Big Data Workshop: dplyr and San Fran Crime

Miles McBain

2 August 2016

Table of Contents

# Intro

In this practical we're going to use geolocated data to motivate data analysis using the dplyr package. dplyr is the latest iteration of the [Split-Apply-Combine Strategy for Data Analysis](https://www.jstatsoft.org/article/view/v040i01) in R. While this strategy is not new, dplyr is an important innovation since it allows the strategy to be applied to big data through modular backends. With dplyr one can use the same code to prepare data in R dataframes, [SQL databases](https://cran.r-project.org/web/packages/dplyr/vignettes/databases.html), and [big data platforms (Spark)](http://spark.rstudio.com/dplyr.html).

## The data

The data are around 30 000 crimes from the summer of 2014 in the city of San Francisco. While not inconveniently large, it is plain to see that this type of data that could become so, when collected over a longer time or a larger area.

This is public data that is not from a designed experiment. It needs a bit of work to do some analysis on.

## Questions

1. Can we identify the worst combinations of district, day, and time for crime? Where would you avoid?
2. Can we show visually where crime hotspots are?

# Learning Objectives

The exercise aims to introduce you to R packages for shaping, summarising and presenting data.

# Requirements

To complete this exercise you will need a computer with R Studio installed and the following packages:

* dplyr
* ggplot2
* readr
* leaflet

# Instructions

**Note** The R code in this prac may look different to what you have seen before. If so, Great! You're going to learn something. Feel free to ask for an explanation of **ANYTHING** no matter how trivial it may seem in the practical session. *R for Data Science* by Hadley Wickham is a very useful resource for the concepts covered here. See [data transformation with dplyr](http://r4ds.had.co.nz/transform.html).

## Setting up

1. Install any missing R packages in RStudio: install.packages("dplyr",readr","leaflet","tidyr")
2. Change the and comment the setwd() command below to the folder where you downloaded the workshop datasets.
3. Consider the discussion points in each question and if necessary, write R code to resolve them.

#setwd("~/") #Will need to set this to the right path.

## Load Data

Let's load the data and have a look at what we're dealing with:

library(readr)   
sanfran\_data <- read\_csv("./datasets/sanfrancisco\_incidents\_summer\_2014.csv")

## Parsed with column specification:  
## cols(  
## IncidntNum = col\_integer(),  
## Category = col\_character(),  
## Descript = col\_character(),  
## DayOfWeek = col\_character(),  
## Date = col\_character(),  
## Time = col\_time(format = ""),  
## PdDistrict = col\_character(),  
## Resolution = col\_character(),  
## Address = col\_character(),  
## X = col\_double(),  
## Y = col\_double(),  
## Location = col\_character(),  
## PdId = col\_double()  
## )

head(sanfran\_data)

## # A tibble: 6 × 13  
## IncidntNum Category Descript DayOfWeek  
## <int> <chr> <chr> <chr>  
## 1 140734311 ARSON ARSON OF A VEHICLE Sunday  
## 2 140736317 NON-CRIMINAL LOST PROPERTY Sunday  
## 3 146177923 LARCENY/THEFT GRAND THEFT FROM LOCKED AUTO Sunday  
## 4 146177531 LARCENY/THEFT GRAND THEFT FROM LOCKED AUTO Sunday  
## 5 140734220 NON-CRIMINAL FOUND PROPERTY Sunday  
## 6 140734349 DRUG/NARCOTIC POSSESSION OF MARIJUANA Sunday  
## # ... with 9 more variables: Date <chr>, Time <time>, PdDistrict <chr>,  
## # Resolution <chr>, Address <chr>, X <dbl>, Y <dbl>, Location <chr>,  
## # PdId <dbl>

# Question 1

## Filtering

The question is about crime, yet we noticed from previous Load Data step there are some NON-CRIMINAL records mixed in. We can filter those out using dplyr::filter.

sanfran\_data <-   
 sanfran\_data %>%  
 filter(Category != "NON-CRIMINAL")

* Look at the unique entries in sanfran\_data$Category. Are there other values you might want to filter out?

## Summarisation

We want to see if there are significant times or locations that crime peaks. To do this our data needs to be summarised according to these variables. It looks as though we already have a district vaible: pdDistrict and a day variable: DayOfWeek.

### Crimes by Day of Week and District

Below dplyr::group\_by() and dplyr::summarise() work in tandem to produce the crime summary by day of week and district. summarise() always needs to be called on a grouped data frame. Use View(sanfran\_data\_day) to see the complete output.

sanfran\_data\_day <-   
 sanfran\_data %>%  
 group\_by(PdDistrict, DayOfWeek) %>%  
 summarise(n\_crimes = n())  
head(sanfran\_data\_day)

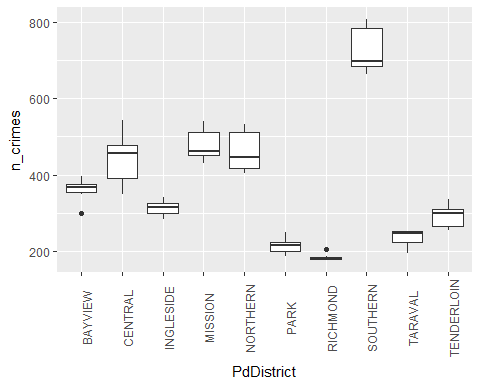
## Source: local data frame [6 x 3]  
## Groups: PdDistrict [1]  
##   
## PdDistrict DayOfWeek n\_crimes  
## <chr> <chr> <int>  
## 1 BAYVIEW Friday 395  
## 2 BAYVIEW Monday 350  
## 3 BAYVIEW Saturday 375  
## 4 BAYVIEW Sunday 299  
## 5 BAYVIEW Thursday 359  
## 6 BAYVIEW Tuesday 377

* What kind of data is this?
* What does the function n() do?
* **Expert:** Can you create a data frame that summarises the crimes by Hour of Day, Day of Week and District?
  + Check out dplyr::mutate() for starters.

## Exploratory Analysis

To visualise the relationship between day, district and number of crimes a boxplot would be suitable, for example:

ggplot(data = sanfran\_data\_day, aes(x=PdDistrict, y=n\_crimes)) +  
 geom\_boxplot() +  
 theme(axis.text.x = element\_text(angle = 90))



* What does the relationship for day look like?
* Try visualising the relationsip between both simultaneously using geom\_tile. E.g. [like this](https://learnr.wordpress.com/2010/01/26/ggplot2-quick-heatmap-plotting/).

## Linear Modelling

The answer to Question 1 is probably clear by now. Depending on your audience it might be sufficent to tell the story with visuals alone. What if we need to determine if the effects are statistically significant? Linear models are a standard statistical tool for this. How can we apply them in a way that will work for big data?

The biglm package in R allows one to fit linear models to big data in much the same way as the traditional lm or glm.

What kind of model might you fit in this case to explain n\_crimes? Is there extra information you might need? Where would you get it if so?

Here's an example model fit:

library(biglm)  
lin\_model <- biglm(data = sanfran\_data\_day, formula = n\_crimes ~ DayOfWeek + PdDistrict)  
summary(lin\_model)

## Large data regression model: biglm(data = sanfran\_data\_day, formula = n\_crimes ~ DayOfWeek +   
## PdDistrict)  
## Sample size = 70   
## Coef (95% CI) SE p  
## (Intercept) 387.5857 350.7110 424.4605 18.4374 0.0000  
## DayOfWeekMonday -41.8000 -76.2932 -7.3068 17.2466 0.0154  
## DayOfWeekSaturday -11.9000 -46.3932 22.5932 17.2466 0.4902  
## DayOfWeekSunday -21.0000 -55.4932 13.4932 17.2466 0.2234  
## DayOfWeekThursday -40.8000 -75.2932 -6.3068 17.2466 0.0180  
## DayOfWeekTuesday -45.7000 -80.1932 -11.2068 17.2466 0.0081  
## DayOfWeekWednesday -29.9000 -64.3932 4.5932 17.2466 0.0830  
## PdDistrictCENTRAL 80.0000 38.7728 121.2272 20.6136 0.0001  
## PdDistrictINGLESIDE -47.5714 -88.7987 -6.3442 20.6136 0.0210  
## PdDistrictMISSION 119.0000 77.7728 160.2272 20.6136 0.0000  
## PdDistrictNORTHERN 102.8571 61.6299 144.0844 20.6136 0.0000  
## PdDistrictPARK -146.0000 -187.2272 -104.7728 20.6136 0.0000  
## PdDistrictRICHMOND -175.1429 -216.3701 -133.9156 20.6136 0.0000  
## PdDistrictSOUTHERN 368.2857 327.0585 409.5129 20.6136 0.0000  
## PdDistrictTARAVAL -125.5714 -166.7987 -84.3442 20.6136 0.0000  
## PdDistrictTENDERLOIN -68.7143 -109.9415 -27.4871 20.6136 0.0009

# Question 2

Visualising spatial information on maps can be a powerful way to explore patterns. Historically these types of plots have been labourious to produce and required expensive tools. Luckily, there are many R packages that can facilitate this kind of plot. We will do an example using the R package leaflet.

## Using Coordinates

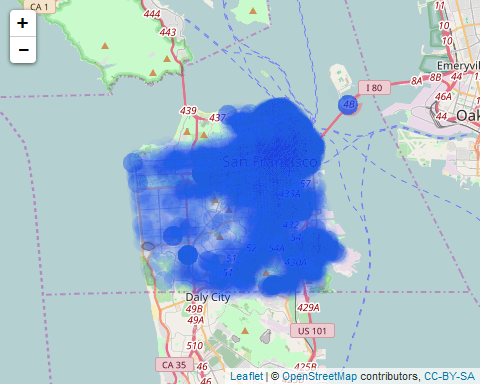
This San Fransisco Crime dataset has some coordinate variables, perfect for plotting on maps. There is the Location variable, which has helpfully been split out into X and Y. X is longitutde and Y is latitude.

* If you copy some Location data into Google maps does it confirm of interpretation of X and Y?

## Plotting with leaflet

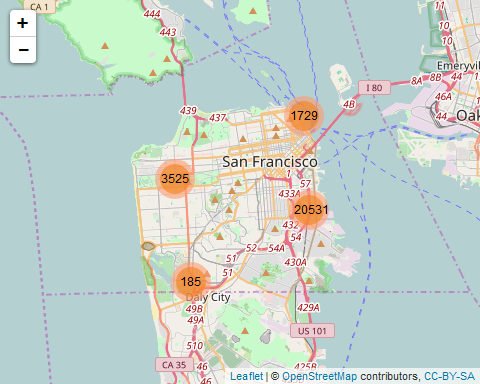
This uses markers with low opacity to generate a heatmap, type plot:

library(leaflet)  
  
sanfran\_map <-   
 leaflet(data = sanfran\_data) %>%  
 addTiles() %>%  
 addCircleMarkers(lng = sanfran\_data$X,   
 lat = sanfran\_data$Y,  
 stroke = FALSE,   
 fillOpacity = 0.02  
 )  
sanfran\_map



While this map automatically creates clusters of incidents:

sanfran\_map\_cluster <-   
 leaflet(data = sanfran\_data) %>%  
 addTiles() %>%  
 addCircleMarkers(lng = sanfran\_data$X,   
 lat = sanfran\_data$Y,  
 stroke = FALSE,   
 clusterOptions = markerClusterOptions()  
 )  
sanfran\_map\_cluster



* Can you spot any outliers or concerning observations using this plot?
* How would you address this?

### Customisation

leaflet is an R binding of a popular Javascript library by the same name. As such, it has many options for customisation. The [documentation](https://rstudio.github.io/leaflet/basemaps.html) is clear and full of examples. A few things you can try:

* Choosing alternate map tiles
* Creating a filter control for crime categories
* Colouring crime categories
* Adding crime details to the cluster plot.