# New York City - Crime Analysis
## Big Data Project


## Project Members:

Harshit Srivastava (hs3500), Joby Joy (jj2196) and Pravar Singh (ps3309)


## Introduction:

We have chosen NYPD crime data set ranging from 2006-2016 to apply big-data tools on & generate insights. The dataset is available for download at https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i. We found that the data set was relatively clean, however there were several NULL/Invalid values in almost all columns. The downloadable file - (NYPD_Complaint_Data_Historic.csv) is 1.3 GB and contains 5,580,036 lines. Its sheer size motivated us to use big data tools such as PySpark for performing analysis.

## Part I - Data Cleaning:

The following table contains the details about each of the 23 columns and the corresponding description for it. This description is provided by NYPD and NYC Open Data and can also be downloaded from the page where the data set is found, the description file is named - NYPD_Incident_Level_Data_Column_Descriptions.csv.

| Column | 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_DT | 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_C D | 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_DE SC | 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 green space 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) |
| X_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) |
| Latitude | Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) |

For part I of the project, we used PySpark to analyze the data set . We created a PySpark script that checks all the columns & identifies whether the values adhere to the constraints of their respective columns, for e.g.- values in Latitude column should all correspond to the latitude range of New York City. Similarly, the Borough names (BORO_NM) should be one of the 5 boroughs of New York City. The script then collects the sum of Valid, Invalid & Null values in every column and gives an output like this:

[('VALID', 5111061), ('INVALID', 468319), ('NULL', 656)]

The following table depicts the number of valid, invalid & null values for each column:

| **Column** | **Valid** | **Invalid** | **Null** |
|---|---|---|---|
| CMPLNT_NUM | 5580036 | 0 | 0 |
| CMPLNT_FR_DT | 5111061 | 468319 | 656 |
| CMPLNT_FR_TM | 5579084 | 903 | 49 |
| CMPLNT_TO_DT | 3713616 | 393633 | 1472787 |
| CMPLNT_TO_TM | 4109777 | 1376 | 1468883 |
| RPT_DT | 5101229 | 478806 | 1 |
| KY_CD | 5580035 | | 1 |
| OFNS_DESC | 5561144 | | 18892 |
| PD_CD | 5575126 | | 4910 |
| PD_DESC | 5575127 | | 4909 |

| | | | |
|---|---|---|---|
| CRM_ATPT_CPTD_CD | 5580028 | | 8 |
| LAW_CAT_CD | 5580035 | | 1 |
| JURIS_DESC | 5580036 | | |
| BORO_NM | 5579572 | | 464 |
| ADDR_PCT_CD | 5579645 | 1 | 390 |
| LOC_OF_OCCUR_DESC | 4356430 | | 1223606 |
| PREM_TYP_DESC | 5544838 | | 35198 |
| PARKS_NM | 12539 | | 5567497 |
| HADEVELOPT | 277818 | | 5302218 |
| X_COORD_CD | 5384167 | | 195869 |
| Y_COORD_CD | 5384167 | | 195869 |
| Latitude | 5384167 | | 195869 |
| Longitude | 5384167 | | 195869 |

**Steps taken to clean the data based on the above analysis:**

- Removed all the rows containing invalid/null values for the following columns :

CMPLNT_FR_DT, CMPLNT_FR_TM, RPT_DT, KY_CD, OFNS_DESC, PD_CD, PD_DESC, CRM_ATPT_CPTD_CD, LAW_CAT_CD, BORO_NM, ADDR_PCT_CD, LOC_OF_OCCUR_DESC, PREM_TYP_DESC, X_COORD_CD, Y_COORD_CD, Latitude, Longitude

Since having a Valid non-empty value in these columns is mandatory and crucial for performing analysis in the next phase of the project.

- For CMPLNT_TO_DT, CMPLNT_TO_TM columns : Retained all the rows with null values in both the columns & copied over the values from CMPLNT_FR_DT, CMPLNT_FR_TM columns respectively since the exact date & time of occurrence of the event is known & to date, to time columns are not applicable. Deleted all other rows having Invalid values for either of the 2 columns or having null for just 1 of the columns.
- For PARKS_NM, HADEVELOPT columns : Retained all the rows containing null values for these columns, since they are optional and have values for only a small fraction of the rows.

After performing calculations, we reached a conclusion that < 10% of the total records were pruned from the dataset which also justifies that pruning was the most efficient method rather than trying to convert these values to valid one's since the size of the affected data set is small.

## Part II - Data Exploration:

As part of our Data Exploration strategy, we employed the strategy of using our cleaned data obtained in phase 1 of the project, to be split into several "slices" of the data to help us identify any patterns and/or anomalies. The explorations and their corresponding visualizations were created using Jupyter Notebook, Pandas, and Matplotlib. They can be re-created by running the notebook visualizations.ipynb in our GitHub.

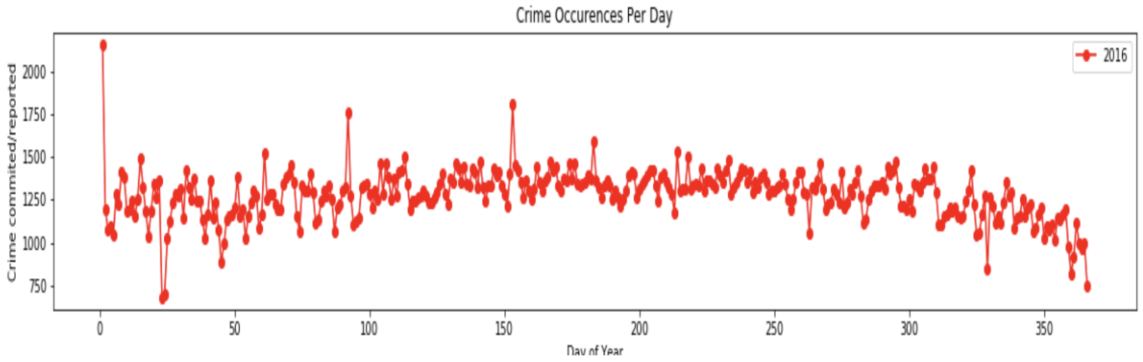
The problem statement's that were addressed as part of the Data Analysis initiatives included:

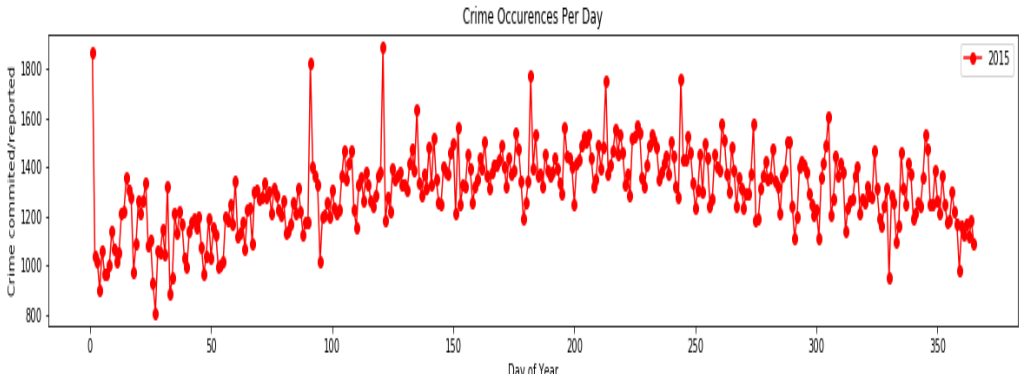
- Total Crime Occurrences across each year in New York City.
- Crime Occurrences across each year per Borough.
- Crime distribution percentage across each Borough every year.
- Crime Occurrences by the Nature of the Crime.
- Crime Occurrences by precinct in each Borough.
- Crime Occurrences by Weather across each Borough.
- Time Series Forecast of Crime Occurrences over the year.
- Average Incidents per day.
- Correlation of the Crime Patterns by Real Estate Demands in Brooklyn.

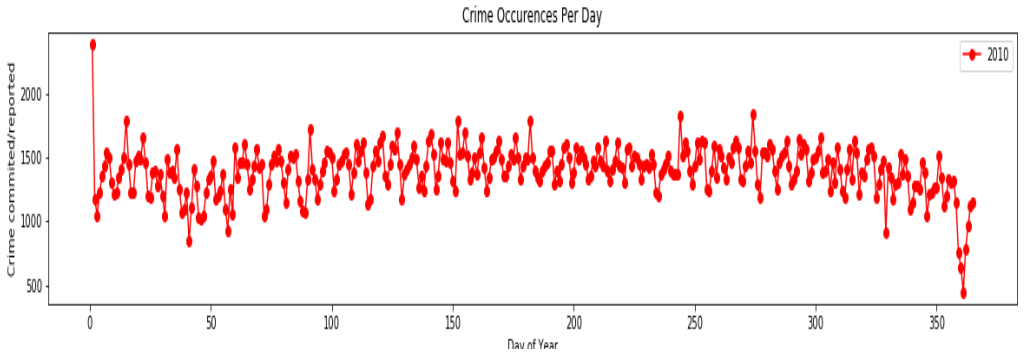
**Observations:**

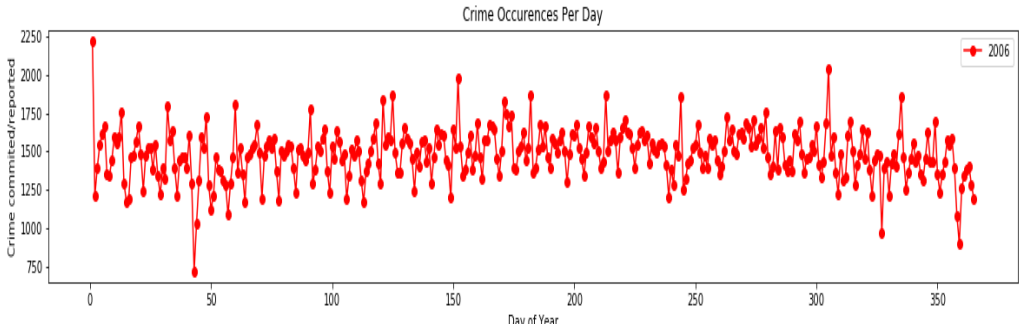
During the course of analysis, our focus was at eliciting the frequency distribution in crime occurrences across time and space, demanding the need to plot each data by year and borough.

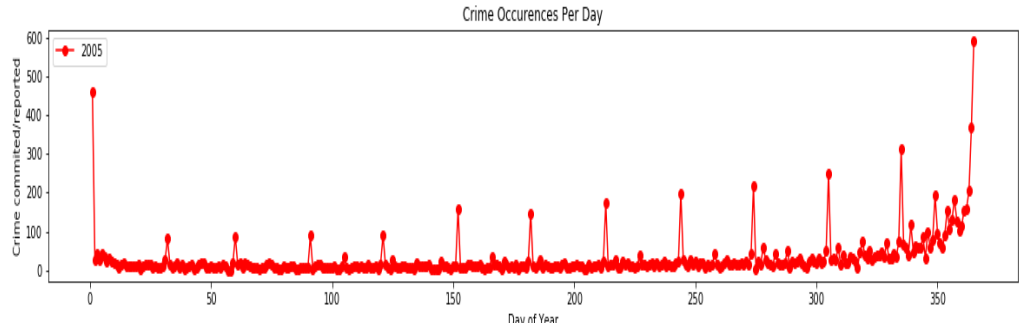
**1: Total Crime Occurrences across each year in New York City**





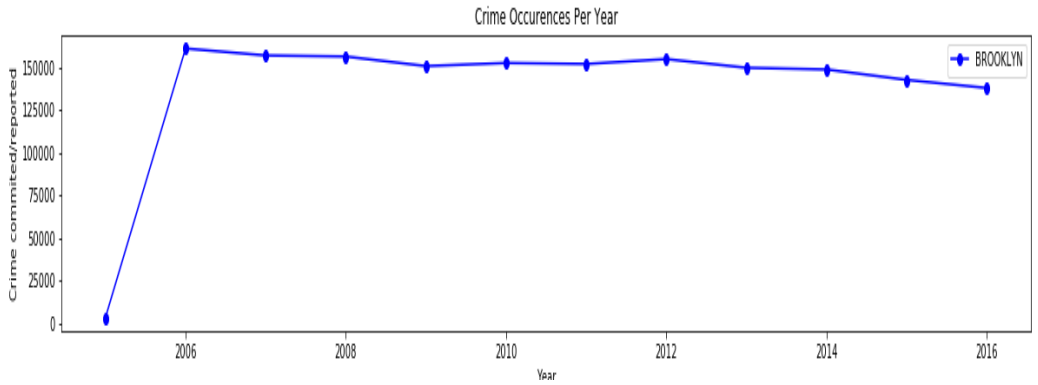



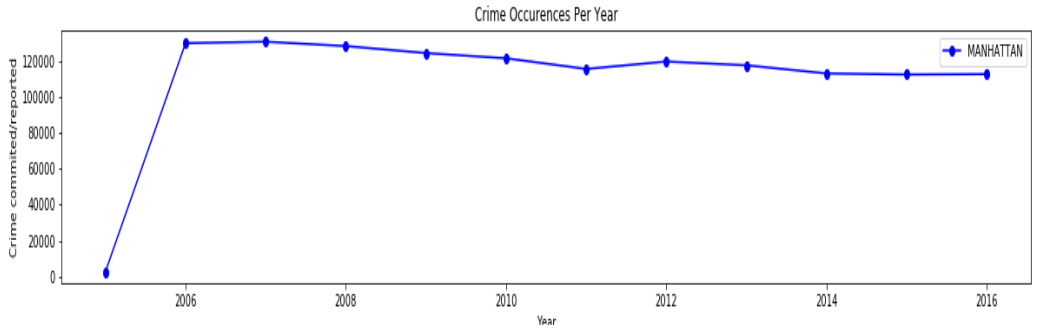

From the graphs, we can infer that the crime count has almost been consistent from 2006 to 2016 across New York city across each year. However strangely we also find that the crime

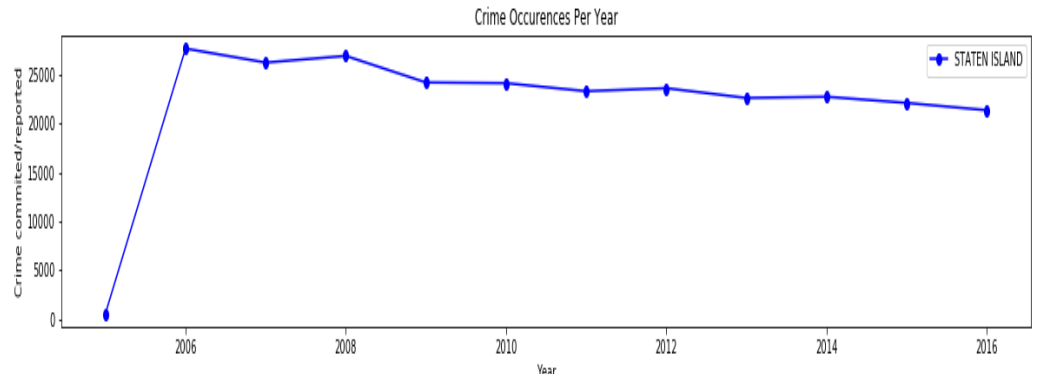
rate in New York was the lowest during 2005 continuing towards an increasing trend only towards the end of the year surmising its due to the holiday season.

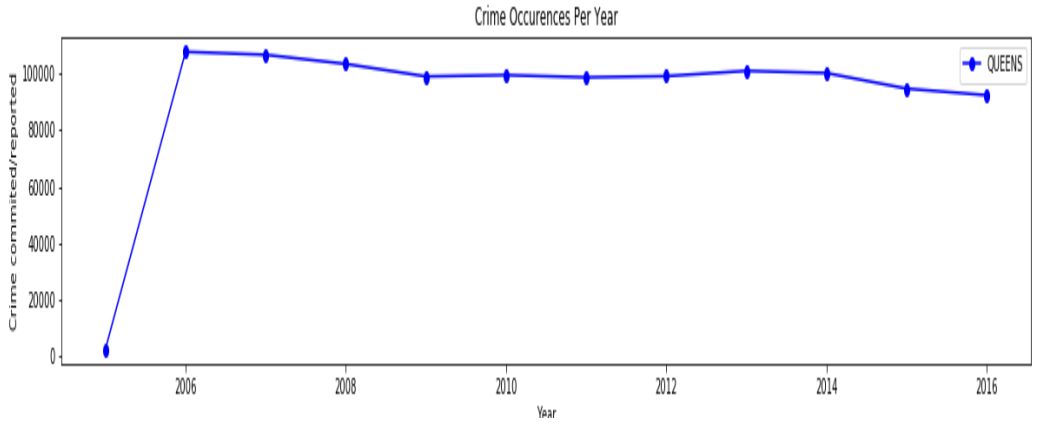
**2: Crime Occurrences across each year per Borough**

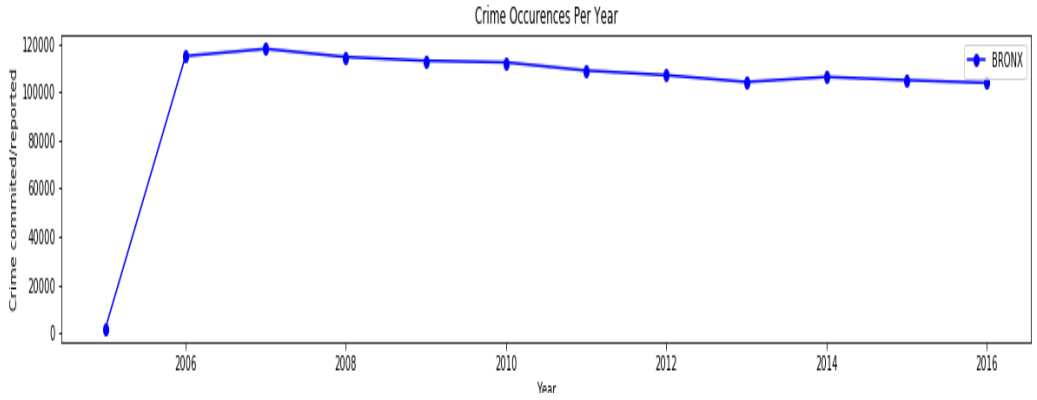
Due to the consistency observed with the Crime patterns across years between 2006 to 2016, we tried to extrapolate the data of Crime distribution across each Borough for every year.



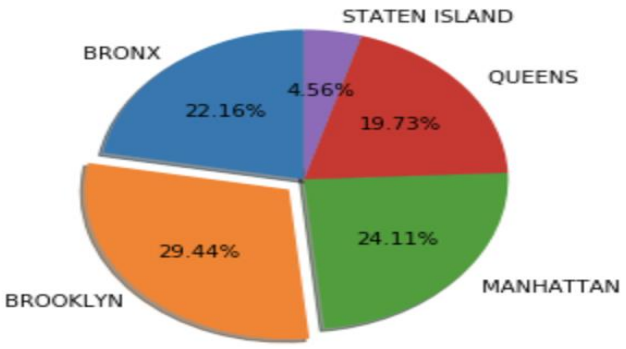





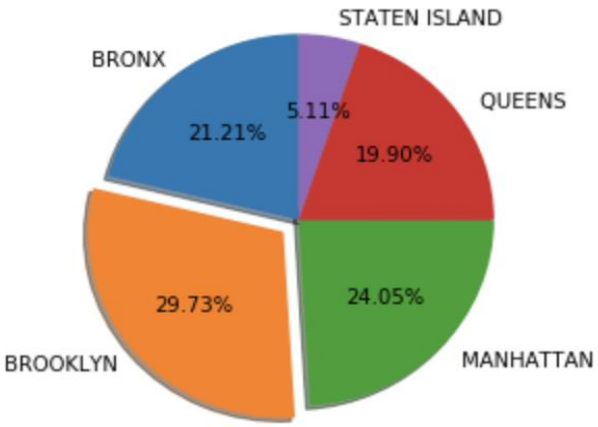




The frequency in the crime occurrences by borough is also observed to be consistent across all the years thus indicating the nature of crime incidences across NewYork for the entire period.

### 3: Crime distribution percentage across each Borough every year




Crime Distribution Per Borough For 2016

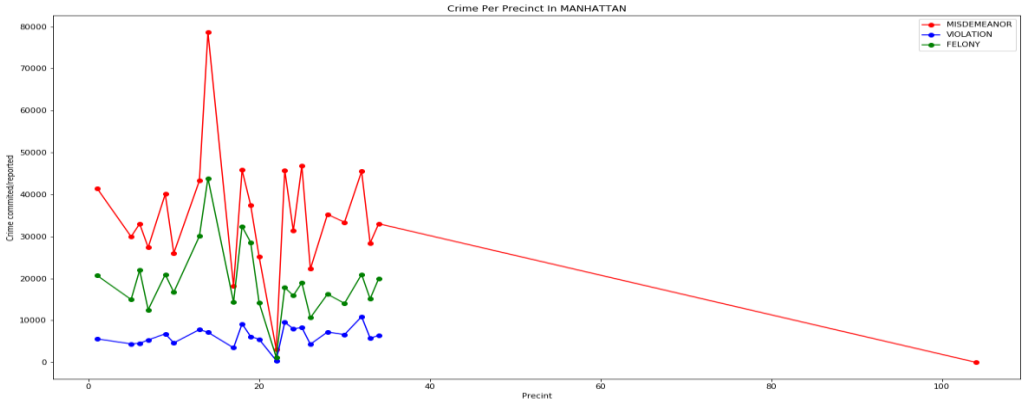


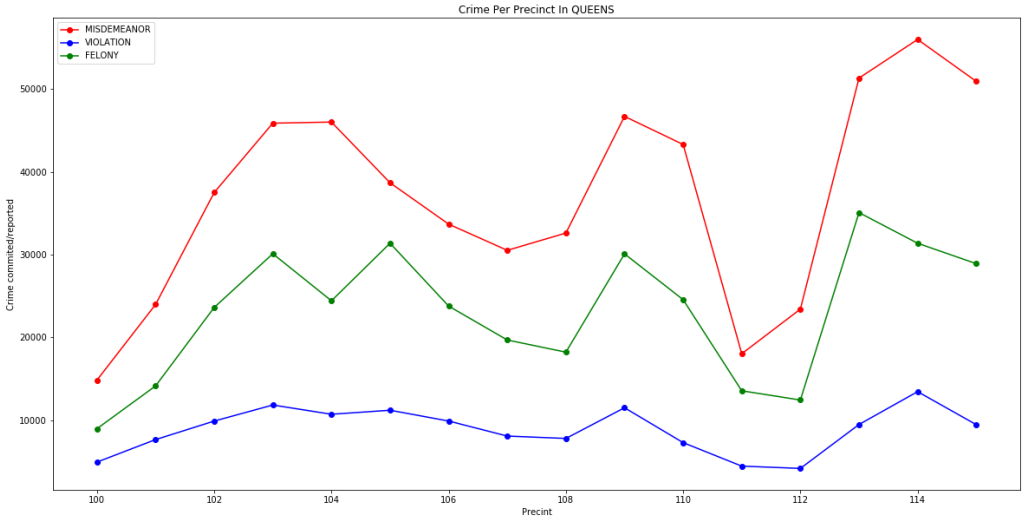

Crime Distribution Per Borough For 2006

The pie chart gives an estimate of the relative distribution of crime occurrences to compare amongst all the Borough's. As observed the maximum crimes are observed to be in Brooklyn which became our basis to do further exploratory analysis between Crime Complaints to Real Estate in Brooklyn.

**4: Crime Occurrences by the Nature of the Crime across Precinct in each Borough**

This endeavor was an attempt to help NYPD to analyze all the busy Precinct's within each Borough's and to share the work load as feasible. The distribution was also made by the nature of Crime so as to indicate the severity of Crimes being worked across the Precinct.
Few plots for the same as illustrated:




From the plot's we see that a precinct in Manhattan rarely gets any incidents reported and also precinct 16 is observed to be overloaded with Misdemeanor's. This will help NYPD to equally distribute the crime load if feasible to the available Precinct's within the Jurisdiction.

### 5: Crime Occurrences by Weather across each Borough

The total count of crimes committed in a day had a seasonal trend in which the crime count went up during the summers and came down during the Winters which implied the Temperature might be playing a big role in Crime incidents as during winters there will be less people on the street to be affected by crime. The possibility of a criminal operating during winters may also be less as a result. So, we decided to combine the Weather dataset from National Oceanic and Atmospheric Administration (NAOO) for the year 2016 and did a Spearman Test on both the datasets to find out Rank Order Correlation between the two datasets for the year of 2016. We got the following values:

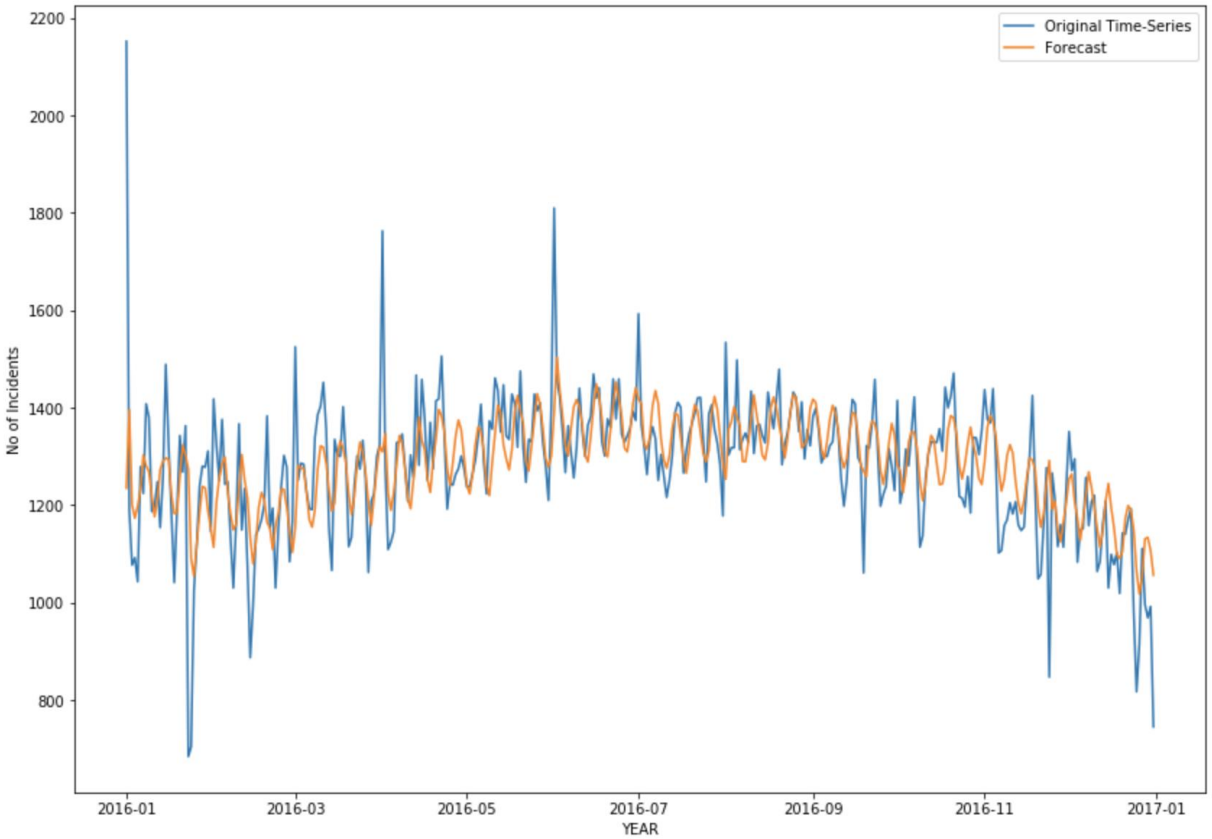
*Sample Size: 365*
*Correlation Coefficient: 0.608*
*P-Value: 2.1195302026e-38*

The Null Hypothesis was 'The 2 datasets are not connected' but after looking at the results, we can safely reject the Null Hypothesis with 95% confidence that the 2 datasets are in fact positively correlated with each other.

**6. Time Series Forecast of Crime Occurrences over the year**

We collected the count of Incidents in a day and mapped it to a particular date. This way we had a count of all Incidents that happened in a day since 2006. As this was a Time-Series, we began Time-Series forecasting to forecast the crime count for the year of 2016 and then we checked the original data to know how accurate the model's performance was. The Time-Series was stationary and it passed the Dickey-Fuller Test of stationarity implying a greater than 95% probability of Time-Series being stationary. We tried many values and got the best result when using the following parameters:

Auto-Regressive Component (AR) – 5
Moving Average Component (MA) – 3

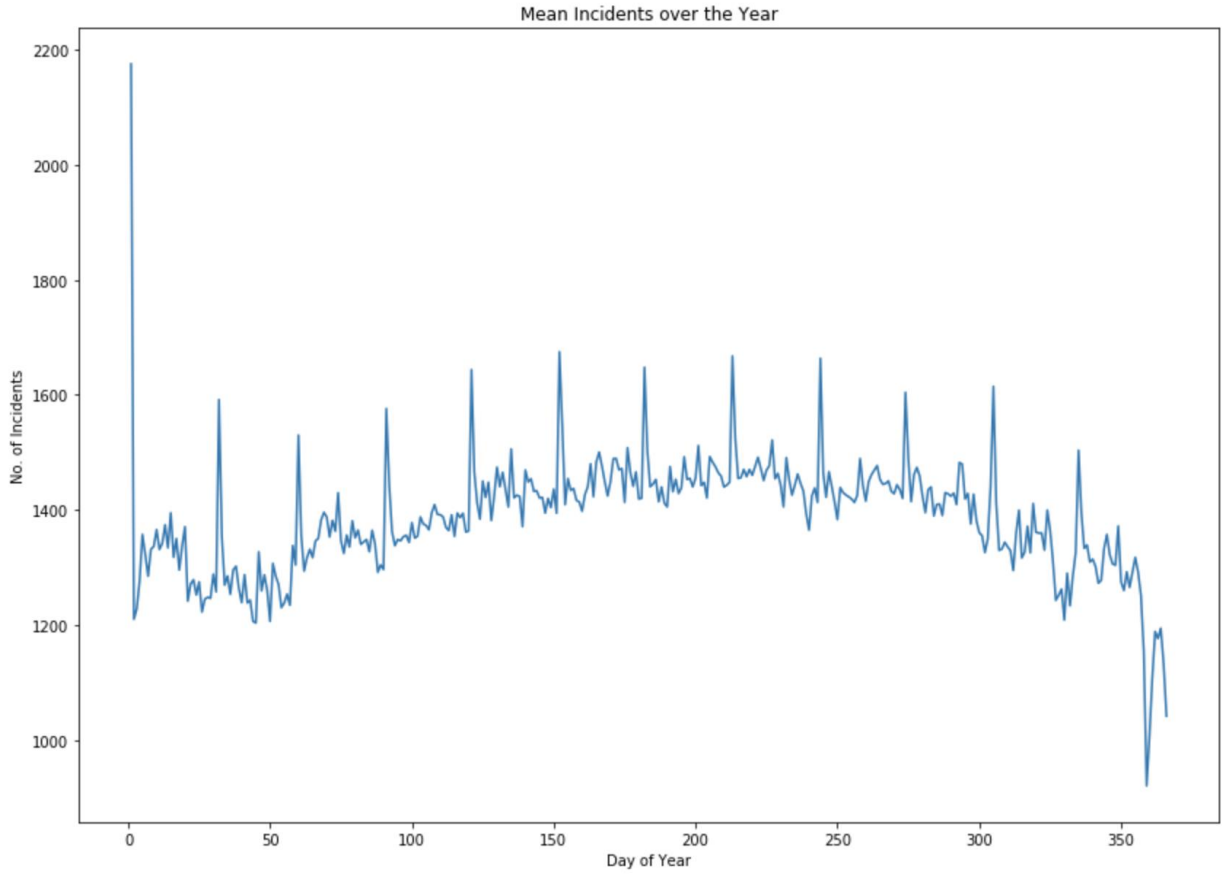


The evaluation of the model was done based on Mean Absolute Error and Mean Forecast Error. The values we obtained for the above parameters are:
Mean Absolute Error: 69.1
Mean Forecast Error: -0.45

The model was able to capture the cyclic trend that is going on, however, it is not able to capture the exact values as there is high variance in the data. Mean Forecast Error of -0.45 suggests that the Prediction was slightly over-estimating the number of incidents as compared to the actual values. This maybe because on an average there were less incidents in 2016 as compared to previous years. MFE is close to 0 which suggests that there is no bias in the prediction i.e. the Incident count for the whole year is roughly equal to the actual count but the predictions for each day is somewhat inaccurate considering the high variance in data.

## 7. Average Incidents in a day

We could see a pattern where the first day of every year had a high number of incidents occurring. Christmas week of every year had less incident count as compared to other days. We decided to aggregate data over the 11 years and take the mean number of incidents happening on each day of every month i.e. we initially had 11 values for January 1st and we took the mean of all those values so that the date January 1st can be represented by 1 single count of Incidents. This way we had Incident count for each day of the year and the plot looked like this:



We can clearly see a sharp rise at the first day of the Year which implies the 1st of January is filled with incidents. The overall Incident count goes up during the summers. During the Christmas week, there is a substantial decline in the number of Incidents occurring which implies that Christmas time is safest in New York.

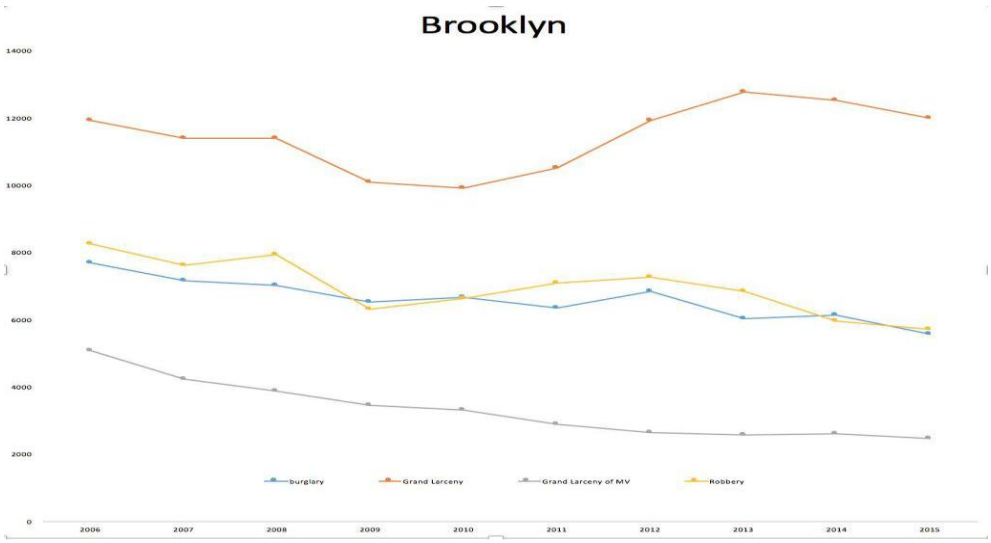
## Part 2: Exploring the Relationship Between Crime Patterns and Real-Estate Demand in Brooklyn

After analyzing the different visualizations of borough-level slices of crime complaints over the last 10 years, we decided to correlate the Crime data with the Real Estate Data in Brooklyn. Our first level of analysis was performed using Brooklyn as we observed that Brooklyn has the highest Crime rate and also there were consistencies observed over Crime type, year and Borough for the last 9 years. Another reasoning that governs our analysis specifically to Brooklyn is that as residents of NYC with friends flocking to the now "hipster" neighborhoods of Brooklyn, it seemed likely that an increase in the safety of valuable items in a given area would go hand in hand with an increase in demand for real-estate in the area.

As part of the analysis, we strategized our exploration by mapping the X and Y coordinates of every crime incident to their corresponding neighborhood. Subsequently we merged this data with the data from real estate

The following plot shows the count of complaints across all of Brooklyn for felonies in the categories of robbery, burglary, grand larceny and grand larceny of motor vehicles.

1. **Number of Felonies in Brooklyn by Category**

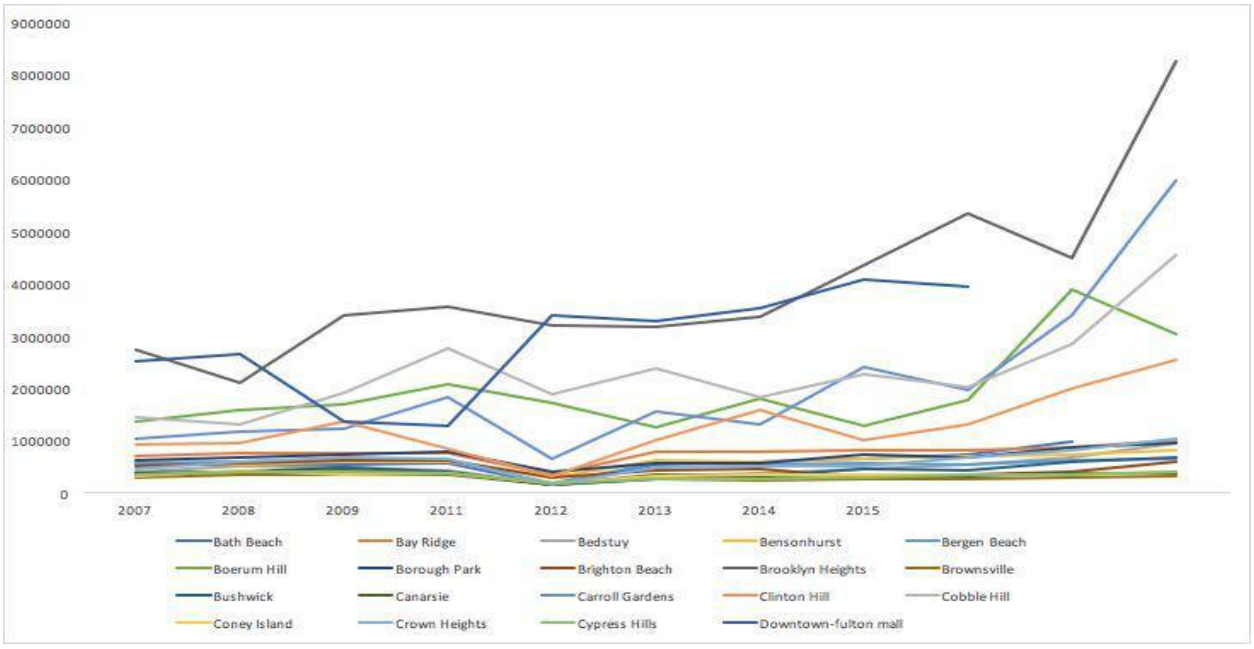


For data on real-estate interest and cost shifts, which we thought should have an impact on crime rates, we turned to the Rolling Sales NYC archives, which break down prices by borough for varying types of sales.

### 2: Real Estate Prices in Brooklyn


We also examined housing prices at the neighborhood-level. The chart below shows the breakdown in real-estate purchases for a sample of Brooklyn neighborhoods.

### 3: Real Estate Prices in Brooklyn by Neighborhood



As evident from the neighborhood breakdown above, the variance can make it seem like Brooklyn contains many cities within itself. We wanted to take a closer look and pinpoint different areas in Brooklyn for crime and real-estate costs. We chose Bedford-Stuyvesant to investigate. This involved using the geopandas library to map shapefiles, available from https://www1.nyc.gov/site/planning/data-maps/open-data/bytes-archive.page,and neighborhood polygons to the city X-Y coordinates to enhance our crime dataset with a neighborhood column. Since the geopandas library wasn't available on dumbo, we broke the files into smaller slices (which can be found in github directory under data/bk_slice) and enhanced them locally before using PySpark to recalculate our counts, this time incorporating neighborhood into our key.

**Table 3: Bedford-Stuyvesant - Average Price of all Real-Estate Sales and Count of Burglary, Robbery, Grand Larceny and Grand Larceny of Motor Vehicles (BRGL)**

**Bedford-Stuyvesant**

10 Year Increase in Sale Cost: **61%**

10 Year Decrease in BRGL: **15%**

| Year | Avg Sale Cost | Count of BRGL |
|---|---|---|
| 2006 | $605,831.34 | 2,053 |
| 2007 | $599,073.63 | 2,133 |
| 2008 | $536,218.40 | 1,957 |
| 2009 | $414,480.32 | 1,781 |
| 2010 | $373,402.03 | 1,915 |
| 2011 | $419,655.64 | 1,978 |
| 2012 | $481,524.01 | 2,086 |
| 2013 | $626,258.19 | 1,812 |

| | | |
|---|---|---|
| 2014 | $804,190.59 | 1,898 |
| 2015 | $ 978,138.33 | 1,750 |

A correlation between real estate prices and the rates of burglary, robbery, and larceny can be seen from our analysis. The up-and-coming neighborhood of **Bedford-Stuyvesant** has seen an increase in real-estate costs as the worries of burglary, robbery, and larceny have decreased over the last decade.

**Individual Contributions**

All team members contributed to brainstorming ideas, analyzing data, interpreting results, and writing. For Part 1, Joby and Harshit divided the columns and wrote the routinues to check the validity of the data. Pravar worked on the scripts to filter the invalid data and generate the cleaned dataset. For Part 2, all the use cases were divided amongst the team members equally. The work involved developing scripts to create data subsets, visualization and exploring additional datasets for deeper analysis.

**Github Repository Link:**

https://github.com/harshit0511/NYC-Crime

**References (Data Sources)**


Crime data:

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

Real-estate:

- "Rolling Sales Data" from www1.nyc.gov

Neighborhood shapefiles:

- "Neighborhood Tabulation Areas" from https://www1.nyc.gov/site/planning/data-maps/open-data/bytes-archive.page


**Acknowledgements**


We'd like to thank Professor Claudio Silva.