Coursera Capstone

IBM Applied Data Science Capstone

The Battle of the Neighborhoods - Report

"Where should I open a grocery store with a restaurant inside in Brooklyn?"

May 2020

Introduction

Shoppers today are craving for new experiences. Apart from shopping they would like to combine different activities when entering a store such as dining or watching movies. This has led many retailers to look for opportunities of opening restaurants in their brick-and-mortal stores. They have figured out that it is a really great way to deliver on experience if we consider that making a dinner relevant to what is sold is a win-win for every retailer. Apart from that spending more time with the consumer and observing his/her behavior is always advantageous to retailers to better understand them. Providing a dining experience in your store will guarantee more time with them. So, the question that arises is where should I open the store to maximize revenue and eliminating competition? Such a business decision needs careful consideration and thorough analysis prior to deciding which is the best location to open it because it is one of the main factors that will determine if the store will be a success or not.

Business Problem/Target Audience

The objective of this capstone project is to provide guidance on a specific retailer XYZ who wants to open a grocery store which will include a restaurant inside. The area of interest is Brooklyn and more specifically the center of it. As a result, the business question that needs to be addressed is where would we recommend that the retailer should open this new grocery store/restaurant? This is a case study that would be of great interest/value to other retailers who look for similar opportunities in the market. Using data science methodology and machine learning techniques like clustering, we will answer the question and suggest the optimum solution.

Data

To address the question, we will use the following data:

- List of neighborhoods in Brooklyn, NY. This is the area of focus of the specific retailer – identifying the optimum location in Brooklyn to open the new grocery store/restaurant is the project scope.
- Latitude and longitude coordinates of these neighborhoods. This is required to plot the maps and get the venue data.
- Venue data, more specifically data related to grocery stores and restaurants. These data will be used to perform clustering analysis on the neighborhoods mentioned above.

Sources of data/manipulation techniques

- The dataset exists for free on the web: https://geo.nyu.edu/catalog/nyu_2451_34572. It contains a total of 5 boroughs and 306 neighborhoods. We will be based on the lab instructions to retrieve the data of New York and then we will focus on Brooklyn.
- We will use geopy library to get the latitude and longitude of Brooklyn.
- We will create a map(folium) to visualize Brooklyn and its neighborhoods.
- Then we will get the geographical coordinates of the neighborhoods in the area.
- Foursquare API keys will be used to get the venue data for those neighborhoods along with many categories of the venue data. We are rather interested in grocery stores and restaurants around to solve the business problem.

What follows

Data section is followed by Methodology section where we will discuss in detail the steps taken in this project, from data analysis to machine learning techniques (K-means clustering) that were applied to finally identify the suggested location.

Methodology

We start by retrieving the list of neighborhoods in the borough of Brooklyn. The dataset exists for free on web (https://geo.nyu.edu/catalog/nyu_2451_34572) for New York. We will be based on the lab instructions to retrieve the data of New York and then we will subtract the area of Brooklyn.

We will use geopy library to get the latitude and longitude of Brooklyn and visualize its neighborhoods using a map via Folium package. We need to do this to make sure that the geographical coordinates correspond to the relative area. Foursquare API credentials (client id and client secret) will be used to get the top 100 venues that are within a radius of 500 meters.

What comes next is: API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Venue data in JSON format will be retrieved and we will extract the venue name, venue category and venue coordinates. We transform the list of Python nested dictionaries into a pandas dataframe. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. We need this preparation prior to proceeding to clustering.

We will run K-means clustering algorithm to cluster the neighborhoods into 4 clusters based on their frequency of grocery stores, Italian and Mexican restaurants thus need to create a dataframe including these types of venue categories only. We do this mainly because we are looking for candidate locations to open a grocery store thus, we need to find the cluster with the lowest frequency in grocery stores. However, at the same time, we need to identify the cluster which has the lowest frequency of Italian/Mexican restaurants since the retailer is interested in offering these cuisines in the restaurant inside the store.

Finally, the location of interest is in the center of Brooklyn rather than the seaside – we must keep this in mind as well prior to selecting the appropriate location. As a result, after we have concluded with the optimum cluster based on frequency of groceries, Italian and Mexican restaurants we focus on the neighborhoods belonging to this specific cluster. We create a dataframe keeping as columns the coordinates of each specific neighborhood in this cluster and we append another line (Brooklyn coordinates). Once we have finalized with the new dataframe we re-run K-means to identify the neighborhoods that are closest to Brooklyn center. We end up with our final cluster which consists of 10 neighborhoods - so if we visualize them in the map we could recommend that the retailer should open the grocery store in one of these and more specifically in Wingate since it is very close to Brooklyn center and has zero frequency of grocery stores Italian and Mexican restaurants.

Results

Running K-means, we end up with 4 clusters based on the lowest frequency of occurrence of grocery stores/Italian/Mexican restaurants:

	Neighborhood	Grocery Store	Mexican Restaurant	Italian Restaurant	Cluster Labels
34	Gerritsen Beach	0	0	0	0
25	East New York	0	0	0	0
29	Flatlands	0	0	0	0
32	Fulton Ferry	0	0.017241	0.017241	0
37	Greenpoint	0.03	0.03	0.01	0
42	Manhattan Beach	0	0	0	0
43	Manhattan Terrace	0.04	0	0	0
44	Marine Park	0	0	0	0
45	Midwood	0	0	0	0
24	East Flatbush	0	0	0	0
47	Mill Island	0	0	0	0
52	Paerdegat Basin	0	0	0	0
55	Prospect Lefferts Gardens	0.019608	0	0.019608	0
57	Red Hook	0.020833	0	0	0
58	Remsen Village	0	0	0	0
60	Sea Gate	0	0	0	0
62	South Side	0.01	0.02	0.01	0
63	Starrett City	0	0	0	0
65	Vinegar Hill	0	0	0	0
49	North Side	0.01	0.01	0.01	0
21	Downtown	0.020619	0.010309	0.010309	0
69	Wingate	0	0	0	0
10	Brownsville	0	0	0	0
19	Cypress Hills	0	0	0	0
18	Crown Heights	0	0	0	0
6	Borough Park	0	0	0	0
20	Ditmas Park	0.022222	0.022222	0	0
5	Boerum Hill	0.021739	0.01087	0.01087	0
17	Coney Island	0	0	0	0
7	Brighton Beach	0.022727	0	0	0
4	Bergen Beach	0	0	0	0
2	Bedford Stuyvesant	0	0	0	0
11	Bushwick	0.013333	0.066667	0.013333	1
54	Prospect Heights	0.013514	0.081081	0	1
26	East Williamsburg	0.014286	0.042857	0	1

28	Flatbush	0	0.1	0	1
64	Sunset Park	0.027027	0.081081	0.027027	1
12	Canarsie	0.125	0	0	2
8	Broadway Junction	0.055556	0	0	2
51	Ocean Parkway	0.05	0	0	2
50	Ocean Hill	0.071429	0.035714	0	2
56	Prospect Park South	0.061224	0.040816	0	2
48	New Lots	0.0625	0	0	2
23	Dyker Heights	0.142857	0	0	2
27	Erasmus	0.08	0	0	2
38	Highland Park	0.076923	0	0	2
59	Rugby	0.111111	0	0	2
40	Kensington	0.088235	0.029412	0	2
39	Homecrest	0.075	0.05	0	2
66	Weeksville	0.055556	0	0	2
14	City Line	0.054054	0	0	2
67	Williamsburg	0.029412	0	0.029412	3
61	Sheepshead Bay	0.035714	0	0.035714	3
3	Bensonhurst	0.033333	0	0.066667	3
1	Bay Ridge	0.022727	0.011364	0.056818	3
22	Dumbo	0	0.016667	0.033333	3
9	Brooklyn Heights	0.02	0.02	0.03	3
46	Mill Basin	0.03125	0	0.03125	3
13	Carroll Gardens	0.02	0	0.11	3
41	Madison	0	0	0.111111	3
15	Clinton Hill	0.021277	0.042553	0.053191	3
36	Gravesend	0	0	0.133333	3
35	Gowanus	0.015385	0.030769	0.061538	3
68	Windsor Terrace	0.038462	0	0.038462	3
33	Georgetown	0	0.033333	0.033333	3
16	Cobble Hill	0.010309	0.010309	0.030928	3
31	Fort Hamilton	0	0.014286	0.042857	3
30	Fort Greene	0.031746	0	0.047619	3
53	Park Slope	0.016949	0.016949	0.050847	3
0	Bath Beach	0	0	0.041667	3

We can see in the table below the average frequency per cluster and category:

	Grocery Store	**Mexican Restaurant**	**Italian Restaurant**
Cluster 0	**0.68%**	**0.38%**	**0.28%**
Cluster 1	1.36%	7.43%	0.81%
Cluster 2	7.92%	1.11%	0.00%
Cluster 3	1.72%	1.03%	5.46%

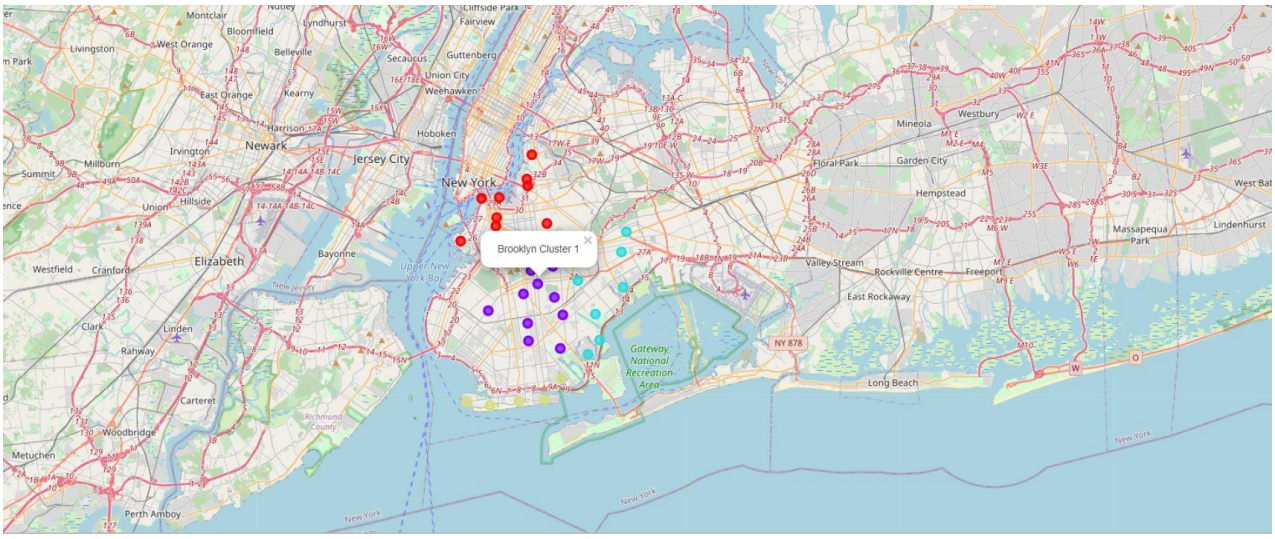
We focus on the cluster 0 because of the lowest on average frequency of groceries, Italian and Mexican restaurants. We can see the relative table below:

	Neighborhood	Grocery Store	Mexican Restaurant	Italian Restaurant	Cluster Labels
34	Gerritsen Beach	0	0	0	0
25	East New York	0	0	0	0
29	Flatlands	0	0	0	0
32	Fulton Ferry	0	0.017241	0.017241	0
37	Greenpoint	0.03	0.03	0.01	0
42	Manhattan Beach	0	0	0	0
43	Manhattan Terrace	0.04	0	0	0
44	Marine Park	0	0	0	0
45	Midwood	0	0	0	0
24	East Flatbush	0	0	0	0
47	Mill Island	0	0	0	0
52	Paerdegat Basin	0	0	0	0
55	Prospect Lefferts Gardens	0.019608	0	0.019608	0
57	Red Hook	0.020833	0	0	0
58	Remsen Village	0	0	0	0
60	Sea Gate	0	0	0	0
62	South Side	0.01	0.02	0.01	0
63	Starrett City	0	0	0	0
65	Vinegar Hill	0	0	0	0
49	North Side	0.01	0.01	0.01	0
21	Downtown	0.020619	0.010309	0.010309	0
69	Wingate	0	0	0	0
10	Brownsville	0	0	0	0
19	Cypress Hills	0	0	0	0
18	Crown Heights	0	0	0	0
6	Borough Park	0	0	0	0
20	Ditmas Park	0.022222	0.022222	0	0
5	Boerum Hill	0.021739	0.01087	0.01087	0
17	Coney Island	0	0	0	0
7	Brighton Beach	0.022727	0	0	0
4	Bergen Beach	0	0	0	0
2	Bedford Stuyvesant	0	0	0	0

We proceed to clustering once again to identify this time the neighborhoods which are closest to Brooklyn center after we have appended in the above table the coordinates of its neighborhood along with Brooklyn ones:

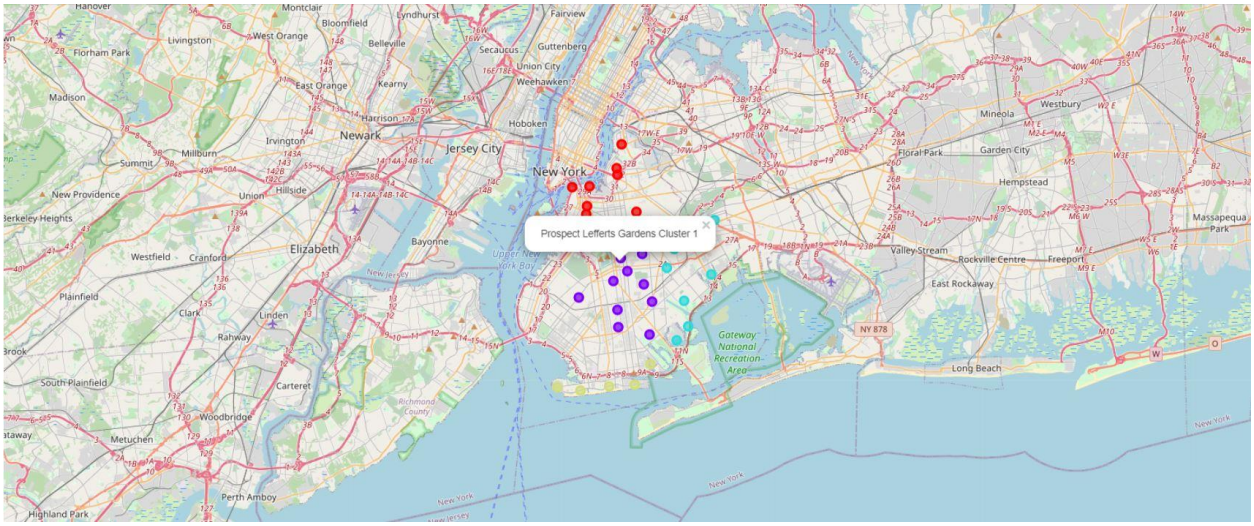
	Neighborhood	Latitude	Longitude	Cluster Labels
34	Gerritsen Beach	40.590848	-73.930102	0
17	Coney Island	40.574293	-73.988683	0
7	Brighton Beach	40.576825	-73.965094	0
42	Manhattan Beach	40.577914	-73.943537	0
60	Sea Gate	40.576375	-74.007873	0
20	Ditmas Park	40.643675	-73.961013	1
43	Manhattan Terrace	40.614433	-73.957438	1
29	Flatlands	40.630446	-73.929113	1
44	Marine Park	40.609748	-73.931344	1
55	Prospect Lefferts Gardens	40.65842	-73.954899	1
24	East Flatbush	40.641718	-73.936103	1
45	Midwood	40.625596	-73.957595	1
18	Crown Heights	40.670829	-73.943291	1
6	Borough Park	40.633131	-73.990498	1
69	Wingate	40.660947	-73.937187	1
2.5	**Brooklyn**	**40.650104**	**-73.949582**	**1**
37	Greenpoint	40.730201	-73.954241	2
21	Downtown	40.690844	-73.983463	2
62	South Side	40.710861	-73.958001	2
5	Boerum Hill	40.685683	-73.983748	2
49	North Side	40.714823	-73.958809	2
32	Fulton Ferry	40.703281	-73.995508	2
57	Red Hook	40.676253	-74.012759	2
2	Bedford Stuyvesant	40.687232	-73.941785	2
65	Vinegar Hill	40.703321	-73.981116	2
25	East New York	40.669926	-73.880699	3
47	Mill Island	40.606336	-73.908186	3
19	Cypress Hills	40.682391	-73.876616	3
52	Paerdegat Basin	40.631318	-73.902335	3
10	Brownsville	40.66395	-73.910235	3
4	Bergen Beach	40.61515	-73.898556	3
58	Remsen Village	40.652117	-73.916653	3
63	Starrett City	40.647589	-73.87937	3

Cluster 1 is the cluster which include Brooklyn center so let's visualize the above clusters in a map to finally identify the optimum locations to address our business question:

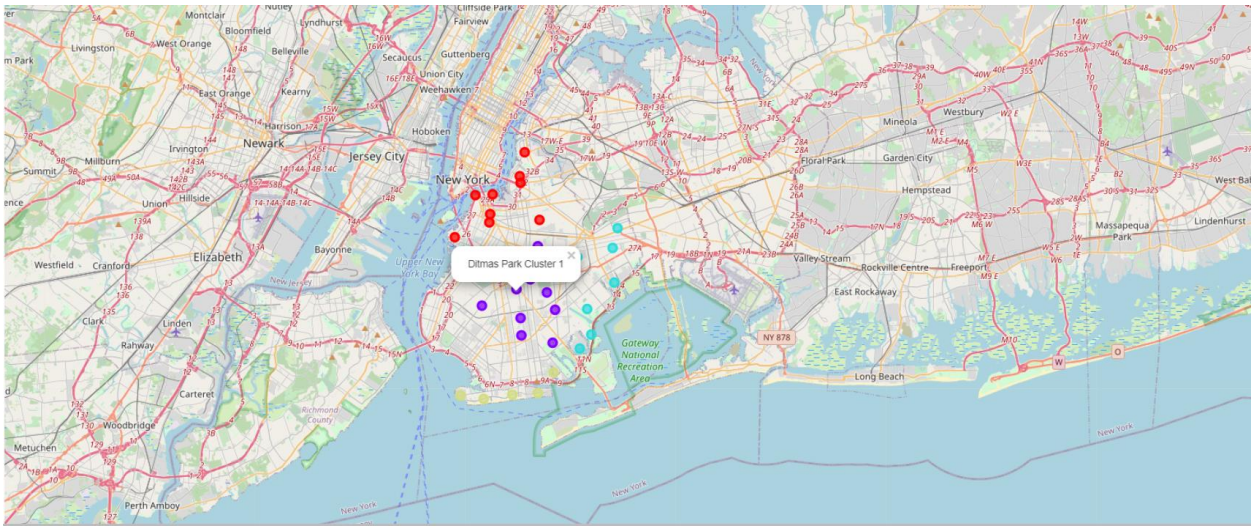


As we can see from the markers in the map above, Prospect Lefferts Gardens, Ditmas Park and Wingate are very close to Brooklyn:

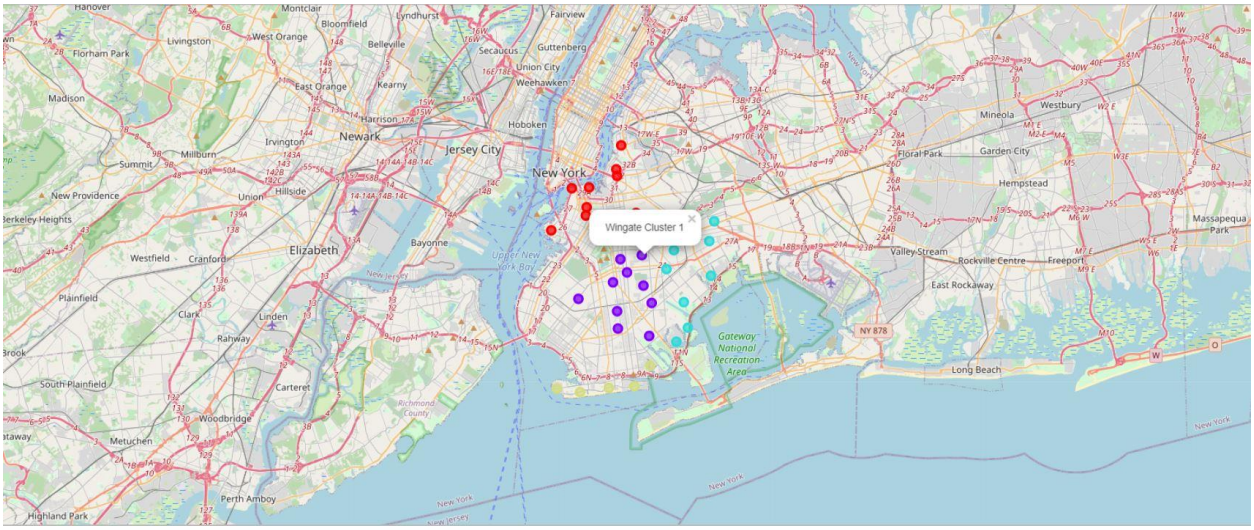
Prospect Leffrerts



Ditmas Park



Wingate



However, as previously mentioned apart from the distance of Brooklyn we should consider the lowest frequency in grocery stores, Italian and Mexican restaurants thus if we recall the average results for these specific neighborhoods, we end up to our final suggestion which is Wingate:

	Neighborhood	Grocery Store	Mexican Restaurant	Italian Restaurant
55	Prospect Lefferts Gardens	0.019608	0	0.019608
69	Wingate	0	0	0
20	Ditmas Park	0.022222	0.022222	0

Discussion/Recommendations

As we have already mentioned in the Results section (K-means first run), the cluster with the highest on average frequency of grocery stores (7.92%) is cluster 2. On the other hand, cluster 0 has very low to zero number of groceries in its neighborhoods (0.68%). If we consider, the Mexican/Italian restaurant's frequency as well, cluster 0 remains the one with the lowest on average frequency for both categories (0.38% and 0.28% respectively). This reveals a great opportunity to XYZ retailer for opening a new grocery store with a restaurant inside, in one of the neighborhoods of this cluster, as the competition from already existing groceries with Mexican or Italian restaurants nearby are eliminated – in most cases turns to zero. Neighborhoods in Cluster 2 are suffering from competition due to high concentration of grocery stores (7.92%) whereas the ones in cluster 1 have the higher frequency in Mexican restaurants (7.43%). Cluster 3 is characterized by the highest frequency in Italian restaurants (5.46%) in comparison with the rest ones. Now, if we focus on the clusters generated - after the second run of K-means - to conclude based on the closest to Brooklyn location it is obvious that the most prevalent cluster is the cluster 1 which includes Brooklyn as well. As we can see from the map the closest neighborhoods are Prospect Lefferts Gardens, Ditmas Park and Wingate (the highlighted ones). Considering that Wingate specifically has a zero frequency of groceries/Italian and Mexican restaurants we can suggest that it could be one of the optimum locations that could be selected for opening a grocery store/restaurant.

In this specific case study, we have considered only two factors frequency of grocery stores, Mexican and Italian restaurants and distance from center however there are plenty of other drivers that we could have incorporated in our analysis such as population of each neighborhood, average income of residents, unemployment rate per region, parking availability nearby, cost of renting in each location, cuisine preferences of the residents and so many others that could undoubtably drive our final decision prior to selecting the desired area for opening the grocery/restaurant. In any case, such an approach/methodology could be adapted and enhanced by other retailers who look for relevant opportunities in the area, taking advantage this time apart from frequency of groceries/restaurants and distance from center multiple additional factors as the ones mentioned above.

Conclusion

To conclude, in the specific case study we have gone through the process of identifying the business problem, specifying, extracting and manipulating the data required, applying machine learning techniques specifically K-means clustering of the data into 4 clusters based on their similarities to end up with providing recommendations to our client, the XYZ retailer, regarding the optimum location to open a new grocery store which will have a restaurant inside serving Italian and Mexican cuisine. To address the business question that was raised in the introduction section, we suggest that most of the neighborhoods in cluster 1 (after second run of K-means) excluding the ones that are in the seaside could be potential selected locations of the retailer. More specifically, the neighborhood Wingate in cluster 1 is the optimum one since it is the very close to Brooklyn and lacks groceries and Italian/Mexican restaurants nearby. Other retailers who look for relevant opportunities could benefit from the insights of the study considering that most of the retailers' goal is to expand their grocery stores to more than one area and of course maximize their revenue by eliminating competition and building their own loyal consumer base.