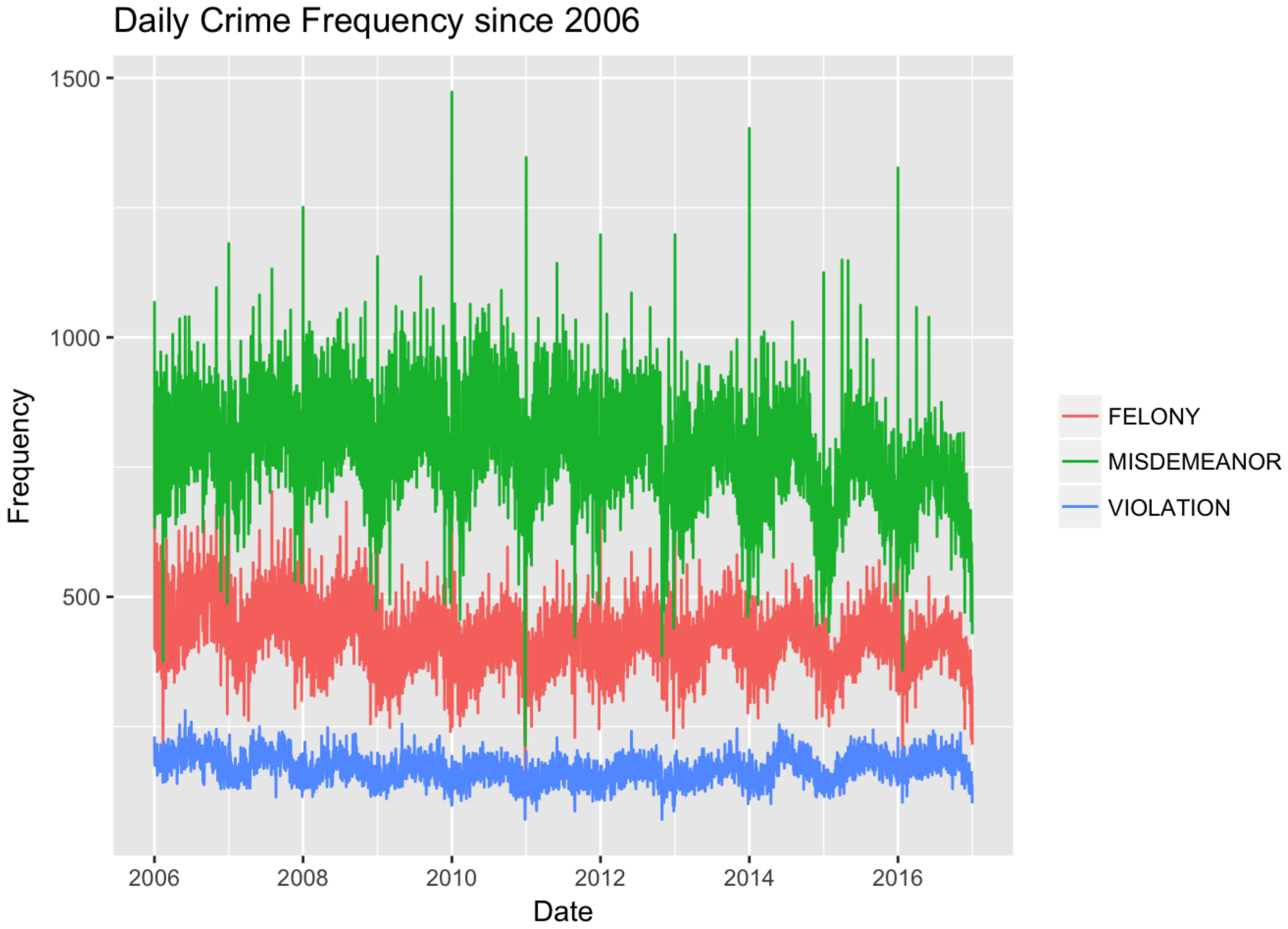
=====Part 4===== Main Analysis

```
#if you have not please install data.table package before run the codes below
#install.packages(data.table)
library(dplyr)
library(ggplot2)
library(forcats)
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 10.8% of 5580035 rows
Read 20.4% of 5580035 rows
Read 29.0% of 5580035 rows
Read 38.5% of 5580035 rows
Read 48.2% of 5580035 rows
Read 58.4% of 5580035 rows
Read 68.1% of 5580035 rows
Read 78.7% of 5580035 rows
Read 88.0% of 5580035 rows
Read 98.4% of 5580035 rows
Read 5580035 rows and 24 (of 24) columns from 1.329 GB file in 00:00:15
```

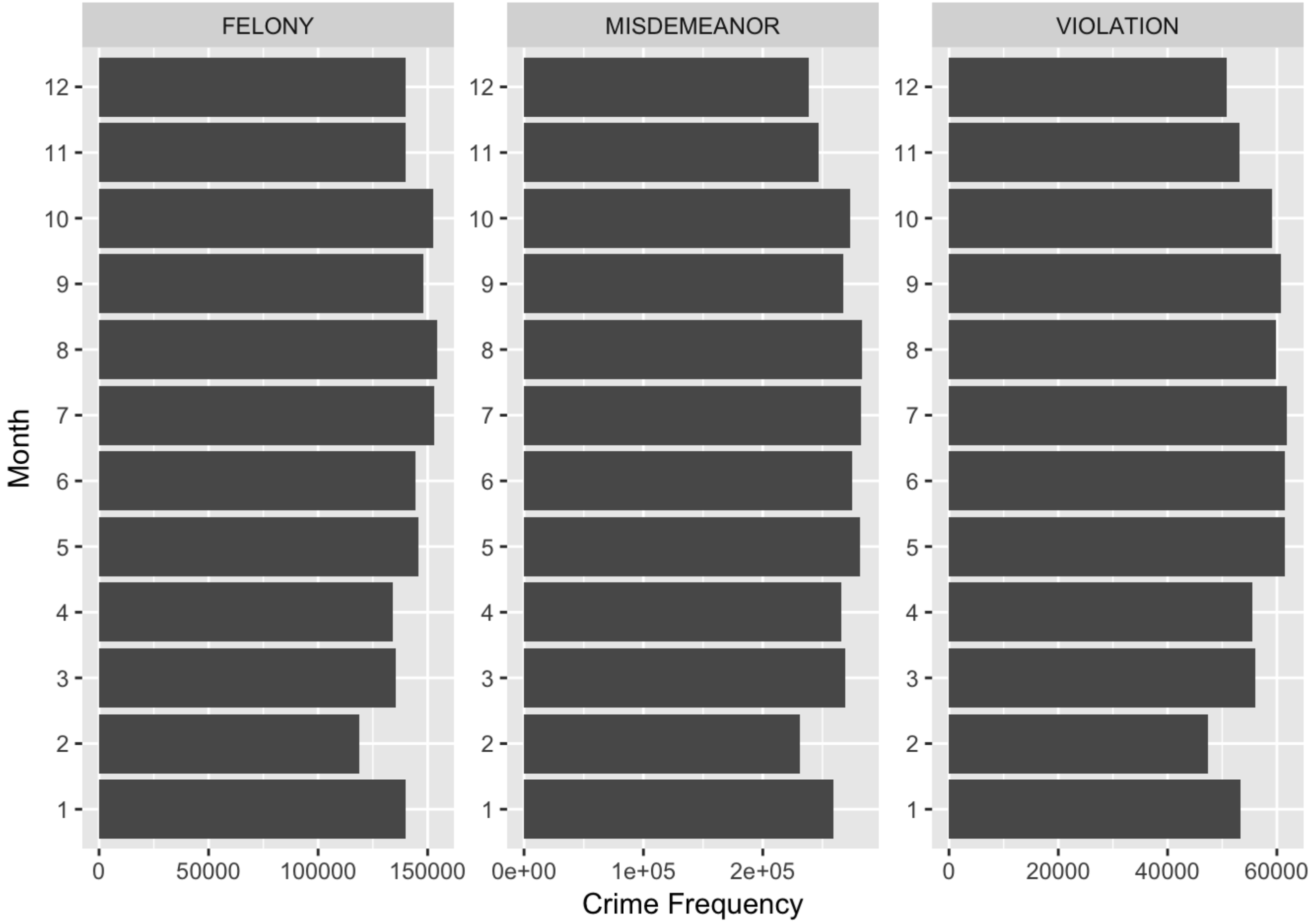
```
#picking non-missing CMPLNT_FR_DT and convert to Date and filter only those after "2006-
01-01", 5560408 obs.
df%>%select(CMPLNT_FR_DT,LAW_CAT_CD)%>%filter(!is.na(CMPLNT_FR_DT))%>%mutate(CMPLNT_FR_D
T=as.Date(CMPLNT_FR_DT,format='%m/%d/%Y'))%>%filter(CMPLNT_FR_DT>=as.Date("2006-01-01"))
->df_Date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

```
#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->byD
ateLaw

ggplot(byDateLaw,aes(CMPLNT_FR_DT,count,color=LAW_CAT_CD))+geom_line()+ggtitle("Daily Cr
ime Frequency since 2006")+labs(x="Date",y="Frequency")+theme(legend.title=element_blank
())
```



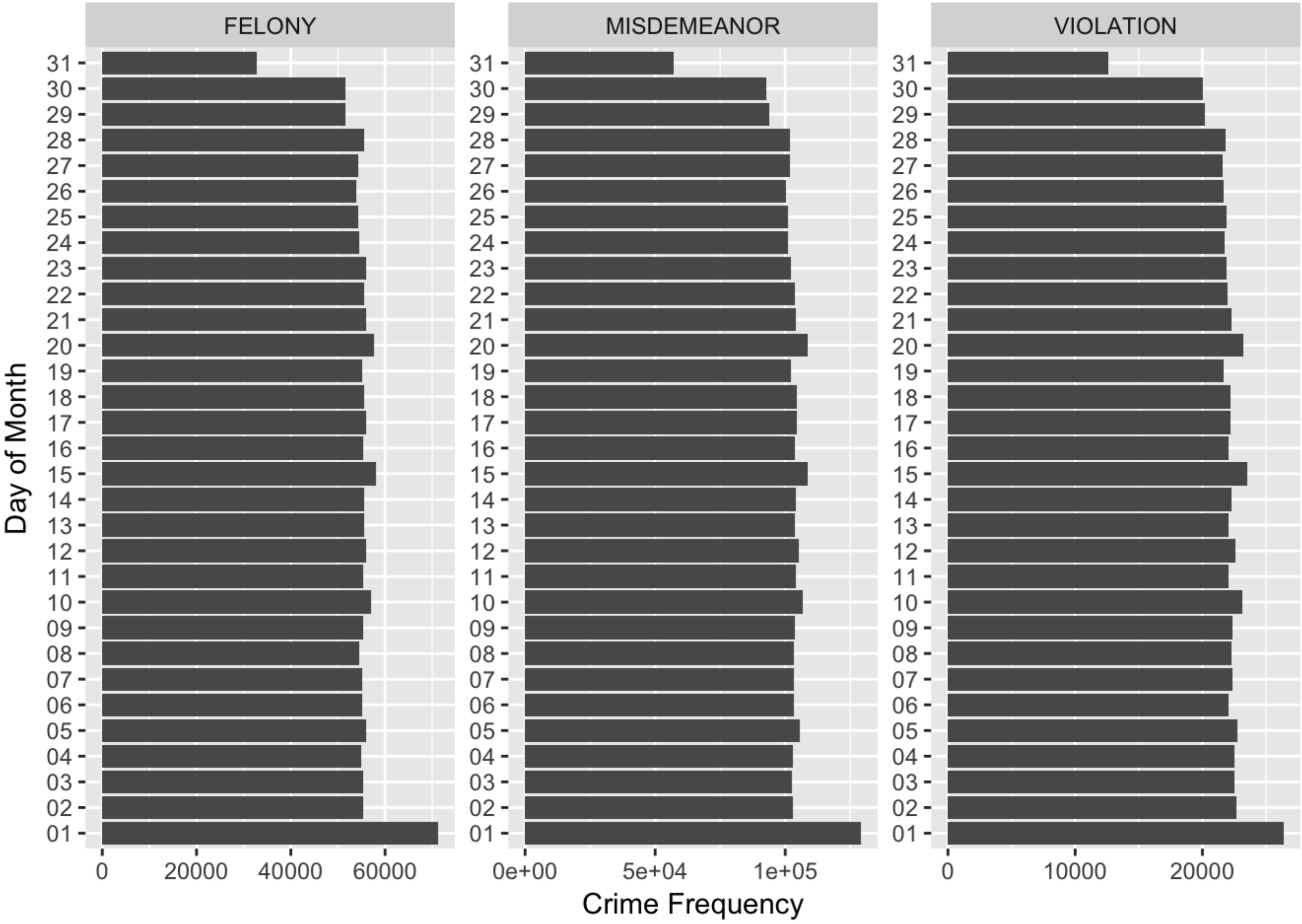
- The crime frequency is decreasing over the years.
- There are obvious annual variation/cycle.

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

byDateLaw_mon%>%ggplot(aes(fct_relevel(Month,"10","11","12",after=9),CntByMon))+geom_bar
(stat="identity")+coord_flip()+ylab("Crime Frequency")+facet_wrap(~LAW_CAT_CD,scales="fr
ee")+xlab("Month")
```



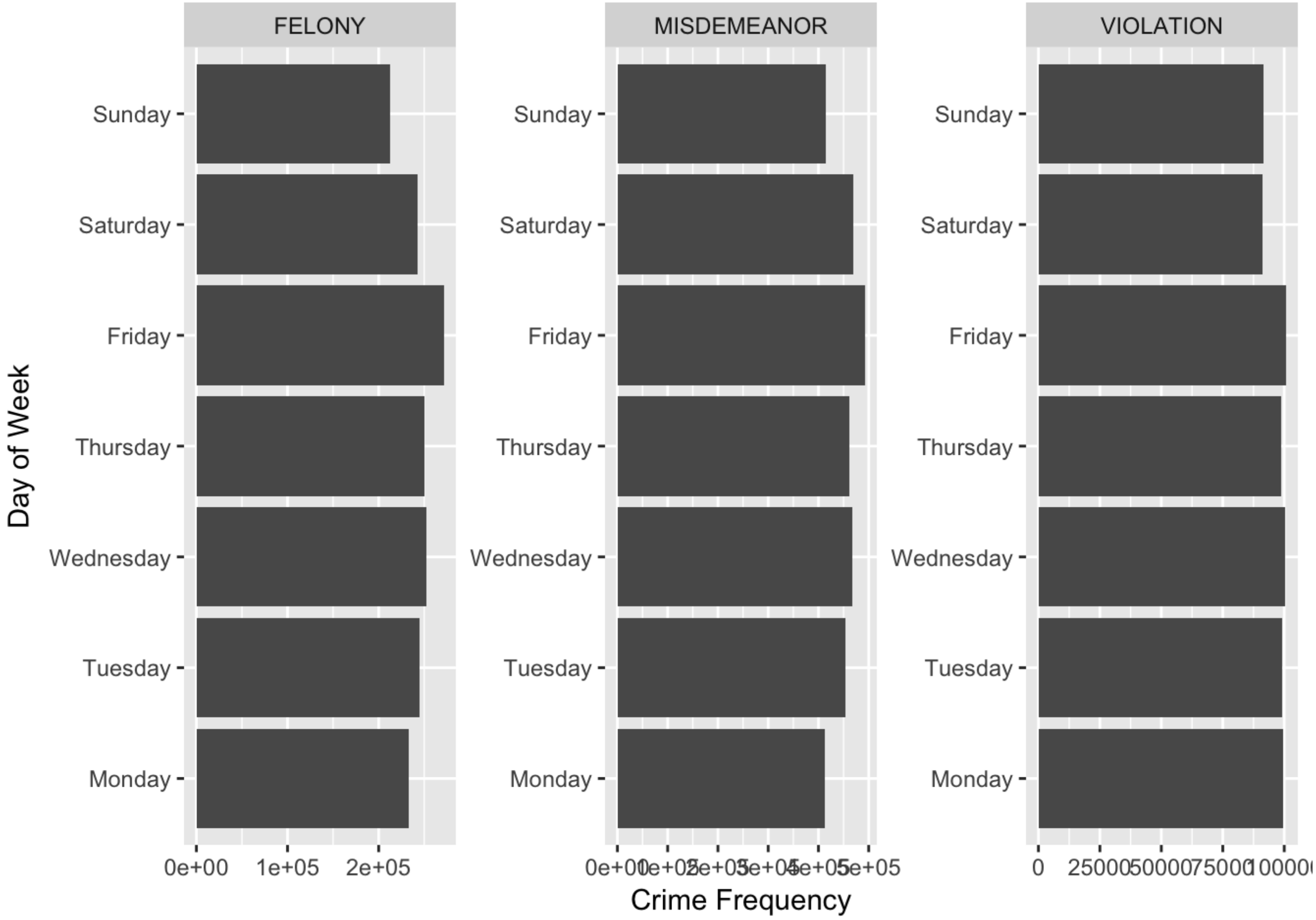
- Indeed by barcharting over the months,we see May-Oct. is a high crime season.
- One insteresting feature is January is a high peak during winter.

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

byDateLaw_day%>%ggplot(aes(Day,CntByDay))+geom_bar(stat="identity")+coord_flip()+ylab("C
rime Frequency")+facet_wrap(~LAW_CAT_CD,scales="free")+xlab("Day of Month")
```



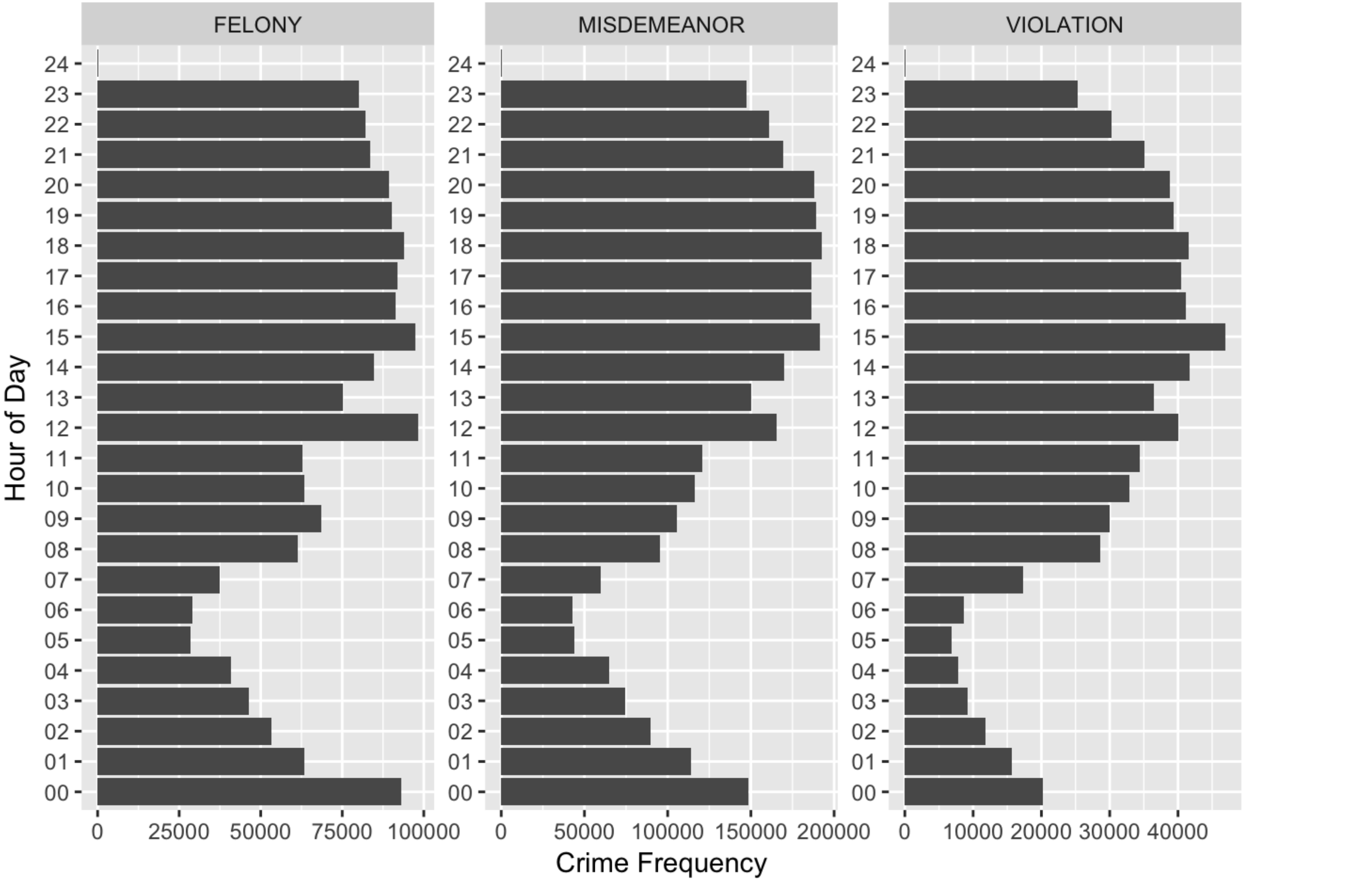
- Not much interesting feature. But end of month and beginning of month stands out. We have to explore.

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

byDateLaw_wkday%>%ggplot(aes(fct_relevel(Wkday,"Monday","Tuesday","Wednesday","Thursday"
,"Friday","Saturday","Sunday"),CntByWkday))+geom_bar(stat="identity")+coord_flip()+ylab(
"Crime Frequency")+facet_wrap(~LAW_CAT_CD,scales="free")+xlab("Day of Week")
```



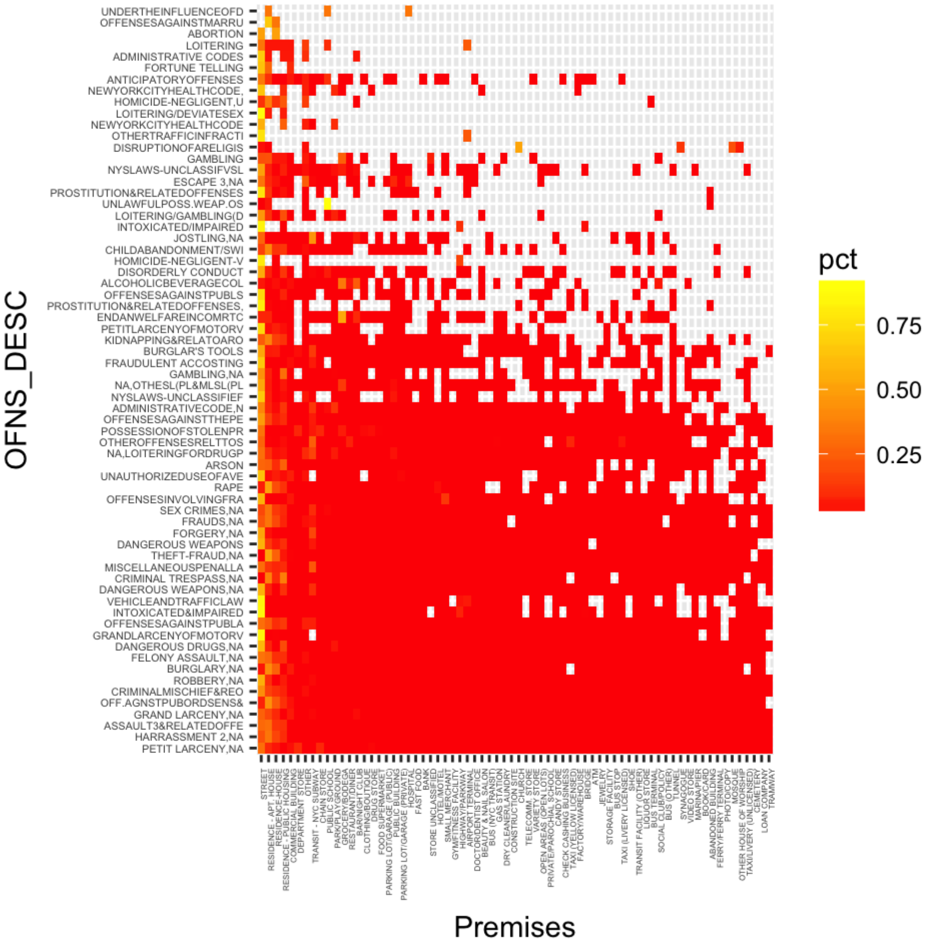
- Violation is low during weekends but same during weekdays.
- Felony and misdemeanor is high on Friday but low on Sunday nad Monday.

```
#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
df_FRTM%>%mutate(Hour=as.factor(substr(CMPLNT_FR_TM,1,2)))%>%group_by(Hour,LAW_CAT_CD)%
>%summarise(CntByHour=n())->byDateLaw_hour

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



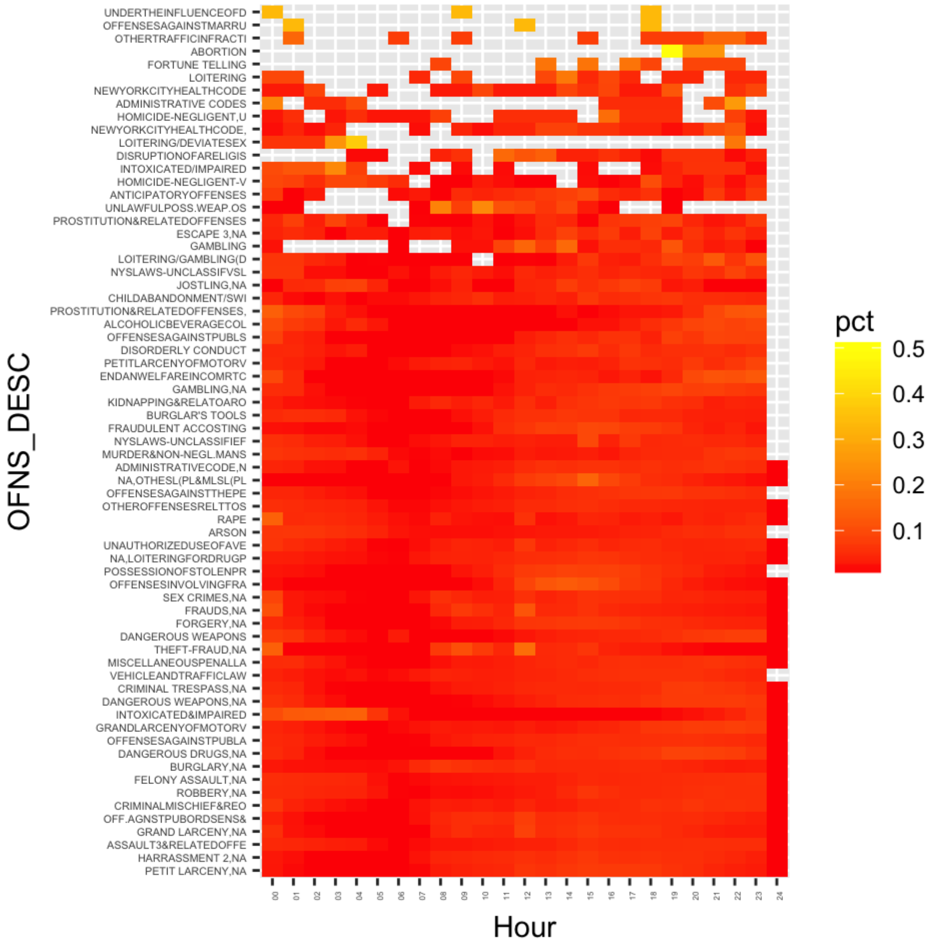
- There is obvious day cycle in the crime occurrence. Early morning has the least crime occurrence while later afternoon has the most crime occurrence.

```
#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)%>%
  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)%>%summarise(count=n())%>%mutate(pct=count/sum(count))->b
yKYbyPREM

#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,in
dex.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"),na.value="blue")+
  scale_x_discrete(label=function(x) abbreviate(x, minlength=20))+coord_flip()+
    geom_tile()+theme(axis.text.x = element_text(size=3,angle = 90, hjust = 1),axis.tex
t.y=element_text(size=4))+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)%>%summarise(count=n())%>%mutate(pct=count/sum(count))->byKYbyFRTM

#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"),na.value="blue"
)+
  scale_x_discrete(label=function(x) abbreviate(x, minlength=20))+coord_flip()+
    geom_tile()+theme(axis.text.x = element_text(size=3,angle = 90, hjust = 1),axis.tex
t.y=element_text(size=4))+ylab("Hour")+xlab("OFNS_DESC")
```



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