Clustering Safest Neighborhoods in New York

By: Harsh Vardhan Jaiswal

Introduction

- According to US Census Bureau an average person moves more than 11 times in their lifetime.
- People who are looking for settling in a new place need to do a lot of research to find the best place.
- Major concern for people is safety in a neighborhood.
- In this problem we will finding the safest neighborhoods in New York City.
- And then clustering those neighborhoods as per their similarity.
- Target audience will be anyone who is looking to settle in New York.

Data Acquisition

- The crime data for New York is available at data.cityofnewyork.us. We will using the latest data i.e. from Jan-Mar 2020.
- For neighborhood data of New York geo JSON file will be used i.e. available here.
- Once both the datasets are cleaned and processed they are merged into a single dataset.

Data Cleaning

Following steps are taken for cleaning the data:

Crime Dataset

- Remove all the data that is not in the first quarter of the year 2020 or has missing information.
- Add Neighborhood column in the dataset using GeoPy API.
- Drop the irrelevant columns from the dataset.
- Prepare a pivot table based on the type of crime. In U.S. the crime is divided into 3 levels i.e. Felony, Misdemeanor and Violation

Neighborhood Dataset

- Neighborhood Data prepared from the geo json file of the New York City. In order to use for analysis we need to convert it into a data frame.

After this we merge the above Crime dataset with the Neighborhood data we get the final dataset.

Methodology

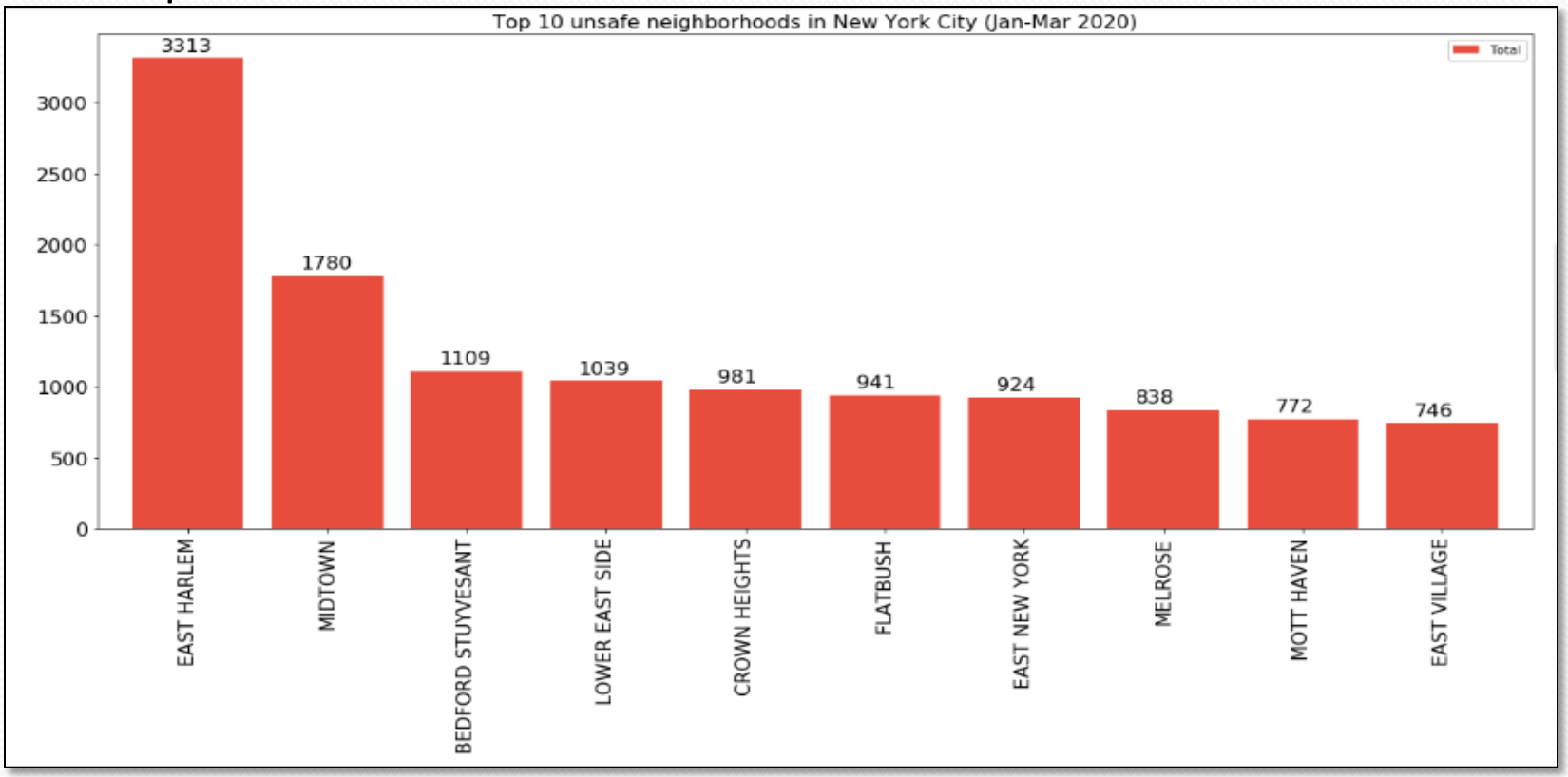
Exploratory Data Analysis

Descriptive statistics for the crime data us the mean, standard deviation, minimum, 25th Percentile, Median (50th Percentile), 75th Percentile and maximum values associated with no. of Felonies, Misdemeanor, Violations and Total no. of crimes reported in the first quarter of 2020 in the city of New York.

	FELONY	MISDEMEANOR	VIOLATION	Total	Latitude	Longitude
count	231.000000	231.000000	231.000000	231.000000	231.000000	231.000000
mean	66.857143	114.844156	35.748918	217.450216	40.704096	-73.952877
std	98.386865	171.142813	52.060226	315.964536	0.097375	0.121677
min	0.000000	1.000000	0.000000	1.000000	40.505334	-74.246569
25%	8.000000	16.500000	6.500000	33.500000	40.622638	-74.010491
50%	33.000000	67.000000	20.000000	130.000000	40.705179	-73.944182
75%	85.500000	155.500000	44.000000	286.500000	40.769667	-73.863736
max	901.000000	1855.000000	557.000000	3313.000000	40.898273	-73.715481

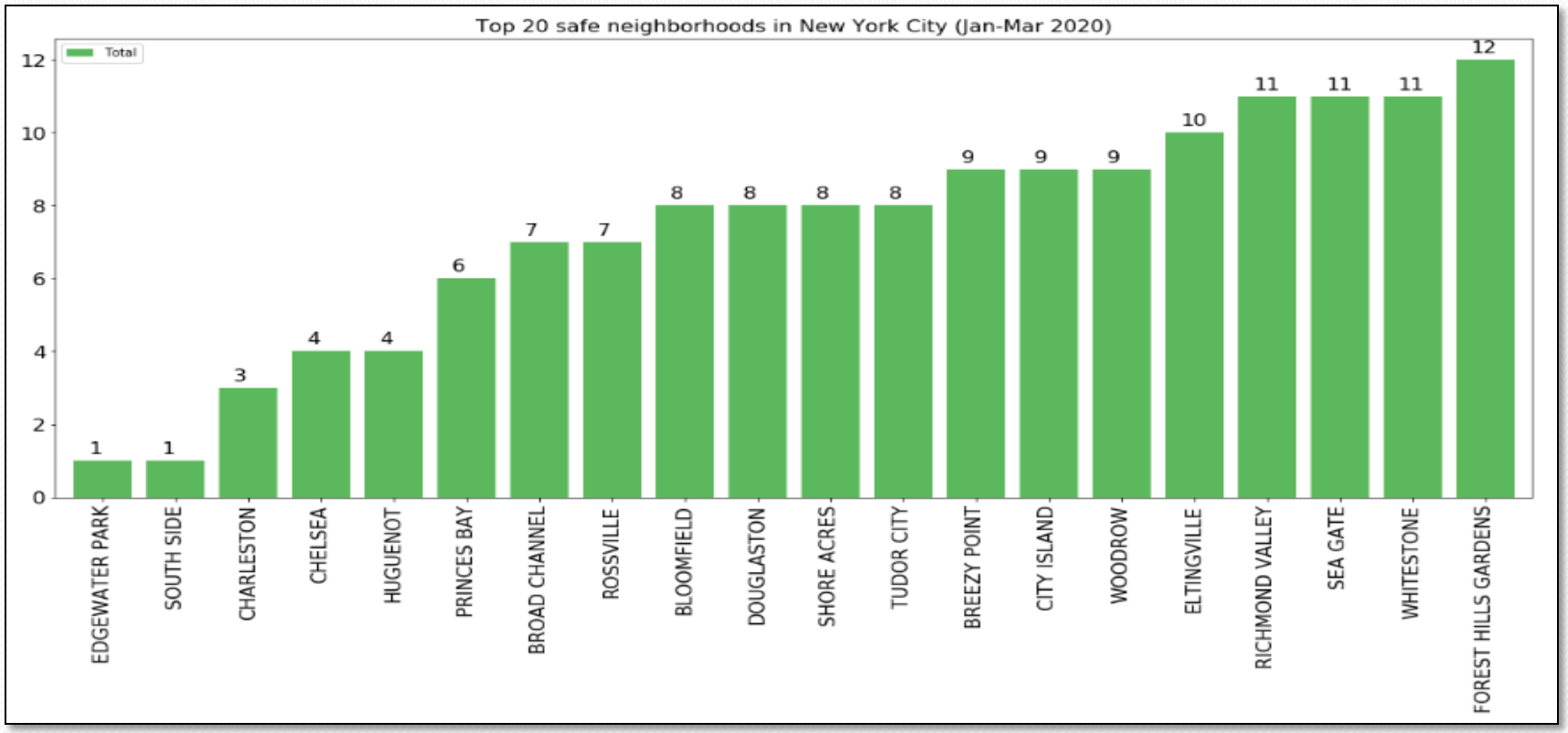
Neighborhoods with highest crime rates

For the analysis 10 neighborhoods with highest crime rate in New York City are taken as per the first quarter of 2020. Following are the bar-chart and table to provide more information in detail.



Neighborhoods with lowest crime rates

For the analysis 20 neighborhoods with lowest crime rate in New York City are taken as per the first quarter of 2020. Following are the bar-chart and table to provide more information in detail.



Modeling

- A separate data frame with only those 20 safest neighborhoods was created for further analysis.
- Then, find all the venues nearby each neighborhood using the Four-Square API.
- The below data frame is in length format. We need to do hot-encoding based on the venue category provided above for all the neighborhoods.

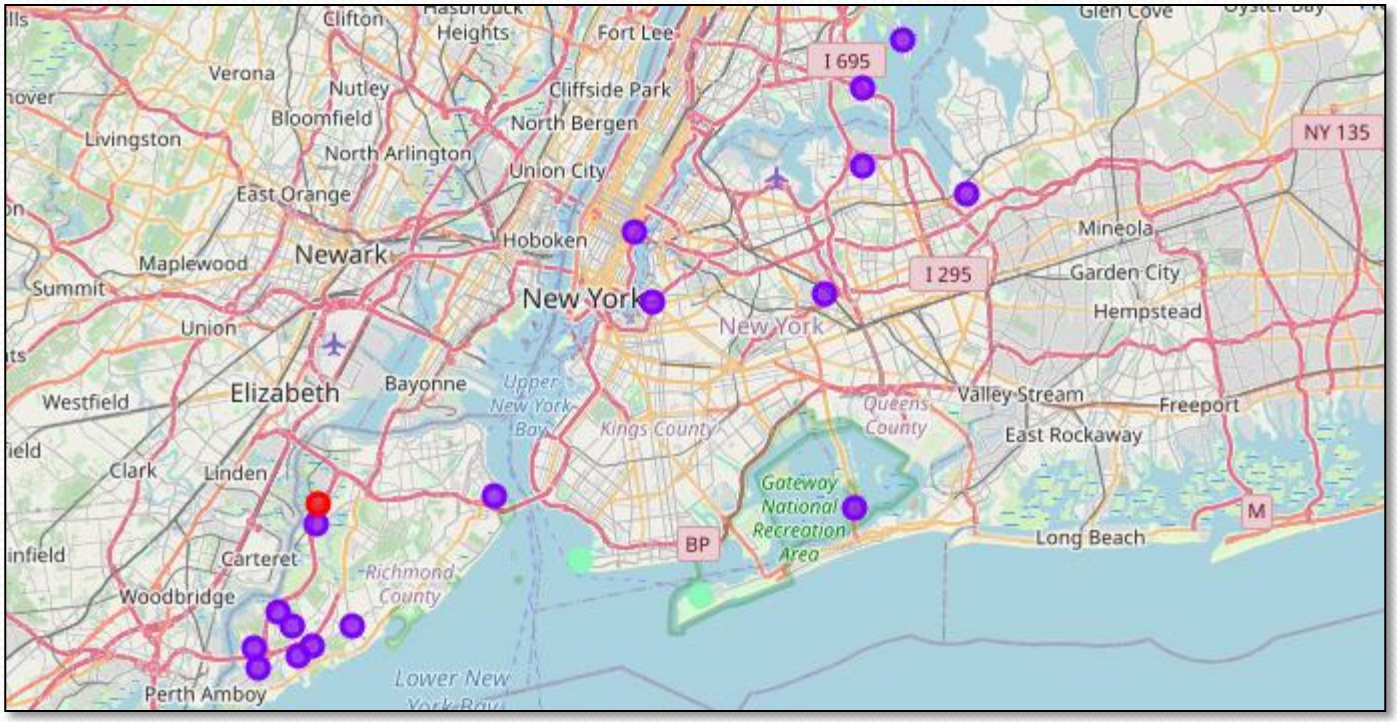
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	EDGEWATER PARK	40.821986	-73.813885	Muscle Maker Grill	40.819391	-73.817298	American Restaurant
1	EDGEWATER PARK	40.821986	-73.813885	Tommy's Pizzeria	40.819573	-73.817482	Pizza Place
2	EDGEWATER PARK	40.821986	-73.813885	The Miles Coffee Bar	40.819462	-73.817352	Coffee Shop
3	EDGEWATER PARK	40.821986	-73.813885	Tosca Café	40.819204	-73.817467	Bar
4	EDGEWATER PARK	40.821986	-73.813885	The Wicked Wolf	40.819688	-73.817359	Pub

- We create a new dataframe from the above to figure out 10 common venues across all the 20 neighborhoods.
- To find similarities in the neighborhoods k-means clustering algorithm will be used.
- In this analysis the neighborhoods will be divided into 3 clusters, this will be done on the basis of type of venues/amenities around the neighborhood.

Results

After running k-means clustering algorithm, the neighborhoods are divided into 3 clusters based on the venues/amenities around them.

Visualizing neighborhoods divided in 3 clusters. Clusters are color coded like Cluster 1 is Red, Cluster 2 is Purple and Cluster 3 is Sea Green.



Cluster 1

- The first cluster is smallest cluster with only 1 neighborhood and has venues like Theme Park, Recreation center, Fast food places and Yoga Studio. This cluster is missing out on basic amenities like supermarket or grocery/departmental stores.

	Neighborhood	Borough	FELONY	MISDEMEANOR	VIOLATION	Total	Latitude	Longitude	Cluster Labels	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
8	BLOOMFIELD	STATEN ISLAND	0	4	4	8	40.605779	-74.187256	0	Theme Park	Recreation Center	Burger Joint	Bus Stop	Yoga Studio	Food Truck	Food & Drink Shop	Food	Fast Food Restaurant	Event Space

Cluster 3

- The third cluster contains 2 neighborhoods it has venues Beach, Restaurants, Departmental stores, Yoga studios, Spa and Sports club and

	Neighborhood	Borough	FELONY	MISDEMEANOR	VIOLATION	Total	Latitude	Longitude	Cluster Labels	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
12	BREEZY POINT	QUEENS	1	6	2	9	40.557401	-73.925512	2	Beach	Trail	Monument / Landmark	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food & Drink Shop	Food
17	SEA GATE	BROOKLYN	4	5	2	11	40.576375	-74.007873	2	American Restaurant	Beach	Bus Station	Spa	Sports Club	History Museum	French Restaurant	Department Store	Dessert Shop	Diner

Cluster 2

The second cluster is biggest cluster with only 17 out of 20 neighborhoods it has venues like Restaurants, Bars/Pubs, Grocery/Departmental stores and other essential amenities.

	Neighborhood	Borough	FELONY	MISDEMEANOR	VIOLATION	Total	Latitude	Longitude	Cluster Labels	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	EDGEWATER PARK	BRONX	0	1	0	1	40.821986	-73.813885	1	Italian Restaurant	Pizza Place	Donut Shop	American Restaurant	Liquor Store	Park	Pub	Coffee Shop	Chinese Restaurant	Sports Bar
1	SOUTH SIDE	BROOKLYN	0	1	0	1	40.710861	-73.958001	1	Bar	Coffee Shop	American Restaurant	Pizza Place	Wine Bar	Breakfast Spot	Yoga Studio	Japanese Restaurant	Sushi Restaurant	Chinese Restaurant
2	CHARLESTON	STATEN ISLAND	2	1	0	3	40.530531	-74.232158	1	Big Box Store	Coffee Shop	Cosmetics Shop	Japanese Restaurant	Furniture / Home Store	Pet Store	Music Venue	Kids Store	Hardware Store	Gym / Fitness Center
3	CHELSEA	STATEN ISLAND	0	1	3	4	40.594726	-74.189560	1	Steakhouse	Spanish Restaurant	Park	Sandwich Place	Dry Cleaner	Food Truck	Food & Drink Shop	Food	Fast Food Restaurant	Event Space
4	HUGUENOT	STATEN ISLAND	0	3	1	4	40.531912	-74.191741	1	Italian Restaurant	Asian Restaurant	Donut Shop	Sandwich Place	Moving Target	Bank	Deli / Bodega	Train Station	Ice Cream Shop	Heliport
5	PRINCES BAY	STATEN ISLAND	2	3	1	6	40.526264	-74.201526	1	Pizza Place	Bank	Pharmacy	Pet Store	Construction & Landscaping	Bagel Shop	Sushi Restaurant	Chinese Restaurant	Dry Cleaner	Food & Drink Shop
6	BROAD CHANNEL	QUEENS	4	3	0	7	40.603027	-73.820055	1	Deli / Bodega	Pizza Place	Sporting Goods Shop	Dive Bar	Other Nightlife	Bus Station	Event Space	Food & Drink Shop	Food	Fast Food Restaurant
7	ROSSVILLE	STATEN ISLAND	1	3	3	7	40.549404	-74.215729	1	Pizza Place	Bagel Shop	American Restaurant	Chinese Restaurant	Liquor Store	Grocery Store	Moving Target	Dry Cleaner	Donut Shop	Deli / Bodega
9	DOUGLASTON	QUEENS	3	4	1	8	40.766846	-73.742498	1	Deli / Bodega	Chinese Restaurant	Bank	Ice Cream Shop	Lounge	Diner	Convenience Store	Donut Shop	Fast Food Restaurant	Shanghai Restaurant
10	SHORE ACRES	STATEN ISLAND	0	5	3	8	40.609719	-74.066678	1	Italian Restaurant	Bus Stop	Intersection	Deli / Bodega	Bar	Supermarket	Gastropub	Furniture / Home Store	Music Store	Food
11	TUDOR CITY	MANHATTAN	3	5	0	8	40.746917	-73.971219	1	Park	Café	Mexican Restaurant	Deli / Bodega	Garden	Seafood Restaurant	Wine Shop	Gym	Greek Restaurant	Dog Run
13	CITY ISLAND	BRONX	2	2	5	9	40.847247	-73.786488	1	Thrift / Vintage Store	Seafood Restaurant	Grocery Store	Boat or Ferry	Diner	Smoke Shop	Pizza Place	Pharmacy	Park	Deli / Bodega
14	WOODROW	STATEN ISLAND	4	3	2	9	40.541968	-74.205246	1	Pharmacy	Grocery Store	Donut Shop	Cosmetics Shop	Coffee Shop	Chinese Restaurant	Miscellaneous Shop	Mexican Restaurant	Diner	Martial Arts Dojo

Discussion

The idea behind the project was to help people to find the safest neighborhood to live with basic amenities which can make people's life easier. For example after the above analysis we can say that Neighborhoods that are present in Cluster 2 and 3 will be ideal for stay as per the various requirements. Like if a person wants to have basic amenities in his/her vicinity then they can easily choose neighborhood from Cluster 2 as per their requirement and if person prefers to have a beach nearby his/her house along with amenities then they can choose from Cluster 3. Ultimately it depends on people's choices and requirements which neighborhood to choose.

Conclusion

Choosing a place to live can be hectic task that takes up significant amount of time, money and effort. But going through technical or data oriented approach can save people significant amount of their resources and effort. This model has turned out to be very helpful shortlisting places and then further narrowing it down based on the requirement. Such models aren’t helpful at the individual level but also at a large scale for example real estate agents or companies can find ideal places for their clients using this approach.

Future Directions

Since in this problem, we are trying to find a suitable neighborhood for a client. We can also take another aspect of a person's requirement i.e. Budget whether a person buying a property or looking for a rented one. We should be able to accommodate that requirement into this model as well. This is something on which one can work on improving the model further.