STATGR5702 Final Project—Crime in NYC

Brent Daniel, Richard Lavery, Anita Pinto, Rashmi Rajaguru, Jingbo Wu

1. Introduction

The topic is choosen to answer some interesting questions:

```
1.  Where are the crimes mostly happening?

2.  When do they happen? Varying with month, with day?

3.  What is the frequency distribution of different crimes?
```

2. Data Description

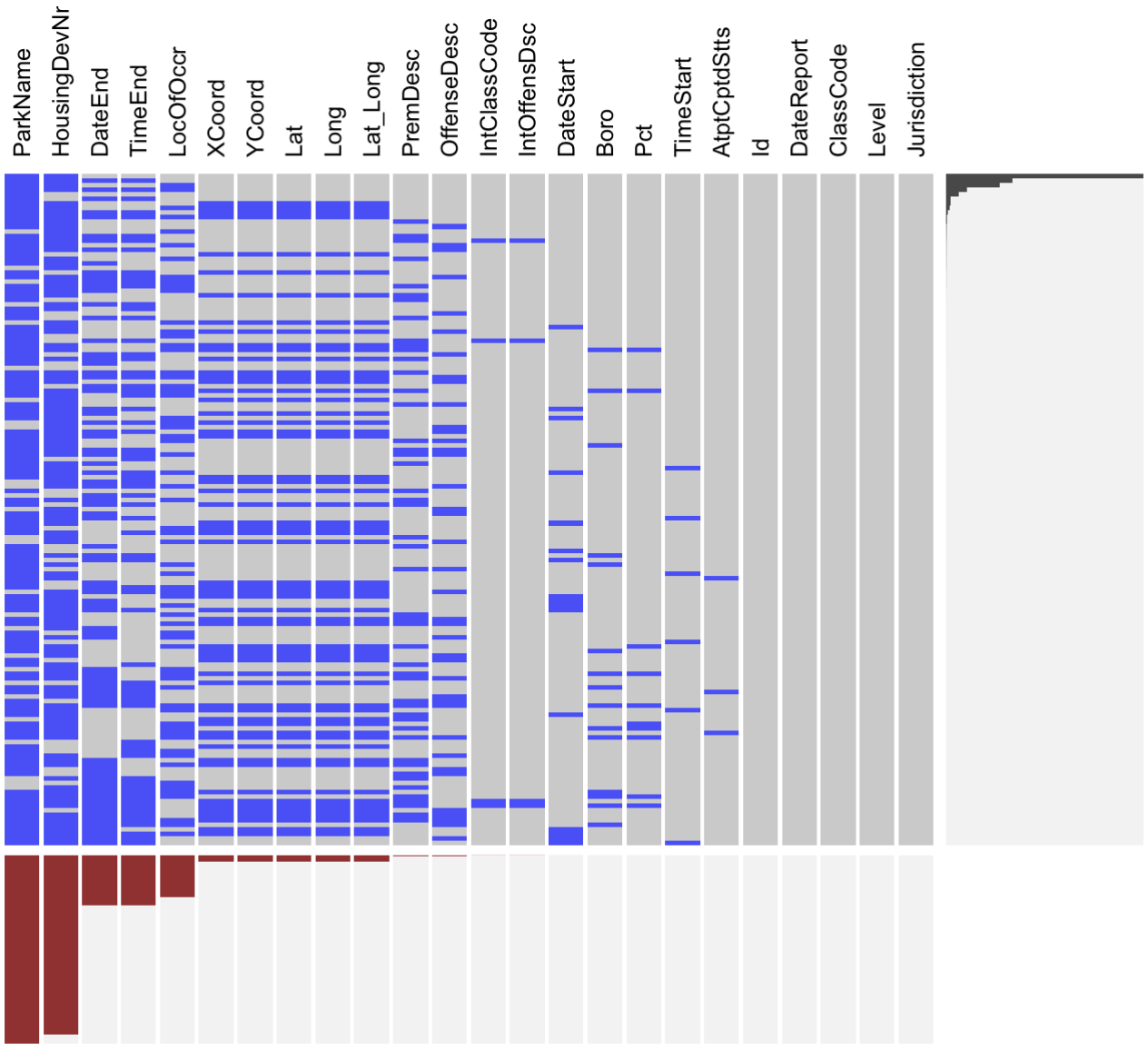
Source of the data:

- https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i (https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i)
- API access via SODA API: https://data.cityofnewyork.us/resource/9s4h-37hy.csv (https://data.cityofnewyork.us/resource/9s4h-37hy.csv)
- This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year (2016). The dataset is created on Novermber 2, 2016. The update frequency is annually.
- The dataset is provided by NYPD and owned by NYC OpenData. It has 5.58 Million rows and 24 variables. Three types of variables included are number, text, and location. The 24 variables are listed below.
- Only valid complaints are included in this release. Information is accurate as of the date it was queried from the system of the record, but should be considered a close approximation of the current records, due to complaint revisions and updates. (NYPDIncidentLevelDataFootnotes.pdf)

3. Analysis of Data Quality

```
##
Read 0.0% of 5580035 rows
Read 10.2% of 5580035 rows
Read 20.3% of 5580035 rows
Read 30.8% of 5580035 rows
Read 41.9% of 5580035 rows
Read 52.9% of 5580035 rows
Read 63.6% of 5580035 rows
Read 74.0% of 5580035 rows
Read 84.6% of 5580035 rows
Read 95.3% of 5580035 rows
Read 5580035 rows and 24 (of 24) columns from 1.329 GB file in 00:00:17
```

-Missing Pattern



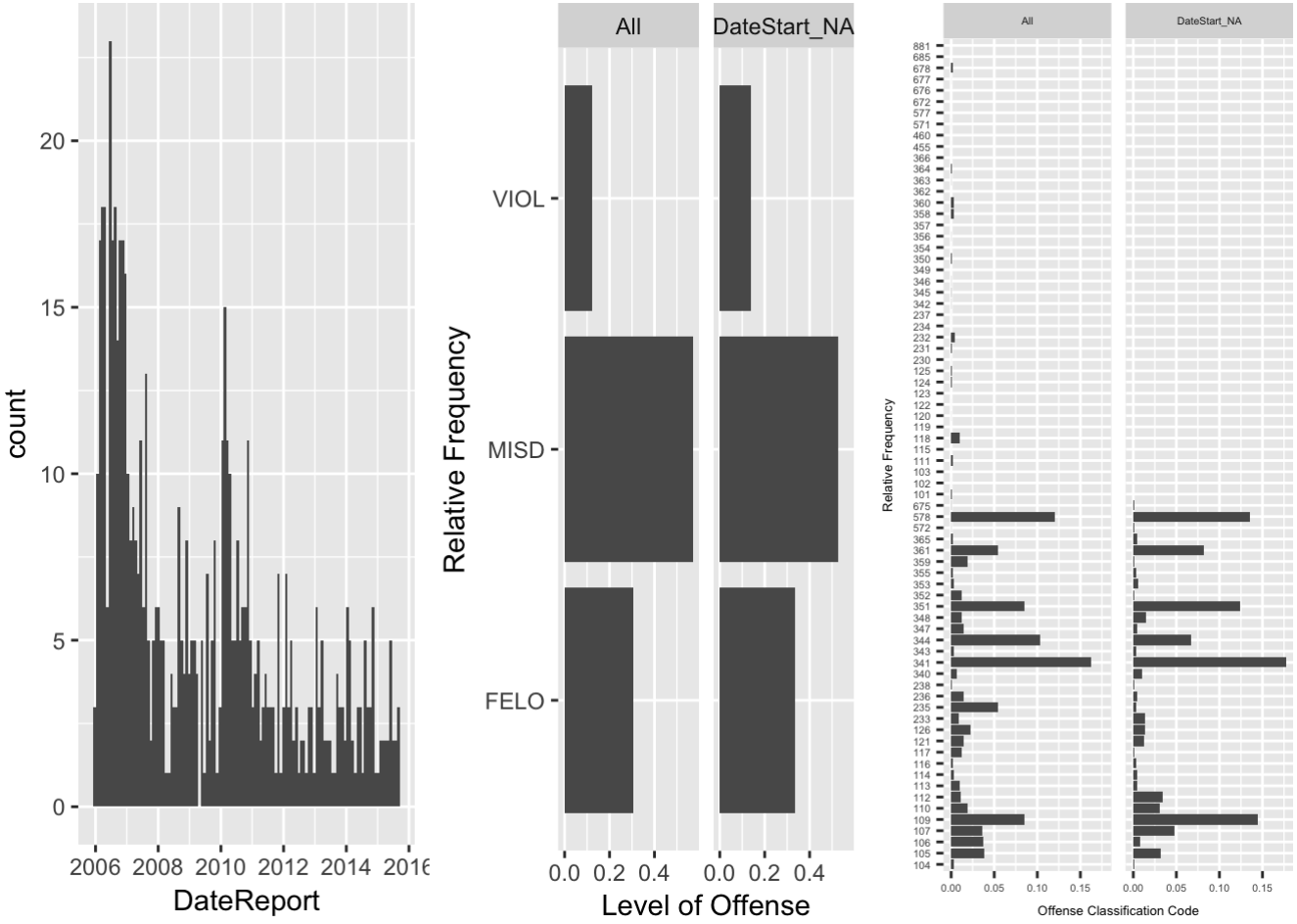
- This dataset has 24 variables and ~5.6 Million rows of complaints/events. It includes all cases with reporting date (DateReport) ranges from 2006-01-01 to 2016-12-31. In this section, we investigate the missing patterns and possible errorness of variables that are important to the understanting of the crime's when, where and what.The overall missing patterns are shown above.

- 5 variables has data all valid. They are complaint number (ID), report date (DateReport), 3 digit offense classification code (ClassCode), level of offense (Level), jurisdiction responsible for incident (JurisDiction).
- AtptCptdStatus is an indicator of whether crime attemped or complelted. Only 7 missing cases; 5483869 coded as completed, and 96159 cases indicated as attempted.
- ParkName are given if the event occurred there. Most of the cases doesn't have this vairable mostly becasue it doesn't apply. How much percent of real missing of park place, we don't know.

-Missing DateStart

There are total of 655 complaints missing DateStart, of which,

1. When looking at the RPT_DT (reporting date) although they look slightly clusted at the beginning around 2006 and less at the ending around 2016, the reporting dates still look pretty even over the period suggesting randomness of the missing against DateReport.
2. The frequency distrinution of Level shares the same pattern of that from all data.
3. The frequency distrinution of ClassCode shares the same pattern of that from all data.



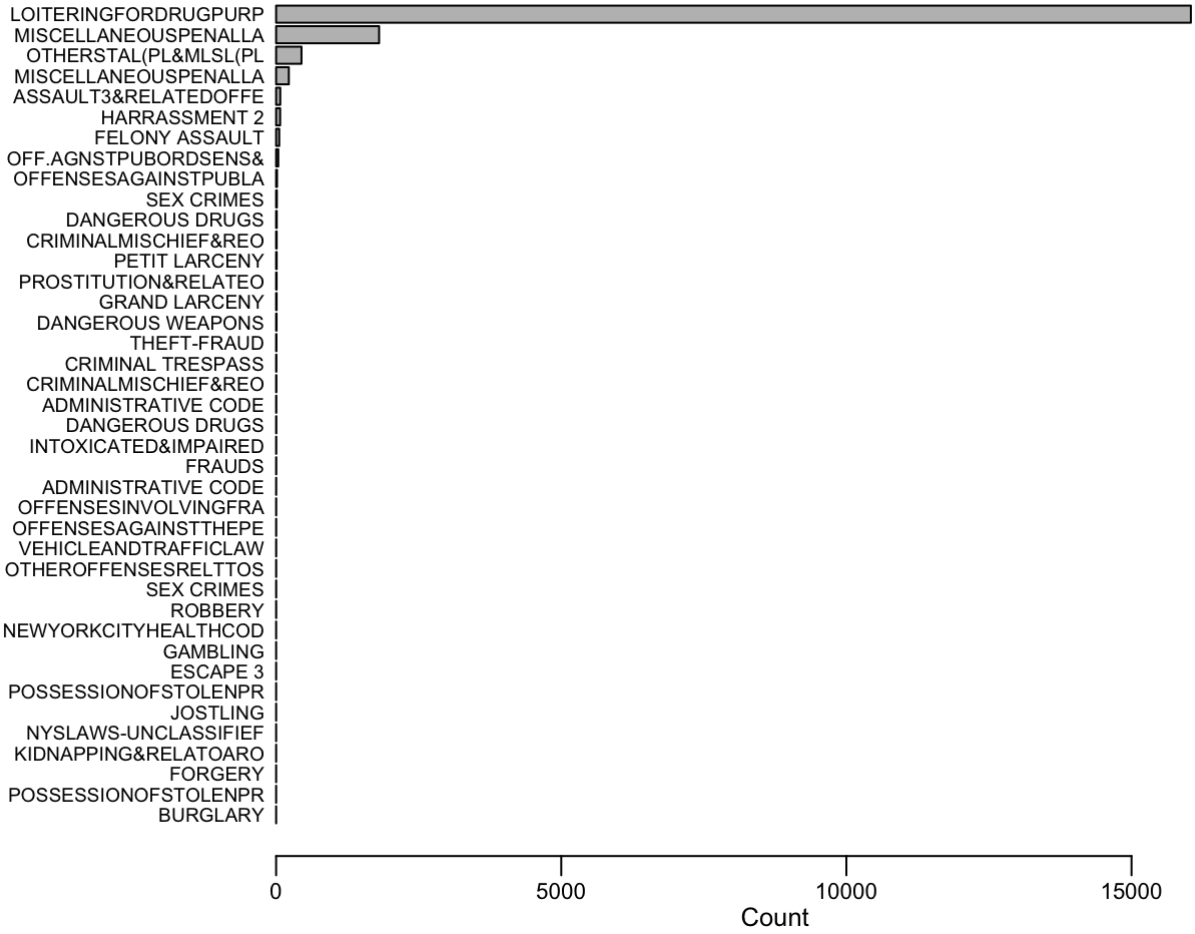
-Mismatch between Pct and Boro

```
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1     6 BRONX         1
## 2     6 MANHATTAN 59559
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1     7 BROOKLYN      1
## 2     7 MANHATTAN 45259
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1     9 BROOKLYN      1
## 2     9 MANHATTAN 67822
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1    13 BROOKLYN      1
## 2    13 MANHATTAN 81145
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro       count
##   <int> <fct>      <int>
## 1    14 BROOKLYN       1
## 2    14 MANHATTAN 129697
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1    23 BRONX         3
## 2    23 MANHATTAN 73154
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1    25 BRONX         1
## 2    25 MANHATTAN 74073
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1    26 BROOKLYN      1
## 2    26 MANHATTAN 37213
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro     count
##   <int> <fct>    <int>
## 1    71 BRONX        1
```

```
## 2    71 BROOKLYN 78909
## # A tibble: 3 x 3
## # Groups:   Pct [1]
##     Pct Boro      count
##   <int> <fct>     <int>
## 1   104 BROOKLYN      1
## 2   104 MANHATTAN     1
## 3   104 QUEENS    81151
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro     count
##   <int> <fct>    <int>
## 1   106 BROOKLYN     1
## 2   106 QUEENS   67367
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro    count
##   <int> <fct>   <int>
## 1   114 BRONX       2
## 2   114 QUEENS 100798
## # A tibble: 2 x 3
## # Groups:   Pct [1]
##     Pct Boro          count
##   <int> <fct>         <int>
## 1   121 BROOKLYN          1
## 2   121 STATEN ISLAND 23804
```

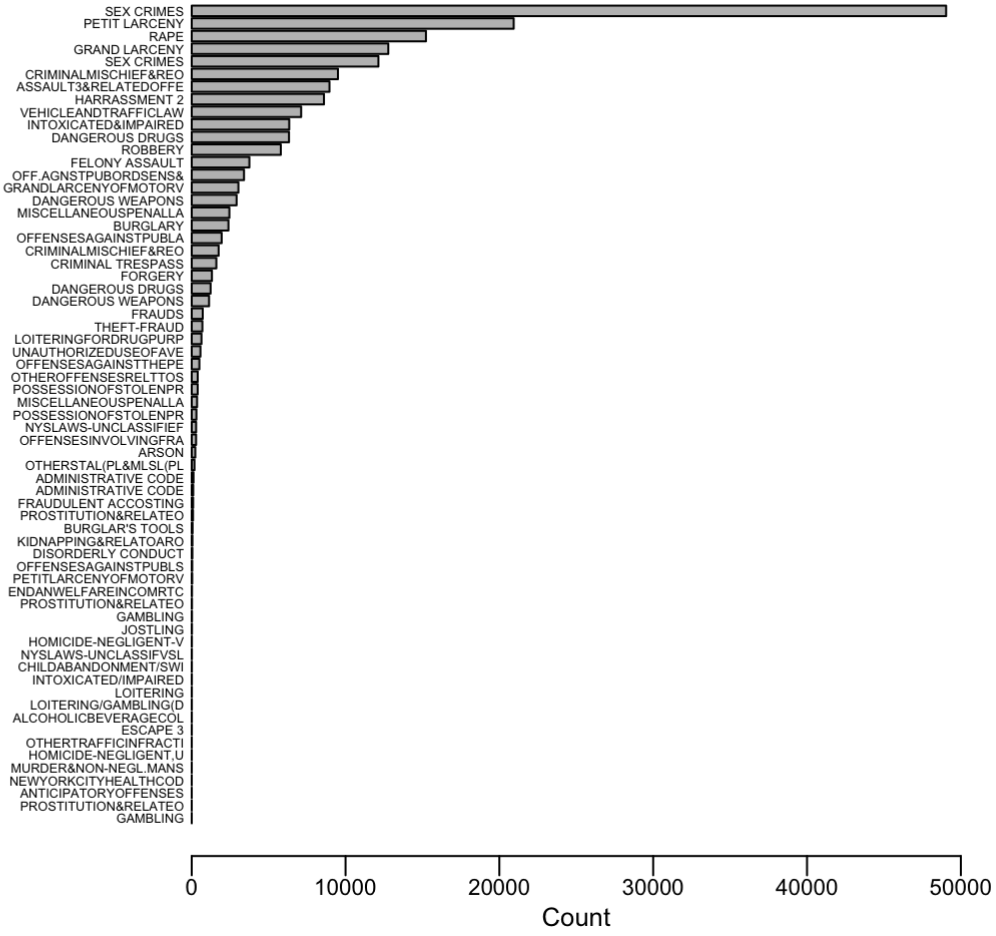
- There are total of 77 distinct precincts in NYC. A list above shows Pct with double/triple borough names.
- There are about 16 cases with precint number not consistent with the borough name. For those precints with double/triple borough names, the # of cases with a 2nd/3rd borough name are only 1-3. This problem can be fixed by correcting the borough names of those rare cases using a map of Pct vs Boro established from dataset.
- 390 cases with missing Pct. 463 cases with missing Boro. Comparing to the total # of cases over 5M, they won't affect the results.

-Missing OffenseDesc



- OffenseDesc (with missing values) is the description of offense corresponding with key code ClassCode which is complete in the dataset. Code and description map each other and valid OffenseDesc can be infered from a map established from the dataset.
- The plot above shows missing counts of the ClassCode categories with OffenseDesc originally missing but now retrieved from the map between ClassCode and OffenseDesc.

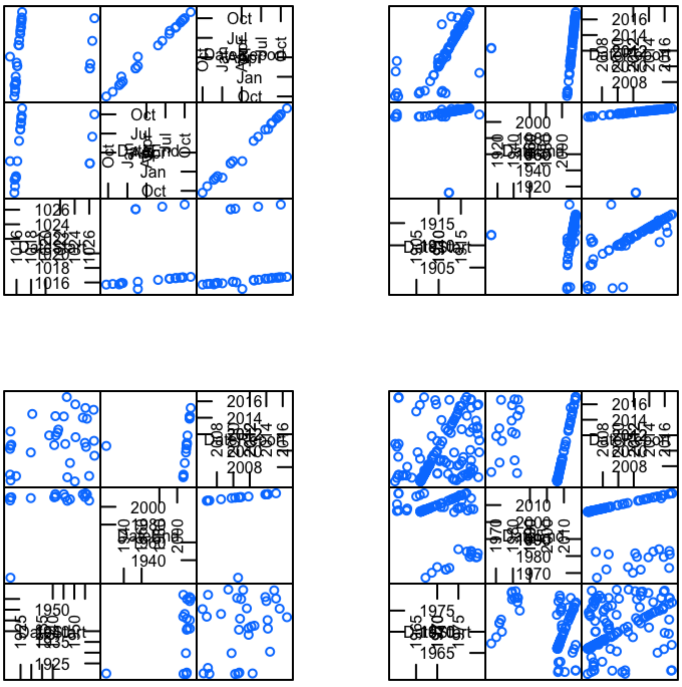
-Missing Geolocation



* The 5 geolocation variables have the same missing pattern as expected. So we only need to look at one of them to examine the missing. In the data document, it stated that “to protect victim identities, rape and sex crime offenses are not geocoded”. We want to see if the missing of geo variables are mostly related with those crime? Is there a lot of missing for other crimes too?

- The missing in geolocation is obviously not random. When examine the spatial pattern of the crimes, we have to bear in mind that particular crimes will not appear on the map due to missing not at random.

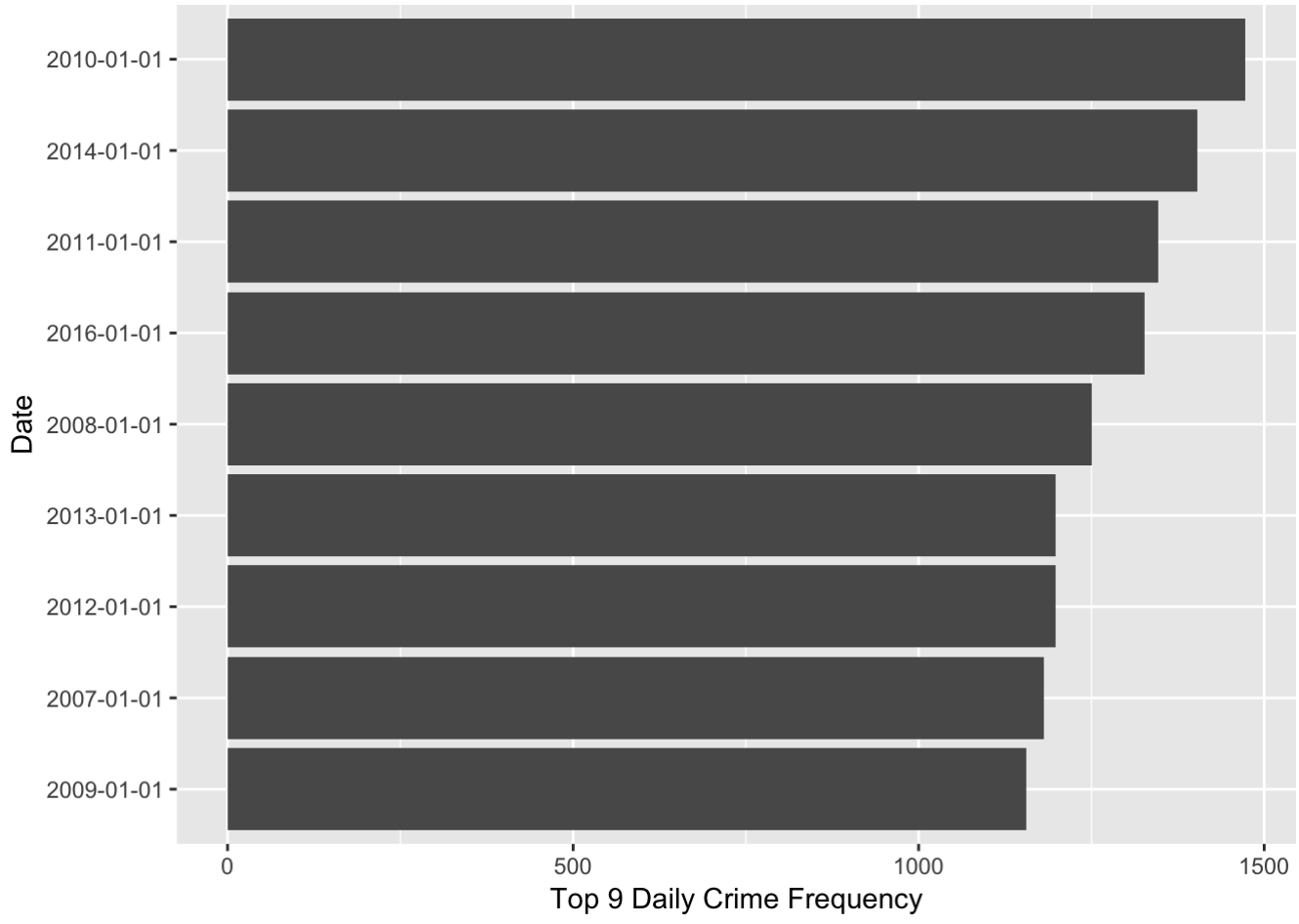
-Errors in DateStart



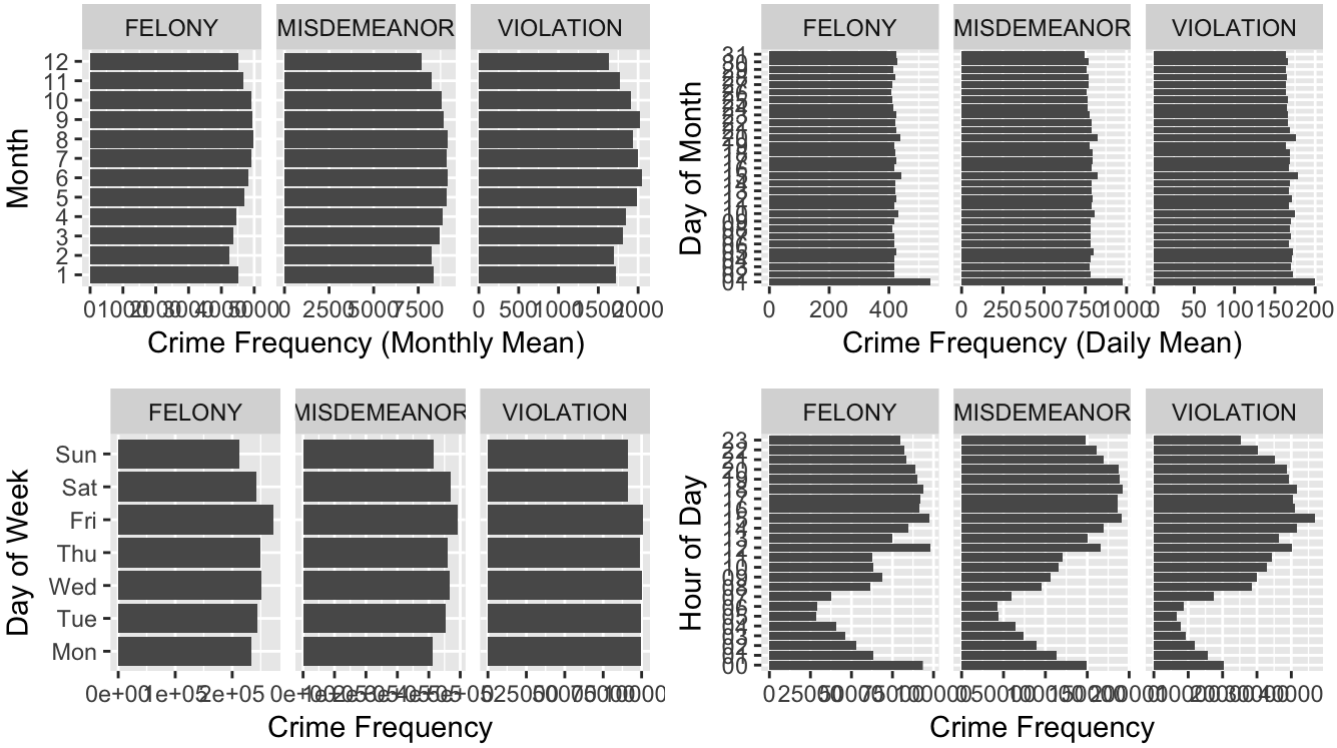
- There seems to be errors in DateStart. It dated back to Year 1015 which is suspicious. But by referncing to DateReport, 2 dates usually have very close month/date. It seems Year1015 may actually be Year2015 due to a typo. DateEnd also suggests so.
- The scatterplot of the DateStart vs DateReport did show some strict linear correlation for many cases during some periods.
- As shown above, the amount of such cases is not large. In our main analysis, we will focus on cases with Datestart after Jan. 1,2006 up until Dec. 31, 2016 in total over 5.5M.

4. Main Analysis


Daily Crime Frequency since 2006 with 30-day running mean
1500
1000
500
Frequency
2006
2008
2010
2012
2014
2016
Date
FELONY
MISDEMEANOR
VIOLATION


2010-01-01
2014-01-01
2011-01-01
2016-01-01
2008-01-01
2013-01-01
2012-01-01
2007-01-01
2009-01-01
Date
0
500
1000
1500
Top 9 Daily Crime Frequency

- The crime frequency is decreasing over the years this is because lots of cases occurred over the years haven't reported yet.
- There are obvious annual variation/cycle. 30-day running mean shows the cycle clearly.
- There are spikes in the misdemeanor category. The top 9 dates with high frequency are shown in the barchart. They are on January 1 on almost each year from 2006-2016 except 2015 which is actually very close behind. These cases seemed like mistakening assigned an occurrence date as January 1 since by examining the relationships between DateReport,DateStart and DateEnd, they don't seem make much sense comparing with others.

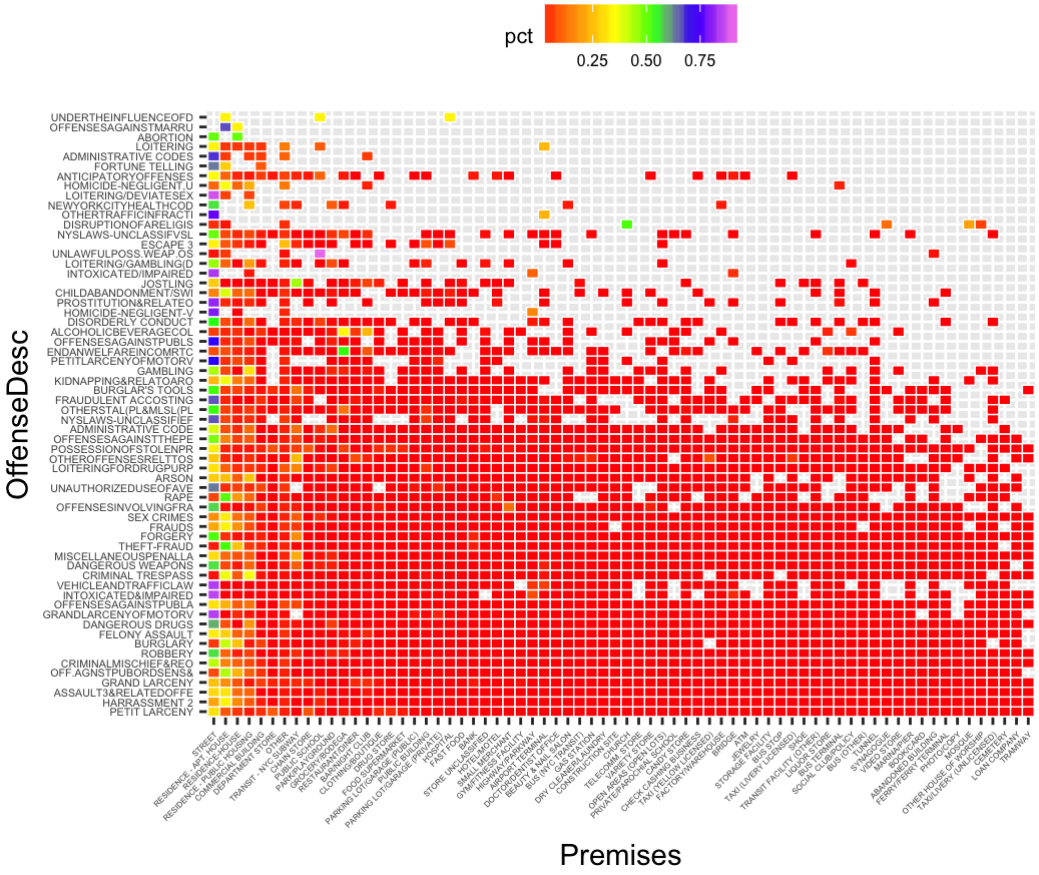


* Indeed by barcharting over the months,we see Jun.-Oct. is a high crime season. * The fake January increasing was due to the errors in the records.

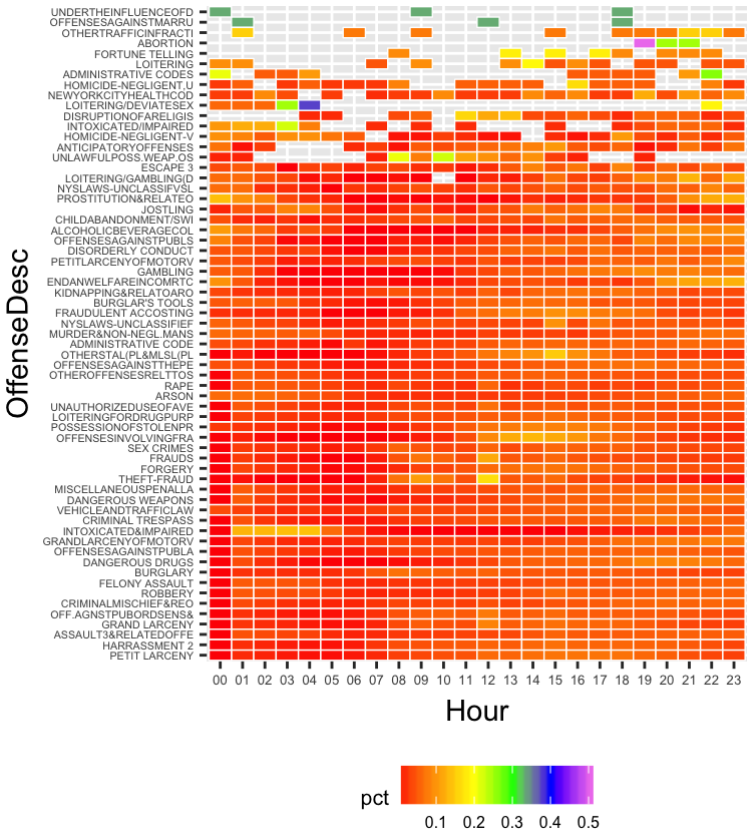
* The spike in Janurary is consistent with the analysis above. * There seemed having a tendency of rounding every 5 day.

* Violation is low during weekends but same during weekdays. * Felony and misdemeanor is high on Friday but low on Sunday nad Monday.

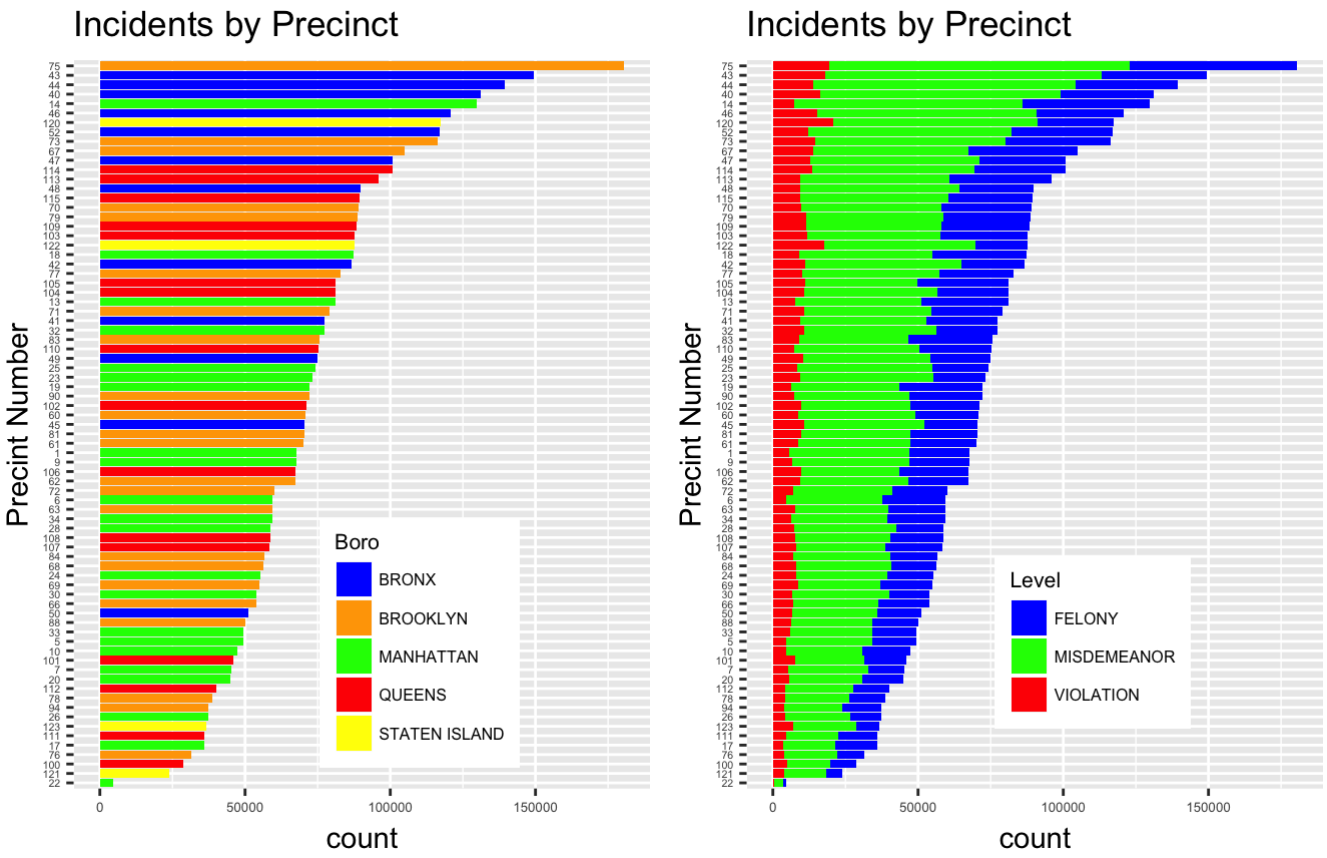
* There is obvious day cycle in the crime occurrence. Early morning has the least crime occurrence while later afternoon has the most crime occurrence.



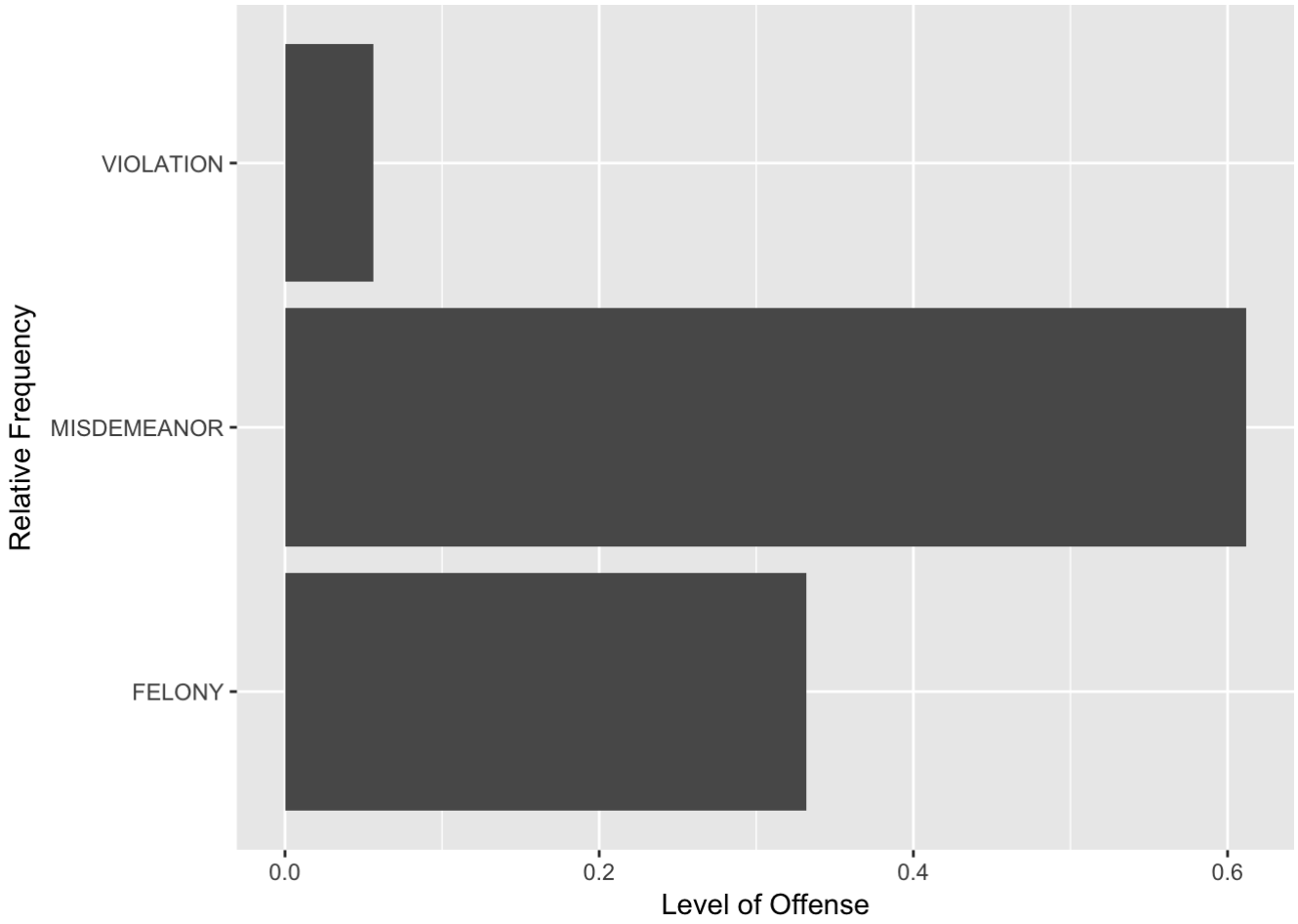
* Doesn't seem having association between crime types and premises.



* Do we see any association between time and certain crime? Do see some high density around middle up right area, which is consistent with the barcharting daily cycle.



- Crime frequency by Pct with either borough name colored or crime Level colored.



- ~12538 cases recorded as occurred in parks/playground or greenspaces. The crime level distribution share the same pattern as the overall data.