



NYU

**TANDON SCHOOL
OF ENGINEERING**

Project: Analyzing NYPD Complaint Data (Historic)

Team Members	NetId
Niranjhana Nayar	nn1024
Priyanka Bhadoriya	pb1826
Suchetha Siddagangappa	ss11436

GitHub path : <https://github.com/pb1826/Big-Data-Project>

Contents

Project: Analyzing NYPD Complaint Data (Historic)	1
Contents	2
Abstract	5
Introduction	6
Part I : Data Cleaning	7
Column Details	7
Descriptive summary of the data quality issues :	8
Strategies for cleaning	9
Code used to recognise NULL values :	9
Code that reports statistical information : (No. of invalid values)	10
Data Cleaning	14
Part 2 : Data Summary	16
Brief Overview	16
OFFENCE DETAILS (KY_CD and OFSC_DESC)	17
CRIMES IN NYC	21
TOP 10 CRIMES IN NYC	22
Borough wise	23
ANALYSIS BOROUGH WISE	24
Bottom 10 Crimes in NYC	25
Borough wise	26
PRECINCT	28
Top 10 Precinct with Highest Crimes Rates	28
75th Precinct	28
CENTRAL PARK PRECINCT	29
MOST UNSAFE PLACES TO LIVE IN NYC	30
Borough Wise	31
SAFEST PLACES TO LIVE	32
Borough Wise	33
HAVDEVLOPMENT CRIME WISE	34
Borough Wise	35
PARKS	37
MOST SAFE PARKS IN NYC	37
MOST UNSAFE PARKS IN NYC	37
Borough Wise	38

TIME	40
TIME AM	41
TIME IN PM	42
Borough Wise	43
Bronx	43
Lowest	43
Highest	44
Brooklyn	45
Lowest	45
Highest	46
Queens	47
Lowest	47
Highest	48
Manhattan	49
Lowest	49
Highest	50
Staten	51
Lowest	51
Highest	52
DAY	53
BOROUGH WISE	54
WEEKDAY	55
BOROUGH WISE	56
DAY CRIME WISE	57
BOROUGH WISE	58
MONTH	60
Borough Wise	60
Borough Wise	63
YEAR	65
Borough Wise	66
YEAR CRIME WISE	68
Borough Wise	69
Geo-Location	71
Bronx	71
Brooklyn	72
Manhattan	73
Queens	74

Staten	75
DATA QUALITY	76
Redundant Data	76
DATA SUMMARY	77
DATA EXPLORATION	79
Key Calendar Dates	79
Halloween	79
New Year	80
Thanksgiving	80
9/11	81
July 4	82
Christmas	82
Memorial Day	83
Labour Day	83
Employment	85
Population	85
Weather	88
Summary	90
Technologies Used	91
Tools Used	91
Individual Contribution	92
References	92

Abstract

The goal of this project is to clean the data in a dataset, analyse it and generate a descriptive summary. Also we have done some data exploration with other data sets. The cleaned data (from the first part) is used to generate a descriptive summary for the features and contents. The dataset used (NYPD Complaint Data Historic) includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2016).

Introduction

In the data cleaning part, we have looked for data quality issues as well as anomalies such as the ones mentioned below.

- 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?
- Are there different kinds of values in the same column (e.g., integers and strings)?
- 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)

In the data analysis part, we have looked at each feature and given a summary about it. Any meaningful understanding from the features was recorded and plotted. We used had cleaned all 24 columns. The key columns we considered for analysis were:-

1. KY_CD
2. OFSC_DESC
3. CMPLNT_FR_DT
4. CMPLNT_FR_TM
5. ADDR_PCT_CD
6. BORO_NM
7. Lat_Lon
8. HADEVELOPT
9. PARK_NM

The other columns either had invalid data or redundant data. We performed analysis overall and then borough wise. First, we obtained we ran the pyspark scripts on dumbo and then we used the generated csvs for the visualization. The visualization was done using jupyter notebook.

In the data exploration we analysed the dataset from NYPD with Employment, Weather, Key Dates and Population. We wanted to find out whether the crime rates had any correlation with the climate, financial crisis, unemployment rates. We did analysis for key dates like Halloween, Labour day, New Year and so on to understand if the heavy police presence makes an impact or is it unnecessary.

Part I : Data Cleaning

Column Details

Color Codes:- Red(ignore), Green(completely valid) and Yellow(validated)

#	Column Name	NULLS	INVALID S	VALID S	TYPE	CHECK FOR INVALID
0	CMPLNT_NUM	0	0	5580035	String(num)	<u>check_col1()</u>
1	CMPLNT_FR_DT	655	18972	5560408	Date	<u>check_date()</u>
2	CMPLNT_FR_TM	48	903	5579084	Time	<u>check_time()</u>
3	CMPLNT_TO_DT	1472786	4938	4102311	Date	<u>check_date()</u>
4	CMPLNT_TO_TM	1468882	1376	4109777	Time	<u>check_time()</u>
5	RPT_DT	0	0	5580035	Date	<u>check_date()</u>
6	KY_CD	0	0	5580035	String(num)	<u>check_id()</u>
7	OFNS_DESC	18892	0	5561143	String	<u>check_id()</u>
8	PD_CD	4909	0	5575126	String	<u>check_desc()</u>
9	PD_DESC	4909	0	5575126	String	<u>check_desc()</u>
10	CRM_ATPT_CPTD_CD	7	0	5580028	String	<u>check_action()</u>
11	LAW_CAT_CD	0	0	5580035	String	<u>check_lawcd()</u>
12	JURIS_DESC	0	0	5580035	String	<u>check_desc()</u>
13	BORO_NM	463	0	5579572	String	<u>check_boro()</u>
	QUEENS	0	0	1105621		
	BRONX	0	0	1666903		
	BROOKLYN	0	0	1331760		
	MANHATTAN	463	0	1331760		
	STATEN ISLAND	0	0	265641		
14	ADDR_PCT_CD	390	0	5579645	String(num)	<u>check_addrpctcd()</u>
15	LOC_OF_OCCUR_DESC	1223392	213	4356430	String	<u>check_loc()</u>
16	PREM_TYP_DESC	35198	0	5544837	String	<u>check_desc()</u>
17	PARKS_NM	5567497	0	12538	String	<u>check_desc()</u>

18	HADEVELOPT	5302218	0	277817	String	<u>check_desc()</u>
19	X_COORD_CD	0	195868	5384167	Coordinates	<u>check_xloc()</u>
20	Y_COORD_CD	0	195868	5384167	Coordinates	<u>check_yloc()</u>
21	Latitude	0	195868	5384167	Coordinates	<u>check_lat()</u>
22	Longitude	0	195868	5384167	Coordinates	<u>check_lon()</u>
23	LatLon	0	195868	5384167	Coordinates	<u>lat_lon()</u>

check_date()	Check if the year is between 2006-2016
check_time()	Check whether the time is in 24 hour format
check_id()	Check whether id is 3 character long
check_desc()	check if the field is a string
check_action()	Valid values are COMPLETED, ATTEMPTED
check_boro()	check if the field is a valid New York borough. Valid boroughs are : MANHATTAN, BROOKLYN, BRONX, QUEENS, STATEN ISLAND
	The precinct in which the incident occurred 1 to 123
check_xloc()	valid x coord range : from 909900 to 1067600
check_yloc()	valid y coord range : from 117500 to 275000
check_lat()	valid latitude range : between -73 and -75
check_lon()	valid longitude range : between 40 and 41

Descriptive summary of the data quality issues :

- Total no. of columns - 24
- 5 columns contain no invalid or null values.
- 14 columns contain NULL values. The highest NULL values are present in column PARKS_NM and the least is present in CRM_ATPT_CPTD_CD.
- 10 columns contain invalid values. Highest no. of invalid values are found in columns x_coord_cd, y_coord_cd, latitude, longitude, lat_lon
- All the columns had only a single data type

Strategies for cleaning

- From the above data it can be seen that CMPLNT_NUM,RPT_DT,KY_CD,LAW_CAT_CD and JURIS_DESC is always valid and can be considered as the most important features.
- CMPLNT_FR_DT should never be null. However CMPLNT_TO_DT can be null in case CMPLNT_FR_DT is present for that record.
- Similar logic applies for CMPLNT_FR_TM and CMPLNT_TO_TM.
- Similarly LOC_OF_OCCUR_DESC, PREM_TYP_DESC, PARKS_NM and HADEVELOPT need not be present for all the fields so cannot be used to eliminate records.
- X_COORD_CD,Y_COORD_CD,Latitude and Longitude conveys the same data as LatLon. Hence, LatLon is used for cleaning the data.
- CMPLNT_TO_DT, CMPLNT_TO_TM, OFNS_DESC, PD_CD, PD_DESC, CRM_ATPT_CPTD_CD, BORO_NM, ADDR_PCT_CD, LatLon are used for filtering VALID data.
- There maybe null values in the CMPLNT_TO_DT, CMPLNT_TO_TM, LOC_OF_OCCUR_DESC, PREM_TYP_DESC, PARKS_NM and HADEVELOPT, X_COORD_CD,Y_COORD_CD,Latitude and Longitude columns.
- Appropriate replacement of these null values are done. ([Code](#))
 - For dates **01/01/1900** was given
 - For time **00:00:00** was given
 - For strings **NA** was given
 - The coordinate fields were never null so it was not replaced with any value
- The final csv is then generated. ([Code](#))

Code used to recognise NULL values :

```
def check_null(x):
    try:
        if x=="":
            return True;
    except ValueError(x):
        return False;
```

Code that reports statistical information : (No. of invalid values)

check_col1()

```
def check_col1(x):
    try:
        if(check_null(x)):
            return "NULL";
        elif re.match('[1-9][0-9]+',x):
            return "VALID", check_type(x);
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
```

check_type()

```
def check_type(x):
    try:
        switcher={
            int: "INT",
            float: "FLOAT",
            long: "LONG",
            str: "STRING",
        }
        return switcher.get(type(x),"NULL")
    except ValueError:
        return "INVALID";
```

check_date()

```
def check_date(x):
    # check if the field is null or if it is between the years 2006-2016
    try:
        if(check_null(x)):
            return "NULL";

        elif(re.match('([0?][1-9]|1[0-2])/([3][01]|1[12])[0-9]|0?[1-9])/([20])(0[6-9]|1[0-6])$',x)):
            return "VALID","DATE";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
```

```

check_time()
def check_time(x):
    #check if the field is null or if it is in 24 hr format
    try:
        if(check_null(x)):
            return "NULL";
        elif(re.match('(?::(?:([01]?\d|2[0-3]):)?([0-5]?\d):)?([0-5]?\d)$',x)):
            return "VALID","TIME";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";

```

```

check_desc()
def check_desc(x):
    #check if the field is null or is a string
    try:
        if(check_null(x)):
            return "NULL";
        elif(type(x)==str):
            return "VALID","STRING";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
    . . .

```

```

check_boro()
def check_boro(x):
    #check if the field is null or in the list of new york boroughs
    try:
        if(check_null(x)):
            return "NULL";
        elif(x in ['MANHATTAN', 'BRONX', 'BROOKLYN', 'QUEENS', 'STATEN ISLAND']):
            return "VALID","STRING";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";

```

```

check_action()
def check_action(x):
    try:
        if(check_null(x)):
            return "NULL";
        elif(x in ['COMPLETED', 'ATTEMPTED']):
            return "VALID","STRING";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";

```

check_xloc()

```
def check_xloc(x):
    #check if item is null or if it is a coordinate
    #X-coordinate (East-West): minimum: 909900; maximum: 1067600
    try:
        x=int(x);
        if(check_null(x)):
            return "NULL";
        elif((909900<x) & (x<1067600)):
            return "VALID","COORDINATES";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
```

check_yloc()

```
def check_yloc(x):
    #check if the item is null or if it is a coordinate
    #for nyc Y-coordinate (North-South): minimum: 117500; maximum: 275000
    try:
        x=int(x);
        if(check_null(x)):
            return "NULL";
        elif((117500<x) & (x<275000)):
            return "VALID","COORDINATES";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
```

check_lon()

```
def check_lon(x):
    #checking if the field is null or if it is a lon between -73 and -75
    try:
        x=float(x);
        if(check_null(x)):
            return "NULL";
        elif((-75<x) & (x<-73)):
            return "VALID","COORDINATE";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
```

check_lat()

```

def check_lat(x):
    #checking if the field is null or if it is a lat in between 40 and 41
    try:
        x=float(x);
        if(check_null(x)):
            return "NULL";
        elif((40<x) & (x<41)):
            return "VALID", "COORDINATE";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
    ...

check_loc()

def check_loc(x):
    try:
        occur = ['FRONT OF', 'INSIDE', 'REAR OF', 'OUTSIDE', 'OPPOSITE OF']
        if(check_null(x)):
            return "NULL";
        elif x in occur:
            return "VALID", "STRING";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
    ...

```

```

check_lawcd()

def check_lawcd(x):
    try:
        law = ['MISDEMEANOR', 'VIOLATION', 'FELONY'];
        if(check_null(x)):
            return "NULL";
        elif x in law:
            return "VALID", "STRING";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
    ...

```

```

check_addrpected()

def check_addrpected(x):
    try:
        if (check_null(x)):
            return "NULL";
        elif (int(x)>=1 and int(x)<=123):
            return "VALID", "INT";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";
    ...

```

```

lat_lon()
def lat_lon(x):
    try:
        x=x.replace('(', '').replace(')', '');
        lat,lon=x.split(',');
        a=check_lat(lat);
        b=check_lon(lon);
        if(a[0]=="VALID" and b[0]=="VALID"):
            return "VALID", "STRING";
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";

check_id()
def check_id(x):
    #check if the field is null or 3 digits long
    try:
        if(check_null(x)):
            return "NULL";
        elif(re.match('[1-9][0-9][0-9]$',x)):
            return "VALID",check_type(x);
        else:
            return "INVALID";
    except ValueError:
        return "INVALID";

```

Data Cleaning

```

cleaned=output.filter(lambda x: x[2][0]=='VALID' and \
x[4][0]=='VALID' and \
x[10][0]=='VALID' and \
x[12][0]=='VALID' and \
x[14][0]=='VALID' and \
x[16][0]=='VALID' and \
x[20][0]=='VALID' and \
x[22][0]=='VALID' and \
x[32][0]=='VALID').map(lambda x: (x[0],\
x[1],x[3],x[5],x[6],x[7],x[8],x[9],x[11],x[13],x[15],x[17],x[18],x[19],x[21],x[23],\
x[24],x[25],x[26],x[27],x[28],x[29],x[30],x[31]));

file=cleaned.map(toCSV);
file.saveAsTextFile("cleaned.csv");

```

Nulls

```

# the coordinate features don't have any null values so it is not been considered
#cmplt_to_dt
cleaned=cleaned.map(lambda x : (x[0],x[1],x[2],x[3],x[4],\
(x[5].replace('','datetime.strptime('01/01/1900', "%m/%d/%Y").strftime('%m/%d/%Y'))),\ 
x[6],x[7],x[8],x[9],x[10],x[11],x[12],x[13],x[14],x[15],x[16],x[17],x[18],x[19],x[20],\ 
x[21],x[22],x[23],x[24],x[25],x[26],x[27],x[28],x[29],x[30],x[31],x[32]) if check_null(x[5])==True else (x));
#cmplt_to_tm
cleaned=cleaned.map(lambda x : (x[0],x[1],x[2],x[3],x[4],x[5],\ 
(x[6].replace('','datetime.strptime('00:00:00', "%H:%M:%S").strftime('%H:%M:%S'))),\ 
x[7],x[8],x[9],x[10],x[11],x[12],x[13],x[14],x[15],x[16],x[17],x[18],x[19],x[20],\ 
x[21],x[22],x[23],x[24],x[25],x[26],x[27],x[28],x[29],x[30],x[31],x[32]) if check_null(x[6])==True else (x));
#LOC_OF_OCCUR_DESC
cleaned=cleaned.map(lambda x : (x[0],x[1],x[2],x[3],x[4],x[5],x[6],x[7],x[8],x[9],x[10],x[11],\ 
x[12],x[13],x[14],x[15],x[16],x[17],x[18],x[19],x[20],x[21],x[22],\ 
(x[23].replace('','NA')),\ 
x[24],x[25],x[26],x[27],x[28],x[29],x[30],x[31],x[32]) if check_null(x[23])==True else (x));
#PREM_TYP_DESC
cleaned=cleaned.map(lambda x : (x[0],x[1],x[2],x[3],x[4],x[5],x[6],x[7],x[8],x[9],x[10],x[11],\ 
x[12],x[13],x[14],x[15],x[16],x[17],x[18],x[19],x[20],x[21],x[22],x[23],\ 
(x[24].replace('','NA')),\ 
x[25],x[26],x[27],x[28],x[29],x[30],x[31],x[32]) if check_null(x[24])==True else (x));
#PARKS_NM
cleaned=cleaned.map(lambda x : (x[0],x[1],x[2],x[3],x[4],x[5],x[6],x[7],x[8],x[9],x[10],x[11],\ 
x[12],x[13],x[14],x[15],x[16],x[17],x[18],x[19],x[20],x[21],x[22],x[23],x[24],\ 
(x[25].replace('','NA')),\ 
x[26],x[27],x[28],x[29],x[30],x[31],x[32]) if check_null(x[25])==True else (x));
#HADEVELOPT
cleaned=cleaned.map(lambda x : (x[0],x[1],x[2],x[3],x[4],x[5],x[6],x[7],x[8],x[9],x[10],x[11],\ 
x[12],x[13],x[14],x[15],x[16],x[17],x[18],x[19],x[20],x[21],x[22],x[23],x[24],x[25],\ 
(x[26].replace('','NA')),\ 
x[27],x[28],x[29],x[30],x[31],x[32]) if check_null(x[26])==True else (x));

```

Part 2 : Data Summary

Brief Overview

- As we looked at the list of data, I realised for the same KYCD, there are multiple offense code. Hence, OFSC_DESC is not a feature that can be relied upon.
- Crimes Borough wise sorted on crime count:-

#	Borough	Count
1	Staten Island	265641
2	Queens	1105621
3	Bronx	1209647
4	Manhattan	1331760
5	Brooklyn	1666903

- PD_CD and PD_DESC combo was a better feature than KY_CD and OFSC_DESC in that it had less repetition of data.
- The different OFSC_DESC field matched to the same KY_CD, more times than with PD_CD and PD_DESC
- However, the data in PD_CD ad PD_DESC are too in detail and would not make much difference to a normal civilian. However, it is the more important feature for law and order.
- For these reasons, we as students being civilians chose to pick KY_CD and OFSC_DESC combo over PD_CD and PD_DESC, else it would make the data they provide more or less redundant.
- For the same reason LAW_CAT_CODE is too broad for a civilian to understand the crime.
- It does not matter to a civilian whether the crime was completed or attempted. They only care about its occurrence.
- Precinct code is useful information for a civilian when choosing where to live in nyc.
- Crimes in parks and the housing areas is also useful for the civilians when choosing their residence.
- PARK_NM field was null for majority as it is only required if a crime had occurred in a park.
- Housing Development field is also null most of the times but it contained useful information when not null.

- Time was also an important field and we did analysis with AM n PM as well
- Day, Weekday, Month n Year we obtained from the CMPLNT_FR_DT
- Location Coordinates field had very important information in finding the most unsafe locations.

OFFENCE DETAILS (KY_CD and OFSC_DESC)

Here is the list of offenses by the key code and description

#	KY_CD	Description	Count
1	102	HOMICIDE-NEGLIGENT-VEHICLE	84
2	103	HOMICIDE-NEGLIGENT, UNCLASSIFIED	33
3	105	ROBBERY	208313
4	106	FELONY ASSAULT	200994
5	107	BURGLARY	201644
6	109	GRAND LARCENY	456326
7	110	GRAND LARCENY OF MOTOR VEHICLE	105101
8	111	POSSESSION OF STOLEN PROPERTY	9549
9	112	THEFT-FRAUD	58031
10	113	FORGERY	53950
11	114	ARSON	14462
12	115	PROSTITUTION & RELATED OFFENSES	54
13	116	SEX CRIMES	28
14	117	DANGEROUS DRUGS	66146
15	118	DANGEROUS WEAPONS	54873
16	119	INTOXICATED/IMPAIRED DRIVING	41

17	120	ENDAN WELFARE INCOMP	40
18	120	CHILD ABANDONMENT/NON SUPPORT	378
19	121	CRIMINAL MISCHIEF & RELATED OF	80478
20	122	GAMBLING	124
21	123	ABORTION	4
22	124	KIDNAPPING	9
23	124	KIDNAPPING & RELATED OFFENSES	2383
24	124	KIDNAPPING AND RELATED OFFENSES	2
25	125	NYS LAWS-UNCLASSIFIED FELONY	4152
26	126	MISCELLANEOUS PENAL LAW	120982

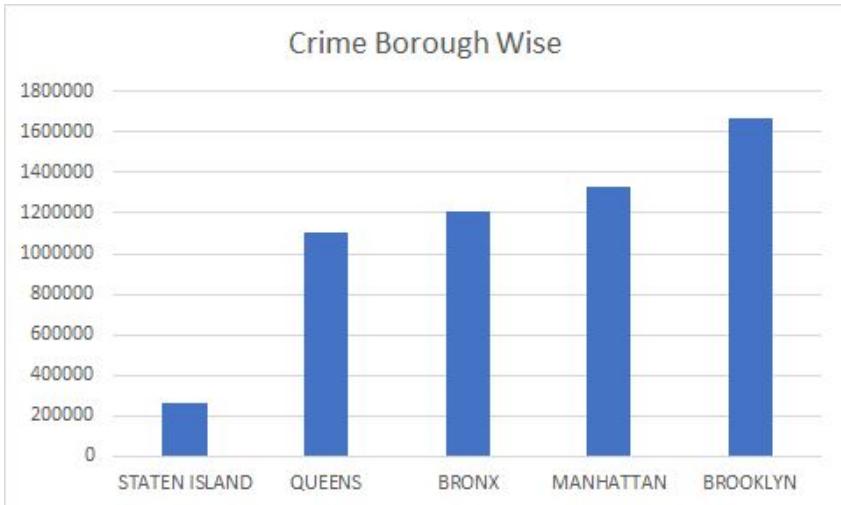
27	230	JOSTLING	334
28	231	"BURGLARS TOOLS"	2684
29	232	POSSESSION OF STOLEN PROPERTY	20892
30	233	SEX CRIMES	48
31	234	PROSTITUTION & RELATED OFFENSES	111
32	235	DANGEROUS DRUGS	297283
33	236	DANGEROUS WEAPONS	75731
34	237	ESCAPE 3	170
35	238	FRAUDULENT ACCOSTING	2764
36	340	FRAUDS	33835
37	341	PETIT LARCENY	880446

38	342	PETIT LARCENY OF MOTOR VEHICLE	909
39	343	OTHER OFFENSES RELATED TO THEF	11090
40	343	THEFT OF SERVICES	3166
41	344	ASSAULT 3 & RELATED OFFENSES	564535
42	345	ENDAN WELFARE INCOMP	133
43	345	OFFENSES RELATED TO CHILDREN	1435
44	346	ALCOHOLIC BEVERAGE CONTROL LAW	840
45	347	INTOXICATED & IMPAIRED DRIVING	73317
46	348	VEHICLE AND TRAFFIC LAWS	58561
47	349	DISRUPTION OF A RELIGIOUS SERV	50
48	350	GAMBLING	2151
49	351	CRIMINAL MISCHIEF & RELATED OF	461883
50	352	CRIMINAL TRESPASS	68827
51	353	UNAUTHORIZED USE OF A VEHICLE	15003
52	354	ANTICIPATORY OFFENSES	96
53	355	OFFENSES AGAINST THE PERSON	13057

54	356	PROSTITUTION & RELATED OFFENSES	706
55	357	FORTUNE TELLING	11
56	358	OFFENSES INVOLVING FRAUD	18664
57	359	OFFENSES AGAINST PUBLIC ADMINI	107042
58	360	LOITERING FOR DRUG PURPOSES	23

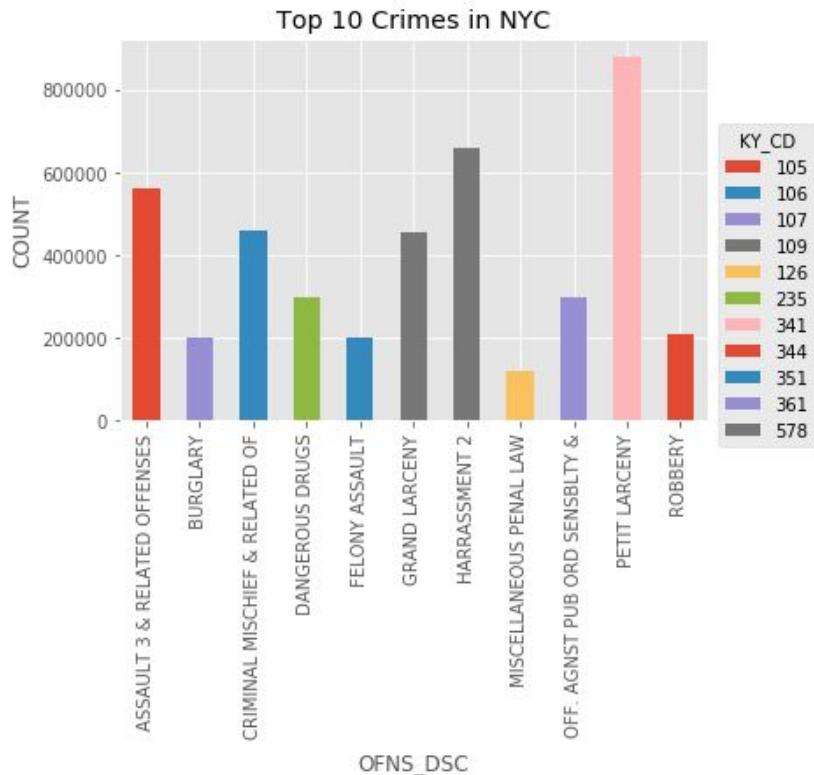
59	361	OFF. AGNST PUB ORD SENSBLTY &	300898
60	362	OFFENSES AGAINST MARRIAGE UNCL	3
61	363	OFFENSES AGAINST PUBLIC SAFETY	833
62	364	OTHER STATE LAWS NON PENAL LA	5562
63	364	AGRICULTURE & MRKTS LAW-UNCLASSIFIED	222
64	364	OTHER STATE LAWS NON PENAL LAW	8
65	365	ADMINISTRATIVE CODE	10739
66	366	NEW YORK CITY HEALTH CODE	45
67	455	UNLAWFUL POSS. WEAP. ON SCHOOL	62
68	460	LOITERING/DEVIATE SEX	16
69	571	LOITERING/GAMBLING CARDS, DIC	210
70	572	DISORDERLY CONDUCT	860
71	577	UNDER THE INFLUENCE OF DRUGS	3
72	578	HARRASSMENT 2	660027
73	672	LOITERING	14
74	675	ADMINISTRATIVE CODE	1409
75	676	NEW YORK CITY HEALTH CODE	33
76	677	NYS LAWS-UNCLASSIFIED VIOLATION	20
77	677	OTHER STATE LAWS	242
78	678	MISCELLANEOUS PENAL LAW	8671
79	685	ADMINISTRATIVE CODES	19
80	881	OTHER TRAFFIC INFRACTION	10

CRIMES IN NYC



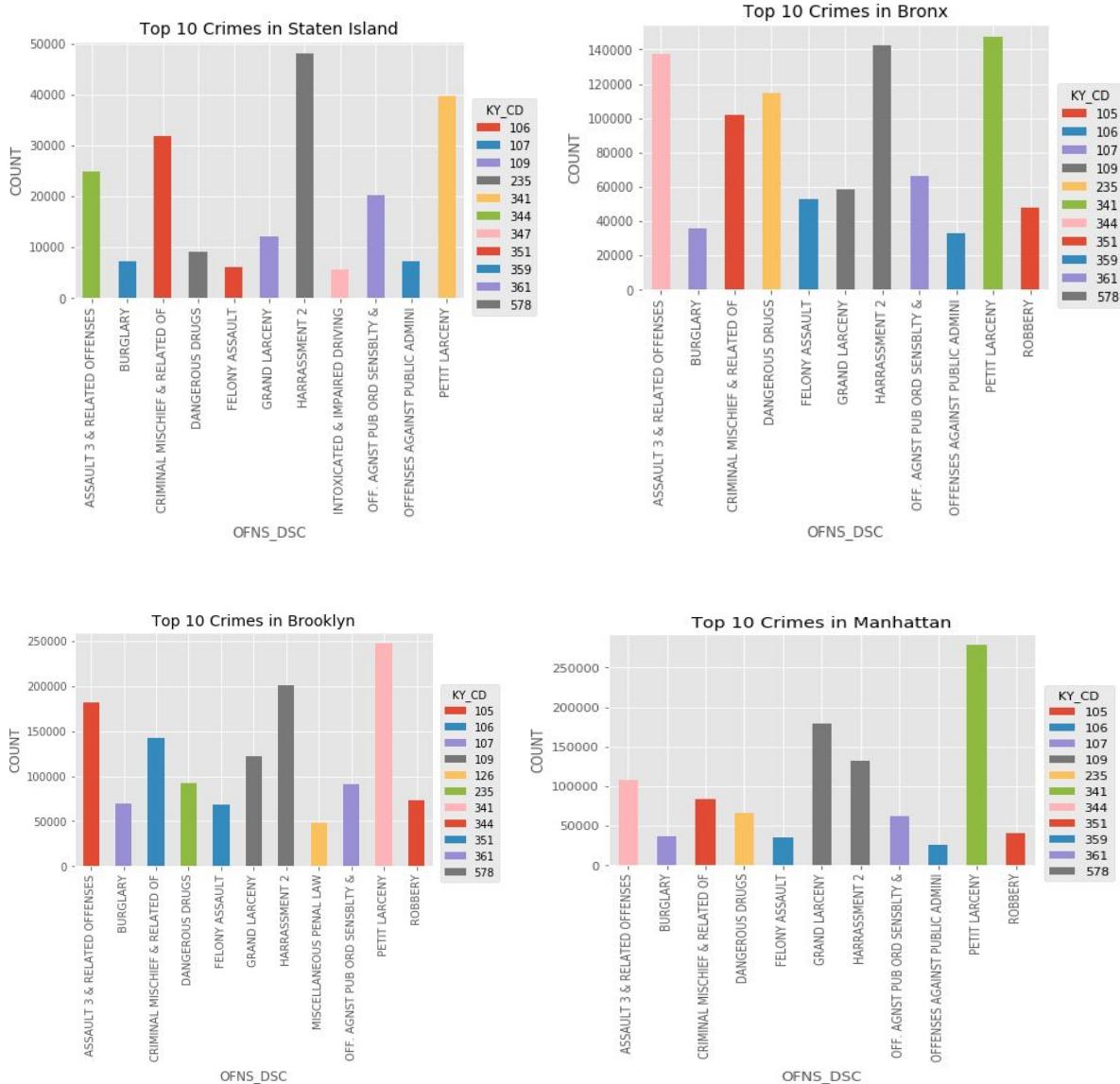
- From Data Cleaning Phase we already knew this was the count per borough.
- We wanted to in detail what were the crimes were and which was the most dangerous of the boroughs

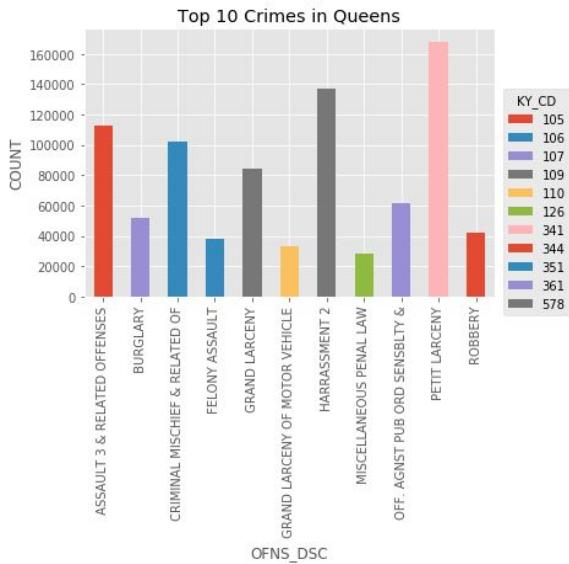
TOP 10 CRIMES IN NYC



- The highest crime is petit larceny which is basically stealing someone's personal property. This is an expected crime.
- Second is the harassment on a second degree which is also expected. Part and parcel of a new yorker's life.
- Same goes for Assault on 3rd degree
- Burglary, Robbery, Grand Larceny are all lower than the top 3, which was surprising
- Felony Assault is not the highest crime which is in contrary to all the NYPD crime dramas that are shown on the Big Screen and TV shows.

Borough wise



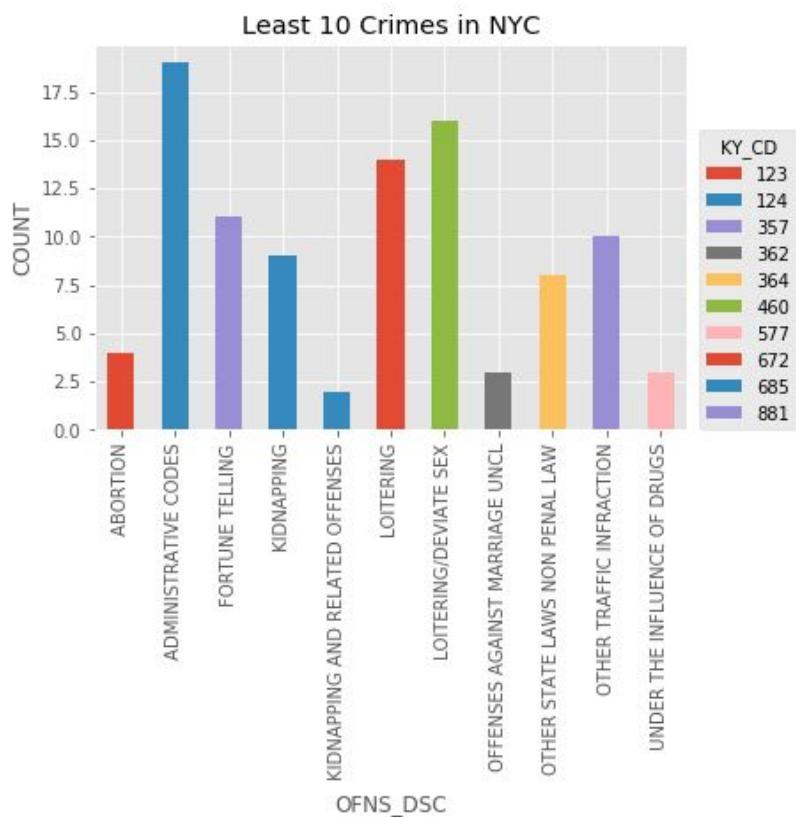


ANALYSIS BOROUGH WISE

- Staten Island
 - Harassment on second degree is the top crime.
 - Petit Larceny, Assault on 3rd degree follows
 - What is different from the overall crimes is the presence of crimes related to intoxicated driving, public sensibility and against public administration
 - Robbery is not in the top 10
- Bronx
 - Petit Larceny is the top crime.
 - Harassment on 2nd degree, Assault on 3rd degree follows
 - What is different from the overall crimes is the presence of crimes related to public sensibility and against public administration
- Brooklyn
 - Petit Larceny is the top crime.
 - Harassment on 2nd degree, Assault on 3rd degree follows
 - What is different from the overall crimes is the presence of crimes related to public sensibility
- Manhattan
 - Petit Larceny is the top crime.
 - Grand Larceny is the second highest crime which is different from the other boroughs
 - Harassment on 2nd degree, Assault on 3rd degree follows
 - What is different from the overall crimes is the presence of crimes related to public sensibility and against public administration

- Queens
 - Petit Larceny is the top crime.
 - Harassment on 2nd degree, Assault on 3rd degree follows
 - What is different from the overall crimes is the presence of crimes related to public sensibility
- Of all the boroughs, Queens, Bronx and Brooklyn seemed more similar to the overall data
- Manhattan is quite different maybe because of the higher cost of living. More than physical harm, the crimes were driven monetarily.

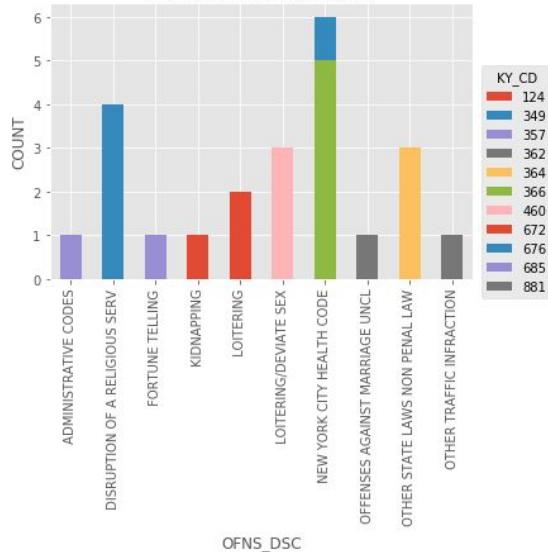
Bottom 10 Crimes in NYC



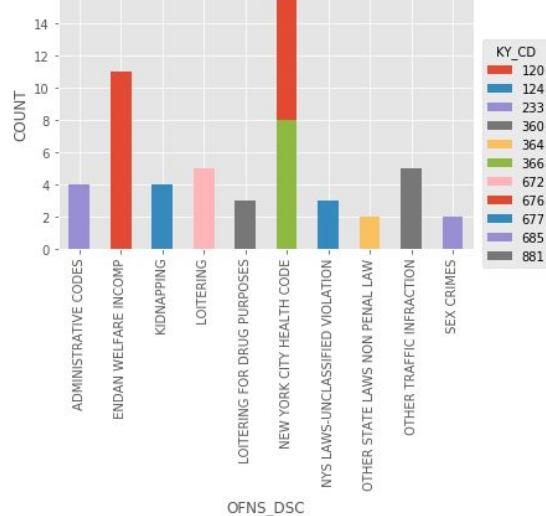
- Kidnapping, abortion and marriage related crimes are on a low in nyc. So it is a safe city for children looking at just that aspect.
- Surprisingly even crime under the influence is low.
- The global image of the NYC thanks to all crime dramas in reel world is that city is has all sorts of crimes and people have this fear of living in this great city.

Borough wise

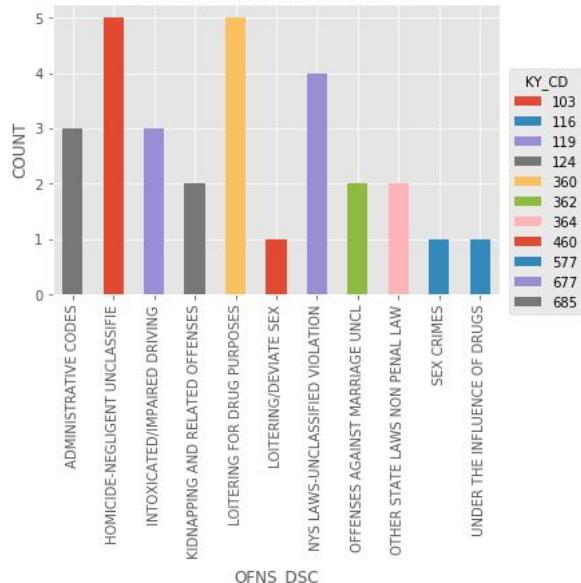
Least 10 Crimes in Bronx



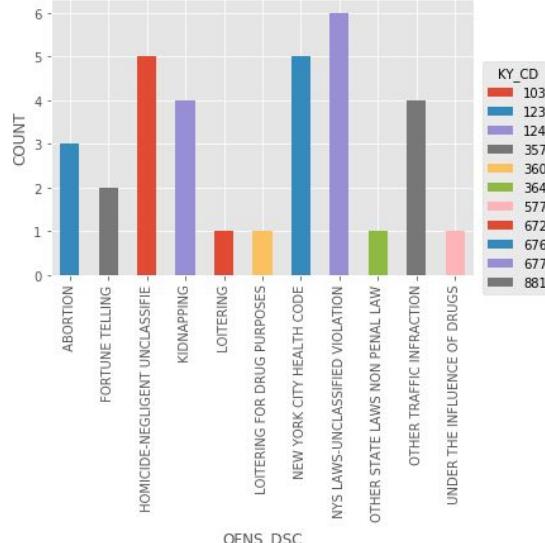
Least 10 Crimes in Brooklyn

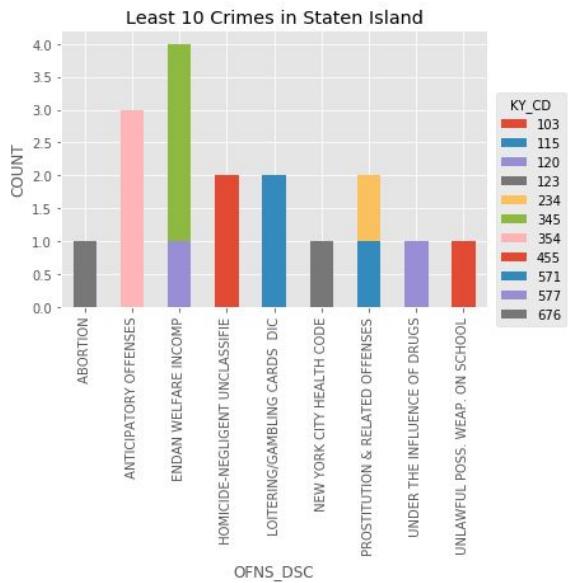


Least 10 Crimes in Manhattan



Least 10 Crimes in Queens

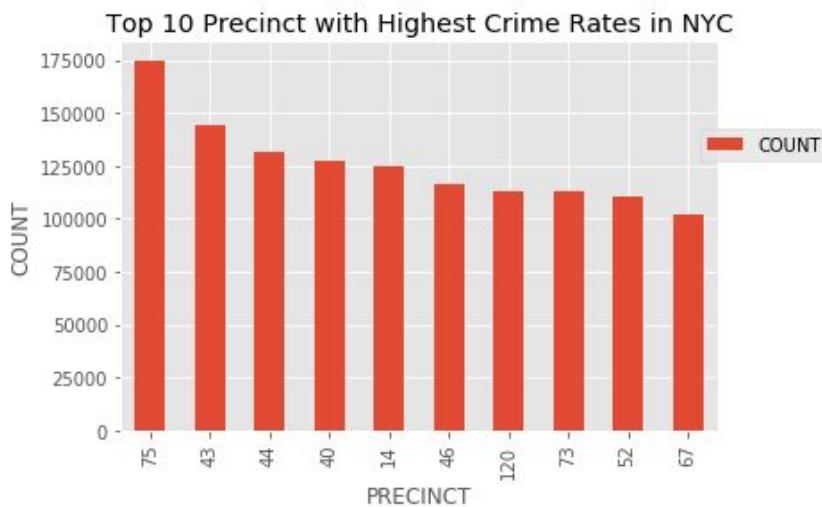




- Bronx
 - Health related crimes, loitering, kidnapping are low crimes
- Brooklyn
 - Health related crimes, Sex crimes, loitering, kidnapping are low crimes
- Manhattan
 - Negligent Homicides are a low crimes
 - Loitering, Sex Crimes, Intoxicated drunk driving are also low
- Queens
 - Drug related crimes are low
 - Abortion and Kidnapping is also low
- Staten Island
 - Health related crimes, loitering, kidnapping, prostitution related are low crimes

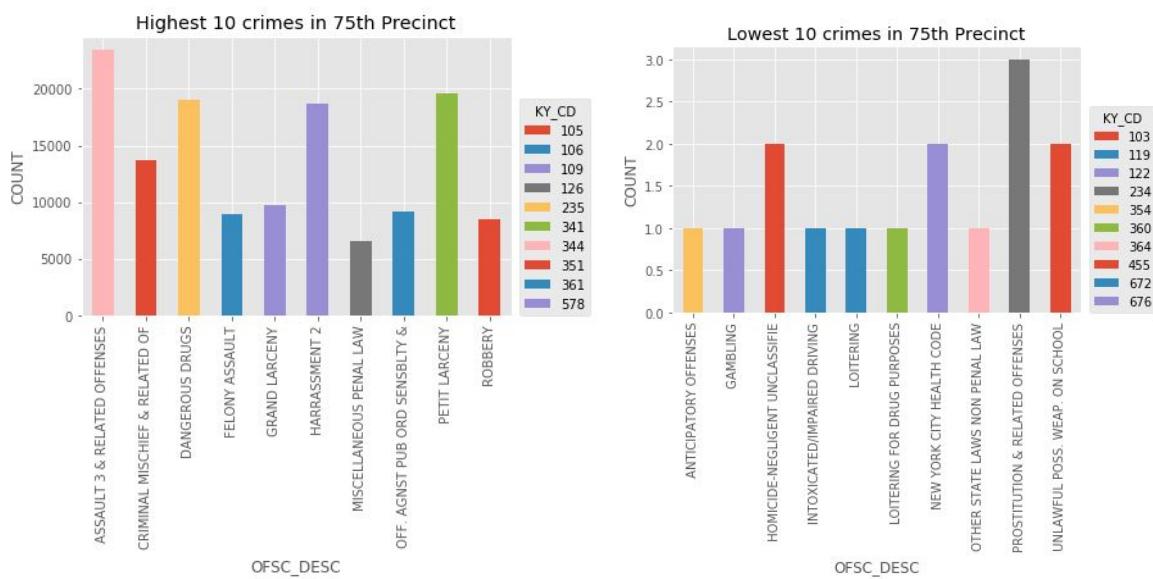
PRECINCT

Top 10 Precinct with Highest Crimes Rates



- 75th Precinct at Sutter Avenue, Brooklyn has the highest amount of crime

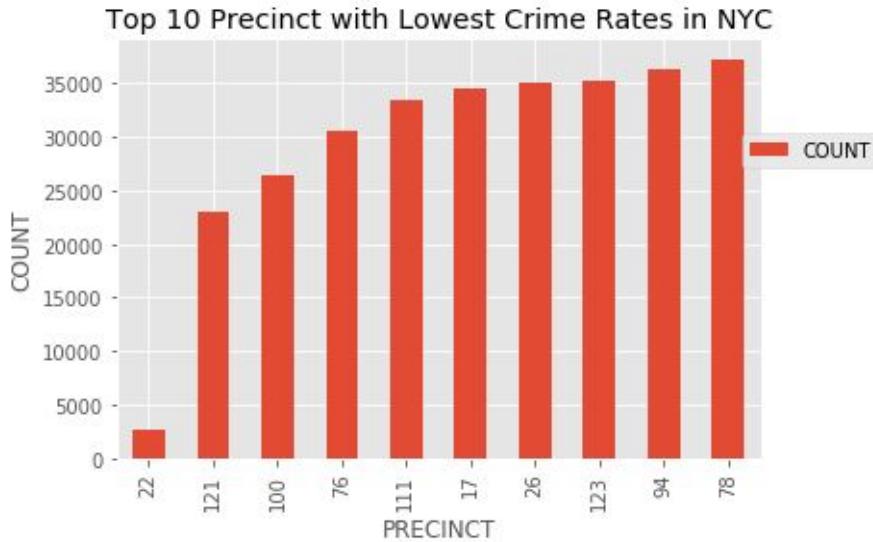
75th Precinct



- Highest crime is Assault on a third degree, followed by Petit Larceny and Harassment on a second degree.
- The crimes have roughly the same part of crimes with Brooklyn as a whole.

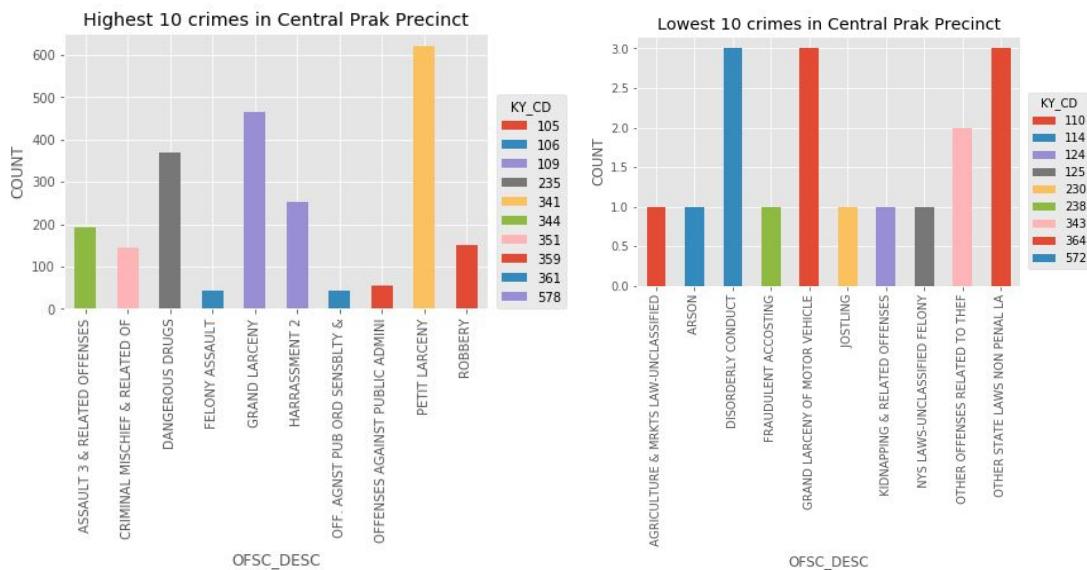
- Loitering, Gambling, Prostitution are among the lowest crimes

Top 10 Precinct with Lowest Crimes Rates



- Central Park Precinct(22) has the least amount of crime

CENTRAL PARK PRECINCT

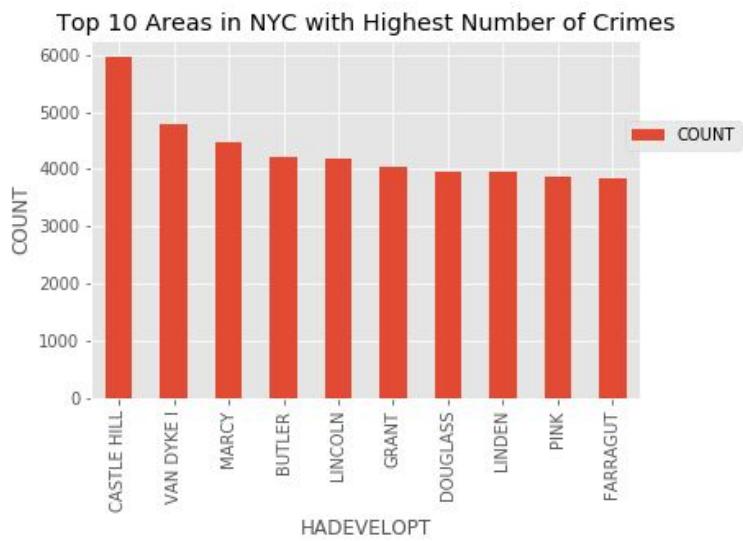


- Like trends of manhattan, petit larceny was the top crime, followed by grand larceny
- The next highest crime was dangerous drugs. Central Park area being public spaces, people would assume using drugs there might not be noticeable.
- Among the low crimes were arson, jostling

HOUSING DEVELOPMENT

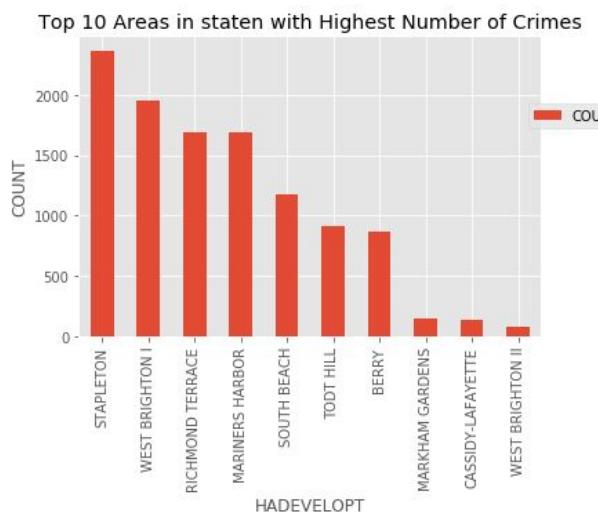
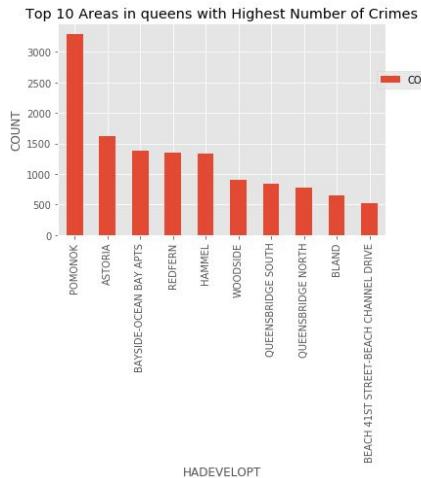
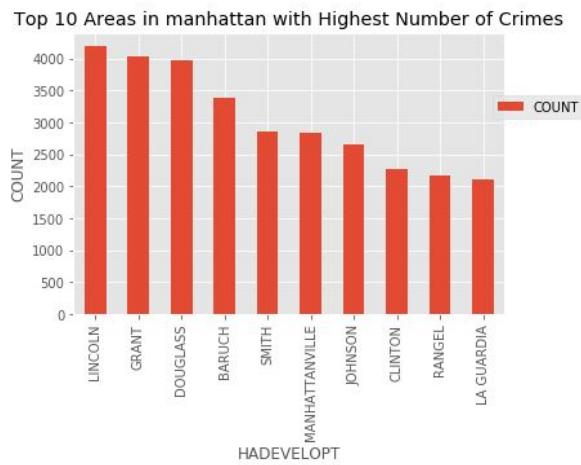
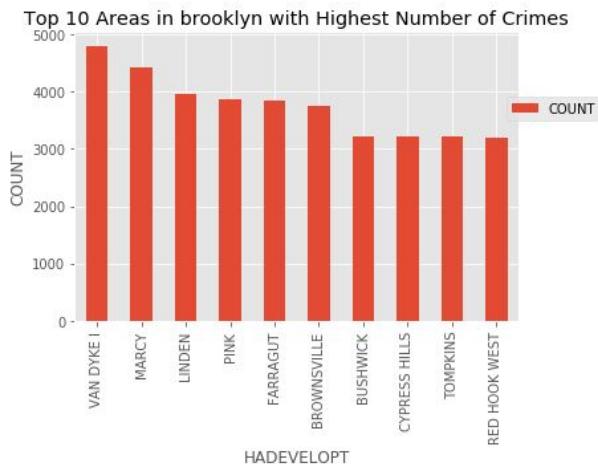
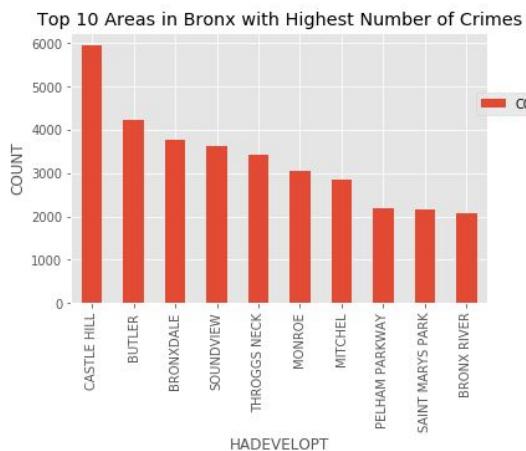
- We used the HAVDEVLOPT field for this graph. This field is blank for a lot of records and the addresses are not exactly neighbourhoods.
- However, it contained interesting information for the rows where it was given.

MOST UNSAFE PLACES TO LIVE IN NYC

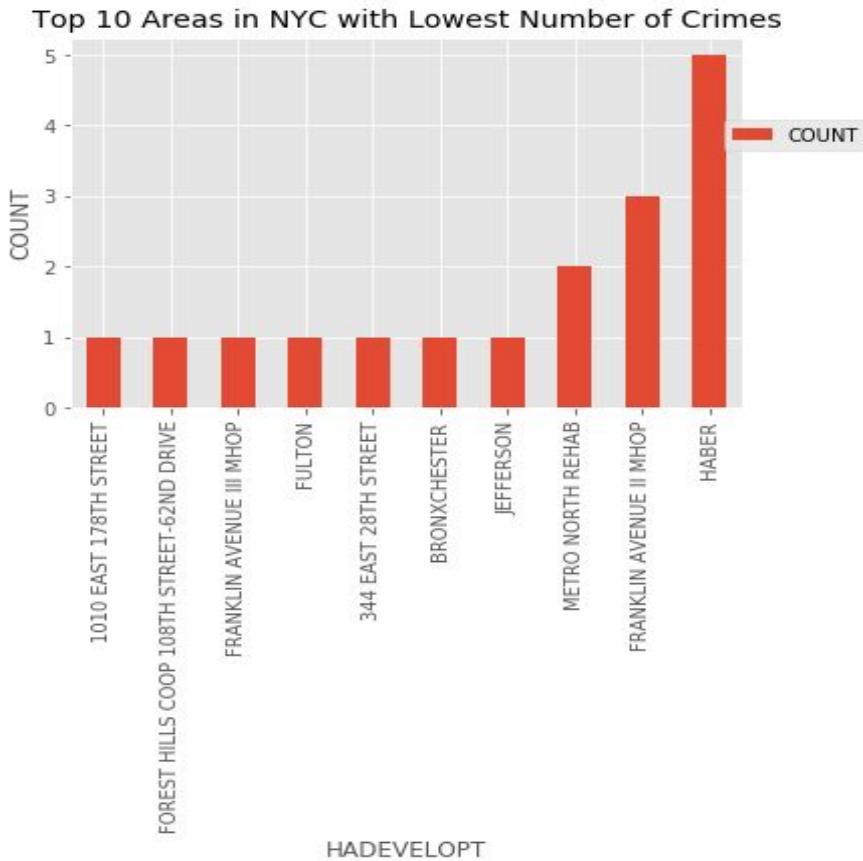


- Most unsafe place to live in NYC is Castle Hill, Bronx
- Followed by Van Dyke and Marcy, Brooklyn
- Butler is in Bronx
- Lincoln, Grant, Douglas to all irony is in Manhattan
- Linden, Pink and Farggut are in Brooklyn

Borough Wise



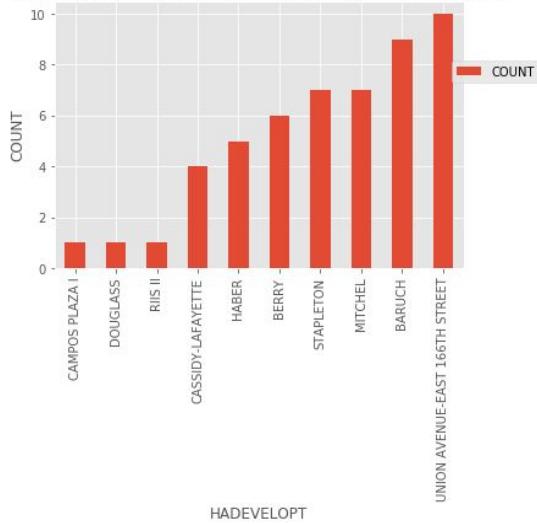
SAFEST PLACES TO LIVE



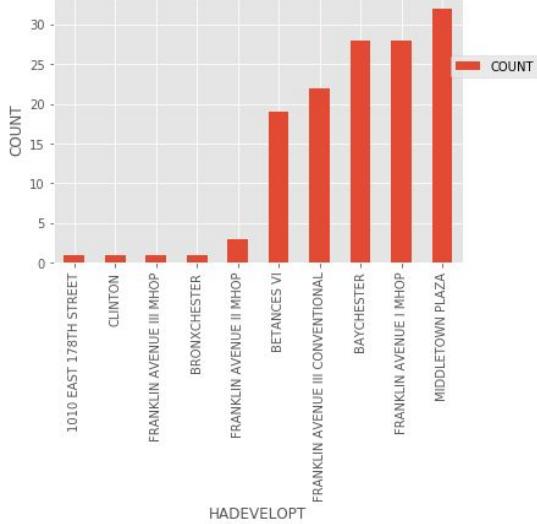
- As you can see some of the locations are not exactly neighbourhoods but addresses
- Among these, 1010 East 178th Street, Forest Hills Coop, Franklin Avenue, Fulton, East 28th Street, Bronxchester, Jefferson reported the least amount of crime
- However, we must remember that areas that never appeared on this list might be even more safer

Borough Wise

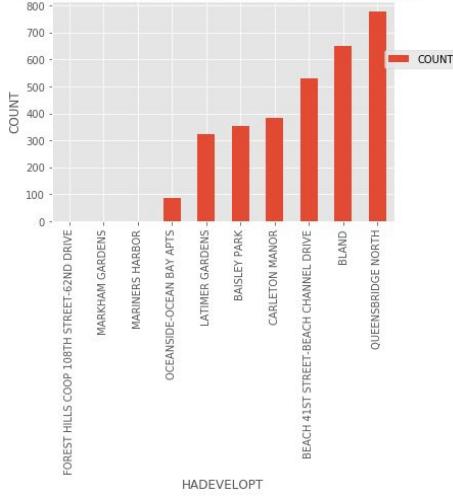
Top 10 Areas in brooklyn with Lowest Number of Crimes



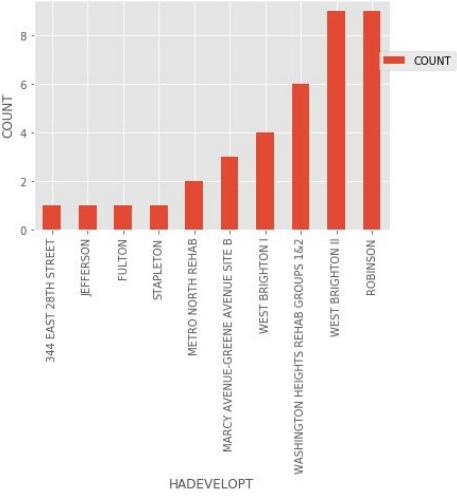
Top 10 Areas in Bronx with Lowest Number of Crimes



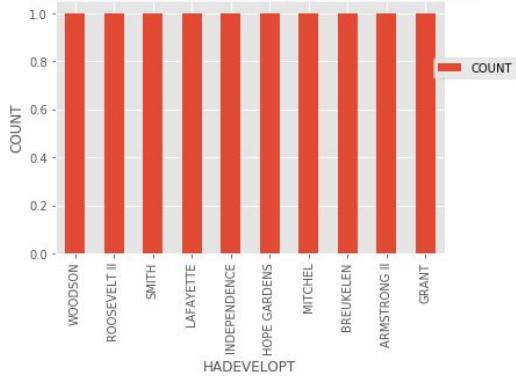
Top 10 Areas in queens with Lowest Number of Crimes



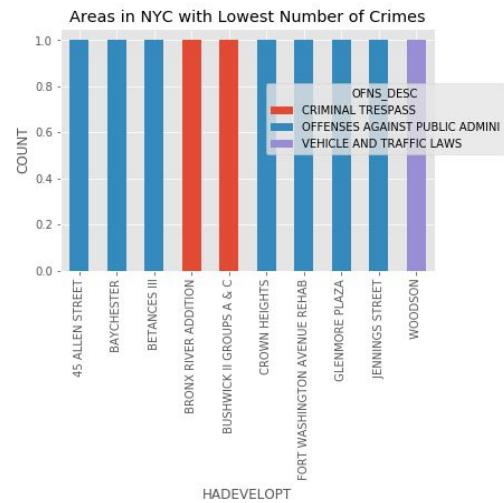
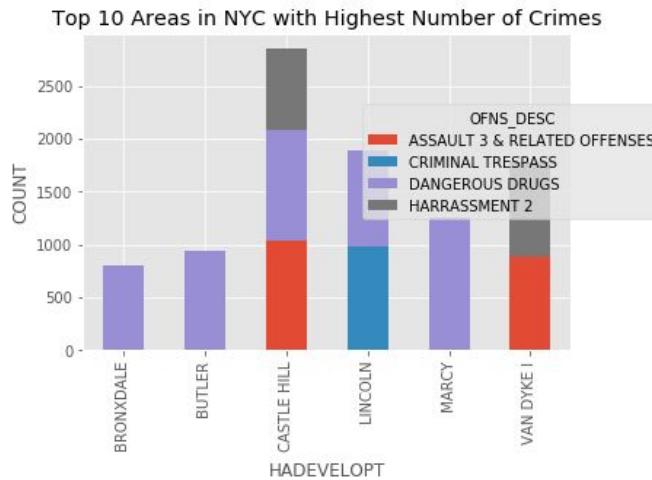
Top 10 Areas in manhattan with Lowest Number of Crimes



Top 10 Areas in staten with Lowest Number of Crimes

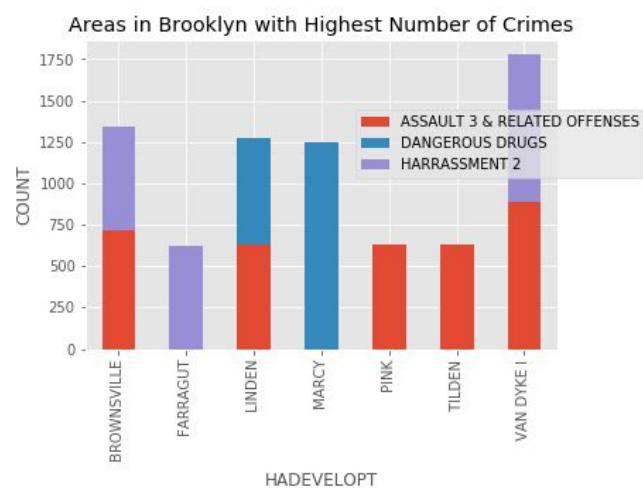
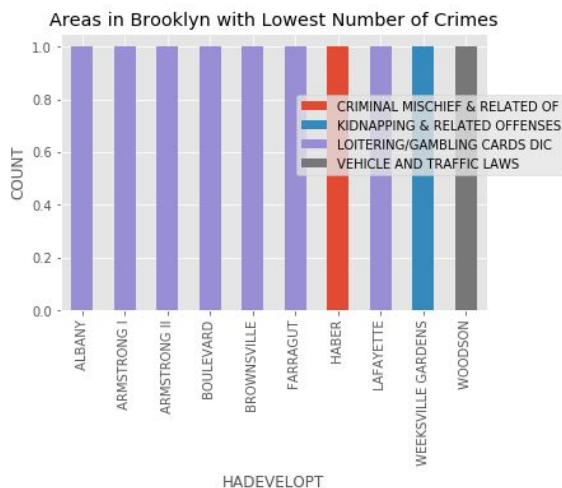
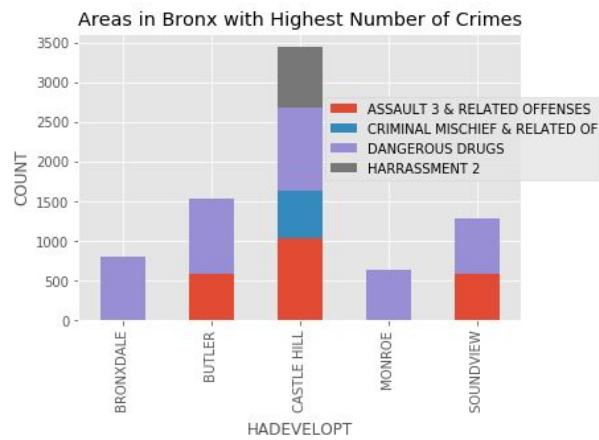
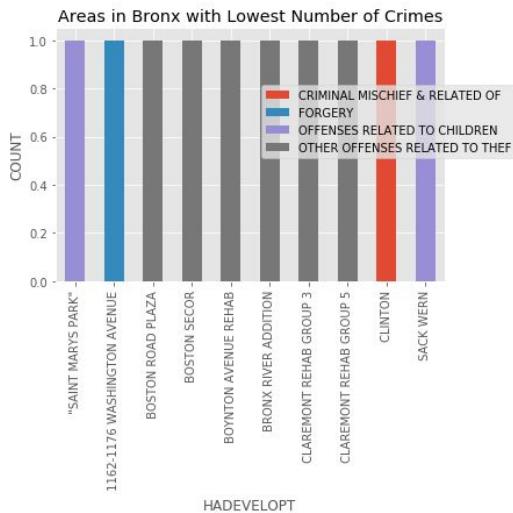


HAVDEVELOPMENT CRIME WISE

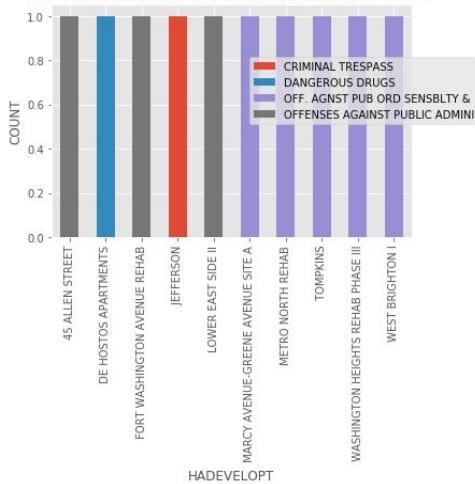


- The top crimes in Castle Hill are Assault on a 3rd degree, Harassment on a second degree and dangerous drugs
- In bronxdale,butler, marcy the top crime is dangerous drugs
- For Van dyke it is Assault on a 3rd degree and Harassment on a 2nd degree
- 45 Allen Street, Baychester etc have least crimes
- Ironically Crown Heights and Bushwick II are on the least crimes. They are considered 2 very unsafe places in brooklyn generally.

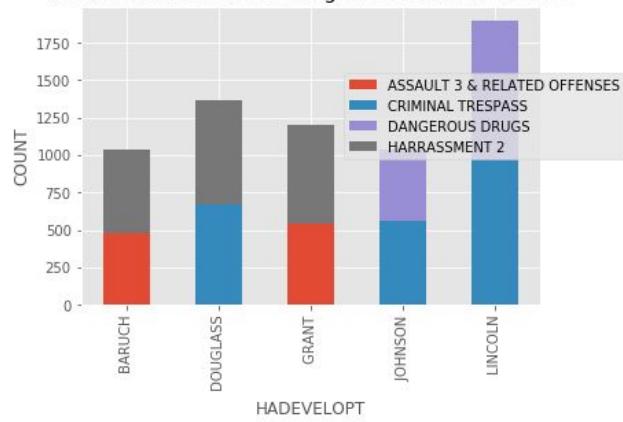
Borough Wise



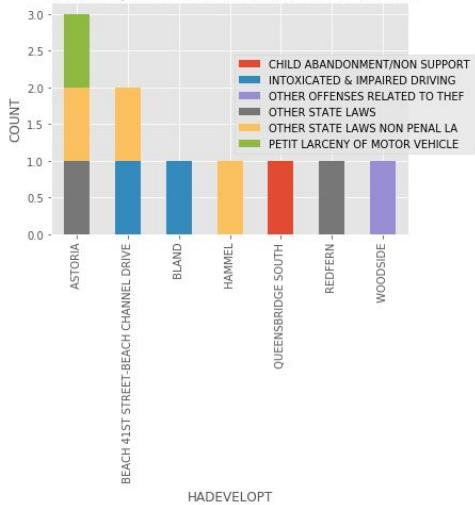
Areas in Manhattan with Lowest Number of Crimes



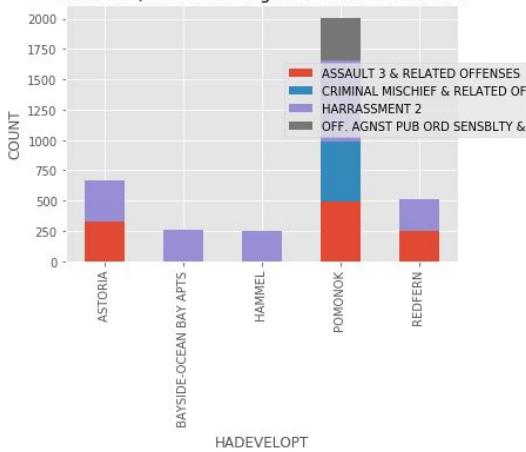
Areas in Manhattan with Highest Number of Crimes



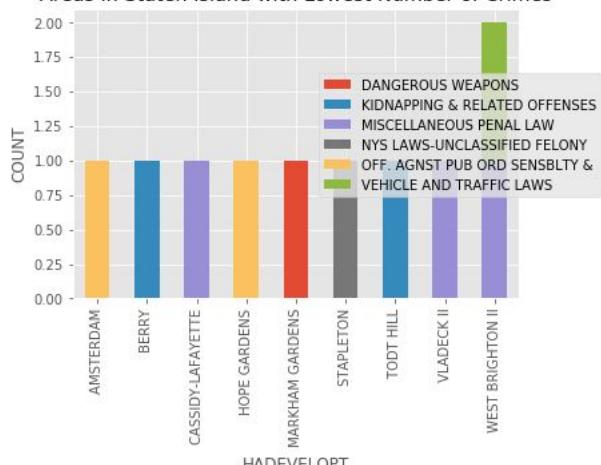
Areas in Queens with Lowest Number of Crimes



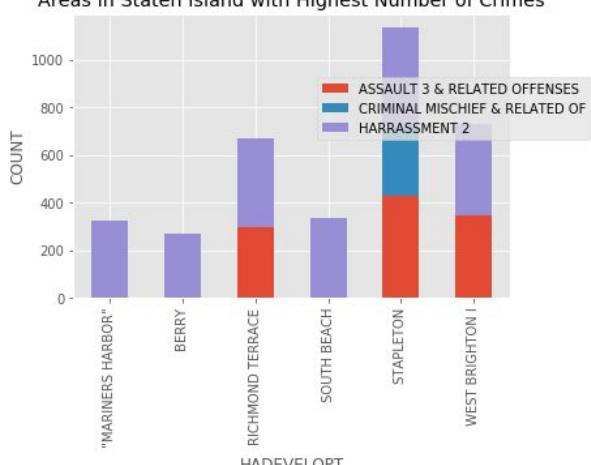
Areas in Queens with Highest Number of Crimes



Areas in Staten Island with Lowest Number of Crimes



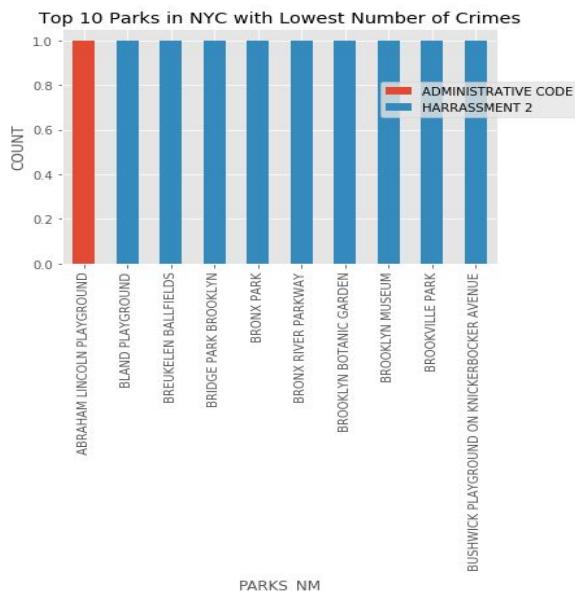
Areas in Staten Island with Highest Number of Crimes



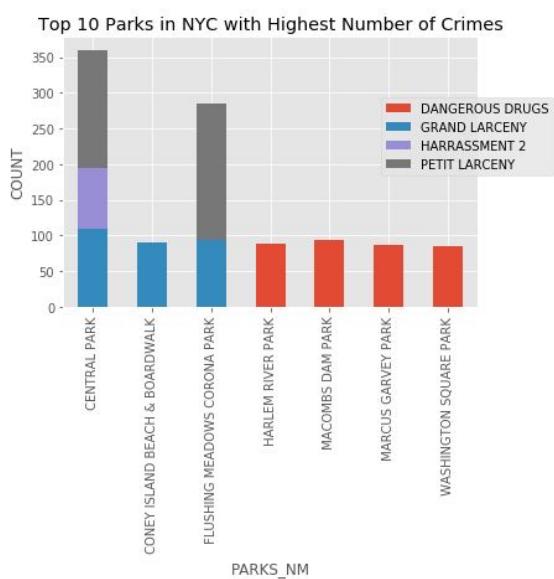
PARKS

- This field was null for majority as it is only required if a crime had occurred in a park

MOST SAFE PARKS IN NYC

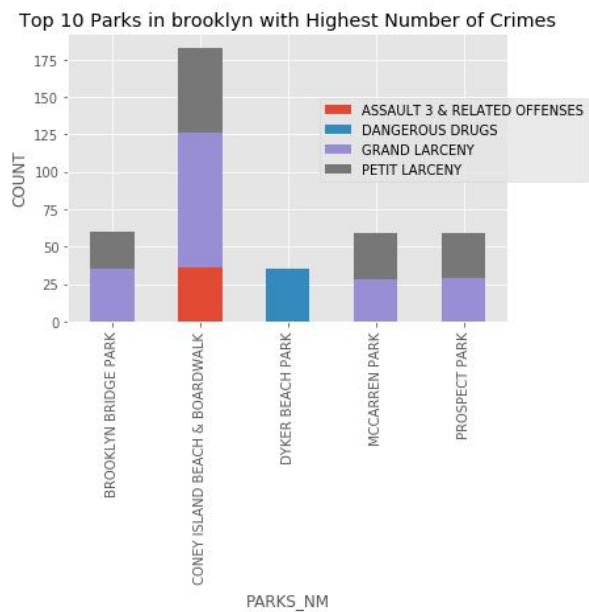
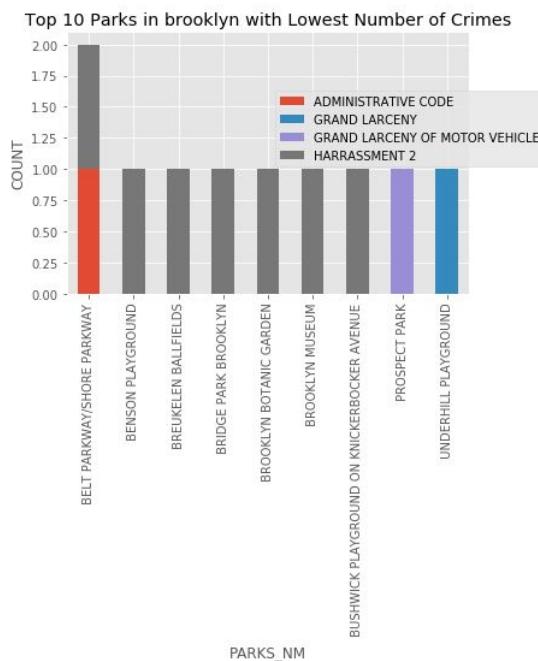
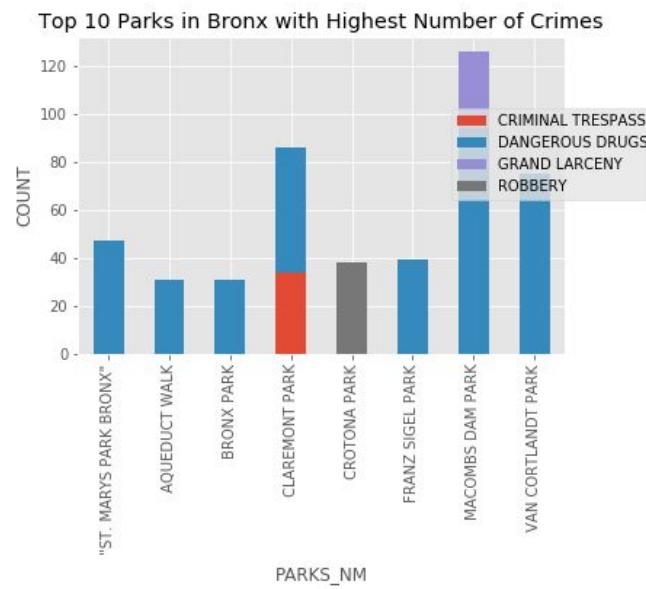
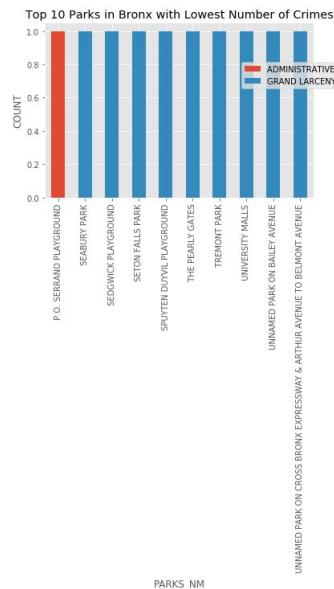


MOST UNSAFE PARKS IN NYC

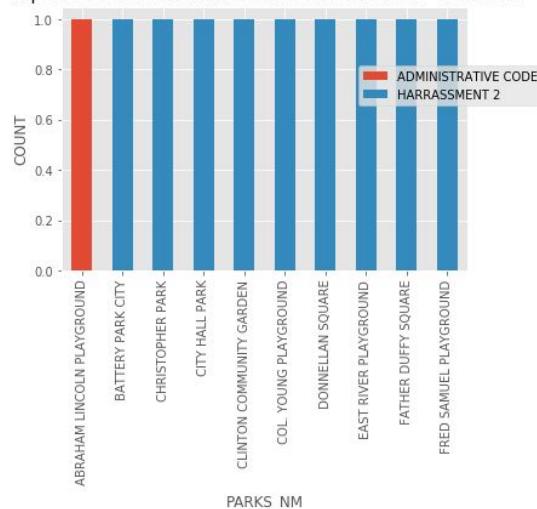


- The most unsafe park is undoubtedly the most famous one, Central Park
- It is followed by Corona Park and Coney Island Boardwalk
- Ironically our very own Washington Square Park is in the top 10

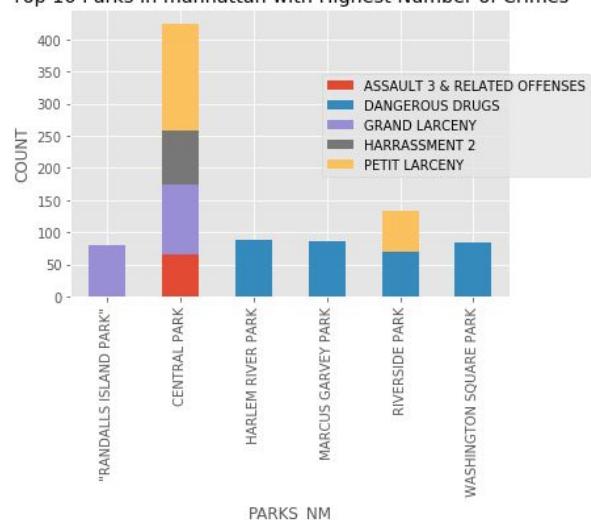
Borough Wise



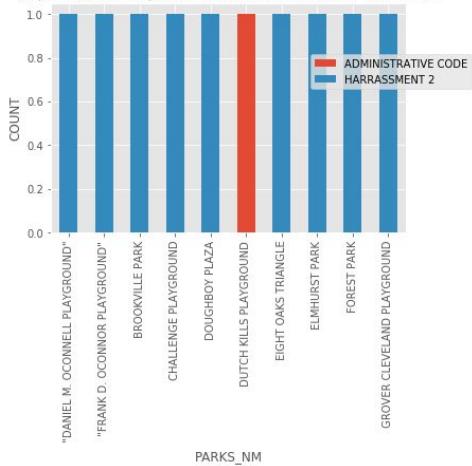
Top 10 Parks in manhattan with Lowest Number of Crimes



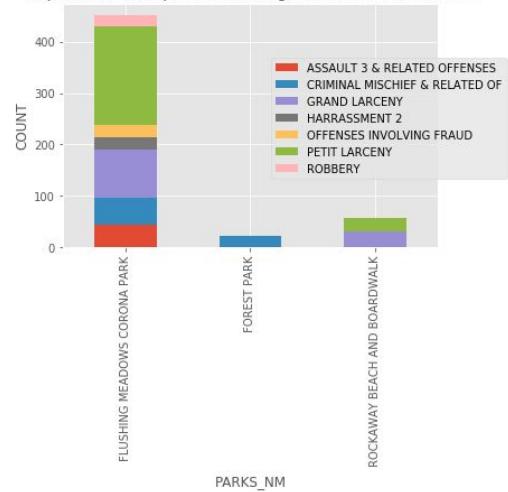
Top 10 Parks in manhattan with Highest Number of Crimes



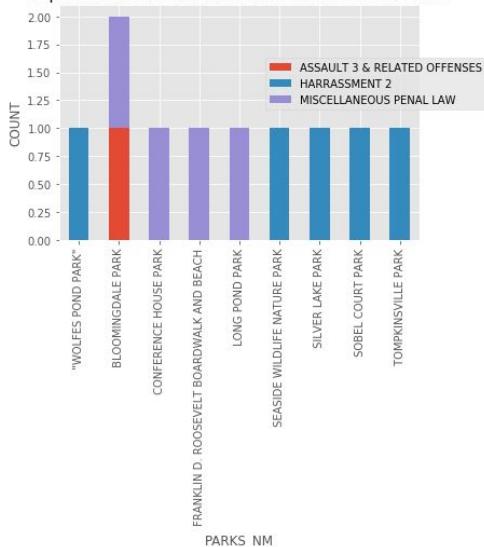
Top 10 Parks in queens with Lowest Number of Crimes



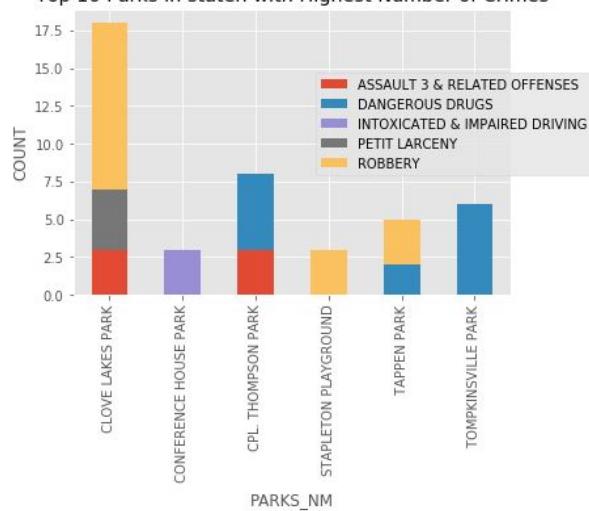
Top 10 Parks in queens with Highest Number of Crimes



Top 10 Parks in staten with Lowest Number of Crimes

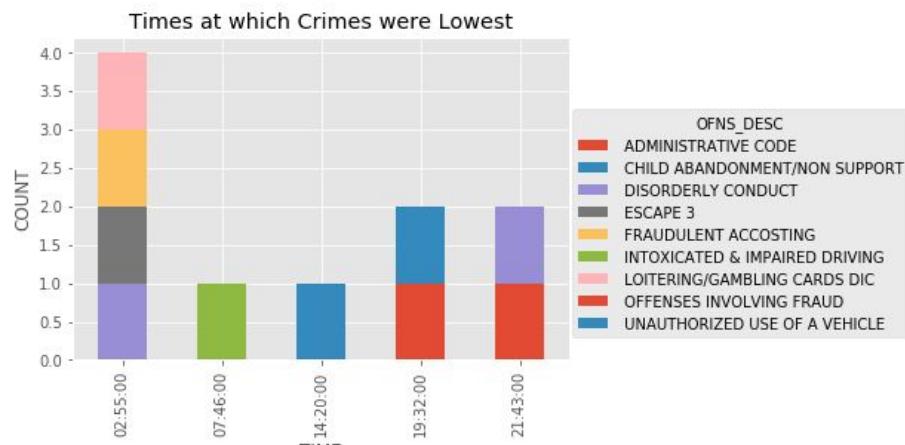
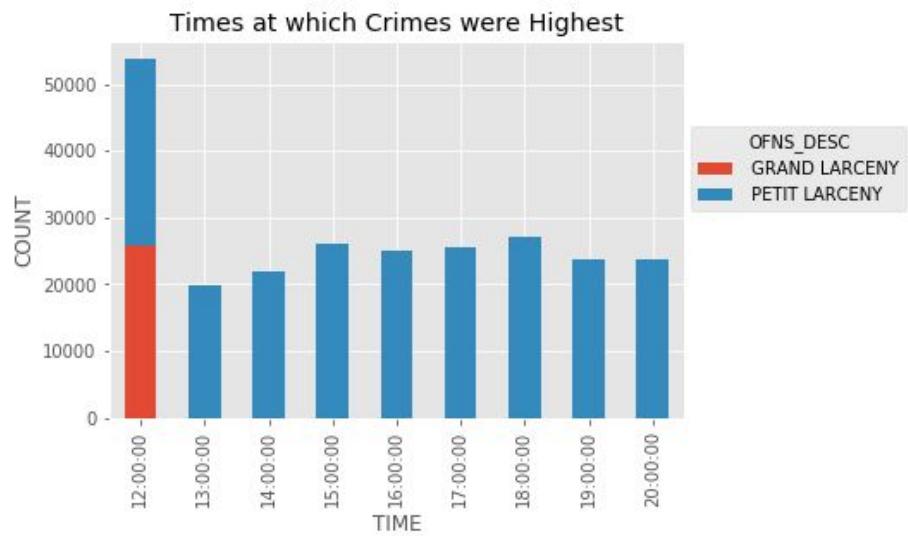


Top 10 Parks in staten with Highest Number of Crimes



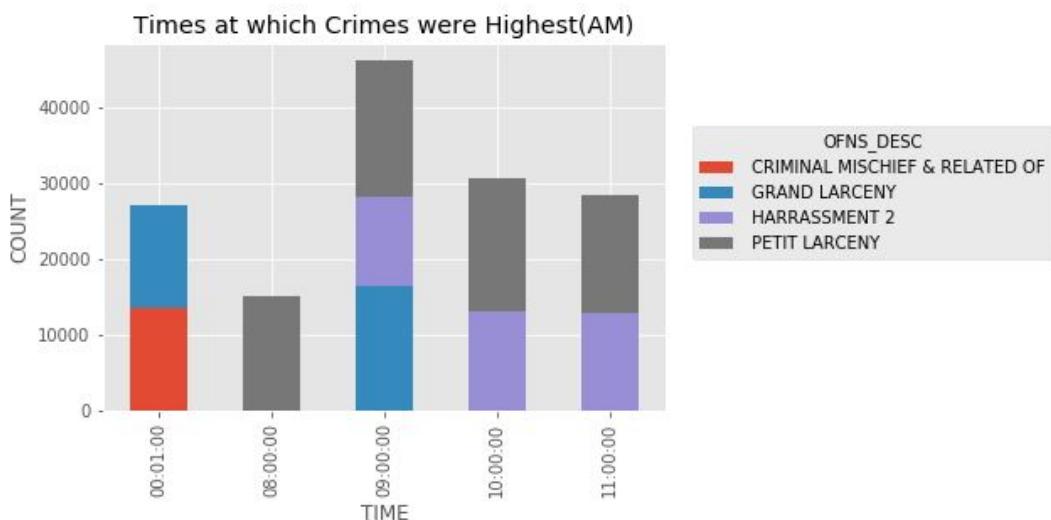
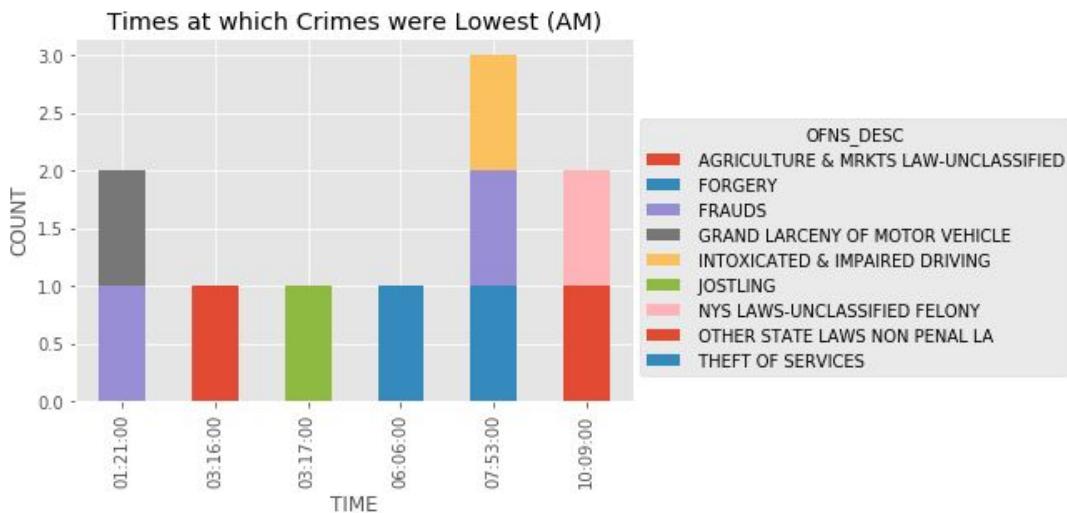
TIME

- Most crimes are generally at 12pm oddly and are mostly larceny. We could assume during day hour most of the people are out and chances of crime occurrence is higher than during mid night as most people stay at home.



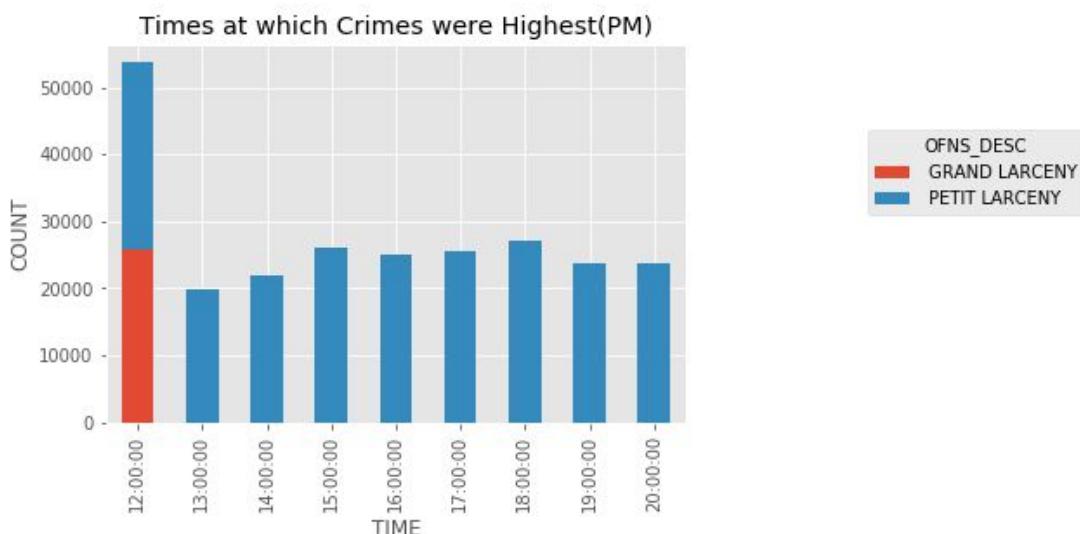
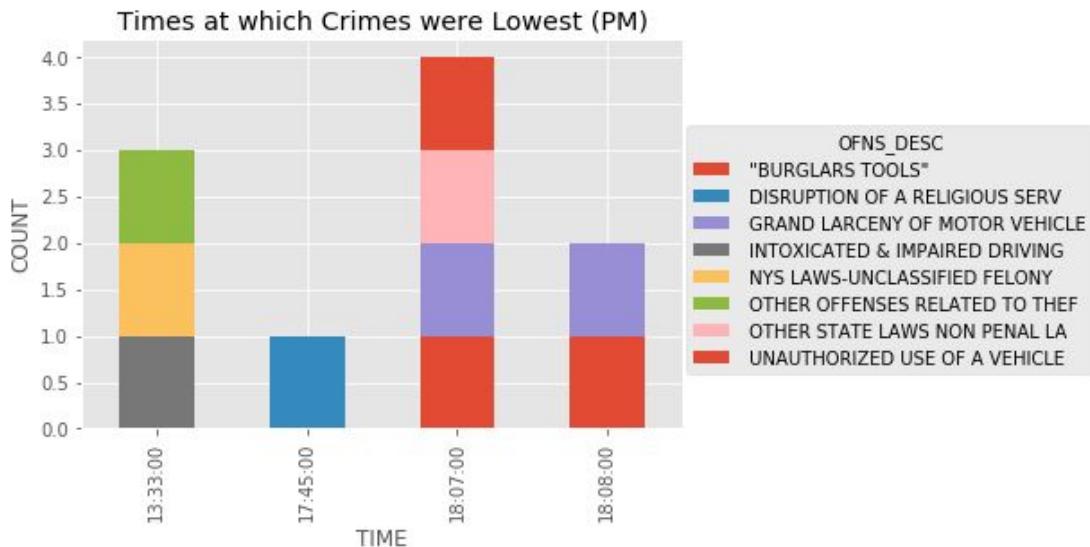
- It was lowest 7 am
- For better analysis we divided into AM and PM and did borough wise

TIME AM



- 9 am sees the most amount of crimes and mostly it is larceny
- 3am, 6am the crimes are relatively

TIME IN PM

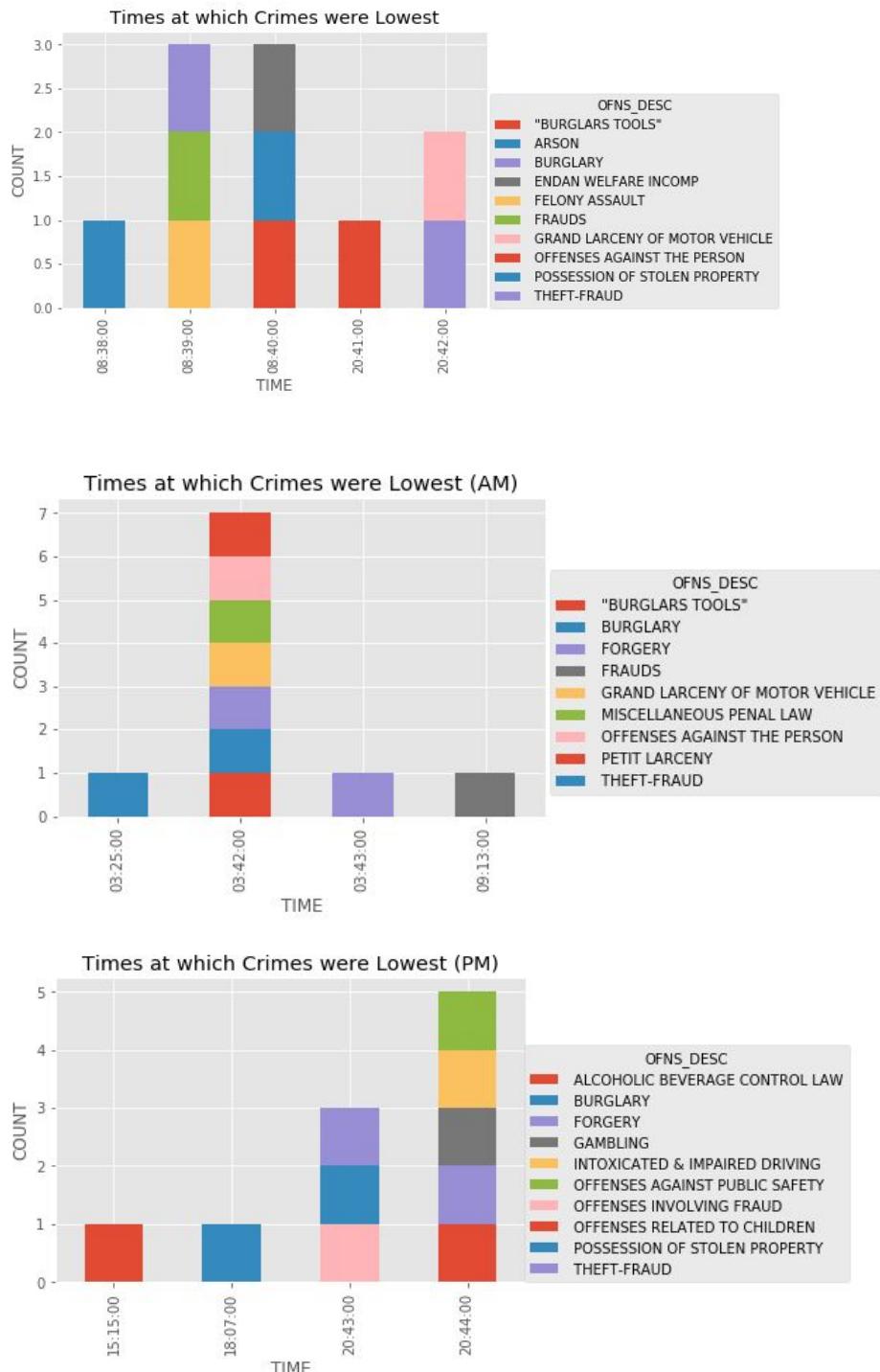


- Crime is highest at 12pm and is of the kind of larceny
- It is lowest at around 5pm

Borough Wise

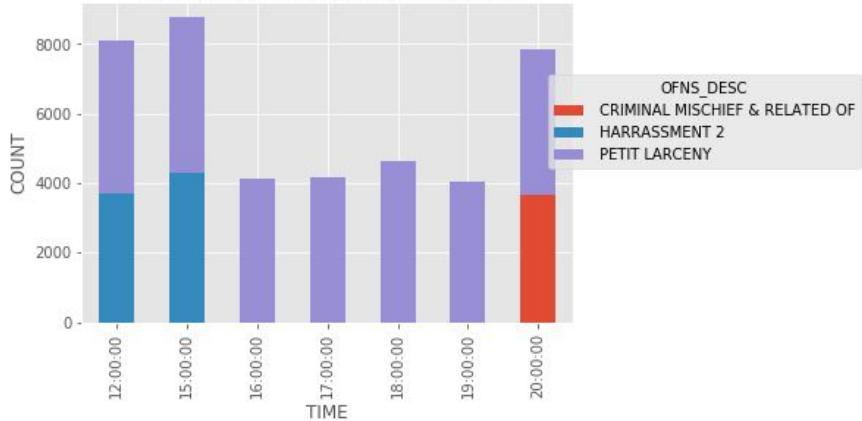
Bronx

Lowest

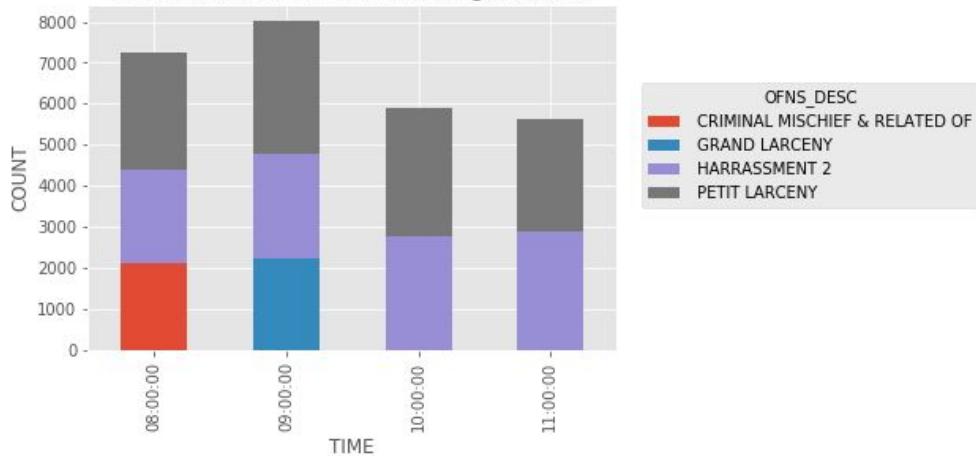


Highest

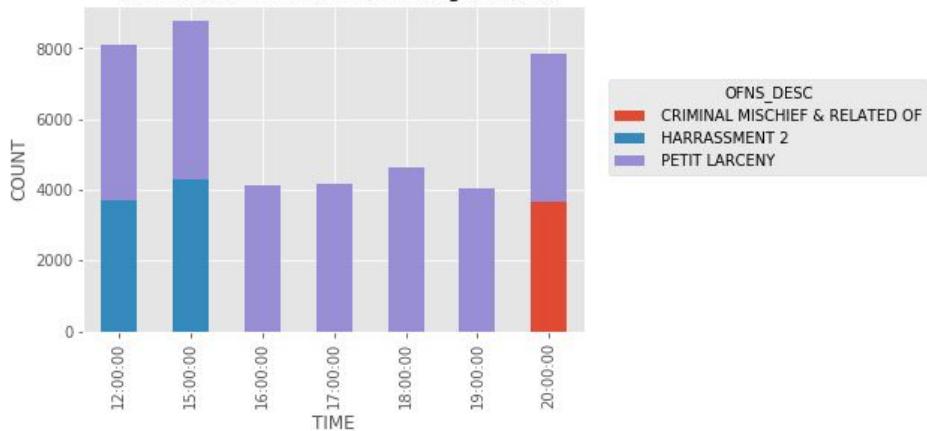
Times at which Crimes were Highest



Times at which Crimes were Highest(AM)

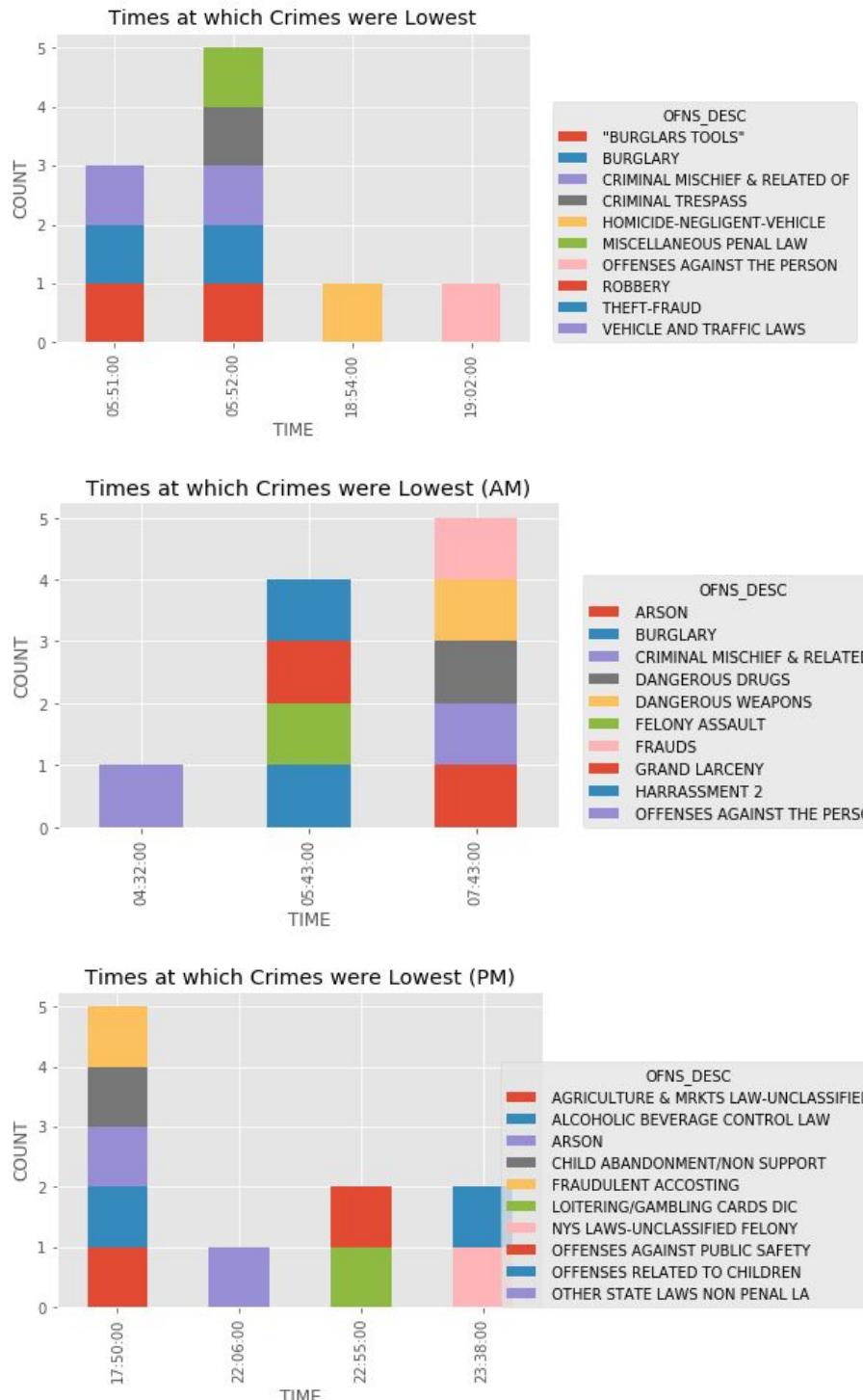


Times at which Crimes were Highest(PM)



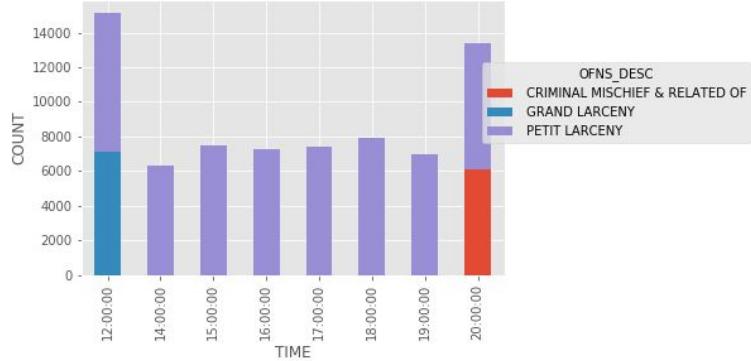
Brooklyn

Lowest

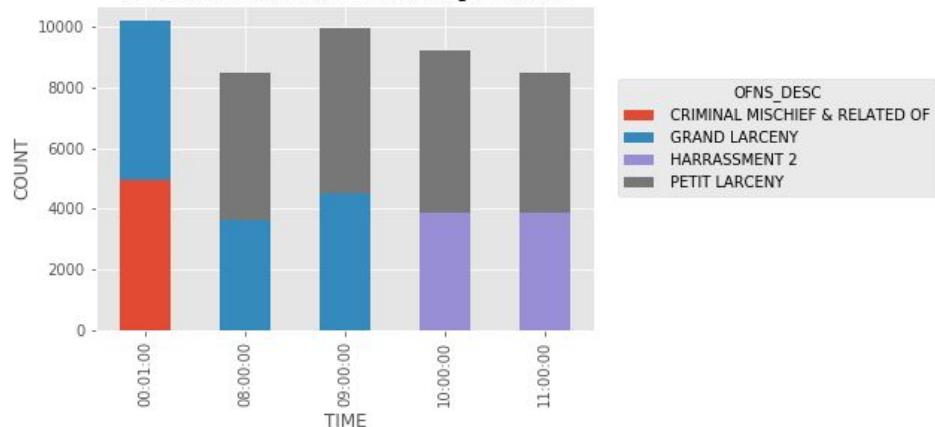


Highest

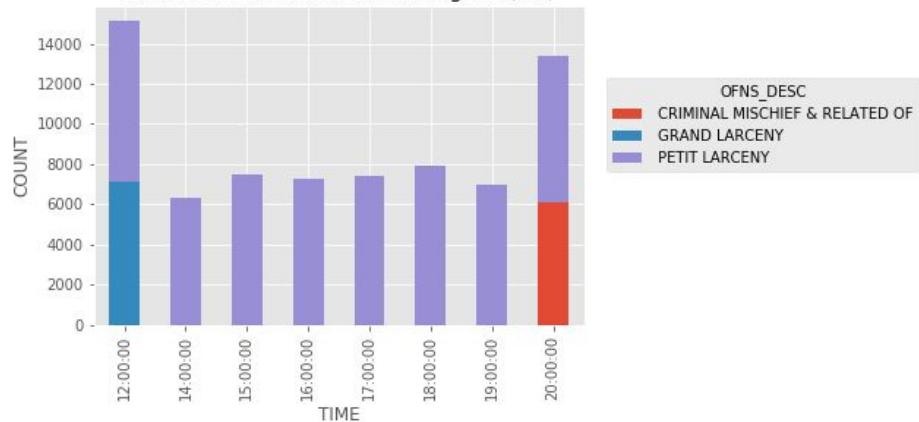
Times at which Crimes were Highest



Times at which Crimes were Highest(AM)

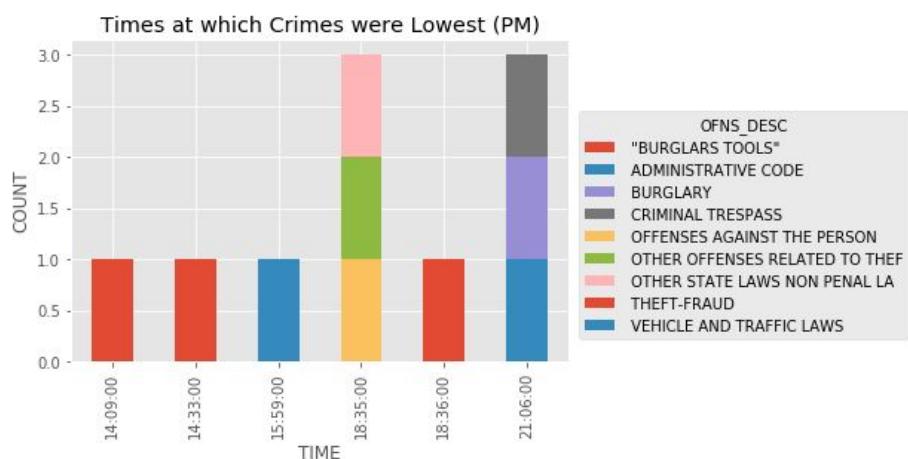
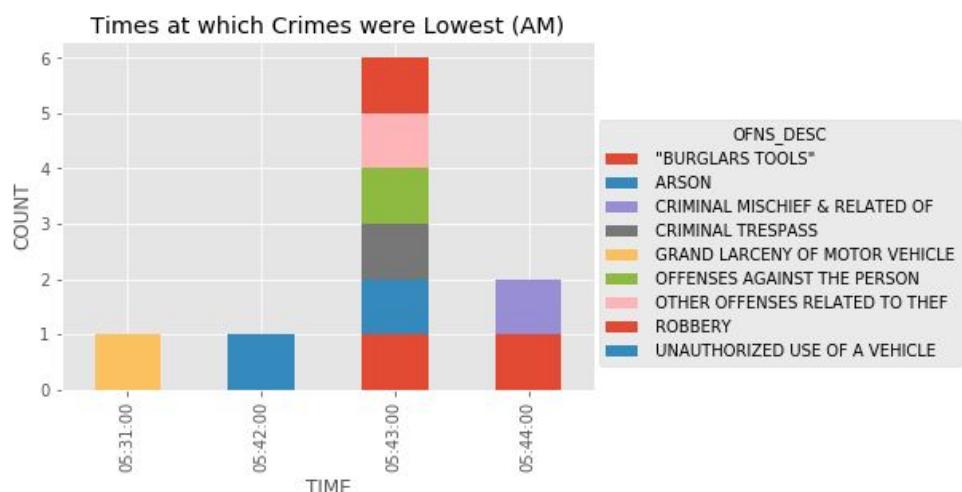
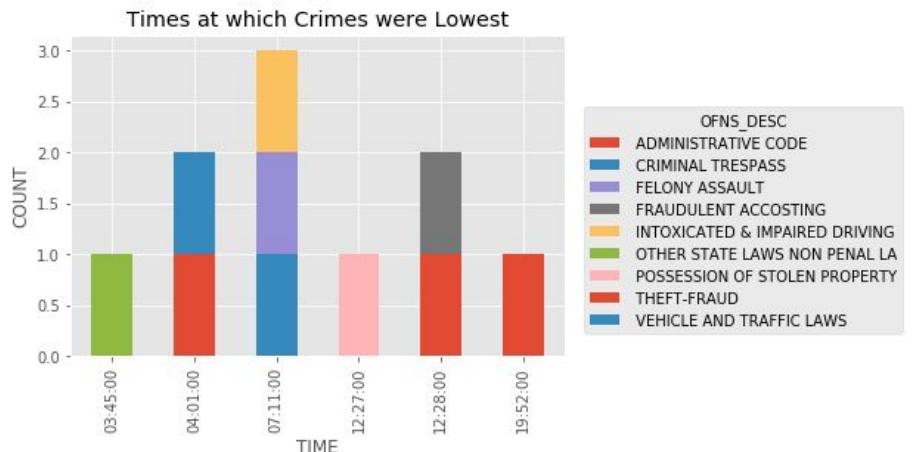


Times at which Crimes were Highest(PM)

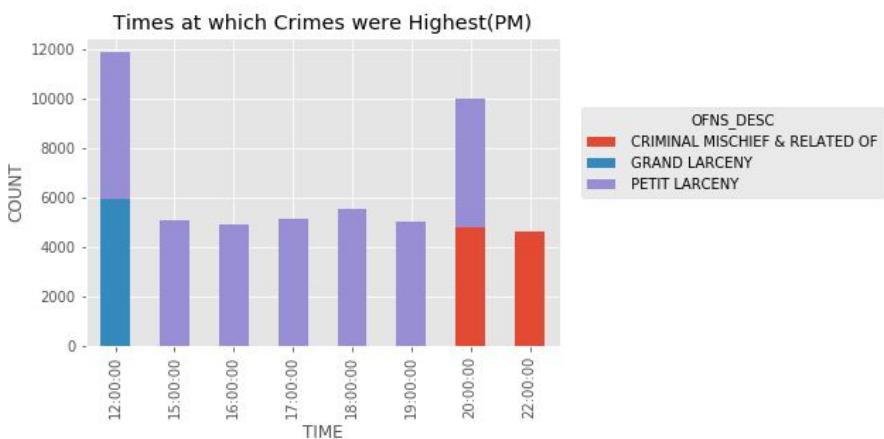
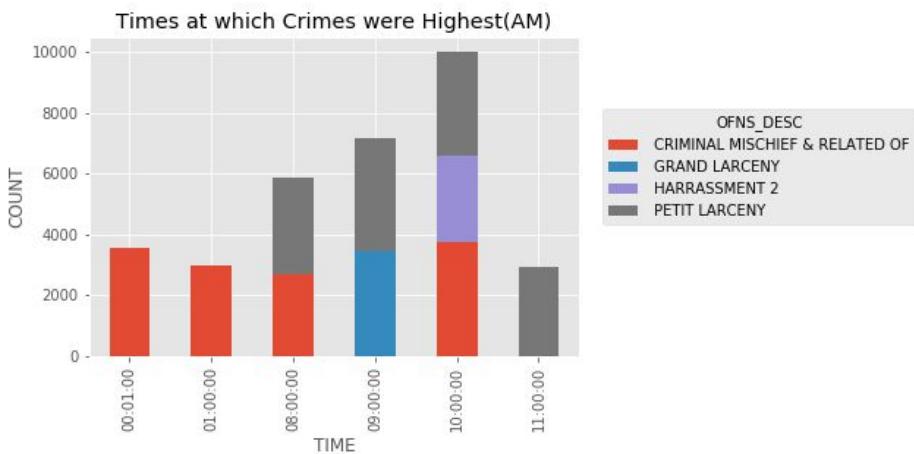
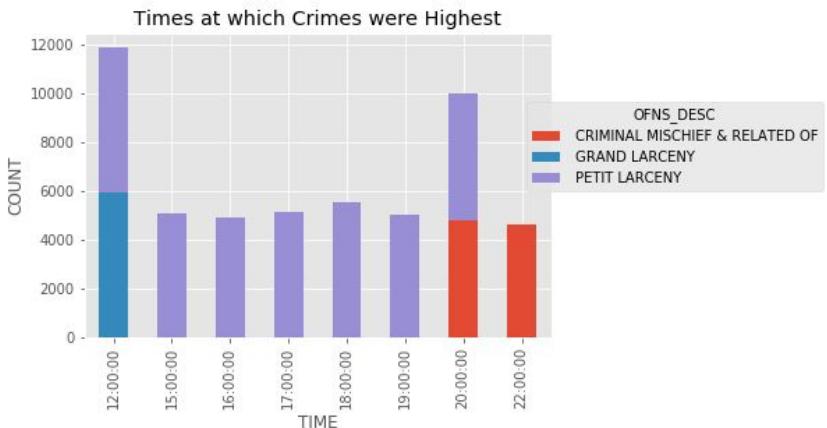


Queens

Lowest

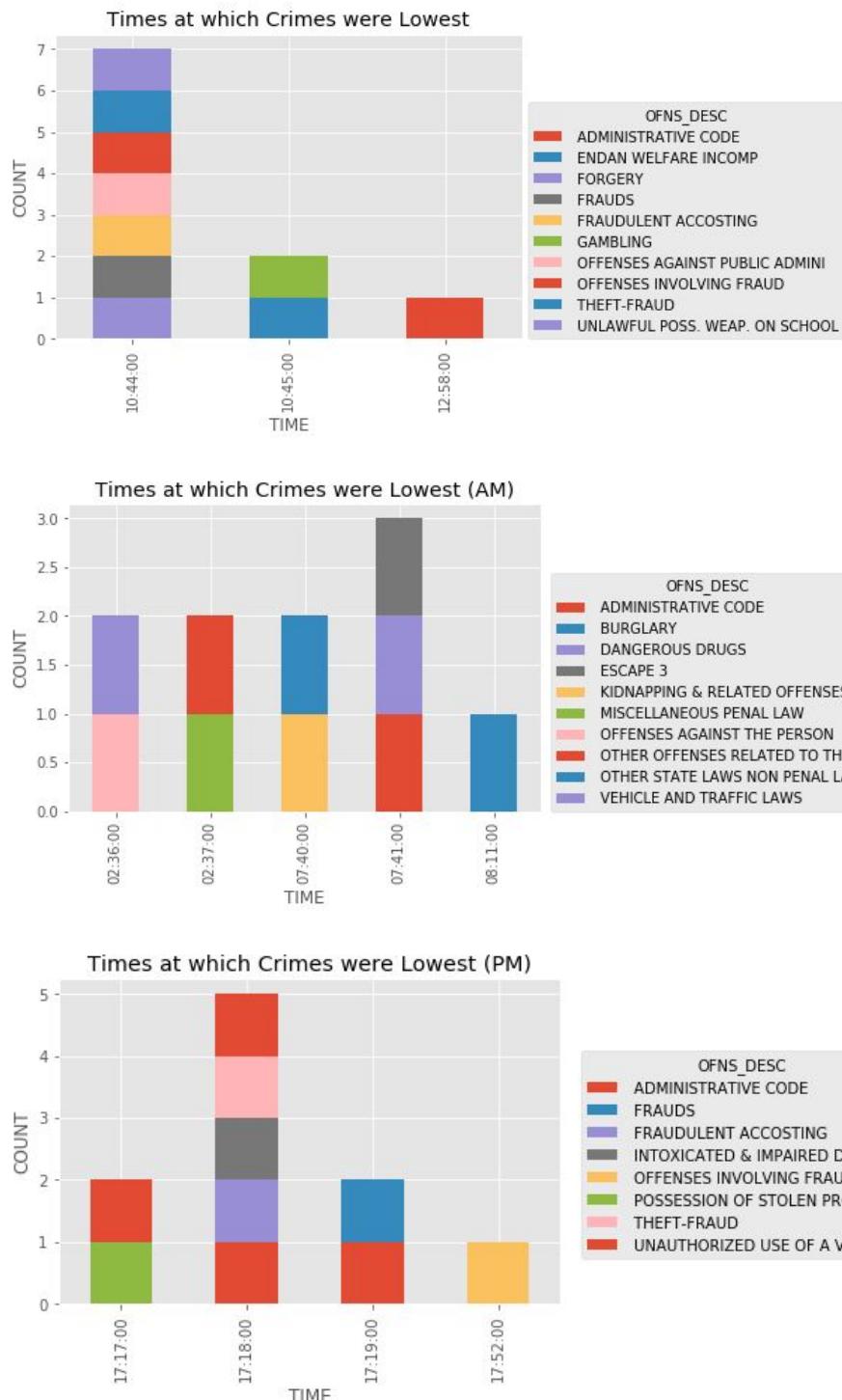


Highest



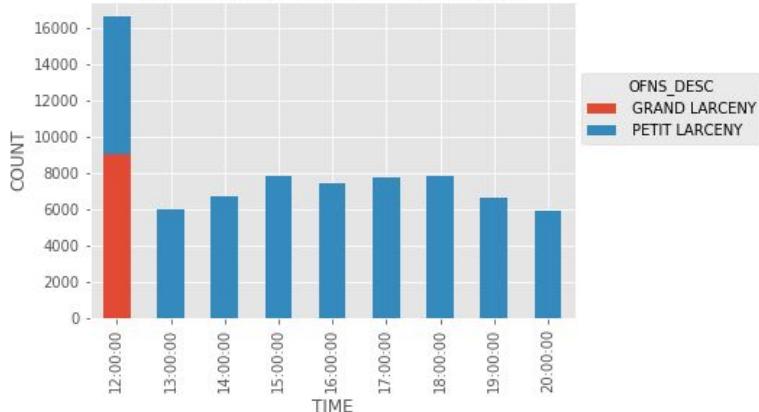
Manhattan

Lowest

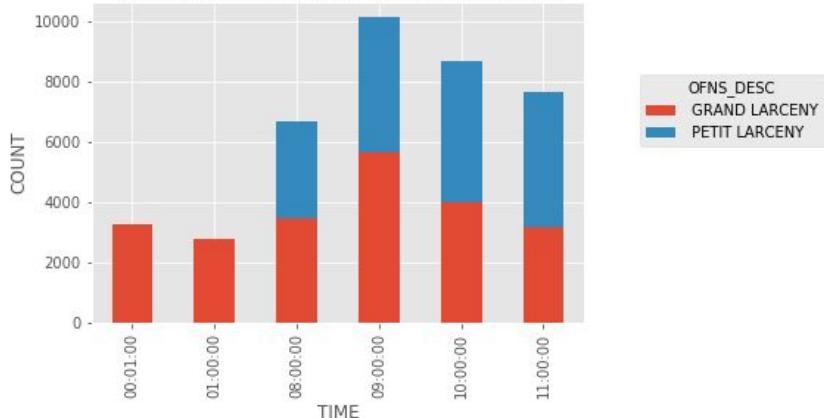


Highest

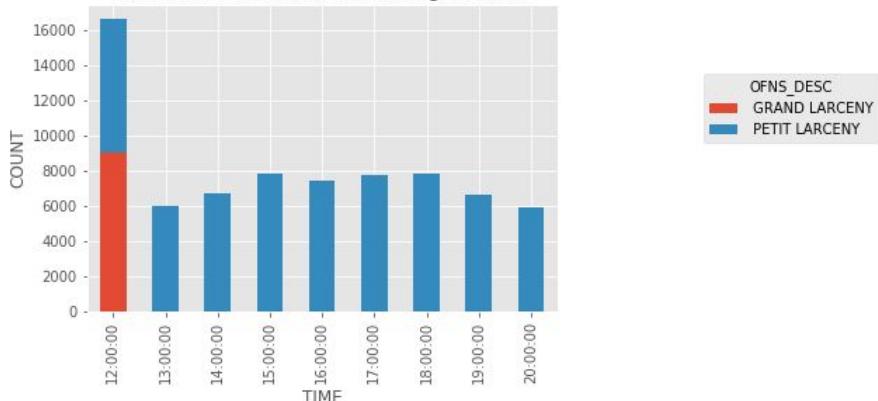
Times at which Crimes were Highest



Times at which Crimes were Highest(AM)



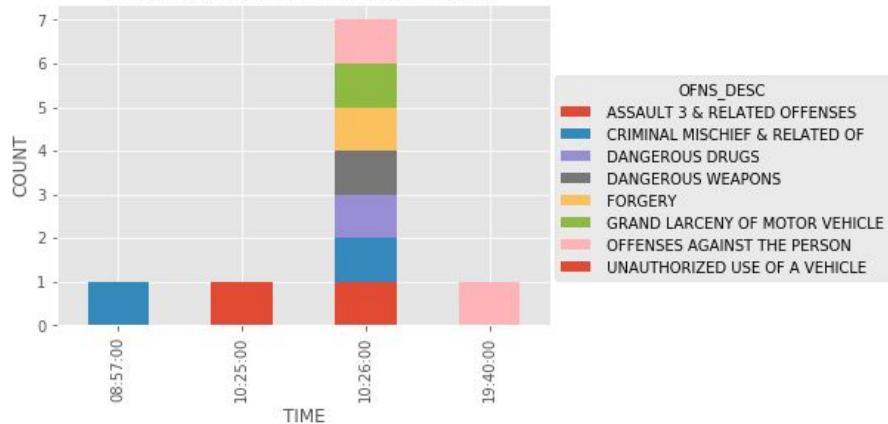
Times at which Crimes were Highest(PM)



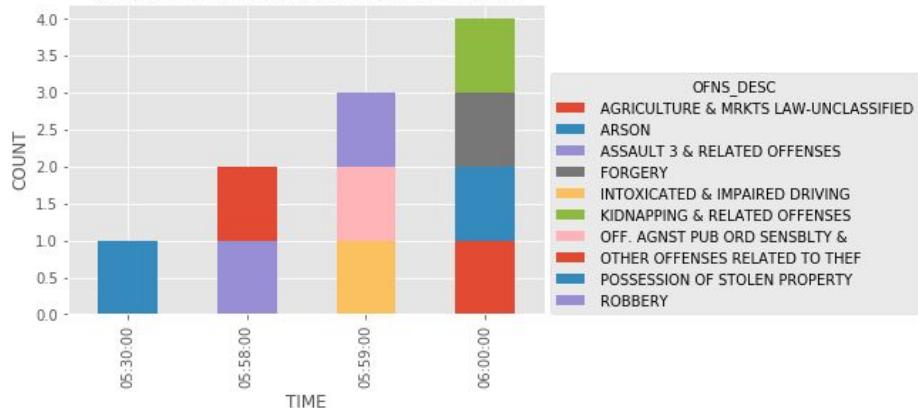
Staten

Lowest

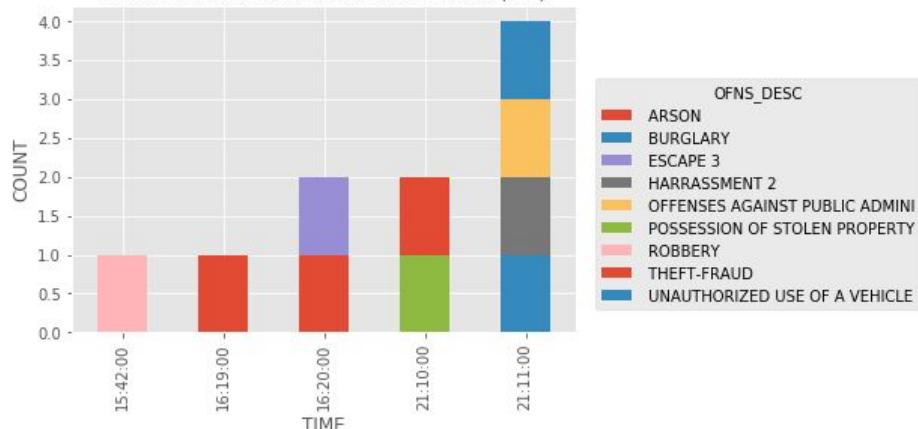
Times at which Crimes were Lowest



Times at which Crimes were Lowest (AM)

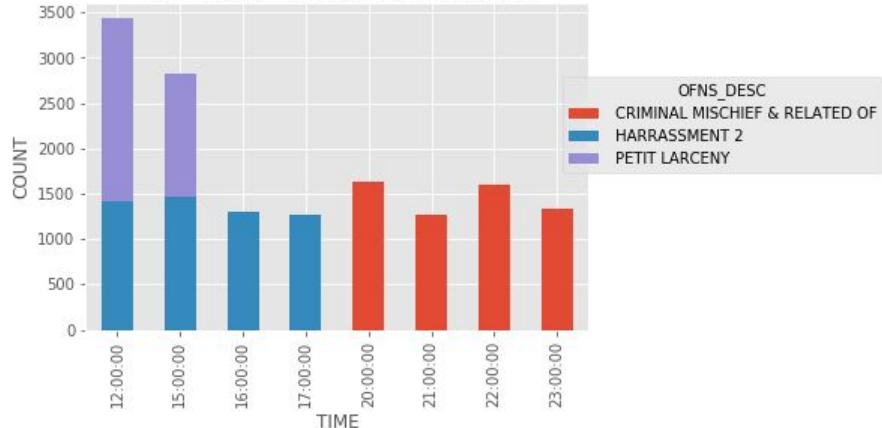


Times at which Crimes were Lowest (PM)

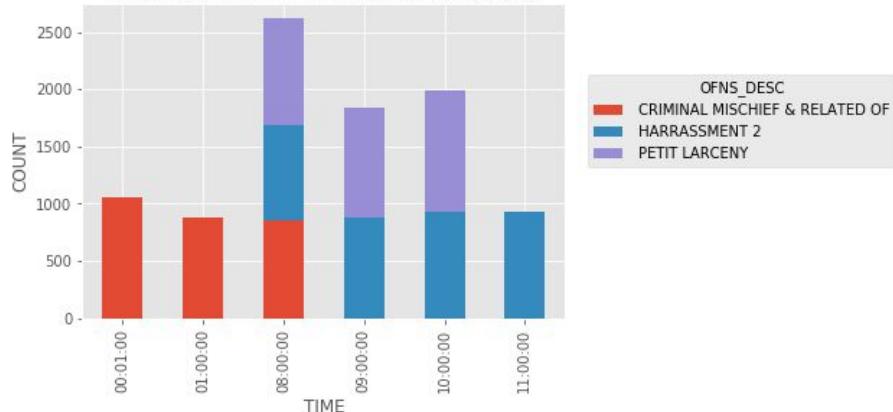


Highest

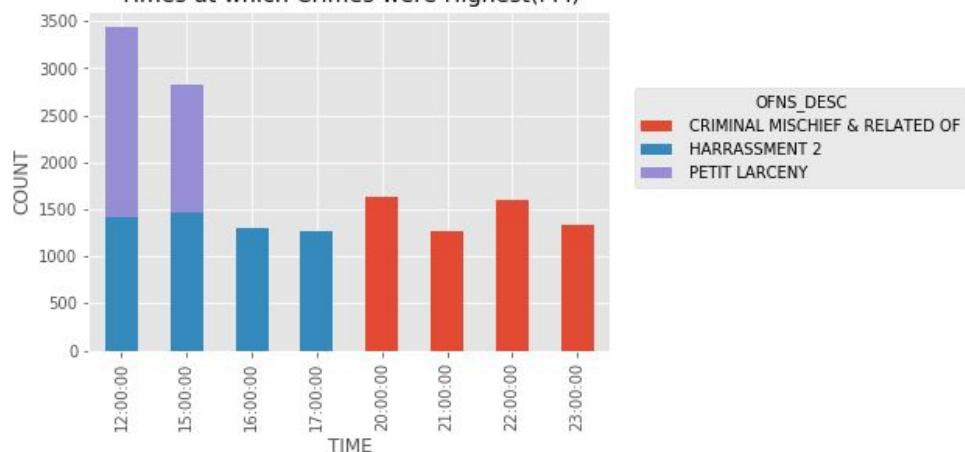
Times at which Crimes were Highest



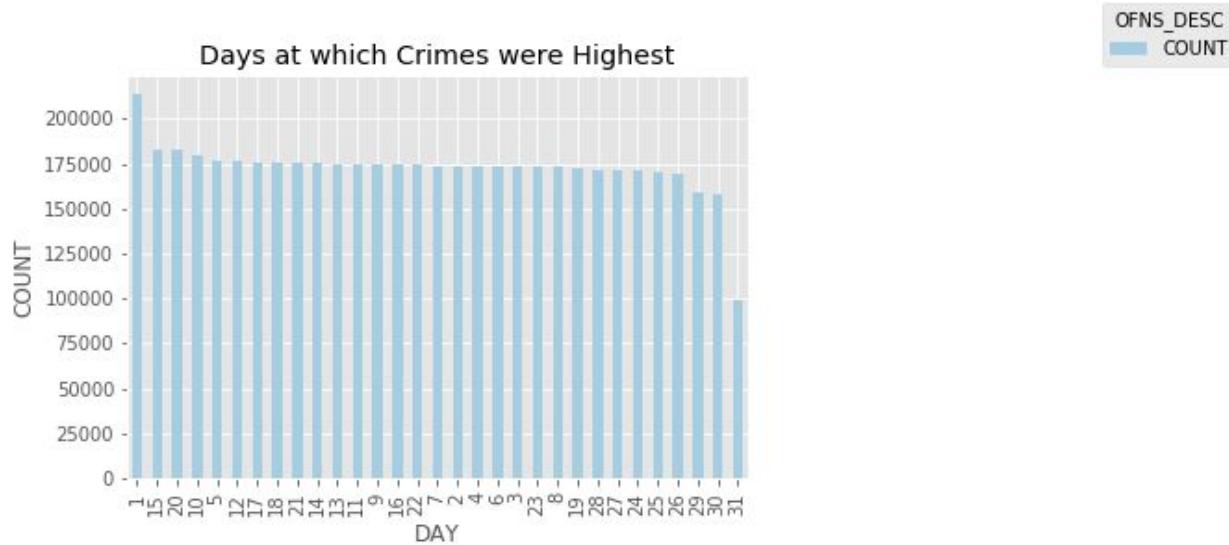
Times at which Crimes were Highest(AM)



Times at which Crimes were Highest(PM)



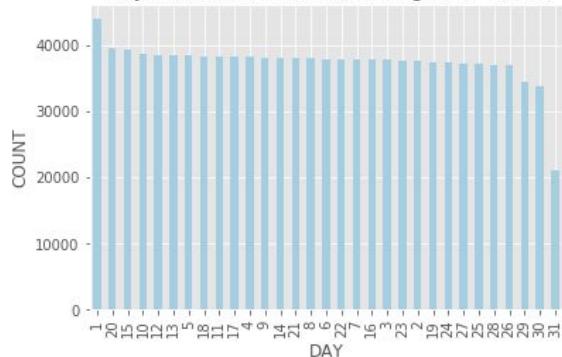
DAY



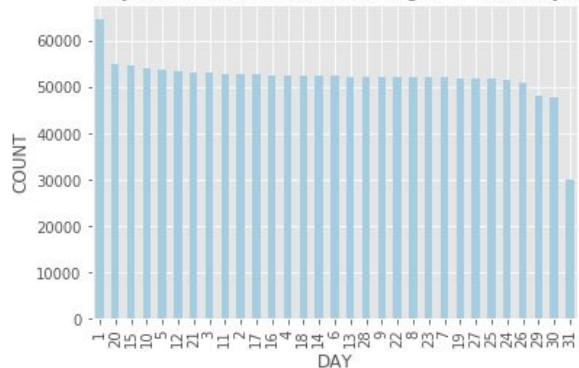
- Crimes are least on 31st and most on 1st
- It maybe because since 31st is not in all the months, the combined reports might be filed on 1st
- The second highest date is 15, followed by 20th
- 30th is the second smallest
- Paychecks are generally issued on 1st and 15th. This maybe a reason

BOROUGH WISE

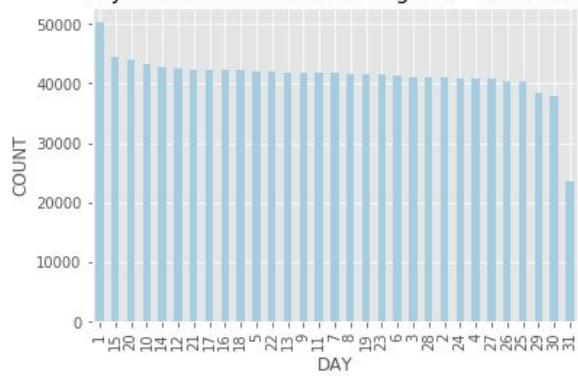
Days at which Crimes were Highest - Bronx



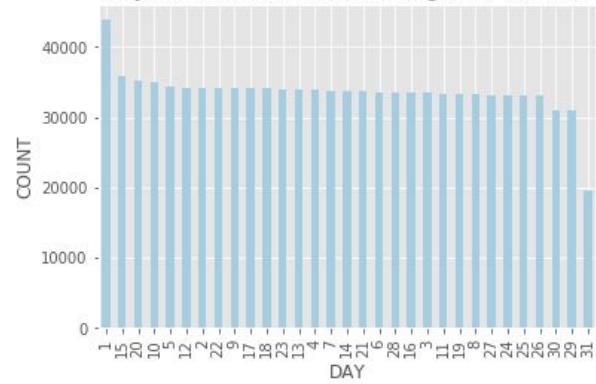
Days at which Crimes were Highest - Brooklyn



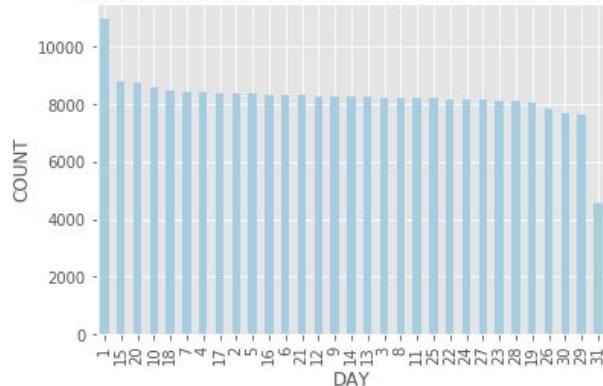
Days at which Crimes were Highest - Manhattan



Days at which Crimes were Highest - Queens



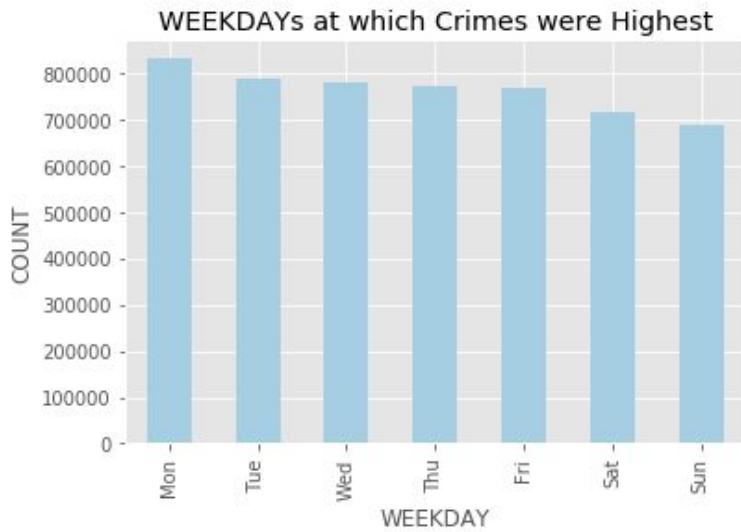
Days at which Crimes were Highest - Staten Island



- For Bronx and Brooklyn, after 1st, 20th is the next highest crime reported date.
- For the other boroughs the second highest is on 15th

WEEKDAY

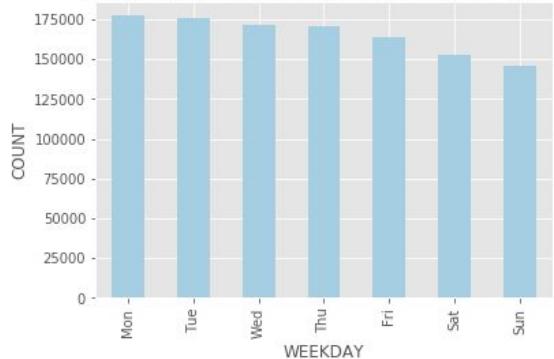
Looking at the Days data, we wanted to check whether the weekday and weekend had an impact on the reporting



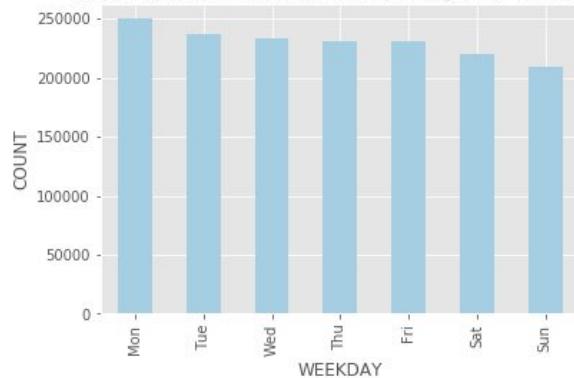
- As expected crime reported was highest on a monday and lowest during the weekends
- All the boroughs also followed the same pattern

BOROUGH WISE

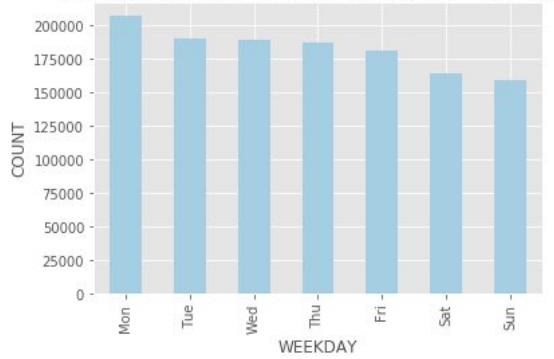
WEEKDAYS at which Crimes were Highest - Bronx



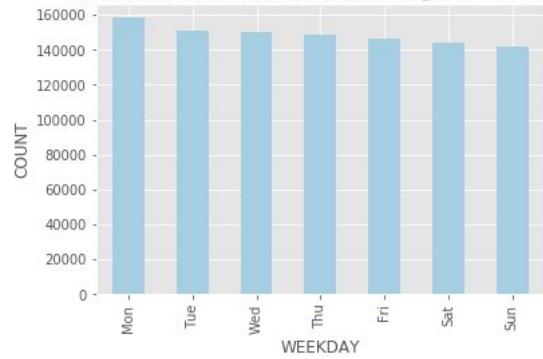
WEEKDAYS at which Crimes were Highest - Brooklyn



WEEKDAYS at which Crimes were Highest - Manhattan



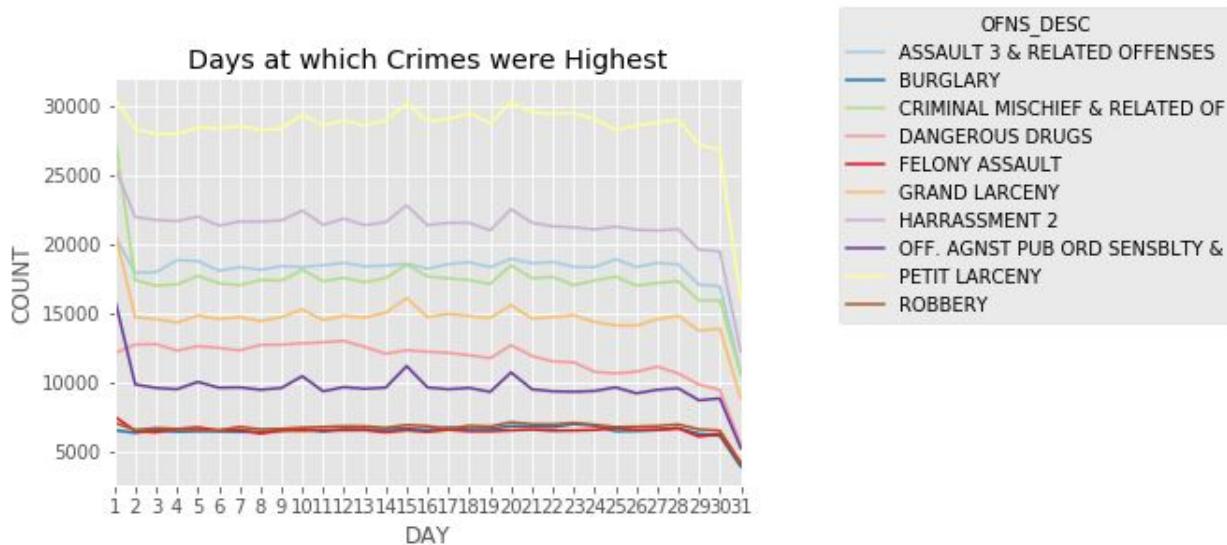
WEEKDAYS at which Crimes were Highest - Queens



WEEKDAYS at which Crimes were Highest - Staten Island



DAY CRIME WISE

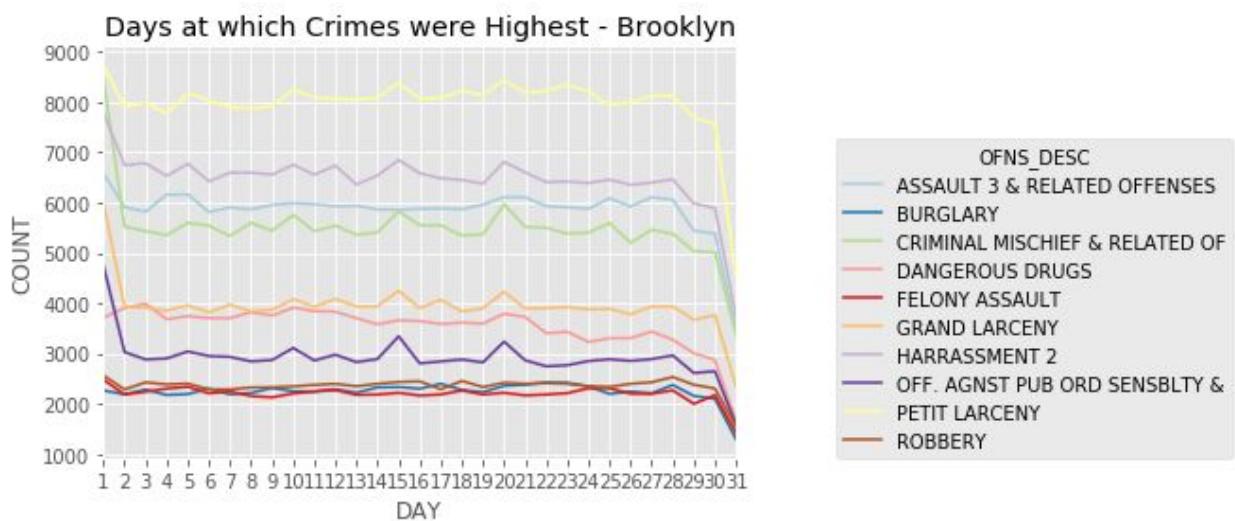
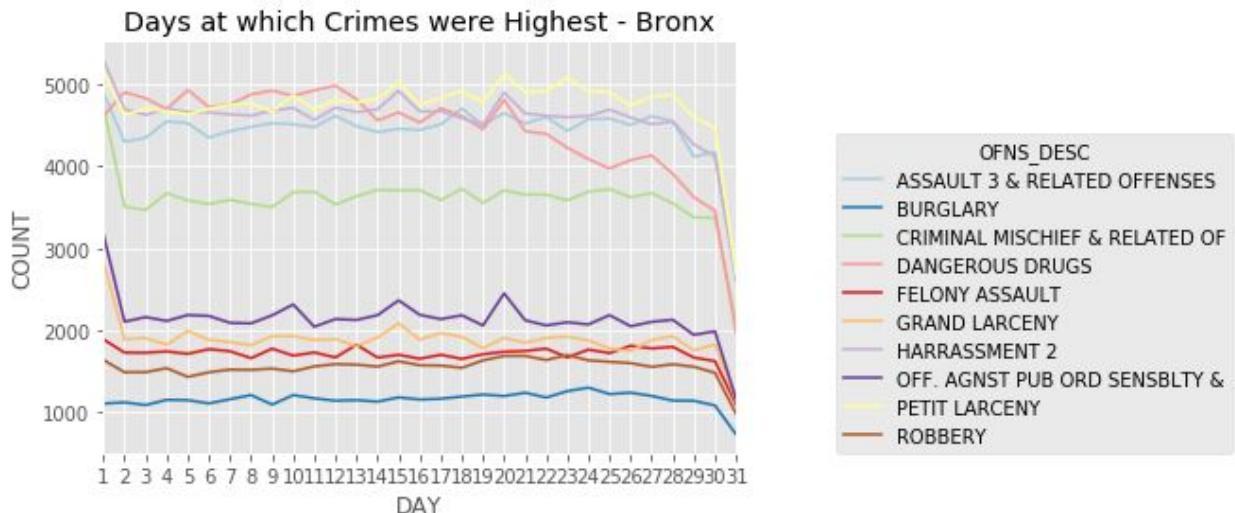


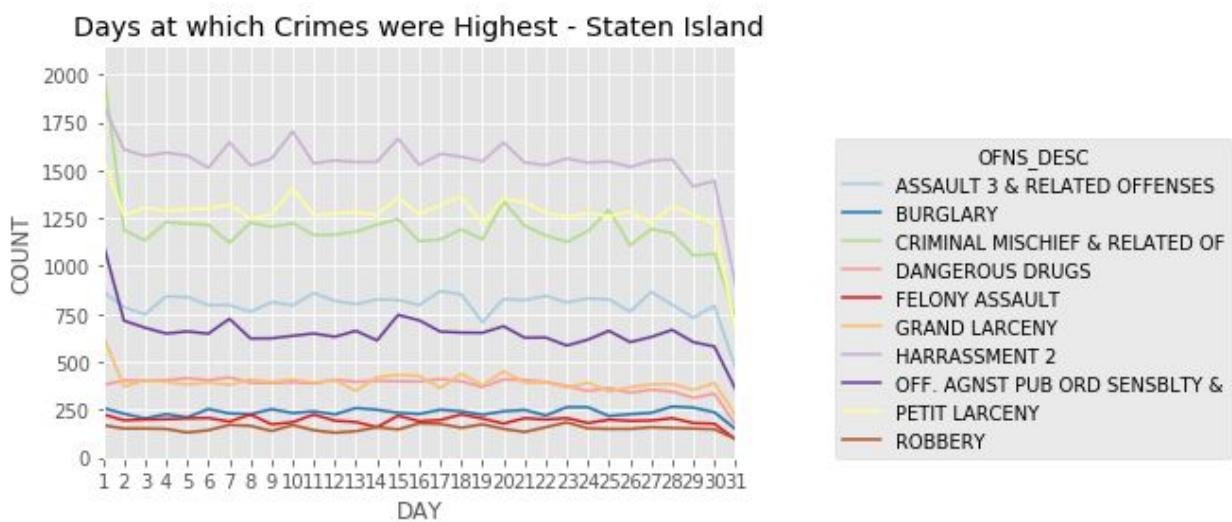
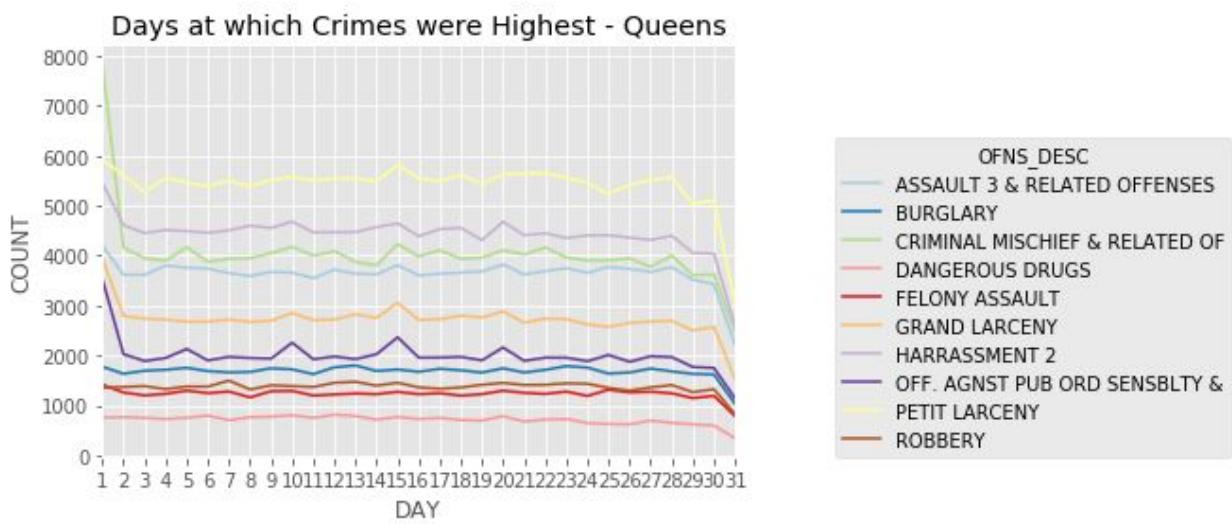
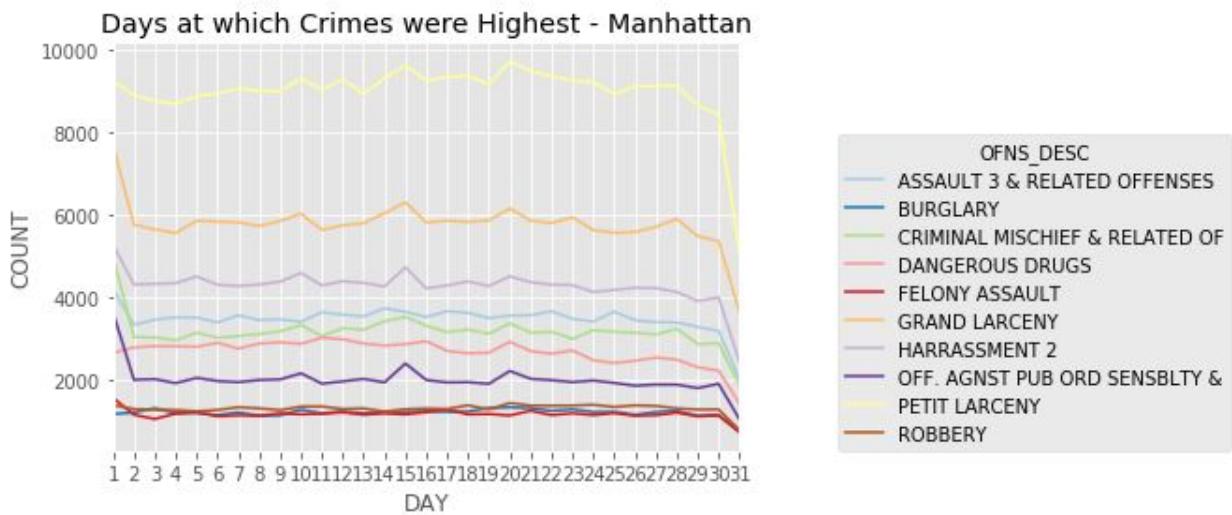
- The graphs are for the top 10 crimes as a graph with all the crimes made the graph too crowded
- There is a peak of the crime on the 1st of every month
- There is a decrease in the end of the month. This maybe due to the fact that Feb has only 28 days and only few months have 31 days
- There is again a rise around the 15th of a month
- It may be because most of the paychecks are issued on 1st and 15th of the month (by weekly) and in the graph also we could observe most of the crime reported were burglary, larceny and theft during that time.

<http://schools.nyc.gov/Offices/EnterpriseOperations/ChiefFinancialOfficer/DFO/PayrollAdministration/KeyDocuments/PayrollCalendars.htm>

- For these reasons we decided to look at month and year more in detail

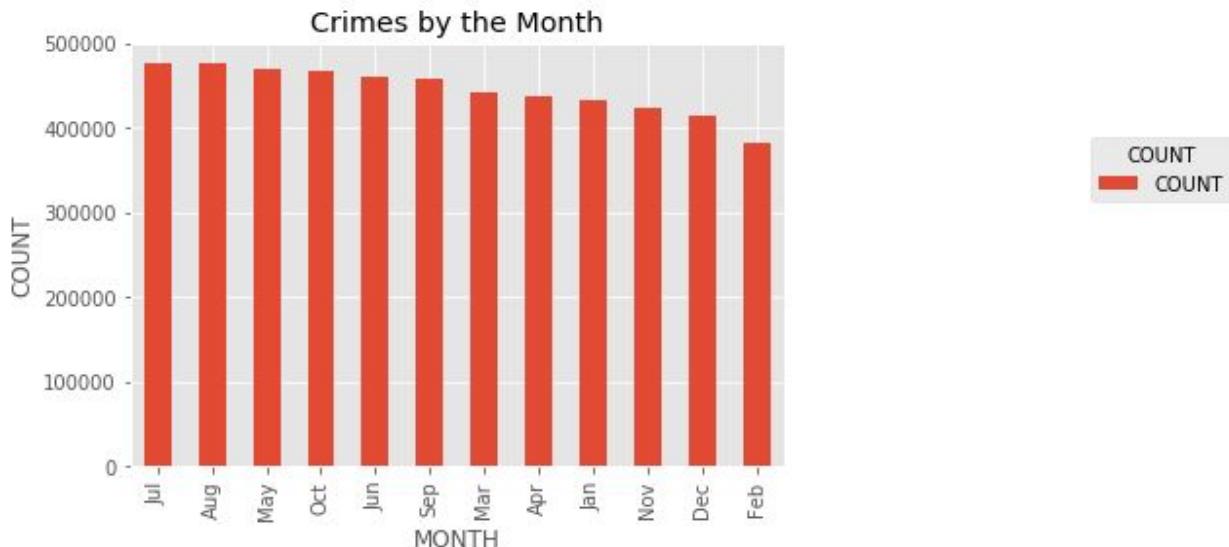
BOROUGH WISE





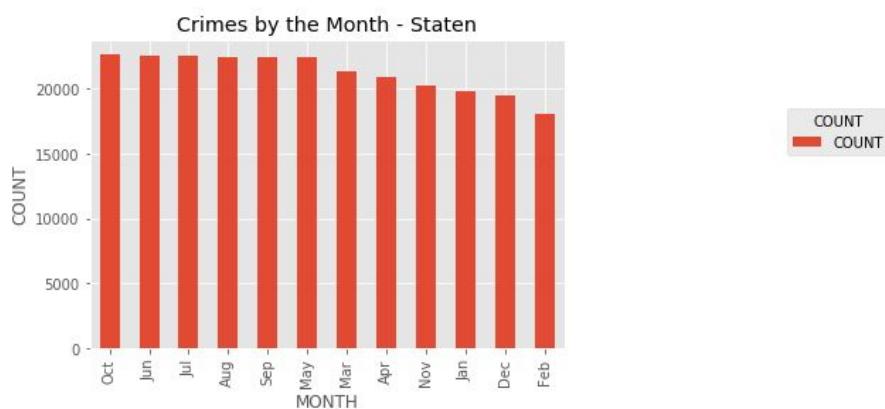
- In Bronx, Brooklyn, Manhattan there was also a peaking around 20th in addition to the ones around 1st and 15th
- For Queens the peaking for crimes was significantly more on 1st

MONTH



- Crimes in July was the highest and was lowest in February
- The climate may have had some impact on this

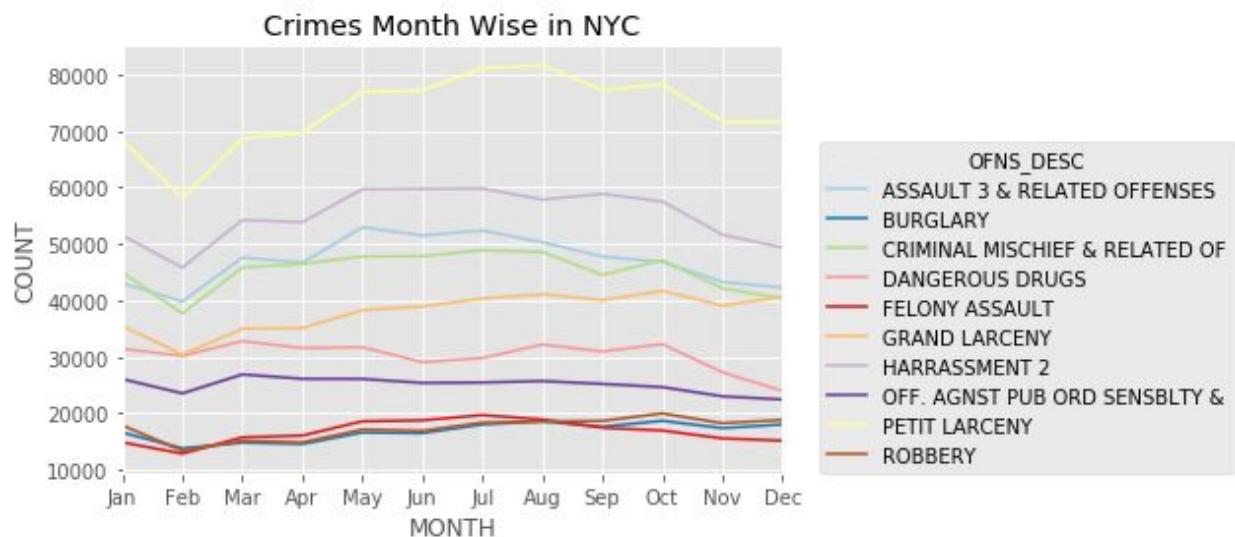
Borough Wise





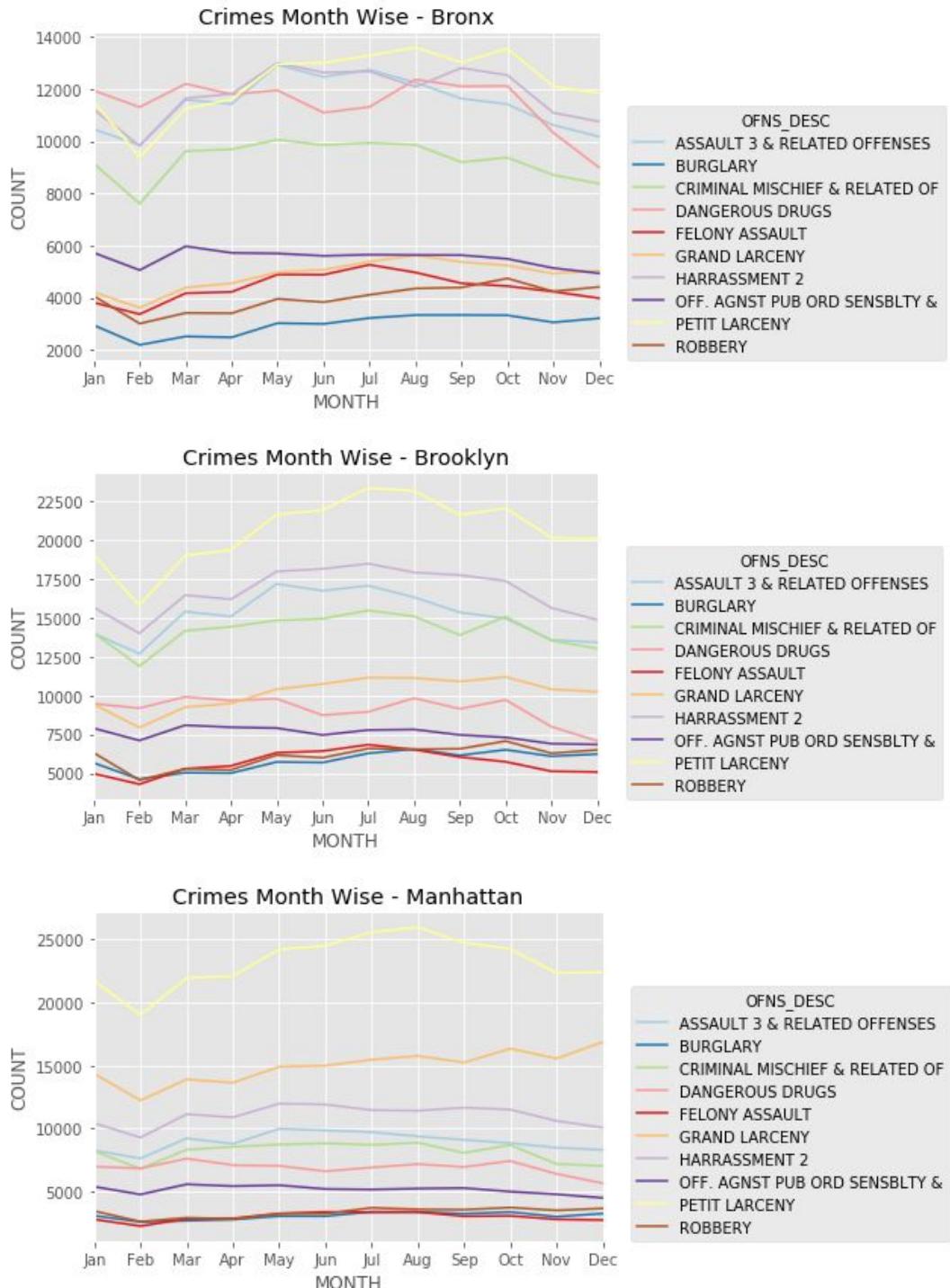
- Staten Island followed a different trend than the rest and reported crimes highest in Oct
- Queens & Brooklyn followed the overall by having the highest crimes reported in July
- Manhattan & Bronx reported the highest in August
- Lowest for all boroughs were in february
- It can be concluded for all boroughs except Staten Island, the highest crimes were reported in the summer months and lowest were in winter months.
- We will further explore this observation in data exploration part, where we will try to analyse the crime data with weather data.

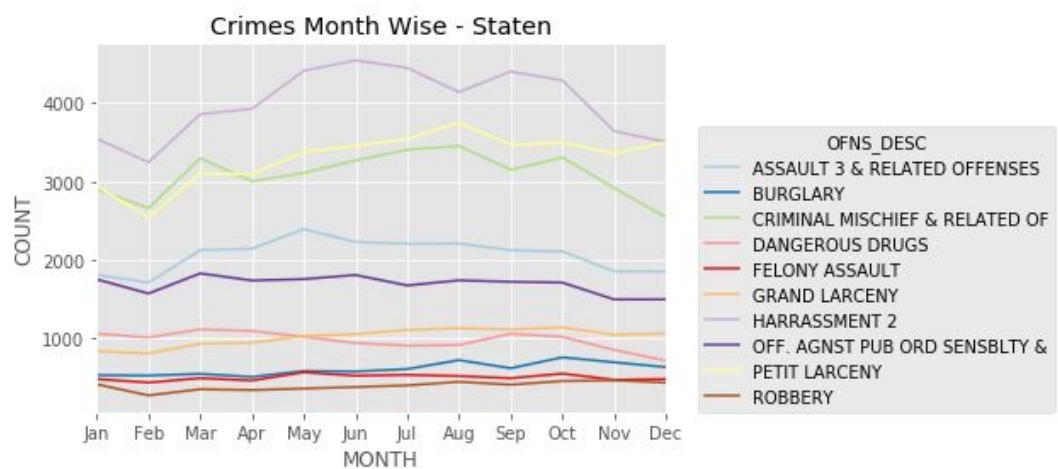
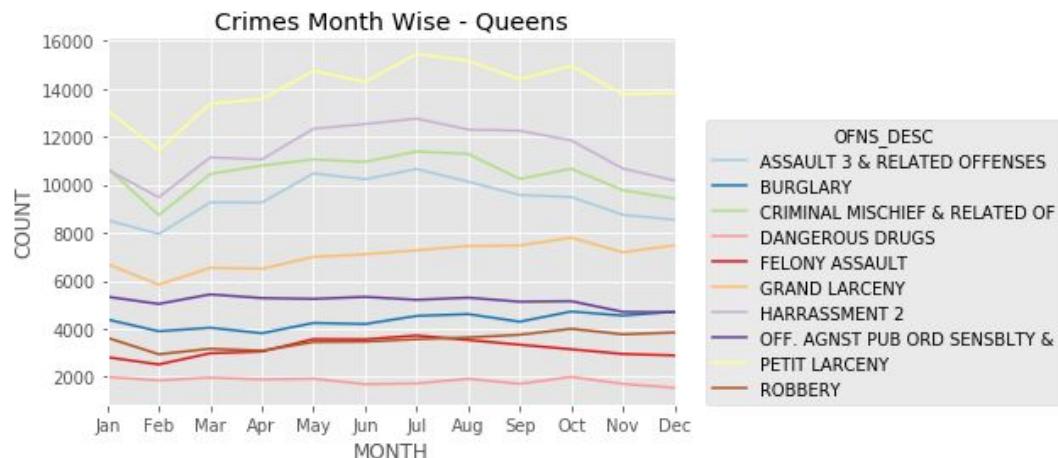
MONTH CRIME WISE



- The graph displays a distribution of the top 10 crimes month wise
- Petit Larceny was at its peak in July
- All the crimes are low in Feb

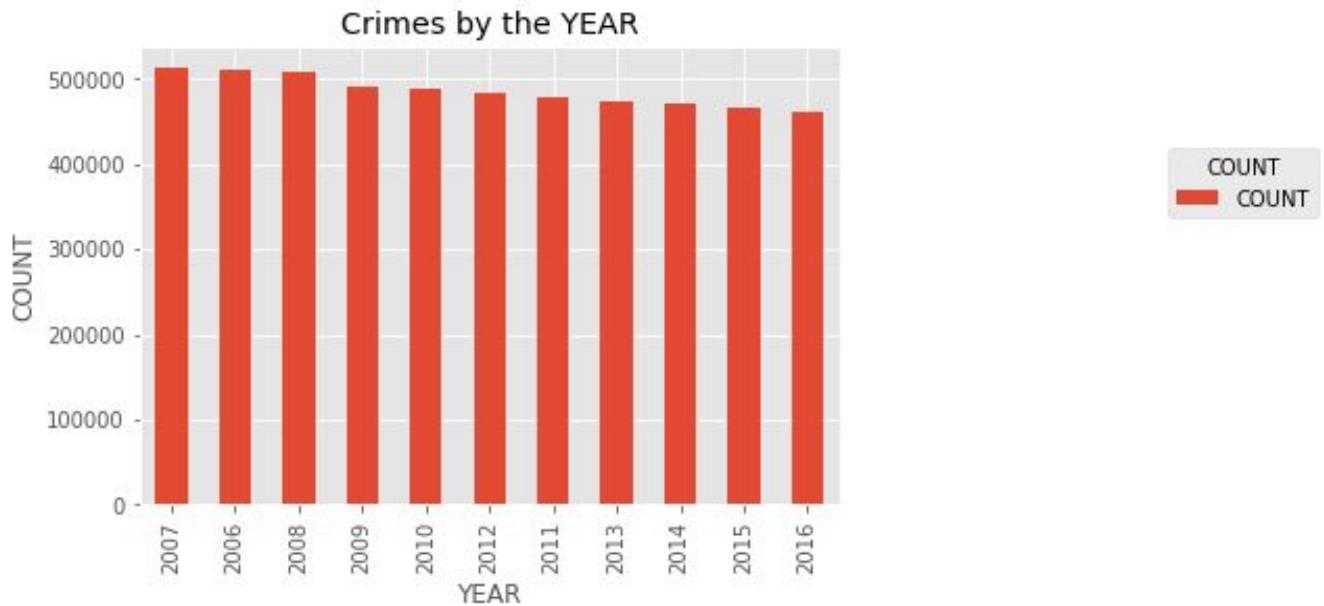
Borough Wise





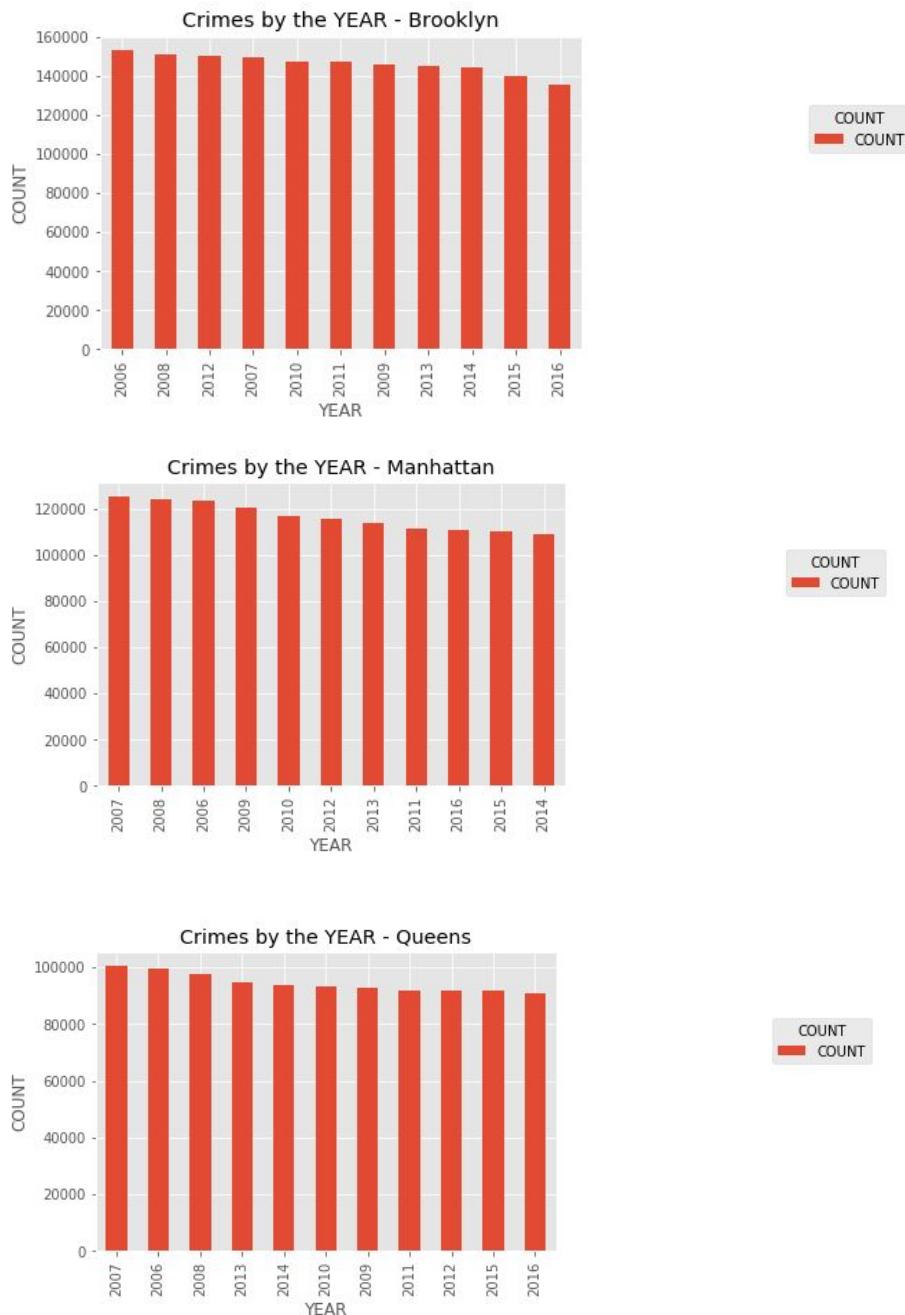
- In Staten Island Harassment is highest in June though overall total crime is highest in Oct
- Queens & Brooklyn have Petit Larceny peak in July
- Manhattan reported Petit Larceny highest in August
- Bronx reported overall crime highest in August however, Petit Larceny peaked in Oct
- Lowest for all boroughs were in february

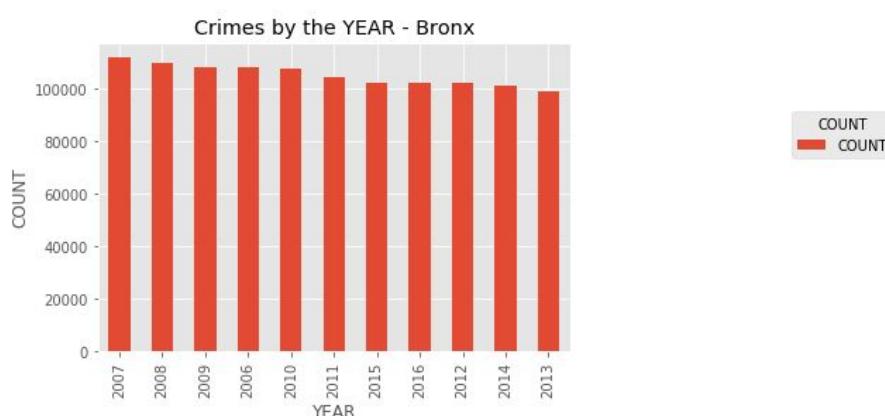
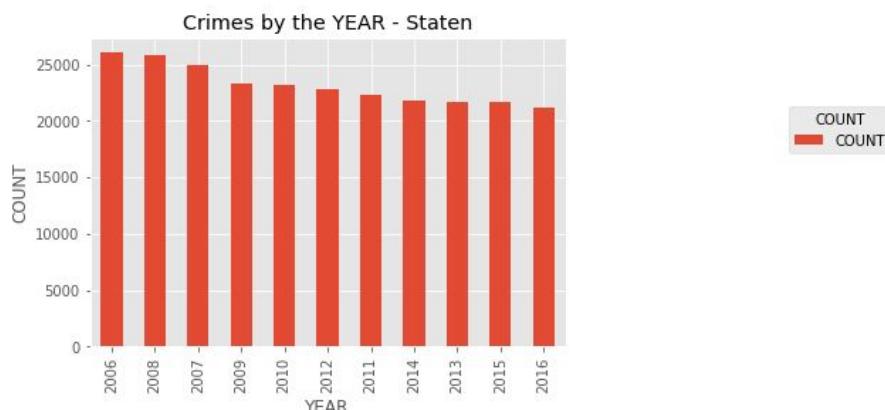
YEAR



- Crimes were highest in 2007
- It maybe because of the financial crisis

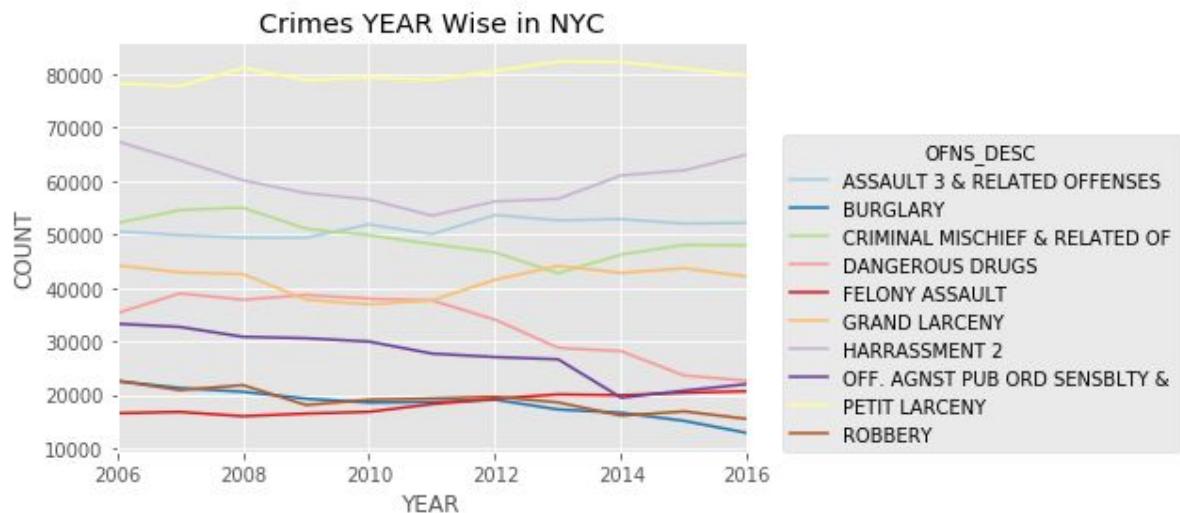
Borough Wise





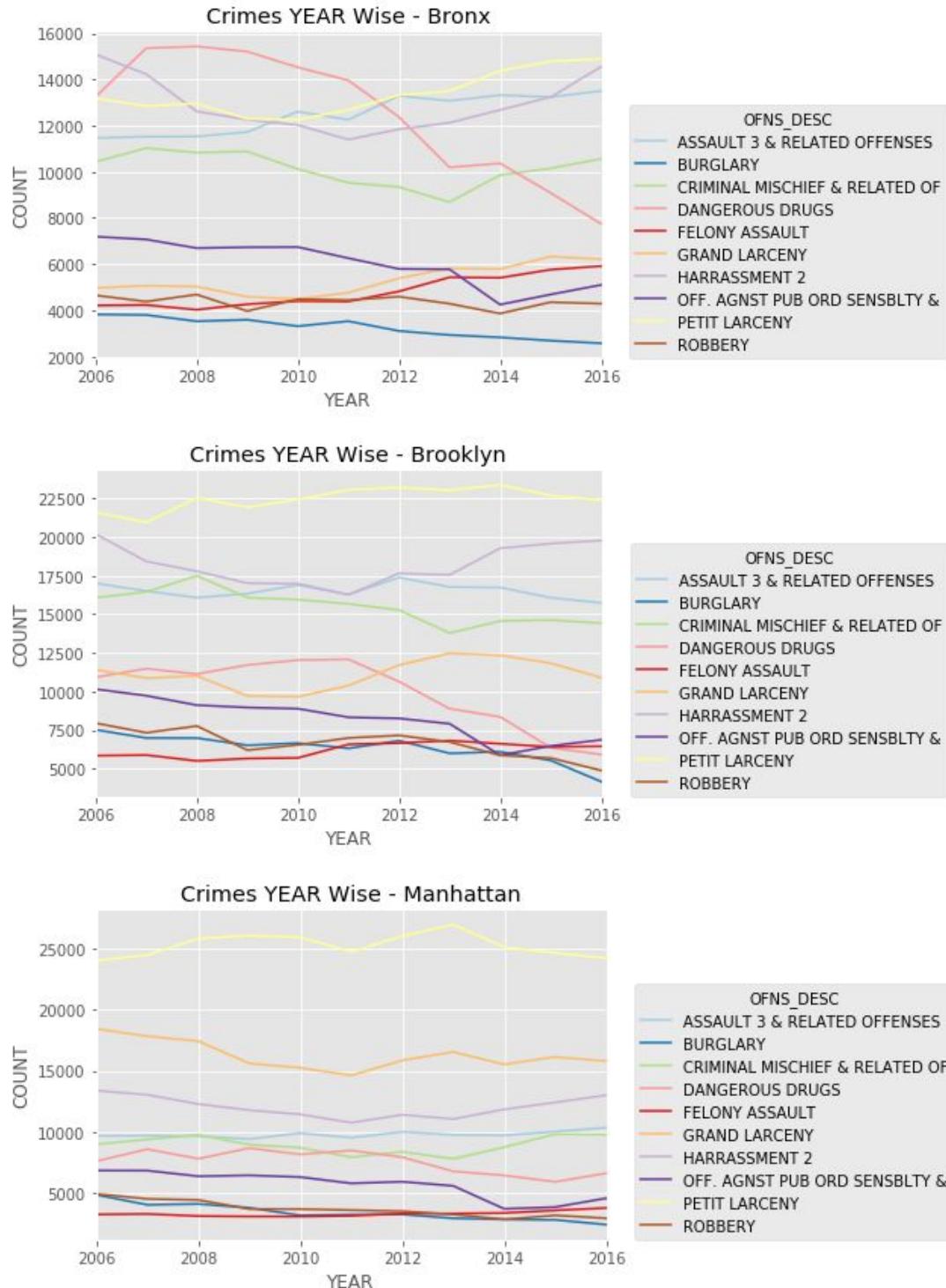
- For Bronx, Manhattan, Queens the year of highest crime is 2007 and lowest was in 2013 for Bronx, 2014 for Manhattan and 2016 for Queens.
- For Staten Island, Brooklyn the year of highest crime is 2006 and the lowest for both is 2016

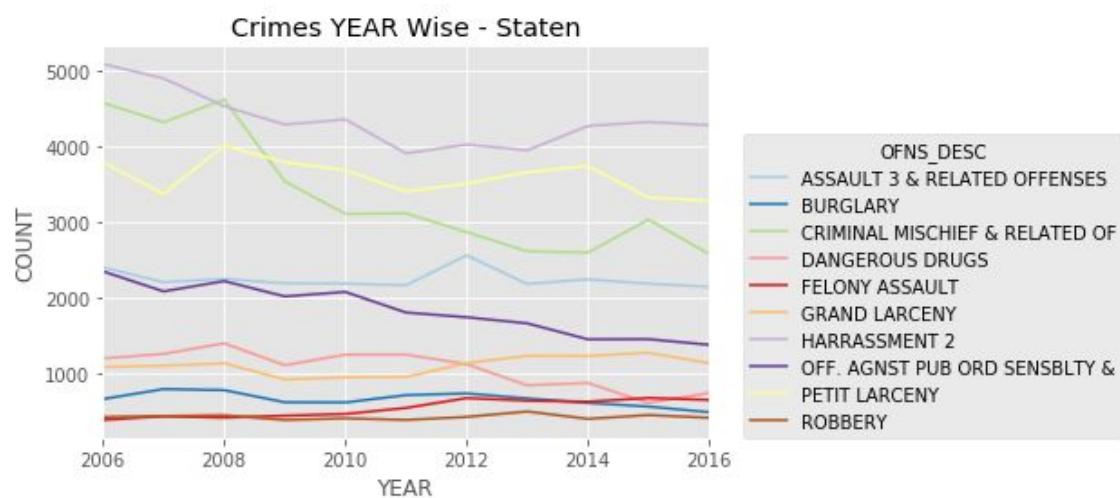
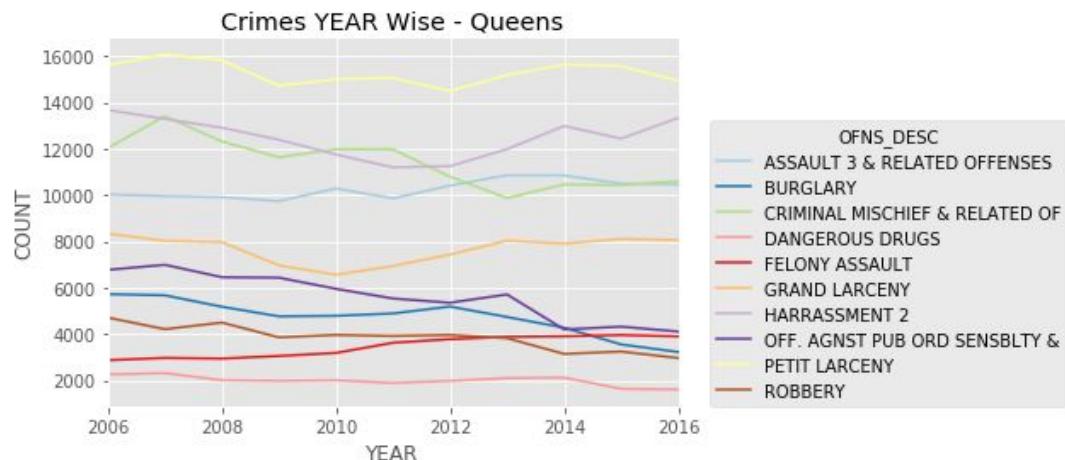
YEAR CRIME WISE



- Petit Larceny was almost stable through the years
- Burglary, Robbery, Dangerous Drugs was dropping
- Felony assault,grand larceny was on the rise
- Harassment 2 dipped low for 2011 but it rose again

Borough Wise



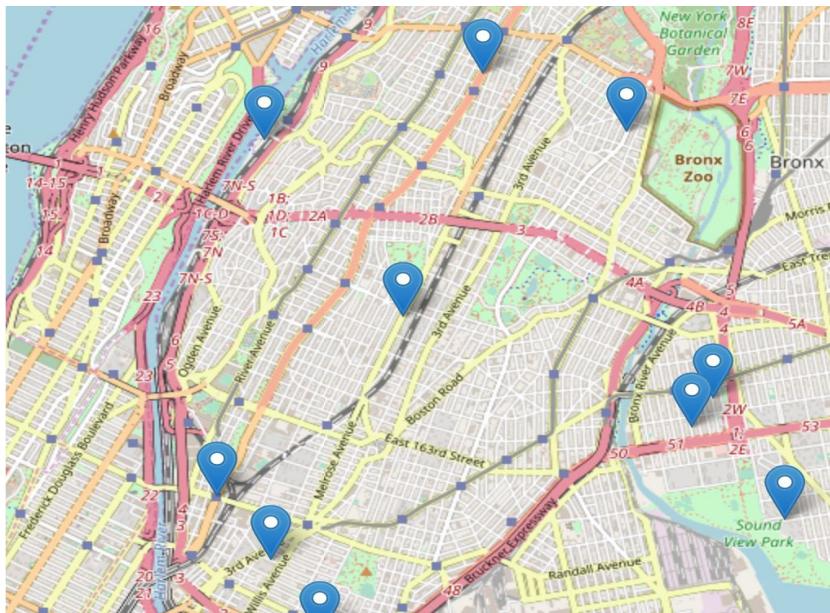


- Dangerous Drugs was steadily dropping in Bronx
- Petit Larceny, Harassment 2 and Assault 3 is on the rise in Bronx
- Crime in general was lowering in Staten Island
- Queens, Manhattan and Brooklyn seemed to roughly follow the same trends over the years

Geo-Location

- To further analyse what are the safer locations in each borough to locate a safe neighbourhood, we took the location dataset and analysed for the highest number of crimes in those boroughs. We identified the locations with highest offense for each borough, mainly assault and larceny were the highest reported crimes.
- The answer we got is a contrary to the answer we got using HAVDELOPT column which was known to be null a lot of times. Which proves HAVDELOPT as an unreliable column. The location column was valid more times than HAVDELOPT.

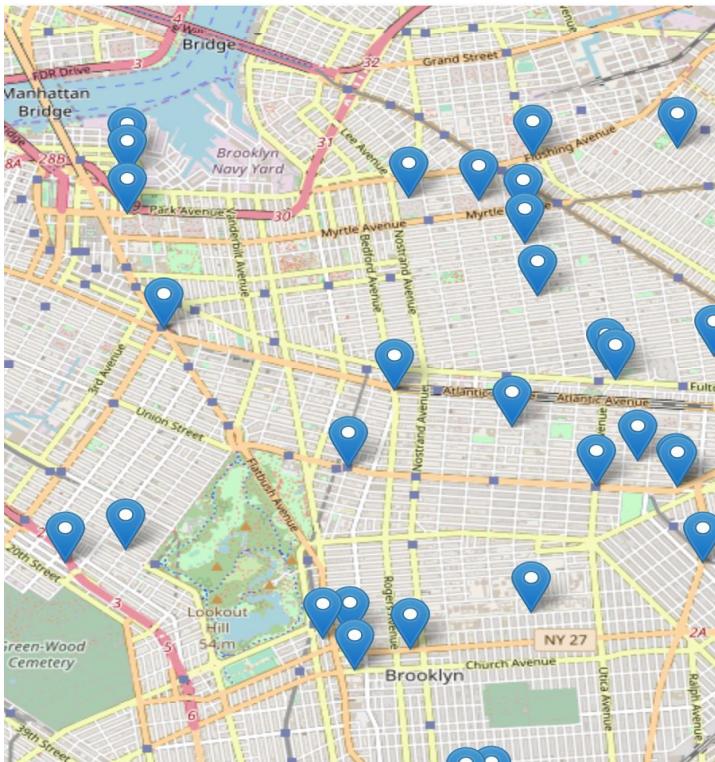
Bronx



- For the top most crime assault we fetched the location coordinates and got the locations against it.

Location Coordinates	Location Name
40.852141187, -73.922375722	Richman Plaza
40.82506399, -73.877397261	Boynton - Watson Ave
40.805501602, -73.916477255	East 137th Street - Cypress Ave

Brooklyn



For Brooklyn we got:

Location Coordinates	Location Name
40.695864034, -73.981822445	223-249 Tillary St
40.664239422, -73.908425011	275 - 293 Dumont Ave
40.667297692, -73.909163377	283 - 291 Stutter Ave

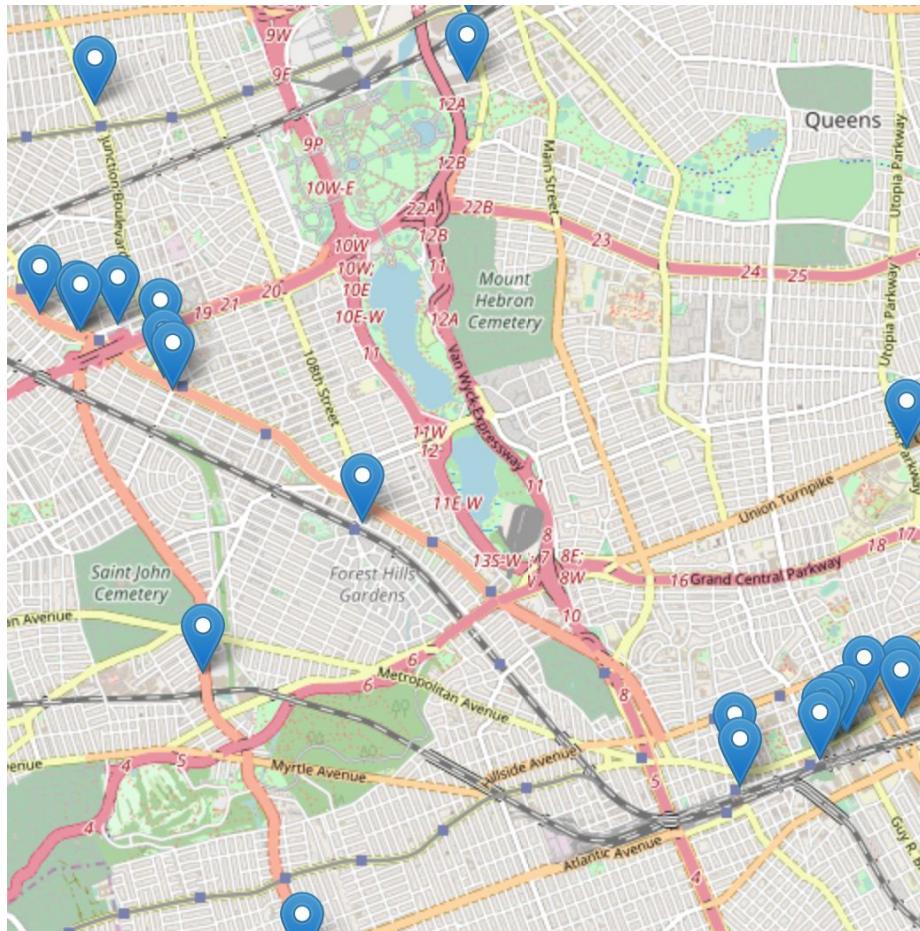
Manhattan



For Manhattan we got:

Location Coordinates	Location Name
40.750430768, -73.989282176	34th Street - 6th Ave
40.756642042, -73.988371577	W 42nd Street - Times Sq
40.76187469, -73.966358084	E 60th Street - 3rd Ave

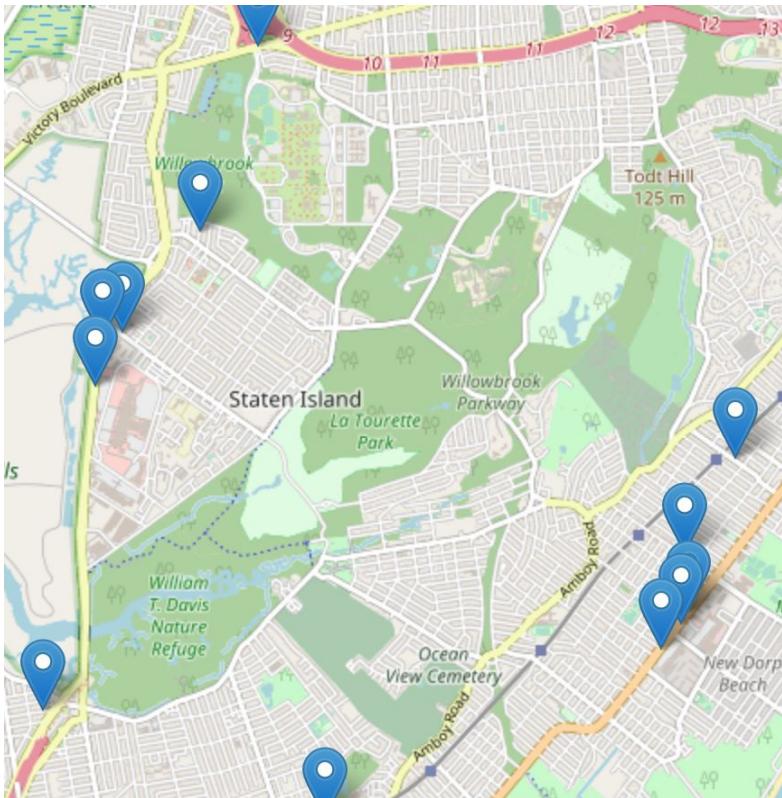
Queens



For queens we got:

Location Coordinates	Location Name
40.735783278, -73.810625803	Parsons Blvd - 65th Ave
40.773793373, -73.936280628	1-1 Astoria Blvd
40.755675739, -73.948039558	Vernon Blvd - 41st Ave

Staten



For Staten Island we got:

Location Coordinates	Location Name
40.615491223, -74.08165623	162-180 Park Hill Ave
40.638984176, -74.171252305	Holland Ave
40.64173826, -74.075678037	5 Bay Street

DATA QUALITY

- HAVDEVLOPT field did not have housing development names. Some rows had street numbers and not even building numbers. It was invalid for majority of the cases.
- For the same KY_CD there might be multiple OFSC_DESC. Which made OFSC_DESC a less reliable field. However, KEY_CD to a normal individual is just numbers, so we had to use them.
 - KY_CD 101 was cleaned out completely as it was an invalid row.
- Also, the OFSC_DESC contained too many punctuations which made making a csv file rather difficult.
- There were multiple Offense Descriptions (OFNS_DESC) for the same Offense Classification Codes (KY_CD). And multiple offense Description implying the same offense, so we correlated them all as one category while doing further analysis.
- In the columns, complaint to time and complaint for time some rows have incorrect 24-hour time format such as instead of “00:00:00”, “24:00:00” is written. Therefore, such rows should be merged while analyzing the crime rate by hour of day.

Redundant Data

- X_COORD_CD, Y_COORD_CD, Latitude, Longitude and Lat_Lon had the same data. So we chose to use Lat_Lon
- KY_CD, PD_CD, LAW_CAT_CD are not the exact same and they convey different meaning. However, they are more or less similar. We chose to go with KY_CD as it is broad enough but not too specific.
- OFNS_DESC, PD_DESC also had similar problems. So, OFNS_DESC was chosen.

DATA SUMMARY

- Crimes Borough wise sorted on crime count:-

#	Borough	Count
1	Staten Island	265641
2	Queens	1105621
3	Bronx	1209647
4	Manhattan	1331760
5	Brooklyn	1666903

- Count wise Brooklyn contributed the most amount of crime.
- Maximum crime offense like assault, larceny, mischief were evident in each of the boroughs.
- The top 10 crimes in NYC are the following
 - PETIT LARCENY (341),
 - HARASSMENT (578),
 - ASSAULT 3 & RELATED OFFENSES (344),
 - CRIMINAL MISCHIEF & RELATED OF (351),
 - GRAND LARCENY (109)
 - OFF. AGAINST PUB ORD SENSIBILITY & (361)
 - DANGEROUS DRUGS(235)
 - ROBBERY(105)
 - BURGLARY(107)
 - FELONY ASSAULT(106)
- Bottom 10 crimes in NYC are the following
 - KIDNAPPING AND RELATED OFFENSES(124)
 - OFFENSES AGAINST MARRIAGE UNCL(362)
 - UNDER THE INFLUENCE OF DRUGS (577)
 - ABORTION (123)
 - OTHER STATE LAWS NON PENAL LAW (364)
 - KIDNAPPING (124)
 - OTHER TRAFFIC INFRACTION (881)
 - FORTUNE TELLING (357)
 - LOITERING (672)
 - LOITERING/DEVIATE SEX (460)

- ADMINISTRATIVE CODES (685)

- There is steady decrease in crime rate in New York city after 2010, with effect of strict and better law policies.
- February month is reported with least number of crime.
- Summer months reported the most amount of crimes.
- We could see similar trend for every month that is maximum number of crimes were reported on 1st and again peaks up on 15th. Last day has least number of crime reported.
- Time hour for most of crimes was maximum around 12 pm that is lunch hour and minimum during midnight hours.
- Crimes are reported the most on Mondays which maybe because it is the first working day of the week.
- The park with the most amount of Crime was Central Park at the same time, Central Park PD had the least amount of crime compared to other PD.
- 75th Precinct reported the most amount of crime.
- Housing Development wise we found Castle Hill had reported the most amount of crimes.
- Year 2007-2008 was reported with maximum number of crimes. This maybe due to the economic crisis.
- We also plotted the maps identifying the areas with maximum crime activities in each borough and came up with coordinates of the locations, thus identifying safe neighbourhood areas.

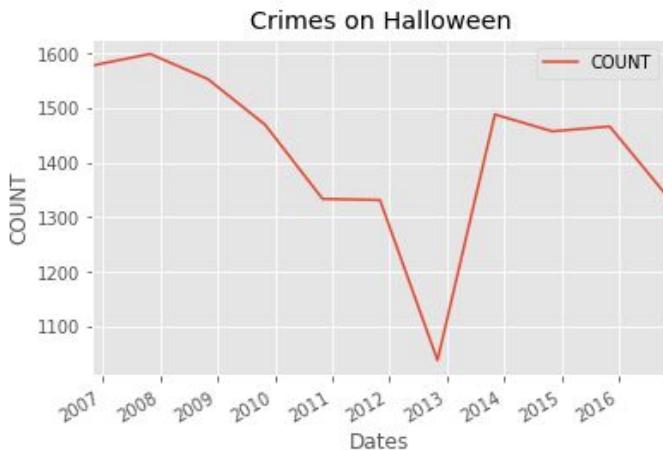
DATA EXPLORATION

Key Calendar Dates

There are certain dates in the year that security in NYC is high. We wanted to explore and see if the crimes for those dates were ever high over the years to cause NYPD to be so vigilant.

We checked the following dates:-

- **Halloween**



- Halloween Parades in 2017 had a few crime attacks
- When we checked the crime history we found out an all time low in 2012
- In 2012, Hurricane Sandy was active. So crime was obviously low.

Source:

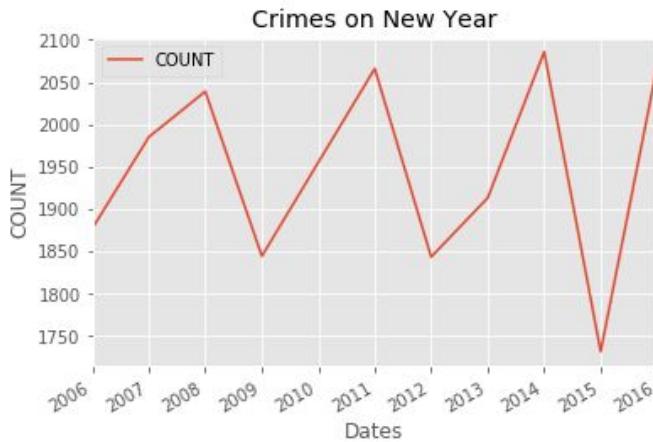
[<http://www.foxnews.com/travel/2012/10/30/in-wake-storm-nyc-village-halloween-parade-cancelled-1st-time-in-3-year-history.html>]

- As the parade was cancelled, this does show a relation to the halloween parade.
- 2008 crimes peaked but that maybe due to the financial crisis.
- Due to economic crisis incidence of robbery may double, and homicide and motor vehicle theft also increase, according to the UNODC report.

Source:

[<http://www.unodc.org/unodc/en/frontpage/2012/February/economic-crises-can-trigger-rise-in-crime.html>]

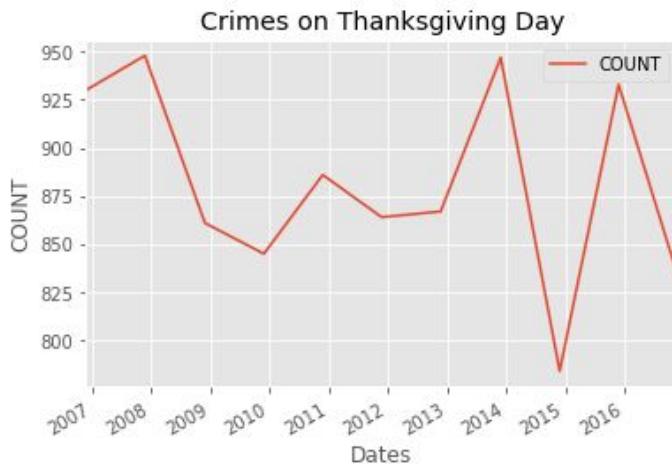
- **New Year**



- Crime on New Year is relatively high
- A million people is supposed to gather to see the ball drop so an increase in crime is obvious
- Lowest was in 2015 maybe due to climate
- The winter of 2014-2015 had reported some record breaking low temperatures.

Source: [https://en.wikipedia.org/wiki/2014%E2%80%9315_North_American_winter]

- **Thanksgiving**



- Thanksgiving Day is shopping day in nyc. Shops are open throughout the night and people are all walking about
- For some reason it was at all time low in 2014 and peaked in 2007 which is related to the financial crisis.

- In 2014, there was a move to pass a legislative to allow performers at Times Square to pose for pictures and collect tips. The crimes might have been as a protest by performers to prove their point.

Source:

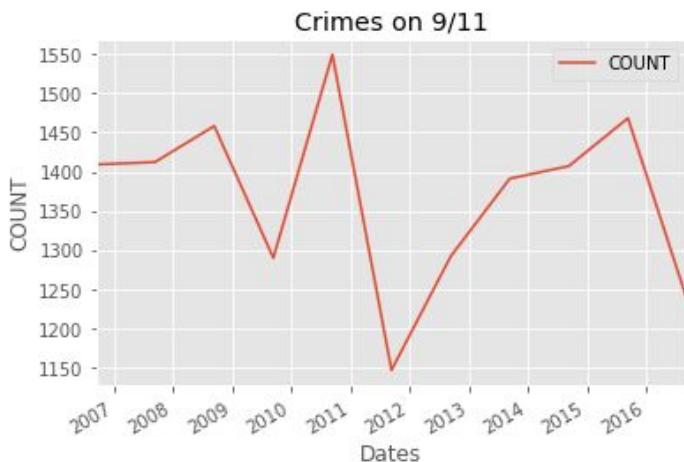
[https://www.washingtonpost.com/blogs/govbeat/wp/2014/08/21/new-york-city-wants-to-make-all-the-costumed-characters-you-pose-with-in-times-square-register-and-wear-id/?utm_term=.d0cedb9be969

[<http://www.nydailynews.com/new-york/nyc-crime/man-dressed-spider-man-punches-nypd-times-square-article-1.1881561>]

- The winter of 2014-2015 had reported some record breaking low temperatures.

Source: [https://en.wikipedia.org/wiki/2014%E2%80%9315_North_American_winter]

● 9/11



- The crimes on 9/11 was highest in 2010 and lowest in 2011
- There was a terrorist attempt on May 10, 2010. This may have caused some disturbance for 2010.

Source:

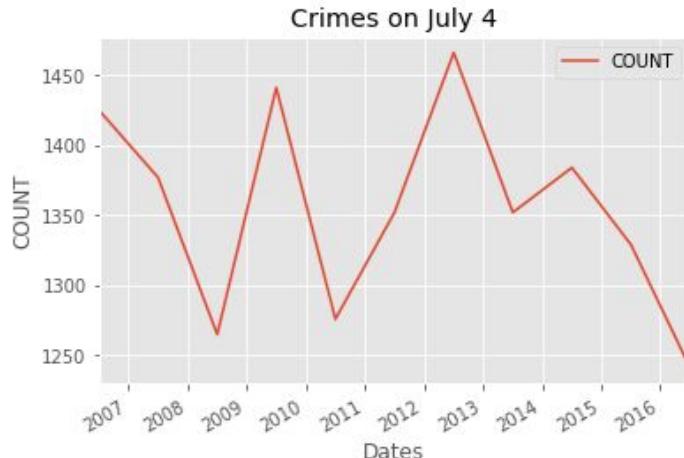
[<http://www.nydailynews.com/new-york/list-terror-attacks-foiled-plots-nyc-article-1.3692532>]

- Osama Bin Laden was killed in May 2, 2011. This and more heightened security might have been a reason for the low. The War Memorial was opened in 2011 which would account for extra police monitoring.

Source:

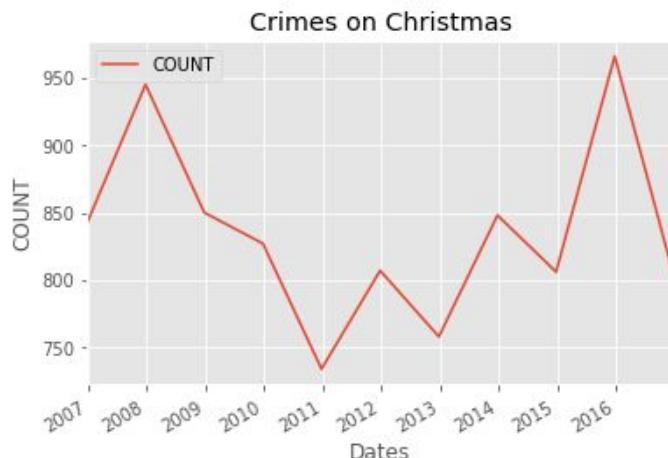
[https://en.wikipedia.org/wiki/National_September_11_Memorial_%26_Museum]

- **July 4**



- Crime was lowest in 2008 and highest in 2012
- This maybe due to the shootings on 2012 Source: [\[https://www.huffingtonpost.com/2012/07/05/nyc-shootings-12-shot-fourth-july_n_1651768.html\]](https://www.huffingtonpost.com/2012/07/05/nyc-shootings-12-shot-fourth-july_n_1651768.html)
- There was a Boston Red Sox vs New York Yankees match in New York in 2008. This with the fireworks would have kept people busy.
Source:[<https://www.baseball-reference.com/boxes/NYA/NYA200807040.shtml>]

- **Christmas**



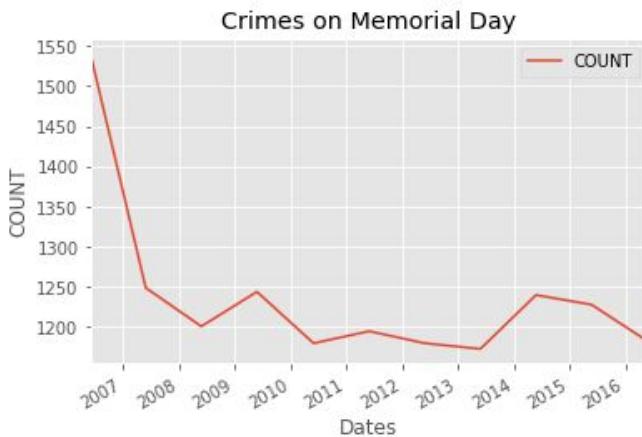
- Crimes on christmas day is generally lower than other days
- It was lowest in 2010 and highest in 2015
- It was the warmest ever christmas on 2015. We have seen a correlation of warm weather and crimes.

Source:

[<http://abc7ny.com/weather/surfers-hit-beach-in-queens-as-nyc-has-warmest-christmas-ever/1134498/>]

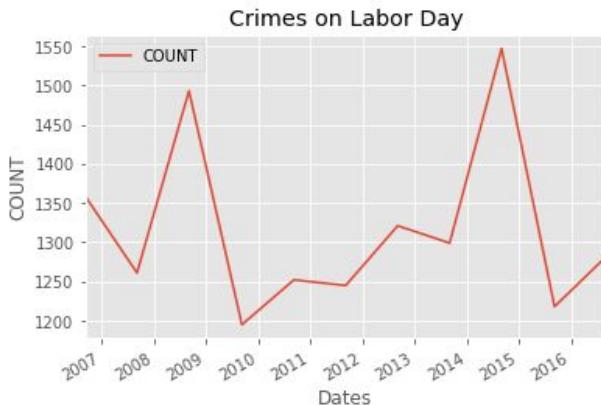
- On 2010, there was a blizzard in NYC. Therefore the crimes would have been low.
Source:[<http://www.nydailynews.com/new-york/boxing-day-blizzard-great-snowstorm-2010-gallery-1.1506?pmSlide=1.53023>]

- **Memorial Day**



- Crime on memorial day has been steadily lowering
- This maybe due to heavy nypd presence

- **Labour Day**



- Crimes were at an all time low in 2009 and highest in 2014
- This could be due to the the Carnival in Brooklyn. Each year on Labor Day, some 250,000 people gather for J'ouvert, according to city officials, and 1.5 million to 2 million people come out for the carnival.
- The west indian day parade happens in crown heights, brooklyn now. It is generally marred with violence.

Source:

[https://en.wikipedia.org/wiki/Labor_Day_Carnival]

- In 2014, the carnival had a lot of incidents.

Source:

[<http://www.nydailynews.com/new-york/media-bashes-carnival-article-1.1938911>]

- In 2015, before the carnival started, there were several shootings and stabbings and the carnival occurred under heavy surveillance. This could attribute to the low crime in 2015.

Source:

[<http://abc7ny.com/news/1-dead-4-wounded-in-jouvert-festival-violence-in-brooklyn/972666/>]

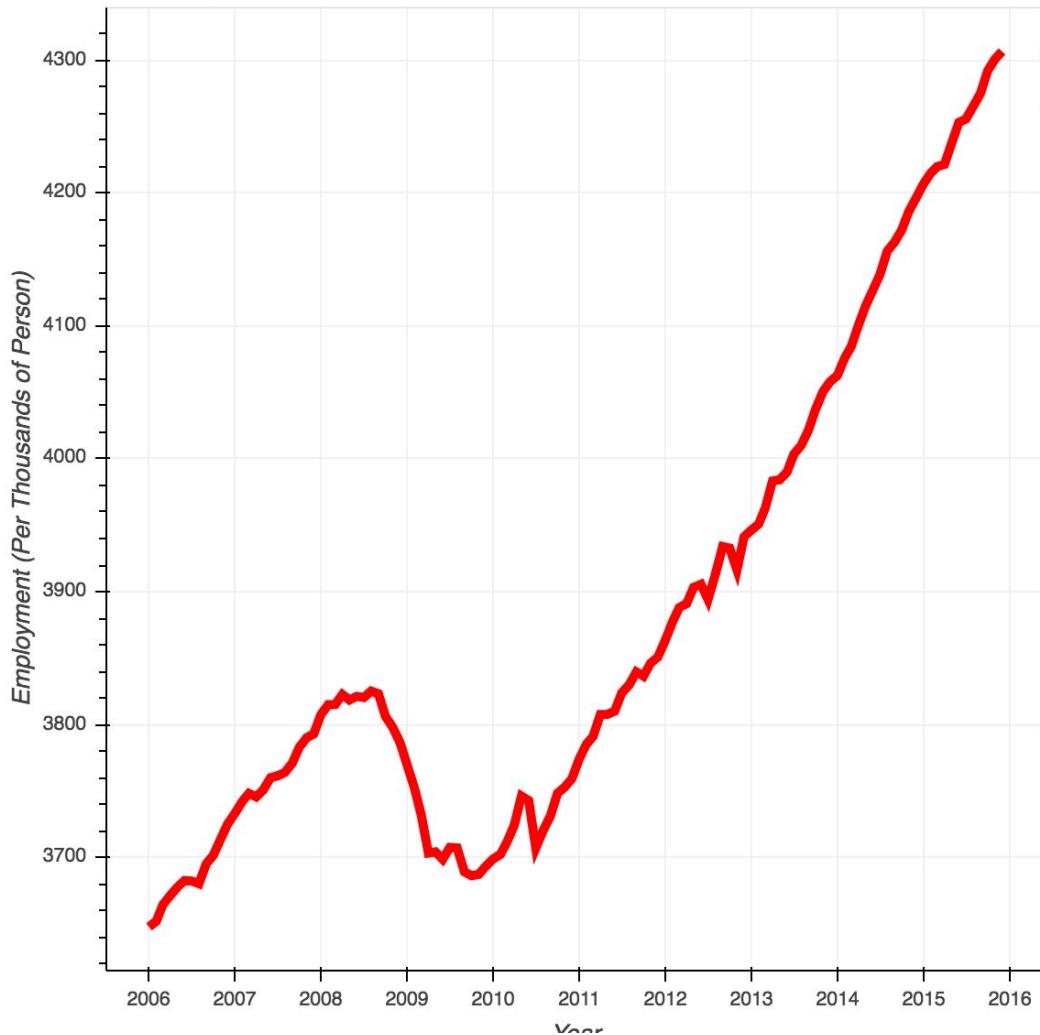
- Crowds were significantly low in 2009

Source:

[<http://www.nydailynews.com/new-york/unions-march-labor-day-parade-crowd-stay-rainy-day-article-1.407331>]

Employment

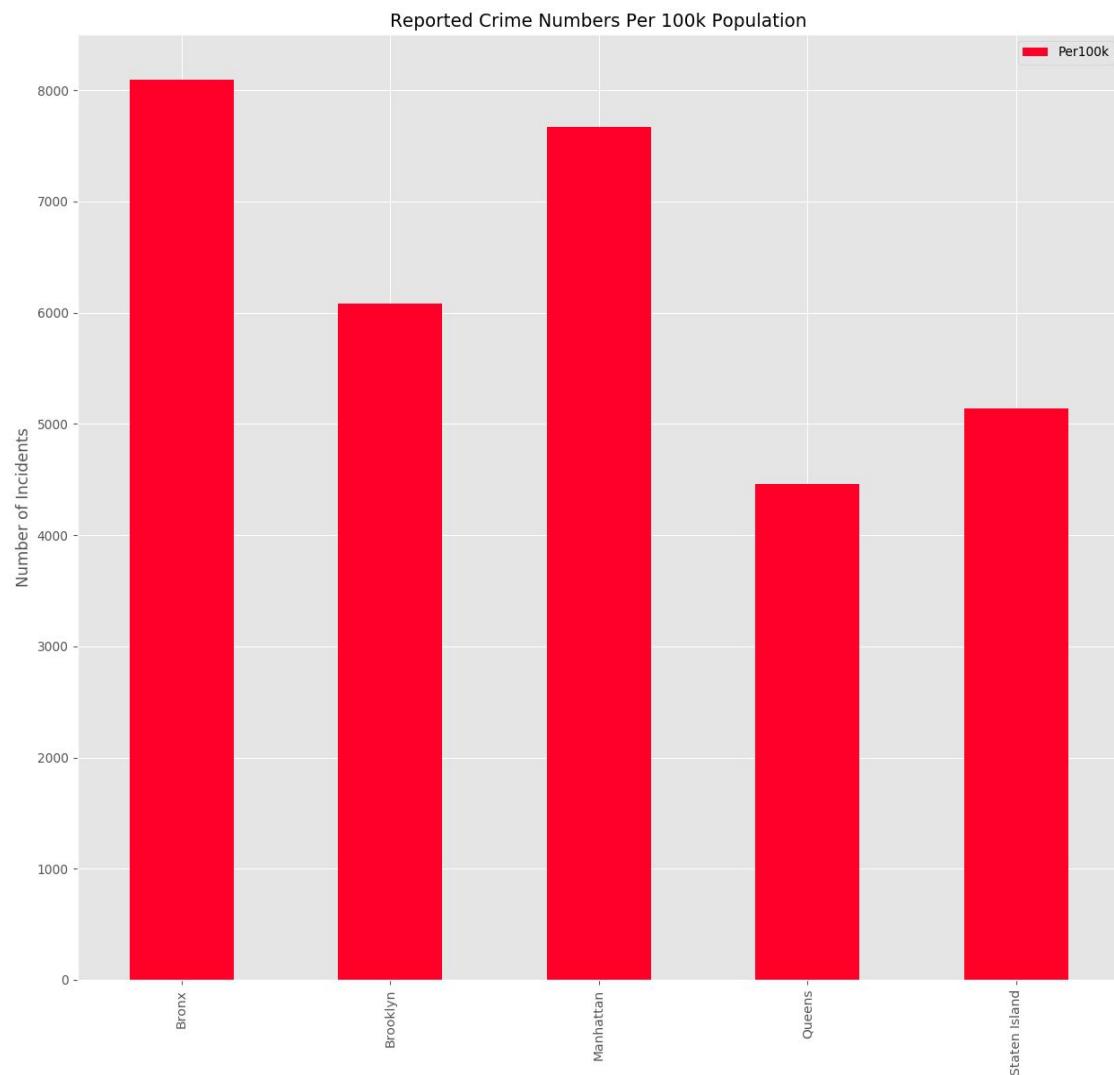
- While observing the graphs for crime rate in New York City, saw a decrease in total criminal offenses after 2011. This behavior could be noted because of following analogies we gathered:
 - There was steady increment in the employment rate after year 2010. Data(nyc-employment.csv) was collected from the **Economic Research Department of the Federal Reserve** of St.Louis and processed in nyc-employment.ipynb.

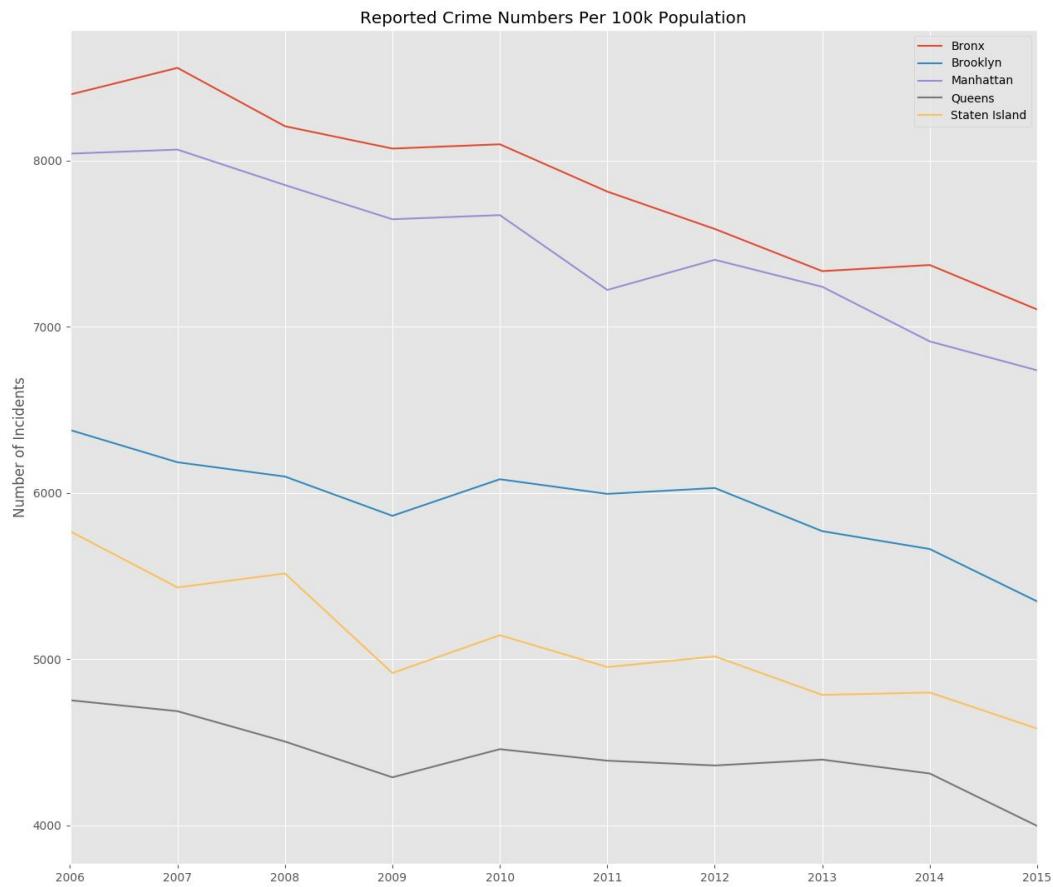


Population

- In further exploration we tried to establish the relationship between *crime rate* and population size and could observe that there was decrease in population between year 2009-2010 and increase in population size after year 2010. We observed in the below graphs crime did not

increase, with increase in population. Estimate of reported crime numbers for each borough.

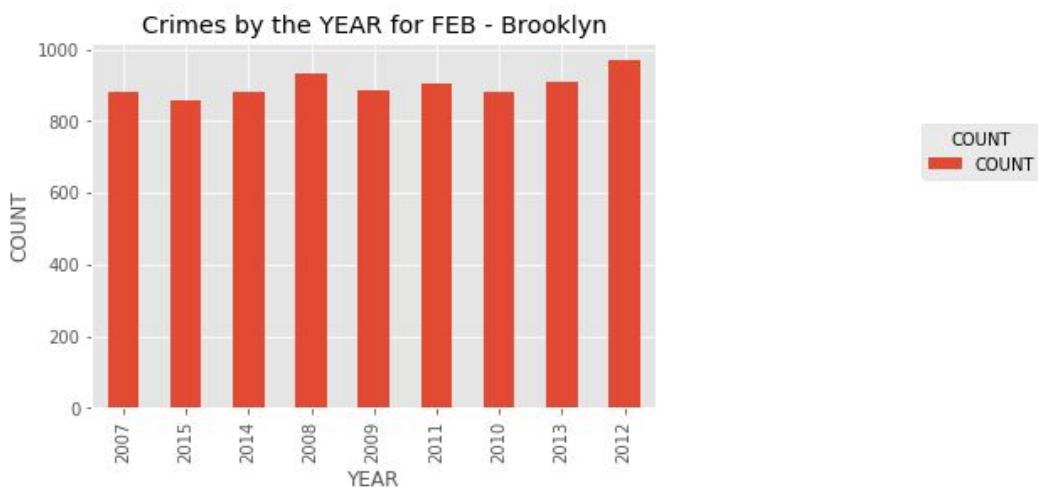
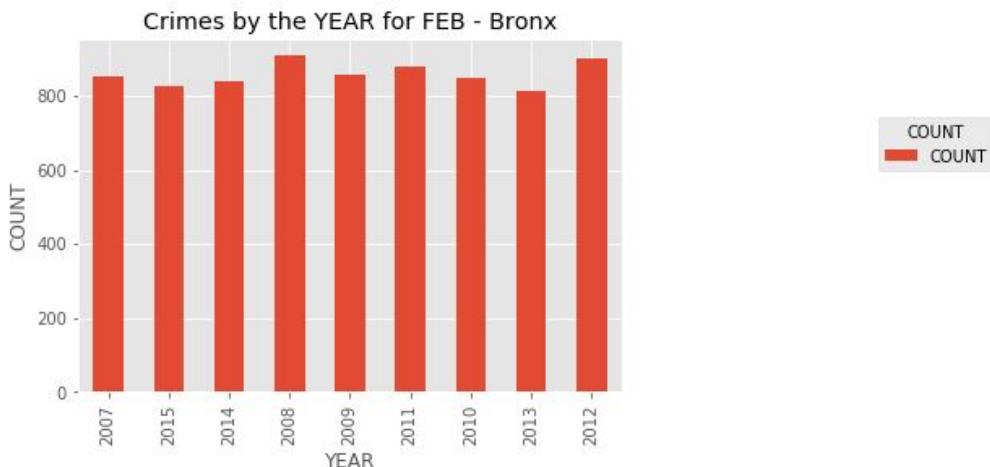




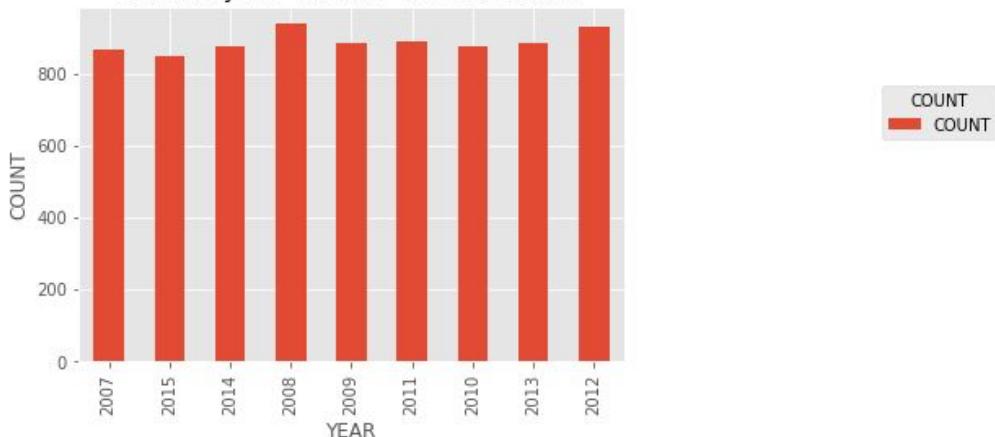
- In the graph we could observe crime rate decreased after year 2010, as the population increase. There could be many factors to mitigate the criminal offenses during the period like strict police patrolling, law and order of maintenance, decrease in unemployment rate as discussed above.
- This dataset was retrieved from (www.health.ny.gov) and except for year 2010 the entire data was an estimated dataset for census provided by U.S. Census Bureau.

Weather

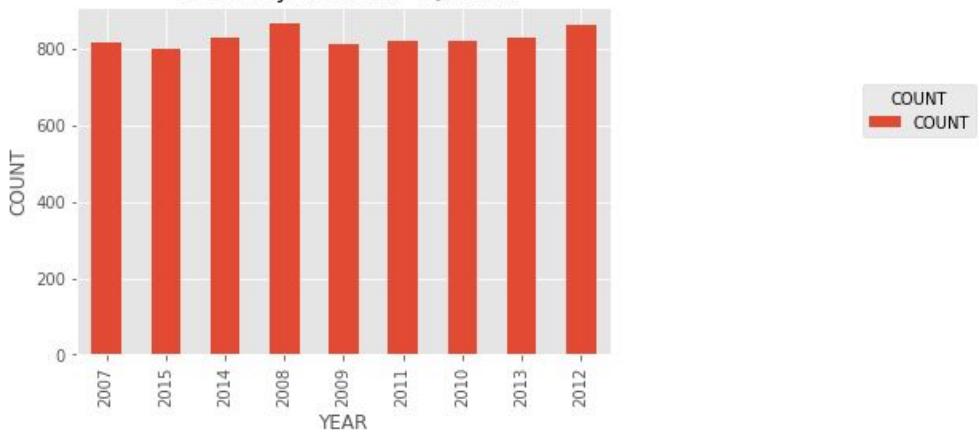
- We collected the dataset for the weather from NOAA's National Centers for Environmental Information and observed that maximum crime rate was reported in summer for past 10 years and least crime was reported in winter(specially February month). We assumed that extreme weather conditions like snow, winter storm can be one of the reasons forced people to stay at home. Due to less number of people stepped out in winter there was less number of crimes reported for winter season.



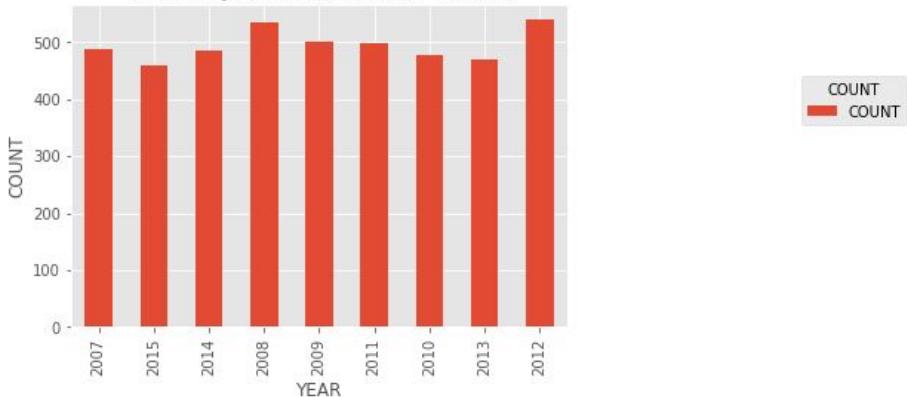
Crimes by the YEAR for FEB- Manhattan



Crimes by the YEAR - Queens



Crimes by the YEAR for FEB - Staten



- In above graphs we can estimate that for each borough crime rate significantly low during February month, thus supporting our hypothesis.

Summary

- There is steady decrease in crime rate in New York city after 2010, with effect of strict and better law policies.
- Maximum crime offense like assault, larceny, mischief were evident in each of the boroughs.
- Count wise Brooklyn contributed the most amount of crime.
- February month is reported with least number of crime.
- Summer months reported the most amount of crimes.
- We could see similar trend for every month that is maximum number of crimes were reported on 1st and again peaks up on 15th. Last day has least number of crime reported.
- Time hour for most of crimes was maximum around 12 pm that is lunch hour and minimum during midnight hours.
- Crimes are reported the most on Mondays which maybe because it is the first working day of the week.
- The park with the most amount of Crime was Central Park at the same time, Central Park PD had the least amount of crime compared to other PD.
- 75th Precinct reported the most amount of crime.
- Housing Development wise we found Castle Hill had reported the most amount of crimes.
- Year 2007-2008 was reported with maximum number of crimes. Mostly due to financial crisis. Due to economic crisis incidence of robbery may double, and homicide and motor vehicle theft also increase, according to the UNODC report.
- We also plotted the maps identifying the areas with maximum crime activities in each borough and came up with coordinates of the locations, thus identifying safe neighbourhood areas.
- There was steady decrease in unemployment rate after year 2010, thus contributing as one of the reason for decrease in crime activities during those years.
- We also estimated the correlation between increase in population and decrease in crime activities during those years. Although the population data used by us was an estimate provided by the US bureau.
- For the weather there was stark difference in crime activities happened in winter than summer, quite lower in number. Storms, blizzard, snowfall and extreme weather conditions could be the possible reason for least crime rate during that season.
- We observed the significant days like halloween parade and Thanksgiving day but to our surprise there was no significant crime rate, may be due to chilly weather or also because of strict police patrolling.

Technologies Used

- We cleaned the NYU complaint historic data and then parsed the useful columns for further analysis like Complaint date, Precinct and Have Development, key code and offense description.
- Crime (in the form of reported incidents) was analyzed in total, borough wise and then top crimes for all boroughs were extracted and their variation along with the variation in total reported incidents was analyzed.
- For further exploration we tried collaborating with other datasets like climate, employment and population.
- Further analysis were compared with datasets from **NOAA's National Centers for Environmental Information for Weather Data, NYS Department of Health for population estimates and Economics Research Division of the Federal Reserve Bank of St. Louis for employment statistics.**

Tools Used

- Pyspark (Apache Spark) for parsing and cleaning large datasets into smaller ones that could be used with pandas.
- Jupyter Notebooks for neat and clean data processing in Python.
- Pandas library for Data Analysis and exploration.
- Bokeh and Standard Matplotlib libraries for visualizations.

Individual Contribution

Niranjhana Nayar:

- Spark scripts for crime data analysis and data cleaning.
- Data visualization on Jupyter notebook.
- Maintained the report with observations.

Priyanka Bhadoriya:

- Spark scripts for crime data analysis and data cleaning.
- Data accumulation and exploration part.
- Maintained the report with observations.

Suchetha Siddagangappa:

- Spark scripts for crime data analysis and data cleaning.
- Maintained the report with observations.

References

To read about the map plots and folium:

<https://blog.dominodatalab.com/creating-interactive-crime-maps-with-folium/>

To read about the mapploit and pandas:

<http://pandas.pydata.org/pandas-docs/version/0.15.0/visualization.html>

<https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html>

https://matplotlib.org/api/_as_gen/matplotlib.pyplot.plot.html

To infer about employment rate and data:

<http://www.deptofnumbers.com/unemployment/new-york/>

<https://fred.stlouisfed.org/series/NYUR>

Article To analyse crime data and correlation with financial crisis for 2007-2008:

<http://www.unodc.org/unodc/en/frontpage/2012/February/economic-crisis-can-trigger-rise-in-crime.html>

Data for last 5 years about probable by weekly pay cheque dates:

<http://schools.nyc.gov/Offices/EnterpriseOperations/ChiefFinancialOfficer/DFO/PayrollAdministration/KeyDocuments/PayrollCalendars.htm>