A Study of Crime Activity in New York City

Chong han ch2905@nyu.edu **Mingzhong Dai** md3797@nyu.edu **Yingbing Wang** yw2848@nyu.edu

Abstract:

This report presents a work of how we handle and analyze crime data in New York in the past 10 years(2006 -- 2015). According to the process we took, this report is delivered mainly in the following four sections:

1. Data summary and data quality section states the data issues we met in this dataset and how we deal with these issues when cleaning. We obtain a clean dataset of 4.6 million entry after cleaning on the original 5.1 million entry dataset, which is 90.89%.
2. Data extraction job is done in the second part. It lists the data name, type and format we want to extract for analyzing along with the specific spark script each job use.
3. In data visualization section we will use the data extracted before. After a deeper inspection of crime type, time and borough aspects, we show how crimes differ in each type and level, distribute variously in daily and month and the similarity and difference each borough have.
4. The last part is data correlation analysis. We find some distribution features of the data, make hypothesis and prove the correctness of them.

Table of Contents

Introduction:

The motivation for us to choose this dataset is that, as international students in a foreign country, the most important things we care about is safety. By observing and analyzing the crime data in New York city, it presents a reliable and statistical results of how safe the city we live is. Besides, we hope our study could provide instructions for government to decrease certain types of crimes and make policies.

This report presents cleaning and statistical analysis on 5.1 million crime entry compiled from the New York City Police Department's records management system. Crime Complaint Reports contain various information like crime type, description, report time, location, completed/attempted etc. in 24 columns. The crime complaint contained in this report represent crimes occurring from January 1, 2006 thru December 31, 2015.

Our work is mainly divided into three parts. Firstly, Data Summary and Data Quality part states the data issues we met in this dataset and cleaning rules we defined when dealing with "bad data". A clean dataset of 4.6 million entry is obtained after cleaning on the original 5.1 million entry dataset. Next, data extracting job is done In the second section Data Extracting Description. It lists the data name, type and format we want to extract for analyzing along with the specific spark script each job use. Data successfully extracted is used in Data Visualization section, in which crime activities are presented in a more visual and intuitional way. After a deeper inspection in crime type, time and borough aspect, we show how crimes differ in each type and level, distribute variously in daily and month and the similarity and difference each borough have.

Roughly, the original crime dataset is 1.3 Gigabytes. It is still more than 1 Gigabytes even after cleaning. It overwhelms a PC to deal with such a big data and get our computer broken down when frequently read and write. With this file distributed and proceed parallelly in Spark, we are able to analyze the data within acceptable time. For the data cleaning job, which aims to check data type and delete invalid information, roughly cost 10-20 minutes depending on the server condition, but on our personal computer, it usually takes more than 1 hours, which is annoying when debugging. Not to mention Data Extracting jobs, in which we need to run dozens of script on the entire dataset.

1. Data Summary and Data Quality

1.1 Type Information

Table 1.0 shows type information and default semantic type for each column, with the script at ***DataCleaning/dumbo_count.py***. Note that we deleted the first line before running, which contains column name information. Thus, the total line number is 5101231.

table 1.0 data summary

	Column name	Type information	Semantic type
0	CMPLNT_NUM	int, 5101231	ID
1	CMPLNT_FR_DT	date, 5100576 null, 655	Date
2	CMPLNT_FR_TM	time, 5101183 null, 48	Time
3	CMPLNT_TO_DT	date, 3709753 null, 1391478	Date
4	CMPLNT_TO_TM	time, 3713446 null, 1387785	Time
5	RPT_DT	int, 5101231	Date
6	KY_CD	int, 5101231	Code
7	OFNS_DESC	string, 5082391 null, 18840	Code
8	PD_CD	int, 5096657 null, 4574	Code
9	PD_DESC	string, 5096657 null, 4574	Description
10	CRM_ATPT_CPTD_CD	string, 5101224 null, 7	Description
11	LAW_CAT_CD	string, 5101231	Level
12	JURIS_DESC	string, 5101231	Department
13	BORO_NM	string, 5100768 null, 463	Borough
14	ADDR_PCT_CD	int, 5100841 null, 390	Code
15	LOC_OF_OCCUR_DESC	string, 3974103 null, 1127128	Location
16	PREM_TYP_DESC	string, 5067952 null, 33279	Place
17	PARKS_NM	string, 7599 null, 5093632	Park name
18	HADEVELOPT	string, 253205 null, 4848026	NYCHA name
19	X_COORD_CD	int, 4913085 null, 188146	X-coordinate
20	Y_COORD_CD	int, 4913085 null, 188146	Y-coordinate
21	Latitude	decimal, 4913085 null, 188146	Latitude
22	Longitude	decimal, 4913085 null, 188146	Longitude
23	Lat_Lon	string, 4913085 null, 188146	Latitude & Longitude

1.2 Data Quality Discussion

1.2.1 Problems

1. Are there non-empty values that represent missing data (e.g., NULL, N/A, UNSPECIFIED, TBA, (999)999-9999) If so, how many in each column?
In this crime dataset, there is not such value.

2. Are there different kinds of values in the same column (e.g., integers and strings)?
No, there are not.

3. Are there suspicious or invalid values in columns? (e.g., a negative value in a price field; for spatial data, coordinates outside the city perimeter; an invalid zipcode)
Yes, there are. Details are shown as following table.

4. Are there surprising (or suspicious) events in the data? (e.g., a day with too few taxi trips or too few noise complaints). Note that you may have to aggregate the data in different ways to search for these events.
Yes, there are. For instance, there are time marked as 24:00:00, which is invalid.

1.2.2 Data Cleaning

We treat a record legal as long as all of its 24 columns are valid. Especially, when tiny percentage of data is null, we will treat it as invalid, and delete the record in the future, such as column 1, 2, 8, 9, 10, 13, 14, 16, 20, 21, 22, 23 (we count from column from 0). For other columns containing large number of null values, as column 3, 4, 15, 17, 18, we treat null values as valid input, but still we will mark it as “null”. The script for generating clean data can be found on DataCleaning/dumbo_clean_data. After deletion, we have 4636592 legal records, which is 90.89% of raw data.

table 1.1 data assumption and results

	Column name	Assumptions and Results
0	CMPLNT_NUM	The base type is int. All data are valid.
1	CMPLNT_FR_DT	There are 655 null values, which will be deleted. Also, we will focus on the data later than 2006, thus year before 2006 will be deleted.
2	CMPLNT_FR_TM	There are 48 null values, we will delete them.
3	CMPLNT_TO_DT	There are 27.28% null values. We will not delete the null values
4	CMPLNT_TO_TM	There are 27.28% null values. We will not delete the null values
5	RPT_DT	All cells are valid.
6	KY_CD	A cell in column 6 corresponds to cell in 7. We only allow one description of one code. Thus we will delete the less part.
7	OFNS_DESC	As described above

8	PD_CD	A cell in column 8 corresponds to cell in 9. We only allow one description of one code. Thus we will delete the less part.
9	PD_DESC	As described above
10	CRM_ATPT_CPTD_CD	There are only 4574 null values, we will delete them.
11	LAW_CAT_CD	All cells are valid. MISDEMEANOR: 2918574 VIOLATION: 615234 FELONY: 1567423
12	JURIS_DESC	All cells are valid
13	BORO_NM	The values are BRONX: 1103514 STATEN ISLAND: 243790 BROOKLYN: 1526213 NULL: 463 MANHATTAN: 1216249 QUEENS: 1011002. We will delete null values.
14	ADDR_PCT_CD	We will delete null values
15	LOC_OF_OCCUR_DESC	There are 22.1% null values. We will not delete them.
16	PREM_TYP_DESC	We will not delete the null values.
17	PARKS_NM	There are 99.85% null values. We will not delete the null values.
18	HADEVELOPT	There are 95.04% null values.We will not delete the null values.
19	X_COORD_CD	There are 3.69% null values. We will delete them.
20	Y_COORD_CD	There are 3.69% null values. We will delete them.
21	Latitude	There are 3.69% null values. We will delete them.
22	Longitude	There are 3.69% null values. We will delete them.
23	Lat_Lon	There are 3.69% null values. We will delete them.

Moreover, we found that, in some cases, the end time is early than start time, corresponding to column 6, 7, 8, 9 which we will treat as invalid data.

1.2.3 Data Labeling

For each value of raw data, we will mark their type information, semantic information, and valid or not. For this part, the code is in DataCleaning/dumbo_cell_information.py. For instance, a data in column 1, “12/31/15”, will be marked as “12/31/15_Date_Date_valid”, while “12/32/15” will be marked as “12/32/15_Date_Date_invalid”

2. Data Extracting Description

2.1 Code for Extracting Data

All the data gathered by our group are stored in the file Big-Data-Project/DataStatistics. The data and the corresponding scripts to get the data are generally divided into three parts. Part1 contains the data and scripts concerning area(etc. Brooklyn). Part2 contains data and scripts concerning KYCD. Part3 contains data and scripts concerning the level of crime. For every python script, it has a output file with the same name but a suffix of '.out' in the results file. The script and data could be described with the table below. (The scripts in Big-Data_Project/former_version are the scripts which used as formerly. It exists only as a copy of the former code. You should not use these code when you reproduce the data.)

Part	Script	Output File	Description
Part 1	area_total_amount.py	area_total_amount.out	The total number of crime records in each area
	area_year_amount.py	area_year_amount.out	The number of crime records in each area in different years from 2006 to 2015
	area_month_amount.py	bronx_month_amount.out brooklyn_month_amount.out queens_month_amount.out manhattan_month_amount.out statenisland_month_amount.out	The number of crime records of every month in different years from 2006 to 2015 in each area
	area_level_amount.py	area_level_amount.out	The number of crime records of different level in each area
	area_top5_KYCD.py	bronx_top5_KYCD.out brooklyn_top5_KYCD.out queens_top5_KYCD.out manhattan_top5_KYCD.out statenisland_month_amount.out	The top 5 KYCD which have most crime records in each area
Part 2	KYCD_datime_amout.py	KYCD_daytime_amount.out	The number of crime records of each KYCD at different time period of a day. (Daytime: 6AM - 5PM, Night: 5PM - 11PM, Midnight: 11PM - 12AM + 12AM - 6AM)
	KYCD_report_amount.py	KYCD_report_amount.out	The number of crime records of each KYCD grouped by the span of report time and happened time. (etc. how many

			records are reported at the same day as it happened)
	KYCD_status_amount.py	KYCD_status_amout.out	The number of crime records of different status(attempted or completed) of each KYCD
	KYCD_total_amount.py	KYCD_total_amount.out	The number of crime records of each KYCD
	KYCD_year_amount.py	KYCD_year_amount.out	The number of crime records of each KYCD in every year from 2006 to 2015
	KYCD_month_amount.py	341_month_amount.out 344_month_amount.out 578_month_amount.out	The number of crime records of the three KYCD with largest amount in every month of different years from 2006 to 2015
	KYCD_weekday_amount.py	KYCD_weekday_amount.pt	The number of crime records of different day in a week of each KYCD
Part 3	level_status_amount.py	level_status_amount.out	The number of crime records of different status(attempted or completed) of each level of crime
	level_total_amount.py	level_total_amount.out	The number of crime records of each level of crime
	level_year_amount.py	level_year_amount.out	The number of crime records in different year from 2006 to 2015 of each level of crime
	level_month_amount.py	felony_month_amount.out misdemeanour_month_amount.out violation_month_amount.out	The number of crime records in different month of years from 2006 to 2015 of each level of crime

2.2 Output Format of Each Output File

The output format of each script could be described with the table below.

Part	Output File	Format (whitespace is \t)
Part 1	area_level_amout.out	(area FELONY,MISDEMEANOR,VIOLATION)
	area_total_amount.out	(area total)
	area_year_amount.out	(area <2006,2006,2007,...,2014,2015)
	bronx_month_amount.out	(year 1,2,3,...,10,11,12)
	brooklyn_month_amount.out	
	queens_month_amount.out	
	manhattan_month_amount.out	
	statenisland_month_amount.out	
	bronx_top5_KYCD.out	(KYCD amount)
	brooklyn_top5_KYCD.out	
	manhattan_top5_KYCD.out	
	queens_top5_KYCD.out	
	statenisland_month_amount.out	
Part 2	KYCD_daytime_amount.out	(KYCD daytime,night,midnight)
	KYCD_report_amount.out	(KYCD =0,<=7,<=30,<=365,>365 day)
	KYCD_status_amount.out	(KYCD COMPLETED,ATTEMPTED)
	KYCD_total_amount.out	(KYCD total)
	KYCD_weekday_amount.out	(KYCD Mon,Tues,Wed,Thurs,Fri,Sat,Sun)
	KYCD_year_amount.out	(KYCD <2006,2006,2007,...,2014,2015)
	341_month_amount.out	(year 1,2,3,...,10,11,12)
	344_month_amount.out	
	578_month_amount.out	
Part 3	level_status_amount.out	(level COMPLETED,ATTEMPtED)
	level_total_amount.out	(level total)
	level_year_amount.out	(level <2006,2006,2007,...,2014,2015)
	felony_month_amount.out	(year 1,2,3,...,10,11,12)
	misdemeanour_month_amount.out	

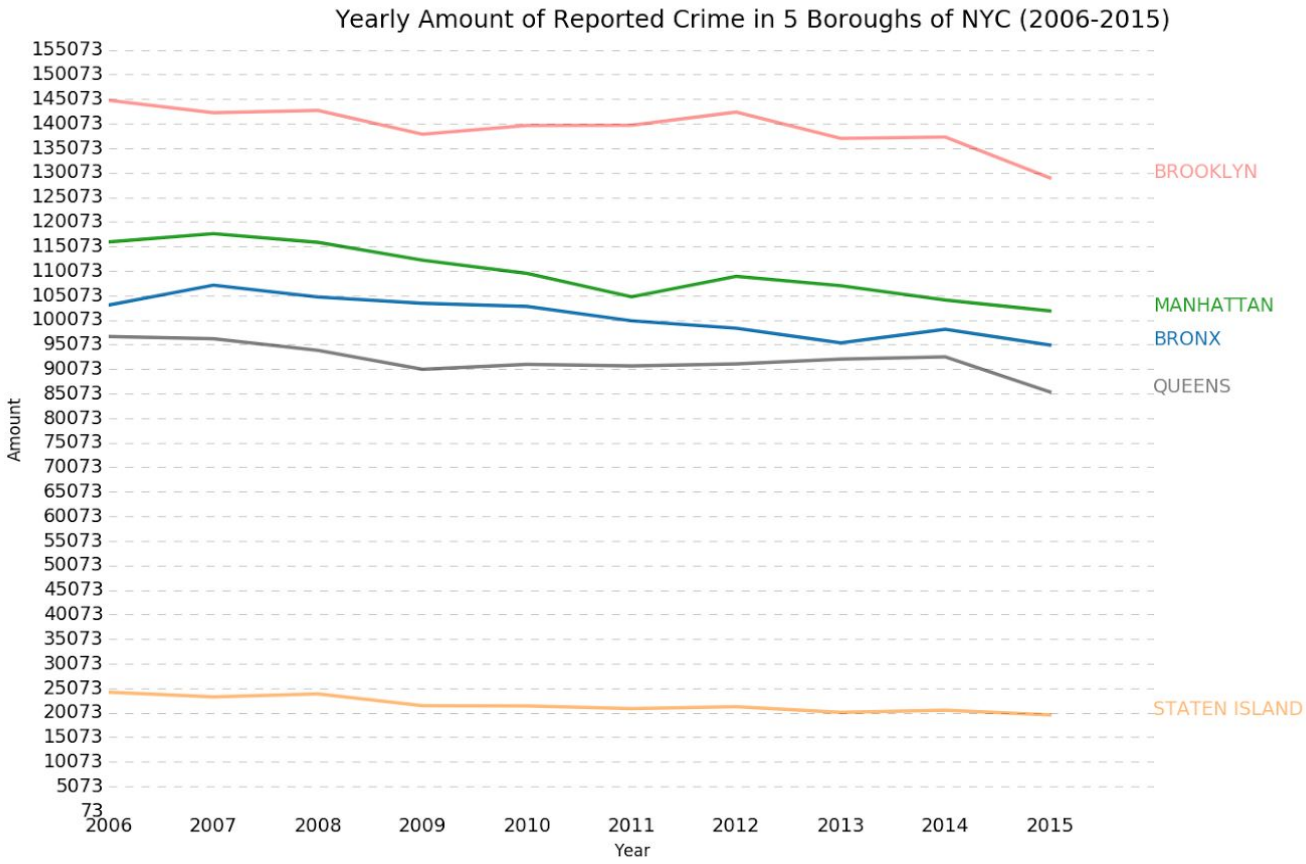
	violation_month_amount.out	

3. Data Visualization

All scripts in this section are stored in Big-Data-Project/DataVisualization/Plot_Code. Figures are stored in Big-Data-Project/DataVisualization/Plots. Each figure below is specified the script used to draw it.

3.1 Overview

From 2006 to 2015, there have been over 5 million crime reported in different boroughs of New York City. Among different boroughs, Brooklyn always has the most crime number, with Manhattan to be the second most criminal-gathering area. The Staten Island has significantly lower crime amount compared with other four borough. We found that the total amount of reported crime has been gradually decreasing during last 10 years in all boroughs, however, to better understand those data, we must dig deeper and try to inspect from different aspects. *[Figure is drew by script: area_year_amount.py]*



3.2 Crime Type and Level Aspect

3.2.1 Crime Type

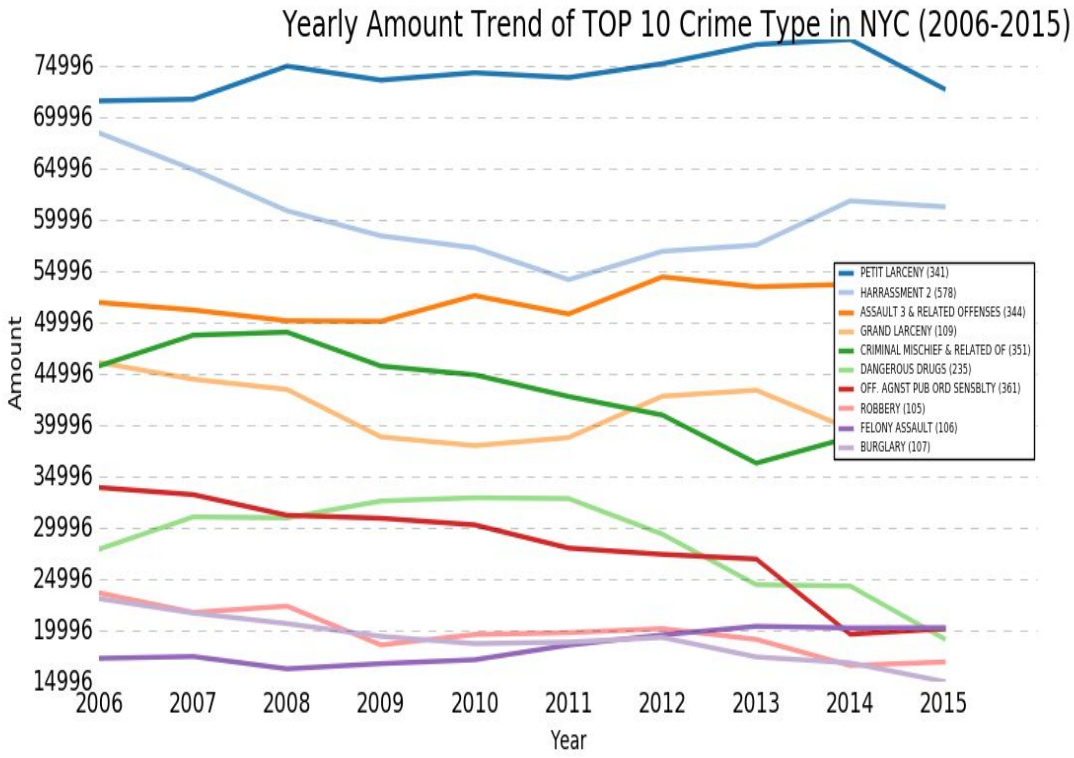
Table 3.0 counts the 10 years data and lists the top 10 crime types of the most happened crime, their amount and percentage of total amount. Among all the data entries, crime entries with these 10 types take up over 77% of the total amount. The first three crime types are PETIT LARCENY, HARRASSMENT 2 and ASSAULT 3 & RELATED OFFENSES, which accounts for nearly 40% percent of total crime. *[Table is drew by directly copying from output file.]*

Table 3.0

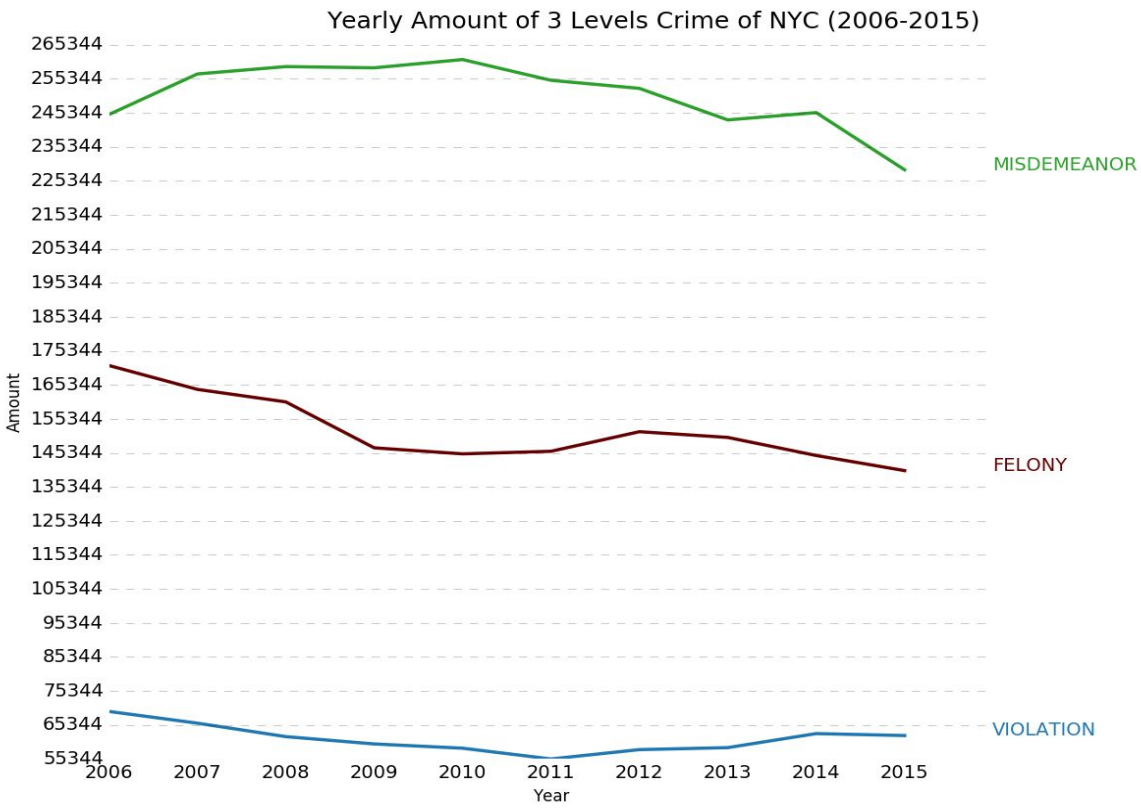
Top 10 Crime Types in NYC (2006-2015)		
Crime Type (KY_CD)	**Amount**	**Percentage(%)**
PETIT LARCENY (341)	612853	16.54409939
HARRASSMENT 2 (578)	433013	11.68927966
ASSAULT 3 & RELATED OFFENSES (344)	386961	10.44609595
GRAND LARCENY (109)	311974	8.421805656
CRIMINAL MISCHIEF & RELATED OF (351)	308793	8.335933873
DANGEROUS DRUGS (235)	207313	5.596459307
OFF. AGNST PUB ORD SENSBLTY (361)	196447	5.303129285
ROBBERY (105)	140670	3.797417098
FELONY ASSAULT (106)	139382	3.762647259
BURGLARY (107)	134921	3.642221598
MISCELLANEOUS PENAL LAW (126)	84904	2.292001857
OFFENSES AGAINST PUBLIC ADMINI (359)	78467	2.118233649
GRAND LARCENY OF MOTOR VEHICLE (110)	66157	1.785922535
DANGEROUS WEAPONS (236)	54264	1.464868425
CRIMINAL MISCHIEF & RELATED OF (121)	53386	1.441166625
INTOXICATED & IMPAIRED DRIVING (347)	52380	1.414009438
CRIMINAL TRESPASS (352)	46284	1.249446598
VEHICLE AND TRAFFIC LAWS (348)	44473	1.200558261
DANGEROUS DRUGS (117)	40812	1.101728774
THEFT-FRAUD (112)	40314	1.088285156
DANGEROUS WEAPONS (118)	38187	1.03086633

3.2.2 Crime Level

However, those 10 crime types appear different trend when we inspecting them in year scale. Figure below shows the yearly amount trends of them. Although most of types present downtrends, crimes of ASSAULT 3 & RELATED OFFENSES and DANGEROUS WEAPONS has been gradually increasing over years. As with HARRASSMENT 2, it has decreased sharply from 2006 to 2011, but began with 2012, it starts to appear an ascending trend. *[Figure is drew by Top10KYCD_year_amount.py]*



Crimes are classed 3 levels of offense: felony, misdemeanor and violation, beyond using KY_CD to define specific crime categories,. From figure below we can see that most of the crimes fall in misdemeanor level, then felony level and violation level crimes are quite fewer. While the amount of misdemeanor and felony appears a trend to decrease from 2011, violation level crimes began to uptrend. *[Figure is drew by level_year_amount.py]*



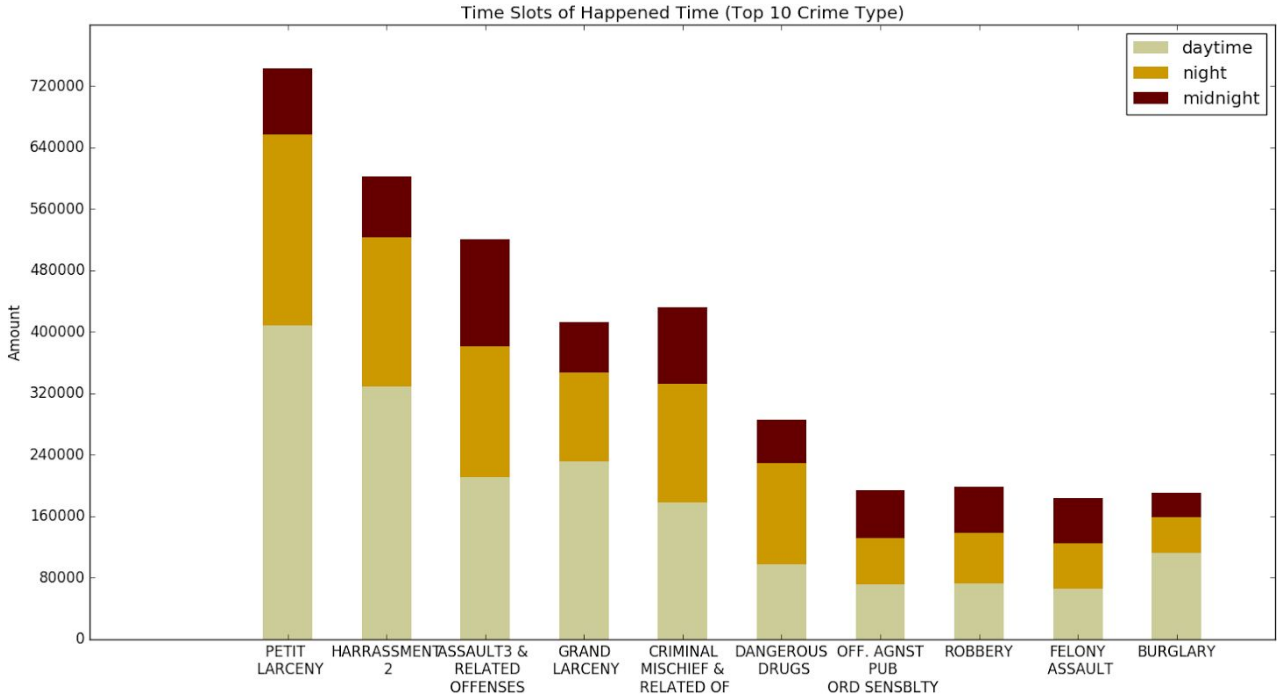
3.3 Time Aspect

3.3.1 Daily Dimension

According to daily routines, we roughly divide a day into three different time slot:

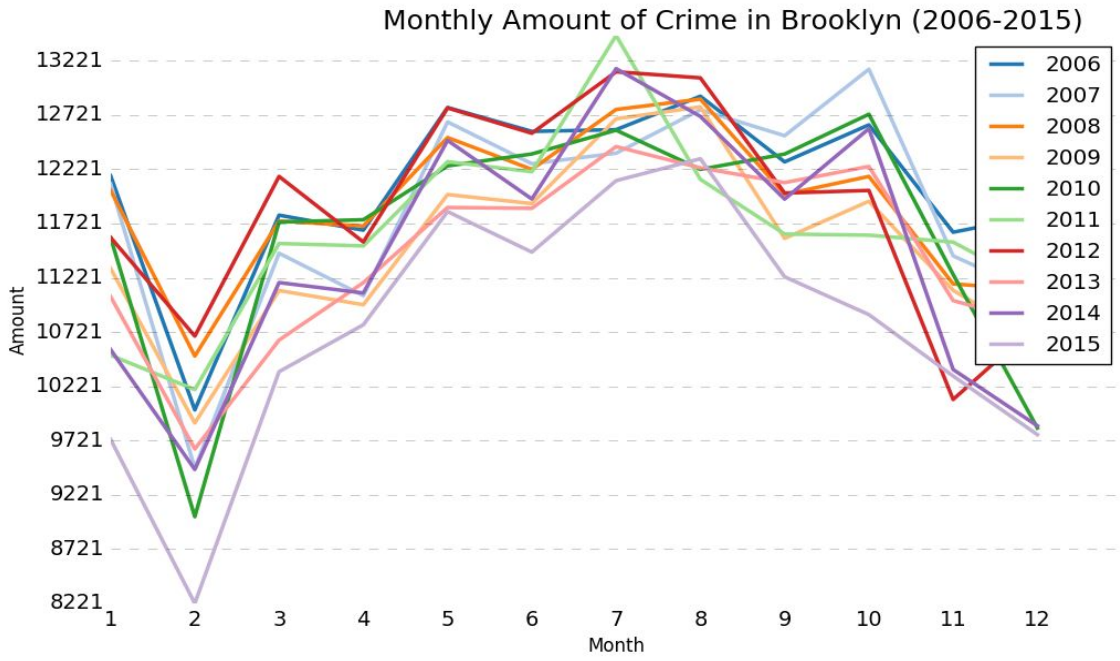
- Daytime: 6AM -- 5PM (11 hours totally). It's the time slot people usually go out to work/study and stay awake.
- Night: 5PM -- 11PM (6 hours totally). It's the time slot people usually get off work/school and come back home.
- Midnight: 11PM -- 12AM and (next day) 12AM --6AM (7 hours totally). It's the time slot people usually fall in sleep.

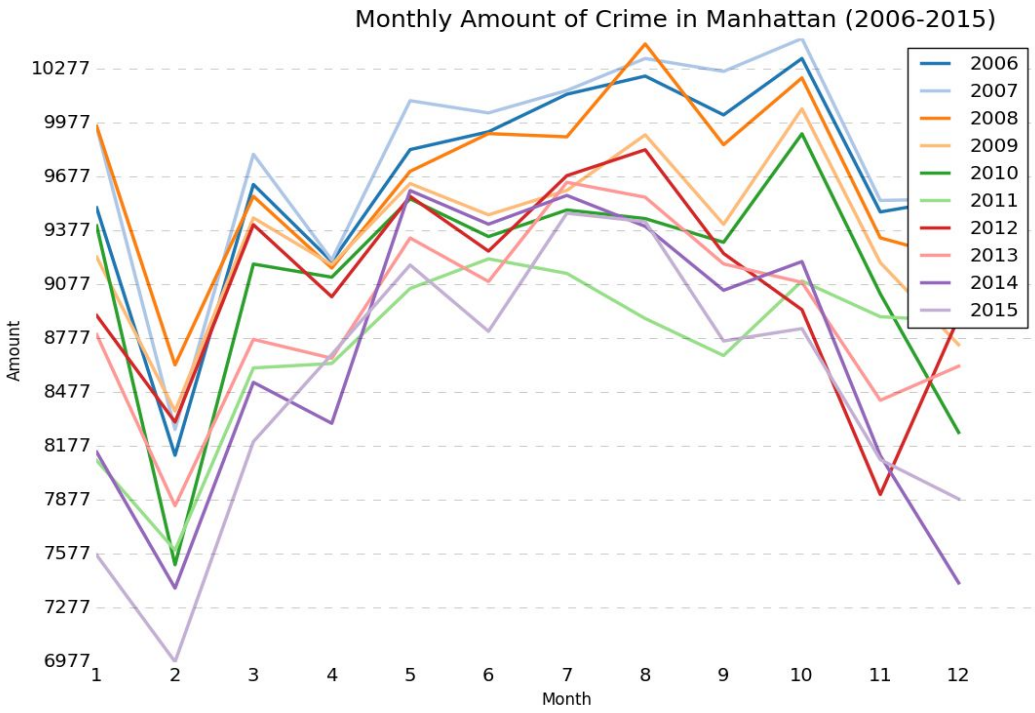
For the top 10 crime type, we inspect their happened time in time slots defined above. PETIT LARCENY, HARRASSMENT 2, GRAND LARCENY and BURGLARY have more than 50% of crime happening in daytime, which is reasonable since daytime slot takes up nearly 50% of a day. However, some crime types present a visibly prone to happen at night and midnight. More than 65% percent of DANGEROUS DRUGS, OFF. AGNST PUB ORD SENSBLTY, ROBBERY and FELONY ASSAULT happened at night and midnight. *[Figure is drew by Top10KYCD_daytime_amount.py]*



3.3.2 Monthly Dimension

Monthly trend of crime in each year is discussed in this section. We pick Brooklyn and Manhattan, which are the two boroughs with the most number of crime, to draw their monthly amount of crime from 2006 to 2015. Interestingly, monthly amount shows the same variation in both boroughs. In each of ten year, February is the trough with the significantly lower crime amount that all of other months. The downtrend usually starts from September of last year and last until February. From February, the crime amount begins to climb and reach the peak in July or August. *[Figures are drew by brooklyn_month_amount.py and manhattan_month_amount.py]*

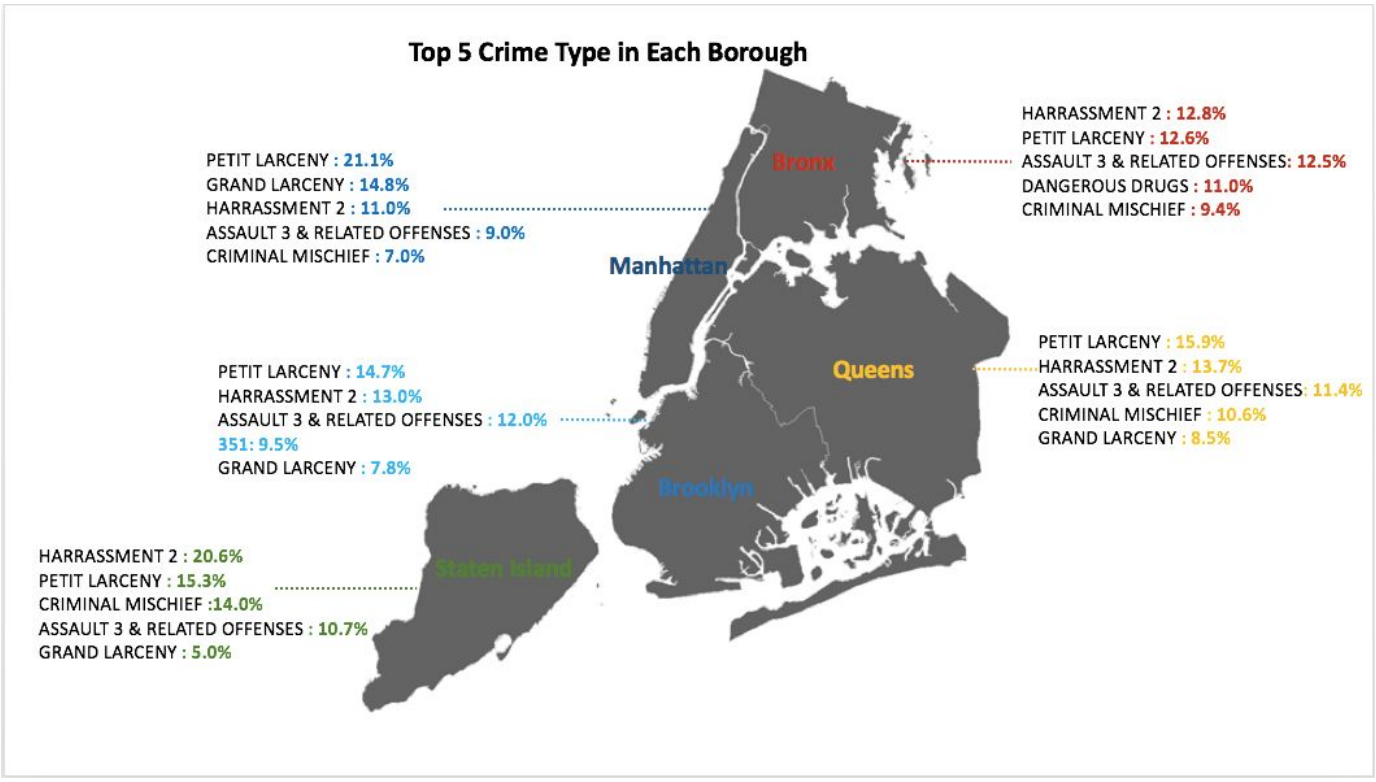




3.4 Borough Aspect

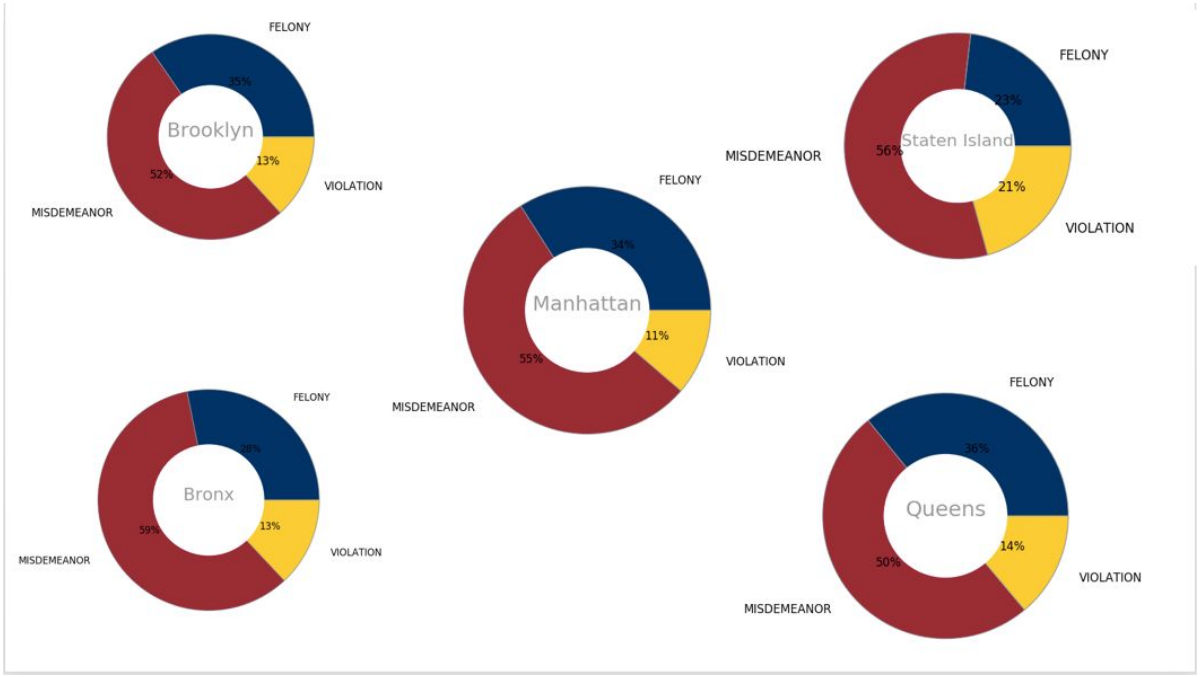
3.4.1 Top 5 Crime Type in Each Borough

For each borough of New York City, 5 crime types with the most amount are given below. Among 70 crime types, these 5 types account for more than 57% percent of total crime in all 5 boroughs. For Manhattan, Brooklyn, Queens and Staten Island, though with slightly different percentage, top 5 crime type are the same: PETIT LARCENY, HARRASSMENT 2, ASSAULT 3 & RELATED OFFENSES, GRAND LARCENY and CRIMINAL MISCHIEF & RELATED OF. In Bronx, DANGEROUS DRUGS appears in top 5 crime type and ranks fourth, which is not shown in top 5 of other boroughs. [Figure is drew by hands]



3.4.2 Crime Level Percentage in Each Borough

All of five borough have 50% - 59% percent of crime falling in MISDEMEANOR level. In five borough, Queens has the most FELONY level crime proportion -- 36%, while only 23% crimes are FELONY in Staten Island. Although Staten Island has fewer FELONY crime, it has the most VIOLATION crime -- it is the only borough with more than 20% VIOLATION level crime, which are all below 14% in other boroughs. *[Figures are drew by brooklyn/manhattan/bronx/queens/statenisland_3level.py]*



4. Data Exploration

After finishing exploring the NYC crime data in sections above, we've observed some interesting facts this dataset reveals, like the crime amount reaches its lowest point in February of each year; why brooklyn always has the most crime amount among five boroughs etc. Such observations encourage us to further dig the possible correlation the crime behavior has with other features.

4.1 Experimental Setup

In this section, we conduct correlation analysis and regression between crime dataset and other datasets we found. Spark is used in cleaning and exacting new data. Pearson correlation and linear regression model are mainly used in this section, and related analysis is done on our local PC. In our experiments below, we consider that two variables are positive/negative correlated if their correlation coefficients are in [0.5,1] or [-1,-0.5], otherwise they are regarded as irrelevant.

4.2 Hypotheses, Analysis and Discussion

4.2.1 Crime Number With Census

In this part, we discuss the relation between crime number and census data. Intuitively, we believe that with more population, crime number will be bigger, as if the crime rate is certain. Code can be found at ***DataCorrelationAnalysis/crime_census.py.*** Census data can be found at https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Census-Tracts/37cg-gxjd/data Thus we make the hypothesis that, the **crime number has positive correlation with census**.

table 4.1.1 crime data with census

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Crime Amount	484876	486652	481191	465156	464569	456062	462190	451760	452786	430955
Census	7993906	8013775	8068195	8131574	8192426	8287000	8365069	8436047	8495194	8550405

Using pearson analysis, we calculate the correlation parameter as a result of 0.86, which indicates strong correlation. **Thus our hypnosis is right**. And the census date could be used in the future regression analysis. For example, we could estimate the crime number once we have census data.

4.2.2 Crime Levels with Census

Based on the result from the first part, we make a further hypnosis that **crime data is has positive correlation with each crime type**, including violation, felony, misdemeanor. We believe that crime that human beings made have certain distribution patterns. Code can be found at ***DataCorrelationAnalysis/three_level.***

table 4.1.2 Crime Data With Crime Levels

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015

Crime Amount	484876	486652	481191	465156	464569	456062	462190	451760	452786	430955
Violation Amount	69324	65869	61916	59758	58513	55344	58098	58679	62809	62233
Felony Amount	171033	164020	160339	146827	145063	145815	151572	149890	144582	140135
Misdemander Amount	244849	256763	258936	258571	260993	254903	252520	243256	245395	228587

As table 4.1.2 shows, we have crime amount and all three level crimes. We found three types crime pearson correlation coefficients are 0.49, 0.87, 0.65 respectively. **That means violation is not so strongly linearly correspond to the crime data, while other two are opposite.** By checking the data again, we find that the number of violation is much smaller, which demonstrates the possibility that violation does not happen often, and has some randomness.

4.2.3 Crime per capita With Graduation Outcomes

Intuitively, a person with education level has the tendency to behave well and abide the law. So we make the **hypothesis** that : **crime per capita has negative correlation with graduation rates.** Graduation Outcomes dataset is available at:
https://data.cityofnewyork.us/Education/Graduation-Outcomes-Classes-Of-2005-2010-By-Boroug/avir-tzek
Graduates are defined as those students earning either a Local or Regents diploma and exclude those earning either a special education (IEP) diploma or GED. Crime per capita is computed from the dataset of our NYC crime dataset. We take the graduation outcome data that students graduated in between 2006 and 2010 in five boroughs and crime per capita every year respectively.

4.2.3.1 Crime per capita With Total Graduation

Our **hypothesis is crime per capita has negative correlation with total graduation outcome.** we compute the correlation coefficient between the total graduation outcome from 2006 to 2010 in each borough and crime per capita in corresponding borough. Entire data table is listed below in table 4.2.3.1. Spark script for extracting data locates at ***DataCorrelationAnalysis/Graduation/2006-2010_borough_total_graduation.py*** Script for Pearson correlation analysis locates at ***DataCorrelationAnalysis/Graduation/crime_per_capita_total_graduation.py***

Pearson correlation is computed as **-0.55**, which is regarded as negative correlated. So we draw the conclusion that our hypothesis is **correct**.

Table 4.2.3.1Graduation Outcome and Crime per capita in 5 borough (2006-2010)

Borough - Year	Total Graduation Outcome (Total Grads of cohort)	Crime per capita in each year
Bronx-2006	0.443	0.076489211
Bronx-2007	0.469	0.079140006
Bronx-2008	0.519	0.076827225
Bronx-2009	0.542	0.075183414

Bronx-2010	0.547	0.074071646

Brooklyn-2006	0.497	0.059445876
Brooklyn-2007	0.519	0.058284357
Brooklyn-2008	0.55	0.058023599
Brooklyn-2009	0.564	0.055435612
Brooklyn-2010	0.588	0.055648748

Manhattan-2006	0.575	0.073489501
Manhattan-2007	0.549	0.074401702
Manhattan-2008	0.58	0.073022302
Manhattan-2009	0.62	0.070890364
Manhattan-2010	0.634	0.068965995

Queens-2006	0.54	0.044485345
Queens-2007	0.568	0.044205551
Queens-2008	0.583	0.042799515
Queens-2009	0.608	0.040601831
Queens-2010	0.637	0.040720209
Staten Island-2006	0.658	0.053005287
Staten Island-2007	0.667	0.050672045
Staten Island-2008	0.673	0.051552617
Staten Island-2009	0.695	0.046048419
Staten Island-2010	0.727	0.045668537

4.2.3.2 Crime per capita With Female Graduation

Luckily, this graduation dataset not only provides a total graduation outcome, but also counting the graduation rate in other perspectives. In this subsection and the one below, we are interested in what relation the female and male graduation outcome have with crime per capita. Our **hypothesis is that: crime per capita has negative correlation with female graduation outcome.** Due to limited space, here we only list data used for bronx in table 4.2.3.2. Five boroughs data are all used in our analysis. Spark script for extracting data locates at ***DataCorrelationAnalysis/Graduation/2006-2010_female_graduation.py***

Script for Pearson correlation analysis locates at
DataCorrelationAnalysis/Graduation/crime_per_capita_female_graduation.py

Pearson correlation coefficient between female graduation outcome and crime per capita is **-0.61** which indicates our hypothesis is **correct.**

Table 4.2.3.2 Female Graduation Outcome and Crime per capita in 5 borough (2006-2010)

Borough - Year	Female Graduation Outcome(Total Grads of cohort)	Crime per capita in each year
Bronx-2006	0.501	0.076489211
Bronx-2007	0.523	0.079140006
Bronx-2008	0.577	0.076827225
Bronx-2009	0.608	0.075183414
Bronx-2010	0.594	0.074071646

4.2.3.3 Crime per capita With Male Graduation

Same as the subsection above, we here hope to find if there is correlation between male graduation and crime. **Our hypothesis is crime per capita has negative correlation with male graduation outcome.** Only the bronx data is listed below as an example for our data format, in our analysis datas of five boroughs are all used. Spark script for extracting data locates at ***DataCorrelationAnalysis/Graduation/2006-2010_male_graduation.py*** Script for Pearson correlation analysis locates at ***DataCorrelationAnalysis/Graduation/crime_per_capita_male_graduation.py***

Pearson correlation coefficient between female graduation outcome and crime per capita is **-0.55** which indicates our hypothesis is **correct.**

Table 4.2.3.3 Male Graduation Outcome and Crime per capita in 5 borough (2006-2010)

Borough - Year	Male Graduation Outcome(Total Grads of cohort)	Crime per capita in each year
Bronx-2006	0.388	0.076489211
Bronx-2007	0.416	0.079140006
Bronx-2008	0.461	0.076827225
Bronx-2009	0.481	0.075183414
Bronx-2010	0.504	0.074071646

4.2.4 Crime per Month With Average Monthly Temperature:

One of the interesting observations from the NYC crime data is that each year the monthly crime amount reaches the lowest point in February and then increases until August after which the amount starts to decrease until February of the next year. Based on such fact **we suppose that crime amount per month has positive correlation with average month temperature.** Weather data is publicly available on http://www7.ncdc.noaa.gov/CDO/dataproduct Here we use the daily temperatures in JFK weather station and compute average month temperature for each month from 2006 to 2015. Spark script for extracting data and computing temperature locates at ***DataCorrelationAnalysis/Weather/2006-2015_monthly_temp.py*** Script for Pearson correlation analysis locates at ***DataCorrelationAnalysis/Weather/manhattan_crime_temp.py and brooklyn_crime_temp.py.***

Based on the same temperature data, we choose two boroughs Manhattan and Brooklyn to compute the Pearson correlation for each. Coefficient between average monthly temperature and crime per month in Manhattan is **0.60** and in Brooklyn such coefficient is **0.78**. These two correlation coefficient is pretty high and prove that our hypothesis is **correct**. Table 4.2.4 lists temperature data we used in 2006 and crime amount per month in Manhattan in 2006. Notice that in our experiment we use the 9 years data from 2006 to 2015.

Table 4.2.4 Average Monthly Temperature and Crime per month In Manhattan (2006)

Year - Month	Average Monthly Temperature (F)	Crime per month
200601	39.03870968	9503
200602	34.71071429	8126
200603	41.20967742	9633
200604	52.44	9204
200605	60.16451613	9827
200606	68.68	9926
200607	76.36774194	10135
200608	75.78387097	10236
200609	66.72333333	10020
200610	56.40645161	10334
200611	50.98	9480
200612	43.73225806	9555

4.2.5 Crime per capita With Individual Income Yearly

It's very natural to associate the amount of crime in a particular area with its local economic status. During the process of analysing the dataset we observe that the amount of crime within a district usually has a trend towards going down with the year increasing gradually. So based on this observation and the association about economy and crime, we **suppose that crime amount per capita has a positive correlation with income of an entire year per capita.** Because we suppose that in general the economic status of an area is somehow becoming better annually. The income data is published on https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t&keepList=t#. And the cencus data could be see at https://factfinder.census.gov/faces/nav/jsf/pages/error.xhtml. (As we use the advanced search in this website, which use a cookie to contains the key word of one search. Open the URL directly may not show the dataset as we search. To find the data the only way is to do the search again, by the key word like 'census' and 'income' and its corresponding area). To get a crime amount of per capita in the main area of New York, we use the former data named **DataStatistics/part1/area_year_amount.out** and the census data on the website. Script for Pearson correlation analysis locates at ***DataCorrelationAnalysis/Income/income_crime_amount_per_capita.py***.
Based on the data we get, we choose the year between 2009 to 2015 to compute the Pearson correlation for the main area in New York. The data is shown in Tables 4.2.5.1 - 4.2.5.5

Table 4.2.5.1 Crime per capita and Income of an entire year per capita in main area of Bronx

Area - year	Income per capita	Crime per capita
BRONX-2009	17301	0.075
BRONX-2010	17575	0.074
BRONX-2011	17992	0.071
BRONX-2012	18049	0.070
BRONX-2013	18171	0.067
BRONX-2014	18269	0.068
BRONX-2015	18456	0.065
Pearson Correlation: -0.96		

Table 4.2.5.2 Crime per capita and Income of an entire year per capita in main area of Brooklyn

Area - year	Income per capita	Crime per capita
BROOKLYN-2009	22959	0.055
BROOKLYN-2010	23605	0.0.056
BROOKLYN-2011	24398	0.055
BROOKLYN-2012	24649	0.055

BROOKLYN-2013	25289	0.052
BROOKLYN-2014	25932	0.052
BROOKLYN-2015	26774	0.048
Pearson Correlation: -0.960		

Table 4.2.5.3 Crime per capita and Income of an entire year per capita in main area of Manhattan

Area - year	Income per capita	Crime per capita
MANHATTAN-2009	60047	0.071
MANHATTAN-2010	59149	0.069
MANHATTAN-2011	61290	0.065
MANHATTAN-2012	61951	0.067
MANHATTAN-2013	62498	0.066
MANHATTAN-2014	63610	0.064
MANHATTAN-2015	64993	0.062
Pearson Correlation: -0.90		

Other area are similar to these these areas, you could see the outcome in our file. As what we could see, the pearson Correlation of these area are quite high, almost above -0.9.

4.2.6 Vehicles collisions amount with Crime amount

Another interesting observation in this project is that we notice vehicle collisions are usually the top 5 KYCD in many area. So we want to know whether these is a relationship between crime and vehicles collisions. The data we use still from https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml and the crime per capita data is similar to the former part. The script for Pearson Correlation of this part is in ***DataCorrelationAnalysis/Income/crash_crime_amount_per_capita.py***. The result is shown in the tables below.

Table 4.2.6 Vehicle Collision with Crime amount In Manhattan (20015)

Year - Month	Vehicle Collisions	Crime
201501	2876	7571
201502	2670	6977
201503	3396	8202
201504	3523	8687
201505	3882	9185

201506	3731	8816
201507	3845	9473
201508	3707	9426
201509	3640	8762
201510	4026	8831
201511	3546	8102
201512	3674	7883
Pearson Correlation: 0.844		

We could find that all these areas are have pearson correlation above 0.7 between the amount of vehicle collisions and crime. So it's positive correlation.

4.2.7 Conclusion

In the above subsection, we have made different hypotheses to find correlation with NYC crime behavior in various aspects, such as census, graduation outcome, temperature, individual income and vehicles collisions. Eight hypotheses are raised, analyzed and proved by us. Under our assumption of that we believe two variables are correlated if they have correlation coefficient in [0.5,1] or [-1, -0.5], we have proved seven out of our eight hypotheses are correct. To give a intuitive feeling, we summarize our conclusion in this correlation table below:

Table 4.2.7 Correlation Table

Relationship	Correlation Coefficient	Conclusion
(Total crime amount, Census) from year 2006 to 2015, all boroughs	0.86	Positive correlation
(Total crime amount , Violation amount) from year 2006 to 2015, all boroughs	0.49	Irrelevance
(Total crime amount , Felony amount) from year 2006 to 2015, all boroughs	0.87	Positive correlation
(Total crime amount , Misdemander amount) from year 2006 to 2015, all boroughs	0.65	Positive correlation
(Crime per capita, Total Graduation Outcome) from year 2006 to 2010, all boroughs	-0.55	Negative correlation
(Crime per capita, Female Graduation Outcome) from year 2006 to 2010, all boroughs	-0.61	Negative correlation
(Crime per capita, Male Graduation Outcome) from year 2006 to 2010, all boroughs	-0.55	Negative correlation

(Crime amount monthly, Average temperature monthly) from year 2006 to 2015, for manhattan	0.60	Positive correlation
(Crime amount monthly, Average temperature monthly) from year 2006 to 2015, for brooklyn	0.78	Positive correlation
(Crime per capita, Individual Income yearly) from year 2009 to 2015	-0.95	Negative correlation
(Vehicles collisions amount, Cime amount), average	0.85	Positive correlation

4.3 Regression analysis

In this section, we make two regression analysis between crime amount and those highly correlated features presented in 4.2.7. Linear regression model is used in all subsections below and all experiments are conducted on local PC.

4.3.1 Crime per month in Manhattan, Average Monthly Temperature and Census

Notice that in 4.2.7 the average monthly temperature and census features present high correlation with crime amount. So here we try to use linear regression to model proper relation between crime per month in Manhattan, average monthly temperature and census. Data in these three dimensions used to model is from 2006-2015. Script for Pearson correlation analysis locates at
DataCorrelationAnalysis/Regression/regression_manhattan_month.py

Table 4.3.1 lists data we used in 2006 as a sample. Our model is that:
Y = Crime per month
X1 = Average Monthly Temperature
X2 = Census in that year

After training we get the fitted model : **Y = 24.10*X1 - 0.04*X2, R-square is 0.44 which indicates a well fitted model.**

Table 4.3.1 Data Sample of year 2006 in Manhattan

Month	Crime per month	Average Monthly Temperature	Census in that year
1	9503	39.03870968	1578171
2	8126	34.71071429	1578171
3	9633	41.20967742	1578171
4	9204	52.44	1578171
5	9827	60.16451613	1578171
6	9926	68.68	1578171
7	10135	76.36774194	1578171

8	10236	75.78387097	1578171
9	10020	66.72333333	1578171
10	10334	56.40645161	1578171
11	9480	50.98	1578171
12	9555	43.73225806	1578171

4.3.2 Crime per year in Manhattan, Individual Income and Census

Here we try to use linear regression to model proper relation between crime per year in Manhattan, individual income and census. Data in these three dimensions used to model is from 2009-2015 (since we only have complete individual income from 2009-2015). Script for Pearson correlation analysis locates at ***DataCorrelationAnalysis/Regression/regression_manhattan_income.py***

Table 4.3.2 lists data we used in 2009-2015. Our model is that:
Y = Crime per year
X1 = Individual income yearly
X2 = Census in that year

After training we get the fitted model: **Y = (-6.73e-07)*X1 + (-6.26e-08)X2, R-square is 0.84 which indicates a well fitted model.**

Table 4.3.2 Complete data used in 2009-2015

Year	Crime per year	Individual income yearly	Census in that year
2009	0.070890364	60047	1583431
2010	0.068965995	59149	1588609
2011	0.065094866	61290	1609789
2012	0.067053976	61951	1624885
2013	0.065618267	62498	1631375
2014	0.063616471	63610	1636966
2015	0.061972566	64993	1644518

5. Individual Contributions

5.1 Chong han

In the first part of this project, I wrote all the code of spark to extract the date we wanted from dumbo. These code are stored in the file named ***DataStatistics/***. And I also write the shell script name ***execute.sh*** in ***DataStatjstjcs/*** to make execute and debug efficiently.

In second part of this project, I analysed the Crime per capita and Income of an entire year as well as the Crash amount and Crime amount.

5.2 Mingzhong Dai

In the first part of the project, I was in charge of the data quality issue and data cleaning task, with assistance from YingbingWang. This part checks input format, and outputs clean data for later usage. Scripts locate in folder ***DataCleaning/***

In the second part, I analysed the Crime number with census,crime levels with census and regression analysis of section 4.2.1 and 4.2.2. Scripts locate in folder ***DataCorrelationAnalysis/***

5.3 Yingbing Wang

During the first part of project, I assisted the job of finding data quality issue and data cleaning. Data visualization and related analysis is done by me. Scripts and plots locate in folder ***DataVisualization/***

In second part, I was responsible for correlation models of crime per capita with graduation outcome(section 4.2.3) and crime per month with average temperature(section 4.2.4). Besides, I also conducted the experiments and analysis of linear regression section (section 4.3). Scripts locate in ***DataCorrealationAnalysis/Weather, DataCorrealationAnalysis/Graduation and DataCorrealationAnalysis/Regression.***

5.4 Cooperation part

Each of us makes contribution to report writing, including abstract, introduction and reference parts. Also we edit github files together.

6. Summary

In the project, we analysis the crime data in new york. Tools and language we used including spark, python, matplot, sklearn. We find interesting features of the opening data and would like to do more analysis in the future.

7. Reference:

7.1 Official Website of New York Crime Data:

https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

7.2 Column Description

Table 4.0 is column desciption, which can be found on:
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

7.3 Census data:

https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Census-Tracts/37cg-gxjd/data

7.4 Graduation Outcome data:

https://data.cityofnewyork.us/Education/Graduation-Outcomes-Classes-Of-2005-2010-By-Boroug/avir-tzek

7.5 Temperature Data:

http://www7.ncdc.noaa.gov/CDO/dataproduct

Table 7.1

Column	Description
CMPLNT_NUM(0)	Randomly generated persistent ID for each complaint
CMPLNT_FR_DT(1)	Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)
CMPLNT_FR_TM(2)	Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists)
CMPLNT_TO_DT(3)	Ending date of occurrence for the reported event, if exact time of occurrence is unknown
CMPLNT_TO_TM(4)	Ending time of occurrence for the reported event, if exact time of occurrence is unknown
5	Date event was reported to police
KY_CD(6)	Three digit offense classification code
OFNS_DESC(7)	Description of offense corresponding with key code
PD_CD(8)	Three digit internal classification code (more granular than Key Code)

PD_DESC(9)	Description of internal classification corresponding with PD code (more granular than Offense Description)
CRM_ATPT_CPTD_CD(10)	Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely
LAW_CAT_CD(11)	Level of offense: felony, misdemeanor, violation
JURIS_DESC(12)	Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external, like Correction, Port Authority, etc.
BORO_NM(13)	The name of the borough in which the incident occurred
ADDR_PCT_CD(14)	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC(15)	Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of
PREM_TYP_DESC(16)	Specific description of premises; grocery store, residence, street, etc.
PARKS_NM(17)	Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)
HADEVELOPT(18)	Name of NYCHA housing development of occurrence, if applicable
X_COORD_CD(19)	X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Y_COORD_CD(20)	Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)
Latitude(21)	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
Longitude(22)	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)