Clustering Based Approach to Enhance Association Rule Mining

MSc Research Project
MSc Data Analytics

Samruddhi Shailesh Kanhere
Student ID: x18190634

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes, Dr. Pramod Pathak

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Samruddhi Shailesh Kanhere
Student ID:	x18190634
Programme:	MSc Data Analytics
Year:	2019-2020
Module:	MSc Research Project
Supervisor:	Dr. Paul Stynes, Dr. Pramod Pathak
Submission Due Date:	28/09/2020
Project Title:	Clustering Based Approach to Enhance Association Rule Mining
Word Count:	6738
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Samruddhi Shailesh Kanhere
Date:	28th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Clustering Based Approach to Enhance Association Rule Mining

Samruddhi Shailesh Kanhere
x18190634

Abstract

Association rule mining algorithms such as Apriori and FPGrowth are extensively being used in the retail industry to uncover consumer buying patterns. However, the scalability of these algorithms to deal with the voraciously increasing data is the major challenge. This research presents a novel Clustering based approach by reducing the dataset size as a solution. The products are clustered based on their frequency and price. Another important aspect of this study is to find interesting rules by performing differential market basket analysis to identify association rules which are likely ignored in the trivial approach. When using a cluster-based approach, it is observed that the same set of rules can be generated by using only 7% of the total 16210 items, which in turn directly contributes to reducing the processing overheads and thus reducing the computation time. Furthermore, results obtained from differential market basket analysis have highlighted a few interesting rules which were missing from the original set of rules. A clustering-based approach used in this study not only consists of frequent items but also considers their contribution to the overall revenue generation by considering its price thus adding value to the state of the art. In addition to this, the least contributing product exclusion rate is also improved from 45% to 93%. These results evidently suggest that the computation cost can be significantly reduced, and more accurate rules can be generated by applying differential market basket analysis. What effects does the product exclusion cause on the business is the question that must be addressed in further researches?

1 Introduction

Retail is one of the vital industries of any economy. According to the research published by retail economics from 2019's nation-wide survey, retail industry valuation in the United Kingdom is about £394 Billion. It employs an estimate of 2.9 million people all over the UK. There are around 306,655 retail outlets in the UK according to this survey, which contributes to 5% of the total GDP of the UK[1]. Being a customer-centric and competitive industry, it highly focuses on serving customers with high-quality products at cheaper rates. Thus, to maximize the profits, one of the methods being used by the retail industry is to try and sell as many items together as possible. With the help of technology, the retail industry is keen on finding interesting buying patterns to serve their customers wisely which in turn will attract more customers and thus improve the business. Increasing

[1] https://www.retaileconomics.co.uk/library-retail-stats-and-facts

the use of technology in terms of payments, cloud services, data warehouses, etc. the retail industry is changing and adapting at a very high pace. During this process, it has managed to gather and create a lot of data based on customer purchases and this data is increasing at an astronomical rate (Gijsbrechts et al.; 2018).

Many analysts have used the likes of Market Basket Analysis to evaluate this data and figure out the likeliness of items to be purchased together. Businesses also use these data-mining techniques to devise cross-selling strategies. Furthermore, this data provides guidance to shopkeepers and managers to place their products in shops and help them in formulating a discount to attract customers in buying more items. These techniques work better when the amount of data is limited. But as the number of transactions ascends the performance of these techniques' plummets. Performance is highly impacted due to an increase in computational cost, which means greater execution times. This creates a scalability issue (Hossain et al.; 2019).

Another major issue with these techniques is that more amount of data it churns higher the complexity of algorithms. Subsequently a huge number of association rules are generated. All the generated rules may not be relevant and useful. Also, explaining the relevance of these rules becomes difficult. To deal with this problem, it is crucial to filter out such associations and only include rules which identify interesting relationships between the products. So, to tackle these important shortcomings of Market Basket Analysis, this paper will focus on scaling down the dataset without losing its attributes and insights (Chun-Sheng and Yan; 2014).

The aim of this research is to investigate to what extent a combination of product clustering and differential market basket analysis can optimize the association rule mining in terms of execution time and the interestingness of association rules. To address this research question, the following research objectives are derived.

1. Investigate the state of the art in the domain of Market Basket Analysis.

2. Design a model that performs clustering and differential market basket analysis.

3. Implement the proposed model and capture the results.

4. Evaluate the performance of the proposed model for optimizing the rule generation process.

This research is carried out by using the transactional data collected from the retail store chain in the UK. The basket data is present in the semi-structured JSON format which makes the pre-processing challenging. The contribution of this research is to address two major issues. One of them is to tackle the scalability issue by using a clustering technique to reduce the items of datasets with top revenue-generating products. Thus, the number of input transactions is reduced thereby reducing the execution times. Another problem addressed by this research is the high volume of trivial association rules. The data will be sampled based on the time of the purchase and the association rules will be derived based on these samples. The results will be compared, which will help in finding interesting rules. This technique is known as differential market basket analysis. This analysis will also contribute to the research in the retail and marketing domain.

This paper is further divided into following sections. Section 2 of this report will focus on the previous work carried out in this domain. Section 3 discusses the methodology employed for this research. It is followed by Section 4 that describes the details about the design and implementation of the research. Section 5 then describes the experiments conducted, the results obtained, and the discussion about them. Following this, Section 6 concludes the research work along with the identification of future work.

2 Related Work

In this section, a critical analysis of studies conducted in the chosen domain is put forth. This study will help in understanding the previous researches, their findings, and their limitations.

2.1 Market Basket Analysis (MBA)

Market Basket Analysis (MBA) is a key technique that has been widely used in the retail industry to calculate the statistical affinity between different items/products. Using this technique, retailers can decide on marketing strategies, shelf arrangement, cross-selling strategies, etc. MBA analyses the purchasing patterns by mining transactions or the shopping baskets. The transaction is a collection of items bought in a single purchase. Several studies mentioned that the use of an MBA allows the retailers to grab every opportunity to make their business work smarter (Zaki and Gouda; 2003; Sorensen et al.; 2017; Valle et al.; 2018).

Kutuzova and Melnik (2018) used several data sources for improving the grocery store recommendation system. The techniques used include collaborative filtering, clustering, and association rules mining. They concluded that the association rule mining gave satisfactory results. The research (Valle et al.; 2019) proposes a probabilistic model for analyzing behaviour of customers. This research was conducted by using supermarket data. In one of the interesting studies, the association between different sports items is identified by using customer purchase data of a Sports Company. The product placement was then changed based on the association rules generated. The sales of a Sports company were significantly increased (Abbas et al.; 2013). The limitation of this research is that the dataset used for this research was very small and thus, results need to be validated with a larger dataset. Thus, it can be seen from the above studies that the Association Rule Mining (ARM) is an important branch of MBA. Moreover, ARM is found to be very useful in practice. The next section discusses the literature related to the ARM.

2.2 Association Rule Mining (ARM)

ARM is the process of identifying which product is purchased with which other product or products. “What” goes with “what” is determined by using ARM. The customer transactions database is used to find these associations (Kurniawan et al.; 2017).

The Apriori algorithm is extensively used for ARM which was first introduced in the year 1993 (Agrawal et al.; 1993). It is the most famous algorithm for ARM because of its simplicity. There are two main steps involved in the Apriori algorithm. The first is to find frequent itemsets. Frequent itemsets are the ones with support greater than threshold support. Second is generating association rules by filtering the frequent itemsets based on the threshold confidence values (Huang et al.; 2020).

Association rules are represented as $X \rightarrow Y$ where X and Y are the individual items. This means that there is some association between items X and Y. In other words, if item X appears in the transaction then there is a high chance that Y will also occur in the transaction. In this rule, X is known as antecedent and Y as consequent (Kotu and Deshpande; 2019).

Support of an item is the number of transactions in which the item occurs out of total transactions. In other words, the support is a relative frequency. The frequent itemset is the one whose support is greater than the minimum support (Kotu and Deshpande; 2019). The support is generally used to filter the frequent itemsets.

The confidence of a rule is defined as a relative probability of occurrence of consequent out of total transactions in which antecedents are present. The confidence of a rule is used to filter the association rules (Kotu and Deshpande; 2019).

Another famous algorithm used for ARM is the Frequent Pattern (FP) Growth algorithm. Several studies compared the two ARM algorithms i.e. Apriori algorithm and FP Growth algorithm. According to these studies, the FP Growth algorithm requires less execution time than that of the Apriori algorithm for large databases (Hossain et al.; 2019; Davanbu and Venkatachari; 2016). However, for smaller databases, the Apriori algorithm is found to be faster. Another ARM algorithm that is similar to the Apriori algorithm is the Eclat algorithm. Researchers compared the performance of these two algorithms when applied to the data of a supermarket in Turkey. Both algorithms yield the same output, though the underlying data structure is different. The Eclat algorithm works really well with smaller datasets according to the researchers (Huseyinov and Aytac; 2017).

Despite the popularity of ARM algorithms, there are some challenges associated with them. The main challenge associated with ARM is the scalability issue. The huge amount of transactional data is generated everyday and processing this huge data creates a scalability issue of ARM algorithms (Gassama et al.; 2017). The next section highlights some approaches used to deal with this issue of scalability.

2.3 Challenges in Association Rule Mining

The execution time of the Apriori algorithm increases exponentially with an increasing number of transactions. Multiple studies used a parallelization approach using a map-reduce framework to deal with this issue (Liu and Lou; 2014). Another approach that is used to solve this issue is the reduction of the database size. One of the researches attempted to reduce the dataset by using the top-selling products to reduce the database. The results showed that 55% of product reduction yielded the same results as that of the whole database (Hossain et al.; 2019). The limitation here is only the frequency of product occurrence is considered for product reduction. The other parameters such as price or quantity of the product might result in a better approach for product reduction (Moodley et al.; 2020). In another research, a parallel version of FP Growth is implemented using Apache Spark which showed improved performance. The authors mentioned that the parallel implementation of FP Growth is complex (Gassama et al.; 2017).

Another challenge that comes with the increasing data is the quality of rules that are generated by processing large databases. The trivial rules are generated as the dataset size increases. Hence, to generate the interesting rules researchers used clustering techniques. Thus, the ARM algorithms are applied to these clusters, and rules are generated for each cluster. The research showed that the interesting rules are generated after clustering (Fong et al.; 2017). Differential market basket analysis is another approach that can be employed to find the interesting association rules.

In conclusion, Apriori, FP Growth, and Eclat are the most famous ARM algorithms.

It is observed from the above studies that there are two major challenges associated with the Association Rule Mining. The scalability issue of ARM algorithms and interestingness of the resulting association rules. This research attempts to find a solution to tackle these issues by presenting a clustering approach for dataset reduction. Moreover, the differential market basket analysis is used to find interesting rules. The next section presents the methodology used for this research.

3 Methodology

This research employs the most widely used data mining methodology i.e. CRoss Industry Standard Process for Data Mining (CRISP-DM). The following steps are followed to systematically conduct this research.

3.1 Business Understanding

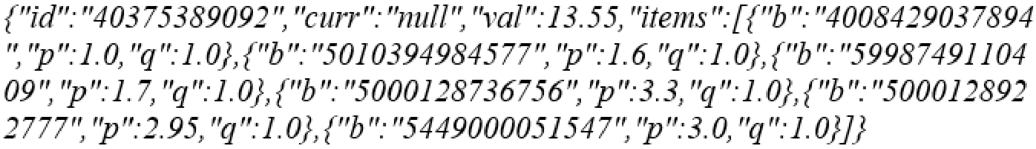
As discussed in the previous section, market basket analysis has been widely used in the retail industry for affinity grouping of the products in the form of association rules. This affinity grouping helps in identifying the purchasing patterns of the consumers. It also uncovers the association between products that are frequently sold together. These associations are useful in deciding the product placement, shelf arrangement, and identifying opportunities for cross-selling of products. These decisions based on associations are beneficial for customer retention and boosting sales. There are two major challenges in Association Rule Mining as previously mentioned. First is the scalability challenges due to the large number of transactions generated every day and the second is filtering the significant rules from a large number of unscrupulous rules (Kotu and Deshpande; 2019). This research attempts to find a solution to these challenges by finding a way to reduce the number of transactions based on clustering and differential analysis. The following subsection provides an overview of the dataset used in this research.

3.2 Data Understanding

The dataset used for this research is the real-world data captured from the retail store chain in the UK. The data files are in the Comma Separated Values (CSV) format. The data includes information about the Products, the Transactions from the stores. The products file has information about all the products that are being sold in the store. This includes the European Article Number (EAN) of the Product which is unique for every product, the Short and long description of the product, its category, and price. The transaction data has attributes such as Basket Id, Total Value of the Basket, Total number of products in the transaction, Time of the transaction, and the Request Basket JSON string which contain the basket data. This basket data is in the semi-structure-nested JSON format. The Figure 1 is the example of one such basket.

This JSON string has information like transaction id, the total value of the basket, and a list of items in the transaction, their price, and quantity. The data of 3 million transactions are available but due to processing limitations, 300,000 transactions are randomly selected. It is observed that in the selected transactions, there are around 16,439 unique items.

The data quality report is generated to know more about the data. Having examined the data quality report, the attributes having missing values or outliers are identified.



```
{"id":"40375389092","curr":"null","val":13.55,"items":[{"b":"4008429037894
","p":1.0,"q":1.0},{"b":"5010394984577","p":1.6,"q":1.0},{"b":"59987491104
09","p":1.7,"q":1.0},{"b":"5000128736756","p":3.3,"q":1.0},{"b":"500012892
2777","p":2.95,"q":1.0},{"b":"5449000051547","p":3.0,"q":1.0}]}
```

Figure 1: JSON example

Some of the important observations are listed below.

1. If the same item is present in two different transactions, then it is found that its price is varying. This variation can be because of an offer, or because the two transactions are from different stores.

2. The maximum length of the transaction, i.e. the maximum number of items purchased in a transaction, is 134.

3. The quantity is found to be negative in a few cases. This could be the result of cancelling the item after scanning it first.

4. For some transactions, the transaction id is missing.

The above observations helped in understanding the pre-processing requirements. The pre-processing and transformation steps are explained in the following section.

3.3 Data Preparation

The data preparation phase is a very crucial phase in the implementation of this research. The data preparation phase involves dealing with the missing data, outliers, inconsistent data, calculating new attributes from the exiting attributes, transforming data, etc. All these data preparation steps are explained in detail in this section.

1. **Handling the semi-structured JSON data.**

   The first and most important step in pre-processing this data is converting the semi-structured basket data to the structured format. This conversion is performed using the 'jsonlite' library in the RStudio. This conversion creates one row per product in the transaction. For example, if a particular transaction contains 5 items, then 5 rows will be created for it.

2. **Handling missing and inconsistent data.**

   As mentioned in the above section, the missing or inconsistent data is present in the dataset. This includes the inconsistent product ids which are present in the transactions but not in the products file, the records with quantity as a negative value, and the records with the missing transaction id's. All these rows with missing or inconsistent data are removed.

3. **Converting the time of the transaction into the categorical variable.**

   The timestamp of the transaction is present in the datetime format. Based on these datetime values, a new categorical variable called 'time_cat' is introduced. The transactions are classified into 4 categories viz. morning, afternoon, evening, and night. This is performed using the 'lubridate' library in the RStudio.

4. **Calculating the frequency and effective price of the product for clustering**

   The new column called 'frequency' is added which has the frequency of each product. As stated earlier, the price of a product is varying in different transactions due to factors such as offers or store locations. Thus, the new column called 'effective_price' is introduced which has the average price of that product that appeared in all the transactions. These two calculated attributes will be used for the clustering purpose.

5. **Removing the transaction containing only one product.**

   The transactions with only one item are removed as these transactions do not contribute to the association rule.

6. **Grouping the transactions based on id**

   While converting data to a structured format, multiple rows were created for each transaction. Hence, the transactions are grouped back based on the transaction id and stored in the different dataframe so that it can be given as input to the different association rule mining models.

7. **Exploratory data analysis**

   Basic exploratory analysis is performed to gain initial insights from the data. Few findings from the exploratory analysis are presented below.

   - Figure 2 shows the top 20 selling products from the overall dataset based on the frequency.

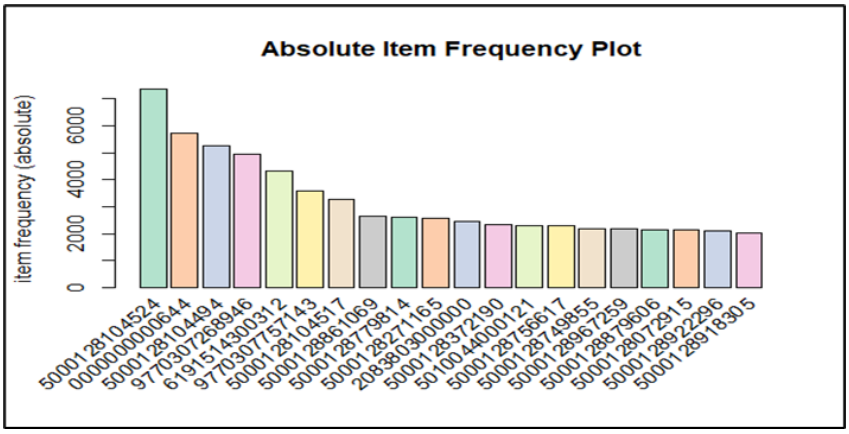


Figure 2: Top 20 Selling Products

   - Figure 3 shows the distribution of transactions based on the time, which clearly shows that a major chunk of transactions occurs in the Afternoon or Evening time.

Thus, the above steps are carried out as a part of the data preparation. After data preparation, the 227,378 transactions are obtained which contain 16210 unique items. In the following section, there is a detailed analysis of machine learning models that are used in this research.

The next section will discuss in-depth about creating and applying the association rule mining model on this processed data.

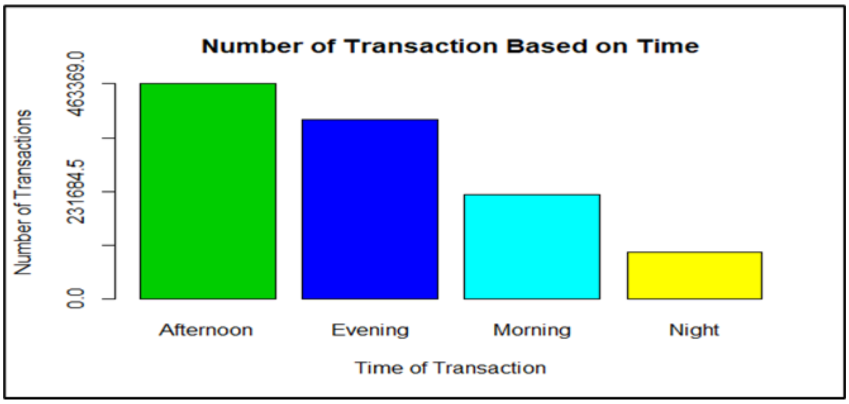


Figure 3: Time based Transaction Distribution

4 Design Specification and Implementation

Once the data is processed, it is time for creating and applying the modeling techniques for further analysis. The next subsection discusses the choice of data models and justification for choosing the selected models.

4.1 Data Modelling

The following models are implemented as a part of this research.

1. **K-Means Clustering**

   K-means clustering is the oldest and the most widely used clustering technique. It is used to group similar entities together. This similarity is calculated based on the parameters provided. It is an unsupervised learning approach that is useful when there is a requirement of grouping the data but the class labels are not known (Yang and Sinaga; 2019).

   In this research, the products are clustered based on their frequency and price. The algorithm is implemented in RStudio using the 'stats' package. To find the optimal number of clusters i.e. K, the Silhouette Plot is used. The Silhouette plot is a visual aid to examine the quality of clusters. The measure called the Silhouette score is calculated which denotes the similarity of an object with its own cluster when compared with the other clusters. This score lies between -1 to +1. The value of K for which this number is close to +1 is the optimal K value[2]. The library used to obtain the Silhouette plot is 'factoextra'. This optimal value of K is then passed to the algorithm. As a result of this, the products are divided into K clusters. Based on the centers of the cluster, the two clusters with higher frequency and higher prices are selected. The products from other clusters are ignored as they are less revenue-generating products. In this research, the role of K-Means clustering is to identify top revenue-generating products. The dataset is reduced based on the results of this model. The results are presented in Section 5.4. The following are the market basket analysis algorithms that are used to generate the association rules.

[2] https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

2. **Apriori algorithm**

   The Apriori algorithm is the most widely used association rule mining algorithm which was first introduced in the year 1994 (Agrawal et al.; 1993). The Apriori algorithm first finds the frequent itemsets and then filters these itemsets to generate association rules. The execution time of this algorithm exponentially increases with the increase in the number of input transactions.

   To implement this algorithm, the data needs to be converted into a specific format. This format is nothing but a boolean matrix with transaction id as rows and products as columns. The columns in a particular row are marked as True if that product exists in the transaction otherwise it is marked as False. Hence, the data is first converted into this format before feeding it to the algorithm.

3. **Frequent Pattern (FP) Growth algorithm**

   The FP Growth algorithm is introduced to deal with the scalability issue of the Apriori algorithm (Kotu and Deshpande; 2019). Despite its complexity, it is widely used for association rule mining. The frequent itemset generation process is improved by introducing a divide and conquer strategy. The data structure called a frequent pattern tree is built. The limitation of this algorithm is that the FP tree generation happens in the main memory which creates a scalability bottleneck when working with large datasets.

4. **Eclat algorithm**

   The Eclat algorithm works in a similar way as that of the Apriori algorithm. It is known as its scalable version. In the Eclat algorithm, the transpose of Aprori's boolean matrix is used. In other words, transaction id becomes columns and the products become rows. This simplifies the frequent itemset calculation process by introducing a depth-first search. Despite this, the algorithm lacks in popularity. Thus, it is interesting to check its performance.

The next section presents the evaluation of various experiments that are conducted as part of this research.

5 Evaluation

This section will detail the experiments that are conducted in this research. It contains the discussion around these experiments and the results obtained by evaluating these experiments. The Market Basket Analysis algorithms do not have well-defined and rigid evaluation metrics. Thus, it is important to perform several experiments and analyze the association rules that are generated, and the time required to obtain these associations. A series of following experiments are conducted starting with the replication of the state-of-the-art.

5.1 Experiment 1: FP Growth algorithm as replication of the state-of-the-art (Hossain et al.; 2019)

Implement Frequent Pattern (FP) Growth algorithm on the state-of-the-art French Retail Dataset and a Bakery Dataset.

The aim of this experiment is to replicate the state-of-the-art (Hossain et al.; 2019). This experiment is performed to verify if the similar results are obtained by replicating the state-of-the-art. This will form a base for conducting further research.

To carry out this experiment, two datasets are used.

- French Retail Store Dataset[3]
- Bakery Dataset[4]

The following pre-processing steps are applied to both the above datasets as described in the base paper (Hossain et al.; 2019). The pre-processed and transformed data is then provided as an input to the FP Growth algorithm. The values of threshold parameters are set as minimum support = 0.01 and minimum confidence = 0.5. The experiment is repeated several times by changing the number of input transactions and the execution time is recorded for each iteration.

After this, the frequency table of all the products is calculated. This frequency table is then used to find the top-selling products. Out of this, 55% of most sold products are selected from both the datasets. The remaining products are ignored as the support is less than the minimum support. The FP Growth algorithm is then applied again to the reduced datasets and execution times are recorded by changing the values of minimum support. Figure 4 and Figure 5 highlight the performance comparison of the FP Growth algorithm in terms of execution time for French Retail and Bakery datasets, respectively.

The results of this experiment showed that the FP Growth algorithm has yielded the same number of rules as the state-of-the-art i.e. 20 for 1000 transactions. The results also showed that the execution time of the algorithm follows a similar pattern as in the state-of-the-art because all the experimental conditions are replicated as described. It can be seen from Figure 4 and Figure 5 that the execution time is reduced when FP growth is used with a product reduction. Following this, the next experiment will discuss the results obtained by replicating a similar approach for the Glantus dataset.

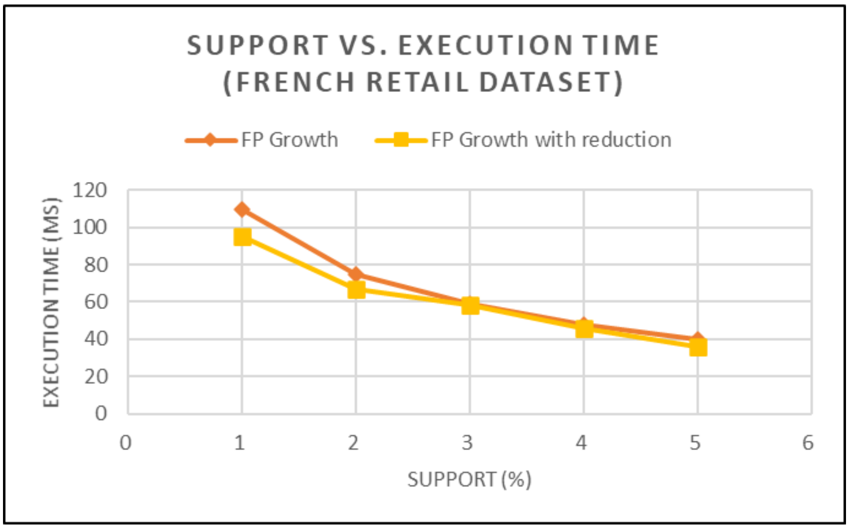


Figure 4: FP Growth Performance Comparison for French Retail Dataset

[3] https://www.kaggle.com/roshansharma/market-basket-optimization

[4] https://www.kaggle.com/sulmansarwar/transactions-from-a-bakery

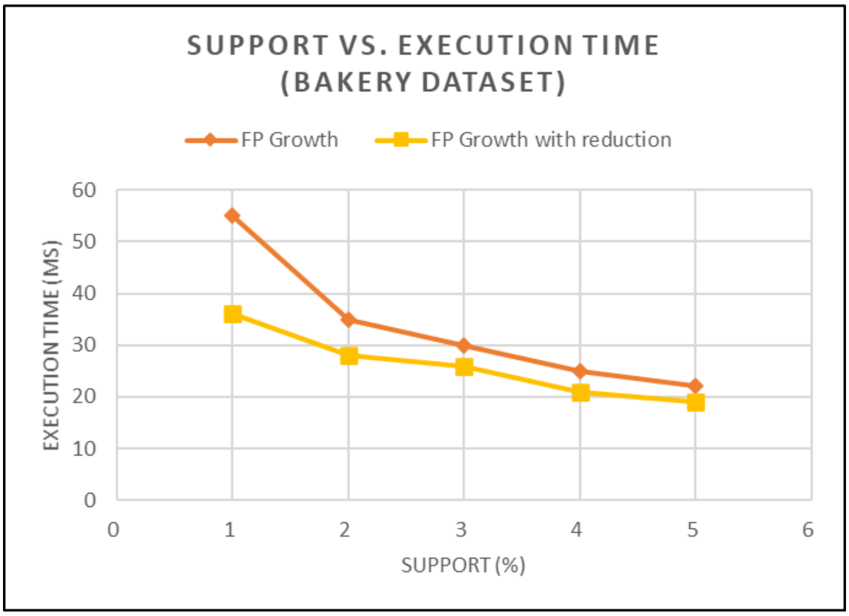


Figure 5: FP Growth Performance Comparison for Bakery Dataset

5.2 Experiment 2: FP Growth algorithm

Implement the Frequent Pattern (FP) Growth algorithm on the Glantus dataset.

The aim of this experiment is to apply the state-of-the-art approach to the Glantus dataset and compare the results with the state-of-the-art. This experiment will check if similar results are obtained when the implementation is replicated for the Glantus dataset.

To carry out this experiment, the Glantus dataset is pre-processed and transformed by following all the steps mentioned in Section 3.3. The threshold values of parameters are taken from the state-of-the-art, which is the minimum support = 0.01 and minimum confidence = 0.5. The FP Growth algorithm is implemented, and results are captured.

The results showed that no rules are generated when the above values of support and confidence are used. The Glantus dataset has about 16,210 unique products and 300,000 transactions. There is no single item that is common in 1% of the total transactions. In other words, 1% of the minimum support is high in the case of the Glantus dataset and cannot be satisfied by any of the products. Thus, no frequent itemsets are generated for this value of support and hence no association rules are generated. In the case of the French retail dataset, there are 7501 transactions with 120 unique items. As there are less unique items they are repeatedly bought in multiple transactions and hence the items can satisfy the minimum support. This is similar in the case of the Bakery dataset as well. Thus, it can be concluded that the minimum support of 1% is not appropriate in this case due to the large number of unique products. The next experiment is performed to determine how association rule mining algorithms perform in terms of execution time.

5.3 Experiment 3: Apriori, FP Growth and Eclat algorithms

Glantus dataset Execution Time Comparison of the Apriori algorithm, Frequent Pattern (FP) Growth algorithm, and Eclat Algorithm

The aim of this experiment is to determine the performance of the chosen algorithms with the change in the number of input transactions. This comparison will be based on the execution time required by each algorithm. This experiment will help in understanding which algorithm requires less time for execution on the chosen dataset. The results of this experiment can be compared later with the results of the proposed approach in the next experiment. This comparison is helpful in determining the efficiency of the proposed approach. Thus, it is important to conduct this experiment.

The chosen dataset i.e. Glantus dataset is pre-processed as per the steps mentioned in Section 3.3. After that, the three algorithms are implemented to find the association rules. Here, the value of the minimum support = 0.0002 and the value of minimum confidence = 0.001. The experiment is performed several times by altering the number of input transactions and the execution times of these algorithms are captured. These captured results are represented in Figure 6.

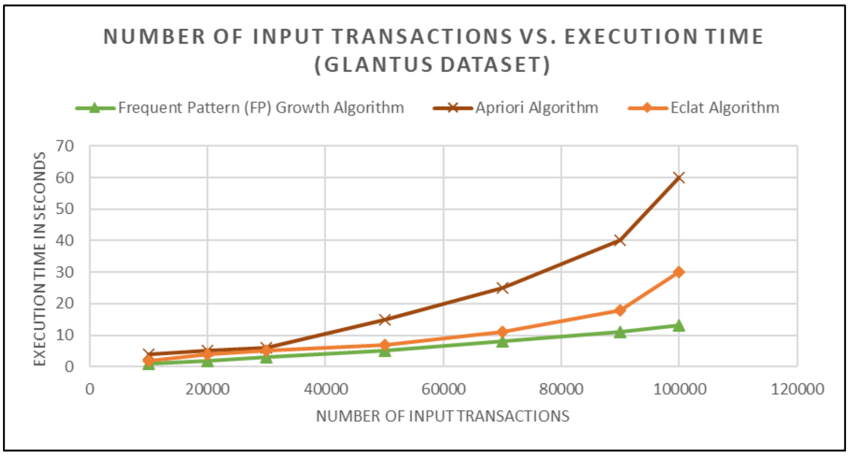


Figure 6: Performance Comparison of MBA algorithms in terms of Execution Time

It can be seen from Figure 6 that the FP Growth algorithm takes less time for execution than the other two algorithms. The execution time required for the Apriori algorithm shows a highly ascending trend as the number of input transactions increase. Also, the Eclat algorithm is faster than the Apriori algorithm. The Eclat and Apriori algorithms yield multiple unnecessary or trivial rules. Hence, it is difficult to extract the interesting rules. In the case of the FP Growth algorithm, the result obtained is more concise. The rules generated are comparatively less in number and non-redundant. The next experiment will check if the clustering based dataset reduction helps in reducing execution times.

5.4 Experiment 4: K-Means Clustering

Implement Clustering based approach for dataset reduction in Association Rule Mining (Glantus dataset)

The aim of this experiment is to determine if the product clustering is efficient for dataset reduction to scale down the processing overhead in Association Rule Mining. This in turn will help to reduce the execution times. As observed in Section 5.3, the execution times of association rule mining algorithms show an ascending trend with an increase in the number of input transactions. However, the tremendous amount of available data creates an issue of scalability. To tackle this issue, it is important to find an appropriate approach to reduce the size of the dataset such that the same set of rules will be obtained after and before reduction.

In this experiment, the K-means clustering algorithm is used to group the items into different clusters. The parameters used for clustering are the frequency and the price of the product. These parameters are computed as described in Section 3.3. Once the data is prepared, it is important to choose an optimal value of K which is done by using the Silhouette plot. As mentioned in Section 4.1, the K value for which the Silhouette score is close to +1 is selected. Figure 7 represents the Silhouette plot, which shows that the optimal value of K = 4.

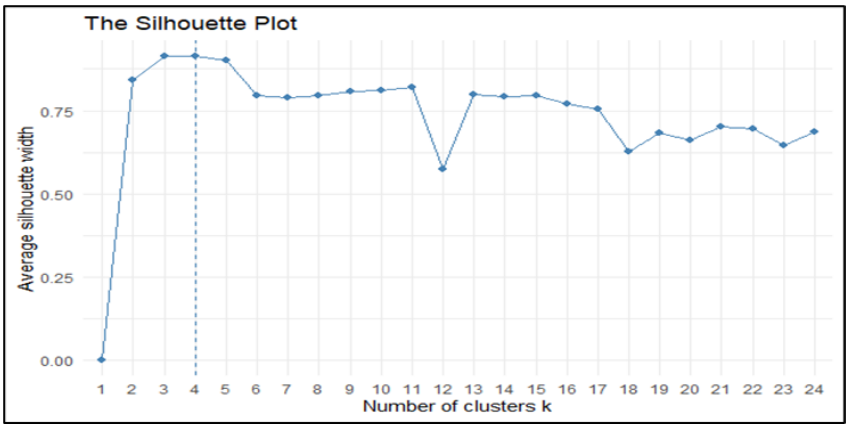


Figure 7: K-means Clustering- Silhouette Plot

After determining the optimal number of K, the K means algorithm is implemented by passing frequency and price as a parameter, and the value of K=4. Table 1 represents the results obtained by clustering. Cluster 2 has a very low frequency, though the price is a bit high. In contrast, other clusters have a very high frequency as compared to cluster 2. Thus, the products belonging to Cluster 2 which is a major chunk of products are less revenue-generating products and can be ignored.

Table 1: K-Means Clustering Results

Cluster	# Products	Centre	
		Frequency	Price
1	7	5840.71429	2.96257
2	14925	27.98968	2.666626
3	213	1303.06573	1.702375
4	1065	413.00845	1.824922

As per the Clustering results, all the products except cluster 2 products are selected. The selected products are 7% of the total products. Thus, the dataset can be significantly reduced with the help of clustering. Now, it is important to check if the same set of rules is obtained or not when this reduced dataset is given as input to the association rule mining algorithms.

It is observed that the same set of rules is generated even after processing the reduced dataset. The next is to compare the performance of these algorithms in terms of execution times. The experiment is conducted several times by increasing the input support values. Based on the reading obtained, the following graphs are plotted. Figure 8, 9, 10 represent the performance comparison of Apriori, FP Growth, and Eclat algorithms with an increase in the input support values. All three figures show that the execution time is reduced significantly with product reduction in the case of all three algorithms. Also, the execution time is reduced with the increase in the support values. Hence, it can be concluded that the product clustering approach can significantly lower the processing overhead thereby reducing execution time. In the next experiment, the differential market basket analysis approach is used to compare the association rules based on the time of the transaction.

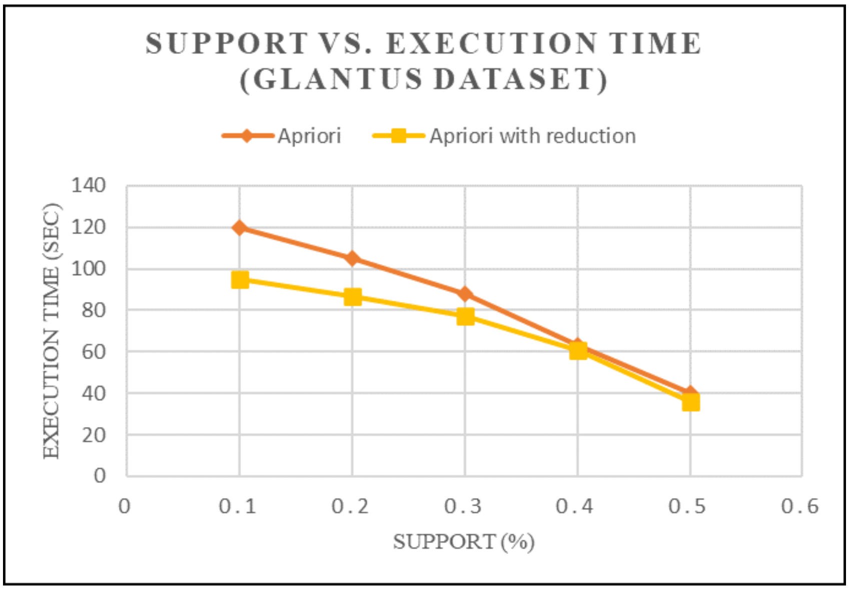


Figure 8: Apriori algorithm- Performance Comparison

5.5 Experiment 5: Differential Market Basket Analysis

Implement differential market basket analysis and Association rule mining on the Glantus dataset.

The idea of differential market basket analysis is to compare the results of association rules across the different groups of transactions. For example, comparing the results from different stores, or from the different seasons of the year. This comparison can lead to the identification of interesting rules which might not be highlighted if the association rules are obtained from the overall transactions.

The aim of this experiment is to perform the differential market basket analysis to assess if the rules obtained from each group are identical or different. The transactions are grouped based on the time of the day. This analysis will help in identifying the effect of time-based sampling of the input transactions on resulting association rules.

To carry out this experiment, all the transactions of the day are grouped into four groups based on the time of the transaction. The four groups are Morning, Afternoon, Evening, and Night. The Apriori and FP Growth algorithms are then applied individually to all four groups of transactions as well as to the all-day transactions to record. The results of each group are then compared to the all-day results. Here, the minimum support = 0.002 and minimum confidence = 0.005.

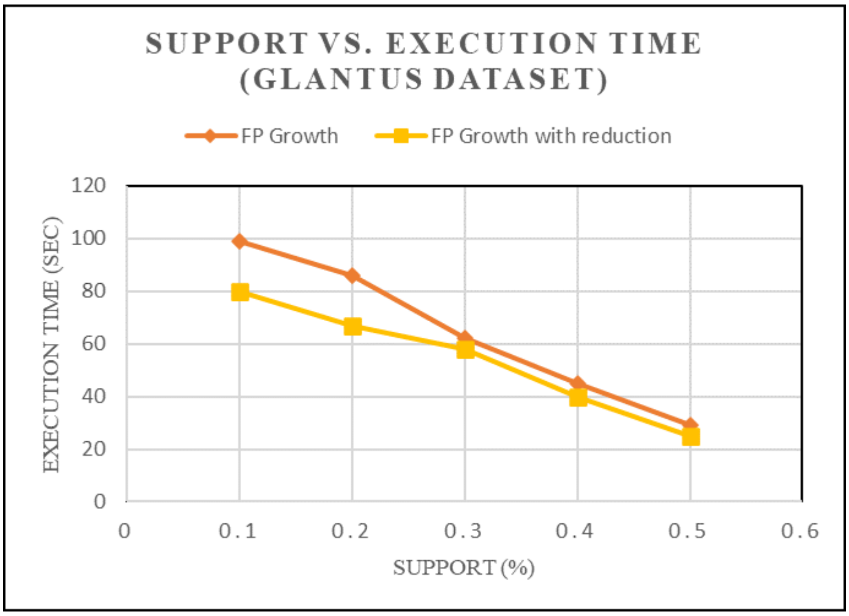


Figure 9: FP Growth- Performance Comparison

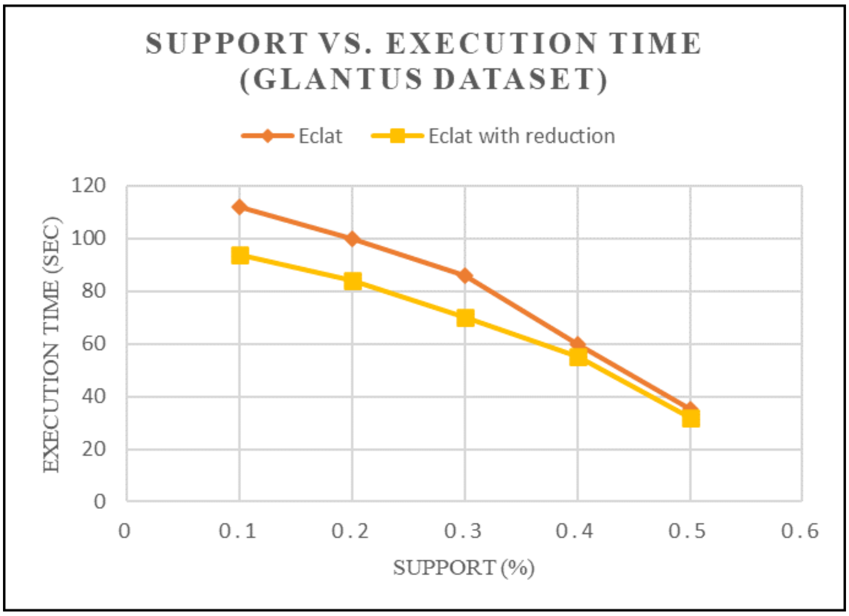


Figure 10: Eclat- Performance Comparison

Table 2 shows the comparison of results obtained from all the groups. It can be seen from the below table that the number of transactions is highest in the afternoon whereas the lowest number of transactions are observed at night.

Table 2: Comparison of Results from different groups

Groups	# Transactions	# Items	# Rules	# Frequent Itemsets
Morning	22294	7894	8	317
Afternoon	36567	10623	54	379
Evening	31290	9909	60	379
Night	9507	6057	54	305
All-Day	99658	13238	77	357

The number of rules generated is different in each case depending on the number of transactions and unique items from these transactions. The results of this experiment also showed that the top-selling products differ in the 'Morning' group.

When the rules of each group are compared with 'All-Day' rules, the rules generated in the 'Afternoon', 'Evening', and 'Night' are comparatively similar to 'All-Day' rules. In the case of 'Morning', the rules generated are different from all the other groups. These are the interesting rules which are obtained by processing the Morning transactions. Also, in this case, when only the 'Afternoon' and 'Evening' transactions are considered, the same set of rules obtained as that of 'All-Day' due to the presence of products that constitute a major part of all transactions. In the following section, all the above experiments and their results are discussed.

5.6 Discussion

This section presents the discussion of the above experiments with respect to the results obtained. This discussion will help in understanding the contribution of this research in this domain.

This research starts with replicating the state-of-the-art (Hossain et al.; 2019) where FP Growth algorithm is implemented on two datasets. As discussed in Section 5.1, the experiment demonstrated that the execution time reduces as there is an increase in the level of minimum support. Also, there is a reduction in the execution time because of product reduction. If the dataset is huge, a large number of frequent itemsets are generated which requires more computation power as well as memory. This approach of product reduction minimizes computation, which in turn reduces the execution time. Hence, it is evident from this experiment that the approach suggested in the state-of-the-art is helpful in minimizing the execution time. Hence, in the second experiment the same approach is followed for the Glantus dataset. As the values of minimum support and minimum confidence were inappropriate, no association rules are generated. This experiment demonstrates that the threshold values of support and confidence should be chosen based on the dataset and its characteristics in order to generate the association rules. To decide these threshold values, the summary of item frequency should be checked. This summary includes the minimum, maximum, and average frequency of the unique items in the dataset. This will help in choosing the appropriate values of minimum support and minimum confidence.

In the third experiment, three association rule mining algorithms i.e. Apriori, FP Growth and Eclat are compared in terms of execution times. As state in Section 5.3, the

FP Growth algorithm resulted in less execution times than other two algorithms. These results are similar to the results obtained in the literature (Davanbu and Venkatachari; 2016; Hossain et al.; 2019) .These three algorithms first generate frequent itemsets, which is followed by the rule generation. The Apriori and Eclat algorithms use a similar approach with multiple database scans which increases the execution times. The only difference between these two algorithms is the underlying data structure. Also, these two algorithms to check for all the possible patterns due to which the larger number of rules are generated. Some of these rules are redundant and inexplicable. In the case of the FP Growth algorithm, the database is scanned only twice which makes its execution faster. Also, the FP Growth algorithm generates a smaller number of rules as it only considers the patterns that are present in the database. It can be seen from this analysis that the FP Growth algorithm is better in terms of both execution time and optimized rules generation.

In experiment 4, the performance of these algorithms is evaluated after clustering based dataset reduction. The results showed that Clustering is an efficient approach for dataset reduction. The clustering selects only 7% of total products which can be used to generate the same set of rules that can be obtained without product reduction. However, in the state-of-the-art (Hossain et al.; 2019) the dataset was reduced using 55% of top-selling products to obtain the same set of rules. This approach goes beyond the state-of-the-art and can effectively reduce the dataset with top revenue-generating products which are just 7% of the total dataset. Thus, the Clustering based on frequency and price is a more effective approach for reduction than the state-of-the-art.

The experiment 5 demonstrates that the differential market basket analysis helps in finding the interesting association rules which might not be obtained by processing the complete set of transactions. This comparison will be helpful in finding a specific set of association rules. The approach similar to differential market basket analysis is used in another research where they found the interesting results from a few groups (Fong et al.; 2017).For example, rules based on seasonal data or based on the week of the day, etc. To calculate the overall association rules, it is not required to process the whole dataset. Instead, the transactions can be grouped based on some criteria, and the major groups can be processed to obtain the same set of rules. This sampling will reduce the computation overhead thereby reducing the execution times and highlighting interesting rules. Following this discussion, the next section will present the key findings of this research.

6 Conclusion and Future Work

In this research, the combination of clustering and differential market basket analysis is implemented to assess the improvement in the performance of association rule mining algorithms. The performance is assessed in terms of execution times and interestingness of rules.

Firstly, the current work in this domain is investigated and the results are verified by replication of the state-of-the-art, thereby achieving the 1st objective that was set. The results showed that the FP Growth algorithm requires less execution time than Apriori and Eclat algorithms. The clustering based approach is designed by deciding frequency and price as clustering parameters. The transaction time is decided as a parameter for differential market basket analysis thus achieving the 2nd research objective. To achieve the 3rd objective, the K means clustering with K=4 is implemented to find the top revenue generating products. Based on the clustering, the dataset is reduced by selecting only 7% of the total 16,210 products. The association rule mining algorithms are then applied to this reduced dataset. The differential market basket analysis is implemented by grouping transactions into four groups i.e. 'Morning', 'Afternoon', 'Evening', and 'Night'. The apriori algorithm is implemented for each of these groups.

The 4th objective is achieved by evaluating the outcomes of the above implementation. The results showed that the same set of rules is generated with and without product reduction. From experimental analysis, it is evident that the dataset reduction reduces the execution times as processing overhead is reduced. The results of differential market basket analysis showed that rules generated from each set are similar except for the 'Morning' group. In other words, interesting rules are obtained from a 'Morning' group. It is thus evident that such comparisons across different groups are helpful in finding interesting patterns. In the future, these comparisons should be performed on broader groups such as seasonal transactions, or the transactions from different stores.

In conclusion, the Clustering based approach for dataset reduction is efficient in dealing with scalability issues. The consequences caused by product exclusion on business should be studied in the future. Moreover, this approach can be used in other domains such as banking, medical, etc. to assess its efficiency.

References

Abbas, W. F., Ahmad, N. D. and Zaini, N. B. (2013). Discovering purchasing pattern of sport items using market basket analysis, *Proceedings - 2013 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2013* pp. 120–125.

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining Association in Large Databases, *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93* pp. 207–216.

Chun-Sheng, Z. and Yan, L. (2014). Extension of local association rules mining algorithm based on apriori algorithm, *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS* pp. 340–343.

Davanbu, I. and Venkatachari, K. (2016). Market Basket Analysis Using FP Growth and Apriori Algorithm: A Case Study Of Mumbai Retail Store, **8**.

Fong, S., Biuk-Aghai, R. P. and Tin, S. (2017). Visual clustering-based apriori ARM methodology for obtaining quality association rules, *ACM International Conference Proceeding Series* **Part F1301**(1): 69–70.

Gassama, A. D. D., Camara, F. and Ndiaye, S. (2017). S-FPG: A parallel version of FP-Growth algorithm under Apache Spark$^{\text{TM}}$, *2017 2nd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2017* pp. 98–101.

Gijsbrechts, E., Campo, K. and Vroegrijk, M. (2018). Save or (over-)spend? The impact of hard-discounter shopping on consumers' grocery outlay, *International Journal of Research in Marketing* **35**(2): 270–288.
URL: *https://doi.org/10.1016/j.ijresmar.2018.01.004*

Hossain, M., Sattar, S. A. H. M. and Paul, M. K. (2019). Market Basket Analysis Using Apriori and FP Growth Algorithm, *2019 22nd International Conference on Computer and Information Technology (ICCIT)* pp. 1–6.

Huang, W., Chen, J., Liu, C., Shi, W., Lin, C., Lyu, X. and Gao, H. (2020). Research on line-loss correlation analysis technology of distribution network based on apriori Algorithm, *Proceedings - 2020 5th Asia Conference on Power and Electrical Engineering, ACPEE 2020* pp. 1399–1403.

Huseyinov, I. and Aytac, U. (2017). Identification of Association Rules in Buying Patterns of Customers based on Modified Apriori and Eclat Algorithms by using R Programming Language, pp. 6–11.

Kotu, V. and Deshpande, B. (2019). Association Analysis, *Data Science* pp. 199–220.

Kurniawan, F., Umayah, B., Hammad, J., Nugroho, S. M. S. and Hariadi, M. (2017). Market Basket Analysis to Identify Customer Behaviours by Way of Transaction Data, *Knowledge Engineering and Data Science* **1**(1): 20.

Kutuzova, T. and Melnik, M. (2018). Market basket analysis of heterogeneous data sources for recommendation system improvement.

Liu, Y. and Lou, Y. S. (2014). Research and application of improved Apriori algorithm based on Hash Technology, *Applied Mechanics and Materials* **668-669**: 1102–1105.

Moodley, R., Chiclana, F., Caraffini, F. and Carter, J. (2020). A product-centric data mining algorithm for targeted promotions, *Journal of Retailing and Consumer Services* **54**(October 2019): 101940.
URL: *https://doi.org/10.1016/j.jretconser.2019.101940*

Sorensen, H., Bogomolova, S., Anderson, K., Trinh, G., Sharp, A., Kennedy, R., Page, B. and Wright, M. (2017). Fundamental patterns of in-store shopper behavior, *Journal of Retailing and Consumer Services* **37**(July 2016): 182–194.
URL: *http://dx.doi.org/10.1016/j.jretconser.2017.02.003*

Valle, M. A., Ruz, G. A. and Morrás, R. (2018). Market basket analysis : Complementing association rules with minimum spanning trees, *Expert Systems With Applications* **97**: 146–162.
URL: *https://doi.org/10.1016/j.eswa.2017.12.028*

Valle, M. A., Ruz, G. A. and Rica, S. (2019). Market basket analysis by solving the inverse Ising problem: Discovering pairwise interaction strengths among products, *Physica A: Statistical Mechanics and its Applications* **524**: 36–44.
URL: *https://doi.org/10.1016/j.physa.2019.03.001*

Yang, M.-S. and Sinaga, K. P. (2019). A Feature-Reduction Multi-View k-Means Clustering Algorithm, *IEEE Access* **7**: 114472–114486.

Zaki, M. J. and Gouda, K. (2003). Fast Vertical Mining Using Diffsets, pp. 326–335.

Answer to Follow on question

1. **What is the reason behind selecting K-means algorithm for clustering in your project?**

   The most important parameter for the selection of K-means is its computation cost. K-means clustering takes less time when the dataset is large. The objective of the research is to reduce the execution times, and hence speed and efficiency of K-means are the important criteria for selection. Apart from these, the other advantages are it is easy to implement and interpret. Also, it works well with real-life data. The important challenge was deciding the number of clusters which is done by using the Silhouette plot. Based on the literature studied, another option was to use the Fuzzy C-means clustering. The problem with this method is, it assigns more than one cluster to each data point which is not desirable in this case. Hence, the efficient K-means algorithm is selected which assigns a single cluster to each data point.