# Question 6

In general, R is a very robust and dynamic tool. It can be used to execute practically all operations related to data analysis and data science. It is only a matter of getting used to its syntax and finding the right packages and functions. There is extensive documentation also available. Sometimes, it is hard to find very specific examples, but mostly, tweaking many of the online standard examples will help.

In comparison to Excel, obviously filtering is tiny bit harder in R. In Excel it is just a matter of checking the right boxes, whereas, coding is required in R. However, filtering columns along row values is much easier as was used to solve Question 5. In addition, data transformation and manipulation such as spreading, and gathering is much easier in R. It can be done in Excel too with a pivot table. But, then applying GETPIVOTDATA() functions in Excel is not very easy. Also, R is much better with vectorized functions, which will be more cumbersome in Excel, as you can just display the data in R, whereas in Excel, it is necessary to store that data as well, which takes up space. Question 5 is almost impossible to do in R. This is for two reasons: firstly, when the data was joined, there were 94,000 rows approximately which would have slowed down Excel; secondly, it must be very hard, if not impossible, to execute an inner join in Excel.

In comparison to Tableau, I will say R is efficient and simple. Tableau is majorly a data visualization tool. It is not good for other purposes. Also, Tableau requires a very standardized dataset whereas R can work much more robustly. This could be seen while doing the forecasting question in Tableau. Although, it may be much simpler to forecast in Tableau, but still overlaying actual data with forecasted data was difficult. R allows more flexibility in manipulating data, but that is why it requires more work as well. Each chart graphics label are usually incorrect and need to be changed manually. Adding an element or color to any chart is not a simple click or drag process but requires coding. This being said, it still has got as good visualization as Tableau, if one is ready to make the effort, but it requires experienced coding, in general.

Summarizing, having worked with R for the first time, I appreciate its robustness and the range of features it provides. It is understandable why it is considered one of the top software to learn for data analytics.

# Question 7

Most of the specific edits that I would like to suggest are more related to syntax rather than changing the question itself.

1. Question 2b made me feel that I need to calculate different percentiles for each year for each plant and then, if the plant satisfied all the criteria, then they were to be selected. The word “every year” made it confusing. It should just have average heat input over all the years.
2. Question 3c asks to name the top 5 EPA regions. I assumed the ‘top’ criterion was gross load but I had to ask.
3. Question 5 asks to design a metric. This was my fault, but I immediately referred to metrics as the heat inputs, gross loads etc. It is my first data course, so I might be naïve but after asking around, many people had the same confusion. It would be good to clarify that the metric needs to be designed based “only” on distance between plants, and number of plants. It was hard to connect the existing metrics with only distance and number.
4. Apart from that, most of the questions were basic filtering and grouping, so there was not much confusion. Although, I would really like to do some sort of forecasting or clustering in the next HW, just to understand how different it is compared to Tableau. All in all, I think it was a great introductory R homework!