JUSTEK : JURNAL SAINS DAN TEKNOLOGI
http://journal.ummat.ac.id/index.php/justek
ISSN 2620-5475
Vol. 8, No. 4, December 2025, pp. 379-387

Clustering of Data on Vegetable Crop Production in the City of Bandung Using the K-Means Algorithm

Ifano Rangga Saputra[1], Cahyono Budy Santoso[2]
[1,2]Information Systems, Pembangunan Jaya University, Indonesia
Email: [1]ifano.ranggasputra@student.upj.ac.id, [2]cahyono.budy@upj.ac.id

ARTICLE INFO

Article History:

Received	: 11-10-2025
Revised	: 07-12-2025
Accepted	: 08-12-2025
Online	: 15-12-2025

Keywords:
K-Means;
Clustering;
Vegetable Production;
Business Intelligence;
Data Mining;
Agricultural Analytics

ABSTRACT

Vegetable farming supports urban food security, and Bandung City is one of West Java's main horticultural centers. However, vegetable production remains unevenly distributed across its sub-districts. This study analyzes production patterns from 2018–2023 using the K-Means Clustering algorithm. The dataset includes 12 major commodities, and the analysis involves data preprocessing, determining the optimal number of clusters using the Elbow Method and Silhouette Score, applying K-Means, and visualizing results through heatmaps and PCA. The findings reveal three clusters: Cluster 0 dominated by potatoes and the "others" category; Cluster 1 dominated by kale; and Cluster 2 dominated by shallots and petsai. These patterns indicate concentrated and specialized production across specific sub-districts. The study concludes that K-Means effectively identifies multi-commodity production similarities and provides strategic insight for Business Intelligence applications in agricultural planning and policy development.

Crossref
https://doi.org/10.31764/justek.vXiY.ZZZ

A. BACKGROUND

Vegetable farming has an important contribution to food security and economic stability of the community (Yusdja & Sayaka, 2017). The city of Bandung is one of the horticultural centers in West Java that produces various vegetable commodities, but the distribution of production shows inequality between sub-districts. This imbalance causes distribution efficiency to decrease, regional planning becomes less optimal, and production potential is not utilized to the maximum. The urgency of this research lies in the need to provide data-driven analysis that is able to comprehensively describe vegetable production patterns so that governments and stakeholders can make more informed decisions (Mulyana et al., 2025).

Previous research has shown that K-Means is widely used in agricultural analysis, such as rice productivity mapping (Farismana, 2024), horticultural clustering analysis (Lianita et al., 2024), and grouping food security indicators (Prastanika & Wijayanto, 2023). This method has also proven to be effective in Business Intelligence-based analysis (Akbar & Octaviany, 2021; Riyanda & Suyanto, 2020). However, the state of the art shows that research on the clustering of multicommodity vegetable production in urban areas, especially the city of Bandung, is still very limited, because most studies focus on one commodity or on rural areas.

The research gap arises from the absence of a comprehensive study that integrates multi-commodity analysis of vegetables, clustering using K-Means, and Business Intelligence visualization to map production between sub-districts. Previous research has not provided a sufficiently in-depth spatial picture of the variation in vegetable production in Bandung. Therefore, the novelty of this research lies in the application of K-Means Clustering to 12 vegetable commodities as well as the BI approach designed to support urban agriculture policies (Dirayati et al., 2025).

Based on these gaps, the purpose of this study is to group sub-districts in the city of Bandung based on the production profile of 12 vegetable commodities using the K-Means Clustering algorithm to produce production pattern information that can be used in distribution planning, regional development, and data-driven decision-making.

B. RESEARCH METHODS

This study uses the K-Means Clustering method based on data mining using Python to analyze the production patterns of 12 vegetable commodities in the city of Bandung for the 2018–2023 period. This method was chosen because it is effective in grouping large-scale numerical data and is widely used in modern agricultural research (Sofyan & Sitorus, 2025). The research process consists of three main stages: pre-processing of data, determination of the number of clusters, and the clustering process.

1. Data Pre-processing

Production data between sub-districts has a varied scale, so normalization is carried out using Standard *Scaler* so that each commodity has a balanced contribution to the clustering process. This stage also handles a zero value that indicates no production in a given region (Noor, t.t.).

2. Determination of the Number of Clusters

Determining the optimal number of clusters (*k*) is an important stage because the results of *K-Means* are heavily influenced by this value. Two methods are used: *the Elbow Method* and *the Silhouette Score*.

a. *Elbow Method*

The Elbow Method calculates *the Sum of Squared Errors* (SSE) for various k values and shows the "elbow" point at *k = 2* as an initial indication of the optimal cluster number.

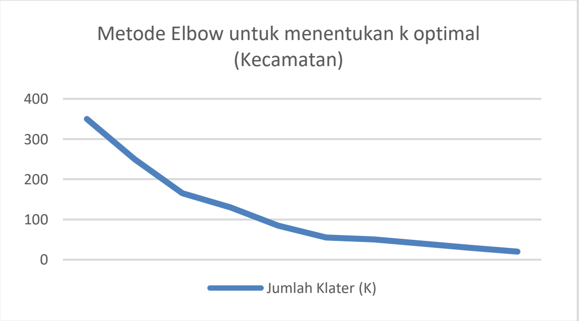

Figure 1. SSE calculation results with *Elbow Method*

Based on the images, it can be seen that the curve has decreased sharply to *k* = 2. After this point, the decline *SSE* become more sloping. This shows that the *Elbow* suggest the optimal number of clusters is 2 clusters (Guntara & Lutfi, 2023).

b. *Silhouette Score*

The Silhouette Score assesses the quality of separation between clusters, and the highest results are also found at *k = 2*.

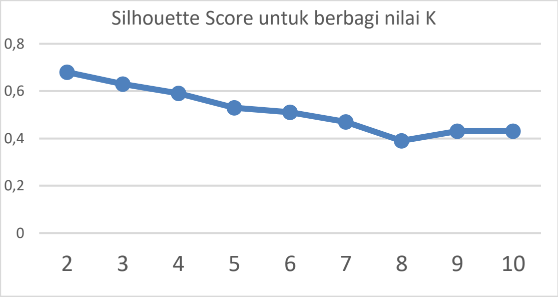

Figure 2. Silhouette Score calculation *results* for various *k values*

Based on Figure 2, the *Silhouette Score* The highest is obtained when the number of clusters is 2. This means that in terms of the quality of separation between clusters, the best results are obtained in *k = 2* (Guntara & Lutfi, 2023).

However, this study assigned three clusters (k = 3) to adjust the analysis *Business Intelligence*, so that the results can be categorized as high, medium, and low production. This approach is considered more relevant to agricultural planning and policy needs (Sjah, t.t.).

3. Clustering Process

Algorithm *K-Means* working by determining *Centroid* start at random, calculating the distance of each data to *Centroid* nearest using distance *Euclidean*, and update the position *Centroid* until the results are stable (Wakhidah, 2010).

The calculation of distances between data uses *Euclidean* distances, with the formula:

$$d(\ x_{i,cj})=\sum_{p=1}^{m}(xip-cjp)2 \tag{1}$$

with x_i as the production data of the ith sub-district, c_j as the centroid of cluster j, and M as the number of *commodities* (Prastanika & Wijayanto, 2023). The objective function of K-Means minimizes the total distance in a cluster through the equation:

$$J=\sum_{J-1}\sum_{Xi\ \in Cj}\|Xi-cj\|2 \tag{2}$$

The K-Means algorithm is implemented through three core stages. First, raw production data is standardized using the StandardScaler to equalize the scale between commodities so that all variables have a balanced contribution in the calculation of distances (Noor, t.t.). Second, the number of clusters is set to three based on the Elbow and Silhouette analysis which shows optimal value candidates as well as the need for high, medium, and low production categorization (Guntara & Lutfi, 2023). Third, the clustering process is carried out iteratively by *calculating* the distance of each sub-district to the centroid using equation (1), determining cluster membership based on the minimum distance, and updating the centroid using the average value of cluster members. This process lasts until the model reaches convergence (Wakhidah, 2010). Through these stages, raw data on vegetable production can be mapped into three final clusters that represent different production patterns between sub-districts in the city of Bandung.

C. RESULTS AND DISCUSSION

1. Visualization of Results

This research resulted in three main clusters of vegetable production in the city of Bandung for the 2018–2023 period. To clarify the grouping pattern, the results are displayed in the form of a table of production averages, *Heatmap*, *scatter plot PCA*, bar chart *Top-5* sub-districts, as well as sub-district distribution. This presentation aims to make the differences in the characteristics of each cluster visible quantitatively and visually, making it easier to identify superior commodities and regions that are the center of production (Akbar & Octaviany, 2021).

a. Production Average Heatmap

The average heatmap of commodities is compiled based on the centroid values generated from the *K-Means clustering process*. The centroid value is calculated using Euclidean distances according to the principle of clustering calculation (Prastanika & Wijayanto, 2023) and is optimized through an iterative process of centroid renewal until the model reaches convergence (Wakhidah, 2010). The determination of the number of clusters was carried out through evaluation using *the Elbow Method* and *Silhouette Score* which showed a stable pattern of data separation in the three main groups (Guntara & Lutfi, 2023). Based on the value of the final centroid, the high production cluster has the largest average especially on the "Other" commodity. The medium production cluster is characterized by the dominance of the kale commodity, while the low production cluster shows the smallest average of most commodities, although Potatoes still contribute relatively higher than other commodities in the cluster. This difference in centroid values is the basis for the preparation of *the heatmap* and describes the different characteristics of vegetable production in each cluster.

b. Scatter Plot PCA

Since the data consists of 12 commodities, dimension reduction using *Principal Component Analysis* (*PCA*) is carried out to project the results of clustering into two dimensions so that the grouping pattern is easier to observe. The results of the projection show that each sub-district forms three clearly separate groups. The high production cluster is concentrated in the area with the largest projection value, the medium production cluster is in the middle position, while the low production cluster is concentrated in the area with the smallest projection value. This separation pattern confirms that *the K-Means algorithm* is able to distinguish sub-districts based on the similarity of their production characteristics consistently, in line with the findings in previous research (Sofyan & Sitorus, 2025).

c. Top-5 Trunk Graph of Districts

The bar graph is used to identify the sub-districts that act as the center of vegetable production in each cluster based on the three main commodities analyzed in the analysis, namely: Other category (high production), Kale (medium production), and Potato (low production).

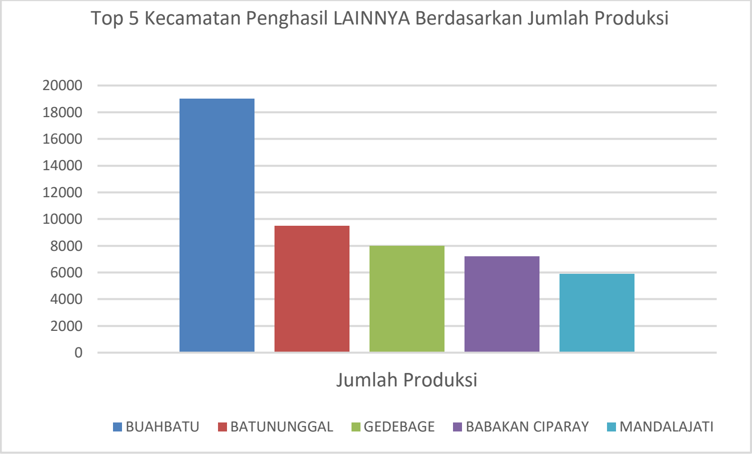

Figure 3. Top-5 sub-districts – Other

This category includes a wide range of horticultural commodities in addition to those specifically mentioned. The graph shows that the five sub-districts with the largest contribution are dominated by Cluster 0, indicating the existence of horticultural production centers centered in several areas with the highest production capacity.

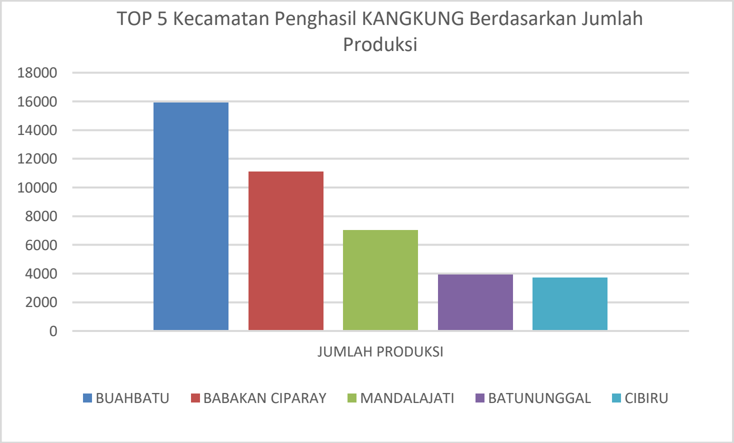

Figure 4. Top-5 sub-districts – Kangkung

The visualization in Figure 4 shows the distribution of kale production according to each sub-district in the cluster. The graph shows that kale has a moderate but more even production level in the sub-districts that are included in Cluster 1. This pattern indicates the existence of areas with favorable agronomic conditions, such as water availability and fast planting cycles, so that kale becomes a relatively stable commodity produced in the region.

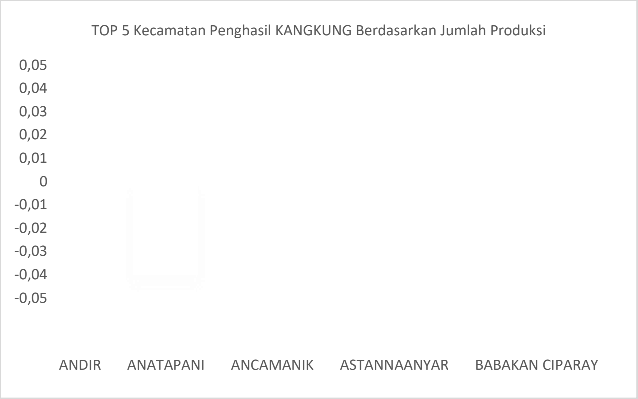

Figure 5. Top-5 sub-districts – Bawang Kentang

The distribution of potato production visualized in Figure 5 shows that this commodity falls into the category of low production, with a small production value and is concentrated in only a few sub-districts. The graph shows that most regions do not have a high production capacity for this commodity, so Potatoes are not a leading commodity for most sub-districts in the city of Bandung. These conditions indicate that Potato cultivation tends to depend on certain ecological and agronomic factors that most regions do not have.

d. District Distribution per Cluster

In addition to being based on commodities, the results of clustering are visualized through graphs that display the distribution of sub-districts in each cluster so that the grouping pattern is easier to observe. The visualization shows that Cluster 0, which is categorized as a high-production group with dominance in the "Other" category, includes the areas of Buahbatu, Batununggal, Gedebage, Babakan Ciparay, and Mandalajati. Meanwhile, Cluster 1 which represents medium production and is dominated by kale consists of the sub-districts of Buahbatu, Babakan Ciparay, Mandalajati, Batununggal, and Cibiru. As for Cluster 2, as a low-production group with a dominance of potatoes, it includes Sukasari, Ujung Berung, Cinambo, Rancasari, Antapani, Bandung Kulon, Cibeunying Kaler, and Cibeunying Kidul. The distribution pattern shown in the graph shows that vegetable production in the city of Bandung is not evenly distributed, but is concentrated in certain areas according to its superior commodities, as also shown in the previous study (Agustian & Mayrowani, 2008).

2. Discussion

The results of the clustering show that vegetable production in the city of Bandung is divided into three main groups with different characteristics. Cluster 0 is categorized as high production with dominance in *the Other category* which includes various horticultural commodities. Cluster 1 is classified as medium production with the main advantage in *kale*, while Cluster 2 is a low-production group dominated by *potatoes*. This pattern shows that the distribution of production is uneven, but rather concentrated in areas with certain agronomic characteristics.

The difference between clusters indicates the specialization of production areas in the city of Bandung. Clusters with high production are generally located in areas with arable land and adequate agricultural facilities, while low clusters are located in areas with land conditions that do not support optimal production. Factors such as water availability and land type also affect the dominance of commodities in each cluster, such as kale in watery areas and potatoes in certain land.

Visualizations in the form of heatmaps, PCA scatter plots, and Top-5 sub-district bar graphs support the findings of this clustering. The heatmap shows the difference in production intensity between clusters, the scatter plot emphasizes the separation of sub-district groups, and the bar graph identifies the main production centers of each commodity. This combination of visualization provides a comprehensive overview of the vegetable production patterns between regions in the city of Bandung.

From the perspective of Business Intelligence (BI), the results of this study have strategic implications for decision-making in the agricultural sector. Clustering can help design a more efficient distribution of crop yields, direct the development of agricultural areas according to superior potential, and become the basis for data-based policies that support regional food security. Thus, the integration K-Means and BI plays an important role in creating an adaptive and data-driven agricultural system (Nurzaman & Sari, 2023).

D. CONCLUSIONS AND SUGGESTIONS

a. Conclusion

This study aims to group sub-districts in the city of Bandung based on the production pattern of 12 vegetable commodities using the K-Means Clustering algorithm, and the results of the analysis successfully show three main clusters that represent different production levels. Cluster 0 depicts areas with high production, Cluster 1 shows moderate production, and Cluster 2 shows low production. These findings prove that the distribution of vegetable production between sub-districts is uneven and has typical agronomic characteristics. Overall, this research is able to fulfill the goal of producing a mapping of vegetable production clusters that are useful in supporting agricultural planning, data-based policy development, and the use of Business Intelligence in identifying production centers in a more targeted manner.

b. Suggestion

Further research is recommended to use longer data periods or recent data so that production trend patterns can be observed more comprehensively. The addition of external variables such as environmental and socioeconomic factors can also enrich the results of clustering. In addition, other clustering methods can be considered to be compared with the results of K-Means so that a stronger picture is obtained about the characteristics of vegetable production in the city of Bandung. The development of an interactive dashboard-based visualization system can also be a follow-up to support decision-making in the agricultural sector.

THANKS

The author expresses his gratitude to Dr. Cahyono Budy Santoso, ST, MMSI, as the supervisor, for his very valuable direction, guidance, and input during this research process. Thank you also to Universitas Pembangunan Jaya for providing support in the form of facilities, data access, and opportunities so that this research and publication can be carried out properly.

Not to forget, the author would like to express his deepest gratitude to parents for the prayers, moral support, and enthusiasm that are always given, so that this research can be completed properly.

REFERENCE

Agustian, A., & Mayrowani, H. (2008). Distribution Pattern of Potato Commodities in Bandung Regency, West Java. *Journal of Development Economics: A Study of Economic and Development Problems*, *9*(1), 96. https://doi.org/10.23917/jep.v9i1.1034

Akbar, R., & Octaviany, M. (2021). Design of Dashboard Visualization and Clustering by Applying Business Intelligence at the Dharmasraya Regency DPMPTSP Office. *Journal of Informatics Education and Research (JEPIN), 7*(3), 340. https://doi.org/10.26418/jp.v7i3.49719

Andriani, R., Setyanto, A., & Nasiri, A. (2020). Evaluation of Information Systems Using Technology Acceptance Model with the Addition of External Variables. *Journal of Information Technology and Computer Science*, *7*(3), 531. https://doi.org/10.25126/jtiik.202073850

Ardiansyah, R., & Hikmawan, M. D. (2020). *Collaboration of Local Organic Farmers as a Form of Food Security*.

Buddhism, and Yoga. (2023). Optimization for Vegetable Crop Production in Indonesia. *Nuansa Informatica*, *17*.

Dirayati, F., Sari, R. A., & Purnomo, R. F. (2025). *Design and Implementation of Internet of Things Based Smart Agriculture Systems to Increase Agricultural Productivity*. *6*(2).

Farismana, R. (2024). Application of K-Means Clustering for Mapping. *Jisamar Journal*, *8*. https://doi.org/10.52362/jisamar.v8i3.1572

Scott, E. (2025). *Clustering Kos with K-Means Algorithm for Venue Recommendations Based on Price and Facilities*. *6*(3), 1990–1995.

Guntara, M., & Lutfi, N. (2023). Cluster Count Optimization in Clustering with KMeans Algorithm Using Silhouette Coeficient and Elbow Method. *JuTI "Journal of Information Technology," 2*(1), 43. https://doi.org/10.26798/juti.v2i1.944

Lianita, E., Pratama, A., & Ulva, A. F. (2024). Application of the K-Means Clustering Method for Mapping Vegetable Productivity Based on Geographic Information System in North Sumatra Province. *Journal of Information Systems and Technology (JustIN), 12*(2), 232. https://doi.org/10.26418/justin.v12i2.72934

Mulyana, M., Nurendah, Y., & Effendy, M. (2025). *Business Intelligence: Concepts and Implementation in Decision Making.*

Noor, M. H. (t.t.). *Master of Informatics Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University, Malang 2024.*

Nurzaman, M. Y., & Sari, B. N. (2023). *The implementation of K-Means Clustering in the grouping of the number of farmers based on sub-districts in West Java Province*. *10*(3).

Prastanika, W. W., & Wijayanto, A. W. (2023). Hard and Soft Clustering Analysis for the Grouping of Indonesian Food Security Indicators 2021. *Journal of Information Systems and Technology (JustIN),* *11*(4), 596. https://doi.org/10.26418/justin.v11i4.68400

Rianti, R., Andarsyah, R., & Awangga, R. M. (2024). Application of PCA and Clustering Algorithm for Higher Education Quality Analysis in LLDIKTI Region IV. *NUANCES INFORMATICS*, *18*(2), 67–77. https://doi.org/10.25134/ilkom.v18i2.211

Riyanda, M. D., & Suyanto, S. (2020). Implementation of Business Intelligence in the Analysis of the Development of Agricultural Products in South Sumatra Province. *Journal of Computer and Information Systems Ampera*, *1*(3), 174–184. https://doi.org/10.51519/journalcisa.v1i3.44

Sjah, Z. (t.t.). *Agricultural policy formulation requires an understanding of farmers' motivations and needs*.

Sofyan, S. N., & Sitorus, Z. (2025). Implementation of data mining for shallot productivity clustering using the K-Means method. *Jatilima : Journal of Multimedia and Information Technology*, *07*, 109–121. https://doi.org/doi.org/10.54209/jatilima.v7i02.1442

Wakhidah, N. (2010). Clustering Using the K-Means Algorithm. *Journal of Transformational, 8*(1), 33–39. https://doi.org/10.26623/transformatika.v8i1.45

Yasmin, P. Y. (t.t.). *Improving the Understanding of Data Visualization in Processing Information*.

Yusdja, M., & Sayaka, A. (2017). *Food and Agriculture Economics in Indonesia*. IPB Press.