Battle of the Neighborhoods:

Finding a Family-friendly Neighborhood in New York City

Applied Data Science Capstone Project

April 26, 2020

Contents

1. Introduction

1.1 Background

Americans are considered to be highly mobile. The average American moves once every five years. Moving is usually one of the most stressful time periods for a family, since they are leaving everything and everyone they know behind. Families need data to help them make decisions regarding where to buy a home, including school quality, safety, location convenience and recreation. This project develops a way for families to make informed decisions when making a move to New York City.

1.2 Problem

This project aims to find the New York City Borough with the lowest crime rate and the best schools for Science, Technology, Engineering and Math (STEM) education, and then find the neighborhoods with the best recreation and convenience options for a family with school-aged children.

1.3 Interest

Any families with school-aged children relocating to New York City would find this information valuable.

2 Data

The New York City data for this effort is from various data sources.

2.1 Crime Data

Data from Kaggle was used for Crime data for New York City from 2013-2015.

https://www.kaggle.com/adamschroeder/crimes-new-york-city

It consists of crimes that occurred in each New York City Borough. It includes attributes such as Borough Name, Date of Occurrence, Data of Report and Level of Offense (felony, misdemeanor, violation). The full list of attributes and descriptions are found in Table 1.

This crime data will be used to find the Borough with the lowest crime rate which would be the safest neighborhood for a family thinking of relocating to New York City.

Column	Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
RPT_DT	Date event was reported to police

KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
JURIS_DESC	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

Table 1. New York City Crime Data

2.2 School Quality Data

Data from Kaggle was also used for the education data. SAT scores for New York City schools for the 2014-2015 school year were provided.

https://www.kaggle.com/nycopendata/high-schools#scores.csv

This data consists of Borough, School Name, Average SAT Math Score, Average SAT Reading Score and Average SAT Writing Score. The complete list of attributes is in Table 2.

The Average SAT Math Score will be used to find the Borough of New York City with the highest test scores, which would allow a family concerned about STEM education to determine the Borough in which to relocate.

Column	Description
School ID	School identification number
School Name	Name of the high school
Borough	Borough where the school is located
Building Code	School building identification number
Street Address	School address
City	School city
State	School state
Zip Code	School zip
Latitude	Latitudinal location of school
Longitude	Longitudinal location of school
Phone Number	School phone number
Start Time	School day start time
End Time	School day completion time
Student Enrollment	Number of Students enrolled in the high school
Percent White	Percentage of Caucasian students enrolled
Percent Black	Percentage of African American students enrolled
Percent Hispanic	Percentage of Hispanic students enrolled
Percent Asian	Percentage of Asian students enrolled
Average Score (SAT Math)	Average score on the SAT Math Exam
Average Score (SAT Reading)	Average score on the SAT Reading Exam
Average Score (SAT Writing)	Average score on the SAT Writing Exam
Percent Tested	Percentage of students tested at the high school

Table 2. SAT Score Data for New York City High Schools

2.3 Location Convenience and Recreation Data

Data from foursquare.com was then used to find venue and recreation (parks) data for New York City. The foursquare API is noted below.

```
LIMIT = 100 # limit of number of venues returned by Foursquare API
radius = 500 # define radius
# create URL
url =
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={
}'.format(
   CLIENT_ID,
   CLIENT_SECRET,
   VERSION,
   neighborhood_latitude,
   neighborhood_longitude,
   radius,
   LIMIT)
url # display URL
```

This data consists of Borough, Neighborhood, Latitude, Longitude, Venue, Venue Category. Venue Category includes Parks, but also includes Restaurants and Grocery Stores.

This data is used to find the neighborhoods in the Borough with the lowest crime rate and best quality STEM education. An example is shown in Table 3.

	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

Table 3. Example of Neighborhood Data from Foursquare.

The data is then used to find the total number of venues, representing neighborhood convenience with respect to grocery stores and other necessities. An example is shown in Table 4.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	St. George	40.644982	-74.079353	A&S Pizzeria	40.643940	-74.077626	Pizza Place
1	St. George	40.644982	-74.079353	Beso	40.643306	-74.076508	Tapas Restaurant
2	St. George	40.644982	-74.079353	Staten Island September 11 Memorial	40.646767	-74.076510	Monument / Landmark
3	St. George	40.644982	-74.079353	Richmond County Bank Ballpark	40.645056	-74.076864	Baseball Stadium
4	St. George	40.644982	-74.079353	Shake Shack	40.643660	-74.075891	Burger Joint

Table 4. Venue Information for a Particular Borough and Neighborhood

The data is the used to find the total number of parks, representing how well the area supports recreation for families. Table 5 provides a representation of this data.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
32	St. George	40.644982	-74.079353	Fort Hill	40.641511	-74.080522	Park
33	St. George	40.644982	-74.079353	Maritime Hospital Quarantine Cemetery	40.641593	-74.077730	Park
42	New Brighton	40.640615	-74.087017	Bocce Courts	40.639800	-74.090000	Park
47	New Brighton	40.640615	-74.087017	Mahoney Park	40.643793	-74.085313	Park
54	Stapleton	40.626928	-74.077902	5050 Skatepark	40.628053	-74.074548	Skate Park

Table 5. Neighborhoods in a Particular Borough with Parks

2.4 New York City Map Data

New York City map data was also obtained using a data set that is available on the web.

https://geo.nyu.edu/catalog/nyu_2451_34572

3 Methodology

3.1 Exploratory Data Analysis

3.1.1 Crime Data

The crime data was analyzed to determine the Borough with the lowest number of crimes in aggregate. The number of crimes for the data set were compiled and a bar chart was developed. Over the time period studied (2013-2015), Staten Island had the lowest number of crimes, with close to 50,000 compared to over 200,000 for the other Boroughs.

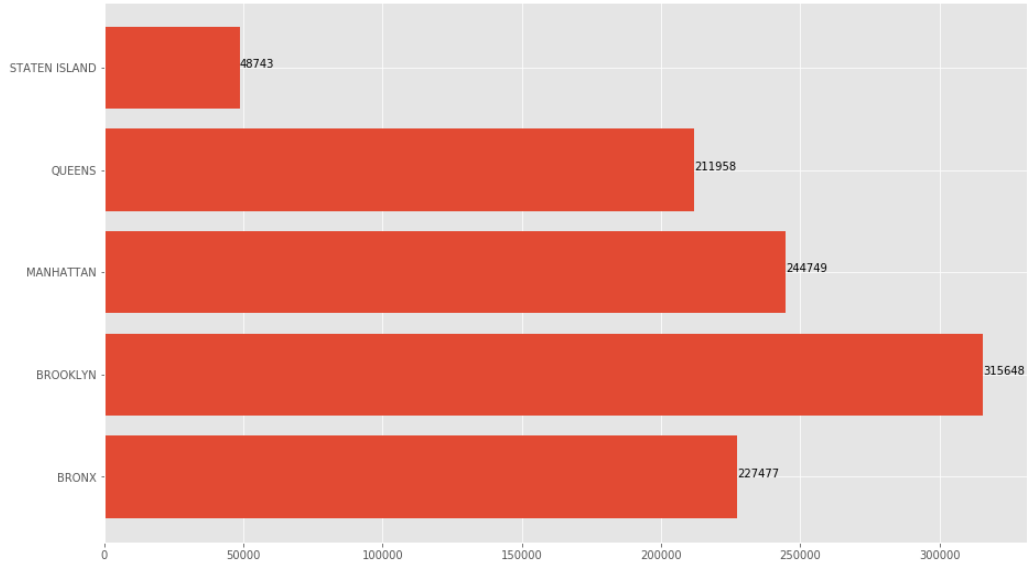

Figure 1. New York City Crime Data by Borough

3.1.2 Education Data

The education data was also examined to determine the Borough with the best SAT Math test scores. A box plot was developed using the data from each Borough. It was determined that Staten Island schools had the best scores overall and should be strongly considered for relocation. The inner quartile range (25%-75%) and median were higher for Staten Island, although the maximum scores were higher in other Boroughs. In Staten Island, a parent would be guaranteed solid results.

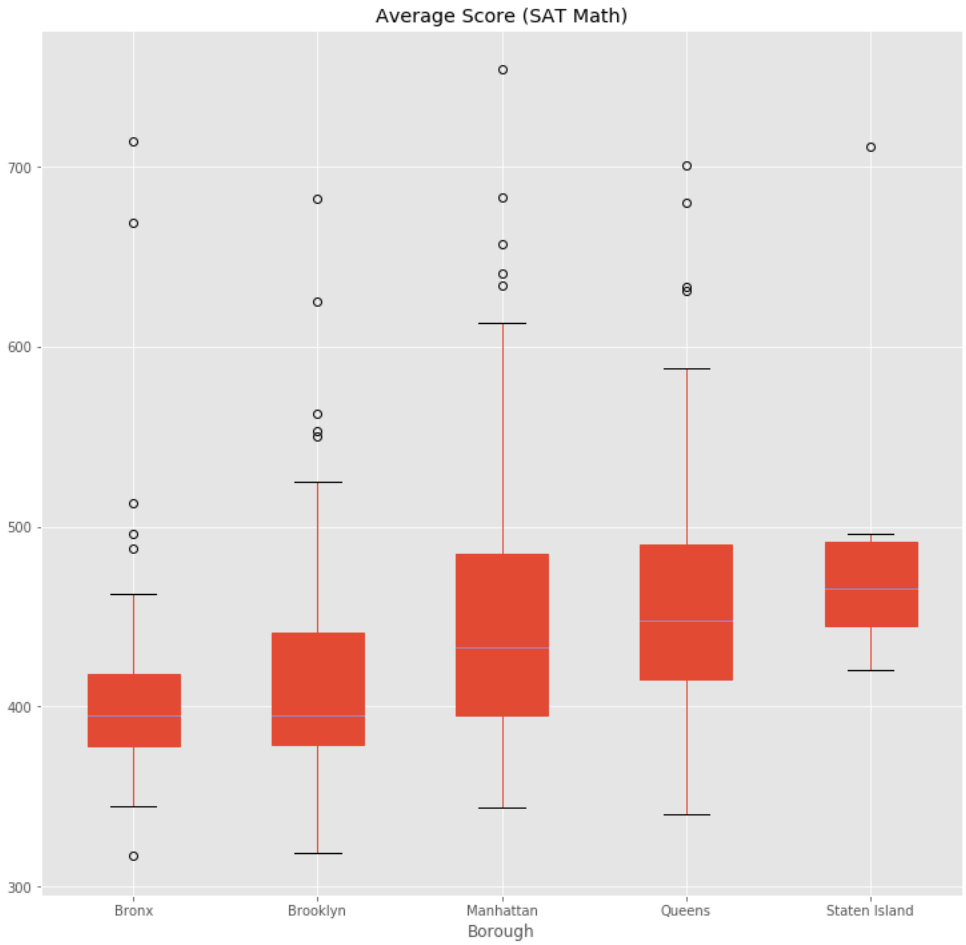

Figure 2. SAT Math Scores by Borough

3.1.3 Maps

New York City consists of five Boroughs: Brooklyn, Bronx, Manhattan, Queens and Staten Island. The map is shown in Figure 3. Staten island is in the southwest portion of New York City. The map is shown in Figure 4.

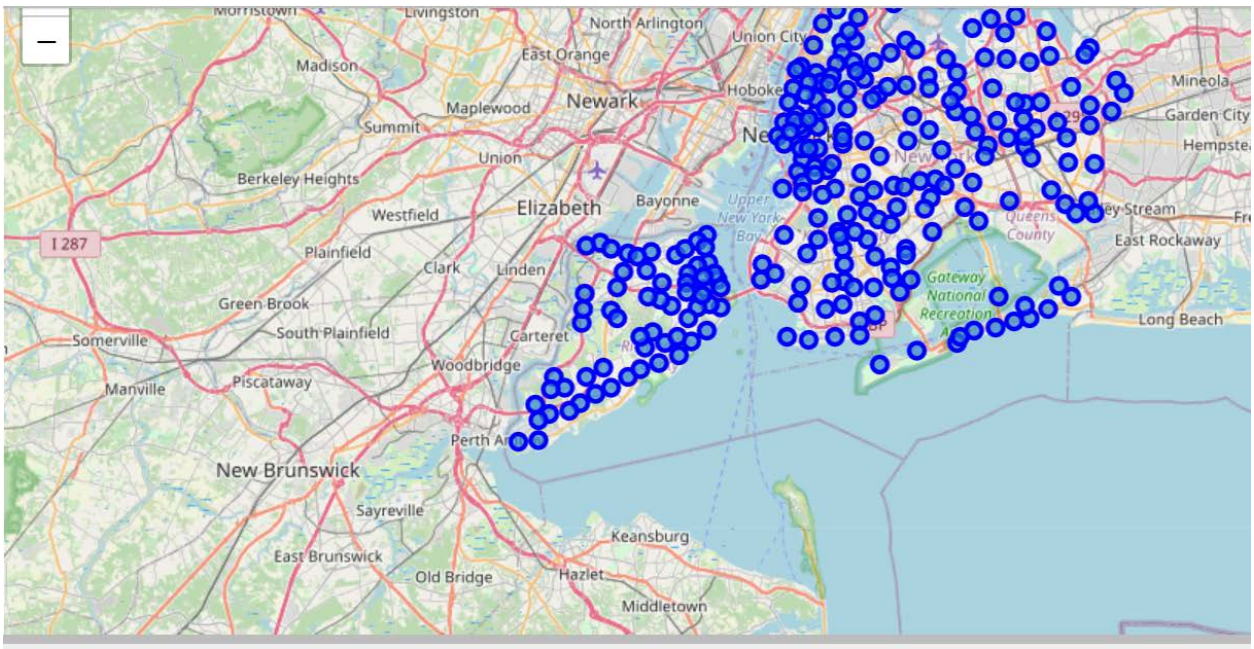

Figure 3. Map of New York City

Figure 4. Map of Staten Island Borough

3.2 Exploratory Data Analysis

Staten Island is a large island with 63 neighborhoods. K-Means clustering was performed as some initial analysis to determine the neighborhoods with the largest number of venues in proximity to a potential home to make it a convenient location for a family.

The Foursquare API was used to obtain neighborhood venue data. One Hot encoding was performed on the venue data to allow it to be better consumed by a Machine Learning algorithm.

In addition to K-Means clustering to help focus the neighborhoods and understand the venue distribution. Data for those neighborhoods was used to perform further analysis. A join of the "Top 10" neighborhoods for venue count table with the neighborhoods that have parks table was performed. Three neighborhoods would be ideal for a family with school-aged children: St. George, Stapleton and Grasmere.

4 Results

The Top 10 neighborhoods for venue count are shown in the table below. For a family, these neighborhoods would be extremely convenient.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bulls Head	45	45	45	45	45	45
1	Eltingville	42	42	42	42	42	42
2	West Brighton	38	38	38	38	38	38
3	St. George	36	36	36	36	36	36
4	Charleston	31	31	31	31	31	31
5	Stapleton	31	31	31	31	31	31
6	Rosebank	29	29	29	29	29	29
7	Dongan Hills	24	24	24	24	24	24
8	Grant City	24	24	24	24	24	24
9	Grasmere	24	24	24	24	24	24

Table 6. Top 10 Neighborhoods by Venue Count

The neighborhoods on Staten Island with Parks are shown in the table below. Ten of the twenty-six neighborhoods on Staten Island have parks.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bloomfield	2	2	2	2	2	2
1	Clifton	1	1	1	1	1	1
2	Concord	1	1	1	1	1	1
3	Grasmere	1	1	1	1	1	1
4	New Brighton	2	2	2	2	2	2
5	Randall Manor	1	1	1	1	1	1
6	St. George	2	2	2	2	2	2
7	Stapleton	2	2	2	2	2	2
8	Todt Hill	1	1	1	1	1	1
9	Tompkinsville	1	1	1	1	1	1
10	Travis	1	1	1	1	1	1

Table 7. Park Count by Borough

Bloomfield	False
Clifton	False
Concord	False
Grasmere	True
New Brighton	False
Randall Manor	False
St. George	True
Stapleton	True
Todt Hill	False
Tompkinsville	False
Travis	False

Table 8. Table Join for Parks and Venues

Based on the information regarding venue count and parks in Table 8, it was determined that either St. George, Stapleton or Grasmere would be the better districts. In the K-Means Clustering results, these neighborhoods were in the same cluster.

Through the K-Means clustering analysis it was determined that St. George was close to Sporting Goods Stores, Clothing Stores, Farmers Market, Bus Stations and Scenic Lookout. There was plenty to keep a family busy. However, St. George has many bars, which may not be the best for a family atmosphere. Meanwhile, Grasmere has bus stop, restaurants and grocery stores. Stapleton has restaurants, discount stores and a bank. Given that Grasmere has restaurants and grocery stores, and also has a park, this would likely be the first choice for a family with school-aged children.

	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	St. George	Clothing Store	Sporting Goods Shop	Bar	Italian Restaurant	Monument / Landmark	Burger Joint	Bus Station	Scenic Lookout	Farmers Market	Donut Shop
2	Stapleton	Mexican Restaurant	Pizza Place	Sandwich Place	Restaurant	Discount Store	Bank	Coffee Shop	Spanish Restaurant	Fast Food Restaurant	Motorcycle Shop
3	Rosebank	Mexican Restaurant	Grocery Store	Italian Restaurant	Pharmacy	Cosmetics Shop	Discount Store	Donut Shop	Eastern European Restaurant	Restaurant	Sandwich Place
4	West Brighton	Coffee Shop	Breakfast Spot	Bus Stop	Italian Restaurant	Music Store	Bank	Bar	Pharmacy	Diner	Café
6	Todt Hill	Trail	Park	Yoga Studio	Electronics Store	Food	Flower Shop	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Fast Food Restaurant
8	Port Richmond	Bar	Bus Stop	Rental Car Location	Donut Shop	Basketball Court	Event Space	Food	Flower Shop	Fish & Chips Shop	Financial or Legal Service

26	Graniteville	Food Truck	Bus Stop	Sandwich Place	Grocery Store	Electronics Store	Flower Shop	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Fast Food Restaurant
27	Arlington	Bus Stop	American Restaurant	Home Service	Coffee Shop	Boat or Ferry	Deli / Bodega	Hookah Bar	Food	Ice Cream Shop	Fish & Chips Shop
29	Grasmere	Bus Stop	Bagel Shop	Ice Cream Shop	Nail Salon	Restaurant	Grocery Store	Park	Bank	Bakery	Business Service
32	Midland Beach	Deli / Bodega	Bus Stop	Pet Store	Beach	Basketball Court	Dessert Shop	Bookstore	Restaurant	Liquor Store	Electronics Store

Table 9. K-Means Clustering - Cluster Number 1

	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	Marble Hill	Sandwich Place	Gym	Coffee Shop	Yoga Studio	Steakhouse	Seafood Restaurant	Tennis Stadium	Supplement Shop	Bank	Donut Shop
1	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	Salon / Barbershop	American Restaurant	Optical Shop	Spa	Coffee Shop	Bubble Tea Shop	Asian Restaurant
6	Central Harlem	Gym / Fitness Center	American Restaurant	Seafood Restaurant	French Restaurant	Chinese Restaurant	African Restaurant	Bar	Southern / Soul Food Restaurant	Bookstore	Beer Bar
14	Clinton	Theater	Gym / Fitness Center	Coffee Shop	Gym	Spa	Hotel	Thai Restaurant	Sandwich Place	Italian Restaurant	American Restaurant
15	Midtown	Coffee Shop	Hotel	Clothing Store	Theater	Pizza Place	Bakery	Gym	Japanese Restaurant	Spa	Mediterranean Restaurant
16	Murray Hill	Sandwich Place	Hotel	Bar	Burger Joint	Gym / Fitness Center	Pizza Place	Coffee Shop	Japanese Restaurant	Chinese Restaurant	Deli / Bodega

Table 10. K-Means Clustering – Cluster Number 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
53	Howland Hook	Pier	Yoga Studio	Electronics Store	Food	Flower Shop	Fish & Chips Shop	Financial or Legal Service	Filipino Restaurant	Fast Food Restaurant	Farmers Market

Table 11. K-Means K-Means Clustering – Cluster Number 3

5 Discussion

The objective of this exercise was to enable families to make an informed decision when moving to New York City. The potential home buyers would be aware of the crime in the area and be able to make a purchase with safety in mind. The home buyers would also be able to understand the quality of the STEM education in a New York City Borough based on the SAT math scores at the high schools in the area. Lastly, the home buyers would be able to consider the convenience of the area based on the general number of venues (including restaurants and grocery stores) and recreation based on the parks in the area. Neighborhoods in Staten Island are the most suitable for a family with school-aged children. Within Staten Island, the neighborhoods of St. George, Stapleton and Grasmere should be considered.

Given that Grasmere has parks, restaurants, grocery stores and a number of other family-friendly venues, it would be ideal.

6 Conclusion

This project helps a family relocating to New York City an understanding of the safety of the Boroughs and the quality of the schools in each Borough. It also provides the ability to make a neighborhood selection based on the convenience of venues such as grocery stores and restaurants as well as recreational access through parks. In the future, this project could more closely examine the school rankings by neighborhood and include non-STEM scores in the analysis.