Anna Brown

Professor Jason Williamson

DS 2002

October 18, 2024

Data Project 1 Reflection

For Data Project 1, I authored a segment of an ETL pipeline that could ingest two formats of pre-defined data sources— a JSON file containing 2023 U.S. Border crossings and a CSV containing D.C. crime incidents in 2024. After fetching these remote data files from a provided URL, I transformed it into a usable format, modified the data frames, then loaded it into a storage destination for export as selected by the user. While I did ultimately produce a successful ETL pipeline, I did also face a couple of challenges that gave me a better sense of areas where I can improve when working with data.

First, working with multiple data formats—specifically JSON and CSV—highlighted the complexities of cleaning and transforming datasets. A major challenge was dealing with inconsistencies in data structures and date formats, which required extra time to standardize. Specifically, I struggled with importing complex JSON files with deeply nested data that contained both dictionaries and lists. These types of files require detailed importation code that surpass Pandas' capabilities (to my knowledge). In addition to these format mismatches, I also encountered frequent connection errors, which made implementing robust error-handling mechanisms essential.

On the other hand, some aspects of the project were more straightforward than I expected. Once I got comfortable with the Pandas library, tasks like filtering, renaming, and reordering columns became relatively easy. Furthermore, exporting the processed data into CSV,

JSON, and SQL formats was surprisingly smooth, thanks to Pandas' user-friendly methods. This seamless exporting capability was a confidence booster, showing me how easily processed data can integrate into different applications and workflows.

In conclusion, the utility built through this assignment will be highly useful for future data projects. Having a structured process for ingesting, processing, and exporting data sets a solid foundation for more complex analysis. This experience will be especially valuable for projects involving multiple data sources, where consistent transformation and formatting are critical. The techniques I developed—like date splitting, column renaming, and ensuring uniformity—will help me handle larger and more complex datasets with greater confidence. Overall, this project reinforced the importance of adaptability and precision in data work, skills that will be crucial in my public policy research and other future endeavors.