The Battle of Neighborhoods

Recommending the best location to open a new cafe in New York

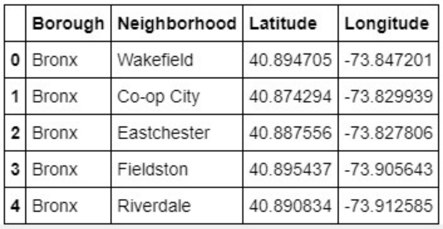
1. Business Problem

A client wants to open a cool new cafe in New York City. She is new to the city and though she loves the whole vibe of the city, she wants to know which would be the best place for her new venture. Since the city has a lot of fast food chains and famous cafes, she is looking for a neighborhood where there are not many famous coffee places - a neighborhood where many people would come to her café and explore. We need to provide her the best location through which she could make the most profit.

2. Data

We would require the location data for New York City to proceed further with our problem statement. Data should describe the geographical location and other details like working hours, cost, ratings, address, etc of all the cafes in the city. With this data, we would be able to cluster the neighborhoods and recommend the right location for our client to open her dream cafe.

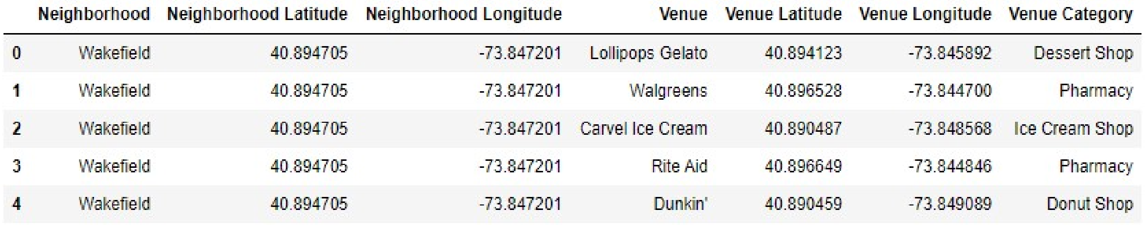
To begin with, we shall be using the New_York data from https://cocl.us/new_york_dataset. It is a geojson file containing all the Boroughs and Neighborhoods of New York along with the geographical co-ordinates. Only the Borough, Neighborhoods, Latitude and Longitude is then filtered out from the json and a new dataframe is created.



	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

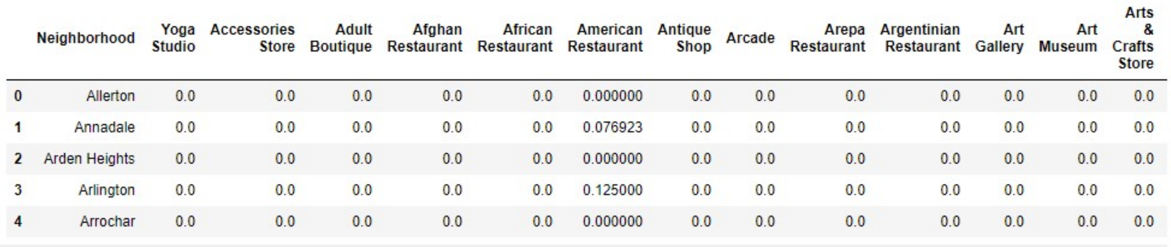
We will then be using the Foursquare location data to explore the neighborhoods. The Foursquare API allows application developers to interact with the Foursquare platform, which is a social location service that allows users to explore the world around them. The API itself is a RESTful set of addresses to which you can send requests and receive a response in json fromat which is easy to work with.

We shall use the explore endpoint for the API calls to obtain the venues near each neighborhood of New York City, with the url https://api.foursquare.com/v2/venues/explore . The response received as a json file is then cleaned and formatted into a dataframe, containing all the venues in each and every neighborhood with the geographical co-ordinates as shown below:



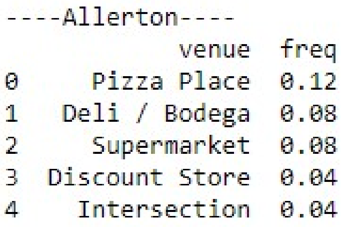
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

This data if further formatted through one-hot encoding to denote the categories of places around each Neighborhood. There are 425 categories amongst the venues and can be denoted as follows:

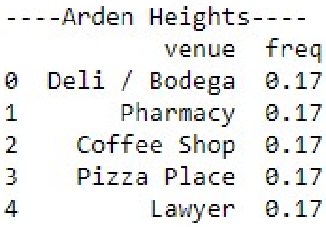


	Neighborhood	Yoga Studio	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.076923	0.0	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	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.125000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

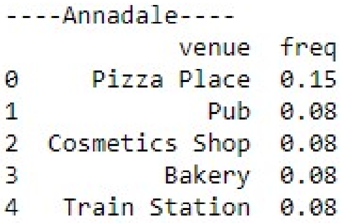
We can also obtain the top 5 venues and their categories in each neighborhood using the Foursquare data. Using this, we could say what kind of places exist in a given neighborhood and the frequency of these places.



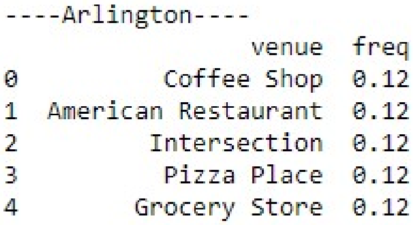
```
----Allerton----
            venue  freq
0     Pizza Place  0.12
1    Deli / Bodega  0.08
2     Supermarket  0.08
3  Discount Store  0.04
4    Intersection  0.04
```



```
----Arden Heights----
           venue  freq
0  Deli / Bodega  0.17
1       Pharmacy  0.17
2    Coffee Shop  0.17
3    Pizza Place  0.17
4         Lawyer  0.17
```

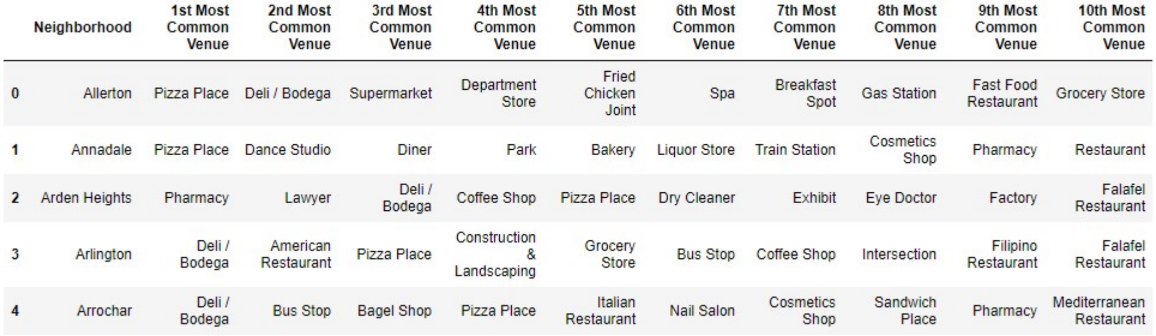


```
----Annadale----
            venue  freq
0     Pizza Place  0.15
1             Pub  0.08
2   Cosmetics Shop  0.08
3          Bakery  0.08
4   Train Station  0.08
```



```
----Arlington----
                 venue  freq
0          Coffee Shop  0.12
1  American Restaurant  0.12
2         Intersection  0.12
3          Pizza Place  0.12
4        Grocery Store  0.12
```

The data has to be now formatted into a dataframe for further analysis. We convert it into a dataframe denoting the categories of 10 most common venues in each of the 301 neighborhoods.



	Neighborhood	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	Allerton	Pizza Place	Deli / Bodega	Supermarket	Department Store	Fried Chicken Joint	Spa	Breakfast Spot	Gas Station	Fast Food Restaurant	Grocery Store
1	Annadale	Pizza Place	Dance Studio	Diner	Park	Bakery	Liquor Store	Train Station	Cosmetics Shop	Pharmacy	Restaurant
2	Arden Heights	Pharmacy	Lawyer	Deli / Bodega	Coffee Shop	Pizza Place	Dry Cleaner	Exhibit	Eye Doctor	Factory	Falafel Restaurant
3	Arlington	Deli / Bodega	American Restaurant	Pizza Place	Construction & Landscaping	Grocery Store	Bus Stop	Coffee Shop	Intersection	Filipino Restaurant	Falafel Restaurant
4	Arrochar	Deli / Bodega	Bus Stop	Bagel Shop	Pizza Place	Italian Restaurant	Nail Salon	Cosmetics Shop	Sandwich Place	Pharmacy	Mediterranean Restaurant

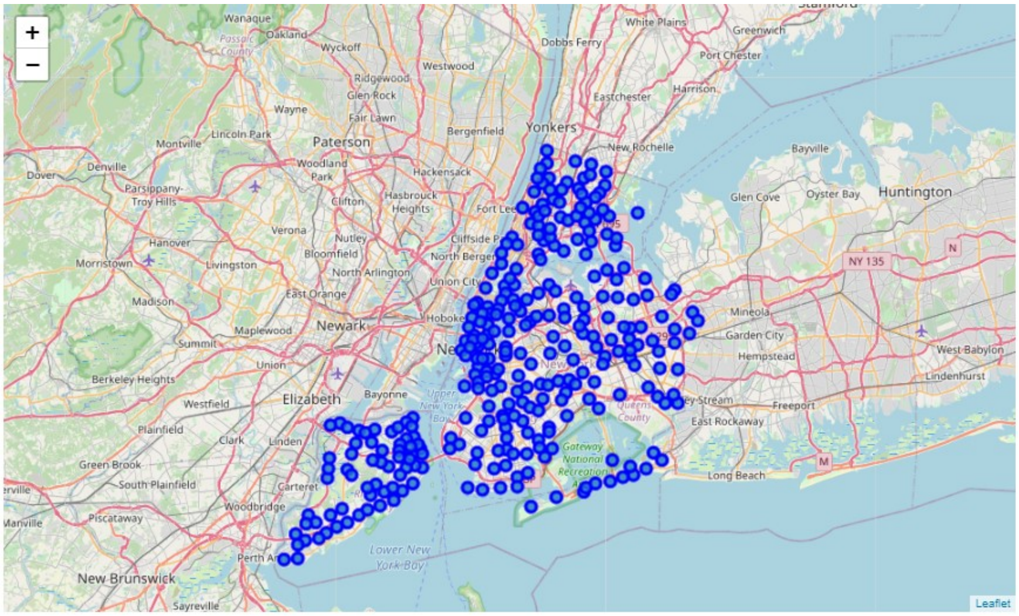
This data is used for clustering the neighborhoods, by which we could arrive at a solution to our problem, i.e., the right place to open the cafe.

3. Methodology

We need to find a neighborhood in the New York City where there are not many famous coffee shops and an establishment of one would bring in huge profits.

For this, we have collected the New York location data and have explored the most common venues in each Neighborhood with the help of Foursquare API. We now have our data in a dataframe as required, with the ten most common venues in each neighborhood of New York. We can now use this data for further analysis using Machine Learning techniques and arrive at a solution.

Below is a map of New York City with all the neighborhoods shown. Out of these, we need to recommend a particular neighborhood for our client to start her new venture.



We will use the k-means clustering algorithm to cluster the neighborhoods based on the types of venues that exist in the vicinity.

K-Means Clustering: Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same cluster are very similar while data points in different clusters are very different. K-means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

The way k-means algorithm works is as follows:

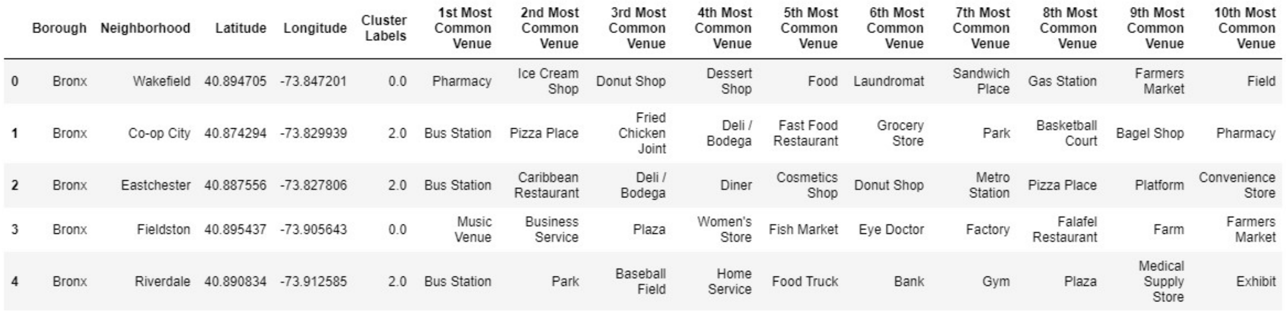
- Specify number of clusters *K*.
- Initialize centroids randomly selecting *K* data points as the centroids.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.
- Keep iterating until there is no change to the centroids i.e. assignment of data points to clusters isn’t changing.

For our problem, we will be using k-means clustering algorithm on our dataset to obtain 5 clusters of the neighborhoods. The neighborhoods within the same cluster will have similar kind of venues. For example, if a neighborhood has too many coffee shops, restaurants and pubs, all the neighborhoods belonging to that particular cluster will have similar surroundings. All the areas are very happening and are busy streets where people hangout frequently.

By applying k-means clustering on our dataset and then analyzing the resulting clusters, we would be able to make our recommendations to our client.

4. Results

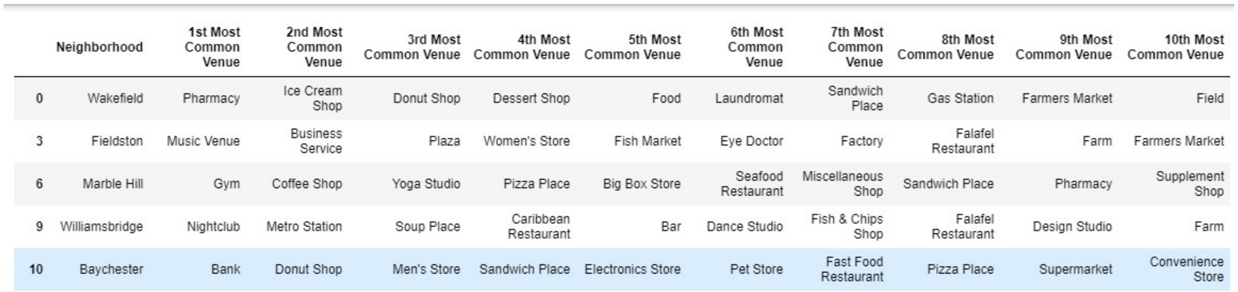
K-Means Clustering is used to form clusters from the Neighborhood data we have gathered. We have created 5 clusters based on the ten most common venues of each neighborhood. A glimpse of the resulting dataframe is seen below, the cluster labels are also added to denote which cluster a particular neighborhood belongs to.



	Borough	Neighborhood	Latitude	Longitude	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	Bronx	Wakefield	40.894705	-73.847201	0.0	Pharmacy	Ice Cream Shop	Donut Shop	Dessert Shop	Food	Laundromat	Sandwich Place	Gas Station	Farmers Market	Field
1	Bronx	Co-op City	40.874294	-73.829939	2.0	Bus Station	Pizza Place	Fried Chicken Joint	Deli / Bodega	Fast Food Restaurant	Grocery Store	Park	Basketball Court	Bagel Shop	Pharmacy
2	Bronx	Eastchester	40.887556	-73.827806	2.0	Bus Station	Caribbean Restaurant	Deli / Bodega	Diner	Cosmetics Shop	Donut Shop	Metro Station	Pizza Place	Platform	Convenience Store
3	Bronx	Fieldston	40.895437	-73.905643	0.0	Music Venue	Business Service	Plaza	Women's Store	Fish Market	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market
4	Bronx	Riverdale	40.890834	-73.912585	2.0	Bus Station	Park	Baseball Field	Home Service	Food Truck	Bank	Gym	Plaza	Medical Supply Store	Exhibit

The clusters are then analyzed further to understand the types of places present in the neighborhood.

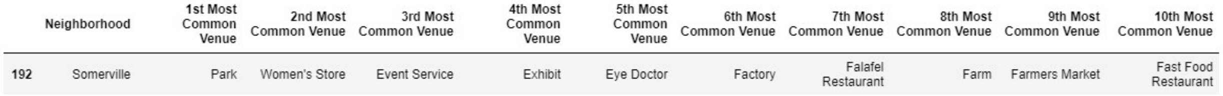
Cluster 1:



	Neighborhood	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	Wakefield	Pharmacy	Ice Cream Shop	Donut Shop	Dessert Shop	Food	Laundromat	Sandwich Place	Gas Station	Farmers Market	Field
3	Fieldston	Music Venue	Business Service	Plaza	Women's Store	Fish Market	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market
6	Marble Hill	Gym	Coffee Shop	Yoga Studio	Pizza Place	Big Box Store	Seafood Restaurant	Miscellaneous Shop	Sandwich Place	Pharmacy	Supplement Shop
9	Williamsbridge	Nightclub	Metro Station	Soup Place	Caribbean Restaurant	Bar	Dance Studio	Fish & Chips Shop	Falafel Restaurant	Design Studio	Farm
10	Baychester	Bank	Donut Shop	Men's Store	Sandwich Place	Electronics Store	Pet Store	Fast Food Restaurant	Pizza Place	Supermarket	Convenience Store

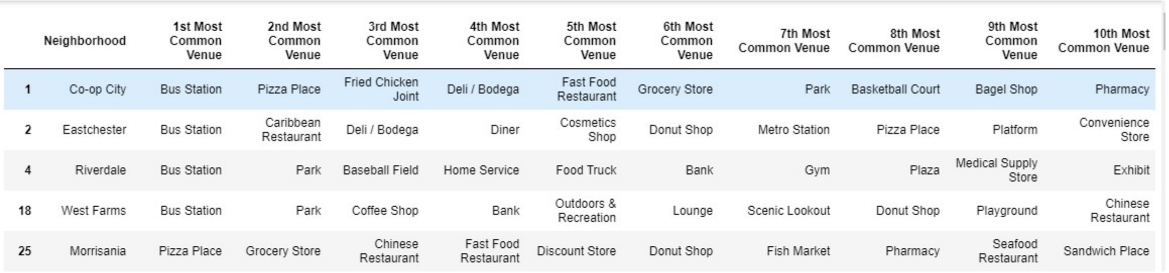
(displaying only top 5 rows as the resultant dataframe is too large)

Cluster 2:



	Neighborhood	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
192	Somerville	Park	Women's Store	Event Service	Exhibit	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant

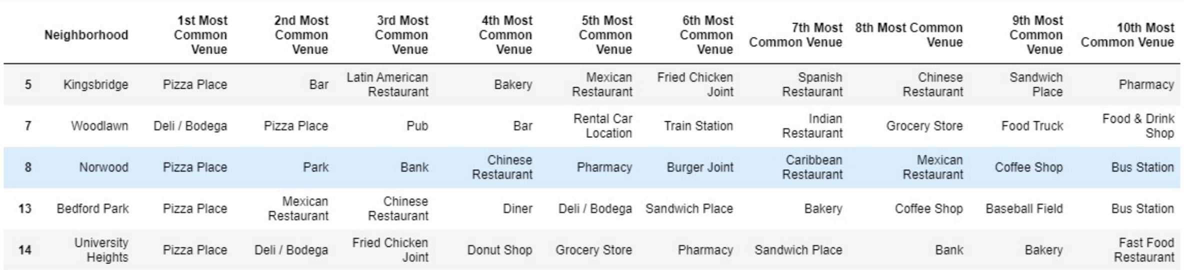
Cluster 3:



	Neighborhood	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
1	Co-op City	Bus Station	Pizza Place	Fried Chicken Joint	Deli / Bodega	Fast Food Restaurant	Grocery Store	Park	Basketball Court	Bagel Shop	Pharmacy
2	Eastchester	Bus Station	Caribbean Restaurant	Deli / Bodega	Diner	Cosmetics Shop	Donut Shop	Metro Station	Pizza Place	Platform	Convenience Store
4	Riverdale	Bus Station	Park	Baseball Field	Home Service	Food Truck	Bank	Gym	Plaza	Medical Supply Store	Exhibit
18	West Farms	Bus Station	Park	Coffee Shop	Bank	Outdoors & Recreation	Lounge	Scenic Lookout	Donut Shop	Playground	Chinese Restaurant
25	Morrisania	Pizza Place	Grocery Store	Chinese Restaurant	Fast Food Restaurant	Discount Store	Donut Shop	Fish Market	Pharmacy	Seafood Restaurant	Sandwich Place

(displaying only top 5 rows as the resultant dataframe is too large)

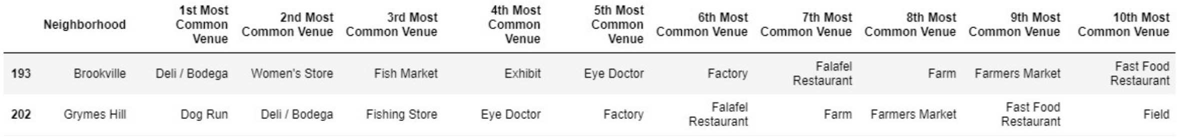
Cluster 4:



	Neighborhood	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
5	Kingsbridge	Pizza Place	Bar	Latin American Restaurant	Bakery	Mexican Restaurant	Fried Chicken Joint	Spanish Restaurant	Chinese Restaurant	Sandwich Place	Pharmacy
7	Woodlawn	Deli / Bodega	Pizza Place	Pub	Bar	Rental Car Location	Train Station	Indian Restaurant	Grocery Store	Food Truck	Food & Drink Shop
8	Norwood	Pizza Place	Park	Bank	Chinese Restaurant	Pharmacy	Burger Joint	Caribbean Restaurant	Mexican Restaurant	Coffee Shop	Bus Station
13	Bedford Park	Pizza Place	Mexican Restaurant	Chinese Restaurant	Diner	Deli / Bodega	Sandwich Place	Bakery	Coffee Shop	Baseball Field	Bus Station
14	University Heights	Pizza Place	Deli / Bodega	Fried Chicken Joint	Donut Shop	Grocery Store	Pharmacy	Sandwich Place	Bank	Bakery	Fast Food Restaurant

(displaying only top 5 rows as the resultant dataframe is too large)

Cluster 5:

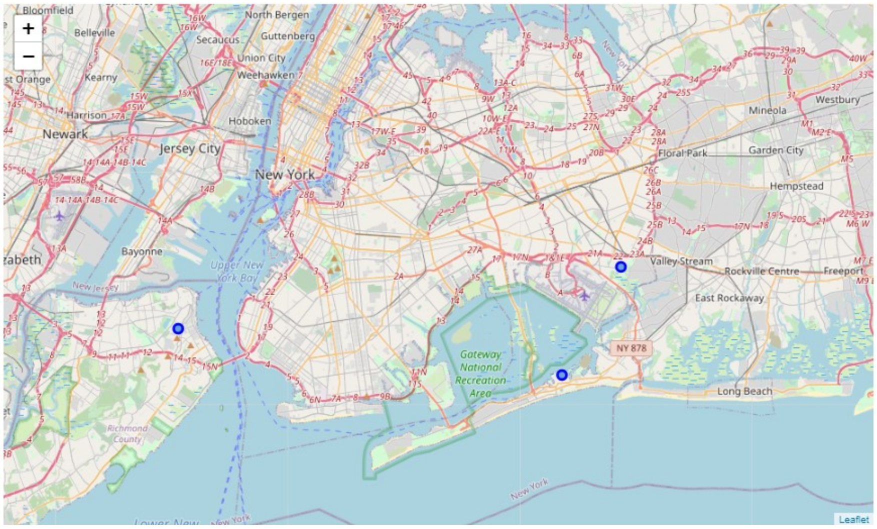


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
193	Brookville	Deli / Bodega	Women's Store	Fish Market	Exhibit	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
202	Grymes Hill	Dog Run	Deli / Bodega	Fishing Store	Eye Doctor	Factory	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Field

5. Discussion

From the above analysis, it is clear that the neighborhoods in clusters 1, 3 and 4 are busy streets with a lot of famous eateries and restaurants. They are already filled with bars, fast food restaurants, dessert places, pizza places, cafes and what not. These neighborhoods are not that suitable for a new café as people already have their favourite spots here and usually hangout in the same places. There is a lot of competition and it would be not that easy to make a famous name if a new café is opened here.

However, clusters 2 and 5 do not have many famous options. Cluster 2 has parks, women's stores, grocery stores, etc and Cluster 5 has farms, eye doctors, factories, dog-run places, etc. People do come here frequently for regular eye check-ups or to run their dogs or women for some pampering. There might also be people working in the factories or farms who daily come to the neighborhood. However, there are no famous eateries or cafes here. If a new café is launched here, it would be great as there are too many people who come to the neighborhood regularly and would want to grab a cup of coffee. The people would also be very glad if they could get some nice place to chill with some great coffee in their vicinity. It would be a great profit for our client as well. Hence, it is recommended to open a cafe in the neighborhoods of these clusters**.**



6. Conclusion

In this study, the neighborhoods of New York City are explored. The most famous venues in each neighborhood are identified and the neighborhoods are grouped into various clusters. On further analysis of the clusters, we have understood the types of places that are famous in a given neighborhood and hence have recommended our client to open her new café in the lesser popular neighborhoods - Brookville, Grymes Hill or Somerville.

The analysis can be extended further to recommend a location for any type of venues to be opened. For example, since there are a couple of Eye Doctors, farms and factories famous in the cluster 5, a new Pharmacy in the neighborhood would do pretty good in case of emergency. Similarly, this model can help for many other recommendations in New York City.