Capstone Project - The Battle of Neighborhoods

Jawharah Almulhim

10 June -2020

1. Introduction:

1.1 Background

New York city also called NYC , it is one of the largest city on USA with huge number of population and diversity of people. The NYC is a main distance for visitors from all over the world . The idea of this project is to explore Manhattan which is often referred to by residents of the New York City area as the City, it is one of the most densely populated of the five boroughs of New York City. Exploring neighborhoods and venues of this borough is handled on this project.

1.2 Problem

The idea of this project suggest a Higley rated parks on Manhattan since mostly, visitors are willing to visit different places to enjoy themselves and parks are one of these places.

Also , ending up with clustering different parks of Manhattan into different clusters with similar features.

1.3 Audience

- 1-Vistors who are willing to visit highly rated parks
- 2-Governemnt agencies when they planning to open a new park ,so new location of park could be close to a highly rated one.

2. Data acquisition and cleaning

2.1 Data sources

To accomplish this project , different data sources are used:

- **New York City data** that contains list Boroughs, Neighborhoods along with their latitude and longitude. Data source : https://cocl.us/new_york_dataset Explanation: the above data set is available for free and it contains main data of NYC like latitude, longitude , boroughs and neighborhoods.
- **Foursquare API service:** using API calls to get neighborhoods and venues of the selected borough as well as detailed information about venues such as tips, likes, rating and more. Such information is necessary for clustering.
- **Pandas data frames** is used to store the results of the API calls and do the operations
- **Geopy** is client which is used to locate the coordinates of addresses using third-party geocoders
- **K-mean clustering :**machine learning tool to cluster the parks on different cluster based on similarities

2.2 Data cleaning

First, data downloaded from https://cocl.us/new_york_dataset in JSON format and the required features are fetched which are brought, neighborhood, latitude and longitude. Then, this data is stored in pandas dataframe.

	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
5	Bronx	Kingsbridge	40.881687	-73.902818
6	Manhattan	Marble Hill	40.876551	-73.910660
7	Bronx	Woodlawn	40.898273	-73.867315
8	Bronx	Norwood	40.877224	-73.879391

Figure 1:Dataframe of New York Data

Also, Multiple operations on dataframe are performed to filter rows to reach the target data which Manhattan's neighborhoods.

Later ,Foursquare API used to make RESTful API calls to retrieve data about venues of the selected borough , Venues retrieved from all the neighborhoods along with their category, latitude , longitude and name.

An extract of an API call is as follows:


```
{'meta': {'code': 200, 'requestId': '5ede334c949393001cde5537'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Marble Hill',
  'headerFullLocation': 'Marble Hill, New York',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 26,
  'suggestedBounds': {'ne': {'lat': 40.88105078329964,
    'lng': -73.90471933917806},
   'sw': {'lat': 40.87205077429964, 'lng': -73.91659997808156}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '4b4429abf964a52037f225e3',
       'name': "Arturo's",
       'location': {'address': '5198 Broadway',
        'crossStreet': 'at 225th St.',
        'lat': 40.87441177110231,
        'lng': -73.91027100981574,
        'labeledLatLngs': [{'label': 'display',
          'lat': 40.87441177110231,
          'lng': -73.91027100981574},
```


Figure 2 Response from Foursquare API call

Lastly, K-Mean clustering was applied on numerical data after it's being normalized , since data normalization helps to interpret features with different magnitudes and distributions equally.

	ID	Lat	Lan	Likes	Rating	Tips	Cluster
0	4a5a4eb2f964a52021ba1fe3	40.792027	-73.959853	109	9.0	6	3
1	4f3c0584e4b0f7c8c775c07e	40.789188	-73.957867	8	8.0	0	4
2	4c841c2ed8086dcb246f8652	40.787786	-73.955924	25	8.5	3	0
3	4b67aad0f964a520265a2be3	40.791591	-73.964795	33	8.6	2	0
4	4d6331414554a0934064afaa	40.788791	-73.955232	16	7.8	1	4

Figure 3 Original Data

```
array([[ 0.87946493,  0.54902537, -0.30578036,  0.7570733 , -0.31312928],
       [ 0.22156757,  0.9432769 , -0.32212336, -0.8758299 , -0.32470717],
       [-0.10341175,  1.32893295, -0.31937256, -0.0593783 , -0.31891823],
       [ 0.77845491, -0.43207911, -0.31807806,  0.10391202, -0.32084787],
       [ 0.12964712,  1.4663409 , -0.32082886, -1.20241054, -0.32277752],
       [-0.96133935, -0.44372359,  3.16217223,  1.90010554,  3.16216719],
       [-0.6523277 , -1.3700646 , -0.29429171,  1.24694426, -0.2938328 ],
       [-0.37045159, -1.15678125, -0.31937256, -0.38595894, -0.31119963],
       [ 2.31841576,  0.5399093 , -0.31969618,  0.10391202, -0.31312928],
       [-0.82579961, -1.41498389, -0.32099067,  0.10391202, -0.32470717],
       [-1.41422029, -0.00985296, -0.32163792, -1.6922815 , -0.31891823]])
```

Figure 4 Normalized Data