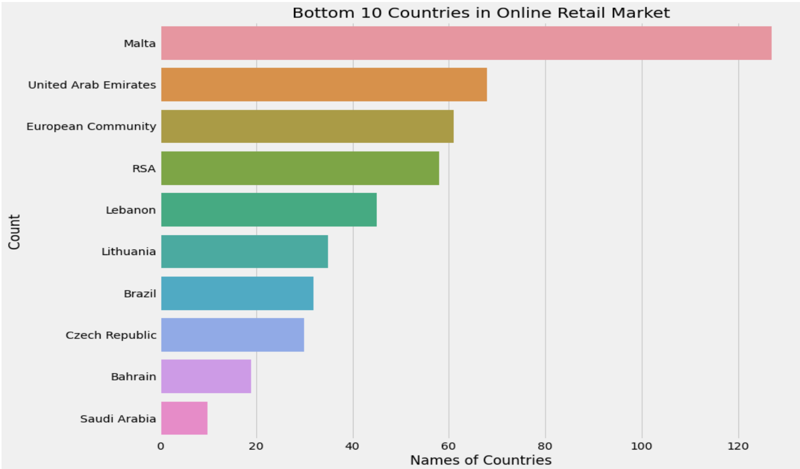
We observe that White hanging heart T-Light holder, Regency Cake Stand, Bag red Retrospot, party bunting and lung bag are purchased most by customers.

2] What are the Top 5 Countries in online shopping?



It is observed that the UK is leading in online shopping.

3] What are the bottom 10 Countries in online shopping?



It is observed that Saudi Arabia lags in online shopping.

5.5 Data PostProcessing

Although the data contains many zeros, we must ensure that any positive numbers are converted to 1s and that everything less than 0 is assigned to 0. This step completes the data's one-hot encoding and removes the postage column.

In [62]:
```
#PostProcessing

def encode_units(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1

basket_sets = basket.applymap(encode_units)
basket_sets.drop('POSTAGE', inplace=True, axis=1)
```

CHAPTER 6
IMPLEMENTATION

6.1 Apriori Algorithm and association mining rule

The first thing that comes to mind is to look for a ready-made algorithm on the scikit-learn website. This approach, unfortunately, is not supported by Scikit-learn. Sebastian Raschka's MLxtend package, fortunately, includes Apriori and FP-Growth algorithms for extracting common item groups for further analysis.

Furthermore, because this is an unsupervised learning method that looks for hidden patterns, data analysis and model construction are not required. This is a tremendous help in some cases of data exploration, and you can use it to prepare the way for a deeper dig into the data by using particular methodologies. Apriori is a prominent method used in association rule learning applications to retrieve frequently observed item sets.

1]Make a list of frequently used itemsets.
Now that the data is appropriately formatted, we can build frequent item sets with at least 7% support (this amount was chosen to ensure that I had enough good examples):

In [63]:
```
#Generate Frequent Itemsets

frequent_itemsets = apriori(basket_sets, min_support=0.07, use_colnames=True)
```

2]Using Frequent Itemset to Create Building Association Rules
The final phase is to create the rules, together with the support, confidence, and lift that go with them:

In [64]:
```
#Building Association Rules Using Frequent Itemset

rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
rules.head()
```

Out[64]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.102041	0.096939	0.073980	0.725000	7.478947	0.064088	3.283859
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.096939	0.102041	0.073980	0.763158	7.478947	0.064088	3.791383
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
4	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.102041	0.094388	0.073980	0.725000	7.681081	0.064348	3.293135

3] Filtering Rules Dataframe
The final step is to use the mlxtend 'association rules' function to create the association rules. Set the minimal confidence level threshold based on the most essential metric (either lift or confidence)

(called min threshold). The Min threshold' as a measure of confidence must be reinstated. When we set the'min threshold' to 1, for example, we will only see rules with 100% certainty.
The difficult aspect now is determining what this means. For example, we can observe that a lot of rules have a high lift value, indicating that they occur more frequently than one might predict given the amount of transaction and product combinations. We can also find other instances when confidence is strong. The domain knowledge will be useful in this section of the analysis.
We can use typical pandas code to filter the dataframe. Look for a large lift (6) and strong confidence (.8) in this case:

```
#Filtering Rules Dataframe

rules[ (rules['lift'] >= 6) &
       (rules['confidence'] >= 0.8) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
17	(SET/6 RED SPOTTY PAPER PLATES)	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.127551	0.132653	0.102041	0.800000	6.030769	0.085121	4.336735
18	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.137755	0.127551	0.122449	0.888889	6.968889	0.104878	7.852041
19	(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.127551	0.137755	0.122449	0.960000	6.968889	0.104878	21.556122
20	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	(SET/6 RED SPOTTY PAPER PLATES)	0.102041	0.127551	0.099490	0.975000	7.644000	0.086474	34.897959
21	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	(SET/6 RED SPOTTY PAPER CUPS)	0.102041	0.137755	0.099490	0.975000	7.077778	0.085433	34.489796
22	(SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY...	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.122449	0.132653	0.099490	0.812500	6.125000	0.083247	4.625850

6.2 Development of Market Basket Recommendation Web application

Flask was used to create the web application. Flask is a micro-web application platform that consists of a set of tools and libraries that make it easier to create online apps with the Python programming language. Pickle library is used to save frequent itemsets that are greater than min support and association rules.

Pickle is used to serialise and deserialize a Python object structure. In this example, a Python object is converted to a byte stream. The dump() method saves the object to the file specified in the parameters. The pickle.load() method opens the file and saves the deserialized data to the model.

1. The Market Basket analysis' home page provides information on the number of goods and the number of rules.
2. The recommendation page allows the user to add things to their cart and displays the items that have been added to the cart as well as a list of recommended products based on the items that have been added previously.
3. The graphs on popular items, departments, aisles, and the findings of the algorithms used for mining the association rules are displayed on the Exploratory Data Analysis page.

Market basket analysis in Amazon

CHAPTER 7
EVALUATION OF ALGORITHMS

7.1 Evaluation of Apriori Algorithm

First, we'll go through the Apriori technique for performing performance analysis in a comparison process. Apriori is the most popular and effective algorithm for mining item sets. Apriori's main premise is to make several runs through data sets or databases that contain transactions or data.

The Apriori method is based on the Apriori property, which states that "a frequent gathering of non-empty item sets must occur." The study also discovered a property that states that if the system fails the minimal support test, all super sets fail the test. The Apriori algorithm employs the Breadth First Search (BFS) method.

It also makes use of the downstream locking feature (any super collection of an unusually broken item is unusual). The transaction database is normally laid out horizontally. In every transaction, the frequency of the item set is calculated.

7.2 Evaluation of F-P growth Algorithm:

When it comes to FP growth, the algorithm uses fragmented and conquering tactics, and the structure of FP data is employed to generate a streamlined transactional database representation. Candidates are not needed to have any standard item sets. Instead of regular patterns, the fp tree is mined. In the early stages of fp development, a list is created and organised in decreasing order of support.

This list is represented by the node structure. Each fp node will have the item name, the support count, and a pointer to a tree node with a comparable item name, in addition to the root node. The fp tree's development is aided by these nodes. Common prefixes can be swapped out throughout the FP tree construction process. The root-to-leaf-node paths are sequentially organised.

Frequent patterns are collected from the FP tree starting in the leaf nodes at this time, when the fp tree is formed. FP Growth uses less memory and has effective storage because of the intended layout. There are two key moves in this game.

1. In comparison, consider the data structure of a compact fp tree.
2. Find objects that appear frequently in the fp tree.

7.3 Evaluation of Eclat Algorithm

Eclat is a database layout approach that is used to mine frequently occurring itemsets. The algorithm is based on depth first search. The data is first represented as a bit matrix in the first phase. If the item is purchased in a specific transaction, the bit is set to one; otherwise, it is set to zero. After that, you'll need to make a prefix tree.

The intersection of the first row with all other rows is used to discover the first item for the prefix tree, and the intersection of the second row with the rows after it is used to generate the second child . All other things are discovered in the same way, and the prefix tree is built in the same way.

Rows that appear infrequently are excluded from further calculations. The depth first search strategy is applied to the prefix tree with backtracking to mine common itemsets. A bit matrix structure is used to store common patterns. Because it employs a prefix tree, Eclat is memory efficient. Because of the small representation, the algorithm is scalable.

7.4 Apriori VS FP growth

SI	Apriori	FP Growth
1.	It is an array based algorithm.	It is a tree based algorithm.
2.	It uses Join and Prune technique.	It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy minimum support.
3.	**Apriori** uses a breadth-first search	**FP Growth** uses a depth-first search
4.	**Apriori** utilizes a level-wise approach where it generates patterns containing 1 item, then 2 items, then 3 items, and so on.	**FP Growth** utilizes a pattern-growth approach means that, it only considers patterns actually existing in the database.
5.	Candidate generation is extremely slow. Runtime increases exponentially depending on the number of different items.	Runtime increases linearly, depending on the number of transactions and items
6.	Candidate generation is very parallelizable.	Data are very interdependent, each node needs the root.
7.	It requires large memory space due to large number of candidate generation.	It requires less memory space due to compact structure and no candidate generation.
8.	It scans the database multiple times for generating candidate sets.	It scans the database only twice for constructing frequent pattern tree.

CONCLUSION

Market Basket Analysis is a concept that originated in marketing and has since been successfully applied to fields such as bioinformatics, nuclear science, immunology, and geophysics. One explanation for MBA's growing popularity in scientific domains is that researchers can test the existence of association rules by adopting an inductive method to thinking.

Many data mining applications rely on association rules to uncover intriguing patterns in large databases. Apriori is the most basic technique for extracting common patterns from transaction databases. The fundamental disadvantage of the Apriori technique is that it is expensive to generate candidate sets, especially when there are a high number of patterns and/or extensive patterns.

Overall, we agree that a recommendation system can have a significant impact on marketing and sales research, which can be used to make strategic company decisions.

FUTURE WORK

Implementing new and advanced mining algorithms, as well as apriori, fp growth for better performance and faster results for sparse datasets, can improve the project. We only employ association rules to use the collective knowledge in the current technique, which entails constructing a model by detecting similarity between customers' product connections and proposing a similar linked item to other items.

In the future, association rules could be used to leverage content-based information, such as discovering similarities between products and proposing products based on comparable product interest. Because the assessment of similarities takes place at the product level, a content-based recommendation system does not rely on a lot of user data.

Perhaps in future work, we might combine the two approaches to create a hybrid strategy that takes advantage of the qualities of both item-based and customer-based approaches. This programme can be expanded to include features like sales monitoring, product tracking, discounting, and price computation, among others.In the future, this strategy could be applied to very big databases when memory space is limited and needs to be increased. It can be fine-tuned even more for better efficiency and performance.

BIBLIOGRAPHY

1]https://www.researchgate.net/publication/343484851_Market_Basket_Analysis_Recommendation_System_Using_Association_Rules#pf17

2] https://www.javatpoint.com/apriori-algorithm-in-machine-learning

3] https://www.softwaretestinghelp.com/apriori-algorithm/

4] https://www.kaggle.com/hassanamin/market-basket-analysis-for-online-retail-dataset

5] https://www.kaggle.com/roshansharma/online-retail-transactions-in-uk

6] http://pnrsolution.org/Datacenter/Vol3/Issue5/116.pdf