**ETL Project Write Up**

The goal of this ETL project was to create a Database that a coffee shop owner could use as a starting point to research local competition.

**Extraction:**

I pulled data from the below three sources to create a Philadelphia Coffee Shop Database:

1. Yelp Fusion API

2. YellowPages using webscraping

3. The L&I Business Licneses (CSV) from the following link: https://opendataphilly.org/dataset/licenses-and-inspections-business-licenses

\*Note: The CSV exceeds 100 MB so it was too large to upload to GitHub. To rerun the Jupyter Notebook, download the CSV locally, rename it "business\_licenses.csv", and place it in the "resources" folder.

**Transform:**

To pull the city/zip codes from YellowPages, I used BeautifulSoup. I stored the outputs in lists, created DataFrames, then concatonated the DataFrames.

For the Yelp Fusion API, I had to use multiple calls to get the desired number of businesses since the limit per call is 50. Once I retrieved over 100 coffee shops, I used json.loads and json.dumps to get the data into a format where I could create a DataFrame. Once I did that for all three pulls, I concatonated the three dataframes to create one df for the Yelp data.

Once I read the CSV into the notebook, I renamed the business name columns to "coffee\_shop", and capitalized the business names within the column so they would have the same format, and merged the dataframes on the "coffee\_shop" column.

To clean the data up, I droped any rows that did not have a Yelp ID, since Yelp had the most info of the three sources. I then deleted all rows that had duplicate phone numbers so that I wouldn't have any duplicate businesses in the table.

**Load:**

To load the data, I created a database in Postgres and then created a table within that database that shared the column names of the DataFrame. I then populated the Postgres table with the combined DataFrame.