ETL Project

Data Selected: <https://www.kaggle.com/Cornell-University/arxiv>, at the time of obtaining the data the JSON file was 2.61gb big. Most of the size is found in the abstracts of each entry, would be interesting to see how big the dataset would be if it included the whole content of each paper. Python took about one whole minute to read the JSON file, utilizing about 80% of full CPU power and almost 100% RAM according to task manager; coincidentally I did upgrade my laptop’s RAM from 8gb to 16gb before the bootcamp began.   
Filtering: I chose to make a SQL table of 5 columns: authors, update date, title, categories, and abstract. For the categories, there are 60303 categories, this many because a paper having multiple categories counts as a different category. Using “.unique()” on the categories column doesn’t yield all of the different categories, it gives a … in the array. I would have liked to make a couple of tables that were divided based on these different categories, but that can be done in Postgres with queries.

Moving the database to PgAdmin took 3 minutes for my laptop to process. The final database was chosen this way for queries to be utilized for further filtering. The only downside is that the titles and abstracts contain LaTeX markdown, which jupyter notebook can convert but PgAdmin can’t.