Advances in Data Science and Architecture

Machine learning with Energy systems

Overview :
Abstract
Introduction
Exploratory Data Analysis
Feature Engineering
Feature Selection
Model Validation and Selection
Final pipeline

Abstract

AdaptiveAlgo Systems Inc. works on solutions to build algorithms and platforms to address energy modeling challenges. And require a solution for energy modelling and is interested in understanding consumer energy usage. As data scientist we build various machine learning models that could contribute to understanding energy usage by appliances and the attributes that contribute to aggregate energy usage.

Introduction

Building energy use prediction plays an important role in building energy management and conservation as it can help us to evaluate building energy efficiency, conduct building commissioning, and detect and diagnose building system faults. In this context, a proper prediction of energy demand in housing sector is very important. We are building a model by analysing different prediction models, feature engineering, and model selection processes. This would in turn aid AdaptiveAlgo Systems Inc. to build algorithms and systems to understand consumer behavior to help them make better decisions. The case study will be conduct an in-depth analysis to provide insights on feature engineering and machine learning with provided dataset.

Exploratory Data Analysis

The Dataset consists of 137 days of energy load on a single house from various source's such as house appliances, temperature and humidity values in different rooms and weather data from nearest airport weather station for every 10 min to capture quick change in the energy usage. On exploring the given dataset, lot of information was analyzed regarding the dependent and independent variables. The target variable in the dataset is 'Appliances'. Apart from target variable, the dataset contains 27 other variables which are considered as predictors before going further with any of the feature selection process. The index of the dataset is DateTime value data type. To understand the data pattern, three additional variables which are extracted from existing ones are added to the dataset. Those are 'Num_sec_midnight' that indicates number of seconds from midnight which indirectly gives you time of each observation, 'Day_status' gives the day of each observation and 'week_status' states whether a observation is week end or week day.

Table illustrating Data variables in the given Dataset:

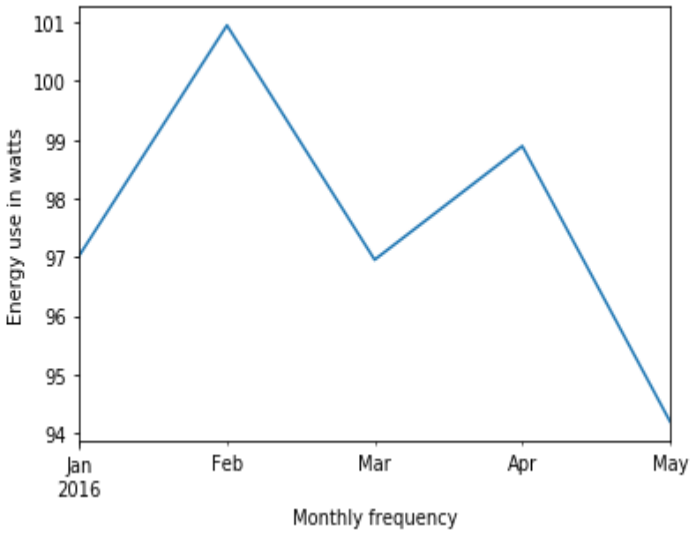
Data variables	Units	Number of features
Appliances energy consumption	Wh	1
Light energy consumption	Wh	2
T1, Temperature in kitchen area	°C	3
RH1, Humidity in kitchen area	%	4
T2, Temperature in living room area	°C	5
RH2, Humidity in living room area	%	6
T3, Temperature in laundry room area	°C	7
RH3, Humidity in laundry room area	%	8
T4, Temperature in office room	°C	9
RH4, Humidity in office room	%	10
T5, Temperature in bathroom	°C	11
RH5, Humidity in bathroom	%	12
T6, Temperature outside the building (north side)	°C	13
RH6, Humidity outside the building (north side)	%	14
T7, Temperature in ironing room	°C	15
RH7, Humidity in ironing room	%	16
T8, Temperature in teenager room 2	°C	17
RH8, Humidity in teenager room 2	%	18
T9, Temperature in parents room	°C	19
RH9, Humidity in parents room	%	20
To, Temperature outside (from Chièvres weather station)	°C	21
Pressure (from Chièvres weather station)	mm Hg	22
RHo, Humidity outside (from Chièvres weather station)	%	23
Windspeed (from Chièvres weather station)	m/s	24
Visibility (from Chièvres weather station)	km	25
Tdewpoint (from Chièvres weather station)	°C	26
Random Variable 1 (RV_1)	Non dimensional	27
Random Variable 2 (RV_2)	Non dimensional	28

Note: Random Variable 1 & Random Variable 2 are randomly added by default to test Boruta feature selection

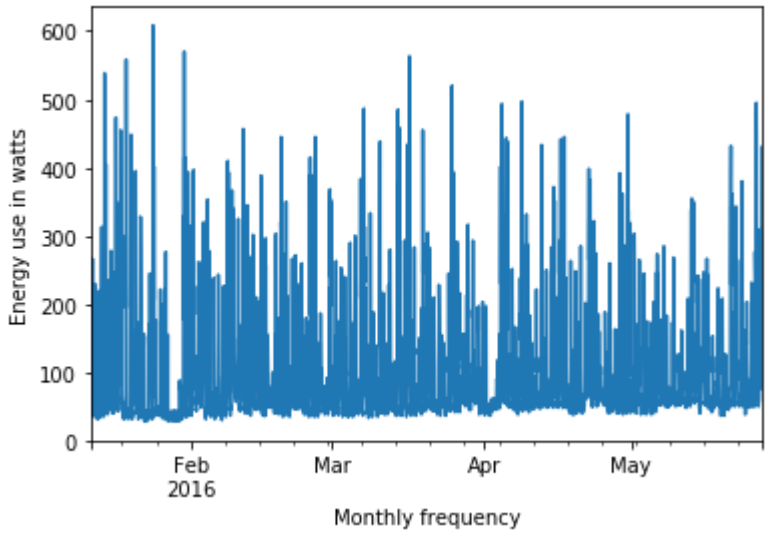
The below plot describes the 'Appliances' variable for 137 days with mean values sampled monthly. The highest energy usage was on February – 2016 compared to other months. Further study is required to get more details.

Plot 1

Date	Mean values
2016-01-31	97.026010
2016-02-29	100.945881
2016-03-31	96.953405
2016-04-30	98.888889
2016-05-31	94.199325

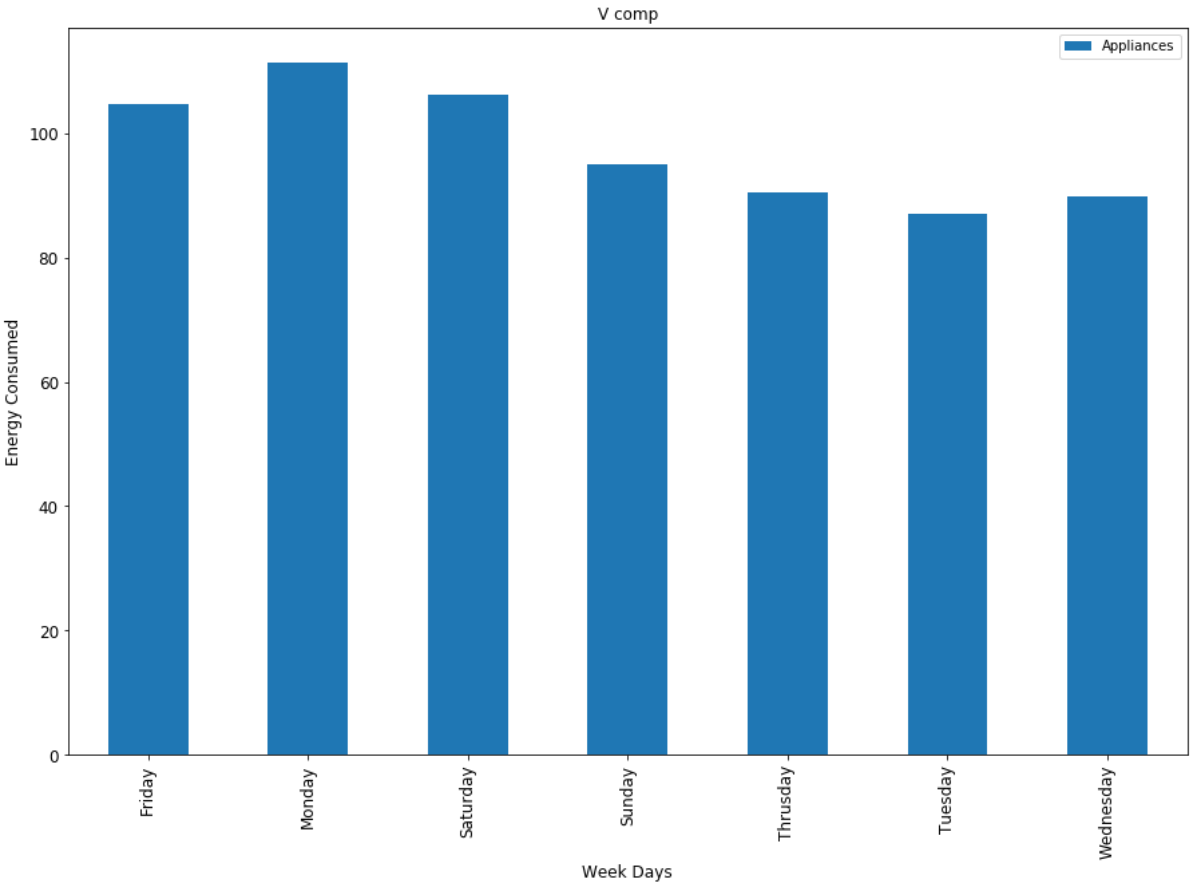


Plot 2



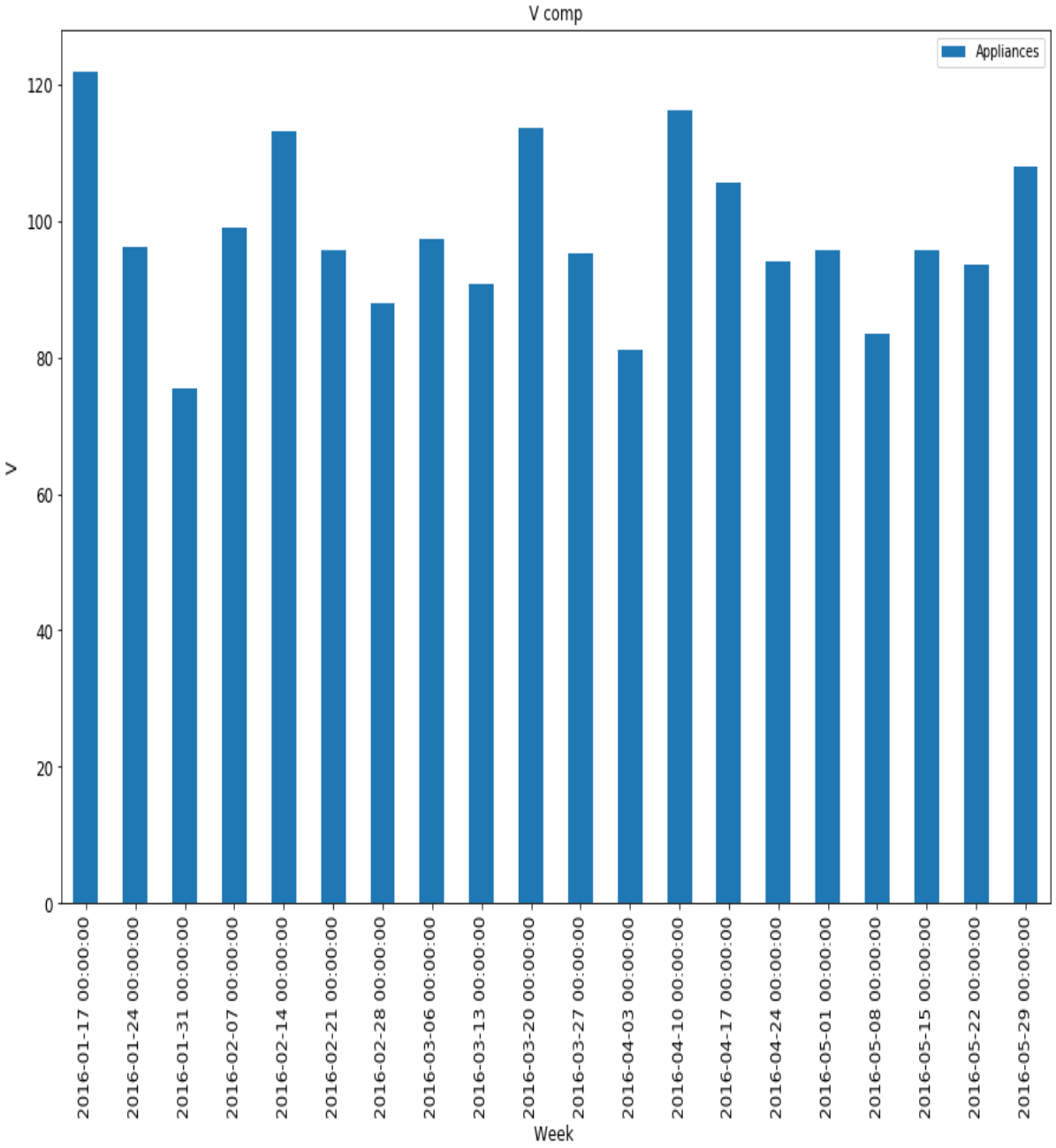
Plot 3 sums up appliances value by weekly for all 5 months and from which we can infer that more of energy was consumed on Tuesday

Plot 3



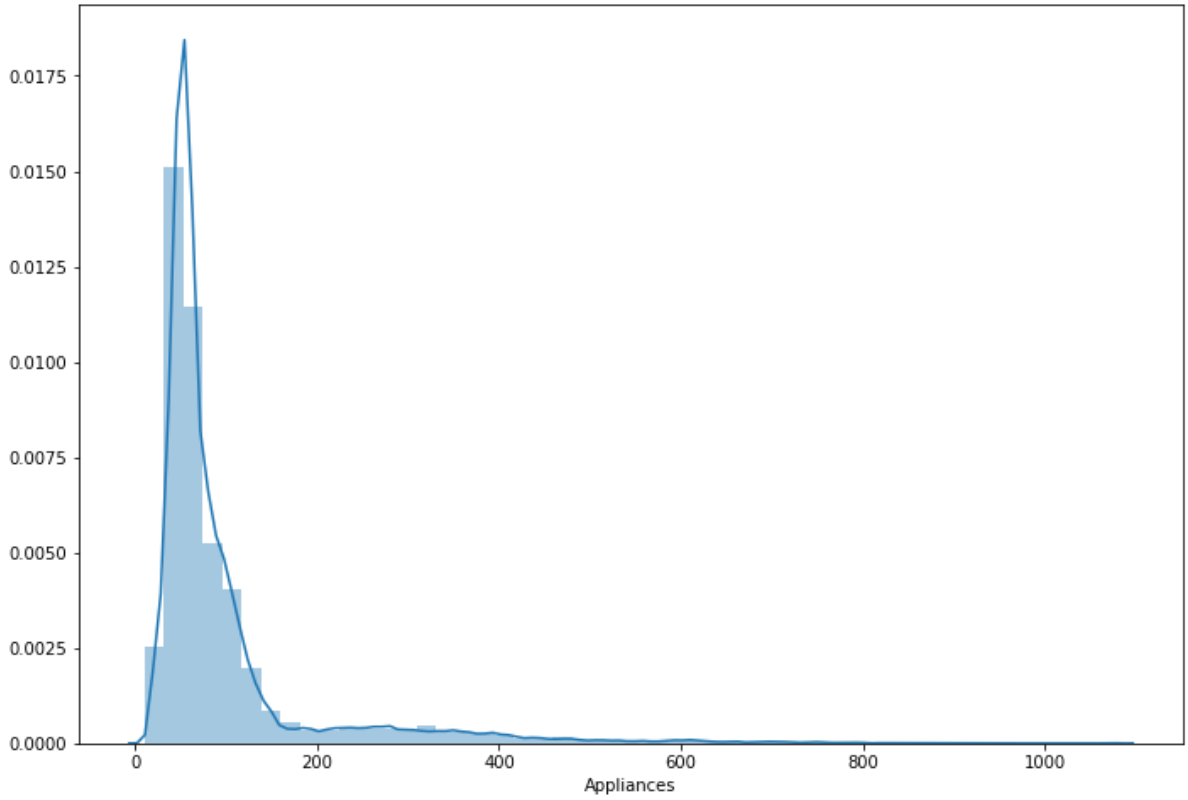
Below ***Plot 4*** gives the energy usage for the 20 weeks on the given data by its mean value.

Plot 4



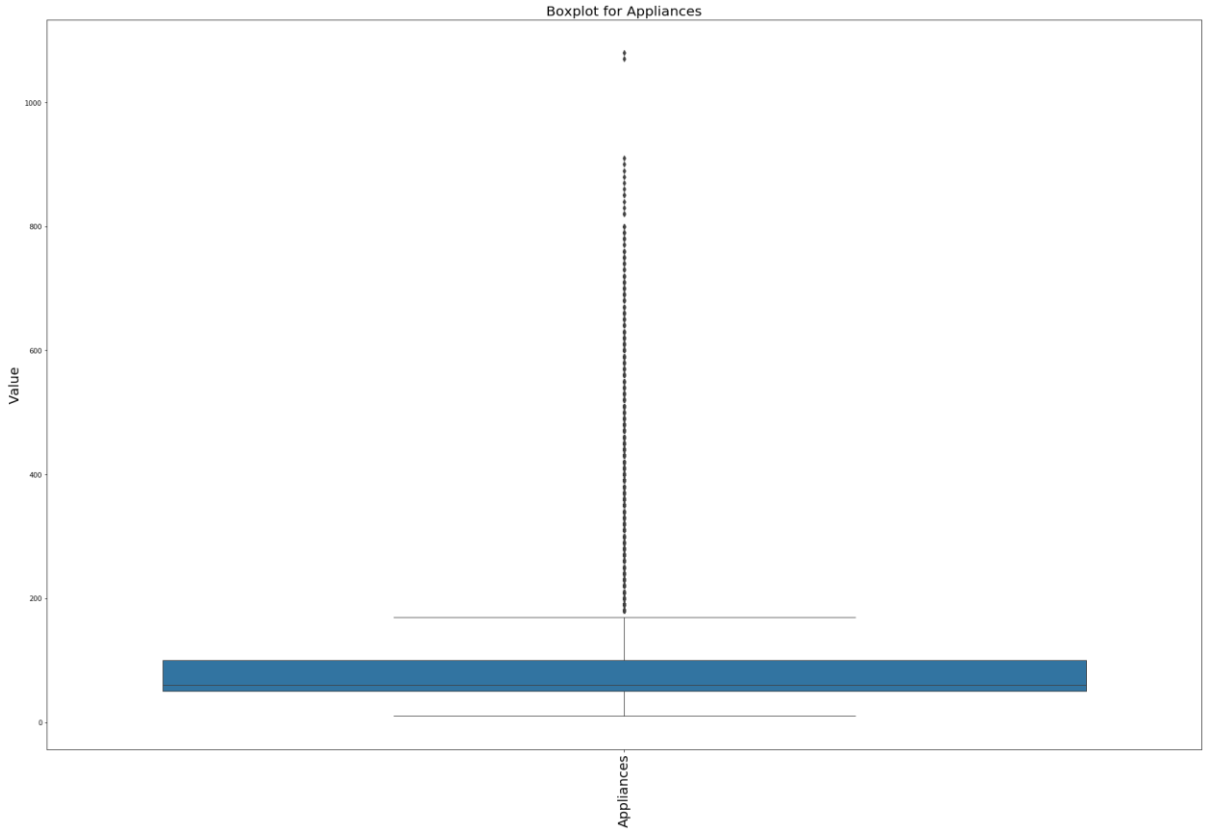
Plot 5 illustrates the data distribution in the target variable. Most of the energy values are between 10 to 180 watts. There are some extreme values in the plot which are few.

Plot 5 – Histogram of Appliance variable



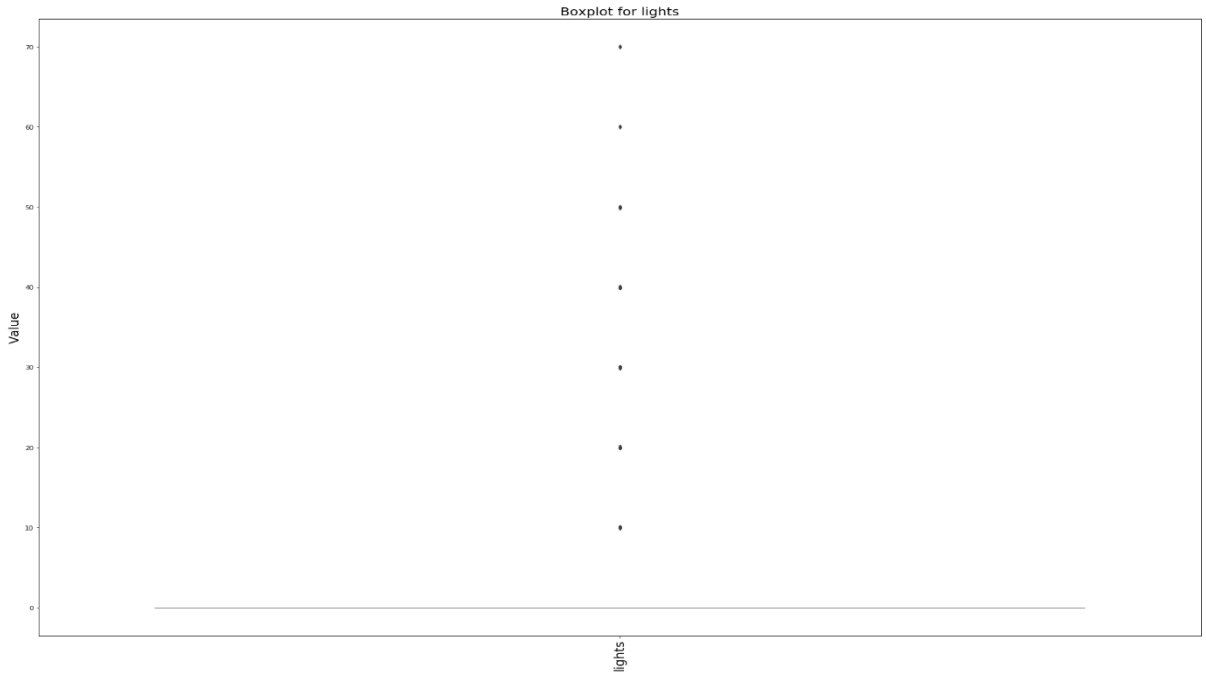
Plot 6 gives boxplot for appliances that shows the extremes values recorded in the data set

Plot 6 – Boxplot for Appliances



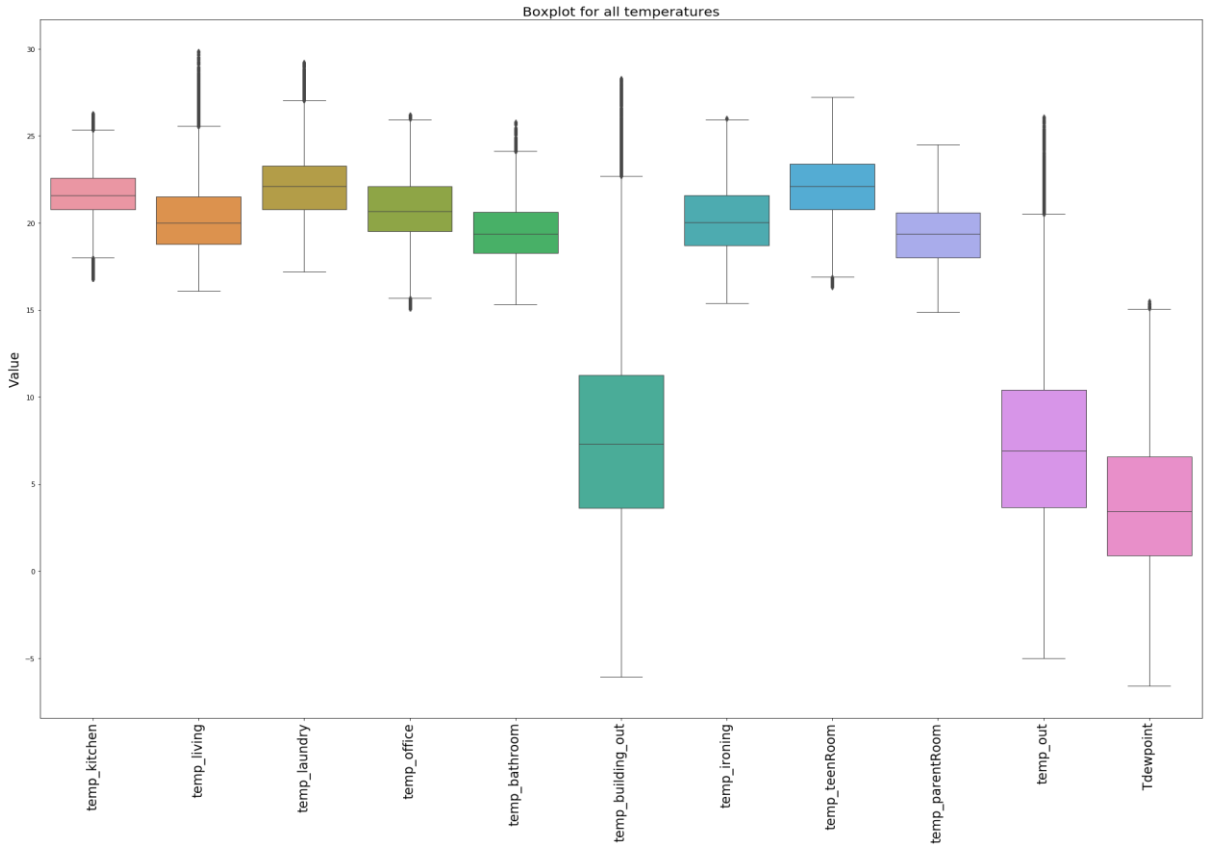
Plot 7 shows the boxplot for Light variable. This clearly shows that 75% of the values are '0'

Plot 7 – Boxplot for Lights



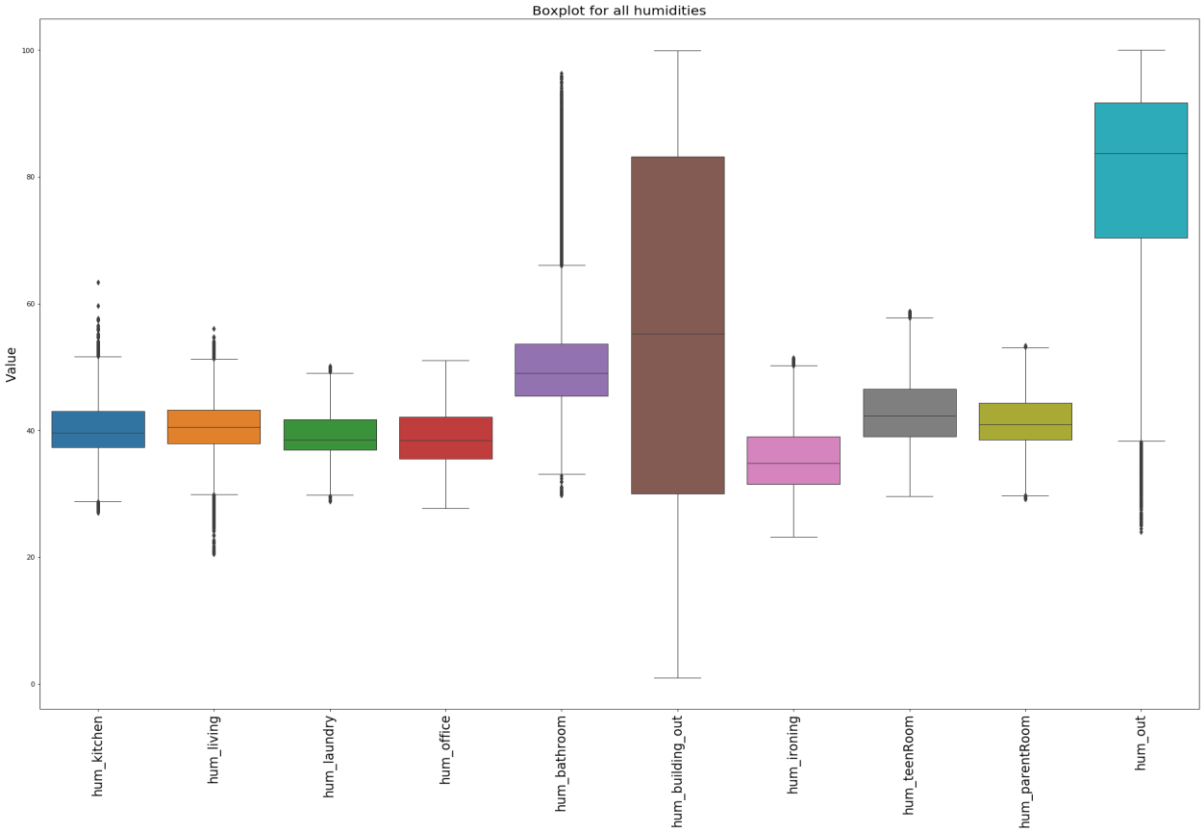
Plot 8 gives the boxplot for all the temperature variable in the Dataset.

Plot 8 – Boxplot for all temperature variables

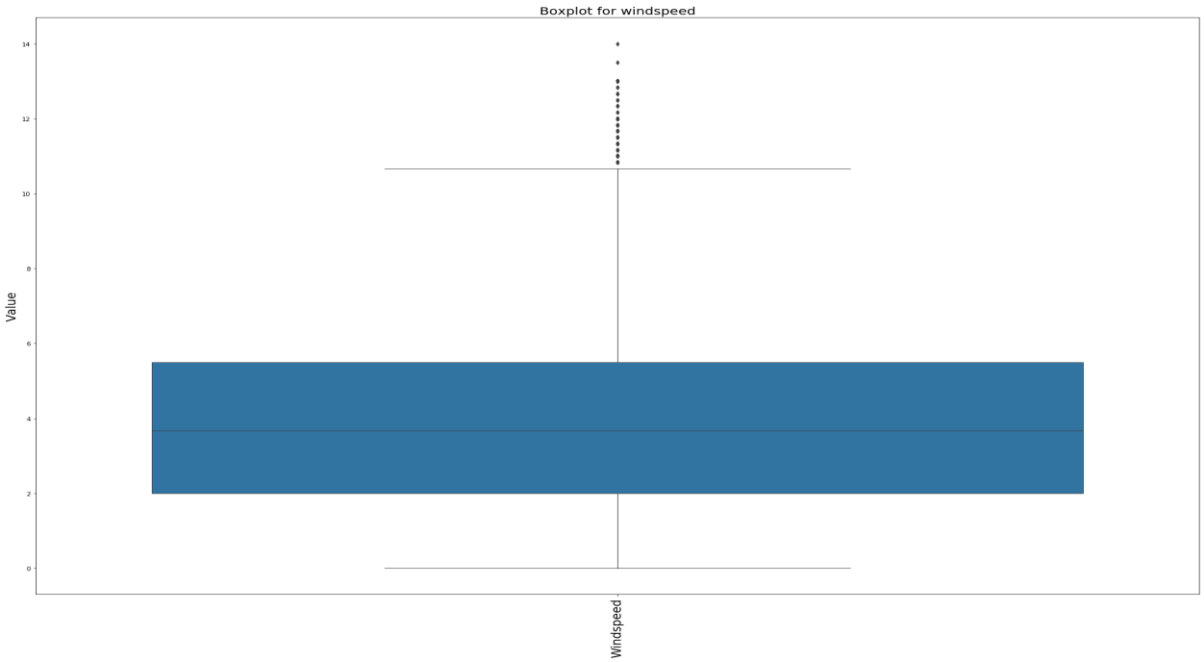


Plot 9 gives the boxplot for all humidity variable in the Dataset

Plot 9- Boxplot for all Humidity variables



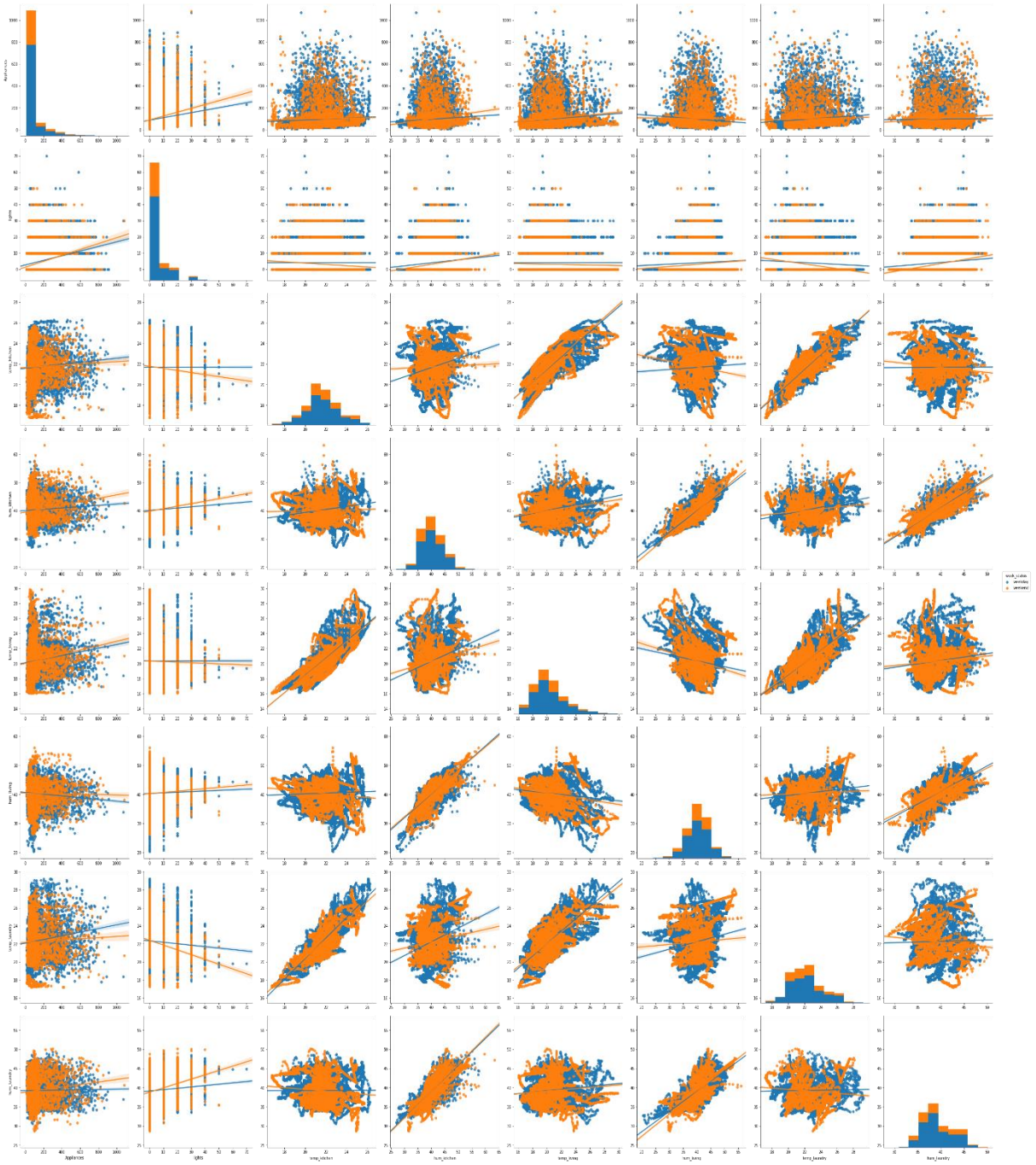
Plot 10- Boxplot for Windspeed



Inorder to find the correlation with each variable, pair plot and correlation matrix was plotted against all 31 variables (including newly added features).

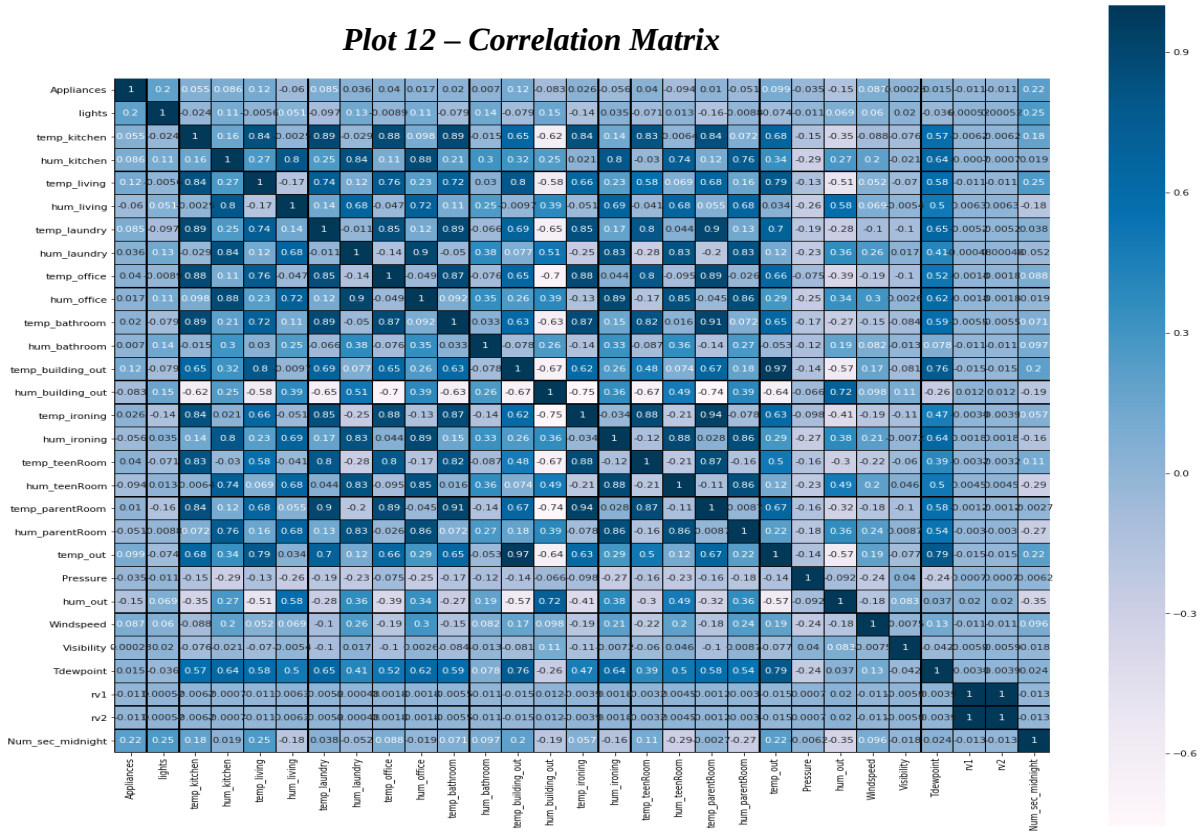
Pair plot are plotted in such a way to discriminate weekday and week-end values. Since there will always be a difference in energy usage between a week-end and weekday.

Plot 11 – One of the pair plot with 8 variables



Inorder to understand the correlation between variables on values bases refer ***Plot 12.*** The below correlation matrix clearly illustrates that there is a high correlation between 'Num_sec_midnight' and 'Appliances' with a value of 0.22. The second largest correlation with 'Appliances' is 'Lights' with a value of 0.19. For the indoor temperatures, the correlations are high as expected, since the ventilation is driven by the HRV unit and minimizes air temperature differences between rooms. For example, a positive correlation is found with 'temp_kitchen' and 'temp_laundry'. And correlation of outdoor temperature with the appliances is 0.12. There is also a negative correlation between the appliances and outdoor humidity/RH6 (−.09. There is a positive correlations between the consumption of appliances and temp_ironing, temp_teenRoom and temp_parentRoom being 0.03, 0.05 and 0.02 respectively. A positive correlation of 0.10 is seen between appliances' consumption and outdoor temperature (Tout) that is, the higher temperatures, the higher the energy use by the appliances. Also there is a positive correlation with appliances' consumption and wind speed (0.09), higher wind speeds correlate with higher energy consumption by the appliances. A negative correlation of −0.15 was found with the 'hum_out', and of −0.03 with pressure. Another important and interesting correlation is between the pressure and the wind speed. This relationship is negative (−0.23)

Plot 12 – Correlation Matrix



Plot 13 – provides the value of all variables for given dates. It helps us compare values on three different dates.

In [72]:
```
def plotDatetimeWise(datetime):
    fig, ax = plt.subplots()
    count = 1
    for time in datetime:
        newdf = Data[Data.columns[0:26]]
        df_player = newdf.loc[time]
        df_player = df_player.astype('int')
        df_player.T.plot.line(figsize = (24,12),ax=ax)
        ax.legend()
        ax.set_xlabel('Columns',fontsize=18)
        ax.set_ylabel('Values',fontsize=18)
        ax.set_xticks(np.arange(len(newdf.columns)))
        ax.set_xticklabels(labels = newdf.columns, rotation=90,fontsize=15)
        count = count + 1
plotDatetimeWise(['2016-01-20 00:00:00', '2016-05-21 11:00:00', '2016-02-10 01:50:00'])
```