Project 2 Final Report

W200 Thursday 4:00pm Session
Alissa Stover, Sophia Skowronski, Ying Hua

1. Context

This project is inspired by an ongoing Kaggle competition. Significant investments have been made to improve energy efficiency, reducing costs as well as emissions. Under pay-for-performance financing, the building owners pay for the difference between actual consumption vs. consumption without any improvement. However, there is no way of really knowing the latter data, and the goal here is to build a reliable model to predict building energy consumption based on building-specific information as well as external factors like weather.

2. Target Questions

The main questions we want to explore are:

- What are some factors that impact energy use?
- Can you predict energy use for a particular building?

While we may not be able to answer these questions fully to be competitive in the Kaggle competition -- the evaluation is based on Root Mean Squared Logarithmic Error (RMSLE) -- we hope to push the limits of our knowledge and do as much as we can with the dataset.

3. Source Data

We are using datasets provided by the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) in this Kaggle competition. The datasets include three years' worth of hourly meter readings from over one thousand buildings at 16 different sites around the world. The dataset also includes weather-related data for those sites. There are several components to the datasets, outlined below.

building_metadata.csv

- This dataset contains information about the buildings from which the meter readings were collected.
- Shape: 1449 rows x 6 columns
- Variables: building_id, site_id, primary_use, square_feet, year_built, floor_count
- Level: one row for each of the 1449 buildings, located across 15 different sites

train.csv

- This dataset includes the meter readings collected from each building at every hour for every day in 2016 (which was a leap year)
- train.csv shape: 20,216,100 rows x 4 columns
    - Given that we have 1449 buildings with 24 readings per day for 366 days, we would expect to have 12,728,016 rows
    - We have ~1.5x as many -- meaning that there are multiple readings from some or all of the buildings, which is probably explained by the fact that there are multiple meter types included in the data
    - When we limited the dataset to only the electricity meter readings, we are left with 12,060,910 rows -- which is much closer to what we expected however we are now missing 667,106 rows of data.
- Variables:
    - buildling_id
    - meter - {0: electricity, 1: chilled water, 2: steam, 3: hotwater}
    - timestamp
    - meter_reading - (in kWh) is the response variable

weather_train.csv

- These datasets contain hourly weather data for 2016 from a meteorological station as close as possible to the site. There is weather data from a meteorological station closest to each of the 16 sites, for 16 stations total.
- weather_train.csv shape: 139,773 rows x 9 columns
    - Given that we have 16 sites with 24 readings per day for 366 days, we would expect to have 140,544 rows
    - It appears that we are missing 771 rows from this dataset with no explanation as to why -- we address how we resolved this missingness below
- Variables: site_id, timestamp, air_temperature, cloud_coverage, dew_temperature, precip_depth_1_hr, sea_level_pressure, wind_direction, wind_speed
- Level: hourly

Additionally, we have two datasets that were intended to serve as test data for the competition, test.csv and weather_test.csv. We did not initially plan to use either, however as we detail below in our discussion of the weather data, in order to merge the meter readings and weather data we performed a timezone adjustment with the weather data. This meant that we eventually used some of the data from weather_test.csv to achieve a more complete merge:

weather_test.csv

- These datasets contain hourly weather data for 2017-18. There is weather data from a meteorological station closest to each of the 16 sites, for 16 stations total.
- weather_train.csv shape: 277,243 rows x 9 columns

    - Given that we have 16 sites with 24 readings per day for 365 days for 2 years, we would expect to have 280,320 rows
    - It appears that we are missing 3,077 rows from this dataset with no explanation as to why -- we address how we resolved this missingness below
- Variables: site_id, timestamp, air_temperature, cloud_coverage, dew_temperature, precip_depth_1_hr, sea_level_pressure, wind_direction, wind_speed
- Level: hourly

For this competition, it is important to look at the features within the training dataset first to develop a model to be used for submission on the test data. Once feature engineering and cleaning has taken place, the training dataset contained the meter readings and weather data for 1449 buildings every hour in 2016:

- Shape: 12,060,311 rows x 29 columns
- Variables: building_id, meter, timestamp, meter_reading, site_id, primary_use, square_feet, year_built, floor_count, time_index, day_of_week, hour_of_day, index, avg, std, outlier, air_temperature, cloud_coverage, dew_temperature, precip_depth_1_hr, sea_level_pressure, timestamp_utc, wind_direction, wind_speed, timezone, country_code, location, dst, local_time
- Level: hourly

4. Initial Exploration and Data Preparation

The weather data corresponds well with the building information. Despite the fact that the data came from the Kaggle competition, we still found a couple of data cleaning issues in addition to some data anomalies. In the section below, we highlight some of our observations about the missing data, data anomalies, and how we addressed them to build a more usable training dataset.

4.1 Missing data

We first looked at the completion of the data provided to us. Right away, we noticed there was a significant amount of missingness which we needed to address.

Narrowing the scope to electricity meter measurements

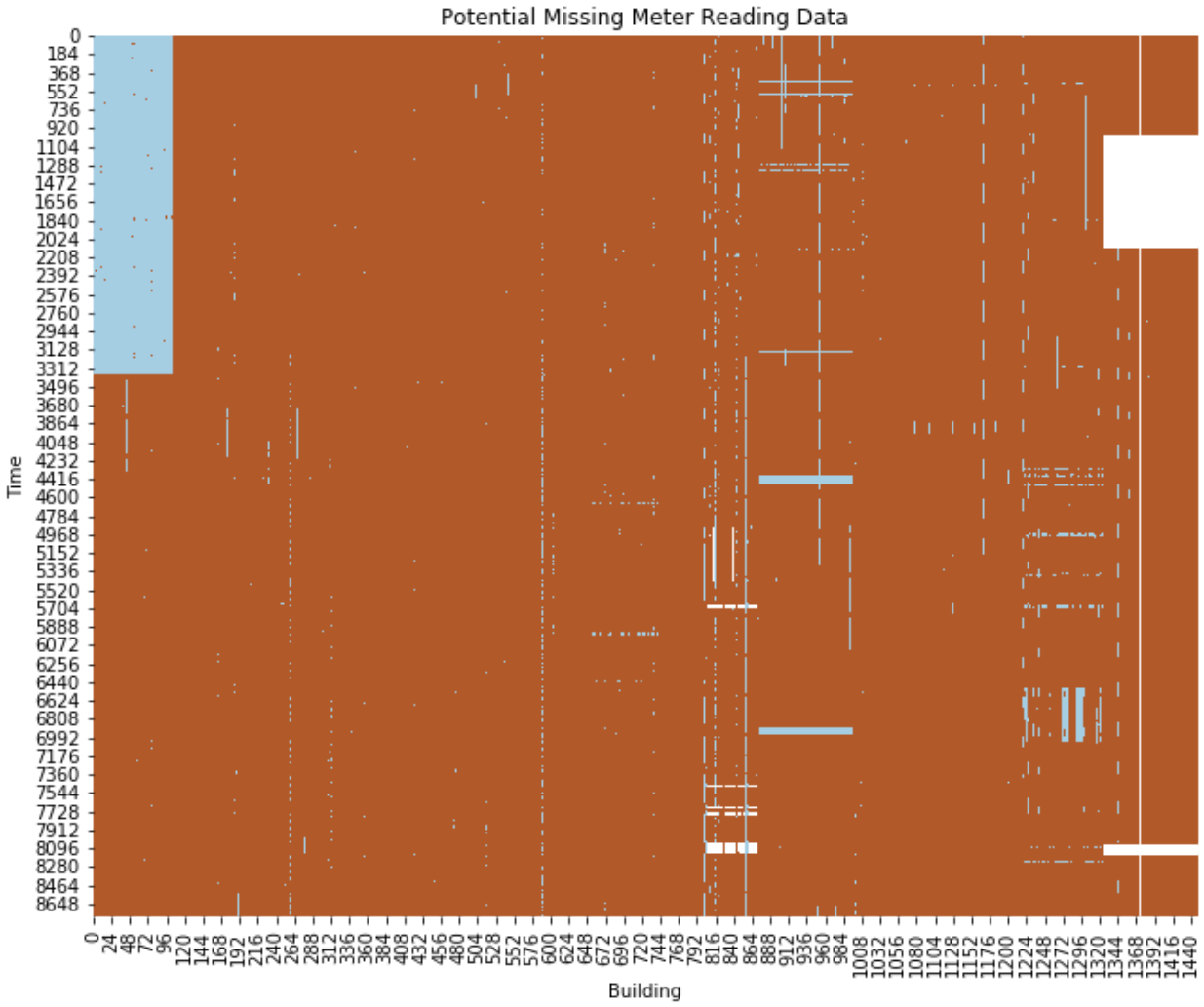
Given the large size of these datasets, we decided to focus on only the electric meter, which has the most complete set of data for the meter_reading variable. It is the response variable we wish to predict and thus one of the most important variables we have available.

Exhibit 1 represents a graph of electric meter reading data. The x-axis represents the individual buildings, and the y-axis represents the hourly meter_reading for an entire year. 2016 was a leap year so there are 24x366 individual data points along the y-axis. Two items immediately jump out: (1) There is missing meter_reading data (i.e. the date stamp is missing, this is indicated by the white space), (2) there are a number of 0 readings (4% of the total meter

reading data), indicated in blue. While some of these 0 measurements may actually represent no energy use (e.g. at 2am in the morning), we suspect that some of these 0 readings were reported erroneously (e.g. blue stripes in the middle of the graph).

Given the importance of having a clean response variable, we prioritized processing of the meter readings. We were more concerned about erroneously reported 0 meter reading data than the missing time stamps, as they are more likely to skew our prediction model. As such, we looked to resolve these potentially erroneous data in the data cleaning section.

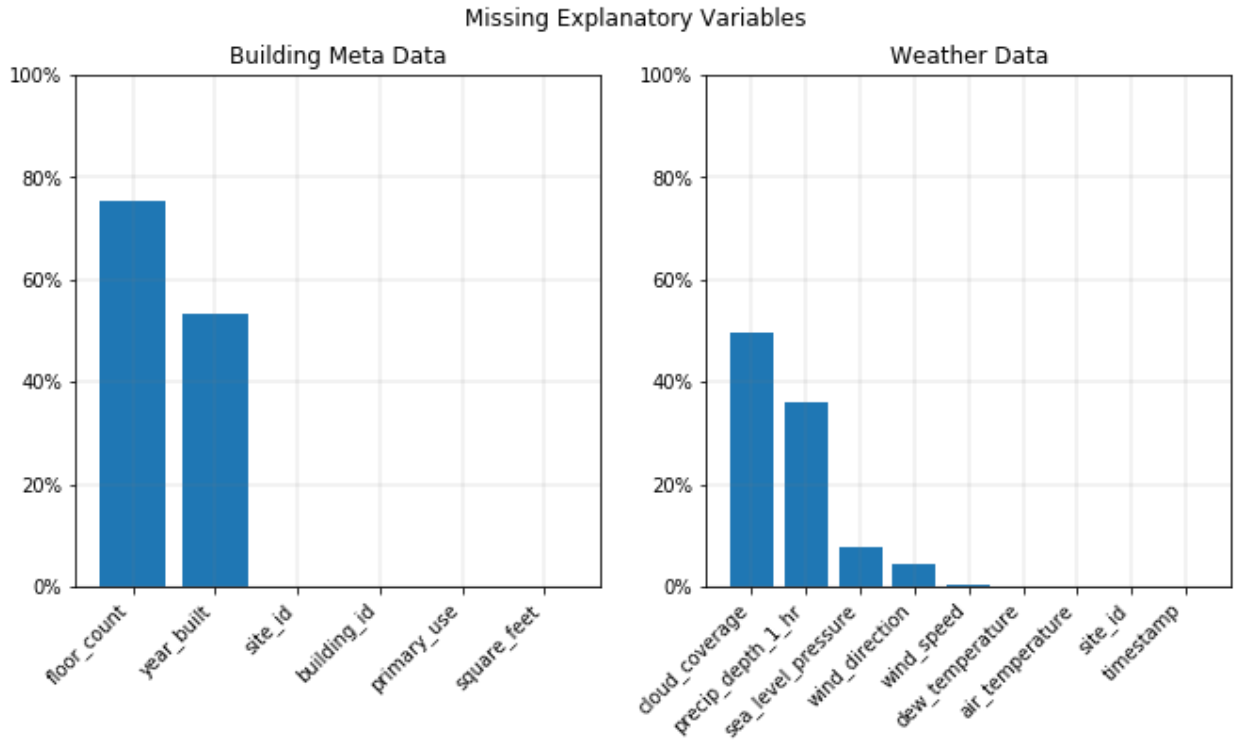
Exhibit 1: A heatmap of meter_reading variable that indicates potentially erroneous data.



Explanatory variables: building and weather characteristics

We then proceeded to evaluate the quality of the data of our dependent, explanatory variables in the building and weather datasets. In Exhibit 2, we display the percentage of missing data for each variable.

Exhibit 2: A bar chart displaying the percent missingness for each variable in the building and weather data.



For building data, the most concerning observation is that over two-thirds of the floor_count variable is missing. Intuitively, we believe square_feet can be a good proxy for floor_count. We noted that if our hypothesis was correct, we would be able to use square_feet to proximate floor_count in our model. If not, we would need to look for ways to impute such missing data.

For weather data, we see that temperature data, both dew_temperature and air_temperature, is the most complete while cloud_coverage is missing almost half of the data. Intuitively, we think temperature will have a bigger impact on energy consumption. As such, we expect that the temperature will be important to explore.

4.2 Imputing the missing data

Response variable: meter readings

We took two approaches to treat issues in the meter reading data. For obviously missing data (i.e. missing timestamps), we decided to ignore it. Because we planned on merging the weather and building data onto meter reading, the missing timestamps would not too strongly impact our predictive model.

For data that was erroneously measured as 0, we believed that we would benefit from using machine learning techniques to impute such data. To do so, we first needed to distinguish which among the 0 readings are missing data and which were truly 0 readings. We defined meter readings as 0 if they were not followed by spikes and did not last longer than 7 consecutive days. From our visual chart (Exhibit 1), we can also see that in site 0 (the first couple of buildings), there were a large chunks of data missing at the beginning of the period. There were some small readings here and there during this period at certain buildings but for the most part, the readings were non-existent. We decided to exclude these data.

After cleaning data in preparation for modelling, we split our clean data into train and test datasets with a 70/30 split. We then tried several different machine learning methods to impute such missing data, eventually settling on using sklearn's linear regression model. To impute erroneous meter reading data, we used building _id, site_id, day_of_week and hour_of_day and achieved a 87% R-square in our test dataset.

Building characteristics

We used a similar methodology to impute missing year_built and floor_count data as well. We note that both of them have a much higher percent of missing data than meter reading: 75% and 53% for year_built and floor_count. While we are able to impute the missing data, we are more inclined to use other variables when it comes to model building. Using a KNN approach, we achieved an accuracy score of 57% and 87% for year built and floor count, respectively, in the test dataset. We think these are good results and decided to use these imputed data in our analysis.

Weather characteristics

We decided to only impute missing air_temperature and dew_temperature data as we expected these to be most relevant for energy use, both from our understanding of weather data and from the fact that these were the most complete variables. To impute missing weather data, we utilized a very different approach from the methodology we used for missing building and meter reading. The reason we used a different method is that the temperature data showed mostly sporadic patterns (i.e. at random hours of the day). This meant that for each building, the values of missing temperature measurements should be bound by the temperatures reported immediately before and after the missing data value.

Given this logic, we first imputed missing temperature data by taking the average of the data before and after such missing data for each building. For data that was still missing after this step (i.e. no data points ahead or after to calculate the average), we used the backfill methodology.

Exhibit 3: A table tracking which techniques we used to treat missing variables in the datasets.

Treatment of Missing Variables				
	Variable Name	% Missing	Treatment	Note
Outcome Variable	meter_reading	4%	Imputation using linear regression	• No NA but we have reason to believe some data reported as 0 are erroneous • Also tried other methods of imputation including KNN, Naive Bayes. Linear regression were the most efficient to run and gave good results
Explanatory Variables	**Building Variables**			
	year_built	75%	Imputation using KNN	• Also tried other methods of imputation including linear regression, Naive Bayes. KNN gave the best results
	floor_count	53%		
	Weather Variables			
	air_temperature	0.03%	Imputation using average of the values before and after; if NA, using backfill	• Because temperature data is bounded by time and specific location, we think this method is most appropriate
	dew_temperature	0.08%		
	cloud_coverage	49%	Ignored	• Most of the weather data did not have significant correlation with the response variable. As such, we priorities other variables to impute. • If we had more time we would impute these using a similar method for temperature
	precip_depth_1hr	36%		
	sea_level_pressure	7%		
	wind_direction	4%		
	wind_speed	0.2%		

4.3 Data anomalies

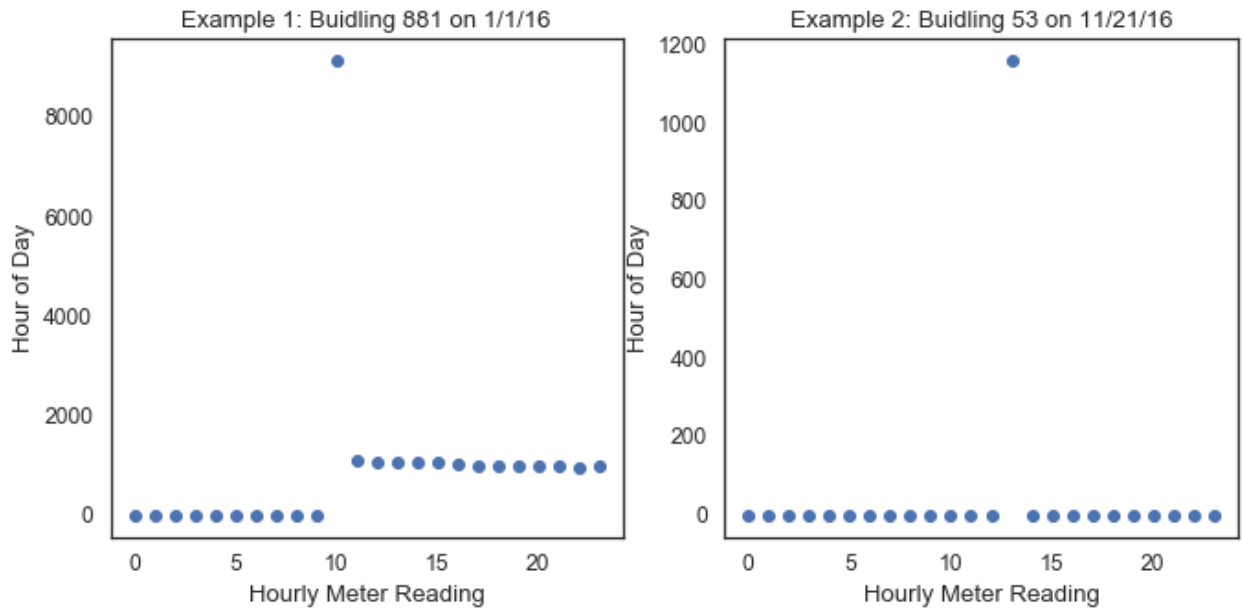
In addition to missing data, we did some sanity checks to make sure the data makes sense. A deeper examination revealed a number of remaining issues:

Response variable: meter readings

Negative measurements**:** 0.2% of meter_reading shows up as negative, some of which are likely due to our imputation method. While there may be a logical explanation for such values (e.g. rebates), we noted that such values would not reflect what we were trying to predict, which was energy use. Given that it is impossible to have negative energy usage in a building and that there were only a few negative readings, we decided to exclude these values.

Meter spikes: there are random spikes of meter_reading data. Often these were followed by 0 meter reading data in the hour before. These data points are likely erroneous and were excluded from the final dataset we used for modelling.

Exhibit 4: The spiking behavior of the meter_reading variable in two buildings.



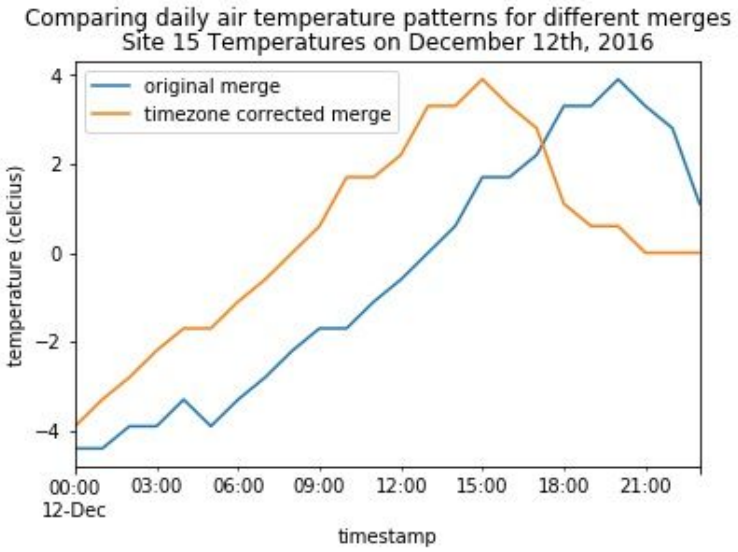
Explanatory variable: weather timezones

Based on the Kaggle discussion boards and our exploration of the data, it appeared that the main variable we intended to use to merge the weather and meter data -- timestamp -- did not reflect the same thing in these two datasets. In the meter data, the timestamps are recorded in local time. However, the meteorological data tracks timestamps in UTC. Merging on these variables without aligning them would make it difficult to measure any relationship between weather and meter readings, since the time of day is a significant factor in electricity use. To achieve a more robust merge, we decided to convert the weather data timestamp variable from universal (UTC) to local time.

We verified that this adjustment improved the merge via visual inspection of the relationship between the timestamp and one of the most easily-interpretable variables of the weather data: air temperature. One would expect that air temperature would rise throughout the morning and reach its peak around midday, then decrease in temperature over night. If adjusting the weather timestamp variable improved our merge, one would expect air temperature to more closely follow this pattern than in the original merge.

Please refer to Exhibit 5 for an example figure from Site 15. The plot shows temperatures rising and falling on December 12th, 2016. One can see that the original merge (plotted in blue) shows a peak temperature very late in the day: close to 9pm. In contrast, the timezone-corrected merge (plotted in orange) shows a peak temperature at a much more logical time: closer to 3pm. This pattern was repeated for all of the sites on various days of the year. Therefore, we decided to use the timezone-corrected merge over the original merge.

Exhibit 5: Comparing air temperature peaks with and without timezone adjustment for Site 15.



4.4 Outliers

A quick look at the distribution of the meter readings show that -- on an hourly basis -- 4% of the data are potential outliers (we defined an outlier as a reading 3 standard deviations away from the average hourly reading). When aggregating data to the daily level, we noticed that the number of outliers were reduced. We decided to examine further where these high readings come from.

Cross referencing primary_use and site_id, we noticed that most of the outliers are from the education sector (accounting for 61% of all outliers). While the magnitude of some of these outliers seems astonishing, we have no reason to believe these were erroneous measurements and thus decided to leave these data points in.

Examining the number of outliers for each of the primary use category, we note that outliers represent a very small portion of the data. Even for Education, which accounts for a large portion of outliers, only 2.8% of the data are outliers.

Exhibit 6: Boxplots displaying potential outliers in the meter reading data when examined hourly or daily.

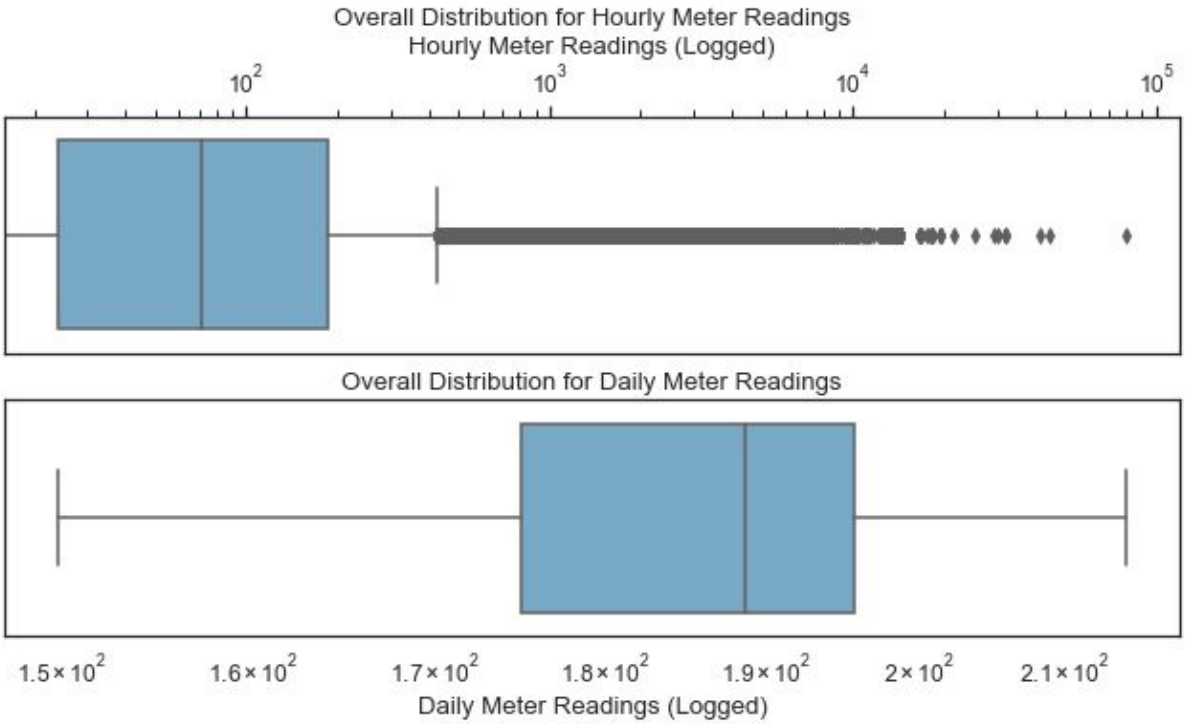
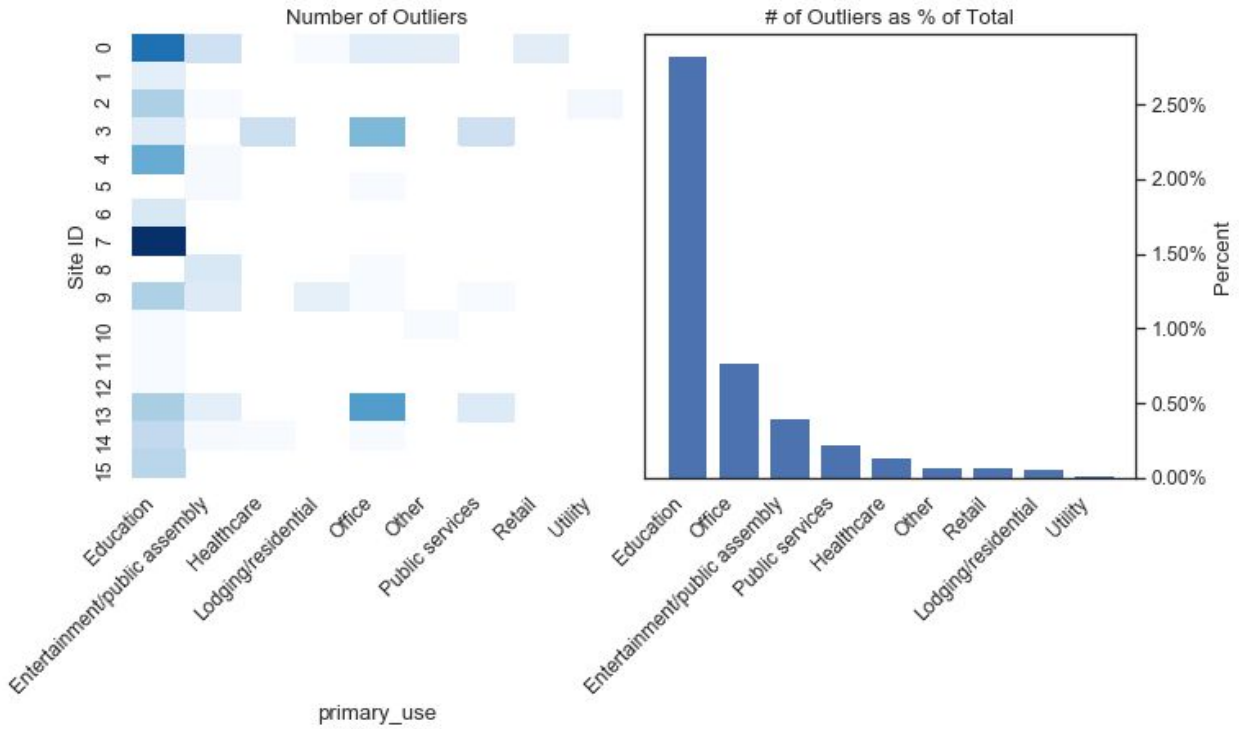


Exhibit 7: A heatmap and bar chart comparing number of potential outliers across sites and primary uses.



5. Exploratory Data Analysis

After we cleaned our data, we moved on to exploratory analysis. In this section, we focus on examining factors that may impact electric meter_reading. We also focus on some interesting observations that were revealed along the way.

After reading a couple of papers that identified which variables were most useful in predicting energy use (especially this one https://www.mdpi.com/2504-4990/1/3/56), we decided to split variables into three categories: meter, temporal and weather. Please see our summary below in Exhibit 8.

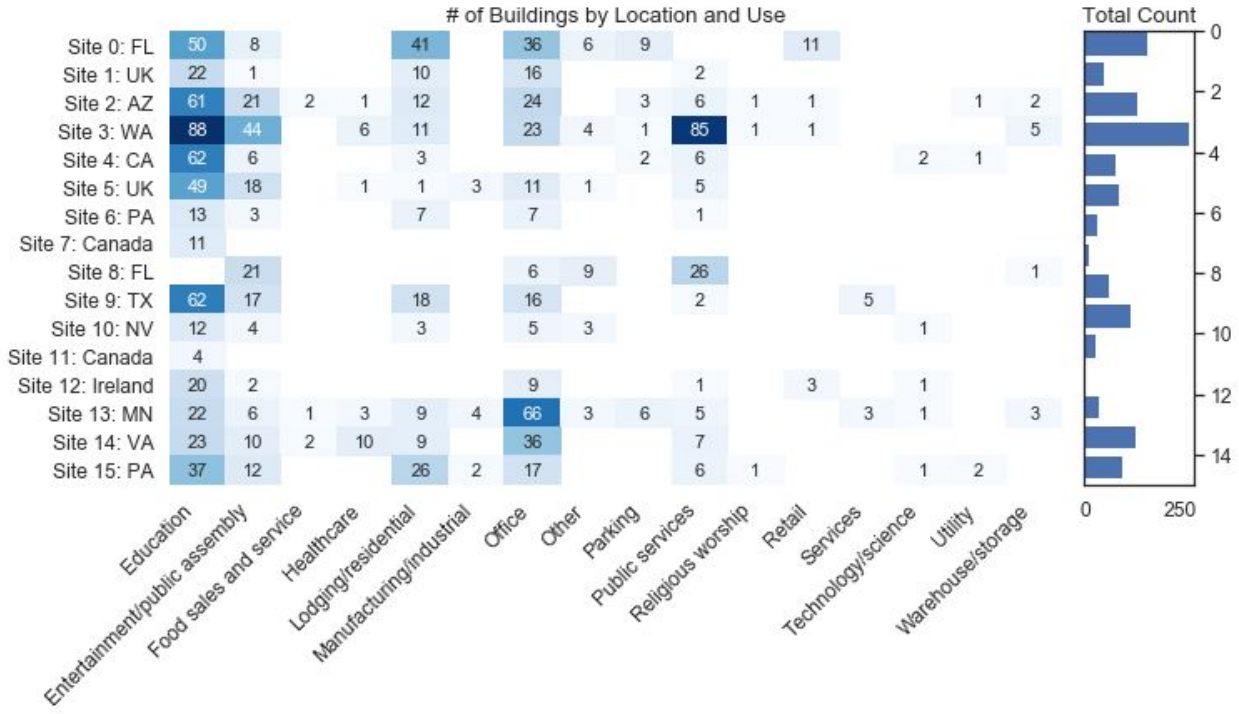
Exhibit 8: A table categorizing each variable and identifying which behaviors are targeted by each.

Category	Variable	Behavior Targeted
Meta	Site location	Location
Meta	Primary use	General category of use
Meta	Building age	Old vs. new buildings
Meta	Square feet	Size of building
Weather	Air temperature	Sensible heating and cooling
Temporal	Time of day	Daily schedule
Temporal	Day of week	Weekly schedule
Temporal	Public holiday schedule	Holidays
Temporal	Schedule type	Seasonal schedule

Meta: Site Location

There are 16 locations, mostly in the U.S. with a couple in Canada and the U.K.. Washington State had the greatest number of buildings, most of which were roughly equally distributed between Education and Public Service.

Exhibit 9: A heatmap and bar chart comparing the distribution of buildings across sites and primary uses.



We were interested in the spread of energy use by site. From Exhibit 9 below, we note that site 8 had most skew in the meter reading distribution. Further examination reveals that site 8 has 5 different industry types. We suspect that since the majority of the buildings are entertainment and have more energy use than other types of buildings, it is dragging up the mean distribution.

Exhibit 9: Boxplots showing differences in the distribution of daily meter readings by site.

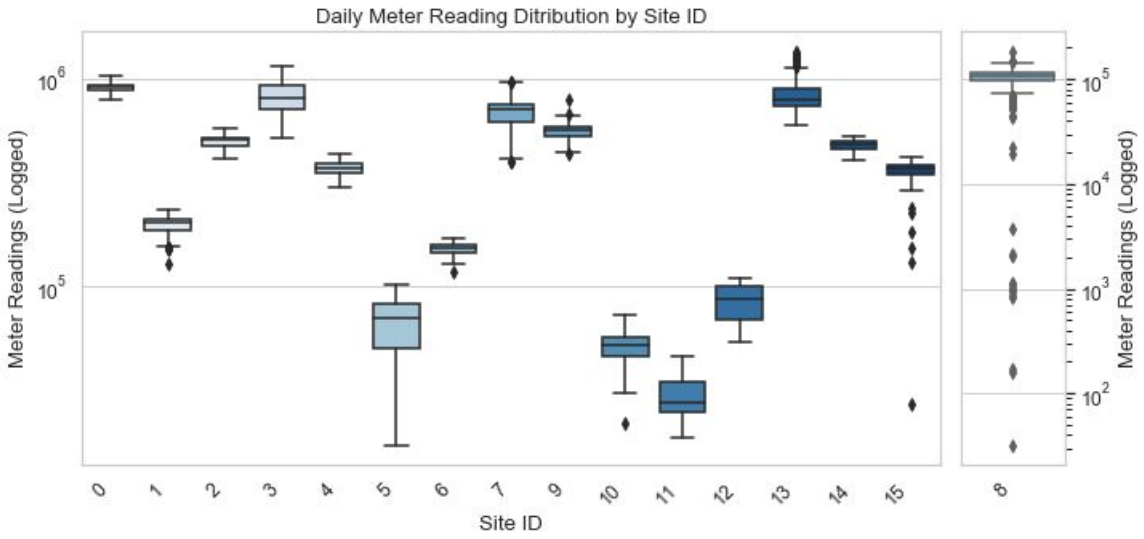
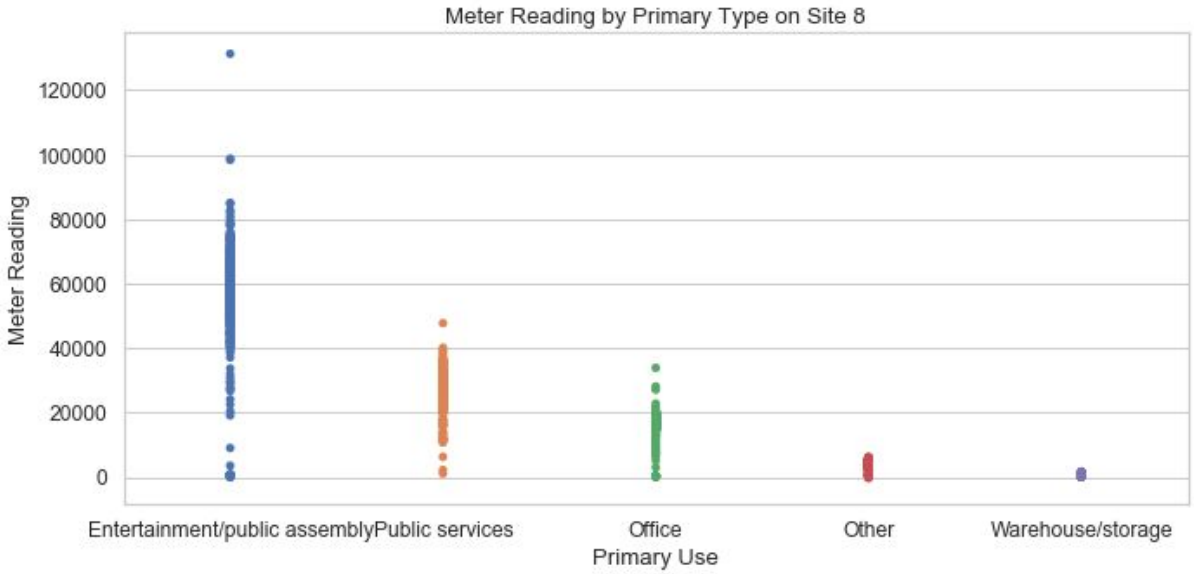
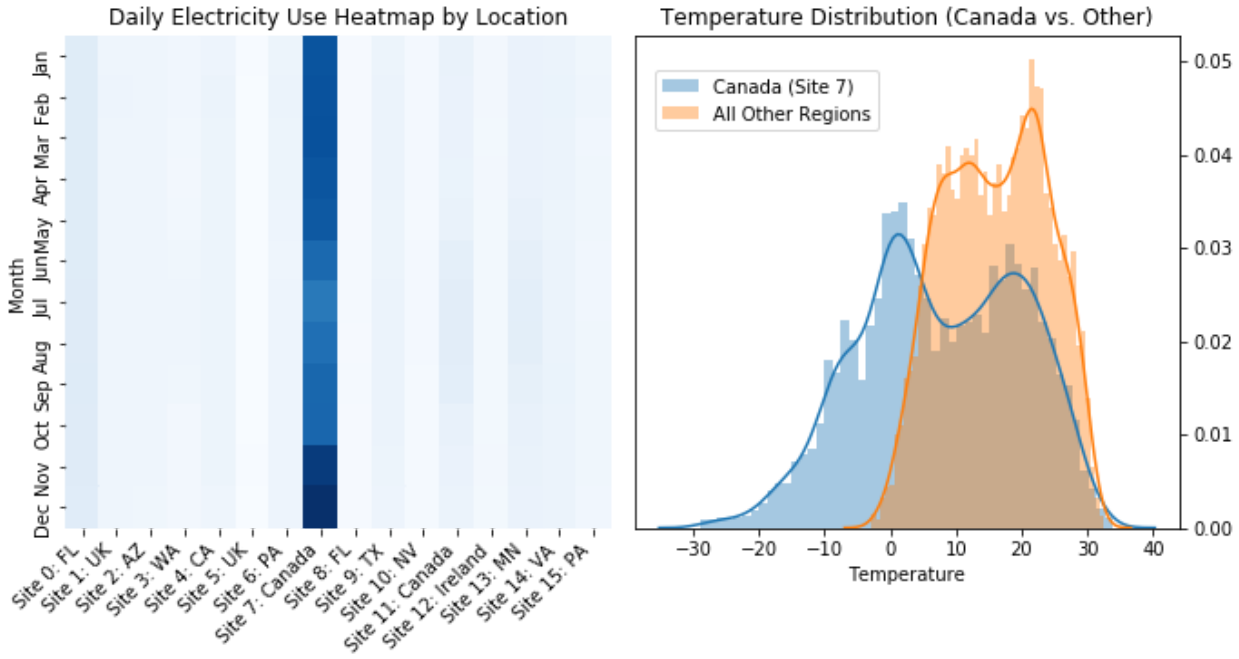


Exhibit 10: Bar chart displaying the distribution of buildings at Site 8 across primary uses.



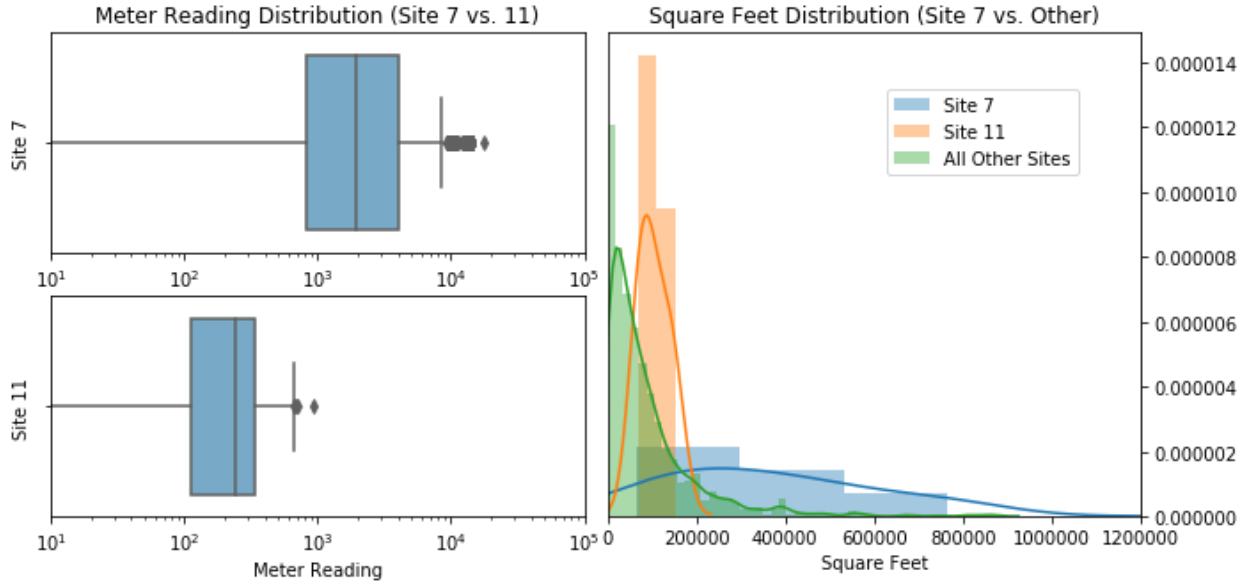
We also wanted to examine electricity usage type by location. When we look at average daily use by month, we noticed that Site 7 (Canada) had by far the most energy use, especially in winter months. We suspected that this was due to temperature. Not surprisingly, when we looked at hourly temperature distribution in Canada versus the average of the other sites, we found that most of the below-freezing temperatures were in Canada.

Exhibit 11: A heatmap comparing electricity use across locations and months. We can see that Canada's cold temperatures could in part explain their high electricity usage.



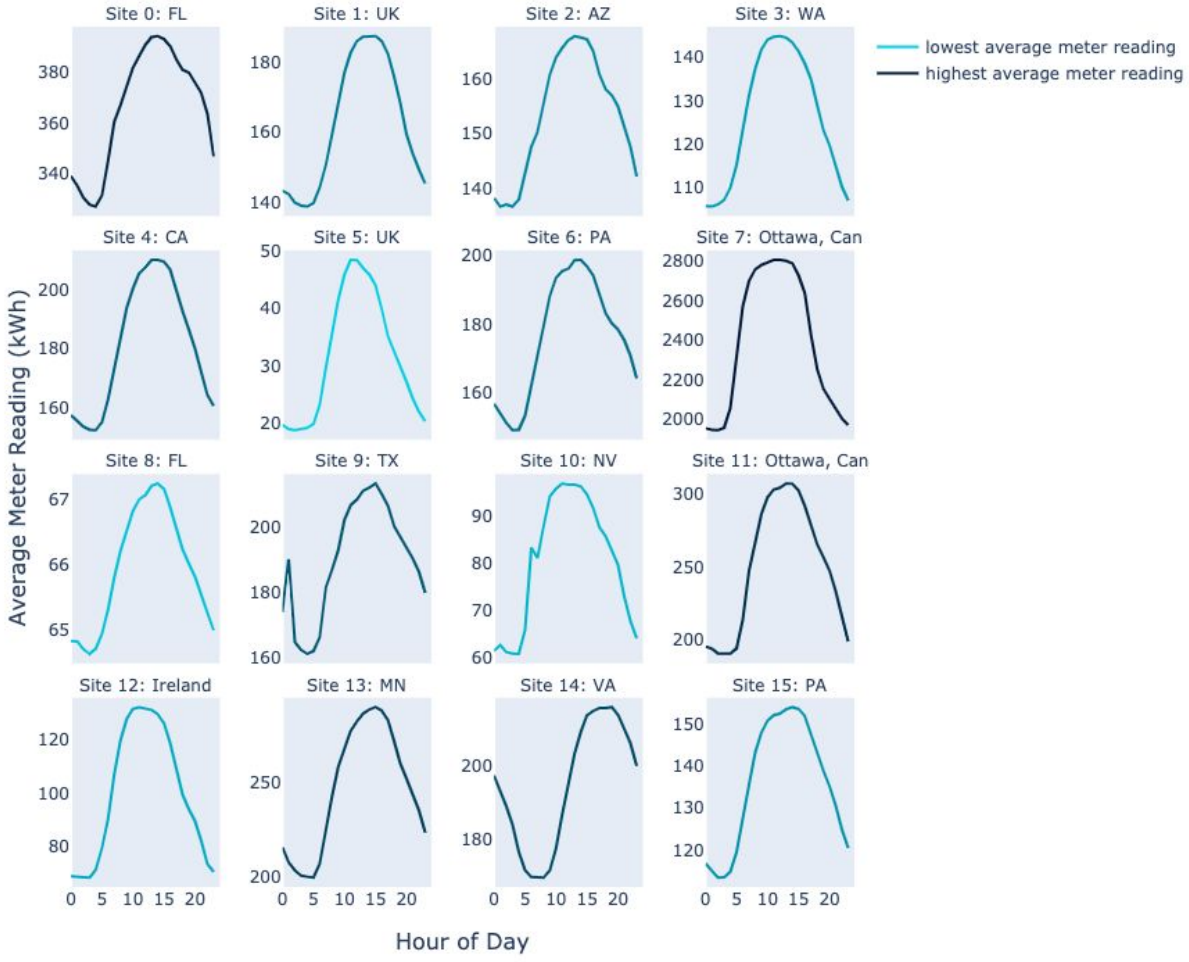
However, upon further investigation we noticed that another Canadian site in the same location does not have nearly as much energy use. We suspected that there were other reasons, beyond temperature, that were driving up the electricity consumption at Site 11. When we dug deeper, we noticed that Site 11 has well above average square footage when compared to Site 7 and the rest of the sites (in fact, 4 buildings at Site 7 had an average size of ~500,000 square feet vs. average building size of 91,000 square feet for the other sites). We believe that this contributed meaningfully to Site 7's electricity consumption.

Exhibit 12: Site 7 has much larger buildings than the other sites, which may explain part of why it has much higher energy usage even when compared to a similar site, such as Site 11.



We noted that different sites had very different patterns of electricity usage throughout the day. The figure below displays the average hourly meter reading at the 16 sites. Overall, many of the sites followed a similar pattern with a dip in the early morning hours (in Sites 9 and 14, this dip was very pronounced), sharp peak in the afternoon, and then a second dip in the evenings. The range of averages across the sites was quite large. As noted above, Site 7 was an extreme outlier at the high end of the spectrum. Site 5 was at the extreme low end. Most sites seemed to stay in the 200-300 kWh range, on average.

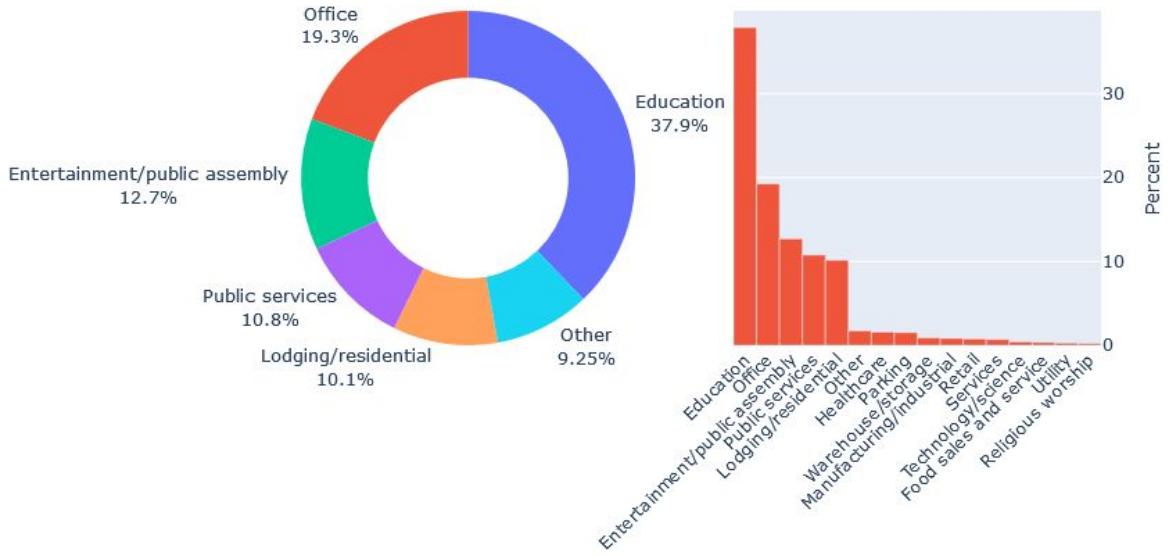
Exhibit 13: A comparison of the pattern and average of hourly meter readings across sites.



Meta: Primary use

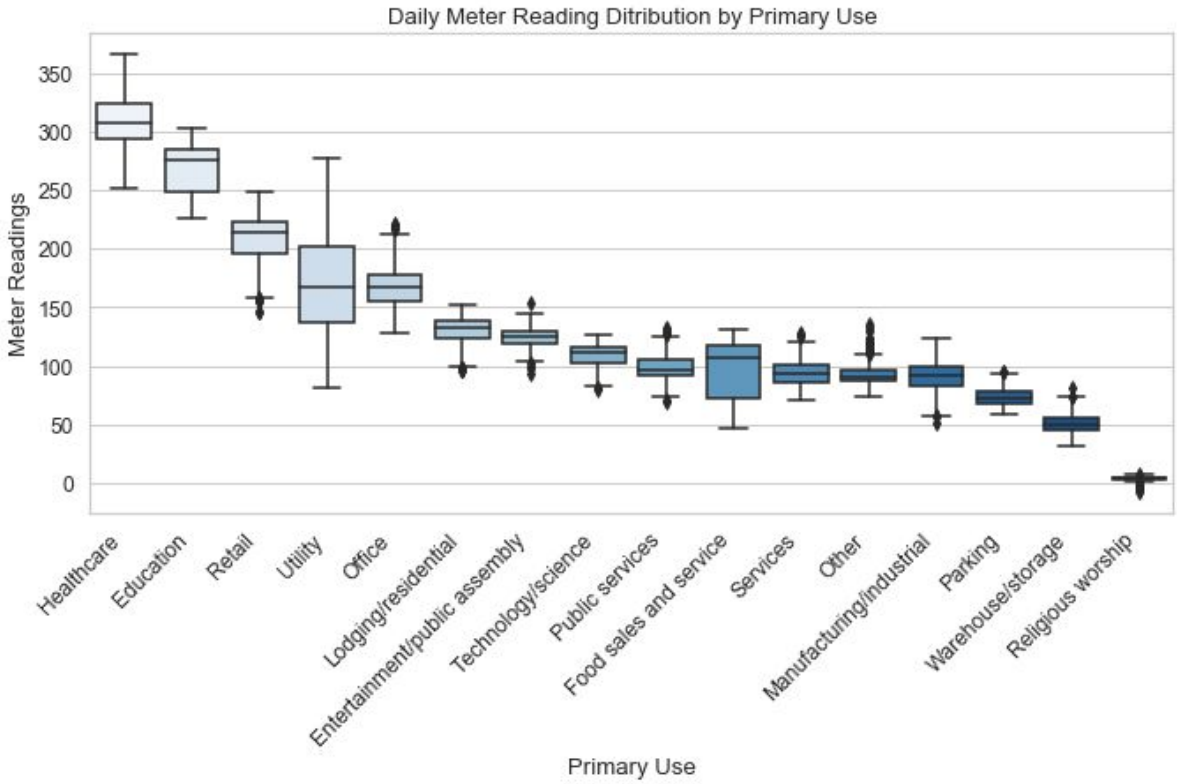
The next variable we wanted to examine was primary use. We noted that close to 40% of the buildings reported have Education listed as their primary use, followed by Office, Entertainment and Public service. In fact, the top 5 uses accounted for 90% of all buildings in the data.

Exhibit 14: The distribution of primary use.



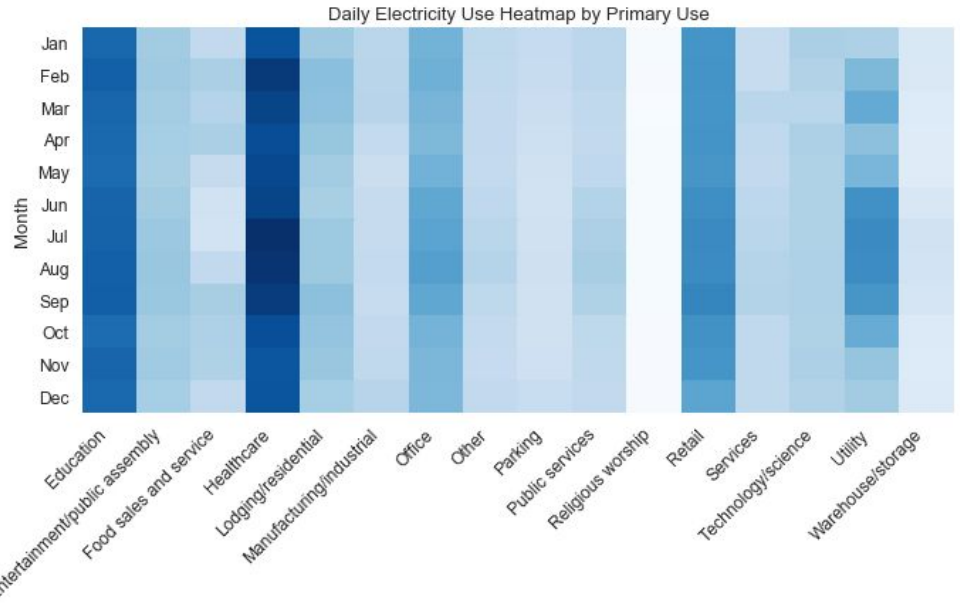
We next examined the spread of average daily use by primary use and noticed that certain industries have a much wider spread of usage than others. For example, healthcare usage falls in a much tighter band than utilities. We noted that religious worship shows up as very skewed with a number of outliers. We think this is mainly due to having a very limited sample size.

Exhibit 15: A comparison of the different distributions of meter readings by primary use.



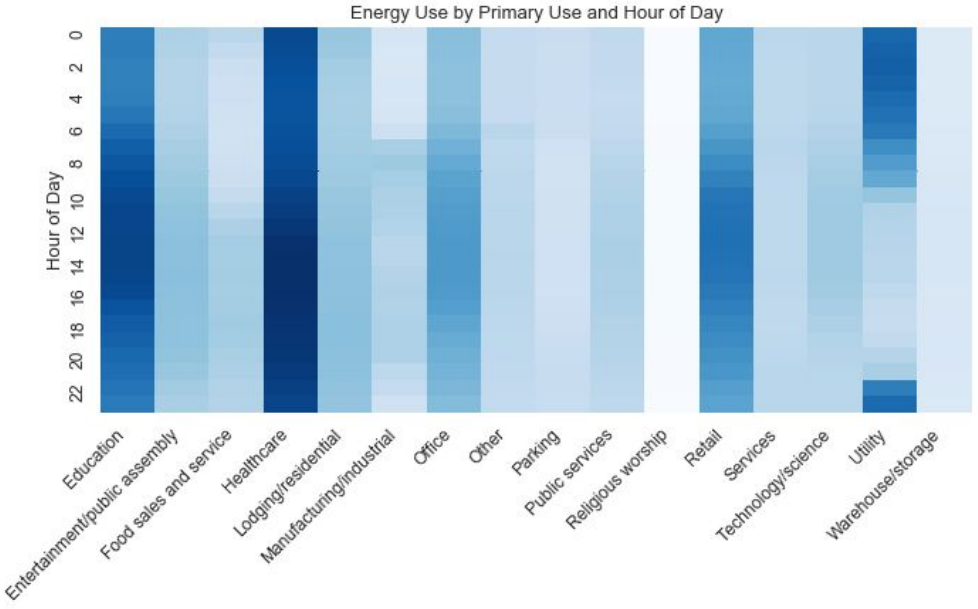
From Exhibit 15 above, we noticed that buildings primarily used for healthcare have the highest amount of energy usage and that those devoted to religious worship have the lowest. We wanted to dig deeper into this source of meter reading variation. Exhibit 16 below shows energy use by primary use in different months. We noted that not only does healthcare show the highest usage, the effect of seasonality is also most pronounced there.

Exhibit 16: A heatmap comparing patterns of electricity use across primary use and month.



We suspected that the different uses of the building would show very different patterns of meter usage throughout the day. For example, buildings used for food sales and service might show higher energy uses towards the late afternoon and evening when people are going to get lunch and/or are off from work. To address this question, we grouped the data by primary use of the building and hour of day, then took the average meter reading for each point. Exhibit 17 displays this grouped data in a heatmap. The diverging patterns suggest that our prediction was correct: meter readings vary widely by primary use of a building. Generally, meter readings seem to rise as the day progresses, then fall in the afternoon -- except for buildings used for utility purposes which show the opposite pattern.

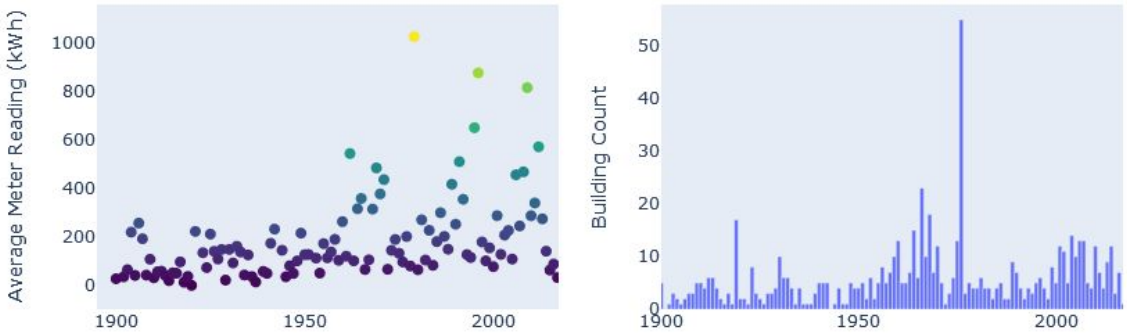
Exhibit 17: A heatmap comparing patterns of electricity use across primary use and month.



Meta: Building age

When we examine the distribution of building ages, we noticed a spike for buildings built in 1976, which accounted for 15% of total buildings. However we did not notice any obvious trend of older building using more energy, as we had originally expected. In fact, there is a bit of an upward trend that shows newer buildings use more energy. We think the trend is likely driven by other factors such as the primary use and/or size of the building.

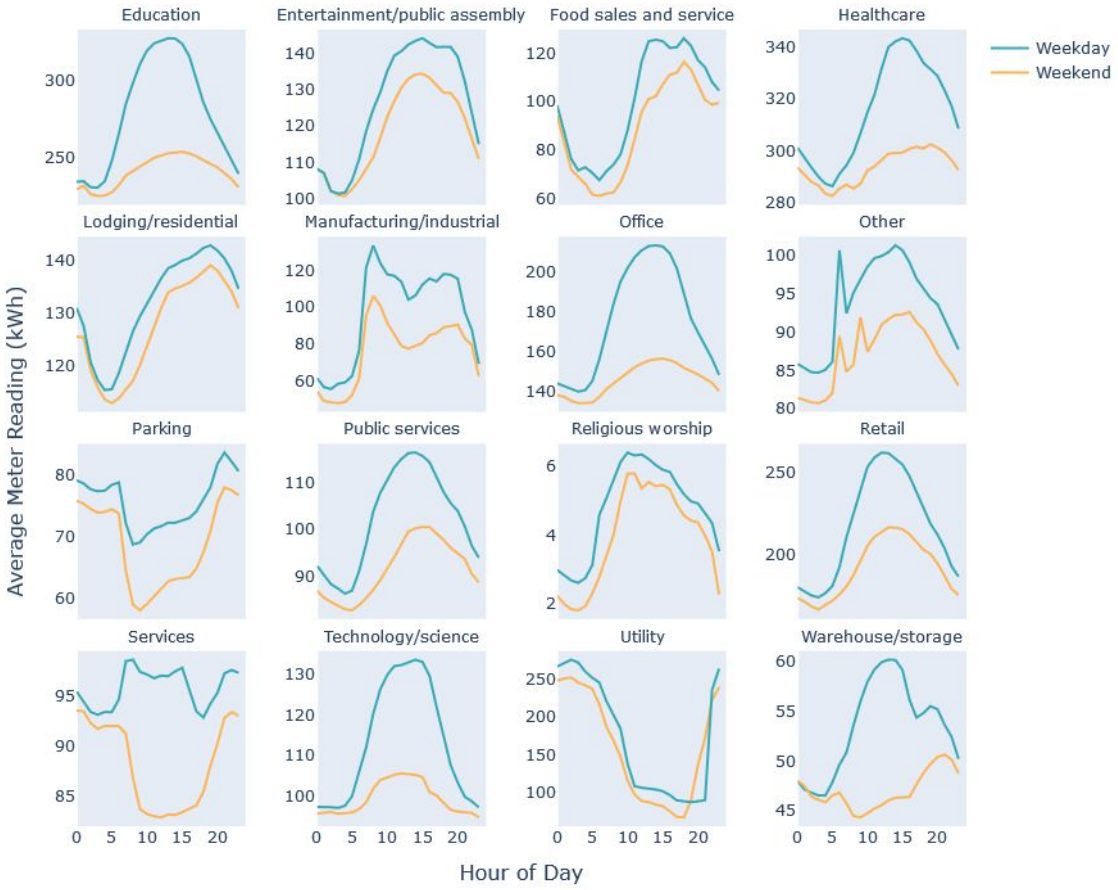
Exhibit 18: The distribution of building age and its relationship to average meter reading.



Temporal variables

We have already explored in earlier charts some of the energy use distribution by temporal variables (see Exhibits 4, 11, 13, 16, and 17). In this section , we will explore these variables in more detail.

Exhibit 19: Comparing average daily energy usage on weekends vs. weekdays (Online Version)



Do industries exhibit similar weekend behavior for energy use? Common sense suggests that for certain industries (e.g. office) we should see higher energy use on weekdays while in other industries (e.g. entertainment) we would expect to see higher energy use on the weekend. This assumes a typical weekday (Monday-Friday) and weekend (Saturday-Sunday) schedule.

Exhibit 20: Examination of the effect of holidays.

Here is another look at planned events in building energy schedules: holidays. Learning to detect non-routine events can strengthen a predictive model. Using the holidays package, we labeled certain days of the year by their regional holidays. Not surprisingly, there is an overall decrease in energy use on those days.

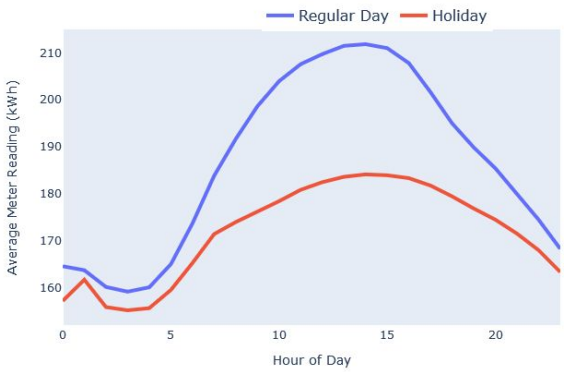
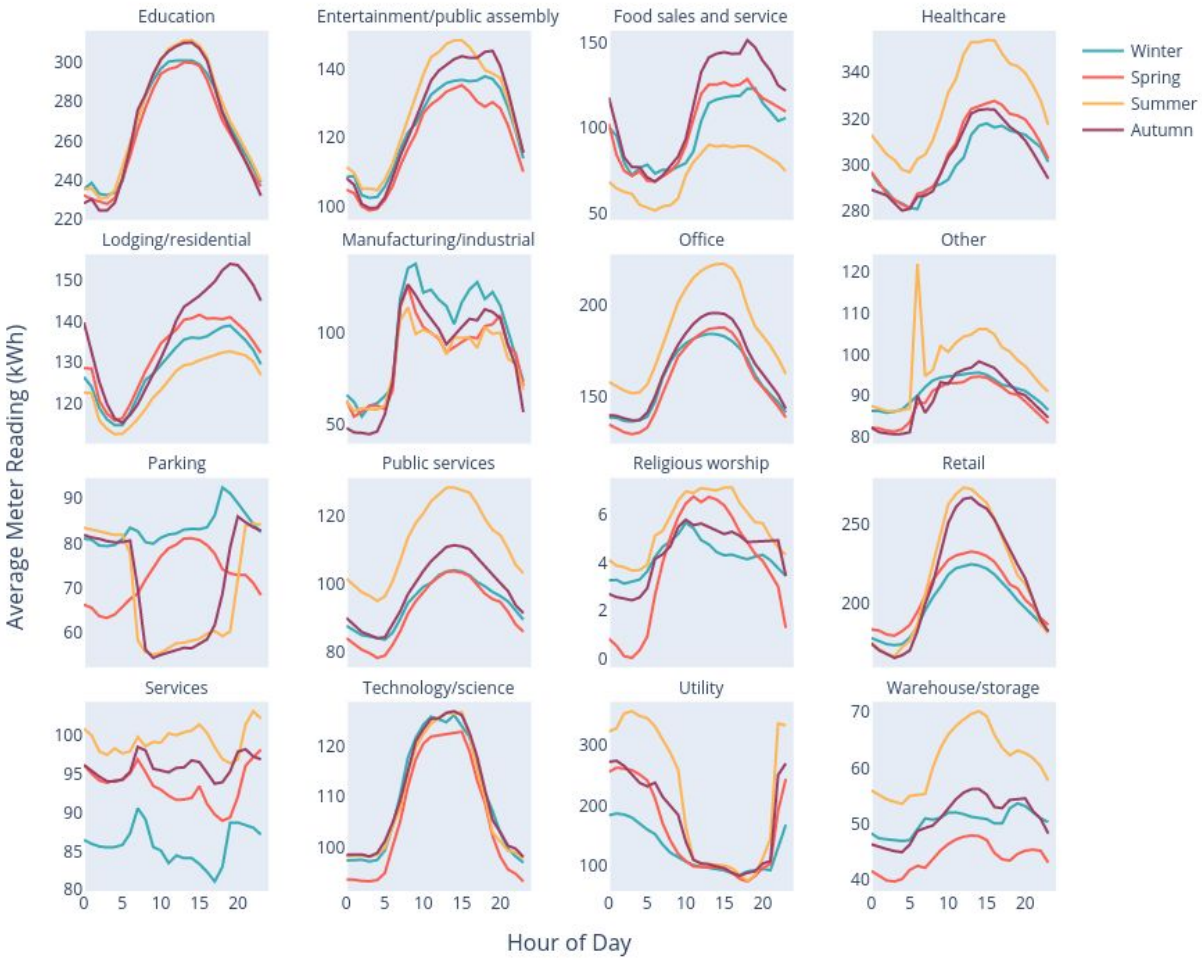
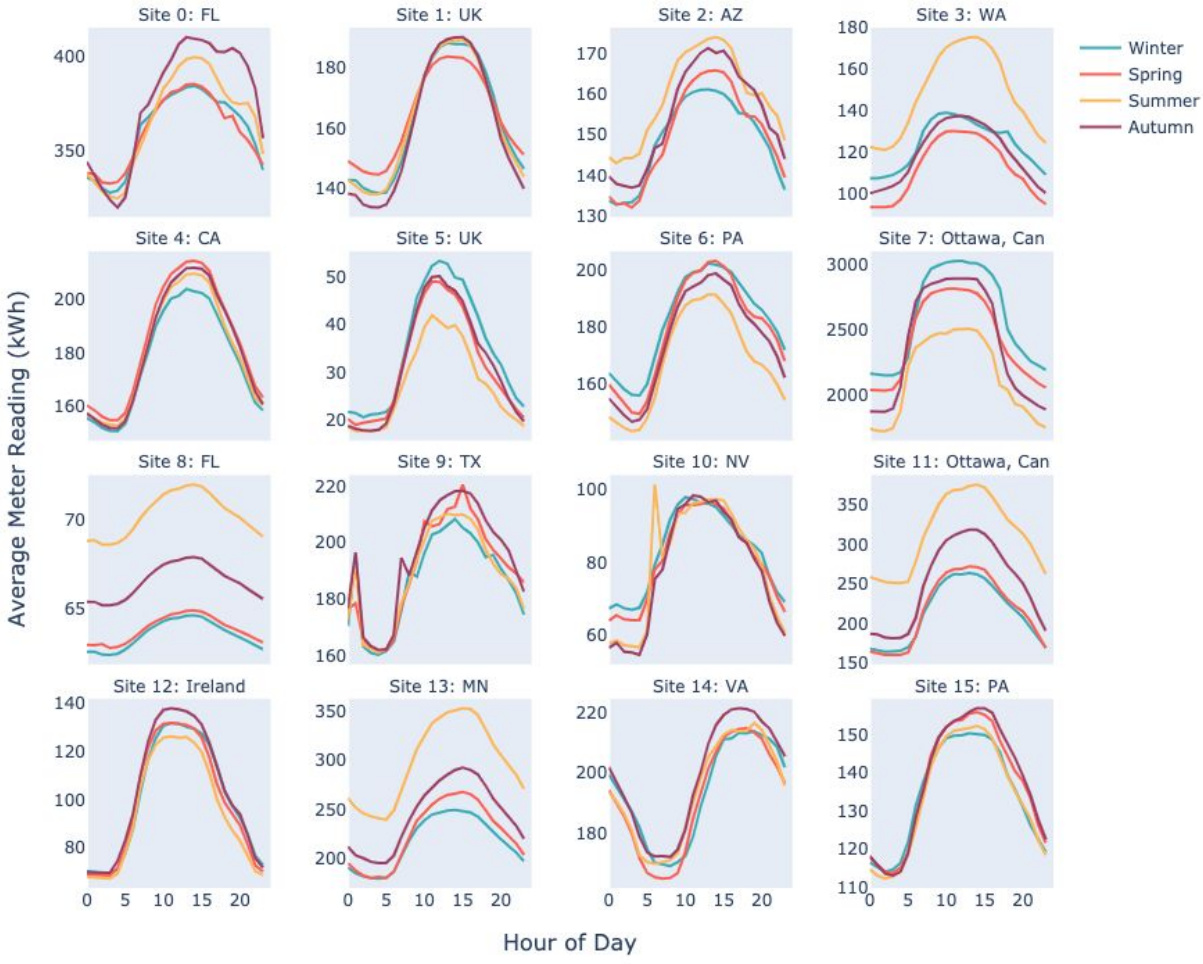


Exhibit 21: Comparing daily average meter readings across seasons and primary use (Online Version)



Exploring the seasonality of daily electricity usage is another interesting lens to understand building schedules. For standard HVAC systems, you can expect that buildings will use more energy as the outdoor conditions get hotter (in the case of cooling) or colder (in the case of heating). You can see from Exhibit 21 that some categories of buildings have a rigid schedule, no matter what the external conditions are. It is surprising to see that education buildings do not have a significant difference between seasons, even though one might initially guess that summer months -- when the typical academic year is not in session -- would show a decrease. For the majority of building types, the summer months result in higher energy usage throughout the day. It may be the case that some buildings switch to other HVAC systems (chilled water, steam, hot water), and that causes unexpected energy trends, but that is out of the scope of this analysis.

Exhibit 22: Comparing daily average meter readings across seasons and site.



When looking at seasonality by site (see Exhibit 22), one can see that in places like Ottawa, Canada with large changes in temperature with changes in season, we also see large differences in energy use. However, in more temperate climates like California, energy usage is not so distinct across seasons. Having extreme weather is associated with larger variations in meter readings; however, when temperatures are at a comfortable medium people are not using as much electricity to stay comfortable.

Temperature

Lastly, we want to explore the temperature variable and see how that impacts meter readings.

By looking at the average energy use by temperature, not surprising to see that the extreme cold temperature typically results in most energy use. This echoes what we noted above when evaluating seasonal changes in energy use across sites.

Exhibit 23: Examining the relationship between air temperature and meter reading.

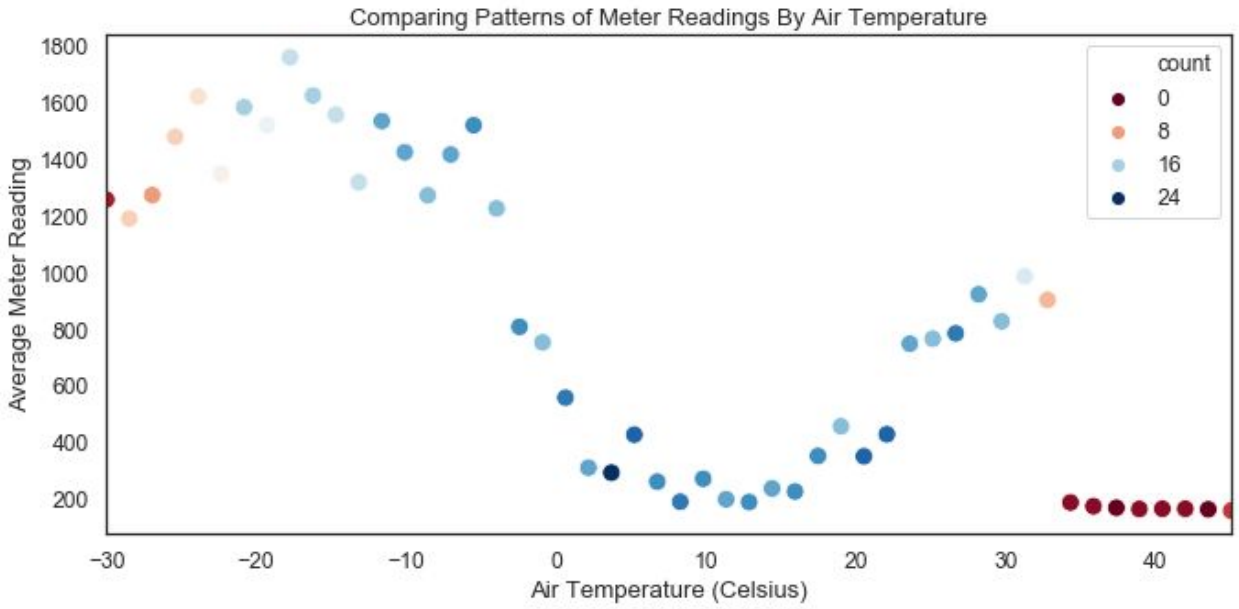
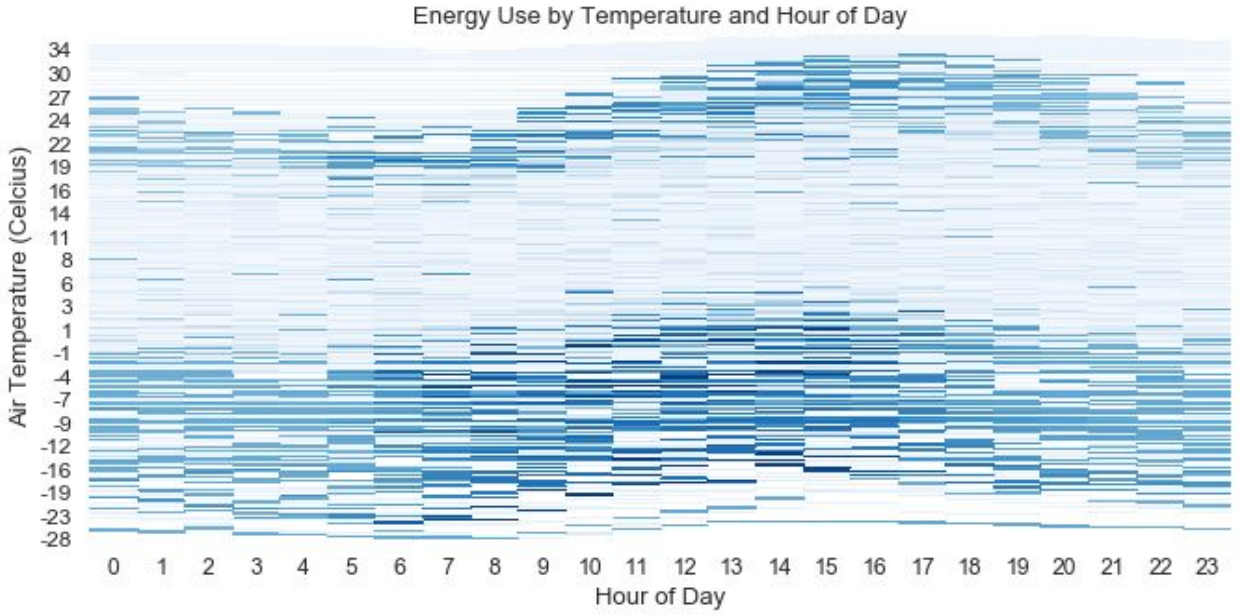


Exhibit 24: A heat map displaying variations in meter reading across different temperatures over the course of the day.



6. Model Building

6.1 Narrowing the scope to predicting the electricity meter

Based on our exploratory analysis above, we felt that we had a pretty good idea of what would or would not have a large impact on the electric meter reading. However, before we started building our predictive model, we decided to log transform our response variable to correct for its strong rightward skew. We also decided to quantify the relationships we had identified through our data exploration efforts, and find the correlation between response (meter_reading) and explanatory variables and within the explanatory variables. We first separated out numeric variables from categorical variables, then calculated Pearson's R for the numeric data. We used Crammer's V to evaluate the strength of association between categorical variables.

What are the building-related variables that correlate most with energy used for any particular building?

Based on our work in EDA, our hypothesized that location (i.e. site_id) primary use, time of day, and seasonality would most impact energy consumption.

To investigate this question, we calculated and compared the correlation between quantitative response and predictor variables. As shown in the correlation matrix below, we can see the location has the strongest correlation to meter readings, followed by square feet, age and primary use.

Which weather variable is most impactful for energy consumption?

We suspected that air temperature would be most helpful in predicting energy consumption. However, as shown in our correlation matrix below, this relationship did not appear when calculating correlations between individual variables. Interestingly, none of the weather data showed any significant correlations to the electricity meter readings. Perhaps these data are more helpful in predicting meter readings with different meter types or are more useful with different modelling techniques than we used here.

Exhibit 25: Matrix displaying the linear relationships between meter reading and our available categorical variables.

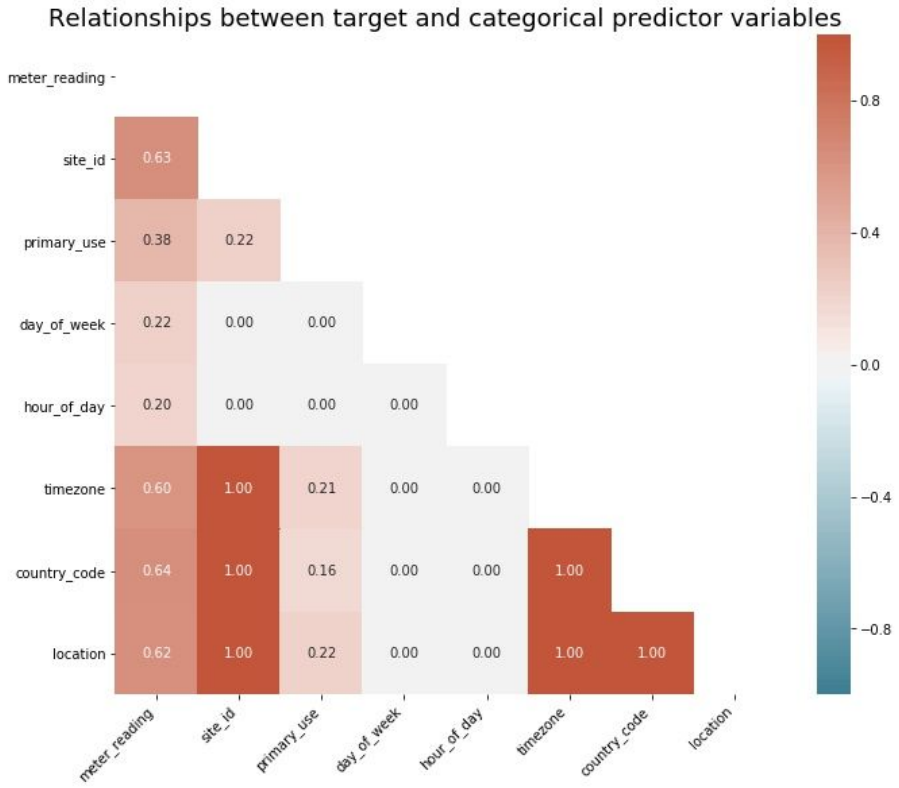
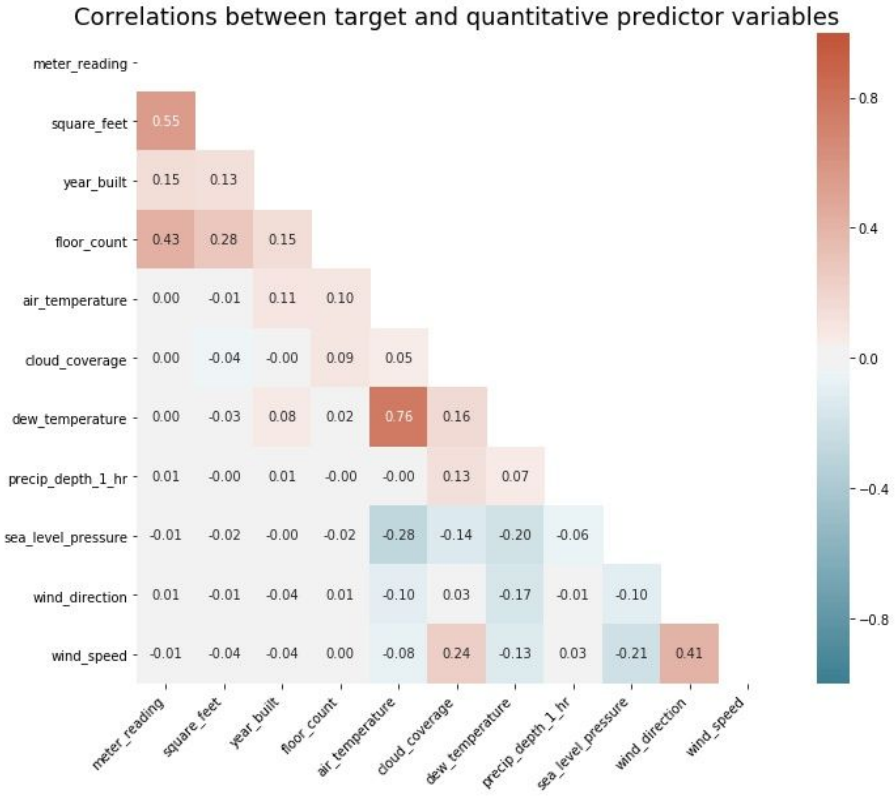


Exhibit 26: Matrix displaying the linear relationships between meter reading and our available numerical variables.



Based on correlation above, we note that variables most correlated with meter_reading are square_feet, floor_count, primary_use, site_id, day_of_week, and hour_of_day.

6.2 Data split for training

To evaluate the effectiveness of different models, we divided our data into training set and test sets with a 70/30 split. We used Root Mean Squared Logarithmic Error (RMSLE) to evaluate the effectiveness of our model on the test dataset. We note that the current top score on Kaggle leaderboard for this competition is 0.93. These scores are for models including all meter types and thus may not be a fair comparison; however, this is our best available benchmark to which to compare our model.

6.3 Model building

Model 1: A simple linear model

Based on the correlation table above, we decided to build a simple linear regression model using square feet, floor count, primary use, site id, day of the week, and hour of day. After converting primary use, site id, day of the week and hour of day to dummy variables, we have

65 explanatory variables in our model 1. Lastly, based on our observation of the positive skew of the meter reading data, we decided to log transform the response variable.

This model produced an adjusted R-Square of 93% which we consider to be pretty strong and with an F- statistic well below 5%, indicating that our model is statistically significant. We note that log transformation improved the explanatory power significantly as without this transformation the same model produced an R-square of 64%.

All of the variables used produced statistically significant coefficients that are not 0.

Exhibit 27: OLS regression results for Model 1.

OLS Regression Results

Dep. Variable:	meter_reading	**R-squared (uncentered):**	0.931
Model:	OLS	**Adj. R-squared (uncentered):**	0.931
Method:	Least Squares	**F-statistic:**	1.848e+06
Date:	Mon, 09 Dec 2019	**Prob (F-statistic):**	0.00
Time:	11:38:55	**Log-Likelihood:**	-1.3281e+07
No. Observations:	8408713	**AIC:**	2.656e+07
Df Residuals:	8408652	**BIC:**	2.656e+07
Df Model:	61		
Covariance Type:	nonrobust		

Using this model, we predicted the meter reading for the test dataset. Comparing results to actual data, we got a RMSLE score of 1.0133.

Model 2: Adding in temperature

For our second model, we added other variables that did not seem to have a strong correlation with energy use but intuitively should impact electricity usage. Specifically, we added in weather data by including air temperature - resulting in a total of 62 explanatory variables. We are using the same log-transformation of the response variable.

Adding air temperature data did not improve the model's explanatory power by much. R-Square remained around 93%. All of the variables used produced statistically significant coefficients that are not 0. RMSLE for this model is slightly better at 1.0127.

Exhibit 28: OLS Regression results for Model 2.

OLS Regression Results

Dep. Variable:	meter_reading	**R-squared (uncentered):**	0.931
Model:	OLS	**Adj. R-squared (uncentered):**	0.931
Method:	Least Squares	**F-statistic:**	1.820e+06
Date:	Mon, 09 Dec 2019	**Prob (F-statistic):**	0.00
Time:	11:58:30	**Log-Likelihood:**	-1.3278e+07
No. Observations:	8408713	**AIC:**	2.656e+07
Df Residuals:	8408651	**BIC:**	2.656e+07
Df Model:	62		
Covariance Type:	nonrobust		

Model 3: What if we throw all features in?

For the last model we decided to include all the clean variables we have (i.e. all building data and the temperature data). We also decided to build a linear regression model using SKLearn, a package used for machine learning. We again log transformed the response variable (meter reading).

SKLearn does not produce a pretty print out of the coefficient but we note the model improved RMSLE to 0.9824, which is close to the Kaggle leaderboard score of 0.93.

6.4 Summary

Exhibit 29 displays chart that shows a summary of variables we used and the RMSLE score we calculated for each model.

Exhibit 29: Matrix displaying the linear relationships between meter reading and our available categorical variables.

Type	Variables	Model 1	Model 2	Model 3
Meta	Square Ft	✓	✓	✓
	Floor Count	✓	✓	✓
	Primary Use	✓	✓	✓
	Site ID	✓	✓	✓
	Year Built			✓
Temporal	Day of the Week	✓	✓	✓
	Hour of Day	✓	✓	✓
Weather	Air Temperature		✓	✓
	Dew Temperature			
Results	RMSLE on test data	1.0133	1.0124	0.9824

7. Conclusion

After an extensive data cleaning effort, we were able to answer the main questions we had about electricity usage in buildings to a satisfying degree.

What are some factors that impact energy use?

During our exploratory data analysis, we found many interesting patterns when it came to meter readings and how the building was being used. We could see, for example, how meter readings would rise and fall over the course of a normal workday in a building, and that when people were out of the office during the weekend, overall levels would be much lower.

We also had some unexpected findings when it came to which types of buildings had the most energy usage -- buildings used primarily for healthcare and education purposes had meaningfully high meter readings compared to others.

Contrary to our expectations, the weather data did not have a strong relationship to meter reading patterns. The largest drivers of how much energy a building used were characteristics about the building (such as its floor count) and location. This is perhaps a positive finding for those interested in sustainability, as we have less control over the weather than we do over factors such as high our buildings are built. However, we did notice that for the buildings in more extreme, cold weather (in Canada), we saw much higher energy usage overall, suggesting that in real life air temperature does have an influence to some degree (even though this was not picked up by our model).

Can you predict energy use for a particular building?

Yes! We managed to create a relatively robust model with a very high R-squared value of 0.93 and excellent RMSLE of 0.9824.

Overall, we found these data to be a trove of insights into energy usage in buildings and are looking forward to continuing to clean these data as well as to extend our efforts to include the other meter types available in the data.

What does this imply for buildings trying to control their energy usage?

Given that we were successfully able to predict a building's electricity use from variables largely outside of their control -- such as their location, usage patterns, and size -- does this mean that it is a futile task for buildings to attempt to improve their energy usage? Well, not necessarily. Although we had a very high R-squared value of 0.93, it is still short of 1 which means that there is still variance in energy use remaining that was not explained by our model. In the plots below, we grouped buildings by their timezone (to represent general geographic location), size, and primary use. One can see that there is still a great deal of variation in their average daily electricity use, suggesting that there may be different practices across buildings that can be leveraged to improve their efficiency. For example, a larger building with the same primary use and in a similar area as a smaller building could have much lower average daily energy usage -- demonstrating that it is not impossible to improve! Furthermore, having a model of baseline energy usage (such as the one described above) can form the basis of data-driven decision-making as to which improvements should be made and evaluating their effects.

Exhibit 30: A comparison of average meter reading across timezone, primary use, and size.