Capstone Project
The Battle of Neighborhoods

IBM Data Science Professional Certificate

Problem

I would like to develop a recommendation service for people who would like to relocated in a city or in the country. Therefore, I aim to find similiar neighborhoods in a city so that people would have a similiar experience and opportunities in the new neighborhood when they move. My target audience for this project is anyone who is looking for a new apartment to move. This project could be implemented as a feature to the current apartment listing websites.

Data Description

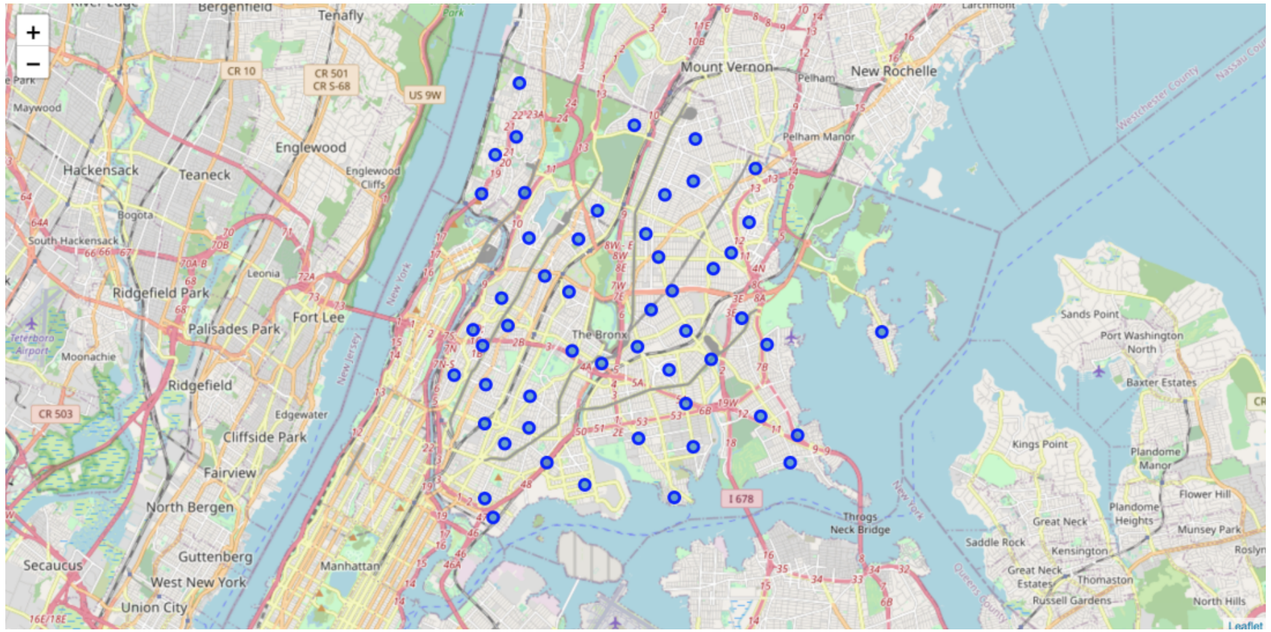
To work on the project, I select New York - Bronx for the area they I will do my analysis. I will compare a selected group of boroughs in this region I will find similiar neighborhoods which have similiar type of venues along with their latitude and longtitude coordinates.

Links for the datasets that I will use:
https://cocl.us/new_york_dataset
https://geo.nyu.edu/catalog/nyu_2451_34572

Methodology / Data Analysis

First, I downladed the data from the links I provided in the data description section. Then, I manipulated the data and identified neighborhoods in Bronx area in New York City. I also utilized Folium library to visualize the neigborhoods in Bronx area.

	Borough	Neighbourhood	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



Methodology / Data Analysis

To find a similar neighborhood, first, I needed to find out available venues in Bronx area. Therefore, I used Foursquare API to find out what kind of venues are available in Bronx area. My method here is to find at least 10 common venues in each neighborhood, so that this would give me a similar neighborhood where people could find same type of venues when they move. I've set the radius as 0,5 km from the center of each neighborhood and pulled the list of venues available within this region, and below is an example of my list.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

Methodology / Data Analysis

In Bronx, there are 1211 venues available and I was able to get details for all those venues using Foursquare API. Then, I created a list of top venues for each borough.

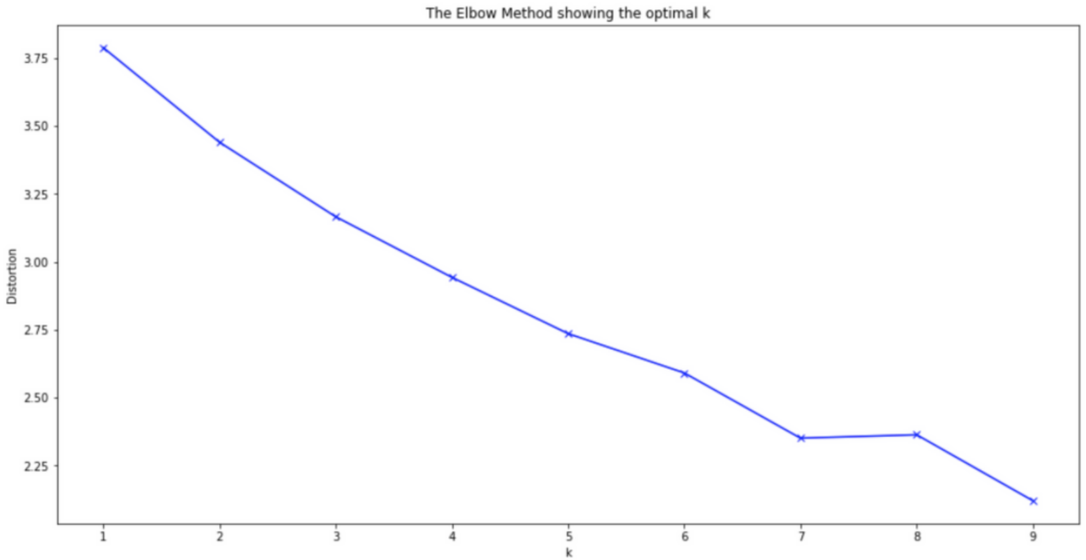
	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Allerton	Pizza Place	Supermarket	Deli / Bodega	Chinese Restaurant	Spa	Intersection	Fried Chicken Joint
1	Baychester	Donut Shop	Mattress Store	Fried Chicken Joint	Pizza Place	Pet Store	Discount Store	Sandwich Place
2	Bedford Park	Diner	Chinese Restaurant	Pizza Place	Mexican Restaurant	Sandwich Place	Supermarket	Spanish Restaurant
3	Belmont	Italian Restaurant	Pizza Place	Deli / Bodega	Bakery	Dessert Shop	Donut Shop	Bank
4	Bronxdale	Convenience Store	Bank	Pizza Place	Coffee Shop	Chinese Restaurant	Performing Arts Venue	Paper / Office Supplies Store

Methodology / Data Analysis

I used K-means clustering to group neighborhoods based on their top venues. This is the logic that I followed to find the similar neighborhoods. At this moment, I classify neighborhoods based on their top venues since top venues have big impact on the character of a neighborhood. For instance, if there are many Chinese restaurants in a neighborhood, this are should be a neighborhood where many Chinese people live.

Methodology / Data Analysis

To do that, I used elbow method and determined the optimal K.



According to elbow method, I identified that the optimal K is 7. Therefore I picked 7 as the optimal number of clusters to create.

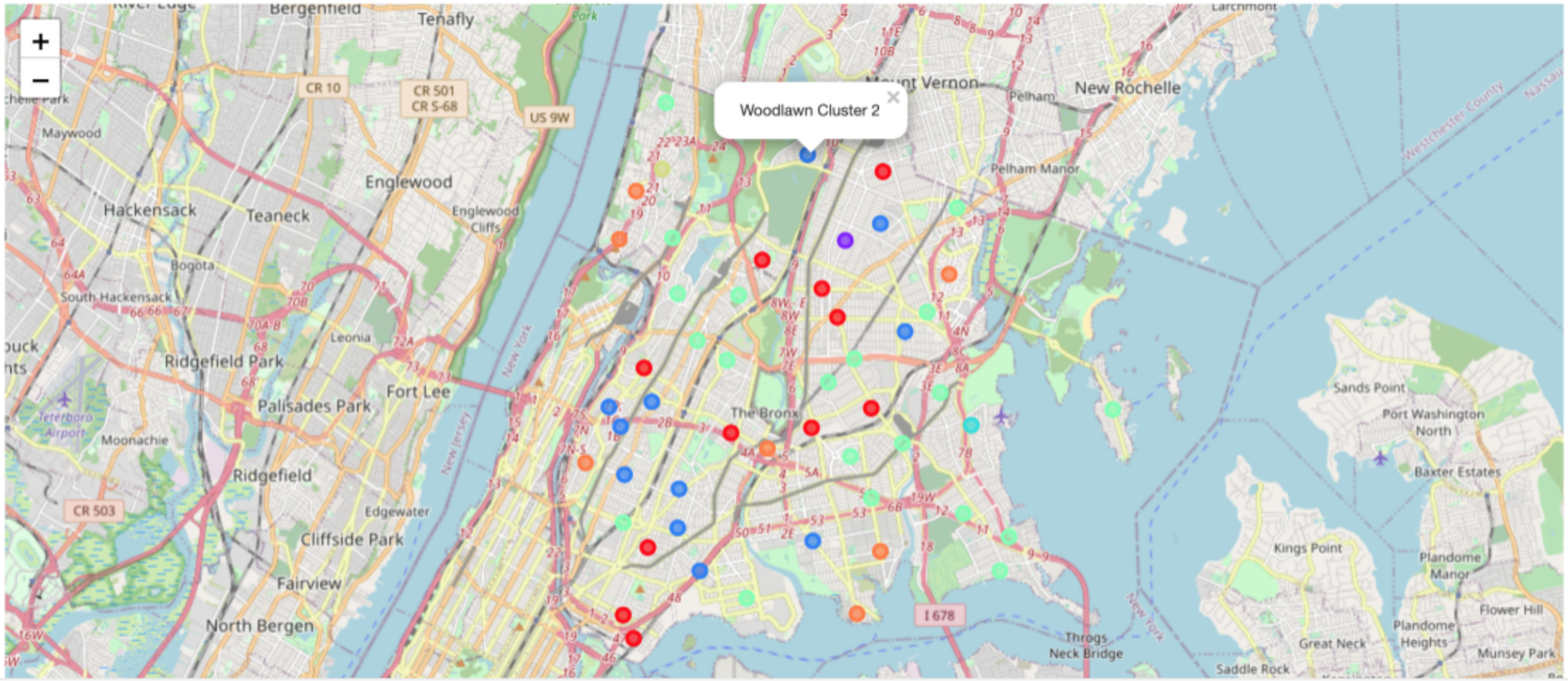
Methodology / Data Analysis

Then I continued with K-means algorithm for k=7. I merged the neighborhood table with the results from K-means algorithm and I found out the following data.

	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Bronx	Wakefield	40.894705	-73.847201	0	Pharmacy	Ice Cream Shop	Gas Station	Deli / Bodega	Pizza Place	Dessert Shop	Sandwich Place
1	Bronx	Co-op City	40.874294	-73.829939	6	Bus Station	Accessories Store	Bagel Shop	Post Office	Pharmacy	Restaurant	Chinese Restaurant
2	Bronx	Eastchester	40.887556	-73.827806	4	Caribbean Restaurant	Bus Station	Diner	Deli / Bodega	Convenience Store	Bowling Alley	Pizza Place
3	Bronx	Fieldston	40.895437	-73.905643	5	Medical Supply Store	Plaza	River	Women's Store	Dive Bar	Fish Market	Fish & Chips Shop
4	Bronx	Riverdale	40.890834	-73.912585	6	Park	Bank	Plaza	Home Service	Playground	Food Truck	Moving Target

Methodology / Data Analysis

And I also used Folium library again to visualize the clusters I identified.



Methodology / Data Analysis

With this analysis, anyone who is considering to move to another part of Bronx can pick a similiar neighborhood which is similiar to the area that he/she is currently living. This analysis can be easily extended to the entire NYC and it can be replicated for all cities.

Conclusion

Utilizing Foursquare API data, we are able to understand certain characteristics of different neighborhoods. This analysis enables us to make data drive decisions while we are looking for different alternatives when we need to relocate. Obviously we can improve the algorithm by adding different types of analysis as well- yet, K-means algorithm, folium mapping tools are very powerful tools that we can make a detailed analysis.

Relocation is a challenge for everyone, and many times even if we find a good apartment we like, we end up in a scenario that we don't like the neighborhood. Therefore, before people make relocation decisions, this analysis will help them a lot to know more about the neighborhood they move in.