Capstone Project

Appliances Energy Prediction

Supervised Machine Learning

Individual project by
Priyanka Shinde

Problem Statement

Predict :- Appliance energy consumption for a house based on factors like temperature , humidity & pressure

Data Exploration

#	Column	Non-Null Count	Dtype
0	date	19735 non-null	object
1	Appliances	19735 non-null	int64
2	lights	19735 non-null	int64
3	T1	19735 non-null	float64
4	RH_1	19735 non-null	float64
5	T2	19735 non-null	float64
6	RH_2	19735 non-null	float64
7	T3	19735 non-null	float64
8	RH_3	19735 non-null	float64
9	T4	19735 non-null	float64
10	RH_4	19735 non-null	float64
11	T5	19735 non-null	float64
12	RH_5	19735 non-null	float64
13	T6	19735 non-null	float64
14	RH_6	19735 non-null	float64
15	T7	19735 non-null	float64
16	RH_7	19735 non-null	float64
17	T8	19735 non-null	float64
18	RH_8	19735 non-null	float64
19	T9	19735 non-null	float64
20	RH_9	19735 non-null	float64
21	T_out	19735 non-null	float64
22	Press_mm_hg	19735 non-null	float64
23	RH_out	19735 non-null	float64
24	Windspeed	19735 non-null	float64
25	Visibility	19735 non-null	float64
26	Tdewpoint	19735 non-null	float64
27	rv1	19735 non-null	float64
28	rv2	19735 non-null	float64

dtypes: float64(26), int64(2), object(1)

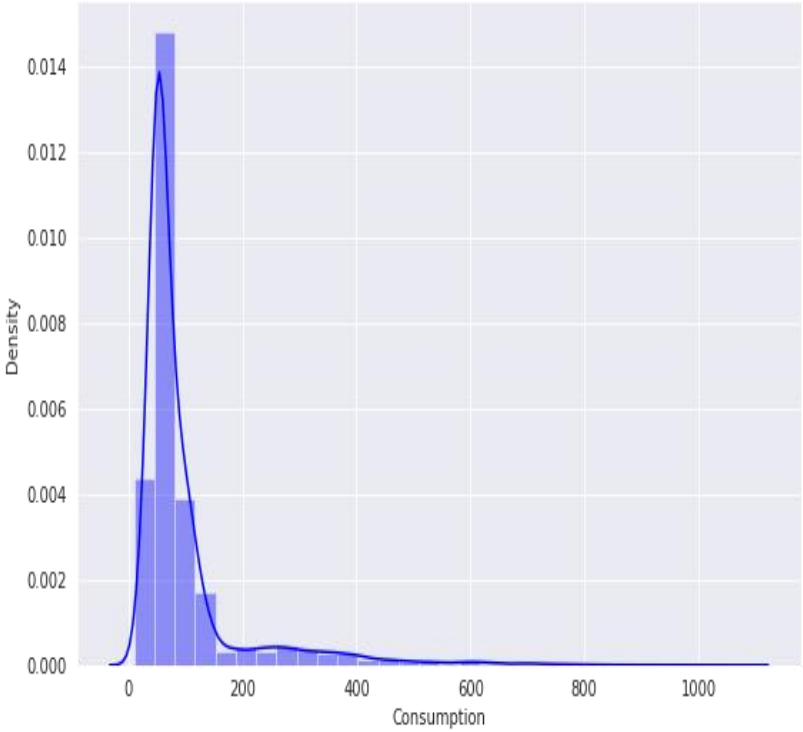
- **The dataset consists of 28 numeric columns and one object column(“date”)**

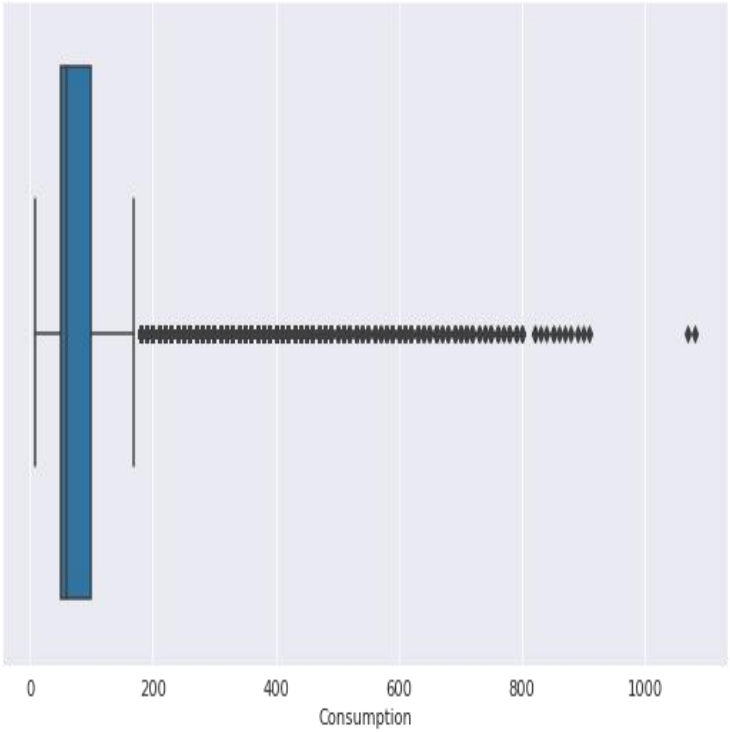
Dataset Shape

```
The number of rows in dataset is =  19735
The number of columns in dataset is =  29
```

- **Dataset contain ZERO NaN and duplicates values.**

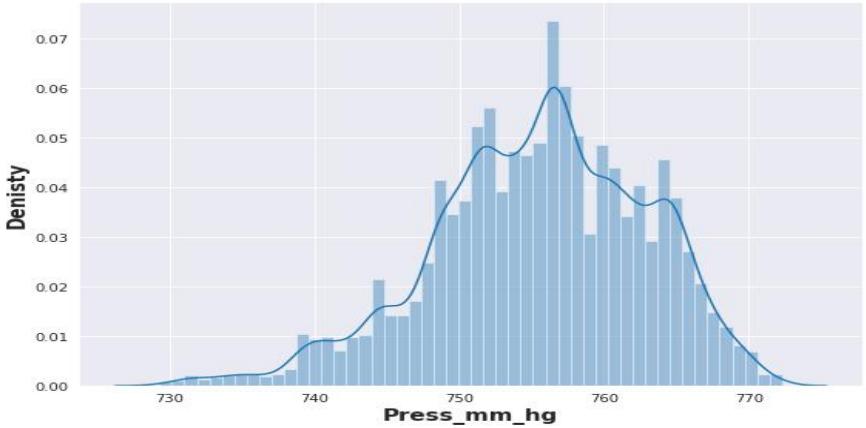
Dependent Variable

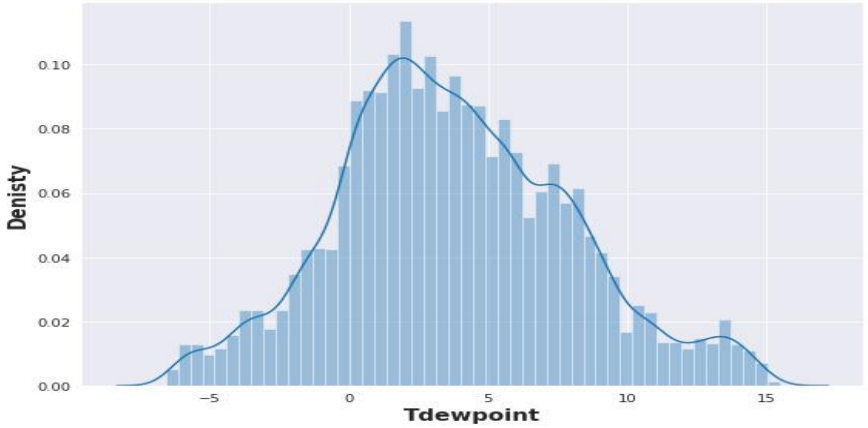


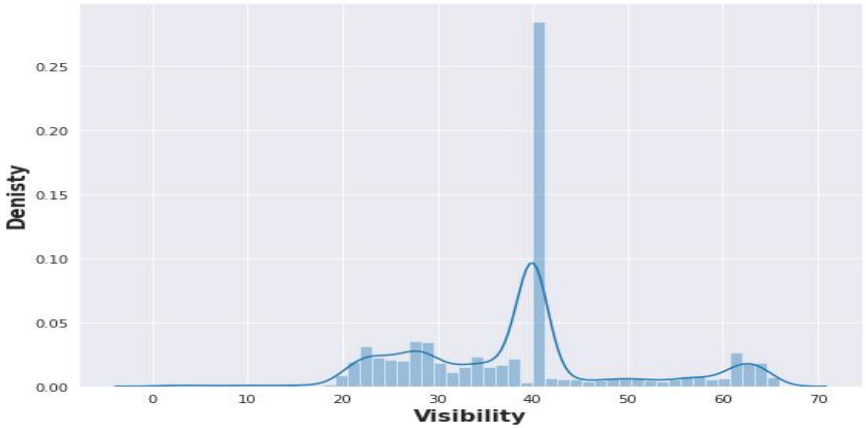


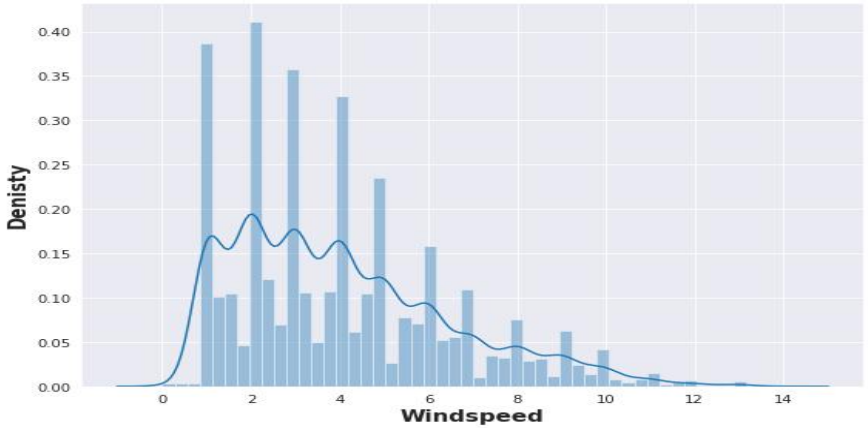
- we see that the 75% of Appliance consumption is less than 100 Wh .
- There are small number of cases where consumption is very high.

Independent Variable

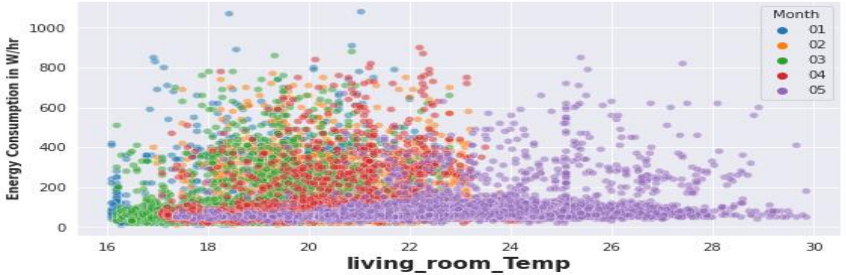


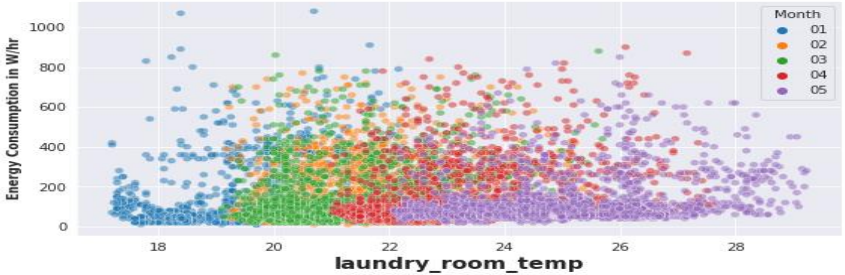


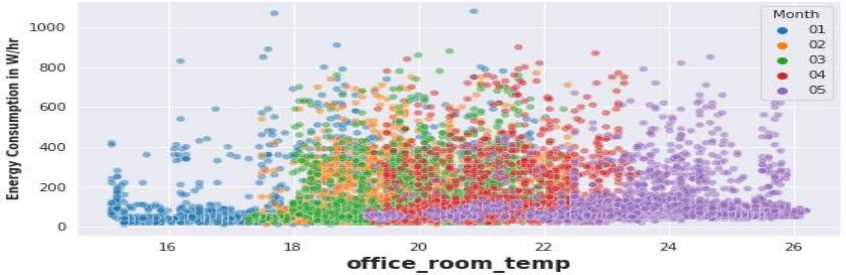


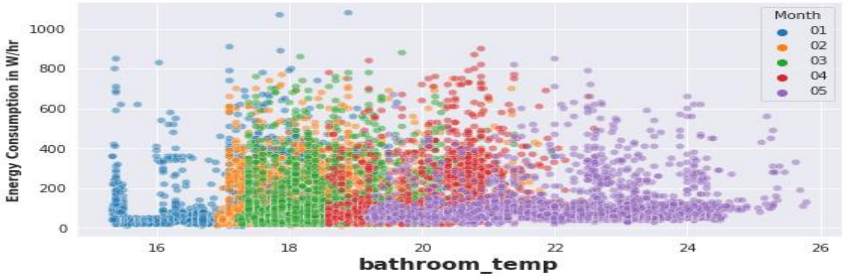


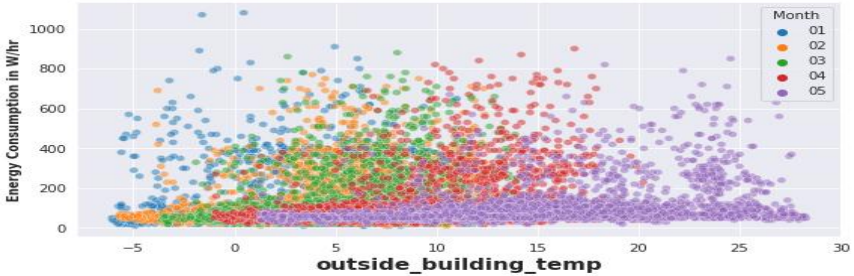
Energy consumption & Temperature



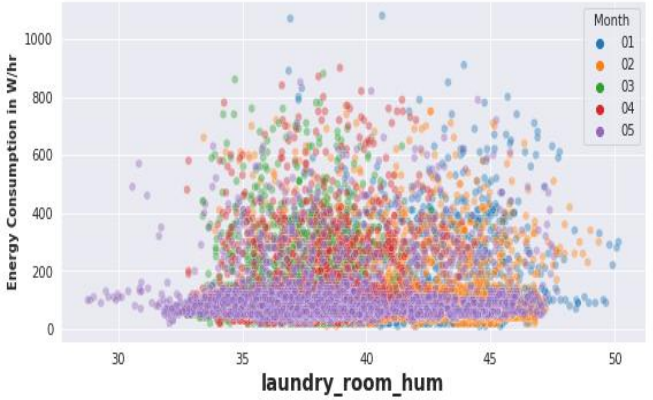


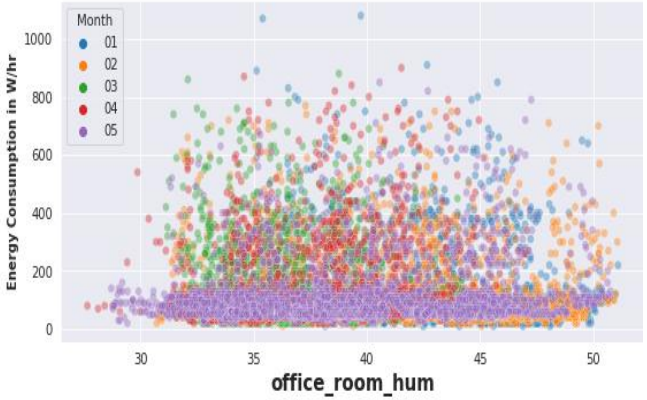




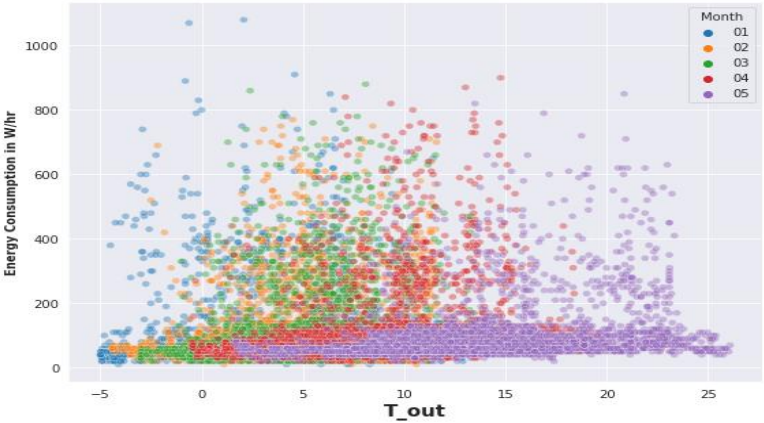


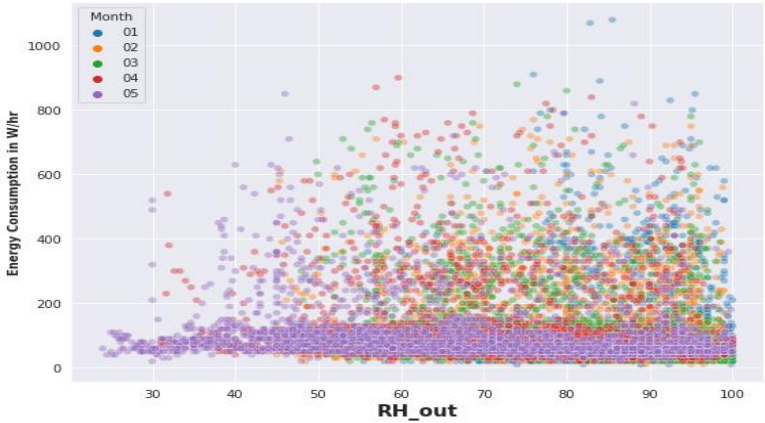
Energy Consumption & humidity

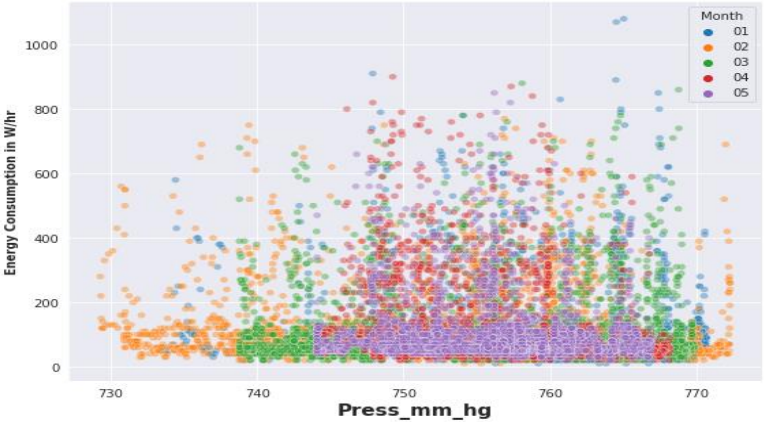


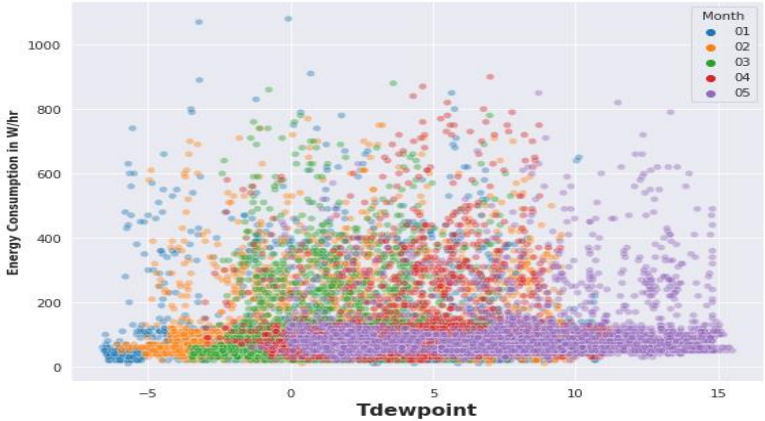


Outiside:- Temperature,Pressure ,Humidity

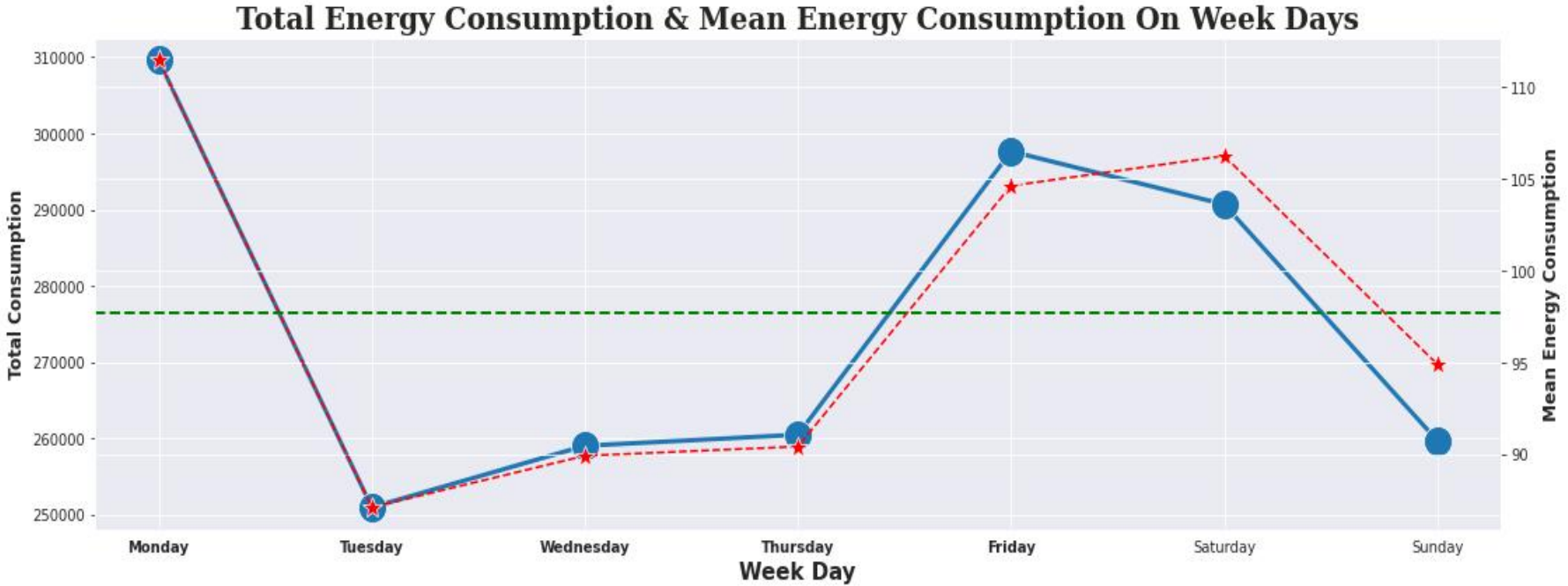






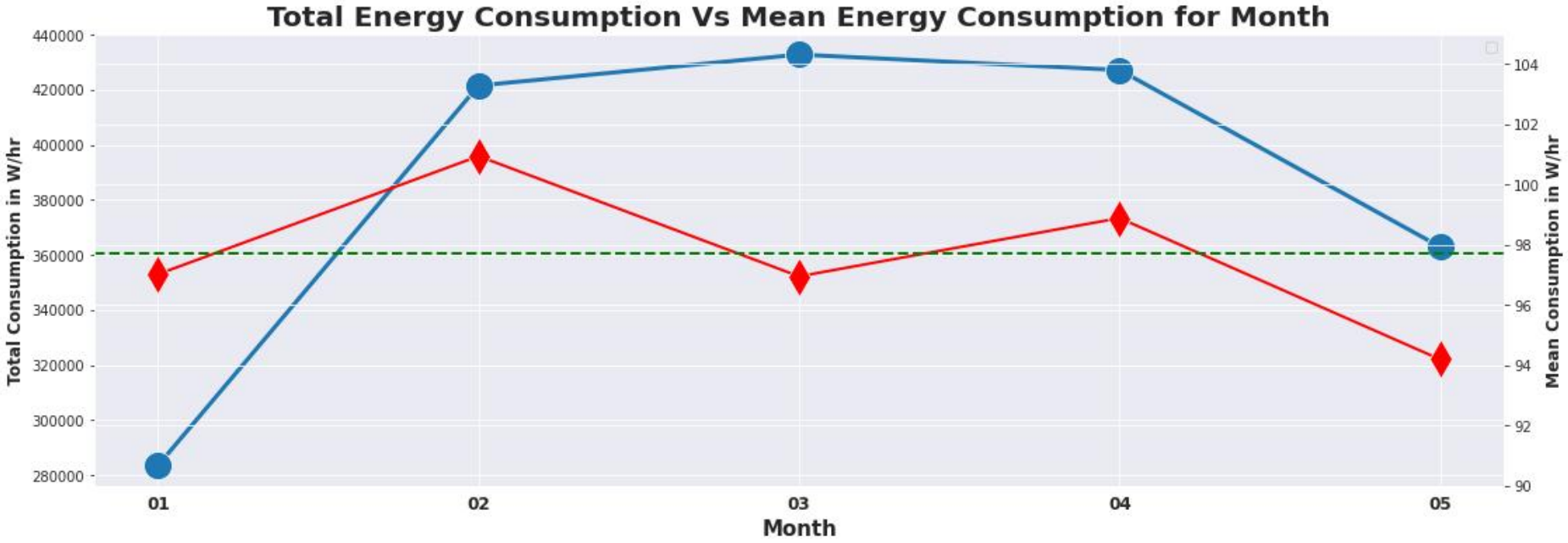


Energy consumption per day



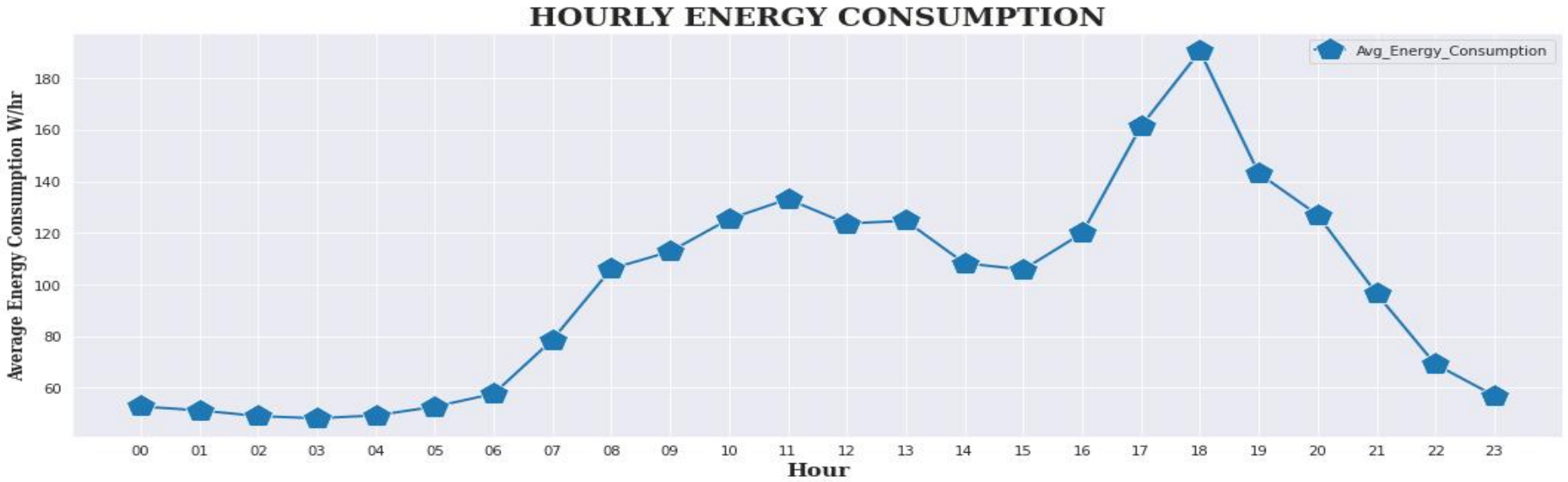
- **Monday has the highest total energy consumption as well as the highest mean energy consumption.**
- **Tuesday, Wednesday, Thursday, and Friday have a mean energy consumption below overall mean consumption and Monday, Friday, and Saturday have a mean energy consumption over overall mean consumption.**

Energy consumption per month



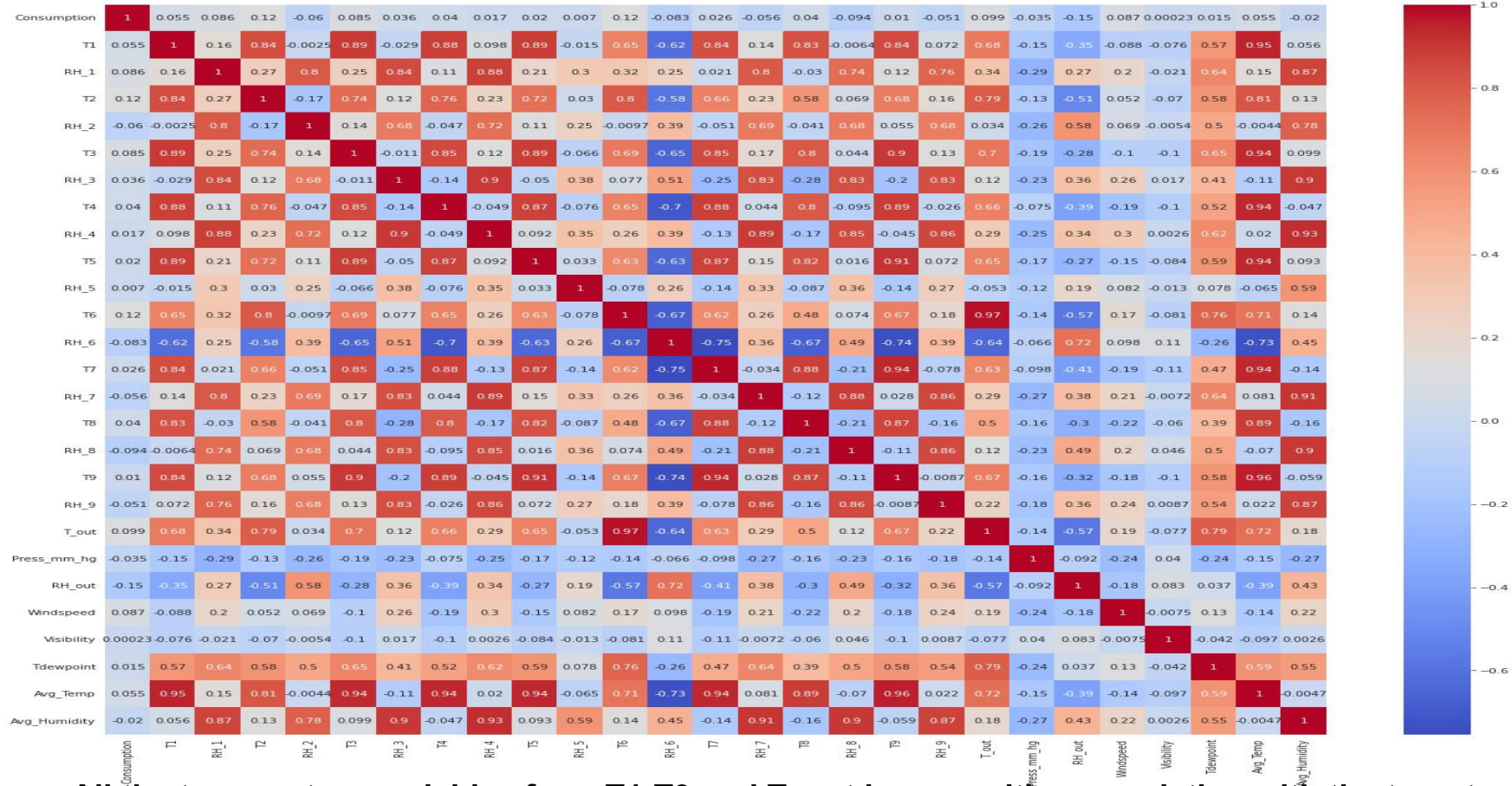
- **February had the highest average of 101 watts per hour, while May had the lowest average of 94 watts per hour.**

Energy consumption per hour



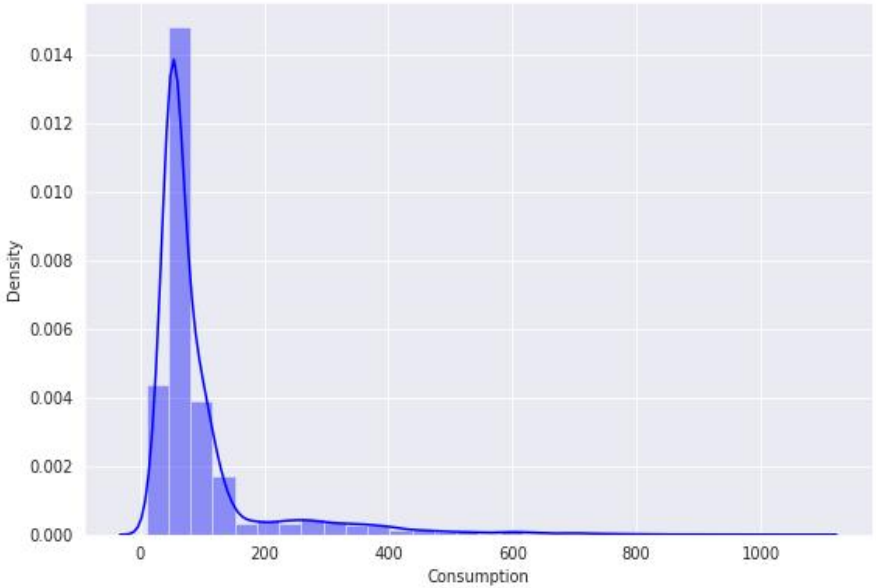
- **The consumption of energy from late night to early morning is very low. This is because appliances are used less during the night.**
- **The consumption of energy from morning until evening is moderate.**
- **In the evening, energy consumption is highest.This is because appliances are used more during the evening.**

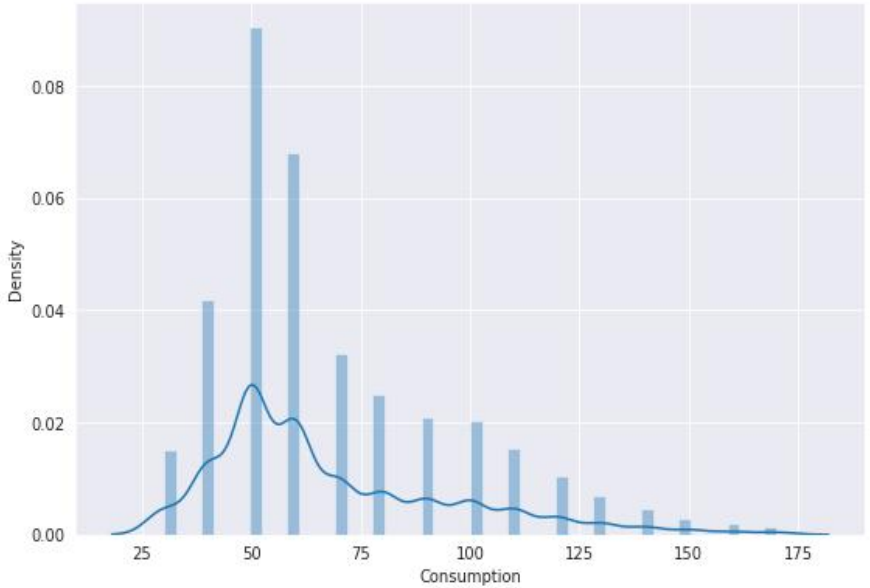
Correlation



- **All the temperature variables from T1-T9 and T_out have positive correlation with the target Appliances**

Outliers





- **Data shape before outlier removal (19735, 32) and after outliers removal (17245, 32).**
- **using IQR we can remove most of the outliers from the data.**

Data preprocessing

- Data destribution for dependable variable Consumption is positively skewed.
- By using Log Transformation we will Normalize the data.
- We have 4.5 months data and dates are in 'Date-Time' format.For computational purpose we will create few new columns for Day,Time,Month,Date,Hour.

Encoding

```
# Dictionary of  week days with numerical values as per the importance based on the avergae consumption
week_days={'Tuesday':0,'Wednesday':0,'Thursday':0,'Sunday':0,'Friday':1,'Saturday':1,'Monday':2}
```

- Using Standard scalar we scale the datapoints.

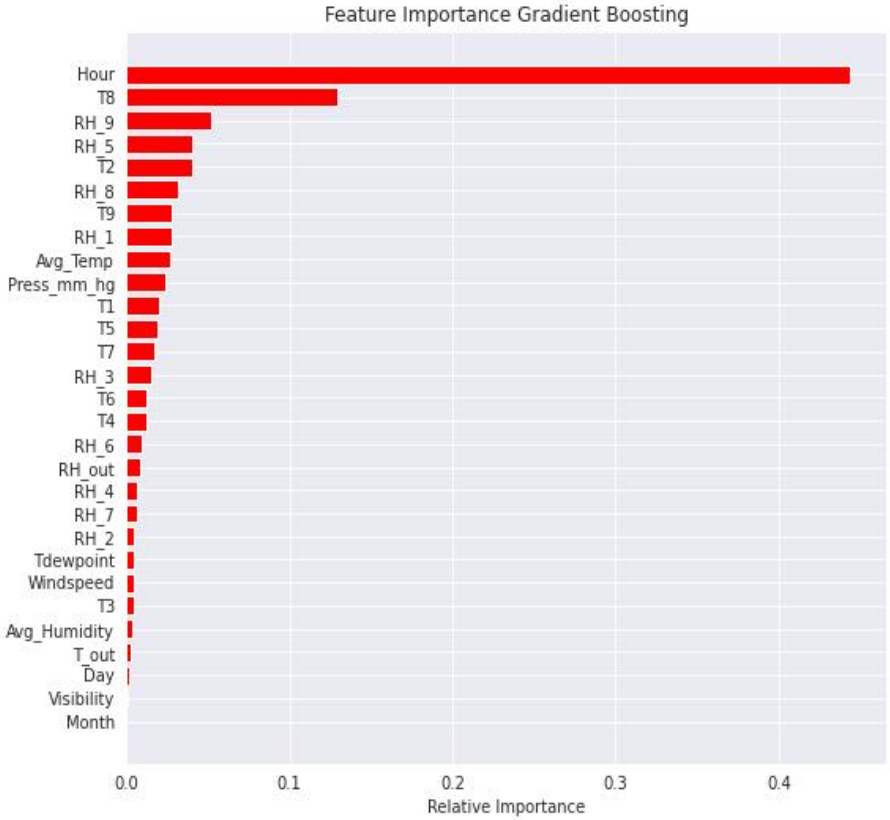
BASELINE MODELS

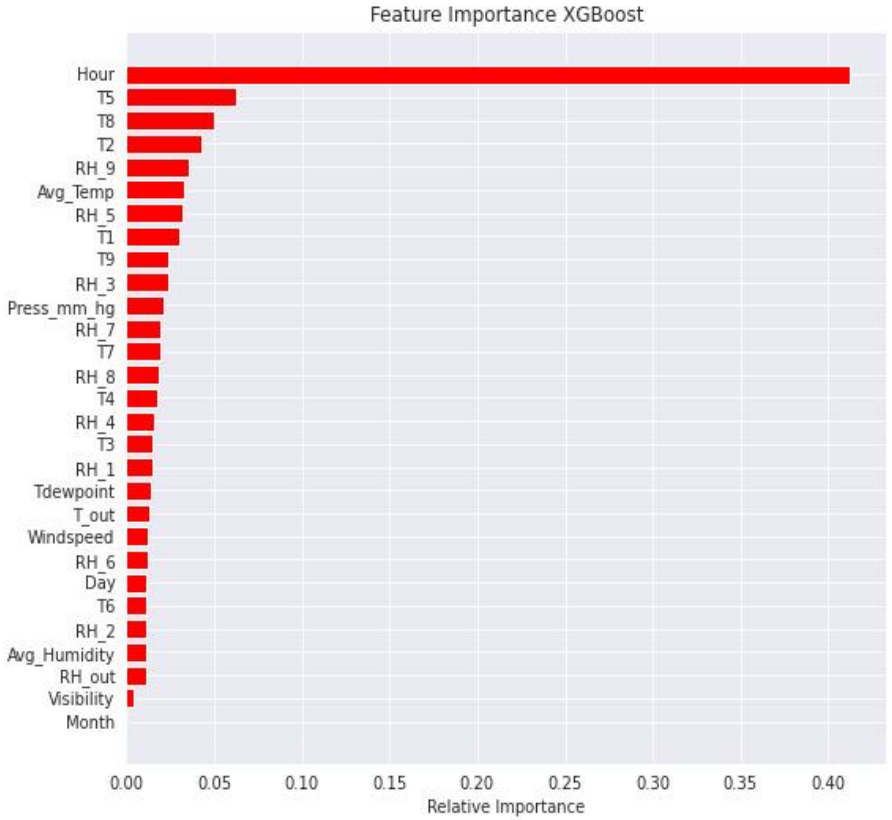
	Model_Name	Train_R2	Test_R2	Train_RMSE	Test_RMSE	Train_MAE	Test_MAE
0	LinearRegression	0.398245	0.405373	21.722216	21.353702	15.165678	15.124388
1	Lasso	-0.030882	-0.032431	28.431429	28.137222	21.112709	21.093452
2	Ridge	0.398246	0.405413	21.722201	21.352975	15.165525	15.122861
3	KNeighborsRegressor	0.768899	0.650593	13.461551	16.368785	8.690979	11.076429
4	SVR	0.635092	0.616631	16.915544	17.145869	11.306851	11.853455
5	RandomForestRegressor	0.946289	0.710631	6.489699	14.896239	3.927154	10.052209
6	GradientBoostingRegressor	0.510505	0.491573	19.591524	19.745360	13.486359	13.816424
7	XGBRegressor	0.505521	0.488247	19.691012	19.809843	13.548481	13.890786
8	LGBMRegressor	0.711819	0.639166	15.032354	16.634290	10.154634	11.535916

- **In order to minimize computational time, baseline models are created to give a general understanding of performance.**

R2 AND RMSE SCORE

Feature importance





HYPER-PARAMETER TUNING

	Model_Name	Train R2	Test R2	Train RMSE	Test RMSE	Train MAE	Test MAE
1	GBMRegressor_with_hyper	0.677	0.599	16.169	17.543	11.048	12.216

	Model_Name	Train R2	Test R2	Train RMSE	Test RMSE	Train MAE	Test MAE
1	SVMRegressor_with_hyper	0.809	0.675	11.976	15.919	8.081	10.715

	Model_Name	Train R2	Test R2	Train RMSE	Test RMSE	Train MAE	Test MAE
1	XGBRegressor_with_hyper	0.871	0.684	10.151	15.471	6.755	10.546

Conclusion

AI

- The Linear Regression model has performed extremely poorly. This is because there is no significant linear relationship between the dependable variable and the independent variable.

- As compared to other models, the Random Forest Regression model has performed best with the highest R*2 scores in both Testing and Training.

- It is also evident from the result that the Random Forest Regression model has more variance as the difference between testing and training r*2 is greater.

- K-Nearest Neighbors Regression model, Support Vector Regression model, and Light Gradient Boosting Machine model also performed well and have low bias and variance compared to the Random Forest Regressor model.

- A Gradient Boosting Regression model (GBM) or the Extreme Gradient Boosting Regression model (XGB) also poorly performed.

- For hyperparameter tuning i used only three models based on the low RMSE SCORE.
- After performing tuning on 1)Gradient Boosting Regression model (GBM),2) Extreme Gradient Boosting Regression model (XGB),3)Support Vectore Machine(SVM) models we observe that the model Extreme Gradient Boosting Regression model (XGB) is the best with low RMSE and High R2 SCORE as compared to other models.

Conclusion

- The top 3 important features are humidity attributes, which leads to the conclusion that humidity affects power consumption more than temperature. Windspeed is least important as the speed of wind doesn't affect power consumption inside the house. So controlling humidity inside the house may lead to energy savings.

- Monday has the highest total energy consumption and tuesday has the lowest.

- The number of days varied from month to month, therefore, the amount of energy consumed varied widely. February had the highest average of 101 watts per hour, while May had the lowest average of 94 watts per hour.

- The consumption of energy from late night to early morning is very low. This is because appliances are used less during the night.

- In the evening, energy consumption is highest.This is because appliances are used more during the evening.

- Windspeed:-On Average windspeed is higher during the day than the night and early morning.
- All the temperature variables from T1-T9 and T_out have positive correlation with the target.
- According to best fit model , the 5 most and least important features

Thank you