Capstone Project - The Battle of Neighborhoods

Applied Data Science Capstone by Alfonso Lopez

Table of contents

Introduction: Business Problem

The **West Stationery Company** (WSC) is expanding their business to the east of USA. During last three years, but it has not been as successful as expected. So, WSC have decided to open at the same time **3 new WSC Stationery Stores** at New York, with the main goal of revert this situation by start having enough presence in one of the most relevant cities of the World.

To support this decision and to select the best neighborhoods for the new stores, WSC has hired us as Data Scientist experts.

WSC has defined this project as ***"The Knowledge Triangle"*** (TKT as secret key inside the company) and many of their resources will focus on it.

This solution will not be easy. The company has given us the following notes (that won't help in the problem resolution):

- WSC has no experience nor information about its market at NY City. Inside the company, there aren't relevant information that could be used. So, the project can be considered as an empty bottle to be filled.
- The renting price of commercial premises at New York is probable the highest of USA. So, the investment must be done with special care.
- There is a short time to finish the work as the opening is expected by **September, the 1st 2019**.
- As there will be only three stores at the beginning, WSC wants not only the best places, but also, they will be more or less geographically distributed at NY City.

With these bare facts, our Data Scientist Team must work in achieving:

The best 3 neighborhoods for opening the new WSC Stationery Stores

Data

Based on definition of our problem, these are the factor that mainly will influence our decision (which neighbors are better for the new stores):

Negative influcence:

- Number of stationery shops and bookstore in the neighborhood.

Possitive incluence:

- Number of schools or universities.
- Number of other education institutions.
- Population.

Additional considerations:

- Distance among selected neighborhoods.

Data Sources

We will use the following data sources for the project:

1. **Geographic Information:** From library ***geopy.geocoders*** we'll obtaing any geographical coordinate from an address. We'll be able to locate New York city (to represent it with the related map). Also, address from venues (FourSquare) we'll help us in selecting the best position for the new WSC stores.

2. **Neighborhoods:** From https://geo.nyu.edu/download/file/nyu-2451-34572-geojson.json, we'll take all neighborhood coordinates. At the beginning, we won't exclude any value. This coordinates will partition the city in enough small regions to allow a valid study. We cannot select the exact renting premise, but we can focus WSC infrastructure deparment in the right neighborhood.

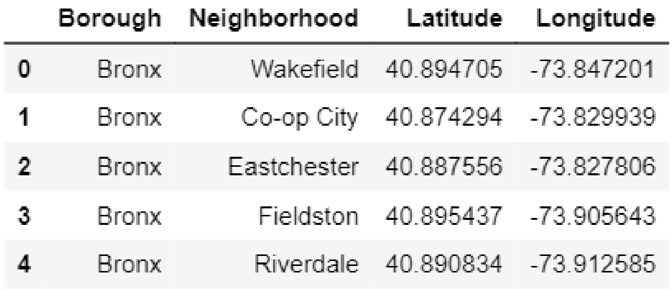
3. **Venues:** From https://www.foursquare.com API, we'll select all related educational centers for each neighborhood. Also, we will detect existing stationery stores and bookstores that will condition the sales for the new WSC stores.

Reading the Neighborhood data and representing it in a NY City map:

From https://geo.nyu.edu we will obtain a detailed dataset of the NY neighborhood, including name and geographical coordinates. We will read a JSON file and will add the information into a new DataFrame.

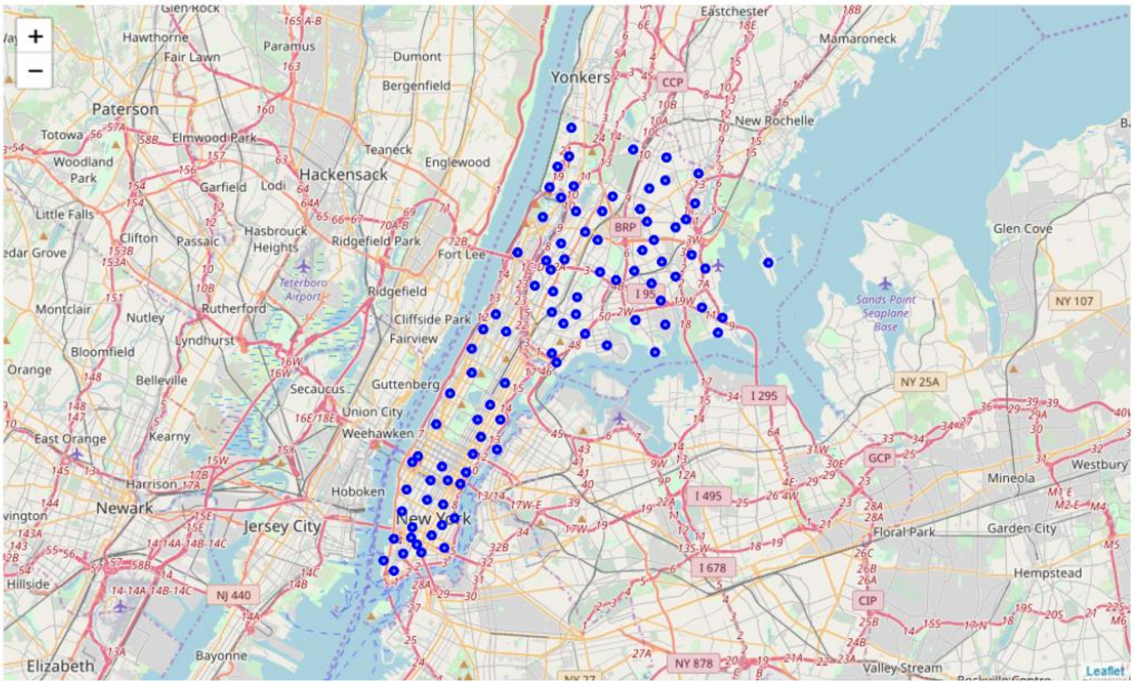
After several meetings, WSC has asked us to focus only in Manhattan and The Bronx boroughs, because they are the most representatives at NY City.

This is an example of the resulting table from geo.nyu.edu:



	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

Map representation of New Yor City (with folium):¶
We can represent all these neighborhoods in the following map:



Foursquare

After having represented all the NY City neighborhoods, let’s use Foursquare API to get info on stationeries, bookstores, different educational centers and so on. All of them related with each of these neighborhoods.

By analyzing the available venues at Foursquare, we have selected the following positive ones (with the related Foursquare code):

- School, HighSchool: 4bf58dd8d48988d13b941735
- University: 4d4b7105d754a06372d81259

In a negative way, these are the selected venues:

- Bookstore: 4bf58dd8d48988d114951735
- Stationery 52f2ab2ebcbc57f1066b8b21

Not all the University related buildings have been selected, nor educational centers. But all stationeries and bookstores will be searched in the Foursquare databases.

Let's explore the neighborhoods in our dataframe.

We will ask Foursquare for venues included in the previous lists. We will use several functions to search these venues given a category and a location (latitude, longitude)

```
Obtaining venues around candidate locations:4bf58dd8d48988d13b941735
 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . done.
4d4b7105d754a06372d81259
 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . done.
Obtaining venues around candidate locations:4bf58dd8d48988d114951735
 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . done.
52f2ab2ebcbc57f1066b8b21
 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
. . . . done.
```

Methodology

In this project we will focus in selecting at least **3 New York City neighborhoods where WSC company should expanse** its stationery store network. There are 5 main boroughs at NY, but we will reduce the study to only two of them: The Bronx and Manhattan (Queens, Brooklyn and Staten Island have been discarded). The methodology is perfectly extendable to new neighborhoods, just by including these new boroughs to our datasets. So, the study is reduced to 92 neighborhoods instead a maximum of 306 possible at NY City.

Every student or office employee is a potential user a stationery, but office employees usually don't manage the buying of stationery material. However, students are especially active buyers in the WSC network. In fact, WSC shops are one of the best rated in sector. So, the analysis will be focused in detecting how cover the **greatest number of schools and universities** with 3 stores. Also, it will be studied the number of **existing stationeries and bookstores** (that directly compete with WSC) to avoid selecting neighborhoods with a huge number of competitor stores.

In a first step we have collected the required venues (schools, universities, stationeries and bookstores around each neighborhood by using the Foursquare API. This information will contain the geographical coordinates.

In second step we will focus on create 3 clusters (using **k-means clustering**) including only the positive (schools and universities). With this distribution it will be possible to detect a centroid point from which WSC could cover a greater number of potential customer (students).

A third and final step in our analysis, we will calculate the density of negative (stationery and bookstores) venues for every neighborhood. Inside WSC, bookstores are considered half competitive than stationery. So, for this density approach, we will use 1.0 and 0.5 as a multiplier for stationeries and bookstores respectively.

Analysis

First, let's create a new DataFrame with the positive venues and count the number of schools and universities for every neighborhood:

```
Total number of Schools and Universities: 3141
```

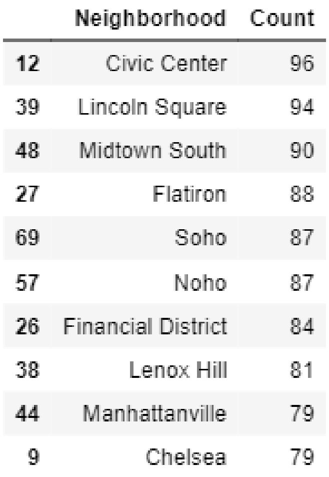
Then, let's do the same with the negative venues (stationeries and bookstores):

```
Total number of Stationeries and Bookstores: 232
```

Clustering schools and universities

Distribution of Schools and Universities by Neighborhood

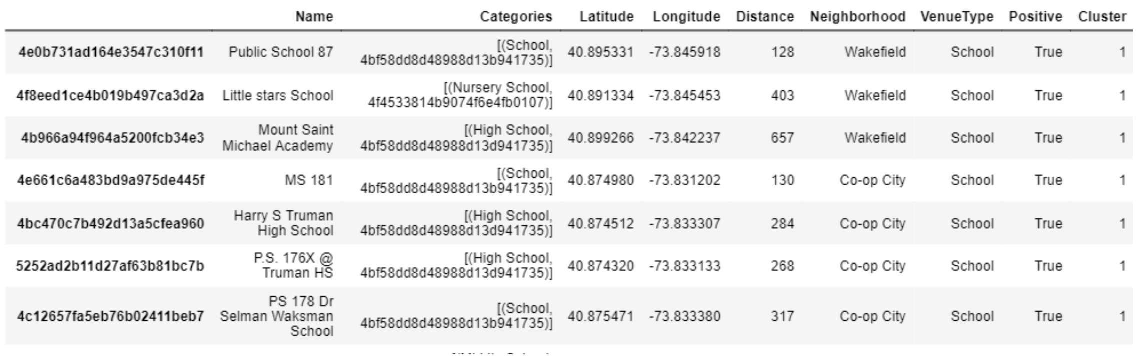
Let's group the results by neighborhood and list the top 10:



	Neighborhood	Count
12	Civic Center	96
39	Lincoln Square	94
48	Midtown South	90
27	Flatiron	88
69	Soho	87
57	Noho	87
26	Financial District	84
38	Lenox Hill	81
44	Manhattanville	79
9	Chelsea	79

Running k-means with k = 3¶

With the following DataFrame, we will obtain 3 groups covering the maximum number of positive venues. Then, we add the cluster calculated to each venue:



	Name	Categories	Latitude	Longitude	Distance	Neighborhood	VenueType	Positive	Cluster
4e0b731ad164e3547c310f11	Public School 87	[(School, 4bf58dd8d48988d13b941735)]	40.895331	-73.845918	128	Wakefield	School	True	1
4f8eed1ce4b019b497ca3d2a	Little stars School	[(Nursery School, 4f4533814b9074f6e4fb0107)]	40.891334	-73.845453	403	Wakefield	School	True	1
4b966a94f964a5200fcb34e3	Mount Saint Michael Academy	[(High School, 4bf58dd8d48988d13d941735)]	40.899266	-73.842237	657	Wakefield	School	True	1
4e661c6a483bd9a975de445f	MS 181	[(School, 4bf58dd8d48988d13b941735)]	40.874980	-73.831202	130	Co-op City	School	True	1
4bc470c7b492d13a5cfea960	Harry S Truman High School	[(High School, 4bf58dd8d48988d13d941735)]	40.874512	-73.833307	284	Co-op City	School	True	1
5252ad2b11d27af63b81bc7b	P.S. 176X @ Truman HS	[(High School, 4bf58dd8d48988d13d941735)]	40.874320	-73.833133	268	Co-op City	School	True	1
4c12657fa5eb76b02411beb7	PS 178 Dr Selman Waksman School	[(School, 4bf58dd8d48988d13b941735)]	40.875471	-73.833380	317	Co-op City	School	True	1

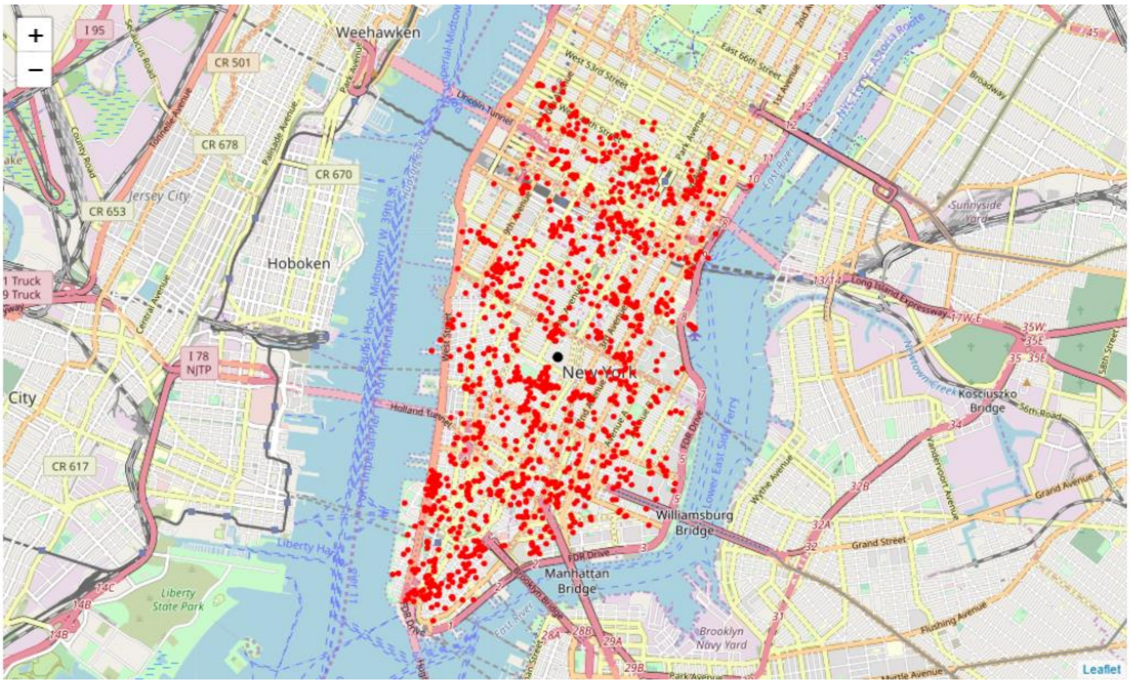
We will add the centroid of each cluster.

kmeans.cluster_centers_

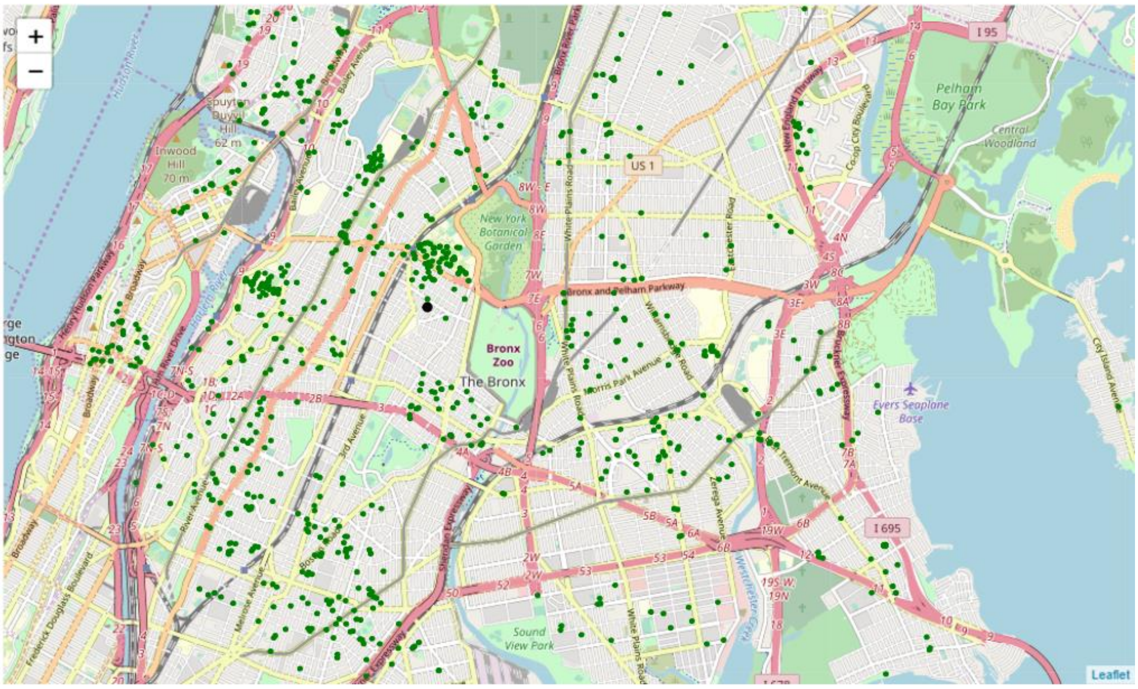
```
array([[ 40.7326154 , -73.9934078 ],
       [ 40.85509272, -73.88831772],
       [ 40.78897246, -73.95933385]])
```

And finally, we can represent each cluster with its centroid point:

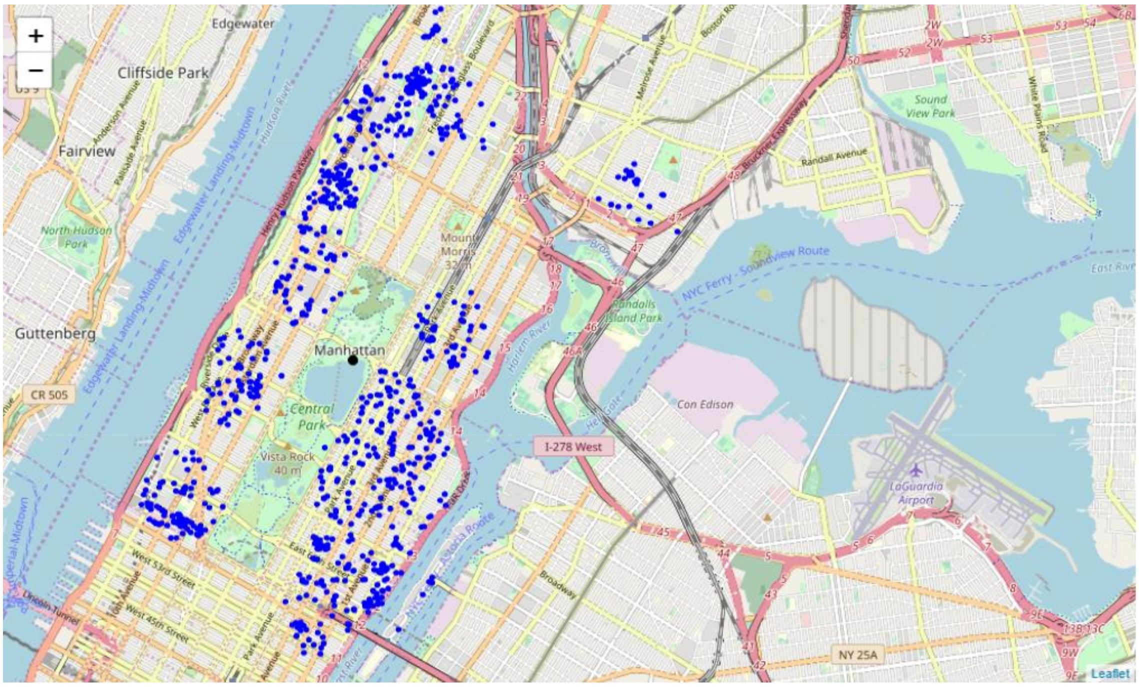
CLUSTER 0



CLUSTER 1



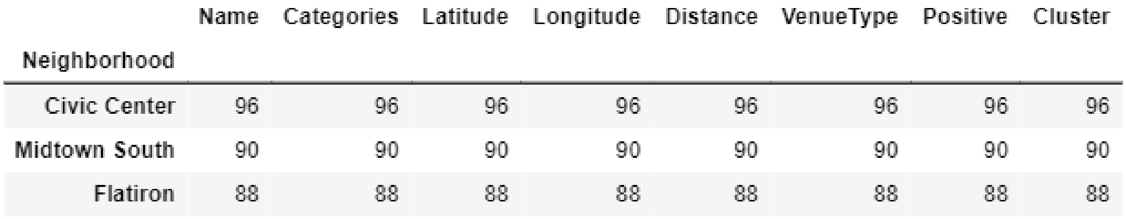
CLUSTER 2



Selecting the clusters

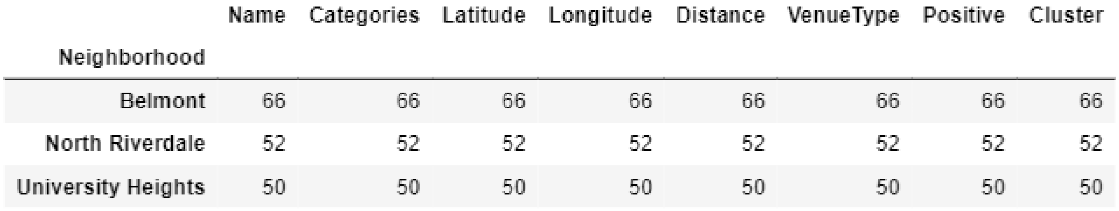
The number of negative venues is not enough for helping in selecting the best neighborhood. So, the best option is select among the neighborhoods with most number of schools and universities for each cluster. we have each cluster divided in the df_cluster_x dataframes.

CLUSTER 0



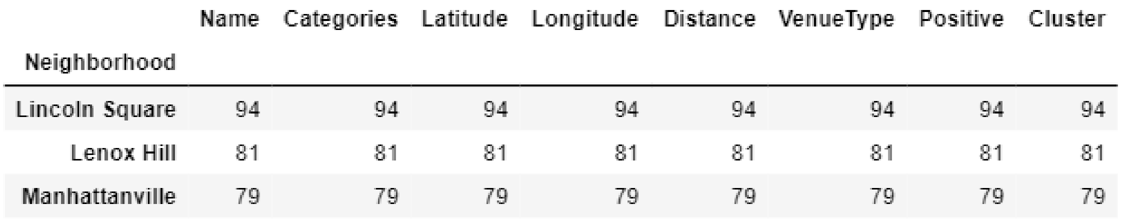
	Name	Categories	Latitude	Longitude	Distance	VenueType	Positive	Cluster
Neighborhood								
Civic Center	96	96	96	96	96	96	96	96
Midtown South	90	90	90	90	90	90	90	90
Flatiron	88	88	88	88	88	88	88	88

CLUSTER 1



	Name	Categories	Latitude	Longitude	Distance	VenueType	Positive	Cluster
Neighborhood								
Belmont	66	66	66	66	66	66	66	66
North Riverdale	52	52	52	52	52	52	52	52
University Heights	50	50	50	50	50	50	50	50

<u>CLUSTER 2</u>



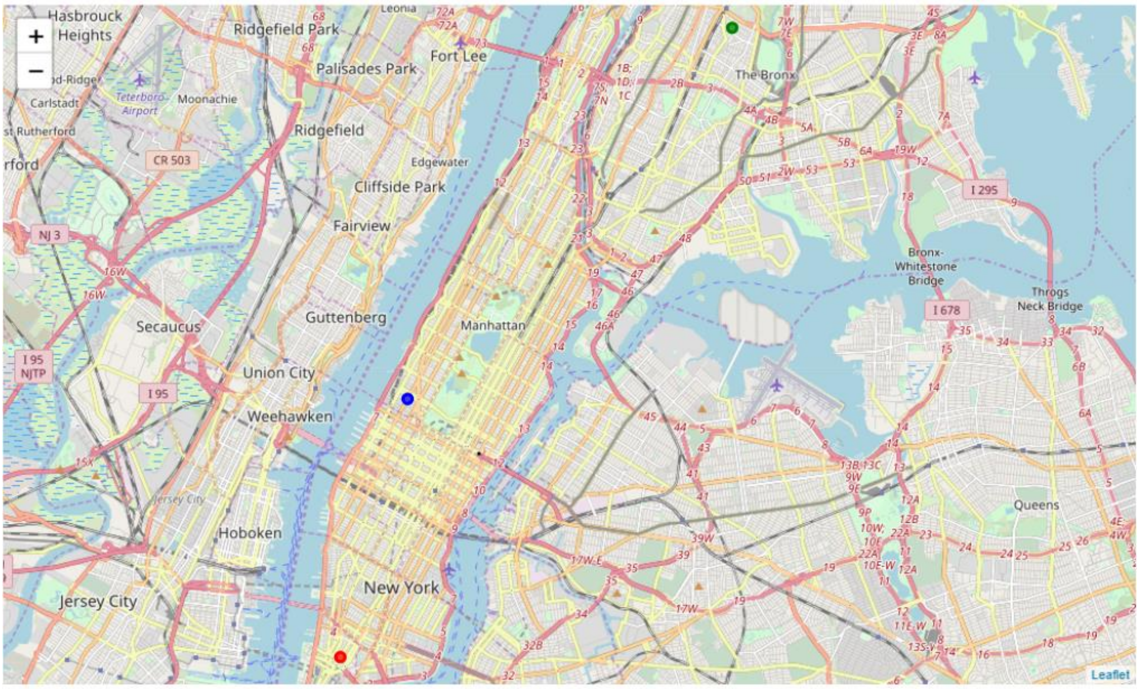
Neighborhood	Name	Categories	Latitude	Longitude	Distance	VenueType	Positive	Cluster
Lincoln Square	94	94	94	94	94	94	94	94
Lenox Hill	81	81	81	81	81	81	81	81
Manhattanville	79	79	79	79	79	79	79	79

Neighborhood selection.

Based in the last results, our recomendation to WSC is to search a renting location at the following neighborhoods:

- Civic Center
- Belmont
- Lincoln Square

The following map represents where are located these neighborhoods at NY City:



Results and discussion

Our analysis shows that there is a great number of educational centers in New York City (~2000 in Manhattan and Bronx boroughs), They are completely distributed through the whole city. So, it is not so important the density or distribution of these venues.

After checking the number of stationeries and bookstores, we have seen that they are not relevant in the study (we have only detected 232 in both boroughs).

So, we have distributed all the educational centers in three main clusters that will help us in the neighborhood selection. With the new distribution, we have selected the three neighborhoods with most educational centers as the best choice. There are also alternatives to these choices, and they will be completely valid as the difference is not so high.

Conclusion

Purpose of this project was to identify New York City neighborhoods from Manhattan and The Bronx boroughs where should be a better choice for deploying the first three new WSC stationery stores. By searching educational centers inside these boroughs and redistributing then into a geographical map, we could discover the high density of these kind of centers inside the city. Clustering those locations was then performed in order to create three major zones of interest for final exploration by stakeholders.

Final decision should be based in these results, but adding it the renting options and price for each one.