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

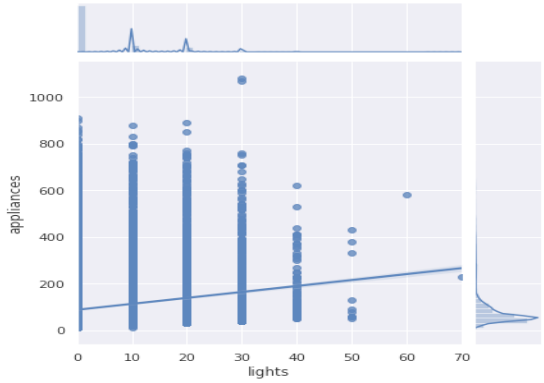
From the Data Wrangling activity, we created the **input.csv** as the final dataset. This has 19735 observations and 30 attributes.

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

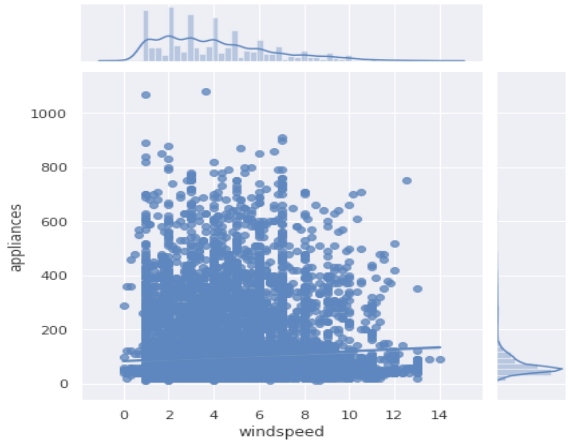
```
# Temperature sensors columns
temp_cols = ["T1", "T2", "T3", "T4", "T5", "T6", "T7", "T8", "T9"]

# 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 –

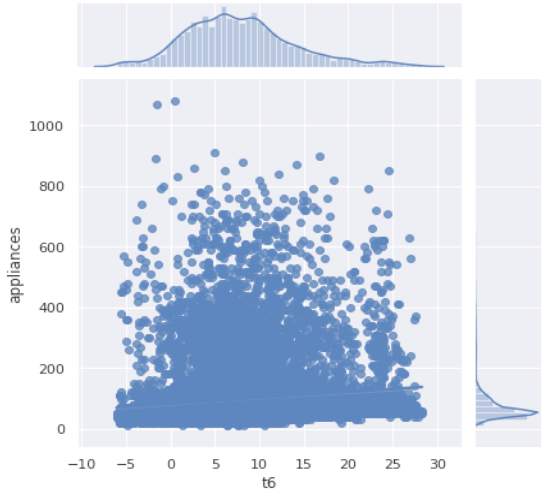
- Between Appliance and Lights



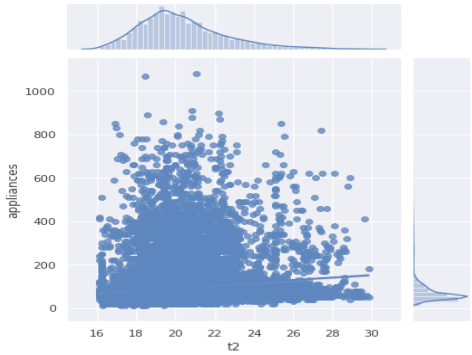
- Between Appliance and Windspeed



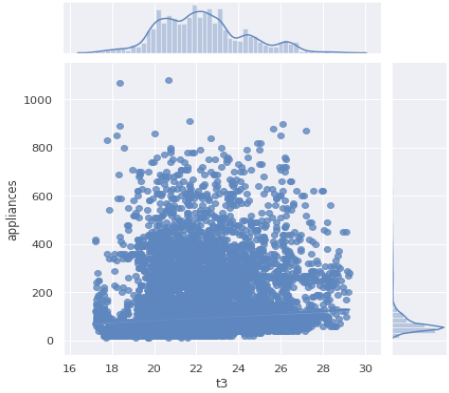
- Between Appliance and T6



- Between Appliance and T2



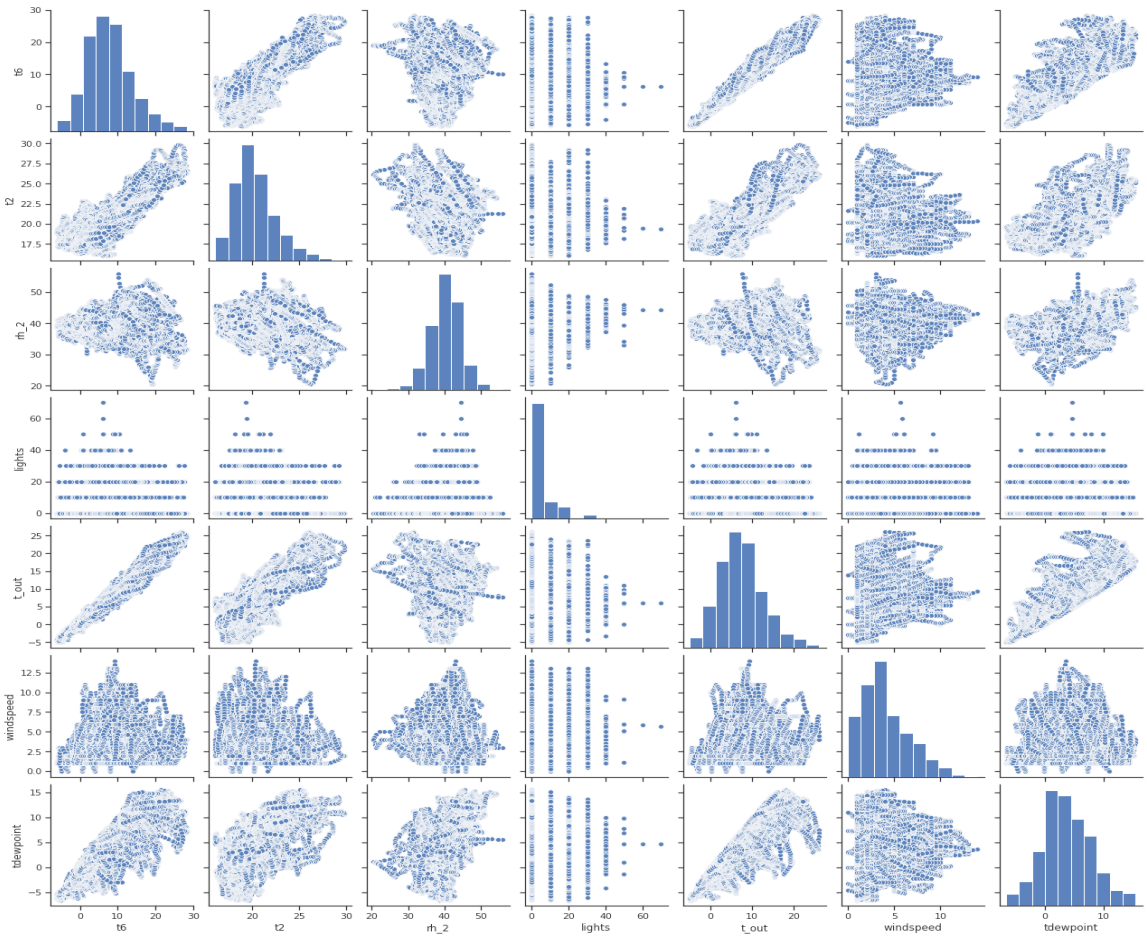
- Between Appliance and T3



Calculate the Correlation between Temperature features -

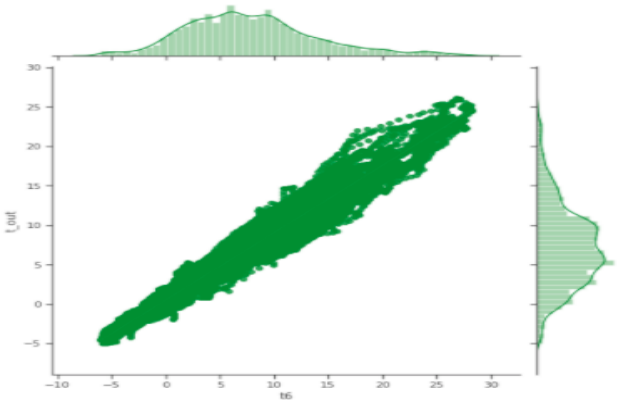
- Correlation between T9 and T1 pearson 0.84 0.00 None
- Correlation between T9 and T2 pearson 0.68 0.00 None
- Correlation between T9 and T3 pearson 0.90 0.00 None
- Correlation between T9 and T4 pearson 0.89 0.00 None
- Correlation between T9 and T5 pearson 0.91 0.00 None
- Correlation between T9 and 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 –



<u>Is there a significant difference between T6 and T_out and impact my future Prediction Models –</u>

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



Run a Two-sided T-test with the following hypotheses:

Null hypothesis: t6 = t_out

Alternate hypothesis: t6 != t_out
Upon Conducting the T-Test – received the below values - Ttest_indResult (statistic=8.675177895656354, pvalue=4.283728402821399e-18)

Result - Given the high p-value: 4.2, hence will not reject the null hypothesis that feature t6 and t_out almost same and redundant.