ap3650_nyc_crime_data_visualization

Anita

March 13, 2018

```
library(data.table)
library(vcdExtra)
library(extracat)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(lubridate)
library(RColorBrewer)


#fread("NYPD_Complaint_Data_Historic.csv",na.strings="",colClasses = c(ParkName="c",HADEVELOPT="c"))->c
#crime_df <- fread("NYPD_Complaint_Data_Historic.csv",na.strings="")
#crime_df_1 <- read.csv("NYPD_Complaint_Data_Historic.csv", header=TRUE)

## Copied from rj2168.rmd for uniform read and variable names

var_names <- c("Id", "DateStart", "TimeStart", "DateEnd", "TimeEnd", "DateReport", "ClassCode", "Offens
               "IntClassCode", "IntOffenseDesc", "AtptCptdStatus", "Level", "Jurisdiction", "Boro", "Pc
               "PremDesc", "ParkName", "HousingDevName", "XCoord", "YCoord", "Lat", "Long", "Lat_Long")

crime_df <- fread("NYPD_Complaint_Data_Historic.csv",na.strings="", col.names = var_names, stringsAsFac
```

```
##
Read 0.0% of 5580035 rows
Read 3.6% of 5580035 rows
Read 6.8% of 5580035 rows
Read 10.8% of 5580035 rows
Read 14.9% of 5580035 rows
Read 19.0% of 5580035 rows
Read 23.1% of 5580035 rows
Read 27.1% of 5580035 rows
Read 31.0% of 5580035 rows
Read 35.3% of 5580035 rows
Read 39.6% of 5580035 rows
Read 44.1% of 5580035 rows
Read 48.4% of 5580035 rows
Read 52.9% of 5580035 rows
Read 57.3% of 5580035 rows
Read 61.8% of 5580035 rows
Read 66.3% of 5580035 rows
Read 70.8% of 5580035 rows
Read 75.3% of 5580035 rows
Read 79.7% of 5580035 rows
Read 84.2% of 5580035 rows
Read 88.5% of 5580035 rows
Read 93.0% of 5580035 rows
Read 97.5% of 5580035 rows
```

```
Read 5580035 rows and 24 (of 24) columns from 1.362 GB file in 00:00:59
```

```
crime_df%>%mutate_if(is.character,funs(factor(.)))->crime_df
```

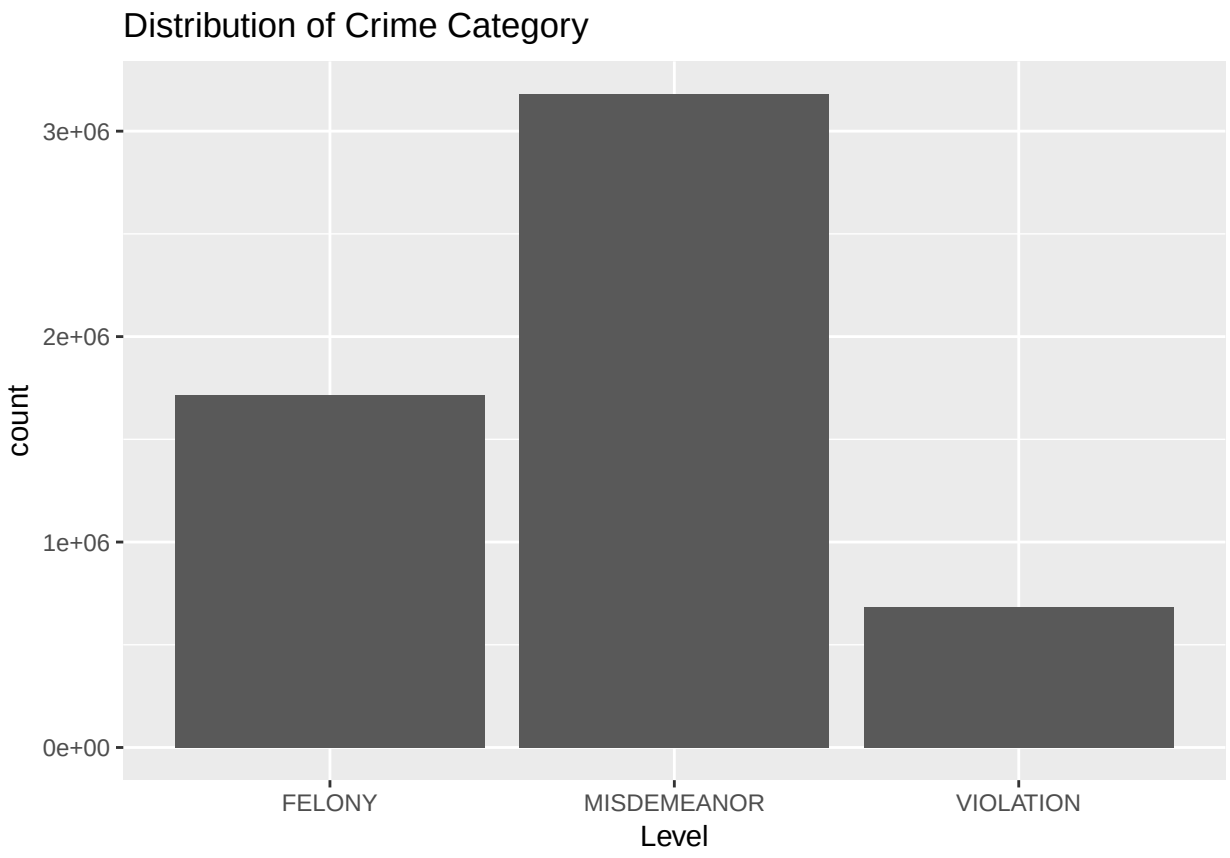
Data Manipulations

```
#Convert dates and times to correct format

#New Variable Names
crime_df$DateStart  <- as.Date(crime_df$DateStart, format='%m/%d/%Y')
crime_df$DateEnd    <- as.Date(crime_df$DateEnd, format='%m/%d/%Y')
crime_df$DateReport <- as.Date(crime_df$DateReport, format='%m/%d/%Y')

crime_df$TimeStart <- as.POSIXct(crime_df$TimeStart, format='%H:%M:%S')
crime_df$TimeEnd   <- as.POSIXct(crime_df$TimeEnd, format='%H:%M:%S')
```

Plots

Warm-up Plot :-) Bar Chart

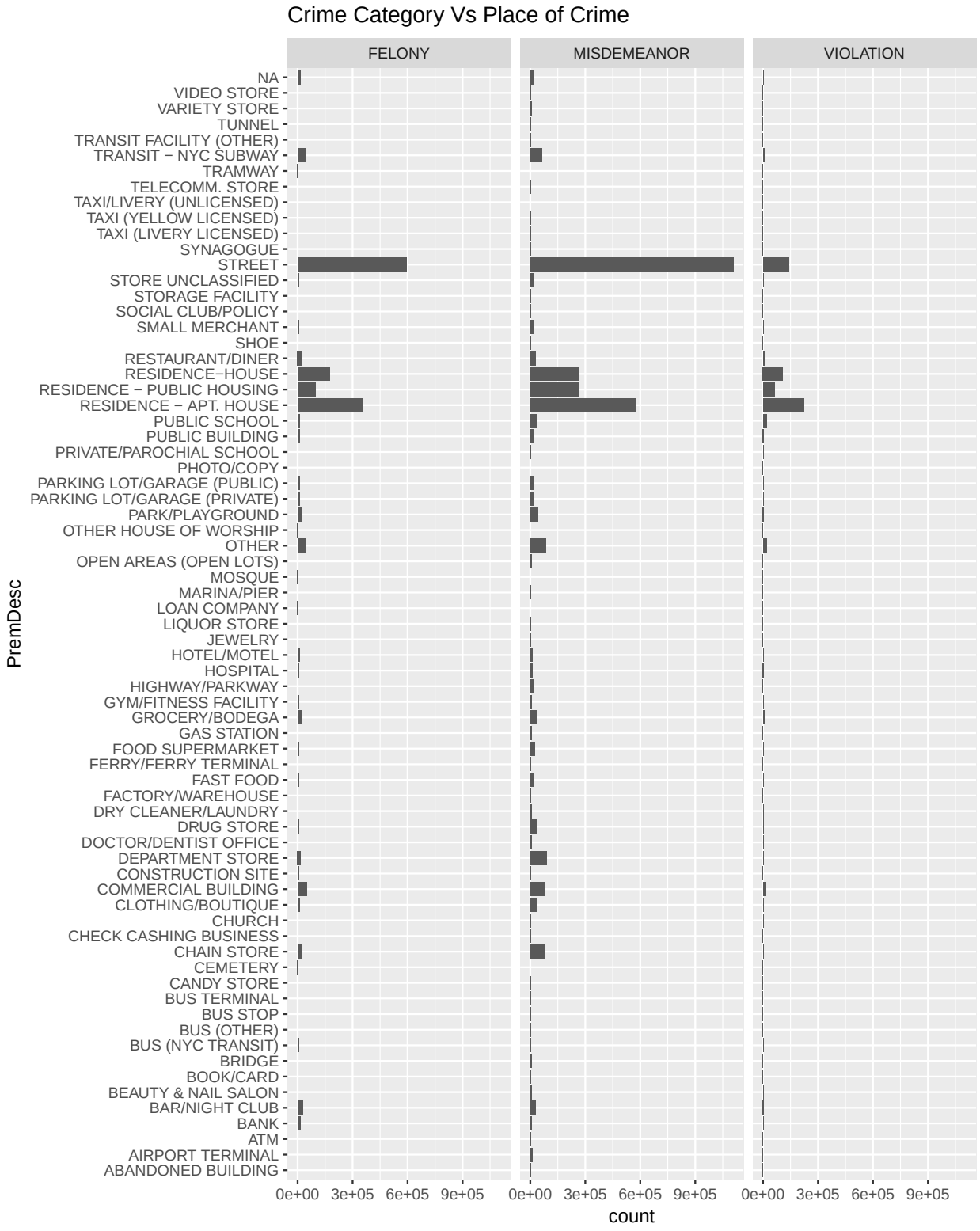
```
ggplot(crime_df,aes(Level)) +
  geom_bar() +
  ggtitle("Distribution of Crime Category")
```



** Observation **
1) Crime Rate of Misdemeanor > Felony > Violation

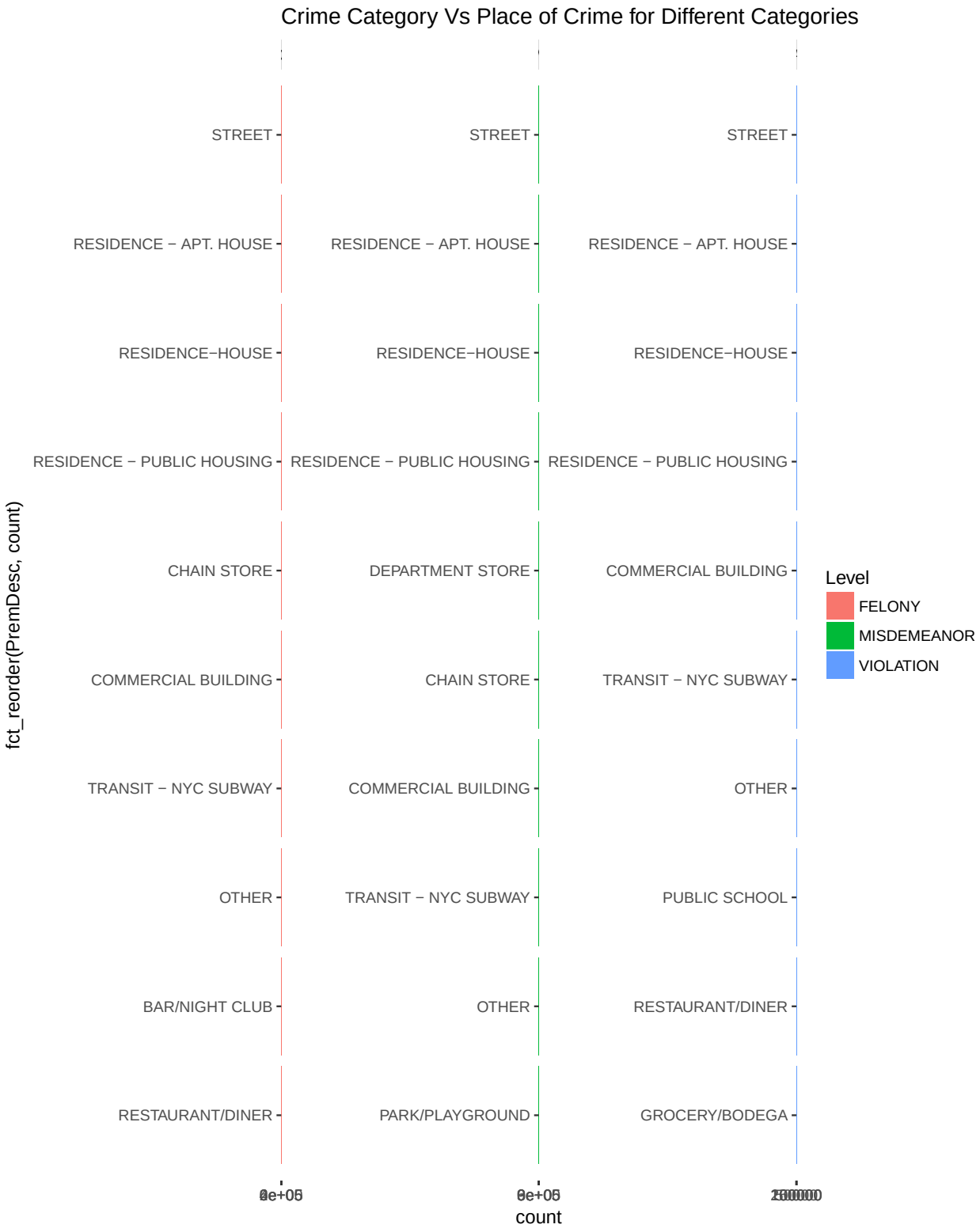
Type of Offense VS Place of Offense

```
ggplot(crime_df,aes(PremDesc)) +
  geom_bar() +
  facet_wrap(~Level) +
  coord_flip() +
  ggtitle("Crime Category Vs Place of Crime")
```




```
crime_place <- crime_df %>%
               filter(!is.na(PremDesc)) %>%
               group_by(Level,PremDesc) %>%
               summarize(count=n()) %>%
```

```
                top_n(n=10, wt=count) %>%
                arrange(Level, count)

ggplot(crime_place, aes(fct_reorder(PremDesc, count), count , fill=Level)) +
  geom_bar(stat="identity") +
  facet_wrap(~Level, scales="free") +
  coord_flip() +
  ggtitle("Crime Category Vs Place of Crime for Different Categories")
```



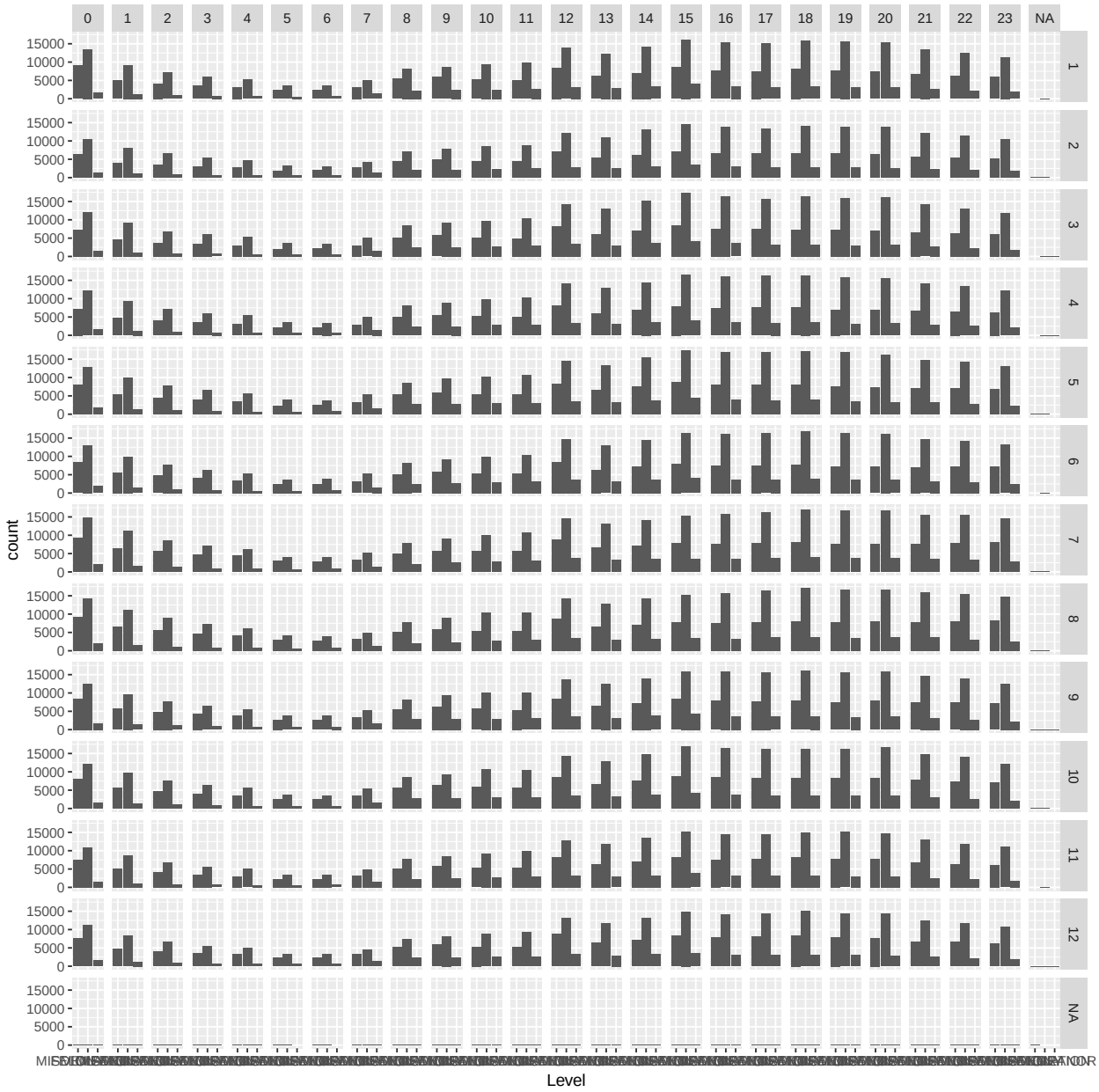
Observations

1) This graph is messed up in order, i cant get the order right for "violation" category 2) Shows top locations for crime under different categories 3) Top 3 locations are almost same for call categories, more violations in Apt-Residence than Street. 4) Also observed violations could be parking / traffic violations, typical locations

include Public school, Diners/Restaurants, grocery etc

Month and Time and Type of Crime

```
#crime_df <- crime_df %>% drop_na()
ggplot(crime_df, aes(Level)) +
geom_bar() +
#facet_wrap(~month(DateStart))
#facet_wrap(~hour(TimeStart))
facet_grid(month(DateStart)~hour(TimeStart))
```

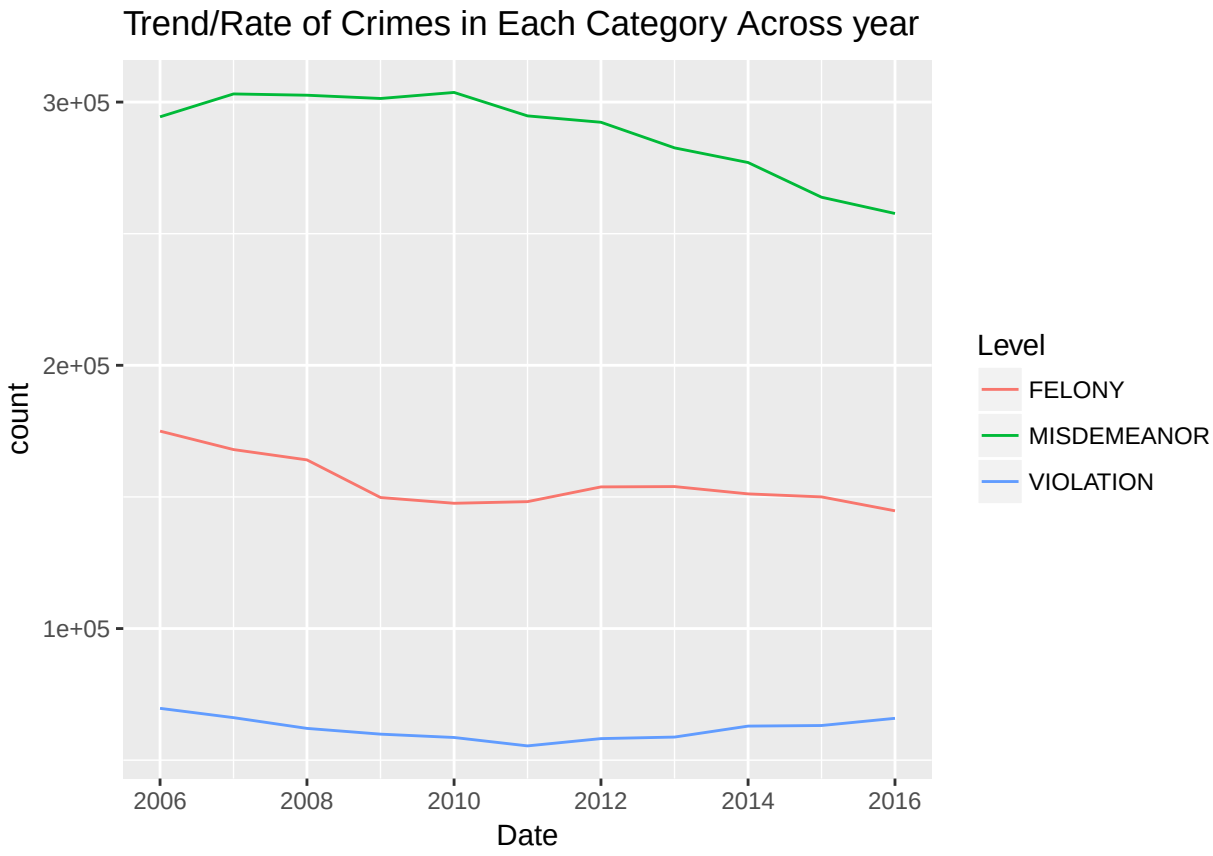


** Inference **
1) Shows less crime during 5 am to 7 am , peak crime 3 pm to 8 pm

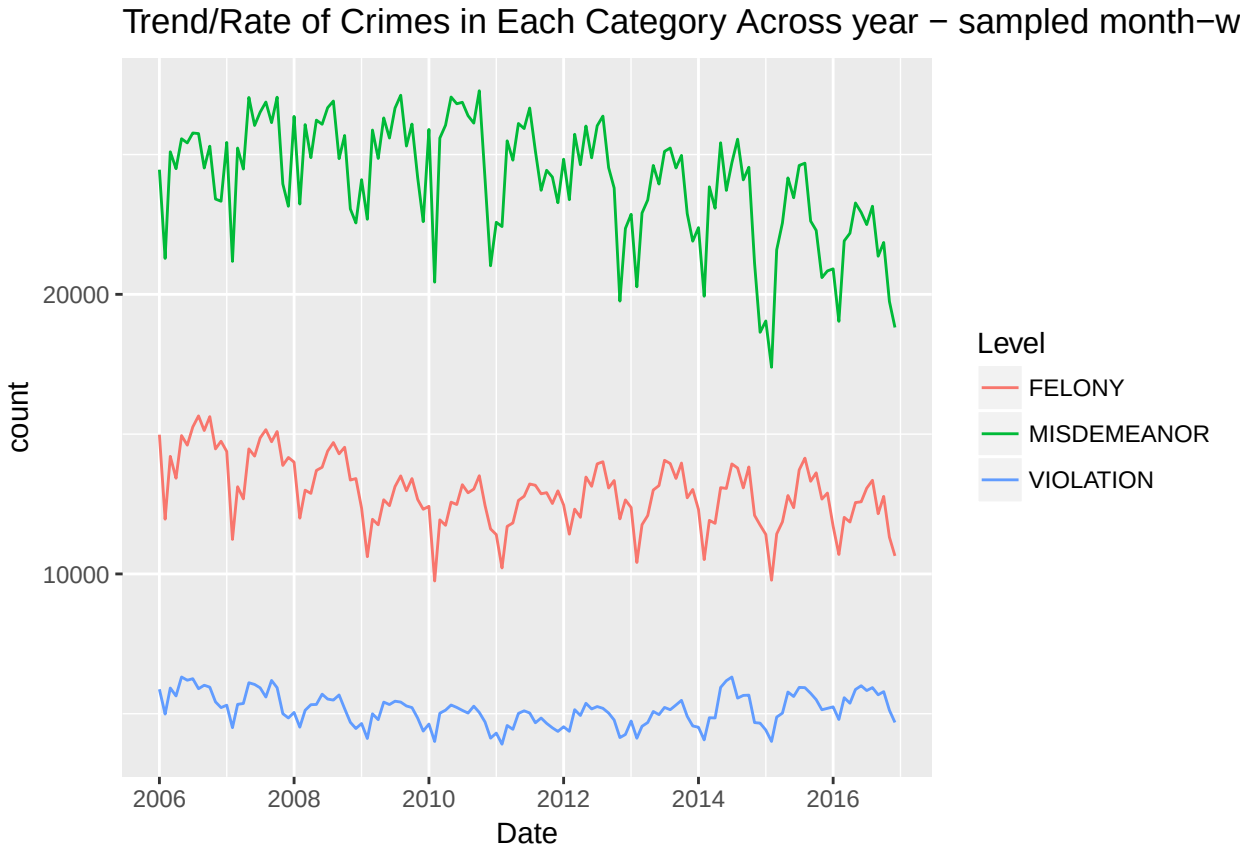
2) The peak and low hourly variances is consistent across months

Time Series - Trend of Crime Rate

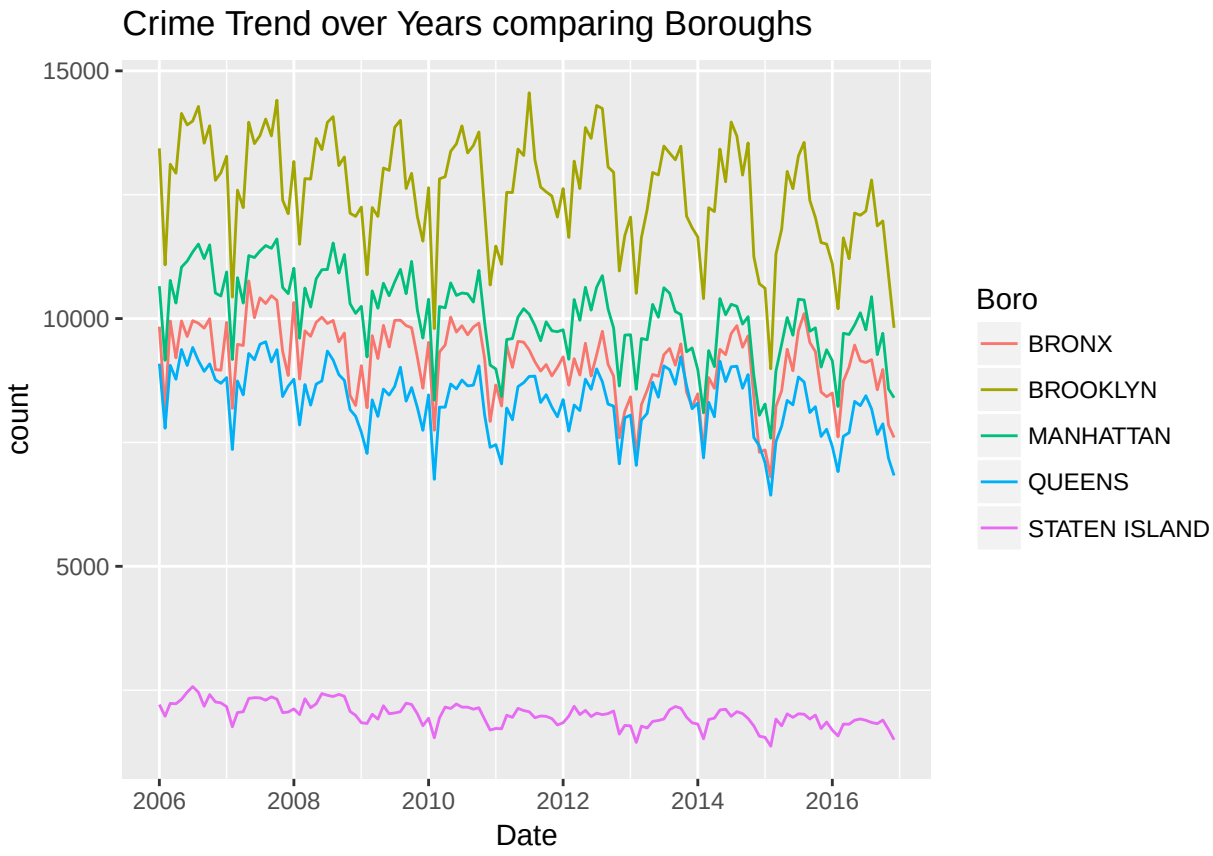
```
crime_time <- crime_df %>%
                filter(year(DateStart)>2005) %>%
                group_by(Date=floor_date(DateStart, "year"),Level) %>%
                summarize(count=n())

ggplot(crime_time, aes(Date,count, color=Level))+
      geom_line() +
      ggtitle("Trend/Rate of Crimes in Each Category Across year")
```



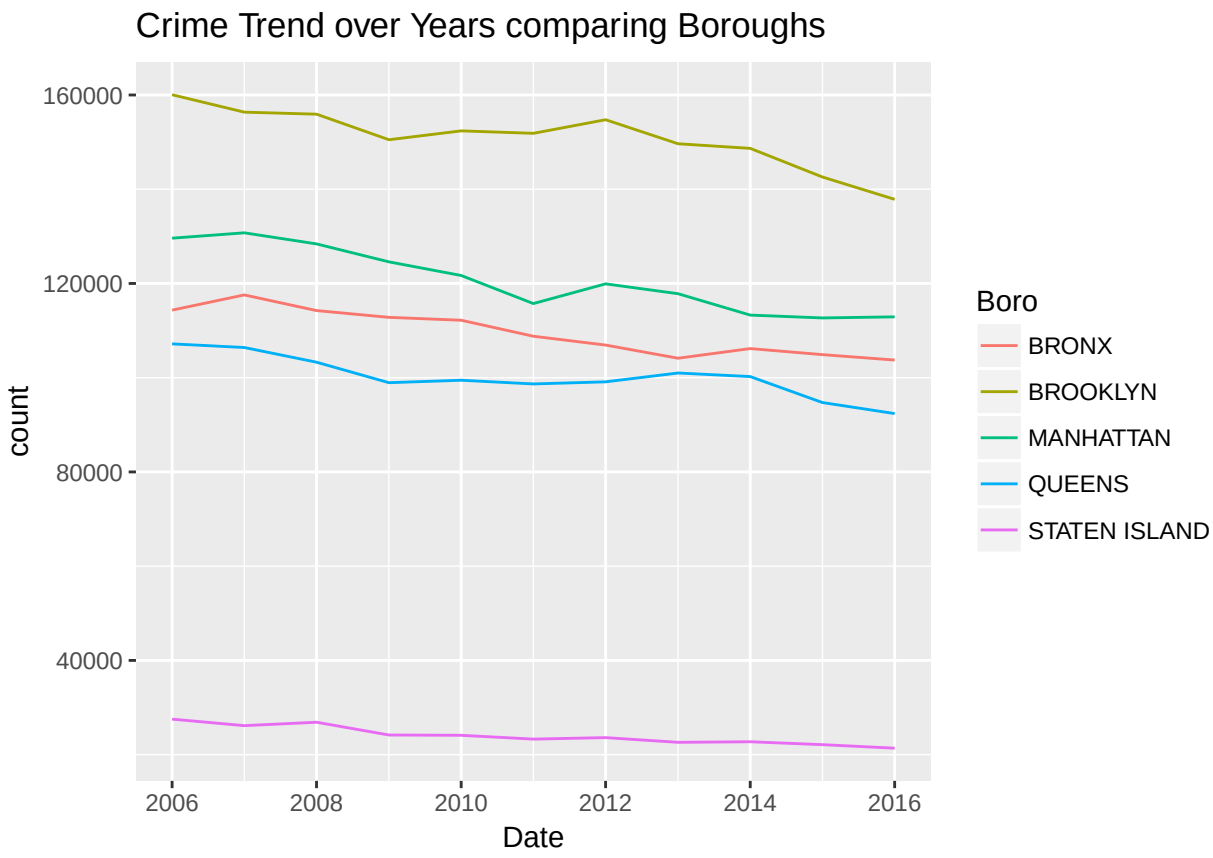

```
crime_time <- crime_df %>%
                filter(year(DateStart)>2005) %>%
                group_by(Date=floor_date(DateStart, "month"),Level) %>%
                summarize(count=n())

ggplot(crime_time, aes(Date,count, color=Level))+
      geom_line() +
      ggtitle("Trend/Rate of Crimes in Each Category Across year - sampled month-wise")
```



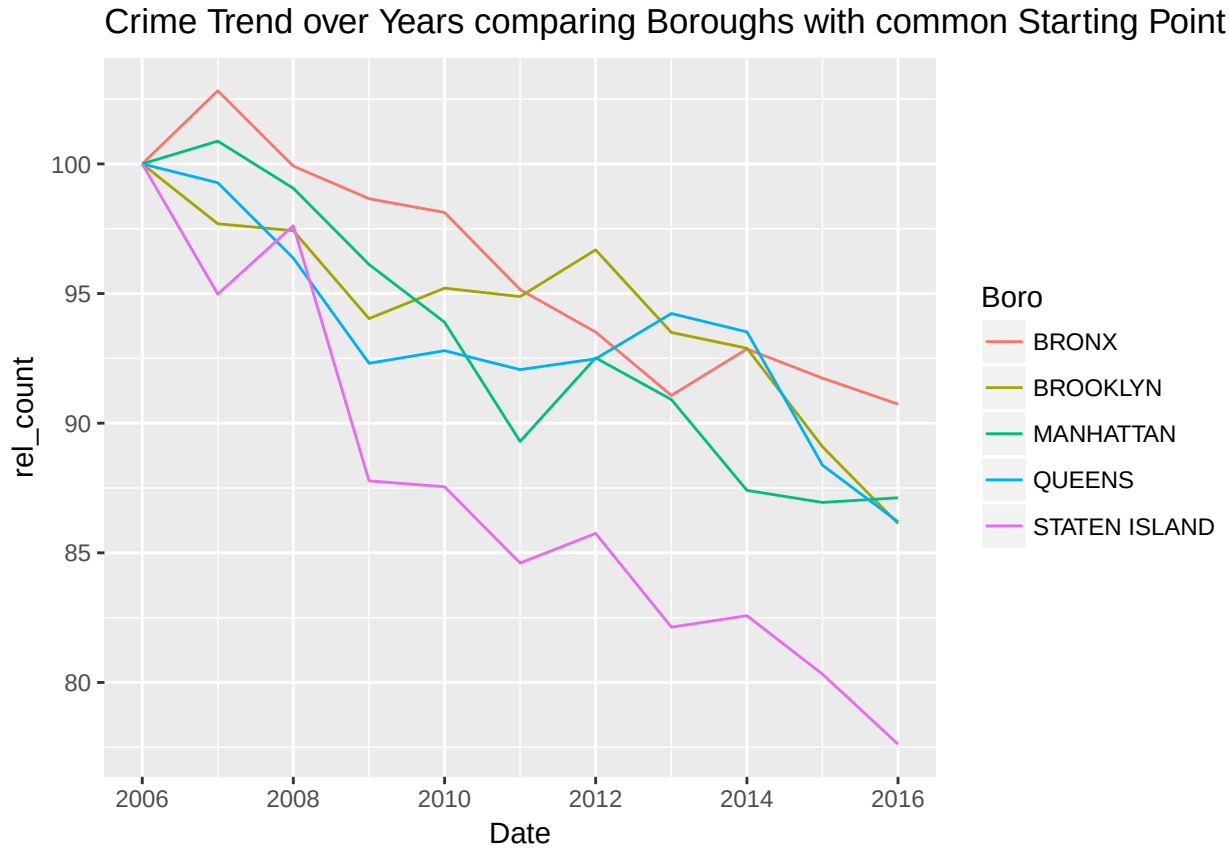

```
crime_boro_m <- crime_df %>%
                  filter(year(DateStart)> 2005 & Boro != "") %>%
                  group_by(Date=floor_date(DateStart,"month"),Boro) %>%
                  summarize(count=n())

ggplot(crime_boro_m,aes(Date,count,color=Boro)) +
  geom_line() +
  ggtitle("Crime Trend over Years comparing Boroughs")
```




```
crime_boro_y <- crime_df %>%
                  filter(year(DateStart)> 2005 & Boro != "") %>%
                  group_by(Date=floor_date(DateStart,"year"),Boro) %>%
                  summarize(count=n()) %>%
                  ungroup() %>%
                  group_by(Boro) %>%
                  mutate(rel_count = count*100/count[1])

ggplot(crime_boro_y,aes(Date,count,color=Boro)) +
  geom_line() +
  ggtitle("Crime Trend over Years comparing Boroughs")
```



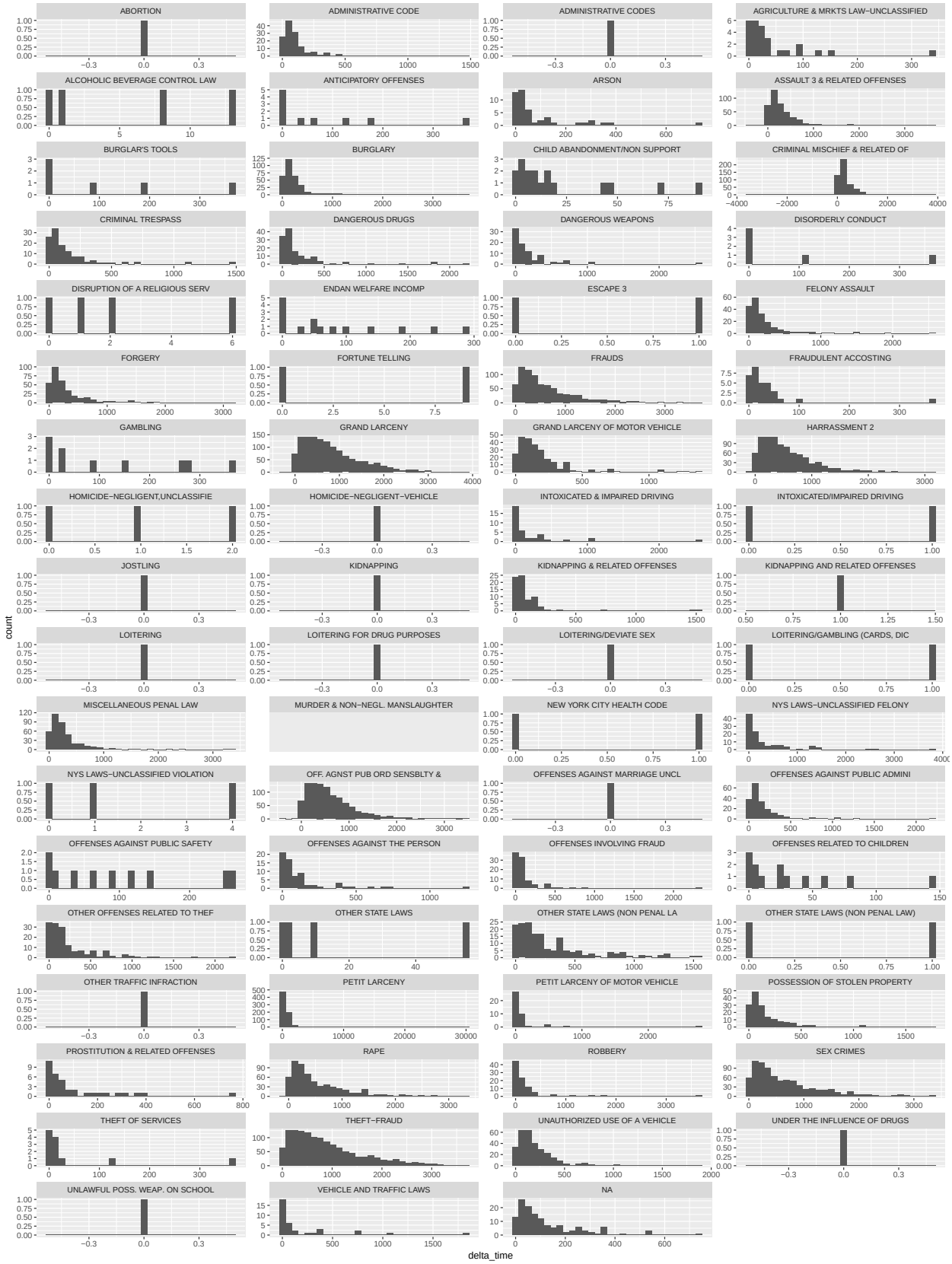

```
ggplot(crime_boro_y,aes(Date,rel_count,color=Boro)) +
  geom_line() +
  ggtitle("Crime Trend over Years comparing Boroughs with common Starting Point")
```



** Inference** * Shows monthly pattern similar to Jingbo's, peaks in middle of year
* Year pattern fluctuates
* Some NM_BORO are empty , found this when plotting without using empty fields, this should be noted in missing data
* Gaps between bororughs reduces towards later years * Using Common Starting point for Crime Trend, we see that in 2008, Staten Island Crime made a huge jump.Need to investiagte why ?

length of Crime Vs Type of Crime

```
#crime_time <- crime_df %>%  drop_na() %>%
crime_time <- crime_df %>%
                  filter(year(DateStart)>2005) %>%
                  mutate(delta_time = as.numeric(DateEnd - DateStart)) %>%
                  group_by(OffenseDesc, delta_time) %>%
                  summarize(count=n())

ggplot(crime_time,aes(delta_time)) +
  geom_histogram(na.rm=TRUE) +
  facet_wrap(~OffenseDesc, ncol = 4, scales="free")
```

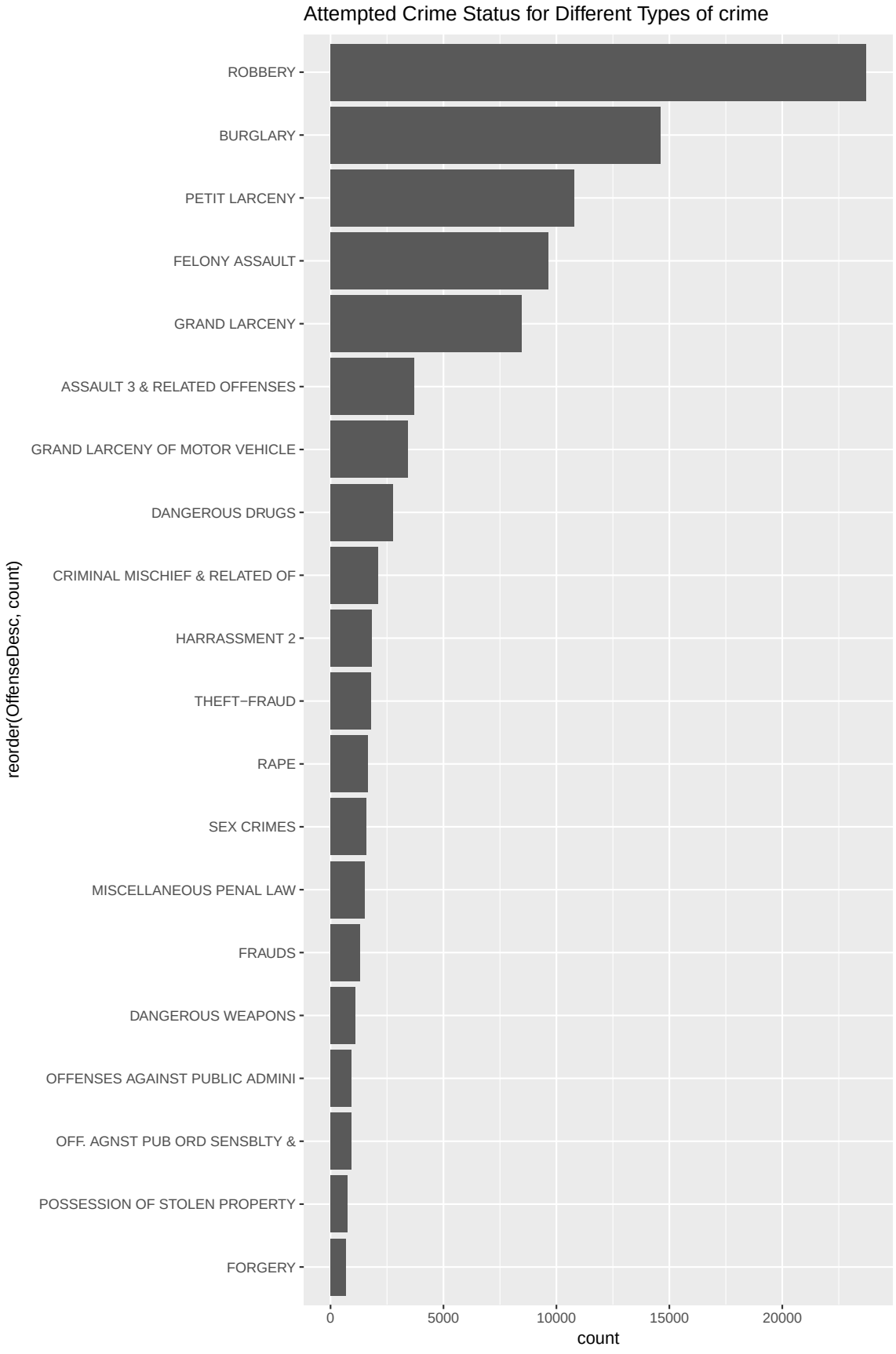


1. There are some cases where there might be typo on "To Date" especially year might be typo
2. Observed larceny(grand and petite have lot of cases)

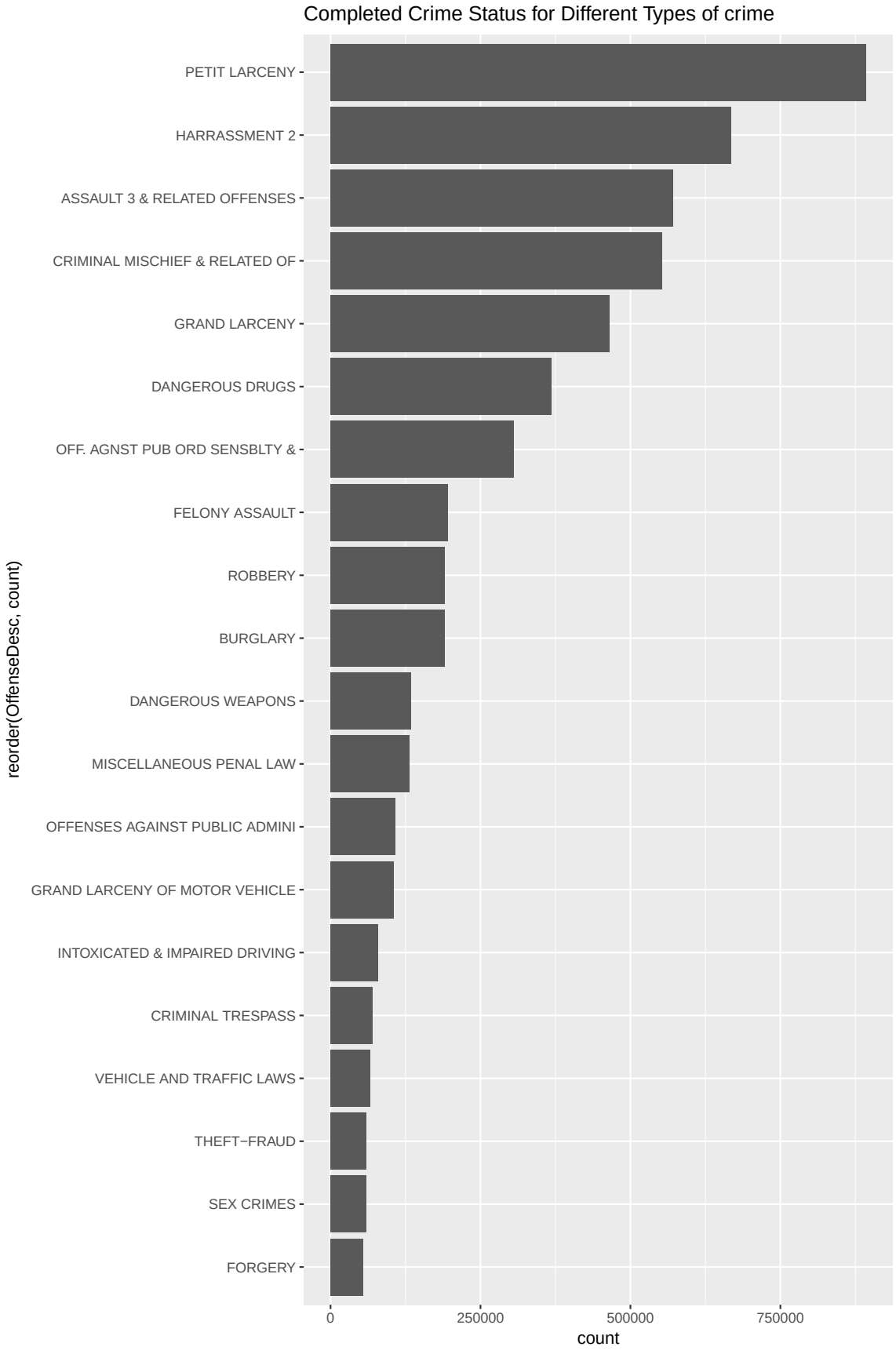
3. There are blank OFfense category 4. I dont like this analysis anymore

Atempted Crime vs Type of Crime

```
crime_stat <- crime_df %>%
                filter(AtptCptdStatus == "ATTEMPTED" & OffenseDesc !="") %>%
                group_by(OffenseDesc) %>%
                summarize(count=n()) %>%
                top_n(n=20, wt=count)

ggplot(crime_stat,aes(reorder(OffenseDesc,count),count)) +
    geom_col() +
    coord_flip() +
    ggtitle("Attempted Crime Status for Different Types of crime")
```


Attempted Crime Status for Different Types of crime
ROBBERY
BURGLARY
PETIT LARCENY
FELONY ASSAULT
GRAND LARCENY
ASSAULT 3 & RELATED OFFENSES
GRAND LARCENY OF MOTOR VEHICLE
DANGEROUS DRUGS
CRIMINAL MISCHIEF & RELATED OF
HARRASSMENT 2
THEFT−FRAUD
RAPE
SEX CRIMES
MISCELLANEOUS PENAL LAW
FRAUDS
DANGEROUS WEAPONS
OFFENSES AGAINST PUBLIC ADMINI
OFF. AGNST PUB ORD SENSBLTY &
POSSESSION OF STOLEN PROPERTY
FORGERY
reorder(OffenseDesc, count)
0
5000
10000
15000
20000
count

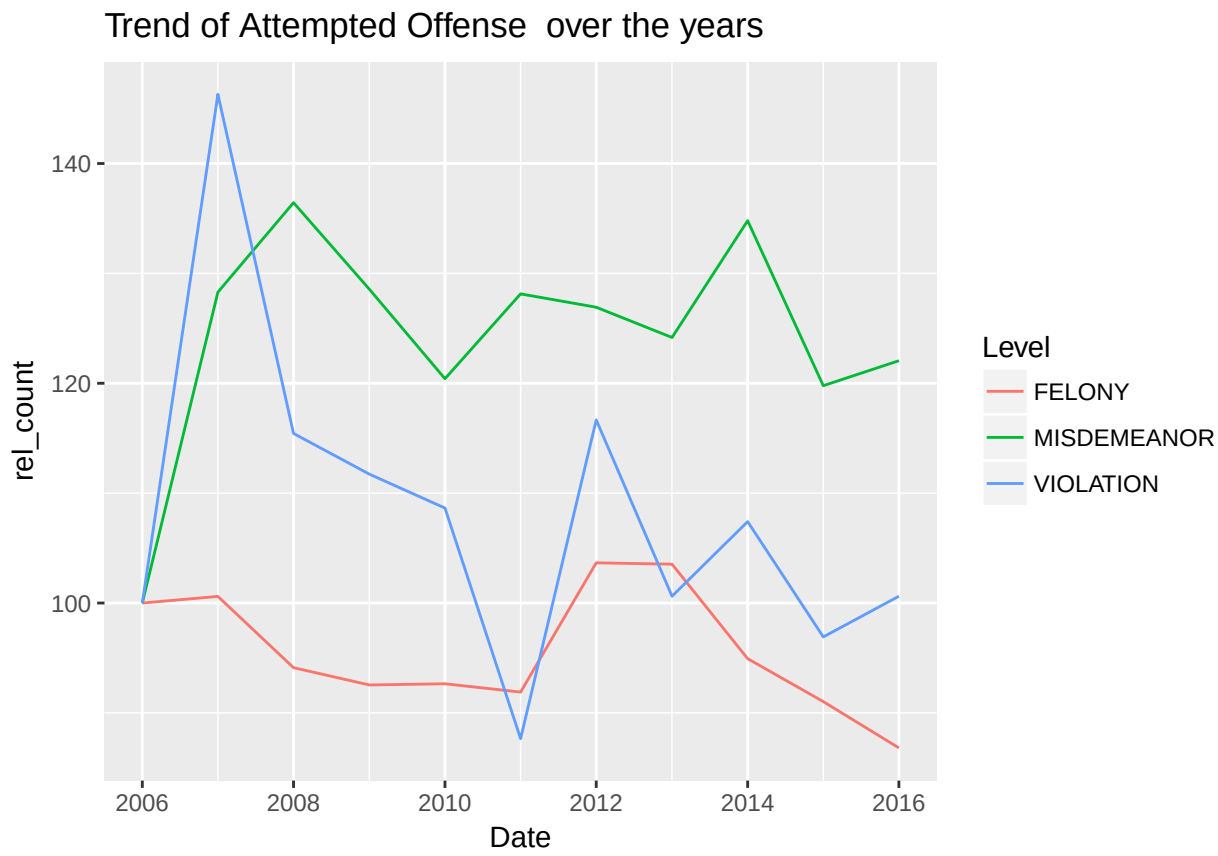
```r
crime_stat_cmtd <- crime_df %>%
                filter(AtptCptdStatus == "COMPLETED" & OffenseDesc !="") %>%
                group_by(OffenseDesc) %>%
                summarize(count=n()) %>%
                top_n(n=20, wt=count)

ggplot(crime_stat_cmtd,aes(reorder(OffenseDesc,count),count)) +
    geom_col() +
    coord_flip() +
    ggtitle("Completed Crime Status for Different Types of crime")
```


Completed Crime Status for Different Types of crime
PETIT LARCENY
HARRASSMENT 2
ASSAULT 3 & RELATED OFFENSES
CRIMINAL MISCHIEF & RELATED OF
GRAND LARCENY
DANGEROUS DRUGS
OFF. AGNST PUB ORD SENSBLTY &
FELONY ASSAULT
ROBBERY
BURGLARY
DANGEROUS WEAPONS
MISCELLANEOUS PENAL LAW
OFFENSES AGAINST PUBLIC ADMINI
GRAND LARCENY OF MOTOR VEHICLE
INTOXICATED & IMPAIRED DRIVING
CRIMINAL TRESPASS
VEHICLE AND TRAFFIC LAWS
THEFT−FRAUD
SEX CRIMES
FORGERY
reorder(OffenseDesc, count)
0
250000
500000
750000
count

** Inference **
1) Total number of offense categories under attempted category is around 55
2) Among top crimes in attempted catgeory is Robbery, Burglar, petit larceny, felony assault, grand lacerny
3) Are these categories the top offense categories too ?
4) For completed categories, the top categories are Petit Larceny, Harrasment 2, Assualt 3, criminal mischief, Grand larceny

Attempted Crime Trend

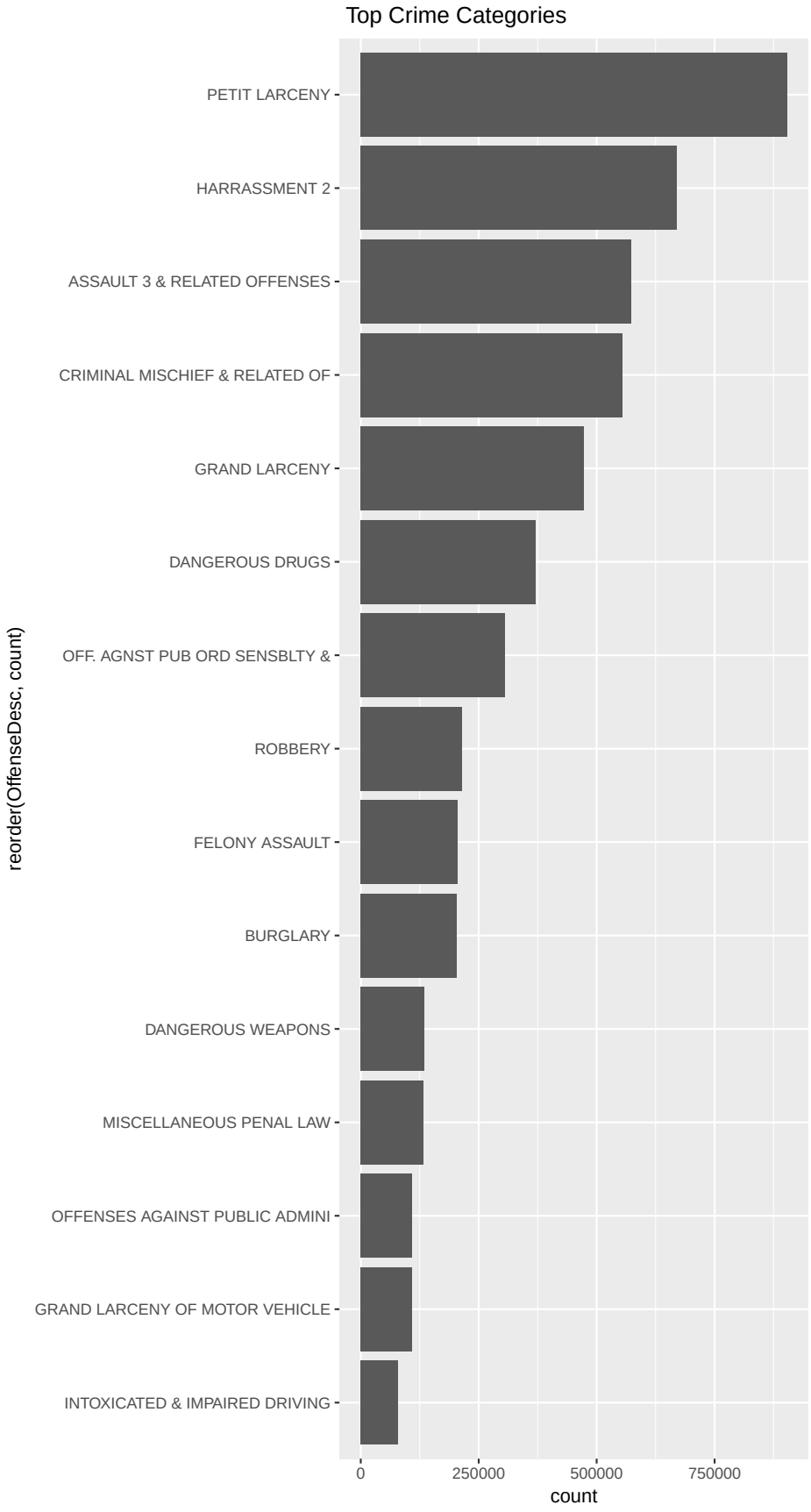
```
crime_stat <- crime_df %>%
                filter(AtptCptdStatus=="ATTEMPTED" & year(DateStart)>2005 & Level != "") %>%
                group_by(Date=floor_date(DateStart,"year"),Level) %>%
                summarize(count=n()) %>%
                ungroup() %>%
                group_by(Level) %>%
                mutate(rel_count = count*100/count[1])

ggplot(crime_stat, aes(Date,rel_count,color=Level)) +
  geom_line() +
  ggtitle("Trend of Attempted Offense  over the years")
```



Inference
1) Over the years, the rate for attempted crime category for Felony is reducing much drastically than violation or misdemeanor

To find Top 10 Crime Categories, mosaic plots building blocks

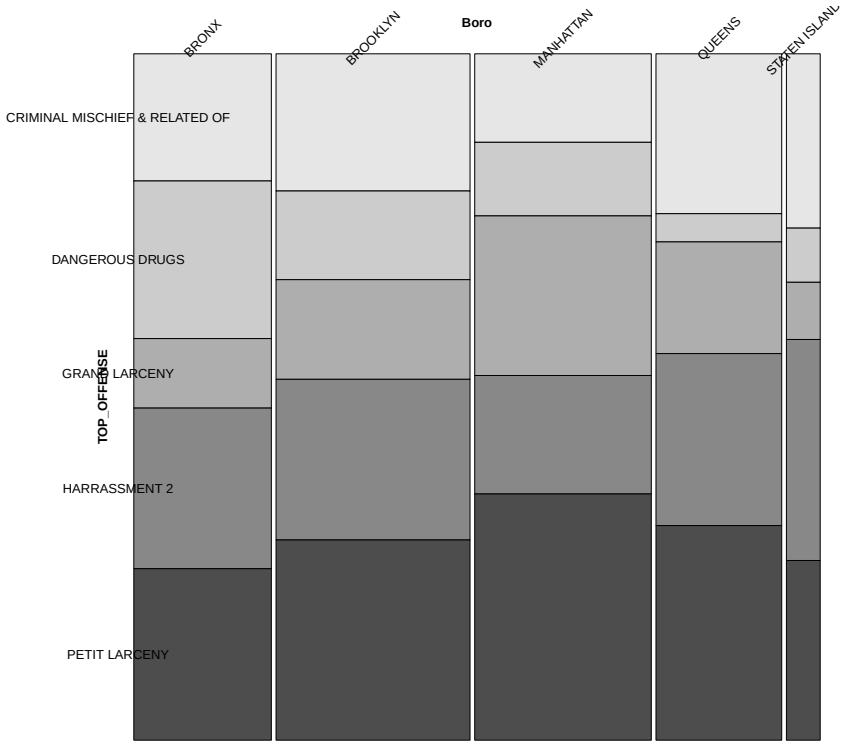
```
crime_top <- crime_df %>%
                filter(OffenseDesc!="") %>%
                group_by(OffenseDesc) %>%
                summarize(count=n()) %>%
                top_n(n=15, wt=count)

ggplot(crime_top, aes(reorder(OffenseDesc,count), count)) +
  geom_col() +
  coord_flip() +
  ggtitle(" Top Crime Categories")
```


Top Crime Categories
PETIT LARCENY
HARRASSMENT 2
ASSAULT 3 & RELATED OFFENSES
CRIMINAL MISCHIEF & RELATED OF
GRAND LARCENY
DANGEROUS DRUGS
OFF. AGNST PUB ORD SENSBLTY &
ROBBERY
FELONY ASSAULT
BURGLARY
DANGEROUS WEAPONS
MISCELLANEOUS PENAL LAW
OFFENSES AGAINST PUBLIC ADMINI
GRAND LARCENY OF MOTOR VEHICLE
INTOXICATED & IMPAIRED DRIVING
reorder(OffenseDesc, count)
0
250000
500000
750000
count

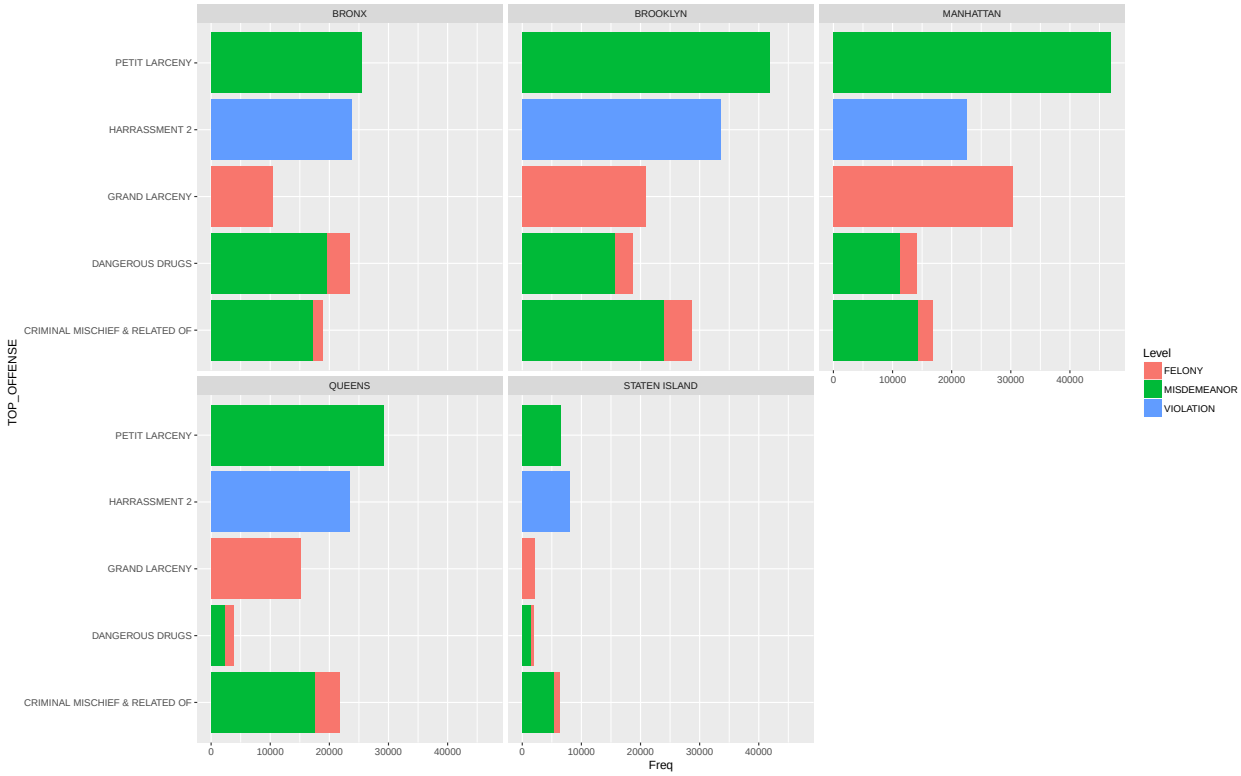
** Inference **
1) The top crime categories are Petit Larceny, Harrasment , Assual, Criminal Mischielf, Grand Larceny , Dangerous Drugs
2) Do we need to find some distribution for these again ??
3) What is miscellaneous Penal LAw ?

Boro, Juris, Crime Categories

```
                #group_by(Boro,Jurisdiction,OffenseDesc) %>%
                #mutate(count=n()) %>%
top_ofns   <- c("PETIT LARCENY", "HARRASSMENT 2", "CRIMINAL MISCHIEF & RELATED OF", "ASSAUALT 3 & REL
crime_sort <- crime_df %>%
                filter(Boro != "", Jurisdiction !="", (OffenseDesc == top_ofns))%>%
                group_by(Boro,Level,OffenseDesc) %>%
                summarize(Freq=n())

crime_sort$TOP_OFFENSE = crime_sort$OffenseDesc[,drop=TRUE]

## I tried changing the code, but I like the above
crime_sort_1 <- crime_df %>%
                filter(Boro != "", OffenseDesc != "") %>%
                group_by(Boro,OffenseDesc) %>%
                summarize(count=n())  %>%
                ungroup() %>%
                group_by(Boro) %>%
                top_n(n=6, wt=count)


#doubledecker(OffenseDesc~Boro, data=crime_sort, gp = gpar(fill = c("grey90", "red")))
mosaic(TOP_OFFENSE~Boro, direction=c("v"), labeling=labeling_border(rot_labels=c(45,0,0, 0)), crime_s
```




```
#doubledecker(TOP_OFFENSE~Boro, data=crime_sort)
ggplot(crime_sort, aes(TOP_OFFENSE,Freq, fill=Level)) +
  geom_col() +
  facet_wrap(~ Boro) +
  coord_flip()
```



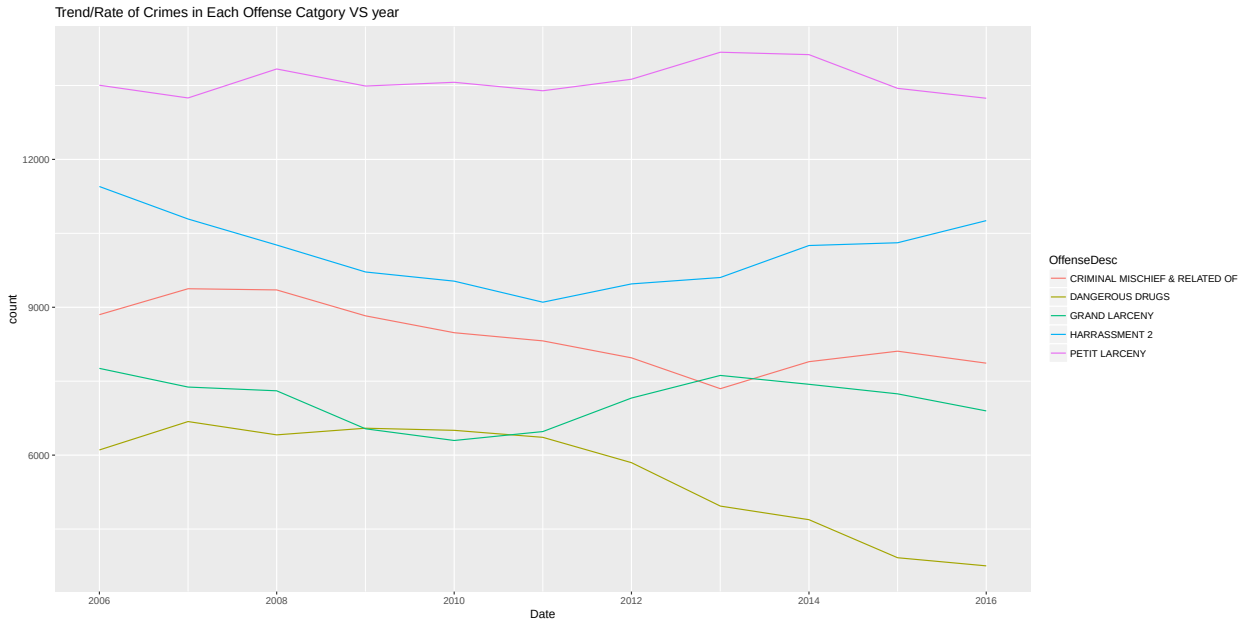
** Inference from Mosaic Plot **
1) How the Top crimes are distributed across Boroughs ?
2) Drugs are more in Bronx compared to Queens
3) Manhattan tops for Petit Larceny and Grand Larceny
4) In staten Island, Harrasment tops among other crimes in staten island
5) From the stacked bar graphs, it shows some comparisions easier. For example in mosaic it is difficult to compare areas of distribution when widths of bars are not same. However Here it clearly shows comparisons among boroughs for each crime as well within boroughs for different crimes
6) It also shows that for Drugs and Criminal mischief the crime category can either be Misdemeanor or Felony depending on severity. In both cases, number of crimes as misdemeanor is more than felony.

** The above plot shows something surprising, the categories are not standard, need to research more. For example, dangerous drugs is under Felony as well as Misdemeanor!! **

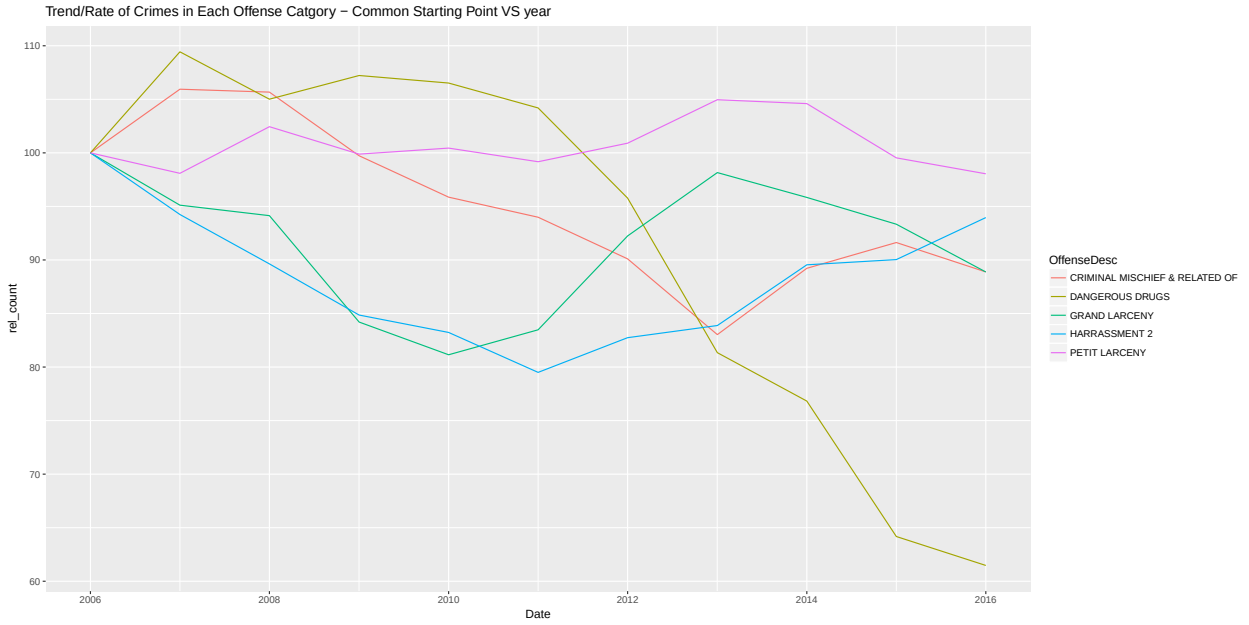
Time Trend of Top OFFENSE Category

```
top_ofns   <- c("PETIT LARCENY", "HARRASSMENT 2", "CRIMINAL MISCHIEF & RELATED OF", "ASSAUALT 3 & RE
crime_time_top_ofns <- crime_df %>%
                       filter(year(DateStart)>2005, (OffenseDesc == top_ofns)) %>%
                       group_by(Date=floor_date(DateStart, "year"),OffenseDesc) %>%
                       summarize(count=n())

crime_time_top_ofns <- crime_time_top_ofns %>%
                       group_by(OffenseDesc) %>%
                       mutate(rel_count = count*100/count[1])

ggplot(crime_time_top_ofns, aes(Date,count, color = OffenseDesc))+
      geom_line() +
```

```
      ggtitle("Trend/Rate of Crimes in Each Offense Catgory VS year")
```




```
ggplot(crime_time_top_ofns, aes(Date,rel_count, color = OffenseDesc))+
      geom_line() +
      ggtitle("Trend/Rate of Crimes in Each Offense Catgory - Common Starting Point VS year")
```



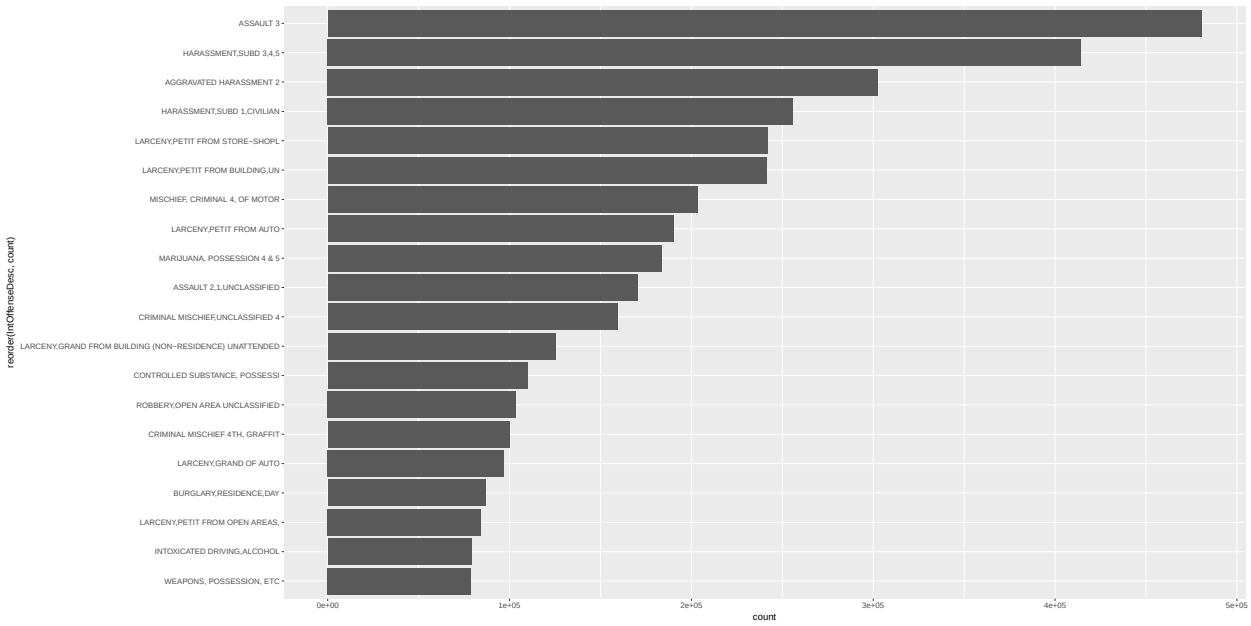
Inference

1) The first graph captures the trend of crimes over the years 2006-2016.
2) The ranks of top crimes have not changed from 2006 o 2016, except in 2013, we see a dip in number of crimes of category “criminal mischief” , similar dip is observed in “Grand Larceny” in 2010 3) To compare the trends of crime by fixing them to a common starting point, we observe that the “dangerous Drugs” category has dipped down drasticaly when compared to other crimes.
4) Harrasment has shown an increase in the last five years.
5) There is slight increase in Criminal Mischief during the period 2013-2015 6) There is a sudden increase in

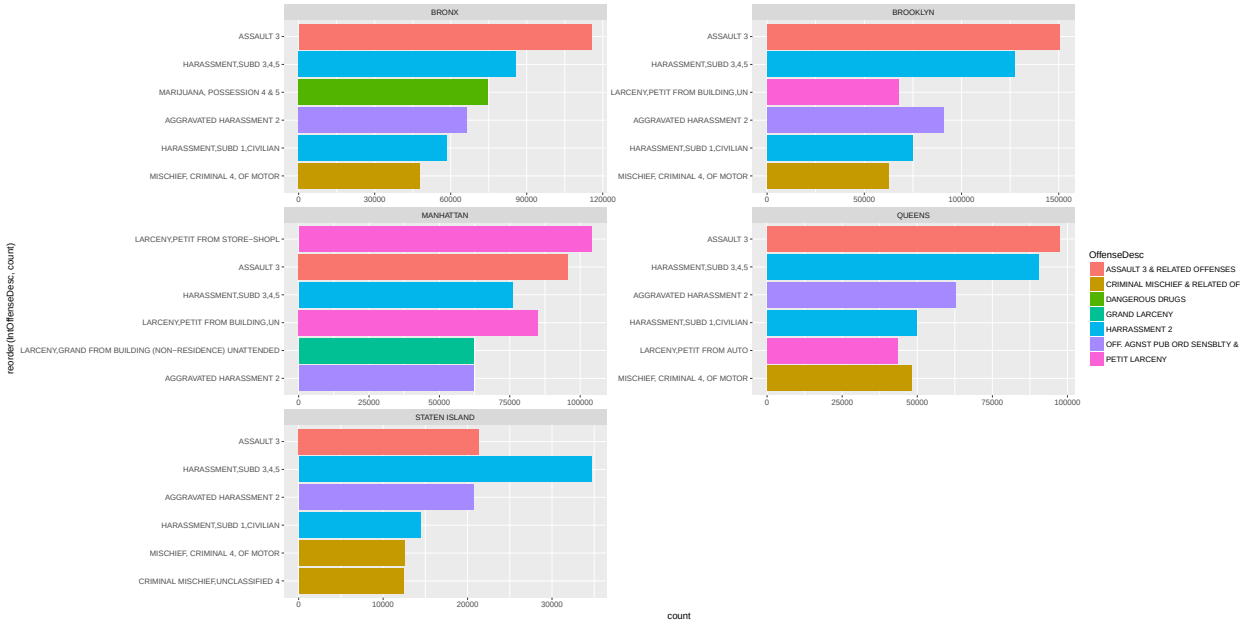
Grand Larceny and Petite Larceny from 2011 to 2013, more than other crimes

Crime PD top

```
  crime_pd_top_all_boro <- crime_df %>%
                    filter(!is.na(Boro)) %>%
                    filter(IntOffenseDesc != "" && !is.na(IntOffenseDesc) && OffenseDesc != "" && Bor
                    group_by(Boro,OffenseDesc, IntOffenseDesc) %>%
                    summarize(count = n()) %>%
                    ungroup() %>%
                    group_by(Boro) %>%
                    top_n(n=6, wt=count) %>%
                    arrange(Boro, desc(count))


  crime_pd_top <- crime_df %>%
                    filter(IntOffenseDesc != "" && !is.na(IntOffenseDesc)  && (Boro != "")) %>%
                    group_by(IntOffenseDesc) %>%
                    summarize(count = n()) %>%
                    top_n(n=20, wt=count)


ggplot(crime_pd_top, aes(reorder(IntOffenseDesc,count), count)) +
  geom_col() +
  coord_flip()
```




```
ggplot(crime_pd_top_all_boro, aes(reorder(IntOffenseDesc,count),count, fill=OffenseDesc)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~Boro, scales="free" , ncol=2)
```



Inference

Interesting to observe the internal classification of top crimes levels
1) These plots are just to analyse the sub categories of Crime/ Offense
2) Petite Larceny is generally from stores and building
3) Different levels of Harassment take the top ranks 2-4
4) Crimes involving Drugs are usually Possession of Marijuana
5) Not sure if you want to use borough comparisions again

total classification of overall crimes (pd_desc) -> 409

Crimes in parks Analysis

```
 crime_parks <- crime_df %>%
                filter(Boro!="",ParkName!="",Level!="") %>%
                group_by(Boro,ParkName,Level) %>%
                summarize(count=n())

 #crime_parks <- crime_parks %>%
 #               arrange(desc(count))

crime_pk <- crime_parks %>%
               group_by(Boro) %>%
               top_n(n=10, wt=count)

 crime_parks_1 <- crime_df %>%
                filter(Boro!="",ParkName!="",OffenseDesc!="") %>%
                group_by(Boro,ParkName,OffenseDesc) %>%
                summarize(count=n())

crime_pk_1 <- crime_parks_1 %>%
               group_by(Boro) %>%
               top_n(n=10, wt=count) %>%
               arrange(Boro,desc(count))
```

```
getPalette = colorRampPalette(brewer.pal(18, "Set1"))

ggplot(crime_pk_1 ,aes(reorder(ParkName,count), count, fill=OffenseDesc)) +
  geom_col() +
  facet_wrap(~Boro, ncol=1, scales="free_y") +
  scale_fill_manual(values = getPalette(18)) +
  # scale_fill_brewer(palette="Set3") +
  coord_flip()
```


```
ggplot(crime_pk_1 ,aes(reorder(ParkName,count),  count, fill=OffenseDesc)) +
  geom_col() +
  facet_wrap(~Boro, ncol=1, scales="free") +
  scale_fill_manual(values = getPalette(18)) +
```

```
  coord_flip()
```

** Inference **
1) Most parks in Bronx have crimes related to Drugs 2) Drugs does not figure among the top crimes in Cental park in Manhatan, where as other parks of Manhatan have Drugs as one of their top category
3) For the parks in Queens, Drugs does not figure as top crimes in any of the parks

trial on ggmap

```
  library(ggmap)

  #NYC <- get_map(location = "new york city",  color = "bw", zoom = 15, source = "google")
```

```
#ggmap(NYC)
#
#ggplot()+geom_point(data = crime_df, aes(x = Longitude, y = Latitude ,colour = factor(Level)))
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.