Clustering of New York neighborhoods for an Election Campaign

Capstone Project

Mabeth Anonuevo

Introduction/Business Problem

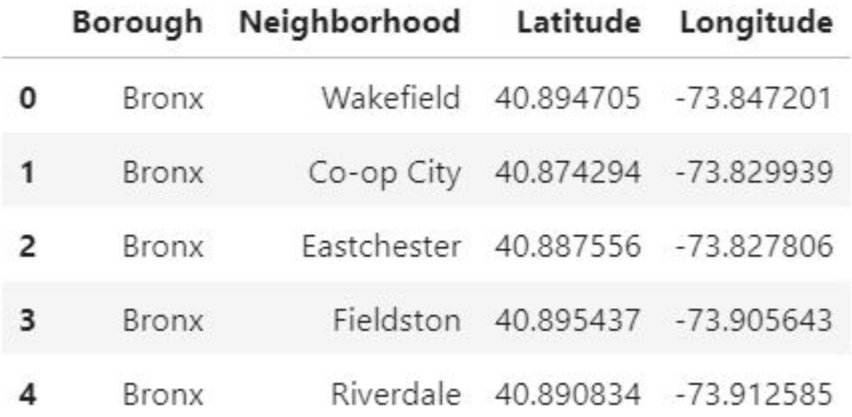
A prospective governor candidate is preparing for his election campaign. Although everyone has an idea of how the New York neighborhoods are, he wants to actually see the data and have a factual basis. He wants to know what neighborhoods are alike, so they can target them not only together but better. The messaging can be the same for neighborhoods that have the same types of places where residents hang out. They want to know that if a certain neighborhood has more parks, the residents there must value the outdoors and would the campaign should have a stronger stance on protecting the environment. If a certain neighborhood has a concentration of universities, then there must be stronger message on what the candidate can do for education.

Data

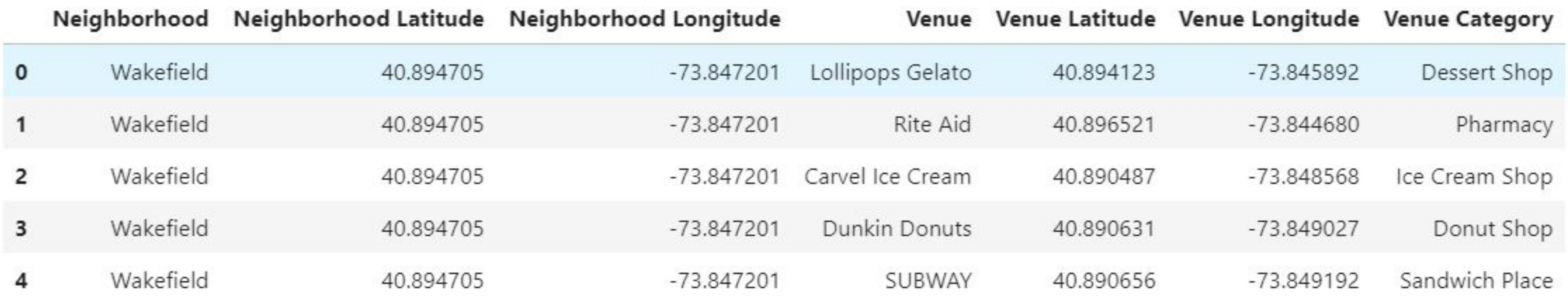
To address the business problem, we would use the New York neighborhood data and merge it with the Foursquare data. The New York neighborhood data would provide us with the neighborhood name, the boroughs, latitude and longitude. From Foursquare, we can get venues information like the venue names and categories. These would help us cluster the neighborhoods and have an idea on how we could provide better messaging for specific clusters.

For this project, the analysis will focus on New York City and its 5 boroughs.

Data



	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896521	-73.844680	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Dunkin Donuts	40.890631	-73.849027	Donut Shop
4	Wakefield	40.894705	-73.847201	SUBWAY	40.890656	-73.849192	Sandwich Place

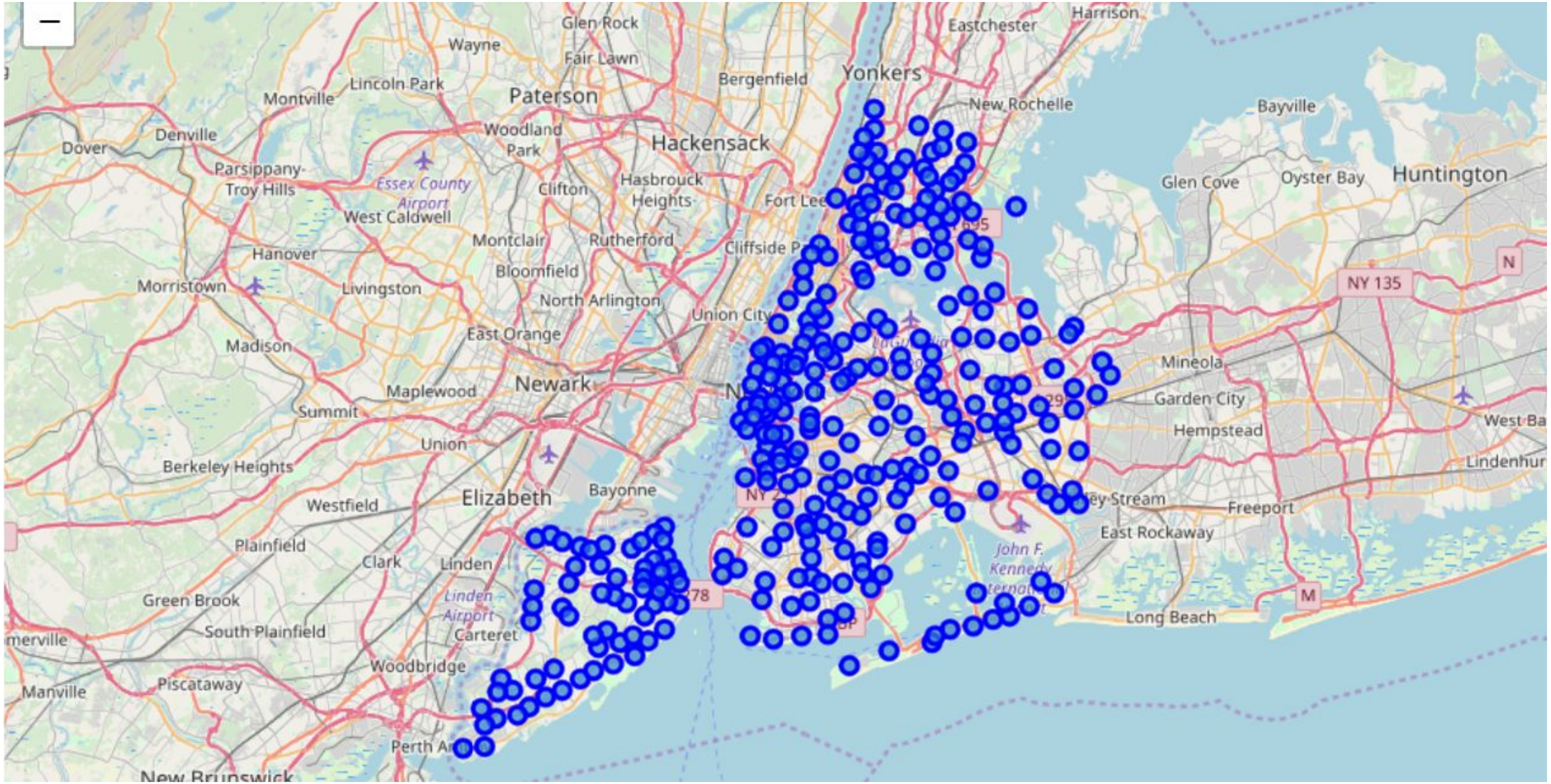
Methodology

- First, we downloaded the necessary packages that will enable us to do data processing, map rendering and machine learning. Some of the packages that we have imported are pandas, json, geocoder, and folium.
- The New York City neighborhood data was taken from https://geo.nyu.edu/catalog/nyu_2451_34572. Once the data was loaded, we looked into the columns and the data types to determine which ones would be relevant to the study. The Features section contains the relevant information (neighborhood, borough, latitude and longitude), so we make sure to extract this portion instead of using the whole dataset.

Methodology

- We then use the loaded data and transformed it into a dataframe. To get the coordinates of New York City, we used the geocoder. We needed to get the coordinates to make sure that the map that we will be creating would be centered on NYC.
- Using the latitude and longitude coordinates, we mapped out all the neighborhoods of New York City. This was done using the package Folium.
- Since the focus of the governor's campaign is to understand the whole city, we didn't do any other filtering. Every neighborhood in NYC was included and mapped.

Methodology: NYC neighborhoods

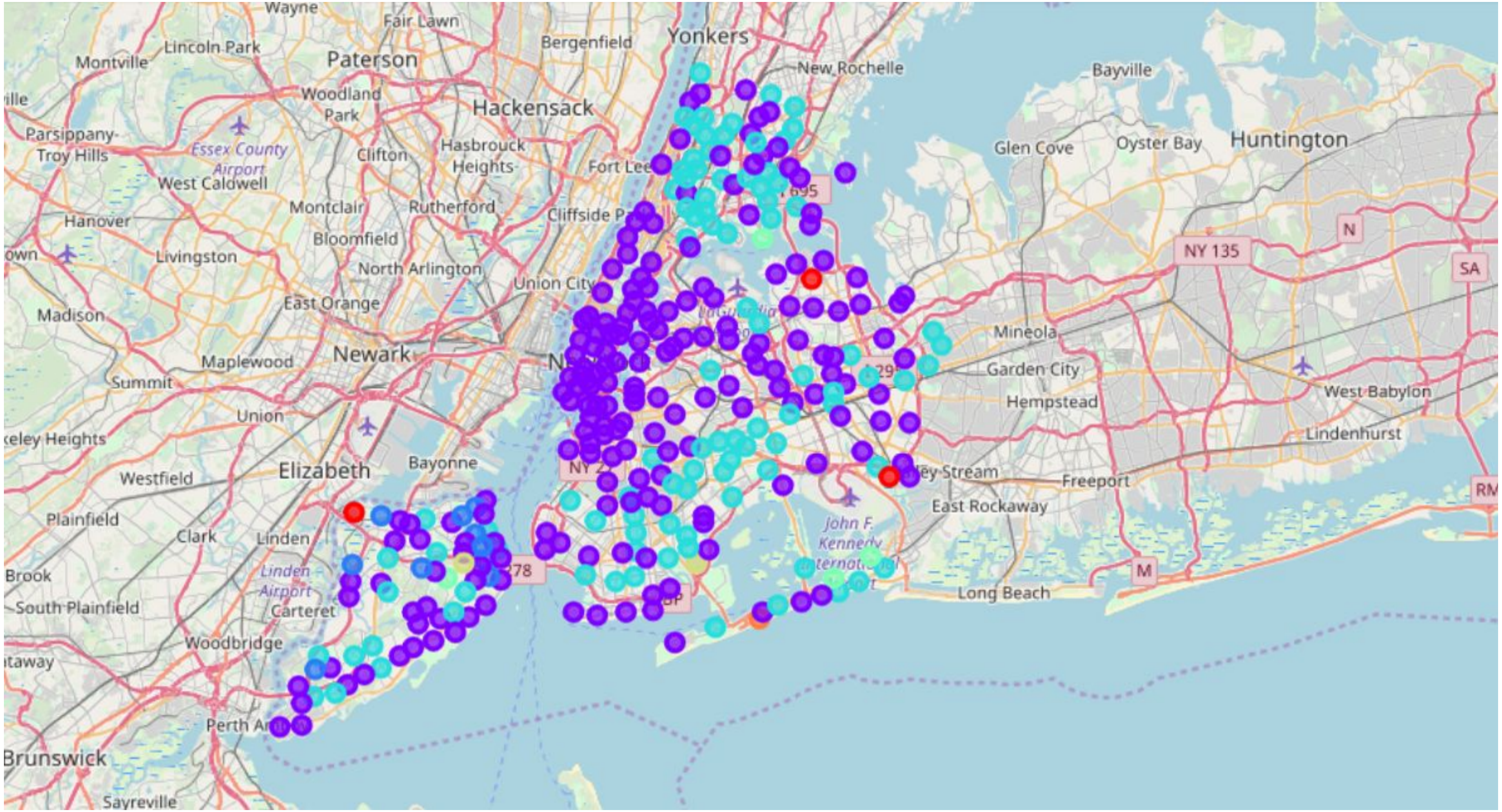


Methodology

- The Foursquare API was used to get the venues in each neighborhood. A limit of 100 was set, so the maximum total of venues associated to each neighborhood will be 100. Only venues in the range of radius = 500 would be included.
- The most important venue data that we used in this study is the venue category. We computed for the frequency of each venue category in each neighborhood and showed the top 10 most common categories in that area. These would be the basis in doing the neighborhood clustering.
- We used the K-means Clustering machine learning algorithm to cluster the neighborhoods. We set the number of clusters to 7. We were then able to get 7 clusters of neighborhoods and provided description that would differentiate the clusters from each other.

RESULTS

Results: Map with clusters



Cluster 1: Quaint neighborhoods with concentrations of immigrants (e.g. Filipino, Middle Eastern)

Neighborhood Count: 3

These are neighborhoods with quaint venues like bubble tea shops, farms, field, and flower shops. Filipino and Middle Eastern restaurants are also popular from which we can infer that there is a customer base for those types of cuisines in that area.

Possible campaign recommendation: Possible campaign topics could be centered on gentrification and immigration.

Cluster 2: Culture centers with diverse restaurants, art, and transportation

Neighborhood Count: 191

This is the largest cluster. The collection of venue categories in this cluster is vast. There is a diversity of restaurant types, museums and transportation. Just from looking at the most common venue categories, these might be the cultural centers/entertainment centers.

Possible campaign recommendation: Campaign must focus on diversity, the arts, and employment issues. Since these clusters are centers of business, the governor must also make sure that he conveys his plans for the economy.

Cluster 3: Areas near Bus Stops and markets

Neighborhood Count: 10

These are neighborhoods where the most common venues are bus stops and farmer's markets. There are definitely still in the city, but probably not in the tourist-centered areas.

Possible campaign recommendation: Have to discuss transportation issues (e.g. traffic issues, fare). Can also discuss nutrition since this area might be more concerned about food and health>

Cluster 4: Melting pot, foodie paradise

Neighborhood Count: 94

This can be similar to Cluster 2 but with a stronger concentration in restaurants. These are neighborhoods where people eat and congregate among each other.

Possible campaign recommendation: Immigration, employment, and urban issues would be good topics for this cluster.

Cluster 5: More suburban places near parks and playgrounds

Neigborhood Count: 4

This cluster is composed of neighborhoods with concentration of parks and playgrounds. From that information, we can infer that these are more suburban places, with amenities catering more to families.

Possible campaign recommendation: From our hypothesis, these neighborhoods will be mainly comprised of families, so issues like maternity leaves, healthcare, and education would be good bets to attract voters.

Cluster 6: Areas near wide open spaces, not city centers

Neighborhood Count: 2

The two neighborhoods in this cluster have unusual common venue categories: lake and farm. We can assume that these are far from the city center and would be more quiet than Clusters 2 and 4.

Possible campaign recommendation: Environmental and energy issues would be good topics since these would be priorities for this cluster.

Cluster 7: Coastal area

Neighborhood Count: 1

There is only one neighborhood in this cluster. The most common venue is a beach, so we can assume that this is more of a coastal area and would have different priorities than the other clusters.

Possible campaign recommendation: Disaster prevention and climate change would be priorities in this area because they are more prone to floods.

Discussion

The recommendations for the campaign are noted with the results. From this clustering exercise, we are able to determine characteristics of neighborhood clusters. Knowing these characteristics, we are able to provide very specific campaign topics that the candidate can focus on. This analysis could be improved by incorporating data like population data and election data (e.g. voter turnout on previous elections). This analysis could also be improved by trying out different clustering algorithms and by trying out different values for k (no. of clusters).

Conclusion

With the use of location and venue data, we can enable a political candidate to learn more about their potential constituents to conduct a better campaign and to better plan for governing. This is an example of how data can lead us to make better decisions.