**S670 Exploratory Data Analysis Final Project Report** **ASHRAE ENERGY CONSUMPTION**

**Group Name**: **Team Vietnam**

**Team members: Amit Kumar Yadav, Mahesh Latnekar, Prashil Negandhi**

**1. Project Motivation:**

Throughout history, buildings have been evolving to address social need. A century ago, the advent of skyscrapers exploited the steel framing technology to overcome the scarcity of real estate in teeming American cities and extended it into a third dimension i.e. vertically. Today’s building industry is entering into a different era of change: minimizing the overall energy consumption of the building and decreasing its carbon footprint. This change to minimize the environmental impact of a building is driven by a social need to conserve and optimize crucial non-renewable resources like coal, petroleum and natural gas. Designers and developers alike are searching for ways to decrease the operating costs of the building by using energy efficient designs and technologies while also increasing the functionality and appeal of their buildings. These “green” trends are very much evident in many of the modern buildings but are lacking in many of the older buildings. This is because these older buildings were constructed when conserving energy wasn’t the top priority of the society.

ASHRAE, founded in 1894, is a global society that advances human well-being through sustainable development. It focuses on promoting building systems that are not only energy efficient but also have the highest standards for functionality. Due to its continuous efforts over the past century, significant investments have been made to improve building efficiencies that reduce costs and emissions. However, one challenge that ASHRAE faces is quantitatively figure out if the improvements being made are working or not. That’s where the Kaggle challenge that we tackled comes in. In this project we came up with simple, interpretable models for metered building energy usage for electricity consumption by combining features of the buildings and their associated weather data. This data comes from over 1000 buildings and had been gathered over a three-year timeframe. ASHRAE hopes to use these models to implement their pay-for-performance financing where the building owner makes payments based on the difference between their real energy consumption and what they would have used without any retrofits.

**2. Relevance:**

Hospitals, offices, industries, etc. are constantly trying to reduce their operating costs as much as possible by eliminating wasteful consumption of resources. This analysis would benefit the building sector by providing an estimation of the expected operating cost.

**2.1 Scientific Motivation:**

It is interesting for the researchers to see how information about the buildings and the associated weather data affects the electricity consumption of a building. And also how much each feature contributes to the overall energy consumption. This analysis creates the source the perform further research and figure out improvements that can be made in each building.

**2.2 Societal Contribution**

The societal contributions of this analysis are manifold. Firstly, this project hopes to quantify the expected electricity consumption of a building by providing a set of metrics. These metrics have aroused after computing tests on each of them to see how much they contribute to predict the electricity consumption of a building. Secondly, it would help ASHRAE by giving them an idea about which buildings and features to focus on to get the maximum benefit. This knowledge also helps the building owners and their users to know how much energy they should be ideally using and if they are using more then they would know which areas to focus on to bring it into control.

**3. Research Questionnaire:**

The main question we hope to answer is: **“How much energy should a building consume?”**. To answer this, we followed a drill down approach by coming up with the following research questions that we aimed to answer using exploratory data analysis methods and simple model fitting on our ASHRAE Energy dataset:

1. **What building features are most likely to predict the electricity energy consumption of a building?**
2. **How much does the weather surrounding a building change the electricity consumption of buildings?**
3. **How well can we predict the electricity energy consumption by factoring in the time (month, day) into our model?**
4. **Finally, how well does all the features combined predict a buildings electricity consumption?**

**4. Dataset Description:**

**Dataset Source: Kaggle**

**(https://www.kaggle.com/c/ashrae-energy-prediction)**

The dataset includes three years of hourly meter readings from over one thousand buildings at several different sites around the world. It is divided into three files as follows:

**Train Data:**

* **building\_id** - Foreign key for the building metadata
* **meter** - 0: electricity, 1: chilled water, 2: steam, 3: hotwater
* **timestamp** - the time of measurement
* **meter\_reading** - Energy consumption in kWh (or equivalent)

**Building Metadata:**

* **site\_id** - Foreign key for the weather files
* **building\_id** - Foreign key for training.csv
* **primary\_use** - Indicator of the primary use of 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 Data:** Weather data from a meteorological station as close as possible to the site.

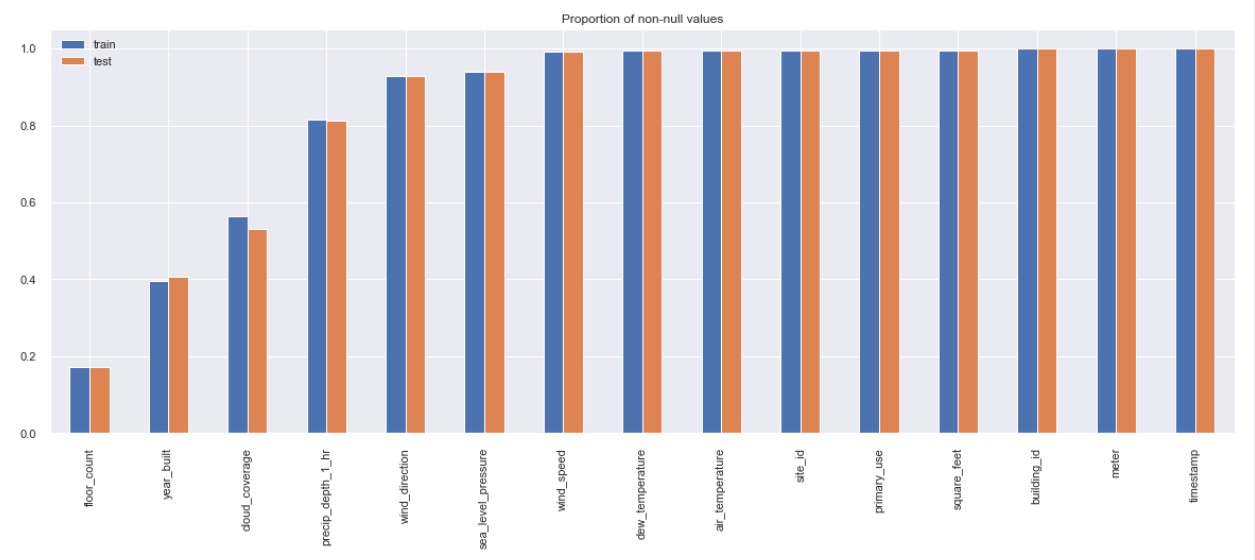
* **site\_id** - Key for reference
* **air\_temperature** - Degrees Celsius
* **cloud\_coverage** - Portion of the sky covered in clouds
* **dew\_temperature** - Degrees Celsius
* **precip\_depth\_1\_hr** - Millimeters
* **sea\_level\_pressure** - Millibar/hectopascals
* **wind\_direction** - Compass direction (0-360)
* **wind\_speed** - Meters per second

**5. Exploratory Data Analysis:**

**5.1 Null Values**

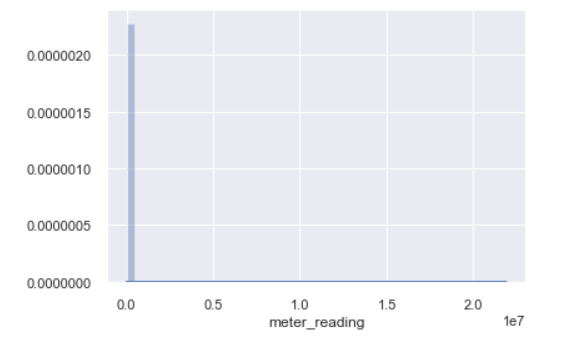
Since this dataset is a real-world dataset, we can expect to have a lot of null values. We found the proportion of non-null values for each of the feature as shown. We find that the feature floor-count has the maximum number of missing values. However, upon further analysis we can see that this feature is not important. This is because we already have the key distinguishing feature of a building that is encoded in the square feet variable which measures the area of the building. Also, one can say that the number of floors would be highly correlated to the area of a building. Also, it would hard to infer the floor-count variable based on other buildings floors. Thus, we decided the drop this variable in its entirety.

We also decided to drop the year built and cloud coverage variables as there were more than half the values missing. Moreover, it wouldn’t have made sense to infer these variables from other observations too. For the other null values, we used the technique of forwardfill

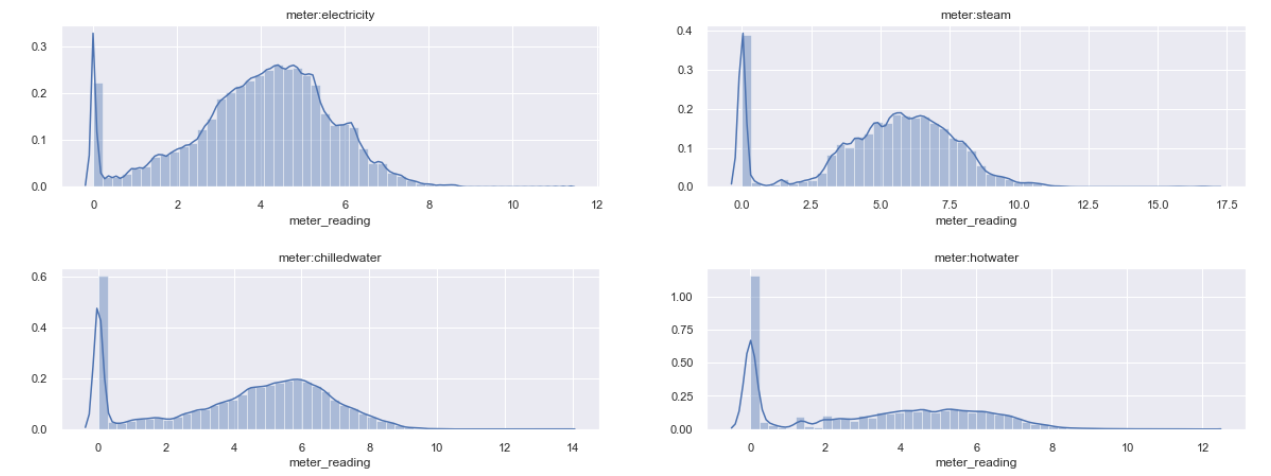


**5.2 Target Variable Distribution:**

We first tried to plot the meter readings without any transformations and couldn’t visualize it as shown below. This is because we have some really high values in it.



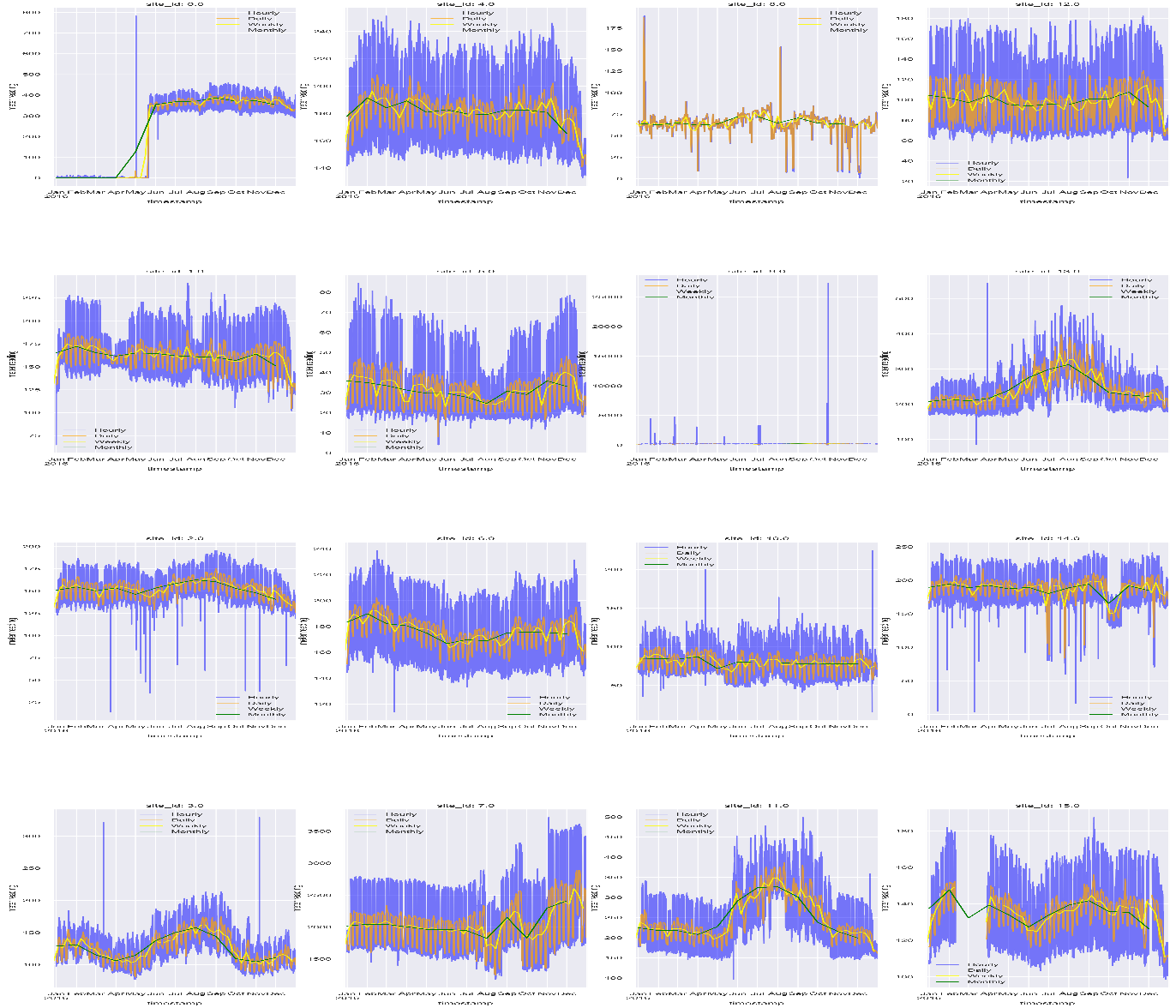
After taking the log transformation of the meter readings we get the following:



As we can see it would be better suited to use log transformations in our modelling. This would help reduce the bias that extremely large values for meter readings would introduce. Let’s see how electricity is distributed over time to see if the time variable would help in making some predictions:



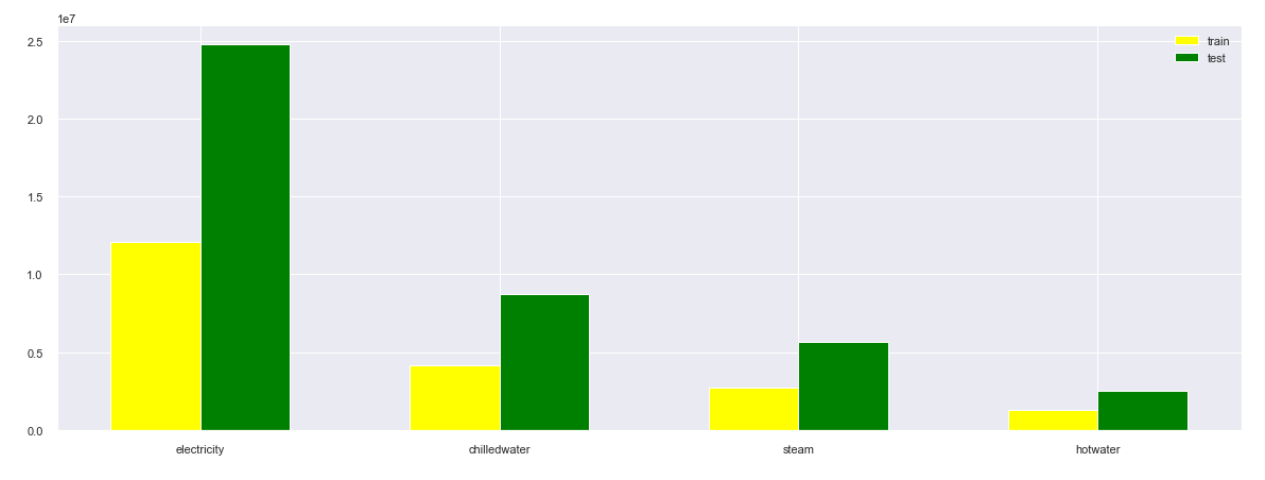
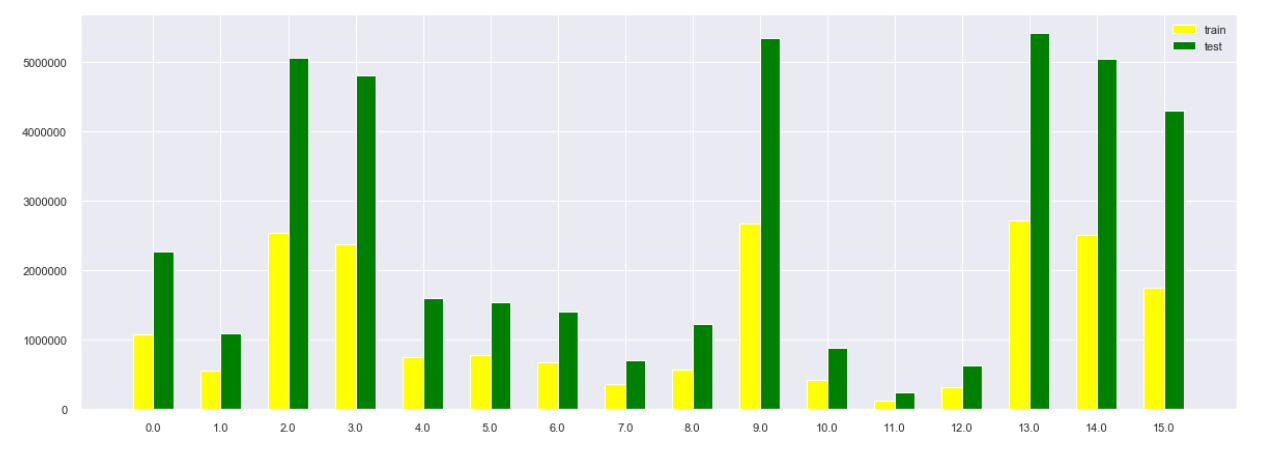
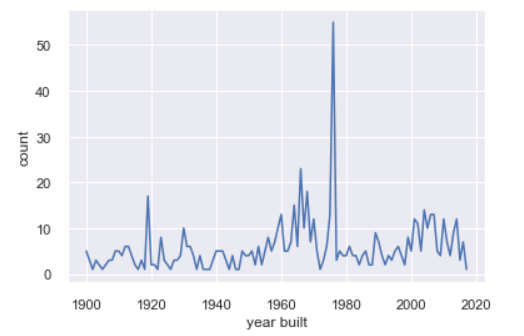
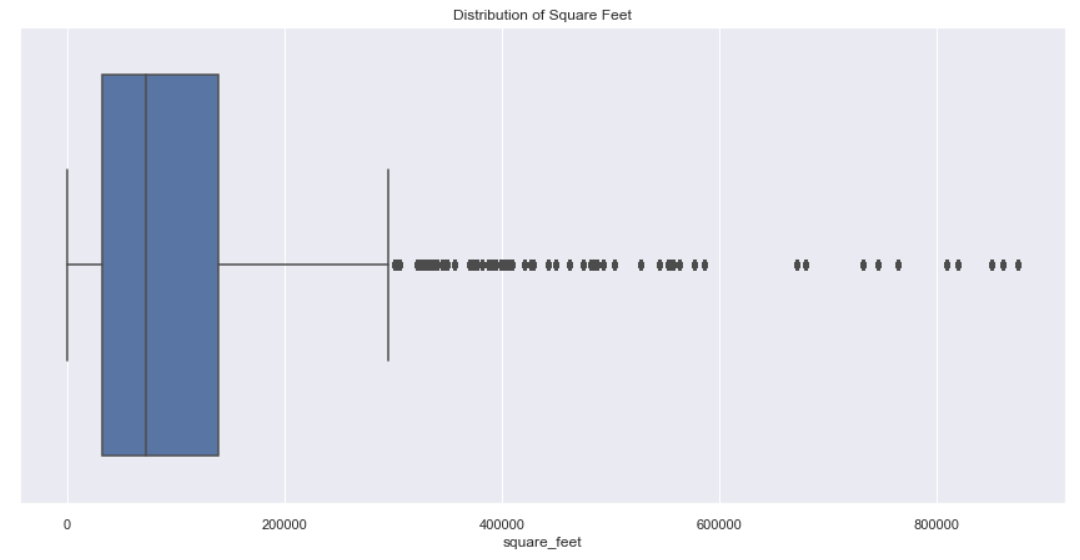
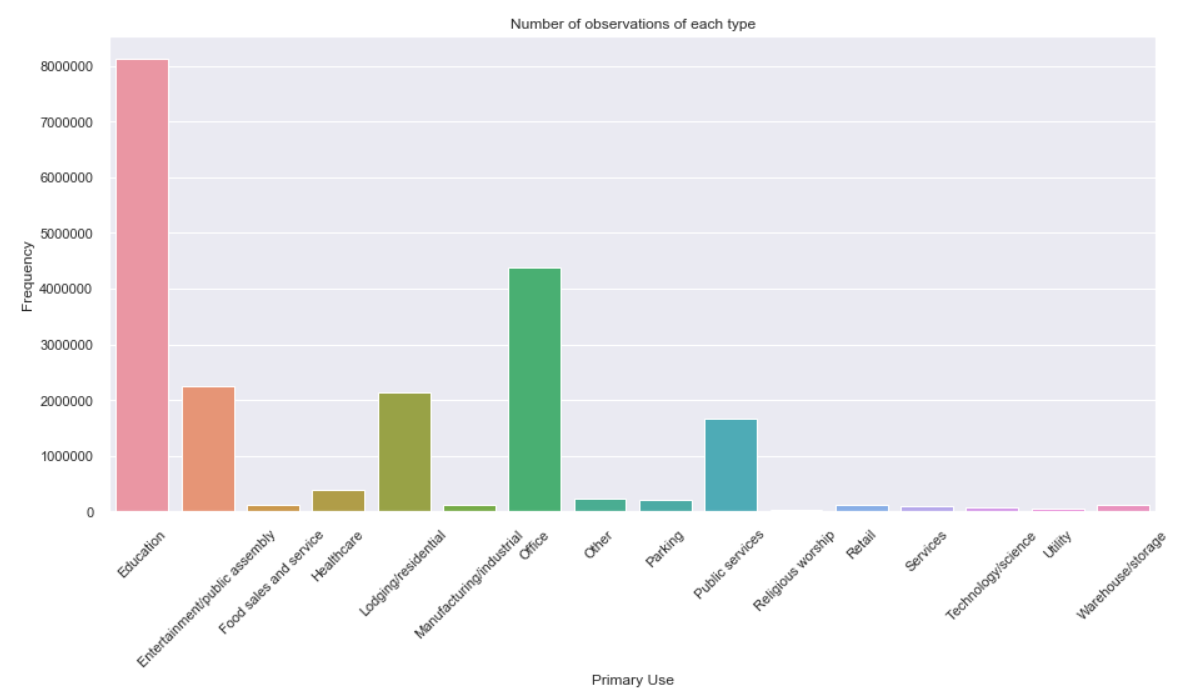
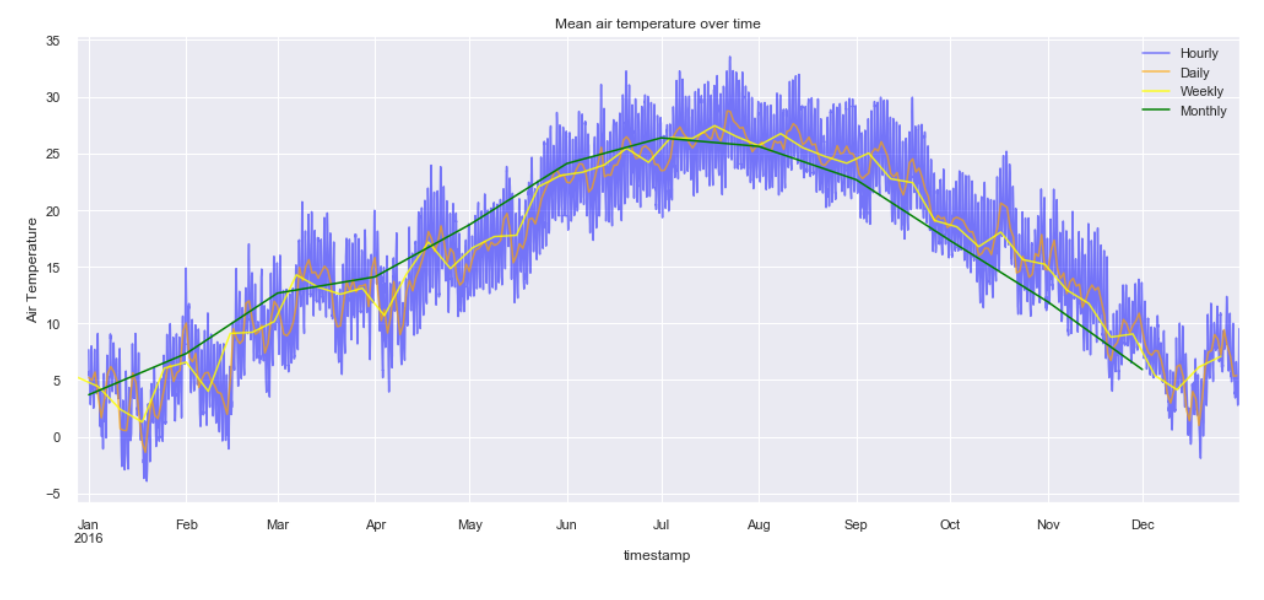
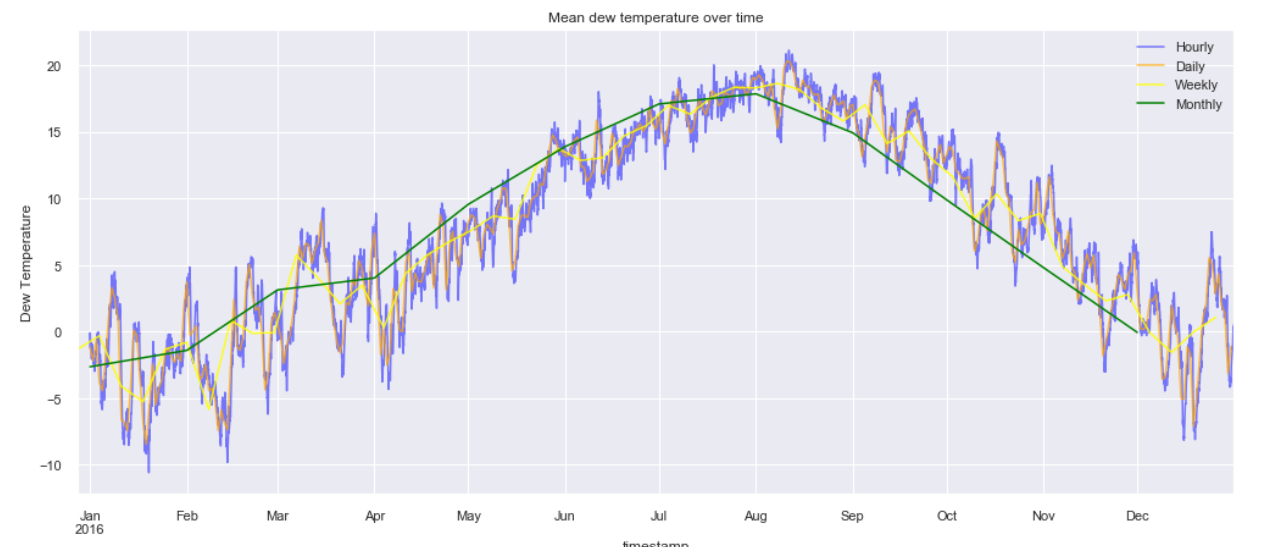
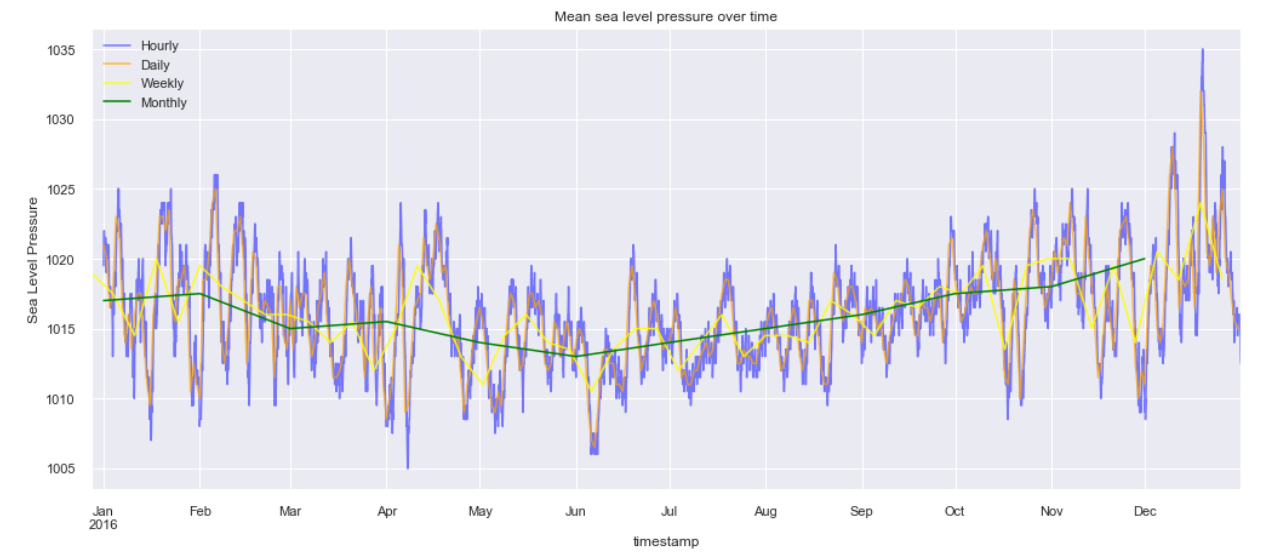
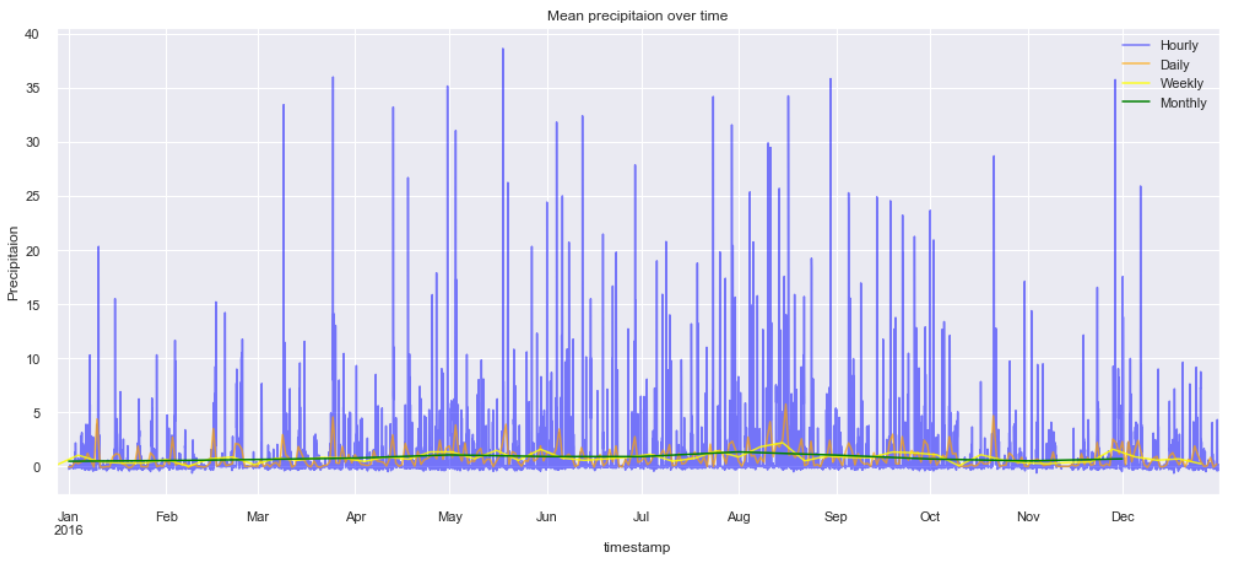
Faceted over each site id:



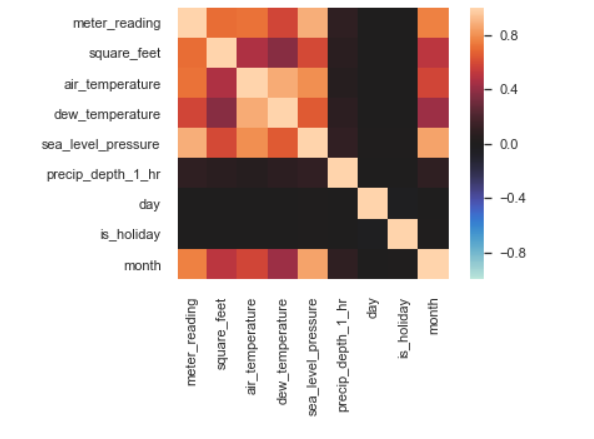
We can see that by plotting the time series we can see some weak correlation of electricity usage with respect to time. However, it’s not very strong. When we go about evaluating different models, we would be better able to understand how much time is contributing to predicting the electricity usage.

**5.3 Explanatory Variables:**

We plotted the distributions of each of the explanatory variables to see how they are changing and see which ones might be a good feature for our model. We have also added three new variables to our data. These variables are: day, month and holiday. Day and month encode the time of the meter readings. Holiday encodes whether that date was a public holiday or not.

* **Meter Readings:** 
* **Site ID:** 
* **Year Built:** 
* **Square Feet:** 
* **Primary Use:** 
* **Air Temperature:** 
* **Dew Temperature:**
* **Sea Level Pressure:** 
* **Precipitation:** 

As we can see from the distributions of dew temperature and air temperature, they follow almost the same distribution and appear to be highly correlated. Also, it makes sense since their temperatures would be almost the same. Let’s find the covariance matrix for these variables and see if our intuition is correct. The covariance matrix is as follows:



As we can see our intuition is correct. Thus, we decided to drop the dew temperature variable in its entirety for simplicities sake.

**6. Data Modelling:**

For modelling our data, we have use the Ordinary Least Squares (OLS) commonly called Linear Regression. Linear regression, as the name suggest follows a linear approach to model the relationship between the scalar response and the explanatory variables. When we consider just one explanatory variable it is called simple linear regression. And when we consider more than one explanatory variable, it is called multiple linear regression.

Linear regression is one of the oldest models and it has been studied rigorously. It is thus used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters. Also, the statistical properties of the resulting estimators are easier to determine.

There are many benefits and drawbacks of using this model. They are summarised as follows:

**Advantages:**

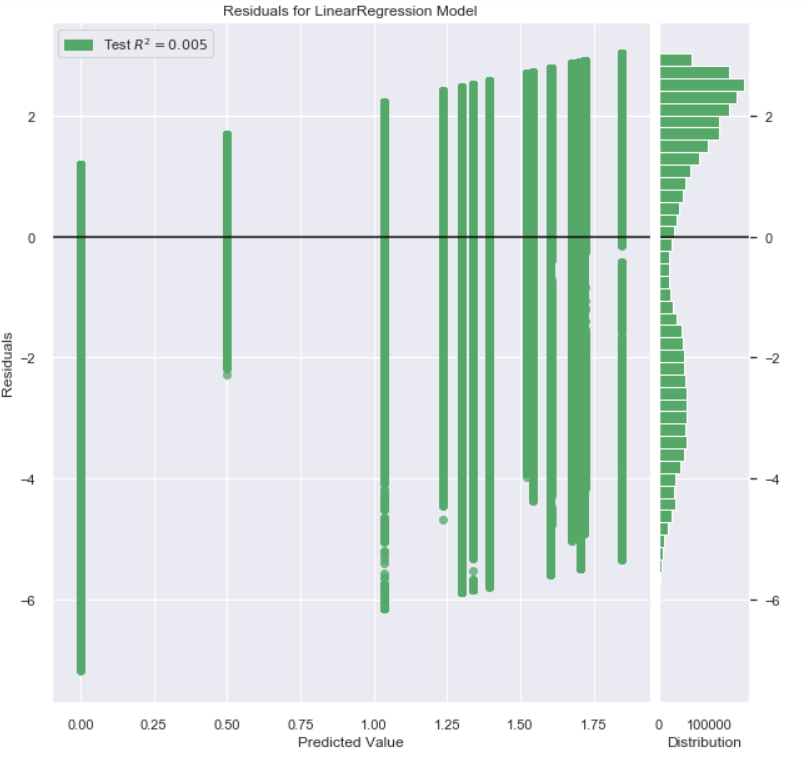
1. Space complexity is very low it just needs to save the weights at the end of training. hence, it's a high latency algorithm.
2. It’s very simple to understand.
3. Good interpretability
4. Feature importance is generated at the time model building. With the help of hyperparameter lambda, you can handle features selection hence we can achieve dimensionality reduction.

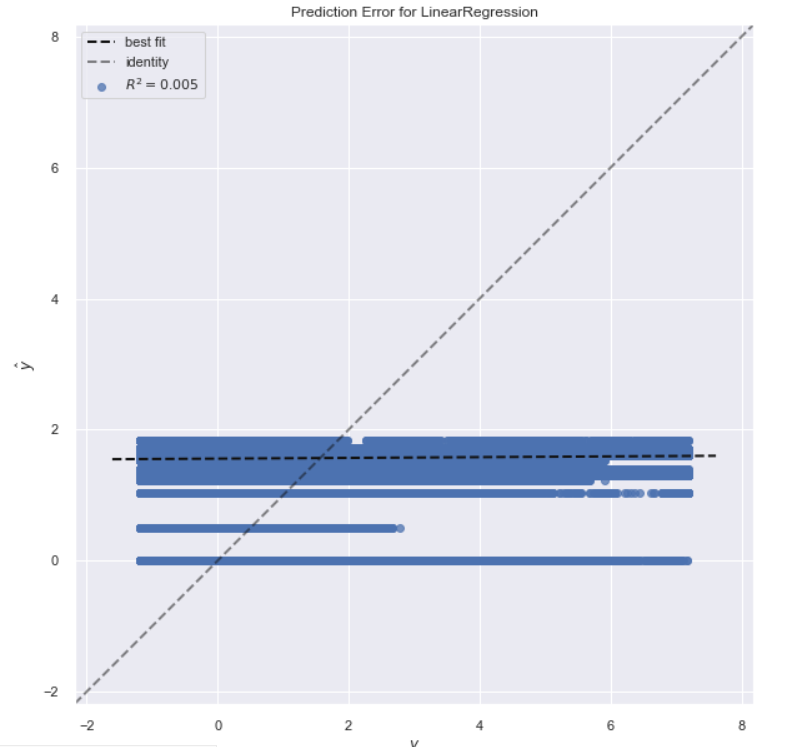
**Disadvantages:**

1. The algorithm assumes data is normally distributed in real they are not.
2. Before building model, multicollinearity should be avoided.
3. It is prone to outliers.

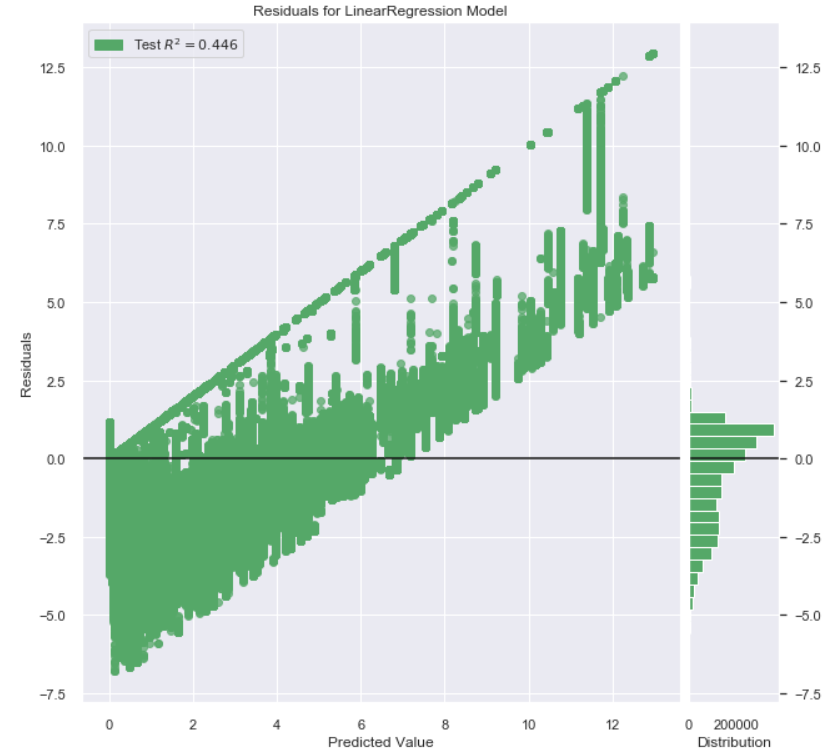
**6.1 What building features are most likely to predict the electricity energy consumption of a building?**

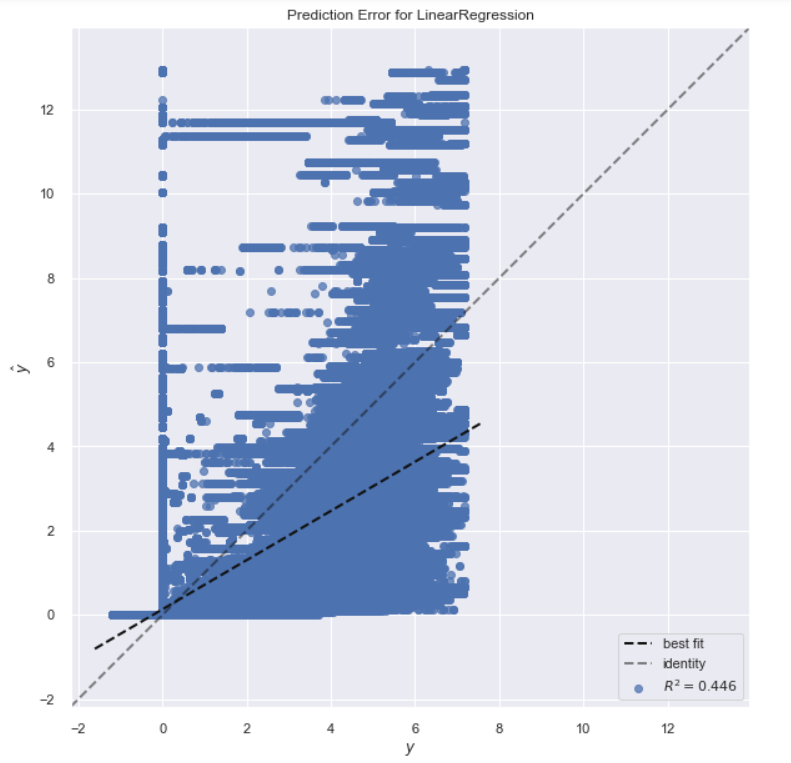
* **Meter reading from primary use of the building:**



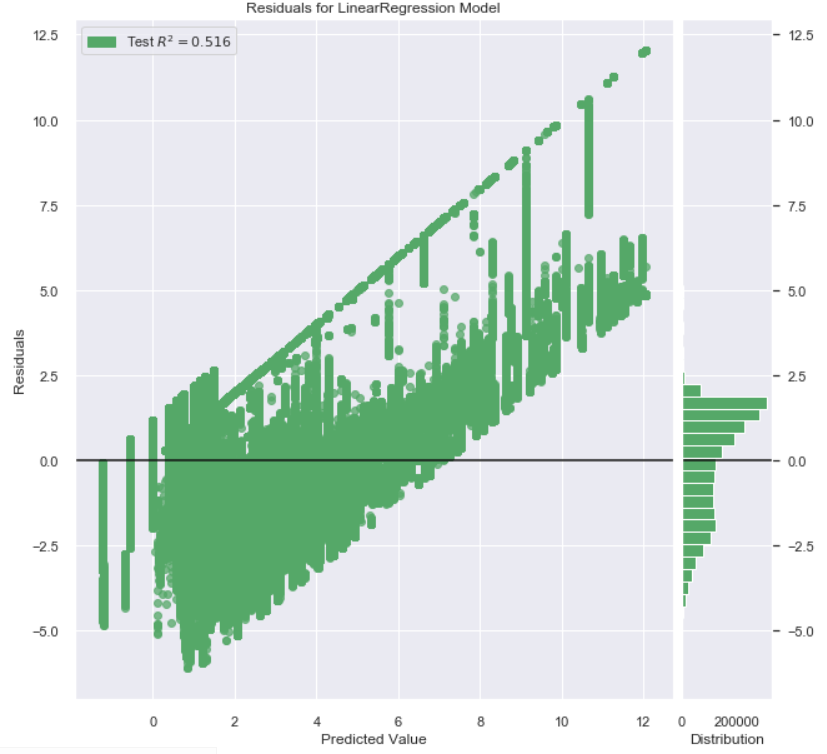


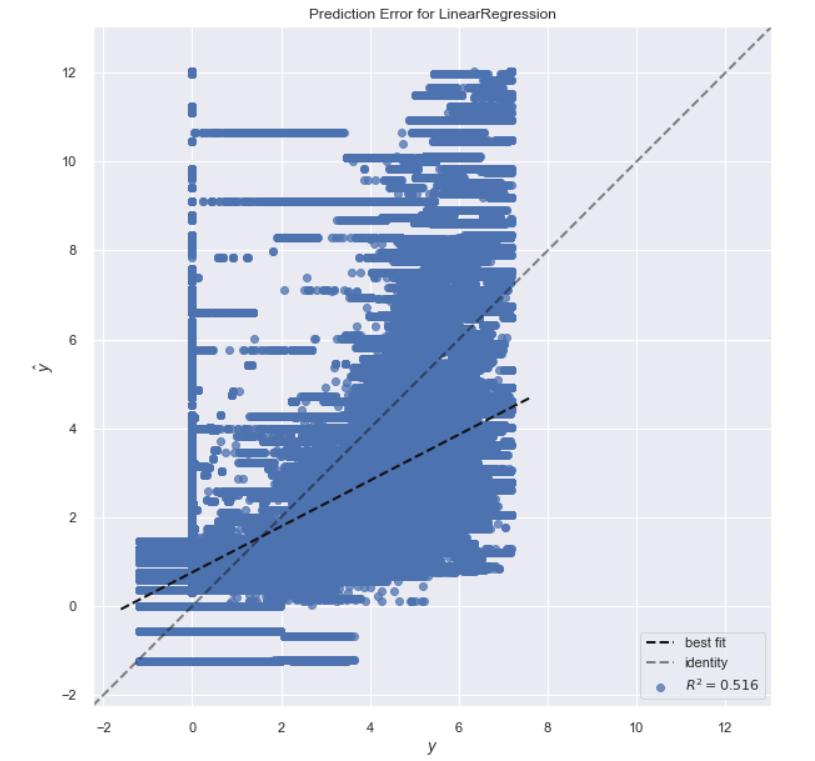
* **Meter reading from square feet of the building:**





* **Meter reading using both primary use and square feet of the building:**



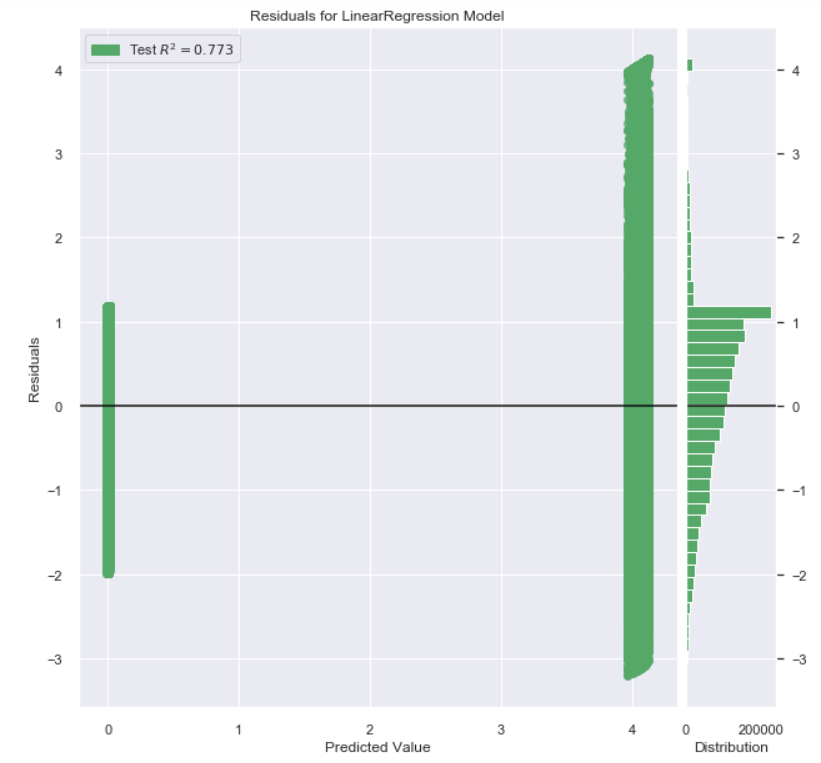


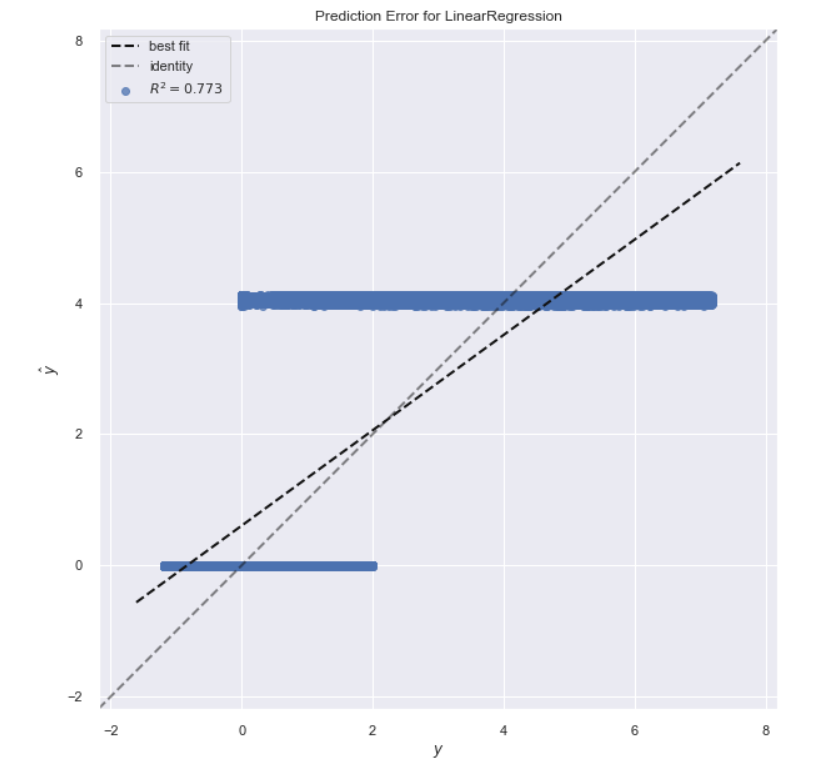
When we use both the square feet and primary use features, the R2 value doesn’t deteriorate very badly from when we used just the square feet variable. We also think that the primary use of a building is an important feature. However, we get a glimpse of the importance of the square feet variable. At the end it boils down to the fact that bigger buildings would obviously use higher amounts of electricity.

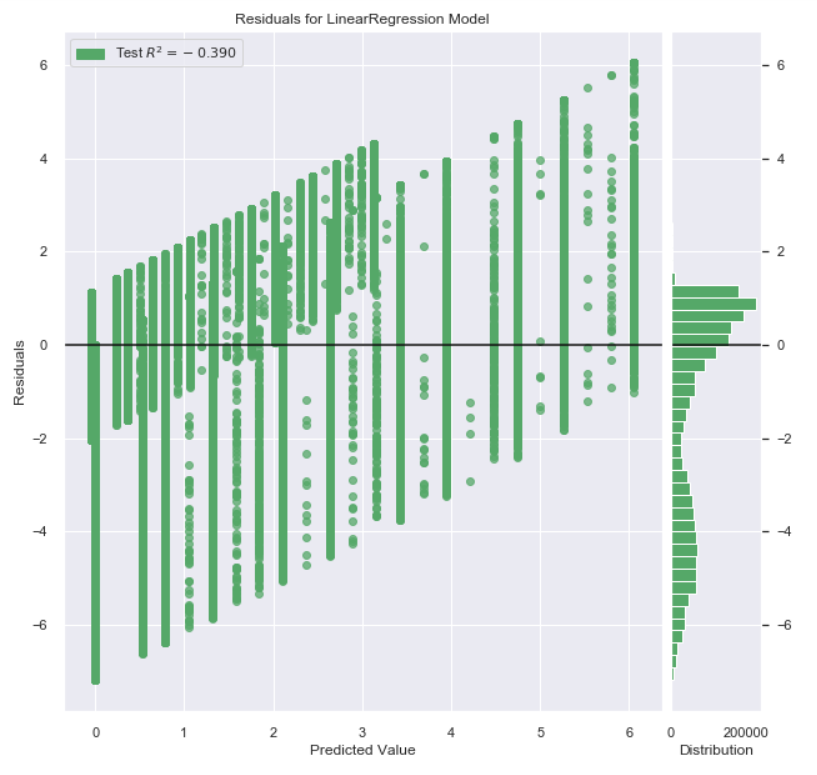
**6.2 How much does the weather surrounding a building change the electricity consumption of buildings?**

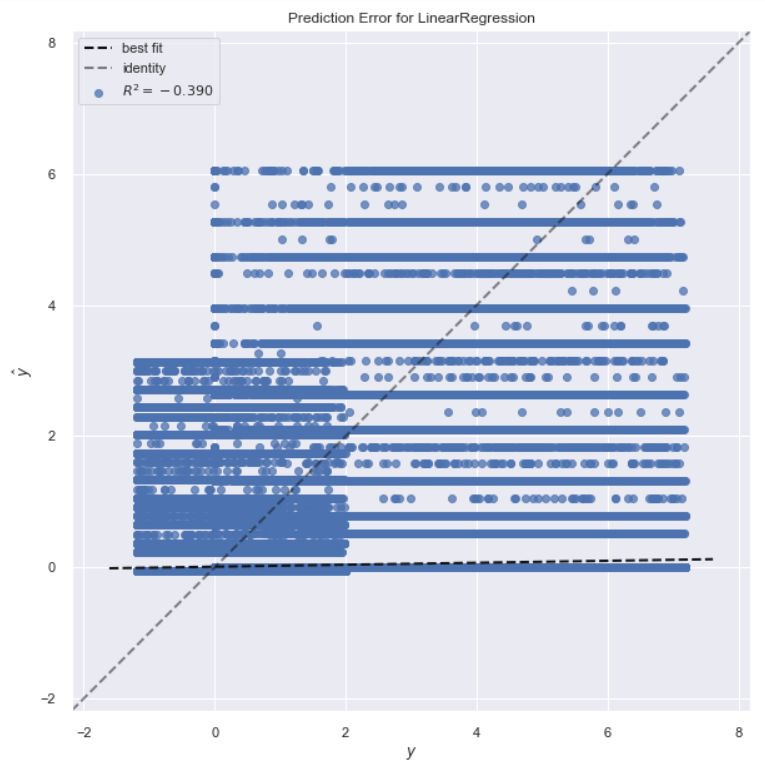
As we can see from the graphs and the R2 values associated with each model, we can say that the weather data contributes a lot to predicting the electricity energy consumption of a building.

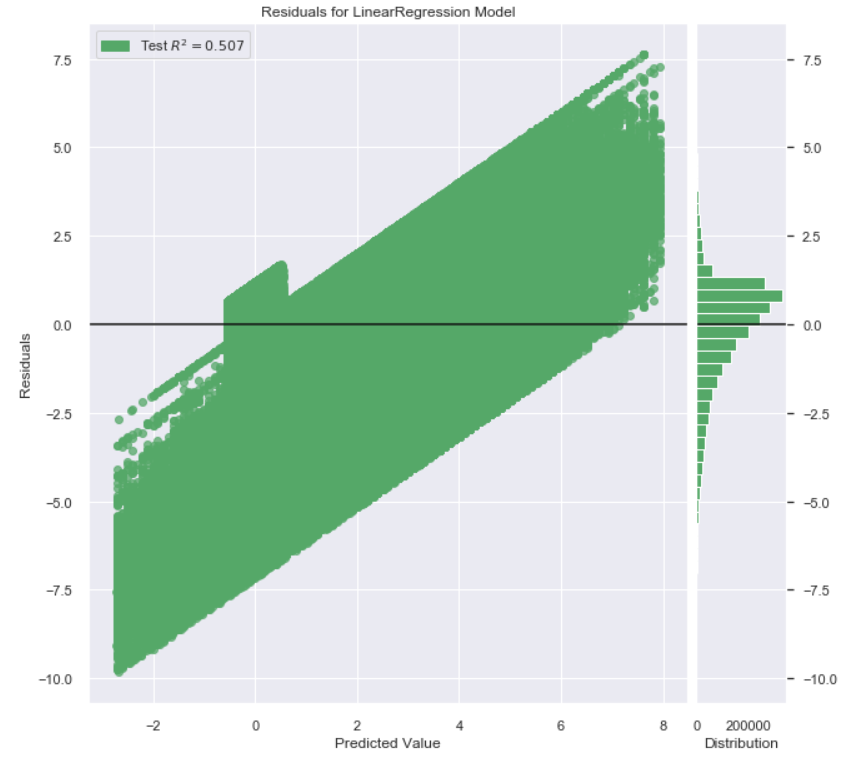
It is also evident that air temperature is a very strong predictor of the energy consumption of a building. However, we felt that the other weather variables are also contributing enough and thus we kept all the weather variables in our model.

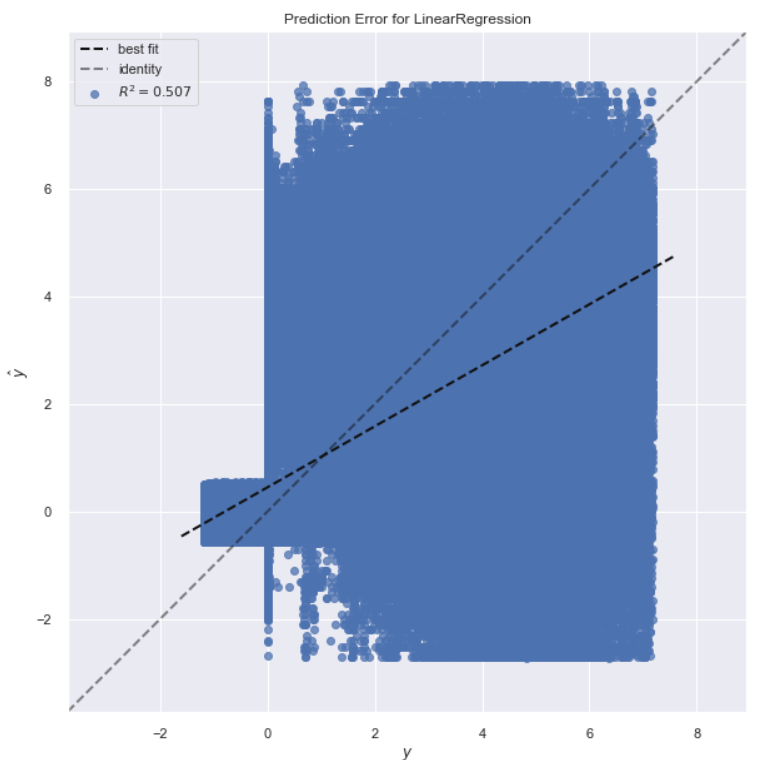
* **Meter reading using sea level pressure:** 



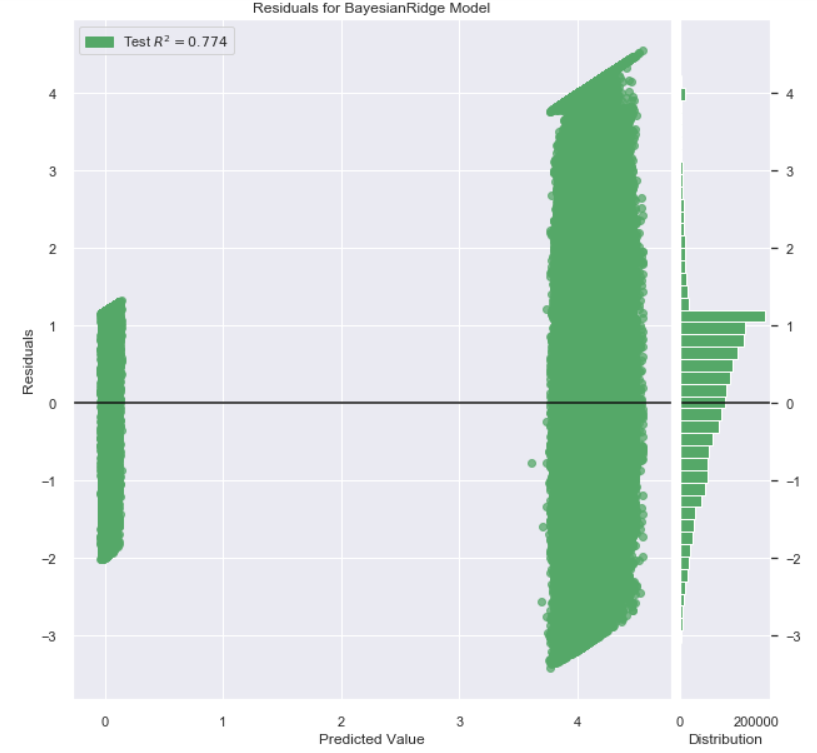
* **Meter reading using precip\_depth\_1\_hr:** 

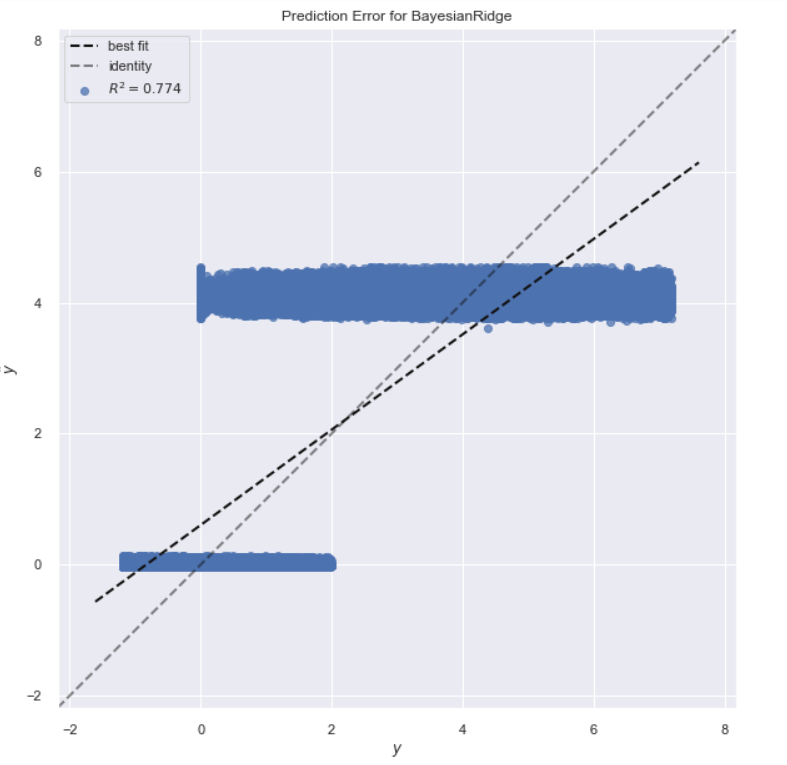


* **Meter reading using air temperature:** 

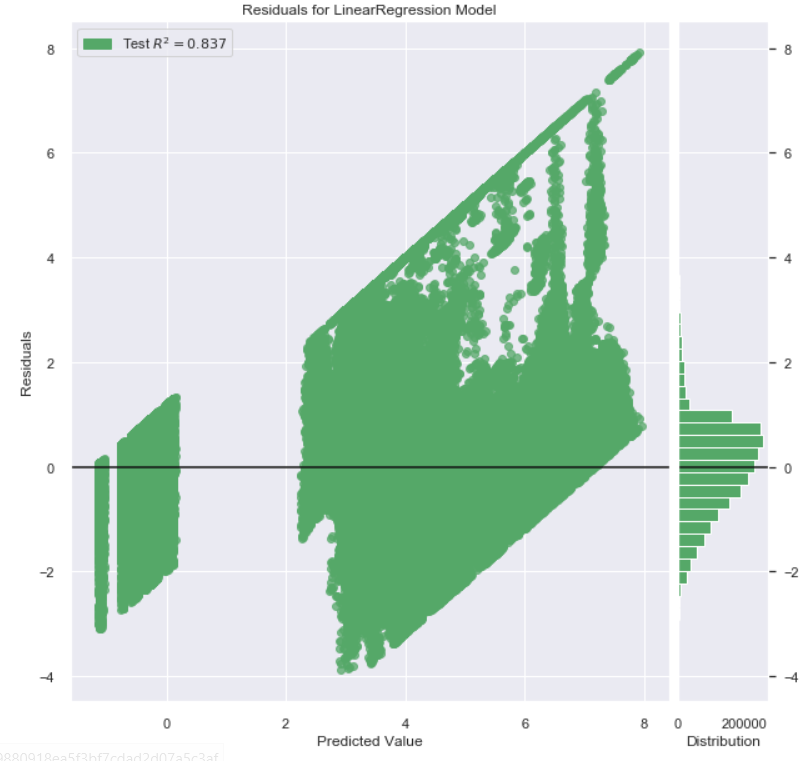


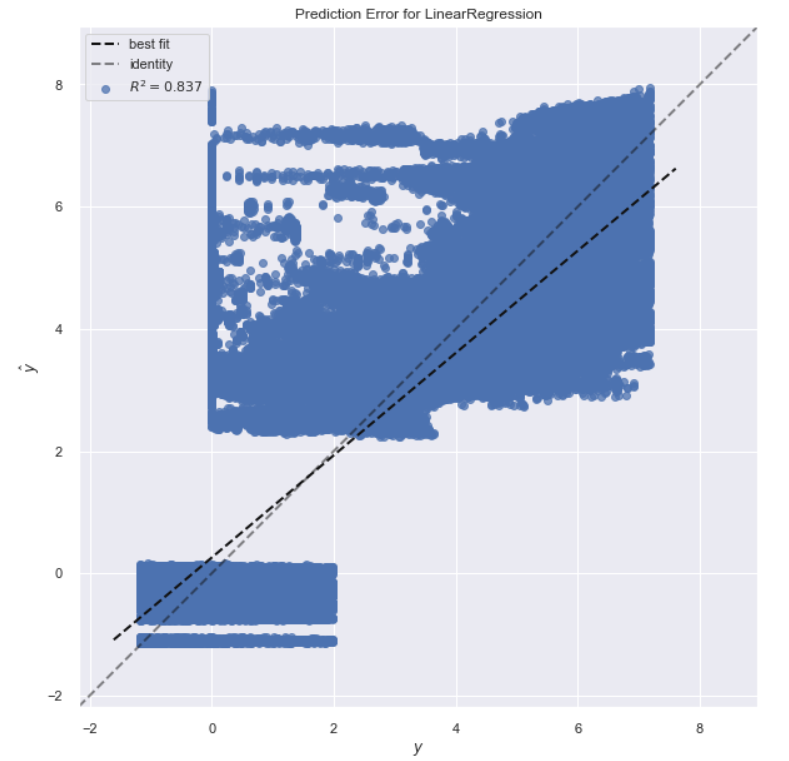
* **Meter reading using all weather variables (without interaction):**





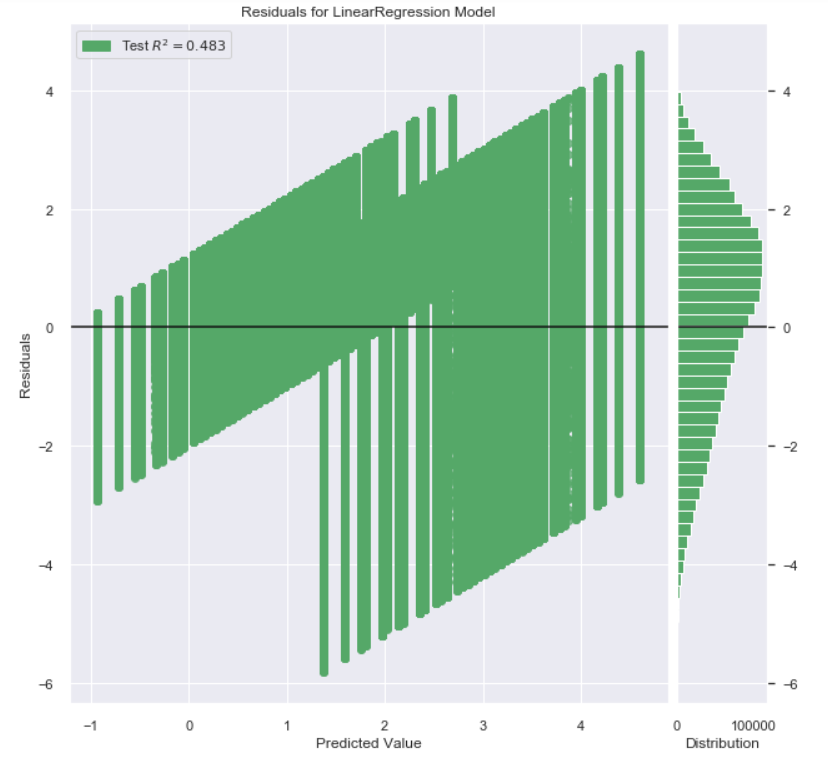
* **Meter reading using all the weather variables and all the building variables (without interaction):**

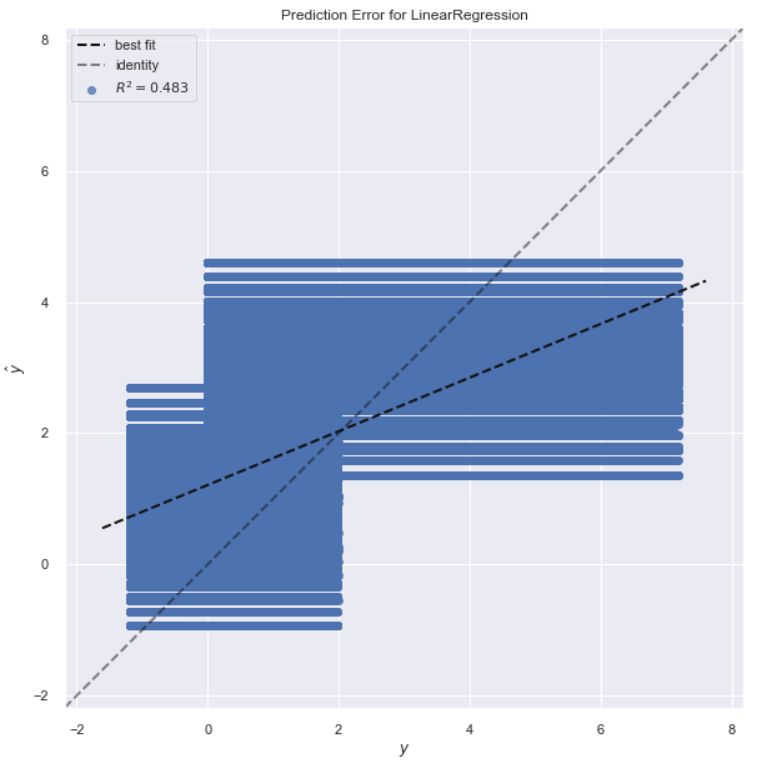




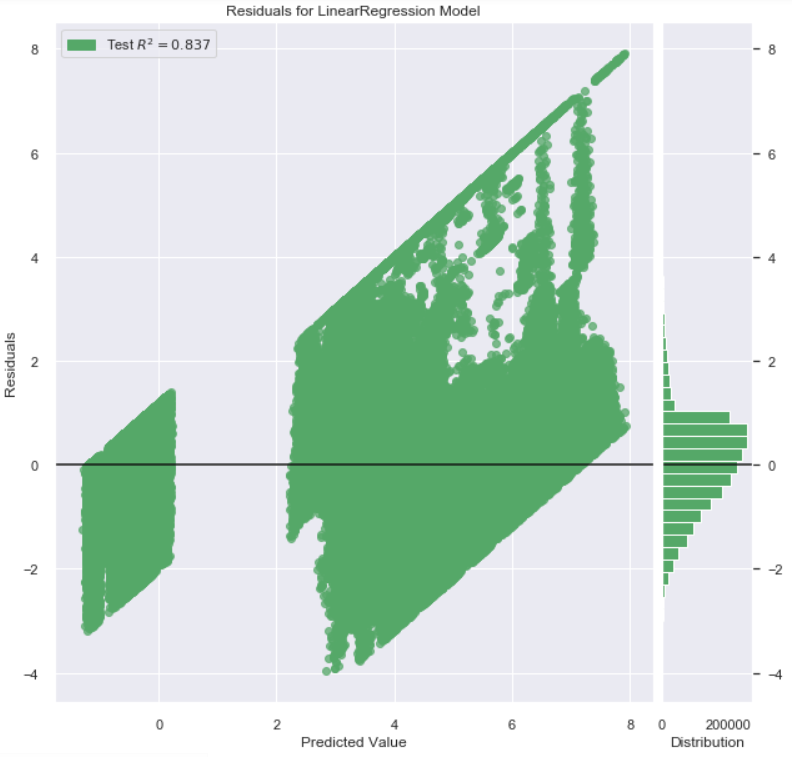
**6.3 How well can we predict the electricity energy consumption by factoring in the time (month, day) into our model?**

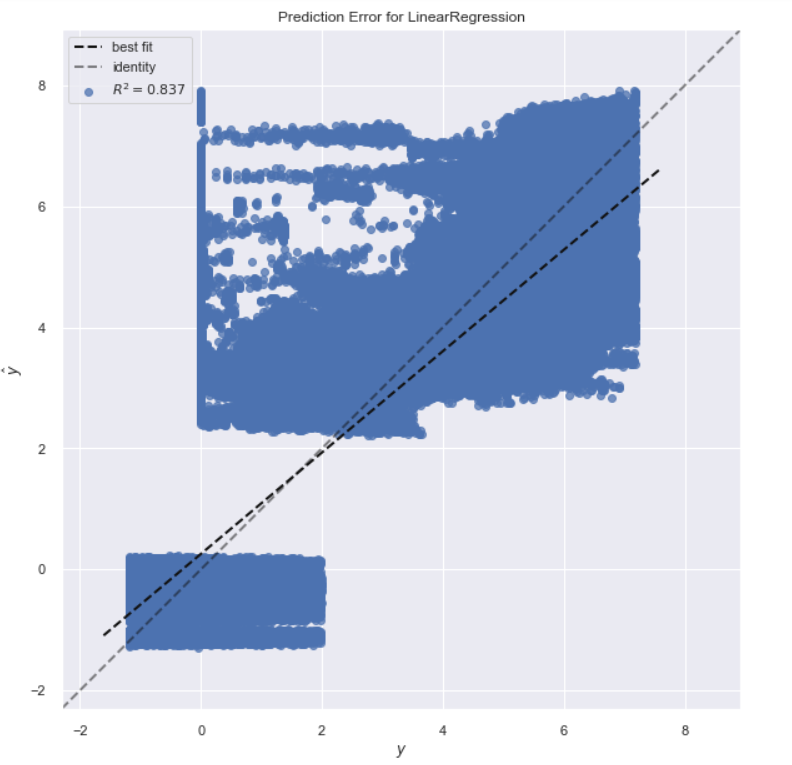
* **Meter reading with the time variables we had engineered: month, day and holiday:**





* **Meter reading using weather, building variables and time variables (without interaction):**





**7. Conclusion:**

Though it appeared at the beginning that there wouldn’t be much significance of our analysis other than the most logical one that bigger buildings use more electricity, we in fact came up with some very interesting observations that would be the base for any future analysis. The most notable discoveries were:

* We concluded that due to the complex nature of the dataset, using a simple model wouldn’t work. We would need to use a far more complex model to be better able to predict meter readings of electricity for all the buildings.
* We also found that adding weather data onto the building metadata made the model predictions a bit better. However, we were not able to effectively add the time variables into our simple model. The R2 value of model decreased when using features engineered to add time information.

**8. References:**

1. <https://www.ashrae.org/about>
2. <https://www1.eere.energy.gov/buildings/publications/pdfs/corporate/bt_stateindustry.pdf>
3. <https://www.eia.gov/todayinenergy/detail.php?id=39092>
4. <https://en.wikipedia.org/wiki/Linear_regression>
5. <https://www.kaggle.com/c/ashrae-energy-prediction/data>