Background:

The *ASHRAE Great Energy Predictor III* competition revolves around developing accurate predictions of metered building energy usage in the following areas: electric, chilled water, hot water, and steam meters. These predictions will be built using variables like the year a building was built, the size of the building (square footage), meteorological data, the number of floors of the building, and primary use. With better estimates of energy-saving investments, investors and institutions will be more inclined to invest in eco-friendly features and upgrades.

The sponsor of the Kaggle competition is the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), an American professional association pursuing the advancement of heating, ventilation, air conditioning and refrigeration. A core component to ASHRAE's purpose pertains to the development and publishing of technical standards to improve building services engineering, energy efficiency, indoor air quality, and sustainable development.

Dataset:

ASHRAE's Great Energy Predictor included five separate data sources from over 1,000 buildings over a three-year timeframe. Fortunately for us, two of these datasets pertained to the test datasets in which ASHRAE would measure performance. We discarded the testing datasets altogether since they did not include meter readings, the variable we are interested in predicting.

Discarding the testing datasets left us with three data sources. The first data source contained information regarding the individual buildings metadata. Buildings metadata contains a site identifier, building identifier, the primary use of the building (education, office, retail, etc.), square footage, year built, and number of floors. The second data source included information such as the building identifier, meter type (electricity, chilled water, steam, hot water), a timestamp, and energy consumption. The final data source contained weather data from nearby meteorological station such as site identifier, air temperature, cloud coverage, dew temperature, precipitation depth, sea level pressure, wind direction, and wind speed.

Existing Notebook Review

A Deep Dive EDA Into All Variables by Jason Zivkovic

Prior to conducting exploratory data analysis (EDA), the train, building medadata, and weather datasets were joined to create a dataframe with 20,216,100 rows and 22 columns.

Missing Data

Missing data was an unavoidable issue with the provided datasets, specifically with missing building and weather variables. This issue will be addressed later on but should not be overlooked.

```r
colSums(is.na(combine_train))
```

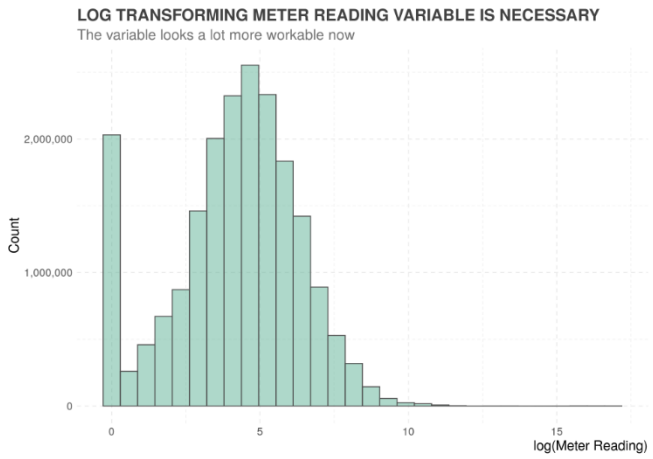
```
##          building_id                meter            timestamp
##                    0                    0                    0
##        meter_reading              site_id          primary_use
##                    0                    0                    0
##          square_feet           year_built          floor_count
##                    0             12127645             16709167
##      air_temperature       cloud_coverage      dew_temperature
##                96658              8825365               100140
##    precip_depth_1_hr   sea_level_pressure       wind_direction
##              3749023              1231669              1449048
##           wind_speed       timestamp_date      timestamp_month
##               143676                    0                    0
##        timestamp_day timestamp_day_number         time_ymd_hms
##                    0                    0                    0
##            time_hour
##                    0
```

Variable Analysis

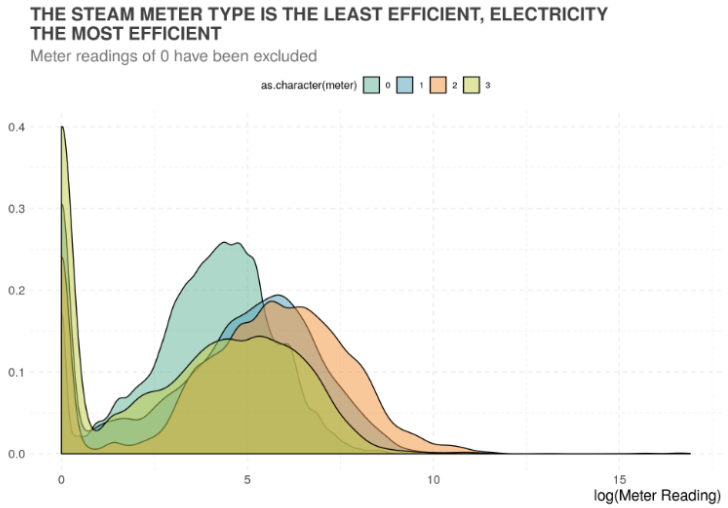
Zivkovic's exploration of each variable began with visualizations of said variables distribution. Meter reading is the most noteworthy of all variables explored, it was heavily skewed by several outliers. The code and output of meter reading's histogram is below. Since there was no way Zivkovic could investigate these outliers he ultimately eliminated all observations greater than 2 interquartile ranges above the mean. However, after doing so the meter reading variable was still heavily skewed and ended up requiring a logarithmic transformation.

Hide

```r
combine_train %>%
  ggplot(aes(x= log(meter_reading + 1))) +
  geom_histogram(alpha = 0.5, fill = "#5EB296", colour = "#4D4D4D") +
  scale_y_continuous(labels = comma) +
  ggtitle("LOG TRANSFORMING METER READING VARIABLE IS NECESSARY", subtitle = "The variable looks a lot more workable now") +
  labs(x= "log(Meter Reading)", y= "Count")
```



Zivkovic utilized several other variables to produce some interesting visualizations. The first visualization I'd like to talk about is meter type and meter reading. Producing this plot showed a distinct correlation between meter type and the amount of energy consumed, steam meter type tends to have higher readings while electricity tends to have the lowest. Zivokovic also created visualizations showing the relationship between energy usage and day, time of day and energy usage, and several other histograms using the buildings primary use, square footage, the year each building was built.



While this notebook didn't go into actually building and tweaking a predictive model, it did provide some great insights into issues that we should be mindful of when we start our machine learning process.

https://www.kaggle.com/jaseziv83/a-deep-dive-eda-into-all-variables