NYPD Incident Level Data Exploration

Minda Fang NetID: mf3308

Qiming Zhang NetID: qz718

Mingdi Mao NetID: mm8688

May 10, 2017

Contents

1 Abstract

The purpose of this report is for the sum up of NYPD incident level dataset from 2006 - 2015. This report includes data integrity issue detecting, data cleaning, data summary and data exploration. [1]

2 Introduction

This is NYU CS-9223 Big data final project report. We select NYPD Complaint Data Historic dataset(from 2006 - 2015) to investigate. To analyze the dataset , we first detect all possible data integrity issue. Then we do the data cleaning work by getting rid of all detected invalid data. After having summary work on the clean data, we collect all the useful result and display them in graphs. We finally have the data exploration work by finding interesting correlations between data and also by showing some anomalies data we find in our crime data analysis.

In the data integrity issue detection part, we analyzed the dataset in five major dimensionalities,null value problem, range problem, unexpected type, mapping problem and other issues. We fixed the error and build a clean dataset based on results of detected issues. In data summary part, we summarized the dataset in 10 main aspects to show what we learned from this dataset. In the data exploration part, we explore 6 main possible correlations, and calculate the correlation factor between these pairs. Finally, we list all the anomalies and outliers we have discovered so far.

3 Part 1: data issues and data summary

3.1 Data issues

3.1.1 Null values

Column null restriction

A lot of fields has null values. Some of them are considered valid, some of them are not.

Column	Allowed to be Null (How to judge if valid or not)
CMPLNT_NUM	Cannot be null and must be a number string
CMPLNT_FR_DT	Can be null (But if this is null and CMPLNT_FR_TM is not null, then this Null value will be regarded as invalid.)
CMPLNT_FR_TM	Can be null (But if this is null and CMPLNT_FR_DT is not null, then this Null value will be regarded as invalid.)
CMPLNT_TO_DT	Can be null (But if this is null and CMPLNT_FR_DT is not null, then this Null value will be regarded as invalid.)
CMPLNT_TO_TM	Can be null (But if this is null and CMPLNT_FR_DT is not null, then this Null value will be regarded as invalid.)
RPT_DT	Cannot be null
KY_CD	Cannot be null
OFNS_DESC	Cannot be null
PD_CD	Cannot be null
PD_DESC	Cannot be null
CRM_ATPT_CPTD_CD	Cannot be null
LAW_CAT_CD	Cannot be null
JURIS_DESC	Cannot be null
BORO_NM	Cannot be null
ADDR_PCT_CD	Cannot be null
LOC_OF_OCCUR_DESC	Can be null
PREM_TYP_DESC	Cannot be null
PARKS_NM	Can be null
HADEVELOPT	Can be null
X_COORD_CD	Can be null but only if OFNS_DESC is Rape or Sex Crime (But if this is null and any one of other Geo info fields is not null, then this Null value will be regarded as invalid.)
Y_COORD_CD	Can be null but only if OFNS_DESC is Rape or Sex Crime (But if this is null and any one of other Geo info fields is not null, then this Null value will be regarded as invalid.)
Latitude	Can be null but only if OFNS_DESC is Rape or Sex Crime (But if this is null and any one of other Geo info fields is not null, then this Null value will be regarded as invalid.)
Longitude	Can be null but only if OFNS_DESC is Rape or Sex Crime (But if this is null and any one of other Geo info fields is not null, then this Null value will be regarded as invalid.)
Lat_Lon	Can be null but only if OFNS_DESC is Rape or Sex Crime (But if this is null and any one of other Geo info fields is not null, then this Null value will be regarded as invalid.)

We deal with all these null value checking Basic_Issue

List of issues we found in null value checking (all of them lead to the whole row to be invalid data)

- 4574 rows of data missing three digit internal classification code(PD_CD). These 4574 rows of data contains all the offense classification code of "101"(MURDER & NON-NEGL) MANSLAUGHTER.
- 463 rows of data has no borough record.
- 390 rows of data has no precinct record.
- 7 rows of data contains empty crime completion record.
- 5415 rows of record contains incomplete date error problem. We first judge if each row of data has incomplete start/end time record. Incomplete means that this start/end date combination has only date or time(one of them is missing)

CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_FM	CMPLNT_TO_DT	CMPLNT_TO_FM
494598165	12/29/2015	06:00:00	(missing)	12:29:00
259351246	(missing)	09:00:00		
546886673	09/18/2014	15:00:00	07/14/2015	(missing)

3.1.2 Range problem

Field range problem is very important in data integrity. Here is how we check if certain field may have range problem

- 4574 rows of data missing three digit internal classification code(PD_CD). These 4574 rows of data contains all the offense classification code of "101"(MURDER & NON-NEGL) MANSLAUGHTER.
- 463 rows of data has no borough record.
- 390 rows of data has no precinct record.
- 7 rows of data contains empty crime completion record.
- 5415 rows of record contains incomplete date error problem. We first judge if each row of data has incomplete start/end time record. Incomplete means that this start/end date combination has only date or time(one of them is missing)

Column	Range
CRM_ATPT_CPTD_CD	COMPLETED or ATTEMPTED
LAW_CAT_CD	FELONY, MISDEMEANOR or VIOLATION
BORO_NM	"MANHATTAN", "BRONX", "BROOKLYN", "QUEENS" or "STATEN ISLAND"
LOC_OF_OCCUR_DESC	"INSIDE", "OPPOSITE OF", "FRONT OF", "REAR OF" or NULL value

We check if each element is valid or not in these columns, we check this all in Basic_Issue **We discover that all data dont have range problem. They are either null value or exactly in the range.**

3.1.3 Unexpected type problem

Every field has its own base type. So we also check some fields whether they have expected type

Column	Type
CMPLNT_NUM	Integer
CMPLNT_FR_DT	DateTime(MM(M)/dd(d)/yyyy)
CMPLNT_FR_TM	DateTime(HH:mm:ss)
CMPLNT_TO_DT	DateTime(MM(M)/dd(d)/yyyy)
CMPLNT_TO_TM	DateTime(HH:mm:ss)
RPT_DT	DateTime(MM(M)/dd(d)/yyyy)
KY_CD	Three Digit Number
OFNS_DESC	String
PD_CD	Three Digit Number
PD_DESC	String
CRM_ATPT_CPTD_CD	String
LAW_CAT_CD	String
JURIS_DESC	String
BORO_NM	String
ADDR_PCT_CD	String
LOC_OF_OCCUR_DESC	String
PREM_TYP_DESC	String
PARKS_NM	String
HADEVELOPT	String
X_COORD_CD	Float Number
Y_COORD_CD	Float Number
Latitude	Float Number
Longitude	Float Number
Lat_Lon	Tuple of Float Number

We deal with all these Unexpected type checking Basic_Issue

List of issues we found in base type checking: (All of them lead to the whole row of data to be invalid)

- **Invalid datetime** We detect if each data combination could be interpreted as real time.

```
def ifIsNotValidDateString(str):
    if not re.match(r"^(0?[1-9]|1[012])/(0?[1-9]|[12][0-9]|3[01])/\d{4}$", str):
        return True;
    array = str.split("/")
    ifCorrectDate = True
    try:
        newDate = datetime.datetime(int(array[2]),int(array[0]),int(array[1]))
    except ValueError:
        ifCorrectDate = False
    finally:
        return not ifCorrectDate

def ifIsNotValidTimeString(str):
    if re.match(r"^(0[0-9]|1[0-9]|2[0-3]):([0-5][0-9]):([0-5][0-9])$", str):
        return False
    else :
        return True
```

Figure 1: Date and time detection

```
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
24:00:00    TEXT    Complaint_Starting_Time Invalid
```

Figure 2: Invalid date time discovered

3.1.4 Mapping problem

One column may be dependent on any other columns. Dependences we have found are below:

- Start time of occurrence must be earlier than end time of occurrence. That is (CMPLNT_FR_DT,CMPLNT_FR_TM) must be earlier than (CMPLNT_TO_DT, CMPLNT_TO_TM).
  No such problem detected.

- To further protect victim identities, rape and sex crime offenses are not geocoded.So when OFNS_DESC is Rape or Sex Crime, X_COORD_CD, Y_COORD_CD, Longitude, Latitude and tuple of (Longitude, Latitude) must all be null values.

And meanwhile, when X_COORD_CD, Y_COORD_CD, Longitude, Latitude and tuple of (Longitude, Latitude) are null values, OFNS_DESC should only be Rape or Sex Crime.
119461 rows of data has such problem. Script for detecting such problem is SexRape_Coordinate_mapping.py

- One offense classification code(KY_CD) can only be mapped to one description(OFNS_DESC). **15 pairs of KY_CD & OFNS_DESC has conflict mapping problem.** Script for detecting such problem is OffenseCodeMapping.py
- One internal offense classification code(PD_CD) can only be mapped to one description(PD_DESC). **No such conflict mapping problem is found.** Script for detecting such problem is PDCodeMapping.py
- One internal offense classification code(PD_CD) can only be mapped to one offense classification code(KY_CD). **183 pairs of PD_CD & KY_CD has conflict mapping problem.** Script for detecting such problem is PD_OFNS_Mapping.py
- The tuple of Latitude and Longitude must be corresponding to the individual value of latitude and longitude. **No such problem is found.** Script for detecting such problem is LATITUDE_LONGITUDE.py
- One precinct(ADDR_PCT_CD) can only be mapping to one particular borough(BORO_NM). **27 pairs of ADDR_PCT_CD & BORO_NM has conflict mapping problem.** Script for detecting such problem is Precinct_BORO_mapping.py

3.1.5 Other issues

- No wrong number of fields record is found.All rows of data contains 24 columns of data.
- No duplicate compliant number record is found. (CMPLNT_NUM.py) No duplicate index
- No wrong report date problem is found. Ex. date time like 2/31/2017 is invalid.

3.2 Data Cleaning

After detecting and finding all possible issues, we need to get rid of these invalid data rows. Besides, the dataset includes data prior to 2006, we get rid of all the data before 2006 and only keep data from 2006 - 2015.

3.2.1 Handling conflicting data

One important aspect in data cleaning is how we decide whether some data should be considered invalid or not when it has conflicts with other data. For example, we can see conflicts in Precinct Num- Borough Name.
If we see such problem, we will look through all the conflicting data and decide a threshold number for invalid number. Here we decide the threshold number to be 100. So if one mapping relationship has appeared less than 100, it is highly possible that it is wrong recorded information.

28 lines (27 sloc) | 697 Bytes

```
(('104', 'BROOKLYN'), 1)
(('104', 'MANHATTAN'), 1)
(('104', 'QUEENS'), 75114)
(('106', 'BROOKLYN'), 1)
(('106', 'QUEENS'), 61494)
(('114', 'BRONX'), 2)
(('114', 'QUEENS'), 91694)
(('121', 'BROOKLYN'), 1)
(('121', 'STATEN ISLAND'), 17374)
(('13', 'BROOKLYN'), 1)
(('13', 'MANHATTAN'), 74442)
(('14', 'BROOKLYN'), 1)
(('14', 'MANHATTAN'), 119414)
(('23', 'BRONX'), 3)
(('23', 'MANHATTAN'), 66690)
(('25', 'BRONX'), 1)
(('25', 'MANHATTAN'), 67479)
(('26', 'BROOKLYN'), 1)
(('26', 'MANHATTAN'), 34274)
(('6', 'BRONX'), 1)
(('6', 'MANHATTAN'), 54904)
(('7', 'BROOKLYN'), 1)
(('7', 'MANHATTAN'), 40885)
(('71', 'BRONX'), 1)
(('71', 'BROOKLYN'), 72088)
(('9', 'BROOKLYN'), 1)
(('9', 'MANHATTAN'), 62164)
```

Figure 3: all borough and precinct mapping

104	BROOKLYN
104	MANHATTAN
106	BROOKLYN
114	BRONX
121	BROOKLYN
13	BROOKLYN
14	BROOKLYN
23	BRONX
25	BRONX
26	BROOKLYN
6	BRONX
7	BROOKLYN
71	BRONX
9	BROOKLYN

Figure 4: Invalid mapping

3.2.2 Collecting key information of potential invalid row of data

For the each mapping problem, we output one .out file illustrating how one row of data will be regarded as invalid.For example:

120	ENDAN WELFARE INCOMP
124	KIDNAPPING
124	KIDNAPPING AND RELATED OFFENSES
345	ENDAN WELFARE INCOMP
364	AGRICULTURE & MRKTS LAW-UNCLASSIFIED
364	OTHER STATE LAWS (NON PENAL LAW)
677	NYS LAWS-UNCLASSIFIED VIOLATION

Figure 5: Invalid mapping

This is the .out file recording all invalid offense code mapping data. In this file, each row has one key-value pair. So as long as KY_CD equals to that key and OFNS_DESC equals to value, that row of data is considered as invalid. This is pretty much like where clause in sql query.
So in the data cleaning script, we load these .out file data.

```
if os.path.isfile('./Other_Issue/InvalidOffenseCodeMapping.out/part-00000'):
    file = open("./Other_Issue/InvalidOffenseCodeMapping.out/part-00000")
    while 1:
        line = file.readline().rstrip('\n')
        if not line:
            break
        [key, value] = line.split("\t")
        if not invalidOffenseCodeDetailRecord.has_key(key):
            invalidOffenseCodeDetailRecord[key] = []
        if value not in invalidOffenseCodeDetailRecord.get(key):
            invalidOffenseCodeDetailRecord.get(key).append(value)
```

Figure 6: data cleaning script

We load these out file in python dictionary data(HashMap) structure or set data structure(HashSet). Then we decide if certain row of data is invalid or not, we will check in these hashmap and hashset to see if key info has certain matches.

3.2.3 Script for data cleaning

We write a script for data cleaning. DataClean.py We have filters for the following issue: 1)If this field has invalid null value 2)If data in this field has invalid base type 3)If data in this field should be in certain range 4)If data in this field has mapping problem

```
cleanedData = lines.map(toStrip) \
                   .filter(isNullValue) \
                   .filter(dataTypeCheck) \
                   .filter(isInSpecificRange) \
                   .filter(mappingCheck)
```

Figure 7: filters for data clean

3.2.4 Results

After we do the cleaning work, we have our new cleanData.csv. All data summary work is based on clean data, not on raw data. The original data has size of 1.3GB. The clean data is 930MB left.[3]

3.3 Data summary

3.3.1 Borough crime

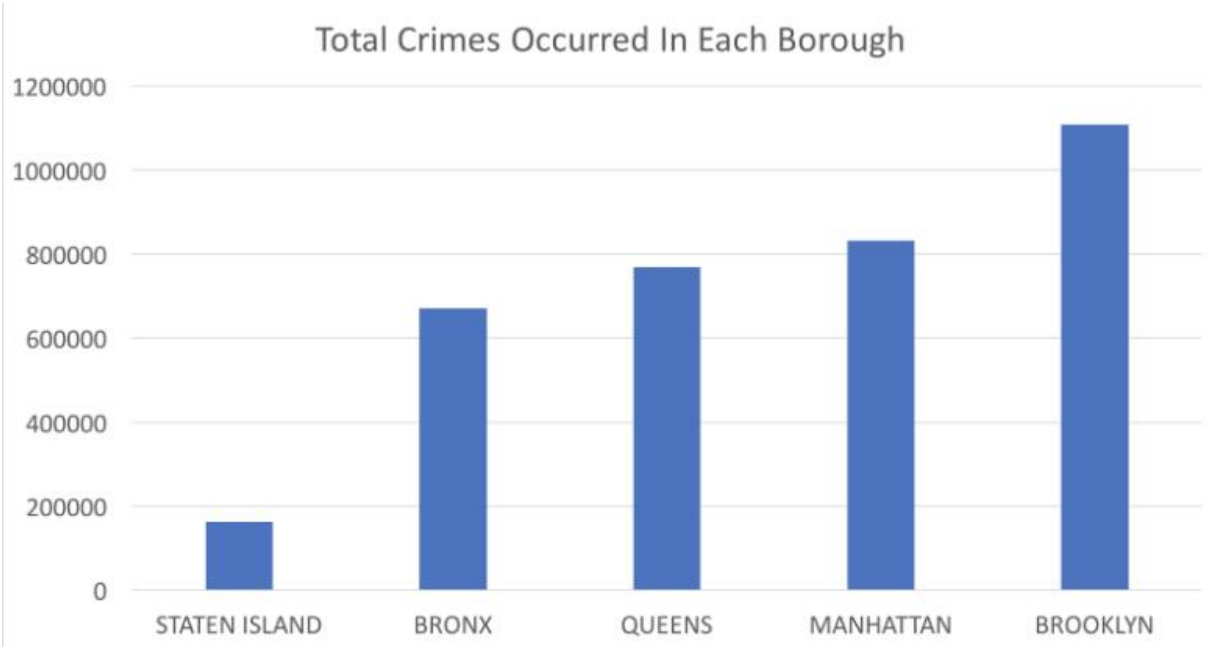

Figure 8: borough crime distribution

As we can see from figure above, Brooklyn takes the first place for number of incidences occurred, which is 1,107,843. Manhattan takes the second place, which is 831,636. Queens and Bronx follow behind, and staten island has the least number of incidences occurred. However, as is shown in [2]NYC distribution 2016, Brooklyn has the most population, and Queens has the second most population, and Manhattan takes the third place, which is contradicted the fact that manhattan has the second number of incidences occurred. Python code in this link(BoroughNameDistribution.py)

3.3.2 Precinct crime

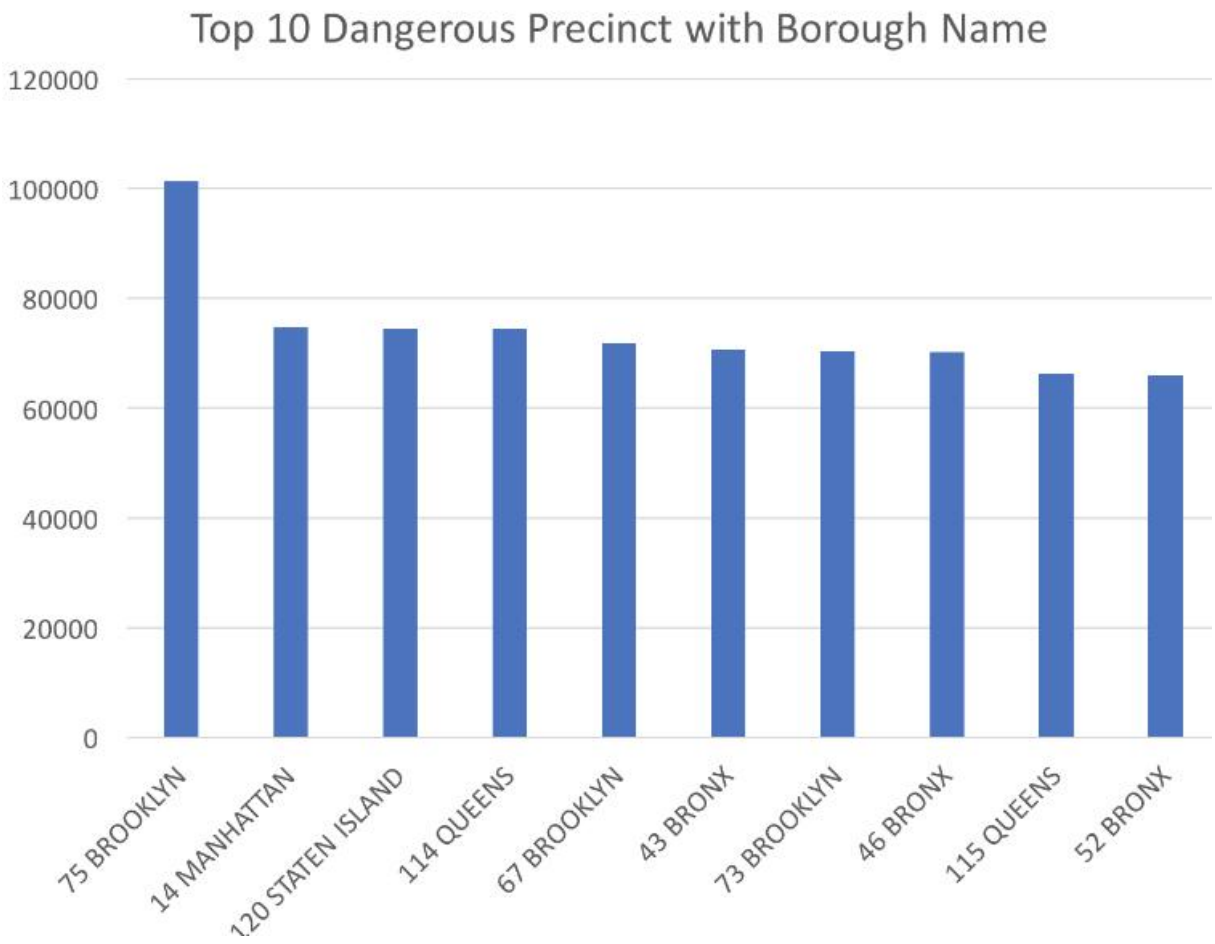

Figure 9: Precinct crime distribution

The figure below shows the top 10 precincts with borough name with respect to number of incidents. From the figure below, we can see the most dangerous precinct is 75 Brooklyn, over 100,000 times of crime happened here, which is far more than the other nine. Meanwhile, Brooklyn has 3 precincts in this top 10 list, which is the same as Bronx. The number of precincts in Manhattan, Queens and Staten Island is 1, 2 and 1 respectively. This distribution can also reflect the borough distribution above, in which Brooklyn takes the first place and Manhattan takes the second.
Python code in this link(PrecinctDistribution.py)

3.3.3 Number of occurrence in each mouth

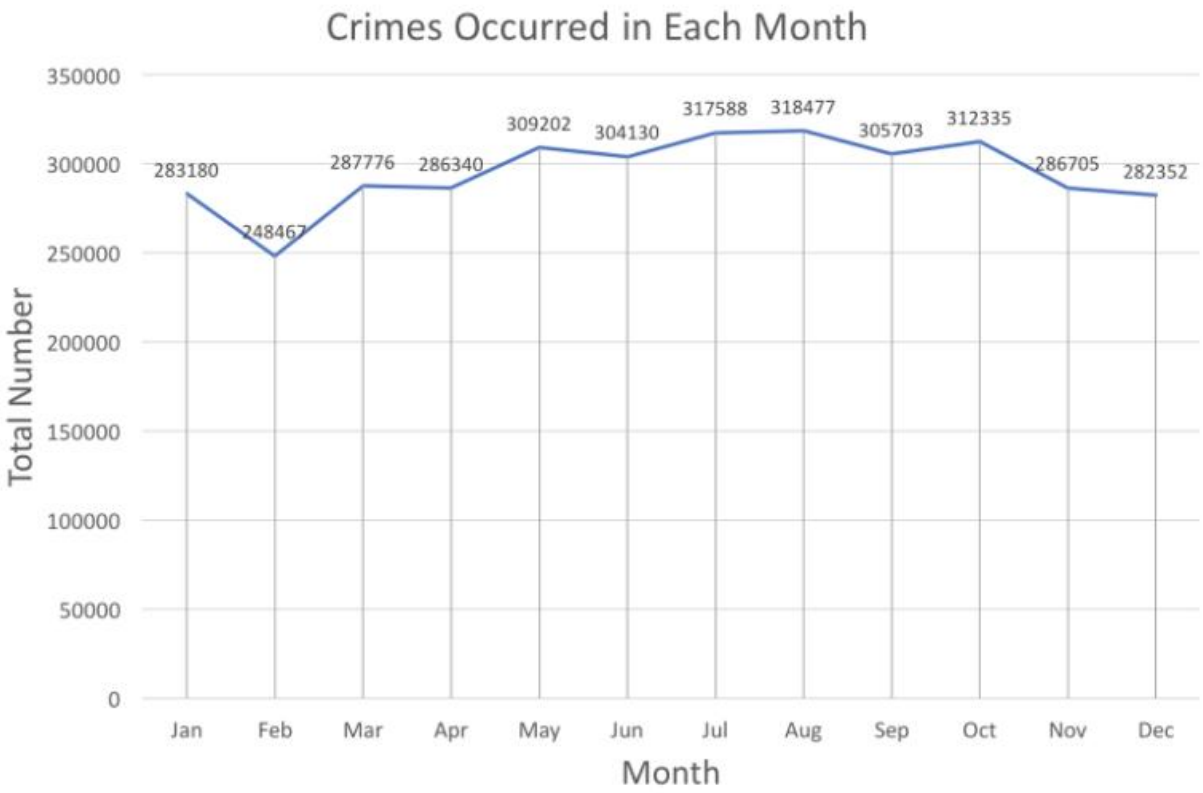

Figure 10: Number of occurrence in each mouth

In this diagram, crime happens least in February and peaks at August. In fact we can see there is an obvious raise from February to May and fluctuates during the next few months. The difference between the max one and the least one is about 70,000. We guess the reason is due to the temperature change, which means criminals also dont want to go outside in winter.
Python code in this link(OffenseYearMonth.py)

3.3.4 Offense numbers in daytime

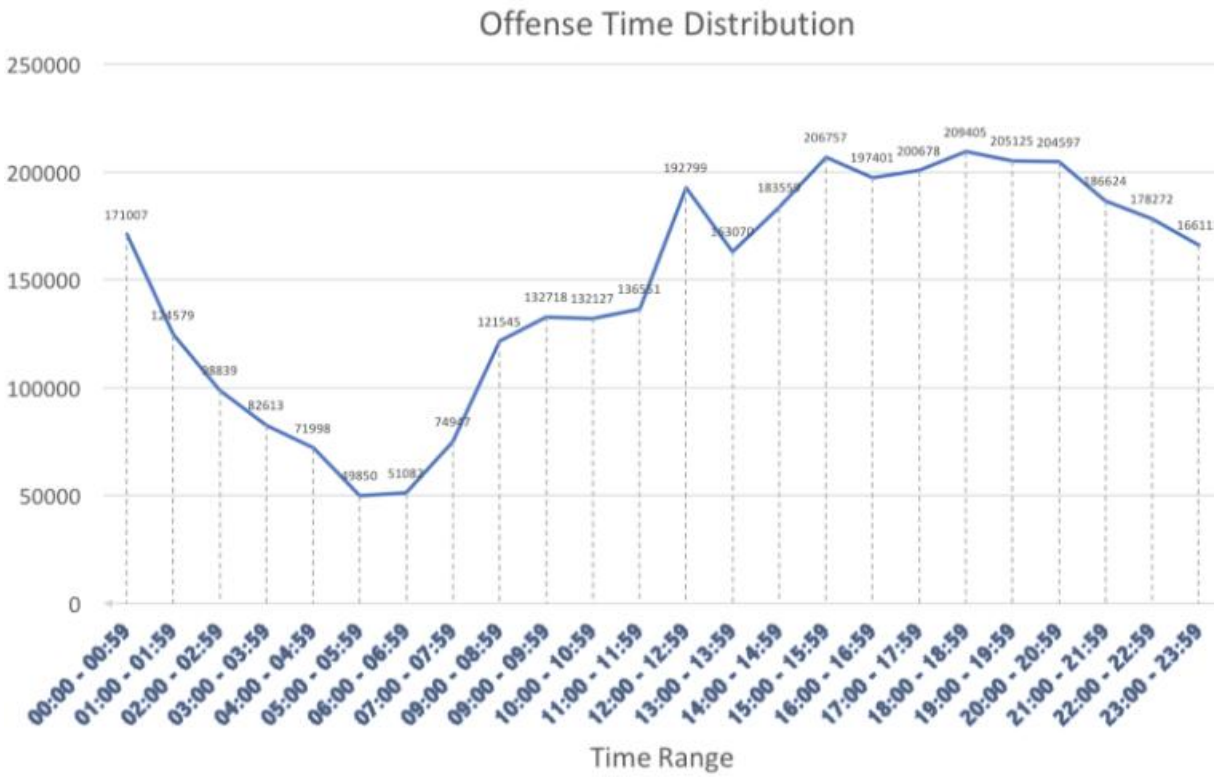

Figure 11: Offense numbers in daytime

In this picture, we can see offence most happen between 15:00 and 21:00, which reminds us to avoid hanging in dangerous area during this period. The crime number decreases sharply during 00:00 to 05:00 and raises significantly during 06:00 to 12:00. Its corresponded to peoples sleeping custom. The maximum occurrence between 15:00 to 15:59 is around 5 times of that minimum between 05:00 to 05:59. Whats interesting is some criminals are night-owls and they will commit offenses to those who are also night-owls. Python code in this link(OffenseTime.py)

3.3.5 Offense level

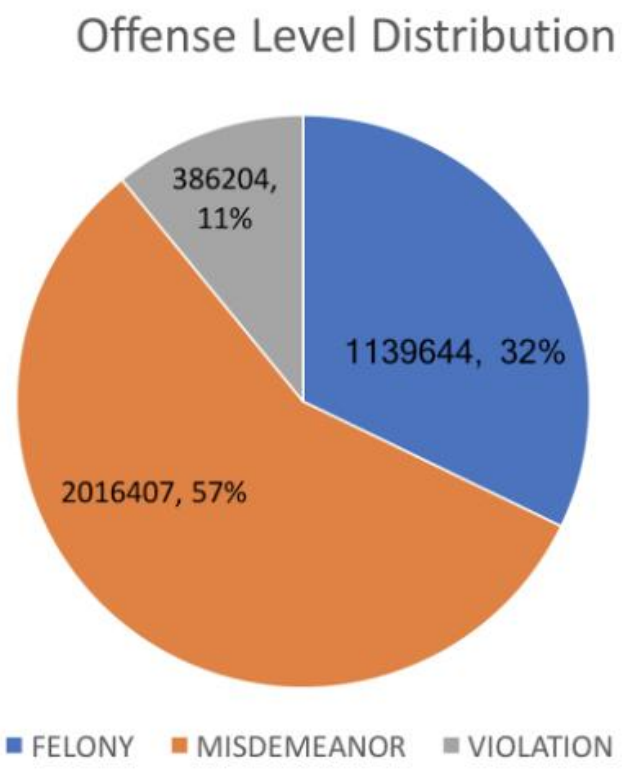

Figure 12: Offense level distribution

The analysis of offense level distribution is quite fitting our assumption. Felony is the most part which follows the misdemeanor, while the former one is about twice of the later one. Despite of felony and misdemeanor, violation seems to be few. We guess its because violation couldnt be counted as a crime and people would choose not to report if they have meet some not so much annoying offense.
Python code in this link(OffenseLevelDistribution.py)

3.3.6 Offense level distribution corresponding to borough

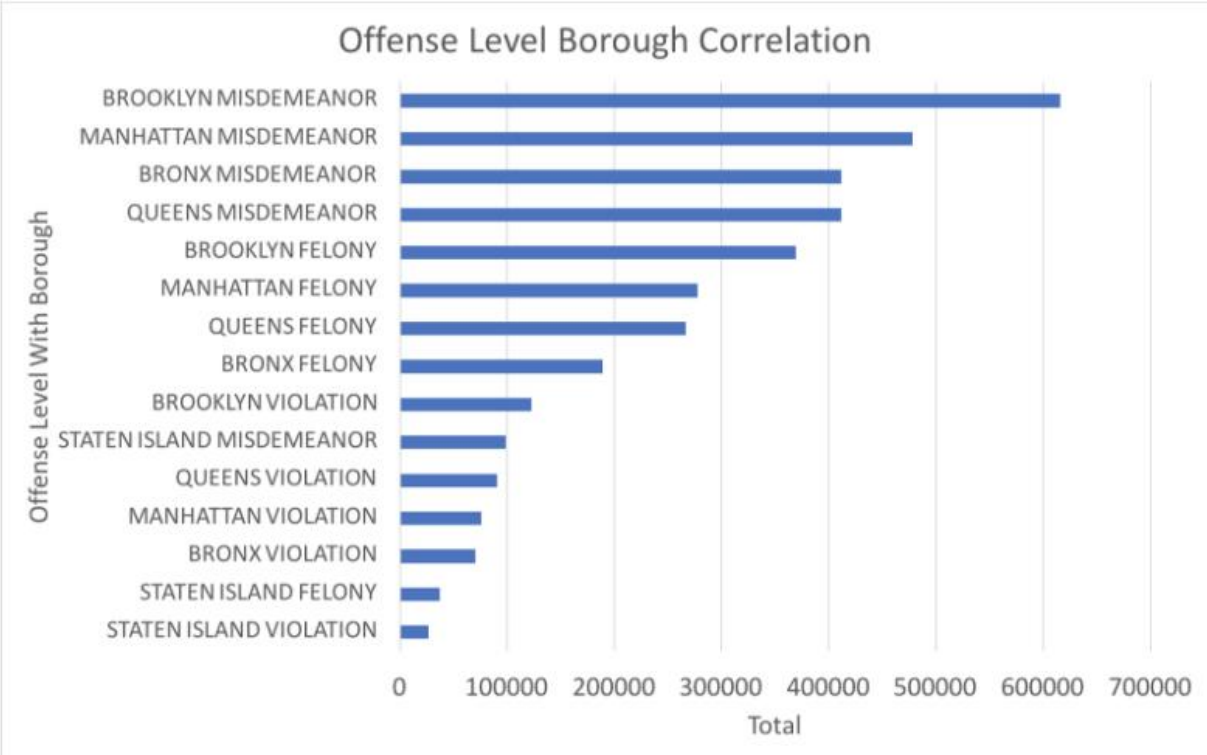

Figure 13: Offense level distribution corresponding to borough

Among the this chart, we can learn that Brooklyn is the most dangerous precinct, no matter comparing misdemeanor, felony or violation, more than 600,000 misdemeanor, 380,000 felony and 80,000 violation happening here. Whats attracting us is that the felony happened in staten island is pretty less than the other counterparts, even less than violation of other four precinct. The ratio among misdemeanor, felony and violation in all five precinct matches the former analysis, misdemeanor is the most and violation is the least no matter in which area.
Python code in this link(OffenseLevelBoroughCorrelation.py)

3.3.7 Offense type

Top 20 Most Occurred Offense Event With Level			
Rank	Offense Detail	Level	Total
1	PETIT LARCENY	MISDEMEANOR	616025
2	HARRASSMENT 2	VIOLATION	378619
3	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR	354209
4	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR	330679
5	GRAND LARCENY	FELONY	321773
6	OFF. AGNST PUB ORD SENSBLTY &	MISDEMEANOR	188292
7	DANGEROUS DRUGS	MISDEMEANOR	185416
8	BURGLARY	FELONY	170623
9	ROBBERY	FELONY	136197
10	FELONY ASSAULT	FELONY	126996
11	GRAND LARCENY OF MOTOR VEHICLE	FELONY	89563
12	MISCELLANEOUS PENAL LAW	FELONY	78140
13	OFFENSES AGAINST PUBLIC ADMINI	MISDEMEANOR	69570
14	CRIMINAL MISCHIEF & RELATED OF	FELONY	56334
15	INTOXICATED & IMPAIRED DRIVING	MISDEMEANOR	51255
16	DANGEROUS WEAPONS	MISDEMEANOR	49441
17	CRIMINAL TRESPASS	MISDEMEANOR	43877
18	DANGEROUS DRUGS	FELONY	40469
19	DANGEROUS WEAPONS	FELONY	34733
20	FORGERY	FELONY	34570

Figure 14: Offense type

Python code in this link(OffenseTypeLevelCorrelation.py)

3.3.8 Specific description of Premises

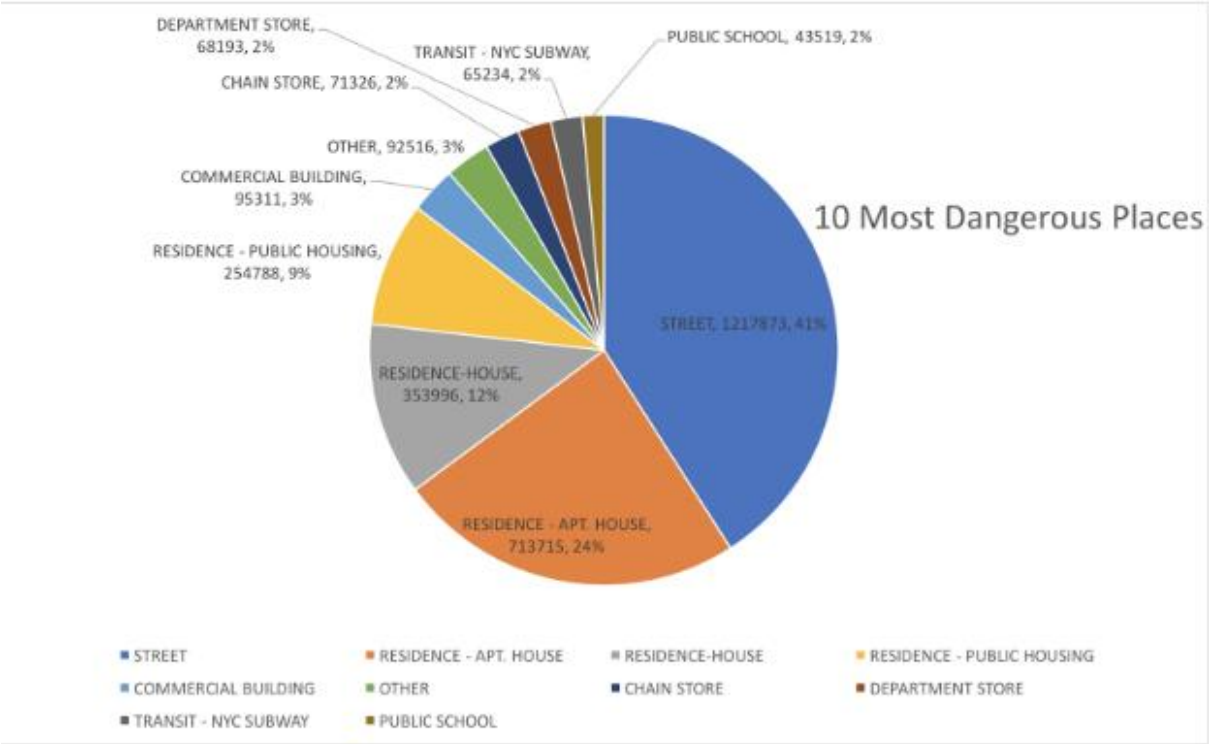

Figure 15: Specific description of Premises

As is shown in the chart above, most incidents happen on the streets and residence houses or apartments, which take up roughly 77%. Thats where people usually gather together or stay in. Some other places include commercial buildings, stores, public transportation, schools.
Python code in this link(PremisesDistribution.py)

3.3.9 Dangerous Drugs and Dangerous Weapons on the Street in each year

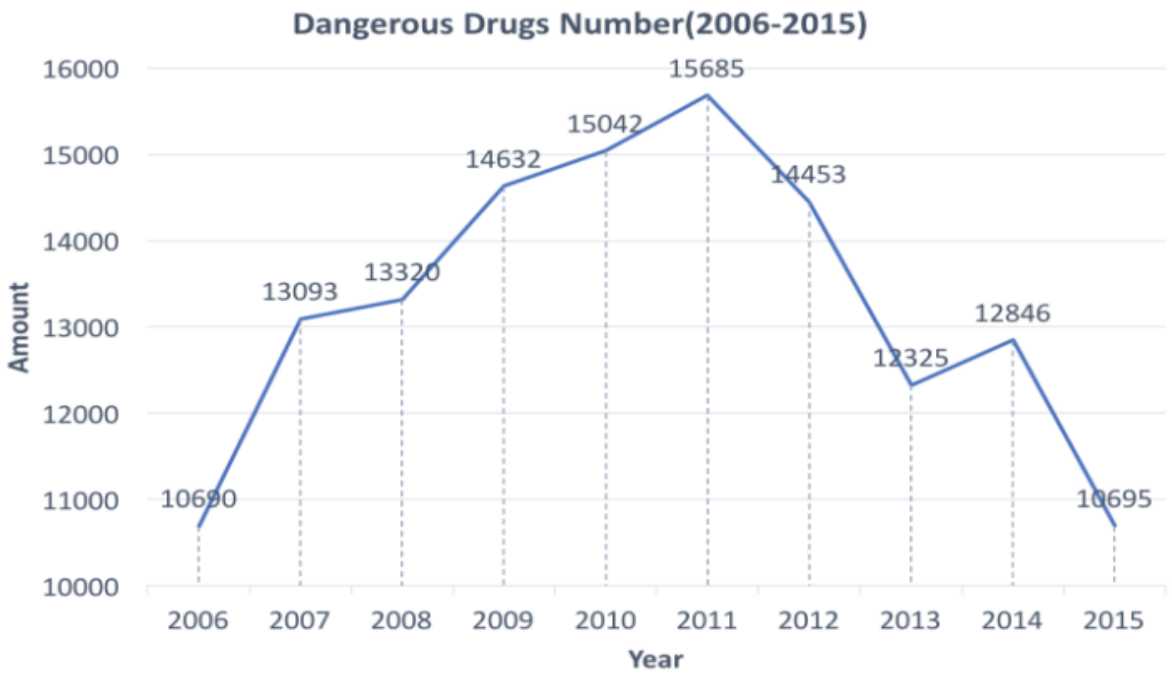

Figure 16: Dangerous Drugs on the Street in each year

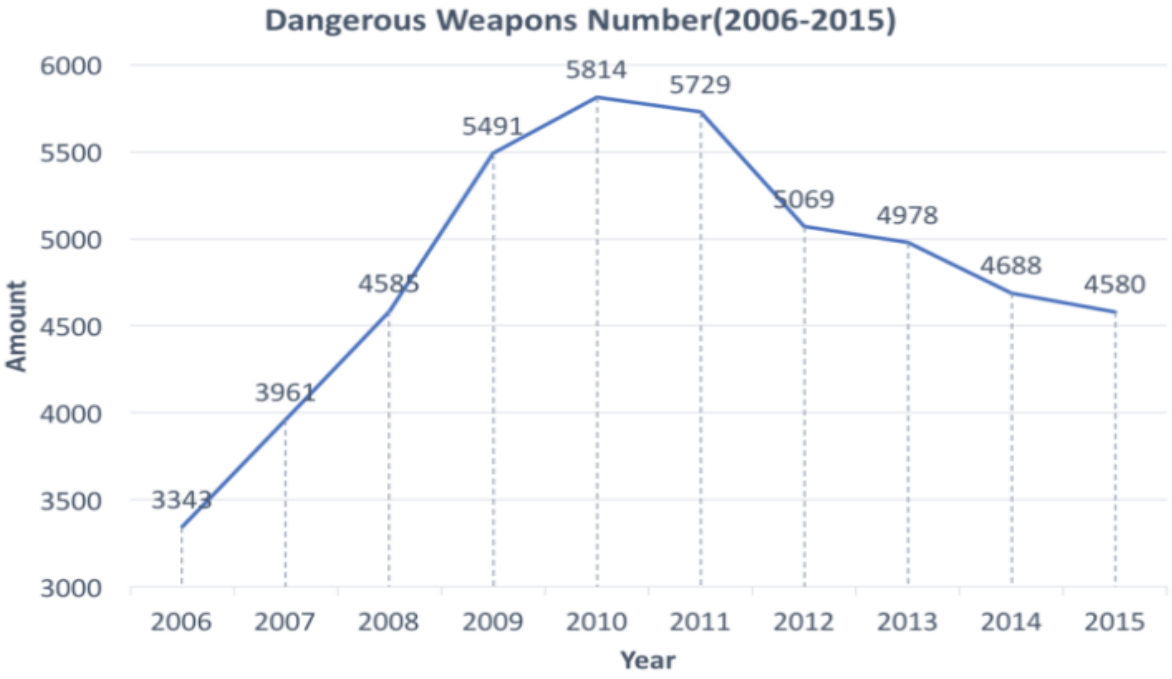

Figure 17: Dangerous Weapons on the Street in each year

This filtered data will be used in analyzing NYC stop-and-frisk policy.
Python code in this link(DangerousDrugsOrWeapons.py)

3.3.10 Midtown Manhattan crime statistics

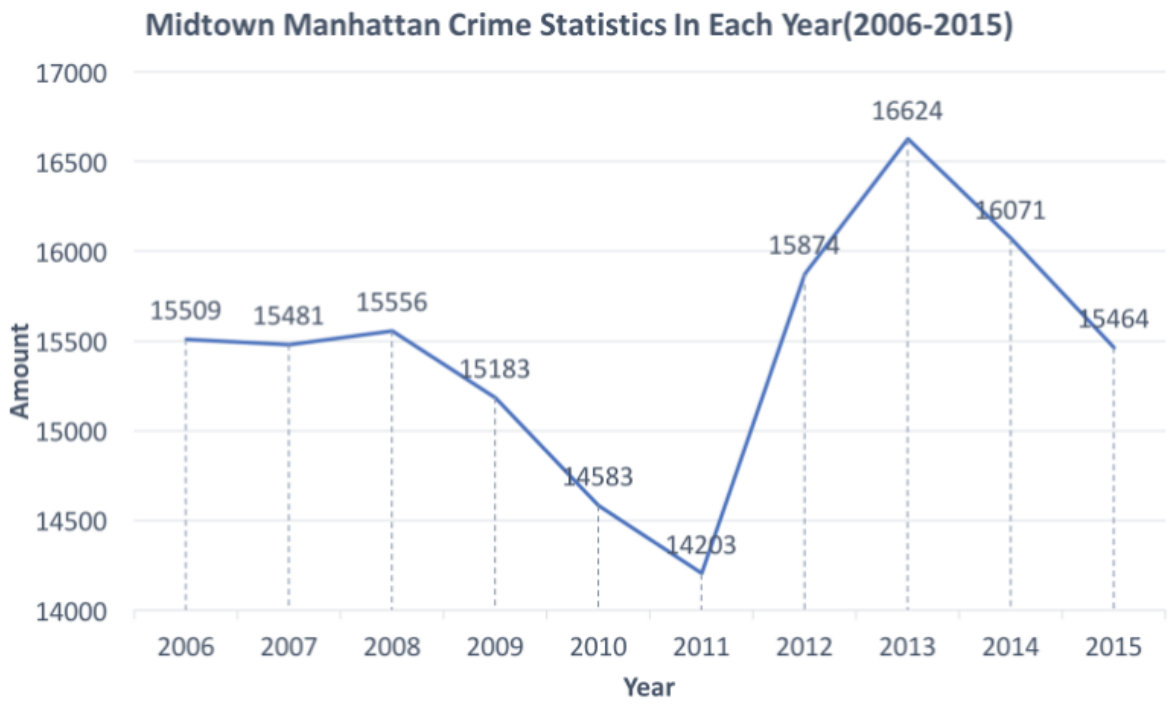

Figure 18: Midtown Manhattan crime statistics each year

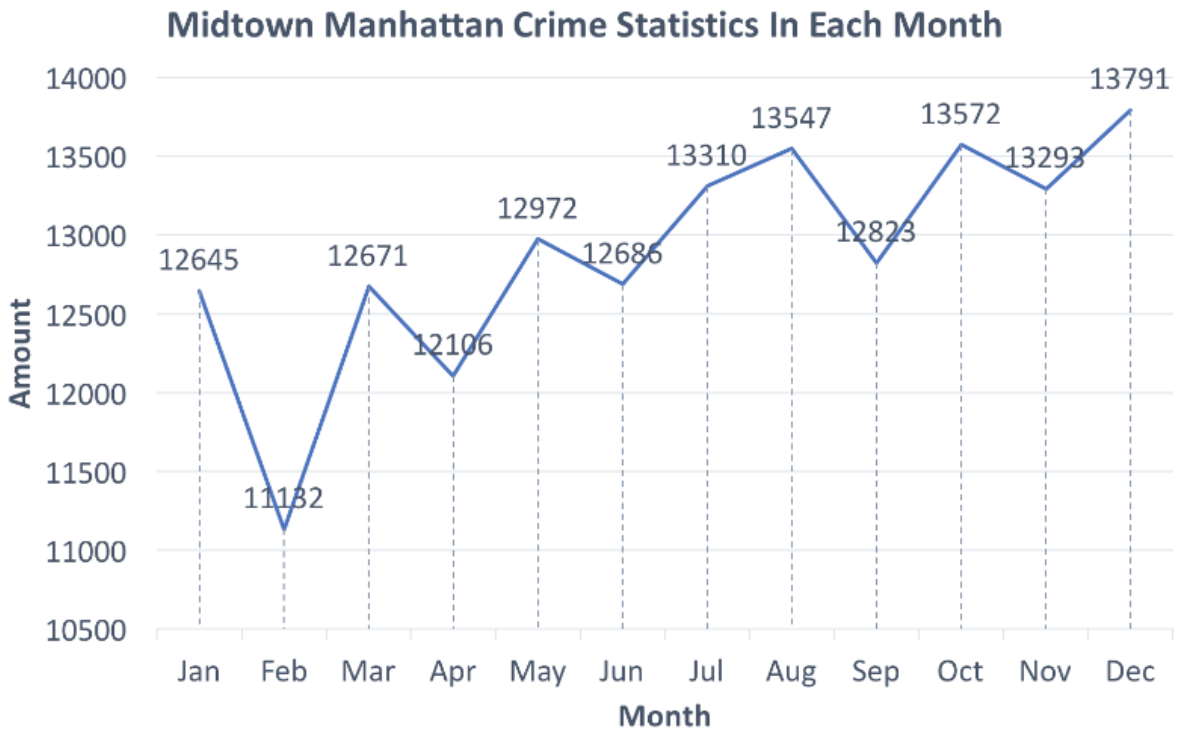

Figure 19: Midtown Manhattan crime statistics each month

This filtered data will be used in analyzing nyc visitors total number influence.
Python code in this link(MidtownManhattanCrimeStatisticsInEachYear.py) and MidtownManhattanCrimeStatisticsInEachMonth.py)

4 Part 2: Data Exploration

4.1 experimental setup in the whole project

In our project, we use apache spark to do the data integrity checking, data cleaning work, and data summary work.
For convenience, we build up our own environment locally to help use test our code easily, and here is the instruction of how we did it.

- Install Java and set up Java environment
- Install Scala and set up Scala environment
- Download Spark from Apache Spark official website(2.0.2 is the version we used here), and extract it to /usr/local/ folder and rename it as spark
- Add Spark environment variables in bash profile:
  export SPARK_HOME=/usr/local/spark
  export PATH=$PATH:$SPARK_HOME/bin
- Navigate to /usr/local/spark, run cp spark-env.sh.template spark-env.sh and edit it as:
  export SCALA_HOME=/usr/local/scala
  export SPARK_MASTER_IP=localhost
  export SPARK_WORKER_MEMORY=4g
- Run spark-shell command to test it and use pyspark to submit job

4.2 The correlation of hypotheses

The formula we use to calculate correlation:

$$Cov(X,Y) = E[X - E(X)][Y - E(Y)]$$

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

4.2.1 Does population census have any influence on the crimes number?

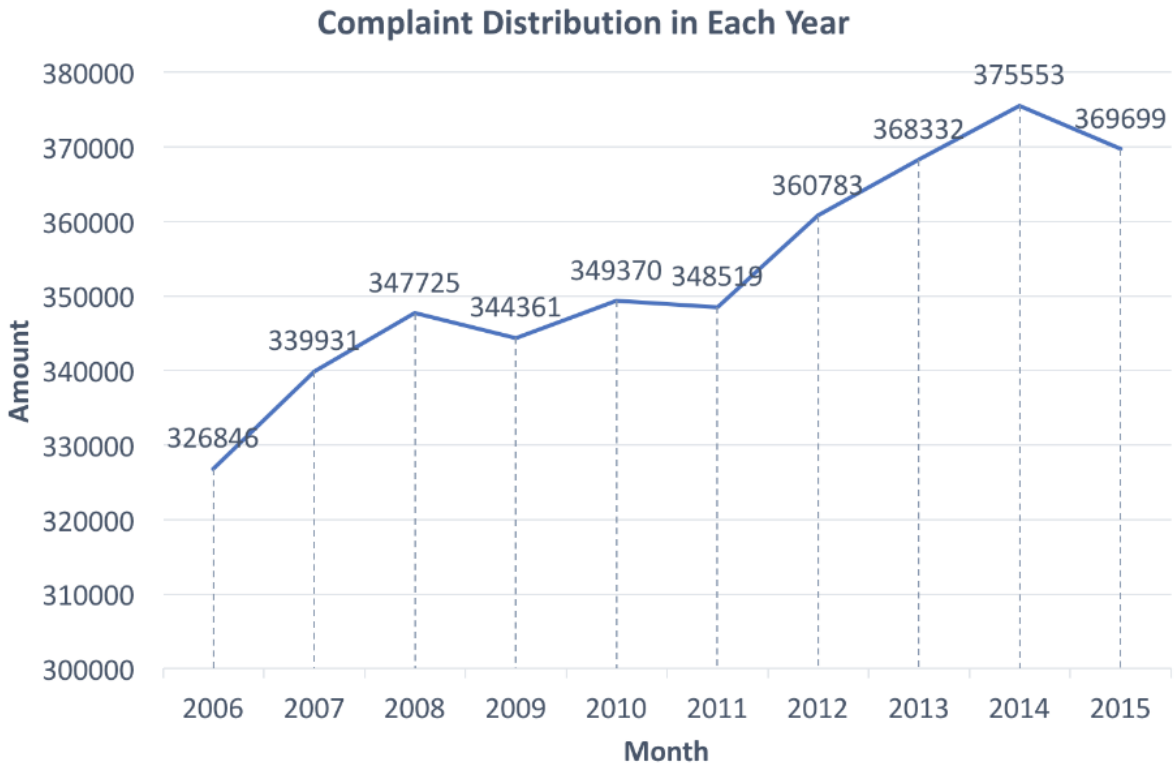

Figure 20: Crime number in each year

We process the crime distribution data in OffenseYear.py, and generate this plot using that data.
From the above picture, we find out that the complaint number has the trend of increasing. So we guess there must be some correlation between the population in NYC and Complaint Total Number.
First, Lets look at the crime statistics in USA from year 2006 to year 2012. [data source link[4]]

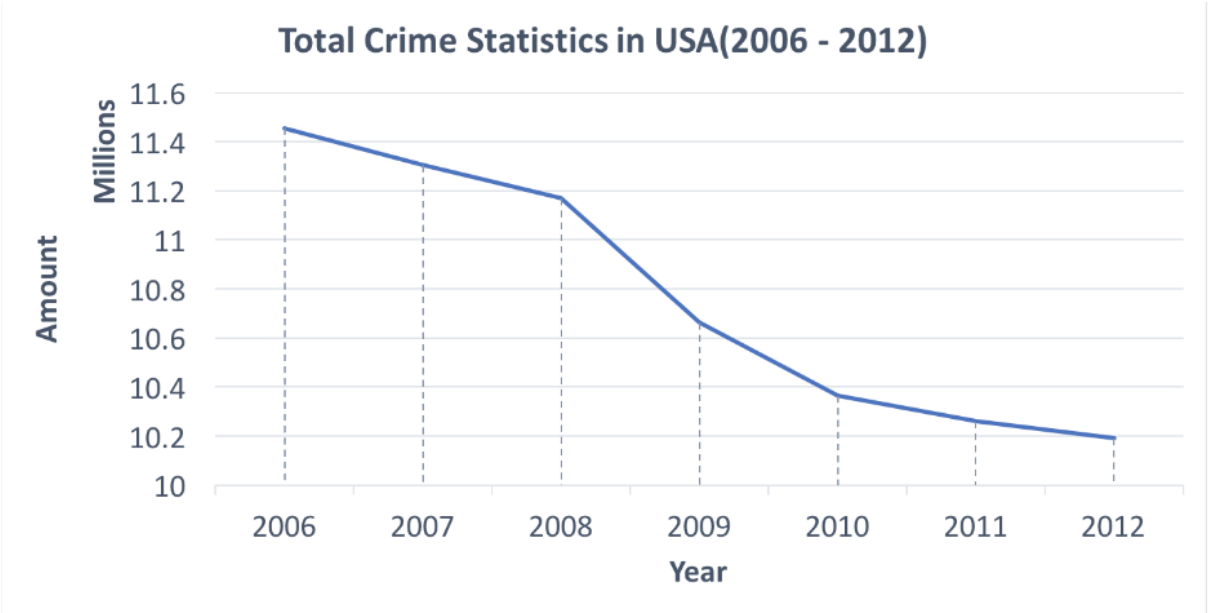

Figure 21: Crime number in USA

From the plot, we can see that the trend of total crime case amount in USA is decreasing. So we can rule out the possibility that NYC crime increasing trend is influenced by the whole nation crime index.

Then we find NYC population data so see if there is any correlation between population and crimes number.

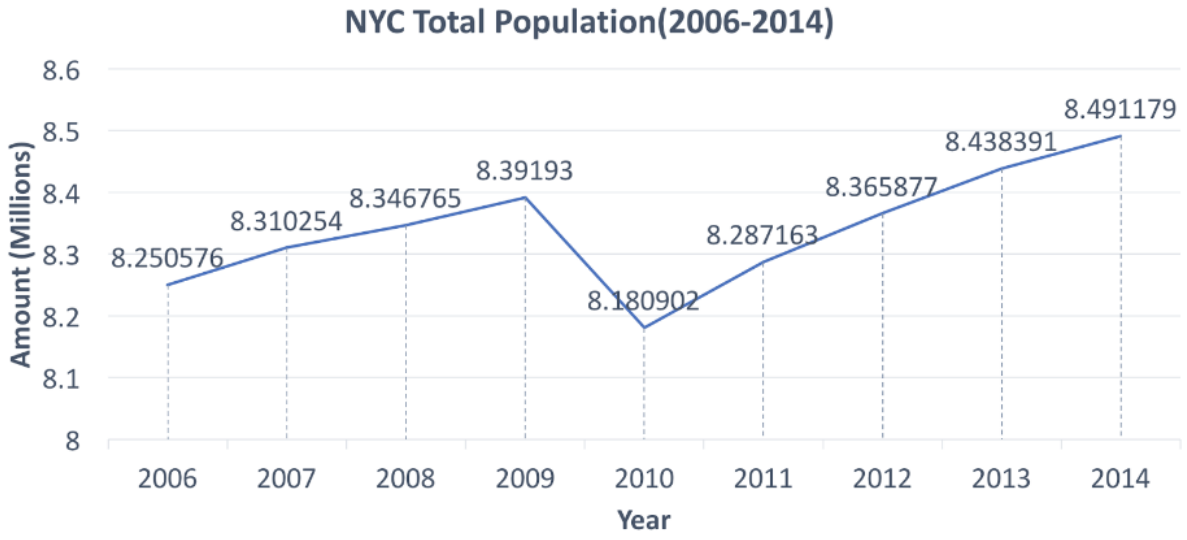

Figure 22: NYC total population

NYC population data comes from Google Public Data. [Data Source Link]
We can see that in NYC, the overall population also has the trend of growing. We conclude that population is at least one of the key factors that influence the trend of crime statistics.
Then we also compute the correlation between population and crimes number.

$$corr(NYCpopulation, NYCcrimesnumber) = 0.72$$

4.2.2 Does weather really influence crime statistics in NYC?

First let us show our generated crime statistics in each month. Python file for that processing has this link[OffenseYearMonth.py]

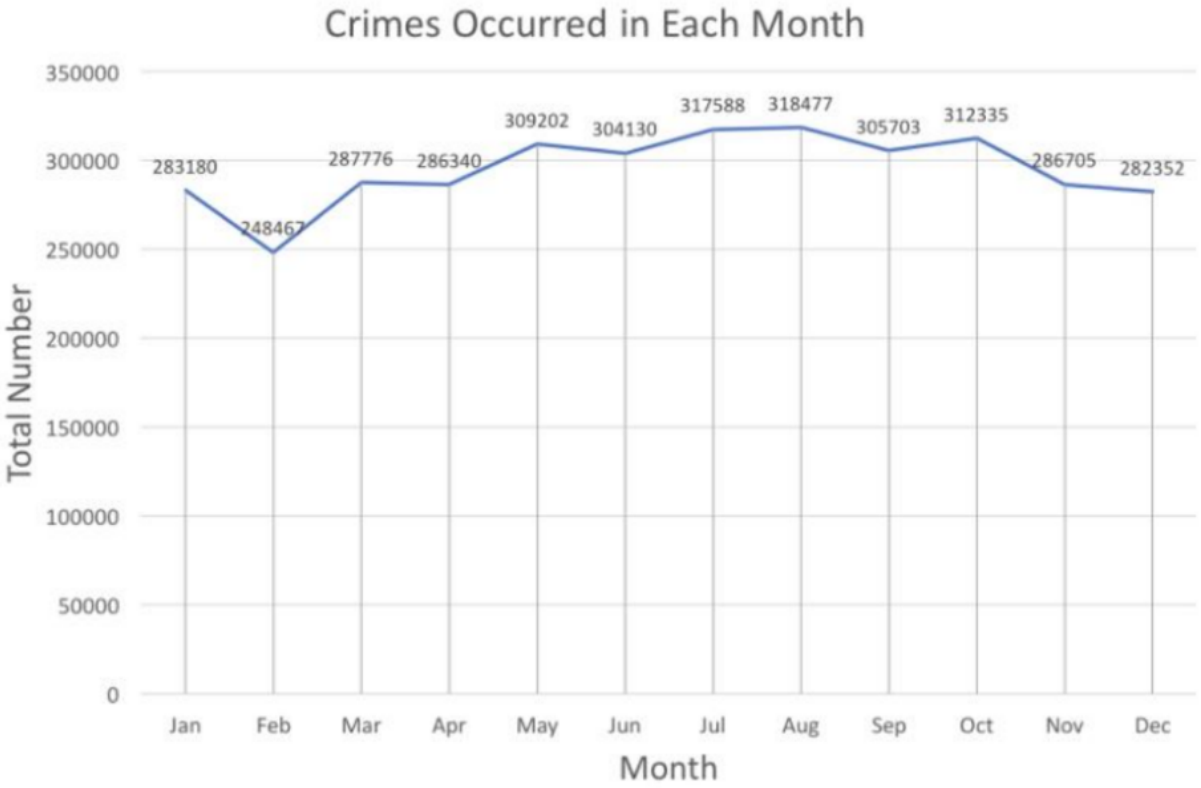

Figure 23: crime occurred in each month

We can see there is an obvious drop in Feb. Thats partly due to the number of days in February is about 2 3 days less in any other months. So if we extend February to 31 days, the total number of crimes reported will be 248867 * (31 / 28) = 275088. This new calculated number, when compared to number of January and March, still has a gap of 10000. And we can also see a downward trend from Oct to Dec. If we see further, this downward trend actually ends in next years February.
From above all, we have our assumption that, this downward trend is influenced by the weather, perhaps by snow falls, or extremely cold weather.

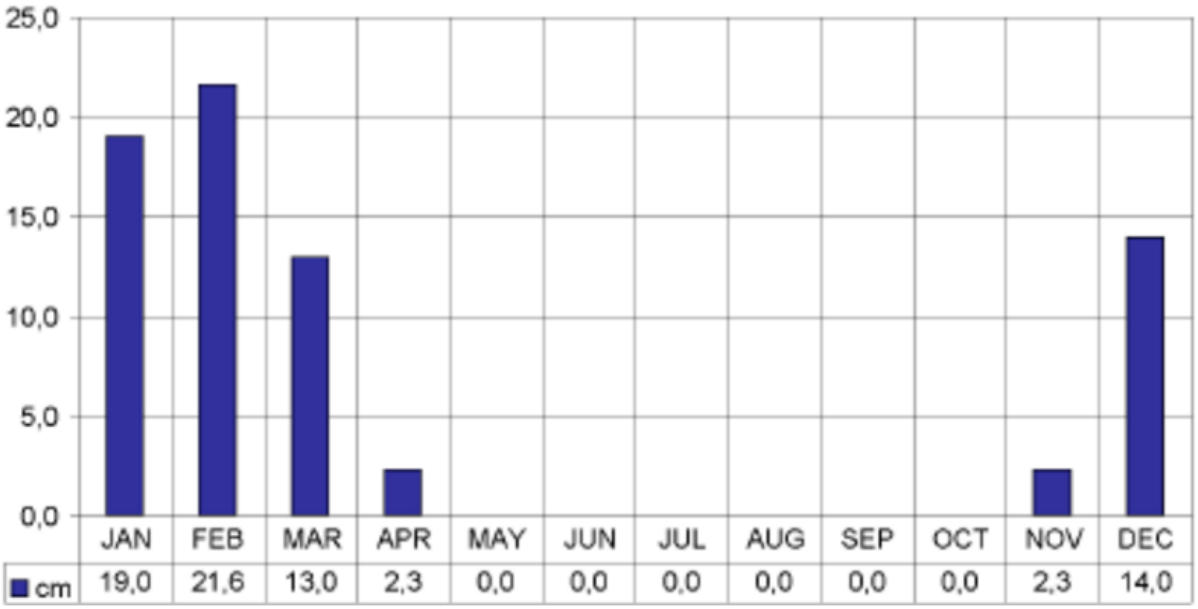

Figure 24: snowfall in each month

Snowfall in cm [link[5]]
First, we can see that months that has snowfall greater than 0.1 cm all have crimes amount less than 300000(if more precisely, all less than 290000).

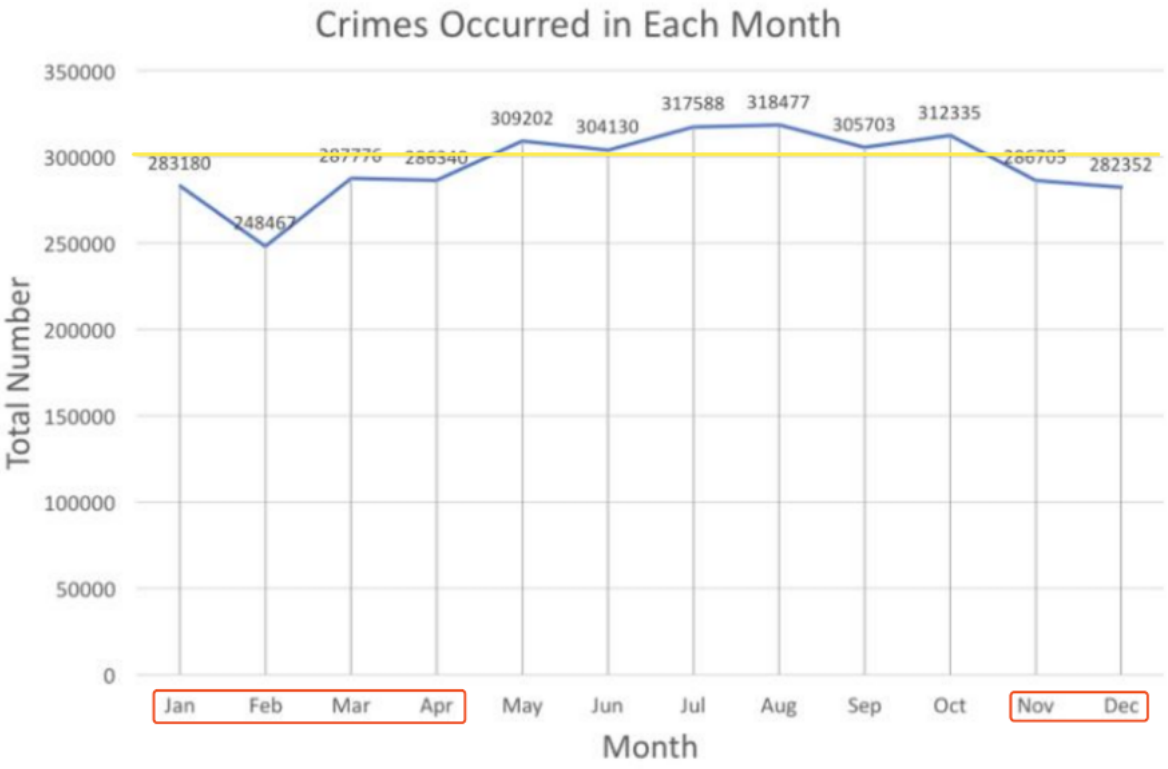

Figure 25: crime occurred in each month

Moreover, the downward trend in crime trend reaches its lowest point in February. Meanwhile, the snowfall amount in cm reaches its highest point(22 cm).

So these analytics confirm our assumptions, which is cold weather could suppress the NYC crime rates to some extent.

Then we also compute the correlation between crimes number in each month and snowfall amount in each month.

$$corr(CrimeNumberinEachMonth, SnowfallamountInEachMonth) = -0.84$$

(note: negative correlation coefficient indicates that one factor increases, the other one decreases)

4.2.3 Does Stop-and-frisk in New York City really help against dangerous drugs and weapons crime?

First, let us show some introductions to this stop-and-frisk policy. The stop-question-and-frisk program, or stop-and-frisk, in New York City, is a New York City Police Department practice of temporarily detaining, questioning, and at times searching civilians on the street for weapons and other contraband. link to explanation of stop-and-frisk[6]

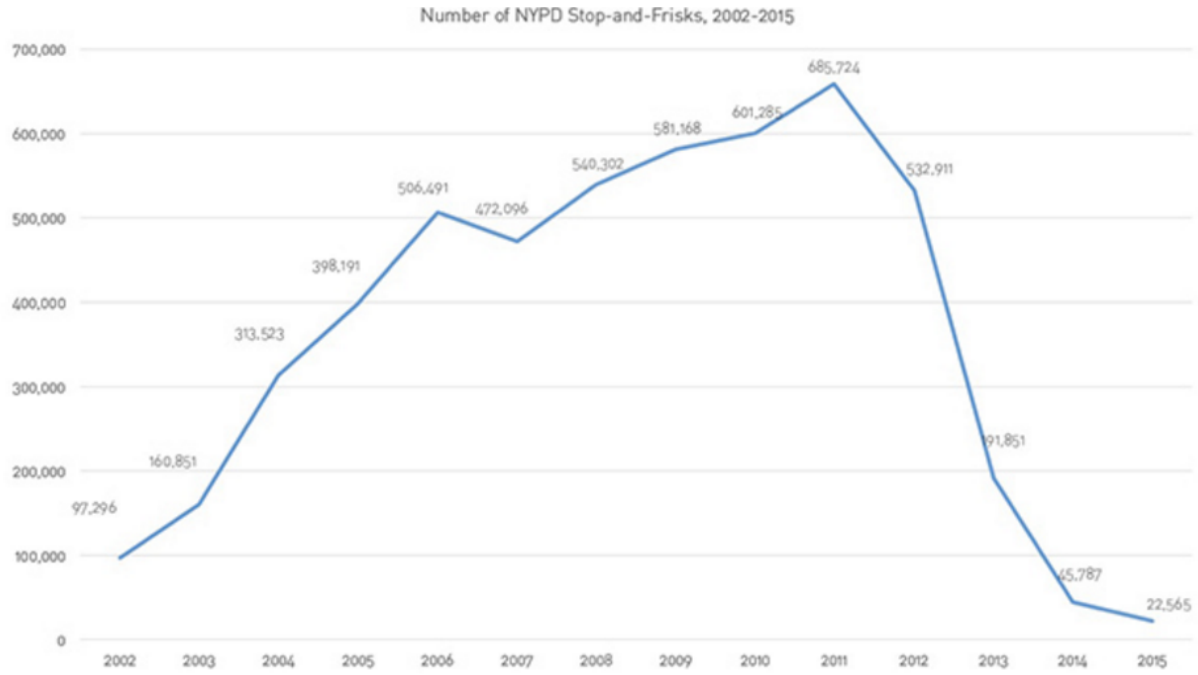

Figure 26: number of stop-and-frisk in each year

So we have our assumption: stop-and-frisk could help against dangerous drugs or weapons on the street.

In order to confirm our assumption, we are going to use our analyzed statistics concerning dangerous drugs and dangerous weapons reported on the streets.

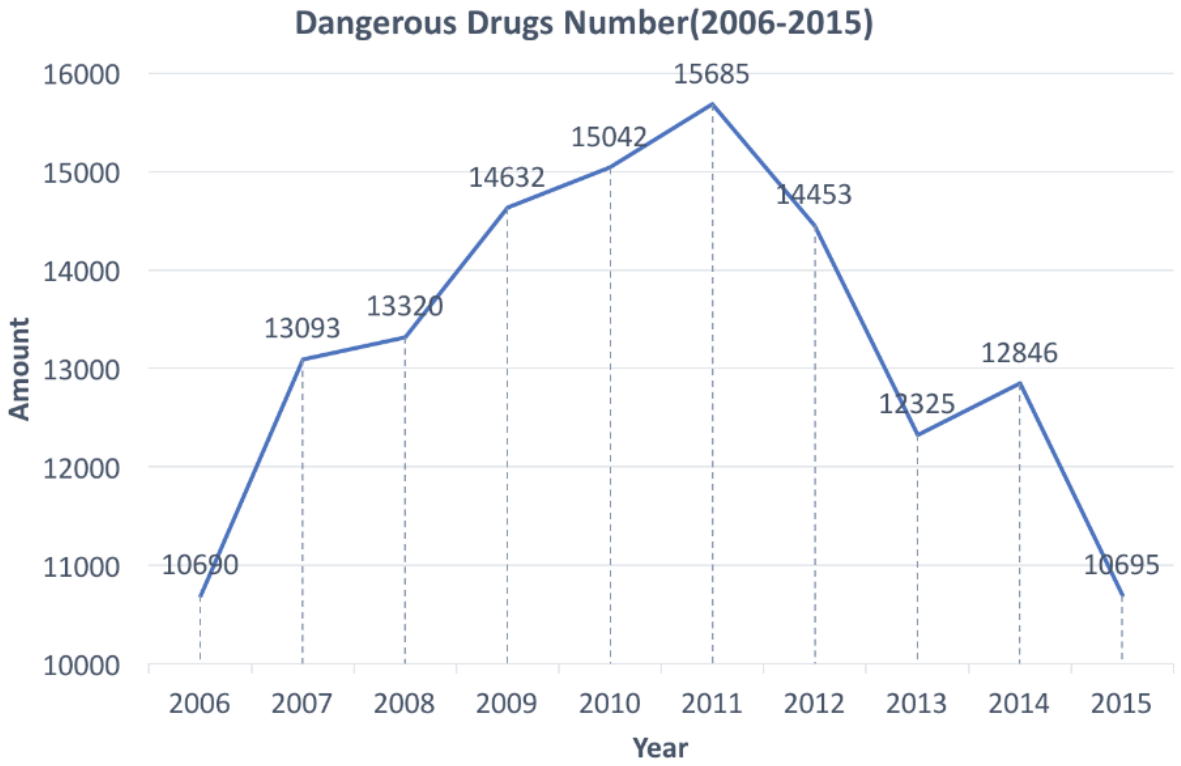

Figure 27: number of dangerous drugs

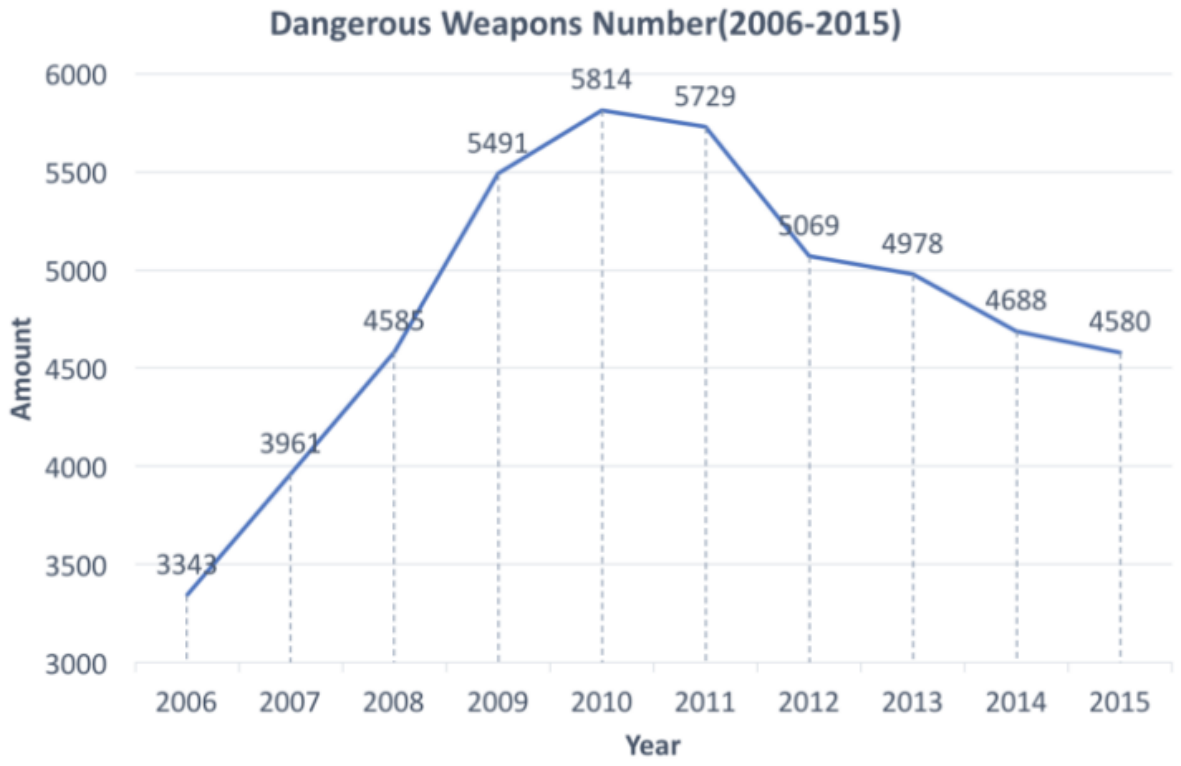

Figure 28: number of dangerous weapons

As we can see, dangerous drugs number on the streets has the trend of increasing from

year 2006 to year 2011. And from year 2012 to year 2015, it has a downward trend. The stop-and-frisk statistics also have the same pattern in the trend.
Meanwhile, Dangerous weapons number also has upward trend between year 2006 and 2011, and has downward trend between year 2012 and 2015.
The number that is going up means that more dangerous drugs and weapons crimes are found. With those plots, we can show that stop-and-frisk in New York City could really help against dangerous drugs and weapons crimes.
We also compute the correlation between stop-and-frisk number in each year and dangerous drugs(weapons) on the streets reported in each year.

$$corr(stop-and-frisk\ number\ in\ each\ year, dangerous\ drugs\ on\ the\ streets) = 0.66$$

$$corr(stop-and-frisk\ number\ in\ each\ year, dangerous\ weapons\ on\ the\ streets) = 0.26$$

We can see that correlation between stop-and-frisk and dangerous drugs is much higher than dangerous weapons.

4.2.4 Fine weather really causes an increase in theft and similar offense ?

When the weather is fine, people are more likely to go out and be absent from their home. You may infer that its more likely for property crimes to happen at residence house since the house is more likely to have no people in it. So we filter out the related data in this python file[link].

Main property crimes includes Larceny, Burglary, and Theft. So when filtering out the useful data, we set the criteria of offense description to have the keyword LARCENY, BURGLARY, or THEFT. And also, we limit the useful data to only have the location concerning residence housing since we dont want property crimes that happened elsewhere to pollute our filtered data.

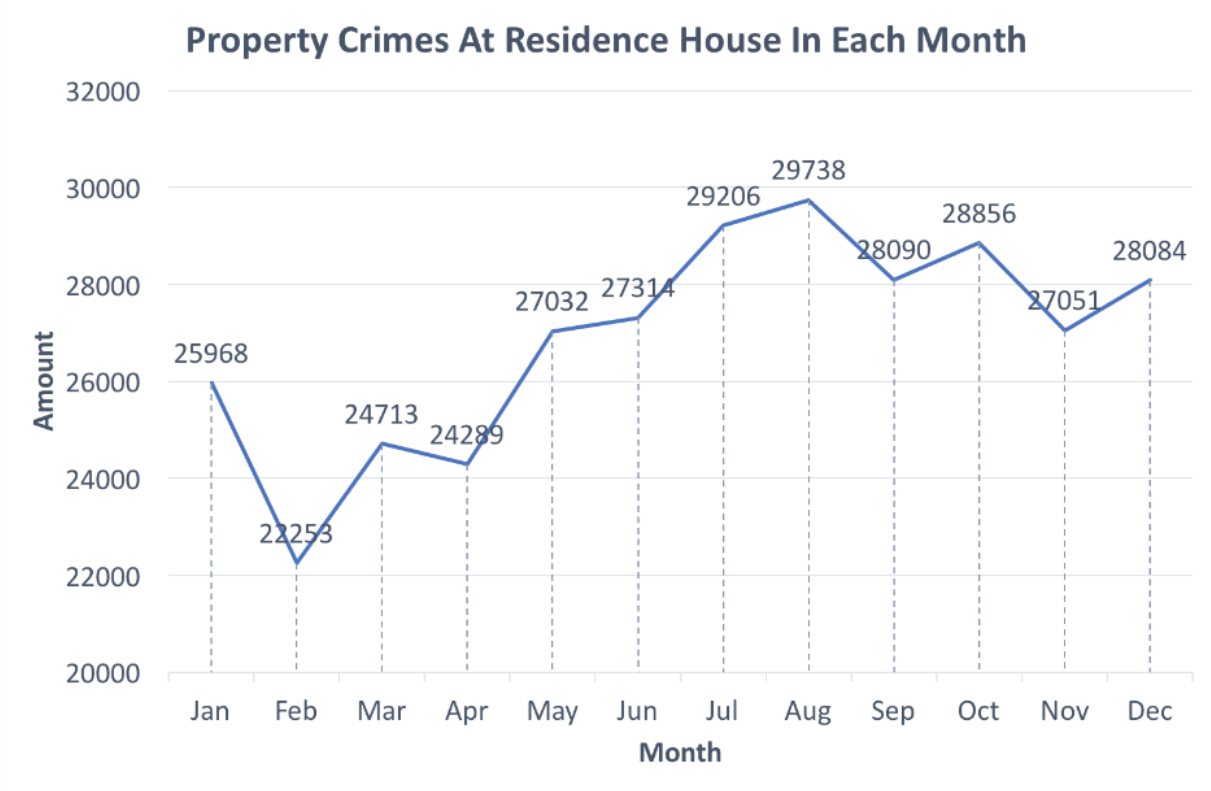

Figure 29: Property crimes in each month

After collecting the related data, we work out the statistics on property crimes at residence house in each month. From the plot, we can see that from January to April, the property crimes number is below 26000 per month. In others months, the numbers are all greater than 27000/month.

In order to show when people will probably be not at home, we searched on the internet and find this survey result. [data source link]

In this survey, 59% of respondents said that they have or will take a vacation during the summer. 46% of those respondents said they will take their summer vacation during July, making it the most popular month for summer travel. More specifically, the second week of July was the most popular week for summer travel. Just 7% of respondents said they took vacations during May. 11% took vacations during June. And 36% plan to take vacations in August.

From this survey, people will more likely to choose summer for vacation time. July and August are most popular during summer time. Property crime statistics also has its peak around July and August.

In all, property crimes are more likely happened in the summer time. People going out for vacations are one of the reasons that causes an increase in theft and similar offenses.

4.2.5 Does the number of tourists in NYC influence the number of crimes in Manhattan Midtown?

Every year, NYC attracts a lot of tourists to come and visit. We find out the most popular places that tourists(or even locals) will go. Source of the data is provided by Foursquare. [data source link[7]]

Top New York City Attractions and Restaurants Visited by Both Tourists and Locals

Attraction	Address	NYC Borough	Type of Attraction
Central Park	59th St to 110th St	Manhattan	Parks
Times Square	Broadway & 7th Ave	Manhattan	Plazas
Bryant Park	E 42nd St	Manhattan	Parks
UrbanSpace Vanderbilt	230 Park Avenue	Manhattan	Food Courts
Shake Shack	691 8th Ave	Manhattan	Burger Joints
Shake Shack	11 Madison Ave	Manhattan	Burger Joints

Figure 30: Top NYC attractions visited by tourists

From the data, we can see that, midtown Manhattan are always crowded with tourists since it is so popular to them. In our original dataset, there is one column that records the precinct code of the crime location area.

We limit our research crime data to midtown Manhattan and its nearby area. So we goto NYPD website to find out all the precinct codes related to that big area. [link]

NYPD | Precincts - Transit Districts - Housing PSAs

Manhattan

1st Precinct	(212) 334-0611	16 Ericsson Place
5th Precinct	(212) 334-0711	19 Elizabeth Street
6th Precinct	(212) 741-4811	233 West 10 Street
7th Precinct	(212) 477-7311	19 1/2 Pitt Street
9th Precinct	(212) 477-7811	321 East 5 Street
10th Precinct	(212) 741-8211	230 West 20th Street
13th Precinct	(212) 477-7411	230 East 21st Street
Midtown So. Pct.	(212) 239-9811	357 West 35th Street
17th Precinct	(212) 826-3211	167 East 51st Street
Midtown No. Pct.	(212) 767-8400	306 West 54th Street
19th Precinct	(212) 452-0600	153 East 67th Street
20th Precinct	(212) 580-6411	120 West 82nd Street
Central Park Pct.	(212) 570-4820	86th St & Transverse Road
23rd Precinct	(212) 860-6411	162 East 102nd Street
24th Precinct	(212) 678-1811	151 West 100th Street
25th Precinct	(212) 860-6511	120 East 119th Street
26th Precinct	(212) 678-1311	520 West 126th Street
28th Precinct	(212) 678-1611	2271-89 8th Avenue
30th Precinct	(212) 690-8811	451 West 151st Street
32nd Precinct	(212) 690-6311	250 West 135th Street
33rd Precinct	(212) 927-3200	2207 Amsterdam Avenue
34th Precinct	(212) 927-9711	4295 Broadway

Figure 31: Manhattan precincts

After some research, we finalize exact the precinct code range to be between 14 and 18. These areas covers most part of midtown Manhattan.

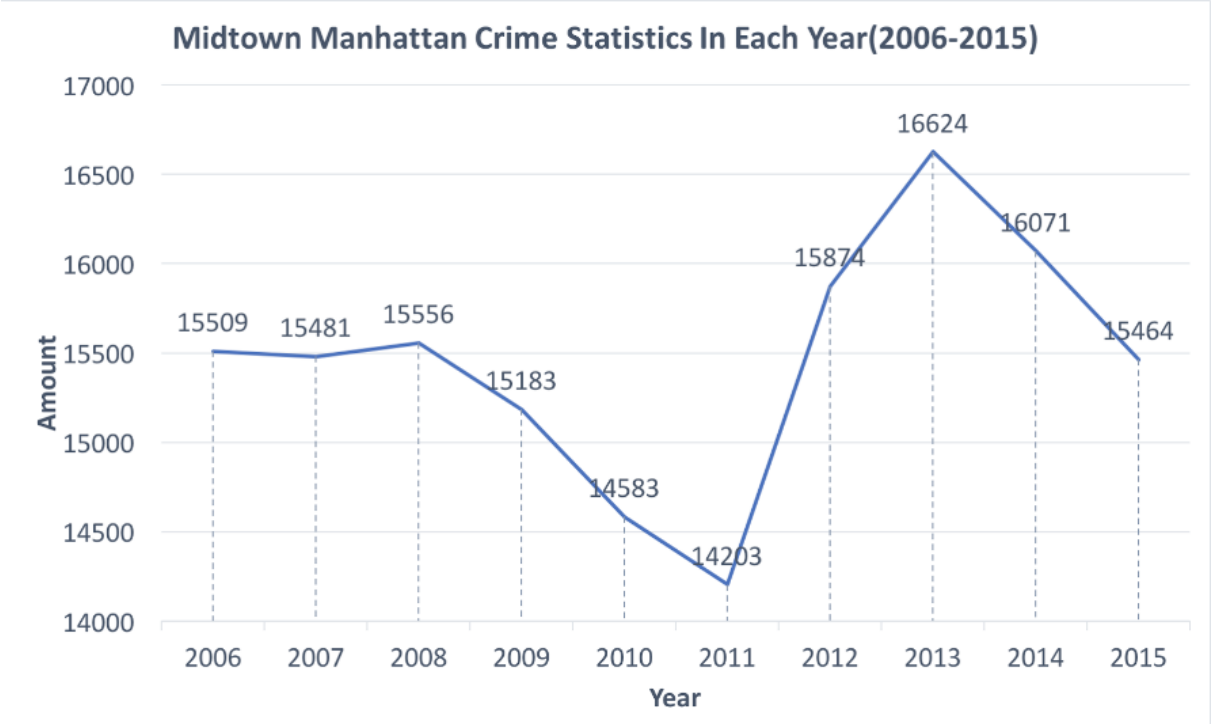

Figure 32: Midtown Manhattan crime statistics in each year

We can see that there is a downward trend from year 2008 to year 2011. Then an upward trend shows up from year 2011 to 2013. Whats really surprising is that there is a quick bounce-back from year 2011 to 2013.

We first think of if there was a deep fluctuation of the number of visitors to NYC between year 2010 to year 2013.

This plot shows the statistics about Tourism in New York City. [link]

Figure 33: total visitors in NYC

We can see that there is no visitors total number fluctuation between year 2009 and

year 2013. So we can rule out the possibility that this abnormal rise in Manhattan Midtown Crime is caused by a sudden fluctuation in visitors number.
Then we look back to our overall crime data distribution in each data.

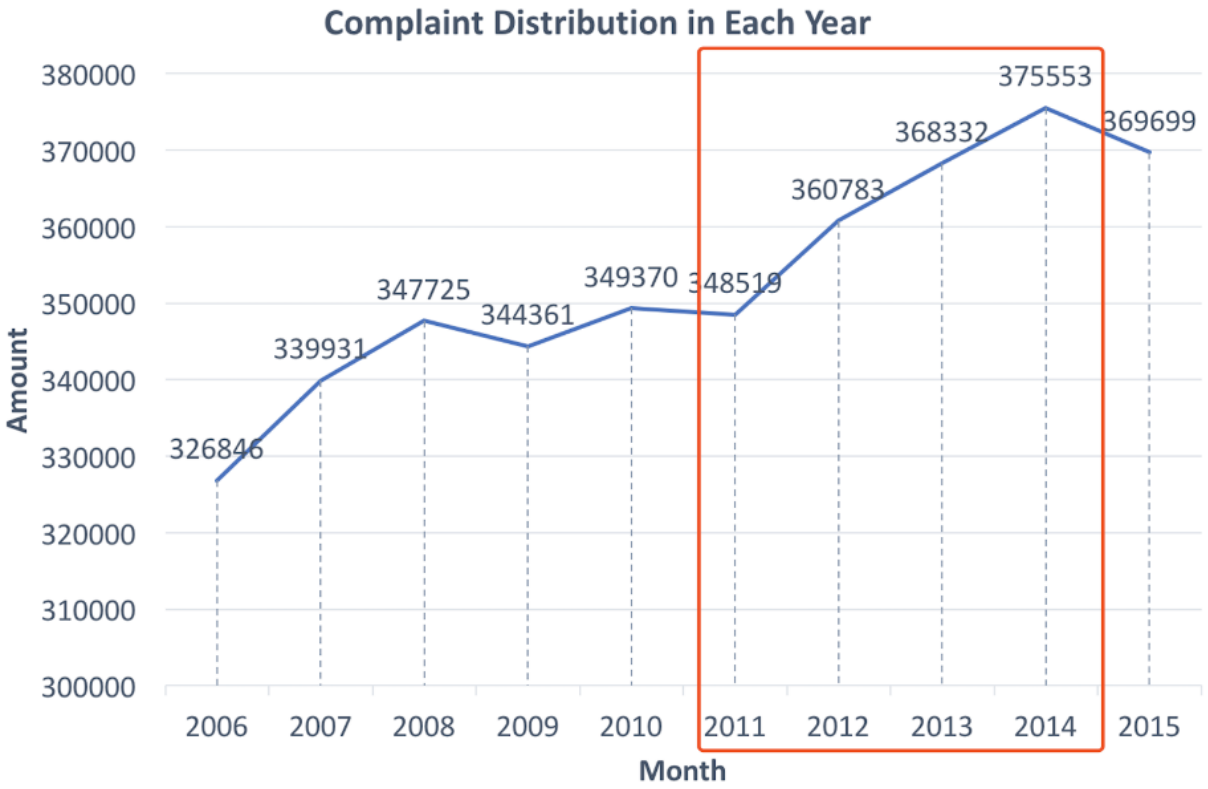

Figure 34: crime statistics in each year

We can see that there is also a quick rise in total crime number from year 2011 to year 2014. So The overall crime index in NYC has increased. As the heart of New York City, Manhattan also suffered increasing number of crimes between year 2011 and year 2013. That explains why there was a sharp upward trend of crimes statistics in Midtown Manhattan from year 2011 to year 2013.

Then we collect Midtown Manhattan Crime statistics in each month. We can see that there was a low point in February. And also there is a upward trend of number of crimes from April to the end of the year.

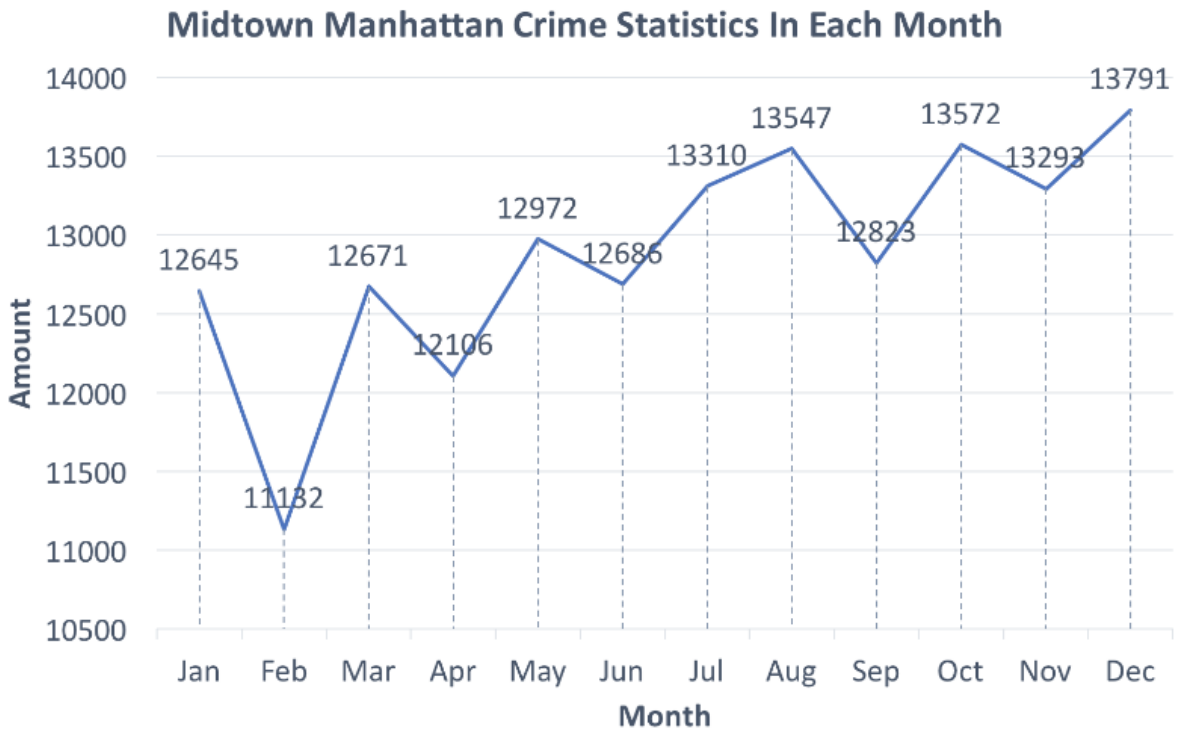

Figure 35: Midtown Manhattan crime statistics in each month

Then we need to know the distribution of visitors coming to NYC in each month. We analyse that distribution from the Manhattan Hotel Occupancy data. [link[8]]

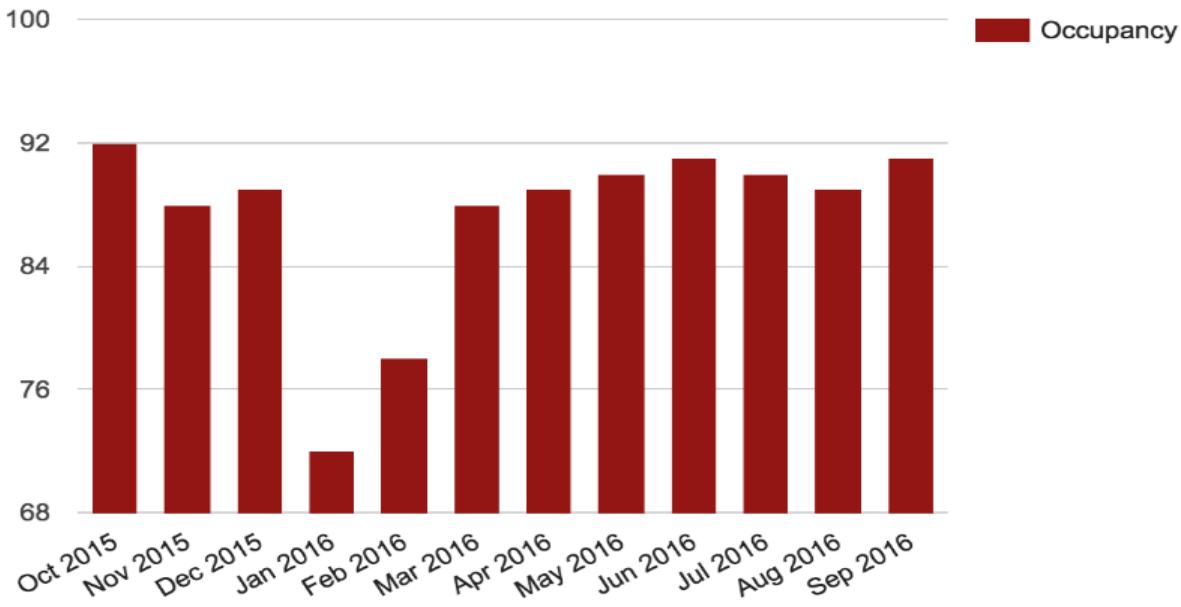

Figure 36: Manhattan Hotel Occupancy

We can see that between June and December, the manhattan hotel occupancy is al-

ways around 90%. This high occupancy shows the increased number of visitors coming to NYC. And we can also see that between June and December, the total number of crimes in Midtown Manhattan are all very high.

We also compute the correlation between number of tourists in NYC and the number of crimes in Manhattan Midtown.

$$corr(number\ of\ tourists\ in\ NYC, the\ number\ of\ crimes\ in\ Manhattan\ Midtown) = 0.52$$

In all, we can show that the number of visitors visiting NYC could greatly influence the crime number in Midtown Manhattan.

4.2.6 Does unemployment rate in NYC greatly influence crime rate?

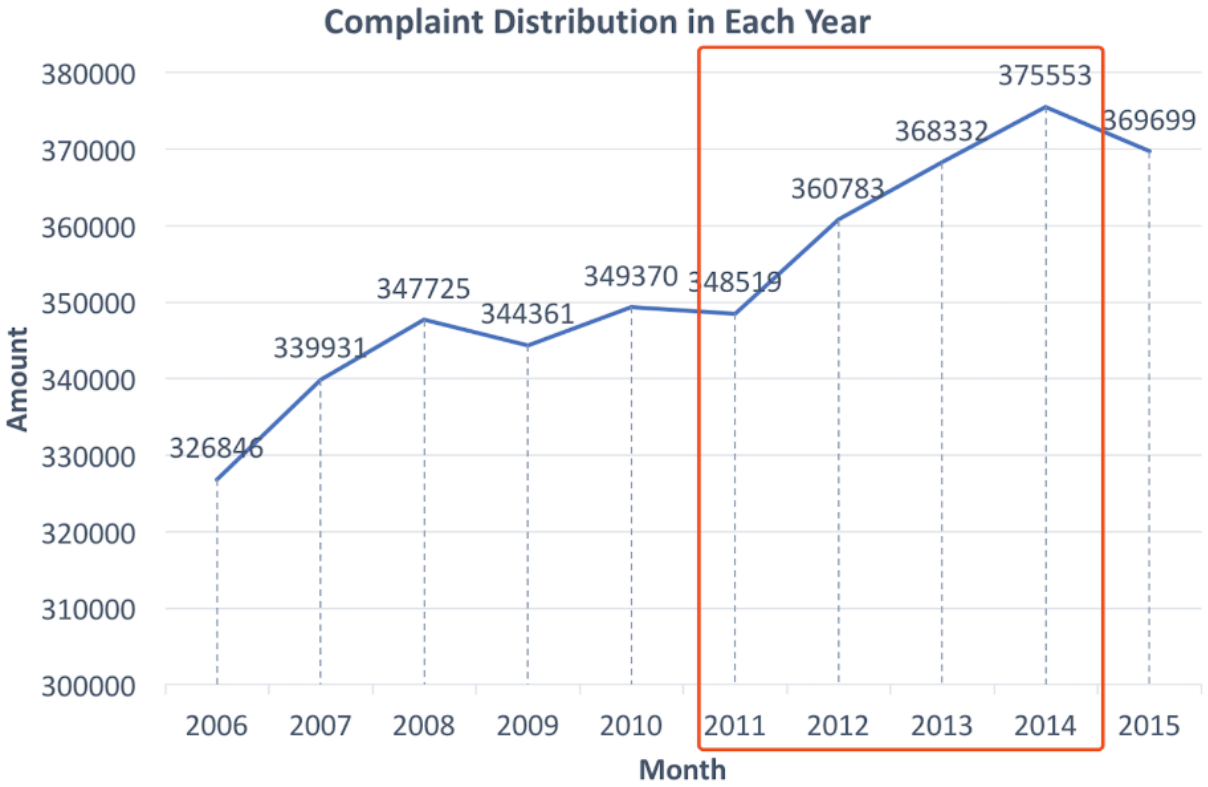

Figure 37: Crime number in each year

We find a quick rise in yearly number of crimes from year 2011 to year 2014.
So we have our assumption that unemployment rate greatly influence the total number of crimes in NYC.

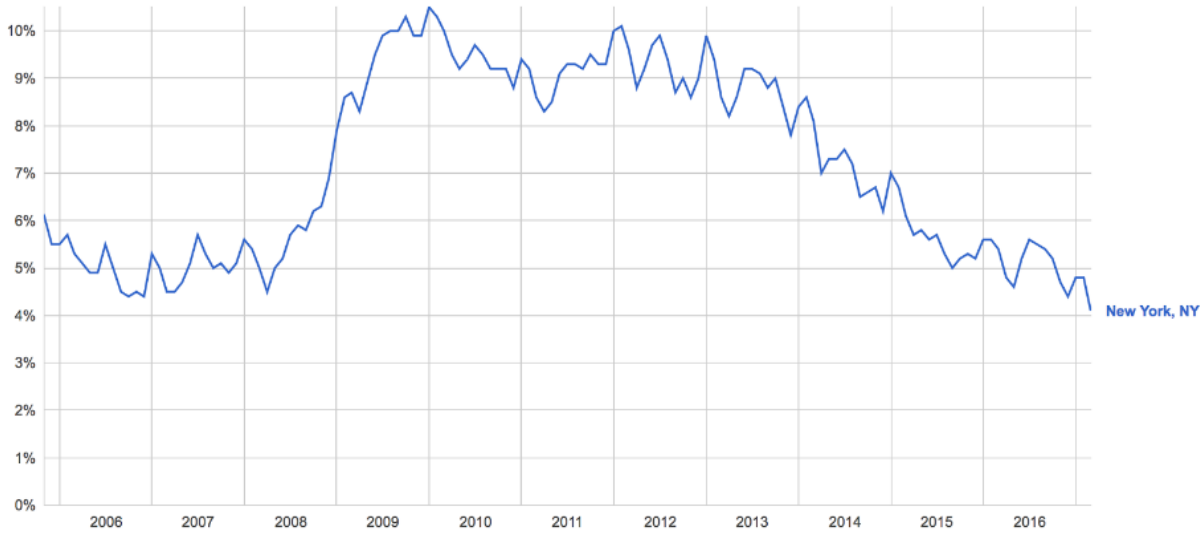

Figure 38: unemployment rate of NYC

This is the unemployment rate of New York City from year 2006 to year 2016. [link[9]]
The unemployment rate exceeds 10% in 2010, and maintained at a rather high ratio around 8.5% for about 2 3 years.
Why crime rate doesnt also reach its highest around year 2009?
Here is our explanation:
If someone loses his job someday, he could live on his savings at least for some period of time.
As the time goes, his saving could be use up gradually. Living in poverty, this person is more likely to be influenced to commit a crime for his living. So when the high unemployment rate lasts for a long time, number of crimes will increase.
Because unemployment rate fluctuation wont immediately influence overall crime index(its effects only shows up after a certain accumulation of time), counting the correlation between the unemployment rate and crime index is not that helpful.
In all, the unemployment rate will greatly influence the crime index in NYC.

4.2.7 correlation summary

factor1	factor2	correlation
NYCpopulation	NYCcrimenumber	0.72
CrimeNumber	Snowfall	-0.84
stop-and-frisk number	dangerous drugs on the street	0.66
stop-and-frisk number	dangerous weapons on the street	0.26
weather	property crimes at residence house	are correlated
number of tourists	number of crimes in midtown Manhattan	0.52

4.3 Anomalies and Outliers

4.3.1 Brooklyn 75th precinct - the most dangerous precinct

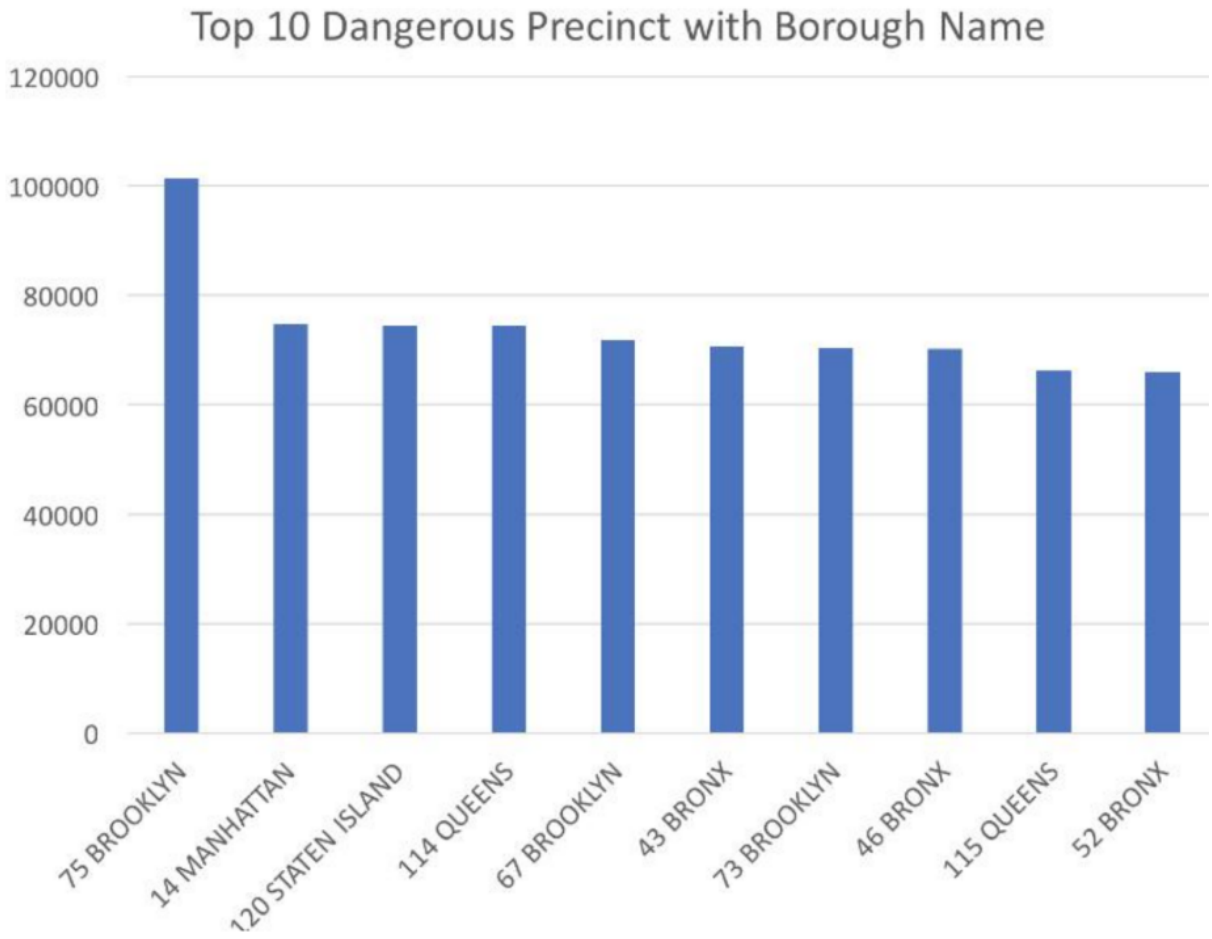

Figure 39: Top 10 dangerous precinct with borough name

As we can see from the figure above, 75th precinct in Brooklyn is the most dangerous place among NYC precincts. And this place is always reported by NYPD for crime incidents.

4.3.2 From year 2011 to 2014, the abnormally fast increase of crime number in NYC(especially for year 2012 when compared to 2011).

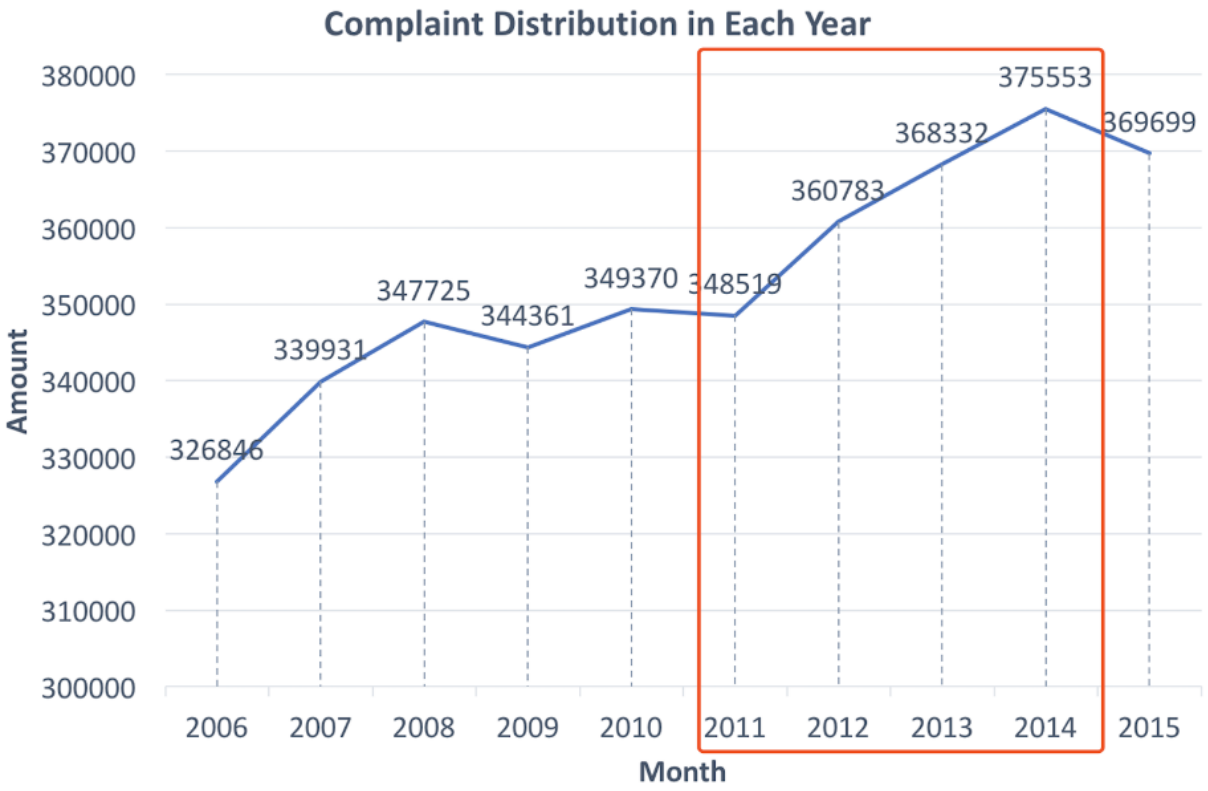

Figure 40: Compliant distribution in each year

We have analyzed this abnormal big jump on crimes total number before. It is due to the unemployment rate raise from year 2008 to year 2010.

4.3.3 The least number of crime - February

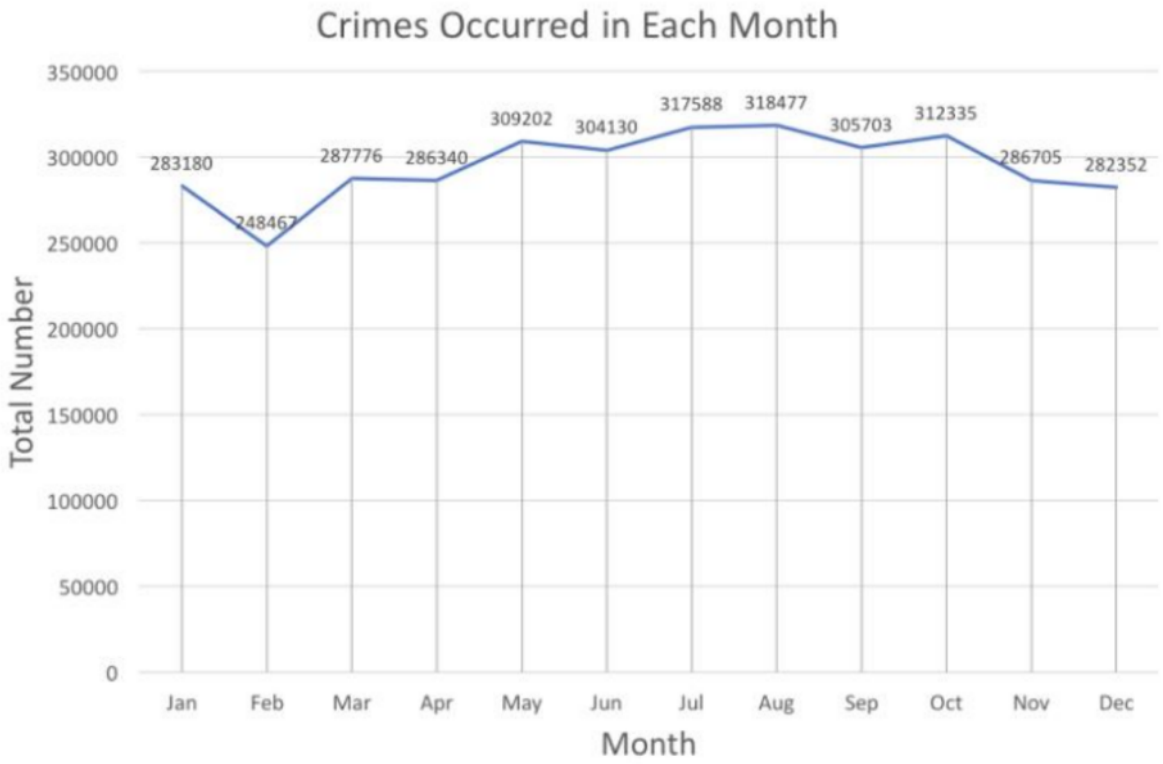

Figure 41: Crime distribution in each month

We have also analyzed this anomaly before. It is due to there are less days in February compared to other months and also the weather in February is also among the worst in the whole year.

4.3.4 Daytime crime number outlier

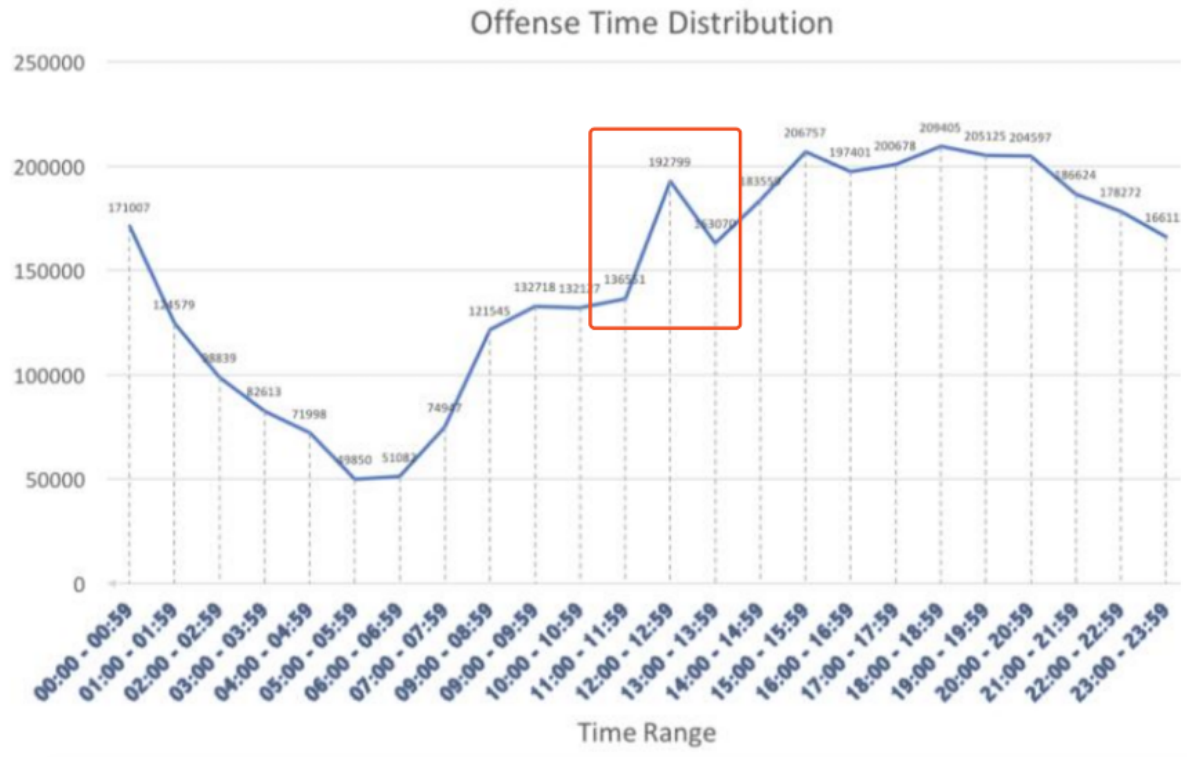

Figure 42: Crime distribution with respect to time

We find out that there is a big fluctuation between 12pm and 2pm. We could not find the reason leading to this big jump. Maybe its partly because of the human beings normal life pattern.

5 Individual contributions

Minda Fang: Participated in writing data issue detecting code. Participated in writing data summary code. Participated in writing the final report. Participated in finding the correlation with other dataset.
Qiming Zhang: Participated in writing data issue detecting code. Participated in writing data summary code. Participated in writing the final report. Participated in finding the correlation with other dataset.
Mindi Mao: Participated in writing data summary code. Accessed dumbo to generate the cleaned dataset. Participated in writing the final report. Participated in finding data outliers.

6 Conclusions

After we finished all analysis in data summary part, we found the following facts about NYC:

- Brooklyn is the most dangerous precinct and staten island is the safest one in whatever analysis dimensionality we use.

- Crime occurs more in the summer than in the winter. Crime number is lowest in February and highest in August.
- For the crime offense time, from 6 am to 7pm, the crime number has an upward trend. From 7pm to 6am, it has a downward trend.
- Misdemeanor level crimes account for 57
- Most dangerous place is the street.

And then we have our data exploration work and have these following interesting findings:

- NYC population fluctuation directly influences the overall crime number.
- Weather influences crime statistics. Cold weather will make people less likely to commit a crime.
- NYC stop-and-frisk policy can help detecting dangerous drugs & weapons crimes on the street.
- Fine weather causes an increase in property crimes.
- Tourists number in NYC influences the crime number in Midtown Manhattan(visitors most possible visiting places).
- Unemployment rate rising greatly stimulated the rising in the crime number.

We find 4 main anomalies and outliers in our analysis:

- Brooklyn 75th precinct is the most dangerous precinct.
- From year 2011 to 2014, yearly crime number rise suddenly and sharply.
- February crime number is the lowest, which is far less than any other months.
- During 12pm to 2pm, there is an abnormal fluctuation in the hourly crime number.

7 Reference

1. Project Code link
2. NYC population
3. Cleaned data
4. Crime statistics in USA from year 2006 to year 2012
5. NYC snowfall record
6. Stop-and-frisk in New York City
7. Top NYC Attractions and restaurants data
8. Manhattan Hotel Occupancy data
9. Unemployment rate of New York City from year 2006 to year 2016