Optimal Location Selection for Fast Food

Establishing a new fast food restaurant is a substantial investment and a franchisee only gets one chance to pick the location to house that investment. Picking a different location could mean the exact same restaurant experiencing a different level of visitorship or operating cost. Any person opening a new restaurant would do well to know the areas with the highest profit potential before they start to put down any groundwork. This investigation aims to provide a model for the potential entrepreneur to identify those areas.

The analysis to produce this model will be kept to a feasible scope by constraining the investigation specifically to the city of Philadelphia. When the investigation is complete, the model will be evaluated for its ability to generalize to areas outside of Philadelphia.

Gathering Model Variables

To measure the potential success of different areas for our fast food restaurant, we will have to rely on observations of current consumers. This means we will need to identify areas currently in Philadelphia where there is a high volume of fast food consumption, and then find the variables that best correlate to those areas.

The set of area variables that we choose to pick from are mostly from the extensive measurements gathered in the US Census' American Community Survey. These data are offered in a convenient set of CSV files by Safegraph free of charge: https://www.safegraph.com/open-census-data. The units of area of the US Census are Census Block Groups (CBGs), and so we will be mapping our fast food restaurants to their corresponding CBGs to line up our measurement variables.

In addition to census variables, we will also make use of Safegraph's recorded visitor count to each CBG, which they also offer free of charge at https://docs.safegraph.com/docs/neighborhood-patterns. Then from outside Safegraphs offerings we will also consult the City of Philadelphia's metadata on crime reports (http://metadata.phila.gov/#home/datasetdetails/5543868920583086178c4f8e/representationdetails/570e7621c03327dc14f4b68d/) to put together a crime count variable for each CBG. Finally we consult Walk Score (https://www.walkscore.com/) for numerical transit and walkability scores to add to our list of predictor variables.

For an output variable we purchase information about the locations of fast food restaurants in Philadelphia and the patronage experienced by each from Safegraph's data shop: https://shop.safegraph.com. To develop our model we will need to tease out the features of the areas that contain the top performers in that list.

Census Data Choices

We use the latest US Census American Community Service from 2016 understanding that, while it will not perfectly reflect the current state of Philadelphia, it is likely to provide us the best possible picture of differences between different city areas.

For our investigation we need to extract observations specific to Philadelphia from the full set of census data. And then, as we would be unable to conduct an investigation of every variable in that survey, we will also need to make some arbitrary up front choices about what subset we do choose to investigate.

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 that can be seen at https://www.census.gov/programs-surveys/acs/guidance/which-data-tool/table-ids-explained.html (ex cbg_b00.csv). Each table is keyed by CBG, so we will need to subset each table to the set of CBGs that are located in Philadelphia.

To filter out this set of CBGs, the Safegraph census data provides the file geometry/cbggeom.json identifying the location of each CBG. We use a streaming json file reader to traverse the large file and extract our set of Philadelphia CBGs. Then as we extract our variables of interest we will subset our data with this initially gathered set of CBGs.

A breakdown of the content in each tables by column ids is available in the Safegraph 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 amalgamate 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. Our final output will combine all these variables into one python Pandas dataframe indexed by CBG.

After this gathering step we find a small subset of our CBGs with entries of 0 in most of our variable columns and choose to remove these rows from our final investigation.

Additional Predictors

After collecting our census information for Philadelphia into one table by CBG, we will also add information to each row about the traffic each CBG experiences. We join the Safegraph data/cbg_patterns.csv neighborhoods patterns tables total raw_visitor_count column to our census table on the CBG index.

We turn to the city of Philadelphia to join a crime count metric to our table of predictors. The CSV of crime occurrences that we use has data from 2006 onwards. To have the best picture of the current state of Philadelphia, we will limit our crime counts to instances recorded in 2018. In order to total how many crimes occur in each CBG we will have to map each recorded Philadelphia crime instance to a CBG first. Using the lat/long location that Philadelphia provides for every crime instance, we call the US Census geocoding service on each location and collect the CBG for each crime. After the completion of this step, we add a total count of instances in each CBG to our table of variables.

To get Walk Score's walkscore and transit score for each CBG, we act on the assumption that the central point of each CBG will have a walk score and transit score representative of the CBG as a whole. We get the central lat/long of each CBG to call Walk Score with from the Safegraph geometry/cbggeom.json file. As Walk Score also requires an address to call their api,

we use Google's map API to get addresses for each CBG's central point. With these two pieces of information we collect the scores for each CBG and add them to our predictor table.

Output Variable

Our output variable that we will build our model around is the current success of fast food locations in Philadelphia. Using the Safegraph data shop we purchase data on each 'limited service restaurant' they have recorded in Philadelphia. As they do not record the CBG of each restaurant we first turn again to the US Census geocoding service to obtain the CBG of each location. Then to each CBG we add the total count of restaurants recorded in that CBG, and the total of visitors experienced by all of the restaurants in each CBGs. We are now ready to do investigation.

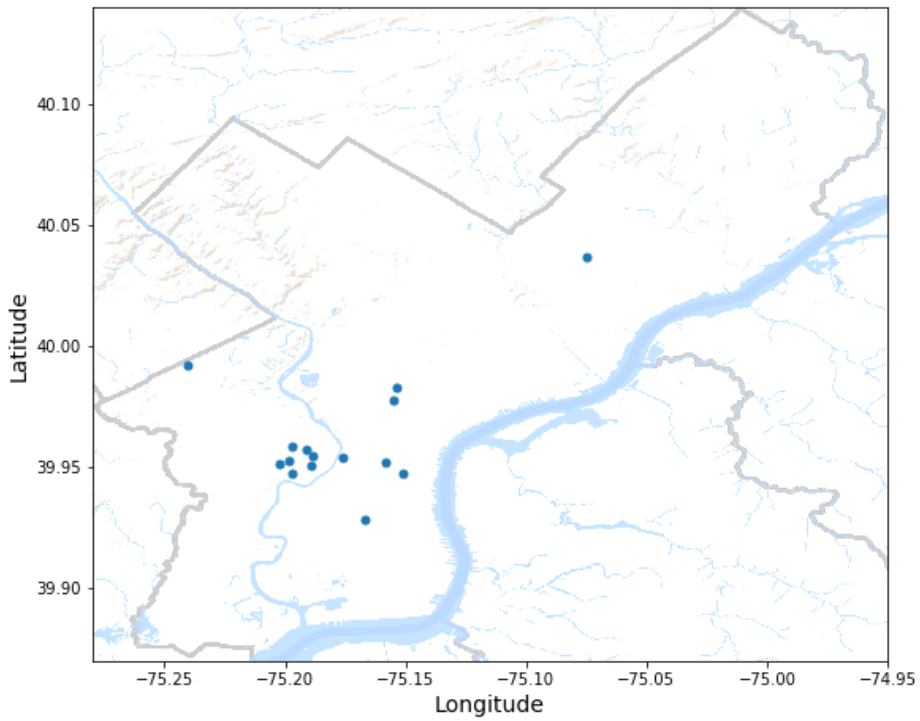
Data Analysis

In analyzing what areas are most successful for fast food establishments, we will not be able to derive any trend from areas that have no recorded fast food restaurants. We therefore start by filtering out the observations of any CBG that does not have any recorded establishment. This leaves 187 CBGs inside of Philadelphia.

Among the remaining CBGs we would like to get an idea of which ones experience the highest consumership and what can be said about them. One of the first things we might want to know is where in Philadelphia those most successful areas are. First we should identify successful CBGs.

We have to draw some arbitrary boundary to define what areas are the successful areas. For a first look, we can define any CBG with average visitorship to establishments at the 75th percentile mark or higher as a successful area. We find that this mark is 667 customers a month. As our ending interest is also where we might place a new restaurant, we narrow this set a little further by making sure if we remove 667 customers off the top of the visitor count in each CBG to represent a potential new entrant, that the average business still experiences at least that amount, meaning there is a good chance such an area can support another establishment.

Mapping these CBGs we get:



Which provides the immediate insight that university students are a primary consumer group to target for fast food as we see the cluster of successful CBGs in ‘University City’. We also spot two successful CBGs North of downtown near Temple University that would support this initial insight.

We look to our observed data to back up this initial assumption by calculating the correlation of each variable to total count of fast food consumers in that area. Indeed we find that, among our gathered predictors, the highest correlated variables to total fast food consumer in a CBG are:

Undergrads and Grads	0.546434
raw_visit_count	0.546264
Crime Counts	0.376914
Transit_score	0.343890
Pop. that Works from Home	0.254979
Total Population	0.245195

With University students leading the way in correlation to success of fast food establishments. Our other metrics seem to correspond directly to potential audience size, but with high volume of visitors seeming to be more relevant than a sitting high population as raw_visit_count and

Transit_score both outweigh ‘Pop. that Works from Home’ and ‘Total Population’. We do also see that high crime counts seem to correlate with high fast food consumption, which might not be exactly what we expected as it seems a risk to the average business, but it may be that low crime is correlated with other community factors that also correlate to high utilization of fast food restaurants. To tease out that exact connection is beyond our scope though, so we will use this as a positive predictor as is.

We do also surface some lighter negative correlations that might be worth using to note the wrong community may be in consideration for a new location, namely, we see the negative correlations:

Married Parent Households	-0.176664
Owner Occupied Households	-0.188451
Single Parent Households	-0.189105

Showing family communities and communities with more established residents make less use of fast food establishments.

A single linear correlation value cannot necessarily capture how important each factor is though. It could be the case for example that while all highly trafficked communities are high in some metric, not all communities that are high in that metric are highly trafficked and this lack of a strict relation may not map well to a single correlation value.

To try to capture this another way, we can group all communities that are high in a metric and see if as a group they capture more than their expected share of the market. We must again arbitrarily draw the line of what ‘high in a metric’ means. This time we choose the 80th percentile and above in each metric, and then map to what percentage of all fast food consumers in Philadelphia are accounted for by for those top percentile. In a truly random world these top 20% would account for the 20% of the total market, so we are interested in the communities that earn significantly more (or less) than their share. The chart below displays these measurements by total market share:

The first thing we see is that the top 20% of CBGs for total fast food consumer count account for almost two-thirds of the total market (more than 3 times their expected random share). A new business owner would do well to be as informed as possible about the characteristics of one of these highly desirable top scoring communities. Some of our other metrics do surface as similarly over represented in their share of the market, though it does seem that they were captured by our correlation values. Again we find that the areas that should capture our eye the most when placing a new establishment are the highly visited areas (those with high visitor count and good transit) and those with high quantities of University students, each capturing more than double their expected market share. At the other end, we again see that family communities are unlikely to be high utilizers of fast food restaurants and should probably be passed over in the placement of a new restuarant.