Energy Usage- Report

Contents

[1. Introduction 1](#_Toc497338412)

[1.1. Problem Statement 1](#_Toc497338413)

[1.2. Dataset 1](#_Toc497338414)

[2. Data Wrangling 2](#_Toc497338415)

[2.1. Handling inconsistent column names and datatype 3](#_Toc497338416)

[2.2. Handling columns with repetitive values 4](#_Toc497338417)

[2.3. Handling the outliers 4](#_Toc497338418)

[2.4. Write the clean data into a new file for further steps 5](#_Toc497338419)

[3. EDA and Data storytelling 5](#_Toc497338420)

[3.1. Multicollinearity 5](#_Toc497338421)

[3.2. Some data facts 7](#_Toc497338422)

# Introduction

## Problem Statement

To Create regression models of appliances energy use in a low energy building.

## Dataset

The data set is at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored with a ZigBee wireless sensor network. 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. Weather from the nearest airport weather station (Chievres Airport, Belgium) was downloaded from a public data set from Reliable Prognosis (rp5.ru), and merged together with the experimental data sets using the date and time column. Two random variables have been included in the data set for testing the regression models and to filter out non-predictive attributes (parameters).

Dataset contain 29 attributes and 19735 observations.

Link – <http://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction>

# Data Wrangling

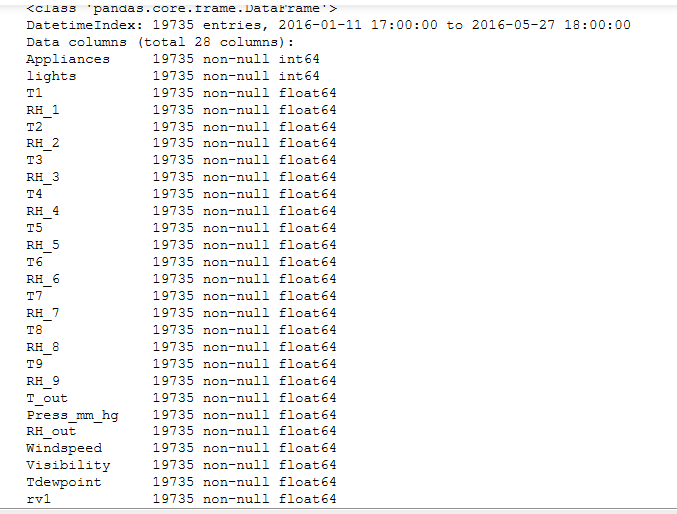
**Goal:** Prepare the dataset for EDA and Modeling

**Tasks performed:**

* Handling inconsistent column names and datatype
* Missing Data handling
* Removal of duplicate rows
* Handling the outliers
* Write the clean data into a new file for further steps

**Data load and description:**

Dataset had 19735 rows and 28 columns



## 

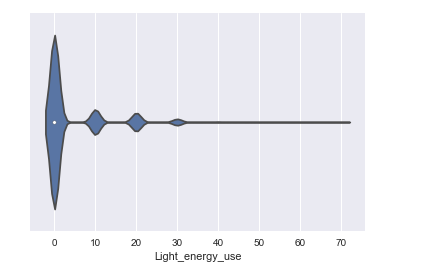
## Handling inconsistent column names and datatype

Columns were with mixed data types. Some of the column names had to be changed for better readability.

Date column specifies the time interval when these observations are captured. I indexed the dataframe with date which gives us better analysis techniques.

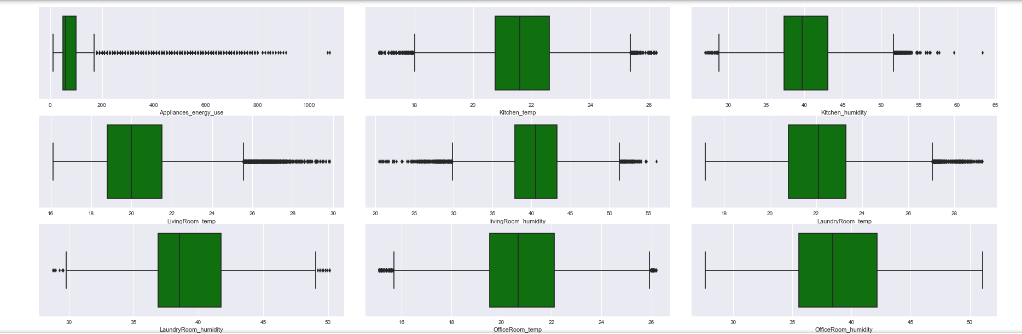
## Handling columns with repetitive values

For the Light\_energy\_use column we had **more than 75% of the data as 0**. So, I preferred to drop this column.



## Handling the outliers

Outliers can be found using various methods. I use Median absolute deviation method and box and whisker to find out the outliers.



Outliers are removed from - Kitchen\_temp, Kitchen\_humidity, LaundryRoom\_temp, LaundryRoom\_humidity, OfficeRoom\_temp, IroningRoom\_temp, IroningRoom\_humidity, TeenagerRoom\_temp,TeenagerRoom\_humidity,ParentsRoom\_temp,ParentsRoom\_humidity and WindSpeed.

516 observations are removed.

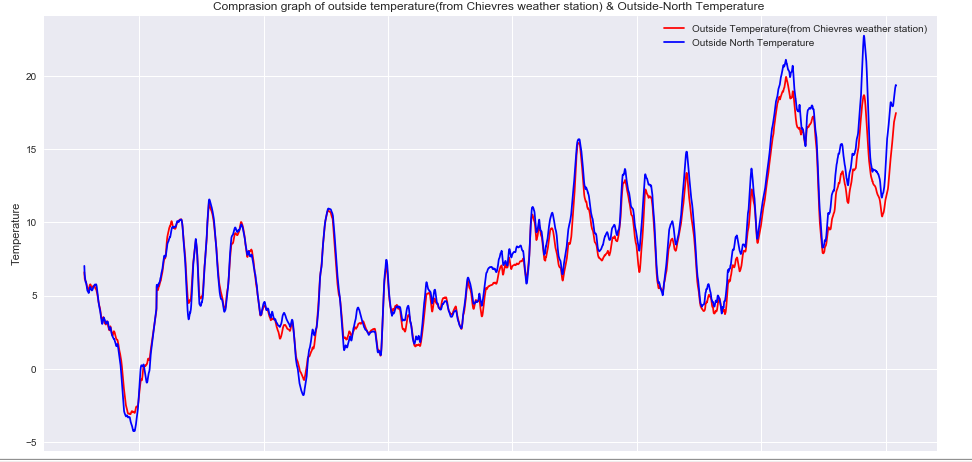
## Write the clean data into a new file for further steps

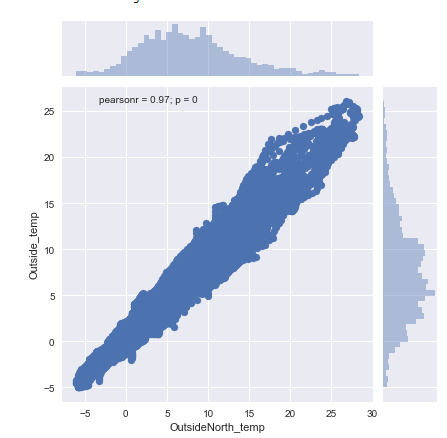
Finally I wrote the data to a new file Backorder\_clean.csv which will be used for further analysis. Cleaned dataset has 1687829 observations and 14 features.

# EDA and Data storytelling

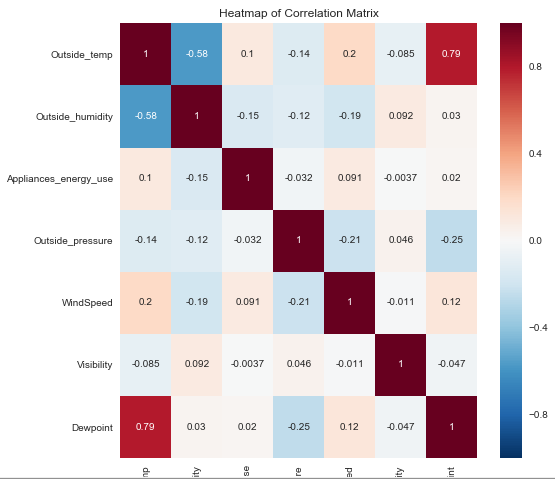
## Multicollinearity

Graph for outside temperature (from Chievres weather station) & Outside-North Temperature is almost similar. **These two attributes can be treated as co-linear.** It could be a problem for our model as multicollinearity increases the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. We will explore this point while modeling.



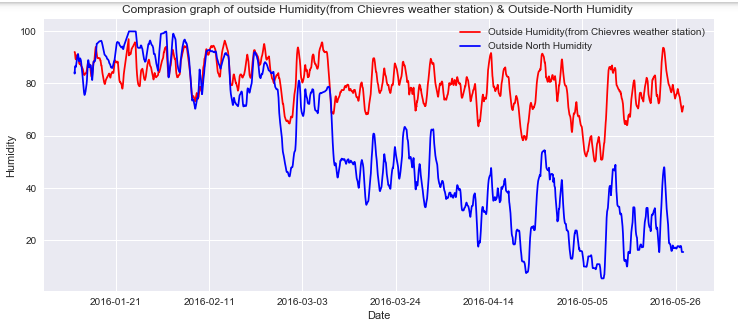


Inside temprature and humidity are corelated.

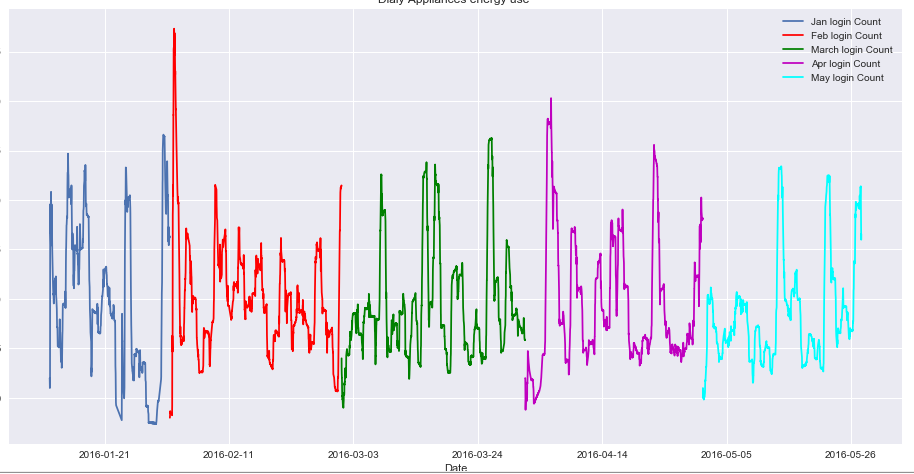


## Some data facts

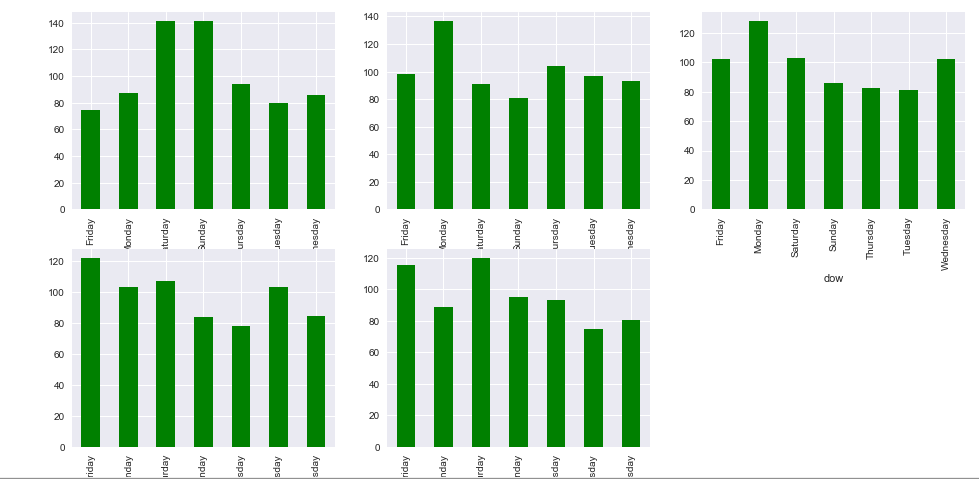
* Outside Humidity (from Chievres weather station) and Building Northside Humidity was almost same but from mid of the march Building northside humidity became significantly lower than the outside humidity. Probably AC could be the reason because days are hotter after mid-March, which will be visible in temperature graph as well.



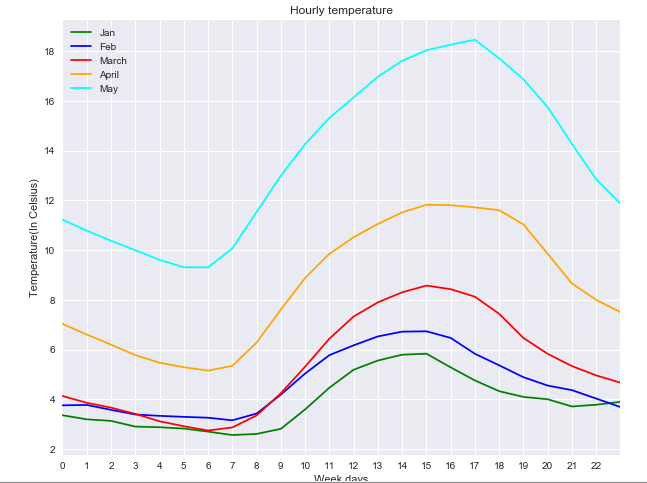
* Some days have higher Energy consumption where some days has less. This pattern needs to be investigated further. April month has the highest energy consumption. Average usage has reached up to 200 Wh in a given day.



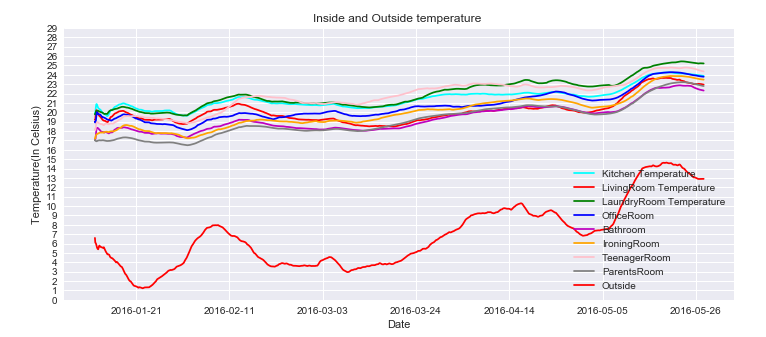
* This is interesting to note here that energy consumption is **not dependent on week days**. Each month has different days as peak and lowest usage of energy. Jan has maximum energy usage on Saturday and Sunday and lowest on Friday, Feb has highest on Monday and lowest on Saturday. Likewise, other months has different days as maximum and minimum energy consumption.



* Hourly Appliances Usage is almost similar for all the months Except Feb has significantly high usage during 10 am – 2 pm and Jan has low usage during 2-4 pm. Lowest energy usage is between 10 pm- 6 am every day. Peak energy usage is between 5 - 7 pm every day.



* Inside temperature is always maintained between 17-25 Celsius. Each room of the home don not have same temperature. Parent's room is always coolest where as laundry room is the always hottest in house.



* Inside Humidity is always maintained. This is because temperature is maintained inside the house. Each room has different humidity level. Bathroom is the most humid where as Ironing room is most dry area which is logical.

