DSGA1004 Big Data Project

Team name: data miner

Team members:
Raochuan Fan (rf1711)
Haichao Wu (hw1551)
Yiran Xu (yx1350)

Abstract:

This project is composed of two parts. First part is data summary and data quality issues. Second part is data exploration. For the first part, we chose Crime dataset as our target. Through careful investigation, we found this is a pretty clean dataset with only 8.17% missing values and very few invalid values. In terms of data summary, we found “misdemeanor” has highest number of records whereas “violation” has lowest number of records. We also found some places are particularly vulnerable to crimes, such as central park and some region in Brooklyn whereas Queens and STATEN ISLAND are relatively safe.

Introduction:

For part I, we focused on data quality and exploration. The problems we are addressing for part one are: 1. checking each value in each column whether it is missing/valid/invalid. 2. checking Data quality of the whole dataset. 3. generating a summary of data. We addressed data quality problems based on principles covered in the lecture “data cleaning”.

For part II, we made several hypothesis between crime number and other features, like time, location, economics and weather, etc. Moreover, we made hypothesis about the different kinds of crime and analyzed correlation between them.

In part I, we use pyspark to deal with invalid or missing data and check other data quality issues. In part II, we created a pyspark.sql dataframe, then mainly used GroupBy method to get summaries of the whole data set. Then we did the analysis and visualization using python and Tableau.

PART I:

1. **Dataset: Crime**
2. **Data Cleaning**

   We notice that there are some typos for VALID records, like
   'OTHER STATE LAWS (NON PENAL LAW)' and
   'OTHER STATE LAWS (NON PENAL LA'.

   We will modify those typos to achieve consistency within VALID values before checking NULL and INVALID values.

3. Check VALID, INVALID, and NULL values for each column

column	C_1 (CMPLNT_NUM)	C_2 (CMPLNT_FR_DT)	C_3 (CMPLNT_FR_TM)	C_4 (CMPLNT_FR_TM)	C_5 (CMPLNT_TO_TM)	C_6 (RPT_DT)	C_7 (RPT_DT)	C_8 (OFNS_DESC)
Base Type	INT	datetime	datetime	datetime	datetime	datetime	INT	TEXT
Smntc Type	persistent ID	date of occurence	Time of occurrence	ending date of occurence	ending time of occurence	reporting date	offense classification code	description of offense
Num valid	5101231	5100569	5101183	3709752	3713446	5101231	5101231	5082391
Num invalid	0	7	0	1	0	0	0	0
Num null	0	655	48	1391478	1387785	0	0	18840
Different kinds of values in the same column	No. All of them are valid.	No. All of them are valid.	No. All of them are valid.	No. All of them are valid.	No. All of them are valid.	No. All of them are valid.	No. All of them are valid.	No. All of them are valid.
Different semantic types in the same column	No.	No.	No.	No.	No.	No.	No.	No.
Valid checking criteria	Integer of 9 digits	Betwee12/31/1850 and 01/01/2017	Of this format:HH:MM:SS	Betwee12/31/1850 and 01/01/2017	Of this format:HH:MM:SS	Betwee12/31/2005 and 01/01/2017	Integer of 3 digits	Non-empty string
Outliers/Interesting/Suspicious values/	None.	Some year is 1015 such as12/04/1015	None.	None.	None.	None.	None.	None.

column	C_9	C_10	C_11	C_12	C_13	C_14	C_15	c_16

	(PD_CD)	(PD_DESC)	(CRM_ATPT_CPTD_CD)	(CRM_ATPT_CPTD_CD)	(JURIS_DESC)	(JURIS_DESC)	(ADDR_PCT_CD)	(LOC_OF_OCCUR_DESC)
Base Type	INT	TEXT	TEXT	TEXT	TEXT	TEXT	INT	TEXT
Smntc Type	internal classification code	description of internal classification	status of crime	Level of offense	Jurisdiction	Borough Name	Precinct	location of occurrence in or around the premises
Num valid	5096657	5096657	5101224	5101231	5101231	5100768	5100841	3973890
Num invalid	0	0	0	0	0	0	0	0
Num null	4574	4574	7	0	0	463	390	1127341 (missing values are empty values and whitespace)
Different kinds of values in the same column	No. All of them are integer.	No. All of them are string.	No. All of them are string.	No. All of them are string.	No. All of them are string.	No. All of them are string.	No. All of them are integer.	No. All of them are string.
Different semantic types in the same column	No.	No.	No.	No.	No.	No.	No.	No.
Valid checking criteria	Three digit integer -> valid	String with length greater than zero-->valid	In ['COMPLETED','ATTEMPTED']	String with length greater than zero-->valid	String with length greater than zero-->valid	String with length greater than zero-->valid	Integer value -> valid	In ‘INSIDE’,'FRONTOF','OPPOSITE

			->valid					OF','REAR OF','OUTSIDE'
Outliers/Interesting/Suspicious values	None.	Some end with “-” or “,”	None.	None.	None.	None.	None.	None.

column	C_17 (PREM_TYP_DESC)	C_18 (PARKS_NM)	C_19 (HADEVELOPT)	C_20 (X_COORD_CD)	C_21 (Y_COORD_CD)	C_22 (Latitude)	C_23 (Longitude)	C_24 (Lat_Lon)
Base Type	TEXT	TEXT	TEXT	INT	INT	DECIMAL	DECIMAL	SET
Smntc Type	description of premises	name of park, , playground or greenspace of occurrence	name of NYCHA housing development of occurrence	X-coordinate for NY State Plane Coordinate System Long Island zone	Y-coordinate for NY State Plane Coordinate System Long Island zone	Latitude with WGS 1984 standard	Longitude with WGS 1984 standard	Latitude and Longitude tuple
Num valid	5067952	7599	253205	4913085	4913085	4913085	4913085	4913085
Num invalid	0	0	0	0	0	0	0	0
Num null	33279 (all missing values are empty values)	5093632 (all missing values are empty values)	4848026 (all missing values are empty values)	188146 (all missing values are empty values)	188146 (all missing values are empty values)	188146 (all missing values are empty values)	188146 (all missing values are empty values)	188146 (all missing values are empty values)
Different kinds of values in the same	No. All of them are string.	No. All of them are string.	No. All of them are string.	No. All of them are integer.	No. All of them are integer.	No. All of them are Float.	No. All of them are Float.	No. All of them are tuple (set in sql)

column								
Different semantic types in the same column	No.	No.	No.	No.	No.	No.	No.	No.
Valid checking criteria	String with length greater than zero-->valid	String with length greater than zero-->valid	Check whether the value is in the list of NYCHA housing[1] development	check int and Bounds: 909126.0155 to 1610215.359	check int and Bounds[2]: 110626.2880 to 424498.0529	check float and Bounds: 40.47 to 41.31	check float and Bounds: -74.27 to -71.75	String with length greater than zero-->valid
Outliers/Interesting/Suspicious values	None. (After sorting alphabetically there is no interesting values.)	None. (After sorting alphabetically there is no interesting values.)	None.	None.	None.	None.	None.	None.

4. Other data quality issues

A. Time Order of Occurrence

We notice that among 3494907 records whose starting date/time of occurrence and ending date/time of occurrence all exist, 531 of them have time order problem, i.e. ending date/time is prior to starting. For records that date and time both exist, we combine them to get the exact time and date of occurrence before comparing.

B. Time Order of Report

Similar to the previous issue, we also notice that among 5100576 records whose starting date of occurrence and reporting date both exist, 2 of them have time order problem, i.e. reporting time is prior to starting.

C. Map from Offence Classification Code to its description

There are 6 Offence Classification Code (KY_CD) has more than one corresponding Description (OFNS_DESC), including 120, 124, 343, 345, 364, 677. For example, code 345 has ‘ENDAN WELFARE INCOMP’ and ‘OFFENSES RELATED TO CHILDREN’.

5. Hard/ambiguous cases

A. We would like to check whether there is any mismatch between X-coordinate, Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104) and Latitude/Longitude. However, it is hard to find correct conversion formula without domain knowledge. Therefore, we chose to check the bounds of X-coordinate, Y-coordinate, Latitude/Longitude.

B. It is ambiguous that which attribute we should trust if there is a violation of attribute dependencies.

C. It is hard to get an exhaustive list of parks,playgrounds and greenspaces in NYC to check whether a park/playground/greenspace is indeed in NYC. The available list of parks in nycgovparks.org is not exhaustive.

D. In column PD_DESC(description of internal classification), there are some value like “GAMBLING 2, PROMOTING, POLICY-” or “LARCENY,PETIT FROM OPEN AREAS,”, which end with “-” or “,”. But these cases appeared not only once, and they match with digit internal classification code. So now we consider these data valid.

6. Data Quality Summary

1. Completeness
   Quotient of number of missing values and records over all represented entities: 8.17%
2. Uniqueness
   There is no identical data records in our dataset.
   Number of records that represent the same entity: 0
3. Timeliness
   For our purpose, there is no records/values that are out of date.
4. Schema Conformance (Syntactic Integrity)
   Number of values that violate format constraints: 0
5. Integrity (Semantic Integrity)
   Number of records that violate integrity constraints: 0
6. Accuracy
   Quotient of number correct values and the overall number of values: almost 0.

7. Data Summary Visualization

For this part, we constructed a spark dataframe, and basically used GroupBy method in SparkSQL to generate summary statistics.

1.The distribution of number of crime records per day.
We only considered crimes after 01/01/2006. Most days have 800 to 1800 crime records, while only 2 days less than 500. The histogram below shows the distribution. The distribution is left-skewed with median around 1400.

Figure 1: histogram of number of crime records per day after 01/01/2006

2. Number of crimes per day

We plotted number of crimes against each day, hoping to find days with too many or few records. To demonstrate this purpose, we only listed the plot for year 2006.

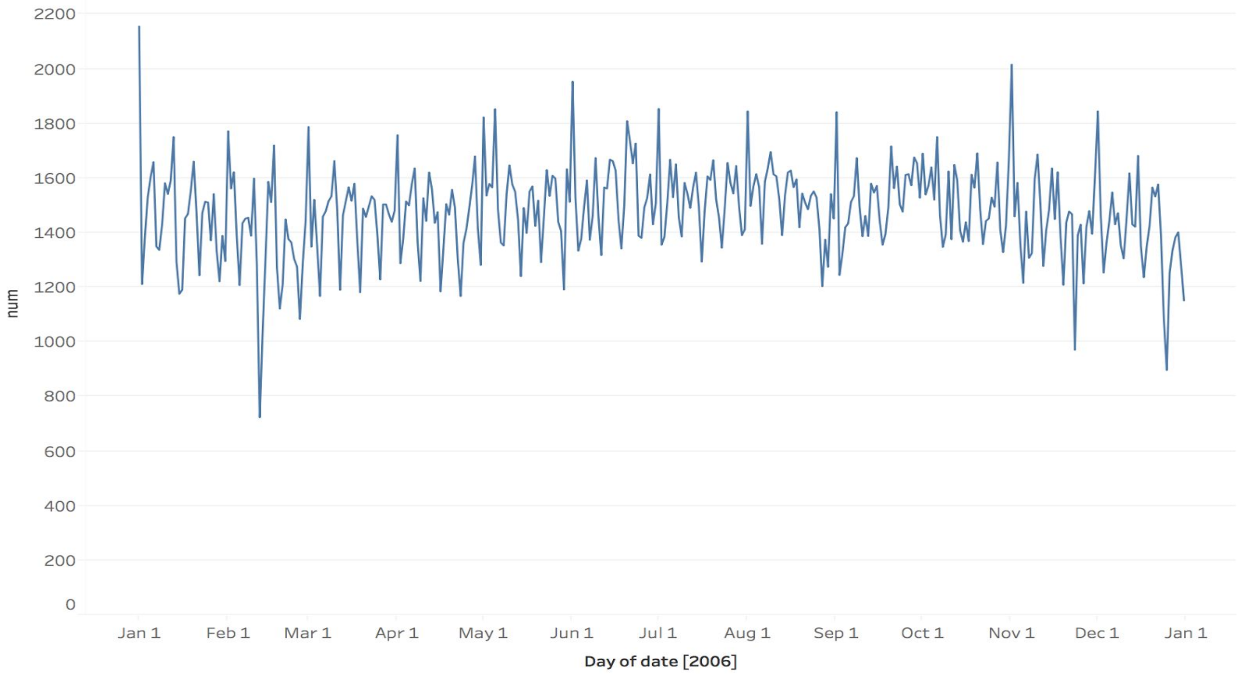


Figure 2: number of crimes each day in Year 2006

year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Too many	-	-	-	-	-	-	-	-	-	04/01 05/01
Too few	02/12	-	-	03/02	12/27	08/28	10/29	-	-	12/31

3. Trend of number of crime records per month

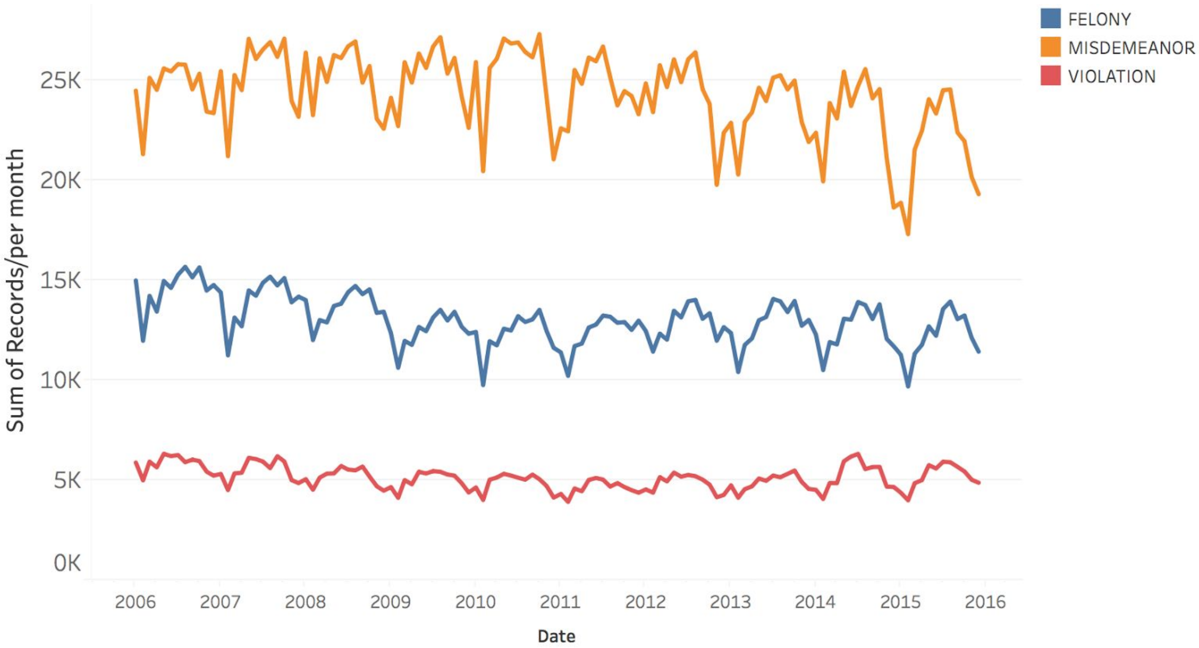


Figure 3: trend of number of crime records per month w.r.t each type of crimes

From figure 3, we can see it seems that number of crimes is lowest at the beginning and end of each year.

4. Number of crime records for each quarter.

From figure 4 below we can see the third quarter has the most crimes for every level of offense.

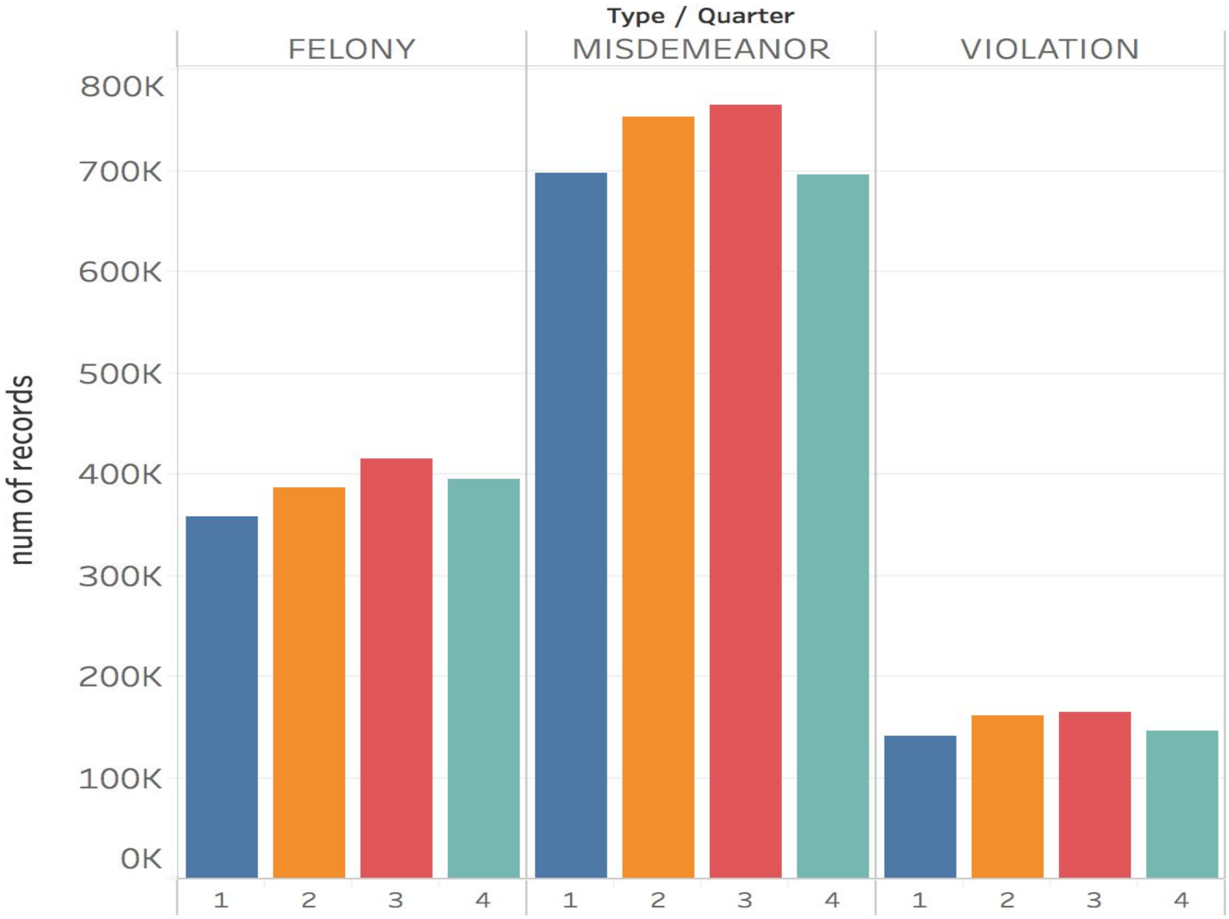


Figure 4: Number of crime records for each quarter

5. Total number of crimes for each year

We can see from figure 5 below that Year 2015 has the lowest total number of crimes.Year 2006 has the highest total number of crimes. Moreover, for every year, “misdemeanor” has highest number of records whereas “violation” has lowest number of records.

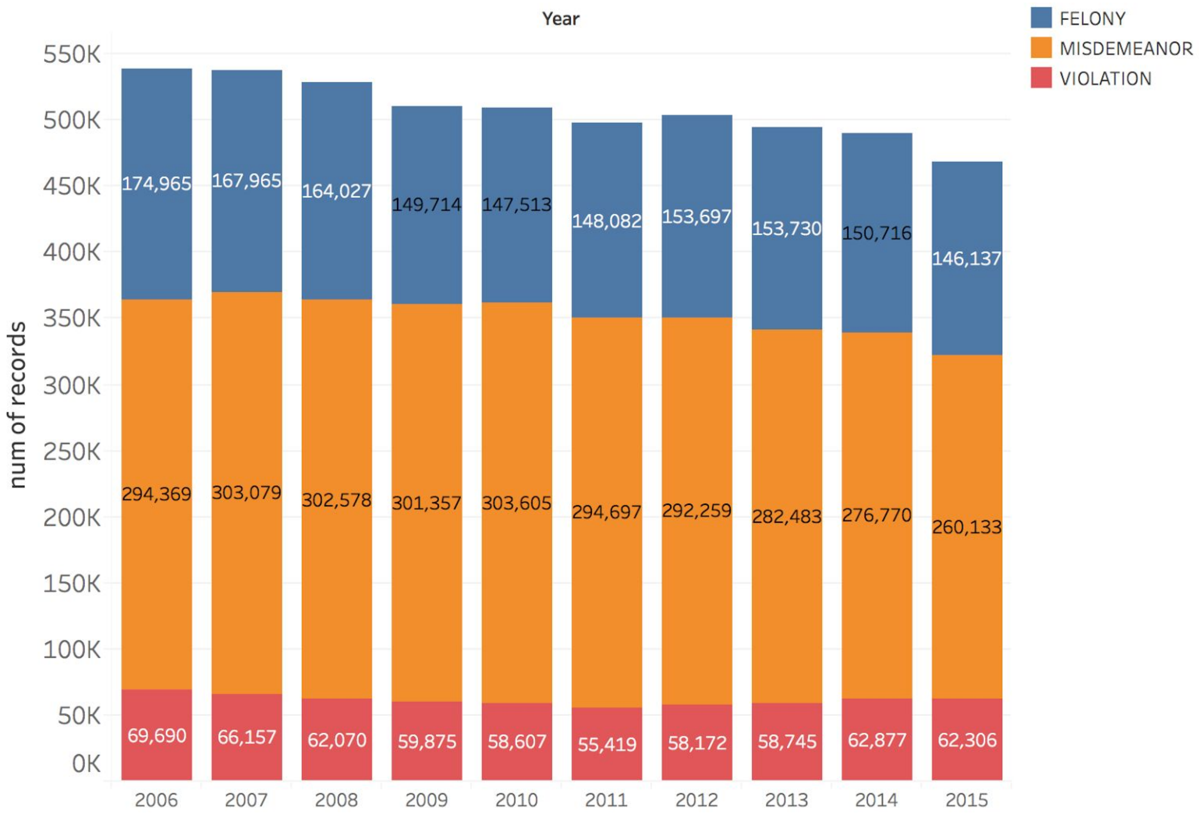


Figure 5: Mosaic plot of total number of crimes for each year w.r.t each type of crime

6. Total number of crimes for each month

We can see from figure 6 below February has the lowest total number of crimes. July and August has the highest total number of crimes.

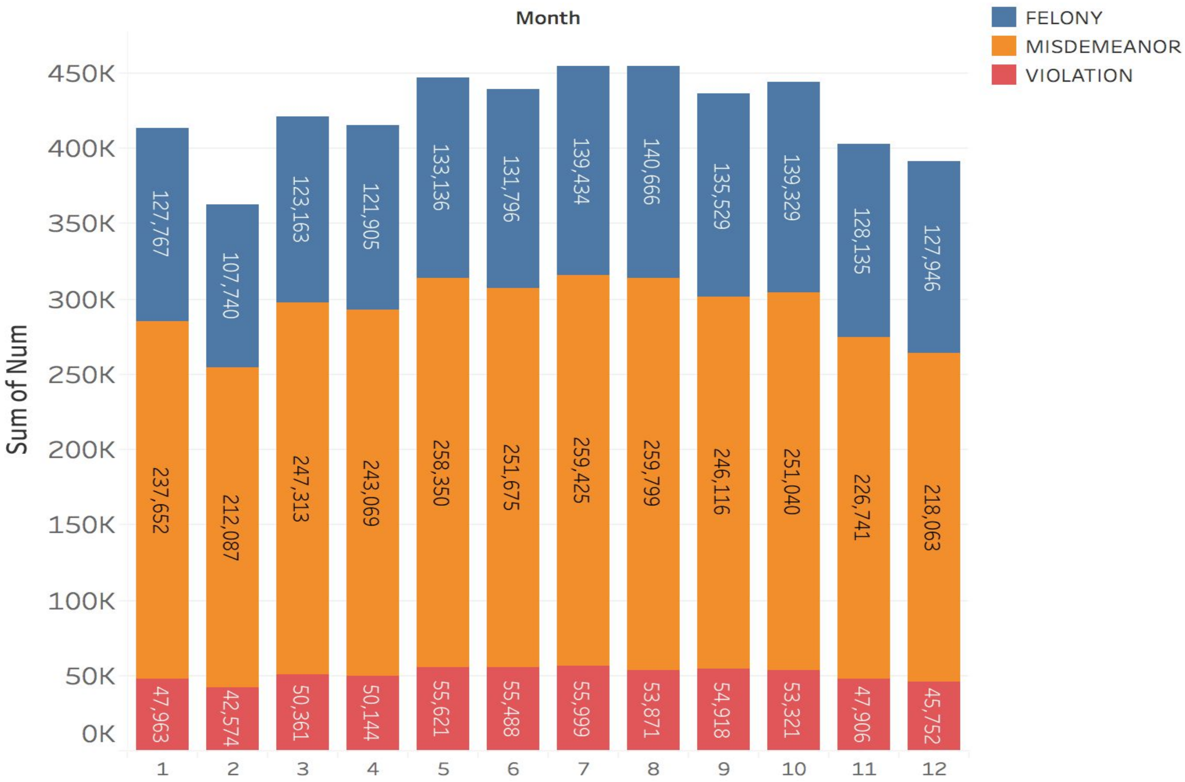


Figure 6: Mosaic plot of total number of crimes for each month w.r.t each type of crime

7. Total number of crimes for each weekday.

We can see from figure 7 below Sunday has the highest total number of crimes. Friday has the highest total number of crimes.

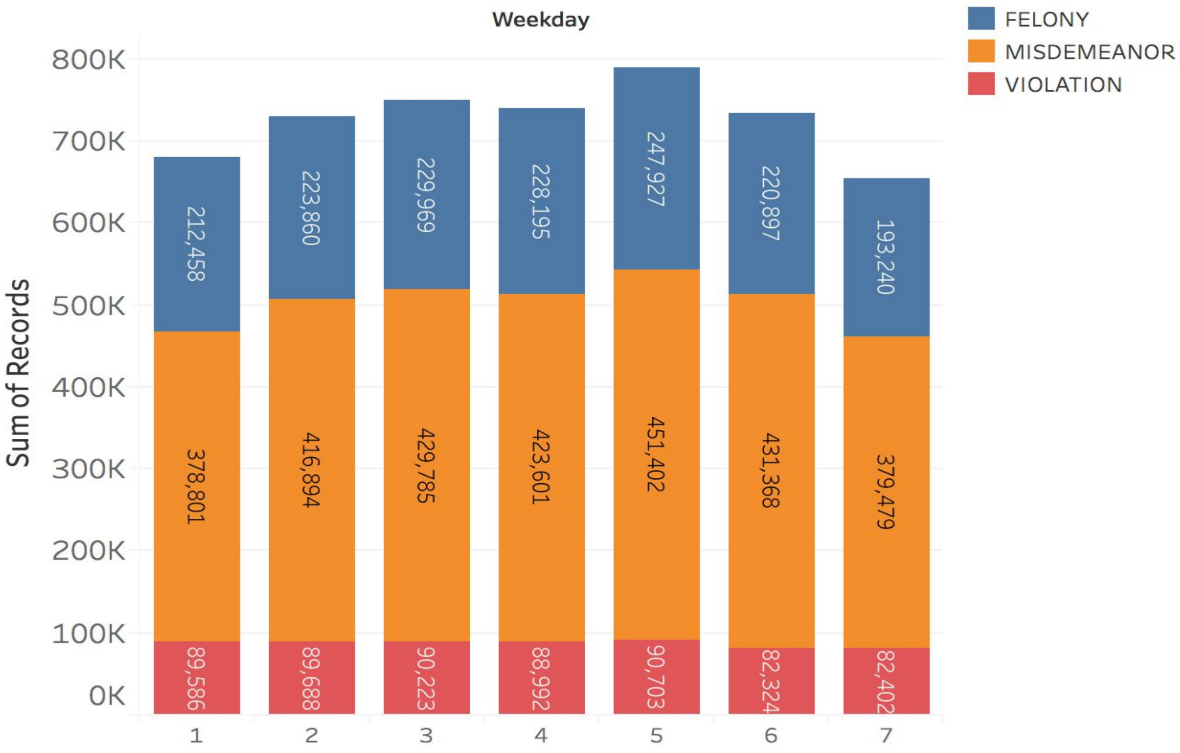


Figure 7: Mosaic plot of total number of crimes for each weekday w.r.t each type of crime

8. Attempted VS Completed

We can see from figure 8 below the majority of crimes in records are completed. It does not necessarily mean there are more crimes out there are completed than those attempted. The reason we have majority of crimes completed in our dataset is that completed crimes are more likely to be reported to NYPD.

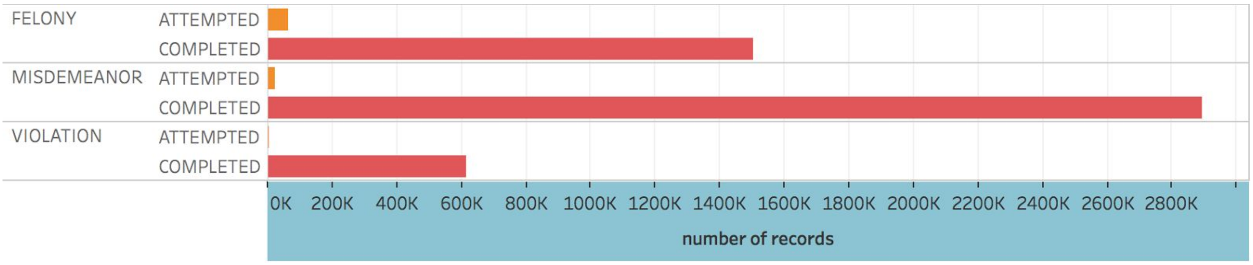


Figure 8: Bar chart of attempted crimes against completed crimes w.r.t each type of crime

9. Number of crime records for each borough

We can see from figure 9 below queens and staten island are relatively safer than bronx, brooklyn and manhattan based on number of records in NYPD.

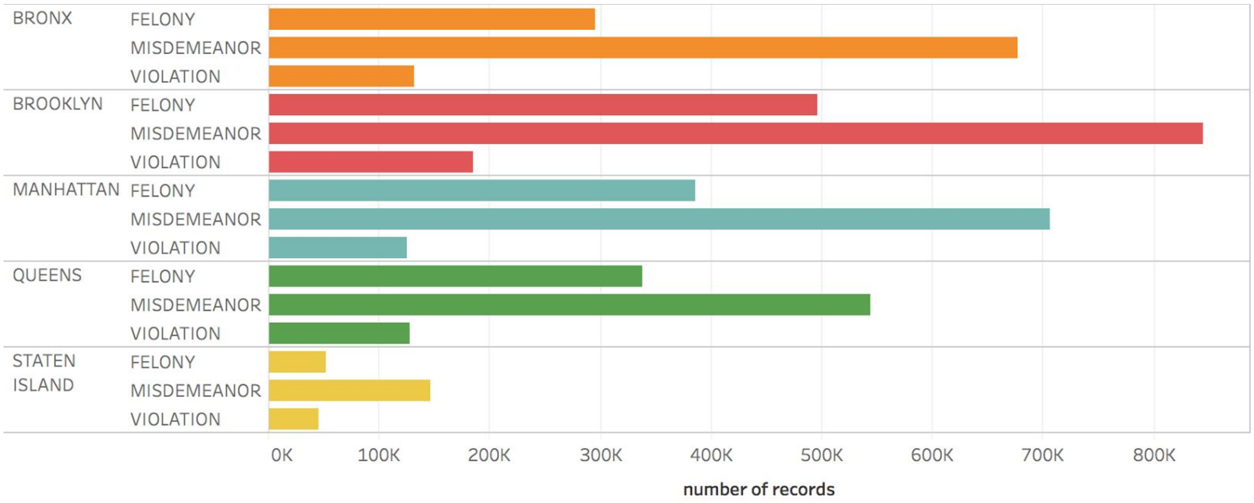


Figure 9: Bar chart of number of records in each borough w.r.t each type of crime

10.Top 20 parks/playgrounds/greenspaces with most crime records
From figure 10 below, we can see central park has the highest number of crime records. Flushing meadows corona park has the second highest number of crime records.

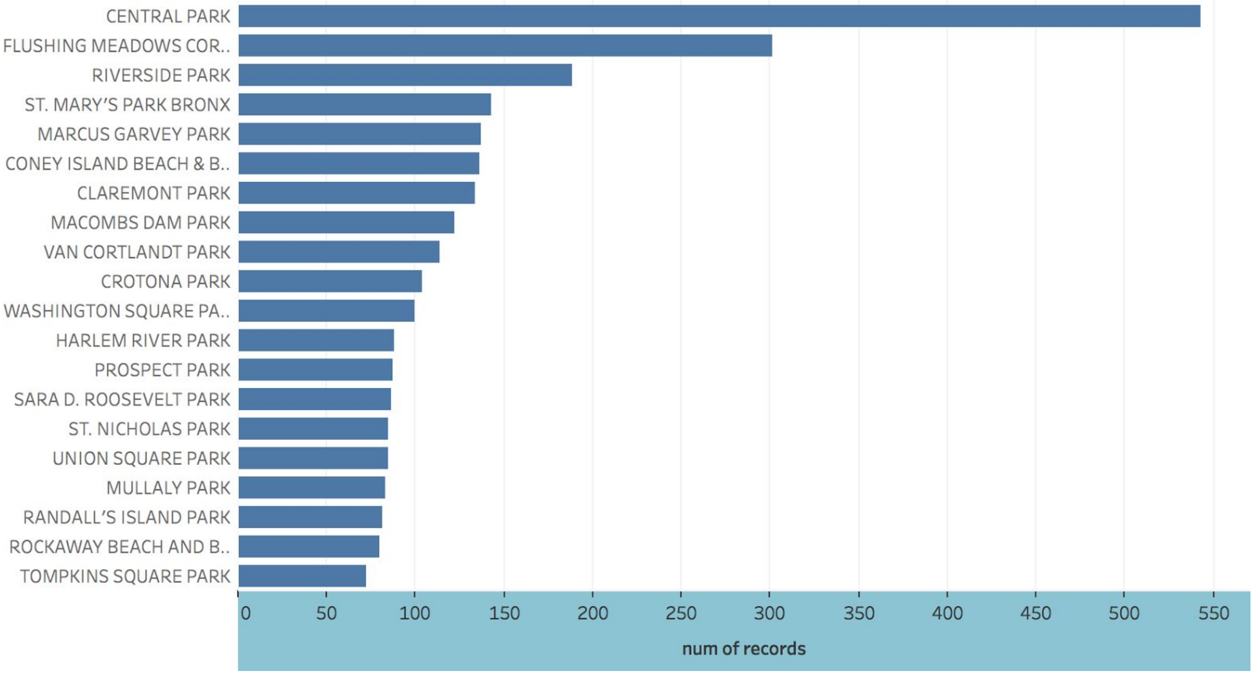


Figure 10: Bar chart of TOP 20 parks/playgrounds/green spaces in NYC w.r.t number of crime records

11. Top 20 Jurisdiction responsible for incident with most crime records
From the figure below, we can see NYPD handles almost all crimes while N.Y.Housing police shares some burden of crimes handling.

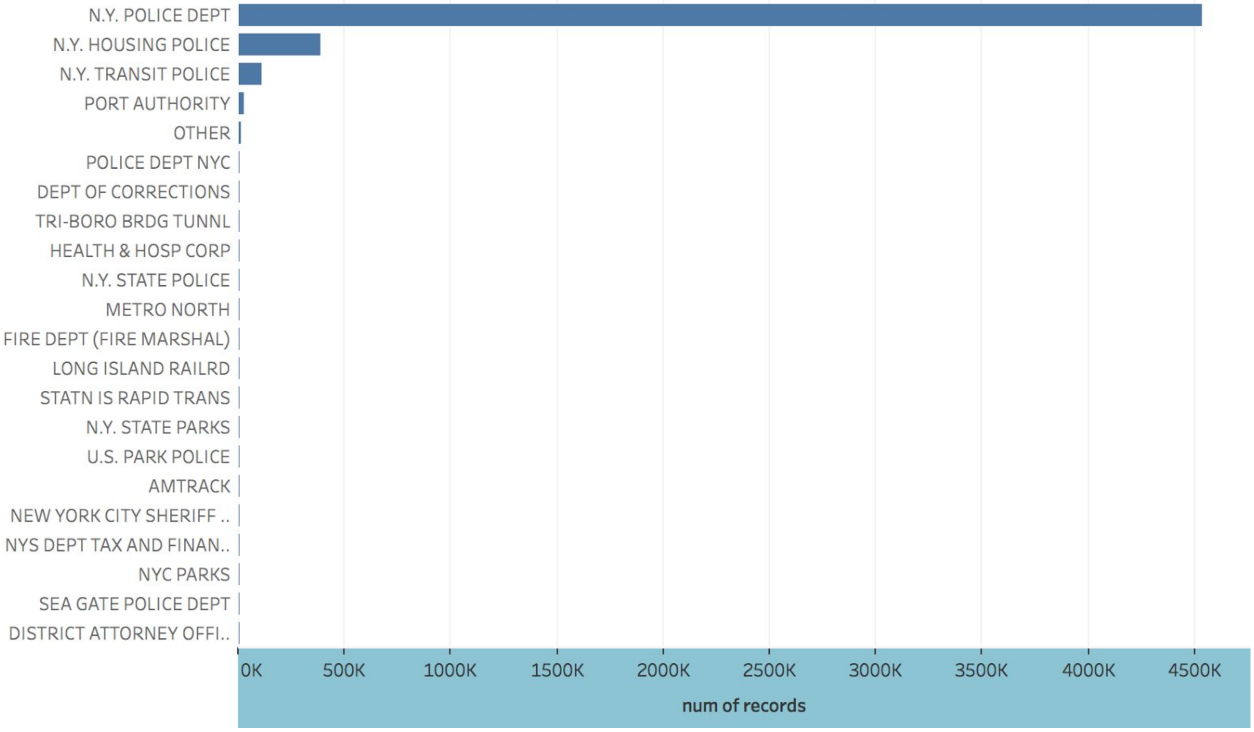


Figure 11: Bar chart of TOP 20 Jurisdiction responsible for incident

12. Top 20 precincts with most crime records.
From the figure 11 below, we can see precinct coded as 75 has the highest number of crime records. Precinct 43 has the second highest crime number. Based on NYC government website, precinct 75 is located in the East New York section of Brooklyn, and precinct 43 is southeast section of the Bronx.

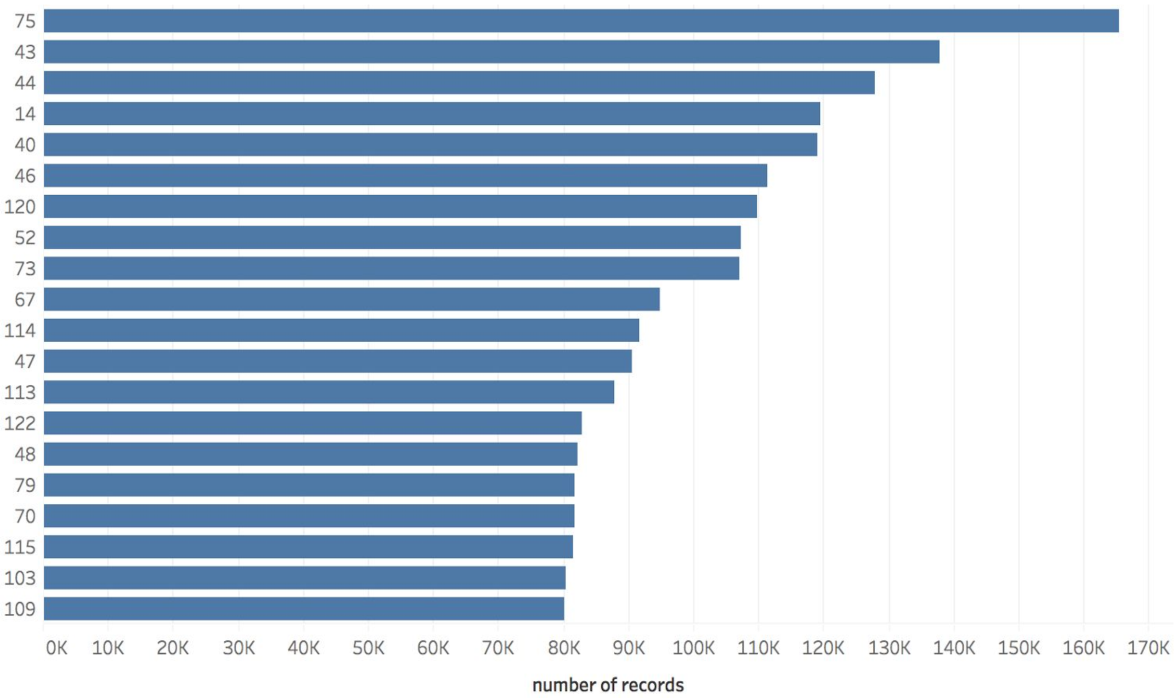


Figure 11: Bar chart of TOP 20 precincts w.r.t to crime records

12.Top 20 premises with most crime records
From the figure 12 below, we can see most crimes happened on the street while the second most crimes happened in residences.

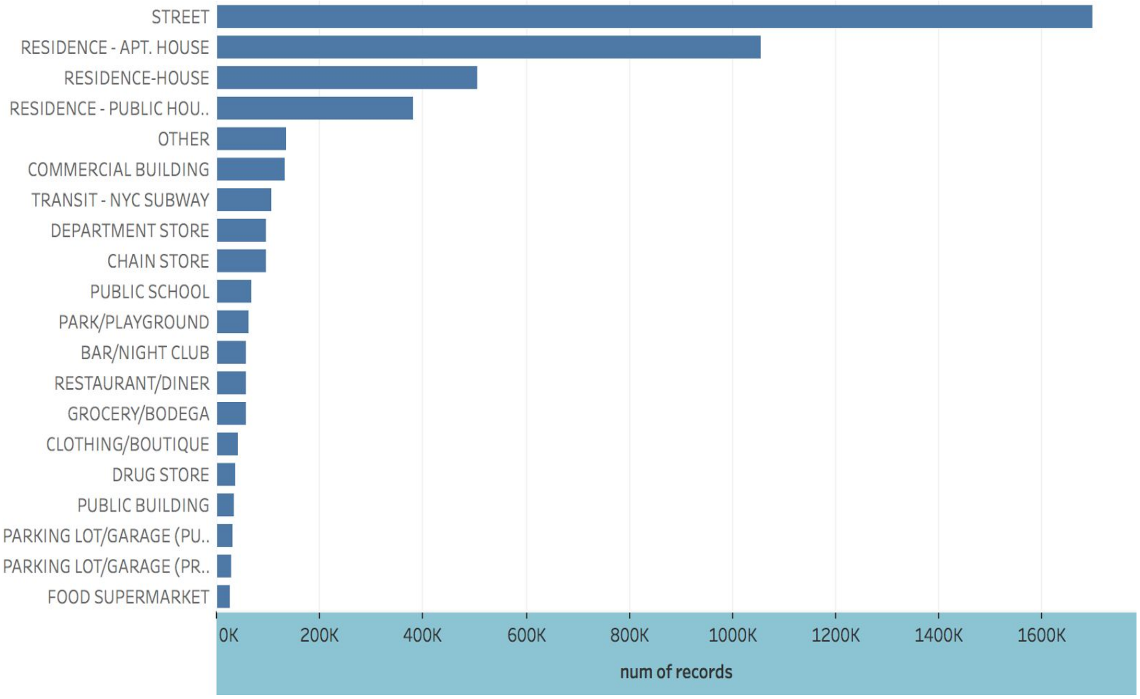


Figure 12: Bar chart of TOP 20 premises w.r.t to crime records

13. Geographical distribution of crimes.
For this and the next plots, we randomly sampled 10,000 records from August 2014.
From this map of density of crimes, we can see Manhattan, Bronx and some region in Brooklyn has the most crimes, which is in accordance with our previous observation.

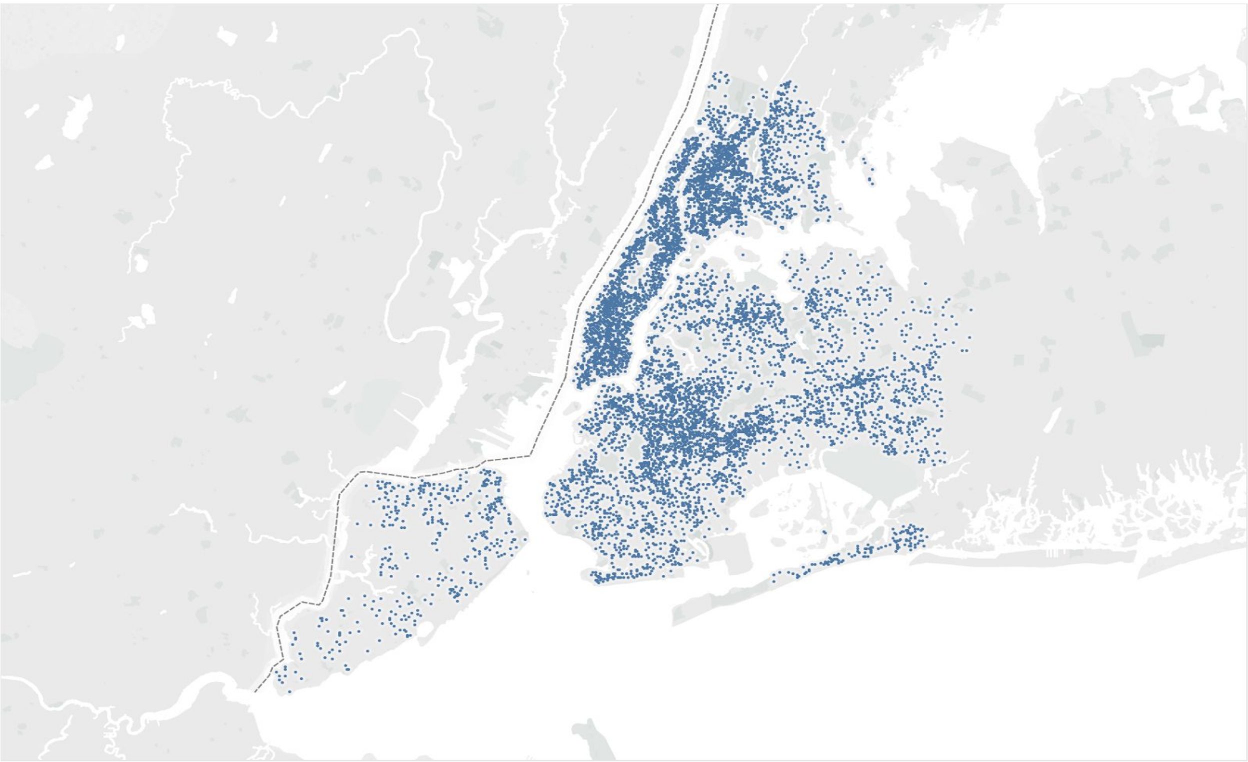


Figure 13: Geographical distribution of crimes in NYC

14. Number of crime record for each postcode.
From the map below, we can clearly see some regions in Brooklyn have the most crimes.

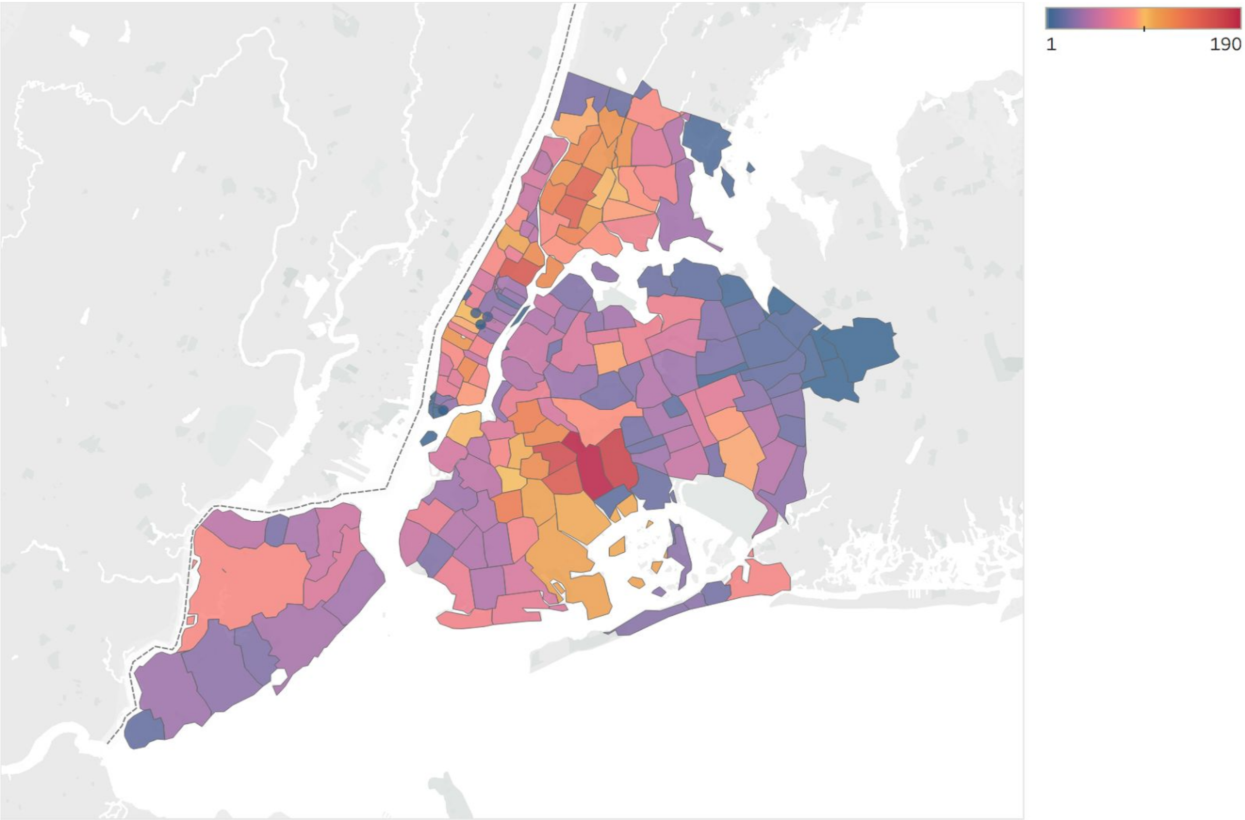


Figure 14: Number of crime record for each postcode.

PART2

Note: Since crime instances with unknown occurrence date are recorded to the first day of that year, we will impute number of crimes on the first day of each year with average number of that year.

Experimental Setup:

We used SparkSQL to perform a lot of data analysis, including group by, count, and so on. After performing the initial data analysis in Spark, we outputted results as csv file and performed data visualization analysis in either Tableau or iPython notebook.

List of Hypotheses we set out to investigate:

1. Time-Variant hypotheses:
    a. The last quarter of a year will result in higher number of crimes.
    b. November and December might have the highest number of crimes.
    c. Friday and weekends have the highest number of crimes.
2. Location-Variant Hypothesis:
    a. Regions with better education will have lower number of crimes per person.
    b. Demographics:

    i. Regions with higher population or population density will have higher number of crimes.
    ii. Regions with higher teenagers ratio will have higher number of crimes per person.
    iii. Regions with higher single-mother household ratio will have higher number of crimes per person.

3. Economics status related hypotheses:
    a. Higher Unemployment rate will result in higher number of crimes.
    b. Regions with low per-capita income will have higher number of crimes per person.
4. Warm weather will result in higher number of crimes:
    a. Crime number is correlated with temperature and wind speed
    b. Crime number is correlated with extreme weather.
5. Looking into details of crime type:
    a. Robbery and frauds tend to happen in regions with higher teenagers ratio.
    b. Robbery and frauds will also happen a lot in regions with higher single-mother household ratio.
    c. Robbery and frauds will happen much less in regions where people have better education

Analyses of Hypotheses:

1. Time-Variant hypothesis:

a. The last quarter of a year will result in higher number of crimes.

Since people have a lot of activities going on in the last quarter of a year such as buying gifts for family, people have stronger incentive to conduct property crime such as theft. Therefore, we hypothesized that the last quarter of a year will result in higher number of crimes.

We made a barchart to see whether our hypothesis is true.

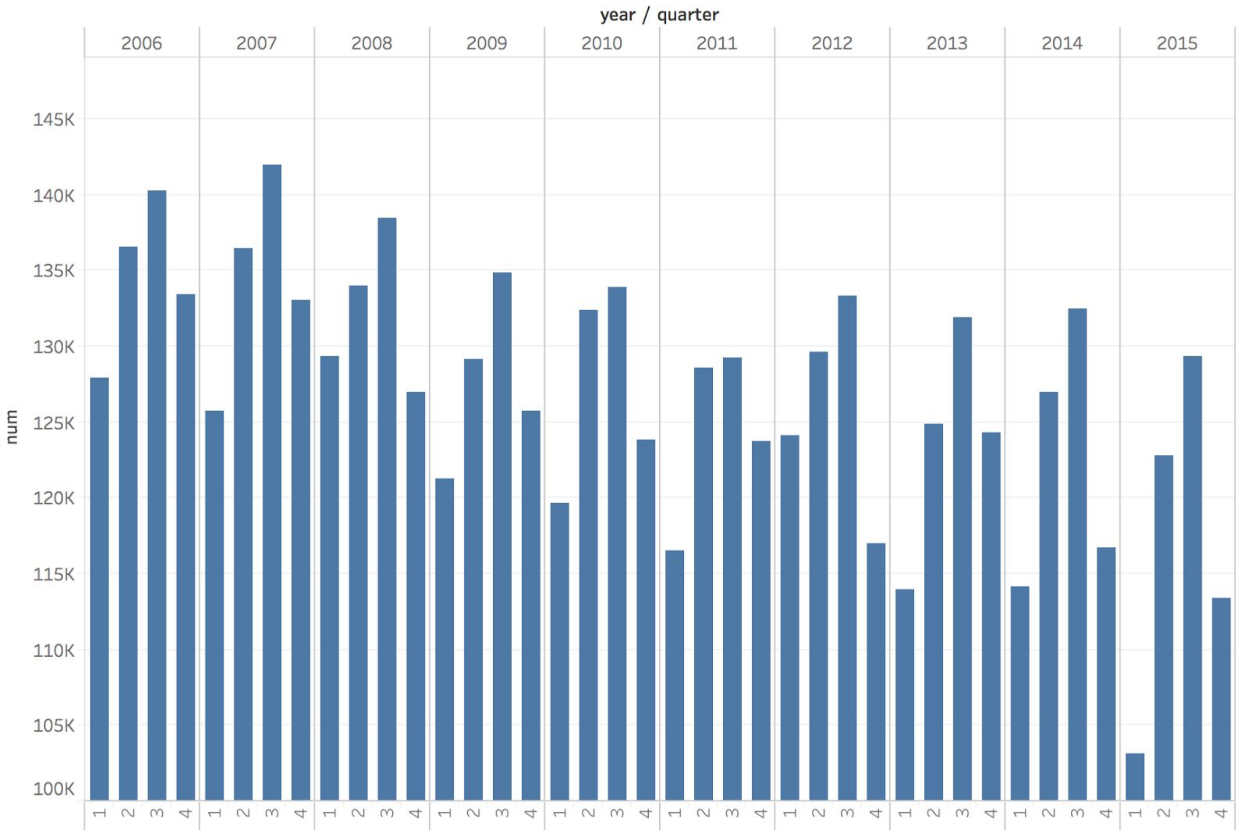


Figure 15: Bar chart of number of crimes each quarter from Year 2006 to Year 2015

We can see from figure 15 that number of crimes is always increasing from the first quarter to the third, the decreasing from the third to the fourth. Therefore, our hypothesis is wrong.

The reason might be that people might want to spend time with families at the end of year rather than conducting crimes and spending all the time in jail. Thus, people might behave well in the last quarter of a year.

The interesting trend that number of crimes is always increasing from the first quarter to the third that make us wonder whether a higher temperature will result in higher number of crimes. We will investigate this whether-related hypothesis later.

b. November and December might have the highest number of crimes.
Because a quarter has three months, the different behaviors of months might be diluted. Therefore, we decide to investigate number of crimes each month over the years between 2006 to 2016.

Since people need money to buy gifts and there are a lot of end-of-year sales, we hypothesized that people have strong incentive to conduct crimes in November and December.

We plotted a time-series plot below to see the trend:

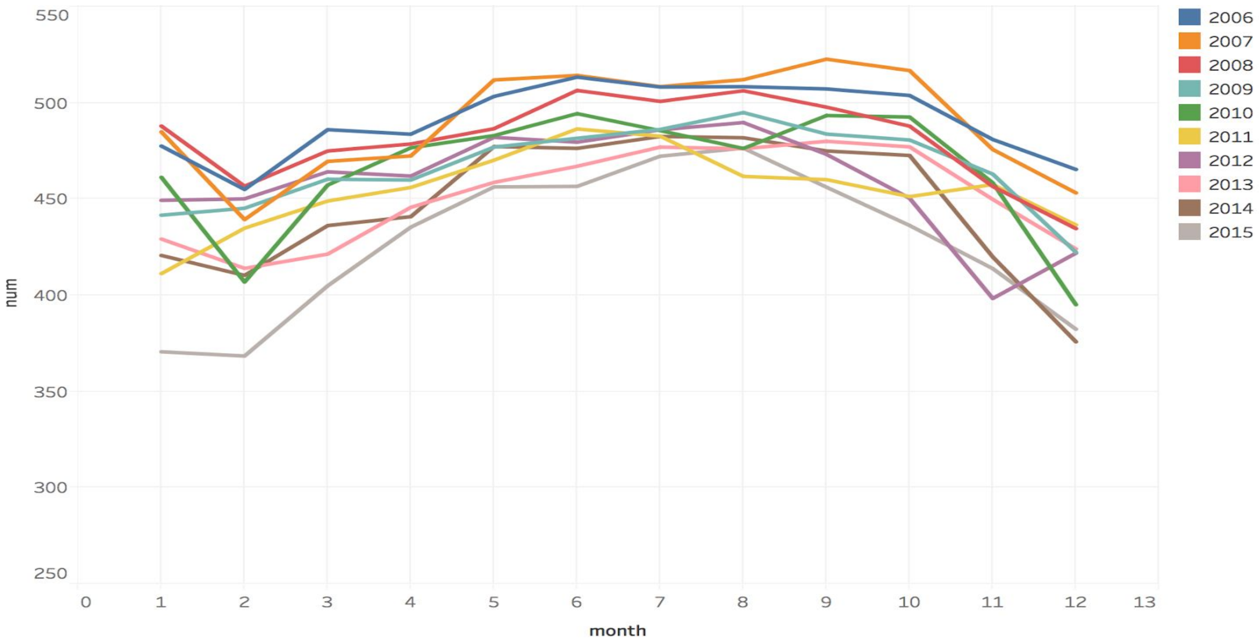


Figure 16: Average number of crimes per day against Month

From figure 16, we can see that generally speaking, February has relatively smaller number of crimes, and there is an increasing trending from February to September, after which number of crimes starts to drop.

This trend contradicts with our hypothesis but agrees with the trend in the quarter bar chart of previous section. The reasons are similar as before. People want to spend time with families at the end of year rather than conducting crimes and spending all the time in jail. Thus, people might behave well in these two months.

c. Week of Day

We hypothesized that Friday and weekends have the highest number of crimes. On one hand, people have more time to conduct crimes.

We made a time-series plot to see the trend:

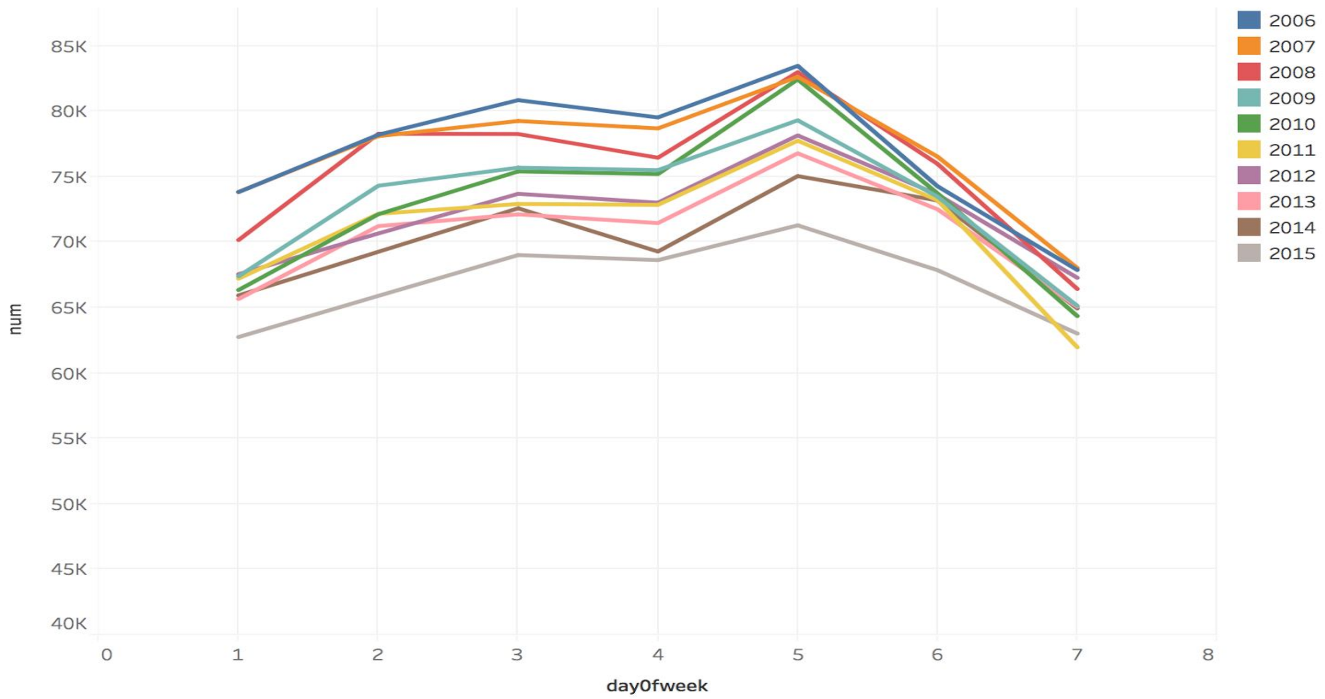


Figure 17: Number of crimes per day against DayofWeek

Generally speaking, there is an increasing trend from Monday to Friday, after which it starts to drop, and Sunday is with least number of crimes. This trend agrees with our hypothesis partially since Friday do have the highest number of crimes. However, weekend actually do not have high number of crimes.

The reason might be that people are less vigilant on Friday so they are more vulnerable to crimes. Villains might need money to spend weekends, so they will conduct crimes on Friday.

2. Location-Variant Hypothesis(Education and Demographical)

Considering that different regions in New York City might have quite different population, we will investigate into number of crimes per person rather than the total number of crimes in one region.

To make data analysis more clear, we used zipcode to partition New York City.

a. Education

The indicator we're using for the education status is average SAT scores per Zipcode. Our hypothesis is that the higher average SAT scores are, the better the education status of that zipcode, which leads to less crimes.

We made two New York City map to assist analysis:

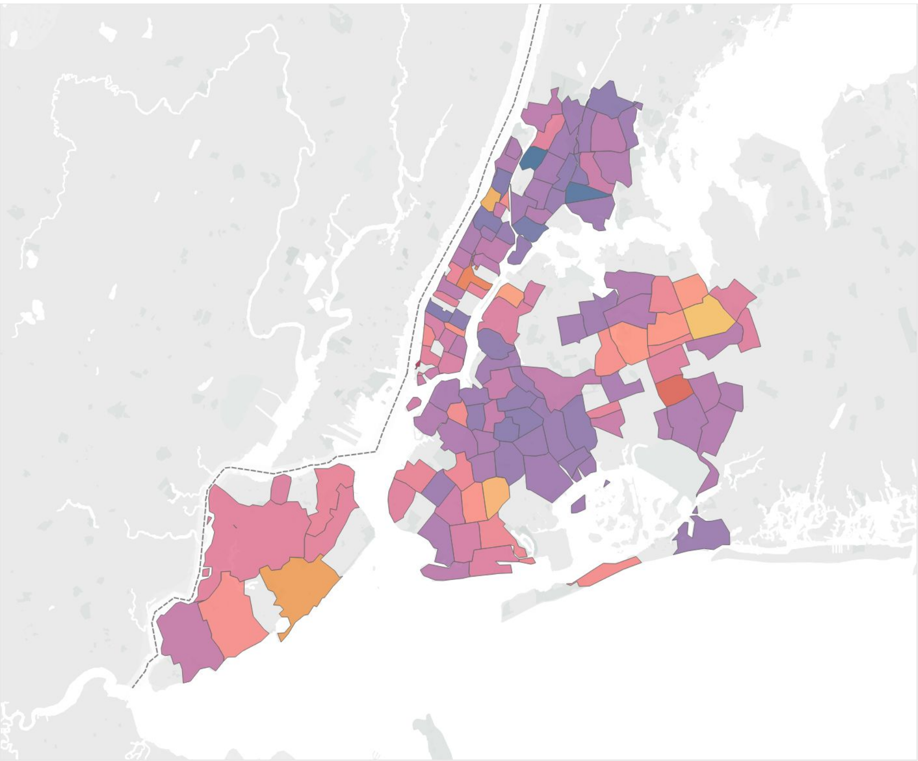


Figure 18: Normalized SAT (Red means high SAT scores)

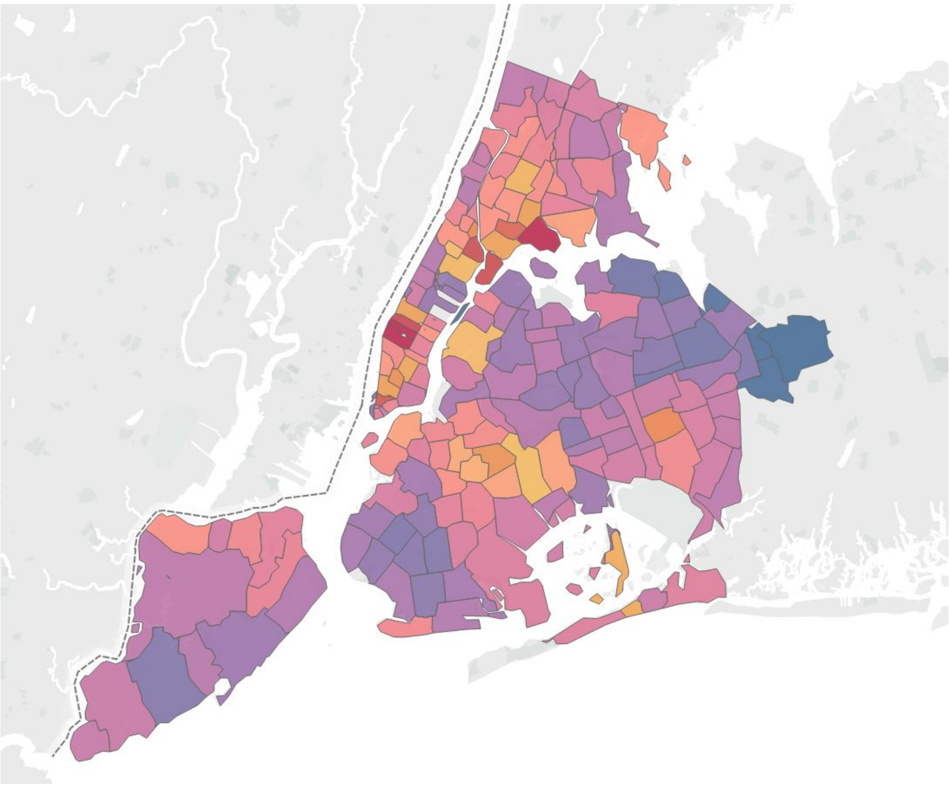


Figure 19: Normalized number of crimes per person (Red means high number of crimes per person)

Then, we made the following plot and analysis.

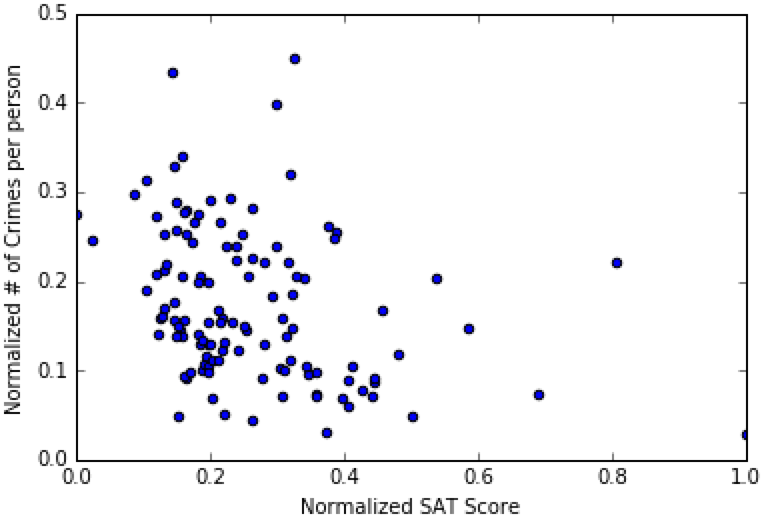


Figure 20: Scatter plot of Normalized SAT scores and number of crimes per person

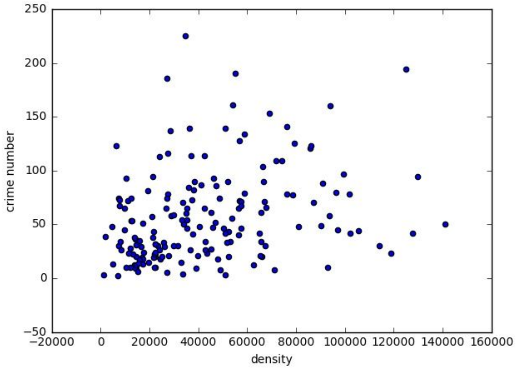
The correlation between the two variables is -0.36, which verifies our hypothesis.

The reasons might be that people with better education have less time and less incentive to conduct crimes.

b. Demographical

i. Population

First, we analyzed the total population and population density in the zipcode area.



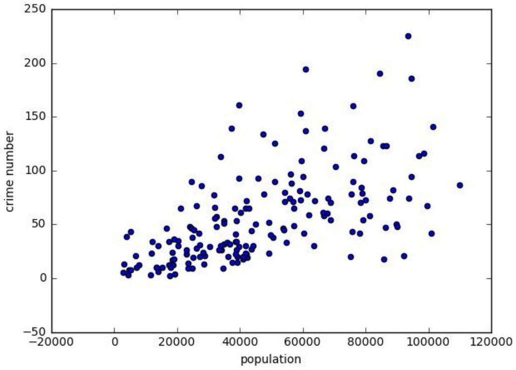


Figure 21: Scatter plot of density and number of crimes

Figure 22: Scatter plot of population and number of crimes

The left plot above is the correlation between population density and crime number. Each point in the graph represent one zipcode area. The right plot above is the correlation between population and crime number. The correlation between population density and crime number is 0.305, and the correlation between population and crime number is 0.592.

ii. Teenager ratio

For this part, we consider teenager(age 15-25) ratio in the zipcode area. Our hypothesis is that teenagers are more likely to conduct crimes. The indicator we're using is teenager ratio per zipcode area.

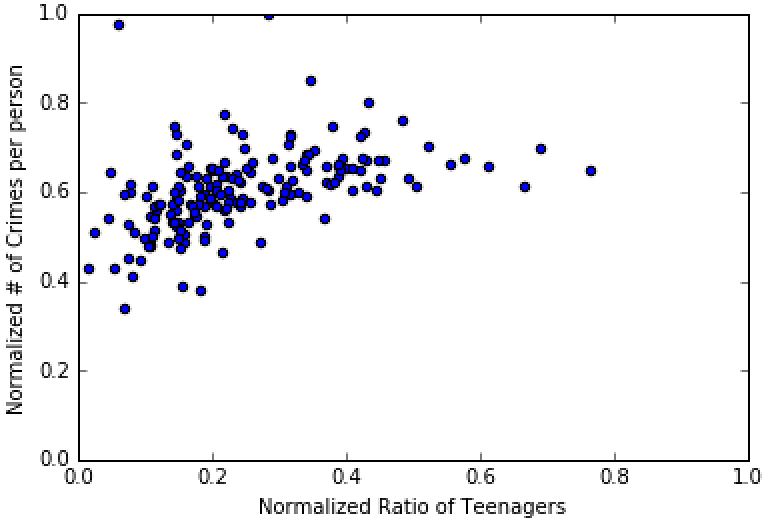


Figure 23: Scatter plot of teenager ratio and number of crimes per person

The correlation between the two variables is 0.39, which verifies our hypothesis.

The reason might be that teenagers are more vulnerable to crimes.

iii. Single-Mother household ratio

For this part, we consider Single-Mother household ratio in the zipcode area. Our hypothesis is that zipcode area with more single-mother households might have more crimes.

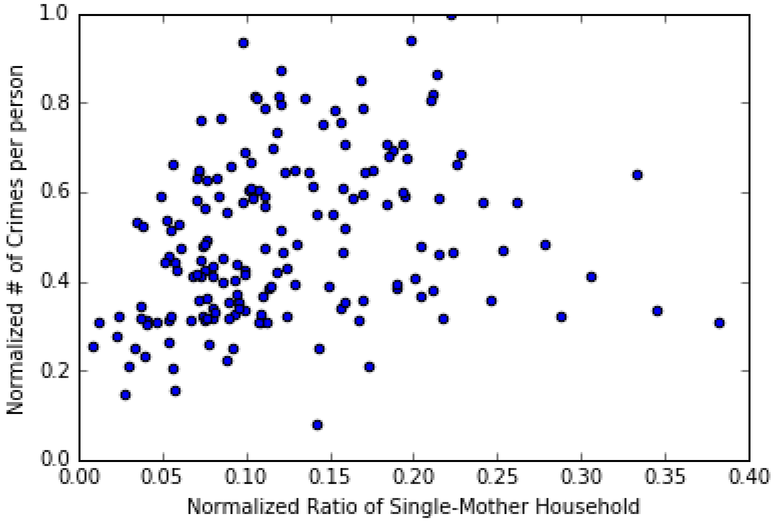


Figure 24: Scatter plot of single-mother ratio and number of crimes per person

The correlation between the two variables is 0.49, which verifies our hypothesis.

The reason might be that single-mother are also quite vulnerable to crimes.

3. Economics

a. Relationship between unemployment rate and number of crimes over years

The hypothesis we had is that there is a positive correlation between unemployment rate and number of crimes. We only selected Year 2006 to Year 2015 for the purpose of visualization.

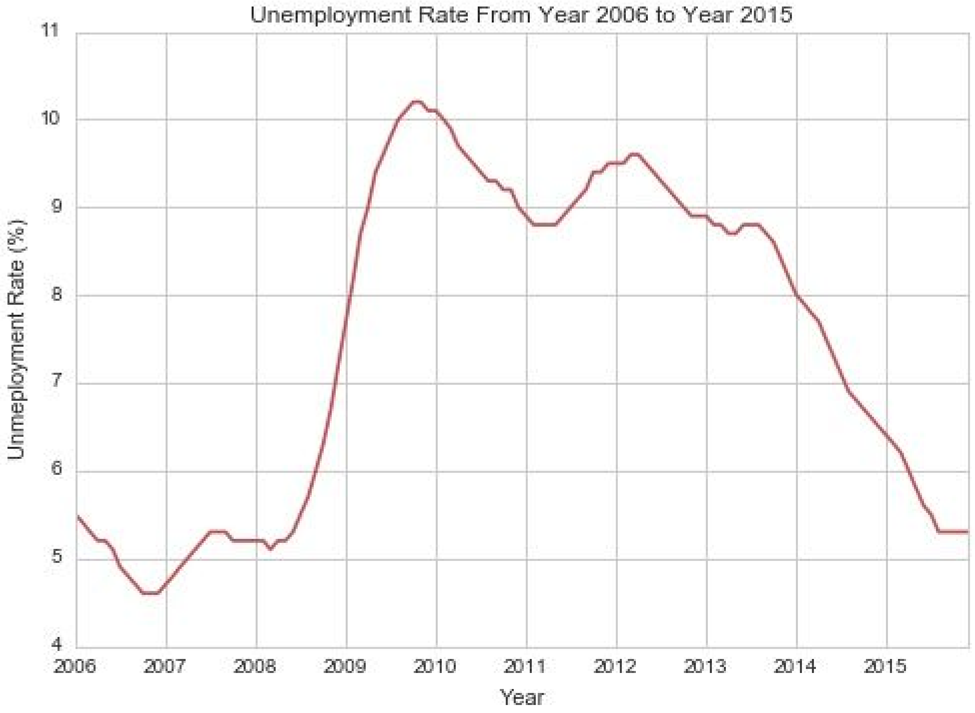


Figure 25: Unemployment rate from Year 2006 to Year 2015

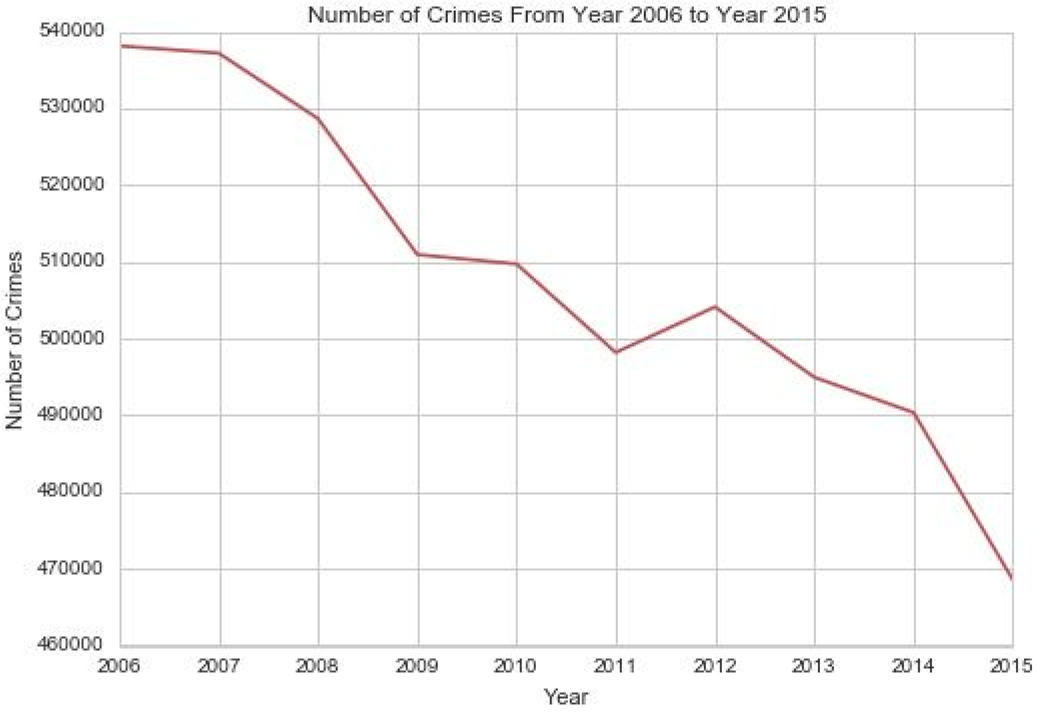


Figure 26: Number of crimes from Year 2006 to Year 2015

From these two graphs, we can not see a clear correlation between unemployment rate and number of crimes. As we can see from the first graph, unemployment rate soared at the year of 2008 due to the Financial Crisis and has gradually went down since year of 2010. However, there is no abrupt rise in number of crimes at the year of 2008 based on the second graph. The number of crimes even went down between the year 2008 to year 2010, which is the opposite direction of unemployment rate.

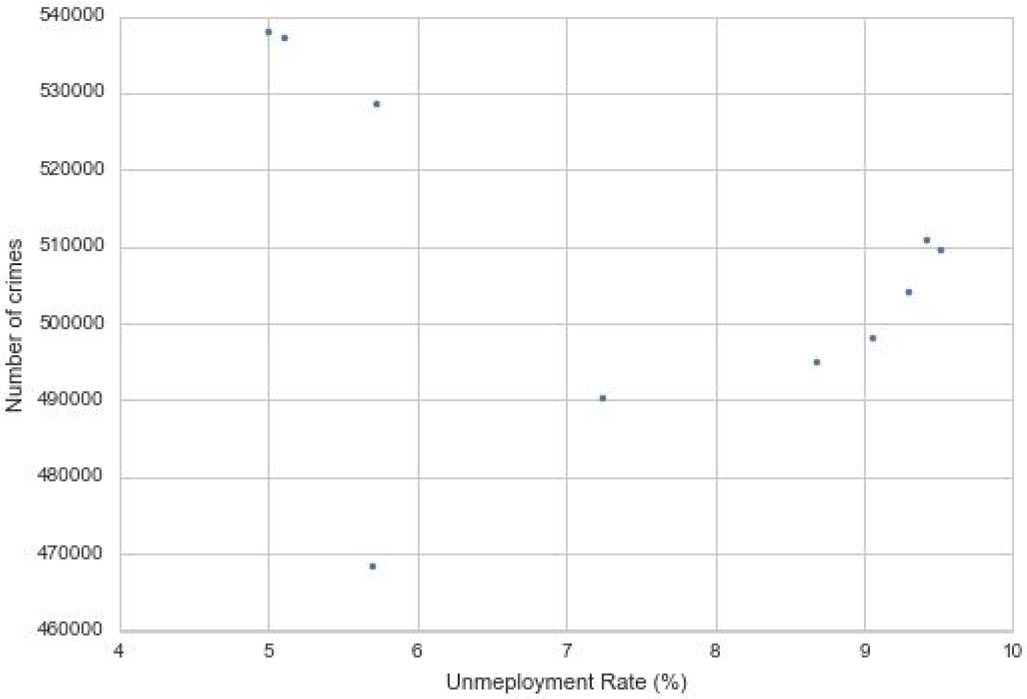


Figure 27: Scatter plot between Unemployment rate and number of crimes

Plotting them in a scatter plot, we can see there is little correlation among these two variables. The correlation calculated is -0.35.

Therefore, our hypothesis is wrong. There is little correlation between number of crimes and unemployment rate. The reason might be that there are only around 30% of crimes are property crime (burglary, larceny, fraud) while the remaining are violent crimes in our data set. While people might have stronger incentive to conduct property crime when they are unemployed, the incentive for violent crime is not strong. People might even be worried that leaving crime records on court would make it even harder to find a job. Thus, people might have tend not to conduct violent crimes. Therefore, we saw a negative correlation between unemployment rate and number of crimes.

b. Number of crimes per person and Per-capita income
The hypothesis we had is that the lower per-capita income a region has, the higher number of crimes will be.

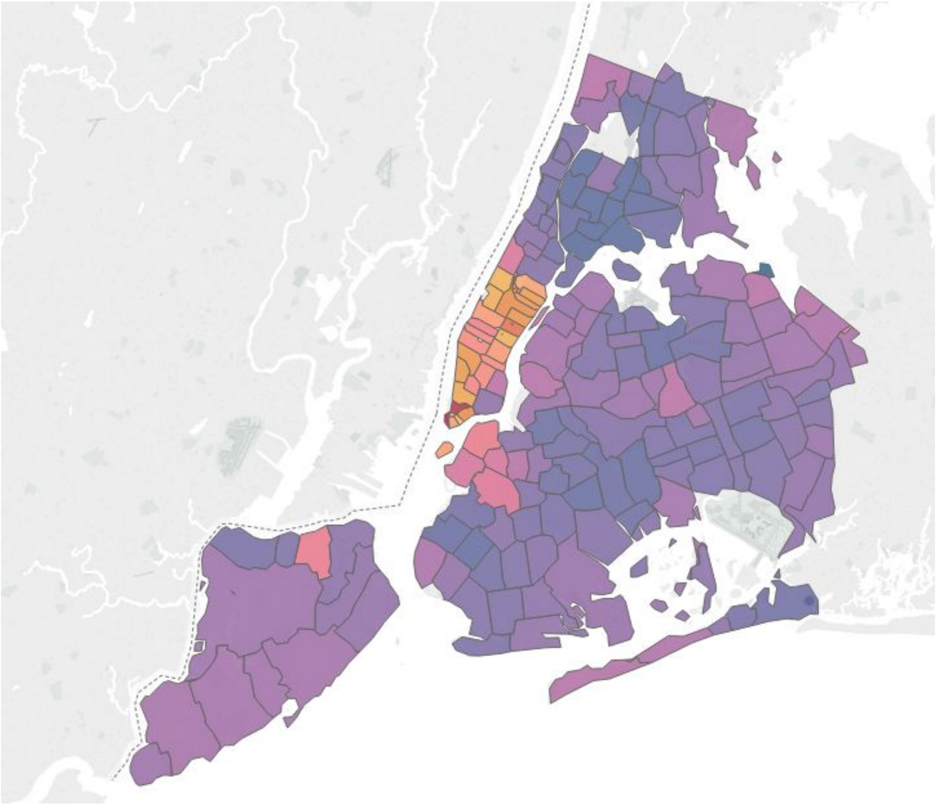


Figure 28: Per-capita income by zip code (The more blue the color is, the lower per-capita income that zipcode region has.)

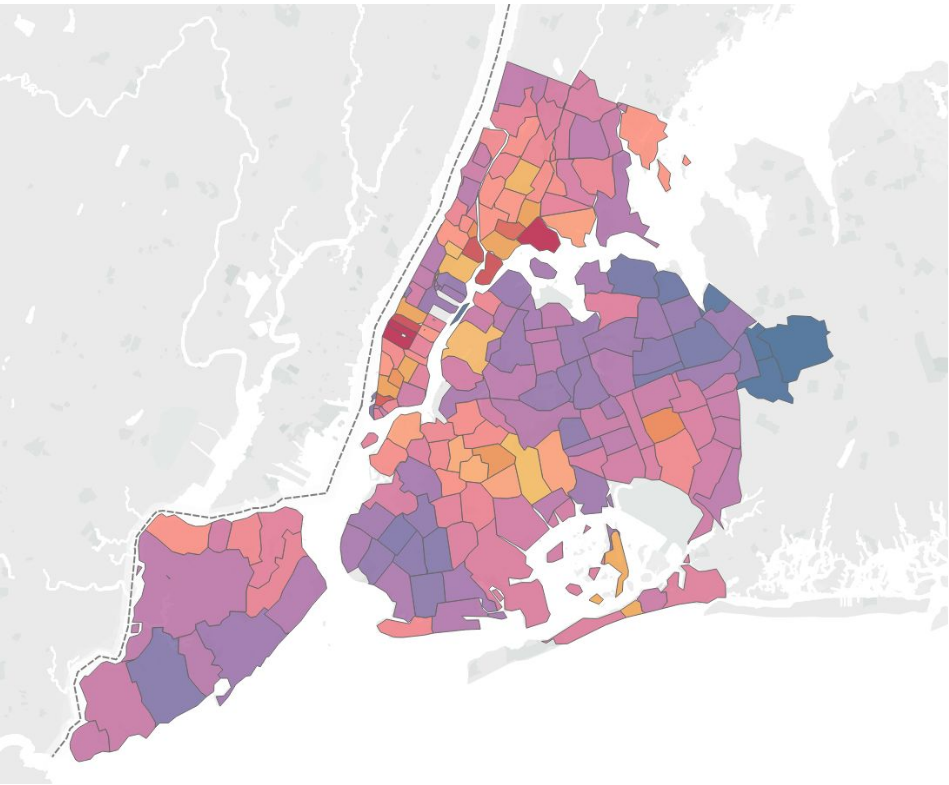
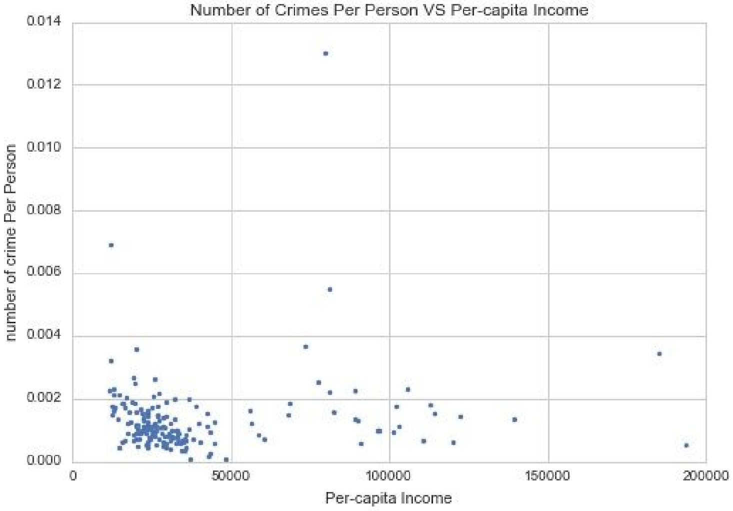


Figure 29: Number of Crimes Per Person by zip code (The more blue a region is, the lower number of crimes that zipcode region has.)

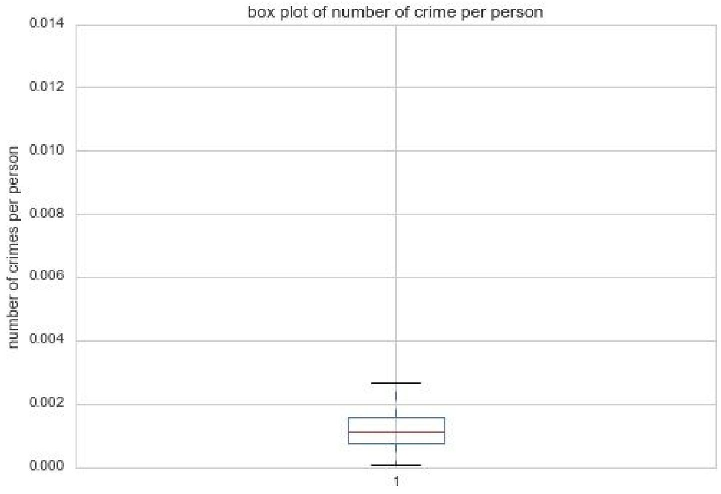
We found that some regions in neighborhoods with low per-capita income such as the red and orange regions in Brooklyn and Bronx indeed have a high number of crimes per person. For example, the red regions in Bronx with the highest number of crimes do have low income per-capita.

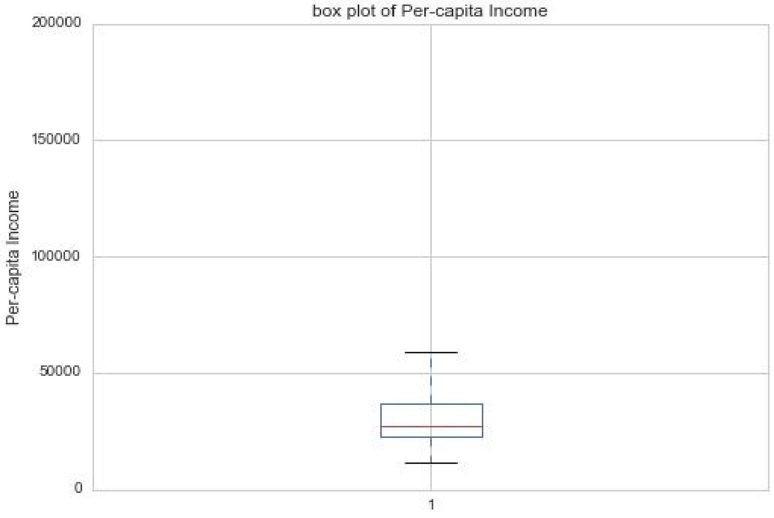
However, a relatively high income per-capita cannot guarantee a low crime number. They might also have high number of crimes per person. To illustrate, one of the red regions in Manhattan corresponds to relatively high-region area in the first graph.

To confirm our observation from visualization, we calculated correlation between number of crimes per person and per-capita income. Before removing any outliers, the correlation is 0.137. The scatter plot is following.



By removing outliers based on boxplots below in both number of crimes per person and per-capita income, the correlation between these two variables is -0.38.





The scatter plot after removing outliers is below:

Therefore, we conclude that in general, the higher per-capita income one region has, the lower number of crimes per person will be. However, for extreme values, this correlation might not hold.

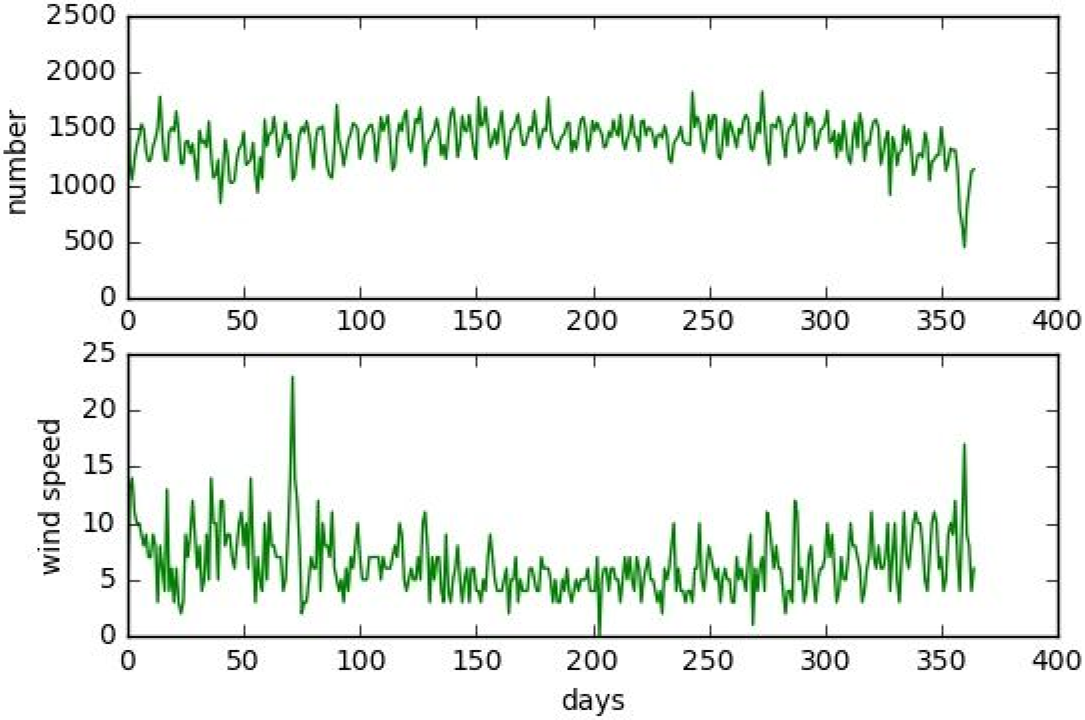
The reason for this observation might be that in general, regions with higher per-capita income have better security. Thus, it is harder to conduct crimes. On the other hand, people living in those regions tend to have higher education, regular job. They have less incentive to conduct crimes. However, since there are many other factors affecting crime rate, sometimes, this observation does not hold.

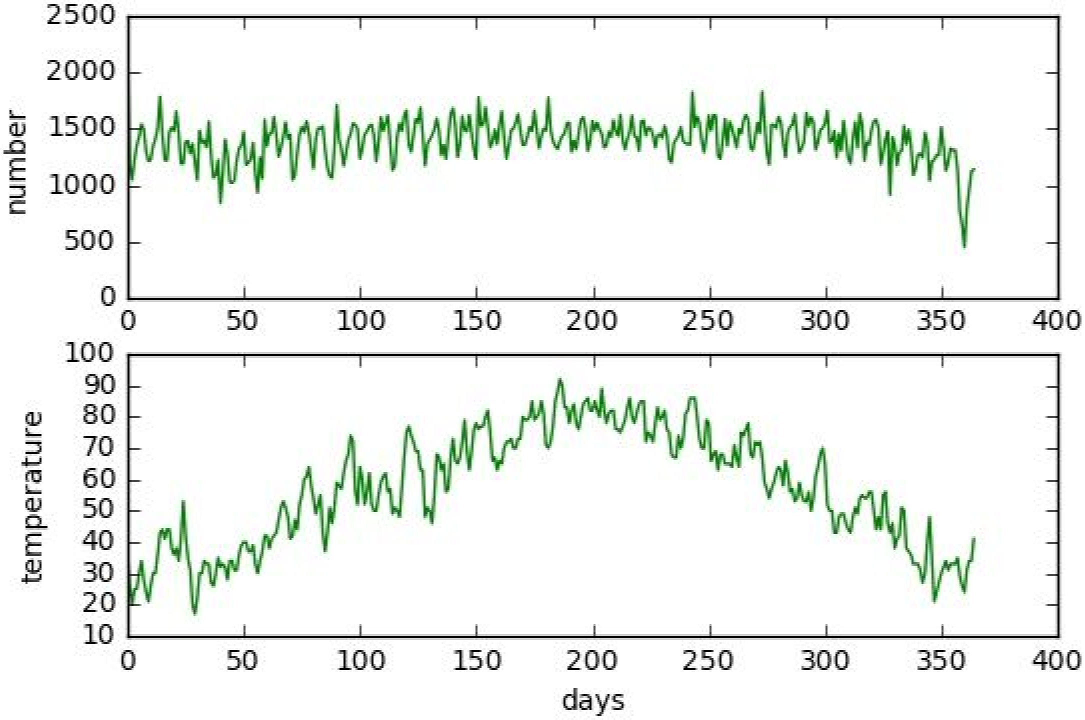
4. Weather and crime

Here we analyzed the correlation between number of crime and weather (temperature and wind speed). We made two hypothesis:

(1) Crime number is correlated with temperature and wind speed.

First, we made two plots between crime number and wind speed or temperature for a year long. For example, for year 2010:


number
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350
400
wind speed
25
20
15
10
5
0
0
50
100
150
200
250
300
350
400
days


number
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350
400
temperature
100
90
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300
350
400
days

From the plot, the correlation is not so obvious. Temperature in every year goes up and then goes down, but the trend of crime number is not obvious, mostly fluctuated. But when we computed the correlation when number of crimes and wind speed or temperature, we found:
The correlation between crime number and temperature for year 2010: 0.439
The correlation between crime number and wind speed for year 2010 : -0.329

The correlation between crime number and temperature is positive, while the correlation between crime number and wind speed is negative. This also explained why the third quarter has the highest number of crimes.

Now, we want to dig deeper and find the correlation between extreme weather and extreme number of crime, which is difficult to observe from the plots.Therefore, we made a second hypothesis here:

(2) Crime number is correlated with extreme weather.

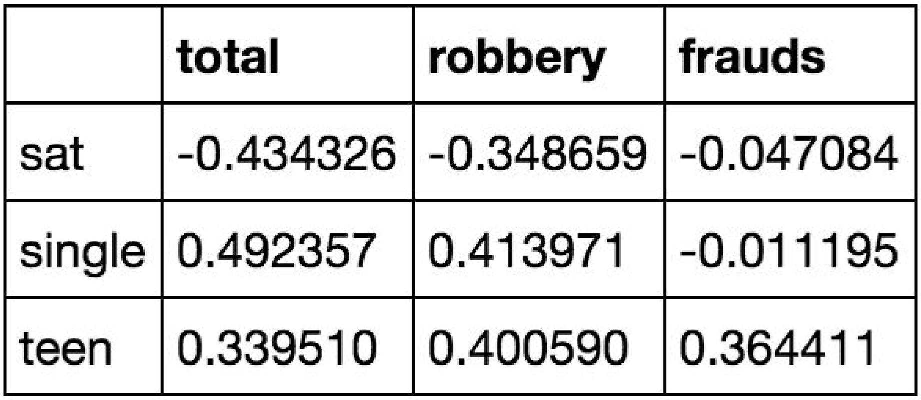
Now we focused on some extreme weather like storm, flood and etc. As we expected, storm or heavy snow strongly influenced crime number. The average crime number is 2378, which is only 70% of usual date. Other extreme weather doesn't strongly influenced number of crimes.

5. Looking into details of crime type:
Different crime types might have different behaviors. Therefore, we continued to investigate details of different crime type. We selected robbery as a representative of violent crime and frauds as a representative of property crime.

We hypothesized that

a. Robbery and frauds will happen much less in regions where people have better education
b. Robbery and frauds will also happen a lot in regions with higher single-mother household ratio.
c. Robbery and frauds tend to happen in regions with higher teenagers ratio.

By calculating correlations, we found the following correlations:



	total	robbery	frauds
sat	-0.434326	-0.348659	-0.047084
single	0.492357	0.413971	-0.011195
teen	0.339510	0.400590	0.364411

For hypothesis a: regions with higher SAT scores indeed have less robbery. However, there is almost no correlation between education and frauds.
For hypothesis b: regions with higher single-mother ratio indeed have more robbery. However, there is no clear correlation between single-mother ratio and frauds.
For hypothesis c: regions with high teenager ratio indeed have more robbery as well as more frauds.

Reasons:
For hypothesis a: regions with higher education tend to have better security. Thus, there is much less robbery.
For hypothesis b: Single-mother are more vulnerable to robbery. Thus, regions with higher single-mother ratio have more robbery.
For hypothesis c: Teenagers are vulnerable to both robbery and frauds. They are relatively weak and less experienced in frauds. Thus, regions with higher teenage ratio have both higher robbery and higher frauds.

Individual Contributions:

Raochuan Fan: Part I: Wrote scripts for column 1 to 8   Part II: Time-base, Education, and Demographical analysis, details of crime type
Haichao Wu: Part I: Wrote scripts for column 9 to 16 Part II: details of crime type, population data and weather data
Yiran Xu: Part I: Wrote scripts for column 17 to 24 Part II: economics-related hypothesis and Time-base analysis, details of crime type

Summary/Conclusion:

For part I: overall, the crime dataset is pretty clean with 8.17% missing values. There are two major data quality issues. One is about time order of crime occurrence or report. In some records, end time or report time are prior to start time. Another one is classification description don't match with classification code.

By data visualization, we found that the distribution of number of crime is a left-skewed distribution with median around 1400. The number of crime records is between 1000-1800 for more than 95% days. When going through the trend of number of crime records, we could see that the number of crime is lowest at the beginning or end of the year and misdemeanor accounts for most of crime, and February has lowest total number of crime. Geographically, Manhattan and Bronx have largest density of crime, and 75 precinct (in the East New York section of Brooklyn) has the most crime records.

For part II:
We made some hypotheses about factors affecting number of crimes and analyzed them by data visualization and computing pairwise correlation. We found data from the following aspects:

1. Time Variant

The distribution of number of crimes has a certain pattern against quarter, month, and day of week, like the third quarter has the greatest number of crimes.

2. Education

We treated each zipcode area as a sample (same as the following steps), and used average SAT score as the indicator of education status. We found that the correlation between the SAT score and number of crime is quite significant.

3. Demographical

We analyzed the correlation of population, population density, ratio of single-mother household, and ratio of teenagers against number of crimes for each zipcode area, and found most of them are significant.

4. Economics

We also found economic factors like unemployment rate and Per-capita income are significantly correlated with number of crimes.

5. Weather

Wind speed and temperature are also correlated with number of crimes.

6. Difference between property crime and violent crime

The last step is to investigate the difference between property crime and violent crime. We used Robbery and Frauds as the example, and found some interesting results.

Data Sources:

1. Unemployment rate: https://www.labor.ny.gov/stats/nyc/
2. Weather data: https://www.wunderground.com/history/airport/KNYC/2017/05/09/DailyHistory.html?req_city=New%20York&req_state=NY&reqdb.zip=10001&reqdb.magic=1&reqdb.wmo=99999
3. Storm Events Database: https://www.ncdc.noaa.gov/stormevents/listevents.jsp?eventType=ALL&beginDate_mm=01&beginDate_dd=01&beginDate_yyyy=2010&endDate_mm=01&endDate_dd=31&endDate_yyyy=2016&county=NEW%2BYORK%3A61&hailfilter=0.00&tornfilter=0&windfilter=000&sort=DT&submitbutton=Search&statefips=36%2CNEW+YORK
4. Income by zip:https://www.incomebyzipcode.com/
5. NYC 2010 census data: http://www1.nyc.gov/site/planning/data-maps/nyc-population/census-2010.page?tab=1
6. List of NYCHA housing development as of 1/1/2016 (close to date of our data records) are obtained from: https://catalog.data.gov/dataset/nycha-development-data-book-9e35e
7. Bounds for X_COORD_CD, Y_COORD_CD, Latitude and Longitude are obtained from: http://www.spatialreference.org/ref/epsg/3628/
8. References for precinct code is obtained from http://www.nyc.gov/html/nypd/html/precincts/precinct_075.shtml and http://www.nyc.gov/html/nypd/html/precincts/precinct_043.shtml

References:

1. Ajimotokin, Sandra, Alexandra Haskins, and Zach Wade. "The Effects of Unemployment on Crime Rates in the U.S." (n.d.): n. pag. Https://smartech.gatech.edu/bitstream/handle/1853/53294/theeffectsofunemploymentoncimerates.pdf. 14 Apr. 2015. Web. 10 May 2017.