Emily Sun and Mariam Seshan

DS 2002 Project 1

10/20/24

Project Reflection

In this project, we designed and implemented a data processor that could ingest data from CSV and JSON sources, convert formats, modify columns, and store the processed data in a local file or SQL database. The project was a valuable learning experience, as we applied what we had learned in class while overcoming new challenges and errors.

One of the main challenges we faced was in modifying columns and generating the data summary. Initially, although our code indicated that the columns were correctly modified, the summary did not reflect these changes. We resolved this issue by testing different variations of the code and referencing previous class work. We also compared calling the dataset from Kaggle versus uploading it directly from an Excel sheet. We discovered that uploading the file produced the expected results, so we continued with that approach. This issue was more difficult than anticipated, especially since we've worked on similar tasks before, and we didn't expect to run into such a problem. We also faced some challenges in ensuring the CSV and JSON files were being correctly transformed into other formats. Sources like Stack Overflow helped with this as we were able to cross reference our issues with similar errors others had run into.

The most satisfying part was successfully converting data from CSV to JSON (and modifying it along with storing it). Once we figured out how to go from CSV to JSON, it became easier to work in reverse, converting from JSON to CSV. In Python, it's always a great feeling when everything finally clicks. This part ended up being more straightforward than we thought, especially considering JSON was less familiar to us at first. Another part we experimented with was moving around the steps of the instructions to test what worked more effectively. We began by testing modifying first before converting, but then went back to converting and then modifying. By playing around with the code, we became more comfortable with what we were doing and more confident in our results. In the end, we went with the process that felt more intuitive and resulted in the outputs we were satisfied with.

The skills we developed in this project will be highly applicable in the future. The ability to quickly ingest, transform, and store data from different sources can support faster analysis pipelines and smoother integration with multiple data environments. For example, in a future project that may involve working with open data sources or APIs, an ETL processor would be valuable for streamlining data preparation and enabling real-time processing.

Overall, this project allowed us to work on technical challenges while also reinforcing the helpfulness of Python. This experience has given us more confidence to apply similar work like this in the future, especially around things that will involve large and varied datasets.