2- Data Wrangling

After performing some EDA visualisation on the entire dataset, this section will explain the data cleaning steps taken in order to prepare the dataset for further in depth analysis. There were two datasets that were used, energy data and weather data.

Github link: https://github.com/y-fawzy/energy-demand-forecast/blob/master/notebooks/2-wrangling.ipynb

Identify irrelevant features and remove from the dataset

→ Examine relevant features /decide action on missing data/outliers

→ **Implement**

1) Energy Dataset

1.1. Identifying irrelevant features

The following features were found in the dataset under a datetime index:

i) Energy generation data from 21 different sources.
ii) Forecasted energy generation from solar and wind energy
iii) Total Load data (‘total load actual’)
iv) Forecasted Load data (‘total load forecast’)
v) Forecasted and actual energy price

The following features are needed from this dataset:

- Total Load data (‘total load actual’)
- Forecasted Load data(‘total load forecast’)

Our model will be aimed at predicting energy demand 24 hours in advance so therefore energy price and generation will not be of use in our analysis.

1.2. Examining relevant features

Examination:

Missing Data:

The ‘total load actual’ columns seems to have 36 hours of missing data. Some of these hours are consecutive hours which if deleted can wipe out half a day worth of data. The dataset has 4 years of data and hence only 4 datapoints for each hour. It would not be suitable to delete this data.

Load data is dependant on the following:

i) The hour of day
ii) The day of the week
iii) The month of the year

A good value to account for the above 3 above points would be to replace missing values with the average of 6 values that represent:

i) 3 weeks of future values and 3 weeks of past values
ii) Same day of week (eg if the outlier is a Monday
iii) On the same hour of day

Example: If the missing value is at 3pm on Monday 15th January 2015. The value would be replaced by the average of 3pm for the last 3 Monday’s and the following 3 Monday’s.

Outliers:

There were no outliers observed in this dataset.

2) Weather Dataset

2.1. Identifying irrelevant features

The following features were found in the dataset under a datetime index:

i) Maximum, minimum and average hourly temperatures in degrees Kelvin.
ii) Pressure (hPa)
iii) Humidity(%)
iv) Wind speed(m/s)
v) Wind Direction (degrees)
vi) Rain amount for the prior 1 hour and 3 hours(mm)
vii) Snow amount for the prior 3 hours
viii) Clouds covering the sky (%)
ix) Weather description for the hour (Codes)

Based on our first EDA, the following features are needed from this dataset:

- Maximum, minimum and average hourly temperature converted to degrees Celsius as it is more relatable to everyday use.
  $$T(C) = T(K) - 273.16$$
- Pressure (hPa)
- Humidity(%)
- Wind Speed(m/s)

2.2. Examining relevant features

2.2.1. *Missing Data:*

There were no missing data points observed in this dataset.

2.2.2. *Outliers:*

i. Pressure

The Pressure column has abnormally high maximum values which seem impossible. This is a data error.

- 1000 hPa is standard atmospheric pressure in the air.

- There were observed values 10,000 hPa and 1,000,000hPa.

As a reference, 10,000 hPA is equivalent to a Sedan car being supported on the palm of a human hand. Imagine 1,000,000 HectoPascal.

- The highest and lowest pressure recorded on earth is 1084hPa and 870hPa respectively.

Values above 1080 and below 870 shall be replaced with the mean of that exact date where values are less than 1080 hPa.

ii. Wind Speed

The wins seed column has an abnormally maximum value which is impossible.

- Fastest wind speed ever recorded on earth is 103 m/s and a category 5 hurricane is 70m/s.

Any wind speed above 70m/s will be replaces with the mean of the day of year over 4 years.