ILKOM Jurnal Ilmiah Vol. 13, No. 3, December 2021, pp. 314-321
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Grouping the Spread of the Covid-19 Virus based on the Positive Number, Population and Area using the K-Means Clustering Method

Asep Muhidin[a,1,*]; Muhtajudin Danny[a,2], Elkin Rivali[a,3]
[a] *Universitas Pelita Bangsa, Jl. Inspeksi Kalimalang No.9, Cibatu Cikarang Selatan, Bekasi, 17550, Indonesia*
[1]*asep.muhidin@pelitabangsa.ac.id,* [2]*utat@pelitabangsa.ac.id,* [3]*elkin.rilvani@pelitabangsa.ac.id*
* *Corresponding author*

Article history: Received October 05, 2021; Revised October 23, 2021; Accepted November 14, 2021; Available online December 31, 2021

Abstract

The Covid-19 virus began to spread in Indonesia in early 2020. Until now, the virus is still spreading in several parts of Indonesia, although it has already started to decline in many areas. This virus is transmitted through direct contact with sputum droplets from an infected person (through coughing and sneezing). The possibility of direct human transmission depends on the number of confirmed positives, the population, and the location. This study looks for the characteristics of the city/district area group based on the population, area, and several residents who were confirmed positive for the Covid-19 virus using the K-Means Clustering method. The clustering process began with finding the best number of clusters (K) using the elbow method and testing cluster members using the silhouette and davies-Bouldin index methods. The best K value results obtained by the elbow method are 3 clusters (K=3). The results of K-Means Clustering with 3 clusters show that in cluster 1, a city or district with a small area with a large population, the number of positive COVID-19 residents is large. In cluster 2, a city or district with a small space with a moderate population, the number of positive people for COVID-19 is small. And in cluster 3, a city or district with a large area with a tiny population, the number of positive COVID-19 residents tends to be moderate.

Keywords: Clustering, Elbow, Davies-Bouldin Index, K-Means, Silhouette

Introduction

2020 is a challenging year for the world's countries; Indonesia is no exception. Starting from China at the city of Wuhan, the coronavirus, later known as Covid-19 (Coronavirus Disease), has become pandemic and spread in the city. Covid-19 (Coronavirus Disease) is a virus of the coronavirus group, namely SARS-CoV-2 or corona virus [1]. In early March 2020, Indonesia first announced that its population was confirmed positive for the Covid-19 virus[2]. The emergence of the Covid-19 virus in Indonesia certainly cannot be ignored. The Indonesian government has carried out various efforts to contain the Covid-19 to block the rate of spread, ranging from physical distancing to Large-Scale Social Restrictions (PSBB) in multiple areas mapped as the epicenter of the spread.

Local government efforts to contain the spread of the virus require strategic information about which regional groups are to be prioritized. In general, the transmission of the virus through direct contact with sputum droplets from an infected person. Areas with a high population density will cause faster transmission of infectious diseases with a more compact and complex distribution chain [3]. It is necessary to analyze whether there is a relationship between positive confirmations and the area and population.

Data mining analyzes large amounts of data to find valuable patterns and rules [4]. Data mining is grouped into estimation, prediction, classification, clustering, and association based on the task. Clustering analysis is a method of data mining used to classify objects so that the objects with the closest similarity to other objects are in the same cluster [5]. Meanwhile, data with different characteristics will be grouped into new clusters. The clustering algorithm process is to find data in the entire set and make it into subgroups, where the similarity of data in the cluster will be maximized. The similarity of data outside the cluster will be minimized [6]. One of the most well-known clustering methods is K-Means which has a simple and efficient algorithm. Therefore, in this study, the K-Means method was chosen for the process of grouping the data on the spread of COVID-19 in cities/districts throughout Indonesia based on the proximity of the positive confirmed number of data to the total population and area.

https://doi.org/10.33096/ilkom.v13i3.1011.314-321

Various studies on clustering the spread of the Covid-19 virus have been studied before. Journals and research that discuss the similarity of clustering algorithms and research datasets are used as references in this study. Research was conducted by Vasilios Zarikas, Stavros G. Poulopoulos, ZoeGareiou, and Efthimios Zervas (2020) with the title Clustering Analysis of Countries using the COVID-19 Cases Dataset [7]. This study grouped countries based on positive cases in a country using the Hierarchical Clustering algorithm. The study results were four groups (clusters) of countries based on the number of positive cases.

Research conducted by Rizkiana Prima R., Yashintia Arien E., and Sutikno with Corona Virus (COVID-19) Cluster Analysis in Indonesia on 2 March 2020–12 April 2020 used the K-Means Clustering Method[8]. This study resulted in 3 clusters, meaning that the optimum value of K is 3. The value of K is obtained from the Elbow and Sillhouette methods, both of which produce the optimum value of K=3.

Another study was conducted by Ahmad Husain Ardiansyah, Wisnu Nugroho, Nurul Hanifatul Alfiyah, Rahmat Aji Handoko, Muhammad Arfan Bakhtiar (2020) entitled Application of Data Mining Using the Clustering Method to Determine the Status of Provinces in Indonesia [9]. This study resulted in 3 clusters with distribution area status, namely Green, Yellow, and Red.

This study focuses on grouping cities/districts based on the proximity of data on the number of positive confirmations, population, and area using K-mean Clustering. The initial value of the number of clusters was determined by the elbow method, and the cluster member test was determined using the silhouette coefficient and davies-Bouldin index (dbi) methods. The clustering results were analyzed descriptively to obtain information on each cluster. This information can be used to determine the government's strategy to prevent the spread of the Covid-19 virus.

Method

This section explains the research processes carried out, including data collection, data preprocessing, finding the best number of clusters, clustering using K-Mean, and descriptive analysis of each cluster. All research processes use the python programming language, google collaborative software, and several libraries to run the K-Means algorithm and visual data.

A. Data Collection

The data used in this research was data extracted from a website (scraping) and manually collected. Data on names of cities/districts, area, and population in each city in Indonesia were obtained from scraping on the website https://id.wikipedia.org/wiki/Daftar_kabupaten_dan_kota_di _Indonesia, which was accessed on September 9, 2021. Meanwhile, the data on the number of positive Covid-19 19 in each city/district was collected manually by accessing the provincial government website with the address covid19/[namaprov].go. Id or the like on a date between 02 to 09 September 2021.

The data result was incorporated into a dataset in a file CSV (Comma Separated Values). Not all local websites provide detailed information on the distribution of positive confirmations for each city, so only part of the data has been collected. The number of data that has been collected is 214 cities in 14 provinces. City/district sample data and data on the province's name, area, population, and some confirmed positive COVID-19 are shown in table 2, and the metadata is explained in Table 1.

Table 1 Dataset structure Distribution of positive confirmations by city/district

Field	Type	Notes
provinsi	varchar	Nama provinsi dari kota/kabupaten
kabkota	varchar	Nama kabupaten atau kota
luas	float	Luas wilayah kota/kabupaten dengan satuan kilometer persegi
penduduk	double	Jumlah penduduk tiap kota/ kabupaten
covid	double	Jumlah penduduk yang terkonfirmasi positif covid-19

Table 2 Sample data for cities/districts, names of provinces, area, population and positive for covid

No	Province	City/district	Large	Population	Covid
1	Jawa Barat	Kab. Bandung	1767.96	3623790	33445
2	Jawa Barat	Kab.Bandung Barat	1305.77	1788336	18747
3	Jawa Barat	Kab. Bekasi	1224.88	3113071	50837
4	Jawa Barat	Kab. Bogor	2710.62	5427068	47160
5	Jawa Barat	Kab. Ciamis	1414.71	1229069	14762
...	...	...	...	...	...
...	...	...	...	...	...

210	Jambi	Kab. Tanjung Jabung Barat	4649.85	322058	2306
211	Jambi	Kab. Tanjung Jabung Timur	5445	222834	1362
212	Jambi	Kab. Tebo	6461	327669	2081
213	Jambi	Kota Jambi	103.54	609620	9349
214	Jambi	Kota Sungai Penuh	391.5	103511	1394

B. Data Preprocessing

Preprocessing is one of the essential steps to process the data according to the method used. Outlier analysis and data transformation were carried out in the preprocessing data research.

1. Outliers

The Outlier is defined as observational data that appears with extreme values, either univariate or multivariate [10]. What is meant by extreme values in observations are far or entirely different from most other matters in the group.

2. Data Transformation

Data transformation was used to change data that suit the data mining process. Normalization technique was performed to scale data values within a specific range of -1 to 1 or 0. One of the normalization methods is the Min-Max method, a simple technique for scaling values based on specified limits [11]. The Min-Max equation is shown in formula 1:

$$X_i = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The stages of the Min-Max method were carried out by finding the smallest and largest values in the city/district that have a positive confirmation distribution dataset. We further used the formula (1) to produce a new dataset with a range of values from 0 to 1.

C. Determination of the optimal number of clusters (K)

In the K-Mean method, the value of K (number of clusters) was initialized randomly so that the grouping of the data obtained could be different [12]. However, if the value obtained randomly for the initializer is not good, the group obtained is not optimal. Several methods can determine the optimal number of clusters to improve the clustering results.

1. Elbow Method

The Elbow method generates information in determining the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point [6]. To get a comparison is to calculate each cluster value's SSE (Sum of Square Error) [13]. With equation 2, the value of SSE on K-Means with bigger number of clusters K leads to the smaller value of SSE.

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in S_k} || X_i - C_k ||_2^2 \quad (2)$$

2. Silhouette Method

Silhouette Coefficient is used to see the quality and strength of the cluster, how well an object is placed in a cluster. This method is a combination of cohesion and separation methods. The silhouette coefficient value calculation results range between -1 to 1. If the silhouette coefficient is getting closer to 1, it means that object i is already in the right cluster [14].

3. Davies-Bouldin Index Method

Davies-Bouldin Index is one of the methods used to measure cluster validity in a grouping method; cohesion is defined as the sum of the proximity of the data to the cluster center point from the cluster being followed. At the same time, the separation is based on the distance between the center point of the cluster to the cluster [15]. The smaller the Davies-Bouldin Index (DBI) value obtained (non-negative >= 0), the better the cluster got.

4. K-Mean Clustering

The K-Means algorithm is used for iterative grouping; this algorithm partitions the data set into several K clusters that have been placed at the beginning. Partitioning the data set was done to determine the characteristics of each cluster so that clusters that have the same features were grouped into one cluster, and those with different characteristics were grouped into other clusters. The following was the flow of the K-Means algorithm:

1) Determine the number of K, where K is the number of clusters to be formed.

2) Set the center point of the cluster randomly; the center point of the cluster is often referred to as the centroid.
3) Equation three was used to calculate each distance of the existing data to each centroid.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (3)$$

With De Euclidean Distance, i is the number of objects, (x,y) are object coordinates, and (s,t) are centroid coordinates (the central point of the cluster).

4) Allocate each object into a cluster based on the minimum distance
5) Determine the new centroid using equation 4:

$$\overline{V}_{ij} = \frac{1}{N}\sum_{k=0}^{N_i} X_{kj} \quad (4)$$

$\overline{V}_{ij}$ is the centroid/average of the ke-i cluster for the ke-j variable, Ni is the amount of data that is a member of the ke-i cluster. Whereas, (i,k) is the clustered index, j is the index of the variable, and Xkj is the data value ke-k in the cluster for the ke-j variable.

6) Go back to steps 3, 4, and 5.

Results and Discussion

A. Outlier Data Result

Based on Figure 1, there is data on the number of positive Covid-19 confirmations that have a value far from the distribution of the data. This data interfered with the model's accuracy, so we deleted all data with a positive Covid value between 0 and 100.

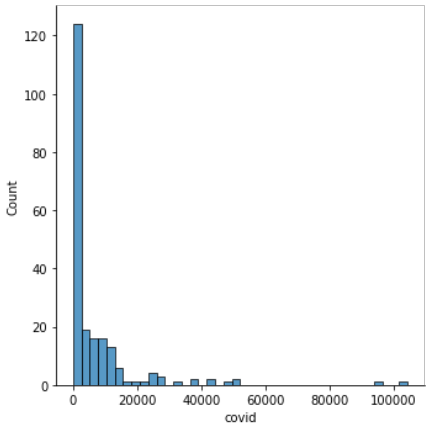

Figure 1 Distribution graph of confirmed covid-19 values

The outlier detection process caused the number of datasets to be reduced to 185 from 214 data lines.

i. Data Normalization

The data normalization process only applies to numeric and integer data. In the dataset shown in table 2 there is float data, namely the broad column. The way to do this was to convert km2 to m2; after that, we looked for the max and min values for each column for the normalization process with formula 1. The results of the conversion and normalization of data are shown in Tables 3 and 4.

Table 3 Dataset of conversion results

No	wide	population	Covid
1	1767960	3623790	33445
2	1305770	1788336	18747
3	1224880	3113071	50837
4	2710620	5427068	47160
5	1414710	1229069	14762
...	...	...	...
...	...	...	...

180	4649850	322058	2306
181	5445000	222834	1362
182	6461000	327669	2081
183	103540	609620	9349
184	391500	103511	1394

Tabel 4 Dataset of normalization results

No	wide	population	Covid
1	0.039566	0.666339	0.320249
2	0.029073	0.326724	0.179084
3	0.027236	0.571841	0.487289
4	0.060968	1.000000	0.451973
5	0.031546	0.223243	0.140810
...	...	...	...
...	...	...	...
180	0.104996	0.055419	0.021178
181	0.123048	0.037059	0.012111
182	0.146115	0.056457	0.019017
183	0.001778	0.108626	0.088821
184	0.008315	0.014981	0.012418

ii. Determination of the initial value of the cluster (K)

Determination of the value of K using the elbow method and each cluster member was tested using the silhouette coefficients and DBI methods. The following Listing program 1 shows the process.

Program Source 1 Searching the best K value

Finding the best K Value Program

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, davies_bouldin_score

score_inertia=[]
score_silhoutte=[]
score_dbs=[]

for k in range(2,6):
  kmeans=KMeans(n_clusters=k)
  members=kmeans.fit_predict(Xst)
  score_inertia.append(kmeans.inertia_)
  silhoutte_avg=silhouette_score(Xst,members)
  score_silhoutte.append(silhoutte_avg)

score_dbs.append(davies_bouldin_score(Xst,members))
```

Figure 2 shows the results for each K for the three methods.

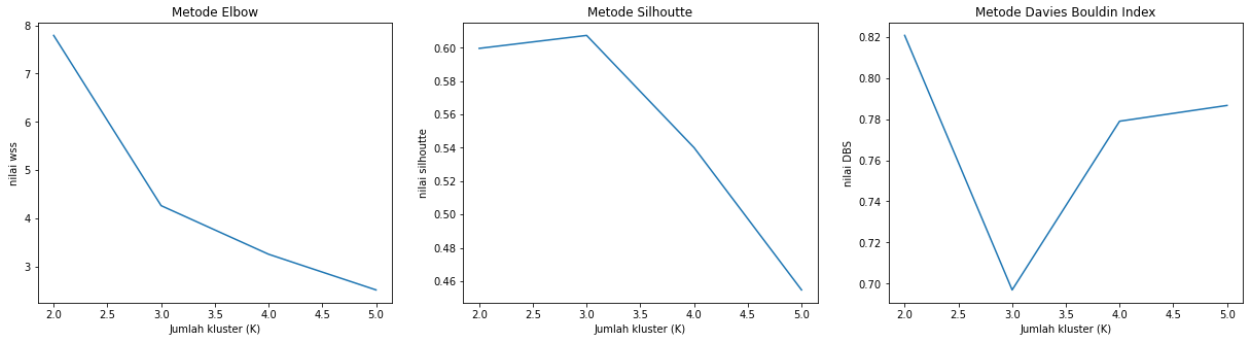

Figure 2 Graph of score elbow, silhouette dan DBI

Based on Figure 2, the best K value for the elbow method is K=3, because the value formed an elbow. In the silhouette coefficients method, the best value of K is K = 3 because that value is closest to 1. And in the DBI method, the best

value is K = 3 because it is the lowest value. From the results of the three methods, it can be concluded that the best value for K is 3 (K=3).

iii. K-Means Clustering Process

The K value obtained in the previous process was used as the initial K value in the K-Means Clustering process. Table 5 shows the results of the cluster center (centroid) for each attribute.

Table 5 Cluster center results

Cluster	wide	popoulation	Covid
1	0.029931	0.407115	0.361306
2	0.077055	0.061904	0.035663
3	0.514465	0.032713	0.055090

The results of grouping provincial data with K-means clustering can be seen in table 6

Table 6 City/district data for each cluster

K	City/district cluster	amount
1	Kabupaten Bandung, Kabupaten Bandung Barat, Kabupaten Bekasi, Kabupaten Bogor, Kabupaten Cianjur, Kabupaten Cirebon, Kabupaten Garut, Kabupaten Indramayu, Kabupaten Karawang, Kabupaten Sukabumi, Kota Bandung, Kota Bekasi, Kota Bogor, Kota Depok, Kota Balikpapan, Kabupaten Tangerang, Kota Tangerang, Kota Tangerang Selatan, Kota Pekanbaru, Kota Batam	20
2	Kabupaten Ciamis, Kabupaten Kuningan, Kabupaten Majalengka,Kabupaten Pangandaran, Kabupaten Purwakarta, Kabupaten Subang, Kabupaten Sumedang, Kabupaten Tasikmalaya, Kota Banjar,Kota Cimahi,Kota Cirebon, Kota Sukabumi,Kota Tasikmalaya, Kabupaten Aceh Barat, Kabupaten Aceh Barat Daya, Kabupaten Aceh Besar, Kabupaten Aceh Jaya, Kabupaten Aceh Selatan, Kabupaten Aceh Singkil, Kabupaten Aceh Tamiang,Kabupaten Aceh Tengah, Kabupaten Aceh Tenggara, Kabupaten Aceh Timur, Kabupaten Aceh Utara, Kabupaten Bener Meriah, Kabupaten Bireuen, Kabupaten Gayo Lues, Kabupaten Nagan Raya,Kabupaten Pidie, Kabupaten Pidie Jaya, Kabupaten Simeulue, Kota Banda Aceh, Kota Langsa, Kota Lhokseumawe,Kota Sabang, Kota Subulussalam,Kabupaten Bantaeng, Kabupaten Bone,Kabupaten Bulukumba, Kabupaten Enrekang, Kabupaten Gowa, Kabupaten Jeneponto, Kabupaten Luwu, Kabupaten Luwu Timur,Kabupaten Luwu Utara, Kabupaten Maros,Kabupaten Pangkajene dan Kepulauan, Kabupaten Pinrang, Kabupaten Sidenreng Rappang, Kabupaten Sinjai, Kabupaten Soppeng, Kabupaten Takalar, Kabupaten Tana Toraja,Kabupaten Toraja Utara, Kota Makassar, Kota Palopo, Kota Parepare, Kabupaten Balangan, Kabupaten Banjar, Kabupaten Barito Kuala, Kabupaten Hulu Sungai Selatan,Kabupaten Hulu Sungai Tengah, Kabupaten Hulu Sungai Utara, Kabupaten Kotabaru, Kabupaten Tabalong,Kabupaten Tanah Bumbu,Kabupaten Tanah Laut, Kabupaten Tapin, Kota Banjarbaru, Kota Banjarmasin, Kabupaten Barito Selatan,Kabupaten Barito Timur, Kabupaten Barito Utara,Kabupaten Gunung Mas, Kabupaten Kotawaringin Barat,Kabupaten Lamandau, Kabupaten Pulang Pisau,Kabupaten Sukamara, Kota Palangka Raya, Kabupaten Paser, Kabupaten Penajam Paser Utara,Kota Bontang, Kota Samarinda, Kabupaten Tana Tidung,Kota Tarakan, Kabupaten Biak Numfor, Kabupaten Jayapura, Kabupaten Jayawijaya, Kabupaten Keerom, Kabupaten Kepulauan Yapen, Kabupaten Lanny Jaya,Kabupaten Nabire, Kabupaten Paniai, Kabupaten Puncak, Kabupaten Puncak Jaya, Kabupaten Supiori,Kabupaten Tolikara,Kota Jayapura, Kabupaten Manokwari, Kabupaten Manokwari Selatan,Kabupaten Raja Ampat, Kabupaten Sorong, Kabupaten Sorong Selatan, Kabupaten Teluk Wondama, Kota Sorong, Kabupaten Agam, Kabupaten Padang Pariaman, Kabupaten Tanah Datar,Kota Bukittinggi,Kota Padang, Kota Pariaman, Kabupaten Lebak, Kabupaten Pandeglang, Kabupaten Serang, Kota Cilegon,Kota Serang,Kabupaten Bengkalis,Kabupaten Indragiri Hilir, Kabupaten Indragiri Hulu, Kabupaten Kampar, Kabupaten Kepulauan Meranti,Kabupaten Kuantan Singingi, Kabupaten Pelalawan, Kabupaten Rokan Hilir, Kabupaten Rokan Hulu,Kabupaten Siak, Kota Dumai, Kabupaten Bintan,Kabupaten Karimun, Kabupaten Kepulauan Anambas, Kabupaten Lingga, Kabupaten Natuna,Kota Tanjung Pinang,Kabupaten Batanghari, Kabupaten Bungo,Kabupaten Kerinci, Kabupaten Merangin, Kabupaten Muaro Jambi, Kabupaten Sarolangun, Kabupaten Tanjung Jabung Barat, Kabupaten Tanjung Jabung Timur,Kabupaten Tebo, Kota Jambi,Kota Sungai Penuh	144
3	Kabupaten Kapuas, Kabupaten Katingan, Kabupaten Kotawaringin Timur, Kabupaten Murung Raya, Kabupaten Seruyan, Kabupaten Berau, Kabupaten Kutai Barat, Kabupaten Kutai Kartanegara, Kabupaten Kutai Timur, Kabupaten Mahakam Ulu, Kabupaten Bulungan, Kabupaten Malinau, Kabupaten Nunukan, Kabupaten Asmat, Kabupaten Boven Digoel, Kabupaten Mappi, Kabupaten Merauke, Kabupaten Mimika, Kabupaten Fakfak, Kabupaten Kaimana, Kabupaten Teluk Bintuni	21

While the number of each city/district in the province for each cluster is shown in Figure 3

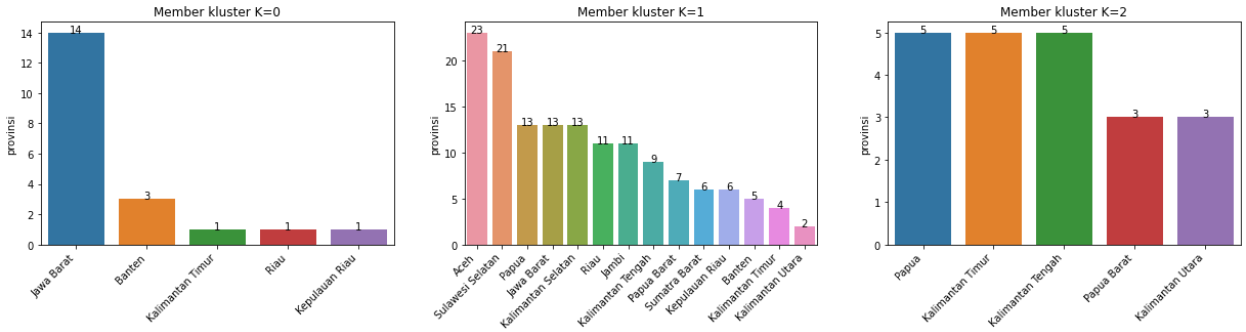

Figure 3, Number of cities/districts each cluster for provincial data

Cluster Descriptive Analysis

From the data center of the cluster (centroid), each cluster can be described by comparing the centroid value (table 5) with the average value of each attribute.

Cluster 1 Analysis

Table 7 shows the comparison between the average value of each attribute and the centroid value of cluster 1.

Table 7 The results of the comparison of the centroid value with the mean

	Mean	*Centroid*	*Description*
wide	0.1216	0.029931	di bawah rata-rata
population	0.0959	0.407115	di atas rata-rata
covid	0.0730	0.361306	di atas rata-rata

From the data in Table 7, the description for cluster 1 is a city or district with a small area with a large population, so the number of positive COVID-19 residents is large.

Cluster 2 Analysis

Table 8 shows the comparison between the average value of each attribute and the centroid value of cluster 2.

Table 8 The results of the comparison of the centroid value with the mean

	Mean	*Centroid*	*Description*
wide	0.1216	0.077055	di bawah rata-rata
population	0.0959	0.061904	di bawah rata-rata
covid	0.0730	0.035663	di bawah rata-rata

From the data in Table 8, the description for cluster 2 from a city or district with a small area, with a medium population shows that the number of people who are positive for COVID-19 is small.

Cluster 3 analysis

Table 9 shows the comparison between the average value of each attribute and the centroid value of cluster 3.

Table 9 The results of the comparison of the centroid value with the mean

	Mean	*Centroid*	*Description*
wide	0.1216	0.514465	di atas rata-rata
population	0.0959	0.032713	di bawah rata-rata
covid	0.0730	0.055090	di bawah rata-rata

From table 9 data, the description for cluster 3 from a city or district with a large area, with a tiny population shows that the number of positive COVID-19 residents tends to be moderate.

Conclusion

This study concludes that the best K value calculated by the elbow, silhouette coefficients, and Davies-Bouldin index methods is 3 clusters. The estimated K value was used for the grouping process using K-Means Clustering.

Descriptive clustering results show that broad and narrow areas tend to be moderate with a small population and the number of positive COVID-19 patients is small. Meanwhile, a small area and a large population indicate many positive COVID-19 patients. The city/district data grouping results can be used for priority handling by local governments.

Suggestions for further research is that it is necessary to analyze the distribution of data on data on areas and populations. And it is required to add an optimization method to the K-Mean Clustering method in determining the cluster center so that the grouping results can be even better.

Acknowledgement

The researcher would like to thank the Directorate of Research and Community Service, Ministry of Research, Technology, and Higher Education for the support given to researchers in the form of research funds to support this research.

References

[1] Lipsitch, M., Swerdlow, D. L., & Finelli, L. (2020). "*Defining the epidemiology of Covid-19—studies needed*". New England journal of medicine, 382(13), 1194-1196.

[2] AchmadYurianto, Direktorat Jenderal Pencegahan dan Pengendalian Penyakit https://www.kemkes.go.id/article/view/20030900003/pasien-positif-covid-19-menjadi-19-orang-kemenkes-fokus-pada-tracing-dan-tracking-contact.html

[3] Desy Noor Permata Sari, YL. Sukestiyarno, "Analisis *Cluster* dengan Metode *K-Means* pada Persebaran Kasus COVID-19 Berdasarkan Provinsi di Indonesia", PRISMA 4 (2021): 602-610, PRISMA, Prosiding Seminar Nasional Matematika

[4] Michael J.A. Berry and Gordon S. Linoff, "*Data Mining Techniques"*, 2nd edition, wiley, 2004

[5] Ediyanto, Novitasari Mara, M., & Satyahadewi, N. (2013). "Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis". Buletin Ilmiah Mat. Stat. Dan Terapannya (Bimaster), 02(2), 133–136.

[6] Taslim, T., & Fajrizal, F. (2016). "Penerapan algorithma K-Mean untuk *clustering* data obat pada puskesmas rumbai". Jurnal Teknologi Informasi & Komunikasi Digital Zone, 7(2), 108-114.

[7] Vasilios Zarikas, Stavros G.Poulopoulos, ZoeGareiou dan Efthimios Zervas(2020), "*Clustering Analysis of Countries using the COVID-19 Cases Dataset*", Elsevier

[8] Rizkiana Prima R., Yashintia Arien E., dan Sutikno (2020), "Analisis Cluster Virus Corona (COVID-19) di Indonesia pada 2 Maret 2020 – 12 April 2020 dengan Metode K-Means Clustering ", Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember (ITS)

[9] Ahmad Husain Ardiansyah, Wisnu Nugroho, Nurul Hanifatul Alfiyah, Rahmat Aji Handoko,Muhammad Arfan Bakhtia, (2020), "Penerapan Data Mining Menggunakan Metode Clustering untuk Menentukan Status Provinsi di Indonesia" , Seminar Nasional Inovasi Teknolog, Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Magelang

[10] Lenny Budiarti, Tarno, Budi Warsito, "Analisis Intervensi Dan Deteksi *Outlier* Pada Data Wisatawan Domestik", Jurnal Gaussian, Volume 2, Nomor 1, Tahun 2013, Halaman 39-48

[11] Chusyairi A. and Saputra P. R. N., 2019, "Pengelompokan Data Puskesmas Banyuwangi Dalam Pemberian Imunisasi Menggunakan Metode *K-Means Clustering*", Telematika, 12(2), pp. 139–148, doi: 10.35671/telematika.v12i2.848.

[12] Achmad Solichin, Khansa Khairunnisa, Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta Menggunakan Metode K-Means, Fountain of Informatics Journal Volume 5, No. 2, November 2020.

[13] Ni Putu Eka Merliana, Ernawati, Alb. Joko Santoso, "Analisa Penentuan Jumlah Cluster Terbaik Pada Metode *K-Means Clustering*", Prosiding Seminar Nasional Multi Disiplin Ilmu, ISBN: 978-979-3649-81-8

[14] Putra A. K. P., dkk. 2015. "Analisis Sistem Deteksi Anomali Trafik Menggunakan Algoritma CURE (*Clustering Using Representatives*) dengan Koefisien *Silhouette* dalam Validasi *Clustering*",. e-Proceeding of Engineering Vol.2, No.2, Page 3837.

[15] Bernad Jumadi D. S., "Peningkatan Hasil Evaluasi *Clustering Davies-Bouldin Index* Dengan Penentuan Titik Pusat *Cluster* Awal Algoritma K-Means", Tesis Program Studi S2 Teknik Informatika Fakultas Ilmu Komputer Dan Teknologi Informasi Universitas Sumatera Utara Medan 2018

[16] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). *Constrained k-means clustering with background knowledge*. In Icml (Vol. 1, pp. 577-584).