Capstone Project-2

Appliances Energy Prediction

By-Md.ImranHaji

Points to discuss:-

- Problem Statement.
- Data summary.
- Understand the relationships between variables
- Hypothesis Testing
- Feature Engineering and Variable _Selection.
- Model Implementation and evaluation.
- Conclusion.

Problem Statement.

It is very important to meet the supply and demand of electricity, as the energy consumption is increasing in recent years. However, when comes to domestic consumers, electricity consumption of house depends on type and quantity of house hold appliances which in turn influenced by weather conditions.

The main objective for this project is to build a predictive model, which could help in predicting the energy consumption. Along with model building ,will look in some feature behavior.

- 1) what is monthly meter reading ?
- 2) What is the meter reading on hourly basis?
- 3) How is the relationship between temperature and humidity ?
- 4) what is monthly average temperature inside and outside ?
- 5) What is the temperature fluctuation in each room ?
- 6) What is daily wind speed and visibility ?
- 7) How was visibility with outside temperature?
- 8) is there any difference in meter reading when compared to inside and outside temperature?

Data summary.

Data-driven prediction of energy use of appliances The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. Each wireless node transmitted the temperature and humidity conditions around 3.3 min. Then, the wireless data was averaged for 10 minutes periods. The energy data was logged every 10 minutes with m-bus energy meters. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru) and merged together with the experimental data sets using the date and time column.
Data Description date time year-month-day hour:minute:second
date - time year-month-day hour:minute:second

- Appliances - energy use in Wh (Dependent variable)
- lights - energy use of light fixtures in the house in Wh (Drop this column)
- T1 - Temperature in kitchen area, in Celsius
- RH1 - Humidity in kitchen area, in %
- T2 - Temperature in living room area, in Celsius
- RH2 - Humidity in living room area, in %
- T3 - Temperature in laundry room area
- RH3 - Humidity in laundry room area, in %
- T4 - Temperature in office room, in Celsius
- RH4 - Humidity in office room, in %

- T5 - Temperature in bathroom, in Celsius
- RH5 - Humidity in bathroom, in %
- T6 - Temperature outside the building (north side), in Celsius
- RH6 - Humidity outside the building (north side), in %
- T7 - Temperature in ironing room, in Celsius
- RH7 - Humidity in ironing room, in %
- T8 - Temperature in teenager room 2, in Celsius
- RH8 - Humidity in teenager room 2, in %
- T9 - Temperature in parents room, in Celsius
- RH9 - Humidity in parents room, in %
- T_out - Temperature outside (from Chievres weather station), in Celsius
- Pressure - (from Chievres weather station), in mm 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
- rv1 - Random variable 1, nondimensional
- rv2 - Random variable 2, nondimensional

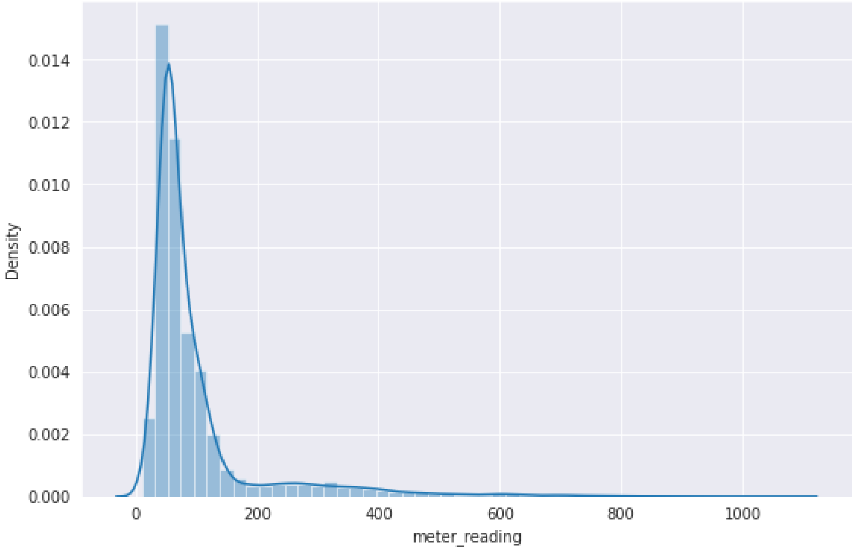
Where indicated, hourly data (then interpolated) from the nearest airport weather station(Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis,rp5.ru. Permission was obtained from Reliable Prognosis for the distribution of the 4.5 months of weather data.

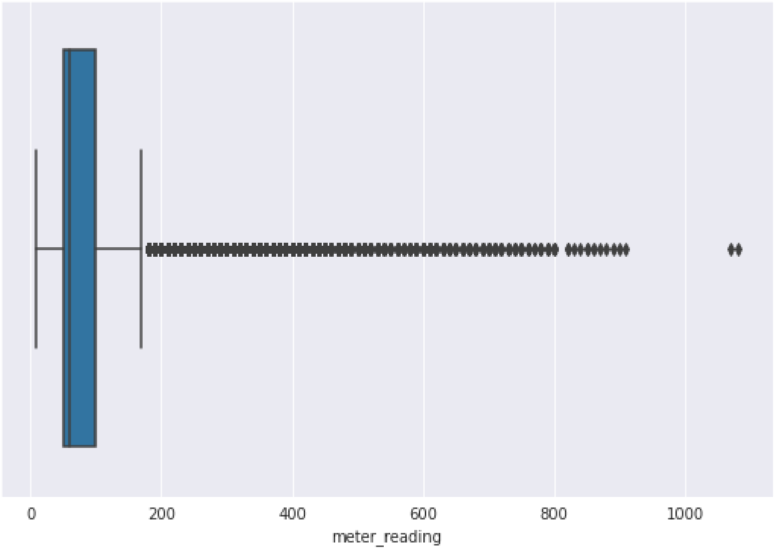
Understand the relationships between variables

Here our dependent variable is Appliances - energy use in Wh

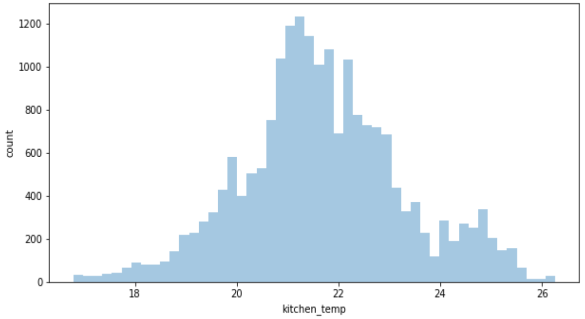
- Meter reading is of ranges from 10 Wh to 1080 Wh

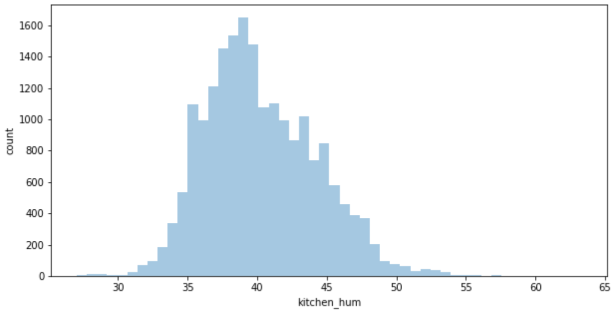
- Our data is right skewed and About 75 % of values lie below 100 Wh, and About.

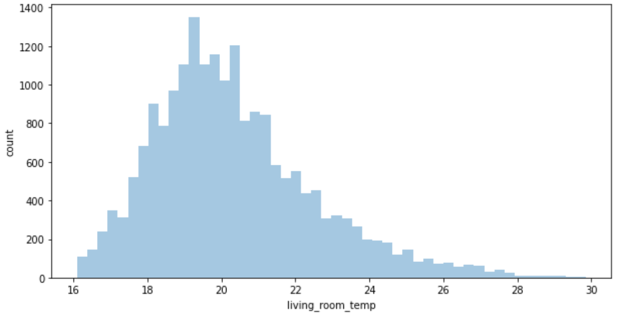


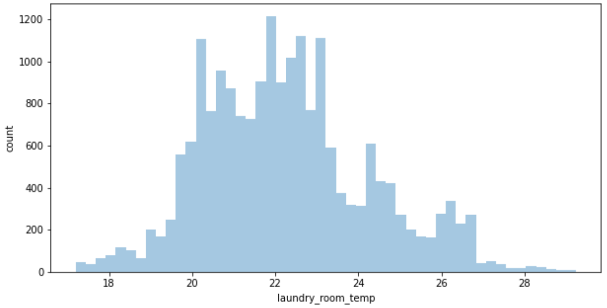


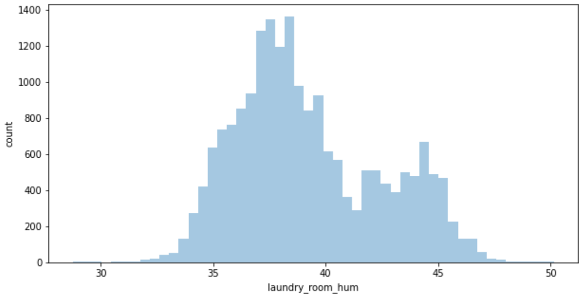
Univariant Analysis:

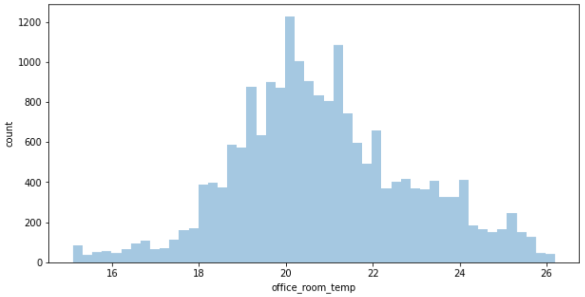


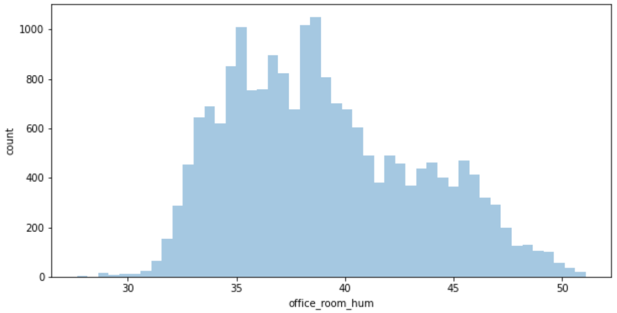


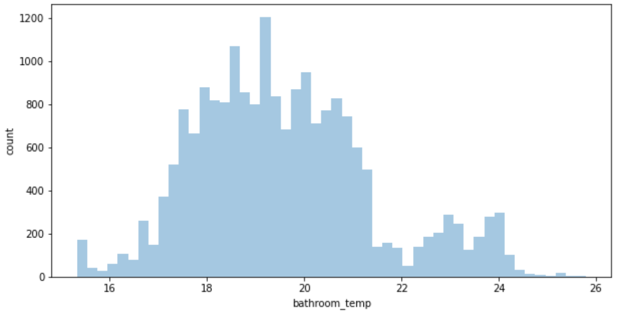


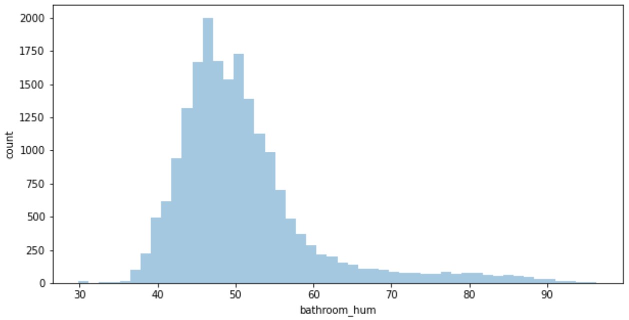


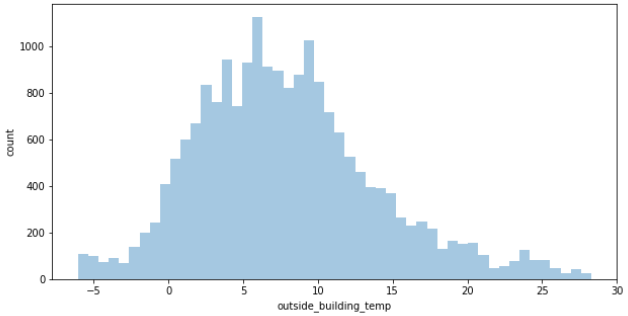


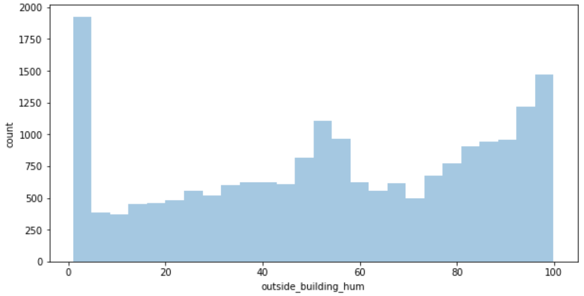


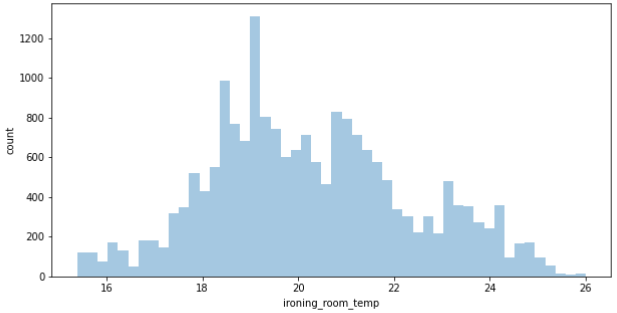


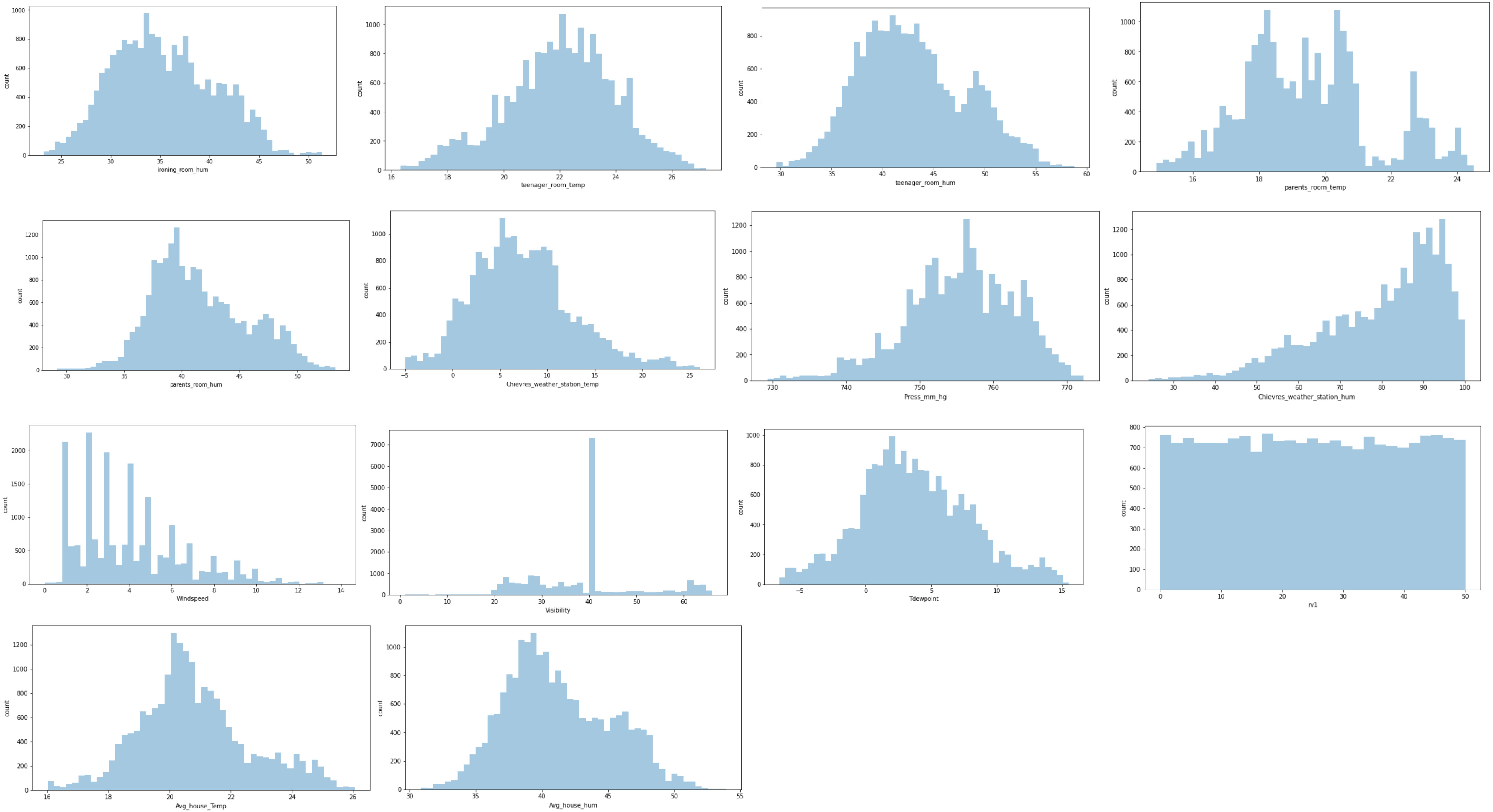











count
ironing_room_hum
25 30 35 40 45 50
0 200 400 600 800 1000
count
teenager_room_temp
16 18 20 22 24 26
0 200 400 600 800 1000
count
teenager_room_hum
30 35 40 45 50 55 60
0 200 400 600 800
count
parents_room_temp
16 18 20 22 24
0 200 400 600 800 1000
count
parents_room_hum
30 35 40 45 50
0 200 400 600 800 1000 1200
count
Chievres_weather_station_temp
−5 0 5 10 15 20 25
0 200 400 600 800 1000
count
Press_mm_hg
730 740 750 760 770
0 200 400 600 800 1000 1200
count
Chievres_weather_station_hum
30 40 50 60 70 80 90 100
0 200 400 600 800 1000 1200
count
Windspeed
0 2 4 6 8 10 12 14
0 500 1000 1500 2000
count
Visibility
0 10 20 30 40 50 60
0 1000 2000 3000 4000 5000 6000 7000
count
Tdewpoint
−5 0 5 10 15
0 200 400 600 800 1000
count
rv1
0 10 20 30 40 50
0 100 200 300 400 500 600 700 800
count
Avg_house_Temp
16 18 20 22 24 26
0 200 400 600 800 1000 1200
count
Avg_house_hum
30 35 40 45 50 55
0 200 400 600 800 1000

Observation:

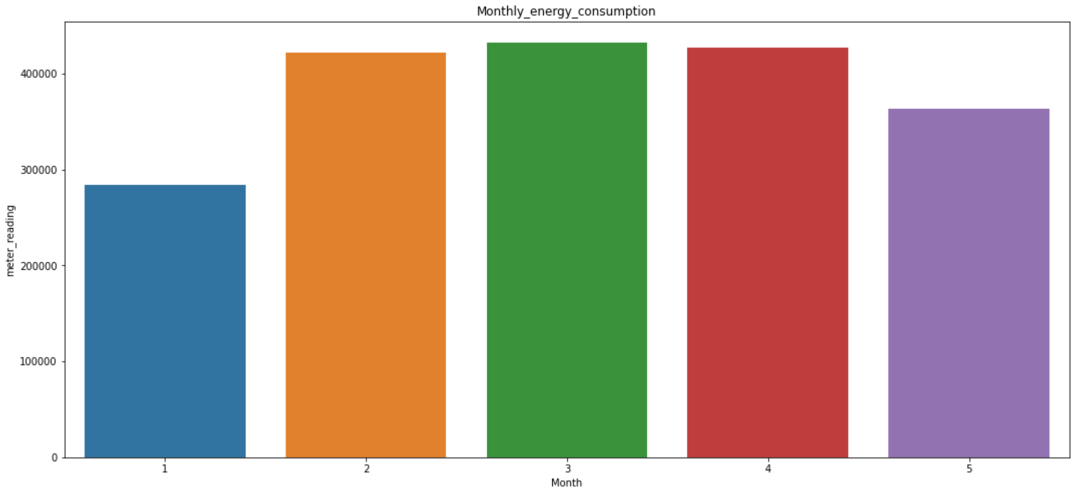
1) Meter reading is right skewed ,with highest count for the meter reading range between 50 to 60.

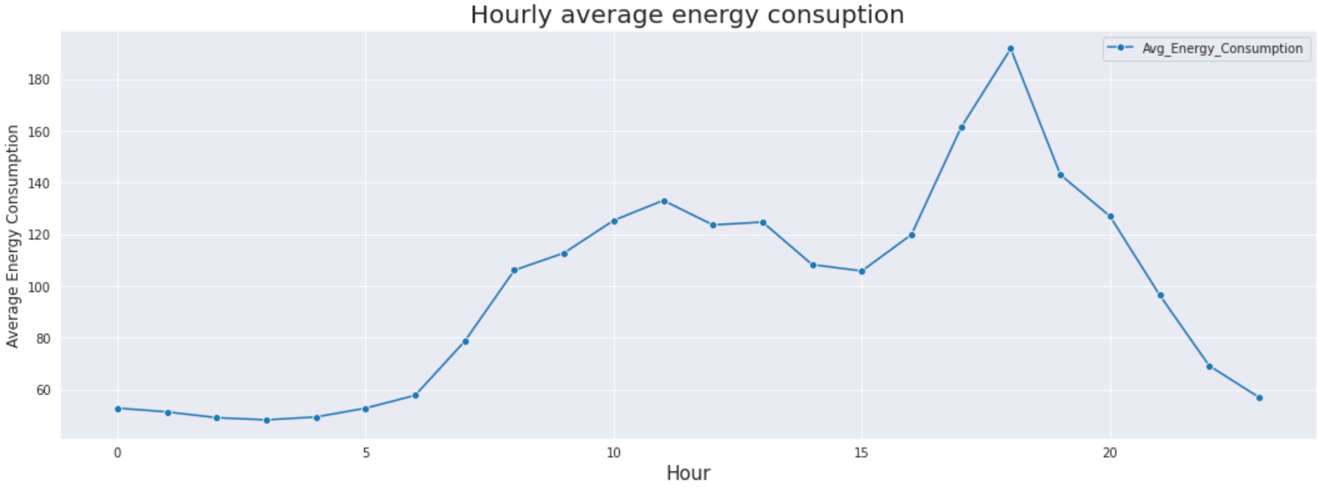
2)Kitchen temperture,living room temperature,laundry_room_temp,office_room_temp, bathroom_temp,ironing_room_temp ,teenager_room_temp,parents_room_temp, and Avg_house_Temp allmost every room had normal distribution with fluctuating range of temperature is between 18 to 24 degree Celsius. **outside_building_temp ,Chievres_weather_station_temp range between -5 to 28**.

3)kitchen_hum,living_room_hum,laundry_room_hum,office_room_hum,roning_room_hum,teenager_room_hum, parents_room_hum and Avg_house_hum had almost a normal distributed plot range between 30 to 55,where most of the count was beween 38 to 42 . 'bathroom_hum **(Right Skewwed).outside_building_hum had a range of humidity from 0 to 100 , with max count between 0 to 5, Chievres_weather_station_hum(left skewwed most of the time between 80 to 100).**

4) Visibility of scale 40 Km was highest.

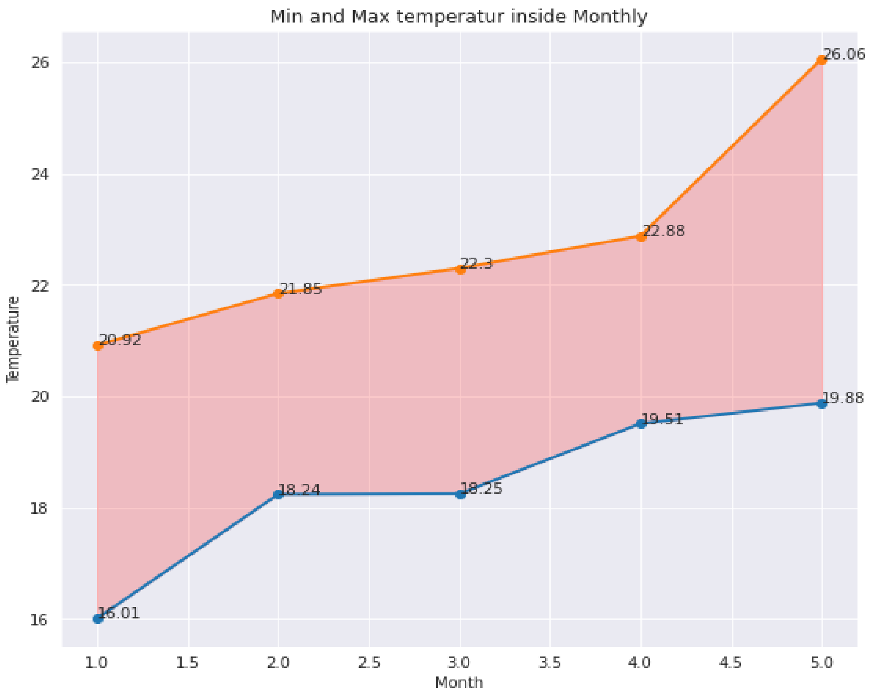
Bivariant analysis

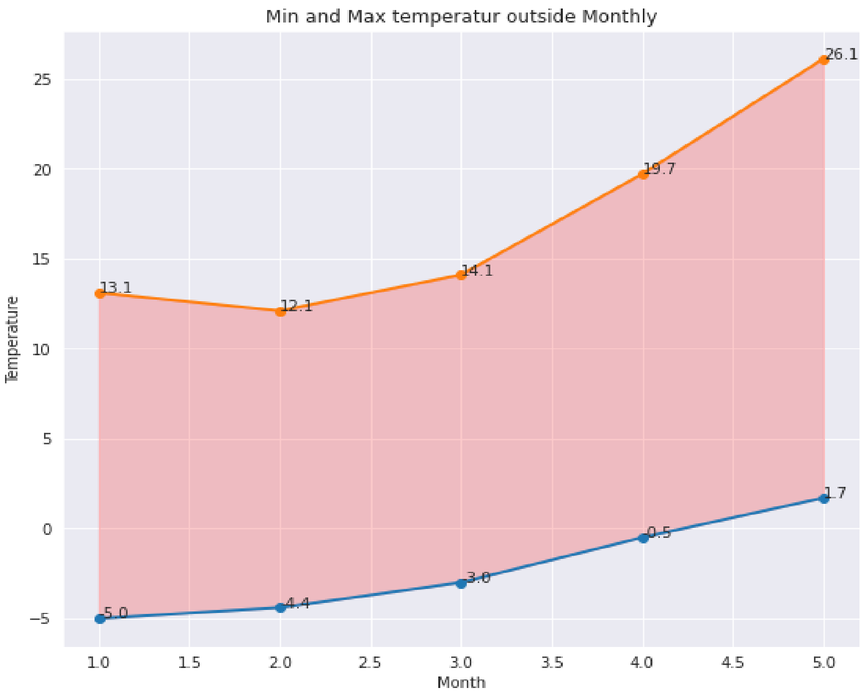




- Month 3 following 4 and 2 had highest meter reading.(data for month 1 is only partial)

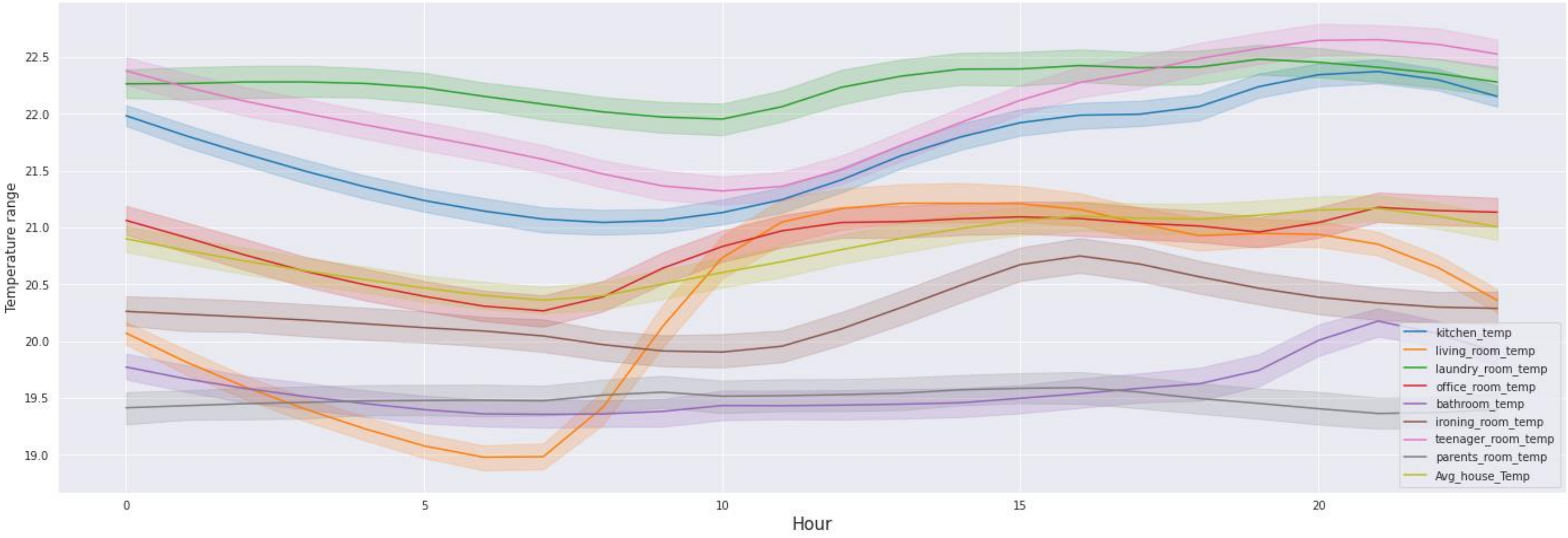
- The average consumption of energy from mid night to early morning is very low, as appliances are used less during the night.
- The consumption of energy from morning till evening hours (6 to 16) is moderate. Where as in the evening(17 to 20), energy consumption is high. This is because appliances are used more during the evening, an was highest at 18

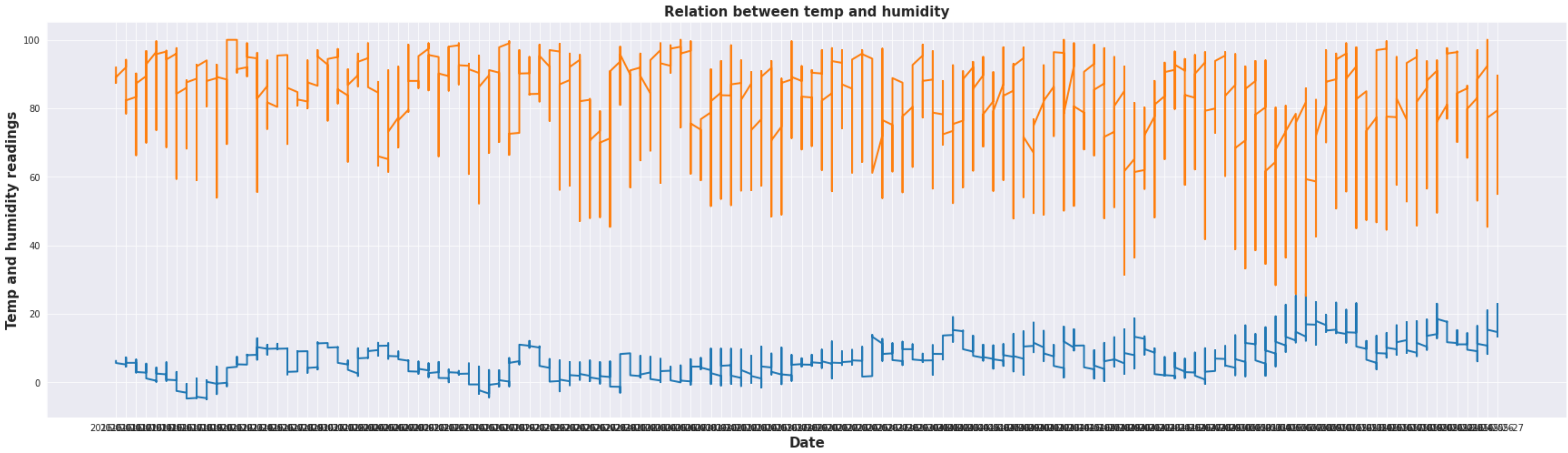




- Average temperature inside was maintained at optimum level.
- Month Jan,feb and march where cool ,January had the low temperature of -5 degree where as may had the highest temperature of 26.1 degrees

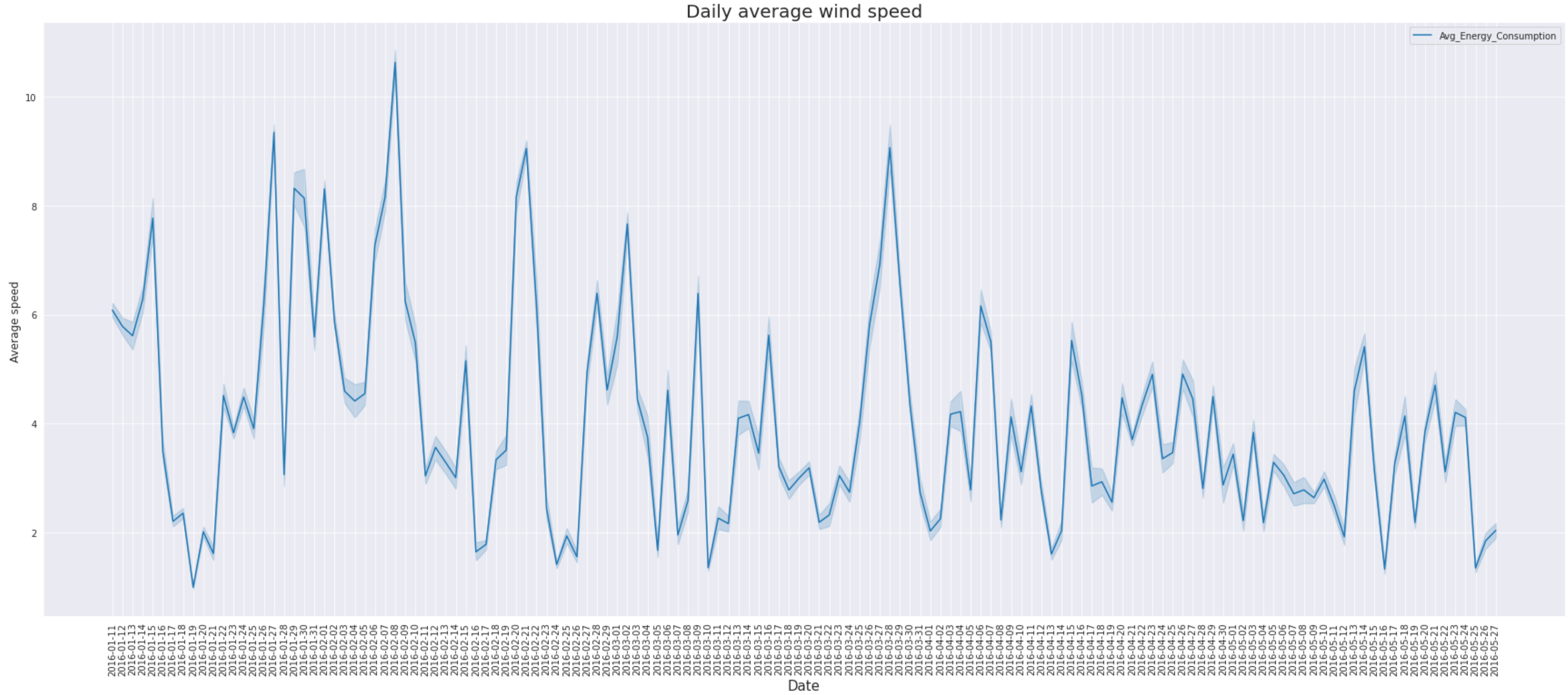
- only living room had some fluctuations other than that almost every room had maintained constant temperature daily.
- parents room temperature was cool, followed by office room and laundry was warmer

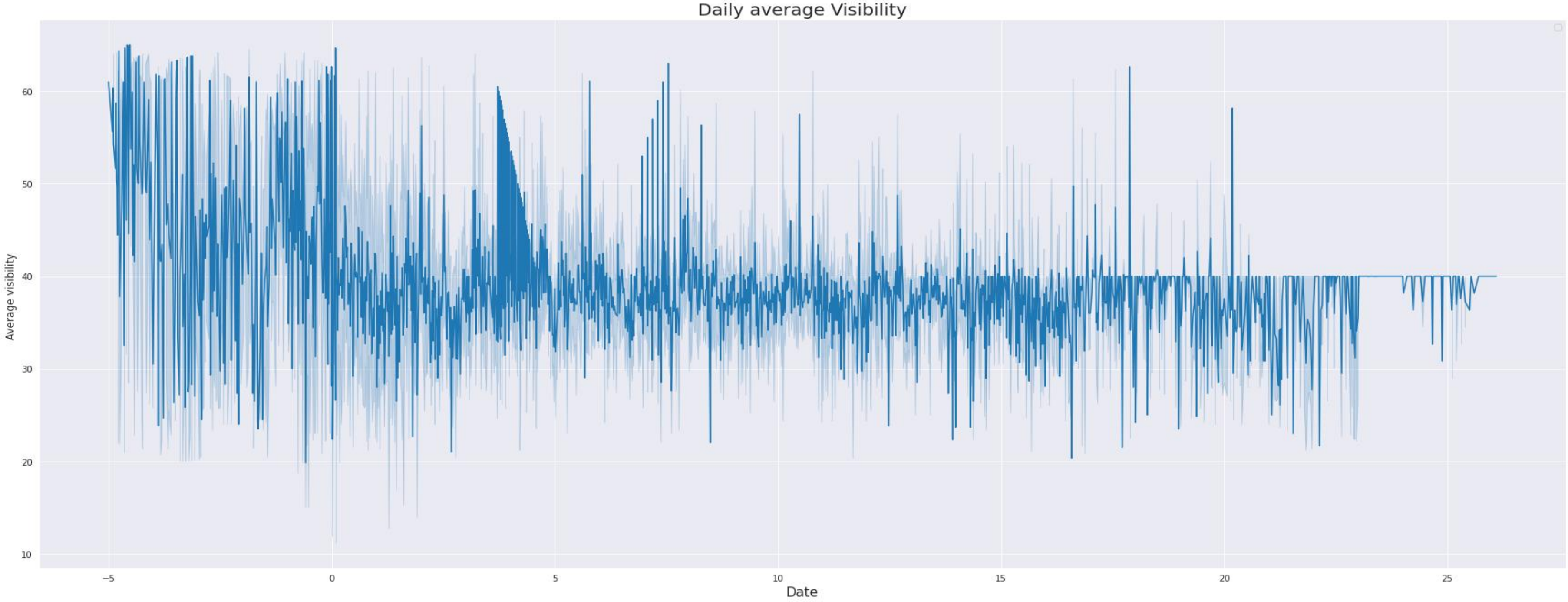




- Temperature and humidity have a inverse relationship.

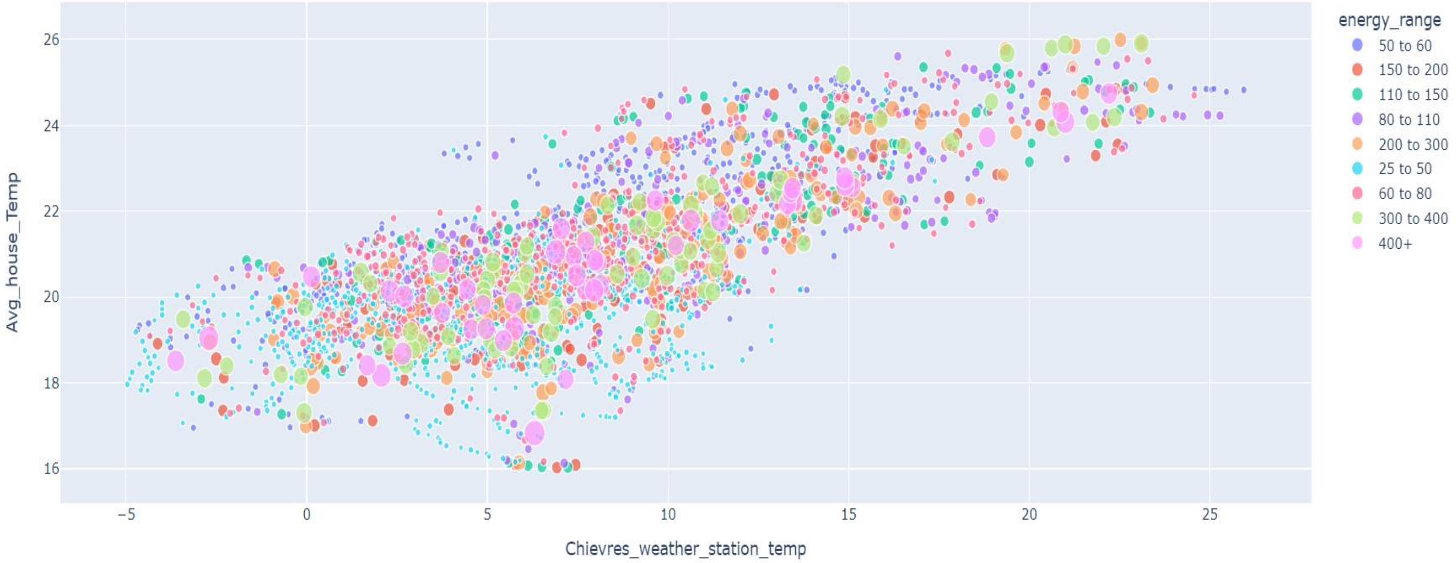
- Windspeed is between 0 to 14 and mostly 0 m/s to 5.5 m/s range fall below 75%.





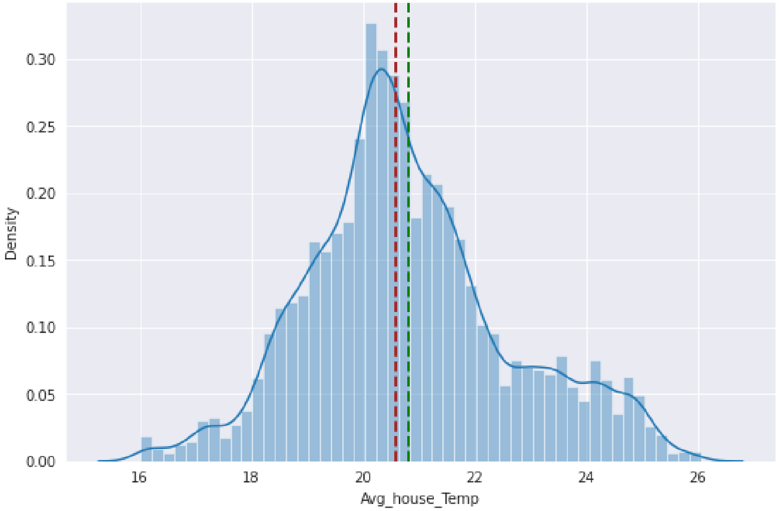
- Most of the time even if the temperature is recording low, visibility was good which is between 20 to 65 KM.

- Even though the out side temperature is low and inside temperature was maintained optimum ,still data is not really clustered ,energy consumption had recorded from low to high range on daily bases.
- Range of 50 to 60 had high count of 799 following 25 to 50 with 706
- Range of 400+ had lowest count of 44



Hypothesis Testing

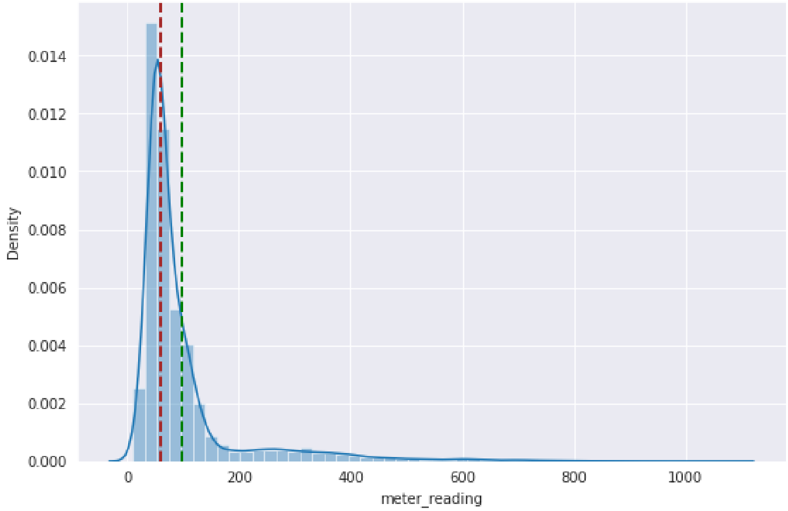
•Null hypothesis = Average temperature inside is almost equal to 21 degree Celsius
•Alternate = Average temperature is not the given values



Since, the plot a normally distributed and mean , median is very close, we can perform z test.

Failed to reject the Null Hypothesis for p = 0.4519667342376532.

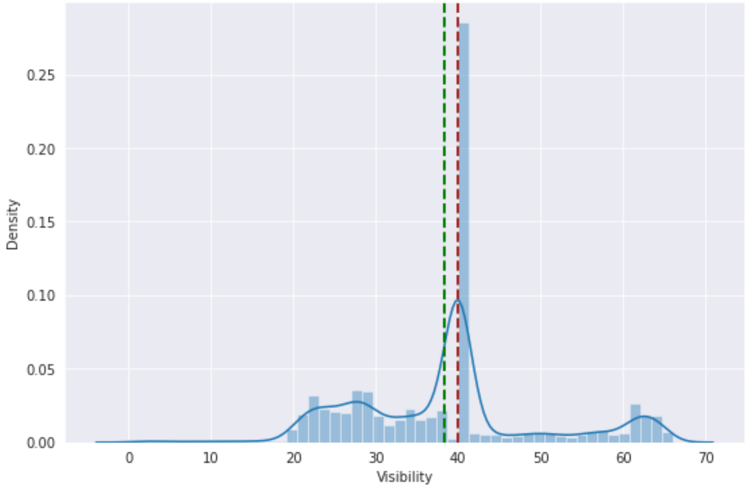
- Null Hypothesis : mean meter reading daily is almost = 97
- Alternate Hypothesis : mean meter reading daily is != 97



The distribution is positively skewed. For a skewed data Z-Test can't be performed.Therefor, for heavily skewed data t-tests can be used even we can take large sample size.

Failed to reject the Null Hypothesis for p = [0.76635109]

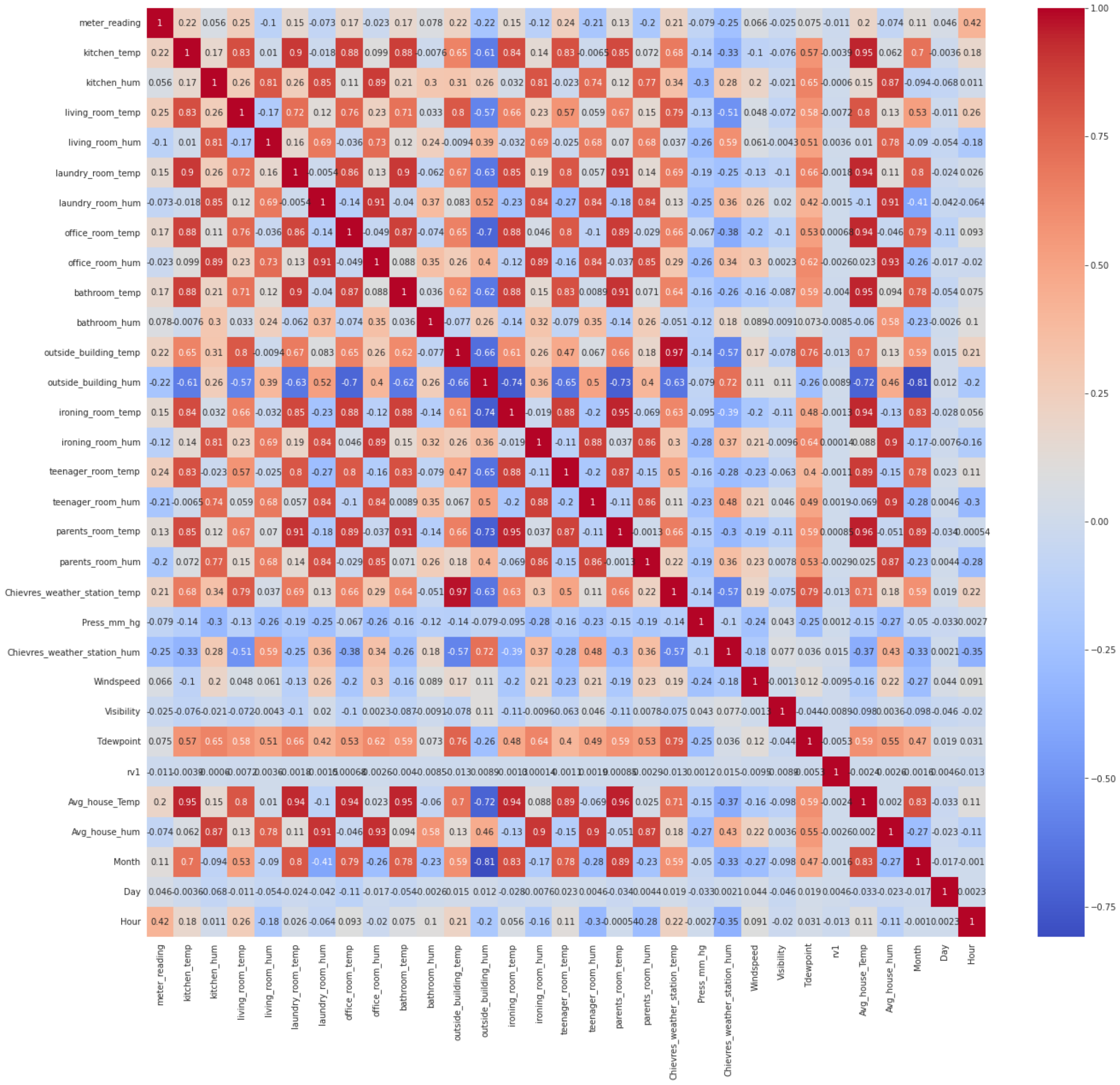
- Null Hypothesis : is visibility = 20
- Alternate Hypothesis : visibility > 20



Since the plot is not really a normally distributed data we will again use t-test.

Null Hypothesis rejected Successfully for p = [1.44731184e-12]

Feature Engineering & Variable _Selection



- The Temperature and humidity levels in each of the rooms are highly correlated.
- However temperature and humidity levels outside the building are negatively correlated.
- There is little correlation between these features and the target variable i.e. Appliance energy consumption

After doing VIF ,below given variables are used for modelling

variables	VIF
outside_building_temp	8.71
Windspeed	3.25
Tdewpoint	4.52
rv1	3.42
Month	6.94
Day	2.95
Hour	3.37

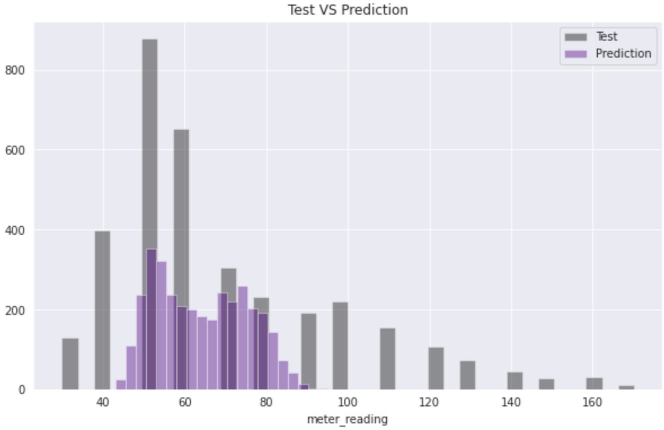
Model Implementation

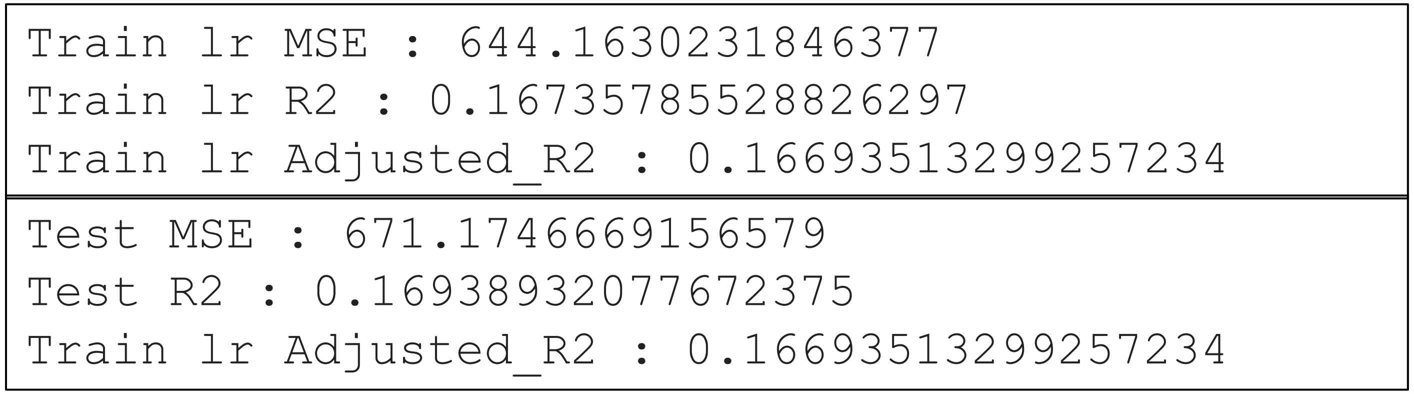
- Since our project is regression, I used only **linear regression** model with regularization
- Lasso and Ridge regression
- Hyper parameter tuning
  - Grid Search CV

Model Evaluation:

- MSE (Mean Squared Error)
- R2 (R — Squared)
- Adjusted R2

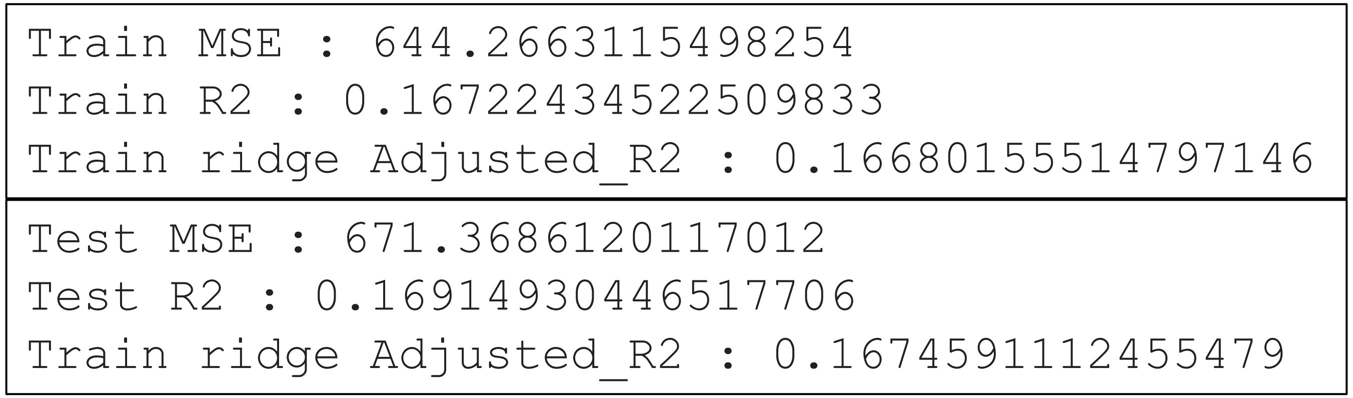
linear regression model with regularization





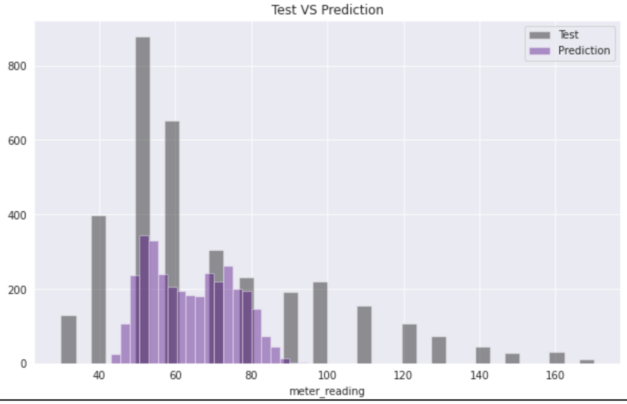
```
Train lr MSE : 644.1630231846377
Train lr R2 : 0.16735785528826297
Train lr Adjusted_R2 : 0.16693513299257234
```

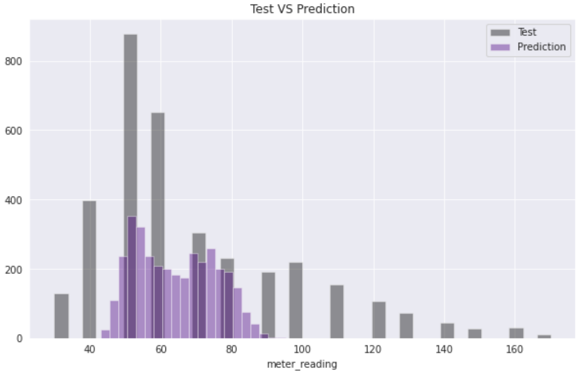
```
Test MSE : 671.1746669156579
Test R2 : 0.16938932077672375
Train lr Adjusted_R2 : 0.16693513299257234
```

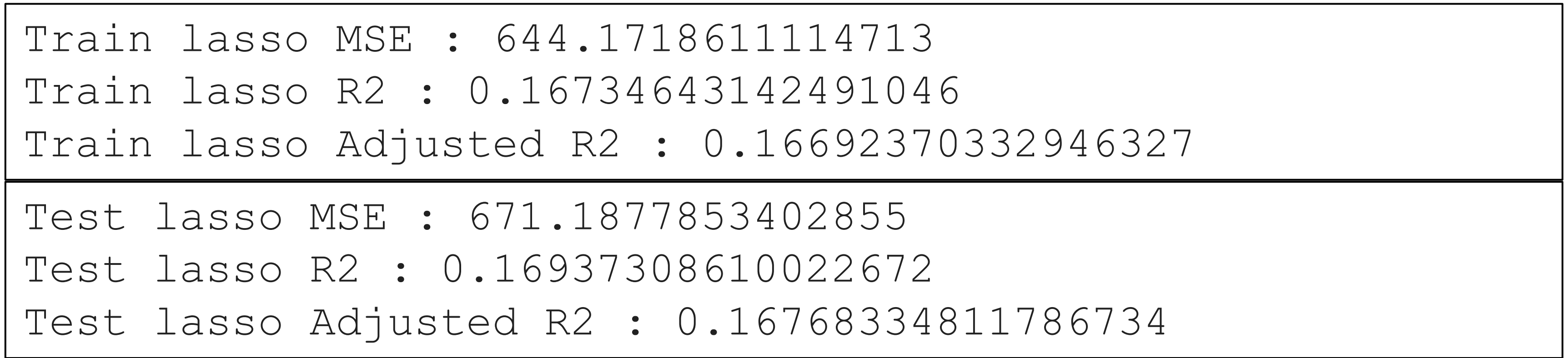


```
Train MSE : 644.2663115498254
Train R2 : 0.16722434522509833
Train ridge Adjusted_R2 : 0.16680155514797146
```

```
Test MSE : 671.3686120117012
Test R2 : 0.16914930446517706
Train ridge Adjusted_R2 : 0.1674591112455479
```



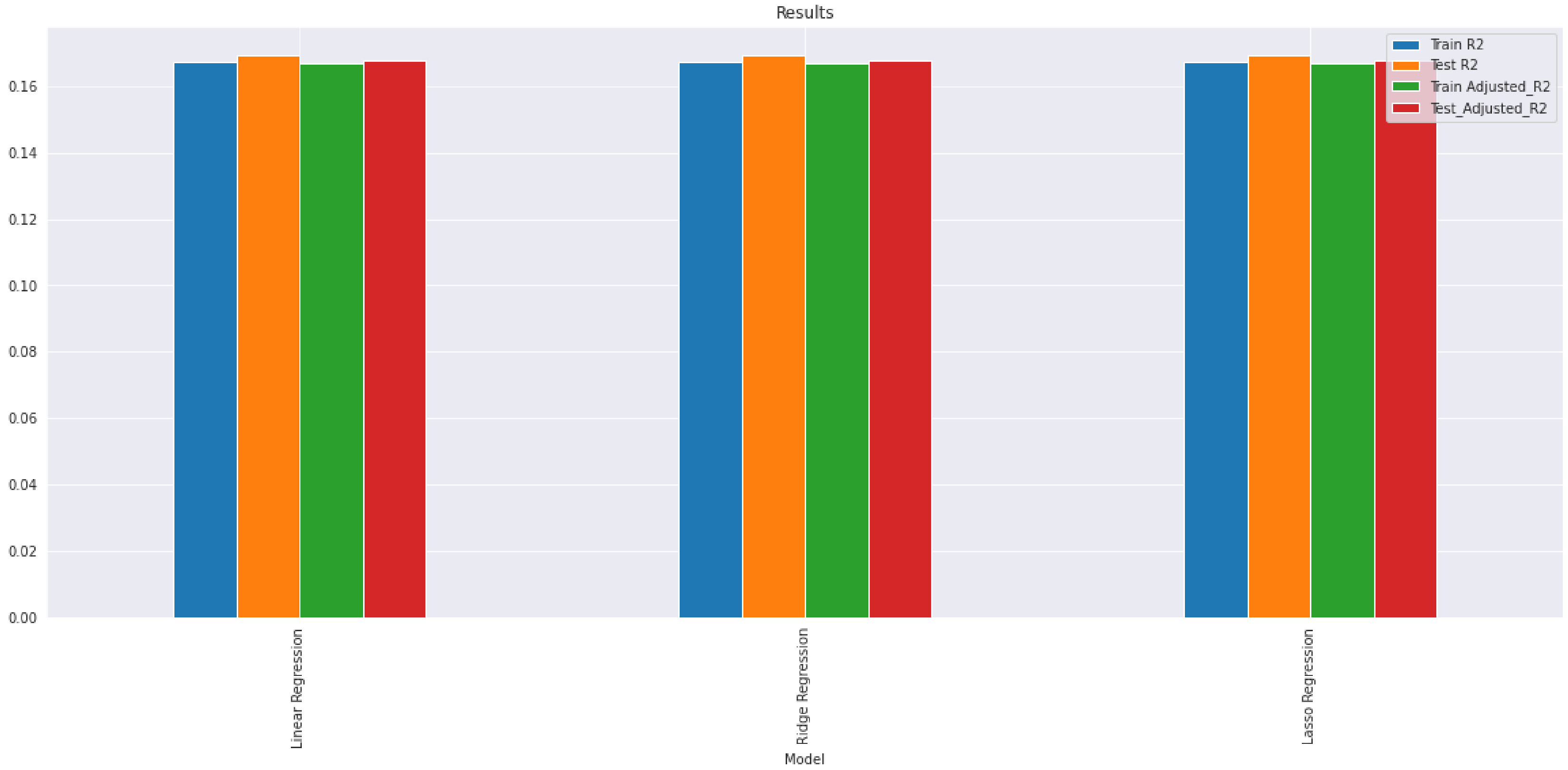




```
Train lasso MSE : 644.1718611114713
Train lasso R2 : 0.16734643142491046
Train lasso Adjusted R2 : 0.16692370332946327
```

```
Test lasso MSE : 671.1877853402855
Test lasso R2 : 0.16937308610022672
Test lasso Adjusted R2 : 0.16768334811786734
```

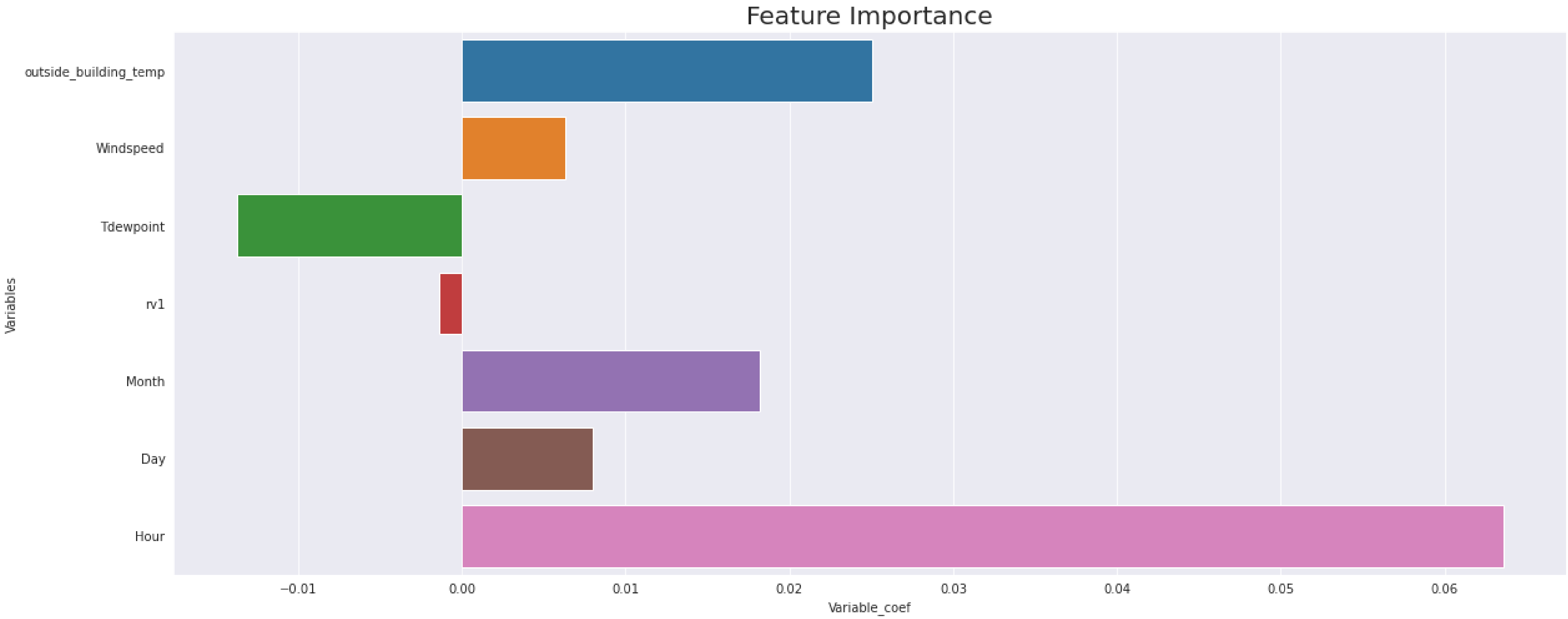
Model Evaluation comparison:



Model	Train MSE	Train R2	Train Adjusted_R2	Test MSE	Test R2	Test_Adjusted_R2
Linear Regression	644.163023	0.167358	0.166935	671.174667	0.169389	0.167700
Ridge Regression	644.266312	0.167224	0.166802	671.368612	0.169149	0.167459
Lasso Regression	644.171861	0.167346	0.166924	671.187785	0.169373	0.167683

Best estimator:

```
lasso_regression.best_estimator_.coef_
```



Conclusion.

- Dependent variable i.e. Appliance energy consumption has little correlation between features and of all the variable only outside_building_temp, Windspeed, Tdewpoint, rv1, Month, Day, Hour are used for model prediction after feature selection process is done.
- Hour variable has large positive correlation in predicting dependent variable .
- Even after Hyper parameter tuning on regularization , our model had almost same results in all regularization.
- Model was good between train and test as results for test and train were almost same.
- In all of these models our r2 score was almost 0.169

References :-

1. mygreatlearning.com
2. GeeksforGeeks
3. Analytics Vidhya
4. Almabetter Notes.