ASHRAE Kaggle Competition

Alissa Stover, Sophia Skowronski, Ying Hua

Introduction

Purpose of competition[1]: create a baseline model for which to compare improvements buildings make to their energy efficiency

Data available for competition: building characteristics, meter readings, weather data

[1]https://www.kaggle.com/c/ashrae-energy-prediction

Data Journey

What are factors that impact energy use? Can you predict it for particular buildings?

Data cleaning

Subset data from 20,216,100 rows → 12,060,910 rows
Merge datasets and clean → 12,060,311 rows

Data exploration and feature engineering

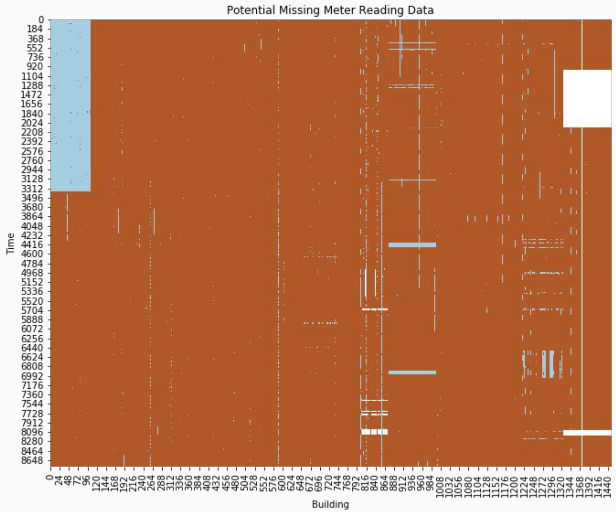
Univariate statistics, examine distributions, create new variables (16 columns → 29 columns)

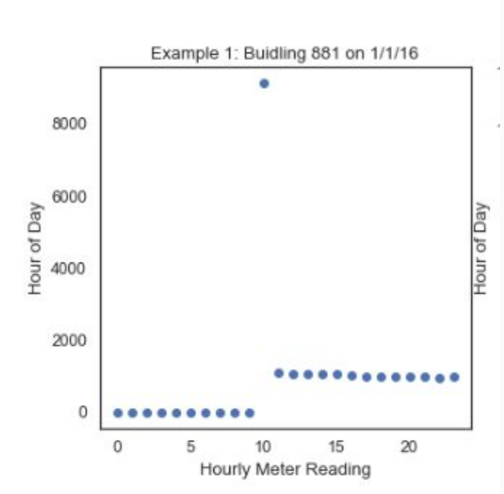
Data modeling

Examining linear relationships between variables & evaluated models with Root Mean Squared Logarithmic Error (RMSLE)

Data prep - Noticed data anomalies

Erroneous data: Some disguised as 0 meter reading; random spike in energy reading





Data prep - Dealing with missing variables

Different Treatments for different types of missing data

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 floor_count	75% 53%	Imputation using KNN	• Also tried other methods of imputation including linear regression, Naive Bayes. KNN gave the best results
	Weather Variables			
	air_temperature dew_temperature	0.03% 0.08%	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
	cloud_coverage precip_depth_1hr sea_level_pressure wind_direction wind_speed	49% 36% 7% 4% 0.2%	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

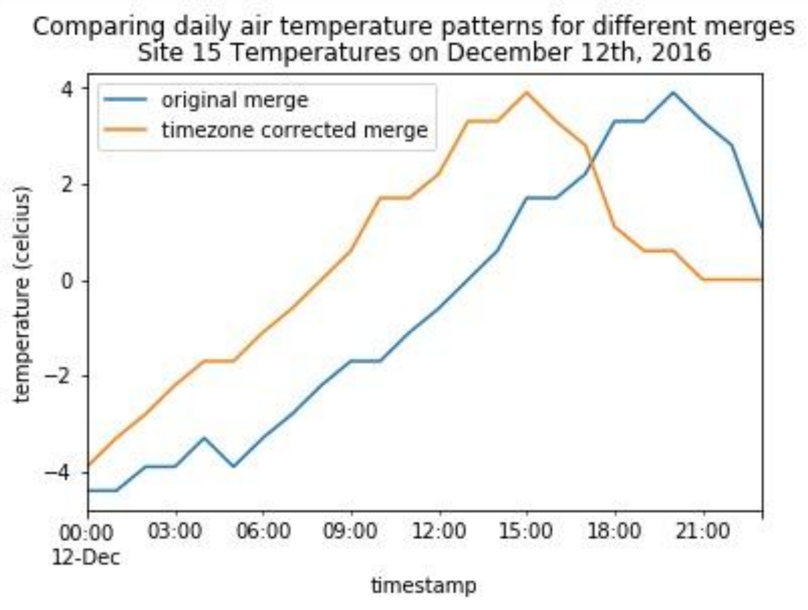
Data prep - Timezone merge

Meter data:

- Variables: building_id, meter, timestamp, meter_reading
- If merged w/ Building data: site_id

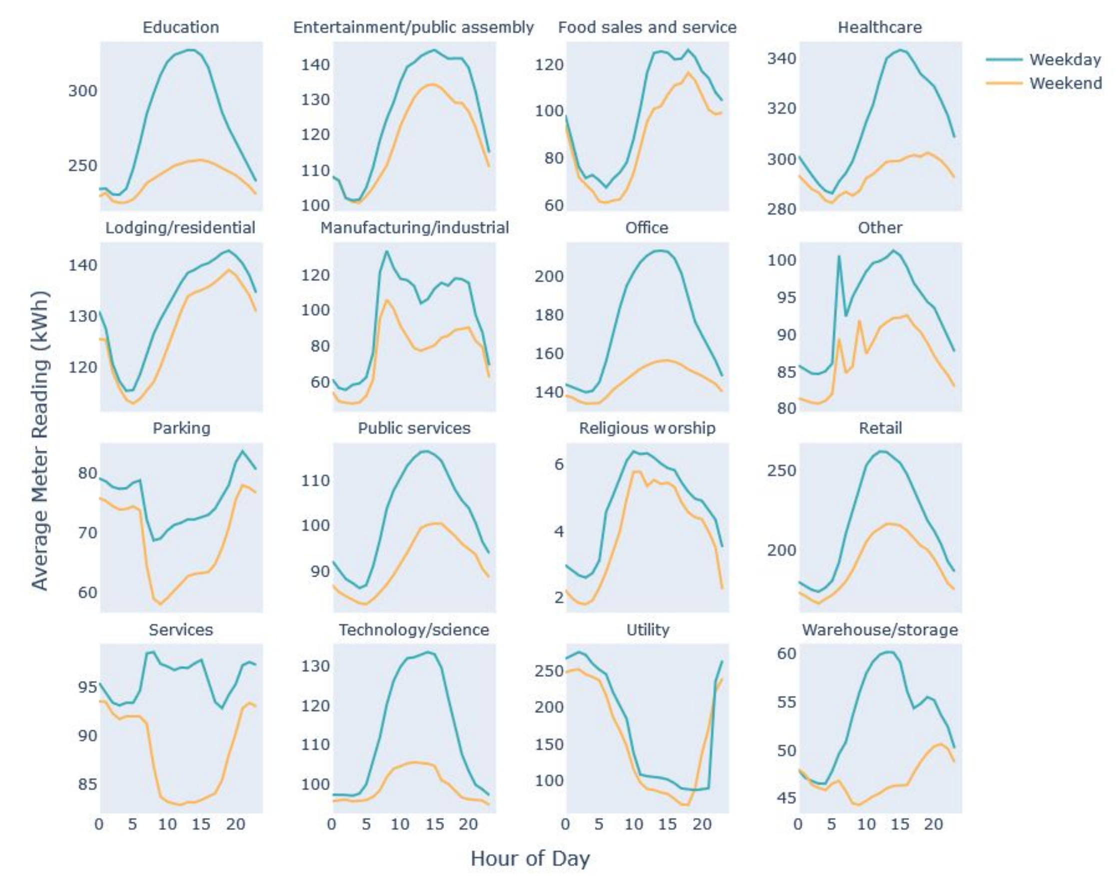
Weather data:

- Variables: site_id, timestamp, air_temperature, cloud_coverage, dew_temperature, precip_depth_1_hr, sea_level_pressure, wind_direction, wind_speed

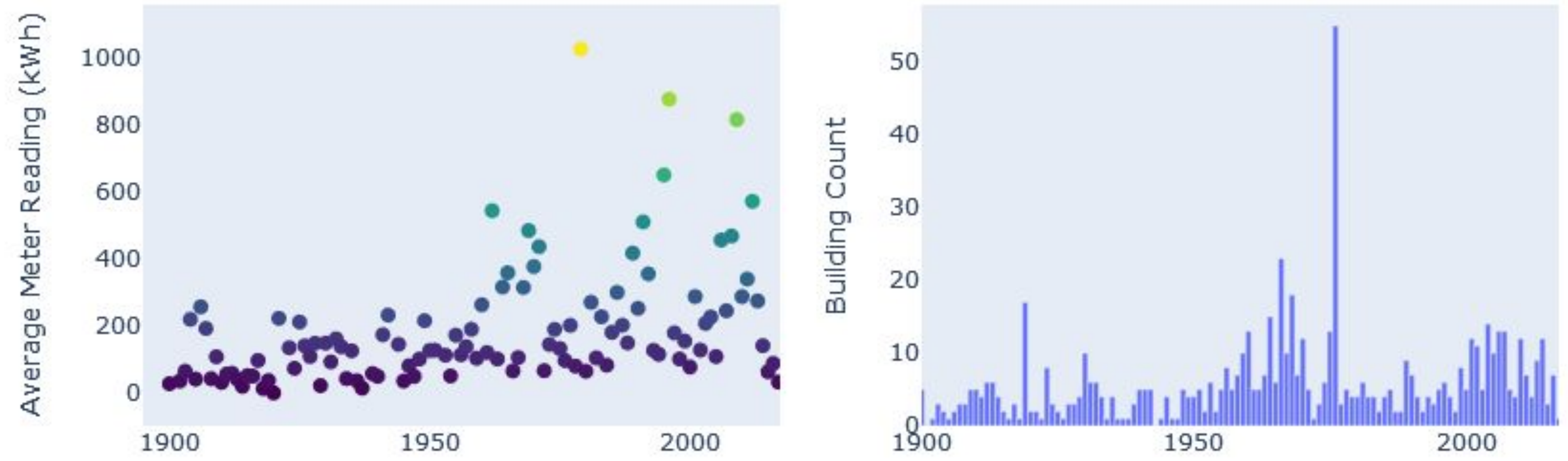


Interesting observations from EDA

Daily patterns vary significantly between primary uses

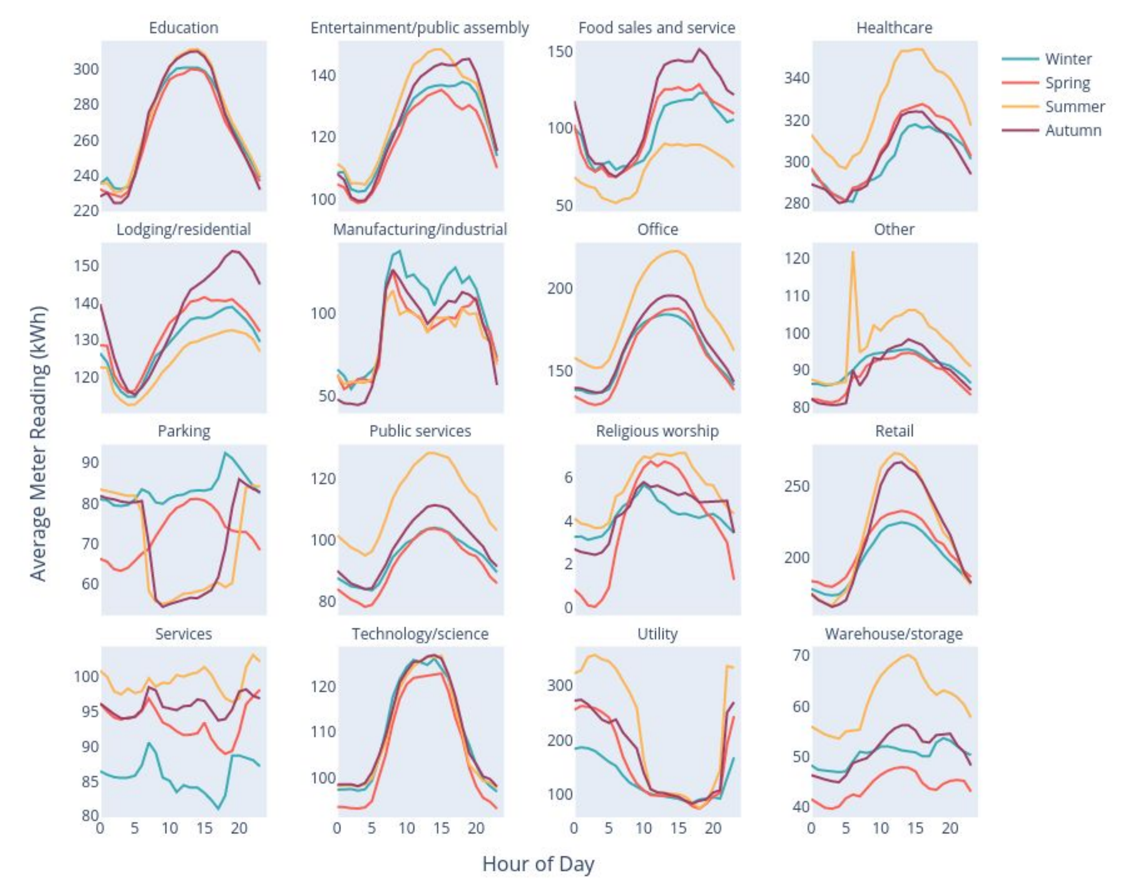


Interesting observations from EDA



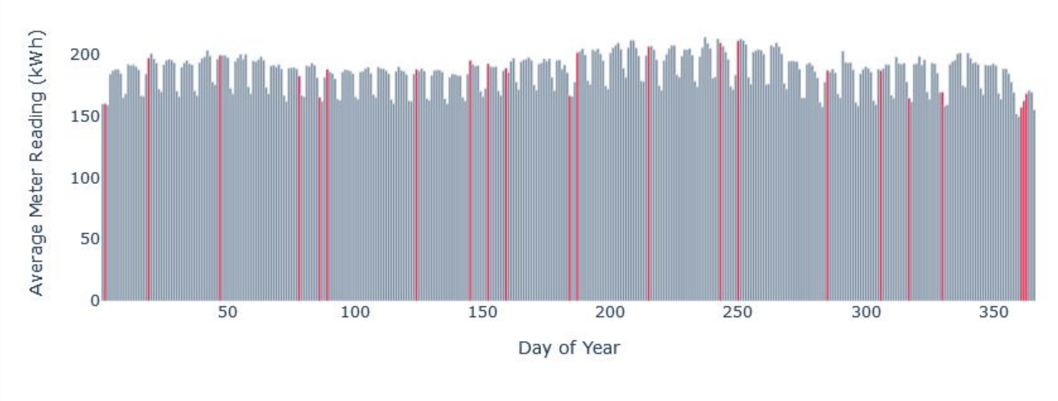
Interesting observations from EDA

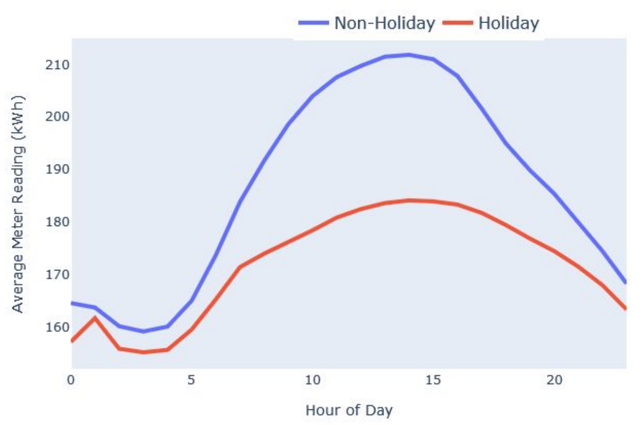
Daily Trends by Season



Interesting observations from EDA

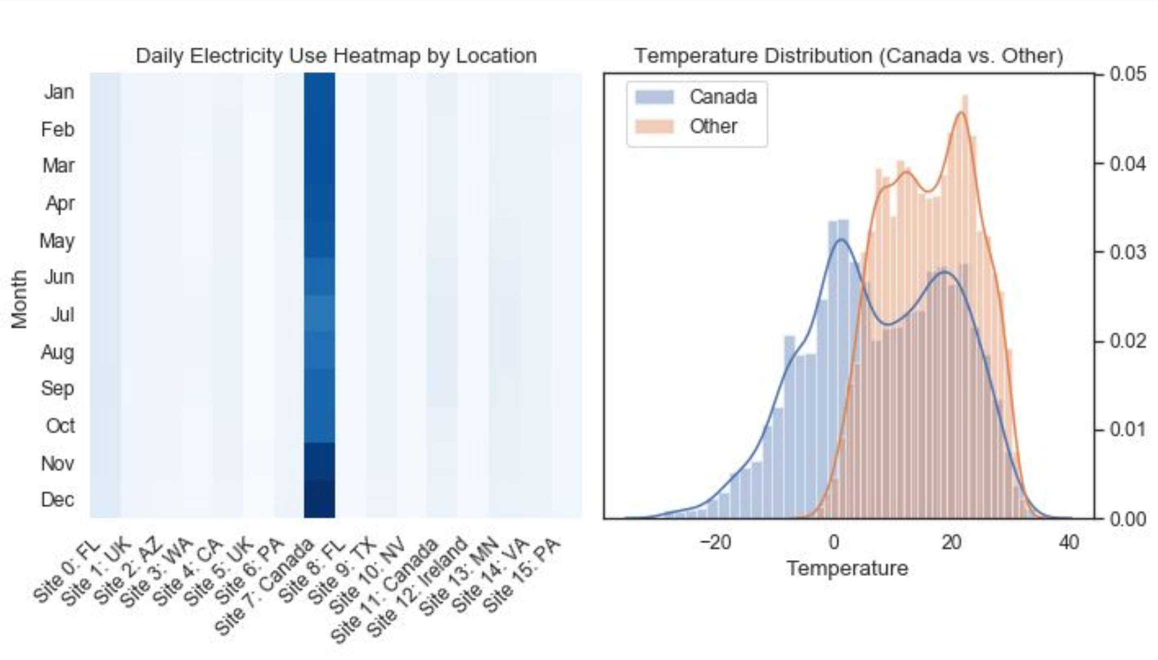
Holidays





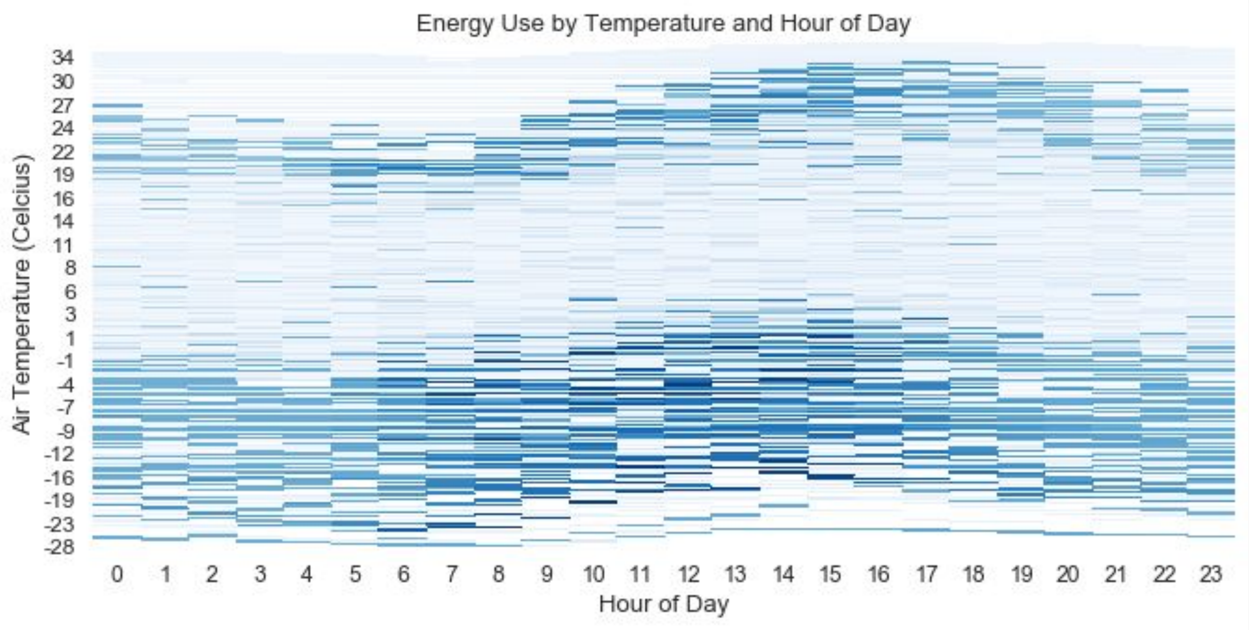
Interesting observations from EDA

Cold weather driving high energy consumption

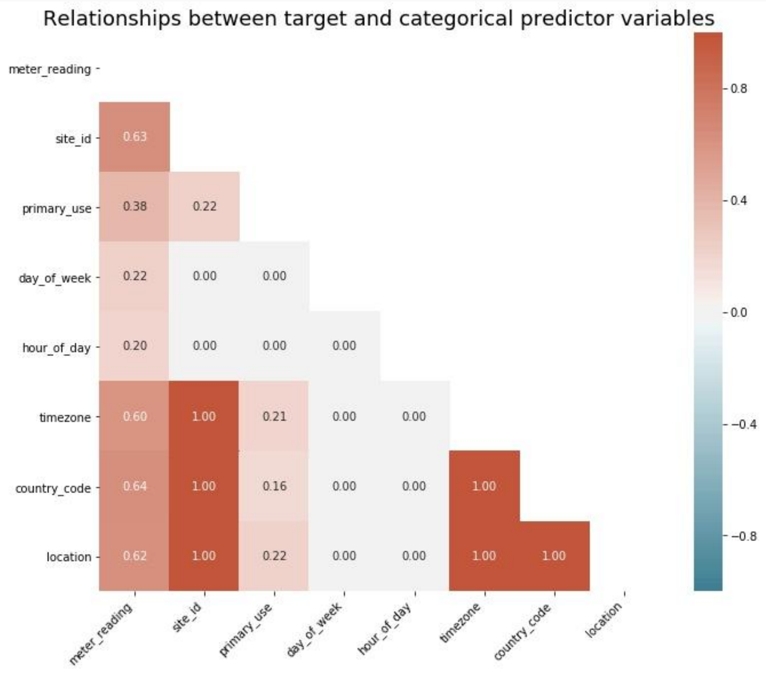


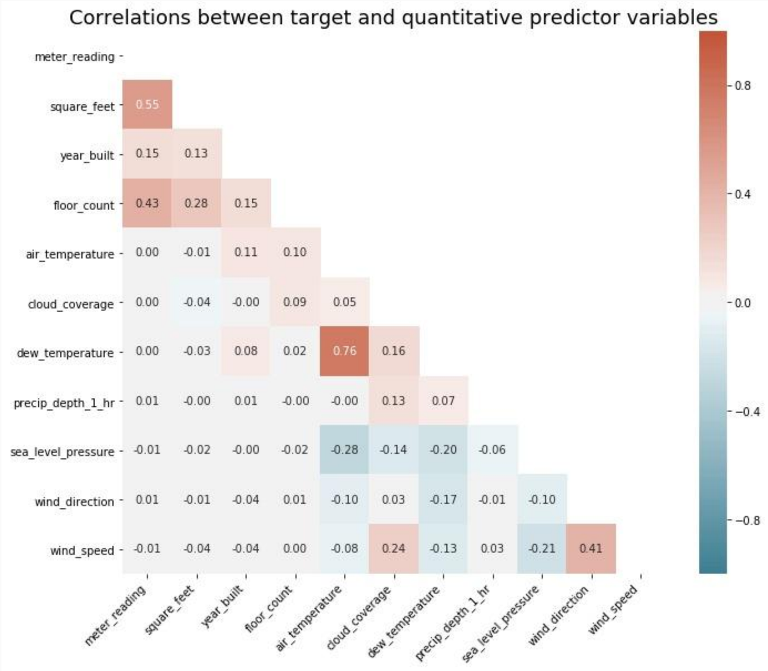
Interesting observations from EDA

Interesting patterns when we look at electricity use by temperature & time



Model Building





Model Building

3 predictive models and good RMSLE scores vs. Kaggle leaderboard (0.93)

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

Key takeaways

- What they say is true: data preparation accounts for 80% of work
- Datetime methods for time series
- Different plotting tools
  - Seaborn is pretty awesome!
  - Plotly is a good intro to interactive visualization
  - Matplotlib(a classic)
- Dabbled a bit into sklearn and ML
  - For imputing missing variables (KNN, Naive Bayes, Linear regression)
  - Building models (tried random forest but ran out of memory)

If we had more time...

- Examine and model all meter types (not just electricity)
- Examine in more detail weather data we did not include in our model (cloud coverage, sea level pressure, wind, etc)
- Try other model building techniques (higher order, ML models)
    - Attempted random tree but ran out of memory