Crime Analytics

Morgane Flauder

November 29, 2015

Don't leave your car alone in Seattle, especially on Wednesdays and Thurdays

For this assignment, I chose to focus on the repartition of the types of crimes in both Seattle and San Francisco. Since there are a lot of crime categories, and they are not the same in the two datasets, I chose to focus on thefts only. I created 10 new categories and I assigned each of the thefts of the two datasets to one of those categories.

The new categories I selected are :

- Burglary
- Embezzlement
- Robbery
- Bike theft
- Other vehicle theft
- Shoplifting
- Car prowl (Breaking into a car to steal what's inside)
- Pickpocket
- Purse snatch
- Other types of theft (Mail theft, extorsion, license plate theft...)

For this purpose, I first created two new datasets that contained only the thefts (one for Seattle and one for San Francisco). Then, I created a new variable "type" and I assigned each theft to its category.

Here is the code I used to transform the data :

```
# Reading files
san_francisco <- read.csv("sanfrancisco_incidents_summer_2014.csv")
seattle <- read.csv("seattle_incidents_summer_2014.csv")

################################################
#  Keeping only the crimes attributed to theft  #
################################################

# Categories corresponding to thefts already in the San Francisco dataset
san_francisco_theft_category <- c("BURGLARY",
                                  "LARCENY/THEFT",
                                  "ROBBERY",
                                  "VEHICLE THEFT",
                                  "EMBEZZLEMENT",
                                  "EXTORSION")
san_francisco_theft_descript <- c("LOST/STOLEN LICENSE PLATE")

# Creating a new data frame san_francisco_theft for San Francisco
# with only the thefts
```

```
san_francisco_theft <- rbind(san_francisco[san_francisco$Category %in%
                                           san_francisco_theft_category,],
                             san_francisco[san_francisco$Descript %in%
                                           san_francisco_theft_descript,])

# To get rid of the unused factors
san_francisco_theft <- droplevels(san_francisco_theft)


# Categories corresponding to thefts already in the Seattle dataset
seattle_theft_category <- c("BIKE THEFT",
                            "BURGLARY",
                            "BURGLARY-SECURE PARKING-RES",
                            "CAR PROWL",
                            "EMBEZZLE",
                            "MAIL THEFT",
                            "OTHER PROPERTY",
                            "PICKPOCKET",
                            "PURSE SNATCH",
                            "ROBBERY",
                            "SHOPLIFTING",
                            "THEFT OF SERVICES",
                            "VEHICLE THEFT")

# Creating a new data frame seattle_theft for Seattle with only the thefts
seattle_theft <- seattle[seattle$Summarized.Offense.Description %in%
                                           seattle_theft_category,]

# To get rid of the unused factors
seattle_theft <- droplevels(seattle_theft)

############################################################
#  Creation of types of theft common in the 2 cities  #
############################################################

# Constructing new categories for San Francisco (new variable : type)
for (i in 1:dim(san_francisco_theft)[1]) {
      # Going through the Category variable
      if (san_francisco_theft$Category[i] == "BURGLARY"){
            san_francisco_theft$type[i] <- "BURGLARY"
      } else if (san_francisco_theft$Category[i] == "EMBEZZLEMENT") {
            san_francisco_theft$type[i] <- "EMBEZZLEMENT"
      } else if (san_francisco_theft$Category[i] == "ROBBERY") {
            san_francisco_theft$type[i] <- "ROBBERY"
      } else if (san_francisco_theft$Category[i] == "LARCENY/THEFT" ||
                 san_francisco_theft$Category[i] == "OTHER OFFENSES") {
            san_francisco_theft$type[i] <- "OTHER TYPES OF THEFT"
      } else if (san_francisco_theft$Category[i] == "VEHICLE THEFT") {
            san_francisco_theft$type[i] <- "OTHER VEHICLE THEFT"
      }

      # Going through the Descript variable
      if (san_francisco_theft$Descript[i] == "ATTEMPTED SHOPLIFTING"
```

```
          || san_francisco_theft$Descript[i] == "GRAND THEFT SHOPLIFTING"
          || san_francisco_theft$Descript[i] == "PETTY THEFT SHOPLIFTING") {
            san_francisco_theft$type[i] <- "SHOPLIFTING"
      } else if (san_francisco_theft$Descript[i] == "ATTEMPTED THEFT OF A BICYCLE"
                 || san_francisco_theft$Descript[i] == "GRAND THEFT BICYCLE"
                 || san_francisco_theft$Descript[i] == "PETTY THEFT BICYCLE") {
            san_francisco_theft$type[i] <- "BIKE THEFT"
      } else if (san_francisco_theft$Descript[i] == "EMBEZZLEMENT FROM DEPENDENT OR ELDER ADULT BY CARE
            san_francisco_theft$type[i] <- "EMBEZZLEMENT"
      } else if (san_francisco_theft$Descript[i] == "ATTEMPTED THEFT FROM LOCKED VEHICLE"
                 || san_francisco_theft$Descript[i] == "GRAND THEFT FROM LOCKED AUTO"
                 || san_francisco_theft$Descript[i] == "PETTY THEFT FROM LOCKED AUTO") {
            san_francisco_theft$type[i] <- "CAR PROWL"
      } else if (san_francisco_theft$Descript[i] == "GRAND THEFT PICKPOCKET") {
            san_francisco_theft$type[i] <- "PICKPOCKET"
      } else if (san_francisco_theft$Descript[i] == "GRAND THEFT PURSESNATCH") {
            san_francisco_theft$type[i] <- "PURSE SNATCH"
      }

}

# Constructing new categories for Seattle (new variable : type)
for (i in 1:dim(seattle_theft)[1]) {
      # Going through the Summarized.Offense.Description variable
      if (seattle_theft$Summarized.Offense.Description[i] == "BIKE THEFT"){
            seattle_theft$type[i] <- "BIKE THEFT"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "BURGLARY"
                 || seattle_theft$Summarized.Offense.Description[i] == "BURGLARY-SECURE PARKING-RES"){
            seattle_theft$type[i] <- "BURGLARY"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "CAR PROWL"){
            seattle_theft$type[i] <- "CAR PROWL"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "EMBEZZLE"){
            seattle_theft$type[i] <- "EMBEZZLEMENT"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "MAIL THEFT"
                 || seattle_theft$Summarized.Offense.Description[i] == "OTHER PROPERTY"
                 || seattle_theft$Summarized.Offense.Description[i] == "THEFT OF SERVICES"){
            seattle_theft$type[i] <- "OTHER TYPES OF THEFT"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "PICKPOCKET"){
            seattle_theft$type[i] <- "PICKPOCKET"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "PURSE SNATCH"){
            seattle_theft$type[i] <- "PURSE SNATCH"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "ROBBERY"){
            seattle_theft$type[i] <- "ROBBERY"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "SHOPLIFTING"){
            seattle_theft$type[i] <- "SHOPLIFTING"
      } else if (seattle_theft$Summarized.Offense.Description[i] == "VEHICLE THEFT"){
            seattle_theft$type[i] <- "OTHER VEHICLE THEFT"
      }
}
```

Let's check the resulting datasets.

```
head(san_francisco_theft, 3)
```

```
##    IncidntNum      Category                     Descript DayOfWeek
## 3   146177923 LARCENY/THEFT GRAND THEFT FROM LOCKED AUTO    Sunday
## 4   146177531 LARCENY/THEFT GRAND THEFT FROM LOCKED AUTO    Sunday
## 11  140738711 VEHICLE THEFT                STOLEN TRUCK    Sunday
##          Date  Time PdDistrict Resolution                  Address
## 3  08/31/2014 23:30   SOUTHERN       NONE 1000 Block of MISSION ST
## 4  08/31/2014 23:30   RICHMOND       NONE      FULTON ST / 26TH AV
## 11 08/31/2014 23:00    CENTRAL       NONE     800 Block of POST ST
##            X        Y                              Location         PdId
## 3  -122.4098 37.78004 (37.7800356268394, -122.409795194505) 1.461779e+13
## 4  -122.4853 37.77252 (37.7725176473142, -122.485262988324) 1.461775e+13
## 11 -122.4158 37.78729 (37.7872932910877, -122.415821891164) 1.407387e+13
##                   type
## 3            CAR PROWL
## 4            CAR PROWL
## 11 OTHER VEHICLE THEFT
```

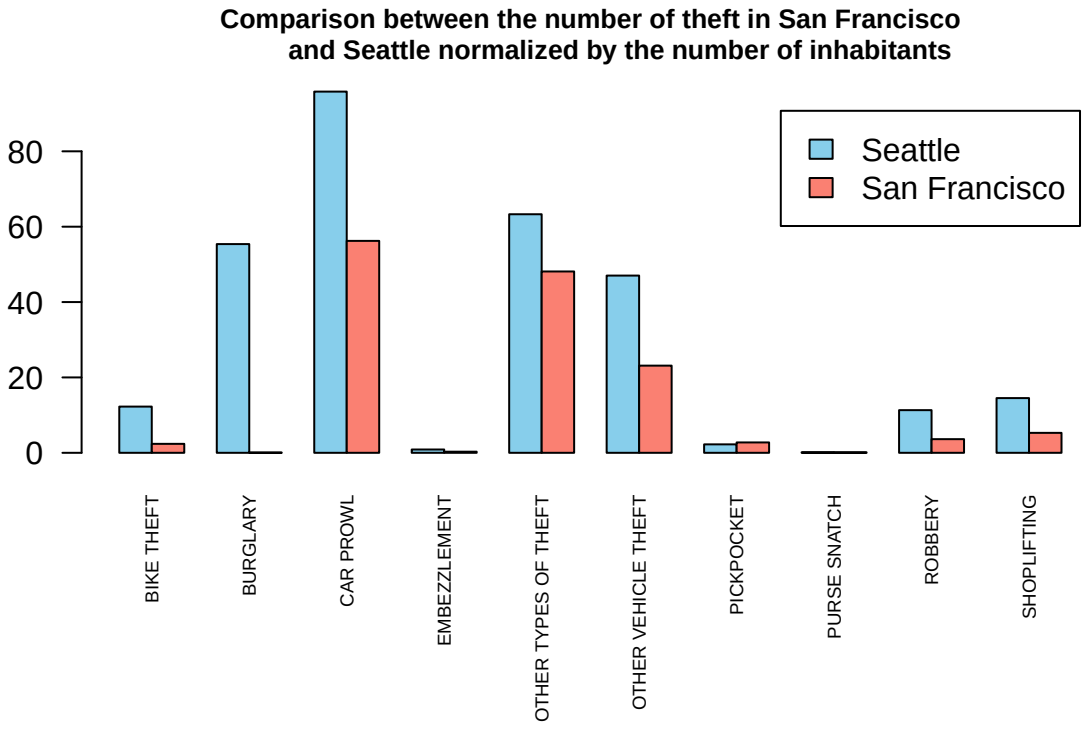
```
head(seattle_theft, 3)
```

```
##   RMS.CDW.ID General.Offense.Number Offense.Code Offense.Code.Extension
## 1     483839             2015218538         2202                      0
## 3     481375             2015210301         2316                      0
## 5     478198             2015207880         2399                      3
##         Offense.Type Summary.Offense.Code Summarized.Offense.Description
## 1 BURGLARY-FORCE-RES                2200                       BURGLARY
## 3         THEFT-MAIL                2300                     MAIL THEFT
## 5          THEFT-OTH                2300                  OTHER PROPERTY
##            Date.Reported Occurred.Date.or.Date.Range.Start
## 1 06/28/2015 10:31:00 AM           06/28/2014 10:31:00 AM
## 3 06/22/2015 09:22:00 AM           08/31/2014 09:00:00 AM
## 5 06/20/2015 11:59:00 AM           06/01/2014 11:59:00 AM
##   Occurred.Date.Range.End      Hundred.Block.Location District.Sector
## 1  06/28/2015 10:31:00 AM       6XX BLOCK OF NW 74 ST               J
## 3                          81XX BLOCK OF 11 AV SW               F
## 5  11/01/2014 12:00:00 PM 77XX BLOCK OF SUNNYSIDE AV N               J
##   Zone.Beat Census.Tract.2000 Longitude Latitude
## 1        J2          2900.301 -122.3647 47.68252
## 3        F3         11300.501 -122.3493 47.52923
## 5        J3          2700.202 -122.3294 47.68596
##                         Location Month Year               type
## 1  (47.68252427, -122.364671996)     6 2014           BURGLARY
## 3 (47.529232299, -122.349312181)     8 2014 OTHER TYPES OF THEFT
## 5 (47.685959879, -122.329378505)     6 2014 OTHER TYPES OF THEFT
```

Let's compare the number of thefts in San Francisco and Seattle. First, we need to normalize the absolute number of thefts. San Francisco has a population of about 850 000 inhabitants, whereas about 650 000 persons live in Seattle (source : Wikipedia). I chose to divide the number of thefts by 65 and 85 in order to obtain the number of thefts per 10 000 inhabitants.

```
# Comparison between the 2 cities with a muliple barplot
comp <- rbind(table(seattle_theft$type), table(san_francisco_theft$type))
rownames(comp) <- c("Seattle", "San Francisco")
comp_norm <- comp
comp_norm["Seattle",] <- comp["Seattle",] / 65
comp_norm["San Francisco",] <- comp["San Francisco",] / 85

par(mar = c(10,4,3,2))
barplot(comp_norm,
        las = 2,
        beside = T,
        legend = rownames(comp_norm),
        cex.names = 0.6,
        col = c("skyblue","salmon"))
title(main = "Comparison between the number of theft in San Francisco
        and Seattle normalized by the number of inhabitants",
      cex.main = 0.8)
```



We can first see that there are a lot more thefts in Seattle than in San Francisco. We can also see that the two categories where there are the largest gaps between the two cities are the burglaries and the car prowls, in other words the break-ins (either in a house or a car).

Since there are very few burglaries in San Francisco, I chose to study the repartition of car prowl by date in the two cities.

In order to do that, I first needed to clean the data once again. I created new datasets which contain only the type of theft, the date of the theft and the number of theft of that type on that day. Since the two datasets are encoded very differently, it is not trivial.

```
################################################
#  Analyzing different types of theft by date  #
################################################

# Conversion of the San Francisco data
dates_temp <- table(san_francisco_theft$type, san_francisco_theft$Date)
san_francisco_theft_date <- as.data.frame(dates_temp)
names(san_francisco_theft_date) <- c("type", "date", "Freq")
san_francisco_theft_date$date <- as.Date(strptime(san_francisco_theft_date$date,
                                                  "%m/%d/%Y"))

# Cleaning Seattle data
dates_temp <- table(seattle_theft$type, seattle_theft$Occurred.Date.or.Date.Range.Start)
seattle_theft_date <- as.data.frame(dates_temp)
names(seattle_theft_date) <- c("type", "date", "Freq")
seattle_theft_date$date <- as.Date(strptime(seattle_theft_date$date,
                                            "%m/%d/%Y %I:%M:%S %p"))

# Because the time is encoded with the date in the Seatle dataset, there are many
# duplicates (one entry for each time of each day for each type of theft)
# The goal is to remove these duplicates, and adding together every type of theft
# that occured on the same day (but at different times)
seattle_split_date <- split(seattle_theft_date, seattle_theft_date$date)
seattle_theft_date_clean <- san_francisco_theft_date
seattle_theft_date_clean$Freq <- rep(0, dim(san_francisco_theft_date)[1])
for (i in 1:length(seattle_split_date)) {
      temp <- as.data.frame(seattle_split_date[i])
      colnames(temp) <- c("type", "date", "Freq")
      temp <- temp[temp$Freq!=0,]
      temp <- split(temp, temp$type)
      for (j in 1:length(temp)) {
            temp2 <- as.data.frame(temp[j])
            colnames(temp2) <- c("type", "date", "Freq")
            seattle_theft_date_clean$Freq[seattle_theft_date_clean$type==temp2$type[1]
                                          & seattle_theft_date_clean$date==temp2$date[1]] <-
                  sum(temp2$Freq)
      }
}
```

Let's check our new datasets.

```
head(san_francisco_theft_date, 3)
```

```
##         type       date Freq
## 1 BIKE THEFT 2014-06-01    2
## 2   BURGLARY 2014-06-01    0
## 3  CAR PROWL 2014-06-01   28
```

```
head(seattle_theft_date_clean, 3)
```

```
##         type       date Freq
## 1 BIKE THEFT 2014-06-01    4
## 2   BURGLARY 2014-06-01   61
## 3  CAR PROWL 2014-06-01   86
```

Now, let's plot the number of car prowls occuring each day in the summer 2014 in Seattle and San Francisco. Once again, we will normalize the number of thefts by the number on inhabitants in each city in order to have the number of car prowls per 10 000 inhabitants.

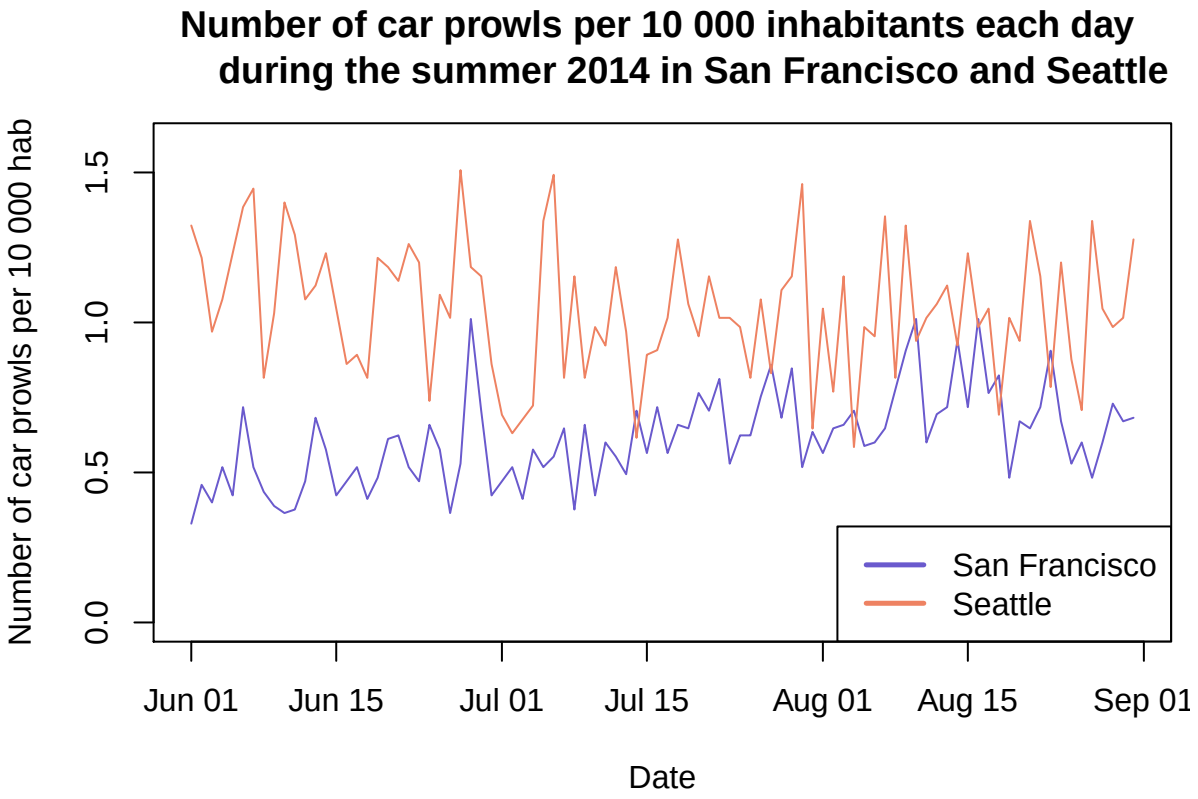
```
par(mar = c(5,4,4,2))

plot(x = san_francisco_theft_date$date[san_francisco_theft_date$type=="CAR PROWL"],
     y = san_francisco_theft_date$Freq[san_francisco_theft_date$type=="CAR PROWL"] / 85,
     type = "l",
     ylim = c(0,1.6),
     col = "slateblue",
     xlab = "Date",
     ylab = "Number of car prowls per 10 000 hab")

points(x = seattle_theft_date_clean$date[seattle_theft_date_clean$type=="CAR PROWL"],
       y = seattle_theft_date_clean$Freq[seattle_theft_date_clean$type=="CAR PROWL"] / 65,
       type = "l",
       col = "salmon2",
       ylim = c(0,1.6))

legend("bottomright",
       legend = c("San Francisco", "Seattle"),
       col = c("slateblue", "salmon2"),
       lty=c(1,1),
       lwd=c(2.5,2.5))

title(main = "Number of car prowls per 10 000 inhabitants each day
      during the summer 2014 in San Francisco and Seattle")
```



We can see that there are clearly more car prowls in Seattle. However, the difference between the two cities seems larger in June than in August. It seems that there are more car prowls in San Francisco in August than in June, and slightly less in Seattle.

We can also see that there is a lot of noise, which suggest that there are some days where there are more car prowls. One theory is that it depends on the day of the week.

In order to analyze the distribution of thefts by the day of the week, we need to add a variable to the datasets : weekday. We also need to compute the mean of the number of thefts for each day of the week.

```
#####################################################################
#  Analyzing different types of thefts by the day of the week #
#####################################################################

# San Francisco

# Adding weekday for San Francisco
san_francisco_theft_date$weekday <- as.factor(weekdays(san_francisco_theft_date$date))
levels(san_francisco_theft_date$weekday) <- c("Monday","Tuesday","Wednesday",
                                              "Thursday","Friday","Saturday","Sunday")
# Calculationg the mean of the total thefts on each day of the week for San Francisco
san_francisco_thefts_per_weekday <- by(san_francisco_theft_date$Freq,
                                       san_francisco_theft_date$weekday,
                                       mean)
# Calculating the mean of the car prowls on each day of the week for San Francisco
san_francisco_car_prowl_per_weekday <-
      by(san_francisco_theft_date$Freq[san_francisco_theft_date$type=="CAR PROWL"],
         san_francisco_theft_date$weekday[san_francisco_theft_date$type=="CAR PROWL"],
         mean)
# Calculating the mean of the burglaries on each day of the week for San Francisco
```

```
san_francisco_burglary_per_weekday <-
      by(san_francisco_theft_date$Freq[san_francisco_theft_date$type=="BURGLARY"],
         san_francisco_theft_date$weekday[san_francisco_theft_date$type=="BURGLARY"],
         mean)


# Seattle

# Adding weekday for Seattle
seattle_theft_date_clean$weekday <- as.factor(weekdays(seattle_theft_date_clean$date))
levels(seattle_theft_date_clean$weekday) <- c("Monday","Tuesday","Wednesday",
                                              "Thursday","Friday","Saturday","Sunday")
# Calculationg the mean of the thefts on each day of the week for Seattle
seattle_thefts_per_weekday <- by(seattle_theft_date_clean$Freq,
                                 seattle_theft_date_clean$weekday,
                                 mean)
# Calculationg the mean of the car prowls on each day of the week for Seattle
seattle_car_prowl_per_weekday <-
      by(seattle_theft_date_clean$Freq[seattle_theft_date_clean$type=="CAR PROWL"],
         seattle_theft_date_clean$weekday[seattle_theft_date_clean$type=="CAR PROWL"],
         mean)
# Calculationg the mean of the burglaries on each day of the week for Seattle
seattle_burglary_per_weekday <-
      by(seattle_theft_date_clean$Freq[seattle_theft_date_clean$type=="BURGLARY"],
         seattle_theft_date_clean$weekday[seattle_theft_date_clean$type=="BURGLARY"],
         mean)
```

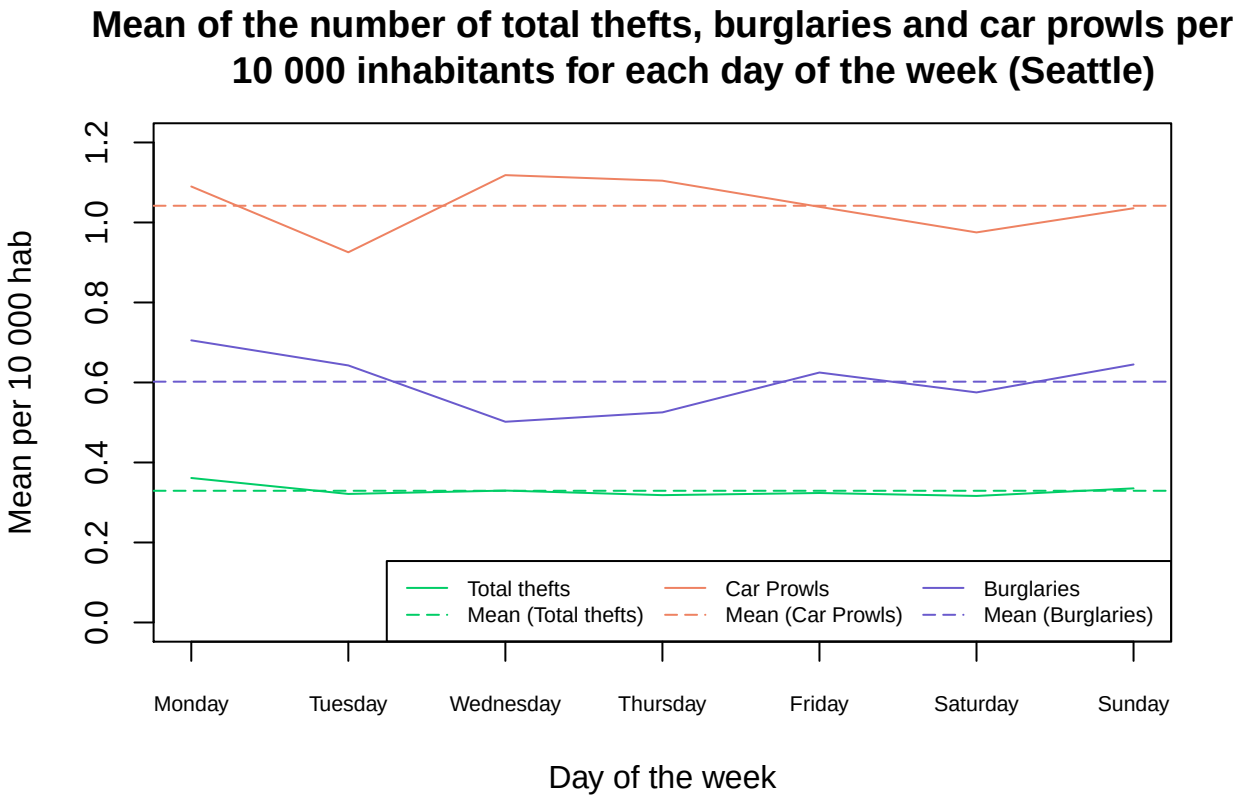
Let's plot the mean of the car prowls, burglaries and total thefts by the day of the week for San Francisco and Seattle, once again after normalization.

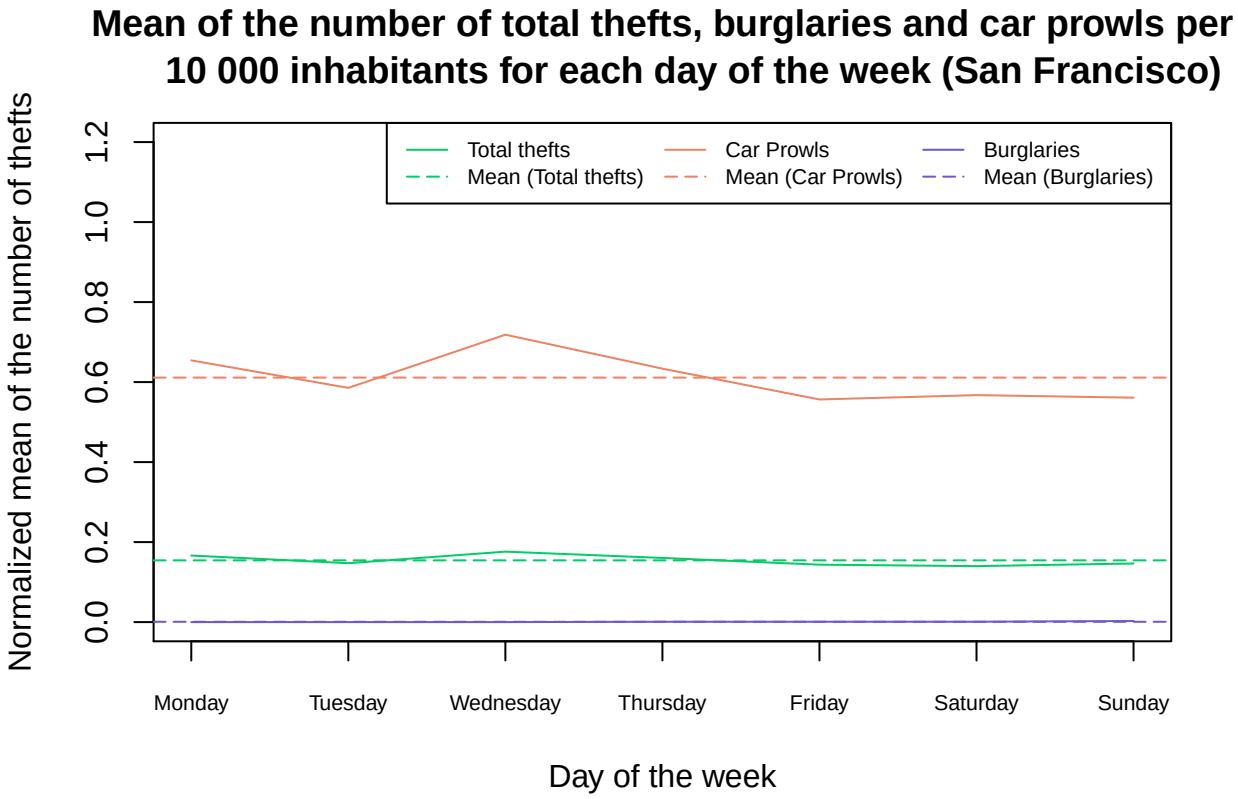
```
par(mar = c(5,4,4,2))

# Seattle plot
plot(seattle_thefts_per_weekday / 65,
     type = "l",
     xaxt = "n",
     ylim = c(0,1.2),
     col = "springgreen3",
     ylab = "Mean per 10 000 hab",
     xlab = "Day of the week")
abline(h = mean(seattle_theft_date_clean$Freq) / 65,
       lty = 5,
       col = "springgreen3")
points(seattle_car_prowl_per_weekday / 65,
       type = "l",
       col = "salmon2")
abline(h = mean(seattle_theft_date_clean$Freq[seattle_theft_date_clean$type=="CAR PROWL"]) / 65,
       lty = 5,
       col = "salmon2")
points(seattle_burglary_per_weekday / 65,
       type = "l",
       col = "slateblue")
abline(h = mean(seattle_theft_date_clean$Freq[seattle_theft_date_clean$type=="BURGLARY"]) / 65,
       lty = 5,
       col = "slateblue")
legend("bottomright",
       legend = c("Total thefts", "Mean (Total thefts)", "Car Prowls",
                  "Mean (Car Prowls)", "Burglaries", "Mean (Burglaries)"),
       col = c("springgreen3", "springgreen3", "salmon2", "salmon2", "slateblue", "slateblue"),
       lty = c(1,5,1,5,1,5),
       lwd = c(1,1,1,1,1,1),
       cex = 0.7,
       ncol = 3)
axis(1, at = 1:7,
     labels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"),
     cex.axis = 0.65)
title(main = "Mean of the number of total thefts, burglaries and car prowls per
      10 000 inhabitants for each day of the week (Seattle)")

# San Francisco plot
plot(san_francisco_thefts_per_weekday / 85,
     type = "l",
     xaxt = "n",
     ylim = c(0,1.2),
     col = "springgreen3",
     ylab = "Normalized mean of the number of thefts",
     xlab = "Day of the week")
abline(h = mean(san_francisco_theft_date$Freq) / 85,
       lty = 5,
       col = "springgreen3")
points(san_francisco_car_prowl_per_weekday / 85,
```

```
       type = "l",
       col = "salmon2")
abline(h = mean(san_francisco_theft_date$Freq[san_francisco_theft_date$type=="CAR PROWL"]) / 85,
       lty = 5,
       col = "salmon2")
points(san_francisco_burglary_per_weekday / 85,
       type = "l",
       col = "slateblue")
abline(h = mean(san_francisco_theft_date$Freq[san_francisco_theft_date$type=="BURGLARY"]) / 85,
       lty = 5,
       col = "slateblue")
legend("topright",
       legend = c("Total thefts", "Mean (Total thefts)", "Car Prowls",
                  "Mean (Car Prowls)", "Burglaries", "Mean (Burglaries)"),
       col = c("springgreen3", "springgreen3", "salmon2", "salmon2", "slateblue", "slateblue"),
       lty = c(1,5,1,5,1,5),
       lwd = c(1,1,1,1,1,1),
       cex = 0.7,
       ncol = 3)
axis(1, at = 1:7,
     labels = c("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday","Sunday"),
     cex.axis = 0.65)
title(main = "Mean of the number of total thefts, burglaries and car prowls per
      10 000 inhabitants for each day of the week (San Francisco)")
```


Mean of the number of total thefts, burglaries and car prowls per 10 000 inhabitants for each day of the week (Seattle)
Mean per 10 000 hab
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Total thefts
Mean (Total thefts)
Car Prowls
Mean (Car Prowls)
Burglaries
Mean (Burglaries)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the week


Mean of the number of total thefts, burglaries and car prowls per 10 000 inhabitants for each day of the week (San Francisco)
Normalized mean of the number of thefts
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Total thefts
Mean (Total thefts)
Car Prowls
Mean (Car Prowls)
Burglaries
Mean (Burglaries)
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the week

We confirm that there are more thefts in general in Seattle than in San Francisco, and that there are almost no burglaries in San Francisco. We can also see that the number of car prowls follow the same pattern in the two cities : there are more car prowls on Wednesdays (and Thursdays for Seattle). However, there are less than average burglaries in Seattle on these days.

To conclude, the number of break-ins (of cars and houses) are very high in Seattle compared to San Francisco. Concerning the car prowls, this is especially true on Wednesdays and Thursdays.