National College of Ireland

Customer Visit Segmentation Based on Clustering and Association Rules

MSc Research Project
Data Analytics

Vishakha Balkrishna Kale
Student ID: X18181643

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes, Dr. Pramod Pathak

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Vishakha Balkrishna Kale
Student ID:	X18181643
Programme:	M.Sc. in Data Analytics **Year:** 2020
Module:	Research Project
Supervisor:	Dr. Paul Stynes, Dr. Pramod Pathak
Submission Due Date:	17th August 2020
Project Title:	Customer Visit Segmentation Based on Clustering and Association Rules.
Word Count:	6974 **Page Count:** 20

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: Vishakha Kale

Date: 16th August 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):	

Customer Visit Segmentation Based on Clustering and Association Rules

Vishakha Balkrishna Kale
X18181643

Abstract

Retail businesses are highly involved with customers, where customers can contribute in the profit and loss of the business thus makes them an important factor to be studied and analysed. Among the various factors studied for customer analysis such as market basket analysis and Customer Segmentation, Customer visit segmentation can also be considered as a meaningful analysis of customers and their shopping visits. Where Customer segmentation explains the motive behind each visit of customer to a retail shop, this research aims at adding a useful analysis to these visits by studying them further with Association rules. The aim of this research is to analyse the customer visits formed with k-means clustering by further analysis using Apriori and Eclat algorithm. This study gives a contribution towards deep analysis of customer visit segmentation using association rules and attempts to improve the performance using Eclat algorithm.

1 Introduction

Customer is always the first priority of every business; it has also proven many times that customer-oriented organisations are successful and ever growing in cooperate world. This was figured out by many organisations and they are trying to implement customer centric approach as their work criteria. To get into the shoes of the customers and try to merge according to the new trend followed by customers is constantly in generating huge profit. According to alias Devi & S.P. Rajagopalan(2012), customer satisfaction is critical in customer relationship management to confirm an increasing graph of customer loyalty and retention. Shrewd buyer being a subject of concern and important for all evolutionary acquiring management, where investors and businesses are willing to empty their pockets. And this is not only just in sales section but also in various other sections like telecommunication (Khamis Mwero Manero et al., 2018). Thus, all the researches are taking a curve towards the customer relationship management which is going above and high into their preferences of the organisations. Customer activities supports to make the procedure very simple and easier as the authors analyse the information/data of customer all performed act or their activities from ample of sources and try to fetch some beneficial perceptions from the same. This method carries a significant value in the business vision and also marketing section could get beneficial information from it.

Through the market basket analysis and customer segmentation practices purchaser is been acquired in present world. Where market basket analysis attention is mainly on customer activities or demand of purchase, customer segmentation aims mainly on the groups of customers and their attributes. Both of these methods are operational in various type of requirements. Like (P.Isakki alias Devi, 2012)(Kaur and Kang, 2016) has operated on market basket analysis to guide marketing team handle their forthcoming activities and ideas. Alias

Devi & S.P.Rajagopalan (2012) has performed on customer segmentation to support the new product launching decisions where knowing customer demands/trends are key points.

Furthermore, apart from customer segmentation and market basket analysis, there is one remarkable concept described and exhibited by researcher (Griva et al., 2018) is Customer Visit Segmentation. The concept involves around recognising the motive of visit of every customer. The researcher has specified that each visit of the consumer is generally has a motive and recognising that could be very vital and helpful in several ways such as shelf organisation, marketing, discount and vouchers etc. Also, if customer visit segmentation utilised further to find information according to the days and parts of the days which can be more useful. In this study, it is executed well using clustering and the details about the customer visits is perfectly mentioned according to the days and part of the day.

Findings has also mentioned that though this study is fine acting out and capable of supporting, it can be further improved by adding this association rule mining into it. Moving with this study ahead, it is proposed in this dissertation the purpose to make customer visit segmentation more successful with the help of clustering as well as association rule mining. This research will elucidate each customer visit cluster in more valuable way. (Balaji et al., 2012) has encouraged to combine association rule with clustering to gain more acquaintance form each cluster in order to make the purpose supportive. This dissertation is the work to be accompanied with the intention to solve the following research question:

• How Association Rule mining contributes in improving clustering-based customer visit segmentation?

To address this research question, the set of research objective derived are as follows:

1. Investigate the state of the art broadly around improvising the customer visit segmentation.
2. Design a model to incorporate clustering and association rule to give out useful insights.
3. Implement the K-means clustering and Apriori algorithm along with all the necessary data handling to give out finest results.
4. Implement Eclat Algorithm to analyse the frequently bought item sets in each visit segment.
5. Evaluate the performance of models along with their contribution to the existing research.

To solve these objective of the research, stepwise methodology is to be followed modelled with KDD. Initial steps to be taken would be Data cleaning, formatting then leading to data mining. Data mining steps to be followed are K-means algorithm for clustering which would ideally give out the customer visit segments stating the motive of the visit. This data would then be followed for association rules. All these data mining methods would be evaluated on the basis of different evaluation method to finally discuss the outcomes of the research and fulfilment of the objectives.

2 Related Work

The concepts behind the motivation of the research are customer segmentation and market basket analysis. As the topic of this research is an amalgam of the both it would be interesting to know the basics of these notions and how they inspired the idea of the research project.

2.1 Customer Visit Segmentation with Clustering and Association Rules

The concept of customer visit segmentation is inspired from the terminology of customer segmentation which is used widely to differentiate customer profiles according to their shopping behaviour. According to (Kansal et al., 2018) Customer segmentation makes easy to handle large data of customers. Segmentation of customers hold high importance as it helps in knowing the pattern of customers according to their categories which helps in planning marketing and business strategies to grow in business.

These customer segments can provide various types of information regarding potential buyers of each segment increasing the value of customer segmentation. Similar experiment has done by (Calvo-Porral and Lévy-Mangin, 2018) to name each customer segment according to the likes and dislikes of the customer regarding product purchase. This study contributes towards the better understanding of customer buying preferences in a specialty store for strategy and marketing.

Customer Visit segmentation is a similar concept which focuses on the type of visit motive for which the customer visits the retail shop in each segment. (Griva et al., 2018) demonstrates the customer visit segments in a highly efficient way using K-means clustering to give out neatly differentiated customer visit segments helping the store management to know at what time of the day and which days of the week most of customers has what motive to visit the shop.

Though the customer visit segmentation assists in a very useful way, there is a scope of further analysis to retrieve maximum knowledge from these visit segments.

Association rule mining is known popularly for market basket analysis which states the products bought together by a customer based on its basket transactions. These rules are used by retail shops to gather the information about the frequently bought products and their chances of being in a basket together.

These benefits of association rules are applied many times by researchers to prove the higher benefit when applied to customer segments. As (P.Isakki alias Devi, 2012) implement association rules using apriori algorithm on customer segments formed using k-means algorithm, states that the research would also help in planning new line of products. (Agarwal, 2017) demonstrates decision making system for supply chain management firms where using association and clustering, the research is aims to help in ease in knowing the demand of the stock for stock items arrangement. The experiment (Silva et al., 2019) is well explaining in terms of benefits of analysing clustered segments using association rules as the researchers implements apriori algorithm on clustered segments formed according to loyalty of the customers. It is observed that the level of loyalty of customers is well analysed with the help of association rules giving out the frequent products bought together by those customers. This theory is backed up by (Miguéis et al., 2011) who implements association rules of customer segments to get the products bought together in a set of 2 explaining that the experiment would help in increasing the loyalty of customers of a retail store.

Clustering based association rule has also seemed to be helpful in reducing the computational cost (Quan et al., 2009) when association rules are applied on smaller segments instead of entire dataset. Along with reducing computational cost it also enhances the results belonging to specific segments.

Importance of clustering-based association rules is explained for online shopping market as well where researchers (Riaz et al., 2014) explains the benefits with demonstrating a product recommendations system. Online segmentation is also benefited from the cluster-based association rules for product recommendation where researcher (Changchien and Lu, 2001) has implemented customer product fragmentation reflecting the flexibility of the concept in the form of its advantages as the system helps in one to one product recommendation of customers and also states the favourite products of customers.

Customer segmentation along with association rule is also flexible with the channel of market where (Liao et al., 2013) experiments the segmentation along with association rule of three channels of products to gather segments with the product buying behaviour in it. It is also observed from the research (Suhail Najam and Hashim AL-Saedi, 2018) that when association rules are applied on cluster segments instead of entire dataset, it helps in gaining the accuracy and the high level of analysis of each segment increasing the understandability.
Where clustering is usually done on customers, we can also see a unique research where segmenting is performed on the store branches. The branch of the retail stores where segmented and association rules were implemented on the segments to get knowledge about the products bought in those store branches to plan separate marketing strategies.
All the above study shows that when association rules mining is collaborated with clustering-based segmentation it assists in retrieving more helpful knowledge from the data analysis.

2.2 Apriori and Eclat Algorithm for Association Rule Mining

When it comes to association rules or market basket analysis, it is seen that apriori algorithm is used on a popular demand. It would be curious to know the benefits of the algorithm and to analyse if the algorithm is in fact the best for all the analysis based on association rules.
As per studies by some research, it is perceived that some of the algorithms such as eclat, performs better than apriori algorithm.
When Apriori and eclat algorithms are implemented on the real-world data by (Robu and dos Santos, 2019) to extract frequent data mining, they are compared on basis of efficiency, performance, support distribution and number of rules generated showing that eclat algorithm has a better performance, better efficiency. Support distribution is observed to be equal, but rules generated by the apriori algorithm are more than what generated by eclat algorithm.
(Kotiyal et al., 2013) has implemented apriori and eclat algorithm to compare the performance for behaviour analysis of the user in web log. It is concluded in the research that eclat algorithm has shown better performance than apriori algorithm with large dataset with the generation of lesser tables and thus taking lesser time.
Another study (Borgelt, n.d.)for comparison of apriori and eclat algorithm has shown the study based on memory usage and execution time tested on 5 datasets. The study concludes the superior performance of eclat algorithm compared to apriori in terms of memory usage on 4 out of five datasets whereas according to execution time it proves to be better in all 5 datasets.
A quantitative study of apriori and eclat algorithm is implemented by researcher (Tanu Jain, 2016) based on R platform where the comparison of the two algorithm is performed based on the datasets of different volumes. The researcher proves by this study that eclat algorithm has shown better performance in both the cases when dataset was smaller as well as when it was large.
Similar to the above studies there are two more studies including the performance comparison of apriori and eclat algorithm in terms of their execution time and memory usage. These studies back up the same observation in most cases where the researchers claim to observe the eclat algorithm performing better than apriori.
As per shown in lot of studies, giving enough evidences, it makes interesting to study if the eclat algorithm makes association based customer visit segmentation better performing model.

2.3 Summary

It is observed from the above research that Customer visit segmentation is an interesting modification of the customer segmentation which can be implemented using K-means clustering. It would be exciting to analyse whether it helps in every other dataset. It is also studied that association rule mining can help further to know about the shopping behaviour of

the customer. The use of clustering-based association rule has inspired to experiment it on customer visit segmentation. For the improvement of the segmentation results, it would be beneficial to classify the basket data according to the categories as per learnt from one of the mentioned research. It is observed that eclat algorithm is most of the times performs better than apriori algorithm becomes a curiosity to experiment it in the research to make the model efficient in performance.

3 Research Methodology

The research methodology is designed according to Knowledge Discovery in Databases methodology. The methodology to be followed as planned would explained in following sections:

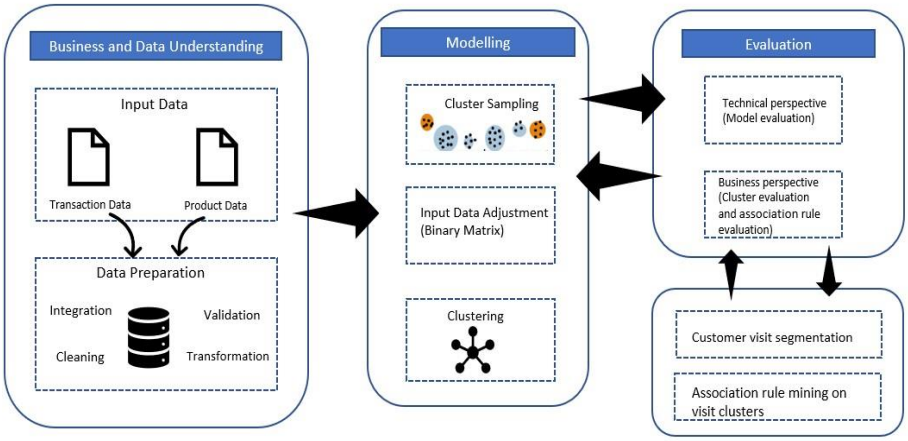

Figure 1 : Customer Visit Segmentation Methodology

3.1 Data Selection

This step involves Business and Data understanding method, Type pf data used in this step would be Transaction data and Products data.

The Research s based on the 'Glantus Dataset' which is provided by the 'NCI-IPP' Team. The data is Real-time data and metadata of the tabular data used is explained in the following tables.

Transaction Data
Date and Time
Store
Value of Basket
Count of items in basket
Items in basket

Transaction dataset provided includes the columns giving information about Date and time of the transaction, Store where the transaction has happened, value of the products bought in the basket, count of the items of basket and a json string detailing about the items in each basket.

Products Data
EAN
PLUCode
RetailLineCode
Department
Section
Subsection
BrandTypeCode
Description
WeightedItem
ShortDescription

The Products data given has information about the products sold by the shop. Among all the data provided, columns related to EAN code(Barcode) of product and department, section and subsection of the product is used in the research.
This data would be stored in SQL server using import export wizard of SQL Server Management Studio.

3.2 Data Pre-processing and Manipulation

The required tables from the data are to be checked for missing or garbage values to increase the quality of data.
The data was analysed to result in some duplicate entries in transaction table. Duplicate transactions were removed. It also contained different baskets with same transaction ID which could result in duplication in results leading to errors hence were removed completely.
As per stated by (Griva et al., 2018), Baskets with Large transaction as well as small transactions could cause the quality of data analysis to get lower as the small baskets does not help in concluding any specific shopping motive and larger baskets can possibly conclude abstract shopping visit with range of categories of products. Due to this reason, Baskets with more than 30 products and less than 3 products were removed to get the data quality raised.

3.3 Data Transformation

On this stage, the data is transformed into the required format of basket data.The data to be convert into the horizontal data with basket ID as a key column and all the items in the basket as separate columns having values 0 or 1 based on their availability in the basket. This sparse matrix data is required for clustering.

The data retrieved from clusters is in same format and used in association rules in Visual studio. For Rstudio, again the format would be different, it would be baset format and it is formatted in RStudio only.

3.4 Data Mining

Data mining algorithms used in the research are K-means clustering, Apriori Algorithm and Eclat algorithm.
K-means clustering is used for customer visit segmentation in visual studio using SQL Server Analysis Services.

Apriori algorithm is implemented to analyse the customer segments formed in terms of the products bought in each segment.
Eclat algorithm is implemented in RStudio to compare the performance with apriori and to test the results to achieve better performing model.

3.5 Evaluation

3.5.1 Clustering Evaluation

Cluster segments formed with K-means clustering are evaluate according to the Sum of Square Error (SSE) values as per retrieved from Elbow Method.
Cluster segments are also evaluated based on the most number of product categories present in that cluster to analyse about the potential visit motive of the segment and to name it.

3.5.2 Apriori Evaluation

Apriori algorithm are evaluated for the association rules retrieved from the mining. The values considered for the evaluation in visual studio are Importance (Lift), Support and Probability (Confidence).

3.5.3 Eclat Evaluation

Eclat Algorithm implemented in RStudio is evaluated based on only support value as the algorithm is very simple in terms of generating item sets that stated the items bought together.

4 Design and Implementation

Figure 2 : Design and Implementation of Research

4.1 Data Transformation and Pre-processing

In this process, the dataset table of Products and Transaction were joined together using left inner join, Data was sampled to remove baskets with less than 3 and more than 30 products and the table was converted to binary matrix with basket ID as a key and each basket Item as a column with values 0 and 1 depending on the availability of the product in that basket. Binary matrix was created using dynamic pivot function of SQL Server with the help of SQL Server Management Studio.

4.1.1 SQL Server Management Studio

SQL Server Management Studio is an open platform for handling and processing SQL Database. It is a convenient and user-friendly software for managing the SQL Databases with ease. SSMS (SQL Server Management studio) is also used in accessing and exploring the data mining results. For Data pre-processing and manipulation SSMS was used as a platform.

4.2 Visit Segmentation using K-means clustering

After the process of Data cleaning, manipulating and sparse matrix formation as per explained in methodology, this data is given as input to form the cluster segments in visual studio of SQL Server Analysis Services.
By creating a Data source and Data views of the table with transaction and products related data, A mining structure is created by selecting Microsoft clustering as a model. Before processing the created model, optimum initial number of clusters were selected as K=6, as per given in the elbow method.

The created model thus processed with the required settings to get a well-formed clustering model detailing about the visit segments of the customers.

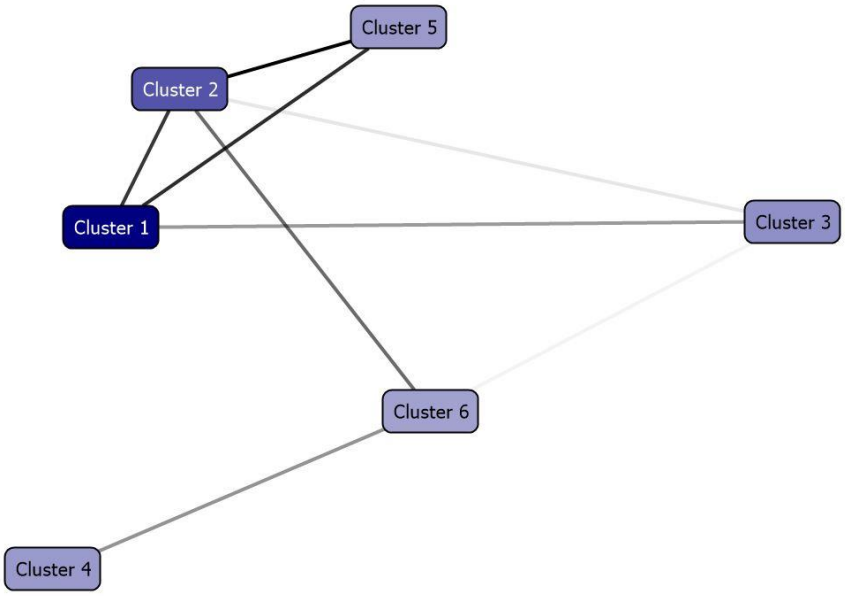

Figure 3: Clusters of Customer visit segments

4.2.1 SQL Server Analysis Services in MS Visual studio

SQL Server Analysis services along with Microsoft visual studio offers a wide range of data mining algorithms to perform. All of these are Microsoft algorithms and allows the researcher to explore the results of the data mining in many forms such as data mining viewer and model viewer.
Data mining steps are performed by accessing the SQL database in Visual studio of SSAS where all the required and transformed database where stored by using connection manager.

4.3 Association Rule Mining

Association rule mining for each customer visit was implemented using two algorithms: Apriori and Eclat. Apriori algorithm was implemented using Microsoft association rules of SQL Server Analysis Services in MS Visual studio. Whereas Eclat algorithm was implemented using RStudio.
The dataset used for this implementation was in the format of basket data with Basket ID as a key column and the products present in the basket in a row.

4.3.1 Association rule mining of clusters in Visual Studio

4.3.1.1 Microsoft Association Rule (Apriori Algorithm)

Microsoft association rule is just an implementation of apriori algorithm with improved visualization and organization in displaying the rules.
Apriori algorithm is used for finding frequent itemset in a dataset with the help of Boolean association rule This algorithm uses a iterative approach to fine a frequent item set with a prior knowledge the item sets hence called as apriori.
The Item sets can be analysed using following parameters:
1. Support (Frequency): Total number of cases with the analysed item
2.Confidence (Probability): probability of the item set happening with respect to all other item set
3. Lift (Importance): The importance calculated for each item set as probability of the item set divided by the compound probability of each item in itemset.

The cluster segments are further analysed using association rule mining in Visual studio. As the cluster segments are formed in SQL Server Analysis Services, it was very feasible to access the transactions involved in each cluster. Using DMX (Data Mining XML) query, the 'cases' involved in each cluster were retrieved and saved as a CSV file to be used further. Then the files were saved on SQL Server in table format for further analysis.
These SQL tables representing the data for each cluster was received in visual studio by connecting to the server. By creating correct data source and data views, new mining structures were modelled as 'Microsoft association rule mining'.

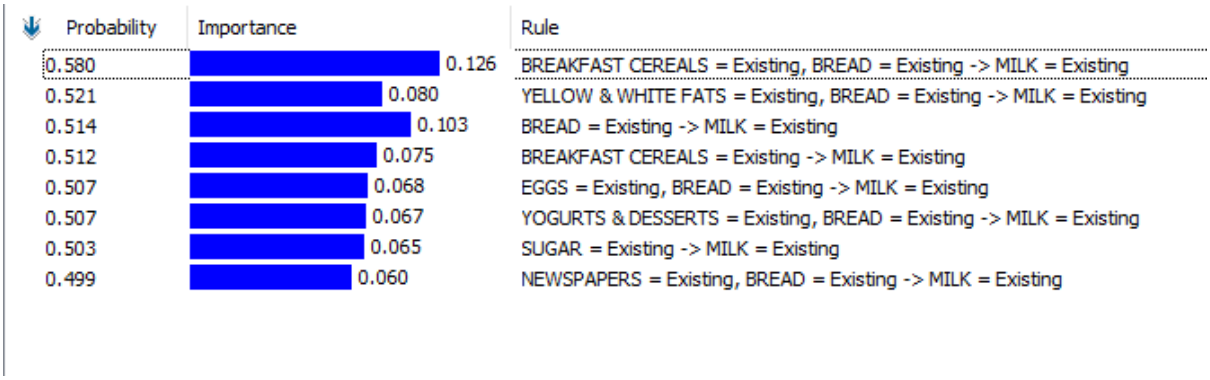

Figure 4 : Association rule Mining with Apriori Algorithm

4.3.2 Association rule mining of clusters in Rstudio

4.3.2.1 RStudio

For the objective to enhance the cluster analysis of customer visit segments by using association, RStudio is used as a platform to perform association rules on each of the cluster formed for exploring the role of association rules in customer visit segmentation.

4.3.2.2 Eclat Algorithm

Eclat algorithm denoted by Equivalence Class Clustering and bottom-up Lattice Traversal Algorithm. It is considered among the popular algorithms used for association rule mining. This algorithm works in a vertical manner (Depth first) and apriori algorithm is which works in a horizontal manner (Breadth first) which makes Eclat faster than apriori algorithm.

Eclat algorithm also has some more advantages over apriori algorithm such as lesser memory requirement, lesser number of scans for computing and better speed being one of them.

Cluster data was further manipulated to create a format accepted by Rstudio to perform association rule with apriori algorithm. The sparse matrix was converted into data in basket form describing each product in basket. This data was then modelled with Eclat algorithm to give further insights of each cluster formed.
The Eclat algorithm gives Item sets brought together than the actual association rule making the association simpler to understand.

```
> inspect(sort(rules, by = 'support') [1:10])
     items                                                  support   transIdenticalToItemsets count
[1]  {CRISPS/SNACKS & NUTS,SOFT DRINKS}                     0.3767262 31590                     31590
[2]  {CONFECTIONERY,SOFT DRINKS}                            0.3483793 29213                     29213
[3]  {MILK,SOFT DRINKS}                                     0.2854366 23935                     23935
[4]  {SANDWICHES & SNACKS,SOFT DRINKS}                      0.2587116 21694                     21694
[5]  {BREAD,SOFT DRINKS}                                    0.1901996 15949                     15949
[6]  {CRISPS/SNACKS & NUTS,SANDWICHES & SNACKS,SOFT DRINKS} 0.1638801 13742                     13742
[7]  {CRISPS/SNACKS & NUTS,SANDWICHES & SNACKS}             0.1638801 13742                     13742
[8]  {CONFECTIONERY,CRISPS/SNACKS & NUTS,SOFT DRINKS}       0.1592411 13353                     13353
[9]  {CONFECTIONERY,CRISPS/SNACKS & NUTS}                   0.1592411 13353                     13353
[10] {BAKERY- INSTORE/CDF,SOFT DRINKS}                      0.1460157 12244                     12244
>
```

Figure 5 : Association rule mining with Eclat Algorithm

4.4 Analysis of the segments along with the association rules retrieved.

The clusters achieved from the K-means clustering are then labelled for a visit motive according to the Items present in the cases. The association rules formed in each cluster visit would be then analysed to know the frequent items bought in each visit.

5 Evaluation

Evaluation for this research is given in terms of technical as well as business understanding. As the base of the research is business intelligence, it has focused on business perspective contribution of the research.
For the evaluation of Cluster segments formed, standard elbow method is implemented to evaluate the inter cluster distance as well as the sum of square error (SSE) values of set of clusters.
Apriori and Eclat algorithms are evaluated based on series of minimum support values such as 0.01%,0.1%,0.5%,1%,5% and 10%. Results of these support values are compared for the following metrics as per referred from literature review:
1. Number of Rules Generated
2. Algorithm performance according to time consumption.

5.1 Evaluation of Visit Segmentation using Clustering.

The Microsoft Clustering is performed for achieving customer visit segmentation as a state of the art.
As per the elbow method, optimal number of clusters defined were K=6 as per the figure 6, Researcher (Griva et al., 2018) had complications in stating the accurate results of elbow method in case of finding the exact elbow for the standard elbow method. In the previous

research where researcher had to test the K values from 6 to 10 to check the most accurate clusters, in this research the elbow was accurately found and implemented.

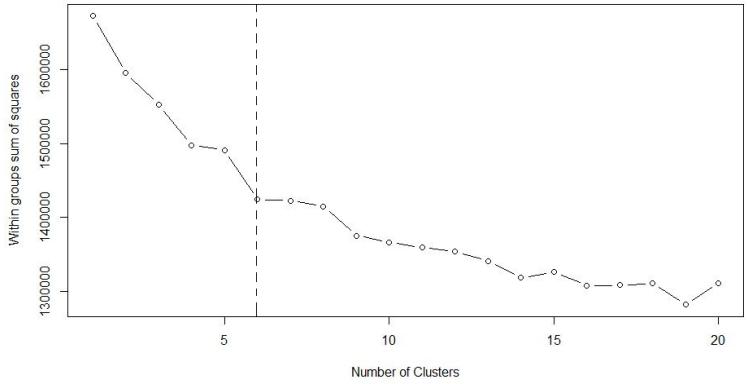

Figure 6 : Elbow Method

The CLUSTER_COUNT was set as 6 in Microsoft Clustering algorithm settings. With the help of sum of square analysis using elbow method, 6 optimal clusters were achieved.

The Clusters retrieved for the segmentation of customer visit motive are concluded as below:

- Cluster 1 can be defined as the **'visit for lunch/dinner'** as the main items bought from the visit are salad, vegetables and chilled produce.

- Cluster 2 can be defined as the **'visit for snacks'** as one of the key items bought in this visit are confectionery, pies and biscuits.

- Cluster 3 had baskets with various types of products which could not conclude any specific shopping motive for the visit hence can be given as '**abstract visit'.**

- Cluster 4 can be defined as '**visit for breakfast'** as most of the baskets included the products bread, cooked meat, morning goods and frozen vegetables.

- Cluster 5 is observed as the motive of '**party visit'** as most baskets had the products with drinks, soft drinks and snacks.

- Cluster 6 can be named as '**Mixed food visit'** as the products observed in the segment are various types of only food products including pet food as well.

Customer visit segmentation has shown accurately formed clusters with the help of exact K value in elbow method achieved.
As the segments achieved from this experiment gives only a basic idea about a visit, It would be necessary to analyse these clusters for further information.

5.2 Evaluation of Microsoft Association Rule Mining (Apriori Algorithm).

Association Rule mining has assisted in further analysis of the cluster segments formed with customer visit segmentation.

Implementation of apriori Algorithm using Microsoft Association rule mining in visual studio had following parameters.

Parameter	Value	Percentage
Min. Support	0.05	5%
Min. Probability (Confidence)	0.50	50%
Minimum Length	2 (Items per rule)	

Table 1 : Parameters for Apriori Algorithm

The following important rules were retrieved for analysis of each cluster representing a customer visit in table 2.

Customer Visit	Interpretation of The Rule
visit for lunch/dinner	If the Customer buys product from [Breakfast Cereal] and [Bread] category, most likely he would also buy product from [Milk] category.
visit for snacks	If the customer buys product of [sandwich and snacks] and [confectionery] category, then most likely he would also buy product from [Crisp/Snacks and Nuts] Category
abstract visit	If the Customer buys product from [Chilled prepared produce] and [Milk] category, most likely he would also buy product from [Vegetable] category.
visit for breakfast	If the Customer buys product from [Potatoes] and [Fruit Tropical] category, most likely he would also buy product from [Vegetable] category.
party visit	If the Customer buys product from [Hot Beverages] category, most likely he would also buy product from [Confectionery] category.
Mixed food visit	If the Customer buys product from [Cheese] and [Soft Drinks] category, most likely he would also buy product from [Cooked Meat] category.

Table 2 : Association Rules for Apriori Algorithm

All the above rules are having 50% of likelihood of happening.
In the research (Griva et al., 2018), The rules generated with the support value as 1% and the confidence values 60% on an average. In this research the rules were generated with lesser confidence (importance) value explaining the values of rules with lesser chances of being true but 50% confidence level gives enough importance for the rules to be used in practical application.
Though, the Analysis of each visit using Apriori Association Rules has shown a better study of the products bought during the visit, Considerable computational time is observed in the implementation. It would be interesting to study if another algorithm takes lesser time and gives better performance.

5.3 Evaluation of Association rule mining of clusters using Rstudio (Eclat Algorithm).

Eclat algorithm is performed to compare with apriori algorithm to achieve a faster performing association rule algorithm.
Eclat Algorithm was set for mining with parameters as stated in table 3

Parameter	Value	Percentage
Min. Support	0.05	5%
Minimum Length	2 (Items per rule)	

Table 3 : Parameters for Eclat Algorithm

Training of Eclat algorithm has shown groups of Items bought together for each visit of a customer in table 4.

Customer Visit	Item list
visit for lunch/dinner	Bread and Milk
visit for snacks	Crisps/Snacks & Nuts and Soft Drinks
abstract visit	Milk and Vegetables
visit for breakfast	Fruit - Tropical, Milk
party visit	Confectionery, Milk
Mixed food visit	Bread and Cooked Meat

Table 4 : Association Rules Generated with Eclat Algorithm

In this research Eclat Algorithm is tested on regards to the computational time consumption. As stated in the research (Griva et al., 2018), the Eclat algorithm took 30 milliseconds to process rules for 200 transactions whereas in this research we can see the number of transactions and the time taken for each visit segment as states in table 5.

Visit Segment	Number of Transactions	Computational Time (milliseconds)
visit for lunch/dinner	125531	13
visit for snacks	83853	41
abstract visit	54166	48
visit for breakfast	49557	8
party visit	49654	3
Mixed food visit	45583	3

Table 5 : Computational Time for Eclat Algorithm according to Visit Segments

Eclat Algorithm here stated the simple rules with minimal parameters indicating the products which are generally brought together in each visit thus, taking lesser computational time than Apriori algorithm.

6 Discussion

The experiment of customer visit segmentation which is a state of the art has shown basic visit segmentation which can be identified as a general visit motive, but the results does not give a confirmation about one motive only. The clustered segments include products of various variety of categories disturbing the judgment of the reason behind visit the shop. This

experiment completely depends on the dataset and the type of products bought and cannot guarantee the perfectly divided visit motive by clustering.
The experiment was followed by the correct number of K Clusters in K-means algorithm, Cluster segments did not find perfect visit motive.
Also, the fact that some of the products cannot confirm any motive and which are bough generally can affect the concept behind deciding the visit motive.

Although, if we consider the basic visit segments for further analysis, Association rule mining has shown excellent rules which can help in finding the products bought in each basket for the visit motive in a time frame.
Following are the number of rules generated for series of minimum support values and the total rules generated.

Minimum Support	Number of Rules Generated					
	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
0.01%	58875	114507	754923	7484551	141191	132382
0.10%	4475	7060	21321	83065	7368	31750
0.50%	646	980	2196	5433	1015	2999
1%	229	398	749	1692	401	1034
5%	8	41	48	110	37	75
10%	1	13	14	28	10	17

Table 6 : Number of Rules Generated in Association rule with range of support

Both Apriori and Eclat algorithm showed same number of rules while showing some performance differences in terms of time consumption as per stated in figure 7.

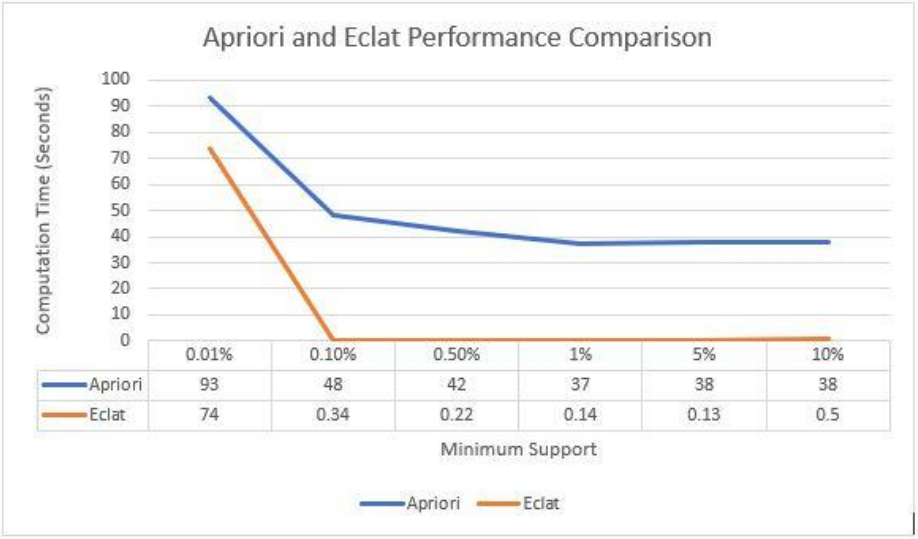

Figure 7 : Performance Comparison of Apriori and Eclat Algorithm.

According to the figure Even if the Number of rules generated are equal, the different algorithm has shown the huge difference between the time taken to compute the rules.
This difference can also be justified by the difference in the platforms of both algorithm (RStudio and Visual Studio). Even if it is seen in the research that Eclat algorithm takes more time to compute then apriori algorithm, it is also backed by the result that it generates lesser

rules and this case do not give enough evidence to be true in this experiment as the number of rules are equal.

This research can be taken ahead in future to compare the different association rules on same platform to compare the results with enough evidence.
It can also be studies if the customer segments show better results with other clustering techniques such as EM technique in Microsoft clustering.
With the future improvements and automation, customer visit segmentation can help in many customer analysis software for Retail shops.

7 Conclusion and Future Work

This research has objective of improving the clusters with the help of elbow method, where the results were accurately achieved with a K value and the clusters were achieved in the K-means algorithm accurately as compared to the previous research.
Research aim of modelling association rules with visit segments resulted into the meaningful contribution to the customer visit segmentation as the rules gave the further insights into the segments with rule having acceptable importance.
Implementing eclat algorithm has shown the frequent item sets instead of rules which were simple to understand with the good support value.
Performance comparison of apriori and eclat algorithm concluded the faster performing eclat model with the same number of item sets as apriori algorithm which is a highly contributing factor of this research.
While the eclat algorithm is proven to be faster than the apriori due to horizontal processing in the studied previous research, in this research it cannot be justified completely considering the difference in the platform for the implementation. It can be used as a future work to compare the performance on the same implementing platform. It would be interesting to compare the both algorithms on RStudio as the platform with R language as the Microsoft association rules has limitation of only apriori algorithm which lead to the limitation for this research.
This research has considerably added the value to the customer visit segmentation model by adding association rules, with the attempt of further improving the model performance it would be beneficial in knowing the customer visit motives with better analysis for retail businesses.

References

Agarwal, R., 2017. Decision making with association rule mining and clustering in supply chains. Int. J. Data Netw. Sci. 11–18. https://doi.org/10.5267/j.ijdns.2017.1.003

Amity University Jaipur Hod (Cse), Amity University Jaipur, Jain, T., 2016. Quantitative Analysis of Apriori and Eclat Algorithm for Association Rule Mining. Int. J. Eng. Comput. Sci. https://doi.org/10.18535/ijecs/v4i10.18

Balaji, S., Srivatsa, D.S.K., Scholoar, R., 2012. Customer Segmentation for Decision Support using Clustering and Association Rule based approaches. Eng. Technol. 3, 5.

Borgelt, C., n.d. Efficient Implementations of Apriori and Eclat 10.

Calvo-Porral, C., Lévy-Mangin, J.-P., 2018. From “foodies” to “cherry-pickers”: A clustered-based segmentation of specialty food retail customers. J. Retail. Consum. Serv. 43, 278–284. https://doi.org/10.1016/j.jretconser.2018.04.010

Changchien, S.W., Lu, T.-C., 2001. Mining association rules procedure to support on-line recommendation by customers and products fragmentation. Expert Syst. Appl. 20, 325–335. https://doi.org/10.1016/S0957-4174(01)00017-3

Griva, A., Bardaki, C., Pramatari, K., Papakiriakopoulos, D., 2018. Retail business analytics: Customer visit segmentation using market basket data. Expert Syst. Appl. 100, 1–16. https://doi.org/10.1016/j.eswa.2018.01.029

Kansal, T., Bahuguna, S., Singh, V., Choudhury, T., 2018. Customer Segmentation using K-means Clustering, in: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). Presented at the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), IEEE, Belgaum, India, pp. 135–139. https://doi.org/10.1109/CTEMS.2018.8769171

Kaur, M., Kang, S., 2016. Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. Procedia Comput. Sci. 85, 78–85. https://doi.org/10.1016/j.procs.2016.05.180

Khamis Mwero Manero, Rimiru, R., Calvins Otieno, 2018. Customer Behaviour Segmentation Among Mobile Service Providers In Kenya Using K-Means Algorithm. https://doi.org/10.5281/ZENODO.1467663

Kotiyal, B., Kumar, A., Pant, B., Goudar, R.H., Chauhan, S., Junee, S., 2013. User behavior analysis in web log through comparative study of Eclat and Apriori, in: 2013 7th International Conference on Intelligent Systems and Control (ISCO). Presented at the 2013 7th International Conference on Intelligent Systems and Control (ISCO), IEEE, Coimbatore, Tamil Nadu, India, pp. 421–426. https://doi.org/10.1109/ISCO.2013.6481192

Liao, S.-H., Chen, Y.-J., Yang, H.-W., 2013. MINING CUSTOMER KNOWLEDGE FOR CHANNEL AND PRODUCT SEGMENTATION. Appl. Artif. Intell. 27, 635–655. https://doi.org/10.1080/08839514.2013.813195

Miguéis, V.L., Camanho, A.S., Cunha, J.F. e, 2011. Mining Customer Loyalty Card Programs: The Improvement of Service Levels Enabled by Innovative Segmentation and Promotions Design, in: Snene, M., Ralyté, J., Morin, J.-H. (Eds.), Exploring Services Science, Lecture Notes in Business Information Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 83–97. https://doi.org/10.1007/978-3-642-21547-6_7

Quan, T.T., Ngo, L.N., Hui, S.C., 2009. An Effective Clustering-based Approach for Conceptual Association Rules Mining, in: 2009 IEEE-RIVF International Conference on Computing and Communication Technologies. Presented at the 2009 IEEE-RIVF International Conference on Computing and Communication Technologies, IEEE, Danang City, Viet Nam, pp. 1–7. https://doi.org/10.1109/RIVF.2009.5174619

Riaz, M., Arooj, A., Malik Tahir Hassan, Jeong-Bae Kim, 2014. Clustering based association rule mining on online stores for optimized cross product recommendation, in: The 2014

International Conference on Control, Automation and Information Sciences (ICCAIS 2014). Presented at the 2014 International Conference on Control, Automation and Information Sciences (ICCAIS), IEEE, Gwangju, South Korea, pp. 176–181. https://doi.org/10.1109/ICCAIS.2014.7020553

Robu, V., dos Santos, V.D., 2019. Mining Frequent Patterns in Data Using Apriori and Eclat: A Comparison of the Algorithm Performance and Association Rule Generation, in: 2019 6th International Conference on Systems and Informatics (ICSAI). Presented at the 2019 6th International Conference on Systems and Informatics (ICSAI), IEEE, Shanghai, China, pp. 1478–1481. https://doi.org/10.1109/ICSAI48974.2019.9010367

Silva, J., Angulo, M.G., Cabrera, D., Kamatkar, S.J., Caraballo, H.M., Ventura, J.M., Peña, J.A.V., de la Hoz – Hernandez, J., 2019. Association Rule Mining for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm, in: Singh, M., Gupta, P.K.,

Tyagi, V., Flusser, J., Ören, T., Kashyap, R. (Eds.), Advances in Computing and Data Sciences, Communications in Computer and Information Science. Springer Singapore, Singapore, pp. 487–497. https://doi.org/10.1007/978-981-13-9942-8_46

Suhail Najam, S., Hashim AL-Saedi, K., 2018. Spam classification by using association rule algorithm based on segmentation. Int. J. Eng. Technol. 7, 2760. https://doi.org/10.14419/ijet.v7i4.18486

Prasanna, S., and D. Ezhilmaran. "Association Rule Mining Using Enhanced Apriori with Modified GA for Stock Prediction." *International Journal of Data Mining, Modelling and Management* 8, no. 2 (2016): 195. https://doi.org/10.1504/IJDMMM.2016.077162.

Ahn, Kwang-Il. "Effective Product Assignment Based on Association Rule Mining in Retail." *Expert Systems with Applications* 39, no. 16 (November 2012): 12551–56. https://doi.org/10.1016/j.eswa.2012.04.086.

Ballestar, María Teresa, Pilar Grau-Carles, and Jorge Sainz. "Customer Segmentation in E-Commerce: Applications to the Cashback Business Model." *Journal of Business Research* 88 (July 2018): 407–14. https://doi.org/10.1016/j.jbusres.2017.11.047.

Gaikwad, Pooja R., Shailesh D. Kamble, Nileshsingh V. Thakur, and Akshay S. Patharkar. "Evaluation of Apriori Algorithm on Retail Market Transactional Database to Get Frequent Itemsets," 187–92, 2017. https://doi.org/10.15439/2017R83.

Heaton, Jeff. "Comparing Dataset Characteristics That Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms." In *SoutheastCon 2016*, 1–7. Norfolk, VA, USA: IEEE, 2016. https://doi.org/10.1109/SECON.2016.7506659.

Maryani, Ina, Dwiza Riana, Rachmawati Darma Astuti, Ahmad Ishaq, Sutrisno, and Eva Argarini Pratama. "Customer Segmentation Based on RFM Model and Clustering Techniques With K-Means Algorithm." In *2018 Third International Conference on Informatics and Computing (ICIC)*, 1–6. Palembang, Indonesia: IEEE, 2018. https://doi.org/10.1109/IAC.2018.8780570.

Shah, Ashish. “Association Rule Mining with Modified Apriori Algorithm Using Top down Approach.” In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT)*, 747–52. Bangalore, India: IEEE, 2016. https://doi.org/10.1109/ICATCCT.2016.7912099.

Sinthuja, M, P Aruna, and N Puviarasan. “EXPERIMENTAL EVALUATION OF APRIORI AND EQUIVALENCE CLASS CLUSTERING AND BOTTOM UP LATTICE TRAVERSAL (ECLAT) ALGORITHMS” 13 (2016): 6.

Venkatachari, Kavitha, and Issac Davanbu Chandrasekaran. “MARKET BASKET ANALYSIS USING FP GROWTH AND APRIORI ALGORITHM: A CASE STUDY OF MUMBAI RETAIL STORE” 8 (2016): 9.

Wen-xiu, Xie, Qi Heng-nian, and Huang Mei-li. “Market Basket Analysis Based on Text Segmentation and Association Rule Mining.” In *2010 First International Conference on Networking and Distributed Computing*, 309–13. Hangzhou, Zhejiang, China: IEEE, 2010. https://doi.org/10.1109/ICNDC.2010.67.

Research Project
Vishakha Kale
x18181643

Q1. What would be the main reason to use Eclat algorithm?

The primary objective behind using another association rule algorithm was to experiment the improvement of the performance of customer visit analysis.
It is studied from some research (Robu and dos Santos, 2019) (Kotiyal *et al.*, 2013) (Borgelt, no date) (Griva *et al.*, 2018) (Jain, 2016) that Eclat algorithm has performed faster than apriori and this theory motivated to experiment the eclat algorithm for customer visit analysis to compare the performance with apriori algorithm.

Q2. What percent of the dataset is used for training of Eclat algorithm, and why?

The entire dataset is used for customer visit segmentation which was further divided into visit segments and the data involved in each visit segment was used to train eclat algorithm for each visit.
The percent of data according to visit segments can be approximately given in Table1

Visit Segment	% of Total Dataset
visit for lunch/dinner	30.7%
visit for snacks	20.53%
abstract visit	13.2%
visit for breakfast	12%
party visit	12%
Mixed food visit	11%

Table 1: Percentage of Dataset for each visit.

Q3. What is a novelty in this work?

Customer visit segmentation is a novel concept given by (Griva *et al.*, 2018) which was the base motivation behind this research. This work contributes towards an extension to the research presented by (Griva *et al.*, 2018) giving a novel experiment of analysis of visit segments with association rule mining and experiment of the performance improvement as well.

Q4. Why RStudio and Visual Studio have huge difference between the time taken to compute the rules? Give more justifications.

The primary reason behind the time difference was given by the difference in the algorithms, but the entire experiment could not be compared on the fact that algorithm was different, as the modelling platforms were also different in this case.
It was not the motive of the research to compare the performance based on platforms and also there are no studies found to back up the theory that visual studio performs faster than RStudio hence it was only assumed that difference in the algorithm performance could be the factor. As RStudio shown faster performance, future work can be considered to experiment

more algorithms on RStudio to conclude if Visual studio was behind the slower performance when compared with eclat algorithm as per stated in the research report.

Q5. What is a rationality to have to set the cluster count as 6?

The cluster count 6 was concluded from the graph of elbow method when applied the method on the dataset.

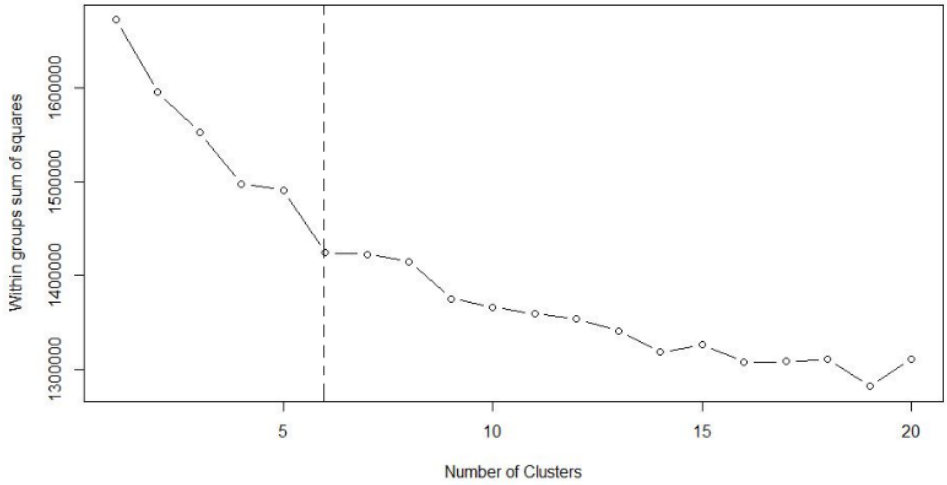

Figure 1: Elbow Method

Elbow is the point of the graph where SSE (Sum of Square) value drops considerably and then it does not reduce with a large measure or almost stays stable.
As the cluster count keeps on increasing, the SSE value keeps decreasing, but the elbow point is considered as the cluster count which would give accurate clusters with acceptable SSE value.
This elbow point is observed at 6 as per given in figure 1 hence the cluster count was rationalised to be 6.

Q6. what is the actual computational time and number of transactions for each visit segment for the apriori algorithm in section 5.3

The Transactions for the apriori algorithm were also same as per the eclat algorithm as given in section 5.3.
The actual computational time is as given in table 2.

Visit Segment	Number of Transactions	Computational Time (Minutes)
visit for lunch/dinner	125531	1.78
visit for snacks	83853	1.65
abstract visit	54166	1.62
visit for breakfast	49557	1.54
party visit	49654	1.54
Mixed food visit	45583	1.47

Table 2: Apriori Algorithm Computational Time

The minimum support value was set as 0.05(5%) for these rules generations as well.

References

Borgelt, C. (no date) ‘Efficient Implementations of Apriori and Eclat’, p. 10.

Griva, A. *et al.* (2018) ‘Retail business analytics: Customer visit segmentation using market basket data’, *Expert Systems with Applications*, 100, pp. 1–16. doi: 10.1016/j.eswa.2018.01.029.

Jain, T. (2016) ‘Quantitative Analysis of Apriori and Eclat Algorithm for Association Rule Mining’, *International Journal Of Engineering And Computer Science*. doi: 10.18535/ijecs/v4i10.18.

Kotiyal, B. *et al.* (2013) ‘User behavior analysis in web log through comparative study of Eclat and Apriori’, in *2013 7th International Conference on Intelligent Systems and Control (ISCO). 2013 7th International Conference on Intelligent Systems and Control (ISCO)*, Coimbatore, Tamil Nadu, India: IEEE, pp. 421–426. doi: 10.1109/ISCO.2013.6481192.

Robu, V. and dos Santos, V. D. (2019) ‘Mining Frequent Patterns in Data Using Apriori and Eclat: A Comparison of the Algorithm Performance and Association Rule Generation’, in *2019 6th International Conference on Systems and Informatics (ICSAI). 2019 6th International Conference on Systems and Informatics (ICSAI)*, Shanghai, China: IEEE, pp. 1478–1481. doi: 10.1109/ICSAI48974.2019.9010367.