Detailed Energy Analysis of Stever House

Raafe Karim Khan
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
+1 (317)-514-2097
rkk@andrew.cmu.edu

Saurabh Mishra
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
+1 (412)-961-6462
saurabhm@andrew.cmu.edu

Roja Malligarjunan
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
+1 (412)-708-9912
roja6055@gmail.com

ABSTRACT

In this paper, we will be looking at a Carnegie Mellon campus building's energy consumption, Stever House, for one year and making suitable recommendations based on the dataset. The paper analyzes energy data by hour, week and day to understand occupant consumption patterns by correlating it with temperature data for the same building i.e. Stever House, and then making suitable recommendations which the Facility Management Services group can implement to lower consumption at peak load, thereby reducing costs.

CCS Concepts

• Mathematics of Computing →Probability & Statistics • Information Systems→ Data Management Systems

Keywords

Energy; Power; Temperature; Dataset; Data; Regression; Tree

1. INTRODUCTION

In the United States, buildings consume about 40% of the total energy generated, which translates to about 39 Quadrillion BTU [2]. To meet future demand, it is necessary for us to understand why, where and how we consume energy in households to reduce peak loads, cut costs and reduce our respective carbon footprint. By understanding consumption patterns, we can predict when demand for electricity will rise and fall based on time of day and then subsequently look at measures which will aid in reducing consumption without depriving consumers for energy or implicitly lowering their current standard of living.

2. PROPOSED APPROACH

The approach of the paper is very simple; we have aggregated power consumption data for a year which we will disaggregate to conduct an analysis which will give us an insight about consumption patterns. The building in question is an undergraduate dormitory which is located at the intersection of Forbes and Morewood Avenue. First we analyze data by every week of the day after resampling for hourly intervals and then subsequently try clustering using a k-means approach and then run a linear regression after normalizing the data. In our approach, the response variable is power in Watts, whereas the key feature variable is temperature in Fahrenheit. In a nutshell, we are looking to conduct exploratory data analysis by seeking what answers the data can give us about occupants' energy consumption with respect to time and temperature.

3. DATASET

The dataset was procured from the buildings meter and portfolio manager could give us temperature values for every instance of data recorded thanks to a sensor system present at the top of the building. The dataset doesn't have equal intervals even though it records a data point every minute so resampling the data every hour was a natural choice in analyzing the data. The dataset has 479,836 rows and 3 columns. The dataset was initially in .txt format and was subsequently converted to a .csv format for ease of conducting analysis. The data points in the data frame were of type string, so using some pandas function we converted all string values to float32 values. Based on the sensors readings, the mean power consumption of the dataset was 110,127.46 W and the mean temperature was 57.23 F. The standard deviation of the power consumed was 34059.13 W, whereas the standard deviation of the temperature was 18.29 F.

4. RESULTS

In the figure below, we can see how the power consumption varies over days of the week. The boxplots help us see the degree of seasonality in the data set. Per our analysis of the box plots, we see some degree of sinusoidal nature in power consumption over the course of the day for every day of the week.

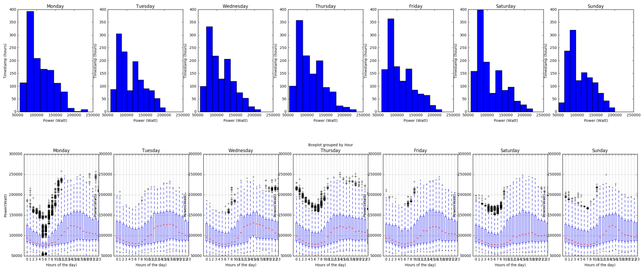

Figure 1. Histograms and boxplots for every day of the week

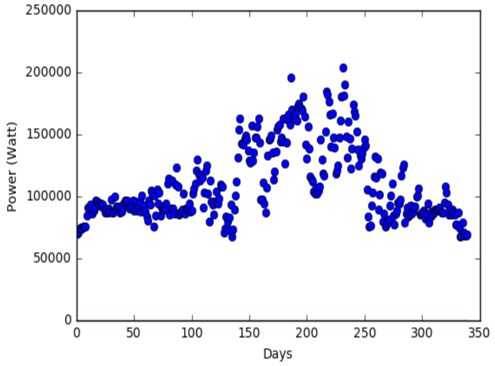

Figure 2. Stem plot for power consumption for the entire year

The above figure illustrates the average hourly consumption over the course of an entire year for four seasons: winter, spring, summer, and fall. The global pattern repeats across seasons, although the magnitude of the peaks and the average usage is higher in warmer seasons than in colder ones. The figure also exposes higher energy usage in peak summer and winter months, which signpost the use of air conditioners (ACs) and electric heaters, respectively. The figure reveals higher energy usage due to the use of gas heaters in colder months. The weekday split of data reveals higher usage during the end of the week-Friday and Saturday. Sundays consume lower energy consumption as students normally take off and are hardly found in the dorms. One of the challenges in this analysis was how to deal with null values post resampling the data-frame. After resampling by the hour, we interpolated the data-frame so that null values can be replaced with some degree of meaningful data points.

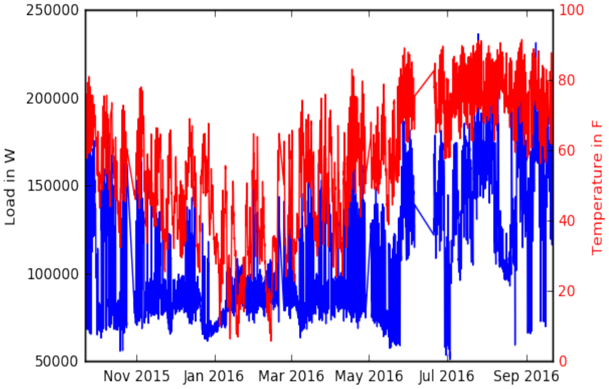

Figure 3. Power and temperature vs time

In figure 3 we see how power and temperature respond as a function of time. To the observer it may seem that there is some degree of correlation, however in subsequent analysis we will be able to determine what the degree of correlation is.

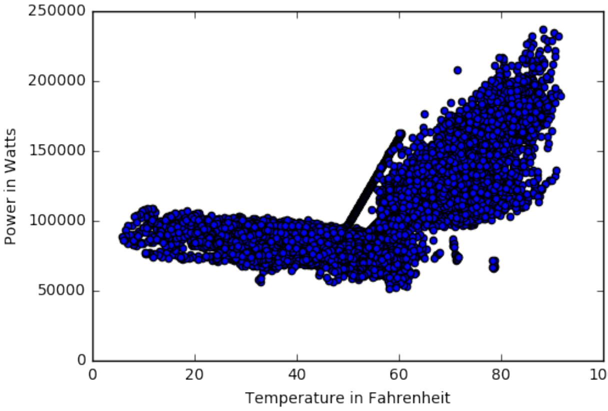

Figure 4. Power vs temperature

If we look at figure 4, we see that for lower temperatures the power consumed is typically within a range. In a way, we can predict that the primary source of energy for heating is natural gas. At higher temperatures, typically above 60 F we see that there is a distinguishable increase in power consumption, which translates to electric cooling being switched on.

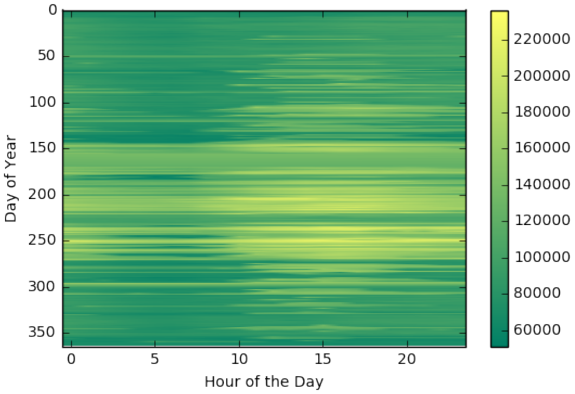

Figure 5. Power heat-map

In figure 5, we can see that towards the mid part of the year and towards the last half of the day, the consumption typically increases, which is evidenced by the yellow region in the figure. The dataset had to be unstacked and grouped by the hour to get the summer heat map. In figure 1, the box plot on the entire data set reveals that the median shifts over time and closely resembles a sine curve. The box plot also shows the number of outliers for each hour of the day. The 24th hour of the day has many more outliers compared to other periods through the course of the day.

Table 1. Brief description of the data frame

Statistical Value	Power (W)	Temperature (F)
Count	8785	8785
Mean	110,127.47	57.23
Standard Deviation	34,059.13	18.28
Minimum	51,127.50	5.86
Maximum	236,427.86	91.62
25%	82,573.60	44.68
50%	98,790.24	58.48
75%	133,541.44	72.38

4.1 Clustering

In this section, we have used an unsupervised algorithm and defined three cluster centroids, one for each cluster. This is an iterative process which follows the following equation:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \parallel x_i^j - c_j \parallel_R^2 [1]$$

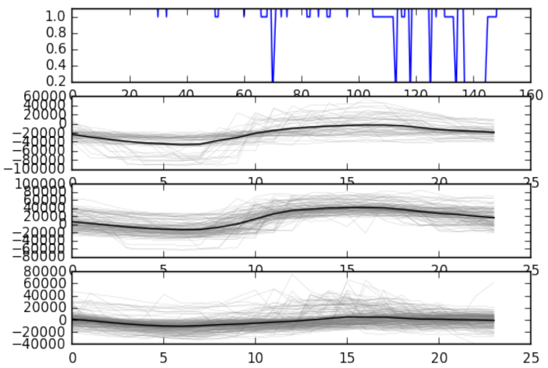

Figure 6. K-means clustering output

From figure 6 we see that the three clusters share a similar pattern in consumption of power. During the first 10 hours of the day, there is a depression in power consumption as evidenced above, however after the 10th hour we see a slight rise in power consumption till the 23rd hour of the day.

4.2 Regression

Using *sklearn*, a library in python for data analysis, we conducted a linear regression. The conditions used for regression were that minimum split samples were 20 with 99 random states and 2 sample leaves.

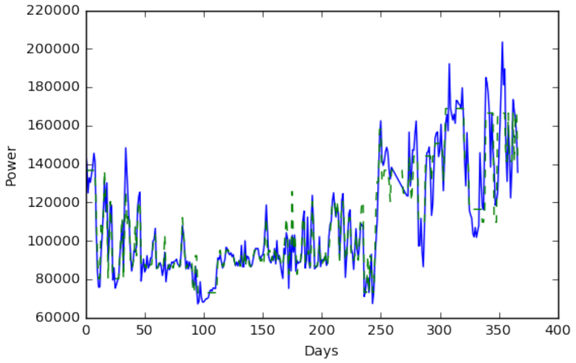

Figure 7. Prediction and training data points after regression

In figure 7 we see how the predicted and expected model look like after regression. The blue line represents the expected data while the dotted green line represents predicted data. We found the R^2 value to be 91.27%. From the regression plot we can see that the predicted plot is more closely aligned with the training data at lower temperatures, however at higher temperatures, we do miss the peaks sometimes.

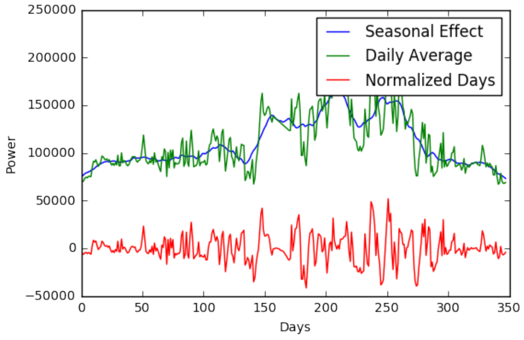

Figure 8. Regression output

In the plot above we can see seasonal effects in the data set as well. However, since we are interested in consumption patterns on weekends and weekdays, we construct a low pass filter to remove seasonality in the data set. The low pass filter is computed over ten days. Furthermore, we plot the normalized power consumption through the year and the daily average as well. As expected, the daily average is simply an amplified version of the normalized curve. As expected, seasonality does effect power consumption because occupant comfort and preferences vary as a function of personal physical characteristics and the un-manipulated feature of temperature.

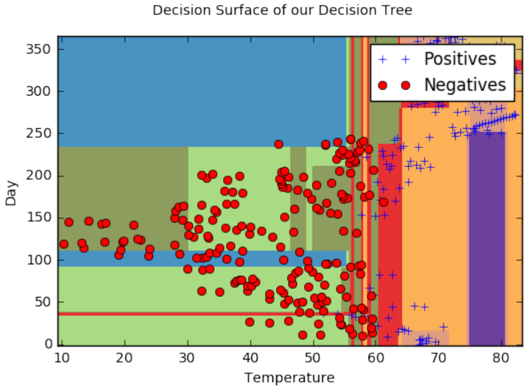

Figure 9. Decision surface of the decision tree

To be able to get this decision tree we first pre-computed the range of the features. In this case the features were temperature and day of the year. Subsequently, we created a mesh grid and then predicted each cell in the mesh. After this step, we plotted the training points and then created the plot in figure 9. From the figure, above we can see how the tree has been split and where the split occurred. As a preliminary step, we split the data frame into two distinct groups. One group had power consumption values less than the median power, while the other group has power consumption values greater than the median. The values assigned to these respective groups were 1 and -1. This is how we generate the positive and negative points in the plot. We can infer from the plot above that power consumption value below the median are typically found at lower temperatures, whereas as the ones with power consumption values higher than the median are located at temperatures above 60 F. As we can see there is some degree of dispersion of data points, which typically takes place after a temperature of 30 F.

5. VALIDATION

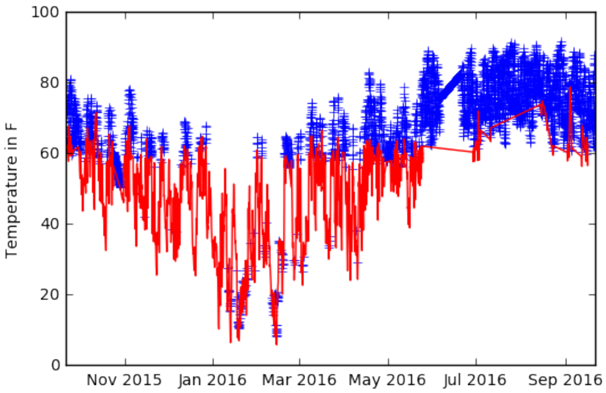

Figure 10. Post-split plot

As mentioned previously, we split the original data frame into two distinct groups based on the median power consumption. Furthermore, we plotted how the power in the two data frames respond to changes in temperature. We see that in figure 10 that there is a great amount of variance i.e. fluctuation in temperature. The pattern is erratic in nature and is most notable in the period between Jan 2016 and Mar 2016. However, as we move to temperatures above 60 F, the variance in temperature is significantly less and power consumption remains above the median for most of the time. The pattern is more stable and we conclude that occupants in Stever house use cooling loads for extended periods of time with little control over the inside temperature. Our building audit confirmed that the heating and cooling is centrally controlled.

To understand this phenomenon better we split the data frame further by power and temperature through the median of these respective variables. We found that in the data frame split by temperature, at temperatures above the median, the co-relation with power is 62.71%, whereas at temperatures below the median, the co-relation is -18.83%. This validates the fact that at higher temperatures, there is more co-relation with power i.e. lesser variance over time and a higher probability that the model will be able to predict accurately. Similarly, for power, for values of power above the median, the co-relation coefficient is 56.73%, whereas for power values below the median the co-relation coefficient is -31.53%.

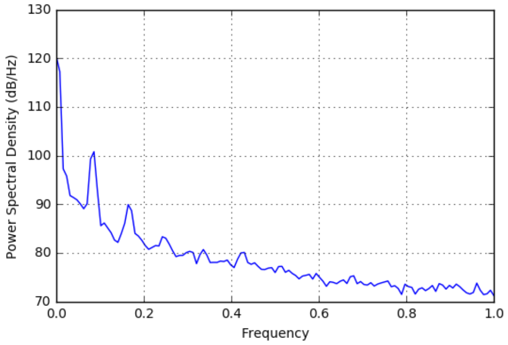

Figure 11. Power spectral density plot for power

In figure 11 however, we see that at lower frequencies, the power spectral density is higher, and as we move towards higher frequencies, the power spectral density decreases almost inversely. It looks like a high pass filter in many aspects.

In the violin plot we can re-iterate our previous claims. Violin plots are like box and whisker plots but they offer a greater sense of variability in the dataset. We can see that there is greater variability at lower values of power, which is typically at lower temperatures rather than at higher temperatures.

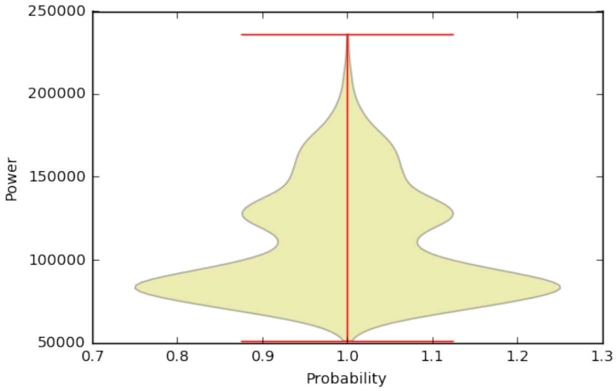

Figure 12. Violin plot for power

6. DISCUSSION

Since the data belongs to a dorm house on the campus, we were expecting it to show a regular power consumption pattern and our analysis does show this in the k means clustering analysis where all the three clusters share a similar pattern during the day. Furthermore, there was a high correlation between power and temperature during summer while during the winter, due to a separate heating mechanism, power shows less correlation with temperature. The consumption is above the median in summer while clearly below the median during the winter i.e. when the temp is less than 60F. We see greater variability at lower temperatures i.e. less than 60 F compared to higher temperatures. Per the decision tree, most of the time of the year the temperature lies between 40 F and 60 F with power consumption below the median. During the winter, there is more variation in temperature, whereas in the summer there is lesser variation. Interestingly, we find that despite the presence of a heating mechanism, power consumption is above the median for select days, which could als0 be outliers.

We also find that disaggregating the data give us more insight into the consumption pattern than aggregated data.

7. FUTURE WORK

In subsequent work, we plan on incorporating a larger dataset with more feature variables so that prediction can be made more accurately, thus making the algorithm robust. In the future, we would also like to conduct more extensive exploratory data analysis to see what other answers the data set can give us about occupant energy consumption. One of our endeavors will be to make a generic algorithm which will help us analyze building data and convert it into a standalone application for use by facility managers.

8. REFERENCES

[1] MATHWORKS, "Clusterdata." K-means Clustering - MATLAB Kmeans. Accessed December 09, 2016. https://www.mathworks.com/help/stats/kmeans.html.

[2] United States Energy Information Administration (EIA), ‘How much energy is consumed in residential and commercial buildings in the United States?’. Updated April 2016. Accessed Dec 2016. http://www.eia.gov/tools/faqs/faq.cfm?id=86&t=1