Tessa Saporito

Professor Tupiza

DS 2002

20 October 2024

Project Reflection

This project had elements that were both familiar to me and challenging. Because of my experience with Python, the modification of the data between loading and storing was the most straightforward aspect of this project. I was able to easily add and remove columns. However, the most challenging element in terms of modification was devising creative ways to alter the already robust datasets. On the other hand, converting and storing were newer to me and thus required more attention, trial, and error.

For converting, I initially attempted to convert from one file type to another by having the user input the initial file type and desired output. However, I realized that this approach made the function longer and more error prone. I found that converting both file input types, JSON and CSV, into a pandas DataFrame first streamlined the process overall, rather than attempting direct file-to-file storage. Once I understood that conversion did not necessarily have to occur before modification, this decision also allowed me to make modifications in DataFrames before conversion and storage. Another challenge was storage. Initially, I tried storing data directly to SQL, but I soon realized SQLite offered a simpler process for saving a DataFrame and still fell within the purview of the project. Leveraging class resources and external online documentation, I was eventually able to work my way through the process of saving a pandas DataFrame to a SQLite database.

Learning how to create an ETL pipeline will prove useful because it equips me with the skills to handle large datasets efficiently and allow me to move efficiently between file types. Converting data into a uniform format ensures that information from diverse sources can be consolidated into a single data storage platform. Additionally, the transformation aspect of ETL allows data to be cleaned and standardized. Overall, the ETL process is useful for optimizing data accessibility and usability, which supports faster and more reliable data analysis.