Data Wrangling for Prediction of Fast Food Demand

The Datasets

We would like to use existing community data and the current presence of Fast Food establishments in these communities to predict areas that are currently underserved by this type of business. We will be obtaining community variables from the Safegraph collection of 2016 census data and for compatibility with this data set will be drawing our community boundaries from the Census Block Groups the data is grouped by. We will then add variables about daily average visitors to these communities, and whether fast food establishments are among the most frequented establishments in the census block groups from the October 2018 Safegraph Neighborhood Patterns dataset. Finally we will use the Safegraph data shop to obtain current total counts of fast food establishments in the listed communities.

Census Data

As the full set of census data will likely be beyond our computational and conceptual capabilities, we will filter our census data to census blocks in the author's current city of residence, Philadelphia. This will require using a streaming json reader over geometry/cbggeom.json file in the Safegraph data folder to obtain the CBGs in Philadelphia county, and then only keeping data entries in those CBGs as we look through the census csvs. We will be making some arbitrary picks from the many available variables in our census information to create a more manageable and general set that we think are best suited to predict success of fast food establishments.

In our downloaded safegraph census data folder, there is a `data` folder containing various csv tables with one census category per table. The tables filenames correspond to their census table ids (ex cbg_b00.csv) that can be seen at
https://www.census.gov/programs-surveys/acs/guidance/which-data-tool/table-ids-explained.html

A further breakdown of the content in each of these tables by their column ids are available in the metadata/cbg_field_descriptions.csv table. We see that there is highly detailed data available, so we will use these field descriptions to help us sum various fields to arrive at high level observations.

Among the information available, we will be amalgamating the following variables:

Census Filename	File Data	Variable Obtained
cbg_b00.csv	Overall Population	Total Population
cbg_b08.csv	Commuting	Total Working at home in CBG
cbg_b11.csv	Household Types	Count of Single occupant households
cbg_b11.csv	Household Types	Count of single parent households (at least one minor present)
cbg_b11.csv	Household Types	Count of two parent households (at least

		one minor present)
cbg_b14.csv	School Enrollment	Count of enrolled undergraduate and graduate students
cbg_b15.csv	Education Attainment	Count of residents with attained high school Diploma or GED
cbg_b15.csv	Education Attainment	Count of residents with Bachelors, masters, or doctorate
cbg_c17.csv	Poverty status	Total Households below poverty line
cbg_b19.csv	Income	Count making 40k and below
cbg_b19.csv	Income	Count making 40k-100k
cbg_b19.csv	Income	Count making 100k and above
cbg_b23.csv	Employment Status	Total employed count
cbg_b25.csv	Housing Characteristics	Total renter occupied households
cbg_b25.csv	Housing Characteristics	Total owner occupied households

Some of these fields are read directly from the census files, and some are summed from various fields. The specifics can be seen in the python notebook in this directory. In the case of each of these variables we are reading in the relevant columns with pandas and producing single column dataframes with our variable, each with an index of the Census Block Groups.

Neighborhood Patterns and Existing Locations

In addition to our census variables we will be adding a few more metrics from other Safegraph obtained data. First for each CBG we will acquire the total monthly visitor count from the Safegraph neighborhood patterns data set. Then from our purchased data set of existing fast food restaurants within Philadelphia, we will use the census geocode service to map CBGs to each establishment, and then will aggregate the amount of already existing fast food establishments in each of the CBGs as well as total fast food establishment visitors for each cbg.

Final Check and Storage

We finally concatenate all our gathered measures into one data frame and do a check for outlying values we might want to remove before storing our final data set. We discover that there are a few rows without data that we will want to remove, and also some additional rows with CBGs well out of the range of the other CBGs that we remove for consistency of data. Having cleared these rows we save the full dataframe as one new csv. As before, all specifics are in the python notebook next to this file.