K-Means Clustering of Venues in the Bronx and Staten Island

A Data Science Project

by David Anim-Addo

# A. Introduction

A.1. Background Discussion

A review of research in public health shows a relationship between the socioeconomics of an urban environment and the well-being of its inhabitants. This is supported by evidence that residents with access to first-class resources have decreased incidence and prevalence of disease as well as overall better health. For example, in a paper by Diez Roux (2003) the influence of neighborhood environments on cardiovascular health is examined. The author mentions that residing in lower income neighborhoods places residents at a higher risk for coronary heart disease. Socioeconomic characteristics have also been linked to other health factors such as smoking, diet, and body mass index. In documenting this problem Diez Roux lists a number of neighborhood features linked to cardiovascular risk. Among these features are access to healthy foods, recreational resources, and transportation. The importance of geographically coded information is discussed as it allows for a wide variety of public health measures to be studied.

More recently, in an investigation by Tabaei, et al. (2018) researchers studied the influence of socioeconomics, food options, and the quality of the built residential area, such as parks and recreation, on glycemic control in adults with diabetes in New York City. Their findings revealed that better diabetes control was associated with residents in advantaged neighborhoods with access to high-quality environmental resources. In addition, when individuals with poor glycemic control moved from disadvantaged neighborhoods to neighborhoods with more resources their health status improved.

In both papers, the need for further research on neighborhood features and their impact on public health is expressed. The construction of the neighborhood environment and its available resources can have long-term health consequences for its residents. The question then becomes about the type of analytical methods useful for gaining further insight in public health. For example, what kind of venues exist across neighborhoods with different socioeconomics? The answer to this question could be a starting point for further research in public health.

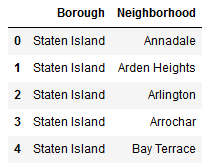
The goal of this project is to use K-Means clustering to examine the venue distribution of two New York boroughs, Staten Island and the Bronx. These areas were selected based on U.S. Census Bureau income and poverty data in order to compare regions of different socioeconomic levels. The venues for this analysis were acquired through the Foursquare API which has nine categories: arts and entertainment, college and education, events, food, nightlife, outdoors and recreation, professional, residence, shops, and traveling. With these categories, neighborhoods in Staten Island and the Bronx will be segmented and clustered to observe how these venues are distributed across the different residential areas. Afterwards, any variations in clusters between the two regions will be explored.

A.2. Data Review

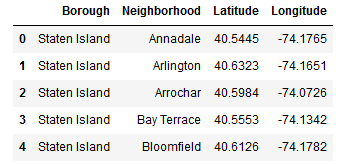
The socioeconomic and population data for this project was collected from the U.S. Census Bureau on New York City income and poverty data (4). This data was used to identify New York boroughs of differing socioeconomic levels for venue cluster analysis. Staten Island has a median household income of $76,244 and the Bronx has a median household income of $36,593. Population data for each borough has been estimated as of July 1st, 2018. The population of the Bronx is 1,432,132 people with 532,487 housing units. The population of Staten Island is 476,179 with 181,199 housing units. The beautifulsoup (1) data scraping method will be used to acquire a list of neighborhoods in Staten Island and the Bronx from Wikipedia (5,6). The geopy library (3) will then be used for geocoding to find the longitude and latitude of the neighborhoods in each borough. Finally, venue category data will be acquired through the Foursquare API (2) to find the most common venues for each neighborhood.

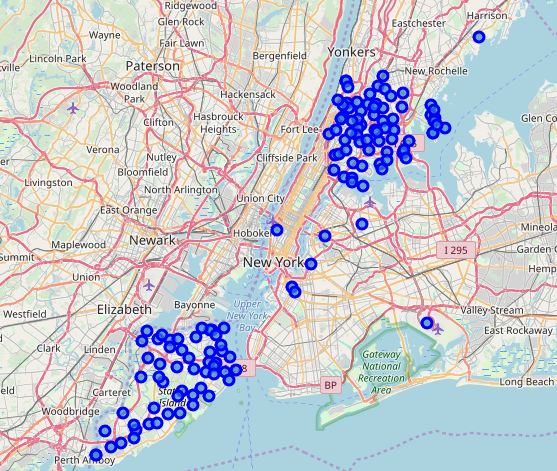
# B. Method

The database used for this project was constructed using available lists of neighborhoods in Staten Island and the Bronx from Wikipedia. These websites were scraped using the beautifulsoup library. The information from the websites was cleaned so that only the neighborhoods were left. The neighborhoods for each borough were then appended into a dataframe.

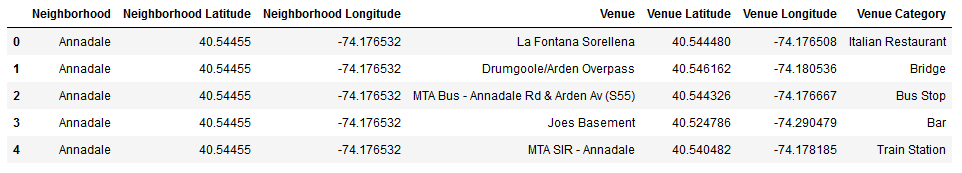


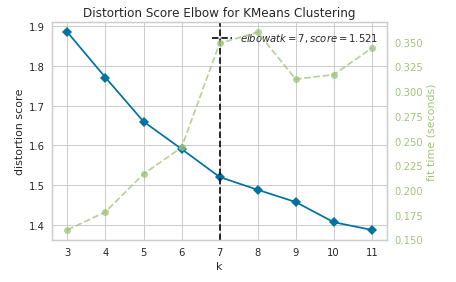
Next, each neighborhood was used for geocoding with Nominatim from geopy. This provided latitude and longitude coordinates for each neighborhood. These coordinates were concatenated with the original dataframe.



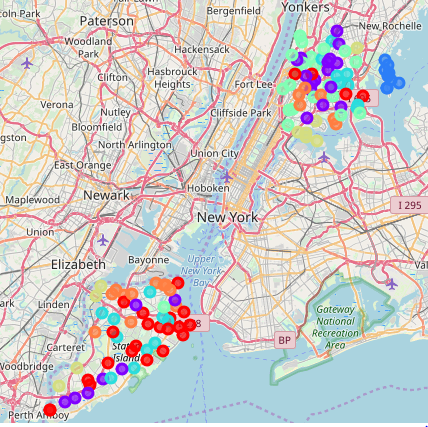
The Folium library was then accessed to plot a map of New York City with circle markers for each neighborhood. Eleven neighborhoods were found to have incorrect location data as determined by their location on the map. These rows were removed from the dataframe.

The Foursquare API was then employed to gather venue data for each neighborhood in Staten Island and the Bronx. The search radius was set to 500 meters with a limit of 100 results. The venue data for each neighborhood was then added to the dataframe.



 Analysis of the dataframe began with grouping the table by neighborhood and finding the count of venues. The onehot encoding method was used for encoding the categorical features. A new dataframe was created to display the top 5 Foursquare venues in each neighborhood. The clustering analysis began with the use of Yellowbricks’ KElbowVisualizer.

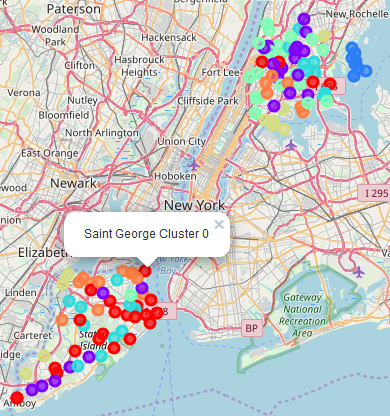
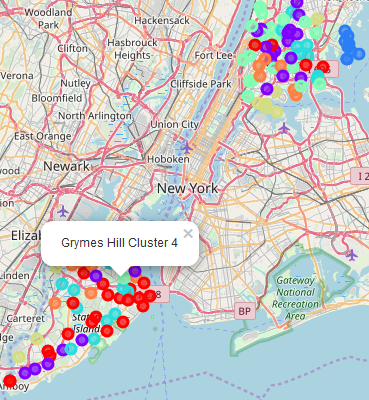
This allowed for the correct number of clusters to be determined prior to Kmeans clustering. The Kmeans algorithm is a common method for unsupervised clustering and will be useful here since there are common venue categories in each borough. Clustering was executed and the cluster labels were attributed to each neighborhood in the dataframe. A second map was created with Folium showing the clusters mapped across the neighborhoods in Staten Island and the Bronx. Each of the clusters was then assessed for the most common venues in order to classify them according to the Foursquare API main categories of arts & entertainment, college & education, event, food, nightlife, outdoors & recreation, professional, residence, shops, and travel.

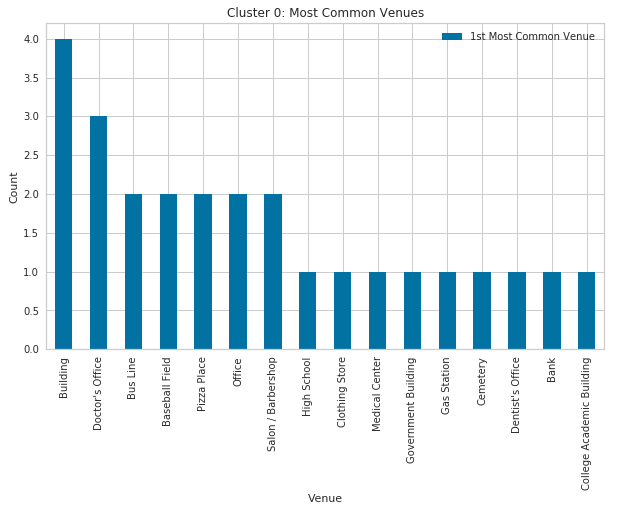


# C. Results

A total of 12,240 venues were returned by the Foursquare API with 473 unique categories. While location data was not available for all neighborhoods, 77 total neighborhoods were analyzed in the Bronx along with 60 in Staten Island. Using the Kmeans Elbow Method it was determined that 7 clusters would be needed for this analysis. After the Kmeans algorithm was executed the clusters needed to be classified according to their most common venues.

|  |  |
| --- | --- |
| **Cluster Number** | **Venues Classification** |
| Cluster 0 | Professional |
| Cluster 1 | Shops & Service |
| Cluster 2 | Travel & Transport |
| Cluster 3 | Professional/Medical |
| Cluster 4 | Residential |
| Cluster 5 | Professional & Service |
| Cluster 6 | Service & Transport |



# 

# D. Discussion

Observations of the cluster map distribution showed that far more residential venues (cluster 4) were present in the Bronx versus Staten Island. Since the Bronx has over double the population and housing units of Staten Island this was expected, however the presence of only one residence cluster in Staten Island speaks of more evenly distributed residential venues. The Bronx has a far higher distribution of residential venues. Since this project was conducted with regards to public health studies, it may be valuable to investigate the impact of many residential clusters on the health of residents in those neighborhoods. Considerably more professional venues (cluster 0) were observed in the Staten Islands borough than in the Bronx. These clusters include mostly business and medical offices. The amount of professional venues in Staten Island does coincide with the higher household median income of the borough. In terms of public health, this could have an impact on residents’ health due to job opportunities and access to medical services. Many businesses offer health benefits to their employees so the number of such venues may also be of interest in public health research.

# E. Conclusion

In this project Kmeans clustering was used to gain insight on the most common venues within a 500-meter radius of neighborhoods in Staten Island and the Bronx. Staten Island has over double the median household income of the Bronx while the Bronx has over double the population of Staten Island. In terms of public health, the clustering of venues in each borough showed potential for impacting residents’ standard of healthy living. This opens the door to more domains of research that can possibly improve public health across all socioeconomic regions of a city.

## Journal References

Diez Roux, A. V. (2003). Residential Environments and Cardiovascular Risk. *Journal of Urban Health*, *80*(4), 569–589. <https://doi.org/10.1093/jurban/jtg065>

Tabaei, B. P., Rundle, A. G., Wu, W. Y., Horowitz, C. R., Mayer, V., Sheehan, D. M., & Chamany, S. (2018). Associations of Residential Socioeconomic, Food, and Built Environments with Glycemic Control in Persons with Diabetes in New York City from 2007-2013. *American Journal of Epidemiology*, *187*(4), 736–745. <https://doi.org/10.1093/aje/kwx300>

## Internet References

1. [Beautiful Soup 4](https://www.crummy.com/software/BeautifulSoup/)
2. [Foursquare API](https://developer.foursquare.com/docs/resources/categories)
3. [Geocoding with Geopy](https://geopy.readthedocs.io/en/stable/)
4. [U.S. Census Bureau QuickFacts: New York County](https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork,bronxcountybronxboroughnewyork,queenscountyqueensboroughnewyork,kingscountybrooklynboroughnewyork,richmondcountystatenislandboroughnewyork,newyorkcitynewyork/HSG010218)
5. [Wikipedia: List of Bronx Neighborhoods](https://en.wikipedia.org/wiki/List_of_Bronx_neighborhoods)
6. [Wikipedia: List of Staten Island Neighborhoods](https://en.wikipedia.org/wiki/List_of_Staten_Island_neighborhoods)