CAPSTONE PROJECT

THE BATTLE OF THE NEIGHBORHOODS:

<u>SHIFTING</u>

Introduction-

This project is aimed at people who want to change their localities.

People need to shift from one area for various reasons. Job transfers and education are two important factors why people may want to change their locations.

But at the same time, changing localities means sacrificing the conveniences which you enjoy in your current location.

And everyone has their preferences when it comes to the ideal location. You may not care about having a fast food stall nearby, but you do care about having a pharmacy nearby.

Rather than searching for homes all over the city, it will be helpful to know which localities to target.

Through this project, I tried to do that.

It will give you an idea of which locality is most similar to your current or ideal location.

Data sources-

To keep the understanding of the project simple, I have only used data related to New York, assuming that a person wants to shift from one community board to another within New York. Later in the report, this assumption has been lifted.

The data used in the report is from the table in https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City

The latitudes and longitudes were searched from google and typed into the source code.

Foursquare API is used to collect information about each community boards.

Jupyter notebook has been used to write the source code.

Methodology-

First, the table was copied from the Wikipedia webpage using the pandas library.

	Community Board(CB)	Areakm2	Pop.Census2010	Pop./km2	Neighborhoods
0	Bronx CB 1	7.17	91497	12761	Melrose, Mott Haven, Port Morris
1	Bronx CB 2	5.54	52246	9792	Hunts Point, Longwood
2	Bronx CB 3	4.07	79762	19598	Claremont, Concourse Village, Crotona Park, Mo...
3	Bronx CB 4	5.28	146441	27735	Concourse, Highbridge
4	Bronx CB 5	3.55	128200	36145	Fordham, Morris Heights, Mount Hope, Universit...

Then, the coordinates of each community board was searched in Google and made into two lists, latitudes and longitudes. These lists were appended to the data frame which was created above.

	Community Board(CB)	Areakm2	Pop.Census2010	Pop./km2	Neighborhoods	Latitudes	Longitudes
0	Bronx CB 1	7.17	91497	12761	Melrose, Mott Haven, Port Morris	40.8197	-73.9132
1	Bronx CB 2	5.54	52246	9792	Hunts Point, Longwood	40.8211	-73.8921
2	Bronx CB 3	4.07	79762	19598	Claremont, Concourse Village, Crotona Park, Mo...	40.7270	-73.9909
3	Bronx CB 4	5.28	146441	27735	Concourse, Highbridge	40.8434	-73.9101
4	Bronx CB 5	3.55	128200	36145	Fordham, Morris Heights, Mount Hope, Universit...	40.8568	-73.9105

With the help of the coordinates and the Foursquare API, the list of venues of each locality was obtained.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx CB 1	40.8197	-73.9132	Landin Mac & Cheese	40.820028	-73.915886	Mac & Cheese Joint
1	Bronx CB 1	40.8197	-73.9132	Xochimilco Family Restaurant	40.818735	-73.916650	Mexican Restaurant
2	Bronx CB 1	40.8197	-73.9132	Blink Fitness	40.819348	-73.910373	Gym / Fitness Center
3	Bronx CB 1	40.8197	-73.9132	Senshi Okami Martial Arts Center	40.819295	-73.914158	Martial Arts Dojo
4	Bronx CB 1	40.8197	-73.9132	Bronx Documentary Center	40.818003	-73.918864	Art Gallery

The count of each kind of venue per locality was taken.
From that, a data frame was created, using the 10 most popular venues for each locality (community board in this case).

	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	Bronx CB 1	Mexican Restaurant	Pizza Place	Sandwich Place	Grocery Store	Donut Shop	Gym	Pharmacy	Fast Food Restaurant	Kids Store	Mobile Phone Shop
1	Bronx CB 10	Pizza Place	Bar	Pharmacy	Fast Food Restaurant	Donut Shop	Italian Restaurant	Bakery	Sandwich Place	American Restaurant	Diner
2	Bronx CB 11	Pizza Place	Italian Restaurant	Deli / Bodega	Sandwich Place	Coffee Shop	Donut Shop	Pharmacy	Bank	Supermarket	Burger Joint
3	Bronx CB 12	Caribbean Restaurant	Pharmacy	Bakery	Pizza Place	Fast Food Restaurant	Deli / Bodega	Donut Shop	Mobile Phone Shop	Supermarket	Gas Station
4	Bronx CB 2	Fast Food Restaurant	Pharmacy	Pizza Place	Grocery Store	Deli / Bodega	Donut Shop	Discount Store	Park	Sandwich Place	Mobile Phone Shop

Now, to understand the similarity, a scoring system is introduced.

The top 10 venues of the known locality are given values from 10 -> 1, ie, the most common venue will get a value 10, and the 10th most common venue will get a value 1.

The same thing will be done to the locality to which we are comparing the known locality.

If one of the ten venues of the known locality matches with a certain venue of the other locality, the values assigned to them are multiplied together.

All those multiplied results are added together, which gives us the total score.

For example,
Let's compare Bronx CB 1 to Bronx CB 12, where Bronx CB 1 is the known locality and Bronx CB 12 is the unknown locality.

By the first statement-

Mexican Restaurant=10 , Pizza Place=9 , Sandwich Place=8 , Grocery Store=7, Donut Shop=6, Gym=5 , Pharmacy=4, Fast Food Restaurant=3, Kids Store=2, Mobile Phone Shop=1

By the second statement-

Caribbean Restaurant=10, Pharmacy=9, Bakery=8, Pizza Place=7, Fast Food Restaurant=6, Deli / Bodega=5, Donut Shop=4, Mobile Phone Shop=3, Supermarket=2, Gas Station=1

By statement three we get-

10x0 + 9x7 + 8x0 + 7x0 + 6x4 + 5x0 + 4x9 + 3x6 + 2x0 + 1x3 = 144

The highest score is 385.

Based on this, a table of similarity score is obtained.

	locality	similarity
0	Bronx CB 4	385
1	Bronx CB 2	305
2	Bronx CB 5	271
3	Bronx CB 1	231
4	Bronx CB 10	224
5	Bronx CB 9	215
6	Bronx CB 6	213
7	Bronx CB 12	184
8	Brooklyn CB 6	178
9	Brooklyn CB 18	178
10	Queens CB 8	171
11	Bronx CB 8	169
12	Brooklyn CB 15	167
13	Brooklyn CB 9	167
14	Manhattan CB 12	160
15	Brooklyn CB 12	159
16	Brooklyn CB 2	151

3 cases have been studied in the above procedure and their histogram and maps are drawn.

- The histogram is used to understand how to analyze the data in the table.
- The map helps us to understand the proximity of the points to the area of interest (office, schools or university, etc), helping us make better decisions.

Results -

Case 1- The known locality is in New York

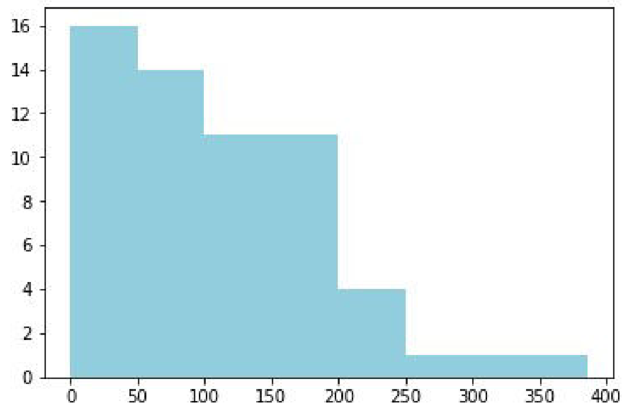
Place considered as known – **Bronx CB 4**

Table obtaned after finding the similarities-

	locality	similarity
0	Bronx CB 4	385
1	Bronx CB 2	305
2	Bronx CB 5	271
3	Bronx CB 1	231
4	Bronx CB 10	224
5	Bronx CB 9	215
6	Bronx CB 6	213
7	Bronx CB 12	184
8	Brooklyn CB 6	178
9	Brooklyn CB 18	178
10	Queens CB 8	171
11	Bronx CB 8	169
12	Brooklyn CB 15	167
13	Brooklyn CB 9	167
14	Manhattan CB 12	160
15	Brooklyn CB 12	159
16	Brooklyn CB 2	151
17	Queens CB 14	150
18	Brooklyn CB 17	144
19	Bronx CB 11	142
20	Bronx CB 7	141
21	Queens CB 2	140
22	Queens CB 12	133
23	Brooklyn CB 16	125
24	Brooklyn CB 4	116
25	Queens CB 9	114
26	Queens CB 4	106

	locality	similarity
27	Bronx CB 3	102
28	Manhattan CB 3	102
29	Queens CB 6	98
30	Queens CB 5	96
31	Brooklyn CB 14	96
32	Brooklyn CB 3	94
33	Brooklyn CB 7	82
34	Queens CB 10	80
35	Brooklyn CB 10	80
36	Brooklyn CB 13	72
37	Manhattan CB 2	71
38	Manhattan CB 11	66
39	Queens CB 11	66
40	Manhattan CB 5	64
41	Brooklyn CB 5	64
42	Manhattan CB 6	58
43	Brooklyn CB 8	48
44	Queens CB 1	48
45	Brooklyn CB 11	48
46	Staten Island CB 3	47
47	Staten Island CB 2	45
48	Staten Island CB 1	38
49	Manhattan CB 9	32
50	Manhattan CB 10	30
51	Queens CB 3	28
52	Manhattan CB 7	9
53	Manhattan CB 4	4
54	Manhattan CB 1	2
55	Queens CB 13	0
56	Manhattan CB 8	0
57	Queens CB 7	0
58	Brooklyn CB 1	0

Histogram of the scores in the above table-



Though the majority of the scores are low, but there is a decent number of localities with 100+ or 150+ scores. So selecting an appropriate locality should be easy.

We can see that the pop-up is displaying the score and the name of the community board.

Case 2- Finding the ideal location

Place considered as known – **Kolkata part 1**

I have used the name Kolkata because I have used the coordinates of Kolkata for the above location, but that is not significant for this specific case.

I manually typed in 10 types of venues which someone may consider to be the properties of his/her ideal locality.

They are-

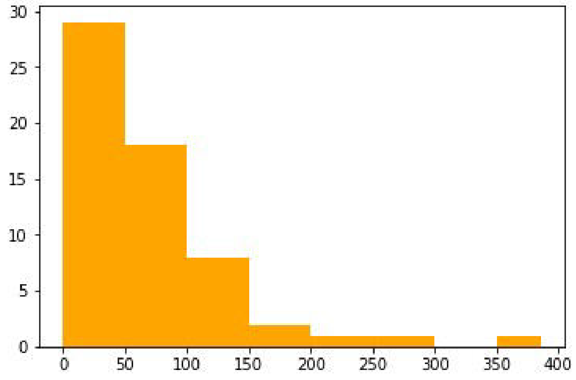
Kolkata part 1	Pharmacy	Fast Food Restaurant	Grocery Store	Indian Restaurant	Tea store	Bakery	Discount Store	Park	Sandwich Place	Mobile Phone Shop

The table of similarity obtained in this case-

	locality	similarity
0	Kolkata part 1	385
1	Bronx CB 2	266
2	Bronx CB 4	223
3	Bronx CB 12	187
4	Bronx CB 10	169
5	Bronx CB 1	140
6	Brooklyn CB 12	132
7	Brooklyn CB 18	123
8	Brooklyn CB 6	123
9	Brooklyn CB 9	116
10	Brooklyn CB 15	116
11	Queens CB 3	105
12	Brooklyn CB 17	104
13	Bronx CB 6	99
14	Bronx CB 9	97
15	Manhattan CB 12	96
16	Bronx CB 7	96
17	Queens CB 2	83
18	Bronx CB 5	80

	locality	similarity
19	Manhattan CB 11	76
20	Queens CB 11	76
21	Staten Island CB 1	75
22	Staten Island CB 2	72
23	Brooklyn CB 2	71
24	Brooklyn CB 7	66
25	Queens CB 12	61
26	Manhattan CB 7	60
27	Brooklyn CB 3	56
28	Manhattan CB 6	56
29	Bronx CB 11	54
30	Queens CB 4	52
31	Brooklyn CB 16	45
32	Queens CB 8	44
33	Manhattan CB 10	44
34	Bronx CB 8	40
35	Queens CB 1	40
36	Brooklyn CB 13	40
37	Brooklyn CB 8	35
38	Brooklyn CB 11	35
39	Queens CB 13	30
40	Brooklyn CB 10	30
41	Manhattan CB 9	30
42	Brooklyn CB 1	27
43	Bronx CB 3	24
44	Manhattan CB 3	24
45	Manhattan CB 1	24
46	Queens CB 9	21
47	Queens CB 7	20
48	Queens CB 14	16
49	Queens CB 6	15
50	Brooklyn CB 4	15
51	Brooklyn CB 14	15
52	Manhattan CB 2	14
53	Queens CB 5	11
54	Staten Island CB 3	9
55	Manhattan CB 4	8
56	Queens CB 10	0
57	Manhattan CB 8	0
58	Manhattan CB 5	0
59	Brooklyn CB 5	0

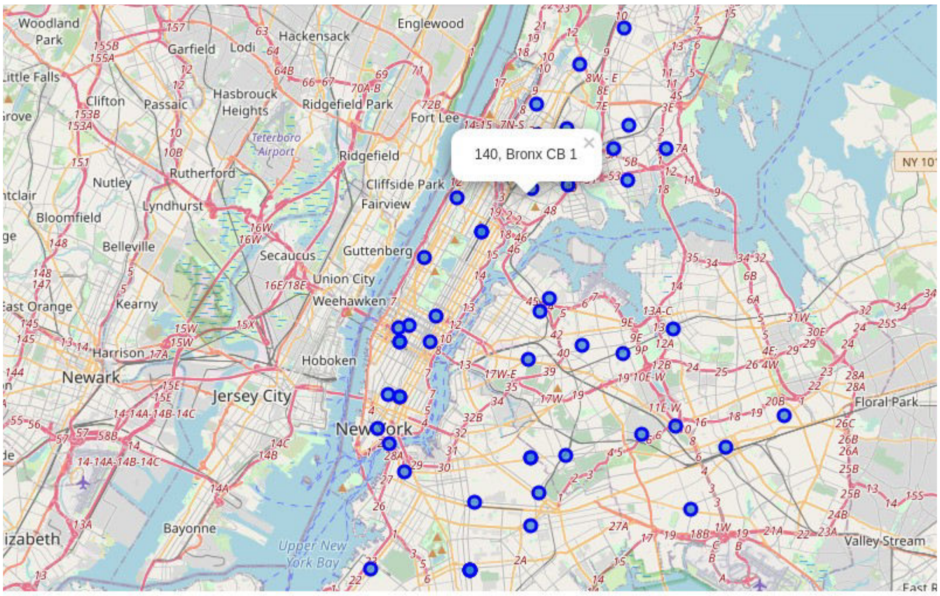
Histogram of the above score table-



We can very easily conclude that this person should primarily focus on Bronx or Brooklyn, as these regions are more compatible with his choice of venues.

But what if he wants to shift to Manhattan(none of the community boards have a 100+ score)?

The chart is still very effective as we see that there are only 4 community boards that don't have a score less than 50; namely 6,7,11 and 12. So it still holds as a very effective filter.



We can now see a second answer to the above question. Bronx CB 1 has a decent score of 140 and is located very near to the Manhattan borough. So it seems to be a very good choice.

Case 3-

Let’s assume that Kolkata part 1 is a real place in Kolkata.

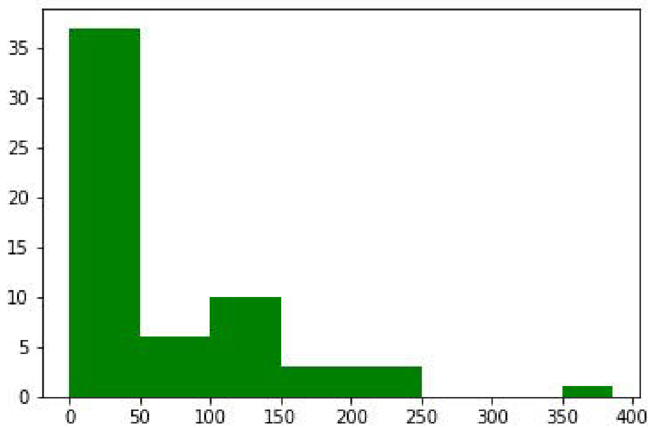
Place considered as known – **Manhattan CB 10**

Table-

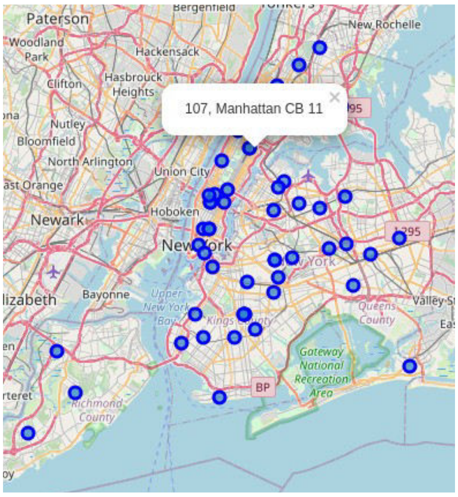
	locality	similarity
0	Manhattan CB 10	385
1	Manhattan CB 5	237
2	Brooklyn CB 5	237
3	Manhattan CB 4	211
4	Manhattan CB 8	179
5	Manhattan CB 1	177
6	Manhattan CB 6	156
7	Manhattan CB 2	138
8	Manhattan CB 3	123
9	Bronx CB 3	123
10	Manhattan CB 7	123
11	Brooklyn CB 3	121
12	Brooklyn CB 11	120
13	Brooklyn CB 8	120
14	Queens CB 11	107
15	Manhattan CB 11	107
16	Brooklyn CB 2	106
17	Staten Island CB 2	96
18	Brooklyn CB 7	86
19	Brooklyn CB 1	80
20	Manhattan CB 9	74
21	Queens CB 2	62
22	Queens CB 3	58
23	Bronx CB 11	48
24	Bronx CB 1	46
25	Kolkata part 1	44
26	Queens CB 1	42
27	Manhattan CB 12	41
28	Brooklyn CB 12	36
29	Brooklyn CB 4	32
30	Bronx CB 12	32
31	Bronx CB 4	30

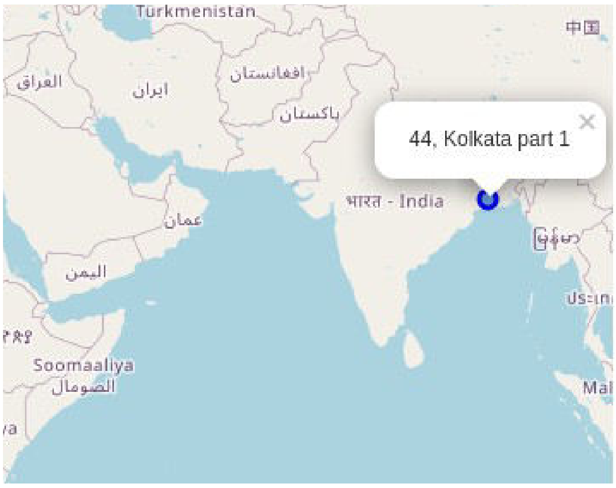
	locality	similarity
32	Bronx CB 5	30
33	Queens CB 13	24
34	Brooklyn CB 10	24
35	Bronx CB 2	21
36	Bronx CB 9	18
37	Bronx CB 10	16
38	Queens CB 7	16
39	Brooklyn CB 6	15
40	Brooklyn CB 18	15
41	Queens CB 14	13
42	Brooklyn CB 9	12
43	Brooklyn CB 14	12
44	Brooklyn CB 15	12
45	Queens CB 6	12
46	Queens CB 12	8
47	Queens CB 5	8
48	Brooklyn CB 16	4
49	Bronx CB 6	0
50	Queens CB 8	0
51	Staten Island CB 3	0
52	Staten Island CB 1	0
53	Queens CB 9	0
54	Queens CB 4	0
55	Bronx CB 7	0
56	Brooklyn CB 13	0
57	Bronx CB 8	0
58	Brooklyn CB 17	0
59	Queens CB 10	0

Histogram-



We can see that 35+ (37 exactly) places have a very low score, so it makes the filtering of data easy.





This case shows us how, with even a little information about a location, we can easily understand whether two localities are compatible, even if they are geographically miles apart.

Discussion-

Through the three cases, we can see why the model is so useful.

The first case was standard, in which the model worked as predicted.
The second case shows us that the map can give us vital information regarding a place, making our decision simpler.
The third case shows that even with a little information about a place, we can easily compare it with other places all around the world.

And the merits of the model doesn't end there.

Say you want to open a Chinese restaurant and want to make it a very important factor, ie, find a region where Chinese food is very popular. All you need to do is put 'Chinese restaurant' multiple times in your preference, making it a very important factor.

Suppose you have a list of 'must not' venues rather than most preferred.
You can still use the same process, just this time eliminate the places with a high score.

There are two scopes of improvement:

- The cost of living has not been considered as it is tricky to assign a weight to it. The concept of affordability is subjective, so it requires a much more complex model to consider that.
- The size of the community boards varies, but we consider the same are area (6.28 sq.km) every time, which may result in outliers. Though it gives a very good estimate of most community boards, it may not work on the larger ones.

Conclusion-

First of all, to avoid confusion, the scores obtained through this program are not indicative of how good or bad a locality is. It just shows how similar they are.

And the rank has no real significance. A locality with a score of 220 can be considered equal to a 280 scorer. Similarly, 0 and 30 are equally bad.

What it acts as is a very basic, yet effective filter.

And I think it will be helpful to a lot of people who have a very significant decision to take.