**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, blackouts often occur in West Sumatra and appear four times a month.

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 specific ones to predict future customer actions or behavior [2]. 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[3][4].

Based on previous research on customer segmentation, electricity consumption is still small. 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 in each group[5]. 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[5][6]. Principal Component Analysis is used to eliminate outliers in the data applied to the clustering model for more accurate grouping [7], [8].

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.

# Literature Review

Previous studies in customer segmentation have explored various dimensions of customer clustering problems [6]– [10]. 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[6].

A study of the marketing context [9]assigns each customer attribute as a dimension and sets each customer as a particle. All customers in the company can form a multidimensional space, defined as the customer attribute space. Mapping relationships between customer attributes and conceptual categories can be constructed using analytical methods or sample learning methods. The analysis method analyzes the attribute character of each conception category that must be possessed, then forms a mapping of the relationship between the attribute space and the conception space. However, many mapping relationships between attribute spaces and conception spaces are unclear, so it is necessary to use sample learning methods to establish mapping relationships. Later this method will also be applied in customer grouping to find customers who use peak load electricity consumption [6]– [12].

A context study of electricity consumption [4] 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 (ToU) 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.

Another study in marketing that combines the LFRM, CLV, and K-Means models explains Customer Lifetime Value (CLV) [10] in each customer segment. The grouping uses the K-Means Clustering method based on the LRFM (Length, Recency, Frequency, Monetary) model. The cluster formation process uses the Elbow and SSE methods with the best clusters = 2 clusters. The CLV value is generated from the multiplication of the LRFM normalization results, and the LFRM weight values ​​are then added up and performed on each group formed. Based on the LRFM matrix, this cluster has a high loyalty value, with the LRFM symbol being a loyal customer (the best segment with a high customer loyalty value). Based on the LRFM symbol, companies can create strategies to retain customers and acquire loyal customers with high profitability. In determining the number of clusters, the elbow method is the best number of groups [8]– [10].

Another study used the Placket-Luce (PL) probabilistic ranking model. Each cluster is represented as a composite of Voronoi cells defined by prototypes and assigned a set of PL label scores that rank the cluster-specific labels. The unknown PL cluster parameters and prototype positions were determined using supervised learning techniques. Cluster membership and ranking for a new instance are determined by its leading members. The proposed algorithm is empirically based on the synthesis of scale and real-life data. The OT-based method is superior to the heuristic-based supervised clustering approach. The proposed PL-based algorithm is also tasked with predicting label rank. The results show that it is highly competitive with ranking algorithms and partially accurate on ranking data [13].

The study of clustering uses a new K-Means technique for unattended change detection in multitemporal satellite imagery using principal component analysis (PCA) and k-means clustering. This study makes the difference in the image partitioned as nonoverlapping, which is arranged to create a space eigenvector. A dimensional feature vector represents each pixel in a different picture. This study detects changes achieved by partitioning the space feature vector into two clusters using the k-means clustering model and then assigning each pixel to one of the two clusters using the minimum Euclidean distance between the pixel feature vectors and the average feature vector of the clusters [7], [8].

Table 1 provides an overview of previous studies that analyzed marketing topics using transaction data or customer history.

Table 1 Reviewed Studies on Customer Segmentation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Article | Business Context | Dataset | Segmentation  Features | Techniques |
| Vucetic et al.,2018 | Marketing | Restaurant customer feedback data period January 1, 2016, to December 31, 2016 | Customer name, type of food ordered, food review, gender | K-Means using PL based algorithm |
| Aziz et al, 2019 | Small Medium Enterprise | SME customers are all pulse server operators AR-Pulsabiz Malang Indonesia period January 1, 2018, to June 30,2018 | ID Customer, Length, Recency, Frequency, Monetary | K-Means Clustering and LRFM Model |
| Ye Jingyi, 2021 | E- Commerce | Online Retail Data Set period 12 January 2010 and 12 September 2011 from UK | Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country, Total Price | K-Means Clustering |
| Sano et al, 2021 | Marketing | Transaction Supermarket data January 1, 2017, to December 31, 2018 | ID Customer, Product Name, Length, Recency, Frequency, Monetary | AEF, RFM, k-means, Ward method, FCM, and the decision tree |
| Antony et al,2019 | RFM Analysis | Sales data of a pharmacy in Palembang period January 2015 until December 2015 | ID Customer, Product Name, Length, Recency, Frequency, Monetary | K-Means Method and eight indexes of validity to determine the optimal number of clusters namely Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Rutkowski Index, Hubert Index, Ball-Hall Index, and Krakowski-Lai Index |
| Puh et al, 2020 | Food Retailing | Questionnaire data consisting of 500 consumers in Croatia in 2020 | Demographic characteristics (Age, Gender, Education, Occupation, Monthly Income in HRK), Product, Frequency, Percentage | Latent Class Model |
| Abdi et al, 2018 | Telecom Company | Customers of a telecom company period January 1,2017 to December 31, 2017 | Socio-demographic attributes (Region, Age, Marital, Address, Income, Education, Employment, Retire, Gender), Behavioral Attributes (Hours of Usage (Longmont, Tollmon, Equipmon, Cardmon, Wiremon), Selected Service (Multiline, Voice, Pager, Internet, Call Id, Call wait, Forward, Confer, Call card, Wireless, Churn)) | K-Means Clustering, Neural network, |
| McLoughlin et al, 2014 | Electricity | Experimental data by installing smart meters to more than 4000 residences in Ireland, period January 1, 2009 to December 31, 2010 | Dwelling type, No. of bedrooms, Age, Social Class, Electronic Type | k-means, k-medoid and Self Organizing Maps (SOM) |
| Li et al, 2012 | Transportation | Historical data from the vehicle sharing platform database at the university detailed data of all customers from November 30, 2015, to November 30, 2017 | User ID, driving mileage, points, discounts and 29 other attributes. The Variable are used User id, current miles, cost, car id | K-Means Method and KLRFMD model |
| Marisa et al, 2019 | 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 |
| Chindyana et al ,2021 | Tourism | Google review rating in 2020 | Id Customer, Gender, Place, Review | K-Means Method and DBSCAN Method |
| Zhao et al,2021 | E-commerce | The customer transaction online customer company in the UK period January 1, 2016, to December 31,2017 | Invoice number, quantity, price, address, and zip code. | K-Means Clustering and RFM Model |

Based on the literature, previous research has done chiefly customer segmentation. The context is more towards marketing and map customer behavior in the future because it affects its marketing strategy. In electricity consumption, previous researchers rarely used customer segmentation techniques in a grouping. In this study, clustering was carried out using the K-Means Clustering method with validation of the number of clusters using the Elbow and Principal Component Analysis (PCA) methods to classify customers based on the consumption of electricity used.

# Method

Figure 1 shows the framework in this study. The framework is adapted from standard methods for building predictive analytical models[15]. 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.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 3 The Process of Data Cleaning

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[5]. 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[10].

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.

* 1. **Evaluation and Validation**

Evaluated the prediction model is performance described earlier, we use the elbow method[17], 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[21]. 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[22]. The larger the number of clusters k, the smaller the WSES value or vice versa. In this study, determine the best number of clusters[9], [18], [20], [22] 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.[8].

* 1. **Model Use and Reporting**

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

# 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

Figure 2 shows the number of clusters based on the principal component analysis (PCA) results used to remove outliers. 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. From Figure 2, the chart starts to decline at point 3 and tells 4.

Chart, line chart, scatter chart

Description automatically generated

Figure 2 The Number of clusters of K with PCA

The analysis was carried out using the Elbow Method at point 3 and point 4 using the K-Means clustering model. Still, using Principal Component Analysis (PCA), the number of clusters is also found at the end 3 points 4. Based on the two validation methods, it is found that without using Principal Component Analysis (PCA) and using Principal Component Analysis (PCA), 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 3 and Figure 4 can be seen below without using Principal Component Analysis (PCA).

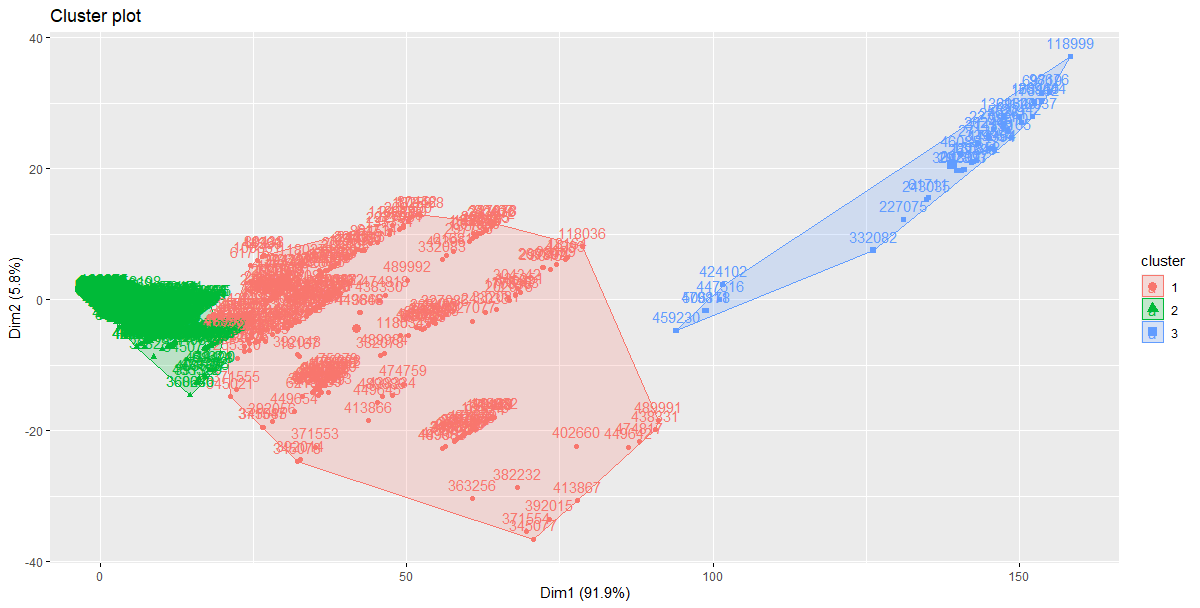


Figure 3 Cluster result of k = 3

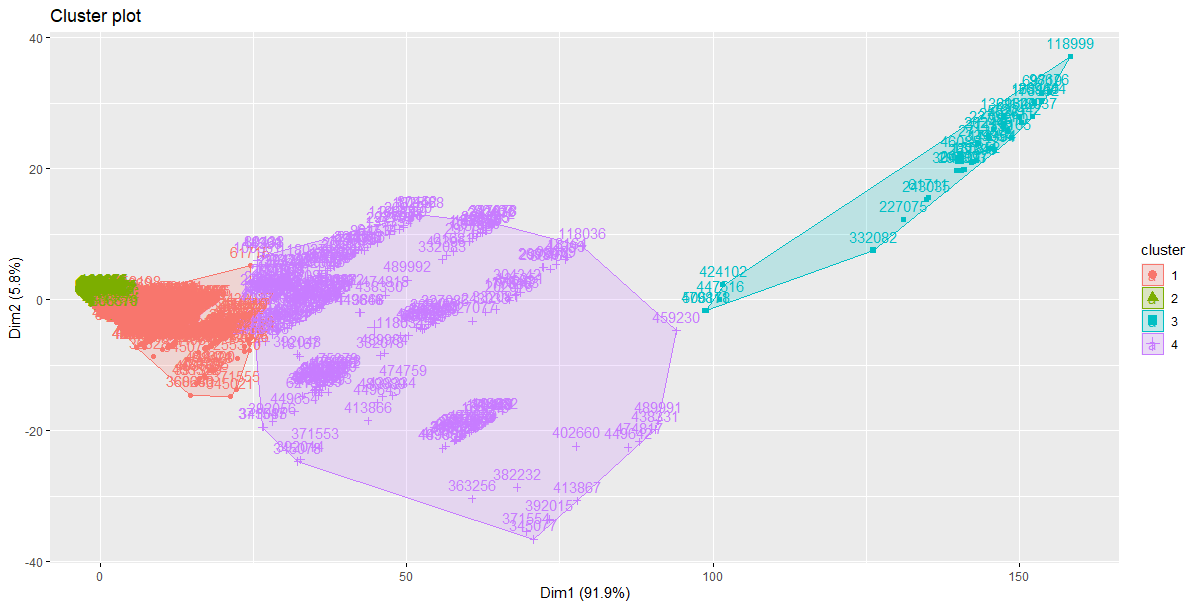


Figure 4 Cluster result of k = 4

Figures 5 and 6 can be seen below using Principal Component Analysis (PCA). Principal Component Analysis (PCA) removes outliers in the data. In figure 5, the results from cluster clustering are distributed in clusters 3 and cluster 2. In figure 6, the results from cluster clustering are distributed in cluster 4 along with cluster 2 and cluster 3.

Chart, scatter chart

Description automatically generated

Figure 5 Cluster result of k = 3 with PCA

A picture containing graphical user interface

Description automatically generated

Figure 6 Cluster result of k = 4 with PCA

Based on the analysis results using the validation of the elbow method and the Principal Component Analysis (PCA) method, it was found that the number of clusters or the correct grouping based on the predictor variables that had been previously selected was three groups using the elbow method, because using the Principal Component Analysis (PCA) method, the distribution of clusters formed from the K-Means Clustering model accumulates at a point, causing the distribution to be ambiguous.

Based on the clustering results, three different customer groups were found, as shown in table 7. The first group represents 937,837 total power 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 power 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 power 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 7 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 8 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] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

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

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

[18] 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.

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

[20] “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

[21] 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.

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