1. Abstract –

**Purpose –** In world, we have a continuous problem for saving Energy in any form like water electricity, etc. This project we are implementing the model for predicting Energy Consumption, if we given condition we are able to predict the energy consumption, we can save or distribute/ repurpose the energy to its rightful usage.

**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

2. Data Wrangling -

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 as mentioned in Fig1.

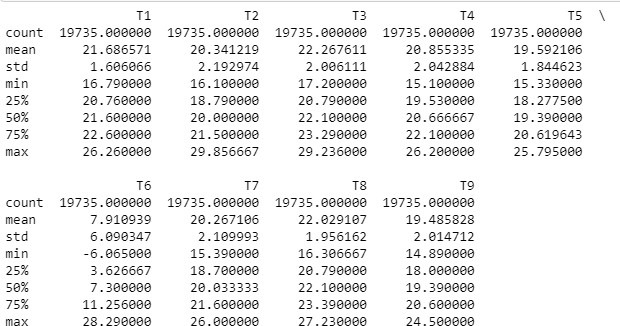


Fig 1

1. 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 as mentioned in Fig2.

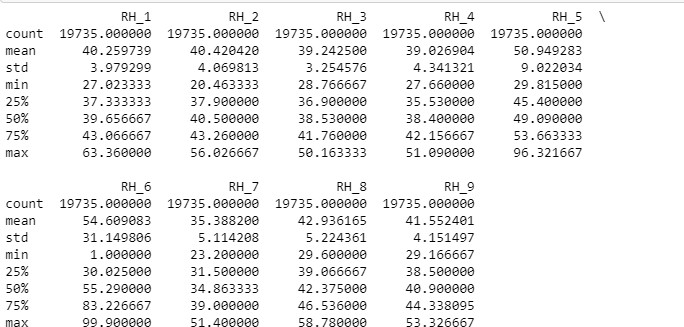


Fig 2

1. 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.
2. 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 as mentioned in Fig3.

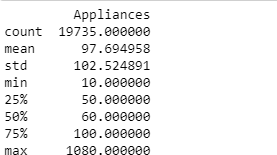
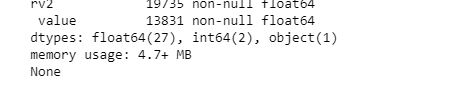


Fig 3

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.



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.

3. Data Visualization -

Divide the data in dimension wise to explore from the input dataset , as defined in below fig 4–

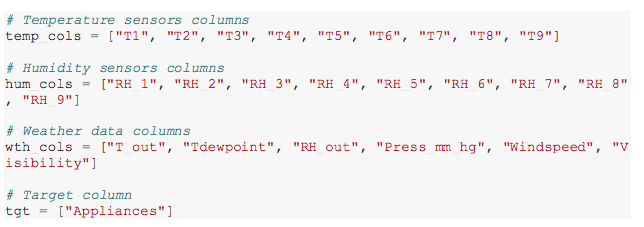


Fig 4

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",** as below fig5

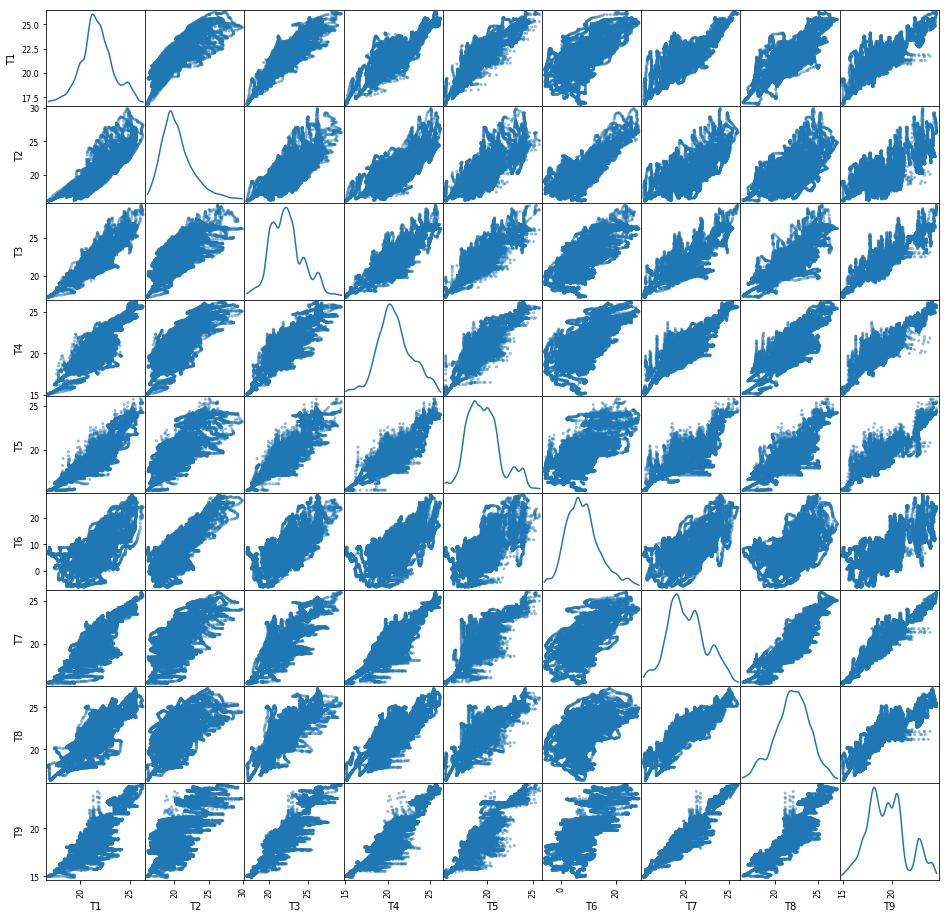
****

Fig 5

* 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.
* From Fig6, we can see that there is a relation between these two attributes but also have some outliers

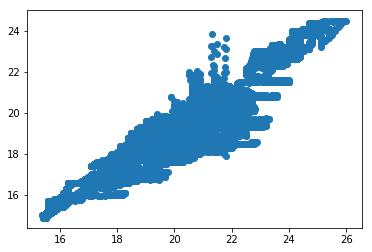


Fig 6

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 as shown in fig 7.

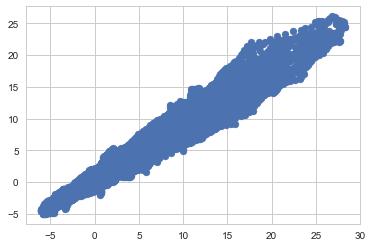


Fig 7

Explore the data using the Weather Dimension , as below in fig 8-

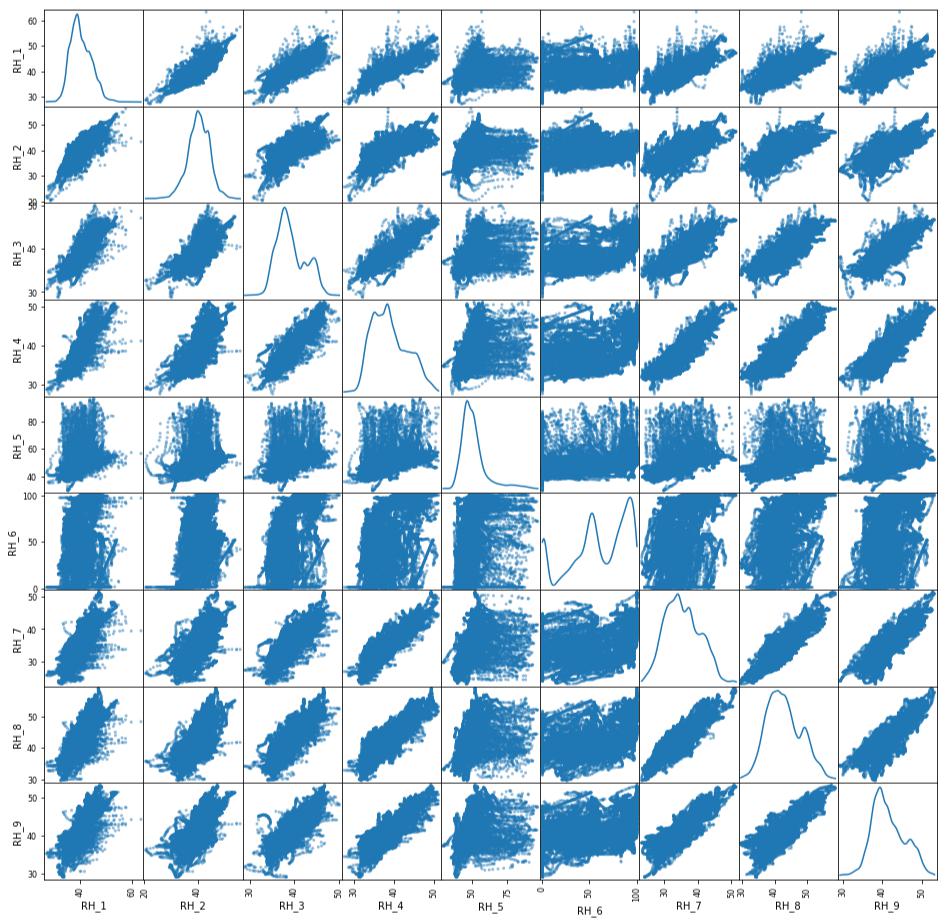


Fig 8

* There doesn’t seems to be having any linearity between any of the attributes.

Lets explore the distribution using the histogram, as mentioned in fig 9 –

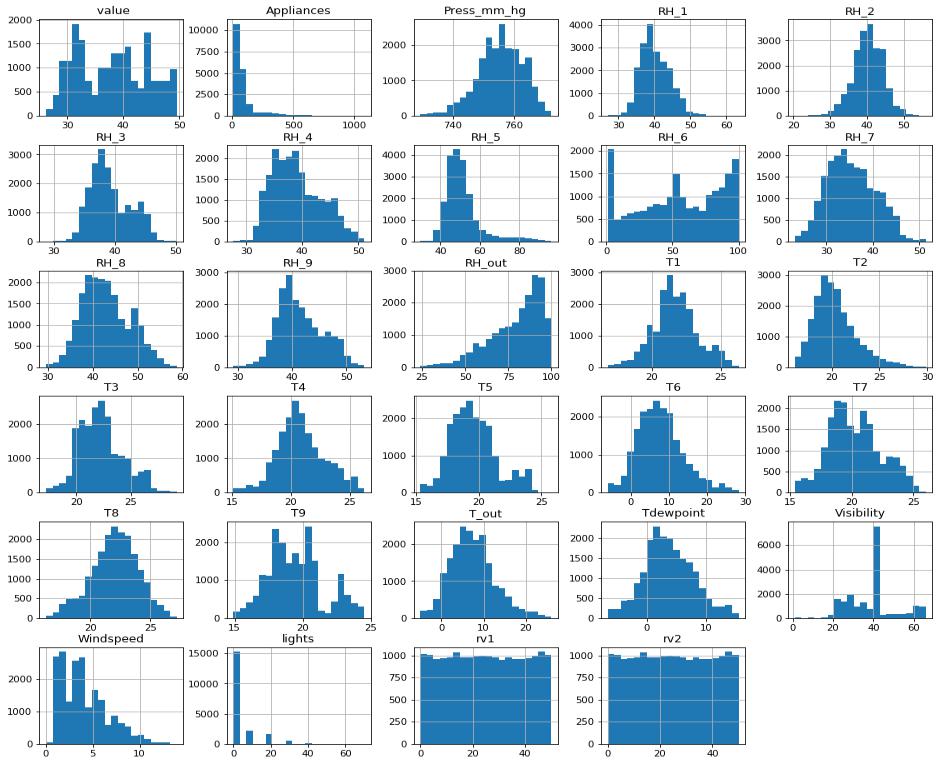


Fig 9

* 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

In Fig 10, On the Target Attribute – Appliance, the below histograms is rightly skewed and most of the data is with 200 KWh.

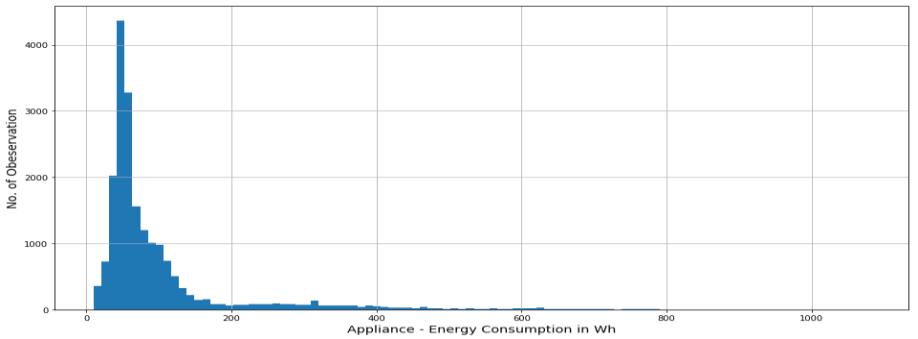


Fig 10

Target variable, Appliances is highly right skewed.

Alternatively exploring using Boxplot – on Appliance Attribute in fig 11, as see that outliers.

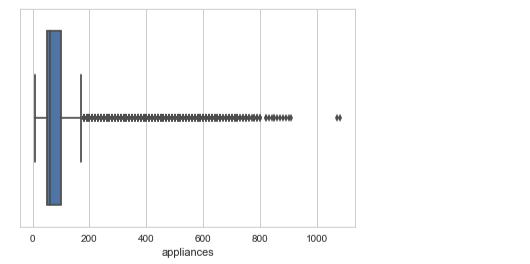


Fig 11

* 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, as shown in fig 12.

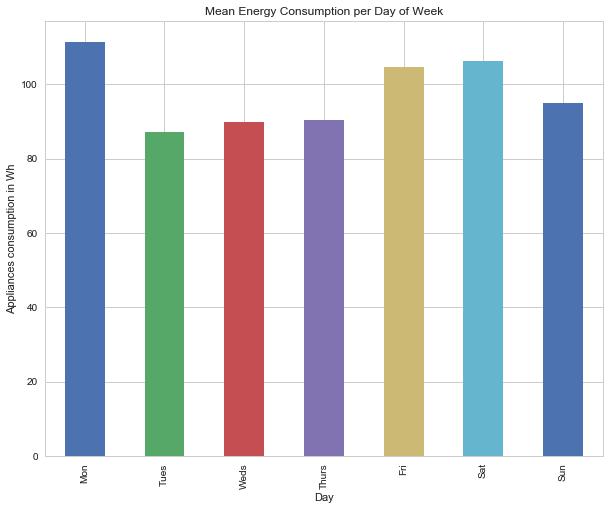


Fig 12

For Monthly Average – On a average, February and April the consumption has been more than other months, as mentioned in fig 13.

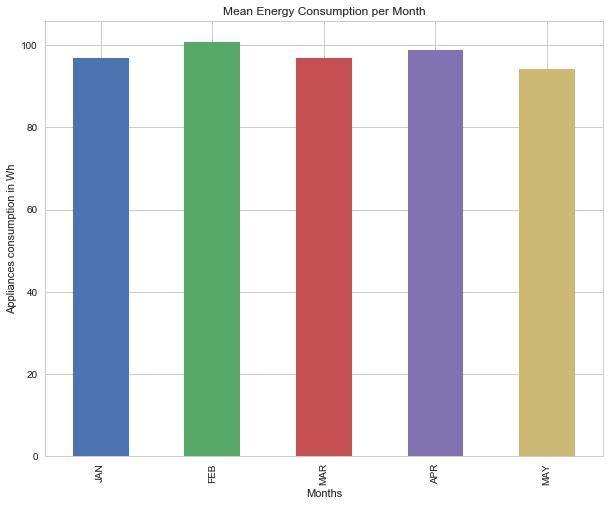


Fig 13

In Fig 14, we can see the time series plot, 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.

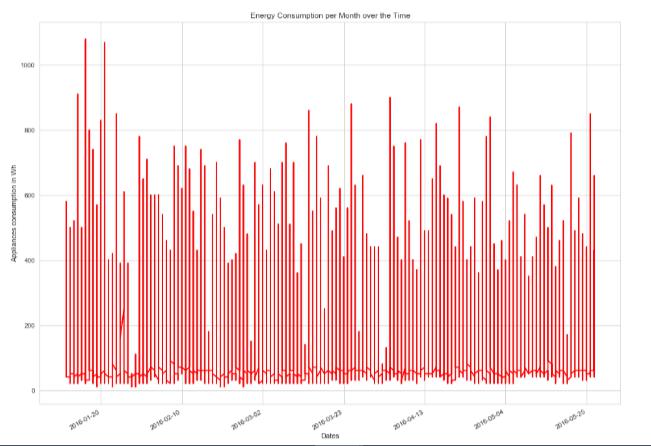


Fig 14

Since the Appliance data captured were rightly skewed, converting the column to Log values to see if it has the normal distribution as shown below in fig 15.

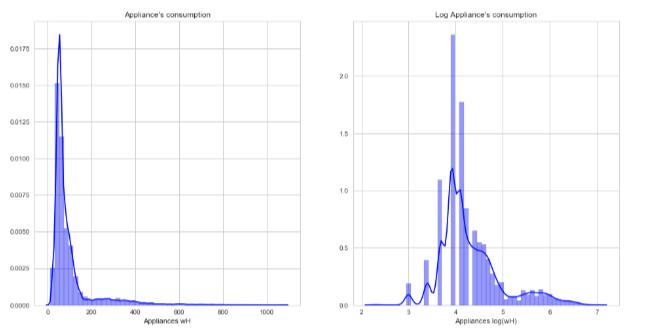


Fig 15

Let’s explore the Correlation plot, in fig 16 –

With Appliance attribute –

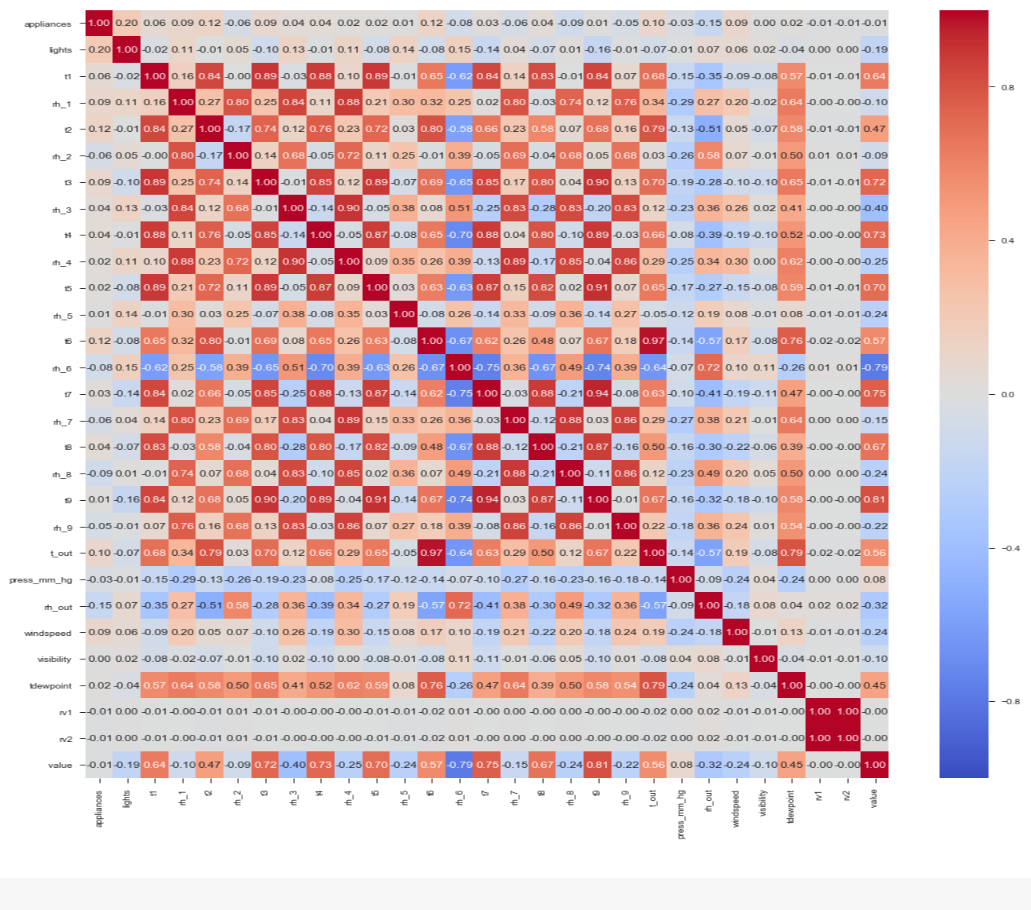


Fig 16

Fig 17 - Scatterplot between appliances and t2

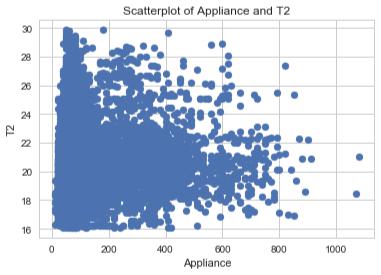


Fig 17

Fig 18 - Scatter plot between Appliance and Lights

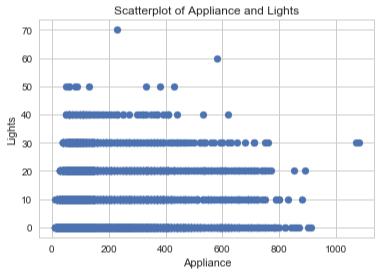


Fig 18

Fig 19 - Scatterplot between Appliance and T6

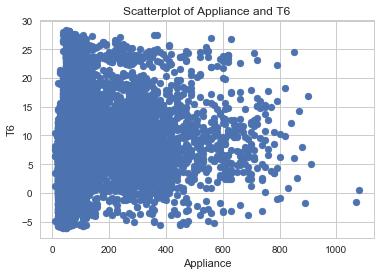


Fig 19

Fig 20 - Scatterplot between Appliance and T\_out

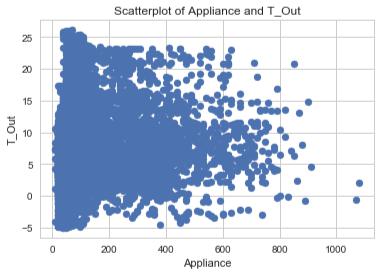


Fig 20

Correlation plot of using Log value of Appliance, as shown in Fig 21 -

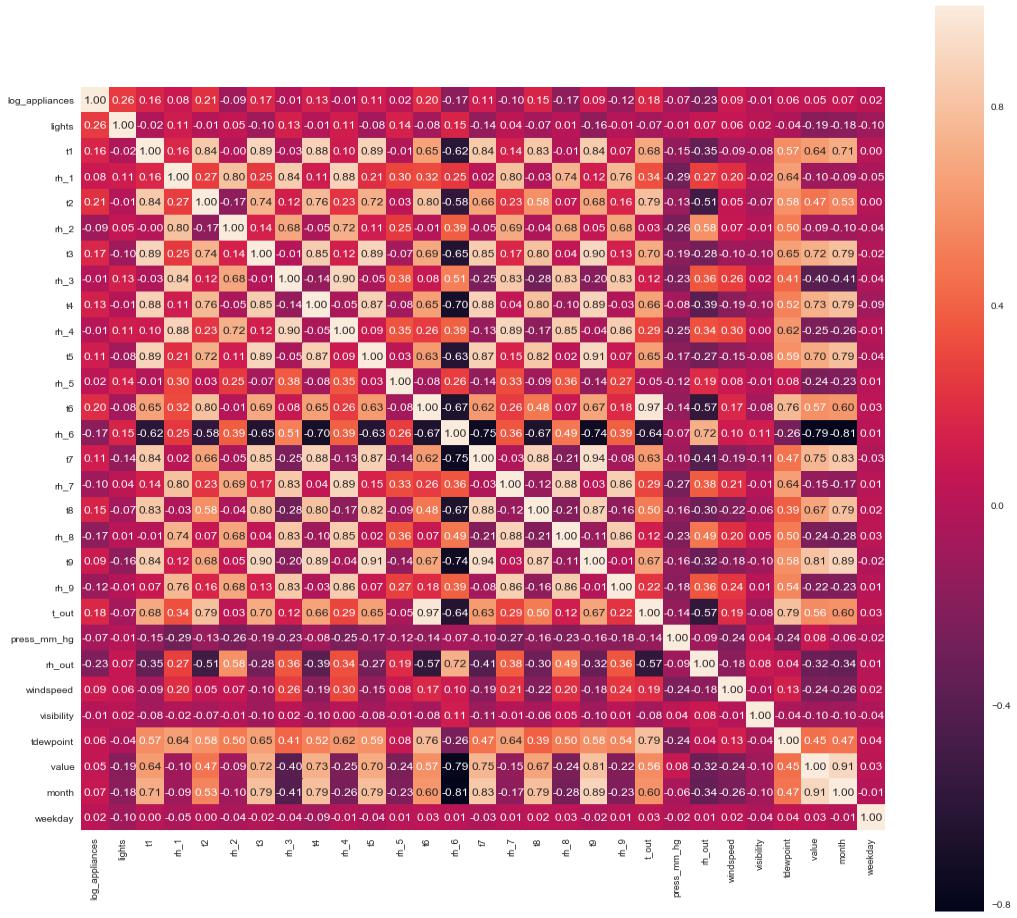


Fig 21

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 be used as features to explain energy consumption otherwise we will have multicollinearity issues.

Fig 22 - Scatter plot of log\_appliance and t1

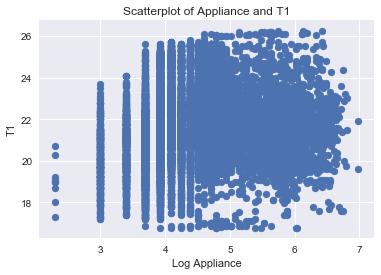


Fig 22

Fig 23 - Scatterplot between log\_appliance and lights

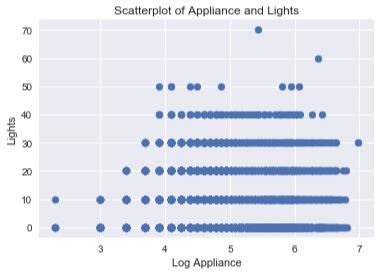


Fig 23

Fig 24 - Scatter plot between log appliance and t\_out

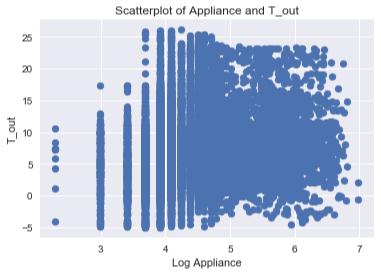


Fig 24

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

* Between Appliance and Lights, as shown in fig 25

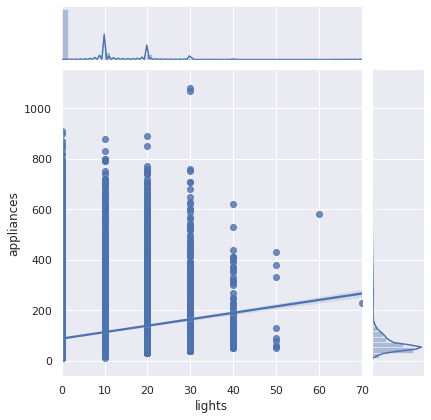


Fig 25

* Between Appliance and Windspeed, as shown in fig 26

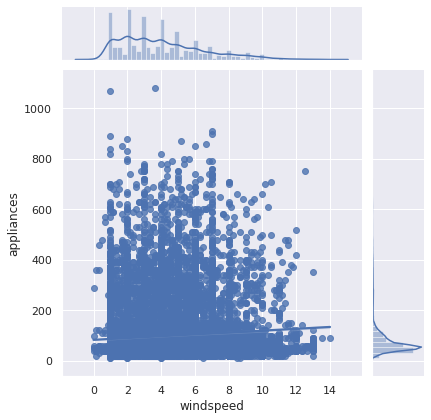


Fig 26

* Between Appliance and T6, as shown in fig 27

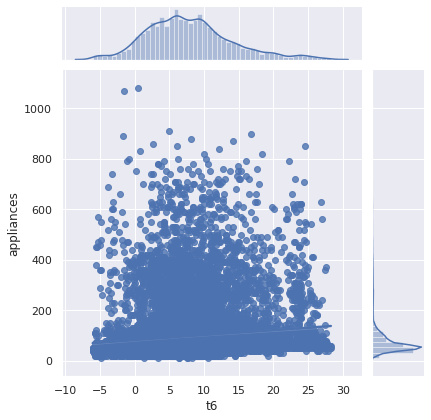


Fig 27

* Between Appliance and T2, as shown in fig 28

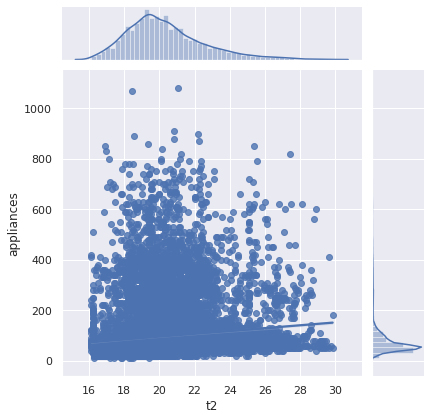


Fig 28

* Between Appliance and T3, as shown in fig 29

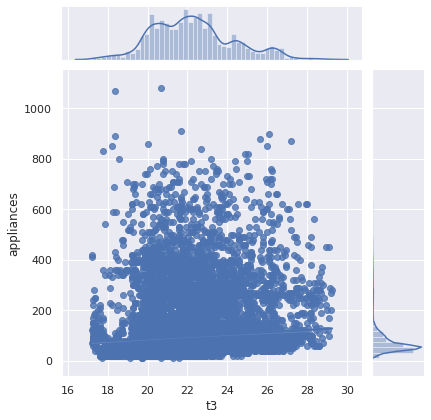


Fig 29

4. Inference Statistics Analysis -

Below are the correlation stats between the temperature dimensions, this will help us to see if any two attributes are redundant or are very similar to each other.

**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. From fig 30, Coefficient table (middle table). We can interpret the t3 coefficient (4.3471) by first noticing that the p-value (under P>|t|) is so small, basically zero. This means that the t3 is a statistical significant predictor of appliance energy consumption.

The regression coefficient for t3 of 4.3471 means that on average, each additional t3 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 𝛽 coefficient is nonzero.

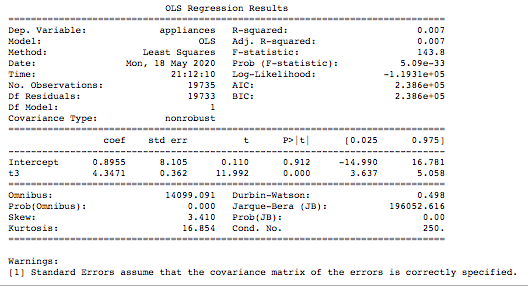


Fig 30

2. From fig 31, Coefficient table (middle table). We can interpret the t3+t6 coefficient (0.4119, 1.8871) by first noticing that the p-value (under P>|t|) is so small, basically zero. This means that the t6 is a statisticall 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 𝛽 coefficient is nonzero.

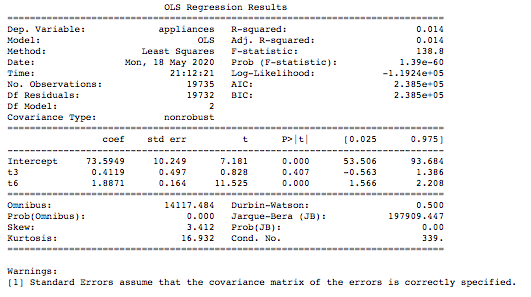


Fig 31

3. From fig 32, 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 t6 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 𝛽 coefficient is nonzero.

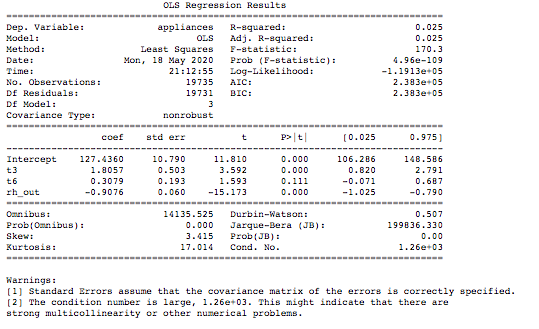


Fig 32

4. From fig 33, 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 statisticall 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 𝛽 coefficient is nonzero.

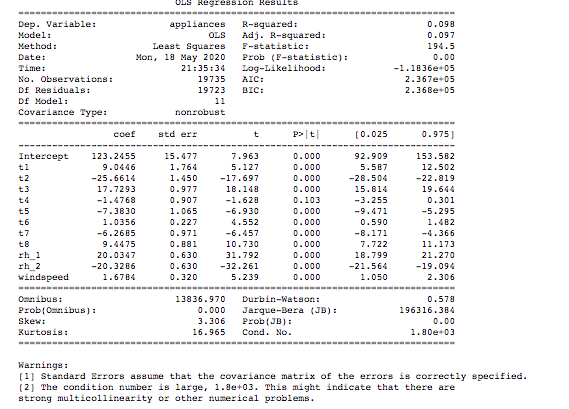


Fig 33

In Fig 35, Pairplot for 't6','t2', 'rh\_2','lights','t\_out','windspeed','tdewpoint' features for their distribution –

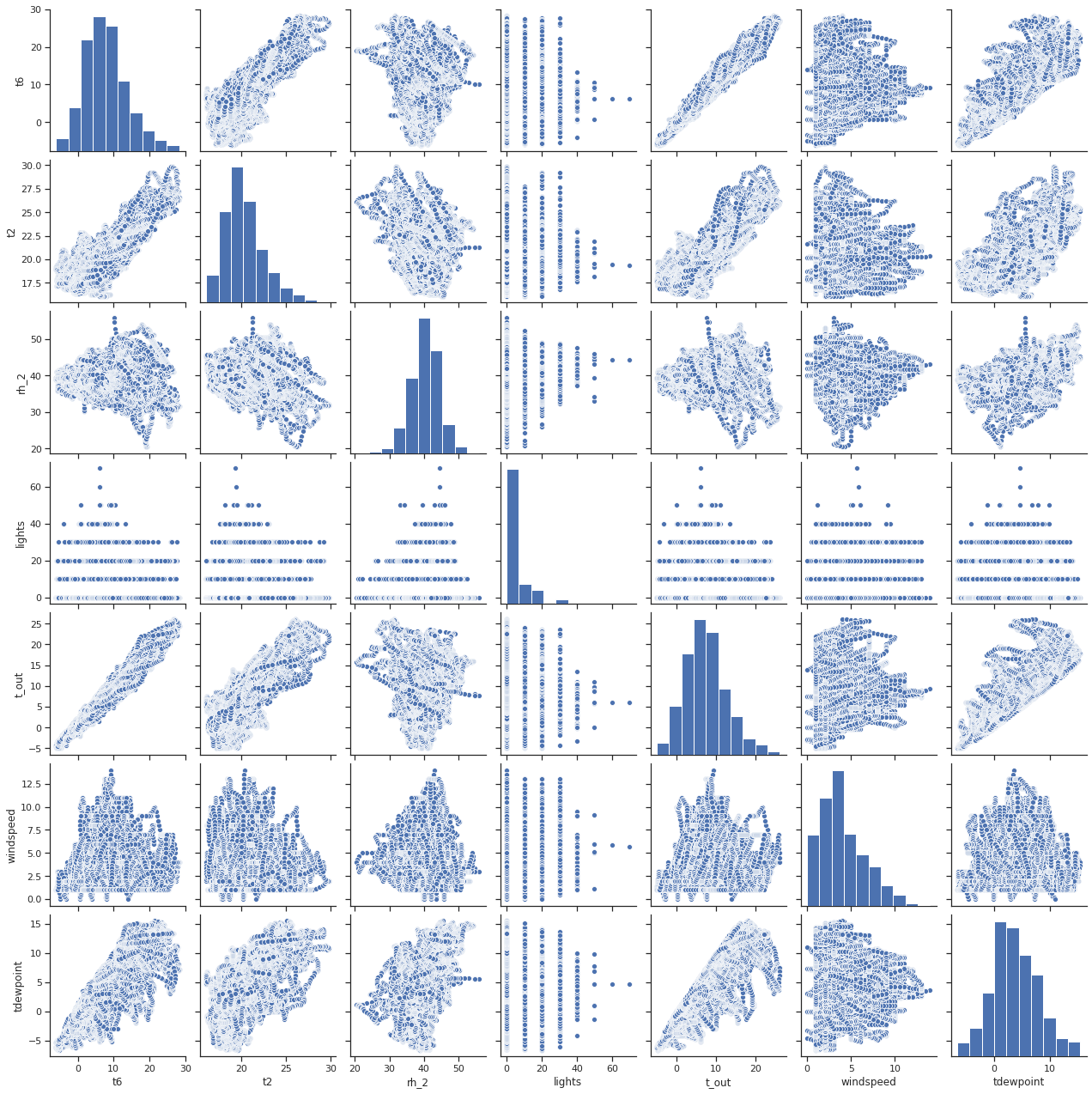


Fig 35

# In Fig 37, Is there a significant difference between T6 and T\_out and impact my future Prediction Models

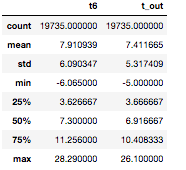


Fig 37

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

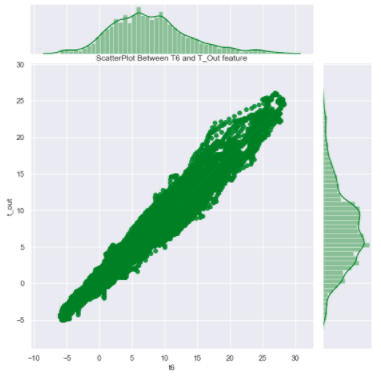


Fig 38

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

# Correlation between all the Features and Target – Appliances – from below Table 12.

|  | **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 |
| **t1** | 0.055447 | 6.449169e-15 |
| **house\_temp** | 0.054740 | 1.411780e-14 |
| **t4** | 0.040281 | 1.507881e-08 |
| **t8** | 0.039572 | 2.683103e-08 |
| **rh\_3** | 0.036292 | 3.402540e-07 |
| **t7** | 0.025801 | 2.890302e-04 |
| **t5** | 0.019760 | 5.503451e-03 |
| **rh\_4** | 0.016965 | 1.715603e-02 |
| **tdewpoint** | 0.015353 | 3.102113e-02 |
| **t9** | 0.010010 | 1.596635e-01 |
| **rh\_5** | 0.006955 | 3.286027e-01 |
| **weekday** | 0.003060 | 6.672580e-01 |
| **visibility** | 0.000230 | 9.741858e-01 |
| **week** | -0.011356 | 1.106606e-01 |
| **month** | -0.011606 | 1.030264e-01 |
| **value** | -0.013535 | 5.725207e-02 |
| **house\_hum** | -0.020075 | 4.799007e-03 |
| **press\_mm\_hg** | -0.034885 | 9.493922e-07 |
| **rh\_9** | -0.051462 | 4.697109e-13 |
| **rh\_7** | -0.055642 | 5.187296e-15 |
| **rh\_2** | -0.060465 | 1.873022e-17 |
| **rh\_6** | -0.083178 | 1.209481e-31 |
| **rh\_8** | -0.094039 | 5.211566e-40 |
| **rh\_out** | -0.152282 | 1.077516e-102 |

Table 12

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

* Create the dataframe with “lights” and “appliances”, as shown in fig 39
* 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

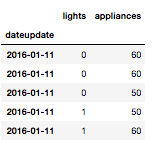


Fig 39

* 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 Statistical 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.