Capstone Project

Applied Data Science Capstone by IBM/Coursera

Introduction: Business Problem

In this project we will try to find an optimal location for a Gym. Specifically, this report will be targeted to stakeholders interested in opening an business in New York City, USA.

Since there are lots of Gym's in New York we will try to detect locations that are not already crowded with business. We are also particularly interested in areas with no business in vicinity. We would also prefer locations as close to city center as possible, assuming that first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on this 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 decission are:

- number of existing business in the neighborhood (any type of business)
- number of and distance to business in the neighborhood, if any
- distance of neighborhood from city center

We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

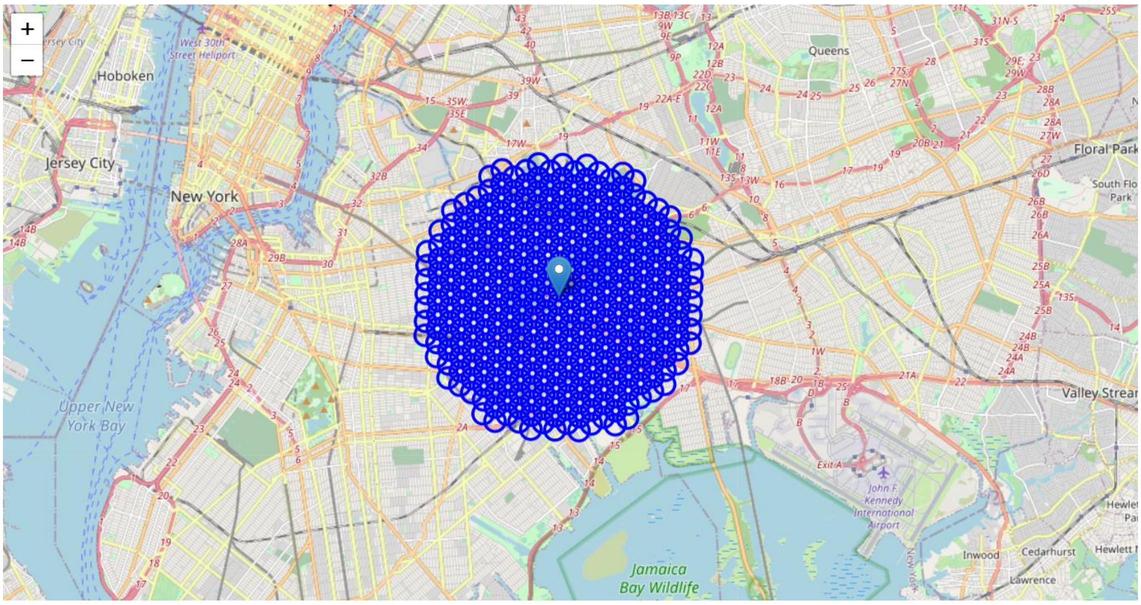
Following data sources will be needed to extract/generate the required information:

- centers of candidate areas will be search and choise and approximate addresses of centers of those areas will be obtained using Foursquare API
- number of Gym's and their type and location in every neighborhood will be obtained using Foursquare API
- coordinate of New York center will be obtained using Foursquare API of well known New York location

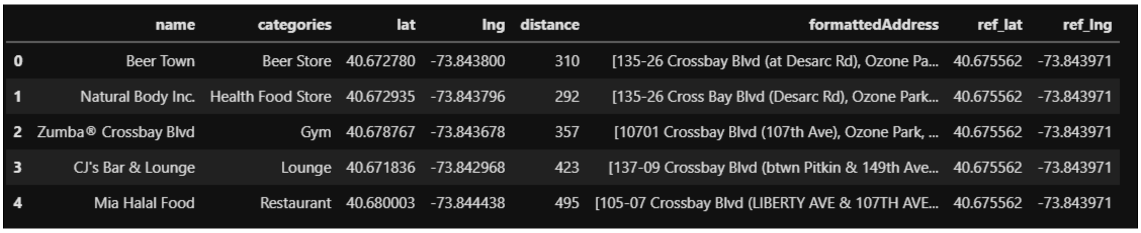
Neighborhood Candidates

Let's create latitude & longitude coordinates for centroids of our candidate neighborhoods. We will create a grid of cells covering our area of interest which is aprox. 12x12 killometers centered around New York city center.

Let's first find the latitude & longitude of New York city center, using specific, well known address and Foursquare API.

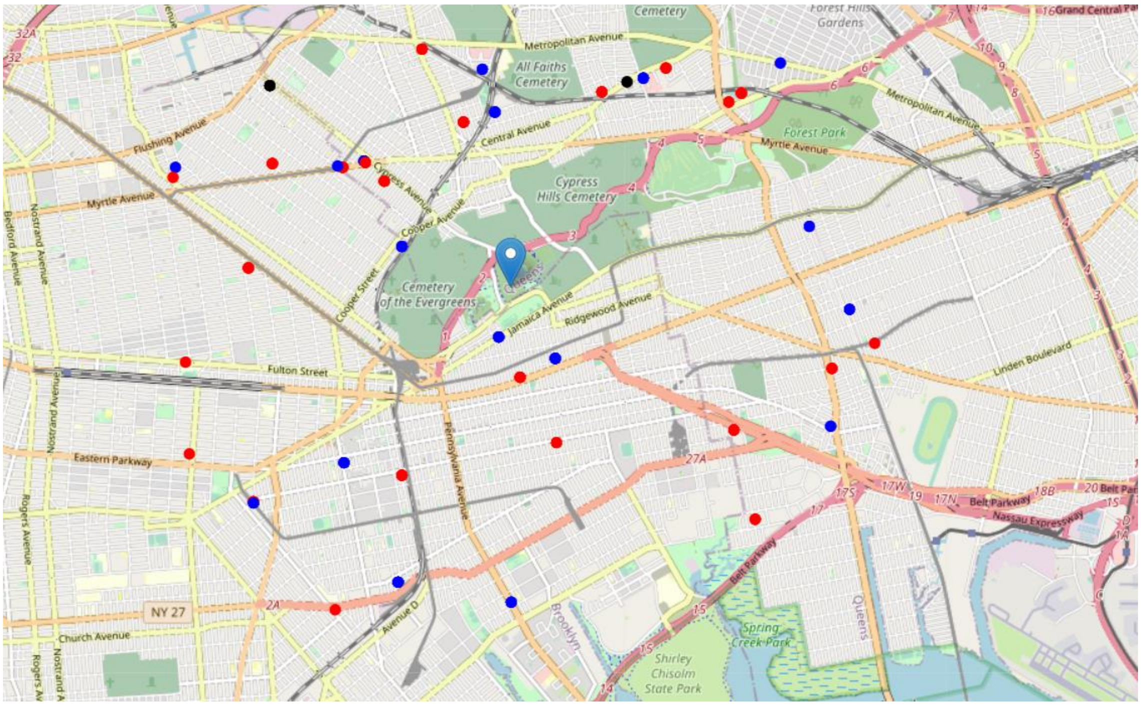


All the locations and observation in this points.



	name	categories	lat	lng	distance	formattedAddress	ref_lat	ref_lng
0	Beer Town	Beer Store	40.672780	-73.843800	310	[135-26 Crossbay Blvd (at Desarc Rd), Ozone Pa...	40.675562	-73.843971
1	Natural Body Inc.	Health Food Store	40.672935	-73.843796	292	[135-26 Cross Bay Blvd (Desarc Rd), Ozone Park...	40.675562	-73.843971
2	Zumba® Crossbay Blvd	Gym	40.678767	-73.843678	357	[10701 Crossbay Blvd (107th Ave), Ozone Park, ...	40.675562	-73.843971
3	CJ's Bar & Lounge	Lounge	40.671836	-73.842968	423	[137-09 Crossbay Blvd (btwn Pitkin & 149th Ave...	40.675562	-73.843971
4	Mia Halal Food	Restaurant	40.680003	-73.844438	495	[105-07 Crossbay Blvd (LIBERTY AVE & 107TH AVE...	40.675562	-73.843971

Position, of all Gym.



Looking good. So now we have all the restaurants in area within few kilometers from Highland Park, and we know which ones are Gym! We also know which restaurants exactly are in vicinity of every neighborhood candidate center.

This concludes the data gathering phase - we're now ready to use this data for analysis to produce the report on optimal locations for a new Italian restaurant!

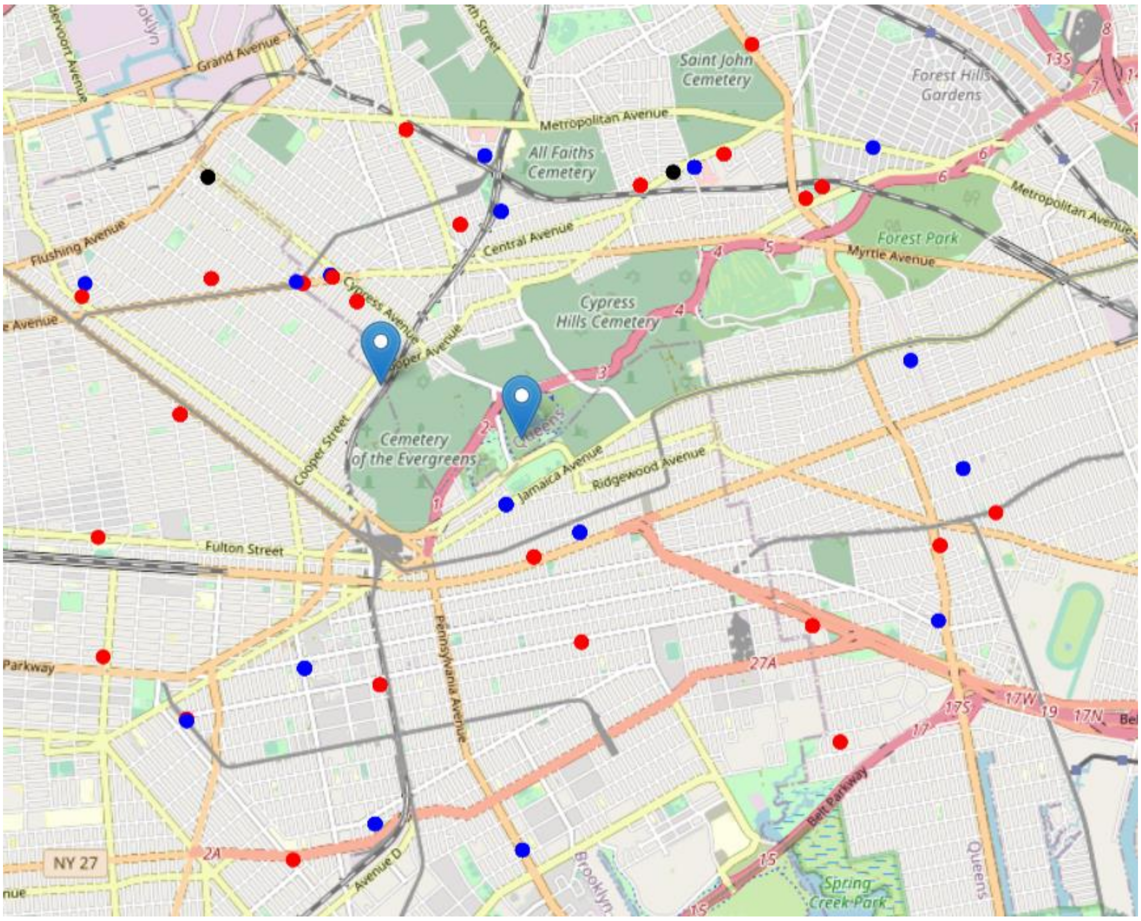
Methodology

In this project we will direct our efforts on detecting areas of New York that have low Gym density, particularly those with low number of Gym's. We will limit our analysis to area ~6km around city center.

In first step we have collected the required data: location and type (category) of every restaurant within 6km from New York center (Highland Park). We have also identified Gym (according to Foursquare categorization).

Second step in our analysis will be calculation and exploration of 'gym density' across different areas of New York - we will use heatmaps to identify a few promising areas close to center with low number of restaurants in general (and no Gym in vicinity) and focus our attention on those areas.

In third and final step we will focus on most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into consideration locations with no more than two gym's in radius of 250 meters, and we want locations without Gym's in radius of 400 meters. We will present map of all such locations but also create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.



Results and Discussion

Our analysis shows that although there is a great number of gym in New York center (~2000 in our initial area of interest which was 12x12km around Highland Park), there are pockets of low gym density fairly close to city center. Highest concentration of gym was detected north and south from Highland Park, so we focused our attention to areas est and west, corresponding to boroughs Ocean Hill and Knews Gardens. Another borough was identified as potentially interesting (Ricmond Hill and Woodhaven, west from Highland Park), but our attention was focused on Ocean Hill and Knews Gardens which offer a combination of popularity among tourists, closeness to city center, strong socio-economic dynamics and a number of pockets of low gym density.

After directing our attention to this more narrow area of interest (covering approx. 5x5km south-east from Highland Park) we first created a dense grid of location candidates (spaced 100m appart); those locations were then filtered so that those with more than two business in radius of 250m and those with an Gym closer than 400m were removed. Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Conclusion

Purpose of this project was to identify New York city areas close to center with low number of business (particularly Gym) in order to aid stakeholders in narrowing down the search for optimal location for a new Gym. By calculating gym density distribution from Foursquare data we have first identified general boroughs that justify further analysis (Ocean Hill and Knews Gardens), and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby business. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decission on optimal Gym location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.