Visualizinig Crime Distribution

Kyu Cho

February 6, 2016

Overview

In this assignment, you will analyze criminal incident data from Seattle or San Francisco to visualize patterns and, if desired, contrast and compare patterns across the two cities.

You may want to consider one or more of the following types of questions when developing your submission.

For either city, how do incidents vary by time of day? Which incidents are most common in the evening? During what periods of the day are robberies most common? For either city, how do incidents vary by neighborhood? Which incidents are most common in the city center? In what areas or neighborhoods are robberies or thefts most common? For either city, how do incidents vary month to month in the Summer 2014 dataset? For either city, which incident types tend to correlate with each other on a day-by-day basis? Advanced What can we infer broadly about the differences in crime patterns between Seattle and San Francisco? Does one city tend to have more crime than the other, per capita? Do the relative frequencies of types of incidents change materially between the two cities? (NOTE: The two datasets do not have the same schema, so comparisons will require some work and some assumptions. This will require extra work, but you will be working at the forefront of what is known!) Advanced For either city, do certain crimes correlate with environmental factors such as temperature? (To answer this kind of question, you will need to identify and use external data sources!)

Introduction

Both San Francisco and Seattle are cities which are rapidly being gentrified causing a change in the social and economic structures of the cities. Therefore, it is likely that crime is also driven by these mechanisms with a clear distinction of which crimes occur in which parts of these cities and during what times of the day. Areas already gentrified are more likely to be affected by property crimes than violent ones.

Preprocessing the Data

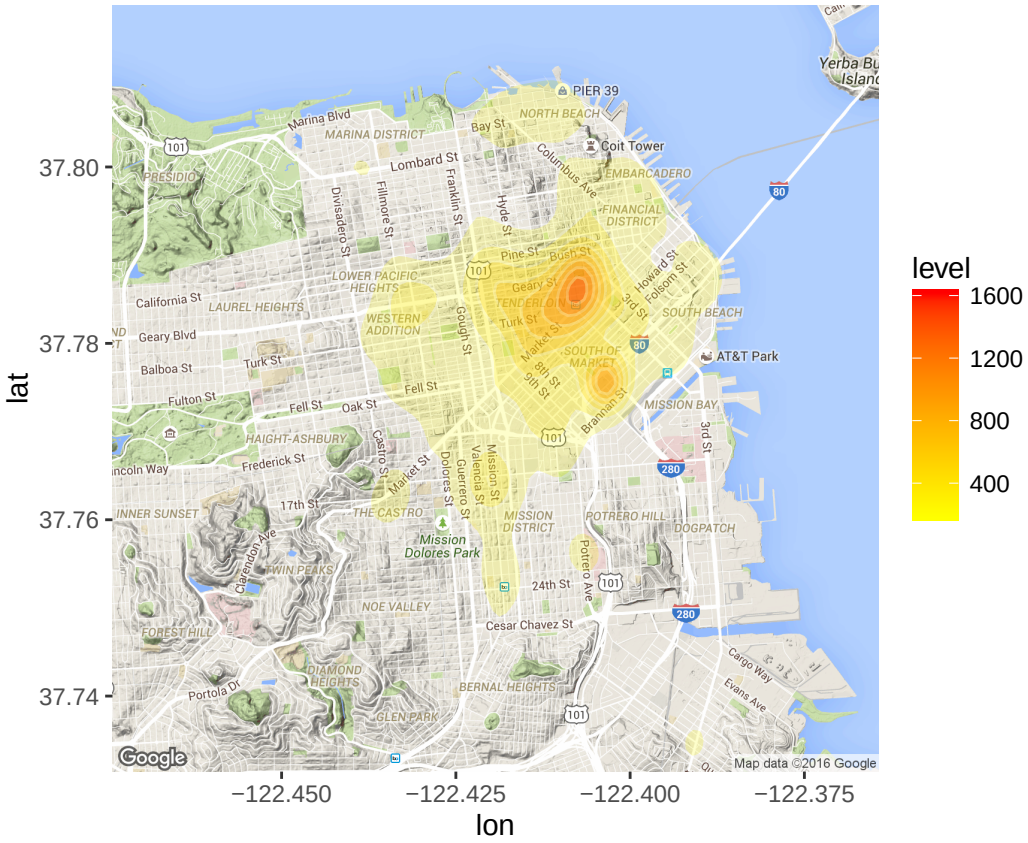
Since both dataset do not aggreed on the catatories, time format etc, I need to preprocess those set to make it easier to explor for the next step. Also, I simplefied into fewer catagories on the types of the crimes.

```
library(knitr)
library(dplyr)
library(ggplot2)
library(lubridate) # Time stamp

setwd("E:/Google Drive/College/20-Data Science at Scale (University of Washington)/3 - Communicating Re

sfo = read.csv("sanfrancisco_incidents_summer_2014.csv",
               colClasses = c("numeric","factor","factor","factor","factor",
                              "factor","factor","NULL","NULL","numeric","numeric",
                              "NULL","NULL"))
sea = read.csv("seattle_incidents_summer_2014.csv",
```

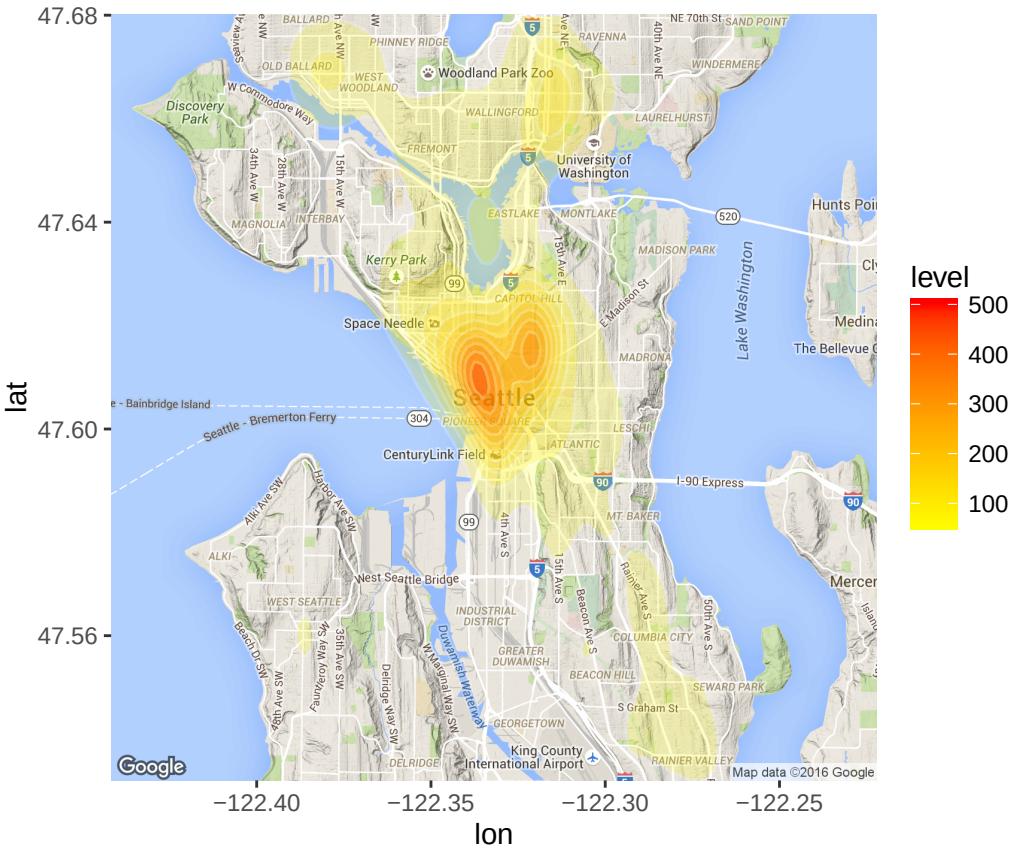
```
                colClasses = c("numeric","NULL","NULL","NULL","factor","NULL",
                               "factor","NULL","factor","NULL","NULL","factor",
                               "NULL","NULL","numeric","numeric","NULL","NULL",
                               "NULL"))

#converting Dates into POSIX classes for plots
colnames(sea)[4] = c("Date")
sea$Date = as.POSIXct(strptime(sea$Date, format="%m/%d/%Y %I:%M:%S %p"))
sea$DayOfWeek = weekdays(sea$Date)
sea = sea[,c(1,3,2,8,4:7)]

sfo$Date = as.Date(sfo$Date, format="%m/%d/%Y")
sfo$Date = as.POSIXct(paste(sfo$Date, sfo$Time))
sfo = sfo[,c(1:5,7:9)]

#tidying up column names
labels = c("IncidntNum","Category","Description","DayOfWeek","Date", "District",
           "Longitude","Latitude")
colnames(sfo) = labels
colnames(sea) = labels

sfo$Category = as.character(sfo$Category)
sfo[sfo$Category == "DRIVING UNDER THE INFLUENCE","Category"] = "DUI"
sfo[sfo$Category == "WEAPON LAWS","Category"] = "WEAPON"
sfo[sfo$Category == "LARCENY/THEFT","Category"] = "BURGLARY"
sfo[sfo$Category == "TRESPASS","Category"] = "BURGLARY"
sfo[sfo$Category == "STOLEN PROPERTY","Category"] = "BURGLARY"

sea$Category = as.character(sea$Category)
sea[sea$Category == "FORGERY","Category"] = "FORGERY/COUNTERFEITING"
sea[sea$Category == "COUNTERFEIT","Category"] = "FORGERY/COUNTERFEITING"
sea[sea$Category == "BURGLARY-SECURE PARKING-RES","Category"] = "BURGLARY"
sea[sea$Category == "STOLEN PROPERTY","Category"] = "BURGLARY"
sea[sea$Category == "BIKE THEFT","Category"] = "BURGLARY"
sea[sea$Category == "SHOPLIFTING","Category"] = "BURGLARY"
sea[sea$Category == "PICKPOCKET","Category"] = "BURGLARY"
sea = sea[!(sea$District == ""),]
sea = sea[!(sea$District == "99"),]


catforcomp = c("ASSAULT","BURGLARY","DISORDERLY CONDUCT","DUI",
               "FORGERY/COUNTERFEITING","PROSTITUTION","ROBBERY",
               "VEHICLE THEFT","WEAPON")
sfo = sfo[sfo$Category %in% catforcomp,]
sfo$Category = as.factor(sfo$Category)
sfo = sfo[order(sfo$Category),]

sea = sea[sea$Category %in% catforcomp,]
sea$Category = as.factor(sea$Category)
sea = sea[order(sea$Category),]

sfo$Hour = hour(sfo$Date)
sea$Hour = hour(sea$Date)
```

Geographic map based on the crime frequnecy

```
library(maps)
library(ggmap)

sfoMap = get_map(location="San Francisco", zoom=13)
seaMap = get_map(location="seattle", zoom=12)

ggmap(sfoMap) +
     stat_density2d(data=sfo, aes(x=Longitude, y=Latitude, fill=..level..), geom="polygon", alpha=0.2) +
     scale_fill_gradient(low="yellow", high="red")
```



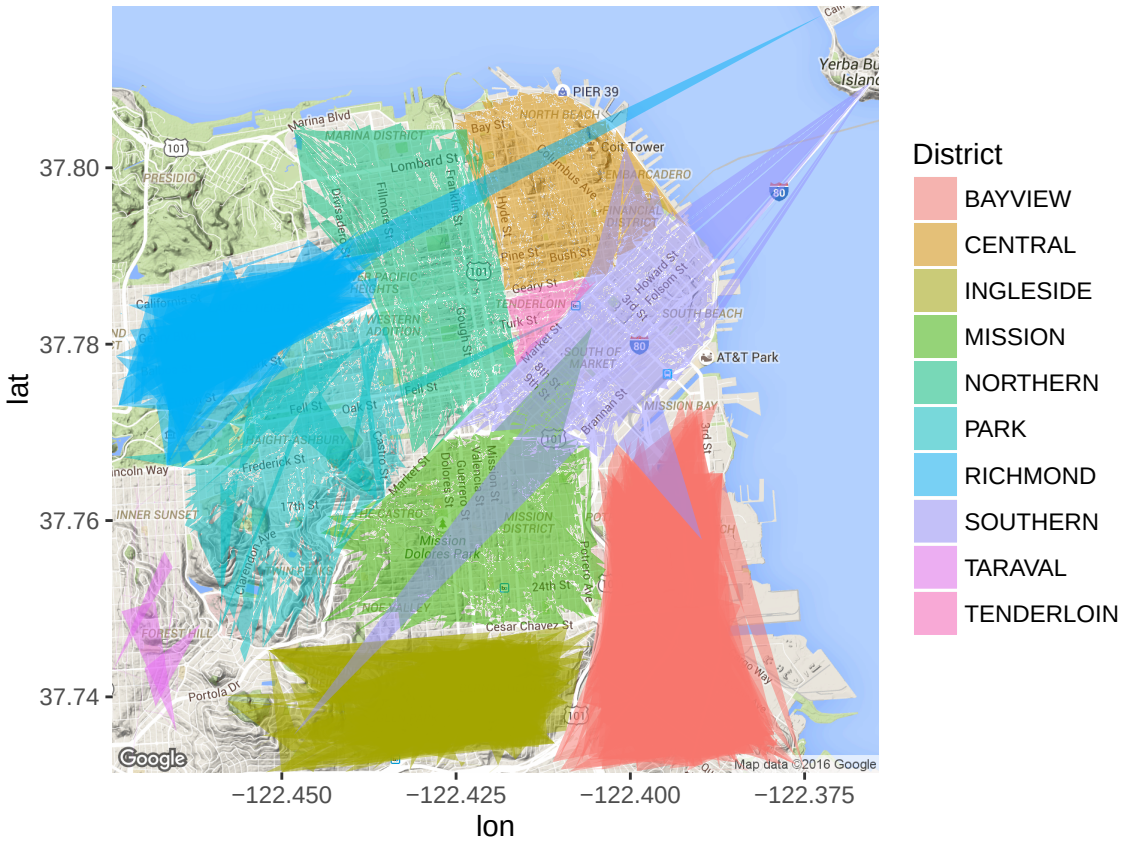

```
ggmap(seaMap) +
     stat_density2d(data=sea, aes(x=Longitude, y=Latitude, fill=..level..), geom="polygon", alpha=0.2) +
     scale_fill_gradient(low="yellow", high="red")
```



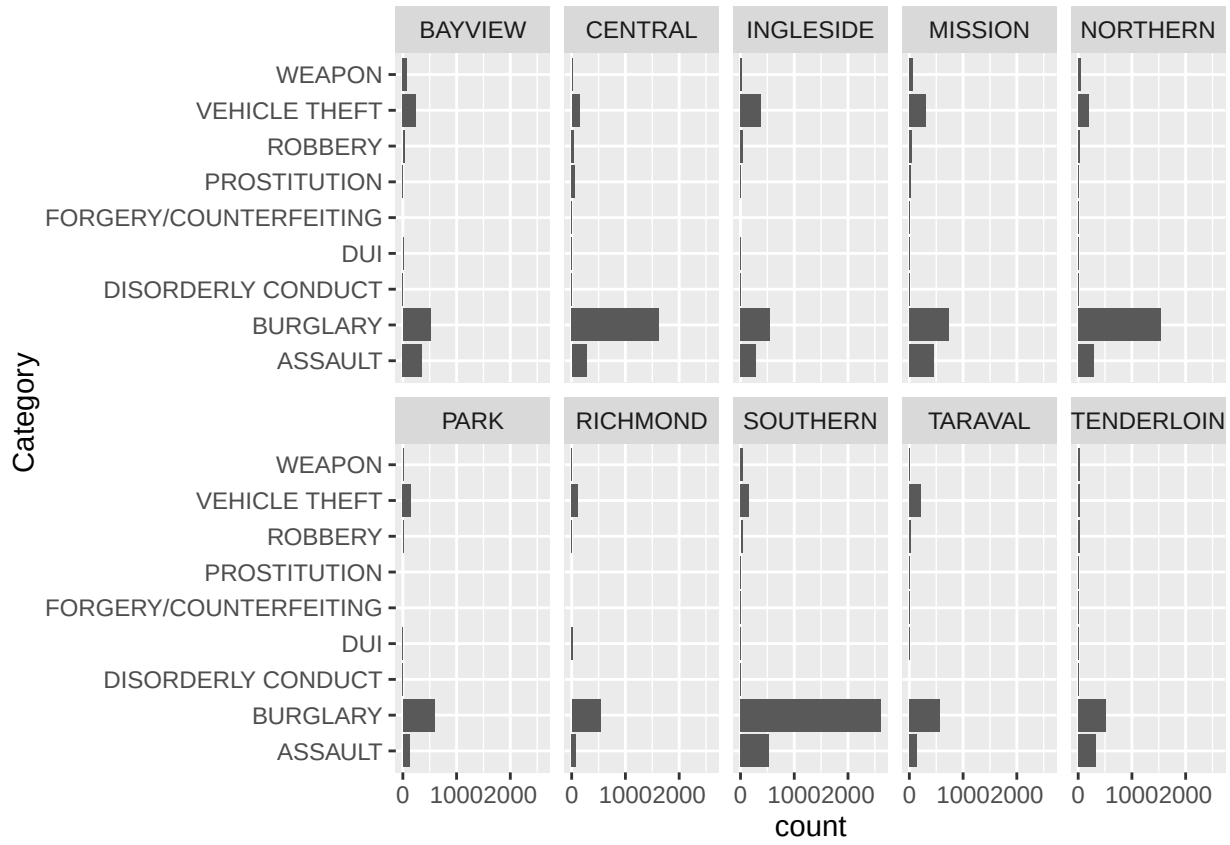
In the map, we do see that the most frequent crime area is centered at the metropolitan area. Which is expected, since relatively larger number of people are living in there compare to other suburban.

Geographic map based on Crime types vs District

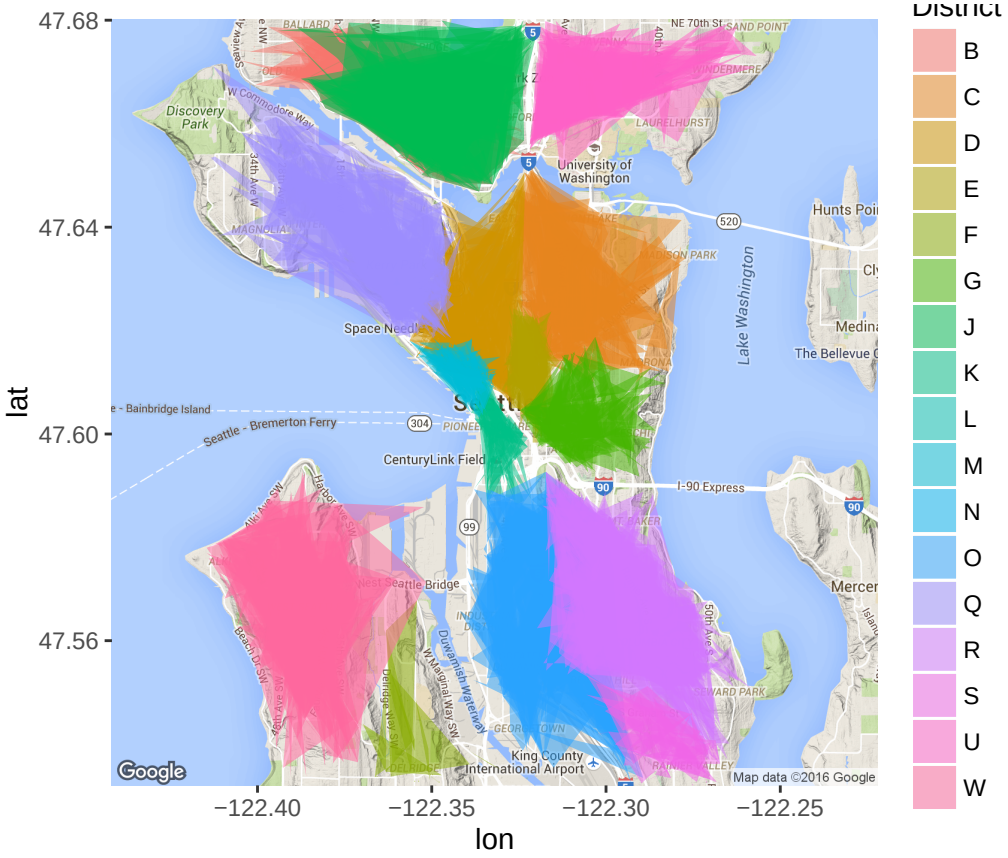
```
ggmap(sfoMap) +
    geom_polygon(data=sfo, aes(x=Longitude, y=Latitude, group=District,fill=District), colour=NA, alpha=
```



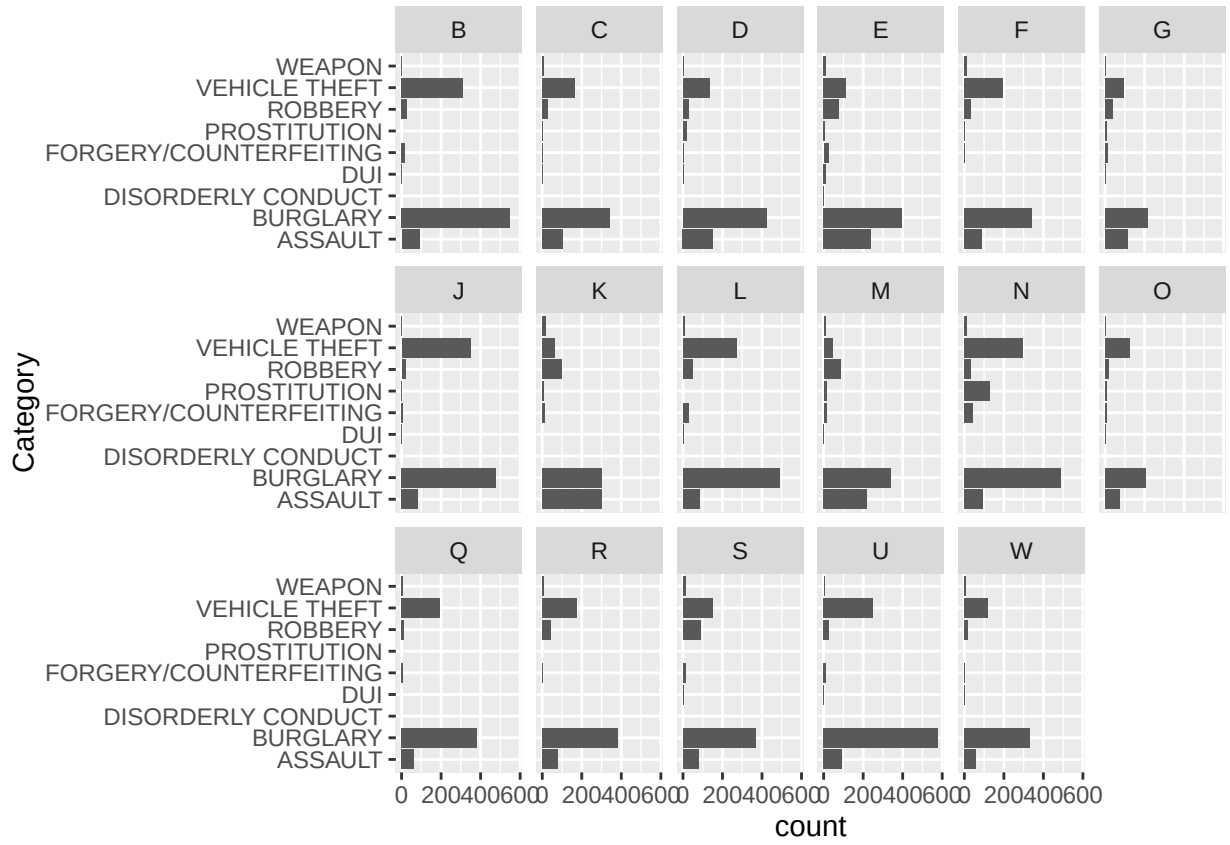

```
ggplot(data=sfo, aes(Category)) + stat_count() +
    facet_wrap(~District, nrow = 2) + coord_flip()
```




```
ggmap(seaMap) +
    geom_polygon(data=sea, aes(x=Longitude, y=Latitude, group=District,fill=District), colour=NA, alpha=
```



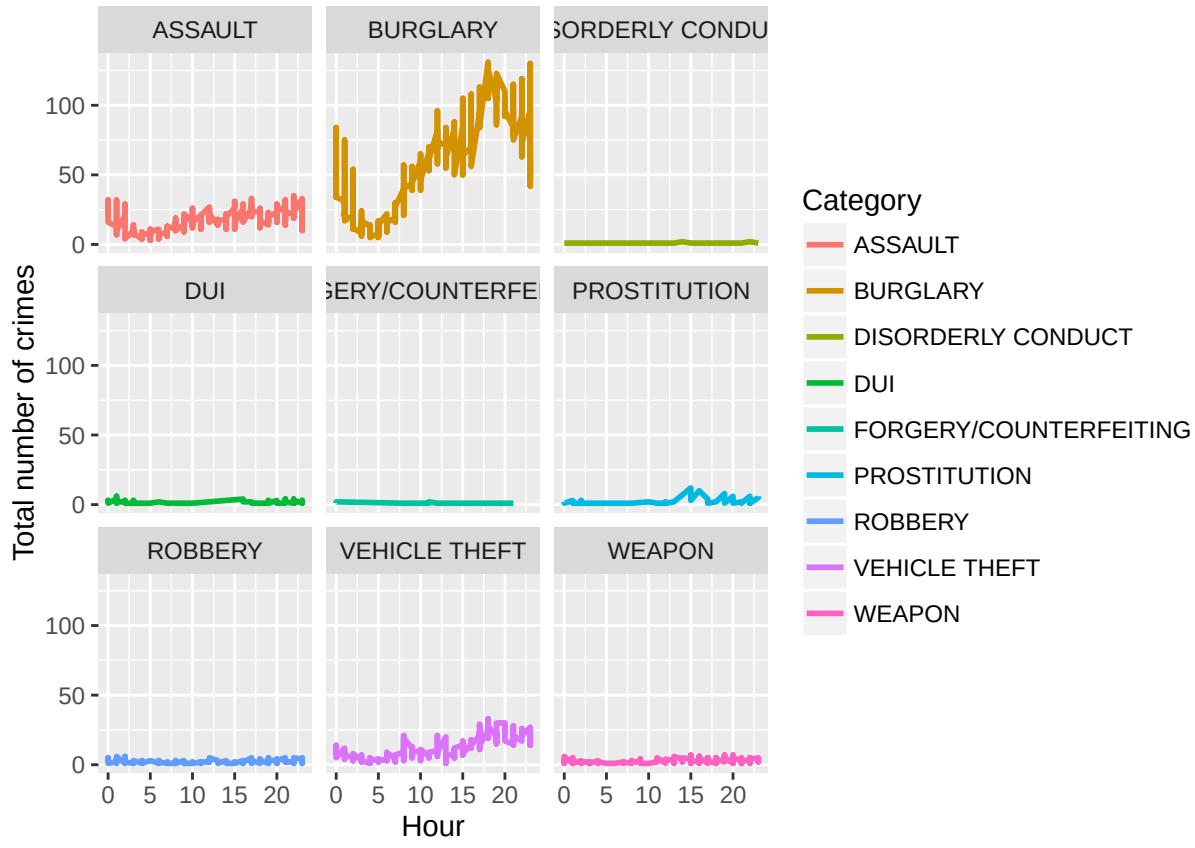

```
ggplot(data=sea, aes(Category)) + stat_count() +
    facet_wrap(~District, nrow = 3) + coord_flip()
```



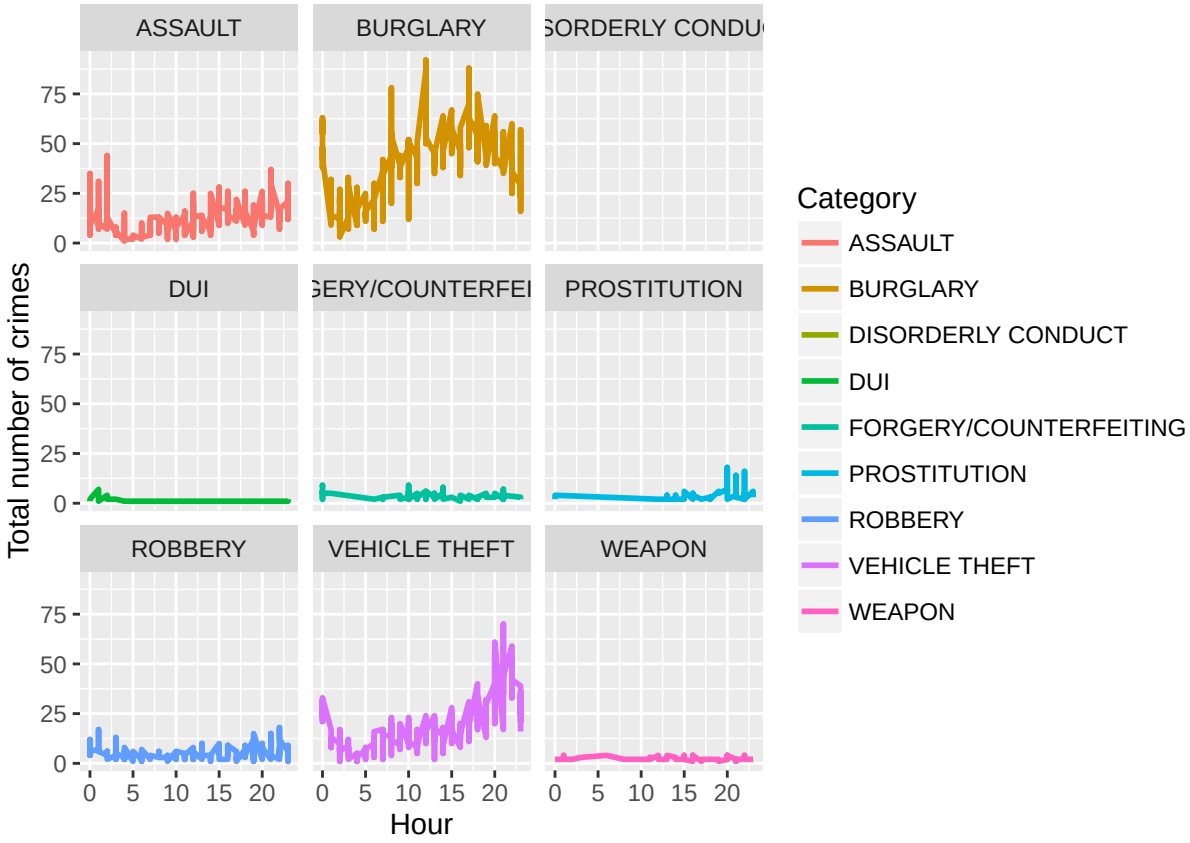
The graph represents that the burglary is the major crime in the San Francisco especially at the southern and central district. In other hadn, in Seattle, burglary, assult and vehicle theft are major crime types especially at the district B, J, and N.

Graph based on Crime types vs Hours

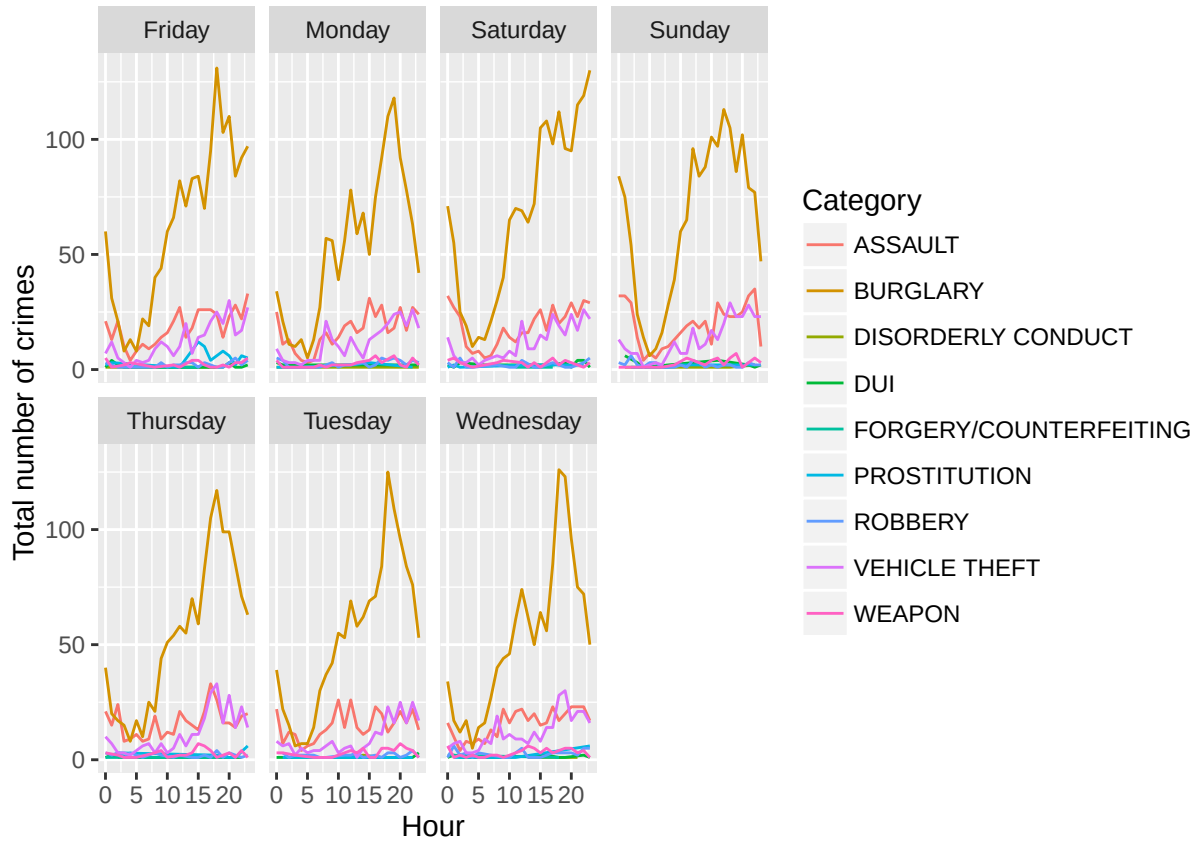
```
sfo_cate = tally(group_by(sfo,Hour,Category, DayOfWeek))
sea_cate = tally(group_by(sea,Hour,Category, DayOfWeek))

ggplot(sfo_cate, aes(x=Hour, y=n)) +
     geom_line(aes(group=Category, color=Category), size=1) +
     facet_wrap(~Category, nrow = 3) +
     ylab("Total number of crimes")
```



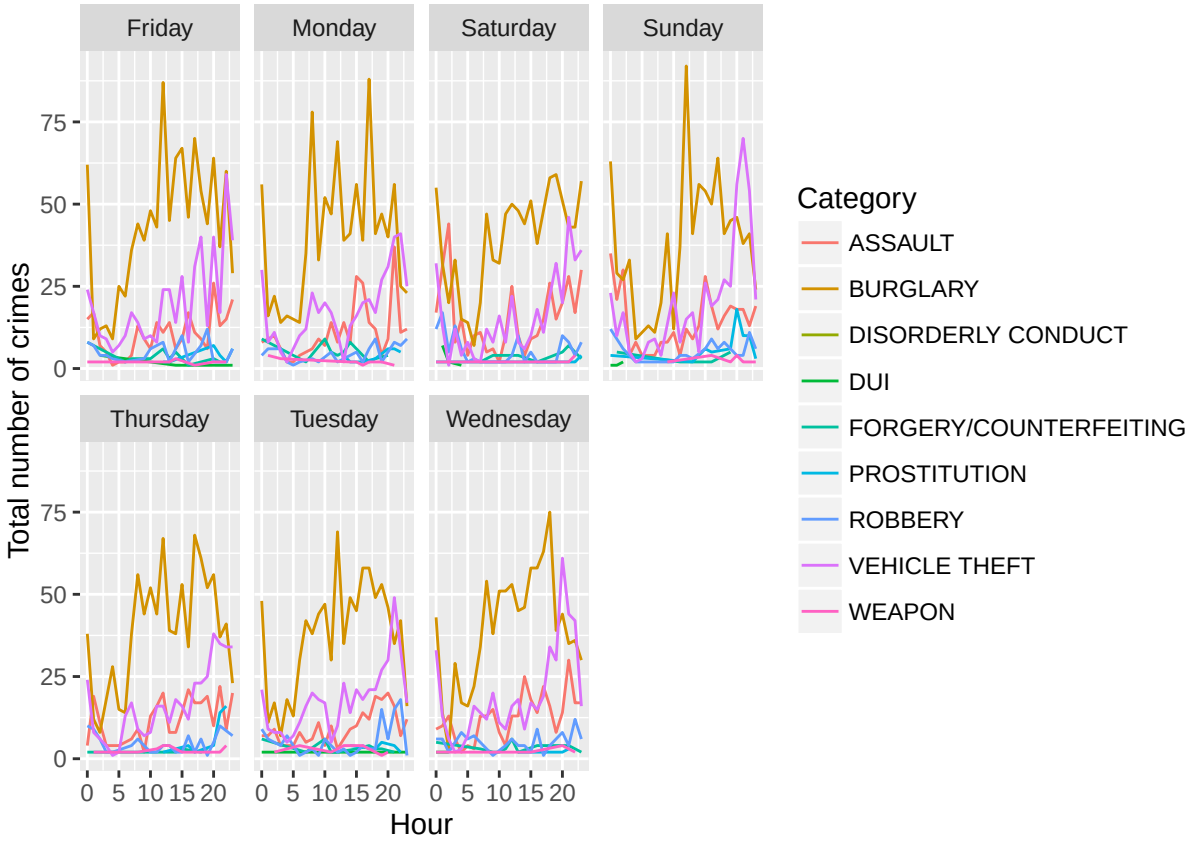

```
ggplot(sea_cate, aes(x=Hour, y=n)) +
     geom_line(aes(group=Category, color=Category), size=1) +
     facet_wrap(~Category, nrow = 3) +
     ylab("Total number of crimes")
```




```
ggplot(sfo_cate, aes(x=Hour, y=n)) +
     geom_line(aes(color=Category)) +
     facet_wrap(~DayOfWeek,nrow = 2) +
     ylab("Total number of crimes")
```



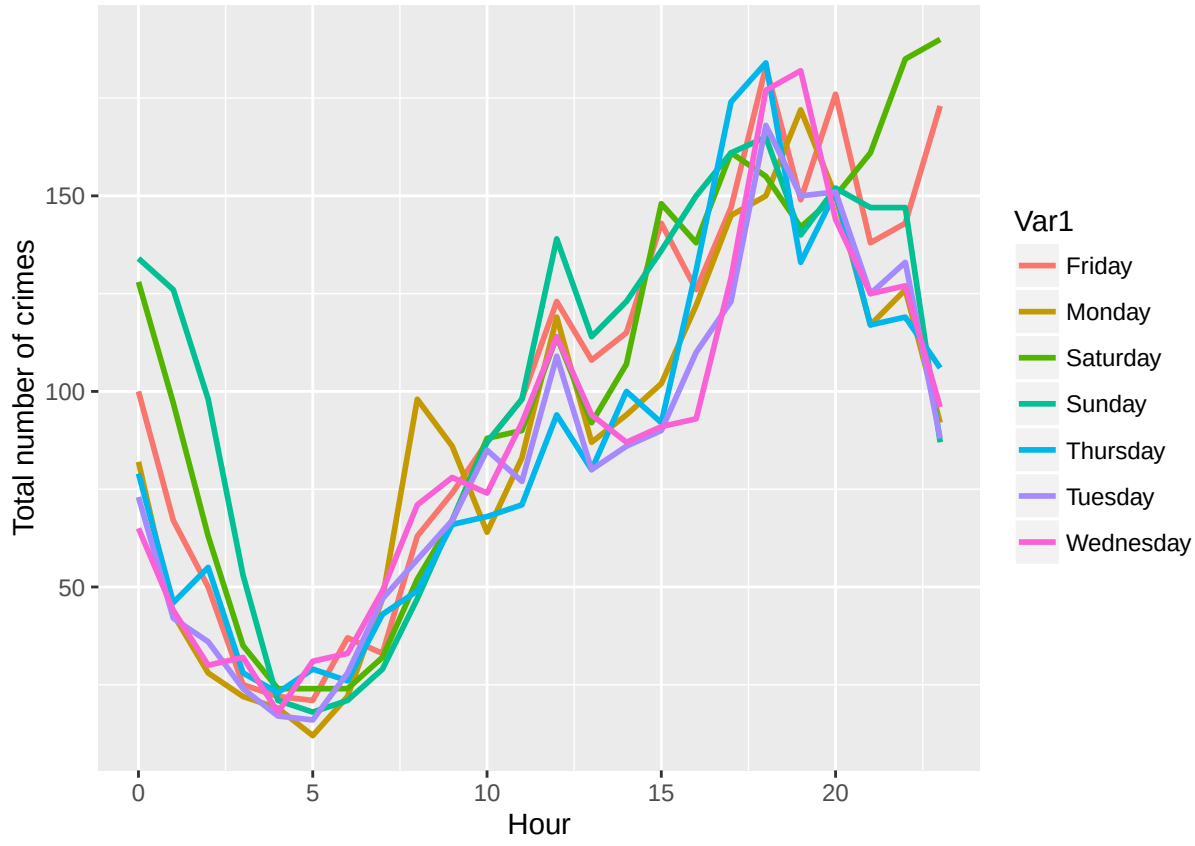

```
ggplot(sea_cate, aes(x=Hour, y=n)) +
     geom_line(aes(color=Category)) +
     facet_wrap(~DayOfWeek,nrow = 2) +
     ylab("Total number of crimes")
```



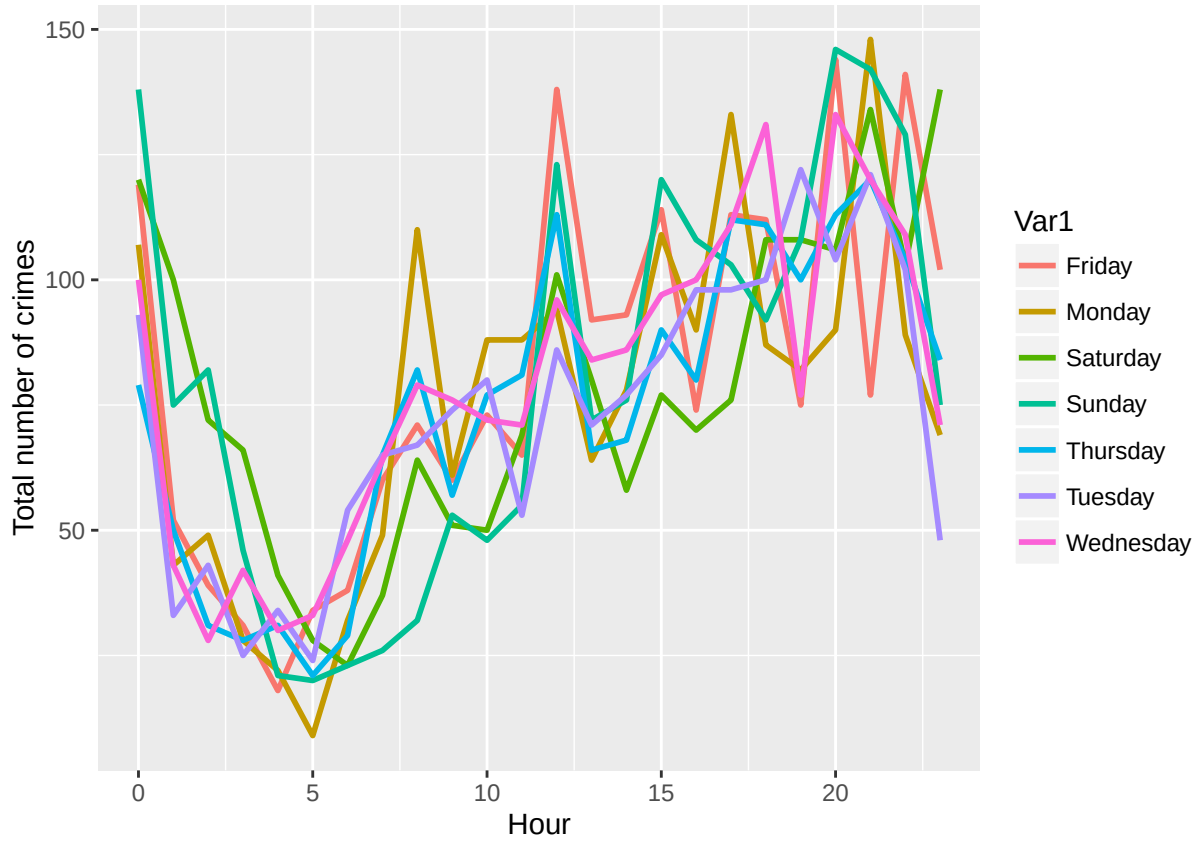
Burglary and weapons assults are the most frequent crime in both cities.

Graph based on Total number of crimes vs Hours

```
sfo_DayOfWeek = as.data.frame(table(sfo$DayOfWeek, sfo$Hour))
sea_DayOfWeek = as.data.frame(table(sea$DayOfWeek, sea$Hour))
sfo_DayOfWeek$Hour = as.numeric(as.character(sfo_DayOfWeek$Var2))
sea_DayOfWeek$Hour = as.numeric(as.character(sea_DayOfWeek$Var2))

# Change the colors
ggplot(sfo_DayOfWeek, aes(x=Hour, y=Freq)) +
     geom_line(aes(group=Var1, color=Var1), size=1) +
     ylab("Total number of crimes")
```



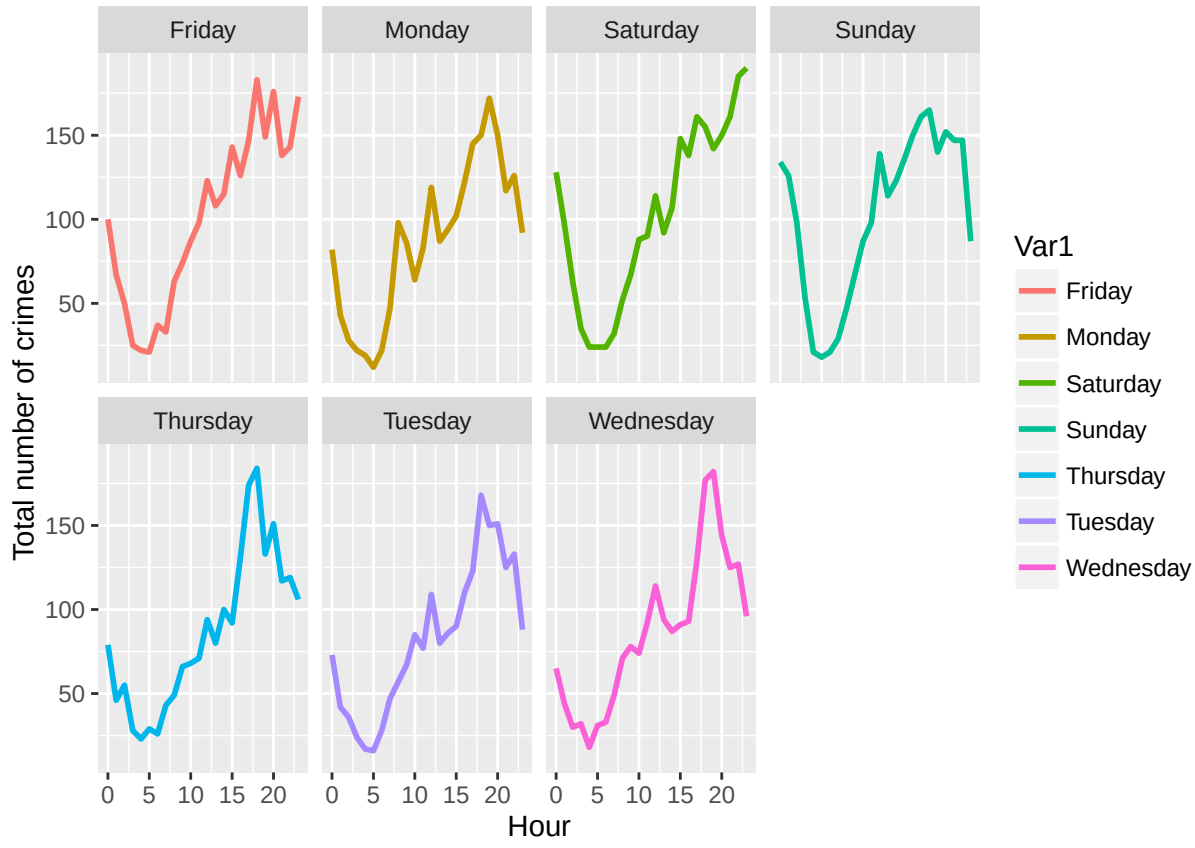

```
ggplot(sea_DayOfWeek, aes(x=Hour, y=Freq)) +
     geom_line(aes(group=Var1, color=Var1), size=1) +
     ylab("Total number of crimes")
```



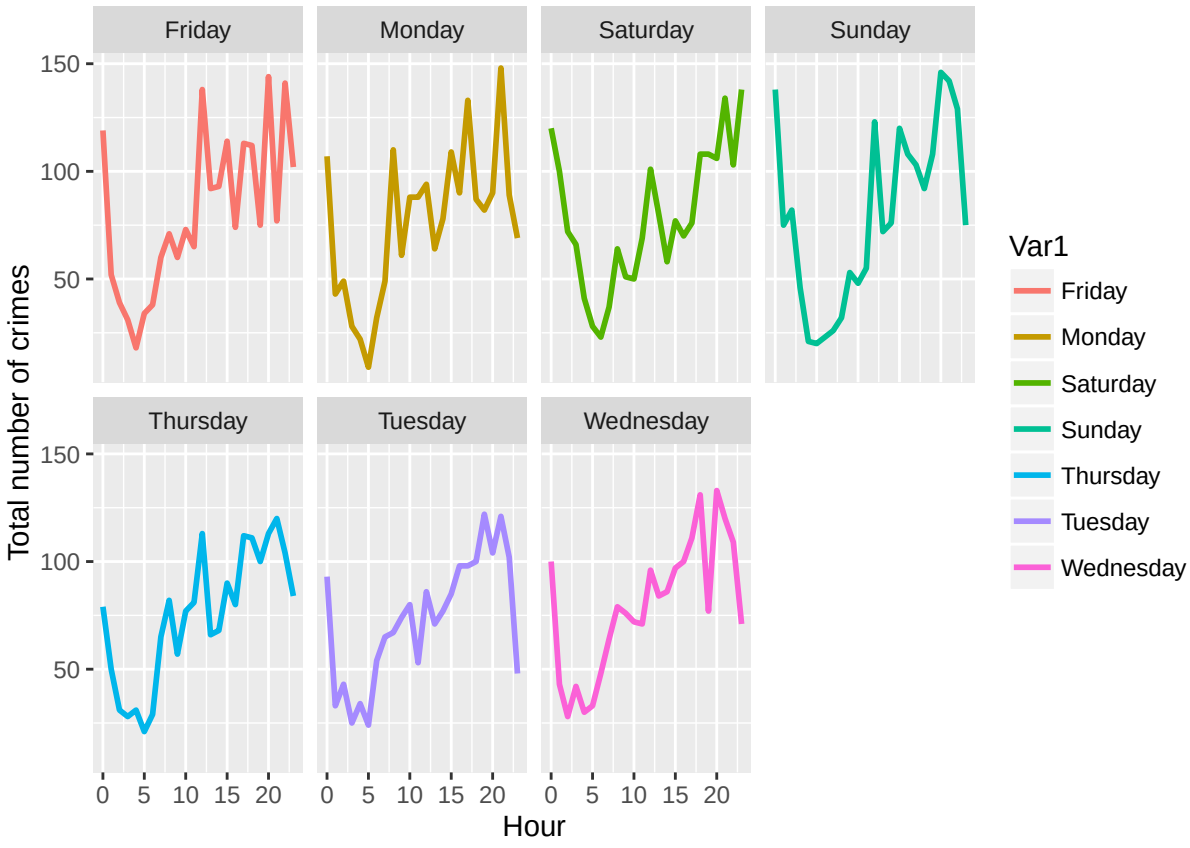
We are seeing that major scrim rate is peek at around 5PM in both city and the lowest scrim rate is showing at 5 AM.

Total number of crimes vs Days

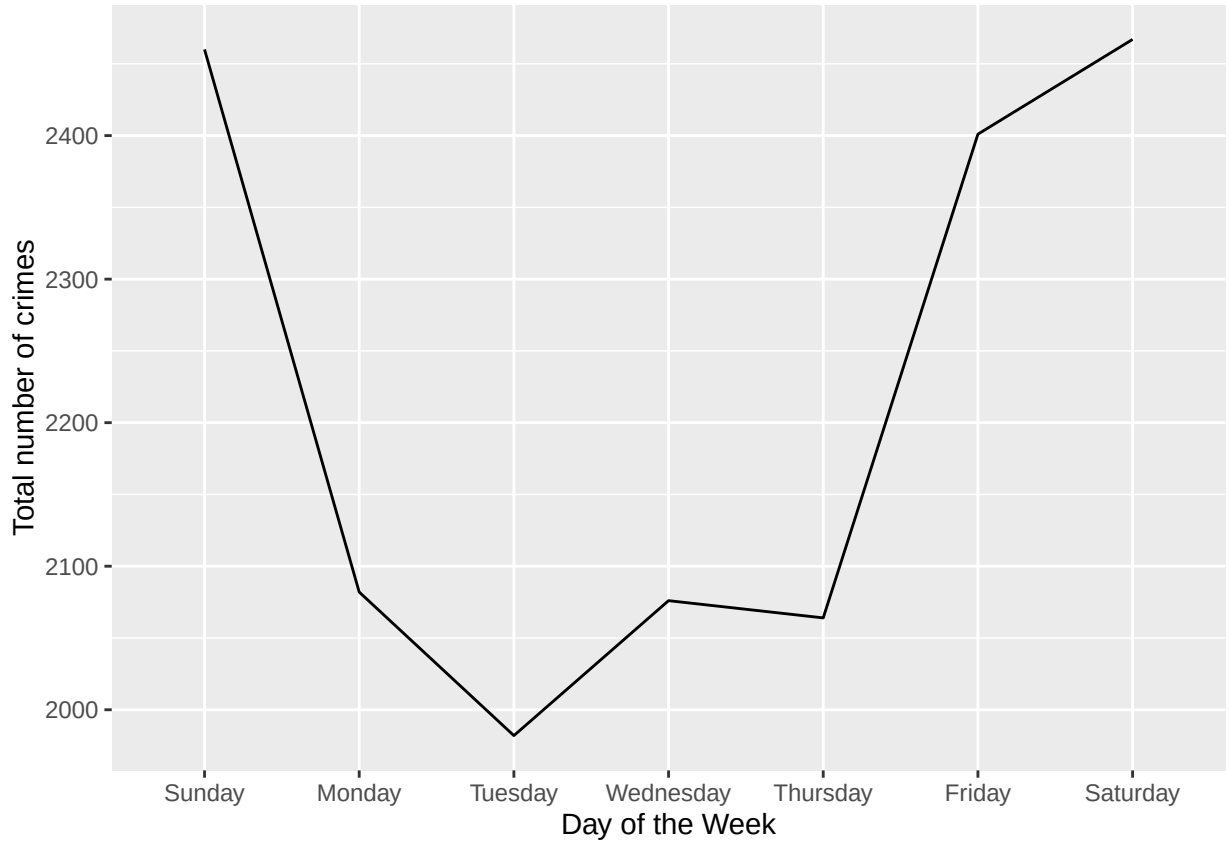
```
ggplot(sfo_DayOfWeek, aes(x=Hour, y=Freq)) +
     geom_line(aes(group=Var1, color=Var1), size=1) +
     facet_wrap(~Var1, nrow = 2) +
     ylab("Total number of crimes")
```



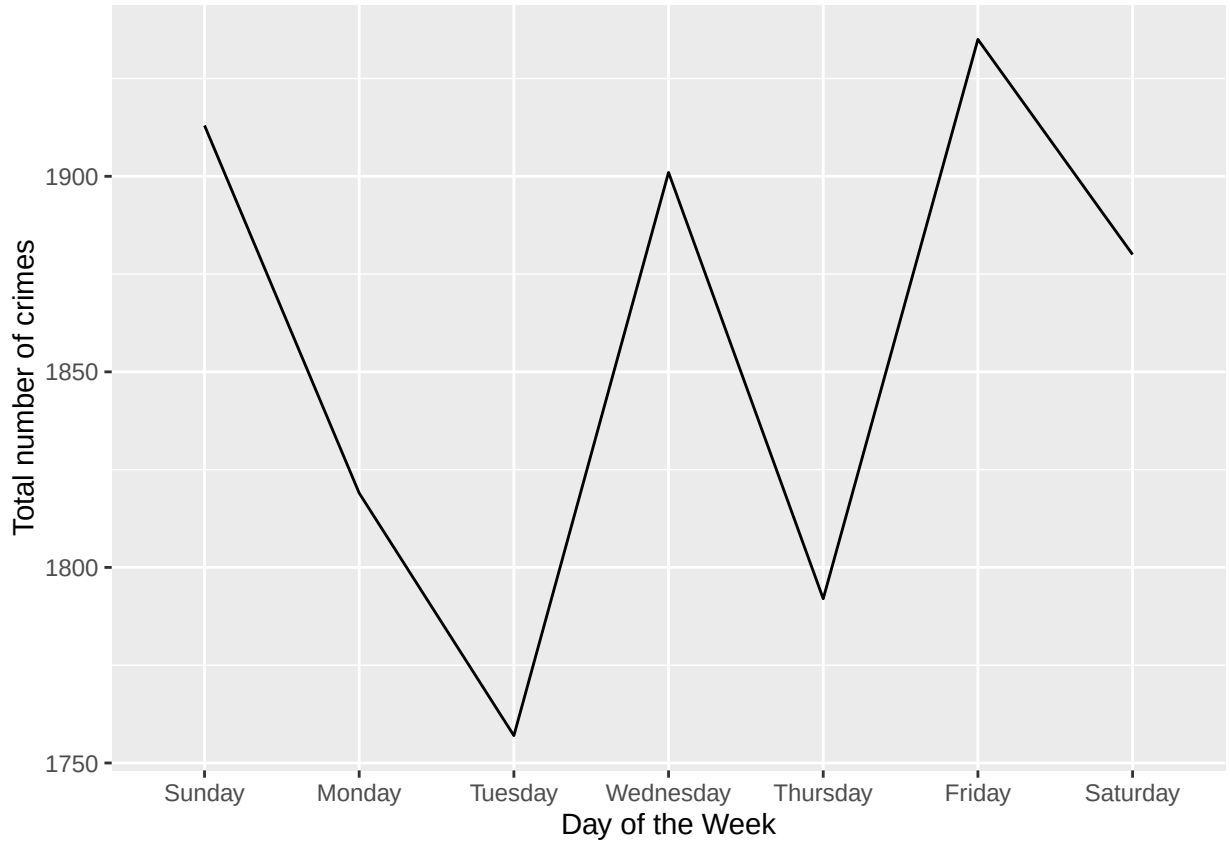

```
ggplot(sea_DayOfWeek, aes(x=Hour, y=Freq)) +
     geom_line(aes(group=Var1, color=Var1), size=1) +
     facet_wrap(~Var1, nrow = 2) +
     ylab("Total number of crimes")
```




```
sfo_DayOfWeek = as.data.frame(table(sfo$DayOfWeek))
sea_DayOfWeek = as.data.frame(table(sea$DayOfWeek))

# Make the "Var1" variable an chronological order instead of an alphabetical order.
sfo_DayOfWeek$Var1 = factor(sfo_DayOfWeek$Var1, ordered=TRUE, levels=c("Sunday", "Monday", "Tuesday", "
sea_DayOfWeek$Var1 = factor(sfo_DayOfWeek$Var1, ordered=TRUE, levels=c("Sunday", "Monday", "Tuesday", "

ggplot(sfo_DayOfWeek, aes(x=Var1, y=Freq)) + geom_line(aes(group=1)) +
     xlab("Day of the Week") + ylab("Total number of crimes")
```




```
ggplot(sea_DayOfWeek, aes(x=Var1, y=Freq)) + geom_line(aes(group=1)) +
     xlab("Day of the Week") + ylab("Total number of crimes")
```




```
df = aggregate(sfo_DayOfWeek$Freq, by=list(Category=sfo_DayOfWeek$Var1), FUN=sum)
arrange(df, desc(x))
```

```
##    Category    x
## 1  Saturday 2467
## 2    Sunday 2460
## 3    Friday 2401
## 4    Monday 2082
## 5 Wednesday 2076
## 6  Thursday 2064
## 7   Tuesday 1982
```

```
df = aggregate(sea_DayOfWeek$Freq, by=list(Category=sea_DayOfWeek$Var1), FUN=sum)
arrange(df, desc(x))
```

```
##    Category    x
## 1    Friday 1935
## 2    Sunday 1913
## 3 Wednesday 1901
## 4  Saturday 1880
## 5    Monday 1819
## 6  Thursday 1792
## 7   Tuesday 1757
```

And finally, crime is committed during a weekend at most.

Conclusion

The data shows quite clearly that violent crimes (i.e. assaults or robberies) occur much more frequently in city centers or areas still in the process of gentrification than other residential neighborhoods. In Seattle, there is also a clear peak of assault crime in the evening hours and especially Friday and Saturday nights. San Francisco has a similar peak in violent crime during those weekend nights, but it is not as clearly defined as in Seattle. When ignoring the day of the week, the number of assaults in San Francisco only fluctuates without a strong peak during the hours of around 8am to 2am. The lack of a distinctive peak in assault rates during the evening/night might indicate missing data (i.e. certain type of crimes categorized as assault in Seattle are classified as burglaries in San Francisco) or different underlying social mechanisms/movements.