Exploring New York City Sports Venues

Thomas Jacobs

13 June 2020

Content

Figures

Tables

1. Introduction

1.1. Background

Nowadays more people are health conscious and take part in various sport activities, not only to reach their ultimate physical potential, but also for reasons such as self-improvement, health improvement, etc.

The sports nutrition and sports equipment market is booming as people are becoming more concerned of what they eat in order to stay healthy and the quality of sports equipment used to reach their peak. Competition is fierce between some of the major players in sports nutrition in North American market – Nestlé S.A, Glanbia plc, Abbott Nutrition Inc, Yakult Honsha Co Ltd, The Coca-Cola Company, Reckitt Benckiser Group Plc, GNC Holdings, and Monster Beverage Corporation, among others. Major players in the sports equipment arena include Adidas, Nike, New Balance, etc.

Therefore, it is critical for a small, start-up company in this industry to ensure that it targets its products at the right target audience for optimal impact.

New York City has a total of 5 boroughs and 306 neighborhoods as indicated in https://en.wikipedia.org/wiki/Boroughs_of_New_York_City.

1.2. Problem

The marketing team of the start-up is highly skilled and experienced in their craft but are not local to New York and are not familiar with the different neighbourhoods. Their marketing budget is constraint and they want to ensure that every cent is optimally invested to get maximum return.

1.3. Value proposition

The aim of the project is to assist the marketing team in understanding their target area in order to develop the best marketing strategy for their portfolio of products. New York City will be clustered into the optimal number of different sports clusters to understand the most prevalent sports associated with each cluster.

2. Data

2.1. Initial datasets

We are interested in what sports activities people engage in and also where the different sports venues are located.

We will use the dataset available from https://geo.nyu.edu/catalog/nyu_2451_34572 which lists all 5 boroughs and the neighborhoods that exist in each borough as well as the corresponding geographical coordinates (latitude and longitude) of each neighbourhood.

We also require access to the Foursquare location data in order to get a dataset of all the different venue categories in the Manhattan area. Foursquare location data API calls will be executed to populate a dataset with all the different venue categories in the different neighbourhoods of Manhattan within a radius of 500m from each neighbourhood geographical coordinates. The dataset will also contain the geographical coordinates for these different venues.

2.2. Cleaning the datasets

We are however only interested in the target audience with includes all sports and health related venues. The full venue categories dataset will then be filtered to only extract the venue categories that are related to sports activities, e.g. gyms, health shops, stadiums, yoga studios, etc.

The sports dataset will contain the neighbourhood name, venue name, venue category and geographical coordinates.

2.3. Exploring the datasets

The sports dataset will be analysed to get a breakdown of the different sports activities in the different neighboorhoods. This will also assist in understanding which sports activities are prevalent.

The sports dataset will be analysed to get a graph of the breakdown of the number of different sports and health related venue categories in New York City's different boroughs. The top 5 venue categories in Manhattan will then be explored to see where they are located.

The next step would be to conduct statistical analysis on the sports dataset to cluster the dataset into the optimal number of clusters that will be crucial in the targeted marketing objective.

3. Methodology

The following methodology was used to arrive at the optimal number of sports clusters (based on our target sports):

a) The New York City json file with borough and neighboorhod geographical coordinates was downloaded and wrangled.
b) The datasets were explored and neighboorhoods visualized using the Folium library.
c) The Foursquare API is used to explore the neighborhoods in each borough and get the different venues in each neigboorhood.
d) The target sports venues were extracted from the datasets and analysed.
e) The sports venue datasets are used to determine the optimum number of clusters for New York City.
f) Finally, the clusters are presented in tabular format and visualized as superimposed color-coded markers on the New York City map using the Folium library.

4. Results

In this section the obtained results for each step will be discussed and the clusters presented.

4.1. Download and Explore Dataset

The json file with the New York City data is downloaded. The neighbourhood data, contained in the features key, is extracted from the json file.

A dataframe with the column names, ['Borough', 'Neighborhood', 'Latitude', 'Longitude'] is created and the dataframe is populated by extracting the relevant data row by row from the neighbourhood data dictionary. The head of the resulting dataframe is indicated in Table 1.

Table 1: Dataframe with New York City data

	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

The dataset is checked to contain the expected number of boroughs and neighborhoods (5 boroughs and 306 neighborhoods).

4.2. Visualize the New York City Map with Neighboorhoods

The next step is to slice the original neighbourhoods dataframe into new dataframes for each of the 5 boroughs, The Bronx, Brooklyn, Manhattan, Queens, and Staten Island.

The geopy library is used to get the latitude and longitude values of New York City and its boroughs as indicated in Table 2.

Table 2: Geographical coordinates

Place name	Latitude	Longitude
New York City	40.7127281	-74.0060152
The Bronx	40.8466508	-73.8785937
Manhattan	40.7896239	-73.9598939
Brooklyn	40.6501038	-73.9495823
Queens	40.7498243	-73.7976337
Staten Island	40.5834557	-74.1496048

The Folium library is used to display a map of New York City with the neighborhoods overlayed on the map as shown in Figure 1. The coding for the legends was done by Colin Talbert @ http://nbviewer.jupyter.org/gist/talbertc-usgs/18f8901fc98f109f2b71156cf3ac81cd.

The neighboorhoods (indicated by color-coded markers: Red – Staten Island, Green - Brooklyn, Blue - Manhattan, Yellow – The Bronx, Orange - Queens) are superimposed on the New York City map by using the geographical coordinates obtained from each neighbourhood dataframes.

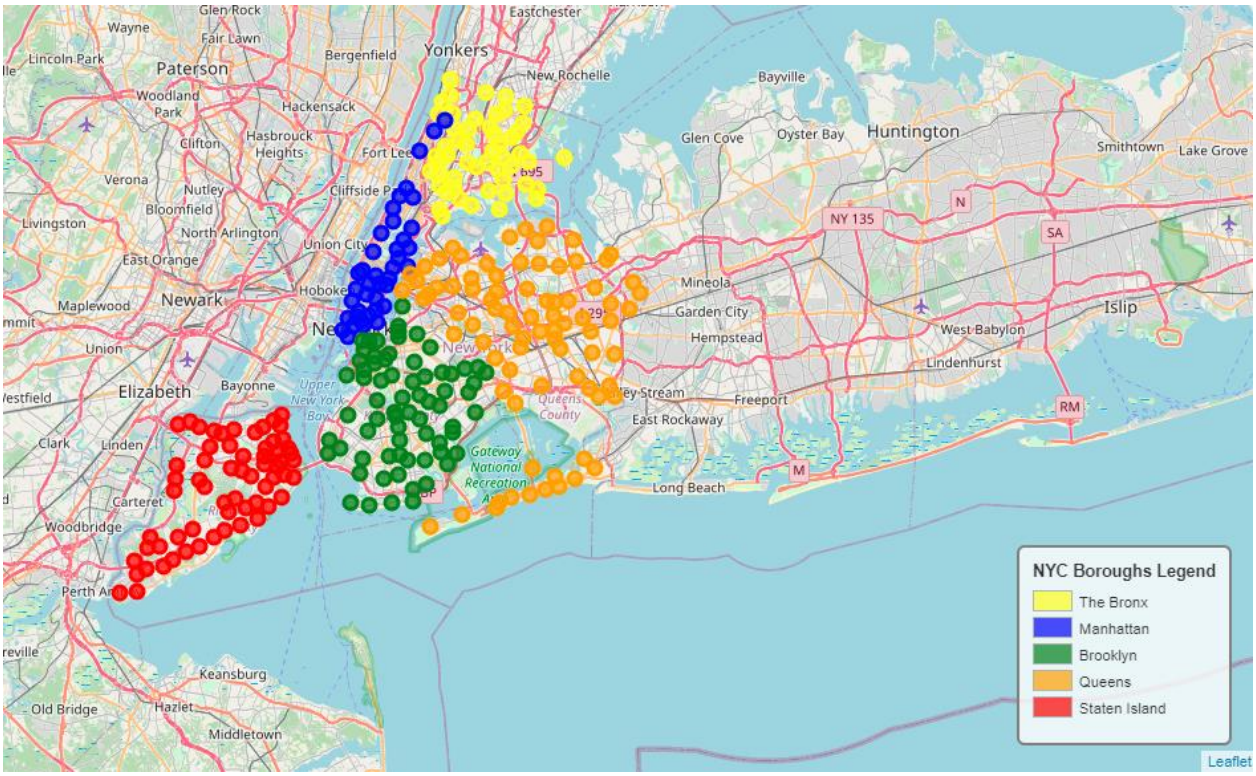

Figure 1: Neighboorhoods of New York City

4.4. Foursquare API datasets

The Foursquare API is used to explore the neighborhoods in each borough. The GET request URL is used to extract the top 100 venues within a radius of 500m of the geographical coordinates of each neighbourhood. These results are populated into a dataframe for each result set. The number of returned venues for each borough is indicated by the first coordinate below:

The Bronx dataframe shape: (1194, 7)
The Manhattan dataframe shape: (3132, 7)
The Brooklyn dataframe shape: (2711, 7)
The Queens dataframe shape: (2088, 7)
The Staten Island dataframe shape: (830, 7)

These datasets are filtered to extract the following target sports venue categories of interest: 'Cycle Studio', 'Athletics & Sports', 'Gymnastics Gym', 'Baseball Field', 'Soccer Field', 'Boxing Gym', 'Martial Arts Dojo', 'Tennis Court', 'Volleyball Court', 'Basketball Court'. The results are dataframes of the target sports venue categories for each of the boroughs. See Table 3 below for an example of the dataframe head (1st 5 rows) for Manhattan dataset.

Table 3: Dataframe head with the sports venues in Manhattan

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Central Harlem	40.815976	-73.943211	Harlem Cycle	40.817201	-73.942592	Cycle Studio
Upper East Side	40.775639	-73.960508	SoulCycle East 83rd	40.776991	-73.955329	Cycle Studio
Yorkville	40.775930	-73.947118	Asphalt Green Field	40.778580	-73.944448	Soccer Field
Yorkville	40.775930	-73.947118	Yorkville Tennis Club	40.778263	-73.946289	Athletics & Sports
Lenox Hill	40.768113	-73.958860	CYC Fitness	40.768988	-73.960797	Cycle Studio

An outer merged is performed on the 5 sports venue categories dataframes to get the number of different sports venues for each of the different boroughs in New York City as indicated in Table 4.

Table 4: Number of different sports venues categories in New York City

Venue Category	Bronx	Manhattan	Brooklyn	Queens	Staten Island	Total
Athletics & Sports	1	2	4	3	6	16
Baseball Field	7	1	5	6	6	25
Basketball Court	2	1	1	4	2	10
Martial Arts Dojo	4	4	4	6	3	21
Soccer Field	1	2	1	0	0	4
Tennis Court	1	5	4	1	0	11
Boxing Gym	0	7	4	1	0	12
Cycle Studio	0	17	4	1	0	22
Volleyball Court	0	1	0	0	0	1
Gymnastics Gym	0	0	2	1	0	3

The plot method of the Matplotlib library is used to display the data in a horizontal bar graph as indicated in Figure 2. The shaded area in the bottom left corner of Figure 2 shows the venue categories that are common across all boroughs in New York City.

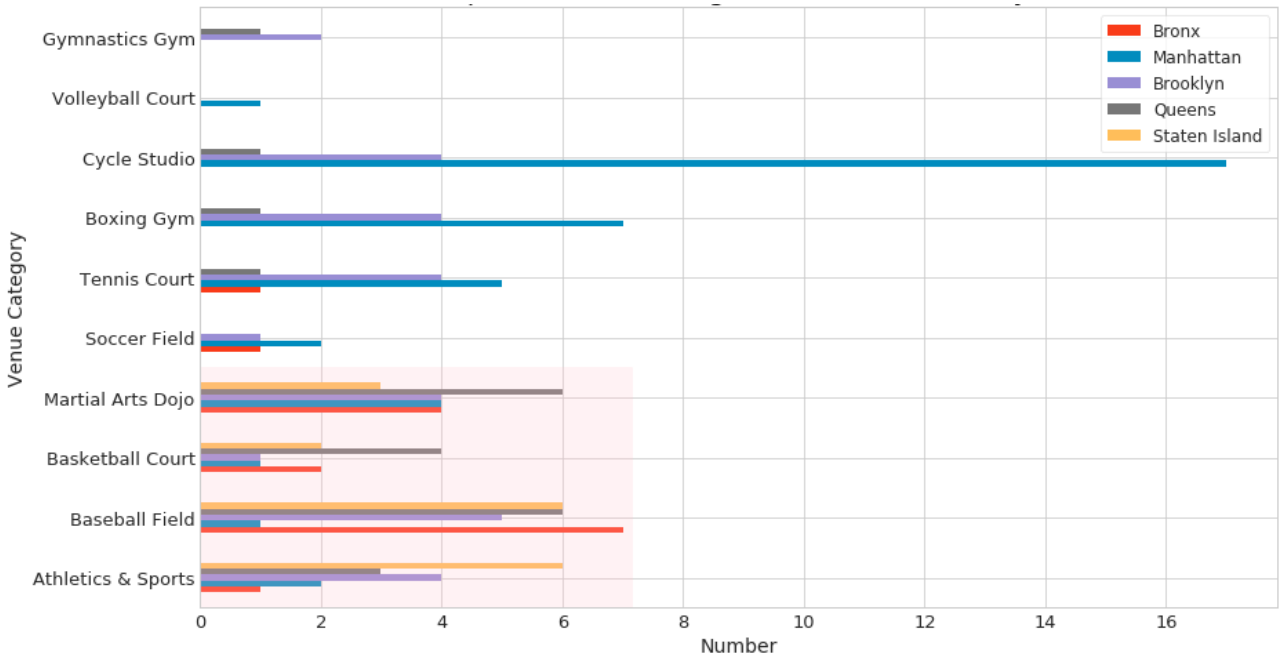

Figure 2: Number of sports venues in New York City

From Figure 3 it is clear that Baseball fields are the most common sports venues in New York city, followed by Cycle Studios and Martial Arts Dojos.

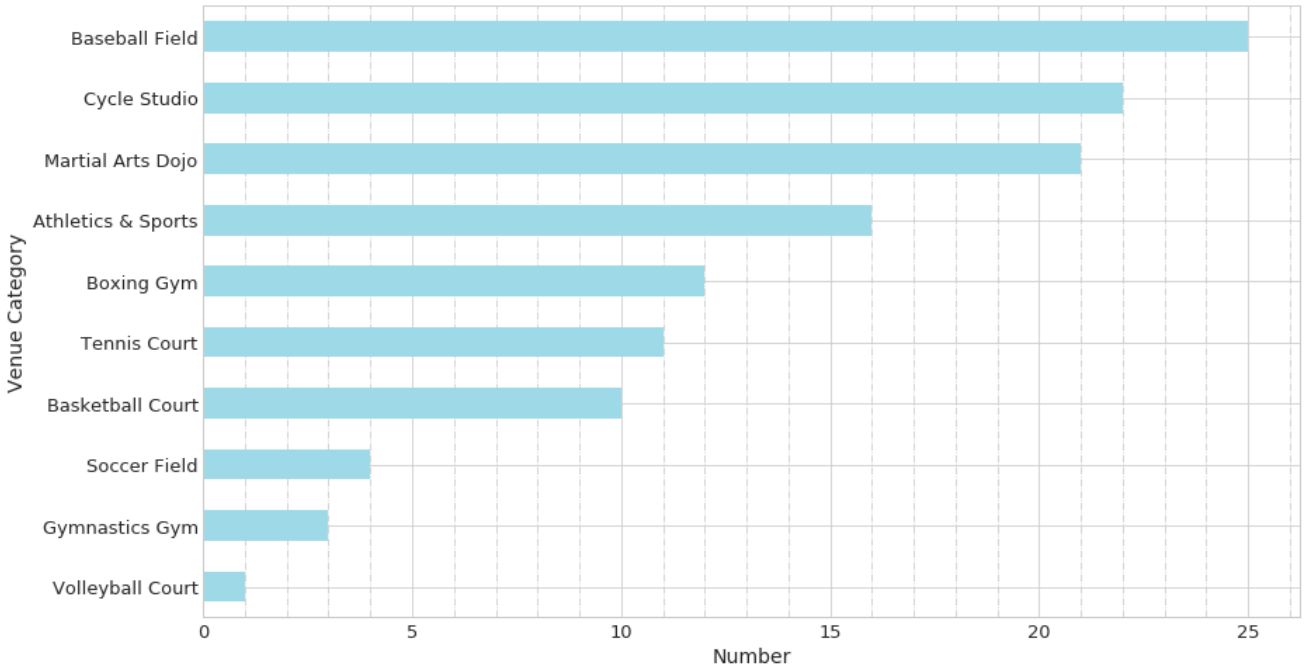

Figure 3: Number of different sports venues in New York City

4.6. Cluster Neighborhoods

The k-means clustering algorithm from scikit-learn's library is used to cluster the each borough and New York City as a whole based on the target sports venues.

Firstly we need to determine the optimum k-value or number of clusters. We do this by assigning a k-value between 1 to 12, initialise k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster centre. As k increases, the sum of squared distance tends to zero. Below is a plot of sum of squared distances for k in the range specified above. If the plot looks like an arm (see Figure 4), then the elbow on the arm is optimal k. From Figure 4, the elbow appears to be at k = 7.

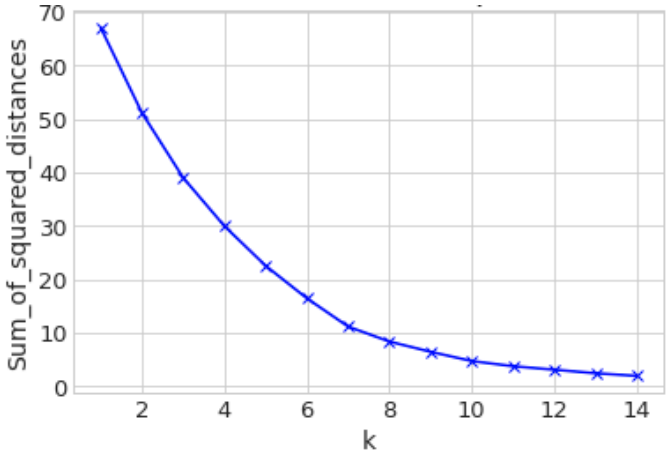

Figure 4: k-value vs. sum of squared distance

The top 5 sports venues for each cluster (by borough and neighbourhood) are indicated in Table 5, below.

Table 5: The 7 Sports Clusters for New York City

Cluster 1

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Brooklyn	Clinton Hill	Boxing Gym	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Brooklyn	North Side	Cycle Studio	Boxing Gym	Volleyball Court	Tennis Court	Soccer Field
Manhattan	Murray Hill	Boxing Gym	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Queens	Murray Hill	Boxing Gym	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Queens	Forest Hills	Martial Arts Dojo	Boxing Gym	Volleyball Court	Tennis Court	Soccer Field

Cluster 2

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bronx	Country Club	Athletics & Sports	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Bronx	Schuylerville	Soccer Field	Volleyball Court	Tennis Court	Martial Arts Dojo	Gymnastics Gym
Brooklyn	Greenpoint	Gymnastics Gym	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Brooklyn	Brooklyn Heights	Cycle Studio	Athletics & Sports	Volleyball Court	Tennis Court	Soccer Field
Brooklyn	Marine Park	Soccer Field	Basketball Court	Baseball Field	Athletics & Sports	Volleyball Court

Cluster 3

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bronx	Kingsbridge	Martial Arts Dojo	Volleyball Court	Tennis Court	Soccer Field	Gymnastics Gym
Bronx	Melrose	Martial Arts Dojo	Volleyball Court	Tennis Court	Soccer Field	Gymnastics Gym
Bronx	Belmont	Martial Arts Dojo	Volleyball Court	Tennis Court	Soccer Field	Gymnastics Gym
Brooklyn	Gravesend	Martial Arts Dojo	Baseball Field	Volleyball Court	Tennis Court	Soccer Field
Brooklyn	Gowanus	Martial Arts Dojo	Volleyball Court	Tennis Court	Soccer Field	Gymnastics Gym

Cluster 4

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Brooklyn	South Side	Cycle Studio	Boxing Gym	Volleyball Court	Tennis Court	Soccer Field
Manhattan	Central Harlem	Cycle Studio	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Manhattan	Upper East Side	Cycle Studio	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Manhattan	Lenox Hill	Cycle Studio	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Manhattan	Lincoln Square	Cycle Studio	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo

Cluster 5

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bronx	West Farms	Basketball Court	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Bronx	Olinville	Basketball Court	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Queens	Corona	Basketball Court	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Queens	Floral Park	Basketball Court	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Staten Island	Grasmere	Basketball Court	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo

Cluster 6

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bronx	Spuyten Duyvil	Tennis Court	Volleyball Court	Soccer Field	Martial Arts Dojo	Gymnastics Gym
Brooklyn	Fort Greene	Tennis Court	Volleyball Court	Soccer Field	Martial Arts Dojo	Gymnastics Gym
Brooklyn	Fort Hamilton	Tennis Court	Volleyball Court	Soccer Field	Martial Arts Dojo	Gymnastics Gym
Manhattan	Lower East Side	Tennis Court	Volleyball Court	Soccer Field	Martial Arts Dojo	Gymnastics Gym
Manhattan	Morningside Heights	Tennis Court	Volleyball Court	Soccer Field	Martial Arts Dojo	Gymnastics Gym

Cluster 7

Borough	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Bronx	Co-op City	Baseball Field	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Bronx	Riverdale	Baseball Field	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Bronx	City Island	Baseball Field	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Bronx	Bedford Park	Baseball Field	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo
Bronx	Mott Haven	Baseball Field	Volleyball Court	Tennis Court	Soccer Field	Martial Arts Dojo

The Folium library is used to display a map of New York City with the different sports venue clusters superimposed (indicated by color-coded markers: Red – Cluster 2, Green – Cluster 3, Blue – Cluster 4, Yellow – Cluster 5, Orange – Cluster 1) by using the geographical coordinates obtained from each neighbourhood dataframes.

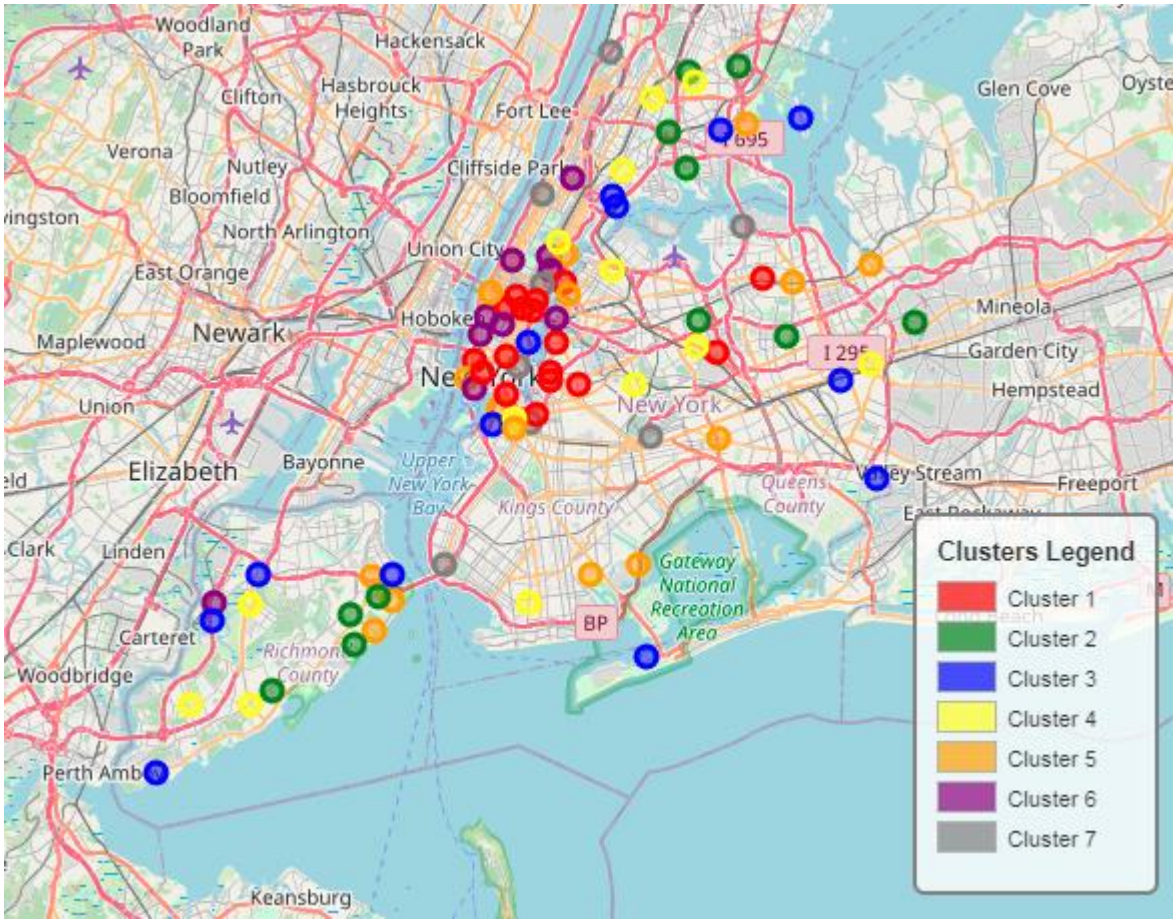

Figure 5: New York City sports venue clusters

5. Discussion

The results obtained indicates that New York City can be divided into 7 sports venue clusters, where each clusters is characterised by a collection of common sports engaged in in the cluster. Baseball Fields (25) are the most common sports venues in New York city, followed by Cycle Studios (22), Martial Arts Dojos (21), Athletics & Sports (16), Boxing Gyms (12), Tennis Courts (11), basket Ball Courts (10), Soccer Fields (4), Gymnastics Gyms (3) and Volley Ball Courts (1).

Cluster 1 and Cluster 2 are mixed between the different sports. The most common sports venues in Clusters 3 and 4 are Martial Arts Dojos and Cycle Studios, respectively. Cluster 5 is predominantly known for Basketball Courts, whereas in Cluster 6 consists of mostly Tennis Courts. The no.1 sports venue in New York City, Baseball Fields, is found in Cluster 7.

By looking at the commonality of sports venues between clusters, the marketing teams can be optimally deployed. For instance, the marketing team specialising in gymnastics should mainly focus on Clusters 6, 3 and 2.

6. Conclusions

The results from this study can be used by a marketing team to decide which sports categories to focus on and in which neighboorhoods of New York City the target markets are. It can also be used to determine in which neighboorhoods opportunities exist for introductions of new sports activities and creating new markets.