A COMPARATIVE STUDY OF CLUSTERING ALGORITHMS AND AN IMPLEMENTATION OF THE APRIORI ALGORITHM IN A RETAIL COMPANY

(Case Study: Product Sales Data)

THESIS

Submitted to International Program Industrial Engineering
in Partial Fulfillment of the Requirements for the degree of
Bachelor of Industrial Engineering at Universitas Islam Indonesia

Arranged by:
Muhammad Rakhmat Setiawan
14522440

INTERNATIONAL PROGRAM
DEPARTMENT OF INDUSTRIAL ENGINEERING
FACULTY OF INDUSTRIAL TECHNOLOGY
UNIVERSITAS ISLAM INDONESIA
YOGYAKARTA
2018

AUTHENTICITY STATEMENT

In the name of Allah SWT, I hereby certify that this research conduct by my creation, unless the citation in which each of those are already mentioned the source and rewrite by myself. If someday this declaration letter is proved plagiarism, Universitas Islam Indonesia has a right to revoke to its confession.

Yogyakarta, 23rd July 2018

METERAI TEMPEL
TGL.
BEF94AFF067376521
6000
ENAM RIBU RUPIAH

Muhammad Rakhmat Setiawan

THESIS APPROVAL OF SUPERVISOR

A COMPARATIVE STUDY OF CLUSTERING ALGORITHMS AND AN IMPLEMENTATION OF THE APRIORI ALGORITHM IN A RETAIL COMPANY

(Case Study: Product Sales Data)

THESIS

Arranged By

Name :Muhammad Rakhmat Setiawan

Student Number :14522440

Yogyakarta, 2nd July 2018

Supervisor

(Dr. techn. Rohmatul Fajriyah)

THESIS APPROVAL OF EXAMINATION COMMITTEE
A COMPARATIVE STUDY OF CLUSTERING ALGORITHMS AND AN IMPLEMENTATION OF THE APRIORI ALGORITHM IN A RETAIL COMPANY

(Case Study: Product Sales Data)

Arranged By

Name	: Muhammad Rakhmat Setiawan
Student Number	:14522440

Was defended before Examination Committee in Partial Fulfillment of the Requirements for the bachelor degree of Industrial Engineering Department

Universitas Islam Indonesia

Examination Committee

Muhammad Ridwan AP, ST., M.Sc., Ph.D

Examination Committee Chair

Dr. techn. Rohmatul Fajriyah

Member I

Agus Mansur, ST., M.Eng.Sc

Member II

Acknowledged by,

Head of International Program

Department of Industrial Engineering

Faculty of Industrial Technology

Universitas Islam Indonesia

(Muhammad Ridwan Andi Purnomo, ST., M.Sc., Ph.D)

DEDICATION

I dedicate my final thesis project for my beloved parents, brothers, and my friends.

I would thank my parents for the support, taking care and anything they gave to me.

I also thank my brothers and friends for the support and motivation.

MOTTO

Fighting has been enjoined upon you while it is hateful to you. But perhaps you hate a thing and it is good for you; and perhaps you love a thing and it is bad for you. And Allah Knows, while you know not. (QS: Al-Baqara 2:216)

Allah does not charge a soul except [with that within] its capacity. It will have [the consequence of] what [good] it has gained, and it will bear [the consequence of] what [evil] it has earned. "Our Lord, do not impose blame upon us if we have forgotten or erred. Our Lord, and lay not upon us a burden like that which You laid upon those before us. Our Lord, and burden us not with that which we have no ability to bear. And pardon us; and forgive us; and have mercy upon us. You are our protector, so give us victory over the disbelieving people. (QS: Al-Baqara 2:286)

So which of the favors of your Lord would you deny? (QS: Ar-Rahman 55:61)

Education is the most powerful weapon which you can use to change the world (Nelson Mandela)

Only they dare to fail that can get success (John F. Kennedy)

Be strong to be useful (Georges Hebert)

PREFACE

Assalamu'alaikum Warahmatullaahi Wabarakatuh

Began with praising to Allah SWT for the gracious mercy and tremendous blessing. Then also Prophet Muhammad SAW who brings people from dark ages into age of enlightenment because of that makes me able to complete the research report entitled "A Comparative Study of Clustering Algorithms and an Implementation of the Apriori Algorithm in a Retail Company (Case Study: Product Sales Data)" timely as the final project that the author already finishes the undergraduate thesis in International Program Department of Industrial Engineering Universitas Islam Indonesia.

The writer did the research project at Pamella Satu Supermarket in Yogyakarta. During the research project and the process of writing this report, the writer cannot finish all the work if there is no help from Allah SWT and all the people that support the writer. The writer wants to say thank you to all the people that are always giving their support, i.e.:

1. Dr. Drs. Imam Djati Widodo, M.Eng.Sc. as the Dean of the Faculty of Industrial Technology of *Universitas Islam Indonesia* in the period 2014 to 2018.
2. Yuli Agusti Rochman, S.T., M.Eng. as the Head of Department Industrial Engineering of FIT UII in the period 2014 to 2018.
3. Muhammad Ridwan Andi Purnomo, S.T., M.Sc., Ph.D. as the Head of International Program of Department Industrial Engineering of FIT UII.
4. My supervisor in this research, Dr. techn. Rohmatul Fariyah for her valuable assistance and guidance to the completion of this report
5. Mr. And Mrs. Sunardi Syahuri as the business owner of Pamella Supermarket Yogyakarta.
6. My beloved parents and brothers who are always giving their prayer, love, and support.
7. My partners in ERP FTI UII Laboratory who still support me to finish this report.
8. My friends in Study of Bioinformatics Research Center who always encourage me to complete this report and lent me a decent laptop for significant data processing.

9. My Family in International Program of Industrial Engineering UII Batch 2014 friends who also support me to finish this report
10. My friends in the International Buddy UII period of 2014 until 2016 who have entrusted me to be a leader for a period and gave me a lot of experience in the international buddy program.
11. My colleague in the Marketing and Communication International Program UII in 2015 until 2017 who has taught me how to work professionally and provides a network and excellent negotiation experience with internal and external stakeholders.
12. My friends in KOSEMA *(Koordinator Seluruh Mahasiswa Angkatan)* batch 2014 of Industrial Engineering Department UII who has been supporting my university life from the beginning until my graduation.
13. My friends while on a student exchange program at Chulalongkorn University who have given me many experiences on how to survive in Thailand for one semester in the beginning of 2017.
14. Office of International Affairs UII that has selected me as a Best Student at Bridging Program and provided me the opportunity to join the 2016 Korean Summer Program in South Korea.
15. For all the people that cannot be mentioned here, may Allah bless you.

The author realized that this report is full of imperfection and employ many shortcomings. Therefore the author is looking forward to the reader’s constructive criticism and suggestions for the sake of improving this report. Also, the author hopes that this research may be beneficial for all parties and hopefully all of what has been done will be rewarded by Allah SWT.

Wassalaamu’alaikum Warahmatullaahi Wabarakatuh

Yogyakarta, 23rd July 2018

Author

Muhammad Rakhmat Setiawan

ABSTRACT

At the retail company, several problems often arise from the sale of the products. The retail company found difficulty to perform the correct promotion strategy for improving their product sales. The information about the customers' segmentation will be beneficial to determine the right business strategy for the retail company. In the study, a detailed discussion on each of three algorithms, i.e., Hierarchical, K Means and Fuzzy C Means Clustering Algorithms and a comparative study between them already investigated. Other than that, the implementation of Association Rules Market Basket Analysis (ARMBA) by using Apriori Algorithm is applied to uncover the associations between transactions. After conducting 100 simulations, the researcher found that Fuzzy C Means is the best clustering algorithm regarding the measure of goodness with the BSS/TSS ratio of 59,34% and average processing time equal to 63,23 seconds. Otherwise, K Means Algorithm has BSS/TSS ratio of 56,7% and average processing time corresponding to 93.83 seconds. There are two customer clusters in which Customer Cluster A and Customer Cluster B with different emphasize to be considered by the business owner. The characteristics of each cluster with these two approaches are the Customers Cluster A tend to buy adult products, and the Customers Cluster B tend to buy baby products. The researcher found that the most promising customers are in Customer Cluster A regarding the comparison of each variable on clustering algorithms result. ARMBA Apriori Algorithm with minimum confidence 80% and support 0,1% has generated 23 rules. Based on these results, therefore, the researcher proposed to the business owner to focus on customer cluster A and provide ARMBA results using Apriori Algorithm so that later is expected to be the basis to formulate business strategy by the business owner.

Keywords: *Data Mining, Hierarchical Clustering, K-means Clustering, Fuzzy C-means clustering, Association Rules, Apriori Algorithm*

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1. Background

Data mining is genuinely an interdisciplinary topic that can be defined in many different ways. In the field of database management industry, data analysis is mainly evolved with several large data repositories. The result yields to the process of data mining. There are several data mining functionalities used to specify the kinds of patterns to be found in the data mining task. These functionalities include characterizations and discrimination, the mining of frequent patterns, associations and correlations, classification regression, clustering analysis and outlier analysis (Han, et al., 2012). Data Mining is an instrumental technique in helping companies to find information that is very important for their data warehouse. The availability of extensive data and the need for information or knowledge as decision support for making business solutions and infrastructure support in the field of informatics engineering is the forerunner of the birth of data mining technique. The information will be used in the decision-making solution in the business environment, for business development itself.

According to Jain, et al. (1999), Data clustering recognized as an essential area of data mining. Data clustering is the process of dividing data elements into different groups as known clusters in such a way that the elements within a group possess high similarity while they differ from the ingredients in a diverse group. Clustering is one of the most exciting and essential topics in data mining that aims to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. The basic concept of cluster analysis is partitioning a set of data objects or observations into subsets. Each subset is unique such that objects in one cluster are similar to one another, yet dissimilar to objects in another cluster. The different cluster may be formed using the same data set by applying different clustering methods (Wu, et al., 2007).

In general, data can be clustered based on the similarity of theme, object or research method. The results of the clustering can show the pattern of similarity or cluster homogeneous from time to time. Clustering analysis results can show us the trends when customers buy the same product and when customers buy different products at the same time. The results of clustering can also display the most and least taken products by customers. Such information is expected to assist the business owner in evaluating the number of product sales at a given time to increase product sales in the future. Clustering analysis depends on the two conditions. First, if there are some formulated strategies for improving product sales, i.e., five strategies, then the data must be clustered into 5 clusters. Secondly, if there are still no strategies formulated for increasing sales, then it will be better for the researcher to cluster the data based on the homogeneity. It means that the number of a cluster will be homogeneous.

Clustering is the process of the grouping of objects based on information obtained from data that explains the relationship between objects with principles to maximize the similarity between members of one cluster and minimize the similarity between classes or groups (Rahmawati, et al., 2016). One of the clustering methods is K-Means. According to Rahmawati, et al. (2016), K-Means can group data in large numbers with relatively fast and efficient computing time. However, the clustering results with K-Means are heavily dependent on the initial cluster. The clustering result with the K-Means method is useful if the initial cluster is right. The Hierarchical Clustering method can be used to overcome the problem of an initial cluster on K-Means algorithm. In this thesis project, the author will combine the K-Means algorithm with Hierarchical Clustering to test the previous research hypothesis is correct or not regarding computation time. The result from Hierarchical Clustering will be used in the initial cluster on K-means clustering. The combination of Hierarchical Clustering and K-Means Clustering methods has been tested by Rahmawati, et al. (2016) and it proves that the algorithm combination is better than just using the K-Means algorithm regarding the clustering result.

Other than that, Fuzzy C Means algorithm is a conventional soft clustering technique, and it has recently been used by researchers for a variety of clustering analyzes. Jim Bezdek first introduced Fuzzy C Means Algorithm in 1981 as an improvement on earlier clustering method. The main advantage of Fuzzy C Means algorithm is that it allows regular membership of data points to cluster measured as degrees in [0,1]. FCM gives the flexibility to express that data points can belong to more than one cluster. According to Tanjung (2016), Fuzzy C-Means is a data clustering technique where the degree of membership determines the existence of each data point in a cluster. He concluded that the clustering analysis could produce output that corresponds to the expected user and the FCM algorithm can segment the data set based on the numeric type of data.

Based on research conducted by Rahmawati, et al. (2016), the researcher would like to prove the quality of the clustering result is also suitable if used for grouping product sales data. Other than that, as has been shown by Tanjung (2016), the researcher also would like to implement Fuzzy C Means Algorithm on product sales data and find out what algorithm is best between the combination of Hierarchical plus K-Means and the Fuzzy C Means Algorithms to process product sales data. In the study, to define the best clustering algorithm is using processing time and Between Sum of Square parameters. BSS value is a measure of the goodness of the classification clustering algorithms has found. Moreover, the researcher will also implement Association Rule Market Basket Analysis using Apriori Algorithm to gain valuable insight from the number of product sales data obtained from one of the retail company in Yogyakarta. According to Astika (2015), she found that Apriori Algorithm can process transaction data to find frequent item-set and association rule that meets transaction limit and able to display regulations in text form. She found that the smaller the transaction limit and the minimum confidence determined, the more rules generated, with the consequence of processing time will be longer than the transaction limit, and the minimum confidence is higher. Therefore, the researcher is very interested in implementing the Apriori Algorithm to product sales data obtained.

In the study, first of all, a detailed discussion on each of these three algorithms is presented. After that, a comparative study of combined Hierarchical plus K-Means and the Fuzzy C-Means Algorithms is done experimentally. The researcher conducts comparative research to contribute to the development of knowledge and find out which clustering algorithm is right for use in customer segmentation analysis, and then later the result can be the basis for the business owner to formulate business strategy in the retail company. By doing a comparative study of combined Hierarchical plus K-Means and the Fuzzy C-Means Algorithms, the researcher will know what clustering algorithm is the best for analyzing product sales data in a retail company. With this research, the researcher also proposed Association Rule Market Basket Analysis by using Apriori Algorithm for discovering interesting relations between items are associated with each other in large databases. The aims to implement this method is to propose the sales tactics using customer transactions already available in the retail company database to the business owner.

1.2. Problem Formulation

The research would critically analyze and conduct a comparative study of combined Hierarchical plus K-Means and the Fuzzy C-Means Algorithms. The researcher conducts comparative research to contribute to the development of knowledge and find out which clustering algorithm is right for use in customer segmentation analysis, and then later the result can be the basis for the business owner to formulate business strategy in the retail company. Other than that, the researcher also proposed Association Rule Market Basket Analysis by using Apriori Algorithm for discovering interesting relations between items are associated with each other in large databases. The aims to implement this method is to propose the sales tactics using customer transactions already available in the retail company database. Thus, the researcher will provide several insights to the business owner regarding the best choice of cluster analysis method and the result of AR MBA by using Apriori Algorithm. Related to the purpose, central research questions addressed in this thesis are:

1. What is the best clustering algorithm between Combined Hierarchical plus K Means and Fuzzy C Means Algorithms to analyze the product sales data?

2. What is the best customer cluster to perform correct promotion strategy for the retail company?
3. What are the results of Association Rules Market Basket Analysis with Apriori Algorithm?
4. What are the shop layout to be proposed based on Clustering and Association Rules Market Basket Analysis results toward the retail company?

1.3. Research Objectives

Based on the problem formulation above, this research is created to fulfill several objectives as mentioned below:

1. Able to choose the best clustering algorithm by performing a comparative study of combined Hierarchical plus K Means and Fuzzy C Means Algorithms toward product sales data.
2. Able to choose the best customer cluster as the basis to formulate the marketing strategy and to support a correct promotion strategy for the retail company.
3. Able to explain the results of Association Rules Market Basket Analysis with Apriori Algorithm to be the basis to formulate the sales strategy in the retail company.
4. Able to propose shop layout based on Clustering and Association Rules Market Basket Analysis with Apriori Algorithm results to be the basis to formulate sales strategy in the retail company.

1.4. Research Limitation

To avoid widespread discussion, the researcher has several limitations, namely:

1. The author only limits the analysis of problems to a comparative study of combined Hierarchical plus K-Means and the Fuzzy C-Means Algorithms following the Knowledge Discovery in Database (KDD) stages as the basis to formulate correct promotion strategy at a retail company.
2. After conducting a comparison of both methods, the researcher will choose the best clustering algorithm and propose the basis to formulate a correct promotion strategy to the business owner.

3. The researcher also applied an Association Rule Market Basket Analysis by using Apriori Algorithm as the basis to formulate the sales strategy using customer transactions already available in the retail company database.
4. In the study, the researcher will use R software to assist the computation on this project.
5. The researcher will conduct data processing for clustering analysis approximately 20,464 rows of product sales data due to RAM limitation.
6. The researcher only includes indicators Sunday, Monday, Tuesday, Wednesday, Thursday and Friday due to ram limitations. Therefore, six days of sales transaction data is taken from 01 October to 06 October 2017.
7. The researcher will perform data processing for Association Rules Market Basket Analysis approximately 39,474 rows of product sales data.

1.5. Research Benefit

Along with the implementation of data mining in the case study. By conducting this research, several benefits can be earned as follows:

1. To provide knowledge previously hidden in the data warehouse, such that it becomes valuable information to increase the product sales in the retail companies by performing Clustering analysis and Association Rules Market Basket Analysis.
2. To understand the best clustering algorithm by completing a comparative study of combined Hierarchical plus K Means and Fuzzy C Means Algorithms toward product sales data.
3. This research could help the retail company to implement clustering algorithm to formulate correct promotion strategy.
4. This research could improve the retail company to apply Association Rules Market Basket Analysis to be the basis to formulate the right marketing and sales strategy using the available data in the local company database.
5. To contribute to the development of knowledge.

1.6. Systematical of Thesis Writing

Furthermore, this thesis writing will be continued as follows:

CHAPTER I INTRODUCTION

This chapter contains the background of the problem, the formulation of the problem, research limitation, research objective, research benefit and systematic writing.

CHAPTER II LITERATURE REVIEW

This chapter explains the literature studies. The review will be conducted in a systematic literature review, from the literature previous research and papers.

CHAPTER III RESEARCH METHODOLOGY

This chapter describes the steps for conducting the research. Those are applied as a reference to keep focusing on the primary goals, which are going to be archived. The study led will be emphasized on the analysis of numerical data, which aims to get a clear picture of a situation based on the data obtained by way of presenting, collecting and analyzing the data. Therefore, it becomes new information that can be used to solve the problems in the investigation.

CHAPTER IV DATA PROCESSING AND ANALYSIS

This chapter explains the processing analysis of the available data systematically. It will be demonstrating the step by step on the selection method and provide the results. A brief analysis would be performed in regards to the results.

CHAPTER V DISCUSSION

This chapter discusses the finding from the analysis based on the available data, the impact of them on the business along with the cited papers and works of literature.

CHAPTER VI CONCLUSION AND RECOMMENDATION

This chapter is a final section that describes the overall conclusions from the results of the study and the suggestion for the future research.

REFERENCES

APPENDIX

CHAPTER II

LITERATURE REVIEW

The literature review studies are divided into two, deductive and inductive. A deductive study will explain the underlying theory that has relation to research that would be conducted from the textbooks, etc. The inductive study will be based on the previous research that already has a reputation. Inductive and deductive studies need to be done to find out the gap between last review and the conducted research and to avoid plagiarism.

2.1. Deductive Literature

2.1.1 Data Mining

Data Mining is often called knowledge discovery in database (KDD). Data Mining is a process that uses statistical, mathematical, artificial intelligence and machine learning techniques to extract and identify useful information and related knowledge from large databases (Luthfi, 2009). Kamagi & Hansun (2014) define that data mining is an analysis of large amounts of data observations to find unknown relationships and two new methods to summarize data for easy understanding and usefulness for data selectors.

Data mining foresees trends and traits of business behaviors that are very useful for supporting critical decision making. Automated analysis performed by data mining exceeds that of traditional decision support systems that have been widely used. Data Mining can answer business questions that traditionally take a lot of time to respond. Data Mining explores databases to discover hidden patterns, searching for predictors of information that may be forgotten by business people as they lie beyond their expectations (Chandra, et al., 2014).

Data mining is not an entirely new field. One of the difficulties of defining data mining is the fact that data mining inherits many aspects and techniques from established areas of science. It aims to improve traditional methods such that can handle the following things:

1. The huge amount of data
2. The high dimension of data
3. The heterogeneous and different properties of data

2.1.2 Stages of Data Mining

As a series of processes, data mining can be divided into several stages illustrated in Figure 2.1. The steps are interactive, and the user is directly involved or with the intermediary of knowledge.

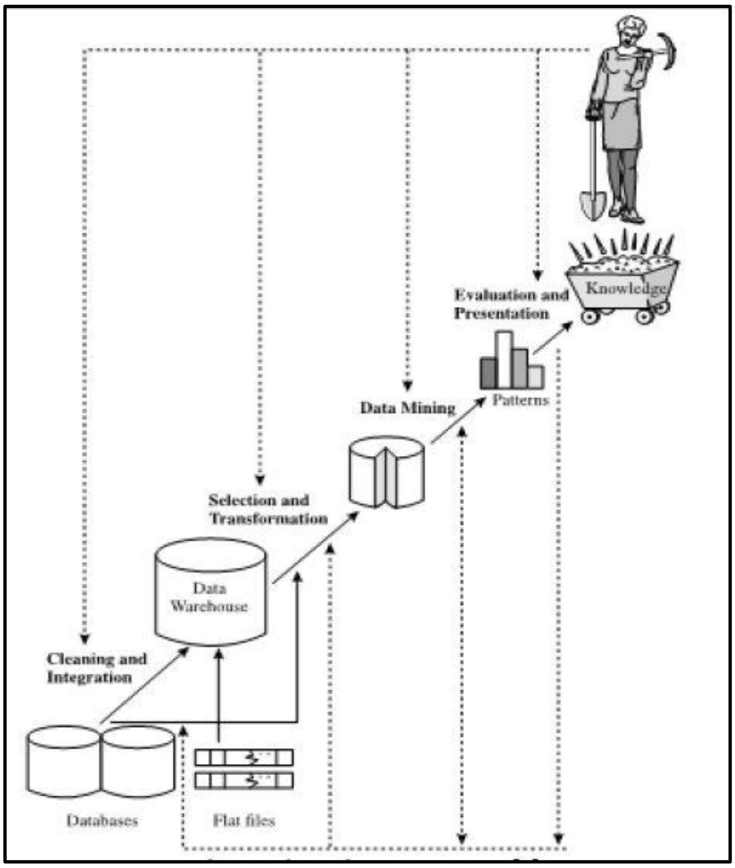

Figure 2.1 Stages of Knowledge Discovery in Databases.

There are six stages of data mining, namely:

1. Data Cleaning

   Data cleaning is a process of removing noise and inconsistent or irrelevant data (Ridwan, et al., 2013). In general, the data obtained, either from company databases or experimental results, have an incomplete data entry such as missing data, incorrect data or also typo data. Besides, there are also data attributes that are not relevant to the data mining hypothesis. Irrelevant data is even better discarded. Data cleaning will also affect the performance of data mining techniques because the data handled will decrease in number and its complexity.

2. Data Integration

   Data integration is a combination of data from various databases into a new database (Ridwan, et al., 2013). Infrequently if the data needed for data mining not only comes from one database but also comes from multiple databases or text files. Data integration is performed on attributes that identify unique entities such as name attributes, product types, customer numbers and so on. Data integration needs to be done carefully because errors in data integration can produce different and even misleading results. For instance, if the combination of data by product type turns out to combine products from different categories, then there will be a correlation between products that do not exist.

3. Data Selection

   Data that is in the database is often not all used, therefore only the appropriate data to be analyzed to be retrieved from the database (Ridwan, et al., 2013).

4. Data Transformation

   Data altered or merged into the appropriate format for processing in data mining (Ridwan, et al., 2013). Some data mining methods require unique data formats before they can be applied. For instance, some standard techniques such as association rules analysis and clustering can only accept categorical data input. Therefore, continuous numerical data needs to be divided into several intervals. This process is often called data transformation.

5. Mining Process

   It is a significant process when methods are applied to find a valuable and hidden knowledge of data. Some techniques that can be used based on the data mining grouping can be seen in Figure 2.2 below:

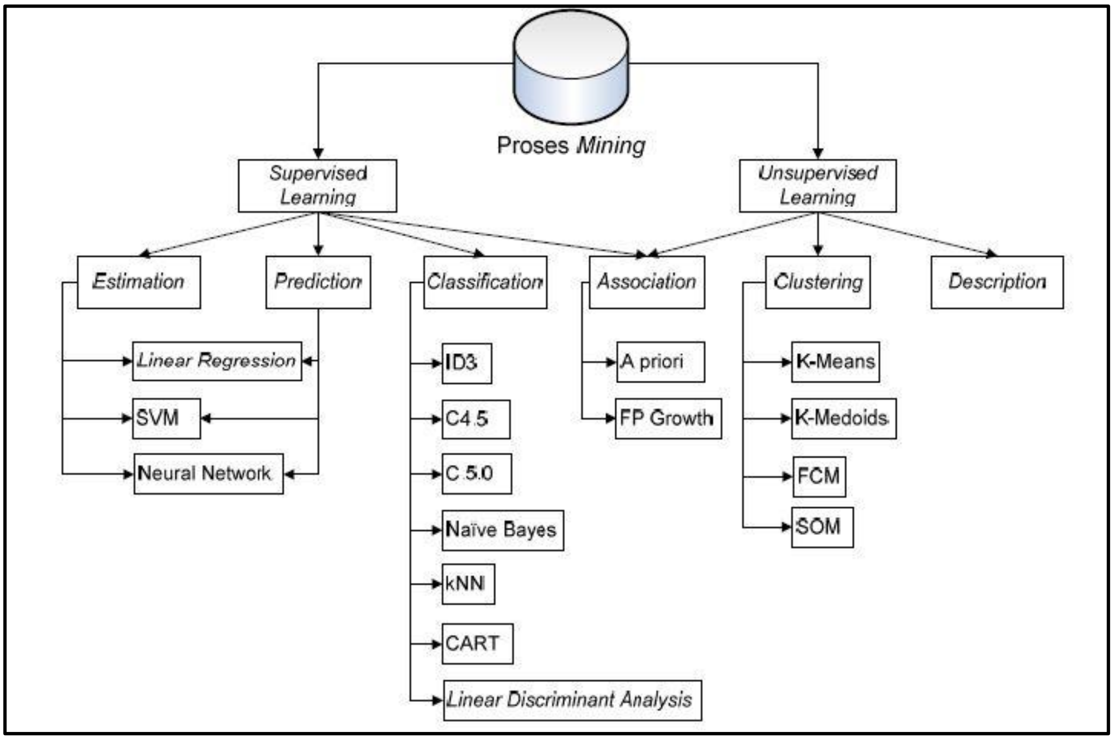

Figure 2.2 Several Methods of Mining Process

6. Pattern Evaluation

   Pattern evaluation is implemented to identify new patterns into the found knowledge base (Ridwan, et al., 2013). In this stage, the results of data mining techniques in the form of typical patterns and predictive models evaluated to assess whether the existing hypothesis is indeed achieved. If the results do not match the theory, then several alternatives can be taken such as making feedback to improve the data mining process, trying other appropriate data mining methods, or accepting these results as unexpected results that may be useful.

7. Knowledge Presentation

   The last stage of data mining process is how to formulate the decision or action from the analysis result obtained. Sometimes this should involve people who do not understand data mining. Therefore, the presentation of data mining results in the form of knowledge that can be followed by everyone is a necessary step in the process of data mining. In the performance, data visualization can also help communicate the results of data mining (Ridwan, et al., 2013).

This Knowledge Discovery in Databases process consists of a series of transformation steps, from pre-processing data and post-processing which is the result of data mining. Data inputs can be stored in various formats such as flat files, spreadsheets, or relational tables and may be in centralized data storage or distributed across multiple addresses. The purpose of preprocessing data is to convert raw input data into an appropriate format for subsequent analysis. The steps taken include repairing dirty or redundant data and selecting records and features relevant to the following data management process. Data can be collected and stored in many ways. Therefore, data processing may take a long time during the entire process of knowledge discovery (Luthfi & Kusrini, 2009).

2.1.3 Clustering

Clustering is the process of grouping or cluster of objects based on information obtained from data that explains the relationship between objects with principles to maximize the similarity between members of one class and minimize the similarity between classes or clusters (Rahmawati, et al., 2016). There are many clustering methods require a distance function to measure similarities between data, as well as means for normalizing the various attributes that the data possess. Some of the most widely known categories of clustering methods are Hierarchical Clustering and K-means Clustering.

2.1.4 Cluster Validation

Each cluster formed has a set of appropriate sizes, such as the value of the cluster validity index (Brock, et al., 2008). Cluster validation is used as a reference in determining the optimal number of clusters. One of the commonly used cluster validation is internal validation. Internal validation uses internal information on the data to assess the quality of clustering. Internal validation includes:

1. Connectivity

   It shows the level of cluster relationship, determined by the number of nearest neighbors.
2. Silhouette Value

   It is the average value of Silhouette for each data. This value measures the level of confidence in the clustering process of each observational data.
3. Dunn Index

   It is the ratio value of the closest distance between observation data in different clusters to the farthest distance intra cluster.

2.1.5 Multicollinearity

Sambandam (2003) describes multicollinearity as a correlation between more than two variables of data and its effect on clustering. A cluster based on the concept of distance, such as the Euclidean distance, in determining its nearest cluster should consider the presence of multicollinearity.

Suppose there are 25 variables, which are highly correlated with each other (more than 0.5). Then the 25 variables are a linear combination of 5 factors or essential variables. Instead of clustering using 25 variables, clustering can be done using only five essential variables. If clustering with 25 variables is done, then the result will be different. Sambandam (2003) demonstrates the influence of multicollinearity on clustering results.

2.1.6 Hierarchical Clustering

According to Gan, et al. (2007) describes hierarchical algorithm divides a data set into a sequence of nested partitions. Hierarchical clustering is divided into two which are Agglomerative Clustering and Divisive Clustering. Agglomerative Clustering classifies data with a bottom-up approach, while Divisive Clustering uses a top-bottom approach. The hierarchical agglomerative clustering method assumes any existing data as a cluster at the beginning of the process, or it starts with every single object in a single cluster, then it repeats merging the closest pair of clusters according to some similarity criteria until all of the data are in one cluster. There is some disadvantage for agglomerative hierarchical clustering which are:

1. Data points that have been incorrectly grouped at an early stage cannot be reallocated, and
2. Different similarity measures for measuring the similarity between clusters may lead to different results.

If we treat agglomerative hierarchical clustering as a bottom-up clustering method, then divisive hierarchical clustering can be viewed as a top-down clustering method. Divisive hierarchical clustering starts with all objects in one cluster and repeats splitting large clusters into smaller pieces. Divisive hierarchical clustering has the same drawbacks as agglomerative hierarchical clustering. A dendrogram, a special type of tree structure, is often used to visualize a hierarchical clustering.

2.1.7 K Means Clustering

According to MacQueen (1967), K Means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (i.e., k clusters), where k represents the number of groups pre-specified by the analyst. It classifies objects in multiple groups (i.e., clusters), such that objects within the same cluster are as similar as possible (i.e., high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (i.e., low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e., centroid) which corresponds to the mean of points assigned to the cluster. As cited in Rahmawati, et al. (2016), there are the steps of the K-Means algorithm:

1. Determine the number of k-clusters to be formed
2. Generate random values for the center of the initial cluster (centroid) as much as k-clusters
3. Calculate the distance of each input data on each centroid using the Euclidean Distance formula until it finds the closest distance of each data with the centroid. The Euclidean Distance is defined as follows

   $$\boxed{d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2}}$$

   Figure 2.3 Euclidean Distance Formula

   With (x_i, μ_i) is the distance between the cluster x with the cluster center μ on the i, x_i is the i weight of the cluster to be searched for, μ_i is the weight of the i at the center of the cluster.
4. Classify each data by its proximity to the centroid (smallest distance).
5. Updating the centroid value. The value of the new centroid is obtained from the average cluster in question by using the following formula:

   $$\boxed{C_k = \frac{1}{n_k} \sum d_i}$$

   Figure 2.4 Formula to update the centroid value

   Where:

   n_k = the amount of data in the cluster

   d_i = the sum of the value of the incoming distance in each cluster
6. Repeat from step 2 to 5 until members of each cluster nothing changes.
7. If step 6 has been met, then the mean value of the cluster center (μ_j) in the last iteration will be used as a parameter to determine the classification of data.

2.1.8 Fuzzy C Means Clustering

Fuzzy C Means (FCM) is one of the most common methods widely used today. FCM was discovered by Dunn (1973) and later developed by Bezdek (1981). FCM is a method of clustering which allows one piece of data to belong to two or more clusters. FCM is frequently used in pattern recognition. It is based on minimization of the following objective function.

$$\boxed{J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \, \| x_i - c_j \|^2, 1 \leq m < \infty}$$

Figure 2.5 Objective Function for Fuzzy C Means

Where m is any real number greater than 1, U_{ij} is the degree of membership of X_i in the cluster j, X_i is the ith of d −dimensional measured data, C_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership U_{ij} and the cluster centers C_j by:

$$\boxed{u_{ij} = \frac{1}{\sum_{k-1}^{C}\left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|}\right)^{\frac{2}{m-1}}} \;, \quad c_j = \frac{\sum_{i-1}^{N} u_{ij}^{m} \cdot \mathrm{x}_i}{\sum_{i-1}^{N} u_{ij}^{m}}}$$

Figure 2.6 The update of membership U_ij and the cluster centers Cj

This iteration will stop when max_{ij} $\{|u_{ij}^{(k+1)}$- $u_{ij}^{(k)}|\} < \varepsilon$ where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m. As cited in Bora & Gupta (2014), the FCM algorithm is composed of the following steps:

1. Initialize U = $[u_{ij}]$ matrix, U
2. Initialize U = $[u_{ij}]$ matrix, U$^{(0)}$
3. At k-step: calculate the centers vectors C$^{(k)}$=$[c_j]$ with U$^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

4. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{|| x_{i-} c_j ||}{|| x_{i-} c_k ||} \right)^{\frac{2}{m-1}}}$$

5. If || $U^{(k+1)}$- $U^{(k)}$ || $<$ ε then STOP; otherwise return to step 2

Data are bound to each cluster using a membership function, which represents the fuzzy behavior of this algorithm. To do that, we have to build an appropriate matrix named U whose factors are numbers between 0 and 1 and represent the degree of membership between data and centers of clusters. For a better understanding, the researcher may consider this simple mono-dimensional example. Given a particular data set, supposed to represent it as distributed on an axis. The instance of mono-dimensional will be shown in Figure 2.7.

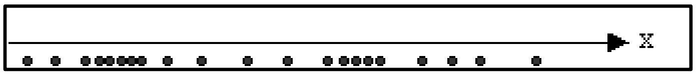

Figure 2.7 The example of mono-dimensional membership function

As shown in Figure 2.7 shows that we may identify two clusters in the proximity of the two data concentrations. We will refer to them using ‘A’ and ‘B’. In the first approach shown in this tutorial - the K-means algorithm - we associated each datum to a specific centroid; therefore, this membership function looked like in Figure 2.8.

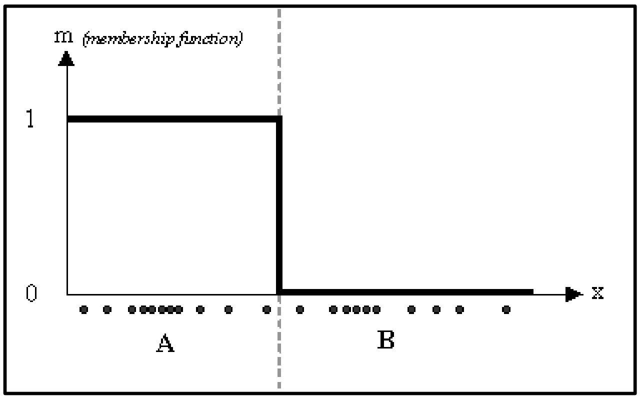

Figure 2.8 The membership function after identified have two clusters

In the FCM approach, instead, the same given datum does not belong exclusively to a distinct cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient and it has been shown in Figure 2.9.

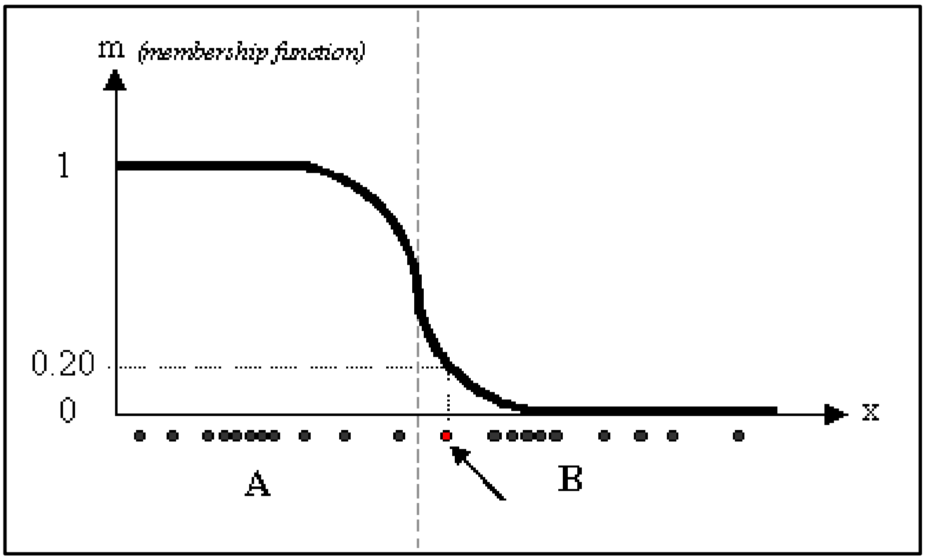

Figure 2.9 The membership function follow a smoother line to indicate that every datum may belong to several clusters

As shown in Figure 2.9, the datum shown as a red marked spot belongs more to the B cluster rather than the A cluster. The value 0.2 of' indicates the degree of membership to A for such datum.

2.1.9 Total and Between Sum of Squares

According to LambdaVu (2014), BSS or TSS value is a measure of the goodness of the classification clustering algorithms has found. SS stands for Sum of Squares, so it's the usual decomposition of deviance in deviance "Between" and deviance "Within". Ideally, you want a clustering that has the properties of internal cohesion and external separation, i.e., the BSS/TSS ratio should approach 1. If the clustering result gives a BSS/TSS ratio of 88.4% (0.884) indicating a good fit. Furthermore, the high similarity within a group will equal to low variance within the cluster or within_SS (WSS) and low similarity between the groups similar to high variation between the clusters, or between_SS (BSS). It means that if the BSS/TSS ratio is 88.4% indicating it has 88.4% low similarity between groups.

2.1.10 Association Rule – Market Basket Analysis (AR-MBA)

Nowadays, the use of barcode and transaction processing machines has been commonly used in retailers (shops or supermarkets). With this machine, retailers can store their transaction data in a transaction database. Each transaction information contains the date and item purchased. This data is referred to as basket data. Market Basket Analysis is one of the essential techniques used by large retailers to uncover associations between items. According to Li (2017), MBA works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

Association Rules are widely used to analyze retail basket or transaction data and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of firm rules. Association rule mining consists of first finding frequent itemsets (sets of items, such as A and B, satisfying a minimum support threshold, or percentage of the task-relevant tuples), from which strong association rules in the form of A and B are generated. These rules also satisfy a minimum confidence threshold (a prespecified probability of satisfying B under the condition that A is satisfied). Associations can be further analyzed to uncover correlation rules, which convey statistical correlations between itemsets A and B (Han, et al., 2012). When it’s implemented in a transaction database, these association rules will be beneficial in defining business strategies such as catalog design, product layout, and designing marketing and promotional campaigns.

One possible example of an association rule, for instance, is that 80% of customers who purchase product A is also buying product B. In this case product A and B are called as frequent itemset. From the frequent itemset, we can specify the association rules between items in the frequent itemset. Finding association rules from a transaction database is not a trivial issue. First, the number of transactions contained in the database is very much. Second, the number of frequent itemset possibilities increases exponentially against the number of item types. Various algorithms can be used for AR-MBA applications, such as Apriori algorithm and FP-growth algorithm. The explanation of both algorithms are as follows:

A. Apriori Algorithm

Apriori is an original algorithm proposed by R. Agrawal and R. Srikant in 1994 for frequent mining itemsets for Boolean association rules. Apriori algorithm is an algorithm to reduce the search space of item combinations so that analysis can be done more quickly. Furthermore, the rules resulting from the apriori algorithm can be identified again to determine which rules can provide more information using the support and lift ratio sizes. Then, the association rules that have been produced can be used as consideration for decision making in business strategy. The two primary processes performed in the Apriori algorithm (Jiawei & Kamber, 2006), are as follows:

1. Join

   In this process, each item is combined with other things until no more combinations are formed.

2. Prune

   In this process, the results of the items that have been combined were then trimmed using the minimum support that has been determined by the user.

B. FP-Growth Algorithm

FP-Growth algorithm is a development of Apriori algorithm. Thus, the shortcomings of the Apriori algorithm are improved by the FP-Growth algorithm. Frequent Pattern Growth (FP-Growth) is an alternative algorithm that can be used to determine the most frequent itemset in a data set (Samuel, 2008). In the Apriori algorithm, a generated candidate is required to obtain frequent itemsets. However, in FP-Growth generate candidate not done because FP-Growth uses the concept of tree development in search of frequent itemsets. That is what causes the FP-Growth algorithm faster than Apriori algorithm.

Characteristics of an FP-Growth algorithm is the data structure used is a tree called FP-Tree. By using FP-Tree, FP-growth algorithm can directly extract frequent Itemset from FP-Tree. Numerous itemset excavation using the FP-Growth algorithm will be generated by generating a data tree structure or called FPTree. The FP-growth method can be divided into three main stages as follows (Jiawei & Kamber, 2006):

1. The step of generating conditional pattern base
2. The phase of the FP-Tree limited generation, and
3. Stage of frequent itemset search.

2.1.11 R software

R is a suite of software used for data manipulation, calculation, simulation, graphics display, and simultaneously as an interpreted programming language. R is derived from the S language, a programming language developed at Bell Laboratory. It is because R is open source licensed, it can be acquired and distributed free of charge under the GNU public license (Wiharto, 2013).

According to Wiharto (2013), R can be run on Windows operating system, Mac OS X, Unix, and Linux. R software can be downloaded from CRAN website (Comprehensive R Archive Network) which is http://lib.stat.cmu.edu/R/CRAN/. R is a programming language that is object-oriented, so all variables, data, functions, etc. are stored in computer memory as objects. R is a programming language type interpreter that is case sensitive. The R design is strongly influenced by two pre-existing computer programming languages, the S language developed by Becker, Chamber, and Wilk, as well as the Scheme programming language developed by Sussman. Thus, the style is very similar to the S language, but Scheme supports its implementation and semantics. The critical thing in R is a self-learning programming environment. Therefore, the purpose of this research is to build an R-based language pipeline for clustering analysis in increasing product sales by combining Hierarchical Clustering and K-Means Clustering.

2.2. Inductive Literature

The right sales strategy is essential in business to increase sales value. Many factors can affect the selling rate of a good sold, such as Luthfi (2009) use data mining with association algorithm to compile a system that can see the pattern of sales of goods which can then be used to develop new sales strategies. From this research found that the business owner can arrange a sales strategy related to the relationship of occurrence of goods simultaneously in a transaction.

Rahmawati, et al. (2016) clustered thesis documents at Chemistry Department, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret (UNS). The clustering method used in theirs is a combination of Hierarchical Clustering and K-Means Clustering. It was found that documents clustering yielded 16 clusters. The cluster results are analyzed by the relationship between the documents and the theme of each cluster. The clustering result is also related to the lecturer who teaches at Chemistry Department. It shows that the lecturer's skill influences the variation of the research theme conducted by the students. It is also known that the number of studies on a subject relates to student interests and lecturer projects in the Chemistry Department.

Tanjung (2016) implemented the Fuzzy C-Means (FCM) algorithm to segment customer data bank. The result of her thesis is a software that can be used as a tool for grouping data bank customers using FCM algorithm. Based on her result test by using the appeal result with 15 data values, the researcher got cluster number (c) by 2, rank or weighting (w) by 2, many iterations (i) 10 as well as the tolerance error (e) is 0.00001. It is concluded that the clustering analysis can produce output that corresponds to the expected user and the FCM algorithm can segment the data set based on the numeric type of data.

The implementation of Data Mining in Sales of Beverage Products at PT. Pepsi Cola Indobeverages has been done by Irdiansyah (2010). Irdiansyah (2010) uses the Clustering method to identify objects that share specific characteristics, and then use those characteristics as "characteristic vectors" or "centroid". Based on the study, found that the clustering method can help PT. Pepsi Cola Indobeverages as an overview for corporate decision making to get the pattern of sales of products produced. For further research, the researcher suggests that the application can be developed further with a more extensive and broader data processing, such that the form really can be used as a tool in more accurate and useful decision making for the company.

Astika (2015) has analyzed the sale of goods in Supermarket Sejahtera Lhokseumawe. The method used in this research was Apriori algorithm to perform association analysis. The required data has been taken from the sales transaction data for a specified period and processed to produce association rules of goods and transactions. From the study, the researcher found that the system can process transaction data to find frequent item-set and association rule that meets transaction limit and able to display regulations in text form. In the analysis of some data, she found that the smaller the transaction limit and the minimum confidence determined, the more rules generated, with the consequence of processing time will be longer than the transaction limit and the minimum confidence is greater.

Felani (2015) has determined the strategy of selling food and beverage products at Toserba Lestari Baru Gemolong. She compared three methods of Data Mining to determine the sales strategies, namely Decision Tree, K-Means Clustering, and Linear Regression which is applied using rapid miner software. The study was conducted with a group of data to determine the percentage of precision value, recall, and accuracy. The study concluded that K-Means method Clustering has better value than other methods on precision and accuracy side, while Linear Regression method has better recall value.

Kamagi & Hansun (2014) also predicted a student's graduation rate with data mining. The purpose of this study is to predict students' graduation with a C4.5 algorithm as a reference for making policies and actions of academic fields (BAAK) in reducing students who graduated late and did not pass. From the research, the category of student's GPA from semester one to semester six, gender, origin of high school, and the number of credits, it can predict the graduation of students with conditions such as quickly pass, pass on time, pass late and drop out using data mining with the C4.5 algorithm. From the research found that category of semester six is highly influential on the predicted outcome of graduation. With the application test result, the accuracy of the graduation prediction acquired is 87.5% from 60 data training and 40 data testing. The results of the graduation predictions from this research can help the board of the study program to find out the graduation status of the students. The analysis can be a recommendation for taking courses for students for the next semester such as thesis and apprenticeship so that the students can graduate at least on time.

Chandra, et al. (2014) has applied data mining to predict customer interest in AJB Bumiputera 1912 Palembang. In the application of data mining conducted on AJB Bumiputera 1912 Palembang by using association rules to find information about the customer's interest based on the criteria of a client relationship to the type of insurance chosen. Based on two data mining applications using XLMiner conducted on the first application, the researcher has determined the minimum value of support at 50 and minimum confidence by 50%, so that it generated six rules. In the second application, the researcher has determined minimum support at 2 and minimum for confidence by 1%, so that it created 67 rules. From the study explained that the higher the minimum value of support and minimum confidence is determined then the better the pattern or rules obtained and vice versa.

Ridwan, et al. (2013) has also implemented Data Mining for student academic performance evaluation. This study focused on evaluating the academic performance of students in the second year and classified of students who can graduate on time or not using the naive Bayes classifier method. The input of this system is student data and student academic data. The sample of students from batch 2005 until batch 2009 will be used for data training and data testing. On the other hand, the student data from batch

2010 to batch 2011 and students who are not yet graduated will be used as the target data. From the study found that the most determinant factor in the student's academic performance assessment is the cumulative GPA, 1st semester GPA, 4th semester GPA, and gender. Therefore, these factors can be used as a university performance evaluation for college managers. Testing data on the student from batch 2005 until batch 2009 using Naïve Bayes Classifier algorithm resulted in precision, recall, and accuracy value of 83%, 50%, and 70%, respectively.

Andriyana (2015) has predicted scholarship recipients based on achievements in SMA Negeri 6 Surakarta. In predicting students who receive the scholarship based on performance, the researcher used three comparison methods which are Naive Bayes, Decision Tree Algorithm ID3, and Linear Regression. The attributes used to consist of Average Score, Gender, Extracurricular, Major, Semester, Total Parent Count, Parent Salary, and Scholarship by using RapidMiner 5. From the study found that based on the number of samples of 305 students, the results of the precision value of the Decision Tree Algorithm ID3 method is better used for this study compared to other methods. While based on the amount of recall and accuracy, Linear Regression is better used than other methods. However, when viewed from the overall results of the predictions of the most influential scholarship recipient is the average score.

CHAPTER III

RESEARCH METHODOLOGY

3.1. Research Object

The object in this research is the product sales data that obtained from a retail company, Pamella Satu Supermarket Yogyakarta, from 1st October until 14th October 2017.

3.1.1 Research Location

This research was taken place in Yogyakarta.

3.1.2 Focus of the Research

This research focuses on a comparison study of combined Hierarchical Clustering plus K-Means Clustering and the Fuzzy C-Means Clustering algorithms and the ARMBA by apriori algorithm toward the marketing and strategy product sales. The R software assists the analysis.

3.2. Research Flow

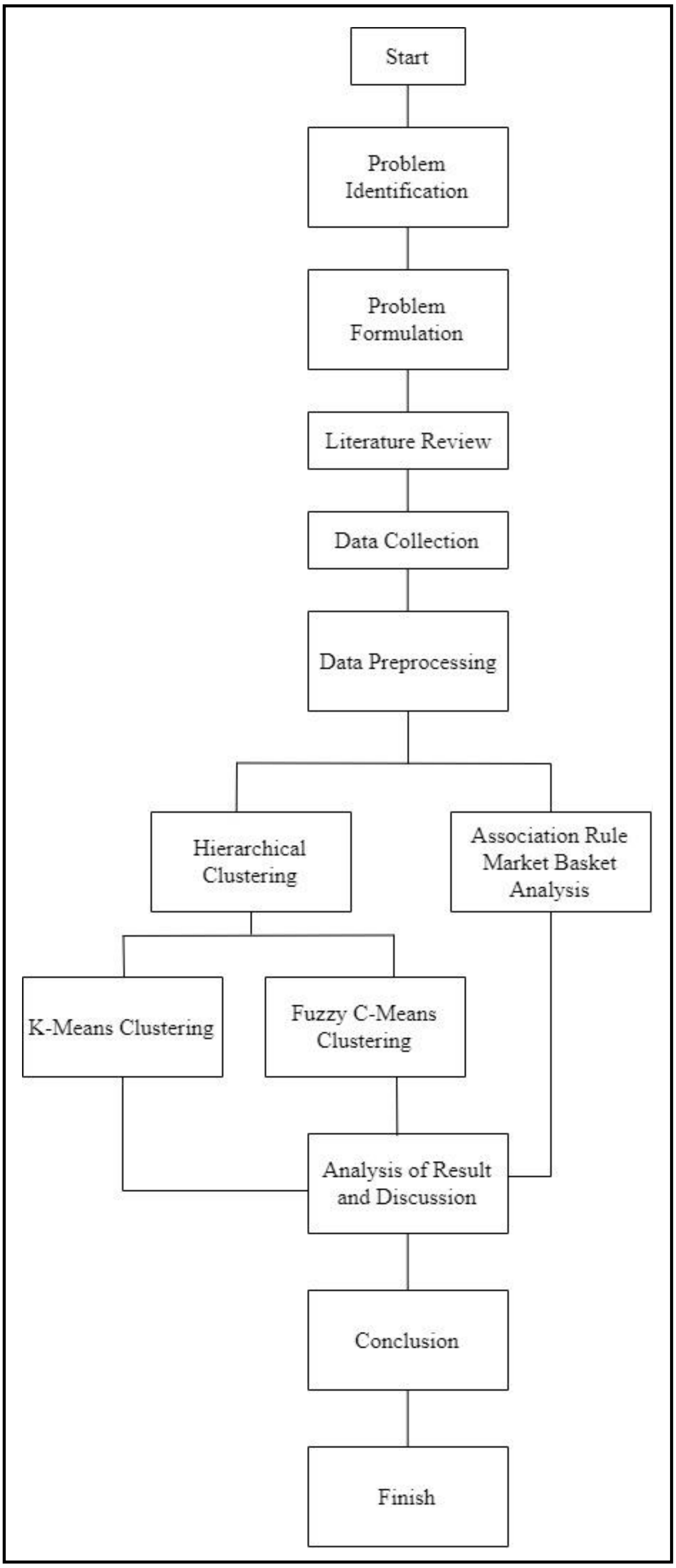

Figure 3.1 Flowchart of Research

3.3. Data Requirement

3.3.1 Primary Data

The primary data is the data that obtained from the observation. It is because the researcher did not conduct an inspection and directly got the product sales data from the retail company database, then there are no primary data taken in this research.

3.3.2 Secondary Data

The secondary data is product sales data that obtained from the retail company, which is the retail data from 1st October until 14th October 2017.

3.4. Method of Data Collection

The appropriate data collection method is to consider its use based on the data type and source. Objective and relevant data with the subject matter of the research is an indicator of the success of the research. The data collection for this research is done in the following way:

1. Interview

   An interview is a technique of data collection by holding question and answer or direct discussion with the marketing department.

2. Literature Review

   A literature review is collecting data by studying problems related to the research object and sourced from manual books, literature compiled by experts to supplement the data required in research.

3.5. Method of Data Analysis

The data analysis methods used are

1. Descriptive analysis method with quantitative approach means that research done is to emphasize the analysis on numerical data (number) which aims to get a clear picture of a situation based on data obtained by presenting, collecting and analyzing the data such that it becomes new information that can be used to explain the problem under research process
2. Inferential analysis by implementing the clustering and association rules algorithm. The researcher uses Knowledge Discovery in Databases (KDD) stages and conducts a comparison study of combined Hierarchical plus K Means Clustering Algorithms and the Fuzzy C Means Clustering Algorithm to get the optimum result of clustering as the basis to perform correct promotion strategy for the retail company.

3.6. Tool

3.6.1 R Software

The researcher will use R Studio software to build R-based language pipeline for a comparison study of combined Hierarchical with K-Means Clustering algorithms and the Fuzzy C-Means Clustering algorithm to increase product sales at a retail company and use it as the basis to perform correct promotion strategy for the retail company.

CHAPTER IV

DATA COLLECTION AND PROCESSING

4.1. Data Collection

This chapter explains the process of data collection. The researcher got a secondary data from Pamella Satu Supermarket Yogyakarta database which contained 3 Product Categories such as Baby & Kids, Hair Care, and Soap. In their database, there are 20,464 transactions that need to be processed in order the researcher be able to find knowledge behind the product sales database by conducting comparison study of combined Hierarchical and K Means Clustering Algorithms with Fuzzy C-Means Clustering Algorithm as the basis to perform correct promotion strategy to increase product sales at a retail company.

4.1.1 User Interview

Initial step on collecting data is interviewing the owner of Pamella Satu Supermarket. The interview is the process of identifying the problem which comes from explanation and information of interviewee. The business owner of Pamella Satu Supermarket explained that there is no customers database in their system. Therefore, each rows transaction and point of sales from product sales data in Pamella Satu Supermarket will be indicated as the customer. In the research, the researcher will conduct clustering analysis based on the behavior of the customer when buying the product. Therefore, the owner can guess several products that need to be considered to be placed on the shelf. It is indicated that there are few products are not very, but the provision of the product is too much. Other than that, the business owner would like to change their sales tactics, or promotion system for their daily product sales will run efficiently and effectively. Thus, the researcher suggested that Clustering Analysis is very appropriate to identify the pattern of product sales to improve their promotion strategy. The researcher will be using six variables on doing Clustering Analysis which are Day, Time, Product Category, Product Sub Category, Sale Quantity, and Price. Based on each variable, the researcher will know the pattern of the customer based on each variable. Therefore, the researcher will get an insight on how to change the business system in the Pamella Satu Supermarket Yogyakarta by comparing combined Hierarchical plus K Means, and Fuzzy C Means Clustering Algorithms. To help the business owner of Pamella Satu Supermarket Yogyakarta to improve their effectiveness of marketing and sales tactics using customer transactions, then the researcher will add Association Rule Market Basket Analysis (AR-MBA) in the research.

4.1.2 Historical Sales Data

In the clustering process, we need to know our goals. The researcher’s goal in the study is to group the customers based on predetermined variables. However, there is no specific information about the customer in the product sales data that the researcher got from Pamella Satu Supermarket Yogyakarta. Thus, the researcher assumed that the transaction (Sale Name column) would be indicated as the customer as shown in Table 4.1.

Table 4.1 Product sales at Pamella Satu Supermarket on October 1-2, 2017

No	Sale Name	Barcode	Product Name	Category	Brand	Variant	Pack Base	Sale Qty	List Price
1	**01-10-2017 08:29/Septi /S7678/T22**	8993417172243	B&B KIDS SP COL RASPBERRY 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	B&B	KIDS SP COL RASPBERRY	100 ML	1	12.225
2	**01-10-2017 09:32/dodo/ S7687/T26**	4801010120223	JOHNSONS BABY COLOGNE SLIDE 125 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSON	BABY COLOGNE SLIDE	125 ML	1	20.700

No	**Sale Name**	**Barcode**	**Product Name**		**Brand**	**Variant**	**Pack Base**	**Sale Qty**	**List Price**
3	**01-10-2017 09:40/agung/S7684/T34**	8993417372223	ESKULIN KIDS MIST COL DONALD 100 ML ALL / NON-FOODS / BABY & KIDS / COLOGNE		ESKULIN	KIDS MIST. DONALD	100 ML	1	12.150
4	**01-10-2017 10:08/dodo/S7687/T39**	8888103007513	CUSSONS BABY COL.SWEET L/VIO 100M	ALL / NON-FOODS / BABY & KIDS / COLOGNE	CUSSONS	BABY COL.SWEET L/VIO	100 ML	1	15.550
5	**01-10-2017 10:43/agung/S7684/T57**	8992694242113	ZWITSAL B COL.F.FLORAL 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	ZWITSAL	B.COL. F.FLORAL	100 ML	1	15.100
6	**01-10-2017 10:55/dodo/S7687/T56**	8991111102719	JOHNSONS BABY COL.SUMMER SW 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSONS	BABY COL.SUMMER SW	100 ML	1	17.475
7	**01-10-2017 11:08/supri/S7683/T5**	8888103007490	CUSSONS BABY COL.SWEET L/VIO 50ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	CUSSONS	BABY COL.SWEET L/VIO	50 ML	1	9.025
8	**01-10-2017 11:20/supri/S7683/T13**	8991111101583	JOHNSONS BABY COL.BRISA 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSONS	BABY COL.BRISA	100 ML	1	17.475
9	**01-10-2017 11:23/Tami/S7690/T81**	4801010120223	JOHNSONS BABY COLOGNE SLIDE 125 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSON	BABY COLOGNE SLIDE	125 ML	1	20.700
10	**01-10-2017 11:36/sriw/S8286/T12**	8888103209054	CUSSONS BABY COL SOFT T/BLUE 50 M	ALL / NON-FOODS / BABY & KIDS /	CUSSONS	BABY COL SOFT T/BLUE	50 ML	1	9.025

				COLOGNE					
No	**Sale Name**	**Barcode**	**Product Name**	**Category**	**Brand**	**Variant**	**Pack Base**	**Sale Qty**	**List Price**
11	**01-10-2017 11:36/sriw/S8286/T12**	8993417182228	B&B SPRAY COLOGNE ACTIVE	ALL / NON-FOODS / BABY & KIDS /	B&B	SPRAY COLOGNE ACTIVE	100 ML	1	15.400
12	**01-10-2017 11:42/arif/S7681/T123**	4801010120223	JOHNSONS BABY COLOGNE SLIDE 125 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSON	BABY COLOGNE SLIDE	125 ML	1	20.700
13	**01-10-2017 12:08/supri/S7683/T30**	8993102684976	BAMBI BABY COLOGNE 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	BAMBI	BABY COLOGNE	100 ML	1	20.375
14	**01-10-2017 12:09/agung/S7684/T91**	4801010120223	JOHNSONS BABY COLOGNE SLIDE 125 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSON	BABY COLOGNE SLIDE	125 ML	1	20.700
15	**01-10-2017 12:21/arif/S7681/T153**	8991111101613	JOHNSONS BABY COL.MORN DEW 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSONS	BABY COL.MORN DEW	100 ML	1	17.475
16	**01-10-2017 12:56/agung/S7684/T100**	8993417172144	B&B KIDS SPR COL RASPBERRY 60 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	B&B	KIDS SP COL RASPBERRY	60 ML	1	7.775
17	**01-10-2017 12:56/agung/S7684/T100**	8993417312229	MASTER KIDS COL SPRAY SUPERMAN 100 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	MASTER	KIDS COL SUPERMAN	100 M	1	12.150
18	**01-10-2017 12:58/dodo/S7687/T89**	4801010127321	JOHNSONS BABY COL HAPPY BERRIES 125 ML	ALL / NON-FOODS / BABY & KIDS / COLOGNE	JOHNSONS BABY	COL HAPPY BERRIES	125 ML	1	19.875

The dataset contains 20,464 rows with nine columns that will be processed using R software. Because each customer can be identified from the data residing in the column “Sale Name” with the specific Transaction code, then the researcher will conduct data transformation or coding to facilitate the researcher in the next data processing.

4.1.3 Historical Sales Data Preprocessing

To meet the goal of this research which the researcher would like to conduct a comparison study of combined Hierarchical and K Means Clustering Algorithms with Fuzzy C Means Clustering Algorithm as the basis to perform right promotion and marketing strategy, then the researcher will complete the first stage of data preprocessing which is Data cleaning. The researcher found that there are several attributes in the Product Sub Category variable is not relevant to be processed. The irrelevant attribute is OTHER. Therefore, the researcher changed OTHER attributes into several new attributes such as Baby Gift, Baby Hair & Body Care, Baby Hair Care, Baby Oil & Lotion, Baby Oral Care, Baby Soap Bar, and Baby Soap Liquid. The researcher will propose this new attribute to Pamella Supermarket in the order they can significantly increase their quality of data analysis. Other than that, the researcher also fixed several attributes in Pack base variable.

4.1.4 Historical Sales Data Transformation (Coding)

The researcher realized that the product sales data provided by the Pamella Satu Supermarket Yogyakarta has not been appropriate for the clustering analysis. Nevertheless, the researcher believed that behind the limited data source for clustering process there would be something worth generating. For instance, if the researcher separated the value in the column “Sale Name” into several parts then the researcher will get information about Day, Time and how many customers shopping at Pamella Satu Supermarket Yogyakarta. From the grouping of these variables, the researcher will gain new insight into what day and what time customers tend to buy the most. Thus, the researcher will be able to conclude about what day and time that the business owner needs to pay more attention in keeping the customer to be served efficiently and effectively by changing the working-hour of each employee to the peak day and time.

To accommodate the goal, the researcher needs to make the data transformation or coding. It will put the original data into the categorical form. The example of data transformation or coding is shown in Table 4.2.

Table 4.2 Data Transformation of Pamella Supermarket on October 1-2, 2017

No	Transaction	DayIND	Time IND	CategoryIND	SubCategoryIND	Sale Qty	Price IND
1	S7678T22	7	1	1	2	1	1
2	S7687T26	7	1	1	2	1	1
3	S7684T34	7	1	1	2	1	1
4	S7687T39	7	1	1	2	1	1
5	S7684T57	7	1	1	2	1	1
6	S7687T56	7	1	1	2	1	1
7	S7683T5	7	1	1	2	1	1
8	S7683T13	7	1	1	2	1	1
9	S7690T81	7	1	1	2	1	1
10	S8286T12	7	1	1	2	1	1
11	S8286T12	7	1	1	2	1	1
12	S7681T123	7	1	1	2	1	1
13	S7683T30	7	2	1	2	1	1
14	S7684T91	7	2	1	2	1	1
15	S7681T153	7	2	1	2	1	1
16	S7684T100	7	2	1	2	1	1
17	S7684T100	7	2	1	2	1	1
18	S7687T89	7	2	1	2	1	1
19	S7678T151	7	2	1	2	1	1
20	S7678T22	7	1	1	2	1	1
21	S7687T26	7	1	1	2	1	1
22	S7684T34	7	1	1	2	1	1
23	S7687T39	7	1	1	2	1	1
24	S7684T57	7	1	1	2	1	1
25	S7687T56	7	1	1	2	1	1
26	S7683T5	7	1	1	2	1	1
27	S7683T13	7	1	1	2	1	1
28	S7690T81	7	1	1	2	1	1
29	S8286T12	7	1	1	2	1	1

In the Pamella Satu Supermarket's Data transformation there are 20,464 rows with seven columns. The researcher will conduct clustering analysis using Transaction that indicated as the customer based on six variables, namely Day, Time, Category, SubCategory, Sale Quantity and Price Indicator. The attribute of each Day Indicator can be seen in Table 4.3.

Table 4.3 The attributes of Day Indicator

No	Indicator	Day
1	1	Monday
2	2	Tuesday
3	3	Wednesday
4	4	Thursday
5	5	Friday
6	7	Sunday

As shown in Table 4.3, the researcher cannot include Saturday into the Data Transformation due to researcher’s computer RAM limitation. Thus, the data will be processed from 1st until 6th October 2017 which started on Sunday through Friday at the beginning of the month. Moreover, the attribute of each Time Indicators can be seen in Table 4.4.

Table 4.4 The attributes of Time Indicator

No	Indicator	Time
1	1	Morning
2	2	Afternoon
3	3	Evening
4	4	Night

As shown in Table 4.4, the researcher would like to know when Pamella Satu Supermarket got a hectic transaction activity in a day. Therefore, the researcher will gain an insight on how to create a correct marketing strategy based on this indicator. Other than that, the attribute of each Product Category Indicators can be seen in Table 4.5.

Table 4.5 The attributes of Category Indicator

No	Indicator	Category
1	1	Baby & Kids
2	2	Hair Care
3	3	Soap

As shown in Table 4.5, Category Indicator will give an insight for the researcher to what product that is sold very often in Pamella Satu Supermarket Yogyakarta. Thus, the researcher can be able to create a marketing and sales tactics on how to provide more product in the product display based on this indicator. Moreover, the attribute of Product Sub Category Indicators can be seen in Table 4.6.

Table 4.6 The attributes of Sub Category Indicator

Indicator	SubCategory	Indicator	SubCategory
1	ACNE SOAP PACK	**15**	DIAPERS & WIPES
2	BABY COLOGNE	**16**	FACIAL SOAP MEN
3	BABY GIFT	**17**	FACIAL SOAP WOMEN
4	BABY HAIR & BODY CARE	**18**	FEMININE WASH
5	BABY HAIR CARE	**19**	HAIR COLORING
6	BABY OIL & LOTION	**20**	HAIR NUTRITION
7	BABY ORAL CARE	**21**	HAIR STYLING
8	BABY SOAP BAR	**22**	HAND SOAP
9	BABY SOAP LIQUID	**23**	HEALTHY SOAP BAR
10	BEAUTY LIQUID	**24**	HEALTHY SOAP LIQUID
11	BODY SOAP BAR	**25**	PACIFIER
12	BODY SOAP LIQUID	**26**	SHAMPOO
13	CONDITIONER	**27**	SHAVER
14	CREAMBATH & HAIR MASK	**28**	TALCUM POWDER

As shown in Table 4.6, SubCategory Indicator has 28 values, and the researcher will know exactly what product that the owner needs to be a concern to get the higher profit. Therefore, the researcher will gain an insight into what product subcategory contributed in the most and least selling in Pamella Satu Supermarket later on. Furthermore, the values of Sale Quantity are already numerical, and then the researcher does not need to transform it into categorical data. Lastly, the values of Price Indicator are generating by using weighting. The maximum price will divide the initial amount. For instance, the initial price is IDR. 12,225 and the maximum price is IDR. 189,825, then it will be 12,225/189,825 equals to 0.006. It means that the Price Weight of this transaction is 0.006. Therefore, the price indicator will be shown in Table 4.7.

Table 4.7 The attributes of Price Indicator

Indicator	Price	Price Weight
1	< IDR 47,456	< 0.25
2	< IDR 94,912	< 0.5
3	< IDR 142,369	< 0.75
4	IDR 189,825	1

As shown in Table 4.7, the price indicator will give us an insight into what is the ideal price to sell a product in the whole transaction in Pamella Satu Supermarket Yogyakarta. Therefore, the business owner will see the pattern on what cost that the customers intend to buy a product in their store.

4.2. Data Processing

4.2.1 Multicollinearity Detection

The general rule of thumb is that VIFs are exceeding four warrant further investigation, while VIFs exceeding 10 are signs of severe multicollinearity requiring correction. The way to calculate Multicollinearity detection can be shown in Appendix 1. The multicollinearity detection of this research is shown in Table 4.8.

Table 4.8 The multicollinearity detection result

TimeIND	CategoryIND	SubCategoryIND	Sale.Qty	PriceIND
1.001545	1.111024	1.094984	1.063622	1.064997

As shown in Table 4.8, the VIFs is less than ten which is around 1.00. It means that there is no sign of severe multicollinearity in the variables of Pamella Satu Supermarket Product Sales Data. Therefore, the six variables can be processed for further clustering analysis. In this case, the researcher can directly conduct clustering and no need to perform PCA analysis in advance.

4.2.2 Hierarchical Clustering and its validation

The hierarchical clustering gives two clusters for the retail customer data of Pamella Satu Supermarket Yogyakarta. The cluster validation is needed to be applied. The results of cluster validation are shown in the Tables 4.9 and 4.10.

Table 4.9 Cluster Validation Measures

		2	**3**	**4**	**5**	**6**
Hierarchical	Connectivity	5.5980	7.0980	10.5254	15.4615	17.7484
	Dunn	0.3060	0.3060	0.1562	0.1942	0.1942
	Silhouette	0.8822	0.8621	0.6643	0.6330	0.6328

Table 4.10 The optimal score for cluster center from cluster validation

	Score	**Method**	**Clusters**
Connectivity	5.5980	Hierarchical	2
Dunn	0.3060	Hierarchical	2
Silhouette	0.8822	Hierarchical	2

As shown in Table 4.9 and Table 4.10 the optimal scores for Internal Validation are Connectivity with 2 clusters, Dunn Index with 2 clusters, and Silhouette with 2 clusters. Therefore, the researcher will use cluster centers equals two as the initial K of K Means and Fuzzy C Means Clustering Algorithms. Other than that, Figure 4.1 shows Dendogram for Hierarchical Clustering.

Figure 4.1 shows that if the researcher cuts the tree in the upward side, then it will have resulted around 2 clusters. It means that the amount of initial cluster center will have the same calculation with the Internal Validation using the Hierarchical method in R studio as shown in Table 4.10. On the other hand, it will make the researcher easier in conducting clustering analysis with many product sales data. Furthermore, if the researcher cut the tree downward approximately until the height of 20,000, then the result still same which the initial cluster centers remain 2. However, if the researcher cut more downward, then there will be many initial cluster centers resulted from the dendogram which makes it more accurate, but more it will be more complicated for further clustering analysis.

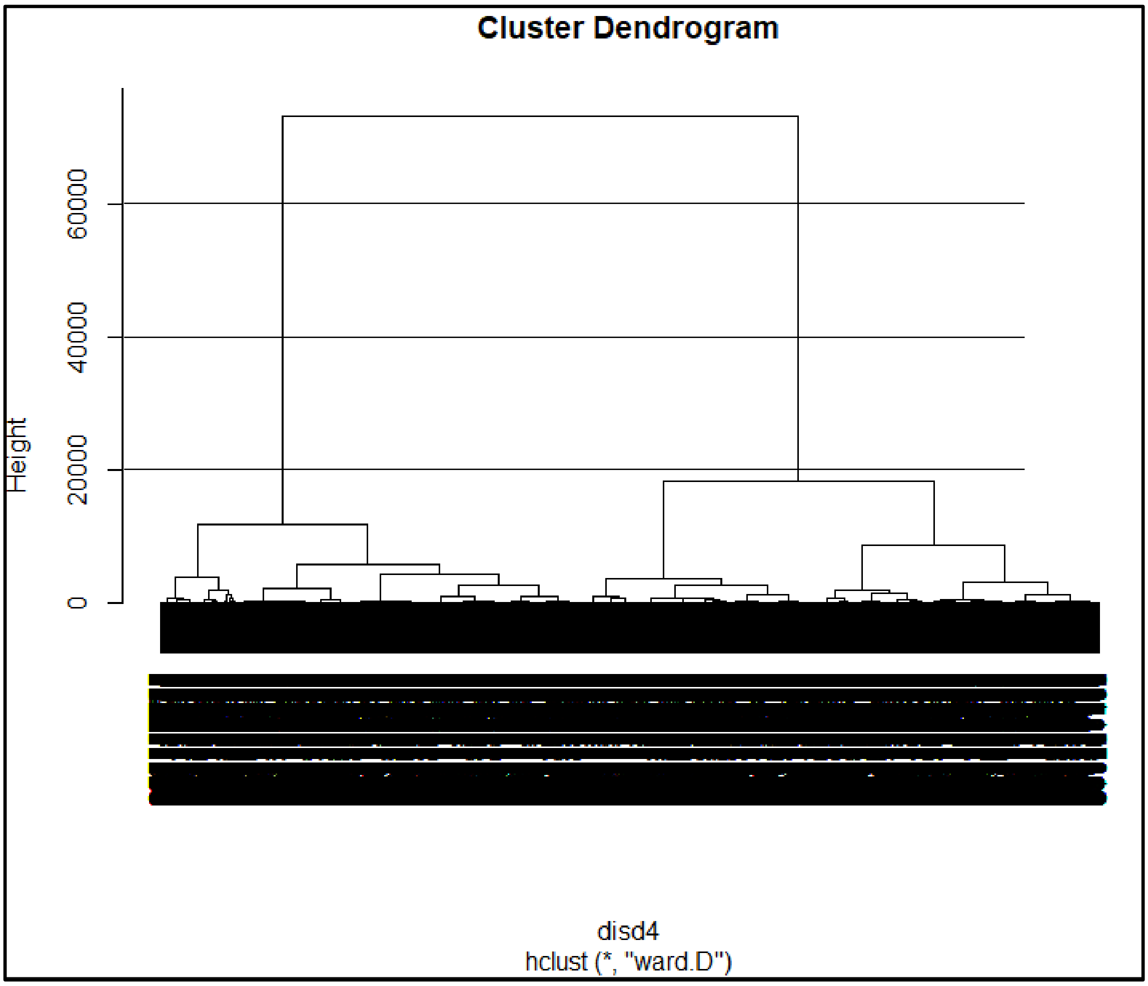

Figure 4.1 Dendogram of cluster based on ward D method of Hierarchical Clustering

4.2.3 K Means Clustering. with K= 2

The clustering will be performed on retail product sales over a period of 6 days starting from 1st October to 6th October 2017. There are 20,464 rows of transactions and each transaction indicated as the customer. The clustering analysis will group each customer into a particular cluster based on six variables such as Day, Time, Category, Sub Category, Sale Quantity and Price Indicators. The process of data processing for K Means

Clustering using R Studio is shown in Appendix 4. The results example of K Means Clustering vector can be displayed in Table 4.11.

Table 4.11 The results example of K Means Clustering

Transaction No	**Cluster**
1	2
2	2
3	2
4	2
5	2
6	2
7	2
8	2
9	2
10	2
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2
19	2
20	2

Table 4.11 shows the K-Means Clustering results that can be identified for each row in the dataset. The visualization of the cluster objects based on the two highest variance components is shown in Figure 4.2. Retracting information from Figure 4.2, there are 11,041 customers belong to cluster 2, and 9,423 belongs to cluster 1.

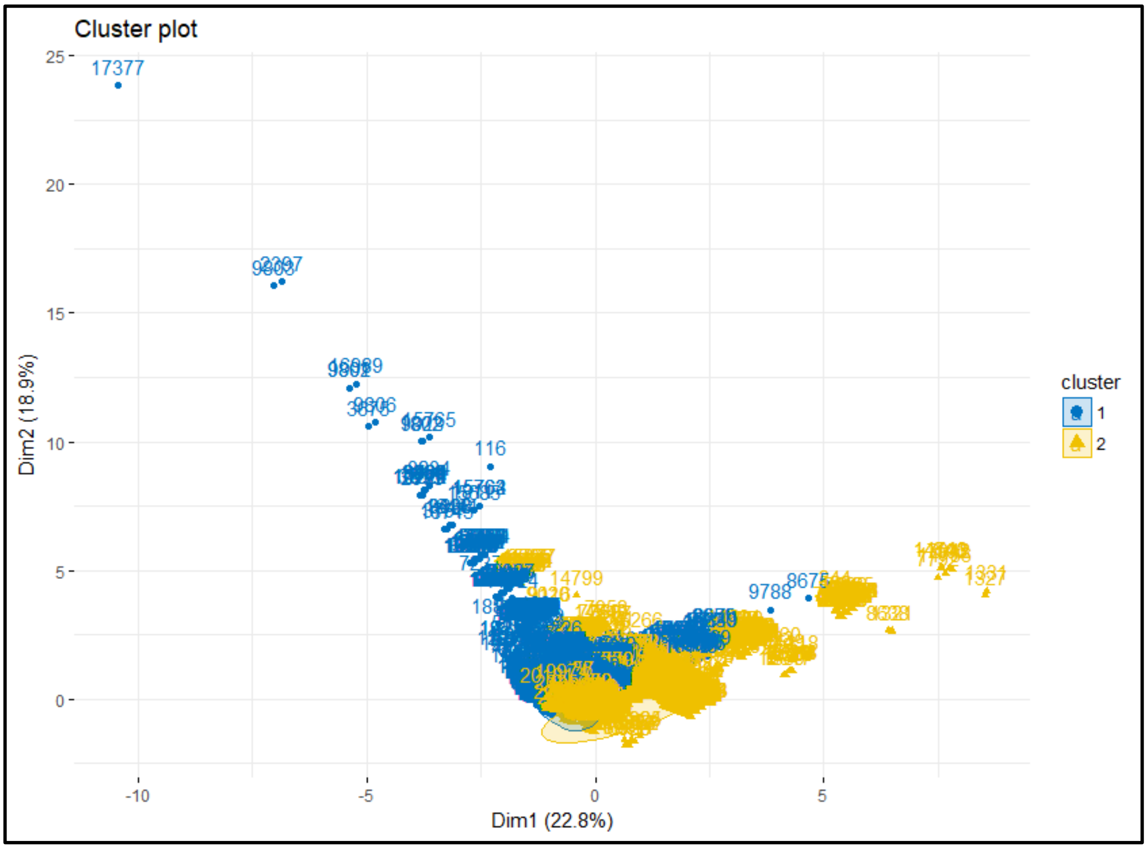

Figure 4.2 Cluster Plot of K Means Clustering with K = 2

4.2.4 Fuzzy C Means Clustering (FCM) with K equals to 2

Starting FCM as well as the other alternating optimization algorithms, an initialization step is required to build the initial cluster prototypes matrix and fuzzy membership degrees matrix. Although this task is usually performed in the initialization step of the clustering algorithm, the initial prototypes and memberships can also be directly inputted by the user.FCM is usually started by using an integer specifying the number of clusters. In this case, the prototypes matrix is internally generated by using any of the prototype initialization algorithms which are included in the package 'inaparc'. Based on the result of cluster validation, then the researcher will use initial cluster center equal to 2. The initial step to conduct partitioning clustering analysis using FCM Algorithm is shown in Appendix 6. Appendix 6 gives a BSS/TSS ratio of 59.34% indicating enough good fit.

Other than that, the Root Mean Squared Deviations (RSMD) is 8.43, and the Mean Absolute Deviations (MAD) is 59.39.

Meanwhile, the data visualization based on the FCM results is shown in Figure 4.3. Contrary to the combined clustering algorithm results, at FCM there are 10,830 customers belong to cluster 1, and 9,634 belong to the cluster 2.

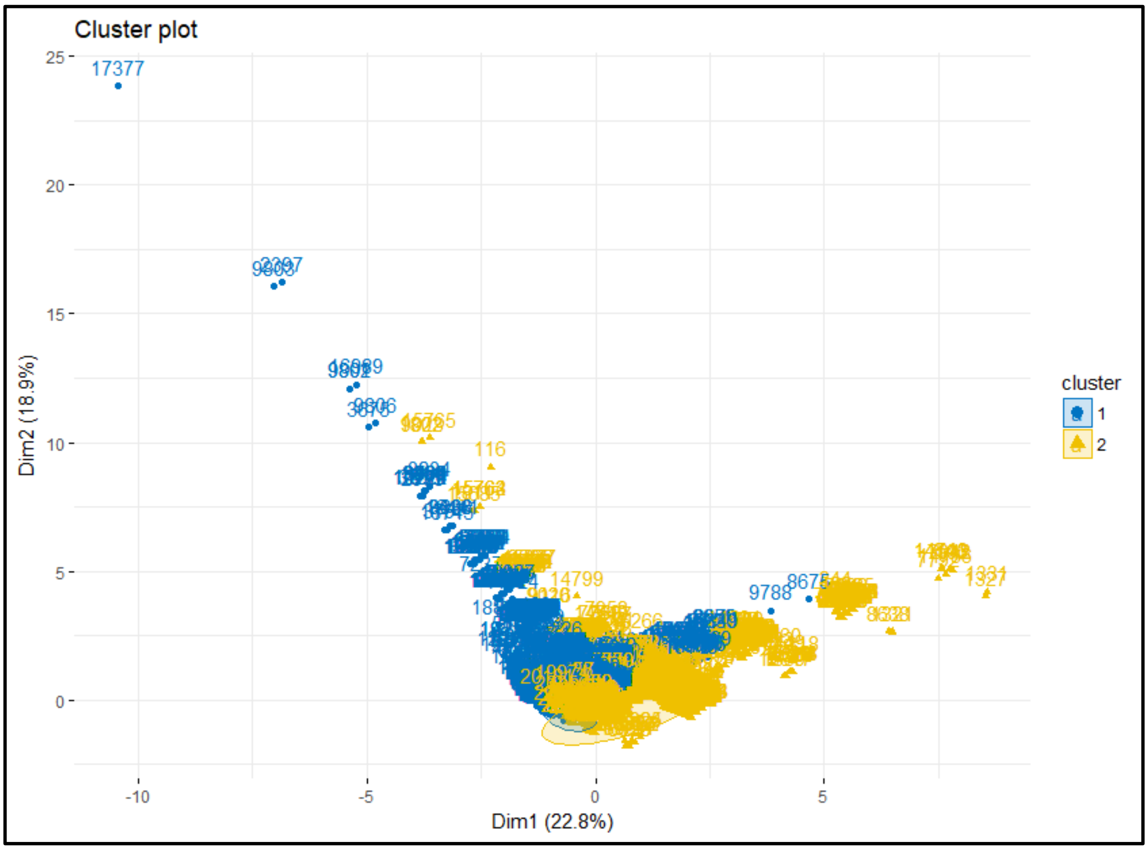

Figure 4.3 Cluster Plot of Fuzzy C Means Clustering with K = 2

The result of which customers belongs to the same cluster between combined clustering and FCM can be shown in Table 4.12.

Table 4.12 The result of which customers belongs to the same cluster based on two approaches

No	Transaction	KMeans Cluster	FCM Cluster
1	S7690T52	1	1
2	S7708T4	1	1
3	S7707T42	1	1
4	S7684T20	2	2
5	S7681T71	2	2
6	S7687T50	2	2
7	S7690T78	2	2
8	S8286T17	2	2
9	S7681T145	2	2
10	S7681T148	2	2
11	S7678T110	2	2
12	S7678T121	2	2
13	S7689T32	2	2
14	S7680T63	2	2
15	S8287T2	2	2
16	S8287T4	2	2
17	S7683T115	2	2
18	S8287T29	2	2
19	S8289T47	2	2
20	S8287T54	2	2
21	S8287T91	2	2
22	S7685T112	2	2
23	S7685T135	2	2
24	S8287T131	2	2
25	S7683T178	2	2
26	S7680T127	2	2
27	S7688T207	2	2
28	S7689T179	2	2
29	S7680T157	2	2
30	S7688T215	2	2
31	S7680T174	2	2
32	S7712T18	2	2
33	S7712T27	2	2
34	S7711T53	2	2
35	S7712T48	2	2
36	S7709T107	2	2
37	S7712T62	2	2
38	S7712T69	2	2
39	S7709T138	2	2
40	S7712T93	2	2
41	S7712T100	2	2

No	Transaction	KMeans Cluster	FCM Cluster
42	S7711T112	2	2
43	S7712T118	2	2
44	S7711T135	2	2
45	S7711T138	2	2
46	S7715T13	2	2
47	S7718T42	2	2
48	S7707T34	2	2
49	S7715T31	2	2
50	S7713T32	2	2
51	S7715T47	2	2
52	S7714T60	2	2
53	S7714T84	2	2
54	S7708T120	2	2
55	S7713T126	2	2
56	S7713T127	2	2
57	S7714T136	2	2
58	S7708T187	2	2
59	S7718T203	2	2
60	S7707T134	2	2
61	S7713T152	2	2
62	S7707T141	2	2
63	S7718T213	2	2
64	S7718T214	2	2
65	S7707T152	2	2
66	S7708T213	2	2
67	S7718T225	2	2
68	S7715T81	2	2
69	S7708T230	2	2
70	S7714T175	2	2
71	S7715T94	2	2
72	S7718T246	2	2
73	S7713T200	2	2
74	S7714T191	2	2
75	S7714T193	2	2
76	S8336T15	2	2
77	S7708T254	2	2
78	S8336T18	2	2
79	S7708T268	2	2
80	S7708T269	2	2
81	S7714T207	2	2
82	S7708T272	2	2
83	S7714T211	2	2
84	S7714T211	2	2
85	S7737T54	1	1

No	Transaction	KMeans Cluster	FCM Cluster
86	S7735T12	2	2
87	S7728T41	2	2
88	S7731T80	2	2
89	S7735T102	2	2
90	S7740T101	2	2
91	S7725T123	2	2
92	S7730T34	2	2
93	S7726T26	2	2
94	S7730T91	2	2
95	S7726T44	2	2
96	S7730T104	2	2
97	S7730T106	2	2
98	S7734T73	2	2
99	S7734T73	2	2
100	S7726T77	2	2
101	S7726T80	2	2
102	S7726T83	2	2
103	S7726T84	2	2
104	S7726T101	2	2
105	S7737T100	2	2
106	S8383T13	2	2
107	S7737T118	2	2
108	S7734T137	2	2
109	S8384T15	2	2
110	S8384T19	2	2
111	S7732T83	2	2
112	S7737T130	2	2
113	S7737T140	2	2
114	S7737T140	2	2
115	S7734T165	2	2
116	S8383T36	2	2
117	S8384T44	2	2
118	S7734T180	2	2
119	S7726T158	2	2
120	S7726T168	2	2
121	S8383T66	2	2
122	S7737T174	2	2
123	S7734T202	2	2
124	S7741T81	2	2
125	S8384T115	2	2
126	S7732T104	2	2
127	S7733T170	2	2
128	S7732T121	2	2
129	S7758T19	2	2

No	Transaction	KMeans Cluster	FCM Cluster
130	S7758T26	2	2
131	S7755T31	2	2
132	S7762T37	2	2
133	S7758T50	2	2
134	S7753T81	2	2
135	S7755T97	2	2
136	S7755T109	2	2
137	S7750T76	2	2
138	S7753T96	2	2
139	S7755T124	2	2
140	S7752T56	2	2
141	S7752T59	2	2
142	S7748T1	2	2
143	S7754T12	2	2
144	S7757T35	2	2
145	S7759T99	2	2
146	S8430T22	2	2
147	S7748T103	2	2
148	S7749T100	2	2
149	S7748T129	2	2
150	S7757T107	2	2
151	S8430T90	2	2
152	S7751T143	2	2
153	S7759T170	2	2
154	S7748T179	2	2
155	S7759T185	2	2
156	S7749T155	2	2
157	S7754T176	2	2
158	S7748T223	2	2
159	S7759T211	2	2
160	S7757T170	2	2
161	S8430T146	2	2
162	S7757T171	2	2
163	S7757T182	2	2
164	S7774T14	1	1
165	S7830T83	1	1
166	S7830T83	1	1
167	S7830T83	1	1
168	S7830T83	1	1
169	S7780T11	2	2
170	S7780T22	2	2
171	S7769T96	2	2
172	S7769T97	2	2
173	S7775T102	2	2

No	Transaction	KMeans Cluster	FCM Cluster
174	S7781T152	2	2
175	S7779T6	2	2
176	S7777T27	2	2
177	S7774T45	2	2
178	S7770T83	2	2
179	S7773T12	2	2
180	S8477T54	2	2
181	S7785T42	2	2
182	S7770T167	2	2
183	S7772T137	2	2
184	S7773T87	2	2
185	S7773T94	2	2
186	S7776T164	2	2
187	S7776T170	2	2
188	S7777T177	2	2
189	S7770T225	2	2
190	S7776T177	2	2
191	S7770T235	2	2
192	S7778T142	2	2
193	S7777T196	2	2
194	S7776T208	2	2
195	S8477T170	2	2
196	S7830T5	2	2
197	S7825T51	2	2
198	S7832T41	2	2
199	S7833T79	2	2
200	S7823T99	2	2
201	S7823T99	2	2
202	S7823T99	2	2
203	S7826T14	2	2
204	S7833T131	2	2
205	S7826T41	2	2
206	S7823T133	2	2
207	S7831T2	2	2
208	S7831T5	2	2
209	S7829T13	2	2
210	S7826T77	2	2
211	S7822T16	2	2
212	S7822T17	2	2
213	S7834T49	2	2
214	S7826T140	2	2
215	S7826T169	2	2
216	S7822T145	2	2
217	S7822T148	2	2

No	Transaction	KMeans Cluster	FCM Cluster
218	S7826T190	2	2
219	S7829T76	2	2
220	S8521T120	2	2
221	S8521T135	2	2
222	S7829T112	2	2
223	S7826T253	2	2
224	S7829T140	2	2
225	S7828T144	2	2
226	S7829T154	2	2
227	S7834T195	2	2
228	S7831T243	2	2
229	S7831T246	2	2

As shown in the Table 4.12, the researcher found that there are 229 customers belongs to the same cluster between combined clustering and FCM.

4.2.5 Association Rules – Market Basket Analysis

Market Basket Analysis is one of the essential techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy. Therefore, in the study, the researcher emphasis on the clustering algorithms which are combined hierarchical with k-means clustering with fuzzy c-means clustering. However, the researcher would like to give a contribution to the business owner that already gave their sales data to identify relationships between the items that their customer bought.

The Association Rules Market Basket Analysis will be based on Transaction, Barcode, Category, Sub Category, Date, Time, List Price, and Sale Quantity. The steps on data processing for product sales with Association Rules Market Basket Analysis using R Studio software are in appendix 8 until appendix 19.

The data summary as shown in Appendix 14 inform that there are 16,594 transactions, and there are 16,622 items. All of these items are the product descriptions in the original Pamella Satu Supermarket Yogyakarta data set.

4.2.5.1 The most and least item frequency plot

The data visualization of the most and least frequent items can be shown in the Figures 4.4 and 4.5. Figure 4.4 shows that the top 5 frequent items that were sold in Pamella Satu Supermarket Yogyakarta are SHAMPOO, DIAPERS & WIPES, FACIAL SOAP WOMEN, HEALTHY SOAP BAR & BODY SOAP BAR. It means that the products were given the high contribution to the product sales of Pamella Satu Supermarket Yogyakarta from 1st October to 14th October 2017. It can be seen that SHAMPOO has 34,105 sales quantity contribution and it is the highest among other products. Other than that, in the customer amount spent, DIAPERS & WIPES has the highest participation in the retail company revenue rather than other products except FACIAL SOAP WOMEN whether its sales quantity is lower than HEALTHY SOAP BAR & BODY SOAP BAR. The least frequent items that were sold in Pamella Satu Supermarket Yogyakarta can be seen in Figure 4.5.

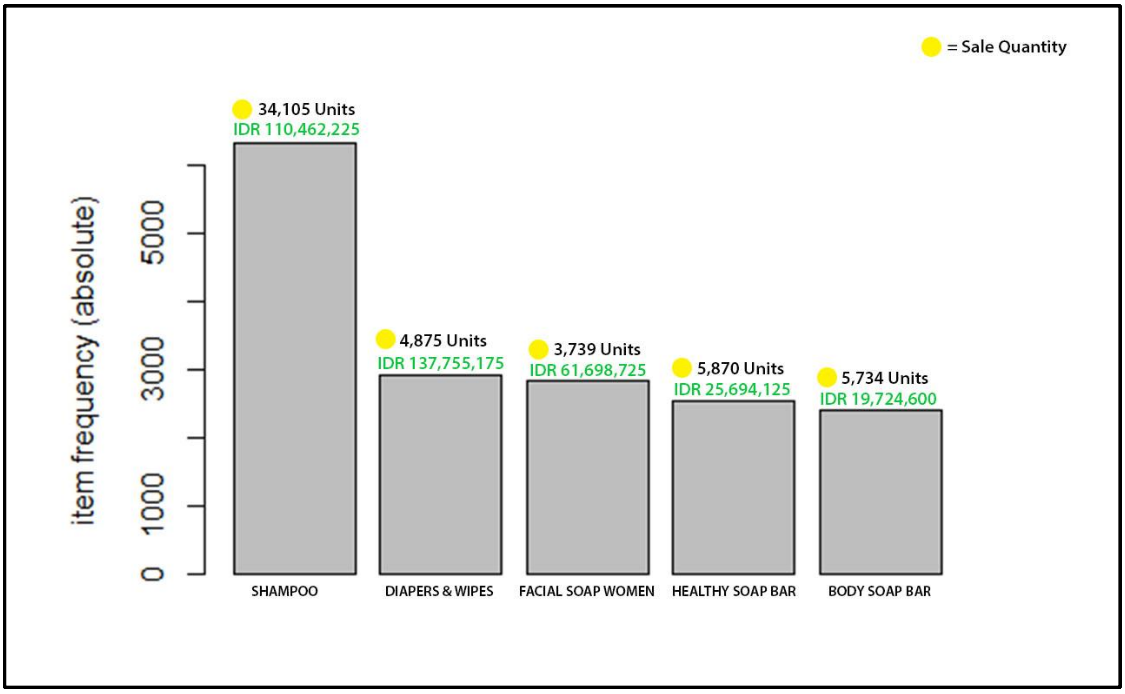

Figure 4.4 The data visualization of Top 5 Item Frequency Plot

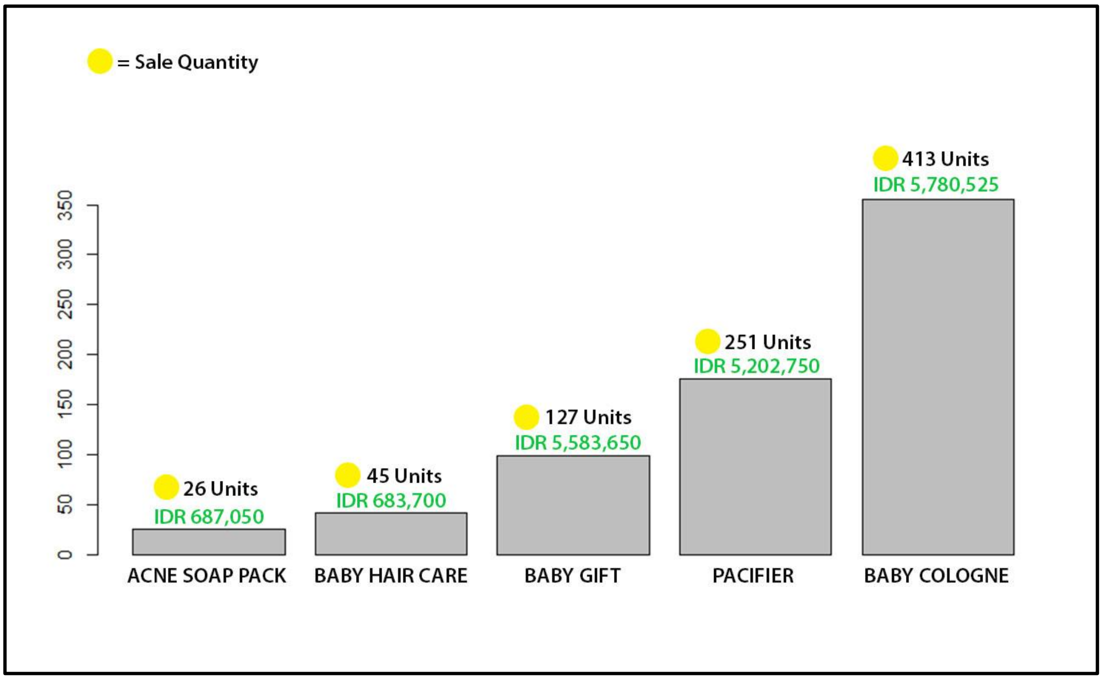

Figure 4.5 The data visualization of the least 5 frequent items

As shown in Figure 4.5, the least five frequent items that were sold in Pamella Satu Supermarket Yogyakarta are BABY COLOGNE, PACIFIER, BABY GIFT, BABY HAIR CARE & ACNE SOAP PACK. It means that the products were giving less contribution to the product sales of Pamella Satu Supermarket Yogyakarta during 1st October to 14th October 2017. It can be seen in Figure 4.5 that ACNE SOAP PACK has 26 units contribution in the product sold and this result indicated that ACNE SOAP PACK become the very major product that gives less input in Pamella Satu Supermarket Yogyakarta. Moreover, the researcher will create several rules in R studio that can be seen in Appendix 16 and Appendix 17.

4.2.5.2 The rules

According to Li, R-bloggers (2017), she determined minimum confidence level equal to 80% and minimum support 0,1% to perform AR-MBA with Apriori Algorithm towards UCI Machine Learning Repository Online Retail dataset. Therefore, the researcher will use similar parameters in this study. There are 23 rules have been generated where the details of the rules can be shown in Table 4.13.

Table 4.13 The generated rules

No	LHS	RHS	Support	Confidence	Lift	Count
1	BABY SOAP LIQUID, CONDITIONER	SHAMPOO	0.001386043	0.9583333	2.512.653	23
2	CONDITIONER, DIAPERS & WIPES, FACIAL SOAP WOMEN	SHAMPOO	0.001024467	0.9444444	2.476.238	17
3	CONDITIONER, DIAPERS & WIPES, HAIR NUTRITION	SHAMPOO	0.001325780	0.9166667	2.403.408	22
4	BODY SOAP BAR, BODY SOAP LIQUID, HAIR NUTRITION	SHAMPOO	0.001265518	0.9130435	2.393.908	21
5	DIAPERS & WIPES, HAIR NUTRITION, HEALTHY SOAP BAR	SHAMPOO	0.001265518	0.9130435	2.393.908	21
6	BODY SOAP LIQUID, HAIR NUTRITION, HEALTHY SOAP BAR	SHAMPOO	0.001506569	0.8928571	2.340.981	25
7	BODY SOAP LIQUID, CONDITIONER, HAIR NUTRITION	SHAMPOO	0.001446306	0.8888889	2.330.577	24
8	CONDITIONER, FACIAL SOAP WOMEN, HEALTHY SOAP BAR	SHAMPOO	0.001386043	0.8846154	2.319.372	23
9	FACIAL SOAP WOMEN, HEALTHY SOAP BAR, SHAVER	SHAMPOO	0.001205255	0.8695652	2.279.912	20
10	BABY HAIR & BODY CARE, FACIAL SOAP WOMEN, HEALTHY SOAP LIQUID	SHAMPOO	0.001144992	0.8636364	2.264.367	19
11	BODY SOAP BAR, CONDITIONER, FACIAL SOAP WOMEN	SHAMPOO	0.001144992	0.8636364	2.264.367	19
12	CONDITIONER, FEMININE WASH	SHAMPOO	0.002229722	0.8604651	2.256.053	37
13	CONDITIONER, HAND SOAP	SHAMPOO	0.002048933	0.8500000	2.228.614	34
14	CONDITIONER, TALCUM POWDER	SHAMPOO	0.001687357	0.8484848	2.224.642	28
15	BABY ORAL CARE, FACIAL SOAP WOMEN, HEALTHY SOAP BAR	SHAMPOO	0.001325780	0.8461538	2.218.530	22
16	BODY SOAP BAR, CONDITIONER, HAIR NUTRITION	SHAMPOO	0.001506569	0.8333333	2.184.916	25
17	BODY SOAP LIQUID, CONDITIONER, FACIAL SOAP WOMEN	SHAMPOO	0.001446306	0.8275862	2.169.848	24
18	CONDITIONER, HAIR COLORING	SHAMPOO	0.001386043	0.8214286	2.153.703	23
19	BABY SOAP LIQUID, HAIR NUTRITION	SHAMPOO	0.001084729	0.8181818	2.145.190	18

No	LHS	RHS	Support	Confidence	Lift	Count
20	CONDITIONER, SHAVER	SHAMPOO	0.002048933	0.8095238	2.122.490	34
21	BABY HAIR & BODY CARE, BABY ORAL CARE, HAND SOAP	SHAMPOO	0.001024467	0.8095238	2.122.490	17
22	BODY SOAP BAR, DIAPERS & WIPES, HAIR NUTRITION	SHAMPOO	0.001265518	0.8076923	2.117.688	21
23	BODY SOAP BAR, CONDITIONER, HEALTHY SOAP BAR	SHAMPOO	0.001506569	0.8064516	2.114.435	25

Those rules show that SHAMPOO is the most frequent by item when a customer purchase other things. The top ten rules can be interpreted as follows:

1. 95.83 % customers who bought “BABY SOAP LIQUID, CONDITIONER” also bought “SHAMPOO.”
2. 94.44 % customers who bought “CONDITIONER, DIAPERS & WIPES, FACIAL SOAP WOMEN” also bought “SHAMPOO.”
3. 91.66% customers who bought “CONDITIONER, DIAPERS & WIPES, HAIR NUTRITION” also bought “SHAMPOO.”
4. 91.3% customers who bought “BODY SOAP BAR, BODY SOAP LIQUID, HAIR NUTRITION” also bought “SHAMPOO.”
5. 91.3% customers who bought “DIAPERS & WIPES, HAIR NUTRITION, HEALTHY SOAP BAR” also bought “SHAMPOO.”
6. 89.28% customers who bought “BODY SOAP LIQUID, HAIR NUTRITION, HEALTHY SOAP BAR” also bought “SHAMPOO.”
7. 88.88% customers who bought “BODY SOAP LIQUID, CONDITIONER, HAIR NUTRITION” also bought “SHAMPOO.”
8. 88.46% customers who bought “CONDITIONER, FACIAL SOAP WOMEN, HEALTHY SOAP BAR” also bought “SHAMPOO.”
9. 86.95% customers who bought “FACIAL SOAP WOMEN, HEALTHY SOAP BAR, SHAVER” also bought “SHAMPOO.”
10. 86.36% customers who bought “BABY HAIR & BODY CARE, FACIAL SOAP WOMEN, HEALTHY SOAP LIQUID” also bought “SHAMPOO.”

.

CHAPTER V

DISCUSSION

5.1. Clustering Analysis

5.1.1 Combined Hierarchical plus K Means Clustering

Clustering algorithms are used to split a dataset into several groups, such that the objects in the same group are as similar as possible and the objects in different groups are as dissimilar as possible. The two clustering algorithms are non-hierarchical such as K Means which is partitioning method used for splitting a dataset into a set of K clusters, and Hierarchical Clustering which is an alternative approach to K means clustering for identifying clustering in the dataset by using pairwise distance matrix between observations as clustering criteria.

However, each of these two standard clustering methods has its limitations. K means clustering requires the user to specify the number of clusters in advance and selects initial centroids randomly. Ward hierarchical clustering is good at identifying small clusters but not large ones. In the study, the researcher already combined the hierarchical with k meant clustering and got a happy result. The result of hierarchical clustering for product sales data from Pamella Supermarket Yogyakarta from 1st until 6th October 2017 is shown in Table 4.11. The optimal scores for Internal validation stated that a proper initial clusters center is 2. It means that the researcher will use k equals to 2 as the primary basis to conduct clustering analysis in the study. From the K Means Clustering result with K = 2, the researcher found that there are 9,423 rows of customer data belong to cluster 1 and 11,041 rows of customer data belong to cluster 2. The aims of this clustering analysis are that to find a right cluster for each variable to run a business effectively and efficiently.

A. Day Indicator as the first variable in Combined Hierarchical plus K-Means Clustering Result with K = 2

In the first variable which is Day, the result is shown in Table 5.1 and Figure 5.1.

Table 5.1 Day variable of K-Means Clustering Result with K = 2

Variable 1	Indicator	Day	Cluster 1 (%)	Cluster 2 (%)
DayIND	1	Monday	17.61	16.38
	2	Tuesday	17.21	18.56
	3	Wednesday	15.10	15.68
	4	Thursday	15.50	15.99
	5	Friday	14.95	13.88
	7	Sunday	19.62	19.53
	Total (%)		**100**	**100**

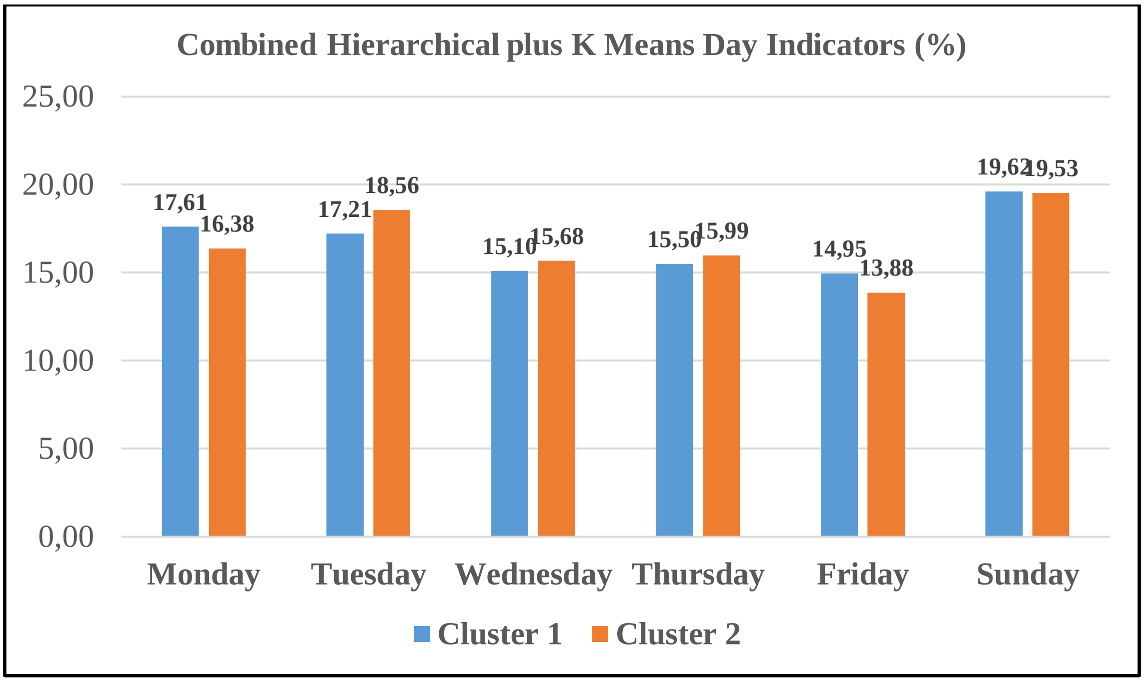

Figure 5.1 Data Visualization of Day variable from Combined K-Means Result with K = 2

As shown in Table 5.1 and Figure 5.1, the researcher found that most of the customers are willing to buy a product in the Pamella Satu Supermarket on Sunday where most of the customers belong to cluster 1 with 19.62% and cluster 2 with 19.53% from the total six days point of sales. The second most days for each cluster are Monday and Tuesday. The least customers in each cluster are on Friday. Based on these results then they can apply 3 marketing strategies as follows:

1. Improving the service for the most customers (Sunday)
2. Improving the service for the second most customer (Monday and Tuesday)
3. Improving the service on the least customers (Friday)

It can be done by managing their employee working hour and try to boost their product sales by giving promotion to a specific product on Friday, Monday, and Tuesday to attract customers to the Pamella Satu Supermarket Yogyakarta.

B. Time Indicator as second variable in Combined Hierarchical plus K-Means Clustering Result with K = 2

The second variable which is Time Indicator are shown in Table 5.2 and Figure 5.2.

Table 5.2 Time Indicator as second variable in Combined K-Means Clustering with K = 2

Variable 2	Indicator	Time	Cluster 1 (%)	Cluster 2 (%)
TimeIND	1	Morning	21.31	20.28
	2	Afternoon	40.44	41.66
	3	Evening	38.25	38.06
	Total (%)		**100**	**100**

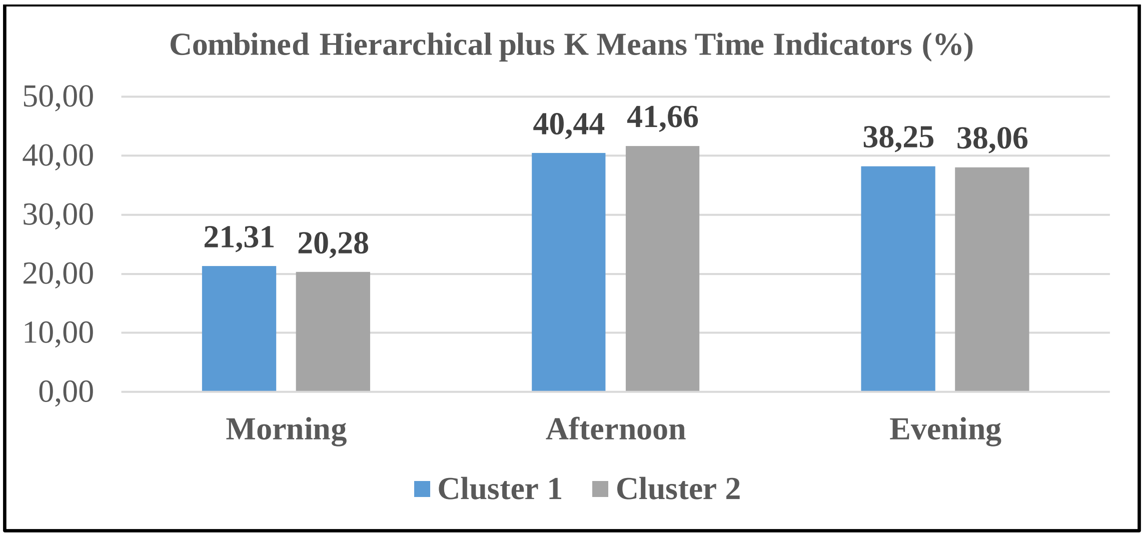

Figure 5.2 The Data Visualization of Time Indicator from Combined K-Means with K = 2

Based on Table 5.2 and Figure 5.2 it can be seen that most of the customers are willing to go for shopping in the afternoon where 41.66% of them placed at Cluster 2, and 40.44% of them set at cluster 1. It means that this clustering result can be a basis for Pamella Satu Supermarket Yogyakarta to give excellent service in the afternoon. The researcher assumed that Pamella Satu Supermarket Yogyakarta could boost their product sales, then the company should be focusing to attract people or the potential customer to come to the shop in the afternoon or giving more marketing strategy to attract customer to visit in the morning as the product sales contribution in the morning is lesser than another time.

C. Category Indicator as the third variable in Combined K Means Result with K = 2

The product category indicator as the third variable in Combined K Means Clustering Result with K = 2 can be shown in Table 5.3 and Figure 5.3.

Table 5.3 The product indicator in Combined K-Means Clustering Result with K = 2

Variable 3	Indicator	Category	Cluster 1 (%)	Cluster 2 (%)
CategoryIND	1	BABY & KIDS	4.55	37.78
	2	HAIR CARE	58.21	6.67
	3	SOAP	37.24	55.56
	Total (%)		**100**	**100**

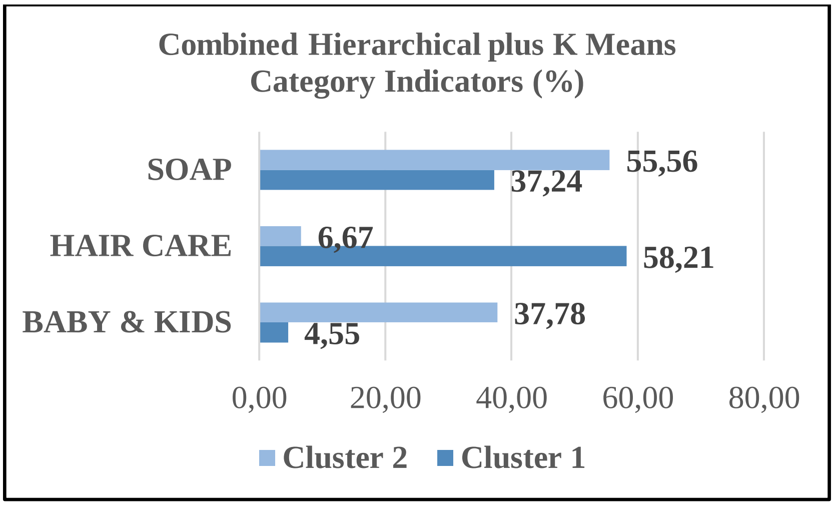

Figure 5.3 The data visualization of product indicator as the third variable in K Means Result with K = 2

As shown in Table 5.3 and Figure 5.3, there are 3 Category Indicators which are BABY & KIDS, HAIR CARE, and SOAP. The researcher found that

1. Cluster 1: most of the customers are willing to buy HAIR CARE
2. Cluster 2: most of the customers are eager to purchase SOAP

To increase the product sales, then Pamella Satu Supermarket Yogyakarta needs to have the ready stock for these two products and concern on the BABY & KIDS Category either because this kind of product category has the low contribution in the product selling.

D. Sub Category Indicator as the fourth variable in Combined Hierarchical plus K Means Result with K = 2

The sub product category indicator as the fourth variable in K Means Clustering Result with K = 2 can be shown in Table 5.4 and Figure 5.4.

Table 5.4 The product sub category indicators in K Means Result with K = 2

Variable 4	**Indicator**	**SubCategory**	**Cluster 1 (%)**	**Cluster 2 (%)**
SubCategoryIND	1	ACNE SOAP PACK	0.00	0.11
	2	BABY COLOGNE	0.00	1.96
	3	BABY GIFT	0.00	0.44
	4	BABY HAIR & BODY CARE	0.00	4.42
	5	BABY HAIR CARE	0.00	0.13
	6	BABY OIL & LOTION	0.00	4.48
	7	BABY ORAL CARE	0.00	4.31
	8	BABY SOAP BAR	0.00	3.19
	9	BABY SOAP LIQUID	0.00	2.48
	10	BEAUTY LIQUID	0.00	4.60
	11	BODY SOAP BAR	0.00	18.20
	12	BODY SOAP LIQUID	0.00	11.21
	13	CONDITIONER	0.08	4.12
	14	CREAMBATH & HAIR MASK	0.00	2.55
	15	DIAPERS & WIPES	0.01	16.38
	16	FACIAL SOAP MEN	0.00	5.15
	17	FACIAL SOAP WOMEN	0.01	14.28
	18	FEMININE WASH	0.00	1.99
	19	HAIR COLORING	2.85	0.00
	20	HAIR NUTRITION	6.16	0.00
	21	HAIR STYLING	3.09	0.00
	22	HAND SOAP	3.95	0.00
	23	HEALTHY SOAP BAR	22.32	0.00
	24	HEALTHY SOAP LIQUID	10.96	0.00
	25	PACIFIER	1.07	0.00
	26	SHAMPOO	42.11	0.00
	27	SHAVER	3.92	0.00
	28	TALCUM POWDER	3.47	0.00
	Total (%)		**100**	**100**

As shown in Table 5.4, there are 28 product sub categories that the researcher intends to conduct clustering analysis. The most wanted product sub category that the customer buy is as following:

1. Cluster 1: SHAMPOO with its sales contribution approximately 42.11%
2. Cluster 2: BODY SOAP BAR with its sales contribution around 18.20%

Further investigation shows that customers at cluster 1 tend to buy the adult product and customer at cluster 2 tend to buy the baby product. This information could be an insight to the owner on how to boost the sale products. The visualization of product sub categories indicator with K Means Result where K = 2 can be seen in Figure 5.7.

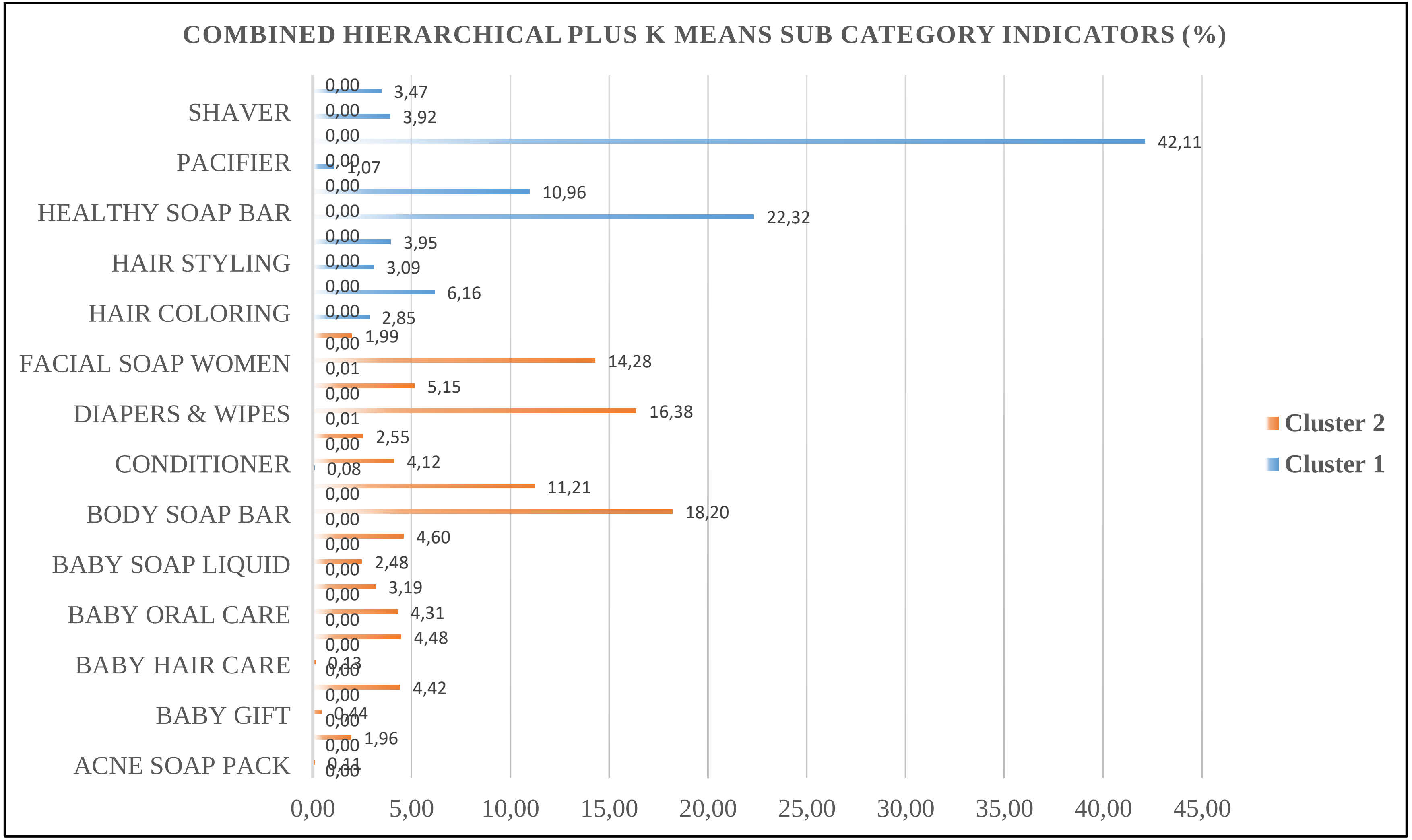

Figure 5.4 The data visualization of Product Sub Category Indicator in K-Means Clustering with K = 2

E. Sale Quantity as the fifth variable in Combined Hierarchical plus K Means Result with K = 2

The sale quantity as the fifth variable in K-Means Clustering Result with K = 2 can be seen in Table 5.5.

Table 5.8 The sale quantity as the fifth variable in K-Means Result with K = 2

Variable 5	Cluster 1	Cluster 2
SaleQty	39,102	39,933

As shown in Table 5.5, the researcher found that the highest contribution of product selling is placed in the cluster 2 although the difference is less significant. This information can be used by the owner to perform marketing strategy to increase their product sells in another cluster. For instance, Pamella Satu Supermarket Yogyakarta can give a promotion which uses marketing strategy of buy 1 get 1 for a specific product that can increase their Sale Quantity.

F. Price Indicator as the sixth variable in Combined Hierarchical plus K-Means Clustering Result with K = 2

The price indicator as the sixth variable in K-Means Clustering Result with K = 2 can be seen in Table 5.6 and Figure 5.5.

Table 5.6 The price indicator as the sixth variable in K Means Result with K = 2

Variable 6	Indicator	Price	Cluster 1	Cluster 2
PriceIND	1	< IDR 47,456	99.22	94.57
	2	< IDR 94,912	0.75	4.12
	3	< IDR 142,369	0.02	1.22
	4	IDR 189825	0.00	0.08
	Total		100	100

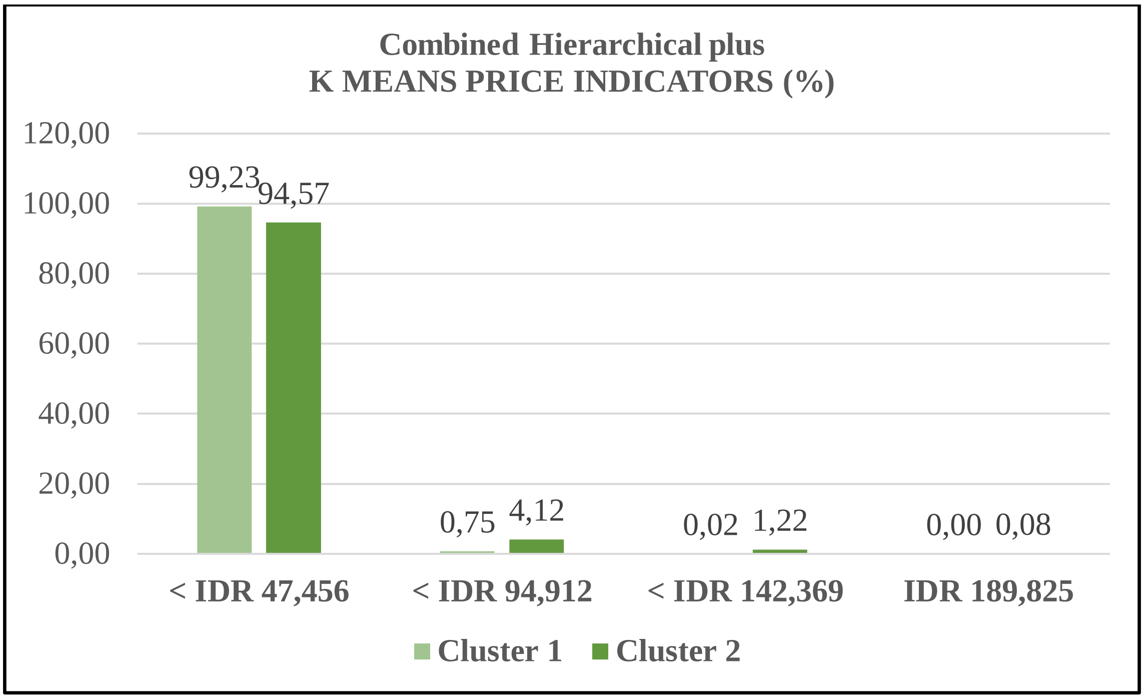

Figure 5.5 The data visualization of Price Indicator in K Means Result with K = 2

As shown in Table 5.6 and Figure 5.5, there are 4 price indicators which are Price Weight < 0.25, > 0.5, > 0.75 and = 1. It means that the price equal to price weight one is equal to IDR. 189.925. If the price weight equal to 0.06 it means that it will be equal to IDR 12,225 and less than IDR. 47.456. The researcher found that in both clusters most of the customers are willing to buy for the product price less or equal to IDR. 47,456. Customers at cluster 2 are willing to buy for the product in range 1-4 as well in comparison to the customers at cluster 1.

In general the characteristics of customers for each cluster from the combined clustering are as in Table 5.7.

Table 5.7 Characteristics of customers comparison

Variables	Cluster 1	Cluster 2
1	9,423 customers Customers are willing to go for shopping on Sunday 19.62%, Tuesday 17.21%, and Monday 17.61%. Customers are reluctant to go on Friday 14.95%	11,041 customers Customers are willing to go for shopping on Sunday 19.53%, Tuesday 18.56%, and Monday 16.38%. Customers are reluctant to go on Friday 13.88%

Variables	Cluster 1	Cluster 2
2	Customers are willing to go shopping in the afternoon 40.44 %, evening 38.25% and morning 21.31%	Customers are willing to go shopping in the afternoon 41.66%, evening 38.06% and morning 20.28%
3	Most of the customers are willing to buy HAIR CARE 58.21%, SOAP 37.24% and BABY & KIDS 4.55%	Most of the customers are willing to buy HAIR CARE 6.67%, SOAP 55.56% and BABY & KIDS 37.78
4	1. SHAMPOO with its sales contribution approximately 42.11% 2. Customers at cluster 1 tend to buy the adult product	1. BODY SOAP BAR with its sales contribution approximately 18.20% 2. Customers at cluster 2 tend to buy the baby product.
5	39,102 quantity	39,933 quantity
6	Customers are willing to pay/buy the product under or equal to IDR. 47,456	Customers are willing to pay/buy the product under range 1-4 particularly under or equal to IDR. 47,456

From Table 5.7 above, it seems that the most promising customers are from cluster 1.

5.1.2 Fuzzy C Means Clustering

The number of clusters used in this method is the same as the previous one. Therefore, the researcher can make a comparison of both approaches. The fuzzy clustering is considered as soft clustering, in which each element has a probability of belonging to each cluster. In other words, each item has a set of membership coefficients corresponding to the degree of being in a given cluster. This is different from k-means and k-medoid clustering, where each object is affected precisely to one cluster. In fuzzy clustering, points close to the center of a cluster may be in the cluster to a higher degree than points in the edge of a cluster. The degree, to which an element belongs to a given cluster, is a numerical value varying from 0 to 1. The fuzzy c-means (FCM) algorithm is one of the most widely used fuzzy clustering algorithms. The centroid of a cluster is calculated as the mean of all points, weighted by their degree of belonging to the cluster: Based on the computation, in cluster 1, there are 10,830 customers and in cluster 2 there are 9,634 customers. The profiling of Fuzzy C-Means Clustering with K = 2 will be described in the following sections.

A. Day Indicator

The first variable is Day Indicator, the profiling can be seen in Table 5.8 and Figure 5.6.

Table 5.8 Day Indicator profiling

Variable 1	Indicator	Day	Cluster 1 (%)	Cluster 2 (%)
DayIND	1	Monday	16.22	17.75
	2	Tuesday	18.53	17.27
	3	Wednesday	15.66	15.13
	4	Thursday	16.06	15.43
	5	Friday	13.87	14.94
	7	Sunday	19.66	19.47
	Total (%)		**100**	**100**

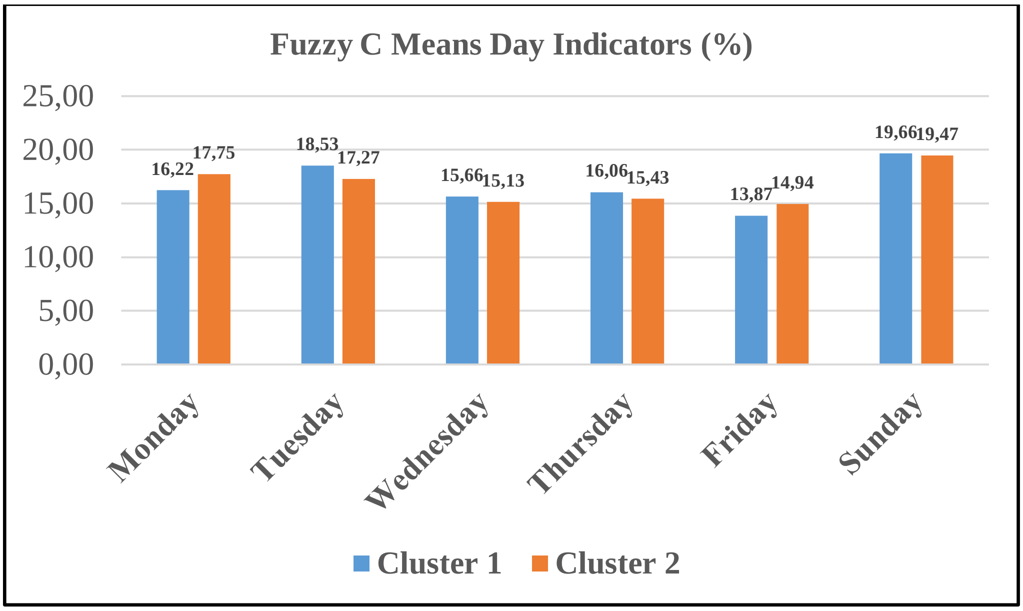

Figure 5.6 The data visualization of FCM Result with K = 2

As shown in Table 5.8 and Figure 5.6, the researcher found that most of the customers are willing to buy a product in the Pamella Satu Supermarket on Sunday where most of the customers belong to cluster 1 with 19.66% and cluster 2 with 19.47% from the total six days point of sales. The second most days for each cluster are Monday and Tuesday. The least customers in each cluster are on Friday. Based on these results then they can apply 3 marketing strategies as follows:

1. Improving the service for the most customers (Sunday)
2. Improving the service for the second most customer (Monday and Tuesday)
3. Improving the service on the least customers (Friday)

It can be done by managing their employee working hour and try to boost their product sales by giving promotion to a specific product on Friday, Monday, and Tuesday to attract customers to the Pamella Satu Supermarket Yogyakarta.

B. Time Indicator

The second variable profiling is Time Indicator can be seen in Table 5.9 and Figure 5.7.

Table 5.9 The time indicator profiling

Variable 2	Indicator	Time	Cluster 1 (%)	Cluster 2 (%)
TimeIND	1	Morning	20.34	21.22
	2	Afternoon	41.75	40.37
	3	Evening	37.90	38.42
	Total (%)		**100**	**100**

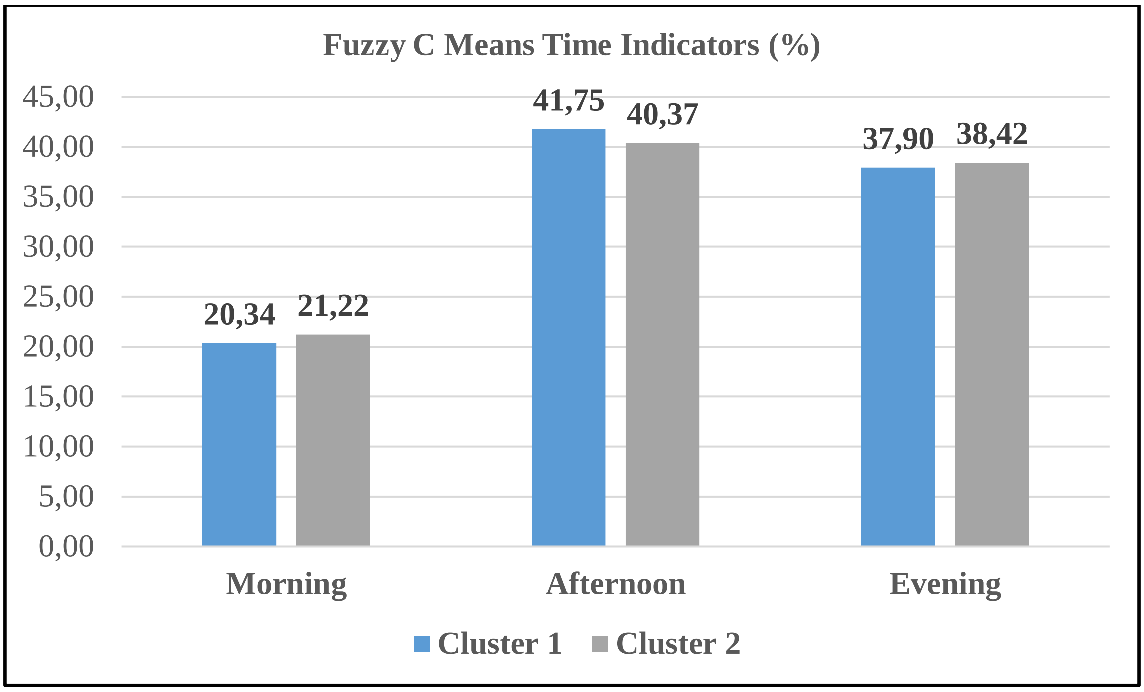

Figure 5.7 The data visualization of Time Indicator using FCM with K = 2

Table 5.9 and Figure 5.7 show that most of the customers are willing to go shopping in the afternoon (41.75%). These result already similar to the previous Combined Hierarchical plus K Means. It means not only Pamella Satu Supermarket Yogyakarta should concern on how to increase product sales in the morning and evening but also to maintain the service in the afternoon such that they can get more customers and customer satisfaction.

C. Category Indicator

The profiling of category indicator as the third variable can be seen in Table 5.10 and Figure 5.8.

Table 5.10 Category indicator profiling

Variable 3	Indicator	Category	Cluster 1 (%)	Cluster 2 (%)
CategoryIND	1	BABY & KIDS	38.52	4.44
	2	HAIR CARE	6.87	56.85
	3	SOAP	54.61	38.71
	Total (%)		**100**	**100**

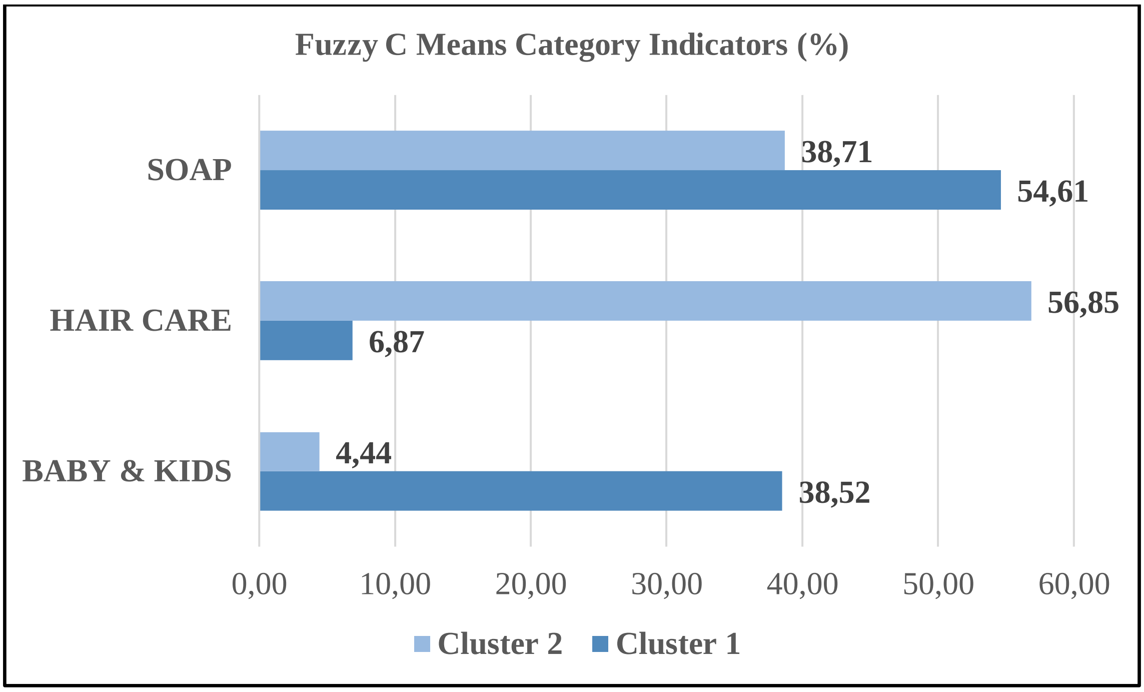

Figure 5.8 The data visualization of category indicator using FCM

As shown in Table 5.10 and Figure 5.8, there are 3 Category Indicators which are BABY & KIDS, HAIR CARE, and SOAP. The researcher found that:

1. Cluster 1: most of the customers are willing to buy SOAP
2. Cluster 2: most of the customers are eager to purchase HAIR CARE.

To increase the product sales, then Pamella Satu Supermarket Yogyakarta needs to have the ready stock for these two products and concern on the BABY & KIDS Category because this kind of product category has the minor contribution in the product selling.

D. Sub Category Indicator

The profiling sub category indicator as the fourth variable are in Table 5.11.

Table 5.11 The sub category profiling

Variable 4	Indicator	SubCategory	Cluster 1 (%)	Cluster 2 (%)
SubCategoryIND	1	ACNE SOAP PACK	0.11	0.00
	2	BABY COLOGNE	1.99	0.00
	3	BABY GIFT	0.45	0.00
	4	BABY HAIR & BODY CARE	4.51	0.00
	5	BABY HAIR CARE	0.13	0.00
	6	BABY OIL & LOTION	4.57	0.00
	7	BABY ORAL CARE	4.40	0.00
	8	BABY SOAP BAR	3.25	0.00
	9	BABY SOAP LIQUID	2.53	0.00
	10	BEAUTY LIQUID	4.69	0.00
	11	BODY SOAP BAR	18.55	0.00
	12	BODY SOAP LIQUID	11.43	0.00
	13	CONDITIONER	4.28	0.00
	14	CREAMBATH & HAIR MASK	2.59	0.00
	15	DIAPERS & WIPES	16.70	0.00
	16	FACIAL SOAP MEN	5.25	0.00
	17	FACIAL SOAP WOMEN	14.56	0.01
	18	FEMININE WASH	0.00	2.28
	19	HAIR COLORING	0.00	2.79
	20	HAIR NUTRITION	0.00	6.02
	21	HAIR STYLING	0.00	3.02
	22	HAND SOAP	0.00	3.86
	23	HEALTHY SOAP BAR	0.00	21.83
	24	HEALTHY SOAP LIQUID	0.00	10.72
	25	PACIFIER	0.00	1.05
	26	SHAMPOO	0.00	41.19
	27	SHAVER	0.00	3.83
	28	TALCUM POWDER	0.00	3.39
	Total (%)		**100**	**100**

As shown in the Table 5.11, there are 28 product sub categories that the researcher intends to conduct clustering analysis. The most wanted product sub category that the customer buy is:

1 Cluster 1: BODY SOAP BAR with its sales contribution approximately 18.55%

2 Cluster 2: SHAMPOO with its sales contribution around 41.19%

Further investigation shows that customers at cluster 1 tend to buy the baby product and customer at cluster 2 tend to buy the adult product. This information could be an insight to the owner on how to boost the product selling. The visualization of product sub categories indicator with Fuzzy C Means Result where K = 2 can be seen in Figure 5.9.

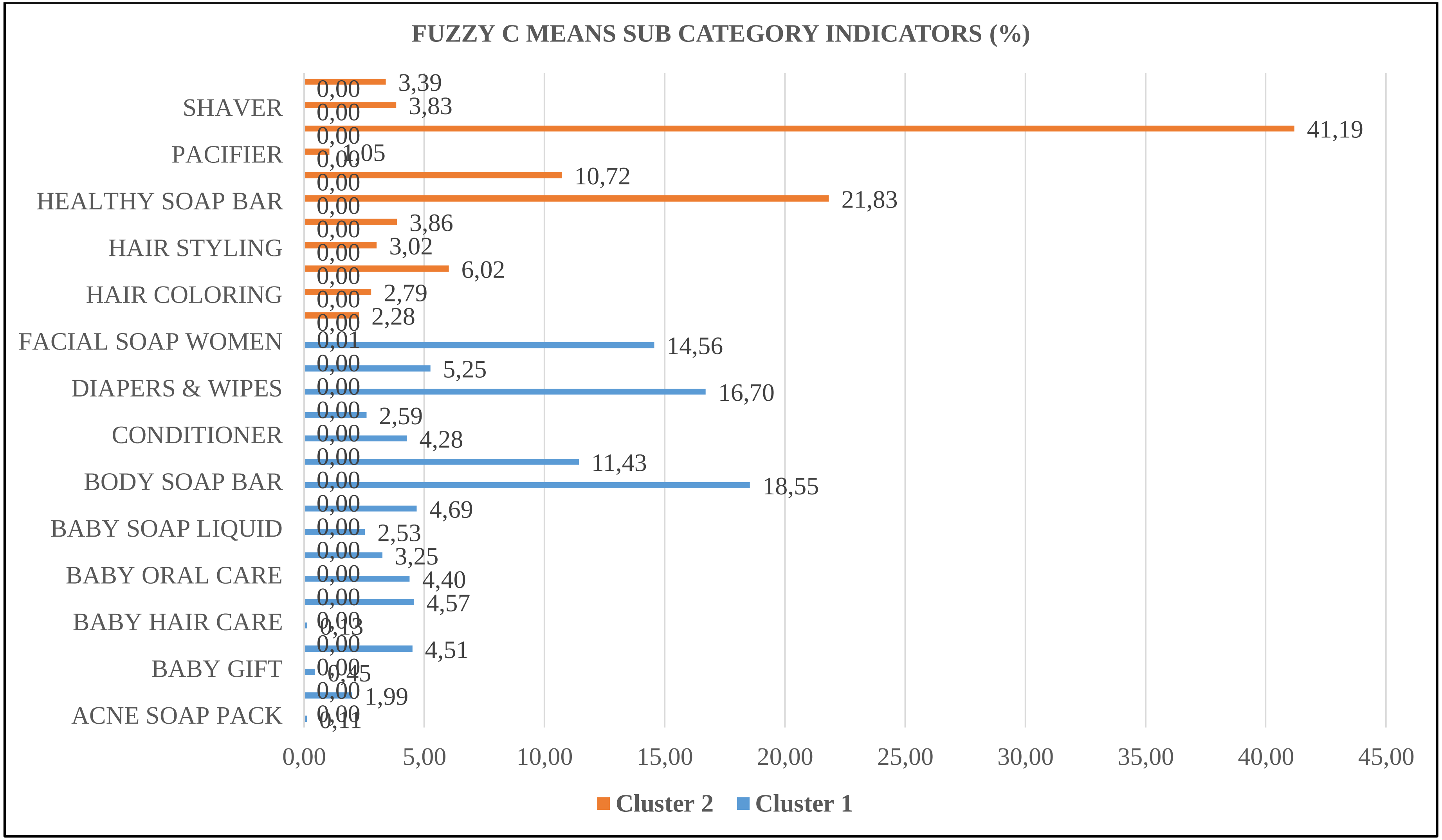

Figure 5.9 The data visualization of product sub category indicator using Fuzzy C Means Clustering with K = 2

E. Sale Quantity2

The profiling of sale quantity as the fifth variablecan be seen in Table 5.12.

Table 5.12 Sales quantity profiling

Variable 5	Cluster 1	Cluster 2
SaleQty	39,933	37,640

As shown in Table 5.12, the researcher found that the highest contribution of product selling is placed in the Cluster 1. The difference is around 2,293 sales. By comparing to the customer for each cluster, then the Pamella Satu Supermarket Yogyakarta can give more attention to customers at cluster 2 to increase their Sale Quantity.

F. Price Indicator

The price indicator profiling can be seen in Table 5.13 and Figure 5.10.

Table 5.13 The price indicator profiling

Variable 6	Indicator	Price	Cluster 1	Cluster 2
PriceIND	1	< IDR 47,456	94.47	99.24
	2	< IDR 94,912	4.20	0.74
	3	< IDR 142,369	1.25	0.02
	4	IDR 189,825	0.08	0.00
	Total (%)		**100**	**100**

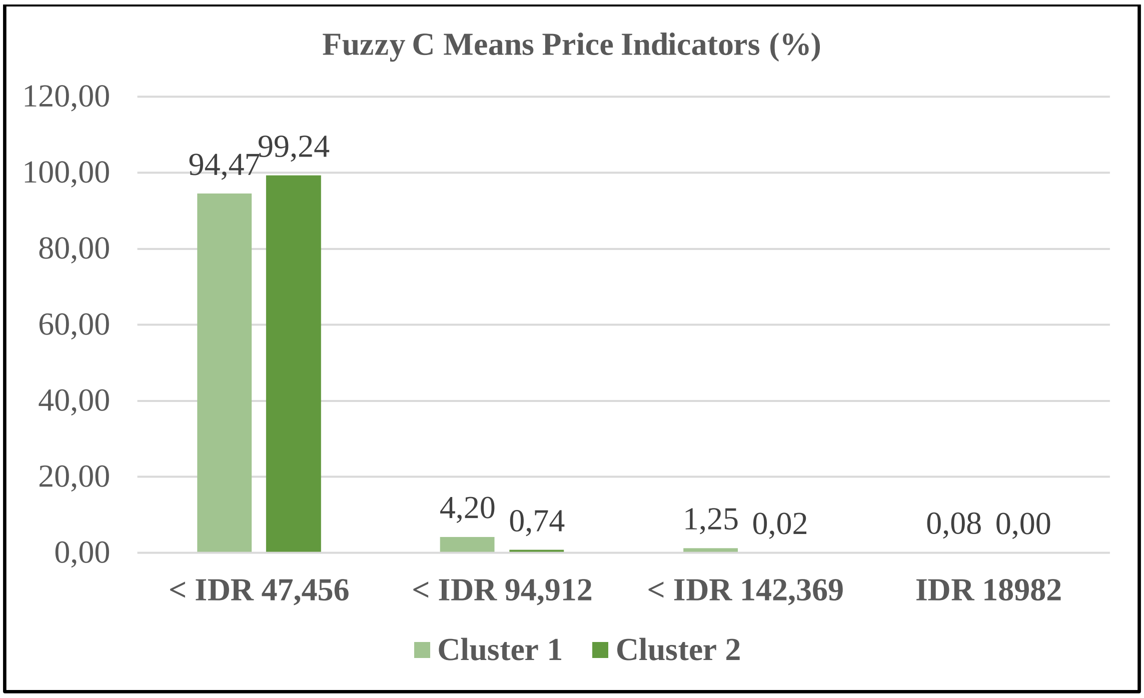

Figure 5.10 Data visualization of Price Indicator using FCM with K = 2

Table 5.13 and Figure 5.10 show that, there are 4 price indicators: Price Weight < 0.25, Price Weight > 0.5, Price Weight > 0.75, and = 1. It means that the price equal to price weight one is equal to IDR. 189,925. If the price weight equal to 0.06 it means that it will be equal to IDR. 12,225 and less than IDR. 47,456. The researcher found that in both clusters most of the customers are willing to buy for the product price less or equal to IDR. 47,456. Customers at cluster 2 are willing to buy for the product in range 1- 4 as well in comparison to the customers at cluster 1. The characteristics of customers for each cluster from the FCM are in Table 5.14.

Table 5.14 Characteristics of customers comparison

Variables	Cluster 1	Cluster 2
0	10,830 customers	9,634 customers
1	Customers are willing to go for shopping on Sunday 19.66%, Tuesday 18.53% and Monday 16.22% Customers are reluctant to go on Friday 13,87%	Customers are willing to go for shopping on Sunday 19.47%, Monday 17.77% and Tuesday 17.27% Customers are reluctant to go on Friday 14,94%

Variables	Cluster 1	Cluster 2
2	Customers are willing to go shopping in the afternoon where 41.75%, evening 37.90% and morning 20.34%	customers are willing to go shopping in the afternoon 40,37%, evening 38.42% and morning 21.22%
3	Most of the customers are willing to buy SOAP 54.61%, BABY & KIDS 38.52% and HAIR CARE 6.87%	Most of the customers are willing to buy HAIR CARE 56.85%, SOAP 38.71% and BABY & KIDS 4.44%
4	1. BODY SOAP BOAR with its sales contribution approximately 18.55% 2. Customers at cluster 1 tend to buy the baby product	1. SHAMPOO with its sales contribution approximately 41.19% 2. Customers at cluster 2 tend to buy the adult product.
5	39,933 quantity	37,640 quantity
6	Customers are willing to pay/buy the product under or equal to IDR. 47,456	Customers are willing to pay/buy the product under range 1-4 particularly under or equal to IDR. 47,456

From Table 5.14 above, it seems that the most promising customers are from cluster 1.

5.1.3 Comparison of Combined Hierarchical plus K-Means Clustering and Fuzzy C-Means Clustering

A. Processing Time and BSS/TSS Ratio Result Comparison

In the study, the researcher would like to choose the best clustering algorithm based on BSS/TSS ratio and processing time. Therefore, the researcher has conducted 100 simulation for both methods. The results can be seen in appendix 20.

Based on Appendix A, it can be seen that FCM has the best result regarding to its processing time and BSS/TSS ratio. It means that in the average its BSS/TSS ratio is 59.34% and it indicates a good fit. It means it has 59.34% low similarity between clusters. The average processing time for FCM is 63,23. Therefore, the researcher concluded that FCM is the best clustering algorithm to perform to support a correct marketing strategy at Pamella Satu Supermarket Yogyakarta rather than Combined Hierarchical plus K Means Clustering Algorithm.

B. Comparison of each variables based on both clustering result

The researcher only includes indicators Sunday, Monday, Tuesday, Wednesday, Thursday and Friday in Day Variable due to ram limitations. Therefore, six days of sales transaction data is taken from 01 October to 06 October 2017. For this reason, researchers can not include Saturday in the analysis when it should be the peak day for the customer to shop at the Pamella Satu Supermarket Yogyakarta. Although Saturday is not included in the analysis, the researcher found a new valuable insight to become the basis by the business owner to formulate the correct marketing strategy in their business. The researcher found that the similar cluster that can be compared each other is Cluster 1 in Combined Hierarchical plus K Means and Cluster 2 in FCM clustering result. Otherwise, the result of Cluster 2 in combined hierarchical plus K Means and Cluster 1 in FCM` has the similar effect based on six variables comparison. The detail of the similarity of each variable based on both clustering result can be seen in Table 5.16.

Table 5.16 The comparison of each variables based on both clustering result

Variable	**Combined Hierarchical and K-Means**		**Fuzzy C Means**	
	Cluster 1	**Cluster 2**	**Cluster 1**	**Cluster 2**
1	Customers are willing to go for shopping on Sunday 19.62%, Tuesday 17.21% and Monday 17.61%. Customers are reluctant to go on Friday 14.95%	Customers are willing to go for shopping on Sunday 19.53%, Tuesday 18.56% and Monday 16.38%. Customers are reluctant to go on Friday 13.88%	Customers are willing to go for shopping on Sunday 19.66%, Tuesday 18.53% and Monday 16.22%. Customers are reluctant to go on Friday 13,87%	Customers are willing to go for shopping on Sunday 19.47%, Monday 17.77% and Tuesday 17.27%. Customers are reluctant to go on Friday 14,94%
2	Customers are willing to go for shopping in the afternoon where 40,44% , evening 38.25% and morning 21.31%	Customers are willing to go for shopping in the afternoon 41,66%, evening 38.06% and morning 20.28%	Customers are willing to go for shopping in the afternoon where 41.75%, evening 37.90% and morning 20.34%	Customers are willing to go for shopping in the afternoon 40,37%, evening 38.42% and morning 21.22%

Variable	Cluster 1	Cluster 2	Cluster 1	Cluster 2
3	Most of the customers are willing to buy HAIR CARE 58.21%, SOAP 37.24% and BABY & KIDS 4.55%	Most of the customers are willing to buy HAIR CARE 6.67%, SOAP 55.56% and BABY & KIDS 37.78	Most of the customers are willing to buy SOAP 54.61%, BABY & KIDS 38.52% and HAIR CARE 6.87%	Most of the customers are willing to buy HAIR CARE 56.85%, SOAP 38.71% and BABY & KIDS 4.44%
4	1. SHAMPOO with its sales contribution approximately 42,11% 2. Customers at cluster 1 tend to buy adult product	1. BODY SOAP BAR with its sales contribution approximately 18,20% 2. Customer at cluster 2 tend to buy baby product.	1. BODY SOAP BOAR with its sales contribution approximately 18.55% 2. Customers at cluster 1 tend to buy baby product	1. SHAMPOO with its sales contribution approximately 41.19% 2. Customer at cluster 2 tend to buy adult product.
5	39,102 quantity	39,933 quantity	39,933 quantity	37,640 quantity
6	99.23 % Customers are willing to pay/buy the product under or equal to IDR. 47.456	94,57% Customers are willing to pay/buy the product under range 1-4 particularly under or equal to IDR. 47.456	94,47% Customers are willing to pay/buy the product under or equal to IDR. 47.456.	99.24% Customers are willing to pay/buy the product under range 1-4 particularly under or equal to IDR. 47.456.

In Table 5.16, it can be seen that there are three variables, namely Time, Sub Category and Price that have similar cluster result. On the other hand, the other three variables also have similar characteristics with the combination of Cluster 1 in combined hierarchical plus K Means and Cluster 2 in Fuzzy C Means. Otherwise, Cluster 2 in combined hierarchical plus K Means with Cluster 1 in Fuzzy C Means Clustering result has the similar characteristics either. Therefore, in the study, the researcher defined that the combination of Cluster 1 in combined hierarchical plus K Means and Cluster 2 in Fuzzy C Means as the Customer cluster A. Besides, the combination of Cluster 2 in combined hierarchical plus K Means with Cluster 1 in Fuzzy C Means clustering result as the Customer cluster B.

By renaming the cluster of customers to clusters A and B, then the researcher can perform a fair comparison such that the results can support the correct marketing strategy for the retail company. The comparison of each cluster are in Table 5.17.

Table 5.17 The cluster comparison

Variable	**Customer Cluster A**			**Customer Cluster B**			**The Best Customer Cluster**
	Cluster 1 (Combined K Means) 9,423	**Cluster 2 (FCM) 9,634**	**Average 9,528.5**	**Cluster 2 (Combined K Means) 11,041**	**Cluster 1 (FCM) 10,830**	**Average 10,935.5**	
1	Sunday: 19.62%	Sunday:19.47%	Sunday: 19.545%	Sunday: 19.53%	Sunday: 19.66%	Sunday: 19.595%	**Cluster A**
	Tuesday: 17.21%	Tuesday: 17.27%	Tuesday: 17.24%	Tuesday: 18.56%	Tuesday: 18.53%	Tuesday: 18.545%	
	Monday: 17.61%	Monday: 17.77%	Monday: 17.69%	Monday: 16.38%	Monday: 16.22%	Monday: 16.3%	
	Friday: 14.95%	Friday: 14.94%	Friday: 14.945%	Friday: 13.88%	Friday: 13.87%	Friday: 13.875%	
2	Afternoon: 40.44%	Afternoon: 40.37%	Afternoon: 40.405%	Afternoon: 41.66%	Afternoon: 41.75%	Afternoon: 41.705%	**Cluster B**
	Evening: 38.25%	Evening: 38.42%	Evening: 38.335	Evening: 38.06%	Evening: 37.90%	Evening: 38.25%	
	Morning: 21.31%	Morning: 21.22%	Morning: 21.265%	Morning: 20.28%	Morning: 20.34%	Morning: 20.31%	
3	Hair Care: 58.21%	Hair Care: 56.85%	Hair Care:57.53%	Hair Care: 6.67%	Hair Care: 6.87%	Hair Care: 6.77%	**Cluster A**
	Soap: 37.24%	Soap: 38.71%	Soap: 37.975%	Soap: 55.56%	Soap: 54.61%	Soap: 55.085%	
	Baby & Kids: 4.55%	Baby & Kids: 4.44%	Baby & Kids: 4.495%	Baby & Kids: 37.78%	Baby & Kids: 38.52%	Baby & Kids: 38.15%	
4	Shampoo: 42.11%	Shampoo: 41.19%	Shampoo: 41.65%	Body Soap Bar: 18.20%	Body Soap Bar: 18.55%	Body Soap Bar: 18.375%	**Cluster A**
5	39102 units	37640 units	38371 units	39933 units	39933 units	39993 units	**Cluster B**
6	Price Weight 1: 99.23%	Price Weight 1: 99.24%	Price Weight 1: 99.235%	Price Weight 1: 94.57%	Price Weight 1: 94.47%	Price Weight 1: 94.52%	**Cluster A**

As shown in the Table 5.17, it can be seen that Customer Cluster A has the higher contribution in the product selling at Pamella Satu Supermarket Yogyakarta. The reason is can be seen at Table 5.18

Table 5.18 Reasoning the best cluster customer

Variable	**Reasoning**
1	People tend to go shopping on the weekend and the percentage of people who are shopping on weekend higher in cluster A than in cluster B
2	People tend to go shopping in the afternoon and evening and the average percentage of people who are shopping in those time in cluster B higher than in cluster A
3	Customer in cluster A tend to buy HAIR CARE and customer in cluster B tend to buy SOAP. Concerning to the price of these products, then the contribution of cluster cluster A higher than in cluster B
4 & 5	Customer in cluster A tend to buy SHAMPOO and customer in cluster B tend to buy BODY SOAP BAR. Concerning to the price of these products, then the contribution of costumer cluster A higher than in cluster B
6	The products price at Pamella Satu Supermarket Yogyakarta mostly at Price Weight 1. Therefore, customer cluster A more meet the objective of the Supermarket market shares than cluster B

Based on the analysis at Table 5.18, then the researcher recommend to the owner to pay attention more to customers in cluster A than in cluster B. Customers in cluster A is is part of Cluster 1 in combined hierarchical plus K Means and Cluster 2 in Fuzzy C Means Clustering. This cluster is a right customers who will contribute better to the supermarket product sale and will need a correct promotion strategy.

C. The business strategies suggestion to the business owner

Based on the previous explanation, the researcher finally can formulate the business suggestion strategies to the business owner, as in Table 5.19.

Table 5.19 The business strategies suggestion to the business owner

Variable	Business Strategy Suggestion
1	The business owner can apply three marketing strategies such as improving the service on the most customers (Sunday), improving the service on the second most customer (Monday and Tuesday), and improving the service on the least customers (Friday). It can be done by managing their employee working hour and try to boost their product sales by giving promotion to a specific product on Friday, Monday, and Tuesday to attract customers to the Pamella Satu Supermarket Yogyakarta.
2	The clustering result shows that it can be the basis for the business owner to concern on how to increase product sales in the morning and maintain the service in the afternoon and evening in the order they can get the customer satisfaction by giving the excellence service and the product stock are ready at that time.
3	Pamella Satu Supermarket Yogyakarta need to have the ready stock of HAIR CARE (for its highest contribution in the product selling in customer cluster A) and BABY & KIDS (because this product category has the least contribution to the product selling among another product categories).
4	Further investigation on Customer Cluster A result shows that customers tend to buy adult products. Therefore, the business owner needs to give attractive promotion to adult sub category products.
5	This information can be used by the owner to perform marketing strategy to increase their product sells in another cluster which has the minor contribution to product selling. For instance, Pamella Satu Supermarket Yogyakarta can give a promotion which uses marketing strategy of buy 1 get 1 for the specific product that can increase their Sale Quantity.
6	The owner need to pay attention on product pricing. Most of the customers of cluster A are willing to buy products at a price less or equal to IDR. 47,456,-

Based on Table 5.19, the reason why the researcher proposed the business strategy above can be seen in Table 5.20.

Table 5.20 The reason of the business strategies suggestion to the business owner

Variable	Reason
1	a. Based on the clustering result on Day variable, the researcher found that Sunday, Monday, and Tuesday are having the significant contribution to the product sales of Pamella Satu Supermarket Yogyakarta. Therefore, the business suggestion can be the basis to formulate correct marketing strategy to boost up the product sales on that Days. b. Otherwise, Friday has the least contribution regarding product sales at Pamella Satu Supermarket Yogyakarta. Although the input is not significant, the owner needs to attract more customer by giving correct promotion strategy on Friday to make their business more feasible. c. In response with the clustering result, then the researcher suggested that Pamella Satu Supermarket Yogyakarta can be focusing on treating the customer well on a significant day by managing their employee working hour and giving promotion to a specific product.
2	Given the fact from the clustering result that more customer is willing to come to Pamella Satu Supermarket in the afternoon, then the business owner needs to maintain their retail service in the afternoon and evening. Therefore, they will achieve more customer satisfaction in doing excellence service, and the product stock is ready at that time.
3	a. Based on the clustering analysis, the researcher found that Customer Cluster A has the significant contribution in buying the product at Pamella Satu Supermarket Yogyakarta. Therefore, the researcher suggested that the business owner need to have the ready stock of HAIR CARE Product Category because it has the highest contribution to the product selling in Customer Cluster A. b. However, given the fact that BABY & KIDS Product Category has the least contribution to the product selling among another product categories in Customer Cluster A, then the business owner need to make attractive promotion such as Buy 1 Get 1 for this BABY & KIDS product Category to boost its product sales at the store.

Variable	Reason
4	Based on the clustering analysis, the researcher found that customer cluster A tend to buy adult products. Therefore, the business owner need to give attractive promotion to adult sub category products such as SHAMPOO and BODY SOAP BAR that have the major contribution in the product sales at Pamella Satu Supermarket Yogyakarta.
5	Given the fact from the clustering result, the researcher found that the sales quantity that already being sold in the Pamella approximately around 37,640 to 39,102 products within six days of operation. Therefore, this information can be the basis for the business owner to increase their performance for selling product by giving promotion buy 1 get 1 for the specific product sales that can improve their Sale Quantity. Other than that, along with increasing the sales quantity, the impact would make the customer buy another product that the location near with the product that already given the promotion.
6	Based on the clustering analysis, most of the customers of customer cluster A are willing to buy products at a price less or equal to IDR. 47,456,-. Therefore, the business owner needs to pay attention to price decision to attract customer to buy the product with the price up to IDR. 47,456.

5.2. Association Rules – Market Basket Analysis

There are 23 rules where the rule length distribution sized in the third item is 7 and fourth item is 16. It means that a length of 4 items has the most rules and the range of 3 items have the least rules. To make it comprehensive, the researcher would like to analyze several results in the Association Rules as one example shown in the Figure 5.11.

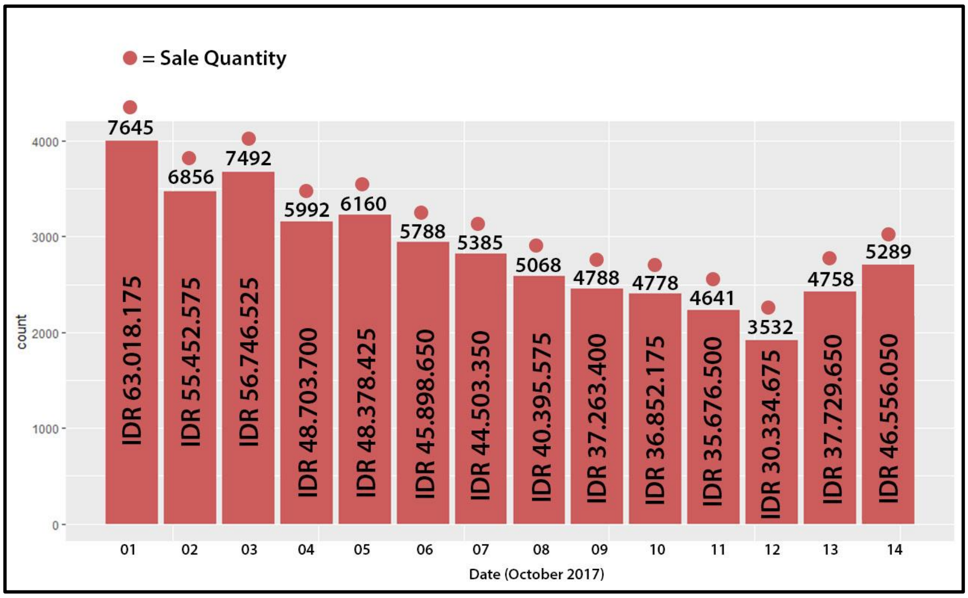

Figure 5.11 The visualization to show the customer amount spend and sale quantity on each date at Pamella Satu Supermarket Yogyakarta

Figure 11 shows about the product sales at Pamella Satu Supermarket Yogyakarta from 1st October to 14th October 2017. The most the peak time for selling product happen on 01st October 2017 which is on Sunday, where the customer amount spent reach IDR 63,018,175, and the sales quantity reaches 7,645 products. The highest selling point is happened between 1st October to 6th October 2017 and the lowest one is on 125h October 2017.

To increase the product selling then the effectiveness of marketing strategy and sales tactics could be improved based on the results from ARMBA. Figure 5.12 show the ten prioritize rules which can be proposed to the business owner. The detail of top 10 rules can be seen in the Table 5.21.

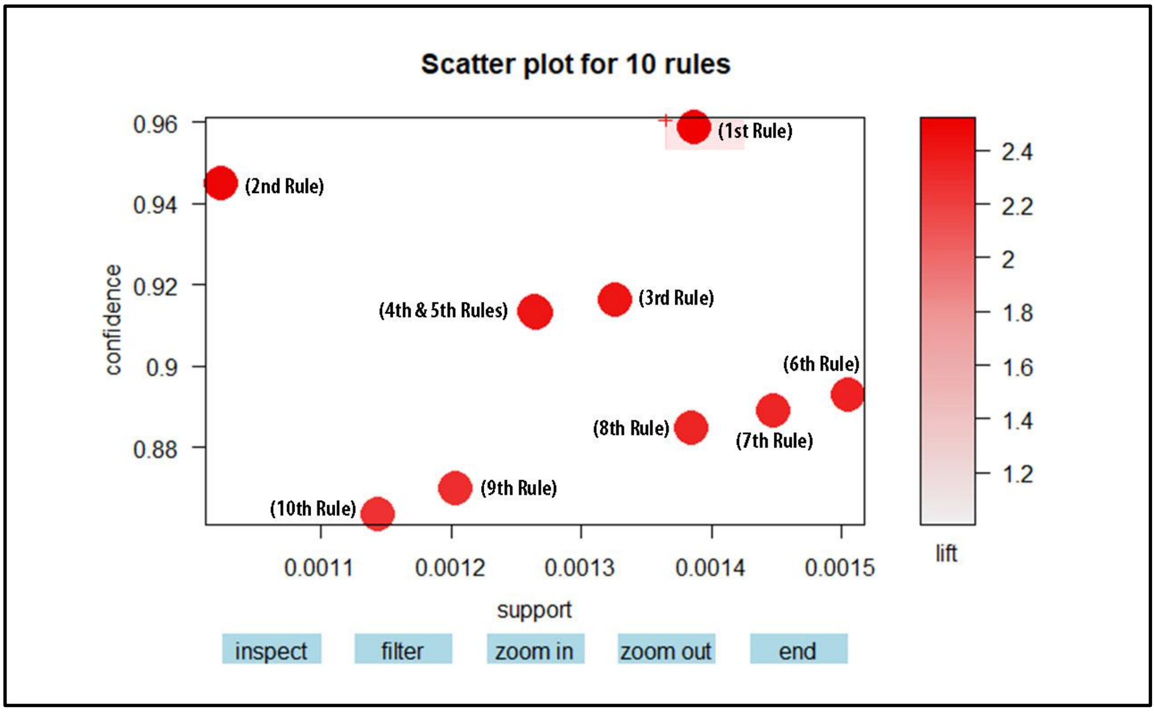

Figure 5.12 The data visualization to plot the Top 10 Rules

Table 5.21 The top 10 Rules prioritize

No	LHS	RHS	Support	Confidence	Lift	Count
1	BABY SOAP LIQUID, CONDITIONER	SHAMPOO	0.001386043	0.9583333	2.512.653	23
2	CONDITIONER, DIAPERS & WIPES, FACIAL SOAP WOMEN	SHAMPOO	0.001024467	0.9444444	2.476.238	17
3	CONDITIONER, DIAPERS & WIPES, HAIR NUTRITION	SHAMPOO	0.001325780	0.9166667	2.403.408	22
4	BODY SOAP BAR, BODY SOAP LIQUID, HAIR NUTRITION	SHAMPOO	0.001265518	0.9130435	2.393.908	21
5	DIAPERS & WIPES, HAIR NUTRITION, HEALTHY SOAP BAR	SHAMPOO	0.001265518	0.9130435	2.393.908	21
6	BODY SOAP LIQUID, HAIR NUTRITION, HEALTHY SOAP BAR	SHAMPOO	0.001506569	0.8928571	2.340.981	25
7	BODY SOAP LIQUID, CONDITIONER, HAIR NUTRITION	SHAMPOO	0.001446306	0.8888889	2.330.577	24
8	CONDITIONER, FACIAL SOAP WOMEN, HEALTHY SOAP BAR	SHAMPOO	0.001386043	0.8846154	2.319.372	23
9	FACIAL SOAP WOMEN, HEALTHY SOAP BAR, SHAVER	SHAMPOO	0.001205255	0.8695652	2.279.912	20
10	BABY HAIR & BODY CARE, FACIAL SOAP WOMEN, HEALTHY SOAP LIQUID	SHAMPOO	0.001144992	0.8636364	2.264.367	19

Table 5.20 shows that to increase the product sales then the business owner of Pamella Satu Supermarket Yogyakarta needs to put of top 10 left-hand side products near shampoo on their store in the order they can boost their product selling. The 10 left-hand side products are: (1) Baby Soap Liquid and Conditioner, (2) Conditioner, Diapers & Wipes, Hair Nutrition, (4) Body Soap Bar, Body Soap Liquid, Hair Nutrition, (5) Diapers & Wipes, Hair Nutrition, Healthy Soap Bar, (6) Body Soap Liquid, Hair Nutrition, Healthy Soap Bar, (7) Body Soap Liquid, Conditioner, Hair Nutrition, (8) Conditioner, Facial Soap Women, (9) Facial Soap Women, and (10) Baby Hair & Body Care, Facial Soap Women, Healthy Soap Liquid.

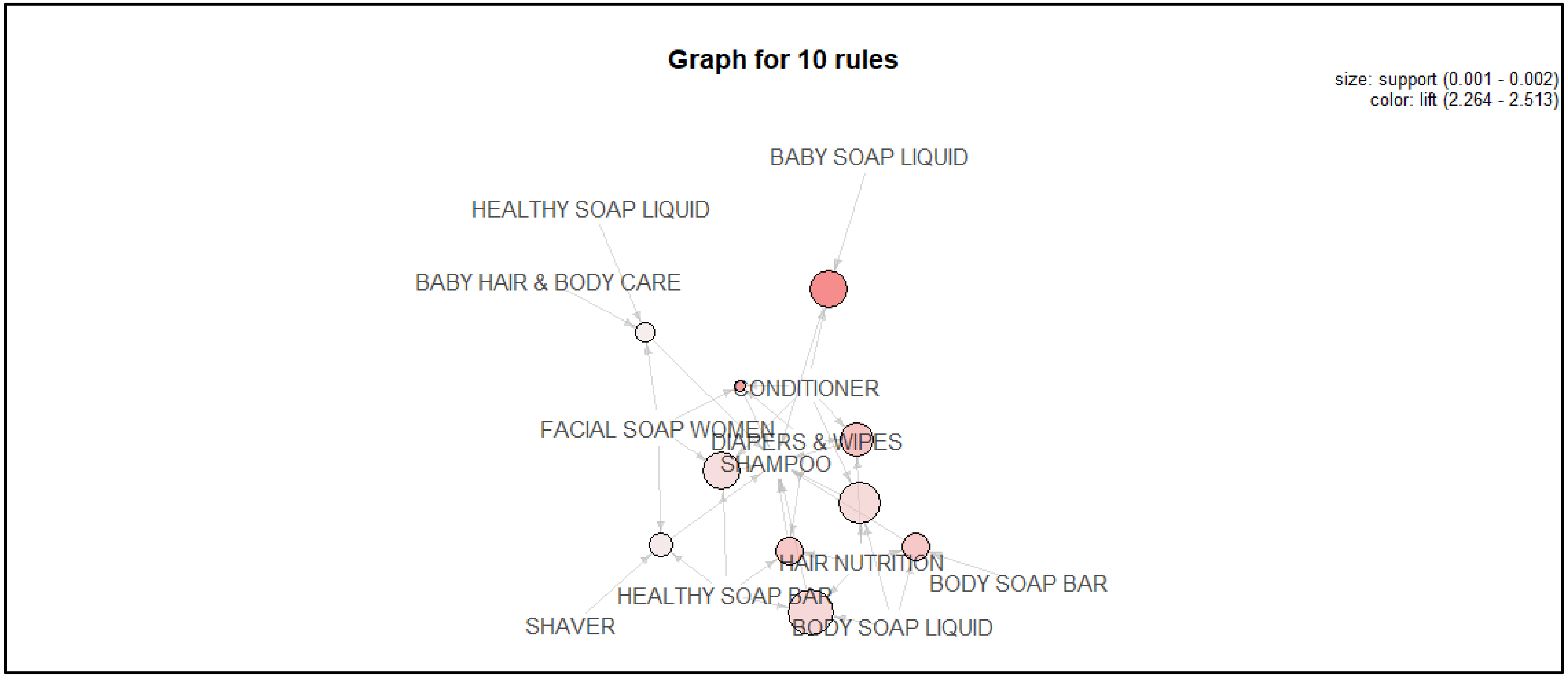

Figure 5.13 The graph of top 10 rules

Figure 5.13 suggests that Shampoo, Baby Soap Liquid, Diapers and Wipes, Hair Nutrition, Body Soap Bar, Body Soap Liquid, Healthy Soap Bar, Facial Soap Women, Baby Hair and Body Care and Conditioner should be put near each other. This is also can be seen in Figure 5.14.

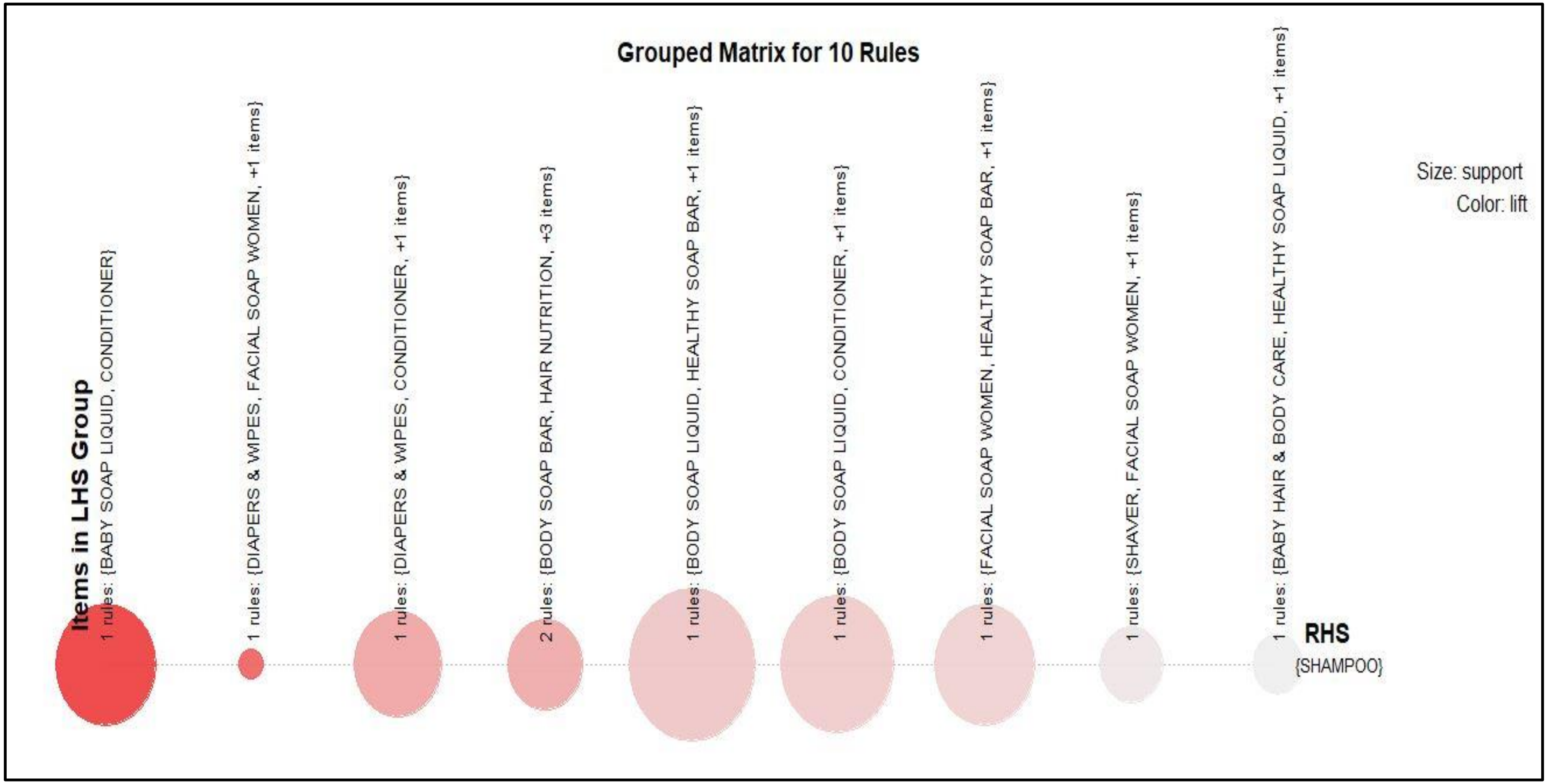

Figure 5.14 The group for 10 rules

In general, the suggestion to the business owner of Pamella Satu Supermarket Yogyakarta are:

1 Create/offer a membership card. The member card will make the owner could evaluate thoroughly the supply and demand at supermarket based on the loyal customers.

2 Cooperate with the bank to provide several attractive promotion such as discount, credit/delayed payment or others.

3 Points reward. Points reward tend to make customer buy more often to get some certain reward. Therefore, it makes the product sales increasing.

4 Re-arranged the layout of the products. Based on the ARMBA analysis, then we need to put products: Baby Soap Liquid, Conditioner, Diapers & Wipes, Hair Nutrition, Body Soap Bar, Body Soap Liquid, Healthy Soap Bar, Facial Soap Women, and Baby Hair & Body Care, near to SHAMPOO. Lay out other products can be put according to other rules.

5 Provide a catalog/software application. Online shopping tend to be a shopping trend among shoppers. Providing a catalogue (hard and or soft copies) will give a chance to customers plan their shopping in advance and will provide you a feedback about what product they need, what they think about the product at the supermarket, etc. This is a very good chance to the owner to get review of their products therefore the owner can understand the need of the customers.

CHAPTER VI

CONCLUSION AND RECOMMENDATION

The last chapter will emphasize the conclusion toward A Comparative Study of Clustering Algorithms and the Implementation of Apriori Algorithm to Give Impact for the Retail Company (Case study: Product sales data). The researcher would like to emphasize the implementation of Association Rules Market Basket Analysis by using Apriori Algorithm to improve the effectiveness of marketing strategy and sales tactics based on the result of clustering analysis. In the clustering analysis, the researcher used 20,464 rows of customer transactions and 39,474 rows for the AR-MBA. The recommendation is constructed in this chapter as well which can be a suggestion for further research.

6.1. Conclusion

Refers to data processing and analysis which has been established in previous chapter, then the conclusion for this research can be constructed as follows:

1. Based on the indicators of time and BSS/TSS with 100 simulations, then the Fuzzy C Means Clustering Algorithm is the best strategy clustering algorithm analyze the product sales data.
2. The best cluster customer (the targeted customers) for the Pamella Satu Supermarket Yogyakarta are the customers who are willing to go shopping on the weekend afternoon/evening, where they tend to buy adult product (SHAMPOO and BODY SOAP BAR) and most of them buying the product price up to IDR 47,456
3. By performing AR-MBA using Apriori Algorithm, the researcher found that there are 23 rules obtained in the R Software. To be the basis to formulate sales

strategy then the business owner of Pamella Satu Supermarket Yogyakarta needs to put products Baby Soap Liquid, Conditioner, Diapers & Wipes, Hair Nutrition, Body Soap Bar, Body Soap Liquid, Healthy Soap Bar, Facial Soap Women, and Baby Hair & Body Care, near to SHAMPOO.

4. By performing clustering and ARMBA, then the owner know the market share and the characteristics of their customers and how to plan the shop lay out of their products in the store to have a correct promotion strategy to increase their product sales.

6.2. Recommendation

Finally, for the future improvement from this study, there is some suggestion which can be constructed for further research. There are some suggestions for future research as follows:

1. Compare other clustering algorithm with Fuzzy C Means Algorithm concerning processing time and sum of squares.
2. Compare the Apriori Algorithm result with other AR-MBA Algorithm.
3. In addition, the next research should use representative data to conduct clustering analysis.

REFERENCES

Andriyana, V. (2015). Perbandingan 3 Metode dalam Data Mining untuk Prekdisi Penerima Beasiswa Berdasarkan Prestasi di SMA Negeri 6 Surakarta.

Arga Felani, D. (2015). Perbandingan 3 Metode dalam Data Mining Untuk Menentukan Strategi Penjualan Produk Makanan dan Minuman pada Toserba Lestari Baru Gemolong.

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective. New York: Plenum Press.

Bora, D. J., & Gupta, A. K. (2014). A Comparative study Between Fuzzy Clustering Algorithm and. *International Journal of Computer Trends and Technology (IJCTT)*, 108-109.

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R Package for Cluster Validation. *Journal of Statistical Software*, 1-22. Retrieved from http://www.jstatsoft.org/v25/i04/.

Dewi Astika, N. (2015). Penerapan Data Mining untuk Menganalisis Penjualan Barang dengan Menggunakan Metode Apriori pada Supermarket Sejahtera Lhokseumawe.

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications.* SIAM.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.* Waltham: Morgan Kauffman.

Hartanto Kamagi, D., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan. 16.

Irdiansyah, E. (2010). Penerapan Data Mining pada Penjualan Produk Minuman di PT. Pepsi Cola Indobeverages Menggunakan Metode Clustering.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A review. *ACM Computing Surveys*.

Jiawei, H., & Kamber, M. (2006). *Data Mining: Concepts and Techniques.* USA: Morgan Kaufmann.

LambdaVu. (2014, January 20). *Quetions*. Retrieved from StackExchange: https://stats.stackexchange.com/questions/82776/what-does-total-ss-and-between-ss-mean-in-k-means-clustering

Li, S. (2017, October 2). Retrieved from R-bloggers: https://www.r-bloggers.com/a-gentle-introduction-on-market-basket-analysis%E2%80%8A-%E2%80%8Aassociation-rules/

Li, S. (2017, September 24). *A Gentle Introduction on Market Basket Analysis - Association Rules*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. Volume 1: Statistics, 281–97). Berkeley: University of California Press.

Octa Chandra, R., Haidar Mirza, A., & Kurniawan. (2014). Penerapan Data Mining untuk Memprediksi Minat Nasabah pada AJB Bumiputera 1912 Palembang. 2.

Pramudiono, I. (2003). *Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data.*

Rahmawati, L., Sihwi, S. W., & Suryani, E. (2016). Analisa Clustering Menggunakan Metode K-Means dan Hierarchical Clustering (Studi Kasus: Dokumen Skripsi Jurusan Kimia, FMIPA, Universitas Sebelas Maret). *ITSmart: Jurnal Teknologi dan Informasi*, 2.

Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Jurnal EECCIS*, 2.

Sambandam, R. (2003). Cluster analysis gets complicated: Collinearity is a natural problem in clustering So how can researchers get around it. *Marketing Research*.

Samuel, D. (2008). Penerapan Stuktur FP-Tree dan Algoritma FPGrowth dalam Optimasi Penentuan Frequent Itemset. *Institut Teknologi Bandung*.

Tanjung, P. A. (2016). Penerapan Fuzzy C-Means Clustering pada Data Nasabah Bank. 30.

Taufik Luthfi, E., & Kusrini. (2009). *Algoritma Data Mining.* Yogyakarta: Andi.

Taufiq Luthfi, E. (2009). Penerapan Data Mining Algoritma Asosiasi. *JURNAL DASI*, 2.

Tussel, F. (2013, January 25). *StackExchange*. Retrieved from stats.stackexchange.com: https://stats.stackexchange.com/questions/48520/interpreting-result-of-k-means-clustering-in-r

Wiharto, M. (2013). Analisis Kluster Menggunakan Bahasa Pemograman R Untuk Kajian Ekologi. *bionature*, 1.

APPENDIX

1. The way to calculate Multicollinearity detection in R Studio

```
library(data.table)
alldata <- fread(“E:/Rakhmat’s Thesis/PamellaDataset1.csv”, header=TRUE)
drakhmat <- data.frame(alldata)
class(drakhmat)
[1] “data.frame”
Dim(drakhmat)
[1] 20464       7
library(car)
vif(lm(DayIND~., data=drakhmat))
kappa(drakhmat)
```

2. Cluster validation using Hierarchical Method

```
library(data.table)
alldata <- fread(“E:/Rakhmat’s Thesis/PamellaDataset1.csv”, header=TRUE)
drakhmat <- data.frame(alldata)
class(drakhmat)
[1] “data.frame”
Dim(drakhmat)
[1] 20464       7
dr4analysis <- drakhmat[,-1]
disd4 <- dist(dr4analysis, method=”euclidean”)
klas1 <- hclust(disd4, “ward.D”)
library(clValid)
intern <- clValid(drakhmat[,-1], nClust=2:6, clMethods=”hierarchical”,
validation=”internal”, maxitems=20464)
summary(intern)
```

3. The way to show dendrogram based on Cluster validation using Hierarchical Clustering Method

From the previous cluster validation code, we need to add more code in R Studio as follow:

```
klas1<-hclust(disd4,"ward.D")
plot(klas1,labels=drakhmat[,1])
```

4. The data processing for K Means Clustering Algorithm

```
library(data.table)
library(car)
library(MuMIn)
library(psych)

alldata<-fread("E:/Rakhmat's Thesis/PamellaDataset1.csv",header=TRUE)
class(alldata)
drakhmat<-data.frame(alldata)
class(drakhmat)
dim(drakhmat)
dr4analysis  <- drakhmat[,-1]
disd4 <- dist(dr4analysis, method="euclidean")
disd4
kmrakhmat  <-  kmeans(disd4, 2)
```

5. The way to compute Data Visualization of K Means Clustering Algorithm with a Cluster Plot

```
#1 Call the required library
library(factoextra)

#2 The way to visualize the Fuzzy C Means Clustering in R Studio
fviz_cluster(list(data = alldata1, cluster=kmrakhmat$cluster),
        ellipse.type = "norm",
        ellipse.level = 0.68,
        palette = "jco",
        ggtheme = theme_minimal())
```

6. The data processing for Fuzzy C Means Clustering Algorithm

```
#1 Call all the required library
library(ppclust)
library(factoextra)
library(cluster)
library(fclust)
library(data.table)
library(inaparc)

#2 Insert the dataset into R Studio
alldata <- fread("E:/Rakhmat's Thesis/PamellaDataset1.csv",header=TRUE)
dim(alldata)

#3 Change dataset to be data frame
alldata1 <- data.frame(alldata[,-1]
#4 Plot the data by the classes of dataset Transaction
Pairs(alldata1, col=alldata[,1]
#5 Initialization
res.fcm <- fcm(alldata1, centers=2)

#6 Initialization by using kmeans++ algorithm
v0 <- inaparc::kmpp(alldata1, k=2)$v
print(v0)
res.fcm <- fcm(alldata1, centers=v0)
res.fcm <- fcm(alldata1, centers=2)
as.data.frame(res.fcm$u)

#7 Initial and final cluster prototypes
res.fcm$v0
res.fcm$v

#8 Summary of clustering result
Summary(res.fcm)
```

7. The way to visualize the Fuzzy C Means Clustering Result and its computation time in R Studio

```
#1 Call the required library
library(factoextra)

#2 The way to visualize the Fuzzy C Means Clustering in R Studio
fviz_cluster(list(data = alldata1, cluster=res.fcm$cluster),
        ellipse.type = "norm",
        ellipse.level = 0.68,
        palette = "jco",
        ggtheme = theme_minimal())

#3 The processing time
proc.time()
```

8. The required R libraries for AR-MBA

a. tidyverse
b. data.table
c. arules
d. arulesViz
e. plyr

9. Data preprocessing and exploring

```
retail <- fread('D:/Thesis/AR-MBA/PamellaDataset2.csv', header = TRUE)
retail <- retail[complete.cases(retail), ]
retail %>% mutate(Transaction = as.factor(Transaction))
```

Transaction	SubCategory	Date	Time	ListPrice	SaleQty
767822	BABY COLOGNE	01/10/2017	08:29	12.225	1
768726	BABY COLOGNE	01/10/2017	09:32	20.700	1
768434	BABY COLOGNE	01/10/2017	09:40	12.150	1
768739	BABY COLOGNE	01/10/2017	10:08	15.550	1
768457	BABY COLOGNE	01/10/2017	10:43	15.100	1
768756	BABY COLOGNE	01/10/2017	10:55	17.475	1
76835	BABY COLOGNE	01/10/2017	11:08	9.025	1
768313	BABY COLOGNE	01/10/2017	11:20	17.475	1
769081	BABY COLOGNE	01/10/2017	11:23	20.700	1
828612	BABY COLOGNE	01/10/2017	11:36	9.025	1
828612	BABY COLOGNE	01/10/2017	11:36	15.400	1
7681123	BABY COLOGNE	01/10/2017	11:42	20.700	1
768330	BABY COLOGNE	01/10/2017	12:08	20.375	1
768491	BABY COLOGNE	01/10/2017	12:09	20.700	1
7681153	BABY COLOGNE	01/10/2017	12:21	17.475	1
7684100	BABY COLOGNE	01/10/2017	12:56	7.775	1
7684100	BABY COLOGNE	01/10/2017	12:56	12.150	1

768789	BABY COLOGNE	01/10/2017	12:58	19.875	1
7678151	BABY COLOGNE	01/10/2017	13:14	16.075	1
7690129	BABY COLOGNE	01/10/2017	13:15	17.475	1
768364	BABY COLOGNE	01/10/2017	13:50	17.475	1
768924	BABY COLOGNE	01/10/2017	14:14	16.075	1
768027	BABY COLOGNE	01/10/2017	14:40	15.050	1
82895	BABY COLOGNE	01/10/2017	16:18	15.400	1
768887	BABY COLOGNE	01/10/2017	16:19	19.875	1
7683109	BABY COLOGNE	01/10/2017	16:26	9.025	1
7683115	BABY COLOGNE	01/10/2017	16:38	7.775	1
7683129	BABY COLOGNE	01/10/2017	17:12	17.475	1
7683130	BABY COLOGNE	01/10/2017	17:18	15.400	1
7688120	BABY COLOGNE	01/10/2017	17:26	12.150	1
768073	BABY COLOGNE	01/10/2017	17:40	17.475	1
7683143	BABY COLOGNE	01/10/2017	17:54	12.225	1
828784	BABY COLOGNE	01/10/2017	18:05	15.500	1
7688143	BABY COLOGNE	01/10/2017	18:29	19.875	1
768091	BABY COLOGNE	01/10/2017	18:40	15.100	2
8289129	BABY COLOGNE	01/10/2017	18:55	12.225	1
8289139	BABY COLOGNE	01/10/2017	19:05	17.475	1
8287115	BABY COLOGNE	01/10/2017	19:09	7.775	1
769554	BABY COLOGNE	01/10/2017	19:13	12.150	1
8287123	BABY COLOGNE	01/10/2017	19:22	7.775	1
769576	BABY COLOGNE	01/10/2017	19:44	15.050	1
8287150	BABY COLOGNE	01/10/2017	20:06	15.100	2
8289180	BABY COLOGNE	01/10/2017	20:18	19.875	1
7683200	BABY COLOGNE	01/10/2017	20:37	12.225	1
7680148	BABY COLOGNE	01/10/2017	20:37	20.700	1
7680176	BABY COLOGNE	01/10/2017	21:26	7.775	1
771737	BABY COLOGNE	02/10/2017	09:37	9.025	1
77083	BABY COLOGNE	02/10/2017	11:45	15.100	1
771298	BABY COLOGNE	02/10/2017	12:03	17.475	1
7711113	BABY COLOGNE	02/10/2017	12:11	15.400	1
77089	BABY COLOGNE	02/10/2017	12:11	15.050	1
7717124	BABY COLOGNE	02/10/2017	12:15	9.025	1
7721112	BABY COLOGNE	02/10/2017	12:42	12.225	1
7717175	BABY COLOGNE	02/10/2017	13:24	9.750	1
77183	BABY COLOGNE	02/10/2017	13:29	17.475	1
770714	BABY COLOGNE	02/10/2017	14:05	19.875	1
770715	BABY COLOGNE	02/10/2017	14:05	17.475	1
771428	BABY COLOGNE	02/10/2017	14:20	15.100	1
770890	BABY COLOGNE	02/10/2017	14:22	12.150	1
770744	BABY COLOGNE	02/10/2017	14:59	15.100	1
771336	BABY COLOGNE	02/10/2017	15:16	8.925	1
771349	BABY COLOGNE	02/10/2017	15:28	15.100	1

771572	BABY COLOGNE	02/10/2017	16:04	15.500	1
771486	BABY COLOGNE	02/10/2017	16:24	17.475	1
770778	BABY COLOGNE	02/10/2017	17:14	17.475	1
7718139	BABY COLOGNE	02/10/2017	17:27	4.725	1
7713123	BABY COLOGNE	02/10/2017	17:37	15.400	1
7713143	BABY COLOGNE	02/10/2017	19:04	15.500	1
7708223	BABY COLOGNE	02/10/2017	19:55	11.775	1
7708225	BABY COLOGNE	02/10/2017	19:57	17.475	1
7708227	BABY COLOGNE	02/10/2017	20:01	16.075	1
7714174	BABY COLOGNE	02/10/2017	20:06	12.150	1
7708234	BABY COLOGNE	02/10/2017	20:21	9.750	1
7722115	BABY COLOGNE	02/10/2017	20:30	16.075	1
833618	BABY COLOGNE	02/10/2017	20:47	11.775	1
7714203	BABY COLOGNE	02/10/2017	20:59	17.475	1
833631	BABY COLOGNE	02/10/2017	21:09	15.400	1
76819	DIAPERS & WIPES	01/10/2017	07:40	5.150	1
76819	DIAPERS & WIPES	01/10/2017	07:40	5.350	1
76819	DIAPERS & WIPES	01/10/2017	07:40	5.325	1
768111	DIAPERS & WIPES	01/10/2017	07:43	98.125	2
768123	DIAPERS & WIPES	01/10/2017	08:05	58.150	1
768129	DIAPERS & WIPES	01/10/2017	08:10	5.350	1
76844	DIAPERS & WIPES	01/10/2017	08:22	32.100	1
76845	DIAPERS & WIPES	01/10/2017	08:25	15.175	1
76848	DIAPERS & WIPES	01/10/2017	08:35	18.500	1
76874	DIAPERS & WIPES	01/10/2017	08:35	2.625	1
768417	DIAPERS & WIPES	01/10/2017	08:52	16.325	1
769015	DIAPERS & WIPES	01/10/2017	08:58	48.025	1
769018	DIAPERS & WIPES	01/10/2017	09:02	24.950	1
768423	DIAPERS & WIPES	01/10/2017	09:03	15.175	1
769024	DIAPERS & WIPES	01/10/2017	09:07	8.700	1
768715	DIAPERS & WIPES	01/10/2017	09:09	5.150	1
769027	DIAPERS & WIPES	01/10/2017	09:11	5.350	1
768430	DIAPERS & WIPES	01/10/2017	09:21	48.075	1
768430	DIAPERS & WIPES	01/10/2017	09:21	15.175	1
767826	DIAPERS & WIPES	01/10/2017	09:28	5.350	1
768433	DIAPERS & WIPES	01/10/2017	09:30	46.275	1
768433	DIAPERS & WIPES	01/10/2017	09:30	74.975	1
768433	DIAPERS & WIPES	01/10/2017	09:30	15.975	2
768728	DIAPERS & WIPES	01/10/2017	09:40	18.050	1
768435	DIAPERS & WIPES	01/10/2017	09:43	51.675	2
767837	DIAPERS & WIPES	01/10/2017	09:43	189.825	1
768159	DIAPERS & WIPES	01/10/2017	09:46	5.325	1
768436	DIAPERS & WIPES	01/10/2017	09:47	5.875	1
768436	DIAPERS & WIPES	01/10/2017	09:47	17.350	1
768436	DIAPERS & WIPES	01/10/2017	09:47	19.200	1

767839	DIAPERS & WIPES	01/10/2017	09:47	11.950	1
769043	DIAPERS & WIPES	01/10/2017	09:54	107.200	1
769043	DIAPERS & WIPES	01/10/2017	09:54	19.200	1
767844	DIAPERS & WIPES	01/10/2017	09:59	19.200	1
768442	DIAPERS & WIPES	01/10/2017	10:02	18.050	1
767848	DIAPERS & WIPES	01/10/2017	10:03	12.575	1
768739	DIAPERS & WIPES	01/10/2017	10:08	15.175	2
769052	DIAPERS & WIPES	01/10/2017	10:12	8.850	1
769052	DIAPERS & WIPES	01/10/2017	10:12	1.825	40
767859	DIAPERS & WIPES	01/10/2017	10:14	10.975	1
768179	DIAPERS & WIPES	01/10/2017	10:21	16.650	1
768453	DIAPERS & WIPES	01/10/2017	10:26	3.275	1
768750	DIAPERS & WIPES	01/10/2017	10:29	18.050	1
768454	DIAPERS & WIPES	01/10/2017	10:32	12.925	1
767871	DIAPERS & WIPES	01/10/2017	10:32	12.575	1
767874	DIAPERS & WIPES	01/10/2017	10:35	54.350	1
768191	DIAPERS & WIPES	01/10/2017	10:39	13.850	1
768191	DIAPERS & WIPES	01/10/2017	10:39	27.425	5
767876	DIAPERS & WIPES	01/10/2017	10:50	84.550	1
76831	DIAPERS & WIPES	01/10/2017	10:54	62.125	1
769070	DIAPERS & WIPES	01/10/2017	11:00	5.325	1
768757	DIAPERS & WIPES	01/10/2017	11:01	8.850	1
767882	DIAPERS & WIPES	01/10/2017	11:03	31.100	1
76835	DIAPERS & WIPES	01/10/2017	11:08	84.550	1
7681105	DIAPERS & WIPES	01/10/2017	11:09	61.175	1
768467	DIAPERS & WIPES	01/10/2017	11:11	46.075	1
769075	DIAPERS & WIPES	01/10/2017	11:11	12.575	1
769075	DIAPERS & WIPES	01/10/2017	11:11	5.350	2
7681106	DIAPERS & WIPES	01/10/2017	11:13	5.150	1
768761	DIAPERS & WIPES	01/10/2017	11:14	69.150	1
769077	DIAPERS & WIPES	01/10/2017	11:16	18.050	1
769077	DIAPERS & WIPES	01/10/2017	11:16	61.000	1
767893	DIAPERS & WIPES	01/10/2017	11:18	8.700	1
768312	DIAPERS & WIPES	01/10/2017	11:18	59.300	1
82865	DIAPERS & WIPES	01/10/2017	11:19	8.700	1
769479	DIAPERS & WIPES	01/10/2017	11:20	52.625	2
768473	DIAPERS & WIPES	01/10/2017	11:24	5.325	1
769481	DIAPERS & WIPES	01/10/2017	11:24	16.650	1
769083	DIAPERS & WIPES	01/10/2017	11:26	84.550	2
769484	DIAPERS & WIPES	01/10/2017	11:30	16.575	1
7692121	DIAPERS & WIPES	01/10/2017	11:31	15.975	1
768317	DIAPERS & WIPES	01/10/2017	11:34	107.200	1
768317	DIAPERS & WIPES	01/10/2017	11:34	15.175	1
7681118	DIAPERS & WIPES	01/10/2017	11:35	49.900	3
828612	DIAPERS & WIPES	01/10/2017	11:36	17.275	1

828612	DIAPERS & WIPES	01/10/2017	11:36	16.575	1
7681121	DIAPERS & WIPES	01/10/2017	11:38	18.500	1
769489	DIAPERS & WIPES	01/10/2017	11:43	52.625	1
768482	DIAPERS & WIPES	01/10/2017	11:46	1.825	6
828617	DIAPERS & WIPES	01/10/2017	11:48	58.150	1
768324	DIAPERS & WIPES	01/10/2017	11:52	19.200	1
768324	DIAPERS & WIPES	01/10/2017	11:52	84.550	2
7681131	DIAPERS & WIPES	01/10/2017	11:54	18.050	1
828628	DIAPERS & WIPES	01/10/2017	11:59	5.875	1
828631	DIAPERS & WIPES	01/10/2017	12:03	5.850	1
768490	DIAPERS & WIPES	01/10/2017	12:05	2.625	1
828635	DIAPERS & WIPES	01/10/2017	12:07	19.200	1
828635	DIAPERS & WIPES	01/10/2017	12:07	43.750	1

```
[ reached getOption("max.print") -- omitted 39308 rows ]
```

10. The result of data pre-processing to show variables and observations for conducting AR-MBA

```
retail$Date <- as.Date(retail$Date)
retail$Transaction <- as.numeric(as.character(retail$Transaction))
glimpse(retail)
```

```
Observations: 39,474
Variables: 6

$ Transaction <dbl> 767822, 768726, 768434, 768739, 768457, 768756, 76835, 768313,
769081, 828612, 82861...
$ SubCategory <chr> "BABY COLOGNE", "BABY COLOGNE", "BABY
COLOGNE", "BABY COLOGNE", "BABY COLOGNE", "BAB...
$ Date <date> 0001-10-20, 0001-10-20, 0001-10-20, 0001-10-20, 0001-10-20, 0001-10-
20, 0001-10-20,...
$ Time <chr> "08:29", "09:32", "09:40", "10:08", "10:43", "10:55", "11:08", "11:20",
"11:23", "11...
$ ListPrice <dbl> 12.225, 20.700, 12.150, 15.550, 15.100, 17.475, 9.025, 17.475, 20.700,
9.025, 15.400...
$ SaleQty <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
```

11. The way to to transform the data into Transaction format in R Studio

```
retail_sorted <- retail[order(retail$Transaction),]
library(plyr)
```

12. The way to split dataset into transaction format

```
itemList <- ddply(retail,c("Transaction","Time"),
            function(df1)paste(df1$SubCategory,
            collapse = ","))

itemList$Transaction <- NULL
itemList$Time <- NULL
colnames(itemList) <- c("items")
```

13. The way to write the data frame into a CSV file

```
write.csv(itemList,"D:/Thesis/AR-MBA/MBA_Pamella.csv", quote = FALSE,
row.names = TRUE)

#Let's have a closer look at how many transactions we have and what they are.
tr <- read.transactions('D:/Thesis/AR-MBA/MBA_Pamella.csv', format = 'basket',
sep=',')
Warning message:
In asMethod(object) : removing duplicated items in transactions
```

14. The result of transaction items in R software using AR-MBA

```
tr
transactions in sparse format with
16594 transactions (rows) and
16622 items (columns)
```

```
summary(tr)
transactions as itemMatrix in sparse format with
16594 rows (elements/itemsets/transactions) and
16622 columns (items) and a density of 0.0001793634
```

Most frequent items:

SHAMPOO	DIAPERS & WIPES	FACIAL SOAP WOMEN	HEALTHY SOAP BAR	BODY SOAP BAR	OTHER
6329	2938	2850	2556	2409	32391

element (itemset/transaction) length distribution:

SIZE												
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	8013	4364	2319	1042	449	208	120	47	15	9	5	2

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	2.000	3.000	2.981	4.000	13.000

```
includes extended item information - examples:
  labels
1      1
2     10
3    100
```

15 The way to know item frequency plot in R Studio software

The way to show top 5 frequent items in R Studio

```
itemFrequencyPlot(tr, topN=5, type='absolute')
```

The way to show least 5 frequent items in R Studio:

```
itemf <- sort(itemFrequency(tr, type = "absolute"), decreasing=TRUE)
barplot(itemf[23:28])
```

16. Create several rules in R studio

```
rules <- apriori(tr, parameter = list(supp=0.001, conf=0.8))
```

Apriori

Parameter specification:

confidence	minval	smax	arem	aval	originalSupport	maxtime	support	minlen	maxlen	target	ext
0.8	0.1	1	none	FALSE	TRUE	5	0.001	10	10	rules	FALSE

```
Algorithmic control:
filter tree    heap   memopt  load    sort verbose
0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 16

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[16622 item(s), 16594 transaction(s)] done [0.01s].
sorting and recoding items ... [28 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [23 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

17. The rules summary computed in R Studio

```
rules <- sort(rules, by='confidence', decreasing = TRUE)
summary(rules)
set of 23 rules

rule length distribution (lhs + rhs):sizes
 3      4
 7      16

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   3.000   4.000   3.696   4.000   4.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.001024   Min.   :0.8065   Min.   :2.114   Min.   :17.00
 1st Qu.:0.001235   1st Qu.:0.8245   1st Qu.:2.162   1st Qu.:20.50
 Median :0.001386   Median :0.8605   Median :2.256   Median :23.00
 Mean   :0.001420   Mean   :0.8629   Mean   :2.263   Mean   :23.57
 3rd Qu.:0.001507   3rd Qu.:0.8909   3rd Qu.:2.336   3rd Qu.:25.00
 Max.   :0.002230   Max.   :0.9583   Max.   :2.513   Max.   :37.00

mining info:
 data ntransactions support confidence
   tr         16594   0.001        0.8
```

18. The way to show visualization of the customer amount spend and sale quantity on each date

```
retail$Time <- as.factor(retail$Date)

        retail %>%

        ggplot(aes(x=Date)) +

        geom_histogram(stat="count",fill="indianred")
```

19 The way to plot the top 10 rules in R Studio

```
topRules <- rules[1:10]
plot(topRules, interactive = TRUE, cex = 2)
plot(topRules, method="graph")
plot(topRules, method = "grouped")
```

20. The simulation result of combined hierarchical plus K Means and Fuzzy C Means Clustering Algorithms

No	Combined Hierarchical plus K Means		Fuzzy C Means	
	BSS/TSS ratio (in %)	**Processing Time (in Second)**	**BSS/TSS ratio (in %)**	**Processing Time (in Second)**
1	56,7	78,52	59,34	75,7
2	56,7	127,88	59,34	63,78
3	56,7	138,57	59,34	62,48
4	56,7	103,13	59,34	64,38
5	56,7	90,05	59,34	63,16
6	56,7	108,32	59,34	65,57
7	56,7	99,85	59,34	64,63
8	56,7	109,2	59,34	64,77
9	56,7	95,83	59,34	63,59
10	56,7	107,09	59,34	63,59
11	56,7	127,99	59,34	64,35
12	56,7	130,86	59,34	64,17
13	56,7	98,86	59,34	62,31
14	56,7	101,4	59,34	64,21
15	56,7	98,6	59,34	64,28
16	56,7	105,47	59,34	65,31
17	56,7	111,81	59,34	64,73

18	56,7	142,31	59,34	61,96
19	56,7	102,66	59,34	60,98
20	56,7	119,9	59,34	63,86
21	56,7	92,64	59,34	64,38
22	56,7	119,28	59,34	62,56
23	56,7	137,88	59,34	64,4
24	56,7	105,05	59,34	59,97
25	56,7	112,82	59,34	63,28
26	56,7	97,49	59,34	61,68
27	56,7	95,84	59,34	62,43
28	56,7	106,99	59,34	64,52
29	56,7	125,29	59,34	64,36
30	56,7	101,79	59,34	60,11
31	56,7	98,26	59,34	62,14
32	56,7	114,99	59,34	62,45
33	56,7	105,01	59,34	64,99
34	56,7	117,17	59,34	64,06
35	56,7	107	59,34	64,86
36	56,7	116,85	59,34	65,25
37	56,7	137,26	59,34	61,29
38	56,7	78,52	59,34	75,7
39	56,7	116,61	59,34	57,13
40	56,7	94,78	59,34	63,3
41	56,7	118,91	59,34	60,46
42	56,7	134,16	59,34	63,55
43	56,7	69,54	59,34	62,45
44	56,7	81,05	59,34	62,72
45	56,7	121,8	59,34	63,49
46	56,7	82,98	59,34	63,22
47	56,7	77,99	59,34	60,64
48	56,7	82,68	59,34	59,26
49	56,7	69,99	59,34	63,52
50	56,7	86,48	59,34	63,79
51	56,7	86,39	59,34	63,52
52	56,7	94,66	59,34	62,66
53	56,7	89,86	59,34	64,54
54	56,7	77,37	59,34	63,03
55	56,7	82,1	59,34	62,35
56	56,7	73,01	59,34	57,53
57	56,7	77,61	59,34	62,25
58	56,7	82,31	59,34	63,52
59	56,7	84,19	59,34	60,23
60	56,7	77,81	59,34	61,91
61	56,7	72,29	59,34	62,81
62	56,7	90,96	59,34	59,65

63	56,7	68,94	59,34	62,43
64	56,7	83,16	59,34	67,37
65	56,7	89,08	59,34	63,92
66	56,7	79,42	59,34	64,82
67	56,7	76,05	59,34	62
68	56,7	72,96	59,34	61,42
69	56,7	77,39	59,34	63,04
70	56,7	81,05	59,34	63,18
71	56,7	75,45	59,34	61,89
72	56,7	77,11	59,34	63,54
73	56,7	82,02	59,34	63,39
74	56,7	71,22	59,34	64,05
75	56,7	89,72	59,34	63,91
76	56,7	89,15	59,34	63,81
77	56,7	74,32	59,34	62,69
78	56,7	80,68	59,34	63,25
79	56,7	79,25	59,34	65,4
80	56,7	85,69	59,34	61,6
81	56,7	76,36	59,34	63,79
82	56,7	81,72	59,34	61,05
83	56,7	79,34	59,34	63,25
84	56,7	83,96	59,34	63,08
85	56,7	87,25	59,34	62,7
86	56,7	85,62	59,34	61,14
87	56,7	83,17	59,34	64,02
88	56,7	69,88	59,34	61
89	56,7	75,62	59,34	65,89
90	56,7	82,03	59,34	64,53
91	56,7	79,38	59,34	61,74
92	56,7	81,94	59,34	61,75
93	56,7	80,06	59,34	63,17
94	56,7	82,56	59,34	64,73
95	56,7	77,92	59,34	63,78
96	56,7	80,03	59,34	63,42
97	56,7	79,97	59,34	68,41
98	56,7	88,39	59,34	66,17
99	56,7	78,57	59,34	67,78
100	56,7	80,57	59,34	57,16
AVERAGE	**56,7**	**93,83**	**59,34**	**63,23**