Food Availability in Indianapolis

Pete Stewart

December 22, 2019

1. **A. Introduction**

   Problem

   The city of Indianapolis, like many large cities in the United States, has experienced a trend in recent years in which many communities have seen supermarkets and grocery stores disappear due to business challenges in sustaining profitability in an industry notorious for low profit margins and high overhead costs. In what has become known as the “food desert” problem, many communities have fewer and fewer options for fresh, affordable and healthy food. The communities that are typically affected by this issue have tended to be more economically challenged and it is often theorized that the lack of affordable healthy food (the type available at a grocery or health food store) combined with prevalent and cheap fast food can compound economic challenges with health challenges in these communities. This is a large and multi-faceted issue but before we can really understand the implications of the issue we need to understand the nature of it and how it manifests in a community. I hope to use foursquare data to get a better understanding of, first that there is a clear issue, and second how the issue manifests itself in terms of the types of grocery venues that are available in different areas of the city.

   B. Interest

   There are many groups that have a business interest in solving this problem. The government has an obvious interest in solving problems for the citizens in these areas. In addition to simply keeping citizens happy there is the question of government healthcare spending. In the United States we have a high cost health system and for many lower income or older Americans the government

pays for a substantial portion of the healthcare cost so understanding the food availability for a population could certainly have an impact on health costs and therefore a large piece of the government budget. If my project is successful and we are able to better understand food availability we could follow this up with a data project to try to connect this data with government health data to see what correlations exist between food availability and health outcomes. Another possible business interest would be food retailers, since this project should offer insights into potential opportunities for grocery businesses. I would especially be interested in a follow up project bringing in data on transportation to see what opportunities might exist for grocery delivery services to those who might not have a car or access to public transit.

2. **Data**

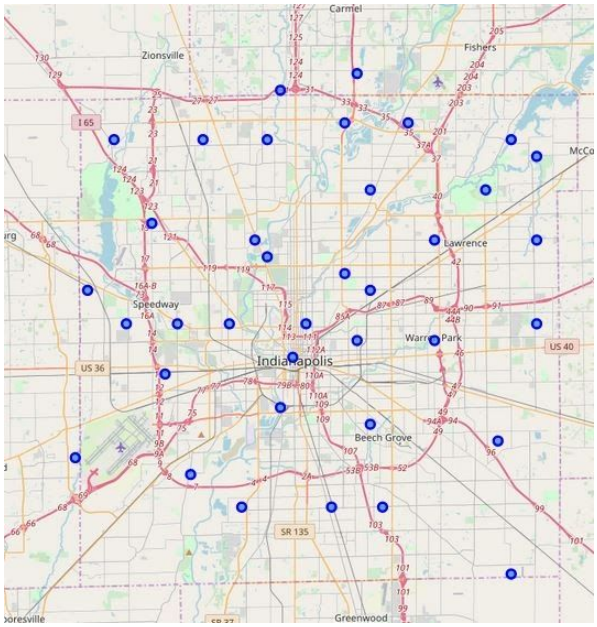
   I've identified several publicly available data sources to be able to segment the city of Indianapolis by zip code. My first source is the Us Postal service website (https://www.unitedstateszipcodes.org) which has a great deal of info on each zip code in the US. I was able to download a CSV file with info on all the zipcodes and I manually sorted it based on counties and saved a copy with just the zipcodes for Marion County which geographically coincides with the city of Indianapolis. Next I used my list of zip codes to pull an even larger pool of data from the python database called uszipcodes (https://pypi.org/project/uszipcode/). This database included north, south, east and west latitude and longitude boundaries for each zip code as well as the central coordinates. It also offered median income info, population, square mileage, and many other demographic statistics. The main source of data on the actual food venues was Foursquare which allowed for searching based on each zipcode's rough boundaries and using their venue categories to pull relevant data for the project. The Foursquare info is very valuable and it is helpful that they make a lot of data available to the public through their developer program. This data provided a good general idea of what was available in my geographic areas of interest.

3. A. Methodology

Downloading the US Postal service zip code file gave me a starting point to gather a list of the zip codes for Indianapolis, which happens to be made up of basically all of Marion County. Next I used the python uszipcode database to create a dataframe with the details I needed for the analysis:

	zipcode	bounds_east	bounds_north	bounds_south	bounds_west	county	cities	housing_units	land_area_in_sqmi	Latitude	Longitude	median_home_value	median_household_income	occupied_housing_units	population	population_density
0	46201	-86.082971	39.791819	39.757465	-86.136878	Marion County	[Indianapolis]	16868.0	5.59	39.78	-86.11	68500.0	26391.0	11977.0	30962.0	5542.0
1	46202	-86.131776	39.800784	39.760050	-86.197006	Marion County	[Indianapolis]	9825.0	5.58	39.79	-86.15	168900.0	31698.0	8149.0	16335.0	2927.0
2	46203	-86.044969	39.763024	39.694590	-86.151173	Marion County	[Indianapolis]	17688.0	13.88	39.73	-86.10	75000.0	30481.0	14870.0	38950.0	2806.0
3	46204	-86.141726	39.784309	39.761219	-86.175804	Marion County	[Indianapolis]	3033.0	1.12	39.77	-86.16	264700.0	39826.0	2459.0	5125.0	4570.0
4	46205	-86.100831	39.848206	39.797326	-86.157192	Marion County	[Indianapolis]	14175.0	6.26	39.82	-86.12	118100.0	36429.0	11194.0	25356.0	4050.0
5	46208	-86.153078	39.871131	39.795494	-86.196808	Marion County	[Indianapolis, Rocky Ripple]	11993.0	6.71	39.83	-86.18	108300.0	31999.0	8879.0	22239.0	3312.0
6	46214	-86.266848	39.823436	39.760697	-86.310898	Marion County	[Indianapolis, Eagle Creek]	11881.0	7.11	39.79	-86.29	116700.0	41095.0	10824.0	24306.0	3420.0
7	46216	-85.989154	39.891035	39.844730	-86.039264	Marion County	[Indianapolis]	1065.0	3.49	39.87	-86.01	137900.0	44479.0	966.0	1697.0	486.0

With this dataframe we can create a folium map to get a look at what Indianapolis looks like and how our zipcodes are spread across the city for those who aren't familiar:



Now with my dataframe of Indy zip codes and the geographical coordinates listed therein it was possible to query the Foursquare database to gather a list of all the venues around town that fell into the selected categories within the square area between the zip code bounds. While this isn't a perfect representation of the zip code geographic areas it does give us a complete listing of venues and also an

idea of the availability for each zip code in the list:

```
(357, 6)
```

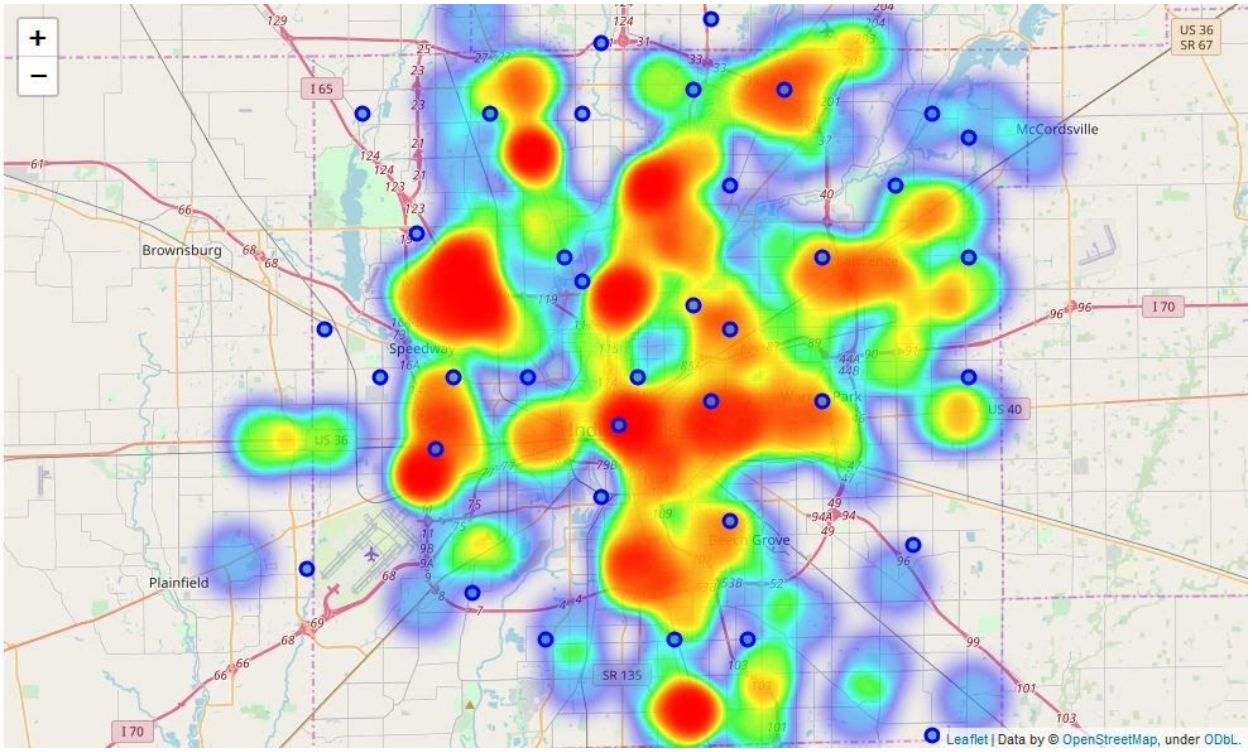
Out[21]:

	Zipcode	Southwest Coordinates	Northeast Coordinates	Venue	Venue Latitude	Venue Longitude
0	46201	39.757465,-86.136878	39.791819,-86.082971	Kroger	39.779	-86.094
1	46201	39.757465,-86.136878	39.791819,-86.082971	Kroger	39.758	-86.115
2	46201	39.757465,-86.136878	39.791819,-86.082971	Safeway	39.789	-86.084
3	46201	39.757465,-86.136878	39.791819,-86.082971	Pogue's Run Grocer	39.781	-86.116
4	46201	39.757465,-86.136878	39.791819,-86.082971	Save-A-Lot	39.768	-86.103
5	46201	39.757465,-86.136878	39.791819,-86.082971	Bienvenidos Supermercados	39.772	-86.107
7	46201	39.757465,-86.136878	39.791819,-86.082971	2 Amigos Mexican Grocery	39.761	-86.113
8	46201	39.757465,-86.136878	39.791819,-86.082971	Denise Scarbrough	39.778	-86.086

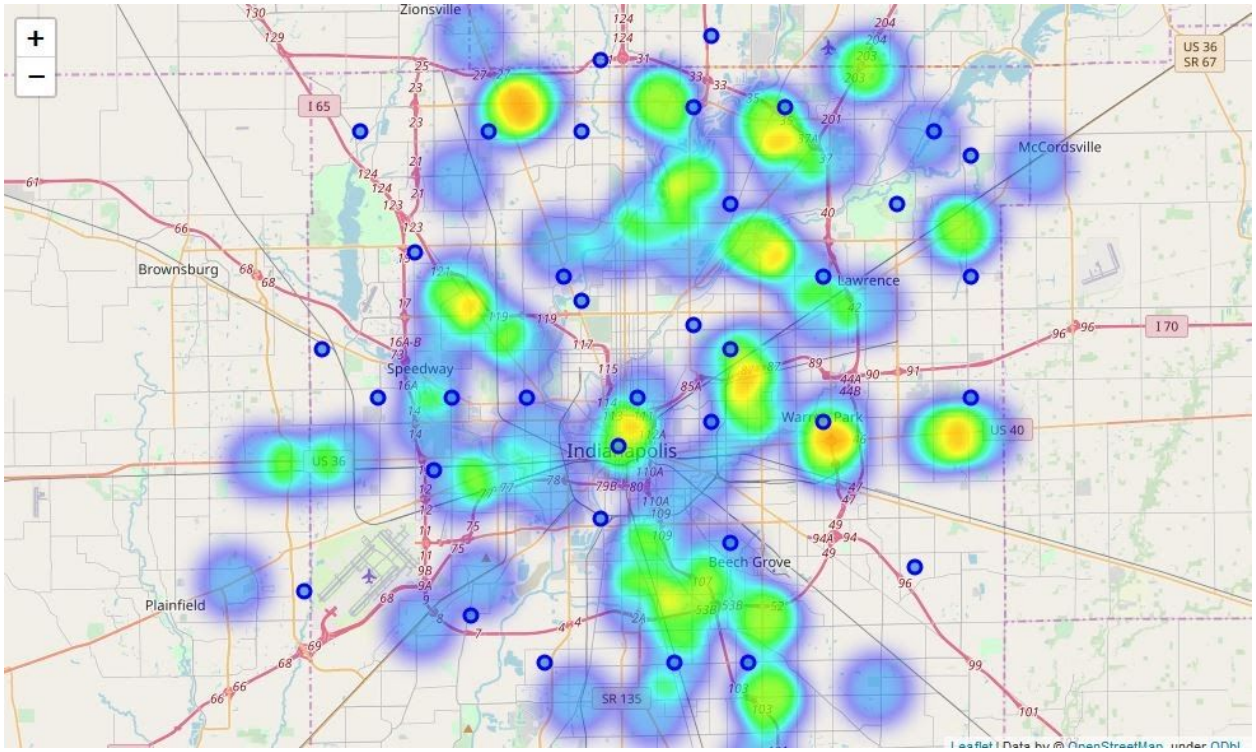
B. Sorting and Visualizing The Data

In order to generate the data frame above I did need to eliminate some duplicate entries that appeared with some of the larger chain stores that had additional listings for gas stations and pharmacies and the like. After these eliminations it brought the list down to 357 venues. There are many duplicates in this list because of the way Foursquare searches using geographical squares so this was an important consideration to keep in mind. In some parts of the following analysis I eliminate duplicates to get a better picture of the overall shape of venues in the city. In other cases where I was looking at the composition of individual zip codes it was important to keep the duplicates since there were many venues that were close enough to the zip code in question that they needed to be considered and not eliminated. Using folium we can generate a map to show all the venues in Indy that match for the Grocery Store and Supermarket categories:

Upon examination of the venue dataframe I could see right away that there was a broad range of venue types in Foursqaure that fit into the 2 categories I picked. I felt that the simplest way to distinguish the venues was to separate the ones that I knew to be the large chain grocery stores and big box stores. After separating this out I felt the list reflected a major demarcation line among the food stores in the city. This could be further granulated with future projects but for now it illustrates an important point about the availability and types of food across the different zip codes. The first map below shows the overall distribution of venues where the red areas have the highest concentration and the areas of no color the lowest:



And then when we look at just the large chains and big box stores:

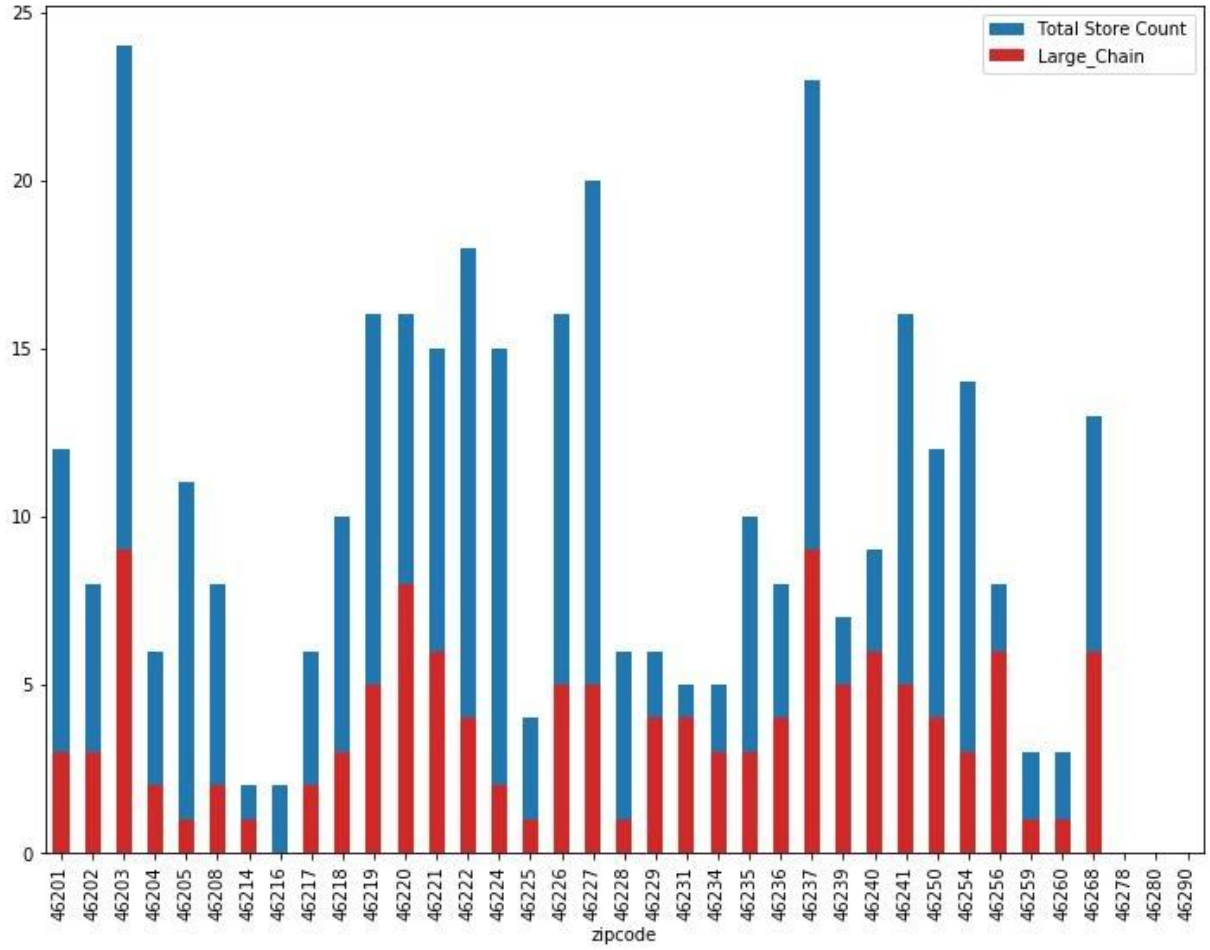


These images demonstrate the way that the smaller venues are more prevalent in some of the inner areas of the map where in the top map we see lots or red

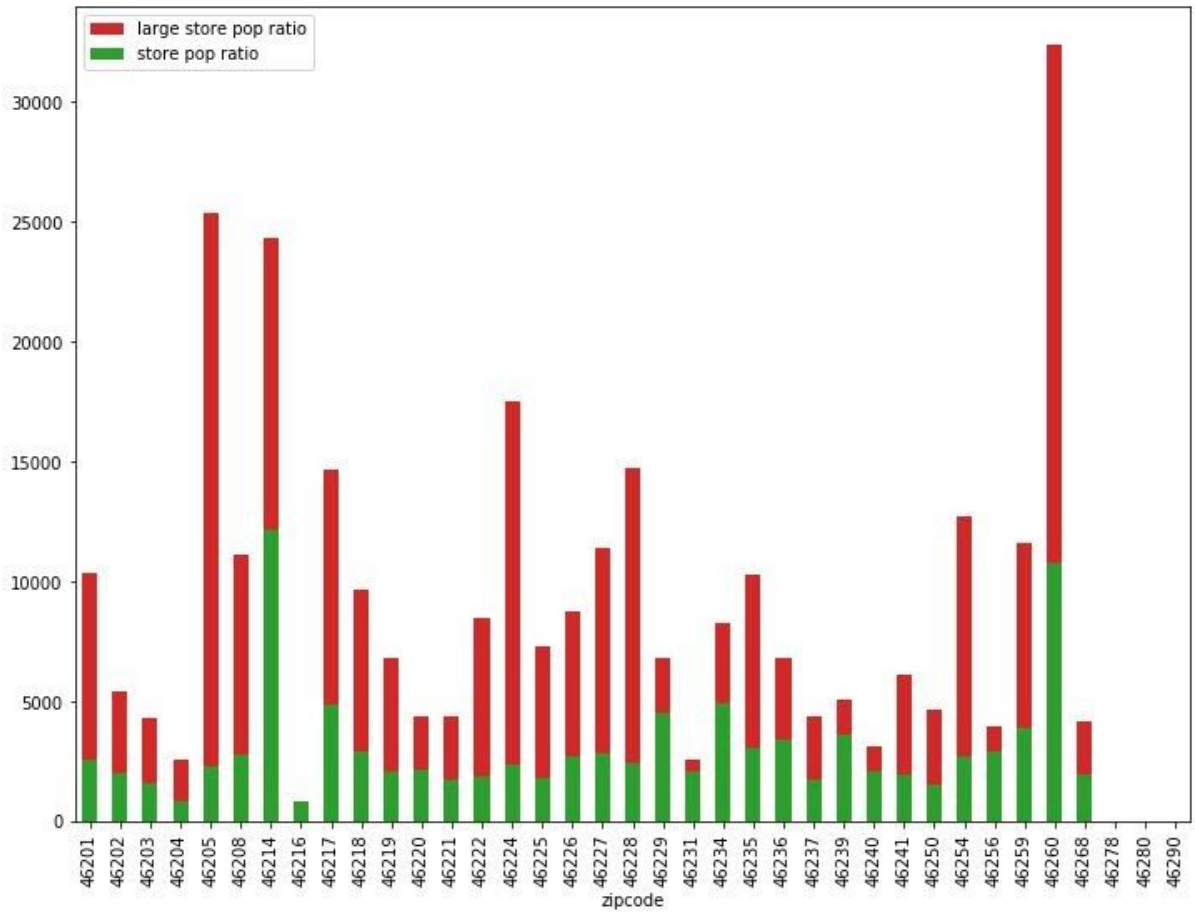
but in the bottom map we see mostly white space with some blue and green spots.

C. Looking at the Venue Info by the Numbers

After creating a category for large chains, I created a new dataframe with the counts for total stores and large chains and the breakdown as illustrated below shows a lot of variance between zip code areas:



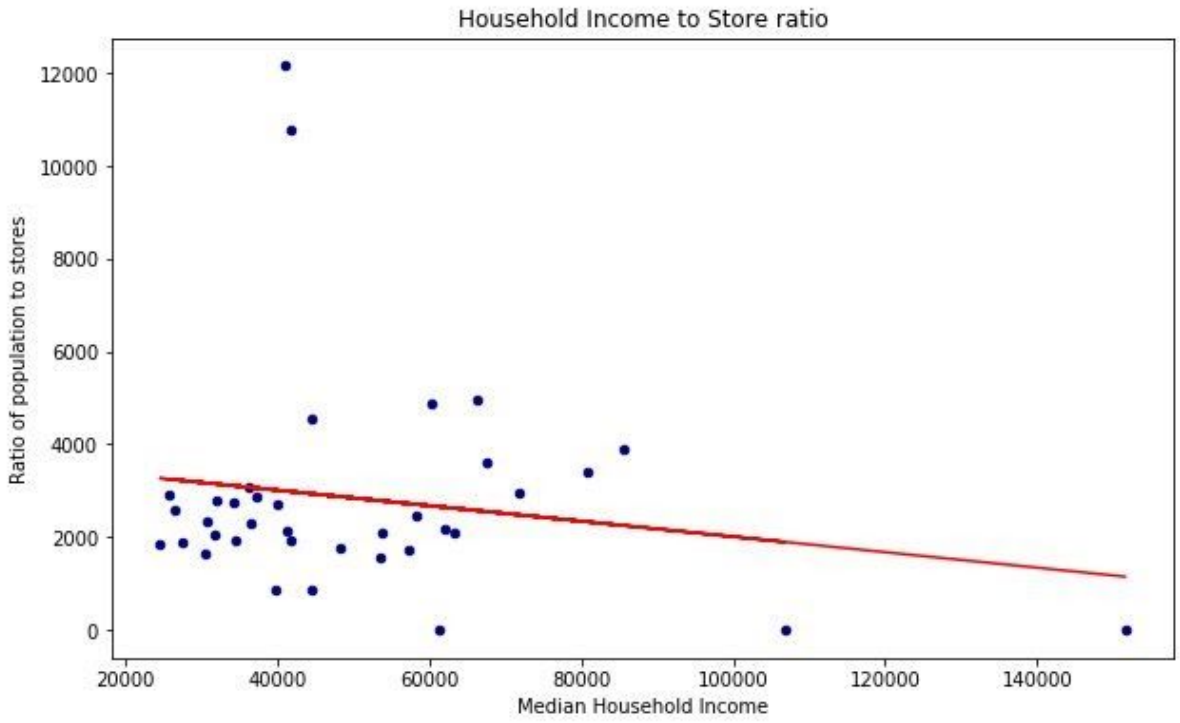
The next step was to look at the stores and large chains per square mile in the zip and then look at the proportion between the population density per square mile and the store density per square mile to see how well served each zip really was. The graph below shows the breakdown in terms of people per store for each square mile:



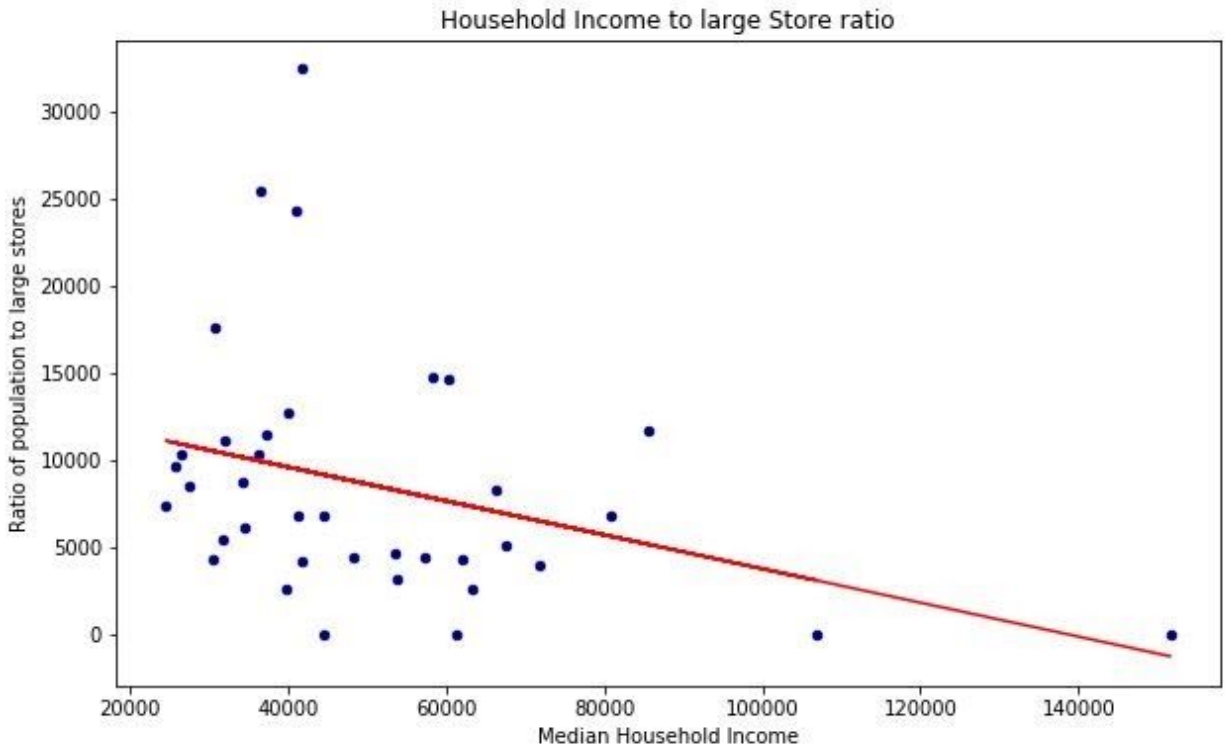
This graph goes a long way to highlighting the zip codes where we have a large population that is not being served well by the existing venue locations.

D. Visualizing Income to Store Availability Ratio

In looking at this data we can plot a regression line to demonstrate a clear relationship between the median income and the overall store availability:



In the above plot we see that the ratio of people for each store goes down as the median income goes up. When we look at just the large chain stores we see this relationship is even stronger:



K Means Clustering

The next step in the process was to use machine learning systems to try to shed more light and help our potential stakeholders to get more value out of the data. One way to do this is to use the K Means Clustering algorithm to try to categorize the zip codes based on the data we have collected. I tried clustering several subsets of the data to see what the system produced but ultimately the simplest way I found was to include median income, population and both total store to population ratio and large store to population ratio. I had the clustering done based on 5 categories and when viewing the dataframe sorted by median household income you can see that the algorithm sorted the zip codes roughly along those lines.

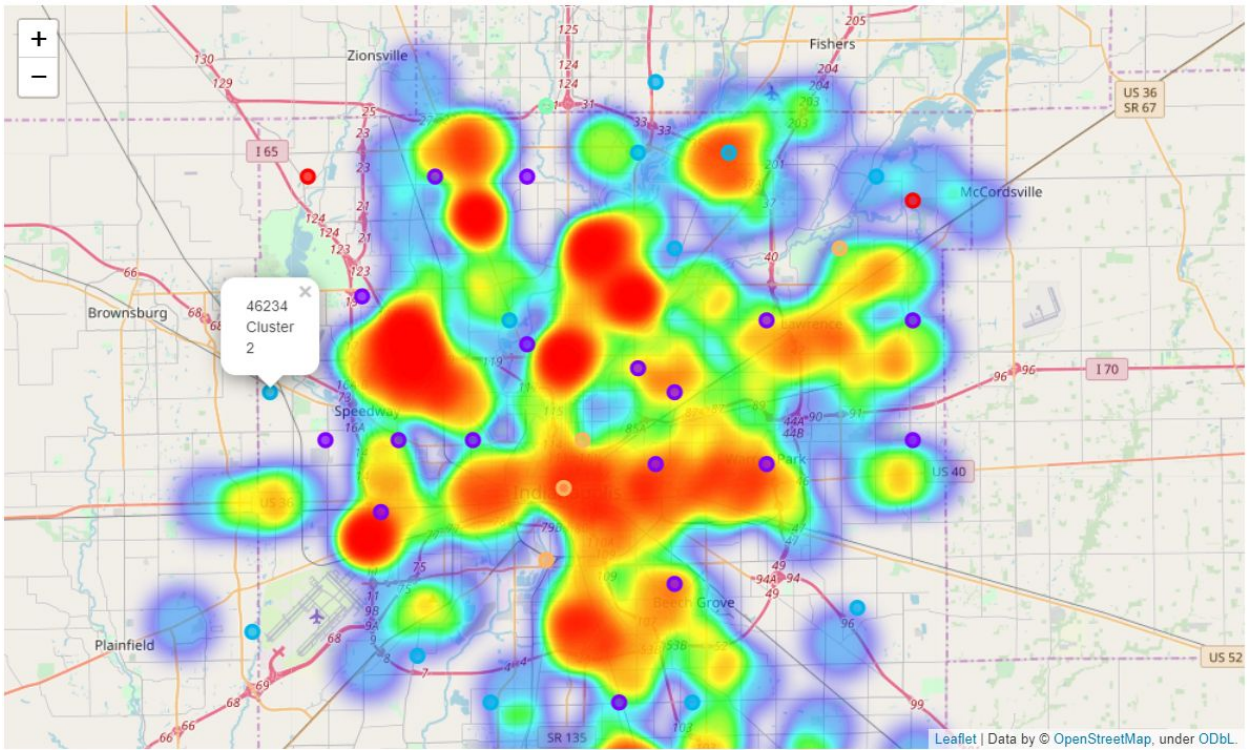
	Cluster Labels	zipcode	median_household_income	store pop ratio	large store pop ratio
15	4	46225	24548.0	1831.130	7324.520
9	1	46218	25666.0	2904.688	9682.293
0	1	46201	26391.0	2581.648	10326.593
13	1	46222	27581.0	1890.380	8506.710
2	1	46203	30481.0	1622.803	4327.476
14	1	46224	30660.0	2340.766	17555.745
1	4	46202	31698.0	2041.582	5444.220
5	1	46208	31999.0	2777.940	11111.760
16	1	46226	34368.0	2743.304	8778.572
27	1	46241	34600.0	1916.600	6133.120
22	1	46235	36213.0	3084.140	10280.467
4	1	46205	36429.0	2304.818	25353.000
17	1	46227	37380.0	2853.362	11413.450
3	4	46204	39826.0	853.067	2559.200
29	1	46254	40046.0	2721.267	12699.247
6	1	46214	41095.0	12158.100	24316.200
10	1	46219	41332.0	2121.244	6787.980
33	1	46268	41718.0	1920.916	4161.985
32	1	46260	41826.0	10796.693	32390.080
7	4	46216	44479.0	848.070	0.000
19	1	46229	44626.0	4544.910	6817.365
12	2	46221	48294.0	1759.128	4397.820
28	2	46250	53568.0	1545.464	4636.392
26	2	46240	53699.0	2080.967	3121.450
24	2	46237	57266.0	1722.468	4401.862
18	2	46228	58208.0	2460.405	14762.430
8	2	46217	60254.0	4891.582	14674.745

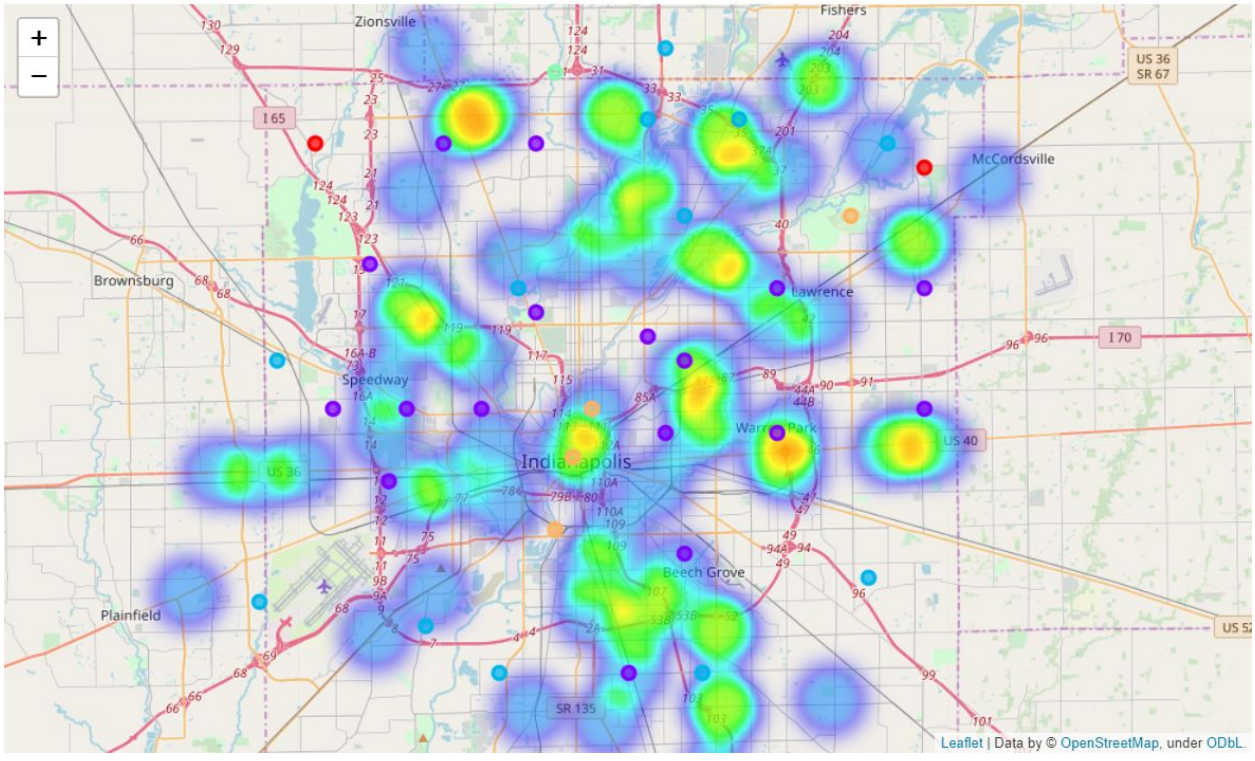
The clustering helps to define the problem depending on who is looking and what population they are concerned with. When we look at this through the lens of food access one would be concerned with addressing the cluster 1 and 4 areas

where there is a high number in the large store population ratio category, these are definitely our most at risk areas where transportation is likely to be an issue and quality food sources are likely to be low. On the other hand this clustering may help businesses that are looking to open in town. An example would be a local organic grocers looking for a higher income demographic where there may be a low population to large store number but a higher overall ratio number showing that the area may not have a lot of smaller boutique and local offerings. In this case they would probably be interested in cluster 2 with numbers like the 46240 numbers where there may be good access to large chain stores but there may be room for the type of store they operate. The advantage with using K means clustering is that it can update as the situation in the city changes over time and keep these interest groups informed about where to focus their efforts.

E. Visualizing the Clusters

The map below shows our cluster groups combined with our overall store location heat map. This offers an even more granular illustration, in our example above, our local organic grocer can see exactly where there are fewer stores currently to pinpoint the ideal location. The second map below is our large store map again with the clusters included and this shows how different the situation becomes for our cluster 1 areas (purple dots) where we can now see how few of the larger stores are available and how the availability of fresh food may be much lower than what the overall store level at first seems to indicate.


Zionsville
Fishers
US 36
SR 67
I 65
McCordsville
46234
Cluster
2
Brownsburg
I 70
Speedway
US 40
Plainfield
US 52
SR 135
Leaflet | Data by © OpenStreetMap, under ODbL.


Zionsville
Fishers
US 36
SR 67
I 65
McCordsville
Brownsburg
Lawrence
I 70
Speedway
US 40
Indianapolis
Beech Grove
Plainfield
SR 135
Leaflet | Data by © OpenStreetMap, under ODbL.

4. **Results**

   As the above data indicates the question of availability of food is a complicated one to consider. The food that people consume has an enormous impact on health and quality of life issues as well as a large economic impact on the community so despite the complexities it is a worthwhile venture to try to understand the issue better. Upon the initial gathering of the store data we saw that there are clearly a great many outlets in the city but scanning through that dataframe anyone who has some familiarity with Indianapolis is immediately aware of the variety of types of stores in the list. While there may be more rigorous ways in a future study to separate the venues, I feel that the stores I seperated into the large chain category reflect the main distinction and also the main concern shared by many in the community that the economic realities of running a large grocery store are causing areas of the community to be left behind. I don't want to pick on any one particular venue but many of the venues that are not in the large chain list are more convenience style stores that specialize in snack foods and processed foods that tend to be less healthy. The data presented here clearly indicate that on a zip code to zip code level in Indy there are clearly very different access levels and that they correlate to economic factors.

5. **Discussion**

   Based on this project I have recommendations for both those who are hoping to alleviate food access problems in the city and for those who are looking for new business opportunities in the community.

   A. **Food Access-** The areas of focus for potential government or civil service organizations looking to help with the issue of food access would be to focus on the cluster 1 purple locations above. These areas fall in the bottom half of the income scale and tend to have more limited access to the large stores. These are also going to tend to be people with more limited access to cars, which in the city of Indianapolis is a significant

problem because of the relative lack of public transit. Obviously access to transportation is outside the scope of this project but would certainly be an excellent area for a follow up study to try to pinpoint further the communities in need of the most aid in this area.

B. **Business Opportunities-** The data above offers insights into economic layout of the city as well as access levels to grocery stores. For stores that are looking to serve a higher median income area they would be well served to look at zip codes in cluster 2 areas. This may be especially true for more specialized stores like high end butcher shops or organic local food outlets since we can see that much of this cluster is well served by the larger big box stores. Another opportunity would be for grocery store delivery providers who have been growing in popularity in recent years. Looking at the areas in both cluster 2 and cluster 1 could help in guiding marketing efforts. Ideally there may be a collaborative opportunity for the government to help offset any additional delivery costs with these organizations in order to help ensure access to those at risk cluster 1 communities.

6. Conclusion

While there is certainly much more to be learned about the nature of food availability in this and other cities I feel that this project has helped to gain an understanding of the nature and variability of access to this basic necessity. As is so often the case it seems that in trying to answer one question I've given rise, at least in my own thoughts to several other questions I'd like to answer. The follow up studies I'd most like to see would be connecting this data to health data, transportation data, and more data about the exact food offerings at each venue. The great thing about this project is that anyone can simply re-run the code and come up with up to date results as the situation changes over months and years. Also it's important to note that all the data in this project is available for free to the public so potentially anyone who wants to apply this data to other US cities can do so as long as they have access to a computer and can create a

free Foursquare developer account. This is a project that can easily be refined and improved with better future data as well as being applied to offer comparative analysis of other US cities.