**INFO 7390- Advance Data Science**

Case: Machine learning with Energy datasets

Course: INFO7390

Advance Data Science & Architecture

PROFESSOR:

Srikanth Krishnamurthy

Submitted By:

Team 8

Birwa Galia

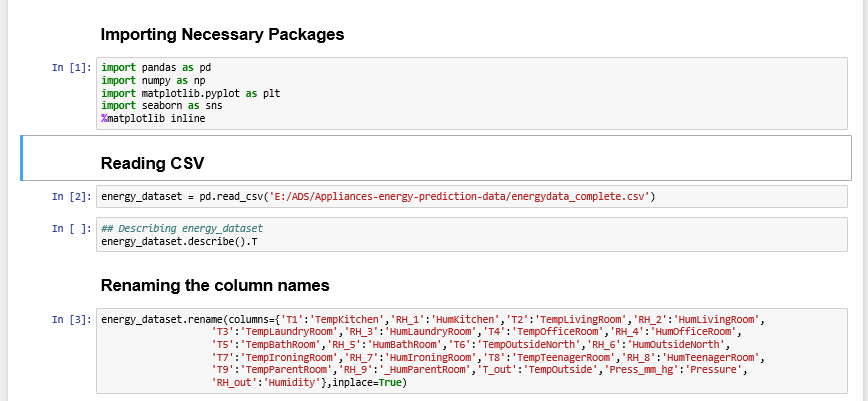
Milony Mehta

Shantanu Deosthale

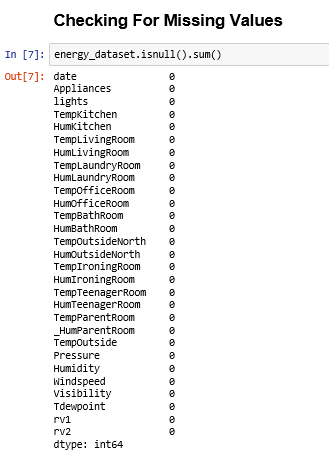
Table of Contents:

1. Introduction…………………………………………………………………………………………………………
2. Part 2: Exploratory Data Analysis…………………………………………………………………………
3. Part 3: Feature Engineering………………………………………………………………………………….
4. Part 4: Prediction Algorithm…………………………………………………………………………………
5. Part 5: Feature Selection………………………………………………………………………………………
6. Part 6: Model Validation and Selection…………………………………………………………………
7. Part 7: Final Pipeline…………………………………………………………………………………………….
8. Part 8: Report and Model development methodology………………………………………….

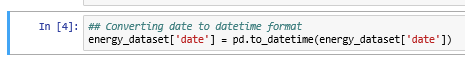
Part 2: Exploratory Data Analysis:



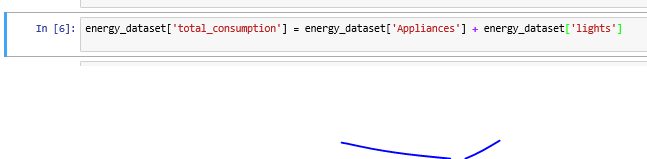
1. Importing Packages
2. Loading the CSV
3. Renaming the column name for better understanding and labels



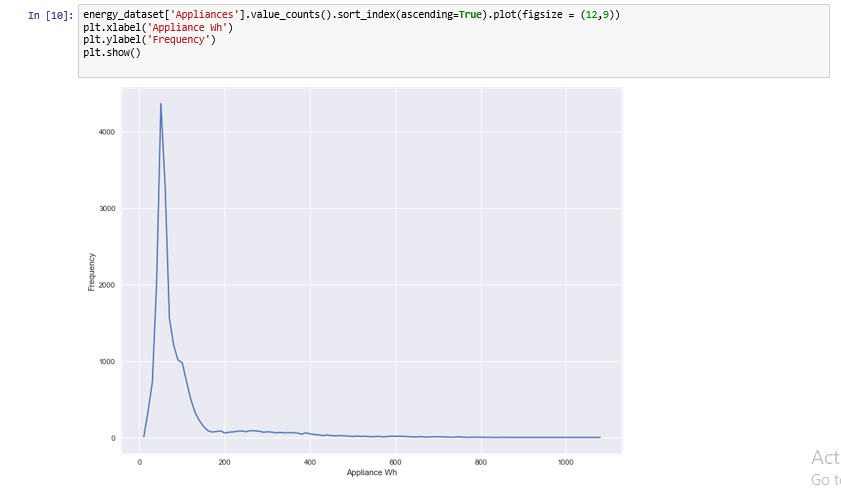
Checking for missing values. There were no missing values found.



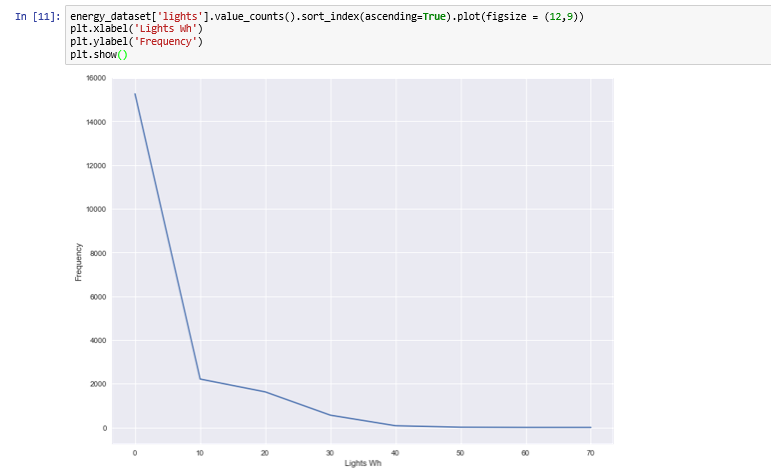
Converting the Object Date column to Datetime Format.



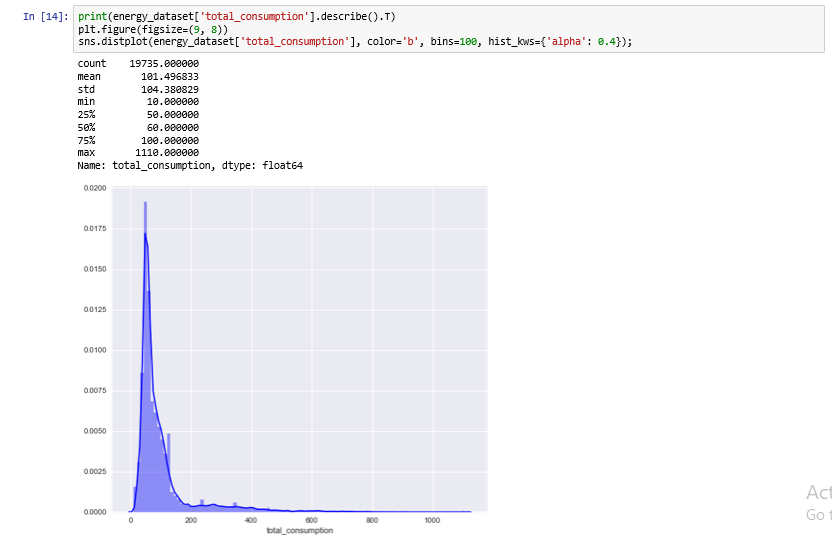
Forming new column “Total\_Consumption” by adding Appliance and lights columns.



Frequency count of Appliances in Wh



Frequency count of Lights in Wh



Distribution Plot of total consumption

A screenshot of a cell phone

Description generated with very high confidence

* Performing feature transformation, converting datetime format to Day, Month, hour, and NSM
* Using seaborn package, plotting stripplot for day vs total\_consumption, Month vs total\_consumption, hour vs total\_consumption
* Stripplot is used to plot the categorical values.

A screenshot of a cell phone

Description generated with very high confidence

It can be seen that the energy consumption is almost the same throughout the week.

A screenshot of a cell phone

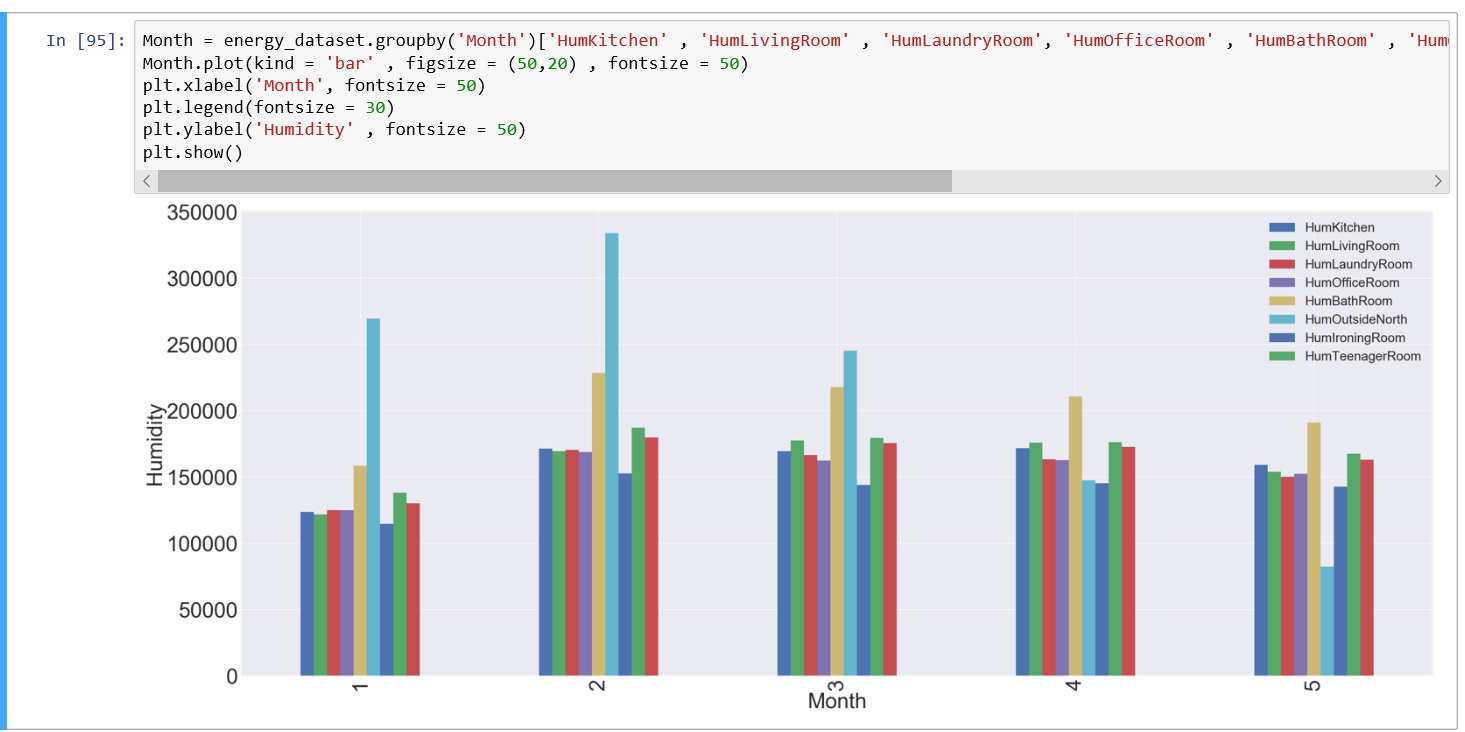
Description generated with very high confidence

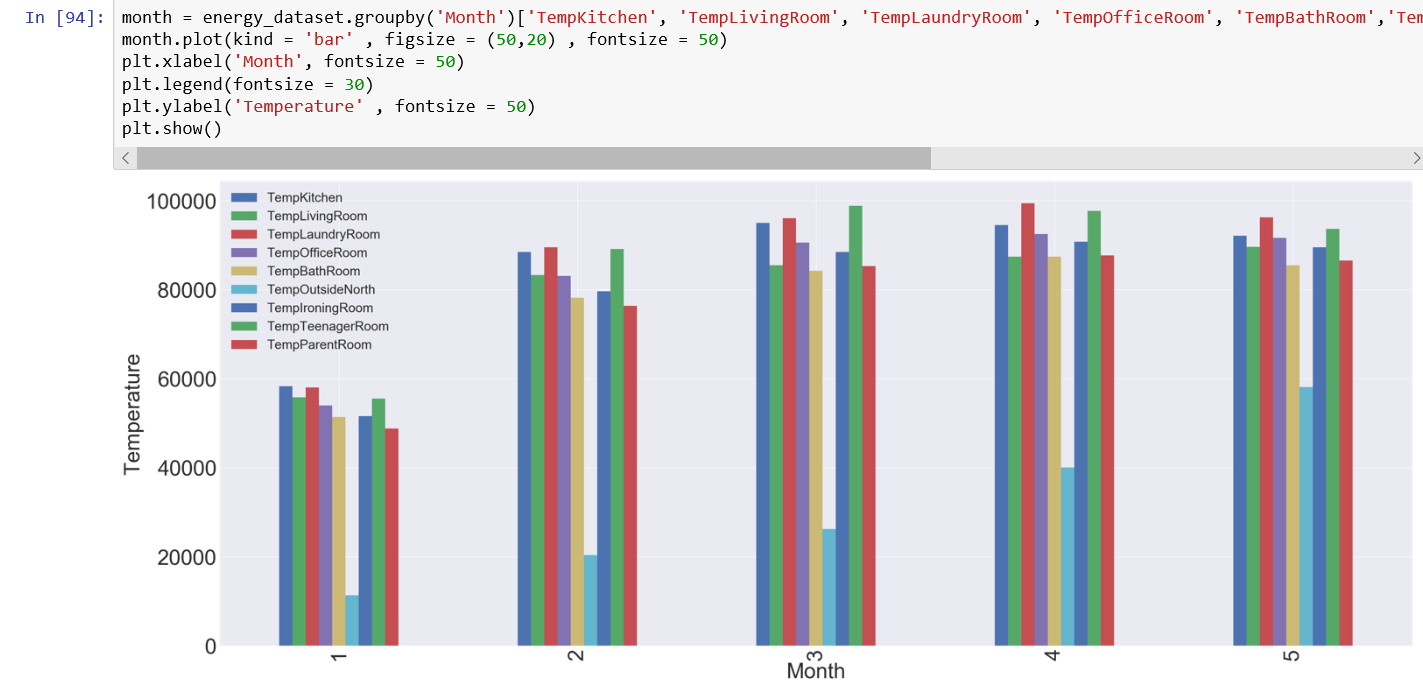
It can be seen that there is an increase in the energy consumption from 6AM onwards and it starts decreasing from 8PM onwards

A screenshot of a social media post

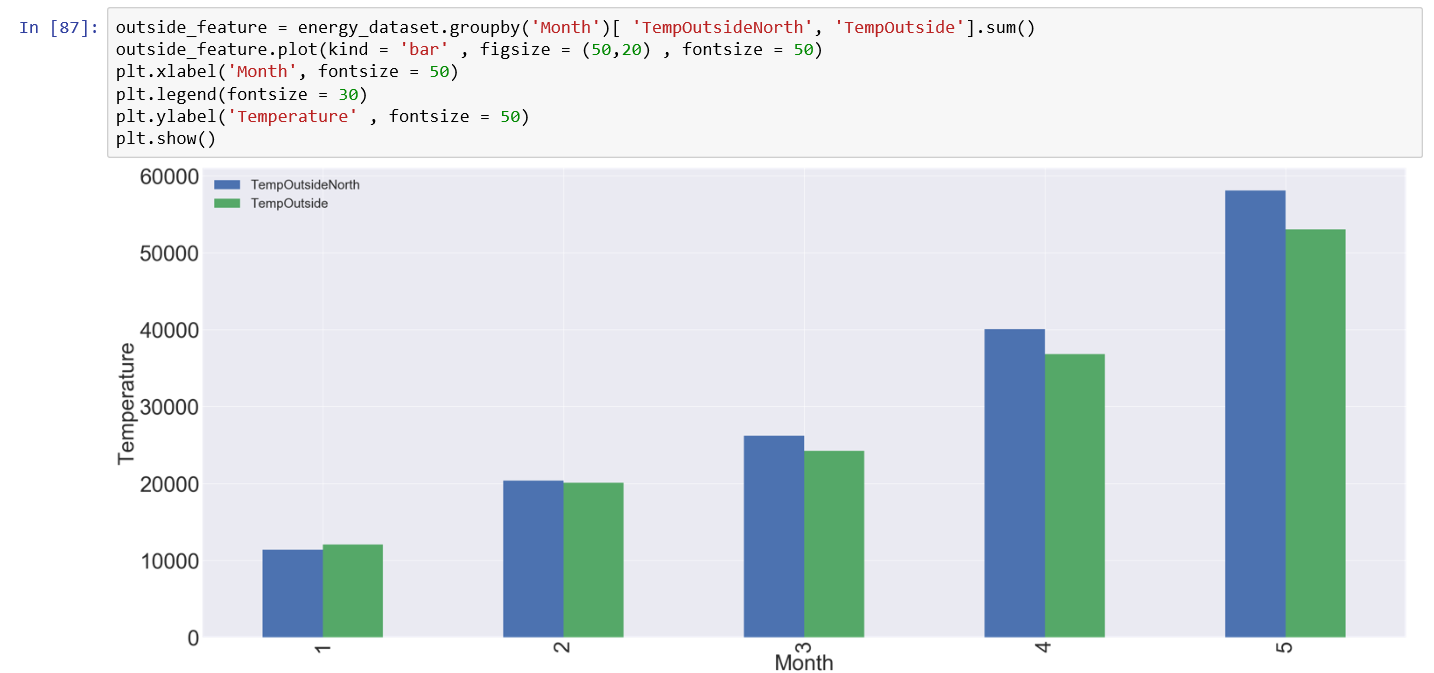
Description generated with very high confidence

It can be seen that the energy consumption is almost the same throughout the months

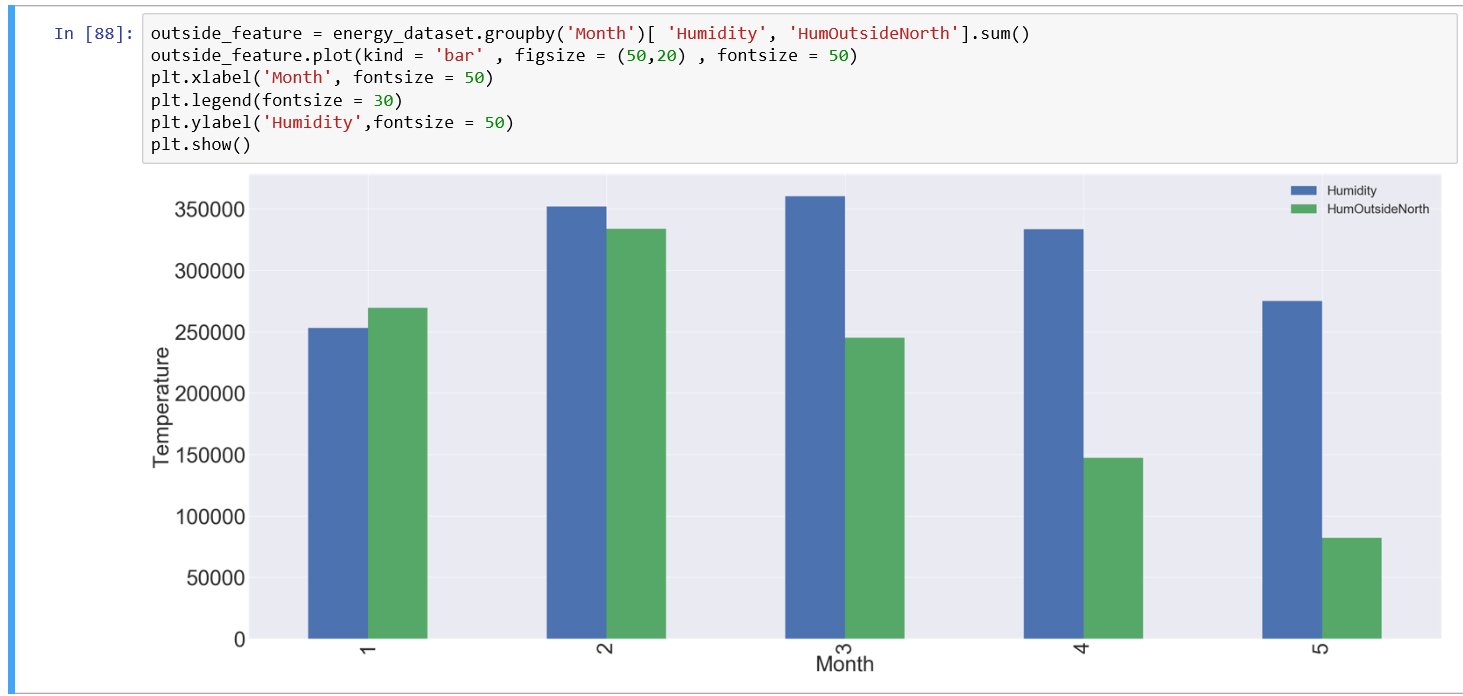


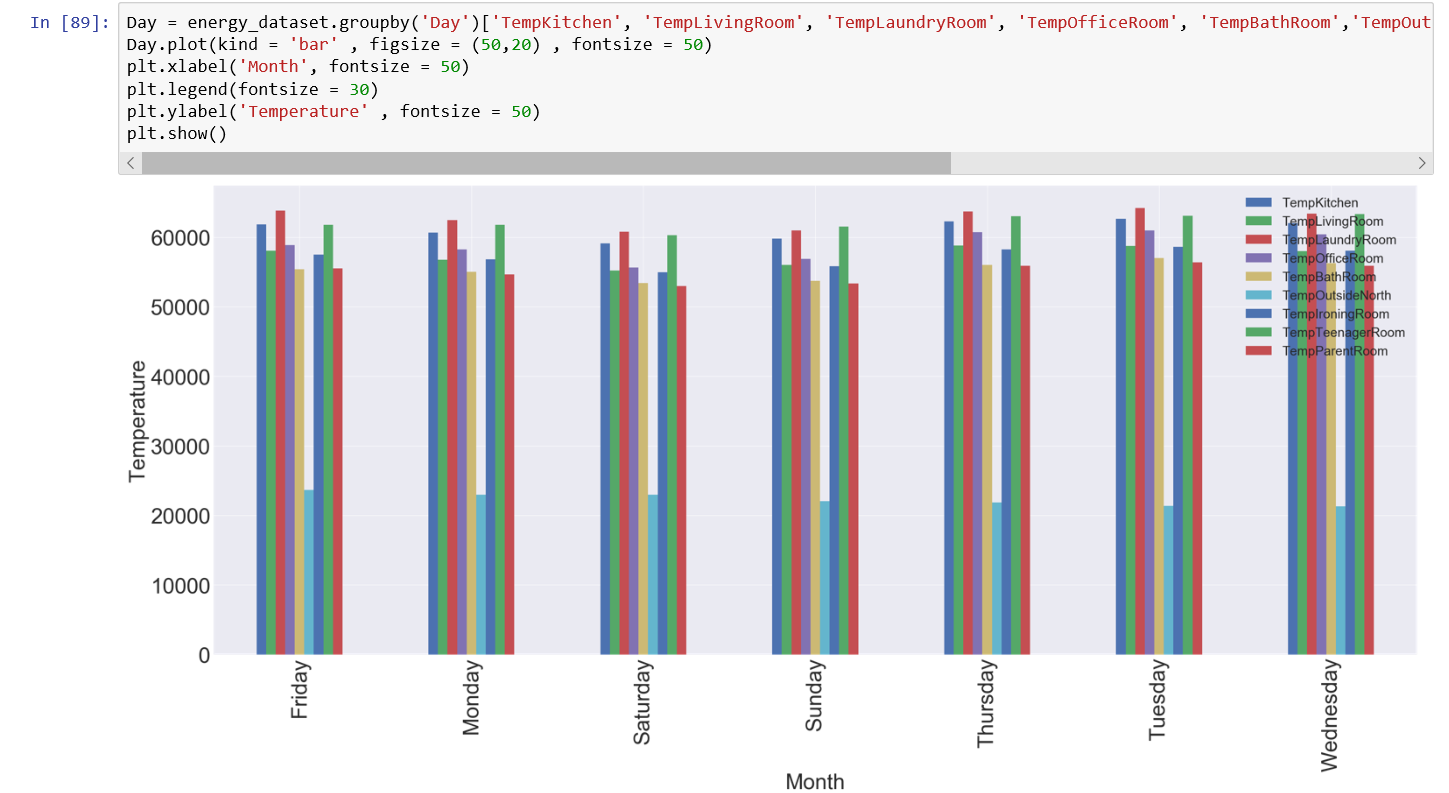


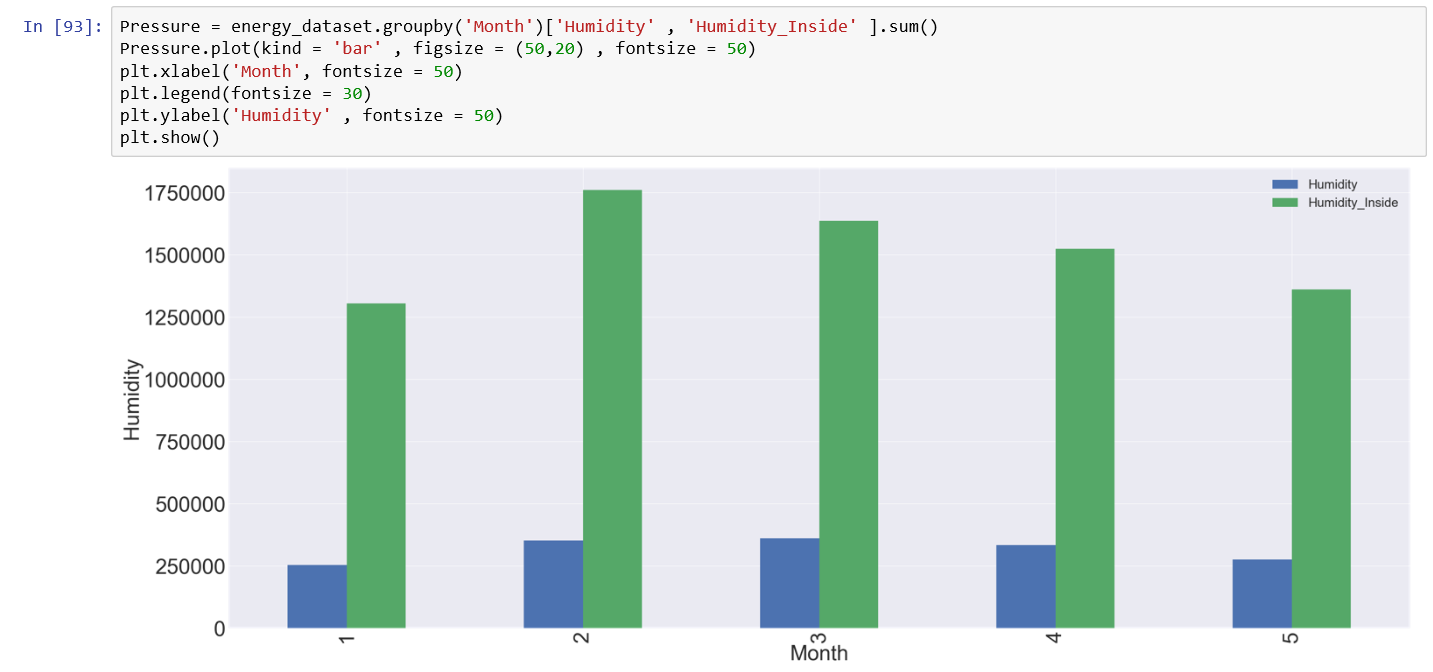
As the month increases the temperature also increases. The temperatures of all the rooms were grouped together



It can be seen that temperatures of both outside as well as on the northern-side of the building increases for the given months thus these temperatures are positively correlated.







Part 3: Feature Engineering:

The data types of each column were checked, converted the object type date to datetime and performed featuring transformation on date column.

A screenshot of a social media post

Description generated with very high confidence

The data was clean as there were no null values

A screenshot of a cell phone

Description generated with very high confidence

Calculating the correlation score

A screenshot of a cell phone

Description generated with very high confidence

Plotting the Heatmap using the correlation score.

A screenshot of a social media post

Description generated with very high confidence

From above heatmap, is it is clear that random variables and Visibility are not correlated to any of the remaining features thus, they can be removed

Part 4: Prediction Algorithms

After getting relevant features from Correlation, the obtained features were applied it to the following models: LinearRegression, Random Forest, Neural Network.

The target variable is “Appliances” and the following scores were obtained.

A screenshot of a cell phone

Description generated with very high confidence

As you can see from the scores, the r2\_test and rmse\_test score of Random Forest Regressor are better than LinearRegression and MLP\_NeuralNetwork, we would recommend Random Forest Regressor as a model based on the features obtained from Correlation.

Part 5: Feature Selection

To get the importance of the features Extra Trees Regressor was used

A screenshot of a social media post

Description generated with very high confidence

The calculated scores are as follows:

A screen shot of a social media post

Description generated with very high confidence

**Tpot Implementation:**

A screenshot of a cell phone

Description generated with very high confidence

With tpot implementation the best pipeline received was ExtraTreesRegressor, and the scores were as follows:

A screen shot of a social media post

Description generated with very high confidence

**Boruta Implementation:**

A screenshot of a social media post

Description generated with very high confidence

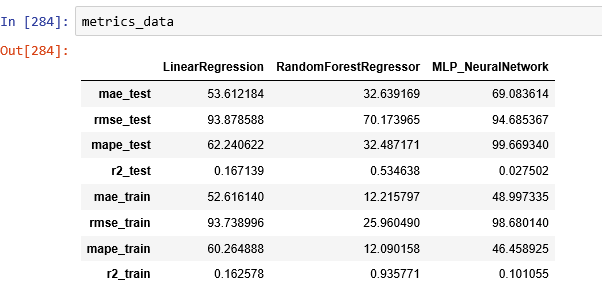
With Boruta implementation, the following best features were received:

A screenshot of a cell phone

Description generated with very high confidence

With rank 1 as the best features and rest as Tentative or Rejected features.

LinearRegression, Random Forest, Neural Network models were applied on the features obtained by Boruta and the scores were as follows:



Stepwise Selection implementation:

A screenshot of a cell phone

Description generated with high confidence

It includes both forward selection and backward elimination.

A screenshot of a social media post

Description generated with very high confidence

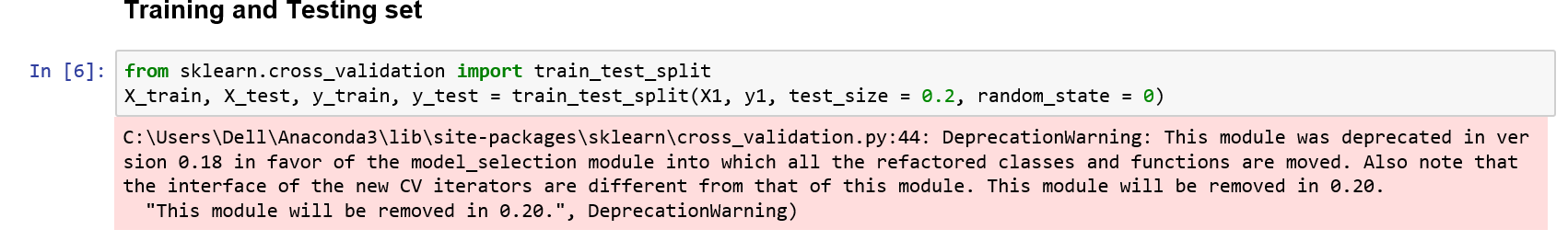
The above features from stepwise selection were received.

LinearRegression, Random Forest, Neural Network models were applied on the features obtained by Stepwise Selection and the scores were as follows:

A screenshot of a cell phone

Description generated with very high confidence

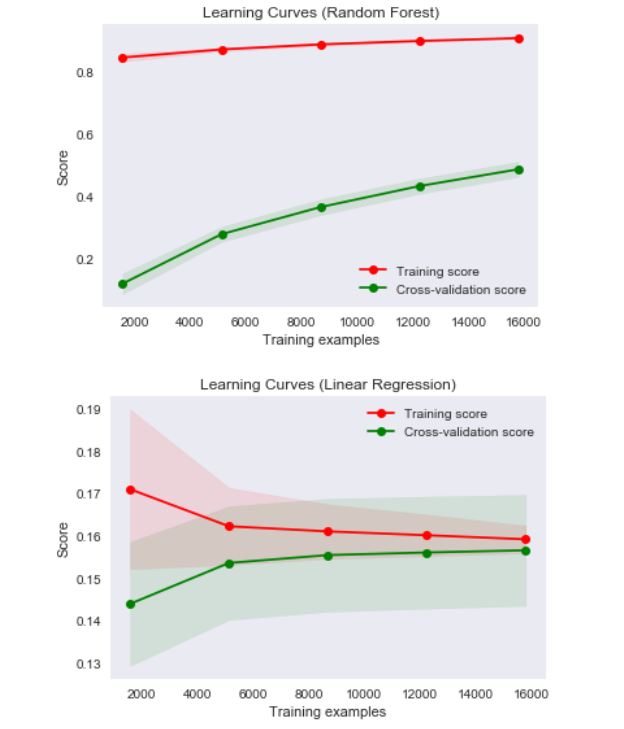
Part 6: Model Selection and Validation

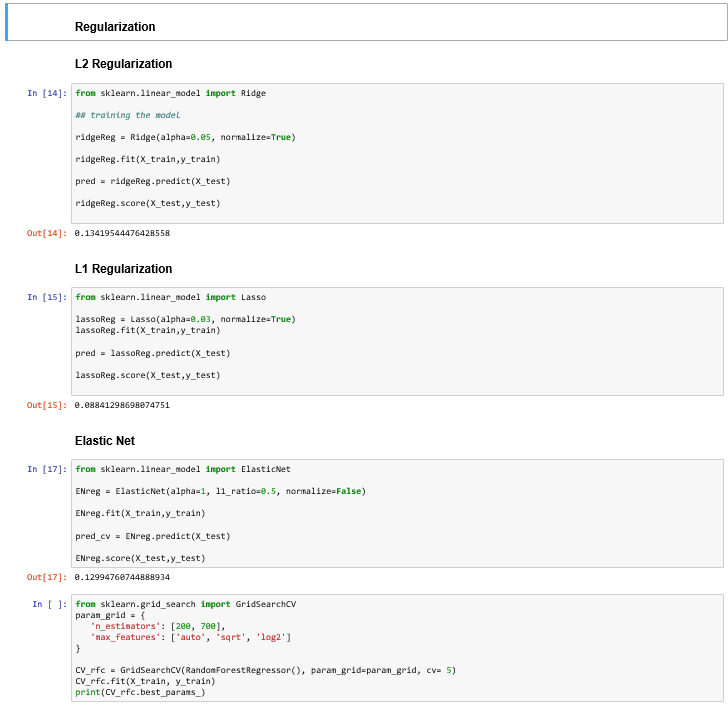


The dataset was split into 80% training and 20% testing data



**Bias Variance Trade-off**





Part 8: Final Pipeline

RandomForestRegressor was selected as our machine-learning algorithm for this dataset.

The features used were evaluated as important using Boruta

