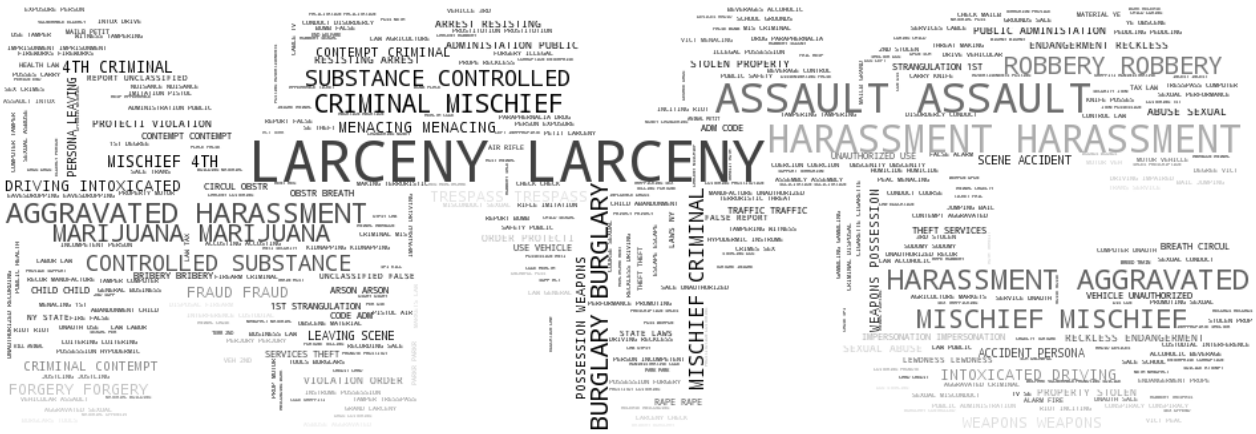
DS-GA 1004 Final Report by Team Sherlock

Weitao Lin, Shangying Jiang, Xue Yang
netid: wl1599, sj2384, xy990

May 10, 2017



Abstract

In this report we present our findings regrading the NYPD Crime Dataset. We first do some data quality analysis for each column in the dataset, including finding the number of unique values, missing values, what's the data type, etc. We break down **Section 1** into several subsections where we give more details about how we do the data cleaning process. Then we give a list of data quality issues in the last subsection. In **Section 2** we do some data exploration on the original crime dataset and other datasets we incorporate. To do this, we propose some hypothesis and do some experiments to test them. Finally we give some conclusions in **Section 3**, followed by the contribution of each team member.

Introduction

Our main goal is to gain some knowledge about the crime happening in New York City, and the relationship between crimes and other latent factors. The problems we are addressing includes: what's the distribution of crimes? Is there any trend in the dataset? What factor causes this trend? etc. We think these problems are important because our findings will benefit related government departments and organization in some way that they can better understand, control and reduce crimes happening in New York City, and thus we can make some contribution to the social security of New York City.

Some big data techniques are required in this project because the original dataset consists of 5.58 million rows and 24 columns, which makes it not applicable to deal with using traditional data analysis tools such as Microsoft Excel, Python Pandas library, R, etc. Our way to do this is use map-reduce techniques to first extract some useful data from the original dataset, then do some analysis on this smaller dataset. Our scripts to run map-reduce tasks and do data analysis are saved in our github repository.

1 Data Summary and Data Quality Issues

In this section we do some investigation on each column of the crime dataset. We show our findings in Figure 1.

	col_name	# unique values	# missing values	% missing values	base type	semantic type	# valid	# null	# invalid
0	ADDR_PCT_CD	78	390	0.01%	string,int	precinct	5100832	390	0
1	BORO_NM	6	463	0.01%	string	borough	5100759	463	0
2	CMPLNT_FR_DT	6372	655	0.01%	datetime,string	date	5081785	655	18782
3	CMPLNT_FR_TM	1443	48	0.0%	string,timestamp	time	5100271	48	903
4	CMPLNT_TO_DT	4828	1391476	27.28%	datetime,string	date	3704859	1391476	4887
5	CMPLNT_TO_TM	1442	1387783	27.2%	string,timestamp	time	3712063	1387783	1376
6	CRM_ATPT_CPTD_CD	3	7	0.0%	string	flag	5101215	7	0
7	HADEVELOPT	279	4848018	95.04%	string	address	253204	4848018	0
8	JURIS_DESC	25	0	0.0%	string	jurisdiction	5101222	0	0
9	KY_CD	74	0	0.0%	int	code	5101222	0	0
10	Lat_Lon	112827	188146	3.69%	string	coordinate	4913076	188146	0
11	Latitude	112804	188146	3.69%	string,float	latitude	4913076	188146	0
12	LAW_CAT_CD	3	0	0.0%	string	code	5101222	0	0
13	LOC_OF_OCCUR_DESC	6	1127339	22.1%	string	description	3973883	1127339	0
14	Longitude	112808	188146	3.69%	string,float	longitude	4686718	188146	226358
15	OFNS_DESC	71	18840	0.37%	string	description	5082382	18840	0
16	PARKS_NM	864	5093623	99.85%	string	name	7599	5093623	0
17	PD_CD	416	4574	0.09%	string,int	code	5096648	4574	0
18	PD_DESC	404	4574	0.09%	string	description	5096648	4574	0
19	PREM_TYP_DESC	71	33278	0.65%	string	description	5067944	33278	0
20	RPT_DT	3652	0	0.0%	datetime	date	5101222	0	0
21	X_COORD_CD	69533	188146	3.69%	string,int	coordinate	4684320	188146	228756
22	Y_COORD_CD	72317	188146	3.69%	string,int	coordinate	4913076	188146	0

Figure 1: plots of detail information for each column except CMPLNT_NUM

We will elaborate how we determine whether a value is valid or invalid in the following subsections for each column.

As for how we assign the base type and semantic types to each value. We use the get_type() function written in utils.py file to return the base data type. For semantic data type, we mainly first get all the unique values of a data column and combine our understanding of these values with our common sense to get the semantic value for each column. In the script, we hard code the semantic type.

Note that when we find a missing value, we assign 'NULL' to it, which may be recognized as 'string' base type. That's why you see some mixed base types in some columns.

1.1 CMPLNT_FR_DT & CMPLNT_TO_DT

These two columns contains the exact date of occurrence for the reported event (or starting date of occurrence) and ending date of occurrence for the reported event, if exact time of occurrence is unknown. We extract all the unique values and their corresponding count from column CMPLNT_FR_DT and CMPLNT_TO_DT, which both record datetime data. After sorting these values, we find there are some skeptical dates and we label them as invalid outliers because they are certainly out of range. Figure 2 shows the details. These outliers are

1. Date 1015-09-16, 1015-09-26, 1015-10-17, 1015-10-27, 1015-11-25, 1015-12-04 from column CMPLNT_FR_DT
2. Date 2090-04-06 from column CMPLNT_TO_DT

	date	count
0	1015-09-16	1
1	1015-09-26	1
2	1015-10-17	1
3	1015-10-27	1
4	1015-11-25	1
5	1015-12-04	2
6	1900-03-10	1
7	1900-05-08	1

	date	count
4819	2015-12-28	912
4820	2015-12-29	831
4821	2015-12-30	882
4822	2015-12-31	722
4823	2016-01-02	1
4824	2016-01-11	1
4825	2016-03-02	1
4826	2090-04-06	1

Figure 2: table showing the outliers found in column CMPLNT_FR_DT (left) and CMPLNT_TO_DT (right)

Figure 3 and Figure 4 plot how many days in a year are recorded having occurrence of crime in these two columns.

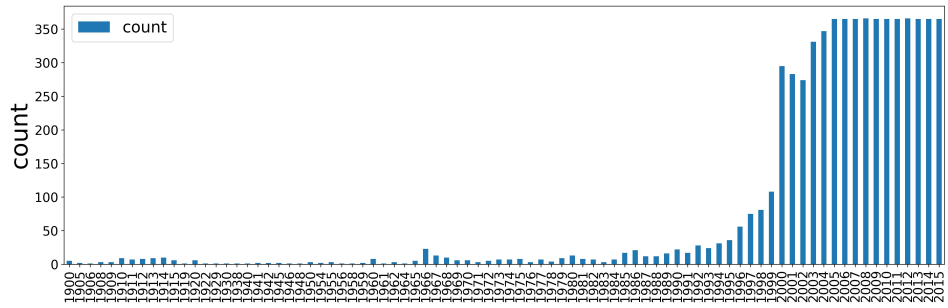


Figure 3: number of days in a year that are recorded in CMPLNT_FR_DT

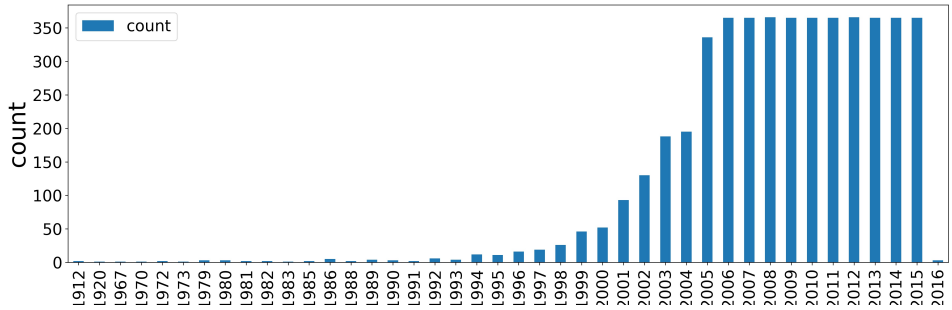


Figure 4: number of days in a year that are recorded in CMPLNT_TO_DT

Looking into these plots, we can find that although some recorded dates are not absurdly out of range, the number of days having occurrence of crime in those years (roughly between 1900 and

2005) are still obviously smaller than the following years.

Since the the official website of the data claims that "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 last year (2015)."

Together with our findings from the plots above. We decide to use 2006-2015 as our final validation interval. Any date before 2006 or after 2015 will be labeled 'INVALID'.

Figure 5 shows how many occurrence of crime in a year between 2006 and 2015.

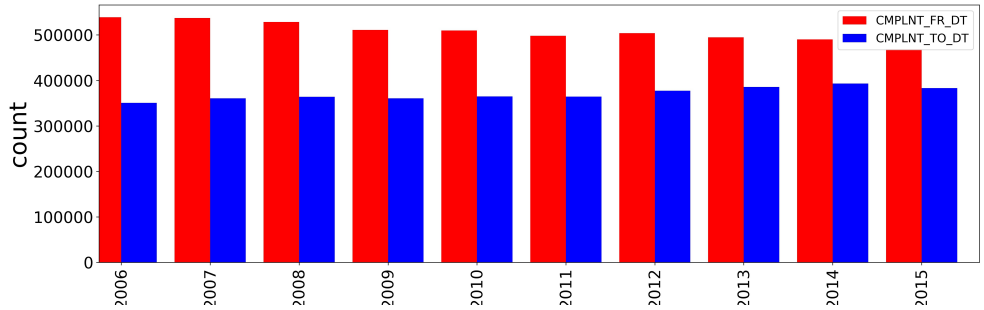


Figure 5: total counts in a year that are recorded in CMPLNT_FR_DT and CMPLNT_TO_DT

From the plot above, we can find an interesting phenomenon: the curve of column CMPLNT_FR_DT goes downward and the curve of column CMPLNT_TO_DT goes upwards. This could be caused by some missing data in column CMPLNT_TO_DT. Note in the previous section we showed that nearly 1/3 of the data in this column is missing. Meanwhile, the downward curve shows that public security of New York City is improving.

1.2 RPT_DT

This column contains the date when event was reported to police.

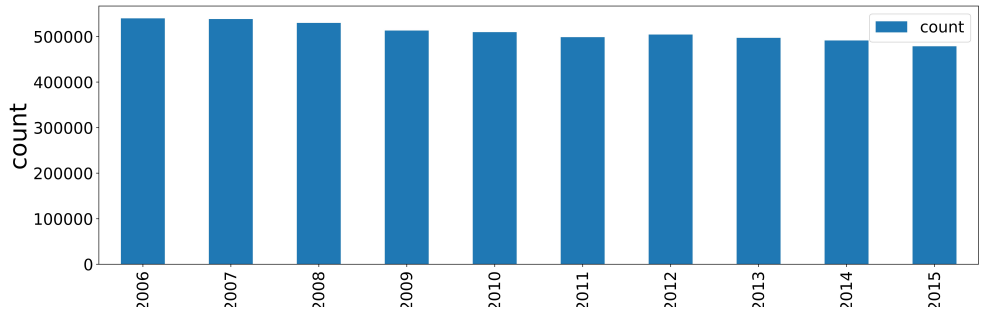


Figure 6: total counts in a year that are recorded in RPT_DT

Figure 6 shows the total number of occurrence of crime reported in a year. This graph also supports the point that the downward curve shows the improvement of public security.

1.3 CMPLNT_FR_TM & CMPLNT_TO_TM

These two columns contain timestamp values of exact and ending time of occurrence. To validate the values in these two columns. We use the following line of code to determine whether a given value is in the valid range.

```
datetime.datetime.strptime(value, '%H:%M:%S')
```

In this way, all values except one (24:00:00) are identified as valid because our function only accept 00:00:00 as midnight. Instead of treating 24:00:00 as outlier, we just convert them to 00:00:00 as it is reasonable for people to get these two mixed up in real life.

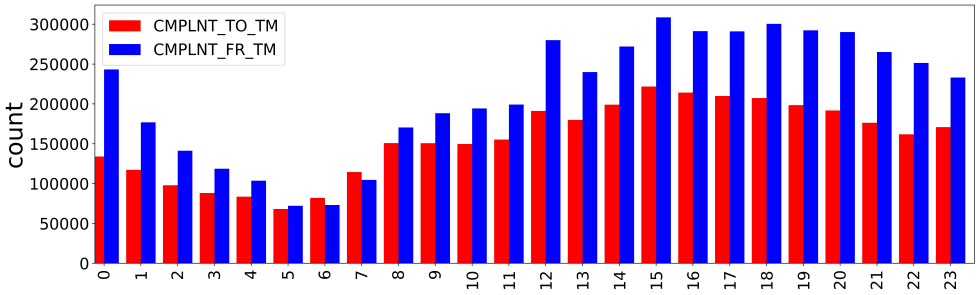


Figure 7: total counts groupby hour that are recorded in CMPLNT_FR_TM and CMPLNT_TO_TM

Figure 7 shows the distribution of number of occurrence against the 24-hour clock. Although a large amount of values are missing in column CMPLNT_TO_TM, we can still find an obvious pattern in this graph: the count of occurrence reaches maximum in late afternoon and minimum around early morning.

1.4 KY_CD & OFNS_DESC & PD_CD & PD_DESC

As these four columns are pre-defined classification code and corresponding description. We assume their values are all valid as long as the codes are three-digits. To help us understand the data, we plot the top 10 KY_CD and PD_CD values order by their frequency in Figure 8 and Figure 9.

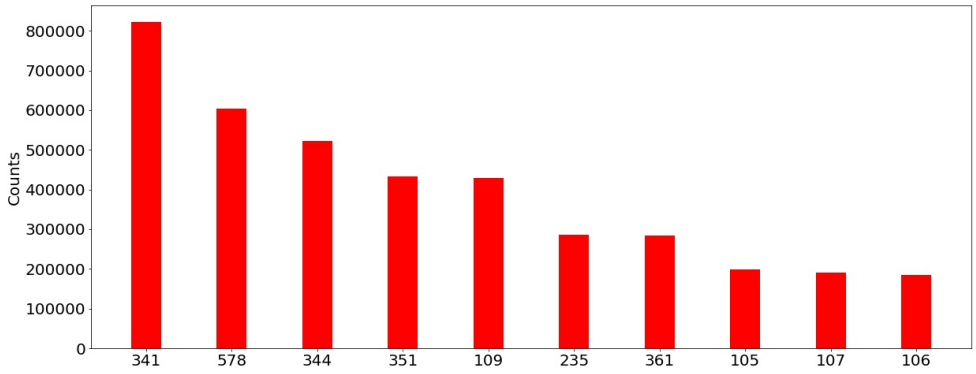


Figure 8: 10 most frequent KY_CD values

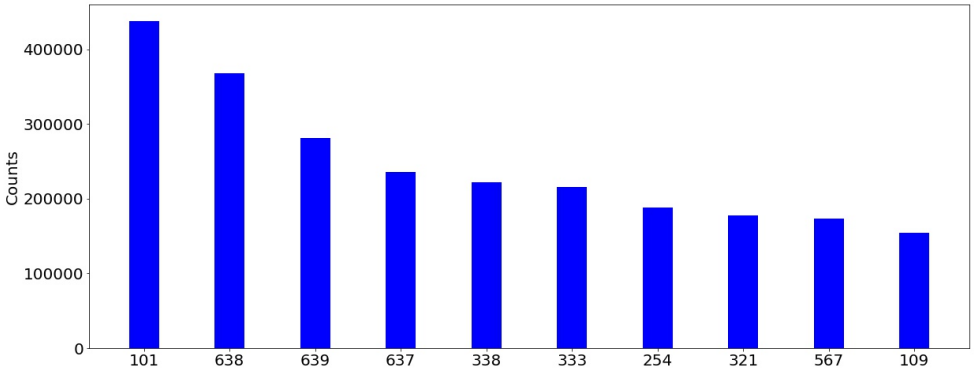


Figure 9: 10 most frequent PD_CD values

KY_CD	OFNS_DESC	PD_CD	PD_DESC
341	PETIT LARCENY	101	ASSAULT 3
578	HARRASSMENT 2	638	HARASSMENT,SUBD 3,4,5
344	ASSAULT 3 & RELATED OFFENSES	639	AGGRAVATED HARASSMENT 2
351	CRIMINAL MISCHIEF & RELATED OF	637	HARASSMENT,SUBD 1,CIVILIAN
109	GRAND LARCENY	338	LARCENY,PETIT FROM BUILDING,UN
235	DANGEROUS DRUGS	333	LARCENY,PETIT FROM STORE-SHOPL
361	OFF. AGNST PUB ORD SENSBLTY &	254	MISCHIEF, CRIMINAL 4, OF MOTOR
105	ROBBERY	321	LARCENY,PETIT FROM AUTO
107	BURGLARY	567	MARIJUANA, POSSESSION 4 & 5
106	FELONY ASSAULT	109	ASSAULT 2,1,UNCLASSIFIED

Figure 10: most frequency codes with corresponding description

Although these two sets of code and description are following different criteria, from Figure 10, we can still conclude that the most frequent types of criminal compliant are: assault, petit larceny, harassment and dangerous drugs.

	top20_type	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
0	CRIMINAL MISCHIEF 4TH, GRAFFIT	4751	8537	9568	9753	9437	10116	8600	7527	10072	10886
1	HARASSMENT,SUBD 3,4,5	39401	36646	34982	34142	34079	33284	35469	36601	40995	41637
2	ASSAULT 3	43953	43877	43274	43519	45600	42315	45122	43909	43782	42353
3	MARIJUANA, POSSESSION 4 & 5	14886	17357	18284	21303	22154	21693	18550	14410	14198	10177
4	LARCENY,PETIT FROM BUILDING,UN	18728	22592	22228	20791	21234	22003	24862	25427	22975	20152
5	CRIMINAL MISCHIEF,UNCLASSIFIED 4	17271	16616	15448	14606	14755	13872	14299	13133	13247	12952
6	LARCENY,PETIT FROM STORE-SHOPL	16432	16927	18518	20530	21911	22133	21940	23410	26490	27067
7	ASSAULT 2,1,UNCLASSIFIED	15453	15546	14399	14851	15135	14747	15553	16047	15981	16207
8	LARCENY,PETIT FROM OPEN AREAS,	8941	8029	7796	6915	6992	7189	7848	7971	7347	7359
9	AGGRAVATED HARASSMENT 2	33837	33078	31098	30763	30024	27748	27250	26760	19442	19819
10	WEAPONS, POSSESSION, ETC	7098	7713	7656	8359	8723	8605	7121	6514	6165	5667
11	INTOXICATED DRIVING,ALCOHOL	7533	8593	8245	7629	6603	5950	6638	8136	7813	5878
12	LARCENY,PETIT FROM AUTO	19774	20783	22626	20411	19204	18002	15580	15009	14205	12092
13	ROBBERY,OPEN AREA UNCLASSIFIED	14525	13117	14286	11505	12395	12452	7847	4814	4204	4394
14	CONTROLLED SUBSTANCE, POSSESSI	12159	12619	11604	10173	9660	10020	9789	9217	9302	8496
15	MISCHIEF, CRIMINAL 4, OF MOTOR	22937	22917	23348	20594	19801	18006	17188	14796	14678	14253
16	LARCENY,GRAND OF AUTO	14734	12111	11450	9676	9344	8407	7147	6344	6562	5894
17	LARCENY,GRAND FROM BUILDING (NON-RESIDENCE) UN...	13383	16248	15370	13915	13892	13700	14699	11344	3911	3705
18	BURGLARY,RESIDENCE,DAY	10340	9602	9348	8910	8218	8320	8527	7141	6594	5397
19	HARASSMENT,SUBD 1,CIVILIAN	29369	28463	26023	24376	23241	20938	21519	20996	20913	19703

Figure 11: Top 20 offense type from 2006 to 2015

We sorted out top20 the most frequent crime types using MapReduce program then we tracked the counts of these crimes from 2006 to 2015. Amongst top20 types, Assault 3,which is a class of Assault, is the most frequent crime type. However Larceny is the most common crime category since it contains Larceny Petit, Laceny Grand and etc. and sum of counts of these larcenies outnumbered the counts of Assault 3. We can see though the counts of many top20 crime types are stable through this decade, some of them are not. For example, the counts of Criminal Mischief 4th, Graffit doubled in 10 years while the counts of Larceny, Grand of auto decreased two third.

1.5 CRM_ATPT_CPTD_CD

The value in these column is an indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely. Thus there are only two unique values (ATTEMPTED and COMPLETED) except NULL value. As this is also an artificially defined variable, we suppose all non-missing values are all valid.

We also find there are less than 2% of crimes that are labeled as "Attempted", while the rest are labeled "Completed".

1.6 LAW_CAT_CD

This column corresponds to level of offense. There are no missing values and all records fall into three different types, we make a pie chart here to show the fraction of every level.

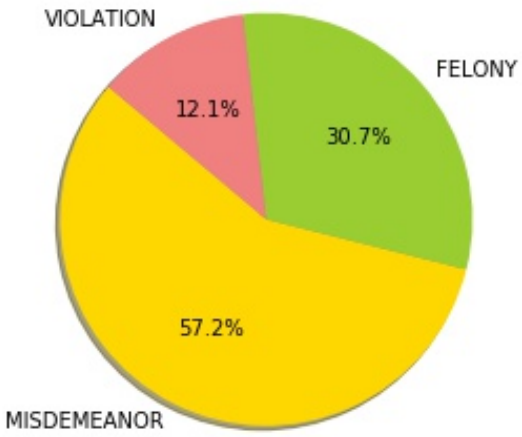


Figure 12: ratio of every level of offense

1.7 JURIS_DESC

This columns tells us which jurisdiction is responsible for incident. By looking at the following pie chart, we can find that most incidents are taken charge by NYPD (around 90%), which makes sense.

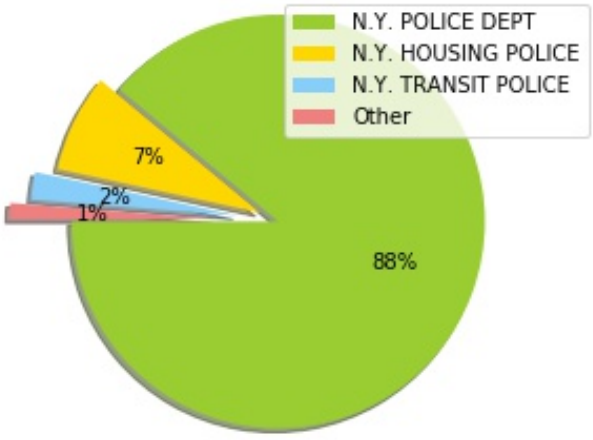


Figure 13: ratio of jurisdiction responsible for incident

1.8 BORO_NM

As we all know, New York City has five boroughs in total. We created a pie chart to display the percent of total number of crimes happened in 5 boroughs. According the figure below, we can see the number of crimes happened in Brooklyn ranks highest, then Manhattan, Bronx, Queens and Staten Island.

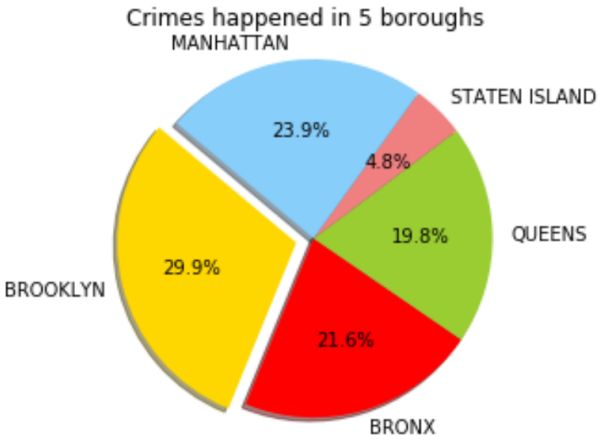


Figure 14: Crimes happened in 5 boroughs

Figure 15 shows that the duration from the start to the end of the incident occurs in 5 boroughs. To be clear, the label in X-axis means that within this period of time but exclude the shorter period of time, for example, within30minutes means within 30 minutes but not within 10 minutes. According to the figure 15,we can see the number of crimes occur within 10 minutes is largest, then within 1 day but not within 1 hour, within 30 minutes but not within 10 minutes. And in different durations of crimes last, Brooklyn still occupies the largest percentage in almost every category, then Manhattan and Bronx.

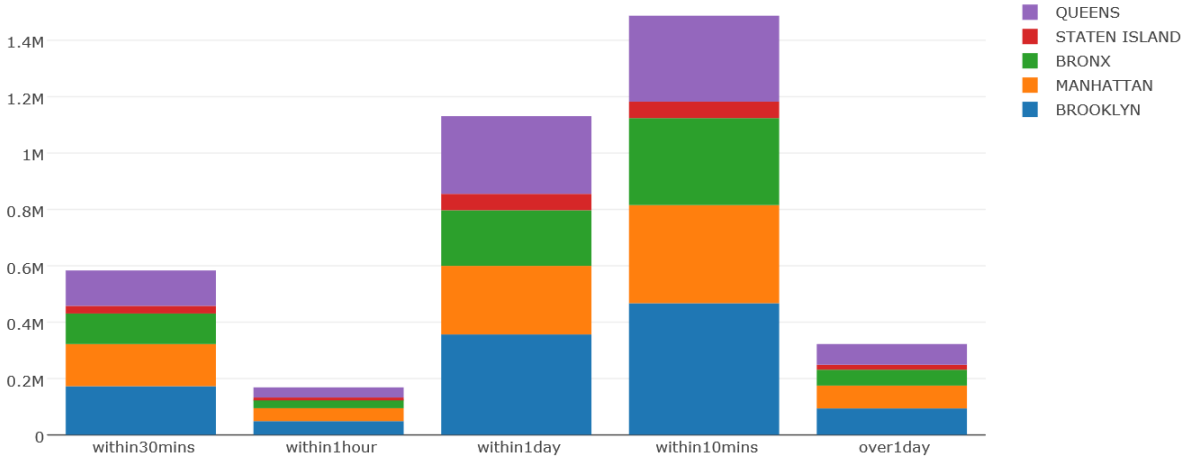


Figure 15: Duration of the incident in 5 borough

Here we make some geographic graphs to show the distribution of crime incidents in space for each borough.

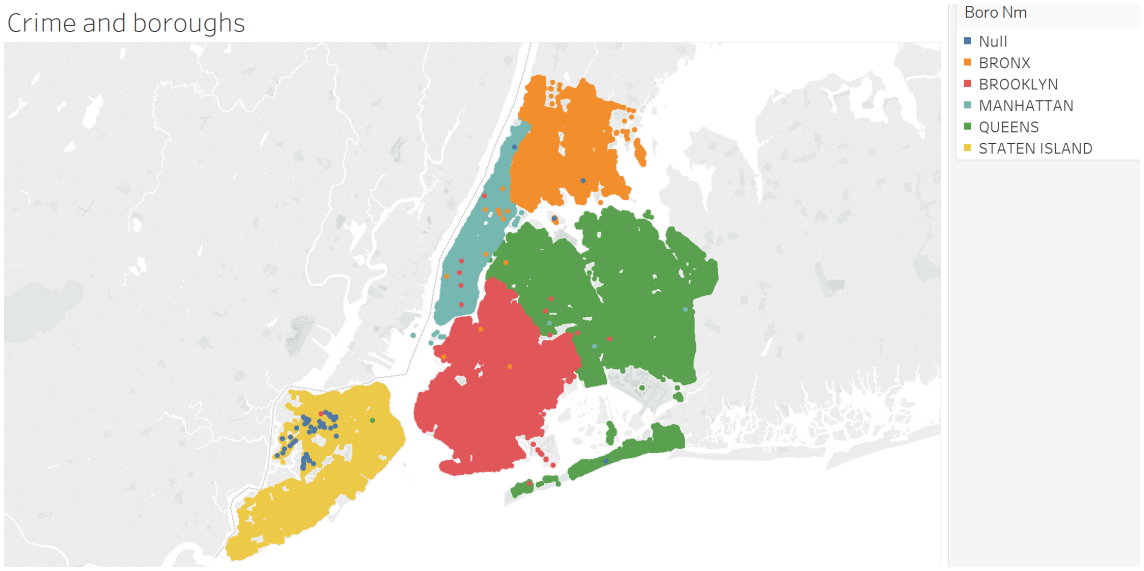


Figure 16: Geomap of NYC offenses

We can see there are few points that are recorded mistakenly. Let's take a closer look.

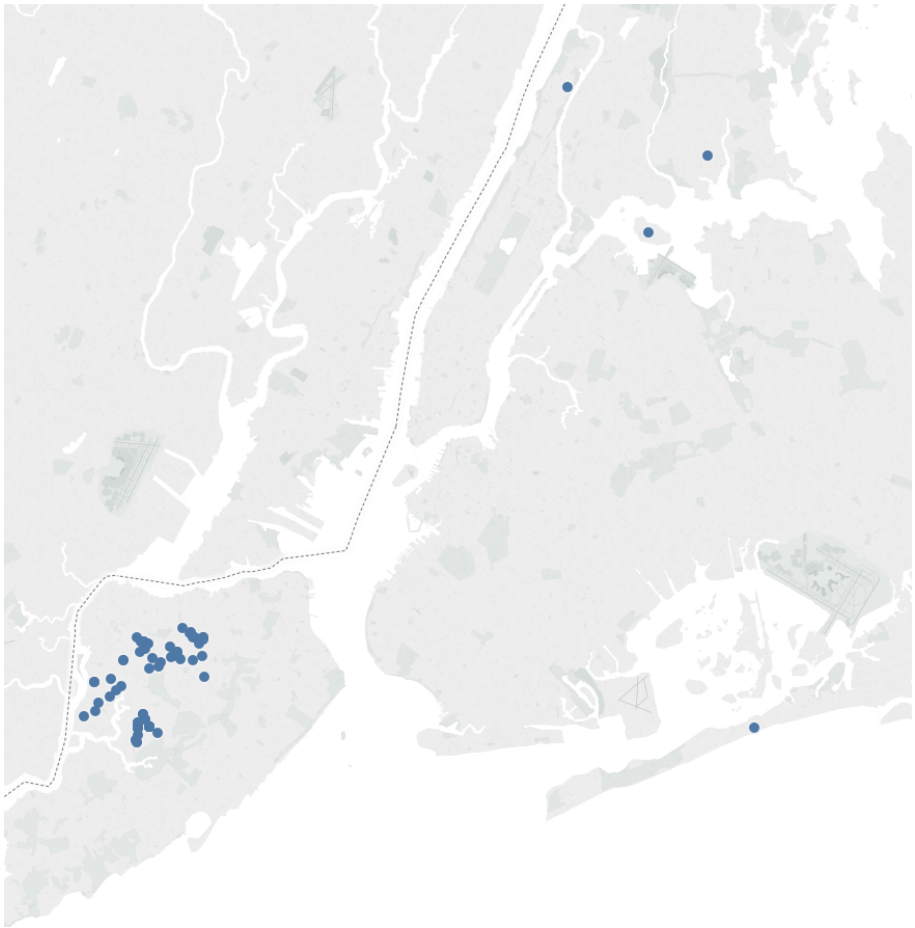


Figure 17: Geomap of offenses without BORO_NM

We can see the most offenses that have no BORO_NM were happened in Staten Island.

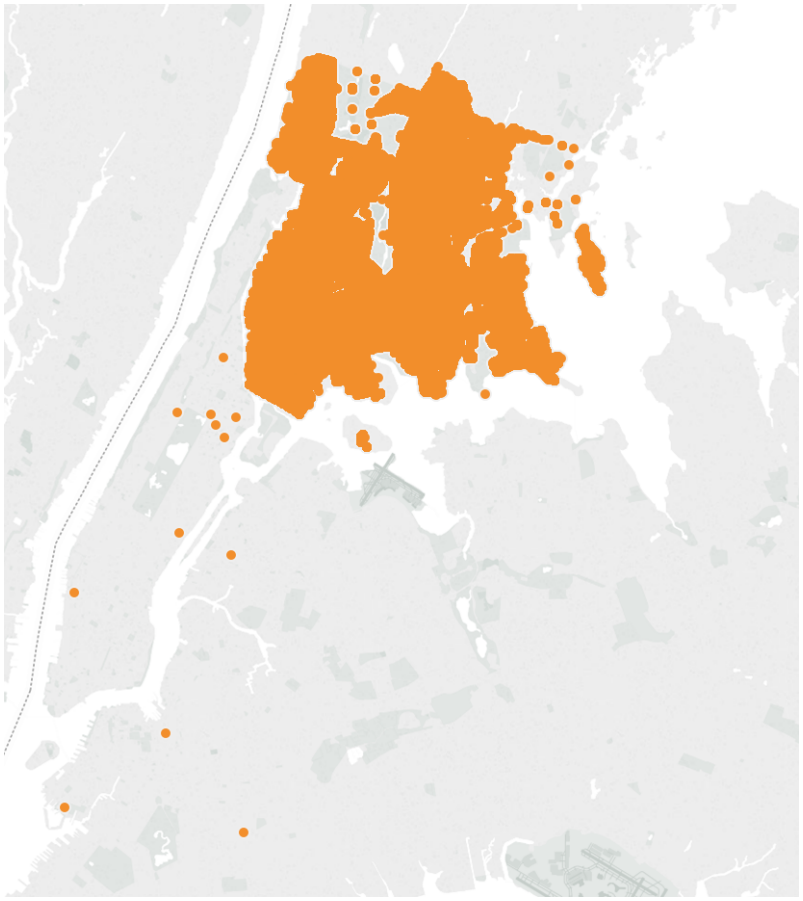


Figure 18: Geomap of offenses in Bronx

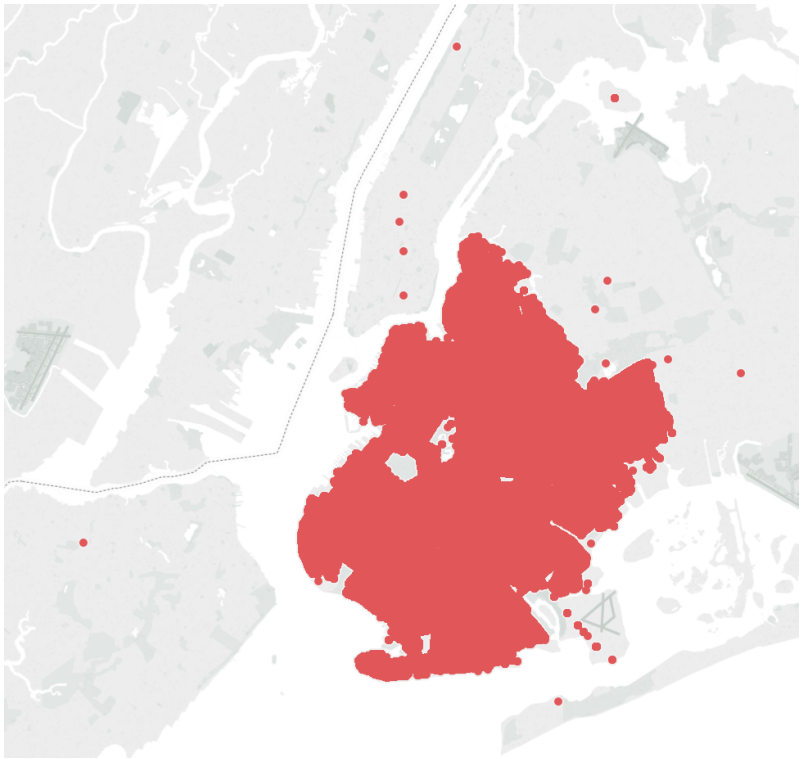


Figure 19: Geomap of offenses in Brooklyn

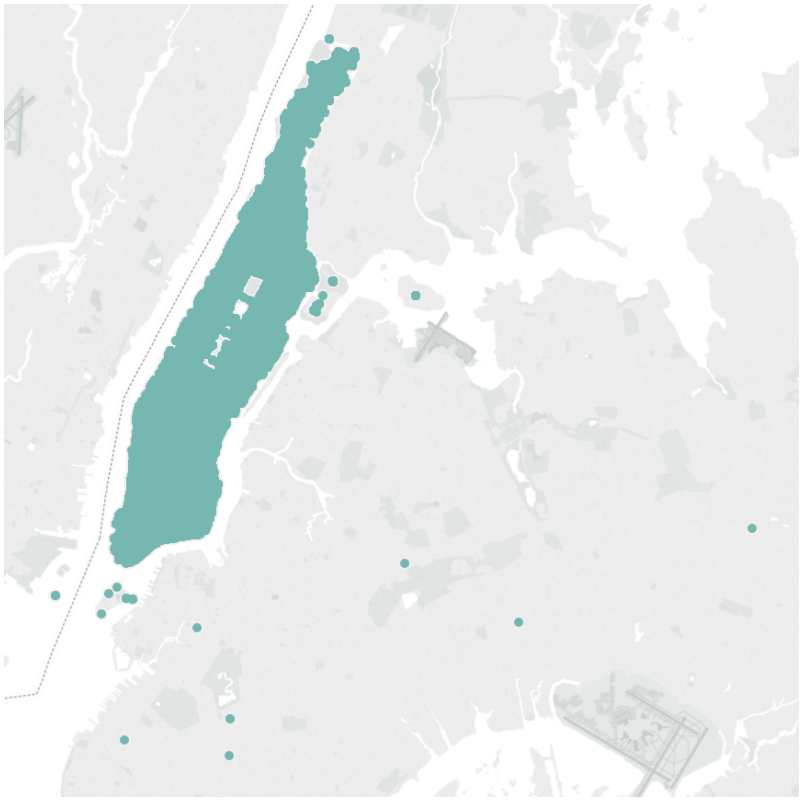


Figure 20: Geomap of offenses in Manhattan

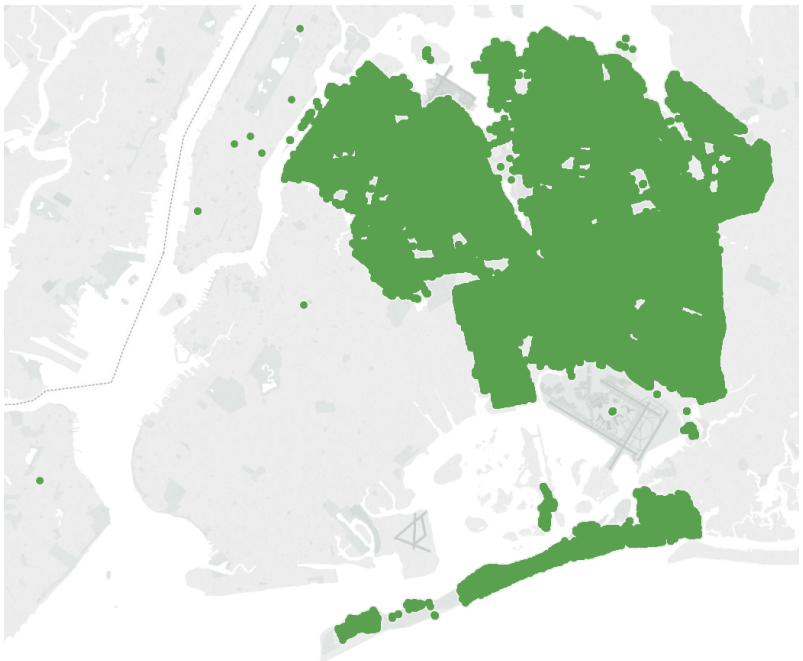


Figure 21: Geomap of offenses in Queens

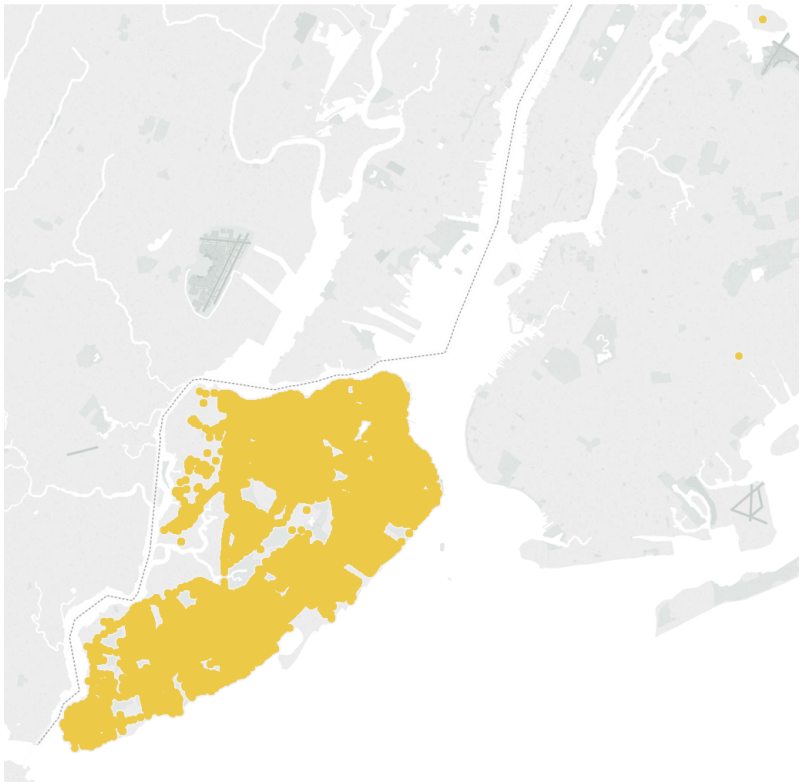


Figure 22: Geomap of offenses in Staten Island

From Figure 16 to Figure 22, we can see for some mis-located offenses, either their borough names are mistakenly recorded or the longitude and latitude are wrong. We will do further research on this data issue, probably in part II.

1.9 ADDR_PCT_CD

The value in this column corresponds to the precinct in which the incident occurred. As we can find all the precincts information on this official website. The way we validate values in this column is to create a set containing all the valid precinct number and check for every value in this column whether it belongs to this column.

An interesting finding here is that, on the official website there is no NO.14, 18 and 22 precinct. Rather, we have Midtown So. Pct. between 13th and 17th Pct., Midtown No. Pct. between 17th and 19th Pct. and Central Park Pct. between 20th and 23th Pct. However in our dataset we do find 14, 18 and 22 in this column. Thus it is very likely that these three numbers correspond to the three precincts we mentioned above.

1.10 LOC_OF_OCCUR_DESC & PREM_TYP_DESC

These two columns together give us location information of occurrence in or around the premises. For these two columns, we assume that all the non-missing values are valid. In the future we may consider combining these two columns together to see if there is some sceptical values such as 'Inside the street'.

1.11 PARKS_NM & HADEVELOPT

These two columns give us the name of NYC park, playground or greenspace of occurrence and NYCHA housing development of occurrence, if applicable. It's plausible that values in these two columns are missing in most of the records. At this stage, we regard every non-missing value as valid value. In the future, we might try to use information from other column to validate data

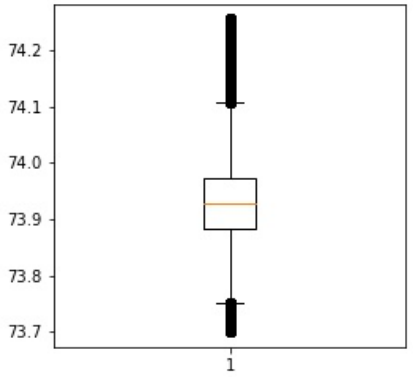
in this column. For example, check whether the park is actually in a specific borough or whether matches the location given by latitude and longitude.

1.12 Longitude & Latitude & X_COORD_CD & Y_COORD_CD

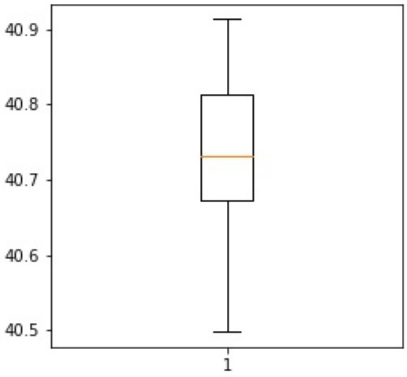
These columns are all coordinates indicating the location of the crime incidents. To validate them, we mainly use boxplot to visualize the data and we regard data points falling outside of the margin as outliers. The range of a boxplot is calculated in this way:

$$\text{range} = (Q_1 - 1.5 * \text{IQR}, Q_3 + 1.5 * \text{IQR})$$

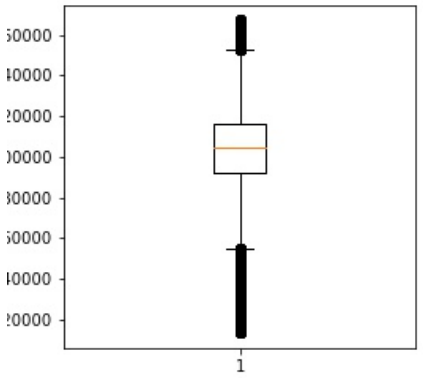
where Q_1 and Q_3 are 25% quantile and 75% quantile respectively. IQR = Q_1 - Q_3.



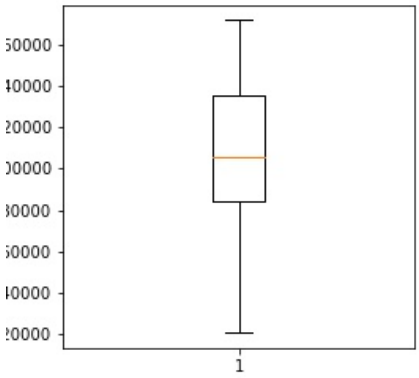
(a) Longitude



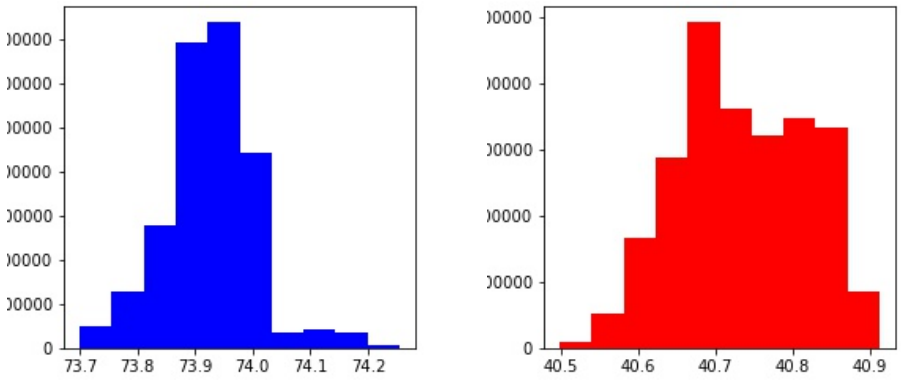
(b) Latitude



(c) X_COORD_CD



(d) Y_COORD_CD



(e) Longitude

(f) Latitude

Figure 23: Box plot of Longitude, Latitude, X_COORD_CD and Y_COORD_CD

From Figure 23, we draw the following conclusion:

1. The distributions of Longitude and X_COORD_CD are almost the same, so are Latitude and Y_COORD_CD. We can assume they are just different from counterparts with respect to scale. (That's why we only include the histogram for Longitude and Latitude here)

2. As both histograms and boxplots indicate, the dispersion of Longitude (X_COORD_CD) is much larger than Latitude (Y_COORD_CD). There are many outliers in Longitude column while Latitude column has no outliers based on our criterion.

1.13 Lat_Lon

This column is just repeating Latitude and Longitude pair. We validate this column by examining whether they are correctly repeating the previous columns.

1.14 Data Quality Issue

For all the data quality issue we found in part I, we list them here:

1. For most columns, there are missing values. Particularly, column CMPLNT_TO_DT, CMPLNT_TM, HADEVELOPT and PARKS_NM have a significant amount of missing values.

2. For column CMPLNT_FR_DT and column CMPLNT_TO_DT, they contains some outliers based on the claim on website that this dataset should only contain data between 2006-01-01 and 2015-12-31.

3. For column CMPLNT_FR_TM and CMPLNT_TO_TM, there are some 24:00:00 values that should be converted to 00:00:00.

4. For column Longitude and X_COORD_CD, some values could be wrong because that location is in some weird place.

5. For column BORO_NM, some values were wrong. For example, there were crimes recorded as happened Manhattan, but the latitudes and longitudes of those data points suggested those crimes were actually happened in other boroughs.

2 Data Exploration

In this section, we attempt to explore the relationship between among columns and datasets. We first give an introduction in Section 2.1 for our experiments. In Section 2.2 we investigate the distribution of crime incidents across different boroughs over time. In Section 2.3 we investigate the distribution of crime incidents across 24-hour time. In Section 2.4 we investigate how the factor of NYCHA affect the occurrence of crime. From Section 2.5 we incorporate our data with census data to further explore how demographic, social and economical factors can influence crime incidents.

2.1 Experimental Setup

1. Cluster Configuration: We use the Dumbo, a 48-node Hadoop cluster, running on NYU HPC.

2. Tools: We mainly use Hadoop map-reduce technique to collect and process data. More details about our map-reduce scripts such as the number of nodes, mappers and reducers are shown in this Github ReadMe. The tools we use to visualize our findings are mainly packages in python such as matplotlib. Details of data visualization are also presented in the ReadMe file.

3. Method of Analysis: To examine our hypotheses, we mainly used Pearson correlation coefficient as score of correlation between variables of our interests. We also used p-value to measure if correlation is significant.

4. Data Sources: Besides the crime data we have, we also incorporate some external datasets. They are:

   (a) NYC Census Data: We use the 2010-2014 American Community Survey (ACS) Profile which contains demographics, social, economic and housing data for each Neighborhood Tabulation Area (NTA) in New York City. To combine this dataset with our crime data, we first aggregate meta data by mapping each data point to a NTA using (longitude, latitude) coordinate based on this NYC NTA shape file. After this step we get counts of crime incidents in each NTA. Then we join this table with the some interesting features from the census data that we manually download for each NTA. Finally we get this merged dataset. The whole process can be reproduced in this ipython notebook

   (b) NYC 2000-2015 Population Data: Since we could only find population data for each borough from decennial census, we used percent of change to estimate the population from 2006-2015 to join our table for crime counts of five boroughs from 2006-2015. Notes: Estimates are based on 2010 Census and reflect changes to the April 1, 2010 population due to the Count Question Resolution program and geographic program revisions. Annual population data reflect estimates as of July 1 in respective year.

2.2 Boroughs Vs. Year

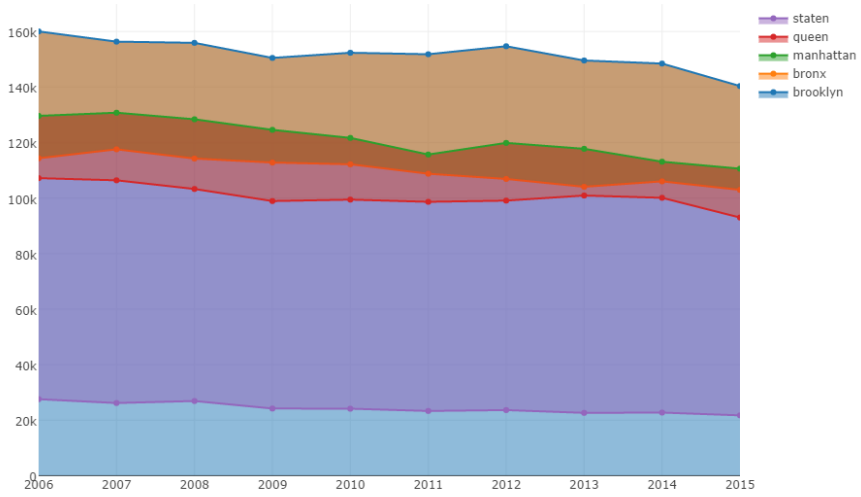


Figure 24: Number of crimes in 5 boroughs each year

Figure 24 is a stacked area chart that shows the number of crimes in 5 boroughs from 2006 to 2015. The y-axis is the number of crime and the x-axis is years. Each line stands for a borough and the colored area below each line shows the amount of increment in crime numbers comparing to the borough beneath. Narrow area means small difference in crime numbers between two boroughs. For example, the difference of crime numbers between Queens and Bronx is small while it is huge between Queens and Staten Island. We found Brooklyn always has the largest number of crimes among 5 boroughs every year while Staten Island always has the fewest. See figure 24.

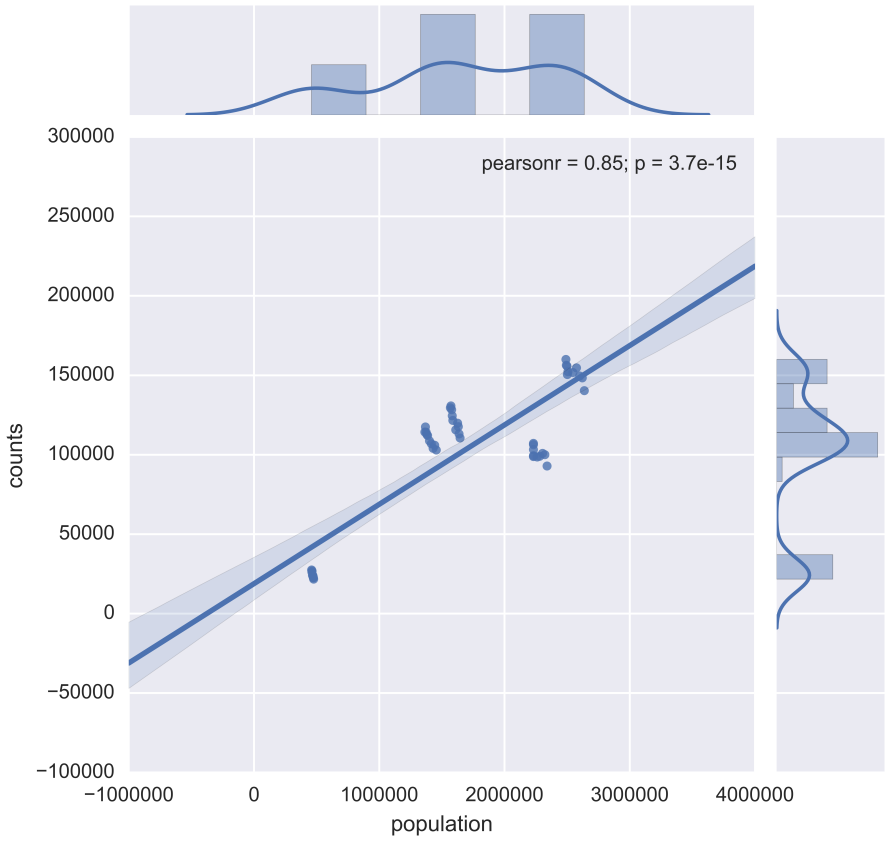


Figure 25: Correlation between the population of boroughs and crime count

Figure 25 shows the correlation between population of five boroughs and crime counts is 0.85 and p-value is significantly small.

2.3 NYCHA Vs. NON-NYCHA

The New York City Housing Authority (NYCHA) provides housing for low- and moderate-income residents throughout the five boroughs of New York City. We compared data points with HADEVELOPT and data points without HADEVELOPT. Our assumption is that crimes occurred around NYCHA housing development would be more dangerous than crimes occurred in other area. We made pie charts to compare percentage of felonies under two situation.

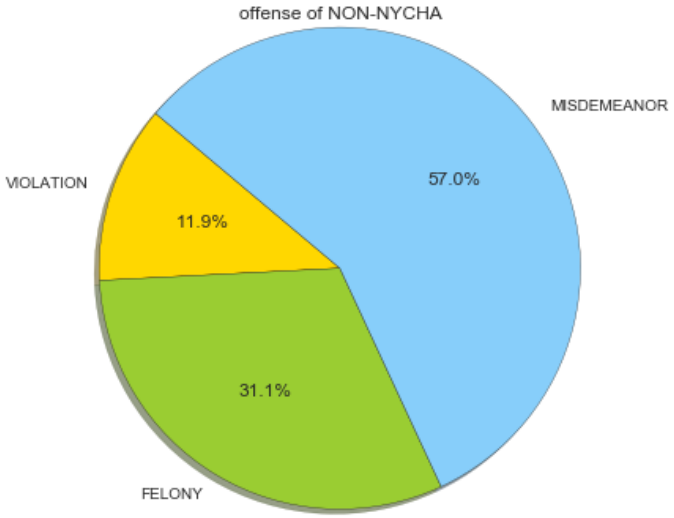


Figure 26: Percentage of LAW_CAT_CD in NON-NYCHA area

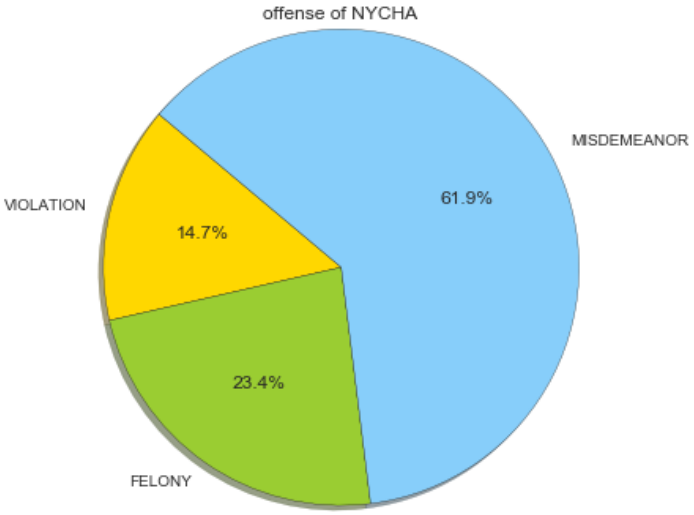


Figure 27: Percentage of LAW_CAT_CD in NYCHA area

From Figure 26 and Figure 27, we can see the percentage of felony in NYCHA area is actually lower than that of NON_NYCHA area. So our hypothesis is disproved. Our another hypothesis about NYCHA development area is that the correlation between population within the area and crime counts would be larger than correlation between overall population and crime counts. To see if out hypothesis is correct, we generate our target correlation and compare it with what we found in Figure 25.

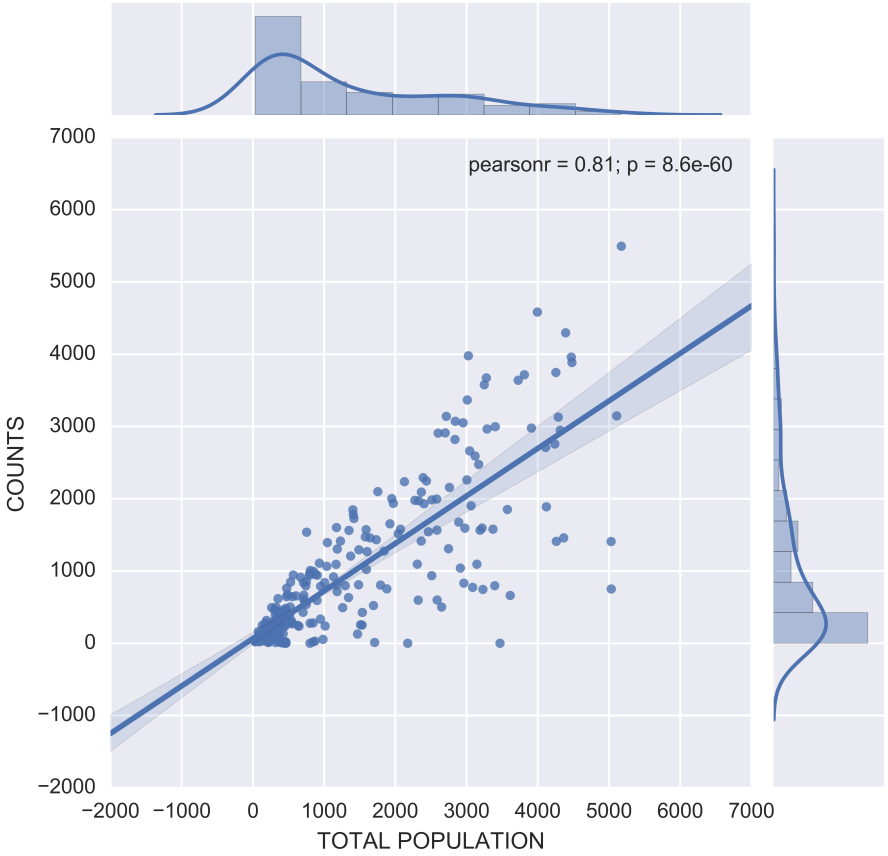


Figure 28: Correlation between the population of NYCHA development area and crime count

From Figure 28 , we can see the correlation is significant with value equals to 0.81 which is lower than the correlation showed in Figure 25. Again, our hypothesis is disproved. We think the reason that our hypotheses about NYCHA development are disproved might be that people live in NYCHA development area will no longer be eligible living there for lower rent if they committed crime. Since residents of NYCHA development have low income, they cannot afford being dispelled from NYCHA housing program.

2.4 Median of Age Vs. Number of Crime

Our hypothesis is that, for a neighborhood, the older the median age is the less crimes will occur. This makes sense because with older median age, the employment rate of a neighborhood tend to be higher, and the population tend to have more stable life style whereas young population tend to be more energetic and unstable.

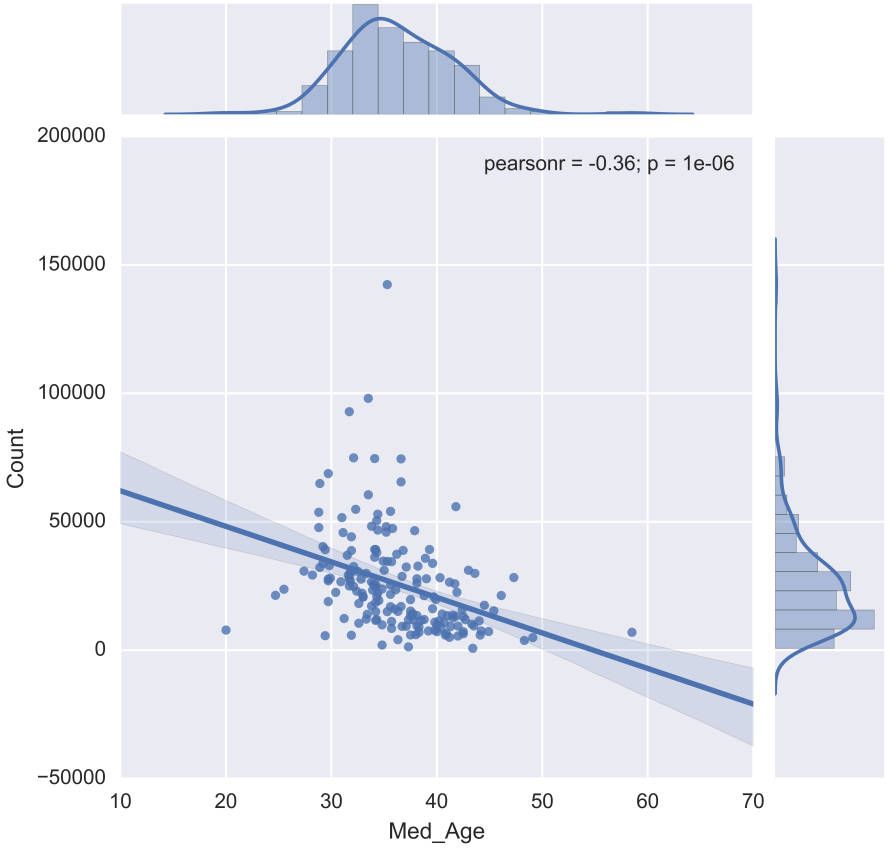


Figure 29: Correlation between the median age and crime count

From Figure 29, we can see crime counts is negatively correlated to median age and p-value is small enough to conclude this correlation is significant. So our hypothesis, as we previously stated, that the older the median age is the less crimes will occur within a neighborhood is proved

2.5 Unemployment Rate Vs. Number Of Crime

Our hypothesis is that, for a neighborhood, the higher the unemployment rate is the more crimes will occur. This makes sense because with higher unemployment rate, more people will be discontented with their life and the neighborhood seems to be less wealthy and safe.

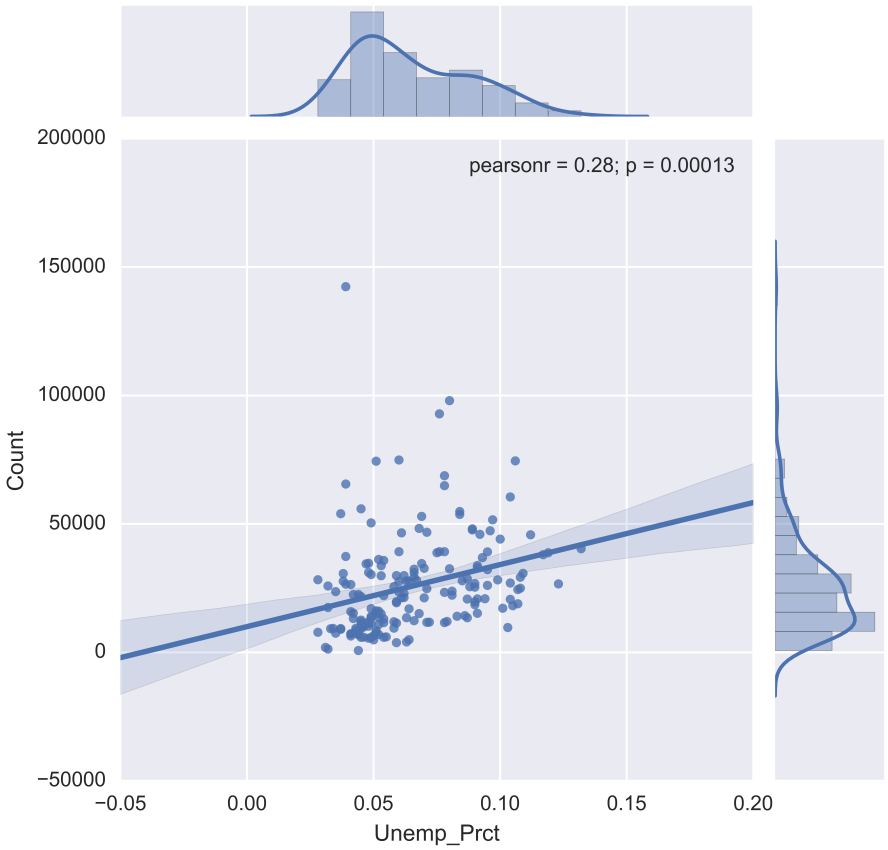


Figure 30: Correlation between the unemployment rate and crime count

From Figure 30, we can see crime counts is positively correlated to unemployment rate and p-value is small enough to conclude that this correlation is significant. So our hypothesis, as we previously stated, that unemployment rate is positively correlated to number of crimes within a neighborhood is proved.

2.6 Female Percentafe Vs. Number Of Crime

Our hypothesis is that, for a neighborhood, the higher the Female percentage is the more crimes will occur. This makes sense because with higher Female percentage, the neighborhood seems to be more vulnerable.

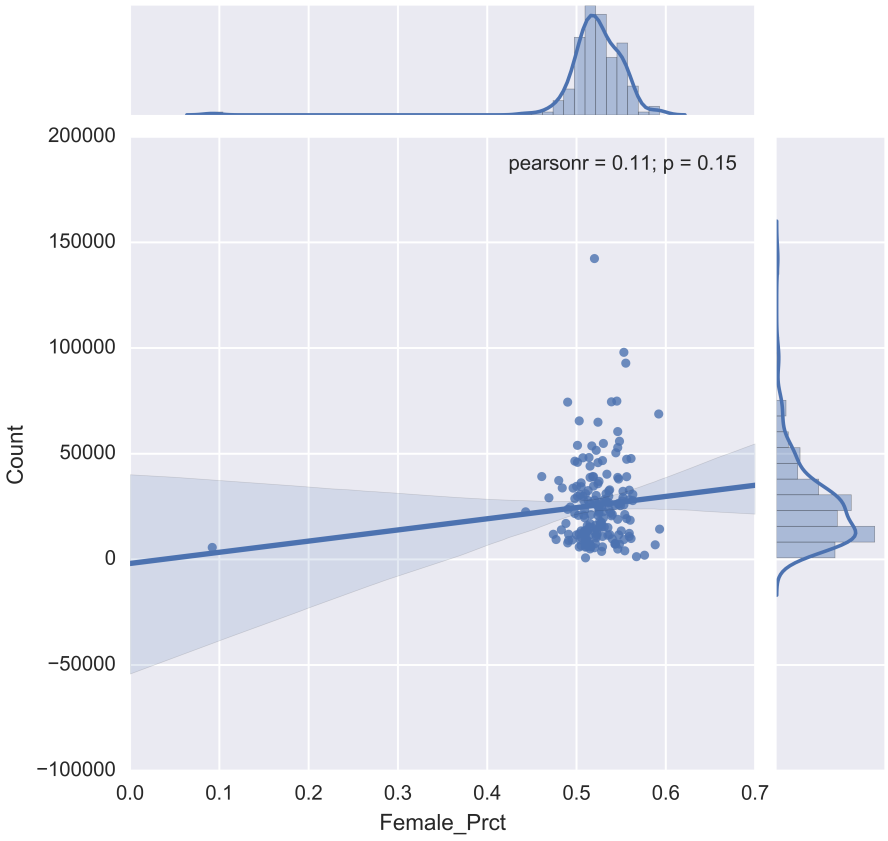


Figure 31: Correlation between the Female percentage and crime count

From Figure 31, we can see crime counts is not significantly correlated to Female percentage because p-value is not small enough to conclude there is correlation between them. So our hypothesis is disproved. But we should notice there is a extreme value that female percentage smaller than 0.1, and the crime counts is less than other neighborhood with higher female percentage. This extreme value makes the correlation appears to be positive though there does not necessarily exists correlation between the Female percentage and crime count generally.

2.7 Percentage Of Black Population Vs. Number Of Crime

Our hypothesis is that, for a neighborhood, the higher the percentage of African American population, the more crimes will occur. This makes sense because for a neighborhood with higher percentage of African American population, which usually is more populous,the unemployment rate tends to be higher and the public security condition generally tends to be worse than other neighborhoods due to higher population density.

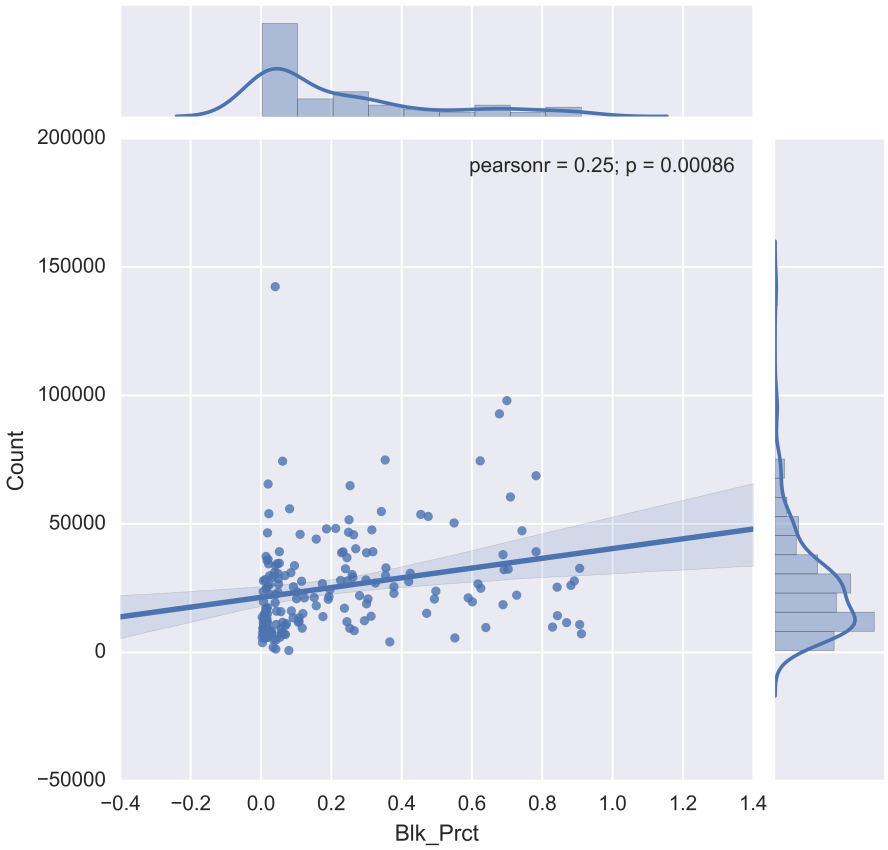


Figure 32: Correlation between the Percentage of African American population and crime count

From Figure 32, we can see crime count is positively correlated to the percentage of African American population in this area, the p-value is small enough to conclude this correlation is significant. So our hypothesis, as we previously stated, that percentage of black people is positively correlated to number of crimes within a neighborhood is proved.

2.8 Percentage of Married Male Vs. Crime

Our hypothesis is that, for a neighborhood, the more married men and the less crimes will occur. This makes sense because for a men married, he has to take the responsibility to take care his wife and his children, and in general, he would not take the risk to commit a crime.

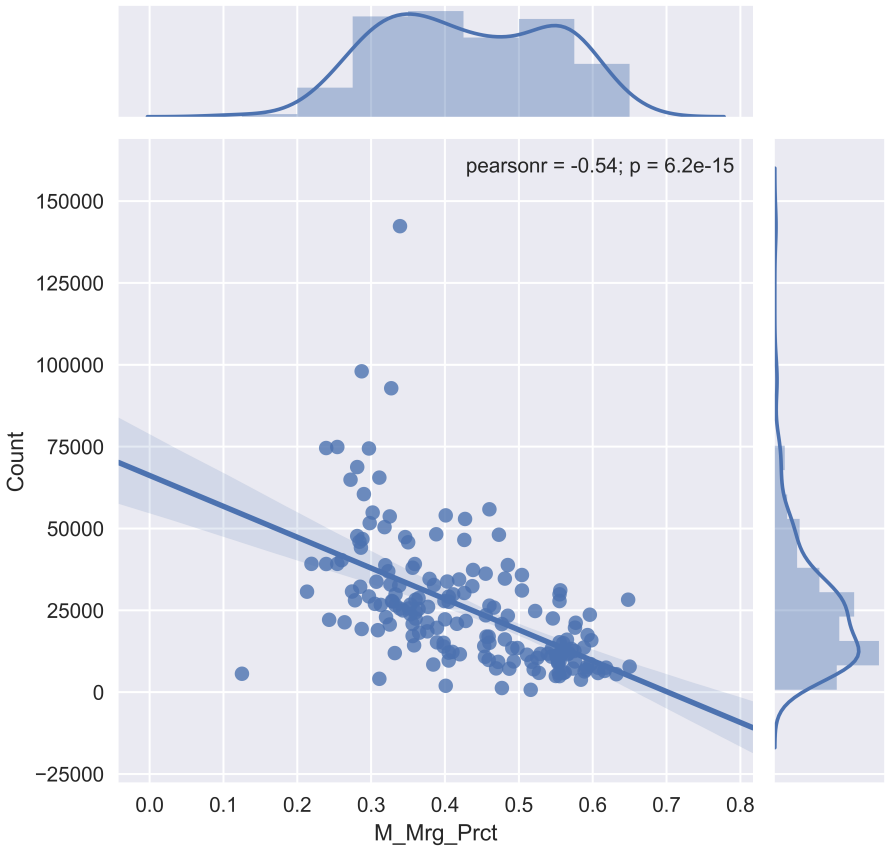


Figure 33: Correlation between the percentage of married men and crime count

From Figure 33, we can see crime count is negatively correlated to the percentage of men married in this area, the p-value is small enough to conclude this correlation is significant. So our hypothesis is proved. As we previously stated, the married male has a family to take care, so he would not like to take the risk to ruin his family and hurt his children.

2.9 Percentage of the Poor Vs. Crime

Our hypothesis is that, for a neighborhood, the more people live below the poverty line, the more crimes will occur at his area. This makes sense because for a person who lives below the poverty line and cannot afford his life, he has to look for various ways to cover his living cost. When he cannot find a legal way to make money, he has to think of a illegal way to get money. Besides, a person with good life seldom commit a crime for money, and he does not want to commit a crime at the expense of losing his good life.

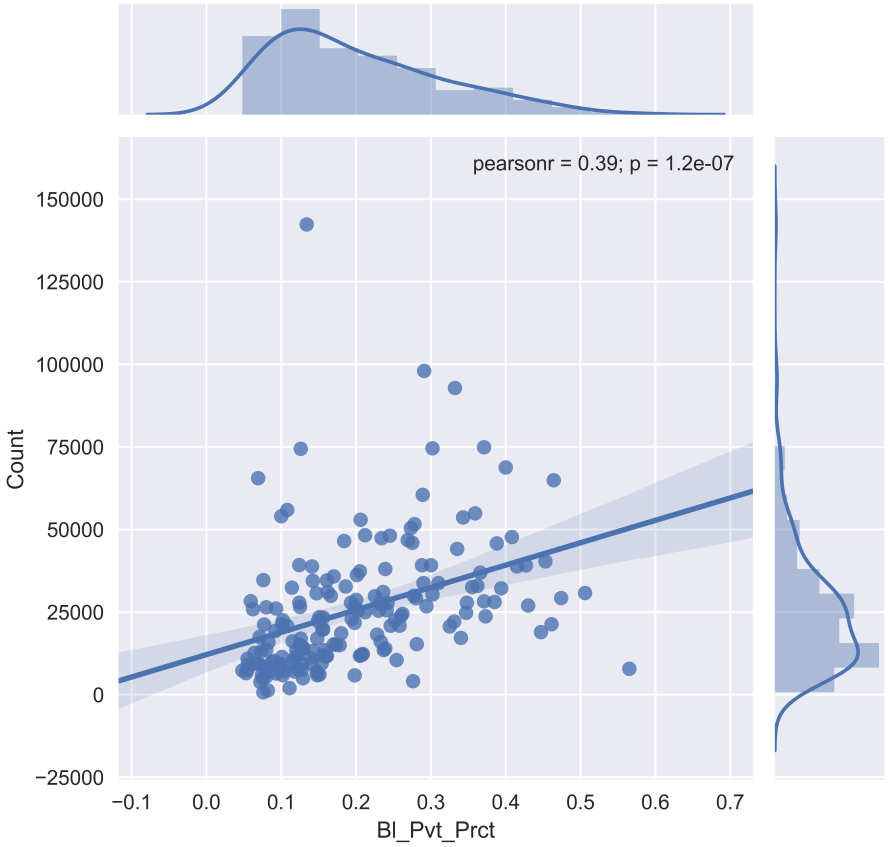


Figure 34: Correlation between the percentage of population below the poverty line and crime count

From Figure 34, we can see crime count is positively correlated to the percentage of neighborhood live below the poverty line, the p-value is small enough to conclude this correlation is significant. So our hypothesis is proved. As we previously stated, the poor people cannot afford their life, so they have motivation to commit crimes, but the wealthy people would not like to commit crimes at the cost of losing their good living condition.

2.10 Percentage of Veteran Vs. Crime

Our hypothesis is that, for a neighborhood, the more veteran live here, the less crimes will occur at this area. This makes sense because veterans ever made contribution to the development and safety of this country. They were just and brave, and feel responsible to their country, so they would not like to destroy the society, on the contrary, they would like to make contribution to preserving the safety of their neighborhood.

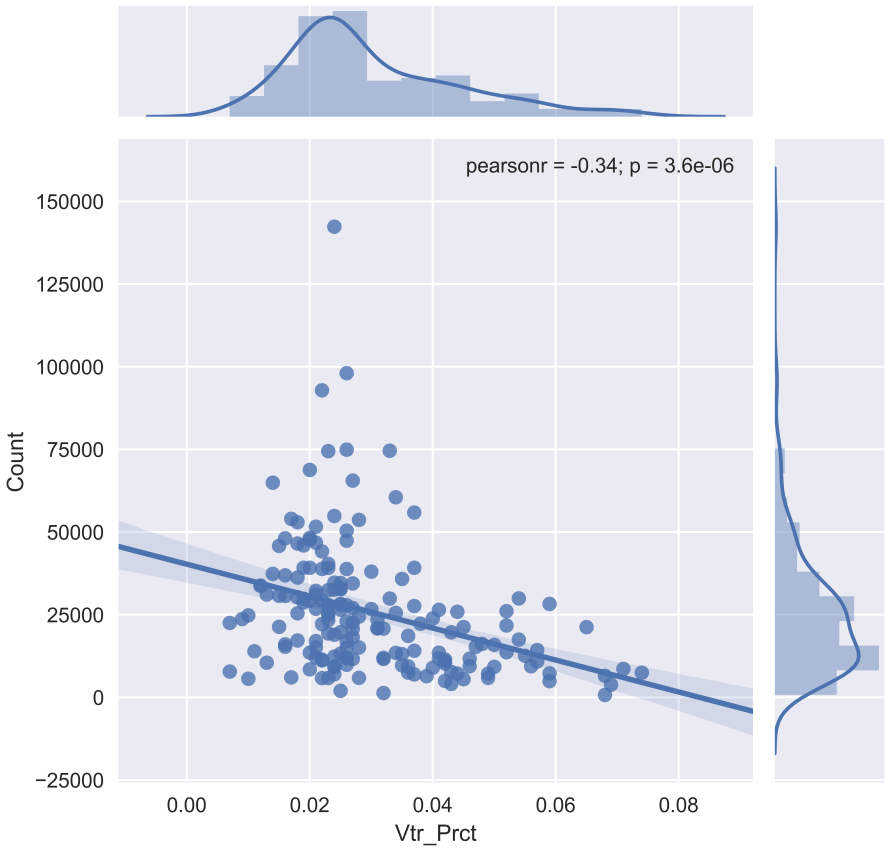


Figure 35: Correlation between the percentage of veteran and crime count

From Figure 35, we can see crime count is negatively correlated to the percentage of veterans, the p-value is small enough to conclude this correlation is significant. So our hypothesis is proved. As we previously stated, the veterans are just and responsible to this country, so they would make contribution to safeguard their neighborhood.

2.11 Percentage of Disabled Vs. Crime

Our hypothesis is that, for a neighborhood, the more disabled people live here, the less crimes will occur at this area. This makes sense because it might be difficult for them to take care of themselves, furthermore, they might not be capable of committing a crime.

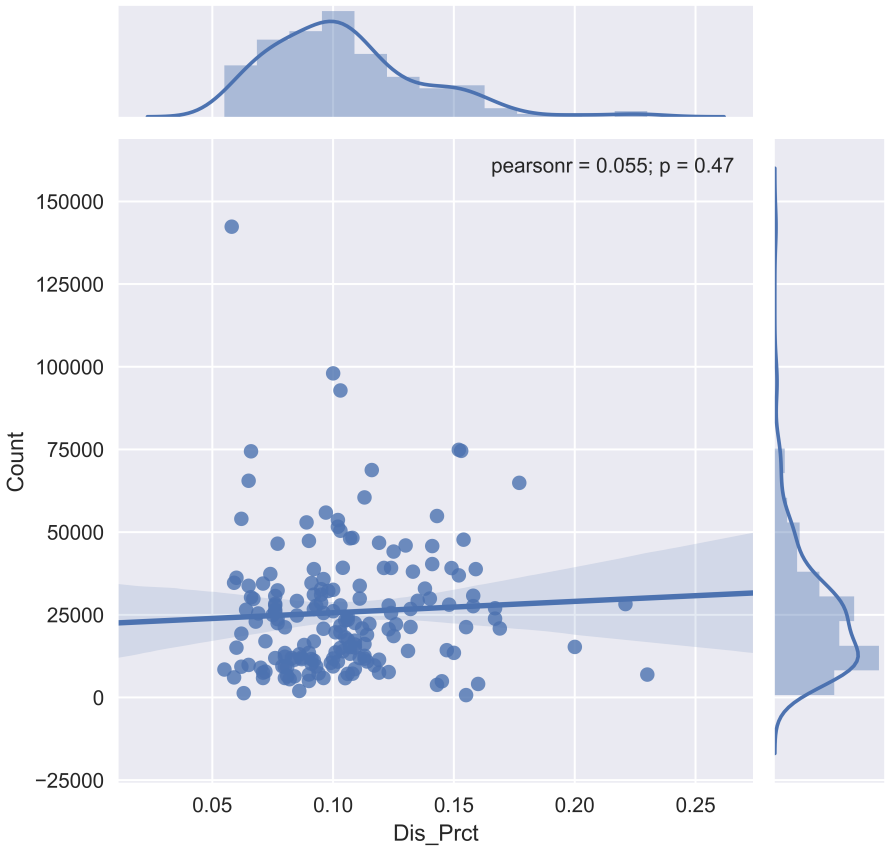


Figure 36: Correlation between the percentage of disabled and crime count

From Figure 36, we can see crime count is not correlated with the percentage of disabled people, the pearson value is close to 0, and the p-value is too large to reject the null hypothesis. So our hypothesis cannot be proved. The reason might be disabled people are not criminals, and whether there are disabled people in this area cannot influence the crime count.

2.12 Percentage of Population with High School Diploma Vs. Crime

Our hypothesis is that, for a neighborhood, the more people with high school diploma, the less crimes will occur at this area. This makes sense because if a person earned a high school diploma, he can tell legal from illegal and he knows the serious and terrible result of committing a crime. Moreover, when a person with high school diploma, it is more possible for him to have a higher level education and find a good job, and enjoy good living condition. Then he does not want to take the risk to lose a good job and his stable and peaceful life.

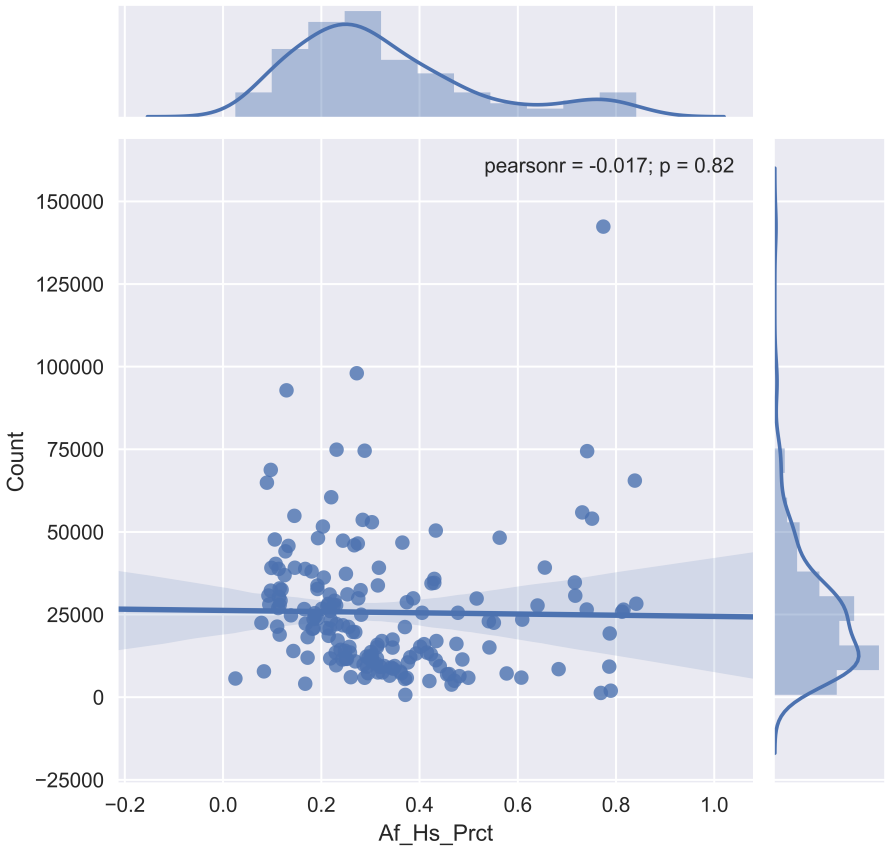


Figure 37: Correlation between the percentage of population with high school diploma and crime count

From Figure 37, we can see crime count is not correlated with the percentage of people with high school diploma, and the pearson correlation value is very small and the p value is very large so that we cannot reject the null hypothesis. In a word, the correlation between the percentage of neighborhood with high school diploma and crime count is not significant, our hypothesis cannot be proved. Therefore, the education is not an important factor that influences the crime commit, and some advanced crime can only committed by smart criminals with high-level education background.

2.13 Percentage of Citizens Vs. Crime

Our hypothesis is that, for a neighborhood, the more citizens, the less crime will occur. This makes sense because a citizen has more job opportunities and enjoy more welfare benefits in the society, so they can have a good life. Citizens are seldom treated unfairly and there are fewer conflicts and troubles they face than those people without citizenship, so they have less motivation to commit a crime than those people without citizenship.

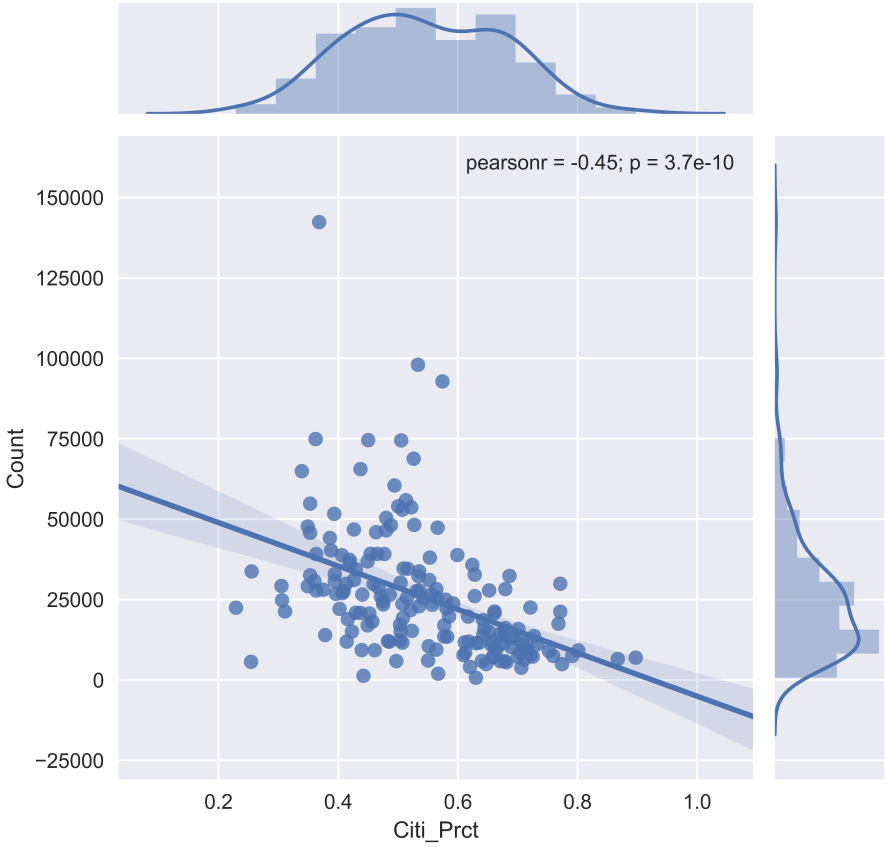


Figure 38: Correlation between the percentage of citizens and crime count

From Figure 38, we can see crime count is negatively correlated to the percentage of citizens in this area, the p-value is small enough to conclude this correlation is significant. So our hypothesis is proved. As we previously stated, citizens can enjoy more welfare benefits and be treated more fairly, the neighborhood with more citizens, the less crimes occur here.

2.14 List of Relationships & Associated Score

	Attributes	Score
crime count	population of borough	0.85
	population of NYCHA	0.81
	median age	-0.36
	unemployment rate	0.28
	percentage of female	0.11
	percentage of African American	0.25
	percentage of married male	-0.54
	percentage of population below poverty line	0.39
	percentage of veteran	-0.34
	percentage of disabled	0.055
	percentage of population with high school diploma	-0.017
	percentage of citizens	-0.45

3 Conclusion

3.1 Summary of Findings

1. As for the quality of the data, we find that most columns have missing values. Particularly, CMPLNT_TO_DT, CMPLNT_TO_TM, HADEVELOPT, LOC_OF_OCCUR_DESC and

PARKS_NM have a large portion of missing values. In addition, based on our validation criterion, we find a lot of invalid values in these columns: CMPLNT_FR_DT, CMPLNT_FR_TM, CMPLNT_TO_DT, CMPLNT_TO_TM, Longitude and X_COORD_CD.

2. As for the distribution of crime, here are some interesting findings:
   (a) Either defined by from date or report date, the curve of total number of crimes per year goes downward from 2006 to 2015. This may indicate the social security of New York City is improving.
   (b) Based on 24 hour clock, Most crimes happen during afternoon and early evening. In contrast, very few crimes happen between 3:00 AM and 6:00 AM.
   (c) Based on the top 10 most frequent KY_CD and PD_CD and their corresponding description, we find the most frequent types of crimes are assault, petit larceny, harassment and dangerous drugs.
   (d) Most (more than 98%) crimes are labeled as "Completed" in the dataset.
   (e) Most (around 60%) crimes are labeled as "Misdemeanor" for the level of offense.
   (f) NYPD are responsible for most (nearly 90%) crime incidents.
   (g) Most crimes (nearly 1/3) happen in Brooklyn. In contrast, much fewer crimes (less than 5%) happen in Staten Island.
3. As for the factors that influence crimes, we find that:
   (a) Population certainly is one of the most important factors, the larger the population is the more crimes occur.
   (b) Median age of a neighborhood also significantly affect the crime counts within that neighborhood: neighborhood with older median age will have less crime counts.
   (c) Unemployment rate has notable impact on crimes. The higher the unemployment rate is, the more crimes will occur.
   (d) The percentage of married male in that area is also an important factor, the more married male live in this area, the less crime occur here.
   (e) The percentage of the veteran in that area also influences the crimes happen within this area: the more veteran, the less crimes occur.
   (f) The percentage of citizens is also an important factor to the crime count: the more citizens within that neighborhood, the less crimes occur.
   (g) The percentage of neighborhood live below the poverty line is also an important factor to the crime count: the more neighborhood live below within that neighborhood, the more crimes occur.
   (h) The percentage of African American population is also an important factor to the crime count: the more African American live within that neighborhood, the more crimes occur.
   (i) The percentage of Hispanic population is also an important factor to the crime count: the more Hispanic live within that neighborhood, the more crimes occur.
   (j) The percentage of Asian population is an important factor to the crime count in a good way: the more Asian live within that neighborhood, the less crimes occur.
   (k) The percentage of White population is an important factor to the crime count in a good way: the more White people live within that neighborhood, the less crimes occur.
   (l) We suppose the the percentage of female or male might be a factor that influences the crime count. However we finally found it is not correlated with crime count.
   (m) We suppose the the percentage of neighborhood with high school diploma might be a factor that influences the crime count, but finally we find that the percentage of neighborhood with high school diploma is not correlated with crime count.
   (n) We suppose the the percentage of disabled neighborhood might be a factor that influences the crime count, but finally we find that the percentage of disabled neighborhood is not correlated with crime count.

3.2 Team Member Contribution

1. Weitao Lin: Column Investigation, incorporate crime data with census data, write up.
2. Shangying Jiang: Column correlation investigation in crime data, correlation investigation between census data and crime data.
3. Xue Yang: Column correlation investigation in crime data, correlation investigation between census data and crime data, write up.

References

[1] NYPD website,
http://www.nyc.gov/html/nypd/html/home/precincts.shtml

[2] Crime data website,
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[3] Census data website,
http://maps.nyc.gov/census/

[4] NTA,
https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page