MSBA 434 Project

Prepared for

Maryam Zokaieinikoo

Department of Operations Management

Case Western University

By

Hruthik Kommuru

Dec 8, 2023

Table of Contents

Introduction

This project delves into segmenting regions or time periods based on energy consumption patterns using k-Means Clustering. It centers around an extensive dataset amalgamating energy consumption records, weather data, and regional demographic information, fostering a comprehensive understanding of energy utilization. The primary goal is to employ k-Means Clustering to categorize regions or time segments based on distinctive energy consumption behaviors. This analysis aims to uncover prevalent trends, anomalies, and potential opportunities for enhancing energy efficiency.

Dataset Overview:

The dataset, accessible through the provided link, encapsulates diverse variables:

- Date and Time: Recorded in dd/mm/yyyy and hh:mm:ss formats, respectively.
- Global Active Power: Represents household global minute-averaged active power, measured in kilowatts.
- Global Reactive Power: Signifies household global minute-averaged reactive power, in kilowatts.
- Voltage: Reflects minute-averaged voltage, measured in volts.
- Global Intensity: Indicates household global minute-averaged current intensity, measured in amperes.
- Sub-metering Categories:
    - Sub_metering_1: Measures energy sub-metering No. 1, associated with the kitchen's electrical appliances.
    - Sub_metering_2: Tracks energy sub-metering No. 2, encompassing appliances in the laundry room.
    - Sub_metering_3: Monitors energy sub-metering No. 3, inclusive of an electric water-heater and an air-conditioner.

Objective

1. **Data Preparation: Organize and structure the dataset for clustering analysis.**

2. **Feature Engineering: Develop relevant features that can influence energy consumption.**
3. **Clustering Analysis: Use k-Means to segment data based on energy consumption patterns.**
4. **Interpretation of Clusters: Analyze and understand the characteristics of each cluster.**

Data Preprocessing

1. **Data Preparation**
    - Read the data
    - Checking and removing the missing values in the whole data because the whole of missing values is 5% of the data.
    - Transforming the data and Normalizing (Converting relevant columns to numeric, Checking and removing NAs, Performing normalization)
    - Grouping the date and time
2. **Feature Engineering**
    - Feature Selection and Clustering:
        - Selected specific features related to energy consumption for k-Means clustering.
        - Performed k-Means clustering with 5 clusters to group similar energy consumption patterns.
        - Visualized the clusters using Principal Component Analysis (PCA) for dimensionality reduction and scatter plots to interpret the clusters.
    - Time-Based Feature Extraction:
        - Extracted additional time-based features (day of the week, hour, minute, second, month) from the 'Date' and 'Time' columns.
        - Aggregated energy consumption features on a daily basis for further analysis.
    - Visualization and Analysis:
        - Plotted energy consumption patterns aggregated on a daily basis.

- Analyzed clusters' characteristics through box plots, line plots, and scatter plots based on different energy consumption metrics and time-related variables.
- Identified anomalies and trends in energy consumption patterns across clusters.

- Insights and Interpretation:
  - Examined how different variables contribute to cluster formation.
  - Explored peak and low energy consumption periods within each cluster across various energy-related metrics and times of the day.
  - Analyzed trends in energy consumption over a 24-hour period for different clusters.

Exploratory data analysis

1. **Initial Analysis:**
   - The initial analysis is done by looking into centroids of the clusters.

	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
1	2.4973899	0.6130877	-0.9762228	2.5274699	-0.003981068	5.87320538	0.5089001
2	0.7276376	-0.0282588	-0.3474241	0.7057025	-0.150357958	-0.14411448	1.3081119
3	2.7422704	0.6831128	-1.1331114	2.7679133	5.782846330	0.15914176	0.5684239
4	-0.5891955	-0.4349845	-0.6045173	-0.5853786	-0.174564329	-0.17276887	-0.7160917
5	-0.6670268	-0.5261292	0.9200027	-0.6731978	-0.178587113	-0.18587473	-0.7158892
6	-0.3537287	1.2647093	0.1935682	-0.3149931	-0.162642577	-0.05940834	-0.5767358

CUSTER-1:

- Slightly below-average global active power and notably lower global reactive power.

- Voltage is slightly below the grand mean.
- Global intensity is moderately high, indicating a consistent but not excessive usage.
- Low sub-metering 1 readings suggest low usage of specific appliances that these sub-meters track (like kitchen appliances and laundry).
- High sub-metering 2 indicates higher usage of appliances that could be washing machine, drier or refrigerator systems.

CUSTER-2:

- Lower than average global active power and reactive power, indicating lower overall energy consumption.
- Higher voltage may indicate light loads on the electrical system.
- Very low global intensity and sub-metering values suggest minimal usage of power and small-scale appliances.
- The negative centroid value for sub-metering 3 might suggest these households typically do not have high consumption appliances or they use them infrequently.

CUSTER-3:

- Very high global active power and reactive power, suggesting heavy power usage.
- The lower voltage could be a result of high power demand from these households.
- High global intensity and extremely high sub-metering 1 readings indicate the use of power-intensive appliances, possibly heating or cooling systems.
- High sub-metering 2 and moderate sub-metering 3 values suggest a variety of appliances contribute to the total load, indicating diverse and intensive usage.

CUSTER-4:

- Lower than average global active power and reactive power, similar to Cluster 2, but with a lower voltage.
- Low global intensity and low sub-metering values across the board indicate low overall power usage.
- These households might have overall lower energy consumption and possibly use fewer heavy-duty appliances.

CUSTER-5:

- Below-average global active power and high global reactive power, which might indicate appliances that have higher inefficiency or standby power consumption.
- Slightly above-average voltage and lower intensity suggest moderate usage.
- Sub-metering 1 and 2 are low, and sub-metering 3 is negative, indicating lower consumption from the main heating system or air conditioners.

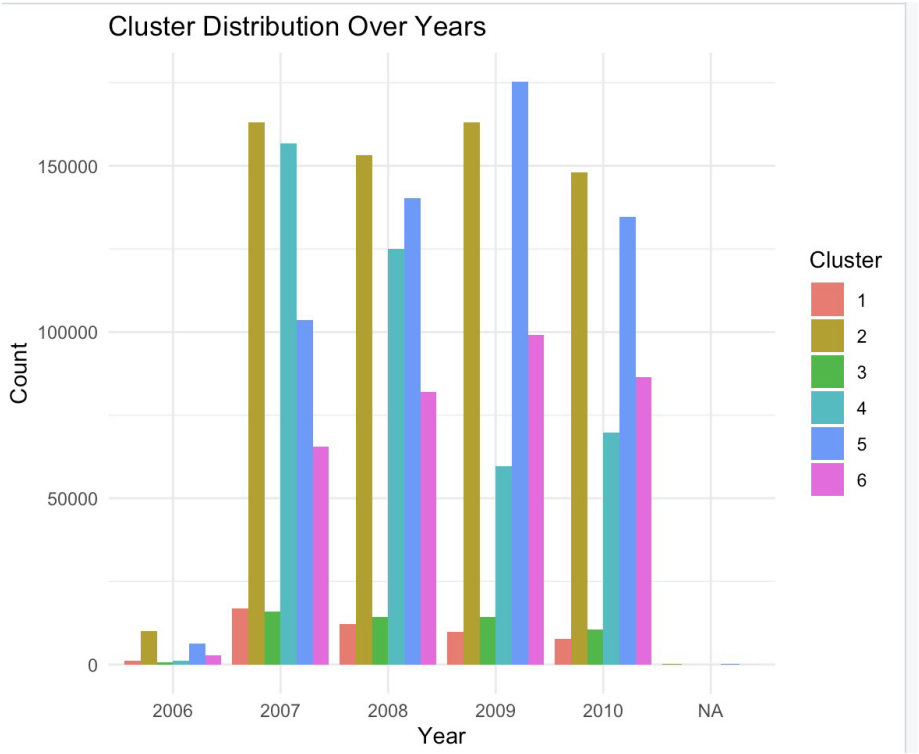
CUSTER-6:

- Above-average global active power and reactive power, indicating higher energy consumption.
- Lower voltage and high global intensity suggest heavy appliance usage, possibly during peak hours.
- Low sub-metering 1 but slightly positive sub-metering 2, alongside very high sub-metering 3 readings, indicate a significant portion of electricity is consumed by appliances or systems not covered by sub-metering 1 or 2, like electric water heaters or space heaters.

Cluster Distribution Over Years

- When looking at the Cluster Distribution Over Years plots, we can see how the prevalence of these clusters changes over time
- Some clusters may increase or decrease, indicating shifts in energy consumption behaviors or

changes in the number of households that fit a particular profile.

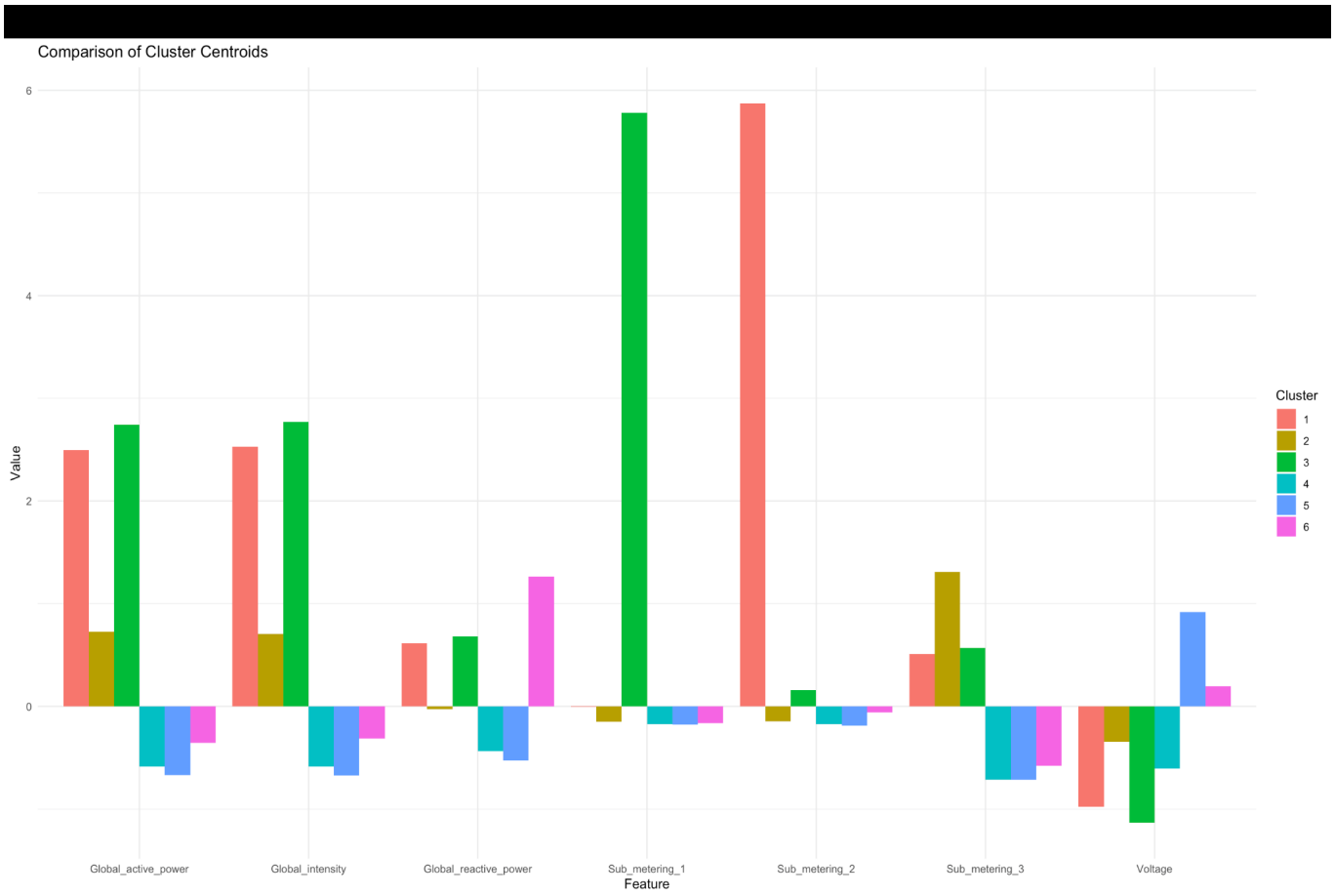


- The evolution of Cluster 2 over the years corresponds to a concurrent rise in the transitions observed in Clusters 5 and 6.
- Cluster 3 over the years from 2007 to 2010 remains the same.

Cluster Centroid Analysis:

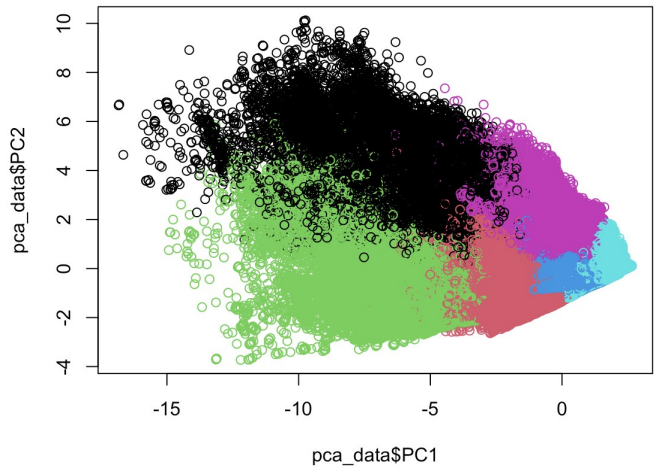
- Analysis based on the cluster centroids for each of the featured variables provides insights on influence and significance of the variable.
- Cluster 3, has highest sub_meetering_1, indicating that the majority of electric consumption corresponds to the kitchen.

- Cluster 1, has highest sub_meetering_2, indicating that the majority of electric consumption corresponds to the laundry room.
- Cluster 2, has highest sub_meetering_3, indicating that the majority of electric consumption corresponds to the water heater and air conditioners..



PCA Plot:

- If the points from different clusters are close to each other, it may suggest similarity in certain features of the clusters.

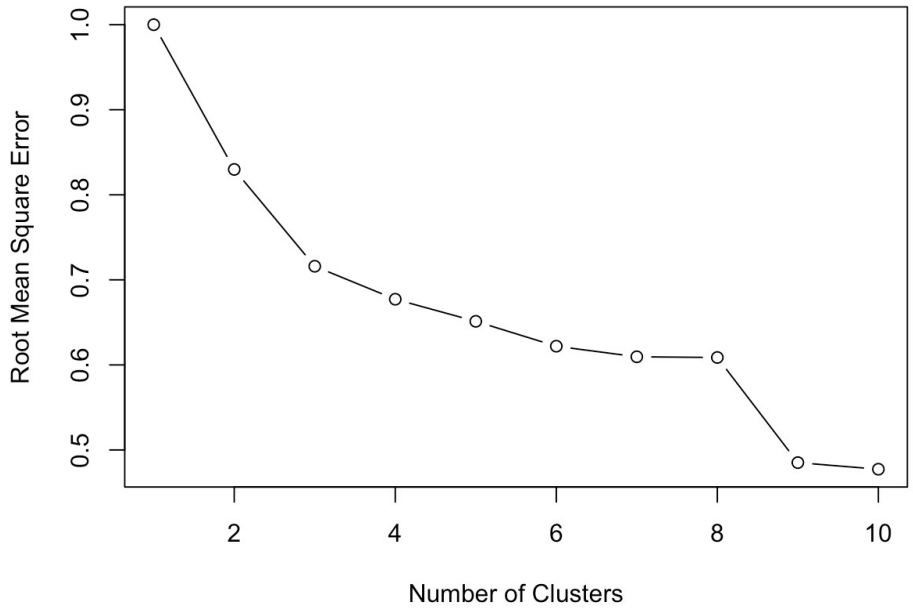


- Cluster 1,3 has some similar overlapping features, while cluster 2, has overlapping features with cluster 3,4.

Model Building and Evaluation

1. **Data Preparation:**
    - The first step involves preparing the data, this involves structuring the given data format into desirable form in order to perform visualization
    - The next step is to check for the missing values and remove them accordingly.
    - Transformation of relevant columns to numeric data type
    - Normalizing the data to ensure consistency and aid in model convergence.

2. **Model Building:**
    - The K-means clustering model is built by identifying the optimal number of clusters using a technique called the elbow method.
    - The optimal K value, 6 is selected for the given model.



    - Using the optimal number of clusters determined in the previous step, to build the K-means model.
3. **Parameters :**
    - The following parameters are considered while building the model. "Global_active_power", "Global_reactive_power", "Voltage", "Global_intensity", "Sub_metering_1", "Sub_metering_2", "Sub_metering_3"
    - Since K-Means is sensitive to the scale of features, all of these parameters are normalized.
4. **Outliers:**
    - Outliers can significantly impact the performance of K-means clustering.
    - The outliers were identified and removed using IQR method.

Conclusion and Insights:-

1. Customer Segmentation
    - Clusters represent distinct groups of households with similar energy consumption patterns.
2. Tailor marketing strategies and service offerings based on the specific needs and behaviors of each cluster. For example, offer targeted energy-saving tips or promotions.

3. Clusters show how energy consumption varies over time. Implement demand-side management strategies to efficiently allocate resources. Adjust energy distribution and supply based on the identified peak usage times for each cluster.

4. Some clusters may have a higher demand for specific resources.
    - Optimize resource allocation by understanding which clusters contribute more to the overall demand. This can guide infrastructure investments and help prevent resource shortages.

5. Grid Planning and Optimization
    - Clusters may have different preferences for energy sources and times of usage.
    - Plan and optimize the power grid infrastructure based on the preferences of each cluster. This can enhance the overall efficiency and sustainability of the energy distribution system.

6. Targeted Energy Efficiency Programs
    - Different clusters exhibit varying levels of energy efficiency.
    - Develop targeted energy efficiency programs for clusters with higher consumption. Provide incentives or educational materials to encourage more efficient energy use.

7. Certain clusters may exhibit distinct patterns of wear and tear on equipment.
    - Implement predictive maintenance strategies by monitoring clusters with higher equipment usage. This can help reduce downtime and improve the reliability of the energy infrastructure.