Case Study 2: Police Data Challenge. Due Friday Nov 3 at 8:59pm (Declaration of intent MUST be submitted by Oct 20)

http://thisisstatistics.org/policedatachallenge/

Official rules of challenge:

http://thisisstatistics.org/wp-content/uploads/2017/09/Police-Data-Challenge-Rules.pdf

Even though the official, challenge allows teams of up to 5, our team limits are still 2-3.

Deadline is 8:59pm Friday, November 3. (Note that we are on Pacific time, so even though the official deadline is 11:59pm, it's in Eastern time, so that's 8:59pm to us)

Declaration of intent:
MUST BE DONE BY Friday, OCT 20

http://thisisstatistics.org/police-data-challenge-2017-declare-your-intent/

Team Sponsor

Michael “Jack” Davis (Lecturer, Simon Fraser University, Burnaby Canada)

Email Body – Declare Your Intent (suggested, you may change)

We are joining this challenge as part of case study assignment in the undergraduate class Stat 440 – Learning from Big Data, at Simon Fraser University.

Submission requirements:

“Teams will be required to submit presentation slides, up to 10 slides. The presentation should be developed as a presentation to the city officials of the selected city. “

“In addition to the stakeholder presentation, teams will be required to submit a document, not to exceed 500 words, describing the technical details of their work.”

So the presentation should include the results of your findings, and the technical document should include the methods, or information about the model. In other words, this is a lot like the report required for Case Study 1, except that the results half is a presentation instead of a technical report.

Case study goal:

“The goal of the presentation is to guide the city in ***making evidence-based decisions on how to improve public safety***. “

Possible solutions:

- Find a way to predict the location of crime in one week based on information from the previous weeks.

- Can you build a method to suggest what times of day, week, or year that a greater police presence would be useful?

- Find the long term trends types, locations, and/or victims of crime. Are certain parts of town getting worse? Better? Are certain crimes becoming more popular?

- Can you predict the priority (non-emergency, low, medium, high) of the call from the other variables like description, district, or incident location? Are there false alarms you can detect (especially 911/no voice)?

- Are there additional datasets, such as weather reports, that could be included any of the above model fits?

Grading:

http://thisisstatistics.org/wp-content/uploads/2017/09/PoliceDataChallengeJudgingRubric.pdf

This is their rubric, mine will be similar, but more focussed on the methods and less on the 'story'.

On the awards:

“Awards will be given in the categories of Best Overall Analysis, Best Visualization, and Best Use of External Data.”

Any team that wins an award will get 100% on the case study. (They will still get critiques and feedback)

Datasets:

Main website
https://www.policedatainitiative.org/datasets/

Data guide:
http://thisisstatistics.org/wp-content/uploads/2017/09/Calls-for-Service-Data-101.pdf

Baltimore
https://data.baltimorecity.gov/Public-Safety/Calls-For-Service-Data-Lense/t3vg-dqh8

Cincinnati
https://data.cincinnati-oh.gov/Safer-Streets/Police-Calls-For-Service-with-Time-Dispatched-Tran/4try-zhpp

Seattle
https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp

Immediate data problems:

VOLUME. The Baltimore data alone has 2.8 MILLION rows. Excel can't open the CSV file to look at these?

What can you do?
1. Download and subset
2. Query the central server

From this post: http://www.devcurry.com/2010/08/free-text-editors-to-open-large-text.html
Try Vim (on Macs) or Notepad++ (on Windows) to be able to see the data.

R may be able to load it if you have a lot of RAM, but it may not be able to do much with it. You could load a portion of the data at a time, run the same analysis on each portion, and then combine the analyses later. This is essentially what is done in the 'mapreduce' algorithm of Hadoop.

```
## Baltimore example.
dat = read.csv("911_Calls_for_Service.csv") ## Takes forever, if it opens at all
dim(dat)  # 2892458 rows, 7 columns

## Possibility 1: Take a sample of 5% of the rows
fraction = 0.05
set.seed(12345) ## Change this to your own number
sample_rows = sample(1:nrow(dat), floor(nrow(dat) * fraction))

dat_sample = dat[sample_rows,]
write.csv(dat_sample,"Baltimore_Fraction_05.csv")
```

```
dat = dat_sample # overwrite the old dat
dat_sample = NULL # To clear RAM
gc() # garbage collection to further clear RAM
```

VARIETY.

Many of the fields are in text.

Some of them are addresses. Can you convert these to some other geographic location?
Try this: http://www.gpsvisualizer.com/geocoder/
Try this: https://cran.r-project.org/web/packages/geomapdata/index.html 'geomapdata' package in R.

Some of these fields are timestamps. Can you convert these into a time variable such as 'number of days after Jan 1, 2000'? This would be useful if you wanted to measure the time between events.
Try this: https://cran.r-project.org/web/packages/datetime/datetime.pdf
'stringi' package in R.

Between the three datasets, different variables are available. How do you compare 'alarm level' in Cincinnati to the 'priority' used in Baltimore.

```
## Cincinnati example for timestamps.
dat = read.csv("Cin_First_Few.csv") ## Just a sample. The first 120 rows or so.

install.packages("stringi")
library(stringi)

head(dat$ARRIVAL_TIME_PRIMARY_UNIT)
# CREATE_TIME_INCIDENT
# ARRIVAL_TIME_PRIMARY_UNIT
# CLOSED_TIME_INCIDENT
# DISPATCH_TIME_PRIMARY_UNIT
# TRANSMIT_TIME_PRIMARY_UNIT

times = stri_datetime_parse(dat$ARRIVAL_TIME_PRIMARY_UNIT, format = "MM/dd/uu
HH:mm:ss a")

?stri_datetime_parse  ## How do we do PHONE_PICKUP_TIME and
INCIDENT_TIME_OF_OCCURANCE


### Finding the difference between times
difftime(times[13], times[1])
times[13] - times[1]
as.numeric(difftime(times[13], times[1]))
as.numeric(difftime(times[13], times[1], units="secs"))
```