When we initially started this project, we wanted to choose a topic that was very versatile and easy to find data for. Considering weather data is collected every day, and has been for years, an abundant amount of data was available and choosing which datasets to use was fairly easy. The first issue we ran into was navigating the API call. Although we had access to the API, figuring out how to access the data within, and in the form it is presented in took a little bit of trial and error. Once we were able to see the format and content of the API we could start figuring out how to merge.

A lot of the challenges we ran into involved merging the two datasets together. As they were pulled from two different sources, we had to initially fix the formatting to only pull data from specific time ranges and states. The API focuses on pulling current data on weather averages across all 50 states, while the CSV file has data for different stations across all 50 states from 2016-2017. For a more realistic comparison, we chose to extract data from the CSV from only March 2016 to compare how weather now compares to weather approximately 10 years ago. We also had to group all the stations in each state together to get state-wide averages and only have one row/value per state. We already had our code for the API set up to pull different weather average data for each state at the current time (March 2025). When merging the data, we were getting a lot of NAs and duplicates in the data so we had to go back and revisit the data cleaning process. Upon further analysis of the dataset, we found that some states were listed by their abbreviation ("DE") instead of their state name ("Delaware"). Because this only occurred for a select number of instances, we just removed those values from the data. We also opted to remove any territories and places that were not states from the data (Puerto Rico, Washington D.C., etc.) for simplicity purposes. In the data pulled from the API, we found that we were missing values for specific states so that we couldn't compare weather now to weather in March 2016. We also opted to remove those states from the CSV data, in order to keep only states that appeared in both sources. After this data cleaning, we were able to merge our data by state and indicate which columns came from the CSV file and which came from the API data.

Aspects of the project that were easier than expected were data analysis and data summarization, though this could have just been because we all have strong statistical backgrounds. It was interesting to see which weather reads were strongly correlated from 2016 to now and which weren't. The data analysis also showed a lot of important trends, specifically that temperature between the two time periods are strongly correlated, while other factors like precipitation and wind speed are not. The aggregation analysis also showed that, on average, temperatures have been higher this March than they were in 2016. There doesn't seem to be much of a difference in precipitation, but wind speed also seems to be higher now than it was in 2016.

A utility like this could be useful for future data projects because we have set this project in a way where we can calculate data trends automatically for all months of the year. We can achieve this just by changing one number which indicates which month we are in. This can be used to expand data projects to analyze trends across a whole year. Additionally, the merged data set has a lot of additional data that we can use as indicators to analyze global warming.