DS 2003 Project One Reflection

I really enjoyed working on this project because it helped me see the potential for real-world application in future data science projects. Exploring Kaggle was interesting, as I discovered many publicly available data sets. However, it was surprisingly challenging to find a JSON file – most of them had thousands of entries, which made it tough to find one that was not multiple gigabytes in size and had a publicly available API key.

I found it much easier to manipulate and work with CSV files. Loading in a CSV file felt intuitive, as I was more comfortable with the format, and separating the data by a comma made everything nicely organized in a table. This simplicity was not there for the JSON file, especially because the API I chose had about 150 pages with 50 records per page. I did not realize the dataset had over 7,000 records, so initially, I was confused when my record count returned only 50. The solution was to implement a loop to aggregate the records from all pages, which was more challenging since I had not worked with this concept in a while. Also, I thought it would be easier to work with all the data in one list instead of spreading it across pages, which took a bit to figure out as well.

Despite these challenges, this project served as an excellent synthesis of the skills we have learned this semester. One particularly, unexpected, valuable experience was learning how to sort through a large number of datasets to find one that fit the project's needs/scope. This process taught me how to assess datasets in terms of size, usability, and format. Furthermore, for any data science project, knowing how to convert files and process data is critical, as is understanding how to handle public APIs and manipulate external data. Overall, this project solidified these skills and provided practical experience that definitely can be applied to future work.