City of Boston: Energy Consumption

Group 9:

Neha Gilson

Aastha Grover

Shuxian Wu

INFO 7390 ADS

Table of Contents

[City of Boston: Energy Consumption 3](#_Toc448669393)

[**Date Cleansing** 4](#_Toc448669394)

[Data Visualization 6](#_Toc448669395)

[Model 6](#_Toc448669396)

[References 8](#_Toc448669397)

City of Boston: Energy Consumption

Problem Statement:

* How can the current energy consumption data be used to monitor, plan and predict the energy usage for the future.
* How can this data be used by business to save money on their energy bills?

Goal: With this project we aim to

* Show the trends in energy consumption by different facilities owned by City of Boston.
* Cluster the accounts based on power usage and type (Commercial, Educational etc.)
* Analyze which facilities are the best and worst energy consumers in each cluster.
* Recommend which facilities need to be monitored and provide suggestion for optimal usage of energy.

Application use case

* Using the trends, predict future usage and send it to the provider for deciding energy distribution at the facilities on an hourly and daily basis.
* Option to view usage trends on monthly, weekly and daily basis.
* Peak hour analysis of daily power consumption.
* Application can be used to study the Bad, Good and Perfect power factor for a particular account.

Work Flow:

**Date Cleansing**

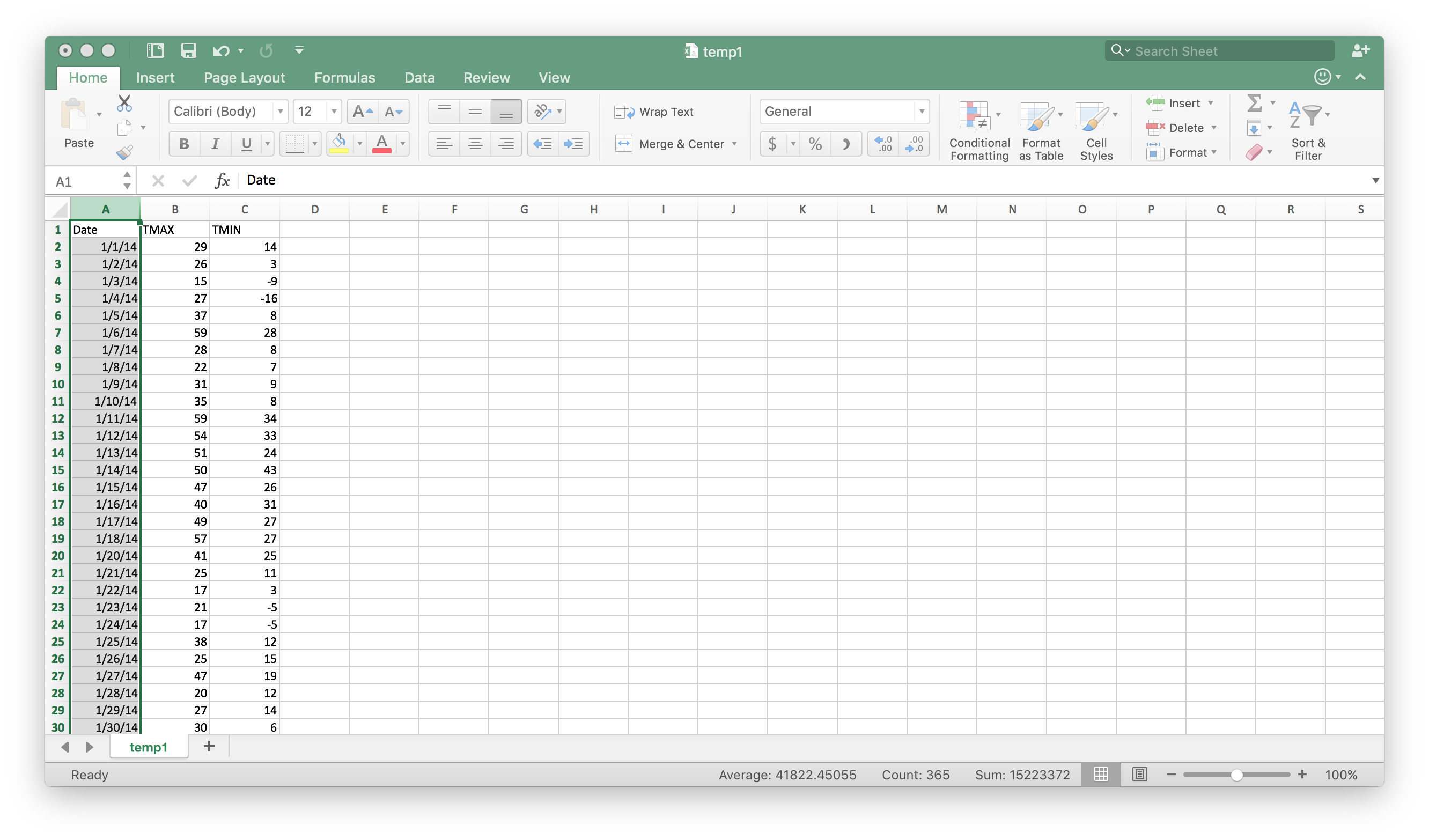
Before preprocessing our data, we did a little research about it and discussed data features we want in our final table. Since we want to do clustering and prediction in our analysis, we separated columns by categorizing them in to “Must have”, “Keep for later” and “features from outside source”.

For “Must have”, there are columns including Date, Hour, hourly Power Factor, hourly kWh, hourly kVarh, Working day, Holiday and Category. For “Keep for later”, we have Day, Month, Account Number and Name of the facilities. Also, we merged weather data from outside source like Wunderground and NOAA, and used it as Temperature MAX, and Temperature Min.

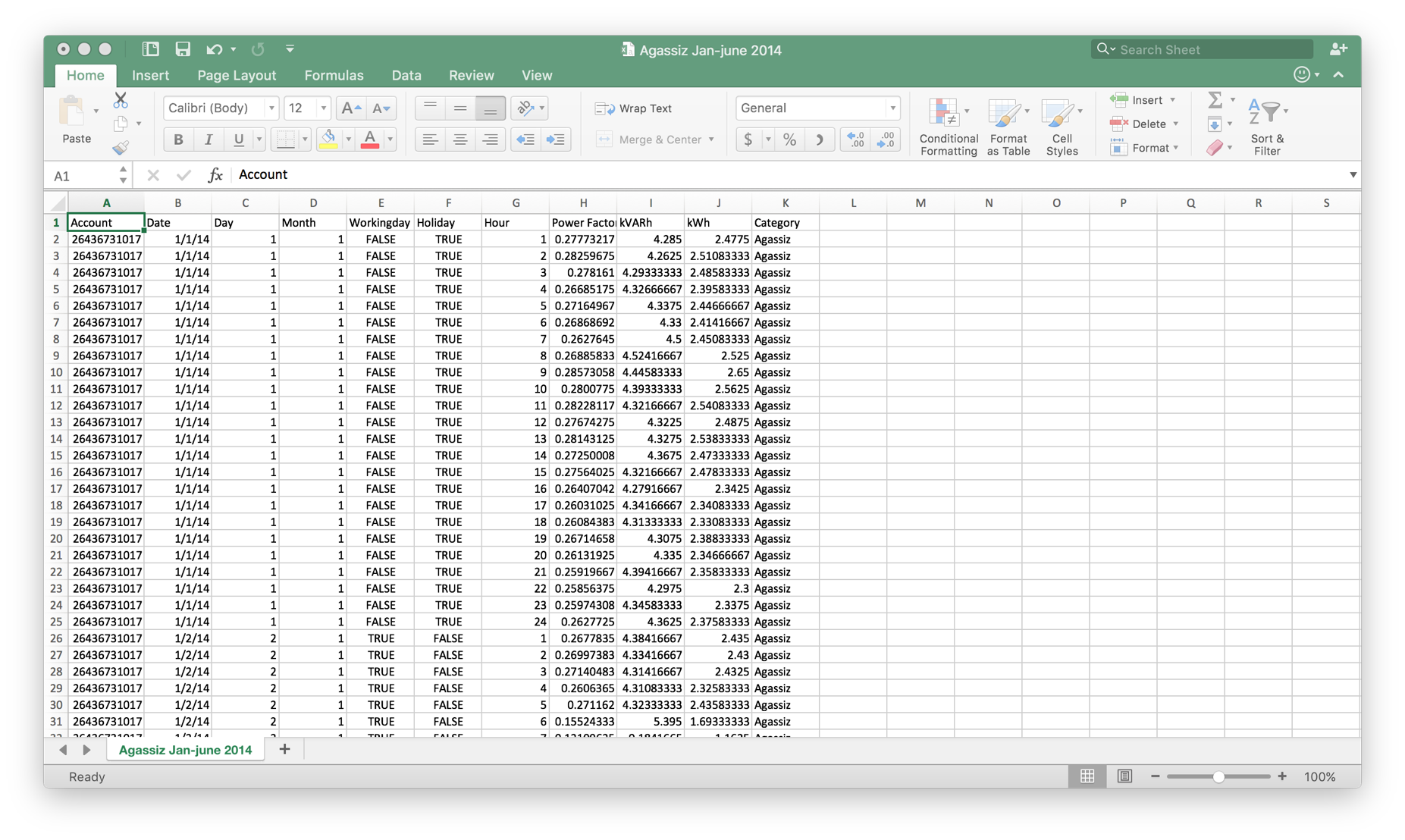
The dataset we are given is noisy with many problems and we used Python 3 with Jupyter Notebook for cleansing it. There are 73 csv files in our dataset. Some of them have N/A value in cells, some of them have duplicates rows, some of them have timestamp for 5 minutes or 15 minutes etc. With Python 3, we mainly use Pandas, Numpy and Datetime modules for processing data.

During the cleansing procedure, basically, we built functions to read csv files, clean function to deal with all the problems with data, such as duplicates, N/A value, Transposing, adding information, merging and concatenating. Our ipython notebook includes more details about the codes we wrote. Please refer to CobEnergyDataClean.ipython file. Below are some screenshots for datasets we using:

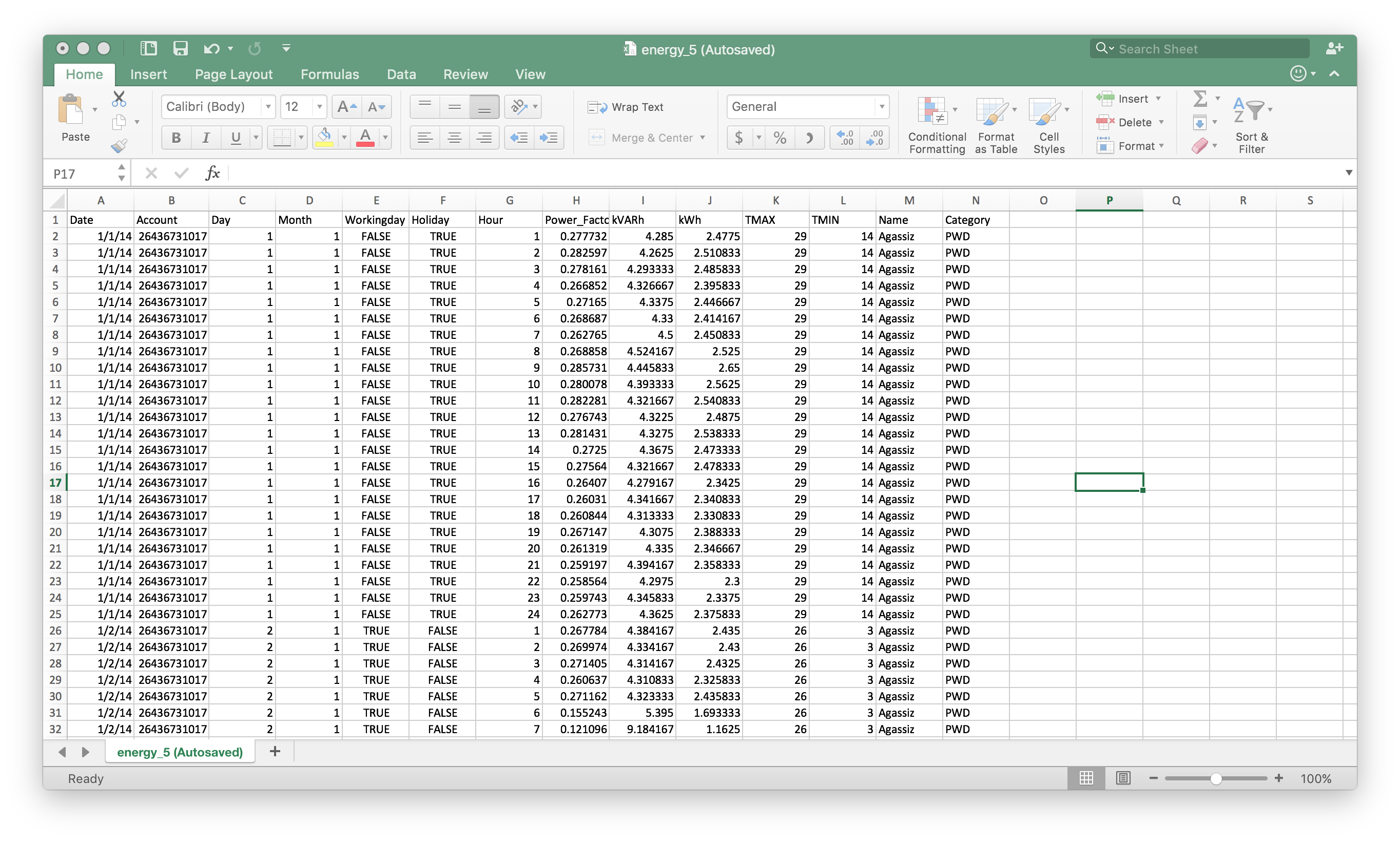
Historical weather data with Max and Min temperature for City of Boston, Year 2014:



CoB Datasets Example Before Merging:



Final Dataset:



Keywords: Python 3, Data Cleansing, Pandas, Numpy, NOAA

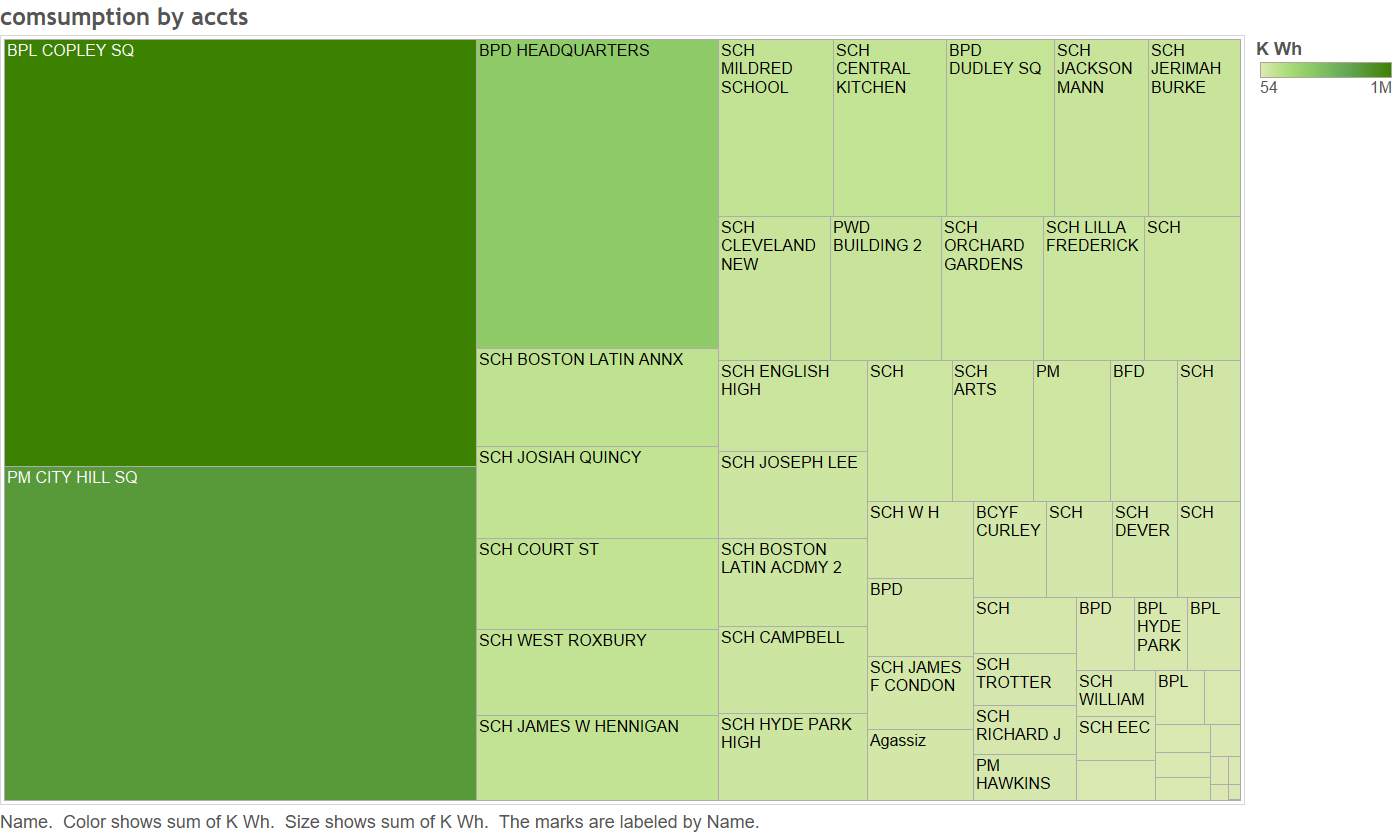
## Data Visualization

We used Tableau to perform data visualization and Tableau Public to share this visualization.

Below are our observations.

Link to tableau Public: <https://public.tableau.com/profile/shuxian.wu#!/vizhome/Final_131/Story1>

Image 1: This figure shows the total energy consumption by different accounts in the year 2014. As seen, in the heat map, Boston Public Library at Copley square has the maximum utilization. When we investigated this further we found that the higher the consumption, higher is the reactive power for such accounts. For such accounts where the KWh and Power factors are high we would recommend to have a lower Reactive power (KVarh). In such a scenario, BPD Dudley square shows a better result as seen in the second image



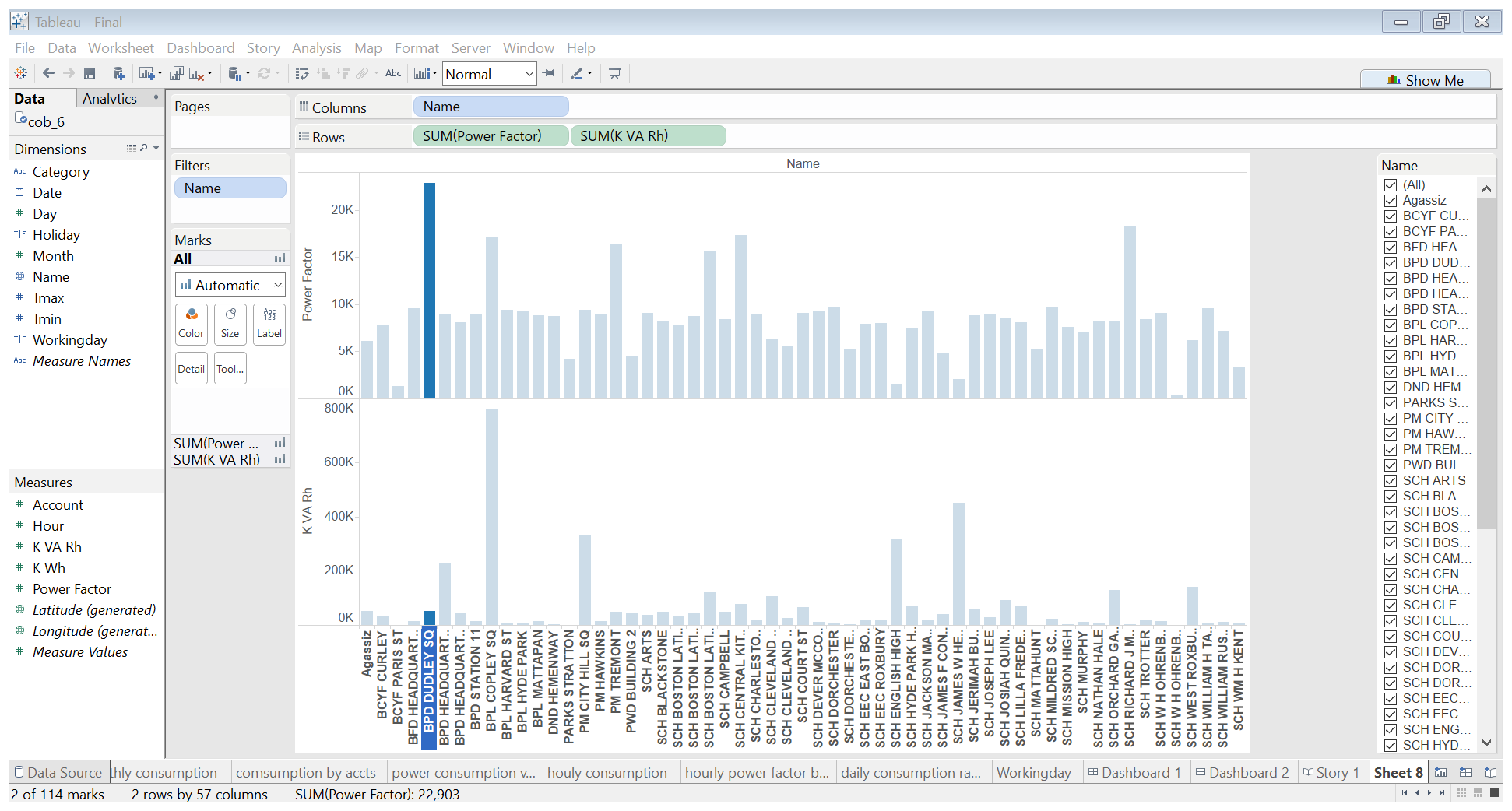


Image 3:

In the below image we tried to visualize the hourly power consumption by all accounts. And as expected, this reveals a curve that shows peaks during the working hours of the day and lows in the off-work hours of the day.

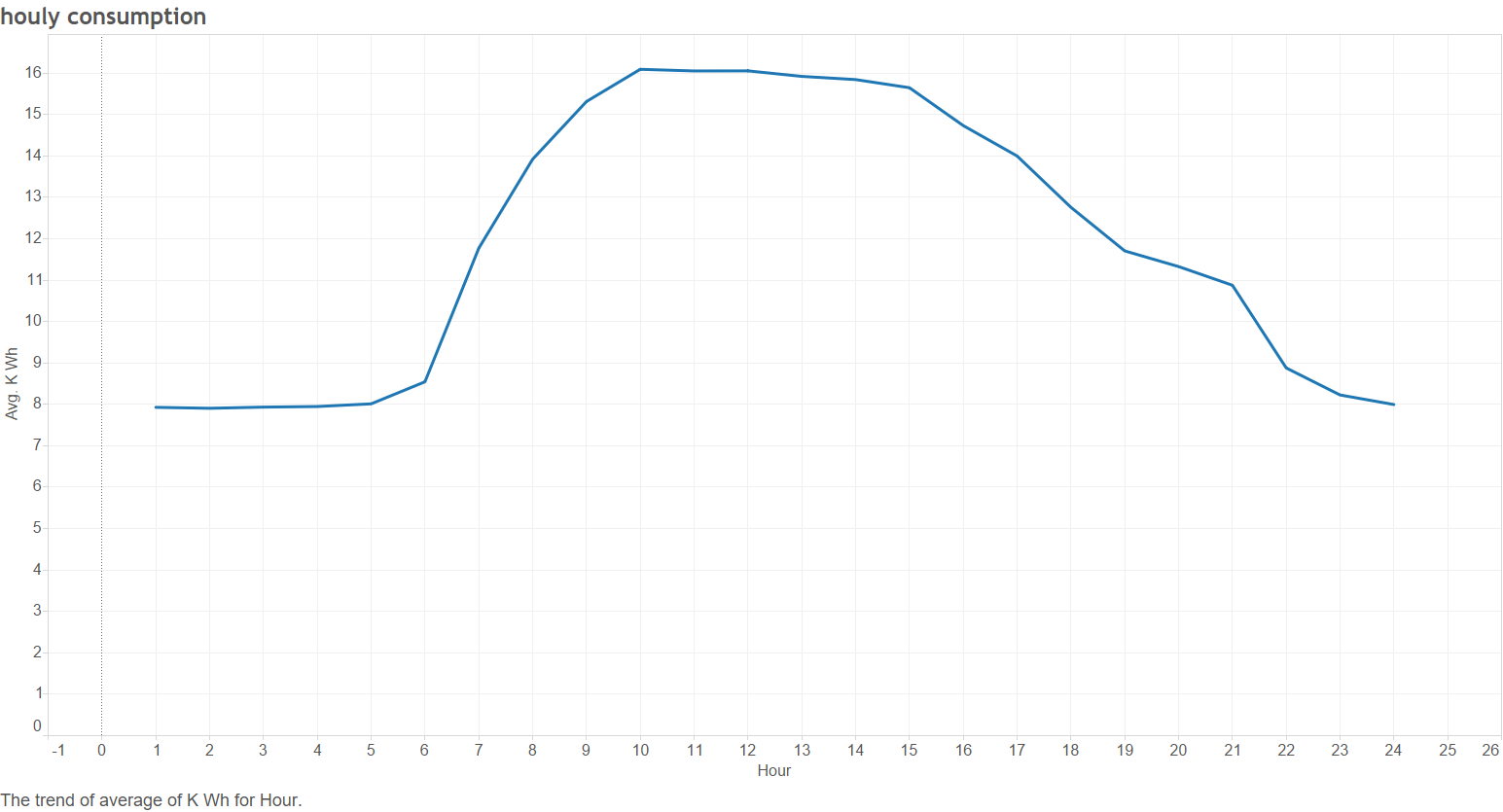


Image 4:

The below image shows the daily power consumption in a month by various facilities. We can see that summer months show a higher range in the energy utilization.

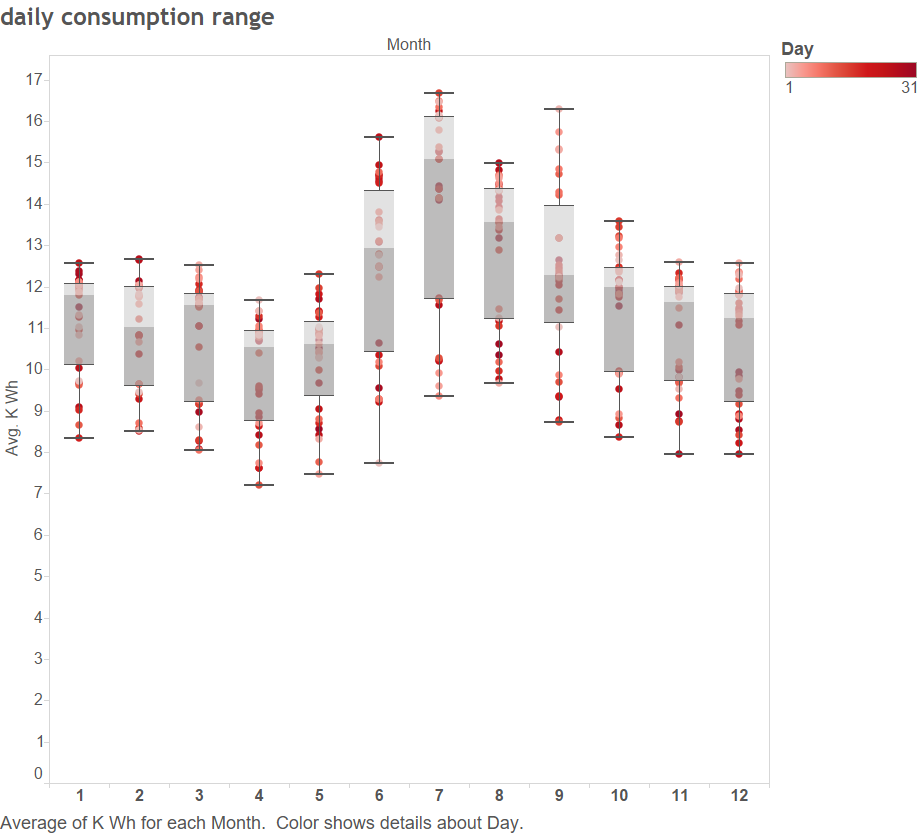


Image 5:

The below image shows the relationship between temperature and energy consumption. We can see a bowl shape being formed. This indicates that whenever the temperature goes above or below average the consumption increases. The blue patches indicate a higher than average consumption of power. Also, we understand that it requires more power to keep a place cool that warm. Therefore, the usage increase as the temperature increases.

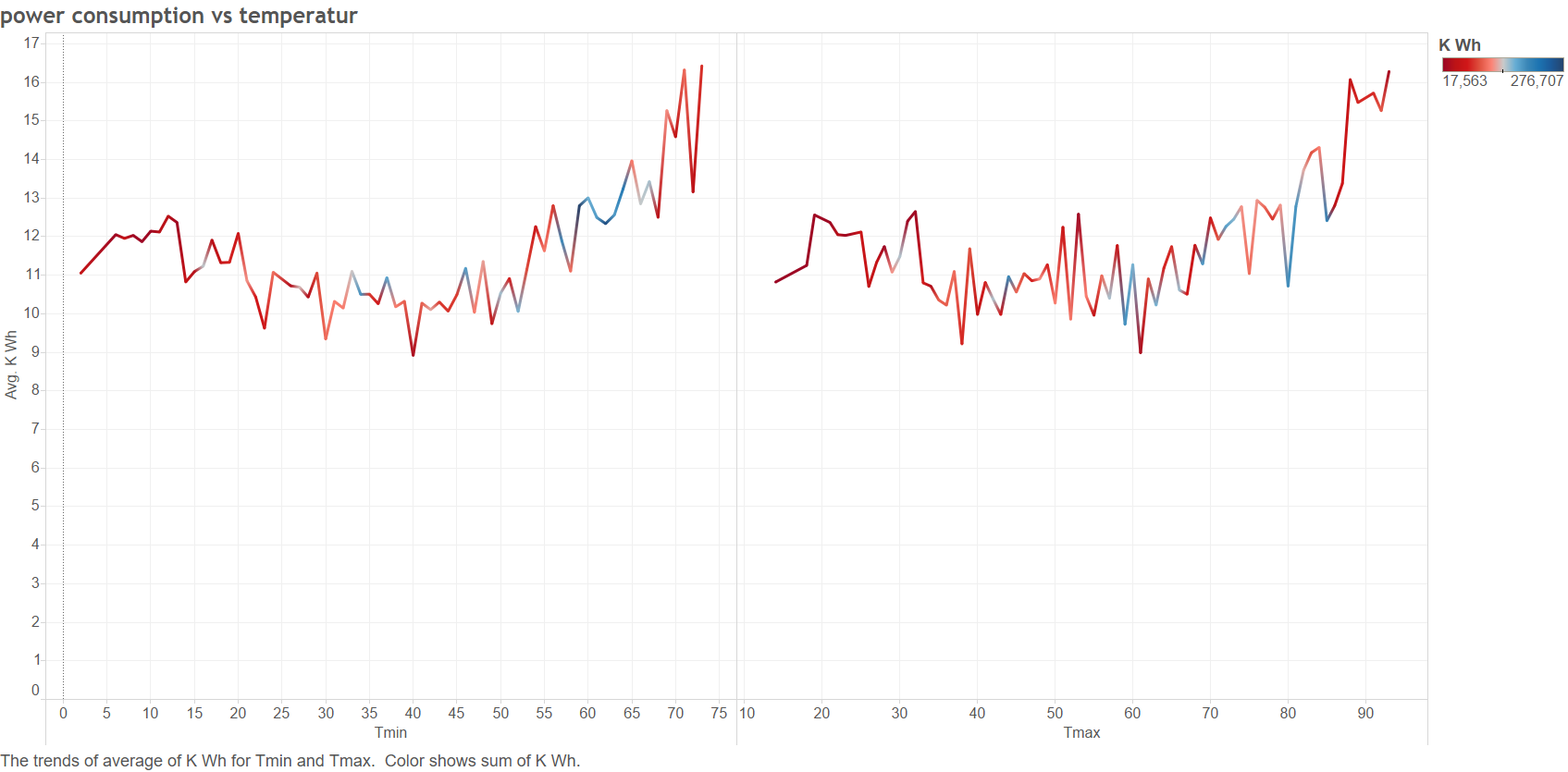
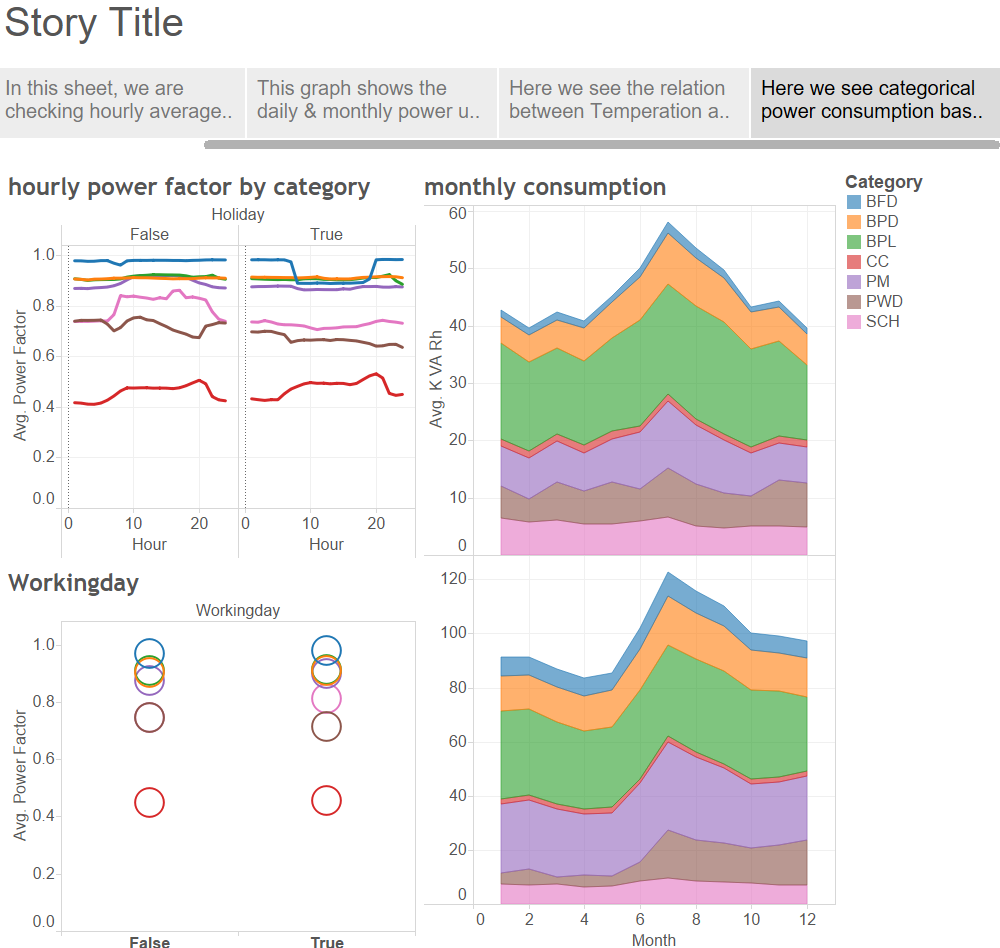


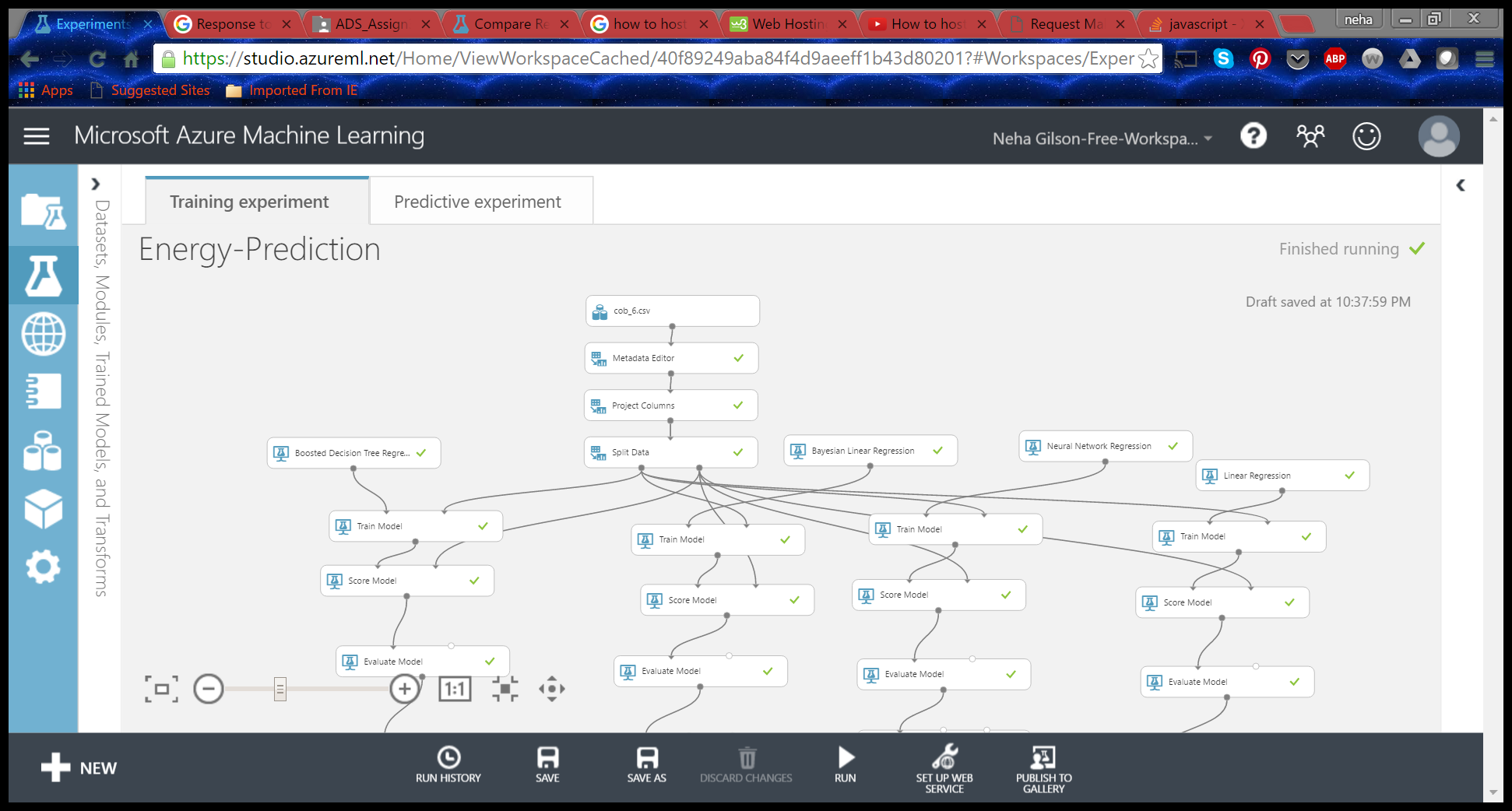
Image 6:

This image tells us power usage by different categories on a working day, holiday and a monthly basis. We see that there is a dip in the power factor by the schools during the holidays which indicates that the schools are unable to utilize the power supplied during the holidays efficiently. Also since they are closed, their usage decreases. We also see that Libraries and Project Management sites as a whole take up major portion of the 3rd graph, indicating a better power factor. We also see that community centers, parks and other public and youth facilities don’t show a major difference in usage which is a little surprising as you would expect them to be more active.



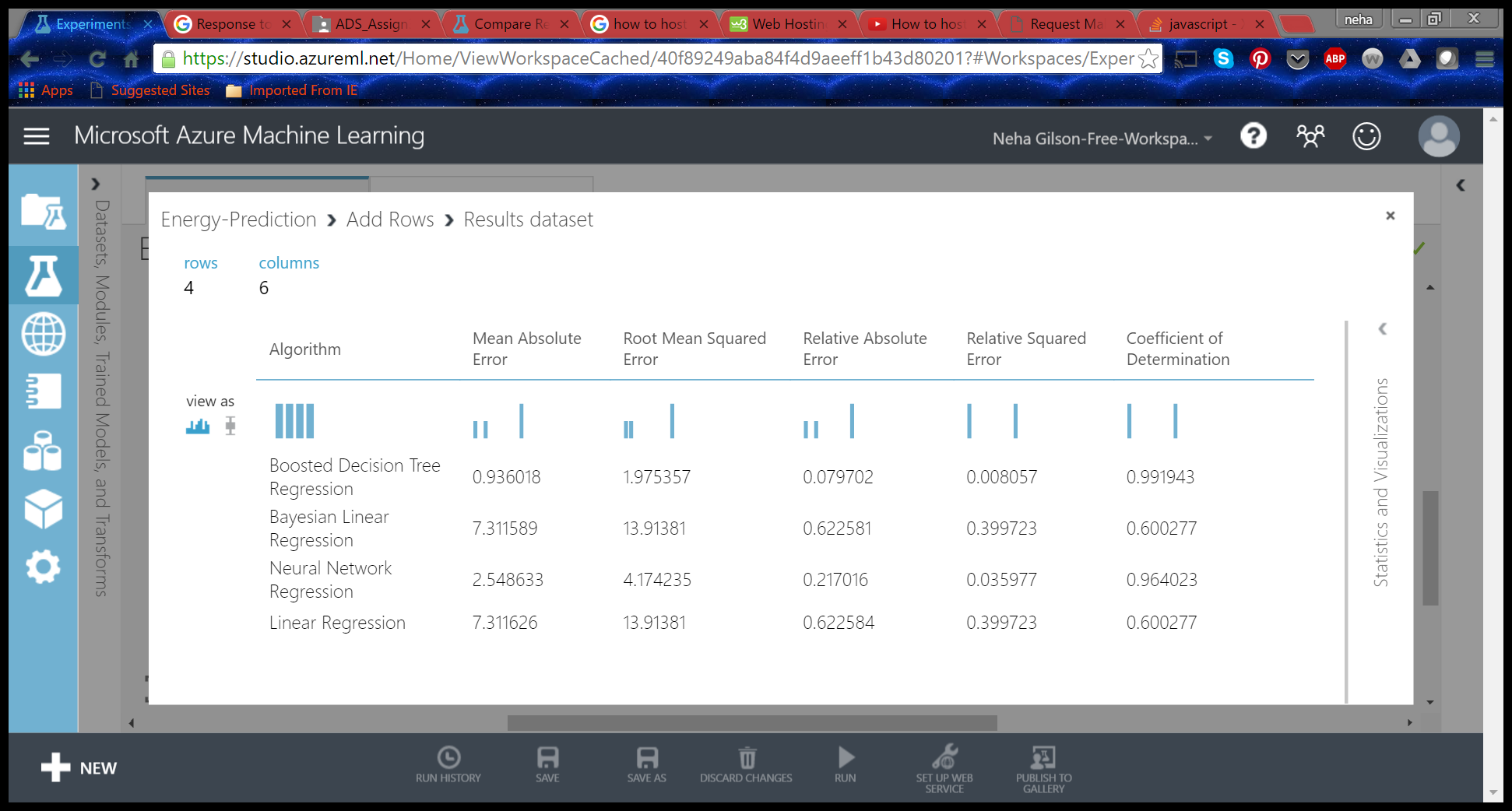
Model

We used Azure Machine Learning Studio to create our prediction and clustering model. We used the built-in functionalities and algorithms available in AML. The regression model we built is shown below. We used multiple algorithms like Boosted Decision Tree, Neural Network Regression and Bayesian Linear regression algorithms. Upon comparing the results of the models, we found that Neural Network gives the best prediction.

****

**The error metrics are as shown below:**

We see that even though Boosted Decision tree regression gives us better results, it indicated the problem of over-fitting. So we found that Neural networks give us a better prediction model with a coefficient of determination of **0.96 and relatively low error rate of 21%.**

****

**We then deployed our model and generated the URL and API key for integration.**

**Recommendations:**

1. **We recommend that Boston public library at Copley square and Project Management facility at City Hill be monitored for its power consumption. The equipment used should be evaluated and replaced if necessary.**

References

EEMS Energy Consumption Assessment Report Final\_tcm3-33503.pdf

Wunderground.com: https://www.wunderground.com/weather/api

NOAA: <https://www.ncdc.noaa.gov/cdo-web/datasets#GHCND>

Power Factor: <https://en.wikipedia.org/wiki/Power_factor>

A guide to reactive power-EDF Energy: <https://www.edfenergy.com/sites/default/files/b2b-guide_to_reactive_power.pdf>