MARKET BASKET ANALYSIS

A Project Report

Submitted to Goa University

In partial fulfillment of the Requirement

For the degree of

Bachelor in Computer Applications

By

Mr. Shirijay Gaons

Mr. Mohanish Khedekar

Ms. Suraksha Gawas

Mr. Ankush Ghadi

Mr. Krishna Ghadi

Ms. Ridhi Jalmi

Mr. Shubham Kolkar

Mr. Nambresh Madkaikar

Ms. Aseeta Melekar

Ms. Manjiri Malik

Guided by

Ms. Nilaxi Chari

SPES's Shri Gopal Gaonkar Memorial Goa Multi-Faculty College

Affiliated to Goa University

CERTIFICATE

This is to certify that the project on

"Market Basket Analysis"

has been successfully completed and submitted by

Mr. Shirijay Gaons
Mr. Mohanish Khedekar
Ms. Suraksha Gawas
Mr. Ankush Ghadi
Mr. Krishna Ghadi
Ms. Ridhi Jalmi
Mr. Shubham Kolkar
Mr. Nambresh Madkaikar
Ms. Aseeta Melekar
Ms. Manjiri Malik

Nilaxi

Ms. Nilaxi Chari

(Internal guide)

Sawant

Mrs. Nisha Sawant

(Project Coordinator)

Prof. (Dr.) Shaikh Mohammad Parvaz Al-Usmani

(Principal)

Patil 24/06/22
(Sameer Patil)

External (Examiner)

SPES
Dharbandora
Goa
Gopal ... Faculty College

SPES's Shri Gopal Gaonkar Memorial Goa Multi-Faculty College
Affiliated to Goa University

DECLARATION BY CANDIDATES

We declare that this project report has been prepared by me/us and to the best of my/our knowledge, it has not previously formed the basis for the award of any diploma or degree by this or any other University.

Roll No	Name	Seat No.	Signature
9312	Shrijay Gaons	0234	
9313	Mohanish Khedekar	0218	
9314	Suraksha Gawas	0238	Gawas.
9315	Ankush Ghadi	0208	
9316	Krishna Ghadi	0215	Ghadi
9317	* Ms. Ridhi Jalmi	—	ABSENT
9320	Shubham Kolkar	0235	
9321	Nambresh Madkaikar	0221	
9322	Aseeta Melekar	0260	
9323	Manjiri Malik	0217	

* Ms. Ridhi Jalmi dropped out during Semester 6.

SPES's Shri Gopal Gaonkar Memorial Goa Multi-Faculty College

Affiliated to Goa University

CERTIFICATE BY THE SUPERVISOR

This is to certify that the project report is the record of the whole work done by the candidates themselves under my guidance during the period of study and that to the best of my knowledge; it has not previously formed the basis for the award of any diploma or degree by this on any other university.

Name of college: SPES's Goa Multi-Faculty College

Programme: Bachelor in Computer Applications (B.C.A.)

Academic Year:

2021-2022

Ms. Nilaxi Chari

Project Guide

ACKNOWLEDGEMENT

We take this opportunity to thank all those who have helped us to complete the project in this stipulated period of time. Our project Market Basket Analysis, consumed huge amount of work, research and dedication, still implementation would not have been possible if we did not have support of many individuals. Therefore, we would like to extend our sincere gratitude to all of them.

We would like to show our gratitude to our internal guide Ms. Nilaxi Chari for her constant support, motivation, inspiration throughout this project.

We are also thankful to Mrs. Nisha Sawant for her help and support as and when required.

We also extend our sincere gratitude towards our Principal, Prof. (Dr.) Shaikh Mohammad Parvez Al-Usmani, for his wholehearted support in this project.

Nevertheless, we express our sincere gratitude to our teaching and non-teaching staff members and friends for their kind co-operation and encouragement throughout this project.

-The Project Team

TABLE OF CONTENTS

I . INTRODUCTION

Market Basket Analysis is a strategy used by merchants to better understand their customers' shopping habits. The supplier's profitability, service quality, and customer satisfaction may all improve because of the effective analysis. Instacart is a supermarket delivery service that operates in the United States. The goal of this project is to use anonymized data from customers' transactional orders to focus on descriptive analysis of customer purchase patterns, items that are purchased together, and units that are frequently purchased from the store, this in turn helps to make reordering and maintaining adequate product stock effectively.

Market basket analysis can be divided into two types: *Market basket analysis that predicts the future*- To assess cross-sell, this kind considers products purchased in order; and *Market basket analysis with a difference*- which takes into account the data from several stores, as well as purchases from various client groups at various periods of the day, month, or year.

If a rule holds true in one dimension (such as store, time period, or customer group) but not in others, analysts can figure out what's causing the exception. These revelations may lead to new product offers that boost sales.

Market basket analysis examples

A well-known example of market basket analysis can be seen on the Amazon website. Amazon displays related products on product pages under the headings "Frequently bought together" and "Customers who bought this item also bought."

Market basket analysis is also applicable to physical stores. If research shows that buying a magazine commonly includes buying a bookmark (which is an unusual combo given that the customer did not buy a book), the book-store might have a selection of bookmarks near the magazine rack.

Algorithms used in the study of market baskets

Association rules are used in market basket analysis to forecast the chance of products being purchased together. The frequency of elements that occur together is counted by association rules, which are looking for relationships that occur considerably more frequently than predicted.

AIS, SETM, and Apriori are examples of algorithms that use association rules. The Apriori approach is frequently used by data scientists in market basket analysis research articles, and it is used to locate frequent items in a database, then evaluate their frequency as datasets grow larger.

Scope of the study

Market basket analysis targets customer baskets in order to monitor shopping for patterns and enhance customer. It is an important element of analytical CRM in retail organizations. By analyzing, recurring patterns in order to offer related goods together, sales can be raised.

Some examples of the use of market basket analysis include:

Product placement: Identifying products that may often be purchased together and arranging the placement of those items (such as in a catalog or on a web site) closely to encourage the purchaser to buy both items.

Physical shelf arrangement: An alternate use for physical product placement in a store is to separate items that are often purchased at the same time to encourage individuals to wander through the store to find what they are looking for to potentially increase the probability of additional impulse purchases.

Up-sell, cross-sell, and bundling opportunities: Companies may use the affinity grouping of multiple products as an indication that customers may be predisposed to buying the grouped products at the same time. This enables the presentation of items for cross-selling, or may suggest that customers may be willing to buy more items when certain products are bundled together.

The main goal of a marketing campaign is to entice customers to visit the shop and buy more than they usually do. Profit margins on promoted items are usually cut; therefore, non-promoted items with the higher profit margin should be sold together with the promoted items. Therefore, the items chosen should make the advertising powerful, sufficient to generate higher sales. In market basket analysis, you analyze purchases that commonly happen together. For example, people who buy bread and butter also buy jelly. It makes sense that these groups are placed side by side in a retail store so that customers can access them easily. When different additional brands are sold together with the basic brands, the revenue from the basic brands is not decreasing but increasing. Market basket analysis targets client baskets so that you can display shopping for styles and improve customer experience. It is an important element of analytical CRM in retail organizations. By analyzing routine styles in order to offer related items together, a pattern can be observed and therefore the sales can be increased. Sales on different levels of goods classifications and on different customer segments can be tracked easily.

If the enterprise is aware of which items are often bought together, they are able to create new offers on those products that allows you to increase the sale of those items or even they can create mixture promotions, which will increase the sale of their products.

The important goal of the project is the detection and analysis of purchase behavior of the customers (items purchased together). Based on variability in demographics and difference in the segment of customers, every store will have different results. So, in our research we will be focusing on a single store and its transactions which will be giving a comparative analysis in different time periods. This way we will be able to investigate the fluctuation of the consumer behaviors on purchases with the change in time.

Analyzing transactional database and discovering association rules have helped the managers to increase the sale of the company. Time series analysis allows you to track the changes in the trends and to keep track of the progress or downfall of any market.

Cross-selling is an e-commerce practice where retailer offer a user additional product that complement those who have already purchased or are about to purchase. For example, cross-selling come up with the product recommendations such as socks, shoelace, or shoe care products.

Upselling, in its turn, is the strategy that consists of suggesting a user more expensive or upgraded product compared to those that they have initially chosen. Upselling example – offering a pair of similar, but more expensive sneakers from the new collection.

Motivation of the study

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding, customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

Market basket analysis encompasses a broad set of analytics techniques aimed at uncovering the associations and connections between specific objects, discovering customers' behaviour and relations between items. It is based on the following idea: if a customer buy a certain group of items, is more (or less) likely to buy another group of items. For example, it is known that when a customer buy beer, in most of cases, buys chips as well. These behaviour produced in the purchases is what we were interested in. We were interested in analyzing which items are purchased together in order to create new strategies that improved the benefits of the company and customers experience.

Objectives of Market Basket Analysis

1. To identify next product that might interest a customer.

   Know your customers better because only they can help you get more lead and more business. Understanding customers is the key to giving them good service, which in turn, results into strong customer relationships. However, understanding the customers' is not easy and most often requires a thoughtful analysis to identify their preferences or purchase patterns so that you can anticipate their needs and exceed their expectations.

2. To generate the association rule from the frequent item sets.

   Agrawal et al. first proposed frequent item set mining for market basket analysis in the context of association rule mining. It analyzes customers' buying habits by finding associations between the different items that customers place in their "Shopping baskets."

3. To improving customer experience.

   Customer experience is the customer's overall sentiment of every interaction that they have with your company. This evaluation runs parallel to your inbound marketing approach, beginning with the first point of contact with the customer and ending with your feedback outlets.

4. To identify the purchasing behavior of the customer.

   To understand the purchasing behaviour of the customers in retail stores it is very important to analyze the customer psychology, the factors which influence a customer for buying certain products/services from the stores and also an analysis of the customer's response towards a sales' promotion is very critical.

5. To increase customer engagement.

   One of the most effective ways to improve customer engagement is to create a customer loyalty program. These act as incentives to reward loyal customers who continually engage with your brand through points, discounts, special gifts, and more.

6. To boost sales.

   Increasing sales is not a one day show as it requires dedication and consistency. Thus, you to focus on a few tactics that will allow you to boost sales and revenue.

7. To understand customers better.

   Knowing and understanding customer needs is at the center of every successful business, whether it sells directly to individuals or other businesses. Once you have this knowledge, you can use it to persuade potential and existing customers that buying from you is in their best interests.

8. To identifies customer behavior and pattern.
   The study of consumer behavior is the research of the behavior of consumers and the methods they employ to decide, purchase (consume), and dispose items and services. This includes the consumers' mental, emotional, and behavioral responses.

9. To develop more effective pricing, product placement, cross-sell and up-sell strategies.

10. It can help predict product sales in specific locations, improving shipping times and warehouse operations.

11. To identify the frequent items from the transaction on the basis of support and confidence

12. Market basket analysis is also used to predict future purchase decision of a customer

Benefits of Market Basket Analysis

Shopping cart analysis can increase sales and customer satisfaction. Using data showing that products are often purchased together, retailers are offering new product bundles to optimize product placement, offer special offers, and promote further sales of these combination. You can create it.

These improvements can bring additional sales to retailers and make the shopping experience more productive and valuable to their customers.

Store layout: to enhance revenue, plan or set up your store according to market basket analysis. The products are put close together so that the buyer notices them and decides whether or not to buy them.

Marketing messages: market basket analysis improves the effectiveness of marketing messages by allowing you to provide relevant ideas to your customers using data.

Recommendation engines: the foundation for constructing recommendation engines is market basket analysis. A recommendation engine is a piece of software that analyses, finds, and suggests material to users who are interested in it.

Application of Market Basket Analysis

Cross selling is a sales method in which a vendor promotes a related product to a customer after he purchases a product. Market basket research enables a business to understand customer behaviour and then cross-sell to them.

Product placement refers to grouping complementary and replacement goods together so that the buyer notices them and buys both of them at the same time, increasing the likelihood of a purchase.

Customer behaviour: market basket analysis helps you understand customer behaviour under different conditions.

Helps with pricing: market basket analysis helps retailers identify the SKUs preferred by a particular customer. For example, infant formula and coffee are often purchased together, and analysts have assigned them more likely to be relevant compared to cookies. Without market basket analysis, retailers usually discount coffee on a particular day, assuming that the coffee is sold at a particular time. However, market basket analysis shows that every time a customer buys milk, he also buys coffee. Whenever, milk and coffee sales are expected to increase, retailers can lower the price of biscuits to increase sales.

SKU display arrangement: a common display format used in supermarket chains is a departmental system in which products are sorted by department. For example, groceries, dairy products, snacks, breakfast items, cosmetics, and personal care products are properly categorized and displayed in various sections. Market basket analysis helps identify items that have a close affinity for each other, even if they fall into different categories. Using this knowledge, retailers can place more affinity items closer to increase sales. For example, if the chips are located relatively close to the beer bottle, the customer will most likely buy both. In contrast, when placed in two extreme locations, customers simply step into the store, buy beer and leave the store, and chip sales are lost.

Promotional coordination: marketers can investigate the buying behaviour of individual customers and relatively reliably estimate the next item they are likely to buy. Today, many online retailers use shopping cart analytics to analyse each individuals buying behaviour. Such retailers can reliably estimate items that individuals can buy at any time. For example, a customer who looks grills will buy meat and barbecue sauce over the weekend. This allows retailers to customize their offers and combine meat and a box of barbecue sauce at a discounted price each weekend to increase purchase frequency.

Identifying sales influencers: all items in a retail store, strong or weak, are related to each other. In most cases, the sale of one item is facilitated by the increase or decrease in the sale of the other item. You can use market basket analysis to investigate buying trends for a particular SKU. The two SKUs have a strong affinity for a period of time and can suddenly decline due to a variety of factors, including rising SKU prices, new brand launches, and the availability of certain brands in SKUs. For example, corona is a consumers' favorite beer and that brand is suddenly removed from the beer SKU, the sales of other beer brands will be stable. This allows marketers to understand the impact of such activities on sales.

Limitations of Market Basket Analysis (MBA)

1. iffy correlation

Averages can be deceiving. You'll hit some speed bumps if you try to replicate a conclusion based on chainwide data to merchandise a single location. The 'beer and diapers' association is a well-known example of this issue in the retail business.

A retail company ran SQL queries against its shop data before 'big data' (the 1990s) and discovered that beer was frequently purchased alongside diapers. This discovery immediately gained traction, and businesses began stocking diapers and beer beside each other on store shelves. Naturally, sales of the two items increased when purchased together.

Naturally, sales for the combination of items increased since businesses were marketing them close to one another in a high-traffic area. The foundation of the issue is that, you can't look for and verify correlations in data that you helped to create. As a result, while market basket research can help you notice a trend, it's tough to judge the veracity of the correlation until you've acted on it.

2. No clear calls to action

Even if there is a true correlation between the two products, determining how to act on this information takes time and talent. Let's imagine you discover that 50% of clients who buy bread also buy rice within two weeks. You may utilise your market basket analysis tool to send a recommendation to every online shopper who purchased bread over the last two weeks. But when should you transmit it so that you may take use of the connection? You won't be able to know from the tool. What can you do with this knowledge in your physical stores? In the ideal world, your consumer has a loyalty card or app that allows you to track their transactions. You might use that to deliver an app-based promotion for the rice

3. Test and learn lag times

Because there are no clear calls to action, businesses who use these solutions must spend time A/B testing whatever actions they take in response to the data. But how do you choose the correlation to examine in an A/B test? It takes a lot of time and effort to test even the strongest associations.

If you want to conduct an in-store cross-promotion, for example, you'll need to re-organise your shelves, alter your planograms, and issue directives to all of your stores.

Following that, you'll need to teach your employees on the new locations and inform them of the promotion. After then, you'll need a reasonable amount of time (weeks? months?) to see if you were correct.If you want to conduct an in-store cross-promotion, for example, you'll need to re-organise your shelves, alter your planograms, and issue directives to all of your stores.

Following that, you'll need to teach your employees on the new locations and inform them of the promotion. After then, you'll need a reasonable amount of time (weeks? months?) to see if you were correct.

Aside from the costs, the entire process is long, from studying the correlations to implementing adjustments, verifying their efficacy, and applying what you have learned in the future.

Data mining

The process of extracting data to spot patterns, trends, and helpful information that may enable the business to require the data-driven call from immense sets of knowledge is termed data Processing. In alternative words, it is the process of finding patterns, information for categorization and conversion into helpful data, that is collected and assembled specially from areas like information warehouses, economical analysis, data processing rule, serving to deciding and alternative information demand to eventually cost-cutting and generating revenue.

Data mining is the act of mechanically sorting out giant stores of data to search out trends and patterns that transcend straightforward analysis procedures. Data processing utilizes complicated mathematical algorithms for knowledge segments and evaluate the likelihood of future events. Data processing is additionally known as data Discovery of knowledge.

Data Mining may be a method employed by organizations to extract specific knowledge from large database to resolve business issue. It primarily turns data into information.

Data Mining method includes numerous kinds of services like text mining, web mining, audio and video mining, pictorial data processing, and social media mining. It's done through package that's easy or extremely specific. By outsourcing data processing, all the work is done quicker with low operation prices. Specialized corporations can even use new technologies to gather knowledge that's not possible to find manually.

Abstract

Market basket analysis (also known as data mining in the field of marketing) is a method for determining correlations between goods / item sets and analysing consumer behaviour based on those associations. It is an integral aspect of a system for determining product placement and planning sales promotion for different segment of customer in order to improve customer happiness and thereby profit. Retailers uses market basket analysis to learn about their customers' shopping habits. As time passes, the customer habits and behaviour changes. We investigate the problem of identifying association rules showing a consistent cyclic variation across time. This issue will allow us to track changing trends. We will be able to predict

customer buying behaviour in a retail sector. The goal of this project is to use anonymized data from customers' transactional orders to focus on descriptive analysis of customer purchase patterns, items that are purchased together, and units that are frequently purchased from the store in order to make reordering and maintaining adequate product stock aptly. It can be done by examining the given data in order to identify and analyse common item-sets in order to define an association rule

1.1 PROBLEM STATEMENT

To study if Market basket analysis affects sales trends

1.2 OBJECTIVES

- Identify customer behaviour
- Apply association rules of Unsupervised learning technique
- Identify dependency of one data item on another data item
- Cluster item into item sets
- Market basket analysis with Apriori Algorithm
- Improve customer experience

1.3 HYPOTHESIS

- Apriori Algorithm gives 100% accuracy for Market basket analysis
- Market basket analysis increases sales of products grouped together
- Market basket analysis can be used by companies to spot sales trends and make better decisions
- Market basket analysis improves the sales of all products
- Mineral waters is the highest selling product in the dataset

II . REVIEW OF LITERATURE AND RESEARCH METHODOLOGY

2.1 LITERATURE REVIEW

1. Building Prediction Model using Market Basket Analysis, 2017, Roshan Gangurde.

In February 2017, Roshan Gangurde published a study on building predictive models with market basket analysis, stating that when they use market basket analysis to come up with product bundles in a retail business, They are using past customer purchase behaviour to predict future purchase behaviour, which is a predictive model. They also stated that by using MBA, premier merchants will be able to attract more customers, raise the value of their market basket, generate more successful advertising and promotion, and much more. The study also recommended that intelligent prediction models be designed and developed in order to provide association rules that may be used in recommendation systems to make them more functional. They developed an improved MBA approach at the end of 2017.

2. Fast Algorithms for Mining Association Rules, 2000, Rakesh Agarwal.

The Apriori algorithm was proposed by Rakesh Agarwal. Apriori was the first associative algorithm proposed, and it has since been employed in future developments in assocuation, classification, and associative classification algorithms. The Apriori algorithm counts transactions on a level-by-level, breadth-first basis. Prior knowledge of common item set properties is used in the apriori process. Apriori employs an iterative method known as level-wise search, in which n-item sets are utilised to investigate (n+1)-item sets. The Apriori property is used to improve the efficiency of the level-wise production of frequent item collections. All non-empty subsets of a frequent item set must likewise be frequent, according to the apriori property. Because of the anti-monotone quality of support measure, the support for an item set never exceeds the support for another item set.

3. Trend Analysis of Association Rules in Different Time Periods, 2016, Kaur and Kang.

Kaur and Kang (2016) used association rule mining to find evolving trends in market data. This research offered a new strategy to periodic mining that will improve the effectiveness of data mining approaches. This research was beneficial in identifying interesting patterns in the data. a huge database that predicts future association rules and provides us with the right methods to figure it out outliers. This research not only shows progress through mining static data, but it also shows a novel technique to mine data to take into account data changes

4. Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules, 2018, Lee, Liu, and Mu.

An MBA was applied to transactional data provided by a Korean shopping mall, that mostly sells food, products, cosmetics, and other items by Lee, Liu, and Mu. The transactional data consisted of 51,080 transactions that occurred between June 2015 and June 2016 and included customer information data like age, gender, customer ID, and so forth. The first stage in this research was to establish a baseline. Using the RFM approach, divide client data into VIP and non-VIP categories. After that, the VIP data is run through an MBA to determine the most effective rules and regulations. The authors present the top ten association rules that shown a high level of confidence. Cosmetics were found in 90% of the cases, and four of them contained cosmetics. To gain a better understanding of the relationship the authors examined the regulations they had collected.

5. A Market Basket Analysis of a cosmetic company, 2020, Hidayata et al.

Hidayat et al. used transactional data from Breillant, an Indonesian online cosmetics business on the Shopee platform, which is an e-commerce online purchasing platform, to do an MBA. 34 sales transactions were included in the transactional data. November 2018 is a month to remember. There were 47 different forms of attractiveness in the deals. There were a total of 126 products in the transactions, making the total number of products 126. The authors don't say how many rules the Apriori algorithm generates in total, but they do say that it's a mix of products with a lot of support and confidence. Customers who purchased Original Liquid Bleaching Seeds and Harva Peeling Gel will get 30 percent confidence and 8.8 percent support when purchasing Castor oil. The goal of this research was to boost revenue for the shop owners by improving sales strategies and product promotion based on the associations discovered.

6. Mining Interesting Rules by Association and Classification Algorithms, 2009, Yanthy et al.

Yanthy et al., says data mining's main purpose is to uncover hidden knowledge from data, and different algorithms have been presented so far. However, not all rules are interesting, and only a small percentage of the generated rules would be of interest to any individual user. As a result, a variety of metrics have been proposed to select the best or most interesting rules, including confidence, support, lift, knowledge gain, and so on. Some algorithms, on the other hand, are good at creating rules that are high in one interestingness metric but low in others. The relationship between the algorithms and the resulting rules' interestingness measurements is not yet evident. The link between algorithms and interesting measurements was investigated in this work. They used synthetic data to ensure that the final outcome is accurate.

7. Market Basket Analysis of Library Circulation Data, 2007, Cunningham et al.

Cunningham et al. established a model for library circulation data, which they use here. The task of discovering subject classification categories that co-occur in transaction records of books borrowed from a university library is tackled using the a-priori market basket method. This information can be important in referring visitors to other parts of the collection that may include papers relevant to their information needs, as well as deciding the physical layout of a library. These findings can also reveal the degree of "scatter" that a classification technique causes in a given collection of documents.

8. Market basket analysis of student attendance records, 2019, Hussain and Hussein.

Hussain and Hussein used a data mining approach, involving the market basket analysis, using the data on student attendance. This analysis contributes towards identifying the student groups who have nearly matching absence records. This kind of similarity may highlight that the students miss classes owing to peer pressure, instead of acceptable reasons. This technique has been tested through the analysis of the student attendance data for more than two thousand students attending a public senior educational institution for a period of one semester. The obtained results were useful in finding the students who are missing classes just because their friends are missing the classes

9. Market Basket Analysis: Trend Analysis Of Association Rules In Different Time Periods, 2015, G. Kapadia.

G. Kapadia conducted a study in 2015 that examined the pattern of consumer behaviour for lifestyle shop products. It provides useful information about the basket's formation. This research aided in product assortments, stock management for likely things sold, promotions for likely items sold, loyal customer discounts, and cross-selling. The study's scope was limited to one store in a certain region, which was a drawback. In the sphere of education, data mining tools are also used. Om Prakash Chandrakar used association rule mining in April 2015 to examine student performance in examinations and anticipate the outcome of upcoming exams. The study's scope was limited to one store in a certain region, which was a drawback.

10. Market Basket Analysis of Consumer Buying Behaviour of a Lifestyle Store, 2011, S. Prakash and R.M.S Parvathi

S. Prakash and R.M.S. Parvathi (2011) suggest a qualitative strategy to mining quantitative association rules. Because the method translates numerical attributes to binary attributes, the proposed methodology is qualitative.

Finding qualitative rules, on the other hand, is the focus of this study. Decision trees, patterns, and dependency tables are the most frequent ways to describe these principles. (William Frawley, 1991, Gregory Piatetsky Shapiro) Categorical attributes are the type of attributes used to mine qualitative rules.

11. Mining Association Rules between Sets of Items in Large Databases, 1993, Rakesh Agrawal, Tomasz Imielinski, and Arun Swami.

Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, 1993 is one of the first articles on association rules to offer a rule mining approach for discovering qualitative rules with no boolean attribute restrictions. The authors put the Association Rules algorithm to the test by using data from a large retailer to see how effective it is.

Many studies have been done on mining sequential patterns in a static database. Agarwal and Srikant were the first to address it. Sequential pattern mining methods are traditionally divided into three categories. Horizontal partitioning approaches that are based on prior knowledge, such as Generalized Sequential Pattern mining, which uses a multiple-pass candidate generation and test methodology in sequential pattern mining. Vertical partitioning methods based on apriori properties, such as Sequential Pattern Discovery using Equivalent classes, use combinatorial properties to break down the original problem into smaller sub-problems that can be solved independently in main memory using efficient lattice search and simple join operations. Prefix-projected sequential pattern mining algorithms, for example, are projection-based pattern growth algorithms that illustrate pattern growth methodology and locate common items after scanning a database once.There are a variety of algorithms available, including closed sequential pattern mining, maximal sequential pattern mining, and constraint sequential pattern mining, in addition to the standard approaches.

12. Sequential Mining: Patterns And Algorithms Analysis, 2006, Parthasarathy et al.

The incremental sequential pattern mining algorithms address fundamental shortcomings of sequential pattern mining algorithms, such as mining patterns from up-to-date databases without removing obsolete patterns. The following are the key incremental sequential pattern mining algorithms: Parthasarathy et al. created the ISM incremental mining algorithm by using an existing database's sequence lattice. All frequent sequences and sequences in the negative boundary are included in the sequence lattice. When additional transactions are added to the database, Masseglia et al suggested another incremental technique, ISE, for incremental mining of sequential patterns. The candidate generation and testing strategy is used in this algorithm. Hang Cheng et al. presented Incspan, a sequential pattern mining approach for incremental databases. The inability of these algorithms to discard obsolete data is a drawback.

Progressive sequential pattern mining is a generalised pattern mining technique that identifies the most current frequently occurring sequential patterns. This approach is unaffected by the presence of old data and works with both static and dynamically evolving databases. The old data has little effect on the patterns. The sliding window is used in this approach to update sequences in the database and aggregate the frequencies of candidate sequential patterns as time goes on. The time stamp across which the algorithm is currently running is determined by the sliding window called period of interest.

13. Market Basket Analysis to Identify Customer Behaviors by Way of Transaction Data, 2009, Kanagawa, Matsumoto, Koike and Imamura.

Market Basket Analysis was probably initially employed by Agrawal, Imieliksi, and Swami (1993) (Agrawal, n.d.) who had a vast collection of consumer transaction data previously collected and the association rules between commodities purchased were discovered. The method was quickly adopted as a standard way for a variety of practical applications in the marketing area (Chen et al., 2005). Kanagawa, Matsumoto, Koike, and Imamura published surveys in which respondents answered open-ended questions about allergic foods (Kanagawa et al., 2009). As a result, scientists discovered that specific food allergies appear to occur in the same individual at the same time. (2017) According to Russell et al. (1999) (Singh and Sinwar, 2017), marketing researchers could utilise Market Basket Analysis to construct multi-category decision-making, theoretical model purchasing decisions involving items.

2.2 METHODOLOGY

The aim of a recommending system is to produce meaningful recommendations for items or products of interest to a collection of users. The fundamental algorithm such as Apriori and FP Growth, collects knowledge about the preferences of people and recognizes that when people buy spaghetti and wine, they are often generally interested in gravies. Association rule is the key part in developing a recommendation engine. The Association Rule produces a number of rules after running on a data set with details from past shopping baskets. Each rule includes a product name collection as an antecedent, one product name as a consequence, and a few class measures, such as antecedent support, consequent support, support, confidence and lift.

Research Approach

The present technological time that we live in has made it feasible for business organisations to accumulate extensive data. Currently, database technology innovation has sufficiently grown to keep these information stacks solid, however, it is significant not to simply keep that information, yet to assess the information to increase the value of the organisation. In today's customer-centered markets, business needs to establish adequate and low advertising techniques that can react to changes in customer perceptions and demands for products. It might also assist business to recognise a whole new market strategy that can effectively target. All together for making key choices on the market strategy, stable, as much as could be and secured proof-based data is needed. With innovation, Data Mining has gotten perhaps the best response to this requirement. Data Mining is the process of refining important data from enormous databases which includes a tremendous assortment of statistical and computational methods, neural network analysis, clustering, classification and summing up information. The computation of association rules, which is one of the Data Mining techniques implemented by Market Basket Research, is part of this project. The analysis is carried out on the Grocery stores' transaction data for the customers. The research goal is to consider the category of product that is likely to be marketed in conjunction by implementing Apriori.

Research purpose

The ultimate purpose of every business is to find better ways to improve the profit for a long run. But, for this research, the aim would be to encountering actual case of dependencies among products chosen by customer. Though several different products could be bought in a single visit to a store like, groceries, we believe that there are no coincidences for these choices. These decisions from several categories results in forming customer's shopping basket. Which with-holds the collection of categories that customer purchased on a specific shopping trip.

Nowadays, Machine Learning is helping the Retail Industry in many different ways. You can imagine that from forecasting the performance of sales to identify the buyers, there are many applications of machine learning (ML) in the retail industry. "Market Basket Analysis" is one of the best applications of machine learning in the retail industry. By analyzing the past buying behavior of customers, we can find out which are the products that are bought frequently together by the customers.

A market basket analysis can provide the information your company's marketing team needs to create more accurate marketing and advertising campaigns. They can use customer purchasing habits to choose which items to group in advertisements, which may attract more customers and increase sales. Market basket analysis helps identify items that have a close affinity to each other even if they fall into different categories. With the help of this knowledge, retailers can place the items with higher affinity close to each other to increase the sale.

Why did we choose MBA

A market basket analysis is a technique that analysts use to look for relationships between items and identify trends. Typically, a market basket analysis shows which items customers frequently buy together, which helps analysts predict what new customers may purchase.

Reasons to choose MBA:

1. Enhance customer satisfaction

Companies may use market basket analyses to increase customer satisfaction by using data about shopping trends to improve the customer experience. Retail analysts may use this information to help their company organize its store effectively, which may make it easier for customers to find the items they plan to buy.

We can also use data from a market basket analysis to create coupons or promotions that fit customer wants and needs, which may increase customer satisfaction. For example, if you conduct a market basket analysis for a grocery store and notice that a majority of customers buy a certain brand of paper towels, you could recommend the store create a sale for that product. This temporary price adjustment may boost customer satisfaction and increase sales.

2. Increase cross-selling

Retail analysts can use the information they learn from market basket analyses to increase sales through strategies such as cross-selling. Cross-selling refers to the practice of displaying items near each other in a store to encourage customers to buy them together. For example, an analyst might recommend that a clothing retailer display a matching purse, belt and hat on the same mannequin to motivate customers to purchase the products at the same time.

3. Improve advertisements

A market basket analysis can provide the information your company's marketing team needs to create more accurate marketing and advertising campaigns. They can use customer purchasing habits to choose which items to group in advertisements, which may attract more customers and increase sales. For example, if an apparel company's customers often purchase umbrellas and rain boots at the same time, the marketing team could create an advertisement that includes both items.

4. Adjust store layouts

Your company may use market basket analyses to configure its store layout effectively, which may encourage customers to purchase additional products. A company might adjust the layout of its store by moving sections of similar items, reorganizing shelves or creating temporary product displays. For example, a supermarket might move its frozen goods section to the back of the store. This can encourage customers to shop in other sections before choosing frozen products, which may increase sales.

What is Association Rule Learning

The association rule learning is a rule-based machine learning approach that generates the relationship between variables in a dataset. It has major applications in the retail industry including E-Commerce retail businesses. Using this strategy, the products sold in an association can be explored and can be offered to customers to buy together. For example, it can be discovered that if the customers have bought onion and potato together, then most likely they have bought tomato also. It can be given a rule in the form of {onion, potato} -> tomato. These rules are called association rules. The Association Learning methods discover these types of rules from the dataset comprising the list of transactions.

What is Apriori Algorithm in Market Basket Analysis?

Apriori is a popular algorithm used in market basket analysis. This algorithm is used with relational databases for frequent itemset mining and association rule learning. It uses a bottom-up approach where frequent items are extended one item at a time and groups of candidates are tested against the available dataset. This process continues until no further extensions are found. It uses the concept of Support, Confidence and Lift where,

Support (items I) = (total no. Of transaction with item I) / (total no. Of transaction)

Confidence (item I_1 ---> item I_2) = no of transaction with item I_1 and item I_2 / no of transaction with item I_1

Lift (item I_1 ---> item I_2) = confidence (item I_1 ---> item I_2) / support (item I_2)

Apriori algorithm

The Apriori algorithm was the first associative algorithm proposed, and it has since been employed in future developments in assocuation, classification, and associative classification algorithms. The Apriori algorithm counts transactions on a level-by-level, breadth-first basis. Prior knowledge of common item set properties is used in the apriori process. Apriori employs an iterative method known as level-wise search, in which n-item sets are utilised to investigate (n+1)-item sets. The Apriori property is used to improve the efficiency of the level-wise production of frequent item collections. All non-empty subsets of a frequent item set must likewise be frequent, according to the apriori property. Because of the anti-monotone quality of support measure, the support for an item set never exceeds the support for another item set.

Antecedent: The items on the LEFT ie., the item which the customer buy.

Consequent: The items on the RIGHT ie., the item which the customer follows to buy.

Association Rule

Association rule is one of the most important Data Mining techniques used in Market Basket Analysis. All fruits are sorted in the same aisle in a Super Market, all dairy products are placed together under another aisle. Hence spending time and intentionally investing resources to place the most necessary items in an organised way not only reduces a shopping time of customers, but also helps customers to purchase the most appropriate items one might be keen in clubbing in their Market Basket. Association rule is related to the statement of "what goes with what". The purchase of products by customers at Super Market are termed as "Transactions. The magnitude of an associative rule can be derived in the existence of three parameters, namely support, confidence and lift(Kanagawa et al., 2009).

Item Sets

Item sets is the collection of all items in a market basket data, I = { i1, i2.. in }. Transaction is the group of all transactions, T= {t1, 12..tn} Every transaction is a specific one and makes up a collection of items from item 1. When there are n items in an itemset it is called n-itemset. For example, it is called as 3-itemset {Oranges, Apple, Bread}. If an itemset doesn't have an item, it is called the set as null (empty). Presenting the itemset with more than one item significantly expands the chances of the rules to be listed.

Support Count

The width of the transaction is measured by the number of items contained in a transaction. Support Count, which refers to the number of transactions involving a particular itemset, is an important component of an itemset.

Support of an item or set of item is that the fraction of transactions in our data set that contains number of that particular item product to total number of transactions. Support gives an idea of how many times an itemset has occurred in the overall transactions. For example, in a retail shop, if we consider 100 customers had visited to purchase products. It was seen that out of 100 customers, 50 of them purchased Product A, 40 of them purchased Product B and 25 of them purchased both Product A and Product B. Support of Product A is 50%, Support for Product B is 40% and Support of Product A and B is 25%. Value of Support helps in considering the rules, which are worth for further analysis on correlation of a products with other existing products in the store. For instance, if we want to note the item sets which occur at least 50 times in 10,000 transactions, then support = 0.005. With low support value, we will not have enough information on how the products are related to each other and thus helps to find "hidden" relationships.

$$Support\ A = \frac{(Number\ of\ Transaction\ that\ Contains\ A)}{(\ Total\ Transaction)}$$

Equation 1: To find Support of an item

Confidence

Confidence is a measure of the likelihood that customer buy product A will buy product B as well. A rule of association is therefore a remark of the form (item set A) (item set B) where A is the precedent and B is the consequence. Confidence gives the probability of Consequence occurring on the cart provided with pre-existing antecedents. For frequently appearing Consequent, it doesn't matter what the customer have it in the Antecedent. The confidence of an Association rule, which results very often, will always be of great value. For example,

$$Confidence(A => B) = P(A \mid B)$$

$$= \frac{(\ Number\ of\ Transaction\ that\ Contains\ A\ and\ B\ )}{(Total\ Transaction\ that\ Contains\ A)}$$

Equation 2: To find confidence between two items

Of the 50 customers who purchased Product A, 25 have purchased Product B. It ensures if somebody buys product A, they could buy product B by the probability of 50%.

Lift

In comparison to a random selection of a transaction, the ratio shows how effective the rule is in finding consequences. In general, lifting is higher than one implies that the rule has some usefulness. A lift greater than one indicates that the presence of A has increased the probability of generating B in this transaction. The role of A has decreases the chances the role of product B occurrence if lift value is smaller than one.

$$\frac{Lift\ (A => B) =\ (Confidence(A => B))}{(Support(B)\ )}$$

Equation 3 : To find Lift for the effectiveness of rule

A lift value of 1.25 implies that chance of purchasing product B would increase by 25%. Fundamental rules of association grant the phenomenon of one item and the inclusion of another. Through the use of Data Analytics, the process of trying to uncover association rules includes the following steps:

Step 1: Set up the data in the transaction presentation. An association algorithm demands input information to be arranged in transaction format tx={i1,i2,i3}

Step 2: Short-list collection of items that often occur. Item sets are object aggregation. An association algorithm considers the most commonly occurred items and excludes the least occurred items, making increasingly relevant the ultimate rule that will be pulled to the next level.

Step 3: Generate the association rules applicable from the item sets. Ultimately the algorithm produces and filters rules based on the measurements being tuned.

Product Recommendation Using Association Rule

In order to recommend a product, the main aim of data mining is to create a model. The model builders must derive information from historical data and represent it in such a way as to be able to adapt the resulting model to new situations. The data sets analysis process extracts useful information on which to apply one or more data mining techniques to discover previously unknown patterns within the data, or find trends in the data which can then be used to recommend trends or behaviour patterns. It is human nature to know what the future holds and to advise. The recommendation covers the forecast of future events by using sophisticated methods such as machine learning based on historical data observed previously. Through using different strategies such as sampling, correlating and so on, historical data is obtained and transformed.

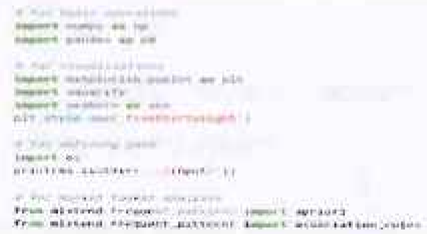
Recommendation system can be split into four stages:

1. The collection and pre-processing of raw data;
2. Convert pre-processed data into an easily achievable form using the selected machine learning method such as Apriori.
3. Create a model of learning (training) using transformed data;
4. Use the previously developed set of association rules to report recommendations to the user:

III. DATA ANALYSIS AND ANALYTICS

3.1 DATA ANALYSIS

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.



```
# for basic operations
import numpy as np
import pandas as pd

# for visualizations
import matplotlib.pyplot as plt
import squarify
import seaborn as sns
plt.style.use('fivethirtyeight')

# for defining path
import os
print(os.listdir('../input'))

# for market basket analysis
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

In the above code, we are importing various libraries which are required in our project. Such as numpy

- NumPy (pronounced /ˈnʌmpaɪ/ (NUM-py) or sometimes /ˈnʌmpi/ (NUM-pee)) is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of hig-hlevel mathematical functions to operate on these arrays. or sometimes /ˈnʌmpi/ (NUM-pee)) is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of hig-hlevel mathematical functions to operate on these arrays.

pandas – We can use pandas to perform various tasks like filtering our data according to certain conditions or segmenting and segregating the data according to preference etc.

matplotlib.pyplot- matplotlib is a cross-platform, data visualization and graphical plotting library for python. Pyplot is a collection of functions that make matplotlib work like MATLAB.

Squarify is library in python which can be used to build treemaps.

Seaborn – is a data visualization library built on top of matplotlib and closely integrated with pandas data structure in Python. Basically, is used for data visualization and exploratory data analysis. It is also used for predictions.

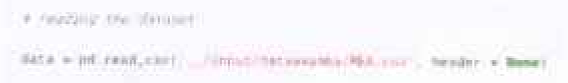
Plt.style.use is used to add style to the plot.

The OS module in Python provides functions for creating and removing a directory (folder), fetching its contents, changing and identifying the current directory, etc. You first need to import the os module to interact with the underlying operating system.

Mlxtend (machine learning extensions) is a Python library of useful tools for the day-to-day data science tasks.

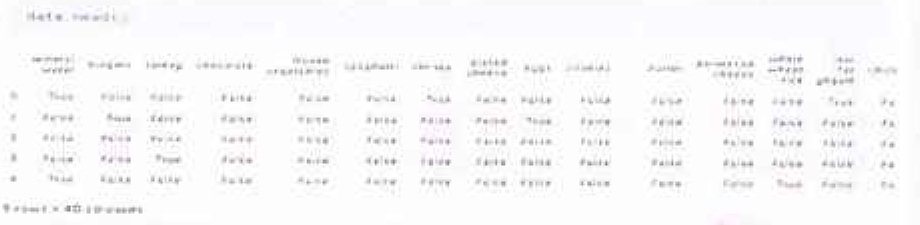
Apriori function to extract frequent itemsets for association rule mining from mlxtend.frequent_patterns import apriori

Function to generate association rules from frequent itemsets from mlxtend.frequent_patterns import association_rules



```
# reading the dataset
data = pd.read_csv('../input/[illegible]/MBA.csv', header = None)
```

It is used to read CSV file using pandas



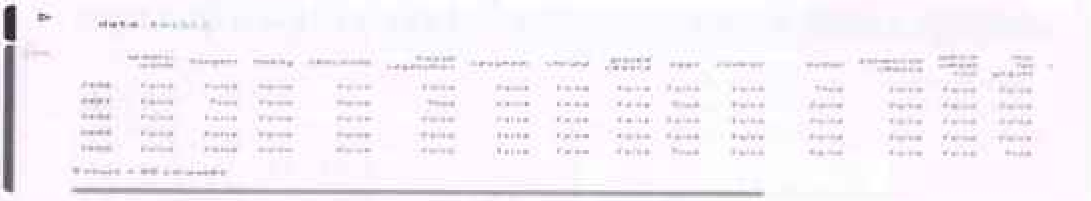
```
data.head()
```

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0	True	False	False	False	False	False	True	False	False	False	False	False	False	True	Fa
1	False	True	False	False	False	False	False	False	True	False	False	False	False	False	Fa
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	Fa
3	False	False	True	False	False	False	False	False	False	False	False	False	False	False	Fa
4	True	False	False	False	False	False	False	False	False	False	False	False	True	False	Fa

5 rows × 40 columns

Data.head() method returns top 5 rows of a DataFrame.

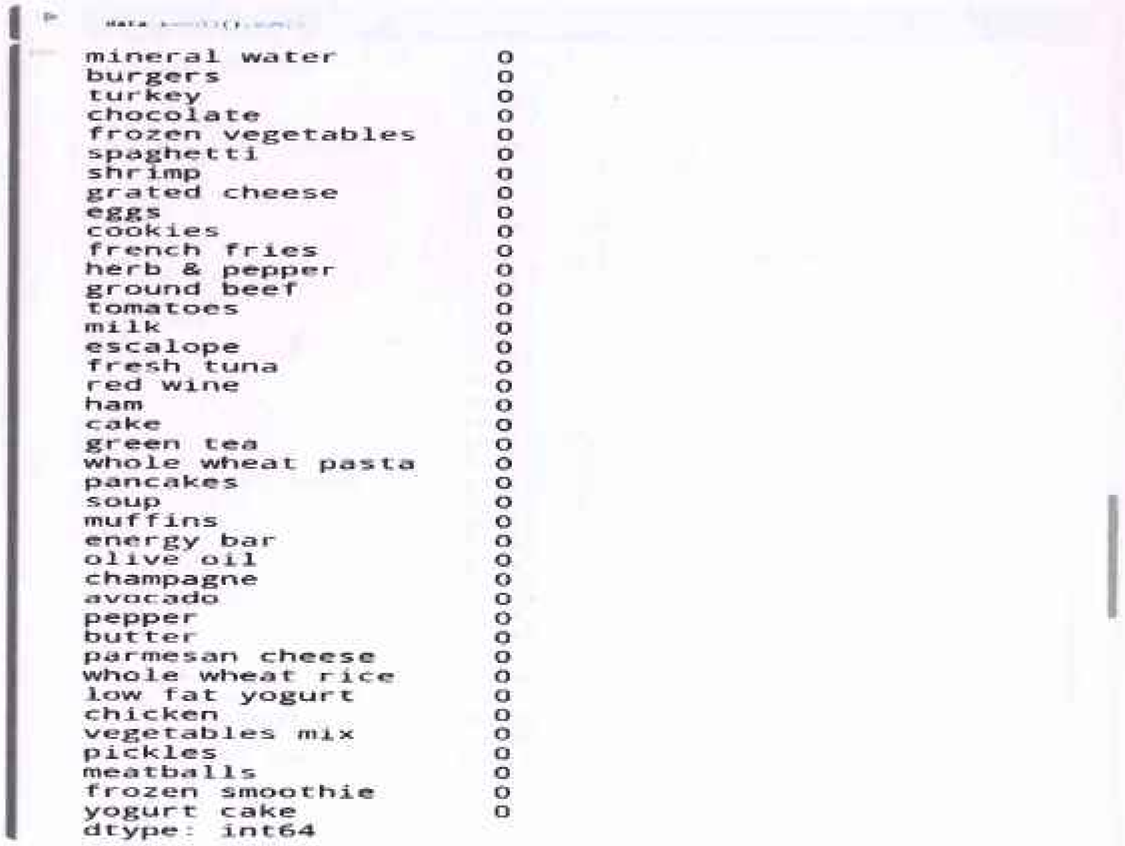
Top 5 rows of data frame/set are returned(displayed). In this, we can observe that numbering starts from 0. So 0 to 4 (total 5 Rows) and all columns (total 40 Columns) and their names Such as mineral water, burgers, turkey, chocolate…… etc.



Data.tail( ) method returns bottom 5 rows of DataFrame.

Bottom 5 rows of data frame/set are returned(displayed). In this we can observe that numbering starts from 7496. So 7496 to 7500 (total 5 Rows) and all columns (total 40 Columns) and their names Such as mineral water, burgers, turkey, chocolate, etc.

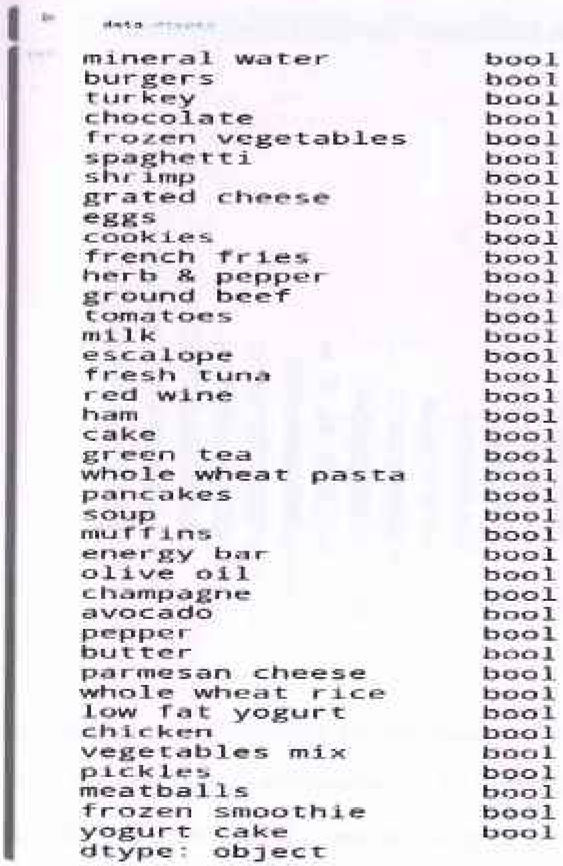
The describe() method is used for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame.



```
data.isnull().sum()
```

```
mineral water        0
burgers              0
turkey               0
chocolate            0
frozen vegetables    0
spaghetti            0
shrimp               0
grated cheese        0
eggs                 0
cookies              0
french fries         0
herb & pepper        0
ground beef          0
tomatoes             0
milk                 0
escalope             0
fresh tuna           0
red wine             0
ham                  0
cake                 0
green tea            0
whole wheat pasta    0
pancakes             0
soup                 0
muffins              0
energy bar           0
olive oil            0
champagne            0
avocado              0
pepper               0
butter               0
parmesan cheese      0
whole wheat rice     0
low fat yogurt       0
chicken              0
vegetables mix       0
pickles              0
meatballs            0
frozen smoothie      0
yogurt cake          0
dtype: int64
```

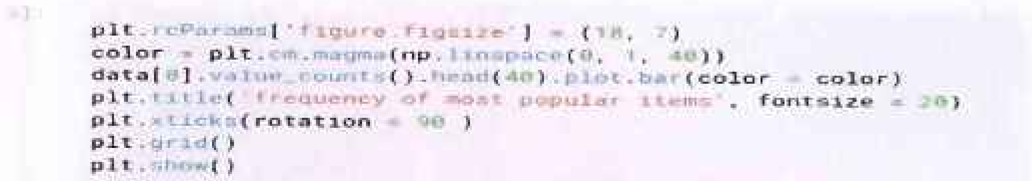
isnull(). sum(). sum() returns the number of missing values in the data set.



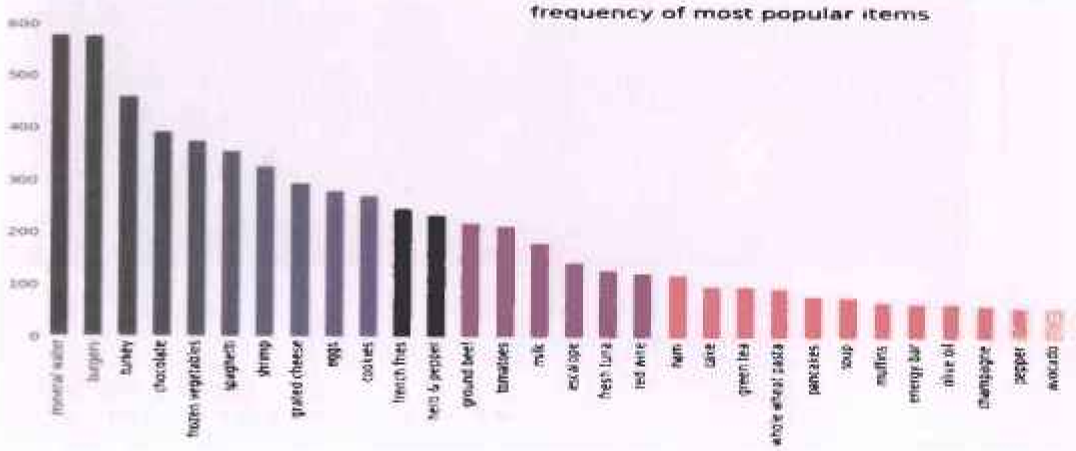
```
data.dtypes
```

```
mineral water          bool
burgers                bool
turkey                 bool
chocolate              bool
frozen vegetables      bool
spaghetti              bool
shrimp                 bool
grated cheese          bool
eggs                   bool
cookies                bool
french fries           bool
herb & pepper          bool
ground beef            bool
tomatoes               bool
milk                   bool
escalope               bool
fresh tuna             bool
red wine               bool
ham                    bool
cake                   bool
green tea              bool
whole wheat pasta      bool
pancakes               bool
soup                   bool
muffins                bool
energy bar             bool
olive oil              bool
champagne              bool
avocado                bool
pepper                 bool
butter                 bool
parmesan cheese        bool
whole wheat rice       bool
low fat yogurt         bool
chicken                bool
vegetables mix         bool
pickles                bool
meatballs              bool
frozen smoothie        bool
yogurt cake            bool
dtype: object
```

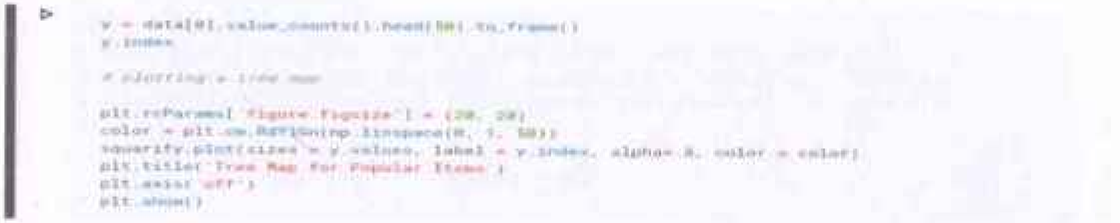
dtypes attribute to find out the data type of each column in the given dataframe.



```
plt.rcParams['figure.figsize'] = (18, 7)
color = plt.cm.magma(np.linspace(0, 1, 40))
data[0].value_counts().head(40).plot.bar(color = color)
plt.title('frequency of most popular items', fontsize = 20)
plt.xticks(rotation = 90 )
plt.grid()
plt.show()
```

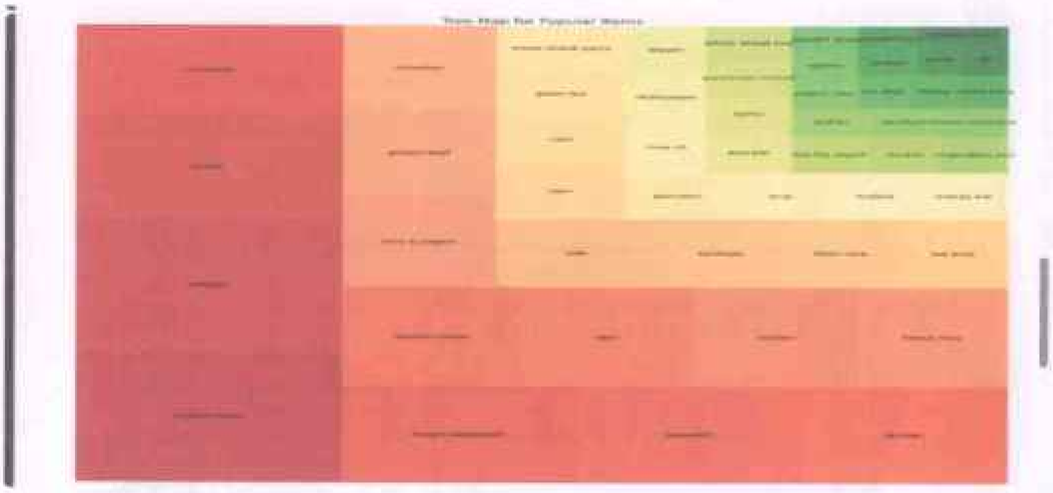


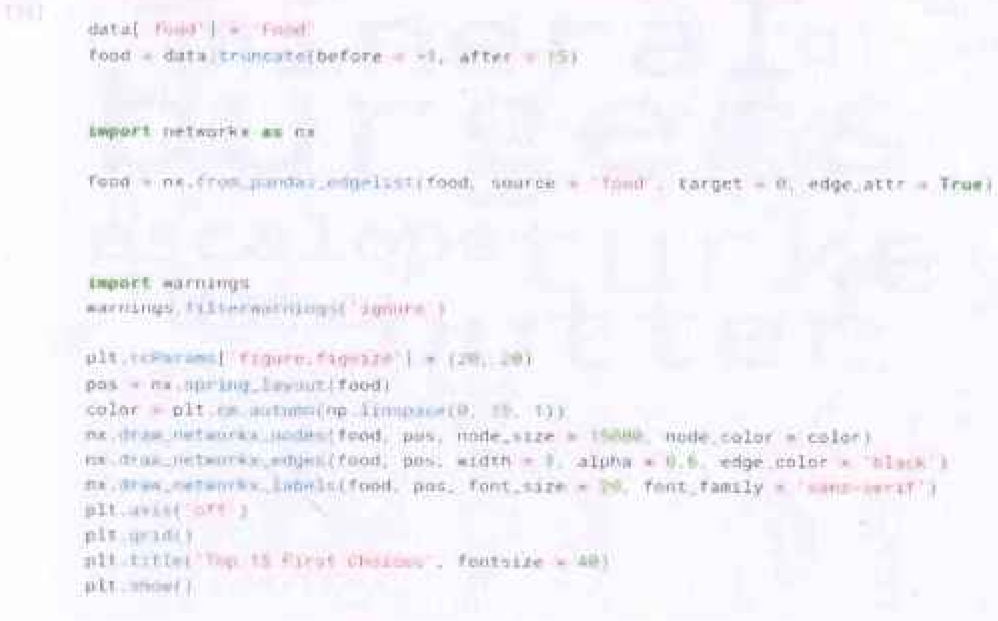
In the above groph we can see the frequency of each product (No. of transaction). At one side we have products in descending order based on frequency (ranging o to 600) in the right side of the graph we can see that the frequency of the products is less and on the left side the frequency of the products is high.



```
y = data[0].value_counts().head(50).to_frame()
y.index

# plotting a tree map

plt.rcParams['figure.figsize'] = (20, 20)
color = plt.cm.RdYlGn(np.linspace(0, 1, 50))
squarify.plot(sizes = y.values, label = y.index, alpha=.8, color = color)
plt.title('Tree Map for Popular Items')
plt.axis('off')
plt.show()
```

In the above code we are going to represent a tree map for popular items bought by the user.

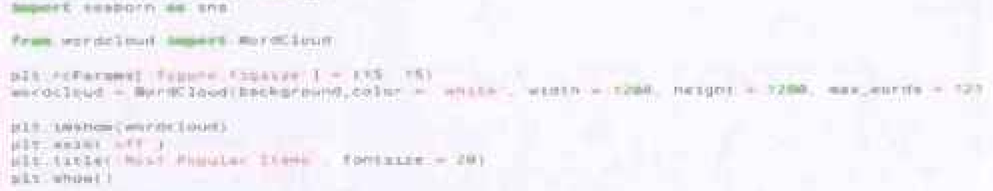




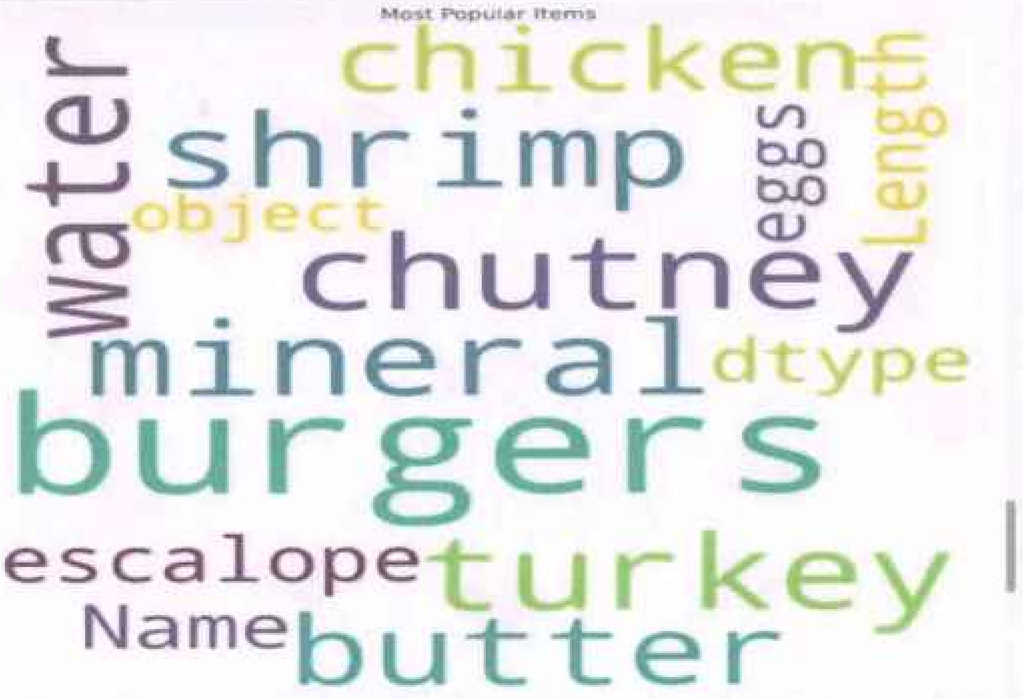
```
data['food'] = 'Food'
food = data.truncate(before = -1, after = 15)


import networkx as nx

food = nx.from_pandas_edgelist(food, source = 'food', target = 0, edge_attr = True)



import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(food)
color = plt.cm.autumn(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(food, pos, node_size = 15000, node_color = color)
nx.draw_networkx_edges(food, pos, width = 3, alpha = 0.6, edge_color = 'black')
nx.draw_networkx_labels(food, pos, font_size = 20, font_family = 'sans-serif')
plt.axis('off')
plt.grid()
plt.title('Top 15 First Choices', fontsize = 40)
plt.show()
```

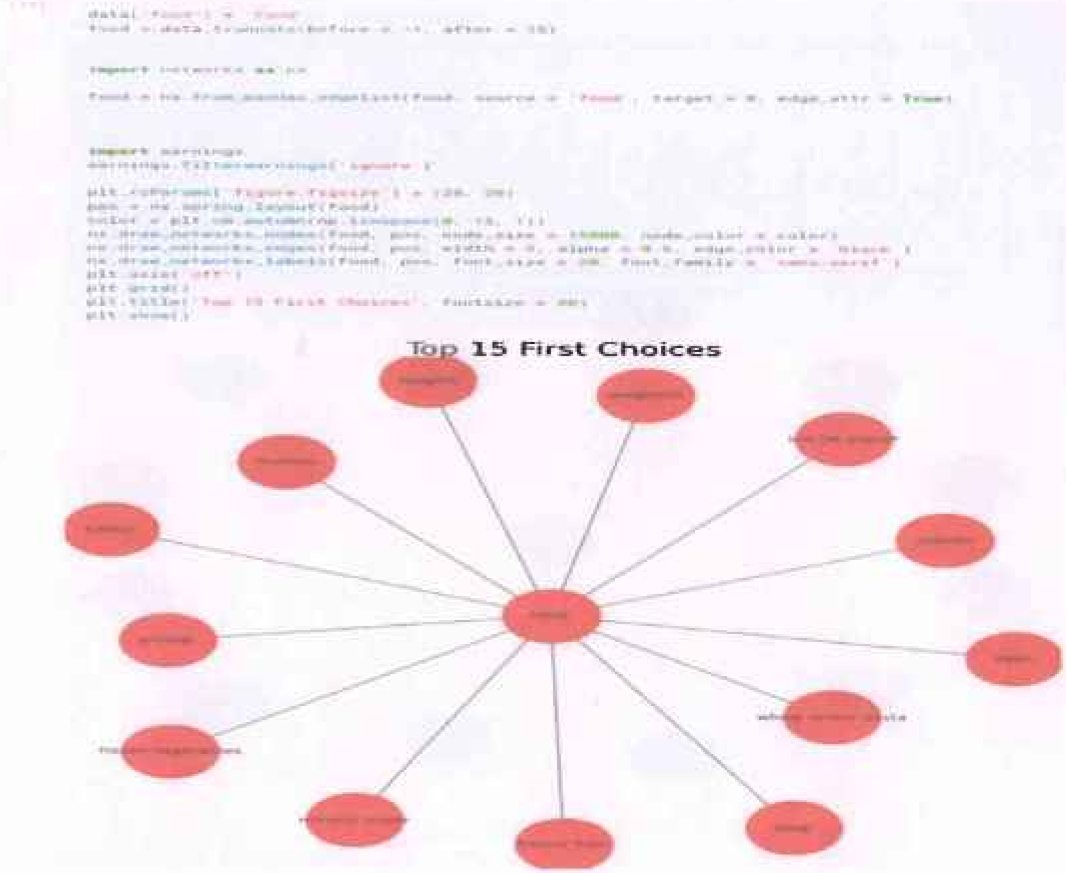
from_pandas_edgelist(df, source='source', target='target', edge_attr=)

Returns a graph from Pandas DataFrame containing an edge list.



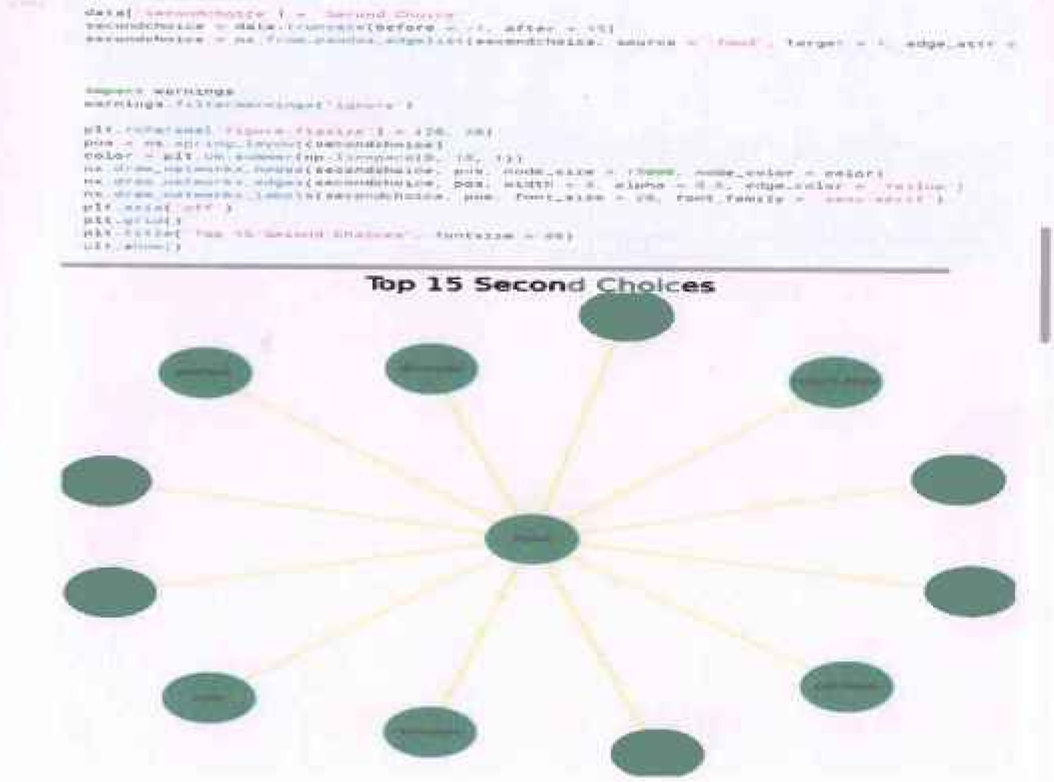
```
import matplotlib.pyplot as plt
import seaborn as sns

from wordcloud import WordCloud

plt.rcParams['figure.figsize'] = (15, 15)
wordcloud = WordCloud(background_color = 'white', width = 1200, height = 1200, max_words = 121)

plt.imshow(wordcloud)
plt.axis('off')
plt.title('Most Popular Items', fontsize = 20)
plt.show()
```





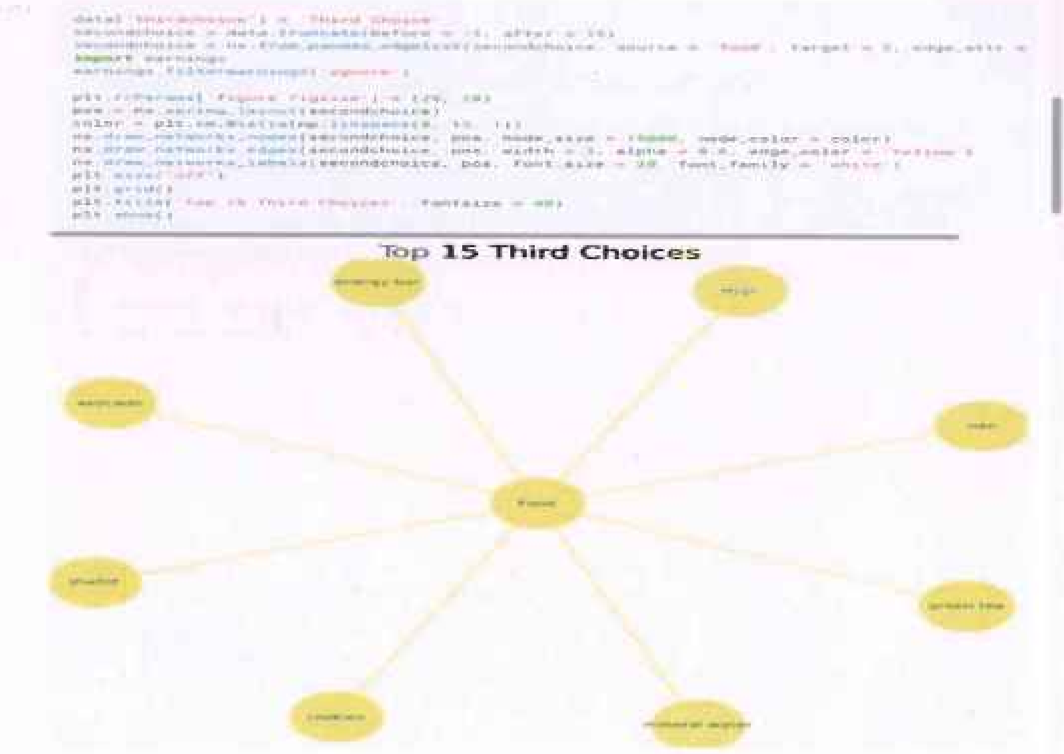
```
data['food'] = 'food'
food = data.truncate(before = -1, after = 15)

import networkx as nx

food = nx.from_pandas_edgelist(food, source = 'food', target = 0, edge_attr = True)


import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(food)
color = plt.cm.autumn(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(food, pos, node_size = 15000, node_color = color)
nx.draw_networkx_edges(food, pos, width = 3, alpha = 0.6, edge_color = 'black')
nx.draw_networkx_labels(food, pos, font_size = 20, font_family = 'sans-serif')
plt.axis('off')
plt.grid()
plt.title('Top 15 First Choices', fontsize = 40)
plt.show()
```

First we are assigning only food column in the food variable. Then, we truncated the food column with (shorting the data by assigning before equals to minus one and after equals to 15) Then we imported networks library as nx then we updated food variable where we assigned food and source in one column food. Then we plotted a network graph by assigning food as a main node and their respected (first 15) attributes/choices.



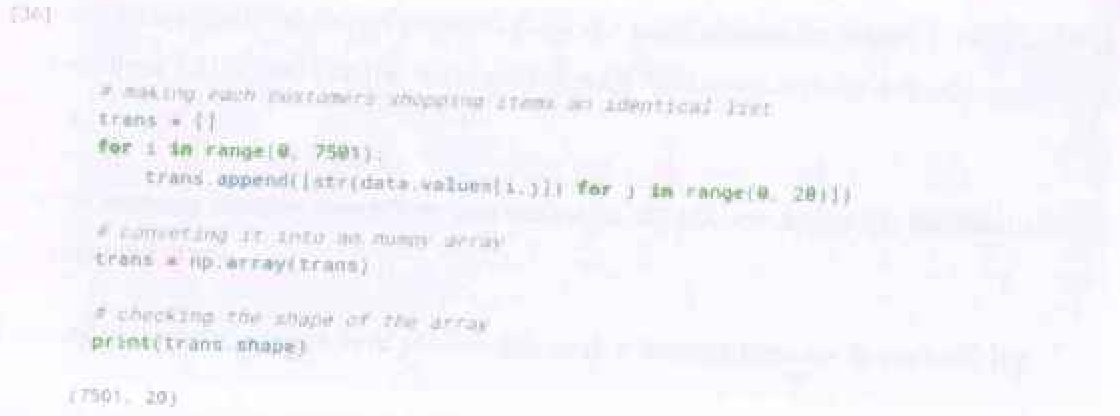
```
data['secondchoice'] = 'Second Choice'
secondchoice = data.truncate(before = -1, after = 15)
secondchoice = nx.from_pandas_edgelist(secondchoice, source = 'food', target = '', edge_attr =

import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(secondchoice)
color = plt.cm.summer(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(secondchoice, pos, node_size = 15000, node_color = color)
nx.draw_networkx_edges(secondchoice, pos, width = 3, alpha = 0.6, edge_color = 'yellow')
nx.draw_networkx_labels(secondchoice, pos, font_size = 20, font_family = 'sans-serif')
plt.axis('off')
plt.grid()
plt.title('Top 15 Second Choices', fontsize = 40)
plt.show()
```

Represent second 12 choices.

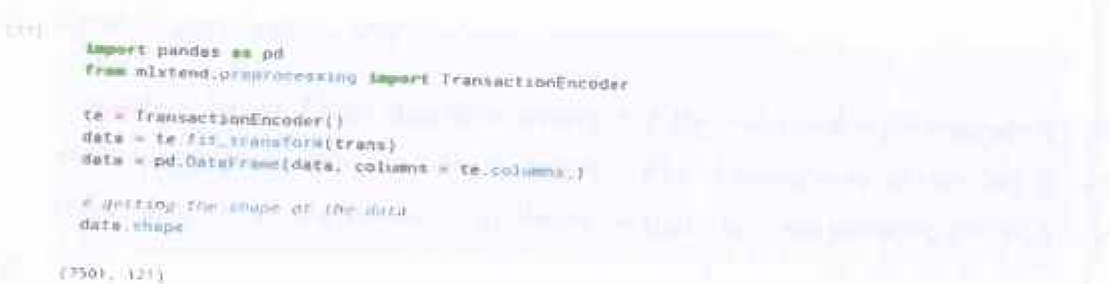


```
data['thirdchoice'] = 'Third Choice'
secondchoice = data.truncate(before = -1, after = 15)
secondchoice = nx.from_pandas_edgelist(secondchoice, source = 'Food', target = 3, edge_attr =
import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(secondchoice)
color = plt.cm.Wistia(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(secondchoice, pos, node_size = 1500, node_color = color)
nx.draw_networkx_edges(secondchoice, pos, width = 3, alpha = 0.6, edge_color = 'yellow')
nx.draw_networkx_labels(secondchoice, pos, font_size = 20, font_family = 'white')
plt.axis('off')
plt.grid()
plt.title('Top 15 Third Choices', fontsize = 40)
plt.show()
```

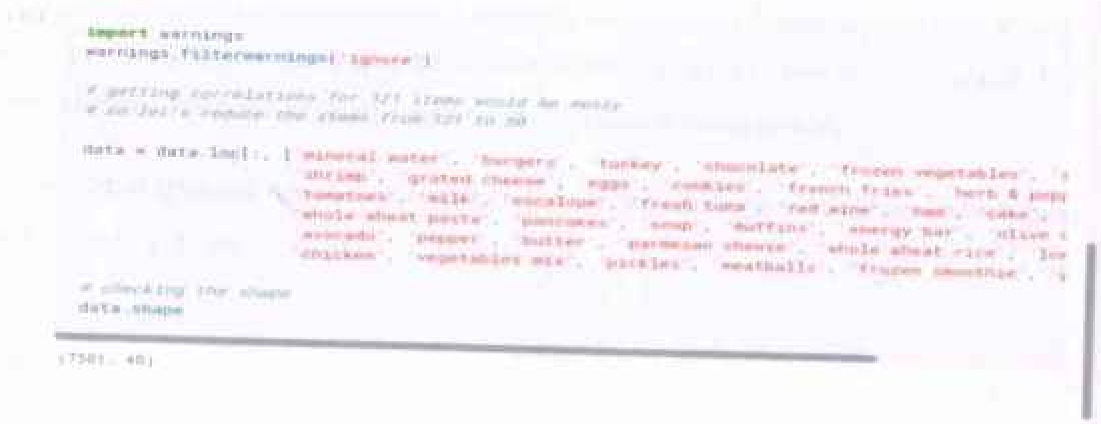
Represent third 8 choices.



```
[36]

    # making each customers shopping items an identical list
    trans = []
    for i in range(0, 7501):
        trans.append([str(data.values[i,j]) for j in range(0, 20)])

    # conveting it into an numpy array
    trans = np.array(trans)

    # checking the shape of the array
    print(trans.shape)

  (7501, 20)
```



```
[37]
    import pandas as pd
    from mlxtend.preprocessing import TransactionEncoder

    te = TransactionEncoder()
    data = te.fit_transform(trans)
    data = pd.DataFrame(data, columns = te.columns_)

    # getting the shape of the data
    data.shape

  (7501, 121)
```



```
    import warnings
    warnings.filterwarnings('ignore')

    # getting correlations for 121 items would be messy
    # so let's reduce the items from 121 to 50

    data = data.loc[:, ['mineral water', 'burgers', 'turkey', 'chocolate', 'frozen vegetables', '
                        'shrimp', 'grated cheese', 'eggs', 'cookies', 'french fries', 'herb & pepp
                        'tomatoes', 'milk', 'escalope', 'fresh tuna', 'red wine', 'ham', 'cake', '
                        'whole wheat pasta', 'pancakes', 'soup', 'muffins', 'energy bar', 'olive o
                        'avocado', 'pepper', 'butter', 'parmesan cheese', 'whole wheat rice', 'low
                        'chicken', 'vegetables mix', 'pickles', 'meatballs', 'frozen smoothie', '

    # checking the shape
    data.shape

  (7501, 40)
```

Data pre-processing

In data pre-processing we gathered data directly from the Kaggle and store it in trans variable. Initially it was having 7501 rows and 20 columns. Then we incorporated these trans variable data into our main dataset and we were getting rows as 7051 and columns as 121.

Then we located only requested columns from our main dataset by reducing the number of columns from 121 to 40. Finally, we were left with 7501 rows and 40 columns in our main dataset.

When a warning matches more than one option in the list, the action for the last matching option is performed.

TransactionEncoder convert item list into transaction data for frequent itemset mining

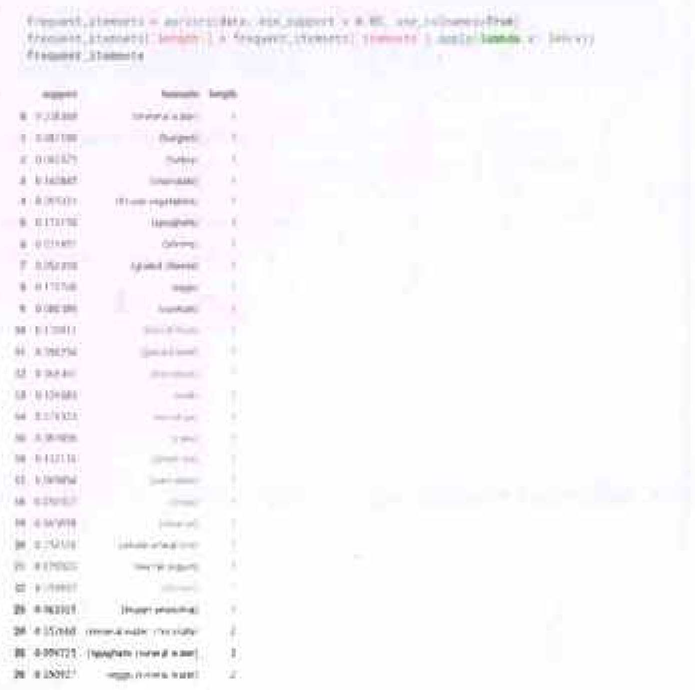
Both -W command-line option and filterwarnings ini option are based on Python's own -W option and warnings.simplefilter

plt.rcParams['figure.figsize'] used to change the size of the figure.

pos = nx.spring_layout These algorithm simulates a force-directed representation of the network treating edges as springs holding nodes close, while treating nodes as repelling objects, sometimes called an anti-gravity force. Simulation continues until the positions are close to an equilibrium.

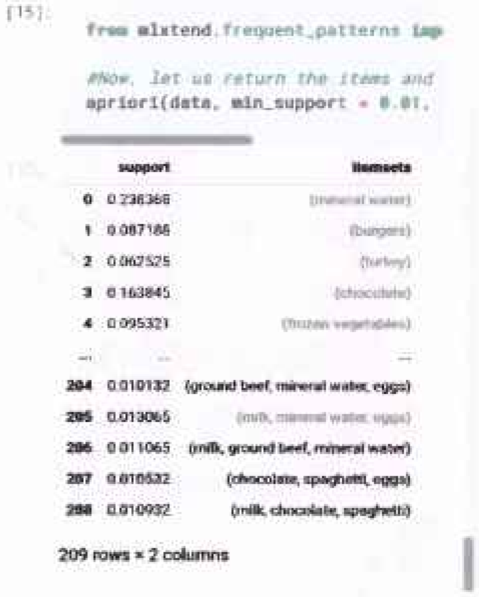
Here we are importing apriori, Apriori Algorithm is a Machine Learning algorithm which is used to gain insight into the structured relationships between different items involved. Then we are returning item and itemsets with atleast 5% support using apriori.

Here we create frequent item sets via apriori and we are adding a new column which stores length of each item sets



```
frequent_itemsets = apriori(data, min_support = 0.05, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x: len(x))
frequent_itemsets
```

	support	itemsets	length
0	0.238368	(mineral water)	1
1	0.087188	(burgers)	1
2	0.062525	(turkey)	1
3	0.163845	(chocolate)	1
4	0.095321	(frozen vegetables)	1
5	[illegible]	(spaghetti)	1
6	[illegible]	(shrimp)	1
7	[illegible]	(grated cheese)	1
8	[illegible]	(eggs)	1
9	[illegible]	(pancakes)	1
10	[illegible]	[illegible]	1
11	[illegible]	[illegible]	1
12	[illegible]	[illegible]	1
13	[illegible]	[illegible]	1
14	[illegible]	[illegible]	1
15	[illegible]	[illegible]	1
16	[illegible]	[illegible]	1
17	[illegible]	[illegible]	1
18	[illegible]	[illegible]	1
19	[illegible]	[illegible]	1
20	[illegible]	[illegible]	1
21	[illegible]	[illegible]	1
22	[illegible]	[illegible]	1
23	[illegible]	(frozen smoothie)	1
24	[illegible]	(mineral water, chocolate)	2
25	[illegible]	(spaghetti, mineral water)	2
26	[illegible]	(eggs, mineral water)	2



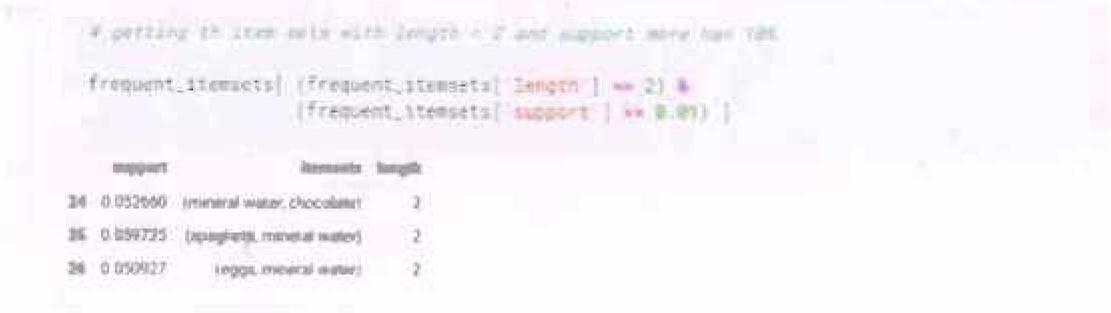
[15]:

```
from mlxtend.frequent_patterns imp

#Now, let us return the items and
apriori(data, min_support = 0.01,
```

	support	itemsets
0	0.238368	(mineral water)
1	0.087188	(burgers)
2	0.062525	(turkey)
3	0.163845	(chocolate)
4	0.095321	(frozen vegetables)
...	...	...
204	0.010132	(ground beef, mineral water, eggs)
205	0.013065	(milk, mineral water, eggs)
206	0.011065	(milk, ground beef, mineral water)
207	0.010532	(chocolate, spaghetti, eggs)
208	0.010932	(milk, chocolate, spaghetti)

209 rows × 2 columns

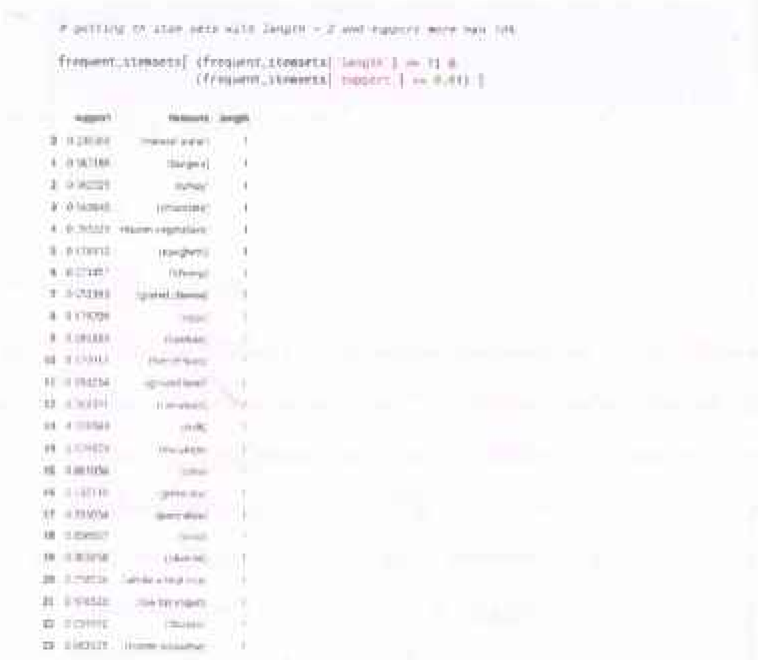
This code returns the item sets with length = 2 and support more than 10%



```
# getting th item sets with length = 2 and support more han 10%

frequent_itemsets[ (frequent_itemsets['length'] == 2) &
                   (frequent_itemsets['support'] >= 0.01) ]
```

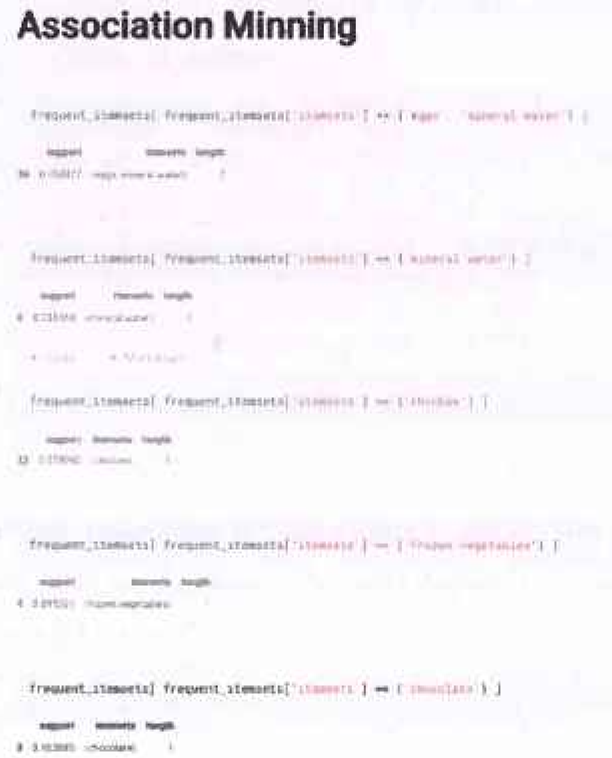
```
     support                     itemsets  length
34  0.052660   (mineral water, chocolate)       2
35  0.059725   (spaghetti, mineral water)       2
36  0.050927        (eggs, mineral water)       2
```

this code returns the item sets with length = 1 and support more than 10%



```
# getting th item sets with length = 2 and support more han 10%

frequent_itemsets[ (frequent_itemsets['length'] == 1) &
                   (frequent_itemsets['support'] >= 0.01) ]
```

```
     support              itemsets  length
 0  0.238368       (mineral water)       1
 1  0.087188             (burgers)       1
 2  0.062525              (turkey)       1
 3  0.163845           (chocolate)       1
 4  0.095321   (frozen vegetables)       1
 5  0.174110           (spaghetti)       1
 6  0.071457              (shrimp)       1
 7  0.052393       (grated cheese)       1
 8  0.179709                (eggs)       1
 9  0.080389             (cookies)       1
10  0.170911        (french fries)       1
11  0.098254         (ground beef)       1
12  0.068391            (tomatoes)       1
13  0.129583                (milk)       1
14  0.079323            (escalope)       1
15  0.081056                (cake)       1
16  0.132116           (green tea)       1
17  0.095054            (pancakes)       1
18  0.050527                (soup)       1
19  0.065858           (olive oil)       1
20  0.058526    (whole wheat rice)       1
21  0.076523      (low fat yogurt)       1
22  0.059992             (chicken)       1
23  0.063325     (frozen smoothie)       1
```



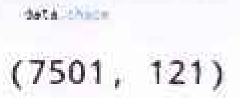
Association Minning

```
frequent_itemsets[ frequent_itemsets['itemsets'] == {'eggs', 'mineral water'} ]

   support              itemsets  length
[illegible] [illegible] (eggs, mineral water)  2

frequent_itemsets[ frequent_itemsets['itemsets'] == {'mineral water'} ]

   support       itemsets  length
[illegible] [illegible] (mineral water)  1

frequent_itemsets[ frequent_itemsets['itemsets'] == {'chicken'} ]

   support  itemsets  length
[illegible] [illegible] (chicken)  1

frequent_itemsets[ frequent_itemsets['itemsets'] == {'frozen vegetables'} ]

   support              itemsets  length
[illegible] [illegible] (frozen vegetables)  1

frequent_itemsets[ frequent_itemsets['itemsets'] == {'chocolate'} ]

   support    itemsets  length
[illegible] [illegible] (chocolate)  1
```

Next we are getting support of each item set that we want.

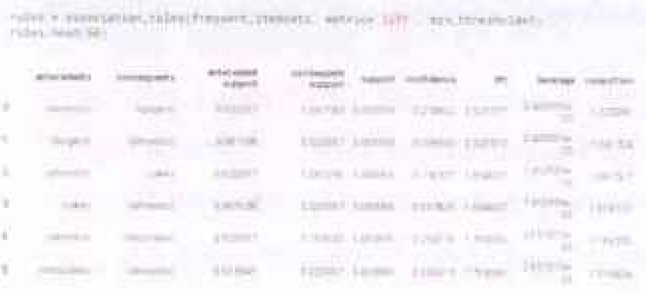
3.2 DATA ANALYTICS

Here we are importing pandas and transaction encoder from mlxtend. Transaction encoder Encodes database transaction data in form of a Python list of lists into a NumPy array. Using TransactionEncoder object, we can transform this dataset into an array format suitable for typical machine learning APIs. Via the fit method, the TransactionEncoder learns the unique labels in the dataset, and via the transform method, it transforms the input dataset (a Python list of lists) into a one-hot encoded NumPy boolean array. Then we turn the encoded array into a pandas DataFrame



```
data.shape

(7501, 121)
```

Now we are again taking the shape of our dataset. Before we had 7501 rows and 20 columns but now we have 121 columns.

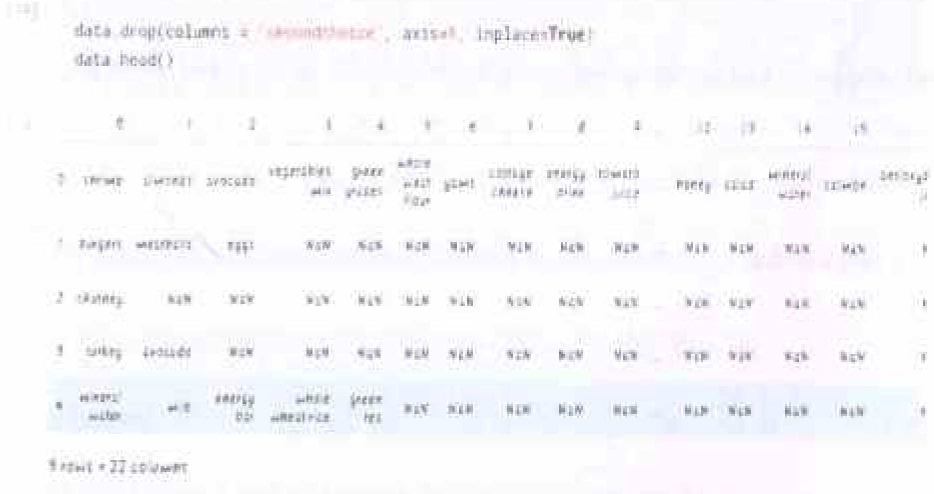


```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
rules.head(50)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

The generate_rules() function allows you to specify your metric of interest and the according threshold. Here we set metric = lift and threshold to 1 because we are only interested in rules that have a lift score of >= 1.

Next we display the first 50 rows using rules.head(50)



```
[ ]  data.drop(columns = 'secondchoice', axis=1, inplace=True)
     data.head()
```

	0	1	2	3	4	5	6	7	8	9	...	12	13	14	15	[illegible]
0	shrimp	almonds	avocado	vegetables mix	green grapes	whole weat flour	yams	cottage cheese	energy drink	tomato juice	...	honey	salad	mineral water	salmon	antioxydant juice
1	burgers	meatballs	eggs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	1
2	chutney	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	1
3	turkey	avocado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	1
4	mineral water	milk	energy bar	whole wheat rice	green tea	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	1

5 rows × 22 columns

With above command, we see that 'secondchoice' has been removed. axis=1 means remove the column.

If we want to delete the columns without having to re-assign the result back to df, we have to specify inplace to True

IV . FINDINGS AND CONCLUSIONS

4.1 HYPOTHESIS TESTING

1. Market basket analysis increases sales of products grouped together – Not rejected

    - Yes it does. for e.g. Men between 30- 40 years in age, shopping between 5pm and 7pm on Fridays, who purchased diapers were most likely to also have beer in their carts. This motivated the grocery store to move the beer aisle closer to the diaper aisle and wiz-boom-bang, an instant 35% increase in sales of both.

2. Market basket analysis can be used by companies to spot sales trends and make better decisions – Not Rejected

    - Yes, companies can use Market basket analysis to discover buying patterns of customers and use this information to make decisions which will increase the profits. MBA is beneficial for both customers and the business

3. Market basket analysis improves sales of all products - Rejected

    - It does helps us in finding the most popular items and item sets bought by customers and also it shows us which product is related to which other product. But, this scenario cannot be applied to all kinds of products in the store.

4. Mineral water is the highest selling product in the dataset – Not Rejected

    - Yes Mineral water has highest frequency in this data set.

4.2 CONCLUSION

Market basket analysis is a series of computations that firms use to figure out what's going on with their sales. The Apriori Algorithm can detect some complementary commodities that are frequently purchased together. Understanding how items sell may be used in a variety of ways, including marketing, cross-selling, and recommendations. While we focused on retail in these examples, every company can benefit from a deeper understanding of how their products move.

Complementary commodities are frequently purchased together, and the Apriori Algorithm can detect this.

From above analysis, we conclude that we were able to identify the frequency of different items in the dataset and we found that mineral water is the most purchased item and tomato sauce is the least bought item in the dataset. Mineral water has support of 0.238368. We chose Apriori algorithm because it effectively generates highly informative frequent item sets and association rules for the data of the supermarket. The frequent data items are generated from the given input data and based on the frequent item stets strong association rules were generated.

4.3 LIMITATIONS OF STUDY

- Our dataset has only 7501 transactions. If the data set was very large (100000 or 1 million) transactions, it would have been very difficult for a normal computer to analyze the data faster. Hence, we had to stick to a smaller data set.
- Every transaction has less item sets so we found less products with relationships.
- NaN (Not a Number) values in this dataset are difficult to remove and might create problems in this study.

4.4 SCOPE OF FUTURE STUDIES

In future research it will be very interesting to do in-depth understanding of the association rules by evaluating the changes in the lift and confidence values, which can be made possible by calculating the standard deviation. This way we will be able to witness the evolution of association rules. Further, we can also find association rules using time series clustering method. Also, we can design and develop an intelligent prediction model to generate the association rules that can be adopted on a recommendation system to make the functionality more operational. These results may use in deciding business strategies to gain more profit in their business in future. A retailer can use this information in marketing, store layout, drive recommendation engines (like Amazons, Flipkart's recommendation of another product.), cross selling of product (selling less sold product with higher sold product in a pair), associate pair selling like bread and butter, giving some discounts on a product etc.

BIBLIOGRAPHY

1) "Knowledge Discovery and Data Mining - IBM. "http://researcher.ibm.com/view_pic.php?id=144.

2) "Association rule learning - Wikipedia. "https://en.wikipedia.org/wiki/Association_rule_learning.

3) "Data Mining: Market Basket Analysis - Albion Research Ltd.." http://www.albionresearch.com/data_mining/market_basket.php.

4) "The Use of Machine Learning Algorithms in Recommender ... - arXiv.org.".https://arxiv.org/pdf/1511.05263.

5) J. Han and M. Kamber. Data Mining: Concepts and Techniuqes, Morgan Kaufmann Publishers, San Francisco, CA, 2001.

6) Berry, M.J.A., Linoff, G.S.: Data Mining Techniques: for Marketing, Sales and Customer Relationship Management (second edition), Hungry Minds Inc., 2004.

7) https://www.geeksforgeeks.org/apriori-algorithm/]Webster, Frederick E. (2011), "The Changing Role of Marketing in the Corporation," Journal of Marketing, 56 (October), 1–17.

8) . Loraine Charlet Annie M.C. and Ashok Kumar D, "Frequent Item set mining for Market Basket Data usinK-ori algorithm" , International Journal of Computational Intelligence and Informatics, Volume 1, No. 1, 2011, pp.14-

9) of Data Mining Techniques in Customer Relationship Management:A Literature Review and Classification Elsevier-Expert Systems with Applications, 36 (2009), pp. 2592-2602

10) Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik (2018). Arules: Mining Association Rules and Frequent Itemsets. R package version 1.6-1. https://CRAN.R-project.org/package=arules

ANNEXURE

```
# for basic operations
import numpy as np
import pandas as pd

# for visualizations
import matplotlib.pyplot as plt
import squarify
import seaborn as sns
plt.style.use('fivethirtyeight')

# for defining path
import os
print(os.listdir('../input/'))

# for market basket analysis
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

data = pd.read_csv('../input/mbaprojectfinal/MBA.csv', header = None)
data.head()
data.tail()
data.describe()

data.isnull().sum()

data.dtypes

import matplotlib.pyplot as plt
import seaborn as sns
```

```
from wordcloud import WordCloud

plt.rcParams['figure.figsize'] = (15, 15)
wordcloud = WordCloud(background_color = 'white', width = 1200, height = 1200,
max_words = 121).generate(str(data[0]))

plt.imshow(wordcloud)
plt.axis('off')
plt.title('Most Popular Items', fontsize = 20)
plt.show()

plt.rcParams['figure.figsize'] = (18, 7)
color = plt.cm.magma(np.linspace(0, 1, 40))
data[0].value_counts().head(40).plot.bar(color = color)
plt.title('frequency of most popular items', fontsize = 20)
plt.xticks(rotation = 90 )
plt.grid()
plt.show()

y = data[0].value_counts().head(50).to_frame()
y.index

# plotting a tree map

plt.rcParams['figure.figsize'] = (20, 20)
color = plt.cm.RdYlGn(np.linspace(0, 1, 50))
squarify.plot(sizes = y.values, label = y.index, alpha=.8, color = color)
plt.title('Tree Map for Popular Items')
plt.axis('off')
plt.show()
```

```
data['food'] = 'Food'
food = data.truncate(before = -1, after = 15)


import networkx as nx

food = nx.from_pandas_edgelist(food, source = 'food', target = 0, edge_attr = True)


import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(food)
color = plt.cm.autumn(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(food, pos, node_size = 15000, node_color = color)
nx.draw_networkx_edges(food, pos, width = 3, alpha = 0.6, edge_color = 'black')
nx.draw_networkx_labels(food, pos, font_size = 20, font_family = 'sans-serif')
plt.axis('off')
plt.grid()
plt.title('Top 15 First Choices', fontsize = 40)
plt.show()

data['secondchoice'] = 'Second Choice'
secondchoice = data.truncate(before = -1, after = 15)
secondchoice = nx.from_pandas_edgelist(secondchoice, source = 'food', target = 1, edge_attr
= True)
data.dropna()
```

```
import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(secondchoice)
color = plt.cm.summer(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(secondchoice, pos, node_size = 15000, node_color = color)
nx.draw_networkx_edges(secondchoice, pos, width = 3, alpha = 0.6, edge_color = 'Yellow')
nx.draw_networkx_labels(secondchoice, pos, font_size = 20, font_family = 'sans-serif')
plt.axis('off')
plt.grid()
plt.title('Top15 Second Choices', fontsize = 40)
plt.show()

data['thirdchoice'] = 'Third Choice'
secondchoice = data.truncate(before = -1, after = 15)
secondchoice = nx.from_pandas_edgelist(secondchoice, source = 'food', target = 2, edge_attr
= True)
import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (20, 20)
pos = nx.spring_layout(secondchoice)
color = plt.cm.Wistia(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(secondchoice, pos, node_size = 15000, node_color = color)
nx.draw_networkx_edges(secondchoice, pos, width = 3, alpha = 0.6, edge_color = 'Yellow')
nx.draw_networkx_labels(secondchoice, pos, font_size = 20, font_family = 'white')
plt.axis('off')
plt.grid()
plt.title('Top 15 Third Choices', fontsize = 40)
plt.show()
```

```
# making each customers shopping items an identical list
trans = []
for i in range(0, 7501):
   trans.append([str(data.values[i,j]) for j in range(0, 20)])

# conveting it into an numpy array
trans = np.array(trans)

# checking the shape of the array
print(trans.shape)

data.drop(columns = 'secondchoice', axis=1, inplace=True)
data.head()

import pandas as pd
from mlxtend.preprocessing import TransactionEncoder

te = TransactionEncoder()
data = te.fit_transform(trans)
data = pd.DataFrame(data, columns = te.columns_)
data = te.fit_transform(trans)

data

data.shape

df = pd.DataFrame(data, columns= te.columns_)
df.head()
```

```
from mlxtend.frequent_patterns import apriori, association_rules
frequent_itemsets = apriori(df, min_support=0.003, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x : len(x))

frequent_itemsets.head(20)

frequent_itemsets[frequent_itemsets['length']>= 2].head(20)

rules = association_rules(frequent_itemsets, metric='lift', min_threshold=1)
rules.head(50)

rules[(rules['lift'] >= 5) & (rules['confidence'] >= 0.4)]

import warnings
warnings.filterwarnings('ignore')

# getting correlations for 121 items would be messy
# so let's reduce the items from 121 to 50

data = df.loc[:,['mineral water', 'burgers', 'turkey', 'chocolate', 'frozen vegetables', 'spaghetti',
             'shrimp', 'grated cheese', 'eggs', 'cookies', 'french fries', 'herb & pepper', 'ground
beef',
             'tomatoes', 'milk', 'escalope', 'fresh tuna', 'red wine', 'ham', 'cake', 'green tea',
             'whole wheat pasta', 'pancakes', 'soup', 'muffins', 'energy bar', 'olive oil',
'champagne',
             'avocado', 'pepper', 'butter', 'parmesan cheese', 'whole wheat rice', 'low fat
yogurt',
             'chicken', 'vegetables mix', 'pickles', 'meatballs', 'frozen smoothie', 'yogurt cake']]

# checking the shape
data.shape
```

```
from mlxtend.frequent_patterns import apriori

#Now, let us return the items and itemsets with at least 5% support:
apriori(data, min_support = 0.01, use_colnames = True)

frequent_itemsets = apriori(data, min_support = 0.05, use_colnames=True)
frequent_itemsets['length'] = frequent_itemsets['itemsets'].apply(lambda x: len(x))
frequent_itemsets

# getting th item sets with length = 2 and support more han 10%

frequent_itemsets[ (frequent_itemsets['length'] == 2) &
              (frequent_itemsets['support'] >= 0.01) ]

# getting th item sets with length = 2 and support more han 10%

frequent_itemsets[ (frequent_itemsets['length'] == 1) &
              (frequent_itemsets['support'] >= 0.01) ]


frequent_itemsets[ frequent_itemsets['itemsets'] == {'eggs', 'mineral water'} ]
frequent_itemsets[ frequent_itemsets['itemsets'] == {'mineral water'} ]
frequent_itemsets[ frequent_itemsets['itemsets'] == {'chicken'} ]
frequent_itemsets[ frequent_itemsets['itemsets'] == {'frozen vegetables'} ]
frequent_itemsets[ frequent_itemsets['itemsets'] == {'chocolate'} ]
```

Gantt-Chart

Task Name	Q4 2021			Q1 2022		
	Sept	Nov	Dec	Jan	Mar	Apr
Planning	■	■				
Research		■	■			
Collection of data-set and design			■	■		
Implementation				■	■	
Follow Up					■	■
Final Result						■