Clustering and Analysis of New York city neighborhoods

Capstone Project - Applied Data Science Capstone by IBM/Coursera

Douglas L. Massuia

September 06, 2020

1. Introduction

1.1. Business Problem

How does New York city neighborhoods correlate to each other and which ones are similar but not necessarily close by physically? How can we group them by venue types and why does that matter?

In this project we will direct our efforts on detecting connections in New York city neighborhoods that aren't visible if you just look at their actual location and coordinate. We want to discover which neighborhoods are related to others based on their venue types, not necessarily those that are close to each other, and how they are spread over the Big Apple.

1.2. Interest

Clustering is considered to be one of the most popular unsupervised machine learning techniques used for grouping data points, or objects that are somehow similar. Cluster analysis has many applications in different domains, whether it be a bank's desire to segment his customers based on certain characteristics, or helping an individual to organize in-group his, or her favorite types of music. Generally speaking though, clustering is used mostly for discovering structure, summarization, and anomaly detection.

This application/algorithm would be able to help solve real world problems like:

- Provide business a tool that would allow them to choose a location to open a new franchise based on their current customer profile segmentation
- Create a customer profile based on their surroundings, or improve your current profiles with more data, better understanding their lifestyle and inclinations
- Provide the public administration sector a tool that would allow them to more deeply understand NYC venues distribution and make better informed choices of what needs to be build/provided and where.
- Select the best neighborhood to buy or rent your new home based on your venue priorities and preferences

2. Data

Data is one of the most important resources a company can have in our current digital era, and yet everyone is currently swimming in a rising tide of data. This data lives in information systems, applications, devices and platforms, and it resides in the surrounding ecosystem – shared by suppliers, supply chain and distribution partners, investors, employees, end-use customers and consumers.

But it’s not just sheer volume. The sources and available types of data also continue to grow exponentially. In fact, 90 percent of the world’s current digital data was produced in the last two years, and the U.S. alone produces upward of 2.6 million gigabytes of internet data every minute.

It's important to make sure the data that is being used to create statistical analysis and models reflect the reality of the situation. For that it is important to take in consideration a few characteristics:

- Accuracy and Precision
- Legitimacy and Validity
- Reliability and Consistency
- Timeliness and Relevance
- Completeness and Comprehensiveness
- Availability and Accessibility
- Granularity and Uniqueness

There are many elements that determine data quality, and each can be prioritized differently by different methods. The prioritization could change depending on the stage of growth of a project.

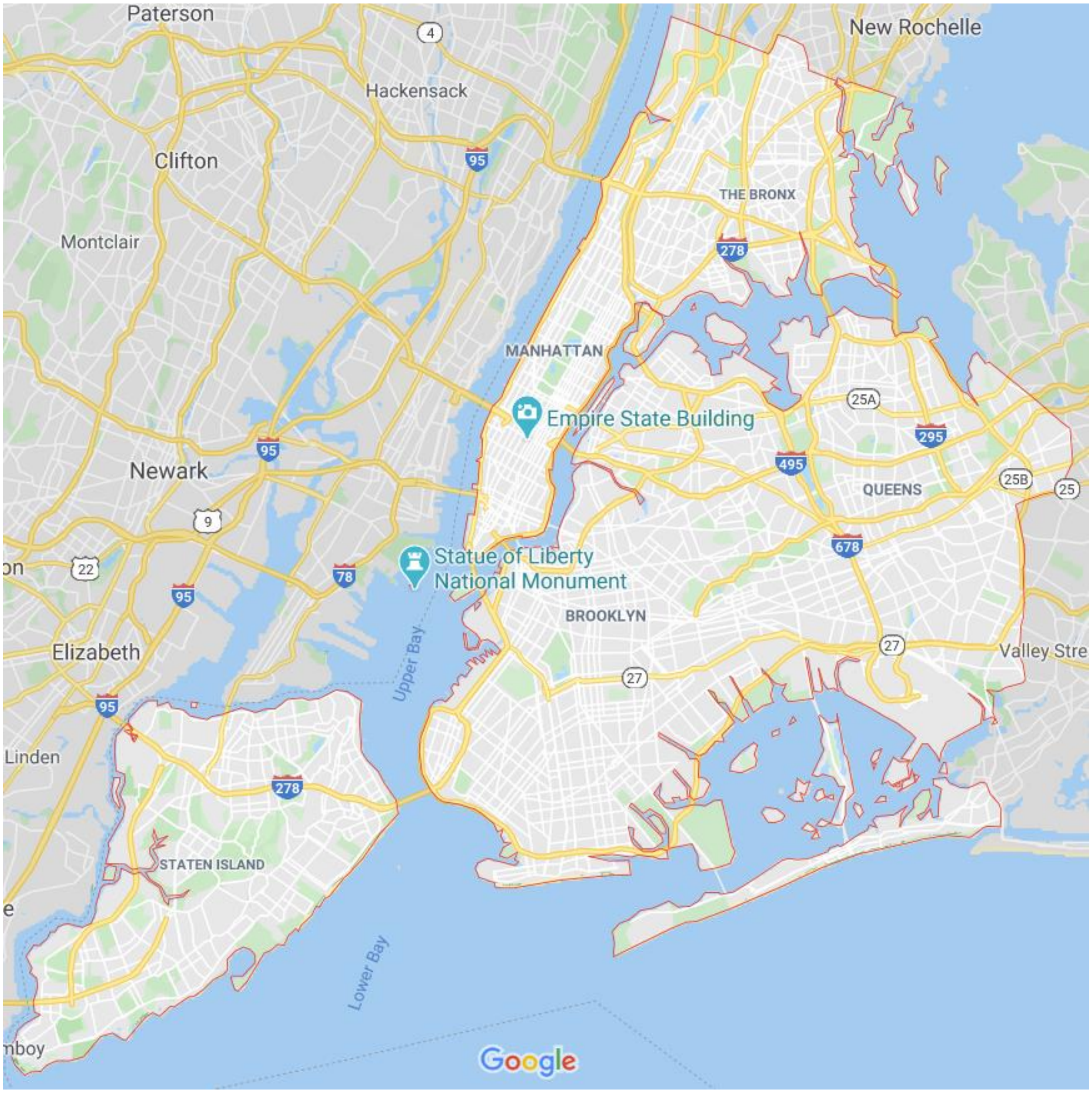
The key is to remember you must define what is most important for your project when evaluating data. Then, use these characteristics to define the criteria for high-quality, accurate data. Once defined, you can be assured of a better understanding and are better positioned to achieve the desired goals.

2.1. Data Source

The data that will be used in this project will come from two different sources:

1) New York University Libraries' portal for GIS data discovery

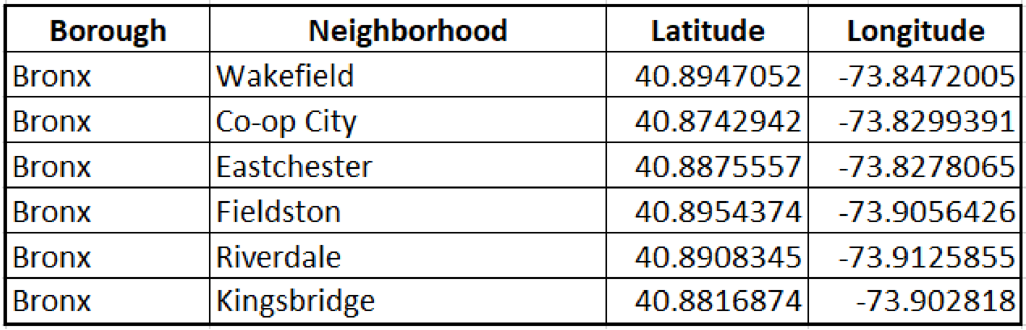
2) Foursquare independent location data platform for Venues



2.2. Get Data

2.2.1 NYU: New York city Neighborhoods, Postal Codes and Coordinates

Link: https://geo.nyu.edu/catalog/nyu_2451_34572



Borough	Neighborhood	Latitude	Longitude
Bronx	Wakefield	40.8947052	-73.8472005
Bronx	Co-op City	40.8742942	-73.8299391
Bronx	Eastchester	40.8875557	-73.8278065
Bronx	Fieldston	40.8954374	-73.9056426
Bronx	Riverdale	40.8908345	-73.9125855
Bronx	Kingsbridge	40.8816874	-73.902818

The DataFrame is composed by 306 Neighborhoods with the following attributes/columns:

- Borough
- Neighborhood
- Latitude
- Longitude

The coordinates will mainly be necessary to correlate with the Foursquare DataFrame for each neighborhood venues.

2.2.2. Foursquare: Venue data for each NY Neighborhood

Use *geopy* library to get the latitude and longitude values of New York City.

In order to define an instance of the geocoder, we need to define a user_agent. We will name our agent ny_explorer.

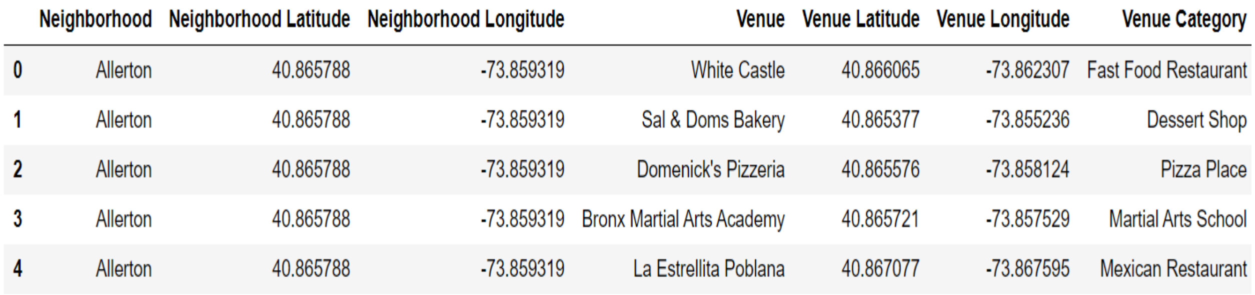
The geographical coordinate of New York City are 40.7127281, -74.0060152

2.2.3. Create getNearbyVenues function

The function getNearbyVenues will get the following information from foursquare databases:

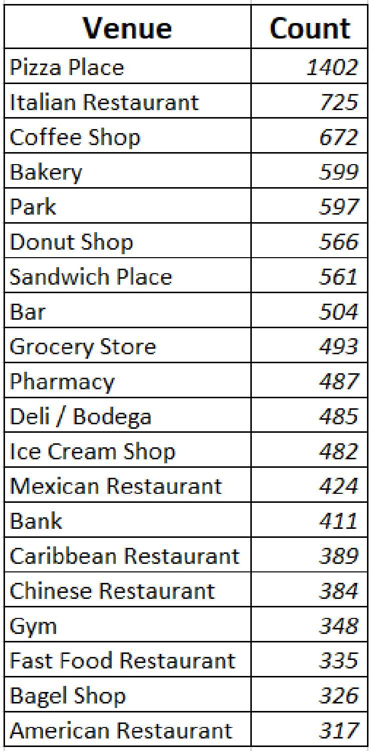
- Venue Name
- Venue Latitude
- Venue Longitude
- Venue Category

We can use it for each neighborhood in our DataFrame and collect the first 100 Venues around it. The result is a DataFrame called nyc_venues with 25.104 rows and 7 columns that has a row for each venue retrieved from the getNearbyVenues as seen below:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allerton	40.865788	-73.859319	White Castle	40.866065	-73.862307	Fast Food Restaurant
1	Allerton	40.865788	-73.859319	Sal & Doms Bakery	40.865377	-73.855236	Dessert Shop
2	Allerton	40.865788	-73.859319	Domenick's Pizzeria	40.865576	-73.858124	Pizza Place
3	Allerton	40.865788	-73.859319	Bronx Martial Arts Academy	40.865721	-73.857529	Martial Arts School
4	Allerton	40.865788	-73.859319	La Estrellita Poblana	40.867077	-73.867595	Mexican Restaurant

Let's take a look at the top 20 most common venues in New York city and their respective counts:



Venue	Count
Pizza Place	*1402*
Italian Restaurant	*725*
Coffee Shop	*672*
Bakery	*599*
Park	*597*
Donut Shop	*566*
Sandwich Place	*561*
Bar	*504*
Grocery Store	*493*
Pharmacy	*487*
Deli / Bodega	*485*
Ice Cream Shop	*482*
Mexican Restaurant	*424*
Bank	*411*
Caribbean Restaurant	*389*
Chinese Restaurant	*384*
Gym	*348*
Fast Food Restaurant	*335*
Bagel Shop	*326*
American Restaurant	*317*

As we can see, Pizza Places, Italian Restaurants and Coffe Shops are the top 3 most common venues, no surprise there knowing the Italian cuisine influence in American culture. In general we can see that restaurants, from all around the world, are the most common venues. Finally, the top 20 venue categories represents about 41% of the total amount of venues types, a respectful count.

3. Methodology

3.1. Introduction

In this project we will direct our efforts on detecting similarities in New York city neighborhoods that aren't visible if you just look at their physical location, we want to discover what neighborhoods are similar to others based on the venue types, not necessarily neighborhoods that are close to each other.

In first step we have collected the required information for the exploratory data analysis (nyc_neighborhoods and nyc_venues). We have limited our analysis to a area ~1.5km around each neighborhood, so that we can get enough venue categories for each neighborhood, improving the accuracy of the model.

The second step in our project is to prepare the data that we'll use to categorize our neighborhoods, as well as creating the machine learning model to cluster them together (using **k-means clustering**). We will finally present a map of all such locations and their respective cluster, identifying how they were connected collectively.

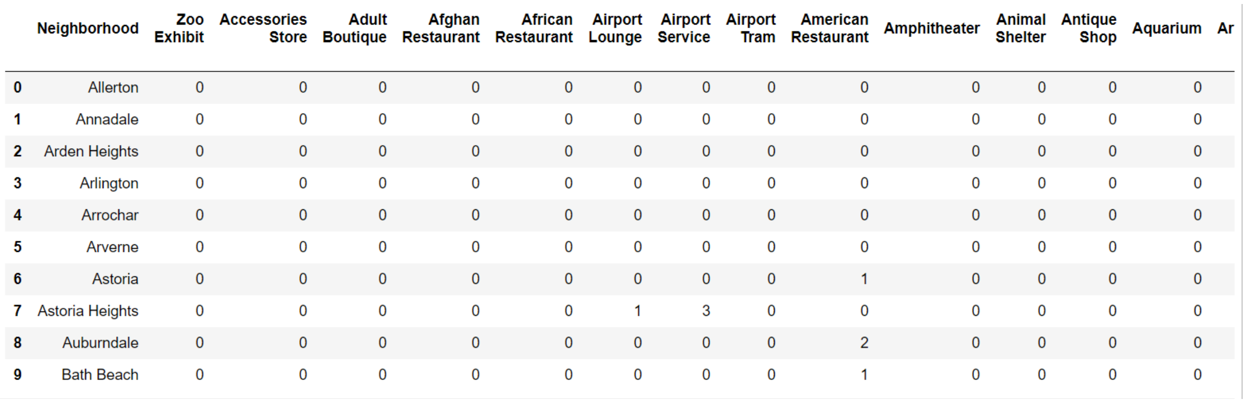
In third and final step we will focus on describing and discussing the results of the clustering model created, understanding why some neighborhoods were defined as they were, what the final outcome can tell us about New York's community and surroundings, as well as wrapping up the project with a final conclusion of the work done.

Let's start with the main variable that will be used for this analysis: the Venue Category, that will be the basis to understand, sort and classify each neighborhood.

3.2. Encoding Categorical Values in "Venue Category"

Because machine learning models require all input and output variables to be numeric, we first need to transform the Venue Category from categorical to quantitative. One great way to do that is through the one-hot encoding, where each bit represents a possible category. If the variable cannot belong to multiple categories at once, then only one bit in the group can be “on”.

We've used the "pandas.get_dummies" function to perform this transformation. There are 451 types of venues in this DataFrame and they'll be the basis to cluster the neighborhoods together. Let's also group all the neighborhoods rows so that all their variables are represented in just one row of data:



	Neighborhood	Zoo Exhibit	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	Airport Tram	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Ar
0	Allerton	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Annadale	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Arden Heights	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Arlington	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Arrochar	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	Arverne	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	Astoria	0	0	0	0	0	0	0	0	1	0	0	0	0	
7	Astoria Heights	0	0	0	0	0	1	3	0	0	0	0	0	0	
8	Auburndale	0	0	0	0	0	0	0	0	2	0	0	0	0	
9	Bath Beach	0	0	0	0	0	0	0	0	1	0	0	0	0	

Each row in the DataFrame above is a neighborhood and each column is a Venue Category, if the value is 0 there are no venue categories of that type in the neighborhood, since the value represents the sum of all venues types.

Finally, with the grouped DataFrame created we can finally start clustering the neighborhoods base on their venues.

3.3. Clustering

A cluster is a group of data points or objects in a dataset that are similar to other objects in the group, and dissimilar to data points in other clusters. In clustering the data is unlabeled and the process is unsupervised.

Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables.

We'll use a K-means, a Partitioned-based Clustering algorithm, that the main objective is to minimize the distance of data points from the centroid of its cluster and maximize the distance from other cluster centroids.

3.4. Clustering Visualization

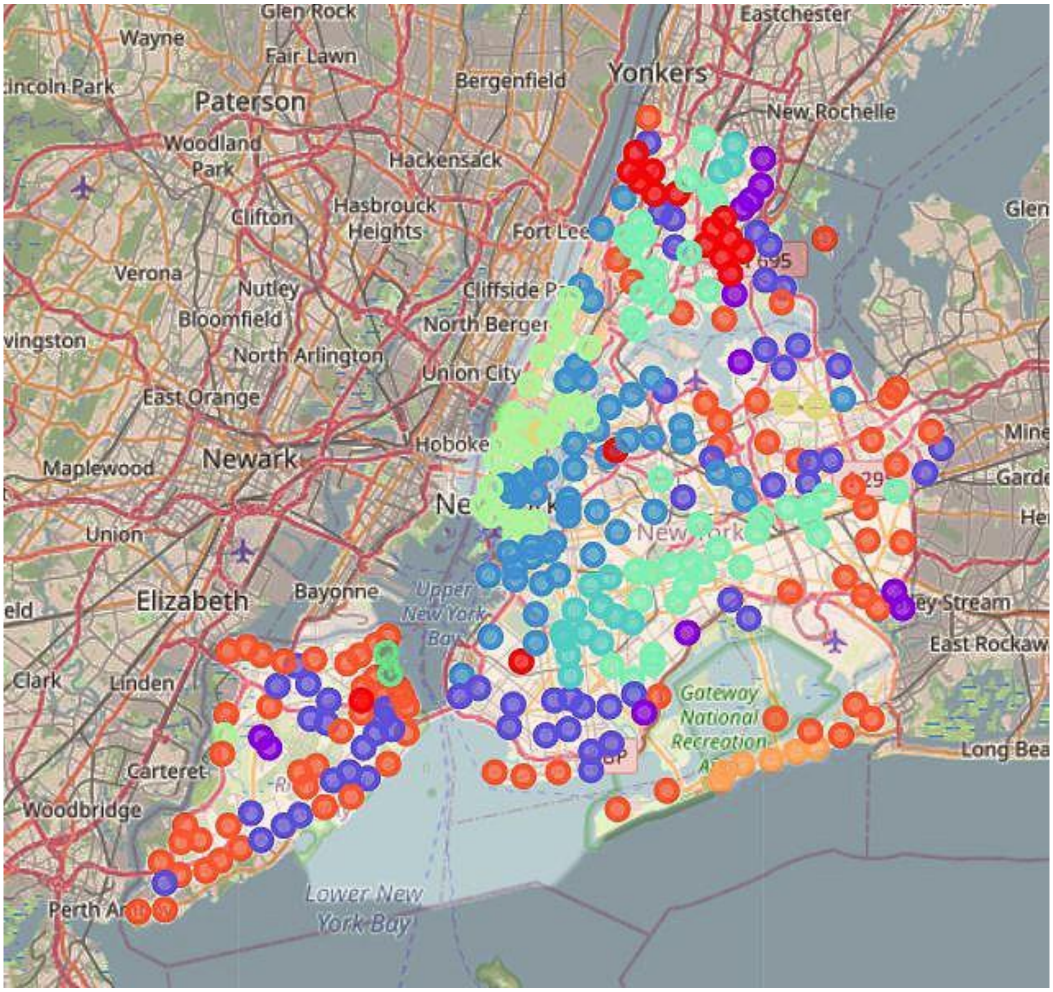
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

3.4.1. Folium Maps

Folium is a powerful Python library that helps you create several types of Leaflet maps. The fact that the Folium results are interactive makes this library very useful for dashboard building.

Folium makes it easy to visualize data that’s been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.



We can see that the visualization of the clustering helps a lot to understand how the neighborhoods were clustered, communicating findings in constructive ways.

This concludes the data exploration, so lets start discussing the results of these clusters and why they were categorized this way in the Results and Discussion section bellow.

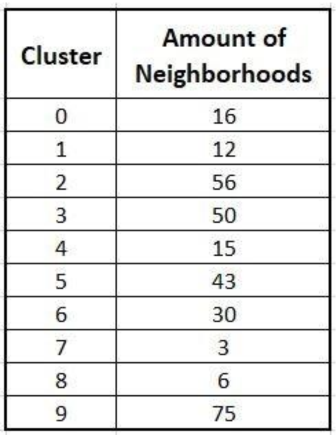
4. Results and Discussion

The Results section of this project represents the core findings of the study derived from the methods applied to gather and analyze information. It presents these findings in a logical sequence without bias or interpretation, setting up the reader for interpretation and evaluation of the data.

A major purpose of the results section is to break down the data into sentences that show its significance to the research question, so lets dig deeper into what neighborhoods are similar to others based on the venue types, not necessarily neighborhoods that are close to each other.

4.1. Understanding each cluster

Let's first understand the size of each cluster:



Cluster	Amount of Neighborhoods
0	16
1	12
2	56
3	50
4	15
5	43
6	30
7	3
8	6
9	75

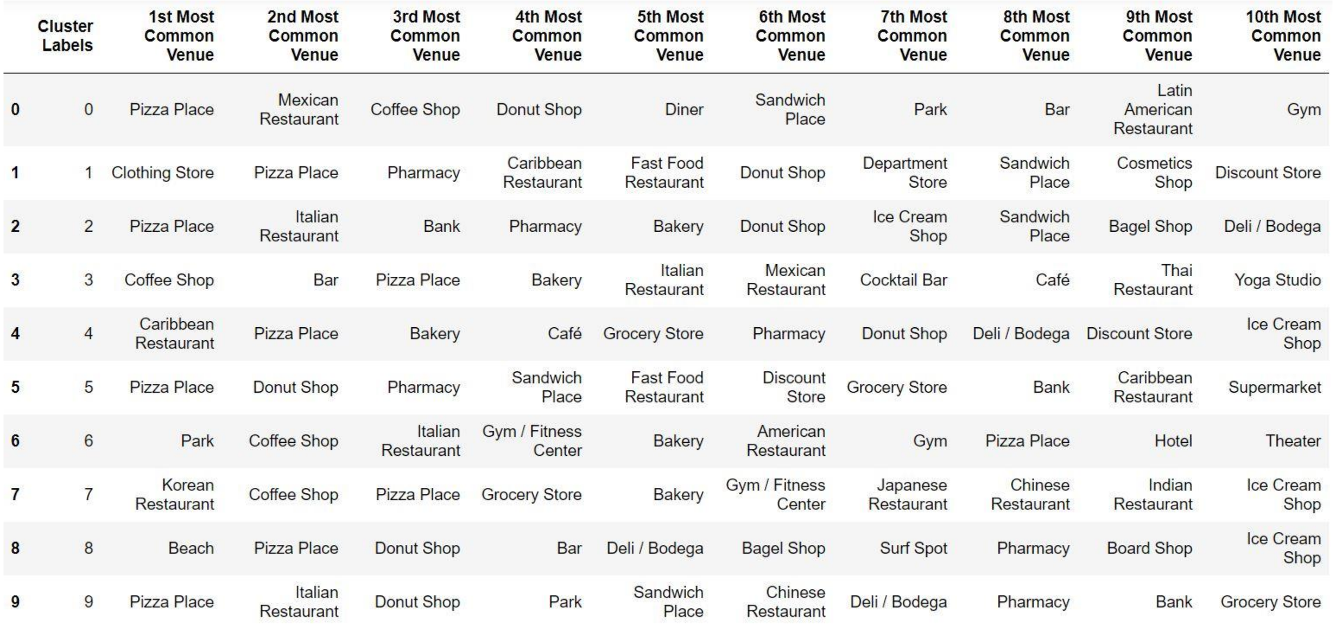
We can see that only 2 clusters have a lower amount of items and the other 8 have a reasonable size.

The clusters are mainly composed by neighborhoods that are close to each other but you can definitely see some items that can be considered outliers by their coordinate disparity.

This means that a neighborhood can be physically far away from their group but it is composed by the same characteristics and venue categories than the other elements inside the same group.

4.2. Understanding each cluster essential venue categories

Let's create a new DataFrame with the top 10 most common venues for each cluster:



	Cluster Labels	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	0	Pizza Place	Mexican Restaurant	Coffee Shop	Donut Shop	Diner	Sandwich Place	Park	Bar	Latin American Restaurant	Gym
1	1	Clothing Store	Pizza Place	Pharmacy	Caribbean Restaurant	Fast Food Restaurant	Donut Shop	Department Store	Sandwich Place	Cosmetics Shop	Discount Store
2	2	Pizza Place	Italian Restaurant	Bank	Pharmacy	Bakery	Donut Shop	Ice Cream Shop	Sandwich Place	Bagel Shop	Deli / Bodega
3	3	Coffee Shop	Bar	Pizza Place	Bakery	Italian Restaurant	Mexican Restaurant	Cocktail Bar	Café	Thai Restaurant	Yoga Studio
4	4	Caribbean Restaurant	Pizza Place	Bakery	Café	Grocery Store	Pharmacy	Donut Shop	Deli / Bodega	Discount Store	Ice Cream Shop
5	5	Pizza Place	Donut Shop	Pharmacy	Sandwich Place	Fast Food Restaurant	Discount Store	Grocery Store	Bank	Caribbean Restaurant	Supermarket
6	6	Park	Coffee Shop	Italian Restaurant	Gym / Fitness Center	Bakery	American Restaurant	Gym	Pizza Place	Hotel	Theater
7	7	Korean Restaurant	Coffee Shop	Pizza Place	Grocery Store	Bakery	Gym / Fitness Center	Japanese Restaurant	Chinese Restaurant	Indian Restaurant	Ice Cream Shop
8	8	Beach	Pizza Place	Donut Shop	Bar	Deli / Bodega	Bagel Shop	Surf Spot	Pharmacy	Board Shop	Ice Cream Shop
9	9	Pizza Place	Italian Restaurant	Donut Shop	Park	Sandwich Place	Chinese Restaurant	Deli / Bodega	Pharmacy	Bank	Grocery Store

The DataFrame above provides a clear view of why these clusters are created and why each element belongs together.

Other main results that are important to highlight:

- "Pizza Place" is in all clusters top 10 common venues
- Cluster 1 main distinctive features are "Clothing Store" and "Pharmacy"
- Cluster 7 has a very wider Asian cuisine influence
- Cluster 8 main distinctive features are "Beach" and "Donut Shop"
- Although "Italian Restaurant" is the second most common item, it is only in 4 out of 10 cluster top 10 common venues, meaning that they might be grouped together instead of being spread all over New York
- Cluster 6 main distinctive features are "Park" and "Coffee Shop"

5. Conclusion

This project had the objective of identifying connections in New York city neighborhoods that aren't evident if only their actual location and coordinates are taken in consideration and explore why those new connections were made.

We initially gathered and cleaned the necessary data from New York University Libraries' portal and Foursquare independent location data platform.

The data was later encoded and clustered bases on their venue characteristics, 10 clusters were created with a total of 302 neighborhoods.

Using Folium as our data visualization tool was possible to analyze their affinities, closeness and correlation.

The results shown is this project can and should be used for many other different purposes of data classification for better decision making.