Three principles for modernizing an undergraduate regression analysis course

Royal Statistical Society
Teaching Statistics Section

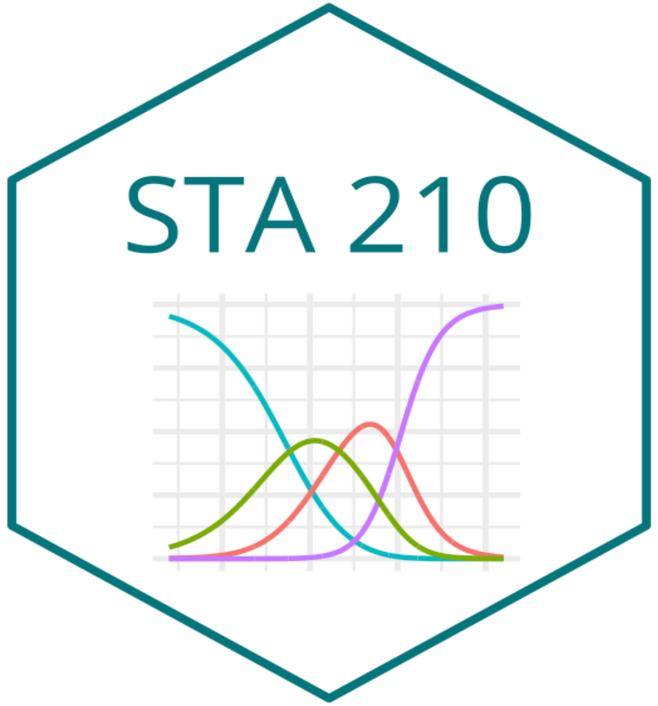
February 7, 2024

Maria Tackett
Duke University

🔗 bit.ly/rss-modernize-regression

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 teams…**"

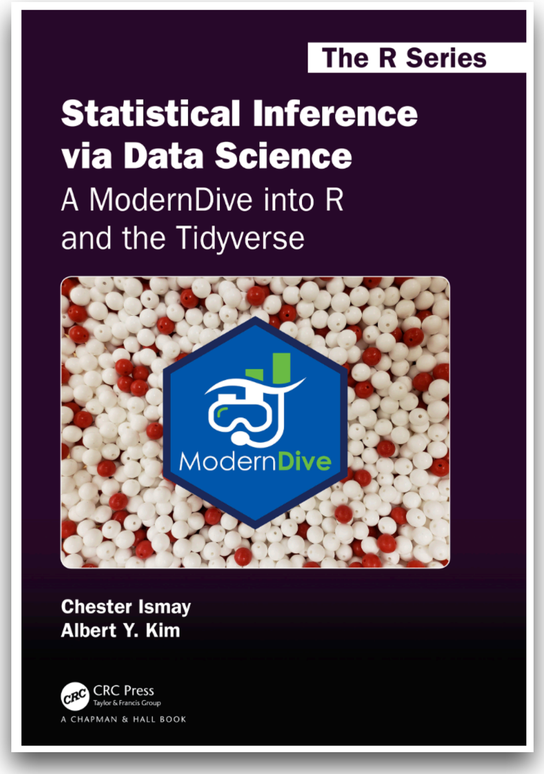
ASA Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines for Undergraduate Programs in Statistical Science*

Observations from final projects

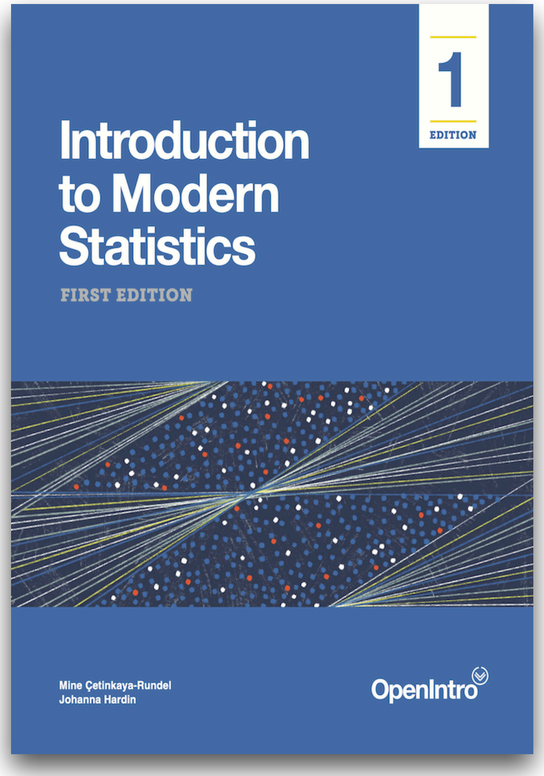
- Final group project throughout second half of the course
  - Use regression analysis to analyze a data set of their choice
  - Produce a written report and presentation

- Challenges students had…
  - Preparing the data for analysis
  - Effectively summarizing model results
  - Making analysis decisions

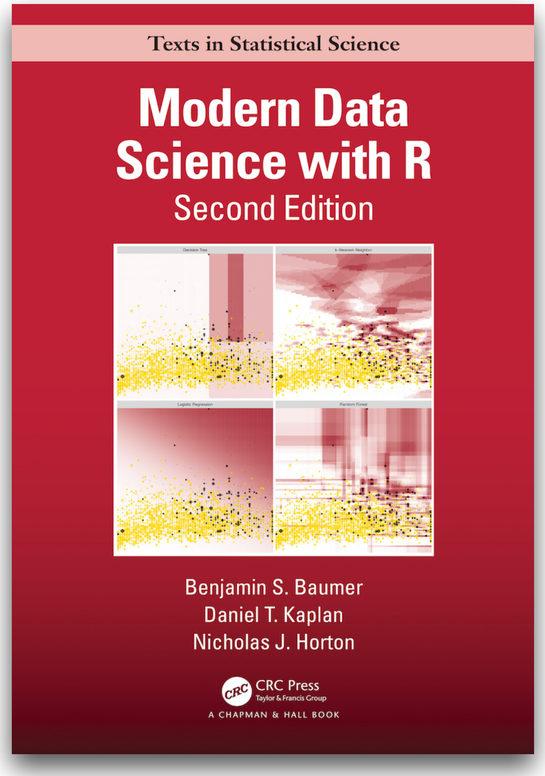
Build on skills from introductory course



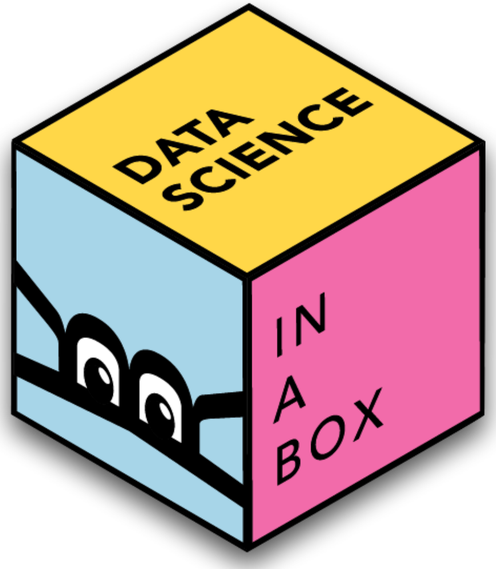
[Modern Dive]



[Introduction to Modern Statistics]



[Modern Data Science with R]



[Data Science in a Box]

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

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
Feature engineering*	Introduction to multinomial logistic regression	
Cross validation*		

Background and Motivation

Three principles

Challenges and next steps

Three principles for modernizing regression

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

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

Benefit for students

✅ Continuity as they continue developing data wrangling skills from introductory courses

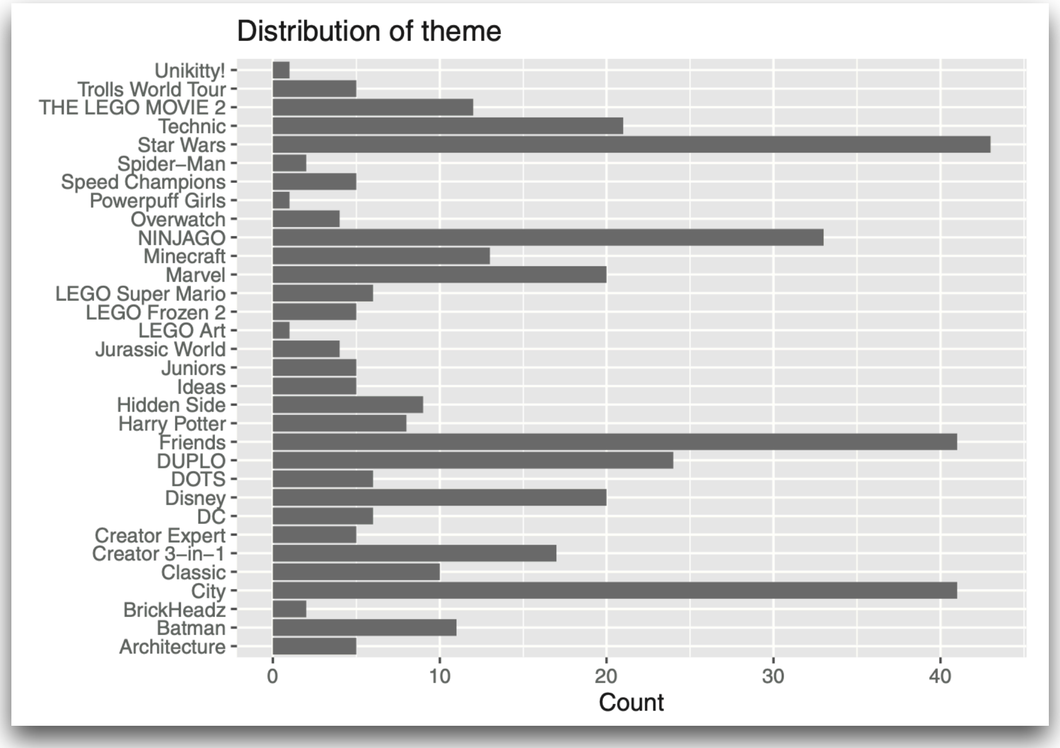
✅ Learn how to use visualizations and summary statistics to make data preparation decisions

✅ Consider the implications data preparation decisions has on the scope of conclusions and potential biases

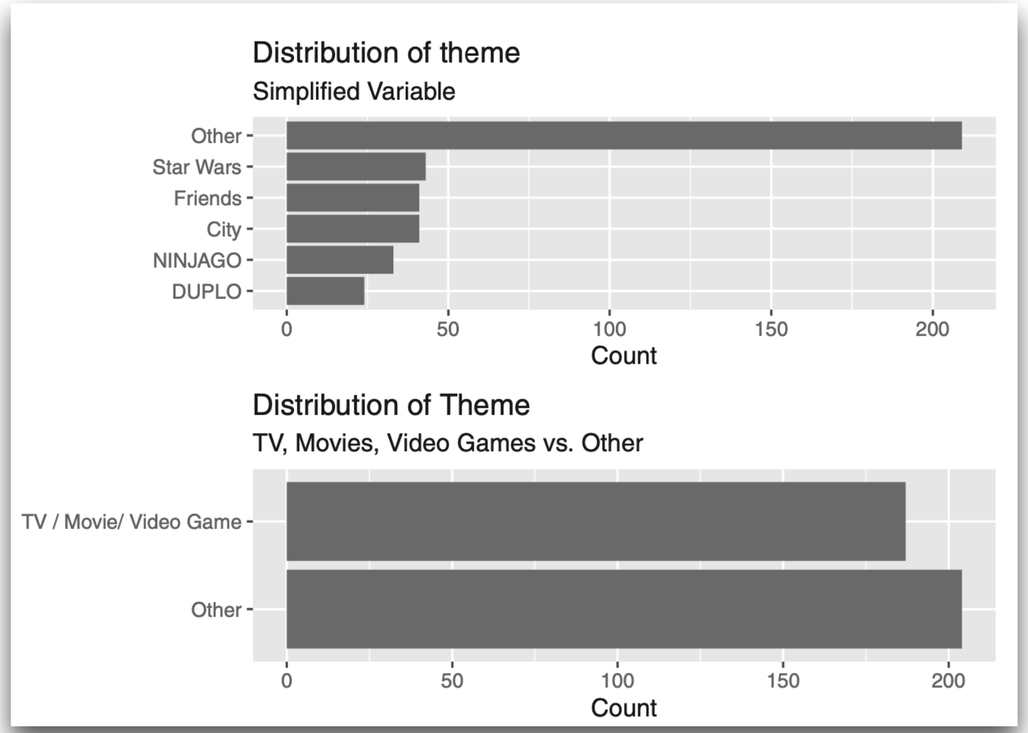
Example: LEGO Themes

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

Original



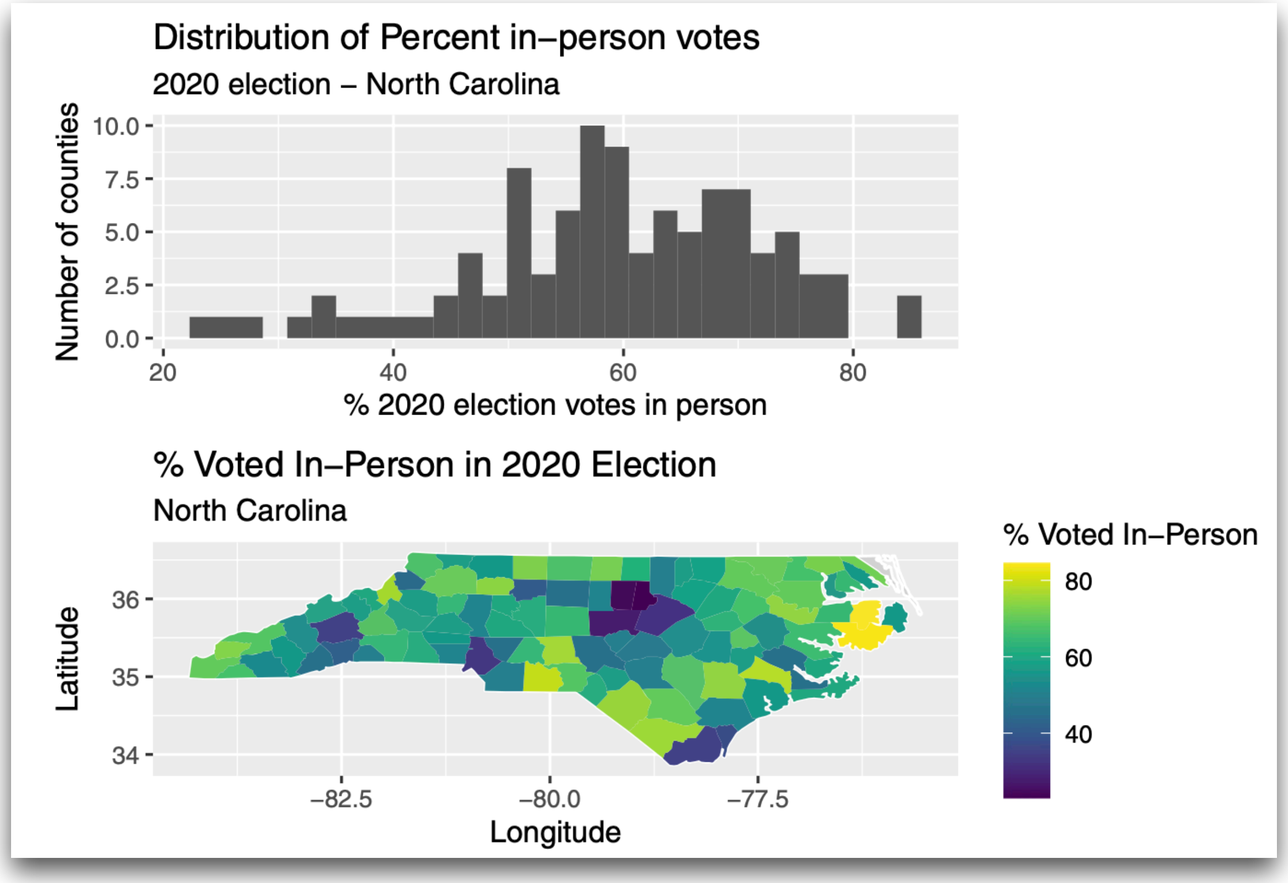
Examples of student strategies



Peterson, A. D., & Ziegler, L. (2021). Building a Multiple Linear Regression Model With LEGO Brick Data. *Journal of Statistics and Data Science Education*, *29*(3), 297-303.

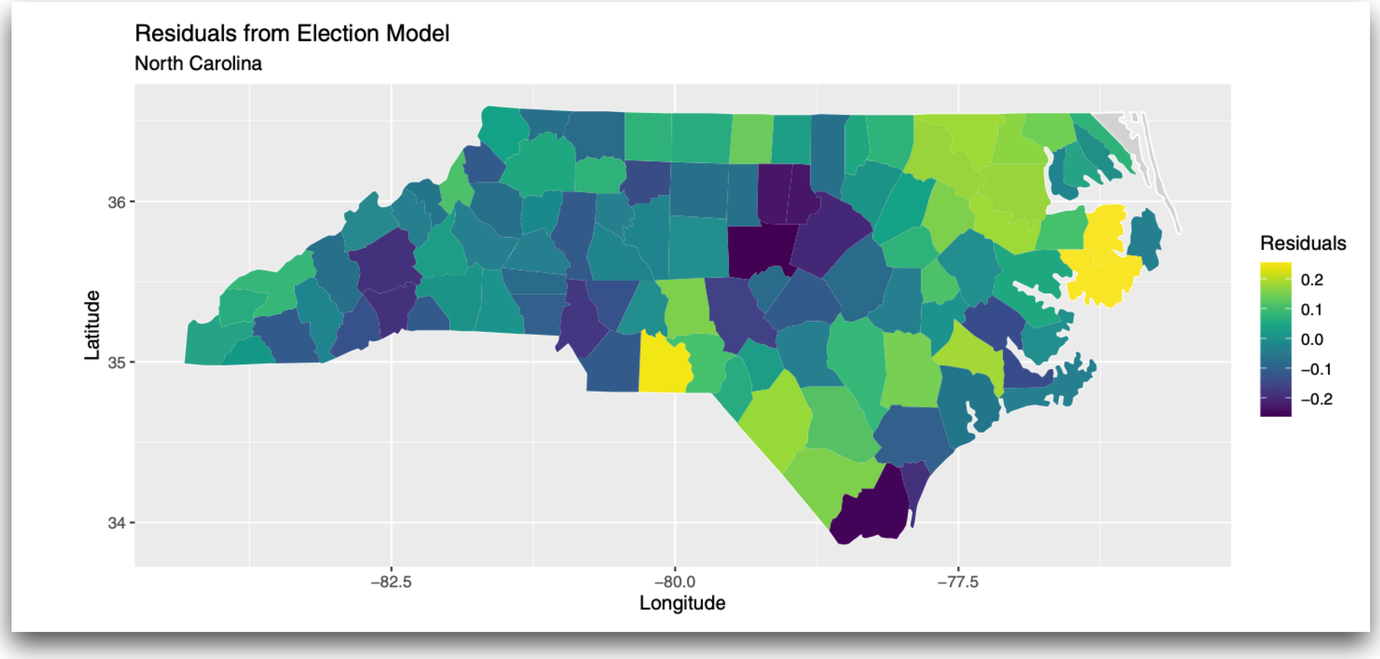
Example: Assessing independence

Students consider potential spatial dependence in North Carolina voting data



Example: Assessing independence

Students consider potential spatial dependence in North Carolina voting data



- *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.*

Resources for finding data

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

Principle 2

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

Benefit for students

✅ Develop computing skills necessary to work with messy real-world data

✅ Learn practices for reproducibility as they’re developing a data analysis workflow

✅ Foundation for learning more advanced computing skills in later courses

Computing toolkit

- Quarto for write up
- 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.

Introducing version control

- Lecture introducing reproducible workflow and computing toolkit
  - Help students understand value early on
- Start with individual assignments and using scaffolding to ease students into the new workflow

Individual assignment

This is a good place to render, commit, and push changes to your lab-01 repo on GitHub. Write an informative commit message (e.g. “Completed exercises 1 - 3”), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane in RStudio should be empty.

Group assignment

Team Member 1: Render the document and confirm that the changes are visible in the PDF. Then, commit (with an informative commit message) both the `.qmd` and PDF documents, and finally push the changes to GitHub.

Team Members 2, 3, 4: Once Team Member 1 is done rendering, committing, and pushing, confirm that the changes are visible on GitHub in your team’s lab repo. Then, in RStudio, click the **Pull** button in the Git pane to get the updated document. You should see the updated name in your `.qmd` file.

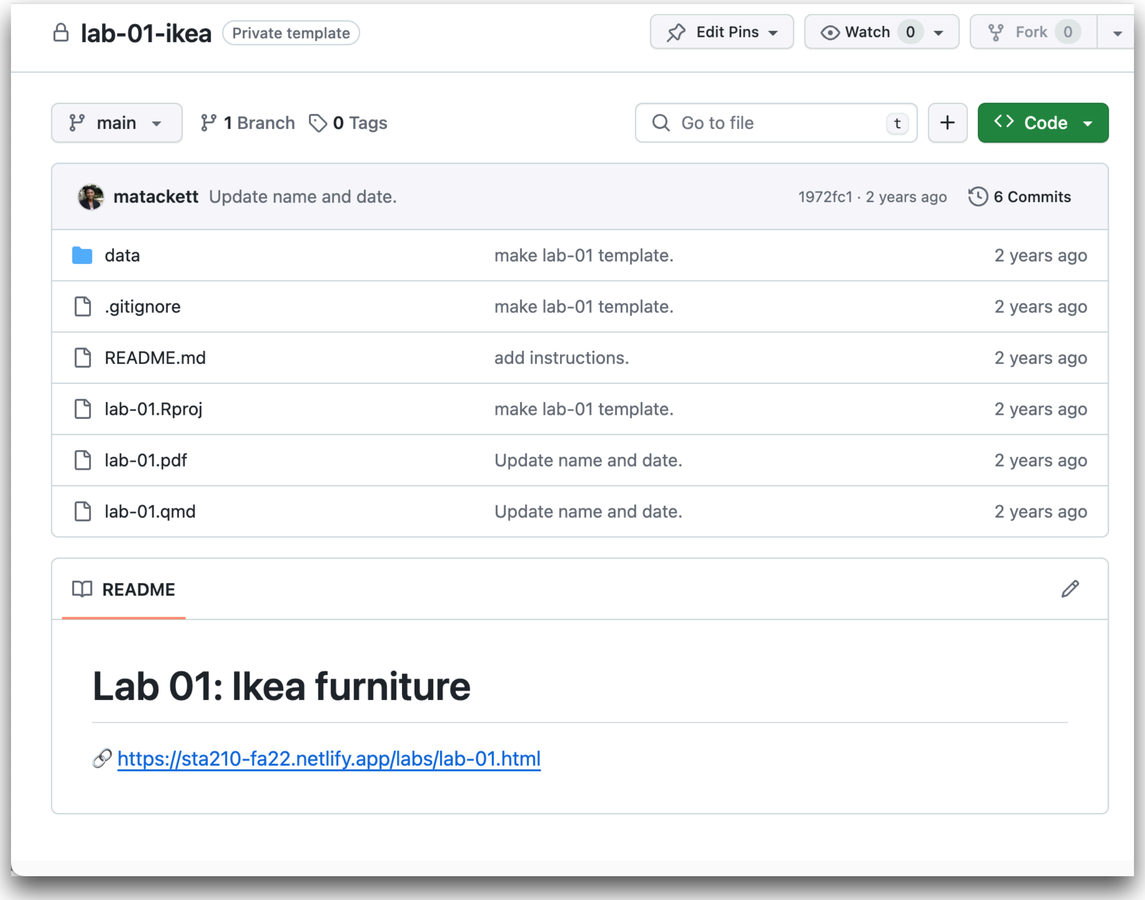
ghclass

- R package for managing courses on GitHub
- Use R functions to distribute assignments, manage teams, collect student work
- Reproducible administrative workflow

Developed by Colin Rundel & Mine Çetinkaya-Rundel

rundel.github.io/ghclass

Assignment workflow



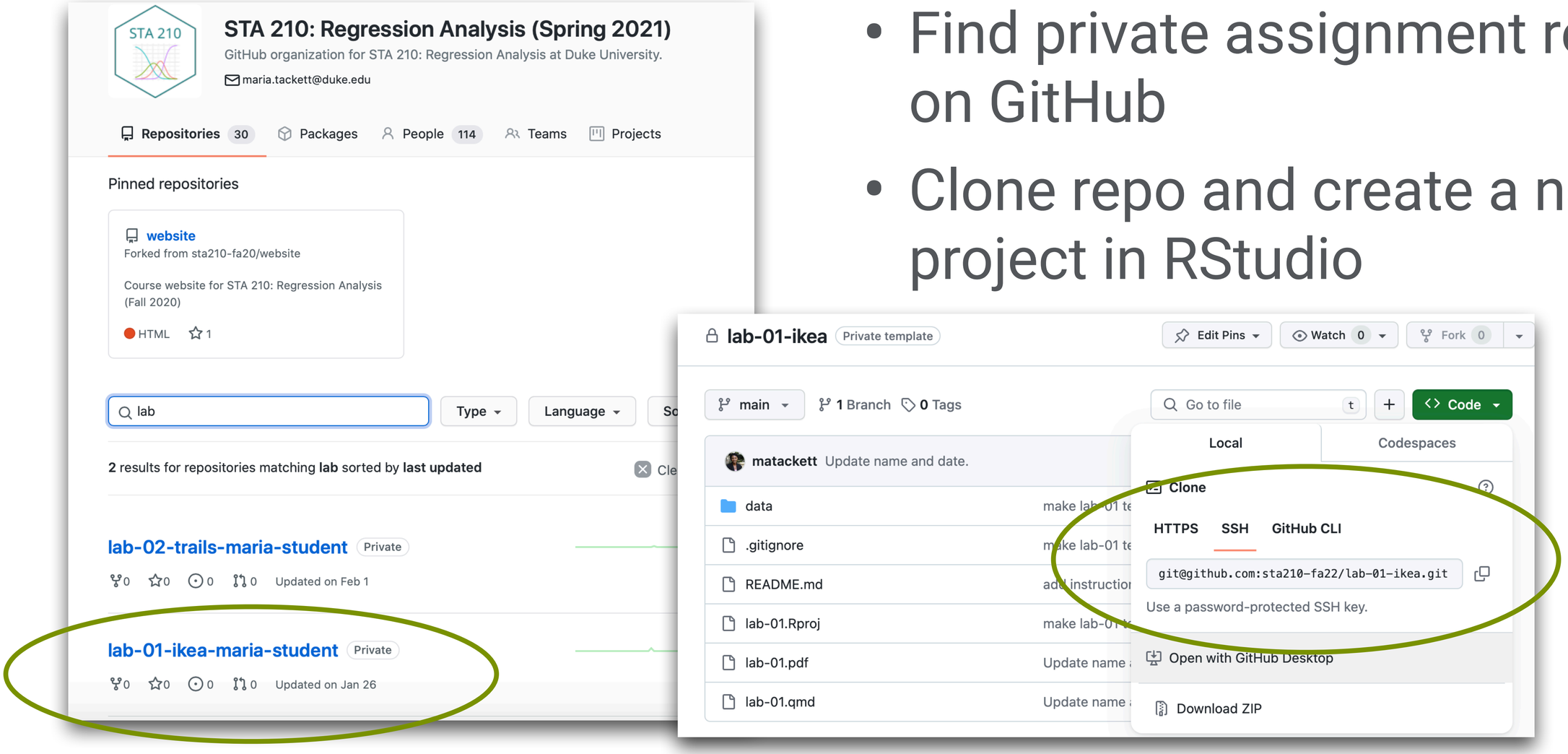
Instructor

- Create starter repo in GitHub.
  - Includes Quarto document, data set, etc.
- Make a copy of the starter repo for each student (or group) using **ghclass**

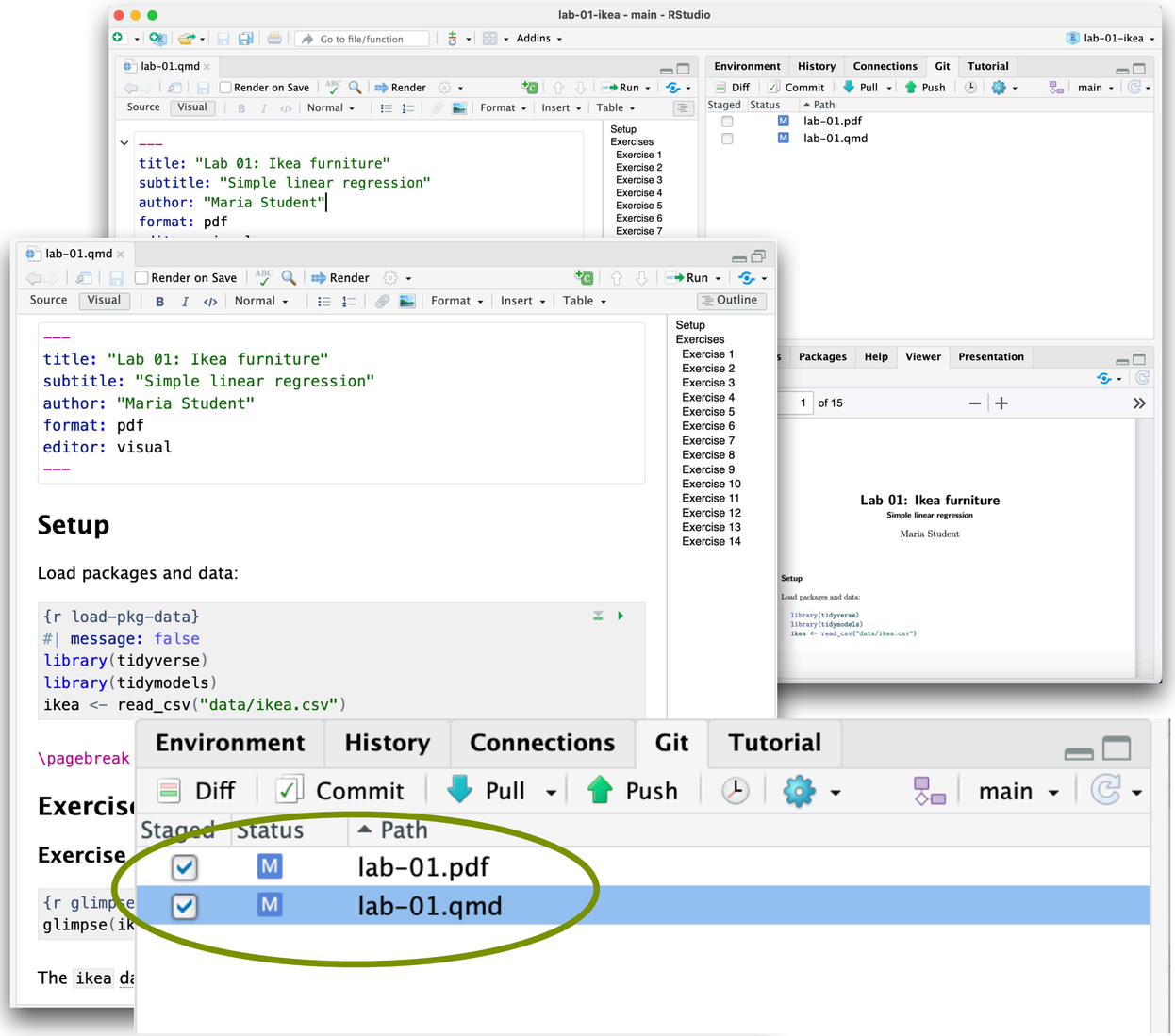
Assignment workflow

Student

- Find private assignment repo on GitHub
- Clone repo and create a new project in RStudio



Assignment workflow



Student

- As they work on assignment:
  - Write code and narrative in Quarto document
  - ***"Render, commit, and push"*** work to GitHub repo
- Final submission in GitHub, learning management system, or online rubric system

Assessment

Individual assignments

- ~ 6 - 10% of grade for formatting, reproducibility, and version control
- Assessed based on regular commits (3+) and informative commit messages

Group assignments

- ~ 10% of grade for formatting, reproducibility, version control, and collaboration
- Each group member’s contribution partially assessed based on commit history

Principle 3

Develop important nontechnical skills, specifically written communication and teamwork

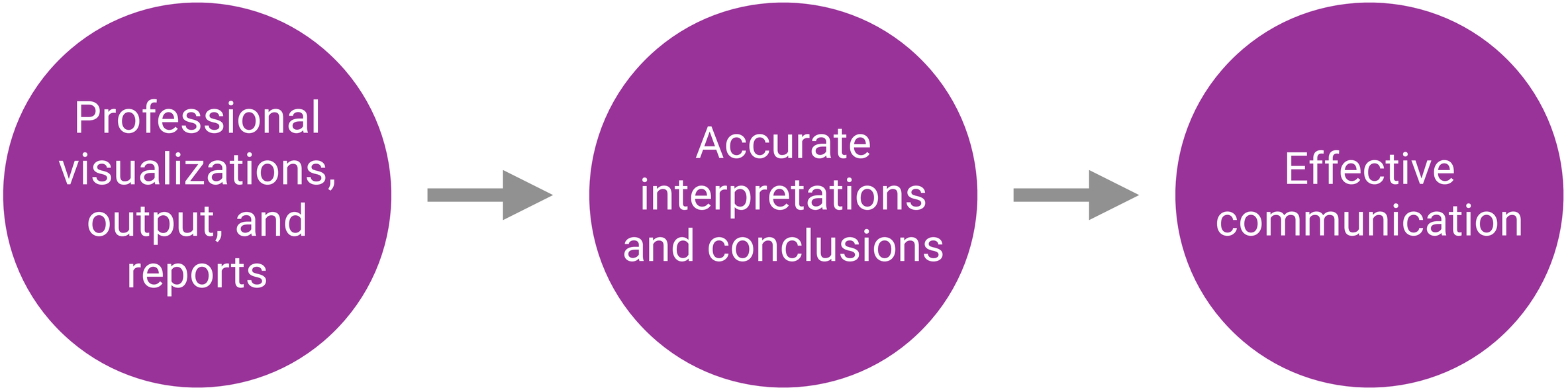
Benefit for students

✅ Learn skills to be a more effective collaborator in and outside the classroom

✅ Produce data analysis reports that can be included in portfolio of work

✅ Opportunity to learn from peers and develop leadership skills

Teaching written communication



Document formatting and presentation

- Points on each assignment dedicated to quality of formatting and presentation of document
  - Writing all responses as cohesive narrative
  - Document formatting suitable for a professional setting
  - Neatly formatted tables and output
  - Informative titles and axis labels for visualizations
- Students provided assignment templates and examples

“What’s the ‘so what’ ?”

- Goal is for students to get beyond basic interpretation to effectively communicating results
  - Writing interpretations in a meaningful way
  - Summarize 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

Teamwork

- Teams of 3 or 4 students assigned based on
  - previous statistics and computing experience
  - major or academic interests
  - trying to give each student at least one potential point of connection with their teammates
- Groups work together throughout the semester on weekly lab assignments and the final project

Teamwork

- The first team assignment includes
  - Completing a team agreement
  - Coming up with a fun team name!

- Teamwork is assessed based on contribution and collaboration
  - GitHub commit history on assignments to assess contribution
  - Periodic team feedback to assess collaboration

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 teams…**"

ASA Undergraduate Guidelines Workgroup (2014), *Curriculum Guidelines for Undergraduate Programs in Statistical Science*

Background and Motivation

Three principles

Challenges and next steps

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 for the course

- Update course topics to better reflect modern modeling
- Teach tidymodels syntax for opportunity to provide more continuity in coding throughout curriculum
  - Mixed success thus far
- Create more writing exercises for individual feedback
- Developing new course that incorporates mathematical theory without losing modern elements of the course

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

- STA 210: Regression Analysis Fall 2023 course website: 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. DOI: 10.1080/10691898.2020.1848485
- Çetinkaya-Rundel, M. (2020), "Data Science in a Box," available at www.datasciencebox.org

Thank you!

✉ maria.tackett@duke.edu

🔗 bit.ly/rss-modernize-regression