Battle of the Neighbourhoods

New York City – Looking for the Best Place to Relocate.

Atharva Deshpande

July 2020

1. Introduction

1.1 Background

When someone or a family is trying to find the best places to live, it's always a good idea to compare cities and if possible, to compare neighborhoods to see if its suites your taste. Safety is a top concern when moving to a new area. If you don’t feel safe in your own home, you’re not going to be able to enjoy living there.

1.2 Problem

The crime statistics dataset of New York City found on data.world has crimes in each Boroughs of NYC in the year 2017. The crime rates in each borough may have changed over time. This project aims to select the safest boroughs in NYC based on the total crimes, explore the neighborhoods of that borough to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

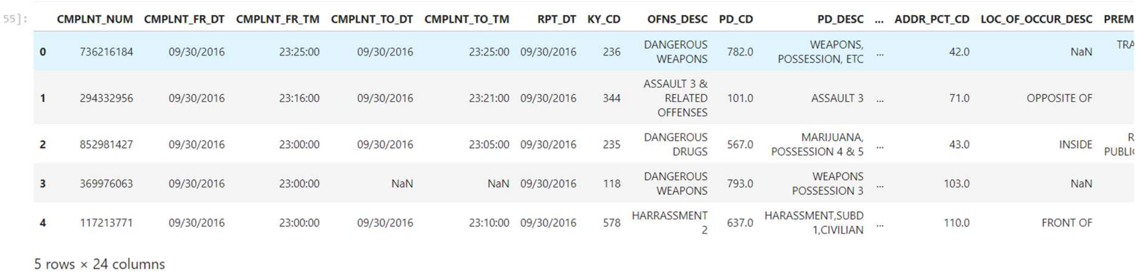
2. Data

There are two main datasets used in this project, one being the NYC crime dataset of the year 2017 and the other dataset contains all the neighbourhoods of NYC with their geographical coordinates.

2.1 Data Sets

2.1.1 New York crime data from 2017

Let’s take a look at this dataset



55]:

	CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	RPT_DT	KY_CD	OFNS_DESC	PD_CD	PD_DESC	...	ADDR_PCT_CD	LOC_OF_OCCUR_DESC	PREM
0	736216184	09/30/2016	23:25:00	09/30/2016	23:25:00	09/30/2016	236	DANGEROUS WEAPONS	782.0	WEAPONS, POSSESSION, ETC	...	42.0	NaN	TRA
1	294332956	09/30/2016	23:16:00	09/30/2016	23:21:00	09/30/2016	344	ASSAULT 3 & RELATED OFFENSES	101.0	ASSAULT 3	...	71.0	OPPOSITE OF	
2	852981427	09/30/2016	23:00:00	09/30/2016	23:05:00	09/30/2016	235	DANGEROUS DRUGS	567.0	MARIJUANA, POSSESSION 4 & 5	...	43.0	INSIDE	R PUBLI
3	369976063	09/30/2016	23:00:00	NaN	NaN	09/30/2016	118	DANGEROUS WEAPONS	793.0	WEAPONS POSSESSION 3	...	103.0	NaN	
4	117213771	09/30/2016	23:00:00	09/30/2016	23:10:00	09/30/2016	578	HARRASSMENT 2	637.0	HARASSMENT,SUBD 1,CIVILIAN	...	110.0	FRONT OF	

5 rows × 24 columns

Each row in this dataset represents a crime that was reported with its details.

Let’s looks at some of the important columns:

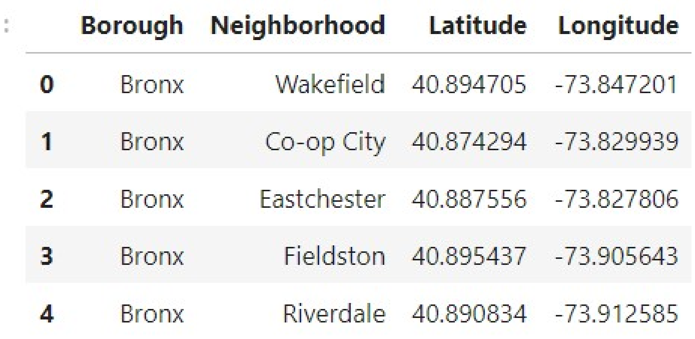
1. CMPLNT_NM: Randomly generated persistent ID for each complaint
2. ADDR_PCT_CD: The precinct in which the incident occurred
3. BORO_NM: The name of the borough in which the incident occurred
4. JURIS_DESC: Description of the jurisdiction code

20 more columns.

We are not interested in most of the columns, for our purpose we only need the BORO_NM column.

2.1.2 New York Neighbourhood’s Data

Let’s take a look at this dataset



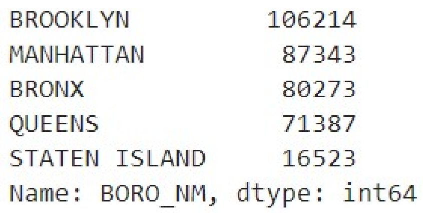
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

This dataset simply contains the list of neighbourhoods in NY with their respective Boroughs and their geographical coordinates

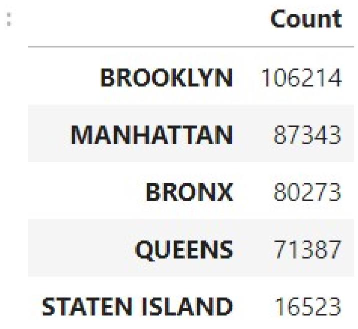
2.2 Data Cleaning

The NY neighbourhoods data set is already clean so we don’t need to perform any cleaning on it.

For the NY Crime dataset we need to simply get the total number of crimes in each borough. We do this by using *pandas value_count* function and save the result in a new dataframe.



```
BROOKLYN         106214
MANHATTAN         87343
BRONX             80273
QUEENS            71387
STATEN ISLAND     16523
Name: BORO_NM, dtype: int64
```



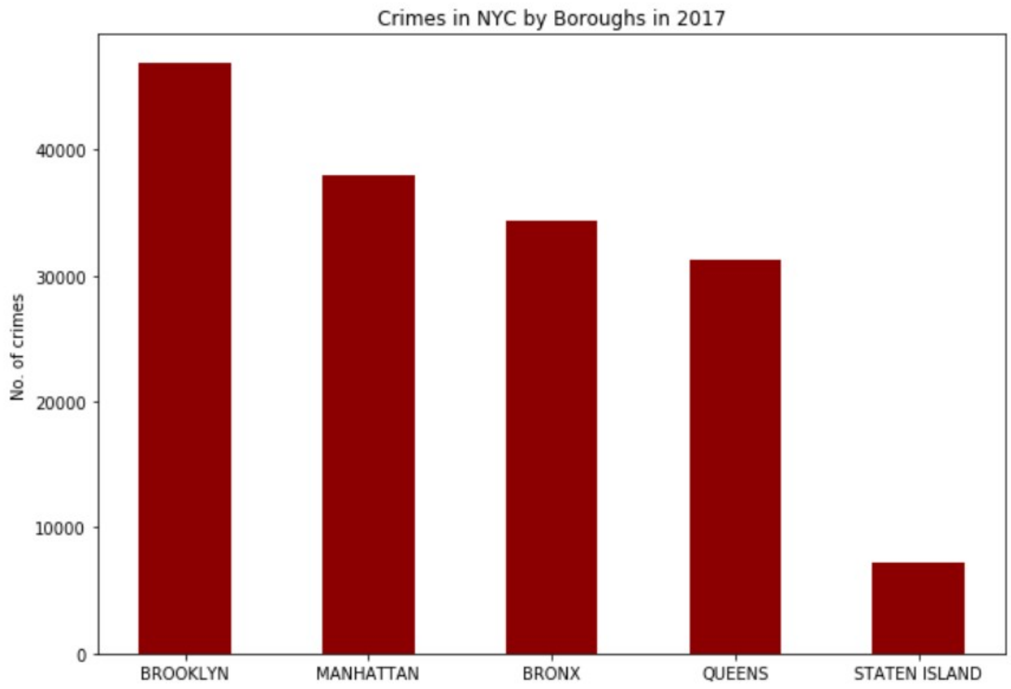
	Count
BROOKLYN	106214
MANHATTAN	87343
BRONX	80273
QUEENS	71387
STATEN ISLAND	16523

3. Methodology

3.1 Exploratory Data Analysis

3.1.1 Let's visualize the borough's total crime reports

Comparing five boroughs with the highest crime rate during the year 2016 it is evident that Brooklyn has the highest crimes recorded followed by Manhattan, Bronx, Queens and Staten Island.



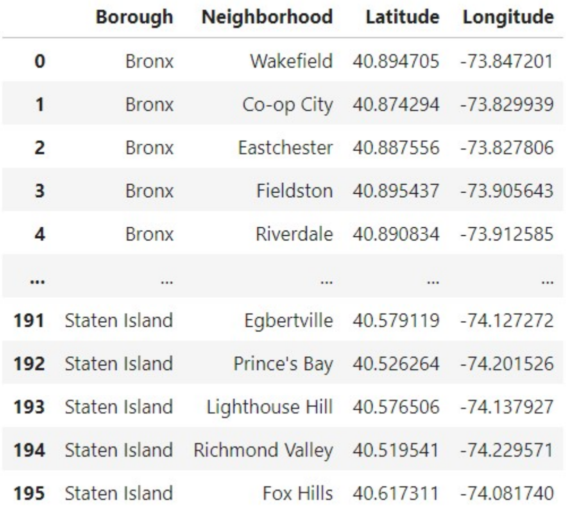
It's evident to see that Staten Island has the lowest crime rate for the year 2017, so it should be the safest borough and we should choose that , buts explore a bit more.

3.1.2 Neighborhoods in Staten Island

Using the *pandas shape* function on our converted df with only Staten Island Borough we realize that there are only 63 neighbourhoods in Staten Island. This is a rather small no of neighbourhoods to choose from. So instead of choosing from only Staten Island lets select the 3 Borough with least crime rate, that is Bronx, Queens and Staten Island.

3.1.3 Visualize the neighbourhoods

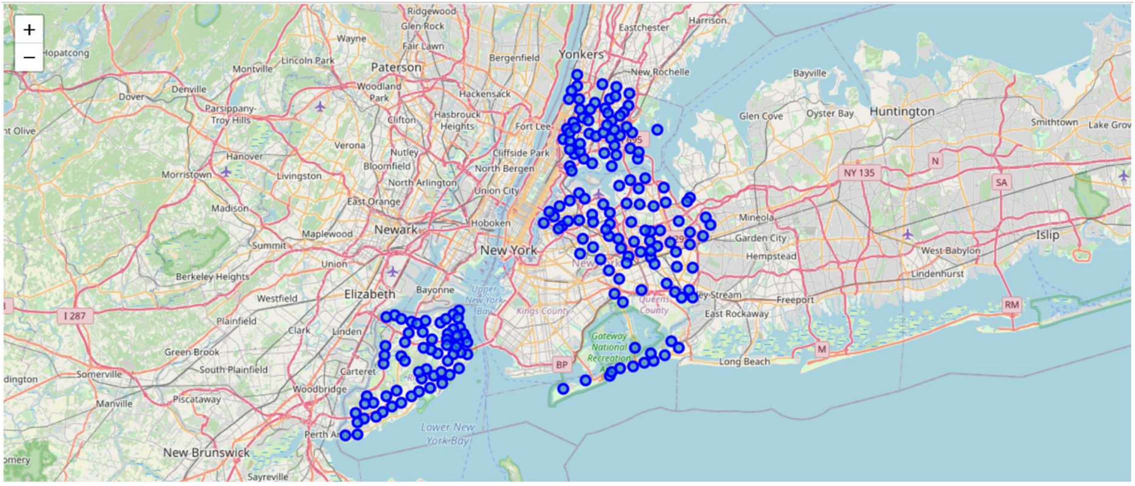
Firstly, we select the neighbourhoods only from these 3 Boroughs, that is Bronx, Queens and Staten Island, and save them in a new dataframe.



	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
...	...	...	...	...
191	Staten Island	Egbertville	40.579119	-74.127272
192	Staten Island	Prince's Bay	40.526264	-74.201526
193	Staten Island	Lighthouse Hill	40.576506	-74.137927
194	Staten Island	Richmond Valley	40.519541	-74.229571
195	Staten Island	Fox Hills	40.617311	-74.081740

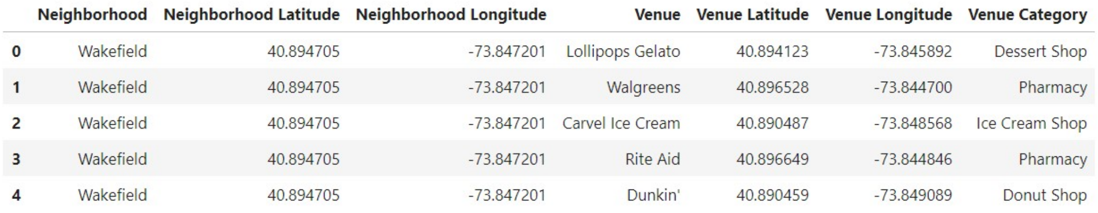
We see that there are 196 neighbourhoods across 3 broughs

The we use the *folium* library to visualize them



3.2 Modelling

Using the final dataset containing the neighbourhoods with the latitude and longitude, we can find all the venues within a 500 meter radius of each neighbourhood by connecting to the Foursquare API. This returns a json file containing all the venues in each neighbourhood which is converted to a pandas dataframe. This data frame contains all the venues along with their coordinates and category.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

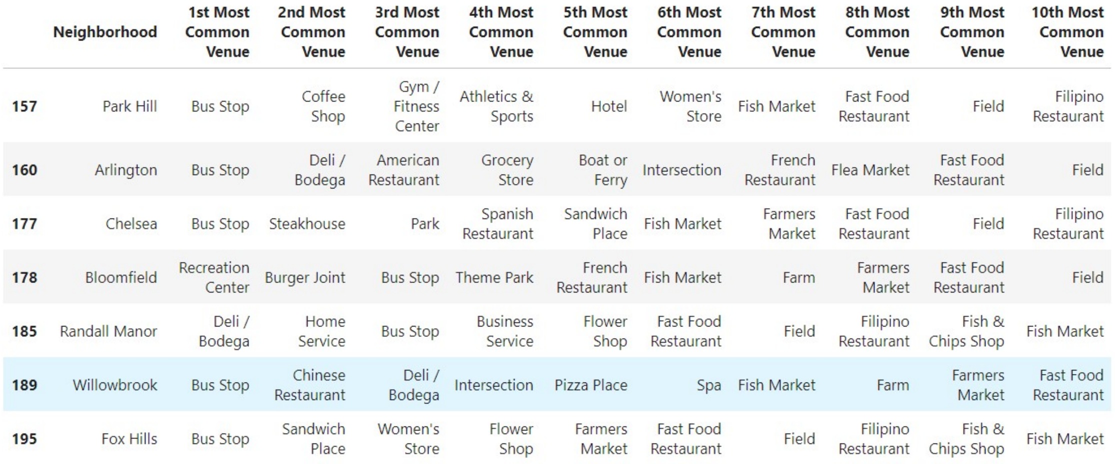
One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighbourhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighbourhoods.

To help people find similar neighbourhoods in the safest borough we will be clustering similar neighbourhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We used the elbow method to find the best cluster size and found 8 clusters to be ideal.

The reason to conduct a K- means clustering is to cluster neighbourhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighbourhood.

4. Results

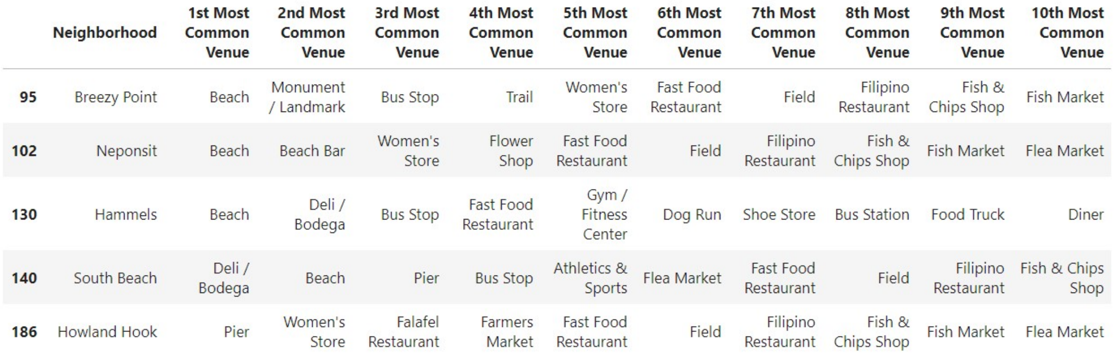
After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. Looking into the neighborhoods in the first cluster



	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
157	Park Hill	Bus Stop	Coffee Shop	Gym / Fitness Center	Athletics & Sports	Hotel	Women's Store	Fish Market	Fast Food Restaurant	Field	Filipino Restaurant
160	Arlington	Bus Stop	Deli / Bodega	American Restaurant	Grocery Store	Boat or Ferry	Intersection	French Restaurant	Flea Market	Fast Food Restaurant	Field
177	Chelsea	Bus Stop	Steakhouse	Park	Spanish Restaurant	Sandwich Place	Fish Market	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
178	Bloomfield	Recreation Center	Burger Joint	Bus Stop	Theme Park	French Restaurant	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Field
185	Randall Manor	Deli / Bodega	Home Service	Bus Stop	Business Service	Flower Shop	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop	Fish Market
189	Willowbrook	Bus Stop	Chinese Restaurant	Deli / Bodega	Intersection	Pizza Place	Spa	Fish Market	Farm	Farmers Market	Fast Food Restaurant
195	Fox Hills	Bus Stop	Sandwich Place	Women's Store	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop	Fish Market

Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Bus Stop, Coffee shops and restaurants.

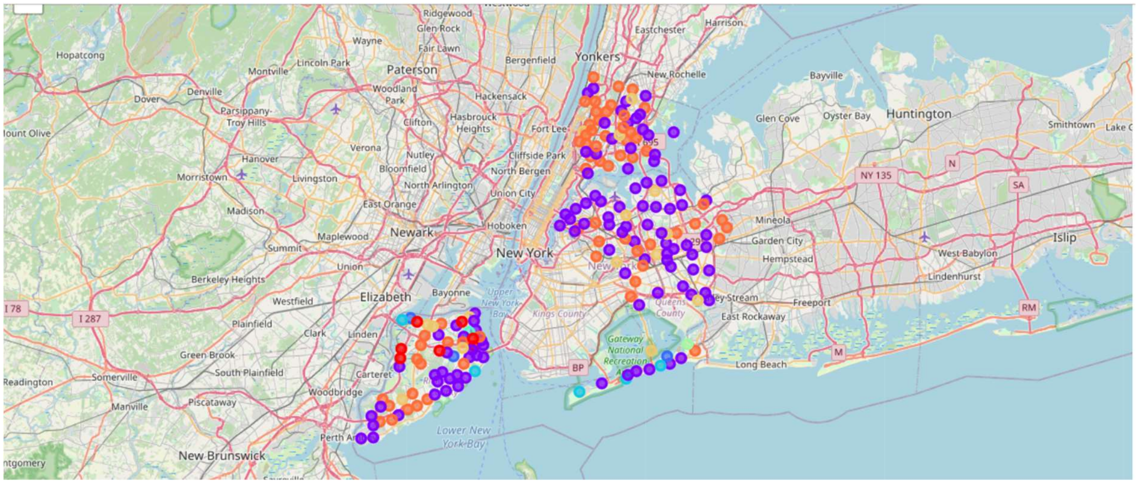
Similarly looking at the 4th cluster we see that it mainly consists of the neighbourhoods with venues such as beach



	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
95	Breezy Point	Beach	Monument / Landmark	Bus Stop	Trail	Women's Store	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop	Fish Market
102	Neponsit	Beach	Beach Bar	Women's Store	Flower Shop	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop	Fish Market	Flea Market
130	Hammels	Beach	Deli / Bodega	Bus Stop	Fast Food Restaurant	Gym / Fitness Center	Dog Run	Shoe Store	Bus Station	Food Truck	Diner
140	South Beach	Deli / Bodega	Beach	Pier	Bus Stop	Athletics & Sports	Flea Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop
186	Howland Hook	Pier	Women's Store	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant	Fish & Chips Shop	Fish Market	Flea Market

Similarly, we can examine each cluster to find out which neighbourhoods suits our best interest by looking at the most common venues.

Finally lets visualize the clustered neighbourhoods using Folium.



5. Discussion

The aim of this project is to help people who want to relocate to the safest borough in New York city, expats can choose the neighbourhoods to which they want to relocate based on the most common venues in it. For example, if a person is looking for a neighbourhood with good connectivity and public transportation we can see that Cluster 1 has and Bus stops as the most common venues. If a person is looking for a neighbourhood with stores and restaurants in a close proximity, then the neighbourhoods in the second cluster is suitable. The choices of neighbourhoods may vary from person to person.

6.Conclusion

This project helps a person get a better understanding of the neighbourhoods with respect to the most common venues in that neighbourhood. It is always helpful to find out more about places before moving into a neighbourhood. We have just taken safety as a primary concern to shortlist the safest boroughs in New York city. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough, such as filtering areas based on a predefined budget.