Real-world Data Analysis in an Introductory Statistics Course: Assignments using Data from the Stanford Open Policing Project

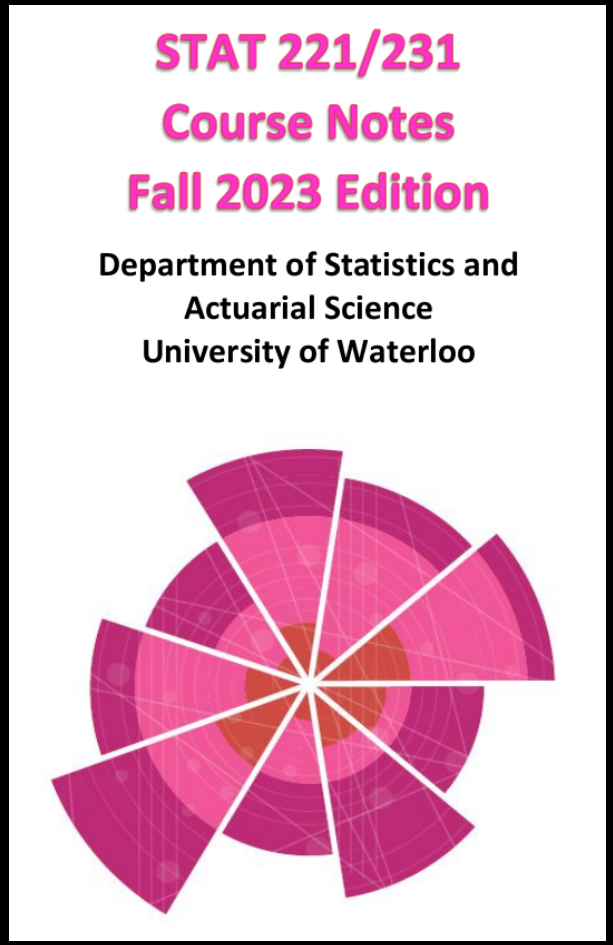
Michael Wallace
michael.wallace@uwaterloo.ca

Assignment dataset downloader (more on this later!):
`shiny.math.uwaterloo.ca/sas/stat231/stopdownloader/`

Slide deck and links available at:
`https://mpwallace.github.io/`

STAT 231: Overview

- Intro stats course (follows STAT 230)
- Enrolment: 400-700 (multi-section)
- Wide range of abilities/backgrounds
- Challenge: student engagement, especially "why is this useful?"

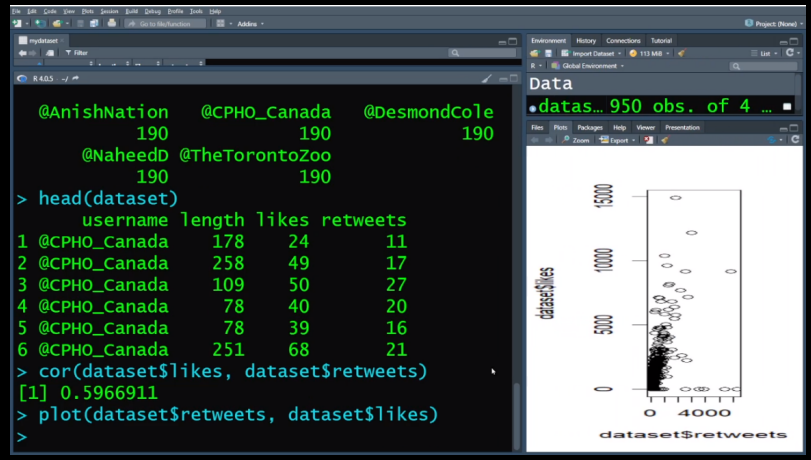


STAT 231: Topics

- Statistical terminology, some principles of study design/analysis
- Numerical/graphical data summaries
- Maximum likelihood estimation
- Likelihood intervals, confidence intervals
- Hypothesis testing (including two-sample t-tests and chi-squared tests)
- Linear regression

Introduce/expand R experience through assignments:

- Prior experience: little/none
- Relatively basic commands/analyses
- Not a coding/programming course!
- Focus: interpretation of results

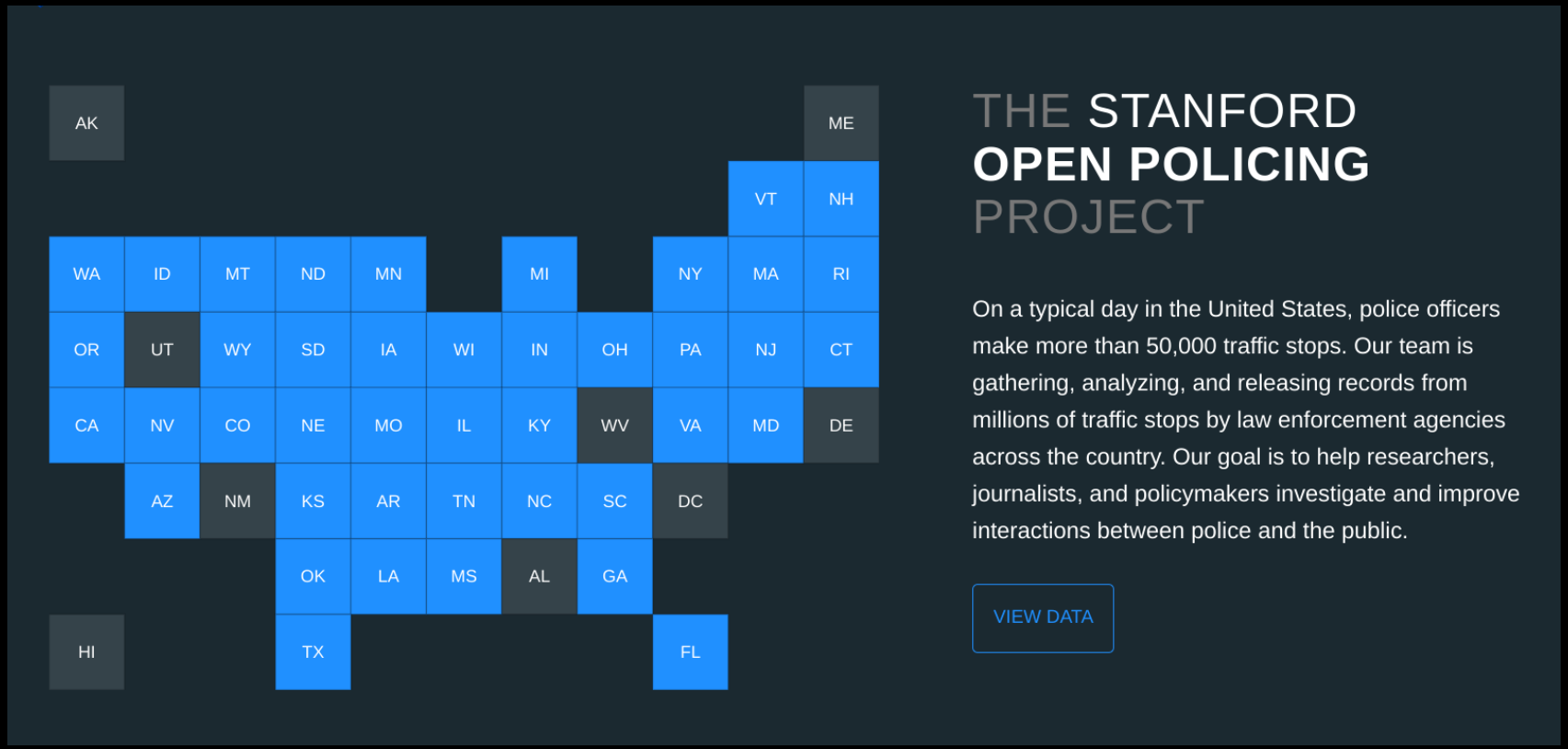


R material primarily covered through tutorial videos

Assignments: Goals

- Realistic data
- Realistic analysis questions
- Broad range of analyses
- Unique datasets per student

Stanford Open Policing Project



https://openpolicing.stanford.edu/

Stanford Open Policing Project

State	Download	Stops	Time Range	Stop Date	Stop Time	Stop Location	Driver Race	Driver Sex	Driver Age	Search Conducted	Contraband Found	Citation Issued	Warning Issued	Frisk Performed	Arrest Made	Reason for Stop	Violation
AR																	
Little Rock	R	13,641	Dec 2016 - Nov 2017	■	■		■	■	■		■						
AZ																	
Gilbert	R	480,599 [2]	Dec 2007 - May 2018	■	■												
Mesa	R	157,015 [2]	Dec 2013 - Sep 2019	■	■		■	■	■			■	■		■		■
State Patrol	R	3,494,153 [2]	Jan 2009 - Dec 2017	■	■	■	■	■		■	■	■	■		■		
CA																	
Anaheim	R	87,876	Dec 2011 - Mar 2017	■											■		
Bakersfield	R	189,685	Mar 2008 - Mar 2018	■	■	■	■	■	■		■						
Long Beach	R	365,924 [2]	Dec 2007 - Dec 2017	■			■	■	■		■						■
Los Angeles	R	5,418,400 [2]	Dec 2009 - Jun 2018	■	■	■	■	■									
Oakland	R	133,405 [2]	Mar 2013 - Dec 2017	■	■		■	■		■	■	■	■		■	■	

https://openpolicing.stanford.edu/data/

Note: Available data varies by location.

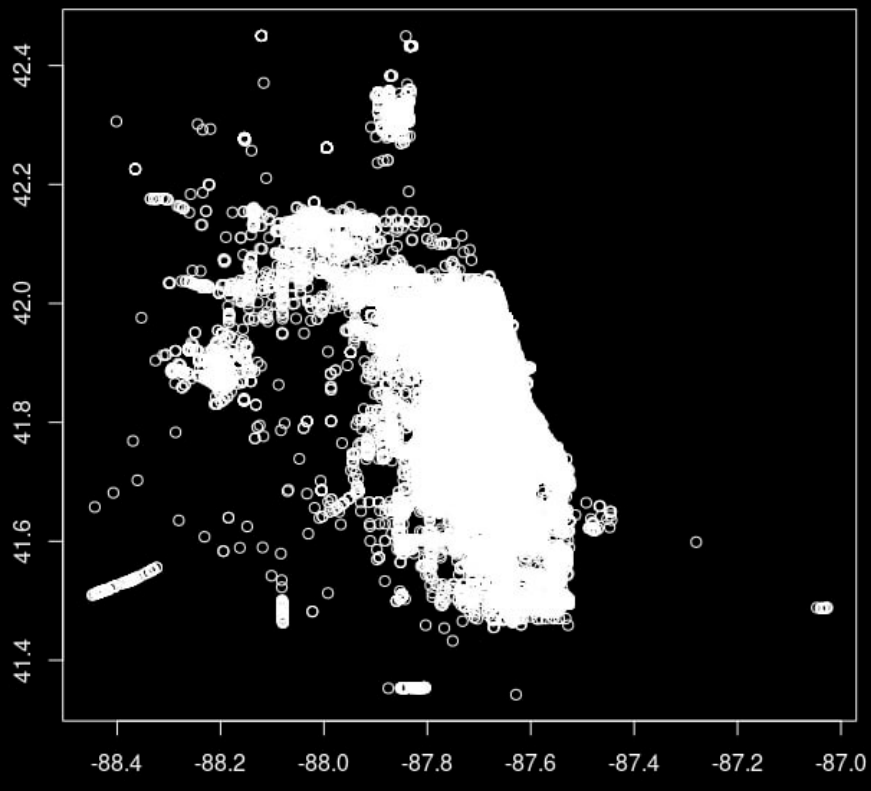
Variates

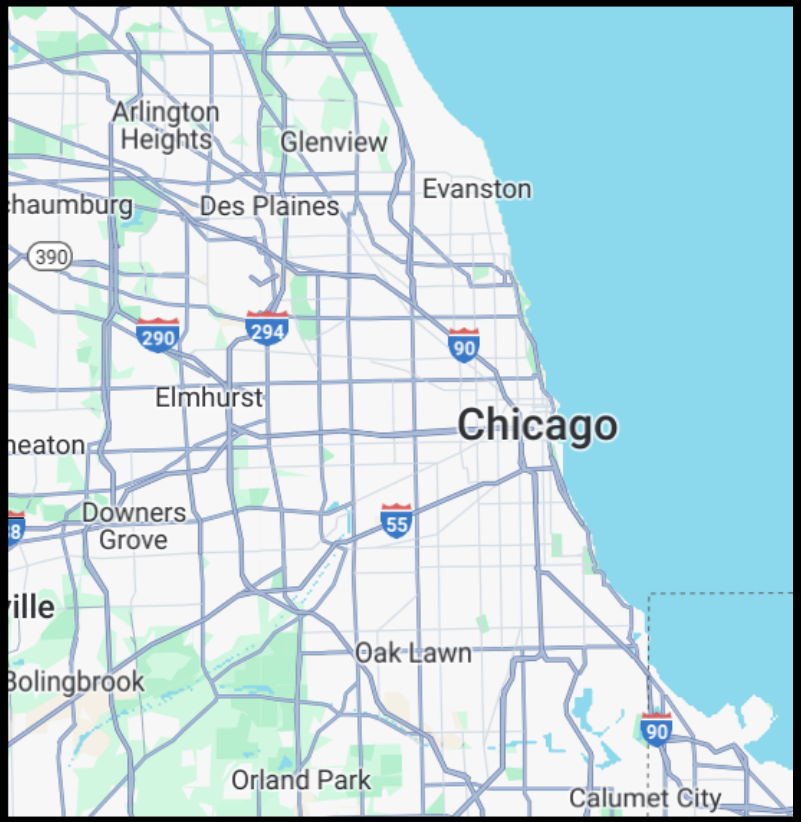
I chose Chicago and San Francisco and the following variates:

Stop data	Time, Date, Location (lat/long)
Subject demographics	Age, Race, Sex, Vehicle make
Offence data	Violation, Outcome

Pre-processing steps:

- Time: Converted to hours after midnight
- Date: Converted to day of week
- Vehicle make: Collapsed to Toyota, Ford, Chevrolet, Honda, Other


42.4
42.2
42.0
41.8
41.6
41.4
-88.4
-88.2
-88.0
-87.8
-87.6
-87.4
-87.2
-87.0


Arlington Heights
Glenview
Evanston
Des Plaines
390
290
294
90
Elmhurst
Chicago
Downers Grove
55
Oak Lawn
Orland Park
Calumet City

Data Processing

Main data issues:

- Missing data:

  $\rightarrow$ Complete case analysis ($\approx$ 2 million traffic stops)

- Sex:

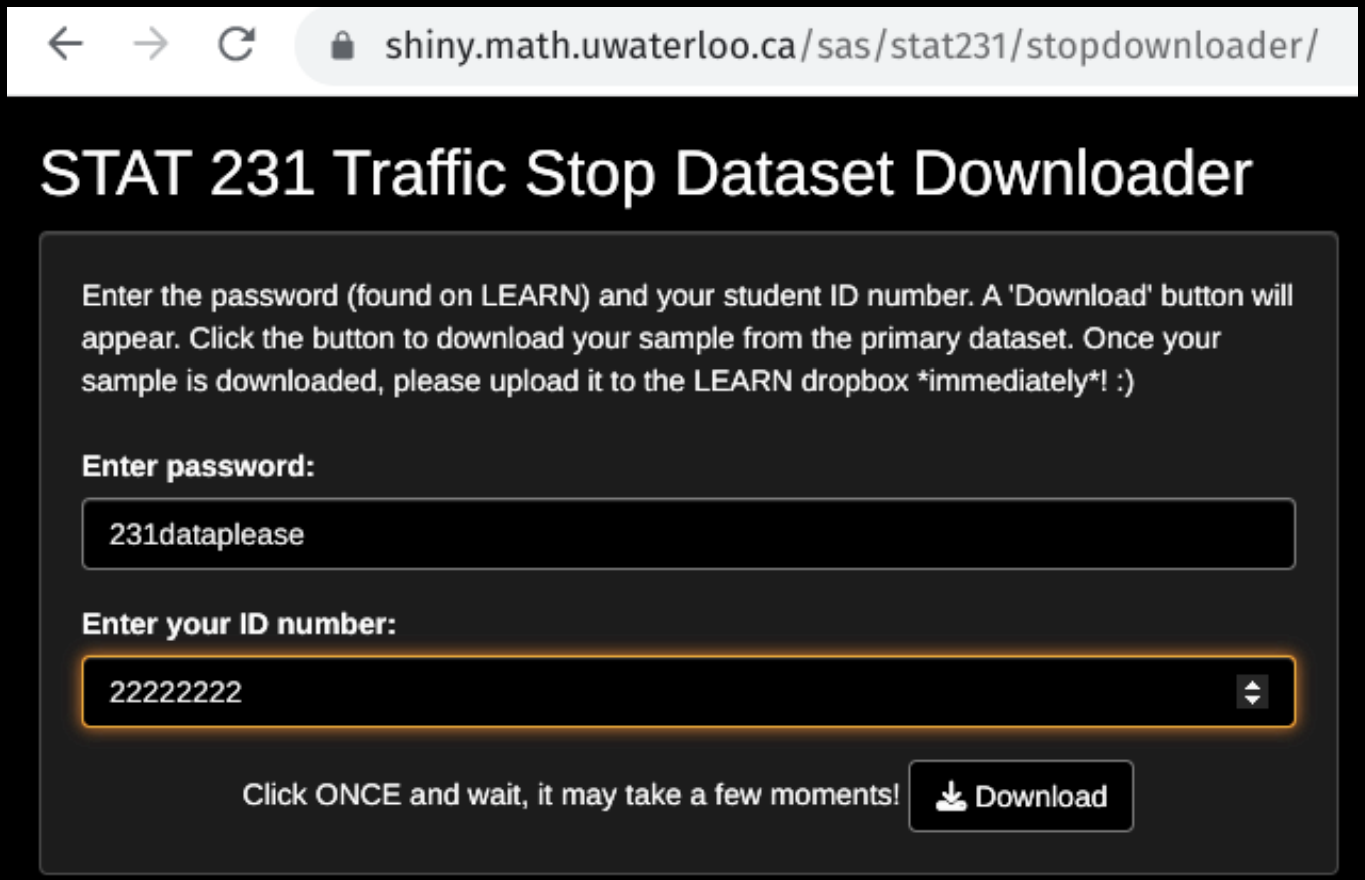
  $\rightarrow$ Coded as Female or Male only

- Race:

  $\rightarrow$ San Francisco: Asian/Pacific Islander, Black, Hispanic, White, Other

  $\rightarrow$ Chicago: Black, Hispanic, White

Generating Unique Datasets

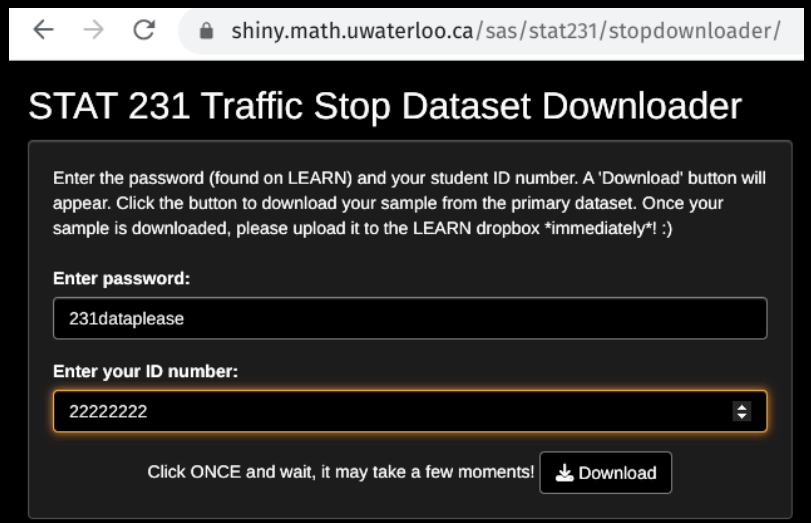


`https://shiny.math.uwaterloo.ca/sas/stat231/stopdownloader/`

Random sample of ≈ 800 entries, tied to student ID number.

Generating Unique Datasets

- Datasets are unique
- But not *too* unique!
- Tied to student ID
- Uploaded to course website for pre-checks



Assignment Structure

Assignment structure:

- 2-4 'Analyses', each focused on 1-2 variates
- Sub-parts of R analysis followed by interpretation, e.g.:

  (a) Generate a bar plot of Subject Race or Vehicle Make stratified by Subject Sex

  (b) Discuss any similarities/differences
- Students are given some choice (e.g., to analyze Race or Vehicle Make)
- Students submit a structured 'Report' and their R code

Grading and Feedback

- TAs verify numerical/graphical results using R code
- TAs review and give feedback on interpretation questions
- 'Debrief' document containing example analysis and highlighting key learning objectives also provided

Conclusions

The good:

- Good engagement/positive student feedback
- Opportunity to discuss broader issues/topics
- Relatively easy grading

The not-so-good:

- Non-trivial setup and pre-processing
- Some analyses a little contrived
- Extra care needed to avoid teaching ‘bad habits’

Links

- **Stanford Open Policing Project** `https://openpolicing.stanford.edu/`
- **Sample Downloader** `https://shiny.math.uwaterloo.ca/sas/stat231/stopdownloader/`

  Password: 231dataplease; ID number $\in [2 \times 10^7, 3 \times 10^7)$
- **My website**: `https://mpwallace.github.io/`
- **My email**: michael.wallace@uwaterloo.ca
- **My socials**: @statacake