Monthly Electric Consumption

Kevin Eng

4/28/2020

Introduction

The goal of this analysis is to explore factors which influence electric consumption at buildings within New York City. To accomplish this, we collect three data sets: one which describes the surrounding weather, one which measures a buildings enery consumption, and one which desribees the physical characteristics of the building.

Ideally, in order to seperate electric consumption due to seasonal trends, hourly data would have been preferable. Indeed the inductive basis for making conclusions with high temporal resolution data would be stronger since we would not have to worry about confounding with say energy differences due to holidays. Alas the highest resolution data set that was found was monthly.

Data

Historical weather data was collected off NOAA's database. Fortunately, they have a simple API which allows for quick querying of basic weather data. A particular aspect of their database is that NOAA only offers access to select daily weather summeries. Hourly, and monthly averages are only given as "normal" averages which the NOAA defines as 30-year averages. The observations include daily minimum temperature, maximum temperature, perciptation, snow fall, and average wind speed.

Information on the physical characteristics of a collection of building complexes are recorded in the NYCHA data set. The dataset also contains a number of administrative details such as whether or not it is a senior development. In database venacular the primary key of this data set is the TDS number. Since the TDS number refers to a building complex is it not possible to identify sub-units within a complex.

Details on the monthly energy consumption can be found in the NY open data website. the website claims the data set only contains readings from 2010 up to March 2019. This is not quite true since in reality it contains readings up to September 2019. The monthly readings are given for an individual meter. This means that there are several readings for a given TDS number since each sub-unit within a complex has its own meter.

Data Processing

All three data sets needed some preprocessing in order to get into tidy form. Since the electric consumption measurements are given on a monthly basis, daily weather readings must be aggregated into monthly data. This can be accomplished by extracting the month and and year from each date record using the handy `year()` and `month()` function from the `lubridate` package. Additionally, it would also nice to have some notion of the typical monthly temperature. We can estimate this by computing the median of the monthly average maximum and minimum temperature using purr's `map()` function.

The NYCHA data set contains a great deal of unwanted administrative details. And as we will see, the data is also non-tidy as some rows contain multiple observations. On a somewhat cosemtic level, the naming convention for the columns is overly verbose and makes for combersome data referencing. We can easily fix these issues using dplyr's `select()` and `rename()` functions. There `TDS` column is somewhat problematic because it is a source of multiple observations and possible data entry errors. It likely that some building designs were reused so several apartment building share the same physical makeup on paper. In regards to possible data entry error, several rows contain `*`'s. To seperate the multiple observations we can use the `separate_row()` function which allows us to split a row into multiple rows based on a delimiter. The `*`'s can be easily dealt with using `str_replace()`. The last issue concerns the column that specifies the number of stories in a complex. Since we do not have the data resulution to seperate sub-units within building complexes we must find a way to reduce the number of recorded stories into one value. An easy approach is to take the mean of all the stories. This can be accomplished by first splitting the recording using `str_split` and then using `map_dbl` to produce the mean.

The electric consumption data, like the weather data, provides higher resolution data then can be used. In order to conform to the NYCHA data, total electric consumption and costs must be grouped by TDS which can be done in a similiar fashion as the weather data. Minor data issues were present where several rows contained missing TDS values. These were easily dropped from the table using `drop_na`.

Finally all three data sets were merged using `inner_join`'s. The weather data was joined to the electricity data using the date. Next, the resulting merged table was joined with the NYCHA data using TDS. Because we used `inner_join`'s the final table is complete and because we set each table to the same "scale" the final table is tidy.

Exploratory Data Analysis

To get a rough idea what what annual energy consumption over time looks like, we group the data by data by year and month and compute the total energy consumption. From there we can create a plot of super imposed monthly energy consumption curves over several years.

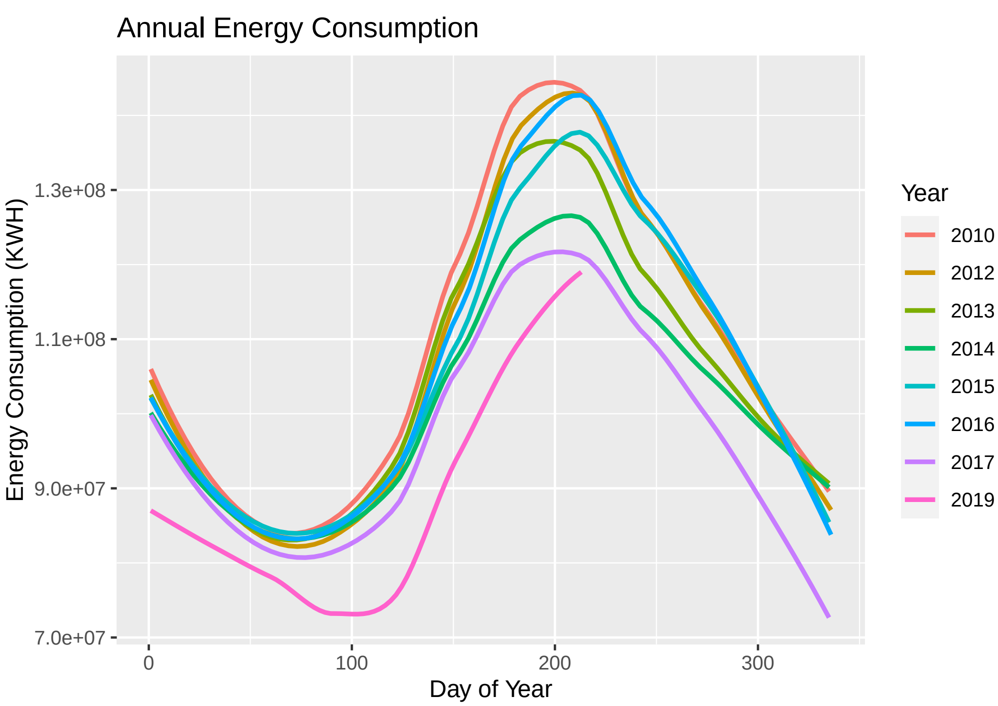


Figure 1:

The visualization shows a striking amount of variation during the summer months. In an optimistic take one might attribute the apparent decrease in electrical consumption to more energy efficient technologies.

Another interesting feature is the magnitutde if the amplitude for each curve. Table ? provides a overview between the month with lowest and highest energy consumption

YEAR	MIN_CONSUMP	MAX_CONSUMP	RATIO
2010	86933325	153667027	1.767642
2012	83199562	156132505	1.876602
2013	83176171	157325627	1.891475
2014	81760396	133032444	1.627101
2015	82900561	147481206	1.779013
2016	83038002	145053870	1.746837
2017	77949025	139148679	1.785124

The ratio indicates there is significant seasonal flucation in power consumption. Likely, the summer months correspond to increased A/C usage. Theoretically, electric heaters could be used during winter months, but a majority of buildings use some sort of radiator system. The stark contrast in power consumption makes it easy to see why, perhaps, during the summer months rolling black outs occur; parts of the eletric grid could be handling close to twice the load.

The goal is now to investigate the why power consumption fluctuates during the summer. One avenue of thought is that improving electrical efficiencies of electrical appliances, or better insulated buildings contribute to fluctuations in summer power consumption. To explore these ideas we need a normalized metric such as KWH per person from which we can compare its distribution across the years. And since we are comparing the distribution of power consumption over the years, the readings for 2019 should be discarded because it is incomplete. Any estimate of the distribution for 2019 will be necessarily screwed due to the missing values.

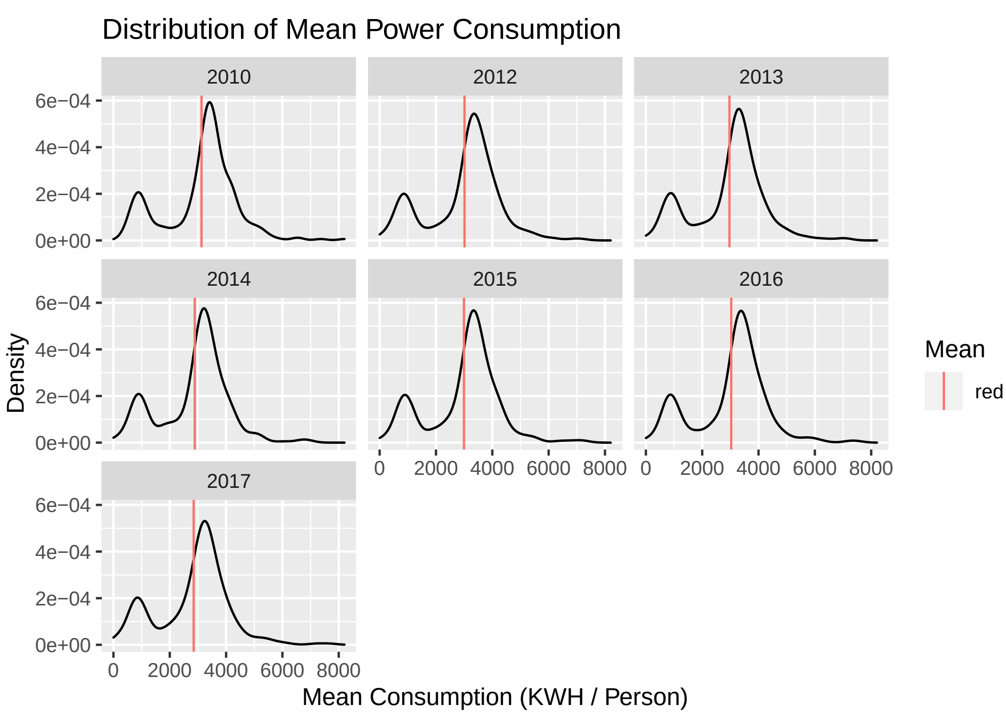


Figure 2:

The apparent bimodal nature of the distributions is a bit perplexing as there is no obvious reason for it. A casual glance at the overall shape of the distributions suggests that a gaussian mixture model may be useful in breaking down the distribution into two groups. For simplicity we choose a gaussian mixture model with two clusters which we can estimate using `normalmixEM` from the `mixtools` package. The `modelr` package allows us to easily group each model; by using `nest()` we can associate each year with a particular mixture model estimate. Figure ? shows each mixture model estimate super imposed on top of the non parametric density estimate. From the mismatch, we can see our data has lighter tails resulting in taller peaks, but the estimated models seem perfectly adquate for exploratory purposes so we proceed with using it.

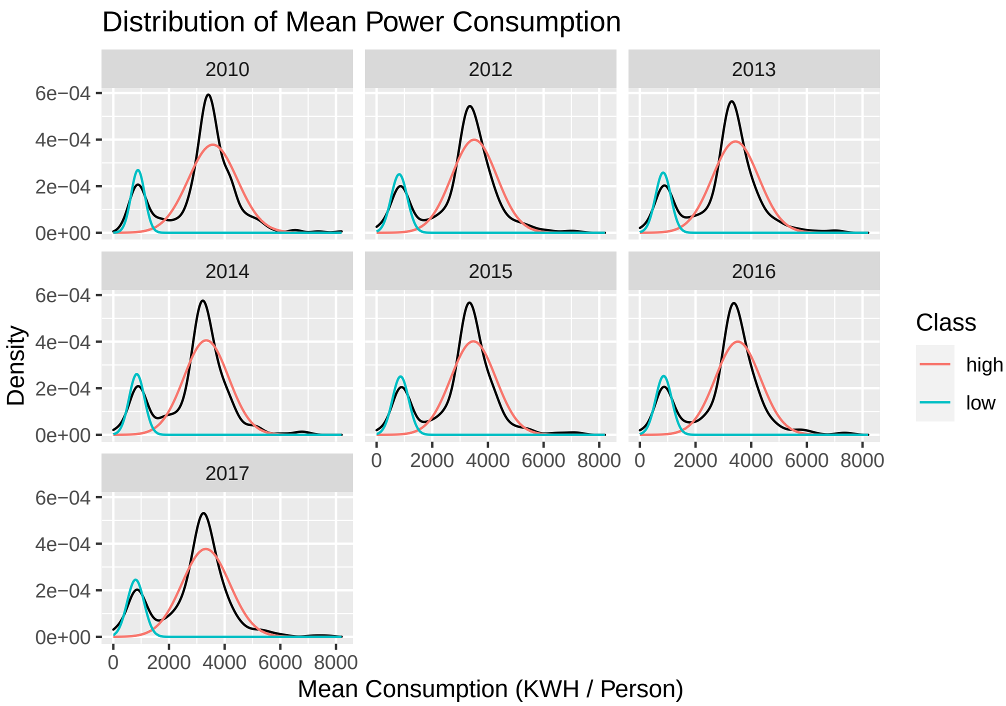


Figure 3:

Figure ? is a table of the numeric estimates for the parameters for each gaussian mixture model. The difference between the estimated means for the first and second component is in the ball park of 2500 KWH. This is a rather stunning conclusion as it implies some new yorkers use twice as much electricity!

YEAR	comp_1_mu	comp_1_sigma	comp_2_mu	comp_2_sigma
2010	884.6576	245.6020	3573.999	880.1385
2012	812.3030	291.2419	3507.857	815.7682
2013	836.9359	277.8140	3435.770	834.9090
2014	842.0521	276.2501	3338.235	806.2767
2015	861.0139	293.6998	3475.179	811.0632
2016	855.2439	290.9703	3520.584	813.5983
2017	799.7874	304.9476	3322.841	859.1821

Since we have calculated mean power consumption by grouping on TDS, it stands to reason that there are specific building characteristics which support more efficient use of electricity. A reasonable starting hypothesis is to suppose that new building are more energy efficient. Material science has done much in the last half century, and modern insulation and building techniques used today are noticibly superior. Figure ? is a density estimate of the distribution of building completion dates. There is a noticible bump in the mid 1980's where there was a surge of new constructions.

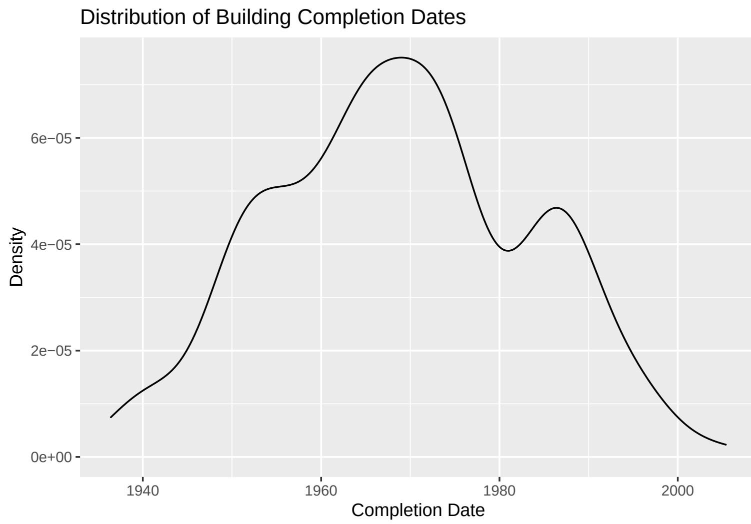


Figure 4:

Figure 5 indicates that our guess isn't quite right, but we aren't completely wrong either. The signicant overlap between high and lower power users even among newer construction suggest building age itself is not enough the explain the difference in energy consumption. However there is a noticble pattern that newer building tend to be more energy efficient. Overall it would seem that there is some characteristic newer buildings tend to have, but which older buildings can also have which save on electrical cost.

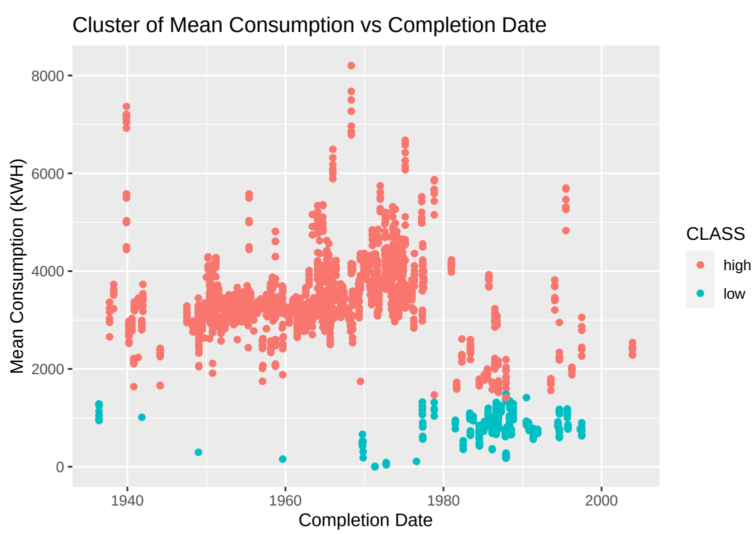


Figure 5:

Next we turn our attention to the weather data. An interesting question is whether not it supports the notion of seasons arriving "late" or "early". In other words, we want to see whether or not ground hogs should be afraid of their shadow. We can investigate this claim by exploring how temperature varies year to year. Using `pivot_longer` we can gather all the temperature readings into one column. This allows us to use ggplot2's facet wrap feature to quickly juxtapose multiple different temperature readings. In figure 2, a derived metric `TDIF` is included which measured the temperature gap between `TMAX` and `TMIN`. Unfortunately for ground hogs everywhere, figure 2 does not support the notion that seasons (at least when observed from a temperature stand point) do not arrive early or late. In fact, it would seem that on average, monthly temperatures are very regular since they only differ by a couple of degrees year to year. However the story is not quite complete because the the `TDIF` graph provides some reason why people believe seasons somtimes don't arrive on time. It may not be a coincidence after all that ground hogs day is a statement about whether or not spring comes early or not. `TDIF` reveals that, by far, the the biggest temperature swings come around late spring (i.e. day 100 - 150) with differences sometimes over 18 degrees. With such a high variance in temperature, it would come as no surprise that some people find it hard to tell when winter ends and spring begins.

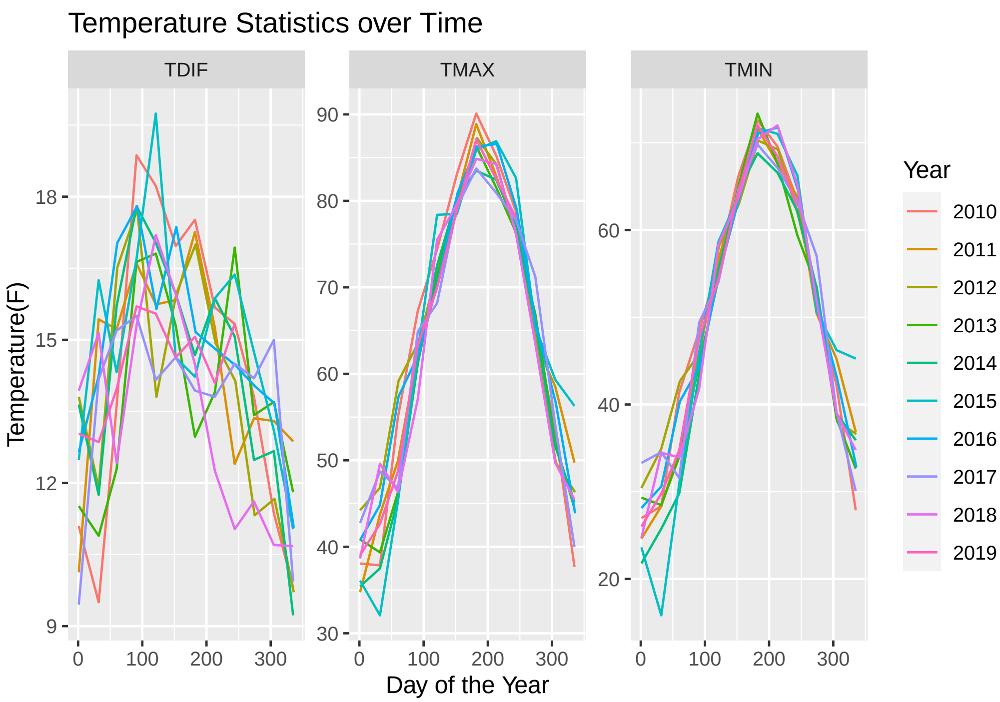
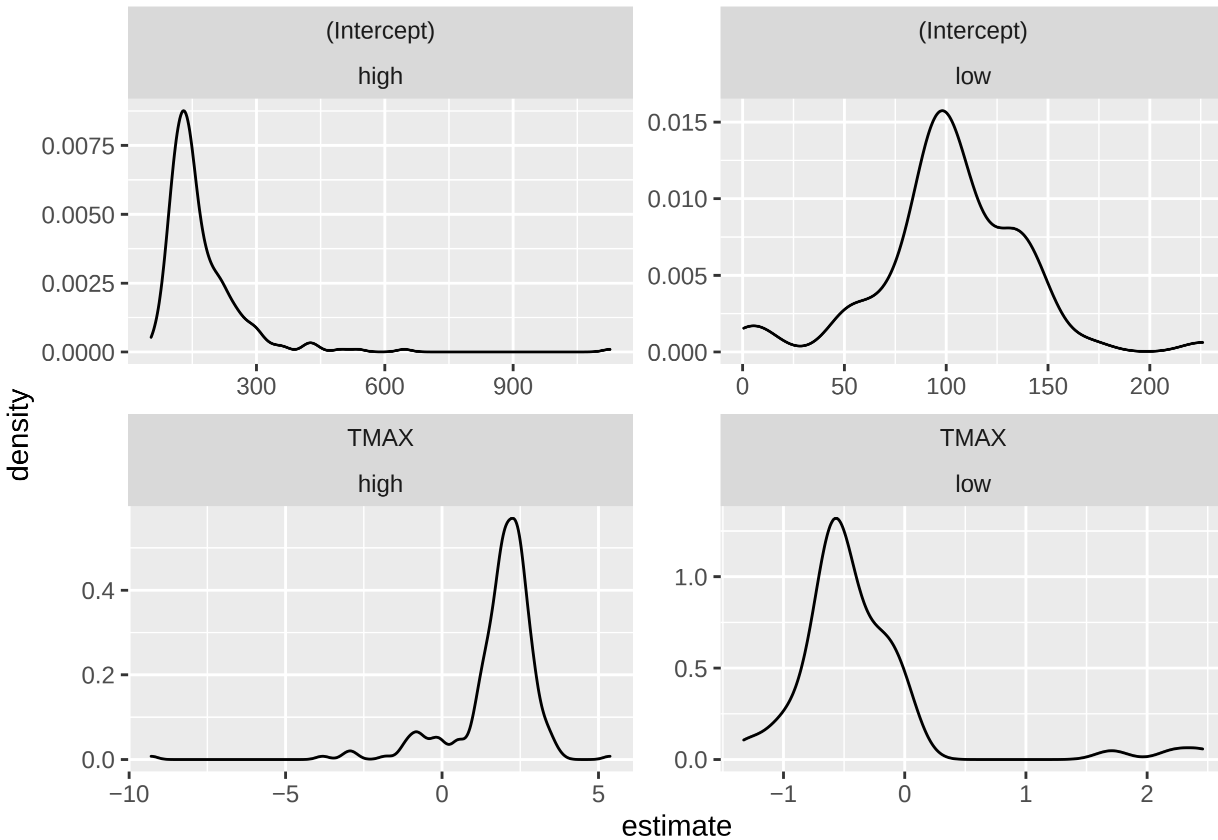


Figure 6:

To more precisely quantify the effects of temperature we can apply a mixed model type analysis. The linear model itself will be simple:

$$\text{KWH Consumption} = \beta_0 + \beta_1 \cdot \text{TMAX}.$$

Here β_0 and β_1 are random and associated with an indiviual TDS.


(Intercept)
high
(Intercept)
low
0.0075
0.0050
0.0025
0.0000
300
600
900
0.015
0.010
0.005
0.000
0
50
100
150
200
density
TMAX
high
TMAX
low
0.4
0.2
0.0
−10
−5
0
5
1.0
0.5
0.0
−1
0
1
2
estimate