Electricity Meter Reading Clustering and Prediction Models on the Building Data Genome 2 (BDG2) Dataset

Jasmine Pinyu Zou

Abstract

The growth of reliance on the usage of electrical and electronic devices have increased the demand for energy production. Precise energy management from understanding the customer electricity usage patterns and forecast on consumer electricity usage assist utility companies to provide better service, policies and to match supply with demand. This study proposes a clustering-based analysis of electricity consumption using K-means clustering algorithm to categorize consumers electricity usage into different levels based on day types (weekday vs. weekend) as well as typical daily consumption patterns throughout a year. This study also attempts to create an electricity consumption prediction model using K-neighbour regressor algorithm and a simple neural network. K-neighbour regressor algorithm was used to predict 6-month of meter consumption ahead with a year and a half of data points, whereas the neural network model utilizes the past 3 months of meter readings at a given time point to predict the next day electricity meter consumption for 201 days. Building meter reading data from the Building Data Genome 2 (BDG2) dataset is used for this study.

1 Introduction

As population grows, the demand for electricity also grows. There is often a gap between the demand and supply in the electricity market [1]. To improve the service efficiency through providing necessary supply to the market, utility companies conduct data analysis on the electricity meter data making attempts to meet the demand, update policy, and optimize the electricity consumption. This study looks into two different types of analysis on the time series electricity meter data: clustering and forecasting. Clustering analysis provides insights on electricity usage patterns of clients, which helps adjust the amount of electricity production at specific times [2]. On the other hand, consumption forecast helps with demand response program, which is a common practice to reduce the stress on the grid and high electricity prices. It is usually implemented with dynamic pricing or other forms of financial incentives to encourage consumers to change consumption behavior, either reducing or shifting the electricity during peak period. Ultimately, it helps balance supply and demand for electric system planners and operators [3].

2 Methodology

2.1 Dataset and Scope

The Building Data Genome 2(BDG2) dataset provided by the Buds Lab at the National University of Singapore is an open-access dataset that consists of 3053 time-series energy meter readings from 1636 buildings [4]. The hourly meter reading data ranges from the beginning of 2016 to the end of 2017. The available meters are chilled water, electricity, gas, hot water, irrigation, solar, stream, and water meters. The buildings in the dataset are grouped by the site and categorized by the primary and secondary space usage. A metadata profile is provided for information on building characteristics such as square footage, location, and climate zone. A weather file including air temperature, wind speed, and cloud coverage is also provided for the sites the buildings are located at. The meter availability is different across buildings. A quick explorative data analysis was conducted on the metadata file as well as individual meter datasets as shown in Figure 1.

This study is scoped to use the hourly electricity meter data of buildings with the most available data and air temperature weather information for analysis. The top ten most data available buildings used for clustering analysis are as follows:

- Lamb Office Bertha
- Lamb Education Willetts,
- Lamb Education Lazaro,
- Lamb Industrial Enrique,
- Lamb Industrial Carla,
- Lamb Assembly Dorothy,
- Lamb Education Eula, and
- Hog Education Casandra.

The animal name is code for the site location; the second word indicates the building type; and the last word is the building name. The building used for electricity consumption analysis is Lamb Office Bertha.

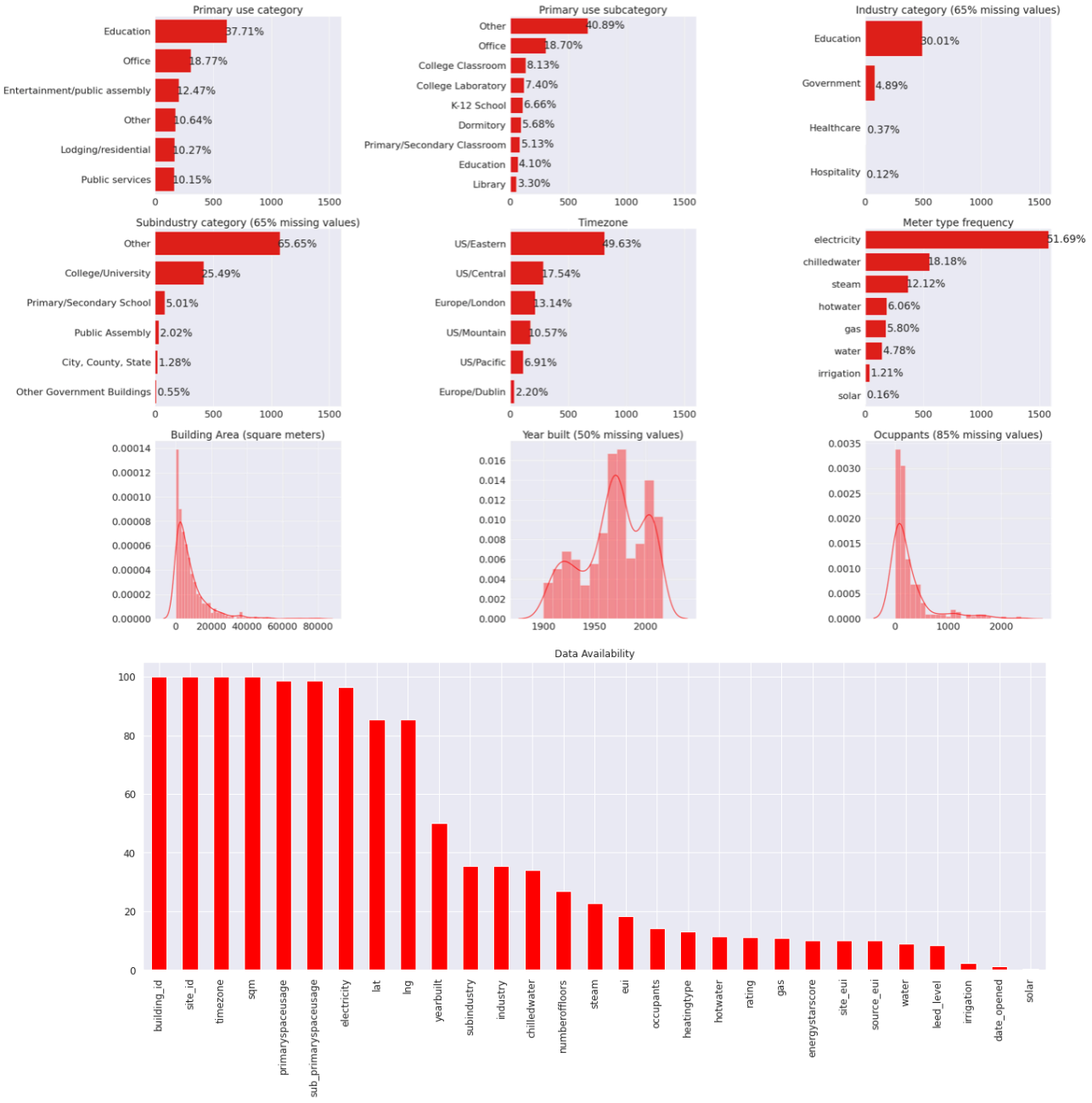

Figure 1 BDG2 Metadata File Overview: BDG2 is a diverse dataset. Education buildings, office used buildings are the primary building types in this dataset. Electricity data is the most available data among the meter data with an availability of 96.5%.

2.2 Methodology

2.2.1 Data Preprocessing

The dataset retrieved from the Buds Lab GitHub repository [5] is cleaned, normalized, and processed to be analysis ready. Faulty meter detection has been conducted and abnormal readings are excluded from the dataset [4]. However, upon further analysis, there are still some missing meter reading values and inconsistent timestamps in the electricity meter and weather data.

Therefore, the data is first preprocessed on missing values using imputation and resampling techniques.

2.2.2 Clustering Analysis

Time series plots for the ten selected buildings are first plotted for pattern exploration. A traditional daily profile analysis with weekday and weekend split is conducted to extract further insights on consumption behaviors. Then, the building consumption data is clustered into 4 different types across the time range using unsupervised K-means clustering algorithm. The algorithm aims to separate the observations in the dataset into k clusters (k = 4 in this study) in which each observation belongs to the cluster with the nearest mean [6].The clustered results are visualized; and further analysis of the clusters such as weekday and weekend breakdown of the cluster profile is provided.

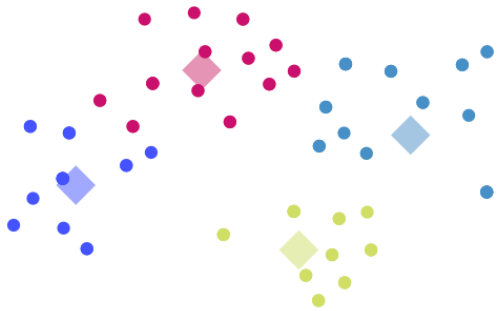

Figure 2 K-Means Clustering Example (k =4); The diamonds are the cluster centroids.

2.2.3 Forecast Analysis

2.2.3.1 KNeighborsRegressor

KNeighborsRegressor package from Scikits-Learn library and neural network package from Keras library are used for the forecast analysis on Lamb Office Bertha electricity meter data. The air temperature information of Lamb site from the weather data file is used as well.

KNeighborsRegressor is a regression algorithm based on k-nearest neighbours; it retrieves some k neighbors of query objects, make predictions by local interpolation of the targets associated with the nearest neighbors in the training set [6].

The KNeighborsRegressor model created in this study predicts six months of hourly electricity meter reading using the past one and a half year of meter reading data. Temporal categorical variables of time of day and day of week are first encoded into categories to allow the model to use the information effectively.

2.2.3.2 Simple Neural Network

A neural network is a black-box machine learning model, containing a network of artificial neurons. Each neuron carries a weight; all inputs are read and modified by the weight and fed into the activation function, which controls the amplitude of the output. The core concept of a neural network is forward and backward propagation. Figure 3 shows a simple neural network architecture with 2 input neurons, 1 hidden layer with 3 neurons, and 1 output neuron. The input neurons can be viewed as features from the dataset that could potentially contribute to the output of interest.

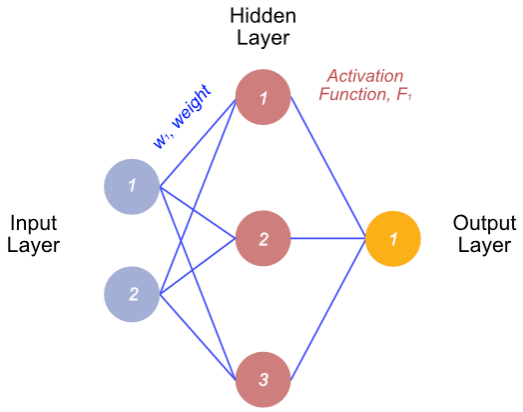

Figure 3 A Simple Neural Network: Each blue line resembles the connections of the neurons modelled by weights. Each neuron in the hidden layer and output layer has an activation function to fire up the input for the next layer (i.e., the activation functions in the hidden layer prepare the input for the output layer).

The neural network model created in this study predicts the day ahead hourly electricity consumption using the past 60 days of meter readding data. That the neural network is used to predict the last 201 days of the electricity meter consumption in the two-year time frame. The neural network constructed for the study contains the following architecture:

- Input layer: 50 input neurons with an input dimension of 1440, ReLU activation function and 0.2 of dropout rate
- Hidden layer: 30 hidden neurons with ReLU activation function
- Output layer: 24 output neurons with linear activation function
- Mean squared error loss function with Adam optimizer and mean absolute error metrics

Dropout rate is used to prevent overfitting to the training dataset by randomly selecting neurons to be ignored during training. Those randomly selected neurons won't be able to contribute to the downstream neurons' activation, nor would it help update any weight updates on the backward pass [8].

ReLU (rectified linear activation function) is a piece wise linear function that outputs the input directly if it is positive, or zero if the input is negative [9] . It is selected to use it as it is one of the most commonly used activation functions on simple neural networks. Linear activation function is used for the output layer as it would return the unmodified output. The number of input and hidden neurons is by arbitrary choices. The input dimension reflects the dimension of the training dataset, which includes 2 months of past meter reading data per reading (2 x 30 x 24 = 1440). There are 24 output neurons as there are 24 hours in a day. Each output neuron is responsible for an hour prediction.

To prepare the data for the neural network, a shifted dataset is created. For each meter reading data *t*, a *{t-1, t-2, …, t-1440}* set of backward shifted past meter reading data is created as the input dataset, X. A *{t+1, t+2, …, t+23}* set of forward shifted future meter reading data is created as the output dataset, Y. The same neural network architecture is used for both input with and without weather air temperature information.

2.2.3.3 Evaluation

Both models use a 70:30 train-test split. Mean Squared Error (MSE) is used to evaluate the accuracy of the forecast models:

$$Mean\ Squared\ Error = \frac{1}{n}\sum_{i=1}^{n}(Y_{observed} - Y_{predicted})^2$$

Where n is the number of data points, $Y_{observed}$ is the observed true values, and $Y_{predicted}$ is the model predicted values of outputs.

Note that the time range of prediction and purposes of predictions are different between the KNeighborsRegressor model and the neural networks. Therefore, no direct comparison will be conducted between these two models. However, for the neural network, Mean Absolutely Error (MAE) is used for model comparison between the weather and no weather neural network model.

$$Mean\ Absolute\ Error = \frac{\sum_{i=1}^{n}|y_{predcited} - y_{observed}|}{n}$$

For bot MAE and MSE, the smaller the values are, the better the model performance it indicates.

3 Data Analysis

3.1 Clustering Analysis

The first part of the clustering analysis is to plot out the daily profiles of each building of interest to check the pattern of meter data. The full two-year time frame is plotted as well as a selected two weeks from 01-30-2016 to 02-14-2016 for a zoomed-in perspective. Jan 30, Jan 31, Feb 6, Feb 7, Feb 13 and 14 are weekends. It appears that there is some standard weekday vs. weekend behaviors and a few basic types of daily patterns. Strong weekday versus weekend profiles can be found in office and education types of buildings. The industry and assembly building shares similar windows of unoccupancy during the two-year time frame.

.

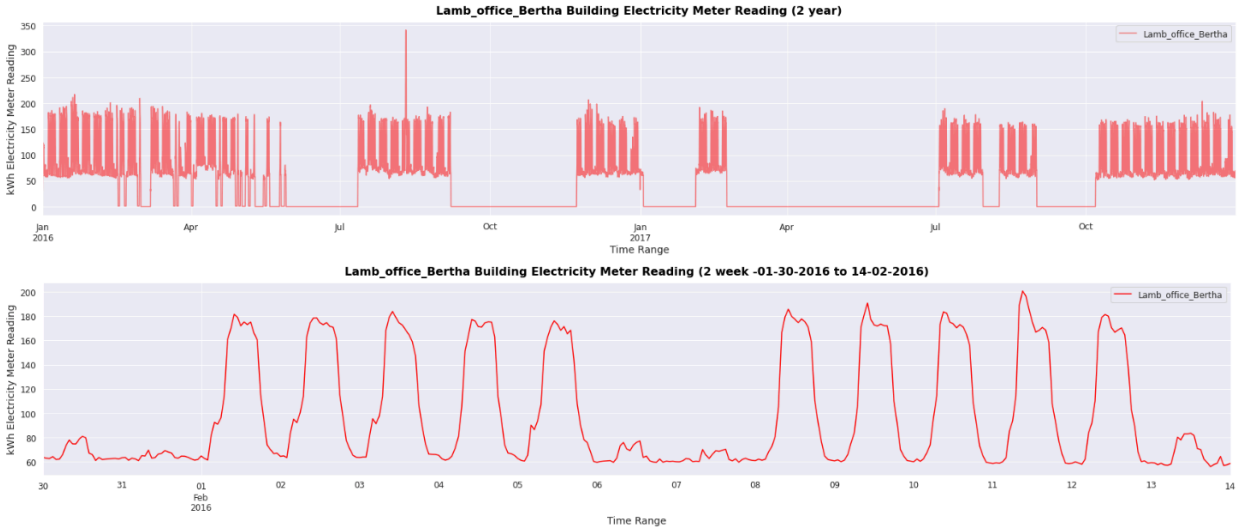

Figure 4 Lamb Office Bertha Time Series Plot. There is a strong weekend vs. weekday profile from the two-week plot. The building seems to have no or low occupancy during certain periods of a year, around summer months and October.

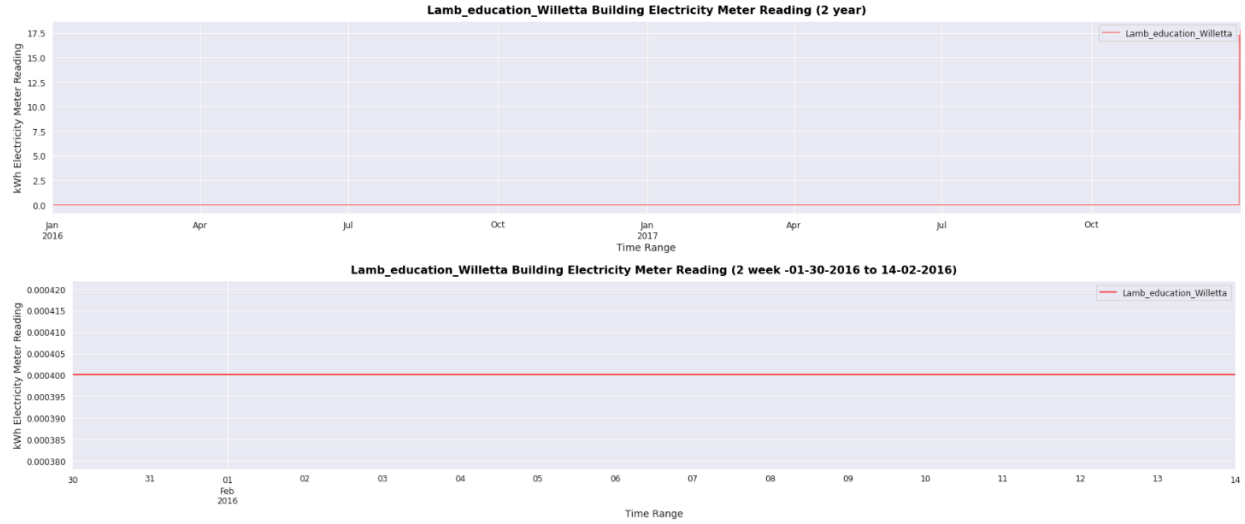

Figure 5 Lamb Education Willetta Time Series Plot. The building seems to be not in use for the majority of the time frame and was in occupied mode at the end of 2017.

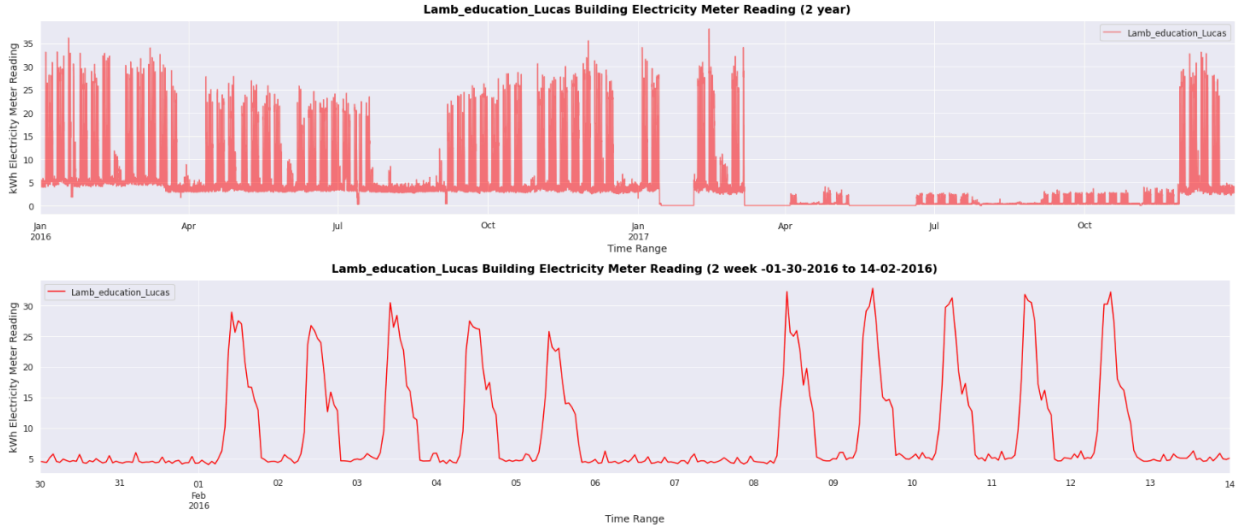

Figure 6 Lamb Education Lucas Time Series Plot. There's a strong weekday and weekend profile shown in the two-week plot. The building seemed to be in low / no occupied mode from March to end of November of 2017.

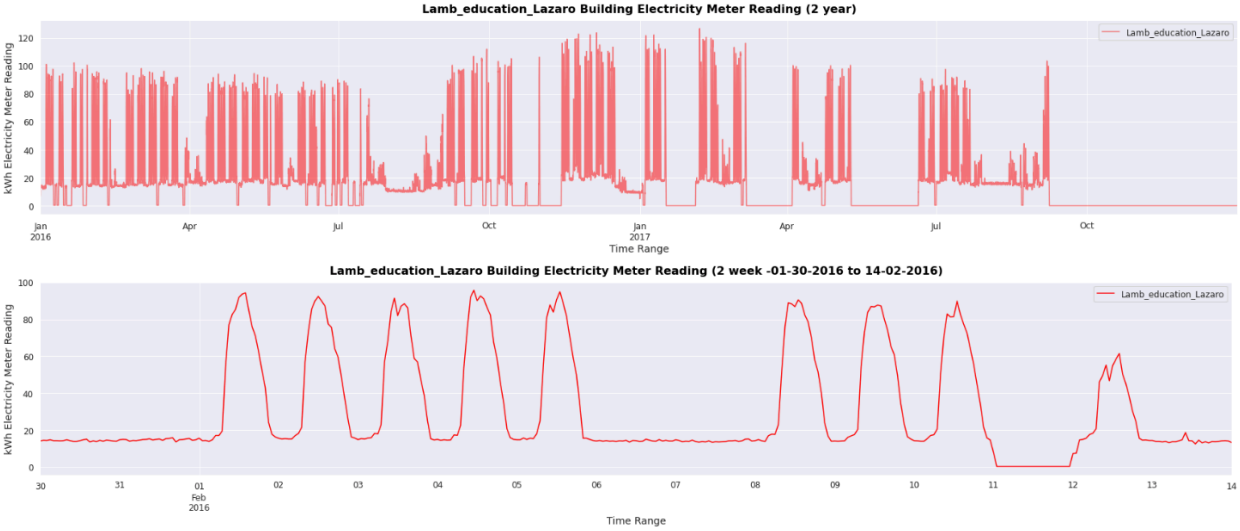

Figure 7 Lamb Education Lazaro Time Series Plot. A weekday and weekend profiles can be observed from the two-week plot. The building has regular patterns of consumption in 2016 and 2017 respectively. The building seemed to be unoccupied after September 2017.

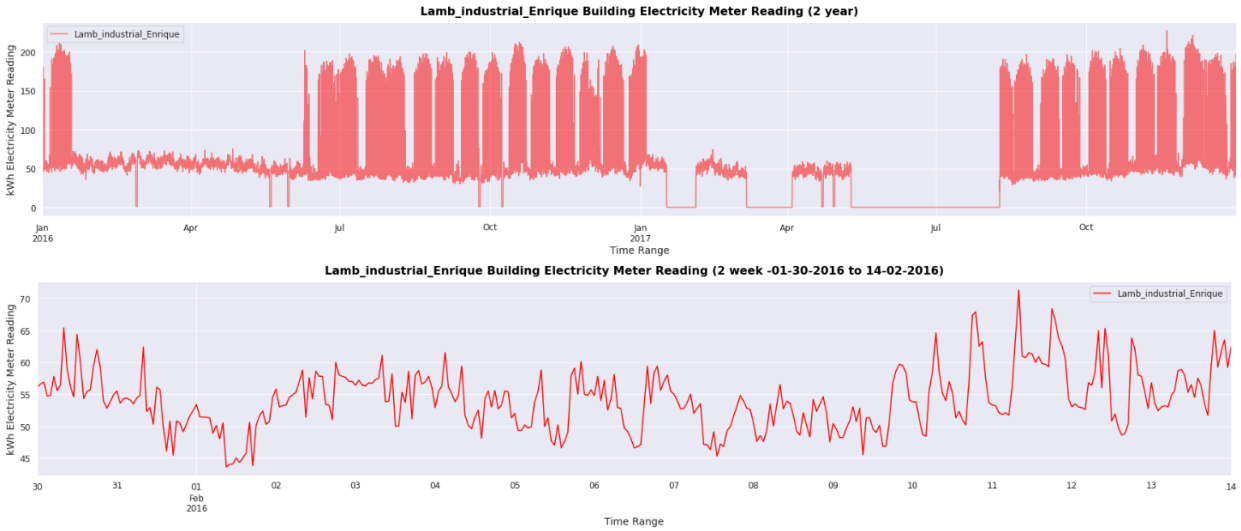

Figure 8 Lamb Industrial Enrique Time Series Plot. The building seemed to be in occupied mode from June to January for each year and low or no occupied mode for the rest of the time. There's no strong weekday vs. weekend profile. Rather, there's a different periodic consumption pattern where each period lasts around half a month or nearly a month, as shown in the 2-year plot.

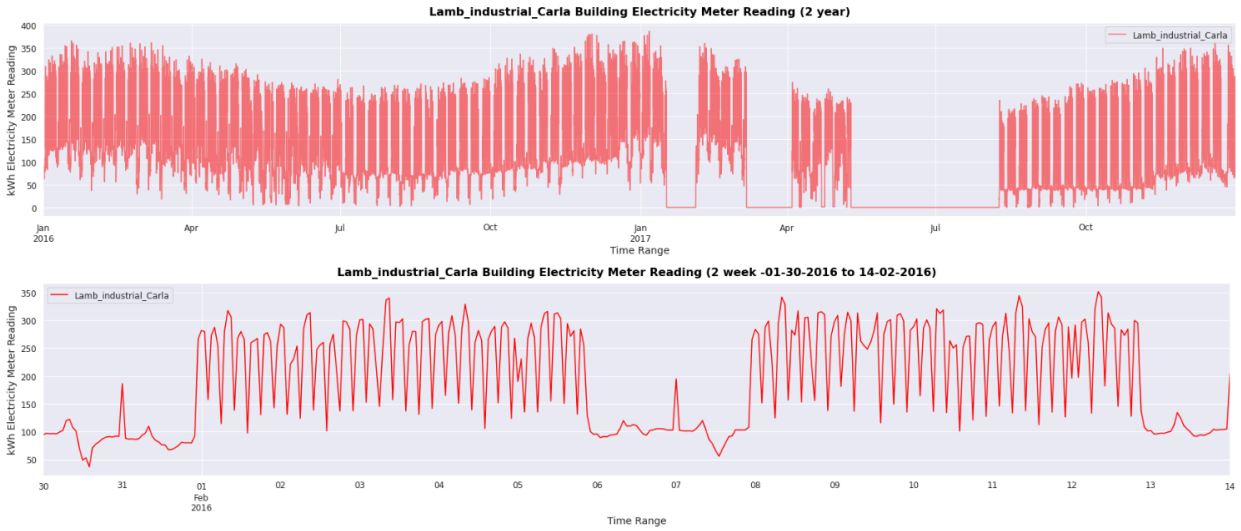

Figure 9 Lamb Industrial Carla Time Series Plot. This industrial building demonstrates different weekday and weekend consumption patterns. The building seemed to be unoccupied for several months in 2017.

Figure 10 Lamb Assembly Dorothy Time Series Plot. This building seemed to have no big differences between the weekday and weekend consumption behaviors. However, there're some seasonal consumption patterns throughout the two-year plotted time frame. This building sharing the similar unoccupied months as the previous industrial type buildings.

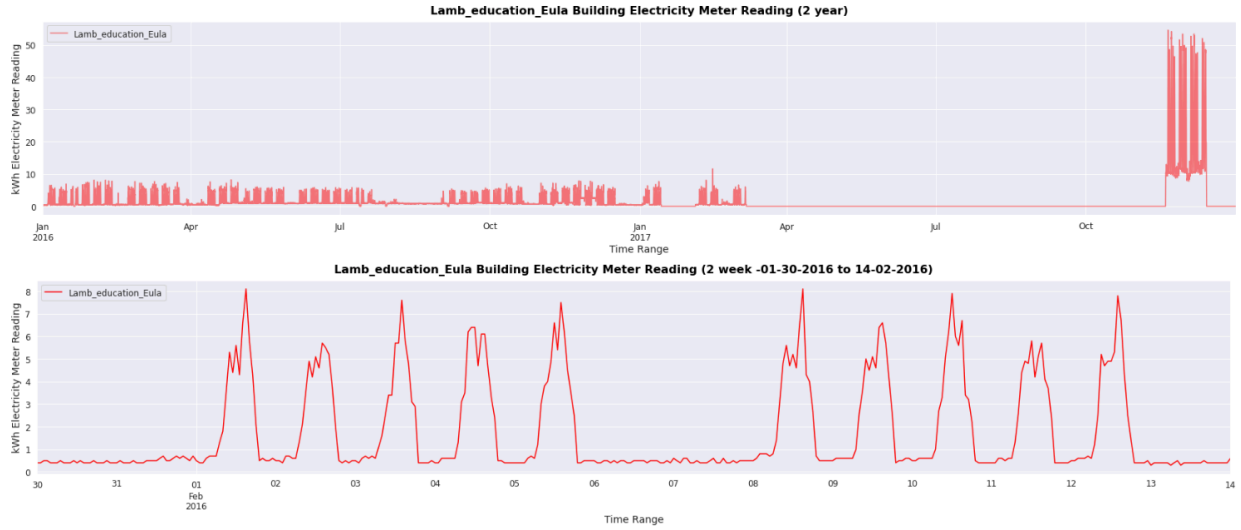

Figure 11 Lamb Education Eula Time Series Plot. This building was not utilized heavily until the end of 2017. It demonstrated separate weekday and weekend consumption behaviors.

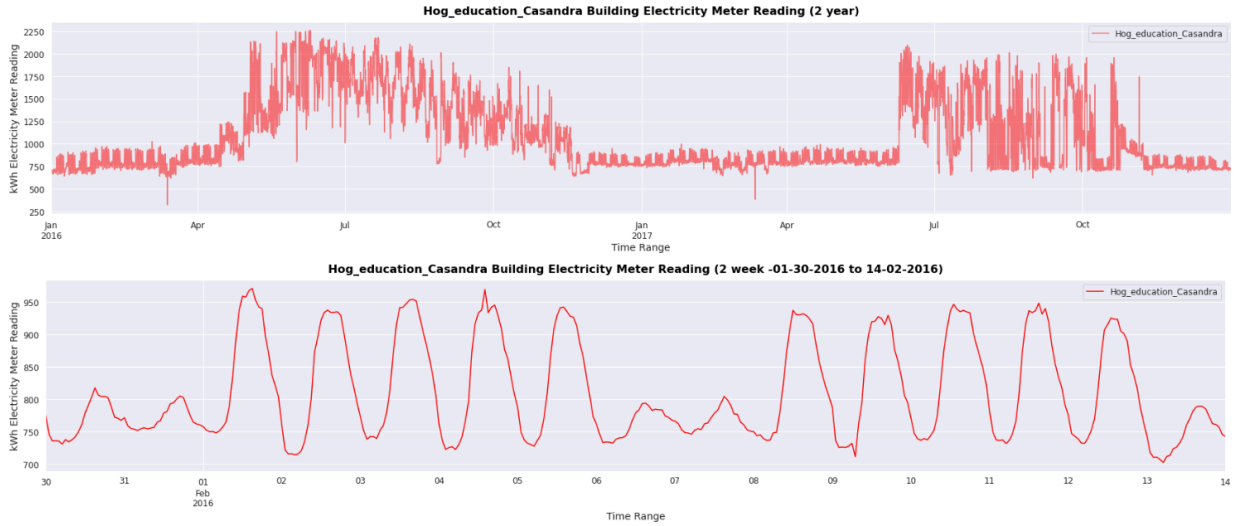

Figure 12 Hog Education Casandra Time Series Plot. This building demonstrated strong weekday and weekend consumption patterns as well as seasonal consumption variations throughout years.

3.1.1 Conventional Clustering Method

As discussed above, there are some weekend vs. weekday consumption behavior. Manually, weekday and weekend daily electricity meter reading plots are created separately to observe further of the patterns. The first plot of each building is the combined weekend and weekday daily profiles. Each line in the plot represents one day of electricity meter reading across the two-year time frame.

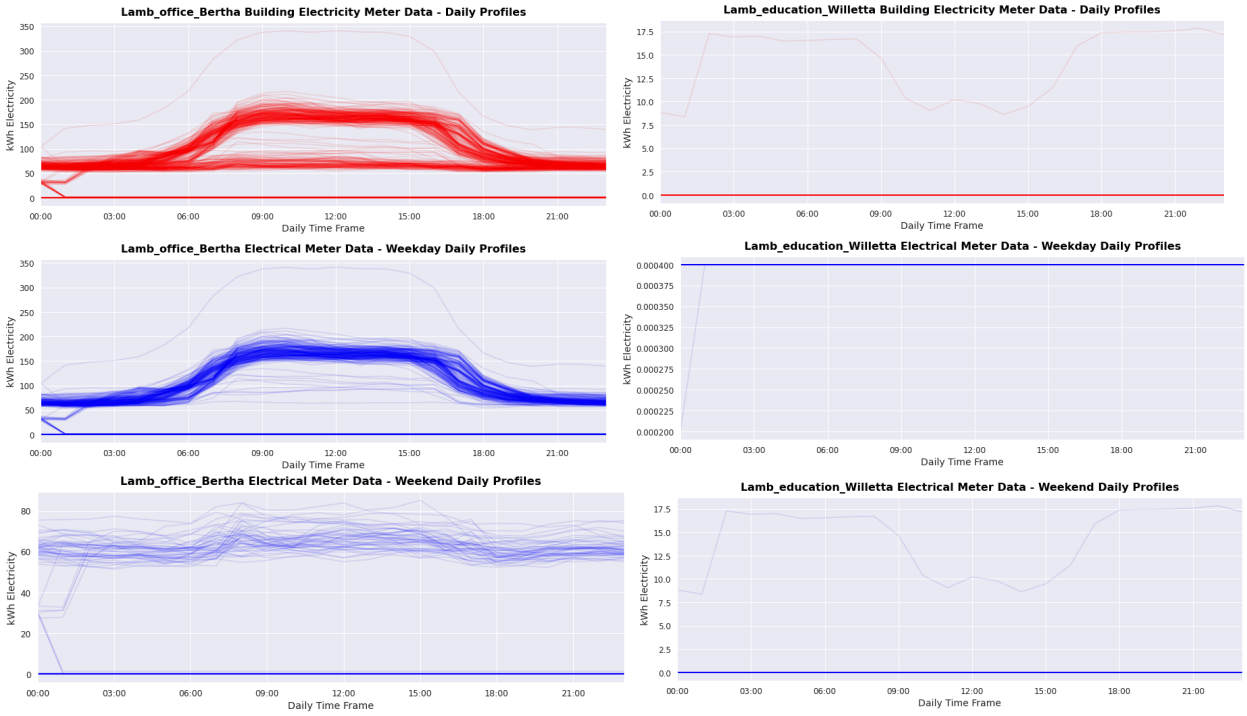

Figure 13 Lamb Education Willetta and Lamb Office Bertha Weekday and Weekday Daily Profiles. Bertha seems to have one major cluster of consumption behavior on weekdays and two clusters of consumption behaviors on weekends. Willetta seems to have one major type of weekday daily profile and two different weekend daily profiles.

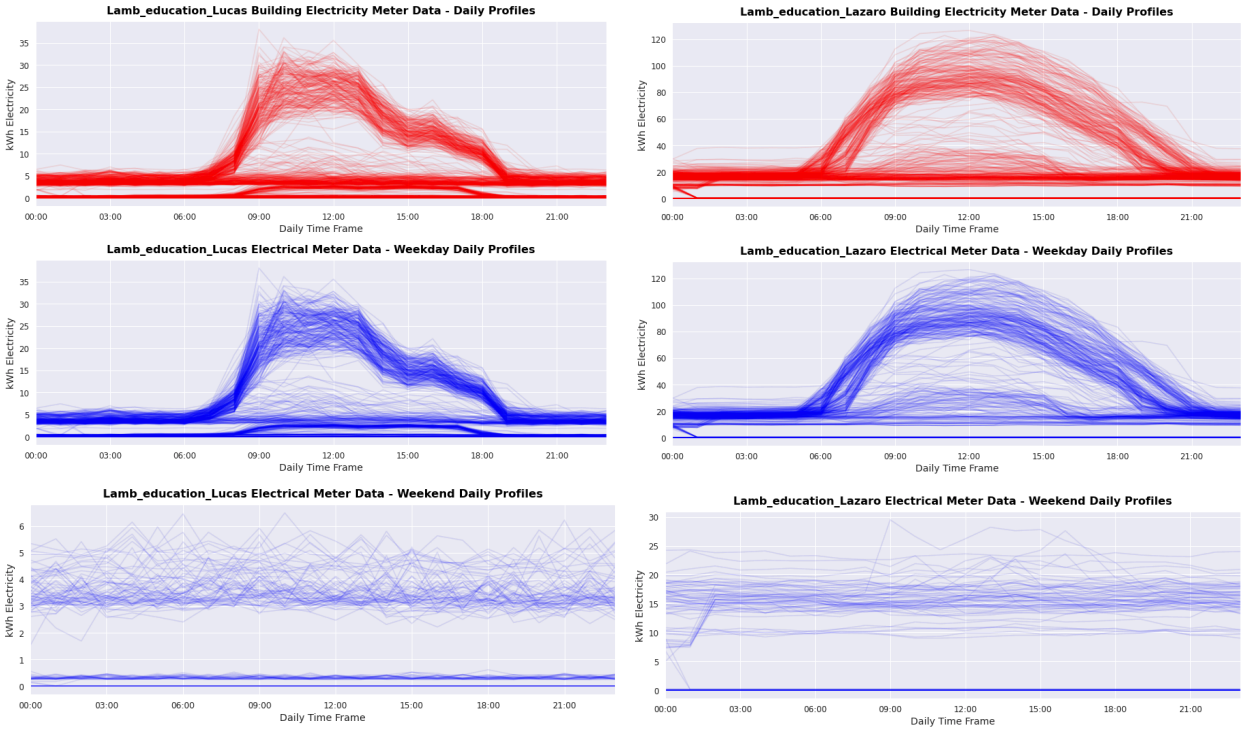

Figure 14 Lamb Education Lucas Building and Lazaro Weekday and Weekend Daily Profiles. Both these education buildings seem to have two major clusters of weekday profiles and two weekend daily profiles.

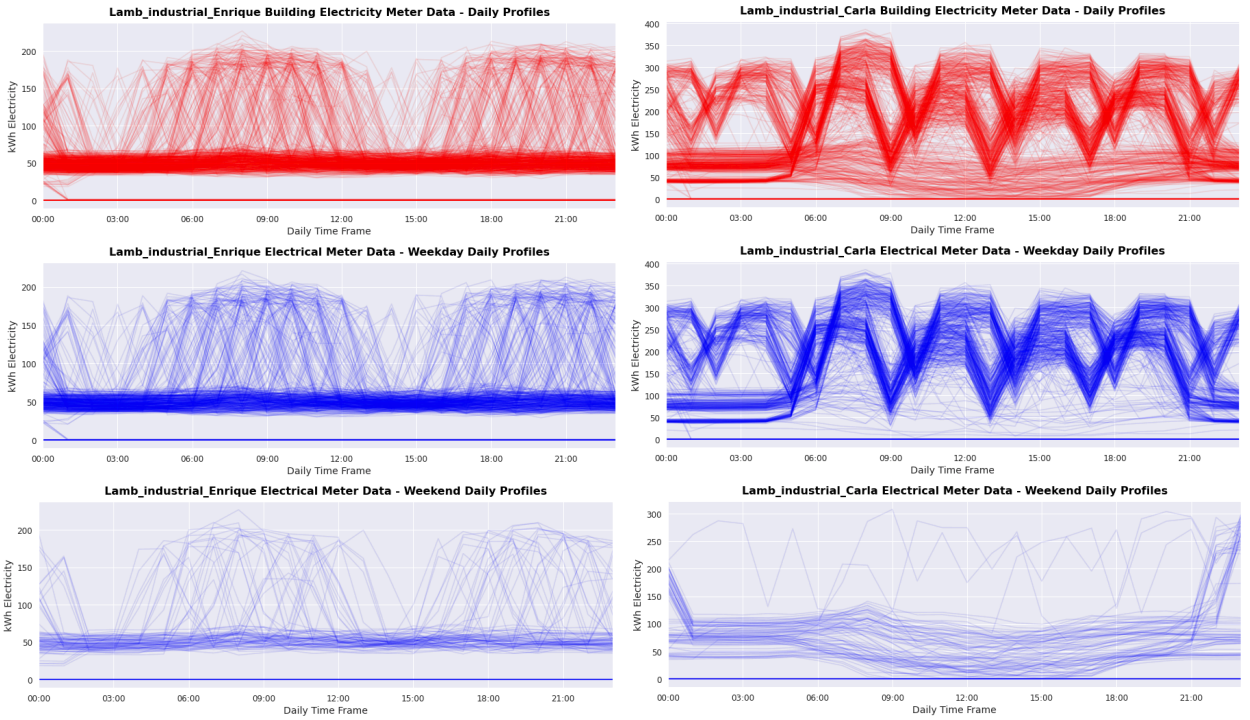

Figure 15 Lamb Industrial Enrique and Carla Weekday and Weekend Daily Profiles. Both industrial type buildings present a more messier weekday daily profile than office and education type buildings. This could be that these buildings are more dependent on shift schedules and occupant behaviors. Carla seems to have two types of weekday profiles whereas Enrique has multiple clusters of weekday profiles. Both buildings on weekends are in low usage most of the time.

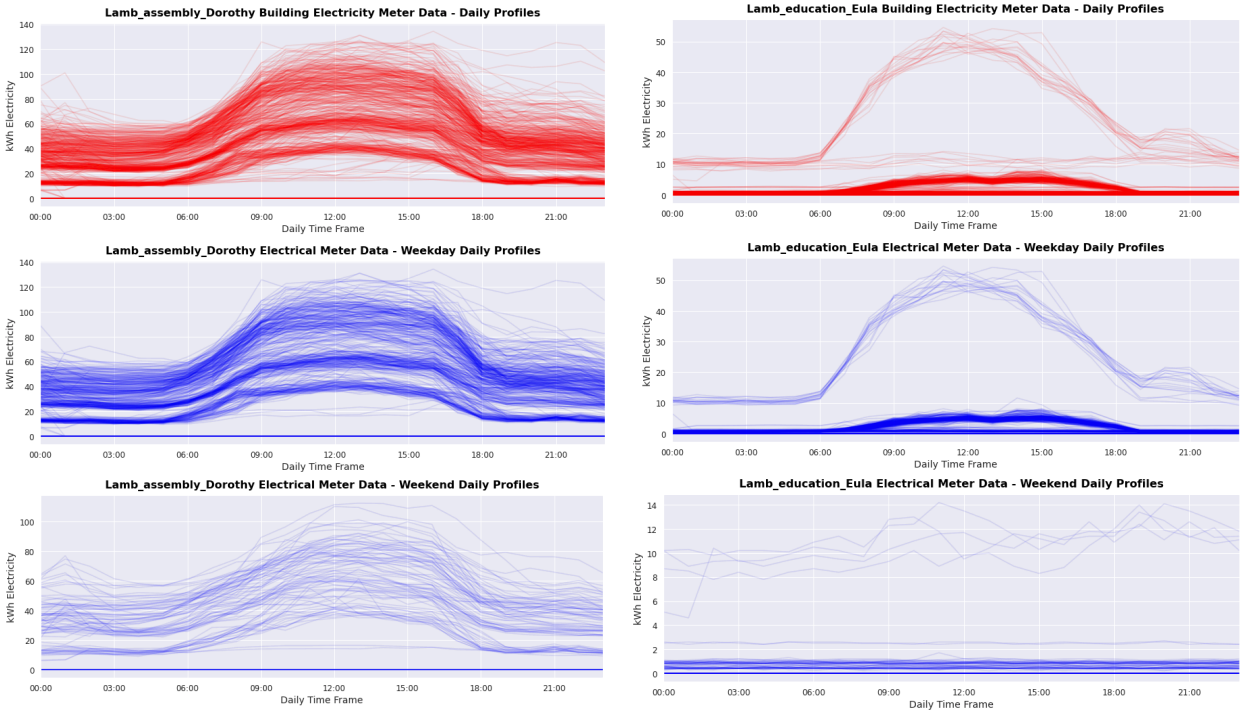

Figure 16 Lamb Assembly Dorothy and Lamb Education Eula Building Weekday and Weekend Daily Profile. Dorothy seems to have three major daily profiles on weekdays and three for weekends. Eula seems to have two major daily profiles on weekdays and two different profiles on weekends.

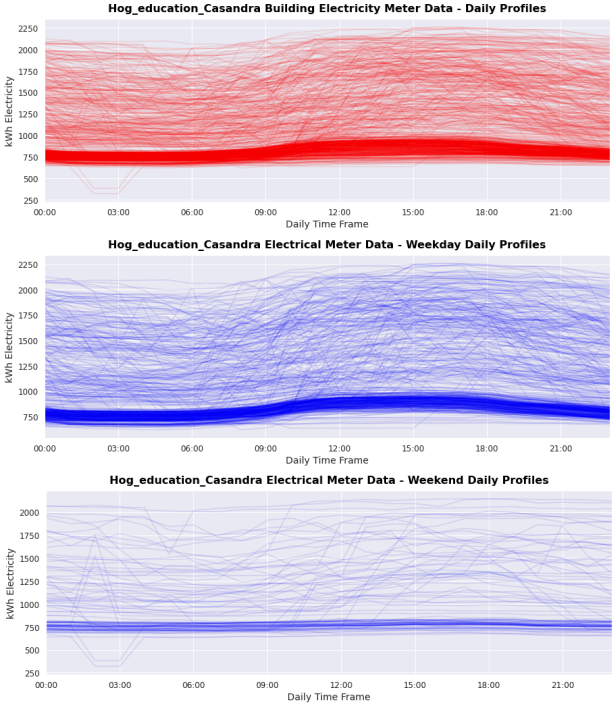

Figure 17 Hog Education Casandra Weekday and Weekend Daily Profiles. This building is located at different site as the other nine. It has a baseline daily profile meter reading with various behaviours throughout the day. This could be a heavily trafficked area where the electricity consumption is highly dynamic and occupant dependent.

3.1.2 K-means Clustering

To observe all the clustering patterns manually takes a lot of time and efforts; it is prone to human mistakes as well. In this part, the results of k-means clustering are summarized in the below figures. All 10 buildings are processed using the same algorithm with cluster numbers set to 4. Only two examples are shown included in this study report. The rest of the charts can be found in the Jupyter Notebook.

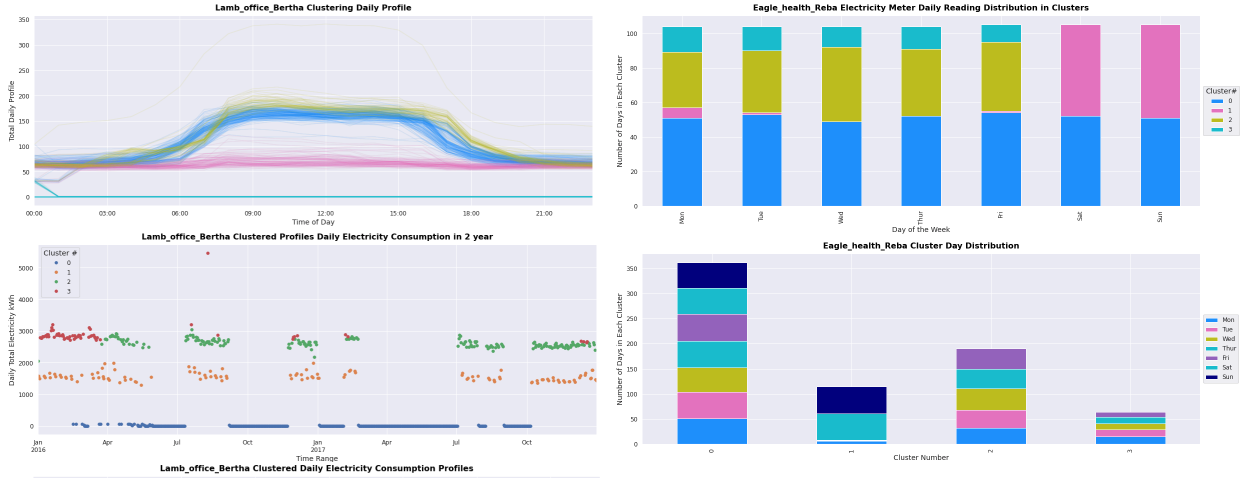

Figure 18 Lamb Office Bertha Clustering Visualization.

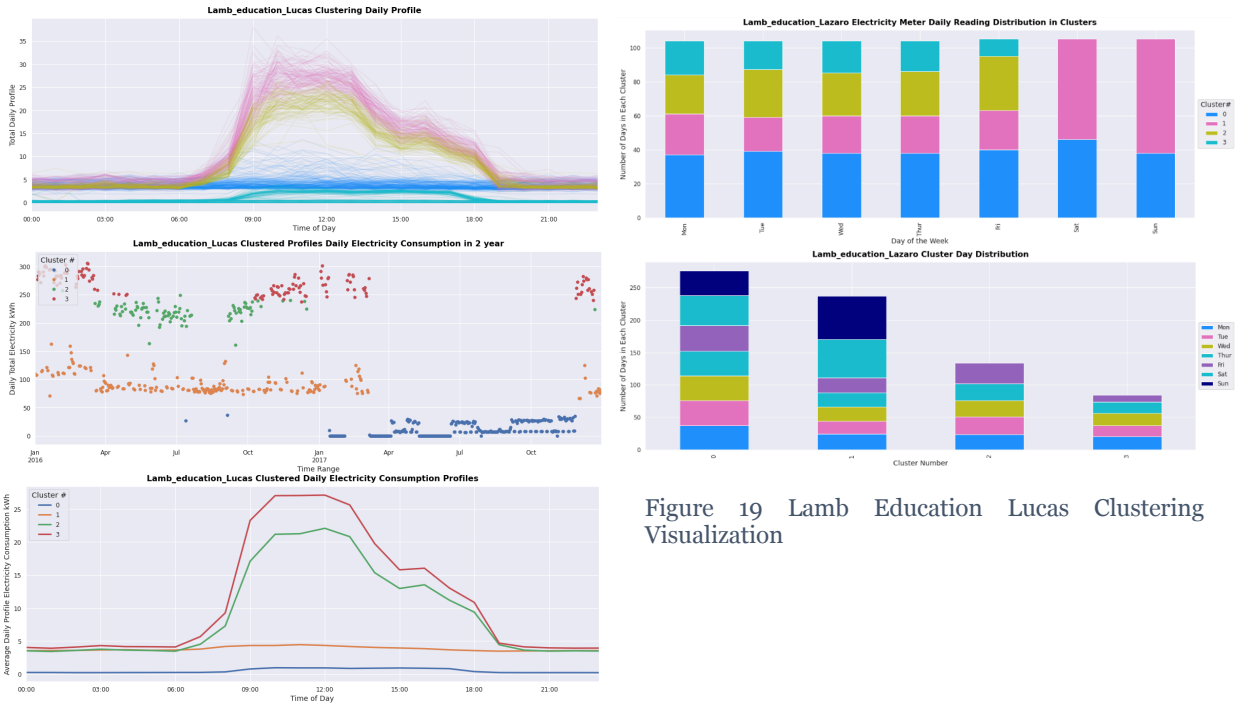

Figure 19 Lamb Education Lucas Clustering Visualization

As shown in figure 18 and 19, the four clusters in daily hourly meter profile are labelled by four different colors in the first plot of each building; an aggregated daily load profile is then plotted across the two-year time frame to demonstrate seasonality ,if any. The third plot on the left of each figure summarized the four clusters of electricity consumption behavior by calculating the average. On the right of figure 18 and 19, the weekday and weekend distribution in each cluster are visualized. The blue cluster #0 appears to be the nighttime consumption profile, and the pink cluster #1 appears to be mostly weekend consumption profile for these two education buildings. For weekdays, there are a mixed of both green #2 and aqua #3. The green cluster #2 could be occupied mode whereas the aqua cluster #3 is when the buildings are shut down with near zero electricity consumption. Note that the top left graph and the right two graphs in each figure share the same clustering labeling and coloring; the left bottom two graphs use a different set of labelling legend, which should be updated to be consistent in future studies.

3.2 Forecast Analysis

The predicted electricity meter readings are plotted against the actual meter reading data for each model. Table 1 summarizes the evaluation metrics for each model. Adding weather to the neural network appear to have a trivial impact on improving the performance, if not worsen.

Table 1 Model Evaluation Metrics

Model	MSE	MAE
KNeighborsRegressor Model	3173.96	-
Neural Network No Weather	1153	20
Neural Network with Weather	1231	22

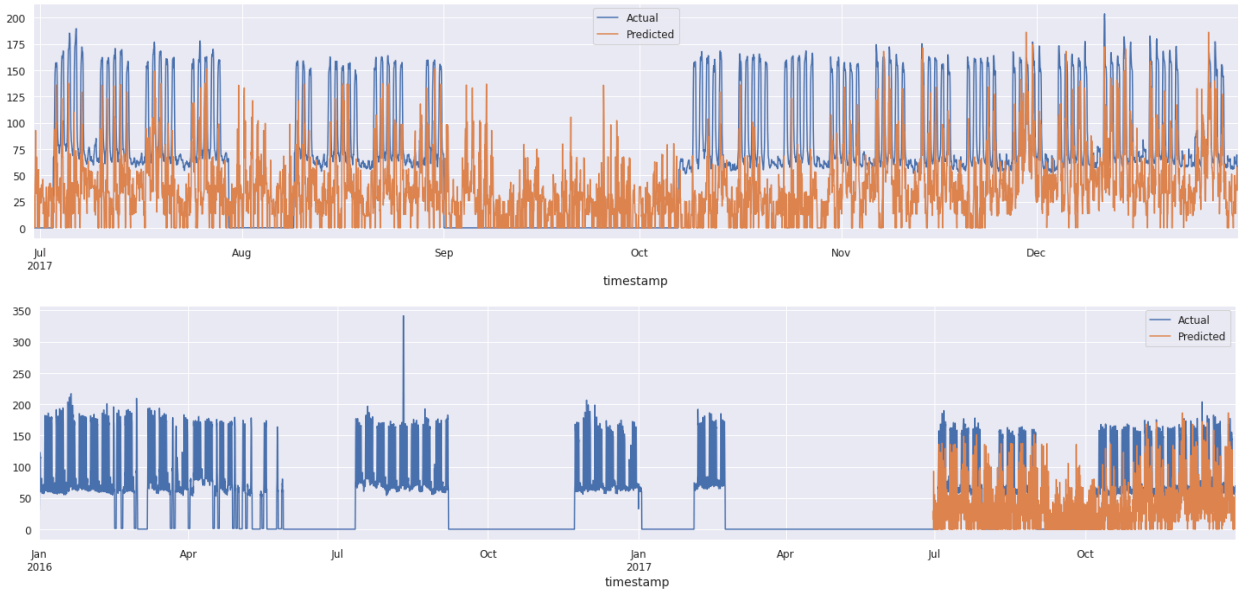

Figure 20 KNeighborsRegressor Model. The predicted 6-month of electricity meter data (orange) is plotted against the actual meter reading data (blue). The model captured the variation pattern to some degree.

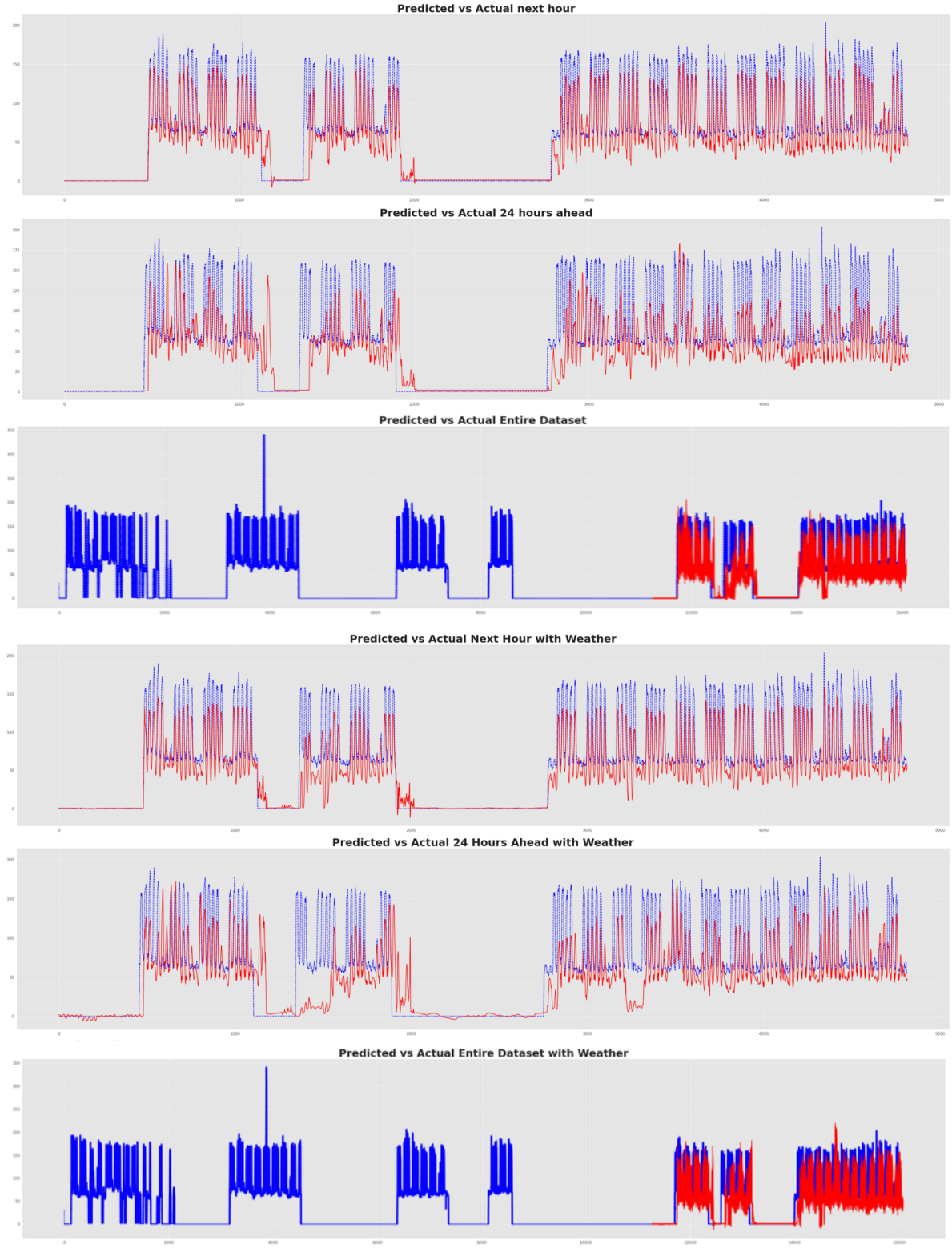

Figure 21 Neural Network Models (With and Without Weather Feature). Red – prediction reading; Blue – actual reading. The next hour prediction for both models align quite well with the actual next hour prediction. However, discrepancies are apparent between the actual and predictions of the next 24 hours can be seen in the second plots of each model. Overall, the predictions follow the general pattern of the actual meter reading consistently and is able to predict low/no occupancy meter readings.

4 Conclusion and Recommendations

Clustering and consumption forecast serves as great tools to help electricity production and pricing planning. This study made attempts on each type of data analysis and have extracted some useful insights. However, the scope of this study is limited. For future studies, this section details the recommendations and improvements. If the below recommendations can be implemented for the electricity dataset, conducting the study on other meter data sets would paint a more holistic picture of building energy consumption.

4.1 Clustering Analysis

Recommendation for future work includes

- Varying the cluster numbers for different buildings to create custom fit; investigate on how well the algorithm captures dynamic building consumption profiles. As observed in some industry buildings in this study, some buildings have a wide range of daily load profiles. Maybe some buildings clusters can't be captured by even setting k equals to 20. They are just always dynamic and occupant dependent; and
- Conducting a clustering-based feature selection with K-means algorithm - clustering across different buildings to investigate how correlated that electricity consumption is with the building type, climate zone and other building characteristics. A principal component analysis would also be interesting as a follow up analysis to confirm the findings.

4.2 Forecast Analysis

Recommendation for future work includes:

- Creating scatter plot of the predicted and actual to see if the model is overall underpredicting or over-predicting ;
- Fine tuning the model performance by conducting a sensitivity on train-test split, hyper-parameter tuning, adjusting training and testing set dates for KNeighborsRegressor Model, and experimenting with different neural network architecture; and
- Further exploration into whether more weather information would improve forecasts.

Reference

[1] P. S. N. H. M. Alexander Tureczek, "Electricity consumption clustering using smart meter data," *Energies,* vol. 11, no. 4, p. 859, 2018.

[2] A. L. F. F. N. S. a. S. R. Yasirli Amri, "Analysis Clustering of Electricity Usage Profile Using K-Means Algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2016.

[3] U.S. Department of Energy , "Demand Response," [Online]. Available: https://www.energy.gov/oe/activities/technology-development/grid-modernization-and-smart-grid/demand-response. [Accessed 8 Feb 2021].

[4] A. K. B. P. Clatyon Miller, "The Building Data Genome Project 2,energy meter data from the ASHRAE Great Energy Predictor III competition," *Nature Publishing Group,* vol. 7, p. 368, oct 2020.

[5] F. M. Clayton Miller, "The Building Data Genome Project: An open, public data set from non-residential building electrical meters," Energy Procedia, 5 September 2017. [Online]. Available: https://github.com/buds-lab/the-building-data-genome-project. [Accessed Feburary 2021].

[6] G. L. ,. M. B. ,. P. ,. ,. M. ,. O. G. ,. V. N. ,. P. P. ,. A. G. J. G. ,. R. L. ,. J. V. ,. A. J. ,. B. H. ,. G. V. Lars Buitinck, "2.3. Clustering," 2011. [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means.

[7] Grupo de Aprendizaje Automático - Universidad Autónoma de Madrid Revision 2db7bb7b., "Scikit-fda," 2019. [Online]. Available: https://fda.readthedocs.io/en/latest/modules/ml/autosummary/skfda.ml.regression.KNeighborsRegressor.html.

[8] J. Brownlee, "Dropout Regularization in Deep Learning Models With Keras," 27 August 2020. [Online]. Available: https://machinelearningmastery.com/dropout-regularization-deep-learning-models-keras/. [Accessed 8 Feburary 2021].

[9] A. A. P. B. E. B. Z. C. C. C. G. S. C. A. D. J. D. M. D. S. G. I. G. A. H. G. I. M. I. R. J. Y. J. L. Martín Abadi, "TensorFlow Documentation," 2015. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/activations/relu. [Accessed 8 Feburary 2021].