Final Capstone Project: Analyzing Venue Information based on Urban Index Values

Kyle Phillips, Capstone Project 2021

Introduction and Business Problem

Background

Across the country, in towns large and small there are restaurants, service stations, hotels, and shops that make up the landscape of life. Obviously, where there are more people, there are more places. Not as obviously, what kinds of places pop up where there are lots of people, compared to the kinds of places that exist in rural communities? This analysis seeks to answer that question.

Business Case

The problem being examined is the one of urban vs. rural areas and the types of venues that might be successful in each. Imagine that a development company has several tracts of recently zoned farmland, but also recently acquired some vacant lots in a re-emerging area of a large city. What types of venues are common in the rural areas that might do well on the farmland? What types of venues are popular in the urban downtowns of populous places? How could someone find out what is already out there?

Interest

A development company who is looking to invest or has recently invested in land might be interested in this analysis to decide what type of venue to put on the land. A city planner might be interested in this data, or a subset, to understand what types of venues could be prevalent in their region, or what kind of venue might be missing in their urban or rural community. As someone who came from a rural area of Ohio, it was interesting to see the similarities and differences between areas of personal experience.

Data Acquisition and Cleaning

Data Sources

There were 2 main sources of data used in this analysis. Using a publicly available dataset from FiveThirtyEight (https://github.com/fivethirtyeight/data/tree/master/urbanization-index), it was possible to examine latitudes and longitudes by state, and also by what the people at FiveThirtyEight are calling the "urban index." This is a measure of how many people are living in a 5-mile radius of other people using census data. This was a desirable data set because it already contained latitude and longitude information, in addition to the state name, and gave the classifications of urban vs. rural areas that the analysis is trying to answer. The data was downloaded and read into the notebook from a .csv file, which also made it convenient to use.

The second data source was using FourSquare venue information, the analysis examined what types of venues existed based on the location from the first data set.

Data Cleaning

Once the data was read in from the .csv file, the dataframe was pared down to only the target state for analysis. The data contained in the final data set included all 2940 census tracts from the state of Ohio. An example of the raw location data is included.

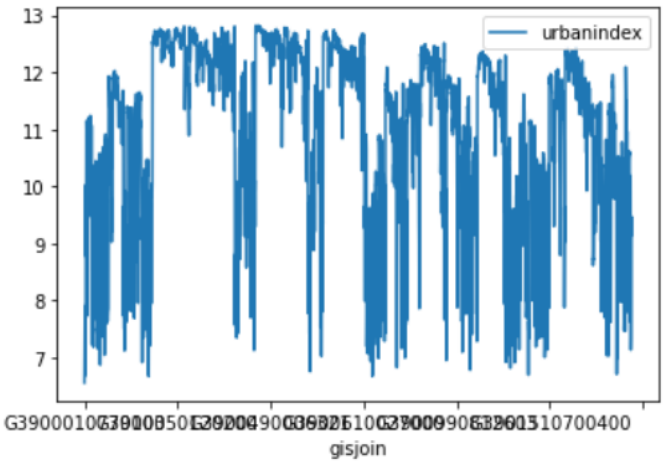
	statefips	state	gisjoin	lat_tract	long_tract	population	adj_radiuspop_5	urbanindex
50023	39	Ohio	G3900010770100	38.95705	-83.35256	4493	701.5263	6.553258
50024	39	Ohio	G3900010770200	38.98275	-83.54929	4998	1151.1370	7.048505
50025	39	Ohio	G3900010770300	38.84060	-83.58295	7133	2701.5280	7.901573
50026	39	Ohio	G3900010770400	38.77373	-83.53587	4149	2701.5280	7.901573
50027	39	Ohio	G3900010770500	38.75594	-83.35669	3567	792.3294	6.674977

From there, the data was usable in this format, although some of the columns were not needed, they were not dropped from the analysis in this step. The Foursquare data was pulled into the analysis at a later step.

Methodology

Determining Rural vs. Urban Locations

The first step was trying to determine what number of urban index indicated the most urban areas for the analysis (the highest on the index) and the most rural (the lowest). The first step to finding out was to graph the data and see if that provided any insights. The graph is below. The data was not organized in any time order, so this plot is not very useful except for to show the approximate bounds of the index values in Ohio.



The descriptive statistics of the data frame were much more useful.

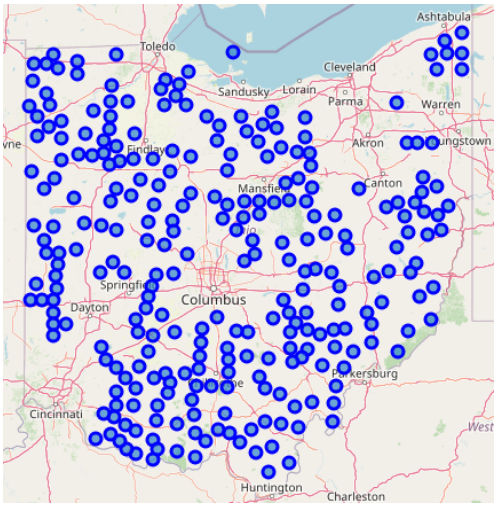
	urbanindex
count	2940.000000
mean	11.031029
std	1.544474
min	6.553258
25%	10.188453
50%	11.499205
75%	12.258868
max	12.822030

From the descriptive statistics, the bounds of the urban index value show that the most urban areas would have a value of over 12.25 and the most rural would have a value under 10.18. This is where the splitting of the data set started.

The first bound was to show the urban areas. The areas were mapped iteratively until they covered a good cross section of the urban areas in the state. The final value ended up covering 706 census tracts. This is any area with an urban index number over 12.28. These areas were mapped to see what areas of the state they covered. The map is below.

The initial urban areas were concentrated in Columbus and Cleveland. The data set was expanded until more of the urban regions were included. The analysis was meant to represent the state and the different cities, if not included, could have skewed the information based on regional preferences, not necessarily by urban or rural characteristics.

The process was repeated for the rural tracts. The concern here is not so much the regionality of the data, but the lack of venues in rural areas to make an adequate sample. The rural sample was also expanded iteratively until the sample covered a good cross section of the state. The final boundary for rural census tracts ended up at 8. The map showing the rural areas is included.



Foursquare Data

The next step was pulling venue information for the rural and urban data sets. The first step was defining the function to get information and then running it for each of the data sets. The rural data set was first. On a first pass, there were 1757 locations in 273 locations. Looking at the descriptive statistics for the number of venues per location, however, showed some locations with significantly more venues than others. This points to a commercial area or some other type of outlier that would skew the types of venues in truly rural areas. See the initial statistics below.

	Venue
count	266.000000
mean	6.605263
std	6.782206
min	1.000000
25%	3.000000
50%	5.000000
75%	8.000000
max	67.000000

The standard deviation is high and the difference in the average location and the max is very large. To make a more homogenous data set, any location with more than 20 venues was dropped. The new data set statistics are below.

	Venue
count	257.000000
mean	5.727626
std	3.830865
min	1.000000
25%	3.000000
50%	5.000000
75%	7.000000
max	19.000000

Dropping the outliers impacted the number of venues, down to 1472, but the data was more representative of the rural areas. It was also a forgone conclusion that there would be less venues in the rural data, but the question is not how many venues are in each environment, but what kinds. This data should be a sufficient sample to answer that question.

The next step was to find the venue information for the urban data. The radius used for the rural data was 5km, the net had to be cast wide to find rural venues. For the urban areas, locations would start to overlap if the radius was 5km, so 1km was used instead. There were 19691 venues in the urban set, again this is to be expected, with more people, there are inherently more places. The descriptive statistics for the urban data set are below.

	Venue
count	706.000000
mean	27.890935
std	24.074901
min	1.000000
25%	10.250000
50%	20.000000
75%	37.000000
max	100.000000

There are outliers in this data set as well, but with so many venues, the information is unlikely to skew the results. The number of venues in the low density areas will not offset the census tracts with many venues.

Results

Types of Venues

A quick count shows that not only are there more venues in number there are more different types of venues in the urban areas. There are 448 unique categories of venues in the urban areas and 199 unique types of venues in the rural areas. The next step was to create a data set that counted the types of venues in each neighborhood. This was done for both data sets. Once the counts were created, the totals for each category were summed up and extracted from the data frames to create the final data sets, these were transposed to put the data in a format that was easy to manipulate and plot. The index was reset and the columns renamed so that the data frames could be joined together for comparison. The top types of venues for each area are included here.

Venue Type	Urban Total
Pizza Place	890
Bar	806
Sandwich Place	614
Fast Food Restaurant	551
Coffee Shop	487
Discount Store	485
Park	456
American Restaurant	453
Bank	417
Convenience Store	397
Pharmacy	374
Grocery Store	354
Gas Station	293
Chinese Restaurant	290
Ice Cream Shop	282

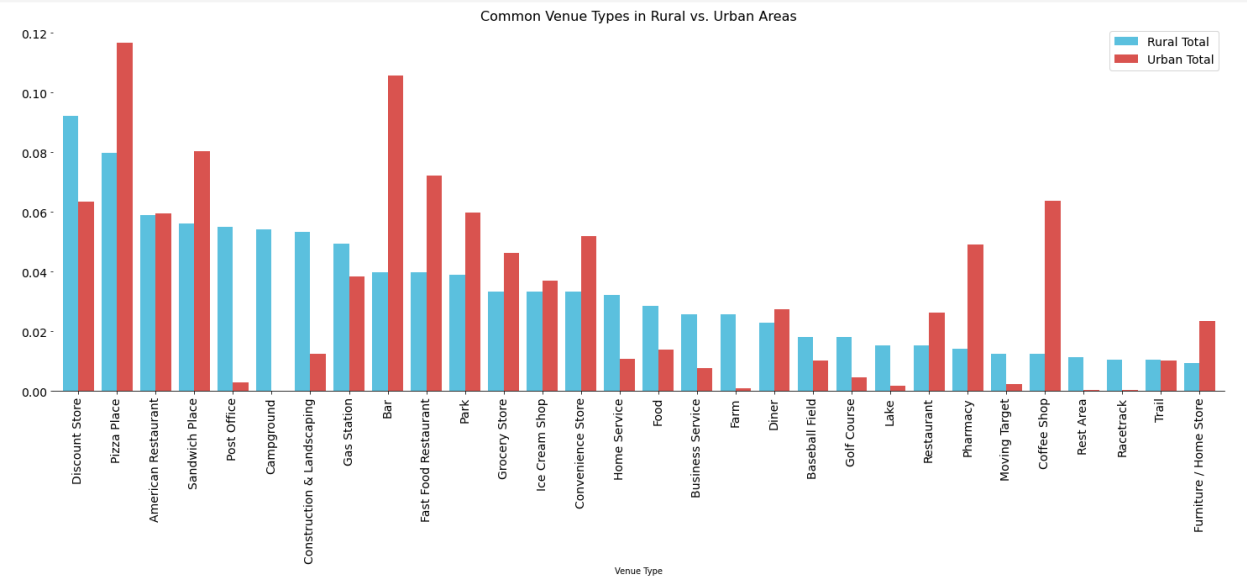
	Total
Discount Store	97
Pizza Place	84
American Restaurant	62
Sandwich Place	59
Post Office	58
Campground	57
Construction & Landscaping	56
Gas Station	52
Bar	42
Fast Food Restaurant	42
Park	41
Grocery Store	35
Ice Cream Shop	35
Convenience Store	35
Home Service	34

To be able to compare the data, it had to be normalized. I chose to do this simply, by dividing each category by the total number of venues. A sample of the normalized data is included here.

2: Urban Top Venue Types *1: Rural Top Venue Types*

Venue Type	Rural Total	Urban Total
Discount Store	0.092205	0.063573
Pizza Place	0.079848	0.116660
American Restaurant	0.058935	0.059379
Sandwich Place	0.056084	0.080482
Post Office	0.055133	0.002884
Campground	0.054183	0.000131
Construction & Landscaping	0.053232	0.012452
Gas Station	0.049430	0.038406
Bar	0.039924	0.105649
Fast Food Restaurant	0.039924	0.072224
Park	0.038973	0.059772
Grocery Store	0.033270	0.046402
Ice Cream Shop	0.033270	0.036964
Convenience Store	0.033270	0.052038
Home Service	0.032319	0.010748

The data was plotted against each other to be able to see how different types of venues are distributed in urban and rural areas. The graph is below.



Discussion and Conclusions

If someone or some company were trying to decide what type of venue to build in Ohio, a good place to start would be a pizza place. In both rural and urban Ohio, one of the most common types of venues is a pizza place. In urban Ohio, there are also a significant number of bars, sandwich shops, fast food places, and coffee shops. While in rural Ohio, the top spots are held by discount stores, pizza places, American restaurants, and sandwich shops.

This data was interesting to review in that it shows some similarities in the urban and rural areas in that pizza places and sandwich shops are both prevalent, but some of the mid-range numbered venues are vastly different, in that urban areas continued to be more commercial and food based, where in the

rural areas the focus quickly shifted to post offices and outdoors related venues. Having personally lived in both rural and urban Ohio, it was interesting to see the results played out with data and how that information both correlated and diverged from personal experience.

Next Steps or Further Investigations

A potential next step for this analysis would be to build a predictive model based on the urban index and see if it would be possible to predict the type of venue that is most prevalent based solely on the urban index number for a location. It would also be an interesting analysis to see how other states urban and rural areas compare to Ohio. Is pizza ubiquitous in the US, or just the Midwest? How do different regions of the country compare? Another question that begs answering is around suburban regions. Do regions that are not urban or rural follow the same patterns?