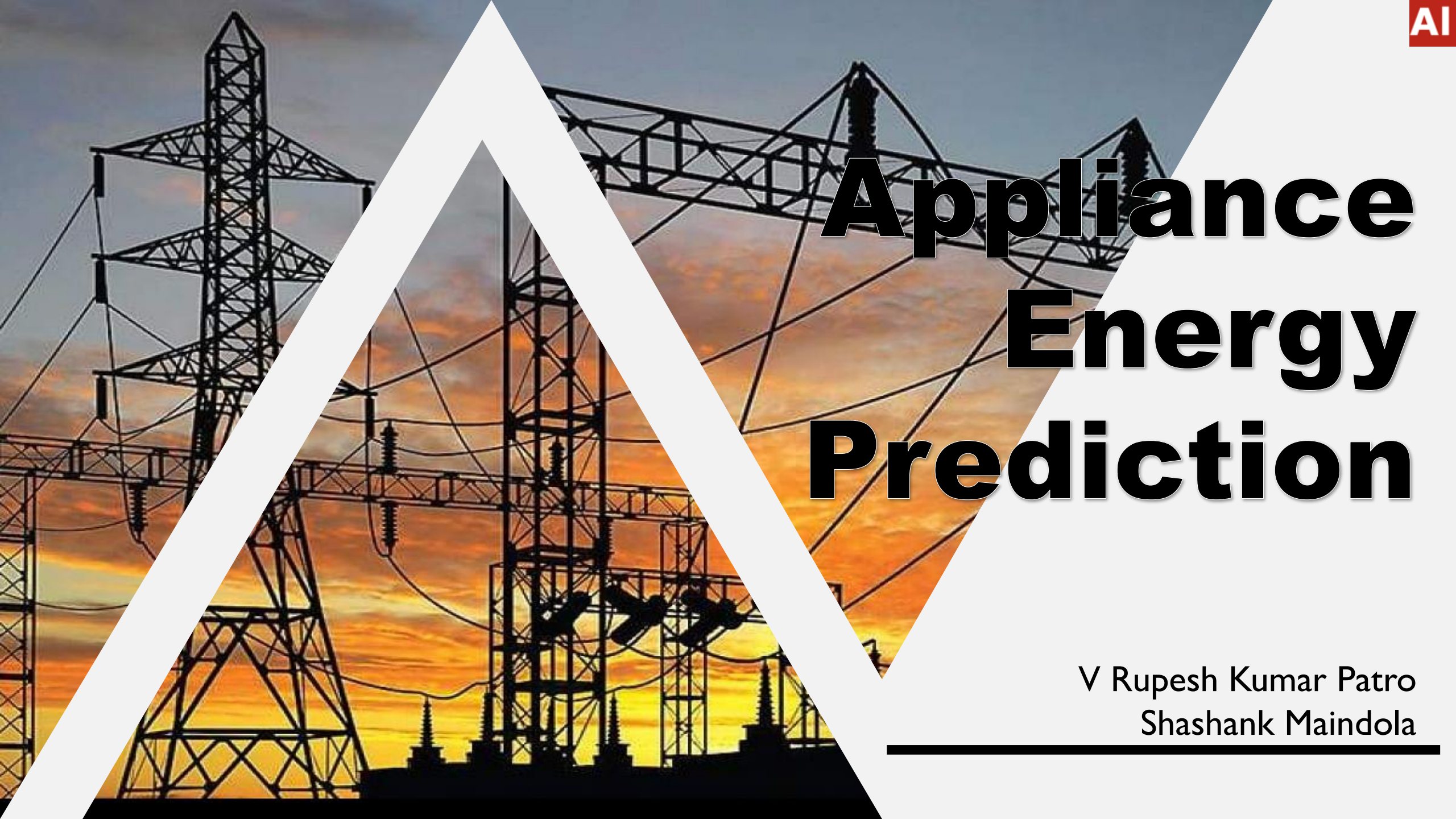


Appliance Energy Prediction

V Rupesh Kumar Patro
Shashank Maindola

Introduction

In this century, the world is being driven towards more cleaner means of energy. The electricity generated using the renewable sources plays a major role. But simply generating and utilizing the electrical energy is not viable, any extra generation will lead to non utilization and less generation will lead to power outages. Hence, we need demand side and supply side management so we can flatten the generation distribution curve. The residential load contributes to 27% of the total energy consumption (**Source:**:Eurostat). Residential household consumption is generally governed by weather conditions. Here we have a data of one of e residential building in Belgium and using the machine learning techniques we have to figure out which algorithm give the best output.

Few of the salient features of our dataset:-

- There are 29 columns and 19735 rows in our dataset.
- Max energy usage of appliance is 1080 and min is 10 watt
- light column having majority of the data 0 values
- Max pressure outside house is 772.3 mm_hg
- Except date column There is none categorical column in the dataset.
- Average temperature outside is about 7.5 degrees. While it ranges from -6 to 28 degrees.
- There is no null or missing values.
- Average humidity outside is higher than average humidity inside.
- Max wind speed is 14 m/s

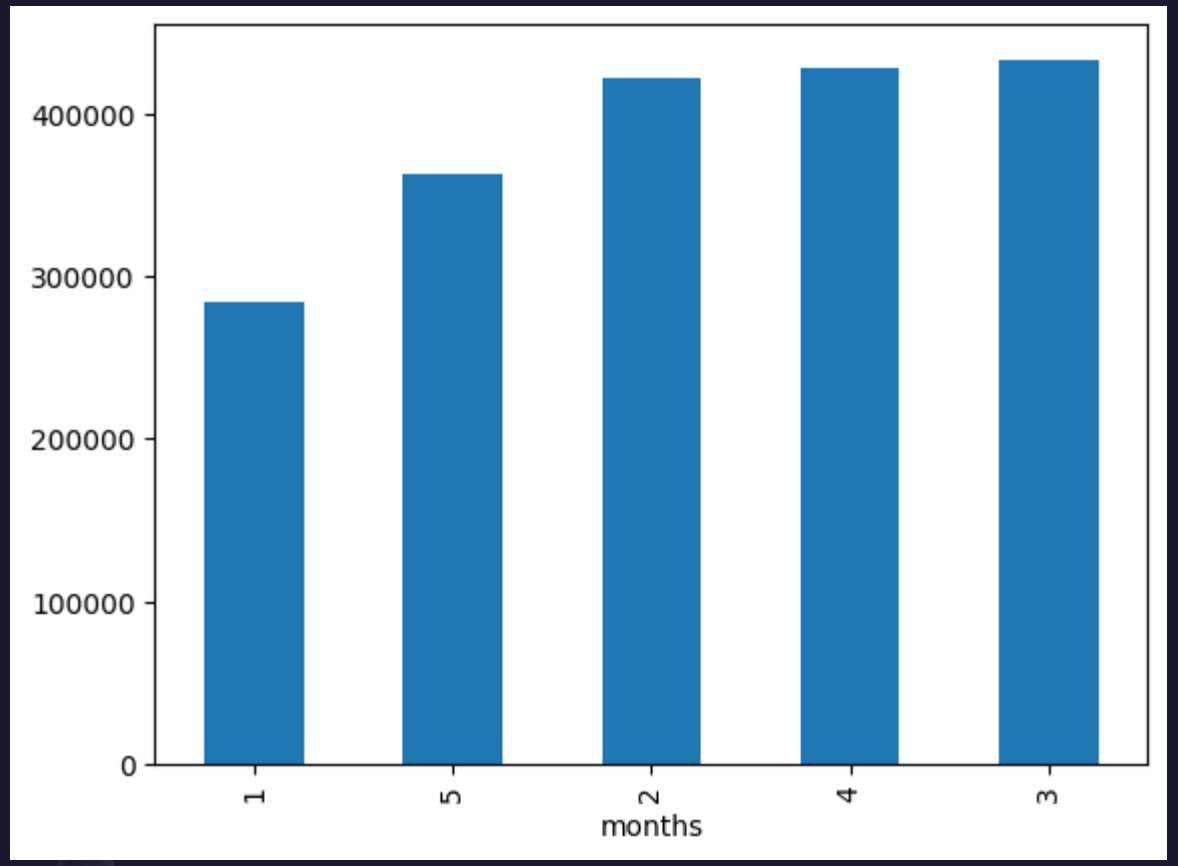
Appliance Energy Prediction using Machine learning

- EDA
- Clean up
- Feature Engineering
- Pre Processing
- Model Implementation
- Model Explanation

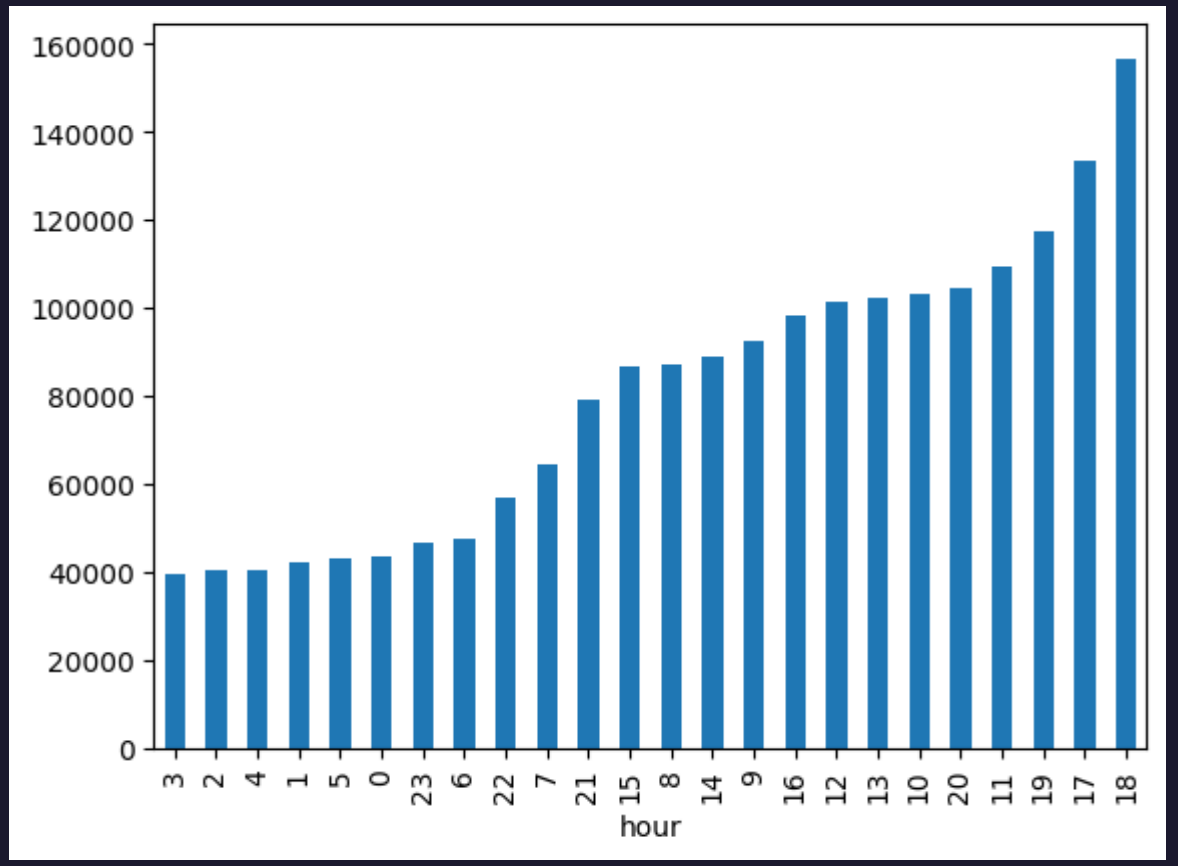
Exploratory Data Analysis(EDA)

We are provided with the data of 4.5 months for a household collected every 10 minutes using zigbee network. It contains temperature and humidity of different rooms, light and appliance energy consumption parameters of the household .We are also provided with visibility, atmospheric pressure, relative humidity, visibility, Dewpoint collected from Chievres weather station, Belgium

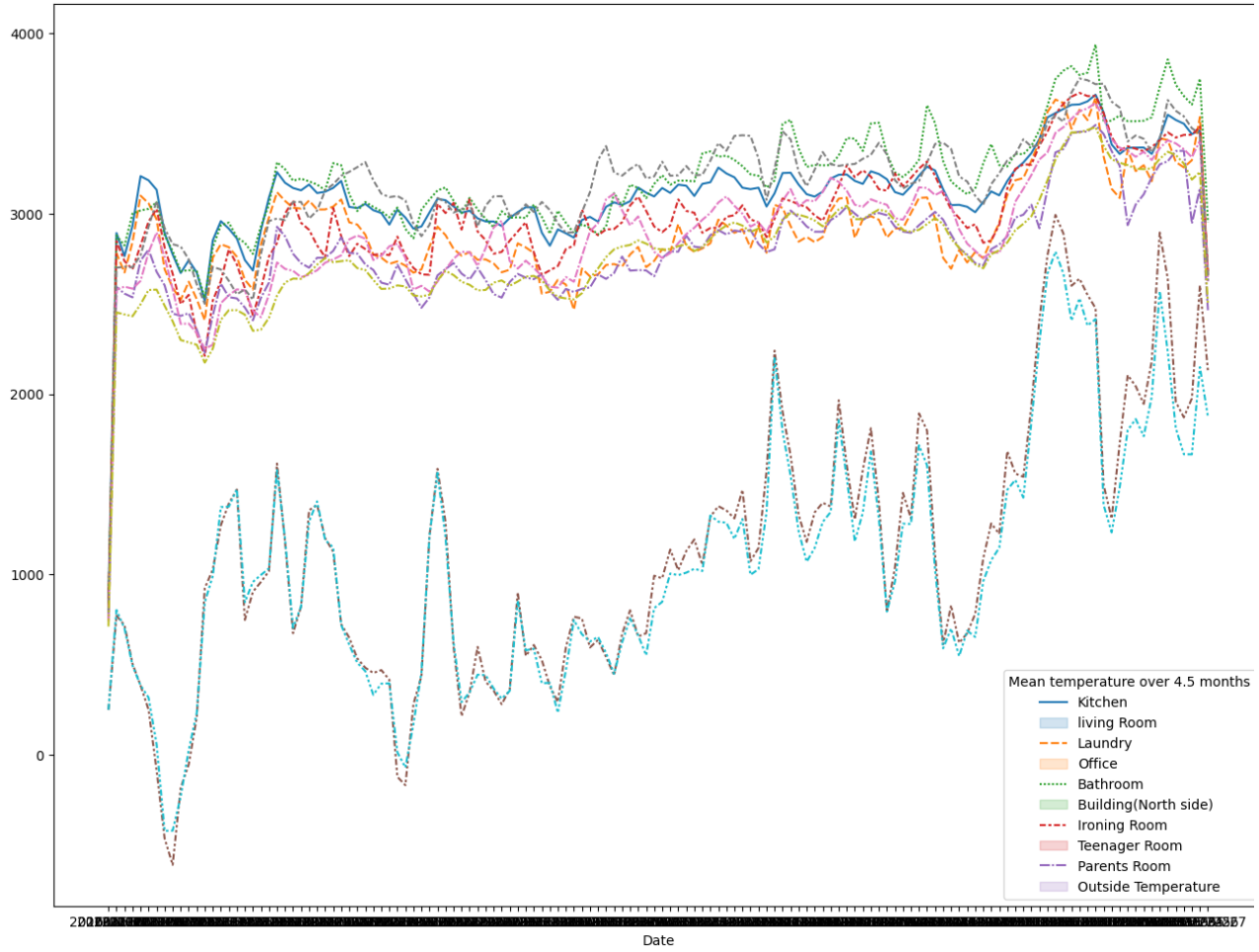
Fields	Description
T1	Temperature in kitchen area, in Celsius
T2	Temperature in living room area, in Celsius
T3	Temperature in laundry room area, in Celsius
T4	Temperature in office room, in Celsius
T5	Temperature in bathroom, in Celsius
T6	Temperature outside the building (north side), in Celsius
T7	Temperature in ironing room , in Celsius
T8	Temperature in teenager room 2, in Celsius
T9	Temperature in parents room, in Celsius
RH1	Humidity in kitchen area, in %
RH2	Humidity in living room area, in %
RH3	Humidity in laundry room area, in %
RH4	Humidity in office room, in %
RH5	Humidity in bathroom, in %
RH6	Humidity outside the building (north side), in %
RH7	Humidity in ironing room, in %
RH8	Humidity in teenager room 2, in %
RH9	Humidity in parents room, in %
To	Temperature outside (from Chievres weather station), in Celsius
Pressure	(from Chievres weather station), in mm Hg
Hg RHout	Humidity outside (from Chievres weather station), in %
Wind speed	(from Chievres weather station), in m/s
Visibility	(from Chievres weather station), in km
Tdewpoint	(from Chievres weather station), Â°C
Appliances, energy use in Wh	Dependent variable

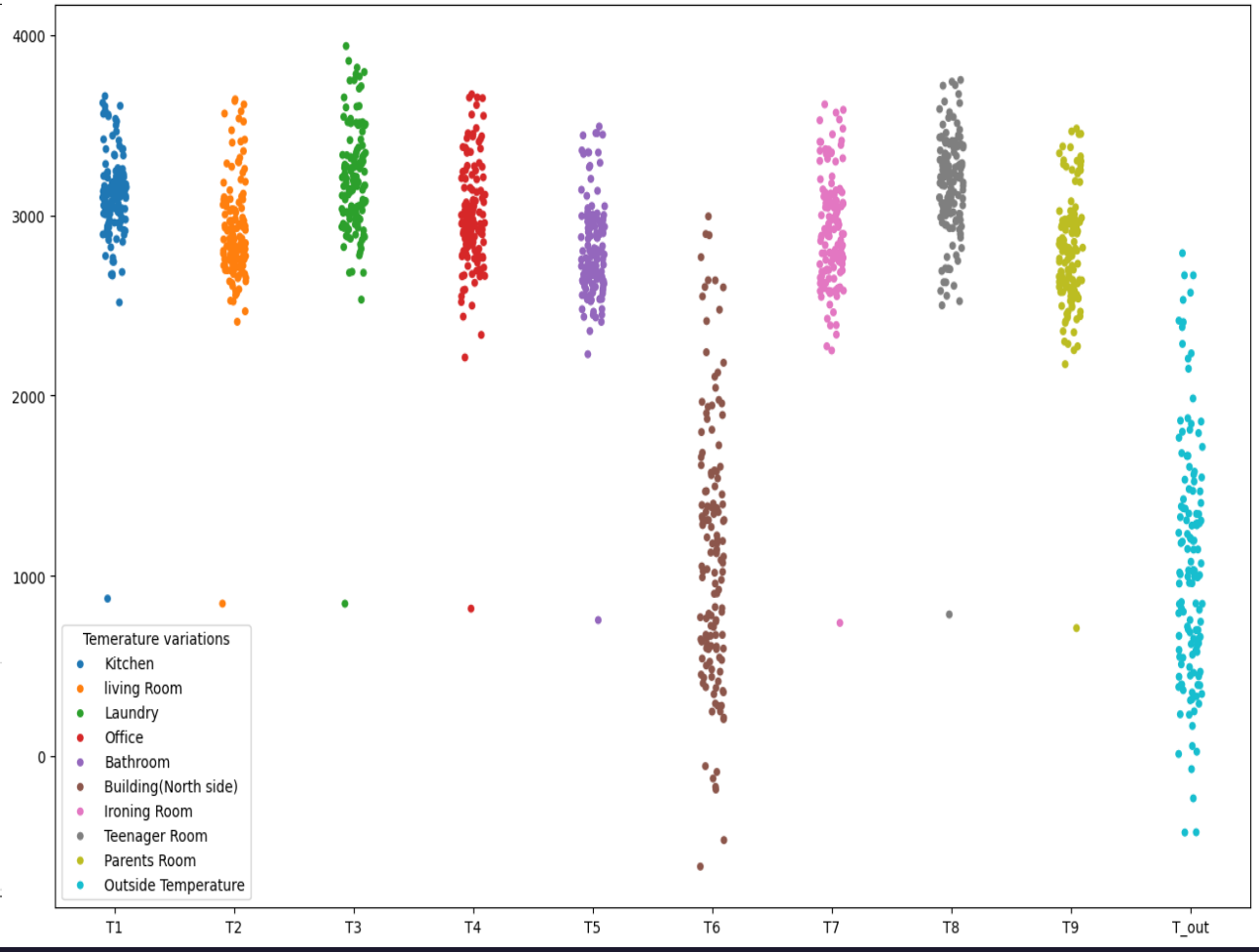


Total Monthly Energy consumption

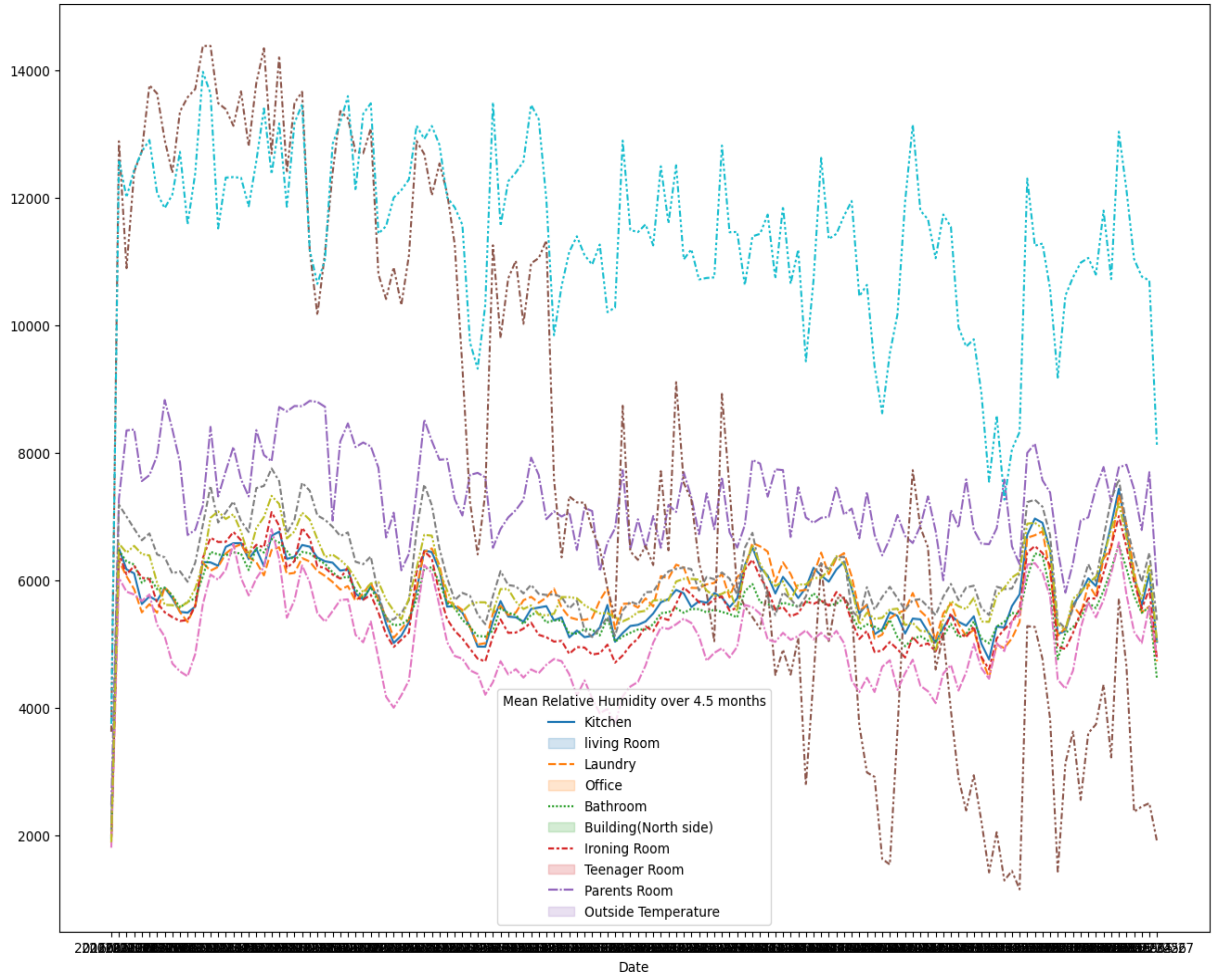


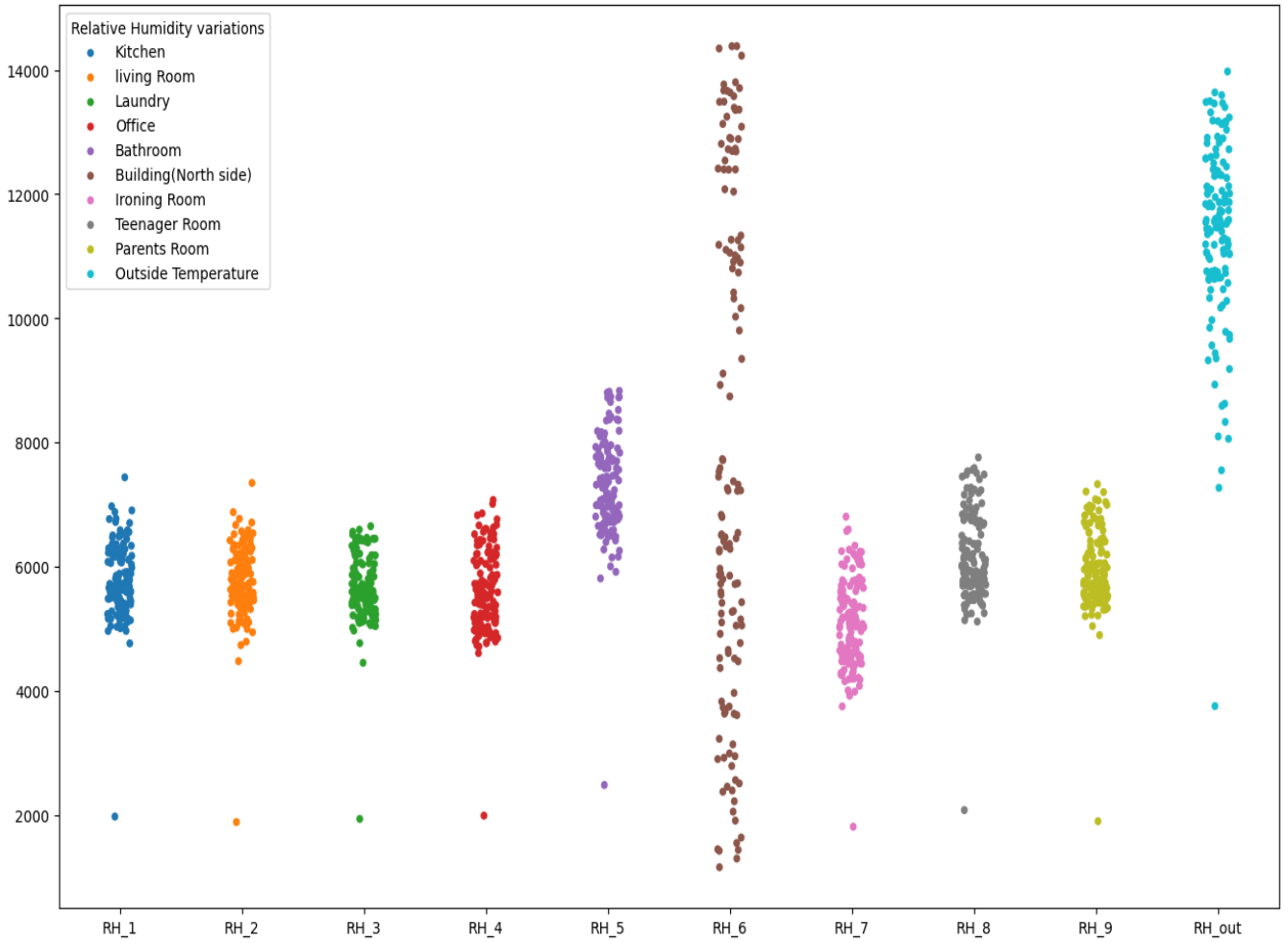
Total Hourly Energy Consumption



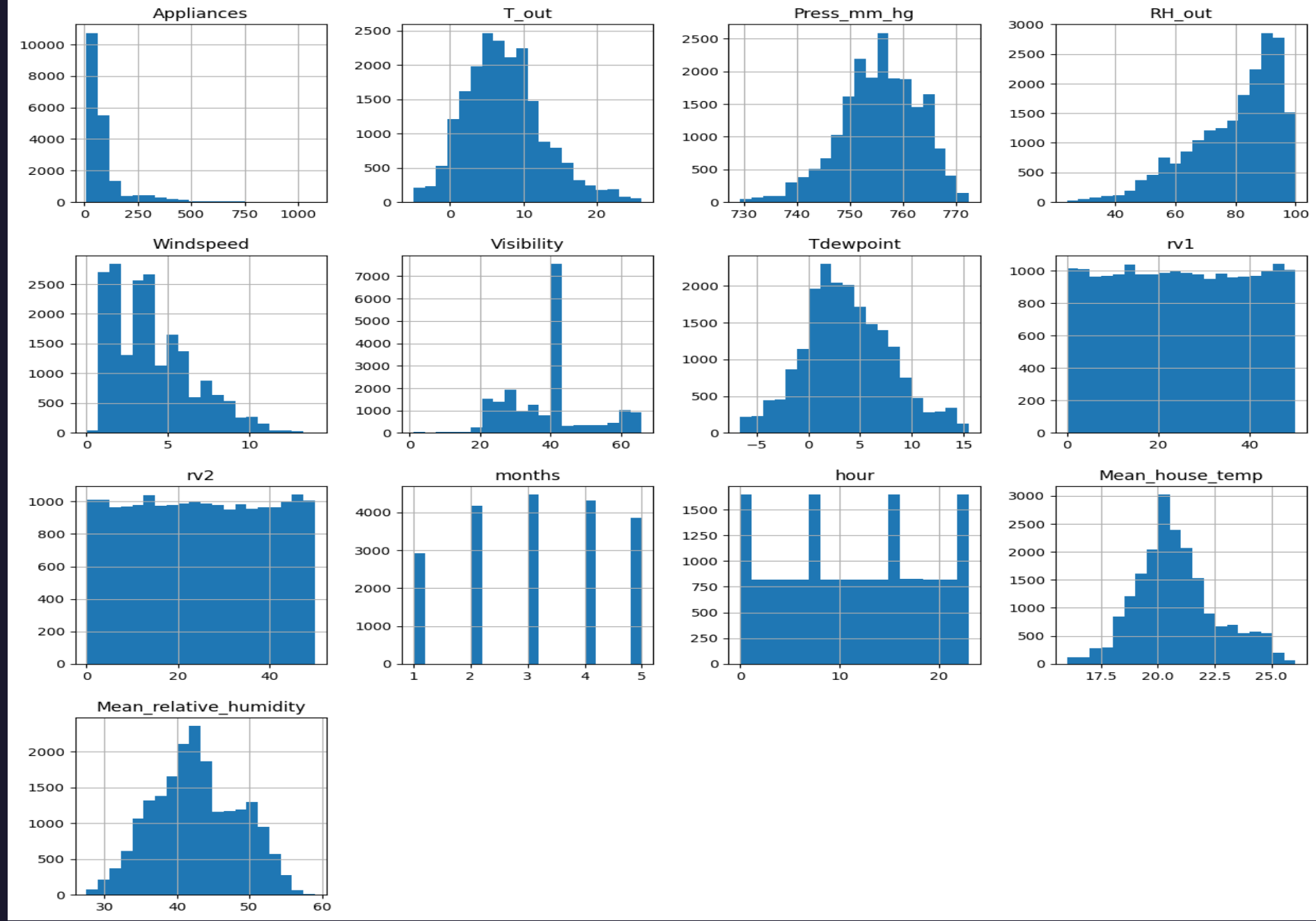


Household and outside temperature

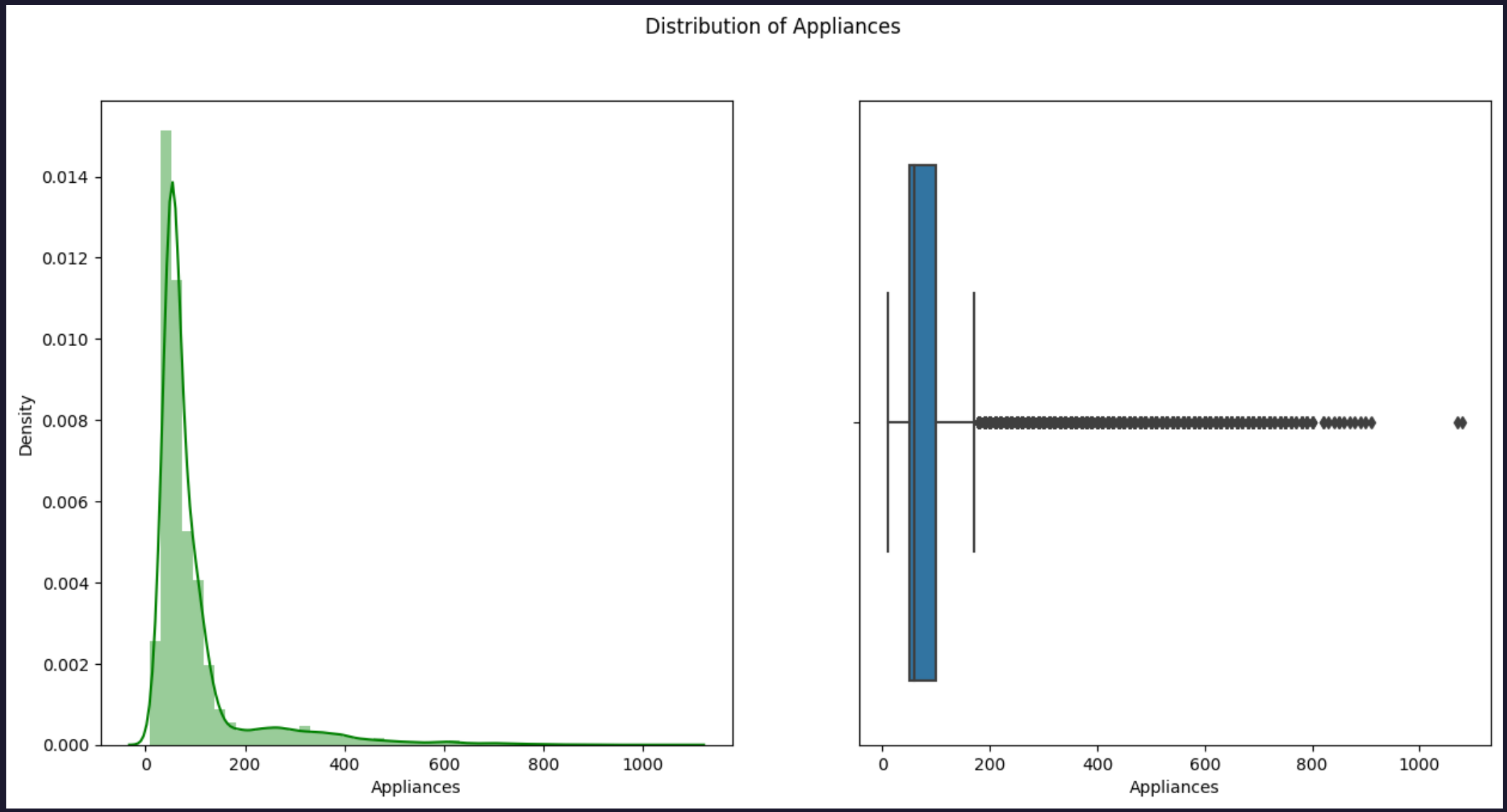




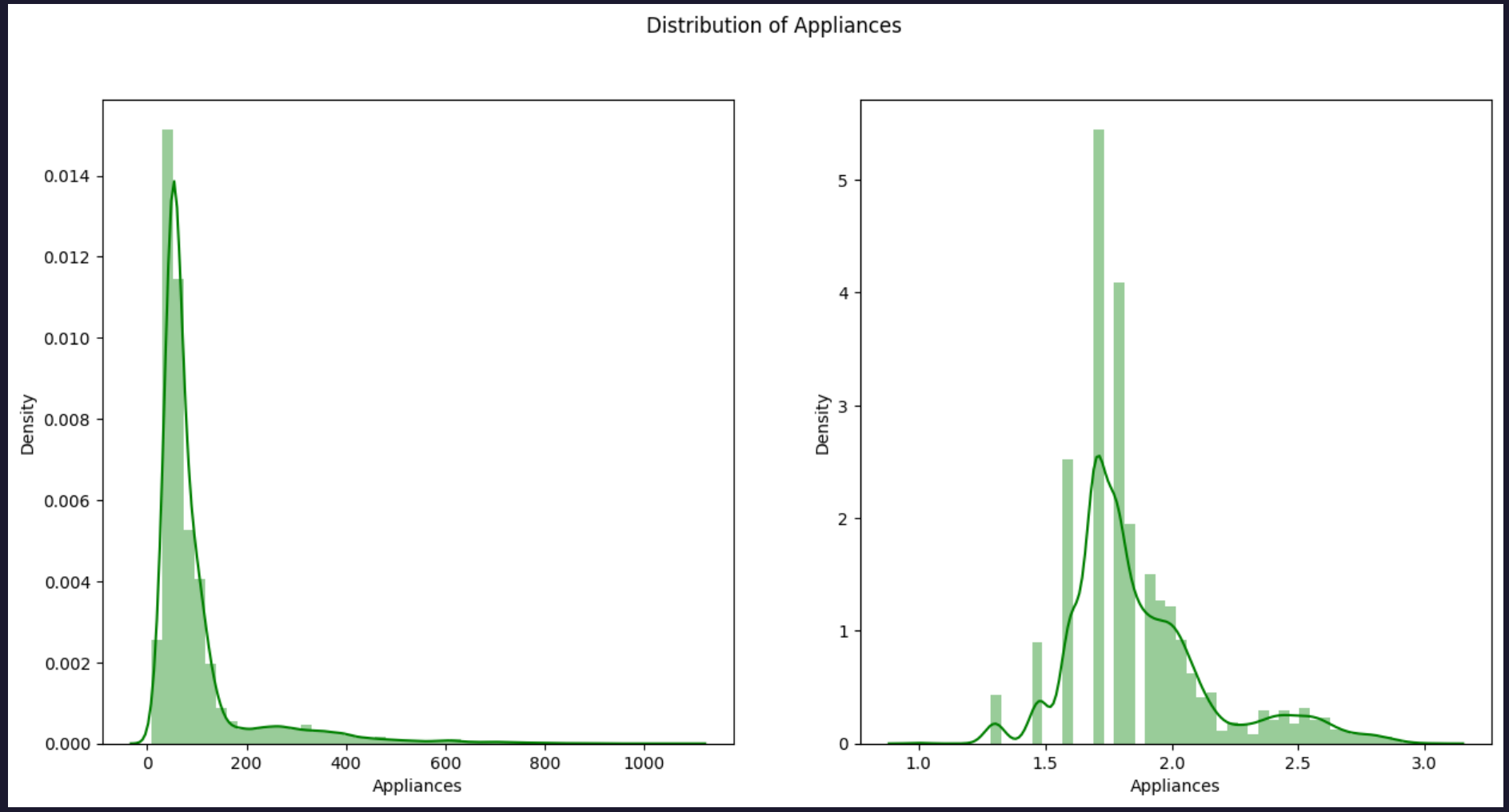
Household and outside Humidity



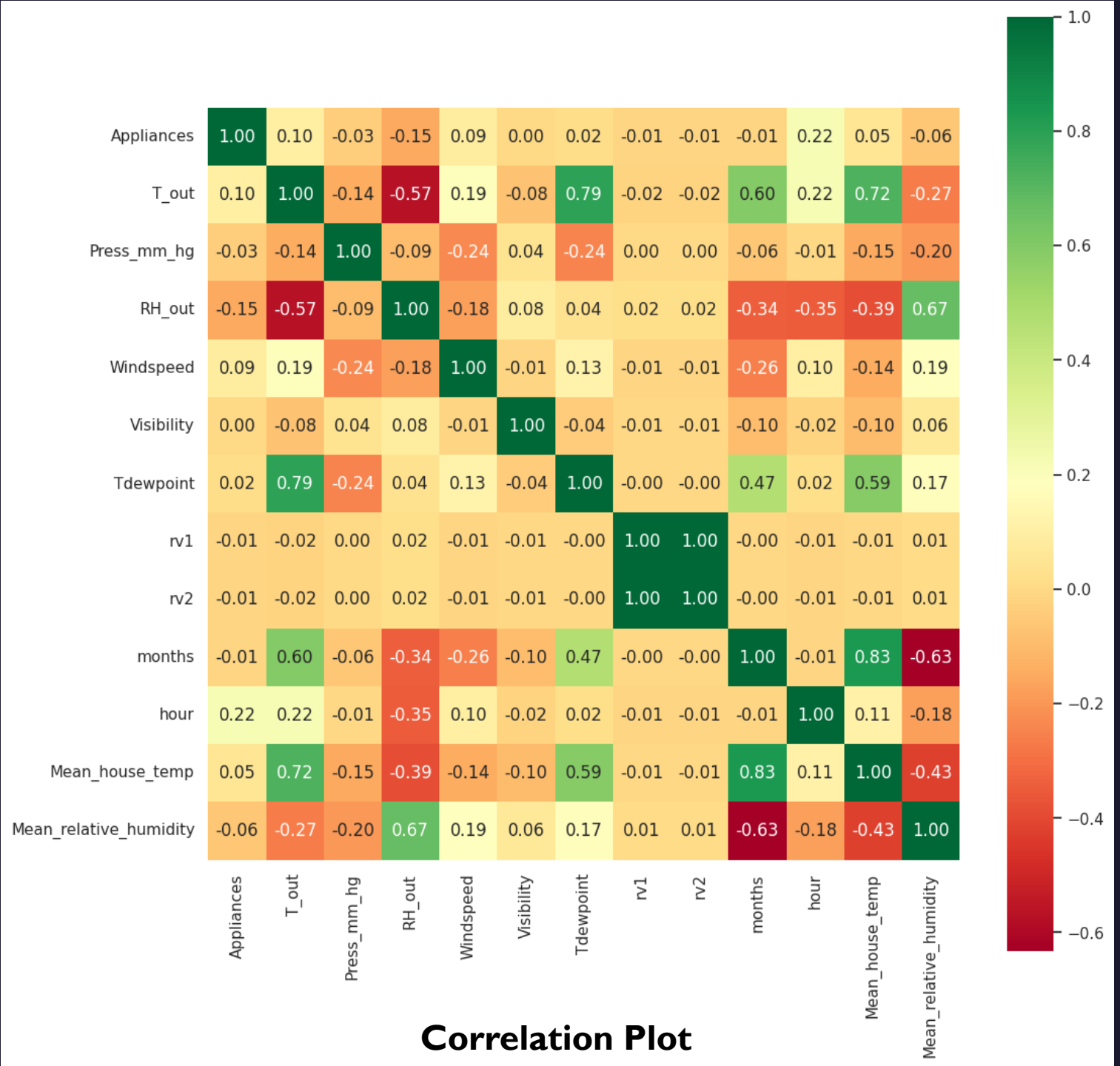
Distribution of all the parameters



Distribution of Appliance Energy Consumption



Our graph is moving towards to y axis as it is positively skewed and we couldn't get any better visualization with these type of graph. Hence we took Log or Square Root or Exponential of the dependent variable and draw the graph.



Correlation Plot

Temperature and Humidity

- The northside temperature as it is similar to north side temperature, hence it verifies that sensor was working perfectly and the data gathered is valid. The temperature inside is more or less same.
- There is a variation in relative humidity of Building (North side) and Outside. The outside humidity is from airport weather sensor so the humidity can be different while the temperature outside the building is different due to the neighborhood factors like landscaping etc., hence we will be ignoring the relative humidity data from the airport.

Seasonal and Hourly consumption Pattern

- It is clearly understood that march has the maximum energy consumption.
- The maximum power consumption is around evening while the least consumption is around the early morning hours.

Distribution

- Press_mm_hg, Visibility, TDewpoint, rv1, rv2, Mean_house_temp, Mean_relative_humidity,months, hour are normal distributed data.
- **Positively skewed(>1)**:- Appliances.
- **Moderately Positively skewed(0.5 to 1)**:- T_out, Windspeed.
- **Normal Distributed(-0.5 to +0.5)**:- Press_mm_hg, Visibility, Tdewpoint, rv1, rv2, Mean_house_temp, Mean_relative_humidity, months,hour.
- **Negative skewed(-0.5 to -1)**:- RH_6.

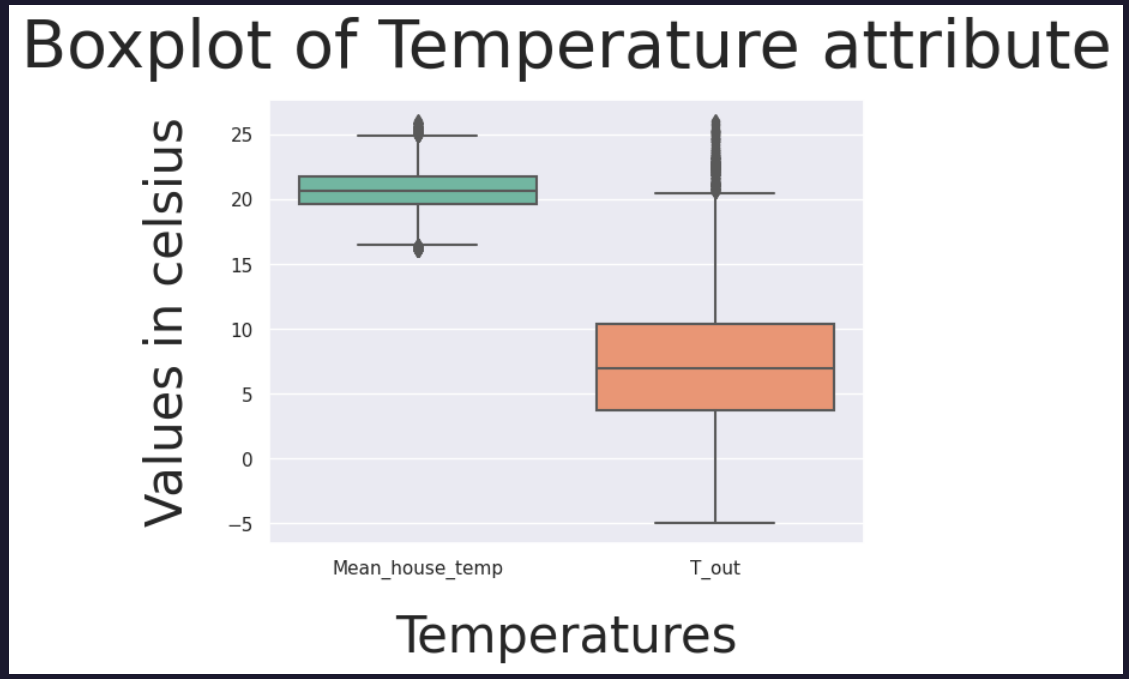
Target Variable

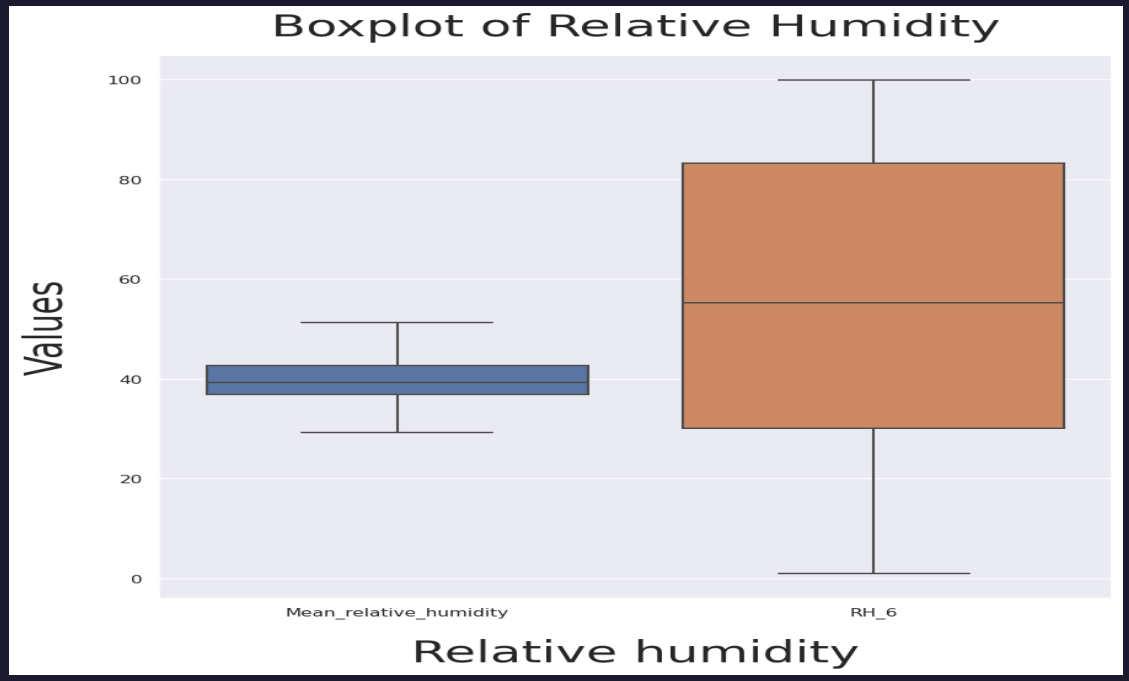
- Dependent variable is right skewed and lot of outliers present in our data set but they are not ignored at there are sometimes situations when the consumption increases here we see there are lot of such instances.
- There is positive correlation between temperature inside and outside with the appliance energy consumption.
- There is low correlation of appliance energy consumption with other variables

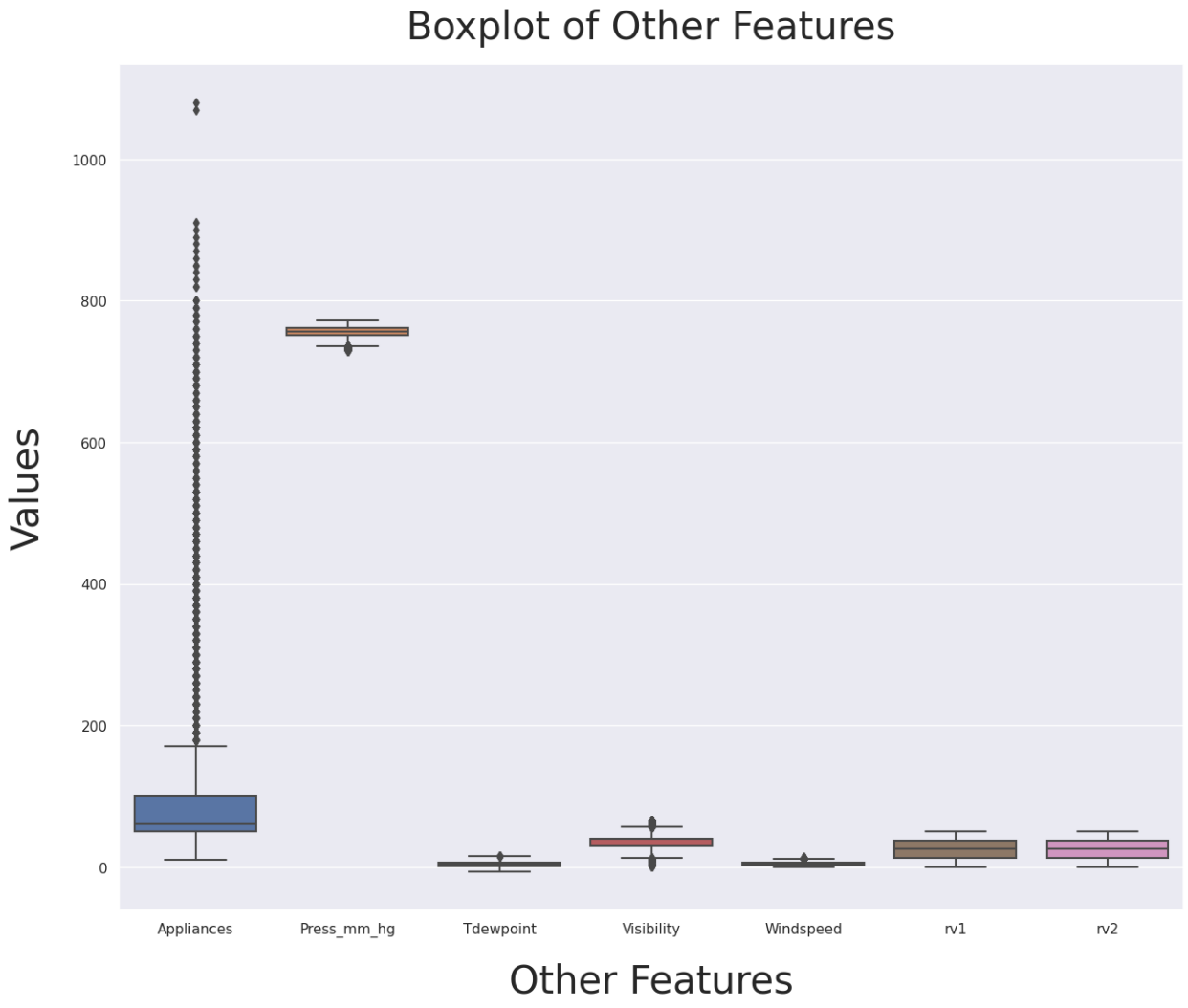
Clean Up

Number of outliers

Appliances	2138
T_out	436
Press_mm_hg	219
RH_6	0
Windspeed	214
Visibility	2522
Tdewpoint	10
rv1	0
rv2	0
months	0
hour	0
Mean_house_temp	512
Mean_relative_humidity	0





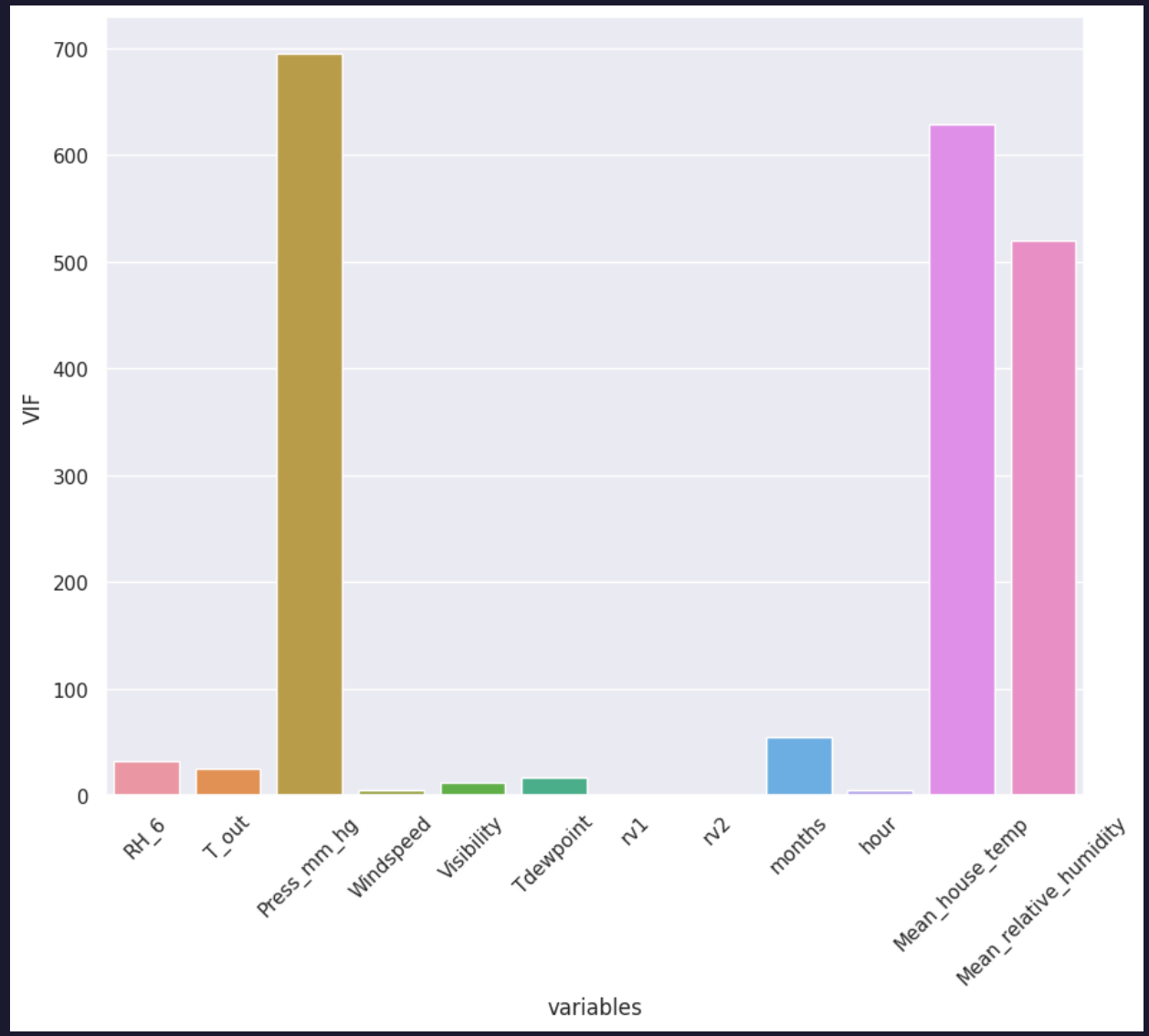


Firstly, we removed certain features like temperature in individual rooms, humidity in individual rooms and replaced them with the mean values for faster model training. We ignored the humidity in the bathroom and humidity at the airport as humidity in bathroom will always be higher than rest of the house and humidity of the airport doesn't represent the humidity of the neighborhood. Temperature outside is almost same as temperature of the northside of the building hence we will ignored temperature out side building. We removed r1 and r2 features as they have infinite VIF. We also removed light feature from our data set.

Secondly, we observe that there are a lot of outliers in visibility and appliance but we will not remove them as there are conditions when there is spike in demand and they are realistic hence we will not ignore them. Rest of the feature have little outliers so we will not remove them

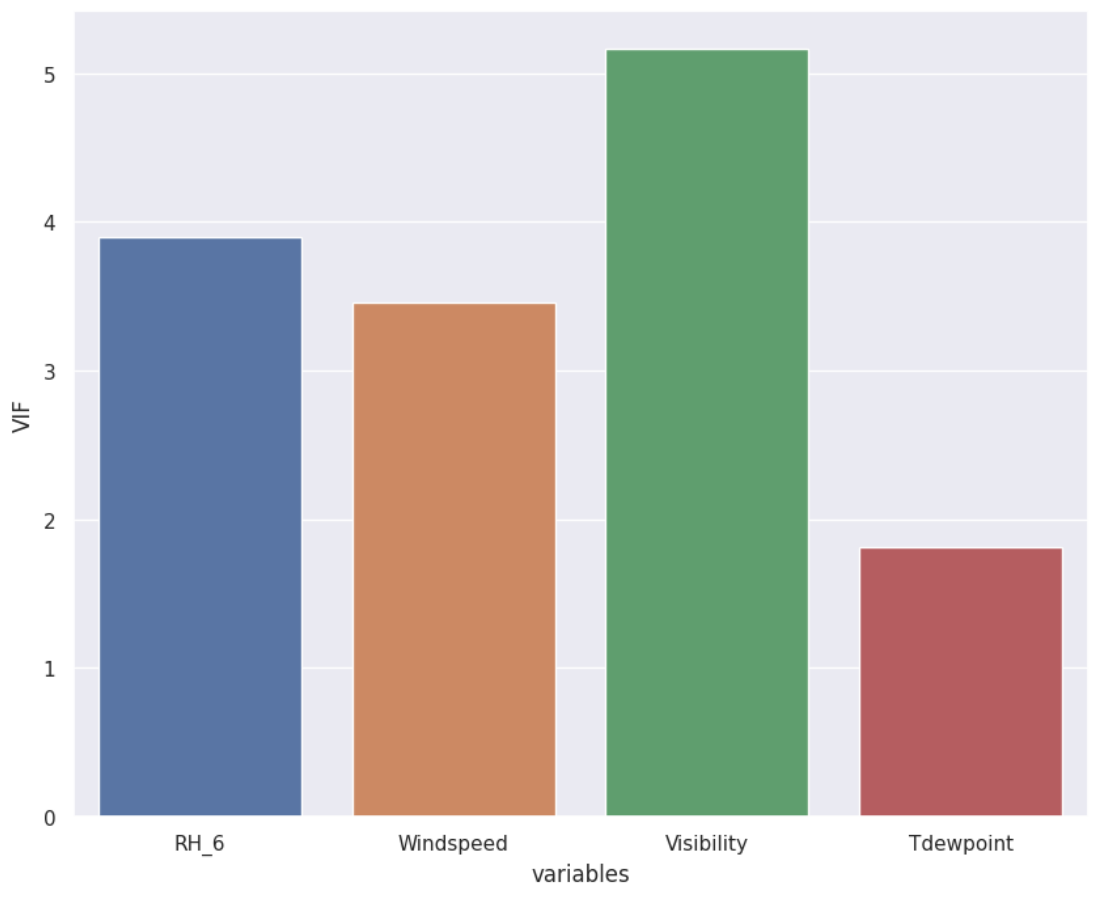
Feature Engineering

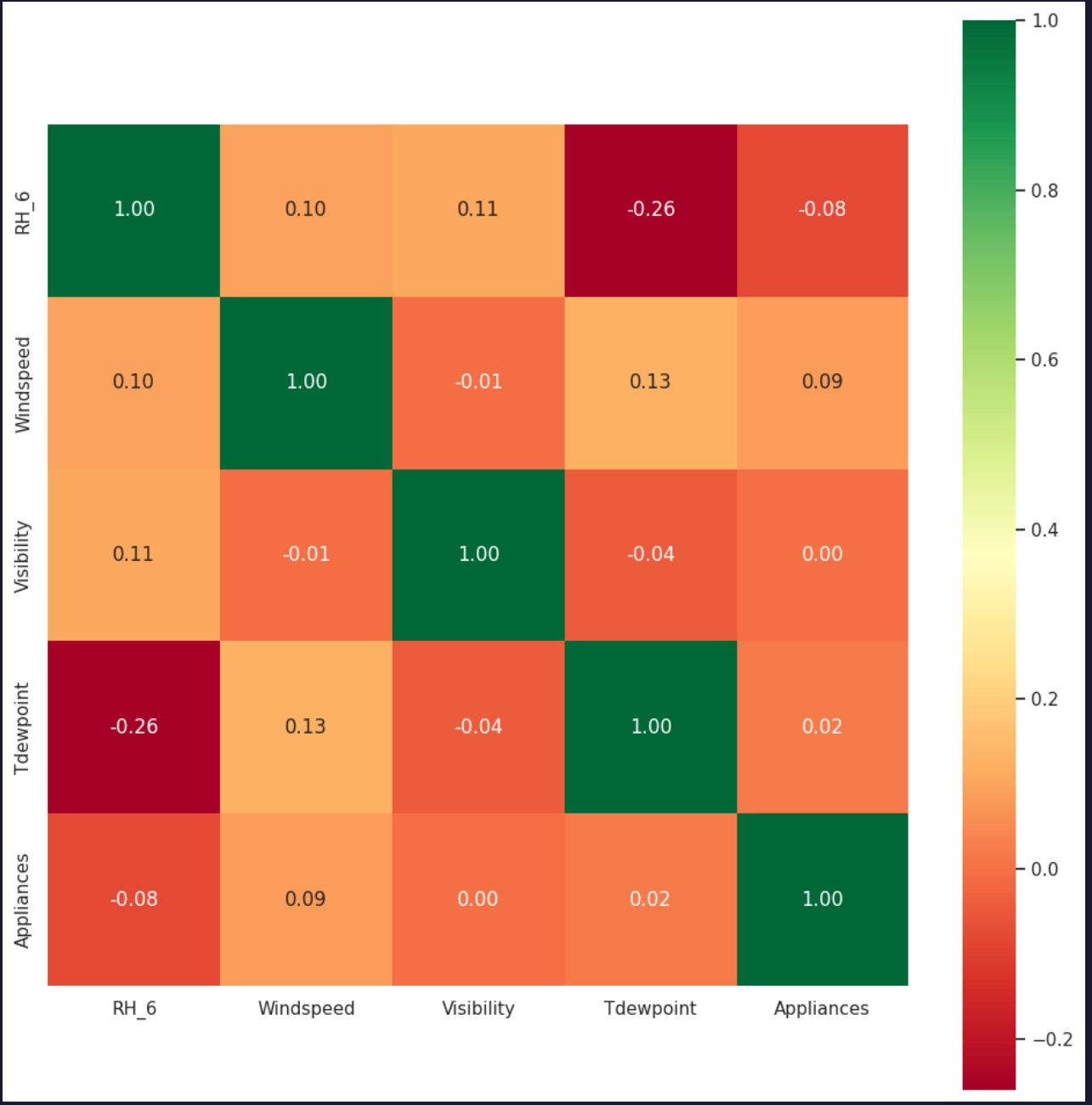
VIF	
RH_6	32.201556153357 96
T_out	24.955611808616 31
Press_mm_hg	694.296480411337 5
Windspeed	5.005980999514861
Visibility	11.771630692843596
Tdewpoint	17.168244694540604
rv1	Infinity
rv2	Infinity
Months	53.957712888827 04
Hour	4.790506544071555 5
Mean_house_temp	627.882678245551 7
Mean_relative_humidity	519.492068056161 2



rv1 and rv2 has infinite Variance Inflation Factor hence we will remove them

	VIF	Multicollinearity
RH_6	3.8924680552202884	← Moderate
Windspeed	3.45875632856162	← Moderate
Visibility	5.159889820397968	← High
Tdewpoint	1.8151746899474195	← Moderate





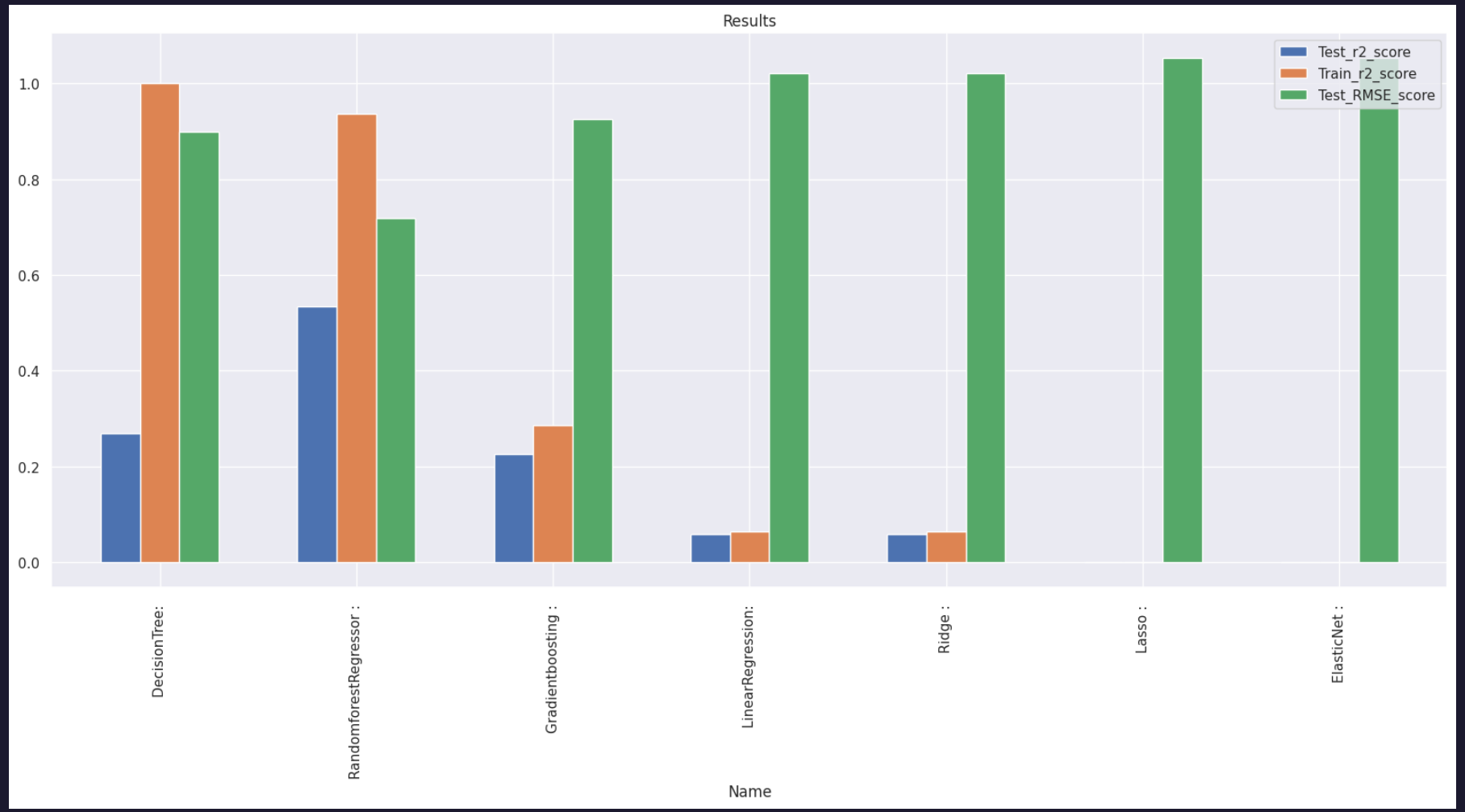
Pre Processing

	count	mean	std	min	25%	50%	75%	max
T_out	19735.0	7.412580	5.318464	-5.000000	3.670000	6.920000	10.400000	26.100000
Press_mm_hg	19735.0	755.522602	7.399441	729.300000	750.933333	756.100000	760.933333	772.300000
RH_6	19735.0	54.609083	31.149806	1.000000	30.025000	55.290000	83.226667	99.900000
Windspeed	19735.0	4.039752	2.451221	0.000000	2.000000	3.666667	5.500000	14.000000
Visibility	19735.0	38.330834	11.794719	1.000000	29.000000	40.000000	40.000000	66.000000
Tdewpoint	19735.0	3.760995	4.195248	-6.600000	0.900000	3.430000	6.570000	15.500000
rv1	19735.0	24.988033	14.496634	0.005322	12.497889	24.897653	37.583769	49.996530
rv2	19735.0	24.988033	14.496634	0.005322	12.497889	24.897653	37.583769	49.996530
months	19735.0	3.101647	1.339200	1.000000	2.000000	3.000000	4.000000	5.000000
hour	19735.0	11.502002	6.921953	0.000000	6.000000	12.000000	17.000000	23.000000
Mean_house_temp	19735.0	20.815611	1.812567	16.012708	19.663000	20.597500	21.764375	26.061940
Mean_relative_humidity	19735.0	39.832333	3.929901	29.264857	36.826714	39.224490	42.698075	51.238571

Our final dataset for model training is split into two one which containing the features for model training(X) and the target variable(y). The two data set are also divided using 20:80 split for training and test the model

For feature scaling we used standardization using standard scalar.

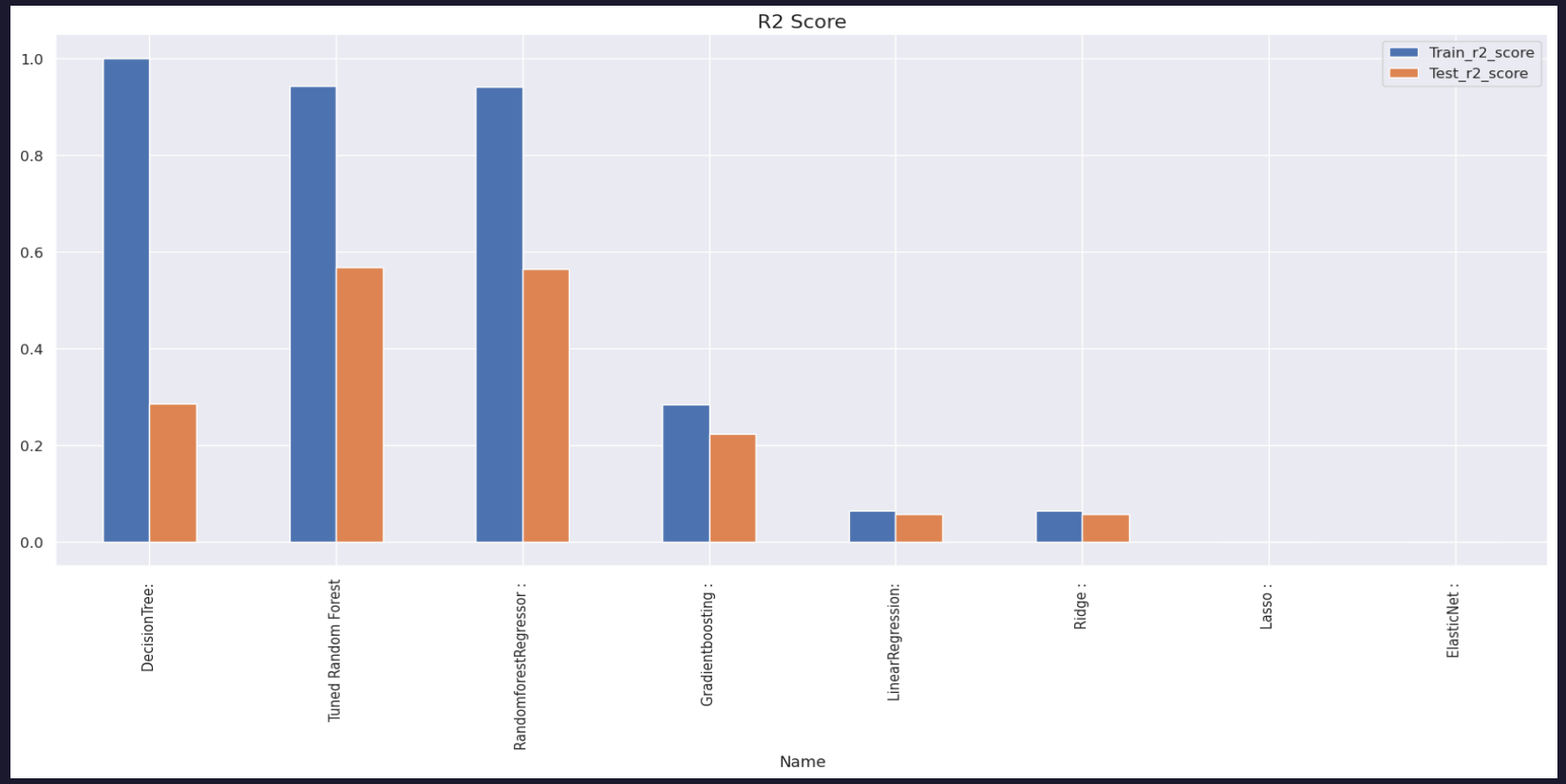
Model Implementation



So after training data Random Forest Regressor has the least RMSE score and a decent R2 score.

Before proceeding to try next models, we try to tune some hyperparameters and see if the performance of our model improves.
Hyperparameter tuning is the process of choosing a set of optimal hyperparameters for a learning algorithm.

Name	Train_r2_score	Test_r2_score	Train_MSE_score	Test_MSE_score	Train_RMSE_score	Test_RMSE_score
DecisionTree:	1.000000	0.286427	4.063988e-34	0.790025	2.015933e-17	0.888833
Tuned Random Forest	0.942690	0.568343	5.730530e-02	0.477904	2.393853e-01	0.691306
RandomforestRegressor :	0.941369	0.565458	5.863053e-02	0.481098	2.421374e-01	0.693612
Gradientboosting :	0.283830	0.224162	7.161700e-01	0.858960	8.462683e-01	0.926801
LinearRegression:	0.063877	0.058172	9.361230e-01	1.042735	9.675345e-01	1.021144
Ridge :	0.063877	0.058173	9.361230e-01	1.042734	9.675345e-01	1.021143
Lasso :	0.000000	-0.000371	1.000000e+00	1.107550	1.000000e+00	1.052402
ElasticNet :	0.000000	-0.000371	1.000000e+00	1.107550	1.000000e+00	1.052402



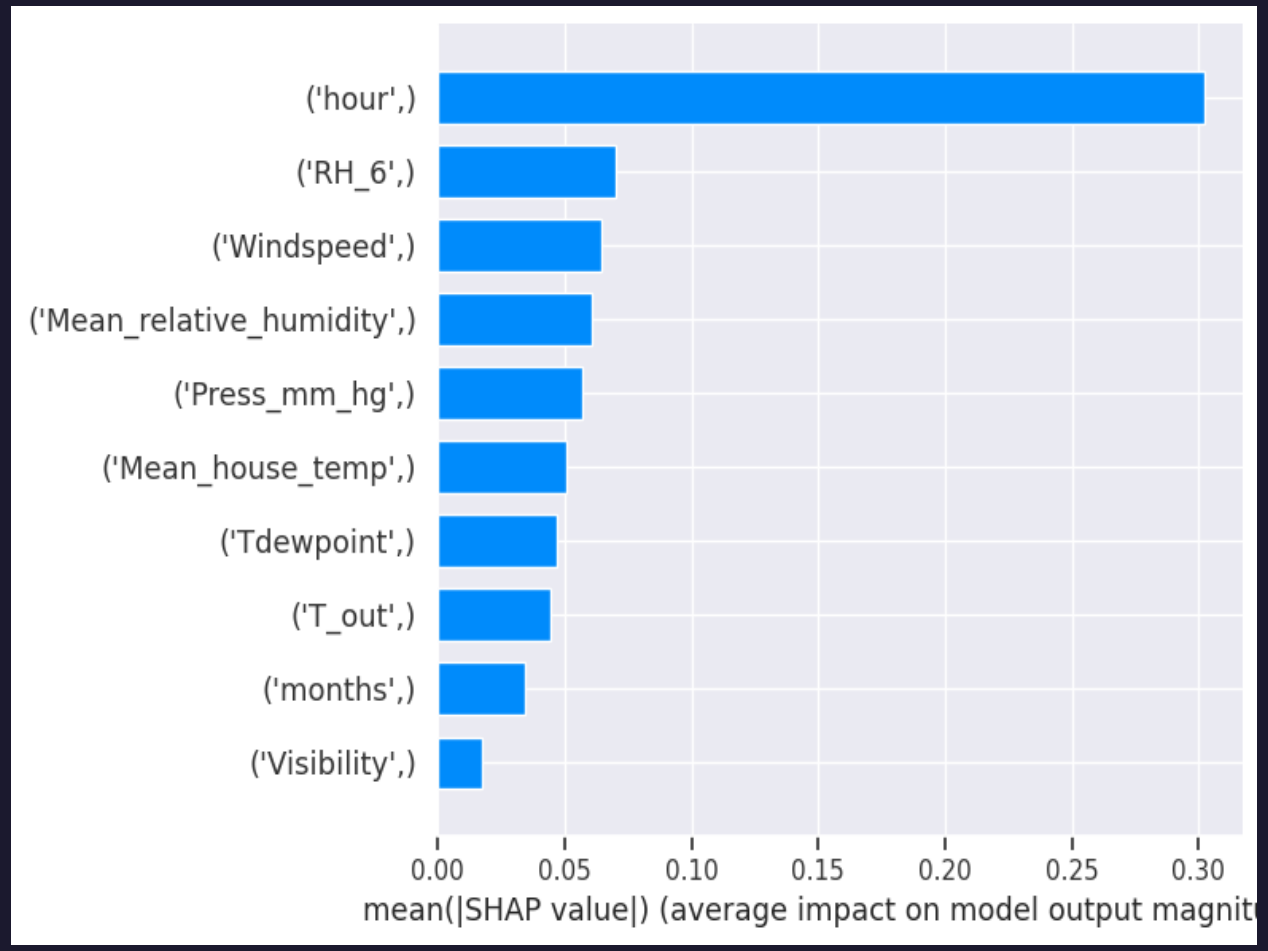
Result

* The Dataset does not contains null values ,but there is very less correlation between features and target variables.
* By fitting all the model get best score in Random Forest regressor , after tuning the hyper parameter using GridsearchCV, GET Train r2 score 0.94 and test r2 score 0.5622 because of improper dataset and less correlation between feature and target variable.
* The performance is low due to like:- no proper pattern of data, less correlation , not enough relevant features.

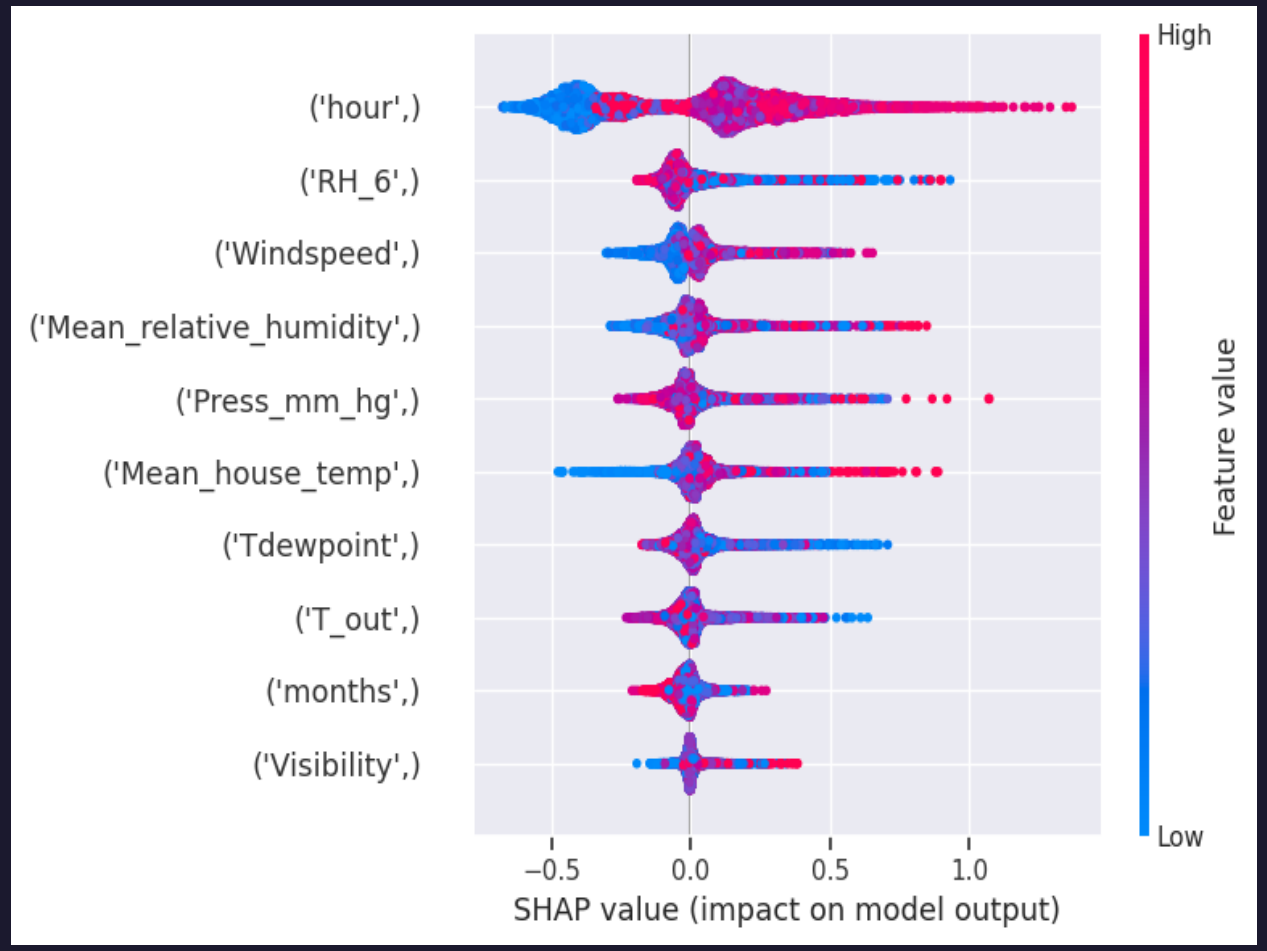
Model Explaination

We used Random Forest Regressor model as it has the most accurate prediction. This SHAP plot gives us the explainability of a single model prediction. Force plot can be used for error analysis, finding the explanation to specific instance prediction. The model is explained as follow:-

- The model output value: -0.51
- The base value: this is the value that would be predicted if we didn't know any features for the current instance. The base value is the average of the model output over the training dataset
- The numbers on the plot arrows are the value of the feature for this instance.
- Red represents features that pushed the model score higher, and blue representing features that pushed the score lower.
- The bigger the arrow, the bigger the impact of the feature on the output.
- The amount of decrease or increase in the impact can be seen on the x-axis.



Bar Summary Plot



Dot Summary Plot

Conclusion

The most important determining factor for energy consumption is the hour of day.

Energy consumption is high in March and low in January, and a rise in temperature results in higher energy consumption.

As a feature, lights are extremely undervalued.

Decreased humidity leads to an increase in electricity consumption. Humidity is proportional to the dependent variable.

We have a high correlation with the dependent variable in the hours column, and many features have less than a 0.1 correlation with the dependent variable in the non linear dataset.

During the evening hours of 16:00 to 20:00, there is a high usage of electricity of more than 140Wh. Electricity use is highest on weekends (Saturdays and Sundays). (more than 25% higher than on weekdays)

The dataset has many outliers and no null values.

Many columns in the dataset are not normally distributed, and the target column is also right skewed.


Thank You