Data Project 1 Reflection

This project was more difficult than I thought it would be. I began with the 'extract' part of this project, and wanted to make sure that the user didn't have to find the file path themselves. This required finding the files in the directory automatically, which was difficult to research since this was done in Google Colab instead of on my device. Another issue that I had with the 'extract' portion of the project was that many times my data would stop working suddenly and produce Value Errors. Even if I had just used the dataset, there were multiple times where I had issues such as 'trailing data' or syntax errors with JSON files when I reran a function. This was most likely caused by how complicated these files are, with multiple lists and dictionaries. This was solved by adding a try and except part of the function to handle those errors. Additionally, I had to redownload the data and reupload it, and this seemed to help with the issues, but it is unclear what was causing the errors in the first place when neither the dataset nor the code changed. Real-world data is often messy and requires extra error messages to ensure that issues are able to be understood by the user. However, I think that the core conversion between datatypes was actually easier than expected because there were prebuilt in functions from the pandas library.

This would be incredibly useful in a project where there are multiple datasets in different data formats that need to be combined. If I want to combine many datasets from different sources, this conversion work can be done automatically, saving time on my end. Additionally, I could combine multiple datasets into different tables, and then create a database and convert them into SQL. This seems like this utility would be present in larger data pipelines, where data needs to be transformed and stored in a consistent manner before analysis or machine learning modeling. It would save a significant amount of time and reduce the chances of human error during wrangling.