What was easier than you thought and what was harder?

Developing the ETL pipeline was easier than I thought overall. This project was likely easier because the extraction step was largely completed for us because the data sets we acquired were cleaned up, and data cleaning is an arduous and time consuming part of data science. The transformations performed were fairly simple and Pandas DataFrames makes this process extremely easy. Saving files through code and bringing SQL into ipynb was a bit difficult to figure out with the syntax, but the documentation out there made this step manageable.

How would a utility like this be useful for other data projects you may encounter?

A utility like this would be useful for other projects by allowing you to select the format you want the data to be, but also allowing you to check what the data looks like during the process; which gives the user the option to use the data or not. This utility would be more helpful, in the specific case of our project, if you could upload your own data, have it visualized, select columns to add or delete, then have more output options. Our project could also be improved upon by have tighter data validation checks like finding missing values, outliers, or records that are not consistent. This would make sure that clean, reliable data gets passed through for analysis. One major benefit for building a project like this is that it has other use cases such as being a template for other data projects which could allow other teams to adapt this project, even the both of us, and build off it based on what’s needed for the projact. These adaptations could include better error handling logic like try, except blocks to ensure that we understand where certain problems arise in the communication process or the cleaning process. Overall, this utility is very useful for taking in data, transforming it, and loading it to a target. This project gave us insight into how an ETL pipeline could be useful outside of data projects as it could be useful for research purposes or other data such as user data.