Part 2: Data and How it will be Used

Data

An MSA is "a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core. … Under the standards, the county (or counties) in which at least 50 percent of the population resides within urban areas of 10,000 or more population, or that contain at least 5,000 people residing within a single urban area of 10,000 or more population, is identified as a "central county" (counties). Additional "outlying counties" are included in the CBSA if they meet specified requirements of commuting to or from the central counties."

Because the definition insists on economic integration, MSAs vary widely in size. The data are drawn from the Census Bureau and include, for each Metropolitan Statistical Area (MSA) of the United States.

- Personal Income per person (Census)
    - A better measure than GDP per person (which includes corporate retained earnings, etc.), this quantity measures the average income received by the residents of that MSA.
- Jobs-Population ratio (Census)
    - Calculated by dividing the number of jobs in the MSA by the population of the MSA. While (a) not all jobs are filled and (b) some people have more than one job, so this is not the same as an employment-population ratio. However, it is a good measure of the economic vitality of an areas.
    - Also notice that some residents may work outside of the MSA while some jobs are filled by people who commute into it. However, the definition of MSA (to include outlying commuting areas) reduces this bias.
- Proportion of college educated and graduate-school educated (Census via SSTI)
    - Share of the population 25 years old or older with university degrees: associates, bachelors, graduate/professional, and “at least bachelors.”
- Population Density
    - Craft breweries often become a neighborhood business. Moreover, the nature of their product is such that driving long distances from or two them is unadvisable. Therefore, it makes sense to assume that higher population densities are correlated with a higher density of breweries.
- Longitude and latitude of MSAs (Statcrunch)
    - If specified criteria are met, a metropolitan statistical area containing a single core with a population of 2.5 million or more may be subdivided to form smaller groupings of counties referred to as "metropolitan divisions."
- Number of breweries within 30 miles (50,000 meters) of the center of the MSA (Foursquare)
    - Foursquare has a sub-category for "breweries". While some of these may be macro-breweries (like Anheuser-Busch), small breweries vastly outnumber this. Focusing on breweries rather than "beer bars" excludes places that serve beer but do not brew it.
    - A call is made to Foursquare using the coordinates provided by SSTI
- Brewery Density: breweries per 10,000 people. Divide number of breweries by the population and multiply by 10,000.

How it will be used

Multiple Regression

Come up with a prediction, using multiple regression, of the amount of people-per-brewery (PPB) that should exist in each MSA. The model to be estimated is

$$BrewDensity = \beta_0 + \beta_1 Income + \beta_2 Jobs + \beta_3 Education + \beta_4 PopDensity + \epsilon$$

The strategy is to then estimate the error, ϵ, which is the part of Brewery Density that is not explained by income per capita, jobs-to-population, educational attainment, or population density. The residual, r, is the estimate of the error. A positive residual is an indication that a particular locality has more breweries per 10,000 people than would be expected given the locality's income, jobs, education, and population density.

We want to locate MSAs that are relatively underserved by breweries but not a craft-beer desert, in the assumption that a sufficient critical mass (not too few breweries per 10,000 people) is necessary to support a new brewery, but that too many breweries would crowd out a new one. Therefore, the residuals will be divided into quartiles. The second quartile (from the bottom), that is, the one composed of MSAs that are "below prediction" in terms of their brewery density but not hugely so will be considered the most promising set of MSAs for a new brewery.

Density Based Clustering

A separate question is whether one should pick a location that is far from other locations or close to another. Because craft breweries sell beer, which makes driving difficult, distance rapidly reduces the attraction of a location. While craft beer can be bottled on the spot and taken far, most people prefer to drink from a brewery they know, from brewers they know. For that reason, a brewery that is 30 minutes away is not a competitive consideration one way or another. On the other hand, density increases the likelihood of friendly government regulation, easy access to suppliers, and a ready-made customer base.

Density-based clustering is a method for collecting locations (or objects) in "natural" clusters. Using Density Based Clustering, we will find MSAs that are similar in income, jobs, education, population density ("determinants"), and brewery density. Some clusters will be high-determinant, high-brewery clusters. Others will be low-determinants, low-brewery clusters. The strategy is to find high-determinant, low-brewery clusters.