The Data Science of Everyday Life

This course will be an introduction to statistics, data management, statistical programming, and reporting for upper-level high school or early college students. The course is intended to impart these skills and understandings through the examination of publicly available data relevant to students' lives. When necessary, we will examine data analyses and reports from other researchers on relevant topics to explore methodology and practice. Where possible, we will directly collect publicly available data, clean it, analyze it, and produce reports on topics relevant to students' daily lives and the social context of the current era, including schooling, crime and policing, economics, housing, bullying, politics, and more.

We will take the issues that matter to students' lives and rigorously examine them with student-driven questions. For example: *Why is there a new epidemic of police shooting unarmed black men? Has Black Lives Matter had any impact on the world? What is the impact of policing in public schools? Are school shootings increasing? Does the lead in the drinking water in city schools disproportionately affect minority students?* All of these questions and their answers are contained within the course readings, and we will address the data and methods of those papers as examples of significant work in this genre. Then we will generate our own questions, find our own relevant data, and generate more answers of our own.

Course Materials

Grolemund, G., & Wickham, H. (2017). *R for Data Science*. O'Reilly Media. *Available for free online.*

R. Free download: http://www.r-project.org/
R Studio. Free download: https://www.rstudio.com/products/rstudio/download/

Additional Papers

See Appendix: we will be incorporating additional research papers as examples of data analysis used to investigate the personal and the political. A full list of papers is included at the end of the syllabus which includes papers we definitely will use, as well as papers that are available as alternatives depending on student interest and current events.

Recommended Resource

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage publications. (I will make copies of these available since it's not available free online)

Learning Objectives

Upon the completion of the course, students will be able to:

- Find, download, explore publicly available data
- Make basic graphs and plots, e.g. histograms, scatterplots, and basic linear models
- Clean data, including merging data sets, generating and removing variables, and correcting for type errors
- Make tables of data
- Conduct basic hypothesis testing with t-tests
- Conduct correlation and regression analysis, with multiple regression, fixed effects, and interactions
- Generate reports of data findings using R Markdown
- Understand the basics of R coding and using R Studio
- Read research papers with an eye toward methodology
- Connect contemporary social research to their everyday lives

Schedule

Week	**Topic**	**Assignment(s)**
Week 1	Introductions, finding data sets	Wickham Chapters 1 and 2 Class activity: downloading R, setting up Rstudio, getting comfortable with typing and coding
Week 2	Data Visualization part 1	Wickham chapter 3 Class activity: graphing city rent prices
Week 3	Data cleaning and transformation part 1	Wickham chapters 5 and 9 Cleaning 2 school shootings data sets
Week 4	Data cleaning and transformation part 2	Wickham chapters 11, 12, 13 Class activity: summary data tables from the Office of Civil Rights
Week 5	Data cleaning and transformation part 3	Wickham chapters 14, 15, 16 Students do a ‘data management’ project to find two datasets, merge them, clean them, and generate a visualization based on their result.
Week 6	Exploratory Data Analysis and Data Visualization part 2	Wickham chapters 7 Latham and Jennings lead in the drinking water EDA Class Activity: Find a new data set, what 5 things can you tell me?
Week 7	Statistics 1: Distributions,	Student data activity: normal distribution of height (and

Week	Topic	Assignment(s)
	Histograms	more!) Midterm Due
Week 8	Statistics 2: Correlation	Class Activity: 50 states data set: finding correlations in national politics Final project introduction.
Week 9	Statistics 4: Hypothesis testing (in theory)	Reading: Emily Rosa paper 1 sample t-test on therapeutic touch
Week 10	Linear Regression 1: Intro	Field Chapter 7 245-260 Gilens and Page on the utility of public opinion to get anything done
Week 11	Linear Regression 2: Multiple regression	Field chapter 7 p261-284 Cho on theories of big city crime
Week 12	Linear Regression 3: qualitative and fixed effects	Field chapter 7 p302-311 Campbell Black Lives Matter reading
Week 13	Linear Regression 4: Interaction & outliers	Return to 50 states data set: models with and without DC as an outlier (R shiny activity?). Teacher pay and Covid Rates interaction?
Week 14	Reporting 1: Markdown	Wickham chapters 26, 27, 28, 30
Week 15	Summary	What else is out there? Shiny, causal inference, ANOVAs, machine learning, etc. Final project due.

Grading and Rubric

The student's grade will be comprised of three pieces. For 30% of the final grade, students will complete in class activities and assignments on a weekly basis, essentially as a mark of active participation and completion. Secondly, to ensure that students keep up with readings as homework assignments, an additional 30% of the grade will come from a weekly reading quiz, which will include a short number of questions taken from the readings as a mark of outside of class participation. Finally, students will complete one midterm project and one final product showcasing their skills from the entire semester–these projects will be the remaining 40% of the student's grade. Late work will be accepted up until the last day of class and will be assessed with a one-time penalty for lateness (25% off value).

Midterm and Final Project

Midterm through the course, students will conduct a small data management project focused on the first several weeks of class, mainly data collection and cleaning. They will find two datasets, merge them, clean them, and generate one data visualization based on their results. Students will receive more explicit instructions and a rubric as the project is introduced in week 5 of the course.

For the final project, students will (a) find publicly available data on a topic of interest, (b) conduct an exploratory data analysis of their chosen data set, (c) clean the data as needed, (d) create some kind of model or finding, (e) include at least 2 data visualizations and (f) write up their process and what they discovered as an RMarkdown report. Rubric will be given to students half way through the course with more explicit instructions and expectations in terms of length, size, and scope.

Additional Papers And Resources

Campbell, T. (2021). Black Lives Matter's Effect on Police Lethal Use-of-Force. *Available at SSRN*.

Card, D., & Krueger, A. B. (1993). *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania* (No. w4509). National Bureau of Economic Research.

Cho, Yong Hyo. (1972). A multiple regression model for the measurement of the public policy impact on big city crime. *Policy Sciences* 3.4: 435-455.

Figlio, D. N. (2007). Boys named Sue: Disruptive children and their peers. *Education finance and policy, 2*(4), 376-394.

Gilens, M., & Page, B. I. (2014). Testing theories of American politics: Elites, interest groups, and average citizens. *Perspectives on politics*, *12*(3), 564-581.

Huang, F. L., & Cornell, D. G. (2019). School Teasing and Bullying After the Presidential Election. *Educational Researcher*, 0013189X18820291

Latham, S., & Jennings, J. L. (2022). Reducing lead exposure in school water: Evidence from remediation efforts in New York City public schools. *Environmental Research, 203*, 111735.

Rosa, L., Rosa, E., & Sarner, L. (1998). A Close Look at Therapeutic Touch. *JAMA, 279*(13), 1005-1010.

Sharkey, P., Schwartz, A. E., Ellen, I. G., & Lacoe, J. (2014). High stakes in the classroom, high stakes on the street: The effects of community violence on student's standardized test performance. *Sociological Science*, 1, 199-220.