Food Scarcity in the Big Apple

Analysis of Food Deserts in New York City

Paul Torres
Coursera — IBM Data Science Professional Capstone
December 15, 2020

Introduction

Food scarcity is a well-known problem in New York City. It refers to the fact that some areas lack healthy food options that are usually provided by supermarkets or fresh food stores. These "food deserts", which they are colloquially called, often lead to higher rates of obesity and obesity-related illnesses. Alternatively, these areas also have a high number of fast-food options. Fast food restaurants are often, and well-deservedly considered unhealthy options because of the ingredients used and the caloric intake averaged on every meal. Supermarkets are one of the few establishments that offer a plethora of fresh produce and healthy options in high quantity.

Business Case

The goal of this project was to use data collected from Foursquare to find areas that had similarities in terms of the number of healthy food options they offer. Government agencies and community advocates would use this information to determine what areas are most in need of subsidized healthy food options.

While subsidizing is often seen as a waste of money, the burden on the healthcare system due to obesity-related illnesses far outpaces money spent on subsidizing low-cost healthy food.

Data

For this project, collecting the following data will be vital:

- List of neighborhoods in New York City
- Latitude and Longitude of the neighborhoods
- Foursquare API calls to gather venues located around those coordinates

In order to acquire all of this data, the project calls for web scraping Wikipedia pages for the neighborhoods in each borough of New York. Once that is done, the next step is using Python package Geoencoder in order to assign coordinates to each of the neighborhoods. The next section will go over the methodology.

Methodology

First, I found five different Wikipedia pages that gave me a list of neighborhoods in each of the five boroughs. These are important because they will be the basis for which we determine our coordinates. We will use the beautifulsoup package from Python to collect the data in order to use the Geocoder package. Using Geocoder, we give each neighborhood a Latitude and Longitude value. This is based on the API we use that gives us exact coordinates.

Next, we used the Foursquare API to find the venues and their categories within a radius of the coordinates. The venue is just another word for establishment. Each venue could be a store, a playground, a landmark, etc. The categories are more important than the names and we use the categories in order to find clusters later on. Removing all venues that were not part of the group that we decided were necessary for the clustering. Venues that were deemed relevant for this project are in categories including:

- Fast Food Restaurants
- Supermarkets
- Grocery Store

All venue entries that were not categorized as the above were removed from the data frame. From there an average was found for each neighborhood. This means that the percentage of fast food stores in the total for food places was found. This applies to all of the options listed above. This method removes some neighborhoods altogether and they will need to be added back in afterward. This resulted in a few NaN values that were the result of not having any values within those categories. They were quickly replaced with zeros and pushed into the clustering phase.

Since the dimensions were narrowed down ahead of time, feature selection was not of the utmost importance. However, principal component analysis (PCA) was done next in order to determine the amount of variance each of the features explained. The results were a matrix of numbers that were determined to be the most The results were used in

KMeans clustering. An important flaw in KMeans is the randomized placement of the centroids. This leads to different clusters than may be found elsewhere. In order to combat that we looped through the random placement of the centroids and how many centroids to use. Looping through the possible number of clusters, the number was narrowed down by looking at the Silhouette Score and the Elbow Curve.

Once the clusters were finalized we map them, using the coordinates, on the Folium. We include the details of their respective scores and the neighborhood names in order to determine the trend among the clusters.

Results

The results of the clustering model told an interesting story. Firstly, after we ran several iterations and concluded the best centroid numbers we found the silhouette scores based on those centroids. The final clustering model that scored second highest was the 3 centroid model. While the 4 centroid model scored higher, it actually captured a lot of noise so I went with the 3 centroid model. *(Image 1)*

The model broke up the clusters based on their supermarket and fast food restaurant scores. The most interesting part of clustering models is that they are unsupervised. This means that any information given by the model must be interpreted by the user.

The clusters broke down as such:

- Cluster 1 — supermarkets are more often available.
- Cluster 2 — supermarkets and fast-food restaurants are equally numerous.
- Cluster 3 — fast-food restaurants are more numerous.

(Image 2)

Discussion

Clustering is one of my favorite unsupervised techniques because it allows for outside knowledge in order to analyze the results.

Based on the knowledge I bring to the project about New York City I can make several conclusions on the clustering locations.

1. Most of the commercial districts seem to have higher percentages of fast-food restaurants. This is obvious, both by their placement in the downtown areas of New York City and by the idea that tourists/commuters would rather have access to faster food than supermarkets. These areas also number the most in the city.
2. Supermarkets are in a surplus in the more residential areas. Even when you give that caveat, the areas with the highest ratio of supermarkets are in higher-income areas like Riverdale in the Bronx — or are away from high traffic areas that serve tourists/commuters to and from work. These clusters number the least.
3. The last cluster is where the percentages are about even. This cluster has a mixture of both other clusters. They are more residential but are still considered high traffic areas for commuters/tourists.

Conclusion

The end of the project brought many questions and many ideas as to what to do next. Combining the information gleaned with this data with census information in order to obtain income and demographic information would be a worthwhile path moving forward.

Appendix

Image 1

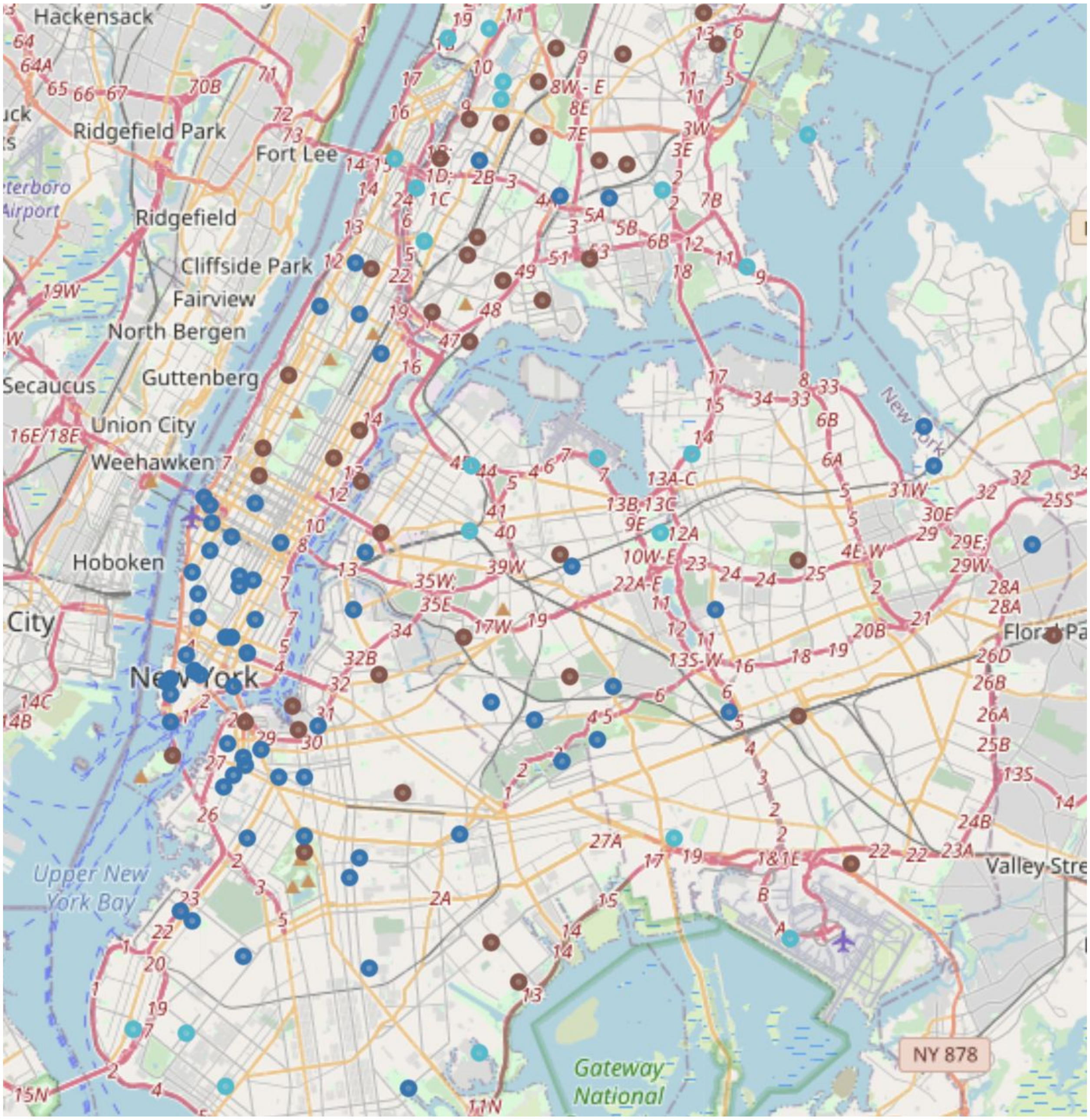
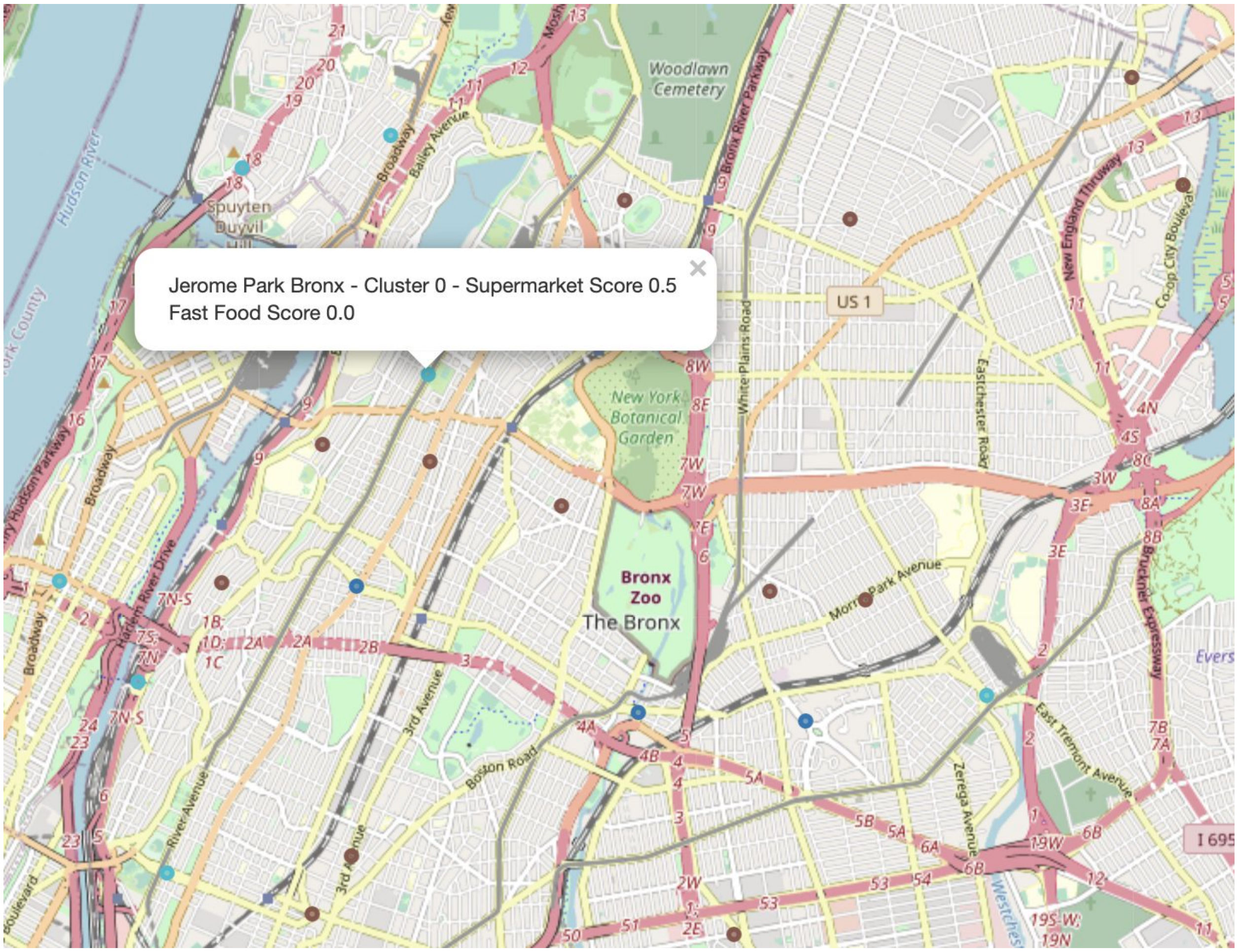


Image 2



Jerome Park Bronx - Cluster 0 - Supermarket Score 0.5
Fast Food Score 0.0