Capstone Project - The Battle of the Neighborhoods

The analysis of the best neighborhoods New York City using data science methodologies.

1 Contents

2 Introduction

2.1 Background

People have their own personal preferences of what they want around their house to live comfortably. When people are moving into a new neighborhood, it becomes difficult to find the best neighborhood that match their needs. Data analysis and machine learning helps solve this problem.

2.2 Problem

Customer A is planning to move to New York City. They have a personal preference of what needs to be close to their home for e.g. – Hospital, Restaurant, etc. They need help finding a neighborhood to move to in New York City with proximity to their needs and preferences.

The objective of this project is to use Machine learning algorithms and Foursquare location to determine the best neighborhood based on Customer A's needs and preferences in New York City.

3 Target Audience

- Anyone planning to move to a new neighborhood
- Real estate agents to help find a new place for their customers.
- This report is targeted to Customer A's preference. But this can be tailored to any person's needs.

4 Data

4.1 Data Sets

The datasets used for analysis for this project are:

4.1.1 New York City data

- Data Source: https://cocl.us/new_york_dataset
- Description: This data set contains Borough, Neighborhoods with latitudes and longitudes. This is used to explore different neighborhoods in New York city.

4.1.2 Foursquare API

- Data Source: https://api.foursquare.com
- Description: This API we will get all the venues in the New York city neighborhood. We can then analyses which neighborhood has the greatest number of venues that match Customer A's preference.

4.2 Data Preparation

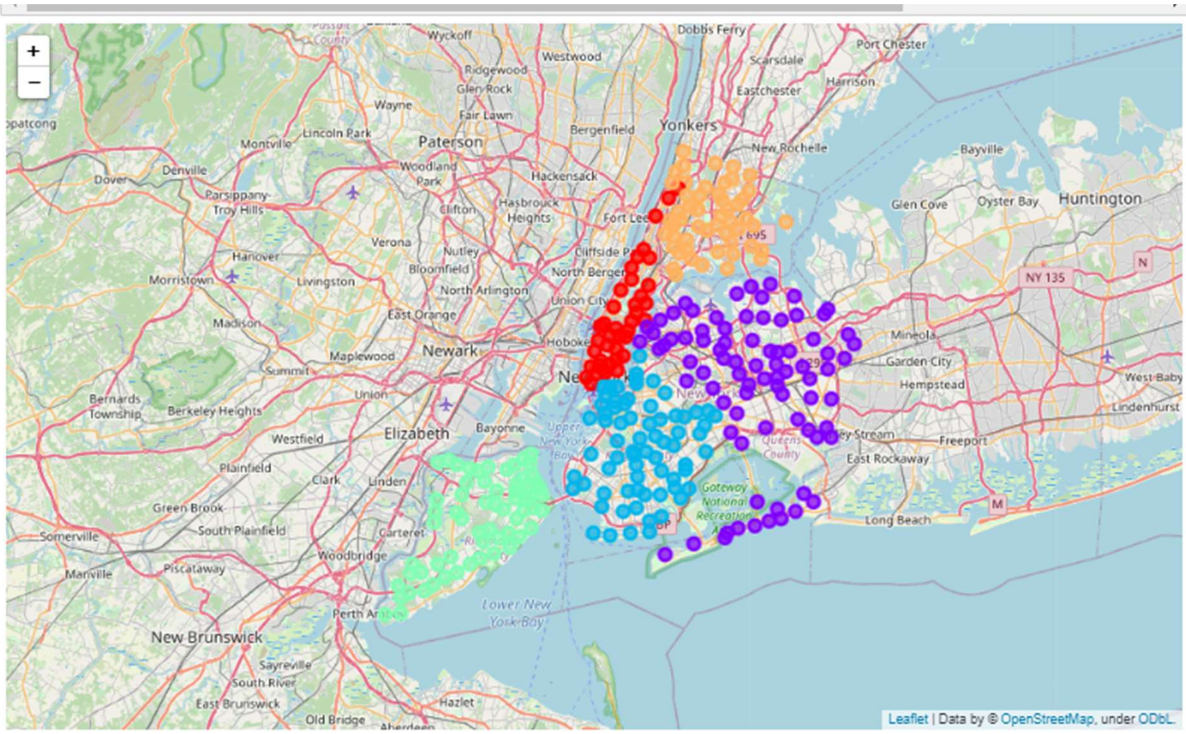
4.2.1 New York City data set

- The New York City data set that contains Borough, Neighborhoods with latitudes and longitudes.
- This is used to explore different neighborhoods in New York city.

	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

4.2.2 Maps

- Python folium library is used to visualize geographic details of New York City and its boroughs.
- The map of New York city is created with boroughs superimposed in different colors as shown:



4.2.3 Four Square API

- The Foursquare API is used to explore the boroughs and segment them.
- The API was set with a of **100 venue** and a radius of **500 meter** for each borough from their given latitude and longitude.
- Sample dataset of the list Venues name, category, latitude and longitude information from Foursquare API.

	Borough	Neighborhood	name	categories	lat	lng
0	Bronx	Wakefield	Lollipops Gelato	Dessert Shop	40.894123	-73.845892
1	Bronx	Wakefield	Rite Aid	Pharmacy	40.896649	-73.844846
2	Bronx	Wakefield	Walgreens	Pharmacy	40.896528	-73.844700
3	Bronx	Wakefield	Dunkin'	Donut Shop	40.890459	-73.849089
4	Bronx	Wakefield	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568

4.2.4 Favorite Categories

- A list of favorite categories is created - **Options**. For our project Customer A's preference is as shown.
- Each of the other categories are buckets to a main category; For e.g. "Doctor's Office", 'Pharmacy' is labeled as **Health**.

```
options = [
                "Doctor's Office", 'Pharmacy',
               'Grocery Store','Supermarket','Deli / Bodega'
               'Coffee Shop','Bakery', 'Donut Shop', 'Sandwich Place', 'Bagel Shop'
               'Italian Restaurant',  'Tex-Mex Restaurant',  'Chinese Restaurant', 'Pizza Place'
            ]
health= [  "Doctor's Office", 'Pharmacy']
grocery = ['Grocery Store','Supermarket','Deli / Bodega'  ]
cafe = ['Coffee Shop','Bakery', 'Donut Shop', 'Sandwich Place', 'Bagel Shop']
resturant = ['Italian Restaurant',  'Tex-Mex Restaurant',  'Chinese Restaurant', 'Pizza Place' ]
```

4.2.5 Merged Data

- The foursquare API data, the New York City data and Favorite category data is merged to a new data set.
- Each Main Category is also assigned a color(this will be used to create a map later)

	Borough	Neighborhood	name	categories	lat	lng	MainCategory	Color
0	Bronx	Wakefield	Lollipops Gelato	Dessert Shop	40.894123	-73.845892	Others	gray
1	Bronx	Wakefield	Rite Aid	Pharmacy	40.896649	-73.844846	Health	purple
2	Bronx	Wakefield	Walgreens	Pharmacy	40.896528	-73.844700	Health	purple
3	Bronx	Wakefield	Dunkin'	Donut Shop	40.890459	-73.849089	Cafe	red
4	Bronx	Wakefield	Carvel Ice Cream	Ice Cream Shop	40.890487	-73.848568	Others	gray

4.2.6 Filtered Merged Data:

- The Merged data is filtered for only those categories that is included in Customer A's preference.

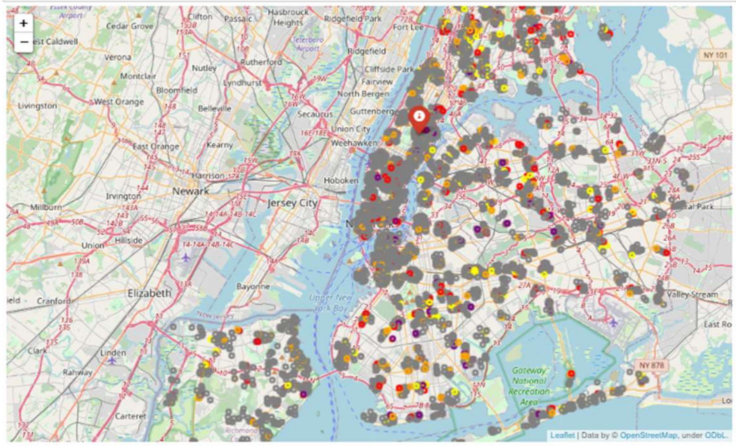
	Borough	Neighborhood	name	categories	lat	lng	MainCategory	Color
1	Bronx	Wakefield	Rite Aid	Pharmacy	40.896649	-73.844846	Health	purple
2	Bronx	Wakefield	Walgreens	Pharmacy	40.896528	-73.844700	Health	purple
3	Bronx	Wakefield	Dunkin'	Donut Shop	40.890459	-73.849089	Cafe	red
6	Bronx	Wakefield	Subway	Sandwich Place	40.890468	-73.849152	Cafe	red
10	Bronx	Co-op City	Rite Aid	Pharmacy	40.870345	-73.828302	Health	purple

5 Methodology

- This section represents the main component of the report. It starts with an exploratory data analysis before we dig deeper into solving the problem and applying machine learning algorithms.
- For the analysis, venues are filtered for only those categories that is included in Customer A's preference.
- One hot encoding is used to narrow the list of the most promising boroughs in the venue data frames.
- Normalized sum is used to determine top borough/neighborhood based on Customer A's preference.
- k-mean cluster analysis of all venues in New York City will provide us the most promising neighborhoods for Customer A.
- The results from Normalized sum and k-mean cluster should give us the best borough/neighborhood based on Customer A's preference.

5.1 Exploratory analysis

In the Exploratory analysis the distribution of venues in New York City was investigated. The result is shown in figure with the corresponding color code explained in table. It can be seen that there is a lot of promising neighborhoods where venues of Customer A's interest are located.



Catergory	Color
Health	purple
Grocery	yellow
Cafe	red
Resturant	orange
Others	gray

5.2 One Hot Encoding

- One hot encoding is used to narrow the list of the most promising boroughs in the venue data frames.
- It is used to count the number of favorite main categories in the data frame.

	Neighborhood	Cafe	Grocery	Health	Resturant
1	Wakefield	0	0	1	0
2	Wakefield	0	0	1	0
3	Wakefield	1	0	0	0
6	Wakefield	1	0	0	0
10	Co-op City	0	0	1	0
12	Co-op City	0	0	0	1
16	Co-op City	0	1	0	0
20	Co-op City	0	0	0	1
29	Eastchester	0	0	0	1
30	Eastchester	1	0	0	0
35	Eastchester	0	0	0	1
70	Kingsbridge	0	0	0	1

5.3 Normalized Score

- A Normalized score is calculated for each Neighborhood based on the count of favorite main categories in the Neighborhood
- Normalized Score/Sum = Count of categories in Neighborhood/Total count of categories

8]:

	index	Neighborhood	Cafe	Grocery	Health	Resturant	Sum	NormalizedSum	Borough	Latitude	Longitude
0	21	Belmont	10	2	1	9	22	0.025346	Bronx	40.857277	-73.888452
1	132	Kingsbridge	9	3	2	7	21	0.024194	Bronx	40.881687	-73.902818
2	263	Woodside	9	6	2	4	21	0.024194	Queens	40.746349	-73.901842
3	32	Bulls Head	4	3	2	9	18	0.020737	Staten Island	40.609592	-74.159409
4	138	Little Italy	9	1	1	7	18	0.020737	Manhattan	40.719324	-73.997305
5	43	Chinatown	7	2	1	8	18	0.020737	Manhattan	40.715618	-73.994279
6	232	Sunnyside Gardens	3	7	3	5	18	0.020737	Queens	40.745652	-73.918193
7	91	Fordham	8	2	3	5	18	0.020737	Bronx	40.860997	-73.896427
8	203	Rego Park	8	2	3	4	17	0.019585	Queens	40.728974	-73.857827
9	95	Fort Hamilton	6	1	2	7	16	0.018433	Brooklyn	40.614768	-74.031979
10	15	Bedford Park	4	3	2	6	15	0.017281	Bronx	40.870185	-73.885512
11	156	Midtown	9	1	1	4	15	0.017281	Manhattan	40.754691	-73.981669
12	126	Jackson Heights	5	5	2	2	14	0.016129	Queens	40.751981	-73.882821
13	119	Homecrest	6	3	1	4	14	0.016129	Brooklyn	40.598525	-73.959185
14	159	Mill Basin	3	2	2	7	14	0.016129	Brooklyn	40.615974	-73.915154
15	176	North Side	5	1	1	7	14	0.016129	Brooklyn	40.714823	-73.958809
16	82	Eltingville	4	2	1	6	13	0.014977	Staten Island	40.542231	-74.164331
17	153	Melrose	3	3	3	4	13	0.014977	Bronx	40.819754	-73.909422
18	0	Allerton	2	3	1	7	13	0.014977	Bronx	40.865788	-73.859319
19	11	Bay Ridge	2	3	2	6	13	0.014977	Brooklyn	40.625801	-74.030621

5.4 Top Boroughs/Neighborhoods based on Normalized Sum

- Top Boroughs are calculated based on the Normalized Score/Sum for the neighborhoods

	Borough	NormalizedSum
0	Bronx	0.279954
3	Queens	0.246544
1	Brooklyn	0.233871
2	Manhattan	0.148618
4	Staten Island	0.111751

	Borough	Neighborhood	NormalizedSum
2	Bronx	Belmont	0.025346
8	Bronx	Kingsbridge	0.024194
73	Queens	Woodside	0.024194
43	Manhattan	Chinatown	0.020737
49	Manhattan	Little Italy	0.020737

5.5 Clustering

- The k-means clustering is used to cluster the neighborhood into 8 clusters.
- In this analysis only favorite venues are considered -this was done by using one hot encoding, then calculated the mean of each venue in each neighborhood and finally grouped the data frame based on the neighborhood.
- The distribution of the clusters is shown in figure
- On further analysis the clusters 6 and 7 look most promising based on customer preference.
- The recommendation would be to move to Queens based on clustering and Normalized sum.

6 Results and Discussion

- For this report both normalized sum and clustering was performed.
- Considering normalized sum analysis, it was found that either Bronx or Queens could be a good choice to move.

	Borough	Normalized Sum
0	Bronx	0.279954
3	Queens	0.246544
1	Brooklyn	0.233871
2	Manhattan	0.148618
4	Staten Island	0.111751

	Borough	Neighborhood	Normalized Sum
2	Bronx	Belmont	0.025346
8	Bronx	Kingsbridge	0.024194
73	Queens	Woodside	0.024194
43	Manhattan	Chinatown	0.020737
49	Manhattan	Little Italy	0.020737

- The k-means provided an insight into similar neighborhoods and narrowed it down to cluster 6 and 7.

Neighborhoods in Queens for cluster 6 and 7

```
cluster6_neighborhoods = cluster6[cluster6['Borough'] == 'Queens']
cluster6_neighborhoods
```

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Bakery	Chinese Restaurant	Doctor's Office	Donut Shop	Grocery Store	Pharmacy	Pizza Place	Sandwich Place	Supermarket	Tex-Mex Restaurant
133	Queens	Howard Beach	40.654225	-73.838138	6	0.000000	0.125000	0.0	0.125000	0.000000	0.375000	0.000000	0.250000	0.125000	0.0
134	Queens	Corona	40.742382	-73.856825	6	0.142857	0.142857	0.0	0.142857	0.000000	0.000000	0.142857	0.142857	0.285714	0.0
135	Queens	Forest Hills	40.725264	-73.844475	6	0.125000	0.125000	0.0	0.125000	0.000000	0.250000	0.250000	0.000000	0.125000	0.0
136	Queens	Kew Gardens	40.705179	-73.829819	6	0.000000	0.272727	0.0	0.181818	0.000000	0.181818	0.181818	0.090909	0.090909	0.0
146	Queens	Woodhaven	40.689887	-73.858110	6	0.000000	0.000000	0.0	0.166667	0.000000	0.333333	0.166667	0.166667	0.166667	0.0
147	Queens	Ozone Park	40.680708	-73.843203	6	0.000000	0.000000	0.0	0.125000	0.125000	0.375000	0.250000	0.125000	0.000000	0.0
149	Queens	College Point	40.784903	-73.843045	6	0.181818	0.181818	0.0	0.090909	0.000000	0.090909	0.181818	0.181818	0.090909	0.0
152	Queens	Auburndale	40.761730	-73.791762	6	0.000000	0.000000	0.0	0.000000	0.000000	0.500000	0.000000	0.000000	0.500000	0.0
155	Queens	Glen Oaks	40.749441	-73.715481	6	0.000000	0.000000	0.0	0.142857	0.142857	0.428571	0.142857	0.142857	0.000000	0.0
161	Queens	Oakland Gardens	40.745619	-73.754950	6	0.000000	0.250000	0.0	0.250000	0.000000	0.125000	0.125000	0.125000	0.125000	0.0
166	Queens	Rochdale	40.675211	-73.772588	6	0.000000	0.250000	0.0	0.250000	0.000000	0.250000	0.000000	0.250000	0.000000	0.0
169	Queens	Rosedale	40.659816	-73.735261	6	0.000000	0.250000	0.0	0.000000	0.000000	0.250000	0.000000	0.250000	0.250000	0.0
175	Queens	Bay Terrace	40.782843	-73.776802	6	0.111111	0.000000	0.0	0.333333	0.000000	0.111111	0.111111	0.000000	0.333333	0.0
190	Queens	Belle Harbor	40.576156	-73.854018	6	0.250000	0.250000	0.0	0.250000	0.000000	0.250000	0.000000	0.000000	0.000000	0.0
191	Queens	Rockaway Park	40.580343	-73.841534	6	0.000000	0.142857	0.0	0.285714	0.000000	0.142857	0.285714	0.142857	0.000000	0.0
263	Queens	Jamaica Hills	40.711460	-73.796465	6	0.000000	0.142857	0.0	0.285714	0.142857	0.285714	0.000000	0.142857	0.000000	0.0
277	Queens	Sunnyside Gardens	40.745652	-73.918193	6	0.055556	0.055556	0.0	0.055556	0.222222	0.166667	0.222222	0.055556	0.166667	0.0

```
cluster7_neighborhoods = cluster7[cluster7['Borough'] == 'Queens']
cluster7_neighborhoods.head()
```

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Bakery	Chinese Restaurant	Doctor's Office	Donut Shop	Grocery Store	Pharmacy	Pizza Place	Sandwich Place	Supermarket	Tex-Mex Restaurant
129	Queens	Astoria	40.768509	-73.915654	7	0.300000	0.100000	0.0	0.000000	0.200000	0.000000	0.300000	0.100000	0.000000	0.0
130	Queens	Woodside	40.746349	-73.901842	7	0.190476	0.047619	0.0	0.142857	0.285714	0.095238	0.142857	0.095238	0.000000	0.0
131	Queens	Jackson Heights	40.751981	-73.882821	7	0.285714	0.000000	0.0	0.071429	0.214286	0.142857	0.142857	0.000000	0.142857	0.0
143	Queens	Ridgewood	40.708323	-73.901435	7	0.300000	0.000000	0.0	0.000000	0.200000	0.100000	0.300000	0.100000	0.000000	0.0
145	Queens	Rego Park	40.728974	-73.857827	7	0.294118	0.117647	0.0	0.058824	0.117647	0.176471	0.117647	0.117647	0.000000	0.0

- After combining these results, we identified one single borough, that is most likely the best choice: is **Queens**.
- Upon further investigation the neighborhood **Woodside** in Queens borough seems a good option based on both normalized sum and clustering

7 Conclusion

- The purpose of this analysis was to identify a borough/neighborhood based on all the categories in the customer (i.e. arks, coffee, bars, restaurants, grocery stores).
- For this report both normalized sum and clustering was performed.
- After combining these results, we identified one single borough, that is most likely the best choice: is **Queens – Woodside**.
- Upon further investigation the neighborhood **Woodside** in Queens borough seems a good option based on both normalized sum and clustering