JRSSEM 2026, Vol. 5, No. 11, June
E-ISSN: 2807 - 6311
P-ISSN: 2807 - 6494

Regional Cluster Analysis Based on Community Consumption Patterns and Quality of Life for Market Segmentation Strategies Using the K-Means Algorithm

Aziz Kurnia Sandy*, Kemal Ade Sekarwati
Universitas Gunadarma, Indonesia
Email: azizsandy751@gmail.com*, ade@staff.gunadarma.ac.id

Keywords:
K-Means;
market segmentation;
data mining;
IPM;
BPS 2025.

Abstract

This study addresses the challenge of regional inequality in Indonesia, where economic growth and consumption patterns do not always correspond to improvements in quality of life. Differences in food expenditure, non-food expenditure, and life expectancy across provinces create the need for more objective regional mapping to support market segmentation strategies. This study aims to classify 38 provinces in Indonesia based on community consumption patterns and quality of life using the K-Means clustering algorithm. The research employed a quantitative data mining approach using secondary data from the Central Statistics Agency in 2025, including life expectancy, average food expenditure, and average non-food expenditure. Data were cleaned, integrated, and transformed using Min-Max Scaling before clustering. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, resulting in three regional clusters. The model produced an Average Silhouette Score of 0.5820 and a Davies-Bouldin Index of 0.6095, indicating valid cluster separation. The results identify three market categories: Basic Priority Areas, High Expenditure Areas, and Stable Growth Regions. Each cluster reflects different purchasing power, welfare conditions, and strategic market potential. The study concludes that K-Means clustering can provide a useful decision-support framework for businesses and policymakers in designing targeted, efficient, and region-based market segmentation strategies.

INTRODUCTION

The economic development of Indonesia by 2025 shows complex dynamics, meaning that national economic growth is not always directly proportional to the equitable distribution of quality of life across provinces. This phenomenon creates challenges for the business sector in determining market expansion and segmentation strategies. Data from the Central Statistics Agency (BPS) in 2025 reveals significant variations in people's per capita expenditure on food and non-food categories, as well as quality of life indicators represented through Life Expectancy (UHH).

Human development in Indonesia by 2025 shows significant achievements, but still presents challenges in the form of variations in quality of life across certain regions. The Human Development Index (HDI) is a composite measure that represents the success of developing human quality of life through the dimensions of longevity and healthy living, knowledge, and a decent standard of living. The standard of living itself is measured through average per capita expenditure that has been adjusted, so that areas with high HDI values

generally show more diversified consumption patterns. In these regions, expenditure allocation is no longer dominated by basic food needs alone, but has shifted significantly toward meeting broader quality of life needs such as access to health and education services (Setiawan, 2022). HDI aims to provide a more comprehensive picture of societal progress by considering the dimensions of health, education, and living standards (Alif & Fahmi, 2025). One of the main indicators representing the level of public health is Life Expectancy (UHH).

Based on the SP2020 Long Form Life Expectancy data, the national average UHH in Indonesia currently stands at 74 years, with DKI Jakarta recording the highest rate at 85.05 years. However, there are still areas with life expectancy far below the national average, such as Mountainous Papua at 54.91 years. This striking inequality in isolated areas highlights that the quality of health infrastructure is not uniformly distributed, underscoring the need for more in-depth regional mapping to understand the actual characteristics of each province (Hou dan Wang 2024; Song *et al.* 2024; Chiarot *et al.* 2025). From the perspective of business information systems, the large volume of raw data from 38 provinces in Indonesia will not yield strategic value unless it is processed into meaningful information. Consumption patterns (expenditure) reflect purchasing power and household economic priorities, while life expectancy reflects the level of welfare and sustainability of a region. There are anomalies in some areas where very high consumption expenditure—driven by logistics costs—does not correspond to an optimal quality of life, as seen in Papua compared to urban areas on the island of Java.

Alongside improvements in quality of life, people's spending patterns have also undergone significant transformation (Estes dan Sirgy 2019). An economic theory known as Engel's Law, put forward by Ernst Engel in the 19th century, states that the proportion of income spent on food decreases as income increases, while spending on non-essential items such as clothing, entertainment, and transportation rises. This law underpins many studies on household consumption and supports the concept of income elasticity (Norshahlan et al., 2023). BPS data from 2025 reinforces this phenomenon by showing that average monthly per capita non-food expenditure in many regions of Indonesia has now exceeded food expenditure. This trend indicates a shift in purchasing power toward lifestyle, health, and education, which requires a more precise market segmentation strategy to be effectively targeted (Akinrinoye *et al.* 2020).

This shift in consumption patterns is characterized by dynamic income elasticity, so that people are no longer focused solely on meeting basic needs, but are increasingly oriented toward comfort, experience, and quality of life (Norshahlan et al., 2023). The large volume of available data often makes it difficult for businesses to map market potential objectively. Without the support of information technology, UHH data and BPS 2025 expenditure data would remain as passive descriptive reports. Extracting strategic knowledge from such reports requires data mining techniques (Moayer 2016; Bharara *et al.* 2017; Kolling *et al.* 2021). The clustering process offers an effective solution for grouping the 38 provinces in Indonesia based on similarities in lifestyle and quality of life (Agneresa et al., 2022; Sibarani et al., 2022).

The use of the K-Means algorithm can be a solution for processing life expectancy data and spending patterns from 2025. The K-Means algorithm is employed for regional clustering due to its ability to handle large datasets efficiently. By processing the variables of life expectancy and spending patterns, the algorithm identifies distinct clusters, such as regions

with a high cost of living but low quality of life, or regions with established premium market potential. In-depth data analysis forms the foundation of an effective Decision Support System, making the use of this method highly relevant. The results of this clustering will not only benefit the government in mapping welfare conditions, but also serve as a guide for industry in designing more effective and efficient market segmentation strategies across Indonesia (Sinaga et al., 2025). To address this complexity, a data mining approach using the K-Means clustering algorithm is needed. This algorithm is capable of grouping regions based on similarities in consumption patterns and quality of life automatically and objectively. By identifying these regional groupings, businesses can conduct market segmentation that is more precise, efficient, and effective in allocating company resources according to the unique profile of each region by 2025.

RESEARCH METHOD

Data Collection

The data used is secondary data sourced from the Central Statistics Agency (BPS) of the Republic of Indonesia for the period of 2025. The data covers 38 provinces in Indonesia with the variables Life Expectancy (UHH) as indicators of quality of life and public health degrees, Average Per Capita Food Expenditure as an indicator of consumption of basic necessities, and Average Non-Food Per Capita Expenditure as indicators of lifestyle, access to education, and tertiary needs. The variables are downloaded on the BPS website page in excel form and collected into one group. The variables that have been collected are then stored in csv form and called in the form of program code. UHH data, food production and non-food production that have been summarized into one in excel can be seen in Table 1.

Table 1. UHH Data, Food and Non-Food Production 2025

No.	Provinsi	UHH	Peng makanan	Peng non makanan
1	ACEH	76,23	741.980,00	557.776,00
2	SUMATERA UTARA	76,47	765.842,00	632.045,00
3	SUMATERA BARAT	77,27	821.552,00	720.558,00
4	RIAU	76,31	844.678,00	765.749,00
5	JAMBI	75,13	775.092,00	702.013,00
6	SUMATERA SELATAN	74,76	691.256,00	592.283,00
7	BENGKULU	75,68	740.786,00	757.958,00
8	LAMPUNG	73,98	663.740,00	575.873,00
9	KEP. BANGKA BELITUNG	75,26	915.697,00	849.854,00
10	KEP. RIAU	80,53	1.063.897,00	1.404.422,00
11	DKI JAKARTA	85,05	1.153.404,00	1.809.008,00
12	JAWA BARAT	75,90	832.486,00	891.881,00
13	JAWA TENGAH	74,77	667.536,00	669.491,00
14	DI YOGYAKARTA	82,48	766.284,00	1.076.928,00
15	JAWA TIMUR	76,13	715.989,00	703.152,00
16	BANTEN	77,25	870.648,00	879.898,00
17	BALI	79,37	827.210,00	1.137.408,00
18	NUSA TENGGARA BARAT	73,97	762.911,00	639.207,00

19	NUSA TENGGARA TIMUR	69,89	559.895,00	468.980,00
20	KALIMANTAN BARAT	72,09	779.641,00	690.293,00
21	KALIMANTAN TENGAH	74,86	849.309,00	746.856,00
22	KALIMANTAN SELATAN	76,10	827.593,00	775.909,00
23	KALIMANTAN TIMUR	79,39	956.141,00	1.161.212,00
24	KALIMANTAN UTARA	74,04	838.916,00	887.355,00
25	SULAWESI UTARA	76,32	718.423,00	692.081,00
26	SULAWESI TENGAH	72,82	648.896,00	607.498,00
27	SULAWESI SELATAN	75,92	664.456,00	694.365,00
28	SULAWESI TENGGARA	74,25	650.919,00	674.166,00
29	GORONTALO	72,62	650.308,00	665.429,00
30	SULAWESI BARAT	71,16	575.436,00	493.877,00
31	MALUKU	74,09	679.006,00	719.492,00
32	MALUKU UTARA	72,52	749.636,00	742.896,00
33	PAPUA BARAT	68,48	878.045,00	827.294,00
34	PAPUA BARAT DAYA	70,55	857.182,00	959.693,00
35	PAPUA	74,69	830.864,00	979.903,00
36	PAPUA SELATAN	69,54	805.371,00	712.048,00
37	PAPUA TENGAH	60,64	874.847,00	584.240,00
38	PAPUA PEGUNUNGAN	54,91	1.256.747,00	685.250,00

Data Cleaning

After the data is collected, the cleanup stage is carried out in Google Colab using the Pandas library. This process aims to ensure data quality through:

1. Missing Value Handling: Ensure there is no empty data in each province.
2. Data Consistency: Align the format of writing the province name and ensure that the data types on the UHH and Expenditure variables are numerical (float or int).
3. Correlation Check between Variables: Ensure that the selected variable has an appropriate correlation between the variables

Data Integration and Transformation

At this stage, several variables are combined into one raw data unit (Integration). Furthermore, Data Transformation is carried out using the Min-Max Scaling method. This transformation must be carried out because there is a large scale difference between the variables UHH (tens) and Expenditure (millions). The formula used is the result of a transformation that will make all the values of the variables in the range of 0 to 1, so that the K-Means algorithm can give a fair weight to each variable in the distance calculation.

Data Analysis (K-Means Clustering)

Clustering is a technique that is widely used in data mining and has many applications in various fields (Sharda et al., 2021). The K-Means Clustering algorithm is one of the non-hierarchical data clustering techniques used to divide a set of data into several clusters (Putri, 2024). Data that have similar characteristics will be placed in the same cluster, while data that has different characteristics will be grouped into different clusters. This way, data differences within a single cluster can be kept to a minimum. The K-Means method works by applying a partitioning algorithm that divides the data into several groups, with the aim of minimizing the

distance between each data and the corresponding cluster center (centroid). This stage is the core of the research using the K-Means clustering algorithm. This process is carried out so that the results of clustering are more optimal and accurate, using the following evaluation approaches and calculation methods.

1. Determining the Number of Clusters
   This stage of analysis uses the Elbow Method by calculating the Within-Cluster Sum of Squares (WCSS) to find the best elbow point, and is supported by Silhouette Score. The Elbow Method is one of the approaches used to determine the optimal number of clusters by observing the pattern of the percentage of cluster formation that forms an angle resembling an elbow at a certain point (Sikana & Wijayanto, 2021).
2. Cluster Center Initialization
   Randomly determine the center point of the initial cluster using the K-Means++ method. By assigning the number of clusters first, the algorithm will randomly select k points as the center of the initial cluster called the centroid
3. Determining the Distance between Data
   The calculation of the distance between each data and the center of the cluster is carried out using the Euclidean Distance method and the distance is calculated to determine the proximity of each data to the centroid of the cluster.
4. Data Allocation
   Group the data by closest distance after the distance is calculated, and then each data point will be allocated to the cluster that has the minimum distance.
5. Iteration
   This stage updates the position of the centroid based on the average of the new cluster members until it reaches a convergent state (no change in centroid position).

Evaluation and Interpretation of Results

The last stage of this study is the evaluation and interpretation stage of the results of data that have passed the clustering stage using the K-Means algorithm. The results evaluation stage was carried out by evaluating the quality of the clustering model produced through two technical metrics and the evaluation results were presented in the form of visualization of the average profile of the characteristics of each cluster based on min-max scaling, interactive visualization of consumption patterns and quality of life, and visualization of the complete distribution of the list of provinces per cluster. The result interpretation stage is the result of interpreting the clustering of data variables that have been grouped and then reprocessed through the profiling stage. The end result of this research method is a strategic map of market segmentation based on the unique characteristics of each cluster. This stage has gone through the previous process, namely calculating the average value of each variable in each cluster and providing a business identity.

1. Silhouette Score

The first metric measures how close each piece of data is to its own cluster compared to the other. A value close to 1 indicates excellent separation.

2. Davies-Bouldin Index (DBI)

The second metric measures the ratio of the average distance within a cluster to the distance between clusters. The smaller the DBI value, the more optimal the resulting cluster. The next stage is a profiling analysis by returning the data value to the original scale to describe

the economic characteristics of each cluster (Basic Priority Area, High Expenditure Area, Stable Growth Area).

3. Visualization

The results of the last evaluation were presented in the form of visualizations made into three forms, namely:

a. Visualize the average profile of each cluster's characteristics based on min-max scaling.
b. Interactive visualization of consumption patterns and quality of life.
c. Visualization of the complete distribution of the list of provinces per cluster.

RESULTS AND DISCUSSION

The results of data collection in this study cover 38 provinces in Indonesia using secondary data for 2025. The focus of the research includes three variables, namely Life Expectancy (UHH), Average Food Expenditure, and Average Non-Food Expenditure. This data provides an overview of the correlation between the quality of life and the purchasing power of people in each region. Data *mining* processing and modeling *of the K-Means clustering* algorithm is carried out by preparing libraries in *python programming*. After the library preparation stage is completed, it is followed by the collection of data that has been downloaded from the BPS website and then grouped into one in the form of *a csv file format* and called in the python program code. The display of the summoned data can be seen in Table 2.

Table 2. Data Display of 38 Provinces 2025

No	Provinsi	UHH	Peng_makanan	Peng_non_makanan
1	ACEH	76,23	741.980	557.776
2	SUMATERA UTARA	76,47	765.842	632.045
3	SUMATERA BARAT	77,27	821.552	720.558
4	RIAU	76,31	844.678	765.749
5	JAMBI	75,13	775.092	702.013
6	SUMATERA SELATAN	74,76	691.256	592.283
7	BENGKULU	75,68	740.786	757.958
8	LAMPUNG	73,98	663.740	575.873
9	KEP. BANGKA BELITUNG	75,26	915.697	849.854
10	KEP. RIAU	80,53	1.063.897	1.404.422
11	DKI JAKARTA	85,05	1.153.404	1.809.008
12	JAWA BARAT	75,90	832.486	891.881
13	JAWA TENGAH	74,77	667.536	669.491
14	DI YOGYAKARTA	82,48	766.284	1.076.928
15	JAWA TIMUR	76,13	715.989	703.152
16	BANTEN	77,25	870.648	879.898
17	BALI	79,37	827.210	1.137.408
18	NUSA TENGGARA BARAT	73,97	762.911	639.207
19	NUSA TENGGARA TIMUR	69,89	559.895	468.980
20	KALIMANTAN BARAT	72,09	779.641	690.293

21	KALIMANTAN TENGAH	74,86	849.309	746.856
22	KALIMANTAN SELATAN	76,10	827.593	775.909
23	KALIMANTAN TIMUR	79,39	956.141	1.161.212
24	KALIMANTAN UTARA	74,04	838.916	887.535
25	SULAWESI UTARA	76,32	718.423	692.081
26	SULAWESI TENGAH	72,82	648.896	607.498
27	SULAWESI SELATAN	75,92	664.456	694.365
28	SULAWESI TENGGARA	74,25	650.919	674.166
29	GORONTALO	72,62	650.308	665.429
30	SULAWESI BARAT	71,16	575.436	493.877
31	MALUKU	74,09	679.006	719.492
32	MALUKU UTARA	72,52	749.636	742.896
33	PAPUA BARAT	68,48	878.045	827.294
34	PAPUA BARAT DAYA	70,55	857.182	959.693
35	PAPUA	74,69	830.864	979.903
36	PAPUA SELATAN	69,54	805.371	712.048
37	PAPUA TENGAH	60,64	874.847	584.240
38	PAPUA PEGUNUNGAN	54,91	1.256.747	685.250

Table 2 is the result of summoning statistical data in the form of a table of provinces in Indonesia which includes three main indicators, namely Life Expectancy (UHH), expenditure on food (Peng_makanan), and non-food expenditure (Peng_non_makanan). The data shows significant variation between regions, for example, the province of DKI Jakarta was recorded as having the highest UHH value of 85.05 with non-food expenditure reaching 1,809,008, while Mountainous Papua had the lowest UHH value of 54.91 but recorded a fairly high food expenditure at 1,256,747. This entire dataset contains 38 data (index 0 to 37) that can be used to analyze the correlation between the level of health welfare and people's consumption patterns in various regions in Indonesia.

Data Cleansing

Data cleaning is one of the most important stages in the data mining process. At this stage, data that is incomplete, duplicate, or inaccurate will be deleted or corrected so that the data analyzed has good quality and is suitable for use (Putri, 2024). After the data collection stage, the data cleaning stage is carried out by handling missing *values,* checking data consistency and checking the correlation between data variables. This process aims to ensure data quality through:

1. Penanganan *Missing Value*

   This stage ensures that there is no blank data in each province. The results of this process can be seen in Figure 3.

```
RangeIndex: 38 entries, 0 to 37
Data columns (total 4 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Provinsi          38 non-null     object
 1   UHH               38 non-null     float64
 2   Peng_makanan      38 non-null     int64
 3   Peng_non_makanan  38 non-null     int64
```

Figure 1. Missing Value Check Results

Figure 1 presents a summary of *the DataFrame* structure with 38 *entries* and 4 columns. The result shows that all columns have the same *number of non-null* as the number of *entries*, so there is no empty data in the dataset.

2. Data Consistency

   This stage harmonizes the format of writing the name of the province and ensures that the data type in the UHH and Expenditure variables is numerical data (*float* or *int*). The result of the data structure obtained is that the float64 data type is used in the "UHH" column to accommodate decimal values, while the int64 data type is applied to the "Peng_makanan" and "Peng_non_makanan" columns to store integer data.

3. Check Correlation between Variables

   This stage ensures that the selected variable has an appropriate correlation between the variables. The results of this process can be seen in Figure 2.

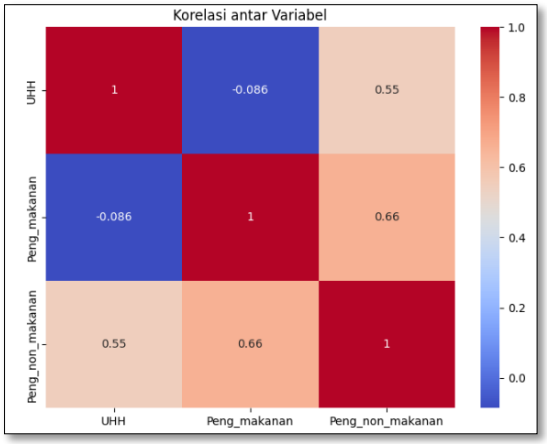

Figure 2. Correlation Checking Results Between Variables

The graphic caption in Figure 2 uses *the coolwarm* color scale to show the strength of the relationship between variables. Dark red color (Figure 1): indicates a perfect positive correlation (a variable relationship with itself). Pink to orange indicates a positive correlation (if one variable goes up, the other tends to go up). Blue indicates a negative correlation (if one variable goes up, the other one tends to go down). Based on the results of the numbers listed in Figure 4.9, the relationship between variables can be explained as follows:

a. Food Expenditure and Non-Food Expenditure (0.66) was the strongest correlation among the different variables in the study data. The number 0.66 indicates a fairly strong positive relationship. Regions that have high expenditure on food needs tend to have high expenditures on non-food needs as well. This usually reflects the level of economic well-being of a region in general.

b. UHH (Life Expectancy) and Non-Food Expenditure (0.55) had a moderate positive correlation of 0.55. This indicates that higher public spending on non-food needs (such

as health, education, and lifestyle) tends to be followed by higher Life Expectancy figures.

c. UHH (Life Expectancy) and Food Expenditure (-0.086) is a negative correlation of near zero (-0.086), which means that there is almost no significant linear relationship between Life Expectancy and the amount of spending on food alone. The blue color on this box shows a slight negative trend, however because the numbers are so small, this relationship is considered very weak or has no direct effect in the *clustering* model but but has a relationship with market strategy analysis that is influenced by habitual patterns in spending.

Data Integration and Transformation

The integration stage is carried out by combining the three main features into a feature and the transformation is carried out using *Min-max sacaling*. This normalization process is crucial in *the K-Means* algorithm because it uses *Euclidean distance* calculations. Without transformation, the expenditure variable that has a large nominal figure will dominate the calculation, so the UHH variable will not have a significant effect. *Min-max sacaling* can be used for all values mapped into the range 0 to 1 without changing the original distribution of the data. The results at this stage can be seen in Figure 3.

--- Data Setelah Min-Max Scaling) ---

	UHH	Peng_makanan	Peng_non_makanan
0	0.707366	0.261297	0.066264
1	0.715328	0.295539	0.121688
2	0.741871	0.375484	0.187741
3	0.710020	0.408671	0.221465
4	0.670869	0.308813	0.173902
5	0.658593	0.188506	0.092015
6	0.689117	0.259583	0.215651
7	0.632714	0.149020	0.079769
8	0.675182	0.510585	0.284228
9	0.850033	0.723255	0.698076
10	1.000000	0.851700	1.000000
11	0.696417	0.391175	0.315591
12	0.658925	0.154468	0.149632
13	0.914731	0.296173	0.453683
14	0.704048	0.223999	0.174752
15	0.741208	0.445938	0.306649
16	0.811546	0.383604	0.498816
17	0.632382	0.291333	0.127032
18	0.497014	0.000000	0.000000
19	0.570007	0.315341	0.165156
20	0.661911	0.415316	0.207366
21	0.703052	0.384153	0.229047
22	0.812210	0.568623	0.516580
23	0.634705	0.400402	0.312348
24	0.710352	0.227492	0.166490
25	0.594227	0.127719	0.103369
26	0.697080	0.150048	0.168194
27	0.641672	0.130622	0.153121
28	0.587591	0.129745	0.146601
29	0.539151	0.022302	0.018579
30	0.636364	0.170927	0.186945
31	0.584273	0.272283	0.204411
32	0.450232	0.456553	0.267393
33	0.518912	0.426614	0.366196
34	0.656271	0.388847	0.381278
35	0.485401	0.352264	0.181390
36	0.190113	0.451964	0.086013
37	0.000000	1.000000	0.161392

Figure 3. Data Results After *Min Max Scaling*

Figure 3 is the *output* of the normalization process using *the Min-Max Scaling method* on three main variables, namely 'UHH', 'Peng_makanan', and 'Peng_non_makanan'. The original values in the dataset have been transformed into the same scale range that is between the numbers 0 to 1. A value of 0 represents the smallest data and a value of 1 represents the largest data in each column. The results of the table show that the data is ready for further processing in the analysis without the dominance of variables due to differences in units or the magnitude of the original number.

Data Analysis

The next stage is the core stage of research using *the K-Means clustering algorithm*. This stage is carried out so that the results of clustering are more optimal and accurate using the following evaluation approaches and calculation methods.

1. Determining the Number of Clusters

   This stage of analysis uses *the Elbow Method* by calculating *the Within-Cluster Sum of Squares* (WCSS) to find the best elbow point, and is supported by *Silhouette Score*. The results of visualization at this stage are in accordance with Figure 4.

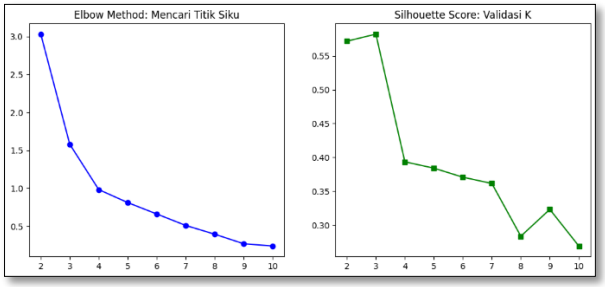

Figure 4. Visualization Results of *the Elbow Method*

   Figure 4 shows two graphs used to determine the optimal number of clusters in data analysis, namely *the Elbow Method* and *Silhouette Score* graphs. The graph on the left (*Elbow Method*), shows a decrease in the inertial value as the number of clusters increases. Significant elbow points or flattening begin to appear in the range of *values k=3* or *k=4. This value* indicates the number of suggested clusters. The graph on the right (*Silhouette Score*) is used to validate the k-value by showing that the highest score achieved at *k=3* is close to 0.60, which indicates that at that point the cluster structure has the most optimal separation and density compared to the number of other clusters.

2. Cluster Center Initialization

   The determination of the center point of the initial cluster is randomly determined using *the K-Means++* method by determining the number of clusters first, then the algorithm will select *k* as a random point to be used as the center of the initial cluster called *the centroid.* Based on the results of the previous visualization, the optimal number of clusters was determined to be 3 (k_optimal = 3). The code uses *the K-Means++* method for intelligent selection of the initial center point for faster and more accurate convergence processes, and sets *random_state*=42 to ensure consistent grouping results each time the program is run.

3. Determining the Distance between Data

   The calculation of the distance between each data and the center of the cluster is carried out using *the Euclidean Distance* method and the distance is calculated to determine the proximity of each data to the *centroid* of the cluster. This stage uses program code running the *kmeans.fit(x_scaled)* command to train the model using the normalized data, then determines the final position of the cluster centers through *cluster_centers_*. The final process of the program code calculates *Euclidean Distance* to measure how far away each provincial data point is from each existing cluster center (*centroid*).

4. Data Allocation

   Group the data by closest distance after the distance is calculated, and then each data point

will be allocated to the cluster that has the minimum *distance*. Theprogram is used to group each provincial data into a certain cluster based on the principle of *minimum distance* that has been calculated in the previous stage. The results of the cluster labels are then stored in a new column named 'Cluster', and an additional *DataFrame* named dist_df is created to summarize the distance ratio of each data to each cluster center along with the determination of the cluster's end.(C_i)

5. Iteration

The iteration process in the *K-Means* algorithm is carried out to minimize the distance between the region data and the within-cluster *sum of squares*, until it reaches a convergent condition where there is no more data transfer between groups. The results of this iteration produce numerical labels in the form of Cluster Numbers 0, 1, and 2. Technically, these numbers are mathematical identities that represent groups of regions with similar characteristics of consumption patterns and quality of life. The reason for using this numerical labeling before converting it into market segment names is to maintain the objectivity of the research results. These cluster numbers serve as a statistical container that summarizes the complexity of data into a simpler structure without removing the essence of the original variable. Using numbers as an initial differentiator, researchers can accurately validate the homogeneity within each group through *centroid values*.

Centroid position update based on the average of new cluster members until it reaches a convergent state (no change in *centroid position*). The program code used serves to display the final *centroid* that has reached the convergence point, which is the position of the central coordinates of each cluster after the repeated shift process is completed. The resulting coordinate values show the average profile of each group (cluster) based on the previously normalized variables. The results at this stage are in accordance with Figure 5.

```
Centroid Akhir Terkonvergensi:
 [[0.6523798  0.28226714 0.20384126]
 [0.0950564  0.72598199 0.12370264]
 [0.88741429 0.71452618 0.7382189 ]]
```

Figure 5. Results *of Converged Final* Centroid Values

Figure 5 is the value of the converged end *centroid* showing the central coordinates of three different groups of regions on the normalization scale. Cluster 0, which has coordinates [0.65, 0.28, 0.20], represents areas with medium to upper life expectancy (UHH) levels but the lowest levels of food and non-food expenditure among other groups. Cluster 1 showed contrasting conditions with values [0.09, 0.72, 0.12], which reflected the region with the lowest quality of life (UHH was close to zero on the normalization scale) despite its very high food expenditure. Meanwhile, Cluster 2 with coordinates [0.88, 0.71, 0.73] is the most superior group of regions that have a very high life expectancy rate supported by the level of non-food expenditure which is also the most dominant, indicating the most optimal level of economic welfare and quality of life. This iteration process ensures that each of those coordinates has reached a stable point and that the position of the *centroid* is no longer changing, thus providing an accurate and convergent representation of the region's characteristics.

Evaluation and Interpretation of Results

The last stage is the evaluation and interpretation of the results carried out in the form of evaluating the quality of the clustering model and translating it into relevant business identities.

Model Quality Evaluation

The evaluation was carried out using two main metrics to ensure the validity of the clusters formed, namely:

1. *Silhouette Score*

   The results of the *Silhouette Score calculation* are in accordance with Figure 6.

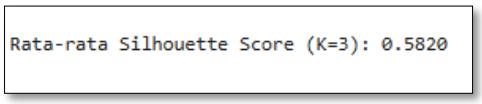

Figure 6. Silhouette *Score Results*

The first metric measures how close each piece of data is to its own cluster compared to the other. Based on the test results, an Average *Silhouette Score* (K=3) was obtained of 0.5820. A value close to 1 indicates excellent separation. This indicates that provinces in Indonesia have been grouped into the right clusters with a high level of consistency of cluster members.

2. *Davies-Bouldin Index* (DBI)

   The results of the DBI value after Figure 7.

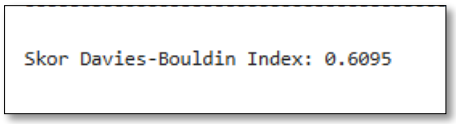

Figure 7. Results of DBI scores

The second metric measures the ratio of the average distance within a cluster to the distance between clusters. The *Davies-Bouldin Index* score obtained from the clustering results is 0.6095. The basic principle of DBI is that the smaller the value (closer to 0), the more optimal the cluster produces, then the value of 0.6095 strengthens the evidence that this clustering model has a very low degree of separation and overlap between regions.

Profiling and Visualization

The next stage is profile analysis by returning the data value to the original scale to describe the economic characteristics of each cluster. The results of the final evaluation were presented in three forms of visualization, namely:

1. Visualization of the Average Profile of Each Cluster's Characteristics (*Min-max sacaling*)

   This visualization provides a comparative overview of the performance of the UHH indicator and the consumption of each cluster on a normalization scale. The results of the visualization can be seen in Figure 8.

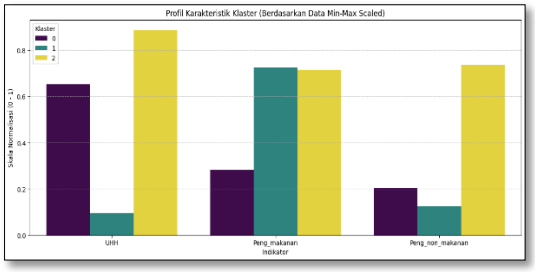

Figure 8. Visualization of the Average Characteristics of Each Cluster

Based on the graph in Figure 10 on the normalization scale (0-1), the results of the analysis show three very contrasting regional patterns. Cluster 0 is characterized as an area with a medium quality of life but has the lowest economic level, as seen from the relatively high UHH value (around 0.65) but with the lowest food expenditure (0.28) and non-food expenditure (0.20) among other groups. On the other hand, Cluster 1 shows a condition of socio-economic vulnerability as evidenced by the life expectancy rate being at its lowest point (0.09) even though food expenditure is quite high (0.72), indicating a large burden of food costs but not followed by adequate health quality. Meanwhile, Cluster 2 represents the region with the highest level of welfare and the most superior profile, as evidenced by the dominant values on all indicators, namely Life Expectancy (0.88), food expenditure (0.71), and non-food expenditure (0.73), reflecting a strong positive correlation between economic ability and the quality of life of people in the region.

2. Interactive Visualization of Consumption Patterns and Quality of Life

   This visualization is through *an interactive 3D Scatter Plot* that maps the linear relationship between life expectancy and people's spending patterns spatially. The results of the visualization can be seen in Figure 9.

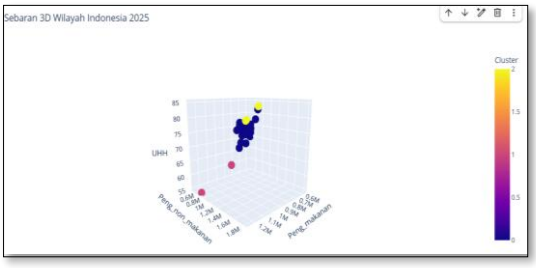

Figure 9. Interactive Visualization of Consumption Patterns and Quality of Life

Figure 9 is a form of interactive 3D visualization produced in this program mapping the distribution of provinces in Indonesia based on three main dimensions, namely food expenditure, non-food expenditure, and Life Expectancy (UHH), which effectively visualizes the regional progress profile through converged cluster coordinates. Cluster 2 emerged as the most superior group of regions with a central point (*centroid*) at coordinates [0.88, 0.71, 0.73], showing a strong correlation of high quality of life values directly proportional to the economic ability of the community to meet the needs of primary and secondary consumption. In contrast, Cluster 1 with coordinates [0.09, 0.72, 0.12] represents areas that are vulnerable, characterized by very low UHH numbers despite having relatively high food expenditure, while Cluster 0 is at the point [0.65, 0.28, 0.20] which describes

areas with a medium level of quality of life but having more efficient or low consumption patterns. Through this 3D visualization, the significant differences between the groups of regions are clearly visible in spatial space and explain that the position of each province is determined by its smallest distance (*Euclidean distance*) to the *centroid* of the respective cluster.

3. Visualization of the Complete Distribution of Province List per Cluster
   This visualization displays a list of provincial members to identify the classification of the region in detail.

a. The UHH indicator and cluster consumption pattern 0 are in accordance with Figure 10.

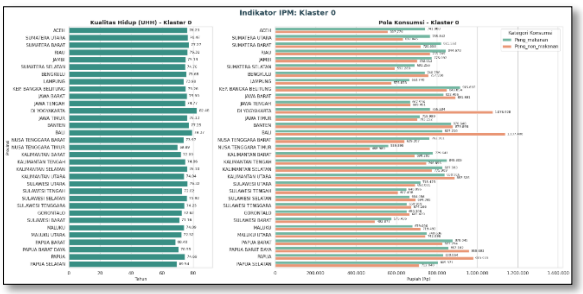

Figure 10. Cluster Indicator 0

Figure 10 is a visualization of data in cluster 0 through two types of bar graphs that distinguish quality of life and consumption patterns in 33 provinces of Indonesia. On the left, a single bar graph shows a Quality of Life (UHH) score. Yogyakarta Province stands out with the highest rate reaching 82.48 years, while West Papua has the lowest figure in the range of 68.48 years. On the right side, the double-bar graph compares consumption patterns between food expenditure (green color) and non-food (orange color), which shows significant diversity; for example, Bali and DI Yogyakarta show a very high dominance of non-food expenditure exceeding 1,000,000 rupiah, while provinces such as Southwest Papua have more dominant food expenditure. This visual table groups provinces in Cluster 0 that have characteristics ranging from 68 to 82 years with varying living costs between regions.

b. The UHH indicator and cluster 1 consumption pattern are in accordance with Figure 11.

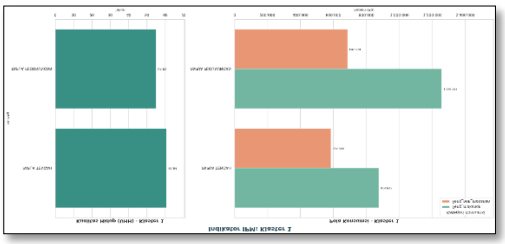

Figure 11. Cluster Indicator 1

Figure 11 is a data visualization in cluster 1 which includes the profile of quality of life and consumption patterns in the Central Papua and Mountainous Papua regions. On the Quality of Life (UHH) graph, it can be seen that the life expectancy rate is relatively low compared to other regions. Central Papua is at 60.64 years and Mountainous Papua at 54.91 years. Meanwhile, on the Consumption Pattern graph, the two provinces show that the

characteristics of the cost of living expenditure on food (green) are much more dominant than non-food expenditure (orange); Mountainous Papua recorded very high food expenditure reaching Rp1,256,247, inversely proportional to non-food expenditure which was only Rp685,253. Cluster 1 describes areas with significant health challenges and economic burdens that focus on meeting basic food needs.

c. UHH Indicators and Consumption Patterns of Cluster 2 are as shown in Figure 12

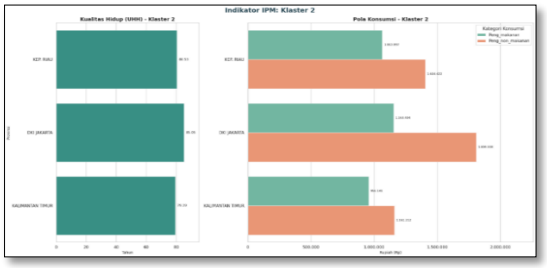

Figure 12. Cluster 2 Indicators

Figure 12 is a visualization of data in cluster 2 which includes the profiles of the provinces of Riau Islands, DKI Jakarta, and East Kalimantan with very high welfare characteristics. On the Quality of Life (UHH) graph, these three provinces show superior life expectancy figures. DKI Jakarta Province leads with 85.05 years, followed by Riau Islands at 80.53 years and East Kalimantan at 79.39 years. The Consumption Pattern graph reveals high levels of expenditure supported by non-food consumption data (orange) consistently dominating above food expenditure (green) across the region. DKI Jakarta even recorded the highest non-food expenditure reaching IDR 1,809,308. Overall, cluster 2 describes the group of regions with the best health standards and a very strong economic capacity of the community to meet needs outside of staple foods.

Interpretation of Results

The interpretation of the results is carried out by the process of describing the characteristics of the clusters of each group by returning the value of the scaled data to the original scale. The results of data processing which are divided into 3 main clusters can be seen in Table 3.

Table 3. Results of Inter-Regional Clusters

PROVINCE	CLUSTER	REMARKS
Aceh	0	Basic Priority Areas
North Sumatra	0	Basic Priority Areas
West Sumatra	0	Basic Priority Areas
Riau	0	Basic Priority Areas
Jambi	0	Basic Priority Areas
South Sumatra	0	Basic Priority Areas
Bengkulu	0	Basic Priority Areas
Lampung	0	Basic Priority Areas
Bangka Belitung Province	0	Basic Priority Areas
West Java	0	Basic Priority Areas
Central Java	0	Basic Priority Areas

In Yogyakarta	0	Basic Priority Areas
East Java	0	Basic Priority Areas
Banten	0	Basic Priority Areas
Bali	0	Basic Priority Areas
West Nusa Tenggara	0	Basic Priority Areas
East Nusa Tenggara	0	Basic Priority Areas
West Kalimantan	0	Basic Priority Areas
Central Kalimantan	0	Basic Priority Areas
South Kalimantan	0	Basic Priority Areas
North Kalimantan	0	Basic Priority Areas
Central Sulawesi	0	Basic Priority Areas
South Sulawesi	0	Basic Priority Areas
Southeast Sulawesi	0	Basic Priority Areas
Gorontalo	0	Basic Priority Areas
West Sulawesi	0	Basic Priority Areas
Maluku	0	Basic Priority Areas
North Maluku	0	Basic Priority Areas
West Papua	0	Basic Priority Areas
Southwest Papua	0	Basic Priority Areas
Papua	0	Basic Priority Areas
South Papua	0	Basic Priority Areas
Central Papua	1	High Production Region
Mountainous Papua	1	High Production Region
Riau Province	2	Stable Growth Region
Dki Jakarta	2	Stable Growth Region
East Kalimantan	2	Stable Growth Region

Based on the analysis of the Life Expectancy (UHH) indicator and public consumption patterns, three categories of business identity were produced as follows:

1. Cluster 0 (Basic Priority Area) / Staple Market.

This cluster consists of areas with an economic structure that focuses on meeting primary needs. A key characteristic of this cluster is a larger proportion of food expenditure relative to non-food expenditure, which suggests that a large portion of household income is allocated to daily consumption. The life expectancy figure at a moderate level indicates the need to strengthen access to health services and purchasing power. Consumers in this region have very high price sensitivity, so business strategies must prioritize affordability. Businesses are advised to provide products in economical packaging to accommodate consumer cash flow and to strengthen distribution through traditional markets and grocery stores.

The strategic implications for governments in this region require an active role in maintaining food price stability (volatile foods inflation) and strengthening social safety nets. The main focus of policy must be directed at increasing access to basic health services to raise life expectancy and providing targeted subsidies to maintain people's purchasing power. For businesses operating in this region, volume market success lies in supply chain efficiency. Given that consumers are highly price-sensitive, companies need to implement a low price point strategy—for example, through sachet or small-size packaging products suited to the daily or weekly income of the community—and ensure product availability in small stalls as the central point of transactions.

2. Cluster 1 (High Expenditure Area) / Special Penetration Market

This cluster includes the regions with the highest nominal per capita expenditure. Its main characteristics are a high Life Expectancy figure as well as a very significant dominance of non-food spending. Economically, these clusters are heterogeneous; It covers metropolitan areas with high lifestyles as well as remote areas with extreme logistics costs that trigger high prices of goods. Strategically, this region has a price sensitivity that tends to be low but demands functionality and stable availability of goods. Business people can implement pricing strategies that adjust *to area-based pricing* and leverage digital approaches to reach distribution points in these high-expense areas.

The strategic implications for the government face a dual challenge in the region. Metropolitan areas are focused on setting up healthy lifestyles and digital infrastructure, while in remote areas with high expenditures due to logistics costs, governments should intervene through connectivity improvements and transportation infrastructure development to reduce price disparities. The implication for business people in this region is that they offer thicker profit margins due to low price sensitivity, as long as the quality and availability of goods are guaranteed. In this region, marketing strategies should be more personalized and data-first (*digital-first*), with an emphasis on product added value and the reliability of delivery services to overcome geographical barriers that may exist.

3. Cluster 2 (Stable Growth Region) / Middle Market

This cluster includes regions with a stable quality of life and strong middle-class purchasing power, which serves as the backbone of national consumption. The spending pattern shows a balance between basic and secondary needs. People have the financial capacity to consume manufactured goods, mid-range automotive products, and modern retail items. Stable life expectancy figures reflect adequate access to health services, supporting sustainable economic productivity. Consumers in this cluster strongly prioritize the balance between price and quality. An effective marketing strategy is to strengthen loyalty through regular promotional programs and ensure product availability across modern retail and online channels to achieve high sales volumes.

The strategic implications for the government in this region include the need to maintain a conducive investment climate and job security to ensure that the middle class remains productive and does not fall into lower income brackets. Improving public facilities and promoting financial literacy education are important, as this region serves as a primary driver of the domestic economy. For businesses, this cluster represents a competitive battleground, where customer loyalty is highly valuable. The most appropriate strategy is to offer products with high value-for-money, supported by membership programs, periodic discounts, and a convenient shopping experience both offline—in supermarkets and malls—and online through e-commerce platforms, in order to capture consumption opportunities across secondary and lifestyle product categories.

CONCLUSION

Based on the results of the analysis and discussion that has been carried out in the previous chapter, several conclusions can be drawn as follows: Implementation of the K-Means Algorithm This research has successfully applied the K-Means algorithm to cluster in 38 provinces in Indonesia based on the variables of consumption patterns and quality of life in

2025. The data processing process carried out on the Google Colab platform through the normalization stage of Min-Max Scaling and optimal K determination using the Elbow Method resulted in the division of the region into three stable clusters. This is evidenced by the results of the evaluation of technical metrics which show an Average Silhouette Score of 0.5820 and a Davies-Bouldin Index score of 0.6095, which confirms that the model has a valid grouping quality with excellent separation between regions. The results of the analysis of the mapping of the region show that each cluster has different characteristics. Cluster 1, which is categorized as an area with a high level of spending, is dominated by non-food spending. This region includes metropolitan areas with strong purchasing power levels, as well as remote areas that face high distribution costs due to geographical conditions. Cluster 2 represents areas with relatively stable economic growth with consumption patterns of basic and non-basic necessities in balanced proportions and become the foundation of people's economic activities. Cluster 0 is an area that is still oriented towards meeting daily food needs with a very high level of sensitivity to price changes. Based on these characteristics, this study formulated a market segmentation approach that is tailored to the conditions of each region. In Cluster 1, the implementation of region-based pricing is considered appropriate to accommodate the relatively high difference in operational costs. Cluster 2, business actors are encouraged to increase sales volume by strengthening consumer loyalty and ensuring product availability in modern retail networks. In Cluster 0, marketing strategies are more focused on providing affordable prices through the use of economical packaging and expanding distribution in traditional markets to maintain competitiveness amid limited people's purchasing power.

REFERENCES

Agneresa, A., Hananto, A. L., Hilabi, S. S., Hananto, A., & Tukino. (2022). Strategi promosi penerapan data mining mahasiswa baru dengan metode K-means clustering. *Dirgamaya: Jurnal Manajemen dan Sistem Informasi, 2*(2), 25–34. https://doi.org/10.35969/dirgamaya.v2i2.275

Akinrinoye, O. V., Kufile, O. T., Otokiti, B. O., Ejike, O. G., Umezurike, S. A., & Onifade, A. Y. (2020). Customer segmentation strategies in emerging markets: A review of tools, models, and applications. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6*(1), 194–217.

Alif, M. F., & Fahmi, A. (2025). Klasterisasi wilayah kemiskinan Jawa Tengah menggunakan K-means berbasis indikator sosial-ekonomi. *Jurnal Nasional Teknologi dan Sistem Informasi, 11*(3), 302–309.

Bharara, S., Sabitha, A. S., & Bansal, A. (2017). A review on knowledge extraction for business operations using data mining. In *2017 7th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 512–518). IEEE.

Chiarot, C. B., Janes, C. R., Abdullah, A. Y. M., Goma, F., Phiri, M., Singini, D., Zulu, R., & Butt, Z. A. (2025). Evaluating the concept of access as a critical dimension of universal health coverage in the context of rural communities of Western Province, Zambia. *SSM–Health Systems, 100106*.

Estes, R. J., & Sirgy, M. J. (2019). Global advances in quality of life and well-being: Past, present, and future. *Social Indicators Research, 141*(3), 1137–1164.

Hou, D., & Wang, X. (2024). Unveiling spatial disparities in basic medical and health services: Insights from China's provincial analysis. *BMC Health Services Research, 24*(1), 329.

Kolling, M. L., Furstenau, L. B., Sott, M. K., Rabaioli, B., Ulmi, P. H., Bragazzi, N. L., & Tedesco, L. P. C. (2021). Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development. *International Journal of Environmental Research and Public Health, 18*(6), 3099.

Maori, N. A., & Evanita, E. (2023). Metode elbow dalam optimasi jumlah cluster pada K-means clustering. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer, 14*(2), 277–288.

Moayer, S. (2016). *Developing a model for competitive advantage through integration of data mining within a strategic knowledge management framework: A deep case study of a global mining and manufacturing company*.

Norshahlan, M., Jaya, H., & Kustini, R. (2023). Penerapan metode clustering dengan algoritma K-means pada pengelompokan data calon siswa baru. *Jurnal Sistem Informasi Triguna Dharma (JURSI TGD), 2*(6), 1042–1053.

Putri, A. (2024). Pentingnya data cleaning sebelum visualisasi: Teknik dan tips. *Teknologipintar.org, 4*(5), 2024–2025.

Setiawan, M. B. (2022). Indeks pembangunan manusia Indonesia. *Jurnal Economia, 18*(1), 1–15.

Sharda, R., Delen, D., & Turban, E. (2021). *Analytics, data science, & artificial intelligence: Systems for decision support* (11th ed.). Pearson.

Sibarani, H., Saputra, W., Gunawan, I., & Nasution, Z. M. (2022). Penerapan metode K-means untuk pengelompokan kabupaten/kota di Provinsi Sumatera Utara berdasarkan indikator indeks pembangunan manusia. *JATI (Jurnal Mahasiswa Teknik Informatika), 6*(1), 154–161.

Sikana, A. M., & Wijayanto, A. W. (2021). Analisis perbandingan pengelompokan indeks pembangunan manusia Indonesia tahun 2019 dengan metode partitioning dan hierarchical clustering. *Jurnal Ilmu Komputer, 14*(2), 66–78.

Sinaga, E., Purba, M., Situmorang, W. R., Nainggolan, G. P., Situmorang, J., & Boot, R. (2025). Elastisitas pendapatan dan pola konsumsi rumah tangga: Studi pada masyarakat perkotaan menengah. *Jurnal Akademik Ekonomi dan Manajemen, 2*(2), 577–587.

Song, C., Fang, L., Xie, M., Tang, Z., Zhang, Y., Tian, F., Wang, X., Lin, X., Liu, Q., & Xu, S. (2024). Revealing spatiotemporal inequalities, hotspots, and determinants in healthcare resource distribution: Insights from hospital beds panel data in 2308 Chinese counties. *BMC Public Health, 24*(1), 423.