Project Summary

In this project, we started by registering on *openweathmap.org* and getting an API key. With the key, we passed our API call for weather data, based on name of location, through the python *requests* library. We pulled the data from the website. The data came in JSON format and needed to be formatted neatly.

We picked out the specific information we wanted considering less than a hundred states both in the US and out. The data were organized in form of different lists; and subsequently collected into a Pandas data frame. We named the headers and formatted the data frame nicely.

By our code, missing values were automatically updated with NA while picking the data from the JSON-format into the data frame. Afterwards, we intended to clean and organize the results more neatly. One challenge encountered here was in updating the missing values with the respective column mean values; we had to change the datatype to float as JSON returned numbers as strings. We discovered that there were no duplicates in the data frame after performing the operation of dropping duplicates. We checked the shape of the new data frame and it was the same as before dropped duplicates.

In data transformation, we dropped the columns “Base” and “Visibility”. This is because they respectively contained “not very useful” data and many missing values (too many to make significant improvement to the dataset or any deductions from it). For “Base”, it was the same data all through and for “Visibility”, there were many NAs. At the end we had a dataset that was clean, easy to read, and ready for analysis.

Overall, I think a data scientist will have to do such as process often because it helps in the data science process to make it more advanced with assistance of an API. This helps a data scientist with analytics by gathering data from sources and impacts the whole analytics lifecycle. Therefore, I think it has a significant impact on data scientists.