Exploring the Use of K-Means Clustering Algorithm in Analyzing Price Trends of Commodities in Eastern Indonesia

Marchel Yusuf Rumlawang Arpipi[1]
[3] School of Informatics Engineering, Faculty of Indormation Technology, Tarumanagara University, Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail: [1]marchel.535210039@stu.untar.ac.id

Abstract

The application of Machine Learning has provided many benefits in decision-making and helps perform analyses in various industrial fields. One of the most frequently used algorithms is K-Means for clustering. This research uses K-Means to analyse commodity price trends in Eastern Indonesia. The objective is to see how K-Means works to group each commodity in a region into several clusters and calculate the average silhouette score generated and then measure the extent to which each sample in a given cluster matches its set compared to the nearest neighbouring collection. It is hoped that through this research, the evaluation of food or commodity prices in Indonesia, especially Eastern Indonesia, can be done better market segmentation, where each group with similar behaviour and characteristics can be treated differently in marketing and pricing strategies.

Keywords—*Food Prices, K-Means, Clustering, Commodities*

1. INTRODUCTION

One of the basic needs of humans is food. With energy, humans can move and do other things according to the energy capacity obtained from the food eaten in one day. Meanwhile, a commodity is a raw material from crops that can be traded at the same value. The price of food available in an area in Eastern Indonesia is still unstable. The threat of natural disasters such as floods, landslides, and other disasters often thwarts harvests so that supplies become perishable and disrupt the distribution of food commodities, especially since production is centred in Eastern Indonesia, causing distribution to take a long time to reach the destination area. This causes food prices to fall into the high-price category. However, the advantages of Indonesia as a large country with a tropical climate, a population of more than 270 million, and located on the equator are the reasons why crops can thrive.

The agricultural sector is very important in the national economy, as seen from the number of people who live and work in this sector [1]. Indonesia has a vast agricultural land, so food security is important to be improved for further economic growth. That way, the availability of healthy and nutritious food can be fulfilled by the state for its people. There is also the challenge of how to cope with rising food prices due to large demand on festive days or risks such as land shortages and declining agricultural productivity. The Covid-19 pandemic that occurred for approximately 2 years has made people who work as farmers experience problems because the harvest is not maximised. Sometimes, people do not realise significant changes in food prices. This may be due to the government's lack of announcing or informing changes in food prices. However, by further studying the characteristics of food commodity prices in each region, the central and local governments can collaborate in making appropriate policies.

The National Strategic Food Price Information Centre (PIHPS-National) is a government agency that provides past and current data. According to PIHPS, staple food items in Indonesia

consist of rice, chicken meat, beef, chicken eggs, shallots, garlic, red chillies, cayenne pepper, cooking oil, and sugar. These food items are the main problem for traditional market communities in predicting the future prices of these food items due to the fluctuations that occur..

Regarding this issue, this study aims to explore the application of the K-Means algorithm in analysing clusters of commodity price trends of a region in Eastern Indonesia. The K-Means algorithm is used for parameter initialisation as it is simple and works well for large datasets when compared to hierarchical clustering [2]. The K-Means algorithm works by randomly assigning a cluster value (k), where the value becomes the centre of the cluster, also referred to as the centroid. The disadvantage of the k-means algorithm is when determining the initial cluster because it depends on the initial data given. given [3]. K-Means is one of the algorithms in data mining techniques that can cluster heterogeneous data because clustering algorithms are only able to recognise homogeneous attribute values. The process of the K-Means algorithm is different from other data mining algorithms such as the Apriori Algorithm which searches for frequent itemsets that often appear with heterogeneous data models and then prunes and counts according to the number of K itemsets. Clustering groups several n objects into k classes based on the calculation of their distance to the cluster centre. [4]. The object cluster is seen from the distance of the object to the closest centre point. After knowing the closest centre point, the object will be classified as a member or not of that category. [5].

Clustering is part of unsupervised learning from the machine learning discipline. Clustering is the process of dividing data into classes or clusters based on their level of similarity. [6]. Several previous research have been conducted, for example research using the K-Means algorithm to map traditional markets based on food commodity prices [1] and clustering fruit export data based on destination countries [7].

2. RESEARCH METHODS

The dataset collected in this study was sourced from the PIHPS website (https://www.bi.go.id/hargapangan/TabelHarga/PasarTradisionalDaerah).

2.1 Data Pre-processing

The dataset that has been downloaded, was put together into an Excel file with a total of 1278 samples and added 1 feature, namely "Date" starting from 2020-01-01 to 2023-06-30 so that there are 11 features in the dataset. Each province is set as *sheet_name*. Missing value handling is done because there are quite a lot of cells in the dataset that have no value. The handling uses the *ffill()* function to fill the missing value with the last value found before, known as forward fill, and the *bfill()* function to fill the missing value with the next value found, known as backward fill.

Table 1 Food Commodities Dataset (sheet_name='Bali')

Date	Rice	Chicken Meat	Beef	Chicken Eggs	Shallots	Garlic	Red Chillies	Cayenne Pepper	Cooking Oil	Granulated Sugar
01/01/2020 00:00										
01/02/2020 00:00	11.350	36.250	108.750	23.900	34.650	26.400	29.500	33.650	13.500	13.600
01/03/2020 00:00	11.350	36.250	108.750	23.500	34.650	27.000	30.500	32.600	13.500	13.600
01/04/2020 00:00										
01/05/2020 00:00										
01/06/2020 00:00	11.350	36.250	108.750	23.500	34.150	27.750	31.750	36.250	13.500	13.700
01/07/2020 00:00	11.350	36.000	108.750	23.500	33.250	30.150	35.450	37.950	13.650	13.750
01/08/2020 00:00	11.350	36.000	108.750	23.500	33.250	30.650	35.350	38.150	13.650	13.800
01/09/2020 00:00	11.500	36.000	108.750	23.500	35.250	30.650	38.400	39.400	13.650	13.800
01/10/2020 00:00	11.500	36.000	108.750	23.500	35.250	31.500	38.500	38.400	13.800	13.850

2.2 Data Processing

After that, rechecking is done to see the results of handling missing values. However, the "Date" feature will be dropped to see the correlation between other features, namely Rice, Chicken

Meat, Beef, Chicken Eggs, Shallots, Garlic, Red Chillies, Cayenne Pepper, Cooking Oil, and Granulated Sugar and then transposed so that clustering can be done.

Table 2 Correlation table (sheet_name='Bali')

	Rice	Chicken Meat	Beef	Chicken Eggs	Shallots	Garlic	Red Chillies	Cayenne Pepper	Cooking Oil	Granulated Sugar
Rice	1	0.535392	0.556169	0.529433	0.044769	0.084742	0.279846	0.020478	0.297801	0.00855
Chicken Meat	0.535392	1	0.783095	0.740624	0.024208	-0.098476	0.443287	0.416433	0.735797	-0.070005
Beef	0.556169	0.783095	1	0.804268	0.021021	-0.12801	0.400997	0.451264	0.873997	-0.166296
Chicken Eggs	0.529433	0.740624	0.804268	1	0.203435	-0.052158	0.326175	0.306905	0.659793	0.060064
Shallots	0.044769	0.024208	0.021021	0.203435	1	0.409911	0.068549	-0.131984	0.139119	0.585251
Garlic	0.084742	-0.098476	-0.12801	-0.052158	0.409911	1	-0.197592	-0.27199	-0.056636	0.383203
Red Chillies	0.279846	0.443287	0.400997	0.326175	0.068549	-0.197592	1	0.592875	0.459628	-0.102969
Cayenne Pepper	0.020478	0.416433	0.451264	0.306905	-0.131984	-0.27199	0.592875	1	0.45821	-0.446608
Cooking Oil	0.297801	0.735797	0.873997	0.659793	0.139119	-0.056636	0.459628	0.45821	1	-0.033592
Granulated Sugar	0.00855	-0.070005	-0.166296	0.060064	0.585251	0.383203	-0.102969	-0.446608	-0.033592	1

2.2 Bibliographic References

a. K-Means Algorithm

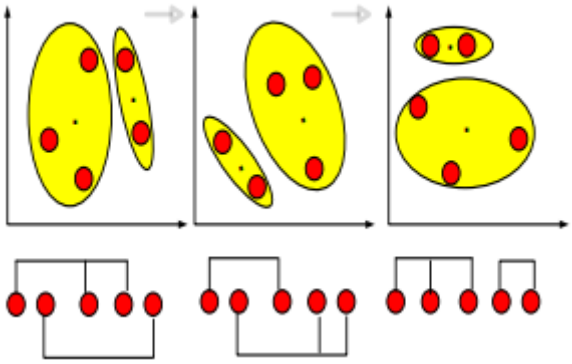

Figure 1 K-Means Method Steps (Source: Implementation Of Data Mining For Potential Customer Selection Using K-Means Algorithm [8])

The K-Means Clustering algorithm is a simple and effective algorithm for finding clusters in data with the following algorithm:

1. Define k as the number of clusters to be formed. Set the cluster centre.
2. Calculate the distance of each data to the cluster centre using the Euclidean equation.

$$D(ik) = \sqrt{\sum_j^m (Cij - Ckj)^2} \quad (1)$$

3. Group the data into the cluster with the shortest distance using the equation.

$$\text{Min}\sum_{k=1}^{k} dik = \sqrt{\sum_j^m (Cij - Ckj)^2} \quad (2)$$

4. Calculate the new cluster centre using Eq.

$$C_{kj} = \frac{\sum_{i=1}^{p} Xij}{p} \quad (3)$$

Where::

X_{ij} ϵ cluster k-th [9]

b. K-Means Algorithm

Clustering on data is a stage to classify the set of data whose class attributes have not been described. the concept of clustering is to maximise and minimise the intra similarity between classes. for example, there is a set of objects, the first process can be clustered into several sets of classes and then into a regular set so that it can be derived based on a particular classification

group. Data is clustered based on similarity. Determining the similarity of two objects in clustering is done by calculating the distance between objects. A widely used clustering algorithm is K-Means. The goal of data clustering is to minimise the objective function defined in the clustering process, and generally always minimises the variation of a cluster and maximises the variation between clusters [4][10].

Figure 2 Research Flowchart

3. RESULTS AND DISCUSSION

Masih dengan *sheet_name*, *k*lasterisasi dilakukan dengan memberikan nilai untuk *n_clusters=5* dan *random_state=0* lalu *cluster_labels* di isi dengan data yang telah ditranspose sebelumnya. Panjang *cluster_labels* yang diketahui adalah 10. Kemudian dibuat DataFrame baru dengan 2 kolom (Commodities and Cluster) dengan indeks mengikuti panjang dari *cluster_labels*.

Table 3 Commodities and Cluster tables

Commodities	**Cluster**	
0	Beras	0
1	Daging Ayam	4
2	Daging Sapi	1
3	Telur Ayam	3
4	Bawang Merah	3
5	Bawang Putih	3
6	Cabai Merah	2
7	Cabai Rawit	2
8	Minyak Goreng	0
9	Gula Pasir	0

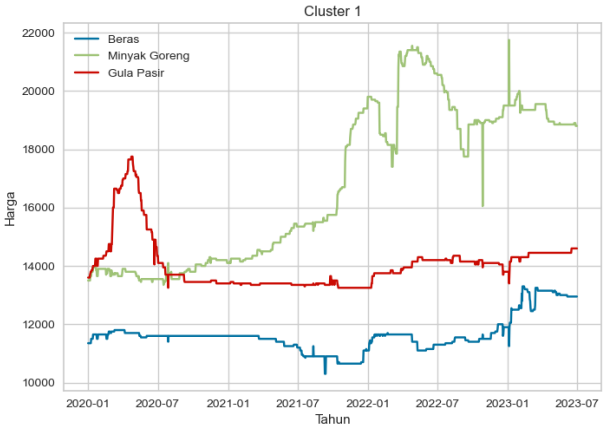

Figure 3 Price Trends for Cluster 1

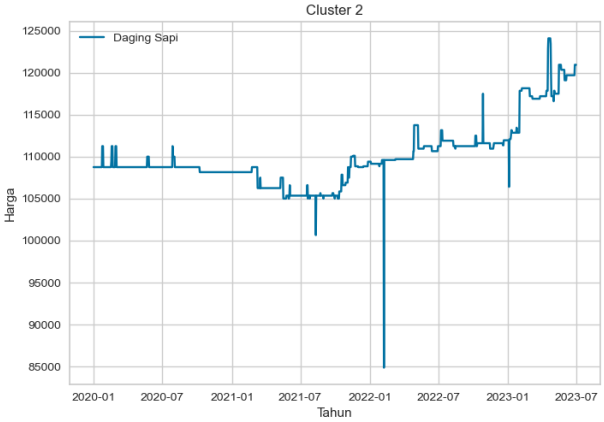

Figure 4 Price Trends for Cluster 2

Table 4 Describe tables for Cluster 1

	Beras	Minyak Goreng	Gula Pasir
count	1277	1277	1277
mean	11635.08222	16768.91151	14046.35865
std	605.236628	2671.698118	884.406699
min	10300	13350	13250
25%	11350	14100	13400
50%	11600	15750	13850
75%	11650	19250	14300
max	13300	21750	17750

Table 5 Describe tables for Cluster 2

	Daging Sapi
count	1277
mean	110010.3367
std	3824.007806
min	84850
25%	108150
50%	108750
75%	111250
max	124100

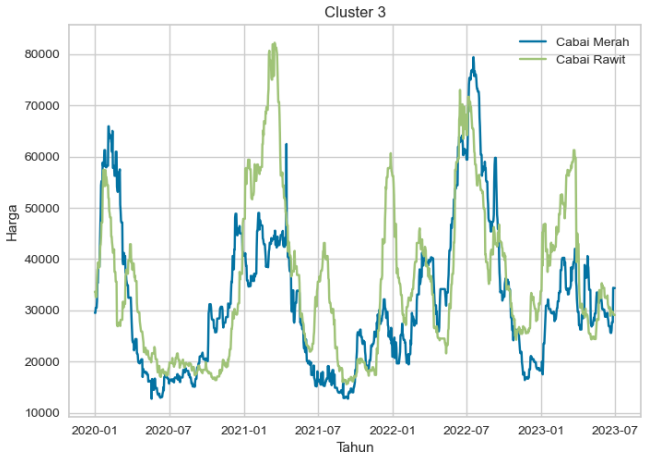

Figure 5 Price Trends for Cluster 3

Table 6 Describe tables for Cluster 3

	Cabai Merah	**Cabai Rawit**
count	1277	1277
mean	31784.37745	35604.65936
std	14258.22319	15353.33967
min	12750	15600
25%	19900	22600
50%	29250	32400
75%	39500	43250
max	79400	82200

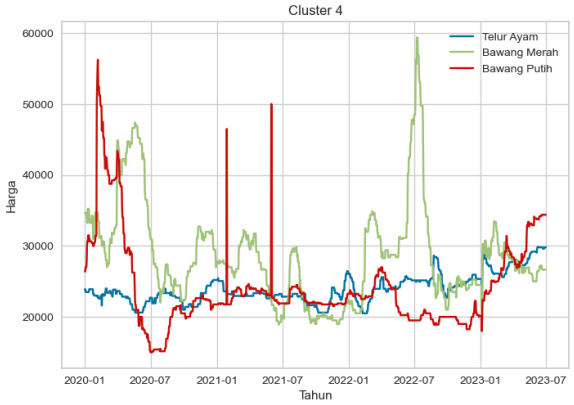

Figure 6 Price Trends for Cluster 4

Table 8 Price Trends for Cluster 4

	Telur Ayam	Bawang Merah	Bawang Putih
count	1277	1277	1277
mean	23980.93187	28622.12216	24137.86218
std	2091.890445	7076.112879	6237.755207
min	20500	18900	15000
25%	22700	24250	20500
50%	23400	27500	22400
75%	25150	31250	25000
max	29850	59400	56250

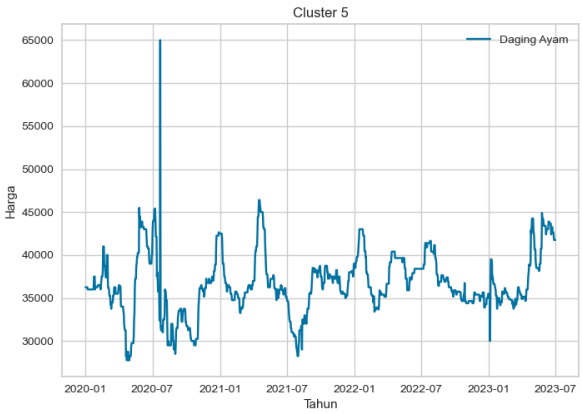

Figure 7 Price Trends for Cluster 5

Table 9 Describe tables for Cluster 5

	Daging Ayam
count	1277
mean	36663.46907
std	3595.339779
min	27750
25%	34900
50%	36250
75%	38400
max	65000

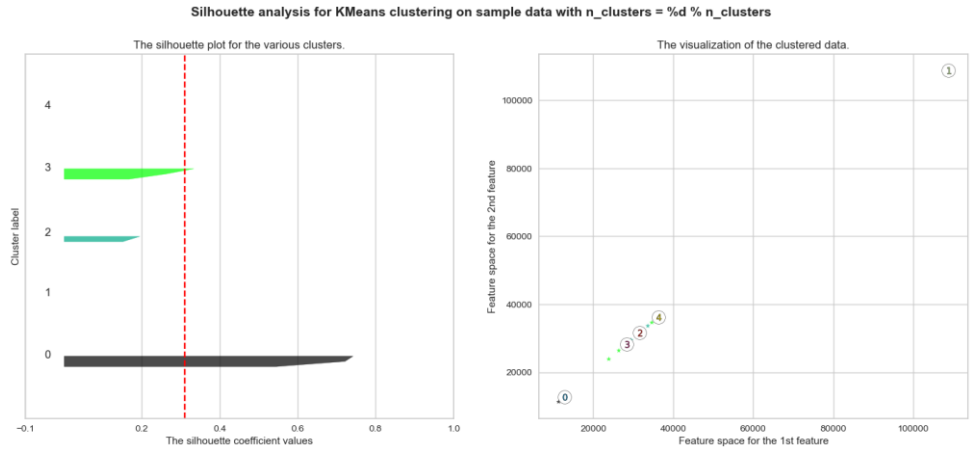

Figure 8 Silhouette analysis for K-Means Clustering on sample data with n_cluster=5

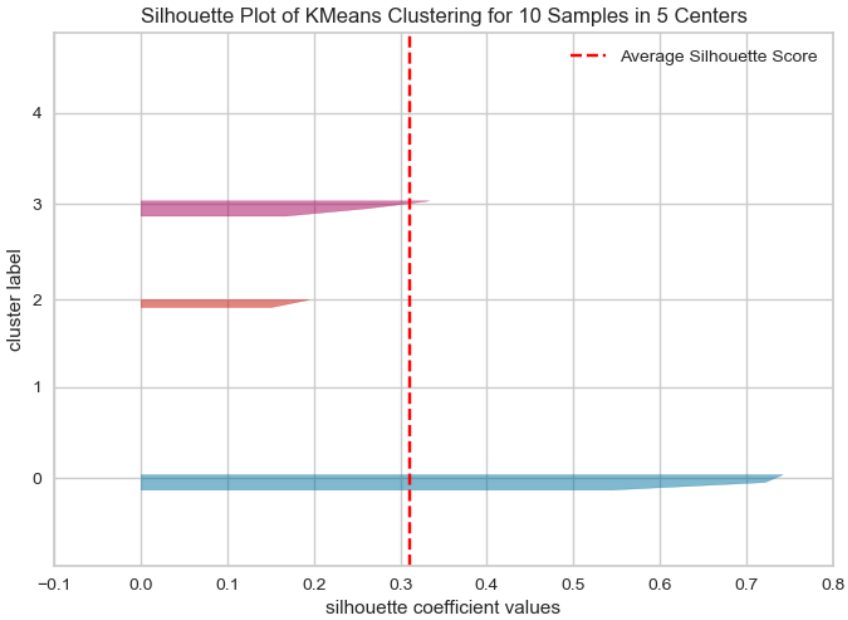

Figure 9 Silhouette coefficient Values

For n_clusters = 5, the average silhouette_score is : 31.1 %. Some comparisons were also made between features with strong positive correlation scores and looking at annual and monthly averages of all commodities.

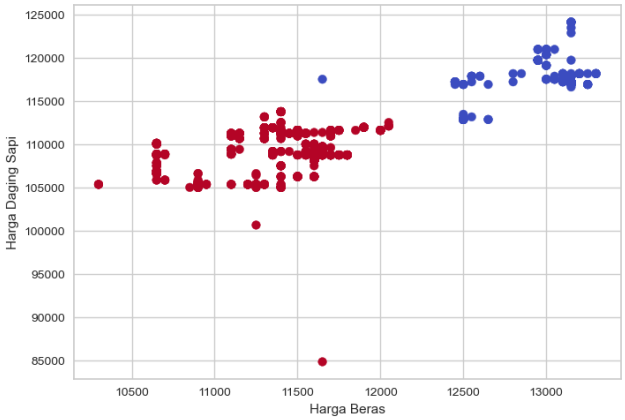

Figure 10 Rice and Beef Price Comparison

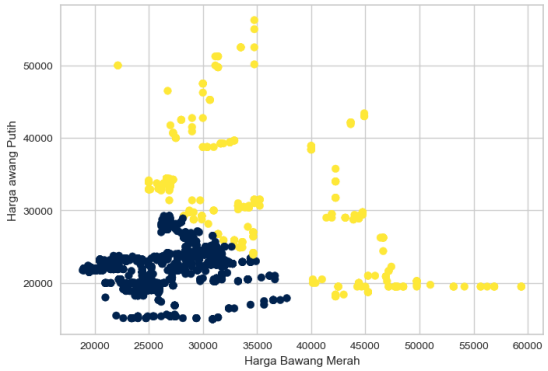

Figure 10 Shallots and Garlic Price Comparison

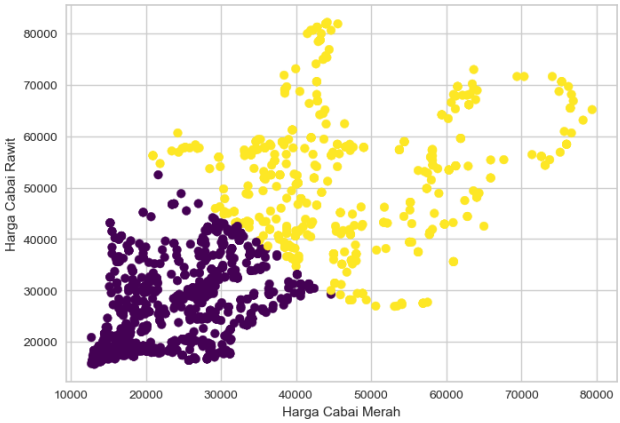

Figure 11 Red Chillies and Cayanne Pepper Price Comparison

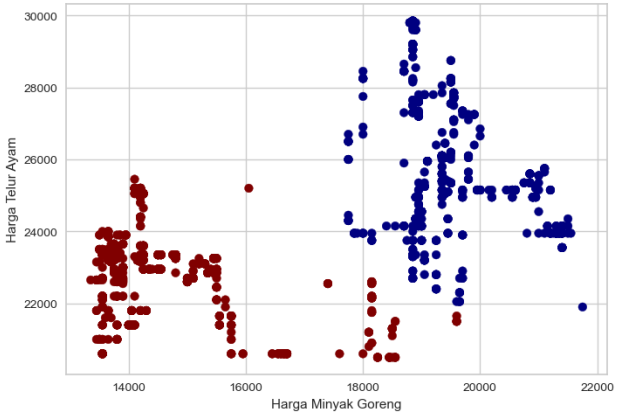

Figure 12 Cooking Oil and Granulated Sugar Price Comparison

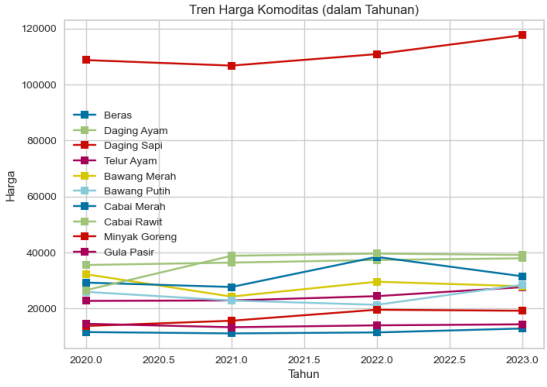

Figure 13 Annual report from 2020 to 2023

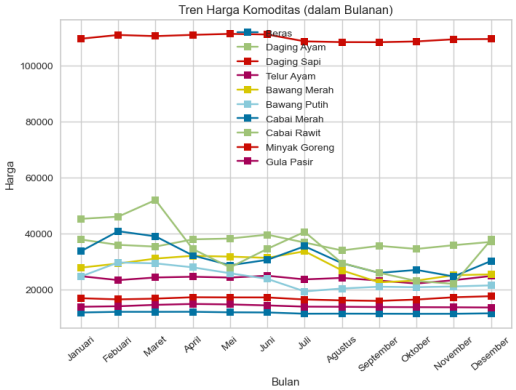

Figure 14 Monthly report

4. CONCLUSION

After the analysis of the food commodities is completed, it can be concluded that several groups of mapped data can be seen through the similarity of the behaviour of each data sample.

The interrelationship between features measured by the correlation method is not enough to show a strong positive relationship, but the comparison as shown in the figure provides an understanding that there is a behaviour that is not much different among the food commodities measured.

One of the relationships that can be seen is a weak correlation relationship, which is indicated by coefficient values that tend to be close to zero. This value indicates that the relationship between some commodities, such as rice and shallots, garlic, sugar, eggs, shallots, and garlic, is so weak that a change in the price of one commodity has little impact on the other. The correlation coefficient between commodities ranges between 0.5 and 0.6 indicating a positive correlation relationship. Price changes in a commodity will have a significant impact on other commodities in a straight line. There is a fairly strong positive relationship between rice and beef, red chilli and cayenne pepper, cooking oil and sugar and shallots and garlic. Therefore, there is a high probability that if the price of one commodity rises, the prices of other commodities will also rise.

THANK YOU NOTE

The author's would like to thank God Almighty and his family who have helped in providing support, encouraging, and continuing to motivate in conducting this research for the sake of increasingly sharpened insights.

BIBLIOGRAPHY

[1] A. Novita and H. B. Seta, *Pemetaan Pasar Tradisional Berdasarkan Harga Pangan Komoditas Menggunakan Algoritma K-Means*. 2021.

[2] J. Hutagalung, Y. Hendro Syahputra, Z. Pertiwi Tanjung, S. Triguna Dharma, and J. I. Pintu Air, 'Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering', *Hal AH Nasution*, vol. 9, no. 1, 2022, [Online]. Available: http://jurnal.mdp.ac.id

[3] J. Nasir, 'PENERAPAN DATA MINING CLUSTERING DALAM MENGELOMPOKAN BUKU DENGAN METODE K-MEANS', *Jurnal SIMETRIS*, vol. 11, no. 2, 2020.

[4] R. Muliono and Z. Sembiring, 'DATA MINING CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS UNTUK KLASTERISASI TINGKAT TRIDARMA PENGAJARAN DOSEN', 2019.

[5] A. Praja, C. Lubis, and D. E. Herdiwindiati, 'DETEKSI PENYAKIT DIABETES DENGAN METODE FUZZY C-MEANS CLUSTERING DAN K-MEANS CLUSTERING', 2017.

[6] F. Yunita, 'PENERAPAN DATA MINING MENGGUNKAN ALGORITMA K-MEANS CLUSTRING PADA PENERIMAAN MAHASISWA BARU (STUDI KASUS : UNIVERSITAS ISLAM INDRAGIRI)', 2018.

[7] H. Atma Negara *et al.*, 'Clustering Data Ekspor Buah-Buahan Berdasarkan Negera Tujuan Menggunakan Algoritma K-Means', *Bina Insani ICT Journal*, vol. 8, no. 1, pp. 73–82, 2021, [Online]. Available: https://www.bps.go.id.

[8] R. R. Putra and C. Wadisman, 'Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means', *INTECOMS: Journal of Information Technology and Computer Science*, vol. 1, no. 1, pp. 72–77, Mar. 2018, doi: 10.31539/intecoms.v1i1.141.

[9] L. Suriani, 'Pengelompokan Data Kriminal Pada Poldasu Menentukan Pola Daerah Rawan Tindak Kriminal Menggunakan Data Mining Algoritma K-Means Clustering', *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 1, no. 2, p. 151, Jan. 2020, doi: 10.30865/json.v1i2.1955.

[10] 'LaporanPenelitianPengelompokanGenKanker2015'.