

Fast Food – ETL Project

Project Team – Kristi McGrath, Todd Auman, Bobby Huffstetler

**Project Objective – Applied ETL to Population Data and Fast Food Restaurants.**

**Data Sources:**

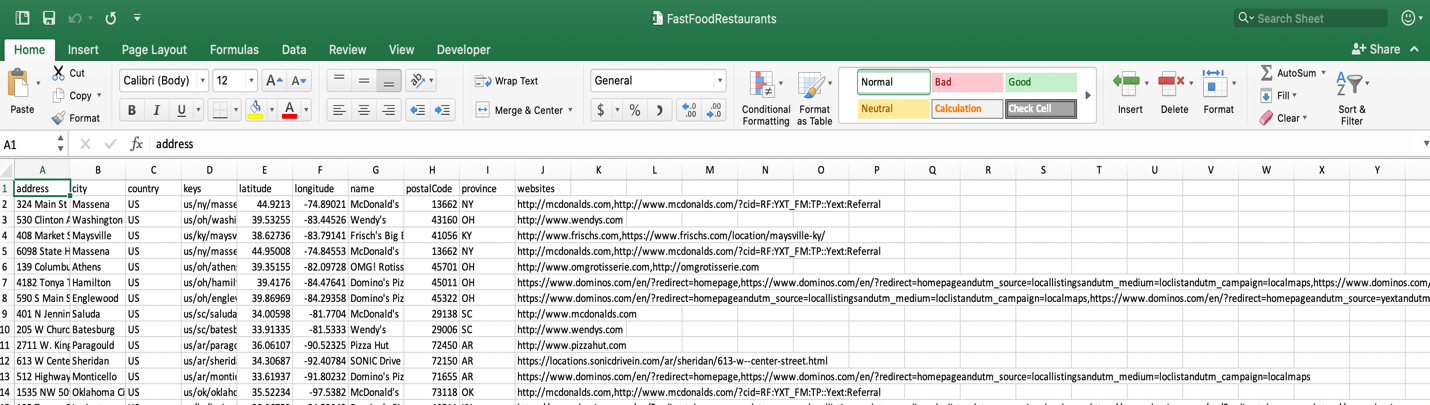
Kaggle.com. Kaggle Fast Food Restaurants Across America dataset. This dataset consists of a sample of approximately 10,000 fast food restaurants from the much larger Datainfiniti’s Business Database.

[Fast Food Restaurants](https://www.kaggle.com/datafiniti/fast-food-restaurants)

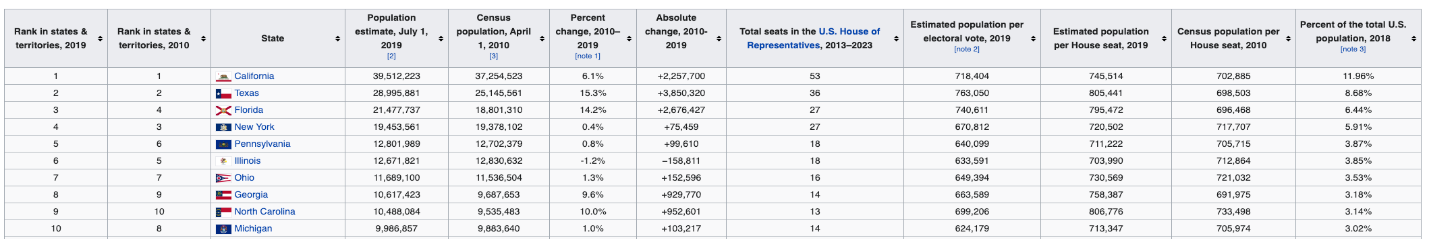
Wikipedia. The ‘List of U.S. states by population’ and ‘U.S. postal abbreviations’ pages.

[Population](https://simple.wikipedia.org/wiki/List_of_U.S._states_by_population)

**Data Extraction:**

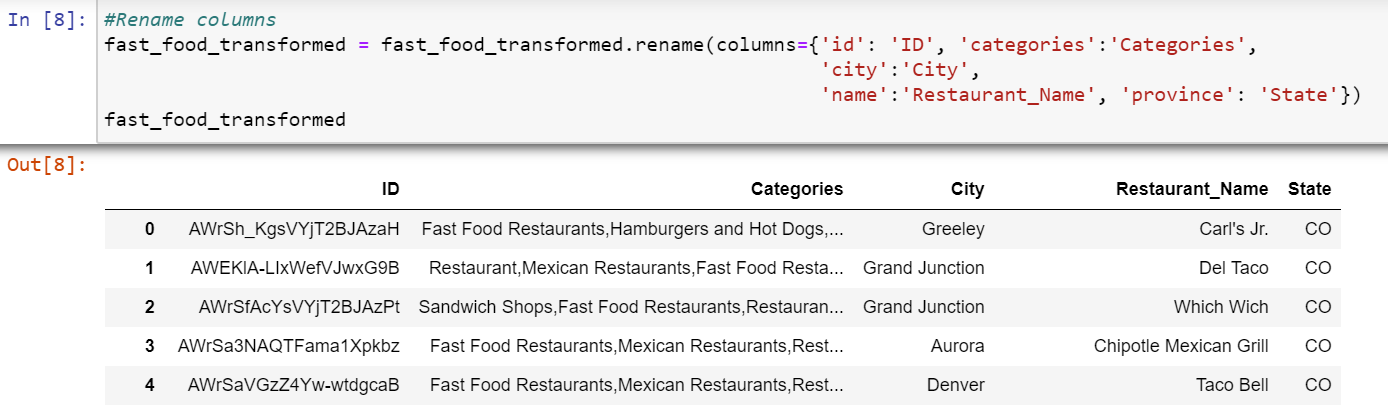
The ‘Fast\_Food\_Restaurants.csv’ file from the Kaggle Fast Food Restaurants Across America dataset (See Resources Folder) was read into the ETL\_main Jupyter Notebook and converted to a dataframe (fast\_food\_df).

The tables from the state population and state abbreviation wiki pages were scraped for their tables and transformed into a Pandas DataFrames (pop\_df and state\_abbr).

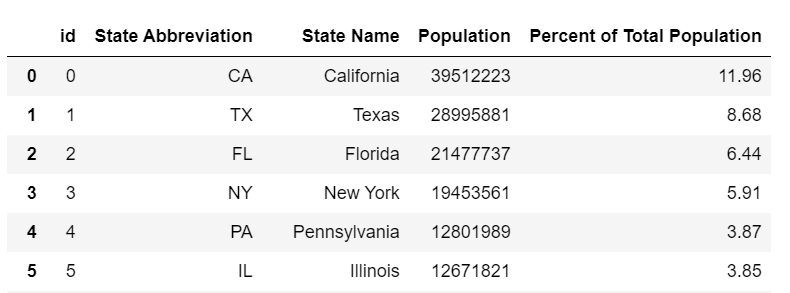
****

**Data Transformation:**

The fast\_food\_df was filtered to remove unwanted columns and the columns were renamed to match those in the restaurants table in the restaurant\_db PostgreSQL database.

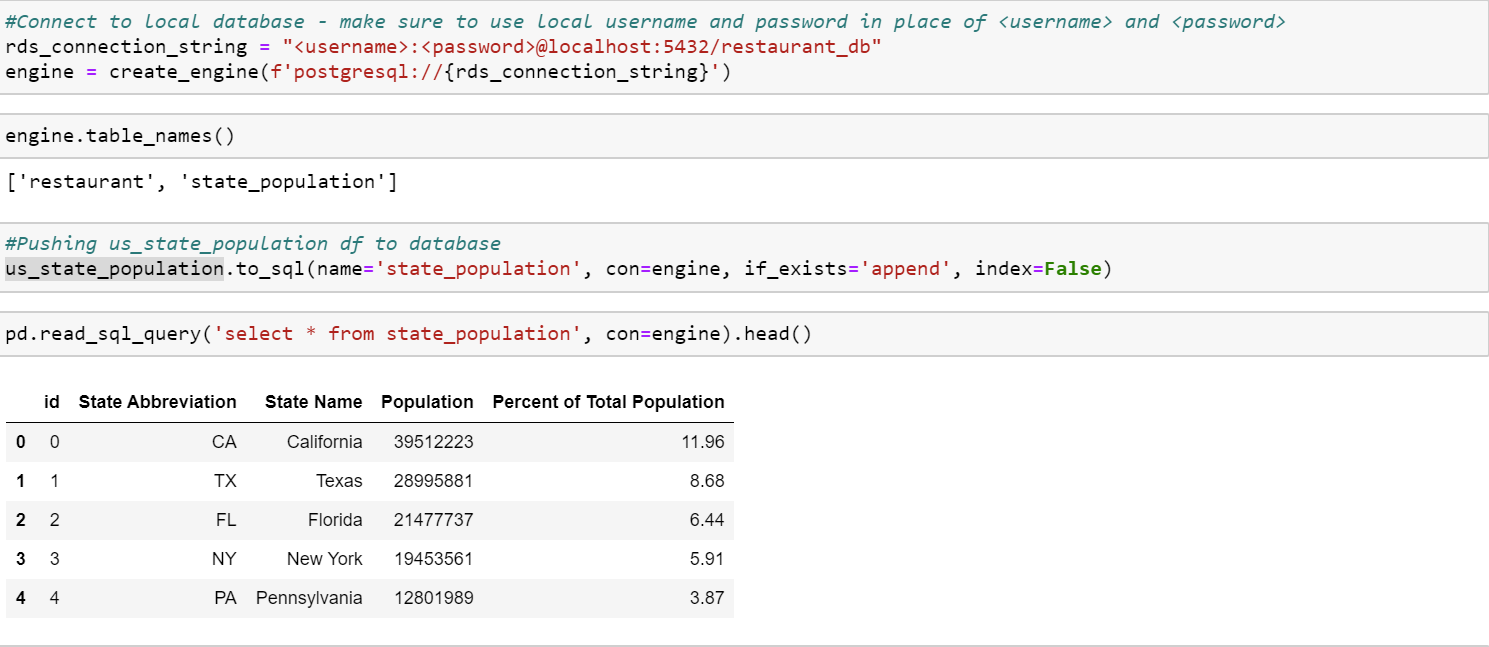


The pop\_df was filtered to remove unwanted columns. Then table rows not corresponding to the 50 official U.S. states were removed. The data in the ‘Percent of the total US population’ column were stripped of the “%” and the data type was converted to ‘float’. Next, the state abbreviation dataframe was merged with transformed state population data frame to get the ‘State Abbreviation’ data column. Finally, the columns were renamed to match those in the state\_population table in restaurant\_db and an index was added to serve as a primary key.



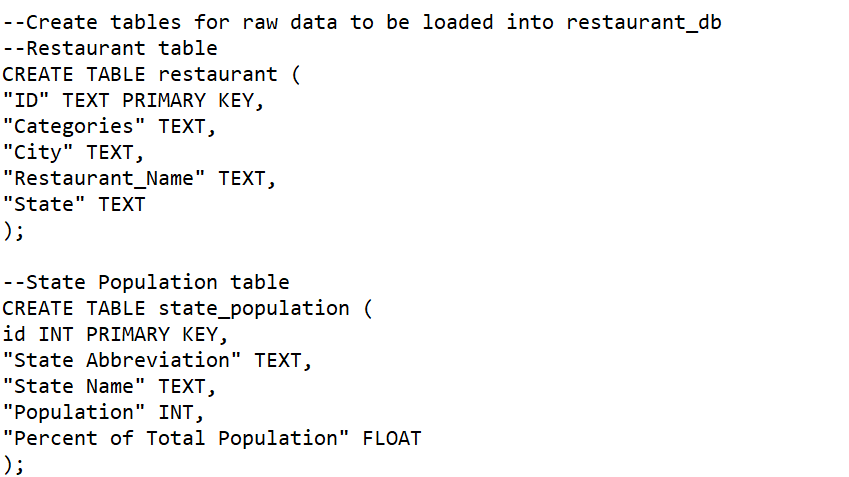
**Data Loading:**

The transformed dataframes (“us\_state\_population” and “fast\_food\_transformed” were loaded into the restaurant\_db using pandas ‘to\_sql’ function.



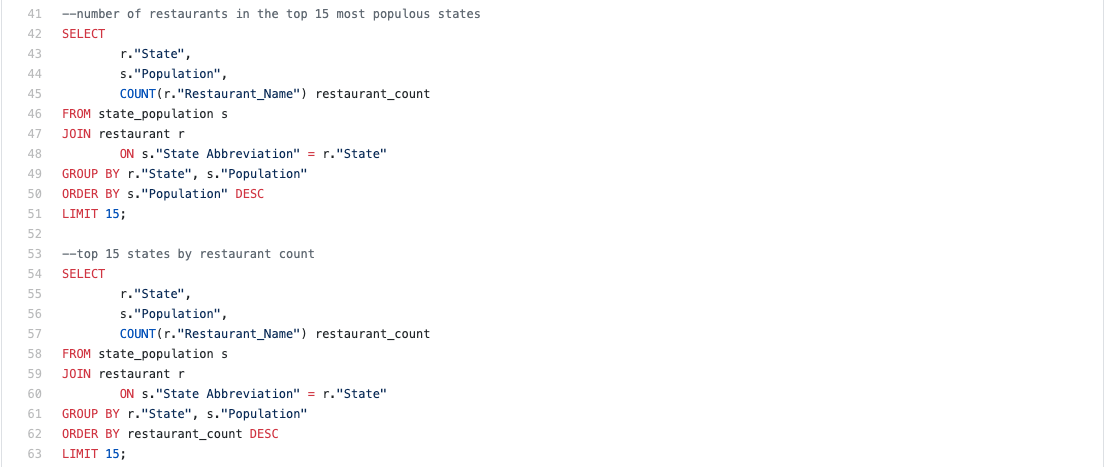
**PostgreSQL Database:**

The restaurant\_db was created in a pgAdmin4 with two tables (‘restaurant’ and ‘state\_population’) according to the following schema:

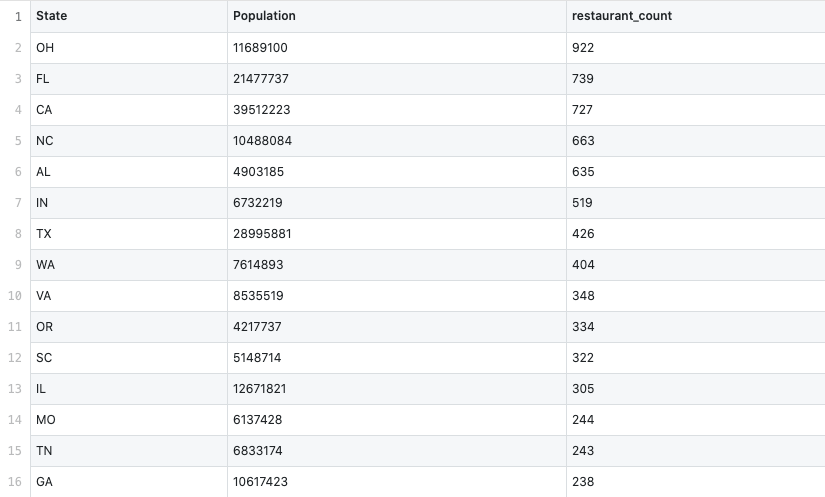


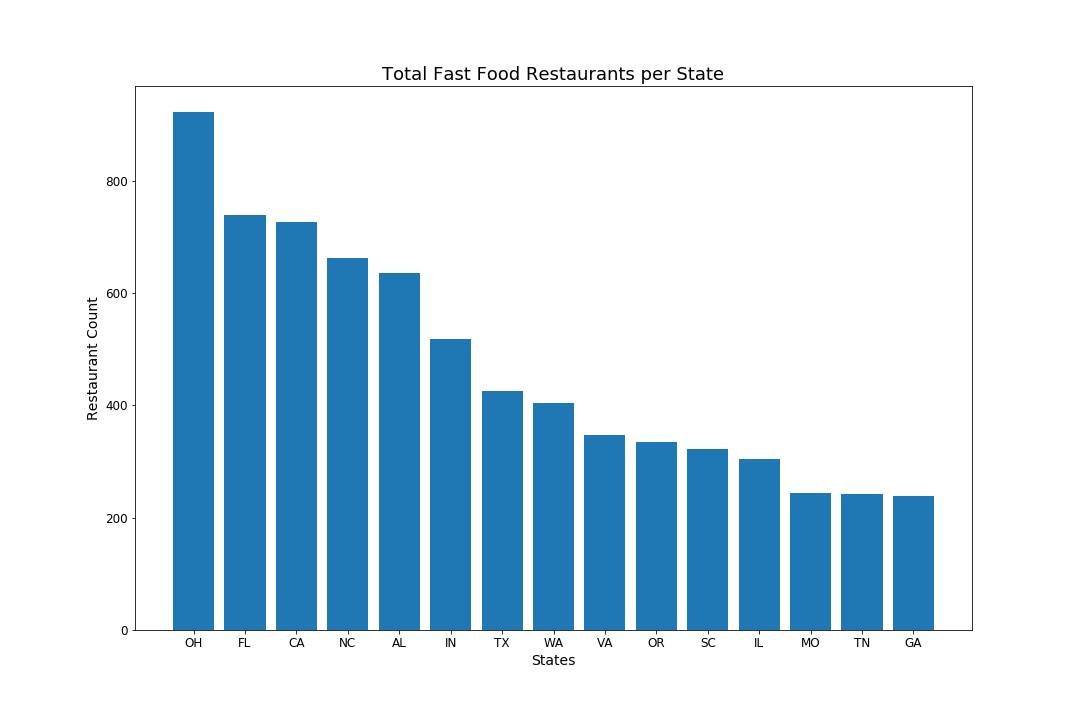
|  |
| --- |
|  |

From these two tables queries can be generated to return the number of restaurants in the top 15 most populous states and the 15 states with the most restaurants in this dataset.

****

See below for the data table of the top 15 states by restaurant count.

****

**Observations**

Ohio ranks 7th in state population but has the highest number of fast food restaurants 922.

California ranks 1st in state population and has 727 fast food restaurants.

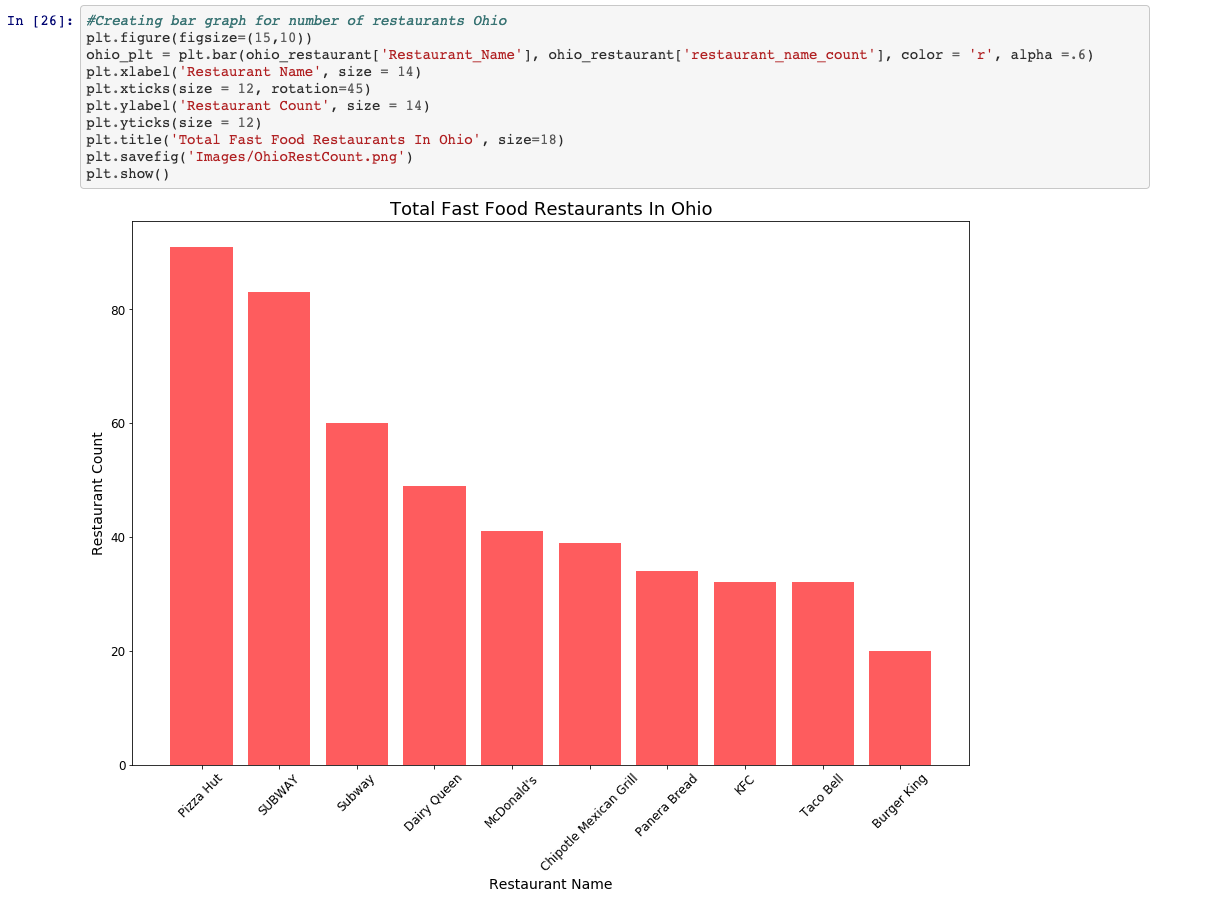
New York ranks 4th in state population has the lowest number of fast food restaurants 173.

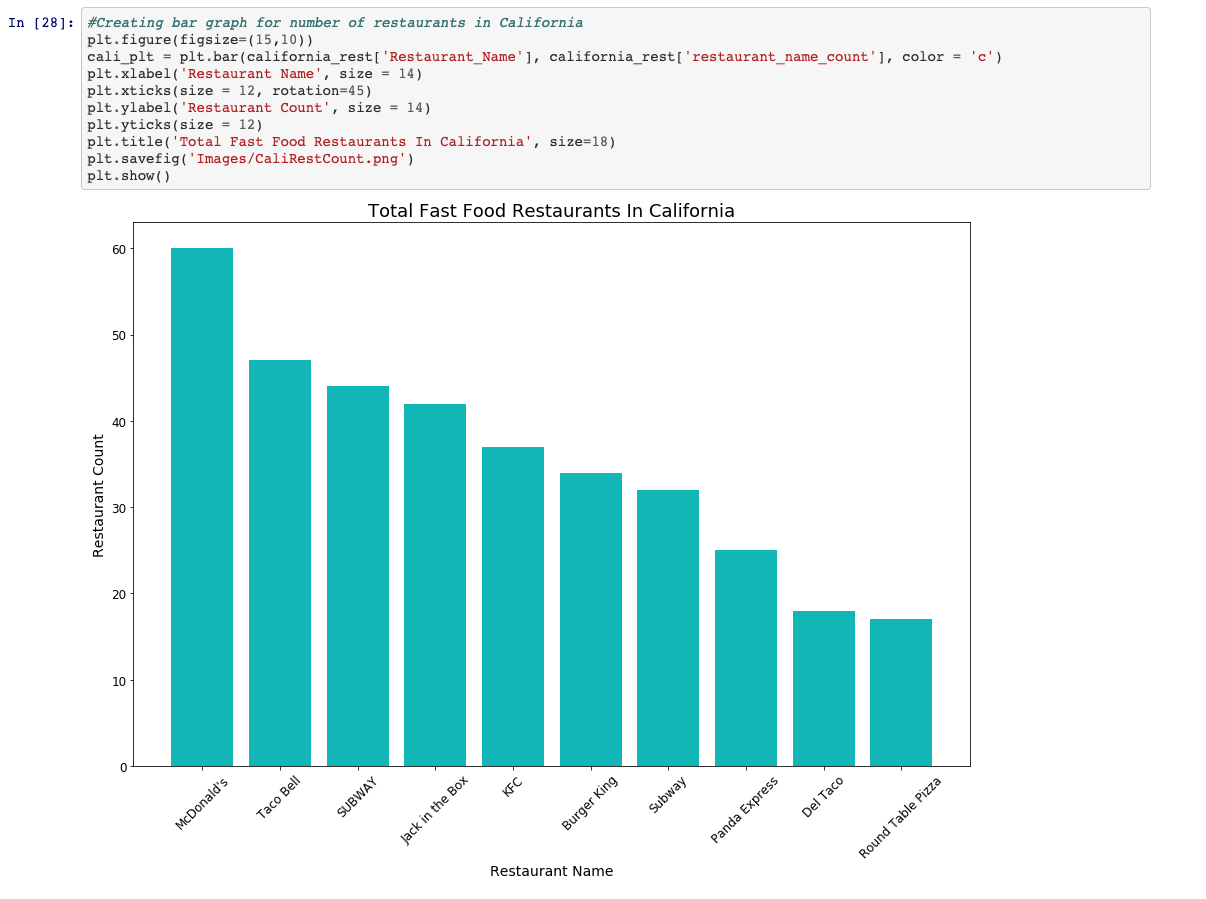
* New Jersey & Virginia had erroneous data for restaurant count in this sample dataset

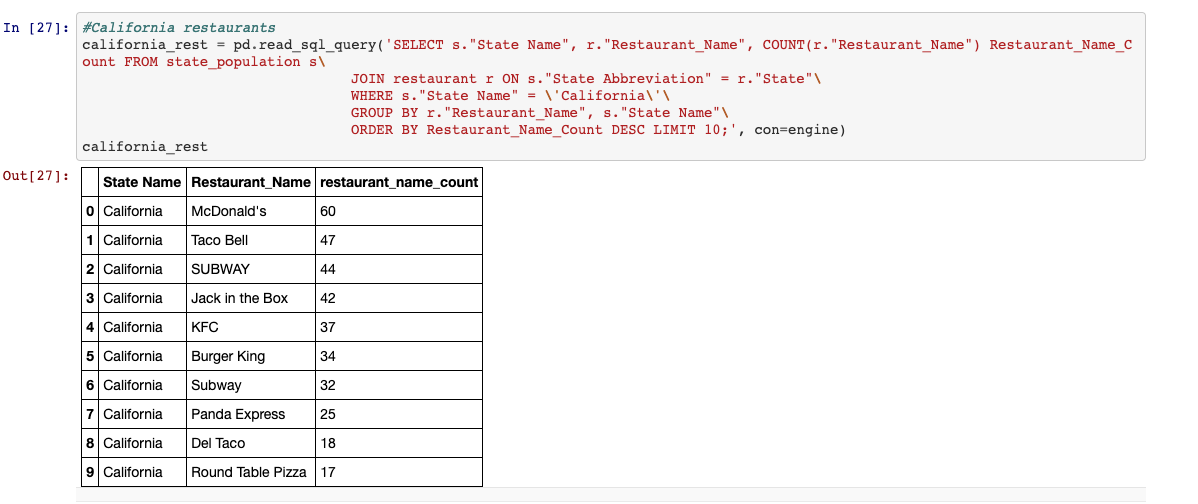
We concluded that population size has little to do with the number of restaurants in the top 15 list.

**Extra Bites**

Once we concluded our research and data merge, we thought it would be interesting to breakdown the top 10 restaurants in the state of Ohio, since they have the most of any state.

****We can see the Buckeye State truly enjoys pizza.

In our data set, California has the largest population. We were curious as to what their top 10 fast food restaurants were.

The Golden State is Lovin’ It. McDonald’s holds the top spot.