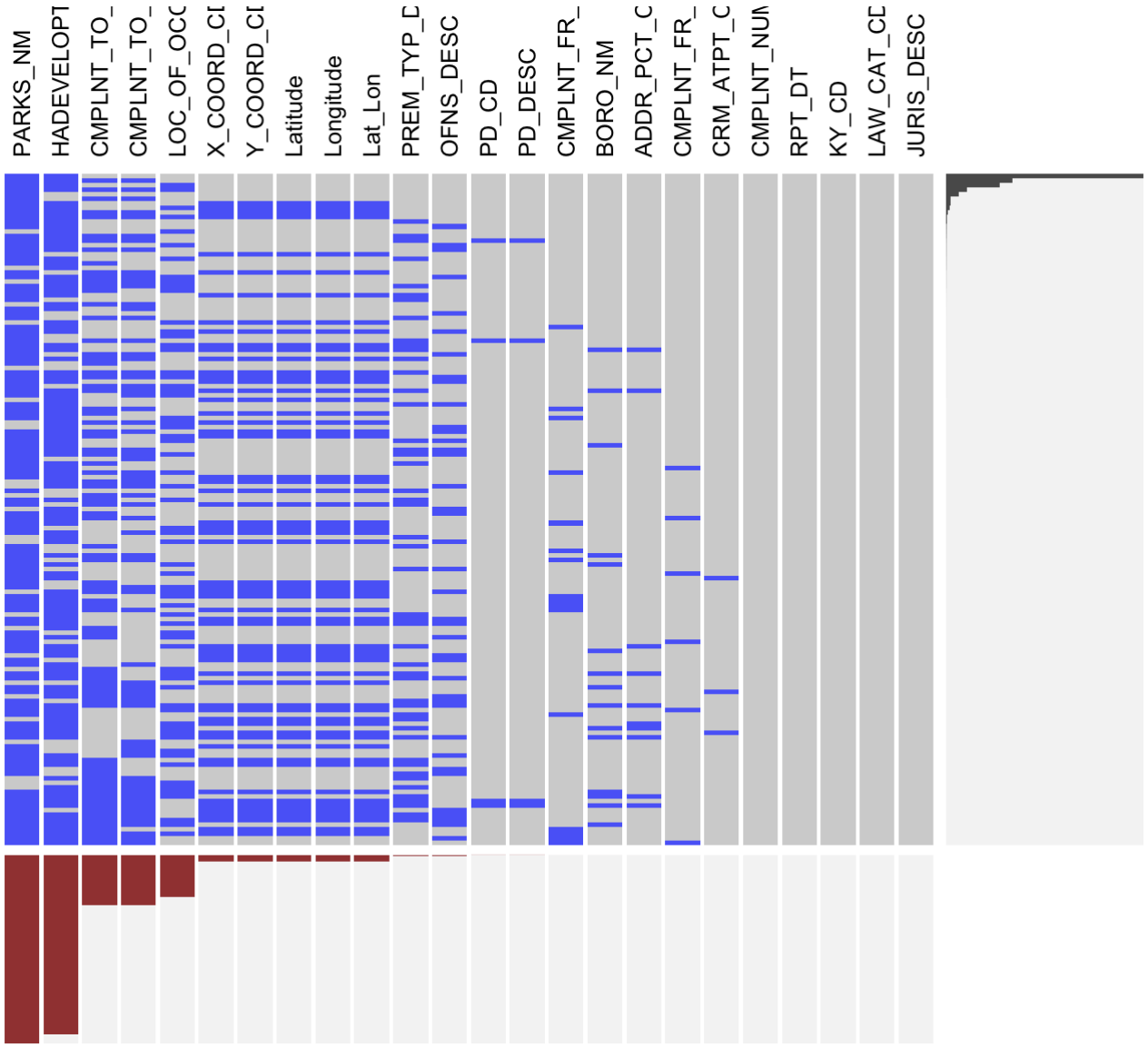
=====Part 3===== Data Quality

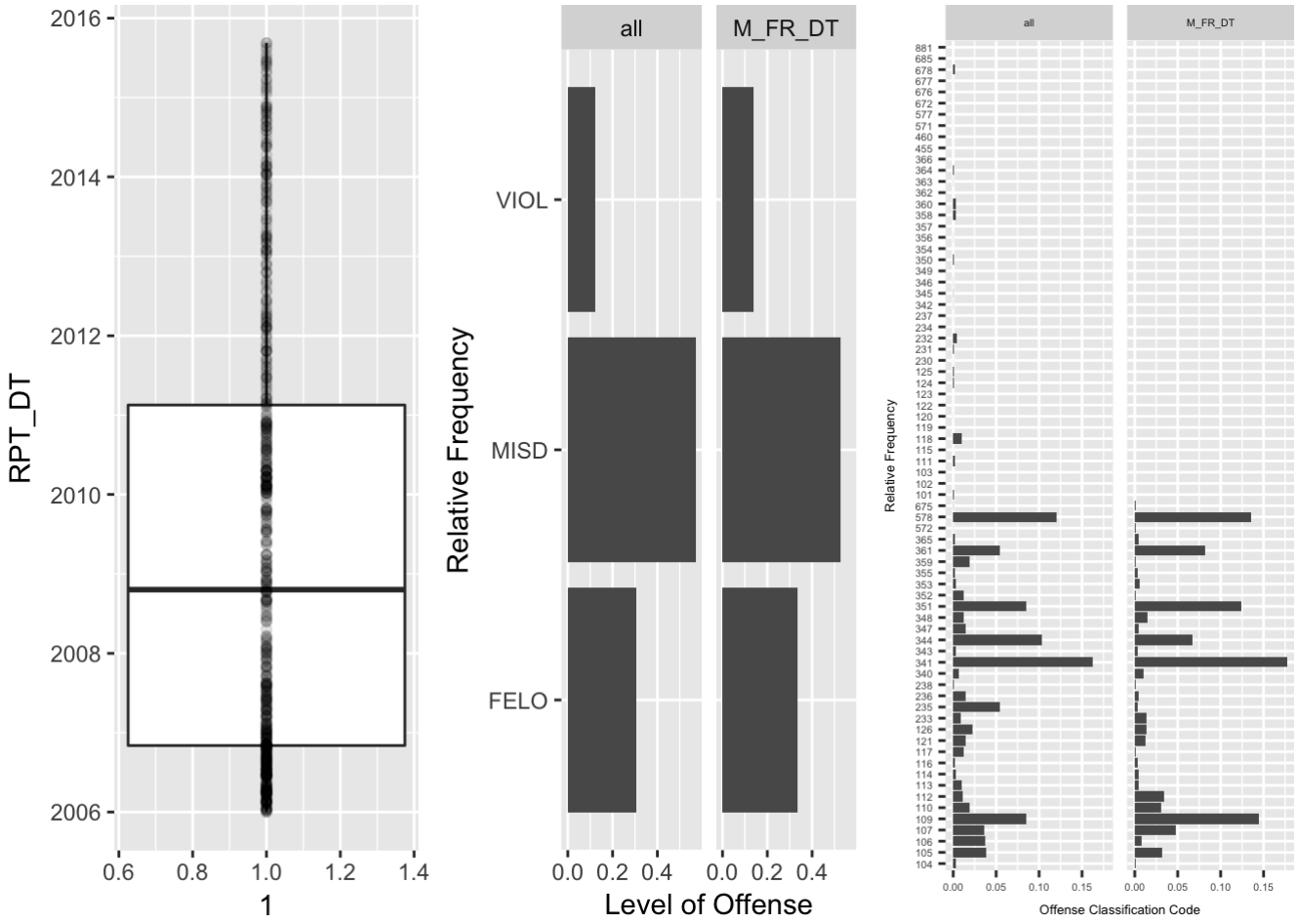
```
##
Read 0.0% of 5580035 rows
Read 10.2% of 5580035 rows
Read 20.1% of 5580035 rows
Read 29.9% of 5580035 rows
Read 39.8% of 5580035 rows
Read 50.7% of 5580035 rows
Read 61.8% of 5580035 rows
Read 72.0% of 5580035 rows
Read 81.9% of 5580035 rows
Read 92.3% of 5580035 rows
Read 5580035 rows and 24 (of 24) columns from 1.329 GB file in 00:00:16
```

===Missing/Error Data Analysis===

This dataset has 24 variables and ~5.6 Million rows of complaints/events. 5 variables has data all valid. They are complaint number (CMPLNT_NUM), report date (RPT_DT), 3 digit offense classification code (KY_CD), level of offense (LAW_CAT_CD), jurisdiction responsible for incident (JURIS_DESC). The variable RPT_DT (the case reporting time) ranges from 2006-01-01 to 2016-12-31. The overall missing patterns are shown below. In this section, we investigate the missing patterns and possible errorness of variables that important to the understanting of the crime's when, where and what.



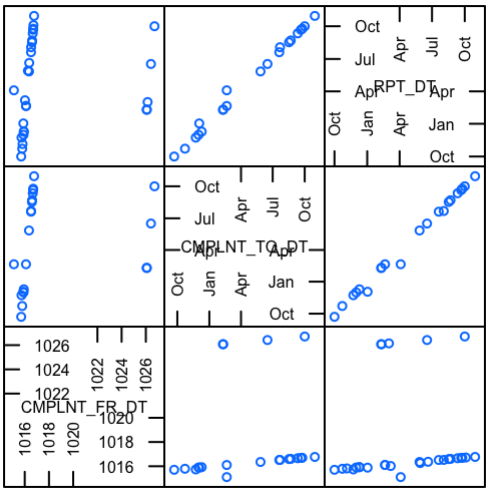
===Missing in CMPLNT_FR_DT===

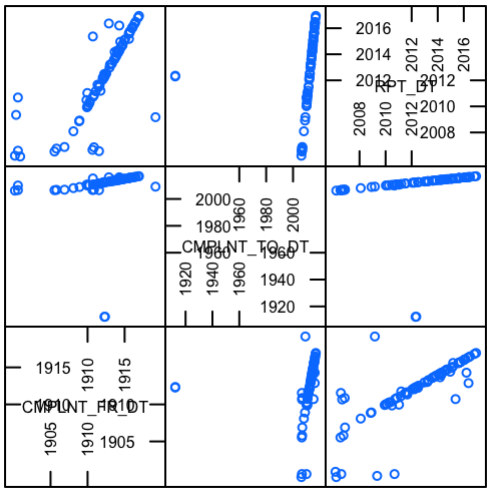


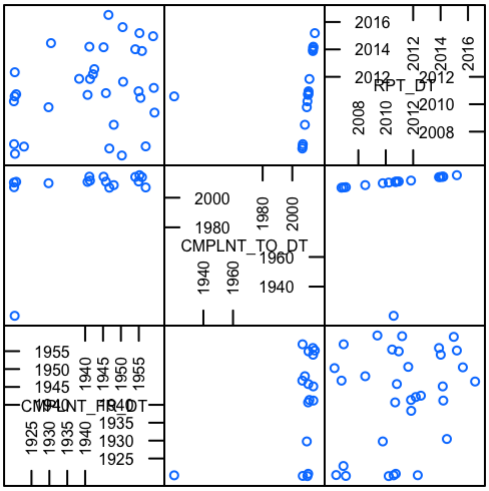
- There are total of 655 complaints missing CMPLNT_FR_DT, of which,

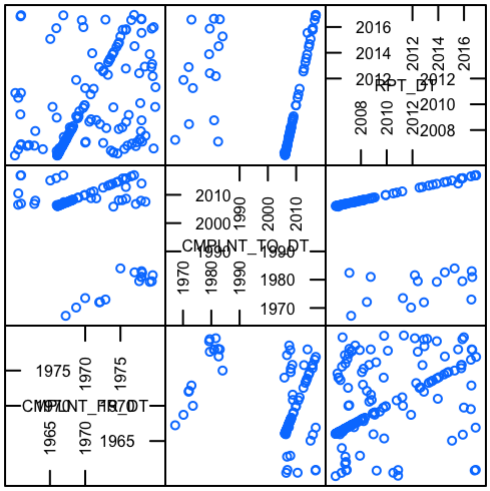
1. When looking at the RPT_DT (reporting date) although they look slightly clusted at the beginning around 2006 and less at the ending around 2016, the reporting dates still look pretty even over the period suggesting randomness of the missing against RPT_DT.
2. The frequency distrinution of LAW_CAT_CD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

   ===Errors in CMPLNT_FR_DT===



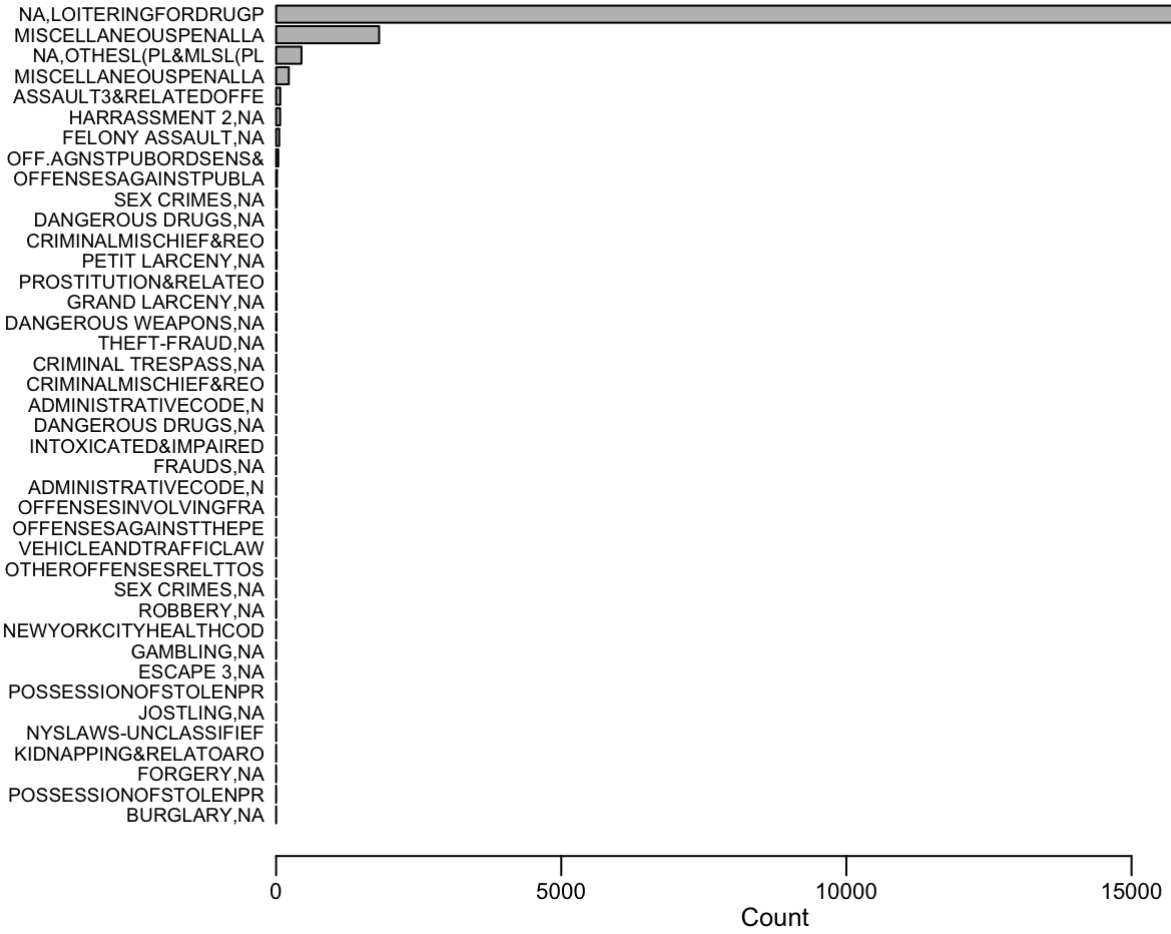






* There seems to be errors in CMPLNT_FR_DT. It dated back to Year 1015 which is suspicious. But by referncing to RPT_DT, 2 dates usually have very close month/date. It seems Year1015 may actually be Year2015 due to a typo. CMPLNT_TO_DT also suggest so. * The scatterplot of the CMPLNT_FR_DT vs RPT_DT did show some strict linear correlation for many cases during some periods.

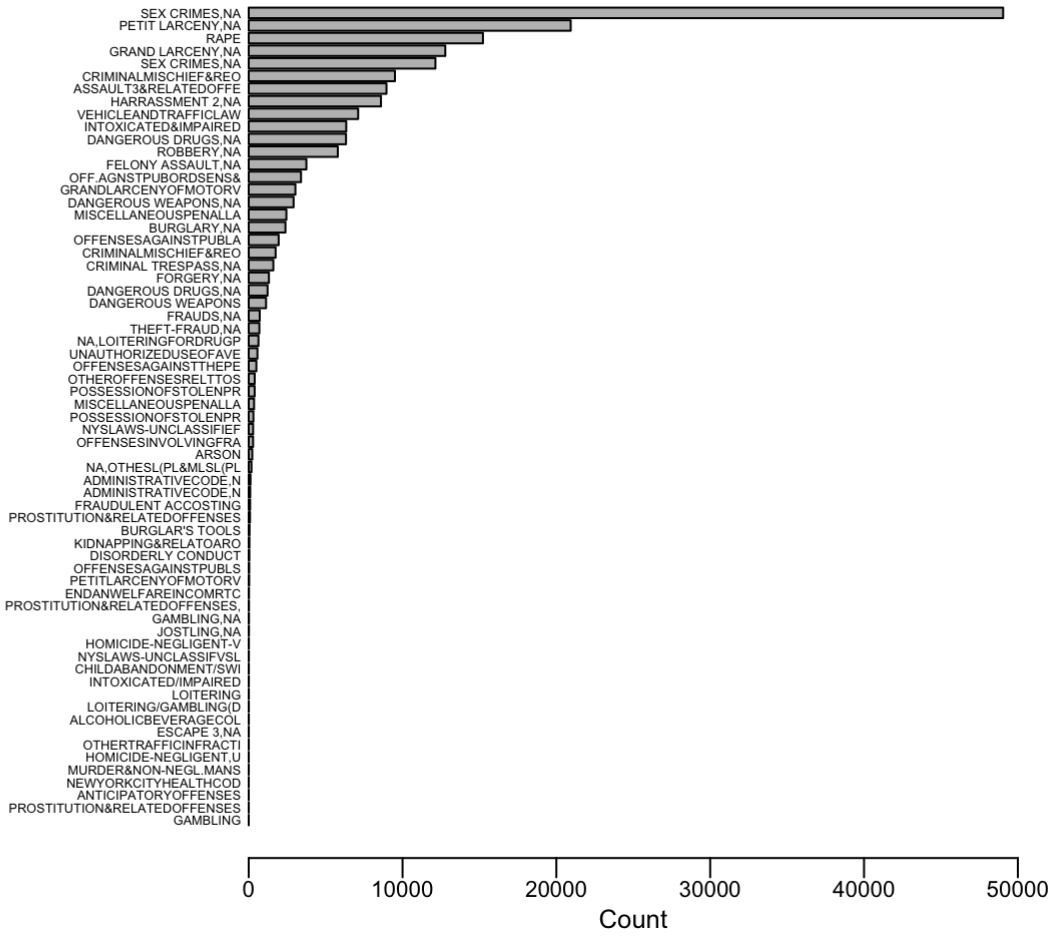
===missing OFNS_DESC===



* OFNS_DESC is the description of offense corresponding with key code KY_CD which is complete in the dataset. (Shouldn't description leads to a code? Why there is missing in description but code is available?) Some case has 2 description but 1 code. some cases have different code but same description. Code and description map each other and valid match can be infered from the dataset. So the missing description can be retrieved from the valid mapping.

- The plot below shows cases with missing OFNS_DESC grouped by KY_CD and then with OFNS_DESC retrieved back from KY_CD.

  ===Missing in 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 mising 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.

  ===Missing in CRM_ATPT_CPTD_CD===

- CRM_ATPT_CPTD_CD is an indicator of whether crime attemped or complelted. Only 7 missing cases; 5483869 coded as completed, and 96159 cases indicated as attempted.

  ===PREM_TYP_DESC===

- 70 levels of description of premises.

  ===PARKS_NM===

- Most of the cases doesn’t not have this vairable mostly becasue it doesn’t apply. How much percent of real missing of park place, we don’t know.

  ===HADEVELOPT===

- Don’t know what does this mean? It’s missing a lot too.

  ===BORO_NM===

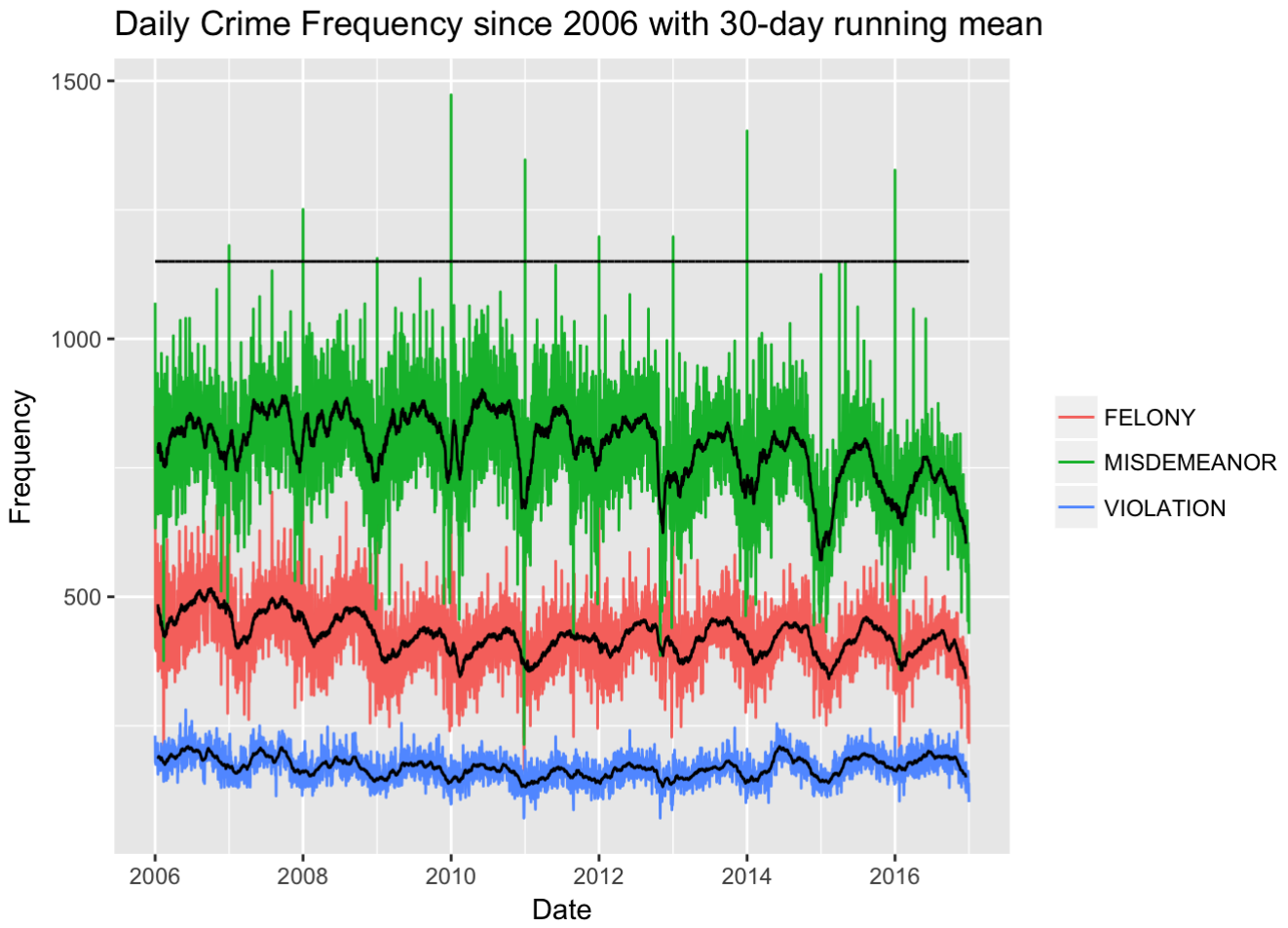
- 463 cases missing BORO_NM, of which 75 has valid location data and 388 doesn’t. Overall, 463 compare to 5M, ignorable.

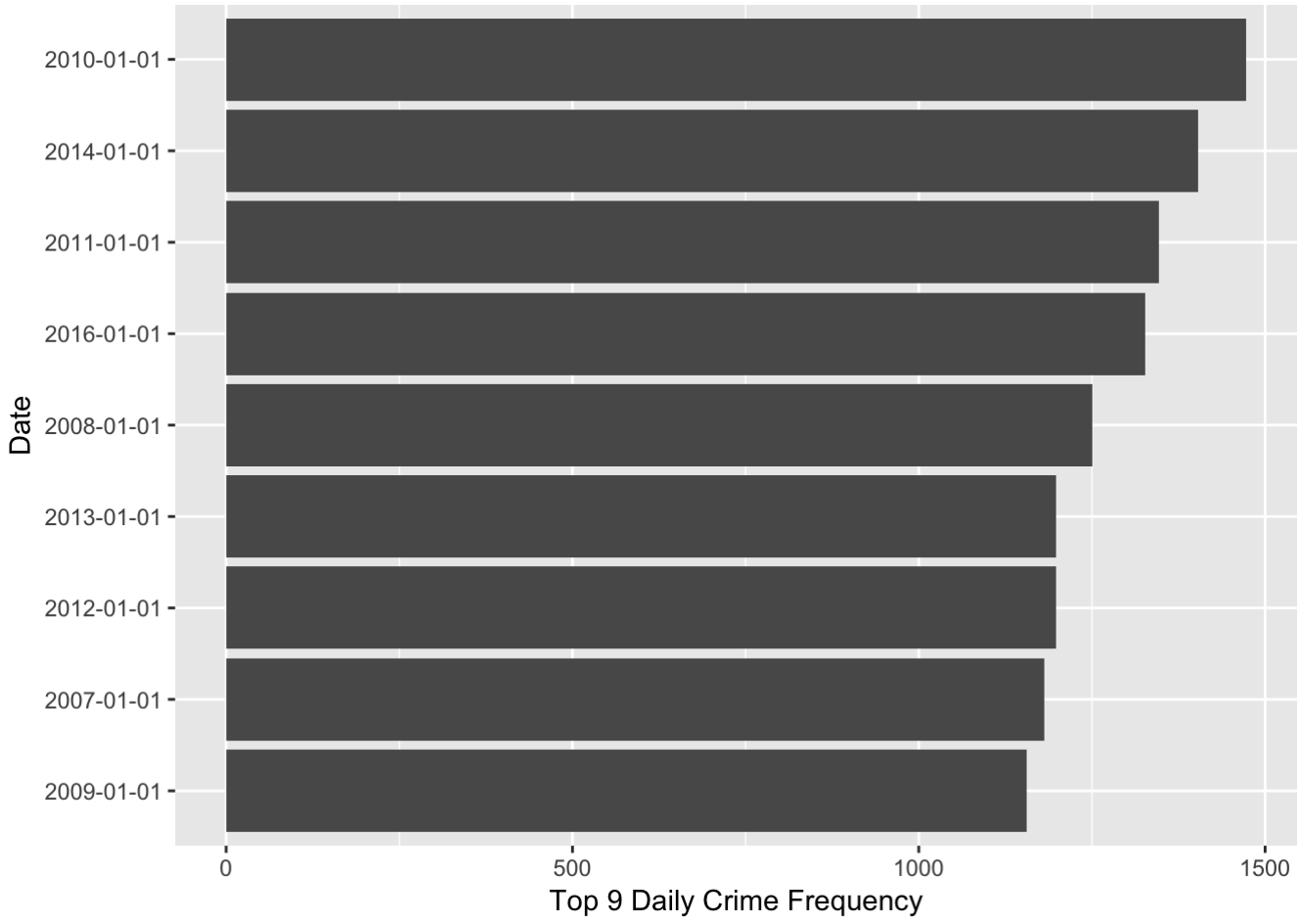
  ===ADDR_PCT_CD===

- 390 missing Ignorable. 77 distinct precincts.
- For some precints, they are counted in more than one borough, i.e., for some cases, they are in one borough while for other cases they are in another borough.

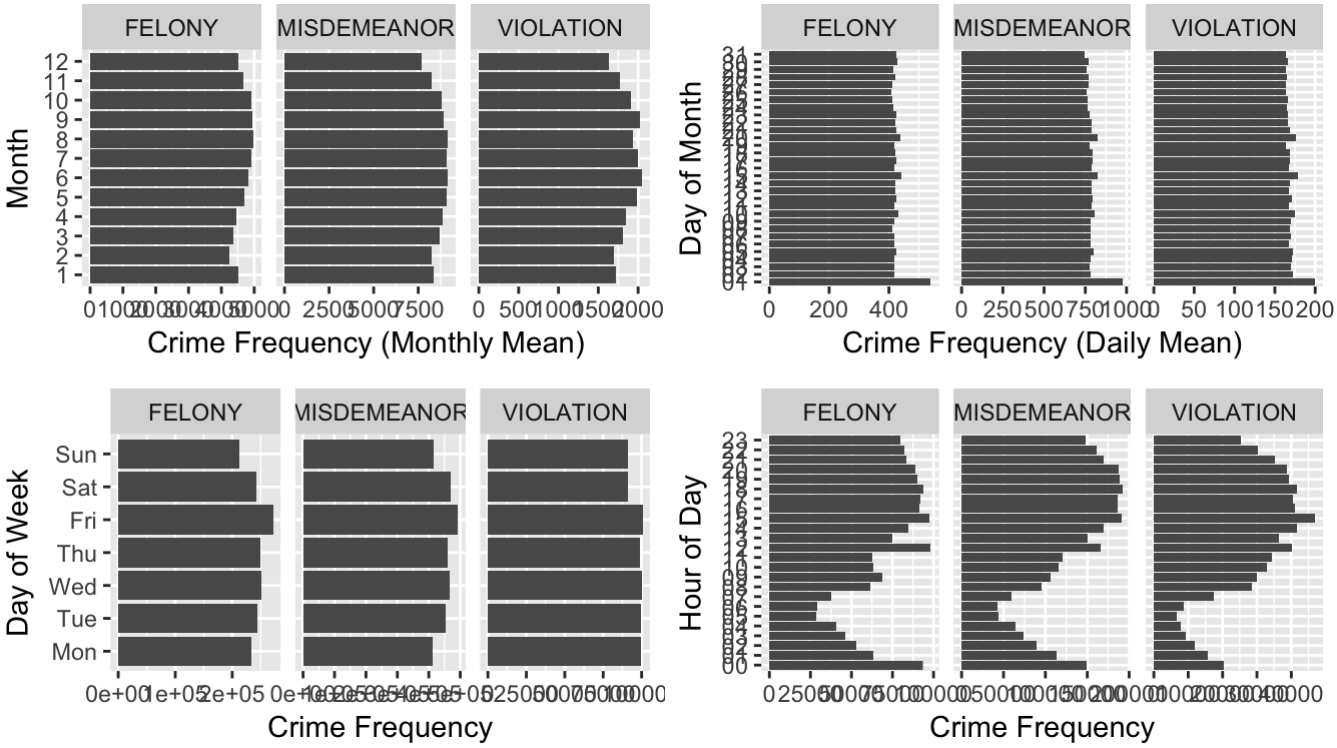
=====Part 4===== Main Analysis

```
##
Read 0.0% of 5580035 rows
Read 10.8% of 5580035 rows
Read 21.1% of 5580035 rows
Read 32.1% of 5580035 rows
Read 43.2% of 5580035 rows
Read 54.3% of 5580035 rows
Read 65.6% of 5580035 rows
Read 76.9% of 5580035 rows
Read 88.4% of 5580035 rows
Read 99.6% of 5580035 rows
Read 5580035 rows and 24 (of 24) columns from 1.329 GB file in 00:00:16
```


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 RPT_DT,CMPLNT_FR_DT and CMPLNT_TO_DT, they don't seem make much sense comparing with others.

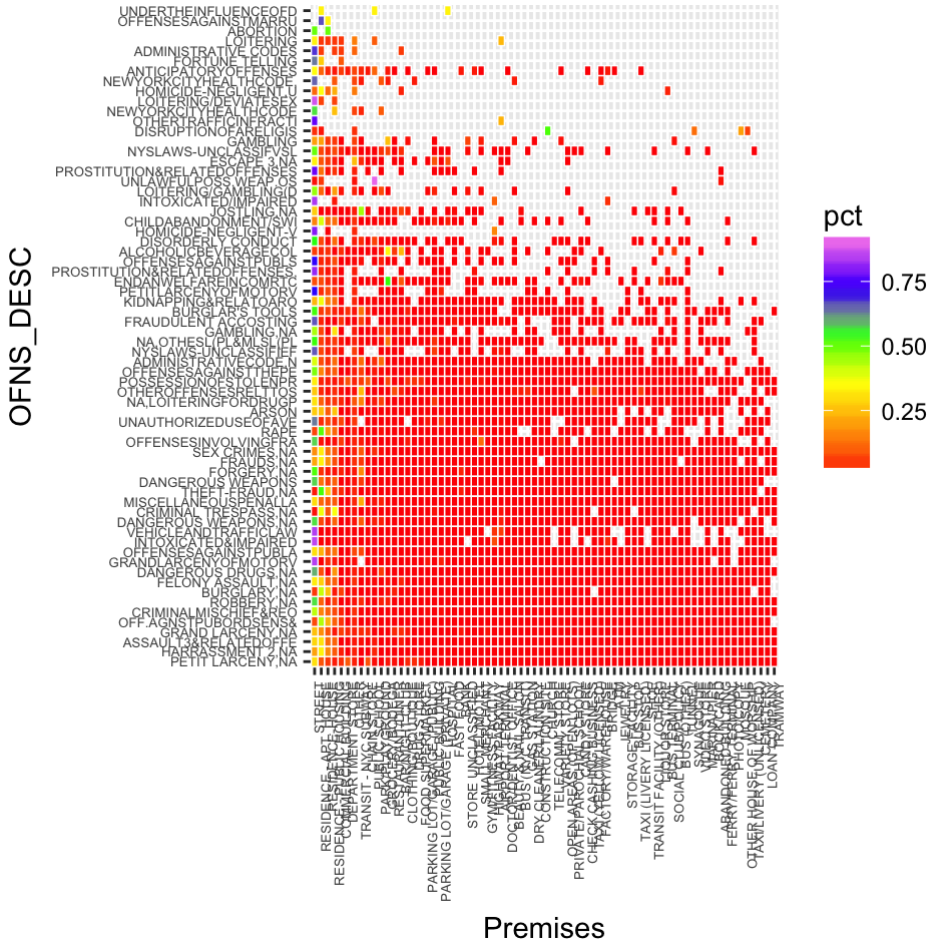


* Indeed by barcharting over the months,we see Jun.-Oct. is a high crime season. * The fake January increasing was due to the errors in the records.

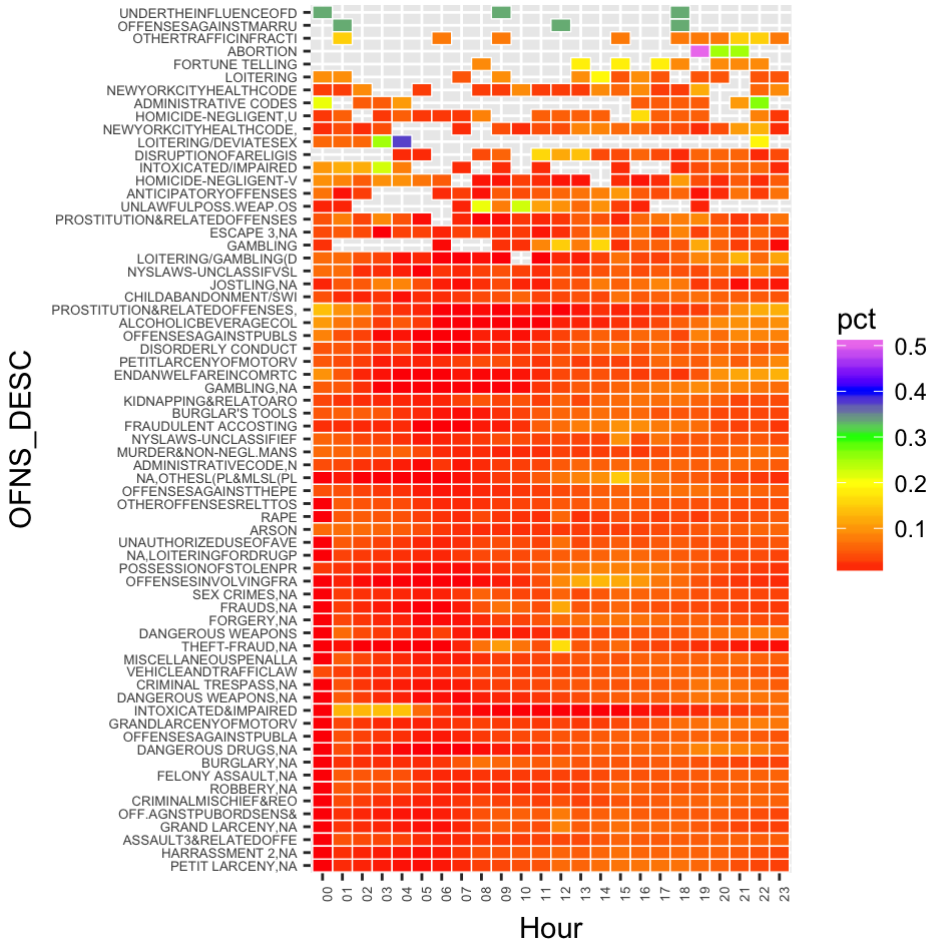
* The spike in Janurary is consistent with the analysis above. * There seemed having a tendency of rounding every 5 day.

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

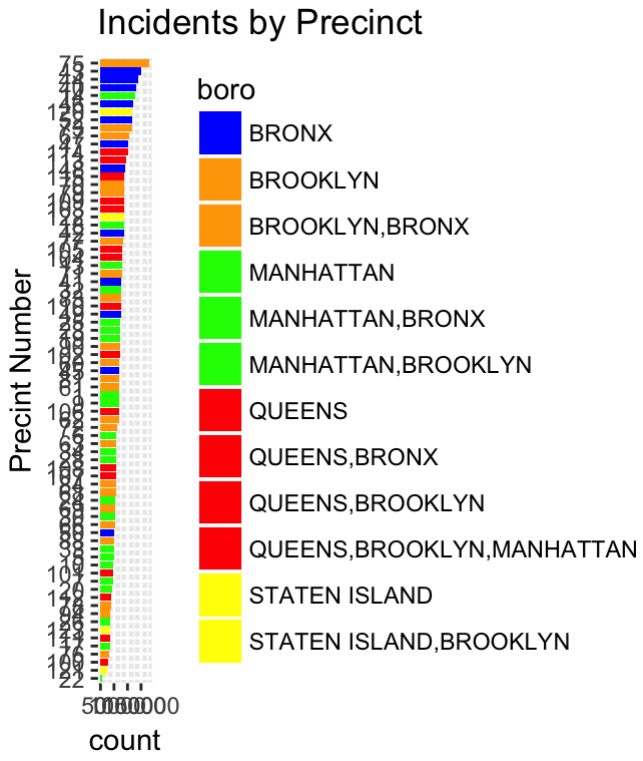
* There is obvious day cycle in the crime occurrence. Early morning has the least crime occurrence while later afternoon has the most crime occurrence.

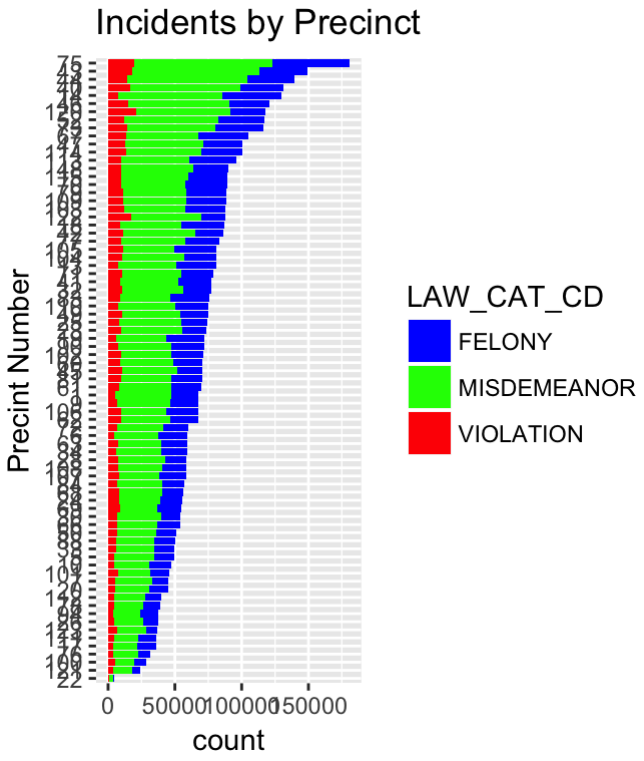


* Doesn't seem having association between crime types and premises.



* 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.

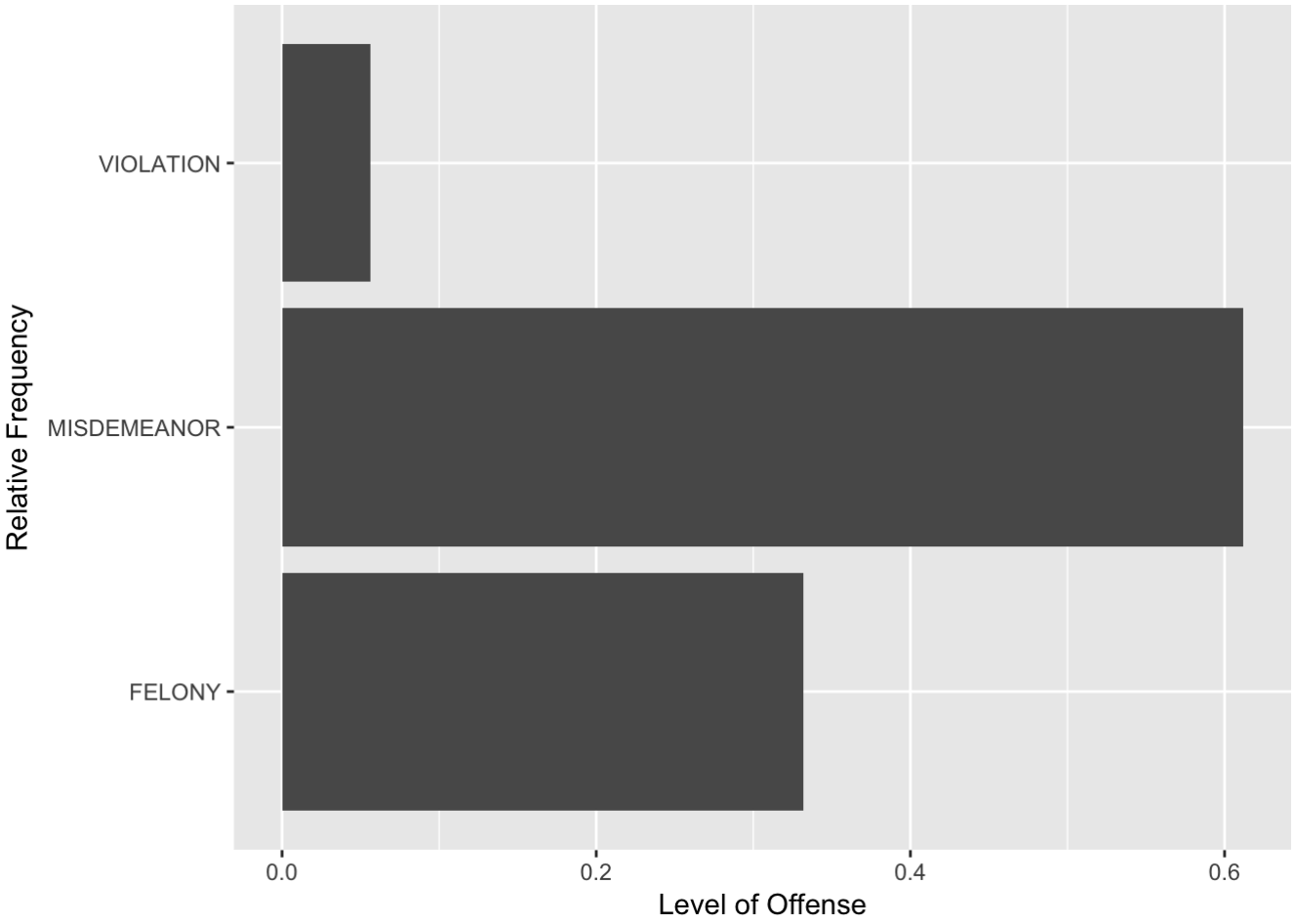




- Similar to Rich's precint plot. But precint number itself doesn't give meaningful information. We can add some meaningful information onto the plot by coloring in borough/location and crime types. Just to see which borough the precints with top crime rates are located, and frequency distribution of 3 crime categories in each precint. Note, there are about 16 cases with precint number not consistent with the borough name (code below will show a list of the precints).
- The borough legends can be modified to 5 borough rather than showing those with double borough names of particular precints.

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

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



* ~12538 cases recorded as occurred in parks/playground or greenspaces. Just a quick peek to see if the crime distribution share the same pattern as the overall data. It is. If needed, we can further investigate into this category.