# 

# 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 ANALYSIS 4](#_Toc15290643)

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

[3.2 CORRELATED FEATURES 4](#_Toc15290645)

[3.3 BASIC UNIVARIATE ANALYSIS 5](#_Toc15290646)

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

3.5 MACHINE LEARNING RESULTS 8

[4.0 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 electricity consumption
* Older buildings will tend to consume more electricity than newer buildings.

# 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 dataset is split into five files, but only three were used for the assignment: train.csv (roughly 620 MB), 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 Site 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 Site ID.

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

## REDUCING MEMORY USAGE

After loading the three datasets into memory, we ran a small function to reduce size of the dataframes.

When python and pandas is used to read and store data in a dataframe, it selects the following default datatypes for each column:

* 64 bit integers (int64), for numeric values with no decimal places;
* 64 bit floats (float64), for numeric values with decimal places; and
* Objects, for string values or variables with mixed data types.

This is a convenient approach, but not always an efficient use of memory. For example, both integers and floats have several different subtypes (16-bit integers, 16-bit floats, 32-bit integers, 32-bit floats, etc) that can be used instead. These subtypes are smaller and require less memory to store.

By looking at each feature’s maximum and minimum value, and then selecting a new (smaller) subtype, we were able to reduce train.csv’s dataframe size to roughly 290 MB.

## CLEANING THE WEATHER DATA

The train.csv dataset did not contain any duplicate rows, or any null values.

Almost all the features in the weather.csv dataset, on the other hand, contained at least some null values (some features had as many as 49% null values, and some as low as 0.039%). Due to the nature of time-series weather data, we felt it appropriate to fill these null values with an interpolation of adjacent values.

## FILTER AND MERGE DATA

This dataset contained records from many different building types: education, office, healthcare, etc. To narrow down the focus of this assignment, we decided to focus on office buildings. Office buildings made up the second most heavily populated building type in this dataset (next to education buildings). This meant that we dropped any records (from the train.csv dataset) that were not related to an office building.

This dataset also contained data from four different energy sources: electricity, chilled water, hot water, and steam. To further narrow down the focus of this assignment, we decided to focus on electricity consumption (mostly because most of the data in the train.csv dataset pertained to electricity records). Our next step was to drop any records (from the train.csv dataset) that were not related to electricity data.

Our final step was to use an inner merge on the dataframes by using the Building ID and Site ID as common features.

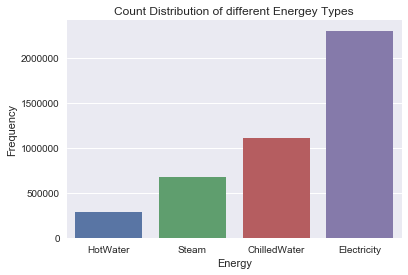
## CLEANING THE BUILDING METADATA

After merging the three dataframes into one, we checked for null values in features formerly from the building\_metadata.csv dataset. Over 90% of the floor\_count records were null, and over 75% of the year\_build records were null. Even though our hypothesis suspected that the building’s age would be a significant factor in predicting that building’s energy use, we decided to drop the feature due to the high number of null values. We also decided to drop the floor\_count feature, since it was similar to the square\_feet feature.

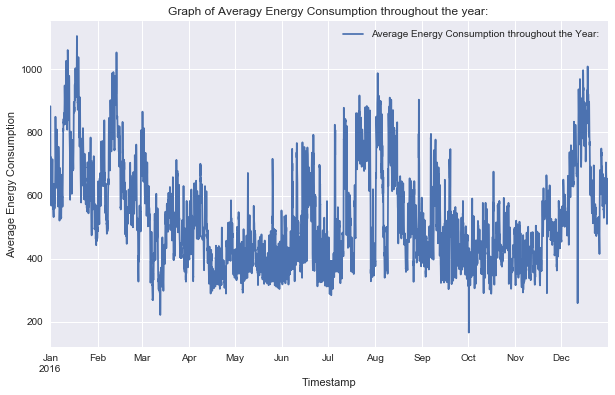
# VARIABLE SELECTION

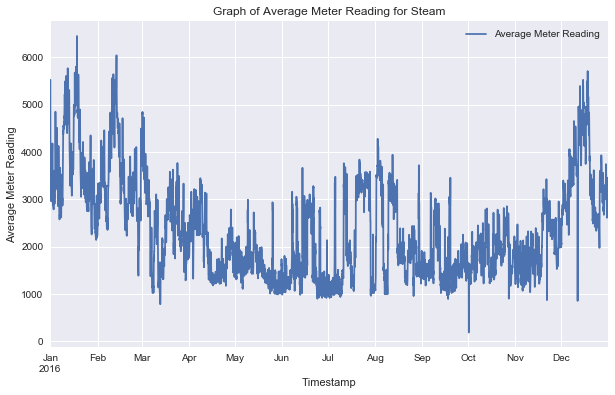
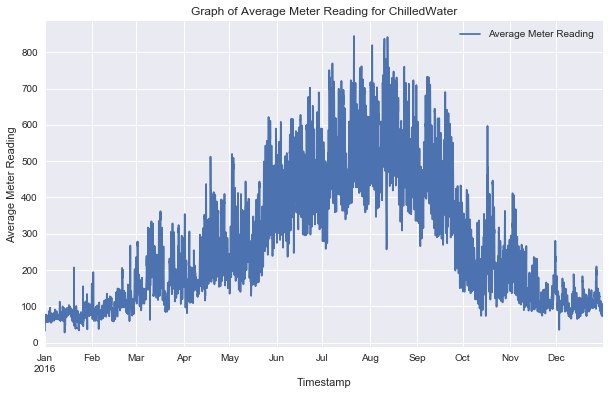
## EXAMINE ENERGY CONSUMPTION THROUGHOUT THE YEAR

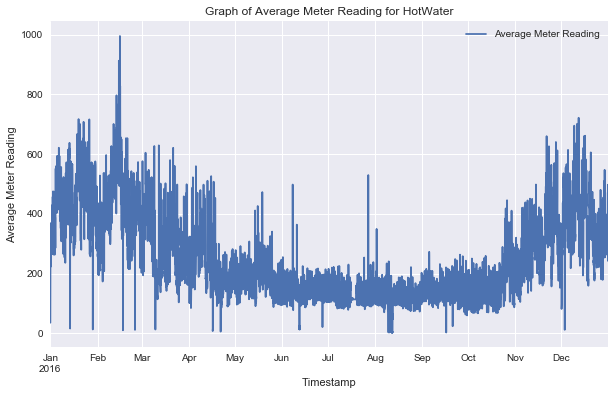
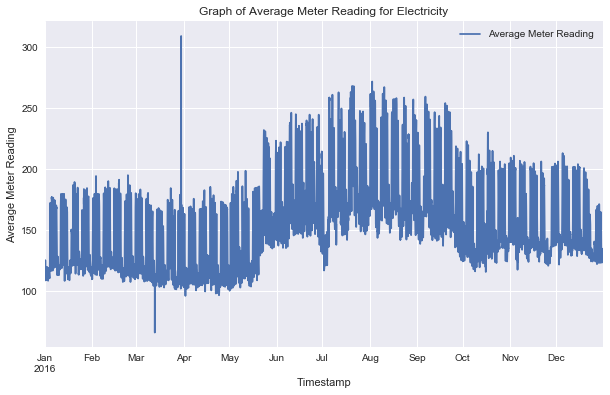
The first step for variable selection involved determining how different energy types behave throughout the year. This would be used to try and find any trends in the data and transform them if needed in order to obtain trends. The first step was to find out the frequency of different energy types using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

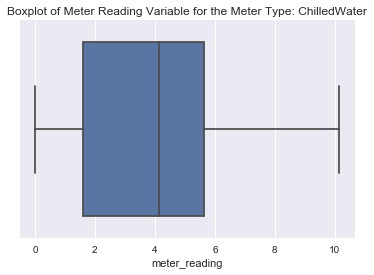
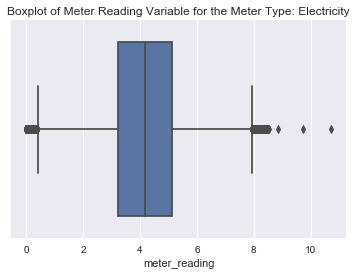


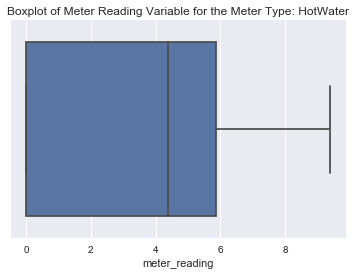
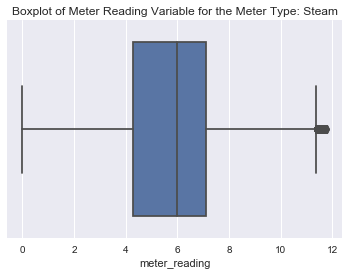




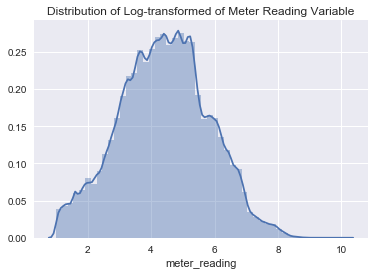
Based on the above graphs, we have determined that the Steam and HotWater follow a trend very similar to the overall trend, however Electricity and ChilledWater are different. However, we can see from the graphs that Steam has far greater values from the other energy types, and it sets the overall trend 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:



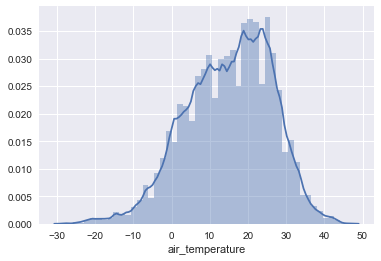
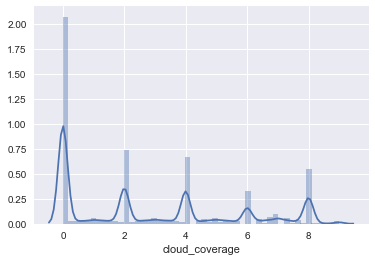
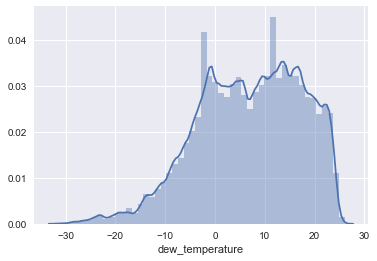
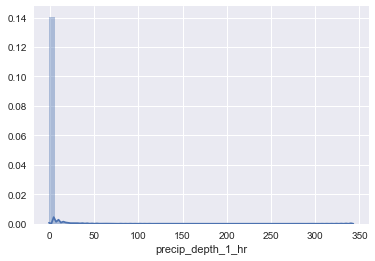
 

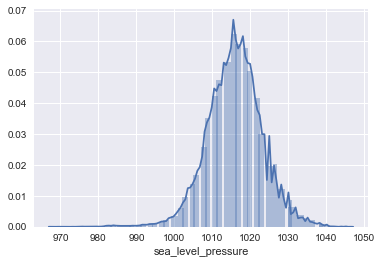
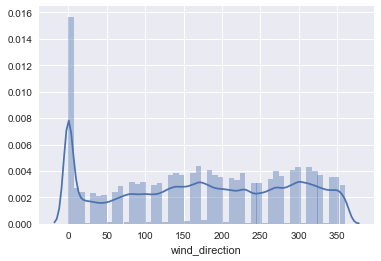
Finally, after these transformations, a near-normal distribution was obtained for the overall energy type consumption:



## EXAMINE WEATHER FEATURES

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

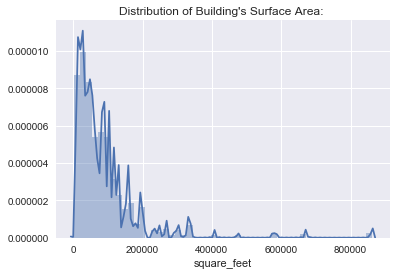
 

Based on the graphs, we can determine that the cloud coverage is different from others in that it is composed of a number of fixed values, and parcip\_depth\_1\_hr has a very large number of 0 values. To that end, these are not very useful for our calculations, and were removed from the data set. The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

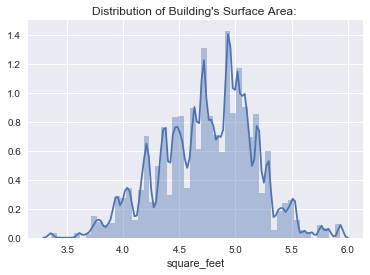
Next, the data frame corr() function was used to find the correlation between the remaining variables. The value of 0.9 was used as the threshold, and any features with the correlation higher than this were dropped as redundant. After applying the function, it was determined that the remaining variables 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:



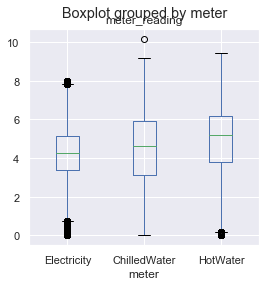
# ANALYSIS

## Multiple Linear Regression

Now that we have cleaned our data set and selected the appropriate variables for our data frame, we will make an attempt at fitting a linear model to the data. Our model equation will be as follows:

Y = b0 + b1X1 + b2X2 + b3X3

Where Y is the meter reading, X1 is the square footage of the building, X2 is the sea level pressure, and X3 is the meter category (Electricity, ChilledWater, HotWater). We have b0 as our y-intercept while the rest of bi are slope constants. First, we have to decide what to do about the meter categorical variable. We start by examining boxplots of each separate value for meter:



Our boxplots give us a clearer indication of the association between meter reading and meter type. We see that Electricity has the lowest average of the 3 types, potentially due to the number of extreme values close to (or slightly above) zero. It seems apparent that there is a visible order in the 3 types, this allows us to utilize polynomial encoding on the meter variable using our boxplot as reference. We will add two new columns to our data frame, mapping as per the table below:

|  |  |  |
| --- | --- | --- |
| **meter** | **meter\_linear** | **Meter\_square** |
| Electricity | 1 | 1 |
| ChilledWater | 2 | 4 |
| HotWater | 3 | 9 |

After fitting our model, we obtain the following results:

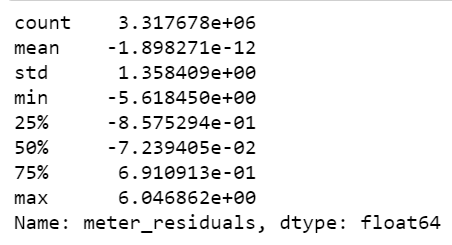
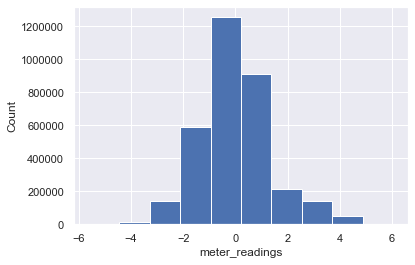
|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | meter\_reading | **R-squared:** | 0.244 |
| **Model:** | OLS | **Adj. R-squared:** | 0.244 |
| **Method:** | Least Squares | **F-statistic:** | 2.673e+05 |
| **Date:** | Thu, 05 Dec 2019 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 11:28:59 | **Log-Likelihood:** | -5.7238e+06 |
| **No. Observations:** | 3317678 | **AIC:** | 1.145e+07 |
| **Df Residuals:** | 3317673 | **BIC:** | 1.145e+07 |
| **Df Model:** | 4 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | -4.3814 | 0.105 | -41.648 | 0.000 | -4.588 | -4.175 |
| **square\_feet** | 1.7575 | 0.002 | 1018.356 | 0.000 | 1.754 | 1.761 |
| **sea\_level\_pressure** | 0.0009 | 0.000 | 8.405 | 0.000 | 0.001 | 0.001 |
| **meter\_linear** | -0.8169 | 0.008 | -107.519 | 0.000 | -0.832 | -0.802 |
| **meter\_square** | 0.2709 | 0.002 | 128.608 | 0.000 | 0.267 | 0.275 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 187360.554 | **Durbin-Watson:** | 1.596 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 257531.267 |
| **Skew:** | -0.523 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 3.877 | **Cond. No.** | 1.43e+05 |

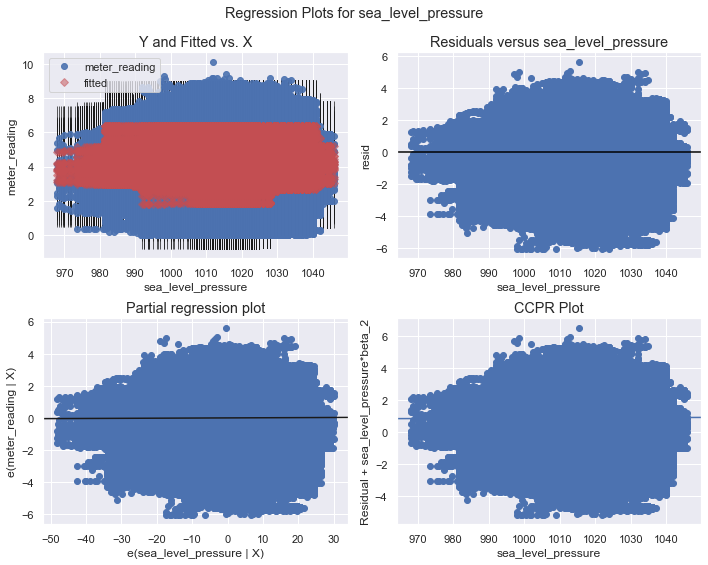
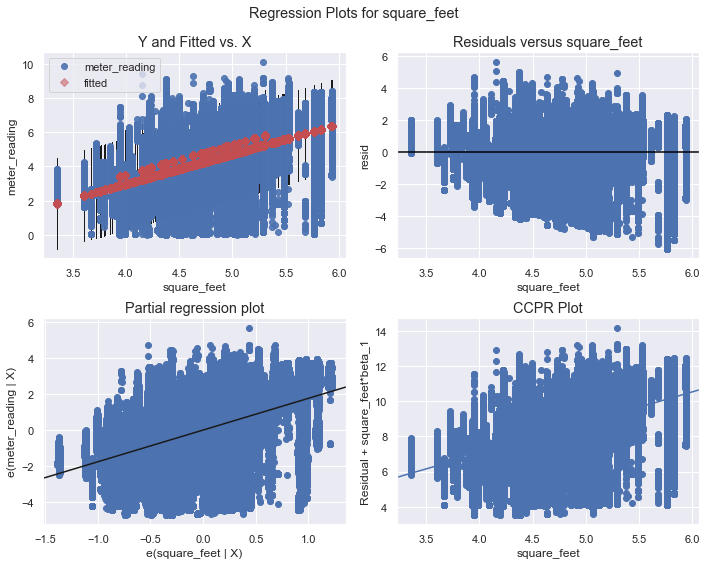
Our obtained linear equation is Y = 1.7575\*X1 + 0.0009\*X2 -0.8169\*X3 + 0.2709\*X4 - 4.3814. The R2 value is 0.244 meaning that the model can only explain approximately 24.4% of the variability in our meter readings. The p-values for both the F-statistic and the t-statistics are all equal to zero, which tells us that there is definitely a linear association between our response and predictor variables.

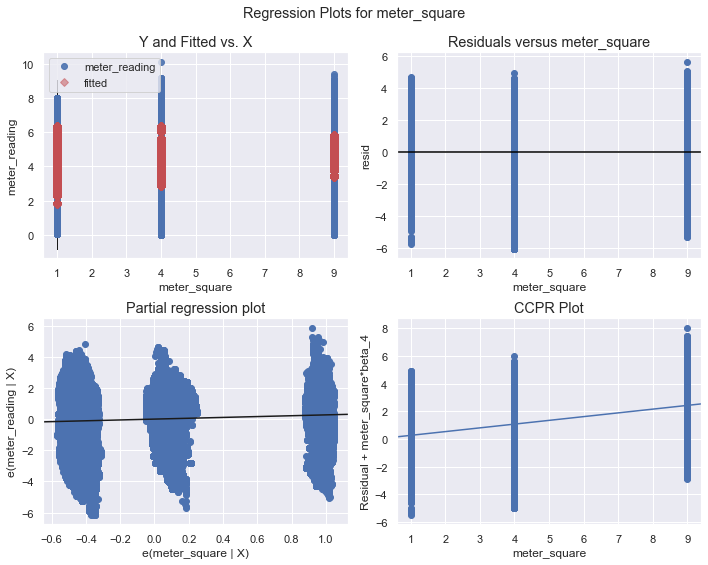
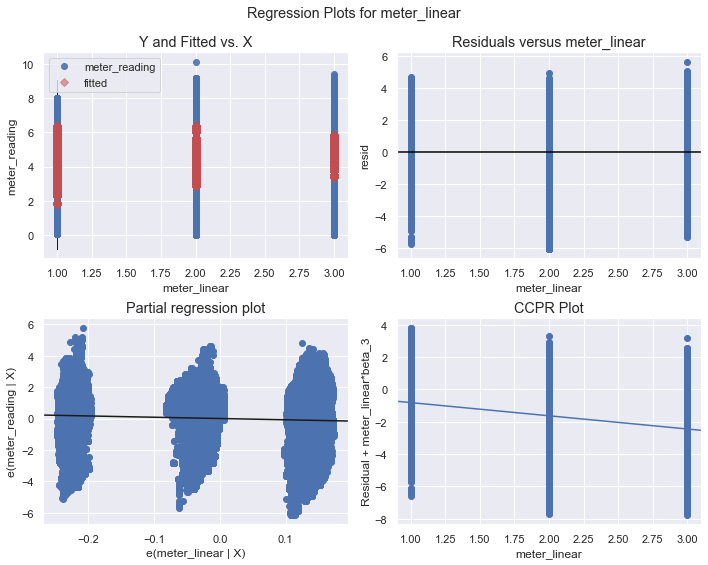
For comparison, we fit a model without the meter variables involved. The results showed that our model weakens (albeit slightly), since the R2 drops, but the p-values remain the same across the board. Suggesting there is a strong association between the meter reading and all of our variables. We will evaluate the residuals using our full model.



The residuals do resemble a normal distribution with a mean and median very close to zero. The errors range from -6 to 6 (based on the maximum and minimum values), although for the most part we can clearly see that the vast majority of errors range from +/- 1. Overall, the residuals paint a much nicer picture of the model then the R2 value in our summary suggests. We can obtain a pretty accurate prediction with this fit the majority of the time, although the goal is obviously to get as close to zero as possible. The obtained RMSE is approximately 1.36, which means that on average our predictions are off by 1.36, which is approximately 3.13% of the average meter reading in our dataset.

Next, we evaluate the regression plots for each of the predictors:





Our regression plots all seem to follow a similar pattern. We are hoping to see constant variability of the residuals throughout, but it appears that we observe an increase in variability as X increases. In other words, if X is small, then it is more likely we are going to have a more accurate prediction for our response. Another consequence of fitting a linear model is that our predictions will always stay close to our average, which is good for keeping a smaller average error, but will never come close to accurately predicting extremely high or low meter readings. Our goal is to find a model that isn't heavily influenced by the size of our predictor values and it may be best to pursue another strategy for this particular set of data.

# 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
* 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