Recruitment Challenge 2020: Accompanying Notes

# Approach Intuition

The intuition behind my approach was to create a simple and logical process to illustrate the trends in electricity demand. My first thought was to examine the data to determine the most appropriate methodology to clean and organise the data so that it would be comparable. Once the data was in similar formats, I would then take simple averages so that the daily data and the historical data could be clearly compared on a ggplot. From there, any improvements with regard to efficiency, usability and cleanliness of the code would be done.

# Rationale

Firstly, my rationale with regards to the initial data exploration was that it is much easier to ensure you are importing correctly once you have seen the data. This was also to investigate any issues that I may encounter, such as differences in time periods and format of the data.

Moreover, my rationale regarding cleaning the data was to morphe the daily data into the same format as the historical for two reasons:

1. The daily data had many unnecessary variables which made it logical to dispose of.
2. The historical data is a larger dataset and if this were comparing a large amount of years/states/etc. it would increase processing times.

The rationale for taking simple averages was an initial approach to see whether the trend would be appropriate. If a simple average did not generate an appropriate trend, then alternative methodologies would be explored.

Finally, visualising the data on a scatter plot was the most logical initial decision. However, a line graph could have had a similar effect. Being able to directly compare the same time intervals, makes the trends immediately apparent.

# Observations

The graph generated (see below) shows the average electricity demand for households for the historical period (financial years 2015-2019) compared to the COVID-19 period (22 Mar to 21 May). The COVID-19 period was chosen due to the availability of data only going back to the 22nd of March. The graph shows that at the maximum peak, the demand is slightly lower for the recent data than the historical average. Furthermore, during the troughs in the very early morning and the middle of the day, the recent data is also lower than the historical data. However, during the morning peak and in transitions from trough to peak, the level of demand is quite similar. This shows that the COVID-19 period has had an impact that during the major peak and the troughs demand is lower, leading to a lower amount of demand overall. This could be explained by increased solar power adoption.

A screenshot of a cell phone

Description automatically generated

# Questions

Q: Why does the historical 2019 csv dataset contain data up to and including 30 Jun 2019?

A: These datasets were extracted as forecasts and are not strictly actuals. Hence, the data for 2019 contains future values.

Q: What methodology should be taken with regard to the historical averages? Should these be seasonal? Should they be arithmetic or geometric? What granularity is appropriate?

A: It is best to initially take the approach of arithmetic averages based on the half hour intervals across all years. This may be skewed by various trends, such as increased

Solar usage in recent times.

# Reflection on Approach

Initially, the code that was written was not the most efficient approach. In hindsight, utilising apply functions straight away would have been more beneficial rather than adjusting the code later. Furthermore, the first version of the code used data frames which were later modified to data.tables. Taking these more efficient approaches from the outset would have saved time adjusting the code later.

In addition to this, it would have been more advantageous to outline and understand the deliverables more clearly from the start as the code was significantly adjusted once this became clearer. Many were able to be delivered using similar functions and processes.

Overall, I believe the approach I took was quite logical but the intricacies of the code could have been improved by taking the more sophisticated approached from the beginning. Further practice in apply functions and data.tables will assist in improving this.

# Further Improvements

The analysis could be improved in a number of ways. Firstly, the averages taken were simple averages and do not account for the seasonal effects of electricity demand. For example, increased demand on hot days due to air conditioning. A seasonal average could have been taken to allow for this effect and further improve the visualisation. Moreover, a geometric average could have been taken to place a higher weighting on more recent observations, whereby the impact of solar power is already considered.

There could also have been more in-depth analysis into the trends and statistical properties of the data. Analysing error distributions and creating ranges for projections are just a couple of examples. Therefore, whilst the model is simplistic in nature, its accuracy and precision can hardly be relied upon in practice.

Finally, additional data would further improve the model. Having data on solar power registrations and usage, weather patterns and events would further improve the model. In addition, since only the last 60 days is kept as daily data, storing a historical database of daily data would prove beneficial.

# Additional resources utilised

<https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html>

[https://ademos.people.uic.edu/Chapter4.html](https://slack-redir.net/link?url=https%3A%2F%2Fademos.people.uic.edu%2FChapter4.html)