Developing a Recommender System for Customers using Apriori Algorithm in Market Basket Analysis

Dissertation submitted in part fulfilment of the requirements
for the degree of Master of Science in Data Analytics
at Dublin Business School

Nishad Abdul Latheef

10382242

Supervisor: Dr. Amir Sajad Esmaeily

MSc in Data Analytics December 2018

Declaration

I, Nishad Abdul Latheef, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this this work is fully compliant with the Dublin Business School's academic honesty policy.

Signed: *Nishad Abdul Latheef*

Date: 06-December-2018

Acknowledgement

It is my great pleasure to express my utmost gratitude and heartfelt thanks to my deeply respected supervisor Dr. Amir Sajad Esmaeily for his treasured guidance and motivation in making it possible complete this thesis. The valuable suggestions he gave throughout the whole process and willingness to guide me without any hesitation is honestly acknowledged.

I would like to also show my deep gratitude to my parents, Mr. Abdul Latheef M. and Sinia K.K., for being there and providing warm support for me during the rough times and motivating me to achieve my goals.

Finally, I would like to thank my friends and other family members who aided me directly or indirectly throughout my research work.

Nishad Abdul Latheef
Dublin, December 2018

Abstract

The aim of the research is to study the possible impact a recommender system can bring about in increasing the revenue for a fully-automated Scan and Go store. The core of the research involves building a recommender system for Customers Using Apriori Algorithm in Market Basket Analysis. This data mining technique makes it possible to create an association between various products with the help of historical transactional data. Once a customer shops an item, the related products could be recommended to them.

Furthermore, the business aspect of introducing a recommender system is also studied. This involves how the revenue can be improved while at the same time satisfying the customers. Basically, the application of recommender system in a physical store is examined.

The secondary research involves the study of various published papers related to relevant topics such as Market Basket Analysis, Recommender systems, Various association rules, use of big data, etc. The knowledge obtained from these papers pave the way for the primary research analysis.

The primary research consists of executing the Market Basket Analysis in CRISP-DM methodology. Further, people reactions to a fully-automated Scan and Go store is also analysed using online survey. All the obtained results are visualised with the help of appropriate tools.

The question of how the recommender system improves the revenue of a company while making their customers satisfied is tried to answer by in-depth analysis of results obtained from the primary research and knowledge obtained from the secondary research.

Contents

List of Figures

List of Tables

1. Introduction

The future seems bright with the modern technologies that uses huge amount of data known as Big-data. The industries are trying to discover better techniques to make their company grow bigger and bigger along with satisfying their customers. The present day's data-torrent inspirers the requisite for new methods to provide people and firms with access to the correct information in the desired amount of time.

In brief, the CEOs of the companies need their employees to build a more personalised techniques of data access and analysis that are adequate of understanding the requisites of people and reacting to these needs in a more reasonable and focused way. The smart use of the recommender systems depicts one of the approaches to building a more personalised data systems which have achieved considerable grip online, especially in the e-commerce industry. At present, major commercial services like Amazon shopping, YouTube, and Netflix help millions of individuals discover what they are seeking for by automatically recommending related items from a deep list of almost infinite possibilities.

These personalised recommendations are the results of simple but effective data mining technique known as Market Basket Analysis (MBA). The retail industry has been using Market Basket Analysis or MBA from long ago to boost marketing performance and to expand their opportunities by putting the desired offer to the desired customer. Retailers utilises the result data obtained from market basket analysis in a variety of ways:

- Laying out the Store: While setting up the store, the products on the particular store level that co-occur together in the result of analysis in close proximity will be placed to boost the shopping experience of the individual customer.
- Effective marketing: Depending upon the prior purchase behaviour patterns through their transactional data or analysis of what is currently in the shopping cart, retailers will be able to market and sell further products to that particular customer.
- Digital Marketing of the products by placing the items on a useful website or products in catalogues.

In general, the main aim of the market basket analysis models is to identify and predict the next product that the individual customer would be interested to buy or to search through the internet. One of the major breakthroughs in the retail-tech industry is the introduction of Scan and Go stores. The Scan and Go technology was developed to produce a much simpler, faster and more favourable shopping experience for the individual consumers. Just like the barcode technology, introduced in the early 1980s, reconstructed the overall efficiency of retail stores, Scan and Go is all set to grow into retail's next great innovative technology that can boost a range of existent retail practices.

Scan and Go is definitely looking like the next logical step in building and improving the in-store experience. This is evident from the more effective technique of self-checkout machines that eliminated the long waiting in line for a shop assistant to scan products and deal with your payments.

1.1 Problem statement

The aim of any commercial firm is to boost their current revenue. The age long dilemma of making the decision that creates a balance between customer satisfaction and revenue growth have disturbed the companies. An enormous amount of money has been spent towards the business intelligence department in order to fix this. With the emergence of high-end technologies like Artificial Intelligence (AI) and automation, the organisations are searching for effective technologies to assist them. The chosen technology should be practical enough to ensure the overall revenue boost along with increasing customer satisfaction.

1.2 Research question and aim

The main aim of this research deals with building a recommender system that understand the purchase behaviour of the customers using transactional data in a scan and go retail store. This can be done by providing the customer a real-time recommendation of products via an application on their device while shopping. The research question that deals with the objective is:

"***How can a recommender system using Market Basket Analysis helps improve the situations for both the company as well as the customers in a Scan and Go store***?"

The sub-research questions that can provide the answers to this research question are:

- What is the best algorithm for building the recommender system in a real world store?
- How effective will the recommender system be on improving the revenue?
- What are the possible observations that can be obtained from deploying the results of the analysis?
- How the people are responding to the changes in retail store brought about by Artificial Intelligence (AI)?

1.3 Dissertation Roadmap

The entire dissertation project is divided into sections of eight as presented in the figure below. In the first six sections, the researched knowledge and evaluation of the model is depicted. This is trailed by a critical self-evaluation section in which the experience throughout this research is mentioned.

The remaining sections of the thesis consist of references and appendices that is utilised throughout the research. The summary of each section is provided after this figure in an order.

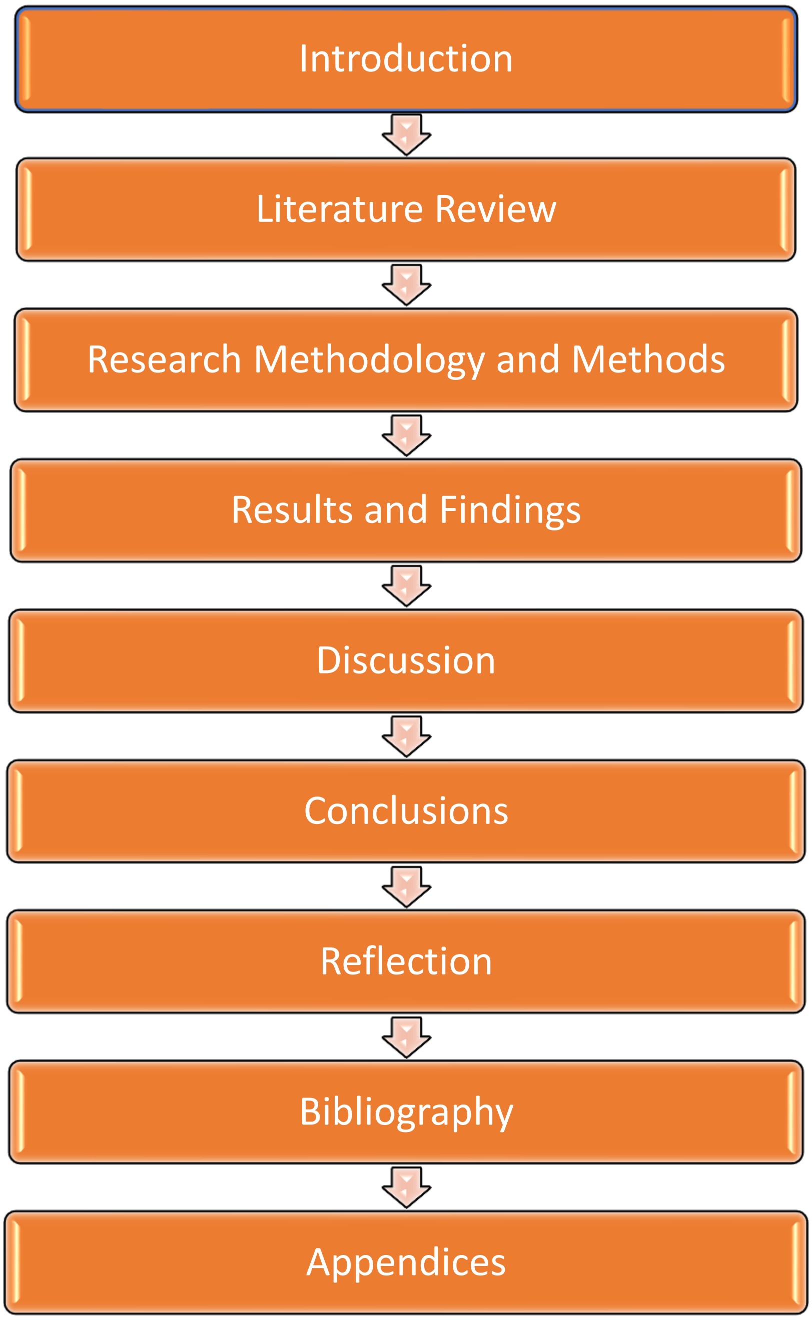


Figure 1: Dissertation Road Map

1.3.1 Introduction

The Introduction section presents the background knowledge about this research topic. It highlights the significance and the motivation behind of this dissertation paper by describing the applied research question, its aim along with the objectives. This section also provides the research scope and limitations.

1.3.2 Literature Review

This particular section contains summary of various articles, blogs and books on topics that are relevant for the proper completion of the dissertations. The topics that are reviewed are Different Mining Algorithms, Recommender systems, Market Basket Analysis, Scan-and-Go stores and the use of Big Data.

This section is significant not just since it designed the entire research question but also it explains the breach within the literature and highlights the necessity of carrying out this research.

1.3.3 Research Methodology and Methods

The Research methodology section provides the method implemented during the execution of the research. This will include Research Design: CRISP-DM, explanation about the chosen algorithm, online survey and its analysis along with the notable ethical problems and limitations in the conducted research.

1.3.4 Results and Findings

This section exhibits the results obtained from data miming process and the findings observed from the online survey that is given under those essential section. Further, these sections includes some sub-sections that is utilised to provide the results and findings so that these can be explored in the discussion section.

The section consist of Exploratory Analysis, Data Pre-processing, Association Rule Mining, Association Rule Visualisation and Online Survey on Amazon Go store.

1.3.5 Discussion

The comprehensive exploration of the provided results in results and finding section is given in this section. In detail, some of the validation of the statements given in literature review is evaluated and also, the possible answers to the research question is tried to be given by discussing the results.

1.3.6 Conclusion and recommendations

In this particular section, the summary of the important conclusions collected from all the results and findings that consist of both primary as well as secondary research along with the relevant points from the discussion section are provided. Furthermore, some suggestions which could be utilised for the future are also given along with its possible limitations.

1.3.7 Reflection

The reflection section provides the relevant details about the researcher along with the self-assessment of the episodes that was experienced during the execution of dissertation process.

1.4 Scope and Limitations

This dissertation paper is expected to develop a recommender system for customers in order to have a more convenient shopping in a Scan and Go store. Here, the developed model will simultaneously try to provide methods for revenue growth of a company as well as satisfying their customers. This technique seems significant as the future of retail store depend upon Automation and Artificial Intelligence (AI). Many bigshot retail companies could adopt this idea to increase their overall revenue.

Figure 2: Recommended Basket

The whole research is carried out in such a way that it focuses on implementing the recommender system in a real-world retail shop. The online recommender systems used by Amazon, Netflix etc. can track down the user's online behaviour, such as browsing history, amount of time spend on a specific product page, to improve their recommendation. However, this is not possible in real-life shopping. So, the recommendation will be purely based on the historical transaction data of the products.

Various data analytic and business intelligence concepts are explored in this paper such as data mining, association rule mining, Apriori algorithm, sentiment analysis, Big data etc. The data mining tools such as RapidMiner and R studio along with the data visualisation tool like Tableau is also used in this research.

However, due to time constraints and the restriction in exploring high-end technologies and personal information of real-life customers, the research is limited. The data mining process in executed to a dataset that cannot be termed as Big Data. Since the companies will have to deal with Big Data in reality, the literature review section also try to cover the scope of this particular idea in the field of Big Data.

1.5 Contributions of the study

Automating the store using Artificial Intelligence is arguably the future of retail industry. Some of the companies already have initiated towards this change. On the other hand, recommender systems have been used is a various way for many years now.

This paper suggests collaboration of recommender system with the Scan and Go store. This might provide a solution for overall increase in revenue for the company along with increasing the convenience of their customers. The developed recommender system will be more like a real-time Virtual Personal Assistant (VPA) that helps out the customers choose what they might require but wouldn’t be able to memorise during their shopping

2. Literature Review

The purpose of this Literature review is to support an academic groundwork for this thesis. In section 2.1 will provide an overview regarding the various Mining Algorithms. Section 2.2 describes the different recommender systems and uses of it. In section 2.3, the history, the definition, and the practical examples market basket analysis will be given. Section 2.4 will focus on the emergence of the Scan and Go stores all around the world. Finally, section 2.5 will tell how big data can be collaborated with this.

2.1 Different Mining Algorithms

The recommendations are created as a result of Market Basket Analysis which is an example of Association Rule Mining. This mining technique identifies the impressive association or relationship between a large set of data items. This section presents a survey on the already existing data mining algorithm for market basket analysis.

An upgraded classification technique which is based on the predictive association rules is suggested by Zhixin et al., [15]. The advantage of associative classification and conventional rule-based classification can be united by using a variant of association classification methods known as the Classification dependent predictive association rules (CPAR). For the creation of the association rule, since almost all of the repeated calculation in CPAR is avoided and more than one literal can be chosen to create multiple rules simultaneously, it is considered to be one of the most efficient way in comparison with the traditional rule-based classification. Even if the advantage mentioned before ignores the duplicated calculation in rule generation, the class rule distribution discrepancy and interruption of erroneous class rules are few of the demerits of the prediction techniques. Further, it is inefficient to cases were no rules are satisfied. To avoid these demerits, techniques such as Class Weighting Adjustment, Centre Vector-based Pre-classification and Post-processing with Support Vector Machine (SVM) are advised by the author of this article. This specific method produced an average improvement of 5.91% on F1 score in comparison with CPAR during an experimental inspection on Chinese text categorization corpus TanCorp.

Qiang et al., [22] provided with an association classification related method on briefness of rules. It was observed that highest classification accuracy and powerful flexibility was produced by associative classification. On the contrary, a demerit of over fitting is being suffered by this associative classification due to the fact that the classification rules fulfilled least support and lowest confidence are restored to the classifier as powerful association rules return. The author of the paper advised a creative association classification method which establishes neatness of rules which is actually an extension of the Apriori Algorithm which acknowledges the inquisitiveness, significance, coinciding relationships among the rules. In comparison with Classification Based on Association and Classification based on Multiple Association Rules, the advised technique was found with better classification accuracy during an experimental observation.

A unique rule weighting method in the classification association rule mining is proposed by Wang et al., [38]. An approach named Classification association rule mining (CARM) is one of the most recent classification rule mining technique which uses classification association rules (CARs) to builds an association rule mining related classifier. Various numbers of rules arrangement approaches have been set up in the near past, that can be labelled as support-confidence, rule weighting, and hybrid. The CISRW (Class-Item Score based Rule Weighting) based rule arrangement technique was found to be doing well by mode of accuracy of classification as a result of a simulation test.

In his paper, Bartik [3] suggested an association related classification for relational data and its utilisation in web mining. The classification according to the mining association rules is found to be a better efficient technique whose classification design is recognisable to humans. The author intended to provide suitable adjustments of the basic association related classification technique which can be beneficial in data collection from Web sources and pages. In this article, the adjustment of the approach and desired reduction of numeric attributes are produced.

A shared Apriori association rule and traditional Apriori mining algorithms for grid-related knowledge detection was suggested by Sumithra et al., [31]. The main objective of this article is to acquire knowledge with the aid of predictive Apriori and shared grid related Apriori algorithms for association rule mining. The applications of an association rules analysis data mining assignment with the aid of Grid technologies is produced by the author. The author also presents an aftermath of application with a contradiction of existing Apriori and shared Apriori. Distributed data mining structure provides an efficient usage of multi-processors and databases to boost the execution of data mining and incorporate data sharing. For determining the efficiency of the mention approach, performance analysis of Apriori and predictive Apriori methods on an approved database have been presented utilising the weka tool. The main aim of grid computing is to provide the firms and application developers the ability to build shared computing environments which can utilise computing resources whenever necessary. Therefore, it can help improve the efficiency and decline the total cost of computing networks by decreasing the time required for the data processing and enhancing the resources and sharing the workloads. Hence, this allows the users to obtain much quicker results on big-scale operations that too with a lesser cost.

In this article, Data Mining Methods for Market Basket Analysis is utilised by Trnka [32]. The technique proposed in this paper is the application of Market Basket Analysis to Six Sigma technique. The Data mining methods provides much more possibilities in the market domain such as Market Basket Analysis (MBA). A variety of statistical methods are used by the Six Sigma technique. With the collaboration of Market Basket Analysis (MBA) to Six Sigma technique, the outcomes can be intensified, and the performance degree of the sigma technique can also be improved. The General Rule Induction (GRI) technique is utilised by the author to develop association rules between the products in the market basket. A variety in between the products are created by these associations. The dependence between the products are depicted by the usage of web plot.

Yanthy et al., [44] proposed a Mining Interesting Rules by Association and Classification method in this article. The objective in data mining is to reveal the hidden information from data and various techniques have been put forth so far. But the disadvantage is that fundamentally not every rule are accurate – only a small part of the developed rules would be of interest to various provided user. Hence, a number of criterions such as lift, confidence, information gain, support, etc., have been advised to discover the best or most interesting rules. On the other side, some methods are better at producing rules significant in one specific interestingness criteria but not so good in another interestingness criteria. The relationship between the methods and interestingness criterions of the produced rules is not definite yet. The relationship between the methods and interesting criterions was explored by the author. The author also utilised synthetic data so that the result of the method is not restrained to specific cases.

Xie et al., [39] put forth the Market Basket Analysis (MBA) Based on Text Segmentation and Association Rule Mining. The scientific decision support for trade market are provided by mining association rules among items customers bought together by utilising Market basket analysis. A creative market basket analysis method by mining association rules on the item’s internal factors which are attained with the aid of automatic words segmentation method is suggested by the author. This method has been utilised for dynamic dishes recommend system and the outcomes also show better accuracy in the experimental procedures.

In this particular article, a market-basket analysis with principal component is suggested by Chiu et al., [8]. Market Basket Analysis is a familiar organisation crisis that can be resolved computationally with the aid of association rules, mined from the transaction data of the customers to decrease the cross-selling outcomes. The market-basket analysis is modelled by the author as a limited mixture density of human utilisation actions which is based on social and cultural actions. The leads to the utilisation of principle component analysis and possibly mixture density analysis of transaction data of customers which was not apparent before. The contradiction of PCA and association rules mined are made by the author from a set of benchmark transaction data of the customers, to find out the similarities and differences between these two data analysis tools.

In this article, Cunningham et al., [9] suggested a Market Basket Analysis of Library Circulation Data. The method of Market basket analysis have recently been utilised widely in examining the individual customer purchasing patterns - especially, in discovering items that are usually bought. The a-priori market basket tool is used by the author of the paper for the purpose of discovering subject classification grouping which coexist in transaction data of books rented from a university library. This analytics results can be used in guiding the users to further parts of the collected that may contain documents that are relevant to their information need, and in discovering a physical layout of the library. These outcomes can also provide insight into the total quantity of scatter that the classification method bring out in a specific collection of documents.

In another article, a mining local association patterns from spatial dataset is provided by Zongyao et al., [29]. Both model and algorithm are put forth by the author of the article for the purpose of mining local association rules from provided spatial dataset even though the reality that spatial heterogeneity may be widely available in realism is entirely considered. The critical element of the suggested model is the calculation of Localized Measure of Association Strength (LMAS) that is used to compute local association patterns. The Spatial association relations are particularly described as spatial relations which are modelled by DE-9IM model. The mining methods for finding out the local association patterns from spatial dataset is also provided by the author. The suggested approach mines reference and target objects that have potential association patterns and executes LMAS for all objects in the reference objects for some relevant spatial relation. Therefore, a LMAS distribution map that depicts association potential variations over the observation area is obtained as the outcome of the algorithm. The spatial interpolation for LMAS is advised to produce a continuous LMAS distribution which can be used to examine hot spots which will in turn give strong association patterns. This suggested approach was practiced in an ecological system research.

Pei et al.., [37] proposed a mining association rules based on Apriori algorithm and application. The mining association rules is an important topic in the data mining research. Focused at two problems of finding out the frequent item-sets in a broad database and mining the association rules from frequent item-sets, the author provides the summary of the analysis that was carried out on mining frequent item-sets algorithm with the aid of Apriori algorithm and mining association rules algorithm with the aid of improved calculation system. The mining association rules approach with the use of support, confidence and interestingness is improved that targeted at producing interestingness inefficient rules and losing helpful rules. More reasonable association rules are produced containing negative items by avoiding unhelpful rules. The proposed approach is used in mining association rules from the 2002 student score list of computer concentrated domain in Inner Mongolia university of science and technology.

A mining association rules with new measure criteria is suggested by Yong et al., [41] in an article. These days, association rules mining from huge databases is a very much interested research area of data mining which is inspired by various application areas. Nevertheless, there are some hurdles in the strong association rules mining that rely upon the support-confidence system. At the beginning, there are enormous number of excessive association rules that are produced, then it is difficult for a user to find out the relevant ones. Then, the interrelation between the factors of a particular application areas is ignored. Hence, creative measure factor named Chi-Squared test and cover must be brought forward to association rules mining, and the more crucial element is the utilisation of Chi-Squared test to decrease the number of rules. The author uses the Chi-Squared test and cover of measures for the purpose of association rules mining for removing the item-sets which are statistic free and the frequent item-sets or rules are produced in the meantime. Hence, the quantity of patterns item-sets are deducted, and it is easier for the user to collect the highly relevant association rules. The outcomes of the test proposes that the Chi-Squared test is effective on reducing the number of patterns through combining support and cover factors. The pattern selection based on Chi-Squared test can eliminate few unwanted attributes and the effectiveness and accuracy of mining association rules are boosted.

Vo et al., [34], in this article, proposed a mining traditional association rules with the help of frequent item-sets lattice. A number of techniques have been drafted for the overall improvement of time in mining frequent item-sets. Nevertheless, the techniques which handle with the time of mining association rules were not researched to the deepest degree. In actuality, in situation were the database consist of many frequent item-sets, the amount of time for mining association rules is much greater than that required for mining frequent item-sets. In this article, the author built a practical case of lattice in mining traditional association rules that will considerably reduce the time for mining rules. This approach consist of two stages: (a) building of frequent item-sets lattice and (b) mining association rules from the lattice. The parent-child linkage in lattice is utilised for obtaining the association rules easily. The observed results shows that the mining rule from lattice is more effective than the direct mining from frequent item-sets by using the hash table.

The optimization of mining association rules with categorical and numeric factors is described by Rastogi et al., [24]. The mining association rules done on huge data sets has acquired serious attention lately. The association rules are used for the purpose of for predicting interrelations between the factors of a relation and consist of practices in marketing and many other retail fields. Furthermore, the association rules that are optimized will provide an effective way to target on the most relevant factors connecting specific attributes. The optimized association rules are permitted to contain attributes that are not instantiated, and the trouble is to discover instantiations in such a technique that the maximization of either the support or confidence of the rule is carried out. In this method, the problems of optimized association rules is briefed in three ways: (a) association rules are allowed to contain disjunctions in extra of features that are not instantiated, (b) association rules are permitted to include a random number of features that not instantiated, and (c) features that are not instantiated can be either numeric or categorical. This particular association rules allows to mine more beneficial data about seasonal and local patterns connecting multiple factors. An effective technique for pruning the search region while calculating the optimized association rules for both numeric and categorical factors is also provided in the paper. The outcome of the experiments proves that pruning methods are efficient for a large number of features that are not instantiated, divisions, and values in the domain of the factors.

An analysis on association rules mining depending on ontology in the field of ecommerce is performed by Wang et al., [42]. The commercial activities carried out with the help of Internet appears to be more and more famous. The collection of essential details by data mining are made possible by enormous number of transaction logs produced by each customer. In this way, association rule mining is very crucial in the ecommerce industry. Nevertheless, there are different difficulties that happen in the current association rules mining schemes. The current traditional methods can't explain these difficulties quite well enough. With the purpose of resolving these problems better, this article suggests association rules mining based on ontology. Usually, researches focus mainly on three parts during the data mining process: (1) approaches of ontology production and values of commodity classification; (2) explaining R-interesting depending on real cases; (3) executing association rules mining based on ontology by enhanced Apriori. Furthermore, this article examines the improved algorithm using the FoodMart2000, Java as the language for development and Jena as the ontology engine, finishes the whole procedure of mining, and approves the validity of the algorithm by the sample of the database.

2.2 Recommender systems

The Recommender Systems are defined as the software procedures and tools that offer recommendations for products which are maximum possibility of interest to a specific user [7,26,25]. The recommendations connect to several decision-making activities, for instance what products to buy, what kind of music to listen to, or what news from the web to read. In its most understandable way, personalized suggestions are presented as lists of products according to its ranked order. In carrying out this ranking, the recommender systems attempt to forecast what the most appropriate items or services are, depending on the consumer’s preferences and constraints.

As e-commerce company websites started to develop, a demanding requirement appeared for producing recommendations resulting from filtering the entire range of existing substitutes. The consumers found it hard to reach at the most suitable selections from the vast variation of products and services that these company websites presented. Some of the various types of recommendation methods are:

- Content-based filtering: Content-based method is a field-dependent process and it underlines more on the evaluation of the various factors of products so as to produce predictions. During the recommendation of documents like web pages, books and music, content-based filtering method is the most effective method. The suggestion is made in content-based filtering methods depends on the consumers profiles using factors mined from the content of the products the consumer has associated with in the past [6,4]. The products which are mostly connected to the positively rated products are suggested to the consumer. The content-based filtering utilises various kinds of models to discover similarity among documents so as to produce relevant suggestions.

- Collaborative filtering: The Collaborative filtering is a field-independent forecast method for content which cannot simply and sufficiently be defined by metadata such as music and movies. The collaborative filtering method works by developing a database of desired choices for products by consumers. After that, it matches consumers with significant interest and choices by evaluating similarities among their profiles to produce suggestions [16]. Those consumers form a group known as neighbourhood. A consumer receives suggestions to those products which he has not been associated with before, but which were previously positively rated by other consumers in his neighbourhood. The suggestions which are created by collaborative filtering can be of either recommendation or prediction. The two types of techniques to carry out collaborative filtering are: memory-based and model-based.
    1. Memory-based: This can be done in two different ways. The first one recognizes clusters of consumers and uses the activities of one particular consumer to predict the activities of other relatable consumers. The second way recognizes clusters of products which have been rated by consumer A and uses them to forecast the activity of consumer A with a different but relatable product B. These ways generally encounter main difficulties with huge sparse matrices, because the number of consumer-product interactions can be really low for creating clusters of high quality.

2. Model-based: These approaches are established on machine learning and data mining methods. The objective is to train the models so that it would be capable of providing predictions. For instance, we could utilise current consumer-product interactions to train a particular model to forecast the top ranked products that a consumer might be interested in the most. One benefit of these approaches is that they are capable of recommending a greater number of products to a greater number of consumers, in comparison with other approaches such as memory-based. We could say this approach have greater coverage, even while dealing with huge sparse matrices.

- Hybrid filtering: This approach is used so as to obtain improved system optimization to dodge some restrictions and difficulties of pure recommendation systems [1,30]. The hybrid techniques is built in such a way such that a blend of algorithms will offer more precise and efficient suggestions than an individual algorithm. This is due to fact that the demerits of one algorithm could be overcome by another algorithm [28]. Utilising multiple suggestion methods can overpower the weaknesses of a single method in a combined model. The combination of techniques can be made possible in any of the following ways: distinct execution of algorithms and merging the result, using some content-based filtering technique in collaborative method, using some collaborative filtering technique in content-based method, producing a combined recommendation system which collectively brings together both the methods.

The recommender systems can be expanded in various ways which consist of boosting the understanding of consumers and products, integrating the background knowledge into the recommendation procedure, backing up multi-criteria rankings, and giving more flexible and less disturbing types of recommendations. Such more inclusive models of recommender systems can offer improved recommendation aptitudes.

2.3 Market Basket Analysis

A variety of opportunities in the marketing domain are created by the Data Mining techniques. The ability of decision making and discovering the behaviour of each of the individual consumers has become critical and challenging dilemma for the firms so as to endure in this competing environment. The most challenging problem which these firms faces is to derive the data from their huge customer databases, so as to acquire competitive advantage.

The author of this article, Yanthy et al [43], presents that the main objective in the data mining process is to explore the hidden knowledge from data. The author also suggested a variety of algorithms. Nevertheless, generally not every rule are relevant and only a small portion of the produced rules seems to be of relevance to any given consumer. Therefore, several methods like confidence, support, and lift have been advised for the generation of the best or most relevant rules. Nonetheless, some algorithms are efficient at producing rules high in one particular factor but inefficient in other.

The technique of Apriori algorithm is suggested by Rakesh Agarwal [2]. The Apriori algorithm is the primary associative rule mining algorithm that was proposed. And all other further development in algorithm such as associative classification, association, classification algorithms have utilised Apriori as a portion of its technique. The Apriori algorithm is defined as a level-based, simple algorithm that mines from the transaction details of the consumers. This algorithm also utilises previous information of frequent item set factors. An iterative method referred to as a level-based search is used by this algorithm, in which (n) - item-sets are utilised to discover (n+1) – item-sets. To boost the effectiveness of the level-based production of frequent item-sets the Apriori property is utilised here. The Apriori property states that all the non-empty subsets that are generated of a frequent item-set should also be frequent. The anti-monotone property of support-measure that states that the support for an item-set will cease to surpass the support for its own subsets. A two-step procedure involves the join and prune processes are carried out iteratively.

The Apriori algorithm is one of the Data Mining techniques that is utilised to discover the frequent items/item set from a given database of information. The algorithm consists of 2 main steps:

a. Pruning

b. Joining

The Apriori property must be taken into consideration as the crucial feature prior to moving forth with the algorithm. The Apriori property defines that If an item named X is connected with an item named Y, then;

Support of (XUY) =minimum (Support of (X), Support of (Y))

Fundamentally, we make use of the values of support and confidence during the evaluation of the relevance of an association rule. This will provide how powerful the connection is between the items in the transaction.

The support of an item is defined as the total amount of transactions that consist of that particular item. The items which ceases to attain the threshold support value are ignored from the next processing. Basically, support controls how often a rule is valid to the provided data set.

The confidence is referred to as the conditional probability which a transaction that contain the items on Left Hand Side (LHS) will mostly contain the item on Right Hand Side (RHS).

The value of confidence defines how often the item on RHS is present in the transaction that contain the items on LHS. During the generation of the rules one should calculate these two factors as it is pretty critical. The value of confidence will help in calculating the dependability of the conclusion provided by the rule.

An innovative association rule mining technique which does not utilise candidate rule production named FP-growth is proposed by Han [13,14]. This method creates an extremely concise frequent pattern tree or (FP-tree) illustration of the transactional data of the consumers. The entire data of transaction is denoted in a tree like structure by at most a single path. The FP-tree is comparatively lesser in size than that of the original transactional data. The generation of the tree involves two database scans. During the primary scan, frequent item sets including their support in all transactions are created and in the next scan construction of FP-tree is carried out.

The procedure of mining the rules is achieved by clustering the patterns with the ones generated from the conditional FP-tree. One restriction of FP-growth technique is that, while dealing with dimensionally bulky database, memory may cease to fit the FP-tree.

The Classification based on Association (CBA) is suggested by Liu [19] as the original Associative Classification algorithm. The popular Apriori algorithm [45] is executed by the CBA so as to find out frequent rule of items. The Apriori algorithm approach usually involves of three major steps;

1. Continuous factors contained in the training dataset becomes rejected.
2. Finding out the frequent rule items.
3. Creation of relevant rules.

The selection of confidence rules with large values are carried out by the CBA to denote the classifier. Furthermore, the CBA employs the highest confidence rule were body finds similarity with the test case to forecast a test case. The outcome of the experiment proves that the CBA provides greater quality classifiers in terms of precision that rule generation and decision tree clustering approaches.

In an article, Phani Prasad J, Murlidher Mourya [23] provides that there are plenty of research studies about the association Rules and already present data mining techniques utilisation for market basket analysis. However, the author primarily targets on the Apriori algorithm and inferences that the algorithm can be improved by expanding it in the forthcoming projects that will deal with reducing the time complexity. The Disadvantages of the algorithm are also pointed out by the author. Nevertheless, it is claimed that there is the way to boost the effectiveness of the algorithm.

2.4 Scan and Go stores

The introduction of the Scan and go stores created a pivotal step in the retail industry. On January 22, 2018, e-commerce giant Amazon opened its first retail scan and go store named Amazon Go at Seattle [11]. It provides the customers the opportunity to shop their desired products from Amazon physically rather than shopping via the Amazon.com website. Nevertheless, these shops are different from other traditional shops such that it doesn't need any cash registers. The customers have to simply walk in, select what they desire and walk out of the store to complete their shopping.

Amazon defines the Amazon Go store as an innovative store where there is no checkout necessary. This basically means, the customers will never have to wait in a queue while shopping at Amazon Go. The shopping is assisted by an Amazon Go application. The whole process gets running by utilising the similar kind of technologies such as deep learning and artificial intelligence that makes the self-driving cars work. The mentioned technology makes it possible to sense when the items are taken or placed back to the shelves. The app also keep an eye on these items in the customer’s virtual shopping cart. The customer’s Amazon account is billed, and a copy of the receipt is received by the customer as soon as they exit the store with their shopped items. In order to use the Scan and Go store one will require the following;

- A verified user’s account.
- A functional smartphone.
- An Application to track the shopping.

The gates of the store will open once the QR code in the smartphone is held against the provided scanning device. The system works by extracting the data from multiple cameras and sensors by utilising a combination of artificial intelligence, machine learning and computer vision. This is to make sure that the customers are only billed for the items they selected from the selves. The usage of computer vision implies that the numerous cameras fitted on the roof of the shop keeps track of the customer in the store. The successful prevention shoplifting and fraud is one of the major challenges faced in this type of store.

Based on the research conducted by The Wall Street Journal, the Amazon is planning to establish 2,000 scan and go stores in the future all over USA. The stores are rumoured to be set-up in multiple formats so as to compete with other retail giants like Walmart.

In the meantime, other retail companies such as Walmart is also planning to upgrade their current system by incorporating the Scan and go technology [35]. Several experimental runs were reported to check the functionality of the customer checkout service in more than 350 Walmart stores across the US.

2.5 Use of Big Data

The term Big Data is defined and opinionated by several authors in their articles. Based on an article, Cuzzocrea et al. [10] represents Big Data as massive quantities of unstructured data generated by high performance tools and application. Big Data is also termed as a social, scientific, and intellectual marvel that depends on the interaction of technology, methodology and analytics by Boyd & Crawford [5]. In another article, Madden [20], Big Data is stated as something that specifies that the data is too large, too quick, or too difficult for available tools and applications to execute. Nevertheless, McAfee & Brynjolfsson [21], in their paper, points out Big Data to the 3 "V" s;

- The Volume for the enormous quantity of data,
- The Variety for the data generation speed, and
- The Velocity for the developing unstructured form of data.

There are several benefits in the usage of Big Data. For instance, Labrinidis & Jagadish [17] provides that the acknowledgment of Big Data has brought about a budding interest for data dependent decision making which is also referred to as evidence based decision making. As a matter of fact, the firm accomplishes on their financial and operational goals when they portray themselves as a data driven firm [21]. Furthermore, LaValle et al. [18] points out to the fact that the top ranked performing firms in their market sector recognizes and utilises the Big Data and analytics as a differentiation tactic. The Big Data technology also has the capability to remodel management sector, since a crucial property of Big Data is the effect it has on how decisions are taken and who takes these decisions [21]. Therefore, executing Big Data and data dependent decisions will definitely lead to a performance improvement for that specific firm.

In the paper, Woo J. and Xu Y. [42], provides the Market Basket Analysis (MBA) Algorithm on Map/Reduce. This approach is association dependant data mining procedure to discover the most frequently appeared group of items in baskets at a store. The provided data set provides that associated products can be coupled by using the Map/Reduce technique. After the products are paired, it can be utilised for further researches by statically evaluating them even successively that is referred to be beyond the paper.

The authors of another paper, Videla-Cavieres I.F. and Sebastián A. [33], proposed an innovative method that utilises graph mining procedures to execute market basket analysis and the utilisation of overlying community recognition algorithms as a frequent item set detection approach. The paper also provided this as an approach to mine valuable information from loads of item sales transaction details. The technique that they described has proved that it can be utilised as an effective approach to find out the frequent item-sets existing in transactional purchase data. These provided knowledge is very important as it is extremely beneficial for retail sector by representing relationships which were not noticeable for the expert of the retail.

3. Research Methodology and Methods

In this section, the fundamental assumption about the research methodology is depicted. The strengths and weaknesses of the suggested research methodology will be provided with relevant reasoning of the activities that are carried out. The usage of CRISP-DM as the research design is highlighted. Next, the selected algorithm is described along with the justification of choosing it. The research ethics are discussed in the following sub-section. Finally, the limitations of the methodology is explained.

3.1 Research Design: CRISP-DM

The entire process of the Market Basket Analysis was carried out using the CRISP-DM data modelling steps. The CRISP-DM methodology is executed by dividing the entire data mining process into six processing steps. The mentioned steps are given as;

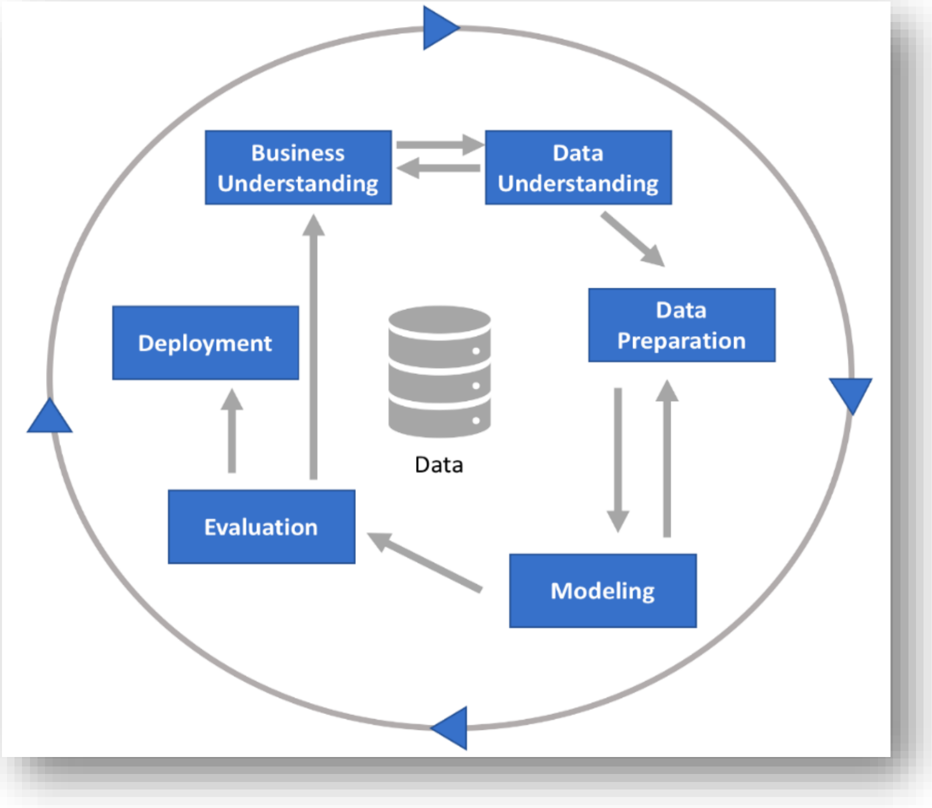


Figure 3: CRISP-DM

3.1.1 Business Understanding

The recommender system provide an assistance while shopping in an e-commerce site. These recommendations are based on user interaction with the site and the previous transactions of similar items.

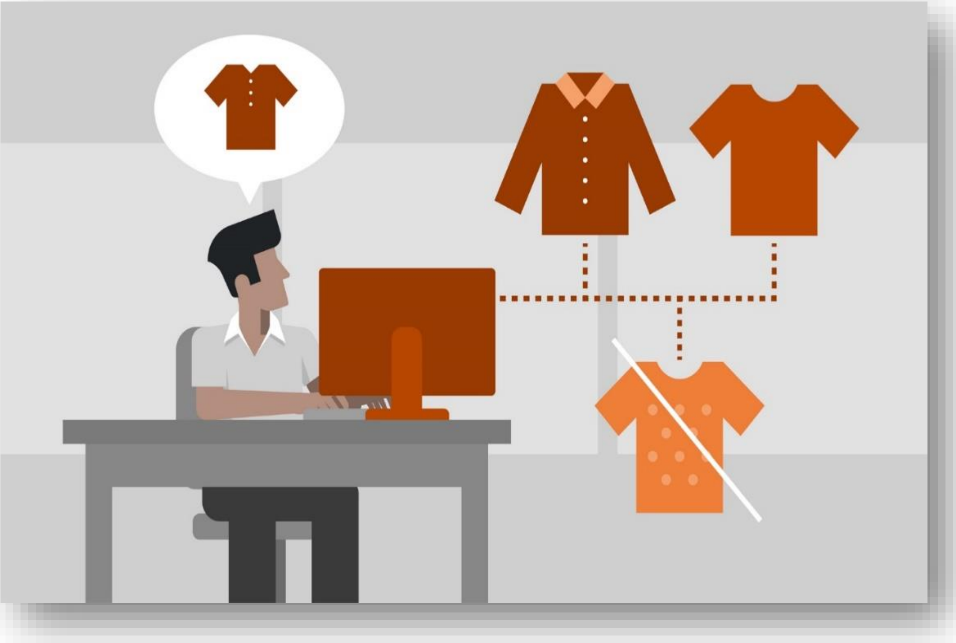


Figure 4: Recommender System

The Scan and Go stores are established with the intention of creating a quicker shopping experience by eliminating long queues at the checkout.

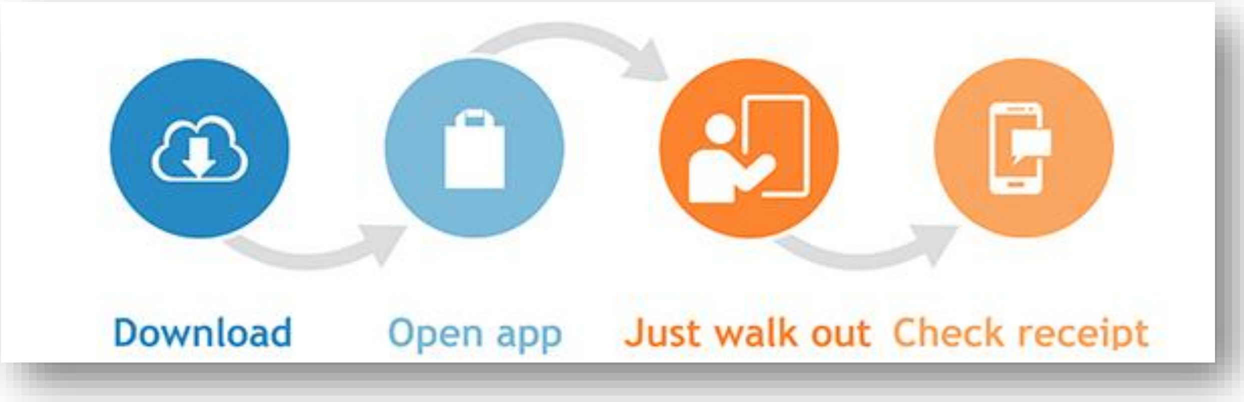


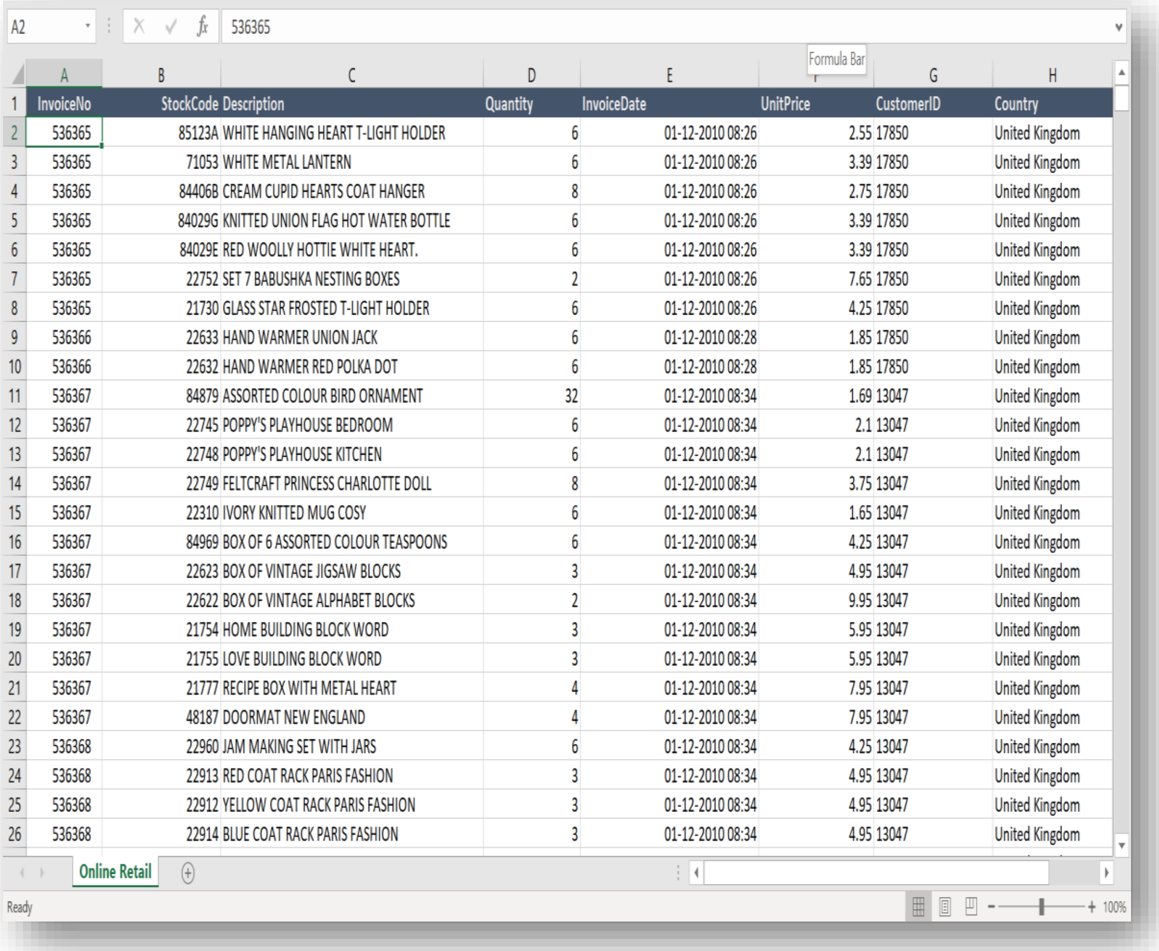
Figure 5: Scan and Go Store (SGS)

This research put forth an idea of collaborating the recommender system with Scan and Go store. The main objective behind developing the suggested model, as mentioned in previous sections, is such that;

- The retail company finds considerable growth in their revenue.
- The customers will be provided a more convenient and faster shopping experience.

3.1.2 Data Understanding

The dataset that is used to build the model is obtained from the 'UCI Machine Learning Repository'. The dataset is named as "Online Retail.xlsx". The dataset, dated from 01/12/2010 to 09/12/2011, consists of transaction data from a Europe-based e-commerce website. The main reason for selecting this particular dataset over the pre-loaded dataset in R is that it is easier to apply data pre-processing on this kind of dataset. The dataset includes 541909 rows and 8 columns of data.



	A	B	C	D	E	F	G	H
1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010 08:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	01-12-2010 08:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01-12-2010 08:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	01-12-2010 08:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01-12-2010 08:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	01-12-2010 08:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	01-12-2010 08:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	01-12-2010 08:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	01-12-2010 08:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	01-12-2010 08:34	4.25	13047	United Kingdom
24	536368	22913	RED COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kingdom
25	536368	22912	YELLOW COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kingdom
26	536368	22914	BLUE COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kingdom

Figure 6: Data Set

1) InvoiceNo: This column indicated the Invoice number. Every transaction is given a unique 6-digit basic number which is nominal. The generated code that begin with the letter 'c' specifies that the transaction is cancelled.
2) StockCode: It defines the product or item code. Every sperate product in allotted a 5-digit basic number which is also nominal.
3) Description: This column simply provides the name of the product or item. It is also nominal.
4) Quantity: The total quantities of respective product or item included in the transaction. This value is numeric.
5) InvoiceDate: The date and time of the invoice is mentioned in this column. The numeric value of the day and time was automatically created during every transaction. One of the sample values of InvoiceDate from the dataset is 01-12-2010 8:34
6) UnitPrice: The unit price of the product or item is given. The numeric value of Product or item price per unit in pound sterling.
7) CustomerID: The ID number of each customer is provided in this column. The nominal value, that is uniquely generated for each customer, is a 5-digit basic number.
8) Country: This column gives the country name. The nominal name of the country where the respective customer lives in provided.

3.1.3 Data Preparation

The whole process of Data Preparation is done with the help of R language in R Studio. The R is a programming language that is commonly utilised among data scientists, statisticians, Machine Learning programmers, and data miners for building a statistical software application and for the purpose of data analysis.

The dataset is contained in an excel format. Before executing Market basket Analysis or Association Rule mining, the excel format of data has to be transformed into transaction data so as to bring together all products or items corresponding to a single invoice are in single row. Basically, the dataset has to be prepared for the modelling. This is made possible by loading the following packages to the R file;

- readxl: This package is used to read excel files to an R file.
- plyr: This package contains tools for various purposes such as dividing, applying and joining of the data.
- dplyr: It is similar to the plyr package. It specifically provides tools for dealing with data frames which is depicted by the letter ‘d’.

- lubridate: It is a R package which makes it simpler to deal with dates and times in R studio.

All the mentioned packages helps to convert the given excel into the form that is suitable form performing the data mining process and generating relevant association rules. This is carried out through the following steps:

1. The excel file is read into the R studio with the help of *read_excel ()* function. This function is present in readxl package.
2. The use of *complete.cases (data)* function makes it possible to ignore rows with missing values in the provided dataset.
3. The Description and Country columns in the dataset is converted into factors column. This is made possible with the help of the *mutate ()* function, which is used to add or edit columns in a data-frame, from the dplyr package.
4. The date and time of transaction is stored separately in different columns. This is carried out by converting the date, present in InvoiceDate, from character format to date format using the *as.Date ()* function. On the other hand, time is extracted from the InvoiceDate column with the help of *format ()* function.
5. The InvoiceNo is transformed from character format to numeric format by using *as.numeric ()* function.
6. The transactions that contain same InvoiceNo and Date values needs to be grouped together. This is carried out using the *ddply ()* function which is a part of ply package. This is stored in a separate data-frame.
7. The InvoiceNo and Date columns are removed from the new Data-frame. This format of transaction details is termed as basket format. This will produce a data-frame that is suitable for the association rule generation.

3.1.4 Modelling

In the modelling phase, the prepared data that has been pre-processed in the previous step will be used to generate the relevant association rules. This is made possible with the aid of *apriori ()* function which is a part of arules package. The syntax of *apriori ()* is given as:

```
apriori(data, parameter = NULL, appearance = NULL, control = NULL)
```

where each argument is given as;

- **data**: This argument denotes the pre-processed data that is to be executed in the Apriori algorithm

- **parameter**: This particular argument is a list of values that control the generation of rules. This list might contain values such as support (supp), confidence (conf), maximum length (maxlen) and maximum time for checking the subset (maxtime). If values are not specified, the function takes in the default values of minimum supp as 0.1, minimum conf as 0.8, maxlen as 10 items, and a maxtime as 5 seconds.
- **appearance**: The item appearance can be made 'restricted' with the help of this argument which otherwise will be 'unrestricted' by default.
- **control**: The overall algorithmic efficiency of the association rule mining algorithm in monitored by this argument.

The summary of the generated association rule could be explored using the summary () function which will give the following details:

- **Parameter Specification**: This will show the specified values of min_sup, min_confidence , and maxlen of item provided in the rule generating function.
- **Total number of rules**: This provides the total number of association rules that was generated.
- **Distribution of rule length**: It will give the total number of rules generated for items with specific length.
- **Summary of Quality measures**: The minimum and maximum numeric values of Support, Confidence and, Lift are specified by this.
- **Information used for creating rules**: The data, support, and confidence we provided to the algorithm.

3.1.5 Evaluation

This particular phase evaluates the grade to which the generated model meets the business aims and finds a way to make the model more efficient. In order to do that we could check the top generated rules using *inspect ()* function. This will provide the support, confidence, lift and count of respective rules.

The same *apriori ()* function could produce stronger association rules by increasing the threshold value of confidence and maximum length of items.

Another step to make the rules more efficient is to remove rules that are subsets to a larger rule. This can be made possible with the help of the following functions:

- ***which()***: This function returns the location of data in the data frame if the given condition is proven as TRUE.
- ***colSums()***: This gives the sum of the rows and columns of data-frames.

- ***is.subset()***: It will find out if the contents of one particular vector consist of all the contents of another.

The appearance argument in the apriori() function makes it easier to find rules related to a specific item. This can be done in following ways:

- In order to find out the products that a customer might buy before buying a particular product *appearance = list(default="lhs", rhs="Product_name")* is applied.
- Similarly, in order to find out the products that was bought by customers together with a particular product *appearance = list(lhs="Product_name", default="rhs")*

The accuracy of a particular rule is given by the value of confidence of the respective rule. Visualising the whole dataset along with the generated rule will provide a much better way to deploy the result. This will be carried out in the next phase.

3.1.6 Deployment

The results can be deployed as visualisation of the data. The various visualisation techniques that will be used is:

1. **Exploratory Analysis of Data**

   The exploratory analysis will give the overall picture of the given data. For this purpose, we the data visualisation tool known as Tableau. The used version of tableau is 10.5.6 for Windows 64-bit. The main reason for choosing Tableau for deployment is because of the following reasons:

   - **Data visualization**
     Tableau is basically a data visualization software tool that is built for business intelligence. Foe the same reason, its build-in technology is used to help complex evaluations, data mixing and dashboard creating for the purpose of producing an attractive visualization. It will also provide interpretations which cannot simply be extracted from going through a spreadsheet. Since it is committed to this specific duty, it has jumped to the topmost of the data visualization tool competition.
   - **Fast production of interactive visualizations**
     In Tableau, one can produce a highly interactive visualisation within a little time with the utilisation of the provided drag-n-drop features of the application. The tableau interface is capable of dealing with limitless variations. On the other hand, it also limits the user from building charts which are against suitable practices of data visualization.
   - **Improved experience for users**

At present, so many diverse types of visualization choices are provided by Tableau that help in improving experience of users. Also, having a pre-knowledge of program coding is not necessary for a user to operate Tableau. So, Tableau is extremely easy to learn for a beginner.

- **Dealing with huge quantities of data**

  The Tableau application can easily deal with huge number of rows of data. Various types of visualization can be produced with the huge quantity of data while making sure that the efficiency of the dashboards are not affected. Furthermore, one of the options in Tableau allows the user to work with online connections to various data sources such as SQL.

- **Utilising other coding language**

  While working with tableau, a user can include other languages such as R or Python that will make sure that the efficiency problems are eliminated, and complex mathematical calculations are carried out with ease. By executing tasks in packages with the help of Python script the load of the tableau software can be reduced.

2. **Association Rule Visualisation**

   The results obtained from the market basket analysis will be visualised using the R studio. The package used for the purpose is arulesViz package. This package is primary built to focus of various visualisation approaches on association rules. With the help of this package, it will be possible to carry out scatter plot, two-key plot, interactive grouped plot, graph plot, parallel coordinate plot etc.

3. **Shiny Application**

   Finally, the results obtained from the R visualisation can be presented as an application with the help of shiny package. The Shiny is a part of R packages which makes it possible to construct interactive web application within R studio with ease.

3.2 Apriori Algorithm

Based on the literature review of various association rule mining algorithms, the Apriori algorithm for Market Basket Analysis is chosen. Considering the limitations and other constrains of the research, Apriori algorithm was chosen because of the following advantages:

- It is efficient for bulky datasets.
- It has the ability to produce shared computing setups.

- It boosts the effectiveness and deducts the cost of the entire process.
- It provides quicker outcome for the individual customers.

The Apriori algorithm was put forth by Agrawal and Srikant in the year of 1994. Apriori is structured to focus on databases that consist of large amount of transaction data. The basic Apriori algorithm utilised the bottom-up technique that makes it possible to extend the frequent subsets one item at a time which is termed as the candidate generation step. Further, clusters of candidates are evaluated against the data. The algorithm dismisses the process when no further relevant extensions are possible. The entire process of Apriori algorithm is based on the following terms:

1. **Itemset:** This is the collection of one or more items or products from the transactions. The term K-item-set denotes a set of k items or products.
2. **Support Count:** The total number of occurrences of a particular item-set is defined as support count

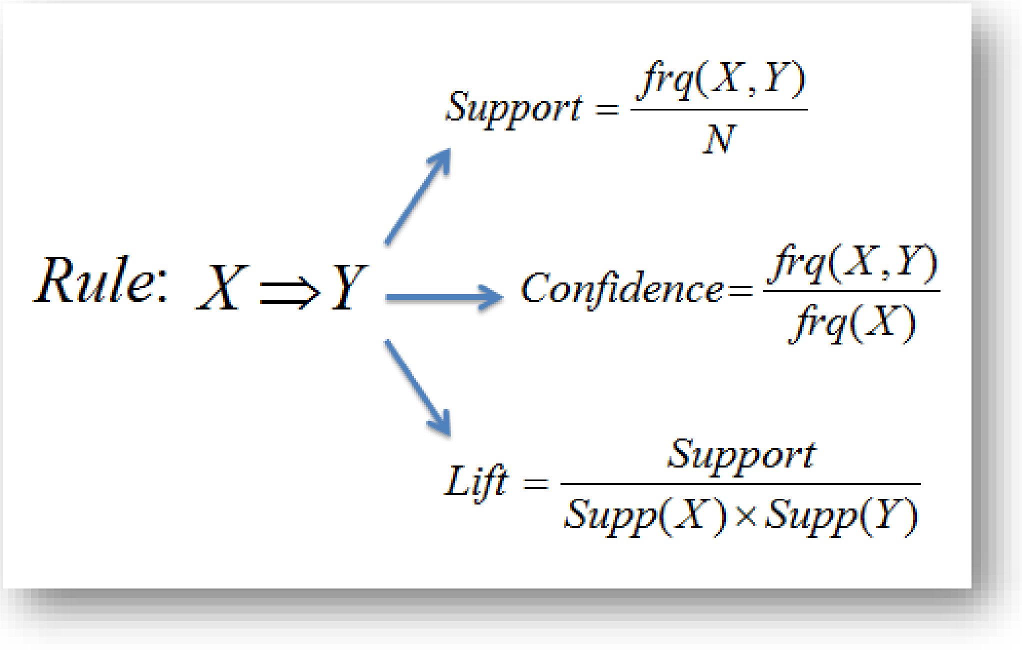

$Rule:\ X \Rightarrow Y$
$Support = \frac{frq(X,Y)}{N}$
$Confidence = \frac{frq(X,Y)}{frq(X)}$
$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

Figure 7: Support, Confidence and Lift

3. **Support:** The ratio of transactions which consist of the specific item-set. Basically, it is the division of the number of transactions of both item A and item B by the total number of transactions, where the total number of transactions is defined by N.

$$Support\ of\ (A \rightarrow B) = P(AUB) = n(AUB)\ /\ N$$

4. **Confidence:** It is defined in such a way that, the confidence for a rule A → B provides the percentage of transaction in which item B is bought along with item A. Otherwise, it is given

as the division of the total number of transactions with both item A and item B by the total number of transactions containing item A.

$$Confidence\ of\ (A \rightarrow B) = P(AUB) \,/\, P(A) = n(AUB) \,/\, n(A)$$

5. **Lift:** The value of Lift provides the relationship between item A and item B in the rule A → B. The correlation states how one item-set, say A, impacts another item-set, say B.

$$Lift\ of\ (A, B) = P(AUB) \,/\, (P(A) * P(B))$$

6. **Frequent Itemsets:** The value of Support and Confidence determines the interestingness of the generated rule. This achieved by setting the minimum support and minimum confidence thresholds. The Item-sets whose support is greater than or equal to the specified minimum support threshold is defined as the frequent itemset.

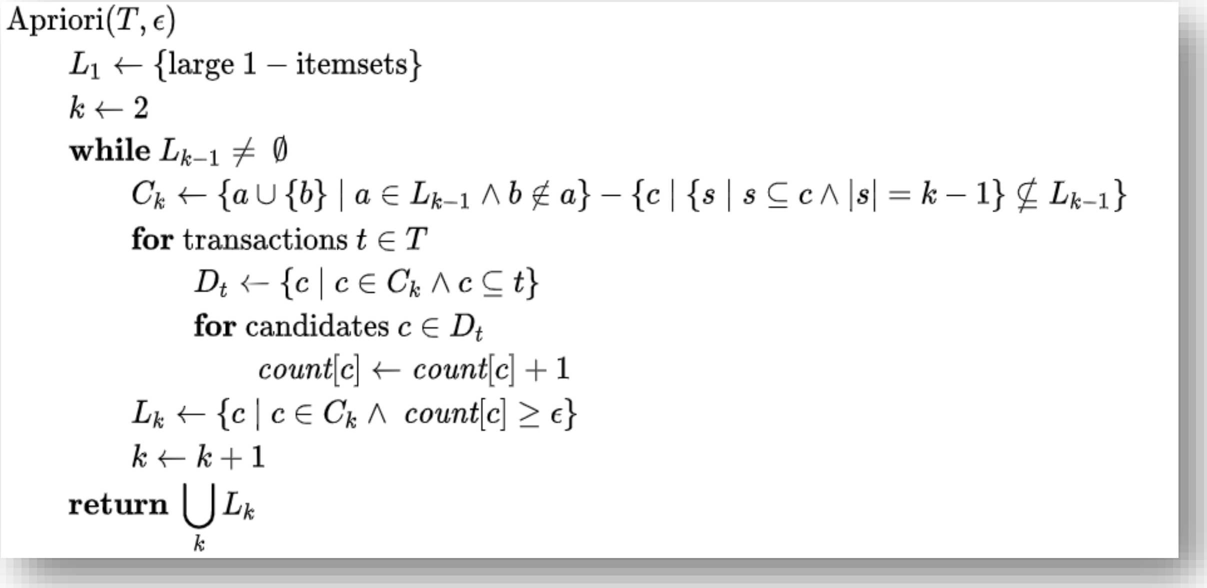
The Association Rule Mining is defined as a two-step technique:

1) **Generating the Frequent Itemset:** This step involves discovering all frequent item-sets that has a support greater than or equal to a pre-determined minimum support count.
2) **Generating the Rule:** This step involves producing all the relevant Association Rules from frequent item-sets. So, this step further involves:
    - Evaluating the Support and Confidence for all the rules.
    - Pruning the rules that does not reach the minimum support and minimum confidence threshold values.

Here, the step involving the Frequent Itemset Generation demands a complete database scan. For the same reason, it is the most expensive step to execute. The Apriori algorithm basically states that:

"*Any subset of the generated frequent itemset should also be a frequent itemset. Otherwise, the superset of an infrequent itemset will not be generated or evaluated.*"

The pseudo code for the Apriori algorithm is provided below for a transaction database termed as T, and with a support threshold given as ϵ. Ck denotes the candidate set for a level of k. The algorithm is expected to produce the candidate sets in every step from the huge item sets of the previous level. The $count[c]$ deals with a section of the data structure which denotes candidate set c, which is usually kept as zero at the beginning.



```
$\text{Apriori}(T, \epsilon)$
    $L_1 \leftarrow \{\text{large } 1 - \text{itemsets}\}$
    $k \leftarrow 2$
    while $L_{k-1} \neq \emptyset$
        $C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\}$
        for transactions $t \in T$
            $D_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$
            for candidates $c \in D_t$
                $count[c] \leftarrow count[c] + 1$
        $L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\}$
        $k \leftarrow k + 1$
    return $\bigcup_k L_k$
```

Figure 8: Apriori pseudo code

The flowchart of the Apriori algorithm is given as:

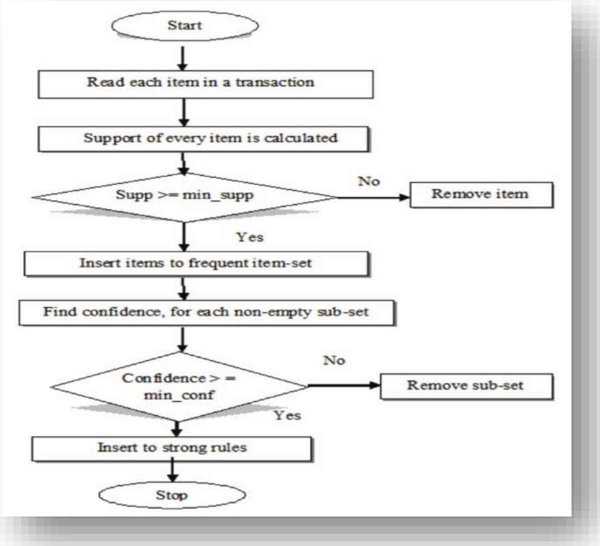
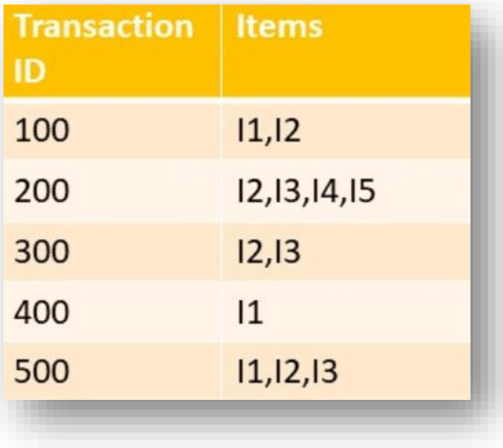

Start
Read each item in a transaction
Support of every item is calculated
Supp >= min_supp
No
Remove item
Yes
Insert items to frequent item-set
Find confidence, for each non-empty sub-set
Confidence > = min_conf
No
Remove sub-set
Yes
Insert to strong rules
Stop

Figure 9: Apriori Flowchart

The entire process of Apriori algorithm can be explained with an example. Consider a transaction data, where items are named as l1, l2, l3,etc. The minimum support and minimum confidence can be assumed to be 2 and 50% respectively.

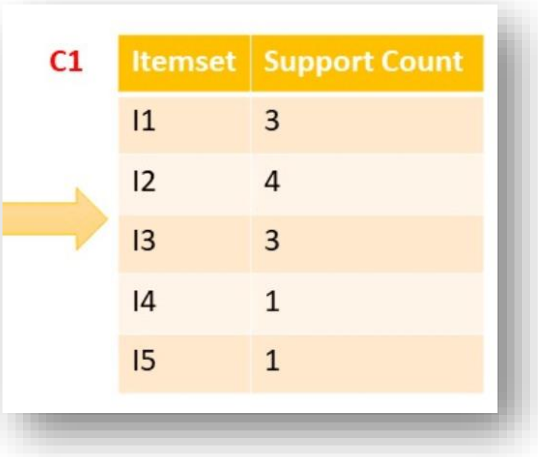


Transaction ID	Items
100	I1,I2
200	I2,I3,I4,I5
300	I2,I3
400	I1
500	I1,I2,I3

Figure 10: Sample Transactional Data

The following steps are carried out in order to get relevant association rules:

1. First you will start with all individual items called candidates and calculate their support count. This is called candidate list generation.

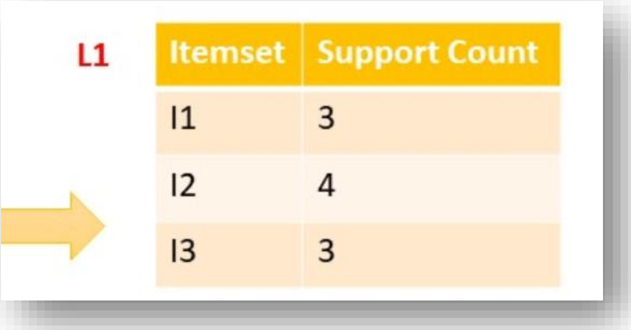


C1

Itemset	Support Count
I1	3
I2	4
I3	3
I4	1
I5	1

Figure 11: Apriori (Step 1)

2. Remove candidates that fail minimum support count. I4 and I5 fail minimum support that is equal to 2. The list is now called L1 containing the frequent Itemsets. Here we have used the Apriori principle: no superset of infrequent itemset must be generated and tested.

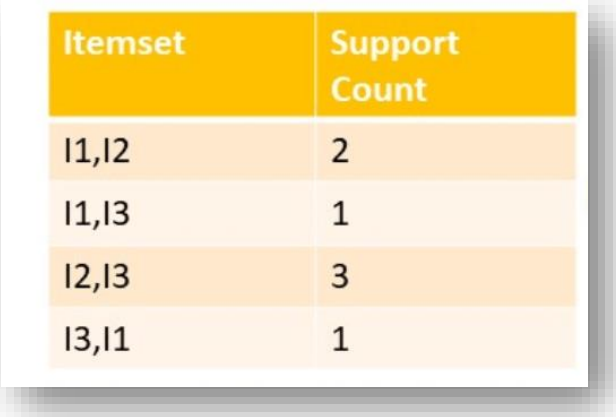


L1

Itemset	Support Count
I1	3
I2	4
I3	3

Figure 12: Apriori (Step 2)

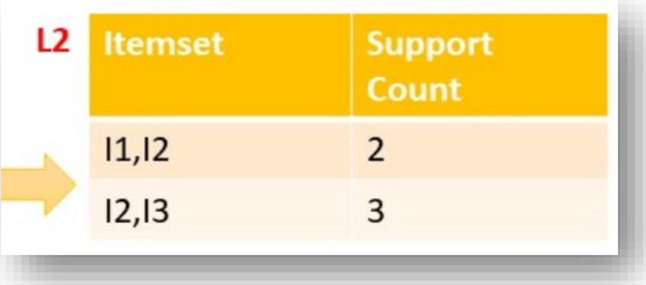
3. Generate second candidate list by L1 Cross join L1. And note support counts. {I1,I2} appear in two transactions together.



Itemset	Support Count
I1,I2	2
I1,I3	1
I2,I3	3
I3,I1	1

Figure 13: Apriori (Step 3)

4. Remove candidates that fail minimum support count.

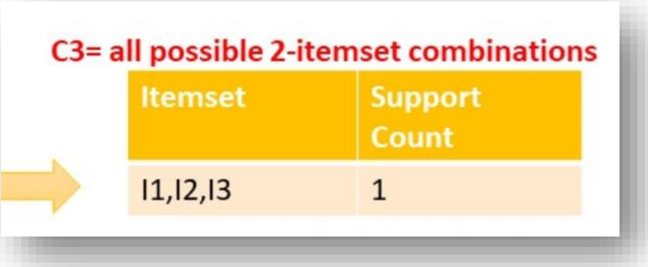


L2

Itemset	Support Count
I1,I2	2
I2,I3	3

Figure 14: Apriori (Step 4)

5. Generate third candidate list by L2 cross join L2. And note support counts. {I1,I2,I3} appear in only one transaction together.



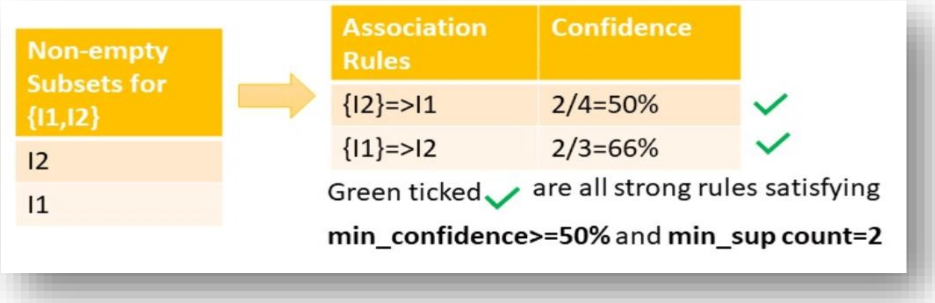
C3= all possible 2-itemset combinations

Itemset	Support Count
I1,I2,I3	1

Figure 15: Apriori (Step 5)

6. L3 is null. L3={} send support count for {I1,I2,I3} fails minimum support. The first step of association rule mining is completed and there will be no C4 candidate list.

The following step in the process will be to gather all the frequent Itemsets. For the same purpose the last non-empty Frequent Itemset is taken, that is L2= {I1, I2}, {I2, I3}. Every non-empty subsets of the item-sets contained in the chosen Frequent Item-set List is considered.

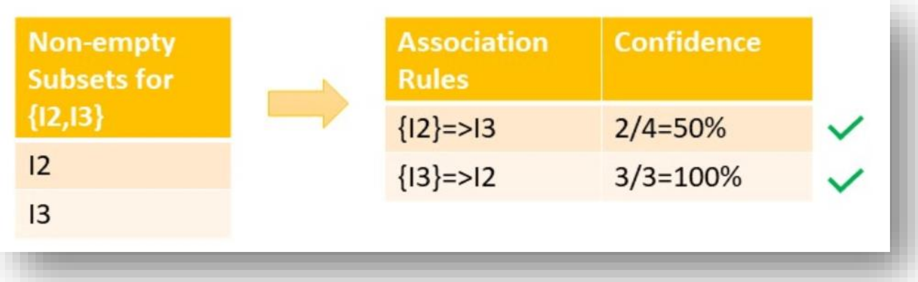


Non-empty Subsets for {I1,I2}
I2
I1

Association Rules	Confidence	
{I2}=>I1	2/4=50%	✓
{I1}=>I2	2/3=66%	✓

Green ticked ✓ are all strong rules satisfying **min_confidence>=50%** and **min_sup count=2**

Figure 16: Apriori (Step 6.1)



Non-empty Subsets for {I2,I3}
I2
I3

Association Rules	Confidence	
{I2}=>I3	2/4=50%	✓
{I3}=>I2	3/3=100%	✓

Figure 17: Apriori (Step 6.2)

The above figures provides the approach used of selecting the appropriate rule. Here, four strong rules are created from the transaction list. In this case, when I2 → I3 is evaluated, it is revealed that the confidence of the rule equal to 50% which denotes that 50% of customers who purchased I2 also purchased I3.

3.3 Online Sentiment Survey

As a part of primary research, the emotions of people to a particular product or service expressed through various social media will be data mined. This was made possible with the help of a website known as Sentione. Sentione is an online application that focuses on social listening by monitoring the online discussion on a particular keyword. The monitoring will be carried out through various online platforms such as youtube.com, facebook.com, Instagram.com, twitter.com etc. Further, the insights of the discussion will be also provided by the application.

The working of the website in based on Text Mining which is described as a process of exploring through and evaluating huge quantity of text data that will be unstructured in nature. Here, the emotions of a particular user is determined by exploring through significant words in the content. The data mining will allow us to find out whether the user is satisfied with the product or service.

In order to access the services of Sentione, a free trial version for 15 days will be used which will be obtained once a user registers to the site. The main objective behind conducting this survey is to find out how people reacts to the concept of Scan and Go shopping. As Amazon Go store is the only fully automated Scan and Go store that is successfully running at the moment, the key word to be searched is "Amazon Go". The advantages of using sentiment analysis for survey is:

- Faster processing as the data used is already present at the moment of survey.
- The survey will be focused only on people that are aware of the concept.

3.4 Ethics of Research

The privacy of the data collected during the primary research will have be an ethical issue if it fails to obey General Data Protection Regulation (GDPR). It is basically a regulation carried out in the existing EU law focused on data protection and privacy. This is applicable for all individuals residing in the European Union (EU) along with those residing in the European Economic Area (EEA).

The dataset collected for the purpose of building the recommendation model is taken from the UCI Machine Learning Repository. This is repository that contain open source datasets.

The online sentiment survey by Sentione also follows the GDPR and the privacy policy along with the checklist is provided in the Appendices.

Further, this thesis paper also follows the ethics demanded by Dublin Business School (DBS) while performing the primary and secondary research.

3.5 Limitations of the methodology

As mentioned in the literature review section, there are some disadvantages in using Apriori algorithm. But the secondary researches show that it is one of the most effective techniques while dealing with large amounts of data. The concept of Big Data analytics was researched but not used due to fact that the successful completion of the particular concept requires high-end systems and paid services which is not feasible for the research. But this will not affect the main objective of the dissertation.

The primary research regarding building the recommender system model will be conducted with the help of data that was posted 6 years ago. The collection of recent data will not successful because of the newly stated GDPR law in EU. However, the dataset used is similar to any other transaction. So, this won't impact the research or artefact that will be produced.

Furthermore, the primary research regarding the online sentimental survey will be limited by the free trial version of the application. More features could have been accessed with the paid service which is not feasible. So, the survey was efficiently conducted with the smart usage of available services.

Finally, as mentioned earlier in the paper, time was a major constrain as the whole research was limited to two months. However, by using online sentiment survey for primary research and by selecting feasible algorithm the research paper along with the artefact is expected to be produced within the time limit.

4. Results and Findings

This section provides the results obtained from carrying out the market basket analysis aa well as the online survey. The subsections of this chapter includes findings from the exploratory analysis of data, pre-processing of data, association rule mining, association rule visualisation, Shiny application and online survey.

4.1 Exploratory Analysis

The dataset that was used for carrying out the Market Basket Analysis was explored using the data exploratory technique with the help of Tableau. The following results were obtained:

4.1.1 The Frequency of Purchase

The bar plot with number of records on Y-axis and different products on X-axis was created. This is to find out the most purchased products in that period. The resulting plot shows that the product with most sale is 'WHITE HANGING HEART T-LIGHT HOLDER' with 2,369 records which is followed by 'REGENCY CAKESTAND 3 TIER' with 2,200 sales.

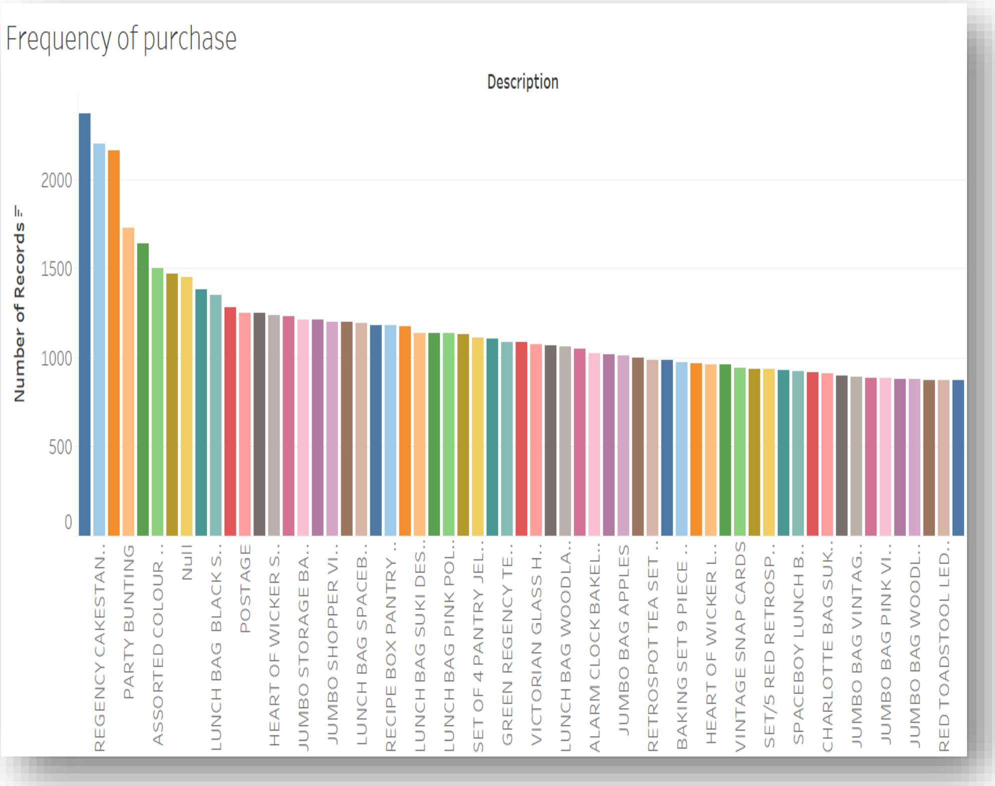
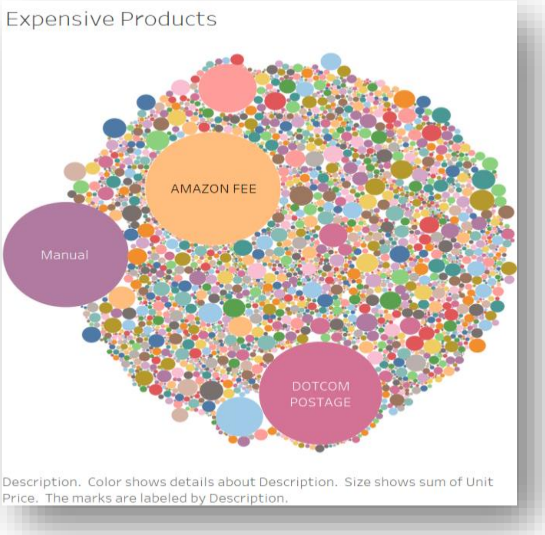


Figure 18: Frequency of Purchase

4.1.2 Most Expensive Product

In order to find the most expensive product, packed bubbles visualisation was used. The size of the bubbles will larger as the product unit price is higher. Each product was given different colours. It was found out that Amazon fee was the most expensive product

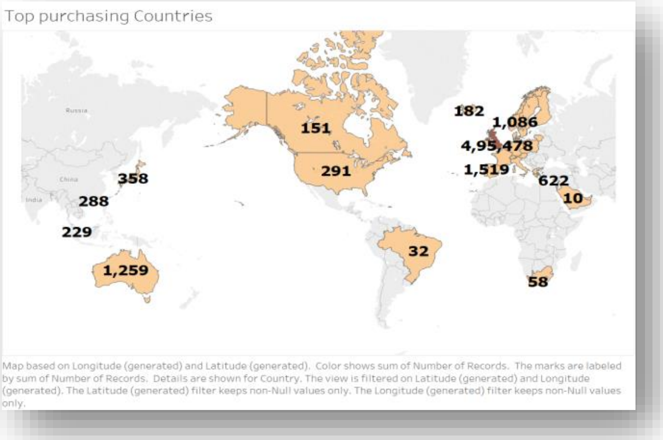


Description. Color shows details about Description. Size shows sum of Unit Price. The marks are labeled by Description.

Figure 19: Most Expensive Product

4.1.3 Top Purchasing Countries

The map plot of different countries indicating the number of records was created. The colour of the country changes from light orange to dark orange as the number of records increases. The country with the greatest number of sales was found to be from United Kingdom with 495,478 records. On the other hand, Saudi Arabia recorded the lowest sales number with 10 sales.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. The marks are labeled by sum of Number of Records. Details are shown for Country. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

Figure 20: Top Purchasing Countries Map

4.1.4 Timeline of Purchase

The timeline of the purchases done during the give period in the dataset was plotted. It was observed that the highest sale was recorder on 5 December 2011 with 5,331 number of records. Whereas, the lowest sale was carried out on 6 February 2011 with 279 records.

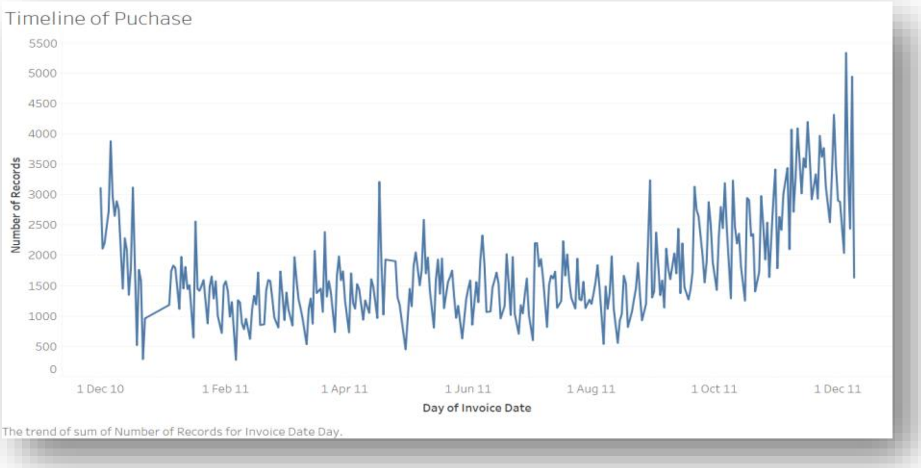


Figure 21: Rush hours of Purchase

4.1.5 Dashboard

The 4 workbooks were combined together to form a dashboard. This will provide a better comparable view of the exploratory analysis.

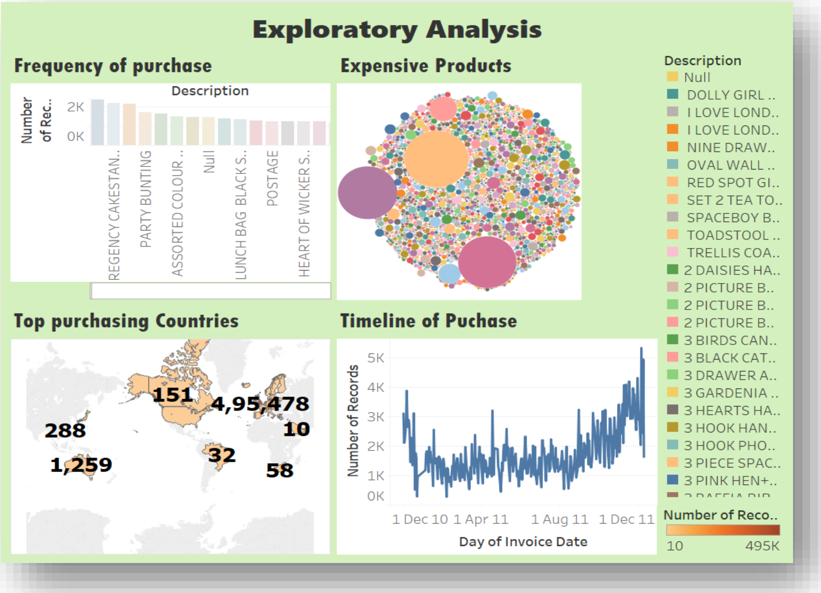
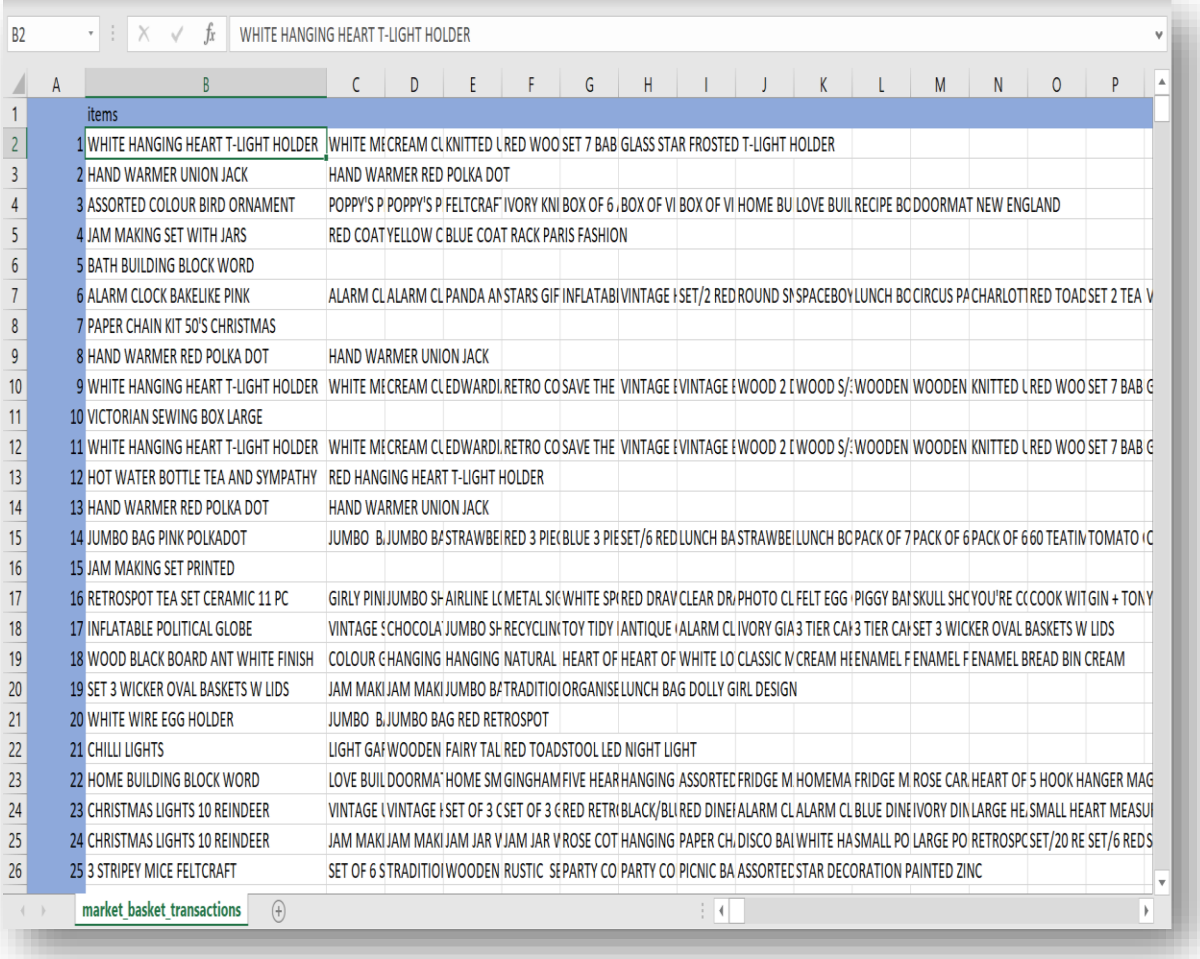


Figure 22: Tableau Dashboard

4.2 Data Pre-processing

The raw dataset obtained from the repository is not apt for the purpose of association rule mining. The data was put through a series of filtering and conversion with the help of some R packages in order to get data in the basket form. The transactions are converted into sparse format that contain 22191 rows of transactions along with 30066 columns of products or items . A sparse matrix, otherwise known as sparse array, is defined as a matrix that contain mostly zero as its elements are. However, the density of the given matrix is considered high when most of the elements contained in it is nonzero. The sparsity of a given matrix is defined as the division of the total amount of zero-valued elements by the total amount of elements. That means the sparsity of a matrix is also same as the value obtained after subtracting the density of the matrix from 1. The *summary()* function provided that the density of the dataset used is 0.0005390256. This value multiplied by the product of number of rows and columns of the matrix will give the total number of items purchased. The figure below shows the prepared data for association rule mining.



B2 | WHITE HANGING HEART T-LIGHT HOLDER

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		items														
2	1	WHITE HANGING HEART T-LIGHT HOLDER	WHITE ME	CREAM CL	KNITTED U	RED WOO	SET 7 BAB	GLASS STAR FROSTED T-LIGHT HOLDER								
3	2	HAND WARMER UNION JACK	HAND WARMER RED POLKA DOT													
4	3	ASSORTED COLOUR BIRD ORNAMENT	POPPY'S P	POPPY'S P	FELTCRAF	IVORY KNI	BOX OF 6 /	BOX OF VI	BOX OF VI	HOME BU	LOVE BUIL	RECIPE BO	DOORMAT NEW ENGLAND			
5	4	JAM MAKING SET WITH JARS	RED COAT	YELLOW C	BLUE COAT RACK PARIS FASHION											
6	5	BATH BUILDING BLOCK WORD														
7	6	ALARM CLOCK BAKELIKE PINK	ALARM CL	ALARM CL	PANDA AN	STARS GIF	INFLATABI	VINTAGE H	SET/2 RED	ROUND SN	SPACEBOY	LUNCH BO	CIRCUS PA	CHARLOTT	RED TOAD	SET 2 TEA
8	7	PAPER CHAIN KIT 50'S CHRISTMAS														
9	8	HAND WARMER RED POLKA DOT	HAND WARMER UNION JACK													
10	9	WHITE HANGING HEART T-LIGHT HOLDER	WHITE ME	CREAM CL	EDWARDI	RETRO CO	SAVE THE	VINTAGE E	VINTAGE E	WOOD 2 D	WOOD S/3	WOODEN	WOODEN	KNITTED U	RED WOO	SET 7 BAB
11	10	VICTORIAN SEWING BOX LARGE														
12	11	WHITE HANGING HEART T-LIGHT HOLDER	WHITE ME	CREAM CL	EDWARDI	RETRO CO	SAVE THE	VINTAGE E	VINTAGE E	WOOD 2 D	WOOD S/3	WOODEN	WOODEN	KNITTED U	RED WOO	SET 7 BAB
13	12	HOT WATER BOTTLE TEA AND SYMPATHY	RED HANGING HEART T-LIGHT HOLDER													
14	13	HAND WARMER RED POLKA DOT	HAND WARMER UNION JACK													
15	14	JUMBO BAG PINK POLKADOT	JUMBO B	JUMBO BA	STRAWBE	RED 3 PIE	BLUE 3 PIE	SET/6 RED	LUNCH BA	STRAWBE	LUNCH BO	PACK OF 7	PACK OF 6	PACK OF 6	60 TEATIM	TOMATO
16	15	JAM MAKING SET PRINTED														
17	16	RETROSPOT TEA SET CERAMIC 11 PC	GIRLY PIN	JUMBO SH	AIRLINE LO	METAL SIG	WHITE SP	RED DRAW	CLEAR DR	PHOTO CL	FELT EGG	PIGGY BA	SKULL SHO	YOU'RE CO	COOK WIT	GIN + TON
18	17	INFLATABLE POLITICAL GLOBE	VINTAGE S	CHOCOLA	JUMBO SH	RECYCLIN	TOY TIDY	ANTIQUE	ALARM CL	IVORY GIA	3 TIER CA	3 TIER CA	SET 3 WICKER OVAL BASKETS W LIDS			
19	18	WOOD BLACK BOARD ANT WHITE FINISH	COLOUR G	HANGING	HANGING	NATURAL	HEART OF	HEART OF	WHITE LO	CLASSIC M	CREAM HE	ENAMEL F	ENAMEL F	ENAMEL BREAD BIN CREAM		
20	19	SET 3 WICKER OVAL BASKETS W LIDS	JAM MAKI	JAM MAKI	JUMBO BA	TRADITIO	ORGANISE	LUNCH BAG DOLLY GIRL DESIGN								
21	20	WHITE WIRE EGG HOLDER	JUMBO B	JUMBO BAG RED RETROSPOT												
22	21	CHILLI LIGHTS	LIGHT GAR	WOODEN	FAIRY TAL	RED TOADSTOOL LED NIGHT LIGHT										
23	22	HOME BUILDING BLOCK WORD	LOVE BUIL	DOORMAT	HOME SM	GINGHAM	FIVE HEAR	HANGING	ASSORTED	FRIDGE M	HOMEMA	FRIDGE M	ROSE CAR	HEART OF	5 HOOK HANGER MAG	
24	23	CHRISTMAS LIGHTS 10 REINDEER	VINTAGE U	VINTAGE H	SET OF 3 C	SET OF 3 C	RED RETR	BLACK/BL	RED DINE	ALARM CL	ALARM CL	BLUE DINE	IVORY DIN	LARGE HE	SMALL HEART MEASU	
25	24	CHRISTMAS LIGHTS 10 REINDEER	JAM MAKI	JAM MAKI	JAM JAR V	JAM JAR V	ROSE COT	HANGING	PAPER CH	DISCO BA	WHITE HA	SMALL PO	LARGE PO	RETROSPC	SET/20 RE	SET/6 RED S
26	25	3 STRIPEY MICE FELTCRAFT	SET OF 6 S	TRADITIO	WOODEN	RUSTIC SE	PARTY CO	PARTY CO	PICNIC BA	ASSORTED	STAR DECORATION PAINTED ZINC					

Figure 23: Prepared Data

4.3 Association Rule Mining

The most important part of the whole dissertation involves carrying out the association rule mining using the Apriori algorithm. For the same purpose, the prepared data was put through the *apriori()* function in R studio. The parameters of the function was set as minimum support to 0.001 and minimum confidence as 75%. The mining was limited by the maximum length of patterns to 10. This process produced as set 69358 rules that followed the parameter specification. The table shows the basic summary of the generated rules.

Table 1: Summary of Generated Association Rules

rule length distribution (lhs + rhs):sizes

2	3	4	5	6	7	8	9	10
141	3385	11413	24723	19517	7210	2175	665	129

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000	5.000	5.000	5.363	6.000	10.000

summary of quality measures:

support	confidence	lift	count
Min. :0.001036	Min. :0.7500	Min. : 9.165	Min. : 23.00
1st Qu.:0.001082	1st Qu.:0.7931	1st Qu.: 20.902	1st Qu.: 24.00
Median :0.001262	Median :0.8444	Median : 26.396	Median : 28.00
Mean :0.001418	Mean :0.8536	Mean : 53.975	Mean : 31.47
3rd Qu.:0.001532	3rd Qu.:0.9062	3rd Qu.: 44.928	3rd Qu.: 34.00
Max. :0.022757	Max. :1.0000	Max. :715.839	Max. :505.00

info:

data	ntransactions	support	confidence
Tran	22191	0.001	0.75

The table provides the details including Parameter Specification, Distribution of rule length, Summary of Quality measures and Information used for creating rules. Out of the 69358 rules that was generated, top 10 rules were inspected. The top 10 rules are given in the following table.

Table 2: Top Association Rules

	lhs		rhs	support	confidence	lift	count
[1]	{WOBBLY CHICKEN}	=>	{DECORATION}	0.001262	1	443.82	28
[2]	{WOBBLY CHICKEN}	=>	{METAL}	0.001262	1	443.82	28
[3]	{DECOUPAGE}	=>	{GREETING CARD}	0.001036	1	389.3158	23
[4]	{BILLBOARD FONTS DESIGN}	=>	{WRAP}	0.001307	1	715.8387	29
[5]	{WRAP}	=>	{BILLBOARD FONTS DESIGN}	0.001307	0.935484	715.8387	29
[6]	{ENAMEL PINK TEA CONTAINER}	=>	{ENAMEL PINK COFFEE CONTAINER}	0.001397	0.815789	385.1741	31
[7]	{LANDMARK FRAME COVENT GARDEN}	=>	{LANDMARK FRAME OXFORD STREET}	0.001487	0.785714	396.2679	33
[8]	{LANDMARK FRAME OXFORD STREET}	=>	{LANDMARK FRAME COVENT GARDEN}	0.001487	0.75	396.2679	33
[9]	{WOBBLY RABBIT}	=>	{DECORATION}	0.001532	1	443.82	34
[10]	{WOBBLY RABBIT}	=>	{METAL}	0.001532	1	443.82	34

The rules suggests that every customer who purchased 'WOBBLY CHICKEN' also purchased ‘DECORATION’. Similarly, every customer who purchased ‘BILLBOARD FONTS DESIGN’ also purchased ‘WRAP’. It was observed that sorting the generated rules based on the value of lift gave more relevant rules. The obtained rules were further evaluated, and the following findings were recorded:

- A set of stronger rules was obtained when the value of minimum confidence threshold was increased.
- The maximum length was extended in order to get a set of extended rules.
- The redundant rules were eliminated by removing smaller subset rules from larger superset rules.
- The rules related to a particular item was found out by specifying the product name in the LHS and RHS.

4.4 Association Rule Visualisation

The generated association rules are visualised with the help of arulesViz package in R. The visualisation provided a better way to explore the data. The following visualisation techniques were used:

4.4.1 Scatter Plot

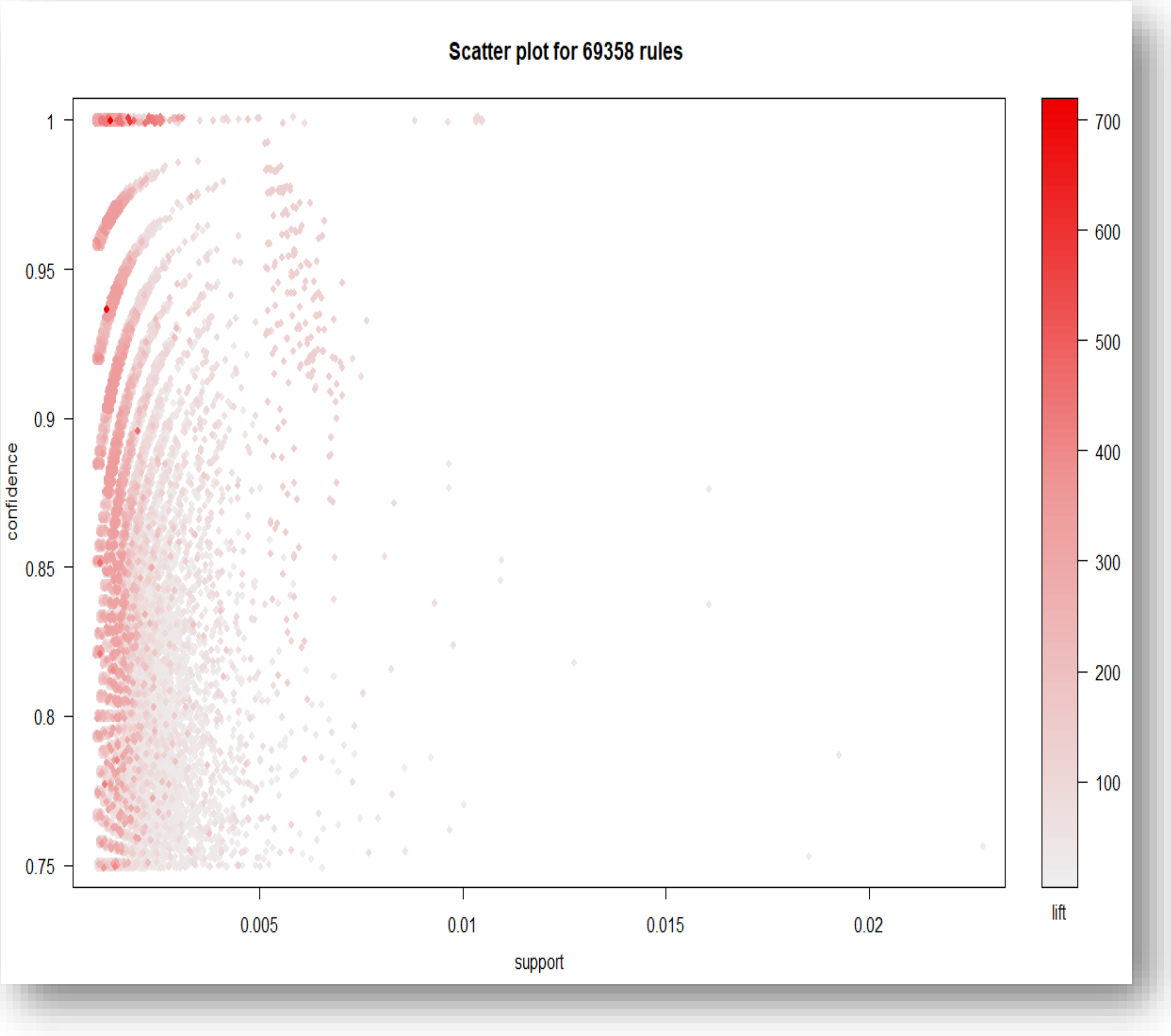


Figure 24: Scatter Plot

The above figure shows the scatter plot. The value of the method argument in the *plot ()* function created a scatter plot of the generated rules. The plot have support as the X-axis and confidence as the Y-axis. The corresponding value of the lift is used for plotting the points. The plot indicates that the rules that have high value of lift will be having a low value of support.

4.4.2 Two-key Plot

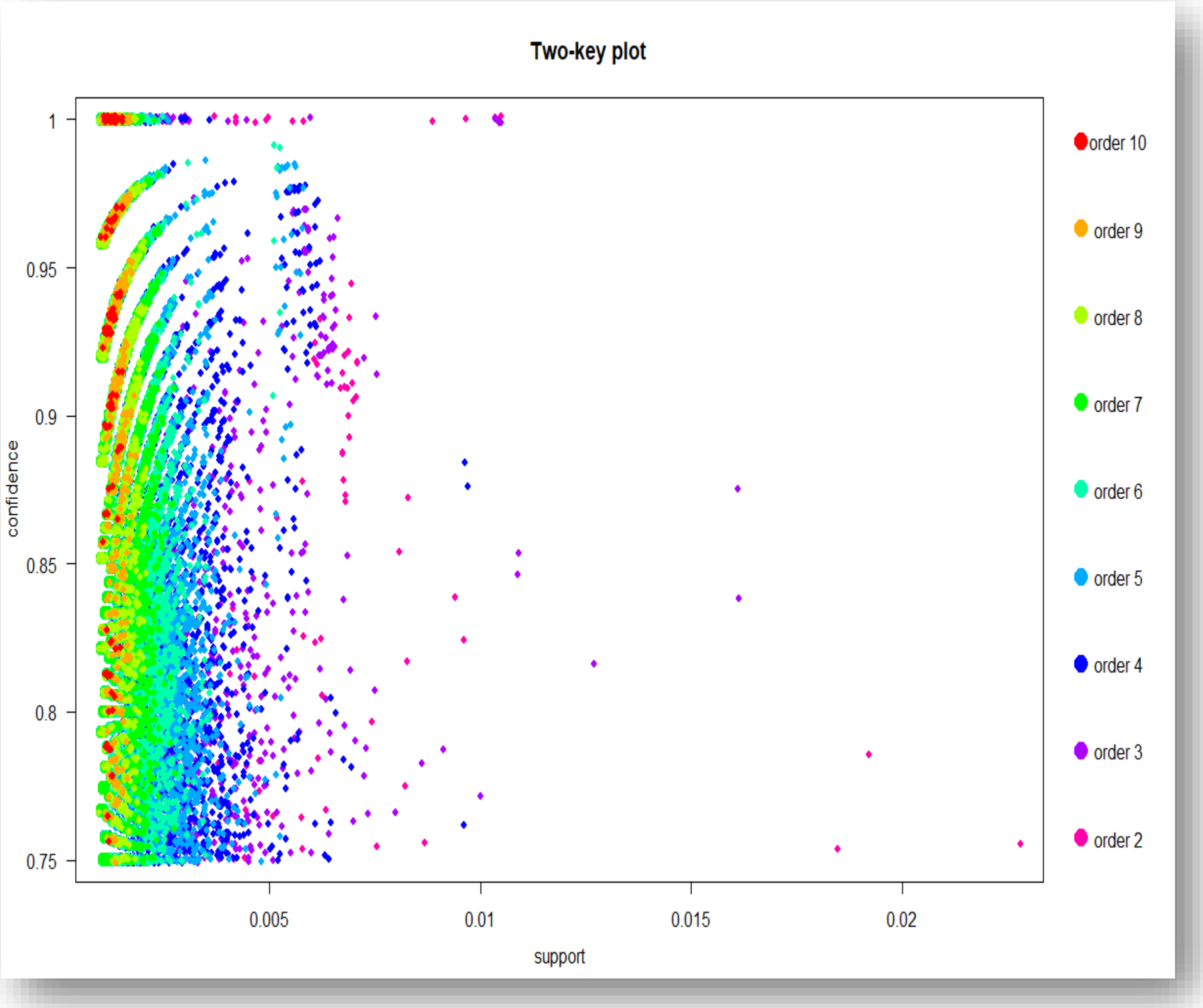


Figure 25: Two-Key Plot

The above figure shows the Two-key plot of the generated rules. This was created by specifying the method argument as “two-key plot”. It is similar to the scatter plot in terms of the X and Y axis representation. For colouring the points, the number of items in the rule or order is utilised.

4.4.3 Grouped Interactive Plot

A grouped interactive plot of the generated rules is created as shown in figure. This was made possible by setting the method argument as “grouped” and Interactive argument as “TRUE”. The various quality measures such as support, confidence and lift of each point can be seen by hovering over the specific point. The Y-axis and X-axis is represented by the RHS and LHS respectively.

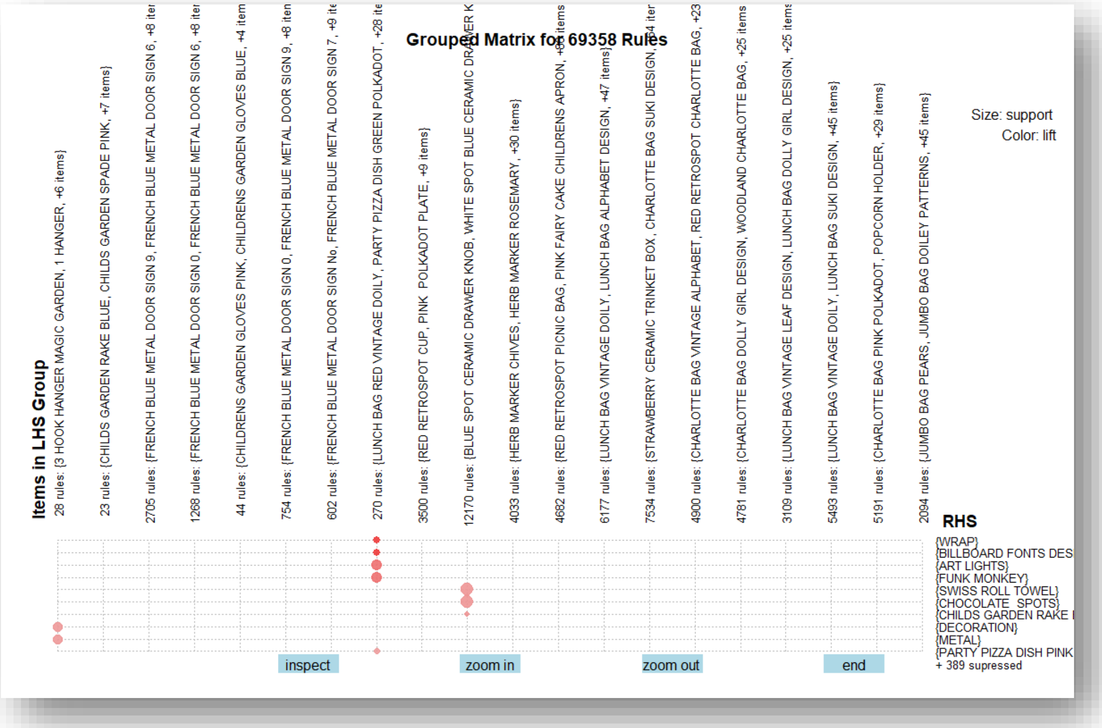


Figure 26: Grouped Interactive Plot

4.4.4 Graph Based Plot

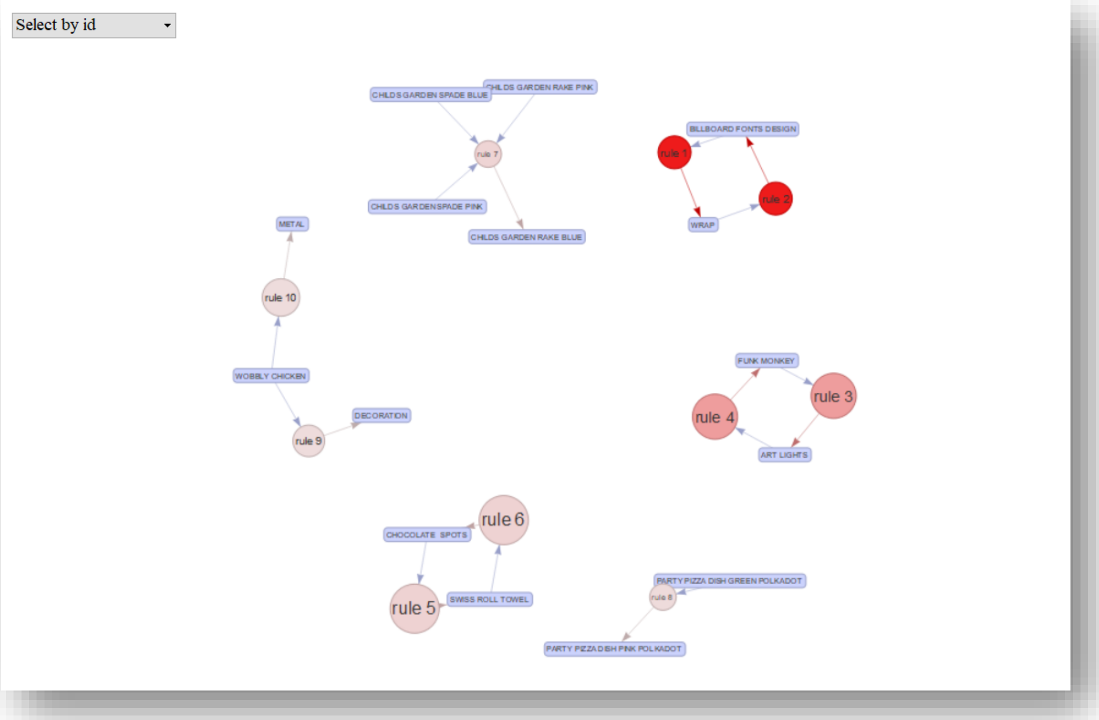


Figure 27: Graph Plot

Another interesting way to represent the rules in a pictorial way is to build a graph-based visualisation. This was created by specifying the method argument as “graph” and engine argument as ”htmlwidget”. As seen in the figure above, this technique utilises the vertices and edges. Here, the vertices are named with the product names, and another set of vertices denotes the item sets or the generated rules. The LHS products are denoted by the arrows directing from product to rule vertices whereas, the arrow directing from a rule to a product denotes the RHS. Usually, the interest measures are indicated by the colour and size of vertices. In order to avoid over plotting, only the top rules were selected to represent the graph.

4.4.5 Parallel Coordinate Plot

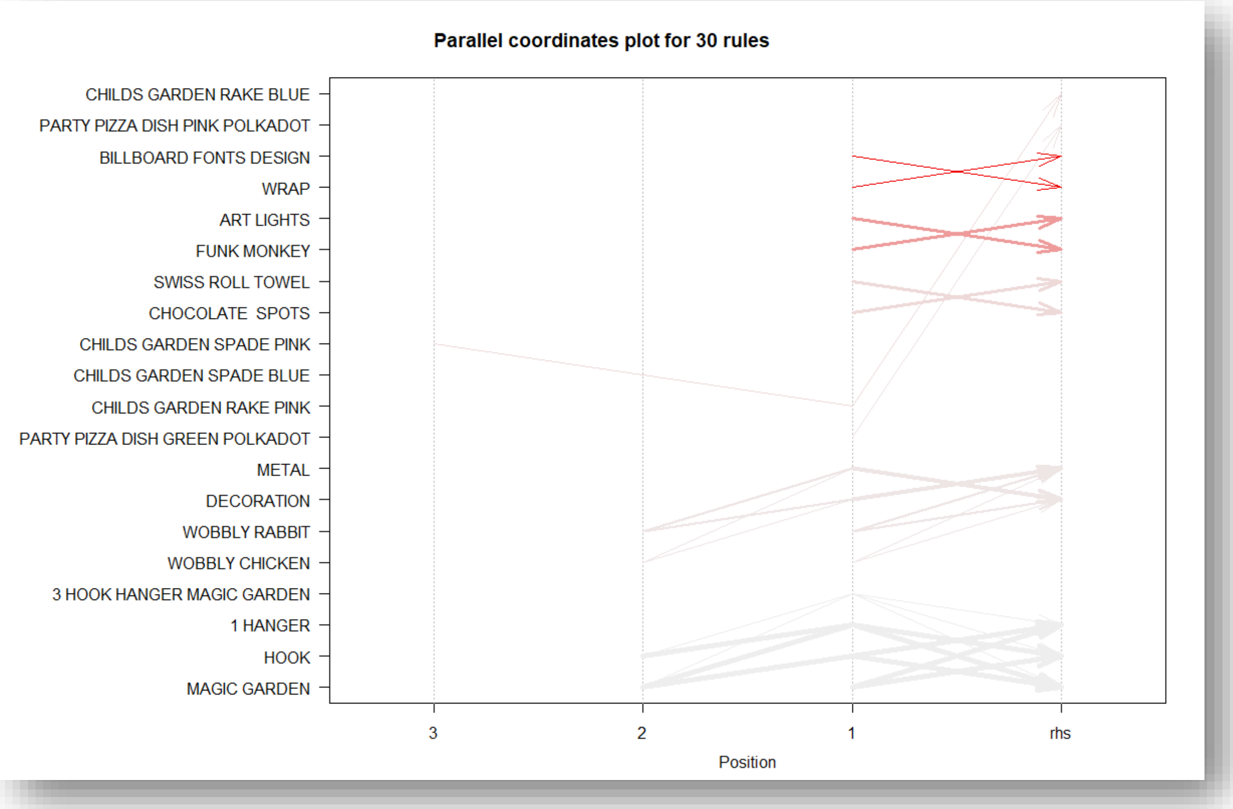
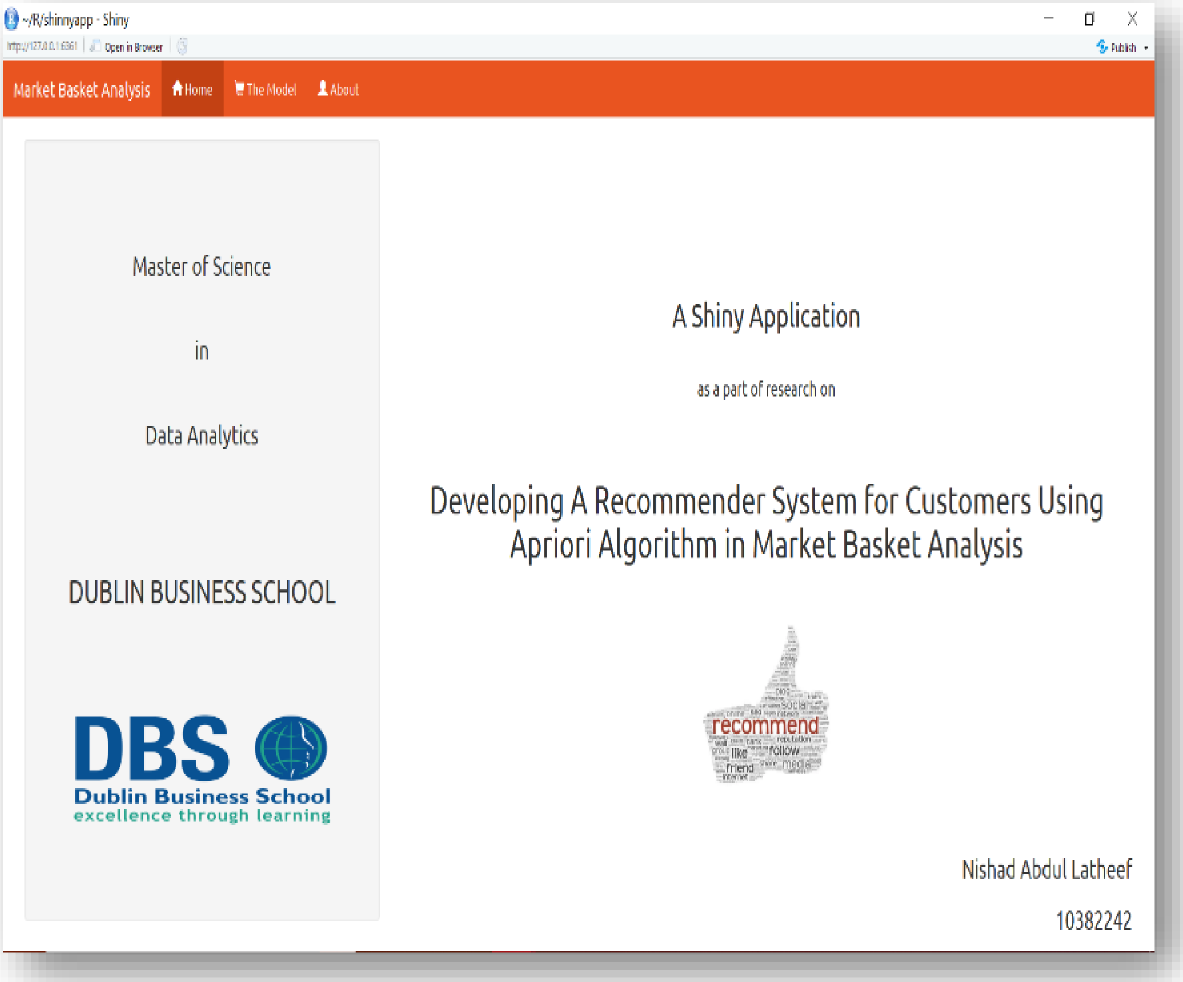


Figure 28: Parallel Coordinate Plot

This above figure shows the parallel coordinate plot of the generated rule. By setting the method argument as “paracoord”, the plot can be created. This technique is employed so as to graphically represent what type of sales would be carried out for those specific items along with other items. Basically, the RHS is the produce that a customer wishes to purchase, and the LHS represents the positions such that the number 2 is the newest purchase to the customer’s basket and the number 1 is the product that the customer purchased just before that.

4.5 The Shiny Application

The process of Market basket analysis is presented in a web application with the help of shiny library in R. This provided a much more interactive way to explore the data. Two slider inputs are provided for both minimum support and minimum confidence that can be varied to produce various inputs. The screenshots of the working application is given below.



Master of Science

in

Data Analytics

DUBLIN BUSINESS SCHOOL

DBS Dublin Business School excellence through learning

A Shiny Application

as a part of research on

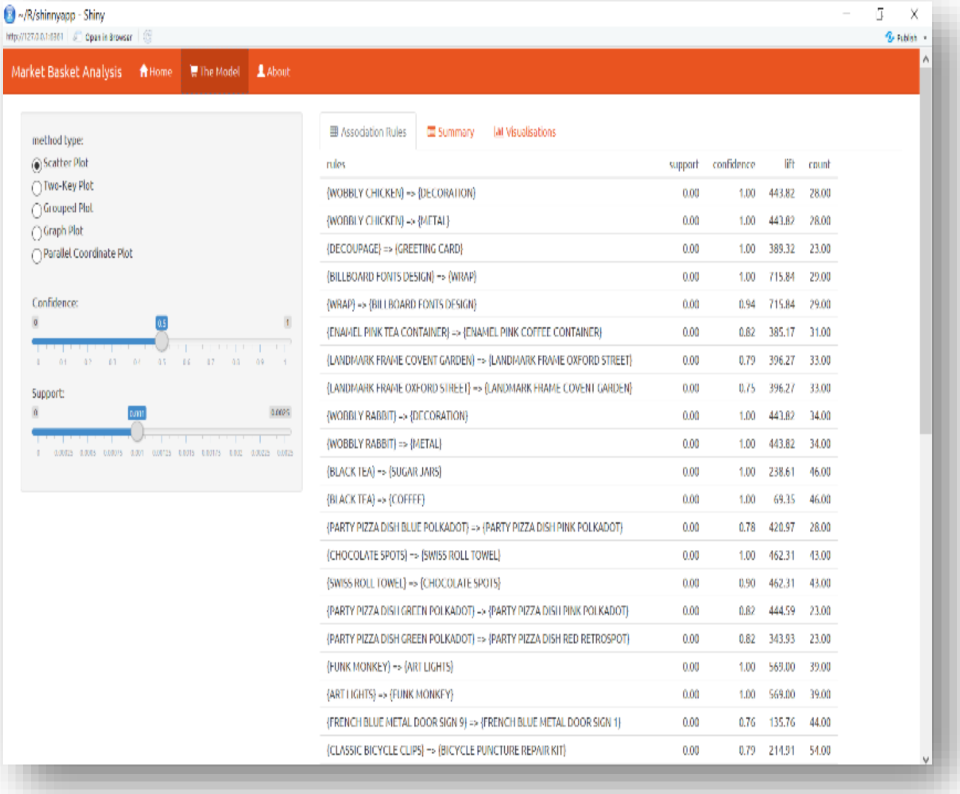
Developing A Recommender System for Customers Using Apriori Algorithm in Market Basket Analysis

Nishad Abdul Latheef

10382242

Figure 29: Home Page (Shiny App)

The Home page, as given above, is the first page that is displayed when the app is opened. This provides the topic of the dissertation along with navigation button that will lead to Model page or About page. The three sub-pages of model page are given below. This include displaying top rules, summary and visualisation of the generated rules in various ways.



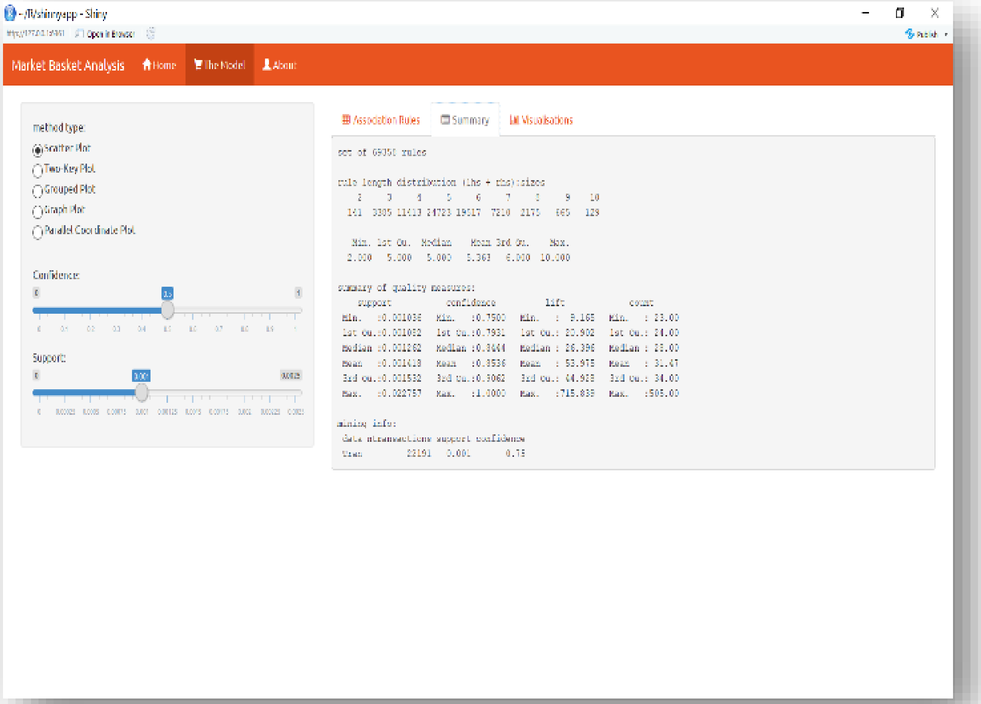
method type:
● Scatter Plot
○ Two-Key Plot
○ Grouped Plot
○ Graph Plot
○ Parallel Coordinate Plot

Confidence: 0.5

Support: 0.001

rules	support	confidence	lift	count
{WOBBLY CHICKEN} => {DECORATION}	0.00	1.00	443.82	28.00
{WOBBLY CHICKEN} => {METAL}	0.00	1.00	443.82	28.00
{DECOUPAGE} => {GREETING CARD}	0.00	1.00	389.32	23.00
{BILLBOARD FONTS DESIGN} => {WRAP}	0.00	1.00	715.84	29.00
{WRAP} => {BILLBOARD FONTS DESIGN}	0.00	0.94	715.84	29.00
{ENAMEL PINK TEA CONTAINER} => {ENAMEL PINK COFFEE CONTAINER}	0.00	0.82	385.17	31.00
{LANDMARK FRAME COVENT GARDEN} => {LANDMARK FRAME OXFORD STREET}	0.00	0.79	396.27	33.00
{LANDMARK FRAME OXFORD STREET} => {LANDMARK FRAME COVENT GARDEN}	0.00	0.75	396.27	33.00
{WOBBLY RABBIT} => {DECORATION}	0.00	1.00	443.82	34.00
{WOBBLY RABBIT} => {METAL}	0.00	1.00	443.82	34.00
{BLACK TEA} => {SUGAR JARS}	0.00	1.00	238.61	46.00
{BLACK TEA} => {COFFEE}	0.00	1.00	69.35	46.00
{PARTY PIZZA DISH BLUE POLKADOT} => {PARTY PIZZA DISH PINK POLKADOT}	0.00	0.78	420.97	28.00
{CHOCOLATE SPOTS} => {SWISS ROLL TOWEL}	0.00	1.00	462.31	43.00
{SWISS ROLL TOWEL} => {CHOCOLATE SPOTS}	0.00	0.90	462.31	43.00
{PARTY PIZZA DISH GREEN POLKADOT} => {PARTY PIZZA DISH PINK POLKADOT}	0.00	0.82	444.59	23.00
{PARTY PIZZA DISH GREEN POLKADOT} => {PARTY PIZZA DISH RED RETROSPOT}	0.00	0.82	343.93	23.00
{FUNK MONKEY} => {ART LIGHTS}	0.00	1.00	569.00	39.00
{ART LIGHTS} => {FUNK MONKEY}	0.00	1.00	569.00	39.00
{FRENCH BLUE METAL DOOR SIGN 9} => {FRENCH BLUE METAL DOOR SIGN 1}	0.00	0.76	135.76	44.00
{CLASSIC BICYCLE CLIPS} => {BICYCLE PUNCTURE REPAIR KIT}	0.00	0.79	214.91	54.00

Figure 30: Display Rules Page (Shiny App)



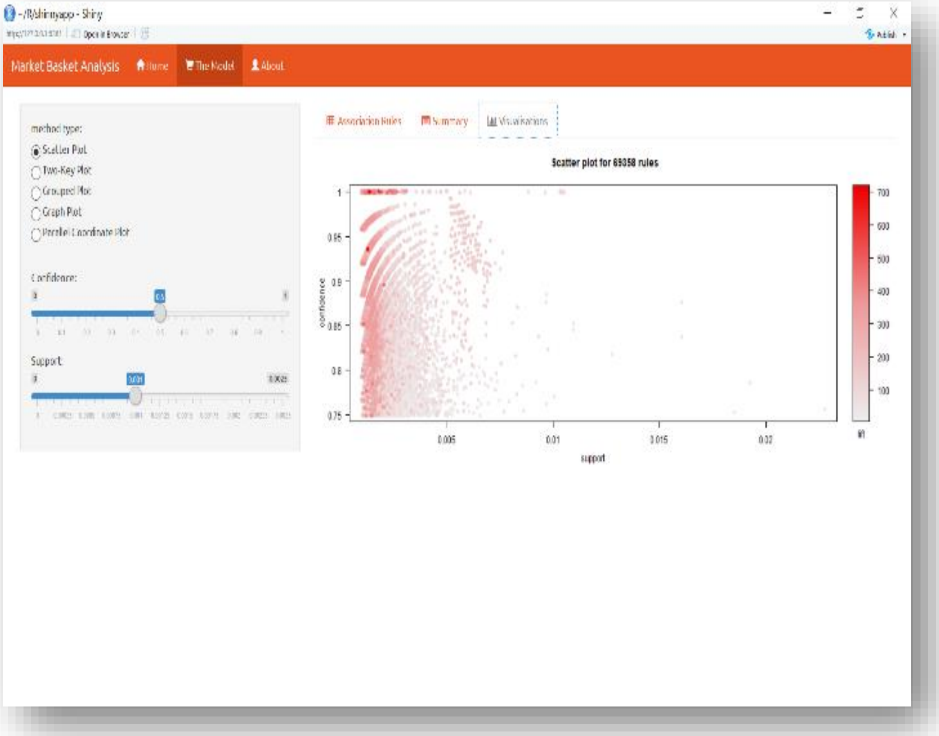
method type:
● Scatter Plot
○ Two-Key Plot
○ Grouped Plot
○ Graph Plot
○ Parallel Coordinate Plot

Confidence: 0.5

Support: 0.001

```
set of 69358 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10
  141  3305 11413 24723 19517  7210  2175   665   129

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   5.000   5.000   5.363   6.000  10.000

summary of quality measures:
    support            confidence          lift              count
 Min.   :0.001036   Min.   :0.7500   Min.   :  9.165   Min.   : 23.00
 1st Qu.:0.001082   1st Qu.:0.7931   1st Qu.: 20.902   1st Qu.: 24.00
 Median :0.001262   Median :0.8444   Median : 26.396   Median : 28.00
 Mean   :0.001418   Mean   :0.8536   Mean   : 53.975   Mean   : 31.47
 3rd Qu.:0.001532   3rd Qu.:0.9062   3rd Qu.: 44.928   3rd Qu.: 34.00
 Max.   :0.022757   Max.   :1.0000   Max.   :715.839   Max.   :505.00

mining info:
 data ntransactions support confidence
 tran         22191   0.001       0.75
```

Figure 31: Summary page (Shiny App)



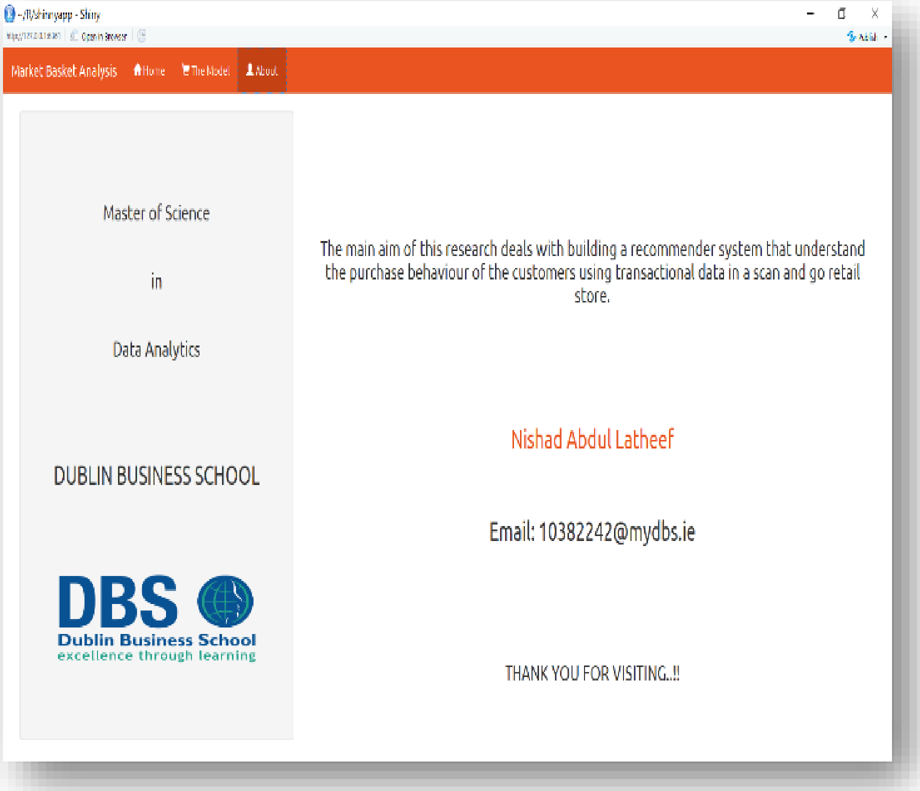
method type:

● Scatter Plot
○ Two-Key Plot
○ Grouped Plot
○ Graph Plot
○ Parallel Coordinate Plot

Confidence: 0.5

Support: 0.001

Figure 32: Visualisation Page (Shiny App)



Master of Science

in

Data Analytics

DUBLIN BUSINESS SCHOOL

DBS Dublin Business School excellence through learning

The main aim of this research deals with building a recommender system that understand the purchase behaviour of the customers using transactional data in a scan and go retail store.

Nishad Abdul Latheef

Email: 10382242@mydbs.ie

THANK YOU FOR VISITING..!!

Figure 33: About Page (Shiny App)

The above Figure shows the About page of the shiny application which provides a short description of the thesis along with details of the creator.

4.6 Online Monitoring Survey on Amazon Go store

As mentioned in the previous section, the conventional method of survey was avoided and a quicker, but effective way of online sentiment analysis on the particular product was carried out. This was made possible with a free trial version of a web application called Sentione. The analysis was executed by searching for the keyword "Amazon Go". This made sure that the survey was done only on those people who are familiar with the concept. All online posts that mentions the specified keyword, during the period from 5th October 2018 to 5th November 2018, was monitored using text mining techniques and the following results were observed:

4.6.1 Timeline of mentions

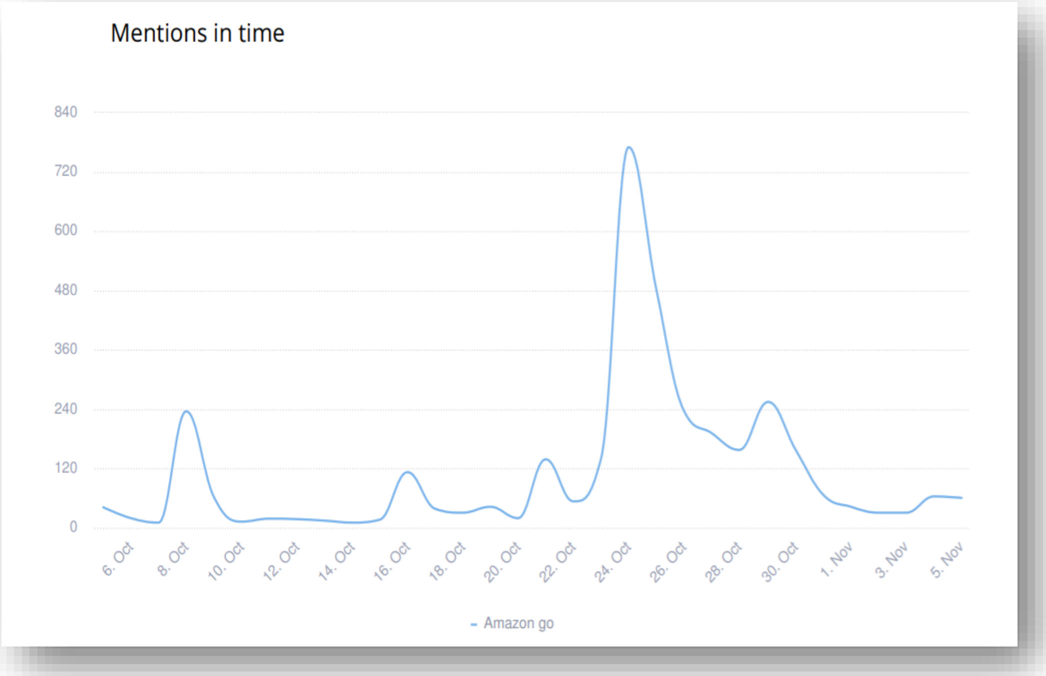


Figure 34: Timeline of Mentions

The timeline of mentions about the amazon go store was plotted as shown above. In the period between 5th October 2018 and 5th November 2018, the greatest number of mentions occurred on 24th October 2018 with about 780 mentions on that day.

4.6.2 Word Cloud

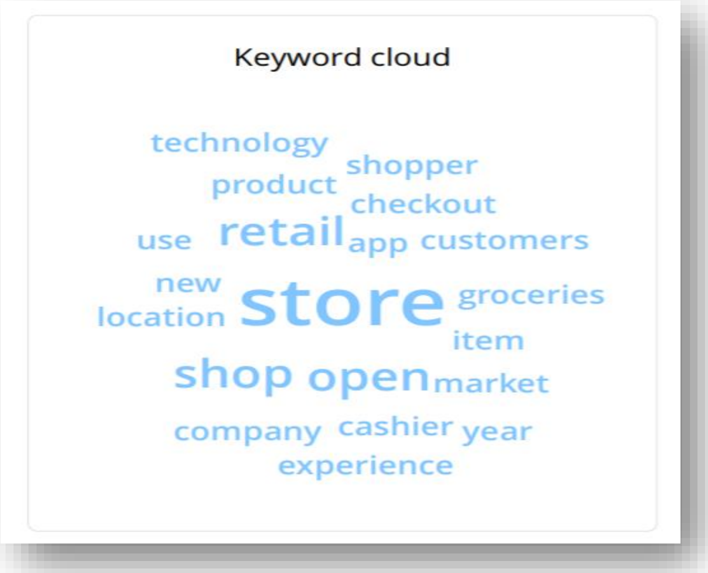


Figure 35: World Cloud

The above figure provides the word cloud of key words mentioned along with ‘Amazon Go’. The words are displayed in the cloud form in a way such a way that, the size of the word indicates the frequency of occurrence of the world. For example, it is clear from the figure that the world “store” has been mentioned the most along with “Amazon Go”.

4.6.3 Sentiment Overview

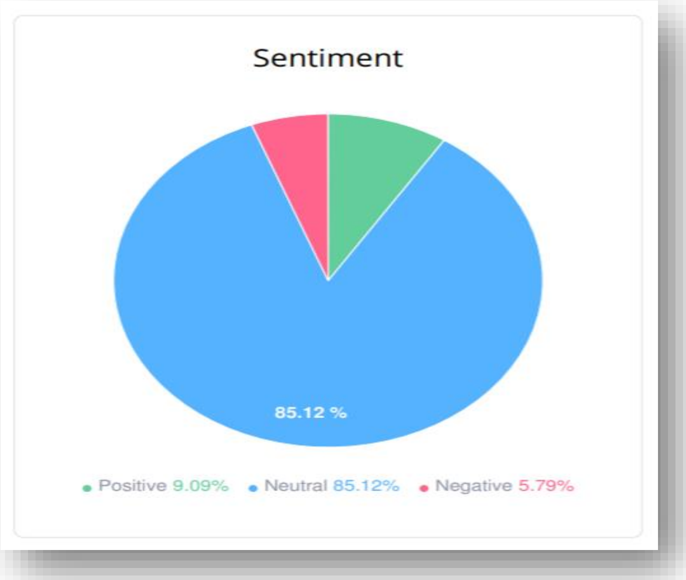


Figure 36: Sentiments Pie chart

The main aim of this survey was to determine how people reacted to the concept of a Scan and Go store. The above pie chart shows that most of the people, i.e. 85%, who mentioned about the product have neutral sentiment towards it. Further, more positive reviews (9.09%) are recorded than the negative reviews (5.79%). This resulted in creating the brand index of 0.61 indicating the brand is doing well as given below.

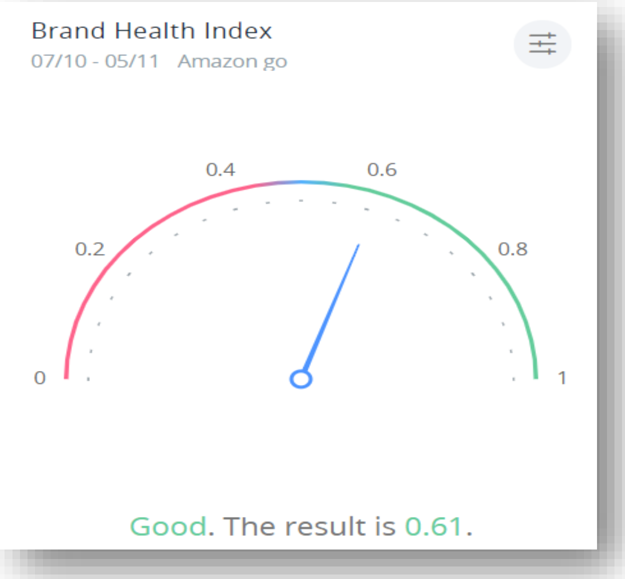


Figure 37: Brand Index Meter

4.6.4 Top Sources of Mentions

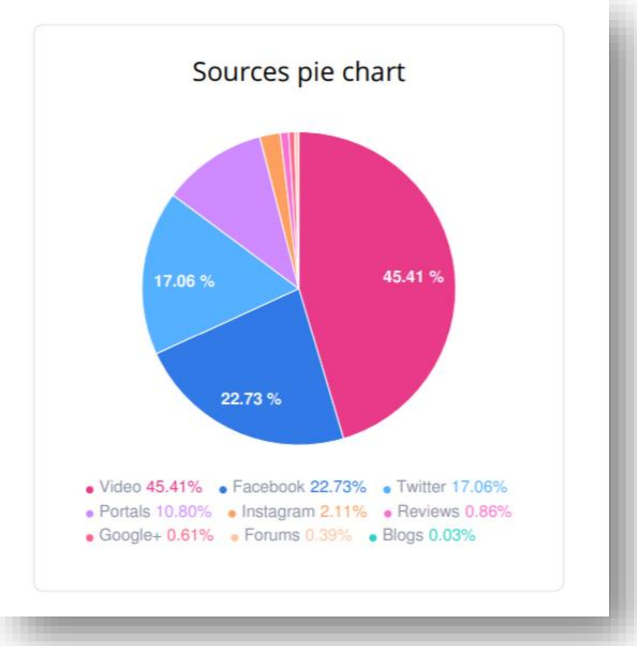


Figure 38: Sources of Mention Pie Chart

The above pie chart provides the distribution of various sources were the keyword was mentioned. From the figure, it can be seen that most of the mentions, i.e. 45.41%, came from video sources such as YouTube. The blogs was recorded as the source from which least mentions were made.

4.6.5 Gender Analysis

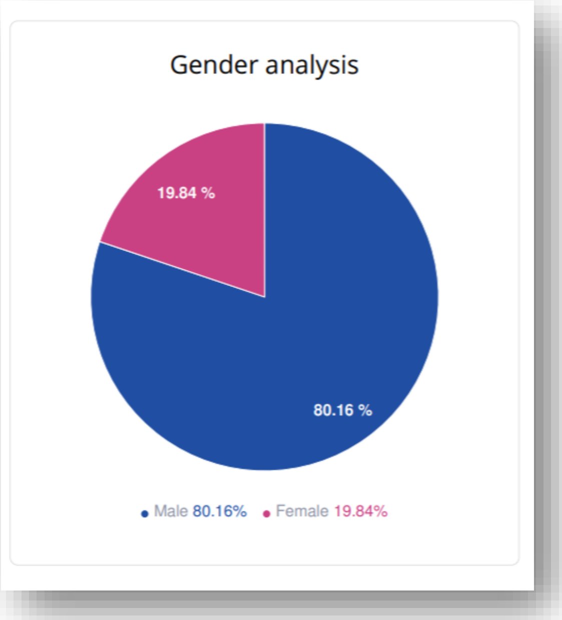


Figure 39: Gender Pie Chart

The gender ratio of people who mentioned about the product was depicted in a pie chart. Most of those people were males with 80.16%.

5. Discussion

The purpose of this section is to provide a critical assessment by discussing the results provided in the previous section. This will include reviewing the findings from building the model as well as from the online survey that was conducted; and relating these findings literature review by explaining the contribution of the work to the particular domain.

In order to provide a better understanding of the main research question, "How can a recommender system using Market Basket Analysis helps improve the situations for both the company as well as the customers?", we have to answer the sub-research questions. By answering these questions, the relationship between the results and literature review can be critically evaluated.

5.1 What is the best algorithm for building the recommender system?

Various algorithms were considered in the literature review section. Out of those, Apriori algorithm for Market Basket Analysis (MBA) was suggested by the majority of papers that was reviewed as a part of secondary research. This was mainly due to fact that, while dealing with large amount of data like transactional data of the customers in a retail store, the Apriori algorithm was proven to be having better performance by providing maximum accuracy in associating the similar products. Nevertheless, the usage of this algorithm does provide the following disadvantages:

- The scanning of the database occurs plenty of times. This means that, the database is scanned each time it executes. This in turn will reflect in deficiency of memory to store the data.
- The processing takes a long time due to fact that the input-output load is not adequate enough. Hence it displays low effectiveness.
- The complexity of time is extremely high.

However, the listed disadvantages can be tackled by executing the following suggested solution to boost the efficiency;

- The items should be clustered into greater conceptual groups. For Instance, white and brown baguette should be joined as "baguette".
- The total amount of scans of the whole database must be decreased.

5.2 How effective will the recommender system be on improving the revenue?

The role recommender system in e-commerce industry was explored in order to discover its impact. Even though the e-commerce bigshot Amazon is yet to reveal their estimates, it is approximated that 35% of people shopping on their website are due to their product recommender system with 60% sales conversion as shown in the figure below [12].

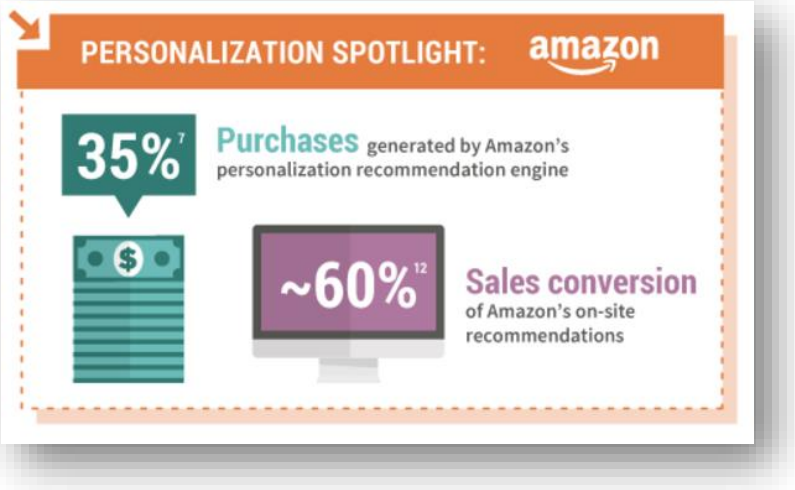

PERSONALIZATION SPOTLIGHT:
amazon
35% 7
Purchases generated by Amazon's personalization recommendation engine
~60% 12
Sales conversion of Amazon's on-site recommendations

Figure 40: According to data published by www.onespot.com

Another industry giant, Alibaba, utilised the Artificial Intelligence (AI) to provide a recommendation of items on the customers' personalized pages during one of the Shopping Festivals conducted in the year 2016. It was published by the company itself that this particular strategy produced 6.7 billion shopping pages that led to a 20 percent boost in the usual conversion rate [36].

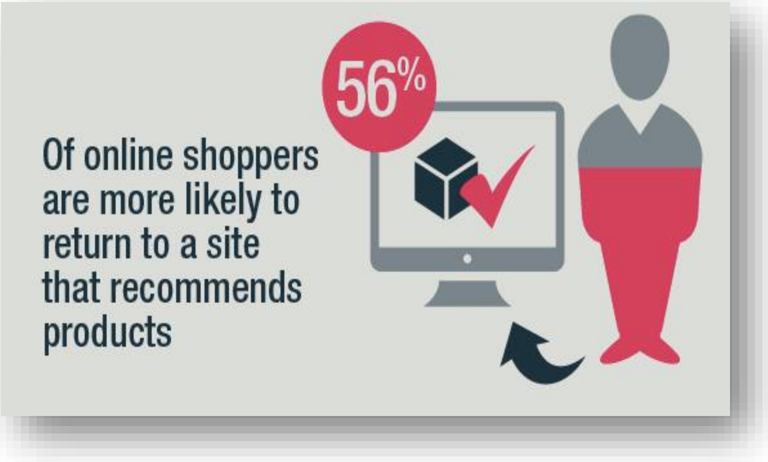

56%
Of online shoppers are more likely to return to a site that recommends products

Figure 41: According to data from invespcro.com

In a data analysis provided by invespcro.com [27], it is mentioned that about 59% of online customers are in favour of the personalization because it makes it easier to discover more relevant products when it is utilised as shown in the above figure.

These figures show that most of the customers finds the recommender system useful while shopping. This suggests that it is an efficient way to ensure revenue growth while satisfying the customers at the same time.

5.3 What are the possible observations that can be obtained from deploying the results of the analysis?

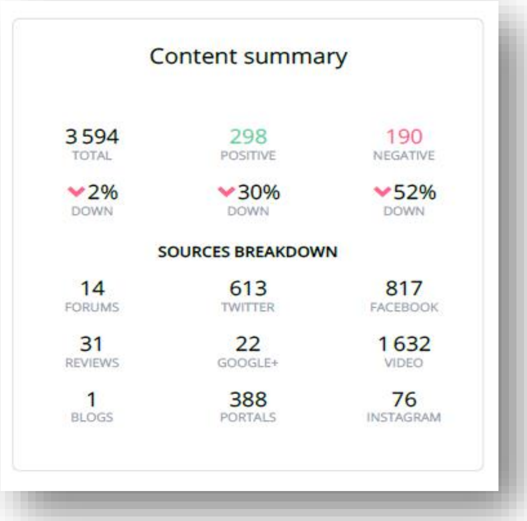
The mentioned dataset was put through the Apriori algorithm model is produce the association rules. These results were deployed to create visualisation of the data. The results of the deployment could be used for tackling various business objectives.

The Frequency of Purchase of all the products can be used to determine the most purchased product. The analysis can help to decide the product that should be stocked more and figure out new strategies to improve the sale for the least purchased product. The most expensive products chart could be compared with frequency of the purchase to form a ratio that can be utilised to maintain a balance between the unit price and sales. The map showing the top purchasing countries will help decide which region to target. The timeline of the purchased made can be utilised to determine the rush-hour that can be further used for better marketing opportunities.

Visualising the generated association rules from the Market Basket Analysis, using various plots such as scatter plot, two-key plot, grouped interactive plot, graph plot and parallel coordinate plot, makes it possible to understand the rules better. The relationship between the product could be easily understood even by a non-technical team member with the help of these visualisations.

5.4 How the people are responding to the changes in retail store brought about by Artificial Intelligence (AI)?

The ultimate success of any innovative technology is to satisfy the customers who are using the particular technology. An online monitoring survey was carried out that focused on individual who are aware of the Scan and Go stores. The summary of the survey is given in the figure below.



Content summary

3 594 TOTAL	298 POSITIVE	190 NEGATIVE
2% DOWN	30% DOWN	52% DOWN

SOURCES BREAKDOWN

14 FORUMS	613 TWITTER	817 FACEBOOK
31 REVIEWS	22 GOOGLE+	1 632 VIDEO
1 BLOGS	388 PORTALS	76 INSTAGRAM

Figure 42: Summary of Online Survey

The brand index of “Amazon Go” store is estimated to be .61, which means that majority of the people are in favour of the Scan and go technology. This suggests that more people all around the world finds this technology useful and prefers the Scan and Go shopping than the traditional shopping.

6. Conclusion

The section of conclusion is constructed by exploring the solutions to aim of the research. This is made possible by answering the sub-research questions that will lead to solution of the main research question. Moreover, a general conclusion of the entire research along with some possible future work is presented in this section.

In this dissertation, a research on collaborating the recommender system to the Scan and Go shopping was carried out. The main objective of this thesis deals with constructing a recommender system which can understand the purchase behaviour of the customers, by utilising the transactional data, in a scan and go retail store. The research question that deals with the objective is:

$$\text{“}\boldsymbol{How\ can\ a\ recommender\ system\ using\ Market\ Basket\ Analysis\ help\ improve\ the\ situations\ for\ both\ the\ company\ as\ well\ as\ the\ customers\ in\ a\ Scan\ and\ Go\ store}?\text{”}$$

The general conclusions that were drawn out from carrying out the whole research, that led to solving the sub-research questions, were;

- Considering the limitation of availability of data such a search history, amount of time spend on a particular product that make the recommender systems of amazon.com more efficient, the Apriori algorithm for association rule mining was observed as the best algorithm for carrying out the Market basket Analysis. This was due to fact that this particular algorithm showed better performance while dealing with large data. Even though the algorithm have some demerits, the solutions to tackle these problems are also suggested. Further, only the transactional data will be made available in real life stores. So, this makes the Apriori the better approach.
- The study on impact of recommender systems in various industries such as retail, music, news etc. shows that this particular technology has a crucial part in increasing the revenue of a company. The numbers shows that there is an estimated 10-25% boost in the revenue as a whole that was brought about by the use of the recommender systems. With efficient usage if the recommender system, the marketing strategies of any industry could be refurbished to their advantage.
- The best way to deploy the results was to use the data visualisation technique. The data visualization utilises an approach of building graphics based on statistics, graphics based on data and other simple plots in order to interactively present the information effectively in a precise manner. An effective visualization of data makes it possible for the users to explore

and rationale about the information. Thus, this approach helps in utilising the complex data to make it more understandable and easier to execute manner. Here, the deployment of data was done with the help of both R language and Tableau. It was found that, by using both these tools the data could be effectively presented. The tableau was used to provide the exploratory analysis. On the other hand, the generated rules were presented with the help of R.

- The summary of the online monitoring survey conducted have revealed that majority of the people, who are aware of the Scan and Go store technology, do have a positive review towards this particular technology. This could be mainly due to be fact that this technology makes it possible to have a more convenient and quicker shopping experience. Companies could use this positivity as a confidence to expand their retail store with the help of Artificial Intelligence.
- Overall, it can be concluded that introducing the recommender system into the Scan and Go stores can indeed satisfy both business as well as the customers at the same time. This should encourage the retail industry to implement the suggested model. The customers can use this Virtual Private Assistant (VPA) to make their shopping more convenient.

6.1 Future Work

The researched work could be expanded in various manners. Some of the suggestions for future work is given below;

- The User Interface (UI) of the recommender could be build using HTML and Bootstrap. The data of the product names and the generated rules can be stored in a Relational database management system (RDBMS) such as My SQL, Oracle, etc.
- The customer satisfaction could be increased by providing useful information about products they intend to purchase and making a virtual interaction. This could be made possible by expanding the Virtual Personal Assistant (VPA) with the help of Artificial Intelligence.
- The recommender system can be trained by Machine Learning techniques in order to predict the personalised shopping pattern of a particular customers. This could be made possible by customer segmentation.
- In considerably large stores, the location of a particular recommended item or product can be displayed in the application. This will further save the time for shopping; thus, making it more convenient.

Reflection

The research that was conducted was of utmost importance for me as an individual since it was one of first major platform to explore and showcase my passion for the Data Science field. So, I, Nishad Abdul Latheef, chose this particular topic such that the core of the research, i.e. building a recommender system, would represent Data Analytics. As a goal to pursue a career as Data Scientist in the field of business and marketing, I have encompassed the business benefits of a firm that would be achieved as an outcome of customer satisfaction.

Since the introduction of a fully-automated Scan and go store using Artificial Intelligence is relatively fresh, the research on collaboration of an existing technology like recommender system into those stores have not been found to my research knowledge. Thus, the presented research work is unique. For the same reason, the knowledge that I produced is a result of appropriate literature review, conducting the Market Basket Analysis and analysis of some primary online survey.

The whole experience of the dissertation was something worth noting down in form of a timeline. Some of key events that took place during the period of dissertations were:

- **First week:**
  The first and foremost thing to do was meeting my supervisor, Dr. Amir Sajad Esmaeily, for the first time. The meeting was scheduled on 9th October 2018 over e-mail. During the meeting, the initial proposal of the research was conveyed to him. It was a valuable session as he explained how to plan for the dissertation.
- **Second week:**
  The main objective of the research was formulated along with the main research questions and sub-research questions. This gave me a better understanding to implement the research. The second meeting with my supervisor was conducted on 15th October 2018. He advised the best way to start the research was to do some secondary research in form of Literature Review.
- **Third week:**
  This week was completely dedicating in carrying out my secondary research. Several pre-published paper on relevant topics were reviewed. The important points were noted down along with their respective reference materials. Further, I had my third meeting with my supervisor on 22nd October 2018. The progress in articulating the findings of my secondary research was discussed. The approach to carry out the core data analytics was suggested by him.

- **Fourth week:**

  The various designs of the methodology was considered, and CRISP-DM was chosen from it. This week mainly involved collecting the relevant data for carrying out Data Mining. I created the free-trail account of Sentione.com for online survey. The appropriate algorithm for executing the Market Basket Analysis was chosen.

- **Fifth week:**

  During this week, I drafted the R code for the entire data mining process. The CRISP-DM methodology was executed for the collected datasets. The online survey on sentione.com was also carried out. The supervisor was always contacted through e-mail for some advices.

- **Sixth week:**

  The results were produced in graphical form so as to explore and discuss the results. I had an uncertainty on drafting the discussion section. So, I contacted my supervisor through phone to get the advice. After this, I was clear on what needs to be done in the discussion section.

- **Seventh Meeting:**

  I did my final writing on the main six chapters of dissertation during this period. The progress in my research was explained to my supervisor on our fourth meeting conducted on 22nd November 2018. The feedback on my work gave me confidence in preparing the whole thesis on submission.

So as to present my experience appropriately, I would like to briefly provide the challenges faced along with strength and weaknesses. The time management skills, hunger to learn and eagerness to finish the work within the deadline helped me in overcoming the difficulties that I faced. However, this being my first master's thesis, a considerable amount of anxiety was encountered during the entire preparation of the thesis. The proper guidance from my supervisor, constant contact of my parents, some useful tips from people who had previous experience on doing such research and finally a whole lot self-motivation helped me in overcoming these anxieties. I used the rest of the time focusing on my graduate career by doing an internship in the field of Data Science. This was a good way to focus on myself and giving a much better chance to finish my thesis on the specified time.

Since the total time for the dissertation was only around two months, I spend between six to eight hours a day on my dissertation. The main motivation behind my dedication towards this was to present my thesis in a sophisticated and professional manner to others when needed that could highlight the knowledge that I gained during my master's course.

References

1. Adomavicius, G., and Zhang, J. (2012). Impact of data characteristics on recommender systems performance. *ACM Trans Manage Inform Syst* 3(1).
2. Agrawal, R., and Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. *Journal of Computer Science and Technology*,15.
3. Bartik, V. (2009). Association based Classification for Relational Data and its Use in Web Mining. *CIDM '09, IEEE Symposium on Computational Intelligence and Data Mining*, Pp. 252 – 258.
4. Bobadilla, J., Ortega, F., Hernando, A., and Gutie´rrez, A. (2013). Recommender systems survey. *Knowl-Based Syst 2013;* 46, Pp. 32- 109.
5. boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. Information, *Communication & Society*, 15(5), Pp. 662–679. http://doi.org/10.1080/1369118X.2012.678878
6. Burke, R. (2002). Hybrid recommender systems: survey and experiments. *User Model User-adapted Interact 2002;12(4).* Pp. 70- 331
7. Burke, R. (2007). Hybrid web recommender systems. *The Adaptive Web, Springer Berlin Heidelberg*, Pp. 377–408.
8. Chiu, K.S.Y., Luk, R.W.P., Chan, K.C.C., and Chung, K.F.L. (2002). Market-basket Analysis with Principal Component Analysis: An Exploration. *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3.
9. Cunningham, S.J. and Frank, E. (1999). Market Basket Analysis of Library Circulation Data. *International Conference on Neural Information Processing*, 2, Pp. 825-830, 1999.
10. Cuzzocrea, A., Song, I.-Y., & Davis, K. C. (2011). *Analytics over Large-scale Multidimensional Data: The Big Data Revolution! In Proceedings of the ACM 14th International Workshop on Data Warehousing and OLA, P*p. 101–104. New York, NY, USA: ACM. http://doi.org/10.1145/2064676.2064695
11. En.wikipedia.org. (2018). Amazon Go. Available at: https://en.wikipedia.org/wiki/Amazon_Go [Accessed 28 Oct. 2018].
12. Faggella, D. (2018). The ROI of recommendation engines for marketing - MarTech Today. [online] MarTech Today. Available at: https://martechtoday.com/roi-recommendation-engines-marketing-205787 [Accessed 23 Oct. 2018].
13. Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proc of the ACM SIGMOD International Conference*, 1, Pp. 1-12.
14. Han, J., Pei, J., Yin, Y., and R. Mao. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach.

15. Hao, Z., Wang, X., Yao, L., and Zhang, Y. (2009). Improved Classification based on Predictive Association Rules, *SMC 2009, IEEE International Conference on Systems, Man and Cybernetics*, Pp. 1165 – 1170.

16. Herlocker, J.L., Konstan, J.A., Terveen, L.G., and Riedl, T. (2004) Evaluating collaborative filtering recommender systems. *ACM Trans Inform Syst , 22(1),* Pp. 5–53.

17. Labrinidis, A., & Jagadish, H. V. (2012). Challenges and Opportunities with Big Data. *Proc. VLDB Endow.*, 5(12), Pp. 2032–2033. http://doi.org/10.14778/2367502.2367572

18. LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2011). Big Data, Analytics and the Path from Insights to Value. *MIT Sloan Management Review*, 52(2).

19. Liu, B., Hsu, W. and Ma, Y. (1998). Integrating Classification and Association Rule Mining. *Department of Information Systems and Computer Science.*

20. Madden, S. (2012). From Databases to Big Data. *IEEE Internet Computing*, 16(3), Pp. 4–6. http://doi.org/10.1109/MIC.2012.50

21. McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review.*

22. Niu, Q., Xia, S., and Zhang, L. (2009). Association Classification Based on Compactness of Rules, *WKDD 2009. Second International Workshop on Knowledge Discovery and Data Mining*, Pp. 245 – 247.

23. Prasad, J.P., and Mourya. M. (2013). A Study on Market basket Analysis Using Data Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, 3(6).

24. Rastogi, R. and Shim, K. (2002). Mining optimized association rules with categorical and numeric attributes. *IEEE Transactions on Knowledge and Data Engineering*, 14, No.1.

25. Resnick, P., and Varian, H.R. (1997). Recommender systems. *Communications of the ACM 40(3),* Pp. 56–58

26. Resnick, P., Iacovou, N., Suchak, M., Bergstrom P., and Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. *Proceedings ACM Conference on Computer-Supported Cooperative Work*, Pp. 175–186.

27. Saleh, K. (2018). Online Shopping Personalization – Statistics and Trends [Infographic]. [online] Invespcro.com. Available at: https://www.invespcro.com/blog/online-shopping-personalization/ [Accessed 27 Oct. 2018].

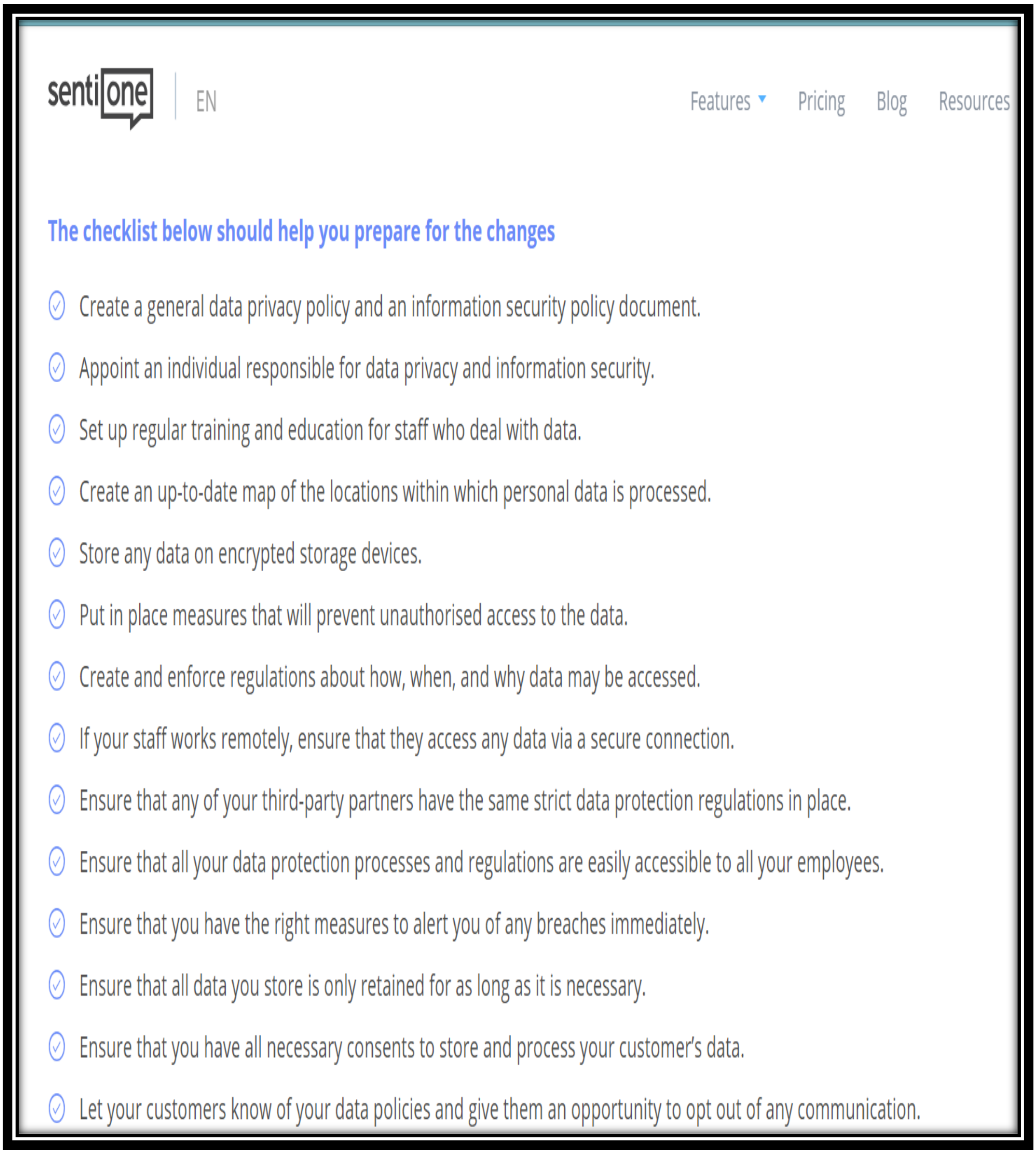
28. Schafer, J.B., Frankowski, D., Herlocker, J., and Sen, S. (2007) Collaborative filtering recommender systems. *Brusilovsky P, Kobsa A, Nejdl W, editors. The Adaptive Web, LNCS 4321. Berlin Heidelberg (Germany): Springer* Pp. 291–324. http://dx.doi.org/10.1007/978-3-540-72079-9_9.

29. Sha, Z., and Li, X. Mining local association patterns from spatial dataset. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 3, pp. 1455 – 1460.

30. Stern, D.H., Herbrich, R., Graepel, T. (2009). Matchbox: large scale online bayesian recommendations. *Proceedings of the 18th international conference on World Wide Web. ACM, New York, NY, USA*, Pp. 20-111.
31. Sumithra, R., and Paul, S. (2010). Using Distributed Apriori Association Rule and Classical Apriori Mining Algorithms for Grid Based Knowledge Discovery. *International Conference on Computing Communication and Networking Technologies (ICCCNT),* pp. 1 – 5.
32. Trnka, A. (2010). Market Basket Analysis with Data Mining Methods. *International Conference on Networking and Information Technology (ICNIT),* Pp. 446 - 450.
33. Videla-Cavieres, I. and Ríos, S. (2014). Extending market basket analysis with graph mining techniques: A real case. *Expert Systems with Applications,* 41(4), Pp.1928-1936.
34. Vo, B. and Le, B. (2009). Mining traditional association rules using frequent itemsets lattice. *International Conference on Computers & Industrial Engineering,* Pp. 1401 – 1406.
35. Wallis, J. (2017). The rise of Scan and Go technology and how it works. [online] Finextra Research. Available at: https://www.finextra.com/blogposting/14606/the-rise-of-scan-and-go-technology-and-how-it-works [Accessed 28 Oct. 2018].
36. Wang, E. (2018). How Alibaba uses artificial intelligence to change the way we shop - Inside Retail Asia. [online] Inside Retail Asia. Available at: https://insideretail.asia/2017/06/07/how-alibaba-uses-artificial-intelligence-to-change-the-way-we-shop/ [Accessed 21 Oct. 2018].
37. Wang, P., Shi, L., Bai, J., and Zhao, Y. (2009). Mining Association Rules Based on Apriori Algorithm and Application. *International Forum on Computer Science-Technology and Applications,* 1, Pp. 141-143.
38. Wang, Y.J., Xin, Q., and Coenen, F. (2007). A Novel Rule Weighting Approach in Classification Association Rule Mining. *ICDM Workshops 2007, Seventh IEEE International Conference on Data Mining Workshops,* Pp. 271 – 276.
39. Wen-xiu, X., Heng-nian Q., and Mei-li, H. (2010). Market Basket Analysis Based on Text Segmentation and Association Rule Mining. *First International Conference on Networking and Distributed Computing (ICNDC),* Pp. 309 – 313.
40. Woo, J. and Xu, Y. (2013). Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 3(6), Pp.445-452.
41. Xu, Y., Zhou. S., and Gong, J. (2005). Mining Association Rules with New Measure Criteria. *International Conference on Machine Learning and Cybernetics,* 4, Pp. 2257- 2260.
42. Xuping, W., Zijian. N., and Haiyan, C., (2007). Research on Association Rules Mining Based-On Ontology in Ecommerce. *International Conference on Wireless Communications, Networking and Mobile Computing,* Pp. 3549-3552.
43. Yanthy, W., Sekiya, T., and Yamaguchi, K. (2009) Mining Interesting Rules by association and Classification Algorithms, *FCST 09.*

44. Yanthy, W., Sekiya, T., and Yamaguchi, K. (2009). Mining Interesting Rules by Association and Classification Algorithms. *FCST '09. Fourth International Conference on Frontier of Computer Science and Technology*, Pp. 177 –182.
45. Yin, X., and Han, J. (2007). CPAR: Classification based on Predictive Association Rules", *Proceedings of the Third SIAM International Conference on Data Mining*, Pp 331-335.

Appendix A: Ethics Checklist



The checklist below should help you prepare for the changes

- ✓ Create a general data privacy policy and an information security policy document.
- ✓ Appoint an individual responsible for data privacy and information security.
- ✓ Set up regular training and education for staff who deal with data.
- ✓ Create an up-to-date map of the locations within which personal data is processed.
- ✓ Store any data on encrypted storage devices.
- ✓ Put in place measures that will prevent unauthorised access to the data.
- ✓ Create and enforce regulations about how, when, and why data may be accessed.
- ✓ If your staff works remotely, ensure that they access any data via a secure connection.
- ✓ Ensure that any of your third-party partners have the same strict data protection regulations in place.
- ✓ Ensure that all your data protection processes and regulations are easily accessible to all your employees.
- ✓ Ensure that you have the right measures to alert you of any breaches immediately.
- ✓ Ensure that all data you store is only retained for as long as it is necessary.
- ✓ Ensure that you have all necessary consents to store and process your customer's data.
- ✓ Let your customers know of your data policies and give them an opportunity to opt out of any communication.

Figure 43: Ethics Checklist of sentione.com

Appendix B: R-Code for Market Basket Analysis

```
#################################################################

#**************************************************************#

#----------------------------------------------------------------------------------------------------#

#########-------- Dissertation Artefact: Nishad Abdul Latheef-------- ##########

##----------------- Market Basket Analysis using Apriori Algorithm ------------------##

#------------------------------------------------------------------------------------------------------#

#***************************************************************#

##################################################################


#Command used to extract and add readxml

#install.packages("readxml")

library(readxl)


#Command used to extract and add tidyverse

#install.packages("tidyverse")

library(tidyverse)


#Command used to extract and add knitr

#install.packages("knitr")

library(knitr)


#Command used to add ggplot2
```

```
library(ggplot2)


#Command used to extract and add lubridate

#install.packages("lubridate")

library(lubridate)


#Command used to extract and add plyr

#install.packages("plyr")

library(plyr)


#Command used add dplyr

library(dplyr)


#Command used to extract and add package arules

#install.packages("arules")

library(arules)


#Command used to extract and add arulesViz

#install.packages("arulesViz")

library(arulesViz)


################################################################

#---------------------------------------------------------------#

# Step 1: Data Preparation

#---------------------------------------------------------------#

################################################################
```

```
#Command used for reading the given dataset in excel format into
#R data frame
retailData <- read_excel('Online_Shopping.xlsx')
#The data might have missing values

#The fuction complete.cases(data) will provide a logical vector
#representing that rows which have no missing values.
retailData <- retailData[complete.cases(retailData), ]

products <- unique(retailData[3])
write.csv(products,"products.csv",
      quote = FALSE, row.names = FALSE)

#The in-built mutate function is a part of dplyr package. It is
#utilised to change the structure of columns in a dataframe.Here,
#Description column is transformed into factor column.
retailData %>% mutate(Description = as.factor(Description))
retailData %>% mutate(Country = as.factor(Country))

#Command used to transform character data to date format.
retailData$DateOfPurchase <- as.Date(retailData$InvoiceDate)
#Command used to obtain the time from the column InvoiceDate
#and keep it in a fresh variable
TimeOfPurchase <- format(retailData$InvoiceDate,"%H:%M:%S")
#Command used to change the format of InvoiceNo into numeric
InvoiceNo <- as.numeric(as.character(retailData$InvoiceNo))
```

```
#Command used to attach newly formed columns TimeOfPurchase and

#InvoiceNo into the given data frame retailData

cbind(retailData,TimeOfPurchase)

cbind(retailData,InvoiceNo)


#Command used to take a peek at the data

glimpse(retailData)

# 406,829 Observastions and 9 variables


#Command used to add plyr

library(plyr)

#Command used to build the transaction data

transaction_Details<- ddply(retailData,

           c("InvoiceNo","DateOfPurchase"),

           function(df1)paste(df1$Description,

                   collapse = ","))

#Checking the result

transaction_Details


#The unwanted columns are removed

#The column InvoiceNo is removed from transaction_Details

transaction_Details$InvoiceNo <- NULL

#The column DateOfPurchase is removed transaction_Details

transaction_Details$DateOfPurchase <- NULL


#The column is renamed to item_list

colnames(transaction_Details) <- c("item_list")
```

```
#The Dataframe transaction_Details is viewed
transaction_Details

#The output of the transaction is stored as in a csv format
write.csv(transaction_Details,"transactions_output.csv",
          quote = FALSE, row.names = FALSE)
#The data is prepared for modelling

#################################################################
#------------------------------------------------------------#
# Step 2: Rule Generation
#------------------------------------------------------------#
#################################################################

#The csv file is loaded into a variable Tran
Tran <- read.transactions('transactions_output.csv',
                          format = 'basket', sep=',')

#The dataframe is viewed
Tran

#The summary of Tran is viewed
summary(Tran)

#The association rules are generated using apriori function by
# setting Minimum Support as 0.001 and  confidence as 0.75.
```

```
Generated_Rules <- apriori(Tran, parameter = list(supp=0.001,
                                        conf=0.75,
                                        maxlen=10))



#The summary of generated rules is viewed
summary(Generated_Rules)

#################################################################
#------------------------------------------------------------#
# Step 3: Evaluation
#------------------------------------------------------------#
#################################################################

#The top 50 rules are inspected
rules1 <- inspect(Generated_Rules[1:50])

#The rules are stored in a file
write.csv(rules1,"Rules_1.csv",
        quote = FALSE, row.names = TRUE)



#Making the rules stronger by limiting the number of items and
# increasing the minimum support
Stronger.Generated_Rules <- apriori(Tran,
                           parameter = list(supp=0.001,
                                        conf=0.8,
                            maxlen=3))
#Stronger rules are generated


#The top 50 rules are inspected

rules_Strong <- inspect(Stronger.Generated_Rules[1:50])


#The rules are stored in a file

write.csv(rules_Strong,"Strong_Rules.csv",

     quote = FALSE, row.names = TRUE)


#Finding the redundant rule

subset.rules <- which(colSums(is.subset(Generated_Rules,

                      Generated_Rules)) > 1)


#Removing unwanted rules

new_Generated_Rules. <- Generated_Rules[-subset.rules]


#Discovering Rules related to a particular item

Focused.Generated_Rules <- apriori(Tran,

                   parameter = list(supp=0.001,

                            conf=0.8),

                   appearance=list(default="lhs"

                           ,rhs="COFFEE"))

#Inspecting the focused rules

inspect(Focused.Generated_Rules)


#Discovering Rules related to a particular item
```

```
Focused.Generated_Rules <- apriori(Tran,

                                parameter = list(supp=0.001,

                                                conf=0.8),

                                appearance=list(lhs="BLACK TEA"

                                               ,default="rhs"))

#Inspecting the focused rules

inspect(Focused.Generated_Rules)




##################################################################

#------------------------------------------------------------#

# Step 4: Deployment

#------------------------------------------------------------#

##################################################################


# The rules that have confidence greater than 50% is filtered out

Fil.Rules<-Generated_Rules[quality(Generated_Rules)$confidence>0.5]


inspect(Fil.Rules[1:10])

#the filtered rules are plotted

plot(Fil.Rules)


#Plotting the two-key method

plot(Fil.Rules,method="two-key plot")


#Plotting grouped interactive plot

plot(Fil.Rules, method = "grouped", interactive = TRUE)
```

```
#Filtering top 10 rules

Fil.rules_2 <- head(Generated_Rules, n = 10, by = "lift")


#Plotting Graph based visualisation

plot(Fil.rules_2, method = "graph", engine = "htmlwidget")


#Saving the graph file

saveAsGraph(head(Fil.Rules, n = 1000, by = "lift"),

          file = "Graph_of_rules.graphml")


#The top 30 rules that have the highest lift is selected

Fil.rules_3 <- head(Fil.Rules, n=30, by="lift")


#plotting Parallel Coordinates

plot(Fil.rules_3, method="paracoord")



##################################################################

#---------------------------The End--------------------------#

##################################################################
```

Appendix C: R-Code for The Shiny Application

```
################################################################

#**************************************************************#

#-----------------------------------------------------------#

######### Dissertation Artefact: Nishad Abdul Latheef ##########

##------------------- Shiny Application --------------------##

#-----------------------------------------------------------#

#**************************************************************#

################################################################


#Command used to extract and add readxml

#install.packages("readxml")

library(readxl)


#Command used to extract and add tidyverse

#install.packages("tidyverse")

library(tidyverse)


#Command used to extract and add knitr

#install.packages("knitr")

library(knitr)


#Command used to add ggplot2
```

```
library(ggplot2)

#Command used to extract and add lubridate
#install.packages("lubridate")
library(lubridate)

#Command used to extract and add plyr
#install.packages("plyr")
library(plyr)

#Command used add dplyr
library(dplyr)

#Command used to extract and add package arules
#install.packages("arules")
library(arules)

#Command used to extract and add arulesViz
#install.packages("arulesViz")
library(arulesViz)

#Command used to extract and add shiny
#install.packages("shiny")
library(shiny)

#Command used to extract and add shinythemes
#install.packages("shinythemes")
```

```
library(shinythemes)


#The csv file is loaded into a variable Tran
Tran <- read.transactions('transactions_output.csv',
                          format = 'basket', sep=',')


#Defining the UI of the app
ui = tagList(


  navbarPage(
    theme = shinytheme("united"),


    "Market Basket Analysis",


    #defining Home page
    tabPanel("Home",
             icon = icon("glyphicon glyphicon-home",
                         lib = "glyphicon"),
             sidebarPanel(
               br(),
               br(),
               br(),
               strong(h3("Master of Science",
                         align = "center")),
               h3("in",align = "center"),
               h3("Data Analytics",align = "center"),
               br(),
               br(),
               br(),
               br(),
               br(),
               h2("DUBLIN BUSINESS SCHOOL",
                  align = "center"),
               h4("Dublin, Ireland",align = "center"),
               br(),
               br(),
               br(),
               HTML('<center><img src="DBS.png"
                    height = "150" width="400" ></center>'),
               br(),
               br()

             ),
             mainPanel(


               br(),
               br(),
               br(),
               br(),
               h2("A Shiny Application", align = "center"),
               br(),
               h4("as a part of research on", align = "center"),
               br(),
               br(),
                       strong( h1("Developing A Recommender System
                             for Customers Using
                             Apriori Algorithm in Market Basket Analysis"
                             ,align = "center")),
                       br(),

                       br(),
                       HTML('<center><img src="Rs.png" height = "150"
                          width="200" ></center>'),

                       br(),
                       br(),

                       h3("Nishad Abdul Latheef",align = "right"),
                       h3("10382242",align = "right")



                       )

                 ),
                 #Defining the model page
                 tabPanel("The Model",
                      icon = icon("glyphicon glyphicon-shopping-cart"
                            ,lib = "glyphicon"),
                      sidebarPanel(
                       #input for selecting the method of graphs
                       radioButtons("meth", "Method Type:",
                     c("Scatter Plot" = "SPlot",
                       "Two-Key Plot" = "TKPlot",
                       "Grouped Plot" = "GPPlot",
                       "Graph Plot" = "GPlot",
                       "Parallel Coordinate Plot"= "PCPlot")),

              br(),

              #Slider inputs for confidence and support
              sliderInput(inputId = "conf",
                     label = "Confidence:",
                     min = 0,
                     max = 1,
                     value = 0.5),
              sliderInput(inputId = "supp",

                     label = "Support:",
                     min = 0,
                     max = 0.0025,
                     value = 0.001, step = 0.00001)

             ),
             mainPanel(

              tabsetPanel(type = "tabs",
                     tabPanel("Association Rules",
                          tableOutput("table"),
                           icon = icon("table")),
                       tabPanel("Summary",
                           verbatimTextOutput("summary"),
                           icon = icon("list-alt")),
                       tabPanel("Visualisations",
                           plotOutput("plot"),
                           icon = icon("bar-chart-o"))
                  )
                 )
             ),

             #defining the About page
             tabPanel("About",
                 icon = icon("glyphicon glyphicon-user"
                      ,lib = "glyphicon"),
                 sidebarPanel(
                  br(),
                  br(),
                  br(),
                  strong(h3("Master of Science"
                      ,align = "center")),
                  h3("in",align = "center"),
                  h3("Data Analytics",align = "center"),
                  br(),
                  br(),
                  br(),
                  br(),
               br(),

               h2("DUBLIN BUSINESS SCHOOL"

                  ,align = "center"),

               h4("Dublin, Ireland",align = "center"),

               br(),

               br(),

               br(),

               HTML('<center><img src="DBS.png"

                    height = "150" width="400" ></center>'),

               br(),

               br()




             ),

             mainPanel(

               br(),

               br(),

               br(),

               br(),

               br(),

               br(),

               h3("The main aim of this research deals

                    with building a recommender system

                  that understand the purchase behaviour

                  of the customers using transactional

                  data in a scan and go retail store."

                  , align = "center"),
                     br(),

                     br(),

                     br(),

                     br(),

                     br(),

                     h2(tags$a(

                      href="https://www.linkedin.com/in/nishad-abdul-latheef-8136ab142/",

                         "Nishad Abdul Latheef"), align = "center"),


                     br(),

                     br(),

                     h2("Email: 10382242@mydbs.ie",align = "center"),

                     br(),

                     br(),

                     br(),

                     br(),

                     br(),

                     h3("THANK YOU FOR VISITING...!!",align = "center")

                     )

                    )

               )

             )


             #Defining the server

             server<- function(input, output) {


               m <- reactive({
                     if(input$meth == "SPlot")
                     {meth <- "scatterplot"}

                     else if(input$meth == "TKPlot")
                     { meth <- "two-key plot"}

                     else if(input$meth == "GPPlot")
                     { meth <- "grouped"}

                     else if(input$meth == "GPlot")
                     { meth <- "graph"}

                     else if(input$meth == "PCPlot")
                     { meth <- "paracoord"}

                     else { meth <- "scatterplot"}

                     return(meth)
                   })

             #Generate top rules
             output$table <- renderTable({

                     c <- input$conf
                     s <- input$supp

                     #The association rules are generated
                     # using apriori function
                     Generated_Rules <- apriori(Tran,
                                           parameter = list(supp=s,
                                                    conf=c,
                                                   maxlen=10))

                     gr <- as( Generated_Rules, "data.frame")
                     head(gr,40)
             })



             # Generate the summary of the generated rules
             output$summary <- renderPrint({


                      c <- input$conf
                      s <- input$supp

                      #The association rules are generated
                      #using apriori function
                      Generated_Rules <- apriori(Tran,
                                            parameter = list(supp=s,
                                                     conf=c,
                                                    maxlen=10))

                      summary( Generated_Rules)
             })
```

```
#Generate visualisations
output$plot <- renderPlot({

    c <- input$conf
    s <- input$supp

    #The association rules are generated
    #using apriori function
    Generated_Rules <- apriori(Tran,
           parameter = list(supp=s,
               conf=c,
               maxlen=10))
    mf <- m()

    if(mf == "graph")
    { Generated_Rules <- head(Generated_Rules,
           n=10, by="lift")}

    else  if(mf == "paracoord")
    { Generated_Rules  <- head(Generated_Rules,
           n=30, by="lift")}

    plot(Generated_Rules, method= m())
```

```
 })


}


shinyApp(ui, server)



################################################################
#----------------------------The End---------------------------#
################################################################
```