NYPD Crime Data Analysis

Yaohan Ke (yk1587)
Wenjie Sun (ws854)
Xinyan Yang (xy975)
Github link: https://github.com/sherylke/ds1004project

Abstract

In this project, we first investigate the data quality of the new york city compliant data available on the NYC open data website. And then, we generated the trend of the data to discover some global trend of this data and carried out some hypothesis about the data. For this part, PySpark is for sure needed to process more than 5 Million records in an efficient way. Also, we generated the trend of the data to discover some global trend of this data and carried out some hypothesis about the data and visualize the trend underlined in this dataset. We found that there seems to be a decreasing trend in the crime frequency by year and across locations the crime frequency also seem to be different. Our hypotheses were that the trend of crime frequency in time may possibly related with unemployment rate and motor vehicle collision frequency through time. And we suspect that the crime frequency in different locations may be correlated with motor vehicle collisions and housing prices in the same regions.

For the second part, we join the data with 3 other open databases (NYC motor collision data available on the NYC open data, NYC housing data available on the Zillow, NYC Unemployment data available on the US labor department website) to prove or disprove these hypotheses.

Finally we find out 2 pairs of correlated attributes:
(Cirme_frequency_by_zipcode, Collision_frequency_by_zipcode, 0.7275),
(Cirme_frequency_by_time, Collision_frequency_by_time, 0.7438);
and 2 pairs of uncorrelated attributes:
(Cirme_frequency_by_zipcode, Housing_price_by_zipcode, -0.1358),
(Cirme_frequency_by_mounth, unemployment_rate_by_mounth, -0.2330).

1. Introduction

New York City was not considered a safe city back in 70s. Over these years, more and more people consider NYC as a safe city to live. Our project collected the crime data from 2006 to 2015 for an in-depth study of these 10 years to understand whether there was more or less crime in the 10-year time frame by leveraging the computation power of PySpark. There seems to be a decreasing trend in the crime frequency by year and across locations the crime frequency also seem to be different, and we would like to find out the reason behind these phenomena and get more insights by finding the relationships with other data sets.

2. Part I: Data summary and data quality issues

To understand how the dataset looks, we would explore it in terms of data summary and data quality issues. Then we would like to find out if there is any trend inside the data.

2.1 Data Summary

The dataset we used is downloaded from NYC Open Data. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of 2015. We can find out more detailed information about each column from the ‘NYPD_Incident_Level_Data_Column_Descriptions.csv’ file.

Column 0 (‘**CMPLNT_NUM’)** gives the randomly generated persistent ID for each complaint. For column 0 we checked if the complaint number is an integer and if it is a unique value in the whole dataset.

VALID	5101231

(However, code:**column0_data_quality.py** and **column0_data_summary.py** require a large physical memory. The reason is in the code, I ran count of each value, which then be left-joined by the value self. I am using this code to make sure each of value is unique (count = 1). However, it seems that these 2 codes don't always work, depending on the hadoop capacity.

Since I have already checked that this column has all unique value (which is qualified for being a primary key), I uploaded a v2 code that doesn't check uniqueness of each value. Instead, v2 only checks if a value is integer.)

Column 1 (**CMPLNT_FR_DT**) gives the exact date of occurrence for the reported event (or starting date of occurrence, if column 3 (‘CMPLNT_TO_DT’) exists). For column 1 we checked if the date is legal and if the cmplnt_from date time is smaller than cmplnt_to date time. Most data has a valid date, a small percent is missing some date, and 4 invalid date.

INVALID	4
VALID	5100572
NULL	655

The below 4 lines are invalid because to_datetime is earlier than from_datetime.

['06/22/2014', '23:00:00', '06/22/2014', '00:0...	INVALID
['11/11/2010', '07:00:00', '11/10/2010', '18:0...	INVALID
['11/15/2009', '10:00:00', '11/14/2009', '12:0...	INVALID
['05/19/2006', '16:00:00', '05/14/2006', '14:0...	INVALID

Column 2 (‘**CMPLNT_FR_TM’**) gives the exact time of occurrence for the reported event (or starting time of occurrence, if column 4 (‘CMPLNT_TO_TM’) exists). For column 2 we checked if the time is legal and if the cmplnt_from date time is smaller than cmplnt_to date time.

23:45:00	TIME	Compliant from time	VALID
23:36:00	TIME	Compliant from time	VALID
23:30:00	TIME	Compliant from time	VALID
23:30:00	TIME	Compliant from time	VALID
23:25:00	TIME	Compliant from time	VALID

It turns out that there are many invalid data in this column.

INVALID	907
VALID	5100276
NULL	48

Most of the invalid data is due to a time format ("24:00:00") is not accepted in Python. This will be corrected by "23:59:59" in the future data analysis.

Column 3 (‘**CMPLNT_TO_DT’**) gives the ending date of occurrence for the reported event, if exact time of occurrence is unknown. For column 3 we checked if the date is legal and if the cmplnt_from date time is smaller than cmplnt_to date time.Since to_date is not required, there are many NULL data.

INVALID	4

VALID	3709749
NULL	1391478

These 4 invalid data is because of from datetime is later than to datetime.

['06/22/2014', '23:00:00', '06/22/2014', '00:0...	INVALID
['11/11/2010', '07:00:00', '11/10/2010', '18:0...	INVALID
['11/15/2009', '10:00:00', '11/14/2009', '12:0...	INVALID
['05/19/2006', '16:00:00', '05/14/2006', '14:0...	INVALID

Column 4 (‘**CMPLNT_TO_TM’**) gives the ending time of occurrence for the reported event, if exact time of occurrence is unknown. For column 4 we checked if the time is legal and if the cmplnt_from date time is smaller than cmplnt_to date time. Since to_date is not required, there are many NULL data. Invalid data is still because of "24:00:00".

INVALID	1380
VALID	3712066
NULL	1387785

These 4 invalid data is because of from datetime is later than to datetime.

['06/22/2014', '23:00:00', '06/22/2014', '00:0...	INVALID
['11/11/2010', '07:00:00', '11/10/2010', '18:0...	INVALID
['11/15/2009', '10:00:00', '11/14/2009', '12:0...	INVALID
['08/31/2009', '23:55:00', '08/31/2009', '24:0...	INVALID
['08/30/2009', '06:30:00', '08/31/2009', '24:0...	INVALID

Column 5 (**‘RPT_DT’**) gives the date the event was reported to police. For column 5 we checked if the date is legal and also is between 2006 and 2015 (since the report is noted that the time range is based on the report time). All the data falls in the valid time range.

VALID	5101231

Looking at the stacked the line chart, the decreasing trend is not very clear. After plotting the trend borough by borough, it is more clear that Brooklyn contributes the most to the decrease, followed by manhattan.

Column 6 (**‘KY_CD’**) gives the three digit offense classification code (the Key Code). For column 6 we checked if the offense classification codes are three digit integers. Data in this column are all valid.

VALID	5101231

Column 7 (**‘OFNS_DESC**‘) gives the Offense Description corresponding with Key Code. For column 7 we checked if the offense description is a valid text object. Some of the data are missing.

VALID	5082391
NULL	18840

There’s actually a problem since column 6 and column 7 are suppose to be matched, but it appears that column 6 has no missing value while column 7 does. It may be some error of data input.

Column 8 (**‘PD_CD’**) gives the three digit internal classification code which are more specific than Key Code. For column 8 we checked if the internal classification code is a three digit integer. Some of the data are missing.

VALID	5096657
NULL	4574

Column 9 (**‘PD_DESC**‘) gives the description of internal classification corresponding with PD code which are more specific than Offense Description. For column 9 we checked if the internal classification description is a valid text object. Again some of the data are missing.

VALID	5096657

NULL	4574

It appears to be reasonable since the number of missing values for column 8 and column 9 are equal, and they are supposed to be matched.

Column 10 (‘**CRM_ATPT_CPTD_CD’**) gives the indicator of whether a crime was successfully completed or attempted, but failed or was interrupted prematurely. For column 10 we checked if the there is only two indicators of 'completed' and 'attempted' representing crime completeness and whether there are missing data in this column. There are some missing data.

VALID	5101224
NULL	7

Column 11 (‘**LAW_CAT_CD’**) gives the three level of offense: felony, misdemeanor, violation. For column 11 we checked if the level of offense is only in three categories of 'felony', 'misdemeanor' and 'violation'. All the data are valid.

VALID	5101231

Column 12 (‘**JURIS_DESC**‘) describes the jurisdiction responsible for incident. (Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.) For column 12 we checked if the jurisdiction description is a valid text object. All the data are valid.

VALID	5101231

Column 13 (‘**BORO_NM**‘) gives the borough in which the incident occurred. For column 13 we checked if the borough in which the incident occurred is one of the five borough names as 'Bronx', 'Manhattan', 'Brooklyn', 'Queens' and 'Staten island'. There are some missing values.

VALID	5100768
NULL	463

Column 14 (‘**ADDR_PCT_CD**‘) gives the precinct in which the incident occurred. For column 14 we checked if the precinct code is an integer and is valid compared to the full list of precinct codes of NYPD (according to http://www.nyc.gov/html/nypd/html/home/precincts.shtml). There are some missing value but the remaining data are all valid.

VALID	5100841
NULL	390

Column 15 (‘**LOC_OF_OCCUR_DESC’**) gives the specific location of occurrence in or around the premises: inside, opposite of, front of, rear of. For column 15 we checked if the premises description is a valid text object. There are many missing values probably because it is sometimes difficult to identify a specific premise location of some crimes.

VALID	3973890
NULL	1127341

Column 16 (‘**PREM_TYP_DESC’)** gives the specific description of premises; grocery store, residence, street, etc. For column 16 we checked if the location description of occurrence is a valid text object as one of the location in 'inside', 'outside', 'opposite of', 'front of', 'rear of'. There are some missing data.

VALID	5067952
NULL	33279

Column 17 (‘**PARKS_NM**‘) gives the name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included).For column 17 we checked if the park name is a valid name of an NYC public park, playground or greenspace (according to https://www.nycgovparks.org/park-features/parks-list?boro=X). There many missing data since only crimes happened in parks would have park names.

VALID	7599
NULL	5093632

Column 18 (‘**HADEVELOPT**‘) gives the name of NYCHA housing development of occurrence, if applicable. For column 18 we checked if the NYCHA Housing Developtment name is a valid name compared to the full list at http://www1.nyc.gov/site/nycha/about/developments.page. There are many missing data since only crimes occurred in NYCHA housing development would have a name in this column.

VALID	253205

NULL	4848026

For column 19 (‘**X_COORD_CD**‘) we checked if the X-coordinate is an integer and if it is in New York City. type: X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104). There is no INVALID data in our dataset although some of them are missing.

VALID	4913085
NULL	188146

For column 20 (‘**Y_COORD_CD**‘) we checked if the Y-coordinate is an integer and if it is in New York City. type: Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104). There is no INVALID data in our dataset although some of them are missing.

VALID	4913085
NULL	188146

For column 21 (‘**Latitude’**) we checked if the Latitude is an float and if it is valid in New York City. type: Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326). There is no INVALID data in our dataset although some of them are missing.

VALID	4913085
NULL	188146

For column 22 (‘**Longitude**‘) we checked if the Longitude is an float and if it is valid in New York City. type: Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326). There is no INVALID data in our dataset although some of them are missing.

VALID	4913085
NULL	188146

For column 23 (‘Lat_Lon’) we checked if the latitude and longitude is an float and if it is valid in New York City. type: (Latitude, Longitude) coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326). There is no INVALID data in our dataset although some of them are missing.

VALID	4913085
NULL	188146

2.2 Data Quality Issues

From the summary of the data we find some quality issues that we can look further into and try to explain the reason and stay cautious when processing data in the future.

Compliant from / to time:
the original data set used “24:00:00” for 12 am midnight. However, this is not accepted by Python datetime package. We have summarized number of occurrence in the data quality part in the report. And for the trend analysis purpose, we have fixed the issue by changing it to “23:59:59”.

Park name / HA Development:
These values are publicly available online. However, after implementing the filter code, we noticed that some “invalid” data was due to a different name convention. We write codes to generate all these potential misclassified invalid names and manually reviewed them.

For column 17 Parks Names, we found the list of NYC park names from the website (https://www.nycgovparks.org/park-features/parks-list?boro=X) and compared our data in this column to the list (we put it in 'parks_nm.txt'). We treat those names not in this list as 'INVALID' and print them out in the 'column17_data_quality_issue.out' file.

After looking into those names from the website we found that those invalid park names are actually not invalid but due to some difference in the naming methods. For example 'White Playground Manhattan' in column 17 is actually that of 'White Playground' in the website, which are essentially the same park. Therefore aftering examine these ‘invalid’ names we confirmed that the names in column 17 are all valid values.

For column 18 we found the list of NYCHA housing development names from the website (http://www1.nyc.gov/site/nycha/about/developments.page) and compared our data in this column to the list (we put it in 'hadevelopt.txt'). We treat those names not in this list as 'INVALID' and print them out in the 'column18_data_quality_issue.out' file.

After looking into those names from the website we found that those invalid names are actually not invalid but due to some difference in the naming methods. For example '33-35 SARATOGA AVENUE' in column 18 is actually that of 'Saratoga Village' in the website, which are probably the same place. Therefore aftering examine these ‘invalid’ names we confirmed that the names in column 18 are all valid values.

Offense classification code, offense description:

There are a few problems for these 2 values. First of all, some codes have multiple descriptions. Partially was due to misspelling. However, some of them are very different. Another issue we uncovered is some descriptions are missing.

Internal classification code and description:

These two columns are more reliable compared to offense classification code and description, as we can find out that these two columns are actually one to one correspondence between each other and there's no multiple correspondence issue that we need to look further into. The only data issue is the missing data. For data analysis purpose, we will majorly rely on this information for the crime type.

Lat-lon Issue:

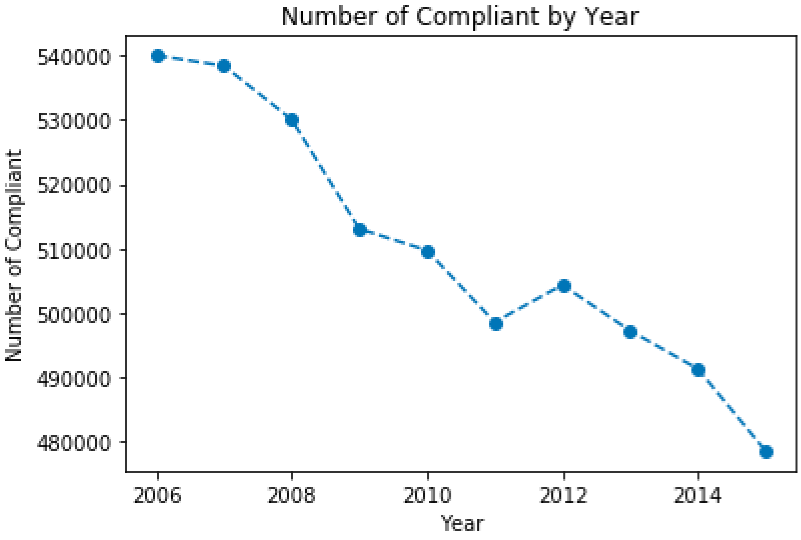
We run the code (col19_23_latlong_key.py) to generate number of crimes by each unique latitude-longitude pair. And then we mapped this latitude longitude to a street address by using geocoder package based on Google Geocoding API. The number one pops up on the list looks odd. It is on the 154 West 34th street. This is the crossing on the Macy’s and Penn Station. We think this datum maybe abnormal. We suspect that people may just enter 10001 (especially tourists) when they report the crime. And in the system, the lat-long is default to W34 Street for any unspecified location in Manhattan (since the zipcode there is 10001).

	lat	**long**	**freq**	**street_name**	**city**	**stree_number**	**zip_code**
26870	40.750431	-73.989282	17232	West 34th Street	New York	154	10001
99042	40.710094	-74.010610	5710	Cortlandt Street	New York	1-27	10007
83628	40.791152	-73.884372	4779	Hazen Street	None	15-0	11370

51470	40.804384	-73.937422	4731	Doctor Martin Luther King Junior Boulevard	New York	140	10035

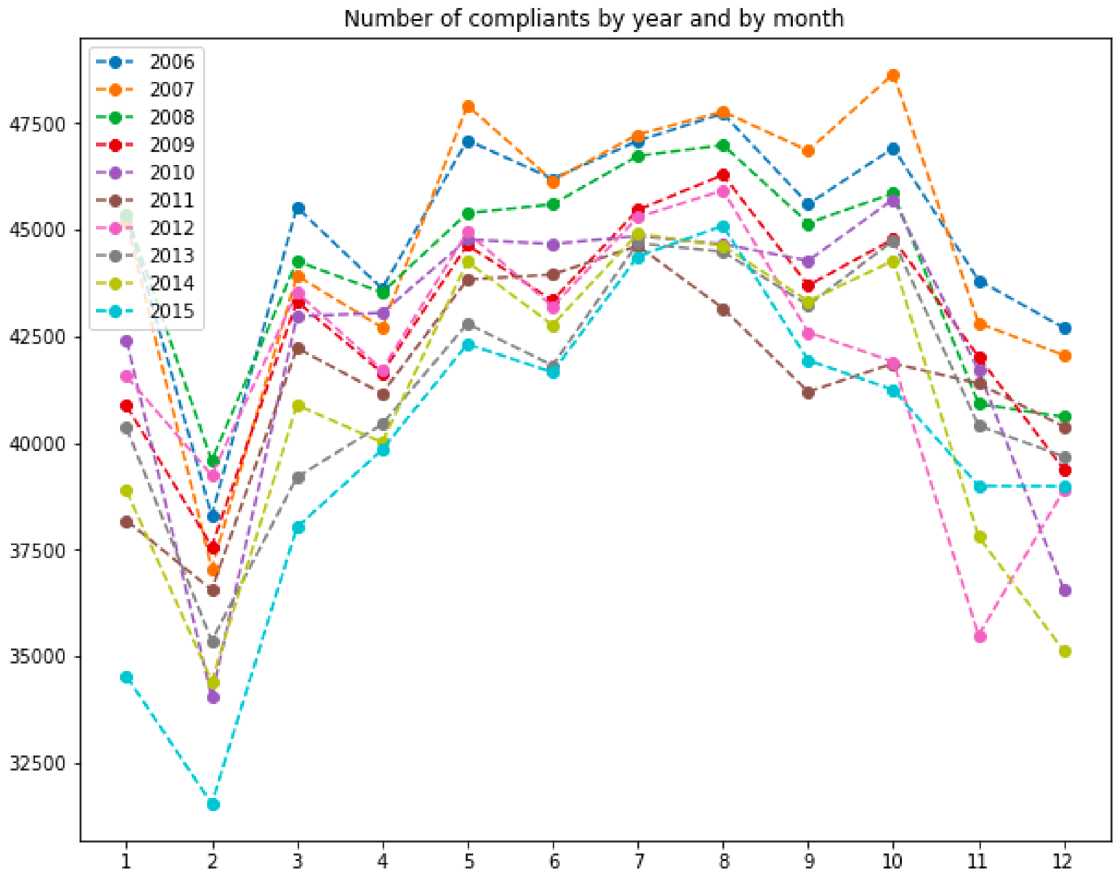
2.3 Data Trends

Number of crimes by year:
From column 5 we can examine the number of crimes by year. Overall, the crime decreased year over year. Our goal is to identify when, where, and possible reasons that the crime decreased. We will discuss this in Part II.

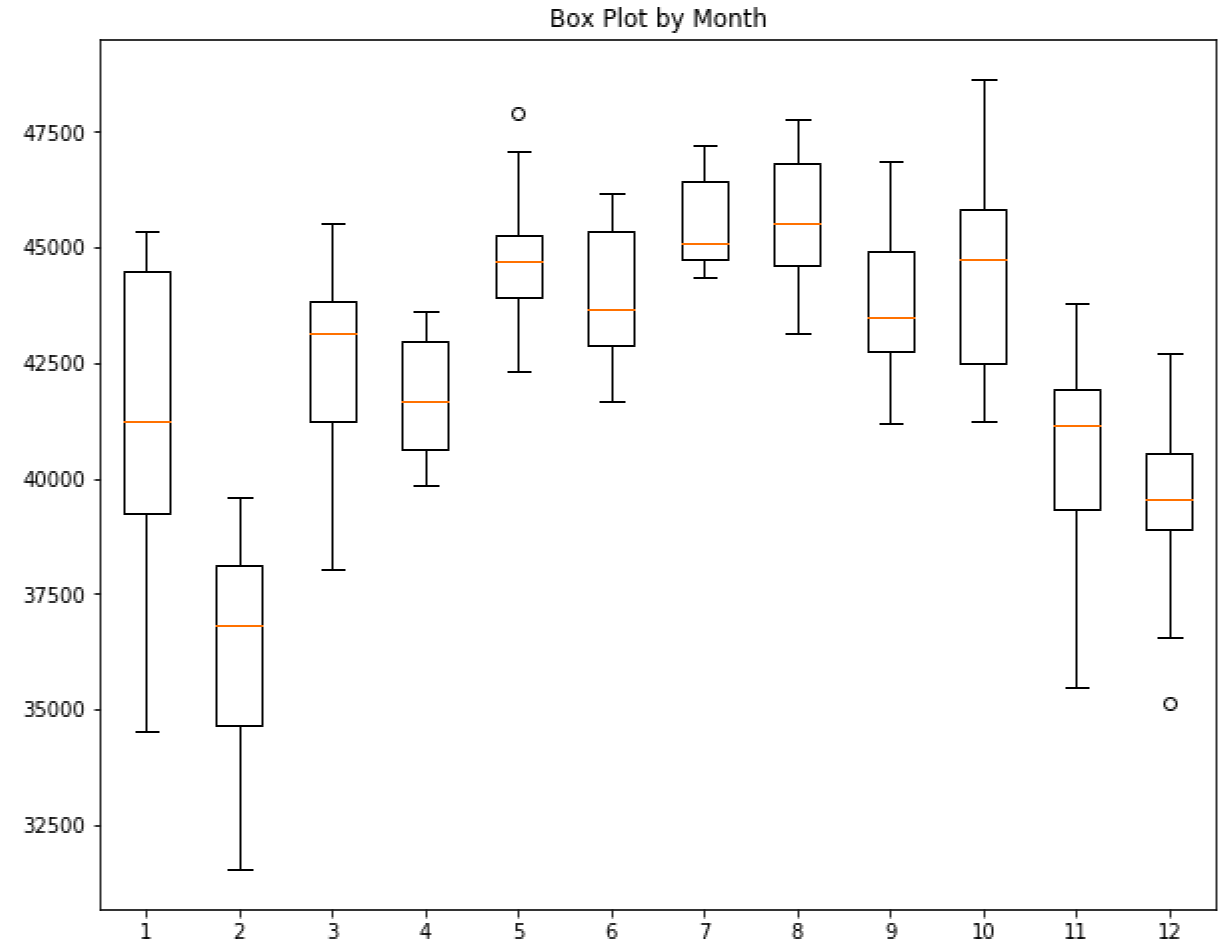


Number of crimes by year and by month:

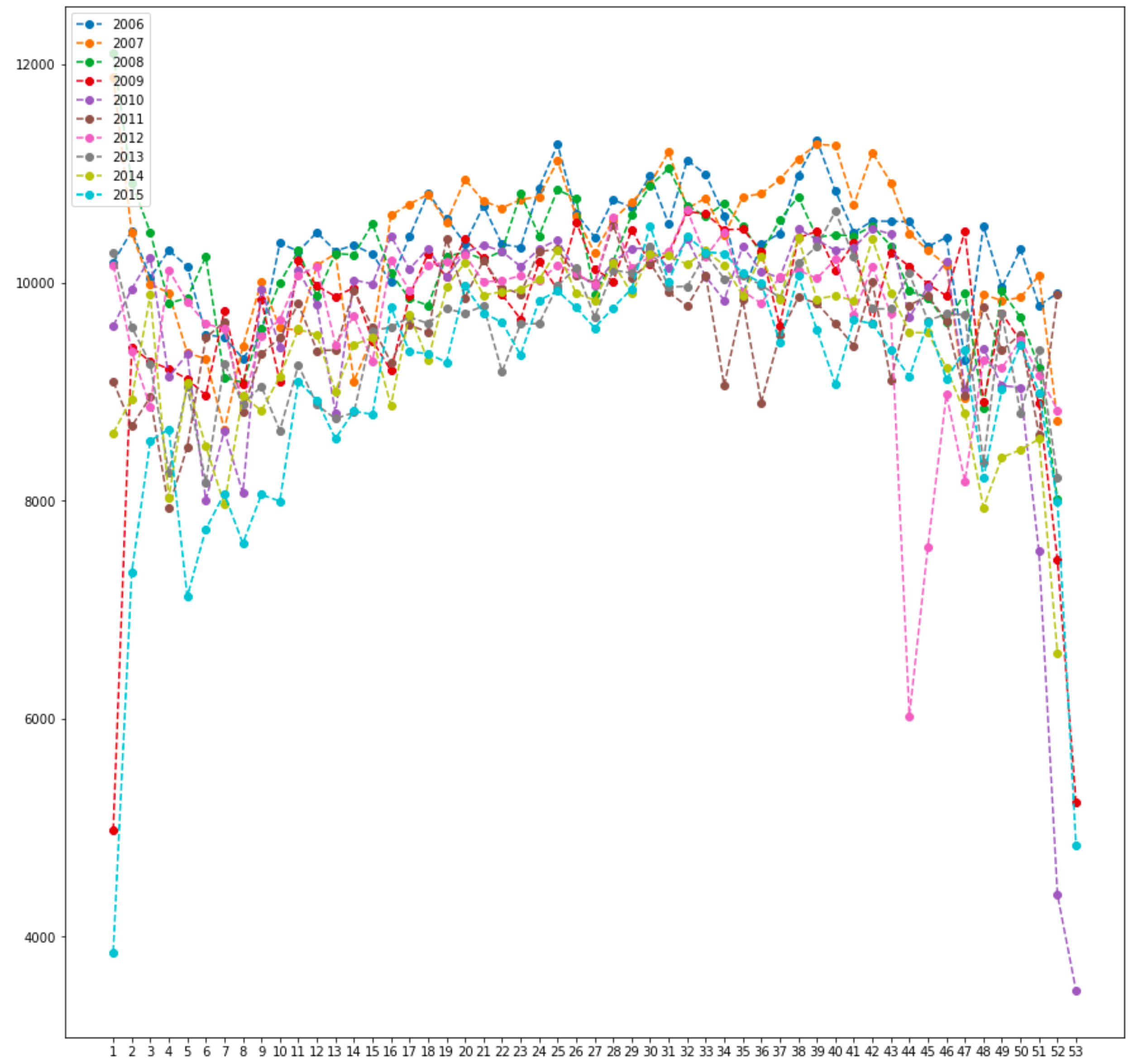
From column 5 we can also examine the number of crimes by each year and each month. It seems that Jan and Feb in 2015 has some abnormal number of crimes compared to the other month.



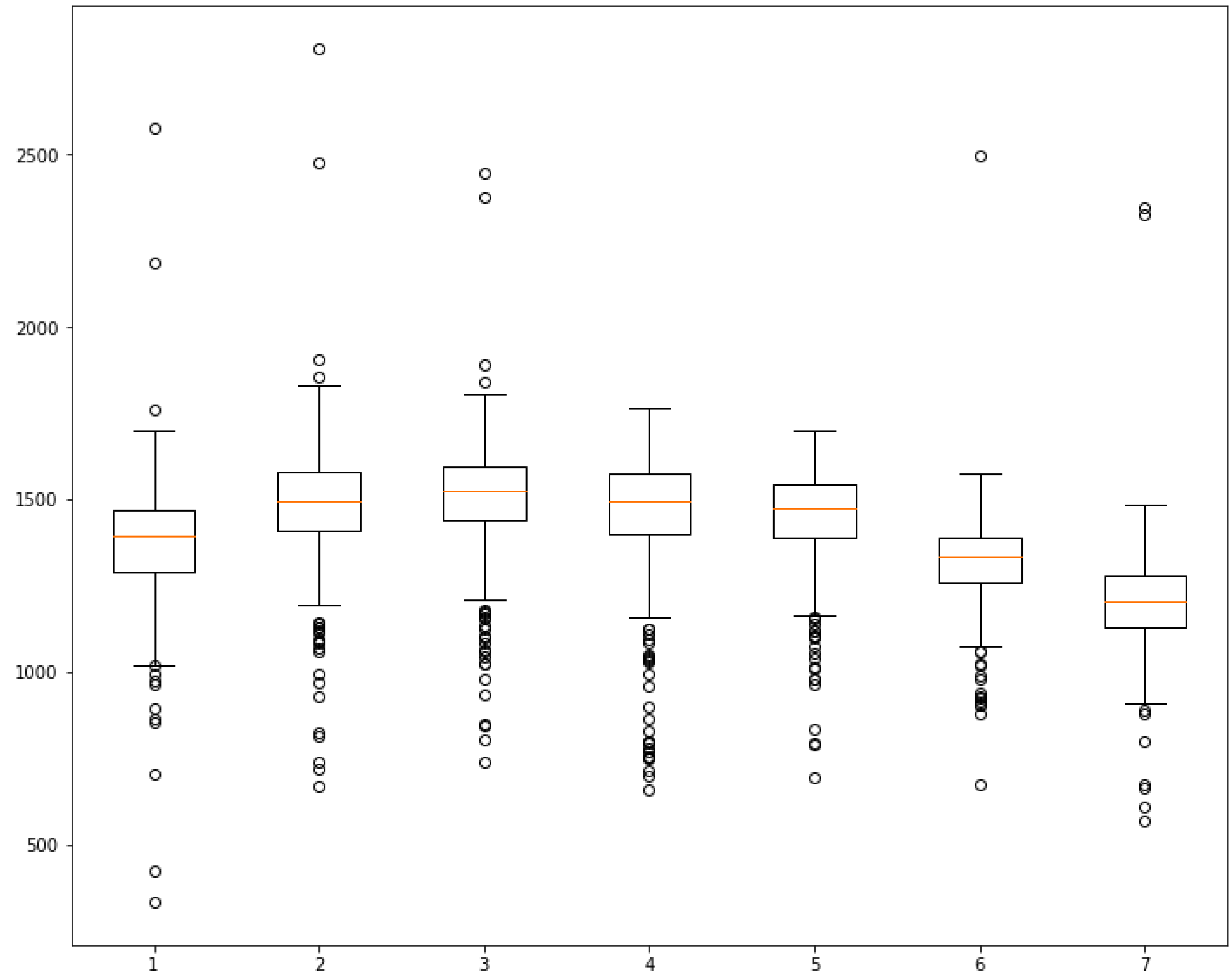
We also plot a boxplot and clearly February has the lowest number of crimes on average (partially is because of a shorter month). It seems there is an outlier in May and December.



Number of crimes by year and by week:
From column 5 we can also examine the number of crimes by year and by week. After plotting by week, it is clear that the week 44 in 2012 has an abnormal trend. It is probably because of hurricane sandy happened during that period.



Number of crimes by day of week:
From column 5 we also generated the number of crimes by year and weeknum in each day of the week. By looking at the boxplot, Tue-Fri is the peak and weekend has less complaints. Also, weekdays have more outliers.

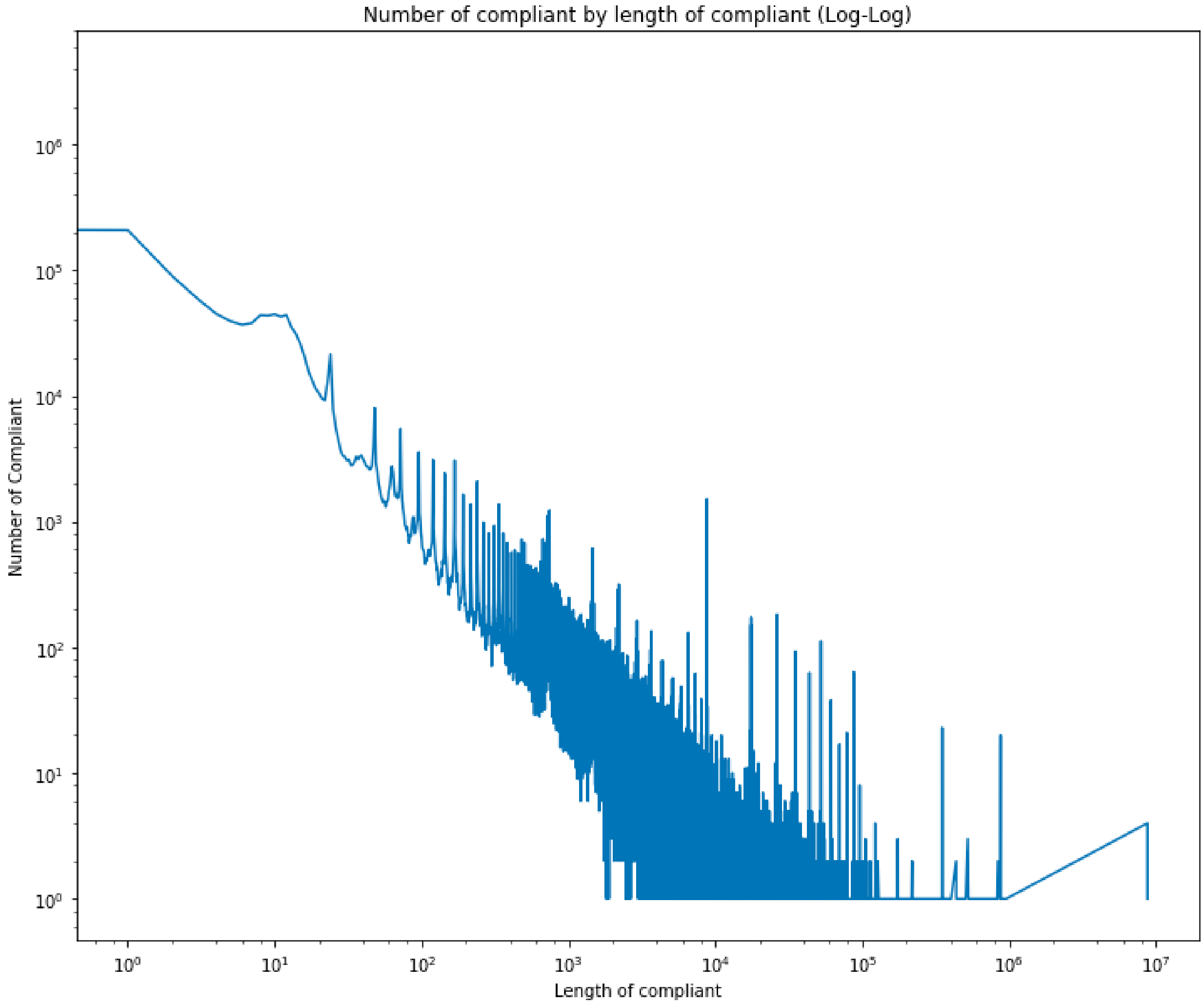


Number of complaints by hour:
From column 1 and 2 we checked number of crimes by hour. The number of crimes is gradually increasing after 7 am and then start decreasing after midnight. However, there are some very high abnormal outliers happened between 12 - 1 am. The outliers are usually between 12 and 1 am on the new year. It makes sense since the NYE celebrations.

	0	**1**	**2**	**3**	**4**
17430	2006	52	7	0	506
17454	2007	1	1	0	596
26214	2008	1	2	0	529
34926	2009	1	4	0	373
52374	2010	53	5	0	475
61134	2011	52	6	0	558

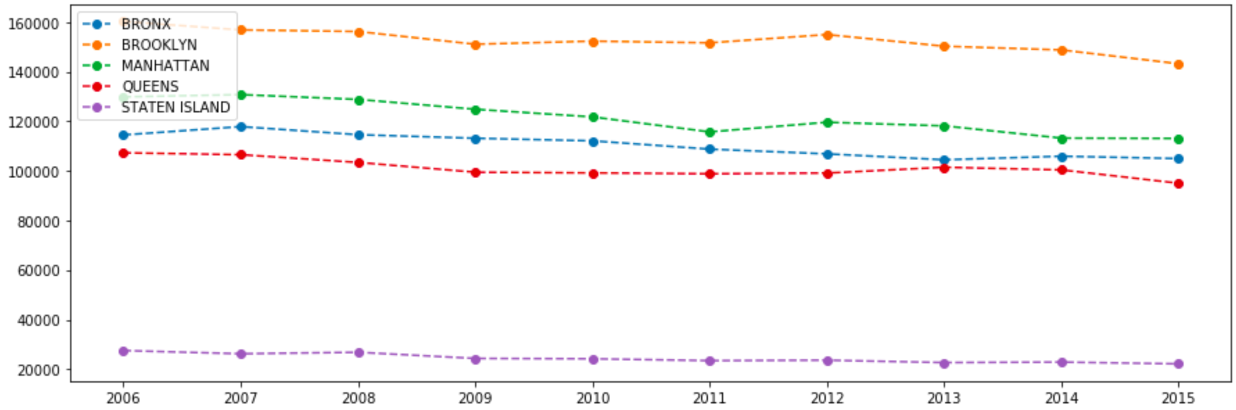
69894	2012	52	7	0	442
69942	2013	1	2	0	401
78702	2014	1	3	0	458
87390	2015	1	4	0	313

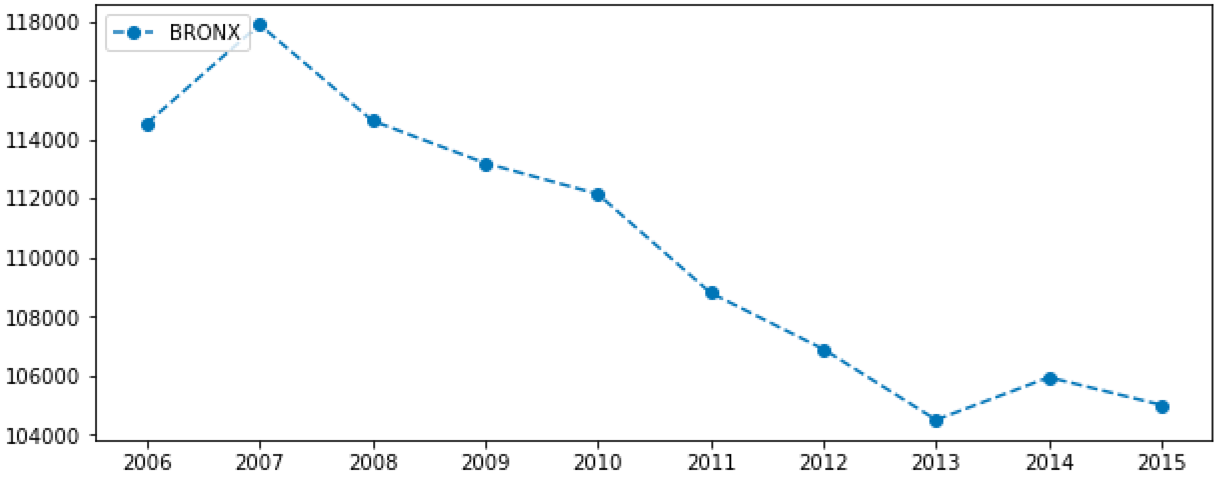
Number of crimes by length of complaint:
We examined the number of crimes by the length of complaint. When the crime has no to_date and to_time, the length will be counted as 0. Overall,number of crimes decays over the length of the compliant after transforming to log-log.

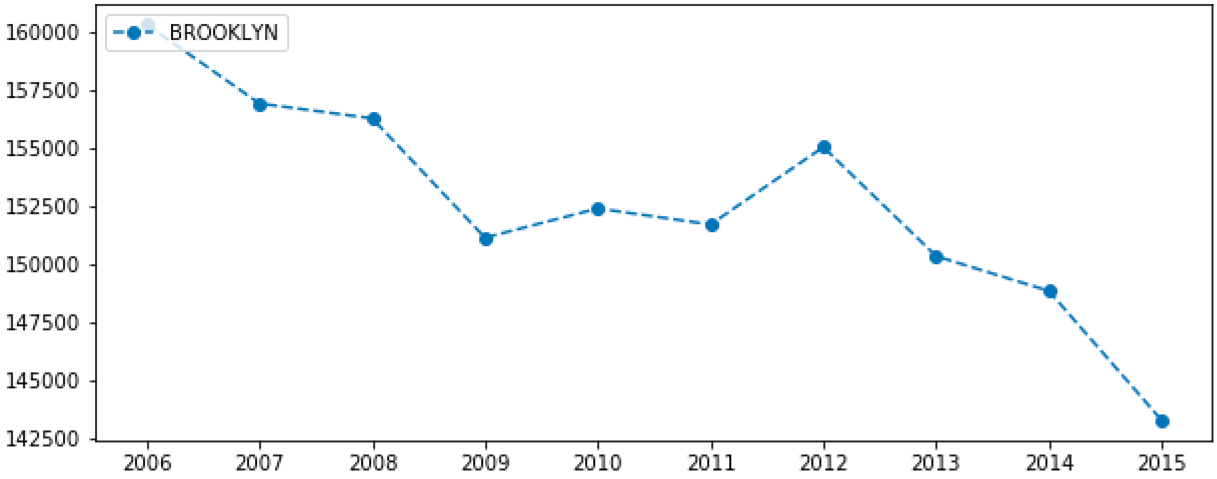


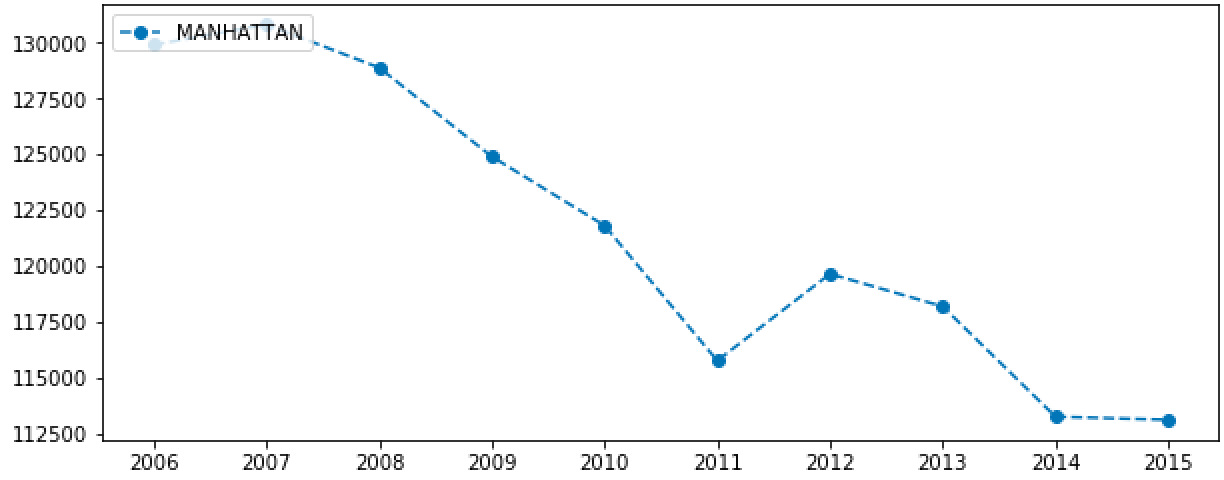
Number of complaints by year and by crime type:
From column 5 and 6 we generated the number of crimes by year and by KY_CD, which stands for the crime type. From the list, the crime 341(‘petit larceny’) is the most popular and did not decrease over the year. It seems that crime 235 (‘dangerous drugs’) and 361 (‘off. against pub ord sensibility &‘) may contribute the year over year decrease as the trend continued going down after 2012.

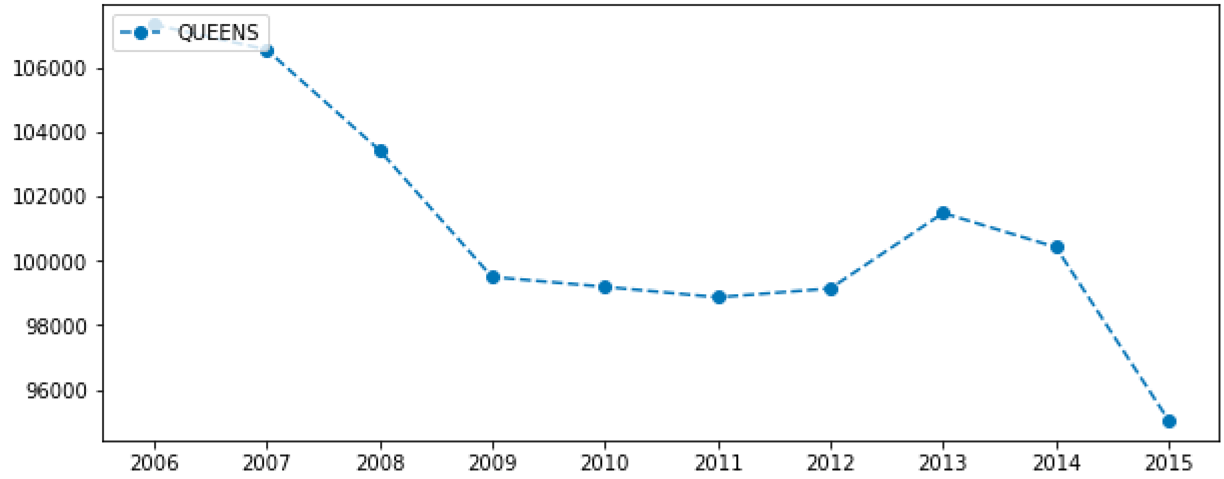
Number of complaints by year and by borough:
From column 5 and 13 we generated the number of crime by number of complaints, year, and borough.

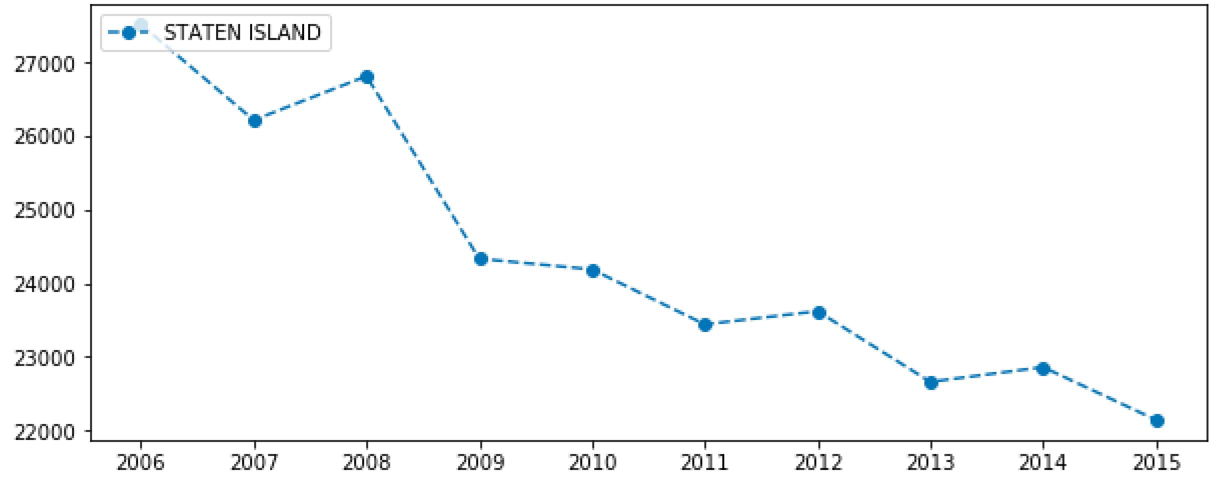








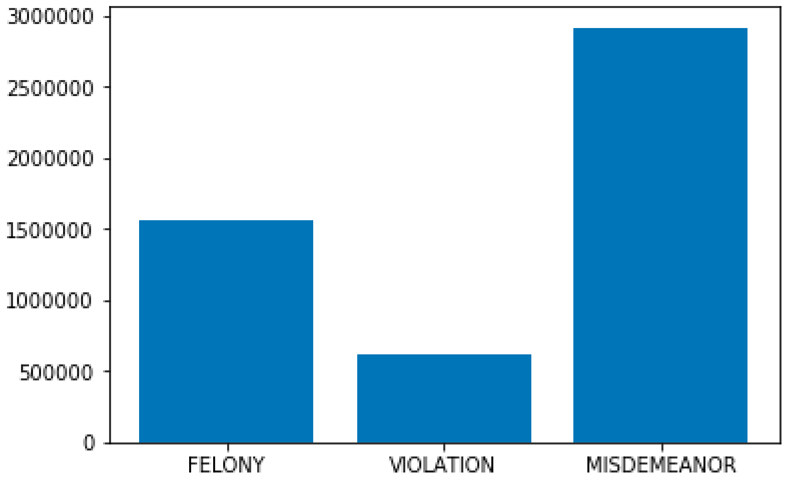




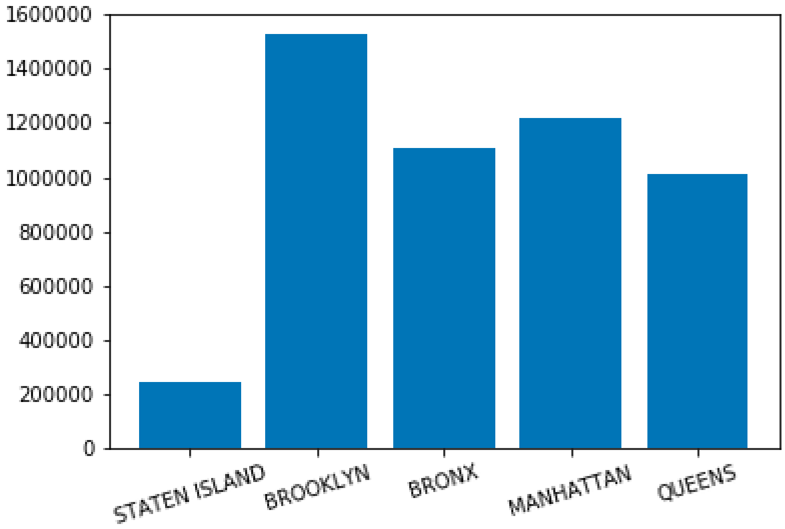
Looking at the stacked the line chart, the decreasing trend is not very clear. After plotting the trend borough by borough, it is more clear that Brooklyn contributes the most to the decrease, followed by manhattan.

Some data distributions:

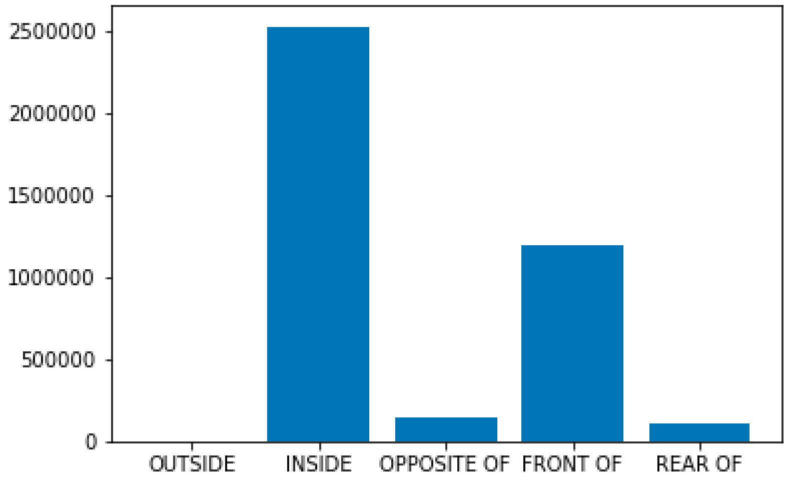
Offense Level:



Borough Names:

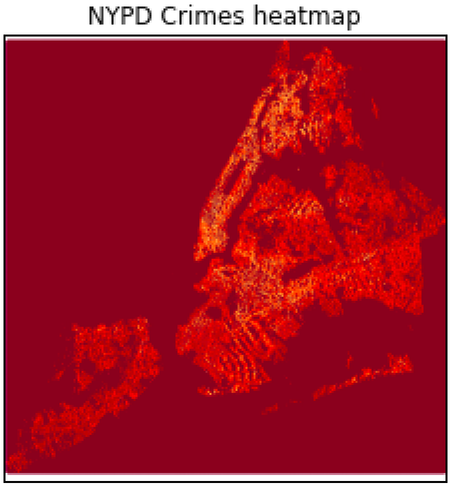


Occurrence Location Descriptions:

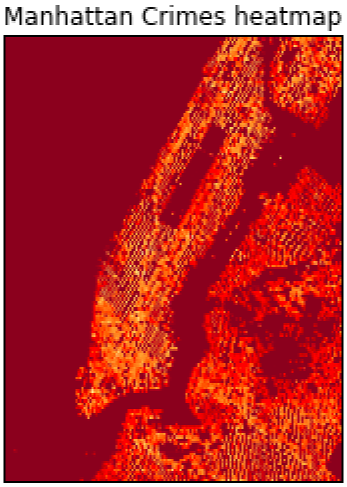


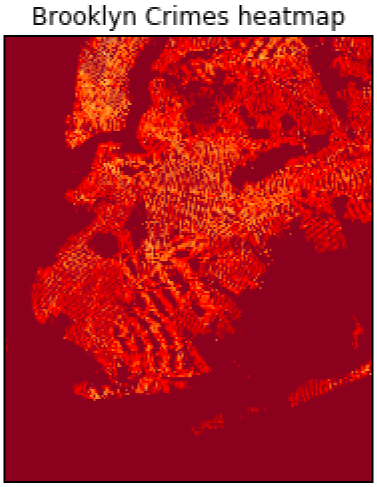
Heatmap of Crimes:
We also plot a heatmap of crimes through history. From the heatmap we can find out whether there is any crime in an invalid location, for example, on the river. It seems that our data are all within the region they supposed to be.

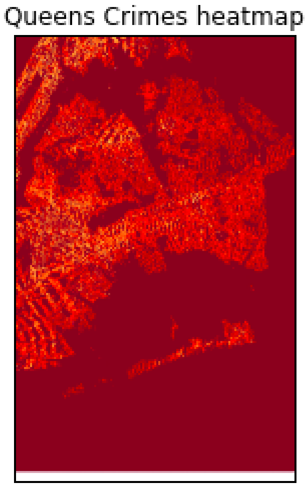
Map of NYC:

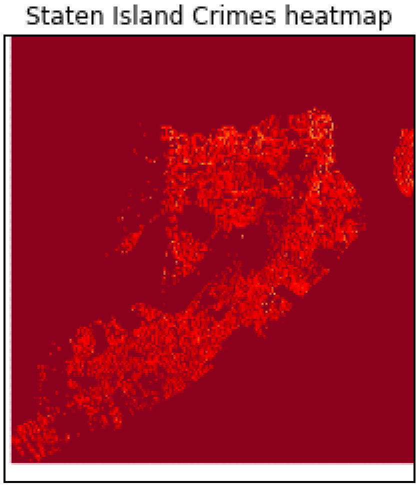


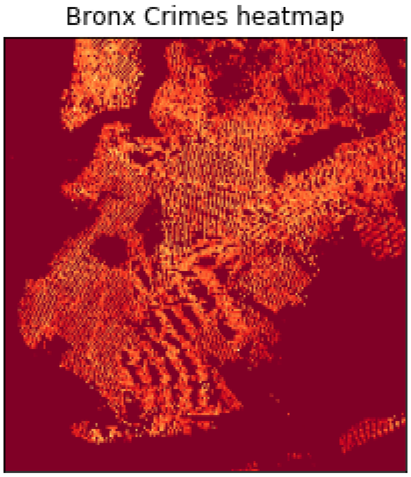
Map of 5 Boroughs:











3 Part II: Data Exploration

.

3.1 Experimental techniques and methods

We used the PySpark for aggregating data from the original dataset and process the data into a reasonable size output. And we use python for re-process the aggregated data and visualize the trend, and validated the statistical significance.

1) PySpark: we ran 4 different experiments in this phase. Each is if the number of crime per month is correlated to the unemployment rate, with an assumption that when the unemployment rate is low, people are more engaged in a regular job and have less chance to theft or steal.
2) if the number of crime per zip code is correlated to to it is house value with the hypnosis that a higher median house value area should have more resources to enforce the security and reduce the crime rate.

3) If the number of crime per zip code is correlated to the motor collision, with the assumption that a place is less safety when people don’t drive well.
4) If the number of crimes per week is correlated to the motor collision, with the assumption that it is a good indicator of a few other factors like weather, average education level in that area, etc.

Methods we used:

1. We ran a pyspark code col19_23_latlong_key.py to aggregate all the unique latitude longitude pair. In this case, there were 120K unique pairs in the data.
2. Then, in order to correctly map the latitude longitude to a zip code, we used a few packages in the ipython notebook such as json and shapely to map this in a csv that maps each lat-long to a zip code. We also used python packages like geopandas, seaborn and mpl_toolkits.basemap for the visualization part of this project.
3. We uploaded the mapping list (zipcode.csv) back to hadoop DFS. In the col5_compared_to_zipcode.py code, we used the Left Outer Join to read the data from the original csv file and map a zip code to the data. And then, we aggregate the number of crimes by time (Year, Month, Weeknum, and Weekday) for further analysis in the ipython notebook.

3.2 List of hypotheses

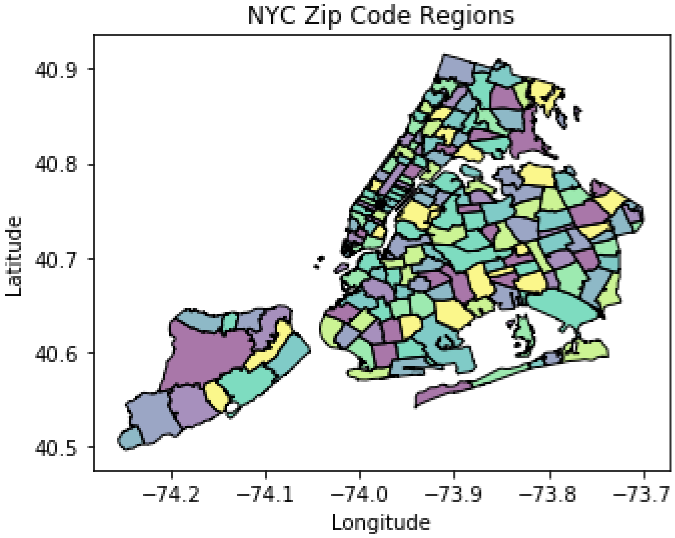
From Part I we have some findings that we want to examine deeper in Part II.

1. Will different locations are inherited with different danger or risk coefficients? How correlated it is? In our case we choose the number of vehicles collision as the danger.
2. Another hypothesis we have to explain the difference of crime frequency among different regions was that maybe some regions with more rich people, they would probably afford to maintain a safer community.
3. The crime frequency is also different through time. It is probably related with the trend in the vehicle collision frequency.
4. The trend through time in crime frequency is probably related with the trend in unemployment rate.

3.2.1. Crime frequency and collision frequency within same regions:

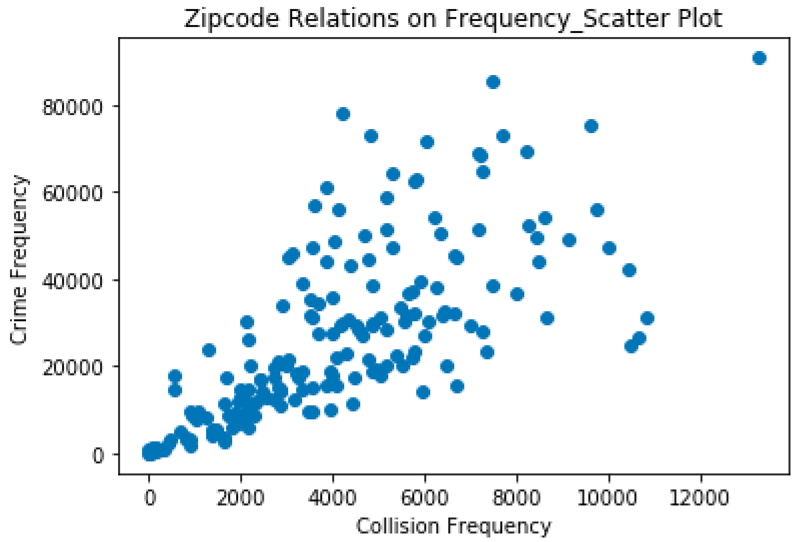
First we found out that the frequency of crimes are different across different locations such as different boroughs or other more specific parts of New York City. Our hypothesis is that maybe different locations are inherited with different danger or risk coefficients (probably due to some historical reasons) so that some regions are more dangerous than other parts of the city. We also found out in some papers that possibly due to the same reason, regions with high frequency of traffic accidents tend to also have high frequency of crimes. And even the time that the traffic accidents happen tend to also coincide with the occurrence of crimes. Therefore we want to compare our crime data with some traffic accident data to figure out whether there is a high correlation between their frequencies of occurrence within some neighborhoods.

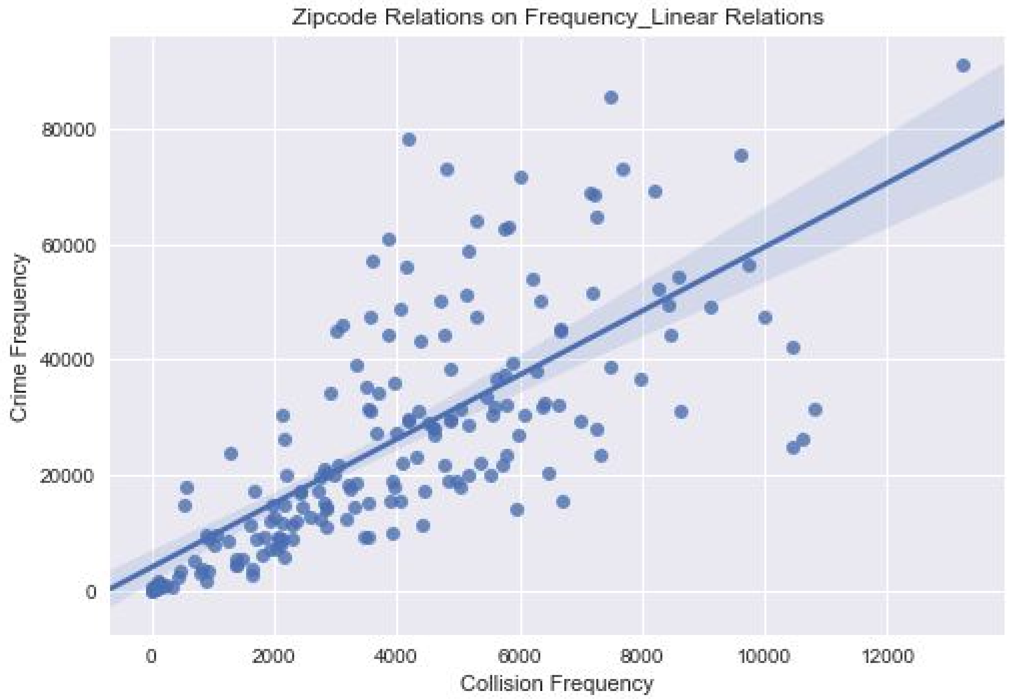
We found ‘NYPD_Motor_Vehicle_Collisions.csv’ from https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95 as our data set for measuring frequencies of traffic accidents in different regions. For the experiment setup we used zip codes as a measure of defining neighborhoods, since it is easier to obtain and understand and there already are zip codes in our collisions data set.



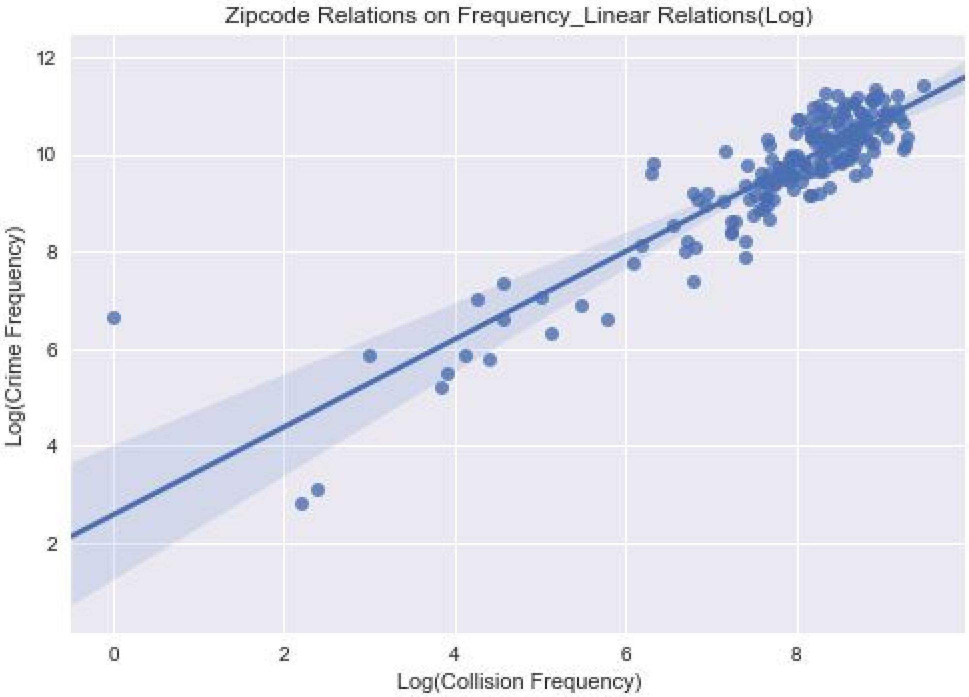
However in our original crime data set there’s only a column of (Latitude, Longitude) but no zip codes information. To encounter this issue we used the Shapefile of NYC Zip Codes - tabulation areas provided by NYC Department of Information Technology & Telecommunications (DOITT) from http://nycdata.pediacities.com/dataset?tags=neighborhoods, which is stored in a geojson file. We used the (Latitude, Longitude) information for each zip code in this file as a boundary measure for our data. Within our original crime data we first generate a list of unique (Latitude, Longitude) pair and then for each pair we define its zip code if it is within one of the boundaries in the geojson file. In this way we are able to uncover one by one each pair’s corresponding zip code. We then merge this unique list back into our original data to get a full data set with a new column for zip codes.

In order to discover whether there really exists a relationship between the frequency of vehicle collisions and the frequency of crimes in the same location, we extracted the unique zip codes and their counts from each of the two data sets. Then we combine these two dataframes into one and compare their frequencies according to zip codes. We calculated their Pearson correlation and obtained a correlation coefficient of 0.7275. We also examined their relations from graphs.

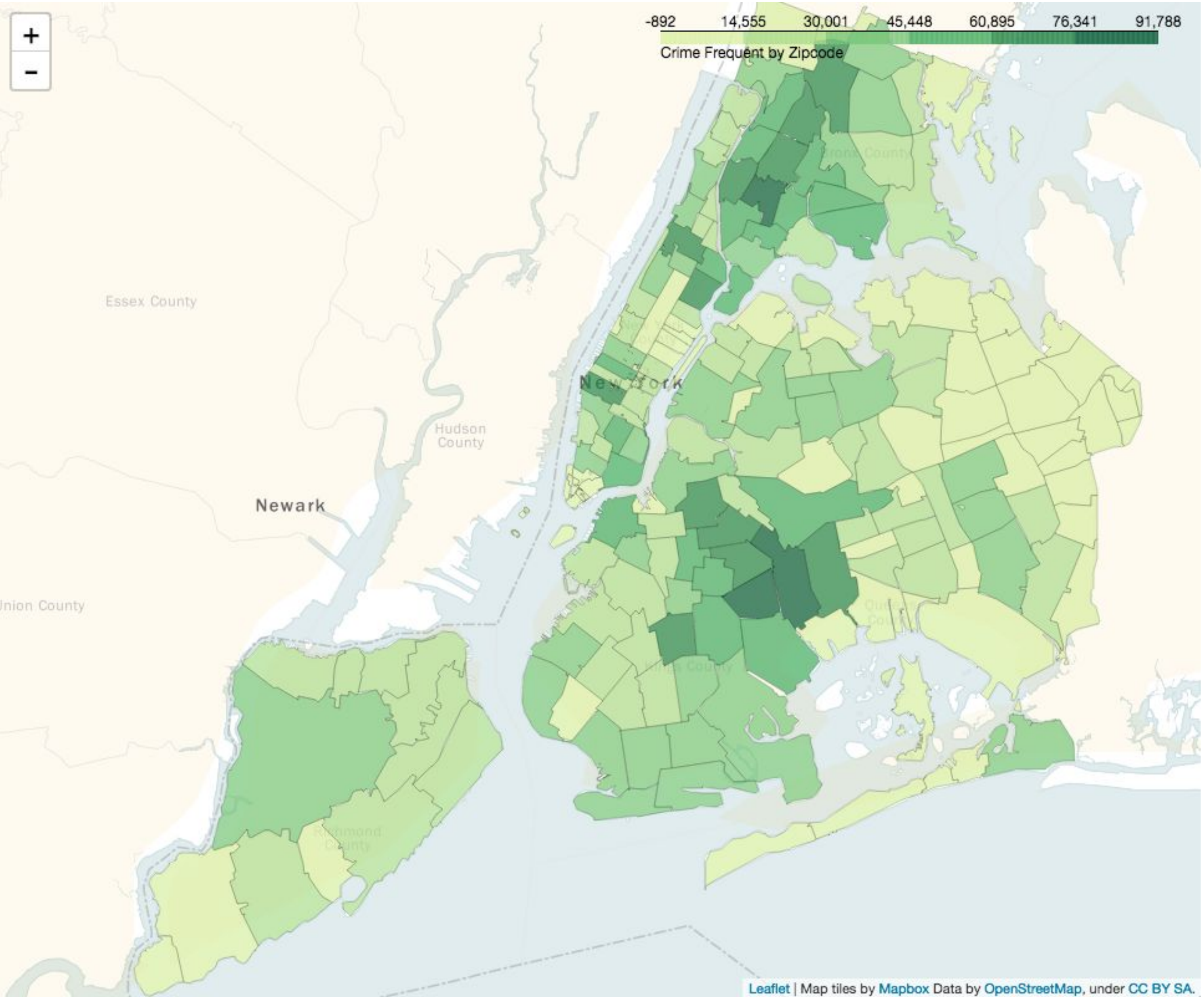




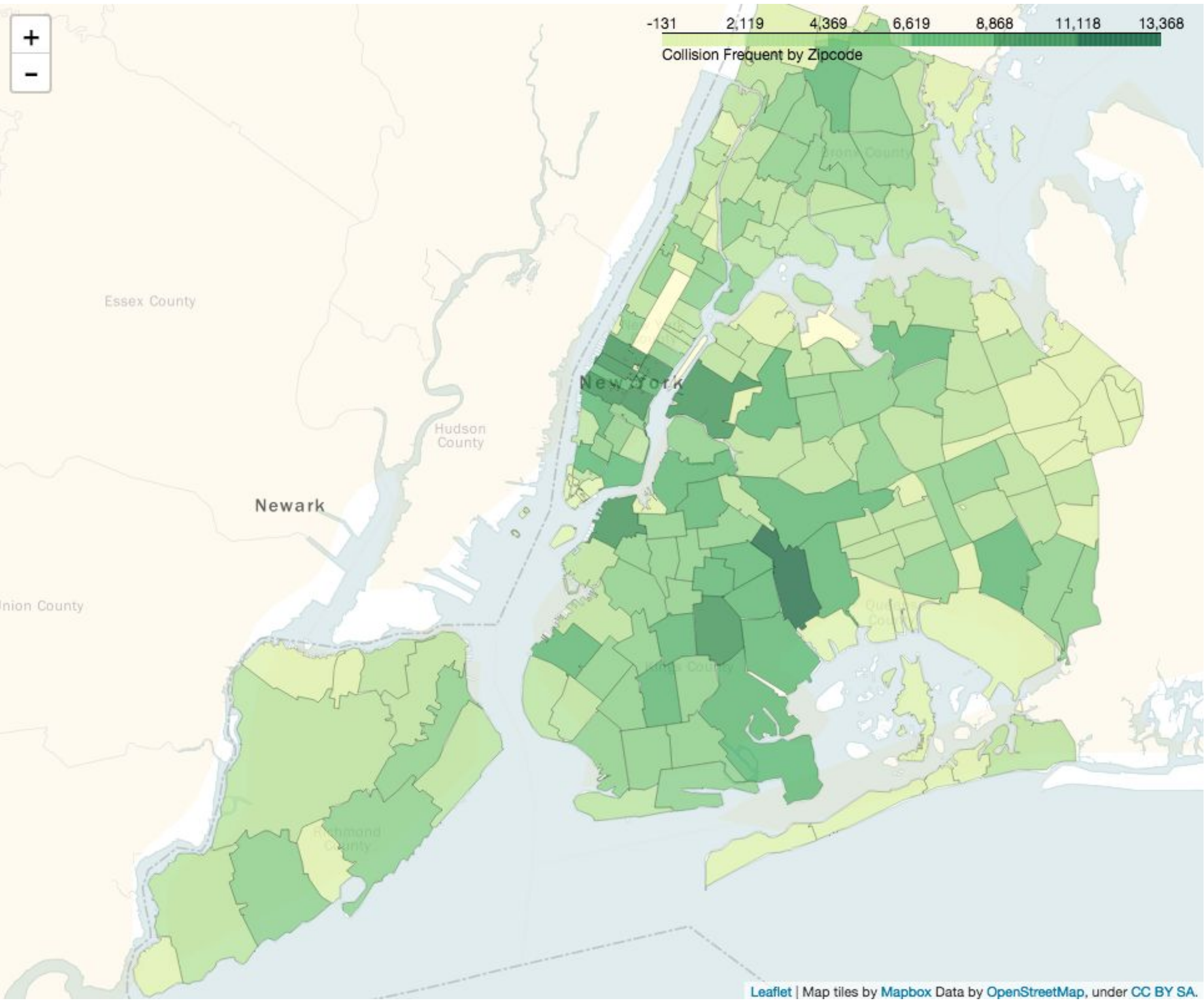
We also take the Log of both frequencies and plot their linear relationship. It seems that it is even clearer about their linear relations.



We plot the map indicating frequencies in different zip code regions for both the crime data and the vehicle collision data.



Crime frequency map



Collision frequency map

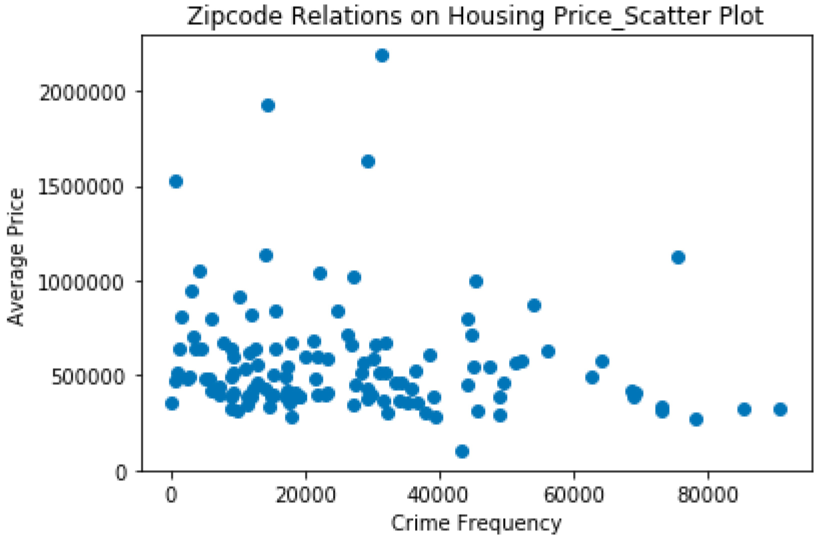
3.2.2. Crime frequency and housing price within same locations:

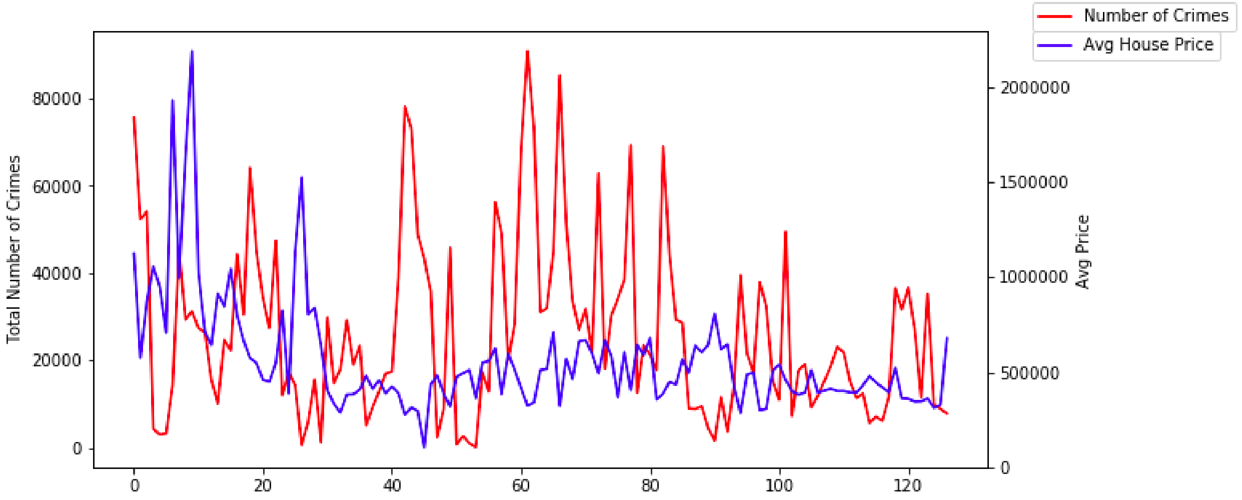
Another hypothesis we have to explain the difference of crime frequency among different regions was that maybe some regions with more rich people, they would probably afford to maintain a safer community. Therefore these communities tend to become places with higher security and crime rates in these locations would possibly be lower than other locations. Our attempt to examine this kind of relationship was to compare crime frequency in different locations with the housing price in the same locations, as we think that places with higher housing prices have possibly more rich people living in.

We obtained housing prices data set from http://files.zillowstatic.com/research/public/Zip/Zip_Zhvi_AllHomes.csv. It also use zip codes to represent different locations. So we then used the above zip code information we generated for our crime data for our comparison work.

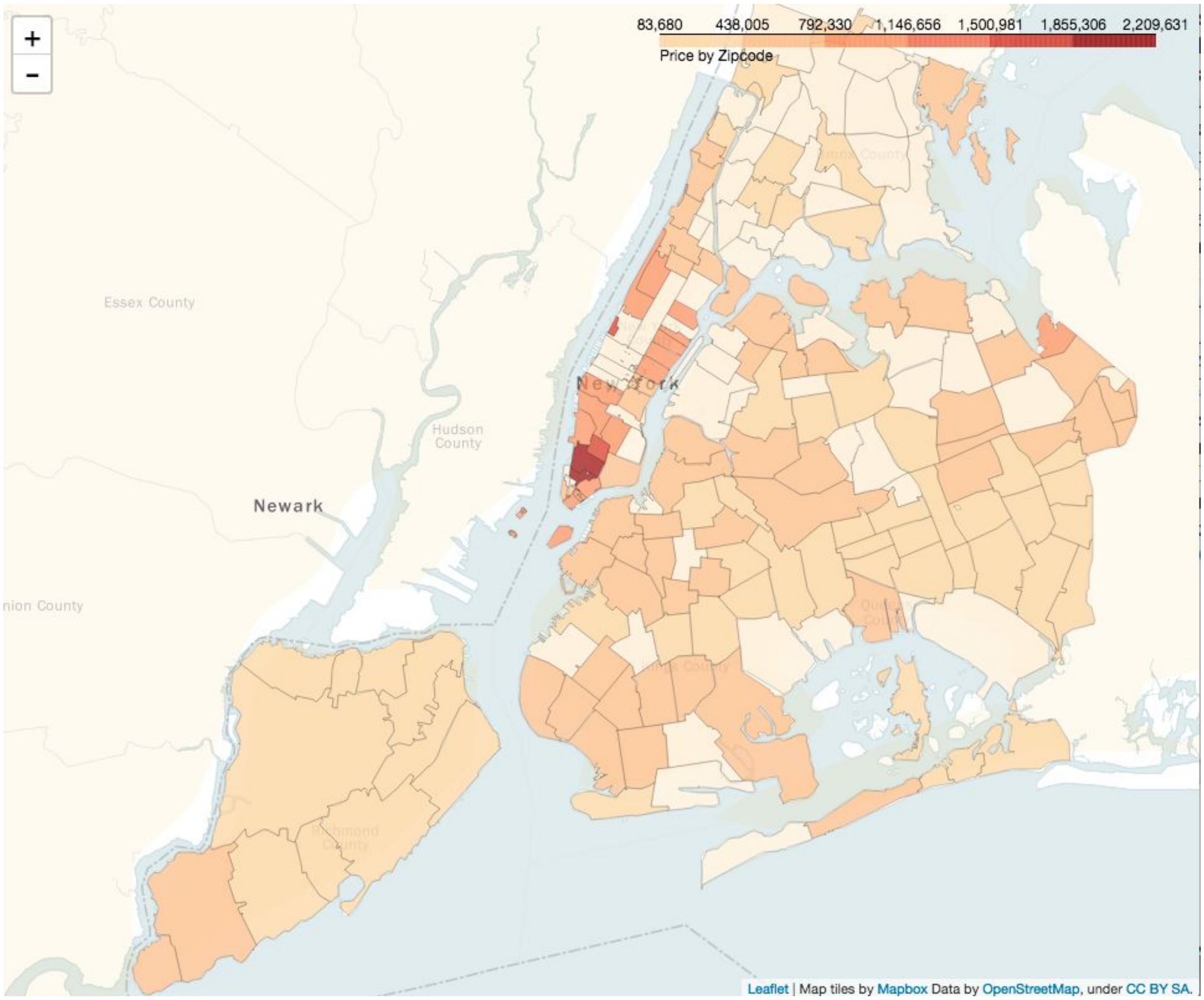
In order to discover whether there really exists a relationship between the frequency of crimes and the housing prices in same locations, we extracted the unique zip codes and their counts from the crime data. The housing price data is already grouped by the zip code locations and each row is the median price within that region. Then we combined these two extracted data together and compared their frequencies according to zip codes. We calculated their Pearson correlation and obtained a correlation coefficient of -0.1357. It seems that the relation between them is not very significant.

We also examined their relationship with graphs.





We can see from the above figures that the relationship between crime frequency and housing prices based on their locations is not significant.



The housing price map also indicates that the correlation between crime frequency and housing price is low.

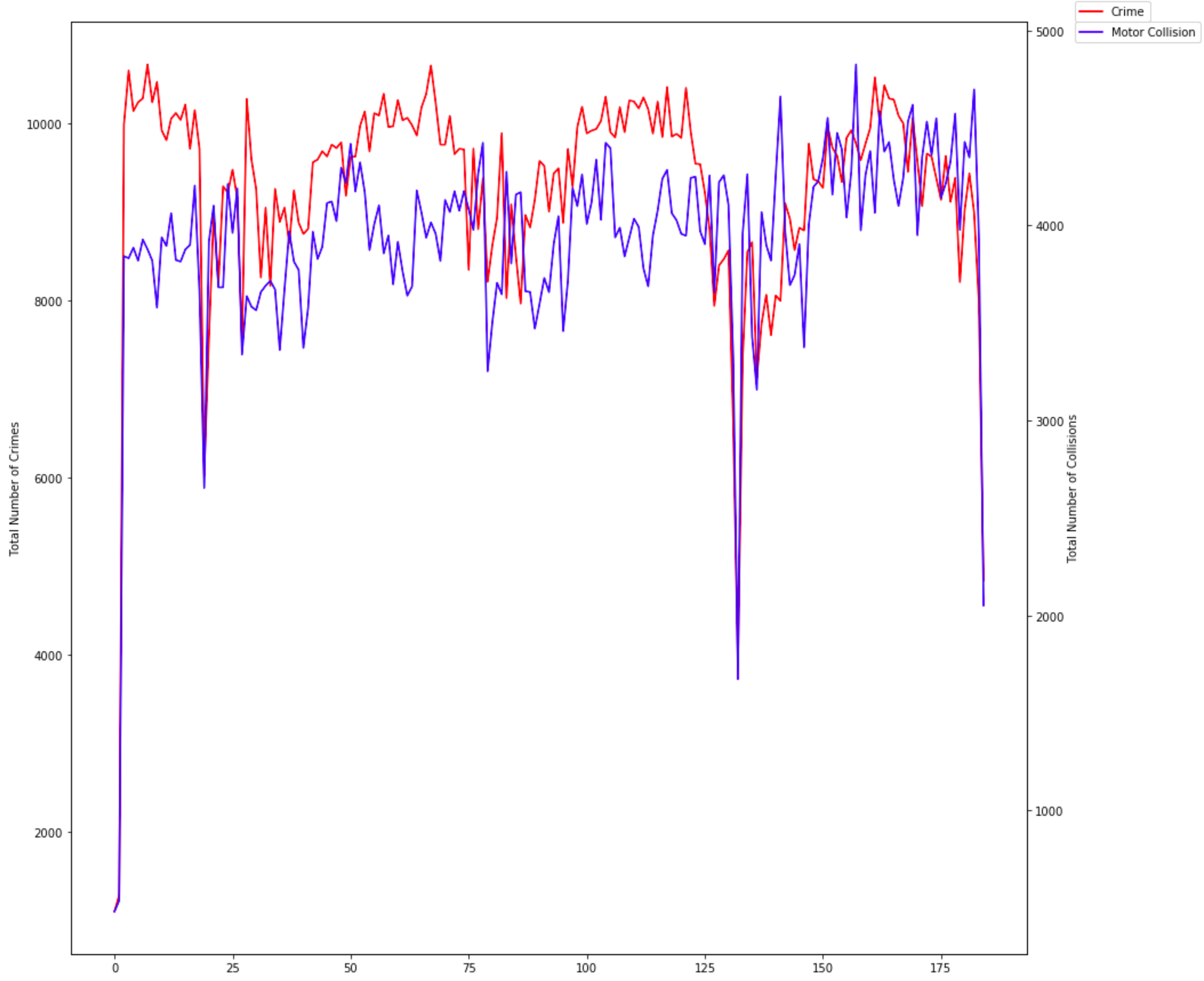
However, since the correlation coefficient is negative, it seems that there may indeed exist a negative relationship between them. And since zip codes may not be a good representation for different communities, maybe it is worth trying to using other methods to separate the locations and examine deeper into the relationship between crime frequency and housing price. Their relationship may become more significant if we can obtain more detailed location information and analyse their relationship.

3.2.3. Crime frequency and collision frequency in the same time:

From Part I we found out that the crime frequency is also different through time. There maybe some potential reason for this trend. One of our hypothesis for this phenomenon is that it is probably related with the trend in the vehicle collision frequency. Therefore we want to compare these two trends and see if there indeed exists a significant correlation between them.

We used the week number (i.e. the index for the week) to examine the trend. We grouped each data by their week number and count their occurrence to extract their trend information. Then we combined these two extracted data together and compared their frequencies in each week number. We calculated their Pearson correlation and obtained a correlation coefficient of 0.7438 . It seems that the relation between them is fairly significant.

We also examined their relationship with graphs.



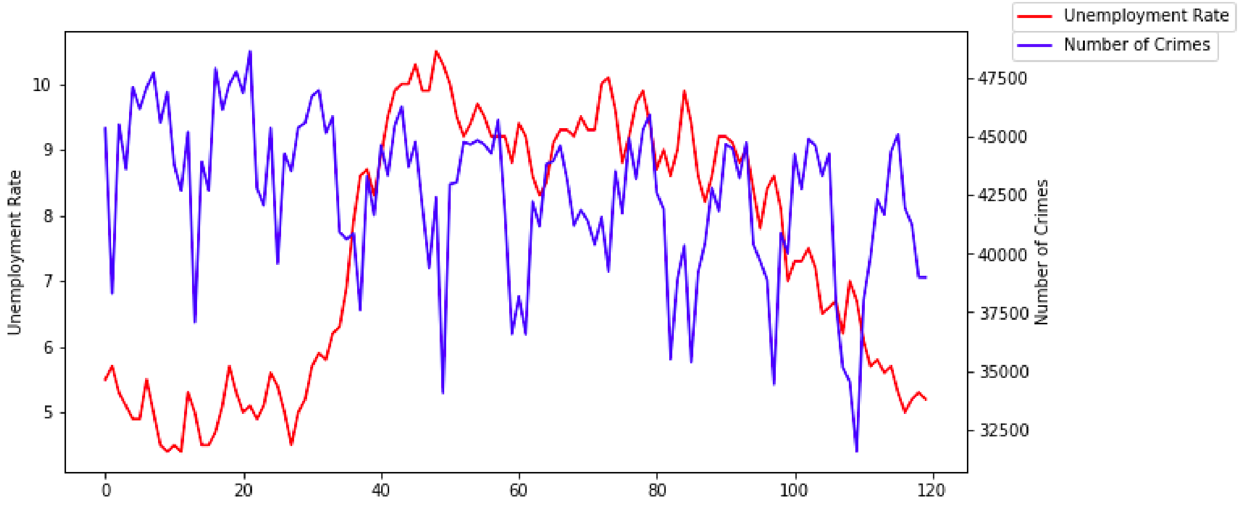
From the figure above we can also see that the correlation in time between crime frequency and vehicle collision frequency is indeed high. They tend to drop simultaneously in some weeks and increase together in other weeks. And it seems that sometimes the vehicle collision series tend to increase or decrease first in time and then the crime series tend to follow up the trend. We think it is an interesting discovery which we could look further into in the future.

3.2.4. Crime frequency and unemployment rate through time

Another hypothesis for the trend through time in crime frequency is that it is probably related with the trend in unemployment rate. Because it is possible that when the unemployment rate is low, people are more engaged in their regular jobs and have less chance to theft or steal, and thus the security quality would likely to improve.

To examine this hypothesis we found the unemployment rate data from https://www.labor.ny.gov/stats/laus.asp. As the unemployment rate data is distributed according to each month, we grouped our original crime data by month and count the frequency. We combined these two and compared them within each month. We calculated their Pearson correlation and obtained a correlation coefficient of -0.2330. It seems that the relation between them is not significant.

We also examined their relationship with graphs.



It can also be seen from the figure above that the two trends seem not have a very strong correlation. However the second half part of the figure seems to represent a relationship between the two trends that when the unemployment rate drops significantly, the crime frequency tend to also go down to a valley. Possible explanation would be that because the seasonal trend of crime frequency is too influential, it would be more reasonable to first get rid of the periodicity in our data and then do the time series analysis. Or we could group both data by year so that possibly the relationship between them would be more significant.

3.2.5 In summary, we can list the result of these hypotheses and use Pearson correlation coefficient as the associated score:
Correlated attributes:
(Cirme_frequency_by_zipcode, Collision_frequency_by_zipcode, 0.7275)
(Cirme_frequency_by_time, Collision_frequency_by_time, 0.7438)
Uncorrelated attributes:
(Cirme_frequency_by_zipcode, Housing_price_by_zipcode, -0.1358)

(Cirme_frequency_by_mounth, unemployment_rate_by_mounth, -0.2330)

4 Individual Contributions

Part I
Col 0 - Col 5: Wenjie Sun
Col 6 - Col 18: Yaohan Ke
Col 19 - Col 23: Xinyan Yang

Part II
vehicle collision: Yaohan Ke, Xinyan Yang, Wenjie Sun
Unemployment Rate: Wenjie Sun, Xinyan Yang
Median House Value: Wenjie Sun, Xinyan Yang, Yaohan Ke

5 Conclusions

This project provides a good summary of the New York City’s crime data trend. It is obvious that year over year, there are less and less crimes happened in the city. With the limit of available data, we were able to prove the correlation between motor collision and crime. Even though there is no proven causal relationship, we think to predict and prevent the crime in the future, NYPD should refer to traffic condition data. Also, the crime rate is very weakly correlated to the house price: the higher the median price is, the less crimes happened in the area. We think this can be better studied if we can obtain data on the neighborhood level. The crime frequency also seems to very weakly correlated with the unemployment rate, and the possible explanation is that because the periodicity in the crime data is too strong it would make more sense to de-cycle this trend first before exploring the relationship, or to group data by year rather by month for analysis.

6 References

[1] Burch, J., and M. Geraci. "Data-driven approaches to crime and traffic safety." *The Police Chief* 76.8 (2009): 18-23.
[2] Chirigati, Fernando, et al. "Data polygamy: The many-many relationships among urban spatio-temporal data sets." *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2016.