A Food Profile for High School Hangouts, Rochester, NY

Karen Lacomis-Cote

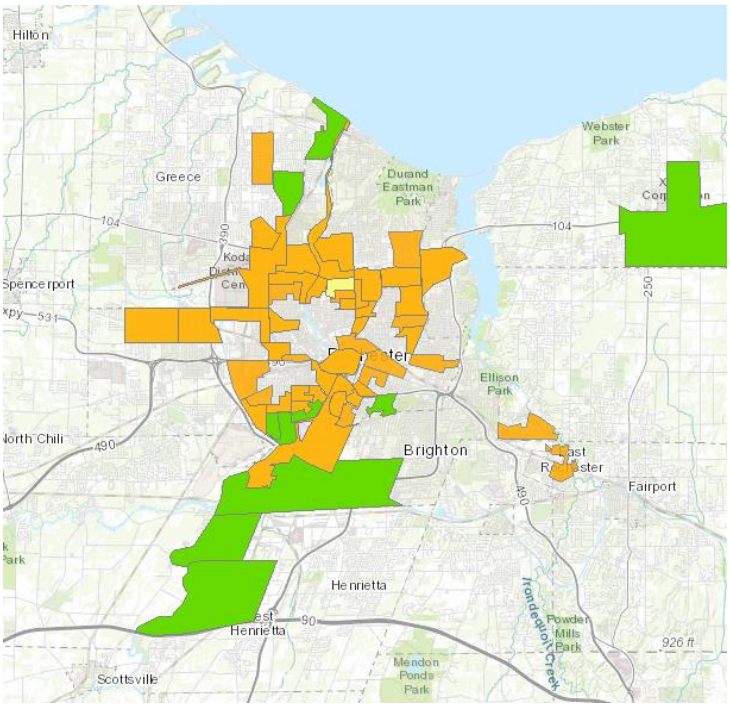
May 20, 2020

1. Introduction

1.1 Background

Food deserts are a well-researched problem in public health. The availability of grocery stores with good food choices within a mile (urban) or 10 miles (rural) of a person's home have a significant impact on health and the ability to eat nutritious meals. Nearly 2.3 million people in the United States live in a food desert.

In cities, lack of grocery stores may require heroic efforts to reach the nearest one, particularly if there is a lack of public transportation. Even in suburbs and more rural areas, the nearest market may be further from one's home than is easily accessible [1]. Rochester, NY, an average sized area in upstate New York, has a large group of food deserts identified by the USDA and noted in green (1 mile access issues) and orange (1/2 mile access issues) [2]:



Easy access to nutritious food is not only limited to households, but also members within those households – creating a different dimension to the food desert problem.

For purposes of this project, the food desert idea will be directed at a smaller subgroup within the population - high school students. There are more than 40 high schools in Monroe County, all with different student and neighborhood profiles. Schools often allow students to leave campus for their lunch period, particularly seniors, who may not have full schedules. Students also look for food after school or after practice, and breakfast options to supplement skipped meals at home. The choices nearby influence their diet - and if the only choices are high-fat, high-sugar fast foods, that diet may be subject to less than ideal inputs. In a high poverty area, quick choices may also replace nutritious home cooked meals if the area is also a food desert. So all the pieces tie together when looking at diet overall.

1.2 Problem

Having data to understand nearly food choices is key to understanding diet. By mapping high school locations, along with food options within a mile radius, this project will analyze the nearby choices for each high school. By clustering the similar option areas together, we can highlight areas where there are more and fewer choices, and more and less nutritious ones as well.

1.3 Audience

The outcome of this analysis could be used in two ways - to help public health and primary care practitioners (pediatricians, primarily) understand what their patients may be eating, and how those available options might translate into their patient population issues like allergies, asthma, and obesity. The analysis could help inform this group how to target interventions for their patients - suggestions for where and what to eat, as an example, to result in better health, and better school and athletic performance.

The other use of this data is business related - finding the types of food that is already available can help define possible new options to fill an open niche. Together, the two uses create a powerful partnership that could bring options into neighborhoods and improve overall health.

2. Data Acquisition and Cleaning

2.1 Data Sources

Information on high schools and their locations can be obtained from Foursquare using API calls to browse the area for venues that are identified as high schools. Food venues around each school can then be retrieved using a separate Foursquare API call.

Poverty levels surrounding each of the high schools can be obtained from the US Census Bureau website. Boundary data for each of the zip code areas in Rochester is available from GitHubOpenDataDE, which supplies files for each state [3, 4].

2.2 Data Cleaning

High school information was sourced from Foursquare, but included some entries that were not actual schools, but rooms within them. Since the rooms were tagged as high schools, they had to be removed from the data set so that there were not duplicate entries representing the same school. Some of the schools in the high school list were K-12, so not technically only high schools. I decided to leave those in, since their high school aged students likely still exhibited the same behaviors as any other high schooler.

Since the school location data was retrieved using a bounding box, some parts of the county were not covered – the choice was to include pieces of other counties, since Monroe County is an odd shape, or to pull the box in to cover the bedroom communities and thereby leave off some of the more outlying towns within the county boundaries. For a first pass, I decided on the bedroom community approach, and left out some of the outlying areas.

Venues local to the high schools were also sourced from Foursquare. Adding these entries pointed out one school as an outlier with no nearby venues. The school itself was very different from other schools due to its location in an affluent community, with most of the students having their own transportation available. Expanding the search area to 2 miles would have pulled in multiple venues, but would have been too broad an area for the other schools. As a result, this school was treated as an outlier and removed from the data set.

There were also redundant venues in the nearby venue list, due to some of the high schools being in close physical proximity to each other. Overall, there were 1021 food venues total (479 unique venues). These were left in place, to ensure a complete list for each high school, but unique counts of venues and venue categories were taken.

The zip code files from GitHubOpenDataDE were convenient sources of boundary data, but the file contained all the zip codes from NY state, and had to be filtered down to the area of interest in Rochester. Once at that level, I discovered that there were some zip codes missing, and was therefore unable to generate a complete map. However, the necessary zip code boundaries were available to cover the high school zip codes. The final generated map has a few gaps, but they do not overly impact the analysis.

2.3 Feature selection

There were many redundant and unnecessary data points in the location set. Information about city, state and country were not necessary and were removed. Specific street address, and a completely formatted address were also redundant (and unnecessary for this analysis) and removed. School Name (all unique and therefore a primary key to the dataset), latitude, longitude and zip code were retained.

For venues, only name, category, latitude and longitude were retained.

Zip code and poverty percentage was held in a separate dataframe for creating a chloropleth map.

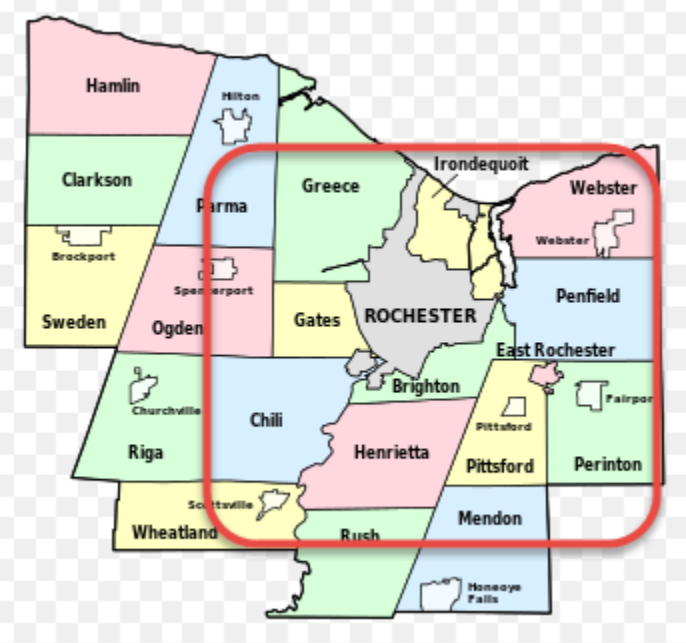
3. Methodology

3.1 Tools Used

Data manipulation and visualization for this project was done using IBM Watson Studio and Skills Network Labs, Jupyter Notebooks and Python. All the initial data for this project was retrieved from Foursquare. Using the search function in a defined area resulted in a list of the local high schools ("High School" is a category within the Foursquare API calls). Maps were created using Folium.

3.2 Area Defined

The defined area covers a portion of Monroe County that encompasses the bedroom communities as shown:



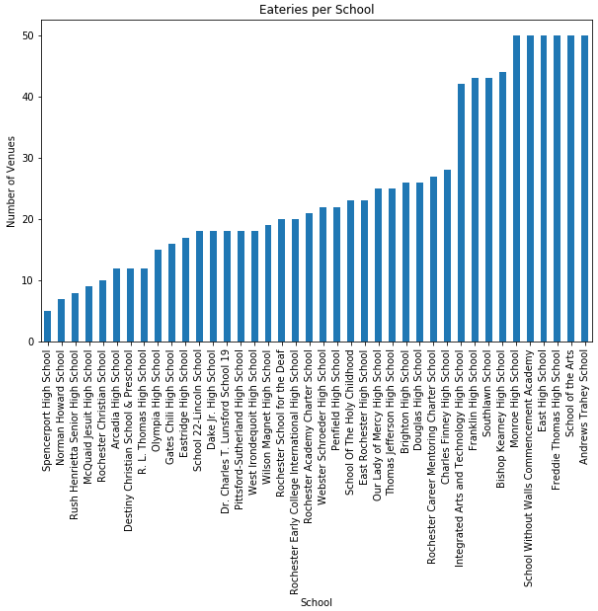
While the county is larger than the boxed in area shown in red, expanding the area would cause areas (and high schools) outside the county to be included, as Monroe County is oddly shaped. For purposes of this analysis, it was sufficient to use the boxed in area, which is the seat of the majority of the county population. A small slice of an adjacent county (Ontario) is included, but that small piece does not contain any high schools. If one were to be included, it could be removed simply by looking for the specific address and zip code, which would fall outside the county line.

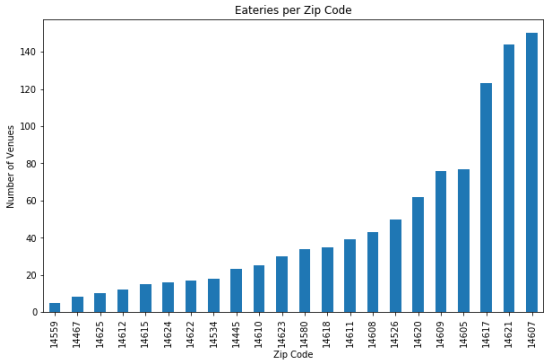
3.3 Distribution of venues

A dataframe of the high schools was created, including latitude and longitude values, which were then used to find nearby venues using the Foursquare API.

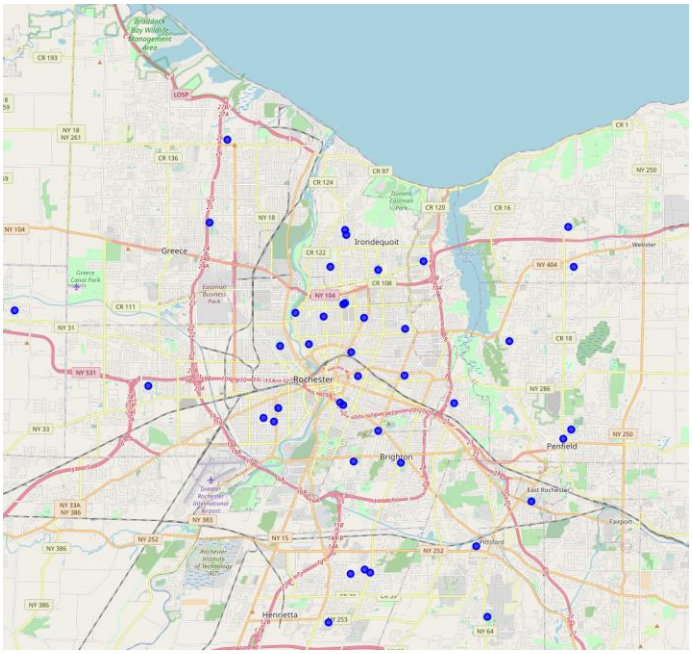
There were not many variables to compare, but some scatter plots were generated to map zip code against venue type – but no relationship was seen.

When counting venues, however, either by school or by zip code, the majority of them were in the city. In fact, many of the city locations topped out at the upper limit of 50 venues, while the suburban schools had a maximum venue count of 30. The reason behind this is not clear and warrants further study, though one might assume that the closer proximity of streets and schools within the city boundaries versus the longer distances in the suburbs accounts for much of this difference.





Using the folium library, a map of the area was created, using the latitude and longitude of the individual high schools to superimpose markers for their locations:



To create clusters of similar schools, first each type of venue was coded for each school – with a 1 if such a venue existed, and 0 if not. The resulting dataframe had 58 columns – one for the school itself, and one for each unique type of venue. Next, the venues were summarized and normalized for each school, and a mean was calculated for the frequency of each category's occurrence for that particular school.

	School	American Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Brazilian Restaurant	Breakfast Spot	Burger Joint	...	Southern / Soul Food Restaurant	Spanish Restaurant	Steakhouse	Sushi Restaurant	Szechuan Restaurant	Taco Place	Thai Restaurant	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Wings Joint
0	Andrews Trahey School	0.040000	0.060000	0.02	0.040000	0.080000	0.0	0.0	0.040000	0.020000	...	0.0	0.000000	0.040000	0.020000	0.02	0.0	0.000000	0.0	0.020000	0.0
1	Arcadia High School	0.000000	0.000000	0.00	0.000000	0.000000	0.0	0.0	0.000000	0.083333	...	0.0	0.000000	0.000000	0.000000	0.00	0.0	0.000000	0.0	0.000000	0.0
2	Bishop Kearney High School	0.022727	0.022727	0.00	0.000000	0.090909	0.0	0.0	0.022727	0.022727	...	0.0	0.022727	0.000000	0.000000	0.00	0.0	0.022727	0.0	0.022727	0.0
3	Brighton High School	0.000000	0.000000	0.00	0.038462	0.153846	0.0	0.0	0.000000	0.076923	...	0.0	0.000000	0.038462	0.000000	0.00	0.0	0.000000	0.0	0.000000	0.0
4	Charles Finney High School	0.107143	0.000000	0.00	0.035714	0.000000	0.0	0.0	0.000000	0.071429	...	0.0	0.000000	0.000000	0.035714	0.00	0.0	0.035714	0.0	0.000000	0.0

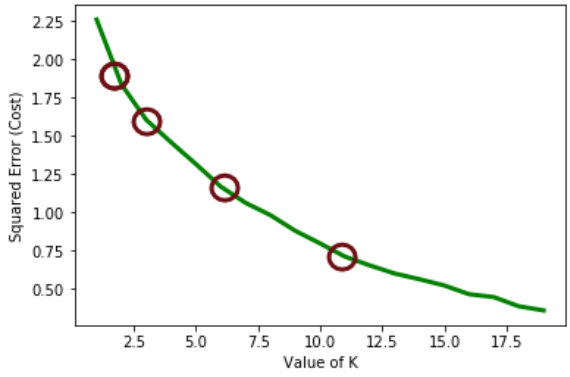
Using this information, two separate steps were performed. First, the top 10 venue types for each school were collected and enumerated in a separate dataframe. This gives a visual of common venues, and was used to compare the schools within clusters once calculated.

	School	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Andrews Trahey School	Bakery	Pizza Place	Asian Restaurant	Diner	American Restaurant	Café	Mediterranean Restaurant	Mexican Restaurant	Noodle House	Sandwich Place
1	Arcadia High School	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Donut Shop	Burger Joint	Diner	Sandwich Place	Café	Food Court	Food
2	Bishop Kearney High School	Fast Food Restaurant	Bakery	Diner	Pizza Place	Sandwich Place	Café	Fried Chicken Joint	Donut Shop	Deli / Bodega	Italian Restaurant
3	Brighton High School	Bakery	Pizza Place	Chinese Restaurant	Food	Burger Joint	Sandwich Place	Salad Place	Deli / Bodega	Mexican Restaurant	Middle Eastern Restaurant
4	Charles Finney High School	Pizza Place	American Restaurant	Italian Restaurant	Burger Joint	Chinese Restaurant	Mexican Restaurant	Café	Thai Restaurant	French Restaurant	Sandwich Place

Second, the dataframe containing the mean values was used to partition schools into distinct groups using unsupervised learning and K-means clustering.

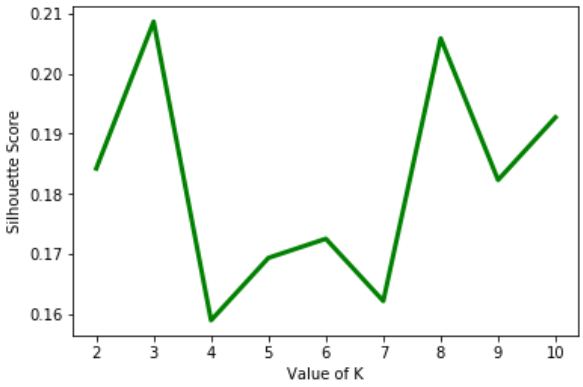
To establish the best value for the number of clusters, the maximum number of clusters was set to 20 when a first run with a max of 10 did not produce an elbow.

However, even with the larger cluster counts, the results were still inconclusive. The graph had no clear elbow – there appeared to be one at 3 clusters and another at 4, a third around 6, and so on. The behavior of the IBM Watson Studio and Skills Learning Labs were also inconsistent. In most cases, this is the graph that was produced, but without warning, some random variations would appear, sometimes with elbows, sometimes without. As the result shown below was the most common result, it was chosen as the representative graph for the exercise.



Use of a different indicator was needed, so the Silhouette score was used. The highest value for the score indicates the best value for K. This time, there was a clear peak for a cluster count of 3. (the score was fairly low at 0.22 (1 is the best), but it was still the best option). A cluster count of 8 was

the next best choice. Again, random variations in this score were seen, so the most commonly resulting graph, shown here, was used to determine K as 3.



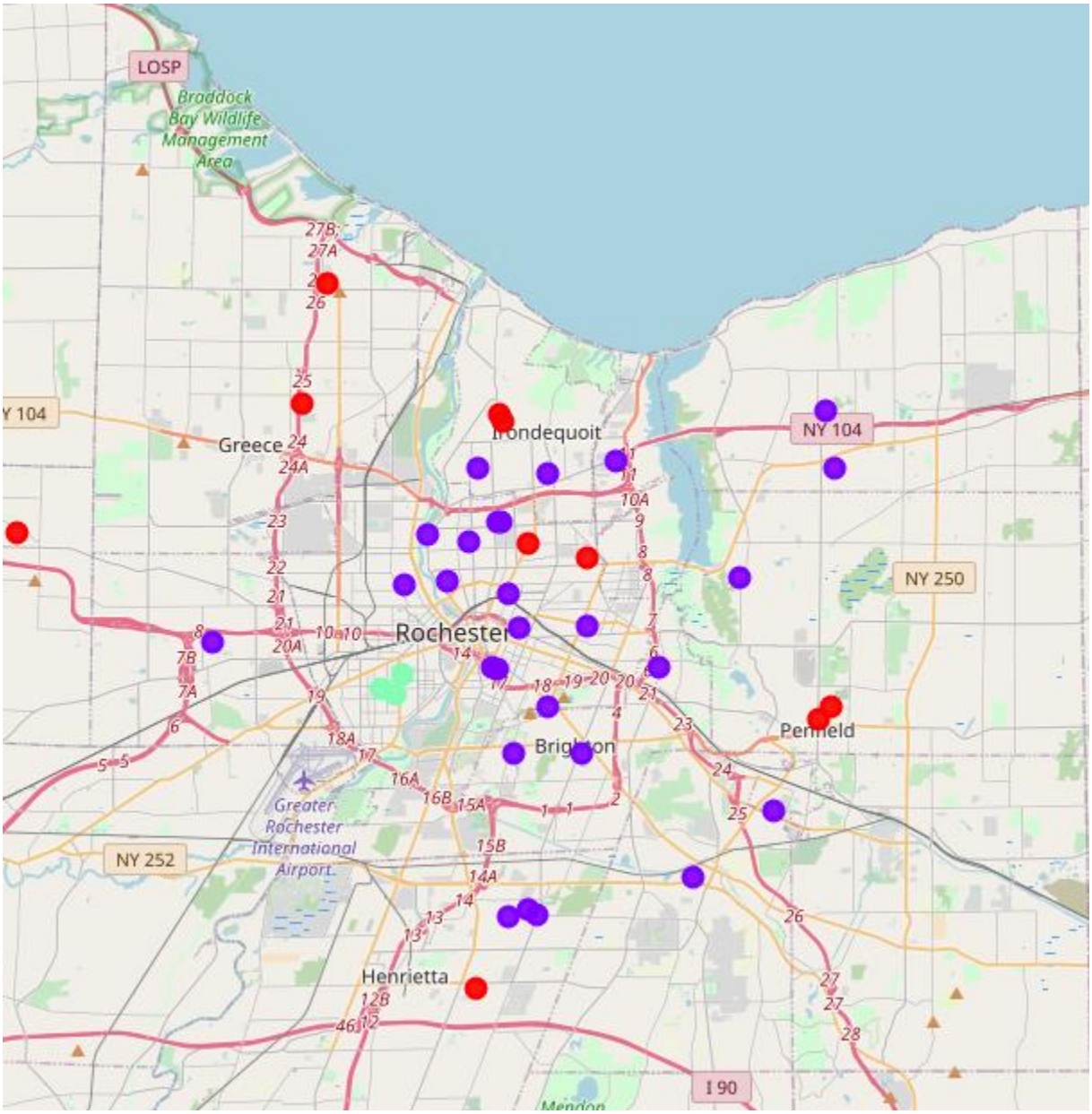
After calculation with a cluster count of 3, the calculated cluster labels are added to the existing data, placing each school into a cluster with others having similar food venue characteristics:

	Cluster Labels	School	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	1	Andrews Trahey School	Bakery	Pizza Place	Asian Restaurant	Diner
1	0	Arcadia High School	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Donut Shop
2	1	Bishop Kearney High School	Fast Food Restaurant	Bakery	Diner	Pizza Place
3	1	Brighton High School	Bakery	Pizza Place	Chinese Restaurant	Food
4	0	Charles Finney High School	Pizza Place	American Restaurant	Italian Restaurant	Burger Joint

The clusters were listed, mapped and graphed for analysis.

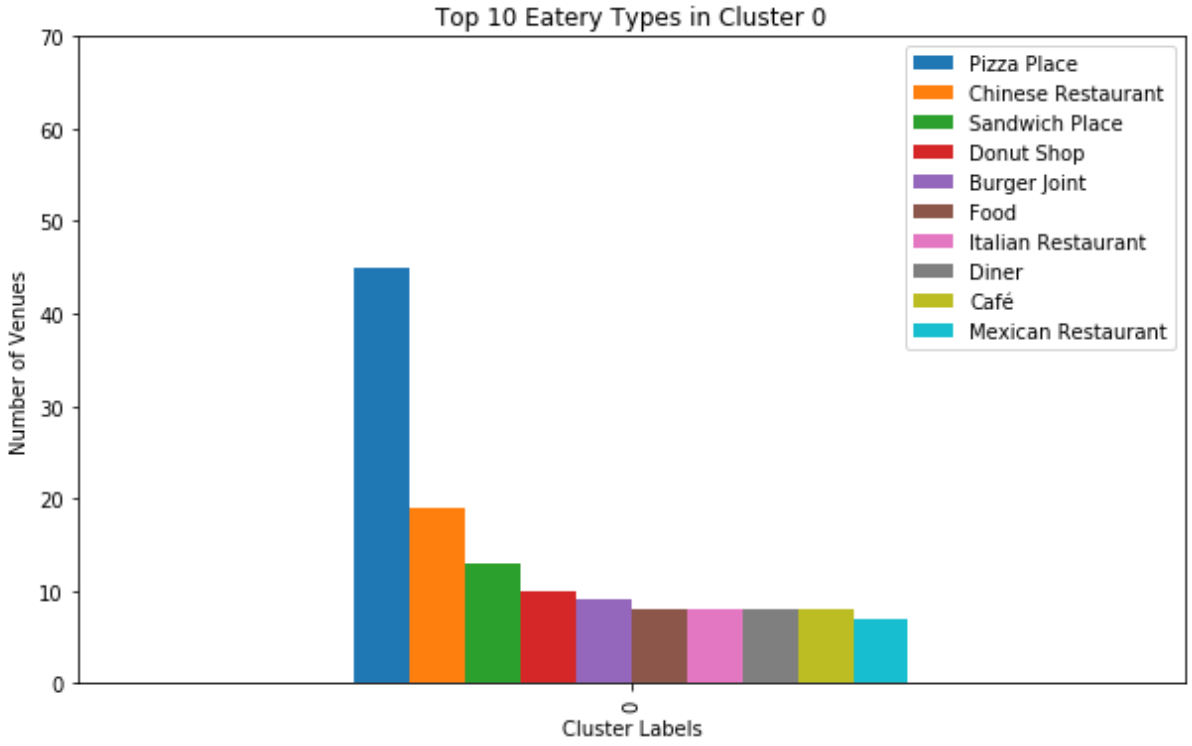
4. Results

Using the merged and clustered data set, a new map is created, color coded for each cluster:

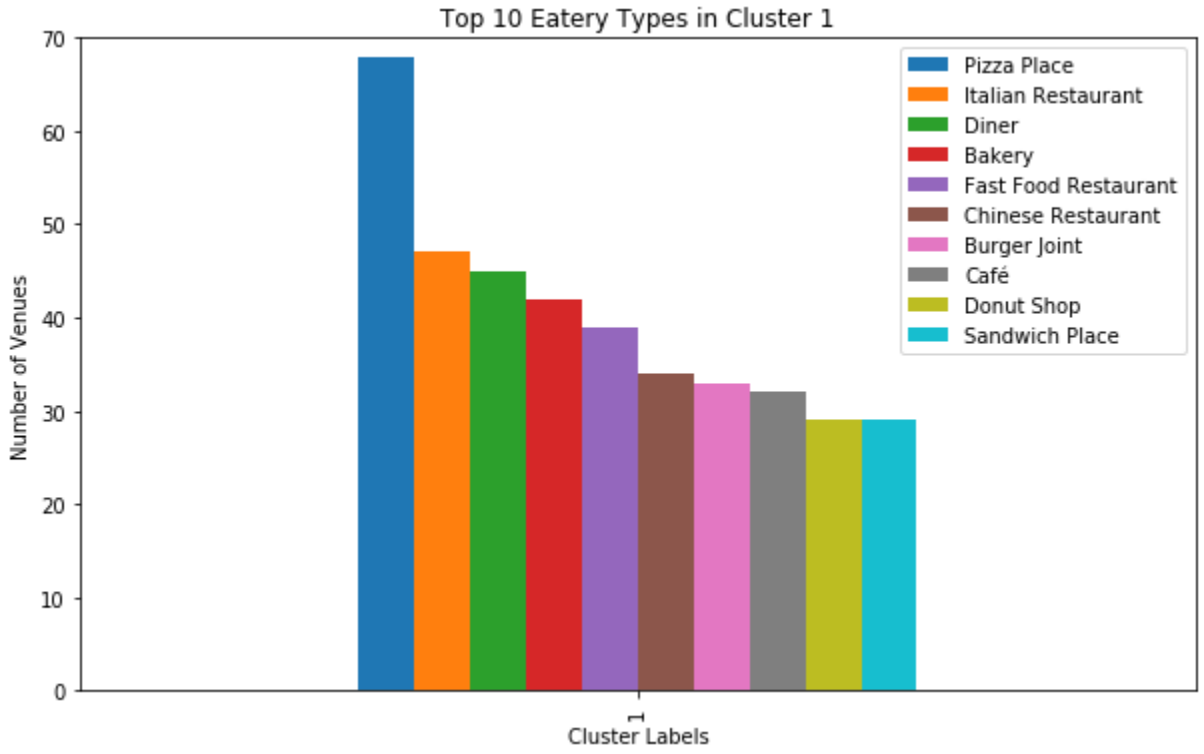


Looking at the clusters, we can plot the top ten venue types for each cluster. More than the top 10 would generate noise, so limiting to the top ten is a better choice.

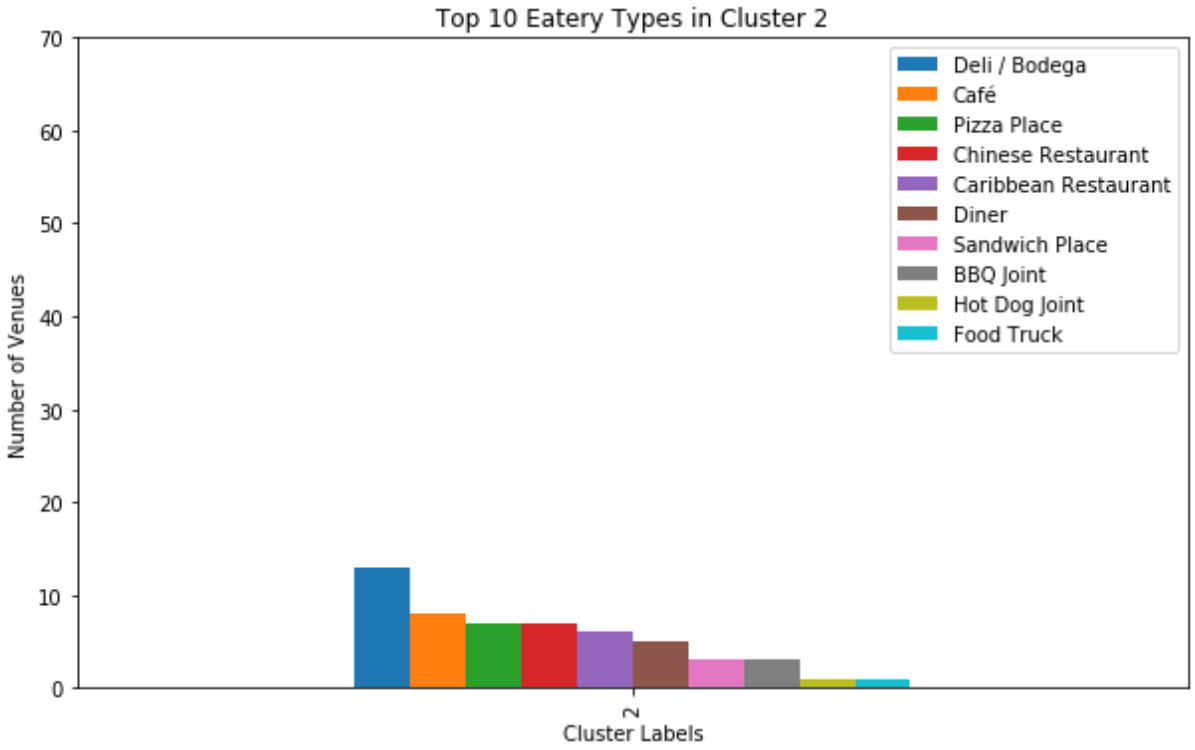
In Cluster 0 (Red markers) pizza places dominate the market, more than double the next closest option.



Cluster 1 (Purple markers) is heavily populated by pizza places, but has a varied landscape, including diners, bakeries, fast food and burger joints and donut shops. Italian and Chinese restaurants are also popular, but are still outnumbered by “quick stop” venues.



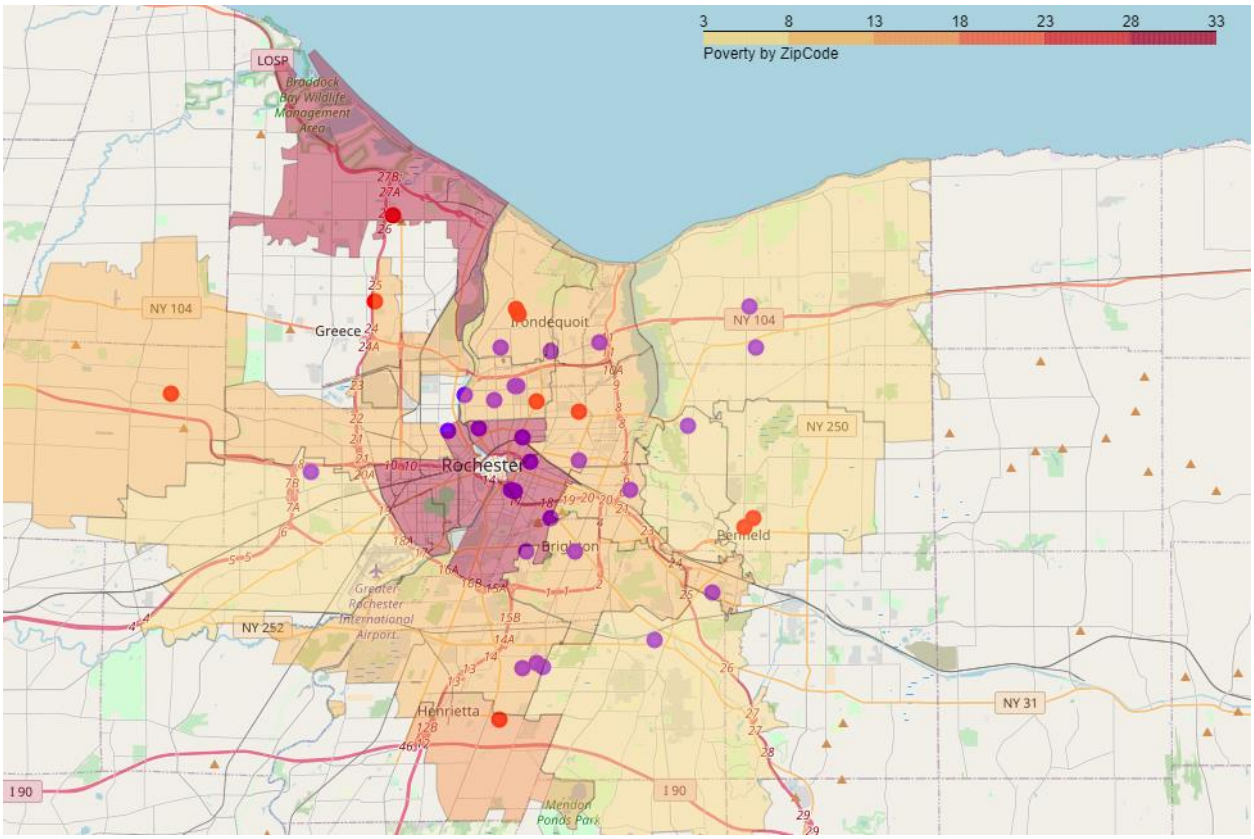
The final cluster, Cluster 2 (Aqua markers) are the venue deserts, with much lower market penetration in any area. The predominant venues are bodegas, which are likely to carry convenience foods, and café's, which may be healthier. The schools in this cluster are within the city, clustered together geographically, and apparently in an area where there is very little on offer.



The clusters were named as follows, based on this information:

Cluster 0: Pizza Party
Cluster 1: Fast Food Heaven
Cluster 2: Corner Store Food Desert

One final step was taken to map the percentage of households in poverty for the zip codes that contain the high schools. The resulting chloropleth map shows the clusters overlaid with colored filters to show impoverished areas.



Discussion

Regardless of cluster, the data clearly shows that the most frequently available choices near each school are not healthy ones! However, it is likely that students in Henrietta would head for pizza, while those in Webster have many more poor options to choose from. This distribution of food venues could lead to further investigation into obesity rates in the different cluster areas to see if there was some correlation between available choices and BMI. Many paths for investigation spring to mind simply by looking at these clusters.

In the broader sense, there is more variety in the purple cluster – but those schools are also closer to major routes, which could influence the choices available, since there is also a market here for travelers on those routes looked for a quick bite. The green clusters (primarily pizza) are in more out of the way locations, so are perhaps catering to their teenage clientele.

Using the choropleth map, there appears to be a tendency for the most variety to be in areas with lowest poverty levels. There is some incursion into the inner city area, which holds downtown office space - another source of lunchtime patrons who would demand (and support) such variety.

As mentioned previously, Monroe County is an odd shape. With a bit more effort, the range of high schools could be expanded to include the entire county, which may create more or different clustering. The bounding box could also be made smaller to focus in on a specific area. It may also be helpful to determine how many students had their own transportation, as that would impact the radius they were able to access in a limited timeframe.

The dearth of options within the city is also an interesting avenue to explore. Cluster 0 points to the southeast side of the city, focusing more on a specific area for investigation than any of the other clusters. There are also specific zip codes within Rochester which have a higher prevalence of poverty and food deserts – the zip code chloropleth helps delineate these areas more clearly.

Conclusion

As the data shows, most high school students have access to a number of choices when looking for a quick lunch or snack near to their school. However, in the majority of cases, those choices represent a less than desirable menu of pizza and fast food. While there are some healthier choices in the "Fast Food Heaven" of cluster 1, it would be an interesting project to see how often the healthier paths are taken over convenience. As an intervention to improve diet, a survey of the different venues in cluster one could be undertaken to create a list of healthier options to offer students.

Cluster 2, with its predominance of pizza places, presents business opportunities to fill some of the open niches. And cluster 0, the high-school snacking desert, could perhaps benefit from some mobile options like food trucks and carts, which could offer healthy choices at prime times of the day.

The result of this analysis provides a wealth of information that can be used in various ways to improve diet, health and access for the high school students and their community.

G. References:

- [1] Food Deserts - Food Empowerment Project
- [2] USDA – Food Access Research Atlas
- [3] Foursquare API
- [4] Github OpenDataDE