This data project was a great opportunity to combine a lot of the skills we have been learning this semester, such as using pandas to clean data, processing CSV and JSON files, and working with SQL and databases. Initially, my partner and I were intimidated by the thought of getting user input and making the project generalized. We thought that it would be really complicated to prepare for all of the different user inputs that might be entered, and were worried about how much we would have to interface with a user. We also thought that it would be hard to make general code that could take an unspecified input and convert it to an unspecified output; we did not understand that we would be able to mount the files in Colab, and thought that we would have to figure out how to make our code take in a file and automatically understand what type of file it was, as opposed to being able to have the user type in what kind of file it was.

Ultimately, the user input aspect of the project was not nearly as difficult as we imagined it would be. It made it easier to operate the code because there was less of a risk of messing up the syntax when the user does not have to call a function or alter anything. We additionally found that testing the code was easier with an input structure, since it was very simple to re-run the code and enter certain things, such as invalid file names.

One thing that was harder than anticipated was figuring out how to structure the project. JP and I found that there were multiple different ways to fulfill the requirements and struggled to decide which strategy would be most efficient, user-friendly, and the least complicated. We also wanted to operate mostly using coding techniques that we learned within this course, since we were most comfortable with them, and a lot of resources we found on Geeks for Geeks or Stack Exchange used methods that we were not familiar with at all. For instance, we did not know whether it would be better to convert from file to file or (such as converting directly from CSV to

JSON) or to instead get the new file by converting from data frame to file. We chose the latter, but faced decisions like that throughout the project that required careful consideration to answer.

I think this project could be extremely useful for other data projects in the future. I am glad to have practiced the techniques from class (converting from one file type to another, data cleaning and processing using pandas, etc.) with real data, and am now confident I could use them again in another context, either with different data or for a slightly different purpose. I will certainly have to convert data from one format to another again and could adapt this project to do so. It was also great to get experience converting data across a wide array of formats, as opposed to merely building an ETL pipeline that exclusively converts from JSON to SQL, for example. This tool and the methods I learned while creating it will be very valuable moving forward.