**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 |
|  | Exact date of occurrence for the reported event (or starting date of |
| CMPLNT\_FR\_DT | occurrence, if CMPLNT\_TO\_DT exists) |
|  | Exact time of occurrence for the reported event (or starting time of |
| CMPLNT\_FR\_DT | occurrence, if CMPLNT\_TO\_TM exists) |
|  | Ending date of occurrence for the reported event, if exact time of |
| CMPLNT\_TO\_DT | occurrence is unknown |
|  | Ending time of occurrence for the reported event, if exact time of |
| CMPLNT\_TO\_TM | 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) |
|  | Description of internal classification corresponding with PD code |
| PD\_DESC | (more granular than Offense Description) |
| CRM\_ATPT\_CPTD\_C | Indicator of whether crime was successfully completed or attempted, |
| D | but failed or was interrupted prematurely |
| LAW\_CAT\_CD | Level of offense: felony, misdemeanor, violation |

|  |  |
| --- | --- |
|  | Jurisdiction responsible for incident. Either internal, like Police, |
| JURIS\_DESC | 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 | Specific location of occurrence in or around the premises; inside, |
| SC | opposite of, front of, rear of |
| PREM\_TYP\_DESC | Specific description of premises; grocery store, residence, street, etc. |
|  | Name of NYC park, playground or green space of occurrence, if |
| PARKS\_NM | applicable (state parks are not included) |
| HADEVELOPT | Name of NYCHA housing development of occurrence, if applicable |
|  | X-coordinate for New York State Plane Coordinate System, Long Island |
| X\_COORD\_CD | Zone, NAD 83, units feet (FIPS 3104) |
|  | Y-coordinate for New York State Plane Coordinate System, Long Island |
| X\_COORD\_CD | Zone, NAD 83, units feet (FIPS 3104) |
|  | Latitude coordinate for Global Coordinate System, WGS 1984, decimal |
| Latitude | degrees (EPSG 4326) |
|  | Longitude coordinate for Global Coordinate System, WGS 1984, |
| Latitude | 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:

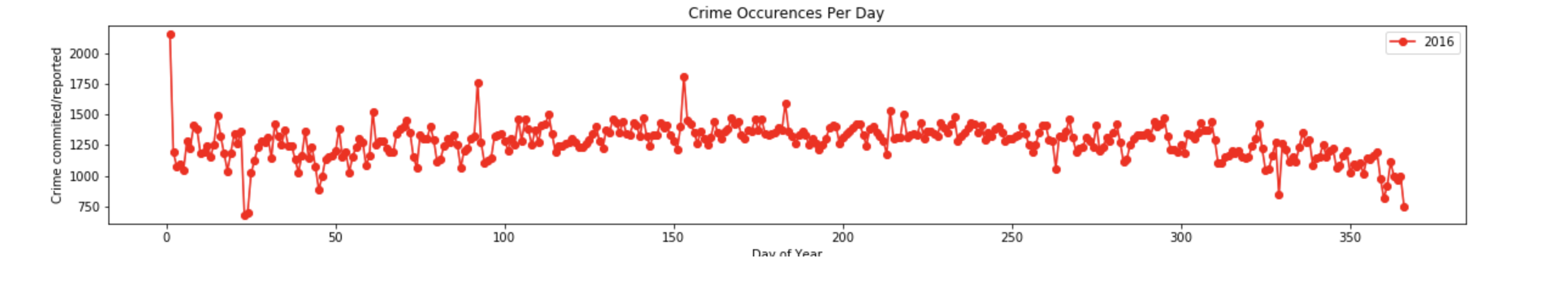
* Total Crime Occurrences across each year in New York City.
* Crime Occurrences across each year per Borough.
* Crime distribution percentage across each Borough every year.
* Crime Occurrences by the Nature of the Crime.
* Crime Occurrences by precinct in each Borough.
* Crime Occurrences by Weather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

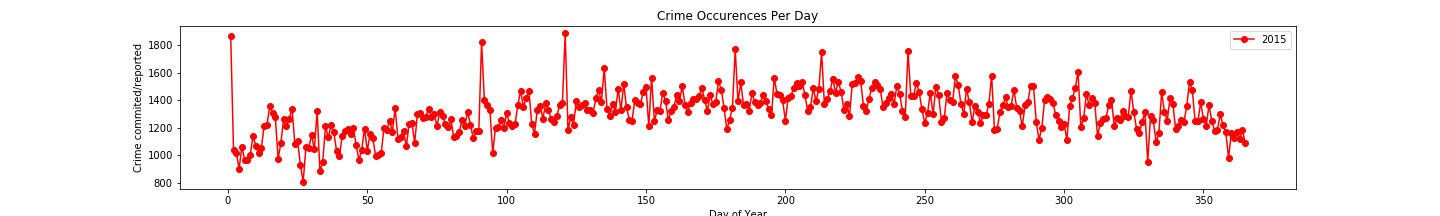
**Summary:**

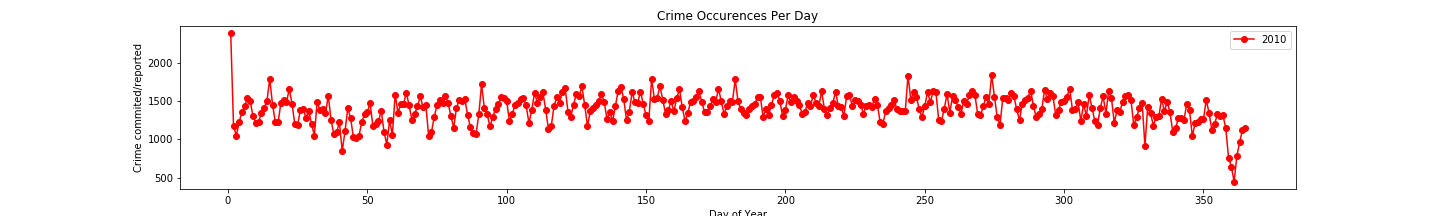
**Observations:**

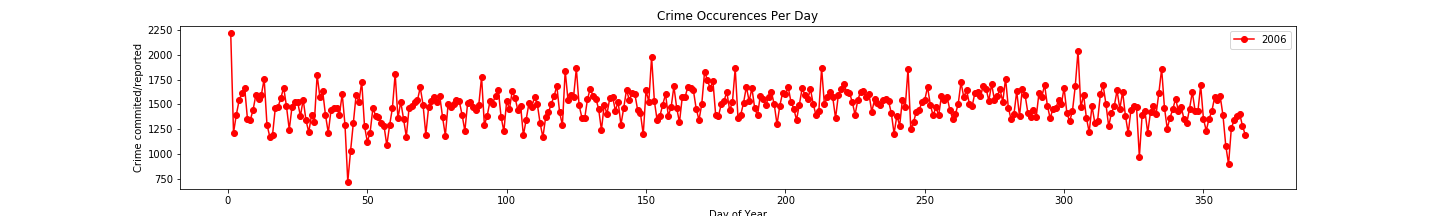
During the course of analysis, our focus was at eliciting the frequency distribution in crime occurrences across time and space, demanding the need to plot each data by year and borough.

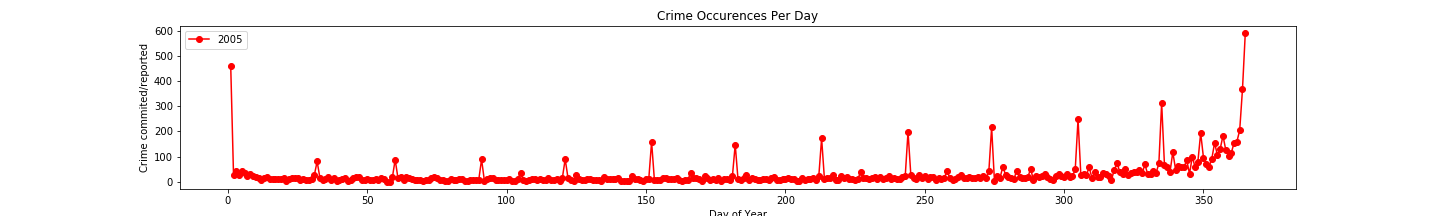
**1: Total Crime Occurrences across each year in New York City**







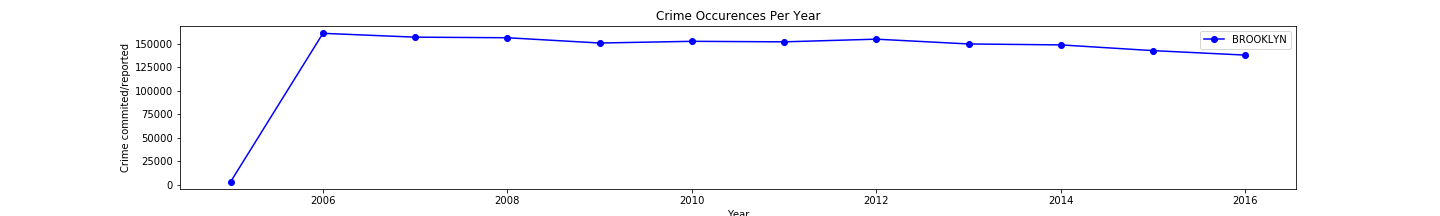


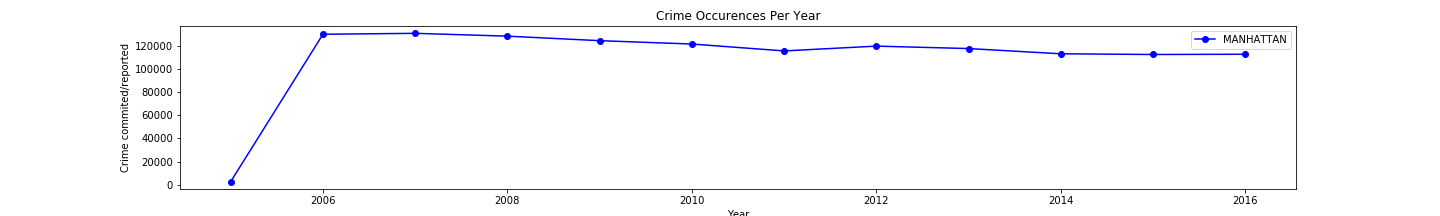


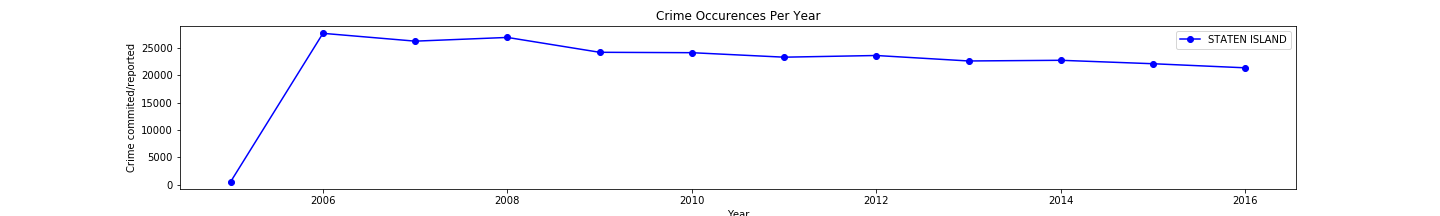
From the graphs, we can infer that the crime count has almost been consistent from 2006 to 2016 across New York city across each year. However strangely we also find that the crime rate in New York was the lowest during 2005 continuing towards an increasing trend only towards the end of the year surmising its due to the holiday season.

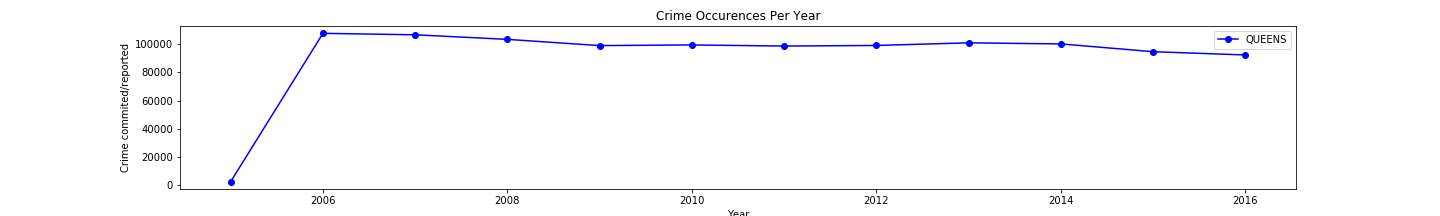
**2: Crime Occurrences across each year per Borough**

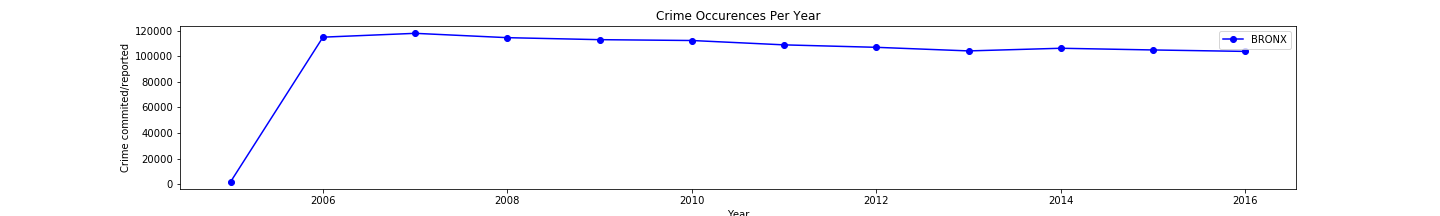
Due to the consistency observed with the Crime patterns across years between 2006 to 2016, we tried to extrapolate the data of Crime distribution across each Borough for every year.





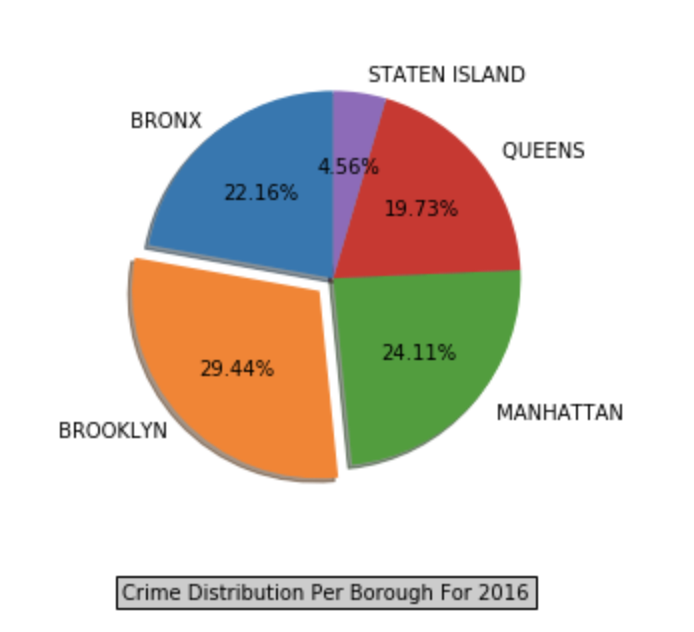


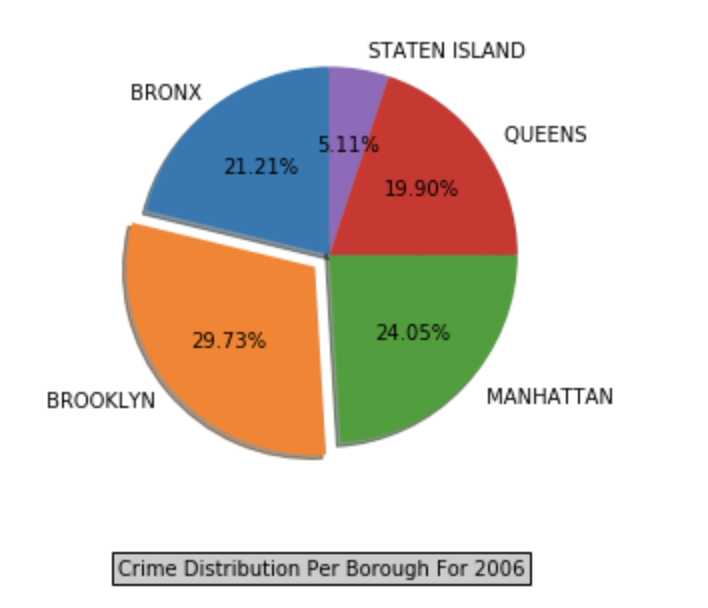




The frequency in the crime occurrences by borough is also observed to be consistent across all the years thus indicating the nature of crime incidences across NewYork for the entire period.

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



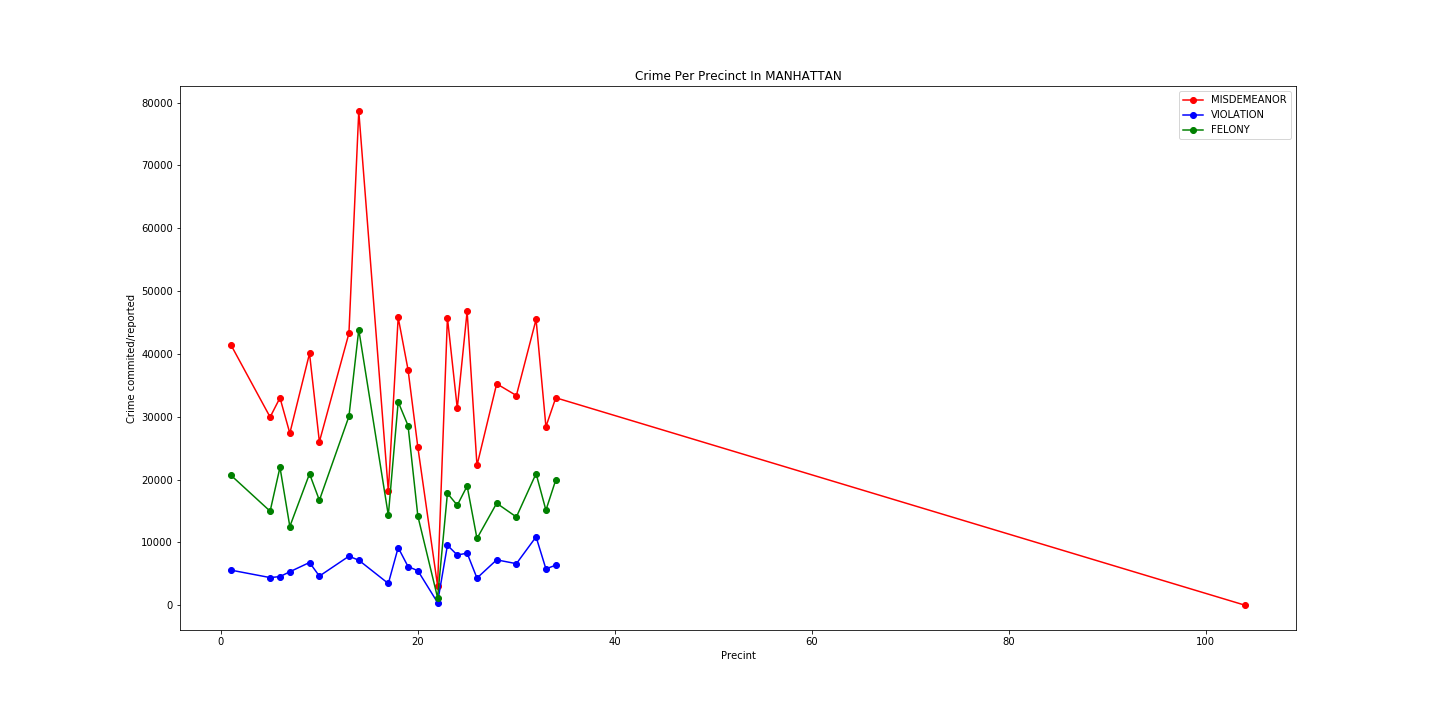


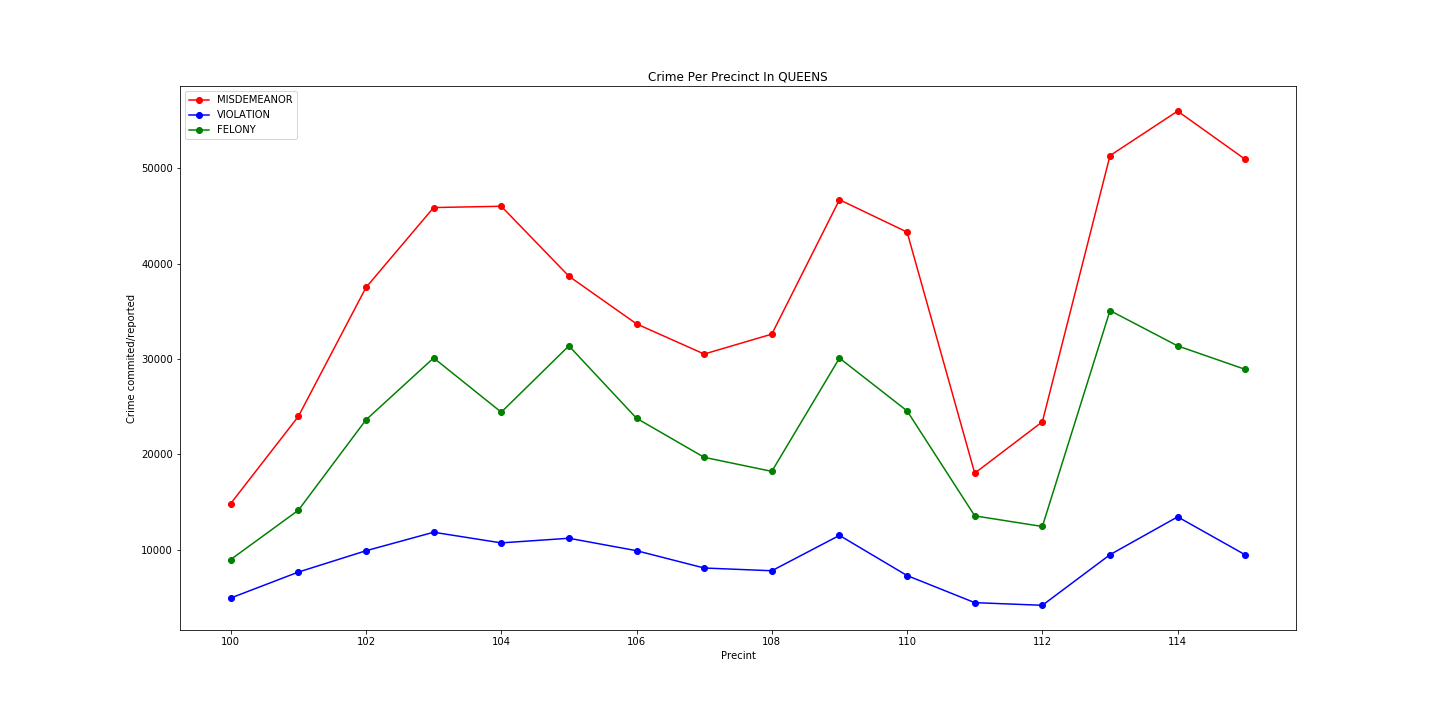
The pie chart gives an estimate of the relative distribution of crime occurrences to compare amongst all the Borough’s. As observed the maximum crimes are observed to be in Brooklyn which became our basis to do further exploratory analysis between Crime Complaints to Real Estate in Brooklyn.

**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**

**To be Filled……….**

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

**To be Filled……….**

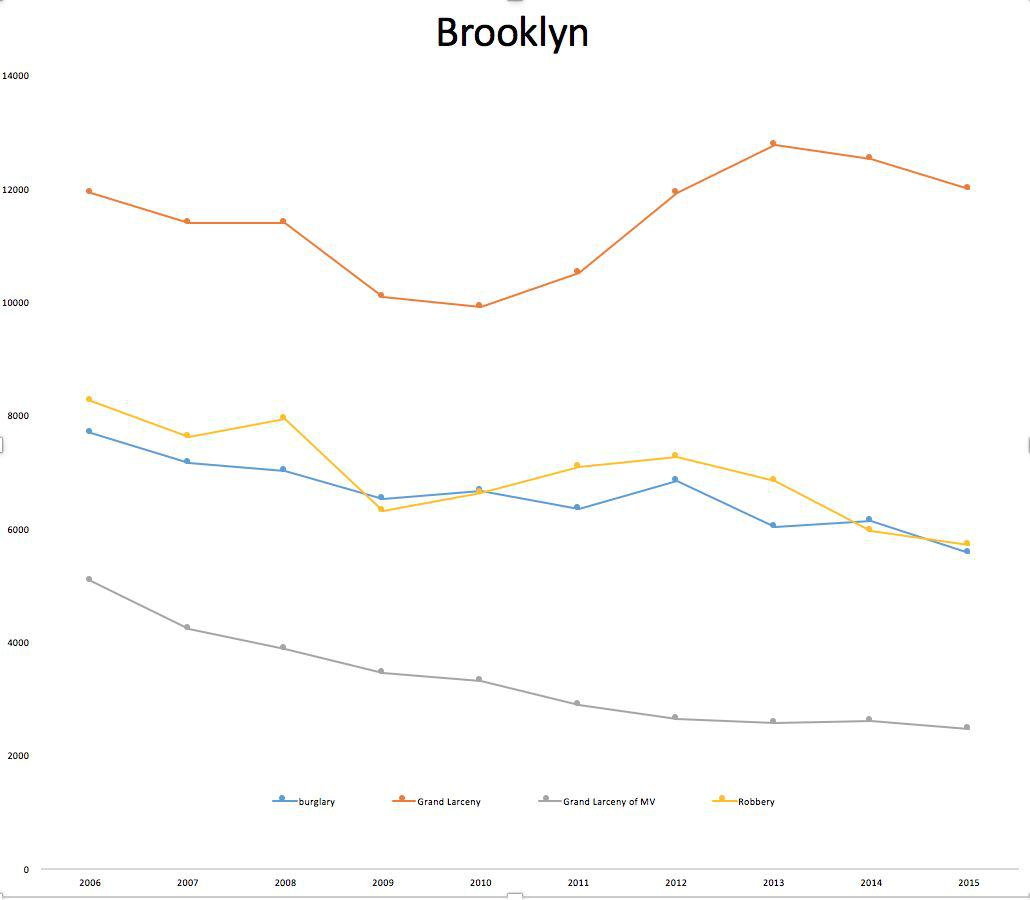
**Part 2: Exploring the Relationship Between Crime Patterns and Real-Estate Demand in Brooklyn**

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 incident

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.

**Figure 12: Number of Felonies in Brooklyn by Category**

****

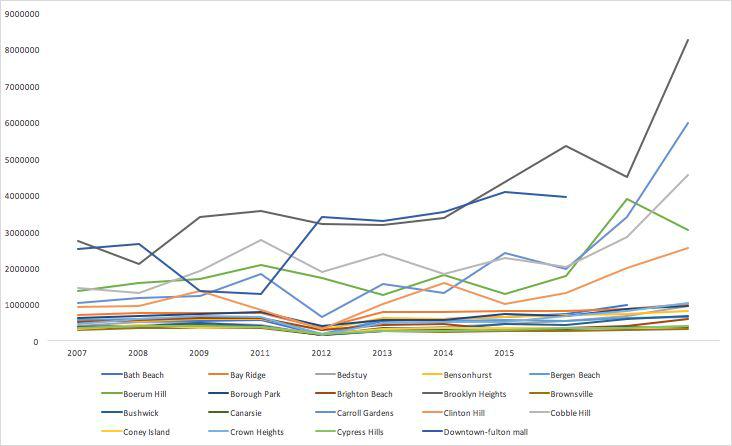
For data on real-estate interest and cost shifts, which we thought should have an impact on crime rates, we turned to the ACRIS NYC archives, which break down prices by borough for varying types of sales. Sure enough, from the chart below we saw there was a steep increase in the average cost of 1-family, 2-family and 3-family homes.

**Figure 13: 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 1-family, 2-family and 3-family homes for a sample of Brooklyn neighborhoods. It was interesting to see the infamous dent in 2012 of real-estate values!

**Figure 14: 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 and Park Slope 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.

While understanding our new housing price dataset, we also learned that the definition of a “home” in NYC sales mean that the purchase is a stand-alone building, not an apartment, which made us seek a more comprehensive dataset from ACRIS that included all real-estate purchase from rentals, condos, to coops and homes. The following analysis includes the latter categorization of sales.

**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)**

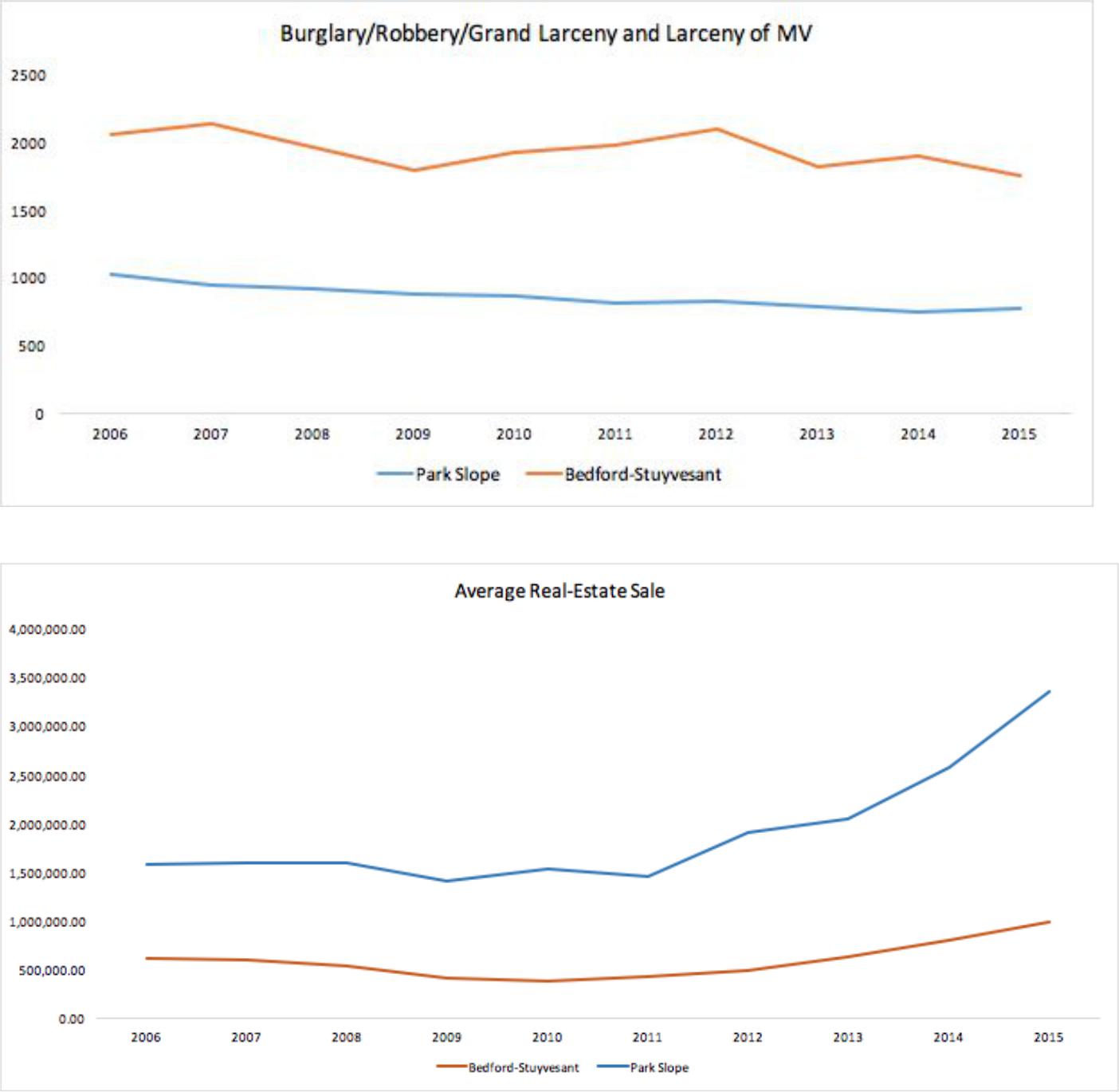
|  |  |  |
| --- | --- | --- |
| **Bedford-Stuyvesant** | |  |
| 10 Year Increase in Sale Cost: | | **61%** |
| 10 Year Decrease in BRGL: | | **15%** |
|  |  |  |
| **Year** | **Avg Sale Cost** | **Count of BRGL** |
|  |  |  |
| 2006 | $605,831.34 | 2,053 |
|  |  |  |
| 2007 | $599,073.63 | 2,133 |
|  |  |  |
| 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 |
|  |  |  |

**Table 4: Park Slope/Park Slope South - Average Price of all Real-Estate Sales and Count of Burglary, Robbery, Grand Larceny and Grand Larceny of Motor Vehicles (BRGL)**

|  |  |  |
| --- | --- | --- |
| **Park Slope** |  |  |
| 10 Year Increase in Sale Cost: | | **112%** |
| 10 Year Decrease in BRGL: | | **25%** |
|  |  |  |
| **Year** | **Avg Sale Cost** | **Count of BRGL** |
|  |  |  |
| 2006 | $1,578,509.14 | 1,018 |
|  |  |  |
| 2007 | $1,588,772.97 | 944 |
|  |  |  |
| 2008 | $1,590,645.56 | 910 |
|  |  |  |
| 2009 | $1,411,799.96 | 875 |
|  |  |  |
| 2010 | $1,533,640.79 | 855 |
|  |  |  |
| 2011 | $1,446,689.03 | 801 |
|  |  |  |
| 2012 | $1,909,316.35 | 817 |
|  |  |  |
| 2013 | $2,044,284.97 | 777 |
|  |  |  |
| 2014 | $2,575,820.84 | 743 |
|  |  |  |
| 2015 | $3,343,544.29 | 765 |
|  |  |  |

**Figures 15 and 16: Park Slope and Bedford Stuyvesant - Comparison of Average Price of all Real-Estate Sales and Count of Burglary, Robbery, Grand Larceny and Grand Larceny of Motor Vehicles (BRGL)**

****

A simple Pearson correlation between real estate prices and the rates of burglary, robbery, and larceny resulted in -.3345 for Bedstuy with a p-value of .3446, while the results for Park Slope were -.6161 and and p-value of .058. As we might have guessed, the up-and-coming neighborhood of Park Slope has seen an increase in real-estate costs as the worries of burglary, robbery, and larceny have decreased over the last decade. Bedford-Stuyvesant, on the other hand, shows this trend less drastically or perhaps hasn’t yet joined the bandwagon of rapidly evolving

Brooklyn neighborhoods. We should note that this is a very naive correlation result as the sample set we drilled down to is too small for statistical significance. This was a lesson learned and side-effect of taking on the complexity of exploring many-to-many relationships within sub-categories of crimes and neighborhood-level slicing of boroughs. Regardless, we discovered the richness of the available urban dataset to test our assumptions about the city we live in, and how to leverage Spark to decimate the effort involved.