Midterm Project DS 2002

Haley Mitchell, Natalie Siewick, Alexa Lathom

In this project, we explored two different datasets containing data about carbon emissions and COVID-19 statistics for provinces in Canada. We used an API call to get COVID-19 data, population numbers, and GDP for each of the provinces in 2022. From there, we discovered a CSV file that held information about the carbon emissions per million metric tons. Finally, to get our full data frame, we combined the data by province.

A challenge that we encountered was fully understanding the instructions in order to perform each piece of the project accurately and meet the grading criteria. Our group had a lot of difficulty understanding how to properly create an ETL pipeline and how to import the data into the SQL database. This meant that we had to exchange many emails with the professor to clarify the instructions. Another challenge we had was including errors in our code. Since we have not specifically gone over errors within this class, we conducted a lot of research to look into the types of errors and when to use each of them properly.

One aspect that was easier than expected was the EDA portion. Although we didn't learn how to create many graphs in class, the internet holds many useful resources for learning how to code them. Our formulated ideas and sketches on how to visualize our data were easily brought to life through research. Communicating within the group was also very easy for us. We met multiple times to discuss our progress and thoughts on how to move forward. It was also very easy to split up our roles and rely on each other to do their part.

Finding viable data to use in the project was one thing that was harder than we expected. When looking through an API, we had to weigh many aspects including how many times we could call it and how easy it was for us to get free access to information. Not only did we abandon many ideas because the API was not up to these standards, but it also took time to try them out because the syntax for outputting data was different for each. Finding a CSV file to merge all of the data was also a problem. Either we couldn't find a common identifier between the two datasets or it would subset the data too much that we couldn't make judgments that would be representative of our topic. For these reasons, finding data was, unexpectedly, the hardest part of this project.

The things that we learned in this project will be very useful in our academic endeavors and professional careers. Many companies use structured query language to store their data, so learning to both write in SQL and navigate creating SQL databases is extremely helpful. Also, although we have all heard the term "API" before, we have never taken the initiative to learn more about it and the potential they have to collect a large amount of data easily. We are very fortunate to learn about APIs and how to make API calls, so that we can apply this knowledge in the future.