Capstone 2 Milestone Report

Problem Statement

Q: How much does it cost to cool a skyscraper in the summer?

A: A lot! And not just in dollars, but in environmental impact.

Thankfully, significant investments are being made to improve building efficiencies to reduce costs and emissions. The question is, are the improvements working? That's where you come in. Under pay-for-performance financing, the building owner makes payments based on the difference between their real energy consumption and what they would have used without any retrofits. The latter values have to come from a model. Current methods of estimation are fragmented and do not scale well. Some assume a specific meter type or don't work with different building types.

Data

Assessing the value of energy efficiency improvements can be challenging as there's no way to truly know how much energy a building would have used without the improvements. The best we can do is to build counterfactual models. Once a building is overhauled the new (lower) energy consumption is compared against modeled values for the original building to calculate the savings from the retrofit. More accurate models could support better market incentives and enable lower cost financing.

Files

train.csv

- `building_id` - Foreign key for the building metadata.
- `meter` - The meter id code. Read as `{0: electricity, 1: chilledwater, 2: steam, 3: hotwater}`. Not every building has all meter types.
- `timestamp` - When the measurement was taken
- `meter_reading` - The target variable. Energy consumption in kWh (or equivalent). Note that this is real data with measurement error, which we expect will impose a baseline level of modeling error. UPDATE: as discussed here, the site 0 electric meter readings are in kBTU.

building_meta.csv

- `site_id` - Foreign key for the weather files.
- `building_id` - Foreign key for `training.csv`

- `primary_use` - Indicator of the primary category of activities for the building based on EnergyStar property type definitions
- `square_feet` - Gross floor area of the building
- `year_built` - Year building was opened
- `floor_count` - Number of floors of the building

weather_[train/test].csv

Weather data from a meteorological station as close as possible to the site.

- `site_id`
- `air_temperature` - Degrees Celsius
- `cloud_coverage` - Portion of the sky covered in clouds, in oktas
- `dew_temperature` - Degrees Celsius
- `precip_depth_1_hr` - Millimeters
- `sea_level_pressure` - Millibar/hectopascals
- `wind_direction` - Compass direction (0-360)
- `wind_speed` - Meters per second

test.csv

The submission files use row numbers for ID codes in order to save space on the file uploads. `test.csv` has no feature data; it exists so you can get your predictions into the correct order.

- `row_id` - Row id for your submission file
- `building_id` - Building id code
- `meter` - The meter id code
- `timestamp` - Timestamps for the test data period

Data Cleaning

Inspect Missing Data

```
inspect_missing(train_data)
```

	Total	Percent
building_id	0	0.0
meter	0	0.0
timestamp	0	0.0
meter_reading	0	0.0
meter_type	0	0.0

```
inspect_missing(weather_train)
```

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

```
inspect_missing(building_metadata)
```

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

Interpolation is used for dealing with missing values esp in weather data.

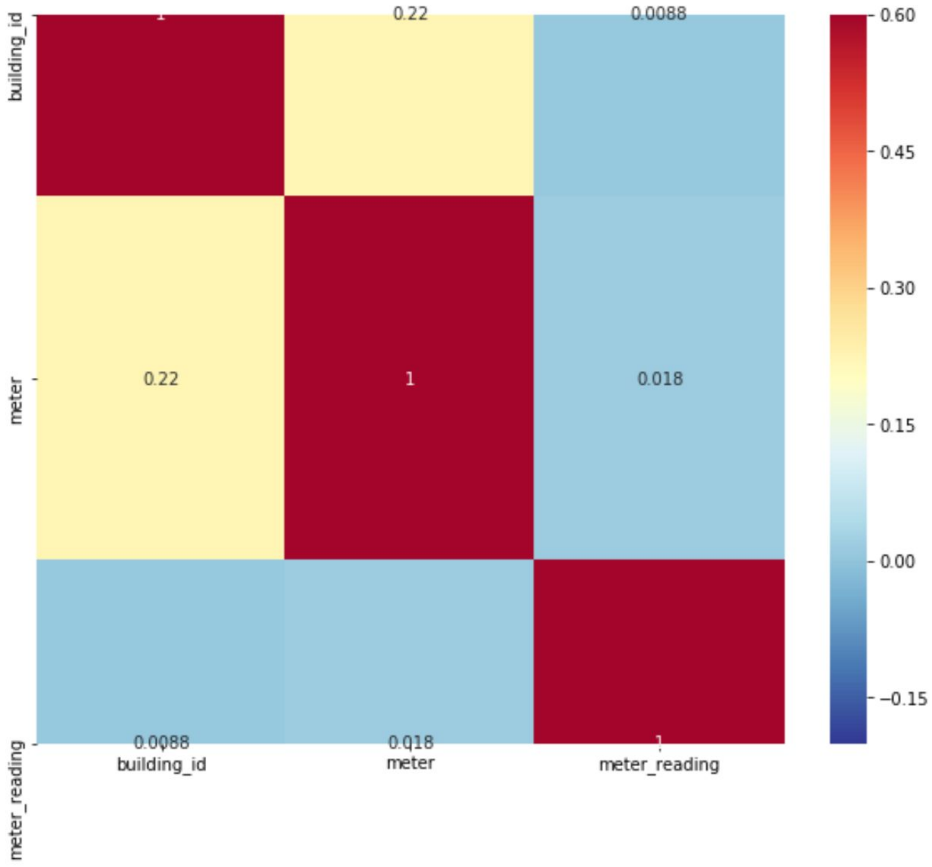
Interpolating Weather Data

```
weather_train = weather_train.groupby('site_id'
weather_train.groupby('site_id').apply(lambda g
```

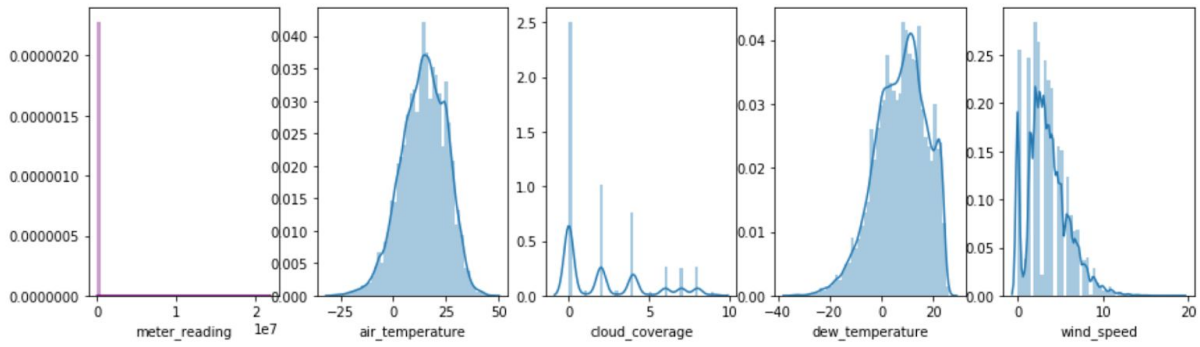
	site_id	timestamp	air_temperature	cloud_coverage	de
site_id					
0	0	0	0	0	
1	0	0	0	0	
2	0	0	0	0	
3	0	0	0	0	

EDA

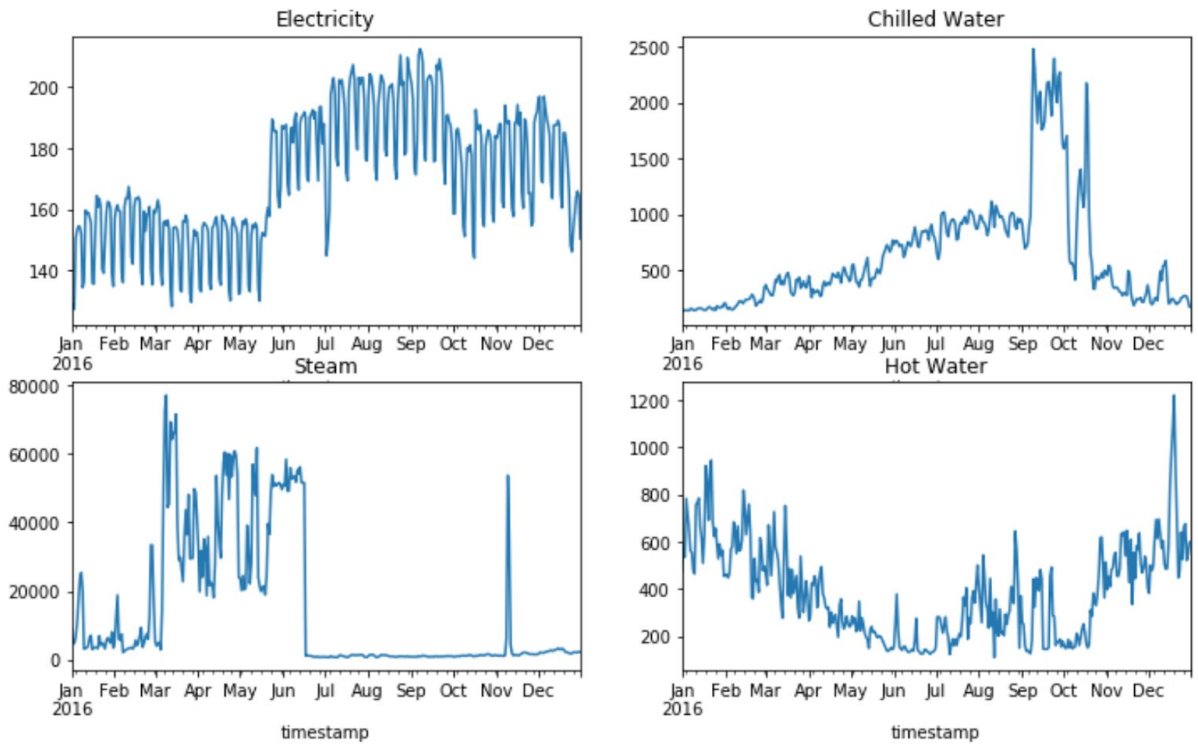
Correlation Matrix



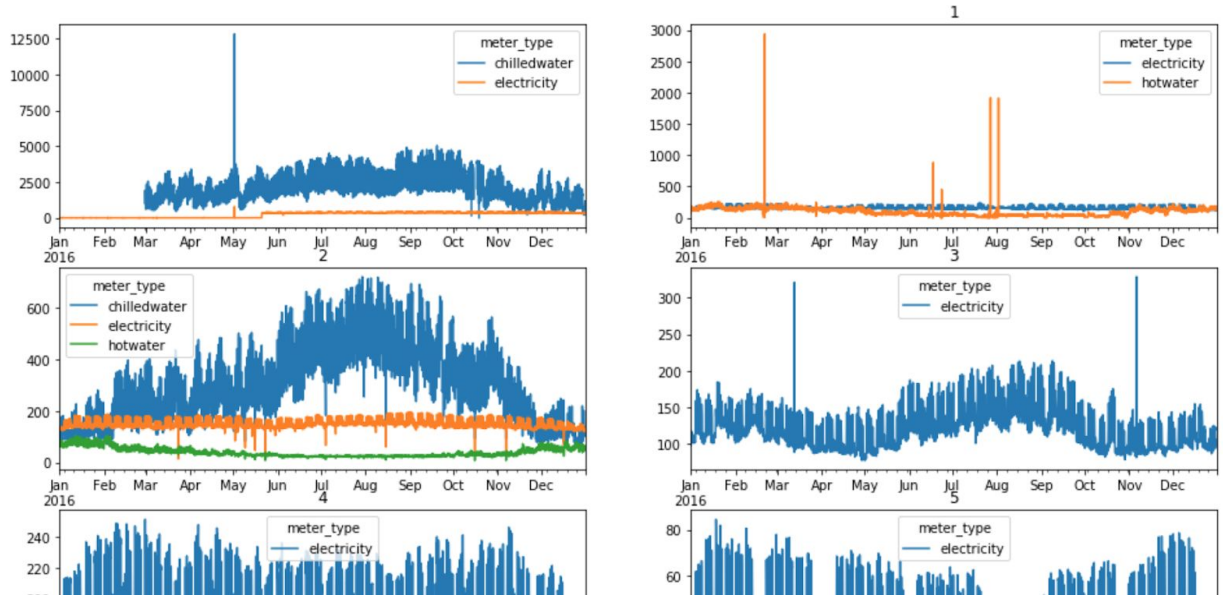
Distribution Plot



Meter wise time series analysis

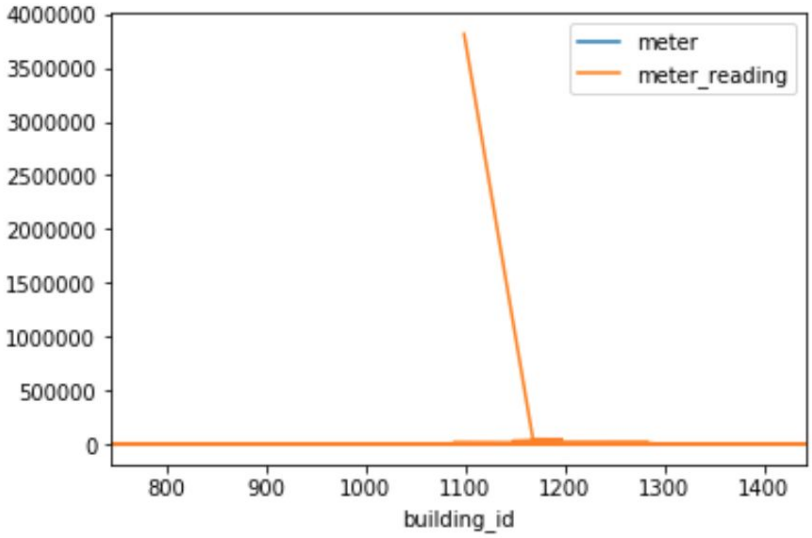


Site id wise meter readings

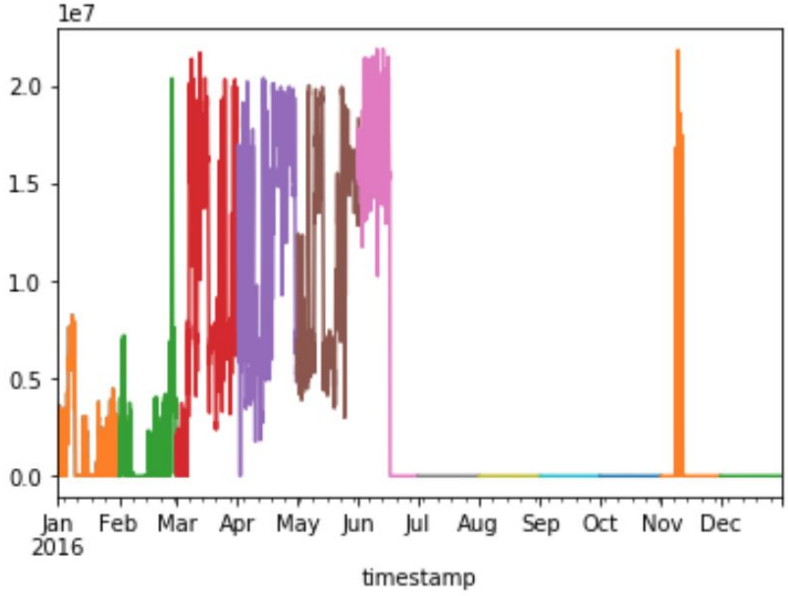


Outlier Detection

Building id wise meter reading



High meter usage for building id 1099



Missing Data for Site id 0