

## Capstone Project 1: Final

### Appliances Energy Prediction -

Business Problem Description – Dataset contains the house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. As per the description on UCI website, 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. Combining this data with the weather data based on the date time columns

Link for the Jupyter file -

[https://github.com/arijitsinha80/Springboard/blob/master/Project/Capstoneproject\\_Phase2\\_Final.ipynb](https://github.com/arijitsinha80/Springboard/blob/master/Project/Capstoneproject_Phase2_Final.ipynb)

Link for PPT –

[https://github.com/arijitsinha80/Springboard\\_CapstoneProject1/blob/master/Capstone%20Project%201\\_Final.pptx](https://github.com/arijitsinha80/Springboard_CapstoneProject1/blob/master/Capstone%20Project%201_Final.pptx)

### Dataset Details –

We have two data sets - **energydata\_complete.csv** and **CrudeOilPrice.csv**. We have taken two different dataset to get better prediction with analyzing the engorge consumed and how was the fuel price during the particular date.

We do not have any missing values in **energydata\_complete.csv**; it has 19735 observation with 29 attributes pertaining to temperature, humidity, light, wind speed, dew, and visibility from local weather channel.

We do not have any missing value in **CrudeOilPrice.csv**, which has the fuel price for respective months and dates. This dataset has 2519 observation and 2 attributes of date and fuel price.

Few Key observation are as below –

1. The dataset is from 2016-01-11 and 2016-05-27; have data starting JAN to MAY of 2016.
2. These are the temperature reading captured inside and outside the house. From the explored reading of each sensor is between 14.89 and 29.85 but '**T6**' is between -6 and 28.29. The possible reason can be its reading are for outside.

	T1	T2	T3	T4	T5	\
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	
mean	21.686571	20.341219	22.267611	20.855335	19.592106	
std	1.606066	2.192974	2.006111	2.042884	1.844623	
min	16.790000	16.100000	17.200000	15.100000	15.330000	
25%	20.760000	18.790000	20.790000	19.530000	18.277500	
50%	21.600000	20.000000	22.100000	20.666667	19.390000	
75%	22.600000	21.500000	23.290000	22.100000	20.619643	
max	26.260000	29.856667	29.236000	26.200000	25.795000	
	T6	T7	T8	T9		
count	19735.000000	19735.000000	19735.000000	19735.000000		
mean	7.910939	20.267106	22.029107	19.485828		
std	6.090347	2.109993	1.956162	2.014712		
min	-6.065000	15.390000	16.306667	14.890000		
25%	3.626667	18.700000	20.790000	18.000000		
50%	7.300000	20.033333	22.100000	19.390000		
75%	11.256000	21.600000	23.390000	20.600000		
max	28.290000	26.000000	27.230000	24.500000		

3. There are Humidity related information as well in the dataset, from the explored reading of each sensor is between 20.46 to 58.79 but 'RH\_5' and 'RH\_6' has max of 96.32 and 99.9.

	RH_1	RH_2	RH_3	RH_4	RH_5	\
count	19735.000000	19735.000000	19735.000000	19735.000000	19735.000000	
mean	40.259739	40.420420	39.242500	39.026904	50.949283	
std	3.979299	4.069813	3.254576	4.341321	9.022034	
min	27.023333	20.463333	28.766667	27.660000	29.815000	
25%	37.333333	37.900000	36.900000	35.530000	45.400000	
50%	39.656667	40.500000	38.530000	38.400000	49.090000	
75%	43.066667	43.260000	41.760000	42.156667	53.663333	
max	63.360000	56.026667	50.163333	51.090000	96.321667	
	RH_6	RH_7	RH_8	RH_9		
count	19735.000000	19735.000000	19735.000000	19735.000000		
mean	54.609083	35.388200	42.936165	41.552401		
std	31.149806	5.114208	5.224361	4.151497		
min	1.000000	23.200000	29.600000	29.166667		
25%	30.025000	31.500000	39.066667	38.500000		
50%	55.290000	34.863333	42.375000	40.900000		
75%	83.226667	39.000000	46.536000	44.338095		
max	99.900000	51.400000	58.780000	53.326667		

4. The max value is 1080wh, whereas 75% of usage is under 100wh. Some of the appliances has high consumption. These can be outliers but, currently keeping them as part of the dataset and not dropping them from the dataset.

5. If we see the statistics for Appliance Attributes, the minimum value is 10 and max value is 1080, and the mean is 97.69 and 75% of records are below 100 KWH. This column has outliers and we will keep them and check during our modeling.

	Appliances
count	19735.000000
mean	97.694958
std	102.524891
min	10.000000
25%	50.000000
50%	60.000000
75%	100.000000
max	1080.000000

When merging the two datasets, in energydata dataset, date is a timestamp and in crudeoilprice dataset, date is a date datatype, so we have to normalize the date, in order for us to merge the two datasets.

1. After the merge, we observe that "values" columns is merged on the dataset, but it doesn't have all the dates values and 5904 records has null values.

```
rv2          19735 non-null float64  
value        13831 non-null float64  
dtypes: float64(27), int64(2), object(1)  
memory usage: 4.7+ MB  
None
```

To solve these null values, we used the “**forward fill**” method and value column was populated with previous day values for the records, which were null and renamed the column to “oilprice”.

The total number of observation is 19735 and 30 Attributes.

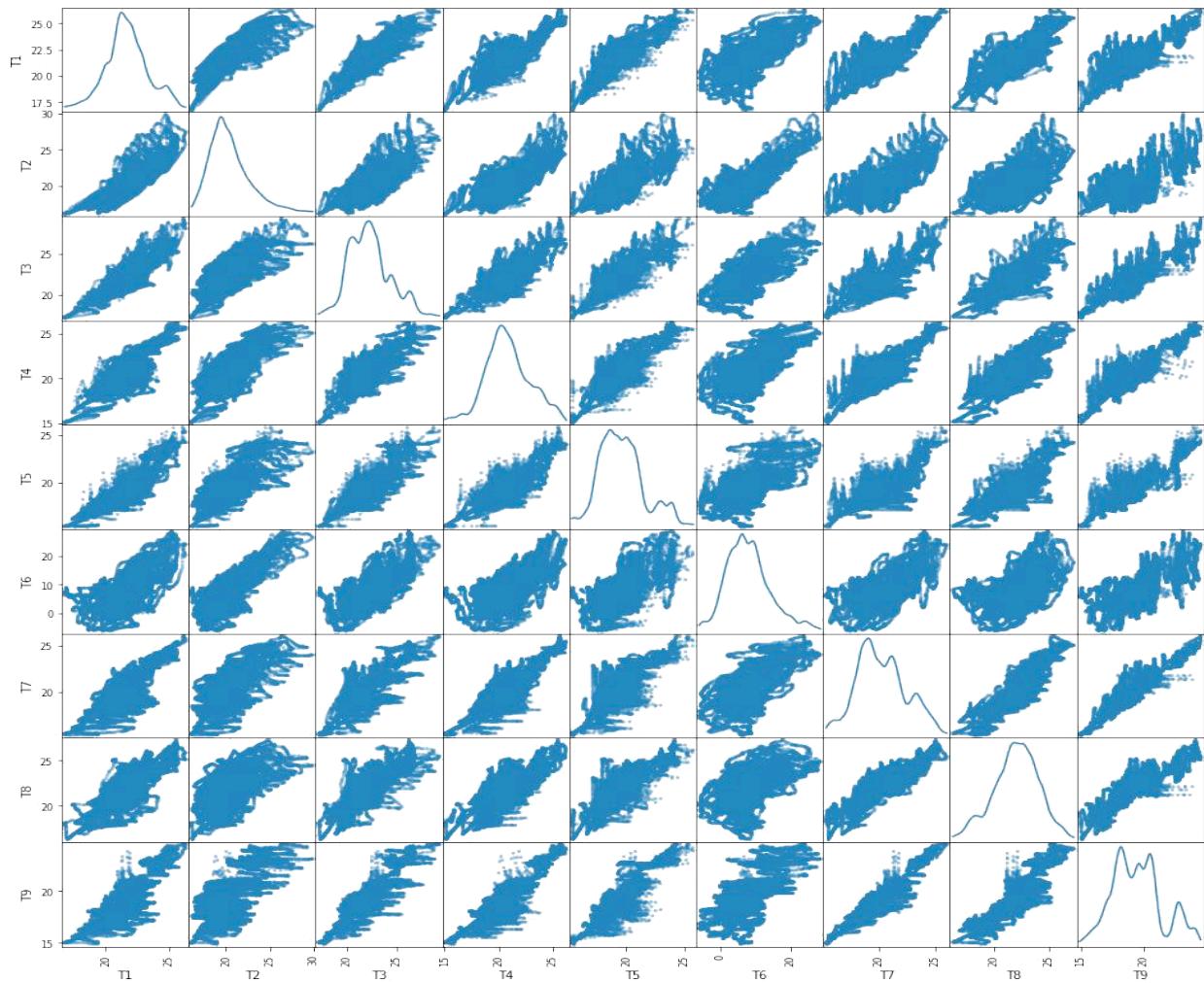
From the Data Wrangling activity, we created the **input.csv** as the final dataset. This has 19735 observations and 30 attributes.

Divide the data in dimension wise to explore from the input dataset –

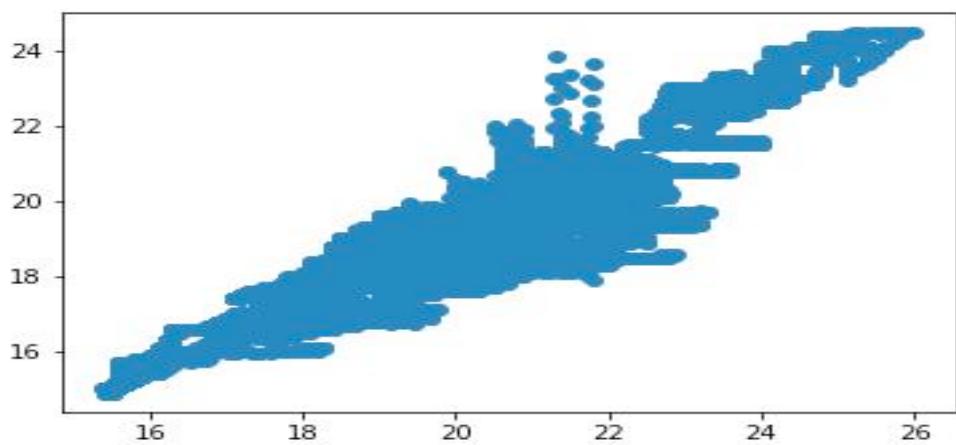
```
# Temperature sensors columns  
temp_cols = ["T1", "T2", "T3", "T4", "T5", "T6", "T7", "T8", "T9"]  
  
# Humidity sensors columns  
hum_cols = ["RH_1", "RH_2", "RH_3", "RH_4", "RH_5", "RH_6", "RH_7", "RH_8",  
, "RH_9"]  
  
# Weather data columns  
wth_cols = ["T_out", "Tdewpoint", "RH_out", "Press_mm_hg", "Windspeed", "Visiblity"]  
  
# Target column  
tgt = ["Appliances"]
```

From the above dimensions, we will start to explore data for each – Plot the scatter matrix for Temperature attributes using method “ **diagonal="kde"** ”

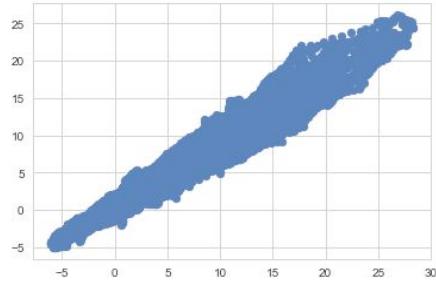
Plot the scatter matrix for Temperature attributes using method “ **diagonal="kde"** ”



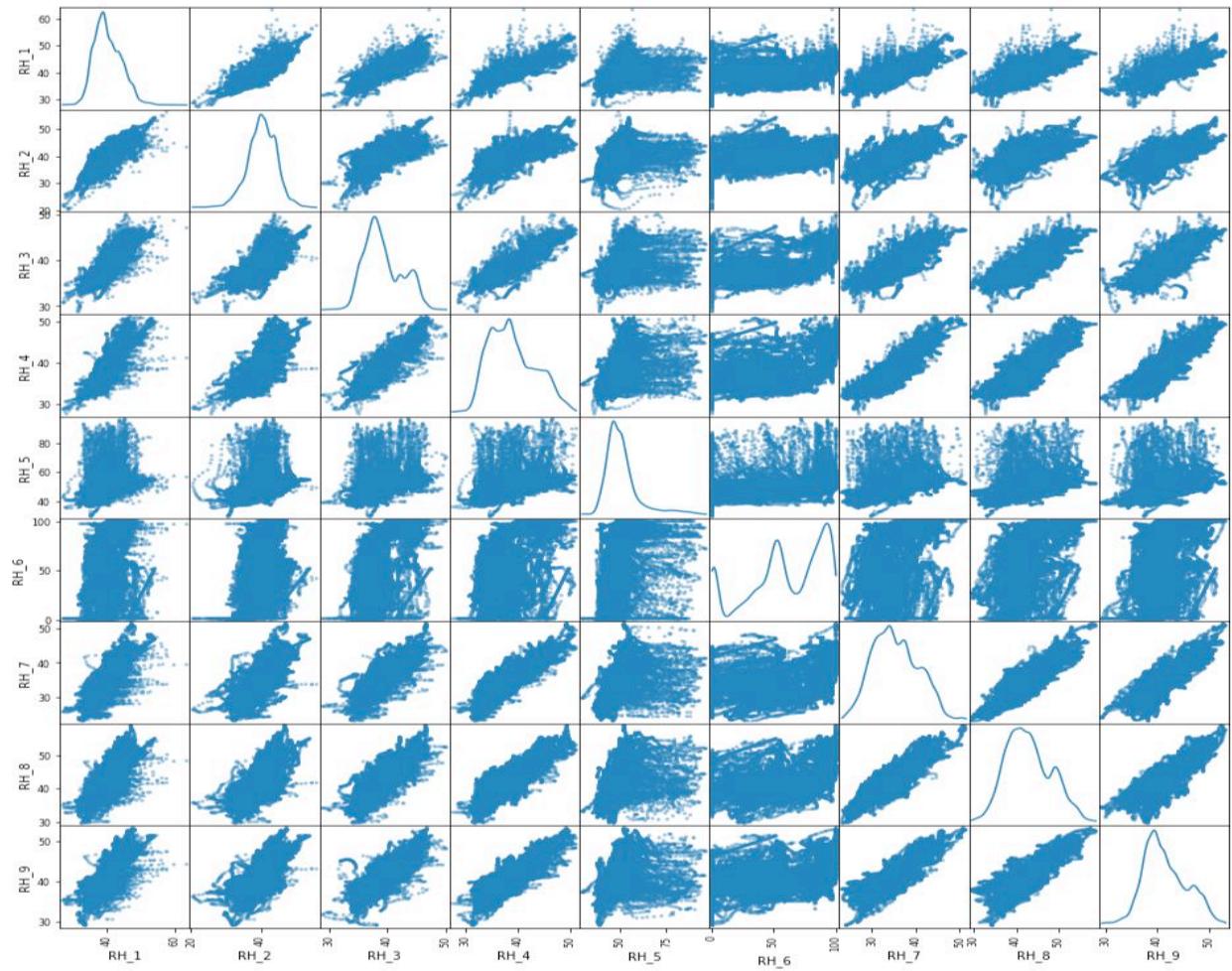
- From the above figure, we can see that there is some linear relation between T7 and T9. Others are having the shape but are not exactly linear.
- There is a relation between these two attributes but also have some outliers



T6 and T\_out is highly correlated, T6 is from the outside the house reading and T\_out is the data collected from weather's site

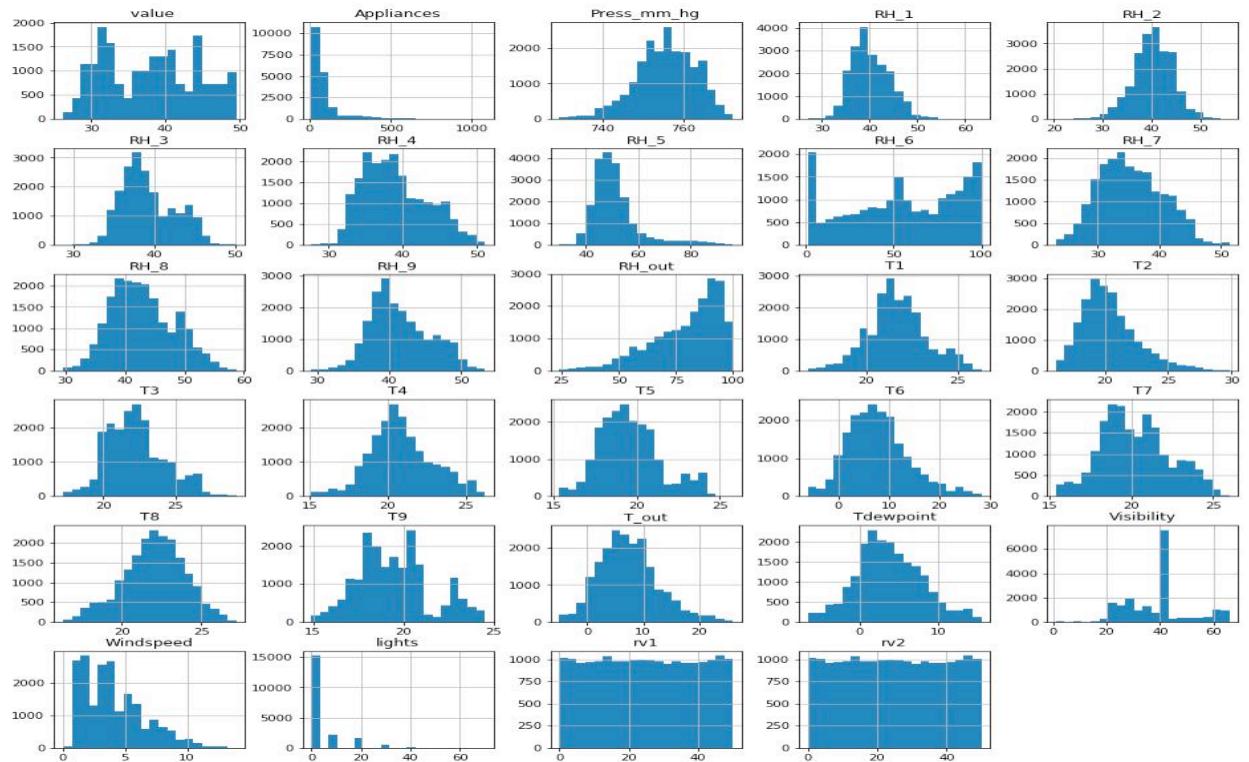


Explore the data using the Weather Dimension -



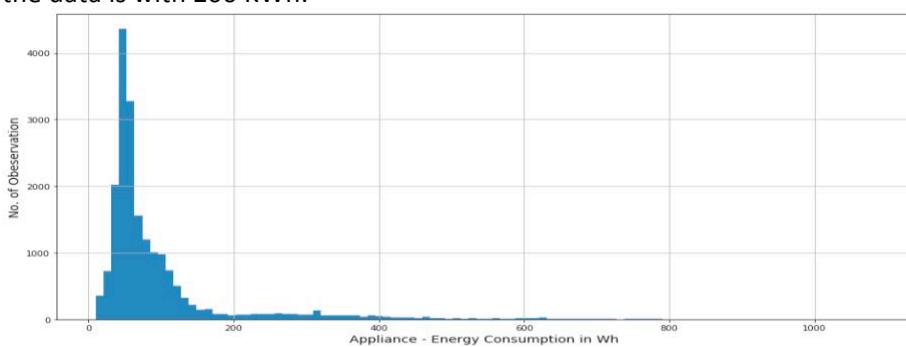
- There doesn't seem to be having any linearity between any of the attributes.

Lets explore the distribution using the histogram –



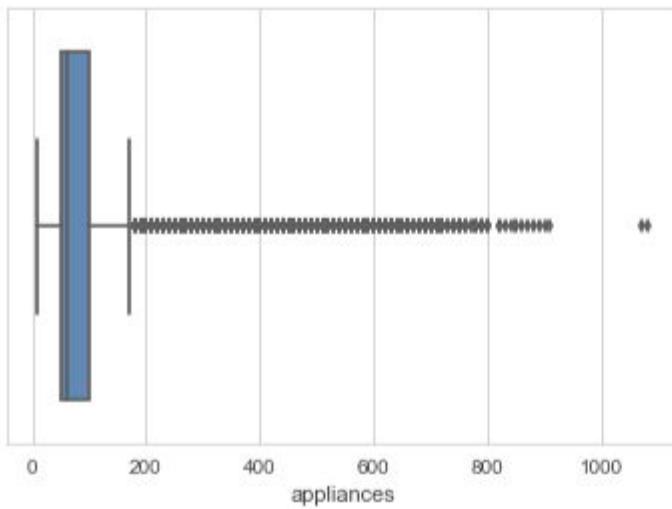
- All humidity values are almost having normal distribution except RH\_6 and RH\_out. In other words the reading from inside the home is having normal distribution.
- All temperature readings follow a Normal distribution except for T9.
- Visibility, Windspeed and Appliances are having skewed data.
- Rv1 and Rv2 are random variables and doesn't seems to be contributing

On the Target Attribute – Appliance, the below histograms is rightly skewed and most of the data is with 200 KWh.



Target variable, Appliances is highly right skewed.

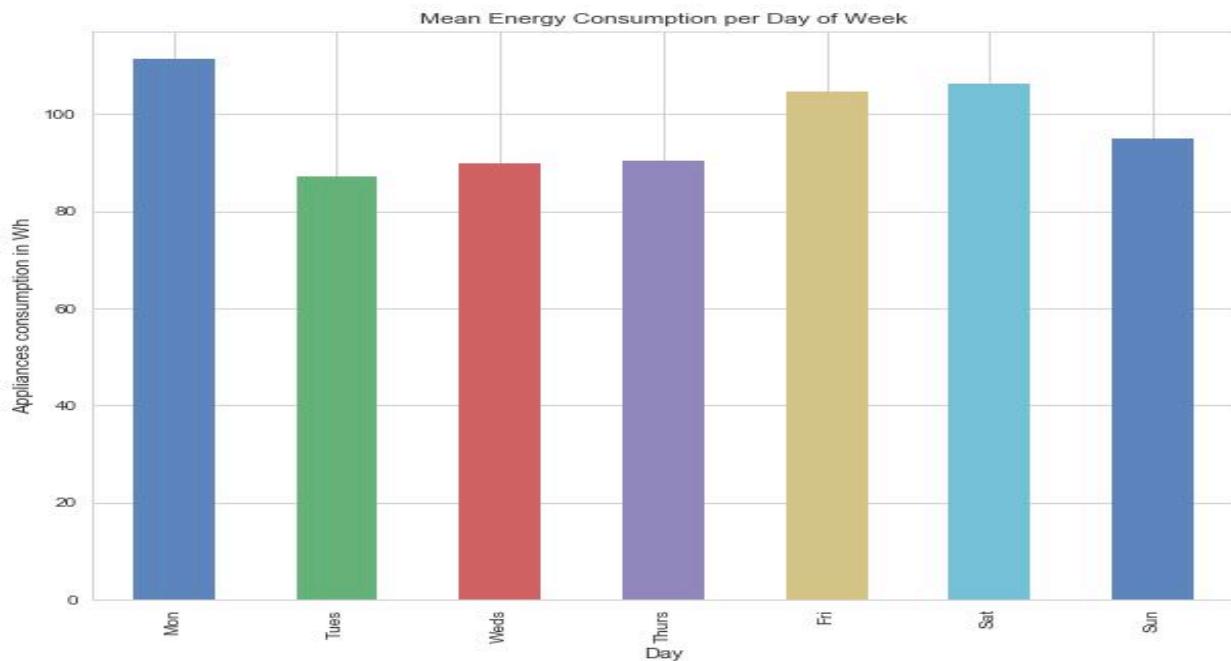
Alternatively exploring using Boxplot – on Appliance Attribute



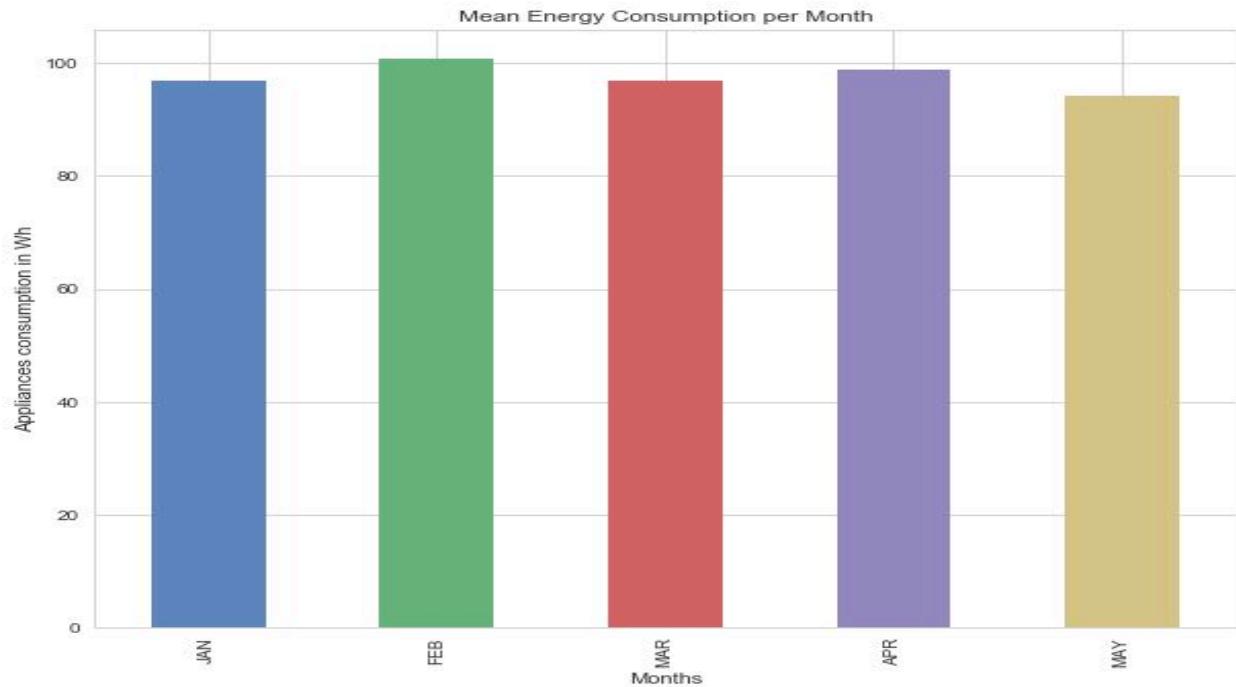
- Percentage of dataset in range of 0-200 KWh is 90.291%

Using the date attributes, created new columns for Month and Weeks using the **datetimestamp** method

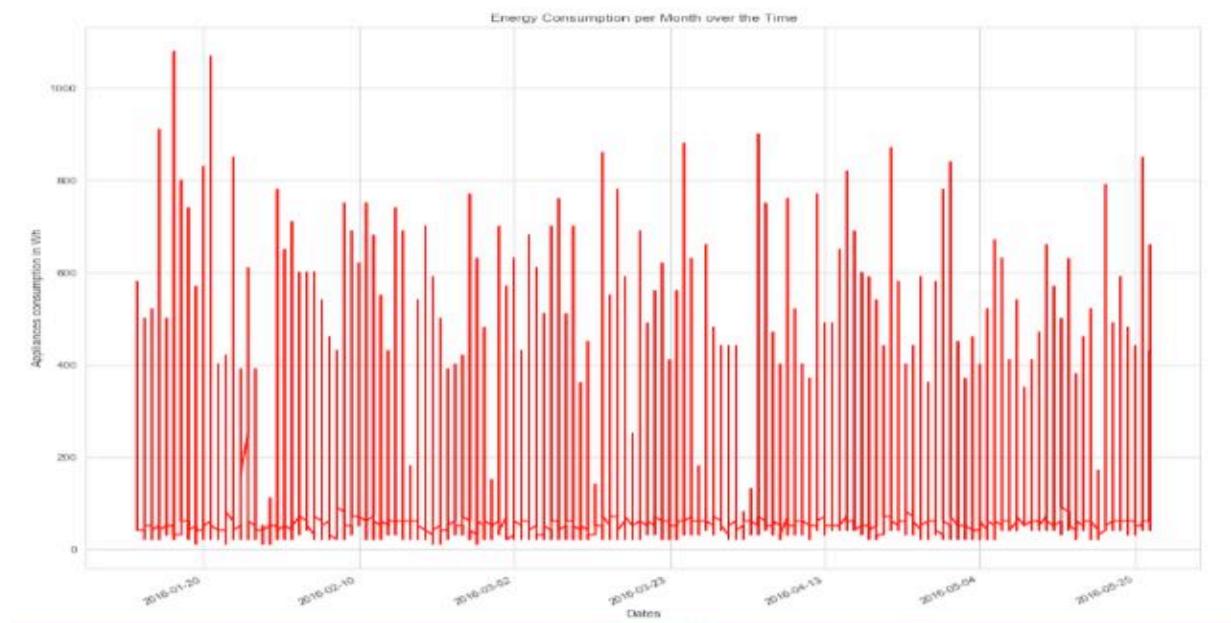
With taking the average on week – Monday the usage has been higher, followed by Saturday and Friday.



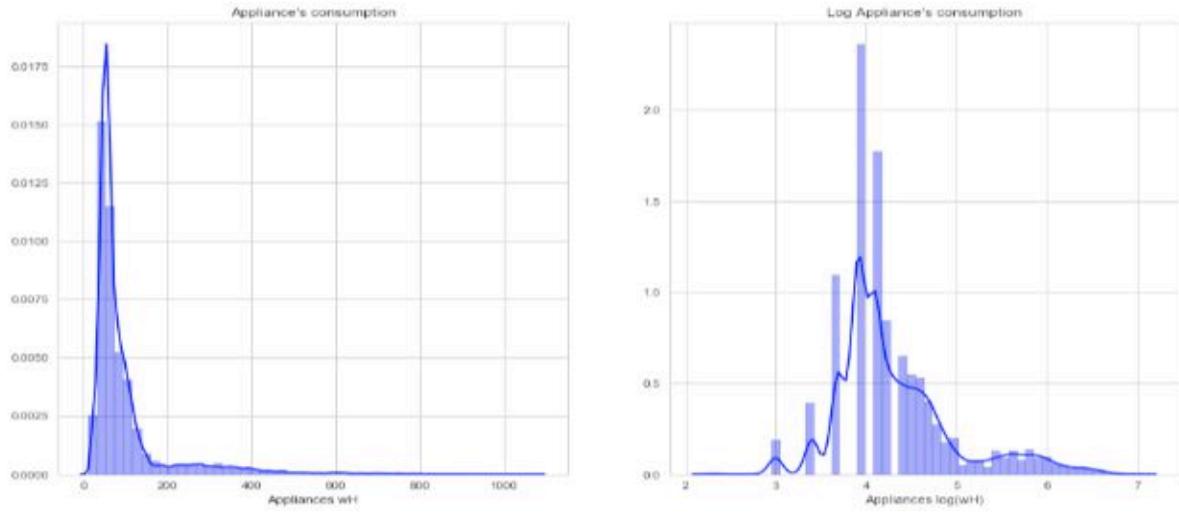
For Monthly Average – On a average, February and April the consumption has been more than other months.



For Day wise consumption – plotting this date wise, energy consumption, In January month there were 2 days when the consumption was more than 1000 KWh.

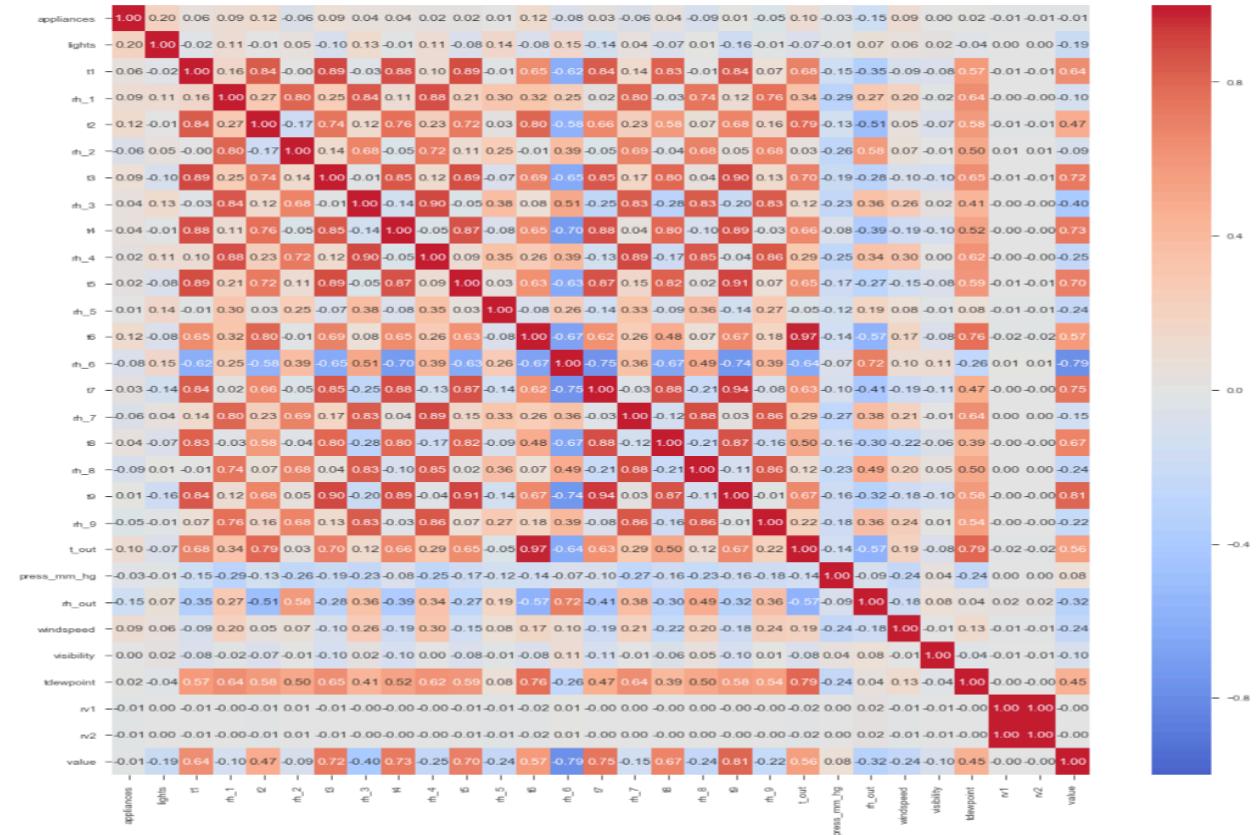


Since the Appliance data captured were rightly skewed, converting the column to Log values to see if it has the normal distribution.

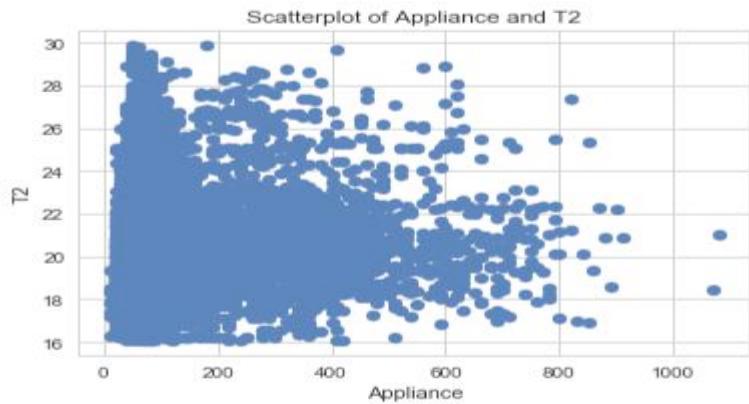


Let's explore the Correlation plot –

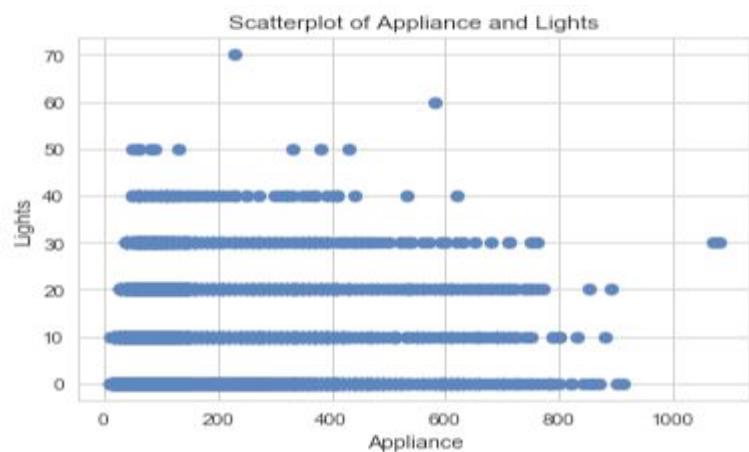
With Appliance attribute –



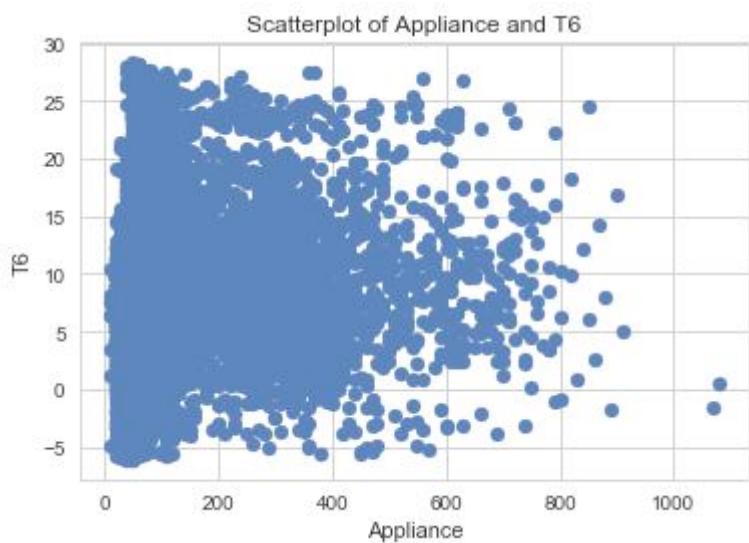
Scatterplot between appliances and t2



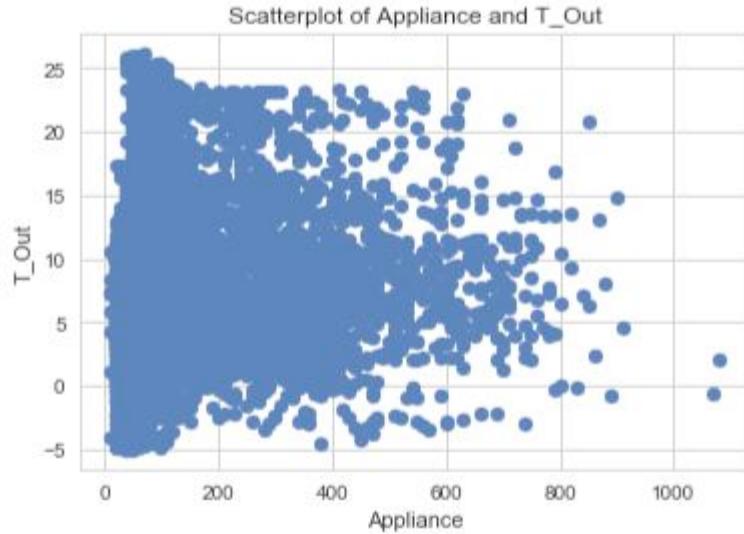
Scatter plot between Appliance and Lights



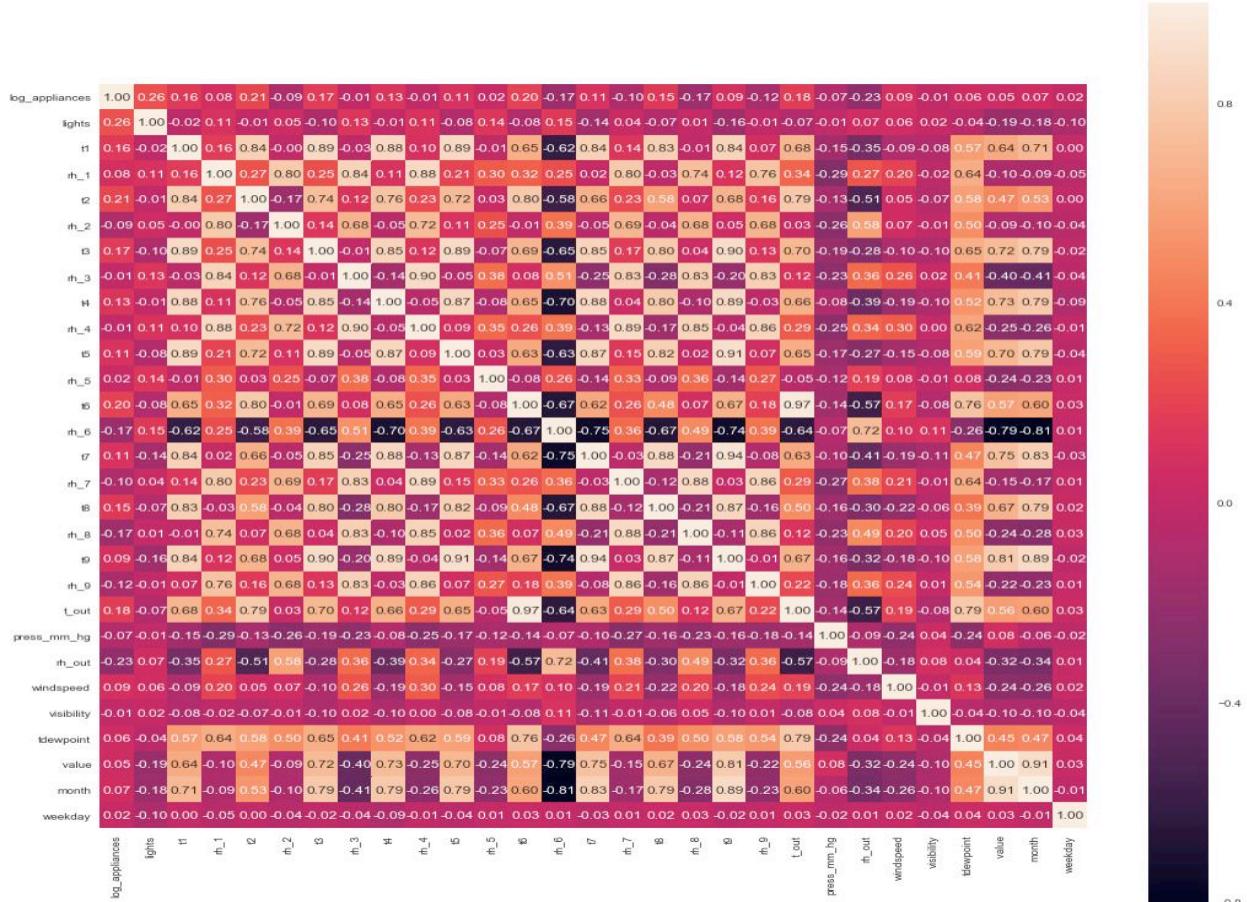
Scatterplot between Appliance and T6



## Scatterplot between Appliance and T\_out



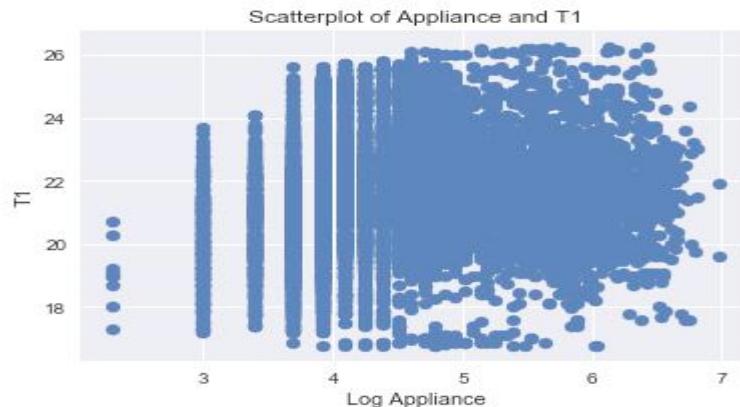
## Correlation plot of using Log value of Appliance -



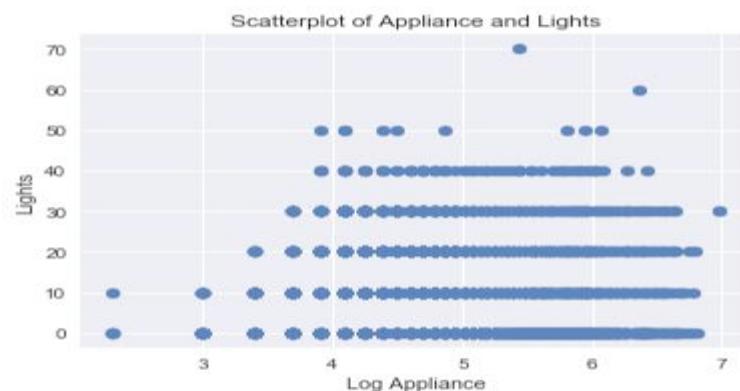
### With Log Appliance

- The most correlated features with energy consumption(log\_appliances) are: lights=0.26, t6=0.20, t2=0.22, t3 = 0.17, t\_out = 0.18, rh\_out = -0.23, rh\_8 = -0.17, rh\_6 = -0.17, windspeed = 0.09.
- In a linear regression problem only linear independent variables can be used as features to explain energy consumption otherwise we will have multicollinearity issues.

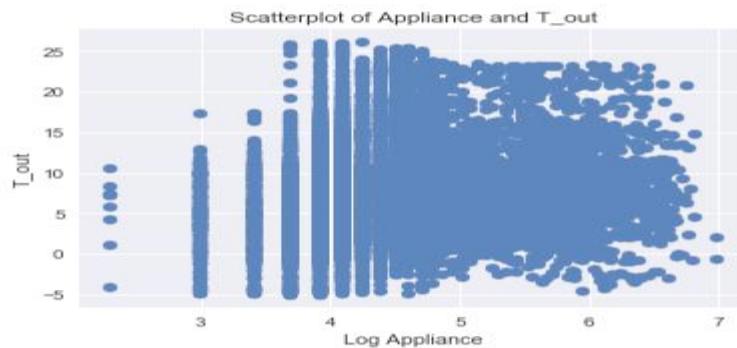
Scatter plot of log\_appliance and t1



Scatterplot between log\_appliance and lights

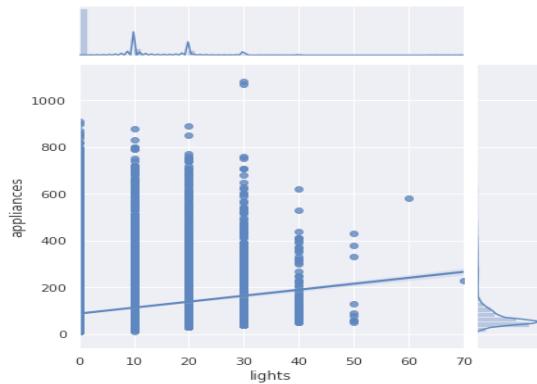


Scatter plot between log appliance and t\_out

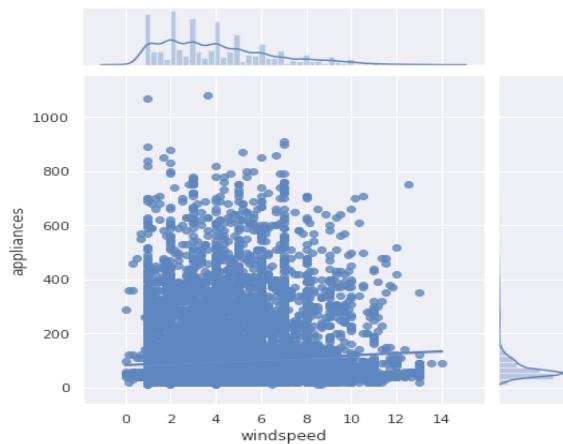


**Variables that are particularly significant in terms of predicting Appliance Energy Consumption based on the correlation matrix –**

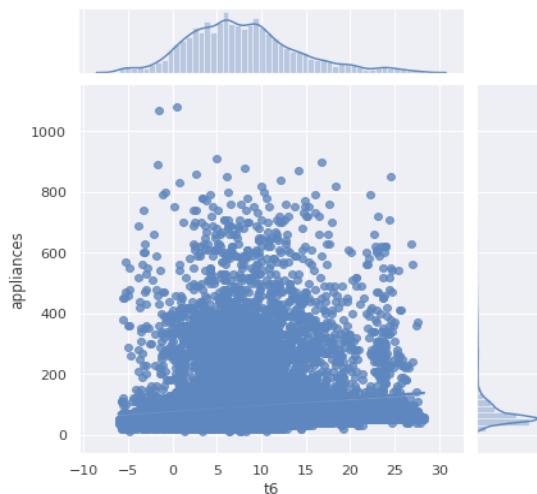
- Between Appliance and Lights



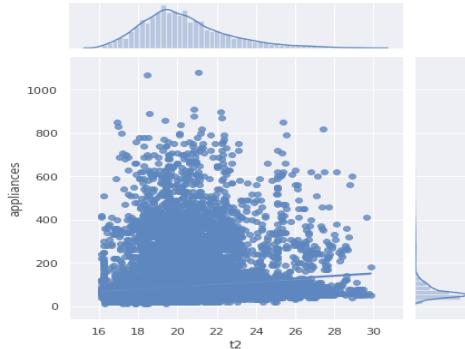
- Between Appliance and Windspeed



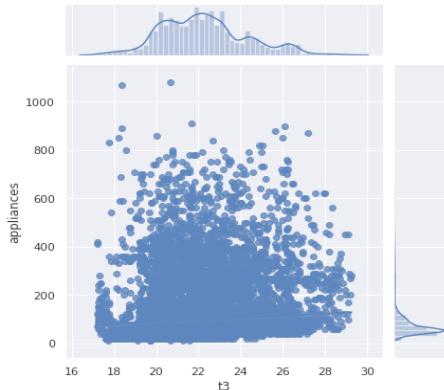
- Between Appliance and T6



- Between Appliance and T2



- Between Appliance and T3



#### Calculate the Correlation between Temperature features -

- |   |      |      |      |
|---|------|------|------|
| • Correlation between T9 and T1 pearson | 0.84 | 0.00 | None |
| • Correlation between T9 and T2 pearson | 0.68 | 0.00 | None |
| • Correlation between T9 and T3 pearson | 0.90 | 0.00 | None |
| • Correlation between T9 and T4 pearson | 0.89 | 0.00 | None |
| • Correlation between T9 and T5 pearson | 0.91 | 0.00 | None |
| • Correlation between T9 and T6 pearson | 0.67 | 0.00 | None |
| • Correlation between T9 and T7 pearson | 0.94 | 0.00 | None |
| • Correlation between T9 and T8 pearson | 0.87 | 0.00 | None |

#### Check, if the Temperature, Humidity and Weather features influences Appliance –

1. Coefficient table (middle table). We can interpret the  $t_3$  coefficient (4.3471) by first noticing that the p-value (under  $P>|t|$ ) is so small, basically zero. This means that the  $t_3$  is a statistically significant predictor of appliance energy consumption.

The regression coefficient for  $t_3$  of 4.3471 means that on average, each additional  $t_3$  temperature is associated with an increase the appliance energy consumption

The confidence interval gives us a range of plausible values for this average change, about (3.637 and 5.058)

$R^2$  is only 0.007, hence t3 doesn't contribute much on the variance. F-Statistic The F-Statistic is 143.8 and the probability for this statistic is 5.09e-33, which is close to 0. We can safely reject the null hypothesis, indicating that at least one coefficient is nonzero.

```

OLS Regression Results
-----
Dep. Variable: appliances R-squared: 0.007
Model: OLS Adj. R-squared: 0.007
Method: Least Squares F-statistic: 143.8
Date: Mon, 18 May 2020 Prob (F-statistic): 5.09e-33
Time: 21:12:10 Log-Likelihood: -1.1931e+05
No. Observations: 19735 AIC: 2.386e+05
Df Residuals: 19733 BIC: 2.386e+05
Df Model: 1
Covariance Type: nonrobust
-----
            coef    std err          t      P>|t|      [0.025      0.975]
-----
Intercept  0.8955     8.105      0.110      0.912     -14.990     16.781
t3         4.3471     0.362     11.992      0.000       3.637      5.058
-----
Omnibus:        14099.091 Durbin-Watson: 0.498
Prob(Omnibus): 0.000 Jarque-Bera (JB): 196052.616
Skew:           3.410 Prob(JB): 0.00
Kurtosis:       16.854 Cond. No. 250.
-----
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```

2. Coefficient table (middle table). We can interpret the  $t_3+t_6$  coefficient (0.4119, 1.8871) by first noticing that the p-value (under  $P>|t|$ ) is so small, basically zero. This means that the  $t_6$  is a statistically significant predictor of appliance energy consumption.

The regression coefficient for t6 of 1.8871, means that on average, each additional t6 temperature is associated with an increase the appliance energy consumption

The confidence interval gives us a range of plausible values for this average change, about (1.566 and 2.208)

$R^2$  is only 0.014, hence t3 and t6 doesn't contribute much on the variance. F-Statistic The F-Statistic is 138.8 and the probability for this statistic is 1.39e-60, which is close to 0. We can safely reject the null hypothesis, indicating that at least one  $\beta$  coefficient is nonzero.

```

OLS Regression Results
-----
Dep. Variable: appliances R-squared: 0.014
Model: OLS Adj. R-squared: 0.014
Method: Least Squares F-statistic: 138.8
Date: Mon, 18 May 2020 Prob (F-statistic): 1.39e-60
Time: 21:12:21 Log-Likelihood: -1.1924e+05
No. Observations: 19735 AIC: 2.385e+05
Df Residuals: 19732 BIC: 2.385e+05
Df Model: 2
Covariance Type: nonrobust

coef std err t P>|t| [0.025 0.975]
-----
Intercept 73.5949 10.249 7.181 0.000 53.506 93.684
t3 0.4119 0.497 0.828 0.407 -0.563 1.386
t6 1.8871 0.164 11.525 0.000 1.566 2.208

Omnibus: 14117.484 Durbin-Watson: 0.500
Prob(Omnibus): 0.000 Jarque-Bera (JB): 197909.447
Skew: 3.412 Prob(JB): 0.00
Kurtosis: 16.932 Cond. No. 339.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```

3. Coefficient table (middle table). We can interpret the t3+t6+rh\_out coefficient (1.8057, 0.3079, -0.9076) by first noticing that the p-value (under P>|t|) is so small, basically zero. This means that the t6 is a statistical significant predictor of appliance energy consumption.

The regression coefficient for rh\_out of -0.9076, means that on average, each additional temperature is associated with an decrease the appliance energy consumption

The confidence interval gives us a range of plausible values for this average change, about (-1.025 and -0.790)

$R^2$  is only 0.025 better than previous, hence t3, t6 and rh\_out doesn't contribute much on the variance. F-Statistic The F-Statistic is 170.3 and the probability for this statistic is 4.96e-109, which is close to 0. We can safely reject the null hypothesis, indicating that at least one  $\beta$  coefficient is nonzero.  $\square$

OLS Regression Results						
Dep. Variable:	appliances	R-squared:	0.025			
Model:	OLS	Adj. R-squared:	0.025			
Method:	Least Squares	F-statistic:	170.3			
Date:	Mon, 18 May 2020	Prob (F-statistic):	4.96e-109			
Time:	21:12:55	Log-Likelihood:	-1.1913e+05			
No. Observations:	19735	AIC:	2.383e+05			
Df Residuals:	19731	BIC:	2.383e+05			
Df Model:	3					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
Intercept	127.4360	10.790	11.810	0.000	106.286	148.586
t3	1.8057	0.503	3.592	0.000	0.820	2.791
t6	0.3079	0.193	1.593	0.111	-0.071	0.687
rh_out	-0.9076	0.060	-15.173	0.000	-1.025	-0.790
Omnibus:	14135.525	Durbin-Watson:	0.507			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	199836.330			
Skew:	3.415	Prob(JB):	0.00			
Kurtosis:	17.014	Cond. No.	1.26e+03			
Warnings:						
(1) Standard Errors assume that the covariance matrix of the errors is correctly specified.						
(2) The condition number is large, 1.26e+03. This might indicate that there are strong multicollinearity or other numerical problems.						

4. Coefficient table (middle table). We can interpret the t1+t2+t3+t4+t5+t6+t7+t8+rh\_1+rh\_2+windspeed, coefficient (9.0446, -25.6614, 17.7293, -1.4768, -7.3830, -7.5650, 1.0356, -6.2685, 9.4475, 20.0347, -20.3286, 1.6784) by first noticing that the p-value (under P>|t|) is so small, basically zero. This means that the t6 is a statistical significant predictor of appliance energy consumption.

The confidence interval of t3 gives us a range of plausible values for this average change, about (15.814 and 19.644)

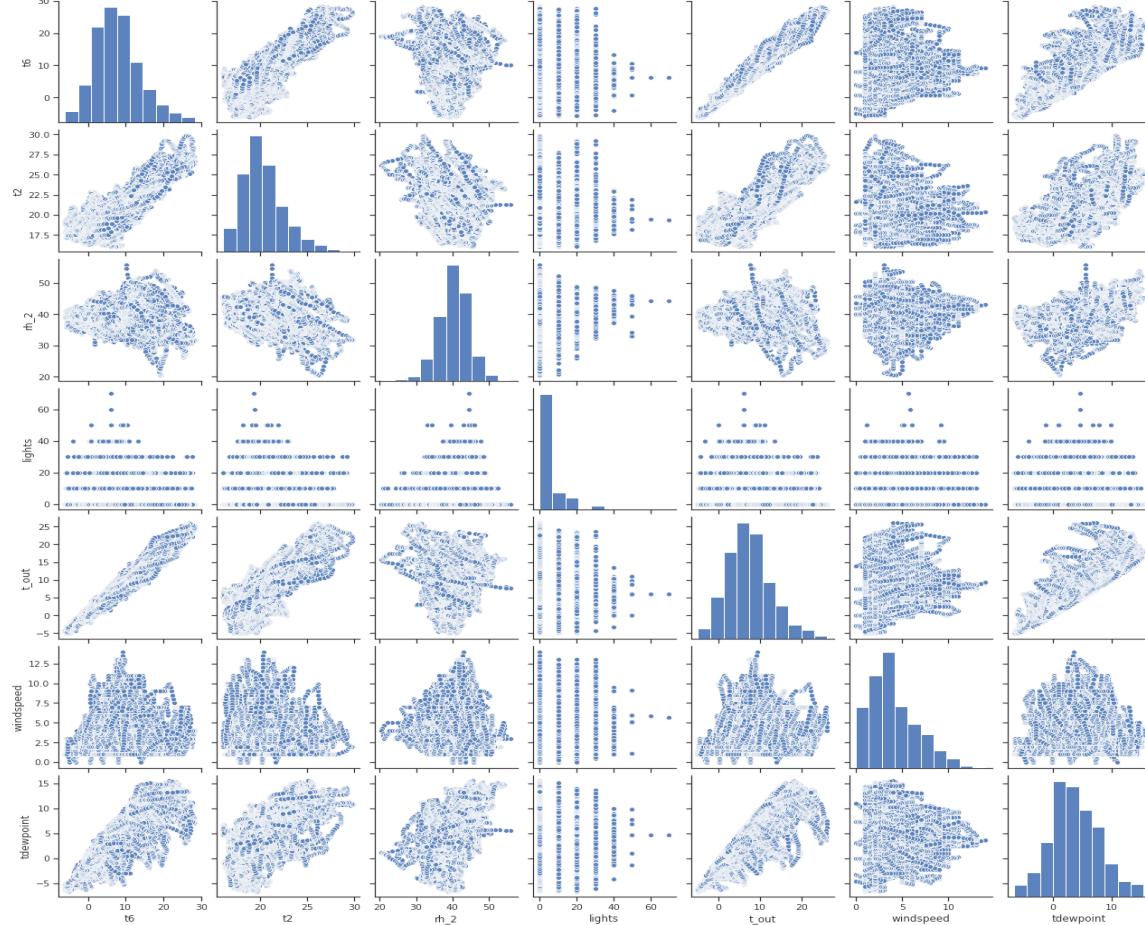
$R^2$  is only 0.098 better than previous, F-Statistic The F-Statistic is 194.5 and the probability for this statistic is 0. We can safely reject the null hypothesis, indicating that at least one  $\beta$  coefficient is nonzero.

```

OLS Regression Results
-----
Dep. Variable: appliances R-squared: 0.098
Model: OLS Adj. R-squared: 0.097
Method: Least Squares F-statistic: 194.5
Date: Mon, 18 May 2020 Prob (F-statistic): 0.00
Time: 21:35:34 Log-Likelihood: -1.1836e+05
No. Observations: 19735 AIC: 2.367e+05
Df Residuals: 19723 BIC: 2.368e+05
Df Model: 11
Df Model: 11
Covariance Type: nonrobust
-----
            coef  std err      t    P>|t|   [0.025]   [0.975]
Intercept  123.2455   15.477   7.963   0.000   92.909   153.582
t1         9.0446    1.764   5.127   0.000   5.587   12.502
t2        -25.6614   1.450  -17.697   0.000  -28.504  -22.819
t3         17.7293   0.977  18.148   0.000  15.814   19.644
t4        -1.4768   0.907  -1.628   0.103  -3.255   0.301
t5        -7.3830   1.065  -6.930   0.000  -9.471  -5.295
t6         1.0356   0.227   4.552   0.000   0.590   1.482
t7        -6.2685   0.971  -6.457   0.000  -8.171  -4.366
t8         9.4475   0.881  10.730   0.000   7.722   11.173
rh_1       20.0347   0.630  31.792   0.000  18.799   21.270
rh_2      -20.3286   0.630 -32.261   0.000 -21.564  -19.094
windspeed   1.6784   0.320   5.239   0.000   1.050   2.306
-----
Omnibus: 13836.970 Durbin-Watson: 0.578
Prob(Omnibus): 0.000 Jarque-Bera (JB): 196316.384
Skew: 3.306 Prob(JB): 0.00
Kurtosis: 16.965 Cond. No. 1.80e+03
-----
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.8e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

```

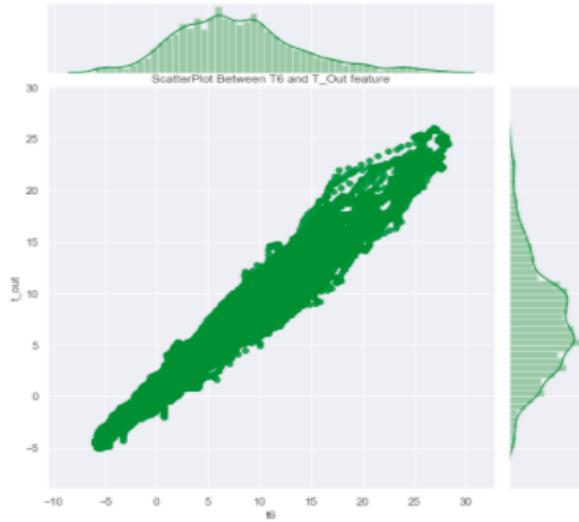
Pairplot for 't6','t2','rh\_2','lights','t\_out','windspeed','tdewpoint' features for their distribution –



**Is there a significant difference between T6 and T\_out and impact my future Prediction Models**

	t6	t_out
count	19735.000000	19735.000000
mean	7.910939	7.411665
std	6.090347	5.317409
min	-6.065000	-5.000000
25%	3.626667	3.666667
50%	7.300000	6.916667
75%	11.256000	10.408333
max	28.290000	26.100000

With Description and plotting the **jointplot** of the two features -



It seems like there is linear relation between 'T6' and T\_out'

**To find the Correlation between all the Features and Target – Appliances -**

	Correlation coefficients	p-value
appliances	1.000000	0.000000e+00
lights	0.197278	2.305108e-172
t2	0.120073	2.784947e-64
t6	0.117638	9.333867e-62
t_out	0.099155	2.624854e-44
windspeed	0.087122	1.456471e-34
rh_1	0.086031	9.639431e-34
weekday_avg	0.085900	1.208067e-33
t3	0.085060	5.086416e-33

	Correlation coefficients	p-value
<b>t1</b>	0.055447	6.449169e-15
<b>house_temp</b>	0.054740	1.411780e-14
<b>t4</b>	0.040281	1.507881e-08
<b>t8</b>	0.039572	2.683103e-08
<b>rh_3</b>	0.036292	3.402540e-07
<b>t7</b>	0.025801	2.890302e-04
<b>t5</b>	0.019760	5.503451e-03
<b>rh_4</b>	0.016965	1.715603e-02
<b>tdewpoint</b>	0.015353	3.102113e-02
<b>t9</b>	0.010010	1.596635e-01
<b>rh_5</b>	0.006955	3.286027e-01
<b>weekday</b>	0.003060	6.672580e-01
<b>visibility</b>	0.000230	9.741858e-01
<b>week</b>	-0.011356	1.106606e-01
<b>month</b>	-0.011606	1.030264e-01
<b>value</b>	-0.013535	5.725207e-02
<b>house_hum</b>	-0.020075	4.799007e-03
<b>press_mm_hg</b>	-0.034885	9.493922e-07
<b>rh_9</b>	-0.051462	4.697109e-13
<b>rh_7</b>	-0.055642	5.187296e-15
<b>rh_2</b>	-0.060465	1.873022e-17
<b>rh_6</b>	-0.083178	1.209481e-31
<b>rh_8</b>	-0.094039	5.211566e-40
<b>rh_out</b>	-0.152282	1.077516e-102

#### Create a New DataFrame to conducting T-Statistics and Changing it to Categorical column –

- Create the dataframe with “lights” and “appliances”.
- Update the feature where the number of lights are 40 or more as 1
- Update the feature where the number of lights are 39 or less as 0

	<b>lights</b>	<b>appliances</b>
<b>dateupdate</b>		
2016-01-11	0	60
2016-01-11	0	60
2016-01-11	0	50
2016-01-11	1	50
2016-01-11	1	60

- When conducted the T-test , T-Statistics is 133.85 and pvalue is 0, hence we can reject the null hypothesis and conclude that there is a statically significant difference.

### **Major Inference –**

- Temperature feature from T1-T9 and T\_out have positive correlation with the target Appliances. For the indoor temperatures, the correlations are high as expected. Four columns have a high degree of correlation with T9 & T3,T5,T7,T8 also T6 & T\_Out has high correlation (both temperatures from outside) . Hence we can remove the T9 and T\_out from the model in next section.
- Weather attributes - Visibility, Tdewpoint, Press\_mm\_hg have low correlation values
- Humidity - There are no significantly high correlation cases for humidity sensors.
- Random variables have no role to play; hence we will remove these features from the model in next section.

### **Model Building and Implementation –**

- As identified earlier – we are dropping the below field –
  - 'date\_x', 'appliances','rv1', 'rv2','t6','t9' from the dataset.
  - Created the X will all the features and Y with the target feature.
  - Using train\_test\_split method, we have done the split of the dataset in 70% Training data and 30% test data.
  - Upon running the Linear regression model- we get the below score for R^2 and RMSE(Root Mean Square Error)

```
Classifier fitted in 1.319 seconds
R^2: 16.678
Root Mean Squared Error: 92.652
```

- Also, with Cross validation, we didn't see the improvement in the performance of the benchmark algorithms –

```
[0.11945252 0.18596483 0.17729608 0.15143344 0.18476264]
Average 5-Fold CV Score: 0.16378190055186861

[0.11837952 0.12052544 0.18454178 0.18941428 0.16061813 0.19884144
 0.13440485 0.17477997 0.17973161 0.19508757]
Average 10-Fold CV Score: 0.16563245972749235
```

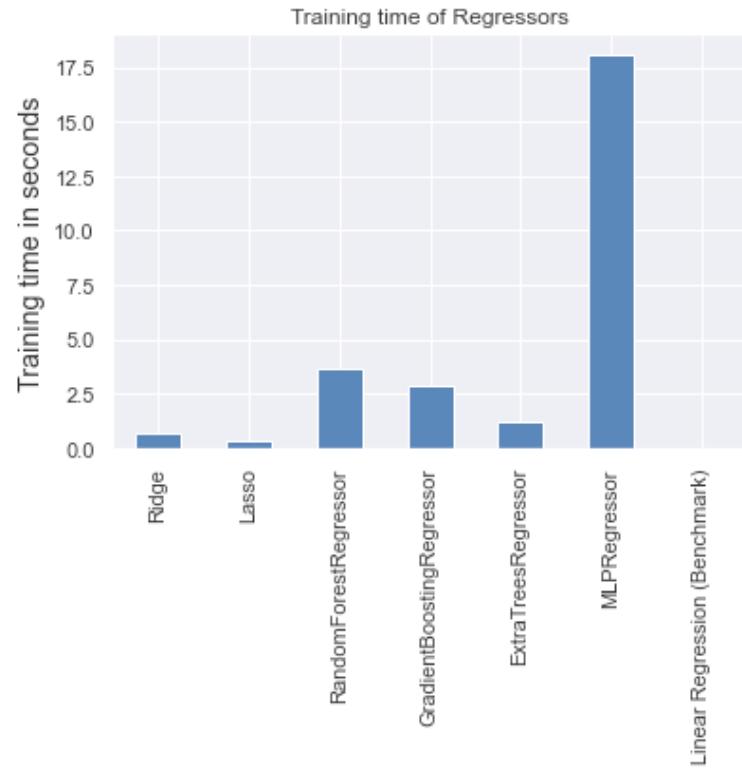
- Now, we will try to scale the data and find best performing model –

1. Dropped the x\_date feature from the dataset, and using the StandardScaler method, scaled the dfactual dataframe.
  2. From the scaled dataset, dropped the 'appliances','rv1', 'rv2','t6','t9'.
  3. Created the Training and Test Dataset with 70-30% ratio.
6. Create the following models with key important features –
- Regularized linear models as an improvement over Linear Regression.
    - Ridge Regression
    - Lasso Regression
  - Ensemble based Tree Regression models, which deal with number of features and outlier data.
    - Random Forests
    - Gradient Boosting
    - Extra Trees
  - Neural networks for non-linear relationships target feature and predictors.
    - Multi-Layer Perceptron

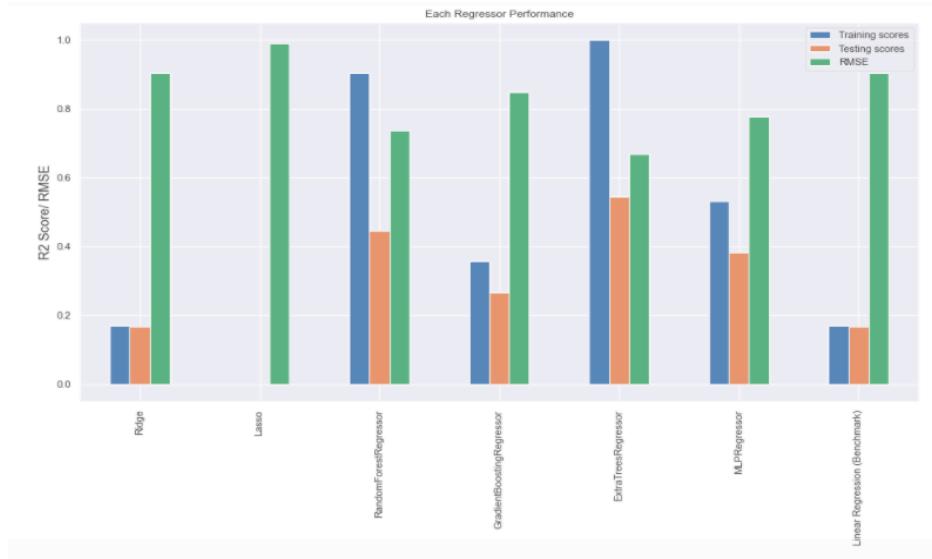
1. Implemented them in a iterative manner with creating different functions
  1. Created function to capture the fit and predict the models and capture the score and accuracy for the models
  2. Created the pipeline and passed all the algorithms to be executed in the above function.
  3. Created a function to display and store the results / outcome.
  4. Below is the results displaying the R^2 and RMSE and time it took to predict.

	Training times	Training scores	Testing scores	RMSE
<b>ExtraTreesRegressor</b>	1.20606	1	0.543191	0.669149
<b>RandomForestRegressor</b>	3.62582	0.904629	0.44562	0.737156
<b>MLPRegressor</b>	18.0836	0.531666	0.383224	0.777534
<b>GradientBoostingRegressor</b>	2.84227	0.357163	0.265424	0.848543
<b>Ridge</b>	0.717286	0.170043	0.166668	0.903784
<b>Linear Regression (Benchmark)</b>	0.0219181	0.169898	0.166489	0.903881
<b>Lasso</b>	0.381927	0	-1.15367e-06	0.990047

2. Comparing the Training time –



3. Plot to compare the performance of the algorithms on datasets



Interpretation –

Least performing Repressor - Lasso Repressor and best performing Repressor - Extra Trees Repressor. Even though Extra Trees Repressor has a R2 score of 1.0 on training set, which might suggest over-fitting but, it has the highest score on test set and also, it's

RMSE value is also the lowest. Clearly, ExtraTreesRegressor is the best model out of given models.

4. Hyper-parameter tuning the best Model – “ExtraTreesRegressor” observed from above step – Using the RandomizedSearchCV, we will find the best estimators and using those estimators we will perform the prediction.

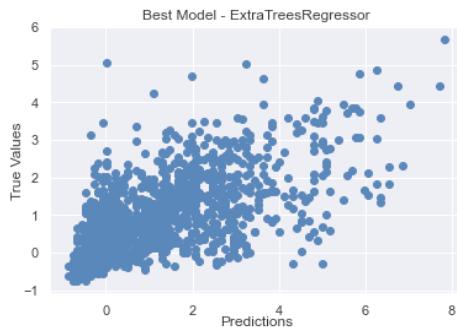
```
RandomizedSearchCV(cv=5, error_score='raise-deprecating',
                    estimator=ExtraTreesRegressor(bootstrap=False, criterion='mse', max_depth=None,
                    max_features='auto', max_leaf_nodes=None,
                    min_impurity_decrease=0.0, min_impurity_split=None,
                    min_samples_leaf=1, min_samples_split=2,
                    min_weight_fraction_leaf=0.0, n_estimators='warn', n_jobs=None,
                    oob_score=False, random_state=79, verbose=0, warm_start=False),
                    fit_params=None, iid='warn', n_iter=20, n_jobs=-1,
                    param_distributions={'n_estimators': [10, 50, 100, 200, 250], 'max_features': ['auto', 'sqrt', 'log2'], 'ma
x_depth': [None, 10, 50, 100, 200, 500]},
                    pre_dispatch='2*n_jobs', random_state=79, refit=True,
                    return_train_score='warn', scoring='r2', verbose=2)
```

```
Parameters of best Regressor : {'n_estimators': 250, 'max_features': 'log2', 'max_depth': None}
```

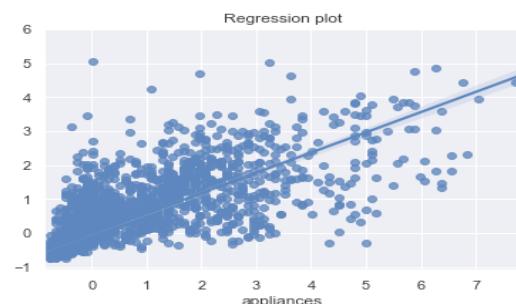
5. Using the best parameter we will fit and predict the training data and predict on test data.

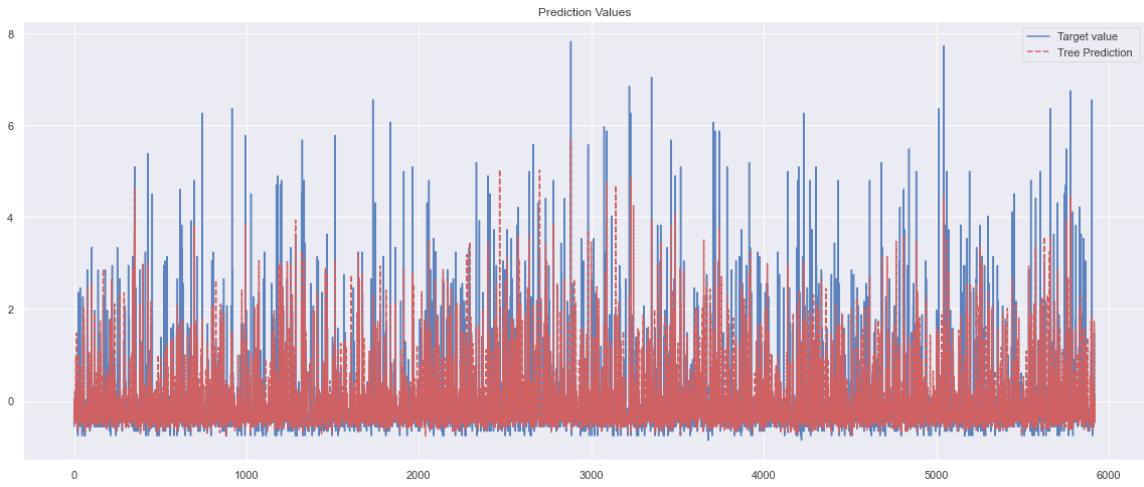
```
R2 score on Training set = 1.000
RMSE on Training set = 0.000
R2 score on Testing set = 0.627
RMSE on Testing set = 0.605
```

6. Plotting the data for y\_test\_s and predicted data –



Using the seaborn plot using regplot function





Overlaying the test data and predicted data, we can see that the prediction is not so accurate.

#### Interpretation from Implementation -

- R2 score improvement compared to Benchmark model = 0.463.
- RMSE improvement compared to Benchmark model = 0.301.
- R2 score improvement compared to without tuned model = 0.086.
- RMSE improvement compared to without tuned model = 0.066.

7. Important features contributing from the data set are as below –

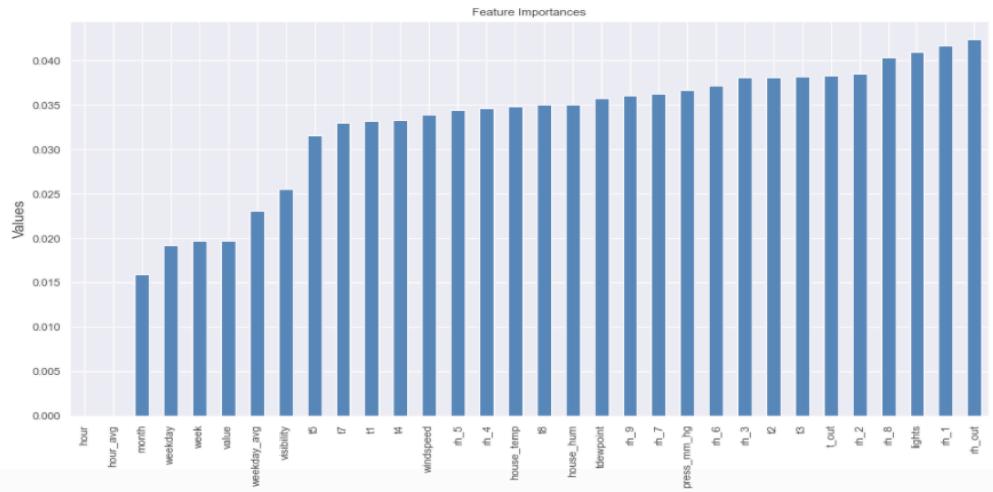
```

Most important feature = rh_out
Least important feature = hour

Top 5 most important features:-
rh_out
rh_1
lights
rh_8
rh_2

Top 5 least important features:-
hour
hour_avg
month
weekday
week

```



## 5. Feature and Model Evaluation-

- Clone the above best model clone with the 'rh\_out', 'rh\_1', 'lights', 'rh\_8', 'rh\_2', 't\_out', 't3', 't2' and do a prediction only with the most important feature, to verify if there is any improvement in the model accuracy.

**R2 Score on testing dataset = 0.519**

**RMSE Score on testing dataset = 0.686**

- Comparing these results with above best performing algorithms – Extratreeregressor

- R2 Score on testing dataset = 0.52
- RMSE Score on testing dataset = 0.69
- Difference in R2 score = 0.109 or 11% loss of explained variance.
- Increase in RMSE = 0.083

The model has not performed better with reduced number of features.

## 6. Conclusion -

- Best Algorithm is ExtraTreesRegressor.
- Top 5 most important features:- rh\_out, rh\_1, lights, rh\_8, rh\_2
- Variance explained on test set = 63%.
- RMSE error = 60.3%
- R2 score improvement compared to Benchmark model = 0.463.
- RMSE improvement compared to Benchmark model = 0.301.
- R2 score improvement compared to without tuned model = 0.086.
- RMSE improvement compared to without tuned model = 0.066.

