CAPSTONE PROJECT- 2

BUILDING MACHINE
LEARNING MODEL
ON

APPLIANCE ENERGY PREDICTION

MD KHALID ANSARI
ALMABETTER TRANIEE

AGENDA

SAVE ENERGY, SAVE ENVIRONMENT, SUSTAINABLE DEVELOPMENT

Table of Content

WHAT, WHY AND HOW?

What – Our goal is to explore the energy consumption in a society building based on appliance energy use given the dataset.

Why – 80% of power supply demand is fulfilled using non-renewable energy sources which are depleting fast. Such data explorations can help minimizing the energy waste and can also be helpful while developing new energy source dependency in future.

How – Given a data set, first step will be descriptive analysis that will tells the past and present energy use, and then predictive analysis where we'll be developing the ML model out of best algorithms available.

Problem Statement

1. Analyzing the affect of weather condition on the power supply.
2. Interpreting the energy required to balance the interior temperature and humidity as per human comfort.
3. How the power supply differs in different phase of time(weeks, days, months)?

Procedure

- **Data Background**– What is the data about?
- **Data Extraction and Cleansing** – Selecting the most essential features out of the bulk of data.
- **Data Exploration-** Doing Univariate and Bivariate analysis to understand the flow of data and how it affect the result.
- **Predictive Modelling** - Selecting the best machine learning algorithm that can predict the output for unseen data set.
- **Data Visualization** – Storytelling with the pictorial representation of the raw data for taking future decisions

Data Background

```
Columns

Index(['DateTime', 'Appliances(Wh)', 'Lights(Wh)', 'Kitchen_temp',
       'Kitchen_humidity', 'Liv_room_temp', 'Liv_room_humidity',
       'Laundary_room_temp.', 'Laundary_room_humidity', 'Office_room_temp',
       'Office_room_humd', 'bathroom_temp', 'bathroom_humd',
       'Outside_build_temp', 'Outside_build_humd', 'iron_room_temp',
       'iron_room_humd', 'teen_room_temp', 'teen_room_humd',
       'parent_room_temp', 'parent_room_humd', 'T_out', 'Press_mm_hg',
       'RH_out', 'Windspeed', 'Visibility', 'Tdewpoint', 'rv1', 'rv2'],
      dtype='object')
```

- A wireless sensor network ZigBee measures temperature and humidity around every 3.3 min. and is averaged over 10 minutes. The energy data was logged every 10 min. with m-bus energy meters and is collected for next 4.5 months which the dataset consists of.

- **Important Columns with value range:**

  Appliance energy: 10 to 1080 Wh
  Lights energy: 0 to 70 Wh
  Indoor temp.: 15 to 29 degrees Celsius
  Outdoor temp.: -5 to 28 degrees Celsius
  Indoor humidity: 20% to 96%
  Outdoor humidity: 1% to 99.9%
  Pressure: 729 to 772 mm of HG

DATA EXTRACTION AND CLEANSING

Steps Involved:

a) Null value treatment

b) Outlier Treatment

c) Column extraction with Heatmap Correlation map

d) Extracting month, weekday, time and date from DateTime

e) Dropped highly correlated independent columns

```
Final Extracted Columns

['Appliances(Wh)',
 'Kitchen_temp',
 'Kitchen_humidity',
 'Liv_room_temp',
 'Laundary_room_temp.',
 'Office_room_temp',
 'Outside_build_temp',
 'T_out',
 'Windspeed',
 'hour',
 'Liv_room_humidity',
 'Outside_build_humd',
 'iron_room_humd',
 'teen_room_humd',
 'parent_room_humd',
 'RH_out']
```

DATAFRAME OVERVIEW (TOP 3 ROWS)

```
             DateTime  Appliances(Wh)  Lights(Wh)  Kitchen_temp  \
0 2016-01-11 17:00:00              60          30         19.89
1 2016-01-11 17:10:00              60          30         19.89
2 2016-01-11 17:20:00              50          30         19.89

  Kitchen_humidity  Liv_room_temp  Liv_room_humidity  Laundary_room_temp.  \
0        47.596667           19.2          44.790000                19.79
1        46.693333           19.2          44.722500                19.79
2        46.300000           19.2          44.626667                19.79

  Laundary_room_humidity  Office_room_temp  ...  parent_room_temp  \
0              44.730000         19.000000  ...         17.033333
1              44.790000         19.000000  ...         17.066667
2              44.933333         18.926667  ...         17.000000

  parent_room_humd     T_out  Press_mm_hg  RH_out  Windspeed  Visibility  \
0            45.53  6.600000        733.5    92.0   7.000000   63.000000
1            45.56  6.483333        733.6    92.0   6.666667   59.166667
2            45.50  6.366667        733.7    92.0   6.333333   55.333333

  Tdewpoint        rv1        rv2
0       5.3  13.275433  13.275433
1       5.2  18.606195  18.606195
2       5.1  28.642668  28.642668

[3 rows x 29 columns]
```

A) NULL VALUE TREATMENT – NO NULL VALUE DETECTED

```
df_final2.isnull().sum()
```

```
Appliances(Wh)        0
Kitchen_temp          0
Kitchen_humidity      0
Liv_room_temp         0
Laundary_room_temp.   0
Office_room_temp      0
Outside_build_temp    0
T_out                 0
Windspeed             0
hour                  0
Liv_room_humidity     0
Outside_build_humd    0
iron_room_humd        0
teen_room_humd        0
parent_room_humd      0
RH_out                0
dtype: int64
```

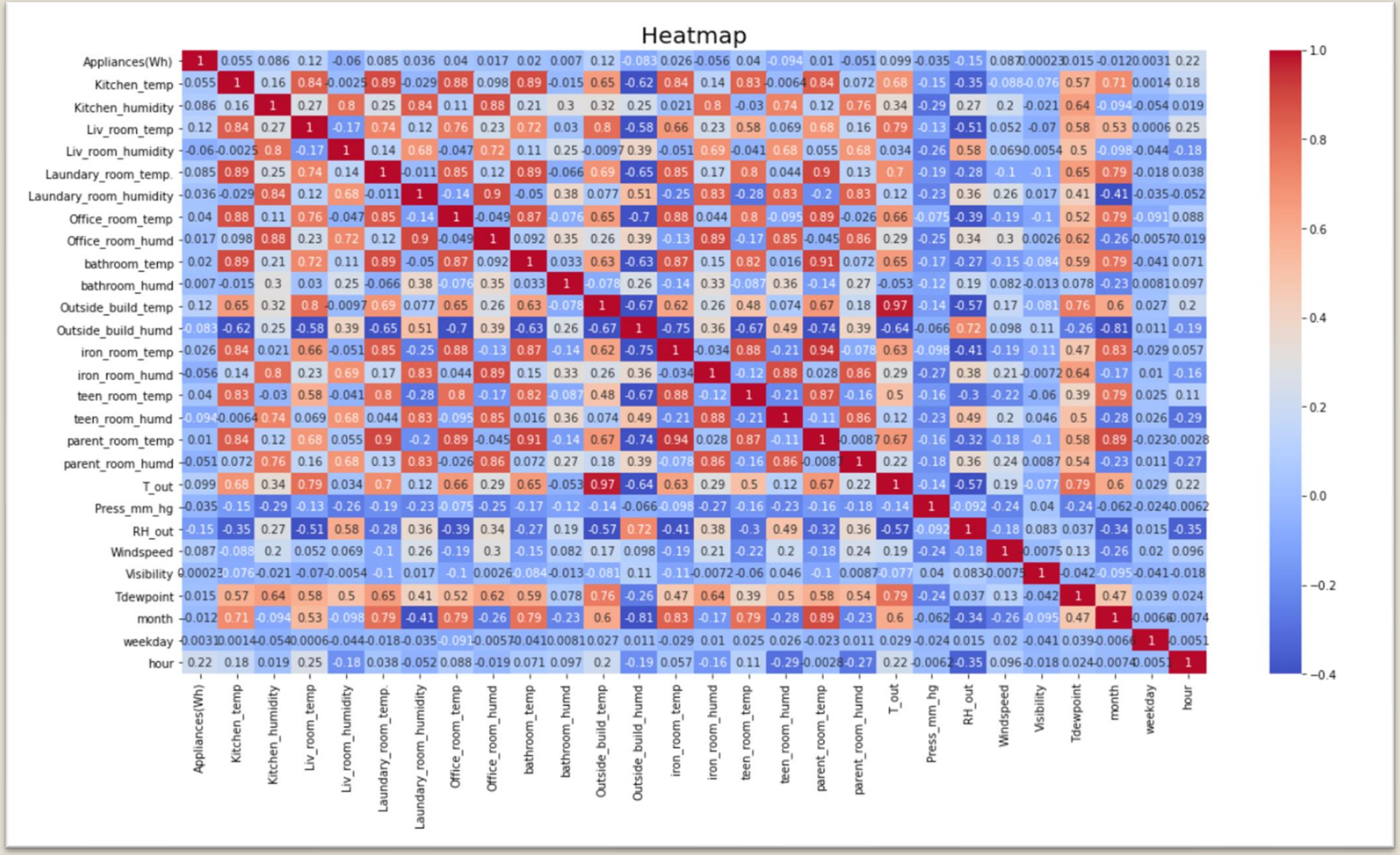
B) Outlier Treatment

```
count     19735.000000
mean         97.694958
std         102.524891
min          10.000000
25%          50.000000
50%          60.000000
75%         100.000000
max        1080.000000
Name: Appliances(Wh), dtype: float64
```

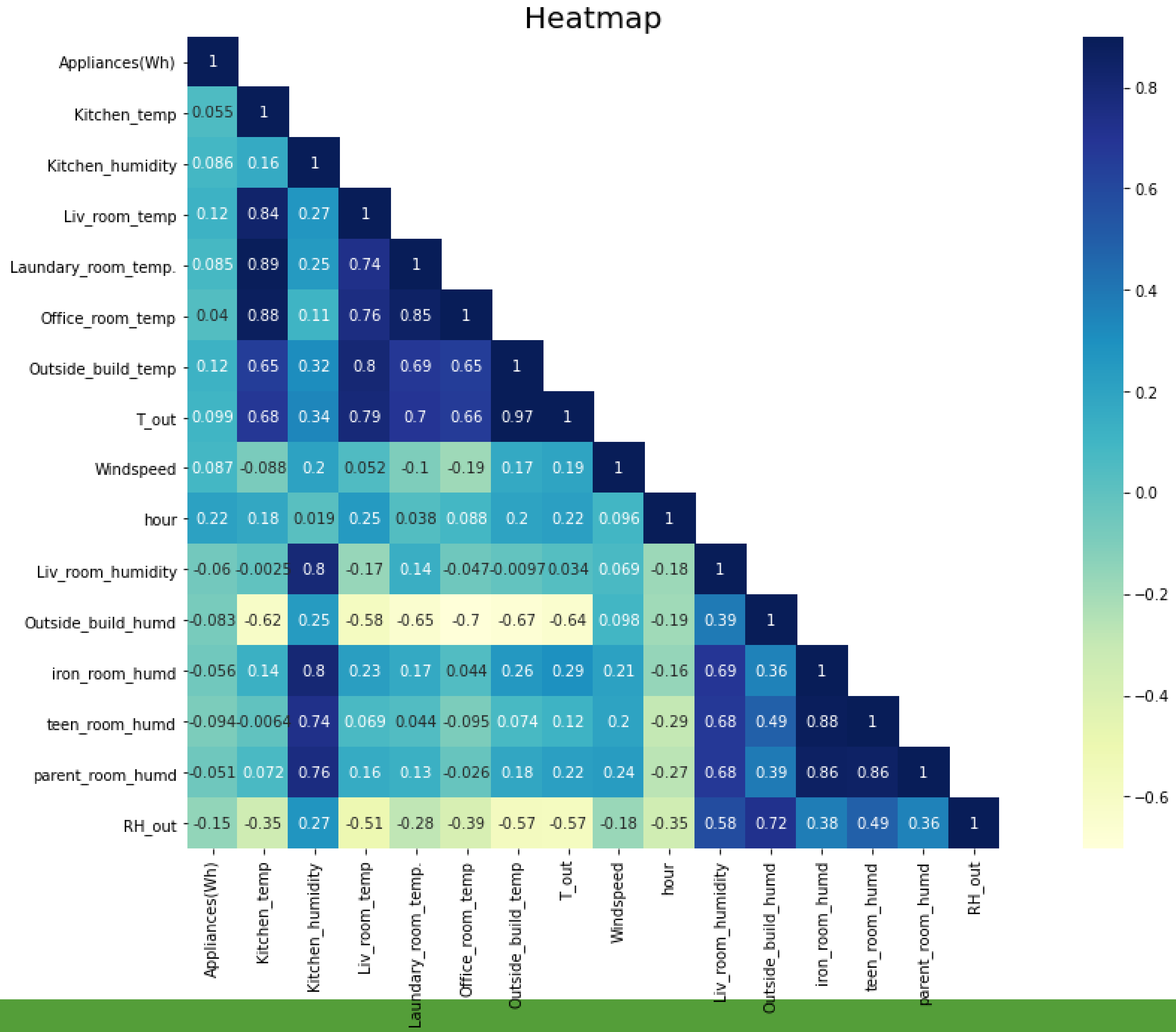
If we look into the describe table for the numeral data, only dependent variable Appliance(Wh) has outliers such that 75% of data below 100Wh, while rest 25% of data is between 100Wh to 1080Wh. But outlier treatment on such data will affect the model negatively and may cause overload power cut when higher energy supply will required. So, no outlier treatment is required in our case.

Features in red region are positively correlated while those in blue region are negatively correlated

C) Correlation Heatmap



Before Log Transformation



Data Transformation

- Heatmap in previous slide showing high multicollinearity in the dataset.
- We decreased VIF value to a large extent except for Kitchen temp. and Laundry room temp.

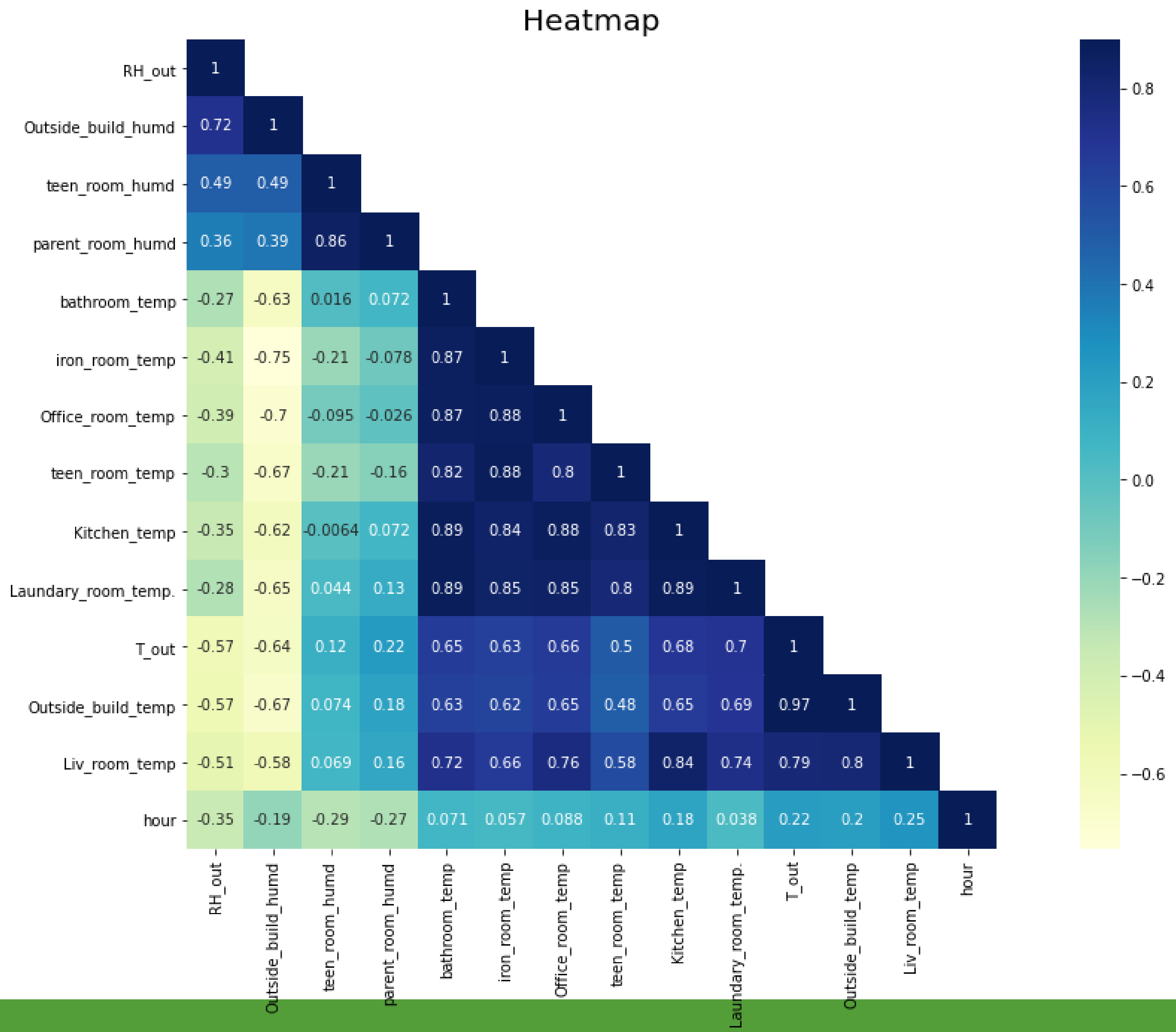
VIF values before Log Transformation

	feature	VIF
0	Appliances(Wh)	2.208654
1	Kitchen_temp	2993.903834
2	Kitchen_humidity	1542.815213
3	Liv_room_temp	2090.967638
4	Laundary_room_temp.	1108.141282
5	Office_room_temp	665.040277
6	Outside_build_temp	84.887421
7	T_out	87.267740
8	Windspeed	4.988243
9	hour	6.610228
10	Liv_room_humidity	1768.070289
11	Outside_build_humd	31.936441
12	iron_room_humd	359.071417
13	teen_room_humd	475.854506
14	parent_room_humd	545.111577
15	RH_out	166.166670

VIF values after Log Transformation

	feature	VIF
0	RH_out	135.575025
1	Outside_build_humd	33.043178
2	teen_room_humd	408.268190
3	parent_room_humd	481.850272
4	bathroom_temp	904.621001
5	iron_room_temp	994.688890
6	Office_room_temp	764.639923
7	teen_room_temp	855.863472
8	Kitchen_temp	2511.735672
9	Laundary_room_temp.	1114.722527
10	T_out	86.464774
11	Outside_build_temp	84.476223
12	Liv_room_temp	721.625450
13	hour	5.882236
14	Appliances(Wh)	51.773923

After Log Transformation and Feature Selection

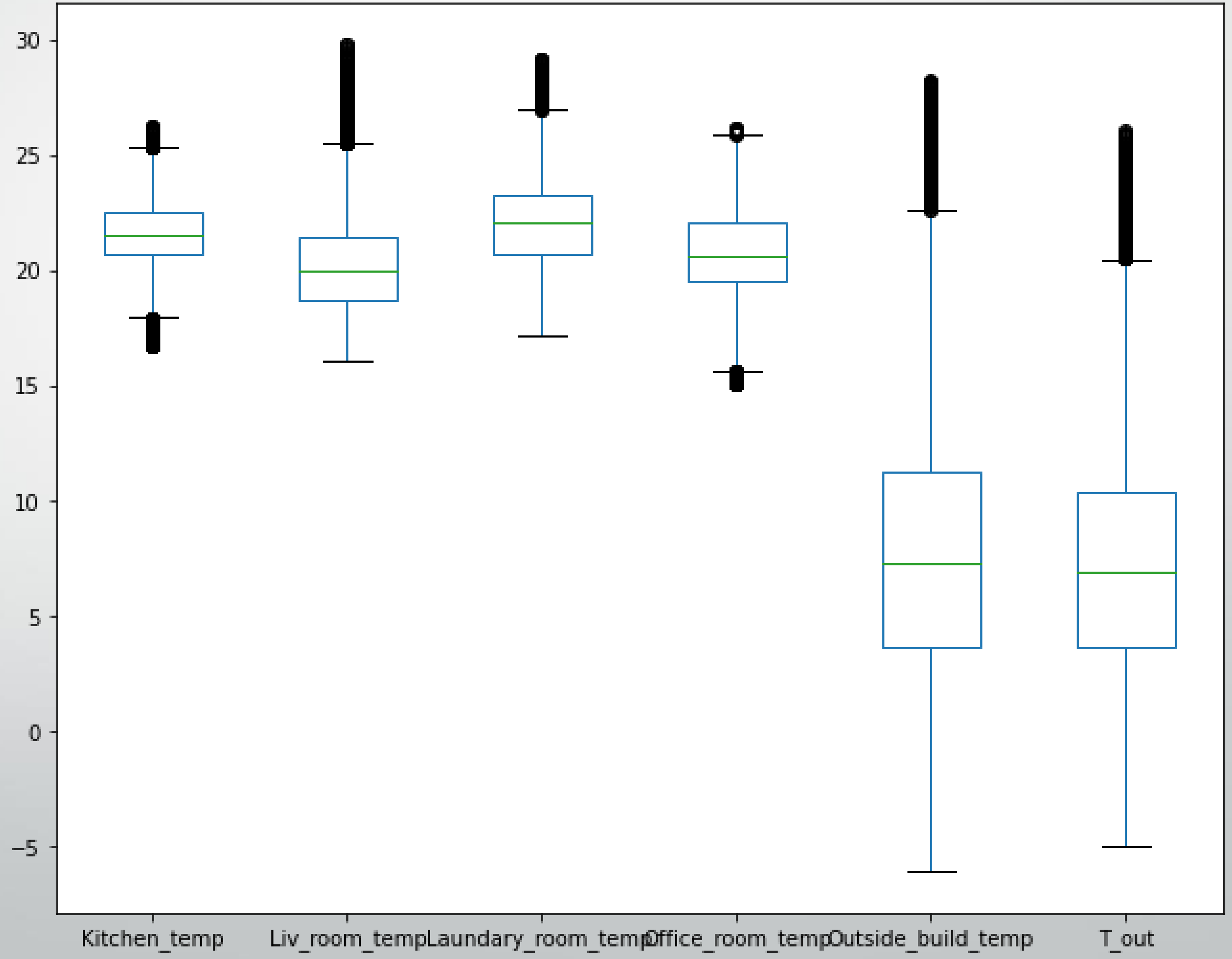


Final set of Columns after feature selection

['RH_out', 'Outside_build_humd', 'teen_room_humd', 'iron_room_temp', 'Office_room_temp', 'teen_room_temp', 'Kitchen_temp', 'Laundary_room_temp.', 'T_out', 'Outside_build_temp', 'Liv_room_temp', 'hour', 'Appliances(Wh)']

Univariate Analysis

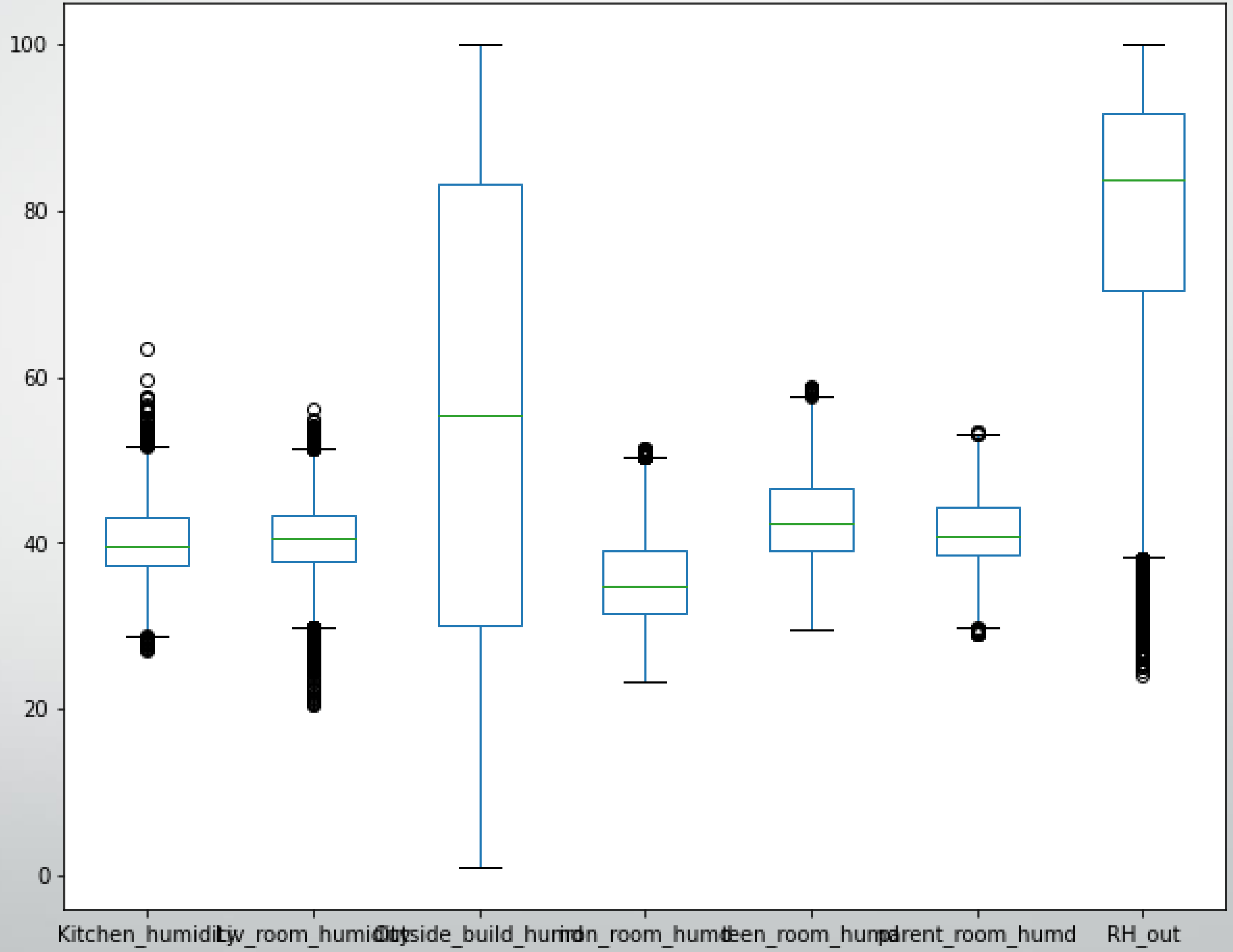
i. Temperature



a) Indoor temperature has low range of 12 from 16^0c to 28^0c, while outdoor temperature has a large range of 31 from -5^0c to 26^0c.

ii. Humidity

Similarly, indoor humidity is in range of 25 to 60%, while that of outdoor is the complete range of 0 to 100%

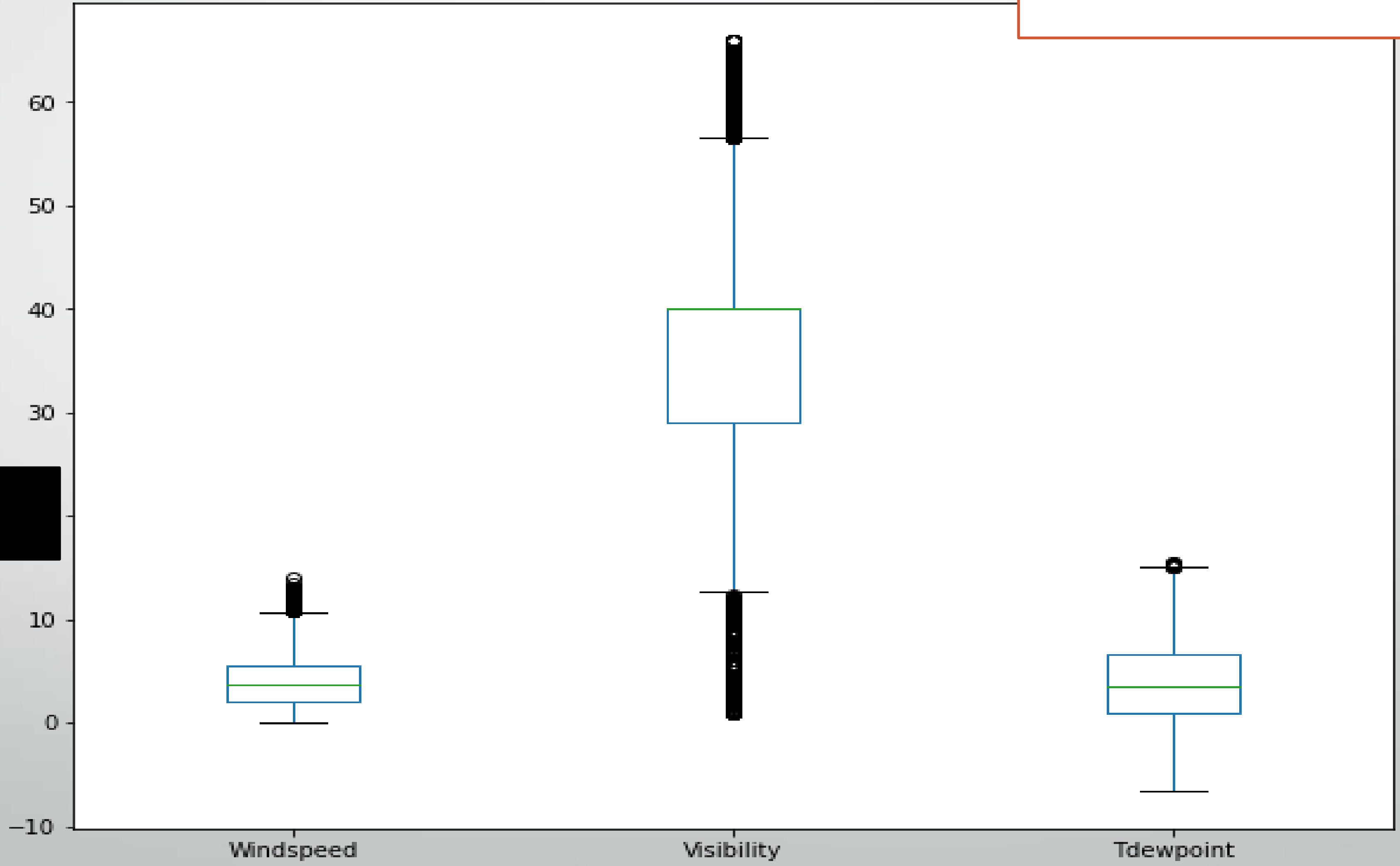


iii. Other Features

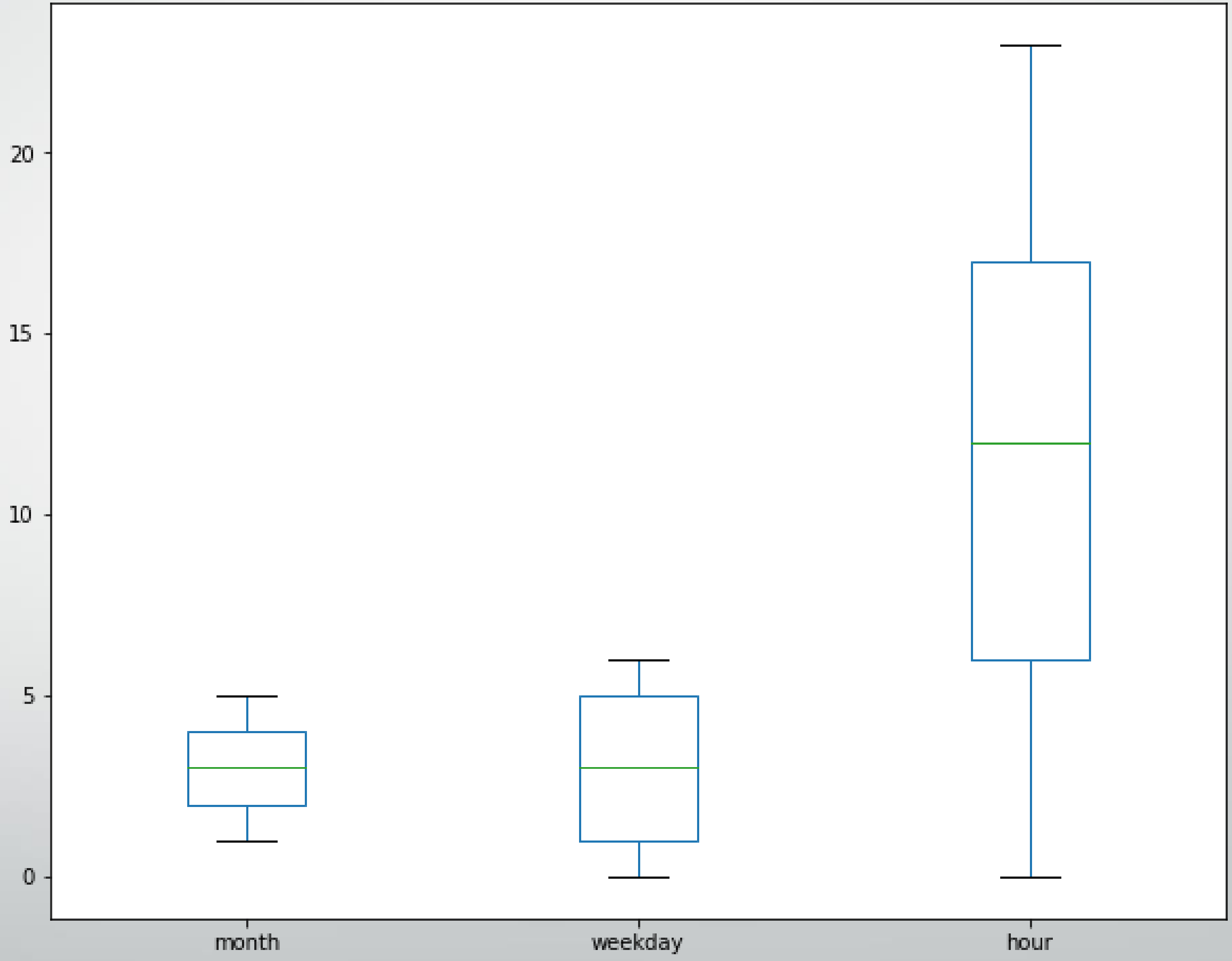
Windspeed range is normal,i.e., 0 to 15m/s

Visibility(Median) = 35 Other graph suggests, visibility is quite low in April.

Dew point temp. ranged from -5 to 15 degree celsius. This tells, 100% humidity achieved in that temp. range.

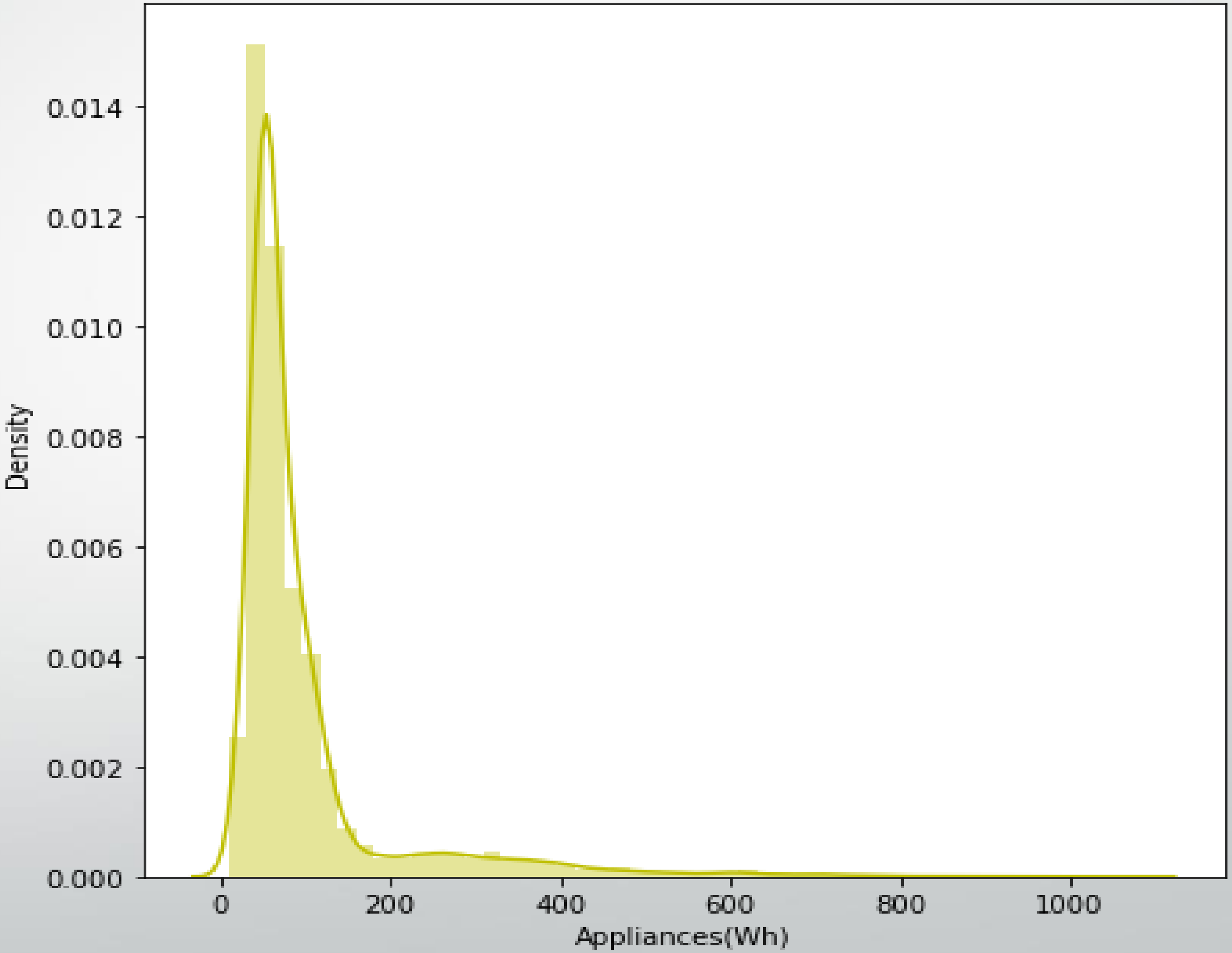


Temporal Features

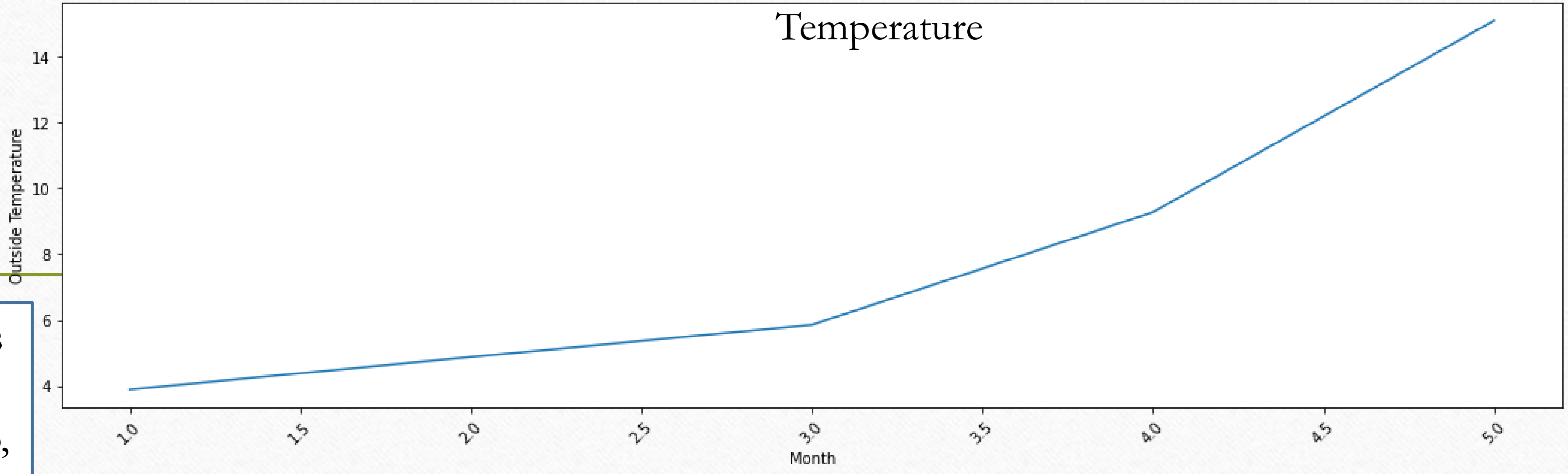


Appliance energy histogram

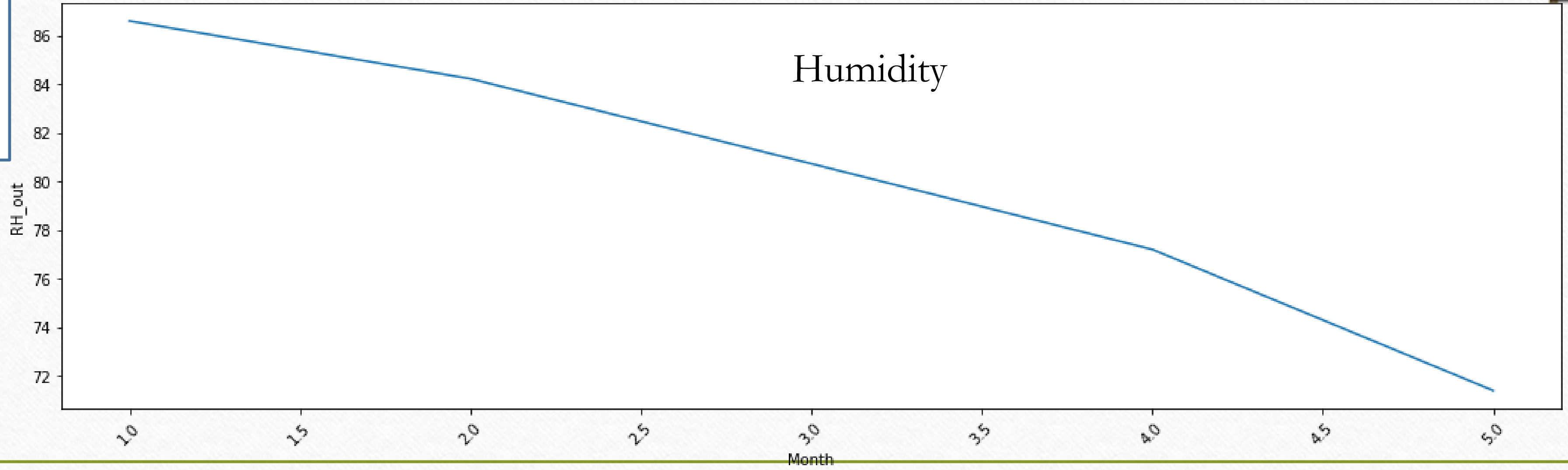
1. For most of the days appliance energy range was 0 to 50 Wh.
2. The high energy use is occasional.



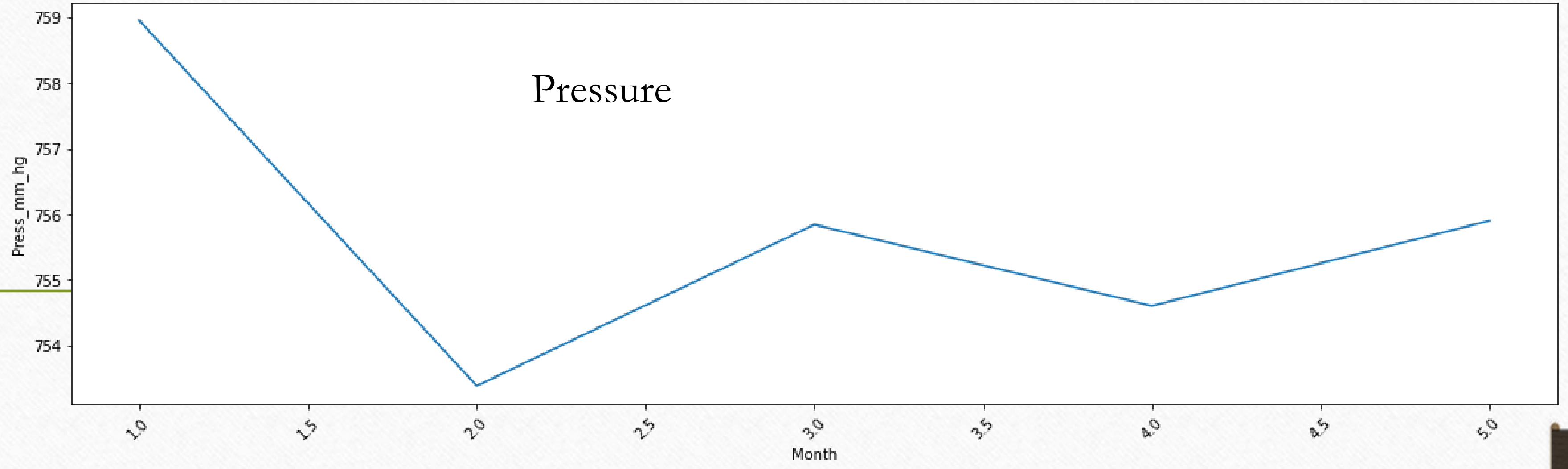
Bivariate Analysis

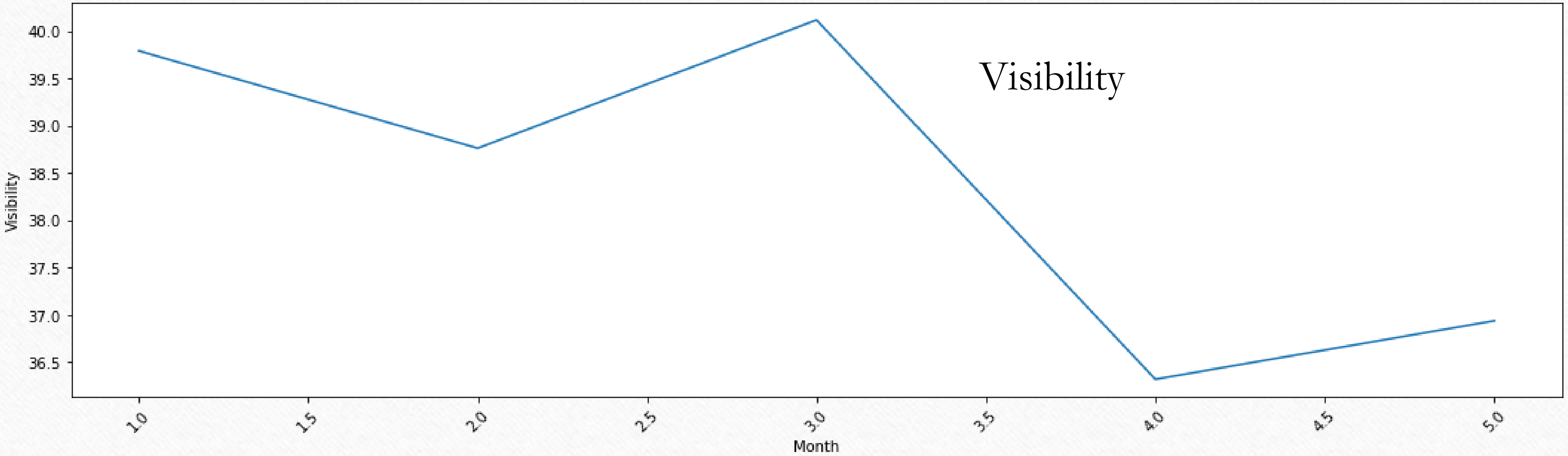


As the temp. increases from January to May, humidity decreases. So, there is inverse relationship in outside temperature and humidity

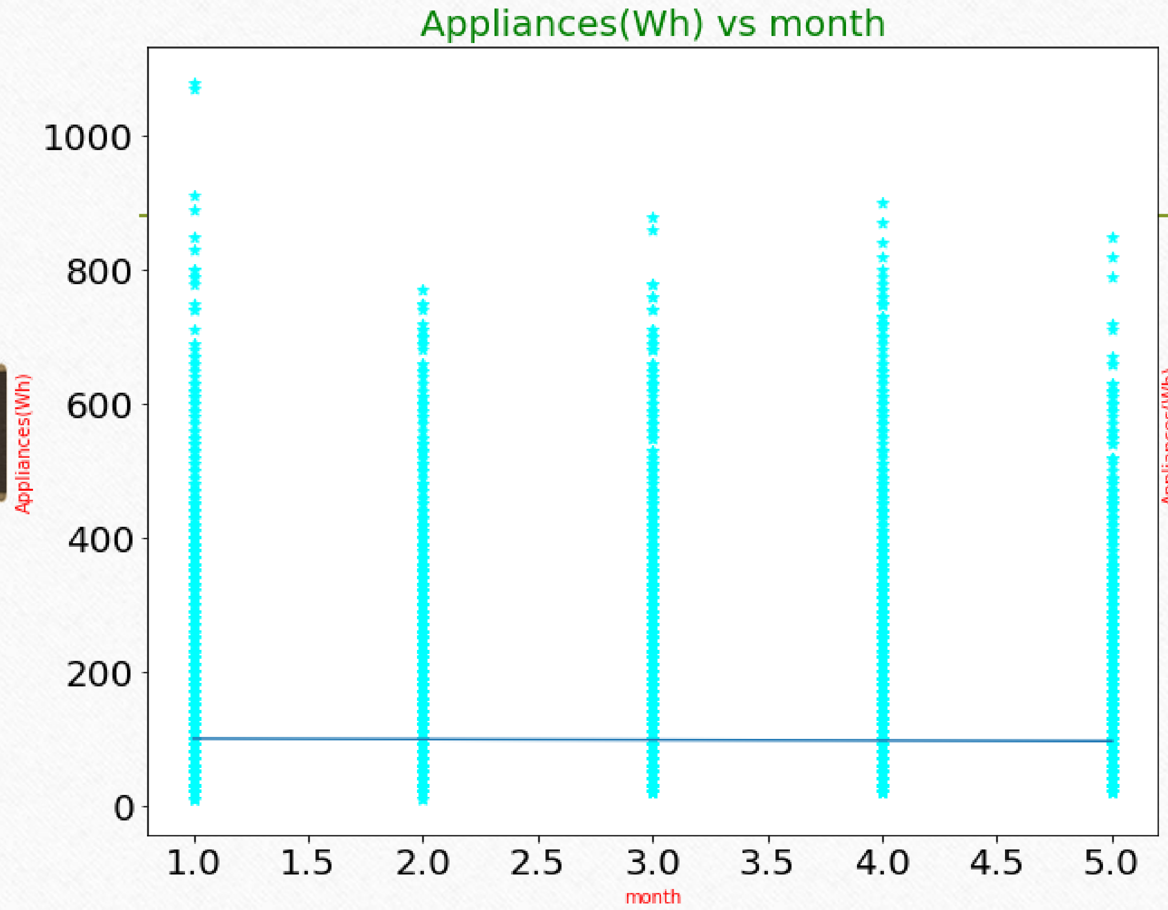


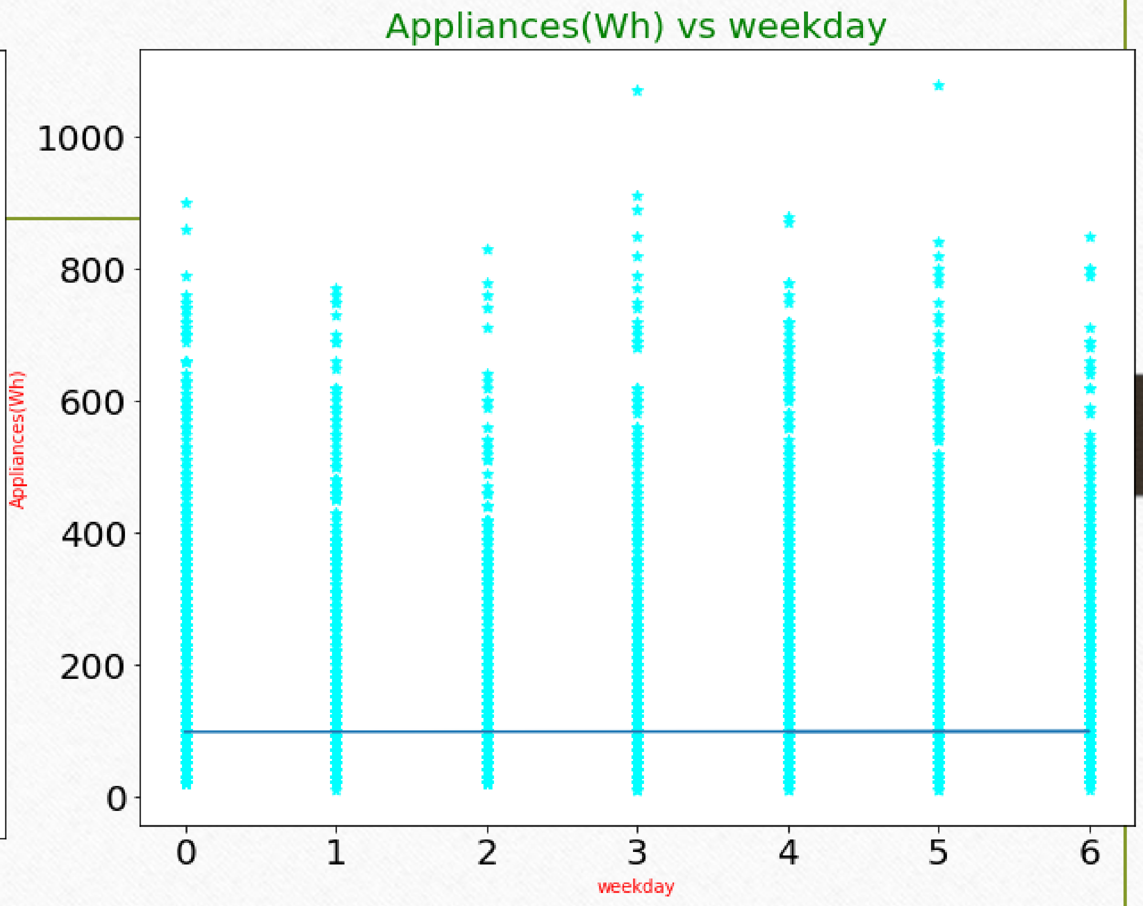
There is weak positive correlation between pressure and visibility.



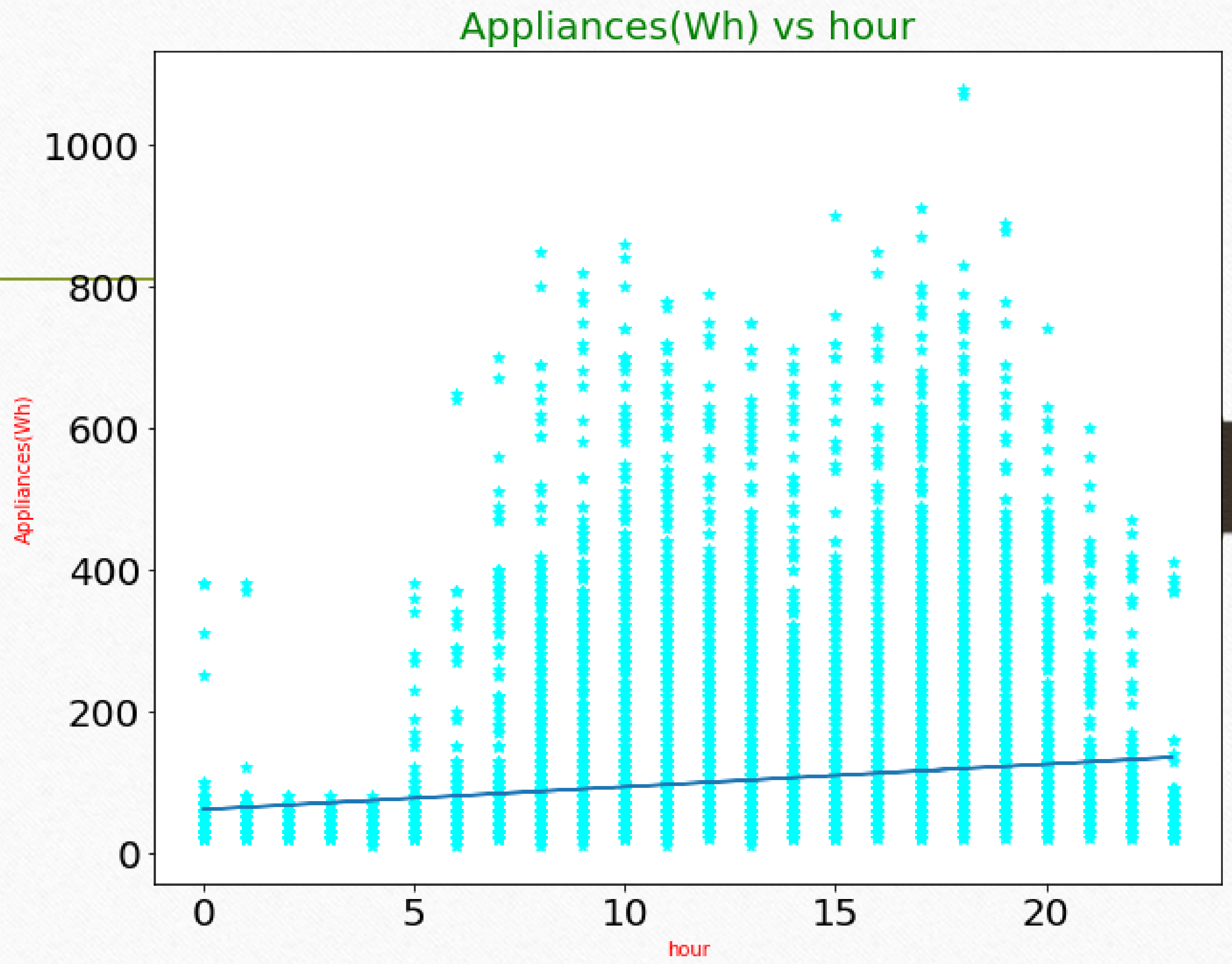


Appliance Energy with Month, Weekday and hour



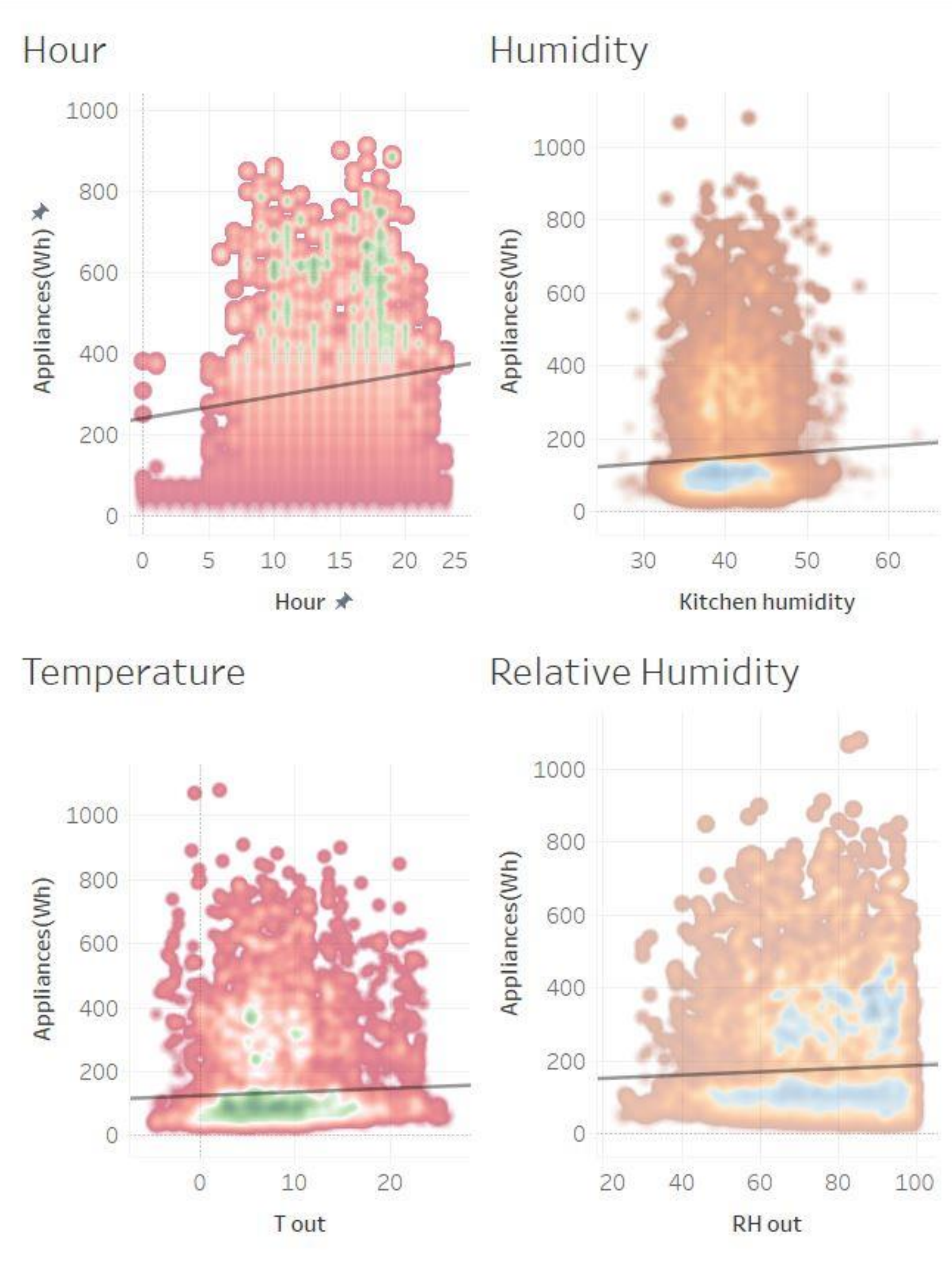


Hour has the highest positive correlation with appliance energy. From the graph, it is also visible that Evening has the highest power consumption.

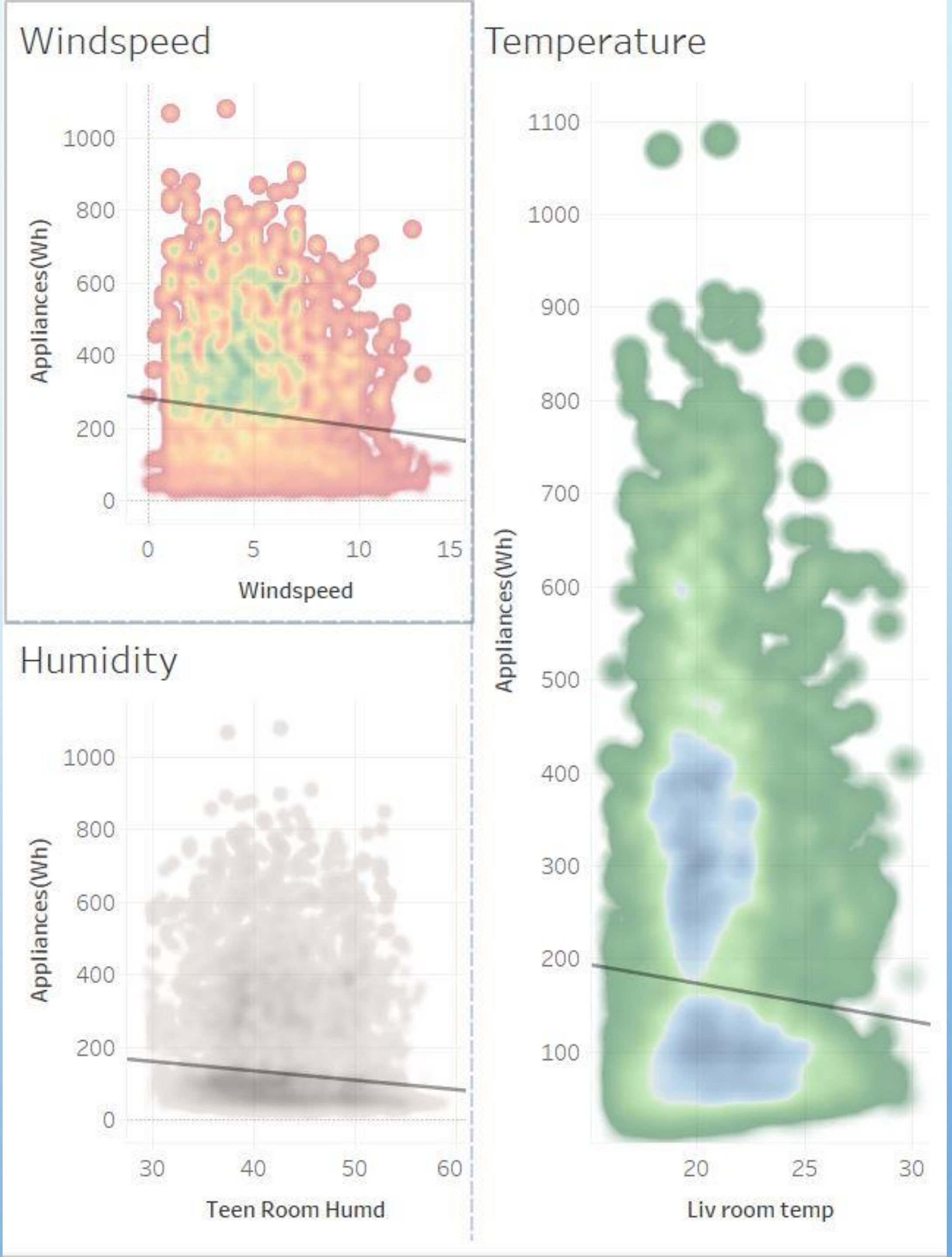


Appliance Energy Variation with highly correlated independent variables:

Upward Trendlines



Downward Trendlines



CONCLUDING EDA BY ANSWERING PROBLEM STATEMENT

1. **Affect of weather condition on power supply:** January has high humidity and low temp. with highest power consumption. May has high temp. and lower humidity with lesser power consumption

2. **Interpreting indoor temperatue-humidity balance:**

Temperature- Appliance energy use is slowly increasing with increase in temp., particularly with living room temp. and outside building temp.

Humidity- Use of appliance increases with increasing kitchen humidity while decreaseses with teen room humidity.

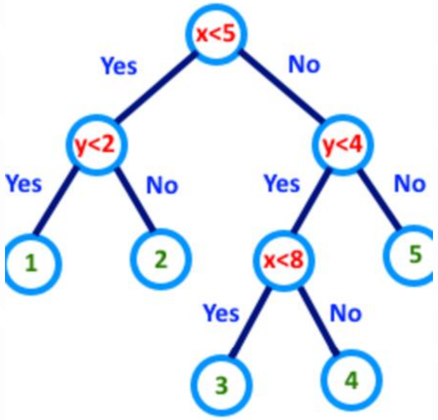
Windspeed, visibility and pressure: During winters, increase in windspeed may cause decline in temperature. So, more heaters may be used. Hence, more energy consumption.

3. **Power supply with hour, week and month:** In a nutshell, February having highest energy consumption 100.9 Wh, Monday having highest average appliance energy use 111.45 Wh and evening at 6 o'clock has maximum consumption 1,100 Wh in every 10 minutes

Machine Learning Algorithm

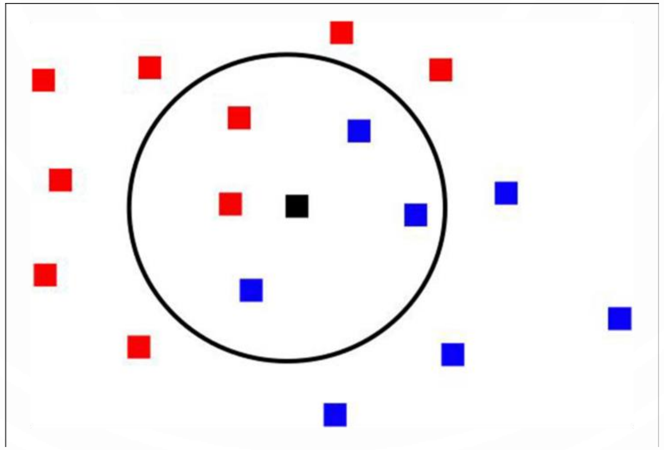
- There is a set of mathematical algorithm which a computer system can operate on the given dataset and can create a model that will predict the future value for unseen data
- ML algorithm used for this regression problem:

1. Decision Tree – Searches for distinct value of each predictor and split into nodes based on lower MSE value.



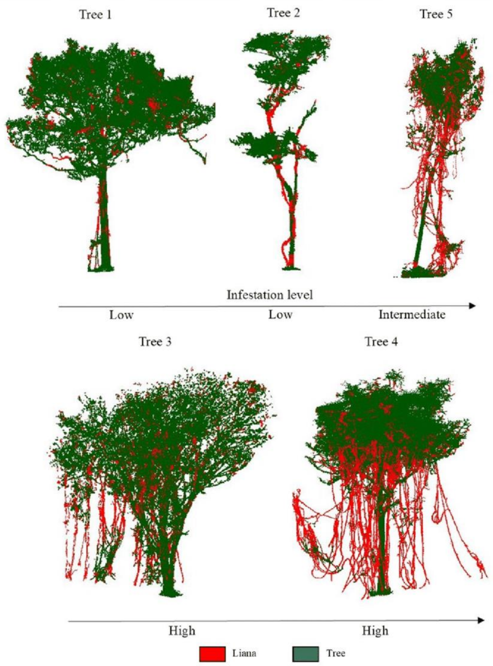
2. Random forest – Aggregate of several decision tree performed on randomly selected dataset replacing everytime.

3. Extra Tree – It first select the most efficient subset features and create a decision tree.



4. kNN(k-nearest neighbors) – Calculate Euclidean Distance of nearest neighbors from all categories and results into the category with least ED.

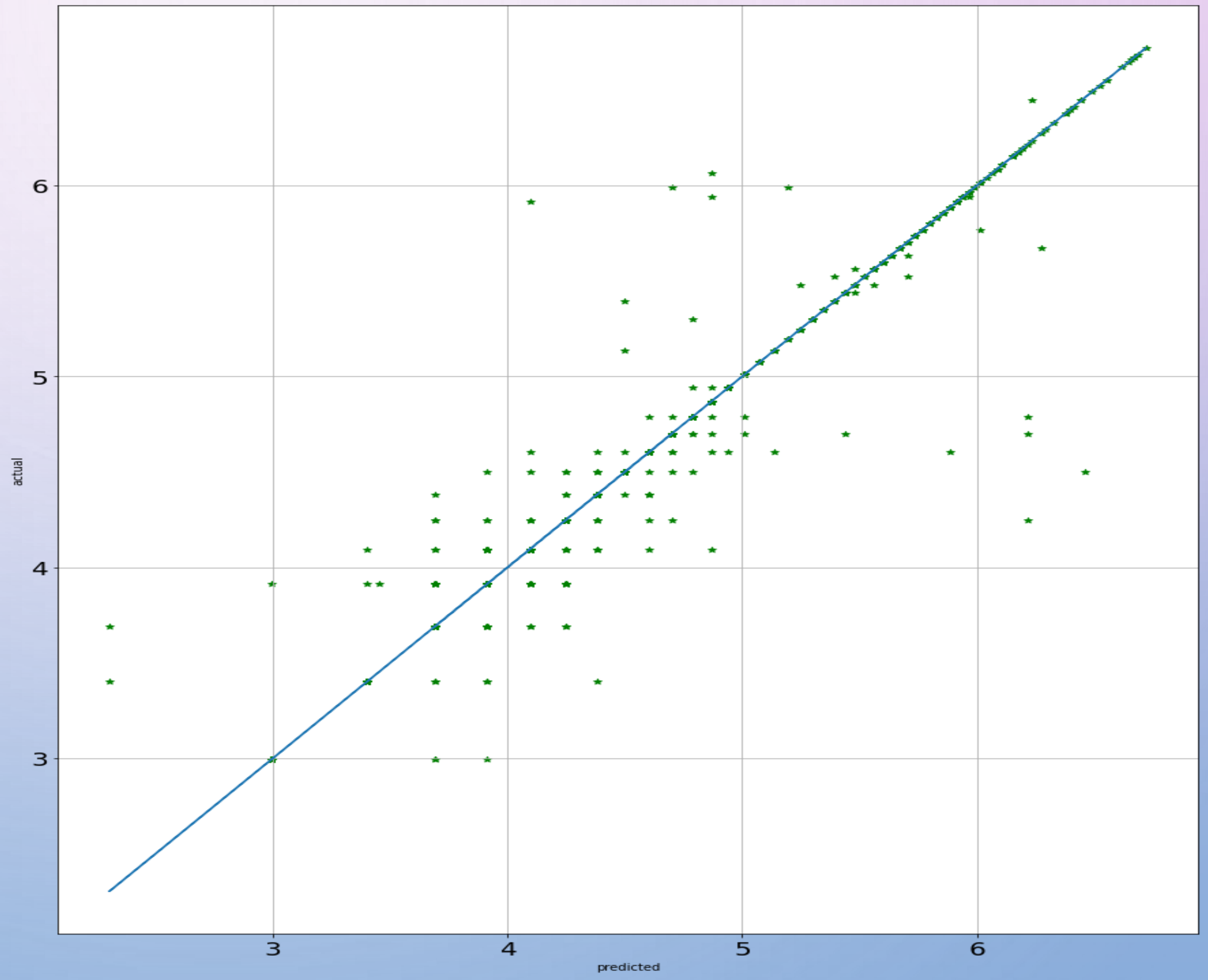
5. XGBoosting - It is a gradient boosting algorithm that uses decision trees as its “weak” predictors.



MACHINE LEARNING MODELS

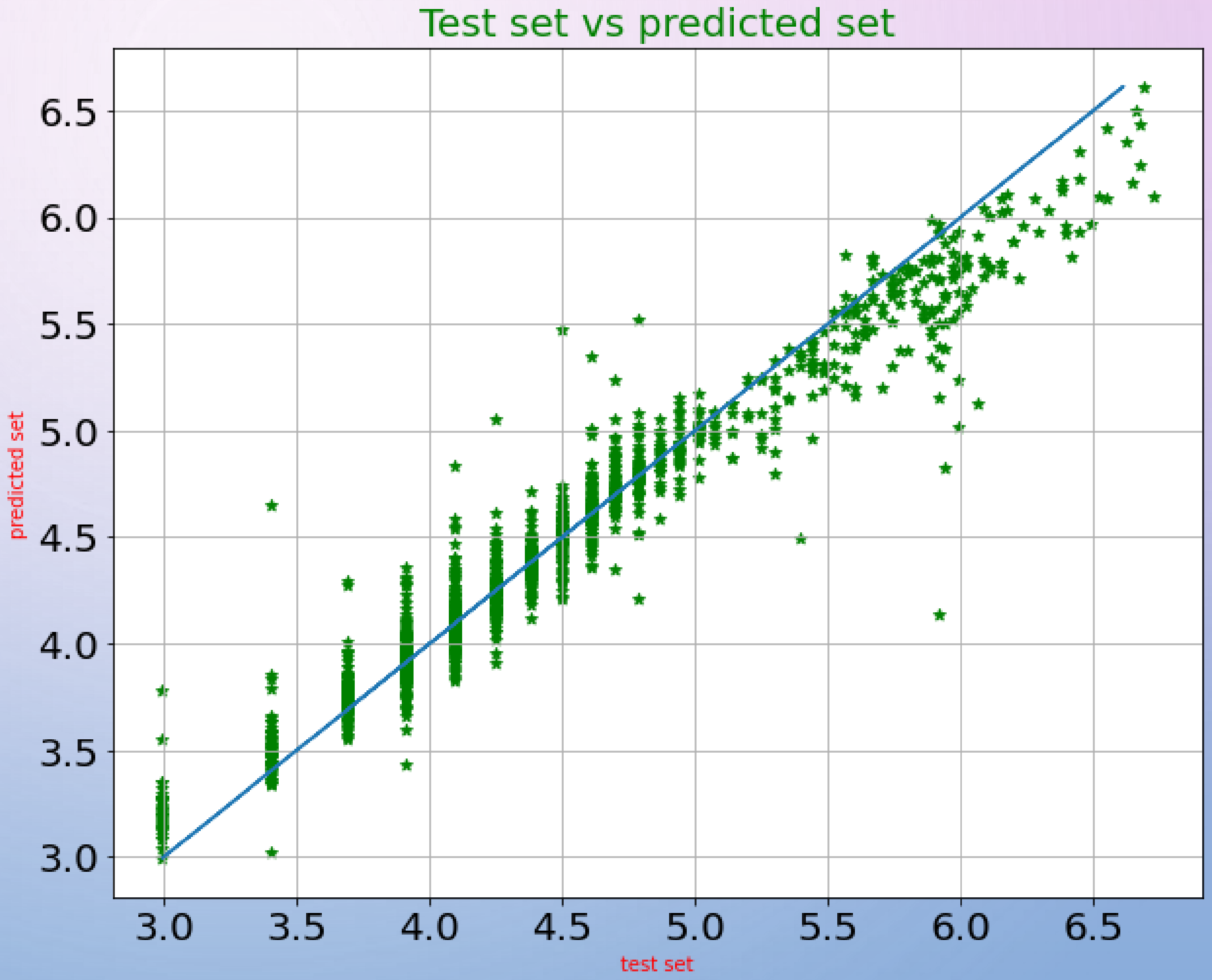
Decision Tree Regressor

```
Train Score:0.9548686074623695
Test Score:0.9496995361202017
MAE:0.028897311000629148
MSE:0.021207167025372713
RMSE :  0.145627
```



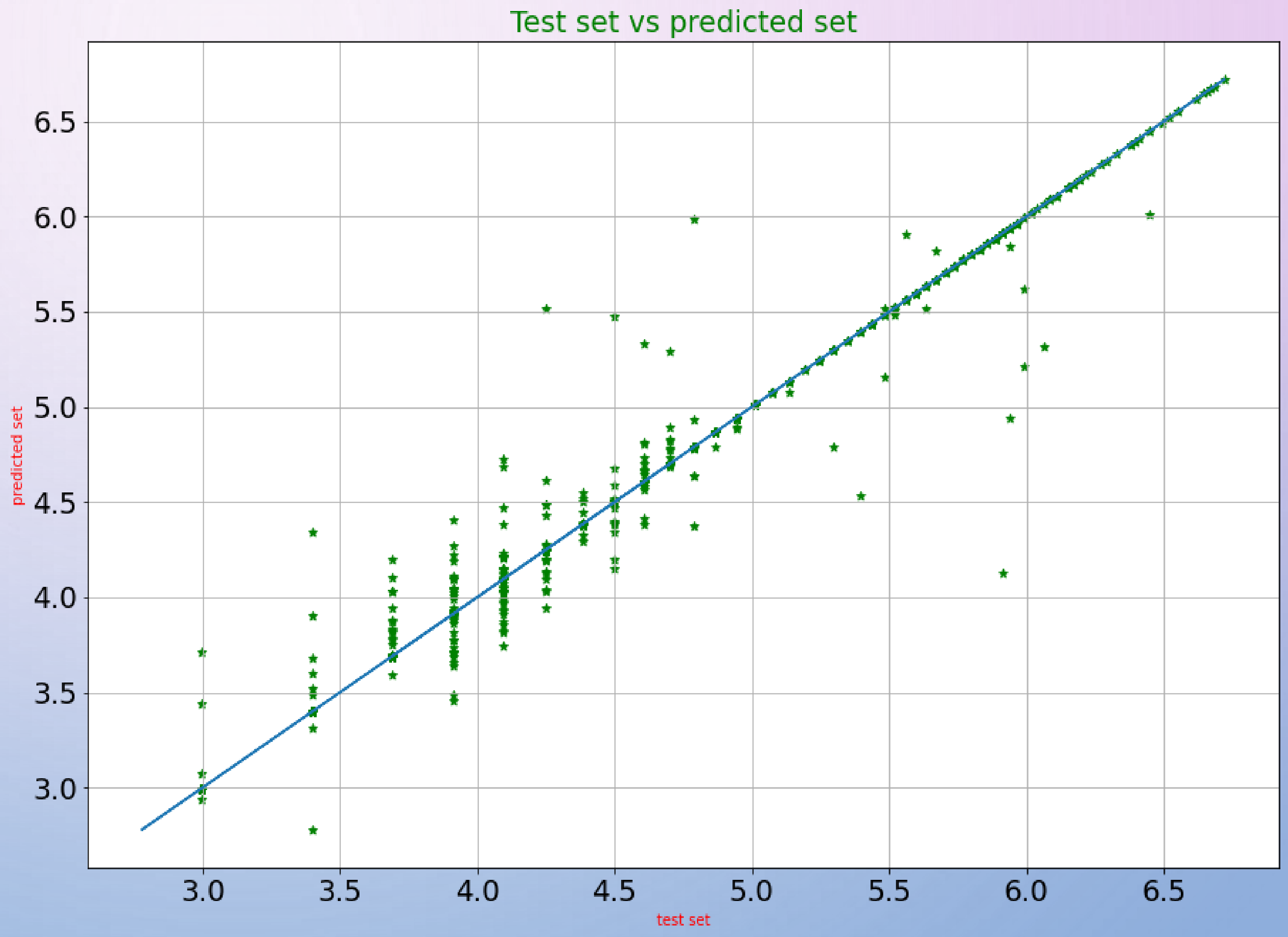
RANDOM FOREST REGRESSOR

Train Score:0.9411667013597079
Test Score:0.942314303475161
MAE:0.09509836016294805
MSE:0.024320853264904685
RMSE : 0.155951



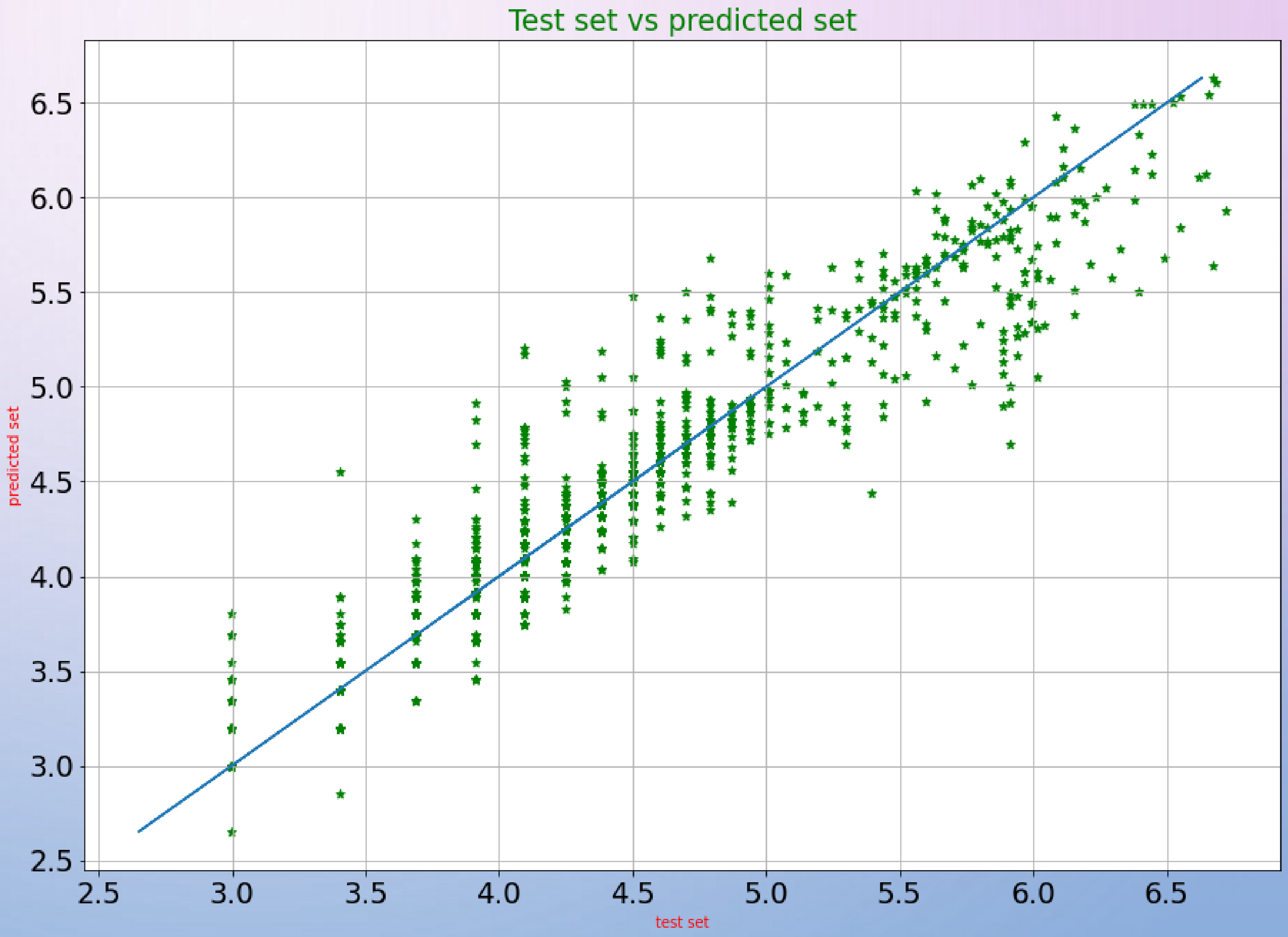
EXTRA TREE REGRESSOR

```
Train Score:0.9782578094542572
Test Score:0.9760687750546259
MAE:0.0200047625395539l
MSE:0.010089638253657189
RMSE :  0.100447
```



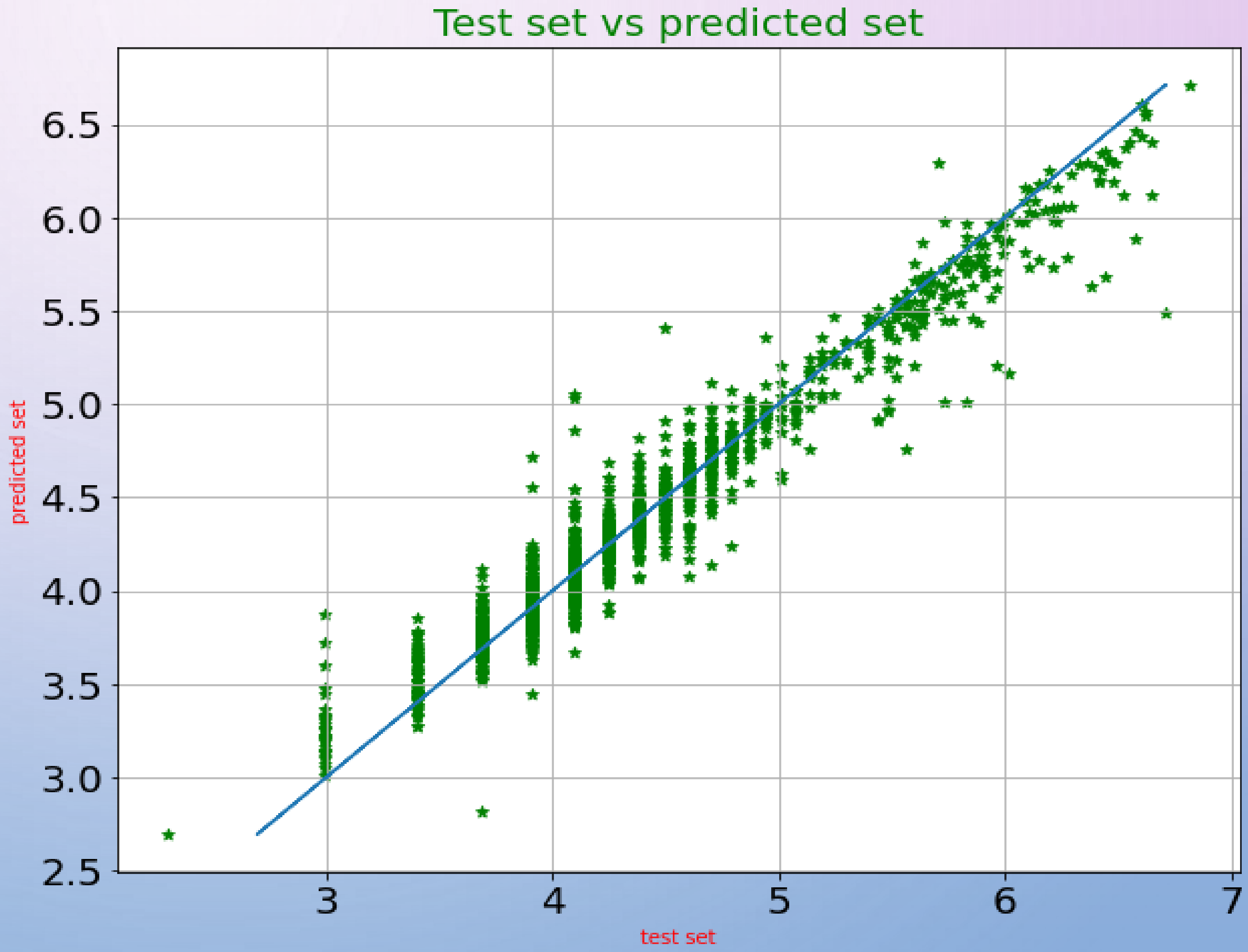
K-Neighboring Regressor

Train Score:0.9650204786674034
Test Score:0.9585396154277893
MAE:0.0256793515667925
MSE:0.01748010321853484
RMSE : 0.132212



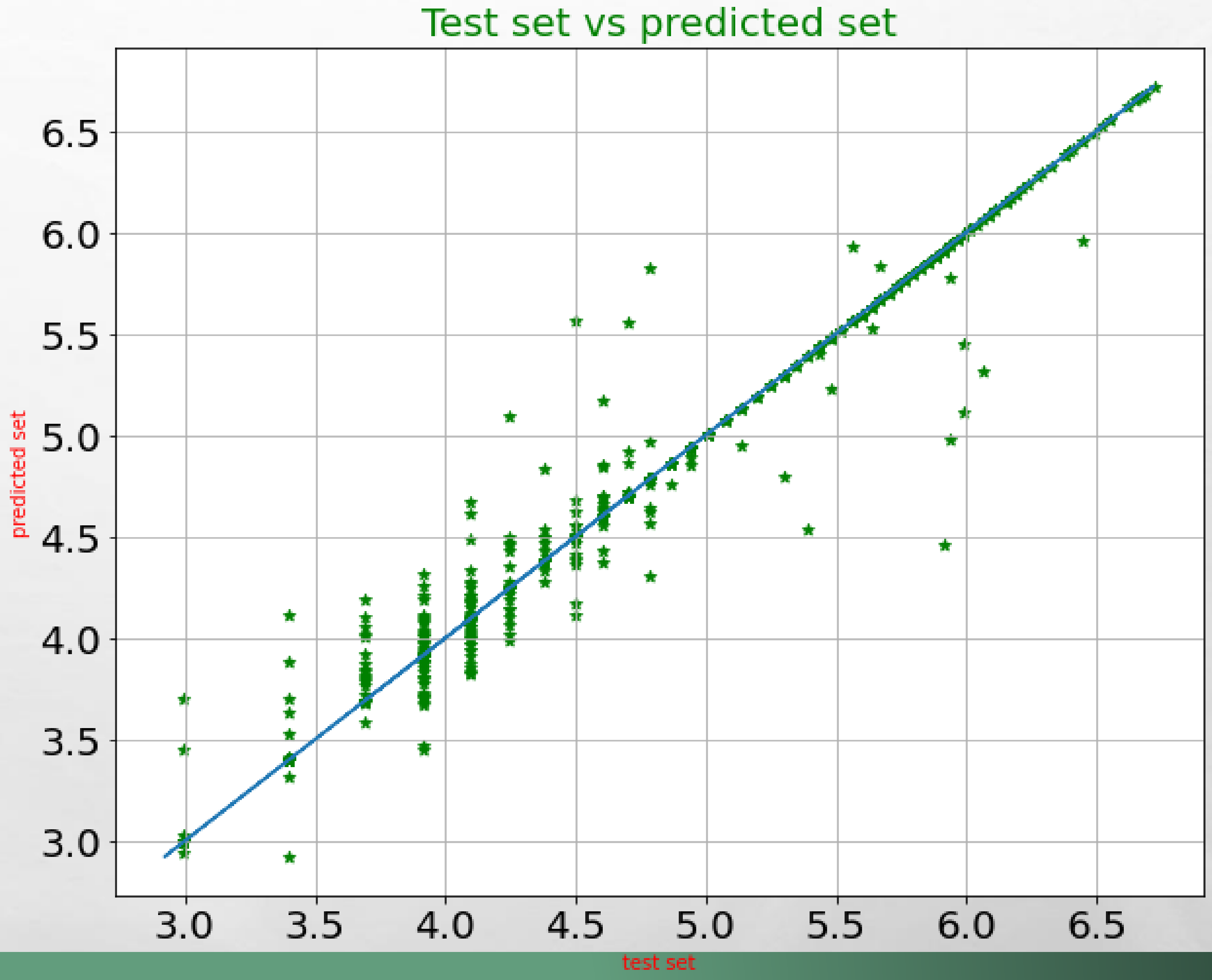
XGBOOSTING REGRESSOR

```
Train Score:0.9342689986343483
Test Score:0.941017013734922
MAE:0.10586207932024369
MSE:0.02480773438117271
RMSE :  0.157505
```



Hyperparameter tuning of ExtraTree Regressor

Best Hyperparameteric values for tuning with ExtraTreesRegressor **:**
'max_depth': 50,
'max_features': 'sqrt',
'n_estimators': 120,
‘random_state’=200



CONCLUSION ON BEST ML MODEL

1. IF WE ORDER THE MODELS AS PER TEST SCORE WE GET, IT CHANGES FROM 94.11% TO 97.6% FROM XGBREGRESSOR TO EXTRATREEREGRESSOR,
2. EXTRATREEREGRESSOR IS THE BEST MODEL OUT OF FIVE WITH MORE THAN 97.6% ACCURACY.
3. AFTER HYPERPARAMETER TUNING, THE RMSE VALUE DECREASED FROM 0.100 TO 0.094 AND THE MODEL IS ENHANCED FROM 97.6 TO 97.88%.
4. TO GET THE BEST RESULT A COMBINATION OF TWO DIFFERENT RANDOM STATES IS ORGANISED FOR TRAIN-TEST SPLIT

HENCE, WE HAVE OBTAINED THE BEST POSSIBLE MODEL FOR OUR APPLIANCE ENERGY DATASET.

A Picture is Worth a Thousand Words

THANK YOU