80/20 Rule on Crime in Cities

Adam Dicken

Sunday, January 10, 2016

My hypothesis is that 80% of crimes are committed in 20% of the area of the city.

The data provided gives location and categorical data about crimes commited in San Fransisco and Seattle.

```
knitr::opts_chunk$set(echo=TRUE, warning=FALSE, message=FALSE)
sanfran <- read.csv("sanfrancisco_incidents_summer_2014.csv",  stringsAsFactors=FALSE)
seattle <- read.csv("seattle_incidents_summer_2014.csv",  stringsAsFactors=FALSE)
```

Both data sets contain a similair number of crimes committed with Seattle being slightly ahead with 32779 crimes vs 28993 for San Fransisco.

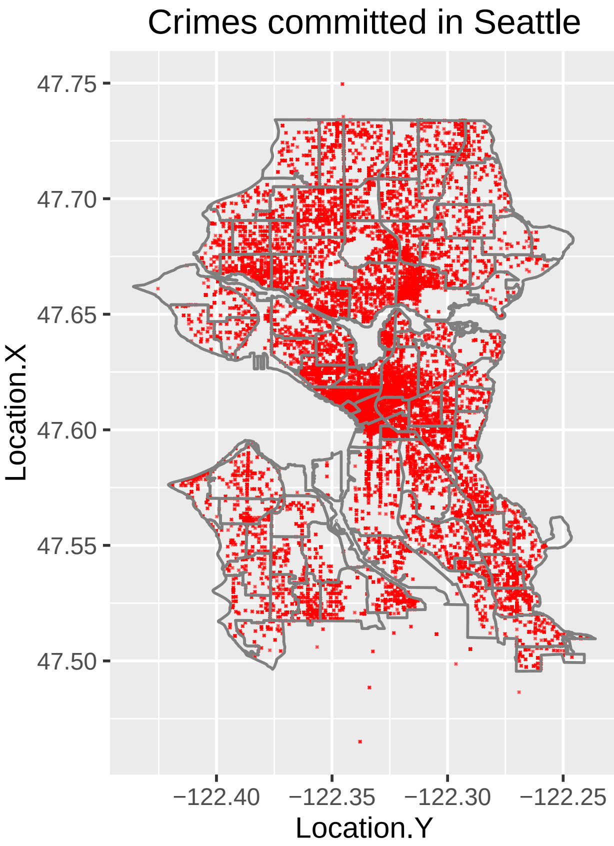
The crimes committed can be shown superimposed over maps of the two cities. The maps are provided by the city data portals as shapefiles, although the one for San Fransisco did not contain lattitude / longitude data so only the one for Seattle is shown. In order to get the location to a good level of accuracy the full (lattitude, longitude) is parsed from the Location attribute provided in both datasets. Note: 2050 rows in the Seattle dataset are missing location and they have been ommitted.

```
library(rgdal)
library(ggplot2)
library(RColorBrewer)
seattleMap <- fortify(readOGR(dsn="seattle_map/WGS84",layer="Neighborhoods"))
```

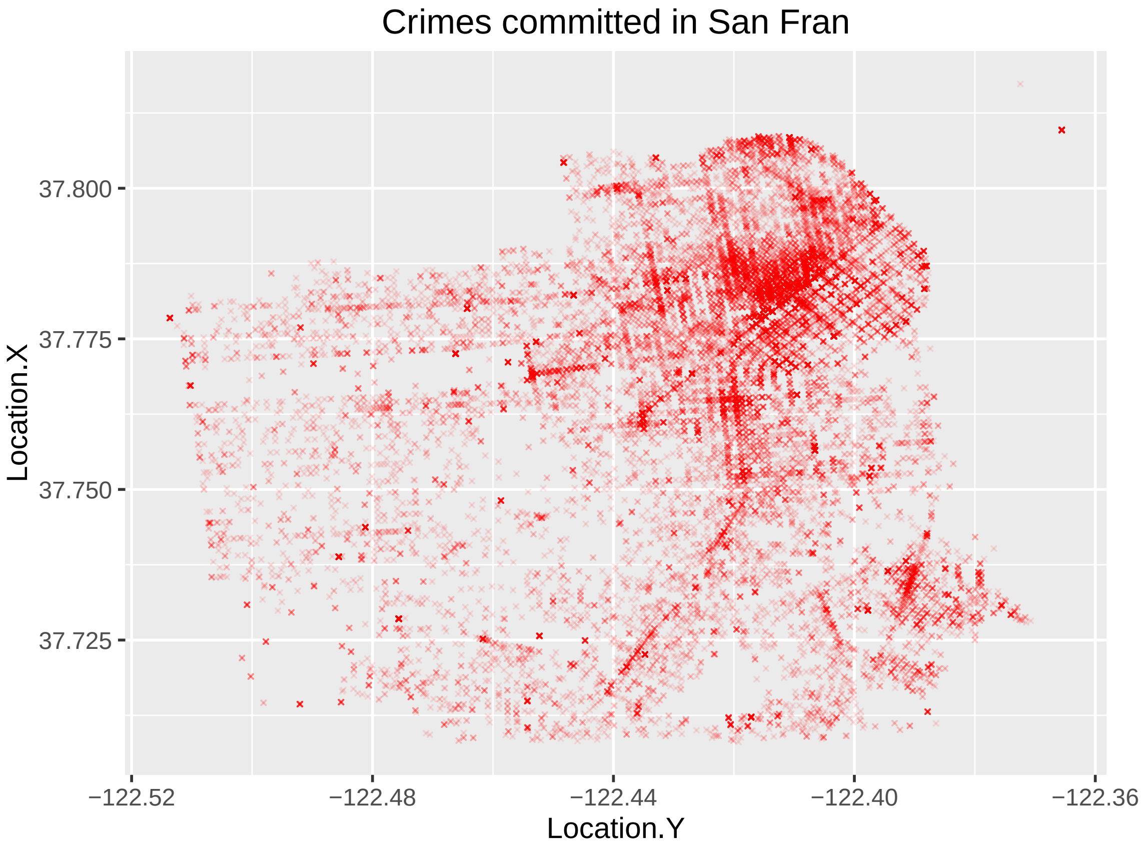
```
## OGR data source with driver: ESRI Shapefile
## Source: "seattle_map/WGS84", layer: "Neighborhoods"
## with 119 features
## It has 12 fields
```

```
# Get accurate lat / long and add as extra column
parseLocation <- function(df){
  for (row in 1:nrow(df)) {
    x <- df$Location[row]
    m <- regexec("^\\((.*), (.*)\\)", x)
    a <- regmatches(x, m)
    df$Location.X[row] <- as.numeric(a[[1]][2])
    df$Location.Y[row] <- as.numeric(a[[1]][3])
  }
  return(df)
}

seattle <- parseLocation(seattle)
sanfran <- parseLocation(sanfran)

# Remove rows with missing data (n.b. missing represented as 0)
seattle <- seattle[seattle$Location.X!=0,]
sanfran <- sanfran[sanfran$Location.Y!=0,]

# Plot data on map
```

```
ggplot(seattle, aes(x=Location.Y, y=Location.X)) +
  geom_point(shape=4, colour="red", size=0.01, alpha=0.5)+
  geom_path(data=seattleMap,aes(x=long, y=lat,group=group), colour="grey50")+
  coord_fixed() +
  ggtitle("Crimes committed in Seattle")
```

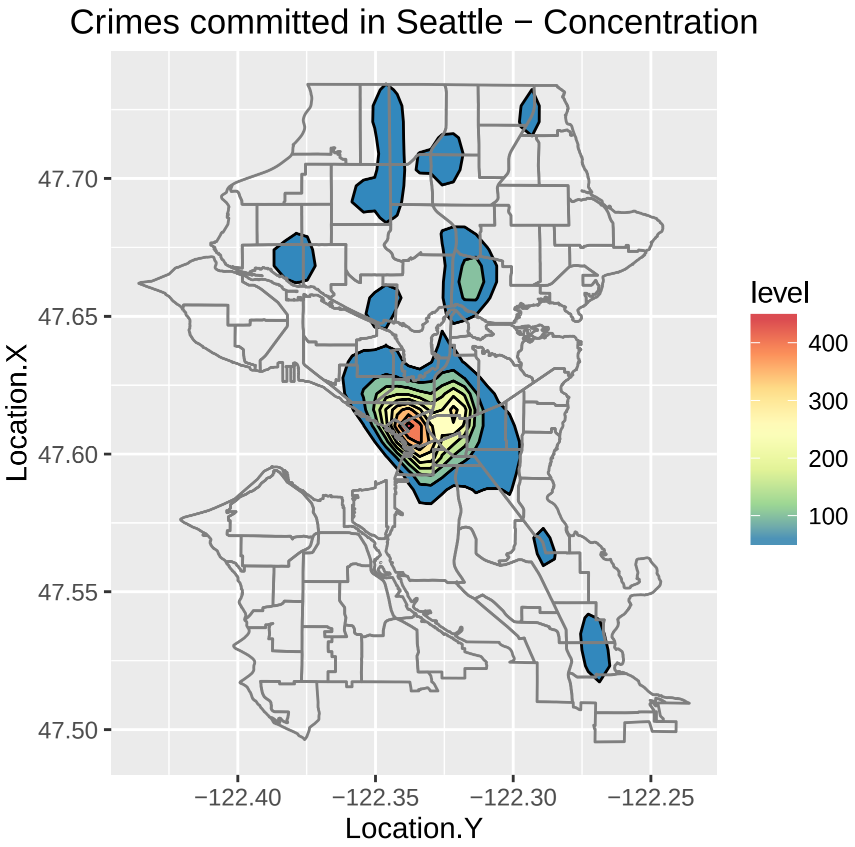


```
ggplot(sanfran, aes(x=Location.Y, y=Location.X)) +
  #stat_density2d(aes(fill = ..level..), alpha=0.5, geom="polygon")+
  geom_point(shape=4, colour="red", size=0.5, alpha=0.1)+
  coord_fixed() +
  ggtitle("Crimes committed in San Fran")
```

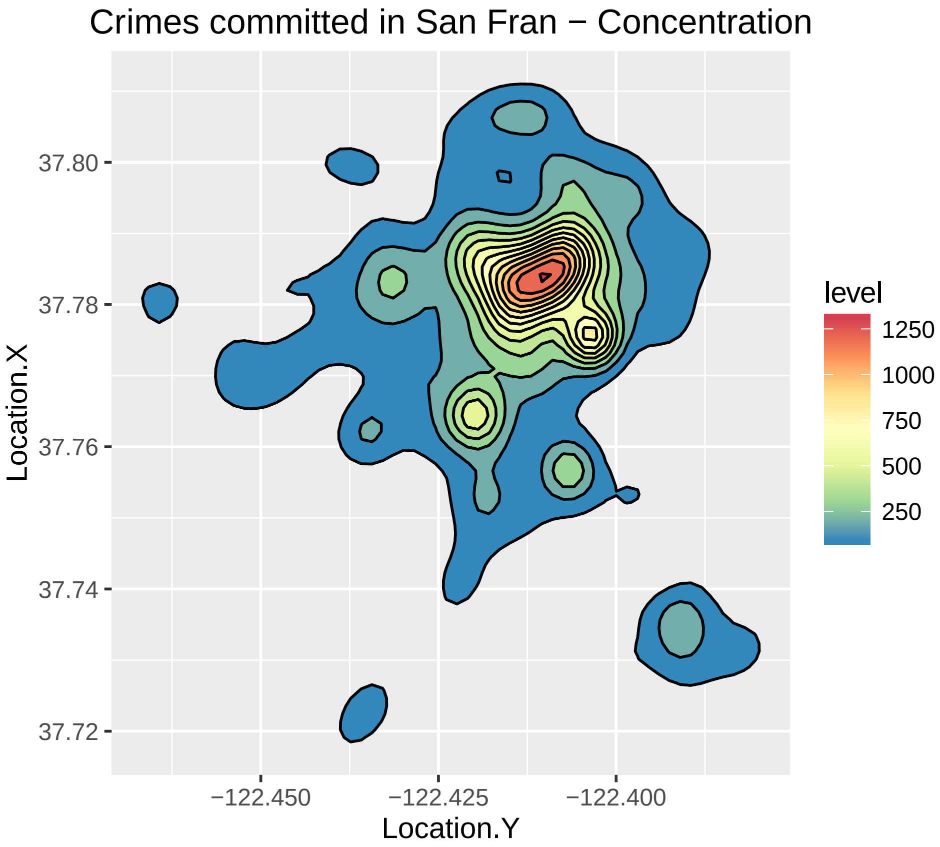


To begin to prove the hypothesis it is useful to plot a heatmap of the crimes committed to demonstrate the concentration in the cities.

```
ggplot(seattle, aes(x=Location.Y, y=Location.X)) +
  stat_density2d(aes(fill = ..level..), geom="polygon", colour="black", n=50)+
  geom_path(data=seattleMap,aes(x=long, y=lat,group=group), colour="grey50")+
  scale_fill_gradientn(colours=rev(brewer.pal(7,"Spectral")))+
  coord_fixed() +
  ggtitle("Crimes committed in Seattle - Concentration")
```



```
ggplot(sanfran, aes(x=Location.Y, y=Location.X)) +
  stat_density2d(aes(fill = ..level..), geom="polygon", colour="black")+
  scale_fill_gradientn(colours=rev(brewer.pal(7,"Spectral")))+
  coord_fixed() +
  ggtitle("Crimes committed in San Fran - Concentration")
```



As can be seen from the above plots the crimes are highly localized in both cities. We will now investigate in detail just how much crime is in the hotspot areas, we will focus our attention on Seattle. Specifically we will use the plot of Seattle above to isolate an area of high density. By using the color scheme and the key we can see a hotspot area of **density greater than 200** right in the middle so lets go ahead and zoom in on that.

```
# As shown on http://docs.ggplot2.org/0.9.3/stat_density2d.html the density plotted previously is calcu
library(MASS)

dens <- kde2d(seattle$Location.Y, seattle$Location.X, n = 50,)
densdf <- data.frame(expand.grid(Y=dens$x, X=dens$y),
 z = as.vector(dens$z))

# Find the point relating to the maximum density
max_density_points <- densdf[densdf$z>200,]
```

The above code snippet gives us a set of points where the highest density of crimes have been committed (the key shows us a density of greater than 200 being yellow to orange to red in colour). We can then define a bounding box using the extreme values from this selection.

```
right_Y <-  max(max_density_points$Y)
left_Y <- min(max_density_points$Y)
top_X <-  max(max_density_points$X)
bottom_X <- min(max_density_points$X)
```

Using the R library Geosphere then allows us to calculate the area of this box in square meters, and it is known that **Seattle has a total area of 369.2 million square meters**.

```r
library(geosphere)
max_density_box <- matrix(c(right_Y, left_Y, left_Y, right_Y,
                            top_X, top_X, bottom_X, bottom_X), nrow=4, ncol=2)
earthRadius <- 6378137
seattle_dense_area <- areaPolygon(max_density_box, a=earthRadius)
seattle_dense_area
```

```
## [1] 5168976
```

```r
seattle_total_area <- 369.2e6
perc_dense_area <- (seattle_dense_area / seattle_total_area) * 100
perc_dense_area
```

```
## [1] 1.400048
```

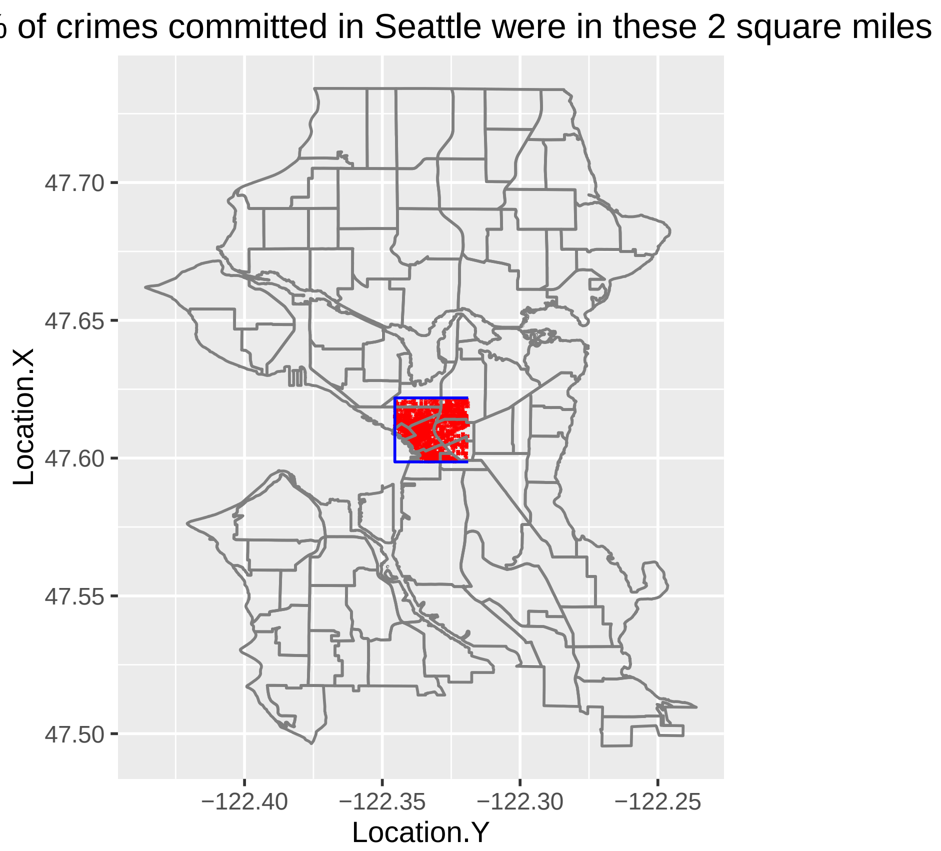
So our selected area of high density accounts for **1.4% of the area of Seattle**, ok so this isn't quite 20% of the area!

We can now determine the total number of crimes which lie inside this area.

```r
crimes <- seattle[seattle$Location.X>bottom_X & seattle$Location.X<top_X &
                    seattle$Location.Y>left_Y & seattle$Location.Y<right_Y,]

perc_crimes_in_dense <- (nrow(crimes)/nrow(seattle) ) * 100
```

So that is 21.4% of crime being committed within 1.4% of continuous area within the city. That is just one isolated location of around **2 sqaure miles** containing 21.4% of all crime!

```r
  ggplot(crimes, aes(x=Location.Y, y=Location.X)) +
    geom_point(shape=4, colour="red", size=0.01, alpha=0.5)+
    geom_path(data=seattleMap,aes(x=long, y=lat,group=group), colour="grey50")+
    coord_fixed() +
    ggtitle("21.4% of crimes committed in Seattle were in these 2 square miles") +
    geom_path(data=data.frame(max_density_box), aes(x=X1, y=X2), color="blue")
```


21.4% of crimes committed in Seattle were in these 2 square miles
47.70
47.65
47.60
47.55
47.50
Location.X
−122.40
−122.35
−122.30
−122.25
Location.Y