New York University

Center for Data Science

DS1004 Big Data Term Project

Crime Analysis in New York City 2006-2015

Yuwei Tu, Zizhuo Ren, Lingshan Gao

05/01/2017

1 Introduction

This report will provide an analysis on the NYC crime incidents that are reported during the time of 2006 to 2015 based on the dataset that is downloaded from NYC Open Data [1]. The first part of the report summarizes the data quality issues as well as solutions to the challenges that these issues may pose on data handling. The second part of the report focuses on data analysis, hypothesis testing and data visualizations on the trends in crime activities and interesting findings from the data.

2 Tools and Individual Contribution

We leveraged big data tools such as Spark and SQL in data analysis and data visualization. Besides, we also employed Plotly and GIS graphic tools to generate the graphics throughout the report. The setup for Spark has been specified to 8G. The amount of off-heap memory to be allocated per executor is set to 2G.

In terms of the individual contribution, as the coordinator of the project, Yuwei Tu has taken the initiative and distributed work according the members' expertise. Besides making sure the group stays on track, she has also set up the framework of the project, contributed to half of the methods in checking columns types in part one, set up Spark environment and SQL graphical tool, performed all the analysis and tests in the temporal aspect, from cleaning data, writing scripts and plotting charts.

Zizhuo Ren, as the lead engineer of the group, has improved and refined the programs of the group to make sure that everyone's script is in consistence with the others'. He has also single-handedly configured all the geographical maps and plots in both Part I and Part II. To optimized the work process, he has discovered and employed KDTree method in assigning each crime incident to a zip code area.

Lingshan Gao has contributed to the other half of the methods in checking columns types in part one. She has cleaned the dataset and collected and gathered demographical data for geographical analysis and hypothesis testing. She is also in charge of plot interpretation and report writeup.

[1] The dataset can be retrieved from https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

3 Data Summary

3.1 Data Type

The table below includes the total count of base type, semantic type and validity of each column in the original dataset.

	Base Type Count	**Semantic Type Count**	**Validity Count**	**Unique Values Count**	**Definition for a Valid Input**
CMPLNT_NUM	INT: 5101231	CASE ID: 5101231	VALID: 5101231	5,101,231	Unique integer
CMPLNT_FR_DT	DateTime: 5100576, TEXT: 655	Start Date: 5101231	VALID: 5100576, INVALID: 9 Null: 655	6,371	A legal date from 1900/01/01 to 12/31/2015
CMPLNT_FR_TM	DateTime: 5100280, TEXT: 951	Start Time: 5101231	INVALID: 903, VALID: 5100280, Null: 48	1,442	Legal 24-hour time format, 00:00:00-23:59:59
CMPLNT_TO_DT	DateTime: 3709753, TEXT: 1391478	End Date: 5101231	VALID: 3709753, INVALID: 9, Null: 1391478	4,827	A legal date from 1900/01/01 to 12/31/2015
CMPLNT_TO_TM	DateTime: 3712070, TEXT: 1389161	End Time: 5101231	INVALID: 1376, VALID: 3712070, Null: 1387785	1,441	Legal 24-hour time format, 00:00:00-23:59:59
RPT_DT	DateTime: 5101231	Report Date: 5101231	VALID: 5101231	3,652	A legal date from 2006/01/01 to 12/31/2015
KY_CD	INT: 5101231	KY ID: 5101231	VALID: 5101231	74	3-digit integer
OFNS_DESC	TEXT: 5101231	KY DESCRIPTION: 5101231	VALID: 5082391, Null: 18840	70	String (Consistency issues are discussed later in the report)
PD_CD	TEXT: 4574, INT: 5096657	PD ID: 5101231	VALID: 5096657, Null: 4574	415	String with a float equals zero. Can be legally convert to 3-digit integers. Code < 123 are legal precinct codes for NYC

PD_DESC	TEXT: 5101231	PD DESCRIPTION: 5101231	VALID: 5096657, Null: 4574	403	String (Consistency issues are discussed later in the report)
CRM_ATPT_CPTD_CD	TEXT: 5101231	LEVEL: 5101231	VALID: 5101224, Null: 7	2	Either "COMPLETED" or "ATTEMPTED"
LAW_CAT_CD	TEXT: 5101231	LAW CATEGORY: 5101231	VALID: 5101231	3	One of the three: "MISDEMEANOR", "FELONY" or "VIOLATION"
JURIS_DESC	TEXT: 5101231	JURIS DESCRIPTION: 5101231	VALID: 5101231	25	String
BORO_NM	TEXT: 5101231	BOROUGH: 5101231	VALID: 5100768, Null: 463	5	One of the five: 'BRONX', 'QUEENS', 'MANHATTAN', 'BROOKLYN', 'STATEN ISLAND'
ADDR_PCT_CD	TEXT: 390, INT: 5100841	PRECINCT: 5101231	VALID: 5100841, Null: 390	77	String with a float equals zero. Can be legally convert to 3-digit integers.
LOC_OF_OCCUR_DESC	TEXT: 5101231	LOCATION: 5101231	INVALID: 213, VALID: 3973890, Null: 1127128	6	One of the five: 'INSIDE', 'OUTSIDE','FRONT OF', 'OPPOSITE OF', 'REAR OF'
PREM_TYP_DESC	TEXT: 5101231	PREMISE: 5101231	VALID: 5067952, Null: 33279	70	String type
PARKS_NM	TEXT: 5101231	PARK: 5101231	VALID: 7599, Null: 5093632	863	String type
HADEVELOPT	TEXT: 5101231	HOUSE: 5101231	VALID: 253205, Null: 4848026	278	String type
X_COORD_CD	TEXT: 188146, INT: 4913085	X COORDINATE: 5101231	VALID: 4913085, Null: 188146	69,532	Legal x-coordinate in NYC
Y_COORD_CD	TEXT: 188146, INT: 4913085	Y COORDINATE: 5101231	VALID: 4913085, Null: 188146	72,316	Legal y-coordinate in NYC
Latitude	FLOAT: 4913085, TEXT: 188146	LAT COORDINATE: 5101231	VALID: 4913085, Null: 188146	112,803	Legal latitude in NYC
Longitude	TEXT: 5101231	LON COORDINATE: 5101231	VALID: 4913085, Null: 188146	112,807	Legal longitude in NYC
Lat_Lon	TEXT: 5101231	LAT LON COORDINATE:	VALID: 4913085,	112,826	In consistent with column "Latitude" and "Longtitude"

3.2 Data Quality Issues and Solutions

3.2.1 Missing Data

We discovered missing data in most features. Some missing data are simply based off facts. For example, if an crime incident happens an area that is not a park, then "PARKS NM" should be left blank. This type of missing values contains meaningful information. However, the missing data in date and location are the more concerning and less easier to handle. This kind of missing values are truly missing information that we are not able to retrieve elsewhere, nor it reflects any truthful facts that help

data analysis. Therefore, we exclude instances that has missing date in "CMPLNT FR DT" and "Lat Lon" while analyzing the time-series and geolocation distribution of the crime events.

3.2.2 Inconsistency

Inconsistency is one of the biggest data quality issues that we have discovered during the process of exploratory data analysis. For example, one would expect columns "CMPLNT FR DT" (complaint from date) and "CMPLNT FR TM" (complaint from time) to both have values or both are null. As have listed in the table, the total number of valid entries for time are even more than the valid entries for dates, which means that we may have some entries that only have a "complaint to" timestamp without the "complaint to" date. For analysis that require both date and time, one way to handle the issue is to keep only the entries that have valid values in both columns.

Another inconsistency issue we have identified is primarily in the "code" and "description" relationship. In the data set, for each "code" column, it is usually followed a corresponding description column to explain what the code represents. Ideally, for each "KY CD" (key code), we would expect to find the description in "OFNS DESC" that uniquely corresponds to the key code. However, we find out that for the same code, there can be multiple different descriptions that are employed. For example, for entries that equal "343" in the "KY CD" column, there are two different descriptions found in "OFNS DESC", which are "other offenses related to theft" and "theft of services". In this case, because both descriptions are related to theft, we can simply replace the description to "theft."

Moreover, for the same description, there could be multiple different codes that correspond to it. For example, both "235" and "117" share the same description "dangerous drugs", and both "236" and "118" have the same description "dangerous weapons." An easy way to solve this issue is to merge these values, such as change all "117" to "235".

4 Data Analysis

After cleaning and preparing the dataset for the analysis, we will observe the trends in both chronological and geographical order without making assumptions. For this analysis, we only selected data between 2006 to 2015 based on the "Complaint Start Date."

4.1 Trends in Time-Series Distribution

4.1.1 Yearly Trend

We summed up all the crime incidents and plotted them against year. The data has shown an obvious trend that the total number of crime events has decreased sharply in the past decade. According to the chart, one may conclude that NYC has become a safer place for living within the last ten years.

4.1.2 Monthly Trend

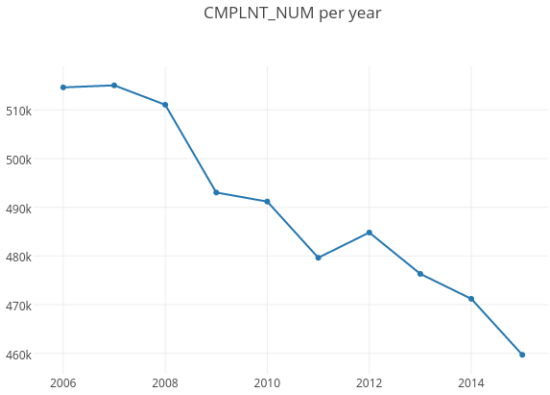
The following chart shows that the total number of crimes distributed over twelve months. The significant dip in February may lead (or mislead) to an easy conclusion that February is the safest month in a year. One may also argue that February is the shortest month among all. But even if we adjust the number of days in a month and use the average instead of total number of crimes, the chart still shows that crime events are more active during summertime.

4.1.3 Daily Trend

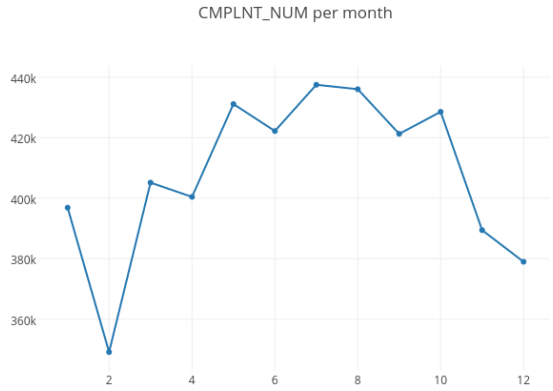
The chart has shown an very interesting pattern that many crimes happened started on the first day of the month and very few incidents started on the last day of the month. It can be due to that when the victims report to the police, habitually people like to describe an event that lasts for a while with "starting from the beginning of the month." However, the interesting pattern is worth being investigated.

4.1.4 Hourly Trend

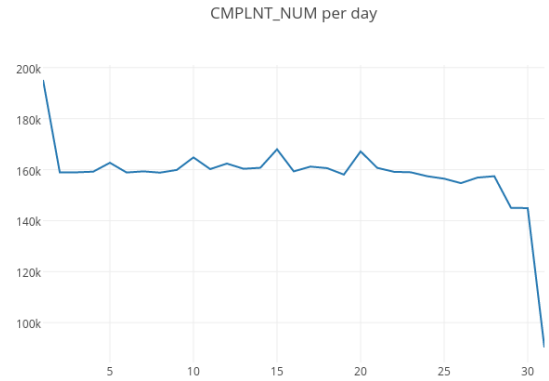
Although it is commonly perceived that crimes are more likely to happen at night. The data tells us a quite different story. According to the data, in the past decade, more crimes happened in the late afternoon to evening, peaking at 3pm to 8pm.



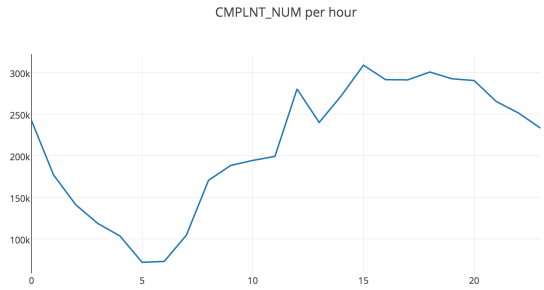
(a) Yearly



(b) Monthly



(c) Daily

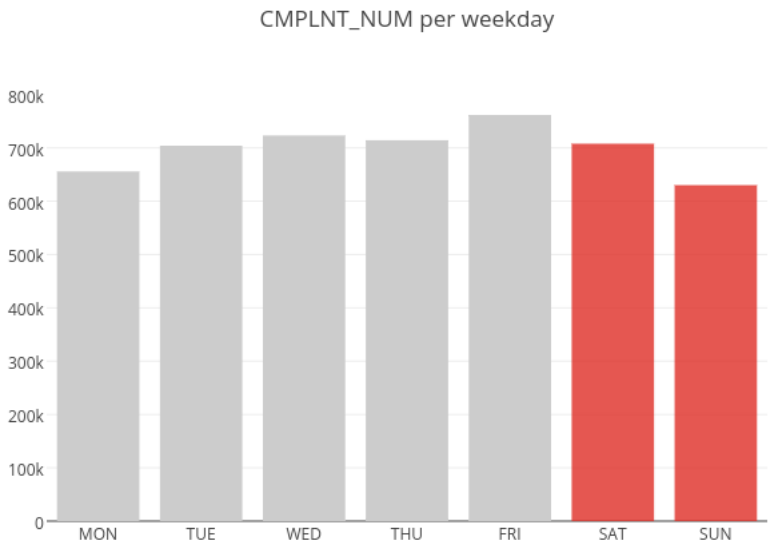


(d) Hourly

Figure 1: Trend in Total Number of Crime by Time

4.1.5 Weekday/Weekend

We are also curious if crime patterns are any different between weekdays and weekends. As shown in the graph below, the total number of crimes that have happened on a weekend is highlighted in coral. At a glance, the number of crimes happened less on weekends than during weekdays, but to validate our assumptions, we conducted the T-test on the difference between weekday and weekend using yearly total. In the table below, "Weekday" denotes the average number of crimes happened on a weekday and similarly, "Weekend" represents the average number of crimes happened on a weekend. It turns out that crime incidents are more likely to happen on weekdays than on weekends at a 95% confidence level.



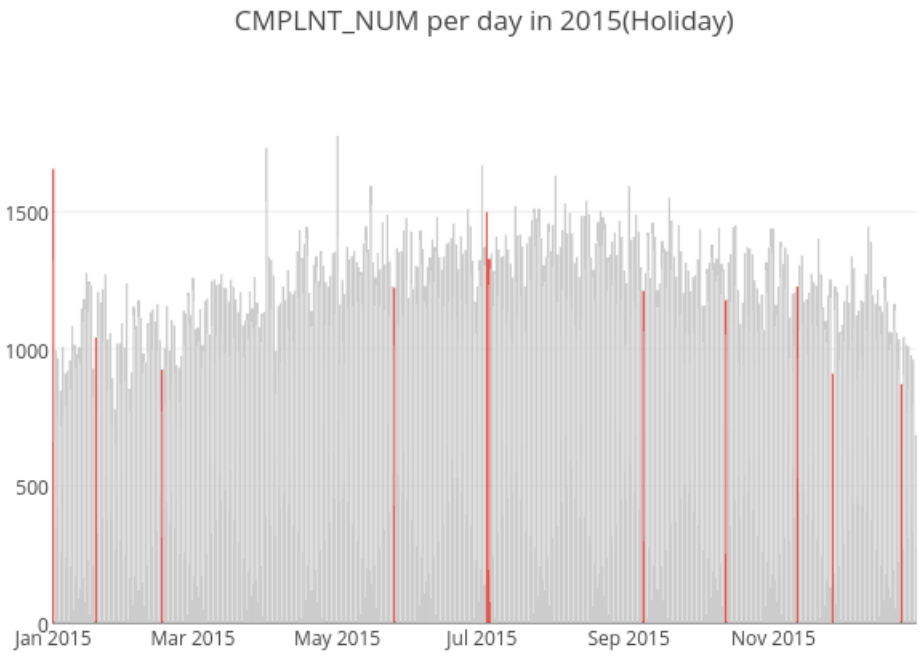
YEAR	Weekday	Weekend
2006	75622.8	68250.5
2007	75412.0	68988.0
2008	74683.4	68817.0
2009	71836.6	66924.0
2010	71628.0	66519.0
2011	69899.0	65077.0
2012	69900.0	67661.0
2013	68795.0	66179.5
2014	67716.4	66315.5
2015	66270.8	64183.0

T-TEST: P-value = 0.003, which means the mean number of crimes during weekday/weekend are significant different in 95% confidence level.

4.1.6 Holidays

Similar to the difference between weekdays and weekends, public holidays (excluding weekends) may also have an impact on criminal activities. To better illustrate and visualize the trend, we used data from the most recent year to plot the following chart. Here we conducted a T-test with a null hypothesis that crime incidents are equally to happen on a holiday or a non-holiday. The T-test has shown that with a 99% level of confidence, a criminal event will happen on a non-holiday. Either that the criminals also take national holidays off, or that people are more obedient to

the law during holidays.

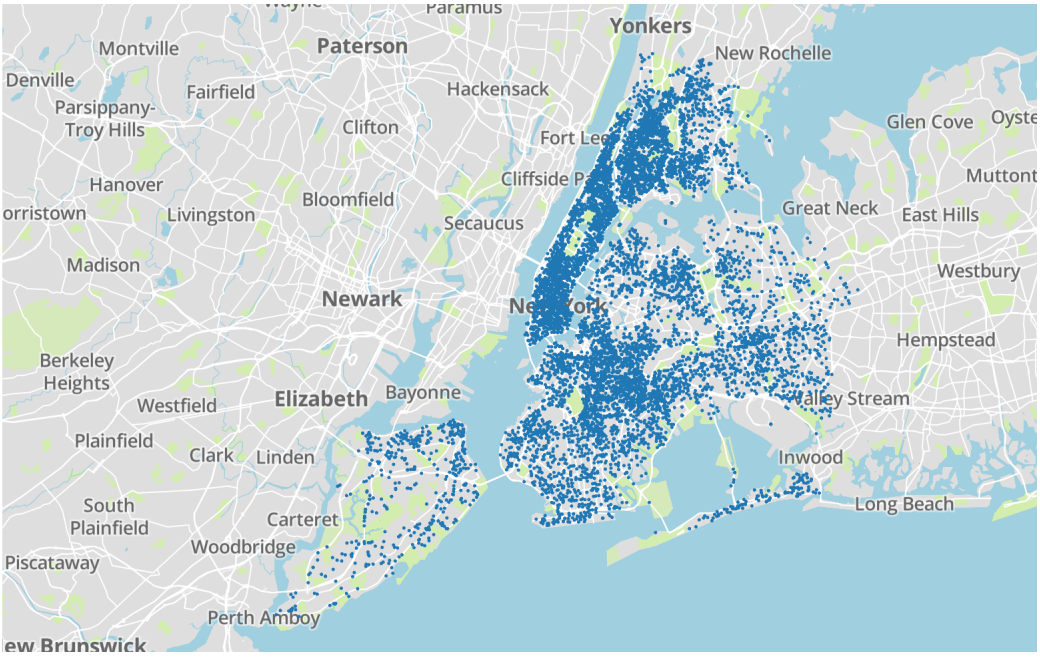


YEAR	Holiday	Workday
2006	1414.0	1290.8
2007	1415.8	1259.5
2008	1400.1	1260.3
2009	1354.7	1225.0
2010	1351.0	1201.8
2011	1317.4	1208.0
2012	1329.4	1183.8
2013	1308.4	1186.0
2014	1292.4	1242.4
2015	1261.7	1187.8

T-TEST: P-value < 0.00001, which means the mean number of crimes during weekday/weekend are significant different in 99% confidence level.

4.2 Trends in Geographical Distribution

To under the crimes distribution across different boroughs in NYC, it is helpful to visualize the crimes on a map according to the location where the crime happened. Drawing everything on the map can get very messy and thus counter the purpose of visualization. Thus, we randomly sampled 10,000 data points and plotted them by their latitude and longitude. As shown below, the distribution is relatively dense in Manhattan and Bronx comparing to other boroughs in NYC.

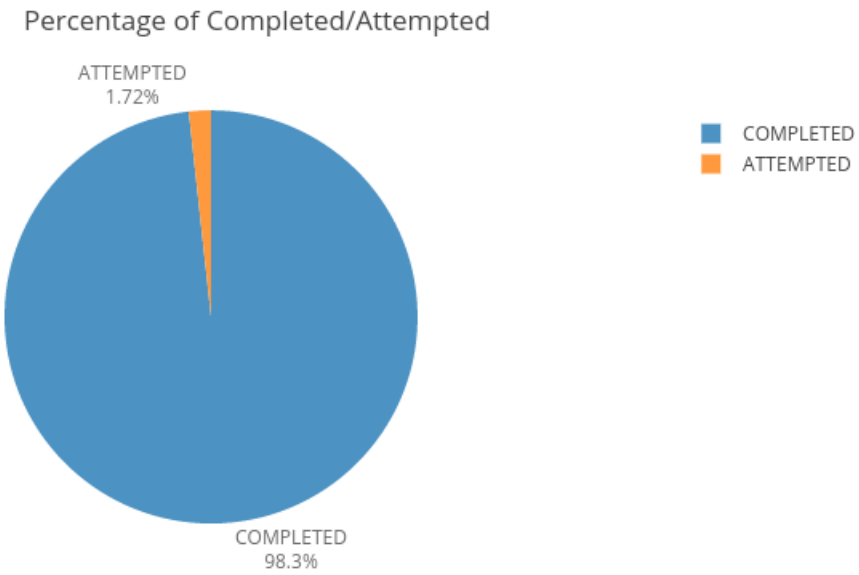


4.3 Distribution by Crime Type

Besides the trends in time and location, the dataset also includes other information that enables us to analyze the criminal activities by different types.

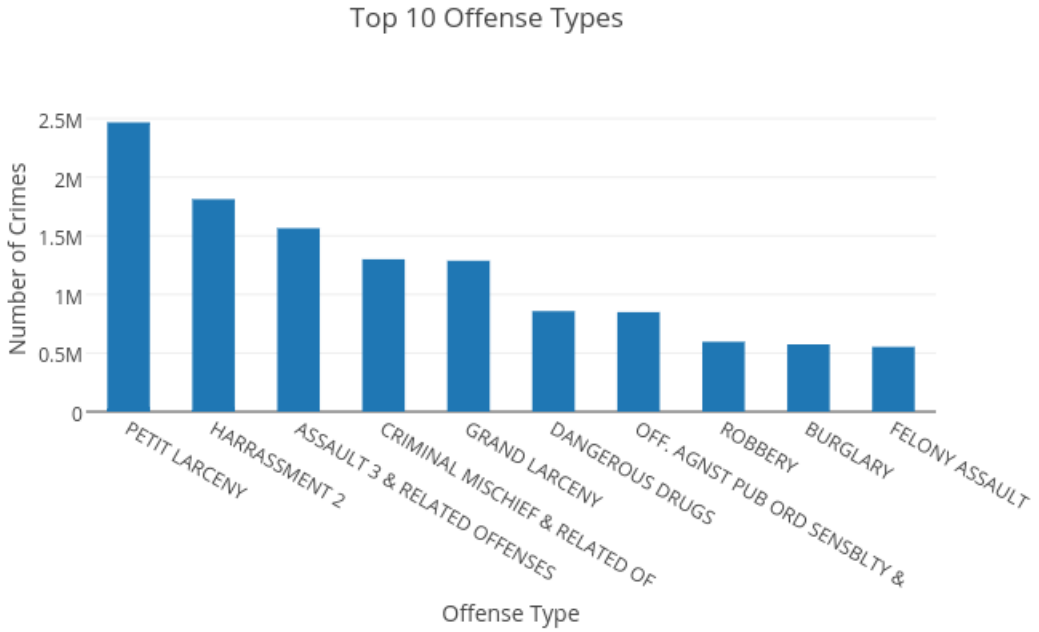
4.3.1 Completed/Attempted

The chart below tells us about 98% attempted crimes have succeeded. Although in reality, it can involve some selection bias as people more likely to report the completed incidents than unsuccessful attempts.



4.3.2 Offense Types

As shown below, the most common type of crime in NYC is larceny, followed by second degree harassment and third degree assault.



5 Hypothesis Tests

So far we have conducted analysis on the crime data and performed T-tests to quantify the significance of the events without making many assumptions. To further explore the trends in the data, we would like to incorporate data from other sources, and employ various statistical methods, such as finding correlations, to validate our assumptions on the crime data. For the purpose of consistency, the assumptions are also made based on time and location.

5.1 Time-dimension Hypothesis Tests

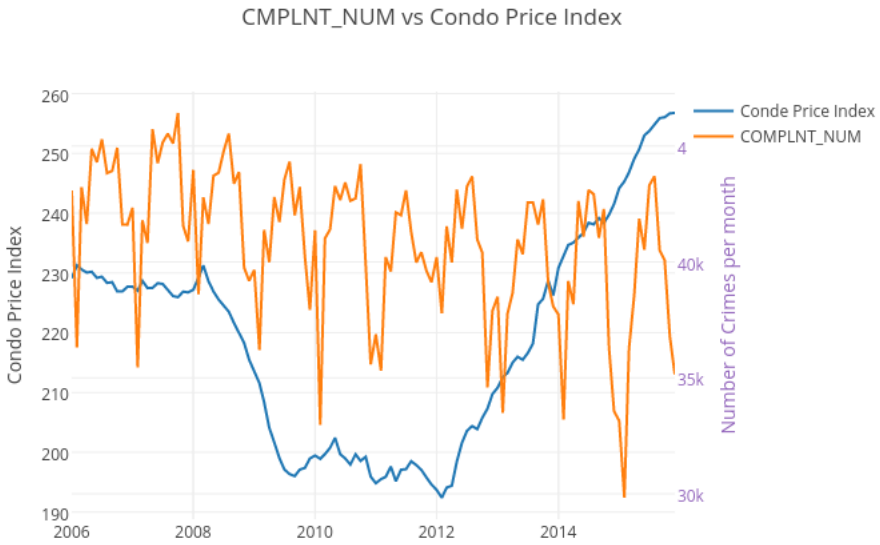
In this part, the majority of external data sources are related to economy. We believe that economy has a strong lagged impact on social stability and municipal security. We will propose hypotheses on the relationship between economy and crimes over-time, and apply statistical method to validate or reject the null hypotheses.

5.1.1 Condo Price Index

1. Source
   Crimes are greatly related to the economic development level for a certain area at a given time. Condo price index[2] represents the economy of a region at large.

2. Trend Plot
   Below we plotted the monthly condo price index for NYC and the total monthly complaints reports per month for the year of 2006 to 2015.

[2]S&P Dow Jones Indices LLC, Condo Price Index for New York, New York© [NYXRCSA], retrieved from FRED, Federal Reserve Bank of St. Louis;
https://fred.stlouisfed.org/series/NYXRCSA, May 3, 2017.



3. Hypothesis
   The hypothesis is that there is a negative relationship between condo price index and number of crimes. Higher condo price index usually indicates economic growth, and thus, a better life quality and higher consumer capability. So the number of crimes should decrease.

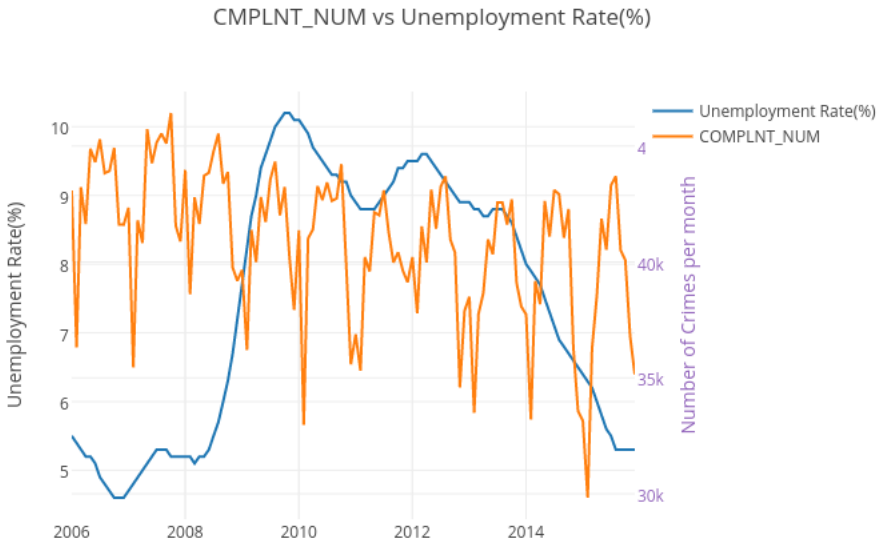
4. Test
   It turns out that the correlation coefficient between number of crimes per month and condo price index per month is -0.055. The correlation coefficient between total number of crimes per year and average condo price index per year is -0.158. Although it does seem to confirm our assumption on the negative direction, the relationship is insignificant.

5.1.2 Unemployment Rate

1. Source
   Another important economic indicator that cannot be overlooked is unemployment rate. We obtained seasonally adjusted NYC unemployment rate for the past decade[3].

2. Trend Plot

[3]New York City Labor Force Data(Seasonally Adjusted Data for Model-Based Methodology) https://www.bls.gov/cps/tables.htm



3. Hypothesis
   We assume that the relationship between unemployment rate and number of crimes is positive. Higher unemployment rate means less disposable income and more instable factors for the society. So the number of crimes should increase.

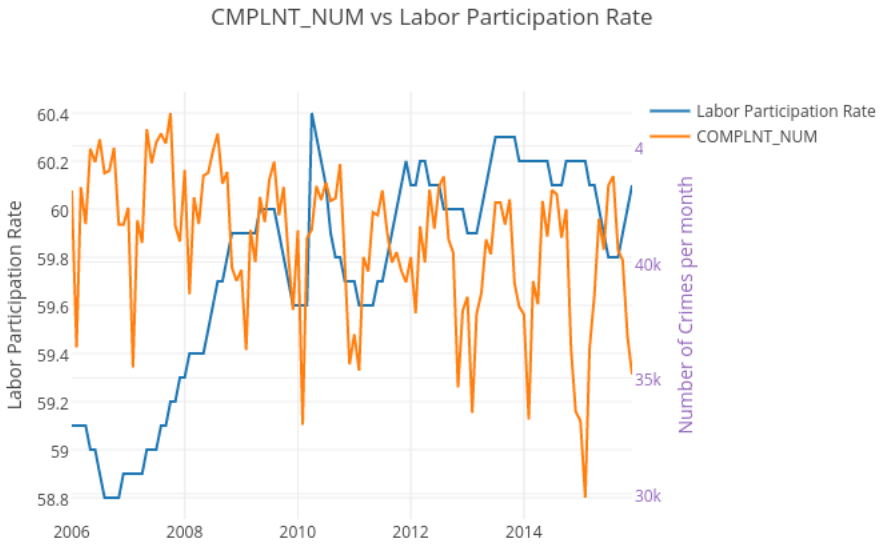
4. Test
   The correlation coefficient between number of crimes per month and monthly unemployment rate is -0.194. the correlation coefficient between total number of crimes per year and average unemployment rate per year is -0.402, which means they have insignificant negative relationship.

5.1.3 Labor Participation Rate

1. Source
   Besides unemployment rate, label force participation rate[4] is also crucial in measuring the stability of a society.

2. Trend Plot

[4]New York City Labor Force Data(Seasonally Adjusted Data for Model-Based Methodology) https://www.bls.gov/cps/tables.htm



3. Hypothesis
   We hypothesize that there is a negative relationship between labor force participation rate and number of crimes. Higher labor force participation rate indicates stronger consumer confidence in the market, which often leads to a boost in economy. Therefore, the number of crimes should decrease.

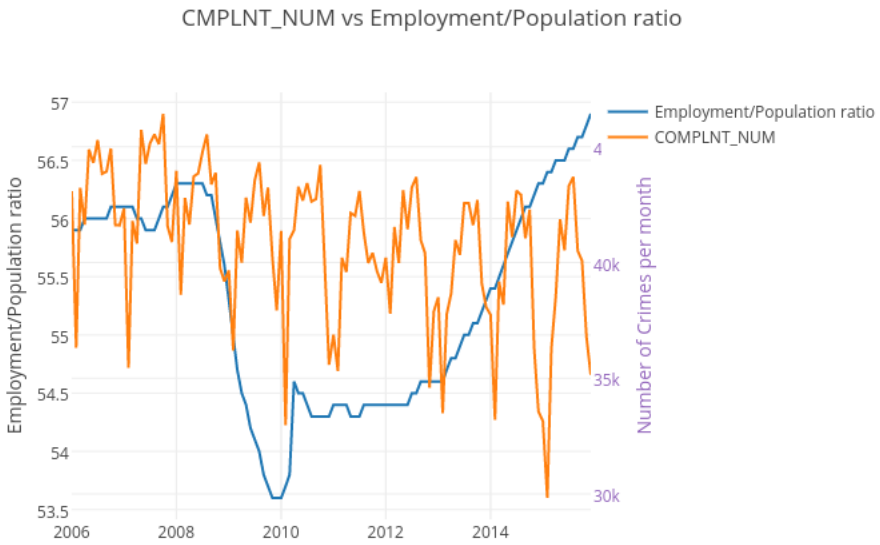
4. Test
   The correlation coefficient between number of crimes per month and labor force participation rate per month is 0.077. The correlation coefficient between total number of crimes per year and average labor force Participation Rate per year is -0.844, which means at a larger scale, they may have a significantly negative relationship.

5.1.4 Employment/Population Ratio

1. Source
   Another measurement for employment is using the actual number of people employed over the whole population rather than the population in the labor force. Thus we will plot the employment/population ratio[5] and the total number of complaints and discover any intrinsic associations they have.

[5]New York City Labor Force Data(Seasonally Adjusted Data for Model-Based Methodology) https://www.bls.gov/cps/tables.htm

2. Trend Plot



3. Hypothesis
   We suppose that the relationship between employment/population ratio and number of crimes is negative. Higher employment/population ratio means more disposable income and thus a more stable society. Logically, the number of crimes should decrease.

4. Test
   The correlation coefficient between number of crimes per month and employment/population ratio per month is -0.351, the correlation coefficient between total number of crimes per year and average employment/population ratio per year is 0.122, which means there is no significant positive/negative relationship.

5.2 Location-dimension Hypothesis Test

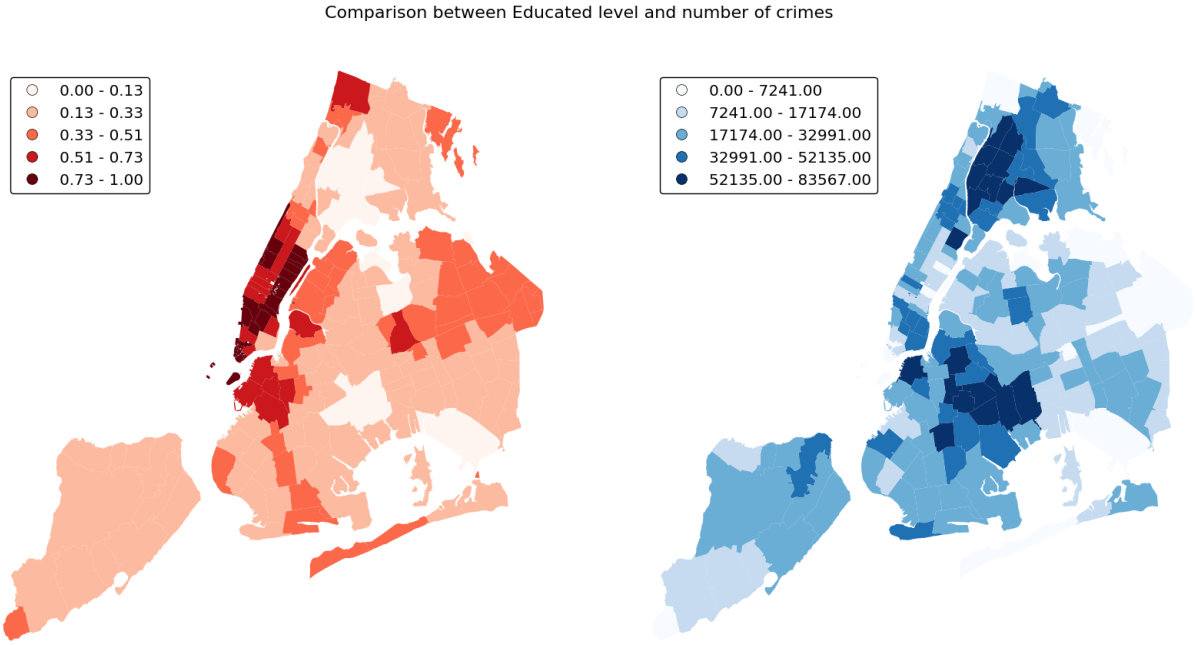
To better understand the variance across regions in NYC using different crime metrics, we assign each incident by its latitude and longitude to the corresponding zip code area. The k-dimensional tree algorithm helps finishing the task with a high accuracy. The American Community Survey[6] estimated demographic data for the year of 2015 is the main contributor to the following hypothesis tests.

[6]Educational Attainment, 2011-2015 American Community Survey
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

5.2.1 Education

1. Source
   Level of education is calculated as the percent of population that holds a bachelor's degree or higher in each zip code area. P(Education Level) = sum of (bachelor's degree + master's degree + professional school degree + doctorate degree) holders / total population for each region. With this method, 1 represents that 100% of the population in the area has a bachelor's degree and higher and 0 means that none has bachelor's degree or higher in the area.

2. Trend Plot
   The map on the left shows the distribution of educated level of the population in each zip code area across NYC. The more highly educated area, the deeper the color is. The map on the right is the distribution of the total number of crimes in each zip code area. Darker blue means a more dangerous area while lighter blue means a less dangerous area.



3. Hypothesis
   Along with increase of education level, the number of crimes will decrease. The hypothesis is derived from comparing the two maps. In the areas that have fairly high education level, such as the upper east side and downtown in the left figure, the corresponding area in the right has a small number of crimes. In the area which has a low education level, take Bronx for example. The level of crimes is at the highest.

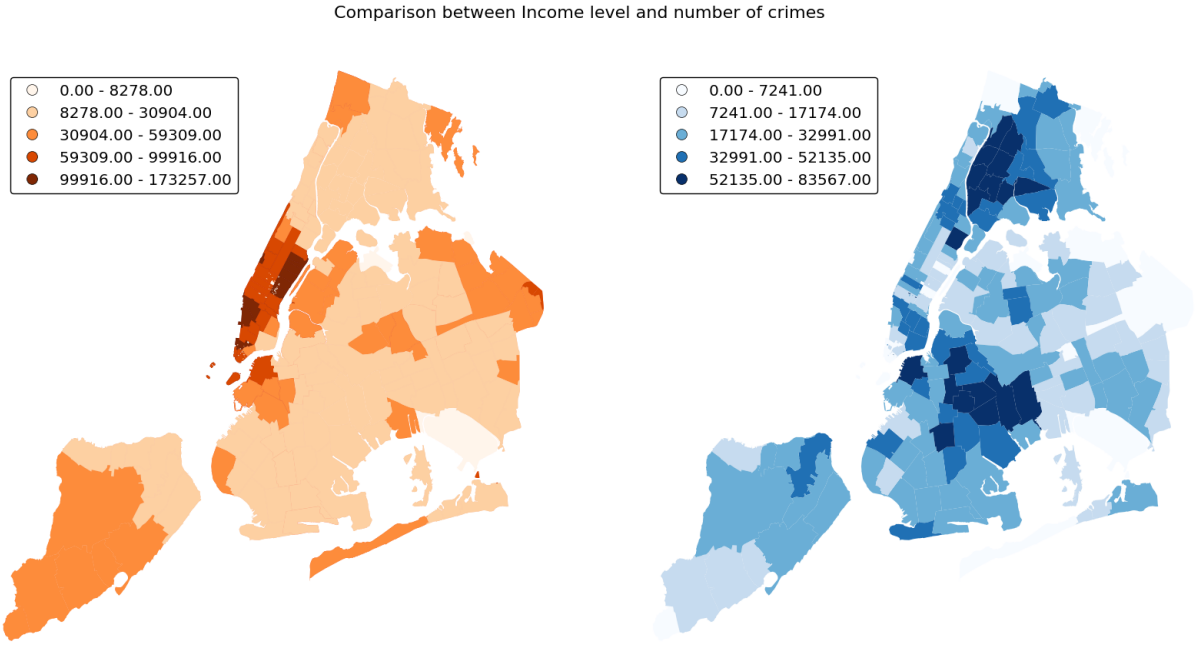
4. Test

For these areas, there are clear negative relationships. However, overall the correlation coefficient of the two variables is -0.136, which is not significant.

5.2.2 Income

1. Source
   We take the median household income[7] for the NYC area and compare them to crime distribution by zip code.

2. Trend Plot
   The map on the left depicts the median household income distribution in NYC and the map on the right is still the crime heatmap.



3. Hypothesis
   We propose that the higher median household income is in an area, the less crimes happen in the area. By simply eyeballing the two maps, we can clearly see that part of Bronx and the areas that connect Queens and Brooklyn has the darkest blue shades in the crime map, and they both correspond to a light orange color on the income map, which means a lower household income.

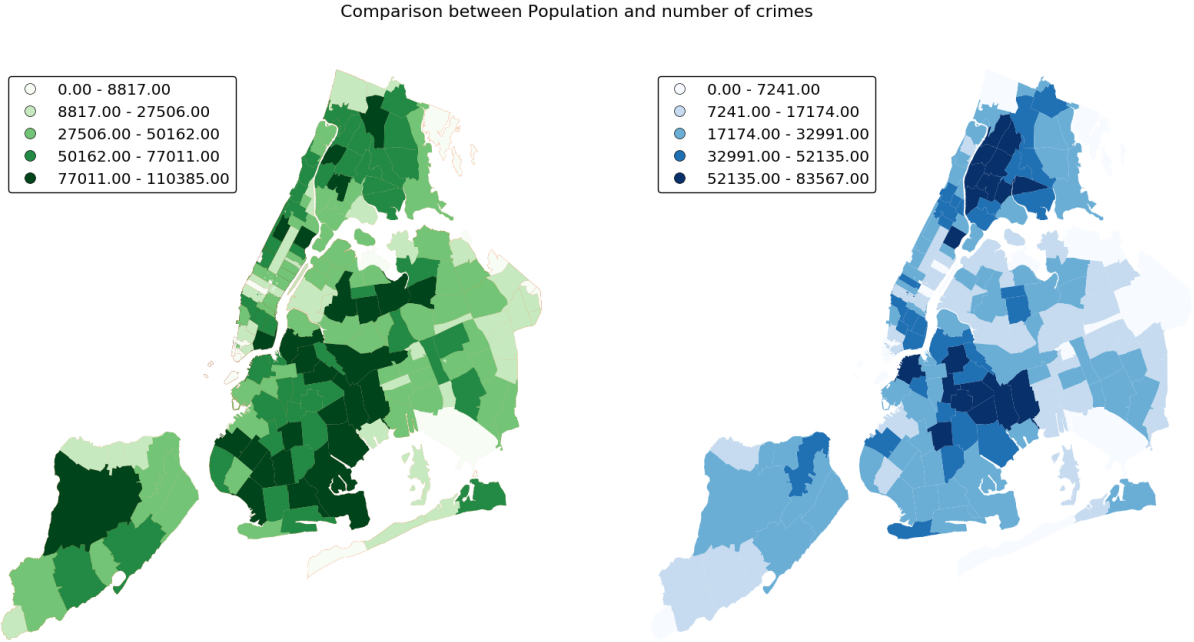
4. Test
   The correlation coefficient of the two variables is around -0.167. Although it has a negative sign as expected, the relationship is not strong enough to draw a conclusion.

[7]Median Household Income, 2011-2015 American Community Survey
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

5.2.3 Population

1. Source Population is an essential and easy to find source that can help us to understand the demographical component of an area. we collected population[8] data that is summarized by zip code in NYC.

2. Trend Plot
   The map on the left is the population distribution by zip code. Darker green means a larger number of total population and a smaller number indicates a smaller number of total population. Again, on the right is the crime map.



3. Hypothesis
   Generally, the population seems to be a good indicator to the number of crimes. The higher population is in a certain area, the more targets that a criminal can have, and thus, more likely that a crime happens in that area. We hence come up with a hypothesis that the population is positive correlated to the number of crimes.

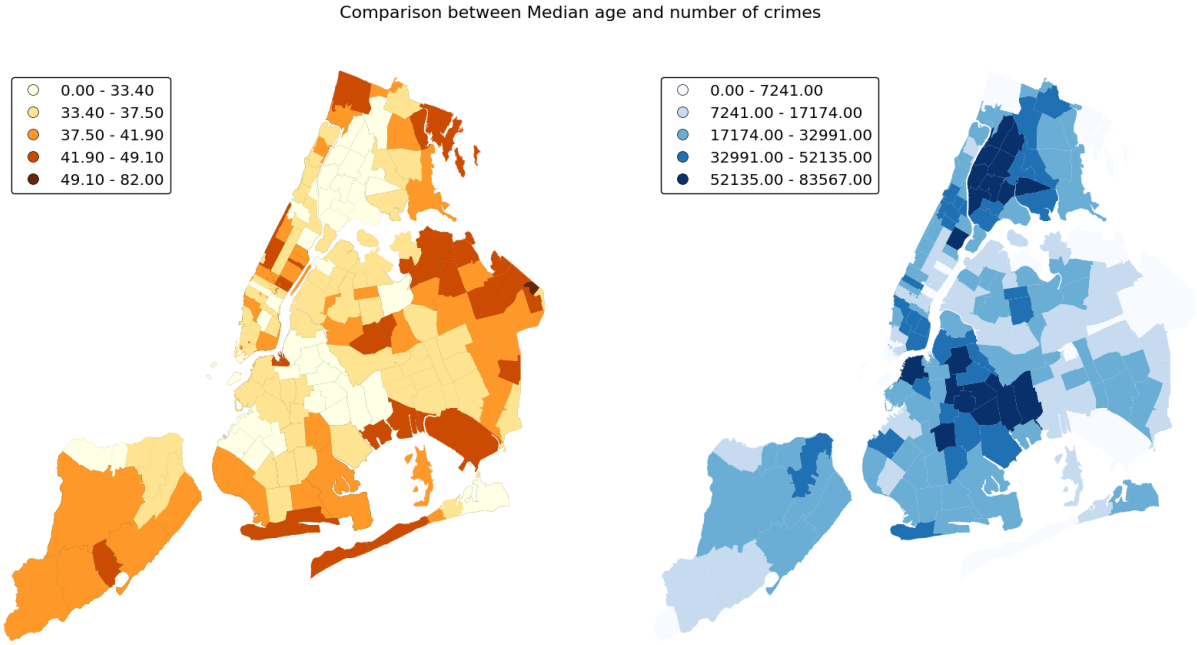
4. Test
   We calculate the correlation between population and the number of crimes. The correlation between these two variables is 0.693 which backs up our hypothesis on the positive correlation.

[8]Population, 2011-2015 American Community Survey
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

5.2.4 Age

1. Source
   We collected median age[9] of the population in each zip code area and analyze the relationship between age and crime by zip code areas.

2. Trend Plot
   The map on the left represents median age in each area, and the map on the right is the crime map.



3. Hypothesis
   The younger people are in a certain area, the more likely crimes happen in that area. In the figure, we can easily find out that areas with more crimes are likely to have a younger median age. Taking Bronx area again for example, the median age of this area is in the range of 0-33 which is the youngest age level and the corresponding total number of crime is also relatively large comparing to other areas.

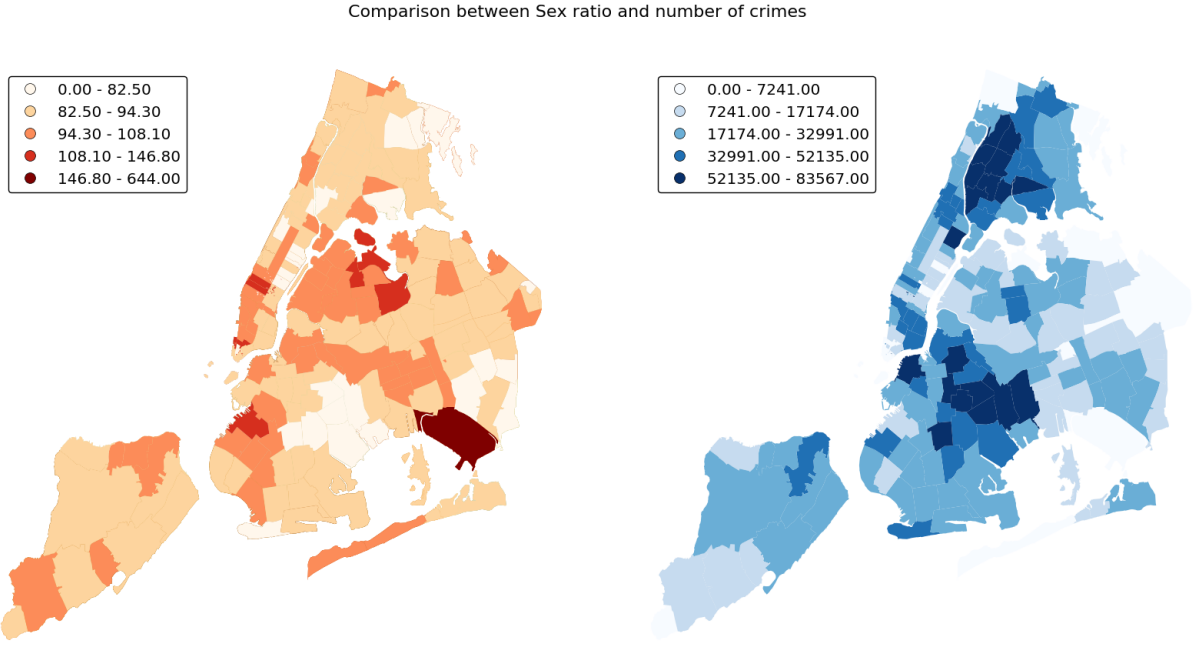
4. Test
   we calculate the correlation between Median age and the total number of crimes. The correlation coefficient is -0.476333, which means median age, which is negatively associated with crime, is a fairly good indicator.

[9] Age and Sex 2011-2015 American Community Survey
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

5.2.5 Sex Ratio

1. Source
   we wonder if gender plays an role in predicting crime incidents. We retrieved the sex ratio data (males per 100 females)[10] for analysis.

2. Trend Plot
   The graph on the left is the sex ratio map. Darker color indicates a lower female percentage in the population decomposition, whereas light color indicates a higher total number of female.



3. Hypothesis
   Our hypothesis is that crimes tend to happen more often in the areas that have a lower sex ratio (more female than male). Stereotypically, females are more prone to be targeted for criminal activities, primarily in sexual harassment, which is ranked number three most frequent crime type among all. By eyeballing the two maps above, it is hard to identify a clear relationship between sex ratio and number of crimes.

4. Test
   The correlation between these two variables is -0.135. Although the negative sign is in line with our expectation, it is not significant enough to validate our hypothesis.

[10] Age and Sex 2011-2015 American Community Survey
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml

6 Conclusion

In this analysis, the most common issues we encountered during data cleaning were missing data and inconsistency. The filtering and clustering method helped resolve some of the issues in the data. We leveraged the big data tool Spark and data management tool SQL to clean the dataset in a reproducible and efficient manner. For analysis, we employed python Plotly package and GIS graphing to help visualize the trends in criminal activities across both time and regions. Over the last decade, the total number of crime had decreased and weekends appeared to have more active criminal activities than other days of the week.

In the end, we referenced various other data sources, from labor statistics to Census data, to testify our hypothesis on the trends. By conducting T-tests and correlation analysis, labor participation rate was significantly and negatively associated with crime activities over time, whereas income and median age were strong indicators to differentiate safety levels across zip code areas.