Capstone Project-2
Appliances Energy Prediction

<u>Team Members</u>

Navin Kodam
Meet Delvadiya
Jyoti Chiluka
Muta Ravali

Content :

1. Defining Problem Statement
2. EDA and Feature Engineering
3. Feature Selection
4. Preparing Dataset for Modeling
5. Apply Model
6. Model Validation and Selection
7. Conclusion

The Dilemma :

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. Two random variables have been included in the data set for testing the regression models and to filter out non-predictive attributes (parameters).

Data Pipeline :

Data processing-1 : In this first part we have removed unnecessary features.

Data processing-2 : In this part, we manually go through each features selected from part 1, and encoded with numerical features.

EDA : In this part, we do some exploratory data analysis (EDA) on the features selected in part-1 and 2 to see the trend.

Create a model : Finally, in this last but not last part, we create models. Creating a model is also not an easy task. It is an iterative process. We show how to start a simple model, and slowly add complexity for better performance.

DATA SUMMARY :

lights : Energy use of light fixtures in the house.
T1 : Temperature in kitchen area
RH_1 : Humidity in kitchen area
T2 : Temperature in living room
RH_2 : Humidity in living room
T3 : Temperature in laundry room area
RH_3 : Humidity in laundry room area
T4 : Temperature in Office room
RH_4 : Humidity in Office room
T5 : Temperature in Bathroom
RH_5 : Humidity in Bathroom

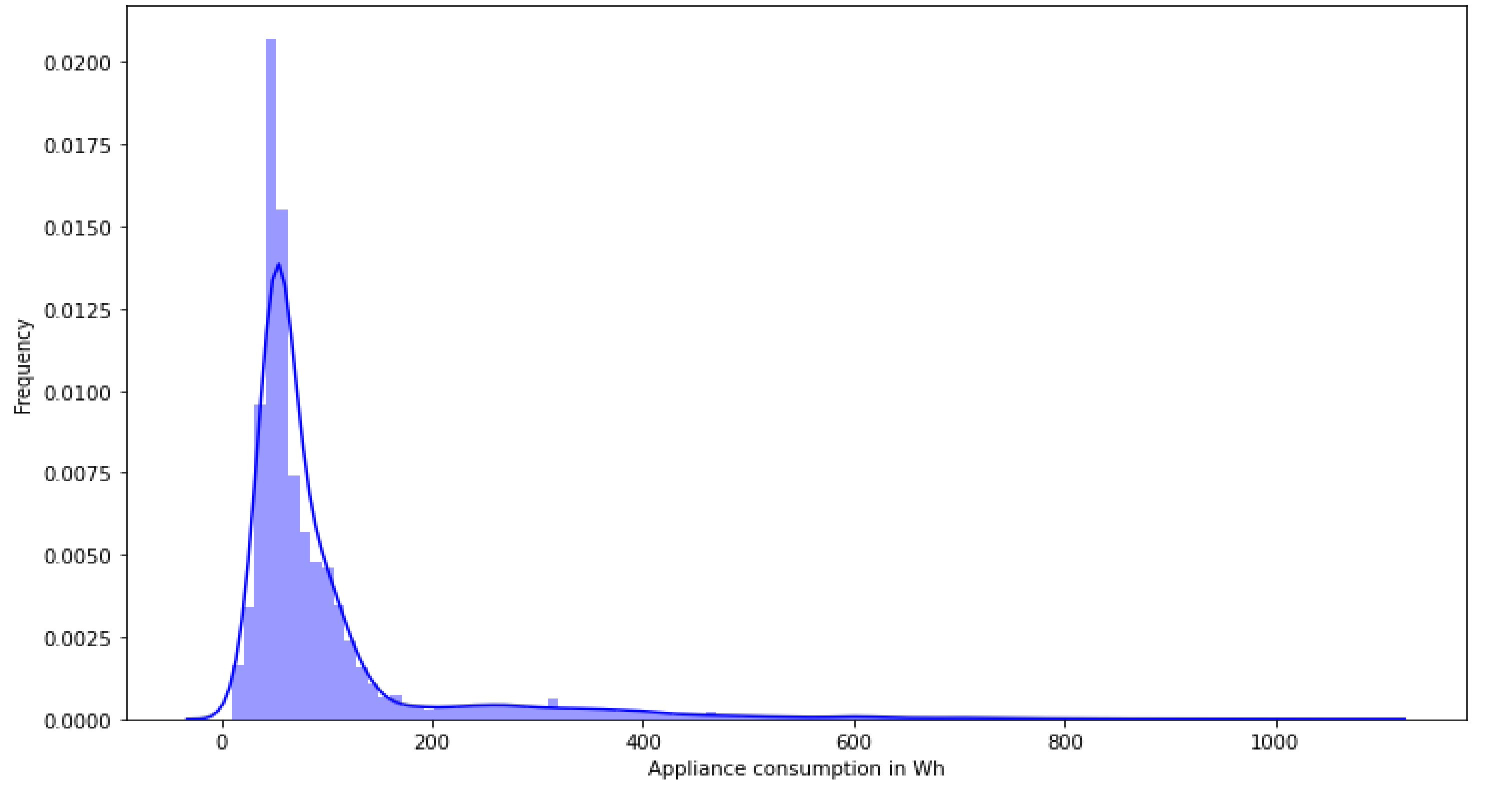
DATA SUMMARY (Contd...) :

T6 : Temperature outside the building (northside)
RH_6 : Humidity temperature outside the building (northside)
T7 : Temperature in ironing room
RH_7 : Humidity in ironing room
T8 : Temperature in teenager’s room
RH_8 : Humidity in teenager’s room
T9 : Temperature in parent’s room
RH_9 : Humidity in parent’s room
T_out : Temperature outside (from Chievres weather station)
Press_mm_hg : Pressure (from Chievres weather station)
RH_out : Humidity outside (from Chievres weather station)

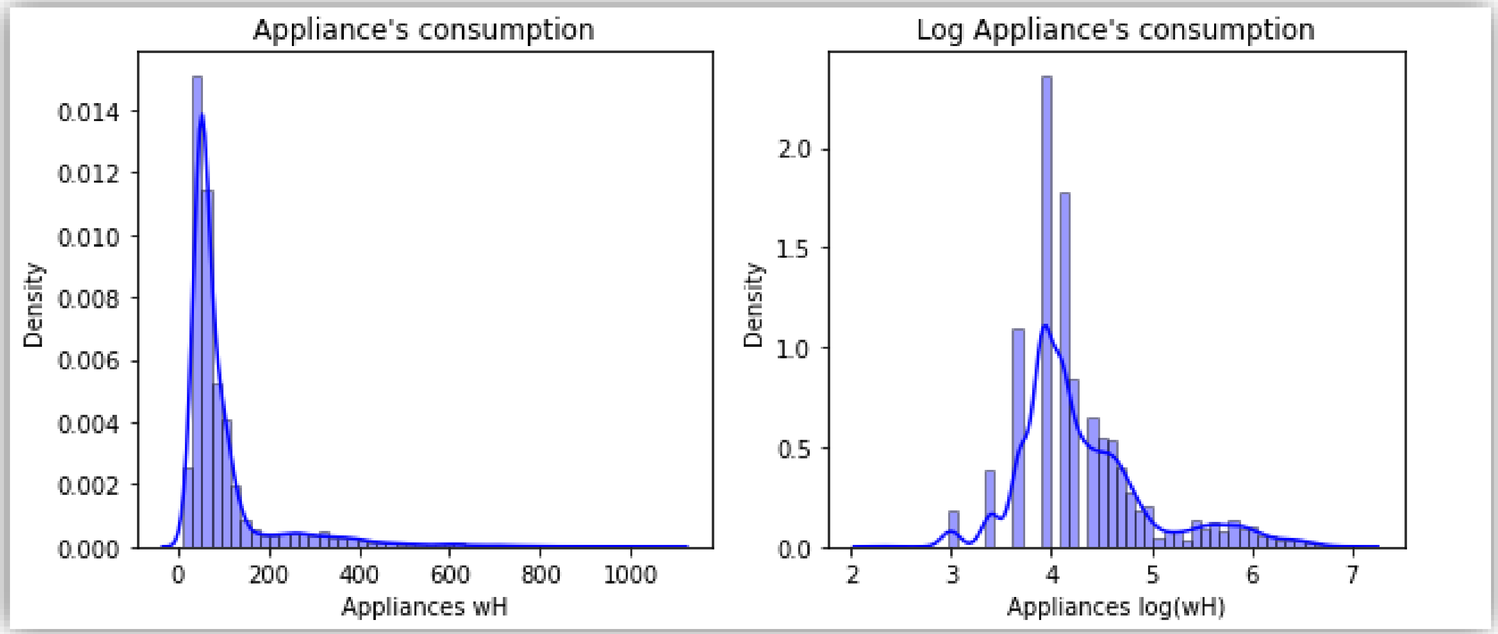
DATA SUMMARY (Contd...) :

Windspeed : Windspeed (from Chievres weather station)
Visibility : Visibility (from Chievres weather station)
Tdewpoint : Tdewpoint (from Chievres weather station)
rv1 : Random variable 1
rv2 : Random variable 2
Date : Date and time format
Appliances : Energy used by appliances

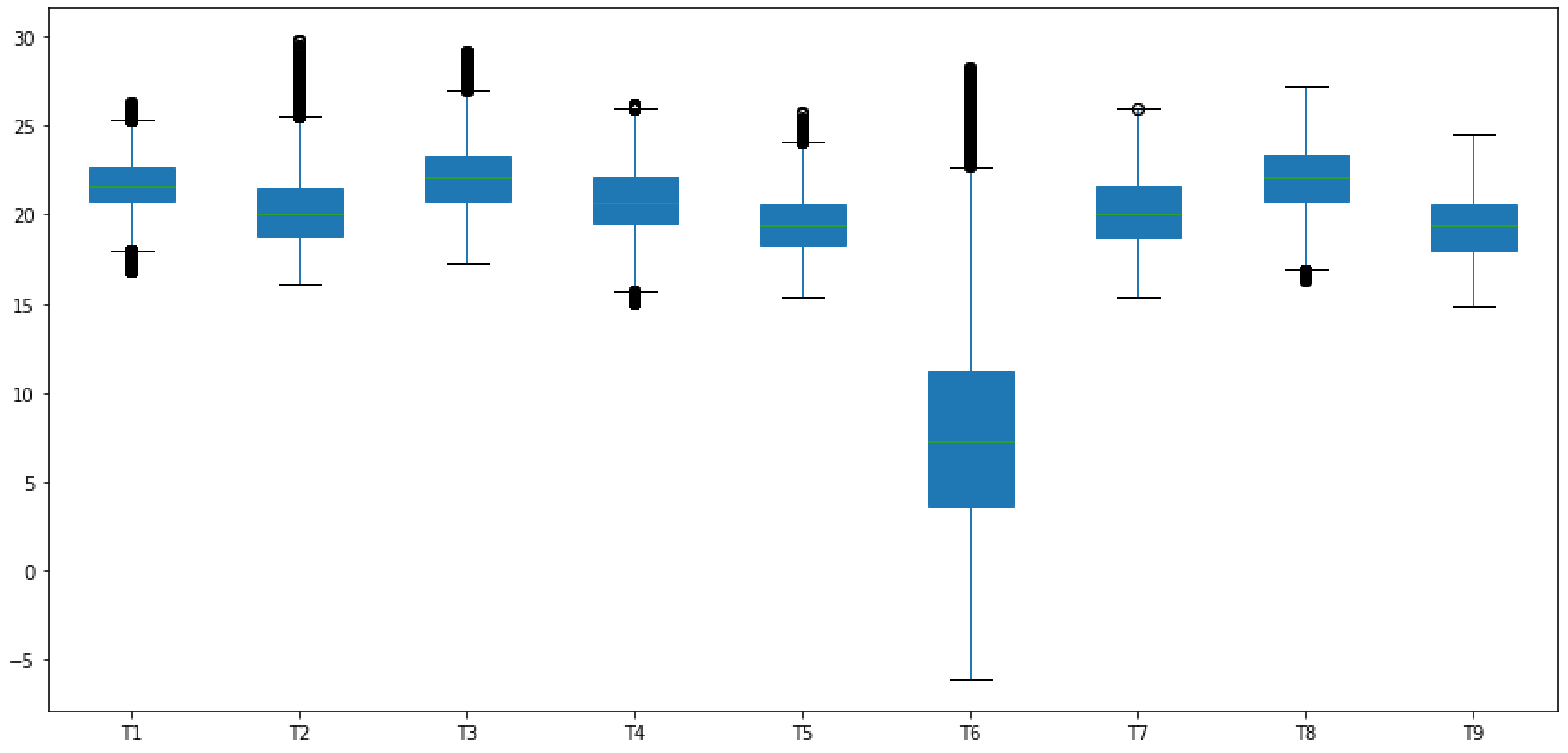
Dependent variable Appliances Graph:



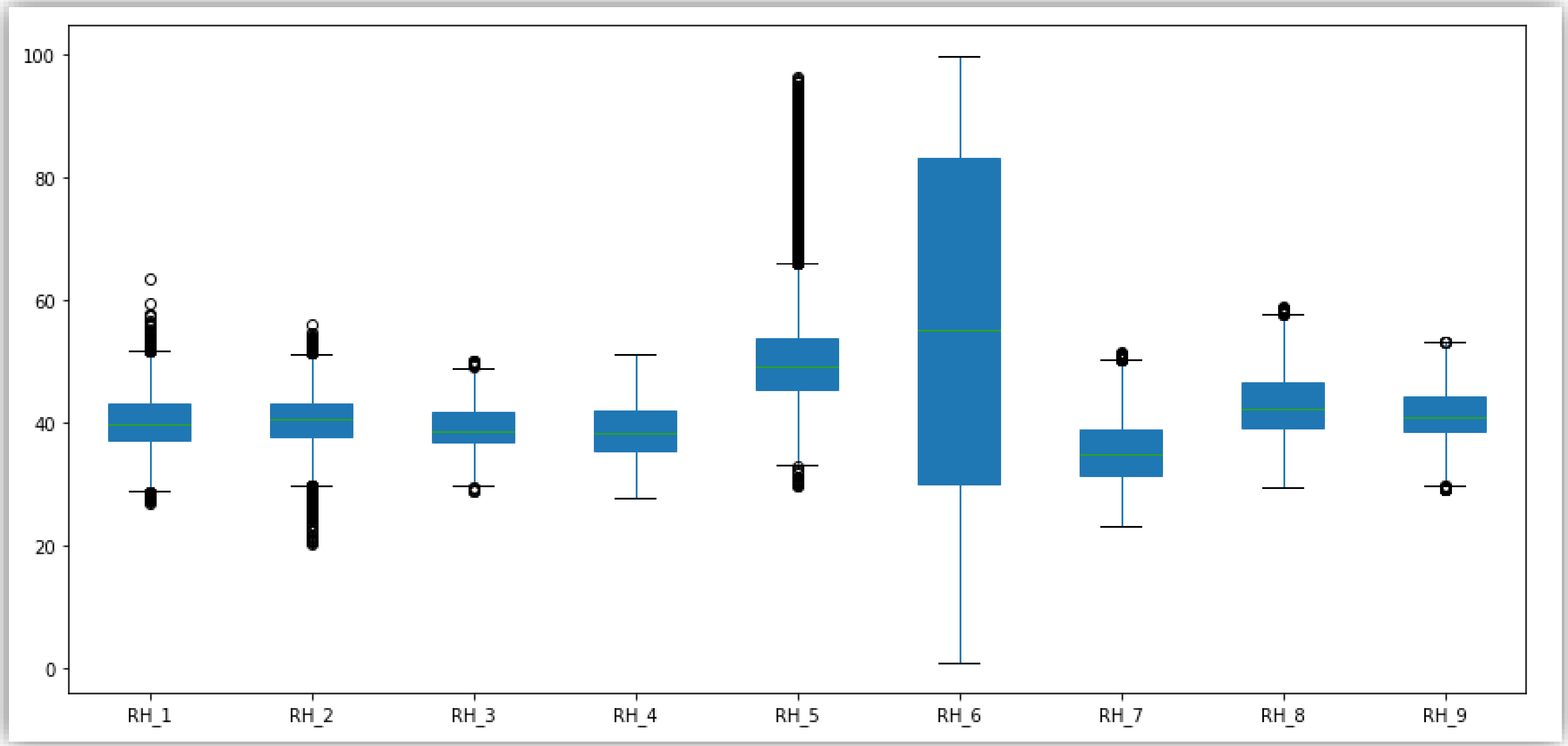
Displot (Appliances wH vs Appliances log(wH)) :



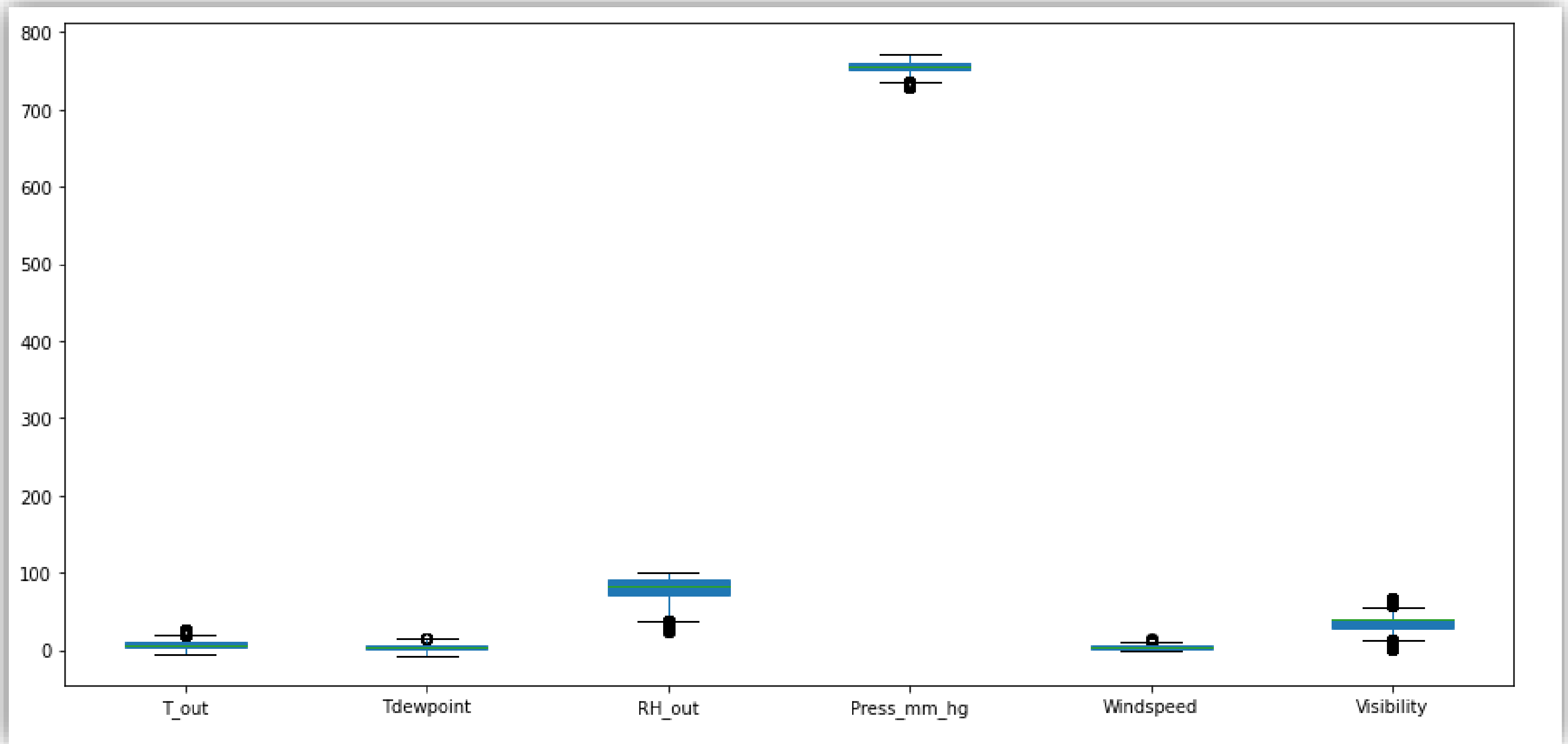
Box plot(Temperature) :



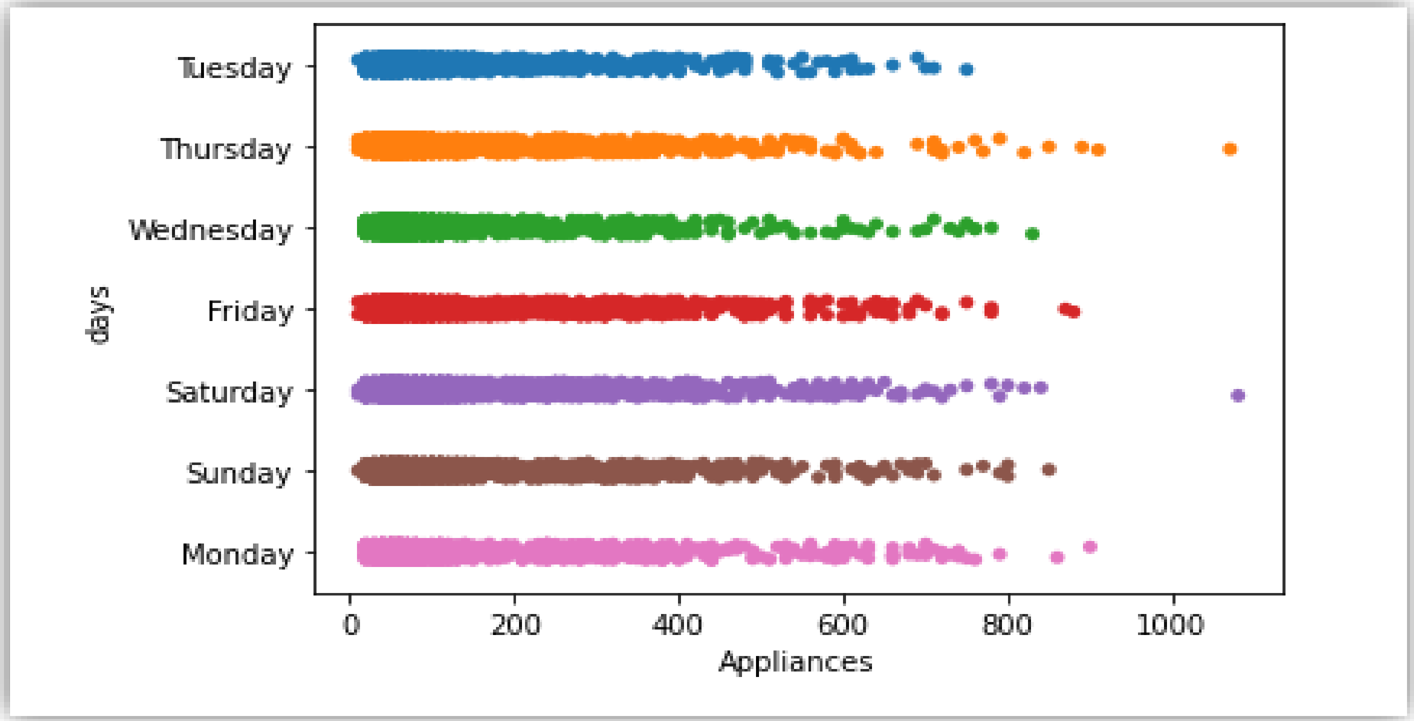
Box plot(Humidity) :



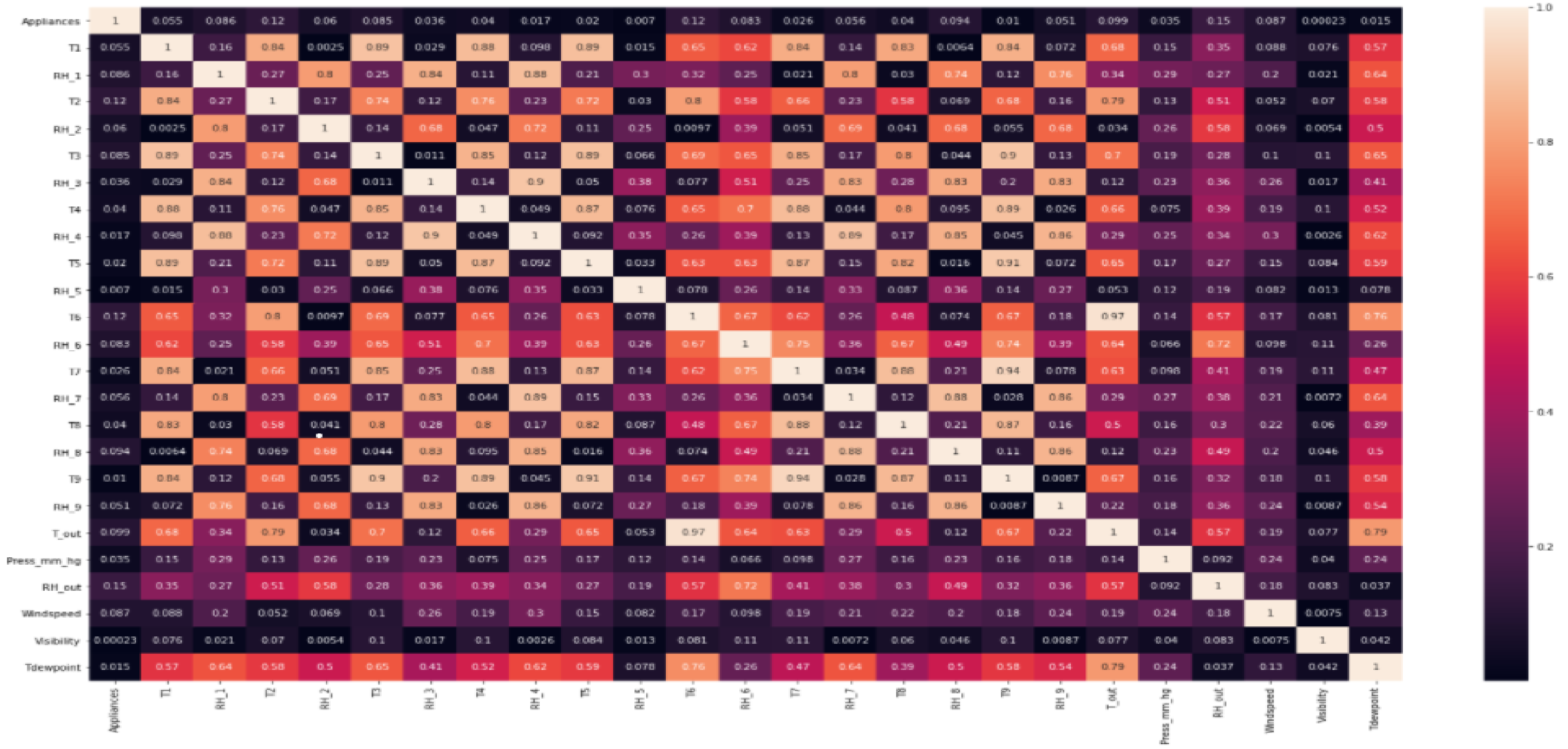
Box plot(Weather) :



Dependent variable count w.r.t. Days :



Correlation :



Preparing dataset for modeling :

Task : **Linear Regression**

Train test split (75%-25%)

Train Set : (14801, 24)

Test Set : (4934, 24)

Dependent Variable :

Appliances

Appliances	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4	T5	RH_5	T6	RH_6
60	19.89	0.475967	19.2	0.447900	19.79	0.447300	19.000000	0.455667	17.166667	0.5520	7.026667	0.842567
60	19.89	0.466933	19.2	0.447225	19.79	0.447900	19.000000	0.459925	17.166667	0.5520	6.833333	0.840633
50	19.89	0.463000	19.2	0.446267	19.79	0.449333	18.926667	0.458900	17.166667	0.5509	6.560000	0.831567
50	19.89	0.460667	19.2	0.445900	19.79	0.450000	18.890000	0.457233	17.166667	0.5509	6.433333	0.834233
60	19.89	0.463333	19.2	0.445300	19.79	0.450000	18.890000	0.455300	17.200000	0.5509	6.366667	0.848933

Reduction of features and multicollinearity

- We had reduced multicollinearity By removing irrelevant and less correlation features and considering these new features
- But these new features are not giving good score
- So we are considering the old features for Further implementation

variables	VIF
RH_5	33.886304
T6	49.249180
RH_6	24.229162
Windspeed	4.624325
Tdewpoint	19.462279
RH_4_T7	100.224776
temp_	55.452478
RH_3_RH_out_RH_1	68.332990
RH_6_RH_7	197.566361

	variables	VIF
0	T1	3604.104348
1	RH_1	1639.816095
2	T2	2490.017329
3	RH_2	2164.338515
4	T3	1239.155390
5	RH_3	1567.762332
6	T4	932.716301
7	RH_4	1357.715241
8	T5	1187.885478
9	RH_5	45.091106
10	T6	88.925465
11	RH_6	40.315447
12	T7	1613.381841
13	RH_7	518.846594
14	T8	975.014239
15	RH_8	568.351963
16	T9	2516.975132
17	RH_9	637.316129
18	T_out	399.738956
19	Press_mm_hg	2084.856382
20	RH_out	1297.930593
21	Windspeed	5.246122
22	Visibility	12.029393
23	Tdewpoint	132.494808
24	rv1	inf
25	rv2	inf

Applying Model (Polynomial Features) :

Fitting on Polynomial features of degrees (1,2,3)

Showing the R2 score and Root mean square error of Train and Test

For degree 1 :

```
Train_RMSE=  93.26197768632537
Train_R2_Score = 0.15188149648747196
```

```
Test_RMSE=  98.69343235583071
Test_R2_Score = 0.13598579386328224
```

For degree 2 :

```
Train_RMSE=  84.31139430405246
Train_R2_Score = 0.3068617908169937
```

```
Test_RMSE=  86.86620652884946
Test_R2_Score = 0.33066106350369224
```

For degree 3 :

```
Train_RMSE=  55.55639468980285
Train_R2_Score = 0.69903540691406
```

```
Test_RMSE=  37.64486125940055
Test_R2_Score = 0.8742940661330362
```

Model Validation & Selection :

Name	Train_R2_Score	Test_R2_Score	Test_RMSE_Score	Train_RMSE_Score
polynomial(degree = 3)	0.699035	0.874294	37.644860	55.556390
RandomForest	0.939304	0.518390	73.684394	24.949095
KNeighborsRegressor:	0.688354	0.440688	79.406315	56.533630
GradientBoostingRegressor:	0.340765	0.233555	92.954056	82.223627
XGBRegressor:	0.327814	0.224481	93.502689	83.027337
Lasso	0.151375	0.129445	99.066258	93.289770
Ridge:	0.151029	0.128090	99.143389	93.308828
SVR:	-0.003329	-0.017707	107.112267	101.437397

Model Validation & Selection (contd...) :

Observation 1 : Support vector regression (SVR) is giving worst r score for these dataset

Observation 2 : As seen in the above slides... Random forest is giving high train score but having less test r score , polynomial(degree = 3) is giving High test r score but having less train r score

Observation 3 : From the above observation we have come to the conclusion that we would choose our model from Random forest or polynomial features(degree = 3)

Model Validation & Selection (contd...) :

- Tuning hyperparameters of Random Forest regressor and we got the best parameters and best estimators

- But the Rmse value of Random Forest is 24.26% and Rmse for Polynomial(degree = 3) is 37.64%

- So we concluded that Random forest is giving the best score than Polynomial for these dataset

```
grid_search.best_params_
```

```
{'max_depth': 60, 'max_features': 'sqrt', 'n_estimators': 250}
```

```
grid_search.best_estimator_
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=60, max_features='sqrt', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=250, n_jobs=None, oob_score=False,
                      random_state=40, verbose=0, warm_start=False)
```

```
grid_search.best_estimator_.score(X_train,Y_train)
```

```
0.9434541975986162
```

```
grid_search.best_estimator_.score(X_test,Y_test)
```

```
0.5864961299637497
```

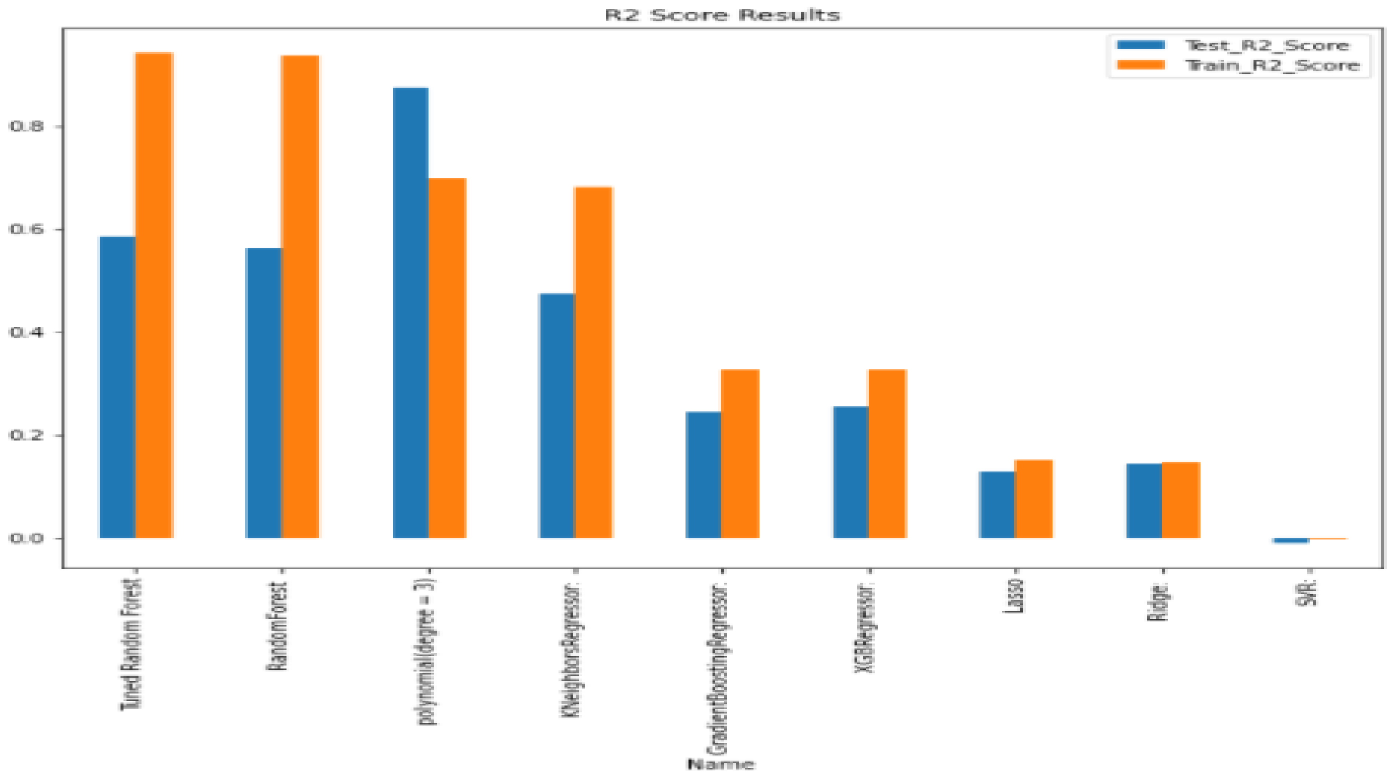
```
np.sqrt(mean_squared_error(Y_test, grid_search.best_estimator_.predict(X_test)))
```

```
65.64558516057404
```

```
np.sqrt(mean_squared_error(Y_train, grid_search.best_estimator_.predict(X_train)))
```

```
24.265303711829215
```

Comparisons of all models :



Conclusion :

- We are getting the good results when we selecting 22 features for the model implementation and dropping lights,rv1, rv2 and Visibility.
- The best algorithm for this dataset is Random Forest Regressor as compared to the rest of the algorithms.
- After tuning the algorithm using Grid Search CV on Random Forest the score is not getting much difference compared to Polynomial Regression, because the correlation between the dependent and independent variables are very low in dataset.
- Feature reduction was not able to give much better accuracy.

Challenges :

- **Mostly, features have very low correlation so feature selection was challengeable.**
- **Most of algorithms doesn’t give good score even after feature engineering.**
- **Computation time in Polynomial Regression.**

Q & A