ILKOM Jurnal Ilmiah Vol. 14, No. 3, December 2022, pp. 284-293
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Implementation of Data Mining using *K-Means Algorithm* for bicycle sales prediction

Ivan Anggriawan [a,1,*]; Wawan Gunawan [a,2]

[a]*Universitas Mercu Buana, Jl. Meruya Selatan No. 1, Jakarta and 11650, Indonesia*
[1]*41516320024@student.mercubuana.ac.id ; [2]wawan.gunawan@mercubuana.ac.id*

** Corresponding author*

Article history: Received November 15, 2022; Revised July 04, 2022; Accepted November 29, 2022; Available online December 20, 2022

Abstract

During the pandemic, to reduce the number of Covid-19 spreads, the government imposed social distancing and work from home (WFH) to reduce community activities outside the home. This rule caused people to have irregular patterns or lifestyles which less physical activity. It can lower the immunity system, increasing the risk of the virus's infection. Therefore, during the pandemic, sports or exercises become one of the activities regularly carried out by the community to increase their immunity. One of the sports activities that can be done to maintain their immunity is cycling. Cycling itself is a light activity that all ages can practice. This occasion is certainly a good marketing target for bicycle-selling companies, but the company sometimes needs help with bicycle stocks that do not match the consumer market target. The purpose of this study is to find out what types of bicycles are in demand by predicting bicycle sales and looking at the desired interests of the community. This study uses the K-Means Clustering algorithm. The results of the K-Means Clustering research are divided into three clusters; Cluster 1, with 209 members with the most interest in mountain bikes; cluster 2, with 787 members with the most interest in folding bicycles; and Cluster 3, with 540 members with bicycle interests. Most of them are city bicycles; from the clustering process above, the Dunn Index validation (Dunn Index) can be obtained with a value of 0.1324532.

Keywords: Bike sales prediction; Data Mining; K-Means; Clustering; RStudio

Introduction

Information technology and business development are developing rapidly, encouraging humans to take advantage of information technology according to their needs. During the COVID-19 pandemic as it is today, it is undeniable that the presence of information technology plays an important role [1] in companies maintaining and developing their businesses.

In the current conditions, exercise is one of the physical activities that must be done during the Covid-19 pandemic. People must remain active even though they work from home. Social distancing and working from home (WFH) tend to make a person have a sedentary lifestyle; studies show that a sedentary lifestyle can reduce the body's immunity, thereby increasing the risk of viral infections (Association of Sports Specialists, 2020) [2]. During the pandemic, cycling is crucial to maintain immunity, so the body is immune from viruses. Cycling is a simple activity carried out by everyone, both children, adults, and the elderly [3]; bicycle-selling companies use this momentum.

The current problem is the difficulties experienced by bicycle-selling companies, where they have stock that needs to follow the target consumer market. Every company, of course, targets the sales it wants to achieve every day, month, or year. Companies need sales forecasting, which can be searched by using trends or predictions to estimate how many sales of their stock types are likely to occur in the coming year. An important factor in sales is to predict the demand for consumer orders that are needed and not excessive [4].

Several studies using K-Means Clustering, including Sufajar Butsianto[1], Nindi Tya Mayangwulan[2], who researched the Application of Data Mining for Car Sales Prediction Using the K-Means Clustering Method. In this study, the dataset used was car sales data. The data collected is car sales data from Gaikindo (Association of Indonesian Motor Vehicle Industries). The data obtained is calculated from data on car sales results in Indonesia from 2015-2019 [5].

https://dx.doi.org/10.33096/ilkom.v14i3.1291.284-293

Then the following research is Najia Salsabila's entitled Classification of Goods Using the K-Means Clustering Method in Determining Stock Predictions of Goods (Case Study: Ukm Mar'ah Jilbab Kediri). The data used is in the form of quantitative data. The data is from historical sales transactions obtained directly from the object through interviews and documentation [6].

In this study, the author will predict the sales of bicycles that are most in demand in the province with the aim that business people or bicycle-selling companies in the region can further optimize the stock of bicycles they have to suit the target market of consumers. The author hopes it will increase revenue from the proceeds of bicycle sales that match the interests of consumers. The author used the clustering method with the K-Means Clustering algorithm in this study. K-Means Clustering is a technique in data clusters that is very well known for its speed in clustering data. However, K-Means Clustering has weaknesses in processing data with many dimensions. Especially for inputs that are non-linearly separable [7].

Method

In this study, quantitative research methods were used, which were made systematically, planned, and structured from the beginning to the creation of the design. The following are the methods of data collection carried out in this study:

1. Interview : Conduct direct interviews with bicycle users and bicycle sellers so that data is obtained precisely and accurately [8].
2. Literature Study : Writing is done by studying various kinds of literature references. Literature study is a data collection technique by collecting and analyzing written and electronic documents related to the title of this final project to become the basis for research and data sources in the implementation of the K-means algorithm clustering in data grouping [8][9].
3. Questionnaire : At this stage, the activities are collecting data by distributing online questionnaires in the form of google forms and offline questionnaires in the form of printed forms.

The data analysis stage, as can be seen in **Figure 1**.

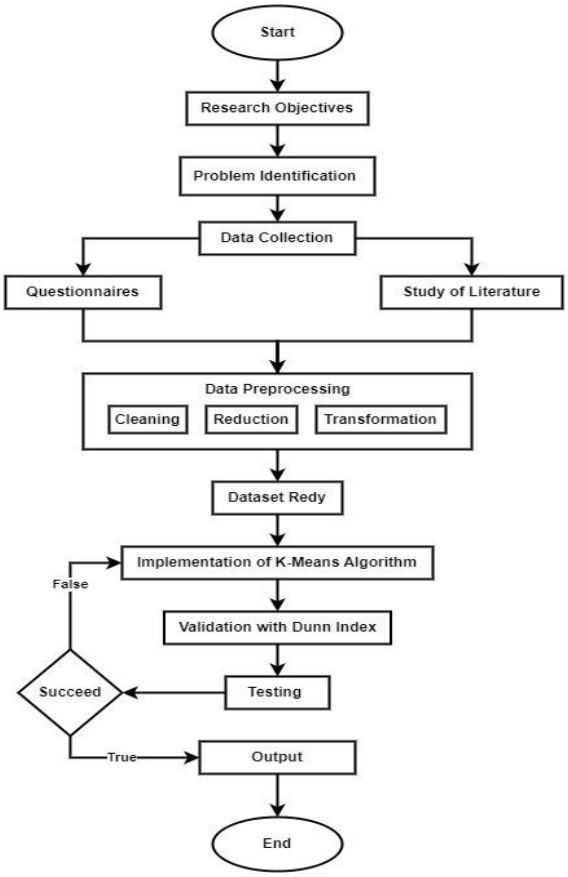

Figure 1. Research Flowchart

A. Clustering

One of the techniques known in data mining is clustering. Understanding scientific clustering in data mining is the grouping of several data or objects into clusters (groups) so that each cluster will contain data that is as similar as possible and different from objects in other clusters [10][11]. The most widely used clustering method is the K-Means clustering method. The main drawback of this method is that the results are sensitive to the selection of the initial cluster center and the calculation of local solutions to achieve optimal conditions. *Cluster analysis* is a multivariate technique with the main objective of grouping objects based on their characteristics. Cluster analysis classifies objects so that each object with the closest similarity to another object is in the same cluster [12].

B. K-Means Algorithm

K-Means can group large amounts of data with relatively fast and efficient computation time. However, K-Means has a weakness caused by determining the initial center of the cluster. The results of the cluster formed from the K-Means method are very dependent on the initiation of the initial center value of the given cluster [5][13]. The clustering process can be started by identifying grouped data using the Euclidean Distance formula.

At this stage implement the k-means algorithm in the coding that we run in R.Studio with the following steps:

a. Specify the value of K as the number of clusters you want to form.
b. Randomly generate the initial centroid (cluster center point).
c. Calculate the distance of each data to each centroid using the correlation formula between two objects (Euclidea Distance), follow **Equation 1**:

$$D_{(a,b)} = \sqrt{\sum_{k=1}^{n}(a_k - b_k)^2} \qquad (1)$$

Where : n is the sum of dimensions (attributes)
ak and bk are the k-th attributes of the data objects *p* and *q*

d. Group each data based on the closest distance between the data and its centroid.
e. Determine the position of the new centroid (k n) by calculating the average value of the data present on the same centroid. Where (n K) is the number of documents in cluster k and I is the document in cluster k, follow **Equation 2**:

$$C_k = \left(\frac{1}{n_k}\right)\sum 1 \qquad (2)$$

f. Return to step (c) if the position of the new centroid with the old centroid, is not the same.

C. Dunn Index

Dunn index is a metric to evaluate the results of clustering. The Dunn Index calculation carried out in this study is calculated based on the average cosine similarity of a title to other titles in a topic group [14].

Dunn index is calculated based on the following **Equation 3**:

$$D_{nc} = {}^{min}_{i=1,..nc}\left\{ {}^{min}_{j=i+1,..,nc}\left(\frac{d(c_i,c_j)}{max_{k=1...,nc}\, diam_{(ck)}}\right)\right\} \qquad (3)$$

Where d(ci, cj) is an unequal function between the cluster ci and cj which is defined as, **Equation 4**:

$$d(c_i,c_j) = {}^{min}_{x\in c_i, y\in c_j}\ d(x,y) \qquad (4)$$

and stationary(C) is the diameter of the cluster that may be considered as a measure of the dispersion of the cluster. Cluster diameter C can be defined as, **Equation 5** :

$$diam(C) = {}^{max}_{x,y\in c}\ d(x,y) \qquad (5)$$

D. Programming Language

R programming is an integrated software facility for data manipulation, simulation, calculation, and graphical demonstration. R can analyze data very effectively and is equipped with array and matrix processing operators. R has graphical display capabilities for displaying its data [15].

Results and Discussion

A. Research Data Set

In this study, the number of datasets used was 1,536 data. The data was obtained from the results of the recap of the distribution of online questionnaires in the form of google forms and offline questionnaires in the form of printed

forms regarding community interest in bicycles in the Yogyakarta area (Yogyakarta City, Sleman Regency, Kulon Progo Regency, Bantul Regency, and Gunung Kidul Regency). The following is the data obtained from the results of the distribution of questionnaires as in **Table 1** :

Table 1. Data Set Results

No	Name	Gender	Age	Domicile	Budget
1	Medi Kusnadi	Male	<20	Bantul Regency	Rp. 2.000.000 - Rp 4.000.000
2	Rohmat	Male	21-30	Sleman Regency	Rp. 4.000.000 - Rp 6.000.000
3	Yusuf	Male	21-30	Bantul Regency	Rp. 4.000.000 - Rp 6.000.000
…	…	…	…	…	…
1535	Sanda putra	Male	21-30	Bantul Regency	Rp. 2.000.000 - Rp 4.000.000
1536	Jesika	Female	21-30	Bantul Regency	Rp. 4.000.000 - Rp 6.000.000

The next stage of the data on **Table 1** will be changed to the data so that it becomes easy to process data. Because some of the data attributes used are non-numeric data, the need to change the data to numeric or by initiating the data can be seen in **Table 2**.

Table 2. Initialize Variable Data to Be Numeric

Variable	Inisialisasi
Budget	
Under Rp. 2.000.000	1
Rp. 2.000.000 - Rp 4.000.000	2
Rp. 4.000.000 - Rp 6.000.000	3
Rp. 6.000.000 - Rp. 8.000.000	4
Diatas Rp. 8.000.000	5
Domicile	
Yogyakarta City	1
Bantul Regency	2
Sleman Regency	3
Kulon Progo Regency	4
Gunung Kidul Regency	5
Gender	
Male	1
Female	2
Age	
<20	1
21-30	2
31-40	3
41-50	4
50>	5

Table 3 below shows the initialization result in **Table 1** of the data set results. The initialization of the numbering helps facilitate the process of processing clustering data. In contrast, in the clustering process in R, the data that can be processed is only numerical or numerical. Therefore, before the data is processed, it is necessary to initialize it so the system can read it. Based on the data obtained, the results of data initialization from the variables in **Table 2** are as follows:

Table 3. Data that has been initialized to numeric

No	Gender	Age	Domicile	Budget
1	1	1	2	2
2	1	2	3	3
3	1	2	2	3

No	Gender	Age	Domicile	Budget
…	…	…	…	…
1535	1	2	2	2
1536	2	2	2	3

B. Proses K-Means Clustering

The next stage of the clustering process

1. Determine the number of Clusters; the first stage is to determine the number of clusters; this system will produce three groups that are identified, *cluster 1*, *cluster 2*, and *cluster 3*.
2. Determine the initial centroid value; the data used has been initialized by numeric or numbering; this initialization facilitates the process of clustering data. Next is to determine the initial centroid value by taking data representing the cluster specified in iteration 1; the initial centroid value of the data that has been initialized into numeric can be seen in **Table 4**.

Table 4. Early Centroid Iteration 1

Kelas	Gender	Age	Domicile	Budget
C_1	2	1	2	2
C_2	1	3	3	2
C_3	5	4	3	2

In Calculate the distance of each existing data to each Cluster, after determining the initial centroid value, the next step is to calculate the distance of each existing data to each Cluster.

$$D_{(x,y)} = ||x - y|| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \text{ ; i = 1,2,3…,n}$$

Notes:

(x) is a data object

(y) is a centroid

Calculate the distance of the Village in each *cluster.*

1) d1,c1 (data 1, centroid 1)

$$d = \sqrt{(1-2)^2 + (1-2)^2 + (2-2)^2 + (2-1)^2}$$

$$= \sqrt{3} = 1.73 \quad \text{(C1)}$$

2) d1,c2 (data 1, centroid 2)

$$d = \sqrt{(1-2)^2 + (1-3)^2 + (2-1)^2 + (2-3)^2}$$

$$= \sqrt{7} = 2.64 \quad \text{(C2)}$$

3) d1,c3 (data 1, centroid 3)

$$d = \sqrt{(1-2)^2 + (1-3)^2 + (2-5)^2 + (2-4)^2}$$

$$= \sqrt{18 = 4.24} \quad \text{(C3)}$$

Determine the members of a cluster; the next step is to determine whether data will be a member of a cluster that has the smallest distance from the center of its cluster. Suppose that for the first data, the smallest distance is obtained at C1 so that the first data will be a member of C1. Perform this Process until the new iteration is the same as the initial iteration.

C. Implementation of K-Means Clustering in RStudio

Researchers will use RStudio software to process data in this implementation and testing. After the bicycle dataset is changed to numerical, a clustering process is carried out with Kmeans in the Rstudio software. The first step is to

enter the bicycle dataset into Rstudio by clicking import dataset and selecting from excel, such as **Figure 2** and **Figure 3.**

```
#### Memanggil data ---------
library(readxl)
data <- read_excel("F:/DATA IVAN/KULIAH/TUGAS AKHIR/Data Hasil Olahan kmeans di R/Dataset Redy 1.xlsx",
                   col_types = c("numeric", "text", "numeric",
                                 "numeric", "numeric", "numeric",
                                 "numeric", "numeric", "numeric",
                                 "numeric", "numeric", "numeric"))
```

Figure 2. Code display bike dataset

Tabel 5. Bike dataset results

No	Name	Gender	Age	Domicile	Budget	Mountain bike	Racing bike	Folding bike	BMX bike	City Bike
1	Medi Kusnadi	1	1	2	2	3	2	3	0	1
2	Rohmat	1	2	3	3	2	3	3	2	2
3	Yusuf	1	2	2	3	2	3	2	2	1
…	…	…	…	…	…	…	…	…	…	…
1535	Sanda putra	1	2	2	2	2	3	1	1	3
1536	Jesika	2	2	2	3	2	1	3	1	1

Next, it displays summary information from the dataset used, with syntax as in **Figure 3.**

```
summary(data)   #menampilkan rincian data tabel
```

Figure 3. Code summary data

```
> summary(data)   #menampilkan rincian data tabel
       No             Nama              Gender          Umur          Domisili         Budget
 Min.   :   1.0   Length:1536        Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.: 384.8   Class :character   1st Qu.:1.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
 Median : 768.5   Mode  :character   Median :1.00   Median :2.000   Median :2.000   Median :2.000
 Mean   : 768.5                      Mean   :1.37   Mean   :2.322   Mean   :2.371   Mean   :2.022
 3rd Qu.:1152.2                      3rd Qu.:2.00   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:2.000
 Max.   :1536.0                      Max.   :2.00   Max.   :5.000   Max.   :5.000   Max.   :5.000
 SEPEDA GUNUNG    SEPEDA BALAP    SEPEDA LIPAT     SEPEDA BMX     SEPEDA KOTA
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
 Median :2.000   Median :2.000   Median :2.000   Median :1.000   Median :2.000
 Mean   :1.559   Mean   :1.604   Mean   :1.595   Mean   :1.346   Mean   :1.593
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :3.000
```

Figure 4. Descriptive results of the table attributes used

Figure 4 Displays the results of the descriptive on the table attribute to be used. In the summary results of the total 11 columns, there is no NA value or missing data because if there is an empty value or no value, we need to discard the missing data so that the data can be processed. Because there is no NA value in the table data, the clustering process can be continued. Then take the columns that will be clustered; the columns that will be clustered include gender, age, domicile, budget, mountain bikes, racing bikes, folding bikes, BMX bikes, and city bikes with syntax as in **Figure 5** and the attributes that will be used in **Figure 6**.

```
data.numerik = data.frame(data[3:11]) #Mengambil kolom yang digunakan
data.numerik #menampilkan data kolom
```

Figure 5. Code retrieves a column in a table

```
> data.numerik = data.frame(data[3:11]) #Mengambil kolom yang digunakan
> data.numerik #menampilkan data kolom
   Gender Umur Domisili Budget SEPEDA.GUNUNG SEPEDA.BALAP SEPEDA.LIPAT SEPEDA.BMX SEPEDA.KOTA
1       1    1        2      2             3            2            3          0           1
2       1    2        3      3             2            3            3          2           2
3       1    2        2      3             2            3            2          2           1
4       2    2        2      1             1            1            1          1           3
5       1    3        5      3             2            2            1          1           1
6       1    4        4      2             3            2            0          2           3
7       1    3        4      3             2            3            1          1           3
8       2    2        2      2             2            2            1          2           2
9       1    2        3      3             2            2            2          2           1
10      2    2        1      2             1            2            2          1           3
```

Figure 6. Table attribute used

The next stage is to look for the value of the distance or distance between objects. To find the distance value between objects can be seen with syntax as in **Figure 7** and the result of the distance value between objects in **Figure 8.**

```
#### Cek Distance ----------
distance <-get_dist(data.numerik) #jarak antar objek satu dengan yang lain
distance
```

Figure 7. syntax looking for distance values

```
> distance
          1         2         3         4         5         6         7         8         9        10        11
12       12        13        14        15        16        17        18        19        20        21        22
23       23        24        25        26        27        28        29        30        31        32        33
34       34        35        36        37        38        39        40        41        42        43        44
45       45        46        47        48        49        50        51        52        53        54        55
56       56        57        58        59        60        61        62        63        64        65        66
67       67        68        69        70        71        72        73        74        75        76        77
78       78        79        80        81        82        83        84        85        86        87        88
89       89        90        91        92        93        94        95        96        97        98        99
100     100       101       102       103       104       105       106       107       108       109       110
111     111       112       113       114       115       116       117       118       119       120       121
122     122       123       124       125       126       127       128       129       130       131       132
133     133       134       135       136       137       138       139       140       141       142       143
```

Figure 8. The result of the distance value between objects

Before carrying out the K-means cluster stage, it is necessary to determine in advance the number of clusters to be formed using the elbow or sillhouete method. The results of determining the number of clusters can both be seen in **Figure 9** and **Figure 10**.

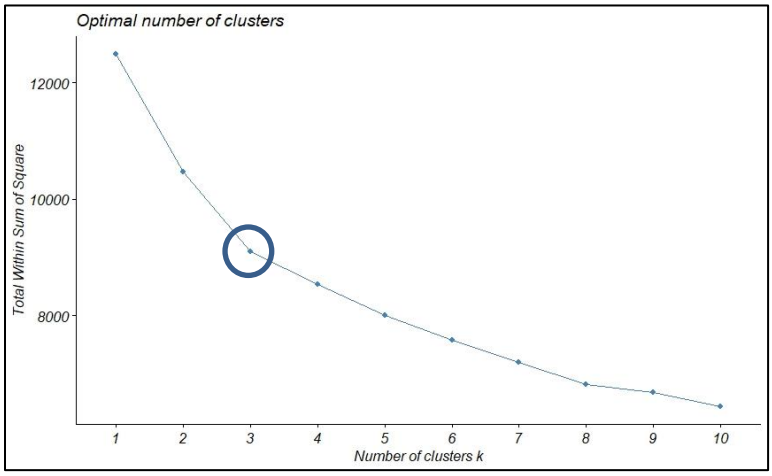

Figure 9. Determination of the number of clusters with elbow

On kmeans, there are considerations to determine the most optimal many k groups. In **Figure 9**, a graph of the elbow method is depicted, which shows the best k value is when k is at number 3. It happens because the best cluster value is when there is a decrease in some cluster values; namely, the line experiences a fault that forms an elbow or elbow when k= three, and then the result of the cluster value will stabilize or slowly decrease. So using this method is obtained the optimal k at the time of being at k= 3.

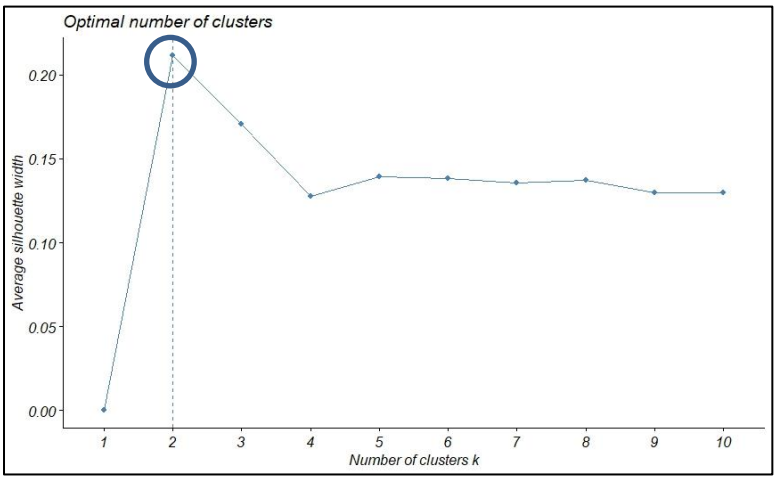

Figure 10. Determination of the number of clusters with sillhouete

The approach averages the value of the sillhouete method to guess the quality of the formed clusters. The higher the average value, the better. Based on the graph in **Figure 10**, many optimal clusters are formed at k = 2. Based on the number of clusters using the elbow method, the result obtained k optimal is at k = 3, and the result of sillhouete obtained k optimal is at k = 2. However, this study will divide into 3 clusters. This study used center = 3 and iteration 25 times, namely with syntax, as in **Figure 11**.

```
kmean <- eclust(data.numerik, "kmeans", k = 3, nstart = 25, graph = FALSE)
kmean
k=data.frame(kmean$cluster)
View(k)
k
```

Figure 11. Clustering syntax with K-means

```
K-means clustering with 3 clusters of sizes 209, 787, 540

Cluster means:
    Gender     Umur Domisili   Budget SEPEDA.GUNUNG SEPEDA.BALAP SEPEDA.LIPAT SEPEDA.BMX SEPEDA.KOTA
1 1.263158 2.320574 2.607656 4.669856      1.669856     1.612440     1.645933   1.373206    1.650718
2 1.398983 2.294790 1.426938 1.637865      1.561626     1.738247     1.640407   1.213469    1.594663
3 1.368519 2.361111 3.655556 1.557407      1.512963     1.403704     1.509259   1.527778    1.568519
```

Figure 12. Results of clustering with K-means

In **Figure 12**, using 3 clusters with the size obtained if you use 3 clusters, namely 209, 787 and 540. Where for the average value of variable cluster 1 on gender is 1.263158, age is 2.320574, domicile is 2.607656, budget is 4.669856, mountain bike is 1.669856, race bike is 1.612440, folding bike is 1.645933, BMX bike is 1.373206, city bike is 1.650718. In cluster 2 gender of 1.398983, age of 2.294790, domicile of 1.426938, budget of 1.637865, mountain bike of 1.561626, race bike of 1.738247, folding bicycle of 1.640407, BMX bicycle of 1.213469, city bicycle of 1.594663. In cluster 3 gender of 1.368519, age of 2.361111, domicile of 3.655556, budget of 1.557407, mountain bike of 1.512963, race bike of 1.403704, folding bicycle of 1.509259, BMX bicycle of 1.527778, city bicycle of 1.568519.

```
Within cluster sum of squares by cluster:
[1] 1486.115 4325.578 3292.306
 (between_SS / total_SS =  27.1 %)
```

Figure 13. Results within cluster sum of squares

Figure 13 within the cluster sum of squares is the distance between objects in the cluster. It can be seen that the distance for cluster 1 is 1496.115, cluster 2 is 4235.578, and cluster 3 is 3292.306, so the distance value is 27.1%. follow **Table 6**.

Table 6. Syntax Description

	Description
Cluster of sizes	size/number of data points on each cluster
Cluster means	average value (centroid) of each cluster
Within_ss	total sum of squares for each cluster
Tot.within_ss	total summation of each ss of within

Total_ss	total sum of squares for the entire data point
Between_ss	the difference in value between total_ss and tot.within_ss
Between_ss/total_ss	the ratio between between_ss is divided by total_ss or also called variance. The greater the percentage, generally the better

Cluster Validity, is a test evaluation process that provides a list of performance criteria values based on cluster centroids, which is useful for knowing how well the clustering process is performing. In this study, cluster evaluation was carried out using the dunn index on the RStudio device. The results of the evaluation can be seen in **Figure 14**.

```
> kmean_stats$dunn
[1] 0.1324532
```

Figure 14. Dunn Indexs Evaluation Results

The results in **Figure 14** show the evaluation value of the validity index, resulting in a dunn index = 0.1324532. For model evaluation, it can be seen from the *average values within* and *average values between clusters*. A good *cluster* is one that has a very small *average within* and has a very large *average between.* The following are given the *average values within* and *average values between* K-Means.

```
#### Nilai av within dan between -----
kmean_stats$average.within
kmean_stats$average.between
```

Figure 15. Syntax Of Within and Between Values

```
> kmean_stats$average.within
[1] 3.342471
> kmean_stats$average.between
[1] 4.317231
> kmean_stats$wb.ratio
[1] 0.7742164
```

Figure 16. Result Value Within, Between and Ratio

Based on **Figure 16**, it is obtained that the *average value within K-Means* of 3.342471 is smaller than the average value between K-Means of 4.317231. So it can be said that the cluster model formed is good. As for the ratio value, it is 0.7742164. After finishing and getting the cluster results, save the cluster result file with xls format. As in **Figure 17**.

```
#### Menyimpan hasil K --------
hasil<-cbind(data.numerik, k)
setwd("F:/DATA IVAN/KULIAH/TUGAS AKHIR/Data Hasil Olahan kmeans di R")
write.csv2(hasil, file = "hasil olahan1.xls")
```

Figure 17. Saving Cluster results

After cluster analysis using the k-means clustering and Dunn index methods, it can be concluded that the results of k-means clustering with Dunn index validation formed 3 clusters with a number of cluster 1 as many as 209 members, cluster 2 as many as 787 members and cluster 3 as many as 540 members with an index value of 0.1324532.

Conclusion

After conducting cluster analysis using the k-means clustering and Dunn index methods, it can be concluded that the results of k-means clustering with Dunn index validation formed 3 clusters. Cluster 1 is 209 members, cluster 2 is 787 members, and cluster 3 is 540 members with an index value of 0.1324532. In Cluster 1, most of them are male (154 people), the dominant age ranges from 21-30 years, the most domiciled in the Yogyakarta area, the average budget spent on buying a bicycle is above 8 million, consumers in this cluster want to use bicycles for sports activities with a total of 131 votes. In cluster 1, the type of bicycle most interested in by consumers is mountain bikes, with a total interest of 121, more significant than other types of bicycles. In Cluster 2, dominant males (473 people) ranging from 21-30 years old, the most domiciled in the Kulon Progo area, the average budget spent on buying a bicycle is under 2 million; consumers in this cluster want to use bicycles for sports activities with a total vote of 465. In cluster 2, the types of bicycles most interested in by consumers are folding bicycles, with a

total interest of 430, more significant than other types of bicycles. In Cluster 3, most of them are male (341 people), ranging from 21-30 years old, and most are domiciled in the Sleman area; the average budget spent on buying a bicycle is under 2 million, and consumers in this cluster want to use bicycles for sports activities with a total vote of 300. In cluster 3, the types of bicycles most interested in by consumers are city bicycles, with a total interest of 290, more significant than other types of bicycles.

References

[1] R. Komalasari, "Manfaat teknologi informasi dan komunikasi di masa pandemi COVID 19," *Tematik*, vol. 7, no. 1, pp. 38–50, 2020, doi: 10.38204/tematik.v7i1.369.

[2] F. K. Hadi, "Aktivitas olahraga bersepeda masyarakat di kabupaten Malang pada masa pandemi COVID-19," *Sport Sci. Educ. J.*, vol. 1, no. 2, pp. 28–36, 2020, doi: 10.33365/ssej.v1i2.777.

[3] T. Hidayat, M. Hudah, and U. H. Zhannisa, "Survey minat masyarakat untuk olahraga rekreasi bersepeda pada masa pandemi COVID 19 di kabupaten Demak," *J. Phys. Act. Sport.*, vol. 1, no. 1, pp. 80–88, 2020, doi: 10.53869/jpas.v1i1.17.

[4] W. S. Herlambang.L, "Analisis peramalan penjualan sepeda dan motor listrik di PT XYZ," *J. Comasie*, vol. 1, no. 1, pp. 130–138, 2021.

[5] S. Butsianto and N. T. Mayangwulan, "Penerapan data mining untuk prediksi penjualan mobil menggunakan metode K-Means clustering," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 3, no. 3, pp. 187–201, 2020, doi: 10.32672/jnkti.v3i3.2428.

[6] N. Salsabila, "Klasifikasi barang menggunakan metode clustering K-Means dalam penentuan prediksi stok barang," *Cent. Libr. Maulana Malik Ibrahim State Islam. Univ. Malang*, p. 89, 2018, [Online]. Available: http://etheses.uin-malang.ac.id/16985/1/14650031.pdf.

[7] I. Vhallah, S. Sumijan, and J. Santony, "Pengelompokan mahasiswa potensial drop out menggunakan metode clustering K-Means," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 2, pp. 572–577, 2018, doi: 10.29207/resti.v2i2.308.

[8] W. Gunawan and B. S. P. Diwiryo, "Implementasi algoritma Fuzzy C-Means clustering sistem crowdfunding pada sektor industri kreatif berbasis Web," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 2, p. 193, 2020, doi: 10.26418/jp.v6i2.38018.

[9] R. Desrianti and H. D. Wijaya, "Implementasi algoritma Fuzzy C-Means pada aplikasi seleksi karyawan digital talent di PT Telekomunikasi Indonesia," *J. Media ...*, vol. 4, pp. 879–888, 2020, doi: 10.30865/mib.v4i4.2267.

[10] D. W. Sari, "Penentuan kriteria dalam memilih sekolah dasar dengan menerapkan K-Means Clustering (Studi Kasus : Wilayah Kecamatan Mampang)," vol. 29, no. 2, pp. 24–28.

[11] W. Utomo, "The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia," *Ilk. J. Ilm.*, vol. 13, no. 1, pp. 31–35, 2021, doi: 10.33096/ilkom.v13i1.763.31-35.

[12] K. Handoko, "Penerapan data mining dalam meningkatkan mutu pembelajaran pada instansi perguruan tinggi menggunakan metode K-Means clustering (Studi Kasus Di Program Studi Tkj Akademi Komunitas Solok Selatan)," *J. Teknol. dan Sist. Inf.*, vol. 02, no. 03, pp. 31–40, 2016, [Online]. Available: http://teknosi.fti.unand.id/index.php/teknosi/article/view/70.

[13] R. NOVIANTO, "penerapan data mining menggunakan algoritma K-Means clustering untuk menganalisa bisnis perusahaan asuransi," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 6, no. 1, pp. 85–95, 2019, doi: 10.35957/jatisi.v6i1.150.

[14] B. Aditya, *Topic Modelling pada Data Artikel Peneliti Penerima Dana PDUPT Menggunakan Gensim*. 2021.

[15] C. Algoritma, A. Carolina, K. Ade, and K. Kunci, "Penerapan data mining dengan menggunakan algoritma C4.5 pada klasifikasi fasilitas kesehatan provinsi di Indonesia," *J. Ilm. Komputasi*, vol. 19, no. 1, pp. 27–38, 2020, doi: 10.32409/jikstik.19.1.153.