Introduction and Business Problem

The "Sport Society" intends to open a new Sport Museum in New York City and wants to identify the most suitable neighborhood. In fact, it would be ideal to locate near stadiums, sport fields, parks and/or sports bars and sporting good shops.

This project aims to indicate to the Sport Society the most suitable neighborhood in which it should build their museum.

Data

Data about New York City (NY) neighborhood, borough and their coordinates come from "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json".

All data about the number and type of facilities set in the different neighborhood comes from Foursquare database. Particularly, the following venues will be analysed: "Athletics & Sports", "Sporting Goods Shop", "Sports Club", "Sports Bar", "Soccer Field", "Tennis Stadium", "Basketball Court", "Volleyball Court", "Baseball Stadium", "Baseball Field", "Park", "Tennis Court".

Neighborhoods will be clustered in according to the presence of the aforementioned facilities and descriptive statistics will be given.

Methodology

The following libraries were imported in Notebook Juptyer (Python 3): numpy, pandas, json, geopy, matplotlib, sklearn.cluster, folium, urllib.request.

New York city dataset (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json) was imported and converted to a panda dataframe. Data about borough, Neighborhood, Latitude and Longitude were retained.

NY coordinates were obtained using Nominatum and then a map of NY showing markers on all the different neighborhoods was made with Folium.

Using Foursquare, a list of venues in each neighborhood, classified on the basis of their category, was obtained. Only the venues in the aforementioned categories (see “Data” section) were retained.

Using sklearn.cluster, a k-means Clustering analysis (unsupervised machine learning) was performed. A number of k-means = 5 was used.

A new map was elaborated, in which the neighborhood were colored according to the cluster they were assigned to.

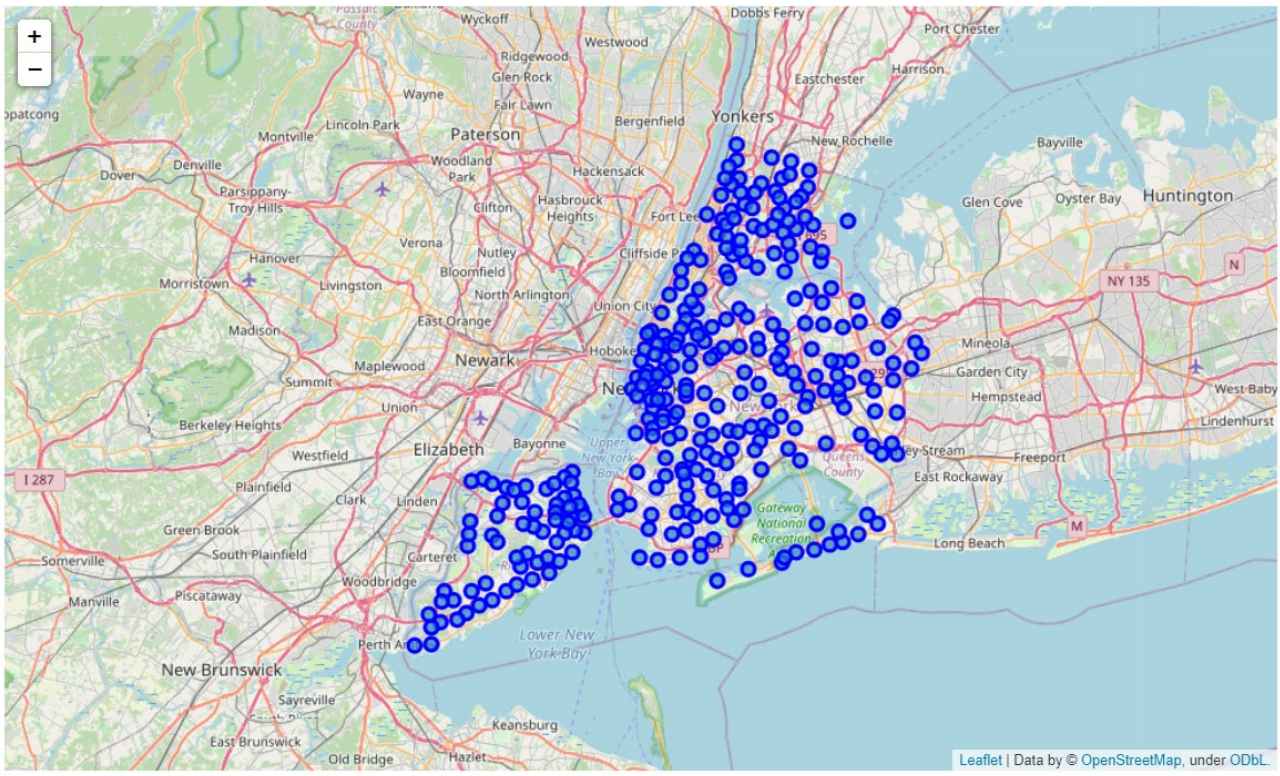
Clusters were then described: the mean number of venues was calculated and plotted as bar plot. The best clusters, according to the average number of venues and the variety of venues, were further described at a neighborhood level, in order to identify the best ones.

A last check was done, analyzing the total number of venues per neighborhood, without considering their clusterization.

Results

The NY dataset included 306 neighborhoods. Nominatum found that the coordinates of NY are latitude= 40.7127281 and longitude=-74.0060152.

Figure 1 show the map of NY with markers on all its neighborhoods.



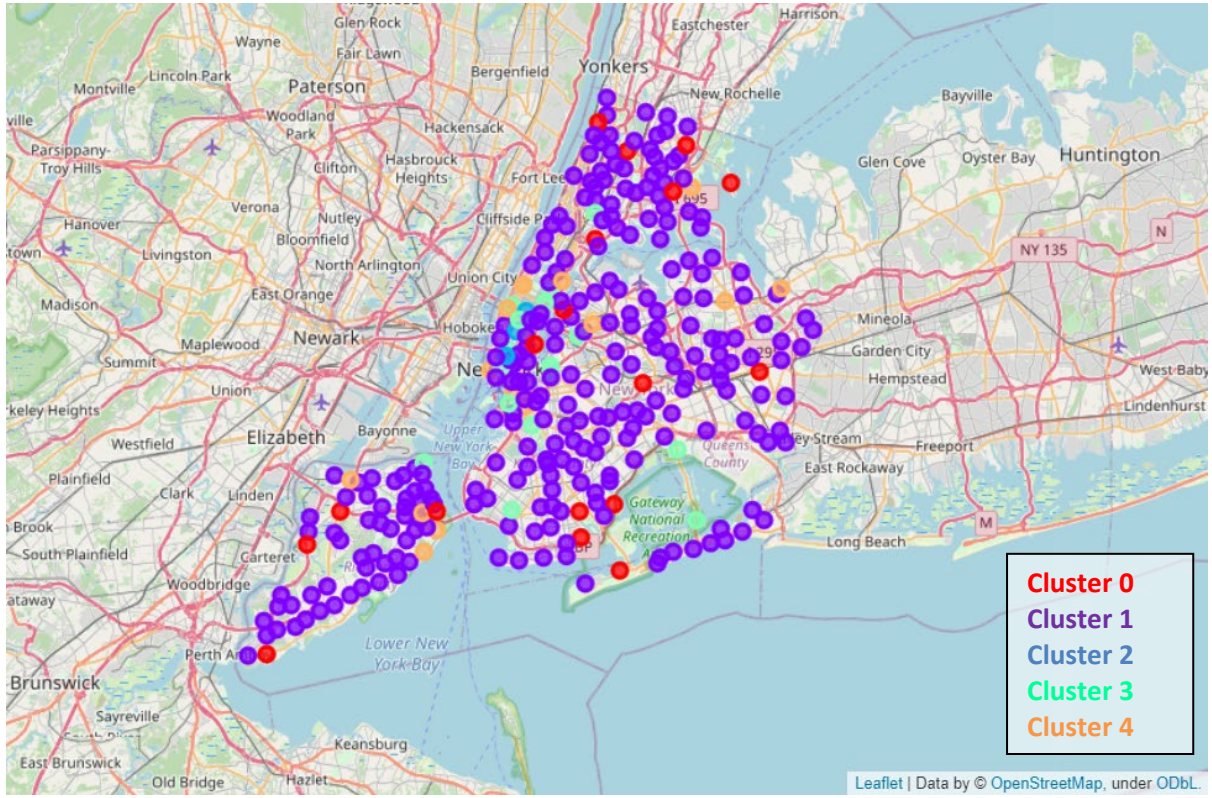
Sport-related venues in all the neighborhoods were identified using Foursquare. The total number of venues in NY are reported in Table 1.

Venue category	Total number of venues in NY
Athletics & Sports	17
Sporting Goods Shop	26
Sports Club	4
Sports Bar	12
Soccer Field	3
Tennis Stadium	2
Basketball Court	11
Volleyball Court	1
Baseball Stadium	3
Baseball Field	20
Tennis Court	12
Total	**111**

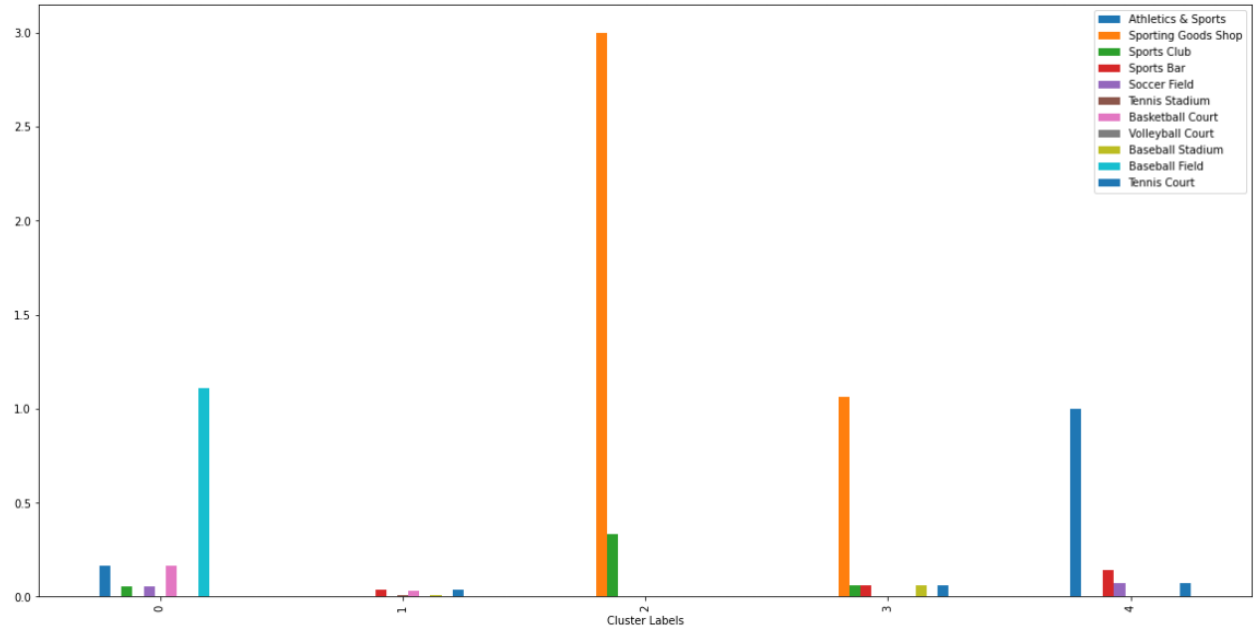
A k-mean clustering with k=5 was performed. Table 2 shows the statistics (number of neighborhoods, means and standard deviation for all venues category) for all the different clusters.

Cluster	N. of neighborhoods	Athletics & Sports		Sporting Goods Shop		Sports Club		Sports Bar		Soccer Field		Tennis Stadium		Basketball Court		Volleyball Court		Baseball Stadium		Baseball Field		Tennis Court	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
0	18	0,17	0,38	0	0	0,1	0,2	0	0	0,1	0,2	0	0	0,2	0,4	0	0	0	0	1,1	0,3	0	0
1	253	0	0	0	0	0,004	0,1	0,04	0,2	0,004	0,1	0,01	0,1	0,03	0,2	0,004	0,1	0,01	0,1	0	0	0,04	0,2
2	3	0	0	3	0	0,3	0,6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	16	0	0	1,1	0,3	0,1	0,3	0,1	0,3	0	0	0	0	0	0	0	0	0,	0,3	0	0	0,1	0,3
4	14	1	0	0	0	0	0	0,1	0,5	0,1	0,3	0	0	0	0	0	0	0	0	0	0	0,1	0,3

A map with markers on clustered neighborhoods is reported as Figure 2.



The mean number of venues per cluster was plotted and is shown in Figure 3.

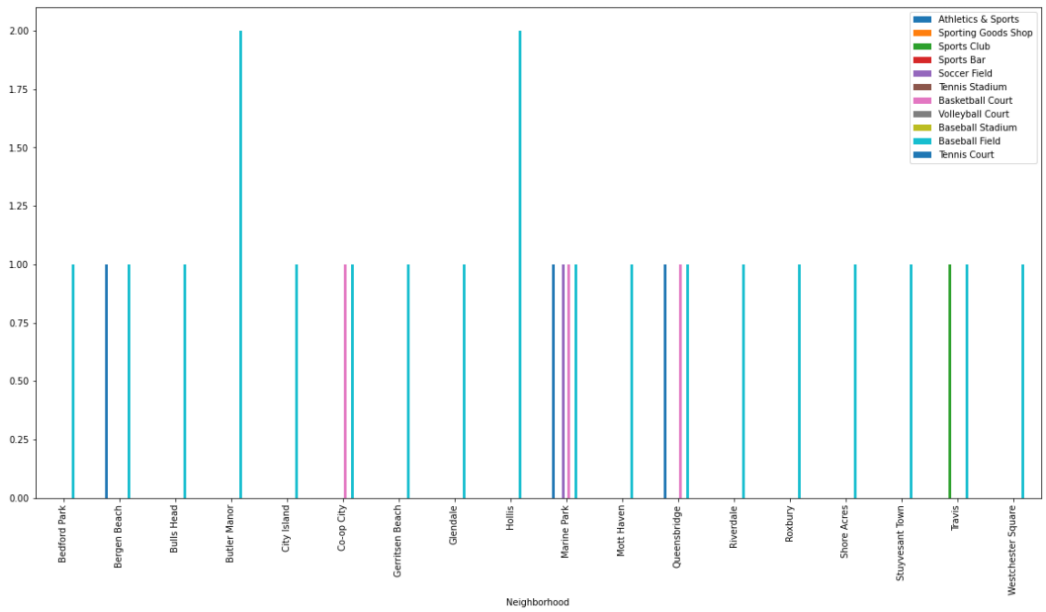


Since cluster 0 and cluster 2 had the highest mean of venues and the most variable category of venues, they were further investigated.

The top 5 **Cluster 0** neighborhoods in terms of number of venues are reported in Table 3.

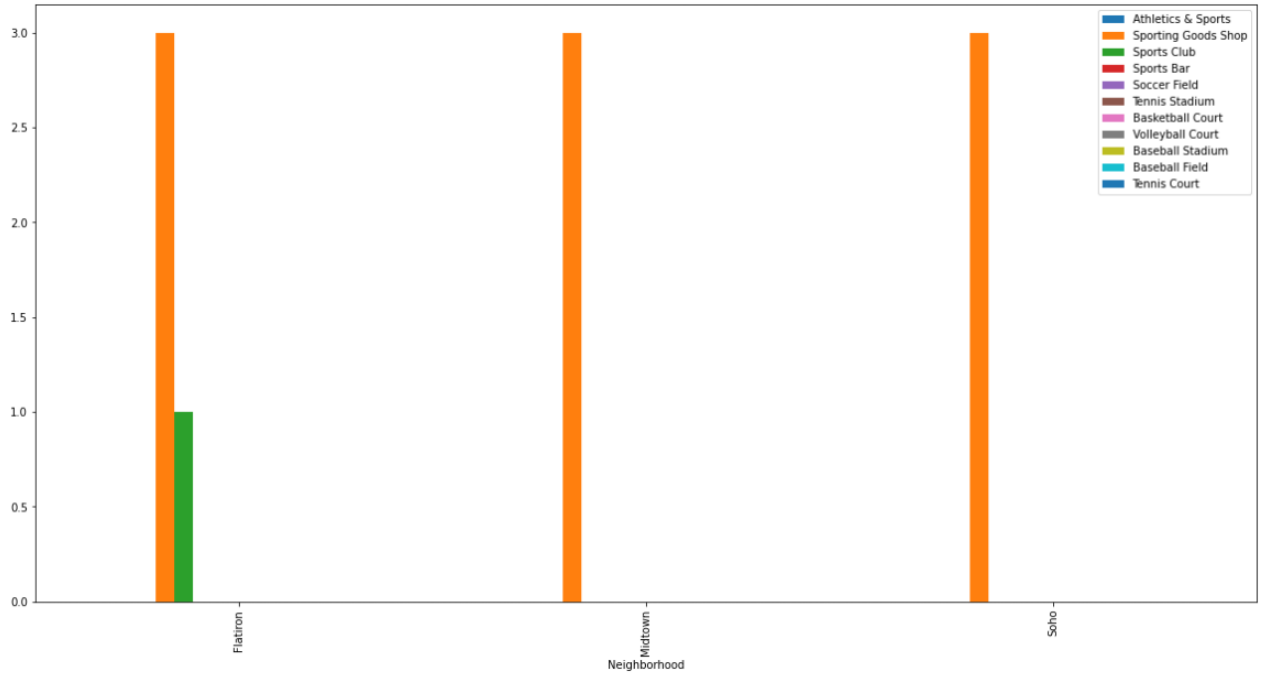
Neighborhood	Athletics & Sports	Sporting Goods Shop	Sports Club	Sports Bar	Soccer Field	Tennis Stadium	Basketball Court	Volleyball Court	Baseball Stadium	Baseball Field	Tennis Court	Tot al
Marine Park	1	0	0	0	1	0	1	0	0	1	0	4
Queensbridge	1	0	0	0	0	0	1	0	0	1	0	3
Travis	0	0	1	0	0	0	0	0	0	1	0	2
Butler Manor	0	0	0	0	0	0	0	0	0	2	0	2
Co-op City	0	0	0	0	0	0	1	0	0	1	0	2

The same results were plotted in Figure 4.



Neighborhoods in Cluster 2 were analyzed in Table 4 and plotted in Figure 5.

Neighborhood	Athletics & Sports	Sporting Goods Shop	Sports Club	Sports Bar	Soccer Field	Tennis Stadium	Basketball Court	Volleyball Court	Baseball Stadium	Baseball Field	Tennis Court	Total
Flatiron	0	3	1	0	0	0	0	0	0	0	0	4
Midtown	0	3	0	0	0	0	0	0	0	0	0	3
Soho	0	3	0	0	0	0	0	0	0	0	0	3



A last check on all the neighborhoods was performed. Flatiron and Marine Park demonstrated to be the neighborhoods with the highest number of venues (n=4). The top 5 neighborhoods in terms of total number of venues belonged to cluster 0 (n=2) and 2 (n=3).

Discussion

A Sport Museum should be built near to sport-related venues, since it is more probable that people interested in sports visit those neighborhoods. The k-means clustering was used to identify groups of suitable neighborhoods, based on the aforementioned data.

Cluster 2 had the higher average number of venues, but it should be considered that most of them fell in the categories “Sport goods shop” (n=9) and, to a lesser extent, “Sport Club” (n=1, in Flatiron Neighborhoods). Even if people interested in sports for sure visit this kind of places, more variety in venues is desirable.

Therefore, cluster 0 neighborhoods may be the best choice for building the Sport Museums. Indeed, they include a lot of different venues, such as “Sports Club”, “Athletics & Sports”, “Baseball field”, “Basketball Court”, and “Soccer Field”. This means that people interested in different kind of sports can be found around these places. Particularly, in Marine Park we can found a basketball court, a baseball field and a soccer field, as well as an “Athletics & Sports” venue.

Conclusion

On the basis of these results, I recommend to build the new museum in a neighbor of cluster 0, and in particular in Marine Park. Moreover, since in cluster 0 neighborhoods there are not sporting goods shops, I would suggest to build a shop inside the museum.