Credit Card Data Analytics

Segmentation and Clustering Using Hierarchical and K-Means Clustering Algorithms

Thomas Januardy
Information System Major, Faculty of Information & Technology, Multimedia Nusantara University, Tangerang City, Indonesia
thomas.januardy@student.umn.ac.id

***Abstract*— Credit cards are one of the developments in payment methods that are currently very intensively used. The use of various credit cards by its users gives rise to a large number of data records that are generated, which can be used for various purposes. Data is one of the most important things in the age of all technology as it is today to help the continuity of a process. It is the set of data that creates the "big data". With a set of data called "big data", we can process an information that helps us in many ways, such as decision making, to determining marketing strategies, especially for market segmentation created by credit card users. With the existence of a segmentation created from the creation of clusters, the data can be used by several companies and organizations to find out the patterns created by each of their customers in using credit cards. With that said, the authors wanted to try to calculate and develop a customer segmentation to define a new marketing strategy. Authors use data mining and exploratory methods against the credit card data and apply two algorithms, both hierarchical clustering and k-means algorithms to make clusters and develop relations between behavioral variables. It is useful for creating a pattern related to new market segmentation which will later be created and can be useful for strategic development and decision making in terms of marketing, especially patterns created by customers in using credit cards.**

Index Terms — Credit Card, Customer, Balance, Credit Limit, Purchase, Clustering, Segmentation, Cluster, Hierarchical, K-means, Data Mining, Plot

I. Introduction

A. Big Data Concept

Data has become one of the important aspects, especially in an all-digital world, where all information is sourced from a data. Every human activity that utilizes technology, almost all of them use data as based infrastructures in order to be used. Data is a set of "raw" facts from an observation or research process, before it becomes the result data, which is information. Data can be divided into quantitative data and qualitative data. After the change of data into information, the fact can be used for many things, one of which is as a strategic analysis material to determine a decision, both qualitative or quantitative datas.

Along with the time, the use of technologies and human needs are increasingly diverse, creating a lot of unstructured datas that are widespreading in irregular forms. This set of data is called "big data". Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. Big data can be analyzed for insights that lead to better decisions and strategic business moves [3]. Big data is an aspect that is often needed by large companies today as a very complex and useful accumulation of data for companies. Big data needs to be analyzed to get value, whether trend, pattern, or any behavior related to people or customers for some companies [4]. The term “big data” refers to data that is so large, fast or complex that it’s difficult or impossible to process using traditional methods. But there is not literally that "big data" means very large amounts of data, but there are some specific criterias that must be achieved in order for a set of data to be referred to as a "big data". Data comes from the rapid growth of volume, but it does rapidly and efficiently processing those data refers to velocity of data processing [5]. That concept of “big data” gained momentum in the early 2000s when an industry analyst, Doug Laney articulated the now-mainstream definition of big data as the “three V’s”:

- Volume: Organizations collect data from a variety of sources, including business transactions, smart (IoT) devices, industrial equipment, videos, social media and more. In the past, storing it would have been a problem – but cheaper storage on platforms like data lakes and Hadoop have eased the burden.

- Velocity: With the growth in the Internet of Things, data streams in to businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

- Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions [3].

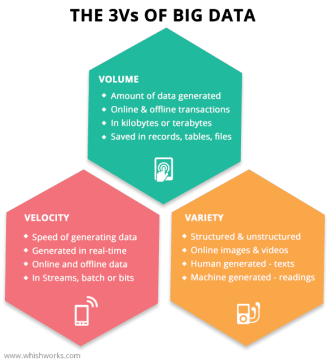

Figure 1. The "3Vs" of Big Data

Big datas are collecting both structured or unstructured data sources beside they are quantitative or qualitative datas. Unstructured data can come from social media (Facebook, Instagram, Twitter posts, etc) [6]. While structured data sources can come from internal database, such from an organization or a company [4]. Many believe that, for companies that get it right, big data will be able to unleash new organizational capabilities and value [1]. The creation and collection of data in today's society is not only limited to functions related to finance and consumers. In short, these companies can have a much more complete picture of their customers and operations by combining unstructured and structured data [2].

The "big data" era has produced a huge trend, and large organizations are working hard to use big data analysis to create values for their companies. With the rapid expansion of data volume, progress and types, the technical and technological development of data storage, analysis and visualization have achieved substantial development. Organizations and companies need to change to embrace this new era of innovation, where business's values can be derived from it, through big data analysis. Companies must adapt to the changing technological climate data creation and collection of today's society, which is not limited to financial and consumer-related functions.

Not only for companies, the use of big data in data management can also be implemented for small events, where they could discover patterns or predict future possibilities. If the data can be managed correctly, it has many uses. Companies could search for strategies and make decisions based on available data, otherwise students and citizens could study the habits of certain people, research, or even predict possible natural disasters based on time or weather's patterns.

B. Problems

In this case, the problem caused by this study is the use of credit cards in this day and age by some customers. In this growing era of technologies, the development of payment methods is increasingly diverse and easier to use. In addition to the existence of physical money, digital money is algo began to widely used because of its effectiveness which is considered easier by the public. One of the payment methods is the use of credit card. Credit cards are now generally used by almost everyone as a method of payment for their transactions, either they are online transactions or offline transactions [7].

No wonder the use of credit cards are increasing, because the requirements for their use are easy and also systematic that are easier than transactions that use physical form of money. As people that using credit cards grow, there are also a lot of "raw" datas from each user that can be managed to get new form of informations. This information is what the case study wants to develop, which is to develop a market segmentation for customers with the aim of formulating a marketing strategy. With this formulation, researchers can find out the behavior of each customer's pattern of purchases, balances, credit limits, and the others so on, from the use of a credit card. For this reason, the study impacted on each customer's credit card usage behavior, which determines which marketing strategies can be used for many purposes.

II. Literature Review

A. Data Mining

The process of searching and extracting information from piles of data with large amounts is the main process of data mining, the main purpose of processing the data is to generate new information [8]. Data mining is the process of discovering interesting knowledge from these large amounts of data stored in database, data warehouse or other information repositories. A number of data mining techniques have already been done on educational data mining to improve the performance of students like regression, genetic algorithm, bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Data mining is also known as Knowledge Discovery in Database (KDD), which is a process that automatically searches data in a very large memory space from data to find patterns using techniques such as association classification or clustering [9].

B. Clustering

Clustering on big data is a stage to classify data sets whose class attributes have not been described. The concept of clustering is to maximize and minimize intra-class similarities. Cluster can also be interpreted as a group. Then, the clustering analysis will basically produce a number of clusters (groups). Analysis needs to be applied to the understanding that a set of certain data actually already has similarities

among its members [10]. Therefore, each member who has similar characteristics is grouped into one or more of a group [11]. The purpose of data clustering is to minimize the objective function set in the clustering process, and generally always minimize the variation of a cluster and maximize the variation between clusters [9].

C. Hierarchical Clustering

Hierarchical methods are one of the clustering techniques by grouping two or more objects that have the closest similar value. Furthermore, the results of the first grouping are grouped again with other objects that have a second similarity value. Thus, it will form a hierarchical construction or based on a certain level such as the structure of the tree (match structure). Dendograms are used to describe the result structure of such hierarchical clusters [12].

Hierarchical grouping method is usually used if there is no information on the number of groups to be selected. The direction of grouping can be divisive (top to down) meaning from 1 cluster to become k cluster or agglomerative (bottom up) meaning from n cluster (from n-existing data) to a k cluster. Hierarchical technique (hierarchical methods) is a clustering technique forming hierarchical construction or based on a certain level such as tree structure [14]. Hierarchical clustering methods consist of complete linkage clustering, single linkage clustering, average linkage clustering, and centroid linkage [13].

Some of the methods of hierarchical cluster techniques used are as follows:

- Single Linkage (Nearest Neighbor Methods): This method defines the distance between two groups into the minimum distance between each single data point in the first cluster and every single data point in the second cluster.
- Complete Linkage (Furthest Neighbor Methods): This method looks for the maximum distance value between objects. It starts with the search for the two objects that are closest to them and the two objects form a new cluster.
- Average Linkage Methods : This method looks for the value of the average distance between objects. It starts with the search for the two objects that are closest to them and the two objects form a new cluster.
- Centroid Method: In the centroid method, the distance between two clusters formed is the distance between the two centroids in both clusters. Centroid is the average distance between objects in a cluster. Centroid values are obtained by doing an average on all cluster members. In this method, the process of recalculating centroids will be carried out every new cluster is formed, until a fixed cluster is formed [12].

From the hierarchical clustering technique, a collection of partitions can be produced, where in the collection there is:

- Clusters that have individual points. These clusters are at the very bottom level.
- A cluster in which there are points that belong to all clusters in it. This single cluster is at the very top level (super-class) [14].

D. K-Means

K-Means is one of the clustering techniques in data mining and modeling process without supervision and method of partitioning data grouping. The data grouped the K-Means method into several groups and each group had characteristics similar or similar to others but with other groups had different characteristics. With the goal of generalizing the difference of each data in one cluster and maximizing the difference with another cluster [15]. The terms in the K-Means clustering algorithm:

1. Cluster: Cluster is a group.
2. Cendroid: Cendroid is the central point for determining auclidian distance.
3. Iteration: Iteration is the repetition of a process, stopping when the iteration results have converged [10].

This method performs group analysis that leads to the division of observation objects into groups (clusters). This method attempts to find the center of the group (centroid) in the data as much as iterations of repairs are performed. This method attempts to divide data into groups so that data of the same character is entered into one group, while data of different characteristics is entered into another group. In general, the basic algorithm of K-Means clustering is as follows:

1. Determining the number of clusters.
2. Placing data into groups randomly.
3. Calculating cluster centers (centroid) by looking for the mean value for each group
4. Place the data to the nearest centroid.
5. Return to step 3, if there is still data that moves clusters or if the centroid value is above the value, threshold, or if the value in the objective function used is still above the threshold [12].

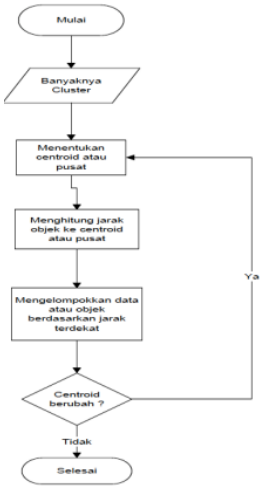

Figure 2. The Flowchart of K-Means

III. METHODOLOGY

A. Object of Research

The object of research that we use as material for the first research is to develop a customer segmentation that using credit cards to define a new marketing strategy. The dataset has met the minimum criteria, specified for research purposes with 18 behavioral variables and about 9,000 data observations of active credit card holders. The behavioral columns contain CUST_ID (chr), BALANCE (num), BALANCE_FREQUENCY (num), PURCHASES (num), ONOFF_PURCHASES (num), INSTALLMENTS_PURCHASES (num), CASH_ADVANCE (num), PURCHASES_FREQUENCY (num), ONEOFF_PURCHASES_FREQUENCY (num), PURCHASES_INSTALLMENTS_FREQUENCY (num), CASH_ADVANCE_FREQUENCY (num), CASH_ADVANCE_TRX (int), PURCHASES_TRX (int), CREDIT_LIMIT (num), PAYMENTS (num), MINIMUM_PAYMENTS (num), PRC_FULL_PAYMENT (num), and TENURE (int). Some int variables such as CASH_ADVANCE_TRX, PURCHASES_TRX, and TENURE need to be numeric into numerical type because currently they are int. Clustering algorithms required numeric datas only in the operation. CUST_ID also needs to be deleted, although it is not needed during the clustering process. Apart from that, they are numeric datas that contain numbers.

B. Method of Collecting Data

Our collection method that we did is by retrieving a dataset from one of the well-known dataset search websites called “Kaggle”. This is because the website has much experienced online dataset, especially in science community and machine learning practitioners. Kaggle also allows us to discover and publish power sets, explore and build models in web-based data science environments, work with data scientists and other machine learning engineers, and enter competitions to solve data science challenges. The reason we also take the dataset in the Kaggle application is because the completeness of the content of the data from the website is very complete and freely accessed. Any people can access the data and can process it. Kaggle has many helpful features such as notebooks as a means for users to display their work data.

C. Research Methods

Research method is a step taken in order to collect information or data and evaluate the data that has been obtained. Research methods provide an overview of the research design that includes research procedures and steps to be done, research time, data sources, and by what steps the data is obtained and the next stage is processed and analyzed to get a conclusion. Because the data that want to be looked for is numerical, the method that can be offered is the management of quantitative data, which is used to group a type of cluster that you want to separate from one customer segmentation to another. As it is known, that people's behavior towards the use of credit cards has a different pattern. So with the formation of a certain cluster, research can take advantage of this separation as a specific grouping on the type of credit card usage to create a new marketing strategy.

The program that will be used to manage datasets is R. R is a programming language as well as a computing program that has been used to support statistical and graph analysis activities. R accessed using RStudio. RStudio is the Integrated Development Environment (IDE) for R.

Before getting into the algorithm, we need to explore the selected dataset. This needs to be done to find out the outside and in the data, whether the selected dataset will be the optimal data to apply data analysis or not, whether the dataset matches the algorithm. Therefore, it is necessary to make observations with data mining in the dataset. Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to,

- numerical analysis,
- pattern matching and areas of artificial intelligence such as machine learning,
- neural networks and genetic algorithms.

While many data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is commonplace to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships [11].

Customer segmentation analysis in this study reveals systematic and directed stages, so it can be known which methods produce the best cluster results. From the overall data, we will see how many and how good are the clusters found in each data from the active credit card holders.

At this stage, data grouping is carried out using a combination of two clustering algorithms, they are hierarchical clustering and the K-means method. Hierarchical clustering algorithm is used to determine the center of the cluster. Furthermore, the cluster center obtained by hierarchical clustering algorithm is used for the process of grouping data using the K-means method. In looking for such data, we will use 2 algorithms to formulate the accuracy that will be used to find the cluster, lead to the purpose of this research.

IV. Results and Discussion

A. Data Validation

Before exploring and processing the data, it is necessary to check the validity of the datas. What is meant to be validity is a condition of the overall observations contained in the dataset. Whether there is data that is damaged or lost, or can also check the existence of duplicated data. This needs to be straightened out first before starting the first data analysis step because checking is an essential step in processing data. One is by way of whether all variable columns have actual data with the other or still exist incomplete, using the "missmap" function in R. The result is that no data is lost at all and the data is one hundred percent complete (Figure 3). In addition, further checking is also carried out using the "plot_missing" function to see statistics for missing data through a plot so that it can be detected. The results show that there are 0.01% of missing data on the CREDIT_LIMIT variable and 3.5% of missing data on the MINIMUM_PAYMENTS variable (Figure 4). Thus, it can be concluded that the data can still be used, with a note that all data observations containing empty or null values must be immediately removed before being used for clustering.

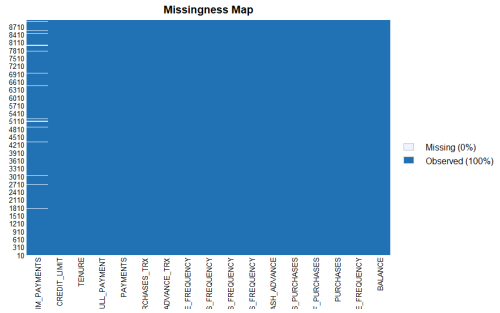

Figure 3. Missingness Map

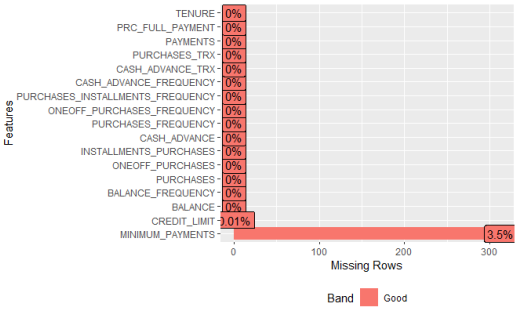

Figure 4. "plot_missing" Map

As for the examination of data validity through Kaggle itself to check the completeness of observation data in each column of variables that the dataset provides. Of the 18 variables, no missing data was found. It can be concluded that all variables are safe to use and ready to be processed to the next stage in R processing. Here is the breakdown of the variable columns from Kaggle:

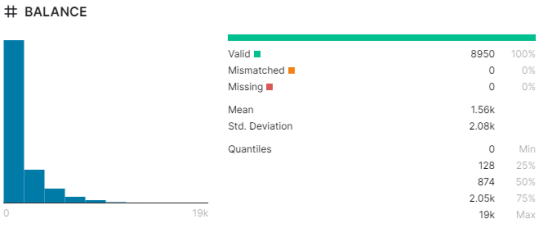

Figure 5. "BALANCE" variable

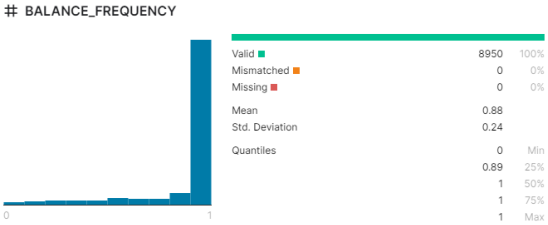

Figure 6. "BALANCE_FREQUENCY" variable

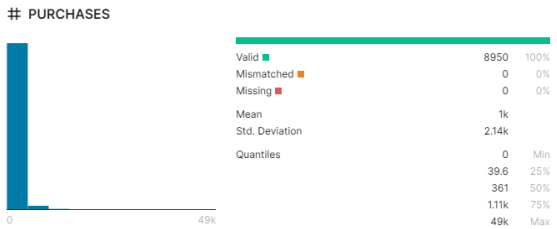

Figure 7. "PURCHASES" variable

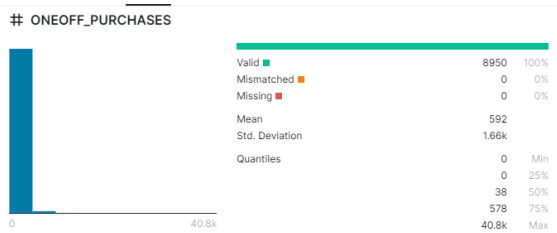

Figure 8. "ONEOFF_PURCHASES" variable

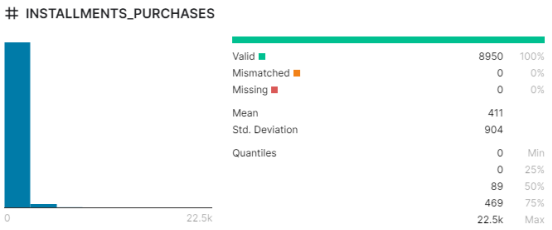

Figure 9. “INSTALLMENTS_PURCHASES” variable

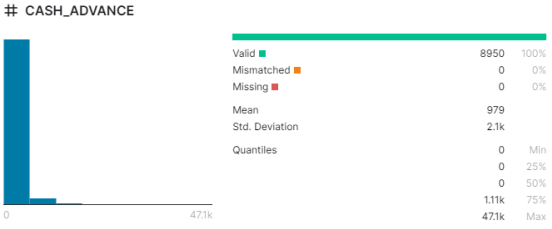

Figure 10. “CASH_ADVANCE” variable

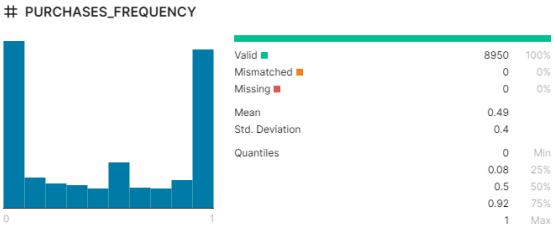

Figure 11. “PURCHASES_FREQUENCY” variable

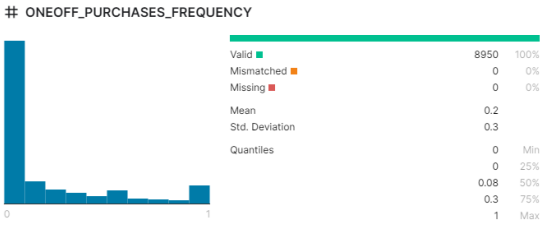

Figure 12. “ONEOFF_PURCHASES_FREQUENCY” variable

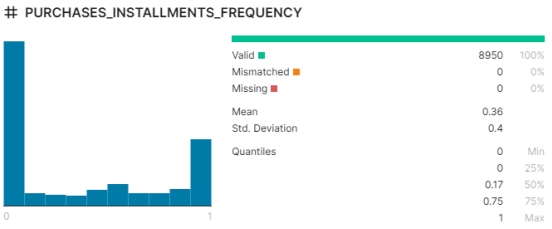

Figure 13. “PURCHASES_INSTALLMENTS_FREQUENCY” variable

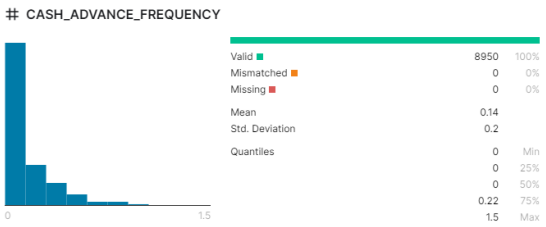

Figure 14. “CASH_ADVANCE_FREQUENCY” variable

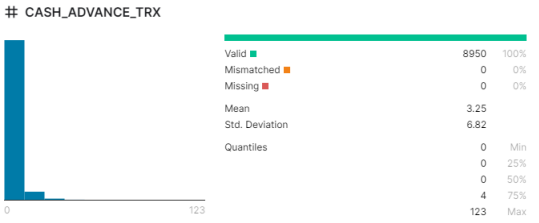

Figure 15. “CASH_ADVANCE_TRX” variable

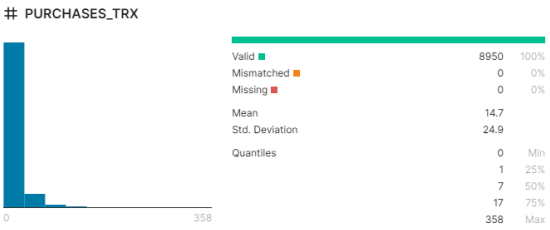

Figure 16. “PURCHASES_TRX” variable

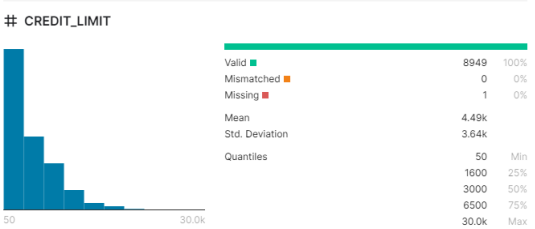

Figure 17. “CREDIT_LIMIT” variable

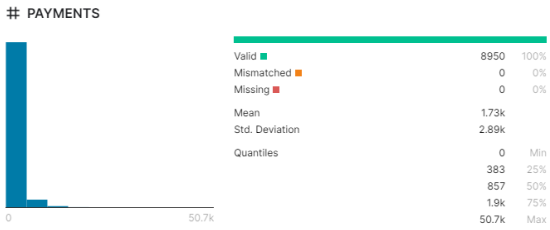

Figure 18. “PAYMENTS” variable

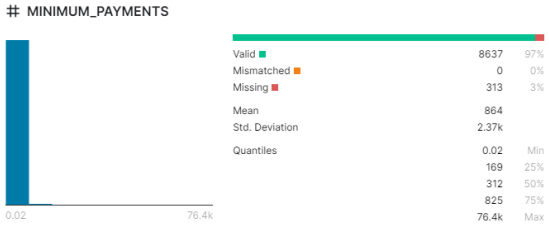

Figure 19. “MINIMUM_PAYMENTS” variable

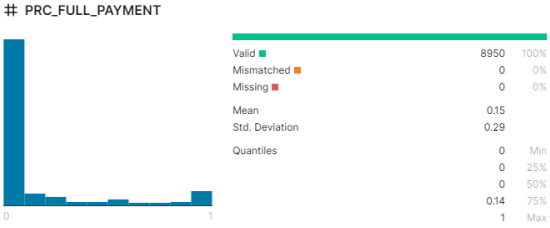

Figure 20. “PRC_FULL_PAYMENT” variable

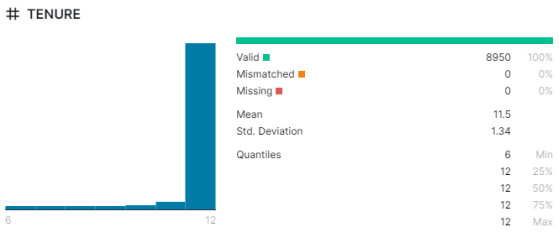

Figure 21. "TENURE" variable

B. *Data Visualization*

1) *Histogram*

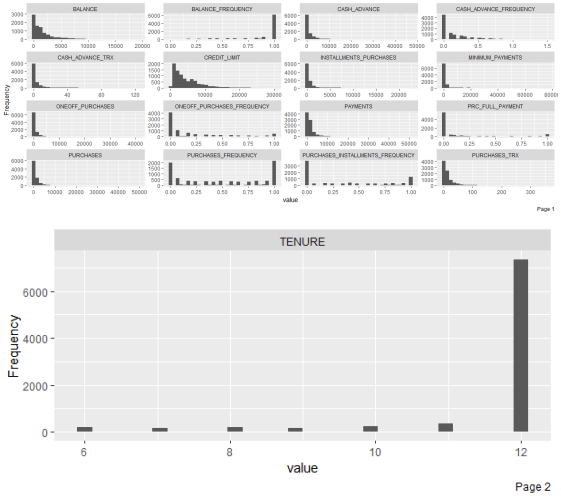

Figure 22. Dataset's Histogram

2) *Plots*

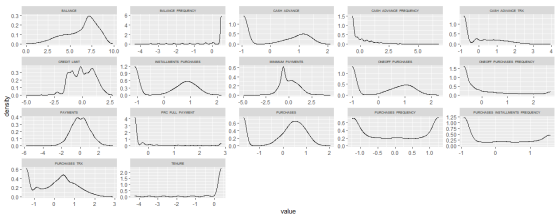

Figure 23. Dataset's Plots

3) *Correlation Plot*

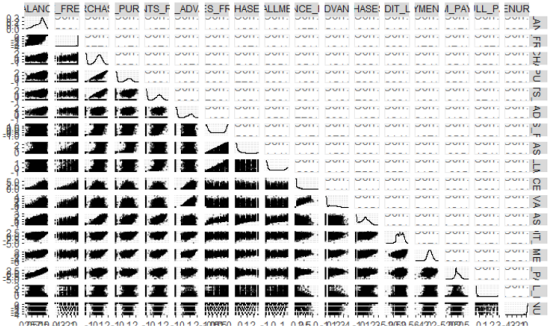

Figure 24. Dataset's Correlation Plot

4) *Plot: BALANCE variable by TENURE variable*

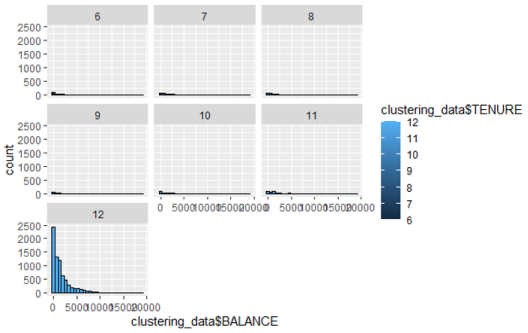

Figure 25. BALANCE variable by TENURE variable's plot

5) *Boxplot: BALANCE variable by TENURE variable*

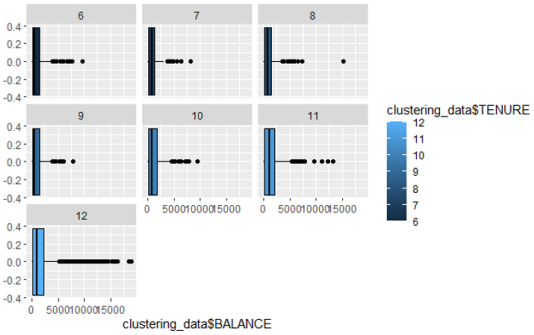

Figure 26. BALANCE variable by TENURE variable's boxplot

6) *Plot: CREDIT_LIMIT variable by TENURE variable*

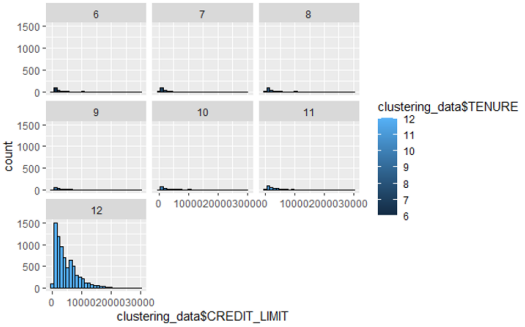

Figure 27. CREDIT_LIMIT variable by TENURE variable's plot

7) *Boxplot: CREDIT_LIMIT variable by TENURE variable*

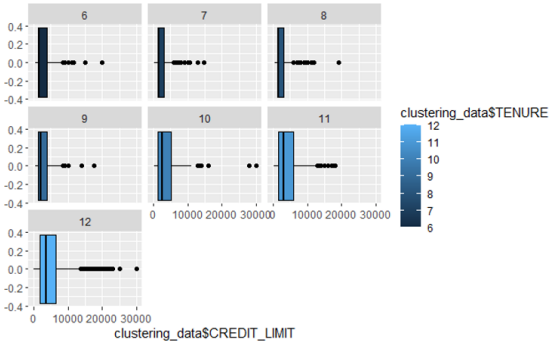

Figure 28. CREDIT_LIMIT variable by TENURE variable's boxplot

C. *Algorithms*

1) *Hierarchical Clustering*

Hierarchical clustering is an alternative approach which builds a hierarchy from the bottom-up, and doesn't require us to specify the number of clusters beforehand. So, for the application of this algorithm, it is not necessary to find the optimal k to determine the initial cluster in a data first. In hierarchical clustering, the categorize of the objects into a hierarchy similars to a tree-like diagram which is called a dendrogram. The distance of split or merge (called height) is shown on the y-axis of the dendrogram.

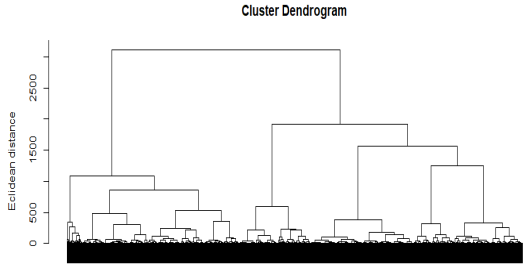

Figure 29. Hierarchical's Dendogram

The application of hierarchical clustering algorithm is an alternative approach that is done, using the "Euclidean Distance" method, because the entire data type supports and facilitates the process of applying hierarchical clustering algorithms. In addition, the algorithm also uses the method of "ward. D" is included in the Euclidean Distance method (Figure 29).

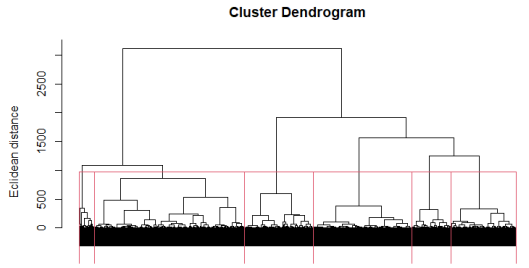

Figure 30. Hierarchical's Cluster of a Dendogram

Regarding the dendogram, 6 clusters seem to be a good size to begin the clustering. The next step that needs to be done is to "cut" the selected dendogram branches to make a total of 6 clusters (Figure 30). From a total of 6 clusters, they will formulate 6 cluster of segmentations that have been created and formulated in relation to observation variables that conclude the behavior of each credit card user.

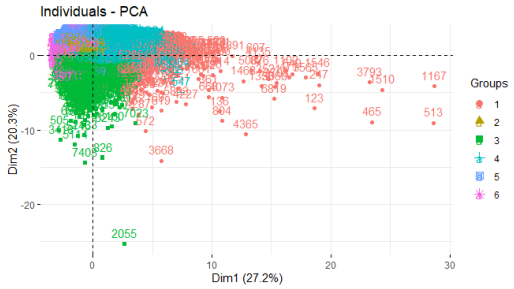

Figure 31. Hierarchical's PCA plot

The PCA plots the data in two-dimensional space (Figure 31). Overall, there are no clear clusters in the data. However, the generated clusters look quite noisy since they are overlapping. Of the 6 clusters, there is no clear cluster that shows its own position by placing a position that has its own segmentation, but rather mixed and related between one cluster to another.

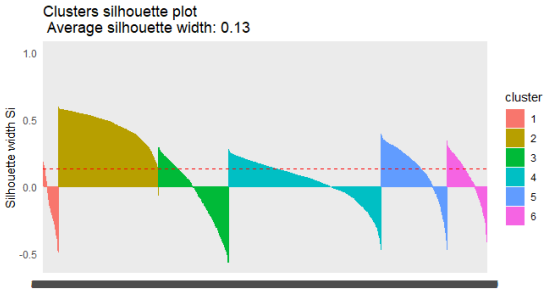

Figure 32. Hierarchical's Silhouette Plot

It shows if an observation is associated with the right (1) or wrong (-1) cluster. The average silhouette width is quite low (0.13). Many observations of 6 clusters that may be in the wrong cluster, considering that there are many overlapping clusters and no clusters are completely clear, but rather mixed up with one another (Figure 32).

2) *K-Means*

K-Means algorithm is an iterative algorithm that tries to partition the dataset into pre-defined K distinct non-overlapping subgroups (clusters), where each data point belongs to only one group. Since using K-means must specify the number of clusters K, there are many ways to specify the optimal number of K. One of them is WSS (Within Sum of Squares) plot that gonna be used in this case (Figure 33).

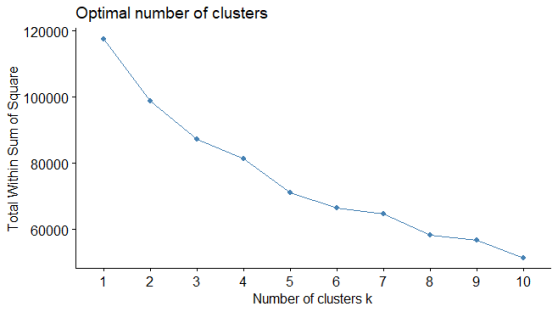

Figure 33. WSS Plot to Formulate Optimal K

Regarding to the plot, 4 clusters seem to be a proper number of clusters. Actually, there is more complete usage found in cluster numbers between 6-10, but 4 provides a more detailed description of the segmentation of each credit card user behavior which can be summarized in the majority into 4 cluster segmentations only.

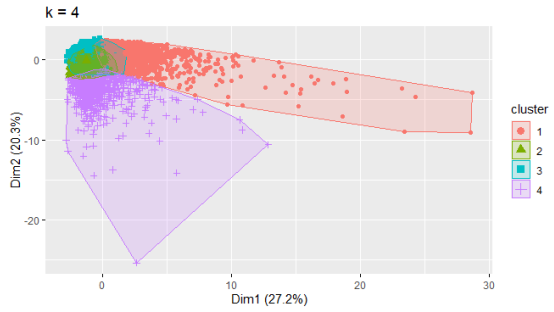

Figure 34. K-Means's Cluster Plot

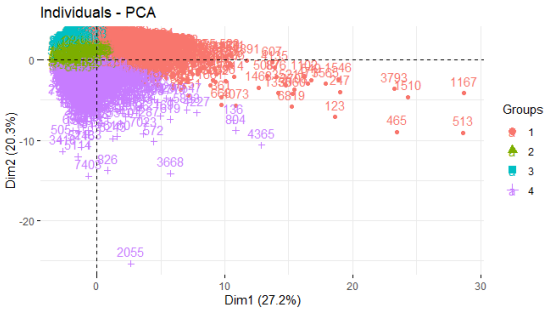

Figure 35. K-Means's PCA Plot

From the two plots above, both cluster and PCA plot, show that the cluster plot with 4 clusters using the K-Means algorithm shows better results than the hierarchical clustering algorithm by "cutting" the dendogram to 6 clusters (Figure 34). The results can be shown that from the 4 clusters displayed show better results with the presence of several clusters that have their own position so as to form their own segmentation (Figure 35). However, the presence of some overlapping and less clear clusters indicate that the cluster characteristics still cannot be maximized to indicate a good cluster.

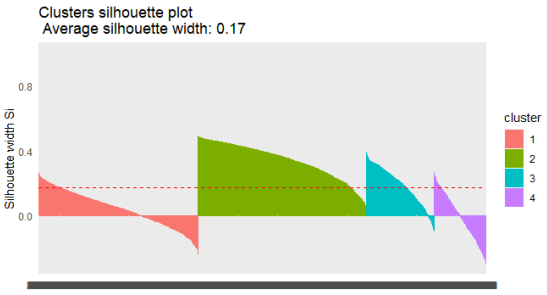

Figure 36. K-Means's Silhoutte Plot

Overall, the result is better than before. However, especially cluster 1 and 3 have still some observations which are still in the wrong cluster. But it´s the best solution with average silhouette width is about 0.17 (Figure 36).

D. Interpretation

In order to interpretate the clusters, grouped boxplots will be used that can be formed from all the clusters that have been found from the K-Means algorithm (4 clusters).

Figure 37. Interpretation's Boxplot

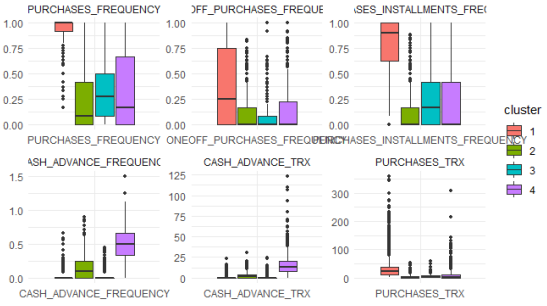

Figure 38. Interpretation's Boxplot 2

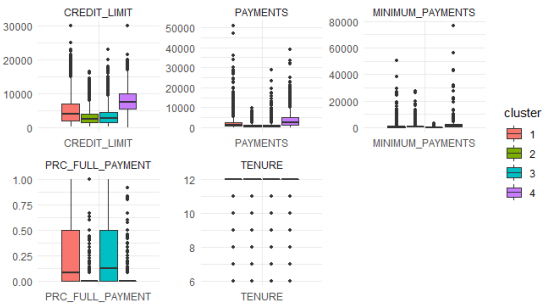

Figure 39. Interpretation's Boxplot 3

The clusters can be interpreted as follows (marketing wise):

- Cluster 1: Frequent user, with (probably) lower income that spends money mostly on consumer goods.
- Cluster 2: Frequent user, with (probably) higher income that spends money mostly on consumer goods.
- Cluster 3: Mid to rare users, with (probably) mid to high income which spends money more for higher priced products with longterm use.
- Cluster 4: Rare user, with (probably) mid to low income which spends money more on consumer goods.

V. Conclusion

Based on the results and information on each algorithm, it can be concluded that the K-Means algorithm is the best algorithm for the clustering. In general, based on data analysis activities that use of credit cards by using the two algorithms that have been described in section result analysis, the application of two algorithms: hierarchical clustering and k-means formulate two different results. However, because the plotting of the K-Means algorithm produces better clusters of 4 clusters than the hierarchical clustering algorithm of 6 clusters, then the mapping of 4 clusters provides a more centralized segmentation and is easier to interpret. The K-Means algorithm gives results with fairly clear clusters and provides separate segmentation for its customers so that they can formulate a new strategy for marketing.

The following is a description of the interpretation of the 4 clusters that have been formulated by the K-Means algorithm:

Interpretation (Marketing)	Cluster			
	1	2	3	4
Frequency	Freqeuent	Frequent	Mid-Rare	Rare
Income	Lower	Higher	Mid-High	Mid-Low
Spending	Normal	Normal	High	Normal

Table 1. Interpretation Table

The study found 4 clusters that can be interpreted as in the table above based on several behaviors, including frequency of use, income, and spending of users's needs who use credit cards. Of the 4 existing clusters, the level of frequency of use is divided into 3 types, including frequent, mid, and rare. Of course, there are some users who don't always or often use their credit cards, so this pattern is created. Then there is the income generated by each credit card user. The levels obtained are also no different, including lower, mid, and higher. Lastly, the needs and needs of each user in meeting their daily needs, which consist of lower, normal, to higher, based on the quality of goods and increasingly complex fulfillment, depending on each customer.

From the table, it can be concluded that each cluster has formulated its respective segmentation for its customers who use credit cards.

- Cluster 1: Frequent user, with (probably) lower income that spends money mostly on consumer goods.
- Cluster 2: Frequent user, with (probably) higher income that spends money mostly on consumer goods.
- Cluster 3: Mid to rare users, with (probably) mid to high income which spends money more for higher priced products with longterm use.
- Cluster 4: Rare user, with (probably) mid to low income which spends money more on consumer goods.

Segmentation of these customers can certainly formulate a new form for a useful marketing strategy. This strategy can be used for companies in providing products or services that suit the most use of credit cards. From the segmentation pattern, it can also be identified that each credit card user has different patterns in their daily lives, ranging from income, expenses, to purchasing patterns for an item, especially those with credit cards. This formulates that with or not a person has a credit card, no matter from what background and position, everyone has a purpose and fulfillment in life that is complex and unique, which only that person knows. Whether it's using a credit card or not, this case can provide an overview of a person's behavior patterns, especially in terms of finances.

But keep in mind also that this research is just a task workmanship and still has many mistakes and shortcomings. All data management and management is not necessarily true with the facts. It is expected that in the future it can be improved and developed again with more diverse data as a task workmanship and portfolio that is useful for others.

Attachments

Here is the R code that has been created in the workmanship and management of data in the dataset "Credit Card Dataset for Clustering" by Thomas Januardy (00000046001):

```r
#Thomas Januardy 00000046001
#Group B5

#Dataset: Credit Card Dataset for Clustering

#---Libraries----
library(tidyverse)  #manipulating and visualizing data (dplyr, purrr, ggplot2, knitr...)
library(readr)      #read in csv files faster
library(cluster)    #clustering algorithms and gap statistic
library(factoextra) #visualization of clustering algorithm results
library(GGally)     #create matrix of variable plots
library(NbClust)    #clustering algorithms and identification of best K
library(caret)      #find correlated variables
library(Amelia)     #missmap: check for missing data
library(DataExplorer) #plot_missing
library(dplyr)      #function %>%
library(ggplot2)    #plotting
library(caTools)    #set seed
library(reshape2)   #melt function
library(ggforce)    #interpretation of clusters


#---Read dataset----
cc <- read.csv("B5_ThomasJanuardy_00000046001.csv", header = TRUE)
#View(cc)

#---View and summarize the structures of the data---
str(cc)
glimpse(cc)

summary(cc)
```

```r
#---Remove "CUST_ID" variable (Non numeric)---
cc <- cc[-1]

#------Check for missing data------
#---check missmap---
missmap(cc)
#No missing data

#---plot missing---
plot_missing(cc)
#CREDIT_LIMIT 0.01% data missing, MINIMUM_PAYMENTS 3.5% data missing

#---Histograms plot---
plot_histogram(cc)

#---Label "TENURE" as a factor variable---
#cc$TENURE <- factor(cc$TENURE, levels = c(6,7,8,9,10,11,12),labels = c('June','July',
'August','September', 'October', 'November', 'December'), ordered = TRUE)

#---As numeric for 3 variables----
cc$TENURE <- as.numeric(cc$TENURE)
cc$PURCHASES_TRX <- as.numeric(cc$PURCHASES_TRX)
cc$CASH_ADVANCE_TRX <- as.numeric(cc$CASH_ADVANCE_TRX)

str(cc)

#---Transform variables for clustering---
transformed_variables <- c("BALANCE", "PURCHASES", "ONEOFF_PURCHASES",
"INSTALLMENTS_PURCHASES", "CASH_ADVANCE", "CASH_ADVANCE_TRX", "PURCHASES_TRX",
"CREDIT_LIMIT", "PAYMENTS", "MINIMUM_PAYMENTS") #numerical data
```

```r
#---Data Cleaning----
clustering_data <- cc %>% drop_na() #drop missing values

plot_missing(clustering_data)


#---Splitting the data (80:20)---
code <- 5 #group B5
set.seed(code) #set seed to 5

samp <- sample(nrow(clustering_data), 0.8 * nrow(clustering_data), replace = FALSE)

trainData <- clustering_data[samp, ] #training data
nrow(trainData)

testData <- clustering_data[-samp, ] #test data
nrow(testData)

#subsetting data by choose TENURE variable = 6 / 12
#subset <- trainData
#subset <- subset(trainData, TENURE == 6 | TENURE == 12)

#str(subset)
#summary(subset)


#---Log(1)'s variables for plotting---
clustering_data_log1 <- cc %>% na.omit() %>% mutate_at(vars(transformed_variables), funs
(log(1 + .))) %>% mutate_at(c(2:17), funs(c(scale(.)))) #remove any missing values in
CREDIT_LIMIT and MINIMUM_PAYMENTS, add 1 to each value to avoid log(0), scale all numeric
var to mean of 0 & sd = 1.
```

```r
#---Histograms plot (after cleaning)---
plot_histogram(clustering_data)

#---Plots---
plots <- as.data.frame(clustering_data_log1) %>%
  gather() %>%                             # make key-value pairs
  ggplot(aes(value)) +                     # values for each variable on x-axis
  facet_wrap(~ key, scales = "free") +
  geom_density() +                         # plot each as density
  theme(strip.text = element_text(size=5)) # shrink text size

plots                                      # print plots

#---Corr_plots---
corr_plots <- ggpairs(as.data.frame(clustering_data_log1),                    # GGally
::ggpairs to make correlation plots
                      lower = list(continuous = wrap("points", 
                                                     alpha = 0.3, size=0.1), # default
point size too big-shrink & change alpha
                                   combo = wrap("dot", alpha = 0.4,size=0.2)
                      )
)

corr_plots                                 # print corr_plots


#---visualization: BALANCE by TENURE---
ggplot(clustering_data, aes(clustering_data$BALANCE,  fill = clustering_data$TENURE))+
  geom_histogram(colour="black")+
  facet_wrap(~ clustering_data$TENURE)
```

```r
#---visualization: BALANCE by TENURE (Boxplot)---
ggplot(clustering_data, aes(clustering_data$BALANCE, fill = clustering_data$TENURE))+
  geom_boxplot(colour="black")+
  facet_wrap(~ clustering_data$TENURE)

#---visualization: CREDIT_LIMIT by TENURE----
ggplot(clustering_data, aes(clustering_data$CREDIT_LIMIT,  fill = clustering_data$TENURE))+
  geom_histogram(colour="black")+
  facet_wrap(~ clustering_data$TENURE)

#---visualization: CREDIT_LIMIT by TENURE (Boxplot)---
ggplot(clustering_data, aes(clustering_data$CREDIT_LIMIT, fill = clustering_data$TENURE))+
  geom_boxplot(colour="black")+
  facet_wrap(~ clustering_data$TENURE)


#---Algorithm 1: Hierarchical Clustering---

#---Euclidean distance---
trainData$TENURE <- as.numeric(trainData$TENURE) #convert TENURE variable as numeric

fit_hc_clust = hclust(dist(scale(trainData), method = "euclidean"), method = "ward")

plot(fit_hc_clust, labels = FALSE, sub = "", xlab = "", ylab = "Eclidean distance")
rect.hclust(fit_hc_clust, k = 6) #dendogram, 6 clusters

#---Choose cluster by cutting them---
hc_cluster = cutree(fit_hc_clust, k = 6) #cutting the dendogram
```

```r
#---PCA plot---
hc_pc <- prcomp(scale(trainData))

fviz_pca_ind(hc_pc, habillage = hc_cluster)

#---Silhoutte plot---
hc_sil = silhouette(hc_cluster, dist(scale(trainData), method = "euclidean"), lable = FALSE
)

fviz_silhouette(hc_sil, print.summary = FALSE) + theme_minimal()

#It shows if an observation is associated with the right (1) or wrong (-1) cluster. The
average silhouette width is quite low (0.13 on width). Many observations probably in the
wrong clusters.


#---Compare table with the original----
table_hc1 <- table(hc_cluster, trainData$TENURE)
table_hc1


#dunn index

set.seed(5)
clmethods <- c("hierarchical")
intern <- clValid::clValid(trainData, nClust = 2:6, clMethods = clmethods, validation =
"internal", maxitems = nrow(trainData))
summary(intern)
# Dunn index annd Silhouette index suggest best clustering is achieved for 2 clusters.

# Greater values of Dunn index indicates better clustering.
```

```r
#Connectivity 3.1290   hierarchical  2
#Dunn 0.2436           hierarchical  2
#Silhouette 0.9087     hierarchical  2


#---Algorithm 2: K-Means---

#---Scaling data---
#use only numerical data
cc_scaled <- trainData %>% dplyr::select_if(is.numeric)
str(cc_scaled)
cor(cc_scaled)

cc_scaled <- scale(x = cc_scaled, center = TRUE, scale = TRUE)
cc_scaled %>% head()

#find the optimal k of scaled "trainData"
fviz_nbclust(scale(trainData), kmeans, method = "wss", k.max = 10)
#k is about 4 to 7, took 4 for easier segmentations.


fit_km = kmeans(scale(trainData), centers = 4) #implementing kmeans

#cluster plot
fviz_cluster(fit_km, geom = "point", data = cc_scaled) + ggtitle("k = 4")

#pca plot
hc_pc <- prcomp(scale(trainData))
fviz_pca_ind(hc_pc, habillage = fit_km$cluster)
```

```r
#cluster silhouette plot
hc_sil = silhouette(fit_km$cluster, dist(scale(trainData), method = "euclidean"), lable =
FALSE)

fviz_silhouette(hc_sil, print.summary = FALSE) + theme_minimal()

#dunn index
clmethods2 <- c("kmeans")
intern2 <- clValid::clValid(trainData, nClust = 2:6, clMethods = clmethods2, validation =
"internal")
summary(intern2)


#---interpretation---

#grouping and factoring the cluster
c = trainData

c$cluster = fit_km$cluster

c_plots = melt(c, id.var = "cluster")

c_plots$cluster = as.factor(c$cluster)
```

```r
#plotting
c_plots %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot(aes(fill = cluster), outlier.size = 1) +
  facet_wrap_paginate( ~ variable, scales = "free", ncol = 3, nrow = 2, page = 1) +
  labs(x = NULL, y = NULL) +
  theme_minimal()


c_plots %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot(aes(fill = cluster), outlier.size = 1) +
  facet_wrap_paginate( ~ variable, scales = "free", ncol = 3, nrow = 2, page = 2) +
  labs(x = NULL, y = NULL) +
  theme_minimal()

c_plots %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot(aes(fill = cluster), outlier.size = 1) +
  facet_wrap_paginate( ~ variable, scales = "free", ncol = 3, nrow = 2, page = 3) +
  labs(x = NULL, y = NULL) +
  theme_minimal()
```


Thank-you Note

This paper will never be completed without the help and guidance of Mrs. Tan Thing Heng, Bsc, Mstats as lecturer of the course of "Data Analysis", Information Systems major, Multimedia Nusantara University. The author would like to thank Mrs. Tan Thing Heng for the availability of time and consultation sessions in the work of this task and paper by providing solutions and guidance so it makes and paper can be completed properly.

References

[1] Davenport, T. H., Barth, P., & Bean, R. (2012). How 'Big Data' is Different. *MIT Sloan Management Review*, *54*(1), 22.

[2] Davenport, T. H., & Dyché, J. (2013). Big Data in Big Companies. *International Institute for Analytics*.

[3] Davenport, T. H. (2013). Big Data: What it is and why it matters. *SAS.* https://www.sas.com/en_id/insights/big-data/what-is-big-data.html

[4] Zulkarnain, N., & Anshari, M. (2016). Big Data: Concept, Applications, & Challenges. *International Conference on Information Management and Technology (ICIMTech),* 307-308.

[5] Schönberger, V. M., & Cukier, K. (2013). Big Data: A revolution that will transform how we live, work, and think. *New York, NY: Houghton Mifflin Harcourt*.

[6] Watson, H. J. (2014). Tutorial: Big Data Analytics: Concept, technology and application. *Communication for the association for information systems, 34*(65), 1247-1268.

[7] Juliantama, A. (2020). Credit Card Fraud Identification Using Machine Learning. *Sistem dan Teknologi Informasi Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung*, 1.

[8] Muliono, R. (2017). Implementasi Algoritma Apriori Pada Data Benchmark Kosarak Dan Mushrooms. *JITE : JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING, 1*(1), 34-41.

[9] E, M., & S, K. (2015). Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang. *Jurnal Bianglala Informatika, 3*(1).

[10] Muliono, R., & Sembiring, Z. (2019). DATA MINING CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS UNTUK KLASTERISASI TINGKAT TRIDARMA PENGAJARAN DOSEN. *CESS (Journal of Computer Engineering System and Science), 4*(2), 273-274.

[11] Budi, S. (2007). Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. *Graha Ilmu*.

[12] Muhidin, A. (2017). ANALISA METODE HIERARCHICAL CLUSTERING DAN K-MEAN DENGAN MODEL LRFMP PADA SEGMENTASI PELANGGAN. *SIGMA: Jurnal Teknologi Pelita Bangsa, 7*(1).

[13] Alfina, T., Santoso, B., & Barakbah, A. R. (2012). Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS). *JURNAL TEKNIK ITS, 1.*

[14] Februariyanti, H., & Santoso, D. B. (2017). HIERARCHICAL AGGLOMERATIVE CLUSTERING UNTUK PENGELOMPOKAN SKRIPSI MAHASISWA. *Prosiding SINTAK 2017.*

[15] Sari, P., Pramono, B., & Sagala, L. O. H. S. (2017). IMPROVE KMEANS TERHADAP STATUS NILAI GIZI PADA BALITA. *SemanTIK, 3*(1), 143-148.