Capstone Project Report

Finding the best place to open a New Fitness Center

Deep Chatterjee
Feb 12,2020

Introduction /Business Problem

New York City was home to nearly 8.5 million people in 2014, accounting for over 40% of the population of New York State and a slightly lower percentage of the New York metropolitan area, home to approximately 23.6 million. Over the last decade the city has been growing faster than the region.As New York is also New York City has been described as the cultural, financial and media capital of the world People now a days are very conscious of their health, hence there is a greater need of places where people can make themselves fit at whatever times of their suiting.

The aim of this project is to discover where is the optimal place to build a gym in the New York City. To answer this question we will analyse which districts have a gym and if there are any similarities between them. We will also see which districts have the least amount of gyms.

The audience of this report should be people interested in starting a new gym in the New York City.

Data

For this project we need the following data :

New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude.

Data source : https://cocl.us/new_york_dataset

Description: This data set contains the required information. And we will use this data set to explore various neighbourhoods of new york city.

Data source: Fousquare API

Description: By using this api we will get all the venues in each neighbourhood. We can filter these venues to get only gym/fitness-center.

GeoSpace data
Data source : Using geopy library to get the location data

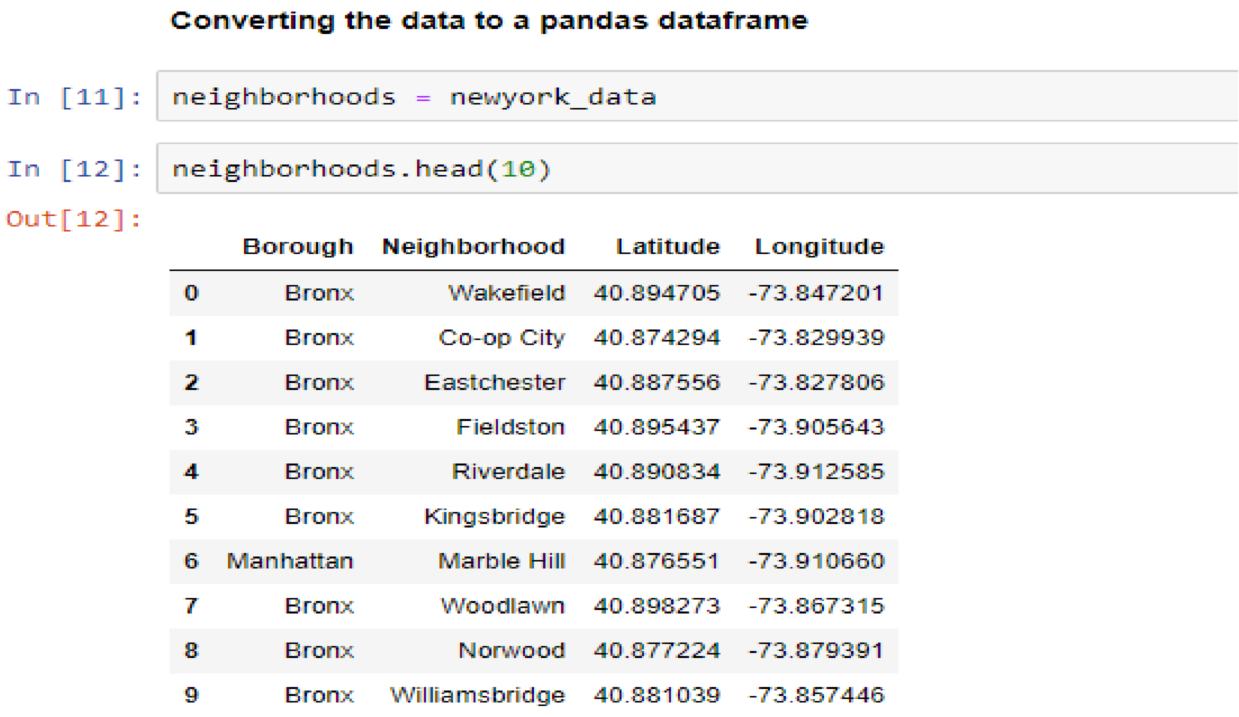
Description : By using this geo space data we will get the New york Borough boundaries that will help us visualize choropleth map.

Code:

1. Data Description and EDA

New York City is often referred to collectively as ***the five boroughs namely:*** Brooklyn, Queens, Manhattan, Staten Island and the Bronx, and in turn, there are hundreds of distinct neighborhoods throughout the boroughs

After data cleaning, a table with New York City 's boroughs, neighborhoods and their respective latitude and longitude was obtained. The table with its first 10 observations after data cleaning and converting into a pandas dataframe for further analysis:



Converting the data to a pandas dataframe

```
In [11]: neighborhoods = newyork_data
```

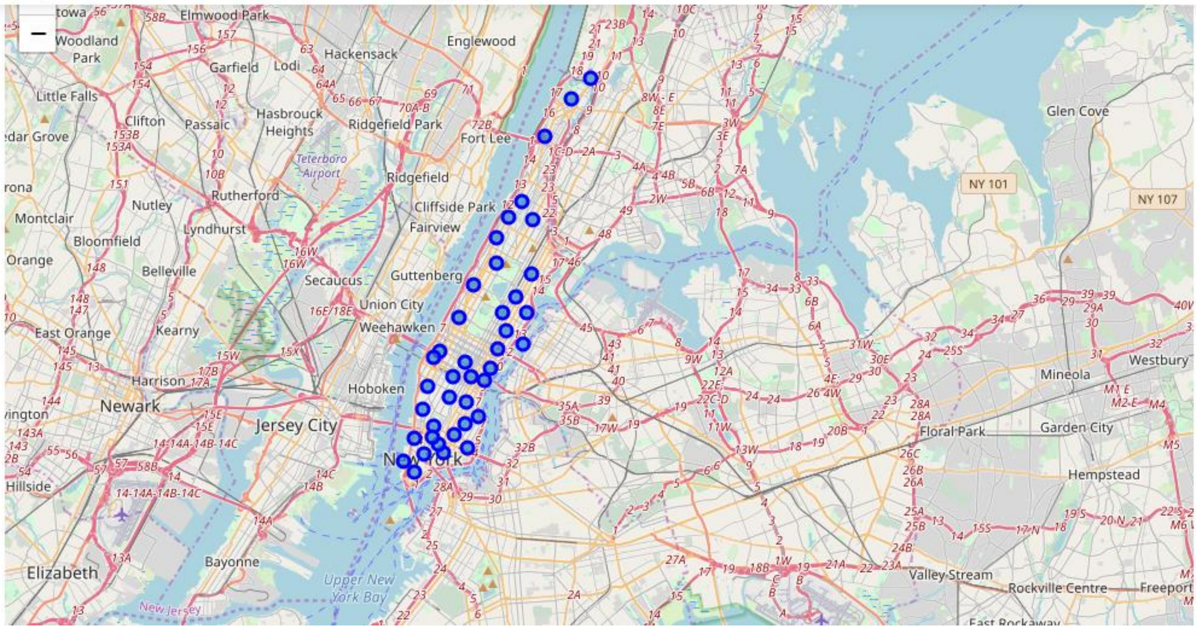
```
In [12]: neighborhoods.head(10)
```

Out[12]:

	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

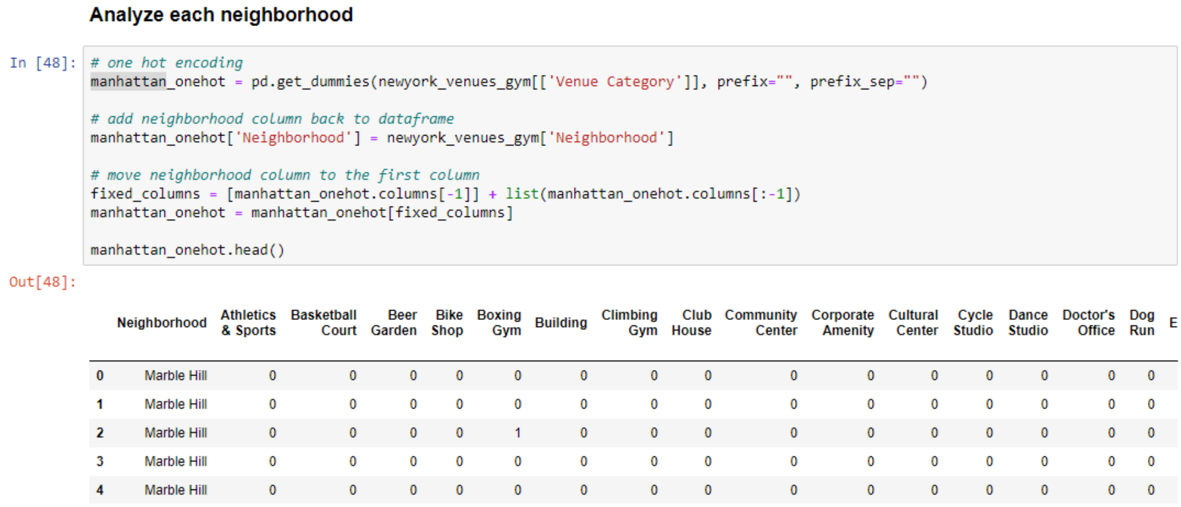
2. Map Visualization and clustering

With help of Folium it was possible to visualize Manhattan Area of New York City and show the neighborhoods:



3. Converting DataFrame for Machine Learning

Analysing each neighbourhood and converting into a dataframe using One-Hot encoding method for further analysis:



Analyze each neighborhood

In [48]:

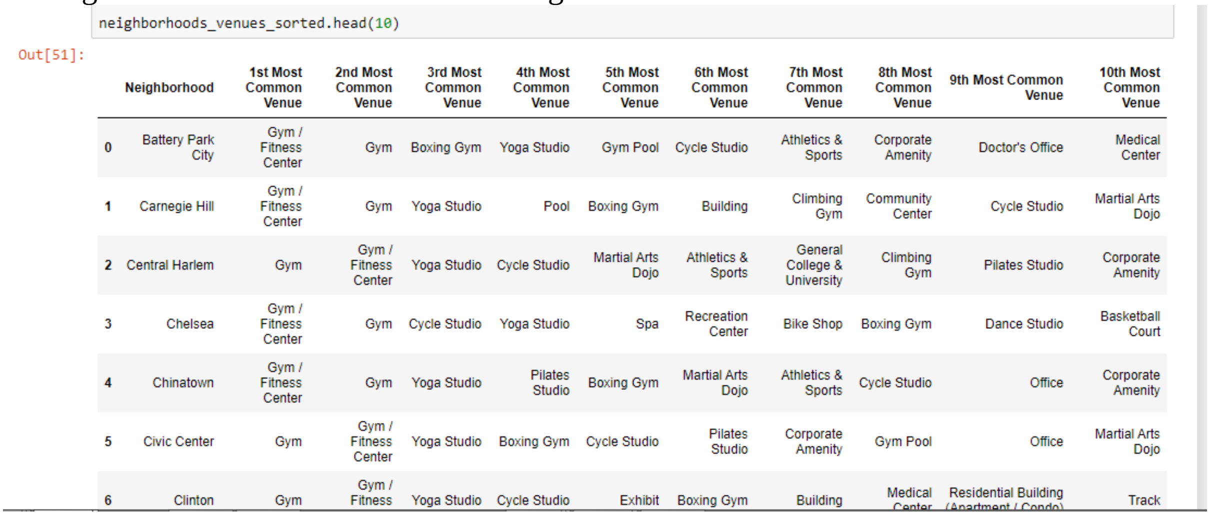
```
# one hot encoding
manhattan_onehot = pd.get_dummies(newyork_venues_gym[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
manhattan_onehot['Neighborhood'] = newyork_venues_gym['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [manhattan_onehot.columns[-1]] + list(manhattan_onehot.columns[:-1])
manhattan_onehot = manhattan_onehot[fixed_columns]

manhattan_onehot.head()
```

Out[48]:

	Neighborhood	Athletics & Sports	Basketball Court	Beer Garden	Bike Shop	Boxing Gym	Building	Climbing Gym	Club House	Community Center	Corporate Amenity	Cultural Center	Cycle Studio	Dance Studio	Doctor's Office	Dog Run	E
0	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Marble Hill	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
3	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Marble Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

.

Sorting most common venues for clustering:



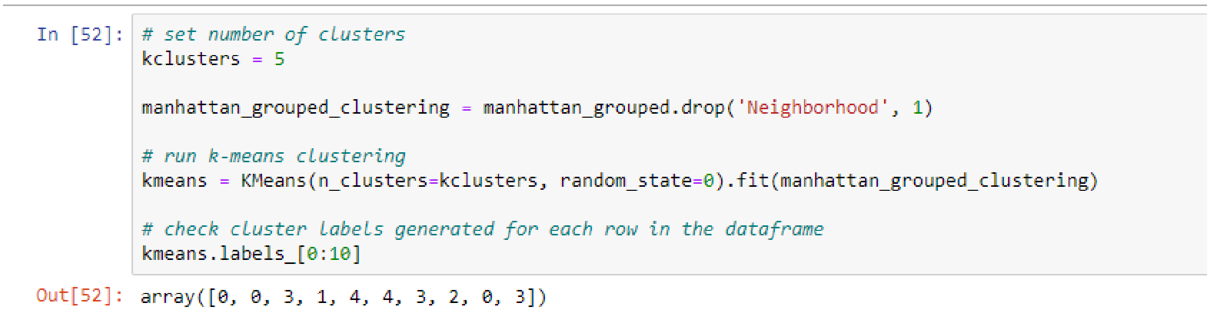
```
neighborhoods_venues_sorted.head(10)
```

Out[51]:

	Neighborhood	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Gym / Fitness Center	Gym	Boxing Gym	Yoga Studio	Gym Pool	Cycle Studio	Athletics & Sports	Corporate Amenity	Doctor's Office	Medical Center
1	Carnegie Hill	Gym / Fitness Center	Gym	Yoga Studio	Pool	Boxing Gym	Building	Climbing Gym	Community Center	Cycle Studio	Martial Arts Dojo
2	Central Harlem	Gym	Gym / Fitness Center	Yoga Studio	Cycle Studio	Martial Arts Dojo	Athletics & Sports	General College & University	Climbing Gym	Pilates Studio	Corporate Amenity
3	Chelsea	Gym / Fitness Center	Gym	Cycle Studio	Yoga Studio	Spa	Recreation Center	Bike Shop	Boxing Gym	Dance Studio	Basketball Court
4	Chinatown	Gym / Fitness Center	Gym	Yoga Studio	Pilates Studio	Boxing Gym	Martial Arts Dojo	Athletics & Sports	Cycle Studio	Office	Corporate Amenity
5	Civic Center	Gym	Gym / Fitness Center	Yoga Studio	Boxing Gym	Cycle Studio	Pilates Studio	Corporate Amenity	Gym Pool	Office	Martial Arts Dojo
6	Clinton	Gym	Gym / Fitness	Yoga Studio	Cycle Studio	Exhibit	Boxing Gym	Building	Medical Center	Residential Building (Apartment / Condo)	Track

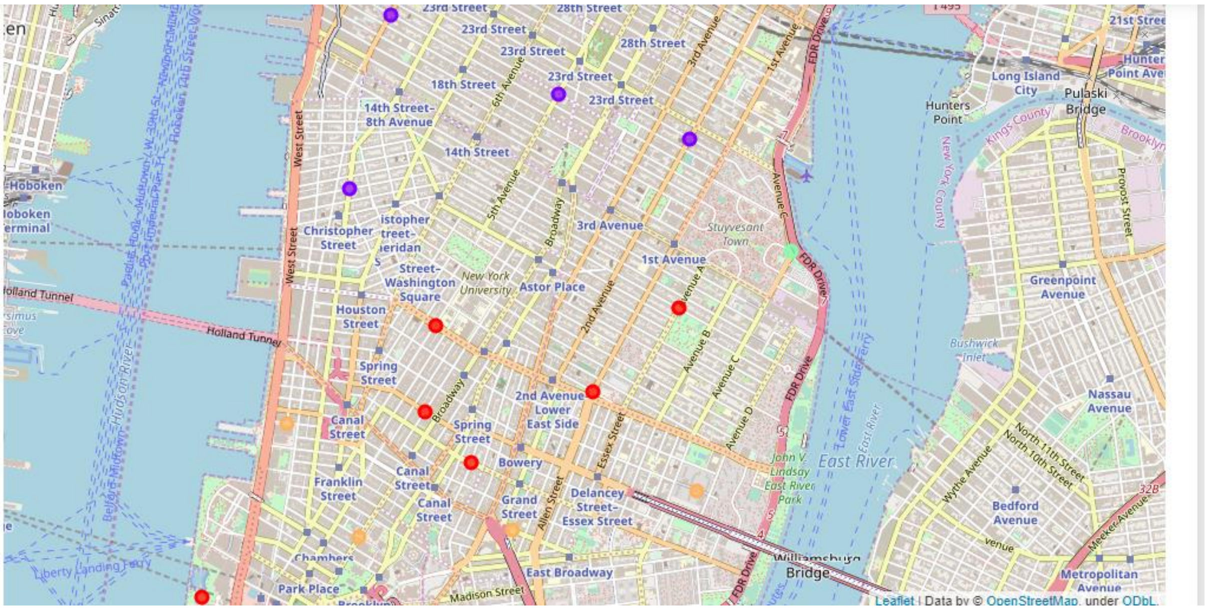
4. **Machine Learning**:

Here I am using K-Means Clustering algorithm to cluster similar venues to find out areas where there is high competition and which are the areas where we can open new Center.



```
In [52]: # set number of clusters
         kclusters = 5

         manhattan_grouped_clustering = manhattan_grouped.drop('Neighborhood', 1)

         # run k-means clustering
         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(manhattan_grouped_clustering)

         # check cluster labels generated for each row in the dataframe
         kmeans.labels_[0:10]

Out[52]: array([0, 0, 3, 1, 4, 4, 3, 2, 0, 3])
```

5. Visualization of Clusters:

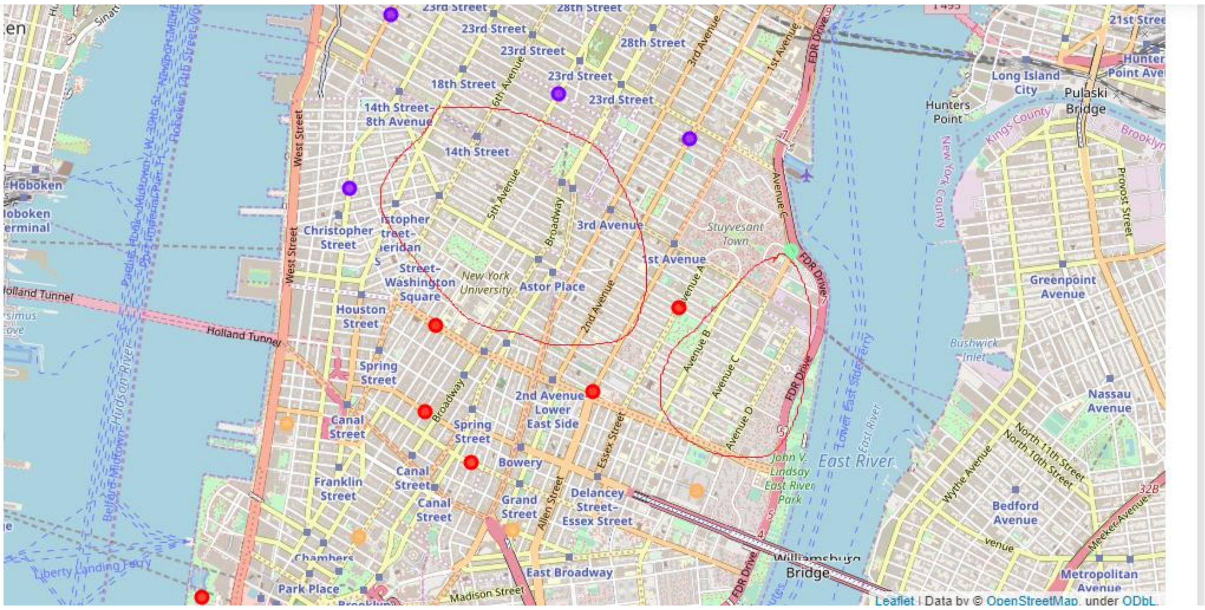


Discussion:

- There is high competition in Lower Manhattan so it is very risky to open business in these areas.
- Soho and East Village have potential to open new centres.
- The above analysis is performed on limited data. Hence, it may be not very accurate. So, if good amount of data is available there is scope to come up with better results.

Conclusion:

Below Circled areas can be considered as to open new Gym/Fitness Center


23rd Street
28th Street
23rd Street
28th Street
23rd Street
23rd Street
18th Street
23rd Street
14th Street-8th Avenue
14th Street
West Street
Hoboken
Hoboken Terminal
Christopher Street
5th Avenue
Broadway
3rd Avenue
Stuyvesant Town
1st Avenue
Washington Square
New York University
Astor Place
2nd Avenue
Holland Tunnel
Houston Street
Spring Street
Avenue A
Avenue B
Avenue C
Avenue D
2nd Avenue Lower East Side
Canal Street
Spring Street
Bowery
Franklin Street
Canal Street
Grand Street
Delancey Street-Essex Street
Essex Street
Allen Street
John V. Lindsay East River Park
East River
FDR Drive
Williamsburg Bridge
Chambers
Park Place
East Broadway
Madison Street
Liberty Landing Ferry
Hudson River
Long Island City
Hunters Point
Pulaski Bridge
Greenpoint Avenue
Nassau Avenue
Bushwick Inlet
North 11th Street
North 10th Street
Wythe Avenue
Bedford Avenue
Metropolitan Avenue
Provost Street
21st Street
Hunters Point Avenue
Leaflet | Data by © OpenStreetMap, under ODbL.