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

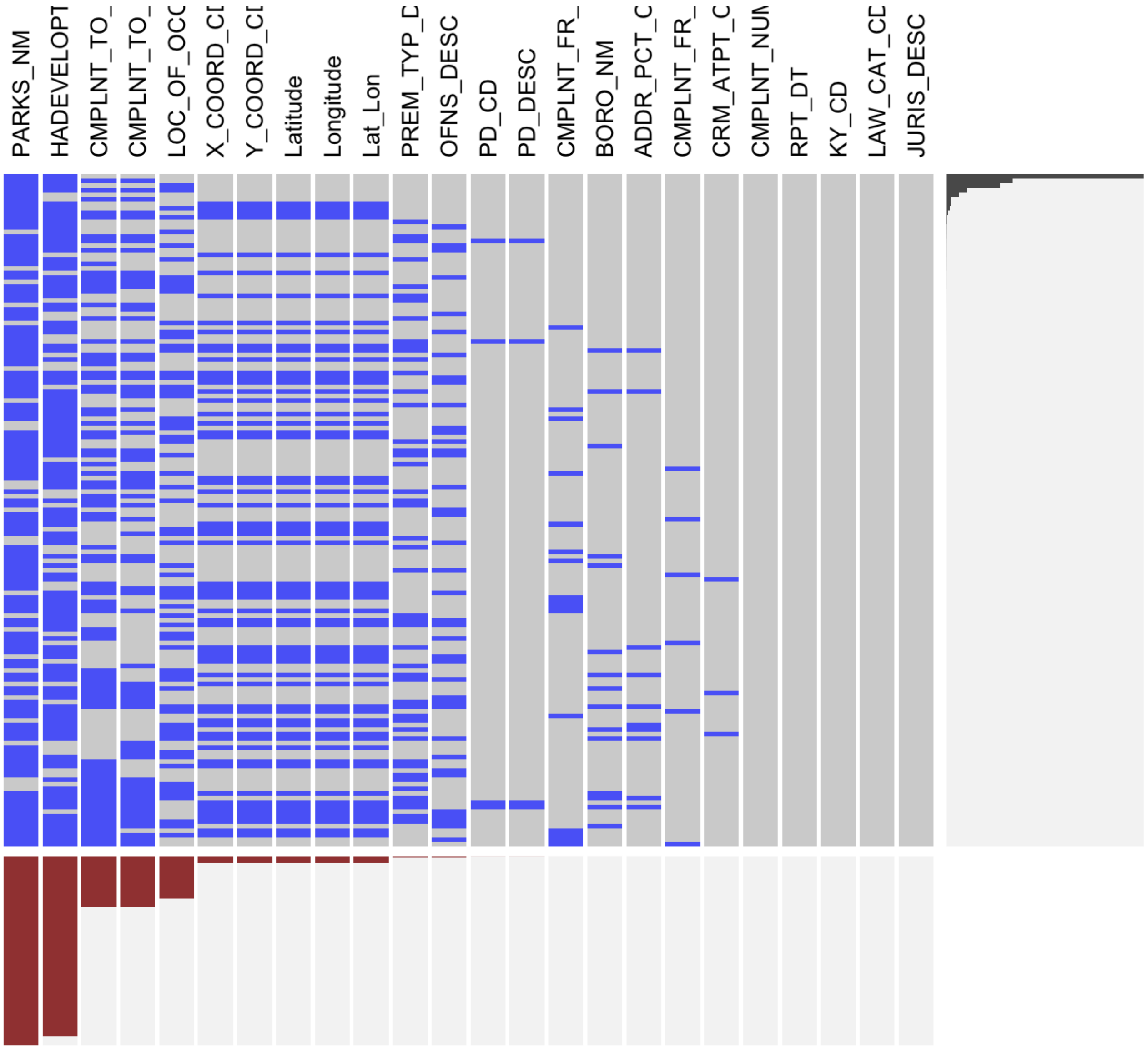
```
#install.packages(data.table)
library(dplyr)
library(tibble)
library(lattice)
library(ggplot2)
library(extracat)
library(gridExtra)
library(data.table)
fread("NYPD_Complaint_Data_Historic.csv",na.strings="",colClasses = c(PARKS_NM="c",HADEV
ELOPT="c"))->df
```

```
##
Read 0.0% of 5580035 rows
Read 10.8% of 5580035 rows
Read 21.0% of 5580035 rows
Read 30.8% of 5580035 rows
Read 40.7% of 5580035 rows
Read 52.0% of 5580035 rows
Read 63.3% of 5580035 rows
Read 74.6% of 5580035 rows
Read 85.8% of 5580035 rows
Read 97.1% 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.

```
visna(df,sort="b")
```



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

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

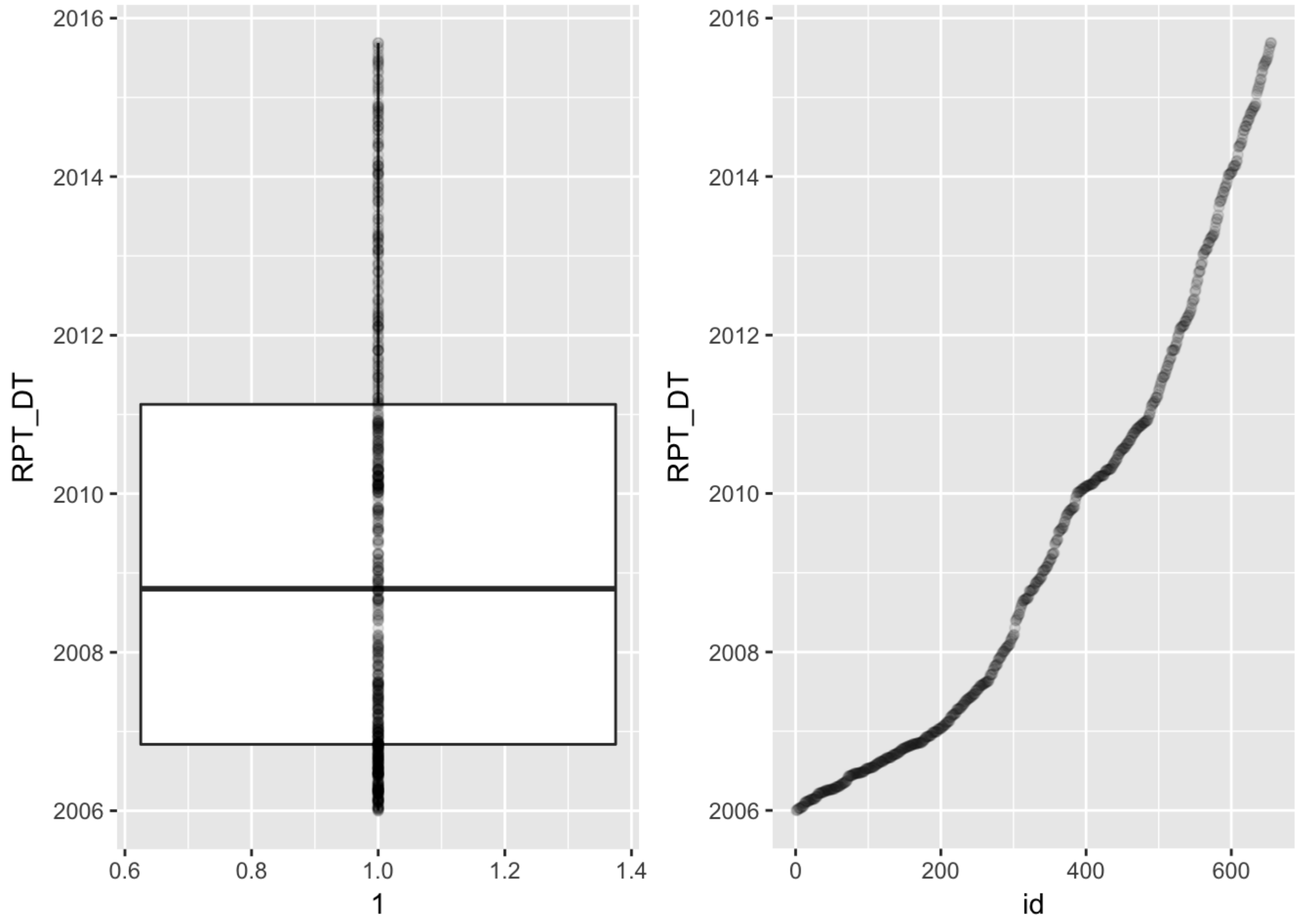
===Missing in CMPLNT_FR_DT===

There are total of 655 complaints missing CMPLNT_FR_DT (the exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists)).

1. Are they evenly/randomly distributed over the RPT_DT range 1/1/2006-12/31/2016 of the dataset?

- For cases missing CMPLNT_FR_DT, although there is a slight tendency of more cases at the beginning around 2006 and less at the ending arounf 2016, the reporting dates still look pretty even over the period suggesting randomness of the missing.

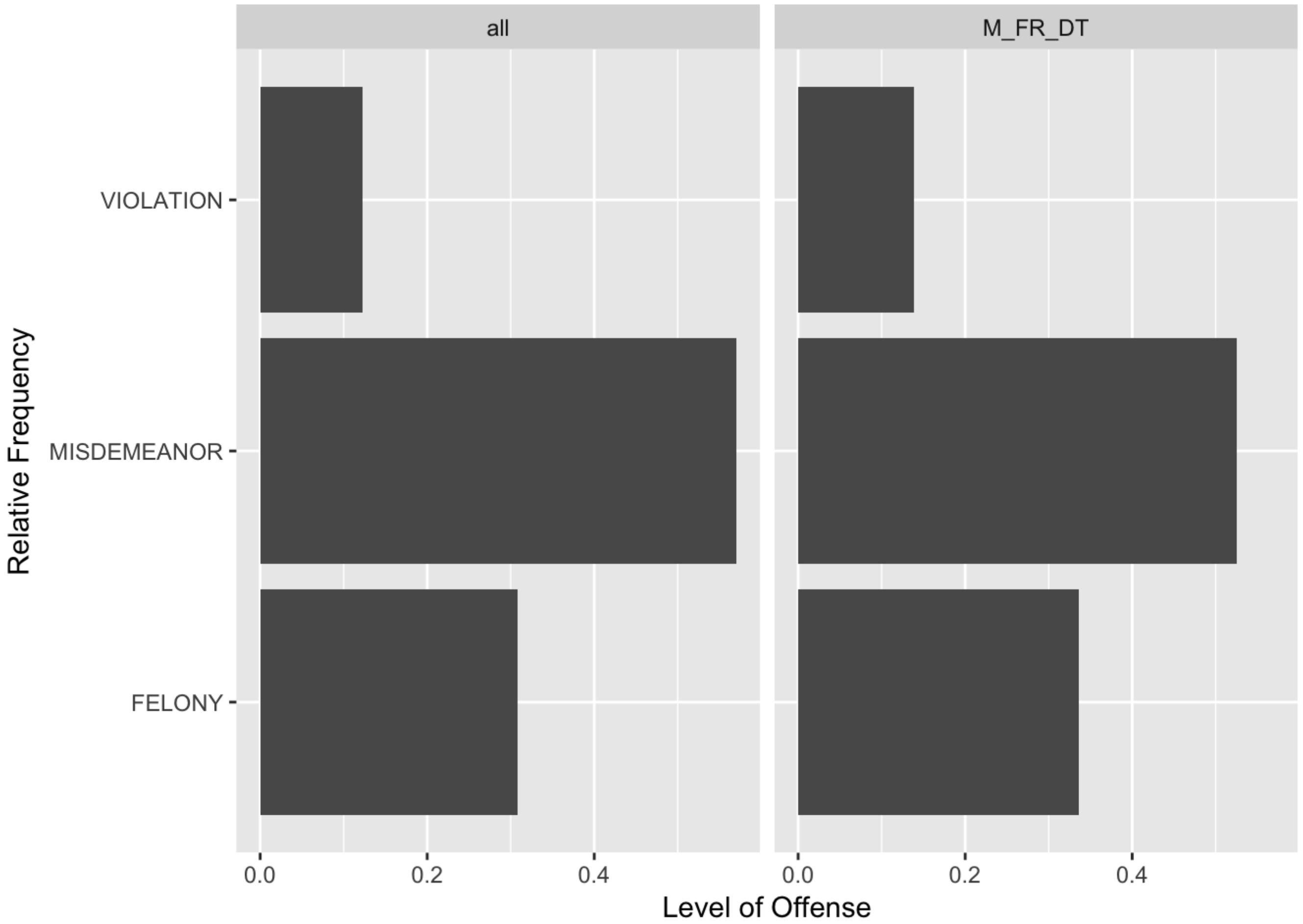
```
#get the reporting dates of cases with starting dates missing
df%>%
  select(CMPLNT_FR_DT,RPT_DT)%>%
  filter(is.na(CMPLNT_FR_DT))%>%
  select(RPT_DT)%>%
  mutate(RPT_DT=as.Date(RPT_DT,format='%m/%d/%Y'))->tmp1

#boxplot of cases with points overlayed
tmp1%>%ggplot()+geom_boxplot(aes(x=1,y=RPT_DT))+geom_point(aes(x=1,y=RPT_DT),alpha=0.1)-
>p1

#scatterplot to have better view of reporting date distribution over cases missing endin
g dates
tmp1%>%arrange(RPT_DT)%>%rownames_to_column(var="id")%>%
  mutate(id=as.numeric(id))%>%arrange(id)%>%
  ggplot()+ geom_point(aes(RPT_DT, y=id),alpha=0.1)+coord_flip()->p2
grid.arrange(p1,p2,nrow=1)
```



2. Are they randomly distributed over LAW_CAT_CD?

- The frequency distrinution of LAW_CAT_CD from the complaints missing CMPLNT_FR_DT shares the same pattern of that from all complaints in the dataset.

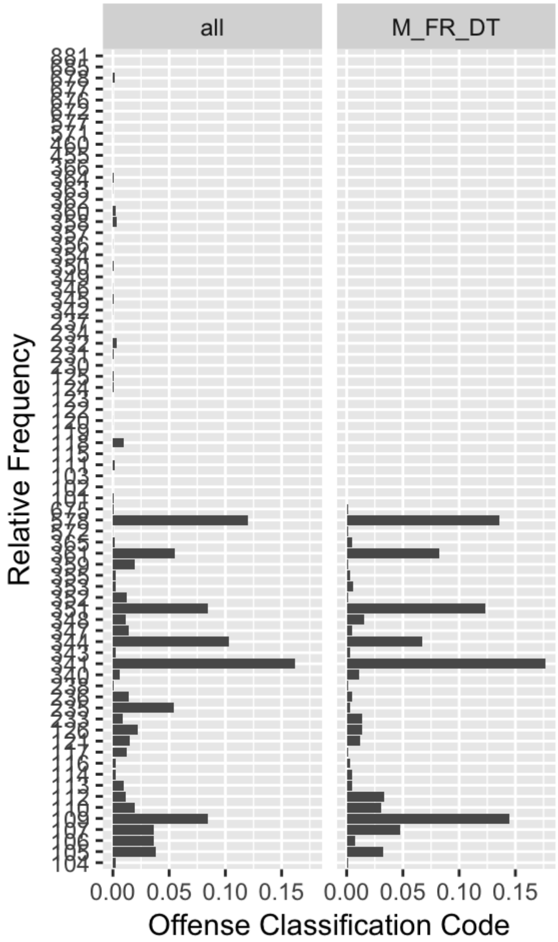
```
df%>%
  select(CMPLNT_FR_DT,LAW_CAT_CD)%>%
  filter(is.na(CMPLNT_FR_DT))%>%
  select(LAW_CAT_CD)%>%
  mutate(LAW_CAT_CD=as.factor(LAW_CAT_CD))%>%
  group_by(LAW_CAT_CD)%>%summarise(count=n())%>%
  mutate(RelFreq = count/sum(count))->tmp3; nr1=nrow(tmp3)
tmp3%>%mutate(type=replicate(nr1,"M_FR_DT"))->tmp3

df%>%
  select(LAW_CAT_CD)%>%
  mutate(LAW_CAT_CD=as.factor(LAW_CAT_CD))%>%
  group_by(LAW_CAT_CD)%>%summarise(count=n())%>%
  mutate(RelFreq = count/sum(count))->tmp5; nr2=nrow(tmp5)
tmp5%>%mutate(type=replicate(nr2,"all"))->tmp5
rbind(tmp3,tmp5)->tmp3tmp5

tmp3tmp5%>%ggplot(aes(LAW_CAT_CD,RelFreq))+geom_bar(stat="identity")+
  coord_flip()+ylab("Level of Offense")+xlab("Relative Frequency")+facet_wrap(~type)
```



3. Are they randomly distributed over KY_CD?

- The frequency distrinution of KY_CD from the complaints missing CMPLNT_FR_DT shares the same pattern of that from all complaints in the dataset.

```
df%>%
  select(CMPLNT_FR_DT,KY_CD)%>%
  filter(is.na(CMPLNT_FR_DT))%>%
  select(KY_CD)%>%
  mutate(KY_CD=as.factor(KY_CD))%>%
  group_by(KY_CD)%>%summarise(count=n())%>%
  mutate(RelFreq = count/sum(count))->tmp2; nrr1=nrow(tmp2)
tmp2%>%mutate(type=replicate(nrr1,"M_FR_DT"))->tmp2

df%>%
  select(KY_CD)%>%
  mutate(KY_CD=as.factor(KY_CD))%>%
  group_by(KY_CD)%>%summarise(count=n())%>%
  mutate(RelFreq = count/sum(count))->tmp4; nrr2=nrow(tmp4)
tmp4%>%mutate(type=replicate(nrr2,"all"))->tmp4
rbind(tmp2,tmp4)->tmp2tmp4

tmp2tmp4%>%ggplot(aes(KY_CD,RelFreq),na.rm=FALSE)+geom_bar(stat="identity")+
  coord_flip()+ylab("Offense Classification Code")+xlab("Relative Frequency")+facet_wrap
(~type)
```



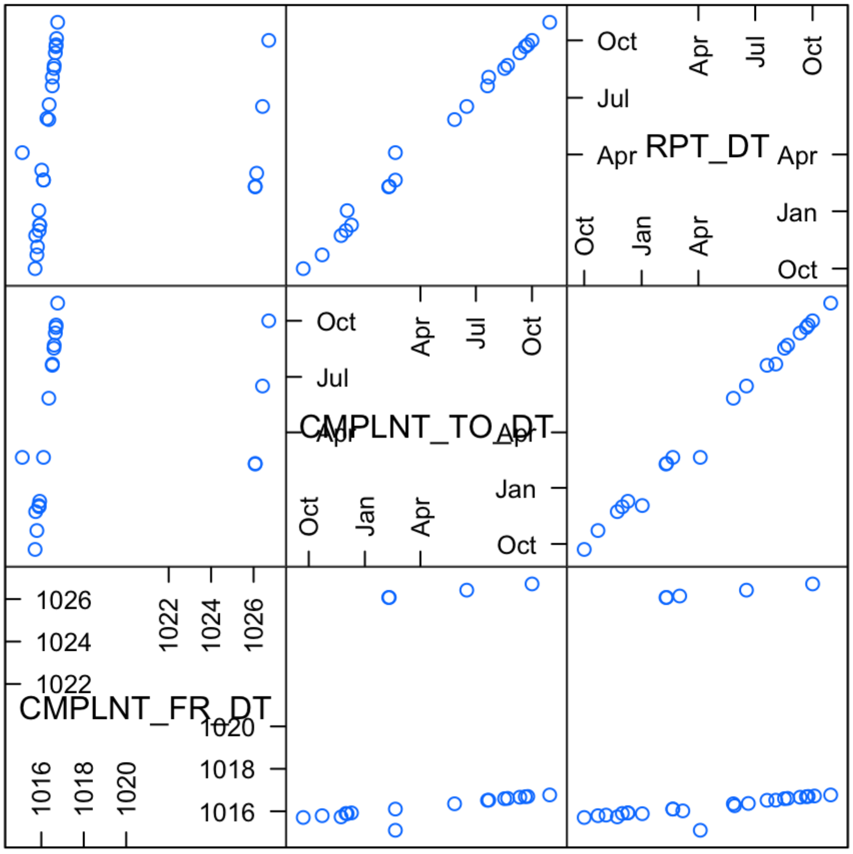
===Errors in CMPLNT_FR_DT===
There seems to be errors in CMPLNT_FR_DT.

- CMPLNT_FR_DT 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.

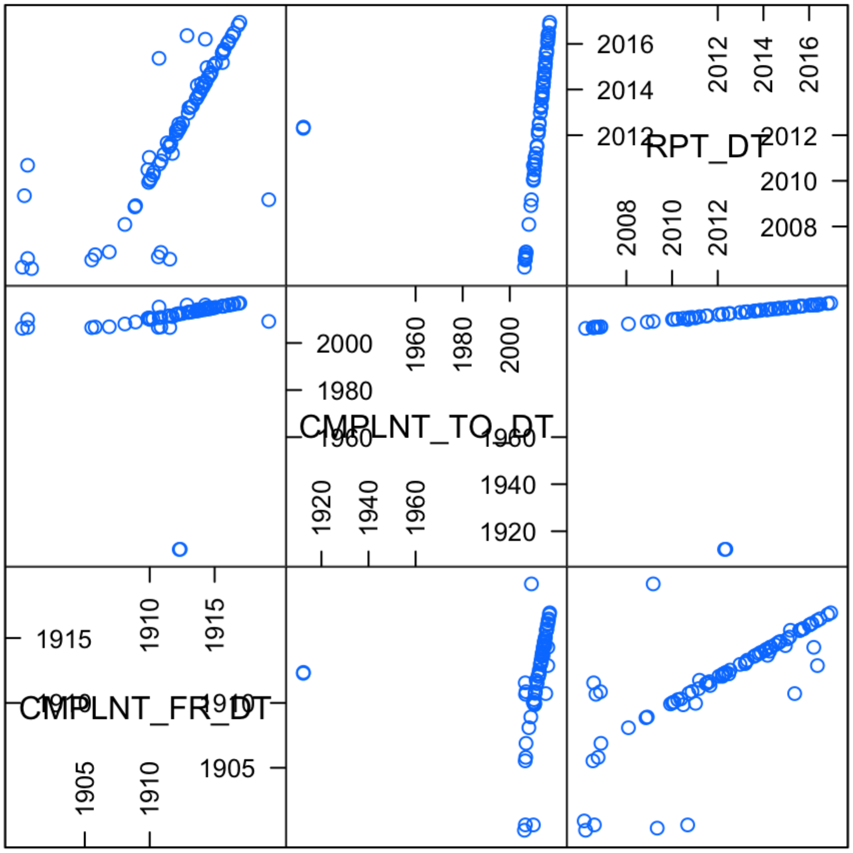
- The scatterplot of the CMPLNT_FR_DT vs RPT_DT did show some strict linear correlation for many cases during some periods.
-

```
df%>%select(CMPLNT_FR_DT,CMPLNT_TO_DT,RPT_DT)%>%
  mutate(CMPLNT_FR_DT=as.Date(CMPLNT_FR_DT,format='%m/%d/%Y'),
         CMPLNT_TO_DT=as.Date(CMPLNT_TO_DT,format='%m/%d/%Y'),
         RPT_DT=as.Date(RPT_DT,format='%m/%d/%Y'))->df_3DT

df_3DT%>%filter(CMPLNT_FR_DT<=as.Date("1900-01-01"))->df_3DT_Year1900
df_3DT%>%filter(CMPLNT_FR_DT>=as.Date("1900-01-01") & CMPLNT_FR_DT<=as.Date("1920-01-01"
))->df_3DT_Year1900to1920
df_3DT%>%filter(CMPLNT_FR_DT>=as.Date("1920-01-01") & CMPLNT_FR_DT<=as.Date("1960-01-01"
))->df_3DT_Year1920to1960
df_3DT%>%filter(CMPLNT_FR_DT>=as.Date("1960-01-01") & CMPLNT_FR_DT<=as.Date("1980-01-01"
))->df_3DT_Year1960to1980
df_3DT%>%filter(CMPLNT_FR_DT>=as.Date("1980-01-01") & CMPLNT_FR_DT<=as.Date("2000-01-01"
))->df_3DT_Year1980to2000
df_3DT%>%filter(CMPLNT_FR_DT>=as.Date("2000-01-01") & CMPLNT_FR_DT<=as.Date("2006-01-01"
))->df_3DT_Year2000to2006

#association between report date and complaint date indicating possible typo in recordin
g the data
splom(df_3DT_Year1900)
```



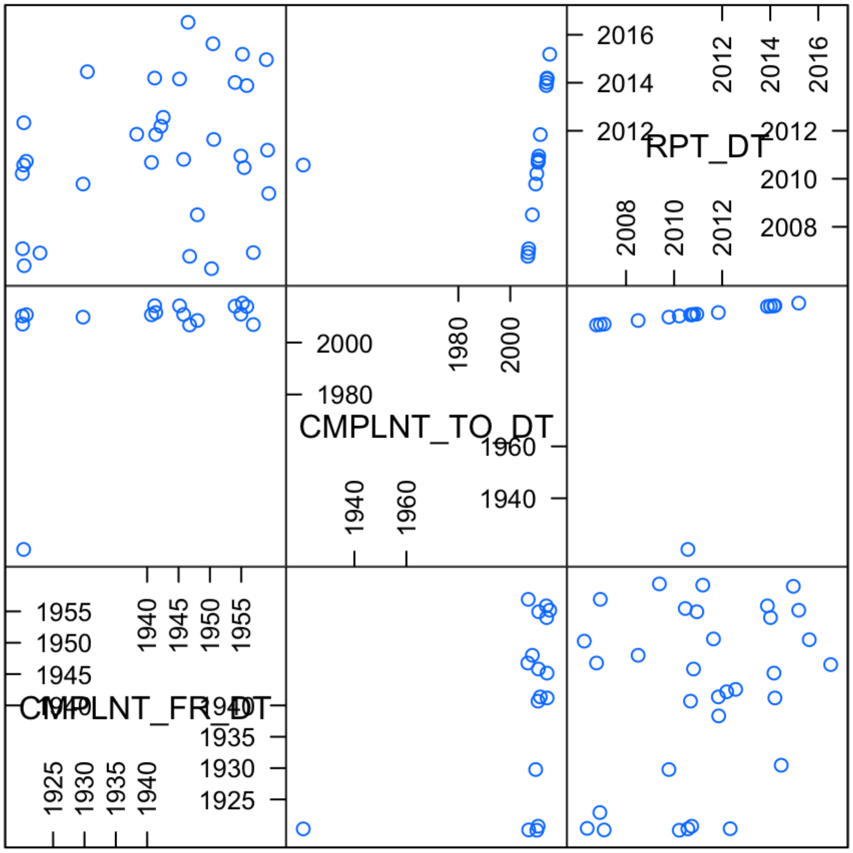
Scatter Plot Matrix

```
splom(df_3DT_Year1900to1920)
```



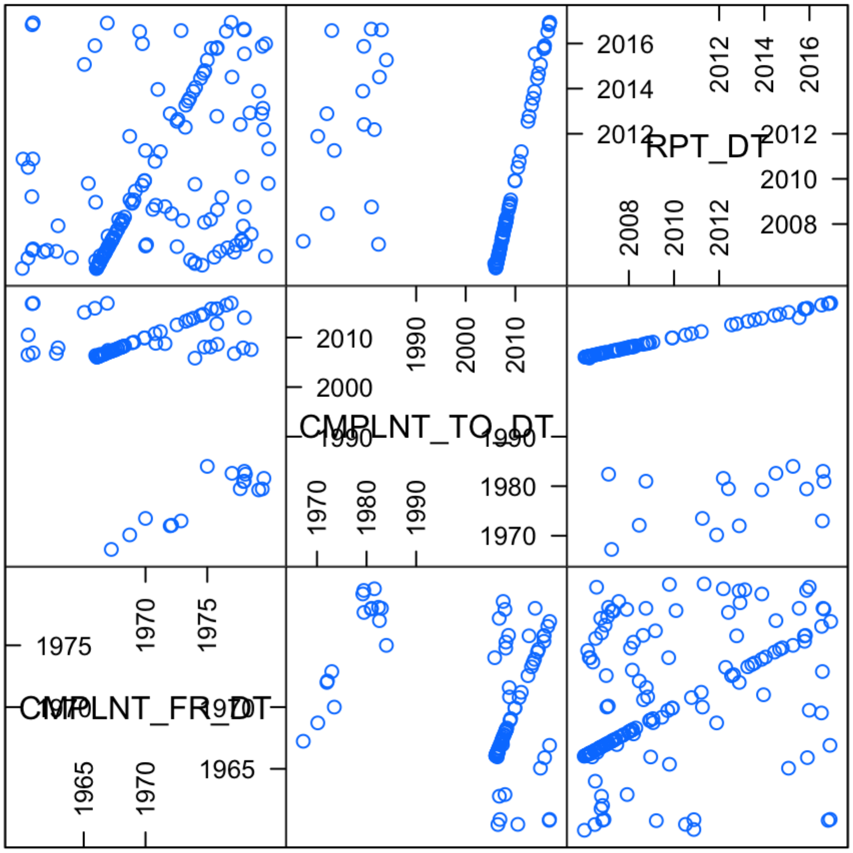
Scatter Plot Matrix

```
splom(df_3DT_Year1920to1960)
```



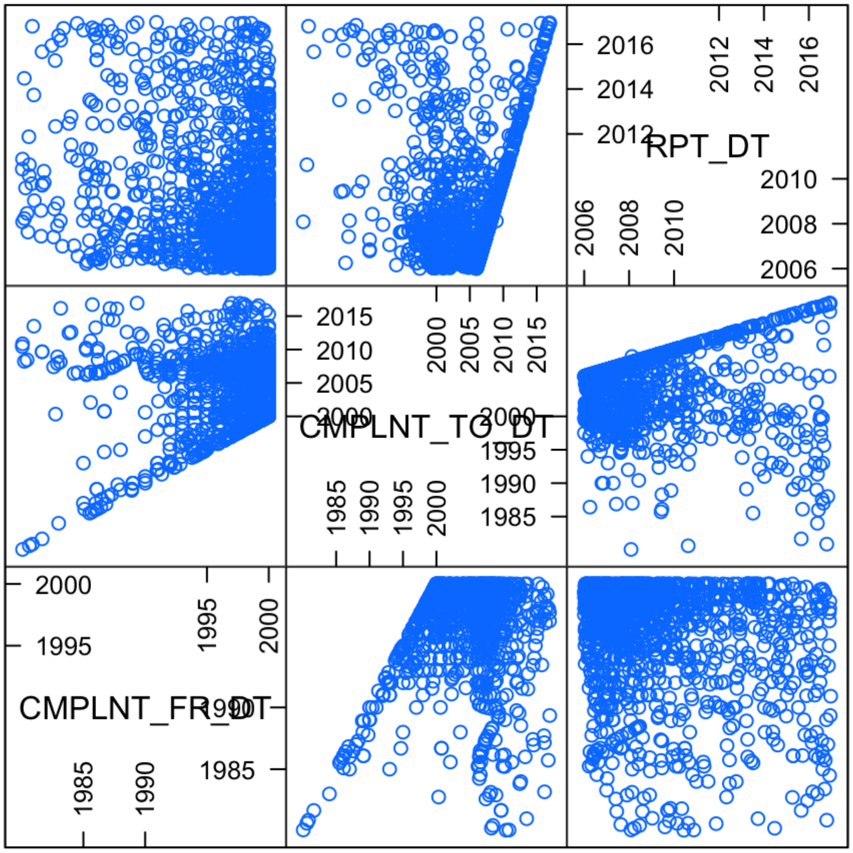
Scatter Plot Matrix

```
splom(df_3DT_Year1960to1980)
```



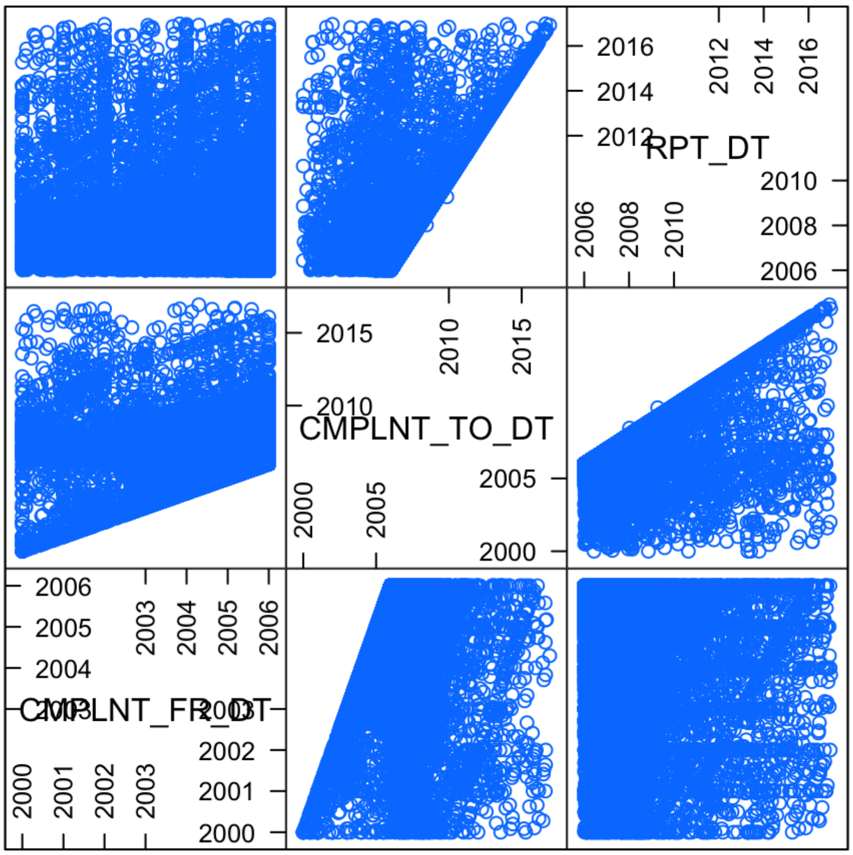
Scatter Plot Matrix

```
splom(df_3DT_Year1980to2000)
```



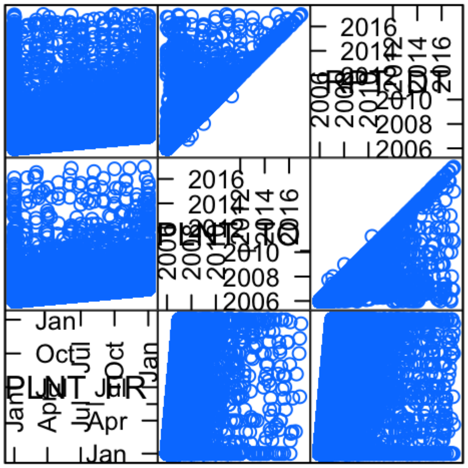
Scatter Plot Matrix

```
splom(df_3DT_Year2000to2006)
```



Scatter Plot Matrix

```
df_3DT%>%filter(CMPLNT_FR_DT>=as.Date("2006-01-01") & CMPLNT_FR_DT<=as.Date("2007-01-01"
))->df_3DT_Year2006to2007
splom(df_3DT_Year2006to2007)
```



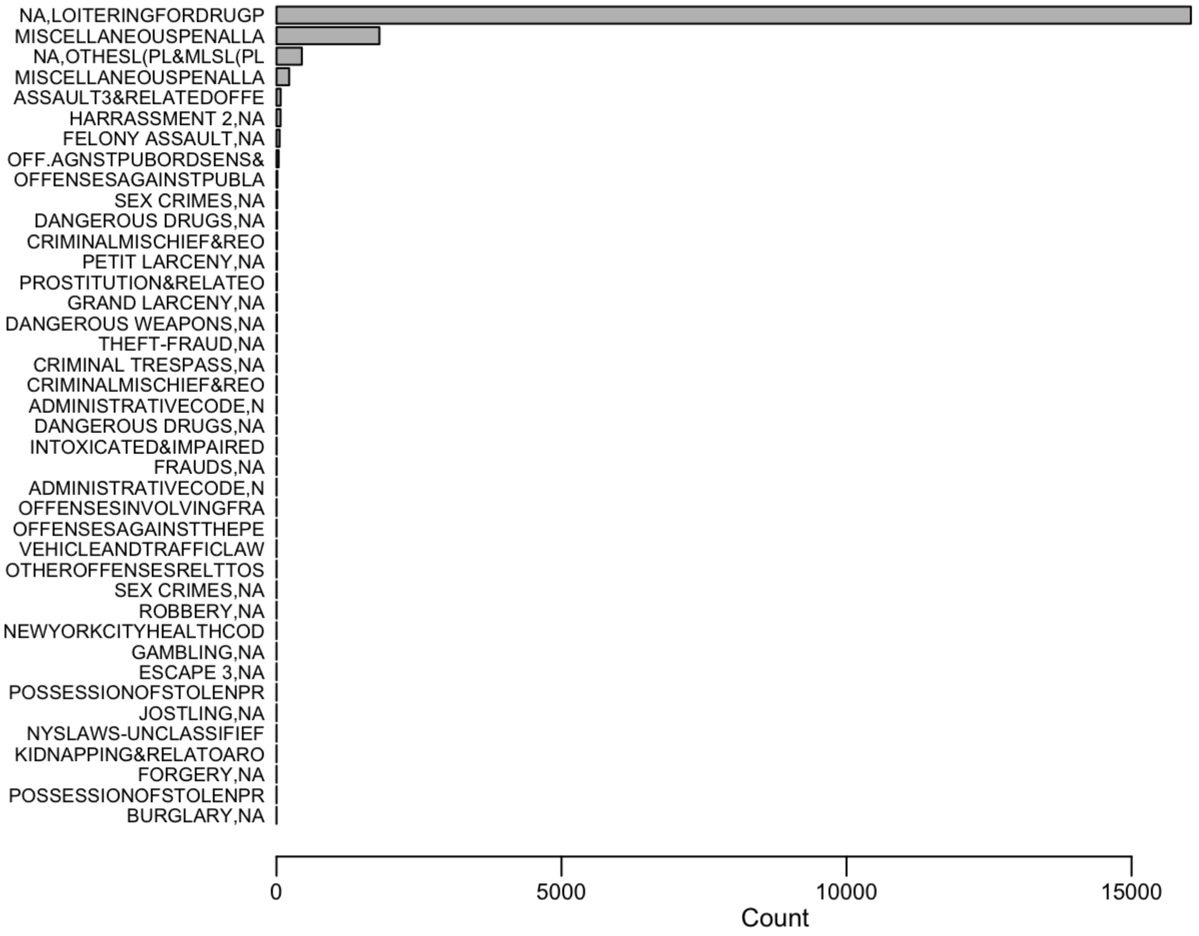
Scatter Plot Matrix

===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.

```
#For cases with missing OFNS_DESC, how they distribute over the code KY_CD
df%>%
  select(KY_CD,OFNS_DESC)%>%
  filter(is.na(OFNS_DESC))%>%
  mutate(KY_CD=as.factor(KY_CD))%>%
  group_by(KY_CD)%>%summarise(count=n())->tmp6

#Matching the code KY_CD with the OFNS_DESC
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

#showing the supposed OFNS_DESC that is missing with its KY_CD
merge(tmp6,match_code_desc,by.x="KY_CD",by.y="KY_CD")%>%arrange(desc(count))->match_byco
unt

par(mgp=c(1,0.3,0),mai=c(0.4,1.8,0.01,0.01))
data2<-match_bycount[order(match_bycount[,"count"]),]
barplot(data2[,"count"],names.arg=abbreviate(data2[,"desc"],minlength=20),cex.names = 0.
6,cex.axis=0.7,cex.lab=0.8,horiz=TRUE,xlim=c(0,17500),las=1,xlab="Count")
```

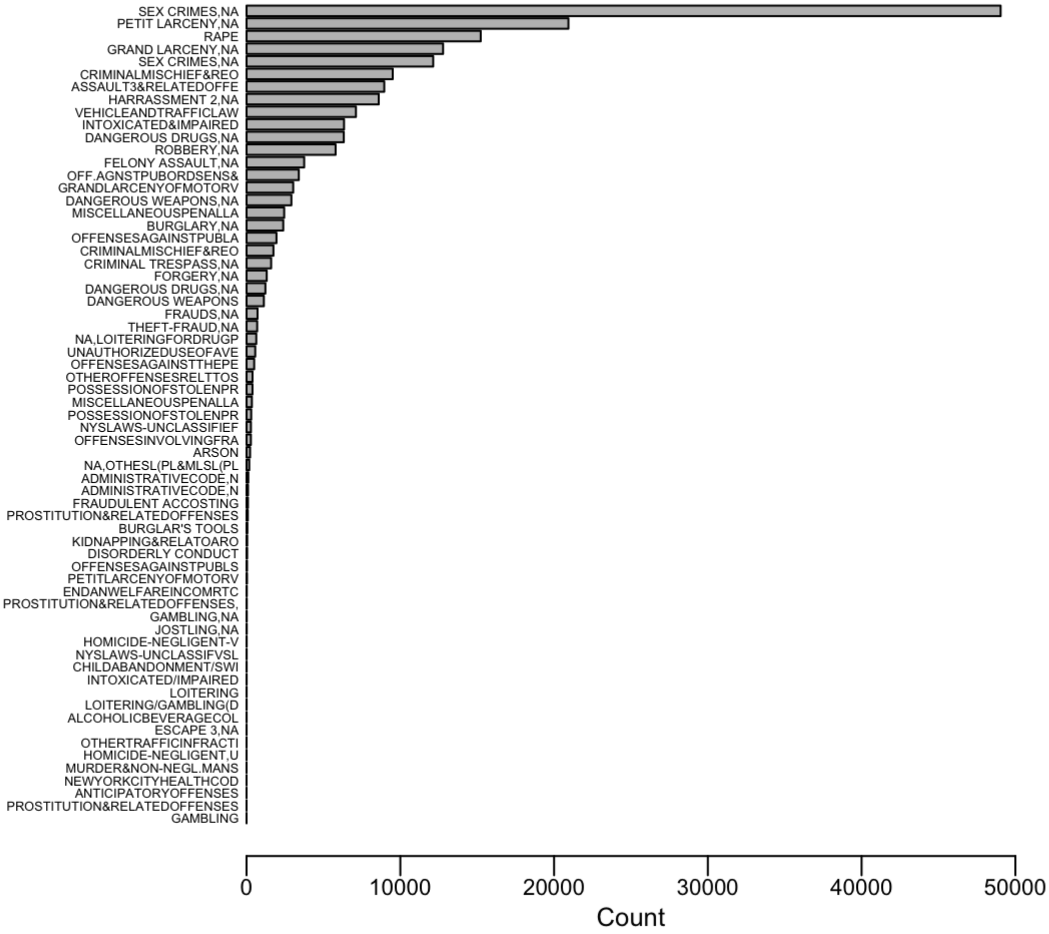


===Missing in geolocation===

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?

- 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.

```
#For cases with missing Latitude, how they distribute over the code KY_CD
df%>%
  select(KY_CD,Latitude)%>%
  filter(is.na(Latitude))%>%
  mutate(KY_CD=as.factor(KY_CD))%>%
  group_by(KY_CD)%>%summarise(count=n())->tmp7

merge(tmp7,match_code_desc,by.x="KY_CD",by.y="KY_CD")%>%arrange(desc(count))->match_byco
unt2

#par(mar=c(4.1,15.1,2.1,2.1))
par(mgp=c(1,0.2,0),mai=c(0.4,2.5,0.01,0.5))
data2<-match_bycount2[order(match_bycount2[,"count"]),]
barplot(data2[,"count"],names.arg=abbreviate(data2[,"desc"],minlength=20),horiz=TRUE,ce
x.names = 0.4,cex.axis=0.7,cex.lab=0.8,xlim=c(0,50000),las=1,xlab="Count")
```



===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.