Capstone Project

The Battle of Neighborhoods Reports.

Zaheer Habib, Coursera final project submission 2019

Introduction.

Description of the background of the Project

New York is the popular city of the united states with an estimate of around population of 8,398,748 which is distributed over 302.6 square miles. Its is also most densely populated city which is in Sothern tip and city is the center of NY metropolitan area.

NYK consists of five boroughs – Brooklyn, Queens, Manhattan, The Bronx, and Staten Island. Many districts and landmark in NYC are famous; including three of the world's ten most visited tourist attractions in 2013, around 62.8 million tourist visited in 2017.

< https://en.wikipedia.org/wiki/New_York_City**>**

"New York City is multicultural. About 36% of the city's population is foreign-born,[23] one of the highest among US cities. The eleven nations constituting the largest sources of modern immigration, to New York City are to New York City are the Dominican Republic, China, Jamaica, Guyana, Mexico, Ecuador, Brazil, Haiti, Trinidad and Tobago, Colombia, Russia and El Salvador".[24]

https://en.wikipedia.org/wiki/Demographics_of_New_York_City

Requirement analysis.

An entrepreneur from China like to invest in New York city (NYC); he is very enthusiastic businessmen having sound business implementation knowledge. So, he contacts with the local technology company to explore the best area in New York which he will invest so that he gets quick return on investment (ROI). Subsequently, to get maximum ROI he pushes developer to find the area in NYC where he will establish his business but only constrain is, he like to invest in the business which is traditionally belongs to Chines culture. Example Chines restaurant, SPA or retail shop which deal in chines herbs. So below are the two requirement which needs to consideration.

- Investment should need to make in NYK best potential borough and its neighborhood area to get maximum ROI
- Technology advise entrepreneur in related to Chinese culture like Chinese restaurant, SPA or retail herbal shop etc.

Problem

The Data that might contribute to determining the explore information to get best location which will further helps to identify the metrics that best describe what kind of business he supposed to put his capital. This project aims to predict the area and type of business which the entrepreneur should needs to invest to get maximum ROI.

Interest.

This project aims to predict the area and type of business which the entrepreneur should needs to invest to get maximum ROI.

Data acquisition and cleaning.

The download Json file is processed accordingly as require and all outliers, irrelevant data are filter out before use for next stage of the analysis to make process more efficient and error free.

Target variables, lat & long is created as require and further filter out at later stages when this is not further required. Tabular data is created throughout the project for analysis data and better understating of the derived information which helps to further drilled down as required.

To consider the problem we get through the below sites to get data

- Forsquare API to get the most common venues of given Borough and neighborhood.
- Coursera Lab NYC borough json data: https://geo.nyu.edu/catalog/nyu_2451_34572

Methodology & Exploratory data analysis

To come up with the finding we were using different concepts, tools and methods which is mention below.

- GitHub repository is used to deposit files and share our work with community
- Tabular data which has main component consist of Borough, …………
- Tabular data which has main component consist of Venue Frequency
- Forsquare API to explore the neighborhood of the borough
- **folium** library to visualize geographic details through choropleth plotting.
- Geolocator is used to get geocode(address) latitude and longitude.

Results:

In summary section, one of our aim is to visualize the requirement to plot in choropleth style; when reviewing all the basic requirement we were consider all these problems, we created a initial map using choropleth style to plot our existence borough and further NEW choropleth style map is created based on derived data which is done according to clusters which is generated through k-means algorithm where this clustered is highlighted according to the venue density in respective brought. This derived information have the information which is help to analysis and created our final map and different tabular information.

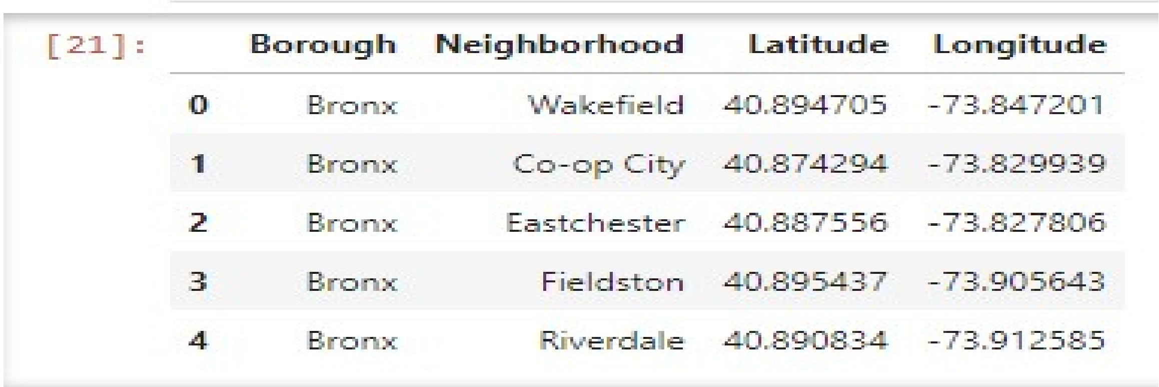
- Borough name
- Cluster name
- Venue
- Frequency
- Lat & Long

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Explring newyork_data json file through features method

```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```

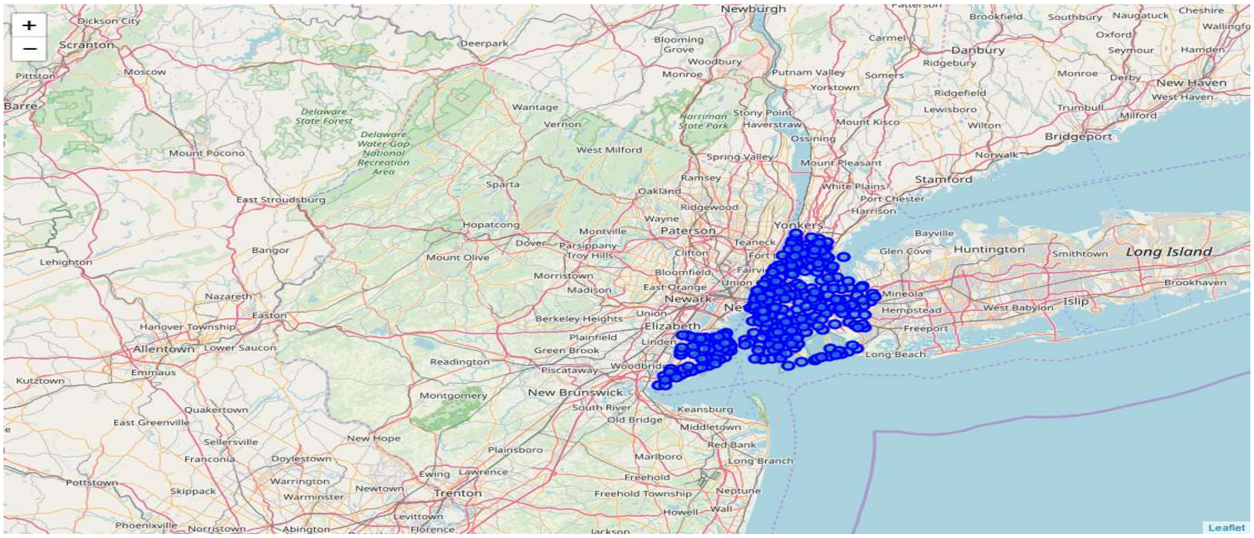
Exploring neighborhoods: total 5 boroughs and 306 neighborhoods hit



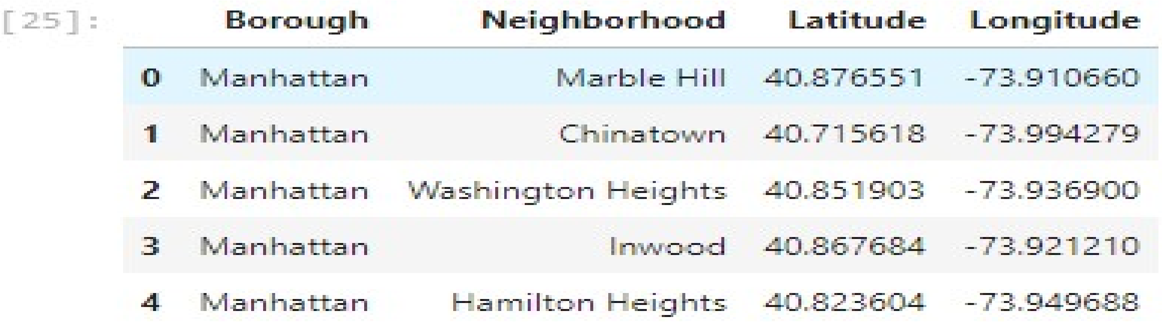
[21]:

	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

Generating NYC map superimposed of neighborhood using Folium library.



simplify the above map and segment and cluster only the neighborhoods in Manhattan

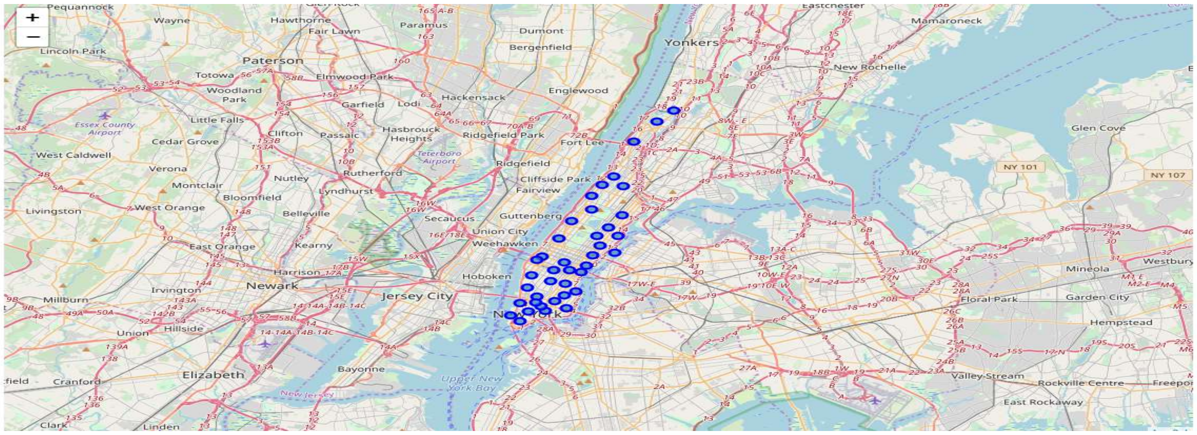


[25]:

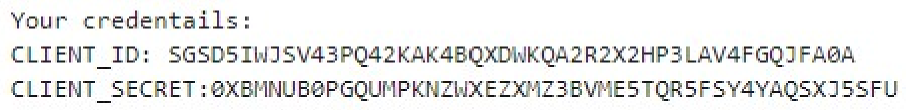
	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Generate Manhattan map.

[27]:

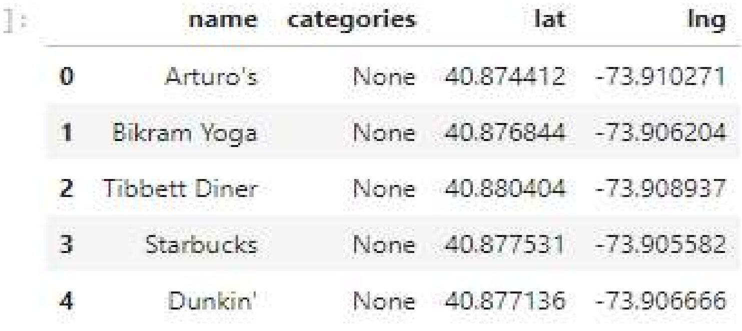


Get credential from Foursquare Credentials and Version



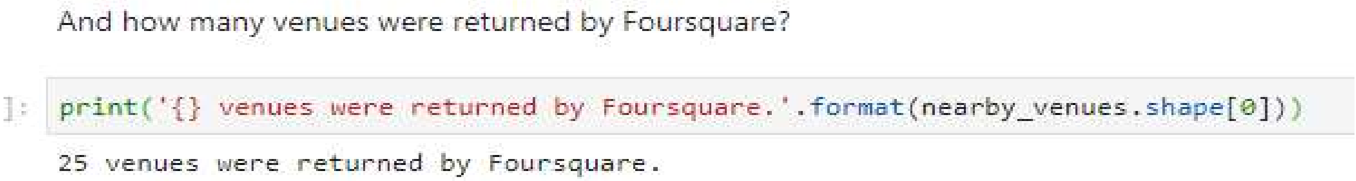
```
Your credentails:
CLIENT_ID: SGSD5IWJSV43PQ42KAK4BQXDWKQA2R2X2HP3LAV4FGQJFA0A
CLIENT_SECRET:0XBMNUB0PGQUMPKNZWXEZXMZ3BVME5TQR5FSY4YAQSXJ5SFU
```

Get Latitude and longitude values of Marble Hill are 40.87655077879964, -73.91065965862981 and generate map 100 topmost neighborhood near 500 radius. When send request to foursquare we get 25 venue.



]:

	name	categories	lat	lng
0	Arturo's	None	40.874412	-73.910271
1	Bikram Yoga	None	40.876844	-73.906204
2	Tibbett Diner	None	40.880404	-73.908937
3	Starbucks	None	40.877531	-73.905582
4	Dunkin'	None	40.877136	-73.906666



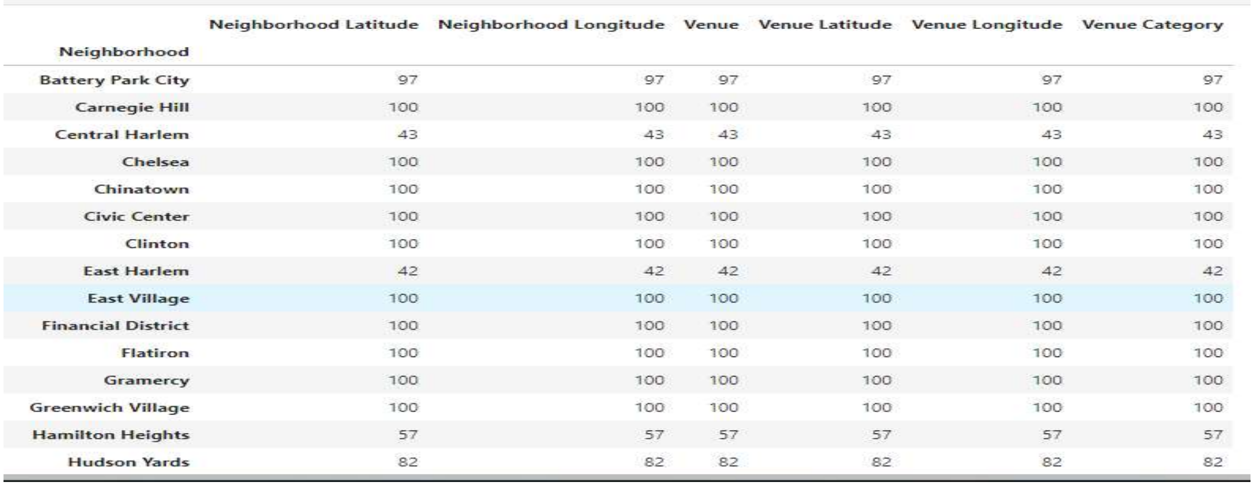
And how many venues were returned by Foursquare?

```
]: print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
25 venues were returned by Foursquare.
```

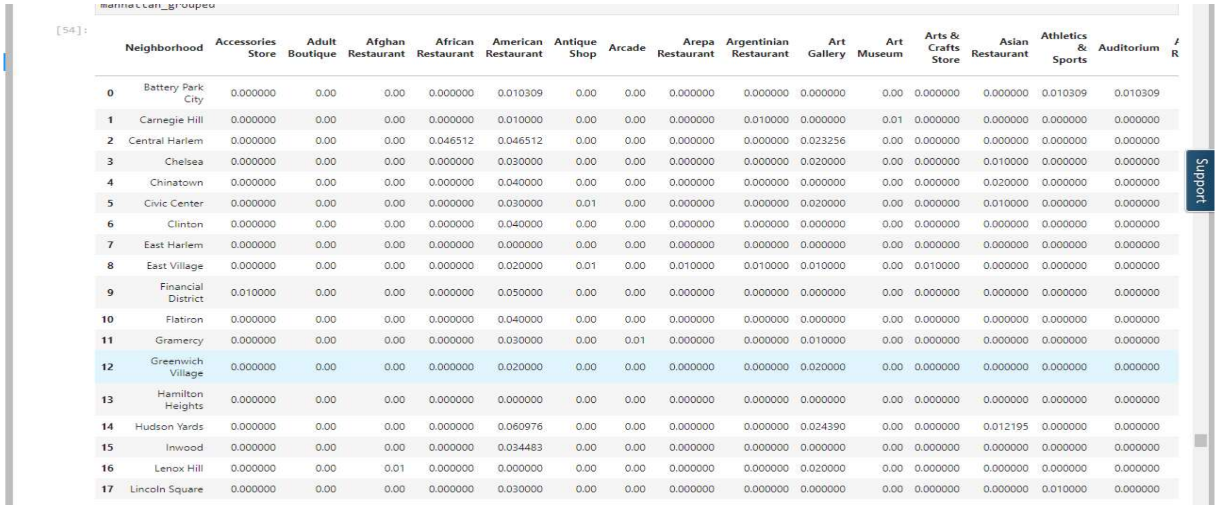
Exploring the Manhattan neighborhood 3312, 7 record hit. While 336 unique category return

-



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Battery Park City	97	97	97	97	97	97
Carnegie Hill	100	100	100	100	100	100
Central Harlem	43	43	43	43	43	43
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100
Civic Center	100	100	100	100	100	100
Clinton	100	100	100	100	100	100
East Harlem	42	42	42	42	42	42
East Village	100	100	100	100	100	100
Financial District	100	100	100	100	100	100
Flatiron	100	100	100	100	100	100
Gramercy	100	100	100	100	100	100
Greenwich Village	100	100	100	100	100	100
Hamilton Heights	57	57	57	57	57	57
Hudson Yards	82	82	82	82	82	82

Group the neighbor hood and take mean of the frequency of each occurrence. Get the new size (40, 337)

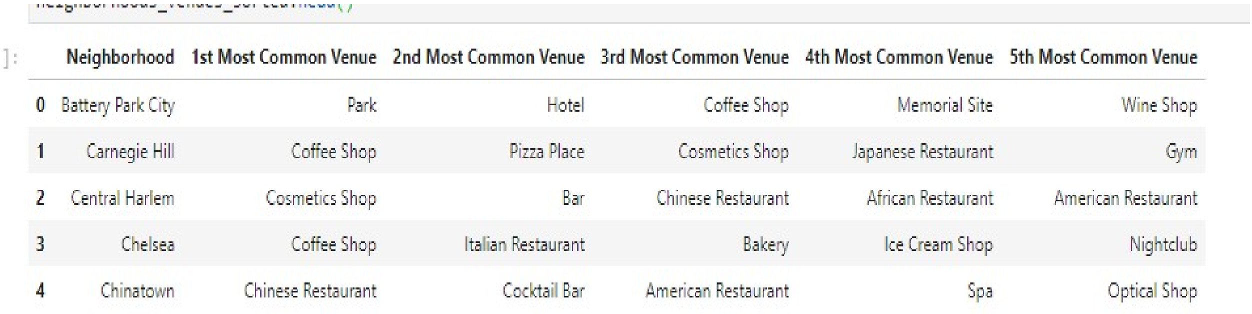


```
manhattan_grouped
```

[54]:

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010309	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.010309	0.010309
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.000000	0.010000	0.000000	0.01	0.000000	0.000000	0.000000	0.000000
2	Central Harlem	0.000000	0.00	0.00	0.046512	0.046512	0.00	0.00	0.000000	0.000000	0.023256	0.00	0.000000	0.000000	0.000000	0.000000
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	0.020000	0.00	0.000000	0.010000	0.000000	0.000000
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.020000	0.000000	0.000000
5	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.01	0.00	0.000000	0.000000	0.020000	0.00	0.000000	0.010000	0.000000	0.000000
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.01	0.00	0.010000	0.010000	0.010000	0.00	0.010000	0.000000	0.000000	0.000000
9	Financial District	0.010000	0.00	0.00	0.000000	0.050000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
10	Flatiron	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
11	Gramercy	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.01	0.000000	0.000000	0.010000	0.00	0.000000	0.000000	0.000000	0.000000
12	Greenwich Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.00	0.000000	0.000000	0.020000	0.00	0.000000	0.000000	0.000000	0.000000
13	Hamilton Heights	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
14	Hudson Yards	0.000000	0.00	0.00	0.000000	0.060976	0.00	0.00	0.000000	0.000000	0.024390	0.00	0.000000	0.012195	0.000000	0.000000
15	Inwood	0.000000	0.00	0.00	0.000000	0.034483	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000
16	Lenox Hill	0.000000	0.00	0.01	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.020000	0.00	0.000000	0.000000	0.000000	0.000000
17	Lincoln Square	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.010000	0.000000

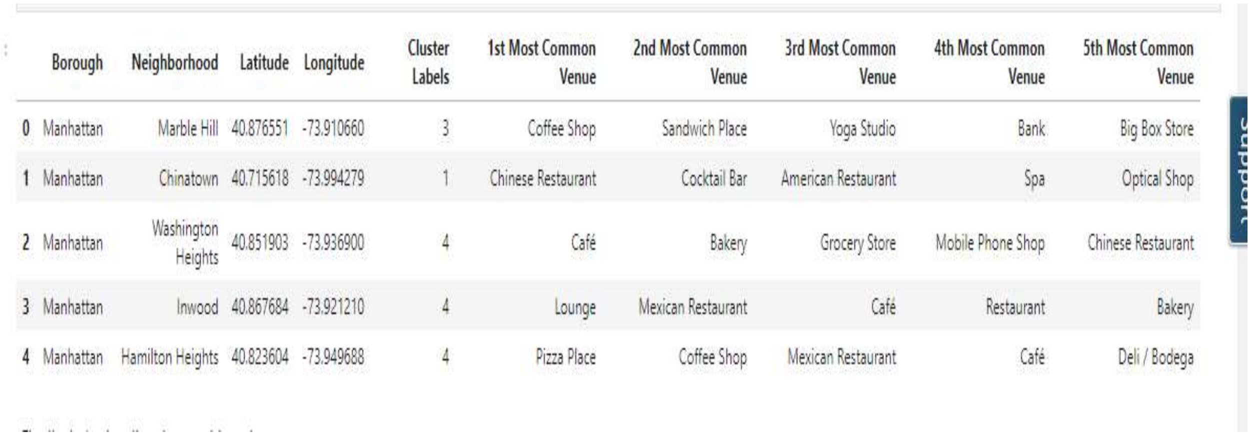
Neighborhood along with 5 most common venue



```
neighborhoods_venues_sorted.head()
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	Park	Hotel	Coffee Shop	Memorial Site	Wine Shop
1	Carnegie Hill	Coffee Shop	Pizza Place	Cosmetics Shop	Japanese Restaurant	Gym
2	Central Harlem	Cosmetics Shop	Bar	Chinese Restaurant	African Restaurant	American Restaurant
3	Chelsea	Coffee Shop	Italian Restaurant	Bakery	Ice Cream Shop	Nightclub
4	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Spa	Optical Shop

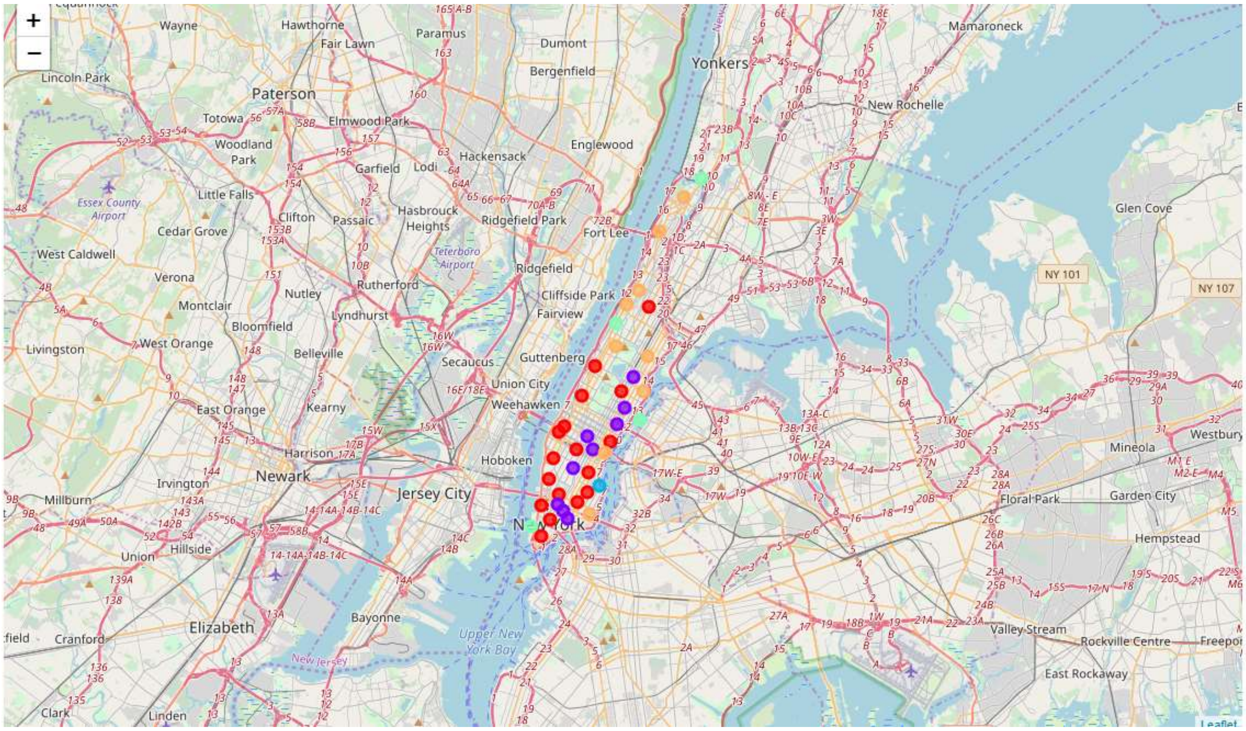
Clustering neighborhood using k-means where k=5



	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
0	Manhattan	Marble Hill	40.876551	-73.910660	3	Coffee Shop	Sandwich Place	Yoga Studio	Bank	Big Box Store
1	Manhattan	Chinatown	40.715618	-73.994279	1	Chinese Restaurant	Cocktail Bar	American Restaurant	Spa	Optical Shop
2	Manhattan	Washington Heights	40.851903	-73.936900	4	Café	Bakery	Grocery Store	Mobile Phone Shop	Chinese Restaurant
3	Manhattan	Inwood	40.867684	-73.921210	4	Lounge	Mexican Restaurant	Café	Restaurant	Bakery
4	Manhattan	Hamilton Heights	40.823604	-73.949688	4	Pizza Place	Coffee Shop	Mexican Restaurant	Café	Deli / Bodega

Visualized the cluster through k-means.

We break are analysis into five cluster and each cluster is color coded for the ease of presentation to understand their neighborhood, we can see that majority of the neighborhood represents different color for different cluster. These neighborhoods have their own cluster (Blue, Red, Purple and Yellow). this color scheme helps to name our cluster based on the venue and neighborhood.



Among five clusters below are the detail.

- Cluster 4 is the largest clusters which having 20 neighborhoods while Cluster 1 second largest having 17.
- Cluster 2 and 5 only 1 neighborhood
- Cluster 3 having 12 neighborhoods.

Discussion

Refer to my introduction part NYC is a big city with a high & diverse population density in a narrow area. The is too many neighborhoods available in respective borough which having high numbers of café, spa, restaurant which are of different type and nature. So, due to complexity of information matrix which is derived in results part very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high-quality results for this metropoltitian city.

We used the Kmeans algorithm as part of this clustering study; set the optimum k value to 5 is used. However, only xx borough coordinates were used and only one borough i.e. Marble Hill is drilled down. For more detailed and accurate guidance, the data set can be expanded, and the details of its neighbourhood is further explored to get more meaningful information which helps us to predict defined problem which is stated in summary part of this report.

We further explore this data to get most frequent visit area (venue) through the frequency it gets visited and which is most place (café, shop, restaurant etc.) is most visited in those area to predict accurately.

We ended the project by creating 5 clusters based on venue and their frequency through plotting them on NY map with choropleth style; this give more broader information about the actual borough and new cluster created through clustering algorithm according to frequency of most visited.

Conclusion

As due to global village and international migration from different part of worlds towards some of big metropolitan cities like Toronto, London, New York, Sidney etc. this migration turns them to diverse population and multicultural community. Through derived results, the investors will predict surely in which part of borough they should engage their capital and what the nature of business (SPA, restaurant, café…) more suitable to get maximum ROI.

In the derive results is quite visible if the Chines entrepreneur invest in borough 'China Town' and type of business he should choose is 'Chines Restaurant' in that area thus; its initial objective can be achieved with higher ROI.

References:

- {https://gist.github.com/doscsy12/5d347e43ae15548677b4adcdd73b87dd#file-house-sales_in_king_count_usa-ipynb}
- https://rpubs.com/Mani/report
- https://www.kaggle.com/madislemsalu/predicting-housing-prices-in-king-county-usa
- https://en.wikipedia.org/wiki/New_York_City
- https://en.wikipedia.org/wiki/Demographics_of_New_York_City