code

December 29, 2015

1 Comparing crimes between San Francicso and Seattle

1.1 Introduction

Two datasets are given for recorded crimes in San Francicso and Seattle during summer 2014. Daylight starts earlier and lasts a little longer in Seattle during this time, so there might be some differences. **This is a very quick and preliminary analysis!**

1.2 Preparing the data

Data are given as comma separated values and converted into data frames easily using the library “readr”.

```
In [7]: require(readr)
        trainSF <- read_csv("sanfrancisco_incidents_summer_2014.csv")
        trainSEA <- read_csv("seattle_incidents_summer_2014.csv")
```

1.2.1 Converting dates into hours of the day

Dates and times are given as character strings, so they have to be parsed first. For some incidents, both start and end time points are given, so a kind of mean is calculated.

```
In [8]: require(lubridate)
        trainSF$hour <- hour(hm(trainSF$Time))
        trainSEA$hourS <- hour(mdy_hms(trainSEA$"Occurred Date or Date Range Start"))
        trainSEA$hourE <- hour(mdy_hms(trainSEA$"Occurred Date Range End"))
        trainSEA$hour <- floor((trainSEA$hourS + pmax(trainSEA$hourE,trainSEA$hourS,na.rm=TRUE)) / 2)
```

1.2.2 Counting incidents by hour

Library “dplyr” is a functional approach to data handling upporting a pipeline syntax https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

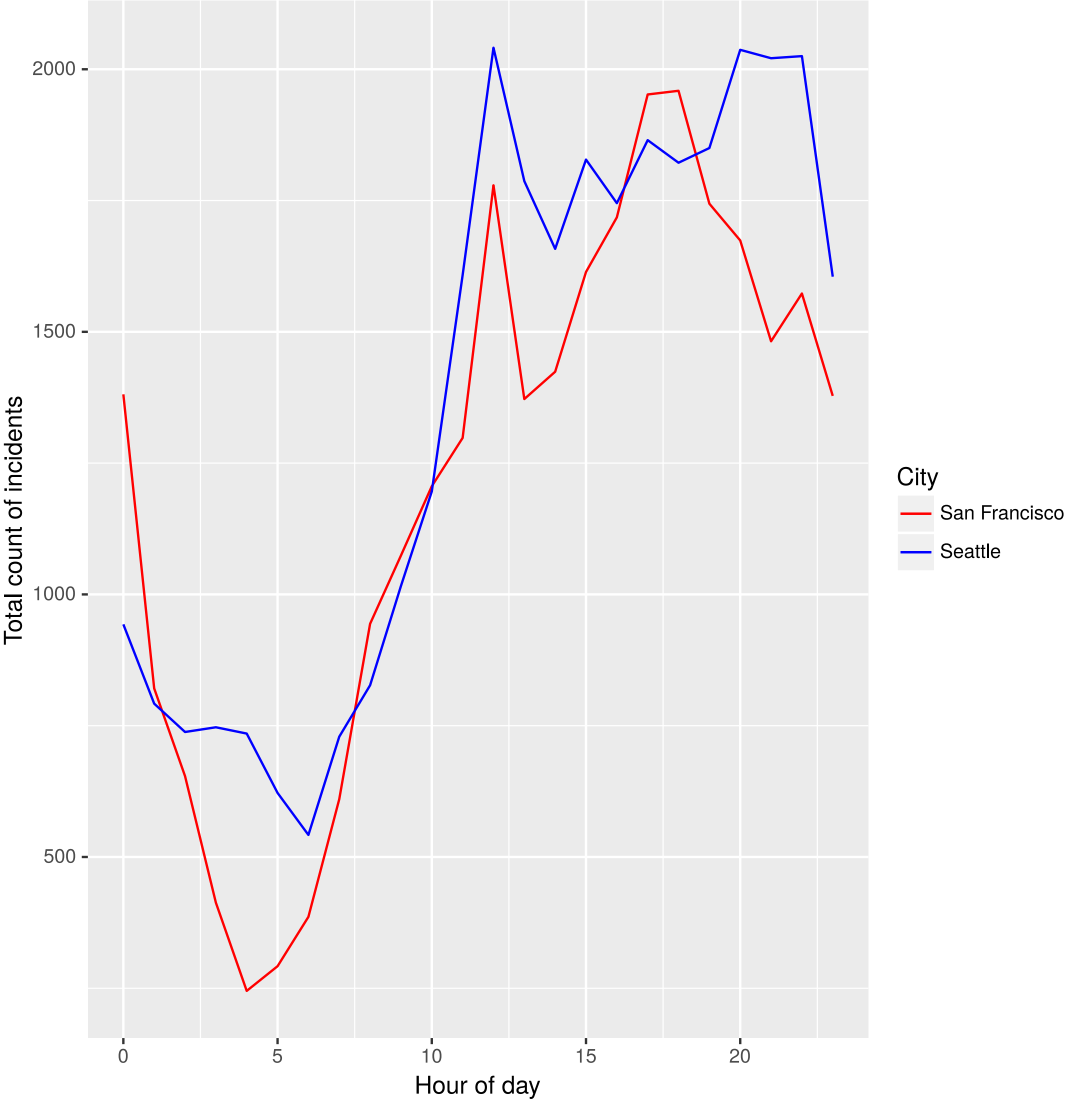
For distinguishing data origins in subsequent figures, columns with fixed city names are added.

```
In [9]: require(dplyr)
        byhourSF <- trainSF %>% group_by(hour) %>% summarise(total = n())
        byhourSEA <- trainSEA %>% group_by(hour) %>% summarise(total = n())
        byhourSF$city="San Francisco"
        byhourSEA$city="Seattle"
```

1.3 Plotting incidents against hours of day

Library “ggplot2” is well documented (http://docs.ggplot2.org/current/) and allows for incremental updates of plots.

```
In [10]: require(ggplot2)
         p <- ggplot(byhourSF, aes(x = hour, y = total, color=city))
         p <- p + geom_line(data=byhourSF)
         p <- p + geom_line(data=byhourSEA)
         p <- p + xlim(0, 23) + scale_colour_manual(values = c("red","blue"))
         p <- p + xlab("Hour of day") + ylab("Total count of incidents") + labs(colour = "City")
         # ggsave("hour.png", p, width=14, height=10, units="in")
         p
```



1.4 Conclusions 1

Regarding the absolute counts of all incidents, counts are comparable in both cities and there is a clear peak during midday in both cities. Data further suggest, that crimes in Seattle occur longer after sunset and still increase when crime rate already drops down in San Francicso. At early morning before sunrise more incidents happen in Seattle.

This is a very quick analysis. Geographic locations (longitudes within time zones) should be taken into account, and relative counts might give an more exact comparison.

1.5 Map incidents on geographic map

Data are taken from the “Open Street Map” project.

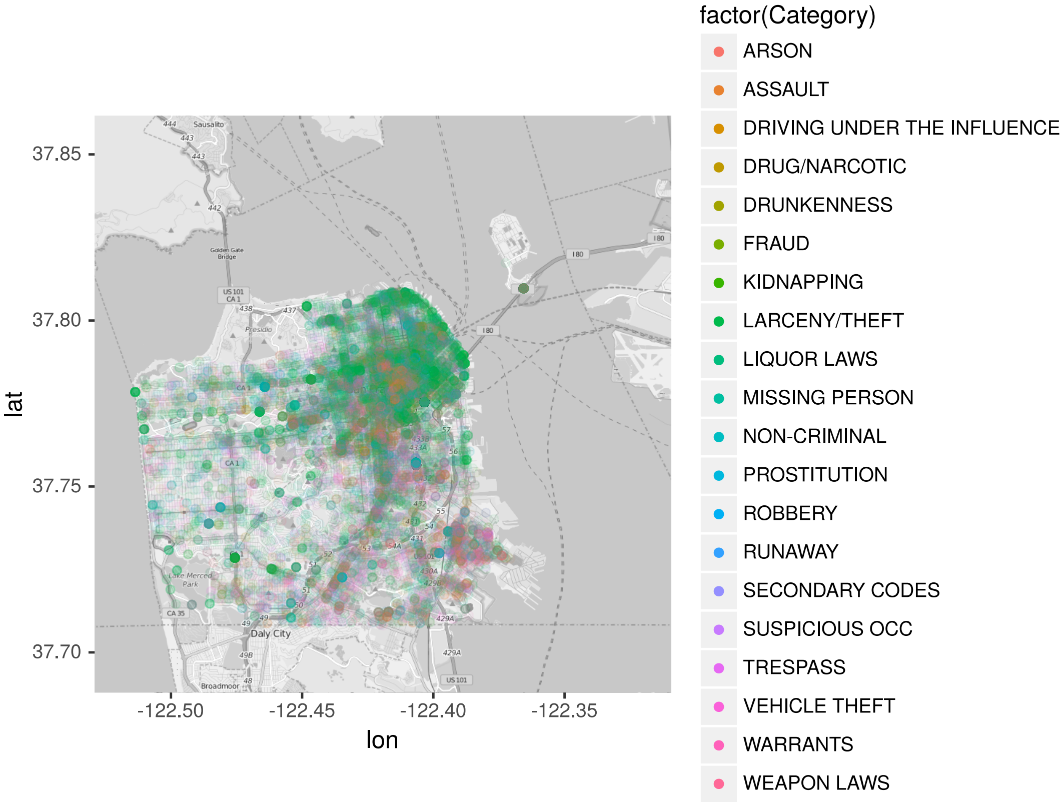
```
In [11]: require(ggmap)
         #mapSF<-get_map(location="sanfrancisco",zoom=12,source="osm",maptype = "roadmap", color = "bw"
         #mapSEA<-get_map(location="seattle",zoom=12,source="osm",maptype = "roadmap", color = "bw")
         #saveRDS(mapSF,"openstreetmapSF.rds")
         #saveRDS(mapSEA,"openstreetmapSEA.rds")
         mapSF <- readRDS("openstreetmapSF.rds")
         mapSEA <- readRDS("openstreetmapSEA.rds")
```

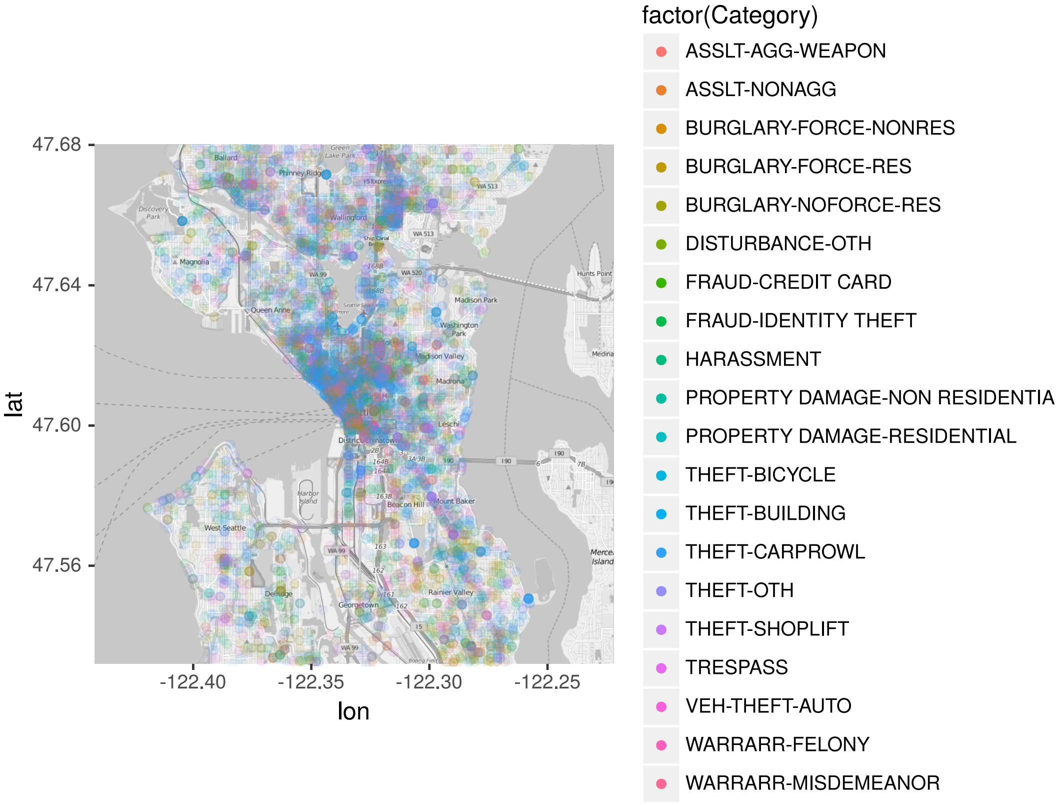
Some ideas taken from
https://www.kaggle.com/benhamner/sf-crime/san-francisco-top-crimes-map/files
More specifically, some categories have to be removed before further analysis.

```
In [12]: trainSEA$Category <- trainSEA$"Offense Type"
         countsSF <- trainSF %>% group_by(Category) %>% summarise(Counts=length(Category))
         countsSF <- countsSF[order(-countsSF$Counts),]
         countsSEA <- trainSEA %>% group_by(Category) %>% summarise(Counts=length(Category))
         countsSEA <- countsSEA[order(-countsSEA$Counts),]
         # This removes the "Other Offenses" category
         topSF <- trainSF[trainSF$Category %in% countsSF$Category[c(1,3:21)],]
         # This removes the "PROPERTY FOUND" category
         topSEA <- trainSEA[trainSEA$Category %in% countsSEA$Category[c(1:8,10:21)],]
```

```
In [14]: pSF <- ggmap(mapSF) +
              geom_point(data=topSF, aes(x=X, y=Y, color=factor(Category)), alpha=0.05) +
              guides(colour = guide_legend(override.aes = list(alpha=1.0)))
         # ggsave("mapSF.png", p, width=14, height=10, units="in")
         pSEA <- ggmap(mapSEA) +
              geom_point(data=topSEA, aes(x=Longitude, y=Latitude, color=factor(Category)), alpha=0.05)
              guides(colour = guide_legend(override.aes = list(alpha=1.0)))
         # ggsave("mapSEA.png", p, width=14, height=10, units="in")
         pSF
         pSEA
```

```
Warning message:
: Removed 7321 rows containing missing values (geom_point).
```


factor(Category)
ARSON
ASSAULT
DRIVING UNDER THE INFLUENCE
DRUG/NARCOTIC
DRUNKENNESS
FRAUD
KIDNAPPING
LARCENY/THEFT
LIQUOR LAWS
MISSING PERSON
NON-CRIMINAL
PROSTITUTION
ROBBERY
RUNAWAY
SECONDARY CODES
SUSPICIOUS OCC
TRESPASS
VEHICLE THEFT
WARRANTS
WEAPON LAWS
lat
37.85
37.80
37.75
37.70
lon
-122.50
-122.45
-122.40
-122.35
Sausalito
Golden Gate Bridge
Presidio
Lake Merced Park
Daly City
Broadmoor



1.6 Conclusions 2

In both cities, most crimes occur in the center of the city, here streets have highest density. However, this is also just a very quick analysis. Differnt types of categories have to be grouped and mapped differently. ggmap provides a suitable function for this "density2d".