

# More than Methods

*Preparing students for data-driven work outside the classroom*

Applied Research & Education Seminar  
April 8, 2024

Maria Tackett  
Duke University

Image credit: Sketchepedia on Freepik



[bit.ly/ares-more-than-methods](https://bit.ly/ares-more-than-methods)

# Courses I teach



[Introduction to  
Data Science](#)



[Regression  
Analysis](#)



[Generalized  
Linear Models](#)

Background  
and  
Motivation

Three principles

Challenges  
and  
next steps

Background  
and  
Motivation

Three principles

Challenges  
and  
next steps

# 2014 ASA Undergraduate Curriculum Guidelines

“...concepts and approaches for working with **complex data**...and analyzing non-textbook data.”

“...students’ analyses should be undertaken in a **well-documented and reproducible way.**”

“...construct effective visual displays and **compelling written summaries**” and “demonstrate ability to **collaborate in team...**”

# 2017 Curriculum Guidelines for Undergraduate Programs in Data Science

“...work with data from a **variety of sources and formats...**”

“...data scientists must **communicate to teammates as well as to those with less intimate knowledge of the project particulars.**”

“...exposure to and **ethical training** in areas such as citation and data ownership, security and sensitivity of data, consequences and privacy concerns of data analysis, and the **professionalism of transparency and reproducibility.**”

# Undergraduate research team

- Interdisciplinary team co-led with Dr. Nichole Schramm-Sapyta in the Duke Institute for Brain Sciences
- 5 - 7 undergraduate students + 1 graduate project manager each year
- **Primary objective:** Understand patterns of healthcare utilization and interactions with the criminal justice system to provide data-informed insights to community stakeholders

# The data

## Bookings in County Detention Facility

<b>unique_id</b>	<b>race</b>	<b>ethnicity</b>	<b>sex</b>	<b>age_in_2020</b>	<b>confined_date</b>	<b>release_date</b>	<b>charges</b>	<b>release_reason</b>
100001	White	Nonhispanic	M	23	11/10/2019	11/12/2019	Theft	Sentence completed
100002	Black	Hispanic	F	20	2/15/2020	2/17/2020	Assault, Burglary	Transfer
100003	Asian	Non-Hispanic	M	28	7/22/2020	7/25/2020	Burglary, Drug possession, Assault	Time served
...	...	...	...	...	...	...	...	...
400001	W	H	F	50	8/12/2020	8/13/2020	Fraud	Secure bond

The tables were generated by ChatGPT for illustrative purposes only.

# The data

## Encounters with Health System

unique_id	race	ethnicity	sex	age_in_2020	encounter_date	chief_complaint	diagnoses
100001	White	Non-hispanic	M	23	9/20/2020	Headache	G44.0, R51
100002	Black	Non-Hispanic	F	20	10/5/2020	Cough	J44.9, R05, J20.9
...	...	...	...	...	...	...	...
400001	White	Hispanic	F	50	5/2/2020	Nausea	R11.0, R10.13, R11.9, R11.10, R11.12

The tables were generated by ChatGPT for illustrative purposes only.

# The data

Bookings

unique_id	race	ethnicity	sex	age_in_2020
100001	White	Nonhispanic	M	23
100002	Black	Hispanic	F	20
100003	Asian	Non-Hispanic	M	28
...	...	...	...	...
400001	W	H	F	50

Health System Encounters

unique_id	race	ethnicity	sex	age_in_2020
100001	White	Non-hispanic	M	23
100002	Black	Non-Hispanic	F	20
...	...	...	...	...
400001	White	Hispanic	F	50

The tables were generated by ChatGPT for illustrative purposes only.

# More than data analysis

- Turn community stakeholder questions into statistical inquiries
- Use a reproducible workflow, with clear and informative documentation
- Present results in reports and presentations to community stakeholders
- Take into account ethical considerations in every analysis decision
- Collaborate with teammates, project manager, and faculty

# Goal: Narrow the gap

Data analysis  
in the  
classroom



Data analysis in  
research,  
internships, and  
jobs

# Goal: Narrow the gap

Data analysis  
in the  
classroom



Data analysis in  
research,  
internships, and  
jobs

# Three principles for modernizing a regression course

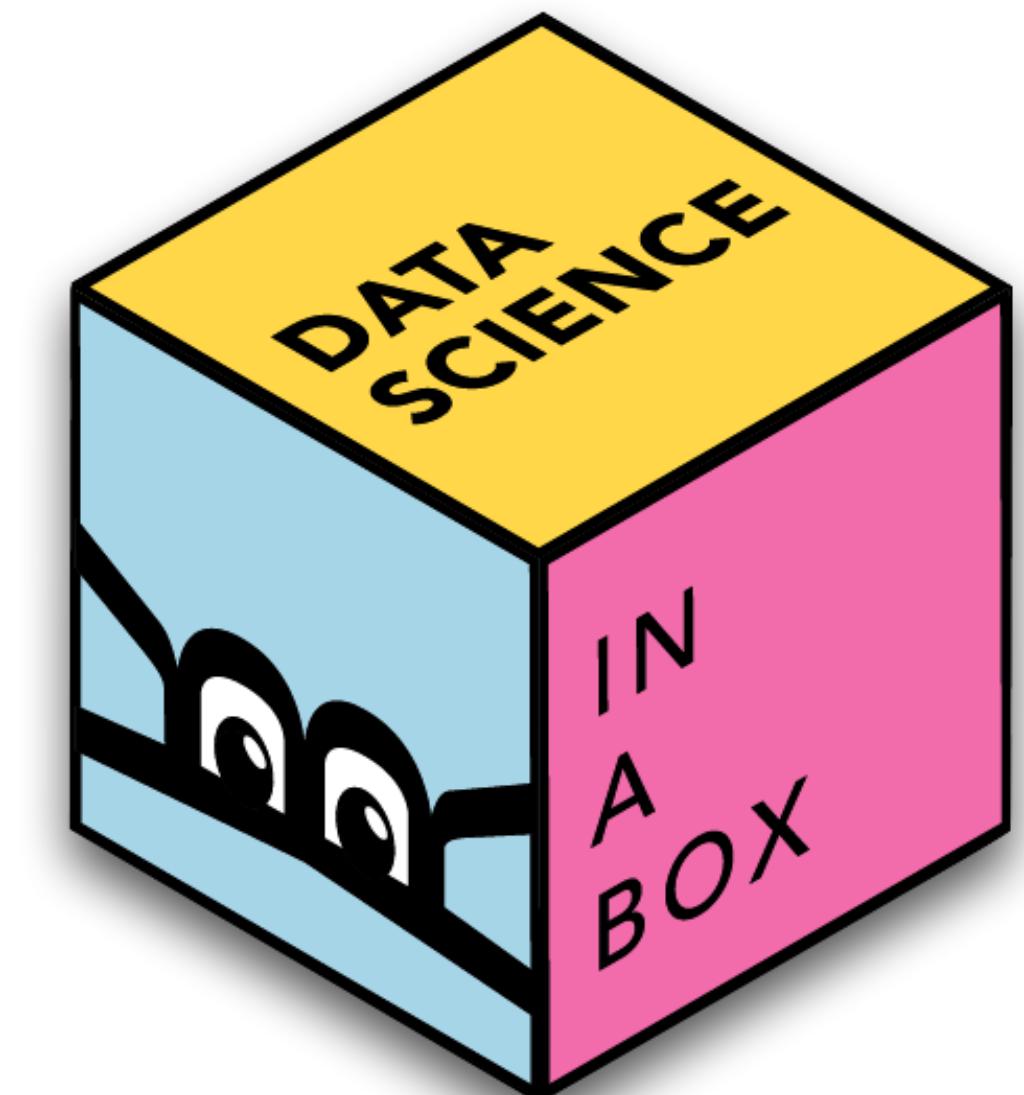
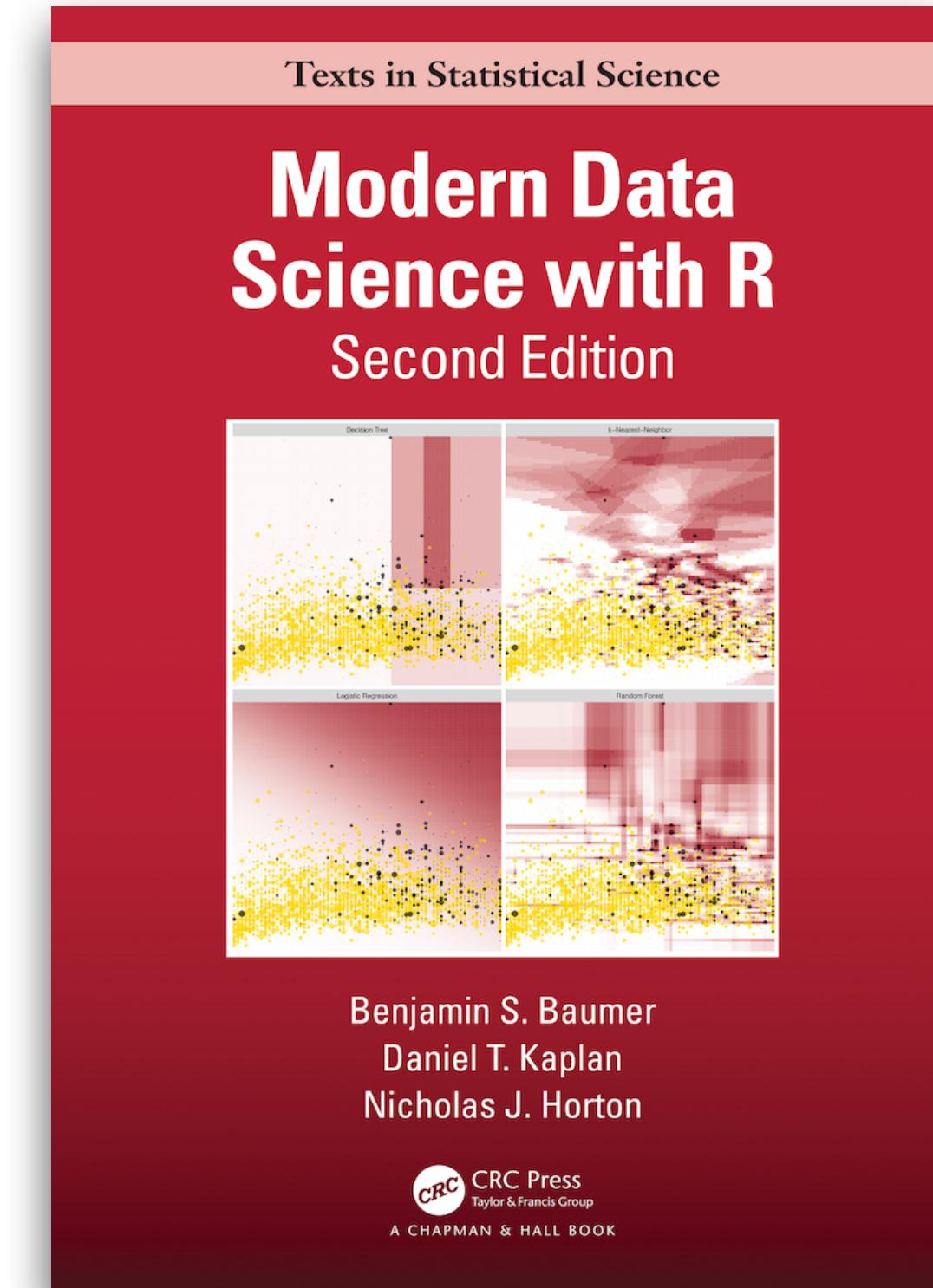
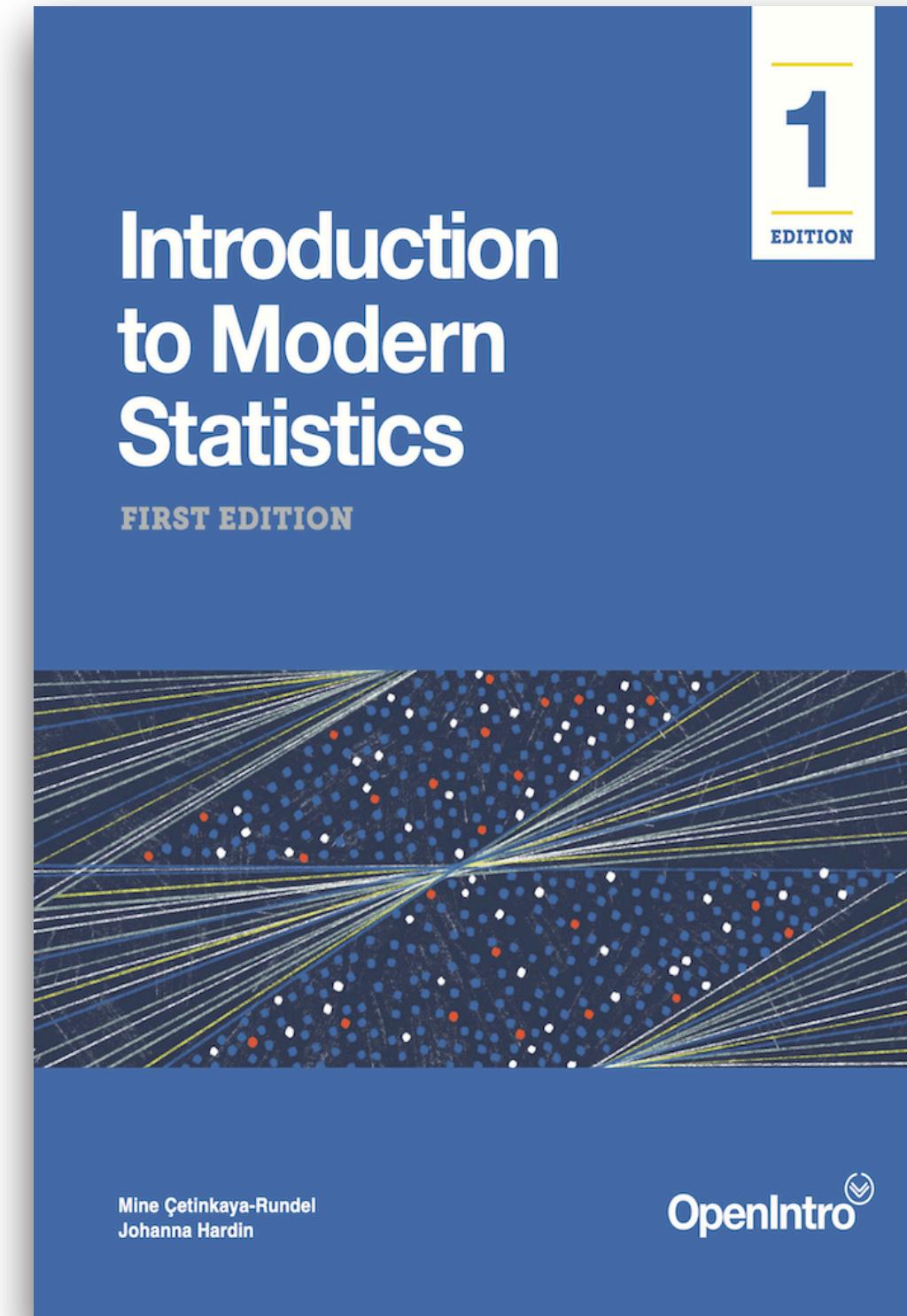
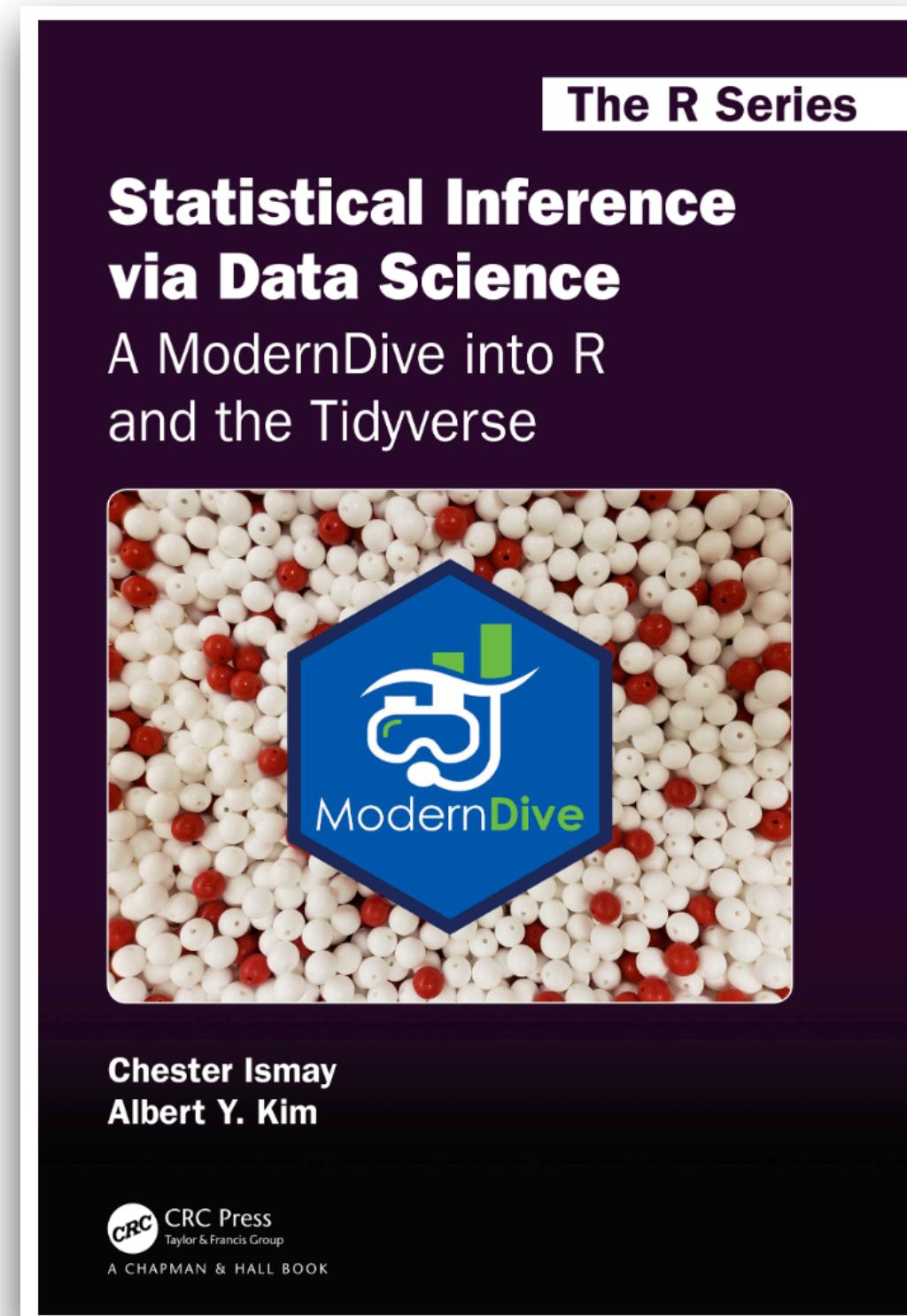
Provide opportunities for students to...

**Principle 1:** Regularly engage with complex (and relevant) real-world data and applications

**Principle 2:** Develop the skills and computational proficiency for a reproducible data analysis workflow

**Principle 3:** Develop important nontechnical skills, specifically written communication and teamwork

# Build on skills from introductory course



Modern Dive

Introduction to Modern Statistics

Modern Data Science with R

Data Science in a Box

Background  
and  
Motivation

Three principles

Challenges  
and  
next steps

# STA 210: Regression Analysis



**Students:** ~100 who have taken introductory statistics, data science, or probability course (majors and non-majors)

**Class Meetings:** 2 lectures with in-class activities and 1 lab

**Teaching team:** instructor, undergraduate and graduate teaching assistants

**Assessments:** labs, homework, exams, final group project

# Learning objectives

By the end of the semester, students will be able to...

- analyze real-world data to answer questions about multivariable relationships.
- use R to fit and evaluate linear and logistic regression models.
- assess whether a proposed model is appropriate and describe its limitations.
- use Quarto to write reproducible reports and GitHub for version control and collaboration.
- effectively communicate statistical results through writing and oral presentations.

# Principle 1

Regularly engage with complex (and relevant) real-world data and applications

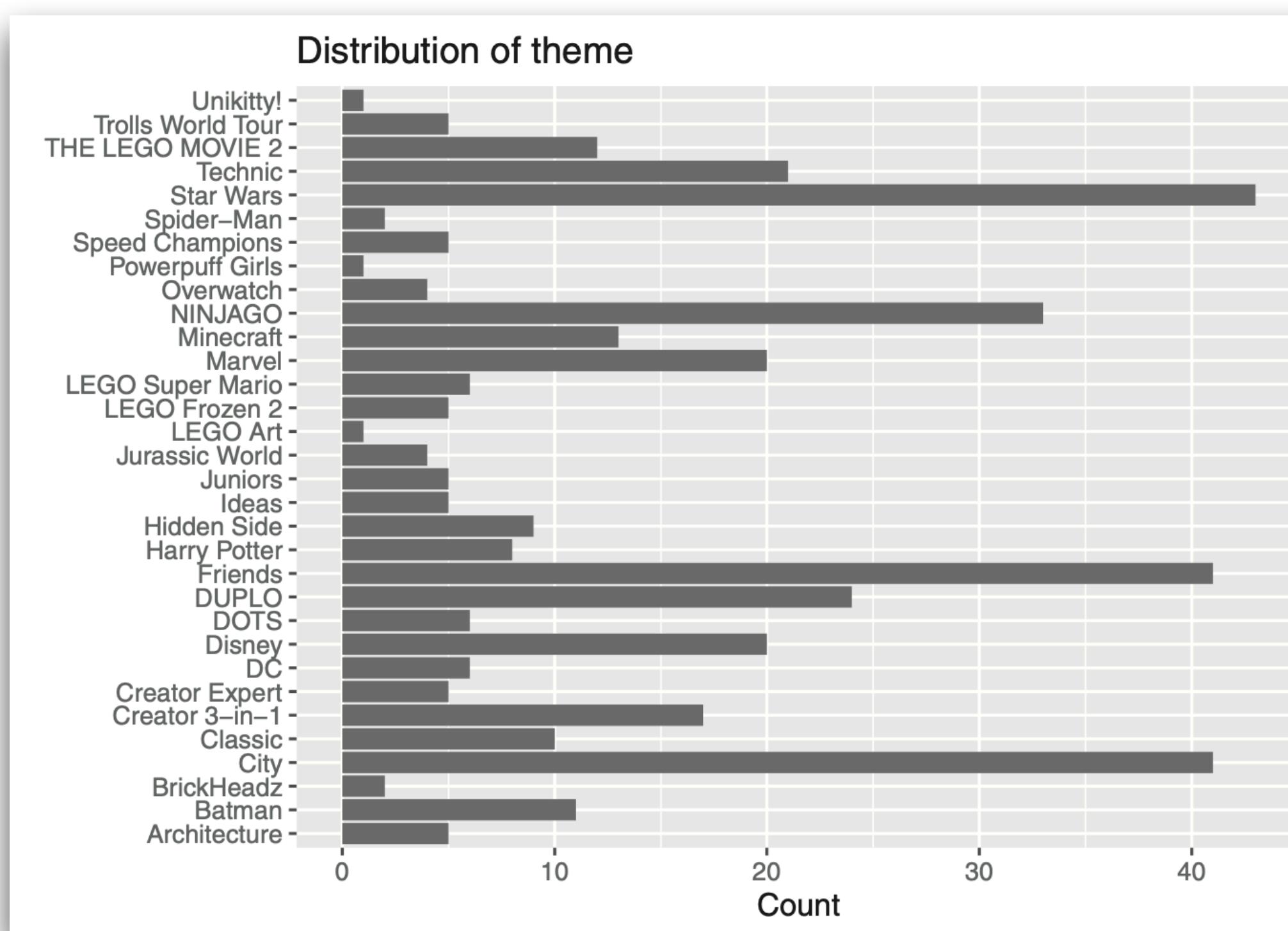
# Real-world data and applications

- “**Real-world**”: relevant and messy data that require some pre-processing before analysis
- **Goals:**
  - Give students exposure to data wrangling required before most regression analysis in practice
  - Demonstrates how regression is used in variety of interesting and relevant contexts
- **Where:** lectures, in-class activities, assignments

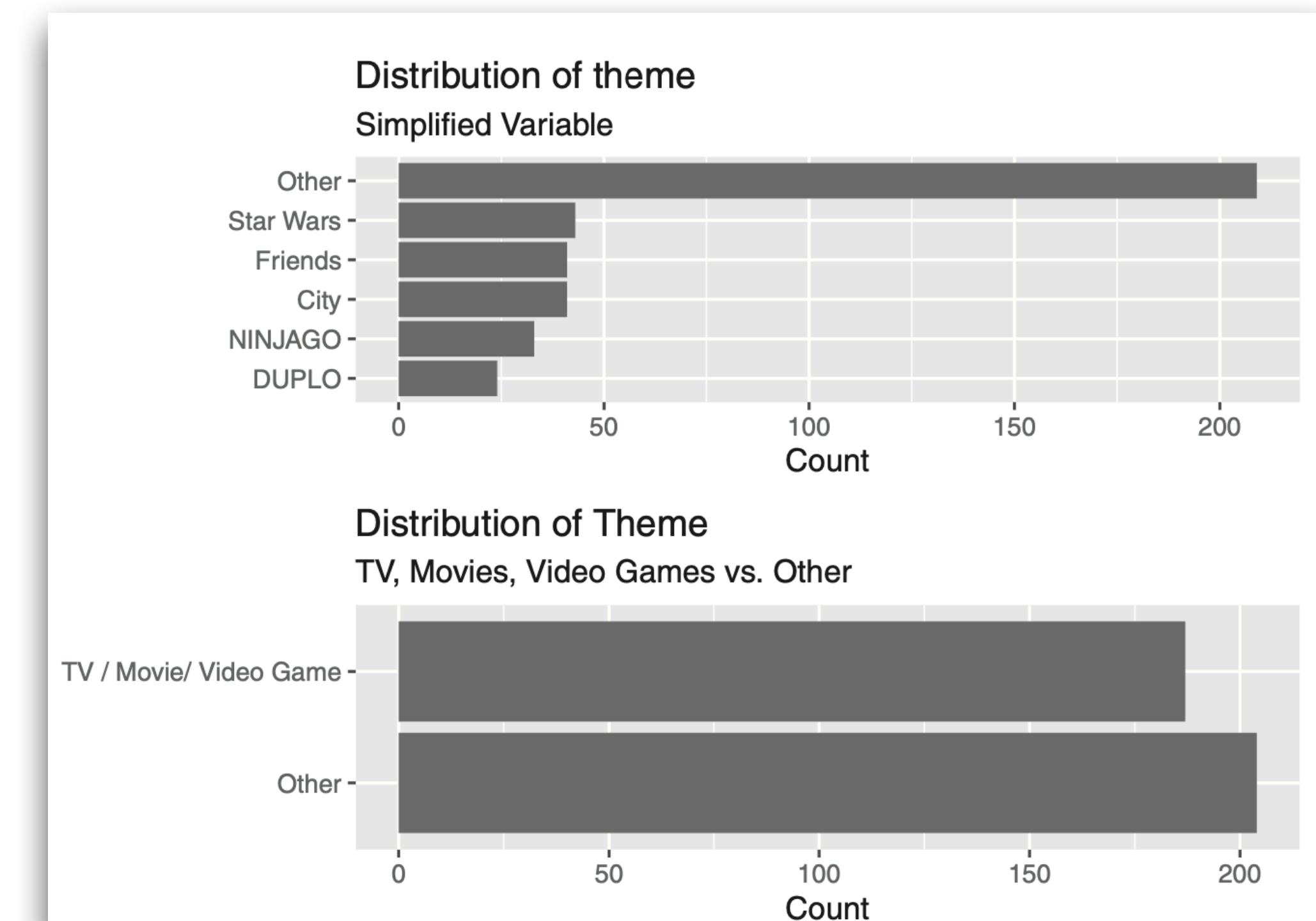
# Example: LEGO themes in-class activity

Students use data from Peterson and Ziegler (2021) to explore strategies to collapse levels of categorical variable

## Original

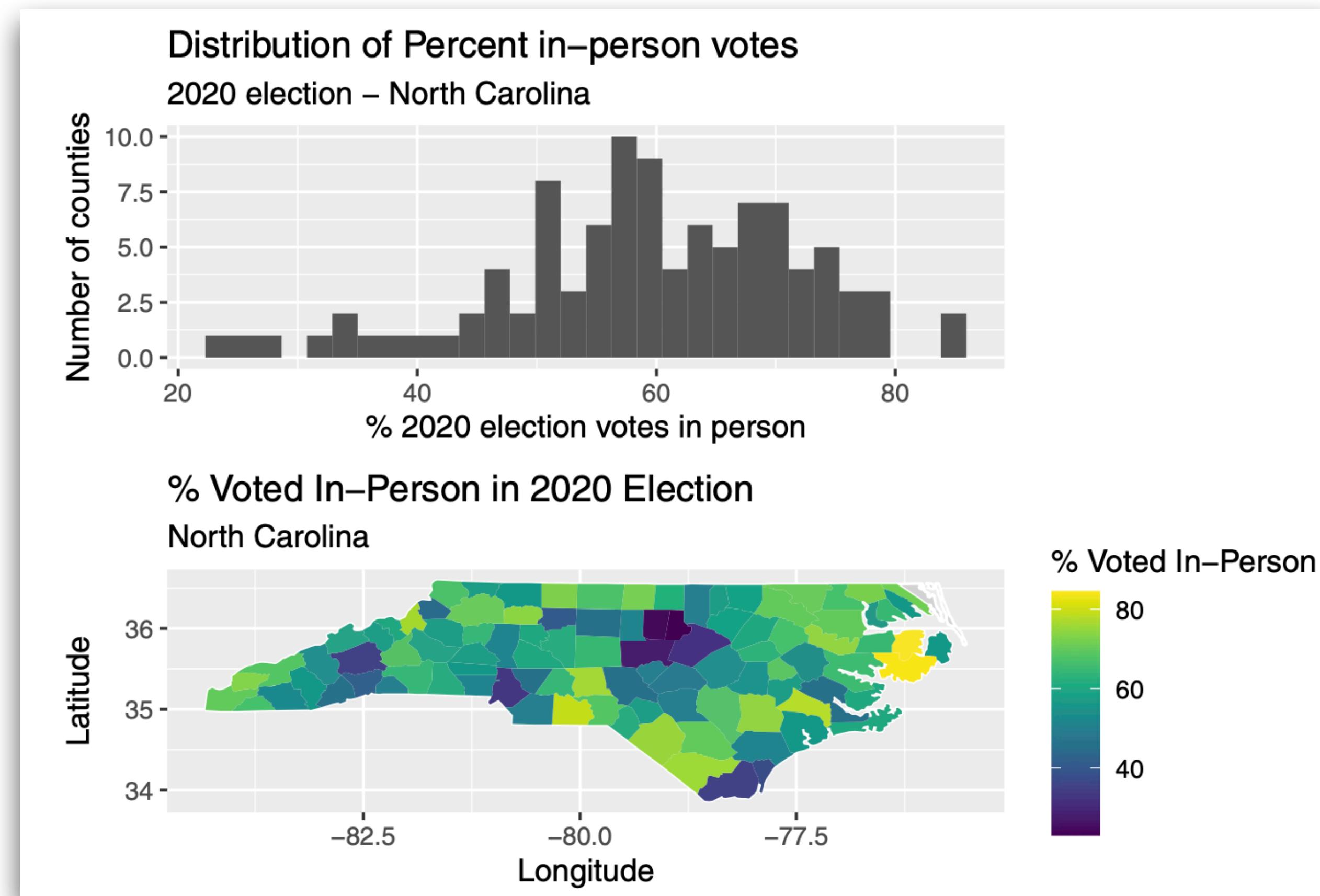


## Examples of student strategies



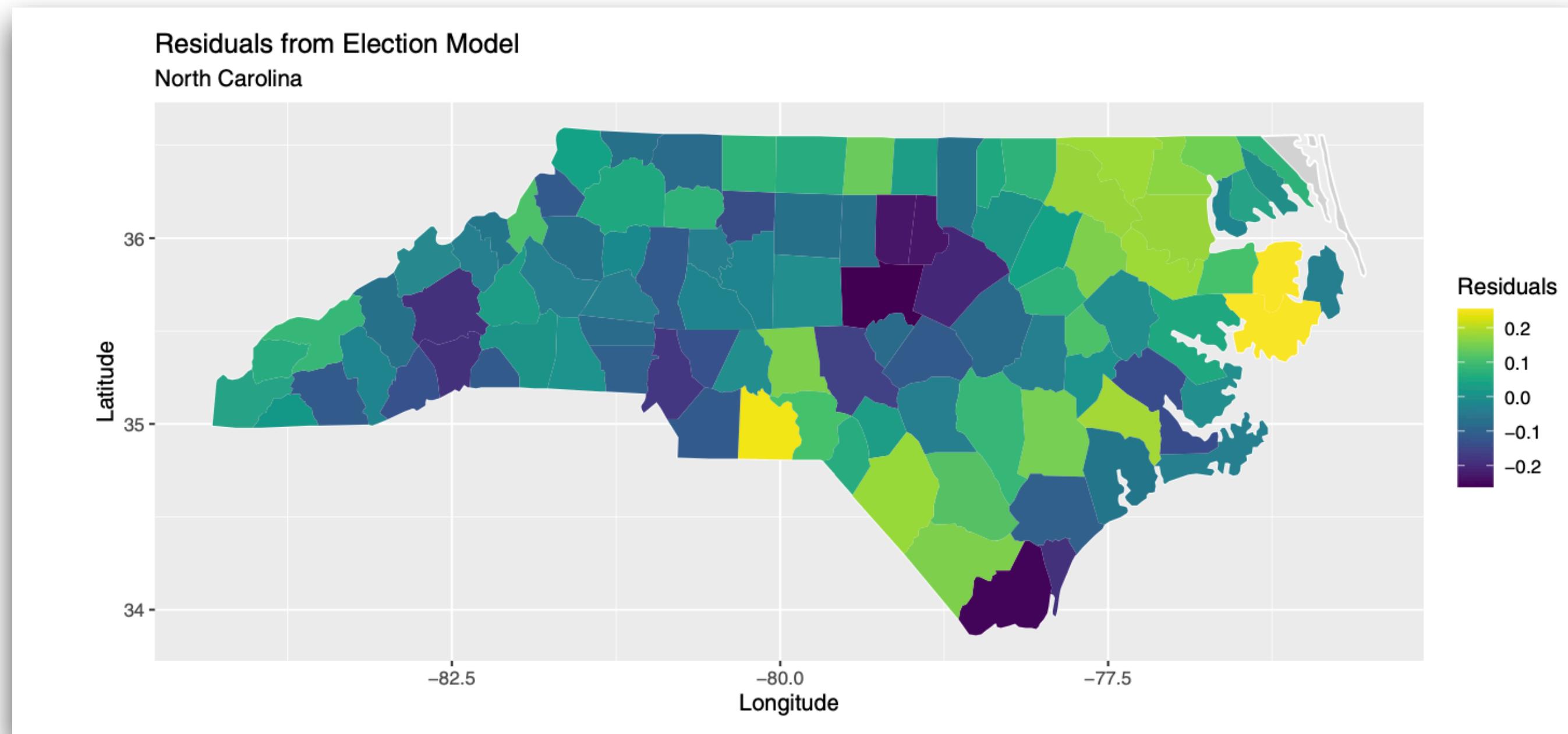
# Example: Assessing independence

Students consider potential spatial dependence in North Carolina voting data from the 2020 presidential election



# Example: Assessing independence

Students examine potential spatial dependence in residuals



- *Briefly explain why we may want to view the residuals on a map to assess independence.*
- *Briefly explain what pattern (if any) we would expect to observe on the map if the independence condition is satisfied.*
- *Is the independence condition satisfied? Briefly explain based on what you observe from the map.*

## Principle 2

Develop the skills and computational proficiency for a reproducible data analysis workflow

# Computing toolkit



- Quarto for assignment write-ups
- Run Git commands using point-and-click interface
- Server-based RStudio\*
  - Git already configured
  - Same set up for all students



- Assign and submit assignments
- Facilitates collaboration on group assignments
- Course management using **ghclass** R package (or GitHub Classroom\*\*)

\*Çetinkaya-Rundel, M., and Rundel, C. (2018), "Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum," *The American Statistician*, 72, 58–65,

\*\*Fiksel, J., Jager, L. R., Hardin, J. S., and Taub, M. A. (2019), "Using GitHub Classroom to Teach Statistics," *Journal of Statistics Education*, 27, 100–119.

# Motivating why reproducibility matters

- Lecture introducing reproducible workflow and computing toolkit
- Students study a case in which lack of reproducible practices had significant negative consequences (Ostblom and Timbers, 2022) and were asked their ideas to make the process reproducible

Use “intervention” and “usual” instead of 0 and 1

Save data in format they know will hold all their data

Have people review the spreadsheet

Enter data in an automated way

Document how they address missing data

Clearly mark missing data

# Reproduce an analysis

- Students work in groups to reproduce one model from an article in a scholarly journal
- They are asked what could have made the process easier

Having a codebook

Fixing typos

Being able to use the same software

Consistency between provided data and the description

Consistency in how variables are handled

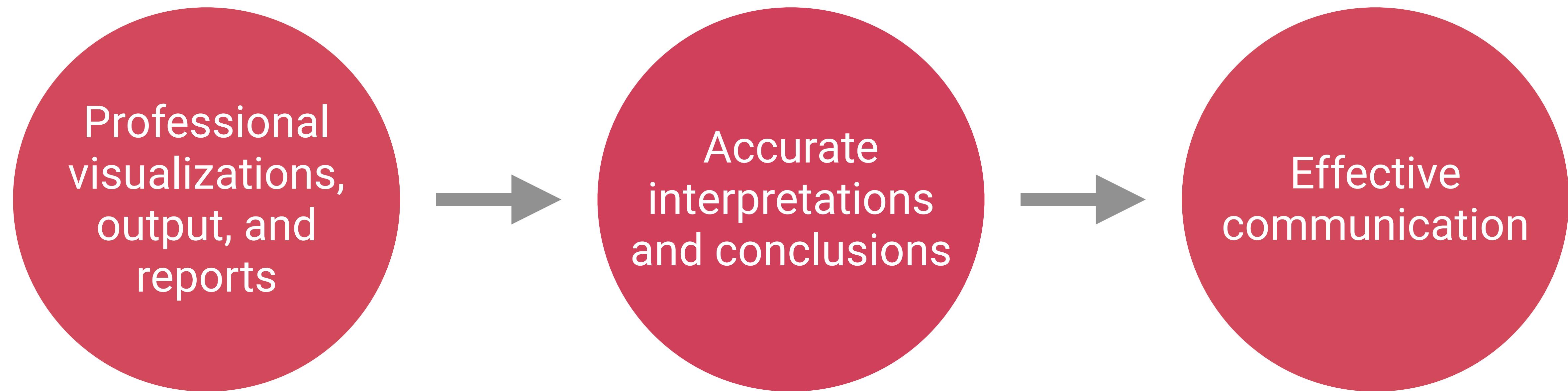
Knowing more about authors' thought process

Using informative variable names and categories

## Principle 3

Develop important nontechnical skills,  
specifically written communication and  
teamwork

# Teaching written communication



# Document formatting and presentation

Points on each assignment for...

- ✓ Writing all responses as cohesive narrative
- ✓ Clearly organized document
- ✓ Neatly formatted tables and output
- ✓ Informative titles and axis labels for visualizations

# “What’s the ‘so what’ ?”

- Goal is for students to get beyond basic interpretation to...
  - writing interpretations in a meaningful way
  - summarizing results to draw conclusions
- Assess analysis and summary separately to more easily identify student misunderstanding
- Do this first in short assignment questions and eventually in final project

# Example: King County, WA houses

Students analyze data about the price and other characteristics of houses in King County, Washington

- *Make a visualization of the price versus square footage with the points differentiated by waterfront. Interpret the visualization*
- *Fit a model with the log-transformed price (see the previous lab to see why we use log-transformed price!) as the response and sqft, waterfront, and their interaction as the predictors.*
- *Interpret the effect of square footage on the price of a house for*
  - *houses with no waterfront view*
  - *houses with a waterfront view*

**Conceptual  
understanding**

# Example: King County, WA houses

Students analyze data about the price and other characteristics of houses in King County, Washington

*Use the results from the previous questions to write a short paragraph (~ 3–5 sentences) about the relationship between square footage and the price of houses in King County, WA, and how (if at all) the relationship differs based on whether the house has a waterfront view. The paragraph should be written in a way that is practical and can be easily understood by a general audience of home buyers.*

**Effective  
communication**

# Addressing challenges

Finding data  
accessible to new  
learners

- Many data sets fail model conditions / require transformations
- Opportunity to get students excited about later units in the course and get exposure to realistic decision-making

Assessing writing

- Difficult to provide detailed individual feedback in large class
- Provide feedback on shorter writing exercises

Training teaching  
team

- Challenging to guarantee consistency in grading across multiple people
- Utilize detailed rubrics and regular meetings for discussions about grading

# What's next?

Include data ethics  
as a core learning  
objective for the  
course

Artificial  
intelligence



# Additional information

Article

## Three Principles for Modernizing an Undergraduate Regression Analysis Course

Maria Tackett 

Pages 116-127 | Published online: 02 Mar 2023

 Cite this article  <https://doi.org/10.1080/26939169.2023.2165989>

[doi.org/10.1080/26939169.2023.2165989](https://doi.org/10.1080/26939169.2023.2165989)

- STA 210: Regression Analysis Fall 2023 course website: [sta210-fa23.netlify.app](https://sta210-fa23.netlify.app)
- Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). [Implementing version control with Git and GitHub as a learning objective in statistics and data science courses](#). Journal of Statistics and Data Science Education, 29, 132-144.
- Alexander, R. (2023). [Reproducible Workflows. Telling Stories with Data](#). CRC Press.
- Çetinkaya-Rundel, M. (2020), “Data Science in a Box,” available at [www.datasciencebox.org](http://www.datasciencebox.org)



# Thank you!

[maria.tackett@duke.edu](mailto:maria.tackett@duke.edu)

Image credit: Sketchepedia on Freepik



[bit.ly/ares-more-than-methods](https://bit.ly/ares-more-than-methods)

# Resources for finding data

- OpenIntro
- TidyTuesday
- FiveThirtyEight
  - GitHub repo
  - R package
- Data is Plural

# STA 210 Course topics

Linear regression	Logistic regression	Looking ahead
Fitting and interpreting linear regression models	Fitting and interpreting logistic regression models	Topics to introduce students to methods beyond the course
Inference	Inference	Missing data imputation
Model conditions and diagnostics	Model conditions and diagnostics	Longitudinal modeling
Categorical predictors, polynomial predictors, interaction terms	ROC curve	Time series
Variable transformations	Prediction and classification	Poisson regression
Model selection	Model selection	Ordinal regression
<b>Feature engineering*</b>	Introduction to multinomial logistic regression	
<b>Cross validation*</b>		