The Battle of Neighborhoods- Report

- **Nishan Sah**

1. Introduction and Business Problem:

New York City is the most populous city and the financial capital of the United States. During early 2020, it became the biggest victim of COVID-19 in the country, a virus that caused a devastating pandemic. A lot of cases and deaths have been reported since then, an extensive lockdown was imposed by the government, the economy was hit very hard and overall, there has been a lot of damage to public life and it is very obvious that the transition to normalcy will be a very hard and tedious process.

New York City is among the most important hubs for businesses and commerce in the world. When the economy normalizes again, the economy will boom as usual and markets will become competitive again.

But that doesn't change the fact that people have been disproportionately affected by the pandemic and a lot of them will be struggling economically. This includes a lot of business owners or people who were about to start a venture, and also for investors who will be skeptical about investing in New York, given that the situation is very volatile and public safety is hard to ensure.

As of now, the situation has been brought to control and the transition period has started. The insights derived from the impact of the virus in different places will allow us to understand where business and investment will be least risky and the return on investment will be reasonable.

1.1 Problem Description:

One of the major businesses that will play an active role in contributing to the economy after the pandemic is completely contained and the lockdown is eased is the restaurant business. There has been an active shortage of eatables and buying groceries from department stores isn't financially feasible for many people, let alone cook at home because of their busy lifestyle. So, people have always been inclined to dine outside and order food (ordering has also been prominent during the lockdown) and now, since people want to avoid contact as much as possible,

cheap restaurants and takeaway food will be the go-to option for many people once life resumes.

With social distancing norms and safety concerns, it will be hard for new ventures to open up. So, after the lockdown it will be important to put safety first while opening up a restaurant strategically, which means to find places least affected by COVID-19. Places that have been least impacted will also do reasonably well in profits, given that get enough customers. But there are also other economic factors to understand while opening up a business like:

1. Average income of the residents.
2. Competition. There are about 50,153 restaurants (as of 2018) in New York City. It is important to analyze the scene of restaurants in the location.
3. Access to raw materials and ingredients. For this, we need to look at locations of Farmer's market and Wholesale markets and how accessible they are.
4. Market saturation.

Amongst many other factors. So, this project will help anyone seeking to open up a restaurant business economically feasible for the people or the ones looking to invest in them as soon as possible by giving them a choice of location where investment will be wise.

Target audience:

1. Anyone looking to start a new, restaurant business in New York City after the lockdown.
2. Anyone looking to invest in existing restaurants in New York City soon.
3. Anyone looking for safe places to eat.

Success Criteria:

The success criteria of this project are contingent upon helping the stakeholder find the safest, yet economically viable and competent neighborhood in the safest borough for a restaurant business by forecasting the effect of COVID-19 in New York City and other economic factors.

2. Data Used in Analysis

Data 1:

New York City will be analyzed in this project and for this we will use the link https://geo.nyu.edu/catalog/nyu_2451_34572 to get required geospatial data in a json file and convert it to a readable format.

Data 2:

Daily statistics of COVID-19 cases in each borough to be used for forecasting and finding the safest borough pulled from: https://github.com/nychealth/coronavirusdata/blob/master/boro/boroughs-case-hosp-death.csv (latest date:26th May 2020)

Data 3:

COVID-19 statistics in each neighborhood to find the safest neighborhood pulled from: https://github.com/nychealth/coronavirus-data/blob/master/data-by-modzcta.csv (latest date: 26th May 2020)

Data 4:

Income statistics in New York City by neighborhood to analyze the economic conditions of the desired location to understand the economic demographic of the chosen location better: https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability

Data 5:

Data on Farmer's market and their locations to find the best proximity to open a restaurant: https://data.cityofnewyork.us/dataset/DOHMH-Farmers-Markets/8vwk-6iz2/data

Data 6:

Foursquare API will be use to explore neighborhoods and generate venue information for each neighborhood to find restaurants nearby and Income statistics in New York City by neighborhood to analyze the economic conditions of the desired location to understand the economic demographic of the chosen location better: https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability competition.

Data 7:

Information on demographics and population of New York City for better understanding and calculations: https://en.wikipedia.org/wiki/New_York_City

3. Methodology

3.1 Part 1: COVID-19 Analysis in New York City

First, I imported two databases, one being New York City's geospatial data and the other being COVID-19 statistics from the government's github repository. The geospatial data was present in a .json file, so initially the file had to be transformed into a readable format. After conversion, we see the following:

Geospatial data:

	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

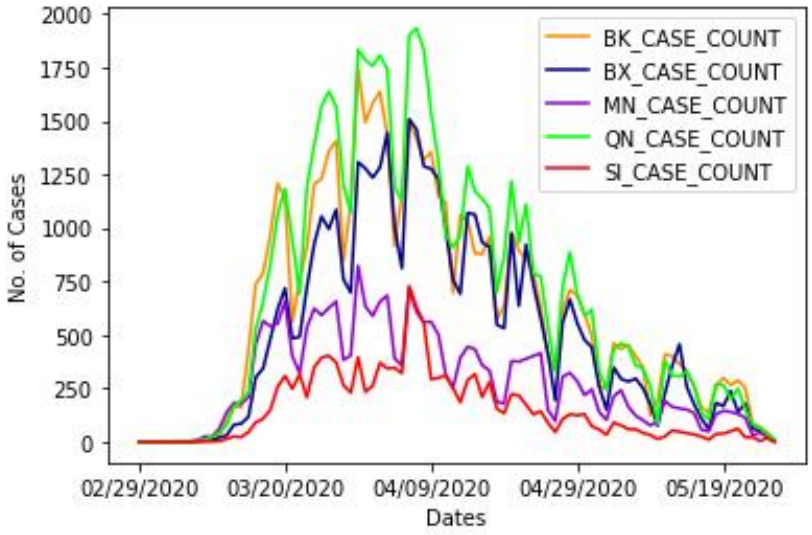
Here, we can see the names of boroughs and their respective neighborhoods with their geospatial co-ordinates. There are **five boroughs in New York City- Brooklyn, The Bronx, Manhattan, Queen and Staten Island**. For reference throughout this report, they will be written as BK, BX, MN, QN, and SI respectively in various instances.

COVID-19 Statistics:

	DATE_OF_INTEREST	BK_CASE_COUNT	BK_HOSPITALIZED_COUNT	BK_DEATH_COUNT	BX_CASE_COUNT	BX_HOSPITALIZED_COUNT	BX_DEATH_COUNT
0	02/29/2020	0	4	0	0	4	0
1	03/01/2020	0	0	0	1	1	0
2	03/02/2020	0	5	0	0	9	0
3	03/03/2020	0	4	0	1	8	0
4	03/04/2020	1	6	0	0	6	0

The table has more rows than seen in the picture, which includes the dates since COVID-19 cases started to be noticed, along with new counts on each day, along with hospitalized count and death count for each borough in its own respective column. However, for the analysis in this project, case counts per day will suffice, as we only want to see the impact on number of cases that has affected the population, not the intricacies of it because they won't help us proceeding further.

So, I cleaned and arranged the dataframe and only included the case counts. Upon plotting a graph, we receive the following result:



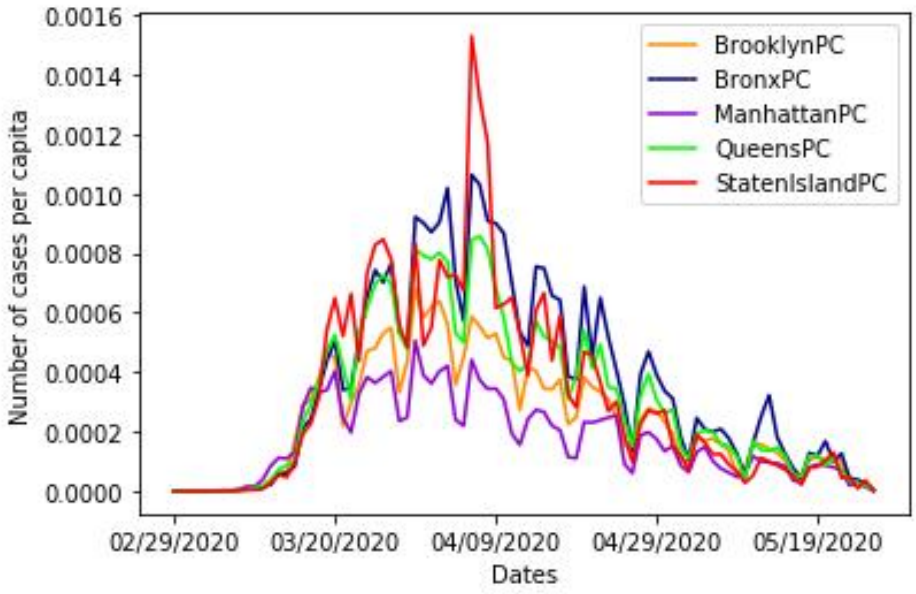
The data has a significant difference between the boroughs, which is not clear enough to understand the impact of COVID-19 in a more holistic way.

For this, I retrieved data of population in each borough which will allow us to calculate the propensity of contracting COVID-19 cases at an individual level, and get a more magnified view of cases in the boroughs. Clearly, the boroughs have diverse areas and some tend to be more populated than others so it is necessary to evaluate the cases at the individual level, which will give us the clarity of the threat posed by the virus over time in New York. For this we'll evaluate the likeliness of contracting the virus each day by diving the number of cases on a given day by the total population. The result, although vague and very low will account for the disproportionate population. We will form new columns for this, for every borough, and see the line plot again.

The following columns are added to the table.

BrooklynPC	BronxPC	ManhattanPC	QueensPC	StatenIslandPC
0.000099	0.000126	6.999422e-05	0.000068	0.000044
0.000033	0.000043	1.841953e-05	0.000037	0.000046
0.000028	0.000038	2.701531e-05	0.000029	0.000008
0.000017	0.000023	1.289367e-05	0.000016	0.000036
0.000002	0.000004	6.139844e-07	0.000003	0.000000

We get the following plot when plot these rows against the dates.

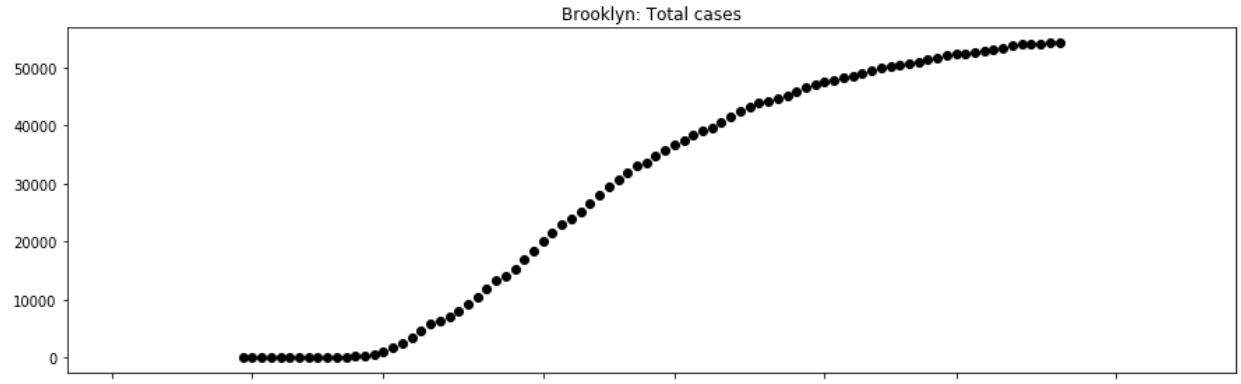


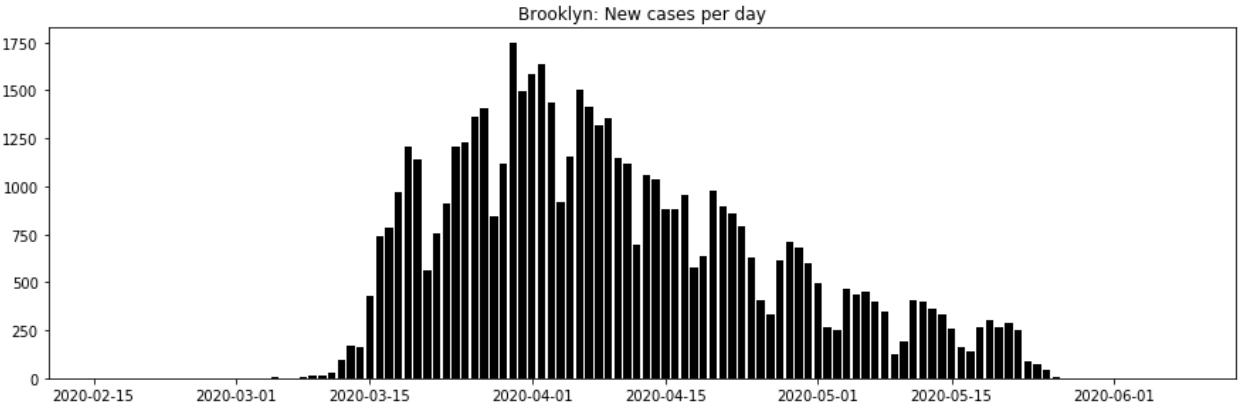
We can see how the data changes when we account for population difference. Staten Island, that was previously low on the graph, suddenly peaks above other boroughs and the visuals change drastically. This shows that individuals are under varying risk to be affected by COVID-19 on an individual level, depending on their location and the number of people around them.

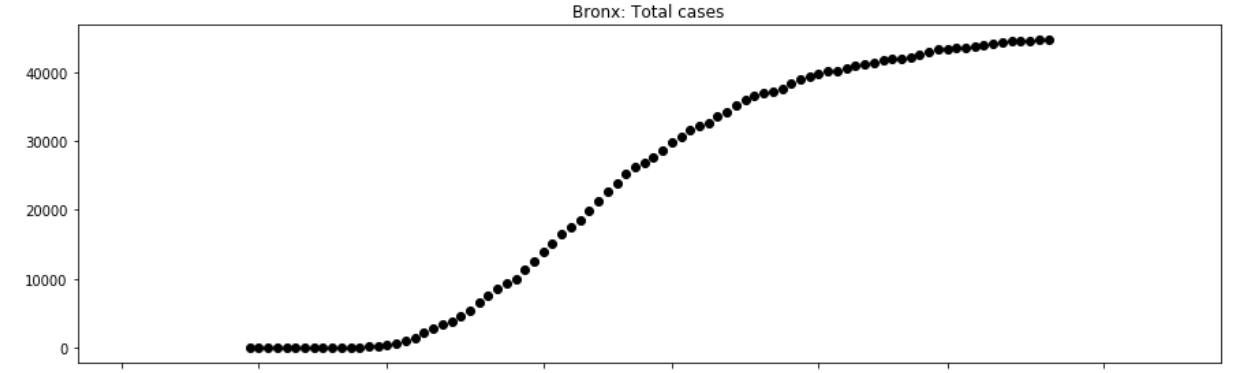
We will now analyze these boroughs individually. For these the data from the main table is taken and appended into new tables for individual boroughs. Then I calculated how the cases have gradually tallied up in each borough everyday by summing all the cases until a given date from the case counts column and for that a new column is needed. For instance, this is what Brooklyn's final table looks like (the column names have been changed from BK_CASE_COUNT to NewCases and from BrooklynPC to CasesPC):

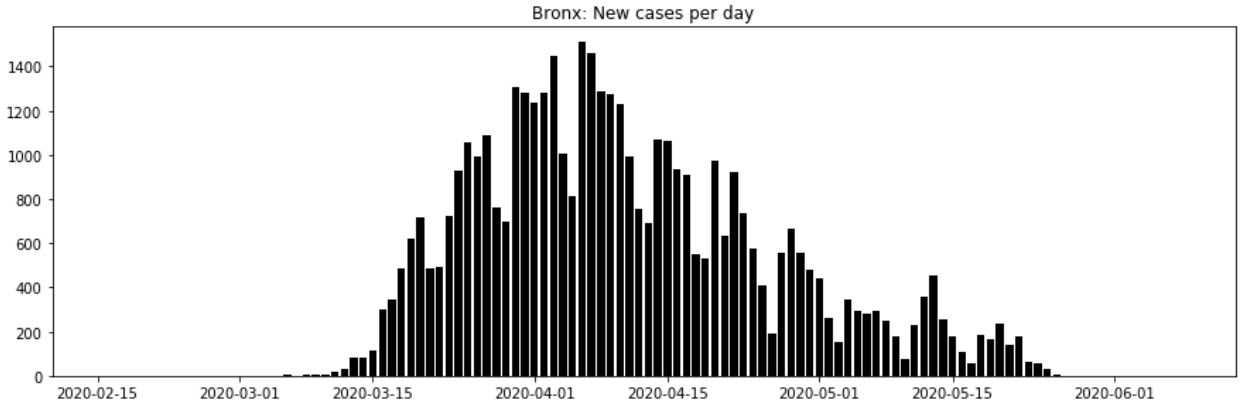
DATE_OF_INTEREST	NewCases	CasesPC	TotalCases
2020-02-29	0	0.000000e+00	0
2020-03-01	0	0.000000e+00	0
2020-03-02	0	0.000000e+00	0
2020-03-03	0	0.000000e+00	0
2020-03-04	1	3.906398e-07	1

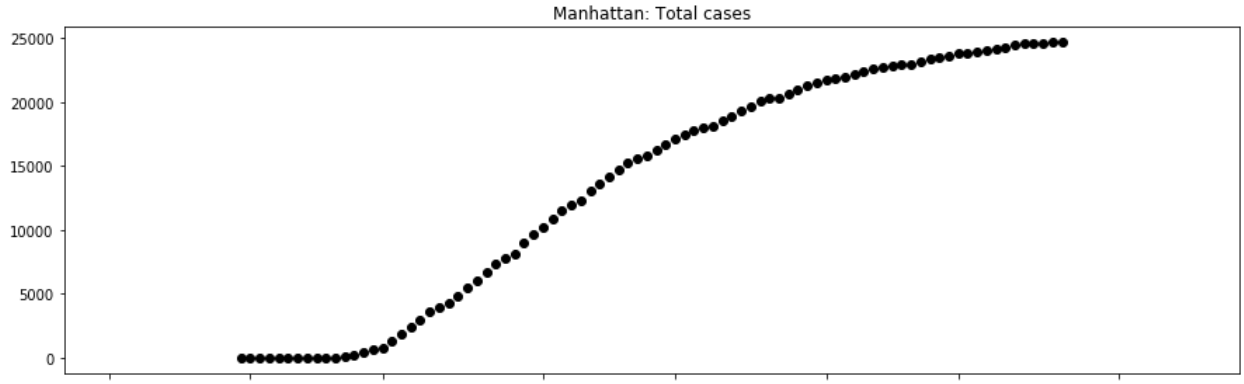
Let us now visualize the trends of total cases and new cases from the data:

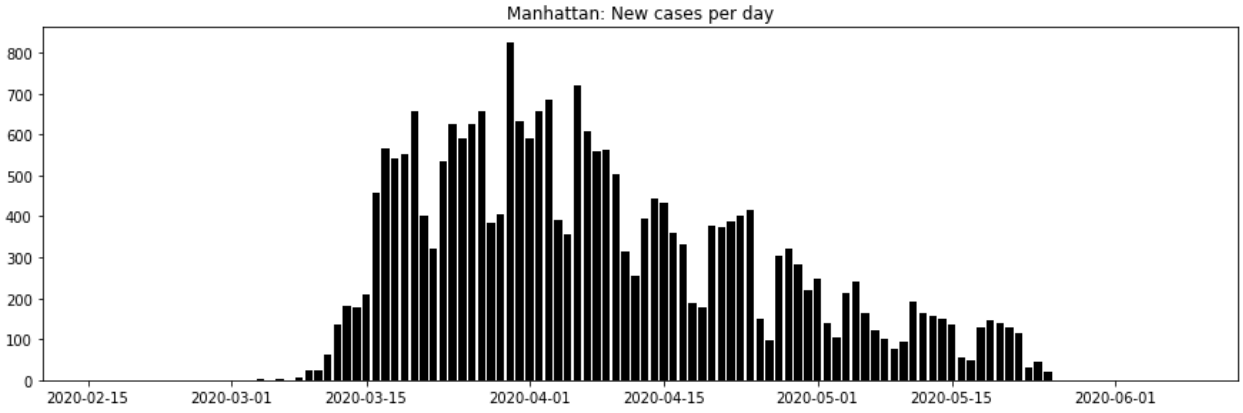

Brooklyn: Total cases
50000
40000
30000
20000
10000
0

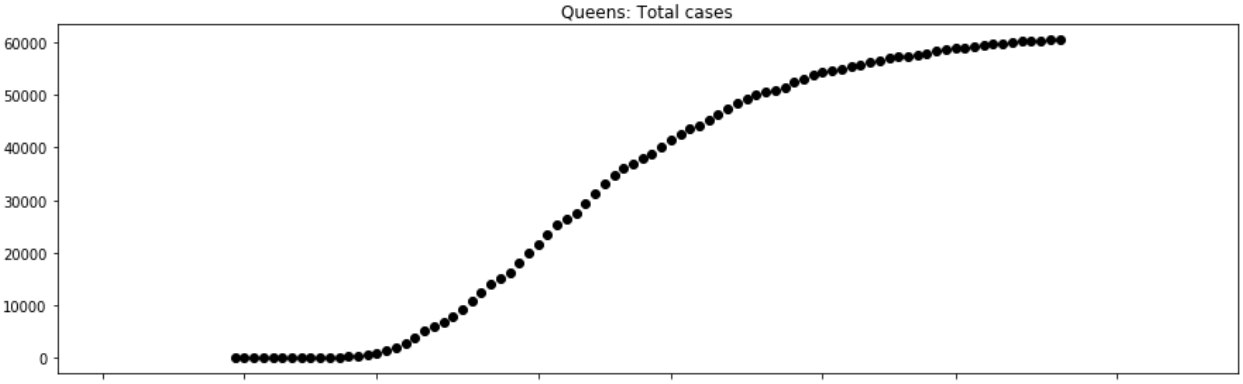

Brooklyn: New cases per day
1750
1500
1250
1000
750
500
250
0
2020-02-15
2020-03-01
2020-03-15
2020-04-01
2020-04-15
2020-05-01
2020-05-15
2020-06-01

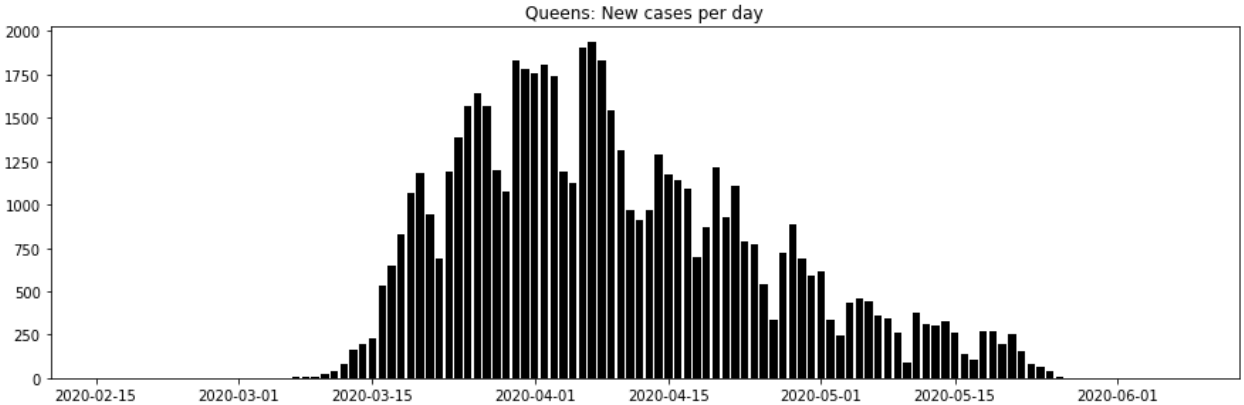

Bronx: Total cases
40000
30000
20000
10000
0

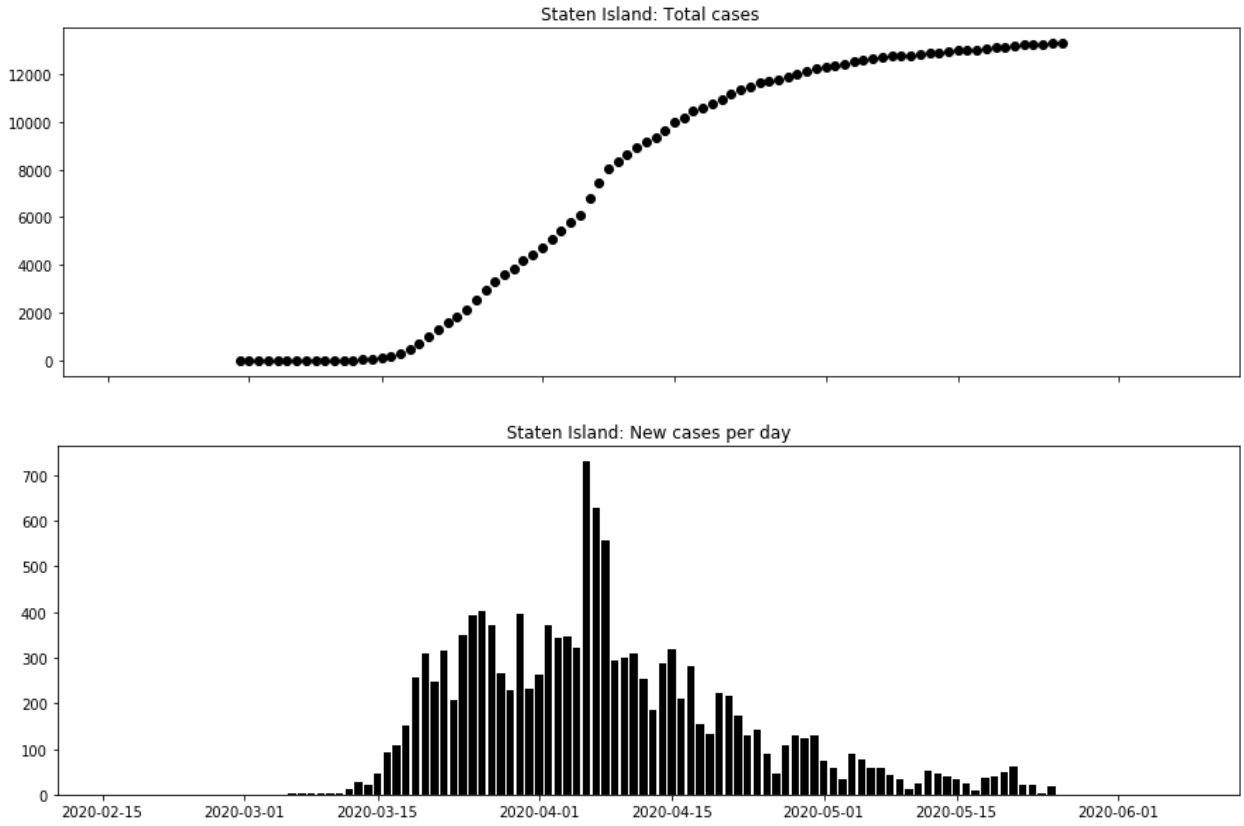

Bronx: New cases per day
1400
1200
1000
800
600
400
200
0
2020-02-15
2020-03-01
2020-03-15
2020-04-01
2020-04-15
2020-05-01
2020-05-15
2020-06-01


Manhattan: Total cases
25000
20000
15000
10000
5000
0


Manhattan: New cases per day
800
700
600
500
400
300
200
100
0
2020-02-15
2020-03-01
2020-03-15
2020-04-01
2020-04-15
2020-05-01
2020-05-15
2020-06-01


Queens: Total cases
60000
50000
40000
30000
20000
10000
0


Queens: New cases per day
2000
1750
1500
1250
1000
750
500
250
0
2020-02-15
2020-03-01
2020-03-15
2020-04-01
2020-04-15
2020-05-01
2020-05-15
2020-06-01



From the graph, we can notice that the number of cases has started to recede and the total cases have nearly saturated, which is a consequence of lockdown being imposed. Queens has comparatively been the worst hit place with the most cases.

Seeing the data, we can say that the ideal time to live in a borough would be when the number of new cases are consistently zero which also obviously means the possibility of not contracting the virus will cease to exist at an individual level (in the most ideal scenario). So, for this we need to use machine learning algorithms and, in this case, we can perform time series analysis to figure out which borough is the safest, and if possible, also find out which borough will be relatively safer as early as possible. Time series will use the existing trends to build a model that will predict the behavior of this virus in the future and the way the cases will exceed/recede as the dates increase.
For this, I have used IBM's SPSS modeler and did a time series analysis using the following stream for all boroughs.

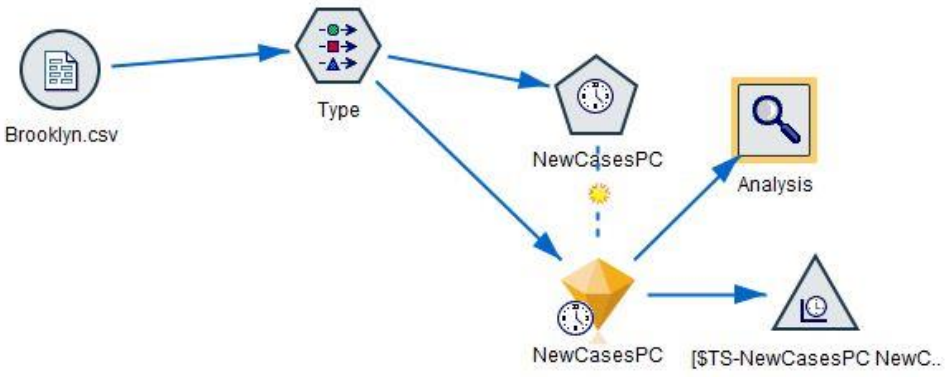
- First, I downloaded bk,bx,mn,qn and si by converting them into csv files.
- In each stream, I first added the respective csv file, then figured out the target variable (NewCasesPC), then implemented a time series model and

received results and made plots with the data from the model. It is represented as $TS-NewCasesPC in the graphs below.

- The forecast was done for 2 months from when the data was available (26th May onwards).

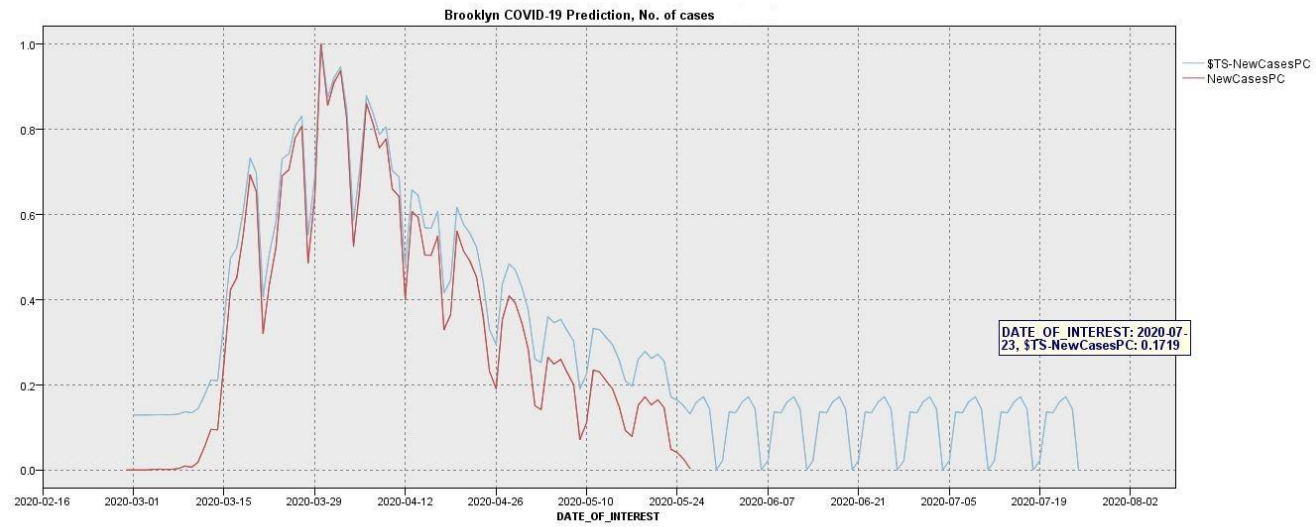
Seeing how important measures have already been taken to contain the virus and a lot of testing has been done, this analysis assumes that number of cases per day have already reached their peak. The data has been normalized for better understanding.

Here's an example of a stream:



Forecast for Brooklyn

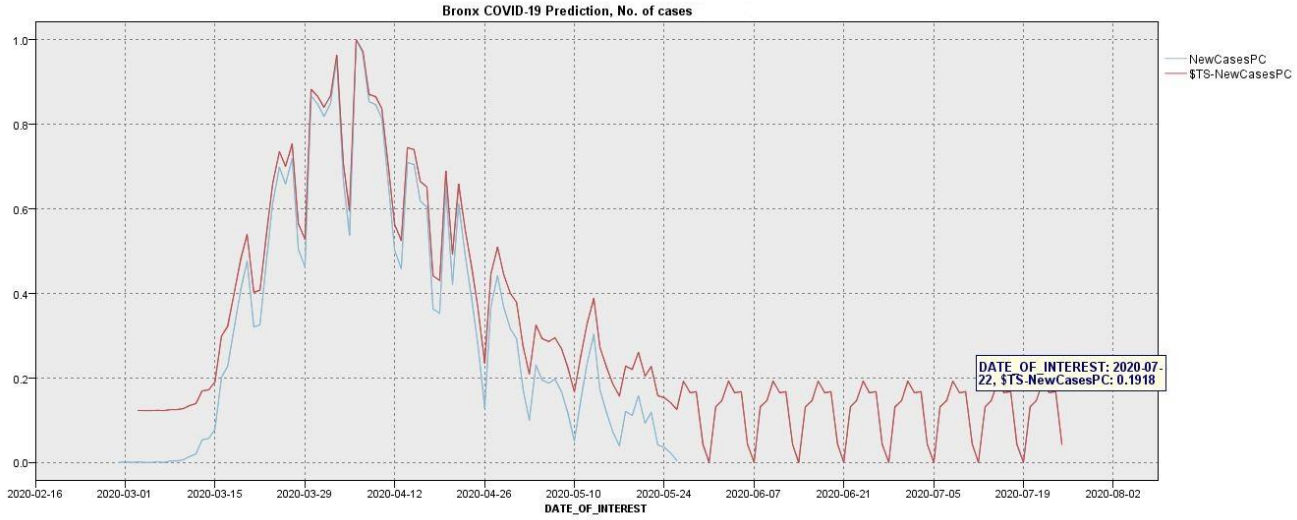
It is observed that there is a repeating pattern from May 27, with the peak being at 0.1719 cases per day and lowest value reaching zero. Which is to say, the likeliness of a new case popping up is 17.19% of the peak value. Starting from May 30, the range is between 0-0.1719 cases per head. Let's see the same for other boroughs.



Forecast for The Bronx

Peak number of cases per day: 0.1918 (19.18% of peak value)

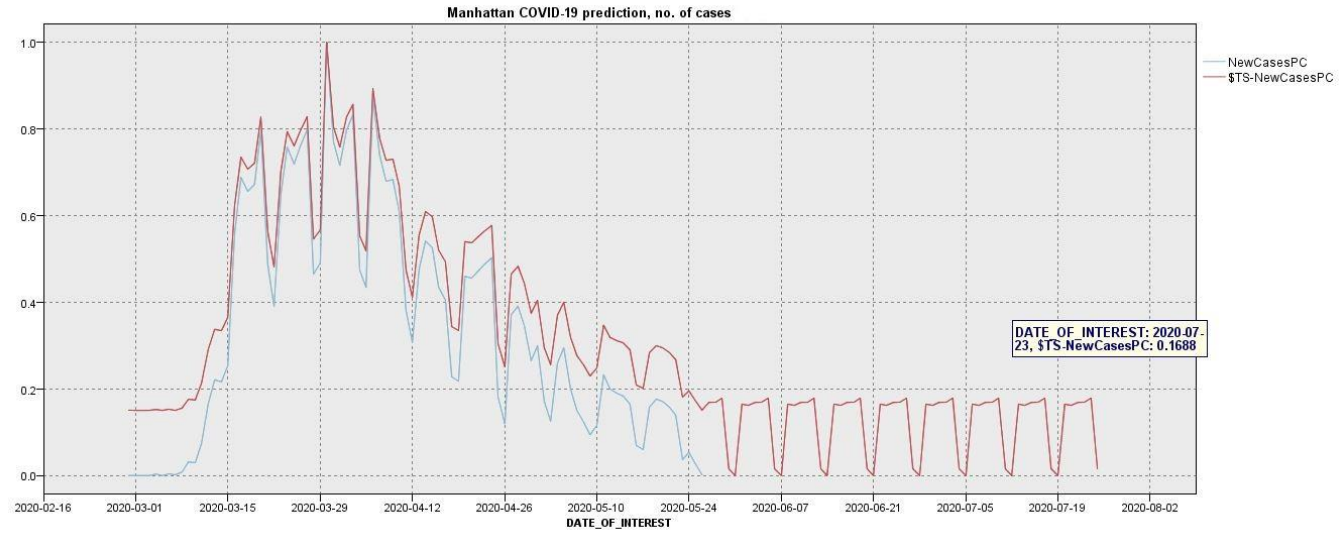
Range: 0 to 0.1918 new cases per day (starting from May 31)



Forecast for Manhattan

Peak number of cases per head: 0.1688 (16.88% of peak value)

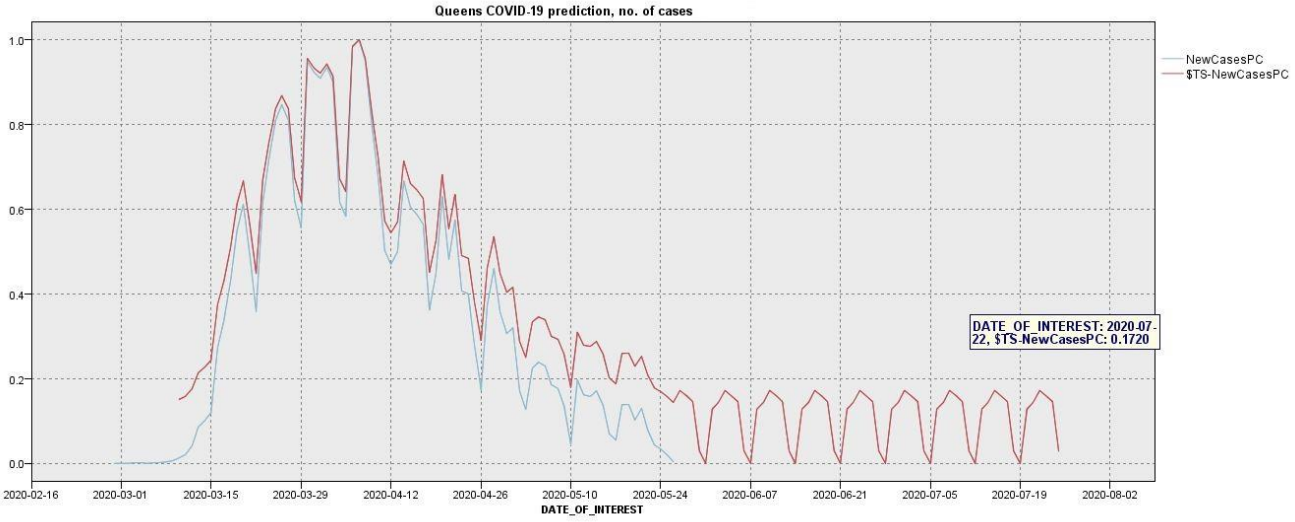
Range: 0 to 0.1688 new cases per day (starting from May 31)



Forecast for Queens

Peak number of cases per day: 0.1720 (17.20% of peak value)

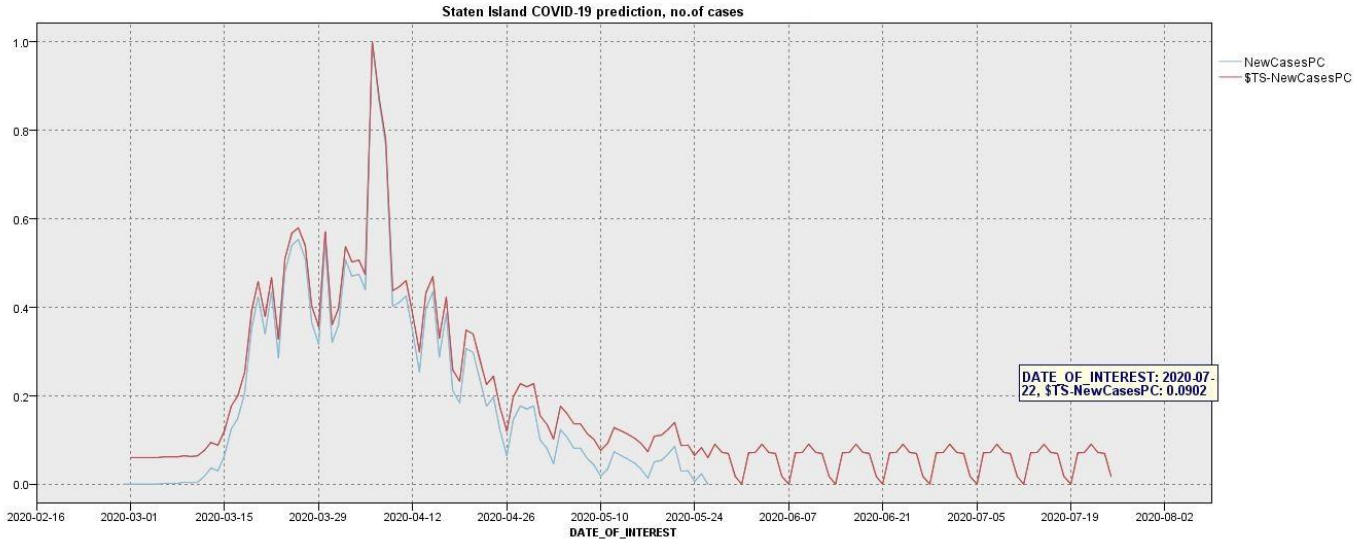
Range: 0 to 0.1720 new cases per day (starting from May 31)



Forecast for Staten Island

Peak number of cases per day: 0.0902 (9.02% of peak value)

Range: 0 to 0.0902 new cases per day (starting from May 30)

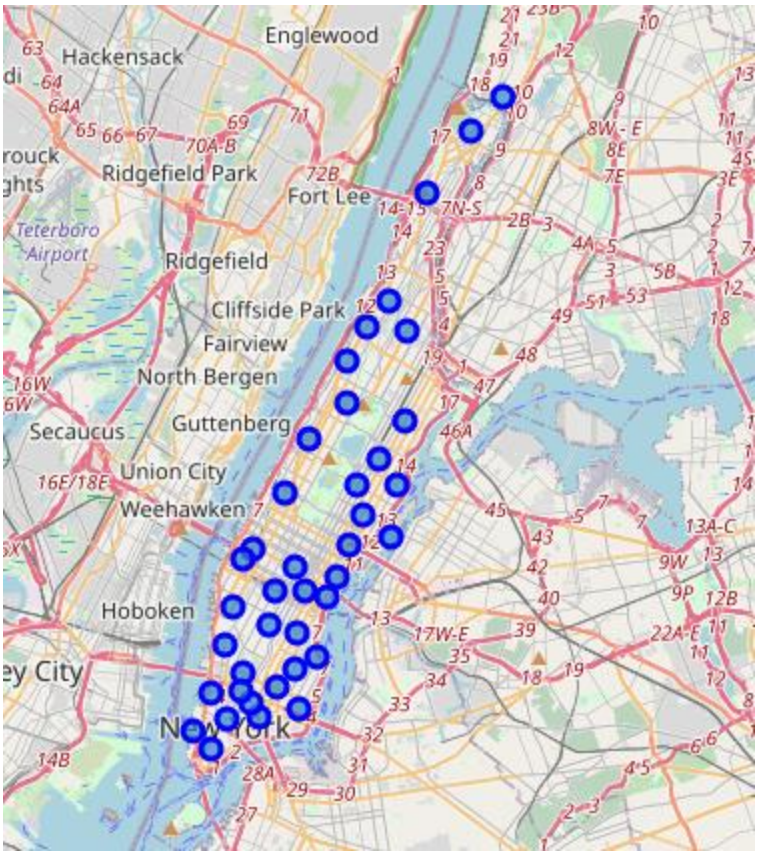


According to the model, the virus has been best contained in Manhattan, as the peak value is much lower when we considered observations of number of new cases per day which also means least propensity to be subjected to the virus, making it the safest borough in New York. I shall now further analyze Manhattan and find the most ideal neighborhood to open a restaurant in.

3.2 Part 2: Neighborhood Analysis

For this I took in data from the geospatial data table from earlier and only used the rows that have Manhattan as the input in the Borough column. In total, Manhattan has 40 neighborhoods.

The visual representation of all neighborhoods in Manhattan map is given below:



Then I imported a new database from the NYC government's github repository, that contains COVID-19 statistics for neighborhoods only.

	MODIFIED_ZCTA	NEIGHBORHOOD_NAME	BOROUGH_GROUP	COVID_CASE_COUNT	COVID_CASE_RATE	POP_DENOMINATOR	COVID_DEATH_COUNT
0	10001	Chelsea - Clinton	Manhattan	358	1519.33	23563.03	21
1	10002	Union Square - Lower East Side	Manhattan	1027	1338.02	76755.41	143
2	10003	Union Square - Lower East Side	Manhattan	445	827.11	53801.62	32
3	10004	Lower Manhattan	Manhattan	31	849.17	3650.61	1
4	10005	Lower Manhattan	Manhattan	61	726.53	8396.11	2

The columns not included in this are POP_DENOMINATOR, COVID_DEATH_COUNT, COVID_DEATH_RATE, PERCENT_POSITIVE

I cleaned this table and only took in the relevant data to Manhattan. The final table looks like this. I removed the column containing zip codes and added the case

counts where the neighborhood name was the same, while taking means of percent positive and case rate.

	NEIGHBORHOOD_NAME	CaseCount	PercentPositive	CaseRate
0	Central Harlem - Morningside Height	3158	24.428000	1867.592000
1	Chelsea - Clinton	2069	18.104000	1535.208000
2	East Harlem	2711	28.145000	2479.360000
3	Gramercy Park - Murray Hill	1404	13.585000	1009.585000
4	Greenwich Village - Soho	709	15.090000	831.026667
5	Lower Manhattan	536	14.864286	832.727143
6	Union Square - Lower East Side	2157	19.040000	1110.376667
7	Upper East Side	2608	17.416667	1344.113333
8	Upper West Side	2559	16.365000	1082.575000
9	Washington Heights - Inwood	5836	27.052000	2148.944000

Upon counting, there are only 16 neighborhoods that have been affected, out of 40. We shall now remove the neighborhoods that have been affected from our further analysis as they are comparatively unsafe, and still have a decent sample space to work with.

We shall now look at these neighborhoods from an economic perspective. Since the sources used to extract data from had different ways of considering neighborhoods in Manhattan, manual aggregation was necessary. In an excel sheet, I added population from 2010, which was when the last census took place and it was projected that the population of Manhattan would increase by 3.3% so I put in the same estimation and rounded off the values to calculate the population of the neighborhoods in 2020. Then I added median rent column to give an impression of (in this instance, for two bedroom apartments) and median income data to provide an idea of economic conditions of these neighborhoods. For neighborhoods that were an aggregation of 2 or more places, average of latitude and longitude values were used and any neighborhood that contained the places affected by COVID-19 were removed as they were in close vicinity and we need the safest place. We will now look at these data intricately. This is how the dataframe looks like at the moment:

Neighborhood	Neighborhood	Median Income	Population 2010	Median Rent	Population 2020	Latitude	Longitude
Lincoln Square	Lincoln Square	120337	61489	5232.5	63518	40.773529	-73.985338
Midtown-Midtown South	Midtown-Midtown South	114491	28630	3950.0	29575	40.751601	-73.985191
Turtle Bay-East Midtown	Turtle Bay-East Midtown	113998	51231	3967.0	52922	40.752042	-73.967708
West Village	West Village	112689	66880	4295.0	69087	40.734434	-74.006180
Yorkville	Yorkville	98840	77942	3250.0	80514	40.775930	-73.947118

After this, using Foursquare API we make a call to get venue details of these neighborhoods. Venues, here imply landmarks, malls, restaurants, theaters or any public place. The data received is comprehensively transformed into a dataframe, that is given below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lincoln Square	40.773529	-73.985338	The Metropolitan Opera (Metropolitan Opera)	40.772742	-73.984401	Opera House
1	Lincoln Square	40.773529	-73.985338	Vivian Beaumont Theater	40.773354	-73.983827	Theater
2	Lincoln Square	40.773529	-73.985338	American Ballet Theatre	40.772668	-73.984476	Performing Arts Venue
3	Lincoln Square	40.773529	-73.985338	Walter Reade Theater	40.773783	-73.983924	Indie Movie Theater
4	Lincoln Square	40.773529	-73.985338	New York Philharmonic	40.772526	-73.983139	Concert Hall

In total, this returns 225 unique venue categories and 927 results in total. Upon further analysis, we get the following results. First, we count individual number of venues in each neighborhood, followed by number of restaurants in each neighborhood. This gives us an insight of what the public life in those areas might look like.

```
Number of venues in Lincoln Square: 98
Number of venues in Midtown-Midtown South: 95
Number of venues in Turtle Bay-East Midtown: 100
Number of venues in West Village: 100
Number of venues in Yorkville: 100
Number of venues in Lenox Hill-Roosevelt Island: 36
Number of venues in Stuyvesant Town-Cooper Village: 18
Number of venues in East Village: 100
Number of venues in Central Harlem South: 46
Number of venues in Hamilton Heights: 60
Number of venues in Chinatown: 100
Number of venues in Central Harlem North-Polo Grounds: 35
Number of venues in Manhattanville: 39
```

```
Number of restaurants in Lincoln Square: 17
Number of restaurants in Midtown-Midtown South: 27
Number of restaurants in Turtle Bay-East Midtown: 35
Number of restaurants in West Village: 25
Number of restaurants in Yorkville: 29
Number of restaurants in Lenox Hill-Roosevelt Island: 9
Number of restaurants in Stuyvesant Town-Cooper Village: 0
Number of restaurants in East Village: 36
Number of restaurants in Central Harlem South: 14
Number of restaurants in Hamilton Heights: 17
Number of restaurants in Chinatown: 36
Number of restaurants in Central Harlem North-Polo Grounds: 2
Number of restaurants in Manhattanville: 1
```

Now, we'll try to understand the preferred cuisines of Manhattan better. Using kMeans algorithm I clustered these neighborhoods as per the kind of restaurant and each cluster will give us an empirical data-based insight of common tastes of the demographic and each cluster will be a different group:

	Borough	Neighborhood	Latitude	Longitude	Cluster	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 Mos Commo Venu
1	Manhattan	Chinatown	40.715618	-73.994279	0.0	Chinese Restaurant	Vietnamese Restaurant	American Restaurant	Asian Restaurant	Malay Restaurant	Mexican Restaurant	Shanghai Restaurant	Dim Sur Restauran
2	Manhattan	Hamilton Heights	40.823604	-73.949688	2.0	Mexican Restaurant	Chinese Restaurant	Indian Restaurant	Sushi Restaurant	Caribbean Restaurant	Japanese Restaurant	Latin American Restaurant	Mediterranea Restauran
3	Manhattan	Manhattanville	40.816934	-73.957385	2.0	Seafood Restaurant	Italian Restaurant	Mexican Restaurant	Indian Restaurant	Chinese Restaurant	Dumpling Restaurant	Ramen Restaurant	Cuba Restauran
5	Manhattan	Yorkville	40.775930	-73.947118	2.0	Italian Restaurant	Sushi Restaurant	Japanese Restaurant	Mexican Restaurant	Chinese Restaurant	Vietnamese Restaurant	Indian Restaurant	Turkis Restauran
8	Manhattan	Lincoln Square	40.773529	-73.985338	1.0	Italian Restaurant	American Restaurant	French Restaurant	Mediterranean Restaurant	Seafood Restaurant	Greek Restaurant	Mexican Restaurant	Chines Restauran
10	Manhattan	East Village	40.727847	-73.982226	2.0	Mexican Restaurant	Japanese Restaurant	Seafood Restaurant	Filipino Restaurant	Greek Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Rame Restauran
13	Manhattan	West Village	40.734434	-74.006180	1.0	Italian Restaurant	American Restaurant	New American Restaurant	Restaurant	Seafood Restaurant	Korean Restaurant	French Restaurant	Mediterranea Restauran

Now, using a dataframe obtained from New York City's Department of Health and Mental Hygiene, we obtained locations of the Farmer's market as well. For our interest, we only look at locations in Manhattan.

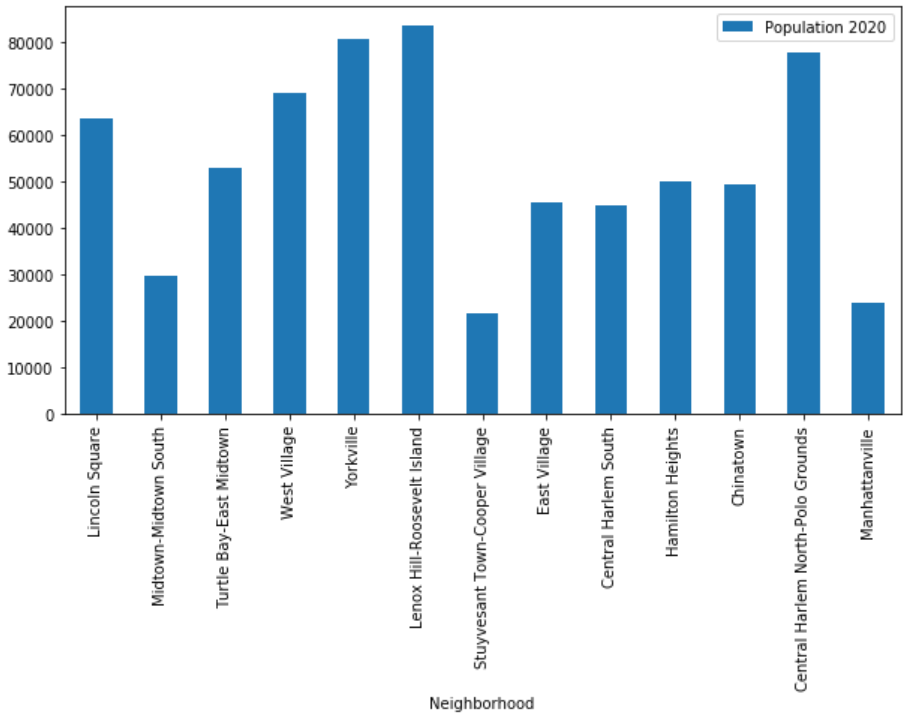
We can clearly see that the neighborhoods are stacked with farmer's market, which means getting ingredients won't be a hassle. There are also supermarkets available everywhere.

4. Results

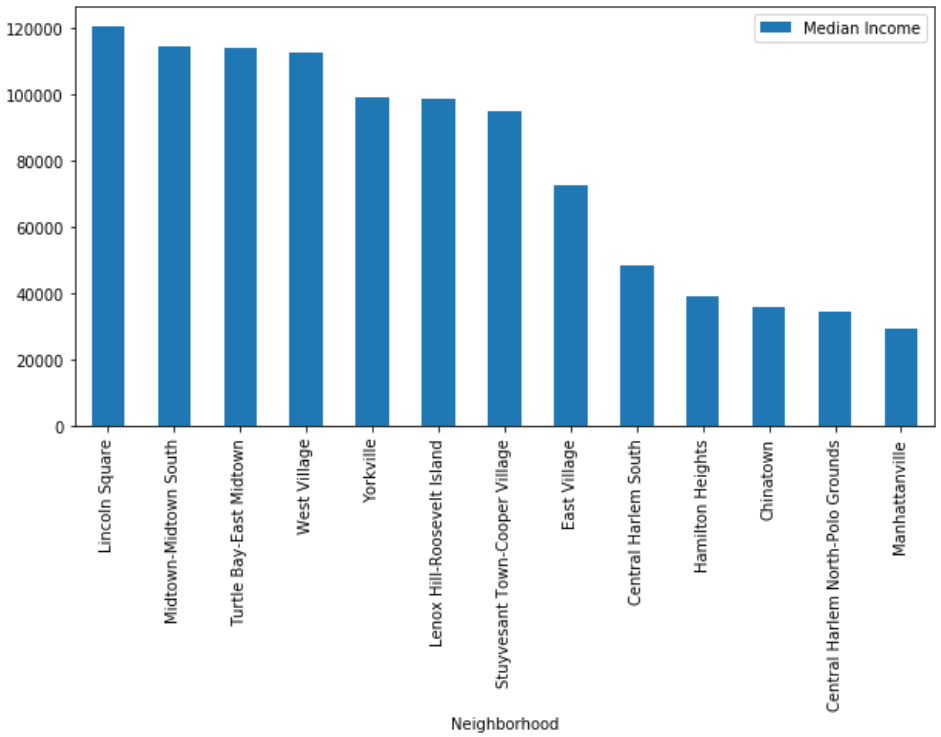
Manhattan was found to be the safest borough in New York City. The following visualizations were obtained:

a. The safest neighborhoods in Manhattan:

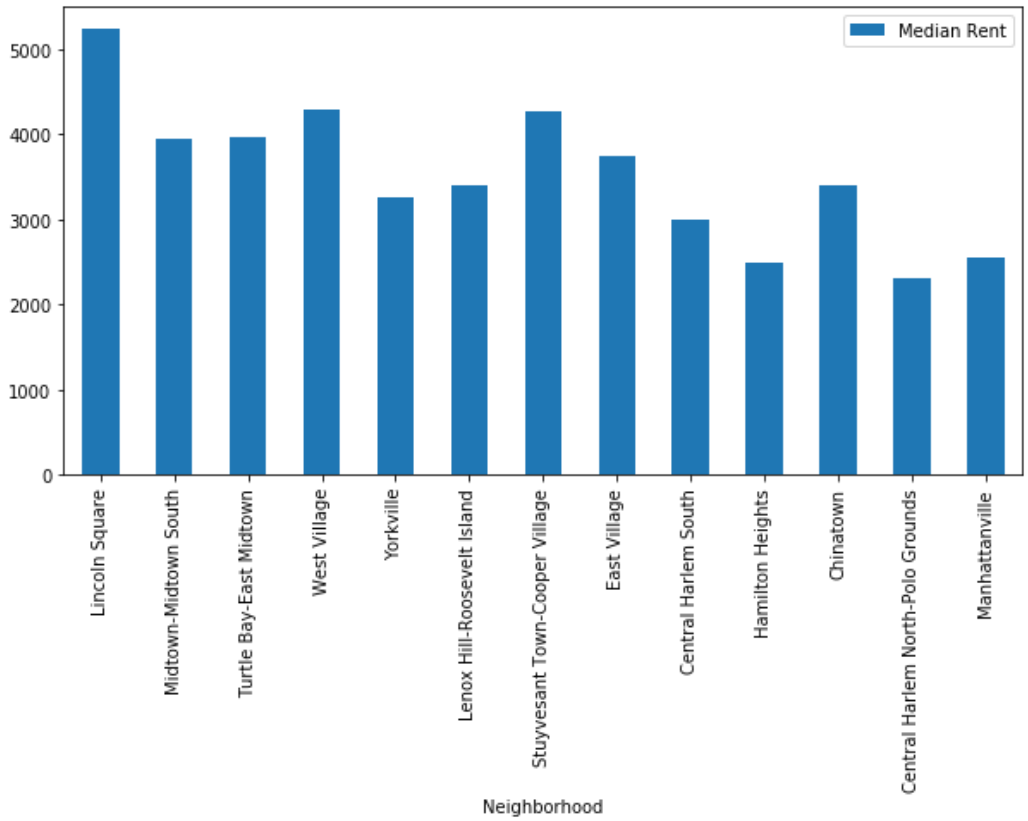
b. Socioeconomic conditions of these neighborhoods:



The range of values is: (83436,21744); Mean value = 53247.69



The range of values is: ($120337,$29182); Mean value=$ 77979.0

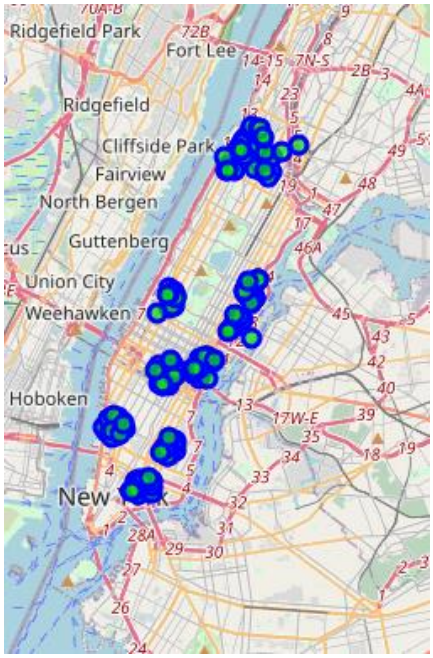


The range of values is: ($5232.5,$2300.0); Mean value=$ 3525.54

c. Visualization of venues and restaurants in these neighborhoods.

Venues

Restaurants



d. Clustered representation of neighborhoods as per restaurants:
Here, red= cluster 0(Chinatown); purple= cluster 1(Lincoln Square, West Village); blue= cluster 2(East Village, Manhattanville, Yorkville, Hamilton Heights)

5. Discussion

Part 1

The chosen locations are very safe till date because there has been no report of COVID-19 cases so far in these locations. Moreover, the effort taken by the New York City government is starting to bear fruit because the new cases have almost stopped popping up. With the same measures, New York City will be free from COVID-19, but that process is dependent on many other factors. It is more likely that normalcy will hit the neighborhoods that we have selected very soon and the economy will start to bounce back from these locations.

Part 2

Economic analysis of neighborhoods:

- Lincoln Square, West Village, Turtle Bay-East Midtown, are well-off neighborhoods in this list with a lot of venues in the city, and the standard of living is high. The population is above average too, so there is a decent scope to get good customers. On the contrary, Lenox Hill-Roosevelt Island and Yorkville, while having big population size and being above-average economically, have more affordable rent in comparison.
- Yorkville also has a lot of venues nearby. Lenox Hill-Roosevelt Island has a comparatively untapped market when compared to other neighborhoods.
- East Village is also financially feasible and has above average quality of living with a lot of venues, but the population is relatively low.
- Manhattanville is not so well-off in comparison and has a sparse population. The market is also quite dire.
- Stuyvesant Town-Cooper Village and Midtown-Midtown South have high standard of living and mostly well-off residents but the population is pretty low.
- Chinatown is not-so-well-off and its population size is about average. But there are a lot of venues.
- Central Harlem North-Polo Grounds has a high population but the residents are not as rich as other places and the standard of living is also the lowest.
- The remaining options are moderate.

Analysis of competition and restaurants/cuisine culture:

- Untapped markets: Lenox Hill-Roosevelt Island, Stuyvesant Town-Cooper Village, Central Harlem North-Polo Grounds
- High competition: Chinatown, East Village, Turtle Bay-East Midtown, Midtown-Midtown South, Yorkville
- Other neighborhoods have moderate competition.
- The most popular cuisine is Italian in these locations, followed by American, Japanese/Sushi, Mexican and Chinese. The competition in opening a restaurant for these cuisines will be very high.
- There is a scope to open more ethnic restaurants in these parts.

- Ideally, having some representation of popular cuisine's in the restaurant will be beneficial, especially Italian cuisines like pizza and spaghetti. It can serve to be a comfort food.
- New Indian, Vietnamese or Filipino cuisine restaurants can potentially compete better in this climate with their existing counterparts, since they are lower in number.
- East Village, Manhattanville, Yorkville, Hamilton Heights have similar tastes.

6. Conclusion

The analysis is performed on limited data and on the assumption that the status quo will remain stagnant and the government's plan of action will work the way it has from the latest date the data was obtained. Arbitrary factors could influence change in our findings. A better insight would have come from analyzing other places that have implemented the rules like New York City. Also, getting to know the precise value of the standard of living would have assisted in a better understanding of the neighborhoods economically, as well as knowing the precise population in 2020. The clusters could have been formed better with a bigger sample space of relevant data.

However, to a fair extent, this analysis is very valid as well and the neighborhoods hence recommended can be explored safely as per the stakeholder's needs.