Business Ecosystem Analysis

Introduction

Finding communities with underdeveloped markets in different business categories can inform investors, entrepreneurs, and public officials of opportunities for growth. Here we will analyze various communities to determine which business segments are least represented.

For example, if most communities of a certain size have X restaurants, a community with X/2 restaurants might be a prime location to consider opening a restaurant.

Data

Foursquare data is used along with census data and the OpenStreetMap API. Foursquare will be used as the source of venue information.

Foursquare

The relevant Foursquare data uses the “explore” endpoint. We will use the venue category information from this data. Example: “Gourmet Shop”. The categories they assign to each business will be the basis for assessing the level of competition of the market in that category.

OpenStreetMaps

This API will give what they call the bounding box for cities.

Example:

"41.327988", "41.373571", "-85.147331", "-85.097333”,
which are the minimum latitude, maximum latitude, minimum longitude, maximum longitude, respectively, of the city. We will use this box as the area to search Foursquare data for venue information.

Census Data

We use the 2017 census data obtained in their sub-est2017_all.csv file to get the populations of the cities that are evaluated.

Example: “Indianapolis” : 872680.

BestPlaces.net

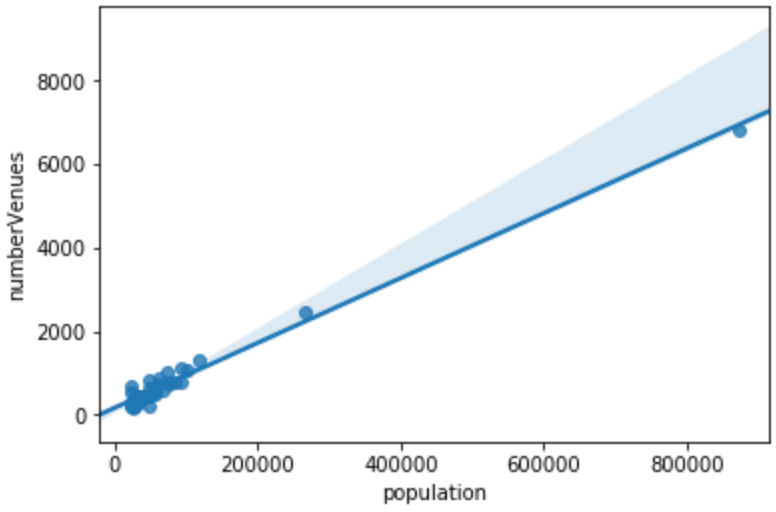
Income per resident was parsed from the web page returned with the url: https://www.bestplaces.net/economy/city/indiana/{}".format(city)

Methodology

With the bounding box geocoordinates, a grid of geocoordinates was constructed to ensure the radius sent to the Foursquare explore endpoint covered every point within the box while keeping the radius size small so as not to reach an API maximum number of venues per query.

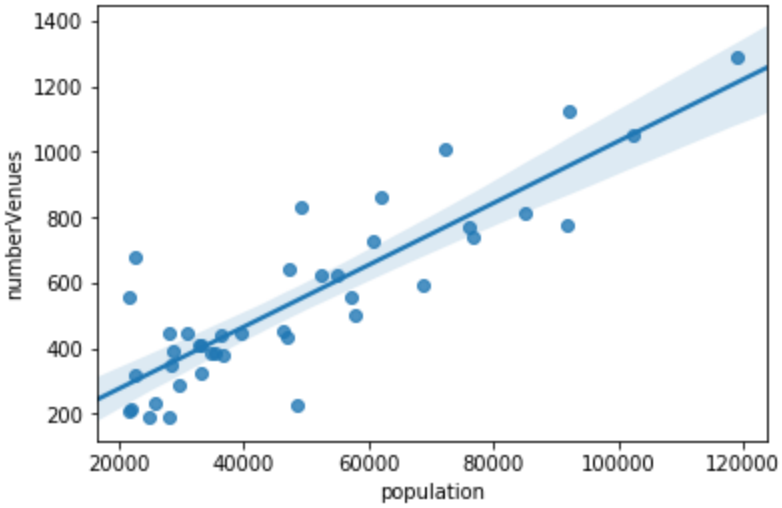
Exploratory Data Analysis

The correlation between overall venue count and population was evaluated and found to be approximately linear.
Linear Regression of relationship between population and number of venues in Indiana cities with a population greater than 20k



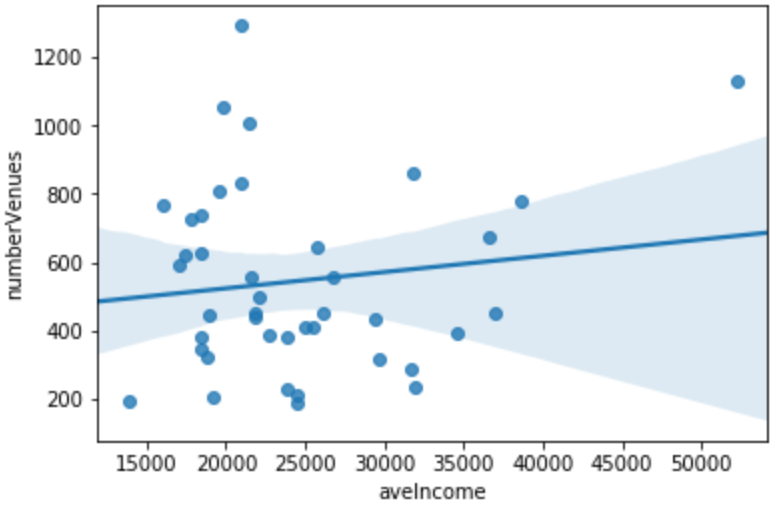
.With y=mx+b, m=0.0078, b=170.

For cities under 200k in population:



With y=mx+b, m=0.0095, b=86.

For cities under 200k in population, correlation between income per resident and number of venues:



With y=mx+b, m=0.0048, b=428.

With income and population showing some correlation with venue count, they were used with sklearn multiple linear regression modeling to determine a trendline.

The predicted number of venues from that model was then subtracted from the actual number of venues. This number was then considered to be a signal that a particular city should be investigated for underrepresentation of businesses, generally.

The same approach was then applied to the business categories that were present in all 43 cities with populations above 20k. Indianapolis and Fort Wayne were excluded as their populations where for beyond the values of the rest of the cities.

Results

Top Ten Cities with fewest venues relative to multiple linear regression model based on income and population. “deltaModel” is the difference between actual number of venues and that predicted by the trendline.

city	population	aveIncome	numberVenues	predictedVenues	deltaModel
Lawrence	48704	23801	226	544.333339	-318.333339
Fishers	91832	38600	777	1008.452024	-231.452024
Franklin	25089	24471	189	324.063321	-135.063321
Brownsburg	25911	31964	235	360.740251	-125.740251
Kokomo	57836	22102	499	623.955084	-124.955084
East Chicago	28215	13868	192	312.640633	-120.640633
Columbus	47143	29396	435	551.196338	-116.196338
Muncie	68625	16986	593	706.025907	-113.025907
Crown Point	29625	31673	289	394.666321	-105.666321
Greenfield	22094	24458	212	295.749183	-83.749183

Most numerous categories in cities over 20k

Category	count
Fast Food Restaurant	1242.0
Pizza Place	1194.0
American Restaurant	841.0
Sandwich Place	761.0
Park	743.0
Bar	728.0
Mexican Restaurant	663.0
Hotel	539.0
Construction & Landscaping	487.0
Discount Store	481.0

Categories in all cities over 20k:

Category	count	numberCities
Pizza Place	1194.0	43.0
Mexican Restaurant	663.0	43.0
Bar	728.0	43.0
Liquor Store	409.0	43.0
Coffee Shop	452.0	43.0
Sandwich Place	761.0	43.0
Fast Food Restaurant	1242.0	43.0
American Restaurant	841.0	43.0
Grocery Store	385.0	43.0
Chinese Restaurant	384.0	43.0
Construction & Landscaping	487.0	43.0
Baseball Field	336.0	43.0
Discount Store	481.0	43.0
Video Store	360.0	43.0
Golf Course	260.0	43.0
Pharmacy	466.0	43.0
Park	743.0	43.0

The categories represented in all cities with a population over 20k where then each analyzed with the same process used to evaluate overall venue count. The top opportunity in those cities is listed below. “Min” represents the difference between the actual number of businesses in that category minus the predicted number.

city	min	minCategory
West Lafayette	-15.634305	deltaModel Fast Food Restaurant
East Chicago	-14.895415	deltaModel Fast Food Restaurant
Lawrence	-13.439069	deltaModel Fast Food Restaurant
Fishers	-12.264502	deltaModel Bar
Bloomington	-10.268434	deltaModel Fast Food Restaurant
Greenwood	-10.082050	deltaModel Bar
South Bend	-8.948254	deltaModel Pizza Place
Elkhart	-8.228488	deltaModel Bar
Evansville	-8.190686	deltaModel American Restaurant
Merrillville	-8.050455	deltaModel Pizza Place

Discussion

With the methodology applied here, opportunities for adding additional fast food restaurants would seem present in West Lafayette, East Chicago, Bloomington, and Lawrence. Fishers, Elkhart, and Greenwood should be looked at for potentially opening up bars. South Bend and Merrillville might be missing some pizza shops and Evansville some “American Restaurants”

Conclusion

.

Foursquare was used as the primary source of venue information. However it was clear that this data is incomplete. Numerous categories where one would expect representation in nearly every city only showed in several. This analysis would be more informative if completed on a higher quality dataset.

The R-squared value for the regression on total number of venues was .75. Other features may offer more explanatory power, particularly demographic and geographic factors.

This analysis makes the most sense when all venues are equal, which clearly isn’t the case.

Revenue data could help ensure venue activity is appropriately weighted.