Capstone Project - The Battle of Neighborhoods

Safety in New York City

Report

Patrícia Falcão

2020, February

Table of Contents

Introduction

Accordingly, to an article from 2019, the average American moves once every 5 years. New York City (NYC) is the most populous city in united states, however is not the safest one. Nonetheless, NYC is a very popular city among migrants and imigrants due to the job opportunities that exist here. In turn, if the job opportunity is important at an early stage, security becomes important when families start having children. Thus, this characteristic become one of the most important when choosing a place to live. As NYC is a big city with so many opportunities, it is useful to know which neighbourhoods are the safest to live in when looking for a home.

This way, this project aims to find the safest borough in New York City based on the crimes committed between 2014 and 2015. Also, it is set as goal to select the 5 most common venues (of the safest borough) in each neighbourhood and cluster them using k-mean clustering.

Data

For this project it was used three different data sources. By using a dataset from Kaggle, it was possible to access to the crimes committed in NYC between 2014 and 2015. In this dataset it was possible to know the date of the crimes ('CMPLNT_FR_DT' renamed 'year'), the type ('OFNS_DESC' renamed 'crimes'), the borough where it happens ('BORO_NM' renamed 'Borough') and the latitude ('Latitude') and longitude ('Longitude').

Second, it was used a Wikipedia page that had the list of NYC boroughs. This table contains some columns: the borough of NYC, the county, the estimate population in 2017, the gross domestic product in billions and per capita, the land area in square miles and square km and the density.

As soon as the more safety borough in NYC was selected, it was used a new Wikipedia page referred to the list of neighbourhoods of the borough. In this case, the borough was Staten Island. Using the names of the neighbourhoods, it was created a new dataframe containing the latitude and longitude of each neighbourhood.

Methodology

- ## Data Cleaning

  The first step when dealing with semi-raw data, it is data cleaning. This way, in the first data source (NYC crime data), it was only considered the crimes from 2015 (most recent). Next, in order to know the total crimes of each type per borough, it was created a crosstab (Figure 1):

crimes	ABORTION	ADMINISTRATIVE CODE	ADMINISTRATIVE CODES	AGRICULTURE & MRKTS LAW-UNCLASSIFIED	ALCOHOLIC BEVERAGE CONTROL LAW	ANTICIPATORY OFFENSES	ARSON	ASSAULT 3 & RELATED OFFENSES	BURGLAR'S TOOLS	BURGLARY	...	RAPE	ROBBERY	SEX CRIMES		Total
Borough																
BRONX	0	212	0	29	35	1	299	13224	33	2662	...	297	4368	999		102712
BROOKLYN	0	404	1	19	20	0	322	16019	71	5491	...	350	5686	1420		140085
MANHATTAN	0	112	0	16	21	0	195	9993	112	2743	...	248	3143	1474	...	110196
QUEENS	2	291	1	13	2	1	146	10507	64	3544	...	235	3260	972		92732
STATEN ISLAND	0	34	0	6	0	0	46	2189	6	560	...	34	456	165		21678

Figure 1 - Part of the table after preparation of the NYC crime data.

Next, using the second data source (the Wikipedia page with the boroughs of NYC), it was used the Beautiful Soup library in order to extract the data from the web page. Next, the borough names in both datasets are checked and then tables are merged (on the borough names), combining all the information in only one dataset (Figure 2). This way it is possible to know the borough with the least and highest crimes recorded in 2015.

	Borough	ABORTION	ADMINISTRATIVE CODE	ADMINISTRATIVE CODES	AGRICULTURE & MRKTS LAW-UNCLASSIFIED	ALCOHOLIC BEVERAGE CONTROL LAW	ANTICIPATORY OFFENSES	ARSON	ASSAULT 3 & RELATED OFFENSES	BURGLAR'S TOOLS	...	VEHICLE AND TRAFFIC LAWS	Total	County	Estimate (2017)[3]
0	BRONX	0	212	0	29	35	1	299	13224	33	...	1201	102712	Bronx	1471160.0
1	BROOKLYN	0	404	1	19	20	0	322	16019	71	...	1716	140085	Kings	2648771.0
2	MANHATTAN	0	112	0	16	21	0	195	9993	112	...	1063	110196	New York	1664727.0
3	QUEENS	2	291	1	13	2	1	146	10507	64	...	2105	92732	Queens	2358582.0
4	STATEN ISLAND	0	34	0	6	0	0	46	2189	6	...	174	21678	Richmond	479458.0

Figure 2 - Part of the merged table.

By analysing the last figure, it is possible to recognise the borough with the lowest crime rate, the safest one. Knowing this, it used the last source (Wikipedia page mentioning the neighbourhoods of the safest borough). It is then created a new dataframe using the name of each neighbourhood. Then, the coordinates of each latitude and longitude are obtained using the Google Maps API Geocoding, resulting in Figure 3. However, since it was not possible to use this google tool to locate all the neighbourhoods (it was a big number), it was only considered 26.

	Neighborhood	Borough	Latitude	Longitude
0	Annadale	STATEN ISLAND	40.544550	-74.176532
1	Arlington	STATEN ISLAND	41.694069	-73.886363
2	Arrochar	STATEN ISLAND	40.598438	-74.072641
3	Bay Terrace	STATEN ISLAND	40.555278	-74.134167
4	Bloomfield	STATEN ISLAND	42.895064	-77.434713
5	Bulls Head	STATEN ISLAND	40.607048	-74.162088
6	Castleton Corners	STATEN ISLAND	40.613160	-74.122365
7	Charleston	STATEN ISLAND	40.536772	-74.237367
8	Chelsea	STATEN ISLAND	40.746491	-74.001528
9	Clifton	STATEN ISLAND	40.620104	-74.077086
10	Concord	STATEN ISLAND	42.535339	-78.730862

Figure 3 - Part of the table that represents the neighbourhoods of Staten Island (the safest borough).

- Exploratory Data Analysis

After cleaning and have a basis to work on, it is important to explore the data. In a first step, applying the describe function, it is possible to know some statistics of the NYC crime data, as shown in figure 4. As it can be denoted, in this table, it is also possible to denote what is the highest ('Petit Larceny') and the lowest ('Fortune Telling', 'Kidnapping', 'Kidnapping and related offenses', 'Loitering', 'Under the influence of drugs') reported

crime. However, since there is a large numerous type of crimes, it can't be showed in this report the whole table.

crimes	ABORTION	ADMINISTRATIVE CODE	ADMINISTRATIVE CODES	AGRICULTURE & MRKTS LAW-UNCLASSIFIED	ALCOHOLIC BEVERAGE CONTROL LAW	ANTICIPATORY OFFENSES	ARSON	ASSAULT 3 & RELATED OFFENSES	BURGLAR'S TOOLS	BURGLARY	...	Total
count	5.000000	5.00000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	...	5.000000
mean	0.400000	210.60000	0.400000	16.600000	15.600000	0.400000	201.600000	10386.400000	57.200000	3000.000000	...	93480.600000
std	0.894427	145.58434	0.547723	8.443933	14.604794	0.547723	113.279742	5175.539373	40.145984	1777.280929	...	43854.210776
min	0.000000	34.00000	0.000000	6.000000	0.000000	0.000000	46.000000	2189.000000	6.000000	560.000000	...	21678.000000
25%	0.000000	112.00000	0.000000	13.000000	2.000000	0.000000	146.000000	9993.000000	33.000000	2662.000000	...	92732.000000
50%	0.000000	212.00000	0.000000	16.000000	20.000000	0.000000	195.000000	10507.000000	64.000000	2743.000000	...	102712.000000
75%	0.000000	291.00000	1.000000	19.000000	21.000000	1.000000	299.000000	13224.000000	71.000000	3544.000000	...	110196.000000
max	2.000000	404.00000	1.000000	29.000000	35.000000	1.000000	322.000000	16019.000000	112.000000	5491.000000	...	140085.000000

Figure 4 - Exploratory Analysis of NYC Crime data.

Next, by analysing figure 5 it is possible to conclude that Staten Island is the borough of NYC with the lowest crime rate in 2015, with 21678 crimes.

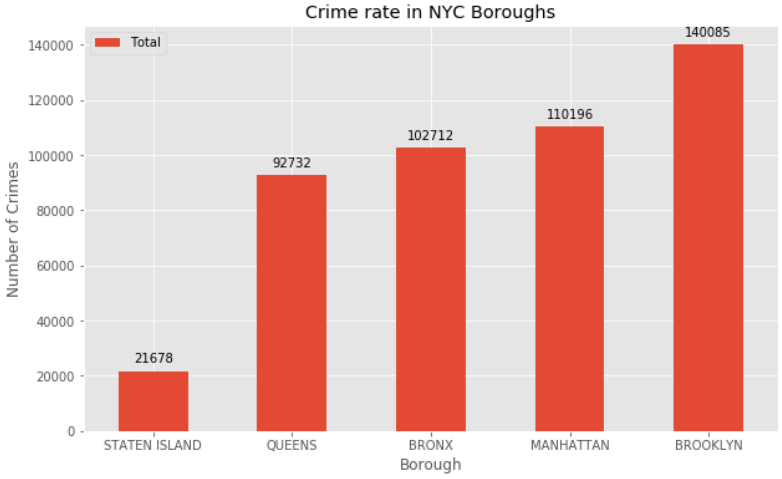

Figure 5 - Crime rate in every borough of NYC.

Also, in order to get an idea of the distribution of some of the crimes in Staten Island, it is plotted 8 types of crimes committed in Staten Island (Figure 6).

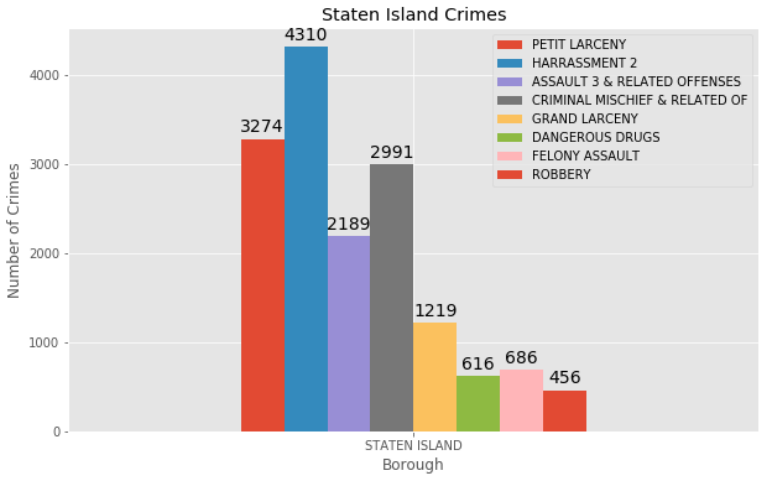

Figure 6 - Distribution of some of the crimes comitted in Staten Island in 2015.

Next, it is used a Wikipedia page referring to the neighbourhoods of Staten Island. However, since it was not possible to geolocate every neighbourhood with Google Maps

API Geocoding, it was considered only 26 of them. It is represented in the next map (Figure 7).

Figure 7 - Geographical positions of some of the neighbourhoods of Staten Island.

- ## Modelling

Finally, using Foursquare API it is possible to find venues 500m away from each neighbourhood of Staten Island considered. Then, the resulting json file is converted to a dataframe, containing all the venues with the coordinates and category (Figure 8).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Annadale	40.54455	-74.176532	Annadale Terrace	40.542555	-74.177187	Restaurant
1	Annadale	40.54455	-74.176532	Annadale Diner	40.542079	-74.177325	Diner
2	Annadale	40.54455	-74.176532	Play Sports Bar	40.540418	-74.177196	Sports Bar
3	Annadale	40.54455	-74.176532	Il Sogno	40.541286	-74.178489	Restaurant
4	Annadale	40.54455	-74.176532	Harbor View Restoration LLC	40.544934	-74.174709	Construction & Landscaping

Figure 8 - Details of some of the venues.

Finally, in order to get better results, it is performed one hot encoding technique on the venues data. The venues are grouped by the neighbourhood and the means are calculated. In the end, the 5 common venues are calculated for each of the neighbourhoods. The final goal of this project is to cluster similar neighbourhoods using 5 – means clustering, so people can find similar neighbourhoods, i.e., regions that have similar characteristics and amenities. As mentioned, the 5 – means clustering will then cluster the neighbourhoods into 5 clusters.

Results

After applying 5-means clustering, it is possible to access to each one of the 5 clusters and know the neighbourhood that are associated to each of the clusters.

The first cluster is the biggest one (Figure 9) with 18 of the 27 neighbourhoods. It can be deonoted that the most common venue in these neighbourhoods are restaurants (Italian, American, Indian, …), Pizza place and Bagel Shop.

	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Annadale	STATEN ISLAND	40.544550	-74.176532	0	American Restaurant	Pizza Place	Restaurant	Liquor Store	Sports Bar
1	Arlington	STATEN ISLAND	41.694069	-73.886363	0	Automotive Shop	Convenience Store	Event Space	Mobile Phone Shop	Trail
2	Arrochar	STATEN ISLAND	40.598438	-74.072641	0	Pizza Place	Park	Deli / Bodega	Bagel Shop	Bus Stop
3	Bay Terrace	STATEN ISLAND	40.555278	-74.134167	0	Italian Restaurant	American Restaurant	Miscellaneous Shop	Playground	Food Truck
4	Bloomfield	STATEN ISLAND	42.895064	-77.434713	0	Ice Cream Shop	American Restaurant	Pizza Place	Lawyer	Flower Shop
5	Bulls Head	STATEN ISLAND	40.607048	-74.162088	0	Bus Stop	Pharmacy	Diner	Baseball Field	Tattoo Parlor
6	Castleton Corners	STATEN ISLAND	40.613160	-74.122365	0	Pizza Place	Bank	Ice Cream Shop	Convenience Store	Salon / Barbershop
7	Charleston	STATEN ISLAND	40.536772	-74.237367	0	Construction & Landscaping	Bakery	Gym / Fitness Center	Deli / Bodega	Rental Car Location
8	Chelsea	STATEN ISLAND	40.746491	-74.001528	0	Café	French Restaurant	Sushi Restaurant	Coffee Shop	Indian Restaurant
12	Egbertville	STATEN ISLAND	40.578622	-74.131570	0	Construction & Landscaping	Italian Restaurant	Bagel Shop	Trail	Cosmetics Shop

Figure 9 - Part of the first cluster.

In turn, the second, thirst ant fifth cluster have only neighbourhood each, what means that are no neighbourhood similar to them.

In the second one (Figure 10), the venues consist of train station, Italian restaurant and movie theatre, yoga studio and flower shop.

	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
25	Huguenot	STATEN ISLAND	40.537328	-74.194588	1	Train Station	Italian Restaurant	Movie Theater	Yoga Studio	Flower Shop

Figure 10 - Second Cluster.

The venues in the third cluster comprehends an indie theather, golf course, trail, café and yoga studio (Figure 11).

	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
22	Grymes Hill	STATEN ISLAND	40.618715	-74.093475	2	Indie Theater	Golf Course	Trail	Café	Yoga Studio

Figure 11 - Third Cluster.

In turn, some of the venues in the fifth cluster are pizza place, Asian restaurant and flower shop.

	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
11	Dongan Hills	STATEN ISLAND	40.597927	-74.098027	4	Pizza Place	Asian Restaurant	Flower Shop	Yoga Studio	Food

Figure 12 - Fifth Cluster.

The last one, the fourth cluster (Figure 13), has 5 neighbourhoods that have common venues such as bus stop and deli/bodega.

	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	Clifton	STATEN ISLAND	40.620104	-74.077086	3	Bus Stop	Grocery Store	Intersection	Deli / Bodega	Train Station
13	Elm Park	STATEN ISLAND	40.631493	-74.148754	3	Deli / Bodega	Italian Restaurant	Bus Stop	Athletics & Sports	Furniture / Home Store
15	Emerson Hill	STATEN ISLAND	40.608585	-74.094564	3	Bus Stop	Acupuncturist	Deli / Bodega	Intersection	Automotive Shop
23	Hamilton Park	STATEN ISLAND	40.641726	-74.090072	3	Bus Stop	Park	Bowling Alley	Mexican Restaurant	Deli / Bodega
24	Heartland Village	STATEN ISLAND	40.588333	-74.157778	3	Gym / Fitness Center	Sandwich Place	Bus Stop	Bus Station	Yoga Studio

Figure 13 - Fourth Cluster.

In the next figure is represented the five clusters in the map, in order to understand its geographical localisation (Figure 14). Each cluster is associated to a colour (red, blue, green, orange and purple).

Figure 14 - Clustered Neighbourhoods in Staten Islands.

Discussion

This project has as aim to help people to find the safest borough in New York City based on the crimes committed between 2014 and 2015 and, based in a cluster technique, also allow people to find neighbourhoods with similarities. This way, as soon as someone want to move to NYC or to relocate inside the city, if one of the wishes is to find a safer borough, the answer is staten island. However, inside this borough, there are a lot of neighbourhoods. So, if someone really want a place that has indie theather, golf course, trail, café and yoga studio, maybe the best neighbourhood is Grymes Hill (third cluster). Conversely, if it desires to have a train station, some Italian restaurant and movie theatre, yoga studio and flower shop, then the best neighbourhood is Huguenot (second cluster). However, the final choice may vary depending on the interests of each person.

Conclusion

Just as mentioned in the introduction section, this works is a big help for people that has as intention to relocate or to move to the safest borough of NYC (This was my criteria, but it could be other one), as well as, have the notion of the similar neighbourhoods of that borough. So, this type of technology can be a decision support, make it easy and more reliable.