**Clustering model of electricity load profile using K-means clustering: A case study of electricity companies in Indonesia**

Radit Rahmadhan1 and Meditya Wasesa2

1,2*School of Business and Management, Institut Teknologi Bandung, Indonesia*

**Abstract.**

**Background**: The increase in the burden of electricity consumption in Indonesia, especially the West Sumatra area, has significantly focused on electricity supply. Developing a predictive model related to a load of electricity consumption is very important in the management of electric power provided by the company.

**Objective**: This study aims to develop a predictive model to determine the possibility of using the profile load used by customers from the provided electrical power.

**Method**: Using customer transaction data of PT. PLN Persero from January 2019 to December 2020, we used an unsupervised machine learning model, namely K-Means Clustering. We assessed clustering using the Elbow and Principal Component Analysis (PCA) methods.

**Results**: In terms of grouping using the K-Means clustering model with validation from the Elbow and Principal Component Analysis (PCA) methods, three groups were found based on the consumption of electricity used by customers, from the findings of these groups they will be ranked so that the company can plan future actions.

**Conclusion**: Clustering using the K-Means Clustering model based on the ranking carried out can help companies optimize the consumption of electricity used by customers.

**Keyword**: *Clustering, K-Means Clustering, Elbow Method, Principal Component Analysis, PCA,*

*Load Profile, Unsupervised machine learning*

# Introduction

The electricity consumption in Indonesia continues to increase from 2015 to 2020 around 98.89%, and business customers dominate electricity consumption [1]. PT. PLN Persero is the only electricity provider in Indonesia that provides higher power for the entire region, including the West Sumatra region, due to the increasing electricity consumption focusing on business customers. This is because it has the potential to increase the company's revenue. However, according to customer information and news obtained, blackouts often occur in West Sumatra and appear four times a month [2].

Based on the data analysis results that have been carried out, power outages cause the average electricity usage time for business customers to be under 50 hours. Based on information from the Commercial Manager of PLN for the West Sumatra Region, the incident was due to customers using power above 200 thousand using a higher peak load electricity usage time than electricity outside peak hours. During off-peak hours, customers rarely use it. Based on these problems, PT. PLN Persero West Sumatra must understand the characteristics of customers so that the use of electricity at times outside the peak load is more optimal based on customer segmentation.

Customer segmentation is one way to understand customer preferences better. According to previous research, customer segmentation refers to grouping customers into similar characteristics that can be used to predict future customer actions or behavior [3]. Customer segmentation is used to predict customer characteristics in buying or using facilities provided by the company by mapping customer characteristics to increase sales or use facilities. That customers use and build relationships and enhance customer commitment to building a solid business[4][5].

Previous research on customer segmentation based on electricity consumption is hard to find. This study uses the clustering method to create customer segmentation. Clustering is part of data segmentation to group large amounts of data into several groups with the same characteristics [6]. Clustering is also widely used to understand customer behavior to increase company profitability. The clustering model used in this study is an efficient K-Means Clustering approach to evaluate customer differences in using electricity consumption. The main factor in selecting this model is the handling of several large data sets, such as data owned by PT. PLN Persero. The dataset used is the customer transaction data of PT. PLN Persero from 2019 to 2020. The data points that will be predicted are installed power at the customer, peak load electricity usage time, peak load electricity usage time. We want to break down the business customers with high potential in peak load electricity consumption into several dimensions. A validation method is needed to determine the best number of clusters to determine the clustering measurements. The Elbow method is used for the correct number of groups by looking at the SSE value by looking at the sloping point of the curve determined[6][7]. Principal Component Analysis is used to eliminate outliers in the data applied to the clustering model for more accurate grouping [8], [9].

This study aims to develop a predictive model of electricity consumption by examining the energy consumption patterns of business customers using the data described previously. We divide consumption behavior into two parts: customers who use peak-load electricity and those who are off-peak, using the K-Means Clustering model in classifying and determining the best number of clusters using Elbow and Principal Component Analysis (PCA) methods. We categorize customers based on their average electricity usage per month. These findings can help companies identify the electricity consumption used by customers to help optimize the power provided by the company.

The flow in this paper will be written as follows. The first part describes the background of the problem, gaps in the research, and the purpose. The second part describes a literature review on customer segmentation carried out in previous studies. Section 3 describes the research method to be carried out. Section 4 will explain the results and discussion. Section 5 will present the conclusions, implications, current limitations, and future research.

# Literature Review

esents focuses on customer segmentation / customer credentials As shown, we categorize related articles based on its business context, dataset, segmentation features, and the segmentation method.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Article |  |  |  | Segmentation  Method |
| [14] | Restaurant | Fcustomer data |  |  |
| [15] | from Malang | A |  |  |
| [4] | in Online Retail from UK |  |  |  |
| [7] | Supermarket | Data set period |  |  |
| [16] | Pharmaceutical marketing in Palembang, Indonesia | Data sed |  |  |
| [13] | Marketing | Data set |  |  |
| [5] | in Ireland |  |  |  |
| [11] | Small Medium Enterprise | The sale of electric pulses period January 1, 2016, to December 31, 2017 | Name Customer, Average of transaction/ week, Payment System | K-Means Clustering |
| [17] |  |  |  |  |
| [18] |  |  |  |  |
| [19] | Simulation of DigSILENT PowerFactory and PMU Nordic Power System | Simulated data from The Kundur and PMU data | - | K-Means and Principal Component Analysis (PCA) |

Previous studies in customer segmentation have explored various dimensions of customer clustering problems [6]– [19]. Many of them use the marketing context as a case study. The K-Means clustering model explores customer grouping by considering the specified product preferences and predicting customer behavior in buying products offered by the company[7].

A study of the marketing context in a restaurant assigns each customer attribute as a dimension and gives each customer as a particle using customer feedback data. This study aims to predict customer satisfaction by mapping customers based on the type of food ordered and food reviews processed using a combined model of K-Means Clustering and PL based Algorithm; the results are used to help companies increase buyers coming to the restaurant [14]. This is also the same in research in Online Retail and the K-Means clustering model by using historical customer data to predict customers will repurchase the company's products by grouping customers based on the amount in buying goods sold [4]. Another study on tourism uses google review data to classify tourists based on the places they visit, which is processed using a combination of methods between K-Means and DBSCAN; this study aims to provide recommendations to tourism providers to improve their services which automatically increase tourist visits [18]. Another transportation survey uses the K-Means Clustering model and the KLFRM model to group customers [17] with the same research objective [[4]]. It also has a similar purpose and model [4] with research on marketing at the Telecom Company [13]. Still, they do not use the KLFRM model but use the Neural Network to classify priority customers after getting the results from clustering.

Another study with a supermarket marketing context [7] with the same objectives and predictor variables [14]uses historical customer data processed by a combination of RFM models to determine the selection of potential data validated by AEF. Then K-means clustering model to map Customers based on the same characteristics [4] is then classified to distinguish potential customers for buyback and then validated using the WARD method. This study uses data from All pulse server operators AR-Pulsabiz ​​Malang, Indonesia, to predict the future of Small Medium enterprises. The potential number of customers who will be the operator by using a combination of the K-Means Clustering model and the LRFM model to group customers to provide services according to priority [15]. Research in pharmaceutical marketing [16]also has the same goal[15], but they use eight validation methods in determining the correct number of groupings.

[11]groups [8]– [10]

A context study of electricity consumption [5] using experimental data by installing 4000 intelligent meters in several homes in Ireland with existing methods used to classify household electricity use, in general, can be divided into four categories, statistics, manipulation, time series, and clustering. Statistical methods have been widely used in the unregulated power market to form a standard load PC. A typical load PC is used for settlement purposes and estimates the amount and Time of Use of electricity used. A series of PCs are manufactured for different market segments (e.g., residential, commercial, industrial) and derived on an average for all customers within a customer class. The study of data simulation[19], data that will be processed by Principal Component Analysis (PCA) to obtain the predictor variables with the most potential to be grouped using the K-Means model.



Based on the literature, previous research has done chiefly customer segmentation. The context is more towards marketing and mapping future customer behavior because it affects the marketing strategy. Previous researchers rarely used customer segmentation techniques in electricity consumption by grouping with validation of the number of clusters using the Elbow Method. From the Principal Component Analysis (PCA) method, previous studies used this method to optimize the K-Means model. Still, in the concept of clustering for customer segmentation, no one has used it, especially in electricity consumption. In this study, clustering was carried out using the K-Means Clustering method to validate the number of clusters using the Elbow method and classify customers based on the electricity consumption used.

# Method

Figure 1 shows the framework in this study. The framework is adapted from standard methods for building predictive analytical models[21]. There are five stages: data collection, data cleaning, selecting relevant predictor variables, determining potential predictive methods, evaluating, validating, choosing the best predictive model, and finally reporting the research results.



Figure 1 Research Framework

**3.1 Data Collection**

In this study, we used data from PT. PLN Persero. The data taken is only 1 area because the fields for each region are the same. The data taken by PLN is the West Sumatra zone. Our research uses customer transaction data from January 2019 to December 2020, consisting of 16,504,228 and 107 data variables. Table 2 describes the data that has been taken from 2 years. Some records were removed from the data set because they showed illogical conclusions, i.e., duplicate records or missing values.

Table 2 Descriptive Statistics of The Datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **Year** | **Row** | **Variable** |
| Customer Transactions history | 2019 | 7,945,689 | 107 |
| Customer Transactions history | 2020 | 8,558,539 | 107 |

* 1. **Data Cleaning**

Customer transaction data collected for two years will be selected based on the potential for prediction. The data has 107 variables and 16,504,228 rows. The first data cleaning removes variables that do not have data variations so that it becomes 49 variables and 16,504,228 rows. The second data cleaning removes variables that do not affect prediction so that it becomes 31 variables and 16,504,228 rows. The third data cleaning removes variables with a bit of variation in data to become 18 variables and 16,504,228 rows.

The four data cleaning variables were chosen with the condition of the customer service area in Padang because of the high number of customers in this area than in other areas. They then chose a group focused on general customers. There were 15 variables and 1,187,934 data. The fifth data cleaning due to data outliers was selected to focus on business customers. This has the potential to be predicted to have exciting data variants and discard two variables due to slight variations so that it becomes 13 variables and 508,934 rows. Table 3 shows the process of data cleaning, and Table 4 shows the results of data cleaning.

Table 3 The Process of Data Cleaning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Period | Filter | | | | |
| Variable | Rates  (Variable) | Group Code  (Variable) | Customer Service Area  (Variable) | Row |
| January 2019 - December 2020 | 49 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 31 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 27 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 18 | - | - | - | 16,504,228 |
| January 2019 - December 2020 | 15 | - | 0 | Padang | 1,187,934 |
| January 2019 - December 2020 | 13 | Business | 0 | Padang | 508,934 |

Table 4 The Result of Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Numeric/Nominal | Data Type | Variable Description |
| ID Customer | Nominal | Integer | Identity of the customer |
| Customer Service Unit | Factor | String | Customer Service Units or service branches provided by the company which are located in 4 customer service centers namely Belanti, Painan, Indarung, Pariaman, Lubuk Basung, Lubuk Sikaping, Koto tuo, Baso, Sijunjung, Sungai Rumbai, Kayu Aro, Sawah Lunto, Batusangkar, Lintau, Lima Puluh Kota and others |
| Data Entry Date |  | Date | Admin enters data per 1 month |
| Rates | Factor | String | B1 means a business that uses electricity from 450 kWh to 5500 kWh, B2 means a business that uses electricity from 6600 to 200 thousand kWh, B3 means a business that uses 200 thousand kwh of electrical power and above |
| Power | Factor | Integer | Power used by customers such as 450 kwh,900 kwh,1300 kwh, 2200 kwh,3300 kwh, 7700 kwh,15400 kwh,132000 kwh,110000 kwh and others |
| Meter Code | Factor | String | M means analogue meter and E means digital meter |
| Flash time |  | Double | Electricity usage time by customer |
| Total KWH |  | Integer | The total of peak load kwh usage and peak external load kwh used by customers |
| KWH Off - Load |  | Integer | KWH used at peak external load by customers |
| KWH Peak Load |  | Integer | KWH used at peak load by customers |
| Discount |  | Double | Discounts given by the company based on the provisions of the company such as using unused kwh by the company or because of a natural disaster |
| Peak Offload Fee |  | Double | Payments made when using Peak Offload |
| Peak Load Fee |  | Double | Payments made when using Peak Load |
| Total Cost |  | Double | The total cost paid by the customer |

* 1. **Choice of Variable**

Power, Meter Code, Flash time, Total KWH, Discount, Peak Offload Fee, Peak Load Fee, Total Cost because this predictor variable is not the one that has the potential to be included in the clustering model. The Power variable shows the electrical power installed by the customer. The KWH Off-Load variable shows customers using electricity from 5 pm to 6 am. The KWH Peak Load variable shows indicators of customers using electricity from 6 am to 5 pm. We grouped based on these three variables to determine the number of customers using peak load times and off-peak load times. Table 5 shows detailed information about the predictor variables.

Table 5 The Result of Data Cleaning

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Numeric/Nominal | Data Type | Variable Description |
| Power | Factor | Integer | Power used by customers such as 450 kwh,900 kwh,1300 kwh, 2200 kwh,3300 kwh, 7700 kwh,15400 kwh,132000 kwh,110000 kwh and others |
| KWH Off - Load |  | Integer | KWH used at peak external load by customers |
| KWH Peak Load |  | Integer | KWH used at peak load by customers |

* 1. **Choice of Potential Method**

This study aims to develop a prediction model with customer segmentation or clustering that can provide accurate predictions of customers who have the potential to use electricity consumption. However, this research still examines the clustering model and its ease of implementation. We use the K-Means Clustering model to group customers.

**3.4.1 K-Means**

Commonly, K-means is one of the well-known unsupervised learning techniques for cluster analysis[6]. Cluster analysis is used to aggregate or divide the data set into several clusters according to the similarity value. The situation in this model is used because this algorithm has simplicity and ease of use, and users can determine the number of clusters themselves. This number of clusters (k) needs to be determined by validation[16]. It starts with randomly generated centroids and iteratively computes new centroids to converge to the last group. The steps in the k-means model are explained as follows[11].

Step 1: Determine the number of clusters with validation

Step 2: Each data point in the data set will be assigned to the nearest centroid, and then a

new centroid is generated.

Step 3: To recalculate a new cluster by assigning all data points to the nearest centroid,

and then a new group is created.

Step 4: The process will be repeated between step 2 and step 3 until the stopping criteria

are met.

**3.4.2 Principal Component Analysis**

Principal Component Analysis is a dimension reduction by considering the variance in the time series into several principal components (PC). Principal components describe the parts of the variation of each element [8], [9], [19].

Variables will be separated based on clusters[19]. The situation in this model is used to determine variables that can be implemented in the predicted model. It begins by entering a predetermined variable, and then processing will be carried out randomly, producing several main components (PC)[9] and iteratively calculating the value of these variables. These results will be validated with the expected value of Principal Variant Explained (PVE). The steps in the principal model are described as follows[8], [9].

**Phase 1**: Variables will be divided based on the main components that have been determined, then variables eligible for NA will be discarded. This process will repeat until the criteria stop[8].

**Phase 2**: The previously selected variable will be set based on the Principal Variant Explained (PVE) value, then this process will be repeated until the criteria stop[9].

* 1. **Evaluation and Validation**

Evaluated the prediction model is performance described earlier, we use the elbow method[22], which is a method used to determine the optimal number of clusters, by looking at the percentage comparison of the number of clusters that will form an angle on the curve[15]– [18]. This method is used in cluster analysis to interpret and perform the correct number of clusters by looking at the SSE value. If the value of the first cluster with the weight of the second cluster forms an angle on the curve or the most significant decreasing value, the cluster value is the best[25]. This method is a visual method that looks at the total intra-cluster variation or the total Within-Clusters Sum of Squares (WSS) function of the number of clusters[26]. The larger the number of clusters k, the smaller the WSES value or vice versa. In this study, determine the best number of clusters[10], [15], [24], [26] By adding a Principal Component Analysis (PCA) model that uses robust dimensionality reduction, which allows the variance in a set of time series to be decomposed into several orthogonal Principal Components (PCs) that explain part of the variance.[9].

* 1. **Reporting**

The number of clusters (k) will be determined from the Elbow Method prediction score. The model will be compared with Principal Component Analysis to determine the best predictive model to be selected and used to assist decision-makers in choosing the correct number of customer groupings in a better way for ranking customers based on processed results.

# Result and Discussion

The research aims to break down the behavior of electricity-using customers into 2, namely customers who use electricity at peak load times and customers who use electricity during peak off-load times by using an unsupervised machine learning model. The model creates a model automatically from the training data set. A non-learning algorithm tries to identify and build patterns that can be grouped based on pre-selected predictor variables. Based on the design created, the first step is to determine the number of clustering using the elbow method to get the best number of clusters (k). Figure 1 shows the number of clusters based on the results of the predictor variables previously described using the Elbow method. The correct number of clusters is determined by looking at the line graph when it slopes. From Figure 1, the chart begins to drop at point 3 and point 4.

Chart, line chart, scatter chart

Description automatically generated

Figure 1 The Number of clusters of K

The analysis was carried out using the Elbow Method at point 3 and point 4 using the K-Means clustering model. The best usage grouping in the electricity consumption sector is at point 3. Still, the analysis results show that at point 4, there are outliers (groups at the dark green point) in the distribution. The study of k-means effects in Figure 2 and Figure 3 can be seen below without using Principal Component Analysis (PCA).

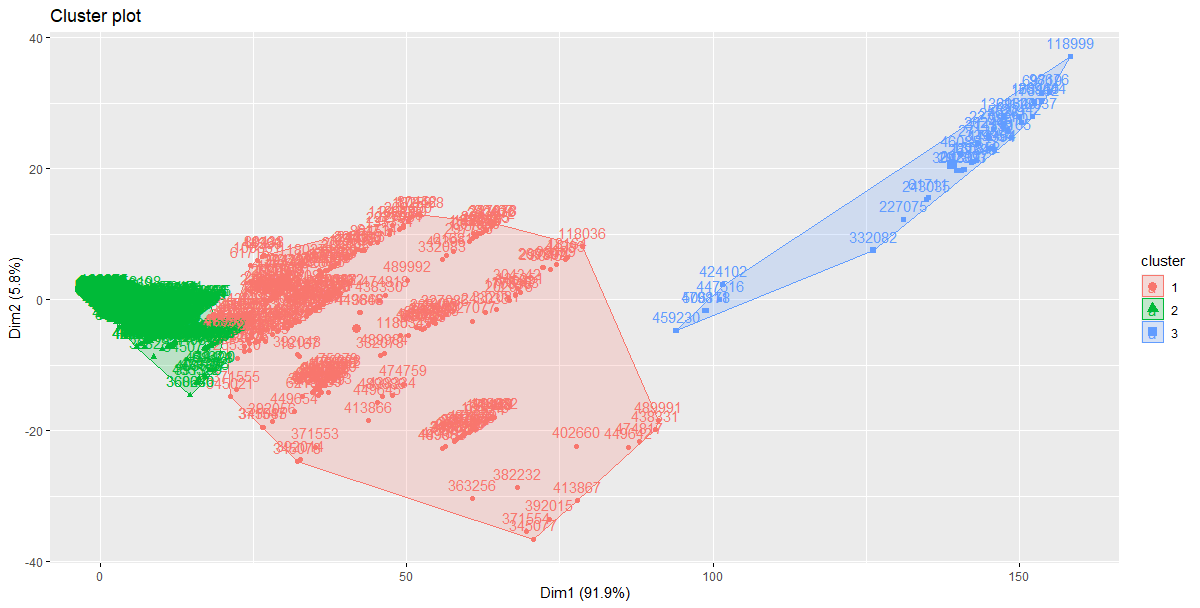


Figure 2 Cluster result of k = 3

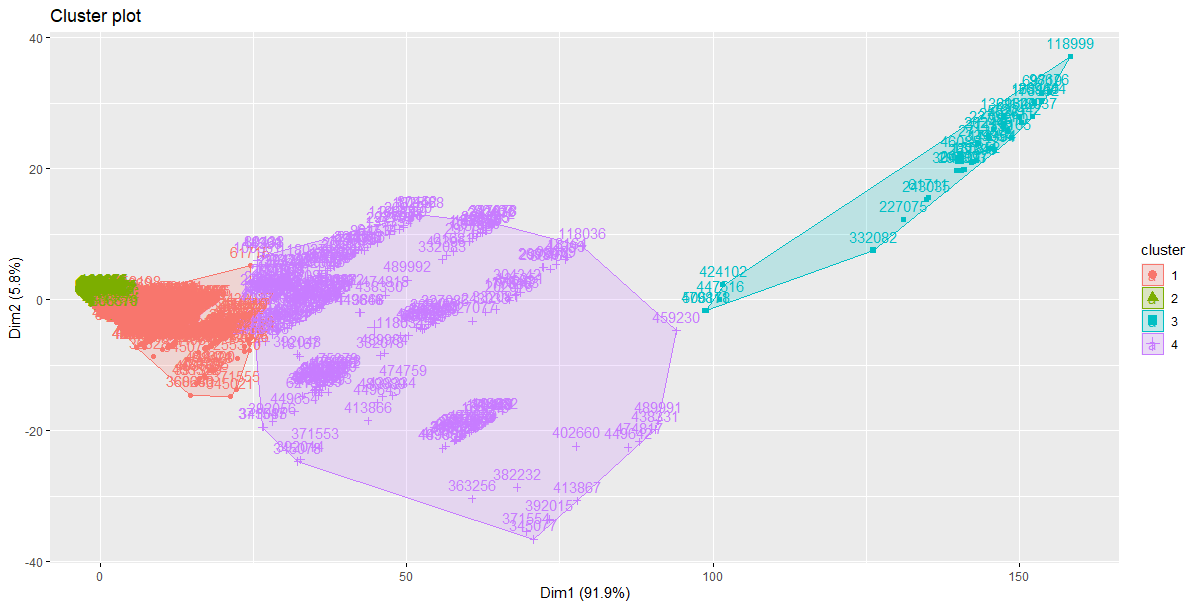


Figure 3 Cluster result of k = 4

The next stage uses the Principal Component Analysis (PCA) model. This analysis is done by reducing the dimensions of the variables that can be applied in the clustering algorithm. PCA will eliminate variables that are impossible to predict. The variables will be broken down using PCA into four Principal Components to be analyzed to see their suitability in the K-Means Algorithm. From this, it can be seen that PC1 and PC2 are the best values in the Principal Component but choose PC1 to determine the variable because they have a variance of about 73%. Table 7 describes the selection of variables that have been carried out by PCA, which can be seen below in the first phase.

Table 7 Possible of Variable with PCA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **PC1** | **PC2** | **PC3** | **PC4** | **Possible** |
| ID Customer | NA | NA | NA | NA | No |
| Customer Service Unit | NA | NA | NA | NA | No |
| Data Entry Date | NA | NA | NA | NA | No |
| Rates | NA | NA | NA | NA | No |
| Power | 3,572 | 0,0554722 | -0,032 | 0,4349 | Yes |
| Meter Code | NA | NA | NA | NA | No |
| Flash time | 1,3462 | 0,8199 | -0,5679 | 0,0238 | Yes |
| Total KWH | 3,8892 | 0,0065 | -0,0045 | 0,0234 | Yes |
| KWH Off - Load | 3,8625 | 0,0567 | 0,0075 | 0,1807 | Yes |
| KWH Peak Load | 3,7234 | 0,5178 | -0,8223 | -0,0567 | Yes |
| Discount | 3,7498 | 0,5678 | -0,8654 | -0,0102 | Yes |
| Peak Offload Fee | 3,7876 | -0,0077 | 0,0057 | 0,3678 | Yes |
| Peak Load Fee | 3,7346 | 0,1101 | -0,0087 | -0,5551 | Yes |
| Total Cost | 3,8771 | -0,0266 | 0,0018 | 0,1150 | Yes |
| Total Variant | 73 % | 11% | 11% | 2% |  |

Based on the results from Table 7, the variables with the value of NA are discarded. The variables that have the matter will be selected in the second stage of the Main Component Analysis. Variables will be chosen by validation from PVE; after that, the variable will be set based on the PVE expected value of about 0.5 and above. The following Table 8 will explain the results of the PVE of the variables that have been selected in the first phase.

**Table 8 The Result Variable With PCA**

|  |  |
| --- | --- |
| **Variable** | **Value** |
| Power | 0,7245 |
| Flash time | 0,5876 |
| Total KWH | 0,5876 |
| KWH Off - Load | 0,6545 |
| KWH Peak Load | 0,6345 |
| Discount | 0,5012 |
| Peak Offload Fee | 0,5321 |
| Peak Load Fee | 0,5234 |
| Total Cost | 0,6234 |

Based on the results of table 8 above, it can be concluded that the nine variables are all possible to be predictors in the clustering model. Therefore, the nine variables will be combined with each variable into the K-means algorithm to get the highest value which can be seen in table 9.

**Table 9 The Combine Variable**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P | FT | TK | POL | PL | POL F | PL F | TC | D | DIM 1 | DIM2 | TV |
| v | v | x | v | v | x | x | x | x | 69,2% | 25,10% | 94,30% |
| v | v | x | v | v | v | v | v | x | 79,70% | 14,30% | 94,00% |
| v | v | x | v | v | v | v | x | v | 65,70% | 14,40% | 80,10% |
| v | v | x | v | v | v | v | v | v | 69,70% | 12,60% | 82,30% |
| v | v | v | x | x | x | x | x | v | 47,30% | 25,10% | 72,40% |
| v | v | v | x | x | x | x | v | v | 57,10% | 20,10% | 77,20% |
| v | v | v | x | x | x | x | v | x | 71,40% | 25,10% | 96,50% |
| v | x | x | v | v | v | v | x | x | 92,50% | 5,10% | 97,60% |
| v | x | x | v | v | x | x | v | x | 93,20% | 4,40% | 97,60% |

*Desc: P: Power, FT: Flash Time, TC: Total KWH, POL: Peak Off Load, PL: Peak Load, POLF: Peak Off Load Fee, PLF: Peak Load Fee, TC: Total Cost, D: Discount, DIM1: Dimension1, DIM2:Dimension2,TV:Total Variant*

Based on the results obtained from table 9, the selected variable is based on the high value of the data variant. The two highest values have the same value, 97.6 per cent. Therefore, choosing this variable is seen from the variance of the first dimension and the second dimension. The higher the number in dimension 1, the better for the conflict in PCA but not for optimizing the number of cluster clusters. Based on this explanation, the selected variables are Power, Peak Off-Load, Peak Load, Total Cost.

The selected variables will be validated with the elbow method to get the best number of groupings. From Figure 4, the best group is at point 3 and point 4 because it is the meeting point when the line is sloping, and the other side of the line starts to rise, but this is not much different from the analysis with K-Means without PCA.

Chart, line chart, scatter chart

Description automatically generated

Figure 4 The Number of clusters of K

Analysis of the K-Mean model using PCA was carried out with validation from the Elbow method at points 3 and 4. The best usage grouping in the electricity consumption sector is at point 3. However, the analysis results are almost identical to the K-Means analysis without PCA. The study of the k-means effect in Figure 5 can be seen below using models from Principal Component Analysis (PCA) and K- Means Clustering.

Graphical user interface, chart, scatter chart

Description automatically generated

Figure 5 Cluster result of k = 3

According to the grouping of business customers from the K-Means analysis without PCA and the K-Means analysis with PCA, it can be concluded that the K-Means analysis without PCA is the best because the data variance is around 97.7% different from the K-Means analysis with PCA, which has a data variance of approx. 97.6%. Even though the data variance on K-Means with PCA has dimension 1 of about 93.2% and dimension 2 of about 4.4%, the results of grouping customers on K-Means without PCA are the best. The following table 9 will compare the best models for business customer segmentation.

**Table 9 The Combine Variable**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Result | | |
|  | Dimension 1 | Dimension 2 | Total Variant Data |
| K-Means without PCA | 91,9% | 5,8% | 97,7% |
| K-Means with PCA | 93,2% | 4,4% | 97,6% |

Based on the results of the analysis using K-Means without PCA and K-Means with the Principal Component Analysis (PCA) method, the correct number of clusters or groupings based on the predictor variables that have been previously selected are three groups using K-Means without PCA with Elbow method validation.K-Means with Principal Component Analysis (PCA) method causes the distribution of clusters formed from the K-Means Clustering model to accumulate at a point, causing the distribution to become ambiguous and make segmentation unclear.

Based on the clustering results, three different customer groups were found, as shown in table 10. The first group represents 937,837 total powers used using total electricity consumption at peak load of 27,827 kWh and total electricity consumption when peak off-load is 115,194 kWh with customers using installed capacity above 10,600 kWh.

The second group describes as much as 4,260 full powers used using total electricity consumption at peak load of 35 kWh and total electricity consumption at peak load of 544 kWh with customers using installed capacity between 450 kWh to 10,600 kWh. The third group describes as much as 2,226,351 full powers used using total electricity consumption at peak load of 123,297 kWh and total electricity consumption at peak load time of 390,803 kWh with customers using installed capacity above 200,000 kWh.

Table 10 The Result of Clustering

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Number of**  **Customer** | **Total**  **Power** | **KWH Peak Off Load** | **KWH Peak Load** | **Installed Power** |
| 1 | 282 | 937,837 kwh | 115,194 kwh | 27,827 kwh | (11,000 - 200,000 )kwh |
| 2 | 508615 | 4,260 kwh | 544 kwh | 35 kwh | (450- 10600) kwh |
| 3 | 37 | 2,226,351 kwh | 390,803 kwh | 123,297 kwh | >200,000 kwh |

Based on table 7, which was previously explained, many customers use electricity at off-peak load times rather than peak load times; therefore, each cluster is analyzed, described as follows.

1. Cluster 1

Cluster 1 consists of 282 business customers using power between 11,000 kWh to 200,000 kWh of energy with a total operating capacity of 937,000 kWh in 1 month. This class reflects the tendency to use electricity at peak off-load times rather than peak loads, thus indicating a too large gap. This shows peak external load usage between 30,000 kWh to 230,000 kWh but peak load usage between 0 to 75,000 kWh. Therefore, the use of electricity outside the peak load that is too high results in suboptimal electricity consumption because the peak load electricity consumption is too little. Figure 7 shows the electricity usage of a business customer.

Chart, box and whisker chart

Description automatically generated

Figure 7 The results from cluster 1

1. Cluster 2

Cluster 2 illustrates the high use of electricity when outside the peak load compared to the peak load, which is used only 1% of the total capable power for the peak load provided each month. This class reflects that business customers use power from 450 kWh to 10,600 kWh. Therefore, this class is like household customers because it has in common some customers use peak load electricity and use electricity outside the peak load, which is 0 kWh. Customers in this class are likely the building has not been occupied for a long time. Figure 8 shows the results of cluster 2.

Chart

Description automatically generated with medium confidence

Figure 8 The results from cluster 2

1. Cluster 3

Cluster 3 describes the highest electricity usage compared to other classes in one month. This class shows peak external load electricity use, which is still high, but peak load electricity consumption is higher than the previous classes. However, this class of business customers mainly uses peak load electricity. Therefore, the possibility of customers in this class continuing to use peak load electricity in the future will be higher, as shown in Figure 9, which is the result of cluster 3.

Chart, box and whisker chart

Description automatically generated

Figure 9 The results from cluster 3

The electrical load described above is characterized by peak load electricity, peak external load electricity, and installed power. The above analysis results found that business customers used a lot of electricity outside the dominant peak load from the three groups above. Business customers dominate electricity usage at peak load in cluster 3. Based on the problems described in section 1, it was found that the high peak load usage was caused by customers using electricity above 200,000 kWh, causing frequent blackouts due to the power being capable of peak loads. This is limited, and this is comparable to the results that have been done previously using the K-means Clustering model.

Based on the analysis results done previously, three customer groupings will be ranked for different actions for each customer. The first group is ranked second because this business customer will potentially one-day use peak loads. Activities that will be possible with the company will educate customers so that the power used does not exceed what is provided. The second group is in the third rank because this business customer uses almost the same electricity as a household. Actions that will be possible are done by giving bonuses to increase electricity consumption. The third group is in the first rank, and this is because these business customers have the potential to increase the company's revenue; therefore, it is vital to maintain these customers by providing education on the use of balanced peak and off-peak loads or giving bonuses to business customers so that they use a balanced electricity load balanced. Table 8 describes the ranking of business customers based on the model results carried out.

Table 12 Business Customer Ranking

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group** | **Number Of Customers** | **Total**  **Power** | **Ranking** | **Ranking Meaning** |
| 1 | 282 | 937,837 kwh | 2 | Important Development-Customer |
| 2 | 508615 | 4,260 kwh | 3 | General Customer |
| 3 | 37 | 2,226,351 kwh | 1 | Important Maintain-customer |

The clustering model detailed in this paper can be applied to any intelligent measurement data set. However, depending on the electrical load in the use of electricity customers, the number of clusters can vary. Finally, a balance is sought in this research paper between peak load and off-peak power consumption, reflecting different ways of using electricity depending on customer usage.

# Conclusion

This paper presents a method of grouping based on the electricity load consumed by breaking down customers who use peak load electricity and off-peak loads for electricity consumption in ​​West Sumatra, Indonesia. The clustering method is applied by the K-Means Clustering model with validation of the number of groupings using the elbow method and the Principal Component Analysis (PCA) method to group data into different electricity usage patterns for each customer. The Elbow method proved to be the most suitable for validation and was used to define customer clustering before performing actions in the clustering model. This model can reflect customer behavior towards consuming the consumed electricity load. In most cases, individual customer characteristics show a positive or negative relationship, with each class showing various patterns of electricity load consumption. As a result, it's possible to group customers and uses electricity based on the electricity consumed.

This finding can inform companies that by grouping customers based on the characteristics of customers using electrical loads, they can improve their predictions that the electrical loads used can be more optimal based on the power that has been provided. In terms of ranked groupings, this research can also help companies act according to the findings that have been made. In terms of the contribution of the literature, this study presents a predictive model using segmentation or customer grouping based on the consumption of electricity used by business customers in the context of a power company. This study only focuses on business customers because they can increase company revenue and only use k-means clustering to explore other clustering methods in further research.

# References

[1] Katadata, “Konsumsi Listrik Nasional Terus Meningkat,” *www.databook.com*, Jan. 09, 2020. https://databoks.katadata.co.id/datapublish/2020/01/10/konsumsi-listrik-nasional-terus-meningkat (accessed Jan. 04, 2022).

[2] Yusrizal, “Sering Mati Lampu, Legislator Minta PLN Segera Bertindak Di Agam,” *Antara Sumbar*, Jun. 17, 2017. https://sumbar.antaranews.com/berita/206333/sering-mati-lampu-legislator-minta-pln-segera-bertindak-di-agam (accessed Jan. 15, 2022).

[3] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, “Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model,” *International Journal of Integrated Engineering*, vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[4] J. Ye, “Analysis on E-commerce Order Cancellations Using Market Segmentation Approach,” in *ACM International Conference Proceeding Series*, Jan. 2021, pp. 33–40. doi: 10.1145/3450588.3450596.

[5] F. McLoughlin, A. Duffy, and M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data,” *Applied Energy*, vol. 141, pp. 190–199, Mar. 2015, doi: 10.1016/j.apenergy.2014.12.039.

[6] R. Bapna, P. Goes, A. Gupta, and Y. Jin, “USER HETEROGENEITY AND ITS IMPACT ON ELECTRONIC AUCTION MARKET DESIGN: AN EMPIRICAL EXPLORATION 1,” 2004.

[7] Z. J. Lee, C. Y. Lee, L. Y. Chang, and N. Sano, “Clustering and classification based on distributed automatic feature engineering for customer segmentation,” *Symmetry*, vol. 13, no. 9, Sep. 2021, doi: 10.3390/sym13091557.

[8] H. Haugdal and K. Uhlen, “Mode shape estimation using complex principal component analysis and k-means clustering,” May 2019. doi: 10.1109/SGSMA.2019.8784556.

[9] T. Celik, “Unsupervised change detection in satellite images using principal component analysis and κ-means clustering,” *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 772–776, Oct. 2009, doi: 10.1109/LGRS.2009.2025059.

[10] F. Marisa, S. Sakinah Syed Ahmad, Z. Izzah Mohd Yusof, T. Mohammad Akhriza, W. Purnomowati, and R. Kumar Pandey, “The Analyze of Relationship between Revenue and Customer Payment Methods in Small Medium Enterprise Based on Clustering K-Means,” in *Journal of Physics: Conference Series*, Jul. 2021, vol. 1908, no. 1. doi: 10.1088/1742-6596/1908/1/012021.

[11] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, “Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model,” *International Journal of Integrated Engineering*, vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[12] S. Janardhanan and R. Muthalagu, “Market segmentation for profit maximization using machine learning algorithms,” in *Journal of Physics: Conference Series*, Dec. 2020, vol. 1706, no. 1. doi: 10.1088/1742-6596/1706/1/012160.

[13] F. Abdi and S. Abolmakarem, “Customer Behavior Mining Framework (CBMF) using clustering and classification techniques,” *Journal of Industrial Engineering International*, vol. 15, pp. 1–18, Dec. 2019, doi: 10.1007/s40092-018-0285-3.

[14] M. Grbovic, N. Djuric, and S. Vucetic, “Supervised Clustering of Label Ranking Data.” [Online]. Available: https://epubs.siam.org/page/terms

[15] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, Fachrudin, and T. M. A. Aziz, “Segmentation model of customer lifetime value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM model,” *International Journal of Integrated Engineering*, vol. 11, no. 3, pp. 169–180, 2019, doi: 10.30880/ijie.2019.11.03.018.

[16] R. Gustriansyah, N. Suhandi, and F. Antony, “Clustering optimization in RFM analysis based on k-means,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470–477, 2019, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[17] H. Li, X. Yang, Y. Xia, L. Zheng, G. Yang, and P. Lv, “K-LRFMD: Method of Customer Value Segmentation in Shared Transportation Filed Based on Improved K-means Algorithm,” in *Journal of Physics: Conference Series*, Jul. 2018, vol. 1060, no. 1. doi: 10.1088/1742-6596/1060/1/012012.

[18] M. Chindyana and L. A. Wulandhari, “Segmentation of tourist interest on tourism object categories by comparing PSO K-means and DBSCAN method,” *Revue d’Intelligence Artificielle*, vol. 35, no. 1, pp. 23–37, Feb. 2021, doi: 10.18280/ria.350103.

[19] H. Haugdal and K. Uhlen, “Mode shape estimation using complex principal component analysis and k-means clustering,” May 2019. doi: 10.1109/SGSMA.2019.8784556.

[20] S. L. Lo, R. Chiong, and D. Cornforth, “Ranking of high-value social audiences on Twitter,” *Decision Support Systems*, vol. 85, pp. 34–48, May 2016, doi: 10.1016/j.dss.2016.02.010.

[21] G. Shmueli and O. R. Koppius, “PREDICTIVE ANALYTICS IN INFORMATION SYSTEMS RESEARCH 1.”

[22] P. Bholowalia and A. Kumar, “EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN,” 2014.

[23] C. D. Rumiarti and I. Budi, “SEGMENTASI PELANGGAN PADA CUSTOMER RELATIONSHIP MANAGEMENT DI PERUSAHAAN RITEL: STUDI KASUS PT GRAMEDIA ASRI MEDIA.”

[24] “Customer Segmentation Model in E-commerce Using Clustering Techniques and LRFM Model: The Case of Online Stores in Morocco.” [Online]. Available: https://www.researchgate.net/publication/284899160

[25] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, “Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty,” *Expert Systems with Applications*, vol. 37, no. 7, pp. 5259–5264, Jul. 2010, doi: 10.1016/j.eswa.2009.12.070.

[26] IEEE Power & Energy Society, IEEE Industry Applications Society, and Institute of Electrical and Electronics Engineers, *2018 IEEE PES/IAS PowerAfrica : 28-29 June 2018.*