Battle of Boroughs and Neighborhoods – Report
Osman Muglu

1. Introduction

1.1 Background

New York City (NYC) is one of the largest and wealthiest metropolises in the world. It has over 8 million people and is the most populous city in the United States of America. NYC can be described as the cultural, financial, and media capital of the world. Partly for this reason, people choose to move here. However, before making a move to NYC it is necessary to know in which neighborhood you want to live. Otherwise, the choice that is made is a pure gamble. In the worst case, if you are not happy with the neighborhood you just moved to, you will have to move again [1].

To prevent this or minimize the chance of a worst-case scenario, it is necessary to do some research. NYC is divided into 5 boroughs; *the Bronx, Brooklyn, Manhattan, Queens, and Staten Island* [1]. And each borough has thousands of neighborhoods. So, how is someone supposed to know which neighborhood to choose to live in?

1.2 Problem

There are a lot of things people consider before buying or renting a new home. The first thing that most people consider is the distance to work and the supermarket. However, based on human instincts the first thing we need to consider when moving to a new area is safety. It is important that when buying a new home to know that the neighborhood is safe. After safety, the second aspect to consider is venues. For human beings to take part in and form a strong and cohesive community, both safety and venues are key aspects. Within this research, we look at the safety of boroughs and the social activities within neighborhoods.

We have found a crime dataset of NYC on Data.World, it has crimes of each borough of NYC in 2016 [2]. The reason that a dataset from 2016 was chosen is due to the scarcity of publicly open data by the NYC government. This research aims to select the safest borough in NYC on total crimes. Moreover, this project explores the neighborhoods of the safest borough based on the most common venues.

Based on the problem description, we have formulated the following two questions:

1. *(1) What is the safest borough in NYC?*
2. *(2) What is the best neighborhood based on venues?*

1.3 Interest

People who are considering moving to NYC will be interested to identify the safest borough and the best neighborhood based on venues.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

The data acquired for this project is a combination of data from different resources.

- The crime data set of this project, is downloaded from Data.World and it contains the crimes per borough in NYC [2]
- The list of Neighborhoods of Staten Island was found on a wikipedia: https://en.wikipedia.org/wiki/List_of_Staten_Island_neighborhoods [3]
- I used Foursquare API to get the most common venues of given borough of NYC [4]

2.2 Data Cleaning and feature selection

2.2.1 NYC Crime dataset

The data preparation for each of the data sources have been done separately. Beginning from the NYC crime data set, only the crimes description and borough description columns were selected. All the other columns within the data set are dropped. Moreover, we counted the values and pivoted the table to see the number of each crime category for each borough. At last, we created a column that counts the total amount of crime for each borough (see gif 1.1).

Crime Category	Borough	ADMINISTRATIVE CODE	ADMINISTRATIVE CODES	AGRICULTURE & MRKTS LAW-UNCLASSIFIED	ALCOHOLIC BEVERAGE CONTROL LAW	ANTICIPATORY OFFENSES	ARSON	ASSAULT 3 & RELATED OFFENSES	BURGLAR'S TOOLS	BURGLARY	...	PROSTITUTION & RELATED OFFENSES	RAPE	ROBBERY	SEX CRIMES	THEFT OF SERVICES	THEFT-FRAUD	UNAUTHORIZED USE OF A VEHICLE	UNLAWFUL POSS. WEAP. ON SCHOOL	VEHICLE AND TRAFFIC LAWS	Total
0	BRONX	170	0	22	32	0	170	10300	27	1947	...	68	265	3149	938	144	644	142	0	1030	80267
1	BROOKLYN	312	0	31	34	0	163	12171	52	3099	...	19	337	3603	1416	9	972	470	1	1311	106202
2	MANHATTAN	64	0	28	7	2	109	7903	79	1824	...	7	220	2185	1376	160	1110	129	0	815	87339
3	QUEENS	176	1	21	2	0	135	7942	45	2401	...	5	249	2237	868	14	780	512	1	1661	71373
4	STATEN ISLAND	37	0	6	0	0	36	1661	2	364	...	3	36	298	141	0	197	85	0	133	16521

Fig 1.1 NYC crime data after prepossessing

2.2.2 NYC safest borough dataset

After the visualization of the crime in each borough, we can find the safest borough. The second source of data is acquired from a list that is available on Wikipedia. Using the list with data from the Wikipedia page we created a new dataset from scratch. The frame is created using the pandas library. The data set contains the names of each neighborhood, borough name and the latitude and longitude coordinates (see fig 1.2).

	Neighborhood	Borough	Latitude	Longitude
0	Annadale	Staten Island		
1	Arden Height	Staten Island		
2	Arlington	Staten Island		
3	Arrochar	Staten Island		
4	Bay Terrace	Staten Island		
...	...	...	...	...
62	Ward Hill	Staten Island		
63	West New Brighton	Staten Island		
64	Westerleigh	Staten Island		
65	Willowbrook	Staten Island		
66	Woodrow	Staten Island		

67 rows × 4 columns

Fig 1.2 NYC dataset of safest borough

The coordinates of the neighborhoods is obtained using Google Maps API geocoding [5]. Even after using the API, we ended up with some missing values. After dropping these values we obtained a ready to use dataset for our research (see fig 1.3).

	Neighborhood	Borough	Latitude	Longitude
0	Annadale	Staten Island	40.544550	-74.176532
2	Arlington	Staten Island	40.632326	-74.165144
3	Arrochar	Staten Island	40.598438	-74.072641
4	Bay Terrace	Staten Island	40.555278	-74.134167
5	Bloomfield	Staten Island	40.612604	-74.178200

Fig 1.3 NYC dataset of safest borough

The column descriptions:

(1) *Neighborhood* **:** Name of the neighborhood in the Borough.
(2) *Borough :* Name of the Borough.
(3) *Latitude :* Latitude of the Borough
(4) *Longitude :* Longitude of the borough

2.2.3 Venue dataset- Foursquare API Data

The next step is gathering data about different venues of the safest neighborhood. To retrieve the data we will use Foursquare. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API. All the gathered data will be sufficient to build the last model. Based on the data we will cluster all the neighborhoods. This clustering happens based on the similar venue categories. After following these steps the dataset is ready to use (see fig 1.4).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Annadale	40.544550	-74.176532	Annadale Diner	40.542079	-74.177325	Diner
1	Annadale	40.544550	-74.176532	Annadale Terrace	40.542555	-74.177187	Restaurant
2	Annadale	40.544550	-74.176532	Il Sogno	40.541286	-74.178489	Restaurant
3	Annadale	40.544550	-74.176532	MTA SIR - Annadale	40.540482	-74.178185	Train Station
4	Annadale	40.544550	-74.176532	MTA Bus - Annadale Rd & Arden Av (S55)	40.544326	-74.176667	Bus Stop
...	...	...	...	...	...	...	...
762	Woodrow	40.543439	-74.197644	Rise Dance Studios	40.542273	-74.197171	Music Venue
763	Woodrow	40.543439	-74.197644	New Dorp Nails III	40.543419	-74.198853	Cosmetics Shop
764	Woodrow	40.543439	-74.197644	MTA Bus - Woodrow Rd & Vineland Av (S56/SIM2/S...	40.543364	-74.197213	Bus Stop
765	Woodrow	40.543439	-74.197644	Espo's Bar	40.542883	-74.201800	Bar
766	Woodrow	40.543439	-74.197644	Citi Taco	40.541220	-74.194345	Mexican Restaurant

Fig 1.4 Venue dataset

The column descriptions:

(1) *Neighborhood :* Name of the Neighborhood
(2) *Neighborhood Latitude* : Latitude of the Neighborhood
(3) *Neighborhood Longitude* : Longitude of the Neighborhood
(4) *Venue* : Name of the Venue
(5) *Venue Latitude* : Latitude of Venue
(6) *Venue Longitude* : Longitude of Venue
(7) *Venue Category* : Category of Venue

3. Methodology

3.1 Required packages

We will be creating our model with the help of Python so we start of by importing all the required packages.

```
import pandas as pd
import requests
import numpy as np
import random
%matplotlib inline
import matplotlib as mpl
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors
import folium
```

```
from sklearn.cluster import KMeans
import geocoder
from geopy.geocoders import Nominatim
from IPython.core.display import HTML
from pandas.io.json import json_normalize
```

Package breakdown:

- requests : Handle http requests
- pandas : To collect and manipulate data in JSON and HTMl and then data Analysis
- matplotlib : Detailing the generated maps
- numpy: to handle arrays
- folium : Generating maps of London and Paris
- sklearn : To import Kmeans which is the machine learning model that we are using.

3.2 Visualizing the neighborhoods

I used python folium library package to visualize geographic details of the Staten Island borough and its neighborhoods on top. I used latitude and longitude values to get the visual as seen in the figure below:

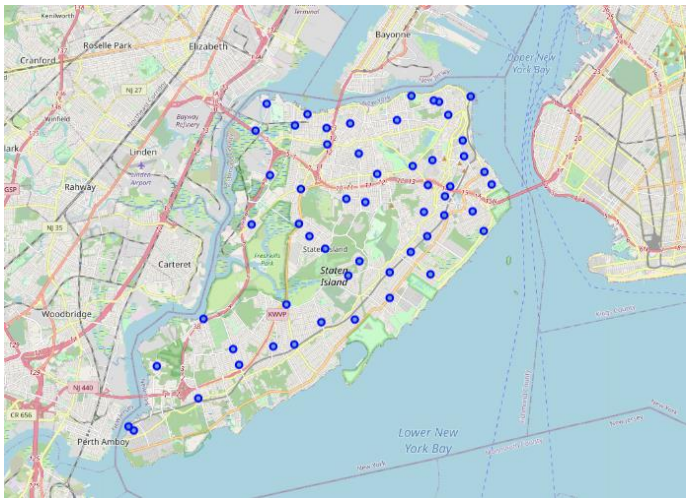

Fig 2.1 map of Staten Island

3.3 Foursquare API

For the exploration of the neighborhoods and to segment we use the foursquare API. The limit is set on 100 venues with a radius of 500 meter for each neighborhood from their given latitude and longitude.

3.4 One hot coding

After obtaining the data with the foursquare API, we use one hot coding to convert categorical data into numeric 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 Neighborhood and the mean of the venues are calculated; finally the 10 common venues are calculated for each of the neighborhoods (see fig. 2.2).

	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	Annadale	Pizza Place	American Restaurant	Business Service	Restaurant	Food	Bus Stop	Bar	Bakery	Diner	Deli / Bodega
1	Arlington	Bus Stop	American Restaurant	Deli / Bodega	Construction & Landscaping	Cultural Center	Donut Shop	Flower Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market
2	Arrochar	Bus Stop	Park	Deli / Bodega	Nail Salon	Bagel Shop	Cosmetics Shop	Pizza Place	Donut Shop	Fast Food Restaurant	Farmers Market
3	Bay Terrace	Food Truck	Playground	American Restaurant	Train Station	Performing Arts Venue	Discount Store	Fast Food Restaurant	Farmers Market	Event Space	Event Service
4	Bloomfield	Hotel	Candy Store	Tea Room	Rental Car Location	Italian Restaurant	Department Store	Spa	Cocktail Bar	Video Store	Discount Store

Fig 2.2 Table top venues

3.5 Model building – K - means

To find similar neighborhoods we will be using K-means clustering. This is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster the all the neighborhoods into 5 clusters. The reason behind using K-means clustering is that, after clustering neighborhoods with similar venues together, people can shortlist the area of their interest based on venues.

4.1 Results - Data Analysis

4.1.1 Statistical summary of crimes

In order to obtain a statistical summary of the NYC crime data we used the describe function of python. This returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime (See fig 3.1).

e Category	ADMINISTRATIVE CODE	ADMINISTRATIVE CODES	AGRICULTURE & MRKTS LAW-UNCLASSIFIED	ALCOHOLIC BEVERAGE CONTROL LAW	ANTICIPATORY OFFENSES	ARSON	ASSAULT 3 & RELATED OFFENSES	BURGLAR'S TOOLS	BURGLARY	CHILD ABANDONMENT/NON SUPPORT	...	PROSTITUTION & RELATED OFFENSES	RAPE	ROBBERY	SEX CRIMES	THEFT OF SERVICES	THEFT-FRAUD	UNAUTHORIZED USE OF A VEHICLE	UNLAWFUL POSS. WEAP. ON SCHOOL	VEHICLE AND TRAFFIC LAWS	Total
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	...	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.00000
mean	151.800000	0.200000	21.600000	15.000000	0.400000	122.600000	7995.400000	41.000000	1927.000000	3.200000	...	20.400000	221.400000	2294.400000	947.800000	65.400000	740.600000	267.600000	0.400000	990.00000	72340.40000
std	108.936679	0.447214	9.659193	16.643317	0.894427	54.123008	3964.934716	28.714108	1006.508569	2.167948	...	27.327642	112.251058	1269.217397	514.771017	79.415364	352.273757	205.563372	0.547723	574.44669	33731.06218
min	37.000000	0.000000	6.000000	0.000000	0.000000	36.000000	1661.000000	2.000000	364.000000	0.000000	...	3.000000	36.000000	298.000000	141.000000	0.000000	197.000000	85.000000	0.000000	133.00000	16521.00000
25%	64.000000	0.000000	21.000000	2.000000	0.000000	109.000000	7903.000000	27.000000	1824.000000	3.000000	...	5.000000	220.000000	2185.000000	868.000000	9.000000	644.000000	129.000000	0.000000	815.00000	71373.00000
50%	170.000000	0.000000	22.000000	7.000000	0.000000	135.000000	7942.000000	45.000000	1947.000000	3.000000	...	7.000000	249.000000	2237.000000	938.000000	14.000000	780.000000	142.000000	0.000000	1030.00000	80267.00000
75%	176.000000	0.000000	28.000000	32.000000	0.000000	163.000000	10300.000000	52.000000	2401.000000	4.000000	...	19.000000	265.000000	3149.000000	1376.000000	144.000000	972.000000	470.000000	1.000000	1311.00000	87339.00000
max	312.000000	1.000000	31.000000	34.000000	2.000000	170.000000	12171.000000	79.000000	3099.000000	6.000000	...	68.000000	337.000000	3603.000000	1416.000000	160.000000	1110.000000	512.000000	1.000000	1661.00000	106202.00000

Fig 3.1 Data summary

4.1.2 Boroughs with the highest crime rates

When we compare the five boroughs we see which one has the highest crime rate during 2016. Brooklyn has the highest crimes recorded followed by Manhattan, Bronx, Queens and Staten Island. Brooklyn has a significantly higher crime rate than the other 4 boroughs (see fig 3.2).

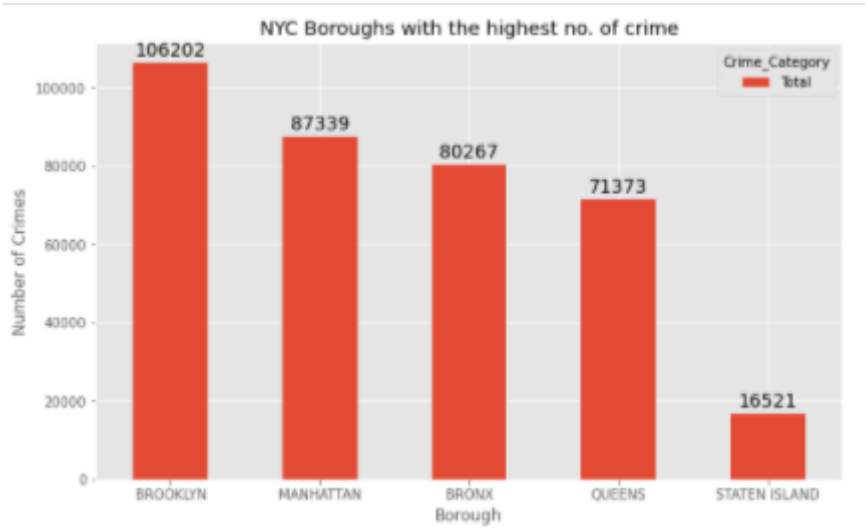

Fig 3.2 Highest crime rate

4.1.3 Boroughs with the lowest crime rates

We already could the borough with the lowest crime rates in the previous figure. However, we visualized a new chart but in a new order. In this new figure we see that Staten Island borough has the lowest crime rates (see fig 2.3).

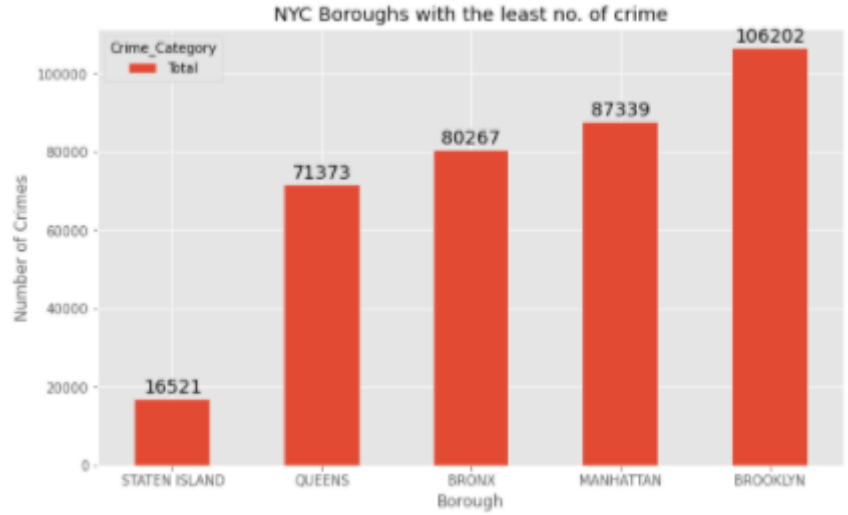

Fig 3.3 Lowest crime rate

4.1.4 Clustering

After running the K-means clustering we can access each cluster created. Here we can see which neighborhoods are assigned to which of the five clusters.

The **first cluster** has 5 neighborhoods in it and these are: Chelsea, Port Richmond, Todt Hill, West New Brighton and Woodrow. By looking at the data we see that the most common venues of these clusters are: bus stops, supermarkets and restaurant serving food from different cultural cuisines.

The **second cluster** in our research has 12 neighborhoods. What is immediately noticeable about this cluster is that, almost all neighborhoods have ranked bus stops as number one most common venue. After, bus stops as most common venues comes: bodega's,

The **third cluster** is the biggest in our study. This cluster has 36 neighborhoods in it. By closely examining this cluster we see that the most common venues in these neighborhoods are: pizza places, cafes, restaurants and parks.

The **fourth** and the **fifth clusters** contain each one neighborhood. The fourth cluster contains the neighborhood; Westerleigh and the top three most common venues are: arcade, yoga studio and discount store. The fifth cluster contains the "Old place" neighborhoods with the next top three common venues: harbor, toll plaza and Yoga studio.

The next image is a visualization of the clusters on the Staten Island map (see fig 3.4). Each cluster has it unique color for the ease of the presentation. We see that the majority of the neighborhoods fall is the blue cluster, which is the third cluster. The colors of the other clusters are as follows: the first cluster has a red color, the second is purple, the fourth cluster is green and the last cluster is yellow.

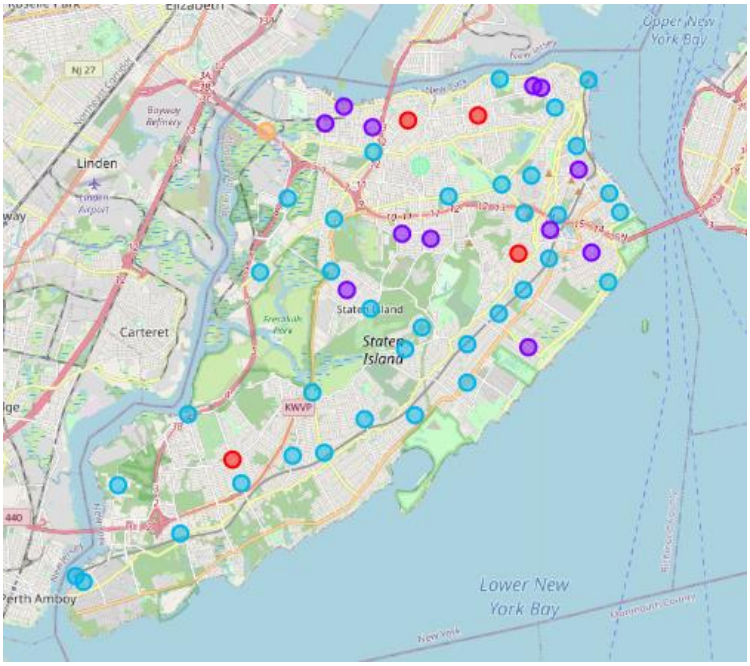

Fig 3.4 Clustered neighborhoods Staten Island

5. Discussion

As we mentioned before in this study, NYC is a very attractive place for people to move to. However, before making a move to NYC it is necessary to know in which neighborhood you want to live. This brings us to the two research questions of this study. These were: “What is the safest borough in NYC?” and “What is the best neighborhood based on venues?”.

In our results we clearly see the safest neighborhood. With the least number of crimes in 2016, the answer is “Staten Island”. In comparison to the other boroughs Staten Island has 5-6 times lesser number of crimes. Selecting a right neighborhood is mostly a subjective choice of someone. However, this study is providing information to help people make that choice based on venues. For example if a person is looking for a location with good connectivity and public transport, the recommended neighborhoods will be in the second cluster. If people are looking for a new home within a neighborhood that offers lot of restaurant and other places were food is served, the third cluster is where they need to look at. The first cluster is recommended for families who are looking for quieter places, but who want to live close to the supermarkets.

6. Conclusion

People want to move to cities big cities in order to make a new living. This project helps people find out which borough in NYC is the safest. It also gives a picture of the neighborhoods, based on venues, within the safest borough. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighborhood. Future projects can focus on adding more data to this research. By example including factors such as average house prices or cost of living.

7. References

[1] NYC Wikipedia

[2] NYC Crime data set 2016

[3] Neighborhoods of Staten Island

[4] Foursquare API

[5] Google Maps API geocoding