Charles V Fisher
ITIN8000 FA2021
HW2 Reflection

Github Repo Link:
https://github.com/cfisherCPL/ITIIN-8000-Assignments-Charles-Fisher

Write a 2-3 short paragraph reflection on your experience while working on this homework. Things to consider including are:

1. What seems to be the pros and cons of CSVs vs JSONs?
2. Did you parallelize your code? How would parallelizing it affect Part 2?
3. What things helped you work well?
4. Are there things you would have done differently if you had to redo the assignment from scratch to make your life easier?

CSVs tend to be easily legible and portable to a number of gui apps for simple inspection compared to JSON. With Pandas installed both are wicked easy to work with, and that ended up being a big chunk of my brain fog on working with all of this. The basic CSV and JSON packages work similarly to java in keeping the document stream open or closed and that made general sense and made some very pretty looking JSON. Took more lines of code than Pandas, but made sense to me. Pandas just sorta...does it all with a function call, but the json it put out was ugly to parse out visually. Still, it just worked when testing by re-writing it to a csv with pandas.

I did not parallelize my code simply due the attempt to get the assignment in not-as-late. However, the flowchart and architecture diagram showcase HOW this could be done. Given that the dataset csvs to be fed in are pretty darn large, reading them each into their own panda object at the same time could have marginally sped up work overall. Pushing the read-in and output selected columns function to asynch using multiprocessing library woulda been pretty darn cool, and most likely necessary for larger datasets.

Pandas. Without a doubt this lil library saved so much hubbub that I'm actually upset I spent so much time trying to NOT use it and do the work exclusively through the csv and json imports.

I should have simply trusted the json output coming from the pandas write function. Even though it looked un-pretty, it is MORE than fully functional and portable for use. Really that just goes to show how parsing large sets of data really isn't a thing for human eyes! I would have liked to get to practice parallelizing this one ahead of our work with MachineLearning. The early version of part 2 didn't take advantage of pandas at all and I wasted days needlessly trying to brute through what should have been a file