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 221/231 Course Notes Fall 2023 Edition

Department of Statistics and Actuarial Science University of Waterloo



- 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)
- $\circ~$ Linear regression

Introduce/expand R experience through assignments:

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

in im im im im jun im im im i M · · · · · · · · · · · · · · · · · · ·	ala jan 198 1000 - 100 - 1000 - 1							0	Rept Direct 1
Enderer -					-	any Core	Des Taure		-0
				-					
					Data	_		_	
@AnishNation	@CPH0_Ca	anada	@DesmondCole		odata	is 91	0 obs.	of 4	
					The Part	****** ***	There is a second		
@NaheedD	@TheToron1								
						_			
> head(dataset)						33	- 0		
username	length lik		tweets						
1 @CPHO_Canada						_	0		
2 @CPHO_Canada							- 8.		
3 @CPHO_Canada					tates (ins		0.	-	
4 @CPHO_Canada					-8		820		
5 @CPHO_Canada							15		
6 @CPHO_Canada							÷		
> cor(dataset\$	likes, data	asetSr	etweets)				2.00		
[1] 0.5966911				-1		-	-	10.00	
> plot(dataset)	<pre>\$retweets.</pre>	datas	et\$likes)				0 40	000	
						da	taset\$	etwe	ots

R material primarily covered through tutorial videos

Assignments: Goals

- \circ Realistic data
- Realistic analysis questions
- Broad range of analyses
- $\circ~$ Unique datasets per student



Stanford Open Policing Project



THE STANFORD OPEN POLICING PROJECT

On a typical day in the United States, police officers make more than 50,000 traffic stops. Our team is gathering, analyzing, and releasing records from millions of traffic stops by law enforcement agencies across the country. Our goal is to help researchers, journalists, and policymakers investigate and improve interactions between police and the public.

https://openpolicing.stanford.edu/

Stanford Open Policing Project

					2º	The	PLOCAHO Dit	r asce	set	Se,	arch Cond	ucced .	ation Issued	ing Issued	erformed MP	de nor
State	Download	Stops	Time Range	St	op Date Str	op Time Sto	PLC ON	Wer Race	wer set Dri	Wer Age 56	arch co	ntrat cit	ation. War	Ins Filsk P	Arrest	Reason JI
AR																
Little Rock	P 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 R	133,405 ²	Mar 2013 - Dec 2017													

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

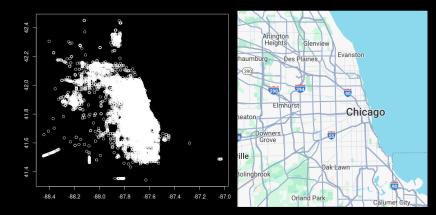
Note: Available data varies by location.

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
- $\circ\,$ Vehicle make: Collapsed to Toyota, Ford, Chevrolet, Honda, Other



Main data issues:

 $\circ~$ Missing data:

ightarrow Complete case analysis (pprox 2 million traffic stops)

 \circ 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

←	\rightarrow	C		shiny.math.uwaterloo.ca/sas/stat231/stopdownloader	/
---	---------------	---	--	--	---

STAT 231 Traffic Stop Dataset Downloader

Enter the password (found on LEARN) and your student ID number. A 'Download' button will appear. Click the button to download your sample from the primary dataset. Once your sample is downloaded, please upload it to the LEARN dropbox *immediately*! :)

Enter password:		
231dataplease		
Enter your ID number:		
2222222		¢
Click ONCE and wait, it may take a few moments!	L Download	

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

Random sample of \approx 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:

- $\circ~$ 2-4 'Analyses', each focused on 1-2 variates
- $\circ~$ 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)
- $\circ\,$ 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



The good:

- Good engagement/positive student feedback
- Opportunity to discuss broader issues/topics
- $\circ~\mbox{Relatively easy grading}$

The not-so-good:

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

- o Stanford Open Policing Project
 https://openpolicing.stanford.edu/
- Sample Downloader https://shiny.math. uwaterloo.ca/sas/stat231/stopdownloader/ Password: 231dataplease; ID number ∈ [2 × 10⁷, 3 × 10⁷)
- o My website: https://mpwallace.github.io/
- My email: michael.wallace@uwaterloo.ca
- My socials: @statacake