Advanced Data Journalism

Instructor

Liz Lucas, associate professor and Houston Harte Chair in Journalism
Office: RJI 218
Hours: https://calendly.com/lucasek-umsystem/30min or by appointment
Email: lucasek@missouri.edu

Class overview

Data is used in journalism in myriad ways: to enhance reporting, to uncover inequity, to unmask corruption, to communicate big picture trends and numbers, to do large scale comparisons, and more. This class will build on the fundamentals of data journalism - finding, evaluating, cleaning and analyzing data – and equip you with the necessary tools to solve more complex problems that you will encounter while working with data in newsrooms.

Taking your data skills to the next level requires 1) special tools and technical skill, and 2) robust critical thinking and decision making. You will never be done developing your skills or your capacity for critical thinking, so think of this class as one early section of a career-long path. Each of you will come to this class with different interests and aptitudes; where you end up depends on where you want to go and how much effort you're willing to invest along the way.

While we will discuss many different techniques and strategies for accomplishing difficult tasks, our primary tool for exploring these concepts will be R, which is quickly becoming a favorite language of newsroom data teams, along with many individual reporters. Using R, you will have the opportunity to engage various datasets and analytical problems and learn to use data with accuracy, efficiency, and clarity. We will touch on these areas:

- Acquiring data from the web via APIs and scraping;
- Parsing, formatting, and otherwise preparing data for analysis;
- Conducting complex analyses with accuracy and integrity;
- Creating reproducible analyses;
- Using automation where applicable;
- Using artificial intelligence where useful;

- Presenting data findings and visualizations with clarity.

Learning objectives

Students will be able to:

- Comfortably use R and RStudio to conduct data analysis on complex data
- Employ the best practices of reproducibility
- Understand and execute the basics of scraping data from websites and using APIs to acquire data
- Use generative artificial intelligence to assist with writing code
- Present data clearly and effectively for storytelling
- Identify biases and shortcomings of common datasets

Your responsibilities

Your final grade is a reflection of your effort and your comprehension of the fundamentals, and will be comprised of the following:

1. Attendance and participation (25%)

You'll need your laptop for every class. I will help you download and install the appropriate software, which shouldn't require anything more than a computer that is reasonably up to date. We'll discuss computers and specs the first week.

The class will move through a series of sequential modules that relate to and build off each other. Because this learning relies on step-by-step progression, and because of the technical challenges posed by the nature of the material, it's very important that you attend class and keep up with the material from week to week.

Your on-time attendance, attention and participation is expected in every class session.

Since illness and life happen, you can be excused from class if you communicate with me ahead of time; I'll allow two such absences for any reason during the semester. You will be responsible for making up the material by talking to classmates or following up with me. Beyond those two absences, or for any unexcused absences (where you do not communicate with me either beforehand or immediately after), you will lose 5% of this

portion of your grade for each missed class. If there are extenuating circumstances that will cause you to miss more class, just come talk to me.

2. Assignments (25%)

Most weeks I will assign practice work that will be due on Mondays at noon. Generally these assignments will be exercises that will help you apply what we've learned in class. There will be 8-10 of them (depending on how we move through the material) and each will be worth 2-3 points, for a total of 25 points.

I will not grade your code (it's OK to turn in code that doesn't work! I will give you feedback on it), but you will lose a point if the assignment is late. I will not accept assignments after class on Wednesday (because we will review it in class), but there will be opportunities to make up for missed assignments.

Writing code is incredibly important in developing proficiency. It can be frustrating, right up until the moment when it starts to make sense and your code does what you want it to do. I strongly encourage you to work through that frustration.

Keep these things in mind:

Failure is necessary. I can almost guarantee that your first attempts will not work, which is how it is for most of us. Failure is an important step in learning.

There is no such thing as perfect code, or one way to write code. Code can be written more efficiently and elegantly (what we call "refactoring" code), but if your code is doing what you want it to do, then you have succeeded. So I encourage you to persevere, and not to be concerned with perfection.

Answer the question. Since we're utilizing code for journalistic purposes, most of the time we'll be trying to answer a question. You can write beautiful code that doesn't answer the question, but that doesn't fulfill the assignment

There are tools out there that will write code for you, such as ChatGPT and Claude. We'll explore when to use these to save yourself time, but a word of warning: don't trust output from artificial intelligence that you can't vet. I've had both ChatGPT and Claude give me bad code before. You must understand the fundamentals of programming before you can use code written by AI responsibly.

3. **Projects (50%)**

Over the course of the semester you will turn in two larger projects (25 points each) that demonstrate your proficiency in coding and your ability to use data for journalistic purposes. I expect your work to be newsroom worthy, something you wouldn't mind handing in to your editor at a newsroom job. Both projects should demonstrate a solid understanding of the fundamentals: know your data, document your work, check your work.

My responsibilities

As someone who has spent 15 years working with data in newsrooms, it is my goal to prepare you for working in a professional setting. I will work hard to make this material accessible to you; to provide helpful feedback on your work; to respond to your requests for help in a timely manner; to take into account your various aptitudes and interests; and to submit a final grade for each of you at the end of the semester that reflects your efforts and your proficiency.

I will be available to help you individually by appointment; sign up for a 30-minute meeting with Calendly or email me about a different day and time. I will try to respond to your emails within 24 hours.

A note about AI tools:

Generative AI may be used in this course. The content you submit is your responsibility. AI-generated code can be wrong (I have experienced this many times). Be sure your work accurately reflects your understanding and avoids these pitfalls. It is very difficult to use AI-generated code if you do not understand it well enough to vet it. If you do use GenAI, tell me so I can understand how you work and offer better feedback.

Semester Schedule

This schedule will be a guide; it may change depending on our pace and circumstances.

Module 1: Introduction to the class and the basics of programming

Weeks 1 - 3

- Discussion: what is data, what is data journalism, where do we see it, how do we evaluate it
- Learning R: computer setup, working in RStudio, basic R syntax
- Reproducible analyses: from the start we'll work in R Notebooks which help foster cleaner, reproducible analyses
- Introduction to Tidyverse: basic analysis concepts and functions

Module 2: Analysis and methodologies

Weeks 4 - 7

- Developing sound methodologies
- Discuss steps for thorough analysis
- Discuss best practices for projects
- Asking and answering good questions

Module 3: Vetting and cleaning your data

Weeks 8 - 9

- Discussing ethics and data bias
- How dirty and/or incomplete is your data?
- Preparing data for analysis

PROJECT 1 DUE (March 20)

Module 4: Acquiring data through advanced methods

Weeks 10 - 13

- Working with data in APIs
- Scraping data from the web

Module 5: Using AI in your workflow

Weeks 14 - 15

- How is generative AI helpful to data journalists?

- What are the ethical considerations of using AI for data journalism?

Module 6: Review and where to go from here

Week 16

- Final project presentations
- Advice, resources, final questions