# 

# ASHRAE ENERGY PREDICTOR

University of Waterloo – Statistics for Data Science – Group Assignment

GROUP 6

Darwing Cara

David Kobayashi

Eric Koritko

Stefan Lazarevic

Patrick McDonnell

December 9, 2019

# EXECUTIVE SUMMARY

# TABLE OF CONTENTS

[1.0 OBJECTIVES 1](#_Toc15290634)

[1.1 ANALYSIS GOALS 1](#_Toc15290635)

[1.2 STARTING HYPOTHESES 1](#_Toc15290636)

[2.0 DATA PREPARATION 2](#_Toc15290637)

[2.1 DATA SOURCE 2](#_Toc15290638)

[2.2 DATA QUALITY 2](#_Toc15290639)

[2.3 PREPARATION FOR ANALYSIS - REDUCING MEMORY USAGE 2](#_Toc15290640)

[2.4 PREPARATION FOR ANALYSIS – ELIMINATING NULL VALUES 3](#_Toc15290641)

[2.5 PREPARATION FOR ANALYSIS – ENGINEERING ADDITIONAL FEATURES 3](#_Toc15290642)

[3.0 VARIABLE SELECTION 3](#_Toc15290643)

[3.1 EXAMINE ENERGY CONSUMPTION THROUGHOUT THE YEAR 3](#_Toc15290644)

[3.2 EXAMINE WEATHER FEATURES 5](#_Toc15290645)

[3.3 EXAMINE SURFACE AREA 6](#_Toc15290646)

[4.0 ANALYSIS 4](#_Toc15290643)

[4.1 PERCENTAGE OF COMPUTERS WITH MALWARE IN THE DATASET 4](#_Toc15290644)

[4.2 CORRELATED FEATURES 4](#_Toc15290645)

[4.3 BASIC UNIVARIATE ANALYSIS 5](#_Toc15290646)

[4.4 UNVARIATE ANALYSIS: INFORMATION VALUE (IV) 7](#_Toc15290647)

4.5 MACHINE LEARNING RESULTS 8

[5.0 CONCLUSIONS 9](#_Toc15290649)

REFERENCES

APPENDIX A: SOURCE CODE (JUPYTER NOTEBOOK)

APPENDIX B: DESCRIPTION OF FEATURES FROM KAGGLE.COM

## LIST OF FIGURES

## LIST OF TABLES

# OBJECTIVES

## INTRODUCTION

As humans, many of us spend most of our time in the built environment – this includes time spent working, shopping, and going to school. These buildings are responsible for nearly 60% of the world’s electricity [1], and most of Canada’s energy production has a greenhouse gas footprint. Since greenhouse gasses are closely linked to climate change, it is in society’s best interest to understand energy use patterns in buildings (and, ultimately, develop ways to reduce energy use).

This project will explore the electricity consumption in several commercial office buildings, and will attempt to identify statistical patterns with characteristics about the building or outdoor environmental conditions.

## ANALYSIS GOALS

The main goal of this exercise is to use statistics and machine learning techniques to predict electricity consumptions in commercial office buildings. We will accomplish this by identifying statistical relationships between the building’s features, the outdoor air conditions, and the building’s electricity consumption (measured by one or more utility meters). Since we are working with time-series data, we will split the data into two parts. The first part will contain data from \*\*\* to \*\*\* (approximately \*\*\*% of the data) and will be used to train the model. The second part will contain data from \*\*\* to \*\*\* and will be used to evaluate the model.

\*\*\*Add quantitative measurement goal here \*\*\*

The secondary goal of this exercise is to, in general terms, identify what key building and outdoor features are the most influential predictors of electrical energy consumption.

## STARTING HYPOTHESES

In a commercial office building, electricity is typically consumed by heating, ventilation, and air conditioning systems (roughly 35%), overhead lighting (roughly 20%), and plug loads such as computers and appliances (roughly 45%). Based on prior knowledge and familiarity with buildings, we will suggest the following starting hypotheses:

* There will be a relationship between the outdoor air temperature and the building’s electricity consumption. Specifically, buildings will consume more electricity on hotter days.
* While there will be a relationship between the building’s size and its electricity consumption (this seems obvious), there will not be much of a relationship between electrical use intensity (electricity consumed per square foot) and building size
* There will be a relationship between the time of day and the building’s energy consumption

# DATA PREPARATION

## DATA SOURCE

We obtained our data from a Kaggle competition hosted by ASHRAE (American Society for Heating, Refrigeration, and Air Conditioning Engineers) [2]. The goal of this competition, hosted in November 2019, is to predict the energy use in various building types using historic usage rates and the observed weather.

The .csv dataset is accessible with a free Kaggle account.

The dataset is split into five files, but only three were used for the assignment: train.csv (toughly 1.4 GB), building\_metadata.csv (roughly 45 KB), and weather\_train (roughly 7.2 MB).

The train.csv dataset contains the historic usage rates for every building. It also contains a Building ID, and a Zone ID. The building\_metadata.csv dataset contains each buildings’ features, indexed by their building IDs. Likewise, the weather\_train.csv dataset contains the observed weather conditions in several climate zones around North America, indexed by a Zone ID.

A full description of the features can be found in Appendix B.

## DATA QUALITY

## PREPARATION FOR ANALYSIS - REDUCING MEMORY USAGE

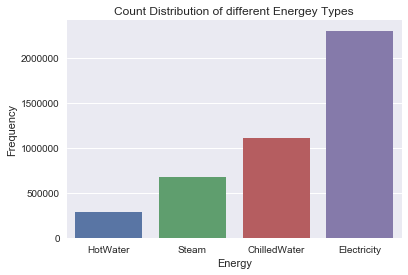
## PREPARATION FOR ANALYSIS – ELIMINATING NULL VALUES

## PREPARATION FOR ANALYSIS – ENGINEERING ADDITIONAL FEATURES

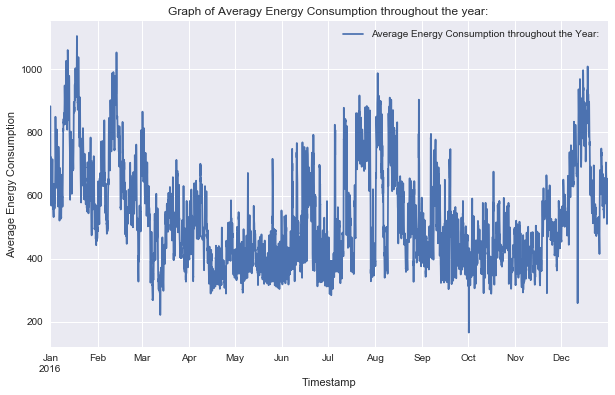
# VARIABLE SELECTION

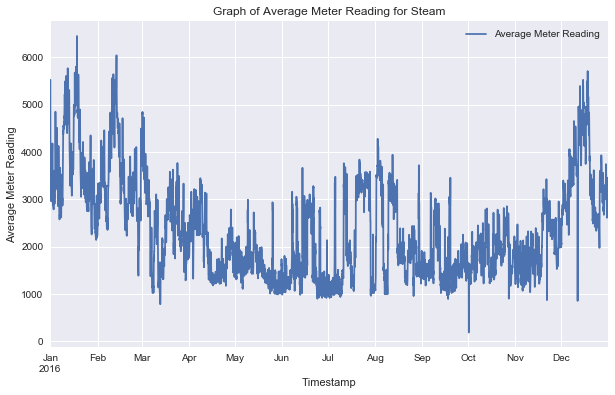
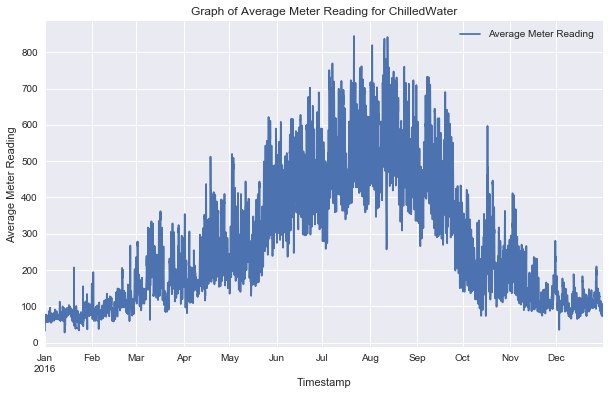
## EXAMINE ENERGY CONSUMPTION THROUGHOUT THE YEAR

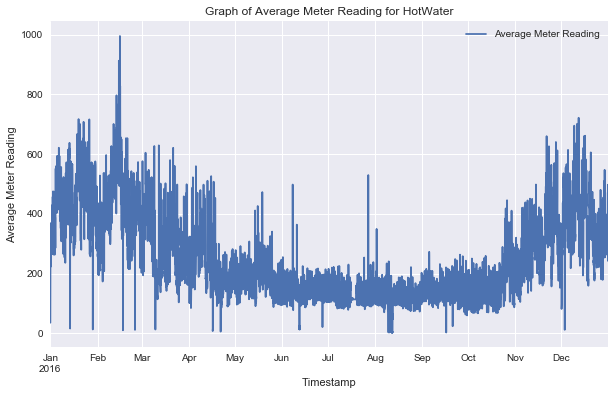
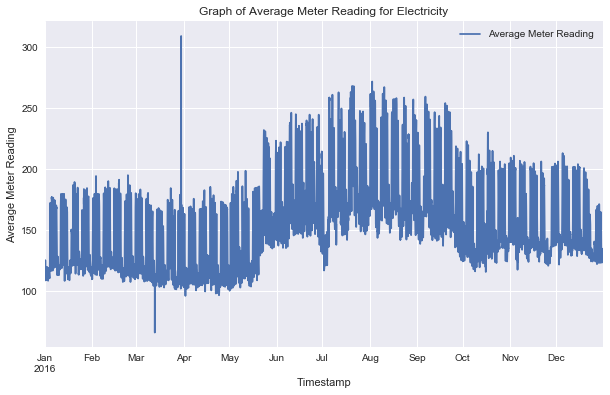
The first step for variable selection involved determining how different energy types behave throughout the year. This would be used to try and find any trends in the data, and transform them if needed in order to obtain trends. The first step was to find out the frequency of different energy types using the counterplot:



Based on the counterplot we can see that Electricity has by far the highest frequency, whereas HotWater has the lowest. Next, we plotted the average energy consumption for all energy types combined, and compared the graph with those of average consumption of each individual energy type:

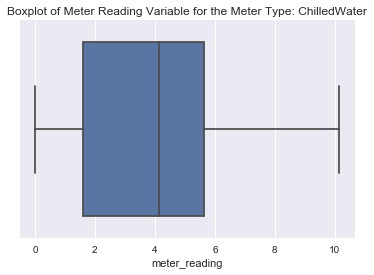
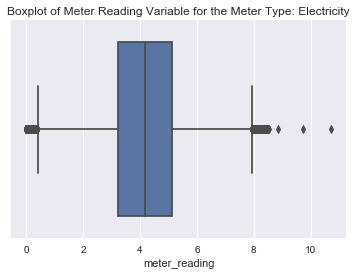


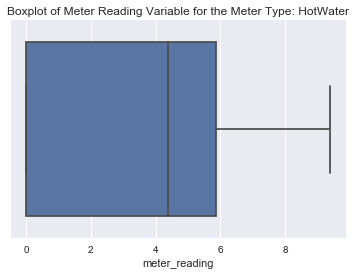
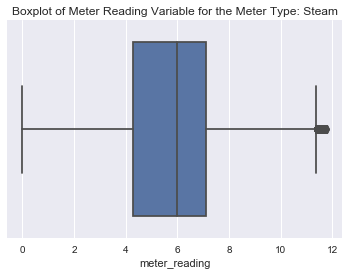




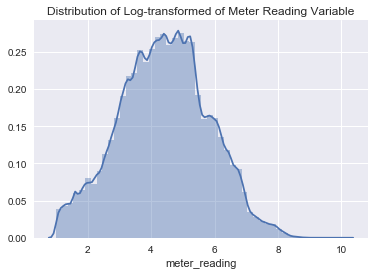
Based on the above graphs, we have determined that the Steam and HotWater follow a trend very similar to the overall trend, however Electricity and ChilledWater are different. However, we can see from the graphs that Steam has far greater values from the other energy types, and it sets the overall trend despite its low frequency. To that end, this energy type was removed from the data set.

After removing steam, the Box Plots were used for the remaining energy types to identify any outliers. The Box Plots have shown a large number of outliers that are less than 1, and greater than 7, especially for the Electricity energy type. These were also removed from the data sets, to get better results:



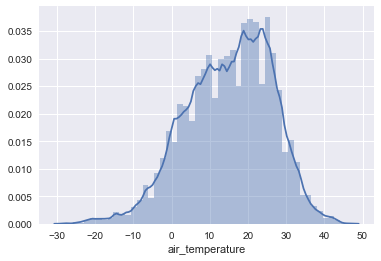
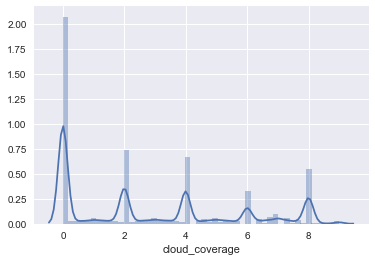
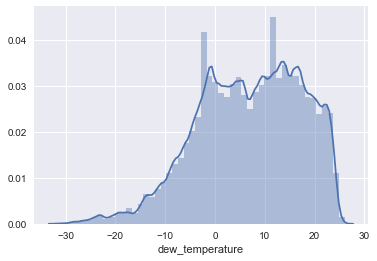
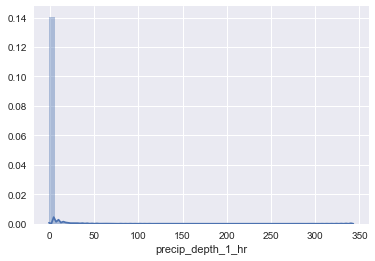
 

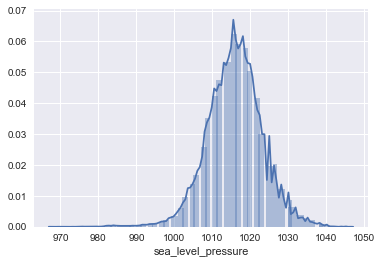
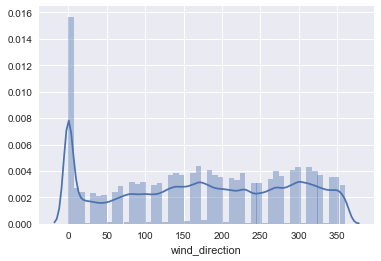
Finally, after these transformations, a near-normal distribution was obtained for the overall energy type consumption:



## EXAMINE WEATHER FEATURES

To examine the different features, we have used the distplots to see how they are distributed. The plots are presented below:

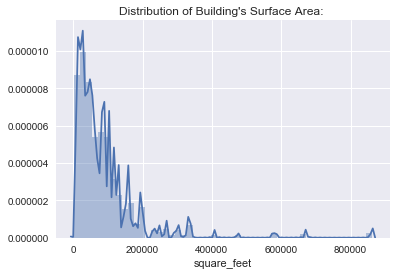
 

Based on the graphs, we can determine that the cloud coverage is different from others in that it is composed of a number of fixed values, and parcip\_depth\_1\_hr has a very large number of 0 values. To that end, these are not very useful for our calculations, and were removed from the data set. The remaining variables all had similar distributions, though they were skewed. The sea\_level\_pressure was the feature that followed the normal distribution the closest.

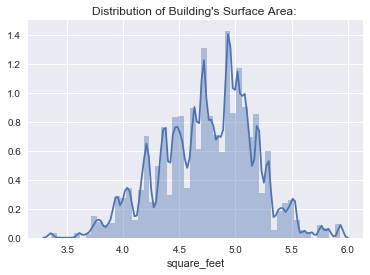
Next, the data frame corr() function was used to find the correlation between the remaining variables. The value of 0.9 was used as the threshold, and any features with the correlation higher than this were dropped as redundant. After applying the function, it was determined that the remaining variables are highly correlated, and thus only the sea\_level\_pressure was kept as it is closest to the normal distribution.

## EXAMINE SURFACE AREA

After plotting the surface area using the distplot, it was determined that it is negatively-skewed:



To that end, the feature was transformed using the log-transformation, and then it gained a near-normal distribution:



# CONCLUSIONS

# REFERENCES

[1] “Energy Use in Buildings,” Commerce Court. [Online]. Available: https://www.commercecourt.ca/sustainability/resources/energy-resources. [Accessed: 07-Dec-2019].

[2] “ASHRAE – Great Energy Predictor III,” Kaggle. [Online]. Available: https://www.kaggle.com/c/ashrae-energy-prediction. [Accessed: 07-Dec-2019].

# APPENDIX A: SOURCE CODE (JUPYTER NOTEBOOK)

# APPENDIX B: DESCRIPTION OF FEATURES FROM KAGGLE.COM

**From:** <https://www.kaggle.com/c/ashrae-energy-prediction/data>

**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).

**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
* 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**

* site\_id
* air\_temperature – Degrees Celsius
* cloud\_coverage – Portion of the sky covered in clouds, in [oktas](https://en.wikipedia.org/wiki/Okta)
* 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