Capstone Project - The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Amoel Geogy

25-7-2021

Introduction

In this project we will try to find the optimal location to open an arcade in Manhattan. Our client has been to almost much all the arcades in Manhattan, but he does not find any of them satisfactory. Being an 80s kid, he lived through the golden age of arcades, and he wants to open one that will satisfy his nostalgia, and show the kids how it really used to be.

The target audience will mostly be early teenagers to young adults. We would need to find the neighborhoods in Manhattan where the density of this crowd is high. It would make sense to search for areas which have a **high number of universities or schools and a low number of arcades**.

Data

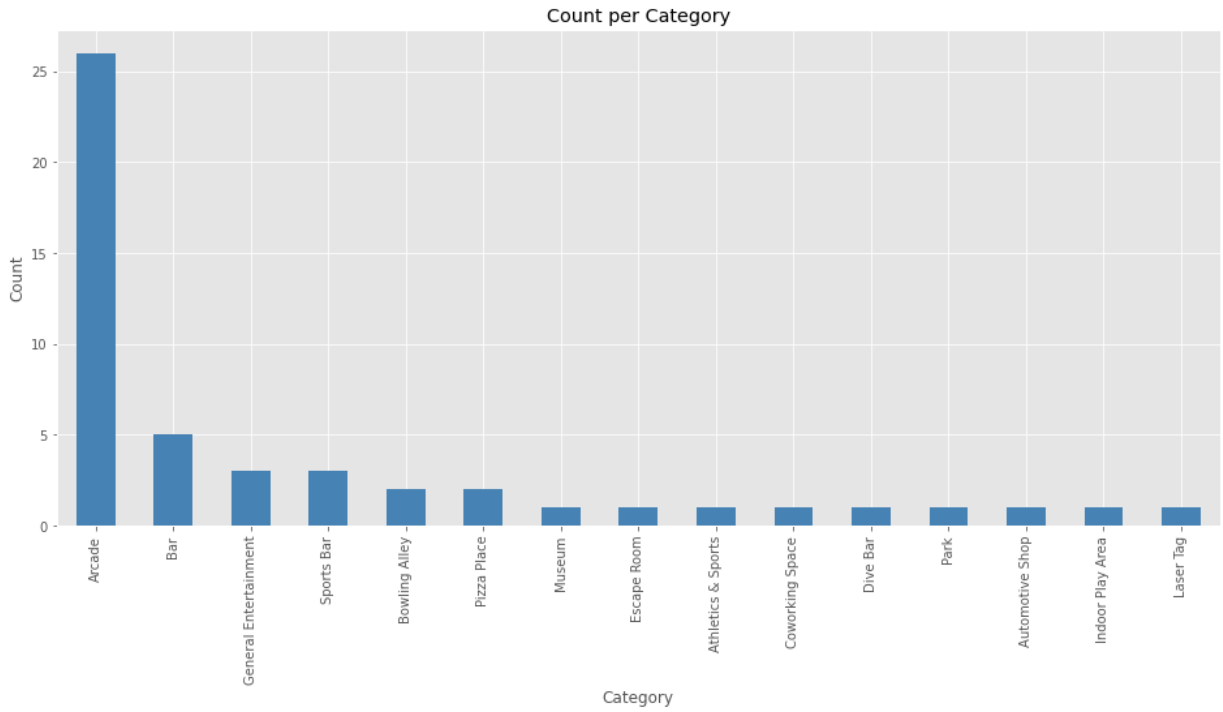
Based on this business problem, we can identify that we need data such as:

- locations of existing arcades
- locations of schools, universities and colleges
- geographical data like the coordinates and areas of each neighborhood in Manhattan

To get the locations of arcades and the academic institutions, we will use the Foursquare API. We will use the categoryids provided in the Foursquare website to search for the data of Arcades and College & Universities. We will retrieve the Name, Latitude, Longitude and Category of these locations.

To get the neighborhood boundaries, we will use the geojson file from this github repository. Furthermore, we will use the geopy.geocoders module for the coordinates of the neighborhoods.

We plotted the Categories of the results in a bar graph. This allowed us to see that we have unwanted categories in the dataset (like Bar, Museum, etc), and we will filter them out later.



Feature Selection:
For both the Arcade and University data, we have selected the Name, Latitude, Longitude and Category as the features to keep. We have dropped the other features.

Arcades:

	Name	Latitude	Longitude	Category
0	The Shelter Uptown	40.825184	-73.942968	Arcade
1	Barcade	40.712061	-73.951239	Bar
2	7B Horseshoe Bar aka Vazacs	40.725014	-73.981330	Dive Bar
3	Barcade	40.744417	-73.994423	Bar
4	The Zarcade	40.694607	-73.907475	Arcade

Universities:

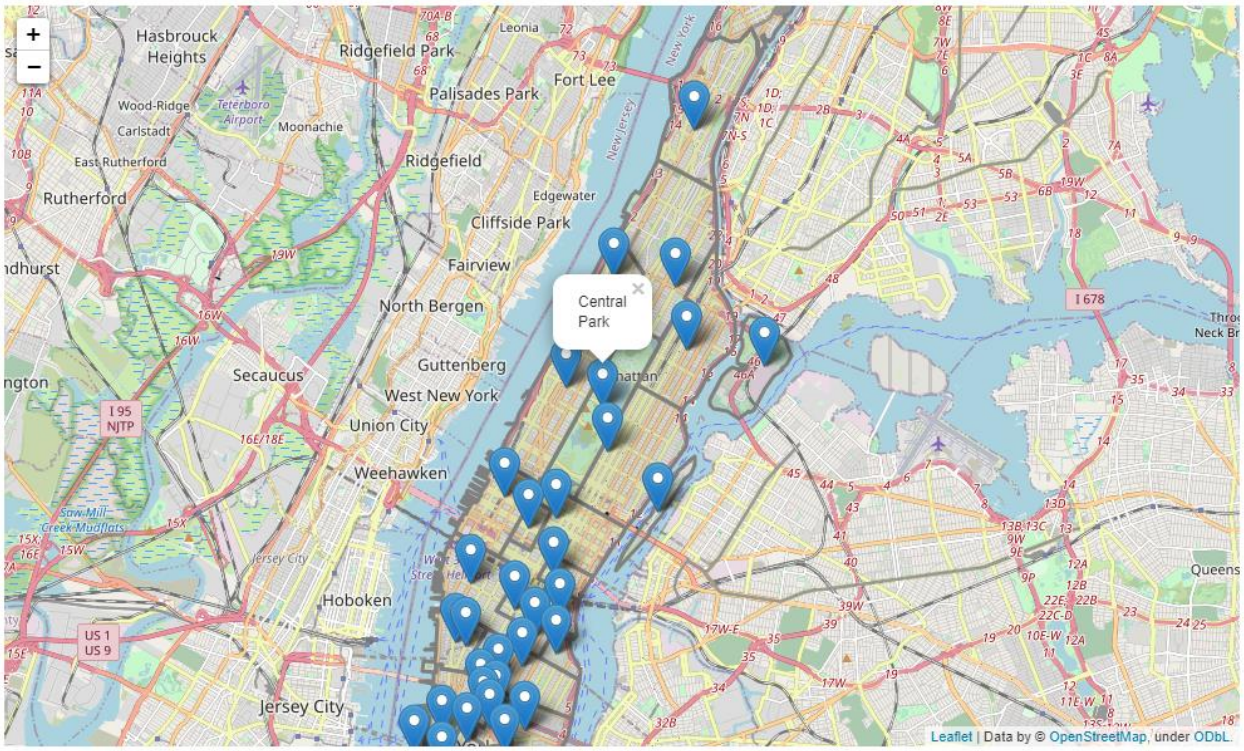
	Name	Latitude	Longitude	Category
0	Icahn School of Medicine at Mount Sinai	40.788249	-73.953110	Medical School
1	New York Presbyterian Hospital Weill Cornell M...	40.764802	-73.954055	Hospital
2	MetroTech Commons	40.694088	-73.985710	College Quad
3	Academy of Our Lady of Good Counsel	41.039467	-73.763842	High School
4	Rutgers University @ UMDNJ	40.740558	-74.191446	University

Using geopy.geocoders, we got the geographical coordinates of the neighborhoods of Manhattan:

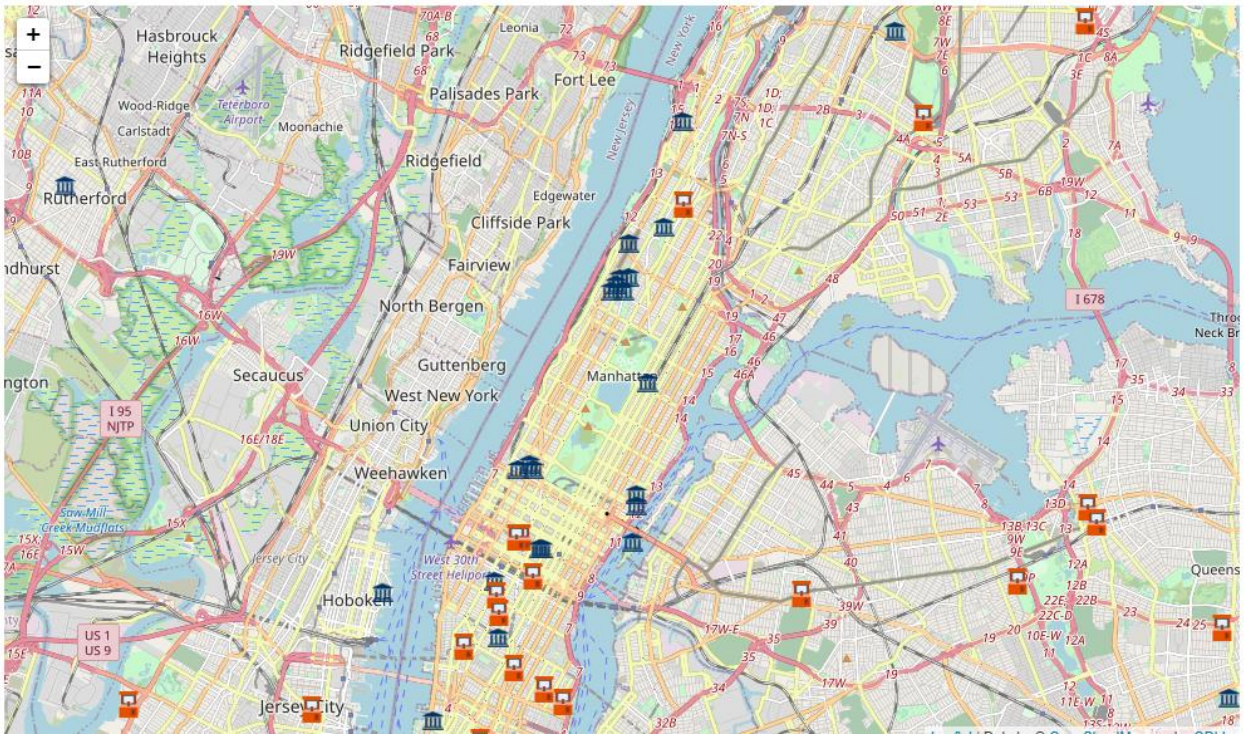
	Neighborhood	Latitude	Longitude
0	Battery Park City	40.711017	-74.016937
1	Central Park	40.782773	-73.965363
2	Chelsea	40.746491	-74.001528
3	Chinatown	40.716491	-73.996250
4	Civic Center	40.713679	-74.002404

We plotted the results using the folium library.

Manhattan neighborhoods:



Universities and Arcades in Manhattan

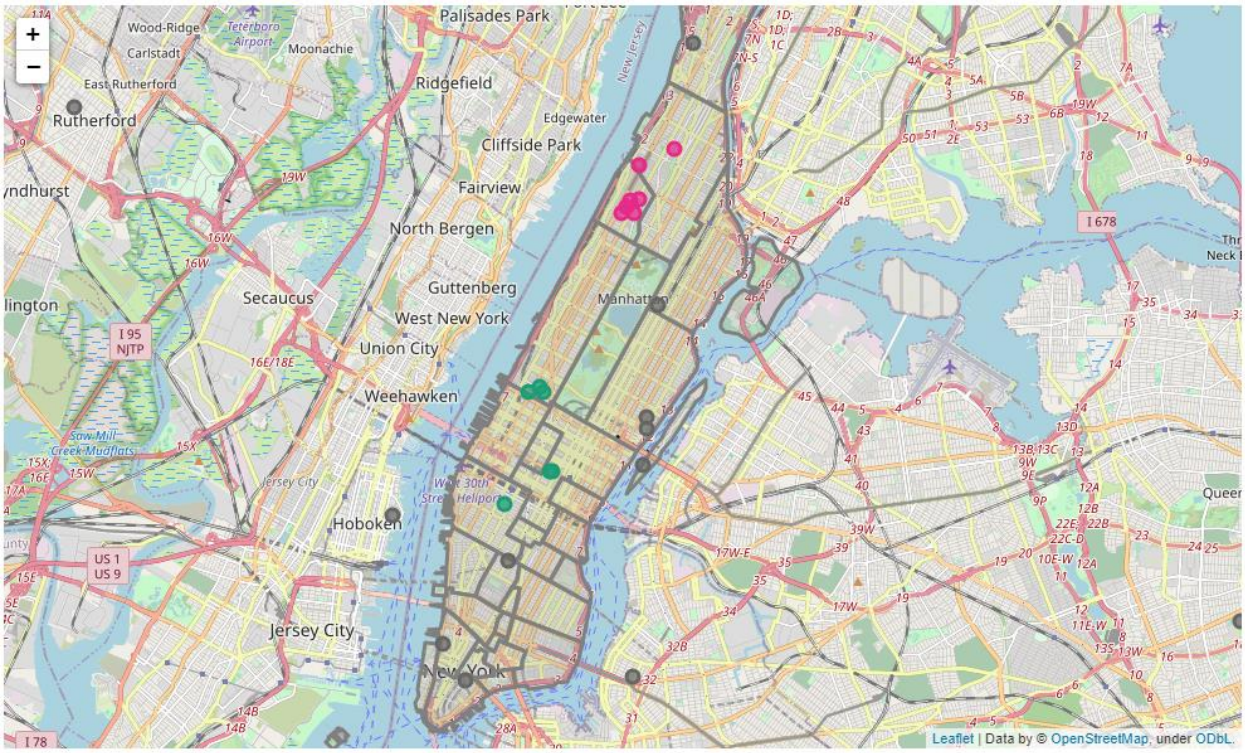


Methodology

We used the location data to find the location with the highest density of Universities. To find this cluster, we will use the **DBSCAN** (Density Based Spatial Clustering of Applications with Noise).

Using recursion, we will use the optimal value of epsilon and choose the min_samples to be = 5. Finally, we will plot the different clusters in different colors and identify the cluster with the highest density. If the number of arcades near this cluster is low, we will find the optimal spot to be near that area.

Analysis



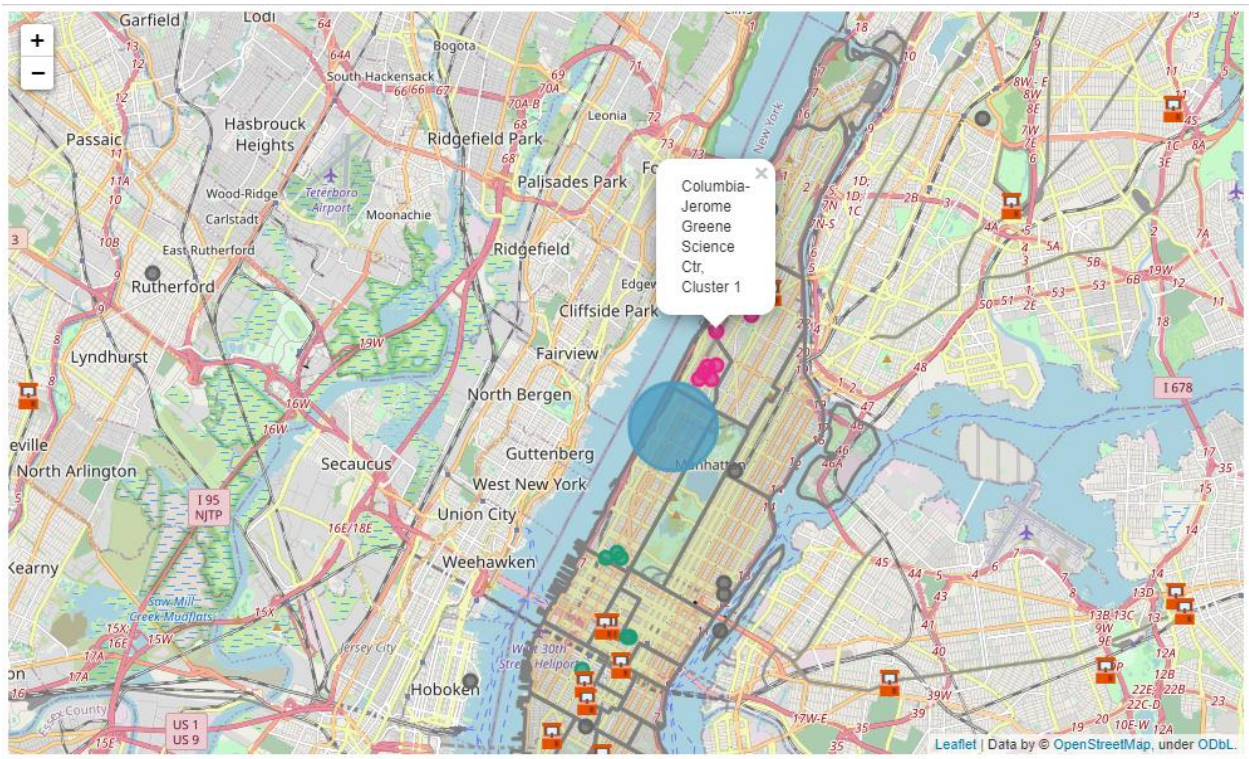
We can see in the output of the DBSCAN clustering algorithm above that the cluster in pink seems to be the densest cluster. We will perform further analysis by adding the arcades back to the map.

Results and Discussion

During the visualization of the data, we found that the northern part of Manhattan had only one arcade. When we plotted the data, we also saw that the density of Universities in the northern part (neighborhoods of Morningside Heights and Harlem) is quite high. This already looks quite promising.

After running a simple DBSCAN on the university locations, we found that indeed, the highest density was to the north of Central Park (colored in green), and the second densest cluster was in the south of Central Park (colored in green). Furthermore, the northern part has fewer arcades than the southern part of the borough.

Conclusion



Using the visualizations, analysis and the results of our clustering model, we can conclude that the ideal location to open an arcade would be just to the south of Morningside Heights. We have found this spot to be in the neighborhood of Upper West Side, in Bloomingdale. The area is marked by a blue circle in the final visualization. We can expect the students of Columbia University to start using this arcade when it is open.