Battle of Boroughs and Neighborhoods – Introduction and Data
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) *What is the safest borough in NYC?*
(2) *What is the best neighborhood based on venues?*

1.3 Interest

People who are considering to move 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 crime 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