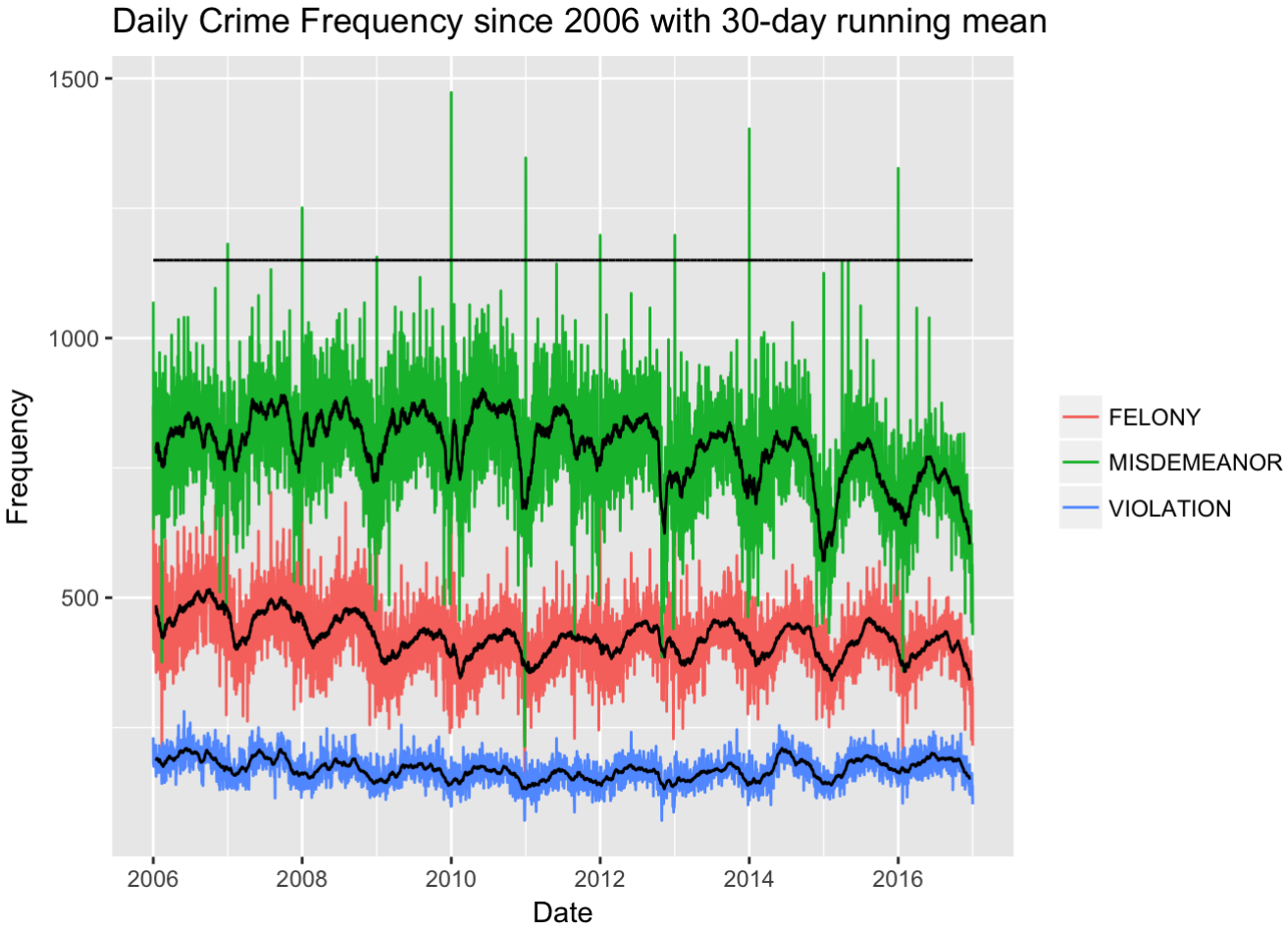
=====Part 4===== Main Analysis

```
#if you have not please install data.table package before run the codes below
#install.packages(data.table)
library(zoo)
library(vcd)
library(dplyr)
library(tidyr)
library(ggplot2)
library(forcats)
library(vcdExtra)
library(gridExtra)
library(tidyverse)
library(data.table)
fread("NYPD_Complaint_Data_Historic.csv",na.strings="",colClasses = c(PARKS_NM="c",HADEV
ELOPT="c"))->df
```

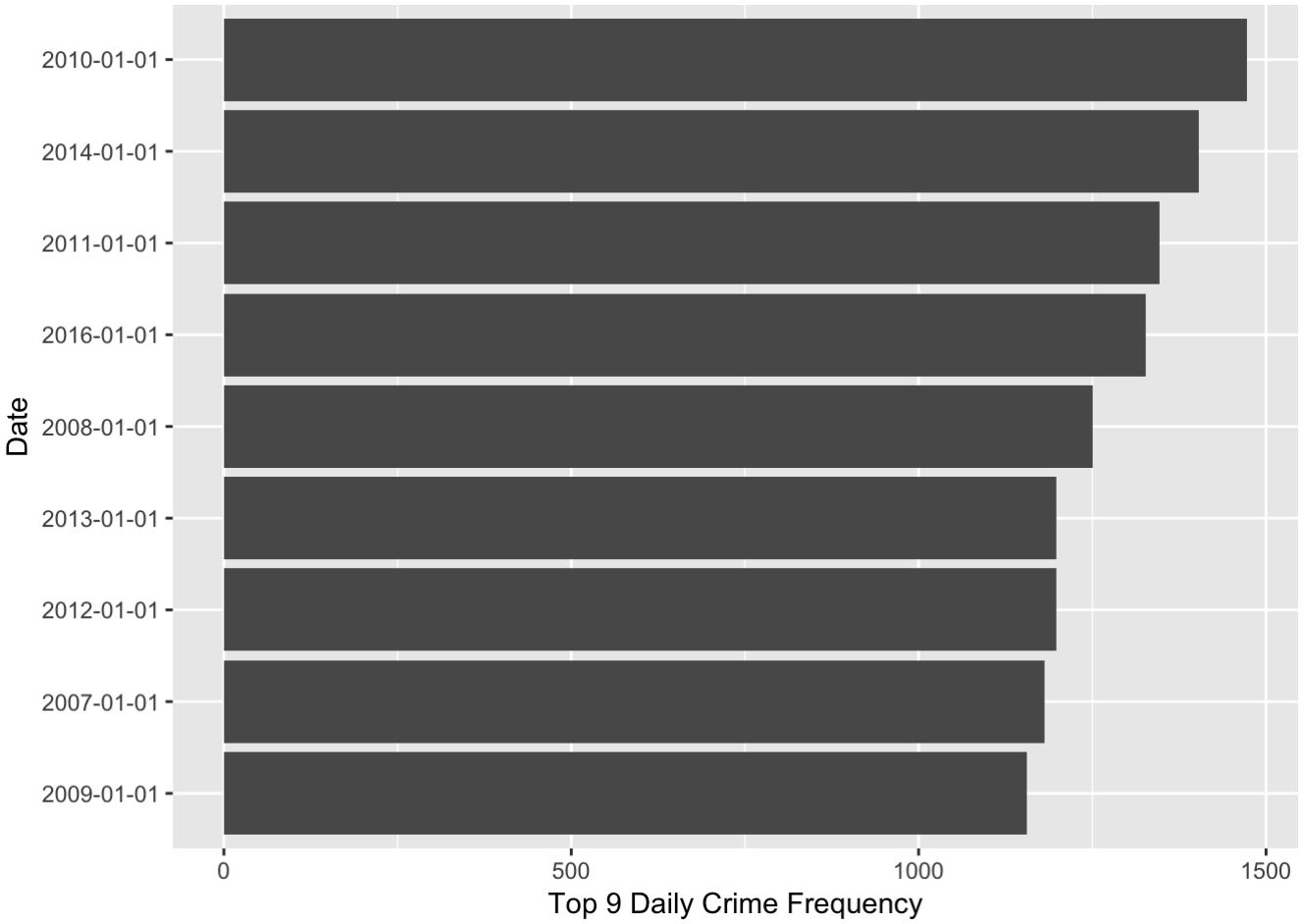
```
##
Read 0.0% of 5580035 rows
Read 9.5% of 5580035 rows
Read 19.0% of 5580035 rows
Read 28.0% of 5580035 rows
Read 38.0% of 5580035 rows
Read 48.7% of 5580035 rows
Read 59.5% of 5580035 rows
Read 70.4% of 5580035 rows
Read 81.2% of 5580035 rows
Read 91.8% of 5580035 rows
Read 5580035 rows and 24 (of 24) columns from 1.329 GB file in 00:00:17
```

```
#picking non-missing CMPLNT_FR_DT and convert to Date and filter only those after "2006-
01-01", 5560408 obs.
df%>%select(CMPLNT_FR_DT,LAW_CAT_CD)%>%filter(!is.na(CMPLNT_FR_DT))%>%mutate(CMPLNT_FR_D
T=as.Date(CMPLNT_FR_DT,format='%m/%d/%Y'))%>%filter(CMPLNT_FR_DT>=as.Date("2006-01-01"))
->df_Date
```

```
#time series of daily frequency of 3 crime categories 2006-2016
df_Date%>%group_by(CMPLNT_FR_DT,LAW_CAT_CD)%>%dplyr::summarise(count=n())%>%ungroup()%>%
group_by(LAW_CAT_CD)%>%mutate(mon_mean=rollmean(count,30,fill=NA))%>%ungroup()->byDateLa
wMean

#daily rate
byDateLawMean%>%ggplot()+
  geom_line(aes(CMPLNT_FR_DT,count,color=LAW_CAT_CD))+
  geom_line(aes(CMPLNT_FR_DT,mon_mean,group=LAW_CAT_CD))+
  ggtitle("Daily Crime Frequency since 2006 with 30-day running mean")+
  labs(x="Date",y="Frequency")+theme(legend.title=element_blank())+geom_line(aes(CMPLNT_
FR_DT,count*0+1150))
```




```
#Top 9 daily rate falls on Jan 1.
byDateLawMean%>%arrange(desc(count))%>%filter(count>=count[9])%>%
  mutate(CMPLNT_FR_DT=as.factor(CMPLNT_FR_DT))%>%
  ggplot(aes(forcats::fct_reorder(CMPLNT_FR_DT, count),count))+geom_bar(stat="identity")
+coord_flip()+ylab("Top 9 Daily Crime Frequency")+xlab("Date")
```



- 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.

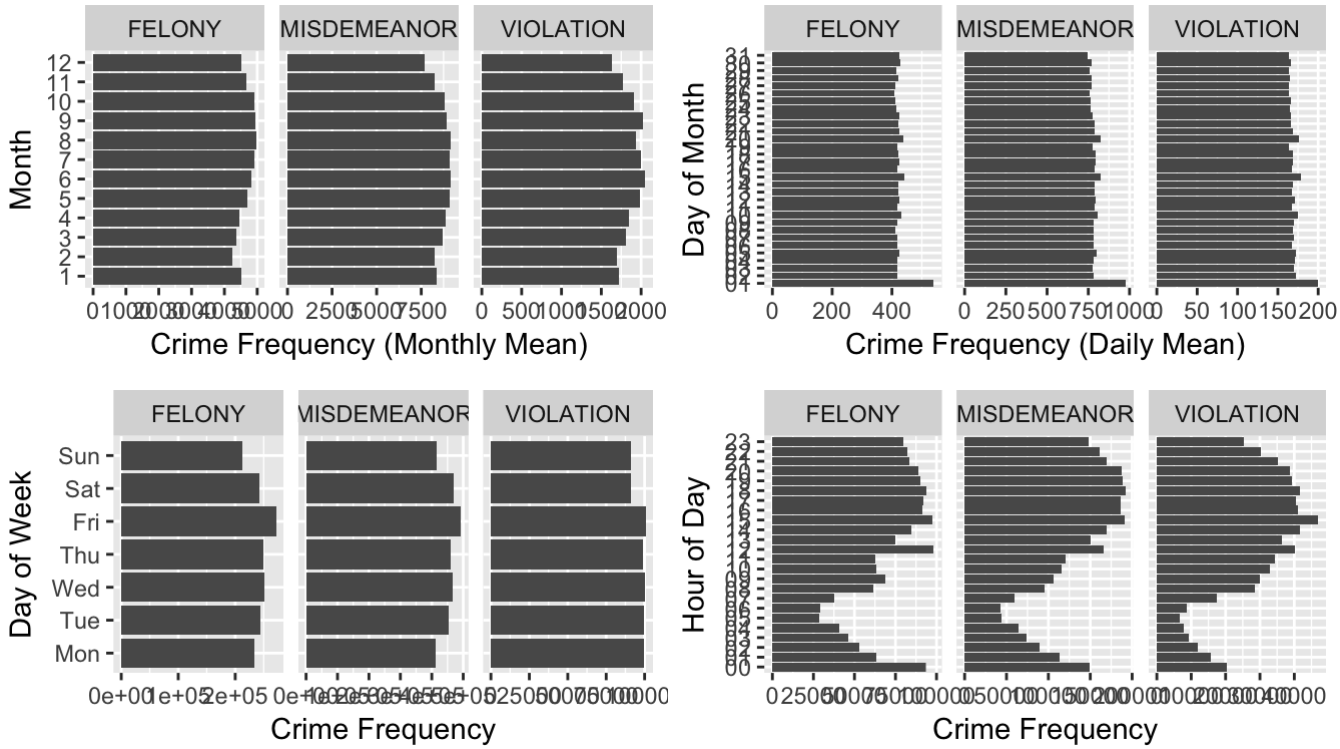
```
#frequency by month
df_Date%>%mutate(Month=as.character(month(CMPLNT_FR_DT)))%>%group_by(Month,LAW_CAT_CD)%
>%dplyr::summarise(CntByMon=n())->byDateLaw_mon

byDateLaw_mon%>%mutate(Days=rep(31,3))%>%mutate(Days=ifelse(Month=="2",28,Days))%>%mutat
e(Days=ifelse(Month %in% c("4","6","9","11"),30,Days))->byDateLaw_mon
byDateLaw_mon%>%ggplot(aes(fct_relevel(Month,"10","11","12",after=9),CntByMon/Days))+geo
m_bar(stat="identity")+coord_flip()+ylab("Crime Frequency (Monthly Mean)")+facet_wrap(~L
AW_CAT_CD,scales="free_x")+xlab("Month")->p1
```

```
#frequency by day
df_Date%>%mutate(Day=as.factor(format(CMPLNT_FR_DT,"%d")))%>%group_by(Day,LAW_CAT_CD)%>%
dplyr::summarise(CntByDay=n())->byDateLaw_day

#Day1-28 has the same total cnts=11yrs*12cnts/yr
#Day 29 cnts=11yrs*11cnts/yr+3cnts (leap yrs)
#Day 30 cnts=11*11; Day 31 cnts=7*11
byDateLaw_day%>%mutate(cnts=rep(12*11,3))%>%mutate(cnts=ifelse(Day=="29",11*11+3,cnts))%
>%mutate(cnts=ifelse(Day=="30",11*11,cnts))%>%mutate(cnts=ifelse(Day=="31",7*11,cnts))->
byDateLaw_day

byDateLaw_day%>%ggplot(aes(Day,CntByDay/cnts))+geom_bar(stat="identity")+coord_flip()+yl
ab("Crime Frequency (Daily Mean)")+facet_wrap(~LAW_CAT_CD,scales="free_x")+xlab("Day of
 Month")->p2
```

```
#frequency by weekday
df_Date%>%mutate(Wkday=as.factor(weekdays(CMPLNT_FR_DT,abbreviate=TRUE)))%>%group_by(Wkd
ay,LAW_CAT_CD)%>%dplyr::summarise(CntByWkday=n())->byDateLaw_wkday

byDateLaw_wkday%>%ggplot(aes(fct_relevel(Wkday,"Mon","Tue","Wed","Thu","Fri","Sat","Sun"
),CntByWkday))+geom_bar(stat="identity")+coord_flip()+ylab("Crime Frequency")+facet_wrap
(~LAW_CAT_CD,scales="free_x")+xlab("Day of Week")->p3
```

```
#picking non-missing CMPLNT_FR_TM
df%>%filter(!is.na(CMPLNT_FR_TM))%>%mutate(CMPLNT_FR_DT=as.Date(CMPLNT_FR_DT,format='%
m/%d/%Y'))%>%filter(CMPLNT_FR_DT>=as.Date("2006-01-01"))->df_FRTM

#Frequency by hour of day, combining hour 00 and hour 24 into hour 00
df_FRTM%>%mutate(Hour=as.factor(substr(CMPLNT_FR_TM,1,2)))%>%group_by(Hour,LAW_CAT_CD)%
>%dplyr::summarise(CntByHour=n())->byDateLaw_hour
byDateLaw_hour$Hour[byDateLaw_hour$Hour=="24"]<-"00"
byDateLaw_hour$Hour<-factor(byDateLaw_hour$Hour)

byDateLaw_hour%>%ggplot(aes(Hour,CntByHour))+geom_bar(stat="identity")+coord_flip()+ylab
("Crime Frequency")+facet_wrap(~LAW_CAT_CD,scales="free_x")+xlab("Hour of Day")->p4
```

```
grid.arrange(p1,p2,p3,p4,nrow=2)
```



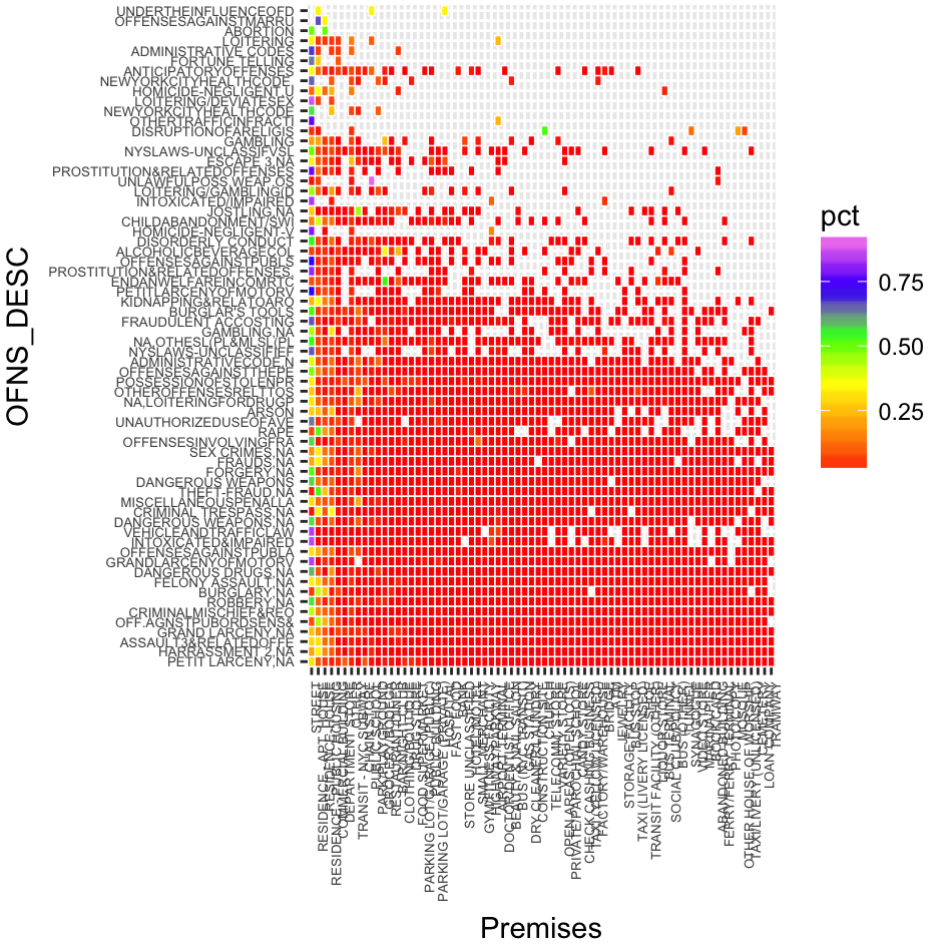
* Indeed by barcharting over the months,we see Jun.-Oct. is a high crime season. * The fake January increasing was due to the errors in the records.

* 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.

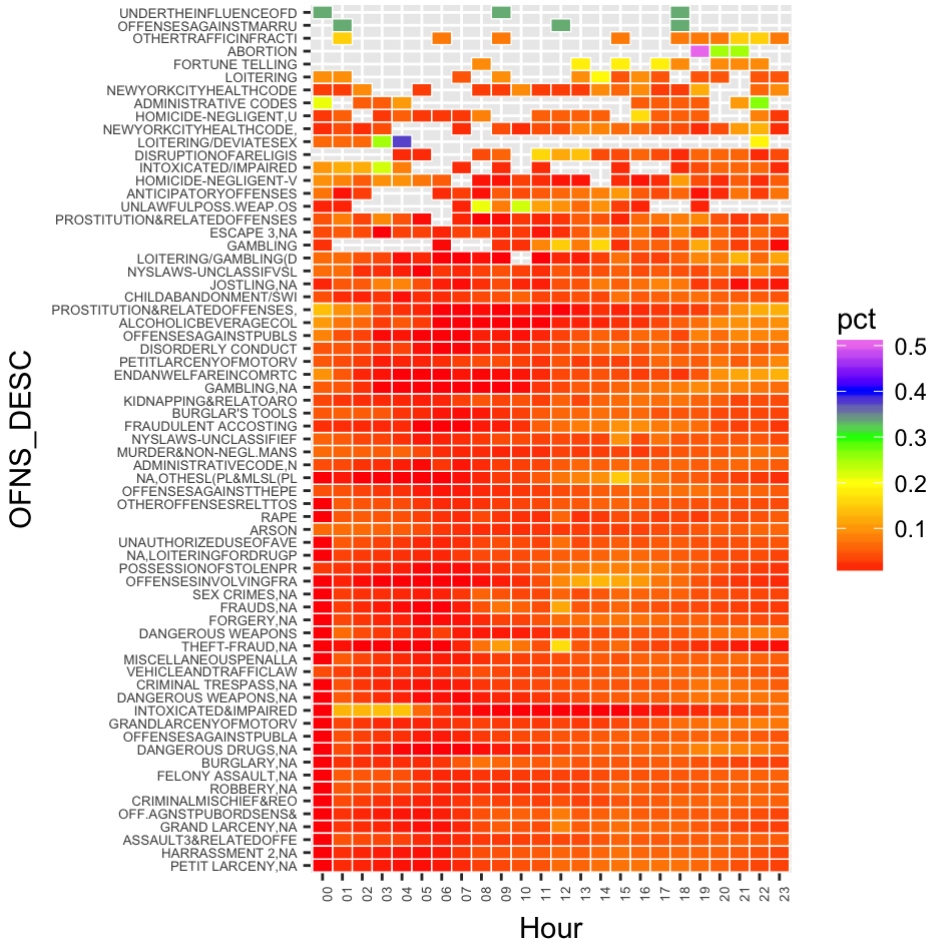
* There is obvious day cycle in the crime occurrence. Early morning has the least crime occurrence while later afternoon has the most crime occurrence.

```
#how the different crime types (OFNS_DESC) associated with different places (a heatmap)
#first filling the missing OFNS_DESC infered from KY_CD
df%>%select(KY_CD,OFNS_DESC)%>%group_by(KY_CD)%>%
  dplyr::summarise(desc=paste(unique(OFNS_DESC),collapse=","))%>%
  mutate(KY_CD=as.factor(KY_CD))%>%arrange(desc)->match_code_desc

df%>%
  select(KY_CD,PREM_TYP_DESC)%>%
  filter(!is.na(PREM_TYP_DESC))%>%
  mutate(KY_CD=as.factor(KY_CD))%>%
  group_by(KY_CD,PREM_TYP_DESC)%>%dplyr::summarise(count=n())%>%mutate(pct=count/sum(cou
nt))->byKYbyPREM

#merging to get OFNS_DESC vs PREM_TYP_DESC correspondence
merge(byKYbyPREM, match_code_desc, by.x='KY_CD', by.y='KY_CD')->byKYbyPREM_match

byKYbyPREM_match%>%group_by(desc)%>%dplyr::summarise(mean=mean(count),na.rm=TRUE)%>%arra
nge(desc(mean))->desc_desc_cnt
byKYbyPREM_match%>%group_by(PREM_TYP_DESC)%>%dplyr::summarise(mean=mean(count),na.rm=TRU
E)%>%arrange(desc(mean))->PREM_desc_cnt

byKYbyPREM_match%>%
  ggplot(aes(fct_relevel(as.factor(desc),as.character(desc_desc_cnt$desc[sort(desc_desc_
cnt$mean,index.return=TRUE,decreasing=TRUE)$ix])),
  fct_relevel(as.factor(PREM_TYP_DESC),as.character(PREM_desc_cnt$PREM_TYP_DESC[sort(PRE
M_desc_cnt$mean,index.return=TRUE,decreasing=TRUE)$ix])),fill=pct))+scale_fill_gradientn
(colors=c("red","orange","yellow","green","blue","violet"),na.value="black")+
  scale_x_discrete(label=function(x) abbreviate(x, minlength=20))+coord_flip()+
    geom_tile(color="white",size=0.25)+theme(axis.text.x = element_text(size=5,angle = 9
0, hjust = 1),axis.text.y=element_text(size=5))+ylab("Premises")+xlab("OFNS_DESC")
```



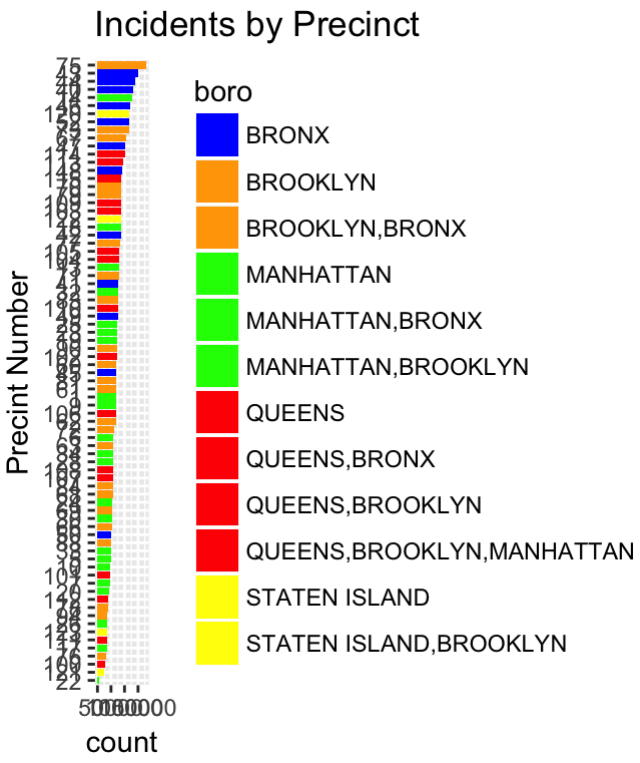
* Doesn't seem having association between crime types and premises.

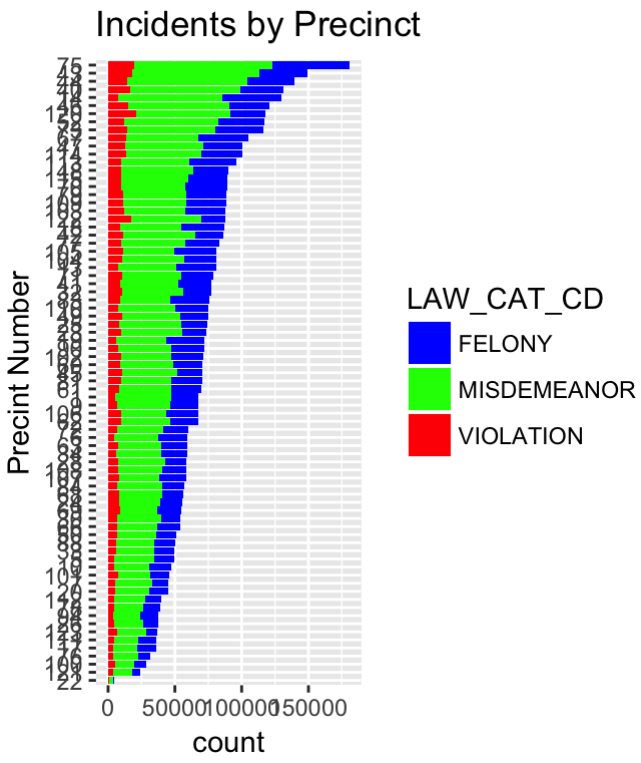
```
#how the different crime types associated with time using heatmap
df%>%
  select(KY_CD,CMPLNT_FR_TM)%>%
  filter(!is.na(CMPLNT_FR_TM))%>%
  mutate(KY_CD=as.factor(KY_CD))%>%
  mutate(Hour=as.factor(substr(CMPLNT_FR_TM,1,2)))%>%
  group_by(KY_CD,Hour)%>%dplyr::summarise(count=n())%>%mutate(pct=count/sum(count))->byK
YbyFRTM
#combining hour 00 and hour 24 into hour 00
byKYbyFRTM$Hour[byKYbyFRTM$Hour=="24"]<-"00"
byKYbyFRTM$Hour<-factor(byKYbyFRTM$Hour)

#merging to get OFNS_DESC vs CMPLNT_FR_TM correspondence
merge(byKYbyFRTM, match_code_desc, by.x='KY_CD', by.y='KY_CD')->byKYbyFRTM_match

byKYbyFRTM_match%>%group_by(desc)%>%dplyr::summarise(mean=mean(count),na.rm=TRUE)%>%arra
nge(desc(mean))->desc2_desc_cnt
byKYbyFRTM_match%>%group_by(Hour)%>%dplyr::summarise(mean=mean(count),na.rm=TRUE)%>%arra
nge(desc(mean))->Hour_desc_cnt

byKYbyFRTM_match%>%ggplot(aes(
  fct_relevel(as.factor(desc),as.character(desc2_desc_cnt$desc[sort(desc2_desc_cnt$mean,
index.return=TRUE,decreasing=TRUE)$ix])),
  Hour,fill=pct))+scale_fill_gradientn(colors=c("red","orange","yellow","green","blue",
"violet"),na.value="black")+
  scale_x_discrete(label=function(x) abbreviate(x, minlength=20))+coord_flip()+
    geom_tile(color="white",size=0.25)+theme(axis.text.x = element_text(size=5,angle = 9
0, hjust = 1),axis.text.y=element_text(size=5))+ylab("Hour")+xlab("OFNS_DESC")
```



* 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.

```
#matching Pct with Boro
df%>%select(ADDR_PCT_CD,BORO_NM)%>%filter(!is.na(ADDR_PCT_CD) & !is.na(BORO_NM))%>%group
_by(ADDR_PCT_CD)%>%
  dplyr::summarise(boro=paste(unique(BORO_NM),collapse=","))%>%
  mutate(ADDR_PCT_CD=as.factor(ADDR_PCT_CD))%>%arrange(boro)->match_pct_boro

df %>% select(LAW_CAT_CD,ADDR_PCT_CD)%>%group_by(LAW_CAT_CD,ADDR_PCT_CD)%>%
  drop_na()%>%dplyr::summarize(count = n())%>%ungroup()->df_pct
merge(df_pct,match_pct_boro,by.x="ADDR_PCT_CD",by.y="ADDR_PCT_CD")%>%arrange(desc(coun
t))->df_pbl

df_pbl%>%ggplot(aes(reorder(ADDR_PCT_CD, count), count,fill=boro)) + geom_bar(stat = "id
entity") + xlab("Precint Number") + ggtitle("Incidents by Precinct") + coord_flip()+scal
e_fill_manual(values = c("blue", rep("orange",2),rep("green",3),rep("red",4),rep("yello
w",2)))->p5
df_pbl%>%ggplot(aes(reorder(ADDR_PCT_CD, count), count,fill=LAW_CAT_CD)) + geom_bar(stat
 = "identity") + xlab("Precint Number") + ggtitle("Incidents by Precinct") + coord_flip
()+scale_fill_manual(values = c("blue", "green","red"))->p6
grid.arrange(p5,p6,nrow=1)
```





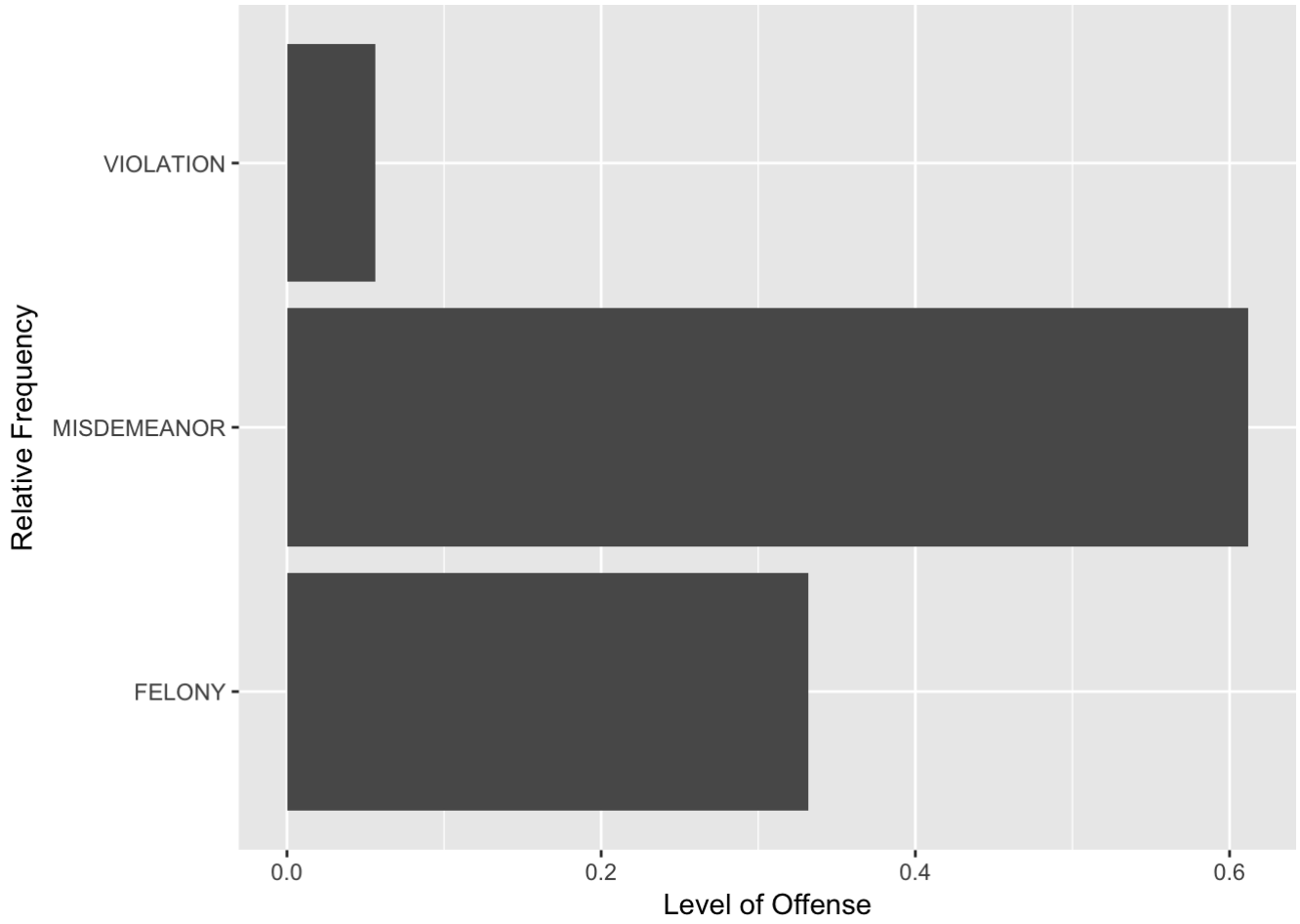
- Similar to Rich's precint plot. But precint number itself doesn't give meaningful information. We can add some meaningful information onto the plot by coloring in borough/location and crime types. Just to see which borough the precints with top crime rates are located, and frequency distribution of 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.

```
df %>% select(ADDR_PCT_CD,BORO_NM)%>%group_by(ADDR_PCT_CD,BORO_NM)%>%drop_na()%>%dplyr::
summarise(count=n())->tmp1
for (i in  1:77) {if(nrow(tmp1%>%filter(ADDR_PCT_CD==(unique(tmp1$ADDR_PCT_CD))[i]))>1)
{print(tmp1%>%filter(ADDR_PCT_CD==(unique(tmp1$ADDR_PCT_CD))[i]))}}
```

```
## # 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
```

```
df%>%filter(!is.na(PARKS_NM))->df_pk
df_pk%>%select(LAW_CAT_CD)%>%group_by(LAW_CAT_CD)%>%dplyr::summarise(count=n())%>%mutate
(RelFreq = count/sum(count))%>%ggplot(aes(LAW_CAT_CD,RelFreq))+geom_bar(stat="identity")
+
  coord_flip()+ylab("Level of Offense")+xlab("Relative Frequency")
```



* ~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.