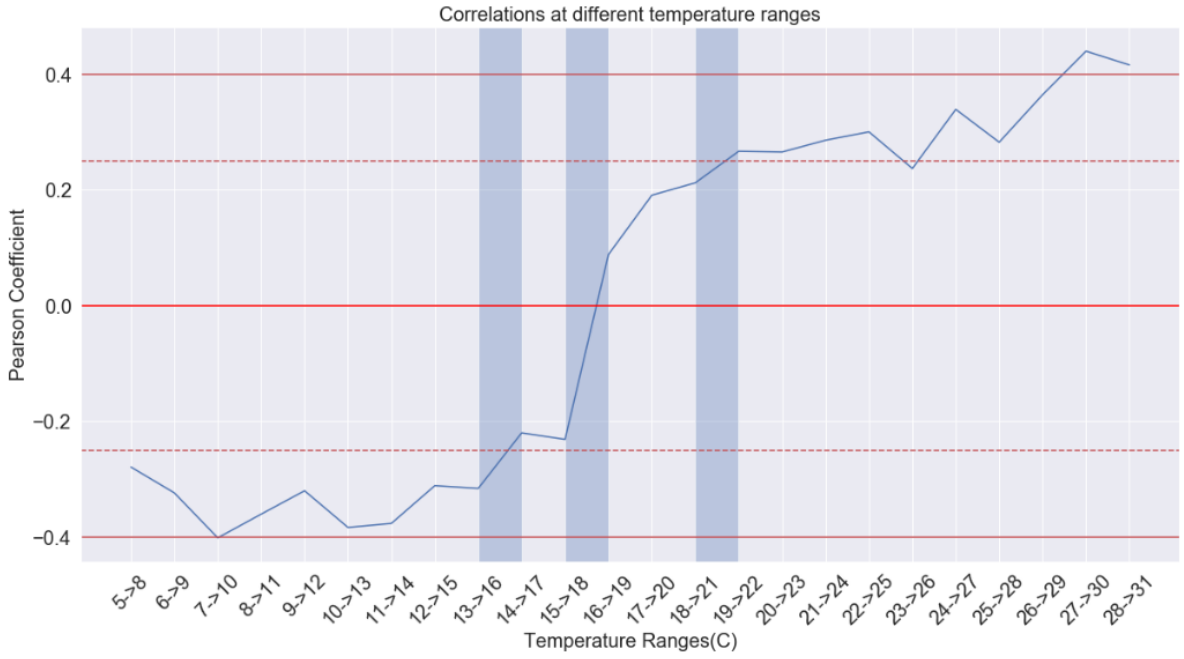
3- EDA (Statistical Analysis)

After performing some EDA visualisation on the entire dataset and cleaning the data, this section will further analyze the datasets using some statistical analysis.

Github link: https://github.com/y-fawzy/energy-demand-forecast/blob/master/notebooks/3-eda(statistical).ipynb

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 is 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:

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

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.

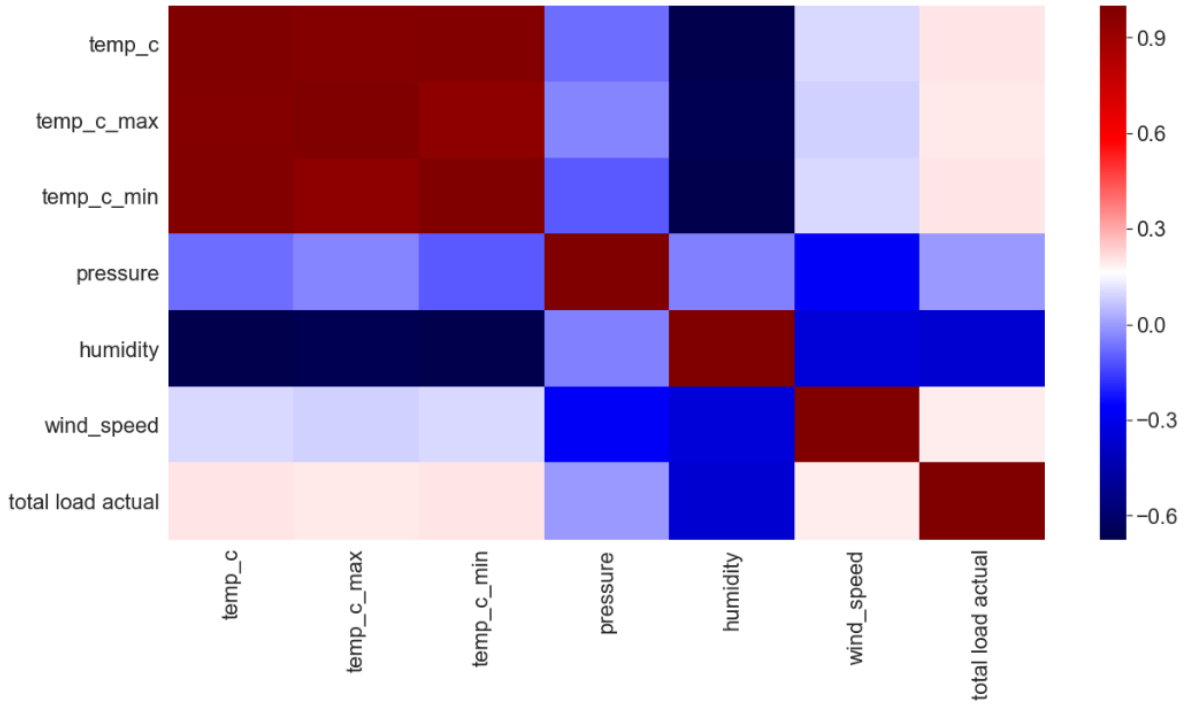
The p-value was observed to be less than 0, meaning that winter and summer do not have the same mean distribution.
This same test was conducted for Spring and Fall, also achieving a p-value less than 0.05.
Hence we can conclude that that a seasonal cycle is in fact 52 weeks.
For the sake of curiosity,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

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

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.