Notes for the challenge

Saul Rathbone-Boschis

July 2022

1 Process

1.1 Working with the csvs

To complete the parts of the task which pertained to the csv files, i made a couple of functions which would open the csv's and define lists to store them, and visa versa.

1.2 Exploring the data

I plotted the daily values for a couple of typical days, and noticed some periodic trend which looks like each daily cycle could be approximated with a 5th or 6th degree polynomial.
I then plotted the daily average for a couple of typical months, and noticed no trend, so i checked the qq-norm plots for the daily averages for all years for any given month, and they appear to be normally distributed.
I then plotted the daily average across the entire data set, and noticed a periodic function.
I then plotted the monthly average, and overlaid them for each year. There does not look to be a trend between the averages relative to the years.

1.3 Building the model

I initially wanted to tackle the problem by:
1) determine the parameters for the normal distribution of daily values for each month.
2) determine the function that best fits the daily power demands for any given day.
3) determine the function that best fits the mean of the monthly average over the course of a year.

However, when I made my first model to predict the next year, and when looking at the output, i notice it looks significantly different to the training data.

I think my assumption that the monthly distributions were normal is incorrect, and also grouping the days into months is a bad assumption.

So instead i used a Gaussian kernel density estimator to sample the daily means.

1.4 Displaying the data

I decided to make 3 plots to display the data. The plots show the change in monthly, and yearly averages over time, as well as an overlay of the models daily average density distribution for each day, and the daily average data from 2020.

2 Observations

There is no trend in yearly average for the years 2015-2019, though in 2020 there is a definite increase.
The daily average values appear to be sampled from a random distribution, though if you zero the half hourly values they match up very well between each day across each year.
There is a trend in monthly average across each year, though there does appear to be many outliers.

3 Suggestions for the future and improvements

At no stage does my code deal with outliers, the model could be made to be more representative of reality if i add some outlier detection and removal in there.
I don't do any cross validation on my entire model (there is some for specific parts, but only to tune hyper parameters), and every change i made to the model was by just eyeballing if it looked right. That would be something that could change.
If i had more time, i could also use the models daily distributions to come up with a probability that the new data set is generated from the same random variable that the historic data sets were generated by.