EAES 494: Data Science and Statistics

Dr. **Gavin McNicol** (gmcnicol@uic.edu)   Fall 2021
Office Hours: By Appointment   M & W
TA: **Joey Pasterski** (mpaste3@uic.edu)   **11 a.m. – 12:50 p.m. in SELE 1280**
Office Hours: **3-4PM Wednesday (SES 2448) or Zoom**   4 credit hours

COURSE DESCRIPTION: This course will introduce you to fundamentals of coding in the free and widely used programming language 'R'. Through hands-on classroom programming and interactive data analysis, you will become familiar with the data science workflow including how to import, tidy, visualize, analyze, and communicate big EaES datasets, including geospatial data, and how to use version control to ensure your work is reproducible, a valued skill in both private industry and academic research. The course will also cover fundamental statistics including data distributions, descriptive statistics, and linear regression. You will need access to a personal computer but there are no math or computer science (coding) prerequisites for this class.

PREREQUISITES/COREQUISITES: None.

OBJECTIVES: As a result of this course, you will be able to:

- Write code with proficiency in the freely available programming language R and its development environment RStudio. We will use RStudioCloud, a browser-based version of this software.

- Use RMarkdown syntax to produce .html and .pdf assignment and project reports that combine code chunks and outputs, and text narrative.

- Use Git and GitHub for version control of your coding projects, both as an individual and as a team collaborator.

- Identify and acquire Earth and Environmental Science datasets online via existing R packages and other sources (e.g., US Geological Survey, NOAA, NASA, etc.)

- Implement each step of the data science workflow to import, tidy, visualize, wrangle, model, and analyze Earth and Environmental Science data.

REQUIRED TEXTS AND RESOURCES:

- Wickham, Grolemund et al., 2021. **R for Data Science (R4DS)**

    o   Free e-textbook: https://r4ds.had.co.nz/index.html

- Çetinkaya-Rundel and Hardin, 2021. **Introduction to Modern Statistics (IMS)**

    o   Free e-textbook: https://openintro-ims.netlify.app/index.html

- Some additional readings from other R Data Science sources.

COURSE POLICIES AND RESOURCES

ATTENDANCE AND PUNCTUALITY: In-person attendance at all scheduled Monday and Wednesday lecture sessions is <u>required</u>. Attendance is essential because each class is a mixture of lecture, interactive exercises, and time to work on lab assignments. In-person class time is the best time to ask questions and troubleshoot problems because we can help you right away. We will take attendance with iClicker Cloud during each class. Before the first week of class, you'll complete a Google Form about your goals for the class and your academic background.

COMMUNICATION: Class materials (e.g. syllabus, links for exercises, labs, and homework, etc.) will be posted on Blackboard and announced via email. You will be invited to a Slack workspace (**EAES494-DataSciAndStats-Fall2021**) for the course including channels for class exercises/labs, RStudio, GitHub, and other troubleshooting. For course logistics, review the syllabus first, and if the information is not there, ask a question on Slack. Only email the instructor or TA if you have tried those options first, or for any personal matters including absences and technical issues.

ACADEMIC INTEGRITY POLICY: As an academic community, UIC is committed to providing an environment in which research, learning, and scholarship can flourish and in which all endeavors are guided by academic and professional integrity. All members of the campus community-students, staff, faculty, and administrators-share the responsibility of ensuring that these standards are upheld so that such an environment exists. Instances of academic misconduct by students will be handled pursuant to the Student Disciplinary Policy.

RELIGIOUS HOLIDAYS: Students who wish to observe their religious holidays shall notify us by the 10th day of the semester of the date(s) when you will be absent. We will make every reasonable effort to honor the request and not penalize the student for missing the class. If you feel unsatisfied with our accommodations, you may request remedy through the campus grievance procedure.

DISABILITY ACCOMMODATION: UIC is committed to maintaining a barrier-free environment so that students with disabilities can fully access programs, courses, services, and activities at UIC. Students with disabilities who require accommodations for access to and/or participation in this course are welcome but must be registered with the Disability Resource Center (DRC). You may contact DRC at 312-413-2183 (v) or 773-649-4535 (VP/Relay).

UIC RESOURCES: If you find yourself having difficulty with the course material or any other difficulties in your student life, don't hesitate to ask for help! Come to us, or if it is about an issue beyond this class, please contact your college advisors or get help from any number of other support services on campus. *The UIC Counseling Center* also offers free individual and couple counseling, and group therapy sessions, free mind-body programs (e.g., mindfulness workshops), and access to psychiatric and integrative medical services. https://counseling.uic.edu/

FACE MASKS: Masks covering both the mouth and nose must be worn at all times by all students, faculty, and staff while on campus and inside any building regardless of vaccination status. If you do not wear a mask, you will be asked to leave the classroom and will not be allowed back in class unless or until you wear a mask. If you have forgotten your mask, you may pick one up from one of the student information desks on campus during the first two weeks of campus. Students who do not comply with the mask wearing policy will be reported to the Dean of Students. Eating and drinking is not allowed in classrooms.

COURSE REQUIREMENTS AND GRADING

Deadlines for all assignments are listed in the course calendar at the end of this syllabus.

Labs (8, lowest score dropped)   70 points

You will complete a lab assignment for each major topic in the course, approximately weekly. These labs let you apply the data science and statistics methods from lecture to EaES datasets. Aim to complete the labs <u>during class time</u> so you can discuss any issues with the instructor or teaching assistant. (10 pt each)

Homework (6, lowest score dropped)   50 points

You will complete a homework assignment for each major topic in the course, approximately weekly. You should aim to complete the homework <u>by yourself</u> after weekly lectures to reinforce skills covered in class lessons and to become comfortable troubleshooting problems by yourself. (10 pt each)

DataCamp Tutorials (6)   60 points

You will complete online tutorial assignments via DataCamp.com. These are a combination of short lectures from R experts and training examples. Plan to spend 2 hours per week on these, however, they will not be assigned every week. These tutorials are not graded, but you will have to complete all assigned course chapters to get points for this component of the course. (10 pt each).

Midterm (1)   50 points

An in-class midterm exam will be assigned mixing short narrative (written) and code responses.

Collaborative Final Project (1)   100 points

The final project for this class will consist of analysis on a dataset of your own choosing. The dataset may already exist, or you may collect your own data using a survey or by conducting an experiment. You will submit a proposal (20 pt), give a presentation (50 pt), and an executive summary (30 pt). More details to be provided later in the course. It is TBD whether the final project will be a team or individual activity.

Total   330 points

TIME COMMITMENT: This course will introduce you to R and data science methods quickly in the first 2 weeks, so that you will gain hands-on experience early in the semester. It may feel like a steep learning curve initially but it will get easier. Expect to spend about 1-2 hours for each lab and homework. You will get longer (2 weeks) for the larger DataCamp assignments that consist of 4 chapters each.

GRADING: To determine your final grade in the course, I will calculate your percentage by dividing your total earned points by the total points possible. These will be converted to letter grades as follows:
              90-100% = A   80-89% = B   70-79% = C   60-69% = D   < 60% = F

ASSIGNMENT SUBMISSION:
- **Class exercises**: Exercises completed during class between lecture portions are <u>not</u> submitted or graded. We can review your work during class as we work through the exercises.
- **Labs, homework, and final project**: All labs, homework, and the final project will be submitted via GitHub, the version control system. The file you submit will be an RMarkdown file (.Rmd).

LATE WORK: Two points will be deducted per calendar day for late lab and homework assignments.

WK Day	Date	Class Agenda/Topics	Asynchronous Work/Due Dates (Due at the end of the day (11:59pm))
		EAES 494 - FALL 2021- COURSE CALENDAR	
		This calendar is subject to change at the discretion of the instructor.	
1 M	8/23	***UNIT 1- WELCOME*** ● *Welcome to Data Science* ● Exercise 1: Global CO2 Emissions	● Reading: R4DS::Ch. 2 - Introduction ● Reading: IMS:: 1. Data Basics

W	8/25	● Brief tech. recap
		● *Meet the toolkit: Programming*
		● Exercise 2: Weather Check
			● **Work on DataCamp 1 (Due 9/1)**
			  ● Intro to basics
			  ● Data frames
			  ● Vectors

2 M	8/30	● **Recap** ● *Meet the Toolkit: Version control* ● **Start Lab 1**: Hello R!	<u>**Continue DataCamp 1 and Readings...**</u>
W	9/1	● **Finish Lab 1**: Hello R! ● **Start Homework 1:** Earthquakes **(Due 9/8)**	● <u>**DataCamp 1 Due**</u>: Intro to Basics, Frames, Vectors ● **Start DataCamp 2 (Due 9/15)** ● Intro. to Data Viz with ggplot2 (All 4 chapters)
3 M	9/6	● **No class - Labor Day**	● **Continue DataCamp 2 (Due 9/15)** ● Intro. to Data Viz with ggplot2 (All 4 chapters)
W	9/8	***UNIT 2 - DATA VISUALIZATION*** ● *Data and visualization* ● Exercise 3a: Class Poll DataViz	● <u>**Lab 1 and Homework 1 Due!**</u> ● Reading: R4DS::3 - Data Visualization

4 M	9/13	• *Visualizing data with ggplot2* • *Visualizing numerical data* • **Start Lab 2:** Plastic Waste	• Reading: [IMS::4 Exploring categorical data](#) • Reading: [IMS::5 Exploring numerical data](#)
W	9/15	• *Visualizing categorical data* • **Finish Lab 2**: Plastic Waste • **Start Homework 2:** Water and sanitation **(Due 9/20)**	• **DataCamp 2 Due:** Intro. to Data Viz (4 chapters) • **Start Work on DataCamp 3 (Due 9/29)** • Intro. to the Tidyverse (4 chapters)

5
M
	9/20	***UNIT 3 - DATA WRANGLING***
		• *Tidy data*
		• *Grammar of data wrangling*
		• *Working with a single data frame*
		• Exercise 4: Great Lakes Wrangling
			• **Lab 2 and Homework 2 Due!**
			    • Reading: [JSS:: Tidy data](#)
			    • Reading: [R4DS::5 - Data transformation](#)

W	9/22	*Working with a single data frame (cont)* Exercise 4: Great Lakes Wrangling	**Continue DataCamp 3 (Due 9/29)** Reading: [R4DS::13 - Relational data](#)
6 M	9/27	Exercise 4: Great Lakes Wrangling *Working with multiple data frames* **Start Lab 3**	Reading: [R4DS::12 - Tidy data](#) Reading: [R4DS::15 - Factors](#)
W	9/29	*Tidy Data* **Finish Lab 3** **Start Homework 3: … (Due 10/6)**	**DataCamp 3 Due!** Intro. to the Tidyverse (All 4 chapters)

7 M	10/4	**_UNIT 1-3: REVIEW SESSION 1_** Survey results Unit 1 -3 Cheatsheet Group exercises	**Work on DataCamp 4 (Due 10/13)** ● _Importing data from flat files with utils_ ● _readr and data.table_ ● _Importing excel data_
W	10/6	**_UNIT 1-3: REVIEW SESSION 2_** Group exercises	**Lab 3 and Homework 3 Due!**
8 M	10/11	**_UNIT 4 - DATA TYPES_** _Data types; Data classes_ **_Continue Group Exercises!_**	Reading: <u>R4DS:: Ch 11 - Data Import</u> Reading: <u>R4DS:: Sec. 16.1-16.3 - Dates & Times</u>
W	10/13	**Start Lab 4: Data types** **Start Homework 4 (Due 10/20): Types and classes**	
9 M	10/18	**_UNIT 5 - IMPORTING DATA_** _Importing data_ **Start Lab 5:** Importing data	
W	10/20	_Recoding data_ **Finish Lab 5:** Importing data **Start Homework 5 (Due 10/27):** Putting it all together!	**Lab 4 and Homework 4 Due!** **DataCamp 4 Due** (3 Importing data chapters) ● _Importing data from flat files with utils_ ● _readr and data.table_ ● _Importing excel data_

10 M	10/25	**UNIT 6 - WRITING R FUNCTIONS** *Functions* **Start Lab 6:**	Reading: R4DS :: Chp 19 - Functions Reading: R4DS :: Chp 20 - Iteration
W	10/27	*Iteration* **Finish Lab 6:** **Start Homework 6 (Due 11/1)**	**Lab 5 and Homework 5 Due!** **Work on DataCamp 5 (Due 11/10)** ● Conditions & Control Flow ● Loops ● Functions ● The apply family
11 M	11/1	**Midterm REVIEW SESSION I** Branches and forks, merge conflicts Unit 4-6 Cheatsheet Group exercises with branches and forks	
W	11/3	**Midterm REVIEW SESSION II**	**Lab and Homework 6 Due!**
12 M	11/8	**MIDTERM EXAM on UNITS 1-6**	
W	11/10	**Overview of Final Projects** Split into teams Choose a dataset Elements of project **(Proposal DUE 11/17)**	**Homework:** Peer review of proposals **DataCamp 5 Due!** ● Conditions & Control Flow ● Loops ● Functions ● The apply family
13 M	11/15	**UNIT 7 - FITTING MODELS IN R** *Fitting and interpreting models* *Modeling nonlinear relationships*	Reading: IMS :: 7 - Linear regression w/ a single predictor

		Models with multiple predictors	Reading: [IMS :: 8 - Linear regression with multiple predictors](#)
W	11/17	**Lab 7:** **Homework 7: (Due 11/24)**	**Work on DataCamp 6 (Due 12/01)** ● Modeling data with the tidyverse (4 chapters)
14 **M**	11/22	***UNIT 8 - HYPOTHESIS TESTING*** *Quantifying uncertainty* *Bootstrapping* *Hypothesis testing* *Inference overview*	Reading: [IMS :: 11 - Hypothesis testing with randomization](#)
W	11/24	**Finish Lab 8:** **Homework 8 (Due 12/1):**	**Lab and Homework 7 Due!** **Work on DataCamp 6 (Due 12/01)** Modeling data with the tidyverse (4 chapters)
15 **M**	11/29	**Work on Final Projects**	
W	12/1	**Work on Final Projects**	**Lab and Homework 8 Due!** **DataCamp 6 Due!** Modeling data with the tidyverse (4 chapters)
16 **M**	12/6	**Final project presentations** (scheduled as a 2 hour exam slot)	
W	12/8	**Final Report Due End of Day Friday!**	