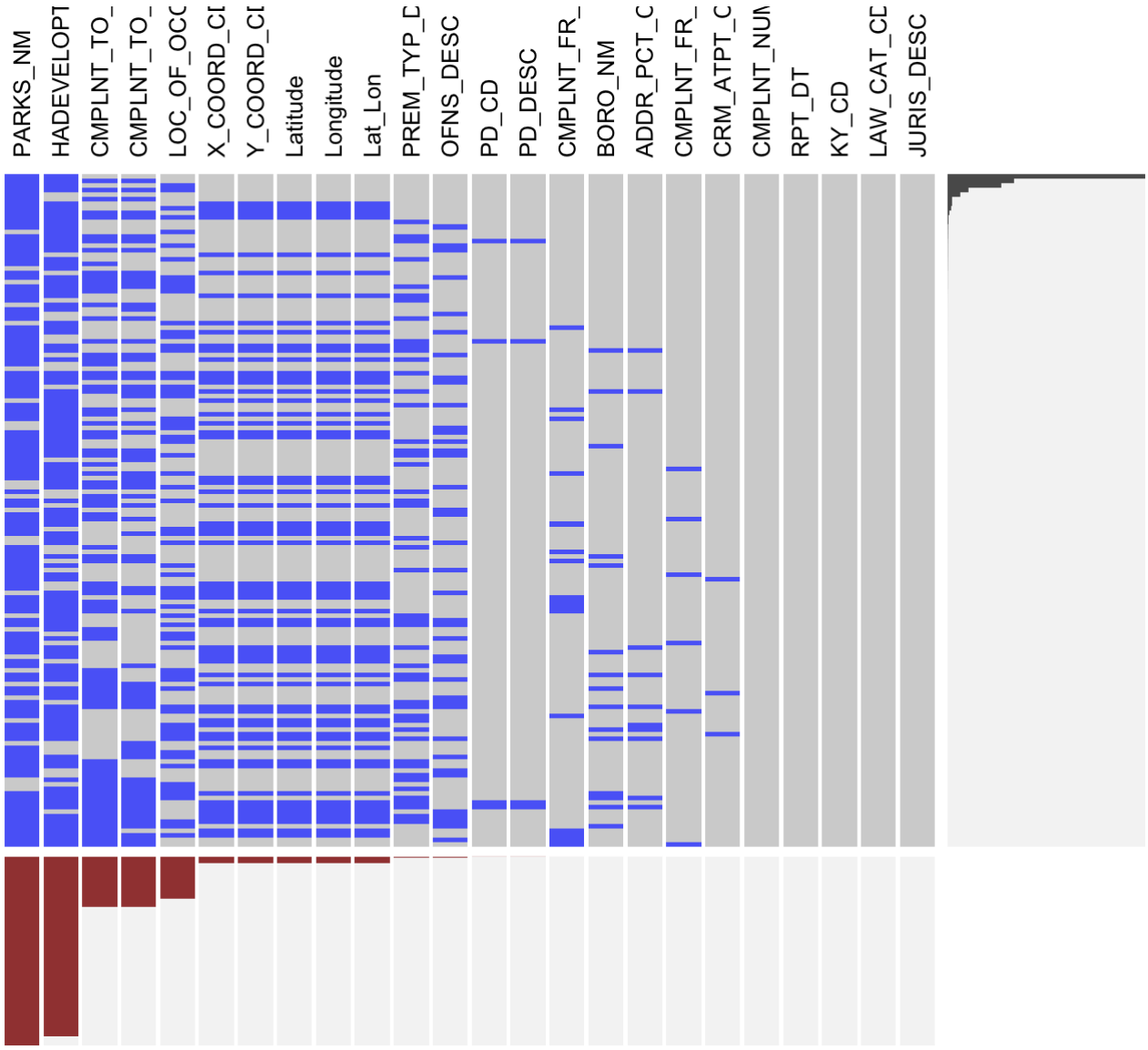
draft

```
#if you have not please install data.table package before run the codes below
#install.packages(data.table)
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 21.0% of 5580035 rows
Read 31.0% of 5580035 rows
Read 40.9% of 5580035 rows
Read 51.1% of 5580035 rows
Read 61.8% of 5580035 rows
Read 73.5% of 5580035 rows
Read 84.9% of 5580035 rows
Read 96.4% of 5580035 rows
Read 5580035 rows and 24 (of 24) columns from 1.329 GB file in 00:00:16
```

```
library(vcdExtra)
library(extracat)
#visna(df,sort="r")
visna(df,sort="b")
```

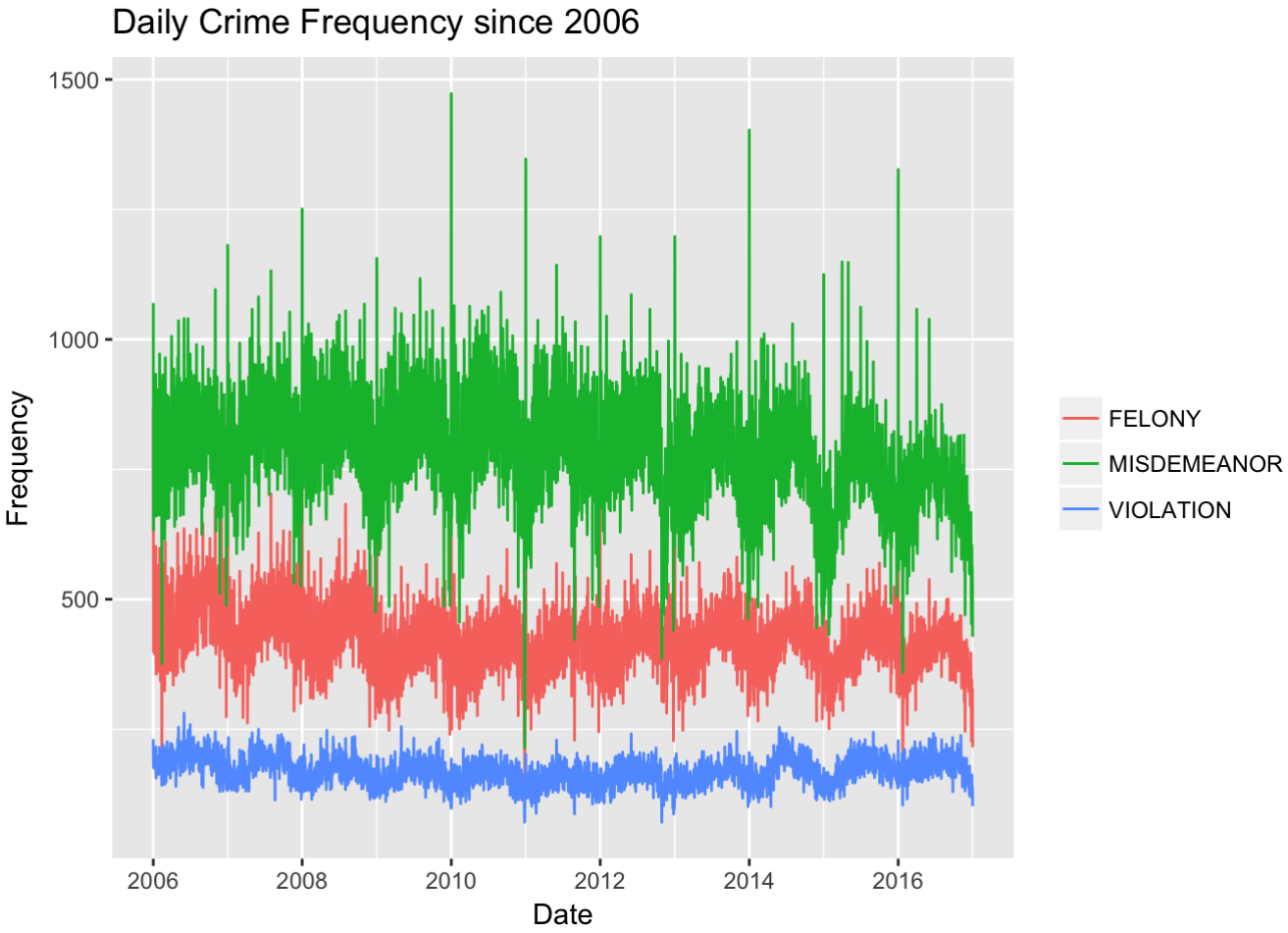


- Now in the missing pattern, the positoning variables are consistent with each other. (X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lat_lon). But the missing in CMPLNT_FR_DT,CMPLNT_FR_TM shows may need further look.

```
#Show missing count and percentage, you can uncomment it if you like to see the statisti
cs.

#for (i in 1:24) message(format(colnames(df)[i],justify="right",width=20),"\t",format(su
m(is.na(dplyr::select(df,i))),digits=7),"\t",sum(is.na(dplyr::select(df,i)))*100/nrow(d
f))
```

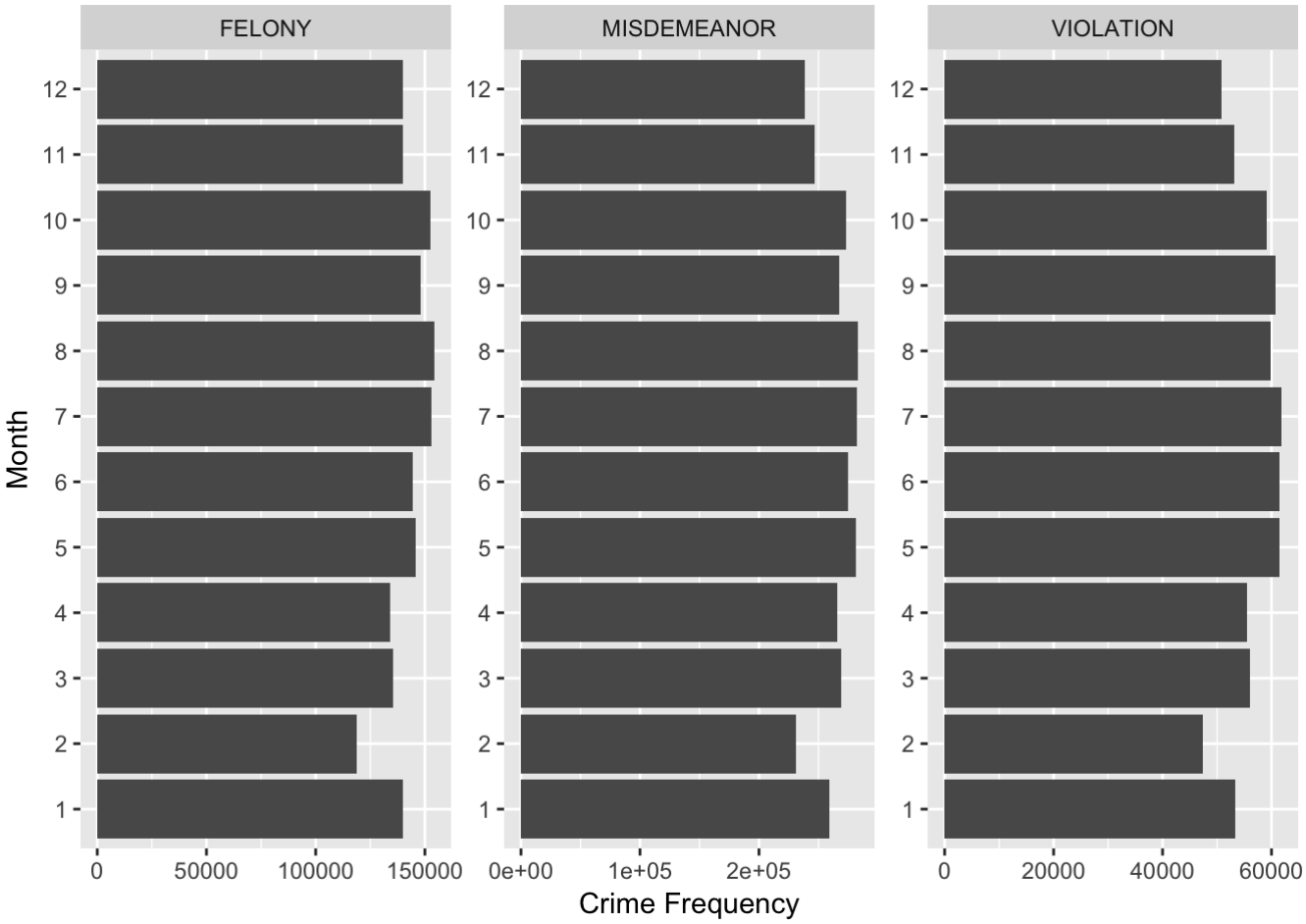
```
library(dplyr)
#picking non-missing CMPLNT_FR_DT and convert to Date and filter only those after "2006-
01-01", 5560408 obs.
df%>%filter(!is.na(CMPLNT_FR_DT))%>%mutate(CMPLNT_FR_DT=as.Date(CMPLNT_FR_DT,format='%
m/%d/%Y'))%>%filter(CMPLNT_FR_DT>=as.Date("2006-01-01"))->df_Date
```

```
library(ggplot2)
#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->byDateLaw

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



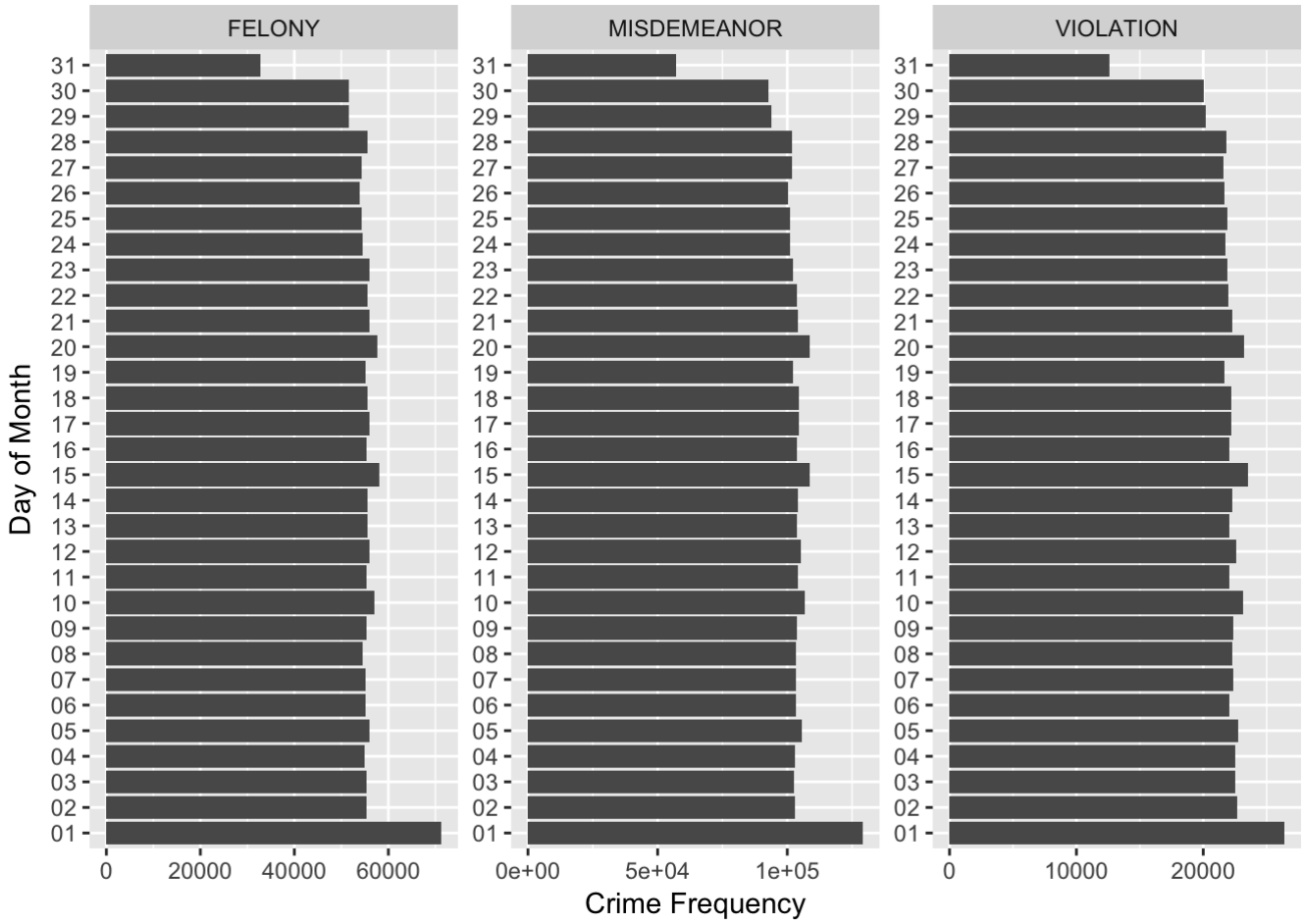
- The crime frequency is decreasing over the years.
- There are obvious annual variation/cycle.

```
library(forcats)
#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="free")+xlab("Month")
```



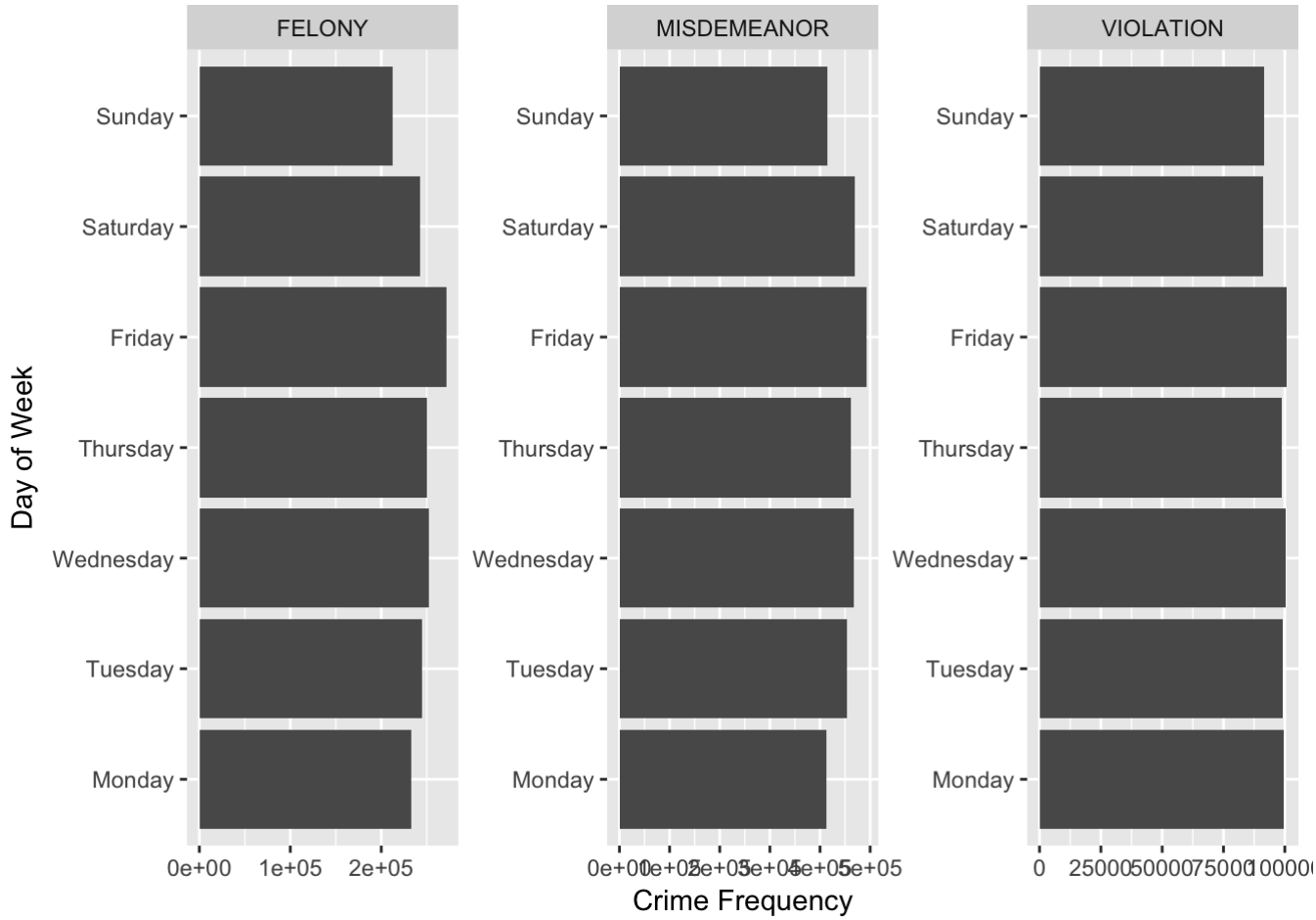
- Indeed by barcharting over the months,we see May-Oct. is a high crime season.
- One insteresting feature is January is a high peak during winter.

```
#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")
```



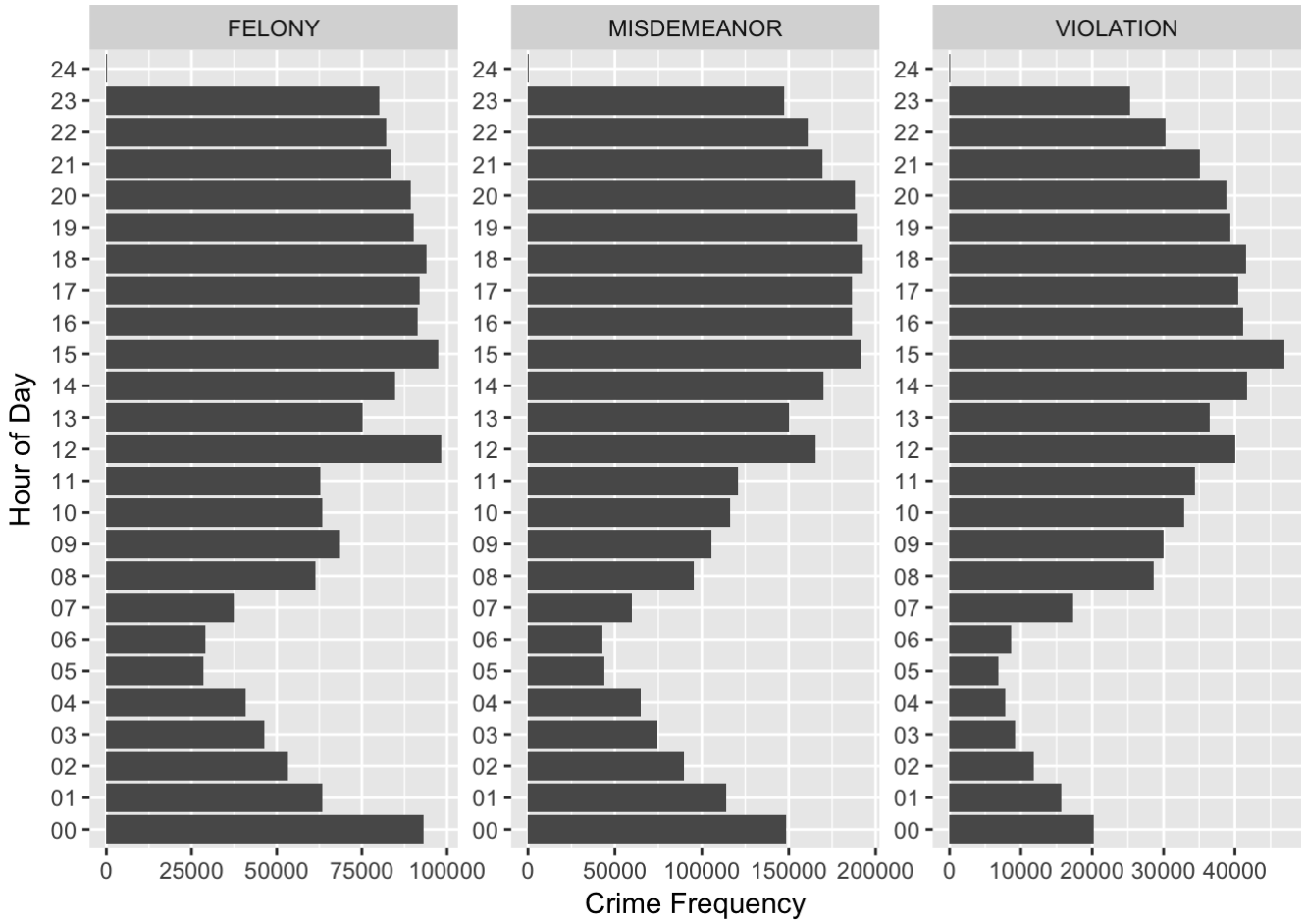
- Not much interesting feature. But end of month and beginning of month stands out. We have to explore.

```
#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.