Our original data sources were a pair of CSV files, one from data world, and one from Kaggle. The data world CSV documented the populations of every city in the United States, while the Kaggle dataset showed the location (city, state, address, and coordinates) of every Chipotle restaurant in the United States. These two datasets were selected to hypothetically compare the correlation (or lack thereof) between population and the presence of a Chipotle location.

All transformations were done in a Jupyter Notebook, and of the steps taken to clean and transform the data, the first was to rename the columns we wanted to keep and to remove the unnecessary columns. Then we created a second data frame of the Chipotle data, grouped it by city, displayed the counts, and merged it back on to the original data frame so that we could display the count of how many Chipotle locations one city had in a single row. After that we renamed the two “state” columns and dropped duplicate rows. With the city population data set, less cleaning was required. All we needed to do was remove the extra columns, rename our columns to match what we had in SQL, and reset our index.

We chose to load our data sets into SQL, where we joined the two tables on city and state to prepare the data for examination. Then, all queries were run in PostgreSQL. A segment of the final table is shown below:

