Capstone Project - 2

Appliances Energy Prediction

By:- Milan Ajudiya

Introduction:-

Today's time with continuous development of cities and the growth of resident construction, the energy consumption is increased in recent years.

The electricity consumption of household is related to type and quantity of household appliances and the Appliances have an influence on the indoor environment, such has temperature, humidity, lights, etc.

Steps:

- Define Problem Statement
- EDA and Feature Engineering
- Feature Selection
- Preparing dataset for modeling
- Apply to model
- Model validation and selection
- Conclusion

Data summary:

The dataset has series of sensors data collected from building in Belgium at interval of 10 minutes for a period of about 4.5 months.

The sensor data consist of temperatures and humidity data of building in different room.

There are sensor that collect data outside of building like pressure, windspeed, visibility and t-dewpoint which is recorded from weather station chievres airport, Belgium.

Data summary:

Data processing :- In this part Removed Unnecessary features.

Data processing :- Go though the each features that are selected from above part and encoded with numerical features.

EDA:- In this I do some Exploratory Data Analysis(EDA) on different features and see the Trend.

Create Model:- In this create some models, I start with simple model and slowly add complexity for better performance.

Data Attributes:

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
RH_1: Humidity in kitchen area, in %
T2: Temperature in living room area, in Celsius
RH_2: Humidity in living room area, in %
T3: Temperature in laundry room area
RH_3: Humidity in laundry room area, in %
T4: Temperature in office room, in Celsius
RH_4: Humidity in office room, in %

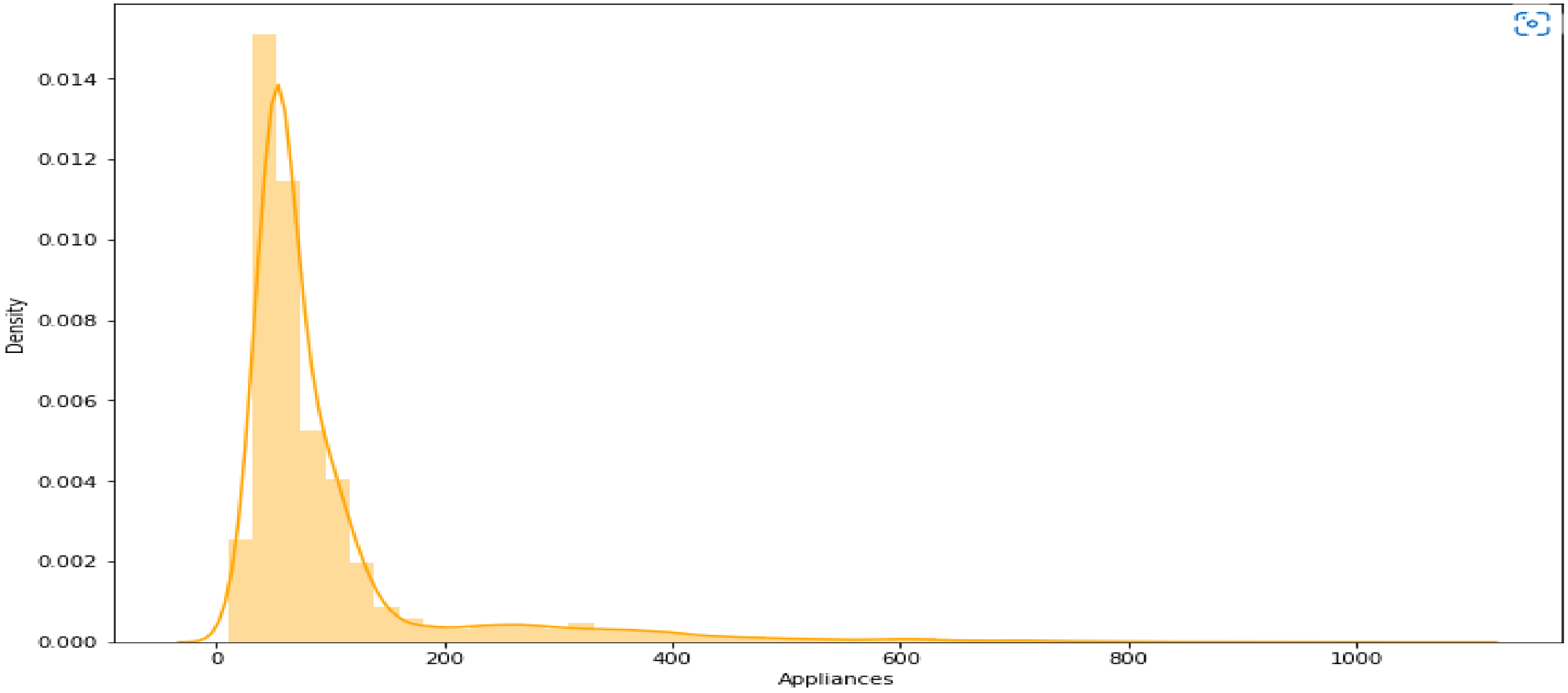
Continue...

T5: Temperature in bathroom, in Celsius
RH_5: Humidity in bathroom, in %
T6: Temperature outside the building (north side), in Celsius
RH_6: Humidity outside the building (north side), in %
T7: Temperature in ironing room , in Celsius
RH_7: Humidity in ironing room, in %
T8: Temperature in teenager room 2, in Celsius
RH_8: Humidity in teenager room 2, in %
T9: Temperature in parents room, in Celsius
RH_9: Humidity in parents room, in %
T_out: Temperature outside (from Chievres weather station), in Celsius

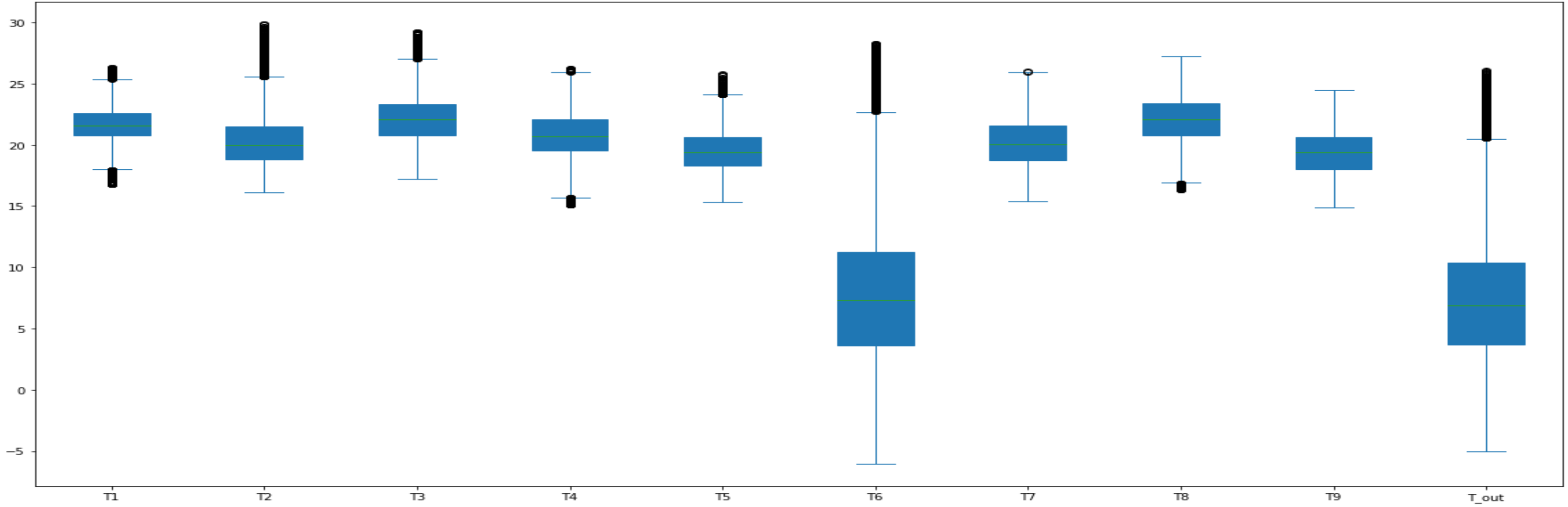
Continue...

Press_mm_hg: Pressure (from Chievres weather station), in mm Hg
RH_out: 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

Dependent variable Distribution:-

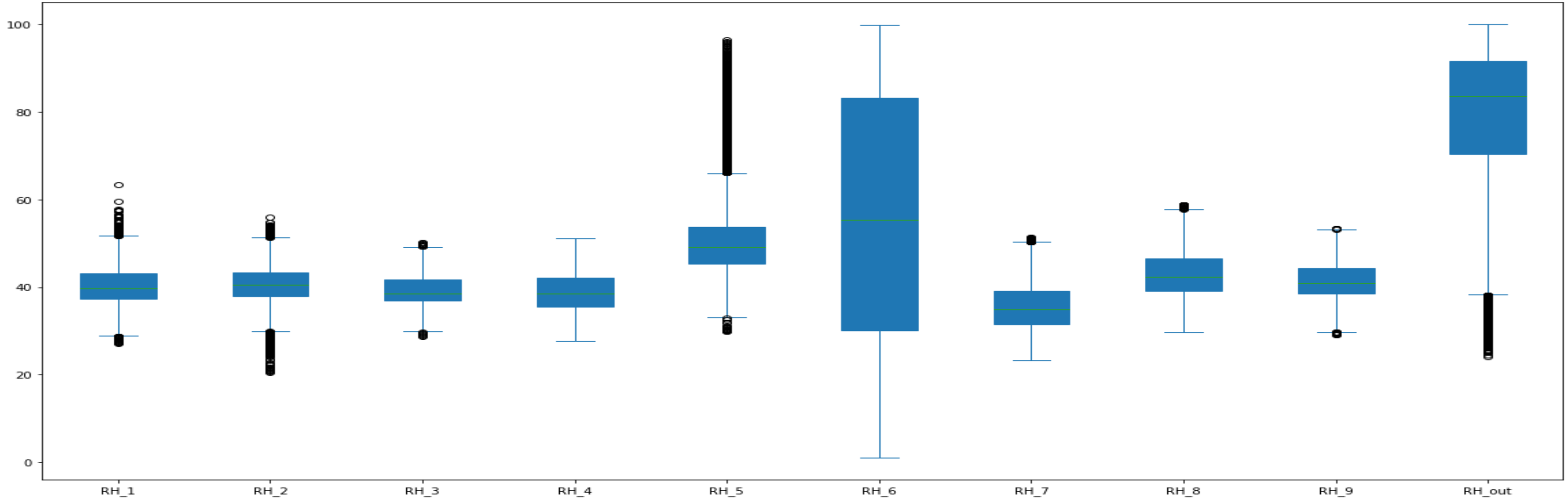


Temperature Box plot:



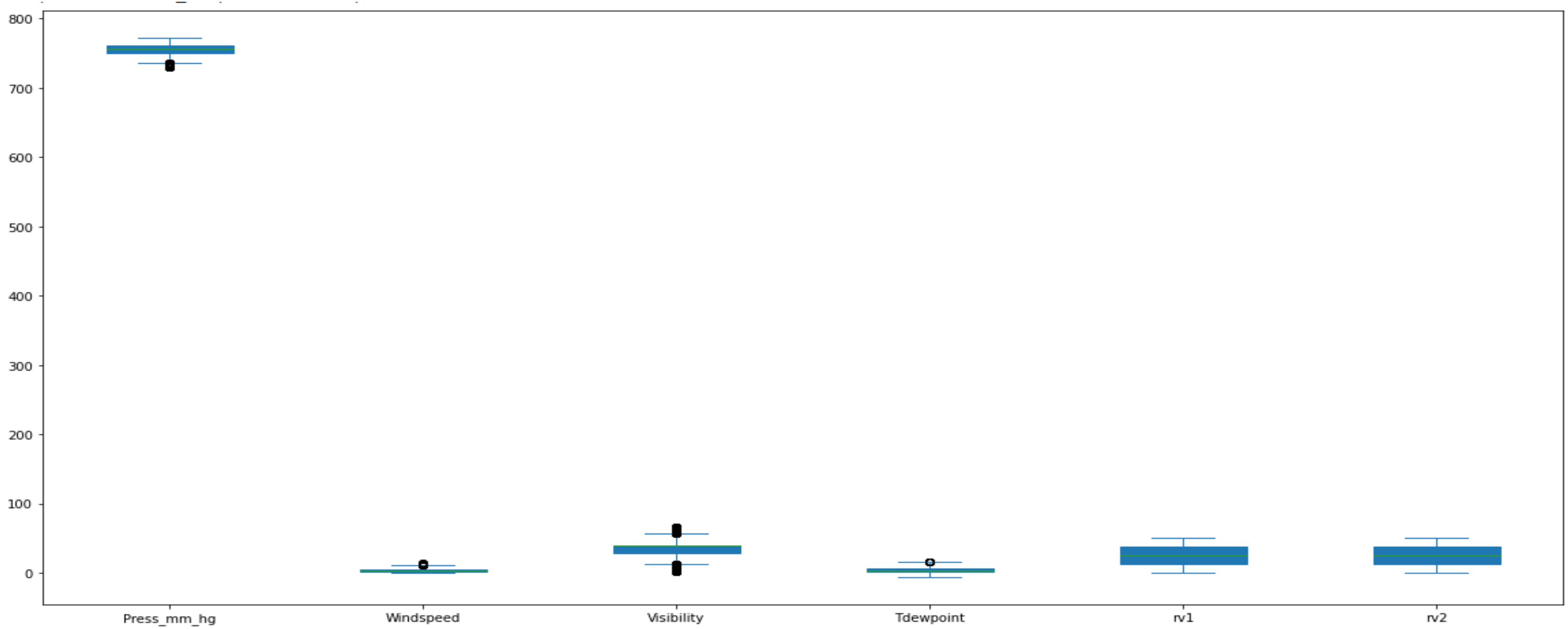
Temperature outside (north side) has min temperature -5 degree and next is T_out(temperature at weather station) is second min temperature.

Humidity Box plot:

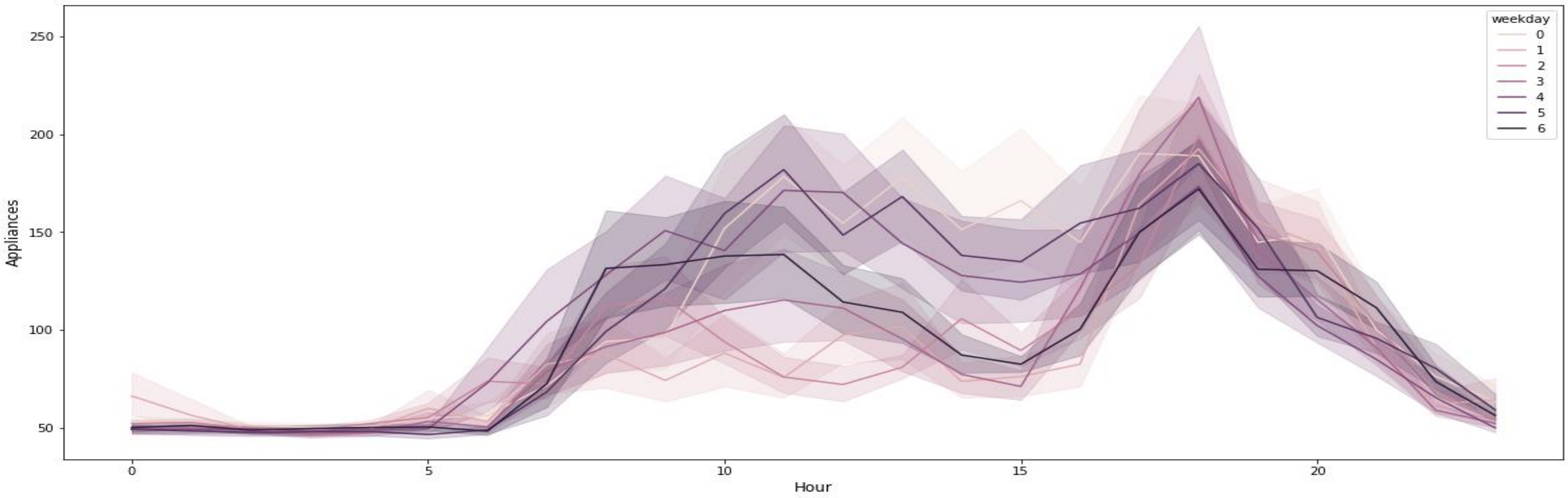


Humidity Outside(north side) has min Humidity and also maximum humidity, while Rh_out(Humidity at weather station) is also maximum.

Weather Box plot:

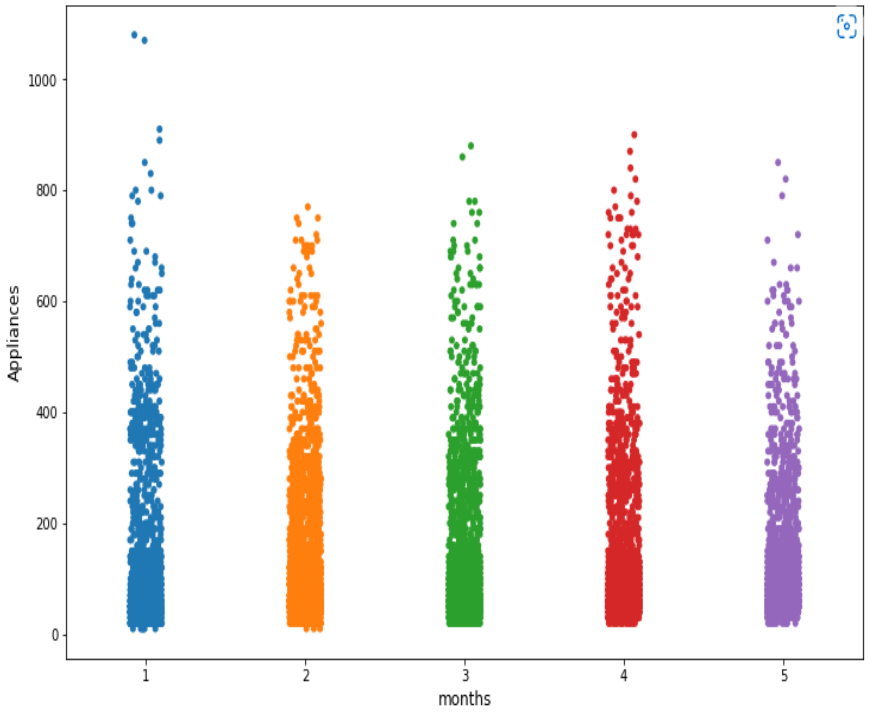


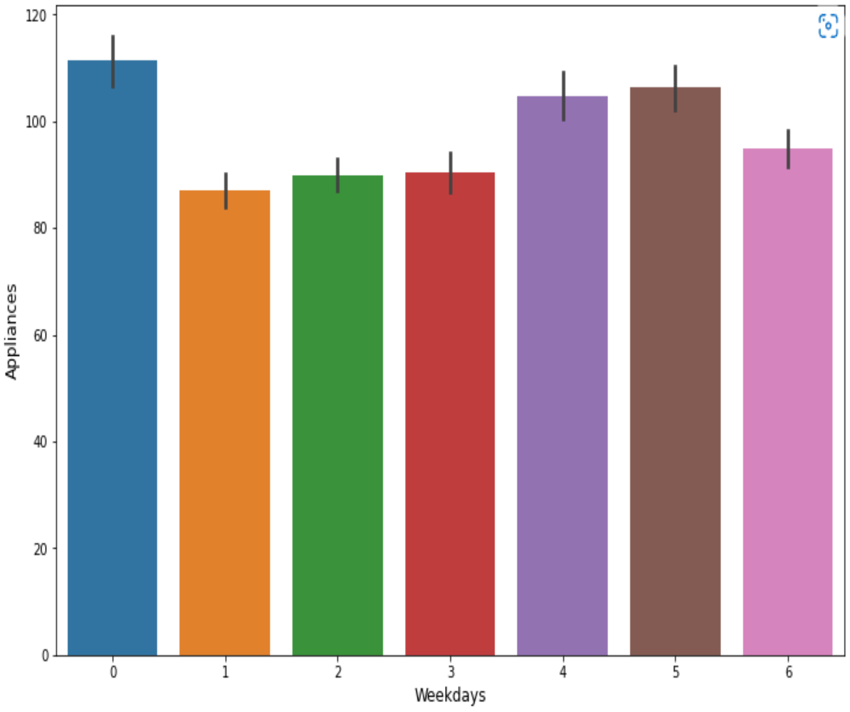
Appliances vs hours:



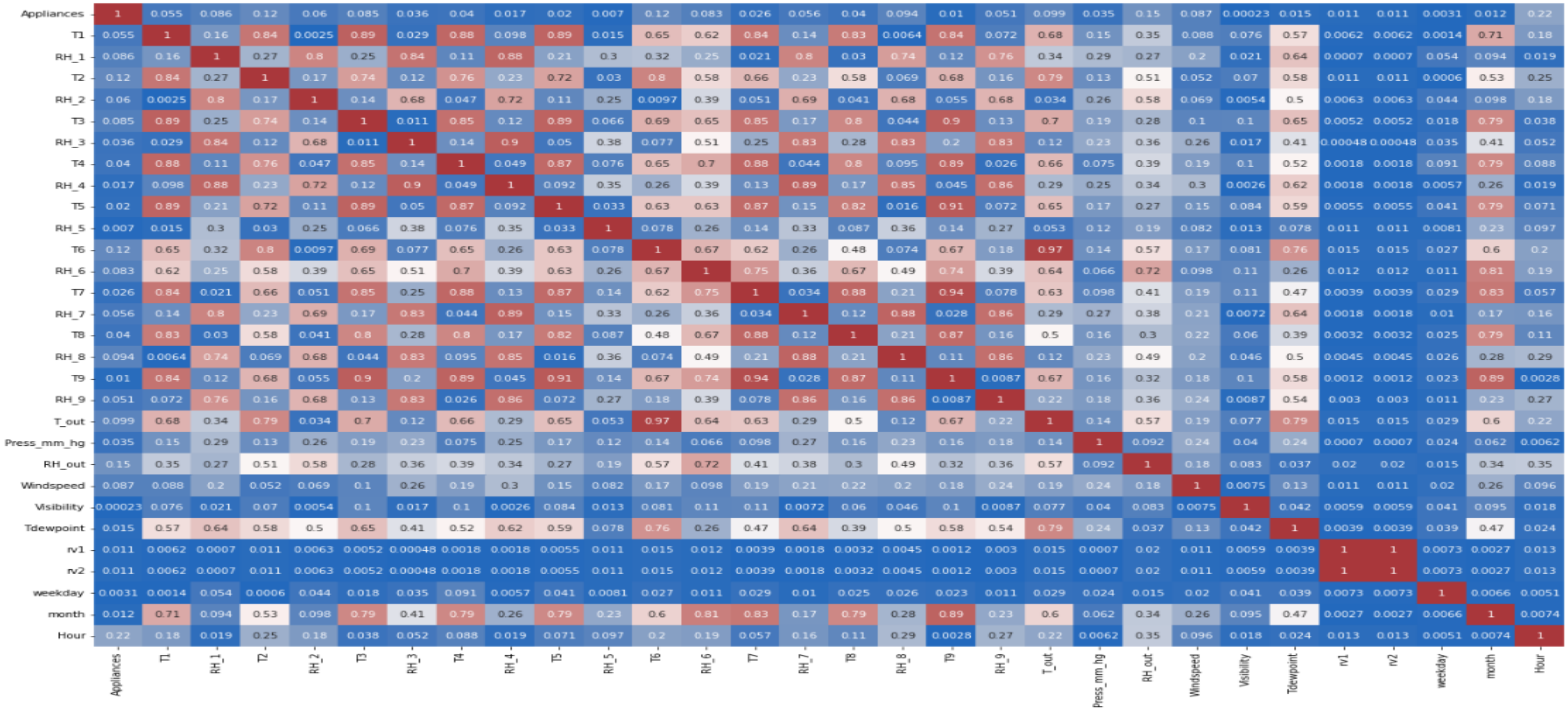
As see in graph Appliances energy consumption is increases after 4 PM and decreases after 7 PM, also understand that in morning at 6 AM to 3 PM energy consumption is moderated.

Appliances vs month and weekdays:





Correlation:



Preparing dataset for Modeling:

Train, test :-(80% and 20%)

Train set :- (15788, 21)

Test set :- (3947, 21)

```
# Import standerdscaler

from sklearn.preprocessing import StandardScaler
scaler=StandardScaler()
x_train=scaler.fit_transform(x_train)
x_test=scaler.transform(x_test)
```

Dependent Variable :- Appliances

The dataset has varying range. Due to different range of features it is possible that some features will dominate the regression algorithm. To avoid this , all feature need to be scaled.

Reduction of feature and multicollinearity:

By using Variance Inflation Factor(VIF) Removed Irrelevant and less correlated features.

Like: rv1, rv2 has infinite VIF so remove that random variables.

	variables	VIF
0	T1	3696.343325
1	RH_1	1671.623725
2	T2	2492.593061
3	RH_2	2166.128604
4	T3	1266.628250
5	RH_3	1594.711214
6	T4	973.109107
7	RH_4	1419.199833
8	T5	1199.624872
9	RH_5	45.913242
10	T6	91.222848
11	RH_6	49.475702
12	T7	1646.451315
13	RH_7	519.852809
14	T8	1002.842397
15	RH_8	632.091594
16	T9	2878.134250
17	RH_9	689.767311
18	T_out	426.761613
19	Press_mm_hg	2162.693341
20	RH_out	1403.216541
21	Windspeed	5.379509
22	Visibility	12.113300
23	Tdewpoint	135.103134
24	rv1	inf
25	rv2	inf
26	weekday	3.584613
27	month	78.534283
28	Hour	7.862823

Model validation and selection:

Applying Linear regression , and Regularized Regression Lasso, Ridge and Elastic Regression.

Then fitting all the model like Random Forest Regressor ,Gradient boosting Etc.

```
from sklearn.linear_model import Lasso,Ridge,ElasticNet
from sklearn.ensemble import RandomForestRegressor,GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
from xgboost import XGBRegressor
from sklearn import neighbors
from sklearn.svm import SVR
```

Model validation and selection(Continue…)

	Name	Train_r2_score	Test_r2_score	Train_MSE_score	Test_MSE_score	Train_RMSE_score	Test_RMSE_score
3	Randomforest :	0.940532	0.546191	0.059468	0.502430	0.243860	0.708823
7	Kneighboursregressor :	0.696022	0.454429	0.303978	0.604023	0.551342	0.777189
4	Gradientboosting :	0.331111	0.230639	0.668889	0.851790	0.817856	0.922925
5	Xgboost :	0.326152	0.236348	0.673848	0.845470	0.820883	0.919494
6	svm :	0.242006	0.196566	0.757994	0.889514	0.870628	0.943141
1	Ridge :	0.135951	0.125259	0.864049	0.968461	0.929543	0.984104
0	Lasso :	0.000000	-0.000371	1.000000	1.107550	1.000000	1.052402
2	ElasticNet :	0.000000	-0.000371	1.000000	1.107550	1.000000	1.052402

Model validation and selection(continue...)

Observation 1:- Lasso and Elasticnet model is giving worst r2 score in this dataset.

Observation 2:- As see in above slide Random forest gives high train r2 score but less test r2 score.

Observation 3:- From above observation Random forest is best model for this dataset.

Model validation and selection(continue...)

Tuning Hyper parameter of Random Forest Regressor and got best parameter and best estimators. and got the r2 score 56% in this dataset.

This is because of low correlation between features and target variable.

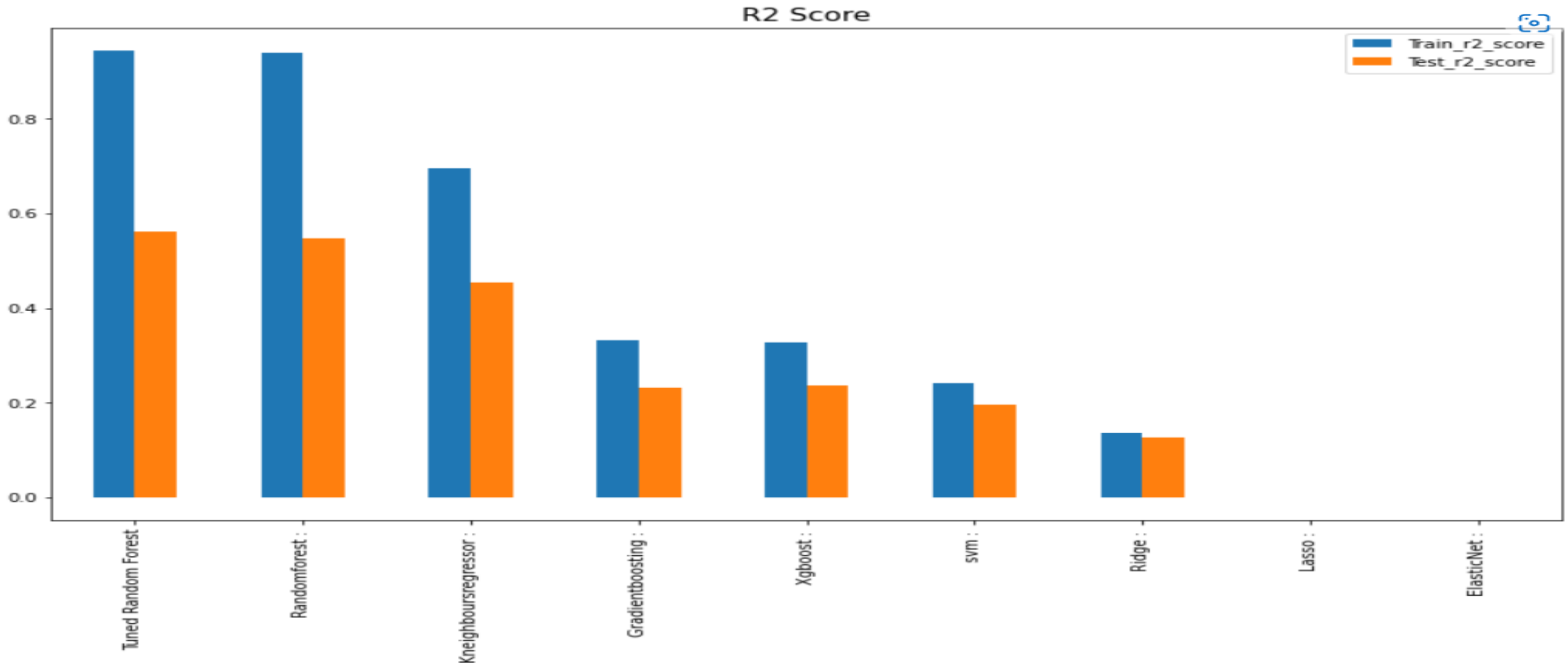
RMSE Value for Random forest regressor is 23% for this dataset.

```
[ ]  1  rf_grid_search.best_params_

{'max_depth': 100, 'max_features': 'sqrt', 'n_estimators': 260}

[ ]  1  rf_grid_search.best_estimator_

RandomForestRegressor(max_depth=100, max_features='sqrt', n_estimators=260,
                      random_state=40)

[ ]  1  y_pred_train=rf_grid_search.best_estimator_.score(x_train,y_train)

[ ]  1  y_pred_train

0.9456129780703171

[ ]  1  y_pred_test=rf_grid_search.best_estimator_.score(x_test,y_test)

[ ]  1  y_pred_test

0.5622271917467065

[ ]  1  Mse_test=(mean_squared_error(y_test,rf_grid_search.best_estimator_.predict(x_test)))

[ ]  1  Mse_test

0.48467559475374034

[ ]  1  np.sqrt(mean_squared_error(y_test,rf_grid_search.best_estimator_.predict(x_test)))

0.6961864655060025

[ ]  1  Mse_train=(mean_squared_error(y_train,rf_grid_search.best_estimator_.predict(x_train)))

[ ]  1  Mse_train

0.05438702192968292

[ ]  1  np.sqrt(mean_squared_error(y_train,rf_grid_search.best_estimator_.predict(x_train)))

0.2332102526255716
```

Model validation and selection(continue...)

Name	Train_r2_score	Test_r2_score	Train_MSE_score	Test_MSE_score	Train_RMSE_score	Test_RMSE_score
Tuned Random Forest	0.945610	0.562220	0.054387	0.484670	0.233210	0.696180
Randomforest :	0.940532	0.546191	0.059468	0.502430	0.243860	0.708823
Kneighboursregressor :	0.696022	0.454429	0.303978	0.604023	0.551342	0.777189
Gradientboosting :	0.331111	0.230639	0.668889	0.851790	0.817856	0.922925
Xgboost :	0.326152	0.236348	0.673848	0.845470	0.820883	0.919494
svm :	0.242006	0.196566	0.757994	0.889514	0.870628	0.943141
Ridge :	0.135951	0.125259	0.864049	0.968461	0.929543	0.984104
Lasso :	0.000000	-0.000371	1.000000	1.107550	1.000000	1.052402
ElasticNet :	0.000000	-0.000371	1.000000	1.107550	1.000000	1.052402

Comparison of all models:



Conclusion:

- Getting Good Result when I selecting 21 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 rest of the algorithms.

- After tuning the algorithm using GridSearchCV on Random forest regressor the score is not getting much difference than the previous, Because of Correlation between dependent and independent variables are very low in this dataset.

Challenges:

- Mostly, Features have low correlation so feature selection is challengeable.
- Most of algorithms doesn't give good score even after feature engineering.

Thank you