Applied Data Science Capstone Peer-Graded Assignment

July 8, 2019

Introduction

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States.

Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass and one of the world's most populous megacities, with an estimated 19,979,477 people in its 2018 Metropolitan Statistical Area and 22,679,948 residents in its Combined Statistical Area. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. The city's fast pace has inspired the term New York minute. Home to the headquarters of the United Nations, New York is an important center for international diplomacy.

As the city grows and develops, it becomes increasingly important to examine and understand it quantitiatively.

Developers, policy makers and/or city planners have an interest in answering the following questions:

1. What neighbourhoods have the highest crime?
2. Is population density correlated to crime level?
3. Using Foursquare data, what venues are most common in different locations within the city?
4. Does a specific neighborhood of New York City really need a new coffee shop?

1 Data

To understand and explore we will need the following City of New York City Open Data:

- Open Data Site: https://data.cityofnewyork.us/
- New York City Neighbourhoods: https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq
- New York City Census: https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp
- New York Boroughs centroids: 'https://raw.githubusercontent.com/gfumarco/Applied-Data-Science-Capstone/master/NYC_Boroughs_LatLong.csv'
- NYPD police precintcs: https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz
- NYPD complaints to date records: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243
- Foursquare Developers Access to venue data: https://foursquare.com/

Using this data will allow exploration and examination to answer the questions.

1.1 Data Details:

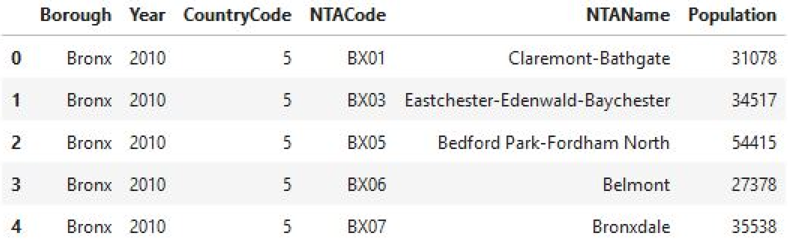
Neighborhood-Tabulation-Area provide a geojson that segment the urban area into the different neighbourhoods.

New-York-City-Population-By-Neighborhood-Tabulation include the population numbers by Neighbourhood population areas.

Columns in this Dataset:

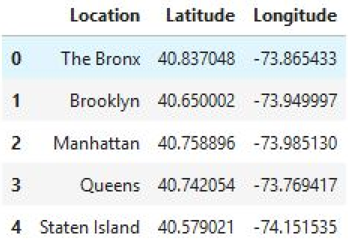
- *Borough*
- *Year*
- *FIPS*
- *County*
- *Code*
- *NTA_ Code*
- *NTA_ Name*

- *Population*



	Borough	Year	CountryCode	NTACode	NTAName	Population
0	Bronx	2010	5	BX01	Claremont-Bathgate	31078
1	Bronx	2010	5	BX03	Eastchester-Edenwald-Baychester	34517
2	Bronx	2010	5	BX05	Bedford Park-Fordham North	54415
3	Bronx	2010	5	BX06	Belmont	27378
4	Bronx	2010	5	BX07	Bronxdale	35538

NYC_Boroughs_LatLong.csv is a compilation of the 5 boroughs centroids.



	Location	Latitude	Longitude
0	The Bronx	40.837048	-73.865433
1	Brooklyn	40.650002	-73.949997
2	Manhattan	40.758896	-73.985130
3	Queens	40.742054	-73.769417
4	Staten Island	40.579021	-74.151535

Police-Precincts provide a geojson that segment the urban area into the different neighbourhoods.

NYPD-Complaint-Data-Current-Year-To-Date is a breakdown of every arrest effected in NYC by the NYPD during the current year. Data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning, each record represents an arrest effected in NYC by the NYPD and includes information about the type of crime, the location and time of enforcement.

Original columns in this Dataset:

- *CMPLNT_NUM*: Randomly generated persistent ID for each complaint
- *ADDR_PCT_CD*: The precinct in which the incident occurred
- *BORO_NM*: The name of the borough in which the incident occurred
- *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

- *CRM_ATPT_CPTD_CD*: Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely

- *HADEVELOPT*: Name of NYCHA housing development of occurrence, if applicable

- *HOUSING_PSA*: Development Level Code

- *JURISDICTION_CODE*: Jurisdiction responsible for incident. Either internal, like Police(0), Transit(1), and Housing(2); or external(3), like Correction, Port Authority, etc.

- *JURIS_DESC*: Description of the jurisdiction code

- *KY_CD*: Three digit offense classification code

- *LAW_CAT_CD*: Level of offense: felony, misdemeanor, violation

- *LOC_OF_OCCUR_DESC*: Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of

- *OFNS_DESC*: Description of offense corresponding with key code

- *PARKS_NM*: Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)

- *PATROL_BORO*: The name of the patrol borough in which the incident occurred

- *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)

- *PREM_TYP_DESC*: Specific description of premises; grocery store, residence, street, etc.

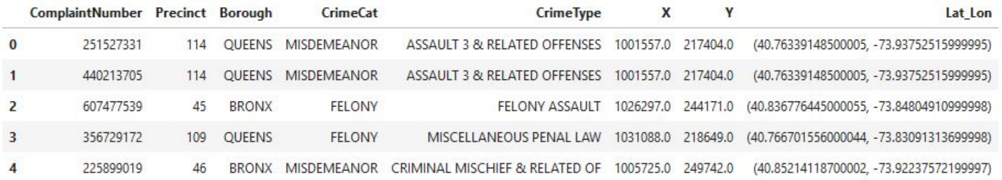
- *RPT_DT*: Date event was reported to police

- *STATION_NAME*: Transit station name

- *SUSP_AGE_GROUP*: Suspect's Age Group

- *SUSP_RACE*: Suspect's Race Description
- *SUSP_SEX*: Suspect's Sex Description
- *TRANSIT_DISTRICT*: Transit district in which the offense occurred.
- *VIC_AGE_GROUP*: Victim's Age Group
- *VIC_RACE*: Victim's Race Description
- *VIC_SEX*: Victim's Sex Description
- *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*: Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- *Longitude*: Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
- *Lat_Lon*: Location

It is not possible to load more than 1000 rows from the NYPD datasite in a jupyter notebook, original dataset has been downloaded, trimmed from some useless columns for this analysis (to keep it under the 25mb maximum size of github) and then loaded on: https://raw.githubusercontent.com/gfumarco/Applied-Data-Science-Capstone/master/NYPD_Complaint_Data.csv



	ComplaintNumber	Precinct	Borough	CrimeCat	CrimeType	X	Y	Lat_Lon
0	251527331	114	QUEENS	MISDEMEANOR	ASSAULT 3 & RELATED OFFENSES	1001557.0	217404.0	(40.76339148500005, -73.93752515999995)
1	440213705	114	QUEENS	MISDEMEANOR	ASSAULT 3 & RELATED OFFENSES	1001557.0	217404.0	(40.76339148500005, -73.93752515999995)
2	607477539	45	BRONX	FELONY	FELONY ASSAULT	1026297.0	244171.0	(40.836776445000055, -73.84804910999998)
3	356729172	109	QUEENS	FELONY	MISCELLANEOUS PENAL LAW	1031088.0	218649.0	(40.766701556000044, -73.83091313699998)
4	225899019	46	BRONX	MISDEMEANOR	CRIMINAL MISCHIEF & RELATED OF	1005725.0	249742.0	(40.85214118700002, -73.92237572199997)

The neighbourhood data will enable us to properly group crime by neighbourhood. The Census data will enable us to then compare the population density to examine if areas of highest crime are also most densely populated.

2 Methodology

The methodology will include:

1. Loading each data set
2. Examine the crime frequency by neighborhood
3. Study the crime types and then pivot analysis of crime type frequency by neighborhood
4. Understand correlation between crimes and population density
5. Perform k-means statistical analysis on venues by locations of interest based on findings from crimes and neighborhood
6. Determine which venues are most common statistically in the region of greatest crime count then in all other locations of interest.
7. Determine if an area, such as Manhattan needs a coffee shop.

3 Loading step

After loading the applicable libraries, the referenced geojson neighbourhood data was loaded from the *NYC (New York City) Open Data site.* This dataset uses block polygon shape coordinates which are better for visualization and comparison. The City also uses Precinct data but the NTA (*Neighbourhood Tabulation Areas*) location data is more accurate and includes more details. The same type of dataset was then loaded for the population density from the NYC Census tracts. This dataset contains the population of each NTA for each year from 2000 to 2010; sadly more recent data wasnt available.

The dataset *NYPD-Arrest-Data-Year-to-Date* was loaded from the NYC Open Data site, under the Public Safety domain, for the analysis. Original version It is an exhaustive set by not including all crimes (violent offenses) nor specific location data of the crime but is referenced by neighborhood.

It's interesting to note the details of this dataset are aggregated by precinct; this made necessary to download the referenced geojson NYPD police precinct data from the NYC (New York City) Open Data site.

This means we can gain an understanding of the crime volume by type by precinct but not specific enough to understand the distribution properties. Valuable questions such as, "are these crimes occurring more often in a specific area and at a certain time by a specific demographic of people?" cannot be answered nor explored due to what is reasonably assumed to be personal and private information with associated legal risks.

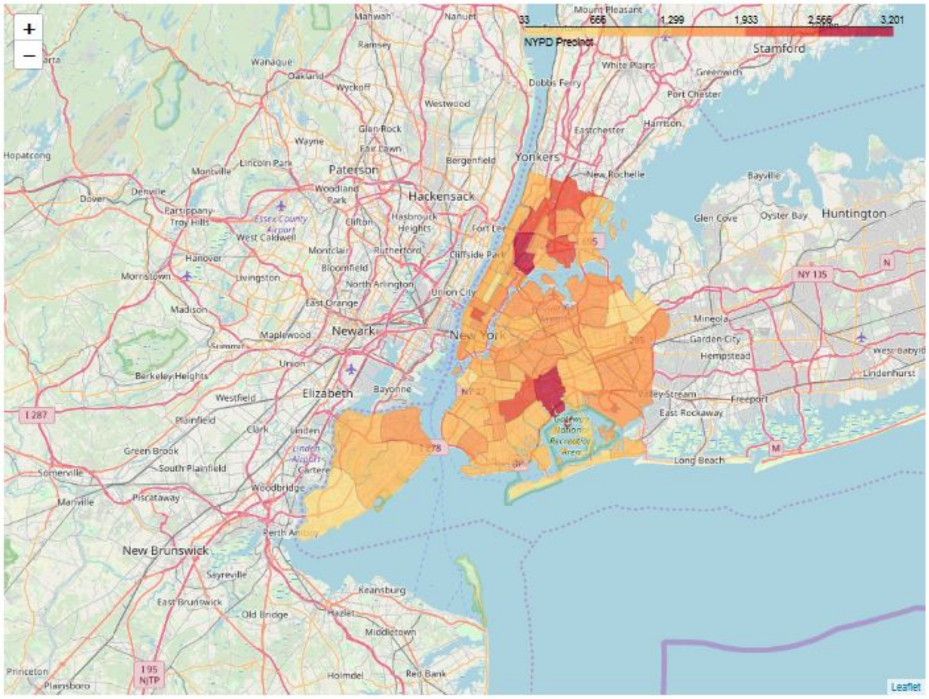
There is value to the city to explore the detailed crime data using data science to predict frequency, location, timing and conditions to best allocated resources for the benefit of its citizens and it's police force. However, human behavior is complex requiring thick profile data by individual and the conditions surrounding the event(s). To be sufficient for reliable future prediction it would need to demonstrate validity, currency, reliability and sufficiency.

4 Exploring the data

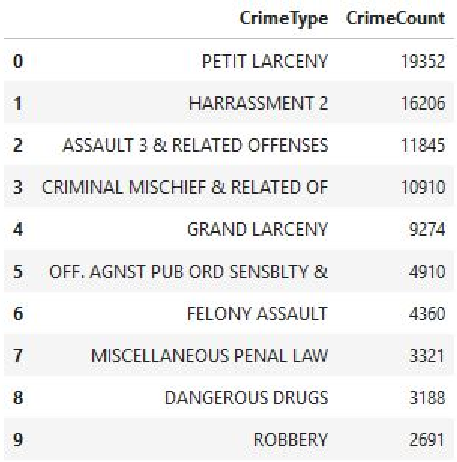
Exploring the count of crimes by precinct gives us the first glimpse into the distribution.

4.1 First Visualization of Crimes

Once the data was prepared, a choropleth map was created to view the crime count by precinct. The region of greatest crime count was found in the Brooklyn and Bronx boroughs.



Examining the crime types enables us to learn the most frequent occurring crimes:



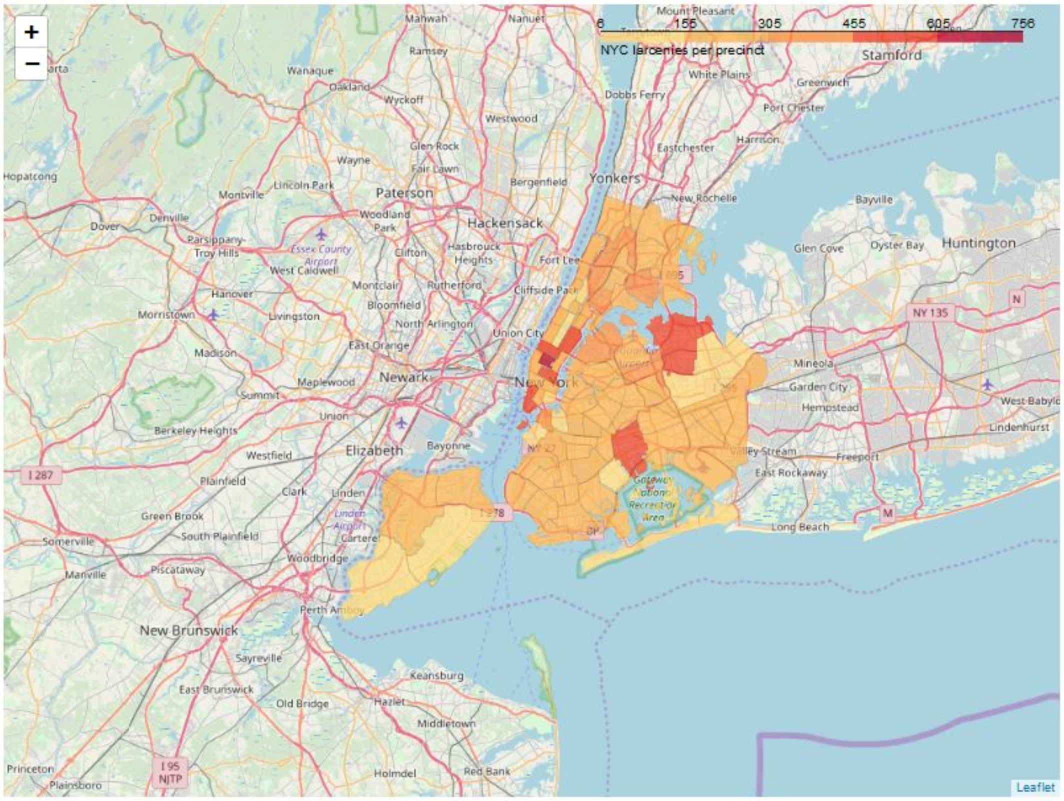
	CrimeType	CrimeCount
0	PETIT LARCENY	19352
1	HARRASSMENT 2	16206
2	ASSAULT 3 & RELATED OFFENSES	11845
3	CRIMINAL MISCHIEF & RELATED OF	10910
4	GRAND LARCENY	9274
5	OFF. AGNST PUB ORD SENSBLTY &	4910
6	FELONY ASSAULT	4360
7	MISCELLANEOUS PENAL LAW	3321
8	DANGEROUS DRUGS	3188
9	ROBBERY	2691

The first and secondo most common crimes are Petit Larceny and Harassment.

4.2 Examining 1st most common crime: Petit Larcenies

Petite larceny is most prevalent in Manhattan as well in the same area as the most frequent crimes. It's interesting to note that Manhattan is mostly commercial.

It would be interesting to further examine if surveillance is a deterrent for larcenies in the Manhattan borough compared to low surveillance in the Brooklyn and Bronx boroughs.

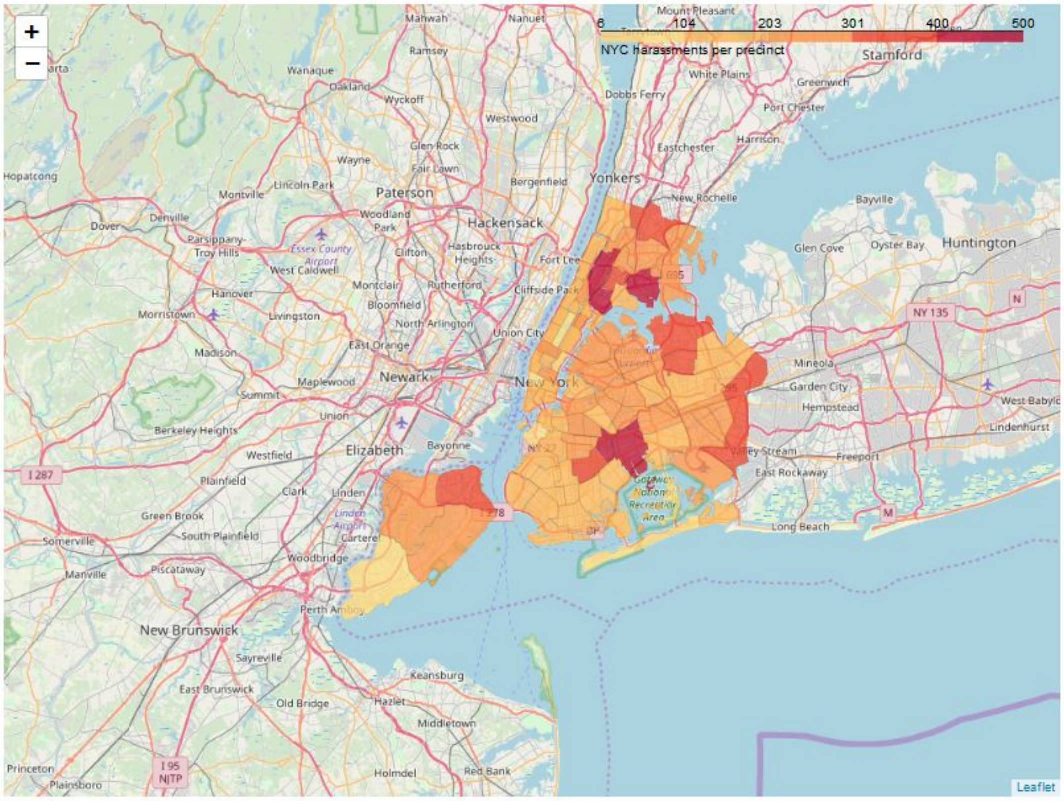


4.3 Examining 2nd most common crime: Harassments

Now we drill into the harassments crimes and plot the choropleth map to see which area has the greatest frequency.

Again, the Brooklyn and Bronx boroughs appear as the most frequent.

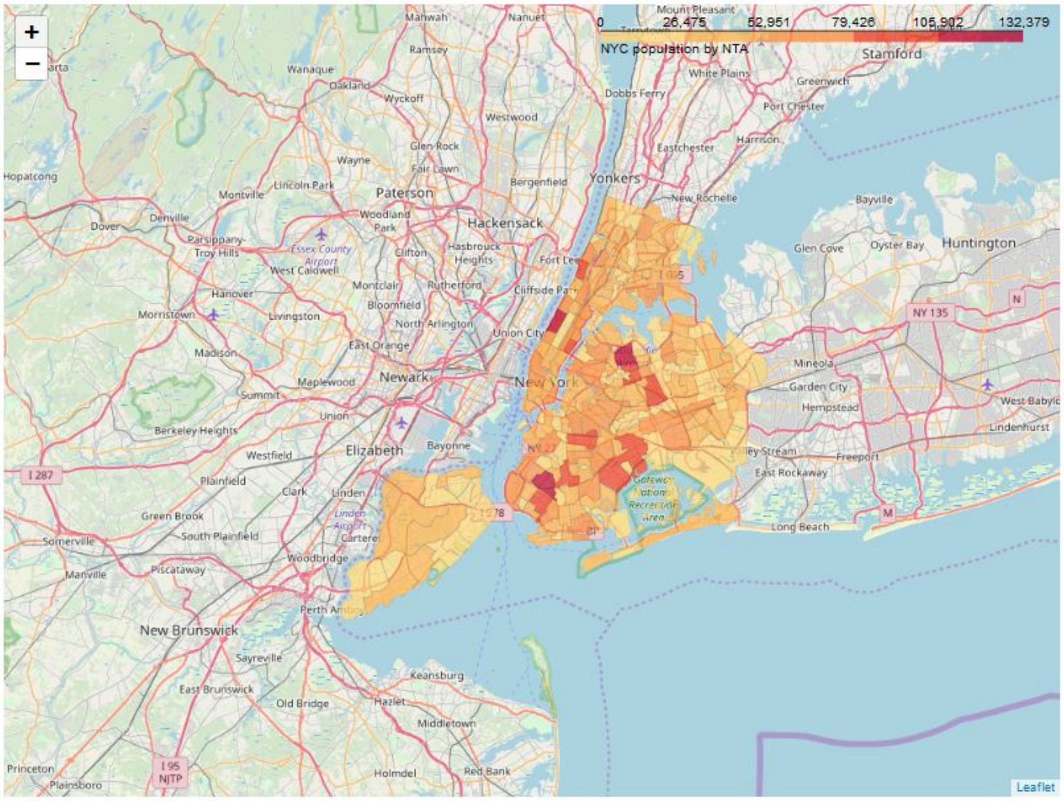
Is this due to population density?



4.4 Examining the Census data

Visualizing the population density enables us to determine that the Bronx has lower correlation to crime frequency than I would have expected.

It would be interesting to further study the Census data and if this captures the population that is renting or more temporary/transient population and/or hotel's check-ins, given the NYC is a commercial and financial hub.

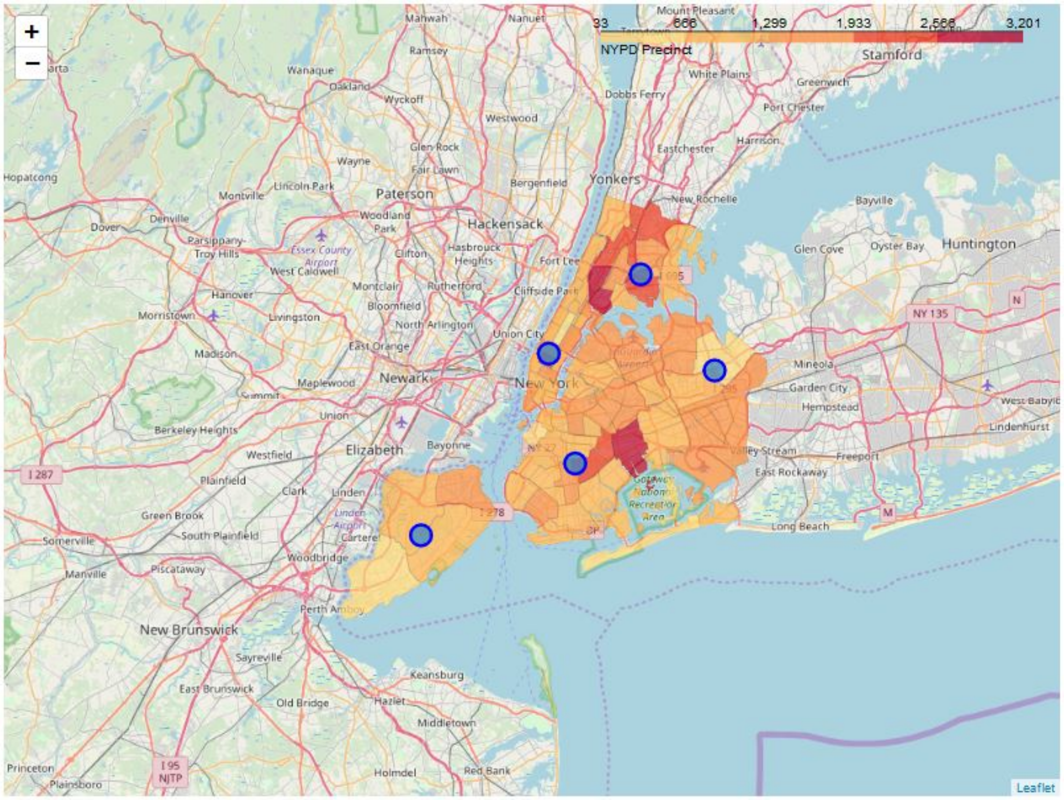


4.5 Look at specific locations to understand the connection to venues using Foursquare data

Loading the "NYC_Boroughs_LatLong.csv" data enables us to perform a statistical analysis on the most common venues by location.

We might wonder if the prevalence of bars and clubs has something to do with the higher crime rate regions.

Plotting the latitude and longitude coordinates of the locations of interest onto the crime choropleth map enables us to now study the most common venues by using the Foursquare data.

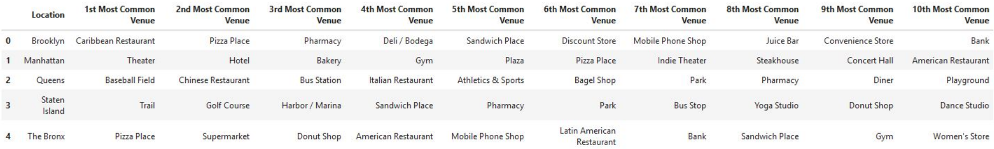


In this picture we plotted the centroids against the total crime choroplet map

4.6 Analysing each Location

Grouping rows by location and the mean of the frequency of occurance of each category we venue categories we study the top five most common venues.

Putting this data into a pandas dataframe we can then determine the most common venues by location.



	Location	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	Brooklyn	Caribbean Restaurant	Pizza Place	Pharmacy	Deli / Bodega	Sandwich Place	Discount Store	Mobile Phone Shop	Juice Bar	Convenience Store	Bank
1	Manhattan	Theater	Hotel	Bakery	Gym	Plaza	Pizza Place	Indie Theater	Steakhouse	Concert Hall	American Restaurant
2	Queens	Baseball Field	Chinese Restaurant	Bus Station	Italian Restaurant	Athletics & Sports	Bagel Shop	Park	Pharmacy	Diner	Playground
3	Staten Island	Trail	Golf Course	Harbor / Marina	Sandwich Place	Pharmacy	Park	Bus Stop	Yoga Studio	Donut Shop	Dance Studio
4	The Bronx	Pizza Place	Supermarket	Donut Shop	American Restaurant	Mobile Phone Shop	Latin American Restaurant	Bank	Sandwich Place	Gym	Women's Store

Frequencies of most 5 common venues:

----Brooklyn----

1. Caribbean Restaurant 0.16
2. Pizza Place 0.10
3. Deli / Bodega 0.05
4. Pharmacy 0.05
5. Sandwich Place 0.04

----Manhattan----

1. Theater 0.32
2. Hotel 0.05
3. Gym 0.03
4. Bakery 0.03
5. Plaza 0.03

----Queens----

1. Baseball Field 0.08
2. Chinese Restaurant 0.06
3. Pharmacy 0.04
4. Diner 0.04
5. Athletics & Sports 0.04

----Staten Island----

1. Trail 0.3
2. Golf Course 0.2
3. Sandwich Place 0.1
4. Bus Stop 0.1
5. Harbor / Marina 0.1

----The Bronx----

1. Pizza Place 0.12
2. Supermarket 0.08
3. Donut Shop 0.07
4. American Restaurant 0.05
5. Mobile Phone Shop 0.05

5 Results

The analysis enabled us to discover and describe visually and quantitatively:

1. Neighbourhood Tabulation Areas in NYC
2. Crime frequency by precinct
3. Crime type frequency and statistics. The mean crime count in New York City for the first two quarters of 2019 is 1364.
4. Crime type count by precinct. The region of greatest crime count was found in the Brooklyn and Bronx boroughs, followed by Manhattan. It would be interesting to further examine if surveillance is a deterrent for crimes in the Manhattan borough compared to low surveillance in the Brooklyn and Bronx boroughs.
5. Crime analysis by precinct and resulting statistics. The most common crime is Petite Larceny followed by Harassment. There is a mean of 251 larcenies in the City.
6. That population density and resulting visual correlation is not strongly correlated to crime frequency. Causation for crime is not able to be determined given lack of open data specificity by individual and environment.
7. Using k-means, we were able to determine the top 10 most common venues within a 1 km radius of the centroid of the highest crime borough. The most common venues in Manhattan are Theaters followed by Hotels and Gyms.
8. While, it is not valid, consistent, reliable or sufficient to assume a higher concentration of the combination of venues predicts the amount of crime occurrence in the City of New York, this may be a part of the model needed to be able to in the future.
9. We were able to determine the top 10 most common venues by location of interest. Statistically, we determined there are no coffee shops within the Manhattan cluster.

6 Discussion and Recommendations

The NYC Open Data Site enables us to gain an understanding of the crime volume by type by area but not specific enough to understand the distribution properties. Valuable questions such as, "are these crimes occurring more often in a specific area and at a certain time by a specific demographic of people?" cannot be answered nor explored due to what is reasonably assumed to be personal and private information with associated legal risks.

There is value to the city to explore the detailed crime data using data science to predict frequency, location, timing and conditions to best allocated resources for the benefit of its citizens and it's police force. However, human behavior is complex requiring thick profile data by individual and the conditions surrounding the event(s). To be sufficient for reliable future prediction it would need to demonstrate validity, currency, reliability and sufficiency.

A note of caution is the possibility Neighborhood Tabulation Areas names could change. The crime dataset did not mention in which specific NTA each crime was committed but used the Police Precinct naming convention. It may be beneficial for the City to note the NTA for easier exploration.

Larcenies and harassments are most prevalent in the same area as the most frequent crimes but also in Manhattan boroughs. It is interesting to note that Manhattan is mostly commercial and financial and has a lower occurrence of less frequent crimes. It would be interesting to further examine if Manhattan's surveillance is a deterrent for crimes compared to low surveillance in the Bronx and Brooklyn boroughs.

It would be interesting to further study the Census data and if this captures the population that is renting or more temporary/transient population and/or hotel's check-ins, given the NYC is a commercial and financial hub.

Given the findings of the top 10 most frequent venues by boroughs centroids, each borough does not have Coffee Shops in the top 10 most common venues as determined from the Foursquare dataset. Given Manhattan has the greatest concentration of theaters, hotels and gyms as venues, it would be safe to assume a coffee shop would be beneficial to the business community and the citizens of NYC.

7 Conclusion

Using a combination of datasets from the New York City Open Data Site and Foursquare venue data we were able to analyze, discover and describe neighborhoods, crime, population density and statistically describe quantitatively venues by locations of interest.

While overall, the New York City Open Data Site is interesting, it misses the details required for true valued quantitative analysis and predictive analytics which would be most valued by investors and developers to make appropriate investments and to minimize risk.

The Open Data Site is a great start and empowers the need for a "Citizens Like Me" model to be developed where citizens of digital NYC are able to share their data as they wish for detailed analysis that enables the creation of valued services.