Access to Fresh Produce in New York City

APPLIED DATA SCIENCE CAPSTONE

ADRIENNE MEEHAN

Introduction

Obesity and diet-related disease are a growing problem in the United States. A healthy diet including fresh fruits and vegetables can decrease the risk of developing overweight, obesity, and other chronic diseases. Many Americans lack access to affordable, healthy produce options.

New York City has both the highest population and the greatest population density in the United States. However, the city also has a significant income disparity with significant differences between neighborhoods. This project examines which neighborhoods have the lowest access to fresh produce and what demographic factors correspond with low access to fresh produce. This information could be used by public health organizations, government programs, and charitable organizations to design interventions to increase access to fresh produce. The information could also be used to examine the factors underlying lack of access to fresh produce to develop interventions.

Data

This project relies on data from Foursquare, New York City's Neighborhood Tabulation Areas (NTA), and American Community Survey demographic data by NTA.

Counts of grocery stores and other stores that sell fresh produce were pulled from Foursquare. Categories that were included as selling fresh produce were "Grocery Stores", "Supermarkets", "Organic Grocery", "Fruit & Vegetable Store", "Farmers Market", "Deli / Bodega", and "Health Food Store." Instances of each of these categories were counted for each neighborhood.

New York City neighborhood boundaries were found from the New York City website Open Data page. The NTA Map includes the geographical boundaries for each NTA in New York City.

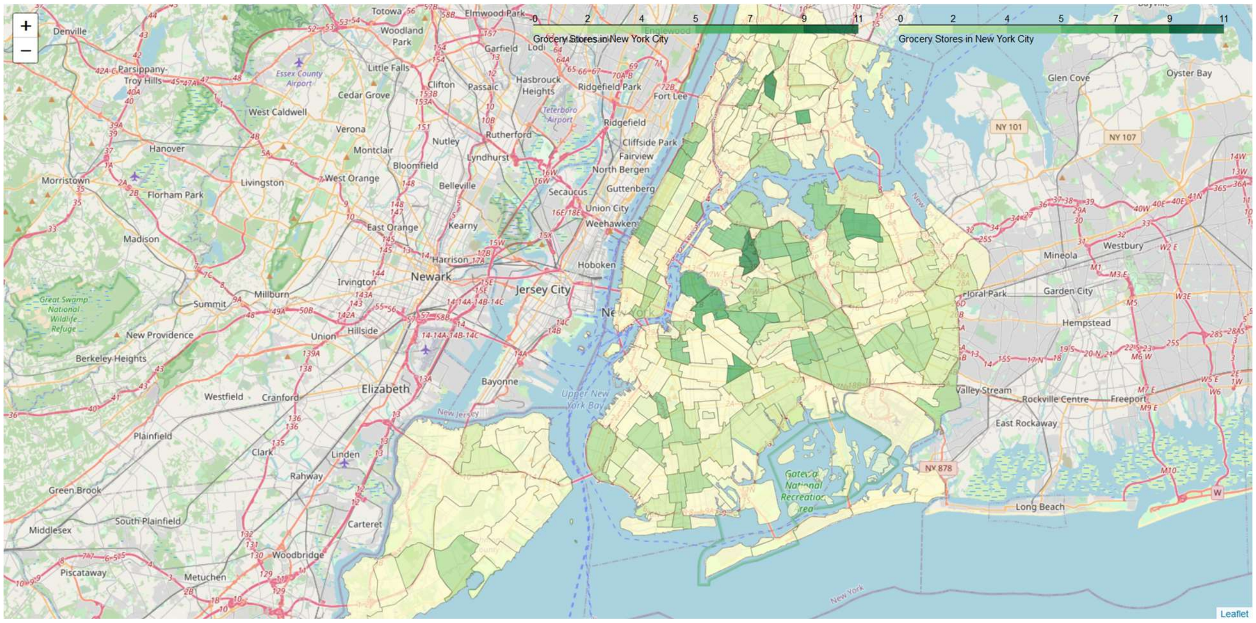
For demographic data for each neighborhood, this report uses the Demographic Profiles of ACS 5 Year Estimates at the Neighborhood Tabulation Area (NTA) level obtained from the New York City website Open Data page. The economic and demographic tables from the 2012-2016 data were used.

Methodology

Exploratory Data Analysis

For this report, the neighborhoods of New York city were mapped and the most popular venues in each neighborhood were categorized. The demographic and economic data for each neighborhood were imported from the American Community Survey. The number of grocery stores in each neighborhood was also mapped. The following categories were included as grocery locations: Grocery Store, Supermarket, Farmers Market, Fruit & Vegetable Store, Organic Grocery, Health Food Store, and Deli/Bodega.

Data were mapped from the ACS datasets to the New York neighborhood data. Some neighborhoods had to be combined in the ACS to maintain anonymity and statistical significance. To account for grouped neighborhoods, venue data were aggregated for these neighborhoods. In other cases, the ACS split neighborhoods so for the analysis, the demographic data were averaged to get an estimate for the entire neighborhood.

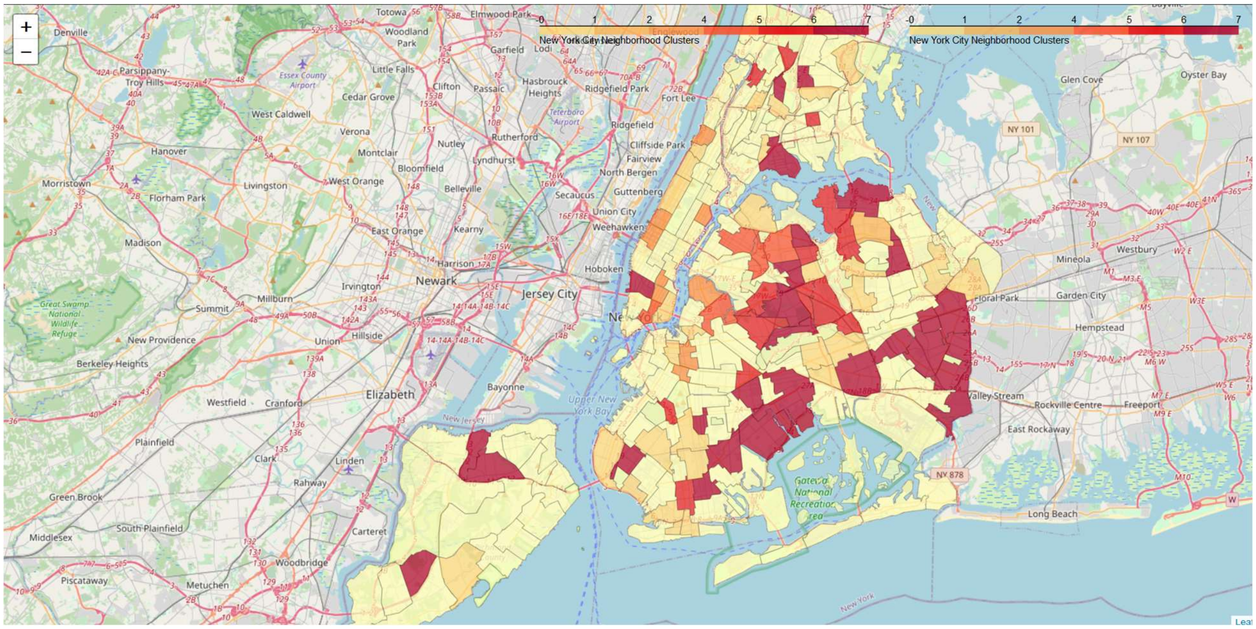


Inferential Statistical Testing and Machine Learning

Several statistical tests were run using the economic and demographic data for each neighborhood to predict number of grocery stores in each neighborhood. The following predictors were included in all models: the percentage of the population identifying as Hispanic; the percentage of the population identifying as Black, Non-Hispanic; median household income; the percentage of the population utilizing SNAP benefits; and the percentage of the population below the poverty level. Two regression models were run, one with raw predictors and the second with normalized predictors. Both models had a very low value for R^2 (0.134 and 0.080 respectively). A KNN model was also fit to this data. Splitting the data into a training and test set (70% training, 30% test) suggested that the highest model accuracy would be achieved with K = 1 resulting in only 23% prediction accuracy.

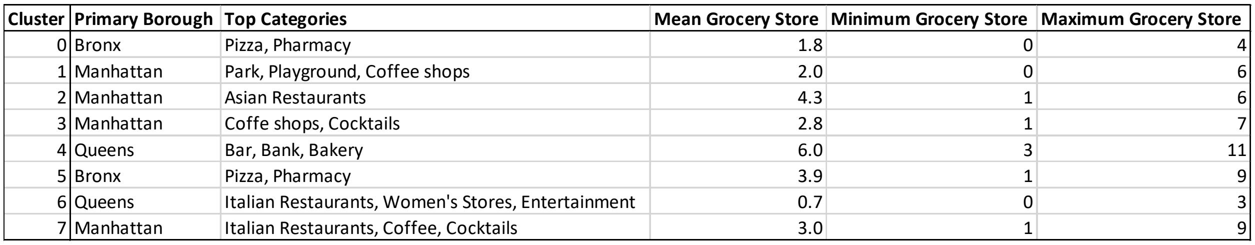
Based on the poor model accuracy using the economic and demographic data, new models were built using most popular venue categories in each neighborhood as predictors. A KNN model was created using these predictors, however using the same technique of splitting the data suggested that it was not a good fit for the data. The highest prediction accuracy was achieved at 26% with K = 5.

KMeans was then used to categorize the data based on most popular venues in each neighborhood and see which clusters tended to have the greatest number of grocery stores per neighborhood. The elbow method suggested that the optimal number of clusters was 8, so a KMeans model was built to categorize the data into 8 clusters.



Results

Eight clusters were identified. The cluster that had the highest mean number of grocery stores was Cluster 4 with a mean of 6.0. Cluster 4 was primarily found in Queens and the top categories included Bars, Banks, and Bakeries. The cluster with the lowest average number of grocery stores was Cluster 6 with a mean of 0.7. This cluster was primarily located in Queens as well and the top categories were Italian restaurants (including pizza restaurants), women's stores, and entertainment venues.



Cluster	Primary Borough	Top Categories	Mean Grocery Store	Minimum Grocery Store	Maximum Grocery Store
0	Bronx	Pizza, Pharmacy	1.8	0	4
1	Manhattan	Park, Playground, Coffee shops	2.0	0	6
2	Manhattan	Asian Restaurants	4.3	1	6
3	Manhattan	Coffe shops, Cocktails	2.8	1	7
4	Queens	Bar, Bank, Bakery	6.0	3	11
5	Bronx	Pizza, Pharmacy	3.9	1	9
6	Queens	Italian Restaurants, Women's Stores, Entertainment	0.7	0	3
7	Manhattan	Italian Restaurants, Coffee, Cocktails	3.0	1	9

Discussion

This analysis has several limitations that may have contributed to the low accuracy of the model. Venue data were determined by finding the coordinates of the center of the neighborhood and counting number of venues within a fixed distance from this center. This method could result in undercounting of venues for neighborhoods with large total area, overcounting for neighborhoods with small total area, and inaccuracies in neighborhoods that are irregularly shaped. A fixed number of venues were counted from each neighborhood. For neighborhoods with a greater number of venues, a large number of grocery stores may be missing from the imported Foursquare data. There might be certain factors that influence whether a grocery store is included on Foursquare. A web search indicates that not all grocery stores are included for some neighborhoods. Finally, number of grocery stores per neighborhood might not be a good indicator of actual access to fresh fruit and vegetables. Access to public transportation, walk scores, and proximity of neighborhoods might have a significant impact on actual access.

Conclusion

Access to fresh produce is important for health and preventing obesity. Unfortunately, Americans do not have equal access to fresh produce. Determining what factors influence access to fresh produce could help public health organizations, local non-profits, and governments increase access throughout the United States.

New York city is highly populous but has significant inequality between neighborhoods, including in access to fresh produce. This analysis clustered New York neighborhoods into 8 clusters to determine which neighborhood characteristics are correlated with greater number of grocery stores.