San Francisco Crime in Summer 2014

I used the Summer 2014 San Francisco criminal incident data. I analyzed the patterns of different categories of crimes regarding to the hours of a day, days of a week, different months. I'd like to see if the frequencies of criminal incidents is correlated to specific time slots such as certain hours of a day, or particular days of a week or certain months. I also plotted the geographic map of incidents to how the incidents distributed in the area.

Step 0, data preparation. There are 34 criminal categories. I selected the top 7 most frequent categories to do the analysis.

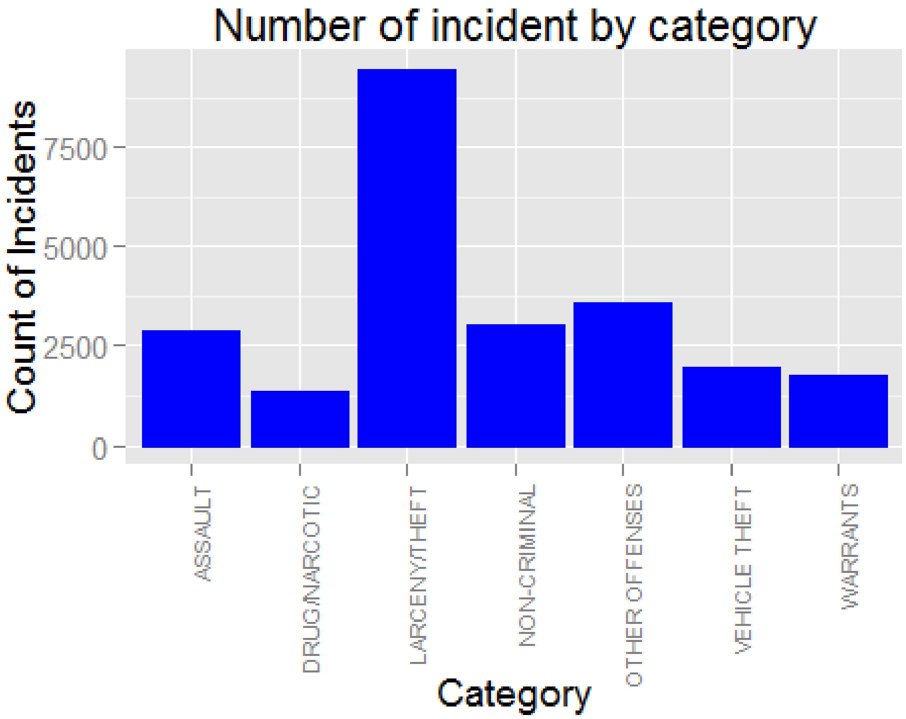
```r
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

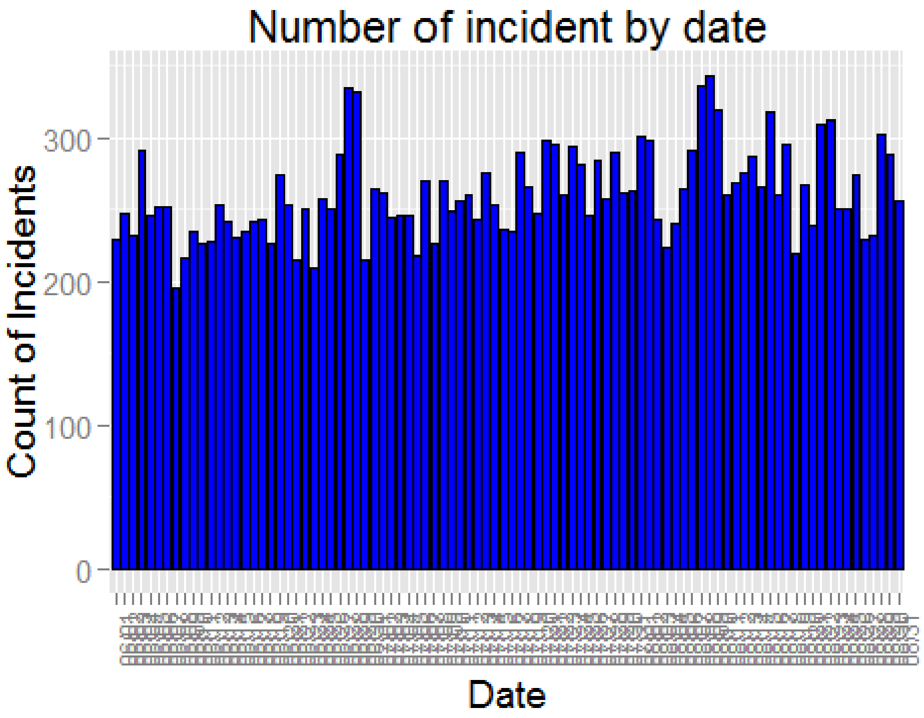
```r
setwd('C:\\Users\\usz003g\\Documents\\R training\\Crime')

d=read_csv("sanfrancisco_incidents_summer_2014.csv",col_names=T)
d0=d
d=d[order(d$Date),]

tb=table(d$Category)
topcat=names(tb[order(tb, decreasing = TRUE)][1:7])
d=d[d$Category %in% topcat,]
d$hour=as.numeric(substr(d$Time,1,2))
d$Month=as.numeric(substr(d$Date,1,2))

d$day[d$DayOfWeek=='Monday']='1.Monday'
d$day[d$DayOfWeek=='Tuesday']='2.Tuesday'
d$day[d$DayOfWeek=='Wednesday']='3.Wednesday'
d$day[d$DayOfWeek=='Thursday']='4.Thursday'
d$day[d$DayOfWeek=='Friday']='5.Friday'
d$day[d$DayOfWeek=='Saturday']='6.Saturday'
d$day[d$DayOfWeek=='Sunday']='7.Sunday'
```

Step 1, number of incidents by Category, Date.

```
#by category
mdf <- d%>%group_by(Category)%>%summarise(count=n())
ggplot(mdf, aes(x = Category, y = count)) +  geom_bar(stat = "identity",
color= "blue",fill='blue')+ labs(x = "Category", y = "Count of Incidents") +
  ggtitle("Number of incident by category") +theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=8))
```



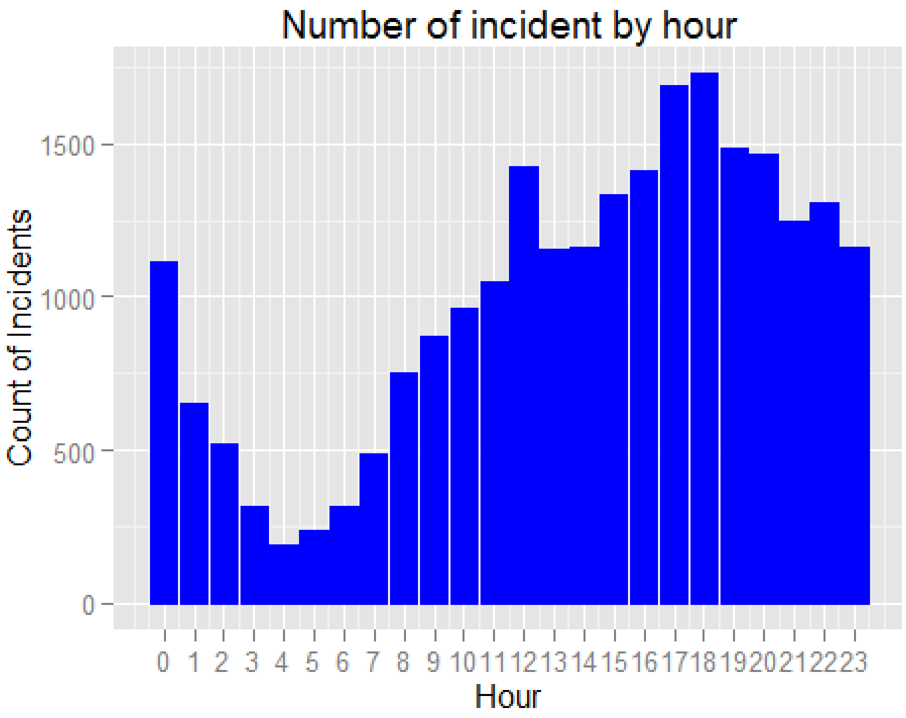
```
#by date
d$date=substr(d$Date,1,5)
mdf <- d%>%group_by(date)%>%summarise(count=n())
ggplot(mdf, aes(x = date, y = count)) +  geom_bar(stat = "identity", color=
"black",fill='blue')+ labs(x = "Date", y = "Count of Incidents") +
  ggtitle("Number of incident by date") +theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=8))
```



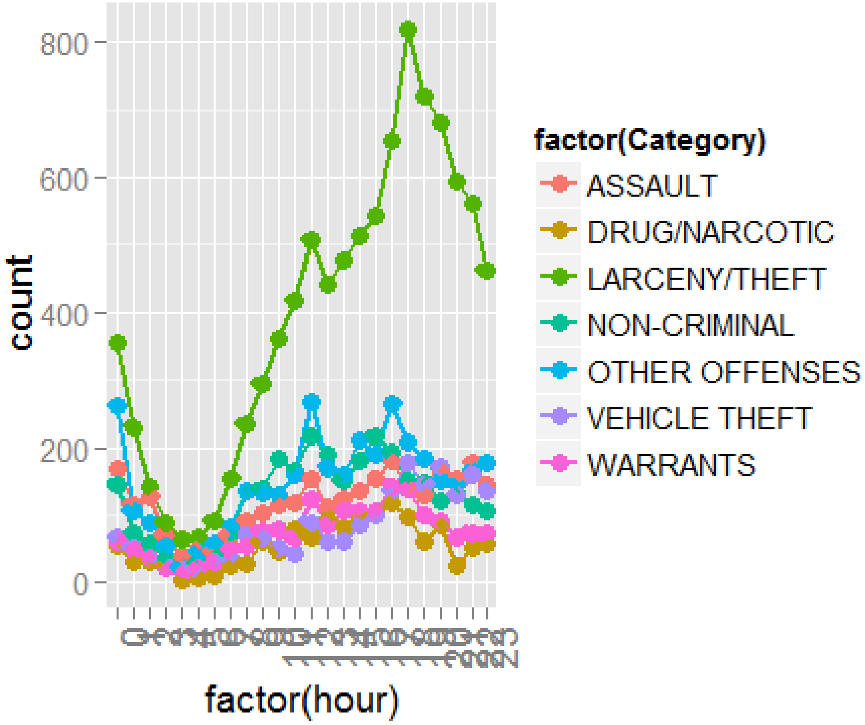
The most frequent type of criminal incident is "LARCENY/THEFT". The are several spikes of incidenst according to the bar char, for example June 26-June 28, August 07-August 09 are two spikes with higher number of incidents.

Step 2, check the patterns of criminial incidents regarding to the hours of the day.

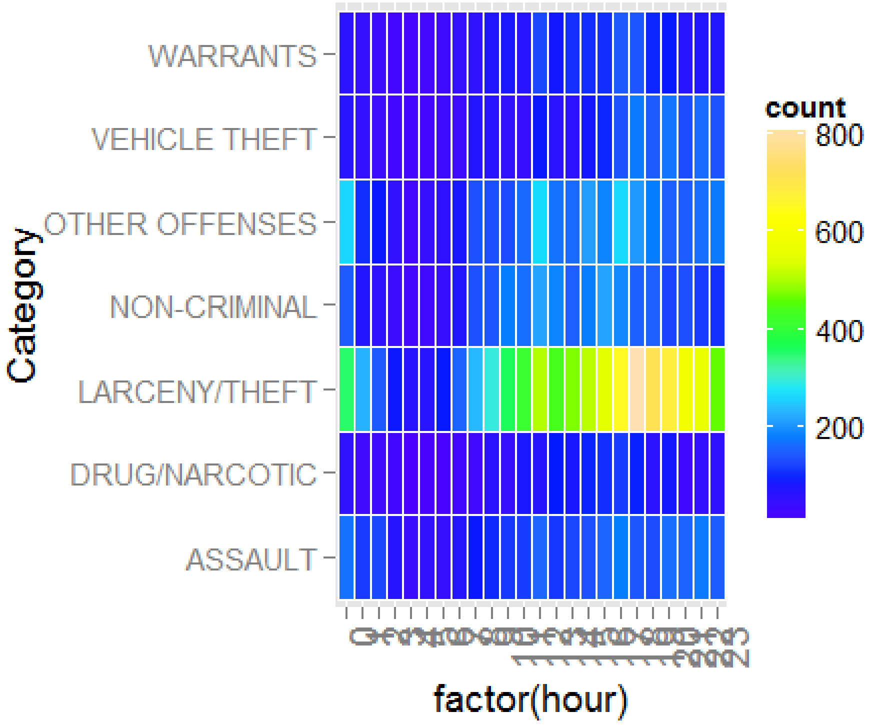
```
#by hour
mdf <- d%>%group_by(hour)%>%summarise(count=n())
ggplot(mdf, aes(x = hour, y = count)) +  geom_bar(stat = "identity", color=
"blue",fill='blue')+ labs(x = "Hour", y = "Count of Incidents") +
  scale_x_continuous(breaks=c(0:23)) +
  ggtitle("Number of incident by hour") +
  theme(plot.title = element_text(size = 14))
```



```
#crime by hour

mdf <- d%>%group_by(hour,Category)%>%summarise(count=n())
#mdf=mdf[mdf$count>1]
#ggplot(mdf,aes(x=factor(income),y=case))+
stat_summary(fun.y=mean,geom="bar",fill='blue')
ggplot(mdf,aes(x=factor(hour),y=count,group=factor(Category),
colour=factor(Category)))+geom_line(size=1) +geom_point(size=3, fill="white")
+geom_jitter(aes(color=factor(Category)))+theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



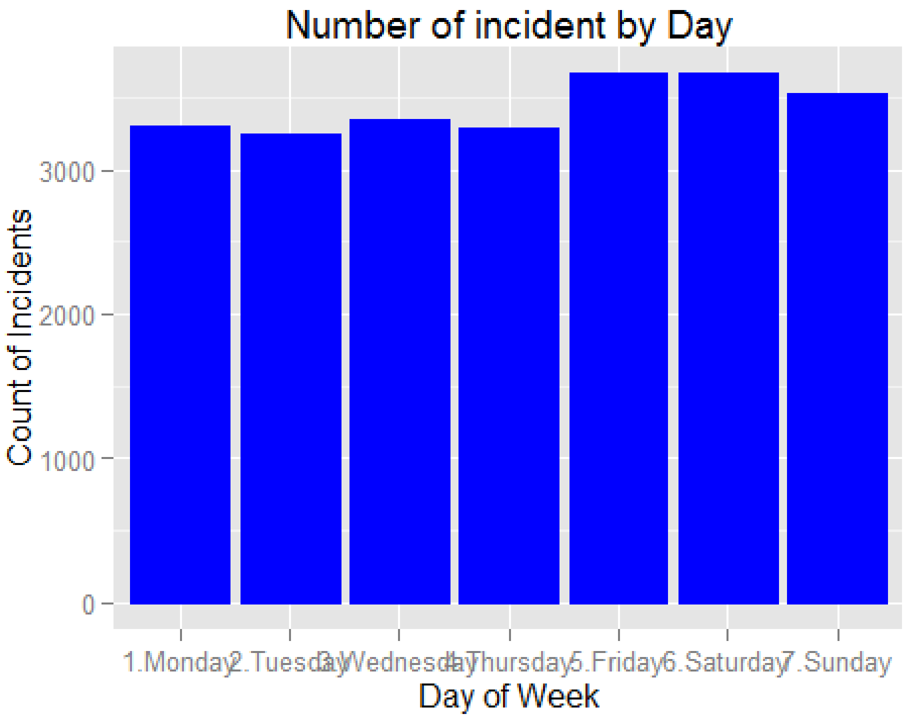
```
ggplot(mdf, aes(y=Category, x=factor(hour), fill=count)) +
geom_tile(colour="white") +
  scale_fill_gradientn(colours=topo.colors(10),
                       guide=guide_colourbar(ticks=T, nbin=50,
                                             barheight=10, label=T,
                                             barwidth=1)) +theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



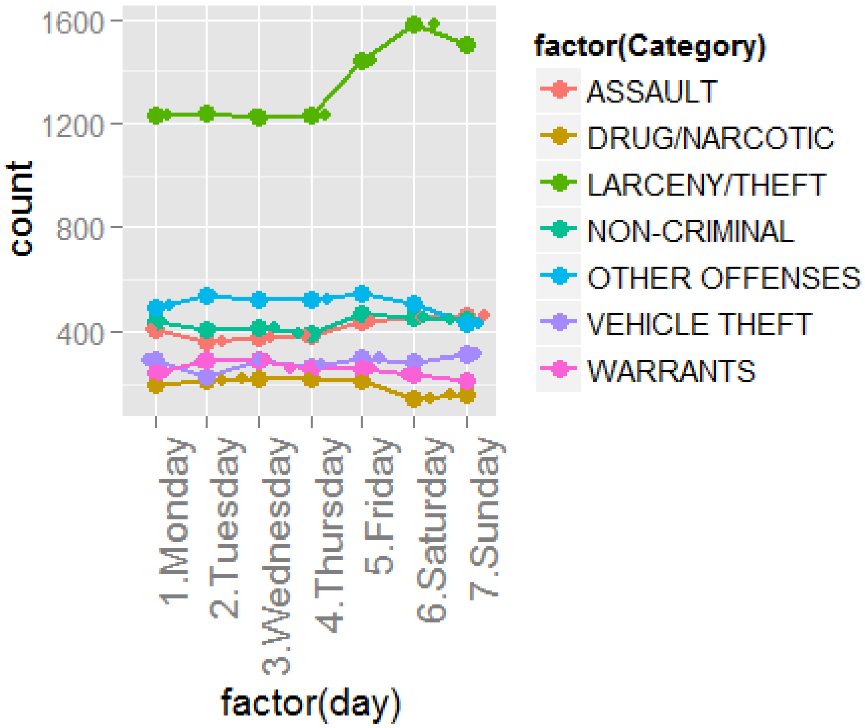
More criminal incidents happened in the afternoons/evenings than mornings.The number of incidents started to increas sharply from 5am in the morning, peaked at the evening around 5pm or 6pm, and then began to drop, least incidents happened in the nights from 1am to 7am,

Step 3, check the patterns regarding to the days of the week.

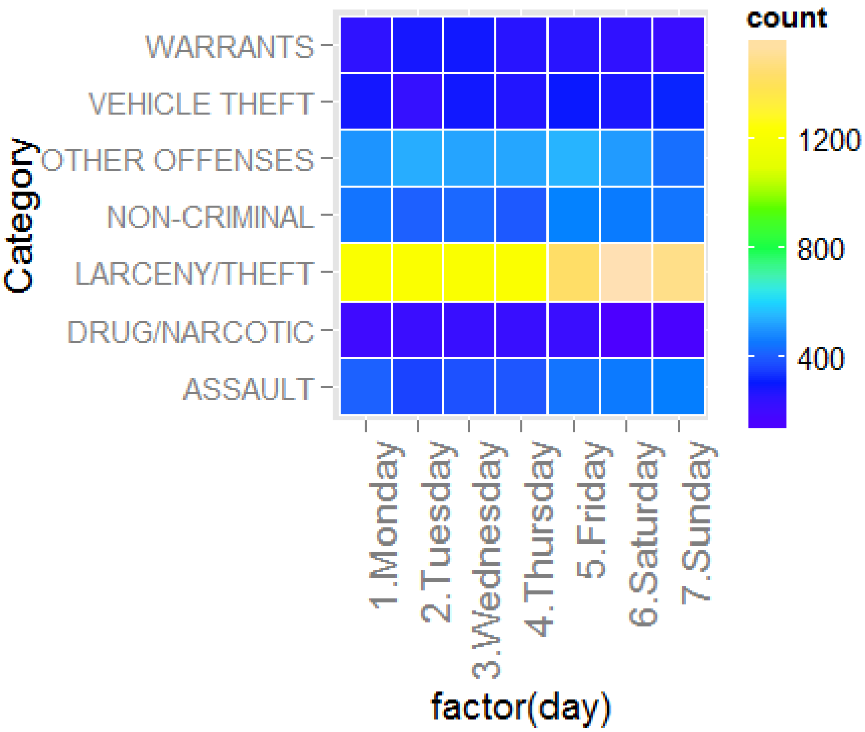
```
#by day
mdf <- d%>%group_by(day)%>%summarise(count=n())
ggplot(mdf, aes(x = day, y = count)) +  geom_bar(stat = "identity", color=
"blue",fill='blue')+ labs(x = "Day of Week", y = "Count of Incidents") +
  ggtitle("Number of incident by Day") +
  theme(plot.title = element_text(size = 14))
```



```
#crime by dayofweek
mdf <- d%>%group_by(DayOfWeek,Category)%>%summarise(count=n())
mdf$day[mdf$DayOfWeek=='Monday']='1.Monday'
mdf$day[mdf$DayOfWeek=='Tuesday']='2.Tuesday'
mdf$day[mdf$DayOfWeek=='Wednesday']='3.Wednesday'
mdf$day[mdf$DayOfWeek=='Thursday']='4.Thursday'
mdf$day[mdf$DayOfWeek=='Friday']='5.Friday'
mdf$day[mdf$DayOfWeek=='Saturday']='6.Saturday'
mdf$day[mdf$DayOfWeek=='Sunday']='7.Sunday'

ggplot(mdf,aes(x=factor(day),y=count,group=factor(Category),
colour=factor(Category)))+geom_line(size=1) +geom_point(size=3, fill="white")
+geom_jitter(aes(color=factor(Category)))+theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



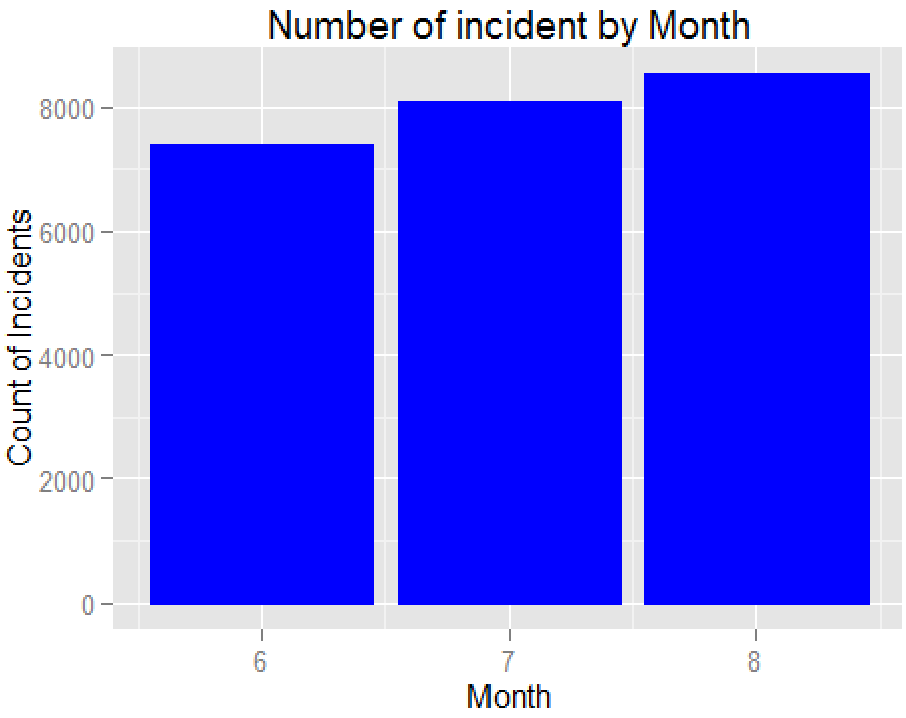
```
ggplot(mdf, aes(y=Category, x=factor(day), fill=count)) +
geom_tile(colour="white") +
  scale_fill_gradientn(colours=topo.colors(10),
                       guide=guide_colourbar(ticks=T, nbin=50,
                                             barheight=10, label=T,
                                             barwidth=1)) +theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



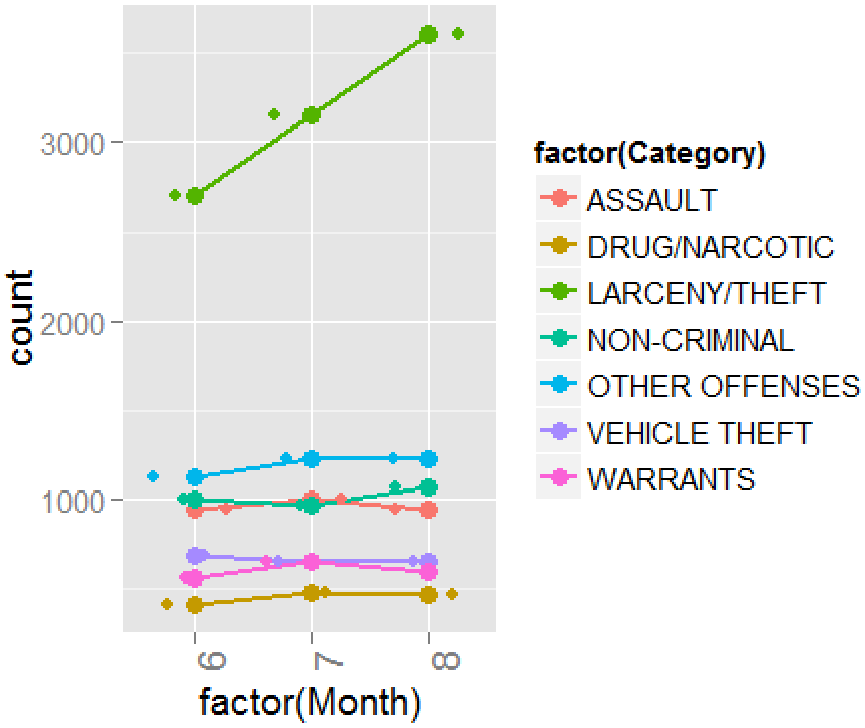
More theft incidents happened on Friday, Saturday and Sunday; but less drug incidents happend on Saturday and Sunday, the left categories are relatively stable across different days.

Step 4, check the incident patterns regarding to months.

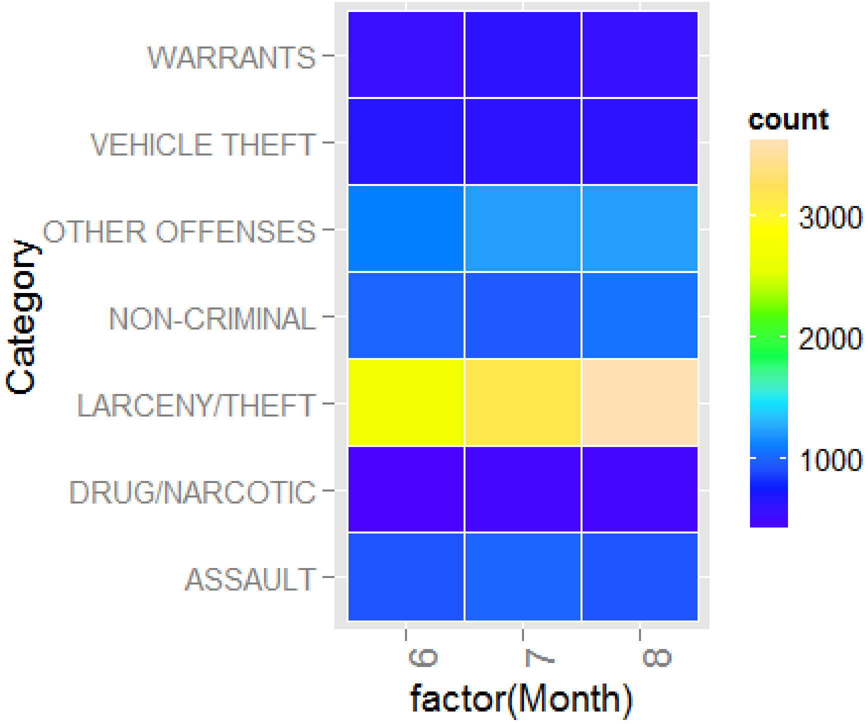
```
#by month
mdf <- d%>%group_by(Month)%>%summarise(count=n())
ggplot(mdf, aes(x = Month, y = count)) +  geom_bar(stat = "identity", color=
"blue",fill='blue')+ labs(x = "Month", y = "Count of Incidents") +
  ggtitle("Number of incident by Month") +
  theme(plot.title = element_text(size = 14))
```



```
mdf <- d%>%group_by(Month,Category)%>%summarise(count=n())
ggplot(mdf,aes(x=factor(Month),y=count,group=factor(Category),
colour=factor(Category)))+geom_line(size=1) +geom_point(size=3, fill="white")
+geom_jitter(aes(color=factor(Category)))+theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



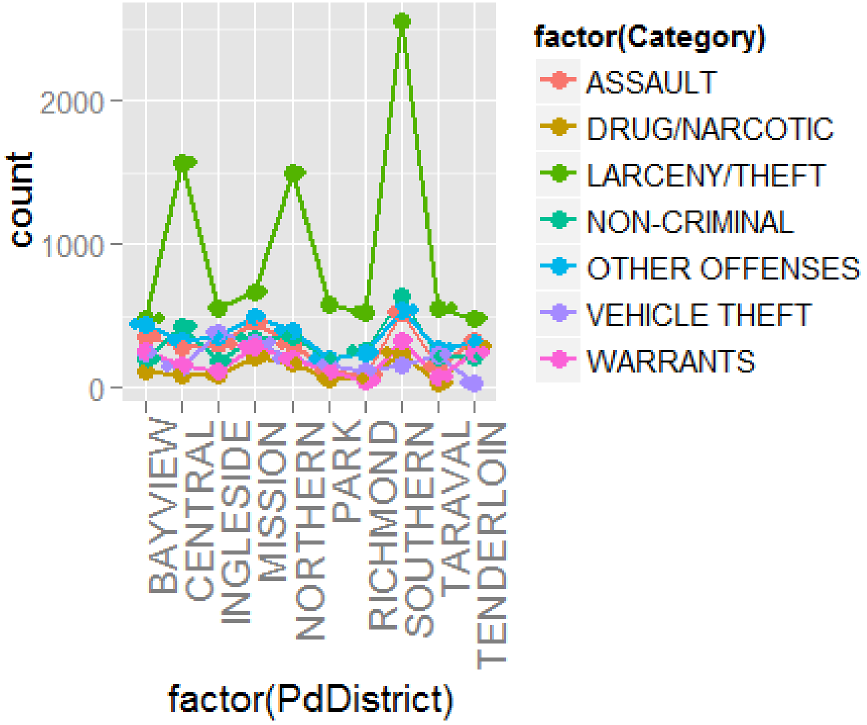
```
ggplot(mdf, aes(y=Category, x=factor(Month), fill=count)) +
geom_tile(colour="white") +
  scale_fill_gradientn(colours=topo.colors(10),
                       guide=guide_colourbar(ticks=T, nbin=50,
                                             barheight=10, label=T,
                                             barwidth=1)) +theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



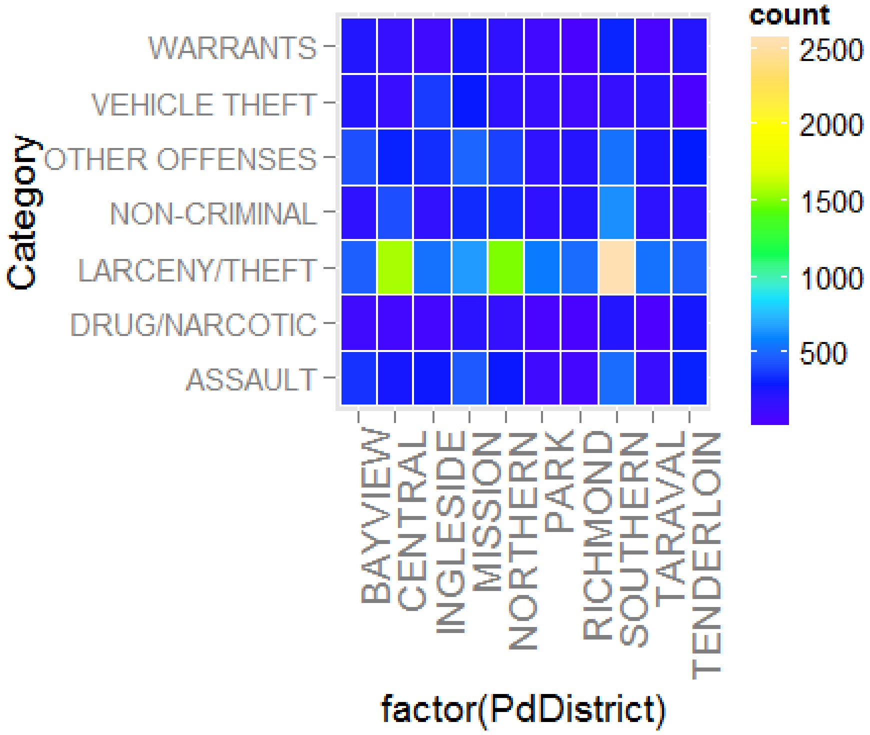
The number of theft incidents increased strictly from June to August. The temperature increased from June to August, and this could indicate that more theft happened as temperature went up. The other types of incidents had no clear pattern.

Step 5, check the incident patterns by Police Districts

```
#PD district

mdf <- d%>%group_by(PdDistrict,Category)%>%summarise(count=n())
ggplot(mdf,aes(x=factor(PdDistrict),y=count,group=factor(Category),
colour=factor(Category)))+geom_line(size=1) +geom_point(size=3, fill="white")
+geom_jitter(aes(color=factor(Category)))+theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```



```
ggplot(mdf, aes(y=Category, x=factor(PdDistrict), fill=count)) +
geom_tile(colour="white") +
  scale_fill_gradientn(colours=topo.colors(10),
                       guide=guide_colourbar(ticks=T, nbin=50,
                                             barheight=10, label=T,
                                             barwidth=1)) +theme(text =
element_text(size=14),axis.text.x = element_text(angle = 90, hjust = 1,
size=14))
```

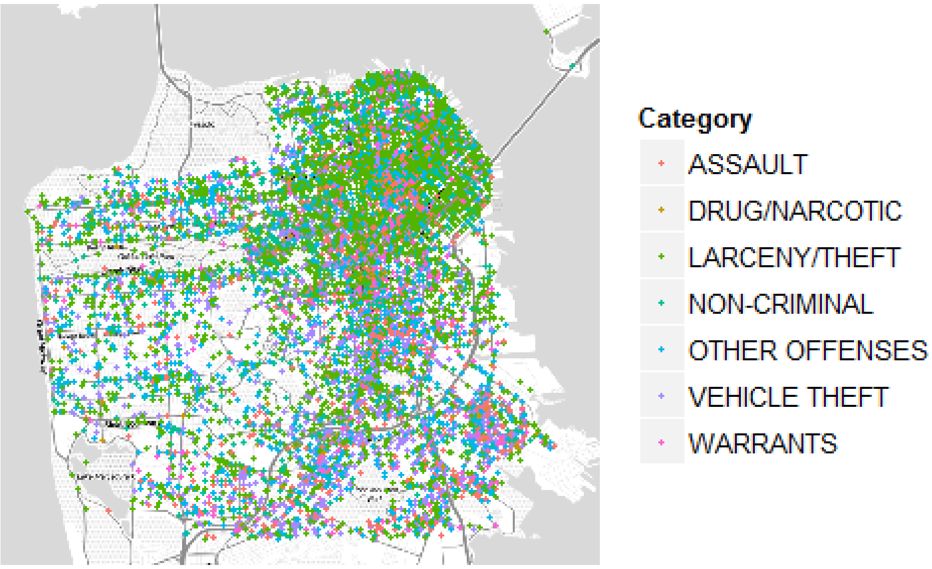


The Southern, Northern and Central police districts had much more THEFT incidents. Southern district had most incidents for almost all the categories.

Step 6, map the geographic locations of the incidents.

```
#map

library(ggmap)
d$Latitude   <- d$Y
d$Longitude  <- d$X

g <- qmplot(Longitude, Latitude, data = d, color = Category, size = I(1))

## Using zoom = 13...
## Map from URL : http://tile.stamen.com/toner-lite/13/1307/3165.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1308/3165.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1309/3165.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1310/3165.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1311/3165.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1307/3166.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1308/3166.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1309/3166.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1310/3166.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1311/3166.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1307/3167.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1308/3167.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1309/3167.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1310/3167.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1311/3167.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1307/3168.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1308/3168.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1309/3168.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1310/3168.png
## Map from URL : http://tile.stamen.com/toner-lite/13/1311/3168.png
```

```
g
```



The north-east conor of San Francisco area had most incidents for that area has the highest population density.

To conclude, more incidents happened during afternoons and evenings, the most frequent type of incidents was THEFT, number of incidents increased strictly from June to August as the temperature went up, most incidents happened in the northeast are of SF due to the high population density.