**Market Basket Analysis using Data Mining Algorithms**

Lavina Kunder

1. **Introduction**
   1. **Background**

Data mining is the process of searching, analyzing and processing large sets of data that have been collected through various sources. In today’s world, these sources have become more and more progressive such as mobile data, software logs, remote sensors etc. Large data sets collected are called big data.

Market Basket Analysis (MBA) is an extremely valuable application of data mining. The term ‘market basket’ basically means associating a few market items together and determining what items are usually bought together. It clubs items that are most likely to be bought together. Analysis of market basket establishes a connection between various datasets. It is sometimes also called Product association analysis.

The main aim of doing this to help the retail industry understand various patterns and customer behavior. It can help the market owner plan their market strategies, promotional items, placement of products. This in return increases their sales and profitability of the business. A good example of this would be online shopping. When items are being sold online, transaction records are simultaneously being created, these records then after analyzing and processing help in adding to the suggestions column usually found at the bottom of most online shopping websites.

There are various methods to do this some of them are Association, in-memory computation framework, FP-Bonsai, Apriori etc. The most common one being Association rule in which a transaction record is used. This lists all the items brought by customers in a single purchase. Through this, items that are frequently got together can be paired together. This gives us an understanding of how the customer thinks and makes it further easier for the retailer to sale their items.

* 1. **Focus/Application**

The main aim of doing this to help the retail industry understand various patterns and customer behavior. It can help the market owner plan their market strategies, promotional items, placement of products. This in return increases their sales and profitability of the business. A good example of this would be online shopping. When items are being sold online, transaction records are simultaneously being created, these records then after analyzing and processing help in adding to the suggestions column usually found at the bottom of most online shopping websites.

* 1. **Target Audience**

The target audience of this study would be small and big retail stores, supermarkets etc. With the data extracted and through data mining algorithms, they will be able to identify what the customers buy frequently, what offers may be put up, how to attract more customers and to keep the existing ones loyal. They would also be able to develop an idea as to what items are usually bought together enabling them to set their offers so as to increase sales. This would also help them in arranging the aisles at their supermarket so as to make it easier to shop for the customers and increase customer satisfaction.

1. **Data Set**
   1. **Data Source**

This dataset taken from the UCI repository of Datasets, focuses on online transactions.

* 1. **Data Background**

It has 8 attributes, namely:

1. Invoice Number
2. Stock Code
3. Description
4. Quantity
5. Invoice Date
6. Unit Price
7. Customer ID
8. Country

There is a total of 15 rows. All these provide information on the transactions by different customers. This data will be further processed and then analyzed.

* 1. **Data Cleaning**

The dataset that was obtained originally consisted of 54,109 rows. This made it very hard for the dataset to be analyzed. As a result, the dataset was set to a concise size of 15 rows to make it easier to identify patterns on a small scale and then they could be applied to larger datasets much later if they prove to be successful in the current dataset.

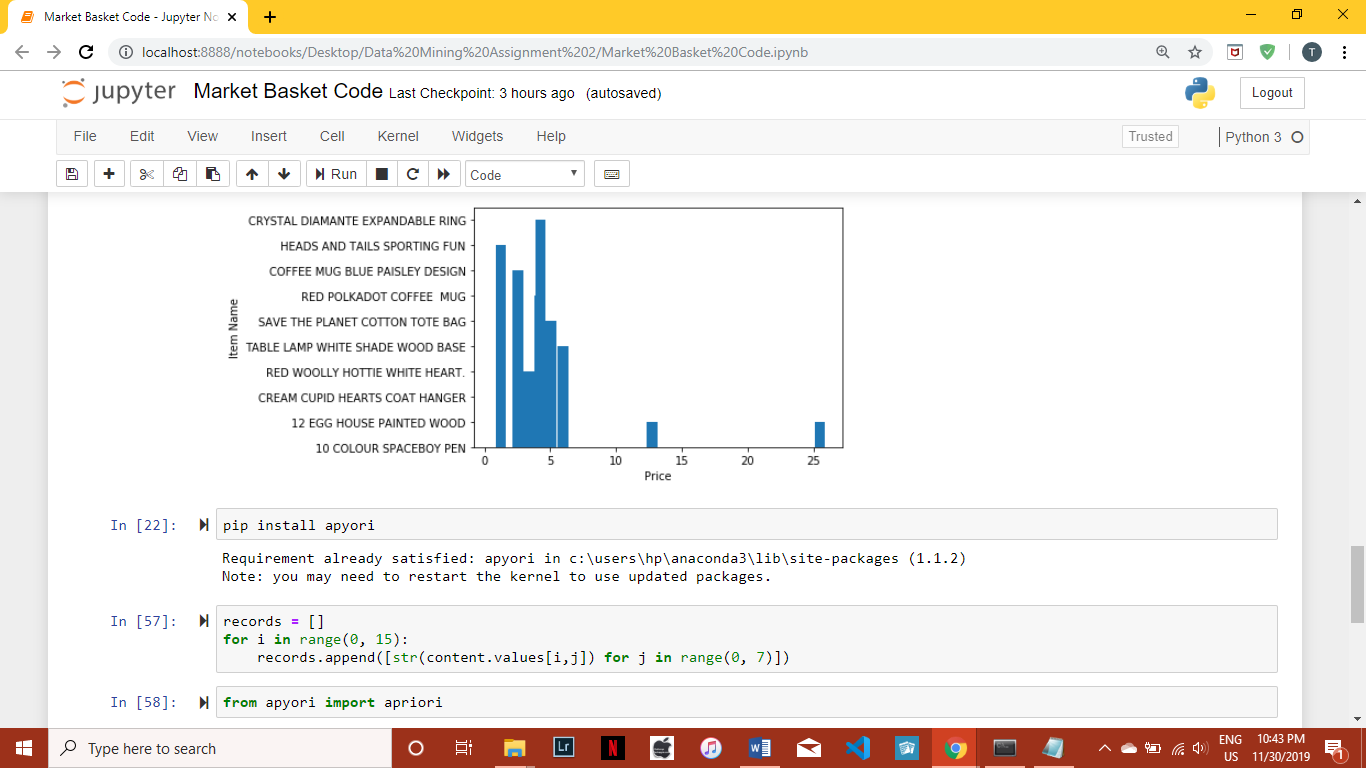
1. **Methodology**

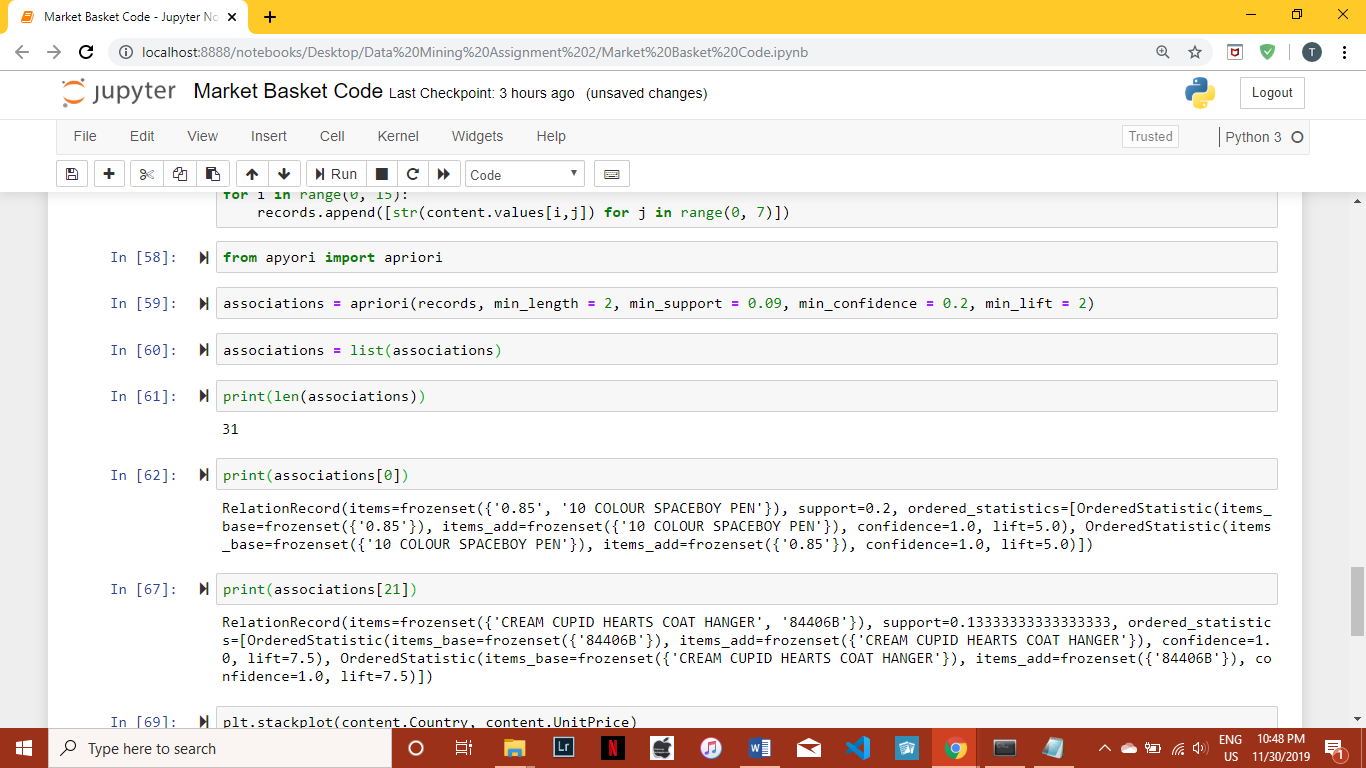
DATA MINING METHODS OR TECHNIQUES USED

* **Apriori Algorithm**

Apriori is an algorithm that is used for frequent itemset mining using association rules of databases. It is based on the concept that a subset of a frequent item set should also be a frequent itemset. This is usually repeated until no more item sets can be formed. It is usually used in situations where there are a lot of transactions, making it perfect for Market Basket Analysis.

In this case it was used to print the most frequent item sets or the items bought together most frequently amongst the existing data that we had.





FLOWCHART

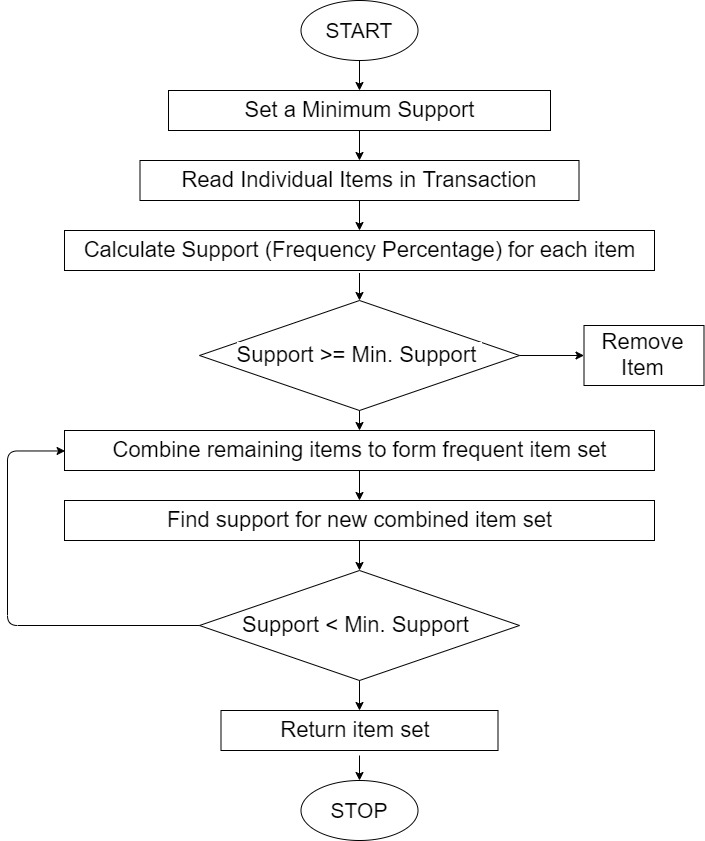


Fig 1

* **Line Plot**

Line Plot is a commonly used graph to show frequency of data for a particular range. A series of data points called markers are used to identify significant points though which a line is drawn.

It is commonly used for continuous datasets and for visualizing data over a period of time.

The best part about this graph is that it can be used to detect anomalies.

In this case it is used to display the quantities of each particular item bought in the whole data set. It helps in visualizing the spread of products and therefore understanding which products should be marketed better.

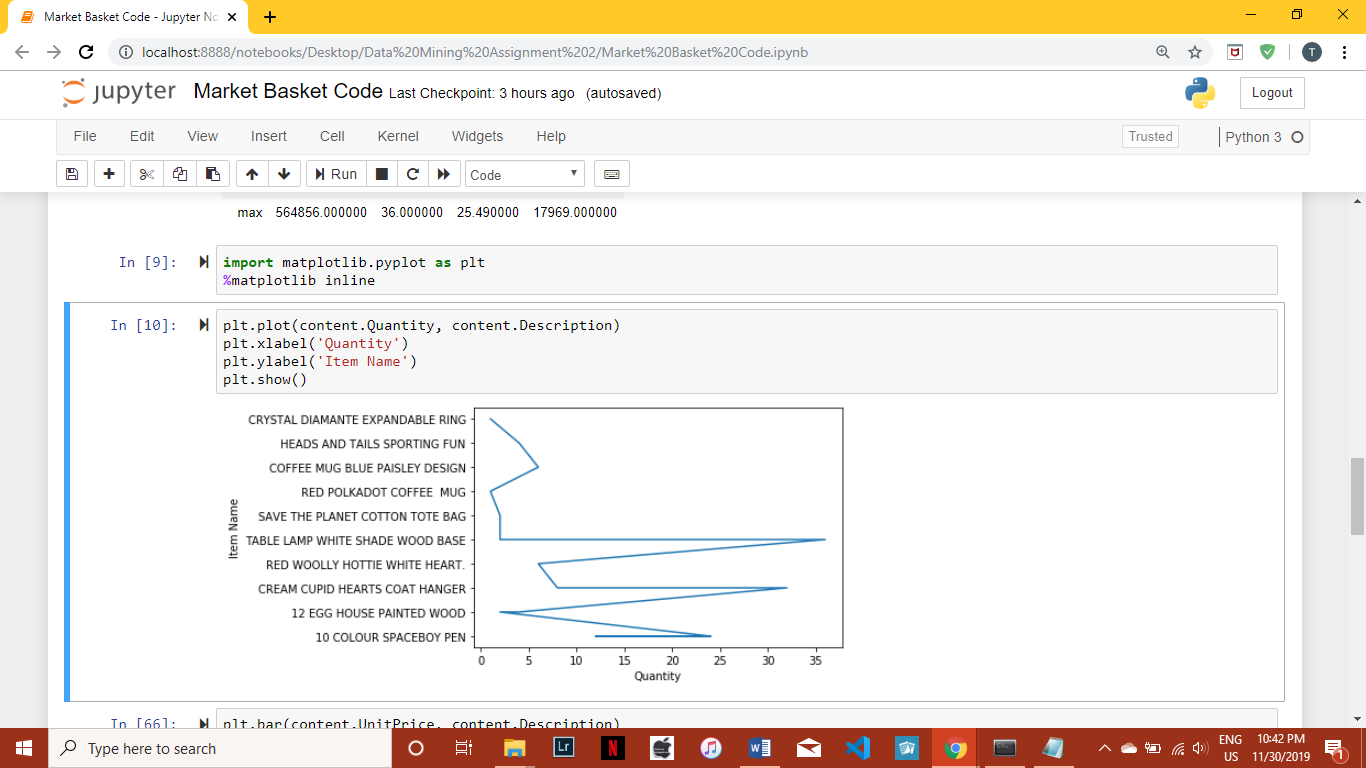


Fig 2

* **Bar Charts**

They are graphs used to compare the values of a variable at a given point of time. The length of each bar is proportional to the value of the item it represents. However, unlike Line Plot, they do not show changes over time. They are used to compare attributes over different groups.

In this case it has been used to identify the price distribution amongst the various items in the dataset. It helps in identifying expensive products and provide an insight on as to how they should be priced so as to increase consumer interest.

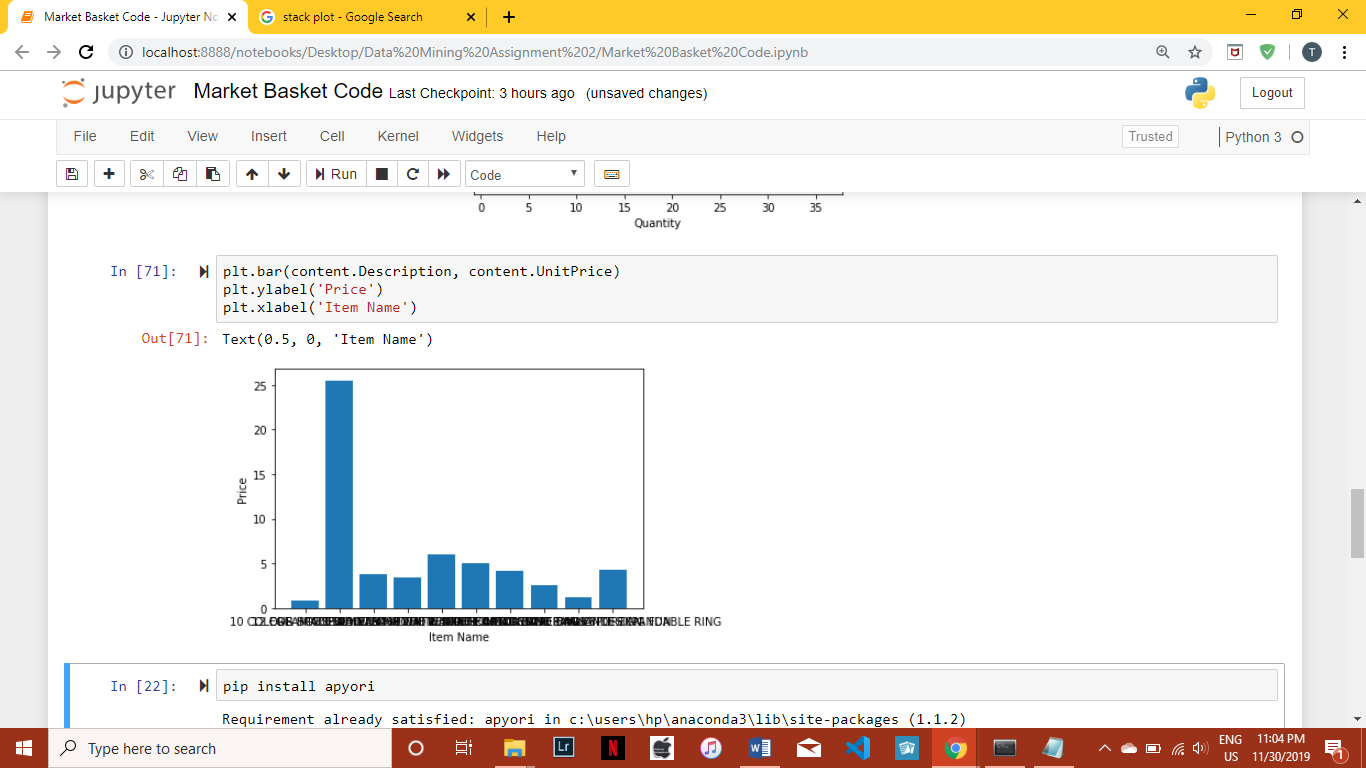


Fig 3

* **Pie Chart**

It is a circular statistical graph divided into slices to illustrate numerical proportions.

There are a number of variations to this graph (Doughnut, Ring, etc.), but the basic type is what is most used. Each slice on the graph represent a proportion of the whole dataset. It provides a clear idea as to how much of the whole belongs to each proportion. (In this case, for top 5 countries with most transactions)

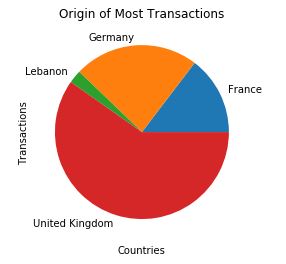


Fig 4

* **Stack Plot**

Similar to the pie chart, the stack plot shows the makeup of part of the unit as well as the whole unit. The constituent of each type is stacked over each other to show relation with the total, thus making visualization easy. It offers a clearer look into each component. In this case it is used to distinguish between the price differences in different countries and the amount of purchases in each of them.

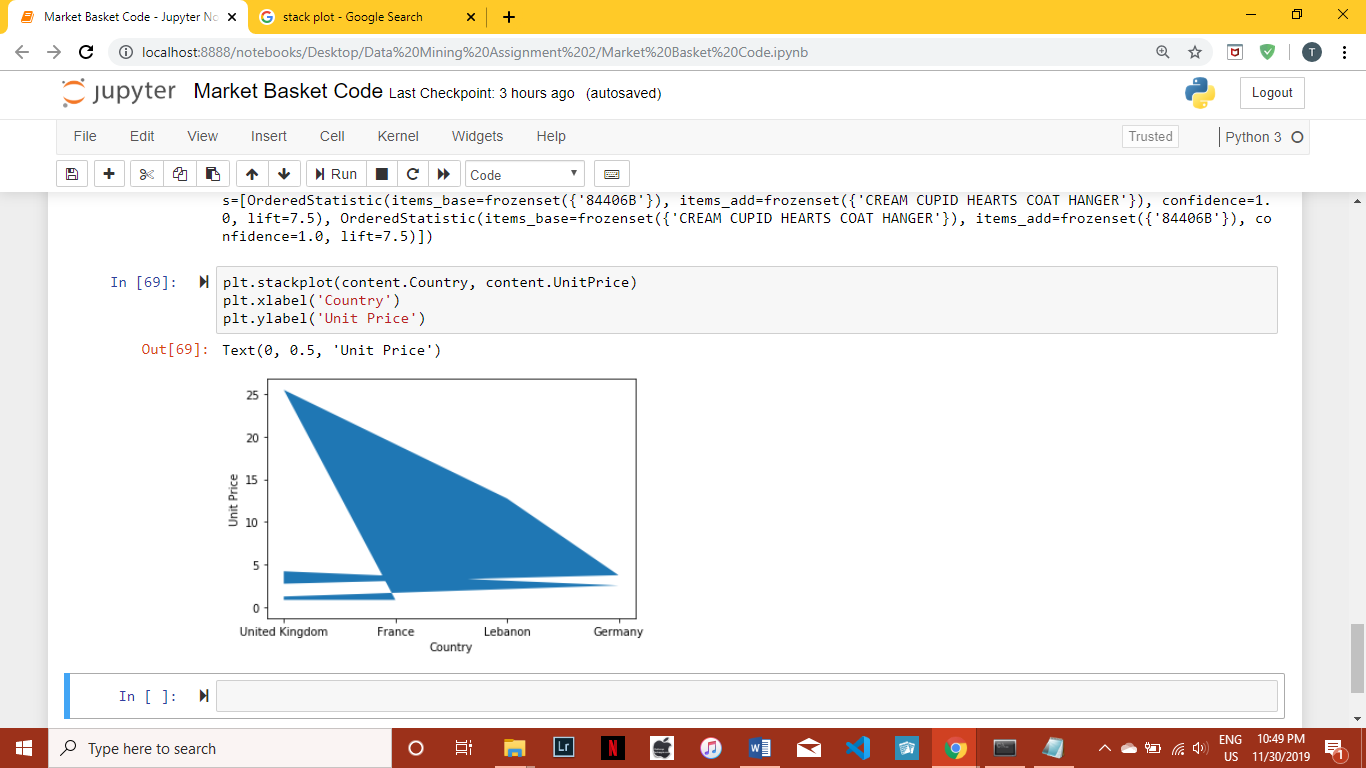


Fig 5

1. **Results**

Through Apriori Algorithm, we were able to identify frequent item sets that point to which items are bought together and their quantities. This helps in setting up better offers and help in selling goods that aren’t bought frequently by pairing them with frequently bought items in the form of offers or sales. The flowchart as per Fig 1, is to provide an idea as to how the process works.

The line plot helped in identifying the sale quantity of different products, therefore enabling us to properly visualize which items are sold the most and which items may need to be marketed better for increase in sales.

The bar chart represented the prices of different products, hence showing the distribution in prices amongst the different products. This could be used to price the items more efficiently so as to keep a proper range between products and help increase customer satisfaction.

The Pie chart focused on where the transactions originated from. It showed the United Kingdom to be in the lead with the most transactions and Lebanon tailing with the least. This helps is providing a general idea as to where further customer relations may need to be developed and where we need to tend to the customers existing satisfactions, so as to ensure they remain loyal consumers.

The stack plot helped in identifying the price differences in different countries with a significant difference between UK and Germany.

1. **Conclusion**

Market Basket Analysis has proved to be very efficient in providing aid to various businesses for maintaining as well as developing relations with new and old customers alike.

It has helped in identifying important trends in the existing data, so as to let companies capitalize on this information for increased profits and customer satisfaction.

The offers, campaigns, marketing strategies, etc. built on this extracted data have proven to be very useful in garnering consumer interest and allowing sales to increase exponentially.

Therefore, Market Basket Analysis has proven to be an indispensable part of the retail industry and through new techniques and practices, will only help businesses reach new heights.