**Capstone Project - The Battle of New York Gyms**

**Author: Bara Pencikova**

**October 2019**

**INTRODUCTION:**

The City of New York, or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 784 km2, New York is also the most densely populated major city in the United States. New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. (source: <https://en.wikipedia.org/wiki/New_York_City>)

**BUSINESS GOAL DESCRIPTION:**

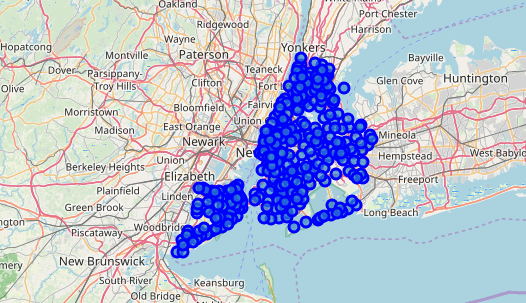
My client is a businessman who spends a lot of time travelling. One of his future business trips will be a long-term stay in New York. Apart from many expected hours sitting at the office my client wants also to stay in a good physical shape, so he wants me to analyze and recommend the best part of New York according to the density of fitness centers, gyms, martial arts dojos and similar facilities. My client doesn’t mind the accommodation, the priority is to have good gyms and/or fitness centers within reach.



**DATA**

For this particular assignment I used the Fourquare data, accessed through API <https://developer.foursquare.com/docs/resources/categories> the category "Gym / Fitness Center" which has this ID: "4bf58dd8d48988d175941735".

For clustering and segmenting I have used the dataset which contains the New York borroughs and neighborhoods with their latitudes and longitudes. The dataset can be found in the NYU Spatial Data Repository pages: <https://geo.nyu.edu/catalog/nyu_2451_34572>



**METHODOLOGY**

I have downloaded the data from NYU Spatial Data Repository (GeoJSON file with geolocations) and converted to Pandas dataframe. After that, using the FourSquare data, I explore neighborhoods of New York borroughs for their most frequent gyms and sporting facilities, converting them to OneHot dataframe and merge with the capital cities data.

Using Machine Learning classification method K-Means, I will classify NY neighborhoods into clusters. Finally, each cluster is inspected for typical venue categories that I will present using Word Cluster visualization method.

**Getting data**

For my assignment I will need the New York borroughs and their associated neighborhoods. I've downloaded the GeoJSON file from the NYU Spatial Data repository: <https://geo.nyu.edu/catalog/nyu_2451_34572>

I will transform the downloaded file into the pandas dataframe:

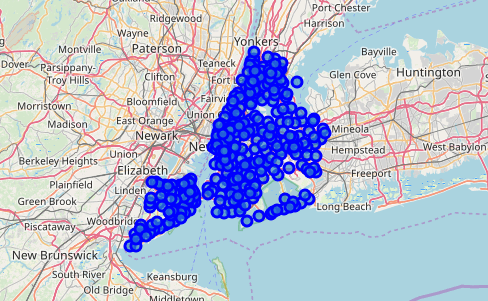
​

|  | **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 |
| **5** | Bronx | Kingsbridge | 40.881687 | -73.902818 |
| **6** | Manhattan | Marble Hill | 40.876551 | -73.910660 |
| **7** | Bronx | Woodlawn | 40.898273 | -73.867315 |
| **8** | Bronx | Norwood | 40.877224 | -73.879391 |
| **9** | Bronx | Williamsbridge | 40.881039 | -73.857446 |

Getting the New York coordinate:

​'The geograpical coordinate of New York are 40.7127281, -74.0060152.'

Creating the map of New York and it's neighborhoods, using the folium library:

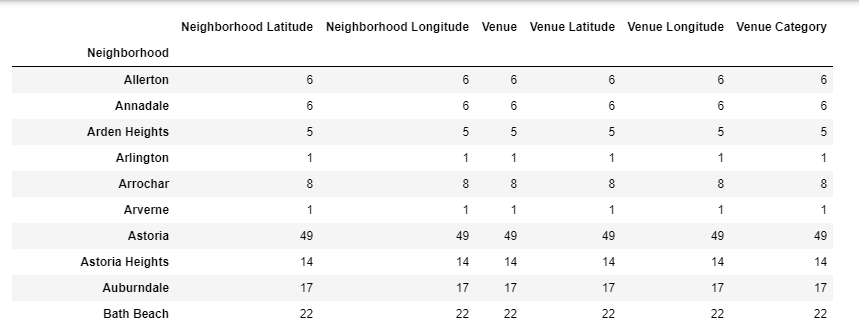


**Choosing the FourSquare Category:**

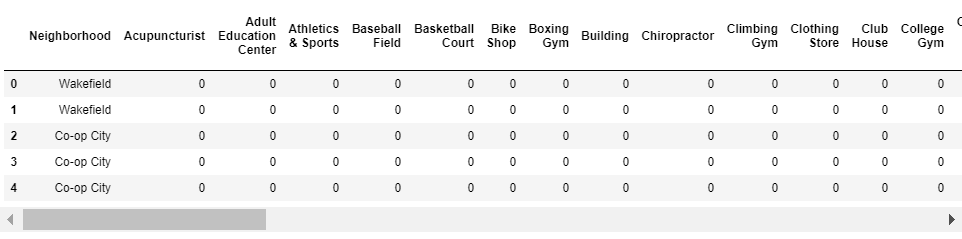
Gyms / Fittnes Centers = 4bf58dd8d48988d175941735

|  | **Neighborhood** | **Neighborhood Latitude** | **Neighborhood Longitude** | **Venue** | **Venue Latitude** | **Venue Longitude** | **Venue Category** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | Wakefield | 40.894705 | -73.847201 | Mount St Michael's Academy (track) | 40.898688 | -73.840709 | Track |
| **1** | Wakefield | 40.894705 | -73.847201 | T Gym & Studio | 40.901266 | -73.838078 | Gym / Fitness Center |
| **2** | Co-op City | 40.874294 | -73.829939 | My Gym | 40.872681 | -73.829400 | Gym / Fitness Center |
| **3** | Co-op City | 40.874294 | -73.829939 | 24 Hour Fitness | 40.867818 | -73.824984 | Gym / Fitness Center |
| **4** | Co-op City | 40.874294 | -73.829939 | bally's total fitness bartow ave | 40.867290 | -73.832603 | Gym / Fitness Center |

Group by Neighborhood:



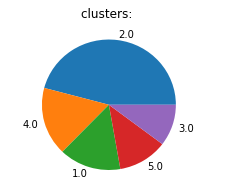
One-Hot Encoding (categorical values are converted into the numerical value for entry in the dataset):



Definning the function that returns the most common venues:



Exploring clusters:



ny\_merged["Cluster Labels"].value\_counts()

2.0 138

4.0 51

1.0 45

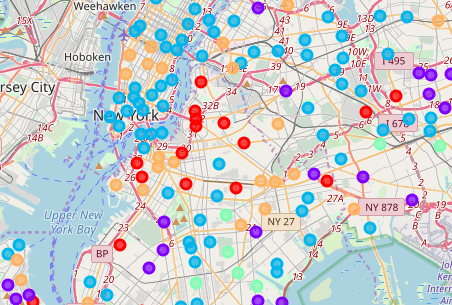
5.0 36

3.0 31

Putting the data into the folium map, each cluster differentiated by the color.

**RESULTS**

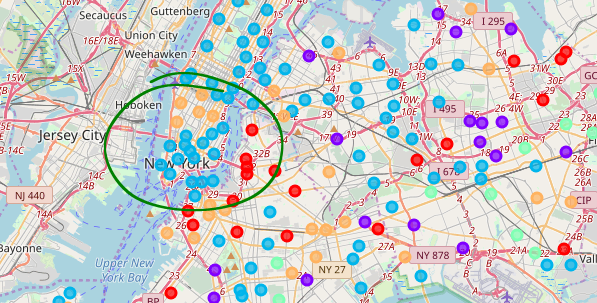
Using the K-Means methodology, the New York neighborhoods were divided into five clusters and displayed on Folium Map (Python). With geolocations added, various colors indicate belonging to a cluster.



Let's use World Cloud to describe the most common venues, that were discovered using the described method and classified into the same cluster by K-Means (Machine Learning algorithm).



As you can see, each cluster is slightly different in terms of most common venues, our client can easily see, what types of sporting activities are most frequently there. However, the original request was to find out the highest density of gyms and fitness centers in New York. As we can see, Cluster 2 is not only the biggest one according the number of venues, it's also the densest one. That's why I would recommend for my client to move to Lower Manhattan, where the density of gyms is the biggest.



**DISCUSSION**

K-Means method works well for classification of neighborhoods. What is debatable are the FourSquare data, as for the project I excluded venues that fall into category of Athletics\_and\_sports, which maybe should be taken into the consideration (category ID is mentioned in the code, it can be used anytime if needed). Category "Gyms / Fitness Centers" is considered. From this dataset, the most common venues categories are used for clustering hence comparing the neighborhoods.

To claim an objective classification you need to understand the data and it takes little experimenting find out the best K for K-Means algorithm, but that's well known problem.

**CONCLUSION**

FourSquare API is very powerful if used correctly, its advantage is that people use it all around the world. I prefer discussion thread for every venue to learn references when I am looking for trying something new.

Using the described method, I learned Manhattan is the borrough with the highest density of gyms / fitness centers / martial arts dojos, which is what our client is looking for. In Manhattan, in particular Lower Manhattan, Cluster 2 venues and Cluster 4 venues are populated, whilst we can see from the WordCloud visualizations, Gyms and Fitness Centers are amongst the most common values (unlike Cluster 3, where the most common value is Gym).

The goal of the project was satisfied, but there is for sure a lot of space for improvements in this area