Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Choosing where to open Zumba Center in NYC

Introduction: Business Problem

In this project we will try to find an optimal location for a Zumba club. Specifically, this report will be targeted to stakeholders interested in opening a Zumba Fitness Club in New York City, The United-States.
Since there are a considerable number of Fitness Club or any Sports Center in New York City we will try to detect locations that are not already crowded with gymnasium. We are also particularly interested in areas with no Zumba Fitness Club in vicinity. We would also prefer neighborhoods with high percentage of Hispanic population, assuming that first two conditions are met.

We want to consider also other factors in our decision making, particularly socio-economic factors. For that we would opt for neighborhoods where unemployment rate is acceptable, average income is medium to high and where average rent price is acceptable.

We will use our data science powers to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decision are:

- ✓ Number of existing Fitness Club in the neighborhood (any type of gym)
- ✓ Number of and distance to Zumba Club in the neighborhood if any
- ✓ Percentage of Hispanic population
- ✓ Unemployment rate, average income, average rent price

Following data sources will be needed to extract/generate the required information:

- ✓ The list of the neighborhoods was be extracted from http://www.infoshare.org/misc/SBANYC.pdf and that I converted it into an excel file.
- ✓ Then we will use the **Shapely** library of python to compute the Coordinates of the centroid of the multipolygon of each neighborhood. The geojson file for that was obtained from https://geodata.lib.berkeley.edu/catalog/sde-columbia-census_2000_032807211977000 **.**
- ✓ The number of gym or fitness club and their type and location in every neighborhood will be obtained using **Foursquare API**. In the **Developer Categories List** and the **Foursquare City Guide**, I have found the Venue types that will be of interests for this project: *Dance Studio, Gym, Gym / Fitness Center, Athletics & Sports, Yoga Studio.*
- ✓ The number of Hispanic population was obtained from csv file downloaded from http://www.infoshare.org and the percentage was obtained by directly calculating it (Number of Hispanic population divided by Total number of population).

- ✓ The Socio-economic data was obtained from csv file downloaded from http://app.coredata.nyc. It contains the *Labor force participation rate,* the *Unemployment rate*, the *Median household income,* and the *Median rent* for each neighborhood.

Data wrangling and exploration

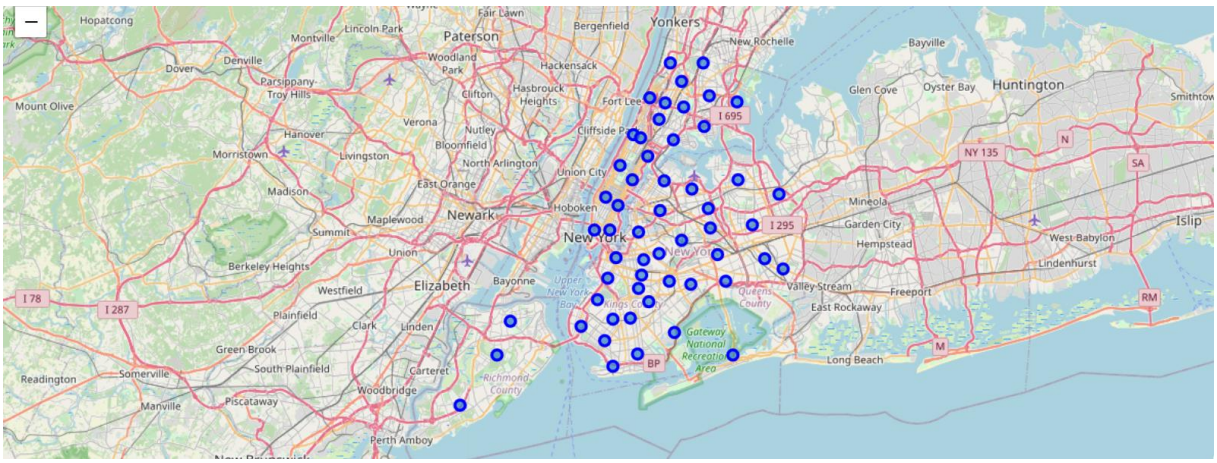
Sub-boroughs areas

First thing I did is downloading the geojson file of all the sub-boroughs areas of New York City and loaded it into Shapely to check it was what I needed.

From the raw geojson data I pulled out the Sub-borough Name and the geometry of the area polygon. I then requested Shapely calculate the centroid location of each area. The resulting data frame looked like this:

	Sub-borough_id	Sub_borough_name	Latitude	Longitude	Coordinates
0	501	North Shore	40.627058	-74.119209	{'type': 'MultiPolygon', 'coordinates': [[[[-7...
1	309	East Harlem	40.794674	-73.935177	{'type': 'MultiPolygon', 'coordinates': [[[[-7...
2	502	Mid-Island	40.593419	-74.137534	{'type': 'MultiPolygon', 'coordinates': [[[[-7...
3	503	South Shore	40.541474	-74.186647	{'type': 'MultiPolygon', 'coordinates': [[[[-7...
4	310	Washington Heights / Inwood	40.853896	-73.932686	{'type': 'MultiPolygon', 'coordinates': [[[[-7...

The Neighborhood in this project is composed of 55 sub-boroughs within 5 boroughs.



Hispanic Population data

This data consists of a simple csv file showing the number of Hispanic/Latino population (the client target of our stakeholders) and its percentage for each sub-borough. The corresponding data frame looked like this:

	Sub-borough Area	Hispanic/Latino	Percentage_hispanic/Latino
0	Mott Haven / Hunts Point	108659	0.6750
1	Morrisania / East Tremont	110555	0.6320
2	Highbridge / S. Concourse	94638	0.6548
3	University Heights / Fordham	95882	0.6832
4	Kingsbridge Heights / Mosholu	95082	0.7126

Socio-Economic data

This data consists also of a simple csv file showing the Labor force participation rate, the Unemployment rate, the Median household income, and the Median rent for each neighborhood.

	Sub-Borough name	Labor force participation rate	Unemployment rate	Median household income	Median rent, all
0	Astoria	0.710674	0.046137	67647.98474	1669.238634
1	Bay Ridge	0.637547	0.047335	69988.79130	1477.653247
2	Bayside / Little Neck	0.602683	0.048616	71492.94040	1843.499810
3	Bedford Stuyvesant	0.639627	0.047344	52896.92904	1239.190585
4	Bensonhurst	0.615441	0.060951	54513.17598	1394.089409

Foursquare

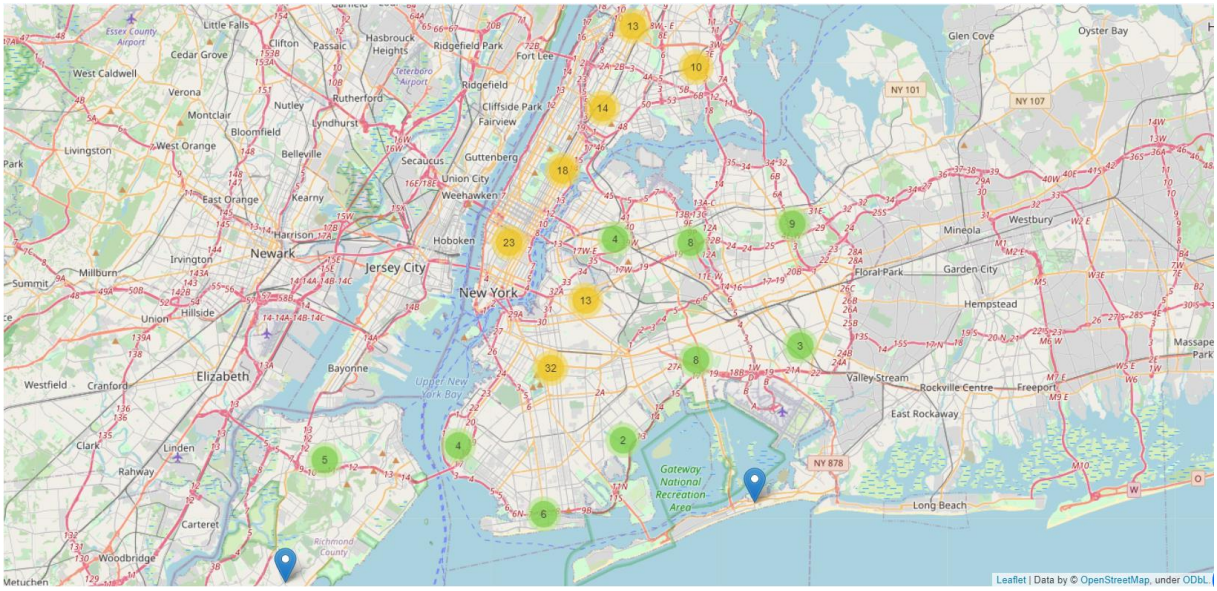
Firstly, I interrogated the Foursquare API for all the venues across New York City considering the coordinates of the center of each area and without any condition yet on the categories.

To pull this data out of the API, I needed to configure a few parameters, such as ignoring opening times. I set to 500 the limit to the number of venues returns in the response.

At the end we get a large (5360 rows) list of all the venues in New York City that was converted into a data frame. But I needed to take from that the ones that might be relevant to this project: sports or fitness center. For that, I screened out the ‘venue category’ to select only the ones that correspond to: Dance Studio, Gym, Gym / Fitness Center, Athletics & Sports, Yoga Studio. This reduced the venue list down to 199 rows (fairy small compared to the total number of venues).

I want to ensure that the list contains unique venues as there is some scope for the API to pass the same venue a few times due to the overlapping of the radius of exploration of each area. So, I decided to drop all duplicates based on the location coordinates. This reduced the venue list down to 174 rows (quite small compared to the total number of venues).

I then wanted to see where in New York City this was all located to see if the venues were spread enough to continue the analysis despite the low number. I mapped the data as clustered marker on a folium map.



To pull the number of possible existing 'Zumba Center', I screen out the 'Venue name' if any contain the word *Zumba*. It returns zero result. So, in a second attempt, I run an API to get the tips of my selected venues and to find if the tips mention anything related to Zumba. But also, it returns zero result. Considering this, I assumed that there was any 'Zumba Center' in New York City.

Methodology

In this project we will direct our efforts on detecting areas of New York City that have **low Sports center density**, particularly those with low number of Gym, Fitness, Recreation Center, Dance and Yoga Studio as mentioned.

In first step we have identified the venues of our interests with the required data: location and category (according to Foursquare categorization).

Second step in our analysis will calculate and explore **'Sports center density'** across different areas of New York City: we will take into consideration locations with ***no more than two sports centers in radius of 1000 meters***. Note that we already saw that *there is no Zumba Center in radius of 2500 meters* (radius used for making the API call in Foursquare*)*. We will use **Choropleth** to identify a few promising areas with low number of sports centers in general in vicinity and focus our attention on those areas.

In third step we will focus on most promising areas and within those we want to define locations with **high Hispanic population density** as those are our customer target. Again, we will use **Choropleth** to reduce the number of promising areas.

In the final step we will create clusters (using **k-means clustering**) using location and socio-economic data to group all the areas. From that we will look specifically the most promising areas from the previous step to identify the optimal neighborhood to propose to our stakeholders for the opening of Zumba place.

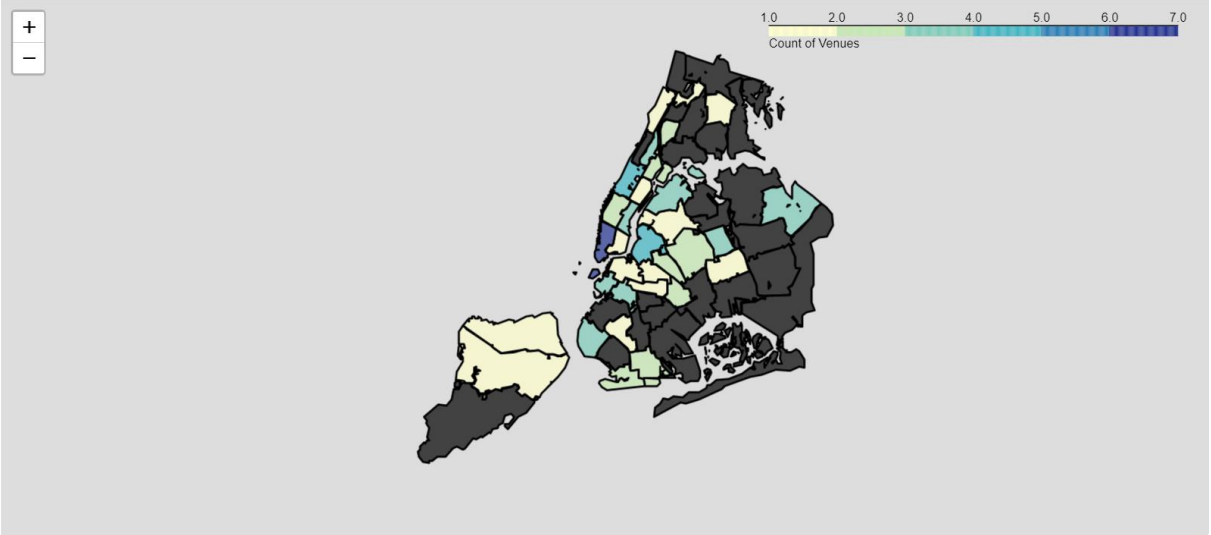
Analysis

Analysis of venues data

I started performing the analysis on the venues data that we got previously.

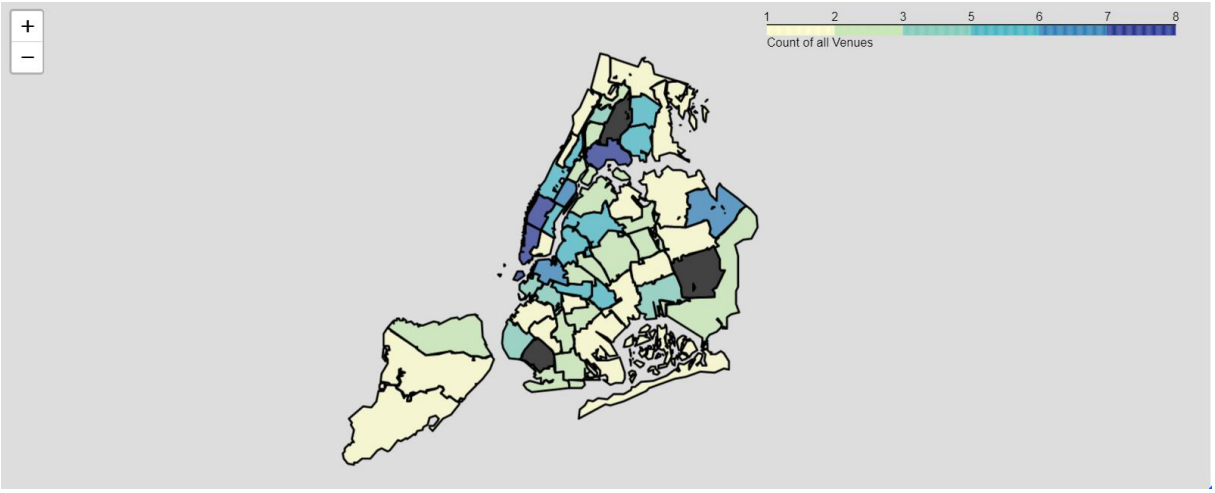
Firstly, I calculated the distance between each venue and its corresponding centroid (Distance in meters). And after, I took the venues that have distance less than 1000m from the center of its

corresponding area. Then I counted the venues for each area and create a shortlist of sub-boroughs. I plotted it on Choropleth as shown below.



The resulted the sub-boroughs that have maximum 2 sports centers located within a radius of 1000m from the center of the area are 21.

But to refine this, I took it a bit further by increasing the scale: radius of 2500m. I counted the number of venues in each sub-borough using the radius defined in the Foursquare API call to identify the area that has less than two sports centers. Then I plotted it again on Choropleth as shown below.



This resulted in 18 areas having less than or equal to two sports venues nearby.

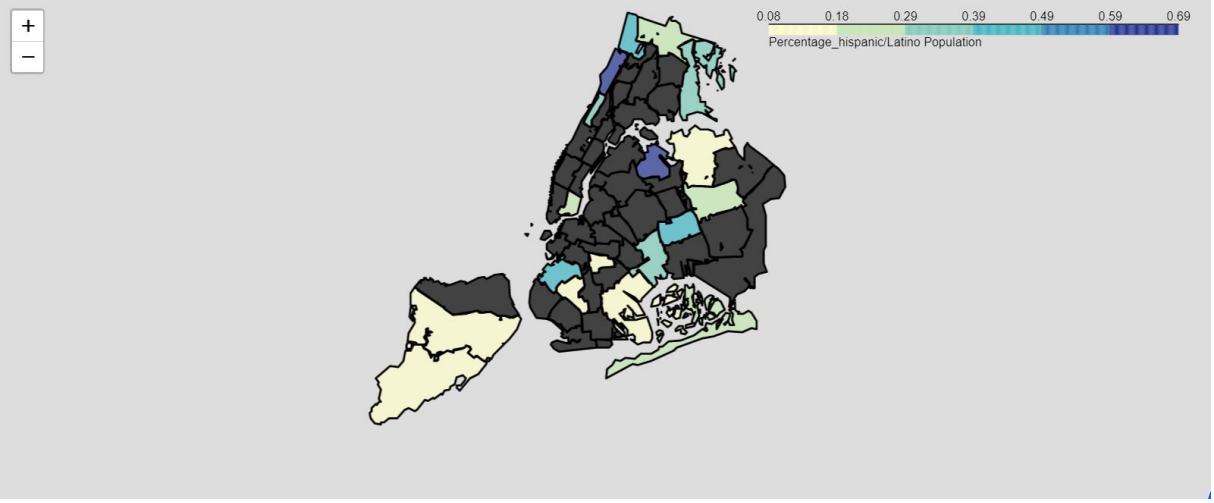
I added those areas to a shortlist.

At this point the number of our promising areas is still high.

	Sub-borough_name
0	Washington Heights / Inwood
1	Sunset Park
2	South Shore
3	Borough Park
4	Rockaways
5	Lower East Side / Chinatown
6	Jackson Heights
7	Morningside Heights / Hamilton Heights
8	Kew Gardens / Woodhaven
9	Williamsbridge / Baychester
10	Flushing / Whitestone
11	Flatlands / Canarsie
12	East New York / Starret City
13	Hillcrest / Fresh Meadows
14	Riverdale / Kingsbridge
15	South Crown Heights
16	Throgs Neck / Co-op City
17	Mid-Island

Analysis of Population data

The next step is to use these reduced area candidates to see where we have the highest number of Hispanic/Latino population (which are the customer target of our stakeholder). Look in the following Choropleth the results.

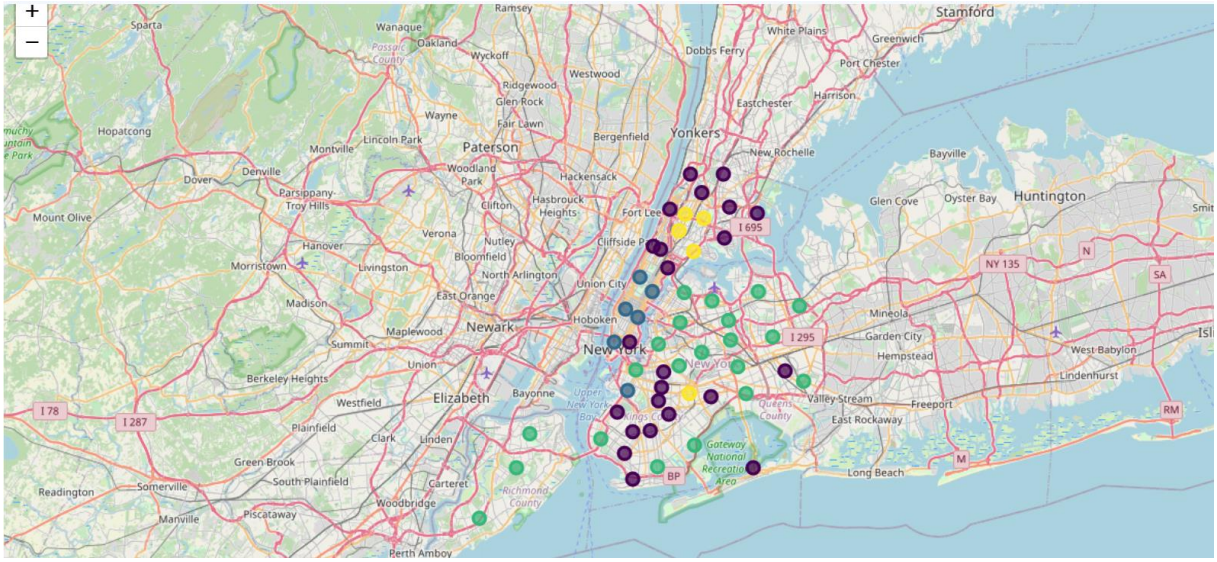


There are two sub-boroughs that have high percentage in Hispanic/Latino population (close to 70%) which are: **Washington Heights / Inwood** and **Jackson Heights**. A third option can be considered with ~48% of Hispanic/Latino population: **Riverdale / Kingsbridge.**

Analysis of socio-economic data

Finally, for the project I performed some data clustering, this was intended to confirm our previously chosen areas, based on other Socio-economic factors. Have in mind that we would rather choose neighborhoods where unemployment rate is acceptable, average income is medium to high and where average rent price is acceptable.

I started with setting the index of the data frame to be the name of the area. Then after performing some pre-processing I normalized the data values and used K-means clustering to see whether it found similarity. I then added the grouping results to the NYC neighborhood data frame and mapped the results.



The average values of each feature for each group are summarized in this data frame:

Clus_km	Labor force participation rate	Unemployment rate	Median household income	Median rent, all
0	0.618321	0.074625	50544.425321	1266.927045
1	0.727545	0.037209	127262.970867	2310.913803
2	0.644337	0.047801	71417.577654	1564.953681
3	0.576035	0.154471	26140.806926	1052.292926

The group can be defined as follow:

CLUSTER 1: Medium to High unemployement rate, Medium to Low income, medium to Low rent price
CLUSTER 2: Low unemployement rate, High income, High rent price
CLUSTER 3: Medium to Low unemployement rate, Medium to High income, medium rent price
CLUSTER 4: Low unemployement rate, High income, High rent price

Retaking the 3 areas selected in the previous step, we can see that:

- **Washington Heights / Inwood** and **Riverdale / Kingsbridge** are from the same group where unemployment is medium to high, median household income is Low, and where rent price is low.
- **Jackson Heights** is part of the group where unemployment rate is medium to Low, median household income is medium to high income, and where rent price is acceptable.

Results and Discussion

The results of the analysis of the venues and the percentage of Hispanic/Latino population show that there are three potential areas that can fit the profile for the scenario:

- **Washington Heights / Inwood** and **Jackson Heights**. They are two areas of high concentration of Hispanic/Latino population on Manhattan and Queens boroughs. They have a total number of 2 fitness/ gym/dance/yoga related venues within 2500m of its centers. I think either would be suitable candidates for locating a Zumba Center. Besides, we have seen that no Zumba centers exist in the whole areas.
- **Riverdale / Kingsbridge**. It is an area that has also only 2 fitness/ gym/dance/yoga related venues within 2500m of its centers. Close to half of its population is of Hispanic or Latino origin.

But the analysis on socio-economic data allowed us to bring down to one the optimal sub-borough: **Jackson Heights.** Yes, because our stakeholders need areas where they can ensure that people will consume their services and areas where the price of rent will be optimal. These conditions are fulfilled by the sub-borough **Jackson Heights.**

The analysis allowed us to observe that New York City has generally few numbers of fitness/ gym/dance/yoga related venues. This can be considered as opportunities to create business related to those activities.

Concerning the fact that any Zumba centers were found, I think its more related to the ways the Foursquare APIs are constructed, because when I searched in Foursquare City Guide, I found few places that says explicitly in the venue name about Zumba, while none of those appeared to the results of my 'explore API call'. Of course, these might be located outside of the radius specified because I do not believe that there is at all any Zumba place in New York City even one. Also, most of the results I got from Foursquare City Guide had the word Zumba in their tips and again those did not appear in my Tips API call. This is just to comment that there might be some rooms for amelioration in the API engine of Foursquare for Developers.

Conclusion

In this study I analyzed the publicly available data from New York City to see if I could use it to determine a good location for a Zumba Club. I think that it has achieved its aim and would hope that improvements can be done on the Foursquare API to be able to get accurate data and perform more accurate analysis. I think that this project can be improved by including other factors such as proximity to office buildings to give better results.