Finding Neighborhoods in New York City with more Health-care access using KMeans Clustering

Arjun M

May 2020

1 Introduction

New York City (NYC), often called The City or simply New York (NY), is the most populous city in the United States. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.
As New York provides a vast field of opportunities, there would be a huge amount of immigrants to NYC. Consider a scenario where a person with some sort health issue forced to move to New York City from some other part of U.S for job or studies. He would be confused to choose between which neighborhood he should choose to stay which will be nearest to Medical centres and Pharmacy, or in case of people who are health conscious and interested in Yoga. This can be a relevant challenge to anyone who is planning to move to NYC from other regions. The challenge is to find a suitable neighborhood in NYC that complies with the demands on access to Medical centres, Pharmacy and Yoga Studio.

2 Data Section

2.1 Data Collection

- Json Data of NYC with list of Boroughs,Neighborhoods of New York with their latitude and longitude is fetched from https://cocl.us/new_york_dataset
- Foursquare API : By using this api we will get all the venues in each neighborhood. We can filter these venues using it. Venues which fall in the category of Medical centres, Pharmacy and Yoga Studio.

2.2 Data Processing

- The json data consists of 'totalFeatures' : 306,which means in dataframe there would be total of 306 records. In each record there will be 'features' which includes id,coordinates,borough..

data = json.load(json_data)

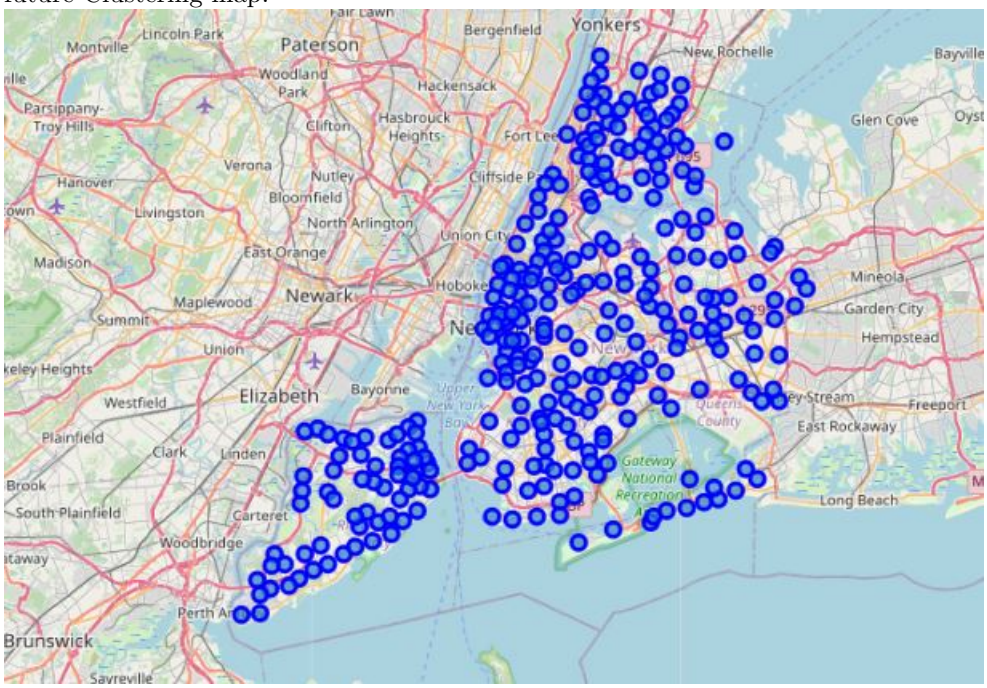

```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'type': 'Feature',
   'id': 'nyu_2451_34572.1',
   'geometry': {'type': 'Point',
    'coordinates': [-73.84720052054902, 40.89470517661]},
   'geometry_name': 'geom',
   'properties': {'name': 'Wakefield',
    'stacked': 1,
    'annoline1': 'Wakefield',
    'annoline2': None,
    'annoline3': None,
    'annoangle': 0.0,
    'borough': 'Bronx',
    'bbox': [-73.84720052054902,
     40.89470517661,
     -73.84720052054902,
     40.89470517661]}},
```


- The json data is normalized to a dataframe with only relevant variables

	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

3 Exploratory Data analysis

- 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. Using this library neighborhoods of the New York City is plotted which can be compared with the future Clustering map.



- We are dealing with 5 boroughs and 306 neighbors in the dataset using which the clustering have to be made

```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(neighborhoods['Borough'].unique()),
        neighborhoods.shape[0]
    )
)
```

```
The dataframe has 5 boroughs and 306 neighborhoods.
```

- Dataframe with every neighborhood which satisfies 'Venue Category' as Medical centres, Pharmacy and Yoga Studio.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
11	Co-op City	40.874294	-73.829939	Rite Aid	40.870345	-73.828302	Pharmacy
111	Kingsbridge	40.881687	-73.902818	Rite Aid	40.885481	-73.900814	Pharmacy
112	Kingsbridge	40.881687	-73.902818	Walgreens	40.878538	-73.904780	Pharmacy
...	...	...	...	...	...	...	...
9579	Prince's Bay	40.526264	-74.201526	CVS pharmacy	40.525814	-74.201656	Pharmacy
9652	Allerton	40.865788	-73.859319	Rite Aid	40.865949	-73.860922	Pharmacy
9698	Kingsbridge Heights	40.870392	-73.901523	Duane Reade	40.867540	-73.896984	Pharmacy
9715	Erasmus	40.646926	-73.948177	The Yoga Studio	40.650000	-73.950000	Yoga Studio
9721	Erasmus	40.646926	-73.948177	Rite Aid	40.650874	-73.950663	Pharmacy

247 rows × 7 columns

- Followed by grouping the dataframe using the name of Neighborhood and does one-hot encoding on the 'Venue Category' for future clustering

	Neighborhood	Medical Center	Pharmacy	Yoga Studio
0	Allerton	0.0	1.0	0.0
1	Annadale	0.0	1.0	0.0
2	Arden Heights	0.0	1.0	0.0
3	Auburndale	0.0	1.0	0.0
4	Bath Beach	0.0	1.0	0.0
...	...	...	...	...
148	Woodhaven	0.0	1.0	0.0
149	Woodlawn	0.0	1.0	0.0
150	Woodrow	0.0	1.0	0.0
151	Woodside	0.0	1.0	0.0
152	Yorkville	0.0	1.0	0.0

153 rows × 4 columns

- k-means clustering : Method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.
  In this case we can treat the n observations as the neighborhoods with 'Venue Category' as Medical centres, Pharmacy and Yoga Studio.
  kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(ny_grouped_clustering)
  where number of clusters is set as 5

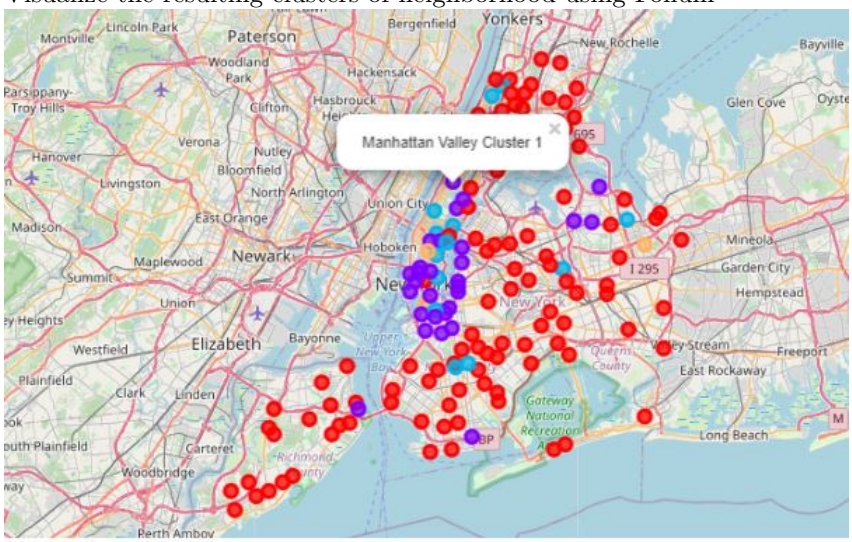
```
kmeans.labels_[0:50]
```

```
array([0, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 1, 0, 0, 4, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 2, 1, 0, 0, 0, 2, 0, 1,
       4, 0, 1, 0, 2, 0])
```

Cluster labels for the first 50 neighborhood is shown above

4 Results

- Visualize the resulting clusters of neighborhood using Folium



- Cluster 1 which is labelled as 0 shows the neighborhood with 1st most common value as Pharmacy, 2nd as Yoga Studio and 3rd as Medical Center

```
ny_merged.loc[ny_merged['Cluster Labels'] == 0, ny_merged.columns[[0] + list(range(3, ny_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Wakefield	0	Pharmacy	Yoga Studio	Medical Center
1	Co-op City	0	Pharmacy	Yoga Studio	Medical Center
5	Kingsbridge	0	Pharmacy	Yoga Studio	Medical Center
7	Woodlawn	0	Pharmacy	Yoga Studio	Medical Center
8	Norwood	0	Pharmacy	Yoga Studio	Medical Center
...	...	...	...	...	...
289	Homecrest	0	Pharmacy	Yoga Studio	Medical Center
290	Middle Village	0	Pharmacy	Yoga Studio	Medical Center
291	Prince's Bay	0	Pharmacy	Yoga Studio	Medical Center
298	Allerton	0	Pharmacy	Yoga Studio	Medical Center
299	Kingsbridge Heights	0	Pharmacy	Yoga Studio	Medical Center

109 rows × 5 columns

- Cluster 2 which is labelled as 1 shows the neighborhood with 1st most common value as Yoga Studio, 2nd as Pharmacy and 3rd as Medical Center

```
ny_merged.loc[ny_merged['Cluster Labels'] == 1, ny_merged.columns[[0] + list(range(3, ny_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
49	Greenpoint	1	Yoga Studio	Pharmacy	Medical Center
52	Sheepshead Bay	1	Yoga Studio	Pharmacy	Medical Center
59	Prospect Heights	1	Yoga Studio	Pharmacy	Medical Center
61	Williamsburg	1	Yoga Studio	Pharmacy	Medical Center
65	Cobble Hill	1	Yoga Studio	Pharmacy	Medical Center
68	Gowanus	1	Yoga Studio	Pharmacy	Medical Center
69	Fort Greene	1	Yoga Studio	Pharmacy	Medical Center
70	Park Slope	1	Yoga Studio	Pharmacy	Medical Center
84	Clinton Hill	1	Yoga Studio	Pharmacy	Medical Center
87	Boerum Hill	1	Yoga Studio	Pharmacy	Medical Center
96	North Side	1	Yoga Studio	Pharmacy	Medical Center
97	South Side	1	Yoga Studio	Pharmacy	Medical Center
103	Hamilton Heights	1	Yoga Studio	Pharmacy	Medical Center
107	Upper East Side	1	Yoga Studio	Pharmacy	Medical Center
115	Murray Hill	1	Yoga Studio	Pharmacy	Medical Center
117	Greenwich Village	1	Yoga Studio	Pharmacy	Medical Center
120	Tribeca	1	Yoga Studio	Pharmacy	Medical Center
122	Soho	1	Yoga Studio	Pharmacy	Medical Center
124	Manhattan Valley	1	Yoga Studio	Pharmacy	Medical Center
128	Financial District	1	Yoga Studio	Pharmacy	Medical Center

- Cluster 3 which is labelled as 2 shows the neighborhood with 1st most common value as Yoga Studio, 2nd as Pharmacy and 3rd as Medical Cen-

ter

```
ny_merged.loc[ny_merged['Cluster Labels'] == 2, ny_merged.columns[[0] + list(range(3, ny_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
6	Marble Hill	2	Yoga Studio	Pharmacy	Medical Center
86	Downtown	2	Yoga Studio	Pharmacy	Medical Center
102	Inwood	2	Yoga Studio	Pharmacy	Medical Center
112	Lincoln Square	2	Yoga Studio	Pharmacy	Medical Center
114	Midtown	2	Yoga Studio	Pharmacy	Medical Center
119	Lower East Side	2	Yoga Studio	Pharmacy	Medical Center
126	Gramercy	2	Yoga Studio	Pharmacy	Medical Center
135	Forest Hills	2	Yoga Studio	Pharmacy	Medical Center
151	Bayside	2	Yoga Studio	Pharmacy	Medical Center
221	Ditmas Park	2	Yoga Studio	Pharmacy	Medical Center
271	Sutton Place	2	Yoga Studio	Pharmacy	Medical Center
274	Tudor City	2	Yoga Studio	Pharmacy	Medical Center
300	Erasmus	2	Yoga Studio	Pharmacy	Medical Center

- Cluster 4 which is labelled as 3 shows the neighborhood with 1st most common value as Medical Center, 2nd as Yoga Studio and 3rd as Pharmacy

```
ny_merged.loc[ny_merged['Cluster Labels'] == 3, ny_merged.columns[[0] + list(range(3, ny_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
30	Parkchester	3	Medical Center	Yoga Studio	Pharmacy

- Cluster 5 which is labelled as 4 shows the neighborhood with 1st most common value as Yoga Studio, 2nd as Pharmacy and 3rd as Medical Center

```
ny_merged.loc[ny_merged['Cluster Labels'] == 4, ny_merged.columns[[0] + list(range(3, ny_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
64	Brooklyn Heights	4	Yoga Studio	Pharmacy	Medical Center
161	Oakland Gardens	4	Yoga Studio	Pharmacy	Medical Center
276	Flatiron	4	Yoga Studio	Pharmacy	Medical Center

From the results table it is clear that depending on the preferences for each Category value :Yoga Studio, Pharmacy, Medical Center the person can choose between suitable neighborhood to move.

5 Future directions

I was able to cluster the neighborhood based on our needed venue category which is related health care. In this study I have set radius=500 and limit=100. In the future work for fetching more venues one can increase these parameters or If one wants the venues in closer vicinity, can decrease the magnitude of parameters. In this study I have only included venue categories like Medical centers, Pharmacy and Yoga Studio. Those who are gym-enthusiasts can 'Gym' as a venue category and do the clustering. 'Spa' and 'Health and Beauty services' can also be added.

6 Conclusion

New York provides a vast field of opportunities, there would be a huge amount of immigrants to NYC. In this study I have focussed on how to choose between a range of neighborhoods to stay once someone have reached New York City, given the person wants best Health care, Pharmacy or he/she is health conscious and interested in Yoga Studio. Output of the clustering shows the neighborhood a person can choose to stay which satisifies his need.