Common Trends In Crime Scenery Of San Francisco and Seattle

Alex Nisenboim

February 5, 2016

Introduction

As was noted in assignment instructions, the inter-city analytics is difficult due to schema discrepancies between the two input data sets. However, there are some common dimensions (i.e. the schemas are in partial agreement) that are intuitive and therefore it is not necessary to provide subtle and speculative justifications to employ them. For example, HOURS and MONTHS can be easily extracted and brought to the identical format in both data sets. This allows one to aggregate data along these dimensions and compare the results in a visual way. This is the primary goal of this work. This project is done using R programming language.

Data Preparation

To satisfy (at least partially) the Reproducible Research condition, in this section we show how there data sets are prepared.

Here San Francisco data set is read and new dimensions (Month and Hour) are added from existing dates

```
dfSF <- read.csv(file = "sanfrancisco_incidents_summer_2014.csv",
                 colClasses = c("numeric", "factor", "character", "character",
                                "character", "character", "factor", "factor",
                                "character", "numeric", "numeric", "character",
                                "numeric"))
dfSF$Hour <- as.factor(substr(dfSF$Time, 1,2))
dfSF$Date <- as.Date(dfSF$Date, "%m/%d/%Y")
dfSF$Month <- as.factor(format(dfSF$Date, "%b"))
```

Then we prepare Seattle data set in a slightly different way, but the same two dimensions are created:

```
dfSEA <- read.csv("seattle_incidents_summer_2014.csv")
MonthCode <- factor(c("Jun", "Jul", "Aug"))
dfMonth <- data.frame(Month = c(6, 7, 8), MonthCode = MonthCode)
dfSEA <- merge(dfSEA, dfMonth, by=intersect(names(dfSEA), names(dfMonth)))
dfSEA$Month <- NULL
names(dfSEA)[names(dfSEA) == "MonthCode"] <- "Month"
dfSEA$Hour <- format(strptime(as.character(dfSEA$Occurred.Date.or.Date.Range.Start),
                              format='%m/%d/%Y %I:%M:%OS %p'), "%H")
```

Incidents By Time Of Day

The idea is to calculate the crime rates by hour of day and juxtapose them for both cities

Calculation Method

Now it is possible to aggregate both sets, counting the number of incidents by hour. Dividing the counts by the total count of incidents we get the rates. This identical procedure is repeated for both data sets. Subsequently we bind them for visualization purposes.

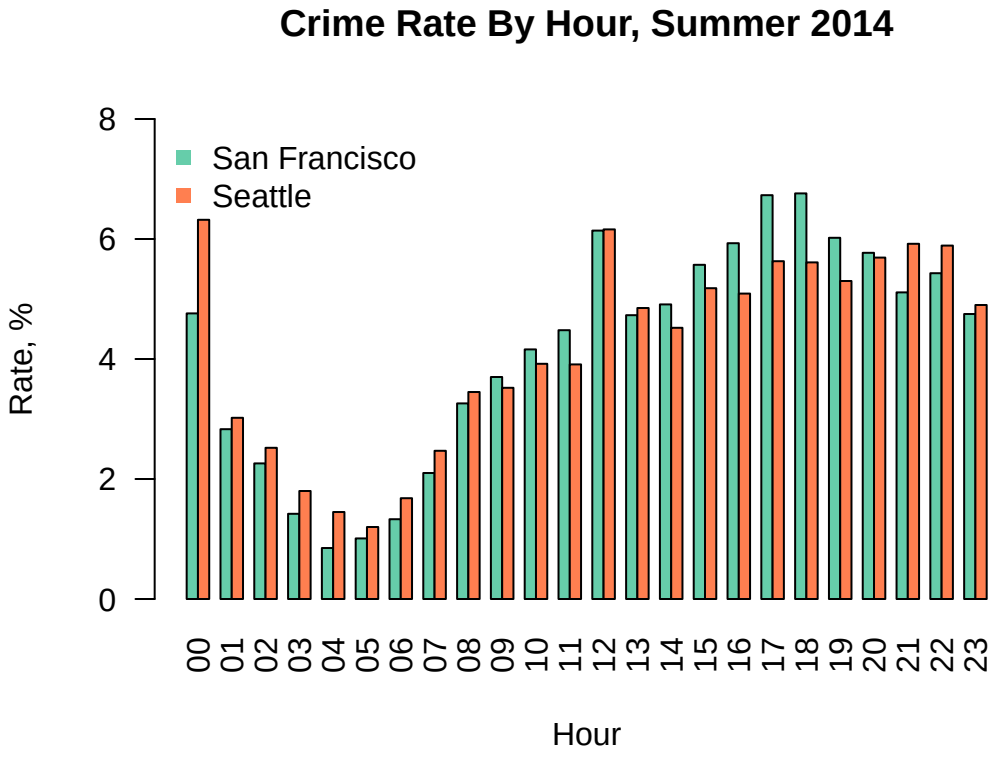
```
dfIncidentsByHour <- aggregate(IncidntNum ~ Hour, data=dfSF, FUN=length)
dfIncidentsByHour$IncidntNum <-
        dfIncidentsByHour$IncidntNum/sum(dfIncidentsByHour$IncidntNum)
names(dfIncidentsByHour)[names(dfIncidentsByHour) == "IncidntNum"] <- "Rate"
dfIncidentsByHour$City <- rep("San Francisco", 24)
dfIncidentsByHourSF <- dfIncidentsByHour

dfIncidentsByHour <- aggregate(RMS.CDW.ID ~ Hour, data=dfSEA, FUN=length)
dfIncidentsByHour$RMS.CDW.ID <-
        dfIncidentsByHour$RMS.CDW.ID/sum(dfIncidentsByHour$RMS.CDW.ID)
names(dfIncidentsByHour)[names(dfIncidentsByHour) == "RMS.CDW.ID"] <- "Rate"
dfIncidentsByHour$City <- rep("Seattle", 24)
dfIncidentsByHourSeattle <- dfIncidentsByHour

dfIncidentsByHourCity <- rbind(dfIncidentsByHourSF, dfIncidentsByHourSeattle)
dfIncidentsByHourCity$Rate <- round(dfIncidentsByHourCity$Rate*100, 2)
```

Here is the result:

```
op <- par(mar=c(5,5,5,5))
barplot(t(matrix(dfIncidentsByHourCity$Rate,nc=2)),
        names.arg=dfIncidentsByHourSF$Hour,
        ylab="Rate, %",
        ylim=c(0, 8),
        xpd = FALSE,
        beside = TRUE,
        col=c("aquamarine3","coral"),
        xlab="Hour",
        main = "Crime Rate By Hour, Summer 2014",
        las = 2)
legend("topleft", c("San Francisco","Seattle"), pch=15,
       col=c("aquamarine3","coral"),
       bty="n")
```



Discussion

The resemblance of trends is remarkable. Although the peaks of crime in San Francisco comes a little earlier (around 5 or 6 PM) than in Seattle (around 9 or 10 PM). The early morning hours are safest to walk both cities and then the dangers are creeping up steadily, reaching a plateau after dark. Lunch hour is also surprisingly dangerous.

Robbery By Time Of Day

The idea is to calculate the crime rates by hour of day for a specific crime and juxtapose them for both cities

Calculation Method

The striking similarity of trends does not change if the data sets are reduced to a particular type of incident, namely ROBBERY. The code below shows a computation method similar to the one described above, but both data sets are filtered to match the particular type of crime.

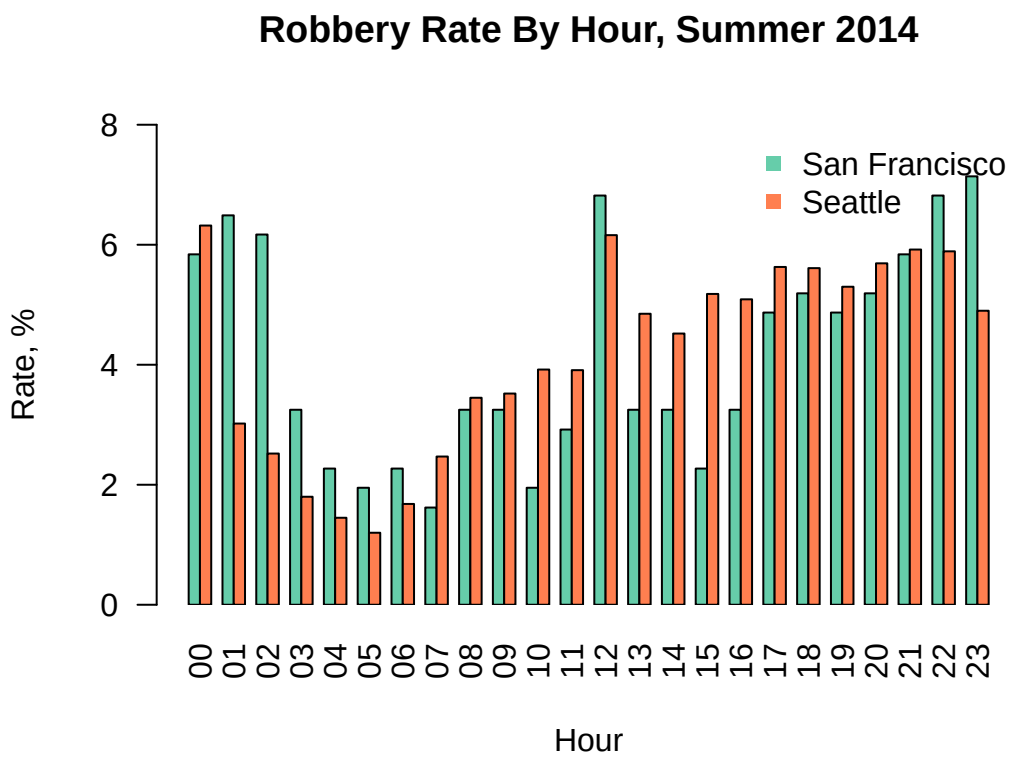
```
dfRobbery <- subset(dfSF, Category == "ROBBERY")
dfIncidentsByHour <- aggregate(IncidntNum ~ Hour, data=dfRobbery, FUN=length)
dfIncidentsByHour$IncidntNum <-
        dfIncidentsByHour$IncidntNum/sum(dfIncidentsByHour$IncidntNum)
names(dfIncidentsByHour)[names(dfIncidentsByHour) == "IncidntNum"] <- "Rate"
dfIncidentsByHour$City <- rep("San Francisco", 24)
dfIncidentsByHourSF <- dfIncidentsByHour

dfRobbery <- subset(dfSEA, Summarized.Offense.Description == "ROBBERY")
```

```
dfIncidentsByHour <- aggregate(RMS.CDW.ID ~ Hour, data=dfRobbery, FUN=length)
dfIncidentsByHour$RMS.CDW.ID <-
        dfIncidentsByHour$RMS.CDW.ID/sum(dfIncidentsByHour$RMS.CDW.ID)
names(dfIncidentsByHour)[names(dfIncidentsByHour) == "IncidntNum"] <- "Rate"
dfIncidentsByHour$City <- rep("Seattle", 24)
dfIncidentsByHourSEA <- dfIncidentsByHour

dfIncidentsByHourCity <- rbind(dfIncidentsByHourSF, dfIncidentsByHourSeattle)
dfIncidentsByHourCity$Rate <- round(dfIncidentsByHourCity$Rate*100, 2)
```

```
op <- par(mar=c(5,5,5,5))
barplot(t(matrix(dfIncidentsByHourCity$Rate,nc=2)),
        names.arg=dfIncidentsByHourSF$Hour,
        ylab="Rate, %",
        ylim=c(0, 8),
        xpd = FALSE,
        beside = TRUE,
        col=c("aquamarine3","coral"),
        xlab="Hour",
        main = "Robbery Rate By Hour, Summer 2014",
        las = 2)
legend("topright", c("San Francisco","Seattle"), pch=15,
       col=c("aquamarine3","coral"),
       bty="n")
```



Discussion

The overall picture does not change for ROBBERY. Although it seems that Seattle is a bit safer between 22:00 PM and 08:00 AM then San Francisco, while the situation reverses itself in the afternoon. Lunch hour is also surprising peak.

Incidents by Month

Here we calculate the crime rates by month and juxtapose them for both cities

Calculation Method

Rates of crime by months, however, break the pattern shown in the previous two sections. The code below aggregates the data sets by month in similar fashion, followed by the actual visualization:

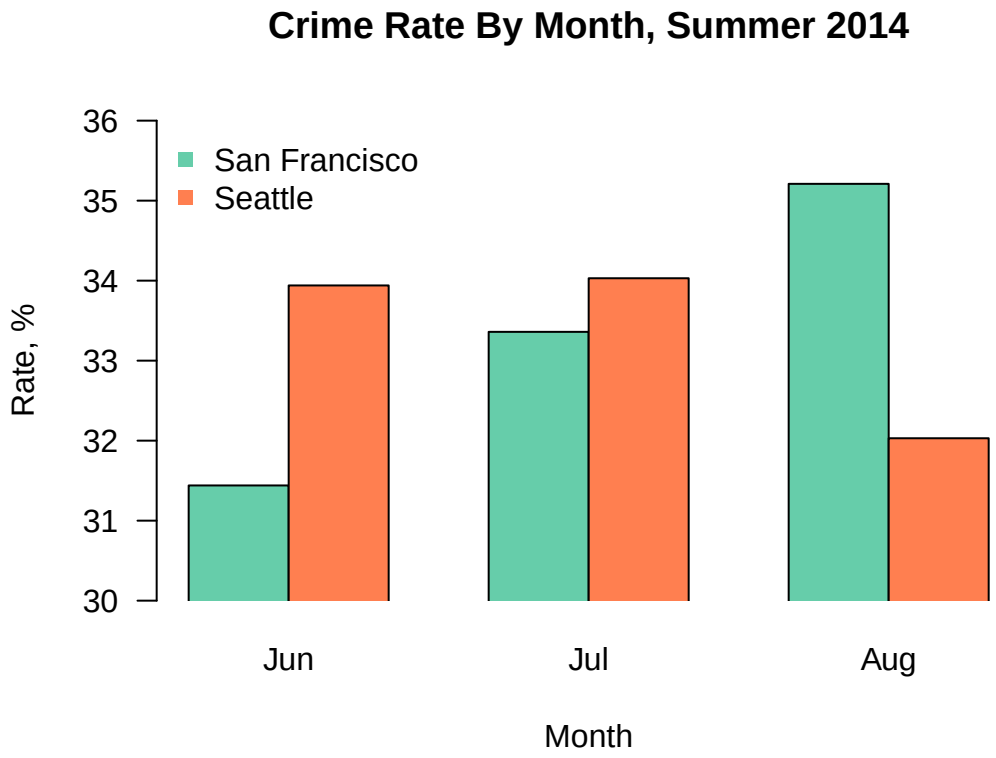
```
dfIncidentsByMonth <- aggregate(IncidntNum ~ Month, data=dfSF, FUN=length)
dfIncidentsByMonth$IncidntNum <-
        dfIncidentsByMonth$IncidntNum/sum(dfIncidentsByMonth$IncidntNum)
names(dfIncidentsByMonth)[names(dfIncidentsByMonth) == "IncidntNum"] <- "Rate"
dfIncidentsByMonth$Month <-
        factor(dfIncidentsByMonth$Month, levels = c("Jun", "Jul",  "Aug"))
dfIncidentsByMonth <- dfIncidentsByMonth[order(dfIncidentsByMonth$Month), ]
dfIncidentsByMonth$City <- rep("San Francisco", 3)
dfIncidentsByMonthSF <- dfIncidentsByMonth


dfIncidentsByMonth <- aggregate(RMS.CDW.ID ~ Month, data=dfSEA, FUN=length)
dfIncidentsByMonth$RMS.CDW.ID <-
        dfIncidentsByMonth$RMS.CDW.ID/sum(dfIncidentsByMonth$RMS.CDW.ID)
names(dfIncidentsByMonth)[names(dfIncidentsByMonth) == "RMS.CDW.ID"] <- "Rate"
dfIncidentsByMonth$Month <-
        factor(dfIncidentsByMonth$Month, levels = c("Jun", "Jul",  "Aug"))
dfIncidentsByMonth <- dfIncidentsByMonth[order(dfIncidentsByMonth$Month), ]
dfIncidentsByMonth$City <- rep("Seattle", 3)
dfIncidentsByMonthSea <- dfIncidentsByMonth

dfIncidentsByMonthCity <- rbind(dfIncidentsByMonthSF, dfIncidentsByMonthSea)
dfIncidentsByMonthCity$Rate <- round(dfIncidentsByMonthCity$Rate*100, 2)

op <- par(mar=c(5,5,5,5))
barplot(t(matrix(dfIncidentsByMonthCity$Rate,nc=2)),
        names.arg=MonthCode,
        ylab="Rate, %",
        ylim=c(30, 36),
        xpd = FALSE,
        beside = TRUE,
        col=c("aquamarine3","coral"),
        xlab="Month",
        main = "Crime Rate By Month, Summer 2014",
        las = 1)
legend("topleft", c("San Francisco","Seattle"), pch=15,
```

```
       col=c("aquamarine3","coral"),
       bty="n")
```



Discussion

We see that while crime rates trending upward in San Francisco throughout the summer, no trends are visible in Seattle.