Introduction

Urban food deserts are areas in large cities where it is difficult, time consuming, expensive, or unsafe for low income households to purchase healthy food. Food deserts typically have a low density of grocery stores or supermarkets, a high density of fast food restaurants, convenience stores, or corner stores, and poor access to public transportation. Using the Foursquare location data and machine learning, neighbourhoods in New York can be evaluated based on the accessibility of healthy food compared to that of unhealthy food. This data would be useful to urban farming/gardening initiatives and potential business owners looking to market affordable healthy food to urban communities.

Data

The Foursquare API will be used to find the amount of grocery stores and supermarkets there are compared to restaurants, fast food, and convenience stores. The geographic data for New York and its neighbourhoods is available for free courtesy of New York University (https://geo.nyu.edu/catalog/nyu_2451_34572). The machine learning algorithm will be an unsupervised classification algorithm because the definition of a food desert is flexible and there is no directory of which NYC neighbourhoods are food deserts.

Methodology

A data frame of every business within a mile radius of each New York neighbourhood in the Bronx and Brooklyn. The businesses were narrowed down to those that can be viewed as a reliable and affordable food source. Artisan shops and restaurants other than fast food restaurants because although they can be used for an occasional meal or food item, low income

families cannot source meals from these places on a day to day basis. Three additional columns were added to the data frame that summed the healthy, unhealthy, and all options respectively.

	Neighborhood	Butcher	Convenience Store	Farmers Market	Fast Food Restaurant	Fish Market	Food Court	Food Truck	Grocery Store	Organic Grocery	Supermarket	Health	Fast	all
0	Allerton	0	2	0	3	0	0	0	1	0	3	4	5	9
1	Bath Beach	0	1	0	2	0	0	0	0	0	0	0	3	3
2	Bay Ridge	0	0	0	0	0	0	0	2	0	0	2	0	2
3	Baychester	0	1	0	1	0	0	0	1	0	3	4	2	6
4	Bedford Park	0	0	1	1	0	0	1	2	0	2	5	2	7
5	Bedford Stuyvesant	0	1	0	0	0	0	0	1	0	1	2	1	3
6	Belmont	0	0	0	0	2	0	0	1	0	0	3	0	3
7	Bensonhurst	1	1	0	1	0	0	0	0	0	1	2	2	4
8	Bergen Beach	0	0	0	0	0	0	0	0	0	1	1	0	1

Figure 1: Sample of Refined Data

A frequency graph of the fraction of businesses that sold affordable healthy food.

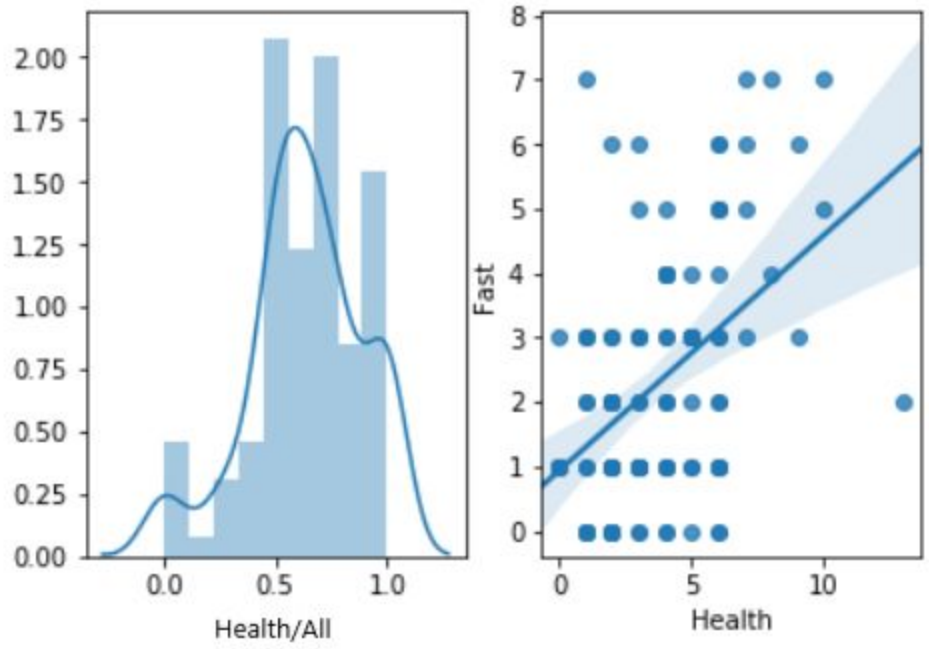

Figure 2: Graphs Showing Distribution of Healthy vs Convenient Stores

A k means algorithm was fit to the “healthy” and “fast” columns. Three clusters were used to portray poor access, bountiful access, and the middle ground. The clusters were

portrayed on a scatter plot of health options vs fast options and, with some notable exceptions, seemed to portray varying levels of food access.

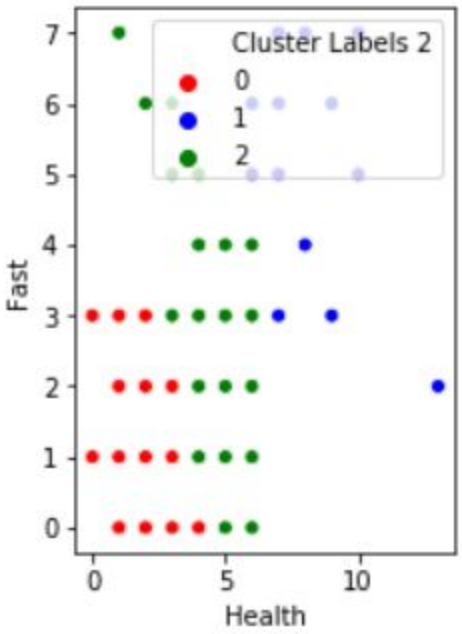

Figure 3: Clustering Algorithm

The different clusters were represented on a map of New York City.

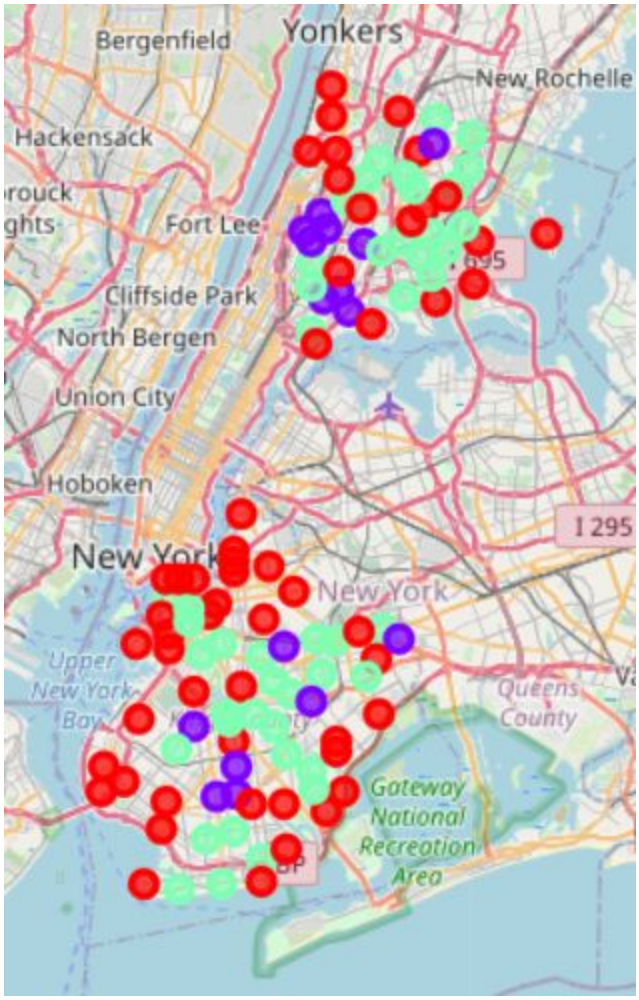

Figure 4: Map of NYC with Clusters Shown

Results

The red dots on the map have poor food access. The green dots have the best access. The best place to place a grocery store would be where there are as few green dots and as many red dots as possible. Neighbourhoods on or near the coast tend to have worse food access, likely because there are fewer businesses within a mile radius. The Northern neighbourhoods of both Brooklyn and the Bronx tended to have poorer food access.

Qualitatively, many red dots surround the Brooklyn Navy Yard, 4th Avenue, Rockaway Parkway, and Van Cortlandt Park.

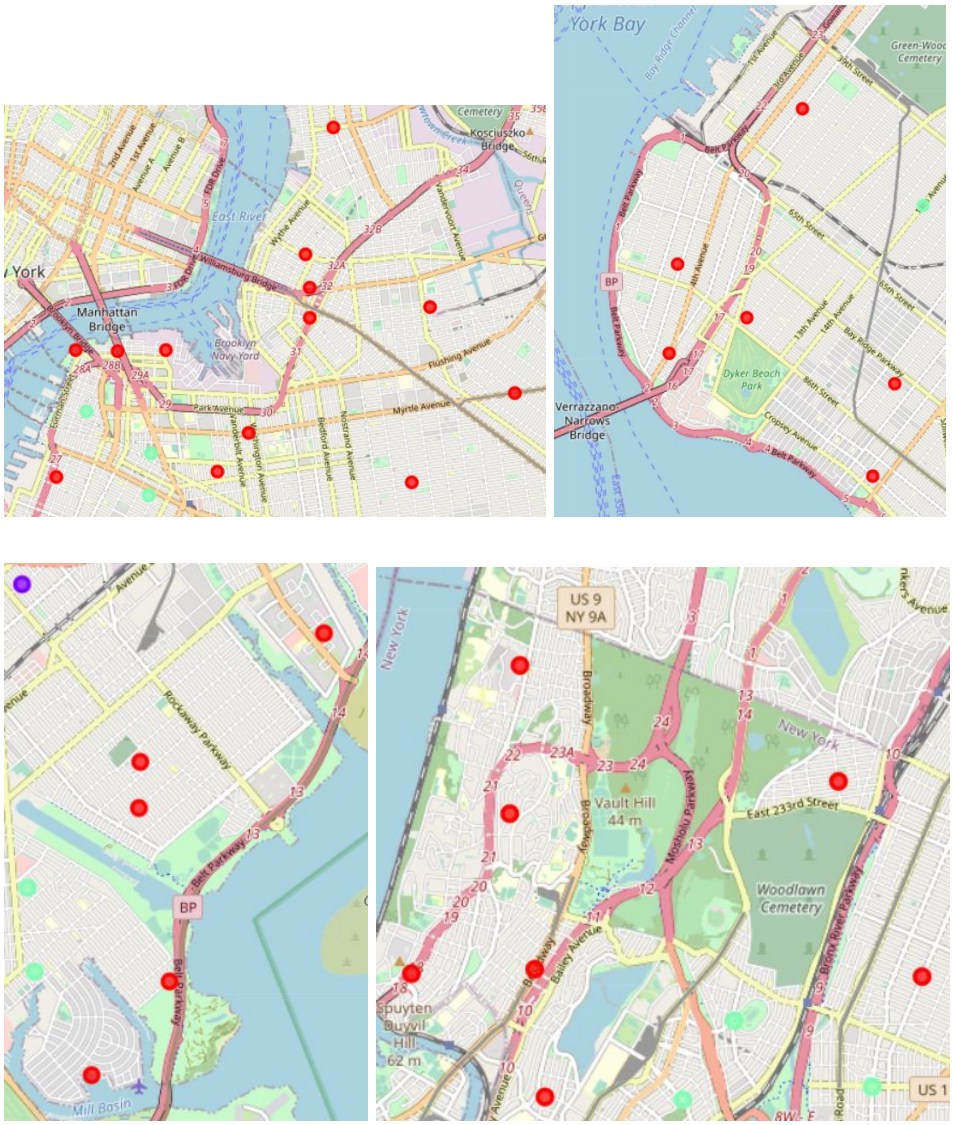

Figure 5: The above mentioned areas of interest. In order.

Discussion

One flaw with the methodology is that it treats all grocery stores as equal. A giant supermarket chain will probably be a better source of healthy food and service a larger radius

than a small grocer. Some convenience stores also strive to provide healthy options, but the methodology lumps them all in together.

The methodology also doesn't account for demographic data like population density. A neighbourhood should require more healthy food services the more people. I assume to a certain extent the number of neighbourhoods and their locations is based on such demographics, so this may have been controlled to a certain extent by whatever system New York uses to designate neighbourhoods. Transportation and safety were also not considered.

The k-means algorithm seemed to do well with two exceptions. In figure three, shown below, the dots at (1, 7) and (3, 6) seem like they should be a lower designation. A different k-value may be more appropriate.

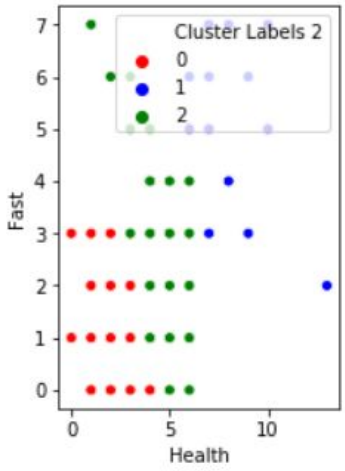

Figure 3

Conclusion

Were I an actual business owner looking to supply healthy food to urban areas, I would follow up this project with a methodology that accounts for population density, and designates a different effective radius and healthiness score to different businesses. Based on the findings of the capstone I would have the follow up focus on the areas of interest identified in the capstone.