Predicting best location for buying home

Ankit Patel
October,2019

1. Introduction

1.1 Background

Consider a scenario, when you want to buy a home in New York City. What you will do? Most of us will reach the internet and property advisor in order to get the answer about property prices and options available to make their final decision. Is this really a wise decision? Most of us will say yes, but it's not.

1.2 Problem

Does the property advisor or the internet considering the liveability index parameters such as medical facilities, ease of travelling and shopping facilities around the suggested area? The answer surely is not, there naïve calculation is just based on the price factor and size of house needed without considering these important factors.

1.3 Interest

We are particularly interested in finding areas in New York city with good liveability index and pocket friendly property rates. Liveability of place considers following parameters

- Medical Facilities
- Shopping & Services
- Food and Restaurant
- Arts & Entertainment Zones
- Outdoor & Recreational activities

2. Data needed

Data needed for our decision making:

- number of Medical Centers in the neighborhood
- number of Arts & Entertainment places in the neighborhood
- number of Shopping stores in the neighborhood
- number of Outdoor & Recreational places in the neighborhood
- number of Food & Restaurants in the neighborhood
- number of Travelling options in the neighborhood
- property price in the neighborhood
- different neighborhood in New York City

We will need geo-coordinates of each neighbor in order to find various liveability index parameters (using foursquare api) described in section 1.3.

Finally, we can know the liveability of that area using the above data i.e. whether the number of Medical Facilities, Shopping & Services, Outdoor & Recreational places and Travelling options are high or not in that area and whether the property price is low or not. Using the above data, we can also compare various neighbors of chosen city and choose the best pocket friendly highly livable area.

Thus, we can achieve our task of predicting best location for buying home.

3. Methodology

3.1 Identifying Neighborhood and property prices

Using this web page https://www.zumper.com/blog/2019/05/nyc-by-square-foot-see-which-neighborhood-gets-you-the-most-space-for-your-money/ we can get different areas of New York city and their property prices. We will use BeautifulSoup to scrap the different neighbors and their respective property prices from the above web page and store it in pandas dataframe. Then we will use HERE maps api to Geocode the neighbors address in order to get Geo-Coordinates of each location.

	Neighborhood	Price_per_sq_foot	Latitude	Longitude
0	West Village	7.68	40.734980	-74.004830
1	Tribeca	7.64	40.718460	-74.008890
2	NoMad	7.63	40.744688	-73.988285
3	Central Park	7.53	40.783920	-73.965840
4	NoHo	7.38	40.729820	-73.991220

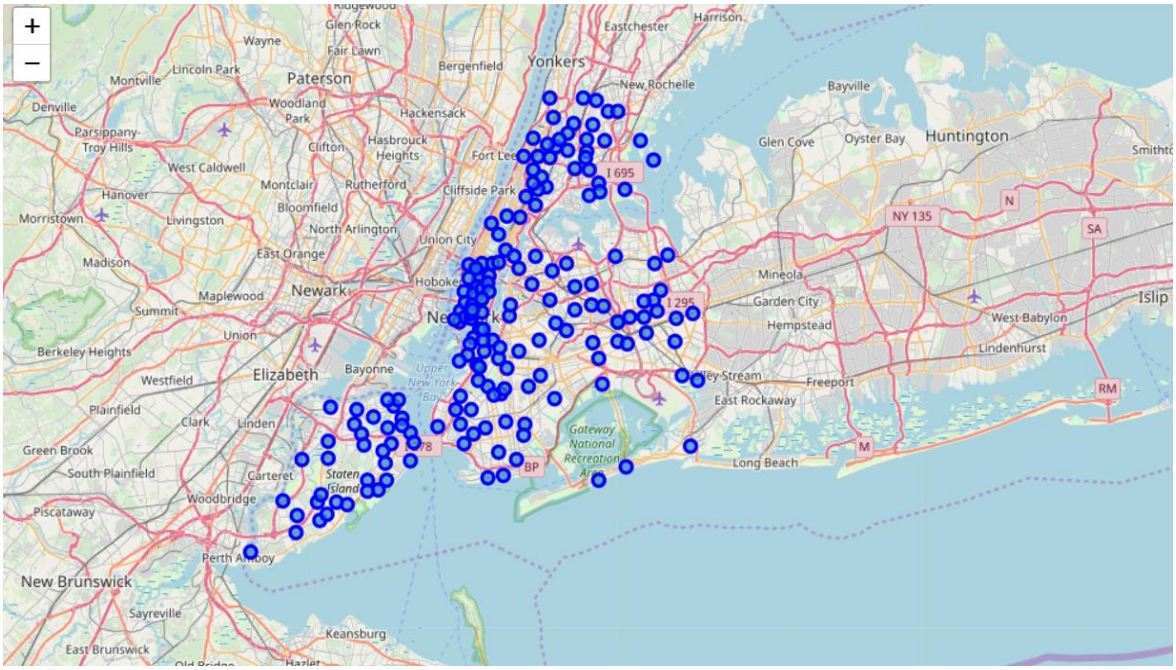


Fig 1: Marker points on different areas of New York City

3.2 Getting liveability index parameters for each neighborhood

After finding the list of neighborhood, their Geo-Coordinates and property prices, we then connect to the Foursquare API to gather information about number of venues inside each neighbor. We are only interested in following categories:

Category	Category id
Arts_Entertainment	4d4b7104d754a06370d81259
Food	4d4b7105d754a06374d81259
Outdoors & Recreation	4d4b7105d754a06377d81259
Medical_Center	4bf58dd8d48988d104941735
Shop_Service	4d4b7105d754a06378d81259
Travel_Transport	4d4b7105d754a06379d81259

The Category id's of different categories are obtained from following web page:
https://developer.foursquare.com/docs/resources/categories

3.3 Statistical view of our data

Clearly from below table we can find that values of column Price_per_sq_foot ranges from 0.71 to 7.68 and that of another lie between 0 to 100. Therefore, there is a need to scale these features before clustering.

	Price_per_sq_foot	Arts_Entertainment	Food	Outdoors & Recreation	Medical_Center	Shop_Service	Travel_Transport
count	185.000000	185.000000	185.000000	185.000000	185.000000	185.000000	185.000000
mean	3.048162	12.205405	36.335135	21.848649	29.637838	50.459459	19.118919
std	1.657277	21.861247	32.228567	30.057532	34.749471	35.948026	26.784708
min	0.710000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	1.000000	11.000000	4.000000	5.000000	19.000000	4.000000
50%	2.450000	3.000000	25.000000	7.000000	13.000000	40.000000	7.000000
75%	3.830000	8.000000	51.000000	23.000000	43.000000	100.000000	18.000000
max	7.680000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000

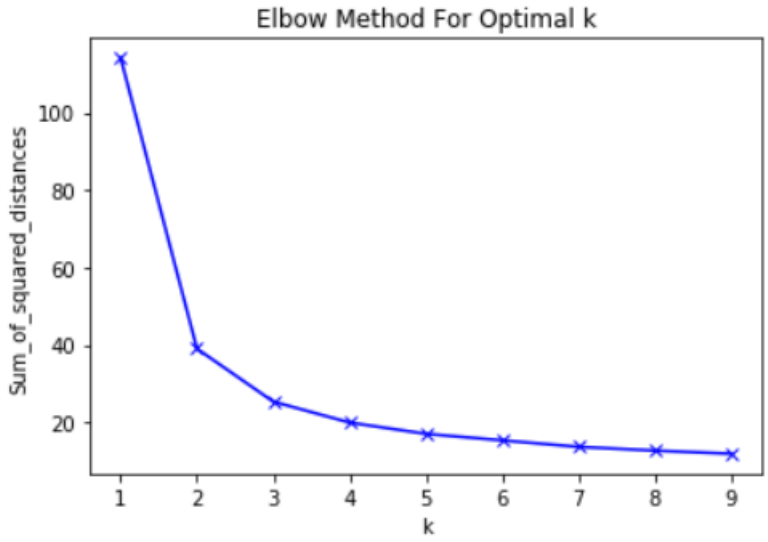
3.4 Applying one of Machine Learning Techniques (K-Means Clustering)

A. We have selected following features for clustering purpose and scaled them:

- Price_per_sq_foot
- Arts_Entertainment
- Food
- Outdoors & Recreation
- Medical_Center
- Shop_Service
- Travel_Transport

B. Finding Optimal k

Using the elbow method optimal value of k is found to be 3



C. K-Means Clustering:

Now with k=3 we can use k means clustering to agglomerate data based on above selected transformed features.

Neighborhood	Latitude	Longitude	Cluster	Price_per_sq_foot	Arts_Entertainment	Food	Outdoors & Recreation	Medical_Center	Shop_Service	Travel_Transport
West Village	40.734980	-74.004830	1	7.68	57	100	67	66	100	39
Tribeca	40.718460	-74.008890	1	7.64	55	100	100	100	100	60
NoMad	40.744688	-73.988285	1	7.63	68	100	100	100	100	100
Central Park	40.783920	-73.965840	0	7.53	23	4	36	43	9	3
NoHo	40.729820	-73.991220	1	7.38	85	100	100	93	100	69

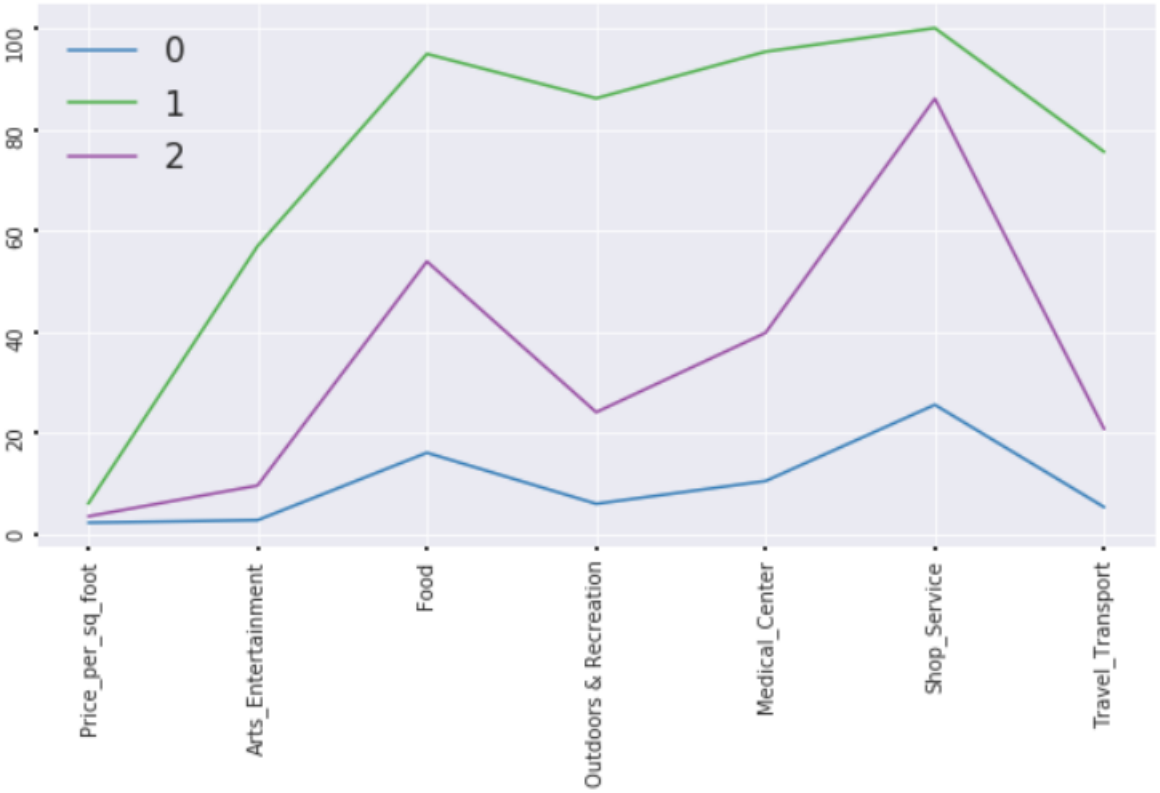
4. Results

4.1 Quick Analysis

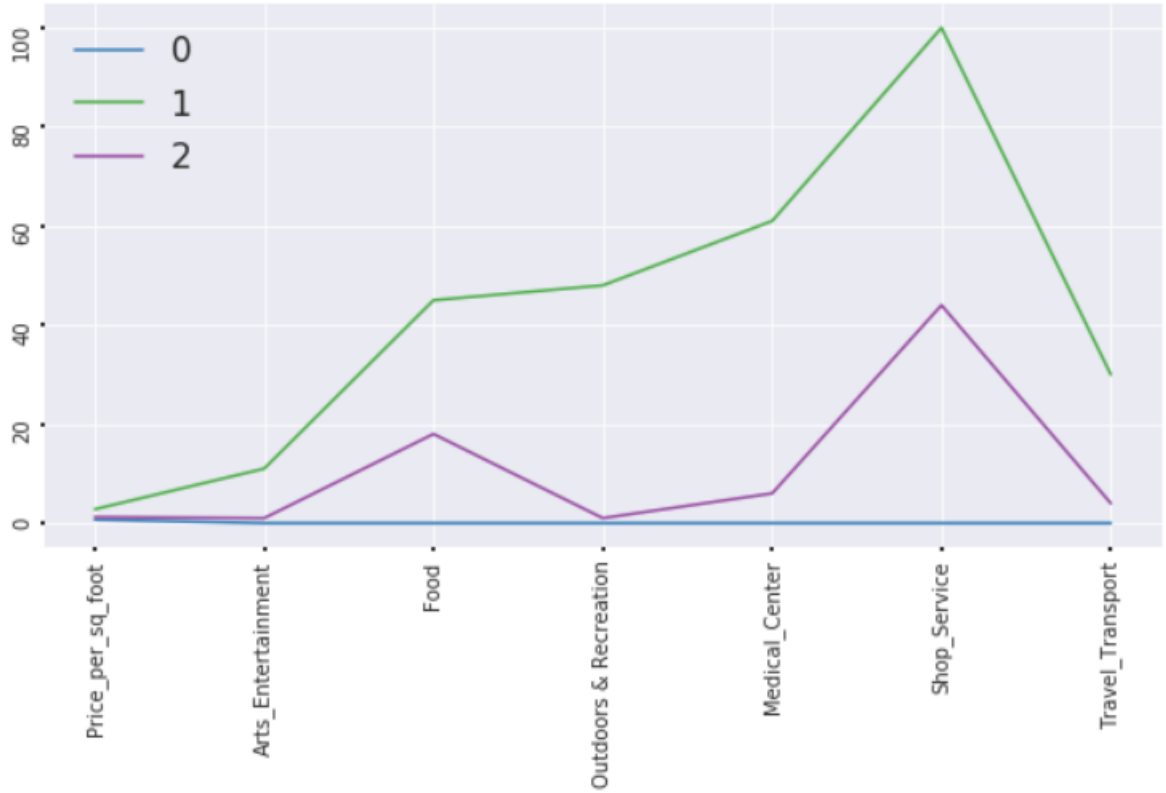
From below two graph we can see liveability of area in each cluster as follows:

- For Cluster 1 liveability is high i.e. number of restaurants, hospitals, travelling, shopping and entertainment options are great but the starting property price in this region is high.
- For Cluster 2 liveability is medium i.e. compromise in either number of restaurants, hospitals, travelling, shopping and entertainment options or less property price, but budget friendly options might also be present in this cluster.
- For Cluster 0 liveability is low i.e. number of restaurants, hospitals, travelling, shopping and entertainment options are less.

Mean value of different parameters in each cluster



Minimum value of different parameters in each cluster



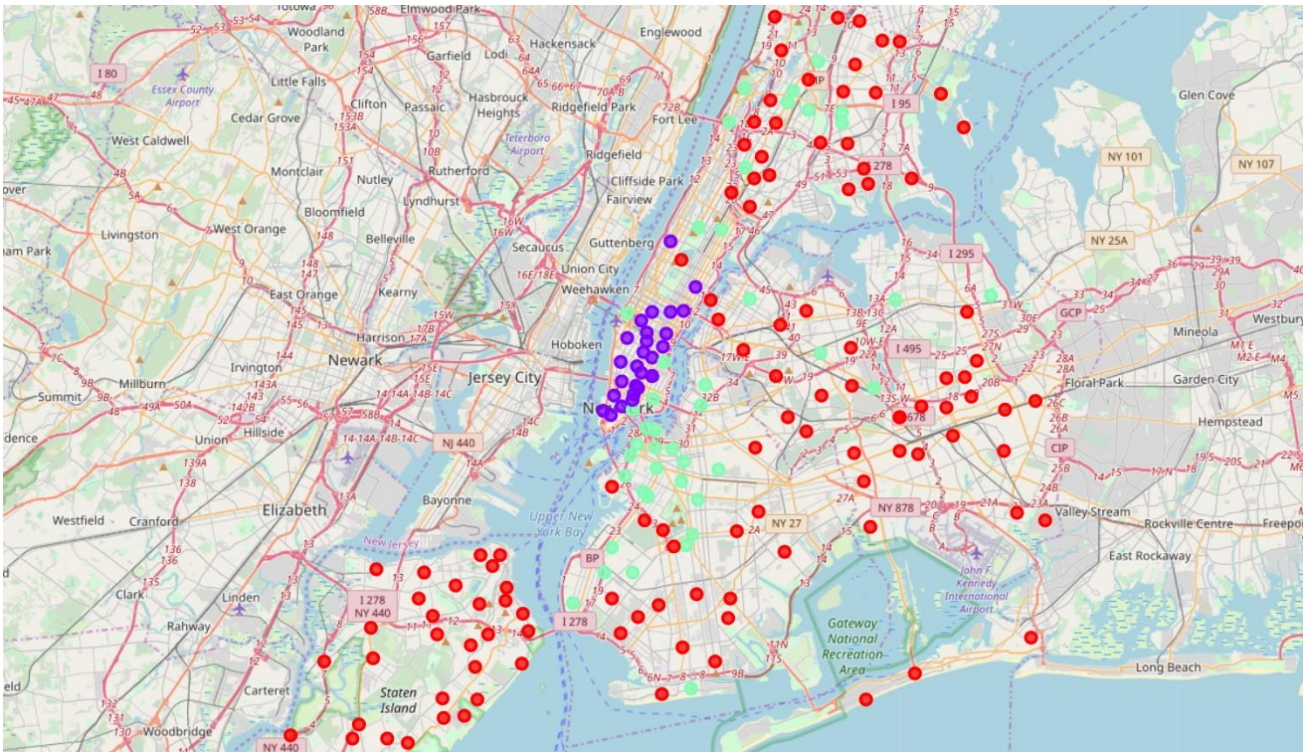


Fig 2: Cluster Marker on New York City Map

Cluster	Color
Cluster 0	Red
Cluster 1	Violet
Cluster 2	Green

Table: Marker color and their corresponding clusters

4.2 Deep Analysis

A. Cluster 1

The property price in this cluster ranges from 2.80 to 7.68 dollar per sq foot which is quite high because the liveability is very high in these places.

If your budget is above 2.80 dollar, some of the best places to buy house in this cluster

- **East New York** with lowest price of 2.80 dollar per sq foot
- **Little Italy** with price of 4.90 dollar per sq foot
- **Turtle Bay** with price of 4.99 dollar per sq foot and so, on

	Neighborhood	Latitude	Longitude	Cluster	Price_per_sq_foot	Arts_Entertainment	Food	Outdoors & Recreation	Medical_Center	Shop_Service	Travel_Transport
68	East New York	40.728040	-73.984990	1	2.80	64	100	61	89	100	62
36	Little Italy	40.718920	-73.996090	1	4.90	70	100	87	100	100	88
34	Turtle Bay	40.759250	-73.965030	1	4.99	27	91	100	100	100	73
33	Chinatown	40.716470	-73.996760	1	5.00	57	100	61	100	100	96
32	Upper East Side	40.770440	-73.957170	1	5.03	11	100	49	100	100	47
30	Midtown East	40.758800	-73.972910	1	5.13	58	100	100	100	100	100
26	Battery Park City	40.711320	-74.015900	1	5.20	23	45	86	95	100	58
23	Civic Center	40.713370	-74.003810	1	5.35	47	76	100	100	100	71
22	Kips Bay	40.742130	-73.977820	1	5.40	16	84	58	100	100	67
20	Garment District	40.754400	-73.991850	1	5.43	100	100	100	100	100	100
19	Upper West Side	40.792510	-73.973210	1	5.53	17	97	48	91	100	30
18	Theater District	40.758890	-73.984820	1	5.59	100	100	100	100	100	100
17	Gramercy Park	40.736830	-73.984600	1	5.59	41	100	92	100	100	69
16	Financial District	40.709020	-74.010610	1	5.66	33	100	100	100	100	100
15	East Village	40.728040	-73.984990	1	5.71	64	100	61	89	100	62
13	Murray Hill	40.748550	-73.976050	1	5.79	17	100	100	100	100	100
12	NoLita	40.723210	-73.995310	1	5.88	82	100	100	98	100	75
10	Soho	40.725200	-74.004150	1	6.74	70	100	85	61	100	64
9	Chelsea	40.746110	-74.000450	1	6.77	73	71	70	92	100	48
8	Greenwich Village	40.732450	-73.994060	1	6.96	76	98	100	100	100	73
7	Bowery	40.722210	-73.993300	1	7.09	97	100	100	100	100	90
6	Flatiron District	40.739420	-73.990350	1	7.10	63	100	100	100	100	98
5	Koreatown	40.748660	-73.988100	1	7.16	65	100	100	100	100	100
4	NoHo	40.729820	-73.991220	1	7.38	85	100	100	93	100	69
2	NoMad	40.744688	-73.988285	1	7.63	68	100	100	100	100	100
1	Tribeca	40.718460	-74.008890	1	7.64	55	100	100	100	100	60
0	West Village	40.734980	-74.004830	1	7.68	57	100	67	66	100	39

B. Cluster 2

The property prices in this cluster ranges from 1.22 to 6.37 dollar per sq foot.

If you budget ranges from 1.22 to 2.80 dollar the best places in this cluster with medium liveability are

- **Concourse** with a price of 1.22
- **Fordham Manor** with a price of 1.80
- **Fordham heights** with a price of 2.28 and so on

	Neighborhood	Latitude	Longitude	Cluster	Price_per_sq_foot	Arts_Entertainment	Food	Outdoors & Recreation	Medical_Center	Shop_Service	Travel_Transport
176	Concourse	40.82764	-73.92534	2	1.22	7	67	28	12	89	10
153	Fordham Manor	40.86476	-73.89536	2	1.80	9	35	11	29	100	23
106	Fordham Heights	40.85894	-73.89886	2	2.28	7	45	8	36	100	18
93	Forest Hills	40.72232	-73.84460	2	2.44	11	48	30	100	100	14

C. Cluster 0

The property prices in this cluster ranges from 0.71 to 7.53 dollar per sq foot.

The lowest property price is in Highbridge, Oakwood, Prince's Bay, Clifton and Woodrock below 1 dollar per sq foot

The best places in this cluster with fair liveability if you budget is below 1.22 dollar are

- **Woodrow** with a price of 0.91
- **Auburndale** with a price of 1.03
- **Morris Heights** with a price of 1.12

	Neighborhood	Latitude	Longitude	Cluster	Price_per_sq_foot	Arts_Entertainment	Food	Outdoors & Recreation	Medical_Center	Shop_Service	Travel_Transport
180	Woodrow	40.54316	-74.19761	0	0.91	2	3	3	7	6	2
178	Auburndale	40.75861	-73.78574	0	1.03	3	24	5	65	32	7
177	Morris Heights	40.84971	-73.91985	0	1.12	3	8	7	8	33	3

5. Discussion

The aim of this project is to find the best places where we can buy home. We found out that

- For Cluster 1 liveability is high i.e. number of restaurants, hospitals, travelling, shopping and entertainment options are great but the starting property price in this region is high.
- For Cluster 2 liveability is medium i.e. compromise in either number of restaurants, hospitals, travelling, shopping and entertainment options or less property price, but budget friendly options might also be present in this cluster.
- For Cluster 0 liveability is low i.e. number of restaurants, hospitals, travelling, shopping and entertainment options are less.

6. Conclusion

This project helps a person in buying home by considering most important liveability parameters apart from property price and suggesting them better pocket friendly alternatives.

Thank You!!!