Power Analysis and Prediction for BEST data using LSTM, K-means, DBSCAN, and GMM

1st Akash Rajanand Alne
CTARA, IITBombay
Indian Institute of Technology Bombay
Mumbai, India
22m0203@iitb.ac.in

2nd Dynaneshwar A Khaire
CTARA, IITBombay
Indian Institute of Technology Bombay
Mumbai, India
22m0206@iitb.ac.in

3rd Panyam Sweeya Goud
Electrical Engineering, IITBombay
Indian Institute of Technology Bombay
Mumbai, India
19d070042@iitb.ac.in

***Abstract*—In this project, we present a comprehensive analysis of power consumption data of BEST using various advanced techniques such as LSTM, K-means, DBSCAN, and GMM. The aim of the project is to provide valuable insights into the power consumption patterns of BEST and to identify any anomalous behavior that may require attention.**

***Index Terms*—K-means, DBSCAN, LSTM, Load data forecasting**

I. Introduction

This report has analysed load power data of BEST(The Brihanmumbai Electricity Supply and Transport Undertaking) region based on weather data which consists of temperature and humidity. Power and weather data are available from 1st Dec 2021 - 31st Dec 2022 (one year and one month). We have used Long Short-Term Memory(LSTM) for supervised, K-means and Density-based spatial clustering of applications with noise(DBSCAN) for un-supervised learning.

II. Dataset

Figures 1,2,3 show the plot of power,temperature,humidity for 15 minute time steps for 1yr plus 1 month data from 2021-12-01 00:00:00 to 2022-12-31 23:45:00.

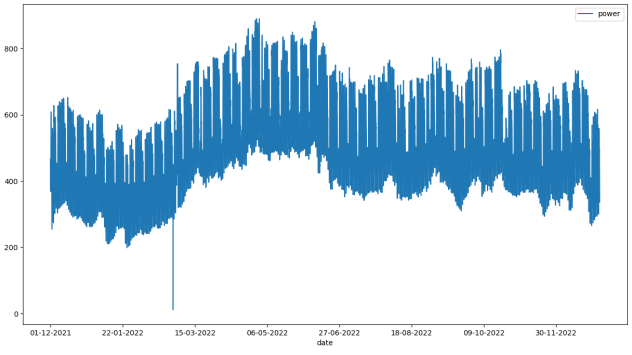

Fig. 1. Power data for 1 year and 1 month in MW sampled at 15 minute interval

Figures 4,5,6 show the box plots of power,temperature and humidity.

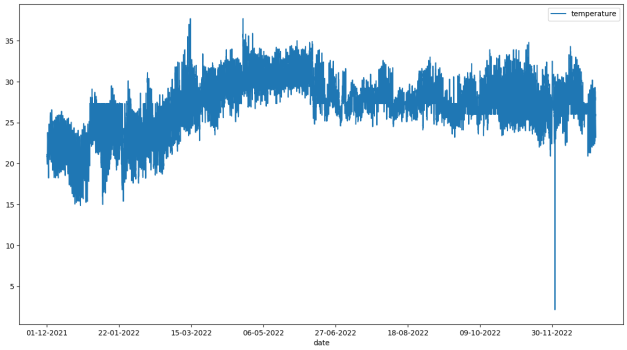

Fig. 2. Temperature data for 1 year and 1 month in °C sampled at 15 minute interval

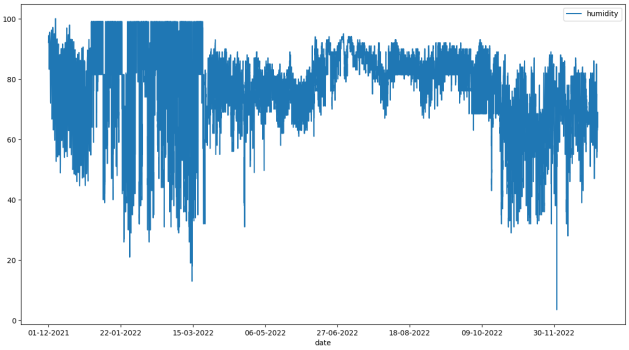

Fig. 3. Humidity data for 1 year and 1 month in % sampled at 15 minute interval

III. Supervised

A. Long Short-Term Memory(LSTM)

This code used is an LSTM (Long Short-Term Memory) neural network model for time series prediction. The model is defined using the Keras Sequential API, which allows for a linear stack of layers to be defined. The LSTM model architecture consists of three LSTM layers, each with 50 units, followed by a Dense layer with a single output unit. The input shape to the LSTM model is specified as (X-train.shape[1], 2), where X-train is the training input data, with shape (number of samples, number of time steps, number of features). Here,

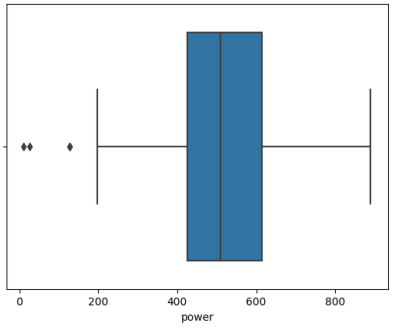

Fig. 4. Box plot of load power

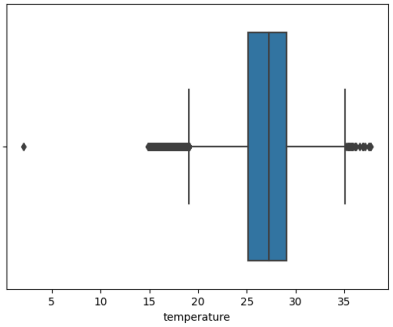

Fig. 5. Box plot of temperature

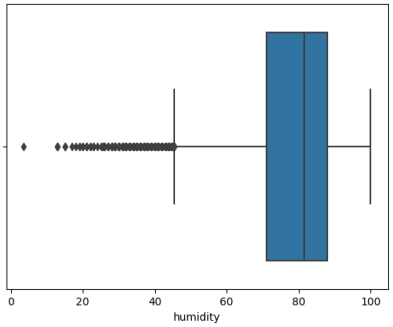

Fig. 6. Box plot of humidity

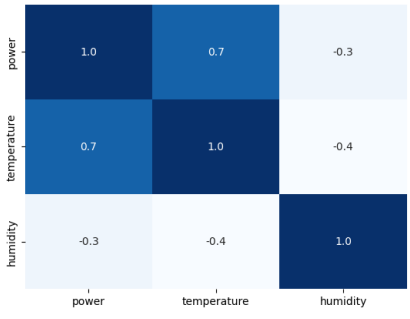

Fig. 7. Correlation

the number of time steps is inferred automatically, while the number of features is set to 2 (temperature and humidity).The 'adam' optimizer is used to minimize the mean squared error loss between the predicted and actual output values.

Model was trained for 1 year i.e, from 2021-12-01 00:00:00 to 2022-11-30 23:45:00 and tested on 2022-12-01 00:00:00 to 2022-12-31 23:45:00. Inputs are temperature and humidity and output is power, which we want to forecast.

Fig 7 shows the correlation among the features(temperature,humidity) and power. Fig 8 has predicted and test power data for the testing period of 1 month. Fig 9 plots error vs sampled frequency. Here the sampled frequencies are varied in multiples of 15 minute and difference of predicted and test data is taken from 15minute...to 15 days. We can observe that the error almost decreases with increased sample time.

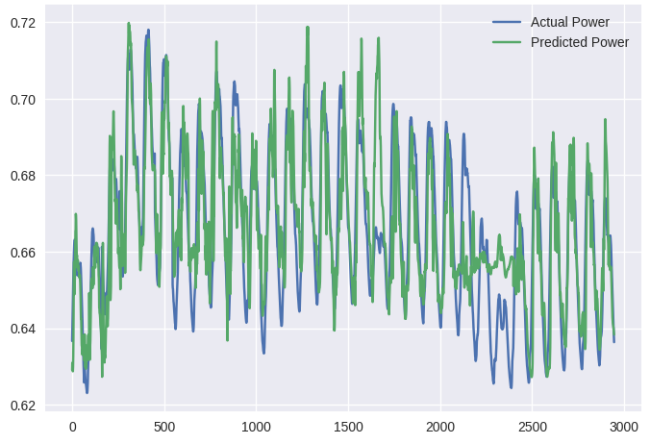

Fig. 8. BEST complete power data

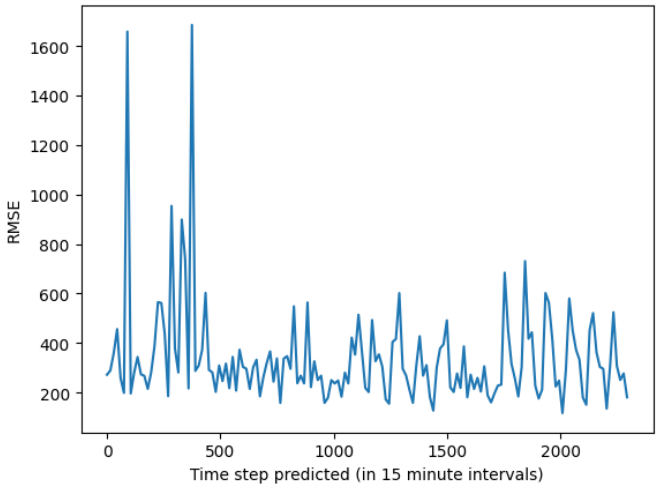

Fig. 9. BEST complete power data

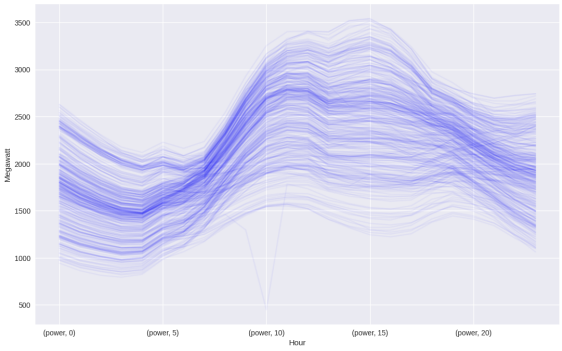

Fig. 10. BEST complete power data

B.

IV. Un-Supervised

A. K-means

K-means clustering is a popular unsupervised learning method for partitioning data into K clusters. The goal is to minimize the sum of squared distances between each data point and its cluster centroid. In this project, we applied K-means clustering to power consumption data over time. We used the elbow method to determine the optimal number of clusters.

B. Density-based spatial clustering of applications with noise(DBSCAN)

DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is a popular clustering algorithm in machine learning and data mining. It is a density-based clustering algorithm that groups together data points that are close to each other in high-density regions and identifies outliers as noise points. The algorithm is able to detect clusters of arbitrary shape and can handle datasets with noise and outliers.

C. Comparision of K-means and DBSCAN

In traditional K-means clustering, each data point is assigned to a specific cluster. However, this method may not be optimal as it can group data into clusters even if they are not strongly correlated. To address this issue, the DBSCAN algorithm can be used to identify outliers within each cluster. [4] The k-means and DBSCAN are unsupervised clustering algorithms used to group data points based on their similarities. In this particular dataset of power consumption of BEST, both models are used to find patterns in the power consumption data and identify clusters of similar power usage. The k-means algorithm partitions the data into k clusters by minimizing the variance within each cluster, while DBSCAN finds dense regions of data points and separates outliers. The comparison between the two models is based on the evaluation of the Davies-Bouldin index, which measures the similarity between clusters and the separation between them. The aim is to identify which model produces more meaningful and accurate clustering results for the power consumption data of BEST.

The data has been segregated into four seasons and the individual DB scores, clusters, and T-SNE based validation have been presented for both k-means and DBSCAN techniques in figures 11 and 12. The optimal number of clusters has been determined by observing the DB score in figure 8, and the corresponding k-means clusters have been shown in the figure. In those cluster plots, the fourth plot has three clusters, while the rest have two. Although the first two plots have three clusters, one of them is flat and cannot be considered as a cluster. The last row of figure 8 shows the clusters after density reduction. The corresponding t-SNE plots display the formed clusters, and out of four t-SNE plots, two match with the k-means clusters, indicating their validation with t-SNE.

The three k-means plots display two clusters, and the gap between them may indicate different electricity users such as the industry and residential sector. The upper cluster in the plots suddenly shifts from the lower cluster around 10 am and again matches with it around 8 pm, suggesting that offices and industries operate in this time period. The fourth plot has three clusters, and the lower cluster has not been observed in the earlier plots. This lower cluster may be because of the low use of air conditioners and cooling fans in the winter season. Figure 12 displays the DBSCAN plots for the four different seasons. The first row shows the DB score of different cluster possibilities. The darker blue zone has the lower DB score, and the selection of the tolerance radius and the number of samples to consider core points are based on it. These parameters are then used for clustering, and t-SNE plots are plotted to validate the given clusters of the data.

The plots of the k-means clustering and DBSCAN for whole dataset has been shown in figure 13. The first row shows the DB score for each type of clustering, by observing these plots the respective hyperparameter are selected.Like number of clusters in k-means as seven and the tolerance as 0.25 and the minimum number of samples as 5 for DBSCAN. In both the plots three clusters formed.But the trend of each cluster is different . The number of optimal clusters indicated in DB score for k-means has not followed in it, but around seven clusters formed in the t-SNE validation.This might because the model wont find more than three clusters in the given data. in the t-SNE of the DBSCAN has one large cluster and other three small clusters. In both plots of clusters the lower two clusters are somewhat at the same location but the upper cluster is shifted to top in the k-means cluster.

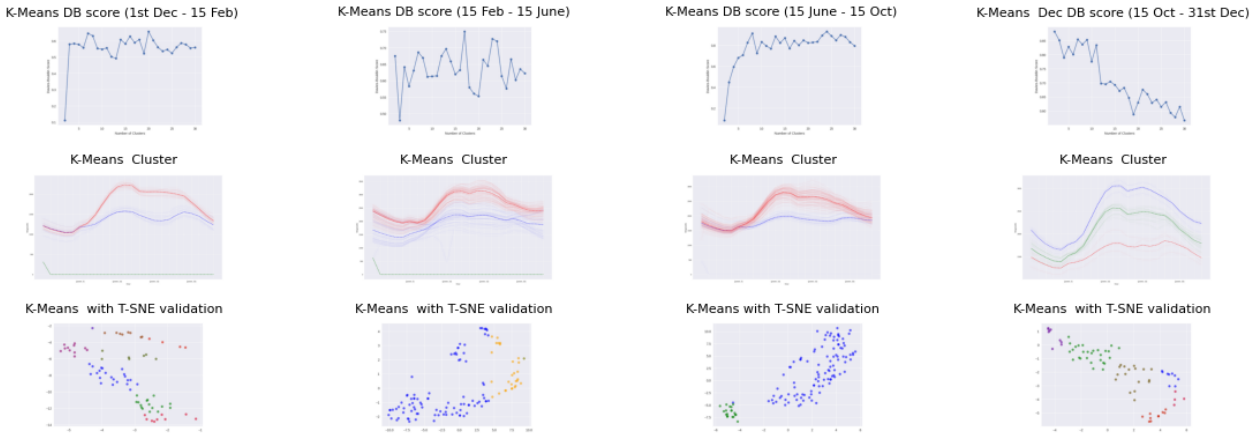

Fig. 11. K-means Clustering

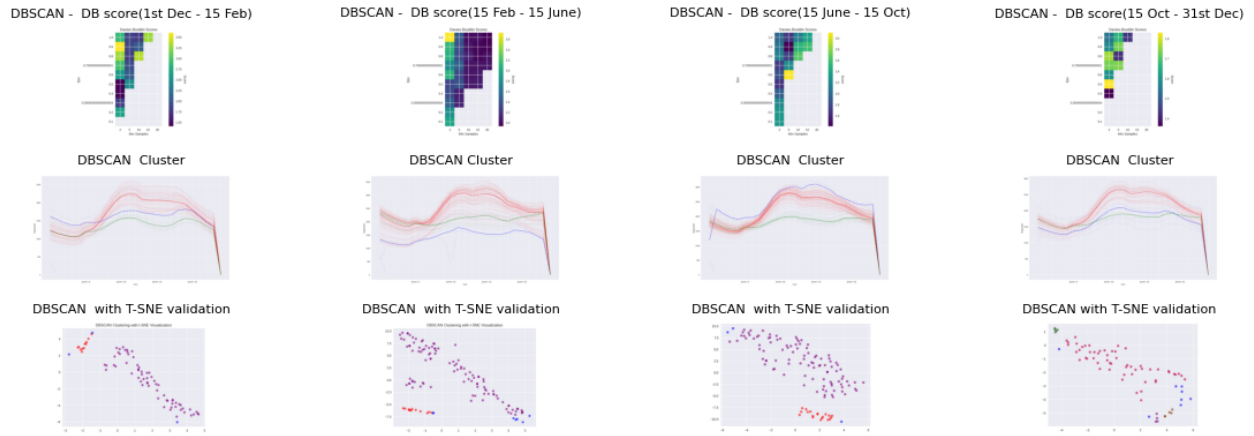

Fig. 12. DBSCAN

D. Anomaly Detection with Gaussian Mixture Model

Anomaly detection is the process of identifying data points that deviate from the normal behavior of a system. It is widely used in various domains, including finance, healthcare, and cybersecurity. Gaussian Mixture Model (GMM) is a statistical model used for clustering and density estimation. It is often used for anomaly detection because it can identify data points that have low probability under the learned density function. In this paper we referred [2], a novel unsupervised anomaly detection algorithm was introduced, which aimed to identify anomalous time points in a given dataset. The algorithm generated anomaly scores for each data point, which were visualized to guide the analyst to important time points. In this code, we are using GMM for anomaly detection in a dataset of power consumption in Mumbai. First, we read in the data from a CSV file and convert the date and time columns to a datetime object. Then, we create two new columns with the time in hours and 15-minute intervals respectively.

Next, we extract the features to be used for anomaly detection, which are the time in 15-minute intervals and power consumption. We fit a GMM with 2 components to the feature data. The probability densities for each data point are then obtained using the fitted GMM.

We set a threshold for anomaly detection by taking the 5th percentile of the density values. Any data point with density value below this threshold is considered an anomaly. We identify the anomalies by selecting the rows in the original dataset corresponding to the anomalous data points. This method is referred from the analysis of [3]

Finally, we plot the data with anomalies highlighted in red. The x-axis shows time in 15-minute intervals and the y-axis shows power consumption. The title of the plot is ”Anomaly Detection with Gaussian Mixture Model”. The code also sets the x-axis limits to 0-96 (24 hours in 15-minute intervals) and sets the x-axis ticks to 4-hour intervals as depicted in Fig 14. As a result, the red patches are showing uneven or unexpected power consumption on particular hours or time interval. And blue is a usual power consumption.

V. Discussion

The LSTM-based prediction model for time series data performed well, with an RMSE of 0.028. When it comes to

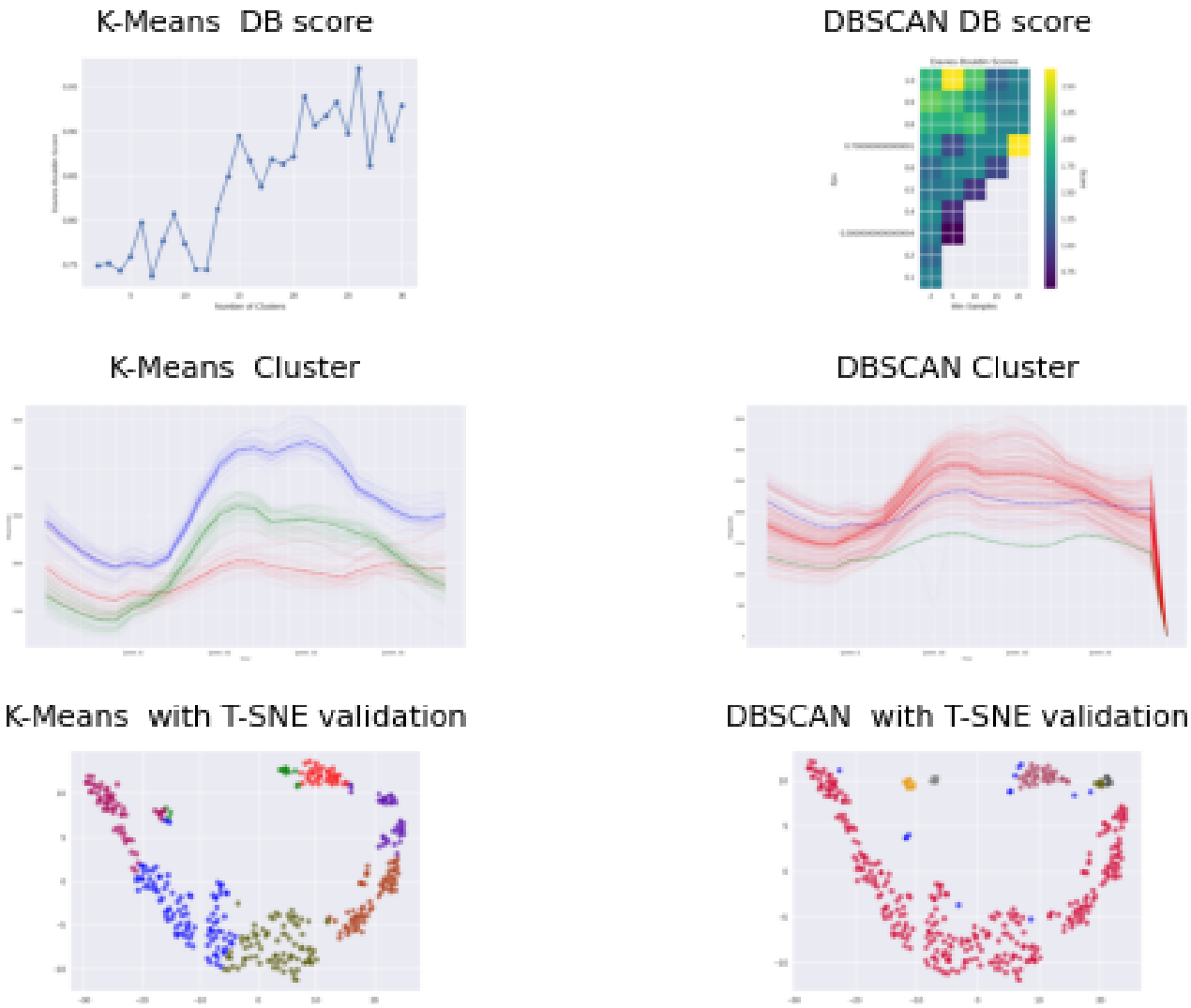

Fig. 13. K-means Clustering

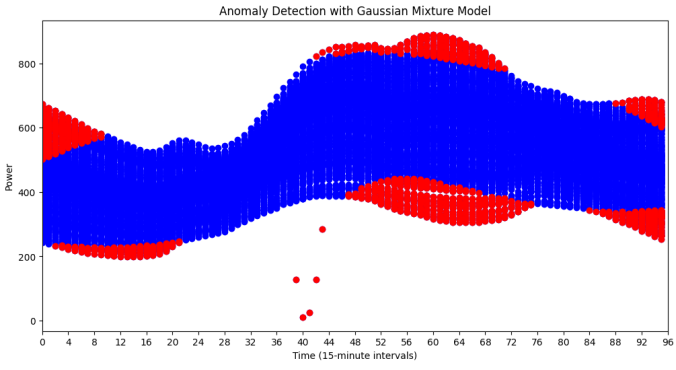

Fig. 14. Anamoly of power consumption

unsupervised clustering, DBSCAN had a slight advantage over k-means in seasonal clustering, as it formed three clusters, while k-means mostly formed two. However, for clustering the entire dataset, both methods formed three clusters. K-means was more sensitive to outliers in the upper cluster than DBSCAN. Both methods showed different performances for smaller datasets, and their results for the whole dataset also differed to some extent.

VI. Conclusion and Future Work

The combination of these techniques can be used to accurately predict future power consumption values, identify clusters of similar patterns, and detect anomalous behavior that may require attention. The results of this analysis can inform decisions related to energy optimization and efficiency, leading to cost savings and reduced environmental impact. We will also implement K-mediod, because median is chosen the representative days we get would be chosen from the original data and will not be a mean of the original data as in k-means.

REFERENCES

[1] Chat-gpt http://towardsdatascience.com/clustering-electricity-profiles-with-k-means-42d6d0644d00

[2] Janetzko, H., Stoffel, F., Mittelstädt, S., Keim, D. A. (2014). "Anomaly detection for visual analytics of power consumption data," Computers Graphics, 38, 27-37.

[3] Laurinec, P., Lóderer, M., Lucká, M., Rozinajová, V. "Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption," Journal of Intelligent Information Systems, 53, 219-239.

[4] Zhang, L., Deng, S., Li, S. (2017, November). Analysis of power consumer behavior based on the complementation of K-means and DBSCAN. In 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2) (pp. 1-5). IEEE.