The biggest challenge for me during this project was finding and using a JSON file or API. During our search, it quickly became clear that csv files are the more common format for available data. The first JSON file I tried to use turned out to only contain metadata such as column names and descriptions. The next one I tried to use was poorly organized and nearly unusable to me. After failing to find interesting and usable environmental data in json format, I decided to try an API call instead. The first site I tried was the EPA's Envirofacts API. The data was interesting but seemed complicated to use and overwhelming for a beginner. After some searching, my partner and I found the public carbon intensity API through GitHub. This was simpler to use and provided clear instructions for calling it, similar to the finance API from class. The data was an informative measure of air pollution and also loosely related to the csv file we planned to use. Unfortunately, though, it only contains data for Great Britain. It would be interesting to compare emissions between countries if a comparable, more comprehensive dataset exists.

After identifying our data, constructing the code was fairly easy with my knowledge of if statements and Pandas. For both the API and the CSV file, the easiest method seemed to be converting to a dataframe, modifying the columns, and then converting to the output format. Ingesting and modifying the CSV file was more straightforward for me because this is what I had more past experience with.

This ETL pipeline will be useful for future projects as it can be modified and used for virtually any csv file or API. It allows for easy viewing, manipulation, and export of data, which could be helpful for the data scraping and analysis component of other projects.