2. Data acquisition and cleaning

2.1 Data sources

● New York City official Incidents Data : **NYPD Complaint Data Current (Year To Date)**
This Data contains a lot of interesting information including the coordinates (longitude, latitude) of each incident which we can exploit using Foursquare API.

● New York City Population Data : **NYC Population by Borough**
This Data will be helpful in calculating the Incidents / Population / Borough

● New York City Census Data : **NYC Census Data**
This Data will allow us to explore The (Demographics - Crime) relation. We will explore this data and extend the results with data acquired from Foursquare API.

● Foursquare API Data : **Foursquare API**
We will be using The Explore feature of Foursquare API to fetch Nearby Venues for each Incidents and analyze the resulting data.

2.2 Data Cleaning :

NYC official Incidents data had a lot of missing values, and some random incidents that didn’t happen in 2020.

The missing values were in both Suspect_Race and Borough features, all the Nans were replaced with ‘Unknown’.

All the Incidents that happened before 2020 were dropped (nearly 5% of the dataset).

Census and Population data were cleaned and merged in the same table, we also added a Population feature to the incidents dataset to be able to calculate the incidents / million / borough rate.

Two Dataframes were extracted from incidents coordinates ‘crimes_by_venue’ and ‘crimes_by_venue_category’, I queried Foursquare for the top 20 venues within 100 meters radius from each incidents and constructed two dataframes one for the crimes / venue and one for the crimes / venue category.

Overall, the data wasn’t very messy and did not require any special Missing Values Engineering.

2.2 Features selection :

NYC official Incidents data contained a lot of columns (35) most of the columns weren't necessary for our project so we dropped all the unneeded features, the Table Below summarize the features selection process:

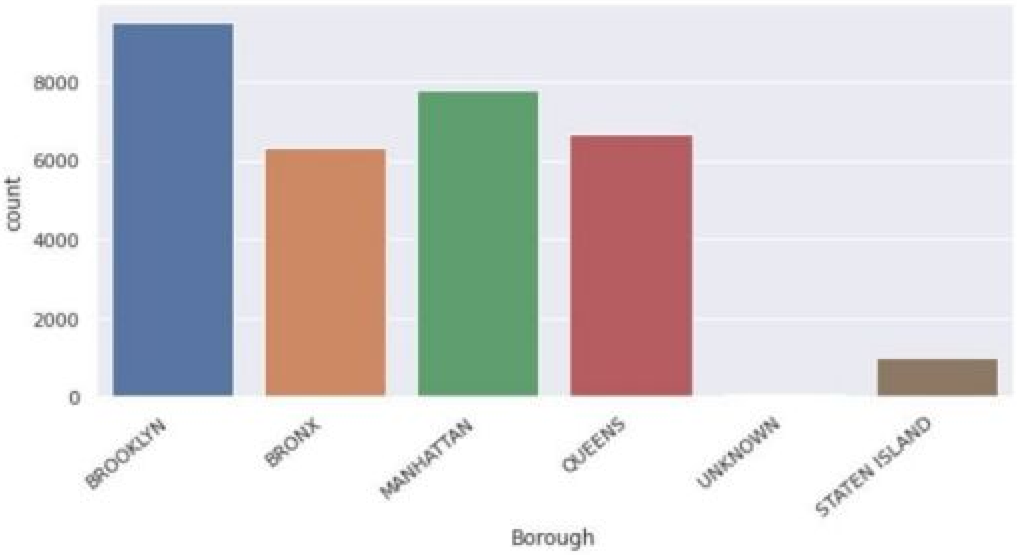
Features	Reason for Dropping
ADDR_PCT_CD, CMPLNT_NUM, CMPLNT_FR_DT, CMPLNT_FR_TM, CRM_ATPT_CPTD_CD, JURIS_DESC , KY_CD, LOC_OF_OCCUR_DESC, PATROL_BORO, PD_CD, SUSP_SEX, SUSP_AGE_GROUP, PREM_TYP_DESC, RPT_DT, X_COORD_CD, Y_COORD_CD.	Irrelevant to our Project.
HADEVELOPT, HOUSING_PSA, JURISDICTION_CODE, PARKS_NM, STATION_NAME, TRANSIT_DISTRICT.	High number of missing values (> 50%)

Kept Featurs	Description
BORO_NM	The name of the borough in which the incident occurred
LAW_CAT_CD	Level of offense: felony, misdemeanor, violation
OFNS_DESC	Description of offense
VIC_AGE_GROUP	Victim's age group
SUSP_RACE	Suspect's race
VIC_SEX	Victim's Sex
Latitude, Longitude	Latitude / Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

3. Exploratory Data Analysis

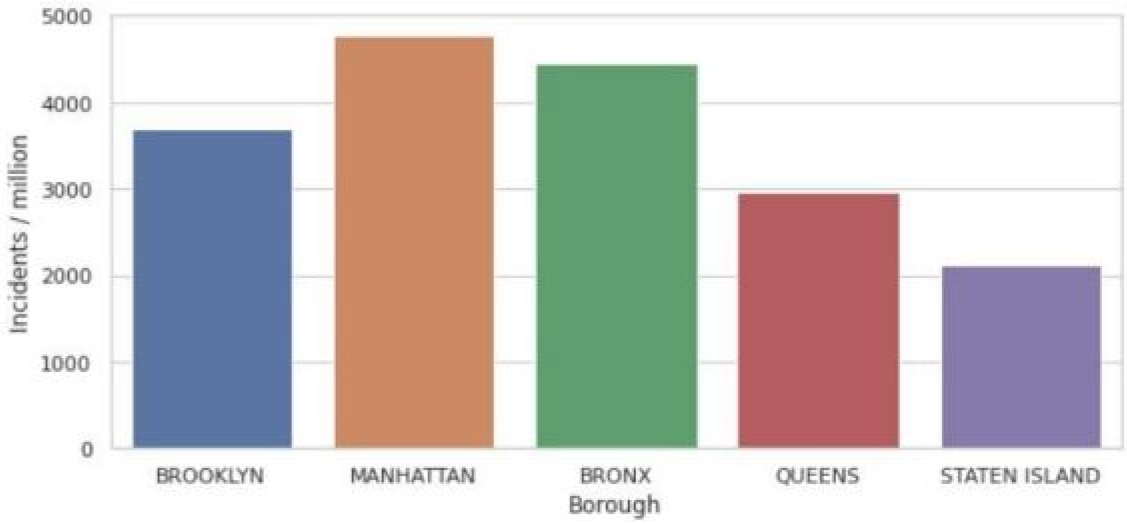
3.1 Crime Distribution across the Boroughs:

We will plot the crime count / borough to get an understanding of the crime distribution across NYC Boroughs



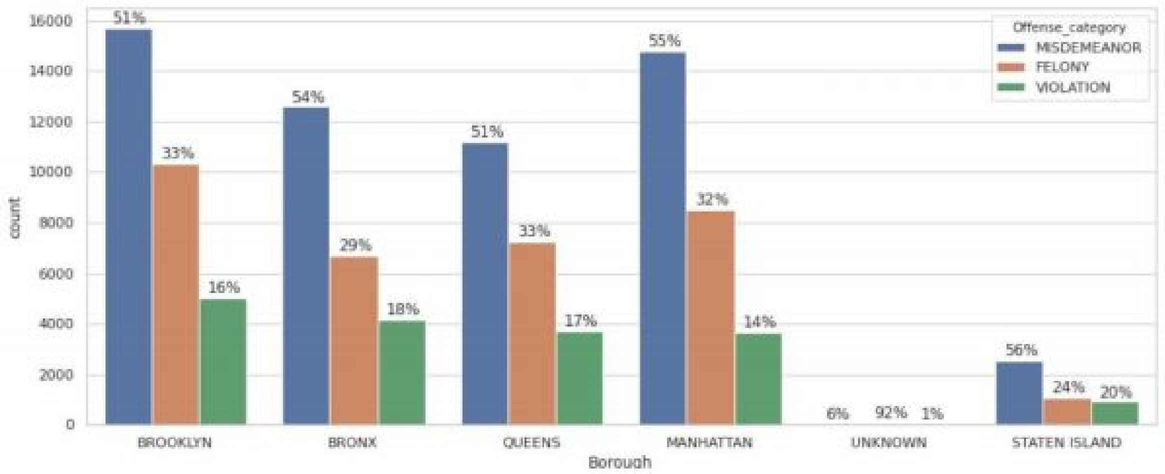
3.2 Crime Distribution across the Boroughs (Incidents / Million) :

Taking into consideration the difference in the population count, it will be more significant if we plot the Incidents / Million..



3.3 Offense Levels Distribution across the Boroughs:

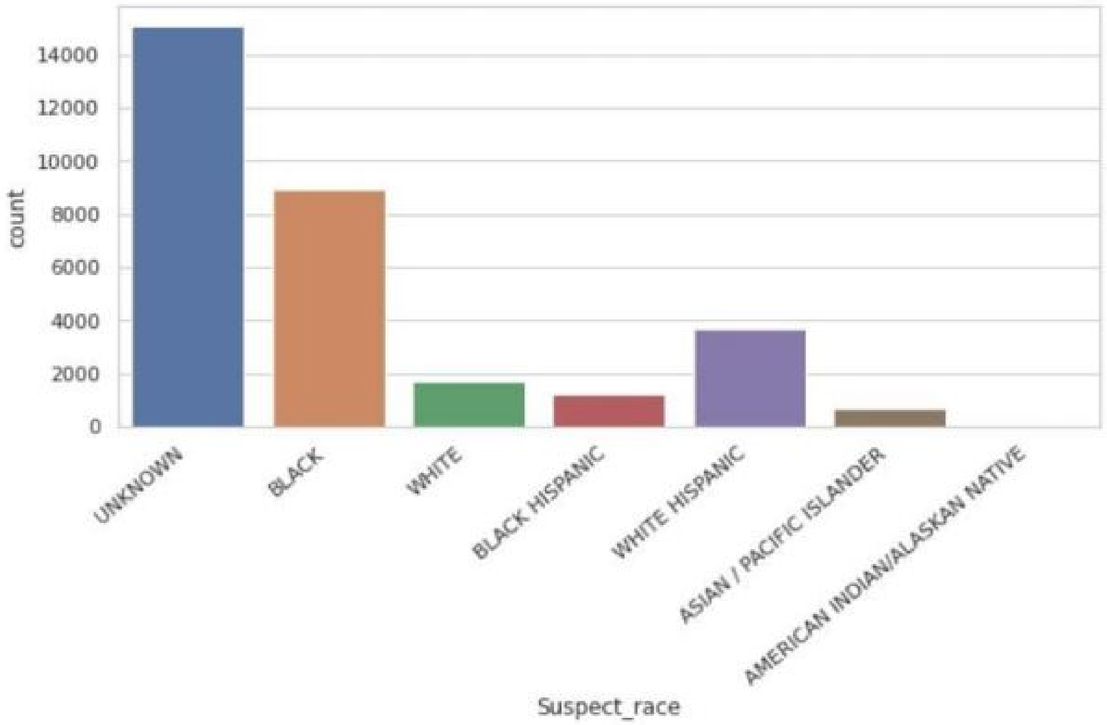
Exploring the distribution of Offense Categories across the borough is a good way to see if some offenses are more common in some boroughs.



- Staten Island has a substantially lower percentage of Felony than the other 4 boroughs (which have about 31%). This could mean that Staten Island is a much more peaceful area, with not only a lower total crime number but also less serious ones. (This also means that the Violation level of crime in Staten Island has a much higher percentage than all the other areas).

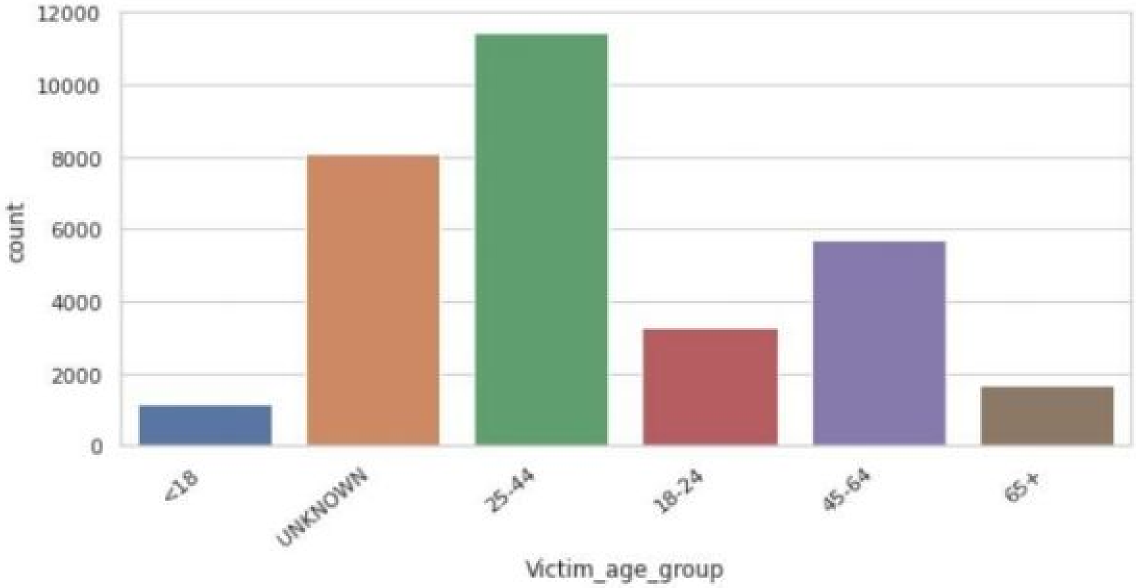
- Bronx also has a smaller percentage of Felony types of crime than the other 3 popular boroughs. This could mean that most of the crime that happens in Bronx is not pressing. This insight can affect the common belief of neighborhood security of Bronx.

3.4 Offenses by Suspect Race :



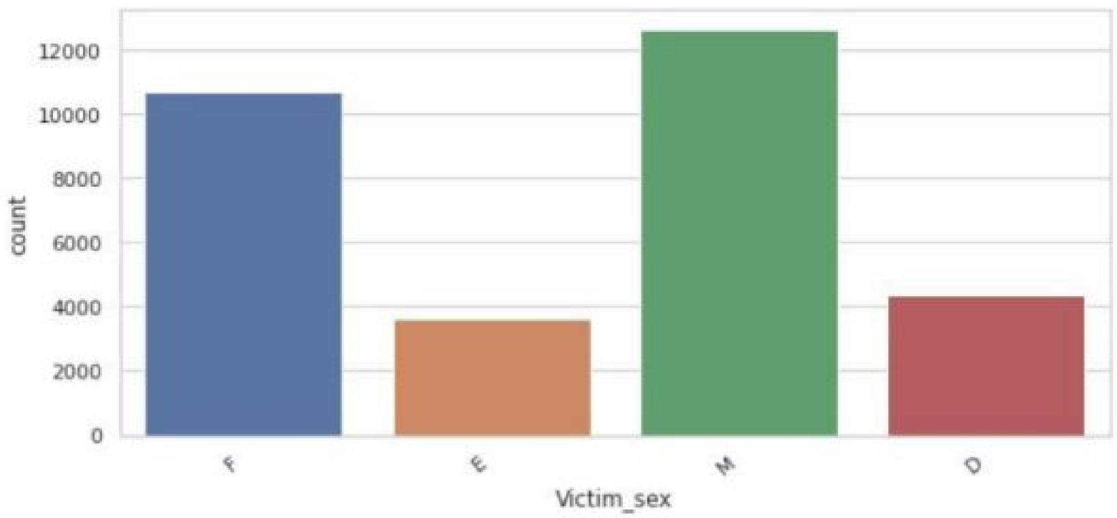
3.5 Offenses by Age Group :

One way to understand the crime dynamics in NYC is to inspect the age groups of the victims.



- Nearly a Third of victims (~11000 / ~31000) are between 25 and 44 years old.

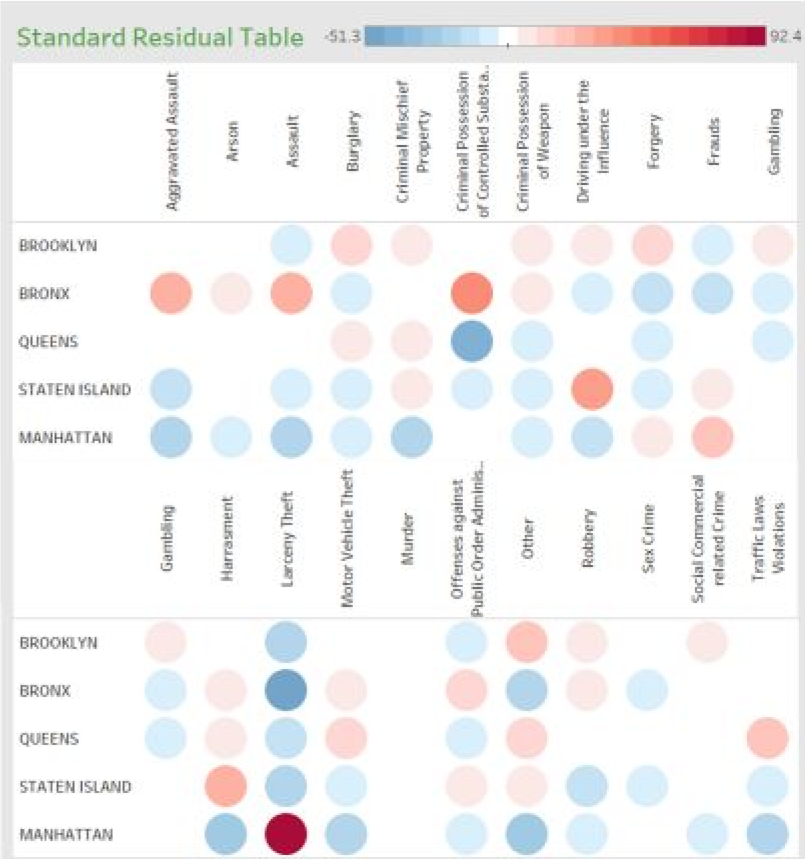
3.6 Offenses by Sex :



(In the Victim_sex feature 'E' and 'D' refers to material objects)

- Victim's sex count doesn't imply much, Male Victims are slightly higher than Females.

3.9 Offense / Borough Standard Residual Table (Plot made in R and Took from Kaggle) :


Standard Residual Table
-51.3
92.4
Aggravated Assault
Arson
Assault
Burglary
Criminal Mischief Property
Criminal Possession of Controlled Substa..
Criminal Possession of Weapon
Driving under the Influence
Forgery
Frauds
Gambling
BROOKLYN
BRONX
QUEENS
STATEN ISLAND
MANHATTAN
Gambling
Harrasment
Larceny Theft
Motor Vehicle Theft
Murder
Offenses against Public Order Adminis..
Other
Robbery
Sex Crime
Social Commercial related Crime
Traffic Laws Violations
BROOKLYN
BRONX
QUEENS
STATEN ISLAND
MANHATTAN