# NYPD Crime Data Analysis

Yurui Mu
Li Lin Qin
Yidi Zhang

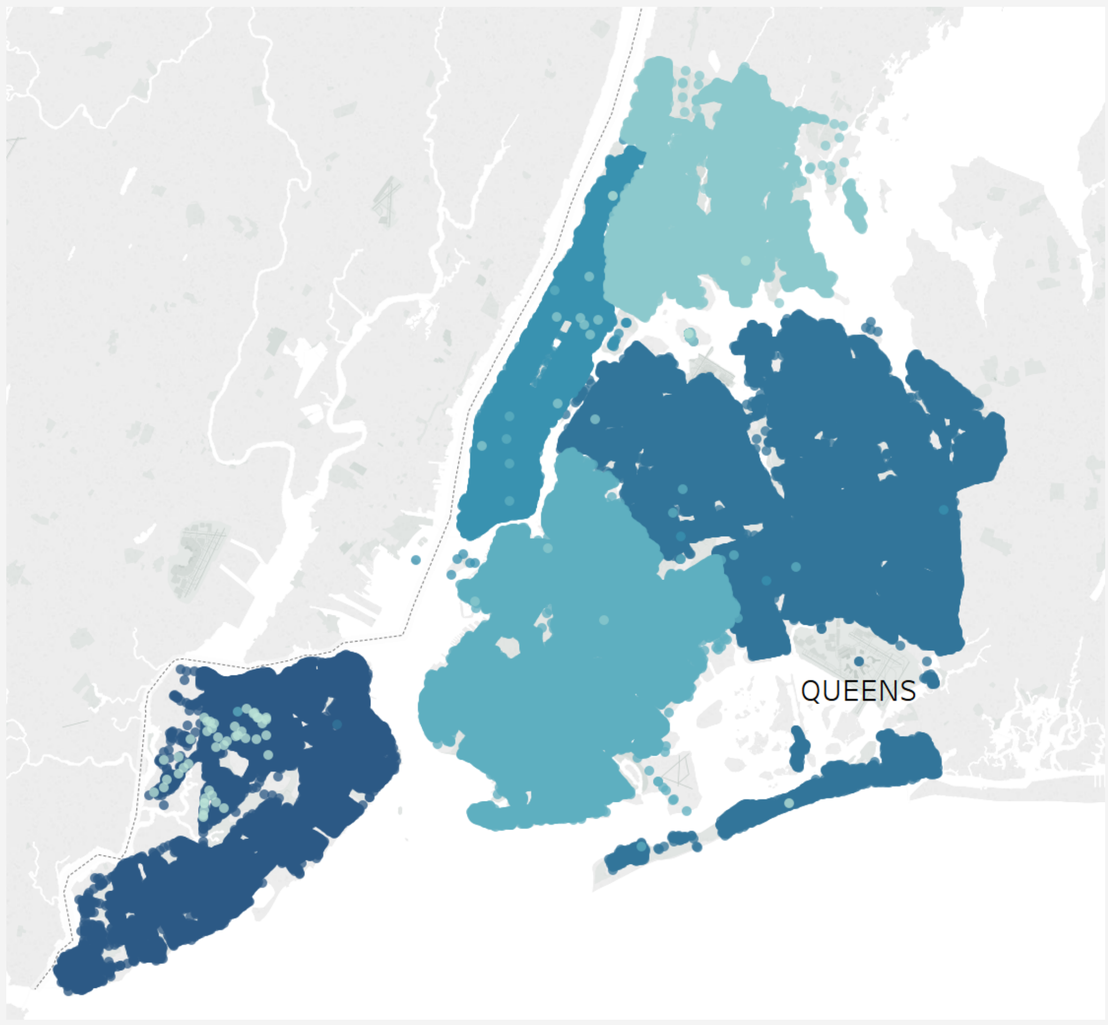
The dataset we will be using is NYPD Complaint Historical Data, which includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2015.

The purpose of this project is to provide a comprehensive analysis for the data. In particular, we will analyze the data from demographic, temporal, synoptic and spatial perspectives to give reasonable explanations for trends and periods.

Results and conclusions can be used to gain a better understanding of the crime trends and help control crime rates in NYC.

For the first part of the project, we analyze the data and generate a descriptive summary of their contents as well as a list of data quality issues.

For the second part, we make a list of hypotheses to gain insights into the data, integrate our selected collection with one or more data sets and look for interesting relationships between them.




Big data infrastructure is especially useful in processing large data that have large volume, variety, and require high velocity. Since the dataset has 5101231 entries and 24 columns, it is ideal to use big data tools.

In this project, we use MapReduce and Spark in operations such as joining tables, and counting values by key.

All components of our project are reproducible.

# Part I: Comprehensive Data Analysis

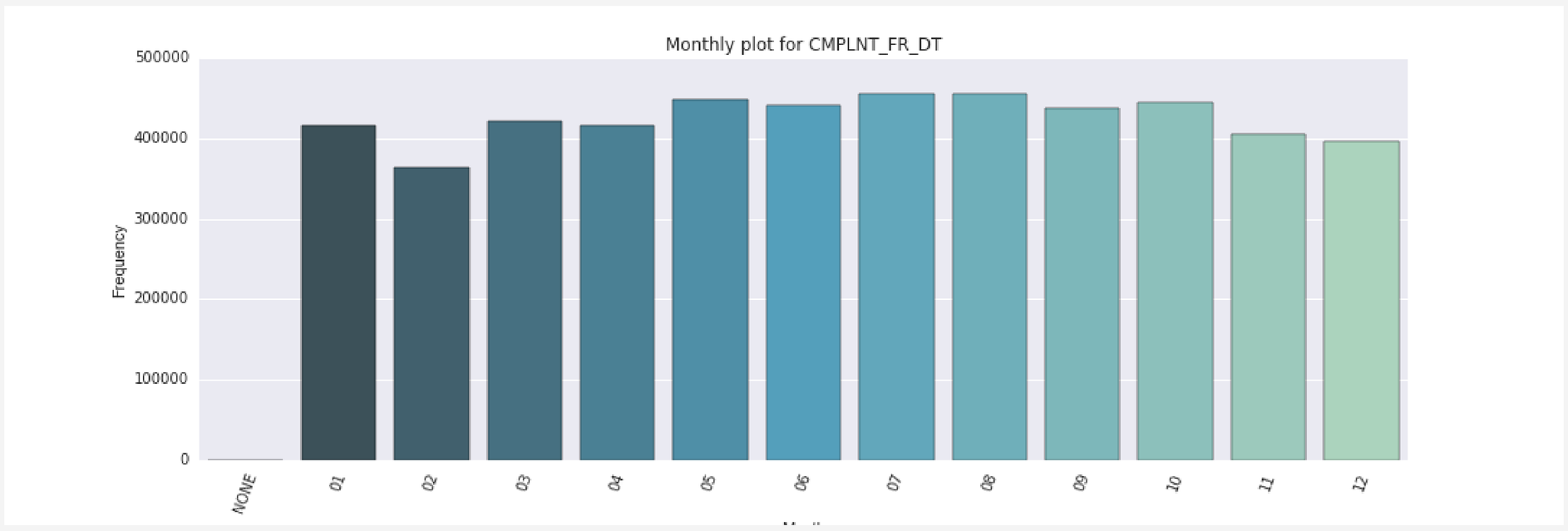
There are a total of 24 columns in the data set.

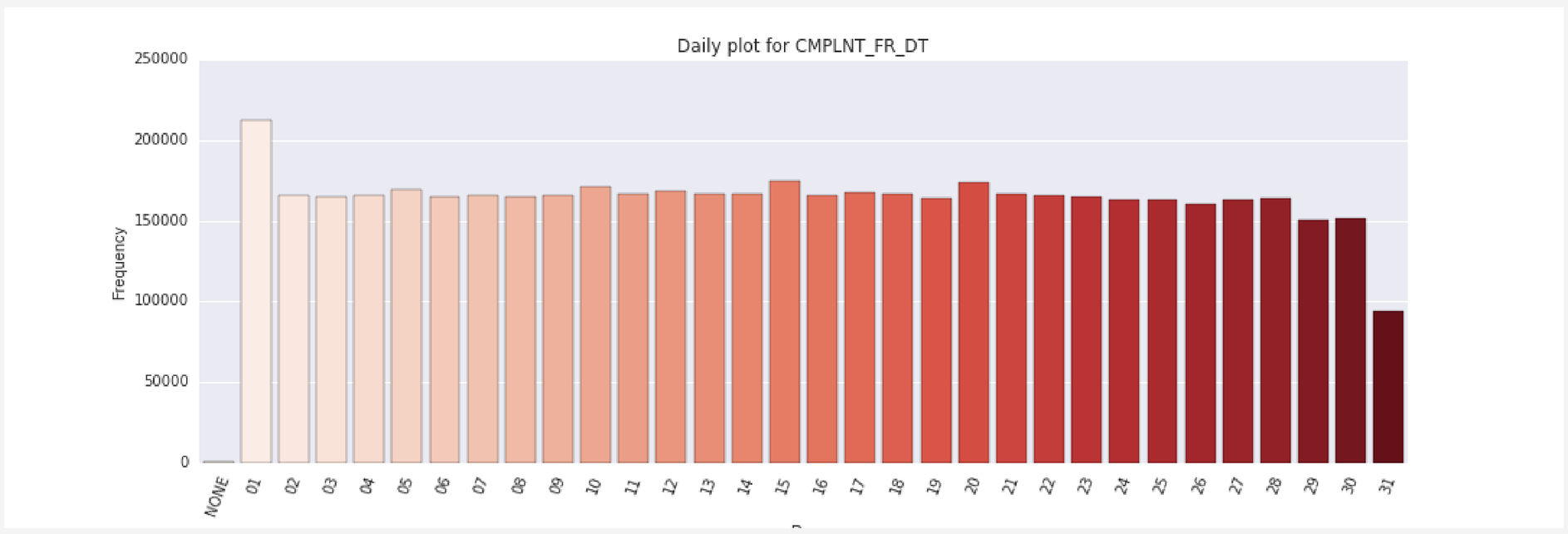
## 1) Complaint Number ID

Randomly generated persistent ID for each complaint. There are 5101231 unique integers assigning to each cases, which is essentially the number of observations in our dataset. All of the entries are valid integers.

## 2) Complaint From Date

The date each event is filed. Using datetime.date() function, there are 5100576 valid DATETIME type dates and 655 null values. Among all, most complaints were filed in 2005-2015. There are also 7 records with dates in 1015, which might be typos, here we still consider them as valid, because they fit in the form of '%m/%d/%Y'. When performing analysis, we should change them into year 2015. In addition, there are also early records dated back to 1900s. Since those records do not seem to be mistakes, we counted them as valid records along with the others.

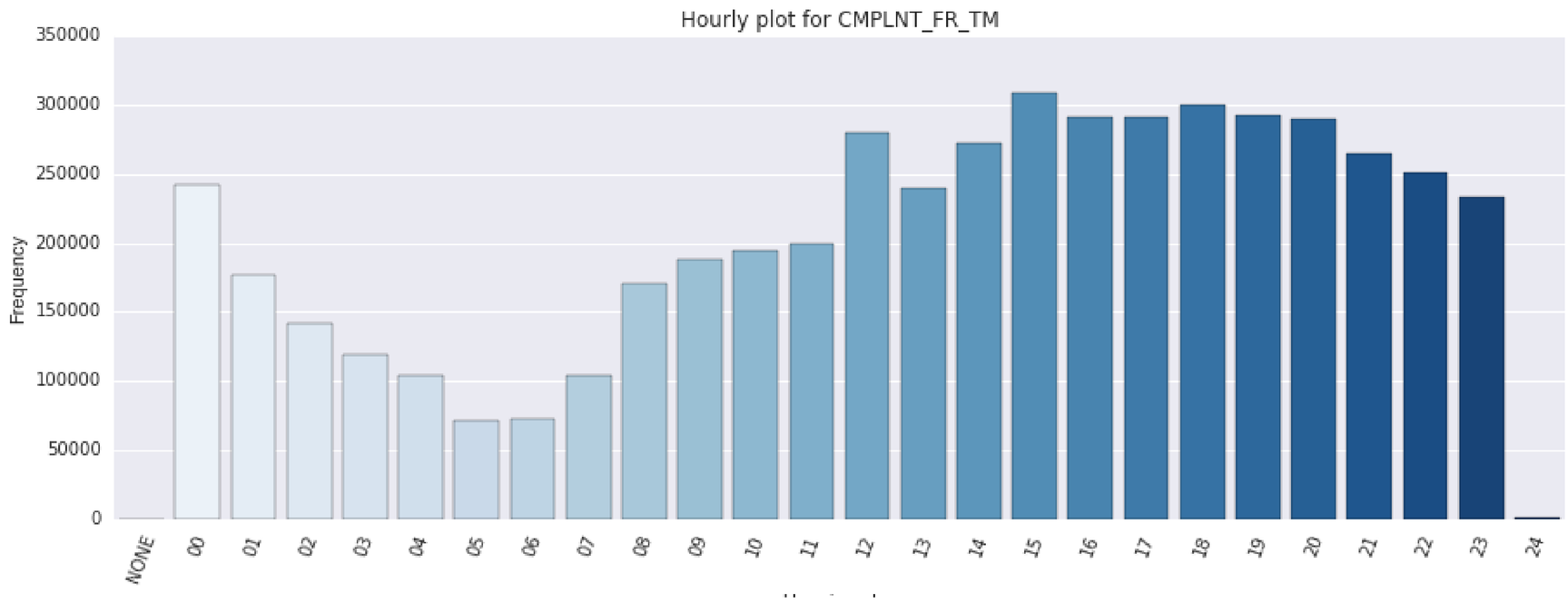



The above graphs indicate counts in a monthly or daily basis over years. The crimes spread rather evenly among all the months, with slight fewer ones in February. We can see that a lot more crimes happened during the first day of the month. Our guess is that it is probably a default number when the exact day of the month is unknown. Furthermore, the 31st of a month had least crimes, for the reason that some months do not have 31st as their last day.

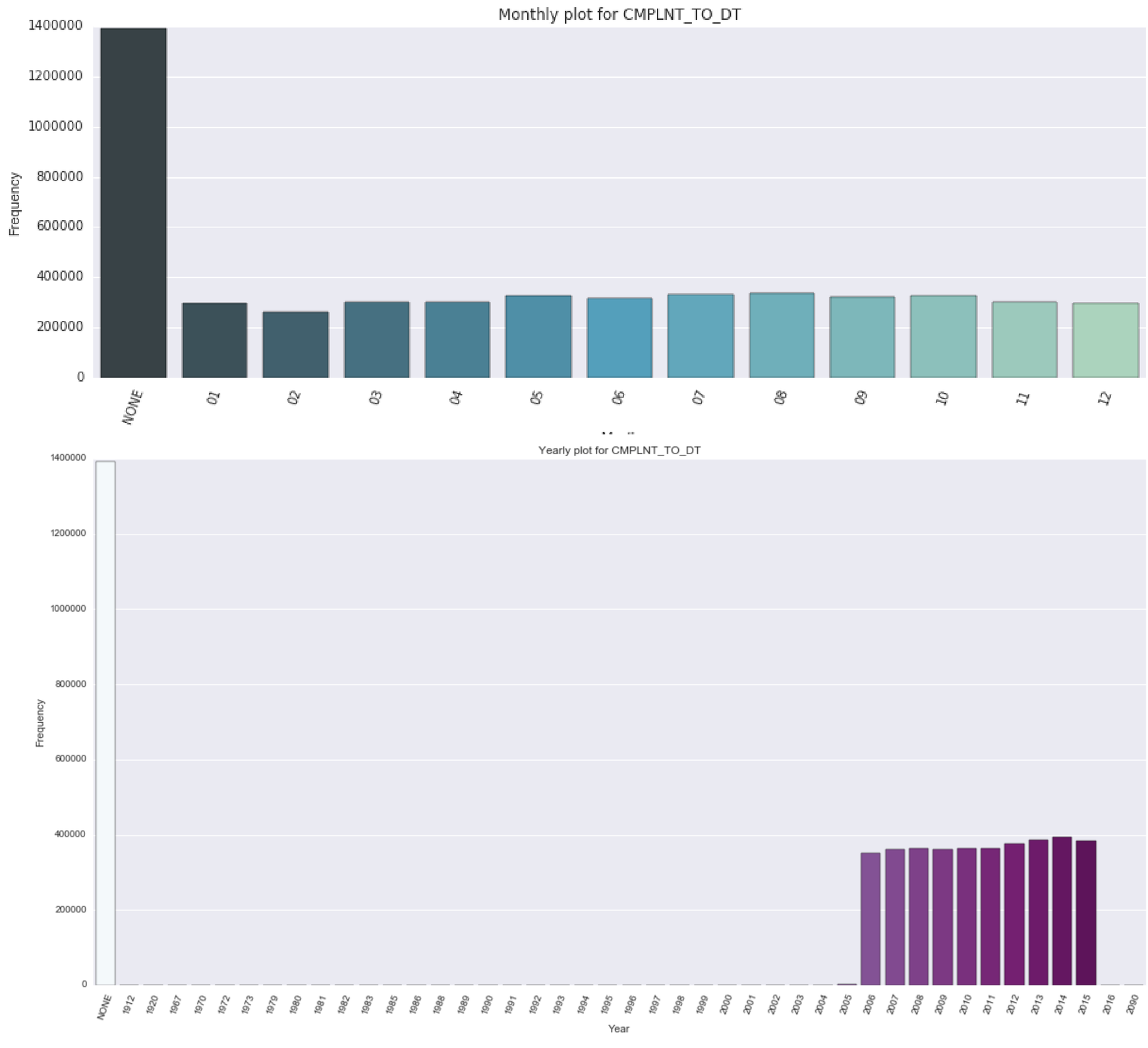
## 3) Complaint From Time

Complaint From Time is the specific time of the day each event is filed. Using datetime.time() function, there are 5100280 valid DATETIME type time, 903 invalid time, and 48 null values. Invalid values come from ‘24:00:00’, whereas midnight is defined as ‘00:00:00’. Thus we are not sure if it happens at midnight or did they simply don’t know about the time. A majority of the crimes happened in the afternoon till later of the day. Fewer crimes took place during midnight or early morning.



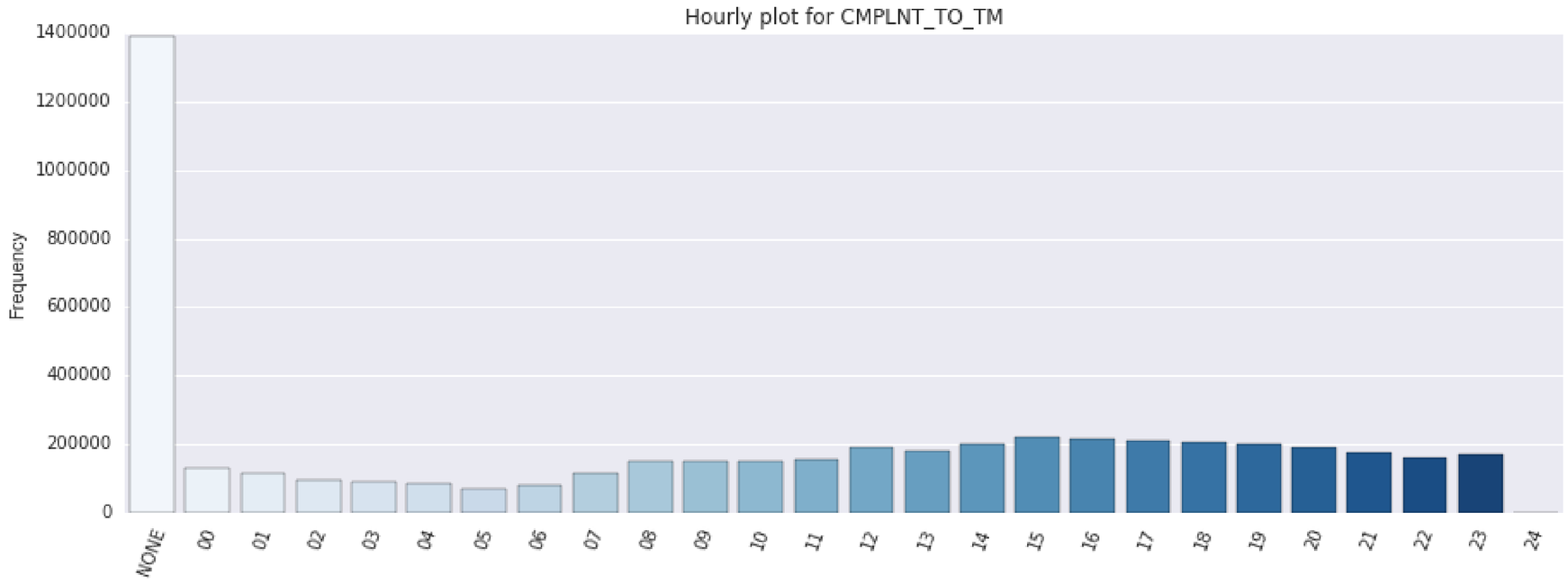

## 4) Complaint To Date

Complaint To Date is the date when event ended. Using datetime.date() function, we found 3709213 valid DATETIME type dates, 540 invalid DATETIME type dates and 1391478 null values. Invalid dates include one record in ‘2090’ and 539 invalid ending dates before starting dates.
Most of the entries are null values, as shown in the graph. The remaining ones are spread relatively evenly among all the months in a year.



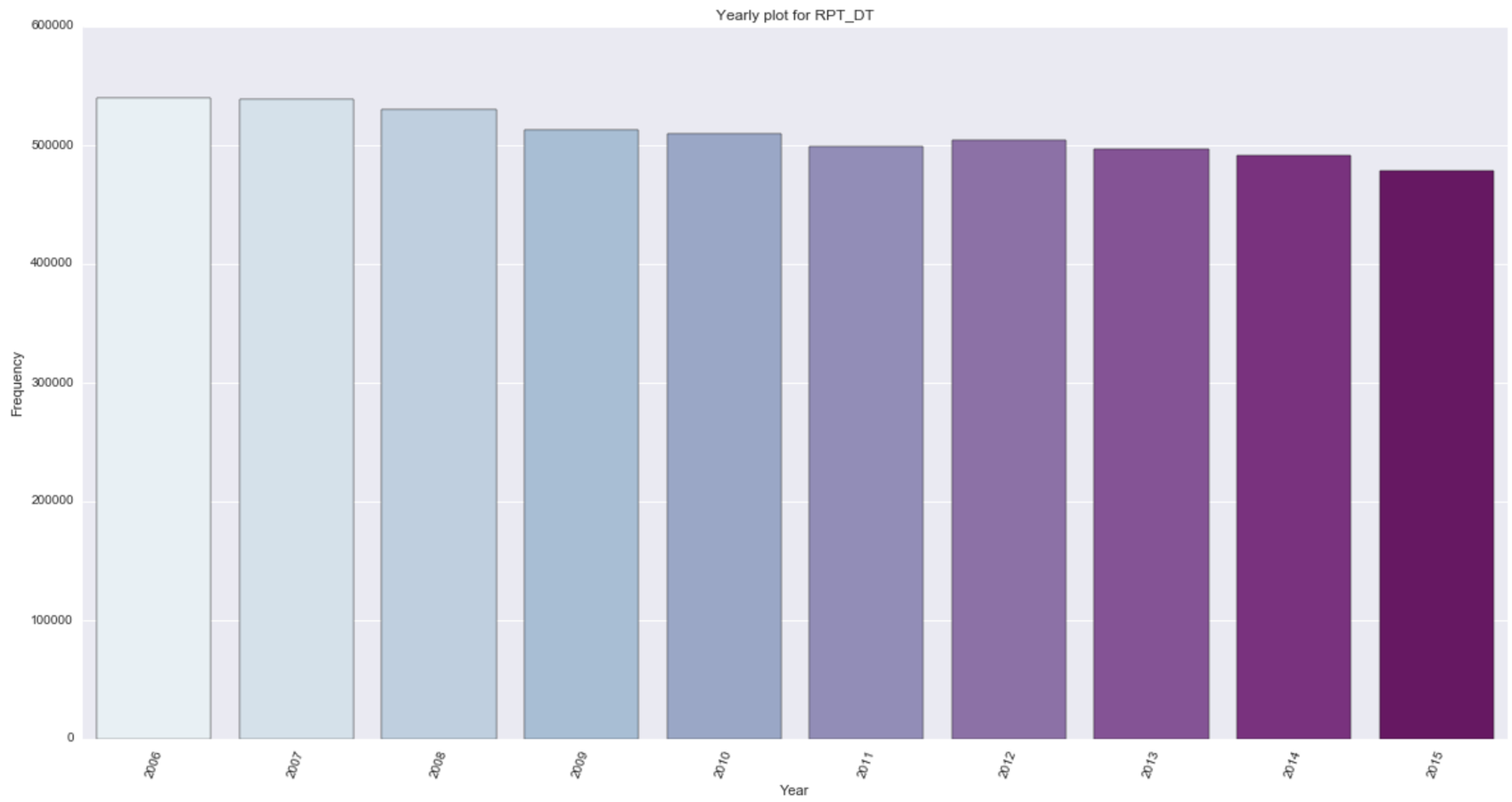
## 5) Complaint To Time

Complaint To Time is specific time of the day when event ended. Using datetime.date() function, we found 3712070 valid DATETIME type time, 1376 invalid time(24:00:00), and 1387785 null values. From the below graph, it is clear that most of the values are null.




## 6) Complaint Report Date

Complaint Report Date is the date police filed reports. In this datetime column, all of the 5101231 values are valid, ranging from year 2006 to 2015. Below is a graph of distributions across these years.




## 7) Key Code

Key Code is the three digits code assigned to each offense category. Type Integer.
There are no missing values. All codes are valid three digits codes, splitting into 74 categories. Then we pair up KY_CD with OFNS_DESC as key, we received 122 pairs of keys. Other than the reason that there are cases with valid key code, while missing offense descriptions. There are also issues due to multiple names of a single key code, like ‘kidnapping’ vs ‘kidnapping & related’ and ‘agriculture & mrkts law-unclassified’ vs ‘other state laws’.

## 8) Offense Description

Offense Description is the semantic description of offenses reported. It includes 70 unique valid string values and 19080 null values. There is a total of 5082151 valid values.

## 9) PD Code

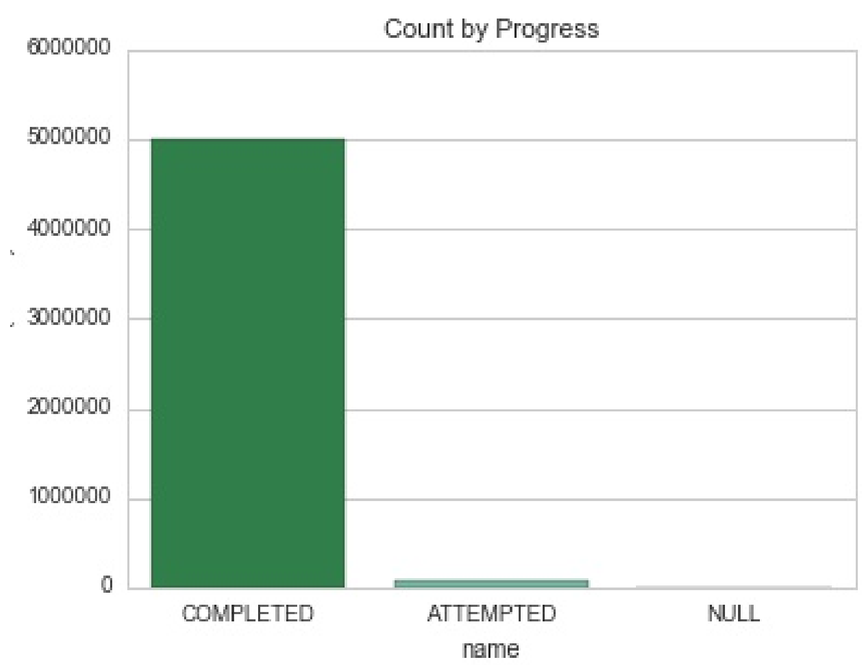
PD Code is a three digit integer representing the internal classification code. This column includes 415 unique valid float values and 4574 null values. There is a total of 5096657 valid values.

## 10) PD Description

PD Description is a text description of internal classification corresponding with PD code. This column includes 403 unique valid string values and 4574 null values. There is a total of 5096657 valid values. Some of the most common descriptions are printed in the below word cloud.

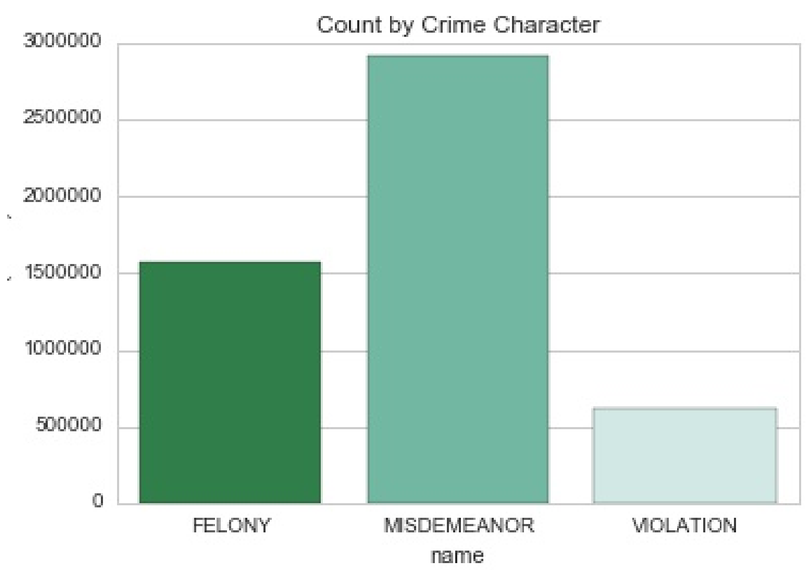

## 11) Crime Attempted Code

Crime Attempted Code is an integer indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely. ‘Completed’ and ‘attempted’ are two values contained in this column, and nearly 80% of the values are ‘completed’. There are 5101224 valid values and 7 null values.



## 12) Law Category Code

Law Category Code is an integer describing the level of offense: whether it is a felony, misdemeanor, or violation. All 5101231 entries are valid.

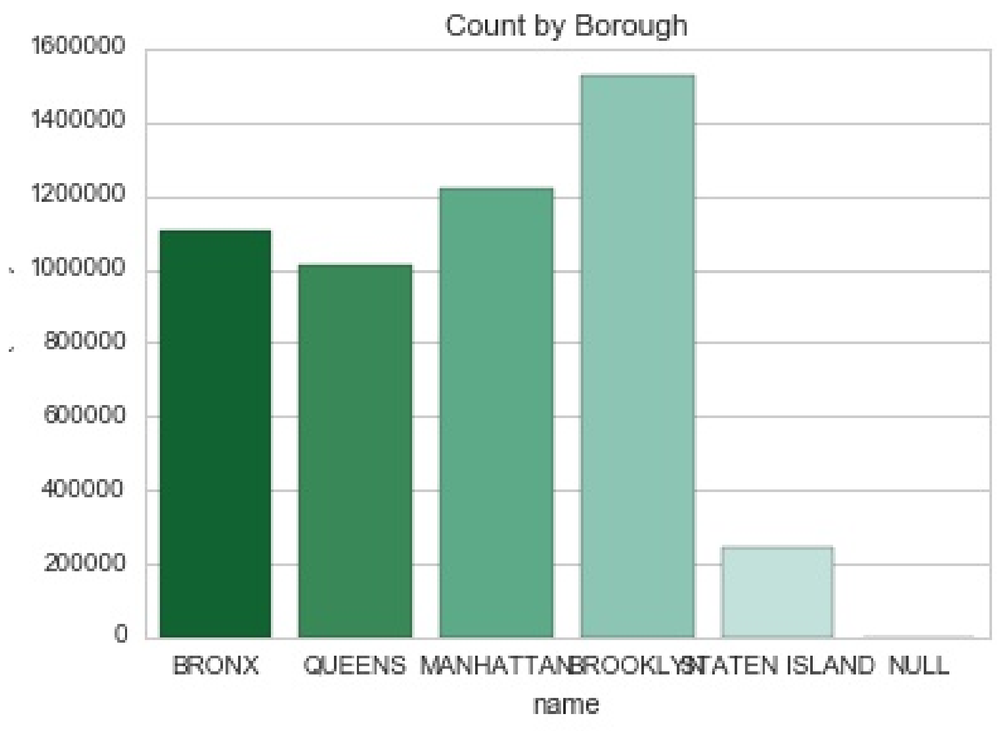



## 13) Jurisdiction Decision

Jurisdiction Decision is a string indicating the jurisdiction responsible for the incident. It has 25 unique values without any invalid or null entries.

## 14) Borough Name

Borough Name is the name of the borough in which the incident occurred. It is a string type. This column contains 5 unique string values, including 'BRONX', 'QUEENS', 'MANHATTAN', 'BROOKLYN', 'STATEN ISLAND', which is five boroughs in New York. There are 5100768 valid entries and 463 null values.

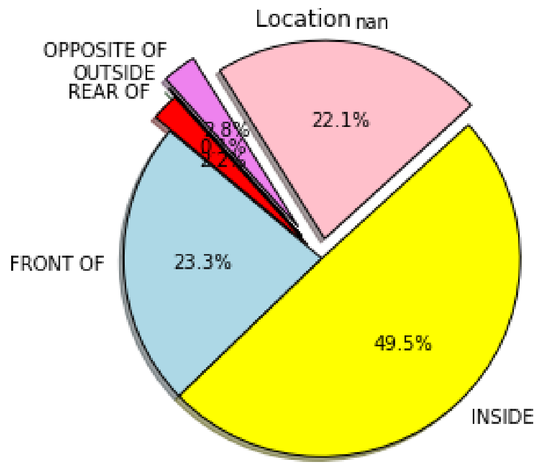



## 15) Address Precinct

The Address Precinct is a float number denoting the precinct in which the incident occurred. There are 5100841 valid entries and 390 null values.

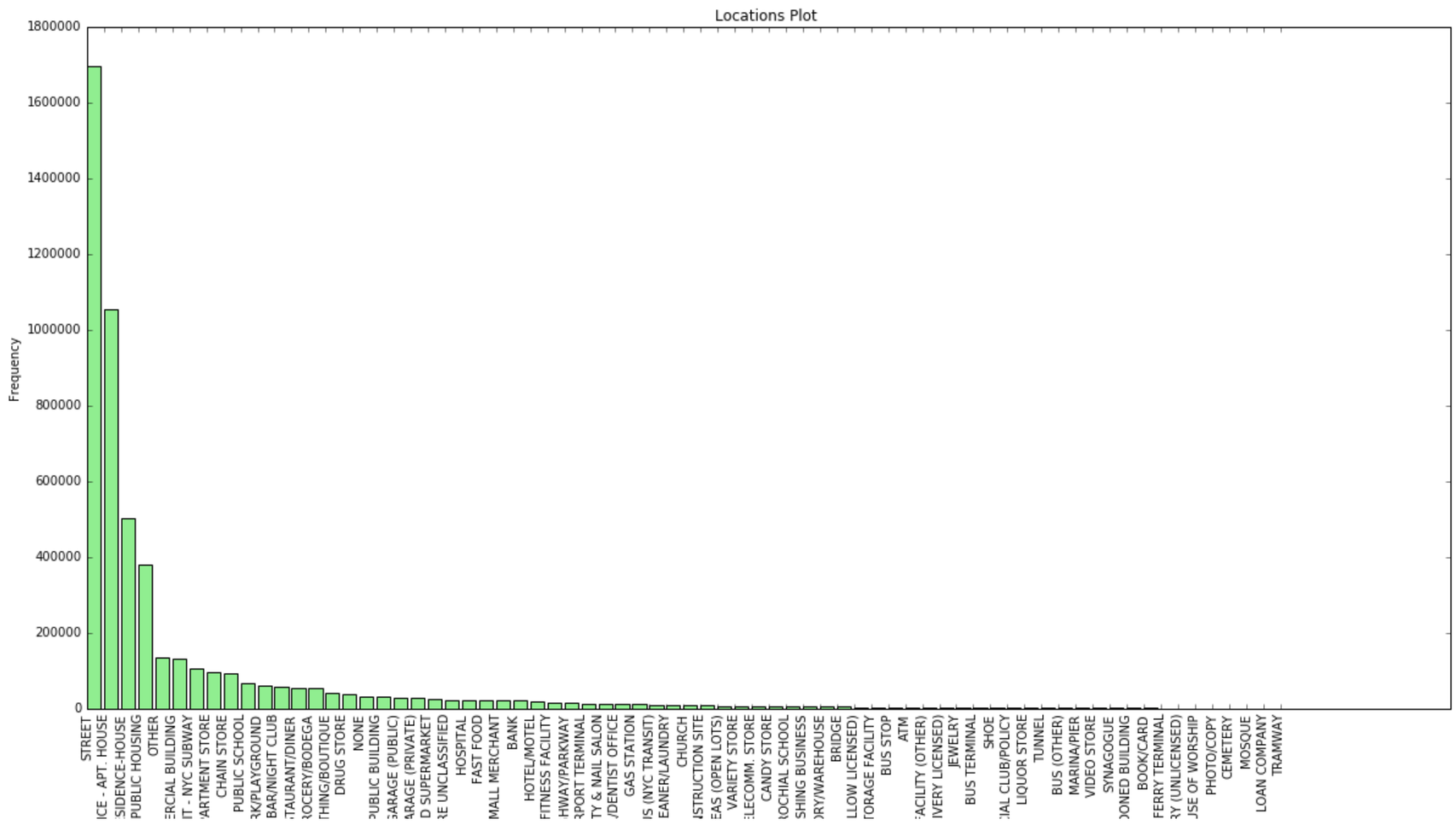
## 16) Location of Occurrence

Location of Occurrence is the relative location of occurred around places. Entries include “front of”, “inside”, “opposite of”, “outside”, and “rear of”, with respective counts of 1189787, 2527543, 140606, 2765,113189. There are also 1,127,128 null values in this column.




## 17) Premises Type Descriptions

Specific description of premises; grocery store, residence, street, etc. There are 33279(0.65%) null values and 5067952(99.35%) valid values. There are 70 valid types in total. We can see that most of the crimes took place on the street.

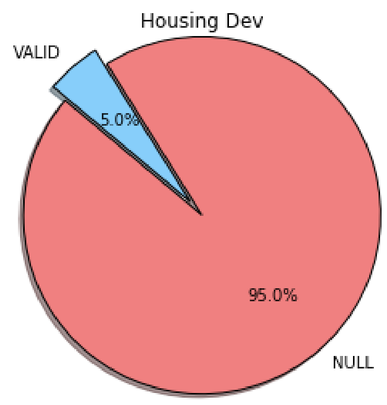



## 18) Park Names

Park names includes text of the name of NYC park, playground or greenspace of occurrence, if applicable. There are 7597 valid entries (0.15%) and 5093634 null values (99.85%). It seems that a majority of crime are happening outside of NYC park area. Among all 863 parks, Central Parks has the most crimes reported: 543. Following by Flushing Meadows Corona Park with 301 crimes, and Riverside Park with 188 crimes reported.

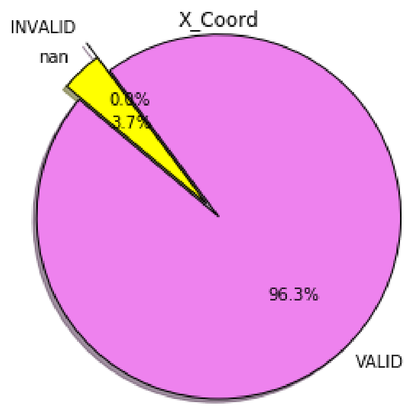
## 19) Housing Development

Hadevelpt is the name of NYCHA housing development of occurrence, if applicable. There are a total of 4848026 null values, as well as 253205 valid ones.  Among all valid values, there are 278 unique types.




## 20) X Coordinate

X Coordinate is a float number representing x-coordinate of the place where the crime happened. There are 4913085 valid entries and 188146 null values. The valid entries are all within the New York area.




## 21) Y Coordinate

Y Coordinate is a float number representing the y-coordinate of the place where the crime happened. Same as the column of x-coordinate, there are 4913085 valid entries and 188146 null values. The valid entries are all within the New York area. The plot should look the same as the above graph for X Coordinate.

## 22) Latitude

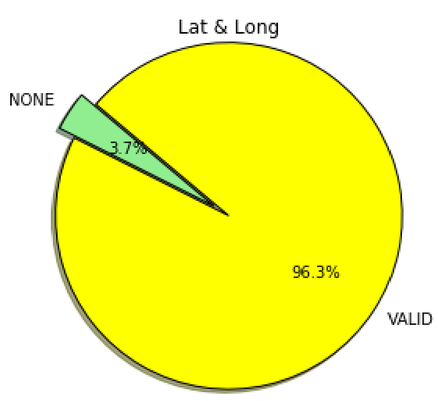
Latitude is a float number depicting the latitude where the crime took place. The valid entries are all within the New York area. There are 188146 empty entries and 4913085 valid counts.

## 23) Longitude

Longitude is a float number depicting the latitude where the crime took place. The valid entries are all within the New York area. There are 188146 empty entries and 4913085 valid counts.

## 24) Lat_Lon

Lat_Lon is a column that combines the two previous column, latitude and longitude to provide a location. It has a form of "(float, float)". So we checked for validity by confirming that the two numbers in the bracket match the two previous numbers in the same row. Besides 188146 empty entries, the remaining 4913085 are all valid.

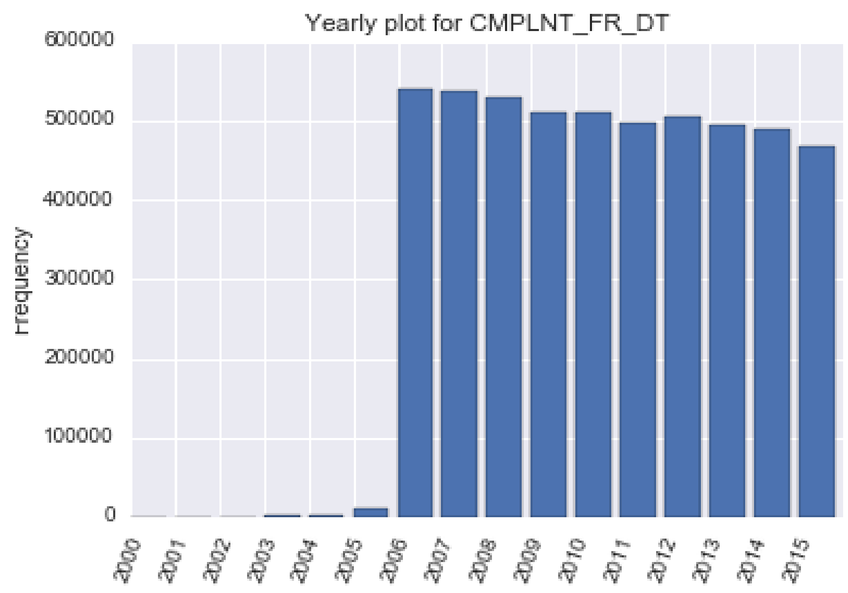


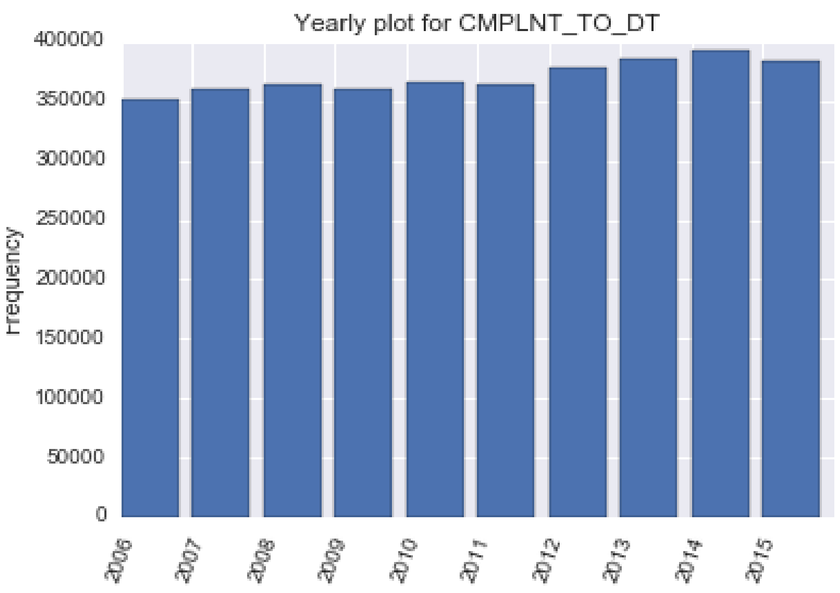

## Data Quality Issues

Besides the quality issues that we mentioned above, there are still some minor changes need to be made on the dataset.

1) PD_CD and PD_DESC should have same number of unique values, however not.

2) There are some invalid values in DATETIME columns, such as 1015 (changed to 2015). And count before 2006 is far smaller compared to the count after 2006. After cleaning, the result are shown as follows:



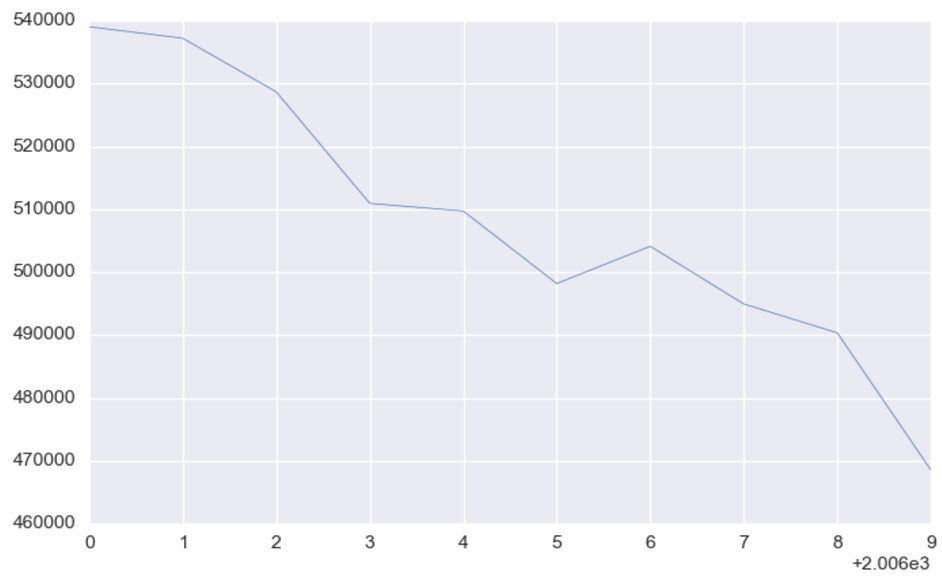
# Part II  Further Analysis with Hypothesis

Now that we have gained insights on NYPD 2006-2015 Crime Dataset, we want to get more information by observing data, raising assumptions and relating other datasets to explain our findings.

In this part, we analyzed NYPD Crime data from demographic, temporal,  synoptic and spatial perspectives to give reasonable explanations for trends and periods shown in crimes dataset.

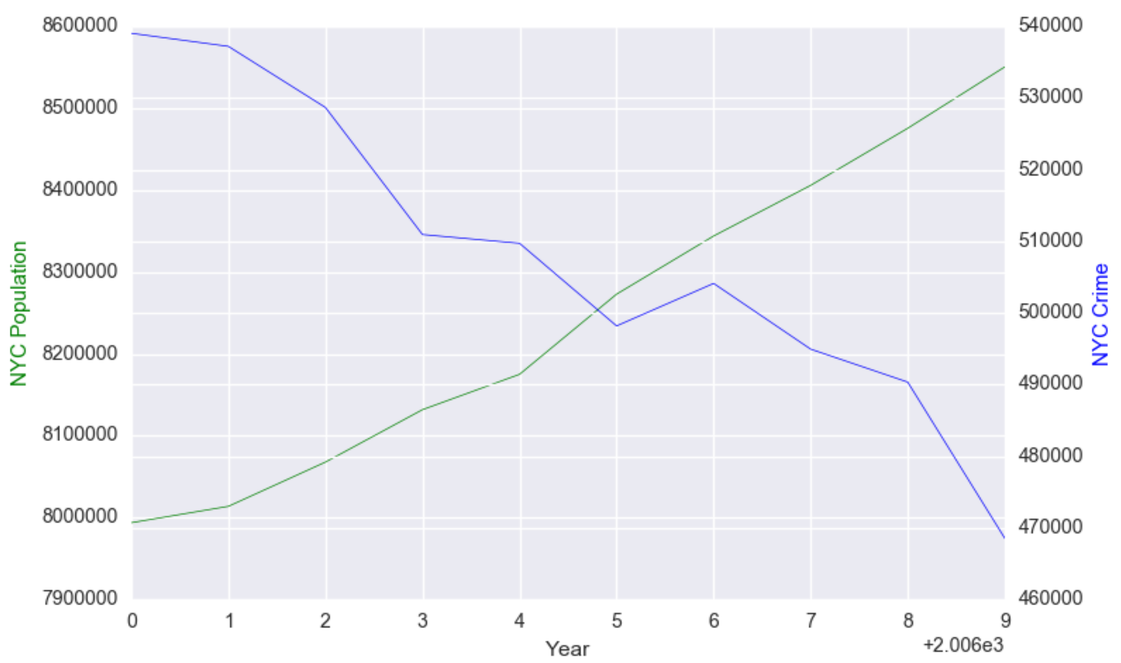
## 1) Demographic vs Crime

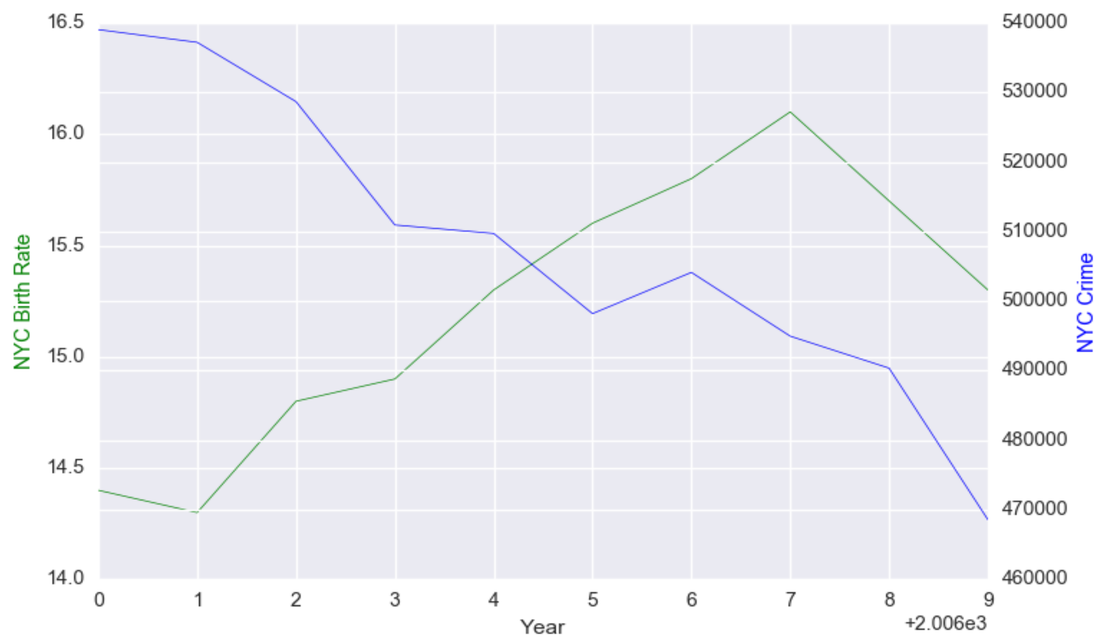
By aggregating our NYC Crime dataset by year, we can see that despite of unusual increase of crimes in year 2012, NYC Crime rates have an overall tendency of dropping through the years. Starting from 2006, where 539977 cases were reported,  to 2015, where  478578 cases were reported, an significant drop of crime cases as big as 12% is casted on the crime dataset, showing that NYC is becoming a significantly safer city to live in.




But is it really this case, we doubted. Is New York City really safer than before? Is it possible that an increasing number of people are moving out of New York City, thus with fewer people staying, fewer crimes happened?

After joining the NYC population dataset from Baruch College,  we found that population keeps growing through the years, and it keeps an average birth rate higher than NY State. Also, we found that crime rate has a negative correlation toward birth rate before 2013.



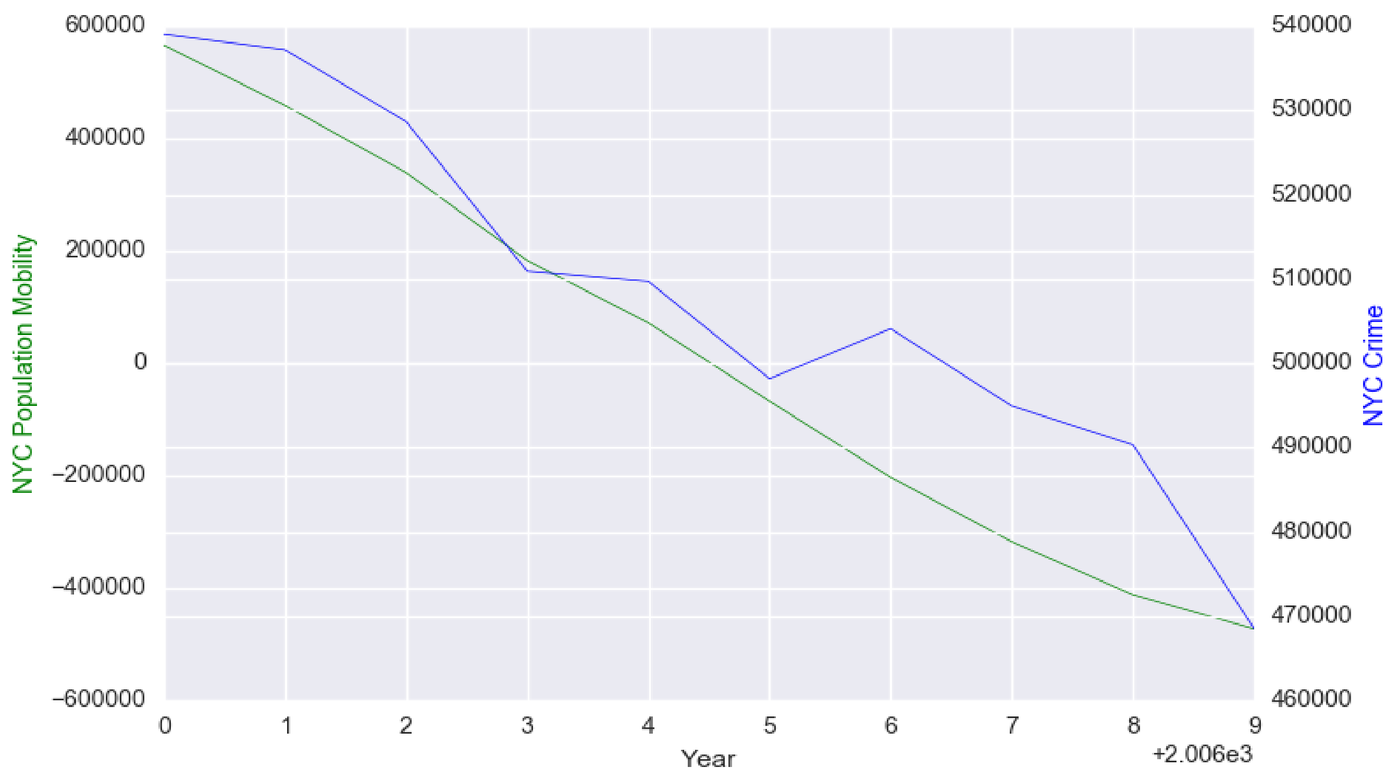
Then we pursued a detailed research on what exactly happened in and after year 2013 that altered the correlation between NYC birth rate and NYC crime rate.

According to a CDC report regarding United States abortion rate surveillance, in 2013, New York City reported 69,840 abortions and 116,777 registered births, which means that city's abortion rate is 60 percent of its birth rate. New York City also recorded the highest abortion rate – 36.3 abortions for every 1,000 women between 15-44, and the highest abortion ratio as well – 598 for every 1,000 live births.

Such a high abortion ratio should be related to NYC birth rate. But the reason of increasing abortion rate and decreasing birth rate after 2013 doesn't necessarily relate to decreasing crime rate. Because in year 2014, pro-abortion New York City Mayor Bill De Blasio pledged to partner with Planned Parenthood to expand his city's abortion businesses and to wipe out pro-life pregnancy centers, which leading to a decreasing birth rate in New York City.

Another aspect of population we researched is New York City population mobility. We deducted the newborns and deaths from each year to get the net number of people moving away from New York City, which gives a nearly perfect correlation with New York City, which shows that New York City is believed to be safer by New Yorkers. And the safer the city is, the less people moving out of the city and the more people moving in.

**The correlation coefficient between NYC crime counts and mobility population is 0.955.**

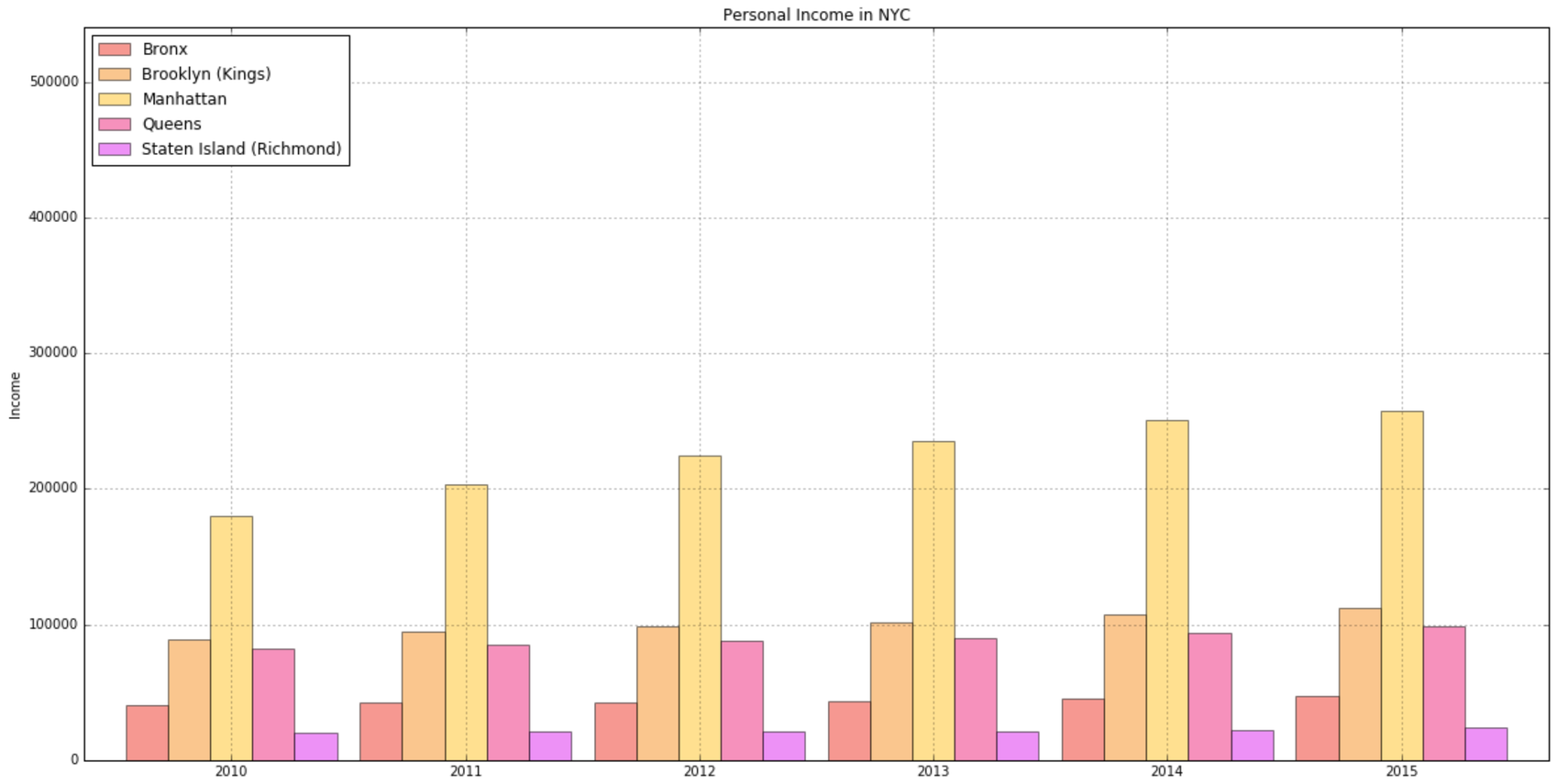


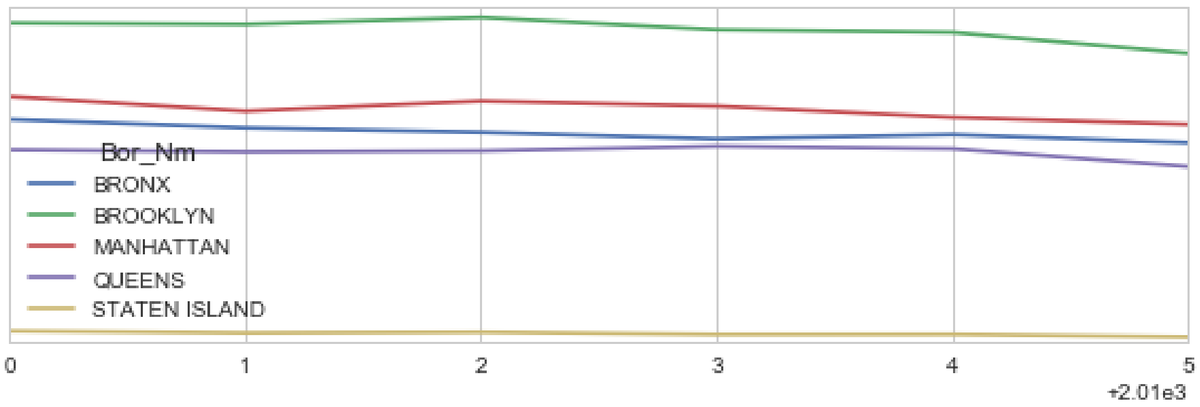

## 2) Income vs Crime

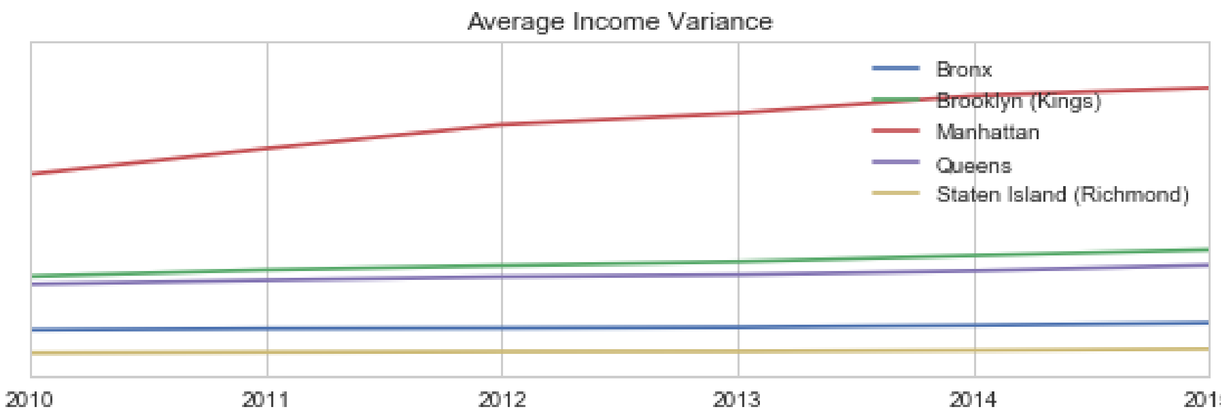
In this section, we explore the relationship between crime rates and the overall income in NYC. As mentioned before, the crime rates in the city is dropping gradually each year. Our hypothesis is that this is related to an increase of personal income each year.

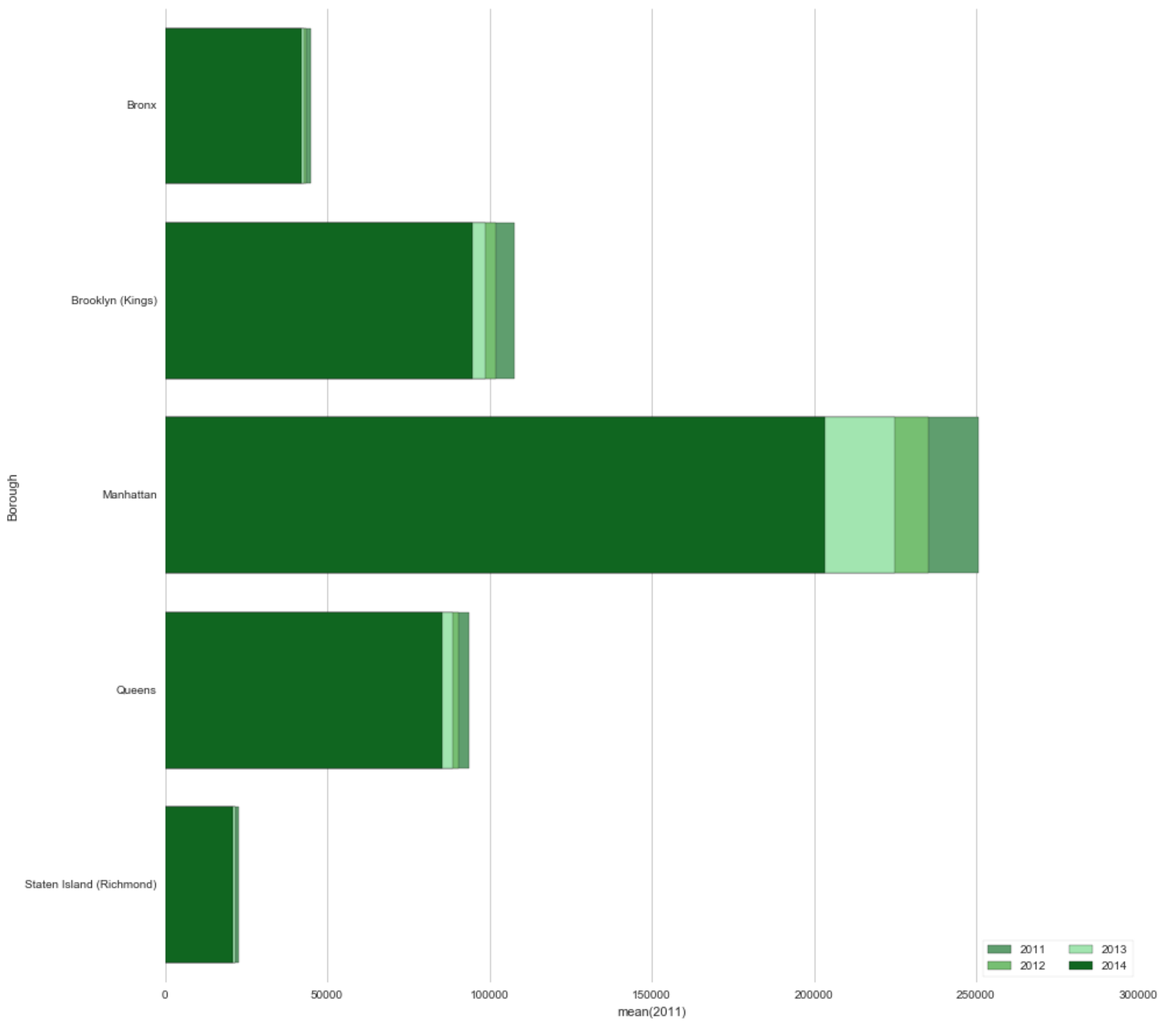
To prove our hypothesis, we analyzed and compare data of personal income obtained from Baruch College.

We plotted the average personal income of each borough in NYC from the year 2010 to 2015. We observe that personal income is indeed increasing gradually each year in all five boroughs.



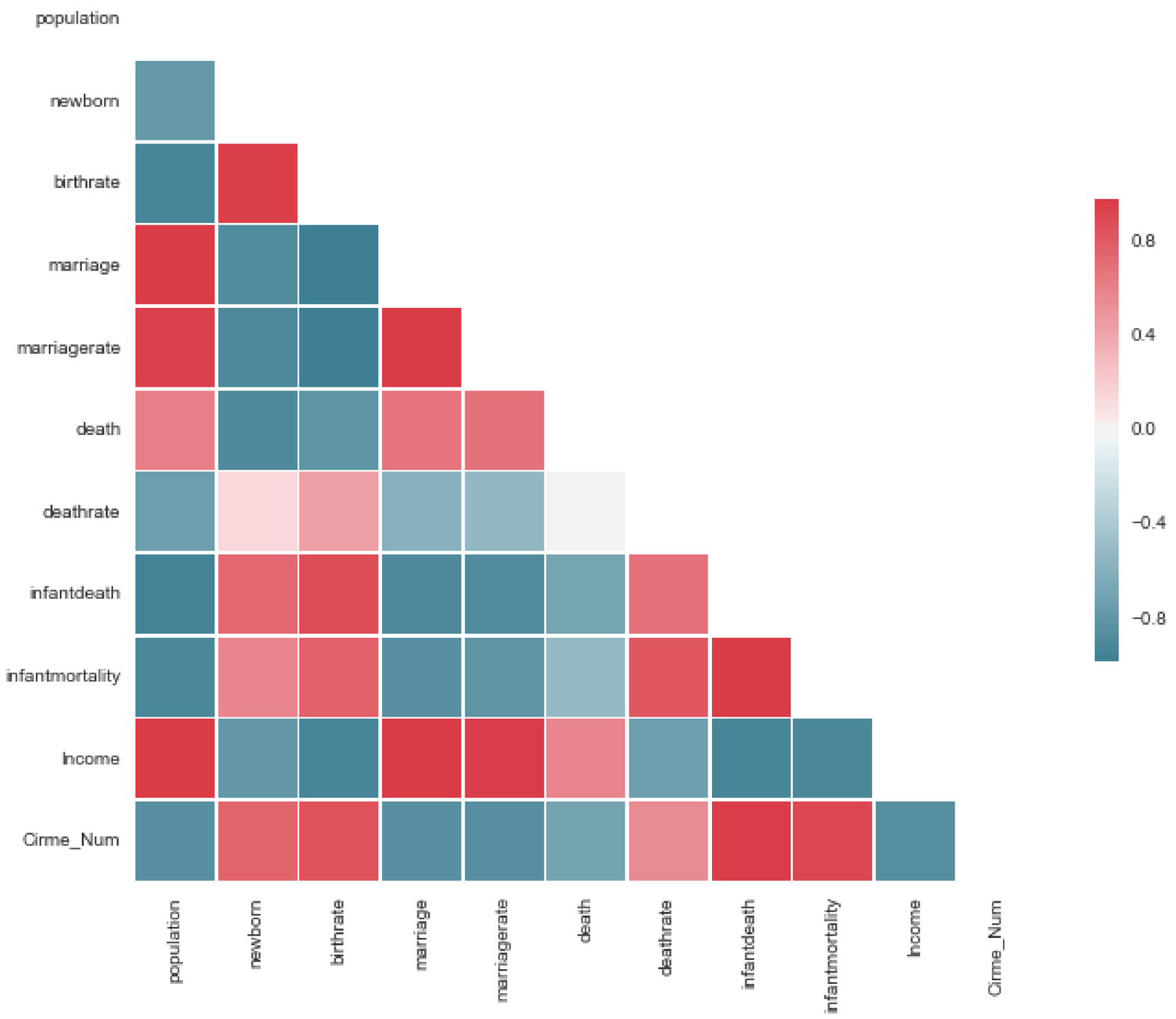




We further confirmed our analysis by joining our crime data with another dataset that includes marriage rates, birth rates, infant mortality, along with variables we analyzed before such as income and population. And we plotted a correlation heat map to give a more straightforward view of the new data.

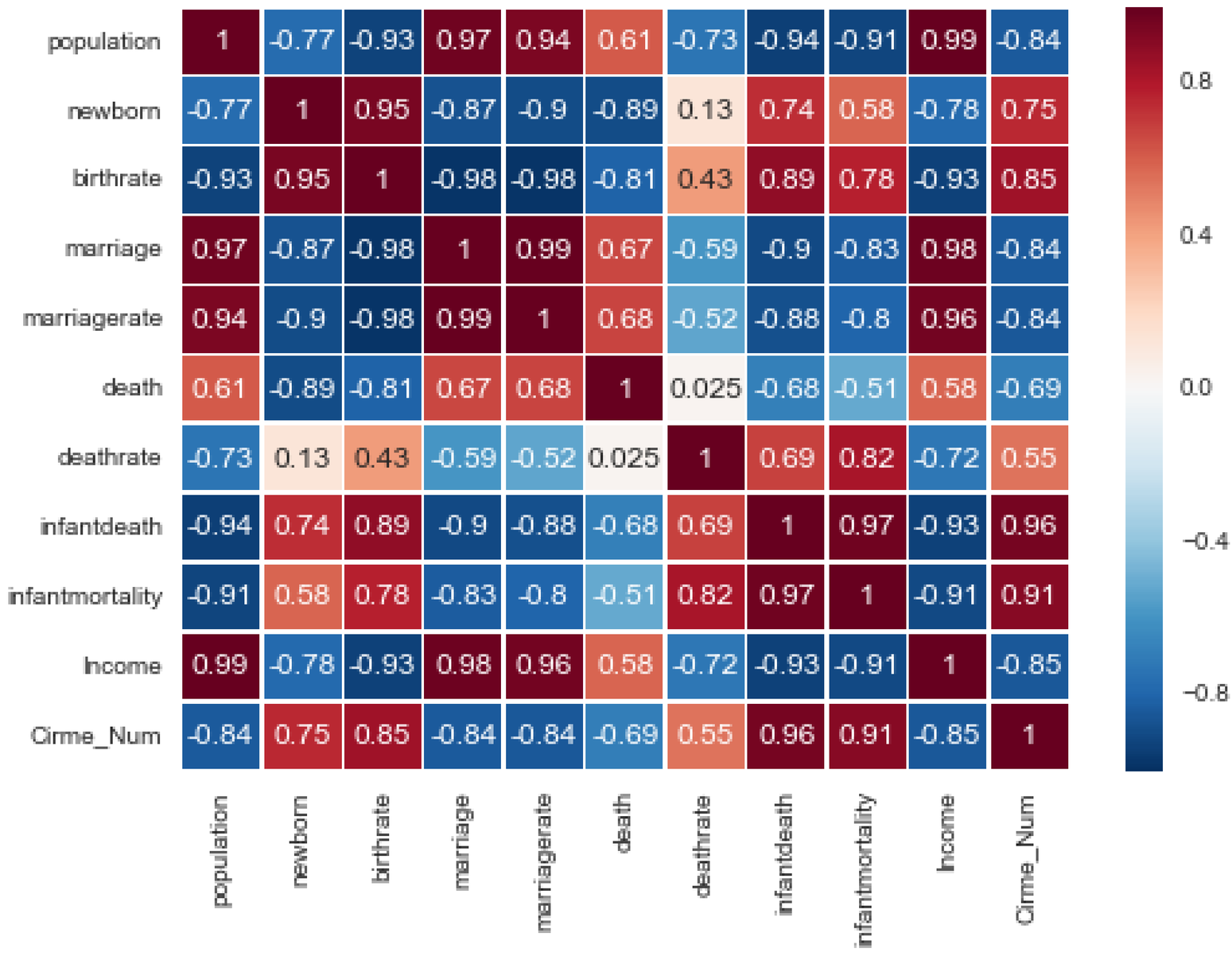
From the graph, the color red indicates a positive correlation while color blue indicates a negative correlation between variables.

The correlation coefficients are the below:



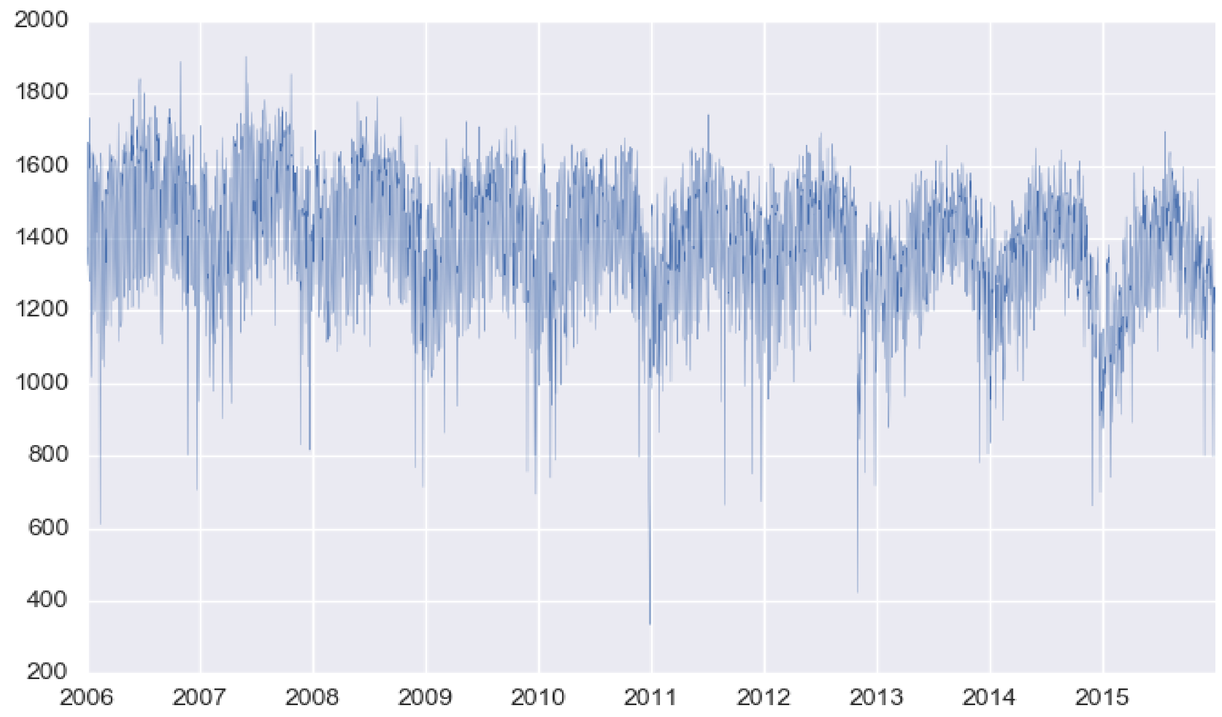

**Some remarkable findings with respect to these correlation coefficients:**

**~ Population vs Crime: -0.84**
**~ Birth Rate vs Crime: 0.85**
**~ Income vs Crime: -0.85**
**~ Marriage vs Crime: -0.84**

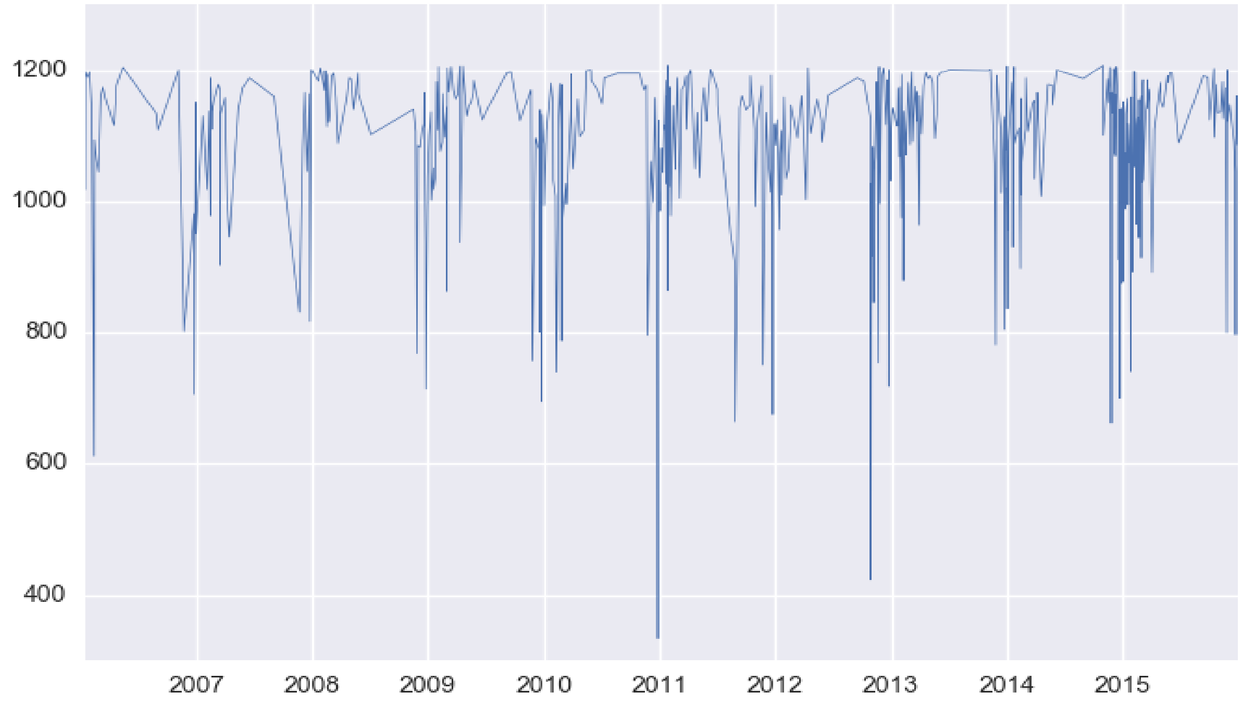


## 3) Temporal vs Crime

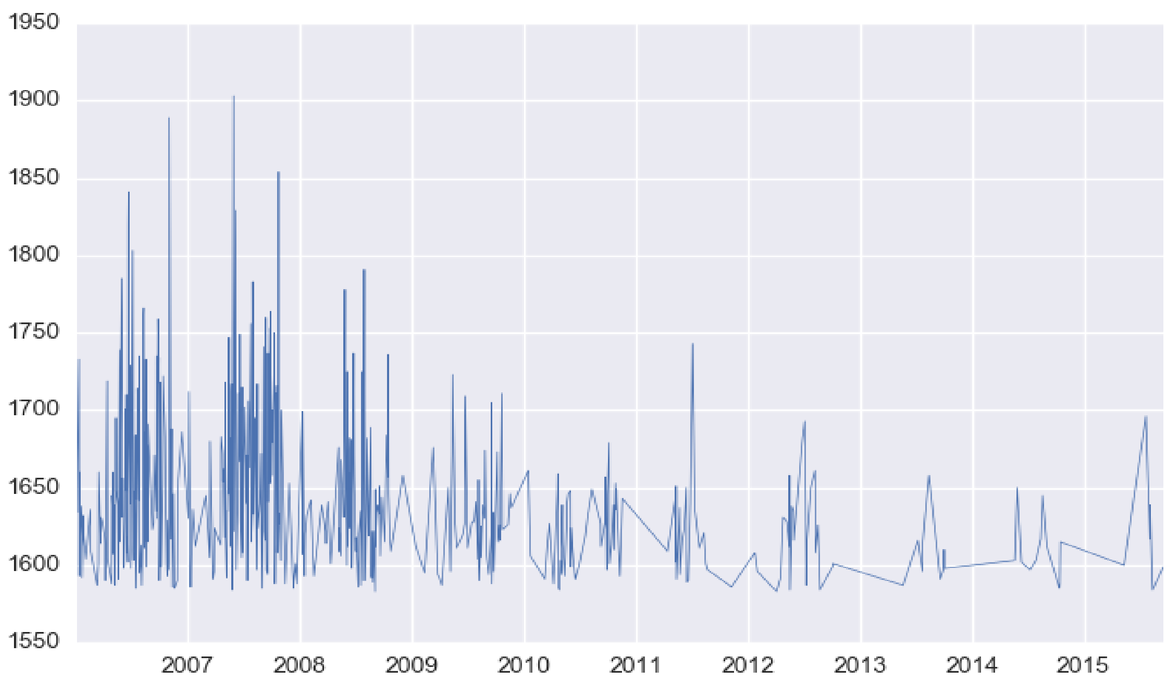
By aggreagating our NYC Crime dataset by report dates, we can see that daily reported crimes shows a periodic pattern with one year period.




We wanted to see the days with minimum number of  daily crimes. We took the days with the 500 least number of crimes. We found that there are sudden drops at the end of year every year. Looking closer into those dates, they are holidays which is Thanksgiving period, Christmas period and New Year period every year. Also, we can see more days with fewer crimes in later years than previous years.

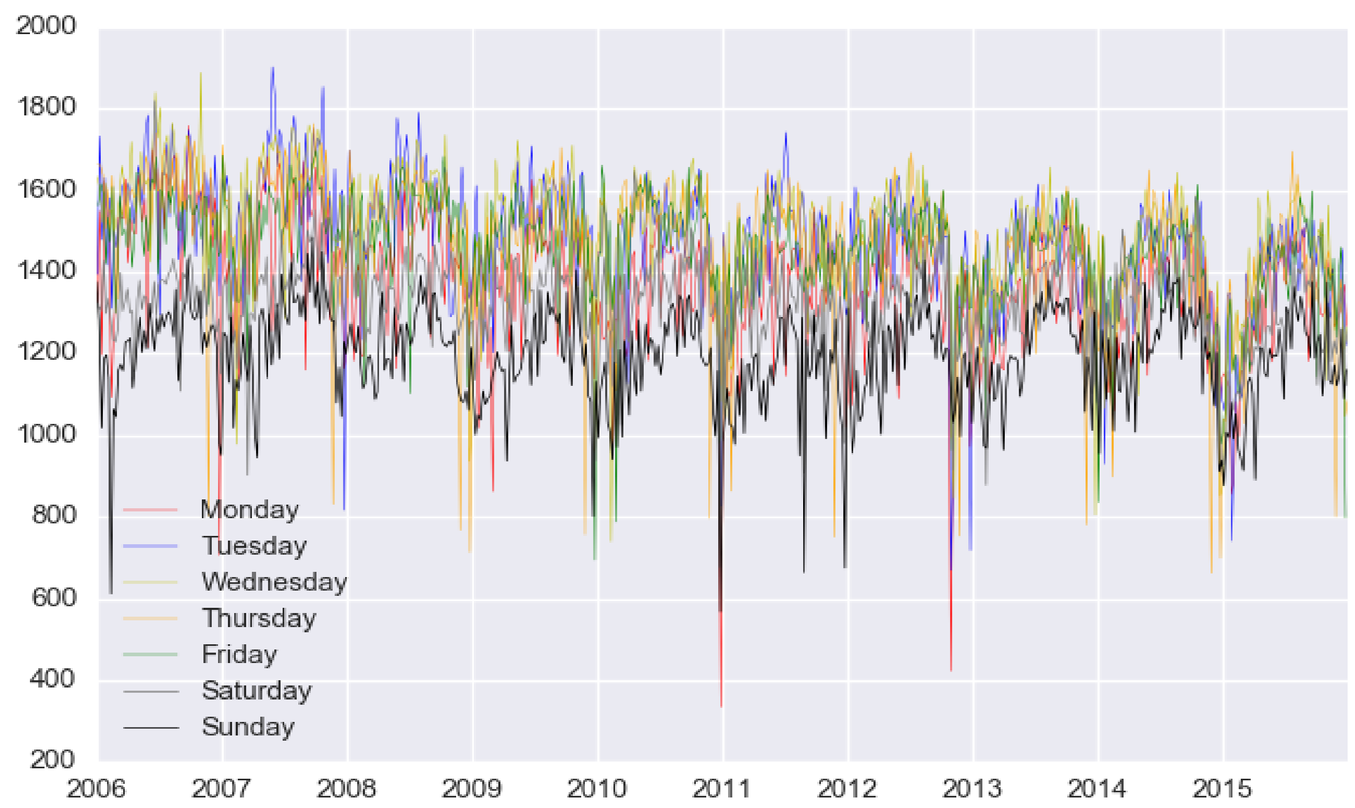



Similarly with maximum daily crimes. We took the days with the 500 highest number of crimes. We found that the highest number of daily crimes crowded in between 2006 - 2009.

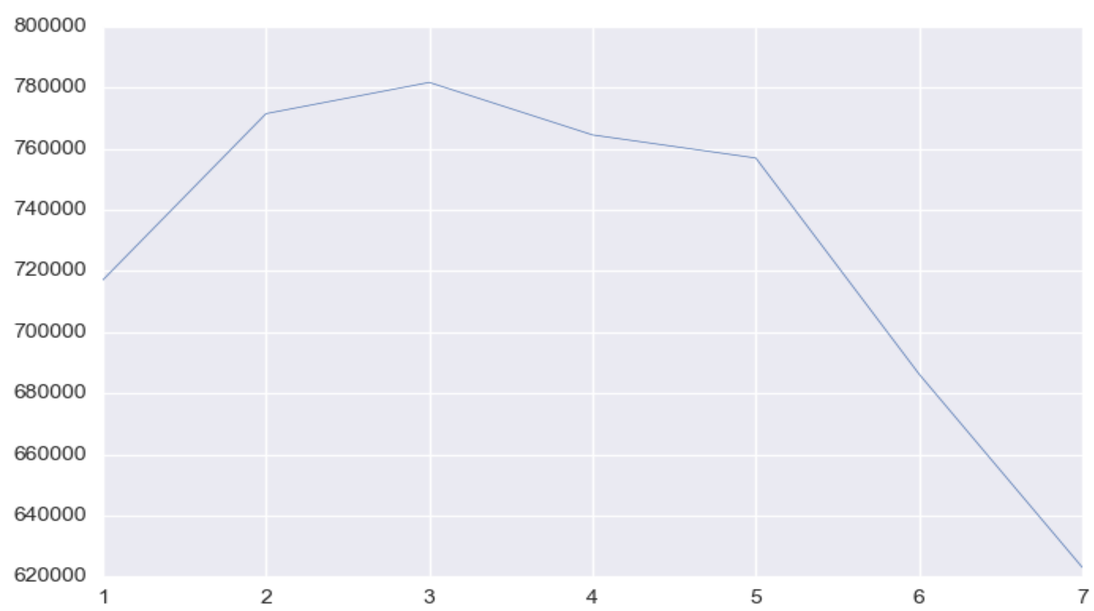


Now we plot daily crime number in 2006 - 2008. We can see that other than Thanksgiving, Christmas and New Year period, all three years have a periodic pattern in each month, with period approximately being 7 days. Thus we want to further discuss the weekday effect on the dataset.

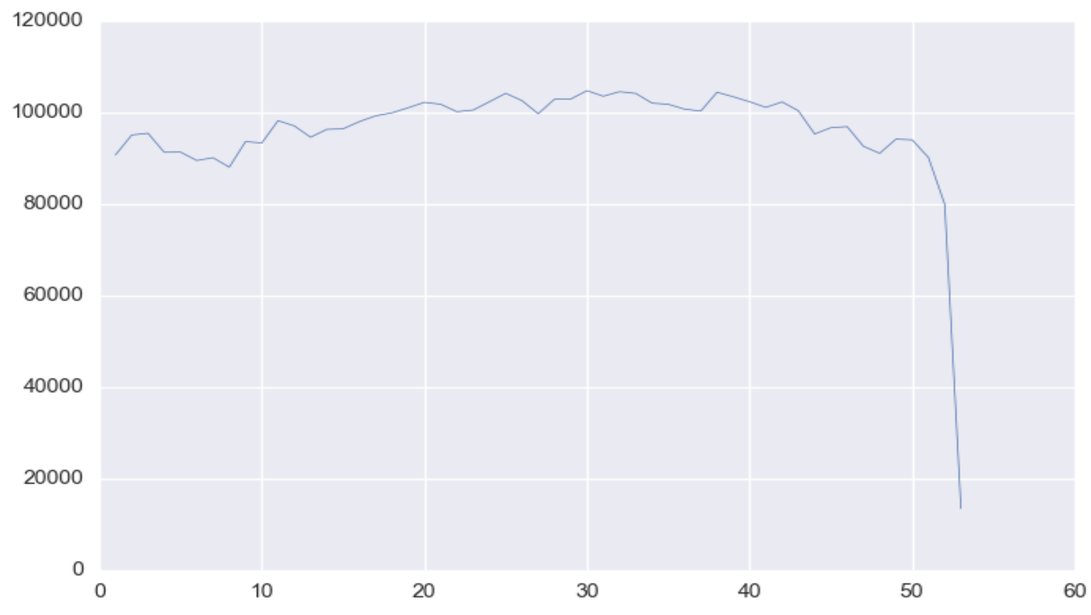

Now we discuss the effects of weekdays throughout the years. We can see that all of the Sundays have fewer crime reported than other days. There could be multiple reasons, either because people tend to commit fewer crimes on Sundays because of religions, resting at home, or because policemen rest at home on Sundays.




Thus we aggregate the crime counts by the day of the week, we could tell that Saturday and Sunday have significantly fewer cases reported.



Thus we eliminate the weekday effect by combining every 7 days to 1 week and aggregate number of crimes by weeks. We can see that the last 3 weeks of all years have decreasing number of crimes reported. The reason of few cases in 53rd week is that date.isocalendar() function assigns only 53rd week to year 2009 and 2015. But other than that, 51st and 52nd week every year still have much smaller number of crimes.



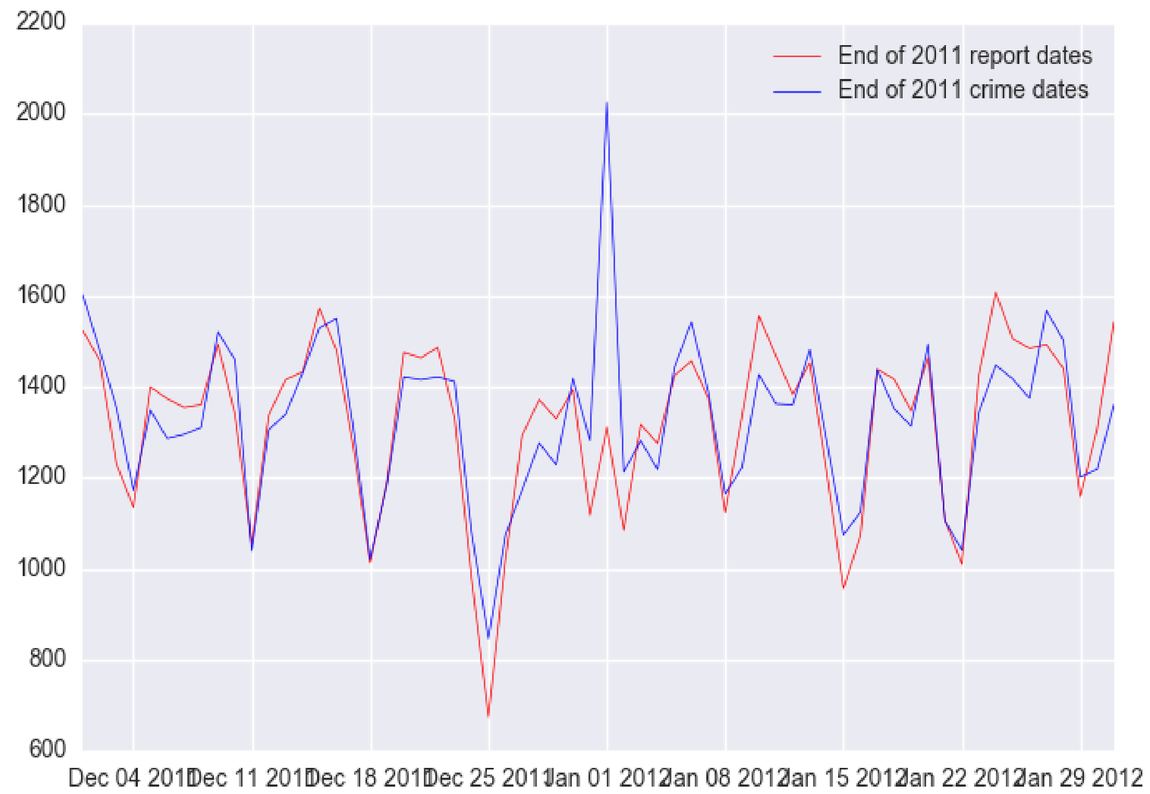

And then we plot to see how number of crimes each week changes through years. At the final weeks of year 2011, although there's a decrease of crimes, the decrease is not as sudden as other years.

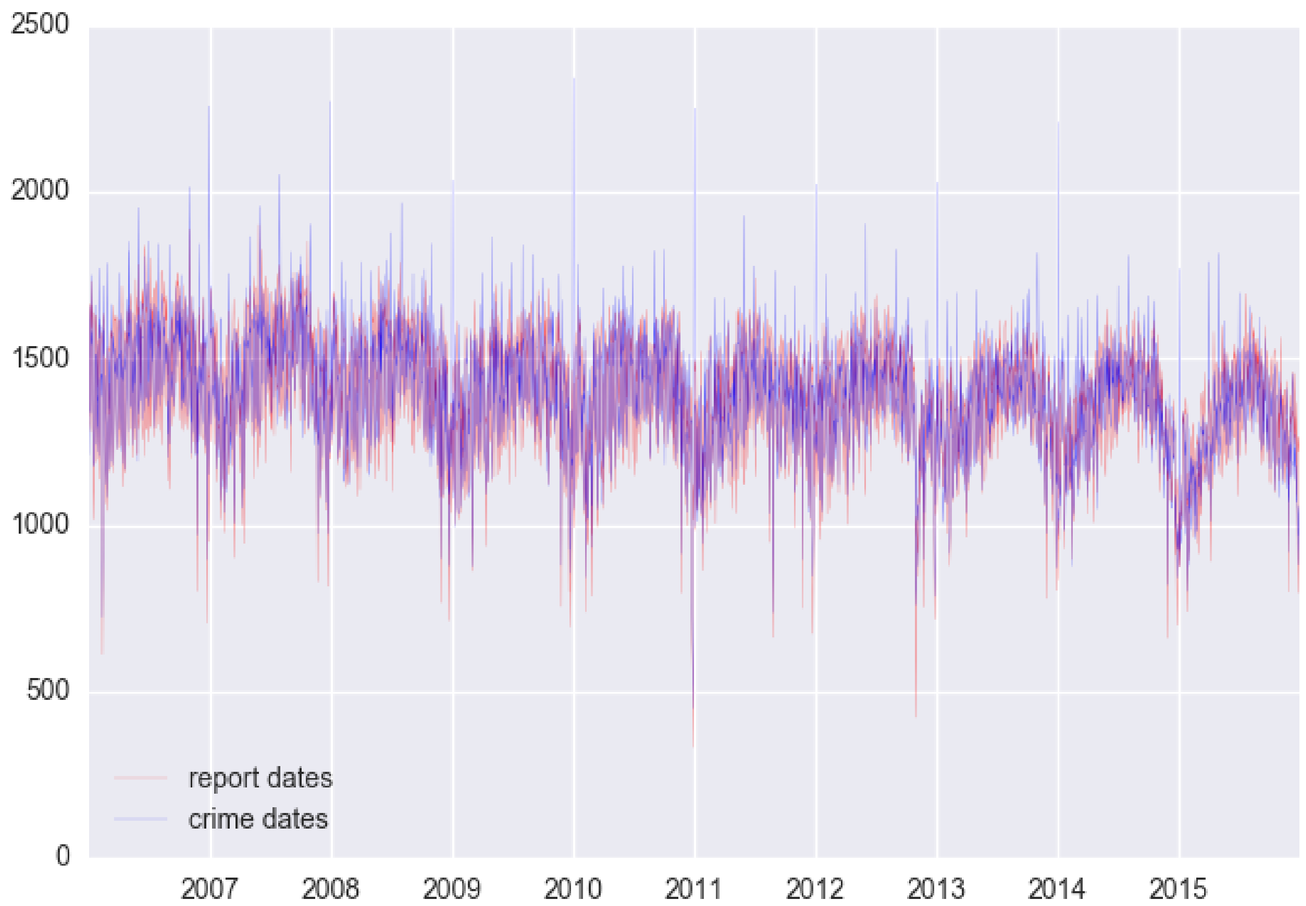

We then want to discuss that whether the trend on report dates synchronize with the trend on crime from dates. We want to explore whether the drop of crime rates during Christmas to New Year's Eve comes from fewer policemen at work or an actual decrease of crimes.

So first we plot report dates vs crime starting dates. We can see that although the drop of crimes on Christmas is synchronized, on New Year, there's a surge of crimes vs a decrease of crime reports.

Then we zoom in to year 2011-2012, other than weekend effect, there is a clear decrease of crimes happened and reported. But on the first day of 2012, there's an clear increase of crimes, much more than reported. Since it is possible that data entry specialists might have filed the unknown starting date to be the beginning of a year, the influence of New Year is still unclear.

Thus we could conclude that decrease of reported crimes on Christmas does not come from fewer policemen on duty. People do commit less crimes on Christmas. On the other hand, the effect of New Year to number of crimes is still unclear.


2200
2000
1800
1600
1400
1200
1000
800
600
End of 2011 report dates
End of 2011 crime dates
Dec 04 2011
Dec 11 2011
Dec 18 2011
Dec 25 2011
Jan 01 2012
Jan 08 2012
Jan 15 2012
Jan 22 2012
Jan 29 2012


2500
2000
1500
1000
500
0
report dates
crime dates
2007
2008
2009
2010
2011
2012
2013
2014
2015

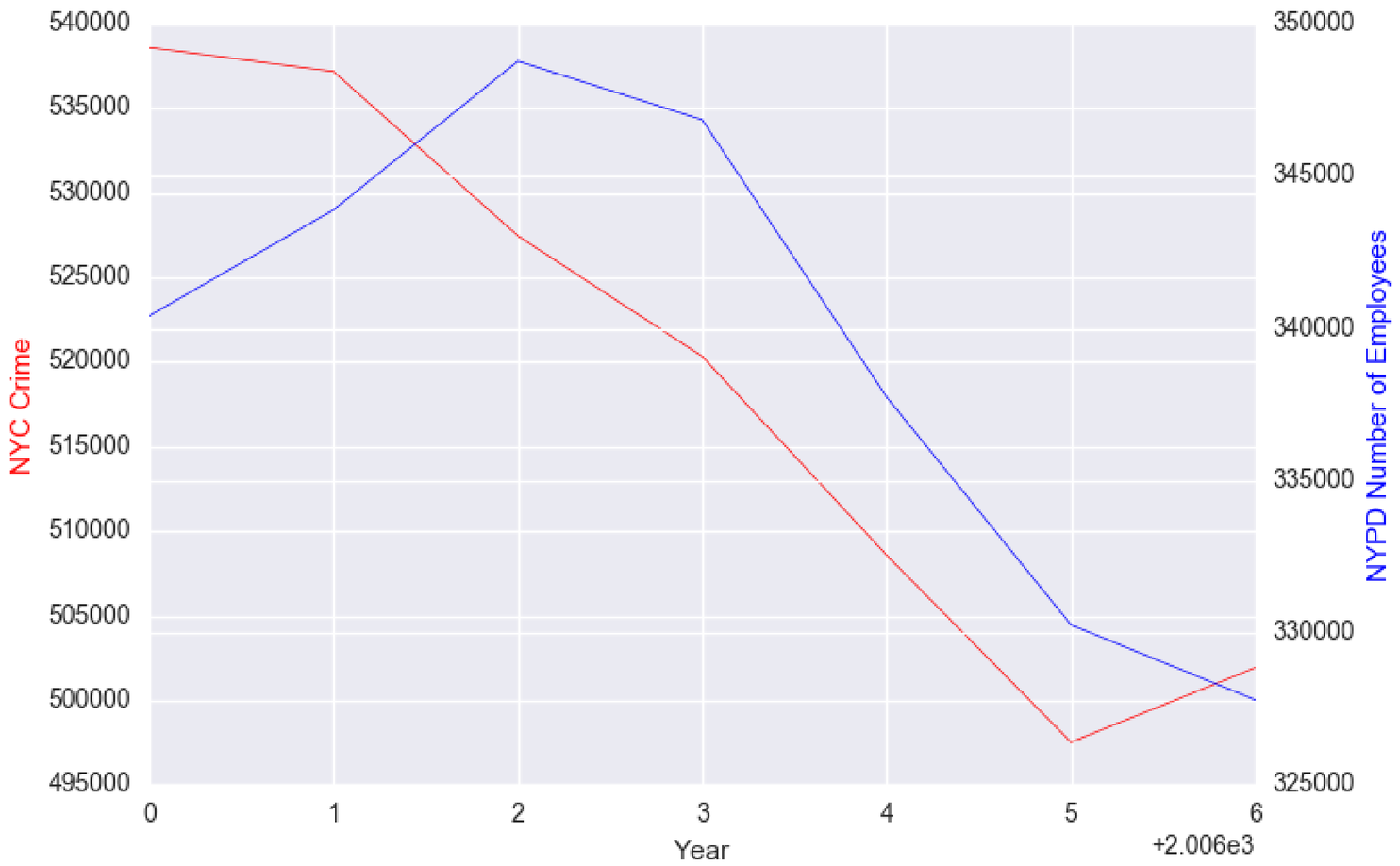
## Workforce vs Crime

We saw that there is a decrease of crimes each year since 2006-2013. We wonder if the decrease of crimes has a correlation with increase of NYPD workforce.

Thus we found the data of NYPD workforce, and examined the relationship between number of NYPD employees and total crimes each year.

**We got a correlation of 0.75542906 between NYPD workforce and crimes.**

We could see that before year 2008, NYPD keeps increasing workforce, resulting in a clear decrease of crimes. After that, NYPD workforce starts to decrease, while keeping a decreasing crime rate as well, until in year 2012, with an increase of crimes. We could conclude that NYPD workforce has a threshold of 330,000 employees to keep crime rates under control.

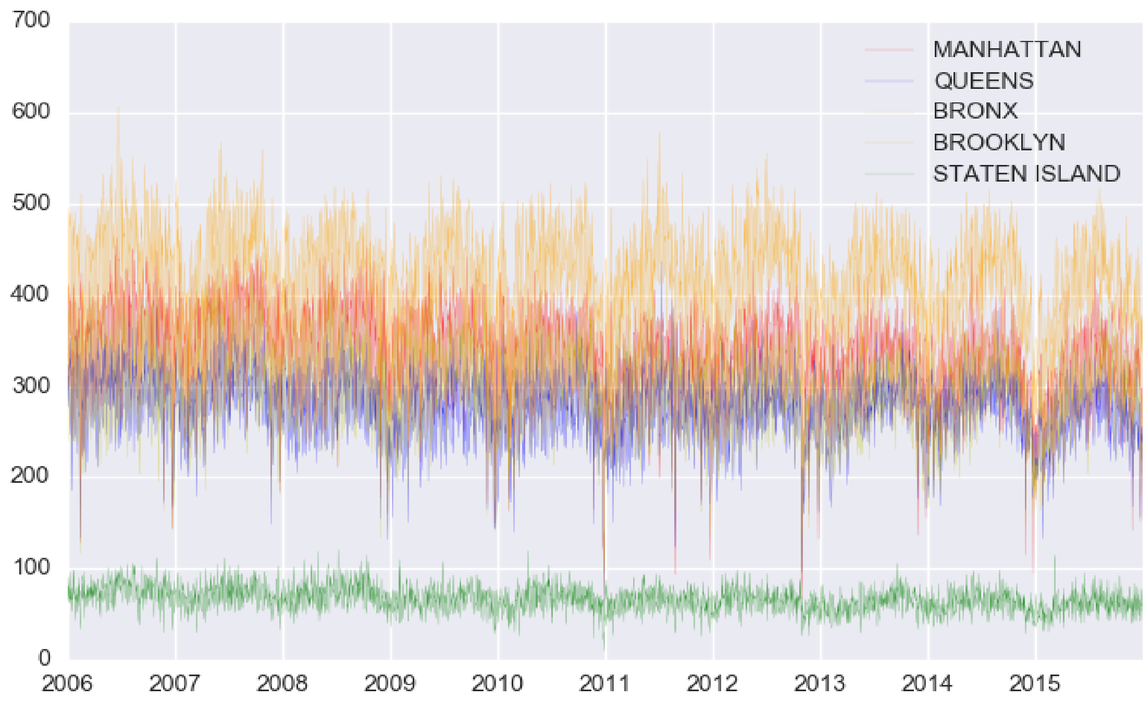


With further work on NYPD Workforce data, we came up with a continuing hypothesis: is it possible that NYPD officers work harder with increasing payments through the years, leading to a safer city?

Although the relationship between 5th percentile, 50th percentile, 95th percentile of NYPD base salary and NYC crime rates is not very clear on our plot. But after we computed the correlation coefficients of these data, we came to know the correlation coefficients are -0.4638507, -0.96516663, -0.10548772. We can see that median of NYPD base salary has a negative correlation with crime rates. The higher pay the majority of police officers get, the lower crime rates.

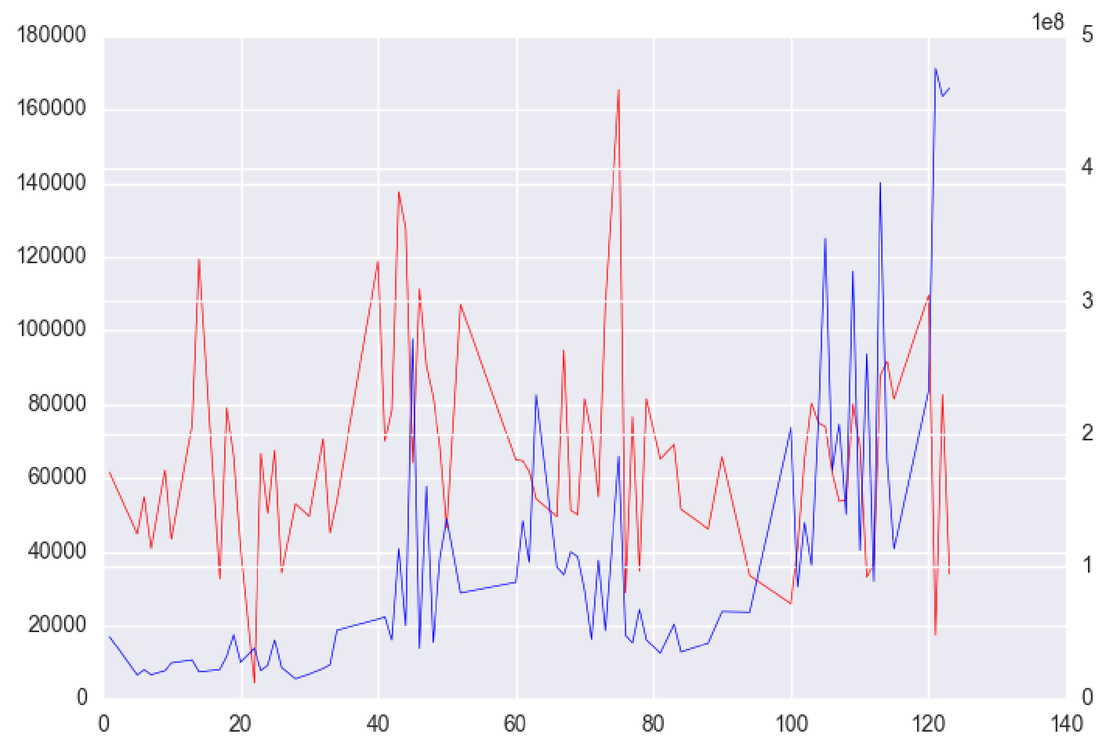
## 4) Spatial vs Crime

By aggregating our NYC Crime dataset by report dates (RPT_DT) and boroughs (BR_NM), we can see that each borough preserves the temporal patterns we found earlier. Overall, Brooklyn borough has the most number of cases, followed by Manhattan, Bronx, Queens, and Staten Island. More bar charts can be found in part one, i.e column summary of our report.




During our research, we found a New York Police Precinct dataset, with area and length of each Police Precinct. We realized that in Brooklyn, Queens and Bronx, each precinct is responsible for much larger area than the Manhattan precincts. So we wonder whether assigning different size of area to each precinct leads to different distribution of crime rates. Is it possible that precincts in Brooklyn can't handle all the cases happened in their large responsible area, and thus have a higher crime rates?

However, after we plot the total count of crimes of each precinct vs the area of each precinct, it doesn't show a clear correlation. **Moreover, their correlation coefficient is 0.00848274.** Thus we disapprove our hypothesis. Crime rates in each precinct do not have a correlation to area of each precinct.



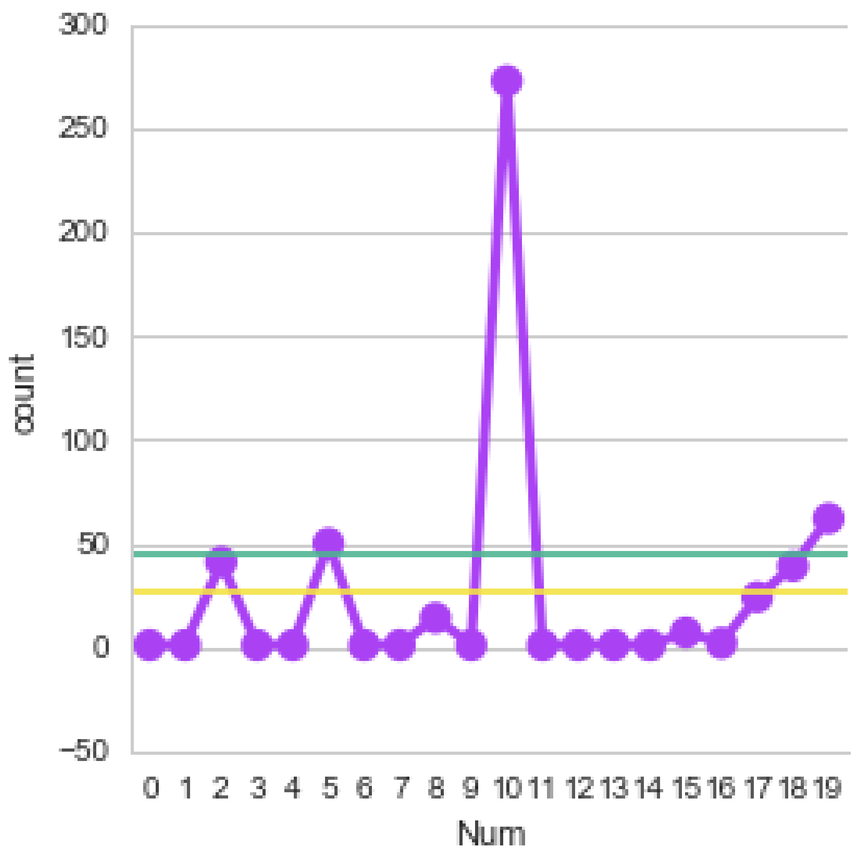
Graffiti vs Crime:

There is research suggesting that graffiti’s presence doubles the probability of people littering and stealing. Our research results support and expand on the "broken-window" theory, which forms the backbone of many crime prevention programs. The theory points out that disorder signs, such as broken windows, graffiti can open the door to individuals breaking social rules.

In the mid-1990s, New York adopted a program, "Quality of Life Campaign", removing city filth, such as graffiti. And the crime in that period did drop.

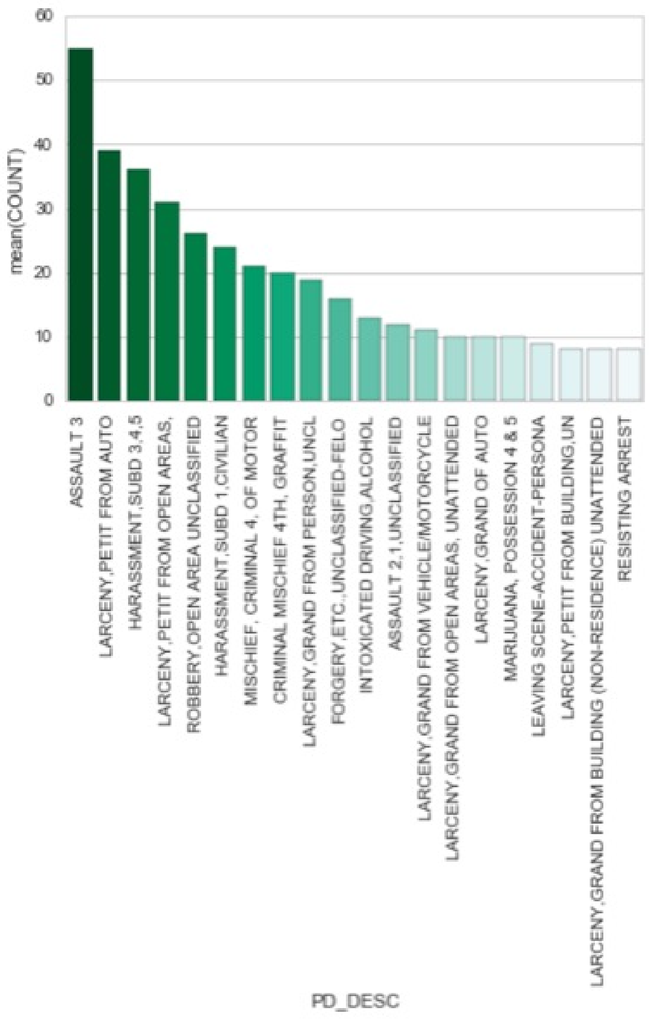
According to this research, we want to discuss whether the presence of graffiti can lead to a rise of crime rate in a neighborhood by joining NYPD data and DSNY Graffiti Information by the X and Y coordinates, and compare the average crime count of the NYPD dataset and the inner joined table, which contain the crime information that happened in the neighborhood with graffiti.

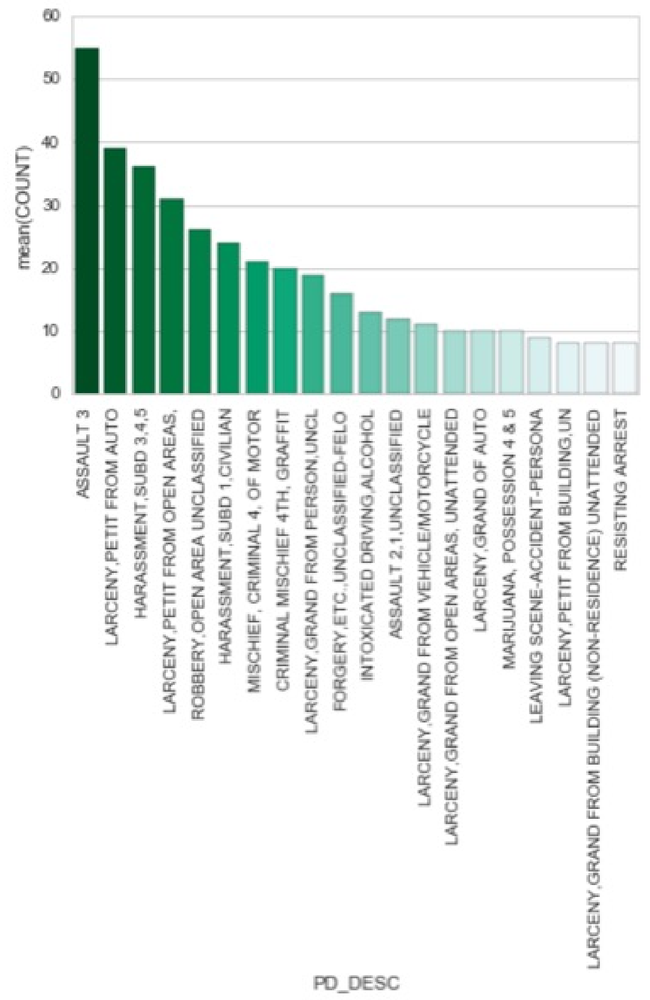
The result is shown as follows:




The green line is the average count in each neighborhood of the NYPD data set, which is 45.33, and the yellow line is the average count in neighborhoods with graffiti, which is 26.95. The purple line reflects the crimes happened in each area.

From this figure, we can find that there is no clear correlation between the graffiti neighborhood and high crime rate area. And the abnormally high crime rate is in the 46 STREET & QUEENS BOULEVARD, which is notable for high crime rates.




These two figures show the top 20 crime description for the joined crime_graffiti table and NYPD table. From the Figure, we can conclude that the top crime type is assault 3 and in graffiti neighborhoods, larceny types crime and robbery types crime are more likely to happen.

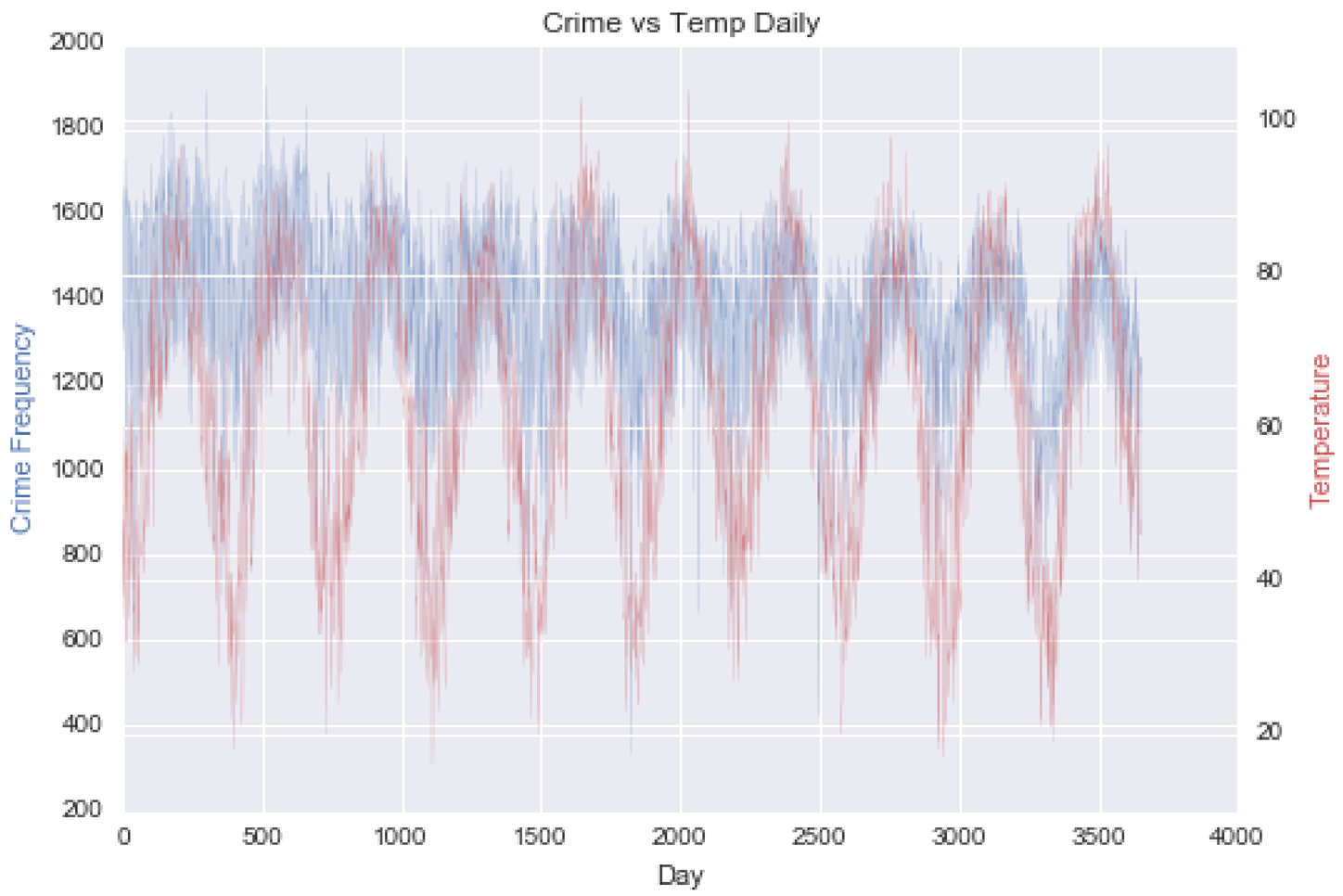
The reasons why our hypothesis is disproved lie in three aspects. First, it may due to small graffiti dataset. Since NYPD data set is 1.2G and graffiti data set is 1.5M, which means, compared to NYPD data, graffiti data provides far less information. Second, the data range in the NYPD is 2006 – 2015, while the graffiti data 2016-2017. The graffiti information cannot explain the history crime happened before 2016. Third, the graffiti information contained may refer to the street art instead of casually graffiti. So the graffiti contained in the graffiti table is unrelated to the crime rate. But through the exploration, we find graffiti presence do have influence on the crime.

## 5) Weather vs Crime

Besides the above mentioned aspects, we also investigated whether temperature of the day has played any role in affecting crime rates in the city. Our hypothesis is that in winter, people will tend to stay indoor to keep warm and consequently lead to a lower crime rate. While in the summer, people are more active and hyped, which may trigger more crimes to occur.

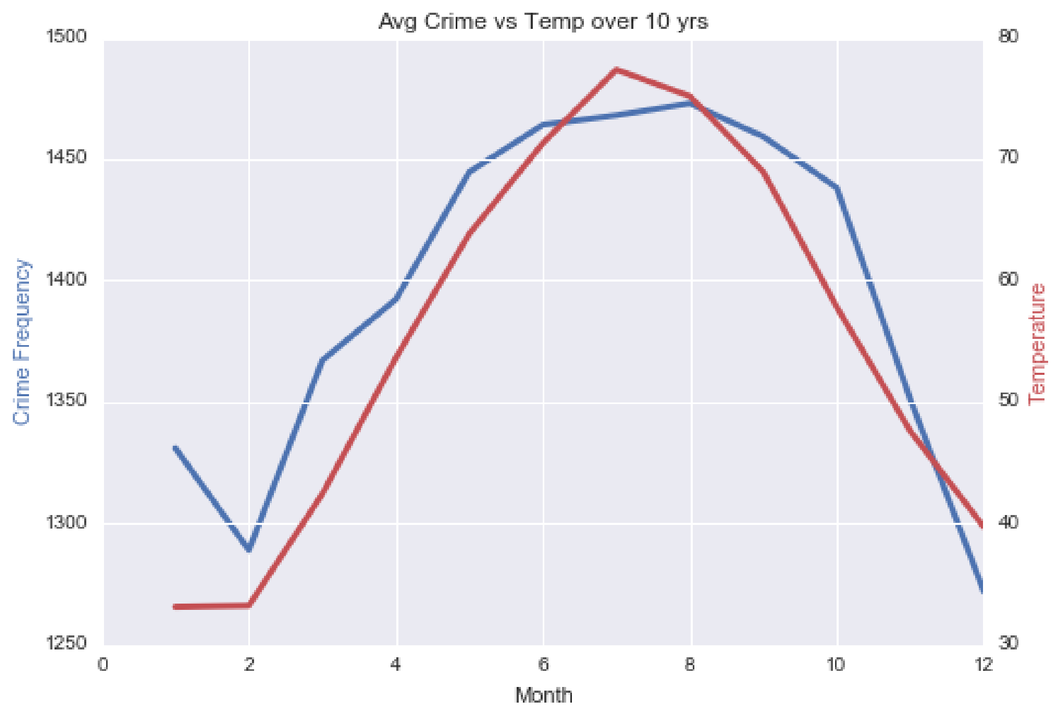
We obtained our weather data from NOAA (National centers for environmental information). The dataset has 7 columns which include max and min temperature, wind speed, and precipitation records in each day from year 2006 to 2015.

We joined the table with our summed crime dataset with dates as keys. The merged dataset has 3652 rows. To prove our hypothesis, we first plotted two lines with temperature and crime counts in each day over the ten years. The resulting graph is surprisingly supportive. The crime counts (blue line) and temperature (red line) are in sync with approximately the same periods. This proves our hypothesis.




Next we aggregated the counts by monthly average over 10 years to provide a more direct graph. Again, the lines have approximately the same overall shape, with peaks during summertime in July and downs in colder days at the very beginning and end of each year.

**The correlated coefficient of temperature and crime is 0.938.**



# Part III: Summary & References

**Summary:**

NYC crime rates have an overall tendency of dropping through year 2006 to 2015. Through comprehensive analysis, we conclude a list of possible features that affect NYC crime rates along with respective correlation coefficients:

(Crime counts, Mobility population, 0.955)
(Crime counts, Population, -0.84)
(Crime counts, Birth Rate, 0.85)
(Crime counts, Income, -0.85)
(Crime counts, Marriage, -0.84)
(Crime counts, Workforce, 0.755)
(Crime counts, Temperature, 0.938)

Furthermore, during holidays which is Thanksgiving period, Christmas period and New Year period every year, the overall crime rate drops. Within a week, Saturdays and Sundays have fewer crimes reported than other days. We also observed that the number of employees in NYPD has a positive correlation with crime rates.

Geographically, Brooklyn borough has the most number of cases, followed by Manhattan, Bronx, Queens, and Staten Island. In areas with more street art and graffiti, we could not find a link between these areas and overall crime rates.

Last but not least, the higher the temperature, the more likely crimes will occur.

**Individual Contributions:**

All of the teammates contributed to data exploration, data cleaning and overall summaries.

Yurui (ym1495): Temporal vs Crime, Demographic vs Crime
Li Lin (llq205): Weather vs Crime, Income vs Crime
Yidi (yz3464): Spatial vs Crime, Income vs Crime

**Experimental Setup & Cluster configuration:**

We used MapReduce for all Part I summaries. Throughout our analysis, one mapper and one reducer are used for each job. The resulting '.out' files are then used in matplotlib libraries to generate plots.

For Part II, we incorporated Spark to optimize our code efficiency for joining large tables together.

## Data Source & References:

1) Original Crime Dataset:

~ NYPD Complaint Data Historic @NYC Open Data (2006 - 2015)
Link: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

2) Demographic Data:

~ Population Statistics by Baruch College
Link: http://www.baruch.cuny.edu/nycdata/population-geography/pop-characteristics.htm

~ Abortion Surveillance — United States, 2013
Link: https://www.cdc.gov/mmwr/volumes/65/ss/ss6512a1.htm

3) Temporal Data:

~ NYPD Workforce Profile Report
Link:
http://www.nyc.gov/html/dcas/downloads/pdf/misc/workforce_profile_report_12_30_2013.pdf

4) Weather Data:

~ NOAA Weather Service
Link: https://www.ncdc.noaa.gov/cdo-web/

5) Income Data:

~ NYC Personal Income by Baruch College
Link: http://www.baruch.cuny.edu/nycdata/income-taxes/personal_income.htm

6) Spatial Data:

~ DSNY Graffiti Information @ NYC Open Data

Link: https://data.cityofnewyork.us/City-Government/DSNY-Graffiti-Information/gpwd-npar
http://www.livescience.com/7599-graffiti-triggers-crime-littering.html
http://www.denverpost.com/2008/07/18/new-study-on-graffiti-crime-correlation/

~ Police Precincts @ NYC Open Data
Link: https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz/data