Capstone Project 1: Statistical Data Analysis

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_stats.ipynb

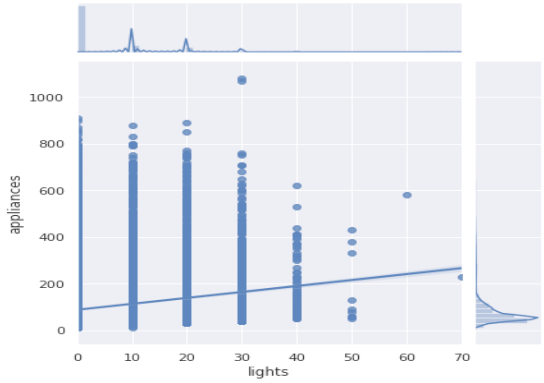
From the Data Wrangling activity, we created the **input.csv** as the final dataset. This has 19735 observations and 30 attributes.

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

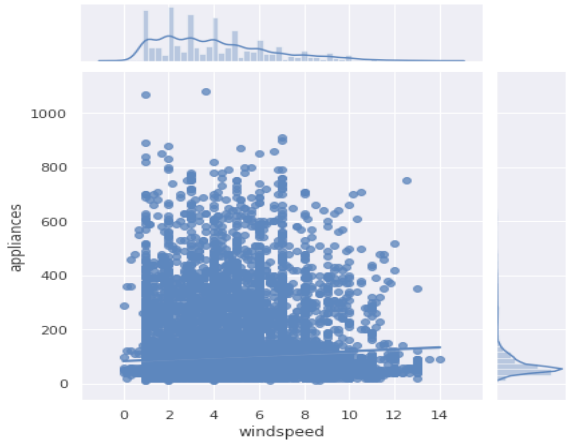
```
# Temperature sensors columns
temp_cols = ["T1", "T2", "T3", "T4", "T5", "T6", "T7", "T8", "T9"]

# 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", "V
isibility"]

# Target column
tgt = ["Appliances"]
```

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

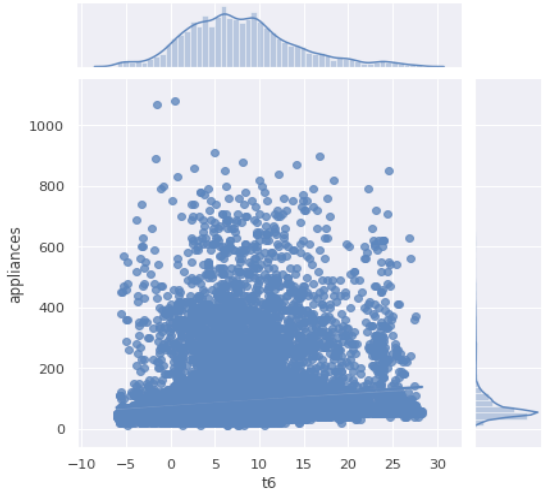
- Between Appliance and Lights



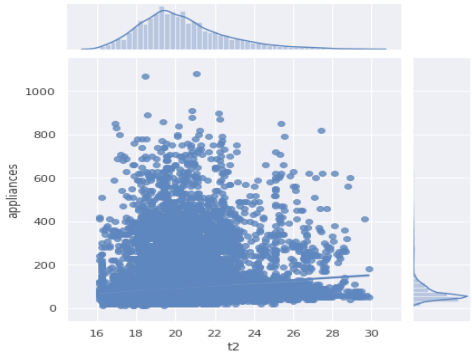
- Between Appliance and Windspeed



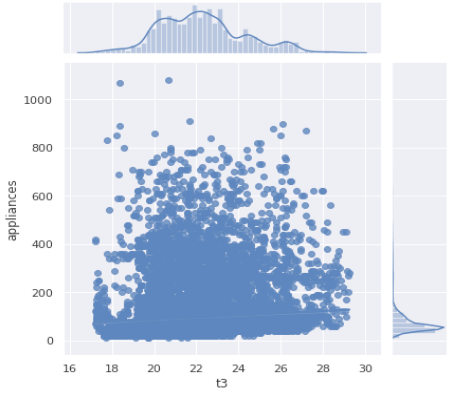
- Between Appliance and T6



- Between Appliance and T2



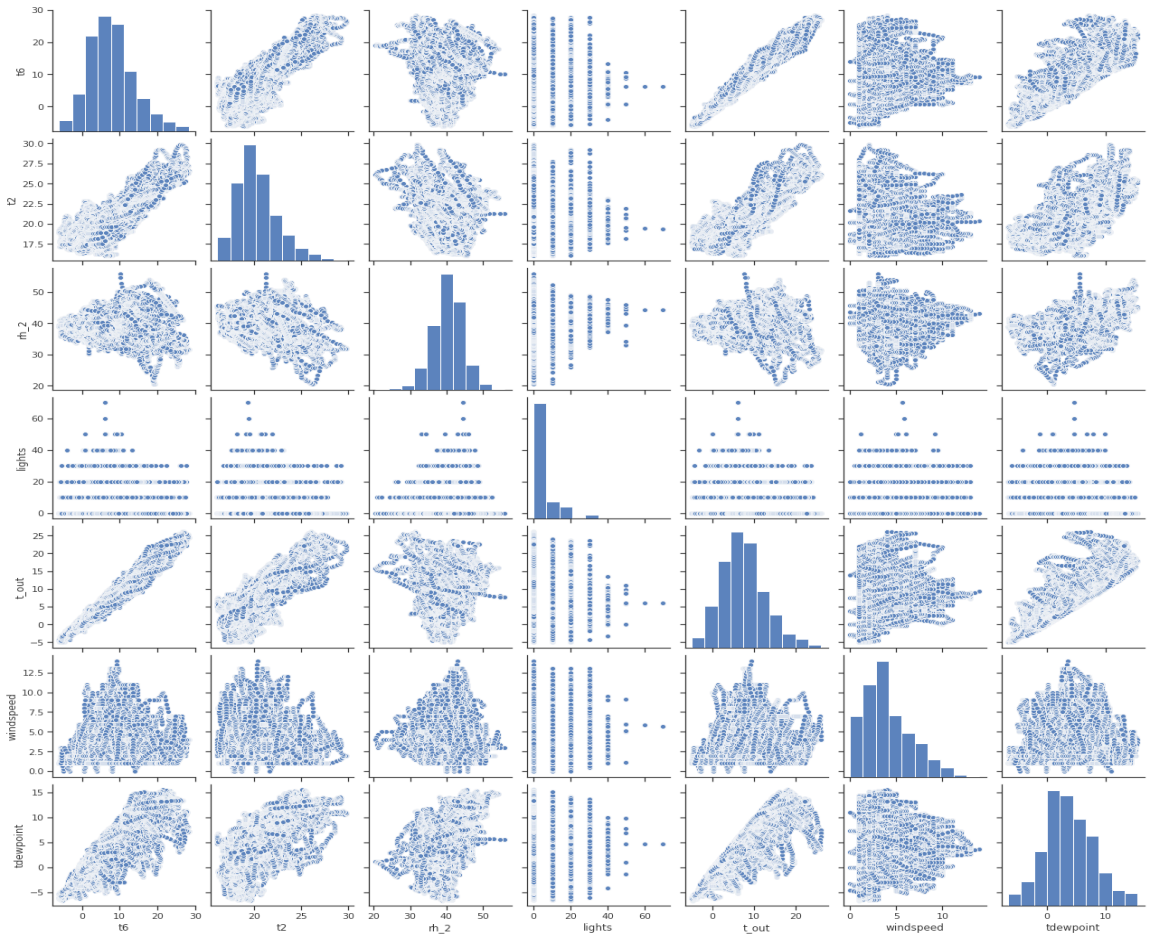
- Between Appliance and T3



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

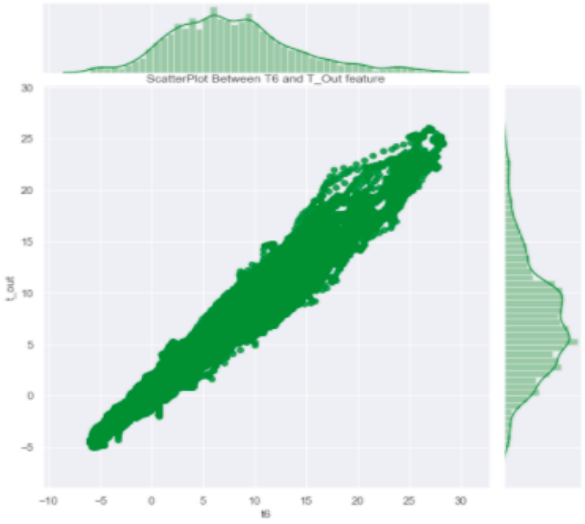
Pairplot for 't6','t2', 'rh_2','lights','t_out','windspeed','tdewpoint' features for their distribution –



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
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

	Correlation coefficients	p-value
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

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

dateupdate	lights	appliances
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.

Inference as Below –

Temperature - All the temperature variables from T1-T9 and T_out have positive correlation with the target Appliances . For the indoortemperatures, the correlations are high as expected, since the ventilation is driven by the HRV unit and minimizes air tempera-ture differences between rooms. 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 T6 & T9 can be removed from training set as information provided by them can be provided by other fields.

Weather attributes - Visibility, Tdewpoint, Press_mm_hg have low correlation values

Humidity - There are no significantly high correlation cases (> 0.9) for humidity sensors.

Random variables have no role to play