Project 2 Proposal

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

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 some building specific as well as external factors (e.g. weather).

Target Questions

The main questions we want to answer is can you predict energy use for a particular building? What are some factors that impact energy use?

While we may not be able to answer these questions fully or to be competitive in the Kaggle competition (the evaluation is based on RMSLE) given to do so probably requires some ML expertise, we hope to push our limit in knowledge and do as much as we can with the dataset.

Source Data

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

There are several components to the datasets (https://www.kaggle.com/c/9994/download-all).

building_metadata.csv

- This dataset contains information about buildings from which the meter readings were collected.
- Shape: 1449 rows x 6 columns
- Variables:
    - site_id
    - building_id
    - primary_use
    - square_feet
    - year_built

    - floor_count
- Data snapshot:

```
1 building_df.head(5)
```

	site_id	building_id	primary_use	square_feet	year_built	floor_count
0	0	0	Education	7432	2008.0	NaN
1	0	1	Education	2720	2004.0	NaN
2	0	2	Education	5376	1991.0	NaN
3	0	3	Education	23685	2002.0	NaN
4	0	4	Education	116607	1975.0	NaN

weather_train.csv & weather_test.csv

- These datasets contain hourly weather data from a meteorological station as close as possible to the site.
- weather_train.csv shape: 139773 rows x 9 columns
- weather_test.csv shape: 277243 rows x 9 columns
- Variables:
    - site_id
    - timestamp
    - air_temperature
    - cloud_coverage
    - dew_temperature
    - precip_depth_1_hr
    - sea_level_pressure
    - wind_direction
    - wind_speed
- Data snapshot:

```
1 weather_train_df.head(5)
```

	site_id	timestamp	air_temperature	cloud_coverage	dew_temperature	precip_depth_1_hr	sea_level_pressure	wind_direction	wind_speed
0	0	2016-01-01 00:00:00	25.000000	6.0	20.00000	NaN	1019.5	0.0	0.000000
1	0	2016-01-01 01:00:00	24.406250	NaN	21.09375	-1.0	1020.0	70.0	1.500000
2	0	2016-01-01 02:00:00	22.796875	2.0	21.09375	0.0	1020.0	0.0	0.000000
3	0	2016-01-01 03:00:00	21.093750	2.0	20.59375	0.0	1020.0	0.0	0.000000
4	0	2016-01-01 04:00:00	20.000000	2.0	20.00000	-1.0	1020.0	250.0	2.599609

train.csv

- This dataset includes hourly meter reading collected from each building
- Shape: 20216100 rows x 4 columns
- Variables
    - buildling_id
    - meter - {0: electricity, 1: chilled water, 2: steam, 3: hotwater}

    - timestamp
    - **meter_reading - (in kWh) the variable we are trying to predict**
- Data snapshot:

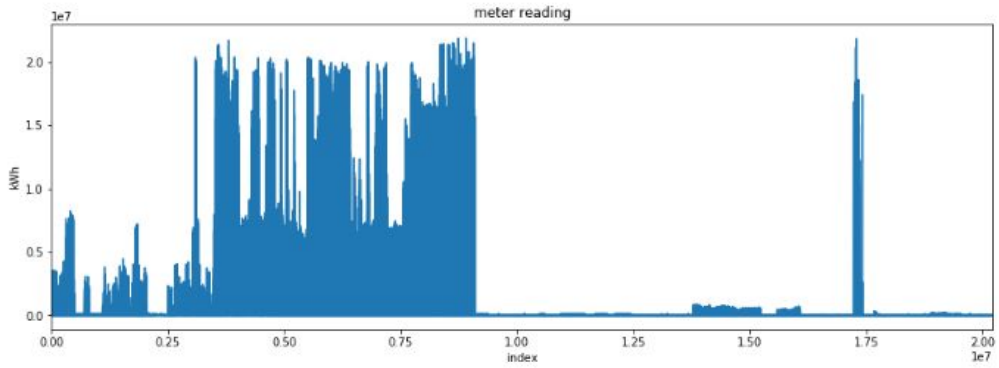
```
1 train_df.head(5)
```

	building_id	meter	timestamp	meter_reading
0	0	0	2016-01-01 00:00:00	0.0
1	1	0	2016-01-01 00:00:00	0.0
2	2	0	2016-01-01 00:00:00	0.0
3	3	0	2016-01-01 00:00:00	0.0
4	4	0	2016-01-01 00:00:00	0.0

Initial Exploration and Data Preparation

Since the datasets are a part of a Kaggle competition, they are generally clean (i.e. we do not expect significant amount of missing data) and the weather data already corresponds well with the building site. That being said, given that this is real data with measurement error, we expect to see some data anomalies.

A quick view of the **meter_reading** data. This is the variable we wish to predict and thus one of the most important variables we will be looking at. There seems to be some large spikes in the dataset, which we wish to explore further.



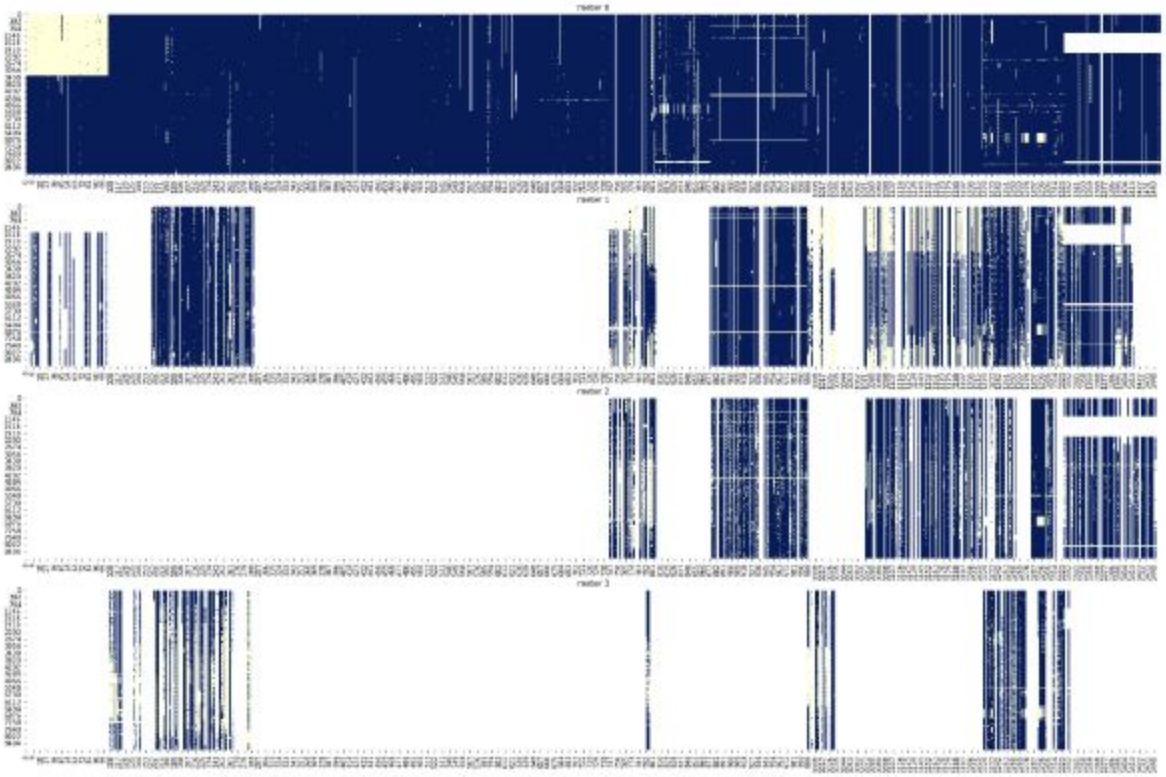
There does not seem to be any NaN data points but there are a number of 0 readings. While we think some of these 0 readings represent no energy used (e.g. at 2am in the morning), we suspect some of these 0 readings are actually missing data.

Below we plot the meter_reading data for 0 values and non-zero values (x-axis represents the 1448 buildings and the y-axis represents time). The graph validates our view that some of these 0 values are missing data. The yellow horizontal lines seen in the first chart is particularly

concerning (i.e. multiple consecutive buildings have 0 readings at the same time). We will need to think about what to do with these “missing” data points.

```
train = train_df.set_index(['timestamp'])
train.head(5)
print ('Number of buldings:', str(len(set(train['building_id']))))
print('% of meter reading that is 0:',\
     "{:.1%}".format(len(train[train['meter_reading']==0])\
          /len(train['meter_reading'])))
```

```
Number of buldings: 1449
% of meter reading that is 0: 9.3%
```



Evaluating the Quality of Data

We then proceed to evaluate the quality of the data of our dependent variables: building and weather data. In the following tables, we display the percent of missing data for each variable.

Weather Data

We see that temperature (both air and dew) data 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 the main variables to explore.**

	Total	Percent
cloud_coverage	69173	49.489529
precip_depth_1_hr	50289	35.979052
sea_level_pressure	10618	7.596603
wind_direction	6268	4.484414
wind_speed	304	0.217496
dew_temperature	113	0.080845
air_temperature	55	0.039350
site_id	0	0.000000
timestamp	0	0.000000

Building Data

Over two-thirds of the floor_count data is missing so we do not expect to use that information. We also believe square_feet can be a good proxy for floor_count. Our hypothesis is that older buildings are less energy efficient and bigger buildings use more energy, so **the most important variables will be year_built as well as square_feet.**

	Total	Percent
floor_count	1094	75.500345
year_built	774	53.416149
site_id	0	0.000000
building_id	0	0.000000
primary_use	0	0.000000
square_feet	0	0.000000

Final report planning

We plan to take a deeper look at the data we have and answer the following questions:

- How to interpret and deal with missing data?
- What are the main variables that correlate most with energy used for any particular building? We suspect age of the building, temperature, and size of the building are the three variables that will most impact energy consumption.

- But we will also take a look at building types and other variables:
    - Are older buildings more or less efficient in energy use than newer buildings?
    - Does the size of the building have a linear relationship with the amount of energy used?
    - Do different types of meters produce different levels of energy consumption?
    - Do building types have an impact on energy consumption?
    - Which weather variable is most impactful for energy consumption?
    - Are there seasonality of energy use? How does that correlate with seasonality in weather patterns?
- Based answers to the questions above, can we build a model to predict energy consumption using certain selection of dependent variables? We will try different models here to see which one is the most effective at predicting energy consumption.