Energy Demand Forecasting – Milestone Report

1. Problem Statement

Why useful Questions?

For Whom?

Statement: ***Forecasting Total Weekly Energy Demand for Spain using weather features***

The more accurate this prediction is, the better it is for humanity. The stakeholders that will directly benefit from the solution to the above statement are as follows:

1. Energy companies
2. Ministry of energy.

But, it will be indirectly beneficial to the rest of the world. How you ask?

i) If energy companies can accurately forecast energy consumption in the future, they will know exactly how much they need to produce.

ii) If they produce too much, the grid can be oversupplied and fail. This is wasted energy, incurred repair costs, incurred energy which is passed on to you and me. It would be an emergency.

iii) If they produce too little, there will be major power cuts. They would have to rush and pay tons of overtime costs to get power up.

iv) If energy companies can forecast energy demand sufficiently, the distribution of energy between fossil fuels and renewable energy sources can be optimized and we will have a much greener planet.

A good forecast of total weekly energy demand will ease the above problems and that is the main aim of this project.

2. Dataset description

i) Where did the data come from?

This dataset was located on Kaggle. https://www.kaggle.com/nicholasjhana/energy-consumption-generation-prices-and-weather. The dataset has 4 years of hourly energy demand, generation by source, energy price and weather data. The energy dataset is representative of all of Spain. The weather dataset represents the 5 most populated Spanish cities. The energy data was collected from ENTSOE a public portal for Transmission Service Operator (TSO) data. Weather data was purchased by the creator of the dataset and made public on Kaggle by himself.

ii) How was the energy dataset wrangled?

The following features were found in the energy 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 were 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 1 week in advance so therefore energy price and generation will not be of use in our analysis. But prior to removing the other features, the generation data was examined and findings will be reported in the EDA section.

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.

iii) How was the weather dataset wrangled?

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)

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.

3. EDA

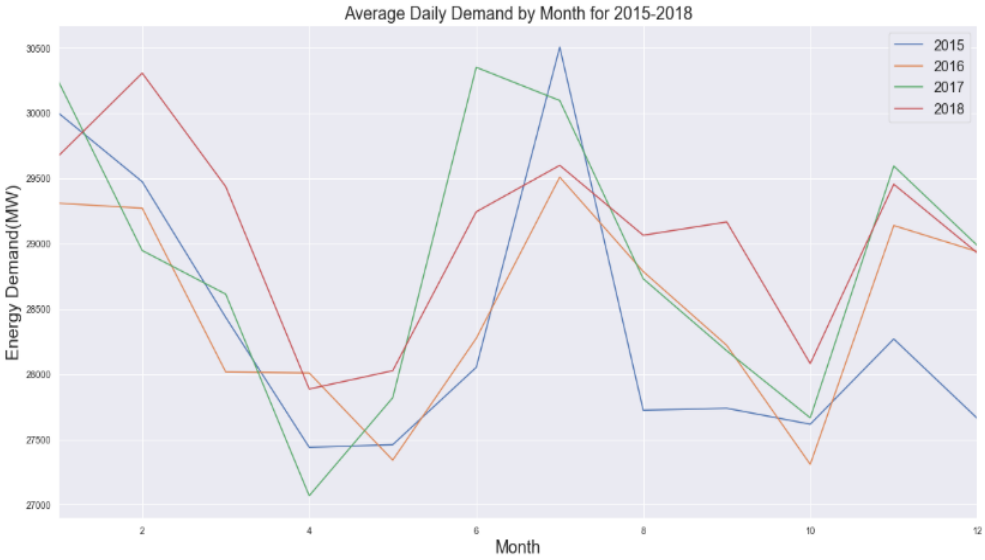
This section was structured by asking a series of questions and answering them with visuals. This section aims on answering questions related to demand and weather features only.

3.1. Energy Demand

1. How did annual energy demand Change between 2015 and 2018?

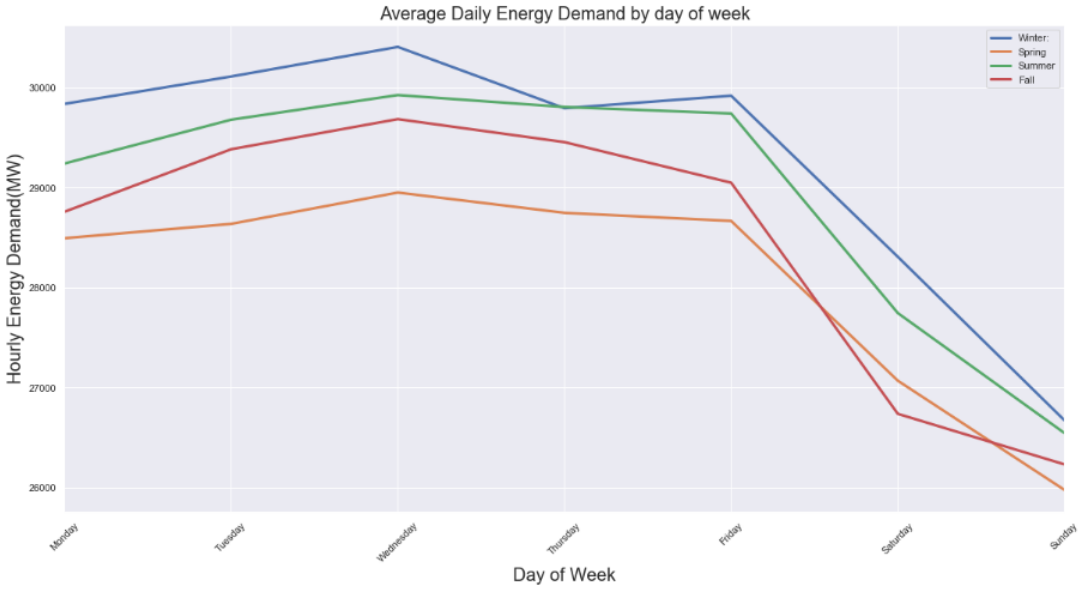
It was observed that energy demand is steadily increasing at a decreasing rate between the periods of 2015 & 2018. The rate of increase has decreased from 1.1% in 2015 to 0.7% in 2018.

2. How does energy demand vary by month?



We see a similar pattern for every year with a cycle of 2 highs and 2 lows. Highs seem to occur during January(Winter peak) and June/July (Summer peak) with lows during April and October(Spring and Fall).

3. How does energy demand change across the week?¶

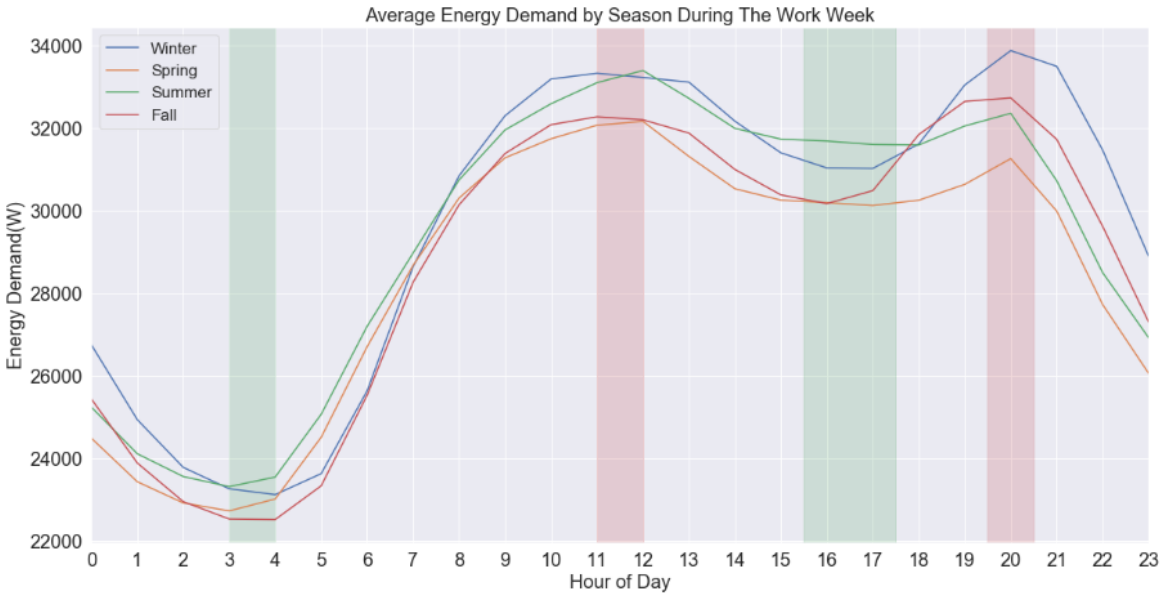


We see a similar pattern during the 4 seasons.

Energy demand peaks Wednesdays, decreasing all the way to Sunday with a significant decrease on Saturday (Weekend). Then a sudden increase on Monday(start of work week).

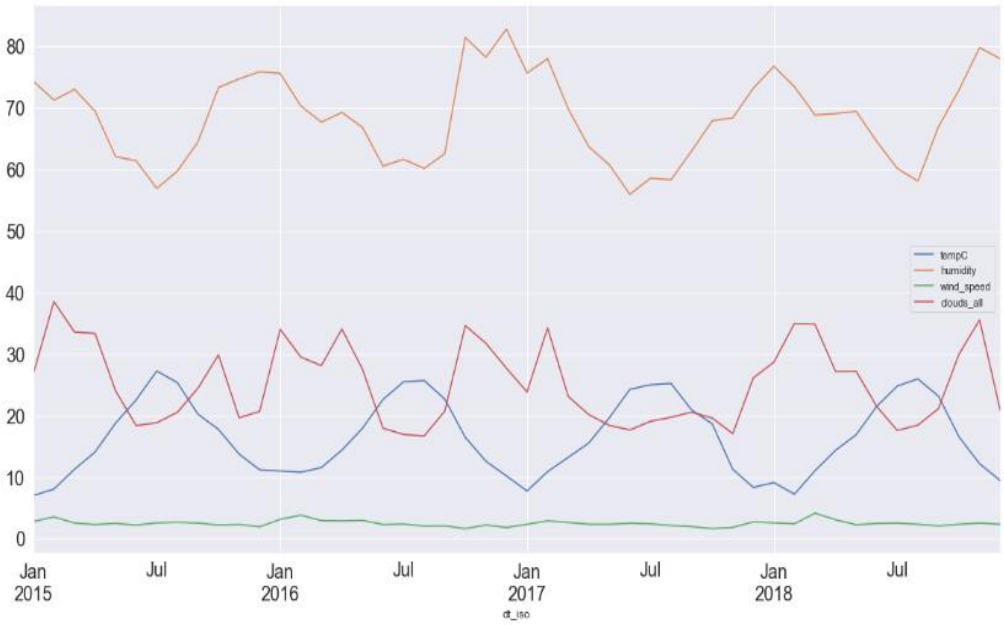
We do see a sharp decrease during winter on Thursdays and increasing to Friday but it unknown why this occurs.

4. How does intraday hourly demand change?



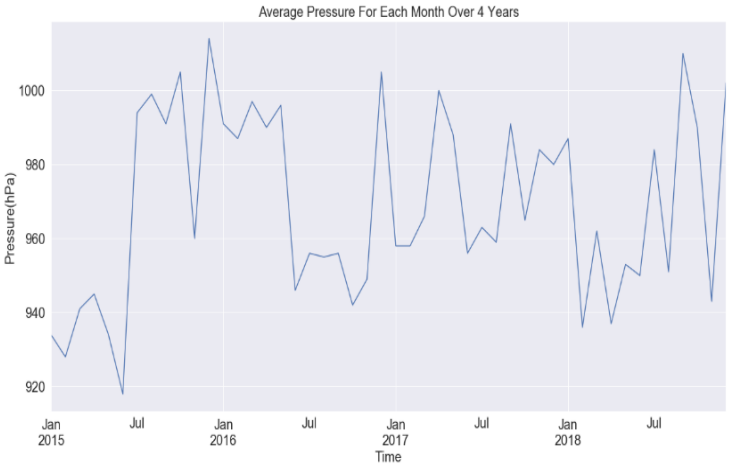
3.2. Weather Features

1. How do weather features change across the months?

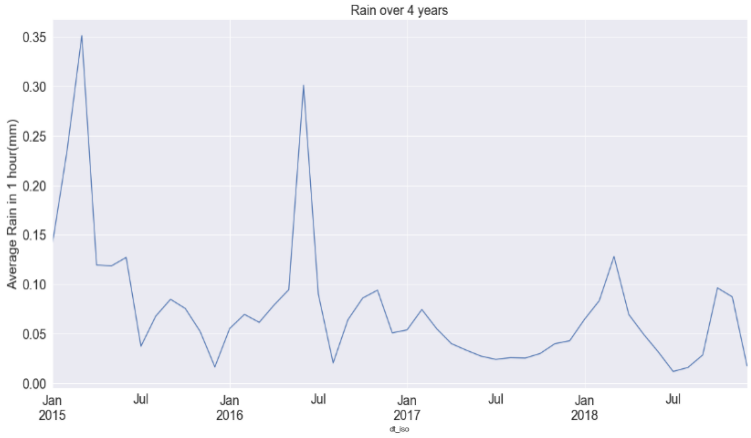


This plot shows how aggregated monthly weather features are moving throughout 2015-2018. Rain, wind speed and pressure cannot be analyzed due to their values having different orders of magnitude.

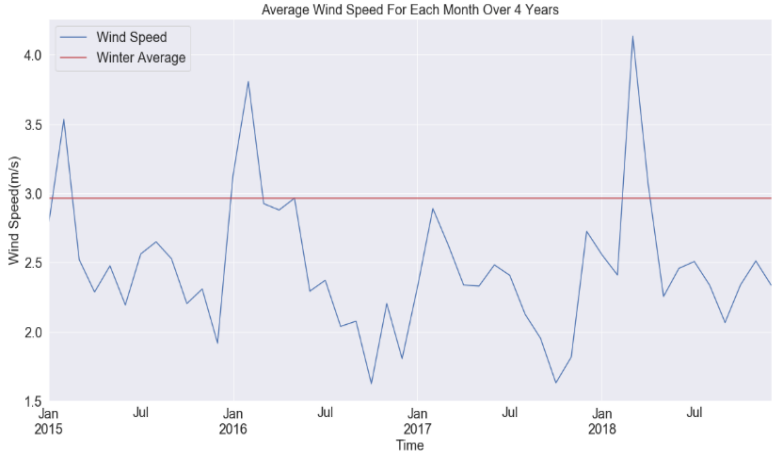
We can see a clear inverse relationship between temperature and humidity as well as temperature and cloud formation.



There is no strong pattern for pressure across the year.

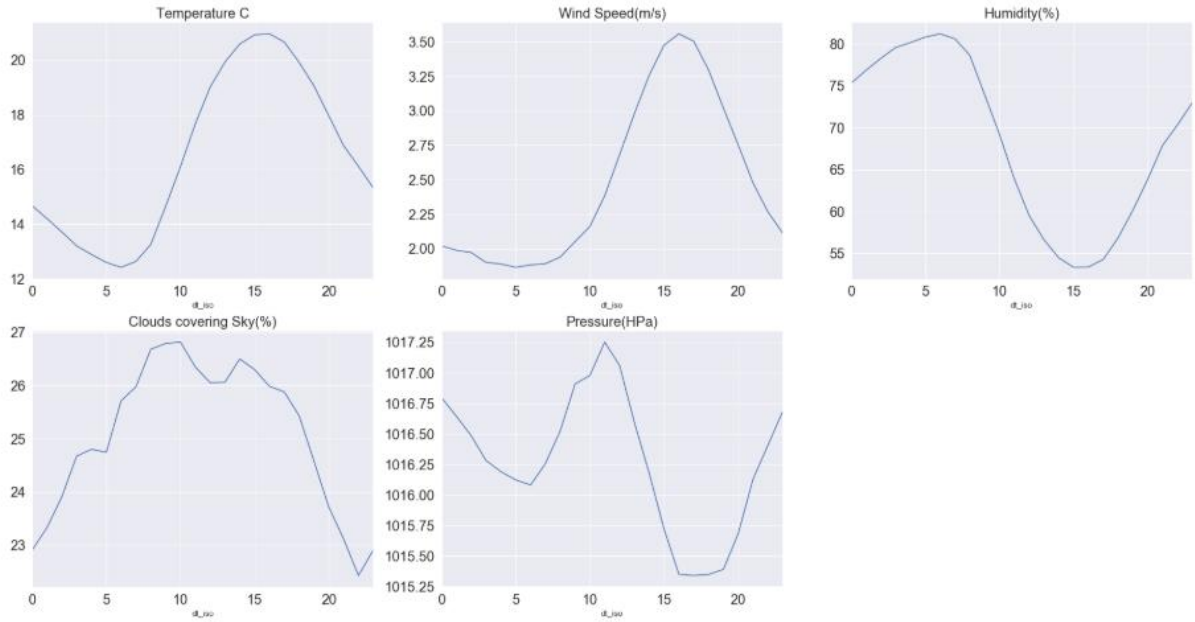


We see a downward sloping trend line for rain across the years. Rain usually peaks in winter but in 2016, there seemed to be a peak in rain during spring which was quite interesting.



We see a consistently repeating wind speed peak in winter. The red line is the winter average wind speed. We have a had a higher than average peaks for all years except for 2017.

2. How do weather features change across the day?



The above graph summarizes how weather features change throughout the day.

Temperature: Increases in the morning (With the sunrise), peaks at around 3 pm, then decreasing while the sunsets through the night reaching a low at around 5am.

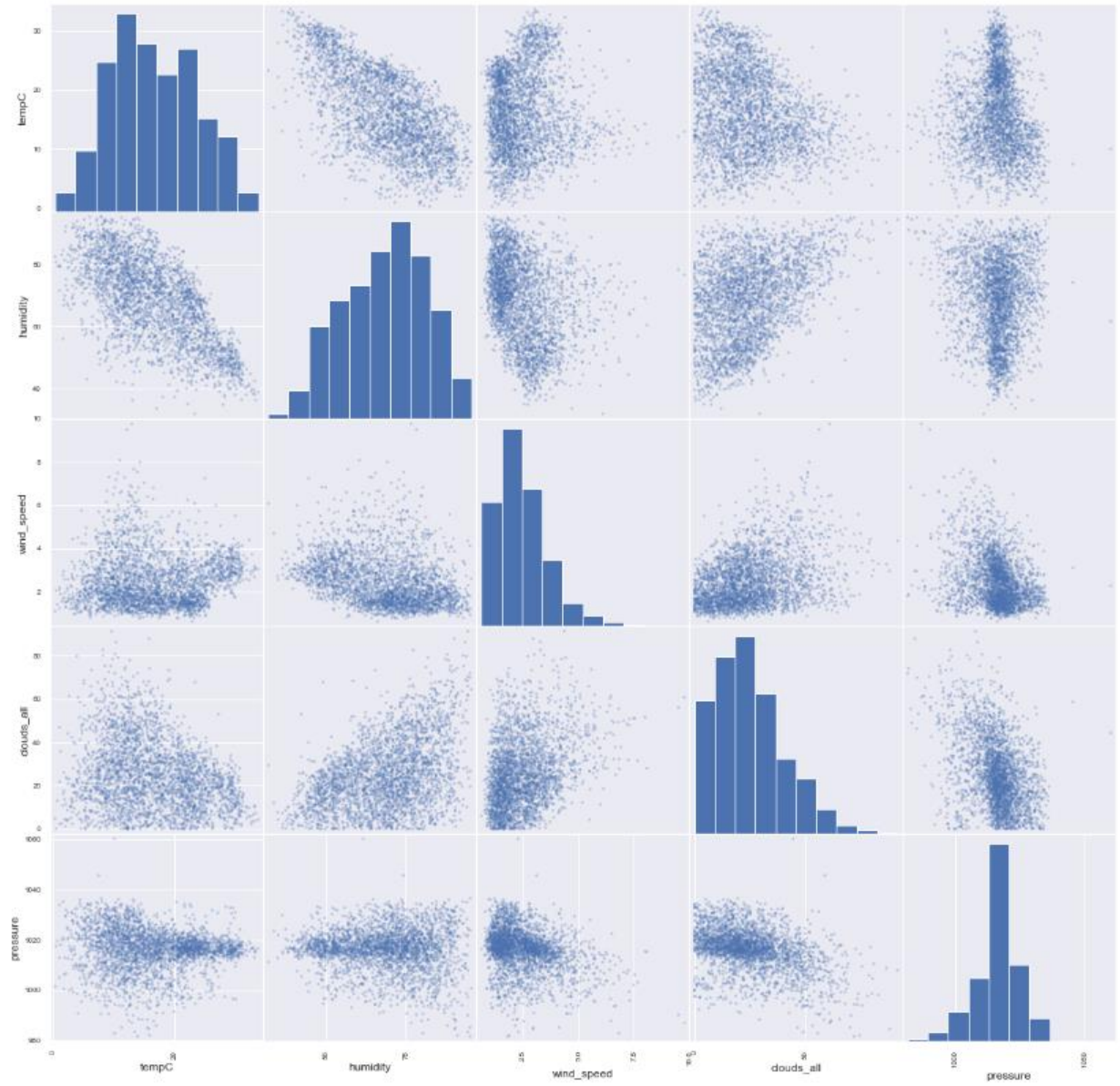
Wind Speed: Pattern is like temperature.

Humidity: Opposes temperature and wind speed.

Cloud Formation: High at 10am, plateauing until 3 pm, decreasing until 9pm then gradually increasing until 10am.

Pressure: 1st High at 11am, 1st low at 5pm, 2nd high at 12am, 2nd low at 6am.

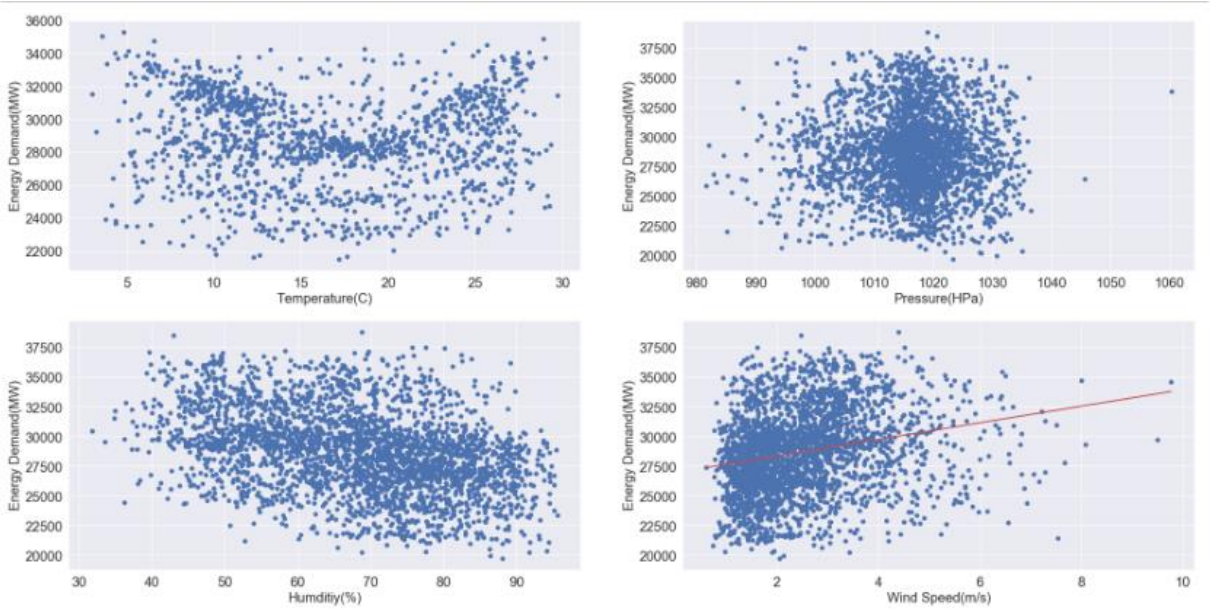
3. How do weather features correlate to each other?



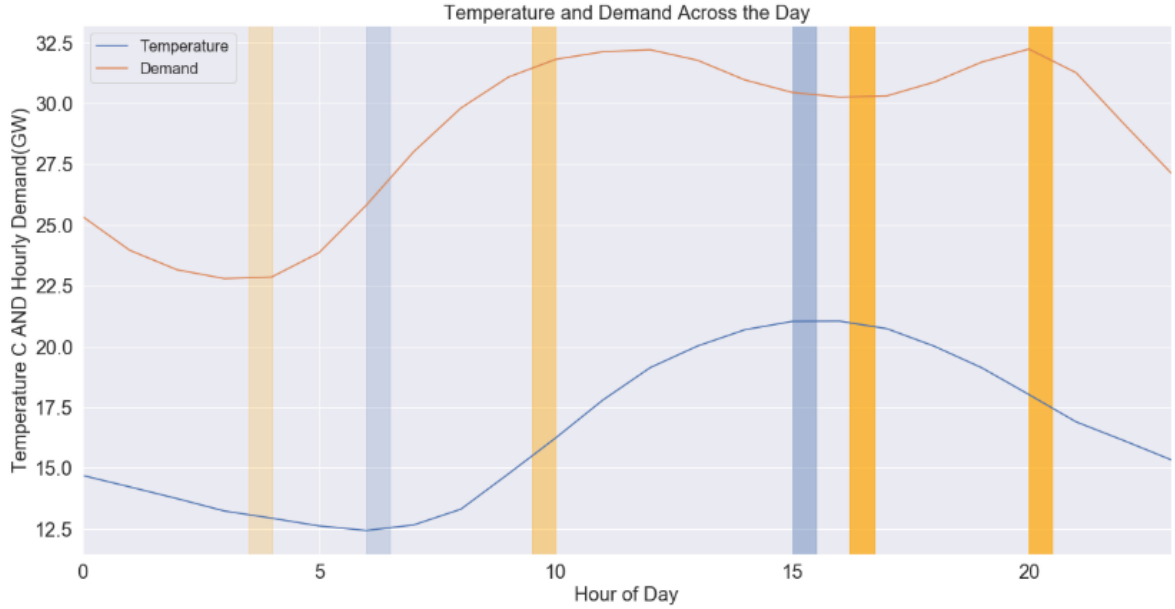
The above graph summarizes how weather features are correlated to each other, this will be better visualized using a heatmap later on. The following is observed:

1. Negative relationship between Temperature and Humidity.
2. Negative relationship between Temperature and Cloud formation
3. Positive relationship between Humidity and Cloud formation
4. Negative relationship between Wind Speed and Humidity
5. Positive relationship between Wind Speed and Cloud formation
6. Negative relationship between Pressure and Cloud formation.
7. Negative relationship between Pressure and Wind Speed.

3.3. Weather Features in relation to Energy Demand

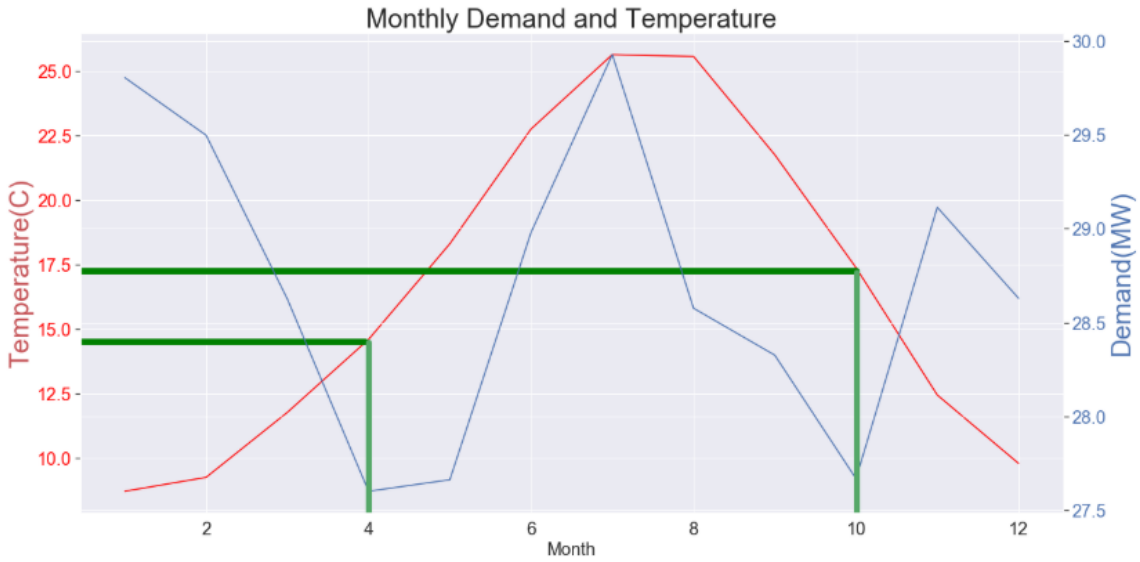


1. Temperature: A parabolic relationship is evident with demand. Demand peaks at around 5C and 30C temperatures but the lows occurs at around 15-20C with almost no change in demand.
2. Pressure: Demand &pressure do not seem to be correlated.
3. Humidity: Demand & Humidity are parabolically correlated which is expected as Temperature & Humidity are negatively related.
4. Wind Speed: A highly variant positive correlation is evident with energy demand.



- Temperature hits a low at 6am -- Demand hits a low at 4am.
- Temperature hits a high at 3pm -- Demand hits a high at 10am.
- Temperature decreasing until 6am -- Demand hits a low at 4:30pm.
- Temperature decreasing until 6am -- Demand hits a high at 8pm.

This analysis suggests that intraday demand movements are not fully predicted by temperature changes. Let's rewind and have a deeper look at monthly demand vs average monthly temperature...



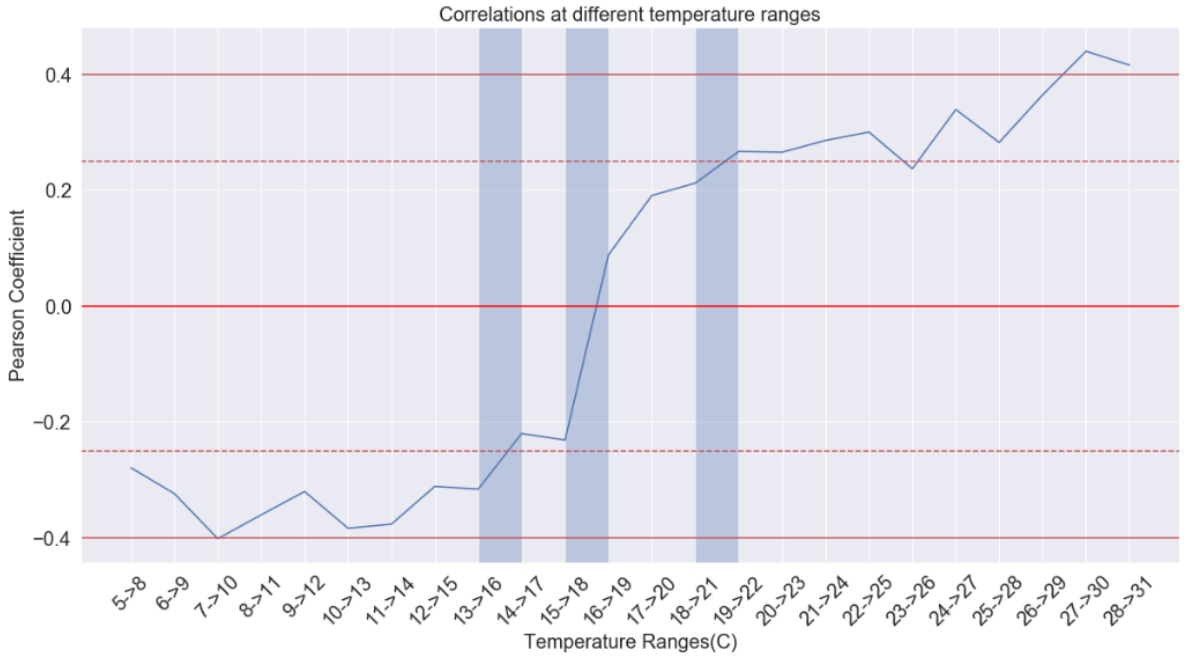
This is just a confirmation of the hypothesis that temperature and demand are related on a larger time scale.

During April, we move from winter to summer, the average monthly temperature at which a demand low is reach is at around 14.5 C As we move from summer to winter, the average monthly temperature at which a demand low is reach is at around 17.5 C.

4. EDA (Statistical)

1. Correlations between Temperature and Demand

A detailed analysis of the correlation between temperature and demand was conducted as a follow up from the previous section. This was done using the Pearson r coefficient and examining how it changed for 3 degrees Celsius temperature changes a shown below.



As can be seen the relationship is not as strong as is expected with maximum magnitudes at 0.4. But the following can be observed:

- Correlation increases above -0.25 from the lows(-0.4) between 13-17C.
- Correlation crosses 0 occurs between 15-19C which means the bottom of demand occurs at this temperature.
- Correlation increase above 0.25 at between 18->22C.
- Correlation is flat at temperatures below 15C at -0.4 and at temperatures above 26C.

Given a larger dataset, the correlation magnitudes are expected to be much stronger than 0.4.

2. Examining the means and standard deviations of the different seasons

An important test to conduct that will help in modelling is whether the high and low seasons can be define by the same mean distribution and standard deviation distribution. If this is the case then when it is time to model, we can define a seasonal cycle as 26 weeks which will ease prediction. We started with the mean distribution of winter and summer.

Given that summer and winter are the high demand seasons of the year due to heating and cooling, do they have the same mean distribution? Our null hypothesis is the following:

"Winter and Summer have the same mean distribution of hourly demand"

The following steps were conducted for this test:

1- Combining the winter and summer demand datasets into one big dataset.
2- Computing the combined mean
3- Shifting the mean of the individual winter and summer dataset to have the same value as the combined mean. This was done by transforming the winter and summer dataset using the following equation:

$$Transformed\ Indivudal\ Season\ Dataset\ (\boldsymbol{Array}) = Observed\ Individual\ Season\ Demand(\boldsymbol{Array}) - Observed\ Individual\ Season\ Demand\ Mean(\boldsymbol{Integer}) + Combined\ Mean(\boldsymbol{Integer})$$

4- Bootstrap 10,000 mean replicates for each **transformed** season by bootstrapping samples.
5- Find the difference of the 10,000 bootstrapped mean replicate datasets.
6- Find the observed difference of the actual means.
7- Test the observed differences against the observed difference by counting the number of bootstrapped replicates that are equal to or greater than the observed difference.
8- Divide the count by the length of the array to obtain the p-value.
9- For a 95% confidence interval, if the p-value is less than 0.05, we shall reject the null hypothesis.

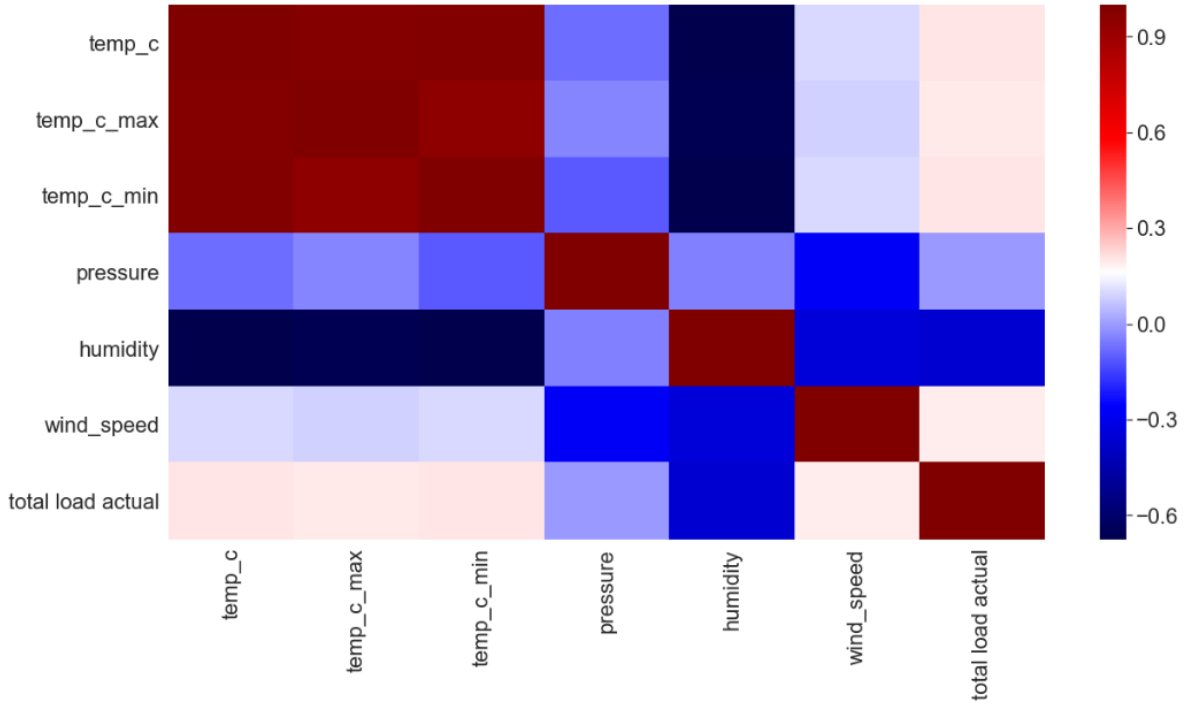
The p-value was observed to be less than 0, meaning that winter and summer do not have the same mean distribution.
This same test was conducted for Spring and Fall, also achieving a p-value less than 0.05.
Hence we can conclude that that a seasonal cycle is in fact 52 weeks.
For the sake of curiosity, a similar test was done by comparing the standard deviations of the different seasons. The transformation was done using the following equation:

$$Transformed\ Indivudal\ Season\ Dataset\ (\boldsymbol{Array}) = \frac{Combined\ Standard\ Deviation(\boldsymbol{Integer})}{Observed\ Individual\ Standard\ deviation(\boldsymbol{Integer})} \times (Observed\ Season\ Dataset(\boldsymbol{Array}) - Observed\ Season\ Mean(\boldsymbol{Integer})) + Observed\ Individual\ Season\ Mean$$

In conclusion the standard deviation and mean distribution of all seasons are different.

3. *Examining the correlations of demand and weather features*

Examination was best visualized using a correlation matrix transformed into a heatmap from Seaborn. The strongest correlation observed are between temperature and humidity, which is a negative correlation.



Demand and temperature are not linearly correlated but there exists a parabolic relationship as seen in the previous section. Demand looks slightly negatively correlated with humidity and completely uncorrelated with pressure.