NYC Crime Data Report

Github Link: https://github.com/pdhandha/BigDataProject

DataSet Used: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

Team Member	Net ID
Vinit Jasolia	vkj223
Priyanshi Shah	ps3296
Parita Dhandha	ppd241

Abstract

The goal of this project is to analyze the top reported offenses for each of the five boroughs of the New York City. The first phase of the project deals with finding data quality issues and cleaning the data appropriately. The second phase deals with analysis part where, we intend to explore these trends and discover the reasons behind the increase or decrease in the number of these criminal offenses.

Table of Contents

1. Introduction

The project provides report on analysis of New York City Police Department (NYPD) complaint data provided by NYC Open Data Initiative. The dataset includes all the valid felony and misdemeanor and violation crimes reported the
Our report provides an analysis of NYPD complaint data provided by NYC Open Data Initiative. The NYPD complaint dataset included all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2015). This dataset was organized into 24 categories with over 5 million records.
We analyzed the number of reported criminal offenses with respect to Boroughs that form New York City. To do so, must first check the quality of data to confirm if it is feasible to be used for analysis. After that, we clean the data column-wise to find and remove all the invalid entries.

2. Data Quality

The NYPD complaint history data set was clean with just a few anomalies. To find the anomalies we did the following steps:

1. Looked at the NYPD_Incident_Level_Data_Column_Descriptions.csv and csv file of the actual data to identify the columns and what are the expected type of values in each column.

2. For each column, we first checked the datatype expected and then checked for the validity of the value each row.

3. For some columns like CMPLNT_FR_DT, a missing or null value would be acceptable so it would be tagged as a VALID entry. However for CMPLNT_NUM (complaint number"), a

null value would not be acceptable as it is the primary key of the file. In such cases, the entry would be tagged as INVALID.

1.1. Summary of missing data in each of the 24 columns is presented below:

Column	Missing Data / Total Number of Missing or Null Values
CMPLNT_NUM	0
CMPLNT_FR_DT	655
CMPLNT_FR_TM	48
CMPLNT_TO_DT	1391478
CMPLNT_TO_TM	1387785
RPT_DT	0
KY_CD	0
OFNS_DESC	18840
PD_CD	4574
PD_DESC	4574
CRM_ATPT_CPTD_CD	7
LAW_CAT_CD	0
JURIS_DESC	0
BORO_NM	463
ADDR_PCT_CD	390
LOC_OF_OCCUR_DESC	1127128 and 213 (spaces)
PREM_TYP_DESC	33279

PARKS_NM	5093632
HADEVELOPT	4848026
X_COORD_CD	188146
Y_COORD_CD	188146
Latitude	188146
Longitude	188146
Lat_Lon	188146

1.2. Consistency Issues in some of the useful columns:

1. There were multiple Offense Descriptions (OFNS_DESC) for the same Offense Classification Codes (KY_CD). Examples of where multiplication descriptions existed for same offense classification:
    1. Code **120** corresponds to CHILD ABANDONMENT & ENDAN WELFARE
    2. Code **124** was used for Kidnapping, Kidnapping & Related Offenses, Kidnapping and Related Offenses. This was another **consistency issue** where all descriptions meant the same thing yet were classified into different categories due to syntactic mismatch.
    3. Code **364** pointed to "Agricultural Offenses" as well as 'Other State Laws',

2. As we already know that the data given is from 2006 to 2015, there were some anomalies in **CMPLNT_FR_DT** (complaint from date) column. A lot of dates outside the **1-1-2006 to 12-31-2015** date range. In addition to the 655 missing values, there were **7** values with incorrect Date format or that fell out of range.

3. For Boroughs **(BORO_NM),** the data was consistent and there were no issues apart from the **463** missing values.

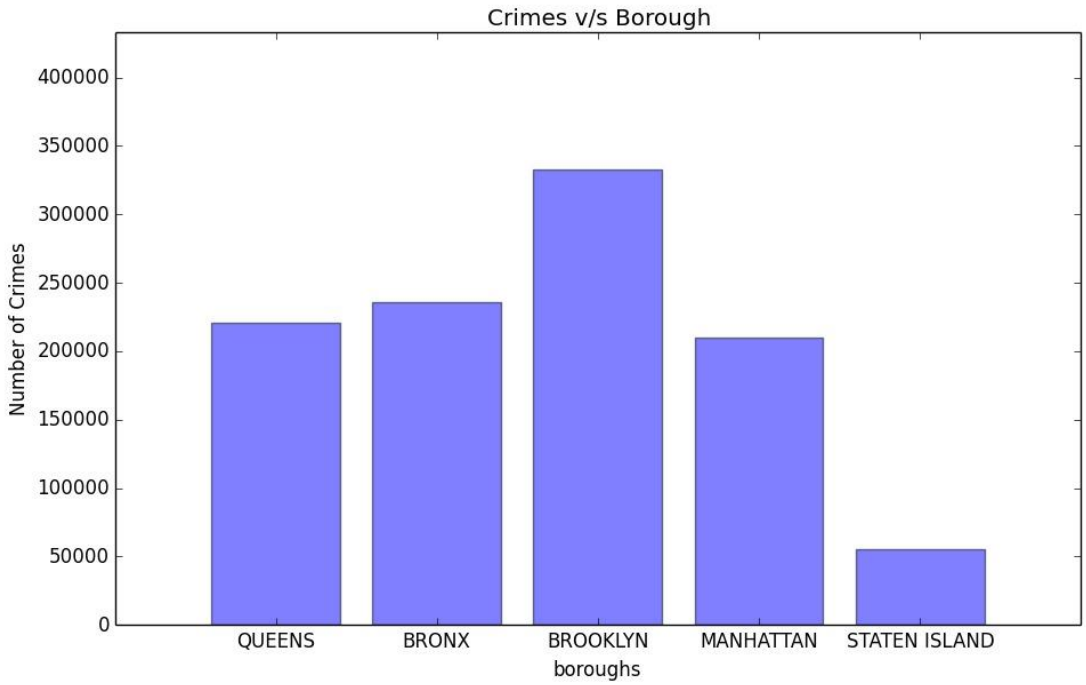
4. **LOC_OF_OCCUR_DESC** (Specific Location of Occurrence in or around the premises) had 213 entries with a 'space' entered: They were technically NULL, but not showing so.

5. In **CMPLNT_FR_TM** and **CMPLNT_TO_TM,** there were time values reported as 24:00:00 whereas they should have been 00:00:00. These entries were marked as INVALID

3. Data Visualization

The following section describes meaningful data visualizations using the NYPD open data.

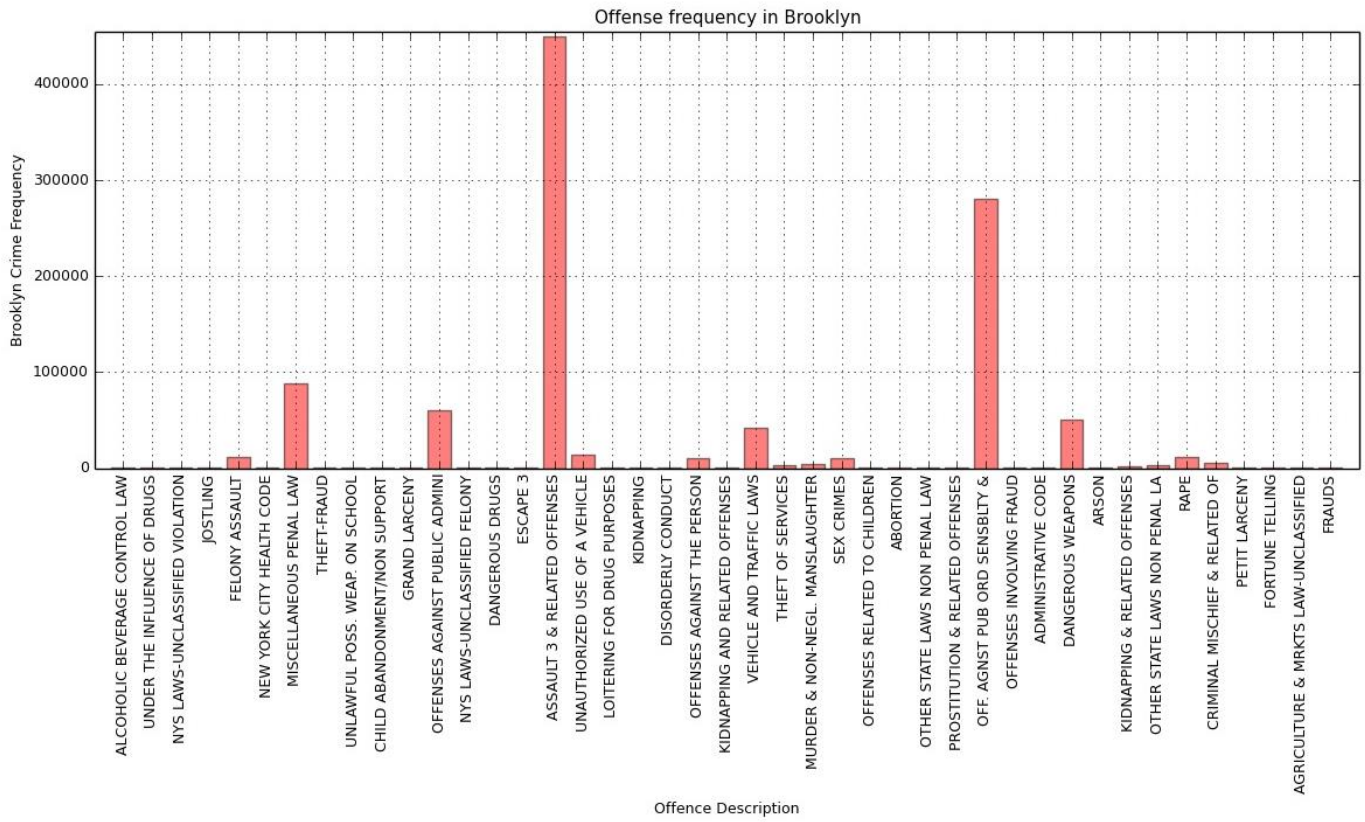
3.1. Analyze the crime frequency in each borough.



Number of crimes in each borough from 2006 to 2015

The above chart tells us that Brooklyn has highest number of crime rate followed by Bronx followed by Queens and Manhattan. The lowest Crime rate is in Staten island.
As Brooklyn has the highest crime rate as per our data source., let's discuss crime in this borough in terms of crime rate over few past years, most frequent in Brooklyn, what times of day are most vulnerable to crime and which type of crime most occurred.

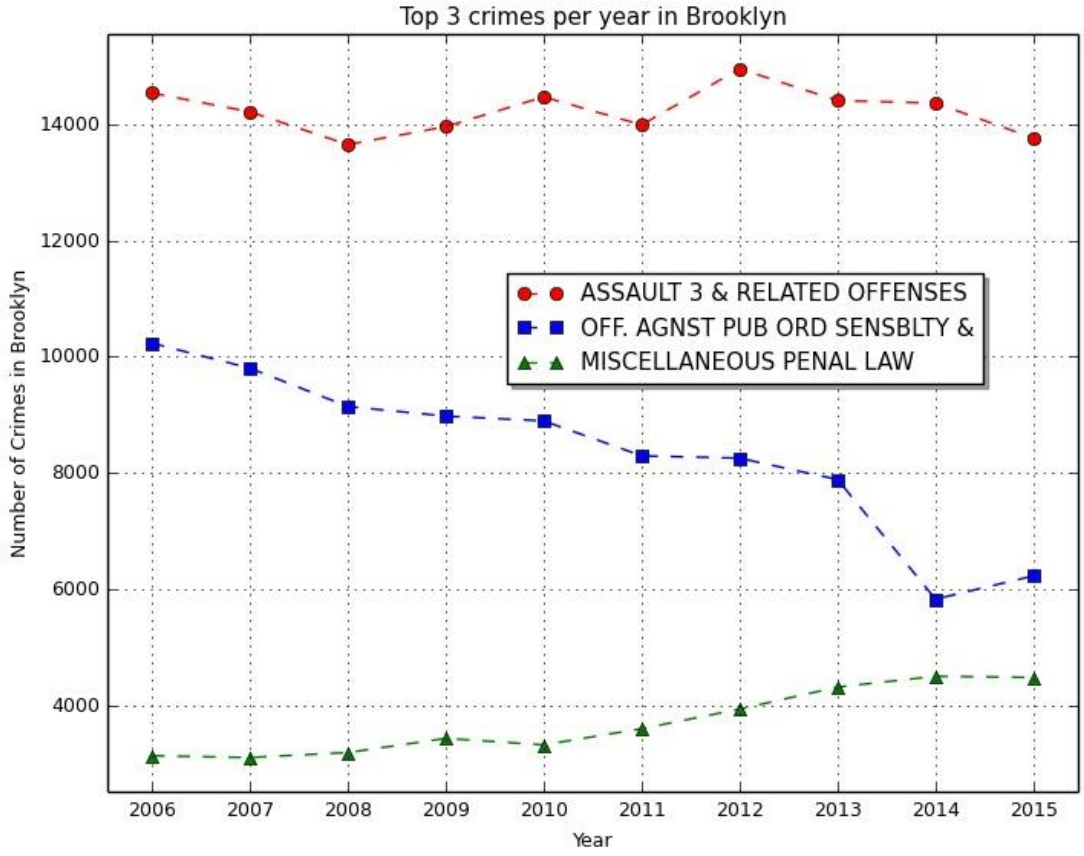
3.2. Top 3 crimes that occurred from 2006 to 2016 in Brooklyn



3.2.1. Offense Description v/s Crime Frequency in Brooklyn

We see that the top 3 crimes are “ASSAULT & RELATED OFFENSES”, “OFFENSE AGAINST PUB ORD SENSBLTY” and “MISCELLANEOUS PENAL LAW OFFENSES”

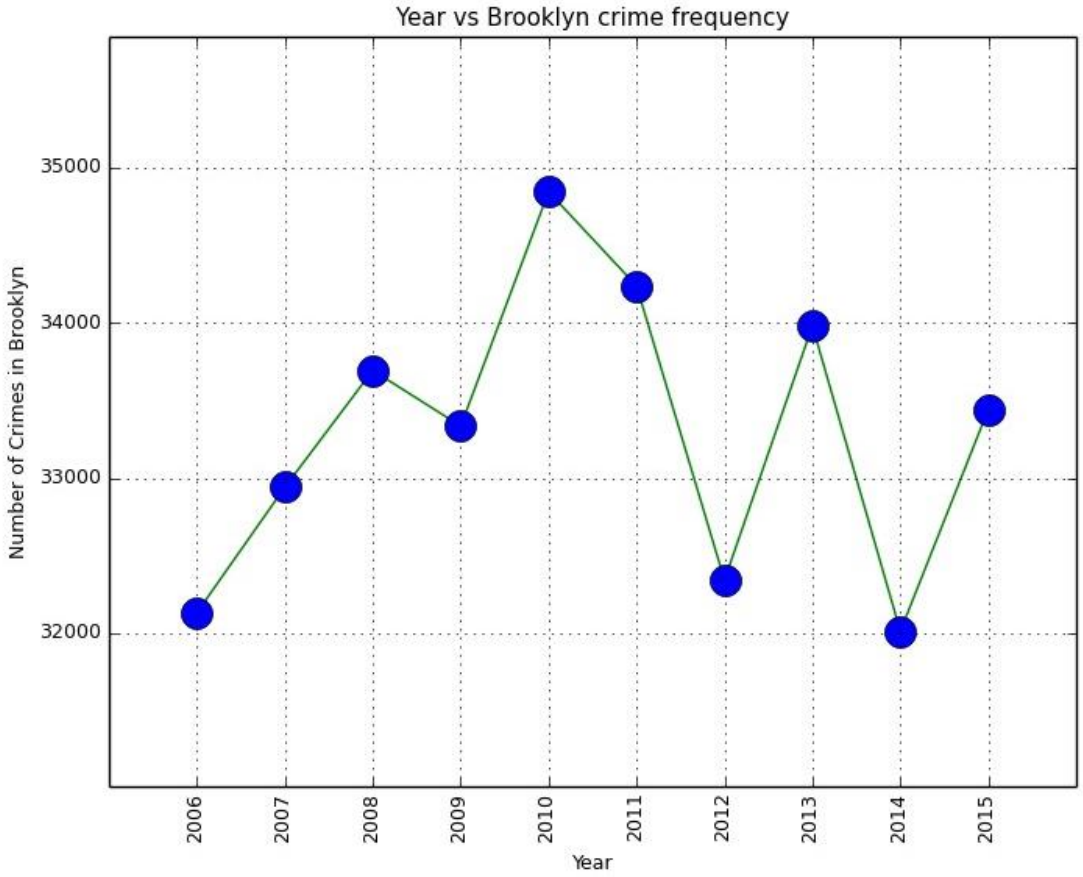
Next, we can analyze the trend of these offenses over the years. That might give some insightful information.



3.2.2. Crime rate of Top 3 Offenses in Brooklyn from 20016 to 2015

There is a clear increasing trend in “MISCELLANEOUS PENAL LAW OFFENSES”. The crime rate goes from approximately 3000 cases in 2016 to 4500 in 2015. The criminal cases for OFFENSE AGAINST PUB ORD SENSBLTY decresases drastically from 10000 in 2006 to 6000 in 2015 and then goes a little higher upto 6200 in 2015. ASSAULT 3 & RELATED OFFENSES is consistently high throughout all the years with some fluctuations.

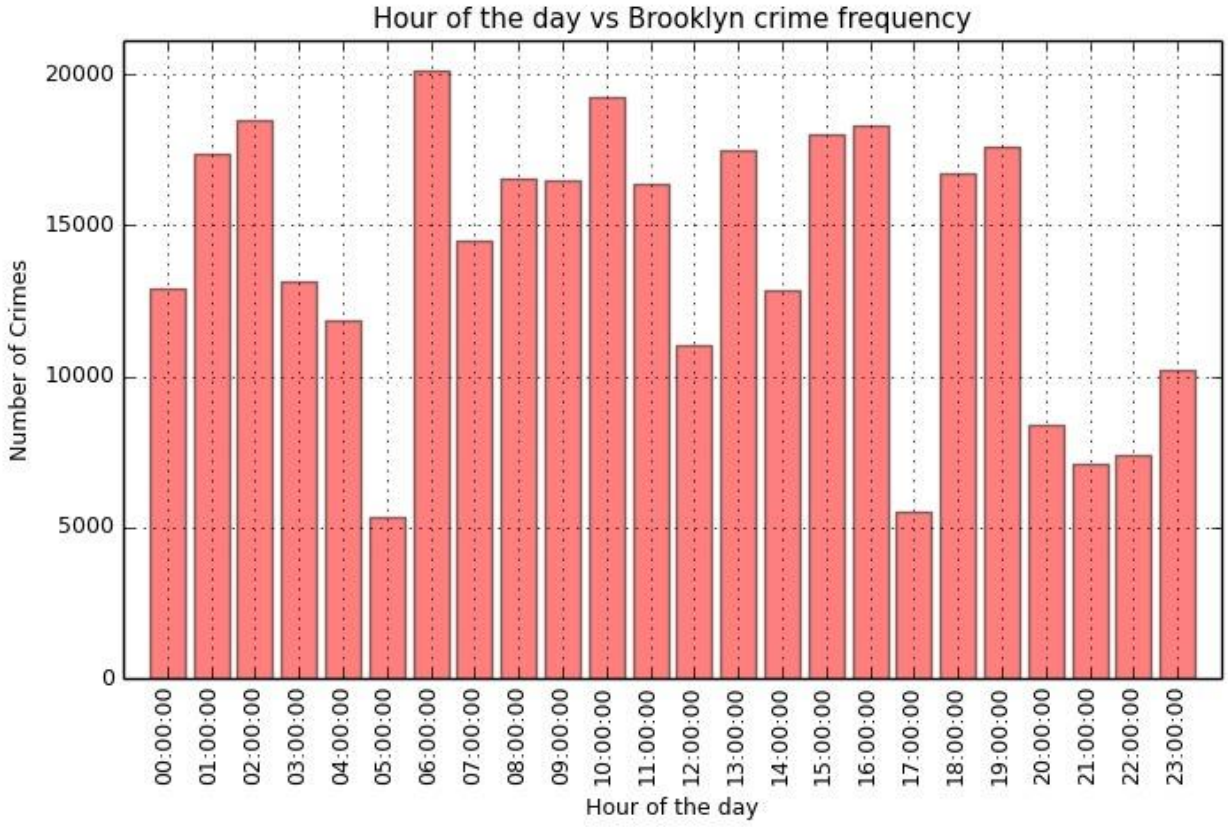
As the above graph gave some useful insights, but only about the crime rate of top 3 crimes. Let us analyze the overall crime rate in Brooklyn from 2006 to 2015



3.2.3. Crime Rate in Brooklyn from 2006 to 2016

The graph clearly shows that the crime rate was at its peak in 2010 after which it decreased significantly through 2011 and 2012 (possibly the highest crime rate alarmed the police and they strove to reduce the crime eventually). However, there is a pattern in the crime rate i.e., it increases and decreased alternately over a period of one or two years.
Since the yearly trend graph was very useful, let is analyze the crime rate on hourly basis in a day.

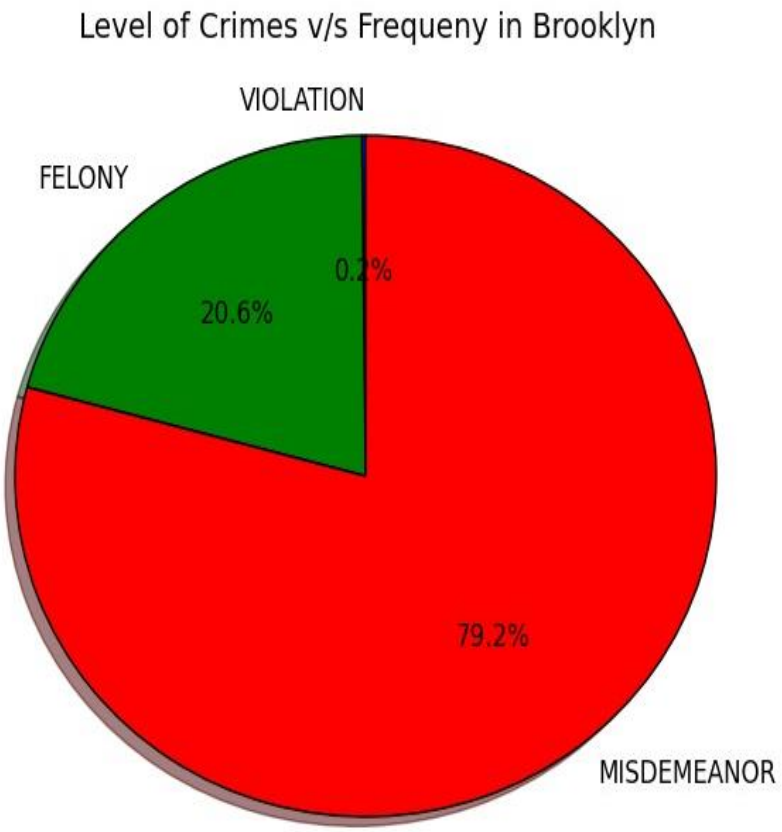
Normally crime rate is higher post-midnight and before sunrise. Let us check if Brooklyn crimes follow the same trend



3.2.4. No. of crimes in Brooklyn on hourly basis in a day from 2006 to 2016

We normally expect crimes to occur after midnight. Our assumption is somewhat correct as the number of crimes is increases suddenly from midnight (00:00:00) to 2am and the decreases gradually from 2am to 5am. Surprisingly the crime rate is pretty high through the day when the sun is out which is not as expected. Hence, it is difficult to draw any certain conclusion from this graph.

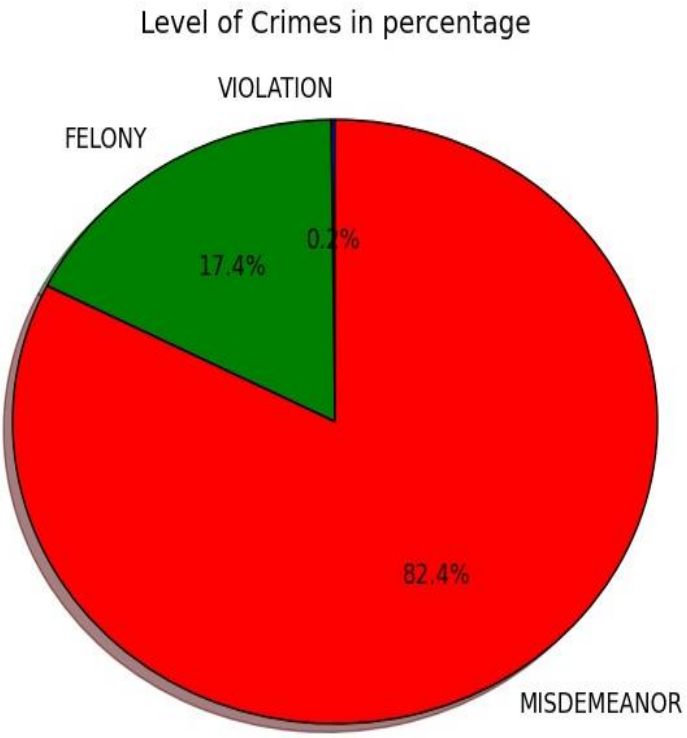
The dataset broadly classifies crimes based on level of offense in 3 categories viz. FELONY, VIOLATION and MISDEMEANOR. Analysis can be done for percentage of each crime based out of whole which may help to take directed actions to reduce them.



3.2.5. Pie chart for Level of offense v/s crime in Brooklyn

MISDEMEANOR is the highest percent of crime in Brooklyn with 79.2% of all crimes and VIOLATION is 0.2% of all the crimes.

3.3.Crime rate v/s Level of offense in overall aggregated crimes in entire New York city

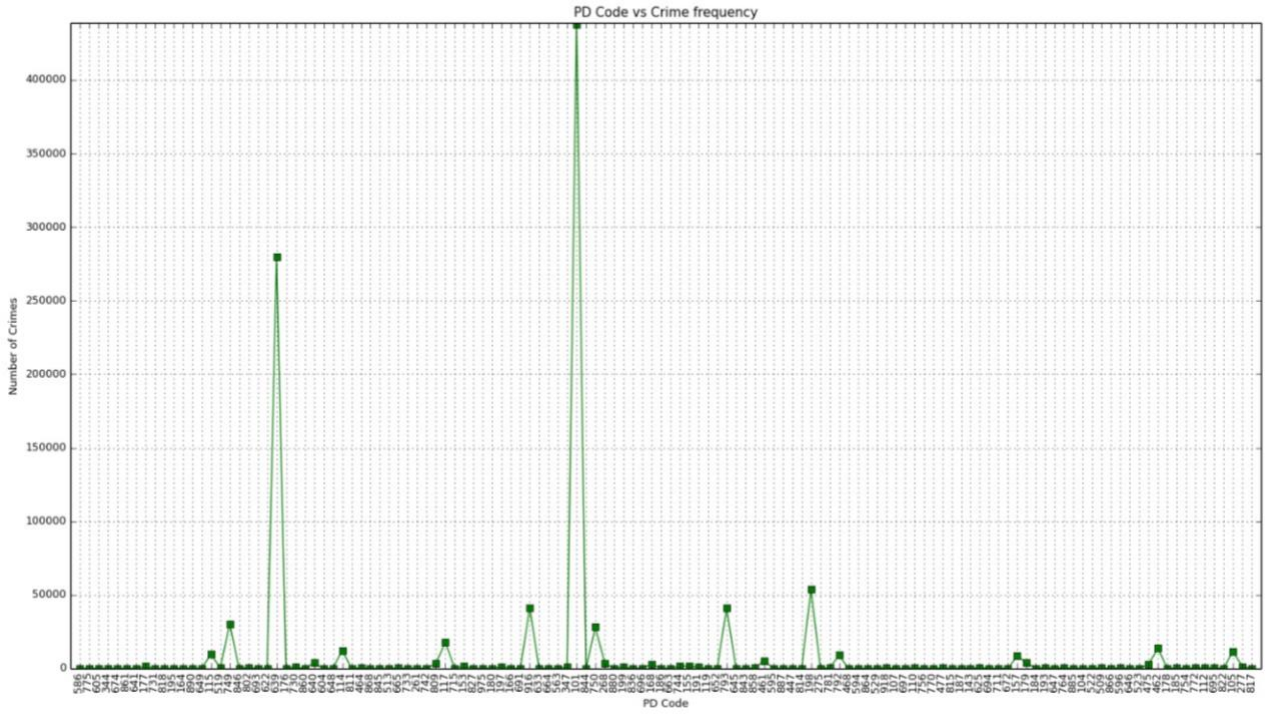


3.3.1. Pie chart for Level of offense v/s crime in NYC

Surprisingly, a similar trend exists for Level of offense in the crimes of all boroughs of NYC aggregated as in the brough of Brooklyn alone. We can conclude that MISDEMEANOR as per law has highest crime rate not only in Brooklyn but overall in NYC.

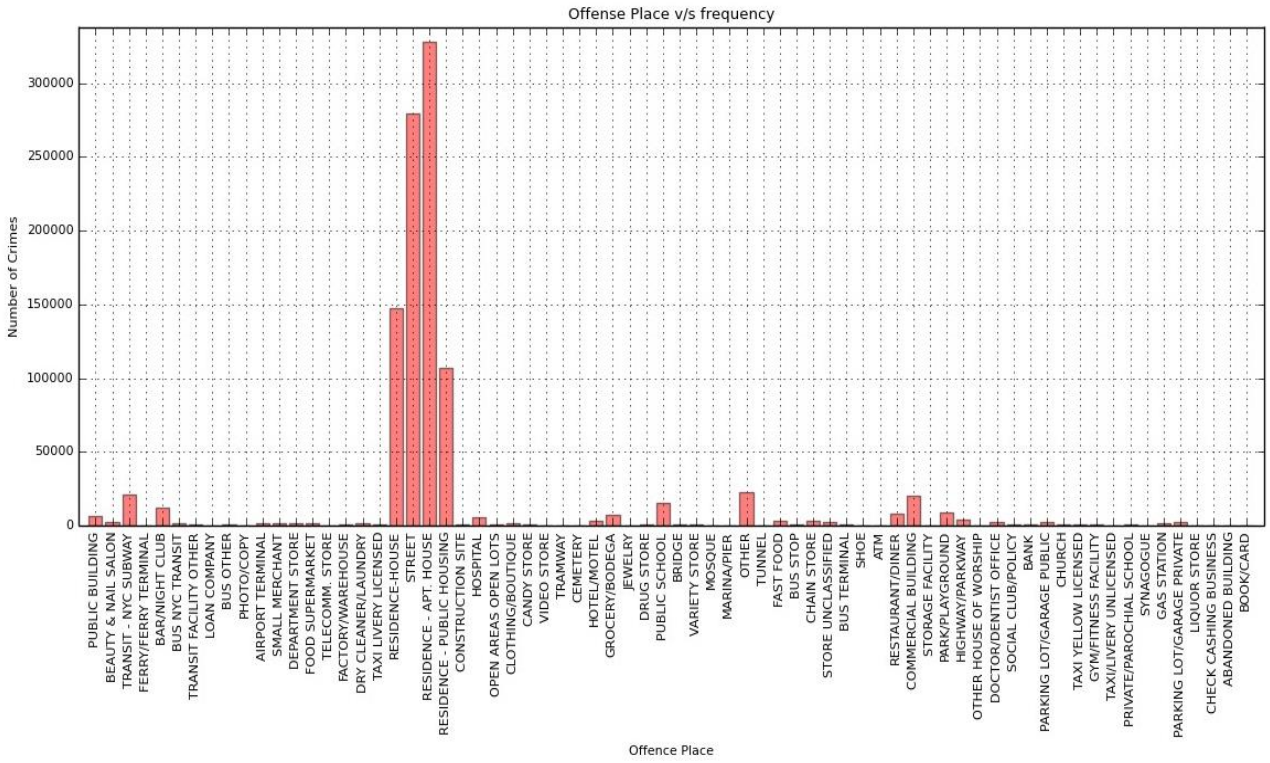
3.4. Crimes in NYC overall on the basis of premises of crimes and crime-rate per crime code.

We have analyzed the Brooklyn borough thoroughly and found a lot of useful information. Let us look at the bigger picture now and analyze the Crimes in NYC overall on the basis of premises of crimes and crime-rate per crime code.



3.4.1. Bar-Chart for crime-rate for each offense-code.

Crime corresponding to offense-code 101 which stands for “MURDER & NON-NEGL. MANSLAUGHTER” has the highest rate in New York City.



3.4.2. Crime rate v/s Crime Premises Bar Chart

From the chart, we can conclude that the most susceptible places for crimes are Residency house, street, apartment house and residence public housing. These places clearly, have very high crime rate in comparison to other premises.

CONCLUSION:

In conclusion, these visualizations gave some insight of various crime patterns. In modern era, Crime Analysis using Big Data and tools, is slowly making its way up the ladder as an efficient method for Crime Prevention and Crime Reduction.

Few hypotheses are easy to deduce but few are almost impossible to deduce and build a pattern on, because of their nature. But some of the interesting conclusions we could draw from the visualization are:

- Majority of crime that occurred belonged to law offense “MISDEMEANOR”
- Residency house, street, apartment house and residence public housing were most prone to crimes in NYC.
- Highest crime rate was in Brooklyn.
- The third most frequent crime in Brooklyn was MISCELLANEOUS PENAL LAW OFFENSES which increased over time.
- Crime rate in Brooklyn showed a consistent pattern of rapid increase and decrease over a period cycle of 2 years from 2006 to 2015.

Looking at past events and learning from recent gathered Data, it is very well understood that we still need to improve skills to Store and Extract Data during an event of crime. It has already been established that crime analysis is essential and useful to police departments and agencies. This will not only help in better analysis but can efficiently and much effectively deduce various new relations between Crime and Other Factors. Thus, reducing the number of Crimes and helping in establish the Innovative, Radical, and Extrapolated approach towards solving Crime.

INDIVIDUAL CONTRIBUTION

Vinit Jasolia:

- Cleaning of Data for Part 1 of the project
- Contributed in report for part 1 of project.

Parita Dhandha:

- Spark Scripts for analysis of overall crime in NYC
- Developed data visualizations to find trends and patterns about place, time, type of crime at NYC-level

Priyanshi Shah:

- Spark Scripts for trend analysis for crime in Brooklyn Borough
- Developed data visualizations to find trends and patterns about top crimes, change in crime rate over time, law category, etc. for crimes in Brooklyn

REFERNCES:

- http://spark.apache.org/documentation.html
- https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-YTD/5uac-w243
- https://matplotlib.org/examples/pie_and_polar_charts/pie_demo_features.html
- https://matplotlib.org/2.0.2/users/pyplot_tutorial.html
- https://docs.scipy.org/doc/
- "Storytelling With Data: A Data Visualization Guide for Business Professionals" by Cole Nussbaumer Knaflic [http://www.storytellingwithdata.com/book/]
- http://www.cnn.com/2013/08/19/justice/new-york-illegal-guns/