Jean-Luc Dinello

Wrangling Safegraph' Census Population and Neighborhood Patterns data

The Consumer patterns data from safegraph comes in a relatively clean form directly from its website. The other nice part about this data is that they also give out the complementary census data set in a far cleaner form factor than the one that comes from the official government site. The census data that is provided from Safegraph is broken down into very granular locations called census block groups. For each block group, populations are estimated for a large range of demographic factors such as age groups, income, ethnicities, service members, etc. This dataset contains over 3,000 columns detailing each census block group. Along with the census population datasets, there is a neighborhood patterns dataset that shows each census block group's top related brands, number of visitors, the visitors home and work census block groups, the median distance travelled by visitors, and its popularity by hour and day. There is also a dataset containing geographic information on each census block group such as the coordinates, the amount of land and water, and the county and state that each cbg is contained in. The final piece of information to tie this all into answering my overarching question is a patterns data set the SafeGraph was kind enough to give me free of charge that details similar features of the census patterns data but is specifically for Chipotle locations. This data also comes with additional features such as the number of Android and iOS users for each location.

The census population data was broken into twelve separate datasets due to its size. The first step was to combine these datasets. The initial step that I took was to glob the datasets together and use a list comprehension to read each dataset into a dataframe. After I had my list of dataframes, I concatenated them together to form the entire population. The column names of the population datasets are originally shown as code names that are detailed in a metadata dataset. In order to explore and analyze the data further I decided that mapping these field names to the dataframe was a worthy endeavor. This was a fairly straightforward mapping because each of the columns in the 'patterns' dataset was translated to a specific field attribute. However the challenging part of this was doing it at the necessary scale. There were over 7,000 features. I created a special function to map the columns with the full field name. After some reindexing, I eventually was able to to concatenate the datasets with the full field name as their column title.

After getting the population data into a single dataframe, I then added the geographic features and the neighborhood insights data. These datasets were far smaller and were indexed by census block group making their additions much more straightforward. I then started to look at the Chipotle data and saw that it was missing the coordinates of each location. Using the requests library along with BeautifulSoup, I scraped this information off of Chipotle's website and merged it with the Chipotle patterns on the street address of each location. Safegraph provided the data with uniform formatting and without null values. With all of the necessary data in a working format, the next step is to explore and analyze Chipotle visitors based on the demographic features of their census block group.