Rachel Leach & Elizabeth Wu
DS 2002

DS 2002 Data Project 1 Reflection

This data science project aims to deliver an ETL (Extract, Transform, Load) pipeline capable of ingesting and processing raw data in the form of JSON or CSV. The data processor is designed to convert data sources into either CSV, JSON, or SQL, allow users to delete specific columns, and store the modified file locally.

One of the initial challenges we encountered was finding an appropriate JSON file to work with. While locating a suitable CSV was relatively straightforward, many JSON files varied widely in size and structure. Some of the available JSON datasets were too small and lacked the complexity needed to effectively test and demonstrate the functionality of our data processor. These discrepancies highlight the importance of understanding the structure and content of datasets before starting the coding process.

Working with the CSV files turned out to be easier than expected, likely due to our previous experience with this format. We found that the tabular format was more intuitive to manipulate and convert. In contrast, printing the JSON data in the correct format proved to be more challenging than anticipated. The nested structure of JSON files required extra attention to ensure that the output was both readable and accurate. Furthermore, maintaining the integrity of the data when converting from CSV to JSON required some trial and error during the debugging process.

A utility such as a data processor can be incredibly beneficial for various future data science projects. By automating the conversion of data formats, it saves time and reduces manual effort, while also expanding the range of datasets available for analysis.The modification features of the processor can be particularly valuable, as they enable users to focus on relevant data, minimizing noise and improving data visualization. Additionally, the processor is scalable, which is crucial in fields like public health, market research, and sports analytics, where larger datasets are common. Moreover, the utility can also be expanded with additional features or integrations, such as additional file formats or user modification capabilities, making it highly adaptable for future projects. Overall, this utility enhances efficiency in data processing and manipulation, making it a valuable asset for any data-driven projects.