Capstone Project – Battle of Neighborhoods (Week 1)

Data Extractions, Cleaning and Processing

A. Data Extraction

1) New York Crime Data

The crime data for the New York City is freely available on the official site of New York City. As the original dataset has data for many previous years, we will be taking the most recent data i.e. the first quarter of 2020 (Jan 2020 to Mar 2020). There are a lot of columns in this dataset hence the columns that we will be using are highlighted with Yellow color. The data set will be containing the following columns:

Column Name	Column Description
CMPLNT_NUM	Randomly generated persistent ID for each complaint
ADDR_PCT_CD	The precinct in which the incident occurred
BORO	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)

2) Neighborhood Dataset

The neighborhood dataset for is prepared using the New York geo json file . This dataset contains the list of all the neighborhoods as per boroughs and their coordinates.

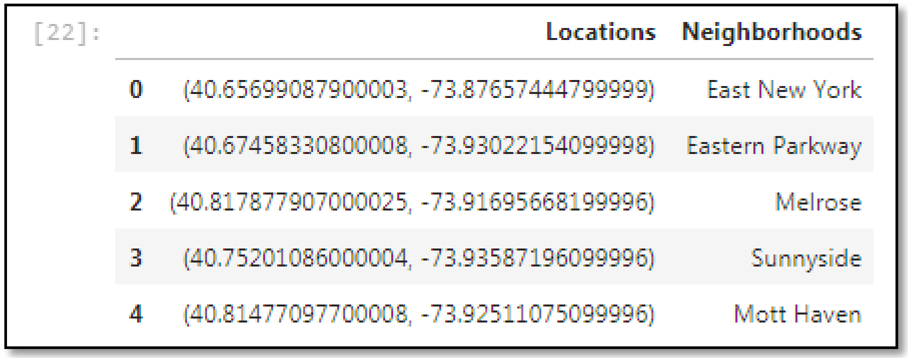
Column Name	Column Description
Borough	Name of the Borough
Neighborhood	Name of the Neighborhood
Latitude	Midblock Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude	Midblock Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

B. Data Cleaning and Processing

Before we can merge the Crime dataset with the Neighborhood data we need to do some data cleaning and processing by following these steps:

1) Remove all the data that is not in the first quarter of the year 2020 or has missing infomation.

2) As you can see that the above column list for Crime data does not have a Neighborhood column. Hence we need to add a Neighborhood column before we can merge the 2 datasets.
In order to do so we need to use GeoPy API and pass the Longitude and Latitude values in it to find the Neighborhood for each row. And then merge it with Crime dataset using Location as the key.



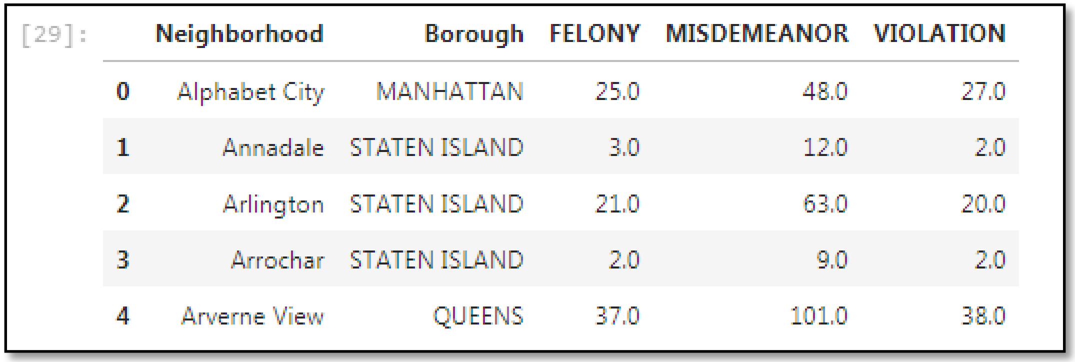
[22]:

	Locations	Neighborhoods
0	(40.65699087900003, -73.87657444799999)	East New York
1	(40.67458330800008, -73.93022154099998)	Eastern Parkway
2	(40.817877907000025, -73.91695668199996)	Melrose
3	(40.75201086000004, -73.93587196099996)	Sunnyside
4	(40.81477097700008, -73.92511075099996)	Mott Haven

Figure 1: Neighborhood dataset generated from GeoPy API

3) Drop the irrelevant columns from the dataset.

4) Prepare a pivot table based on the type of crime. In U.S. the crime is divided into 3 levels i.e. Felony, Misdemeanor and Violation

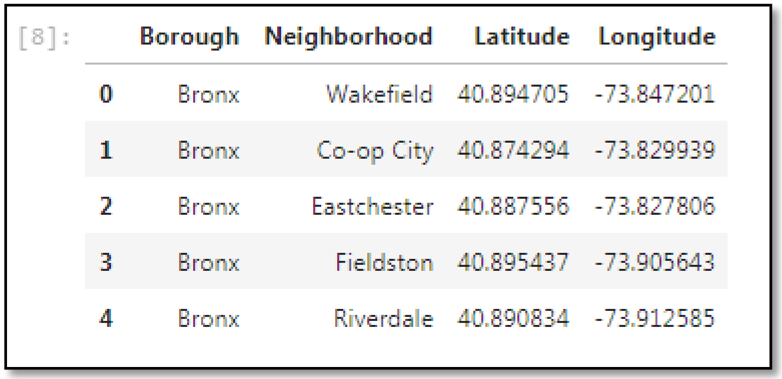


[29]:

	Neighborhood	Borough	FELONY	MISDEMEANOR	VIOLATION
0	Alphabet City	MANHATTAN	25.0	48.0	27.0
1	Annadale	STATEN ISLAND	3.0	12.0	2.0
2	Arlington	STATEN ISLAND	21.0	63.0	20.0
3	Arrochar	STATEN ISLAND	2.0	9.0	2.0
4	Arverne View	QUEENS	37.0	101.0	38.0

Figure 2: Crime data after cleaning and processing

The second data i.e. Neighborhood Data prepared from the geo json file of the New York City. In order to use for analysis we need to convert it into a dataframe.

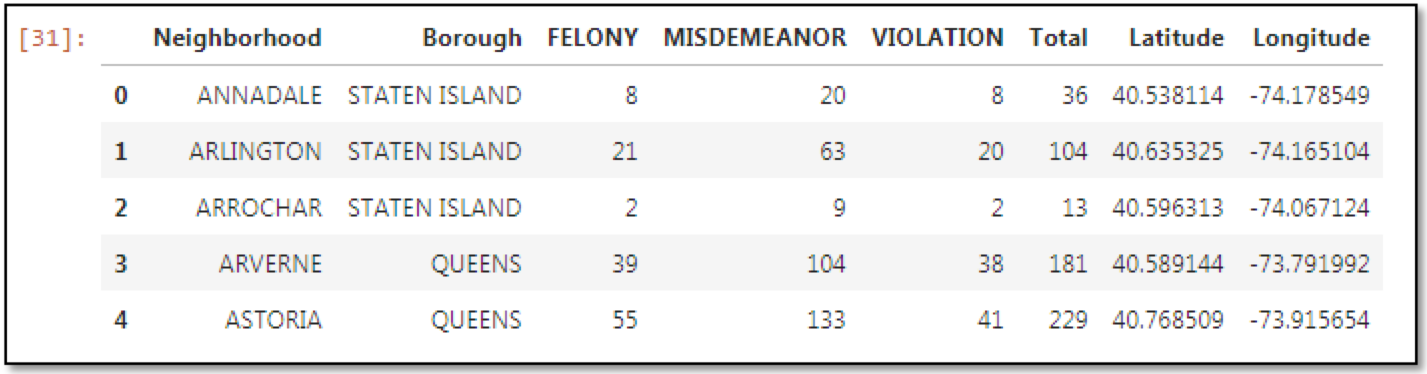


[8]:

	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

Figure 3: Neighborhood data after processing

After this we merge the above Crime dataset with the Neighborhood data we get the following the dataset:

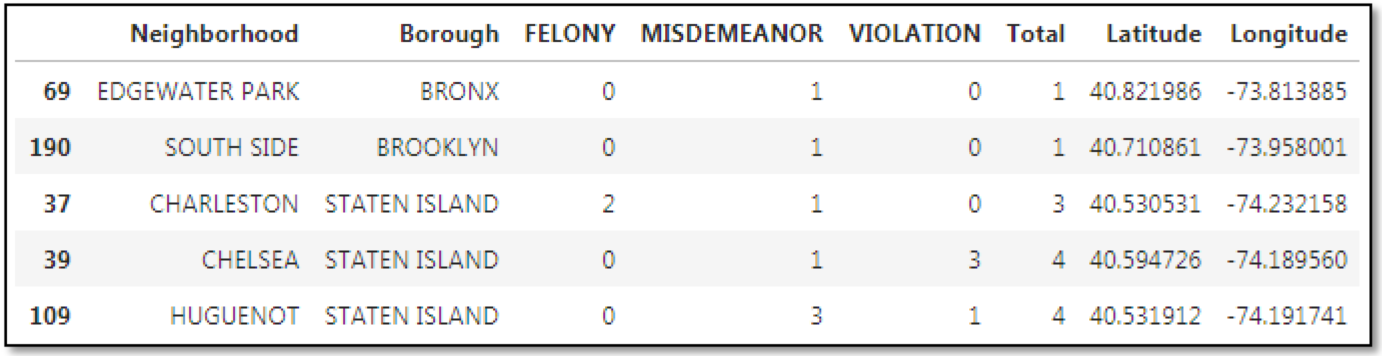


[31]:

	Neighborhood	Borough	FELONY	MISDEMEANOR	VIOLATION	Total	Latitude	Longitude
0	ANNADALE	STATEN ISLAND	8	20	8	36	40.538114	-74.178549
1	ARLINGTON	STATEN ISLAND	21	63	20	104	40.635325	-74.165104
2	ARROCHAR	STATEN ISLAND	2	9	2	13	40.596313	-74.067124
3	ARVERNE	QUEENS	39	104	38	181	40.589144	-73.791992
4	ASTORIA	QUEENS	55	133	41	229	40.768509	-73.915654

Figure 4: Final dataset after merging Crime and Neighborhood data

Using the above dataset we can find the safest neighborhood in New York City which is as follows:



	Neighborhood	Borough	FELONY	MISDEMEANOR	VIOLATION	Total	Latitude	Longitude
69	EDGEWATER PARK	BRONX	0	1	0	1	40.821986	-73.813885
190	SOUTH SIDE	BROOKLYN	0	1	0	1	40.710861	-73.958001
37	CHARLESTON	STATEN ISLAND	2	1	0	3	40.530531	-74.232158
39	CHELSEA	STATEN ISLAND	0	1	3	4	40.594726	-74.189560
109	HUGUENOT	STATEN ISLAND	0	3	1	4	40.531912	-74.191741

Figure 5: Safest neighborhoods

The above data will be used to generate the Venues for each neighborhood using the Foursquare API.