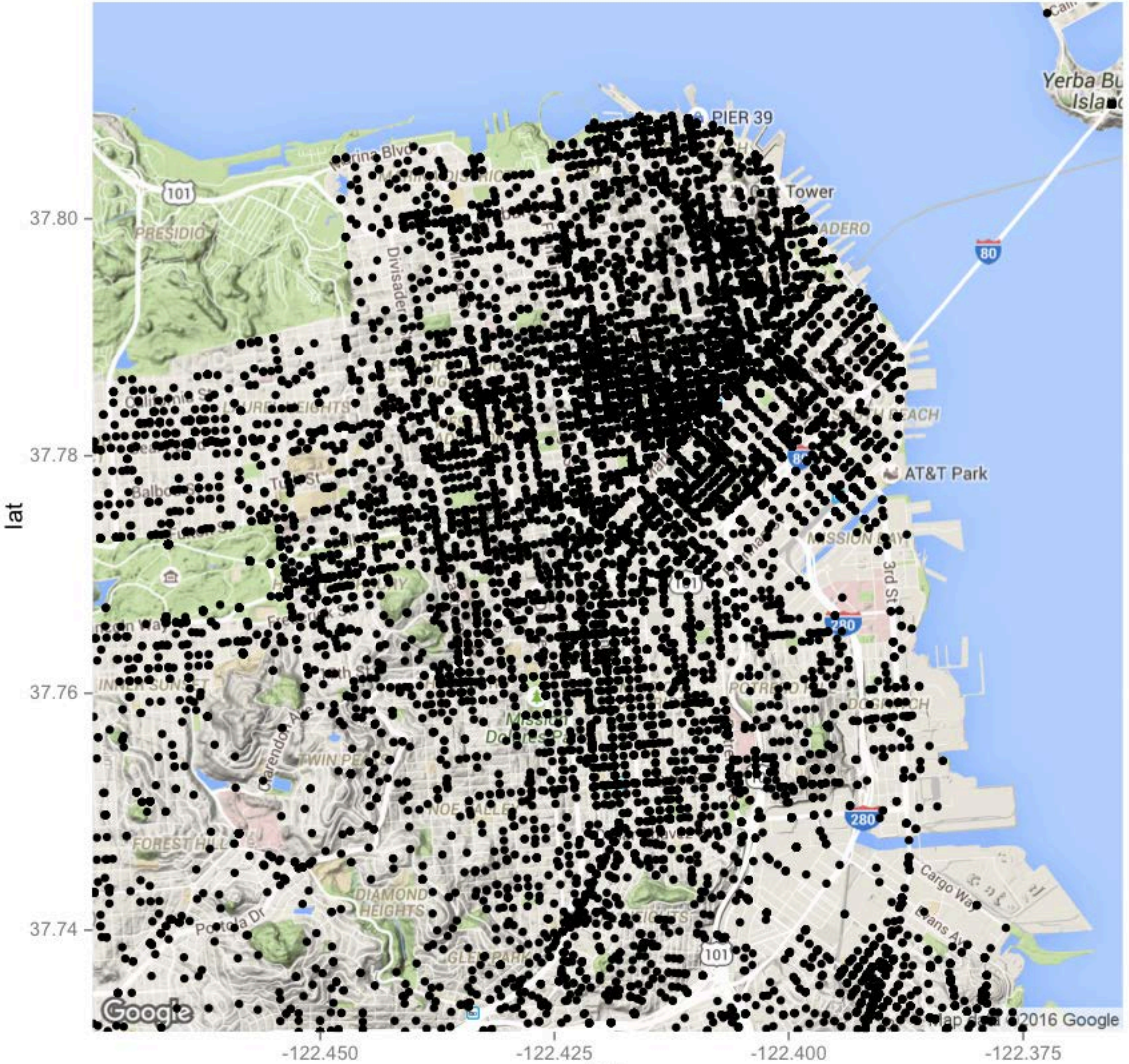
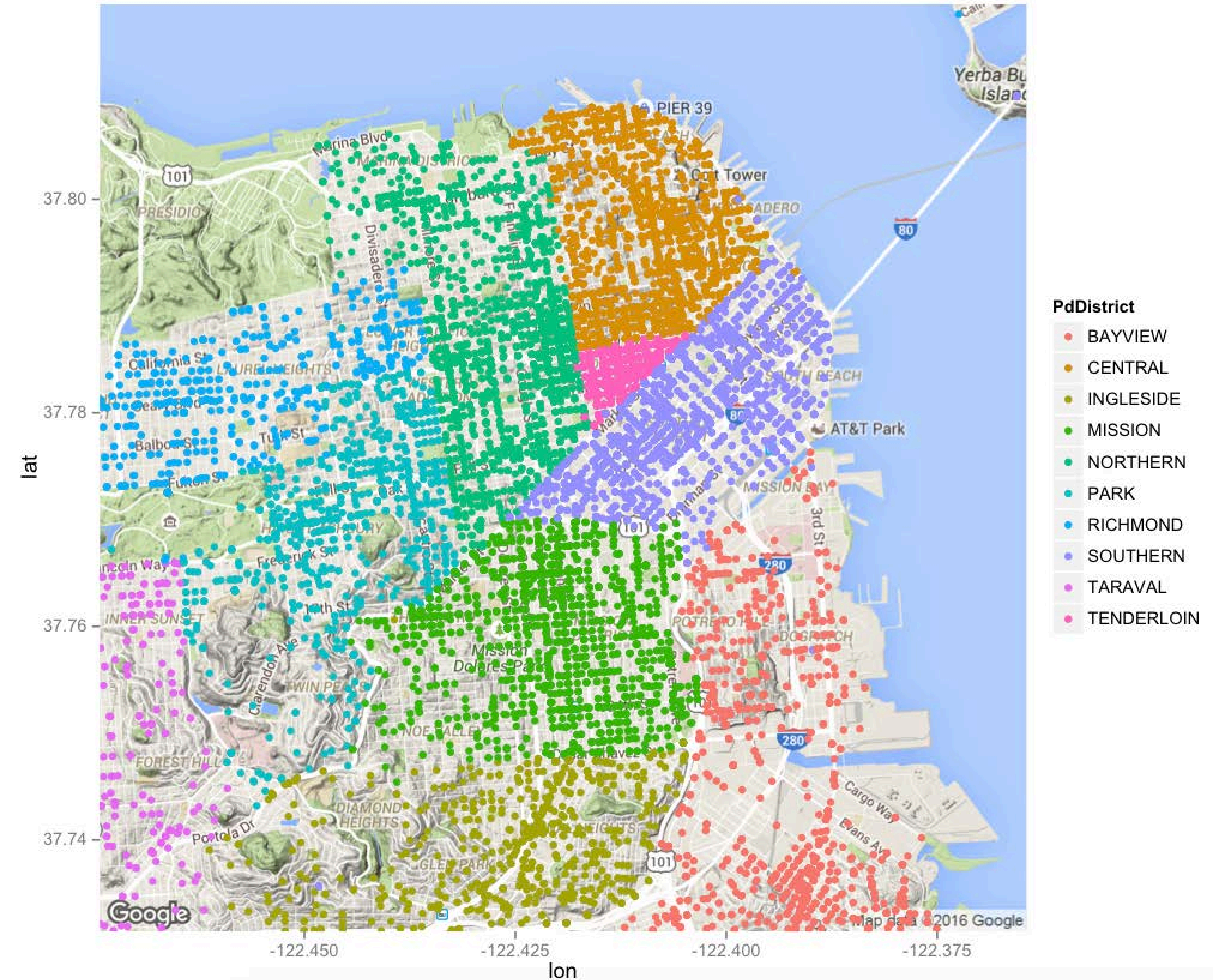
```r
#first, I open the file and make a simple plot of all crimes by district
setwd("/Users/igor/MOOCs/data_manipulation/datasci_course_materials/assignment6")

sanfr = read.csv('sanfrancisco_incidents_summer_2014.csv')

summary(sanfr)

library(ggplot2)

#for this I will need latitude/longitude data, luckily it can be easily transformed
#from $Location column
sanfr$lat = as.double(substr(sanfr$Location, 2, 17))
sanfr$lon = as.double(substr(sanfr$Location, 20, 36))

#ok, this works, but maybe seeing it on an actual map would be more intuitive?
p <- ggplot(sanfr, aes(lat, lon, color = PdDistrict))
p + geom_point()
```


-122.36
-122.40
-122.44
-122.48
-122.52
lon
37.725
37.750
37.775
37.800
lat
PdDistrict
BAYVIEW
CENTRAL
INGLESIDE
MISSION
NORTHERN
PARK
RICHMOND
SOUTHERN
TARAVAL
TENDERLOIN

```
library(maps)
library(ggmap)

sf = get_map(location = "san francisco", zoom = 13)

#now it is much more clear where crime happens most
ggmap(sf) + geom_point(data = sanfr, aes(x = lon, y = lat))
```

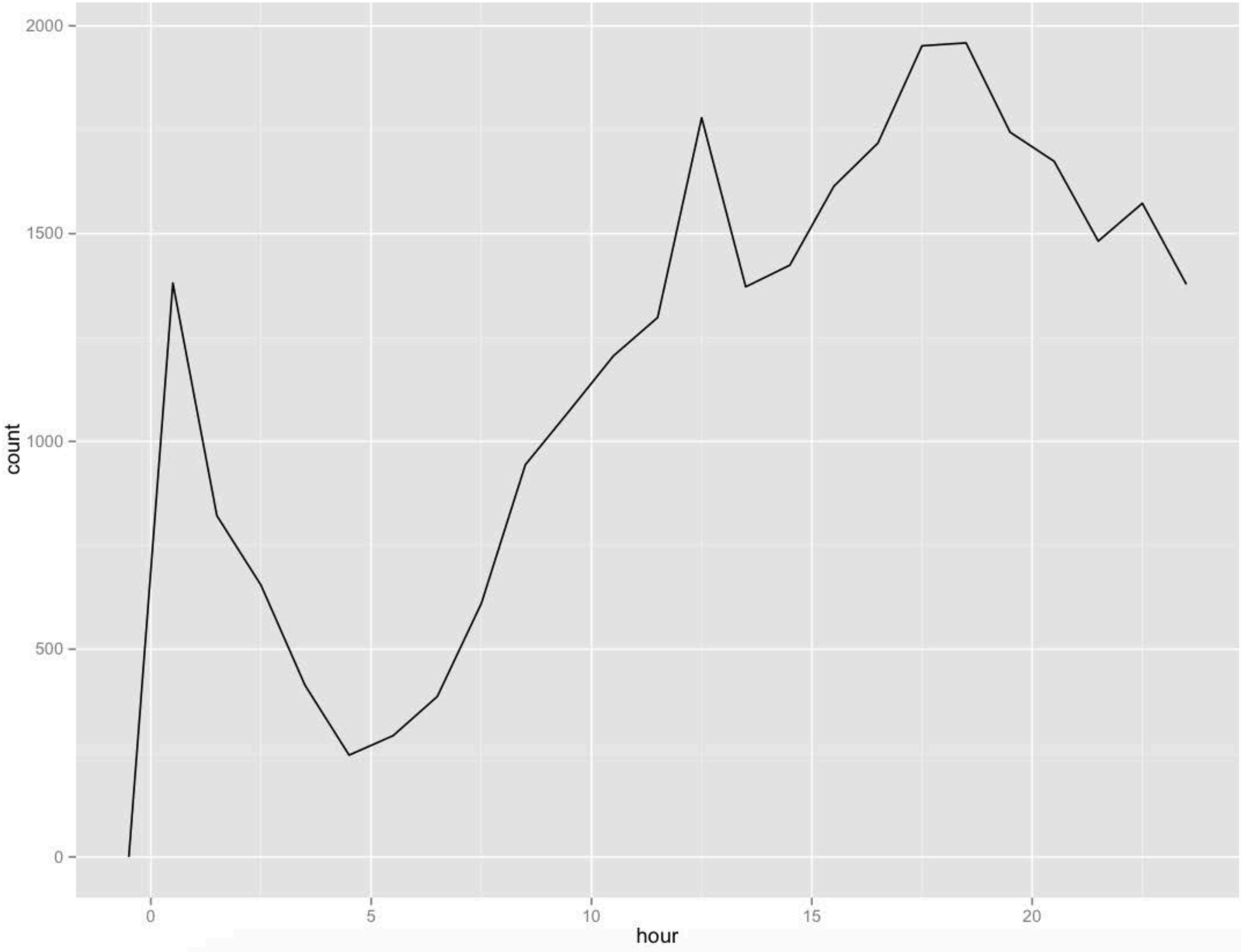

lat
37.80
37.78
37.76
37.74
PIER 39
Tower
PRESIDIO
Marina Blvd
Divisadero
101
80
280
AT&T Park
MISSION BAY
3rd St
Cargo Way
Evans Ave
Portola Dr
Carendon Ave
FOREST HILL
DIAMOND HEIGHTS
Google
Map data ©2016 Google
-122.450
-122.425
-122.400
-122.375
lon

```
#and with districts colored it's easier to send police where it's needed
ggmap(sf) + geom_point(data = sanfr, aes(x = lon, y = lat, color = PdDistrict))
```

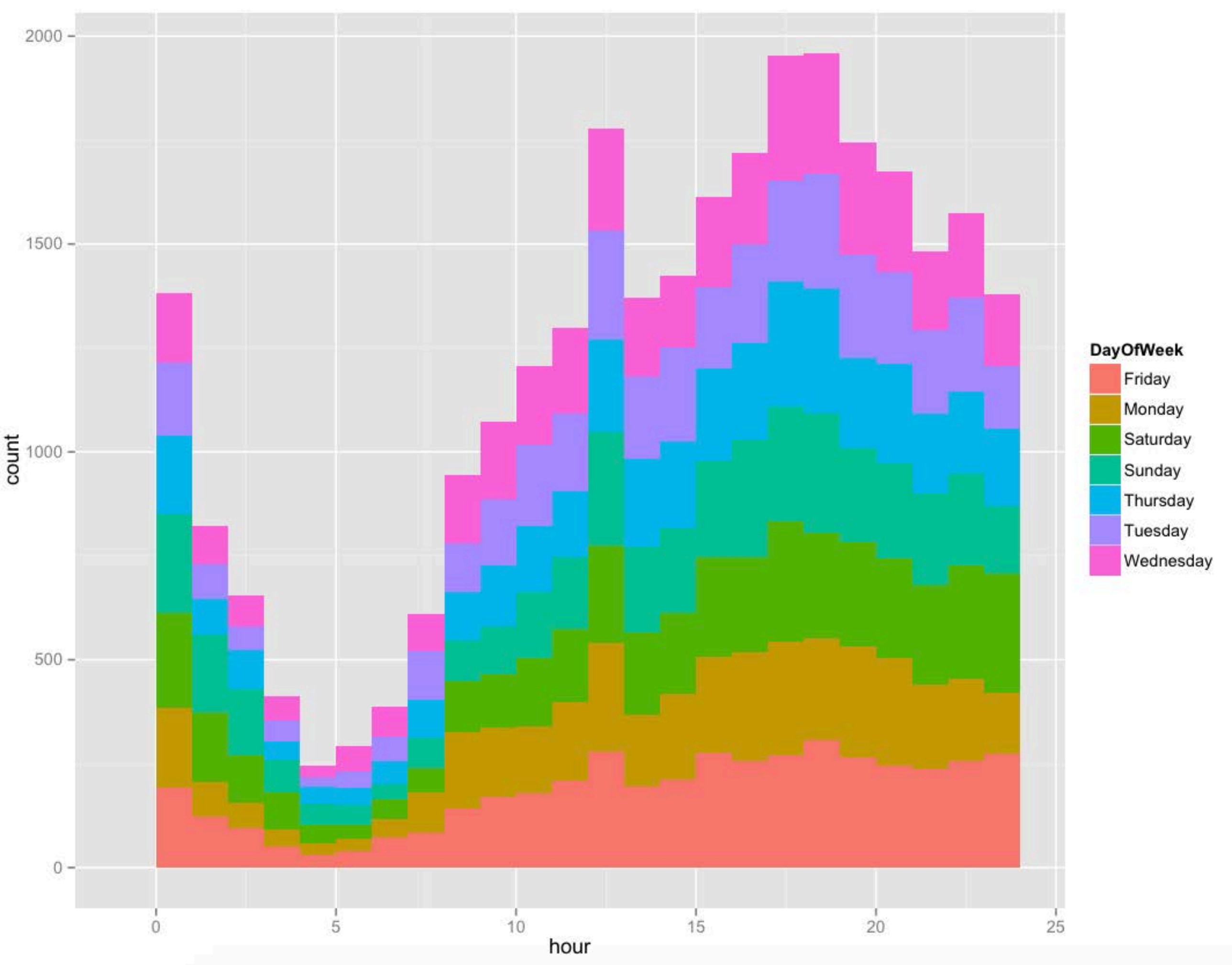


PdDistrict
BAYVIEW
CENTRAL
INGLESIDE
MISSION
NORTHERN
PARK
RICHMOND
SOUTHERN
TARAVAL
TENDERLOIN
lat
lon
37.80
37.78
37.76
37.74
-122.450
-122.425
-122.400
-122.375
PIER 39
Coit Tower
Yerba Bu
Islan
Marina Blvd
PRESIDIO
Divisadero St
California St
LAUREL HEIGHTS
Balboa St
Fulton St
Turk St
Frederick St
INNER SUNSET
Clarendon Ave
FOREST HILL
Portola Dr
DIAMOND HEIGHTS
GLEN PARK
Mission Dolores Park
NOE VALLEY
SOUTH BEACH
AT&T Park
MISSION BAY
3rd St
POTRERO HILL
DOGPATCH
Cargo Way
Evans Ave
80
101
280
Google
Map data ©2016 Google

```r
#extract hours from $Time column
sanfr$hour = as.numeric(substr(sanfr$Time, 1, 2))

#plot it as a line/hist/not sure what better term to use
#apparently 4AM is the safest hour, did not expect that
ggplot(sanfr, aes(hour)) +   geom_freqpoly(binwidth = 1)
```


2000
1500
1000
500
0
count
0
5
10
15
20
hour

```
#histogram of time vs. $DayOfWeek
#the objective of this one is to identify any outliers
#Saturday and Sunday at 1AM show about twice as many crimes as other days
#same thing for midnight on Fraidays and Saturdays
ggplot(sanfr, aes(hour, fill = DayOfWeek)) +   geom_histogram(binwidth = 1)
```


2000
1500
1000
500
0
count
0
5
10
15
20
25
hour
DayOfWeek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday