XXX: Course title here

Fall 2024

Table of contents

1 Course information

Resource	Link
Weekly meetings (online)	Day(s), Time, Timezone
Class website (public)	Link to website here
Questions, discussion, announcements (private)	Link to class repository for student issues here
Class Zoom link	Zoom link here

2 Course description

The primary goals of this course are (1) to teach fundamental skills of "data management," which are important regardless of which programming language you use, and (2) to develop a strong foundation in the R programming language. The course is designed for students who never thought they would become programmers and no prior experience with R is required. For goal (1), most statistics courses teach you how to analyze data that are ready for analysis. In real research projects, data management – the process of cleaning, manipulating, and integrating datasets in order to create analysis datasets – is often more challenging than conducting analyses. For goal (2), R is a free, open-source, object-oriented programming language. R is the most popular language for statistical analysis and one of the most popular languages for "data science" applications (e.g., web-scraping, interactive maps, network analysis). Students will become proficient in data management and R programming through weekly problem sets, which will be completed in groups.

2.1 Extended description

Data management consists of acquiring, investigating, cleaning, combining, and manipulating data. Most statistics courses teach you how to analyze data that are ready for analysis. In real research projects, cleaning the data and creating analysis datasets is often more time consuming than conducting analyses. This course teaches the fundamental data management and data manipulation skills necessary for creating analysis datasets.

The course will be taught using R, a free, open-source programming language. R has become the most popular language for statistical analysis, surpassing SPSS, Stata, and SAS. What differentiates R from these other languages is the thousands of open-source "libraries" created by R users. R is one of the most popular languages for "data science" because R libraries have been created for web-scraping, mapping, network analysis, etc. By learning R you can be confident that you know a programming language that can run any modeling technique

you might need and has amazing capabilities for data collection and data visualization. By learning fundamentals of R in this course, you will be "one step away" from web-scraping, network analysis, interactive maps, quantitative text analysis, or whatever other data science application you are interested in.

The data management and programming skills you learn in this course will transfer to other object-oriented programming languages (e.g., Python).

The course primarily use data and examples from social science research and designed to teach skills that are important for social science research more broadly and also for computational research within the humanities. **This class is welcome to students across the university**.

Recommended prerequisites (encourage, **but not required**)

- One prior introductory statistics course (e.g., undergraduate-level stats course)
- Proficiency in general computer skills is helpful, e.g., downloading files from the internet, renaming files, saving them to a folder of your choosing, finding this folder on your computer, etc.

2.2 Instructor

Instructor Name

- Pronouns:
- Office:
- Email:
- Office hours:
  - Zoom office hours: DAY, TIME, zoom link
  - And by appointment

2.3 Teaching assistants

Name

- Pronouns:
- Email:
- Office hours:
  - Zoom office hours:
  - And by appointment

Name

- Pronouns:
- Email:
- Office hours:
  - Zoom office hours:
  - And by appointment

3 Course learning goals

1. Understand fundamental concepts of object-oriented programming
   - What are the basic object types and how do they apply to statistical analysis?
   - What are object attributes and how do they apply to statistical analysis?
2. Become familiar with Base R approach to data manipulation and Tidyverse approach to data manipulation
3. Investigate data patterns
   - Sort datasets in ways that generate insights about data structure
   - Select specific observations and specific variables in order to identify data structure and to examine whether variables are created correctly
   - Create summary statistics of particular variables to diagnose errors in data
4. Create variables
   - Create variables that require calculations across columns
   - Create variables that require processing across rows
5. Visualize data
   - Create plots using the ggplot2 library
   - Customize plots through color palettes, labels, line shapes, etc.
6. Combine multiple datasets
   - Join (merge) datasets
   - Append (stack) datasets
7. Manipulate the organizational structure of datasets
   - Summarize and collapse observations by group
   - Reshape and "tidy" untidy data
8. Learn guidelines and practical strategies for ensuring data quality when cleaning data and creating analysis variables
9. Become proficient at using GitHub issues– the industry standard platform used by programmers to collaborate on projects– to ask questions about course material and to collaborate with your classmates

3.1 Course structure

Overview. Course structure consists of weekly **asynchronous course materials** and weekly **synchronous meetings**. Each week the class focuses on a particular topic (e.g., creating variables; writing functions). For each weekly topic, students will complete a problem set. Problem sets will be completed in groups and focus on practical application of concepts/skills from the topic of the week.

Asynchronous course materials. Asynchronous course materials will focus on the topic for that week (e.g., processing across rows). Course materials will consist of three types of resources:

1. Detailed lecture slides (HTML) with sample code
2. Pre-recorded video lecture of the instructor working through these slides
3. The ".qmd" file that created the HTML lecture slides.

    - The .qmd file will contain all "code chunks" and links to all data utilized in the lecture. Thus, students will "learn by doing" in that they will run *R* code on their own computer while they work through lecture materials on their own.

Synchronous meetings. Synchronous class meetings will be on Zoom. Attendance during the entire period is required, but students may ask the instructor/TAs for exceptions due to scheduling conflicts.

During synchronous class time, students will have the option of (A) attending live lectures from the instructor or (B) working through lecture materials/problem sets in Zoom breakout rooms in small groups (e.g., problem set groups) or on their own. For the first three weeks of class, students will not have the option of working in Zoom breakout rooms.

For students who decide to work in Zoom breakout rooms, you will use this time to work through course materials (e.g., lecture slides, video lectures) and/or the associated problem set as you see fit. The synchronous workshops are also a great time to ask questions about course material or practical applications. TAs will be moving from one breakout room to the next, providing help. Each group can develop their own approach to how they want to use the synchronous workshop time. Some groups may work relatively independently, while others may work collaboratively. Some groups may agree to work through all asynchronous lecture materials beforehand so they can devote all workshop time to making progress on the problem set. A recommendation is to work through the lecture material before getting started on the problem set.

4 Assignments and grading

Course grade will be based on the following components:

- Weekly problem sets (90% of total grade)
- Participation (10% of total grade)

4.1 Problem sets (90% of total grade)

Students will complete 10 problem sets (the last one due during finals week). Problem sets are due by [TIME] on [DAY], right before the start of class. In general, each problem set will give you practice using the skills and concepts introduced in course materials for that week. For example, after the lecture on joining (merging) datasets, the problem set for that week will require that students complete several different tasks involving merging data. Additionally, the weekly problem sets will require you to use data manipulation skills you learned in previous weeks. Link to problem set expectations and helpful resources HERE.

Problem set groups

- Except for the first problem set, students will complete problem sets in groups of 3. Students abroad are encouraged to form their own group to set aside time to work on the problem sets together.
- Students have the option of not being part of a problem set group.
- The instructional team will assist students in forming groups during the second synchronous class, and you will keep the same group throughout the quarter. However, each student will submit their own assignment. You are encouraged to work together and get help from your group. However, you must understand how to do the problem set on your own, rather than copying the solution developed by group members.
- Since you will be working together, it is understandable that answers to many questions will be the same as those of your group members. However, if there is compelling evidence that a student merely copied solutions from a classmate, this could be considered a violation of academic integrity. That student will receive a zero for the homework assignment.

A general strategy recommended for completing the problem sets is as follows: (1) after the lecture, do the reading associated with that lecture; (2) try doing the problem set on your own; (3) communicate with your group to work through the problem set, with a particular focus on areas group members find challenging.

Grading policies

- For students working in a problem set group, one submission from each problem set group will be chosen at random. The grade on that problem set submission will be the grade for all members of the group.
  - If a member of a problem set group has not submitted the problem set by the time the TAs conduct grading, that submission will be grades separately once it is submitted

  - The lowest problem-set grade will be dropped from the calculation of your final grade.

- Students who are not part of a problem set group will have their problem sets graded individually. A random subset of 4 or 5 problem sets will be graded. For students who work individually, the lowest problem set grade will not be dropped from calculation of final grade.
- Weekly required participation on github will be part of your problem set grade
- Policy on late assignments
  - Problem sets submitted after 11:59PM on [CLASS DATE] will lose one percentage point (e.g., max grade becomes 99% instead of 100%)
  - Starting at 12AM [THREE DAYS AFTER CLASS SCHEDULED TIME] morning, problem sets will lose an additional percentage point for each week-day it is not submitted
    * e.g., for a problem set submitted at 10AM on Monday, the max grade becomes 98%
    * e.g., for a problem set submitted at 10AM on Tuesday, the max grade becomes 97%
  - For late submissions due to an unexpected emergency, you will not lose points. Please contact the instructor and/or TAs and we will work it out together.

4.2 Participation (10 percent of total grade)

Broadly, students are expected to participate by being attentive, supportive of classmates, by asking questions, and by answering questions posed by classmates.

Practically speaking, the vast majority of your participation grade will depend on weekly participation on Github. Each week, students are required to post one communication on Github. This could be asking a question about the problem set, answering a question posed by a classmate, or a post describing something you learned while working through the week's material/problem set. If you post at least one communication on Github each week, you will earn an "A" for participation for the quarter.

In addition, students can work towards an 100% participation grade for the quarter by asking/answering questions during synchronous lecture (e.g., zoom chat) or by consistently being helpful/supportive to your classmates on Github.

4.3 Grading scale

Letter Grade	Percentage
A	93<=100%
A-	90<93%
B+	87<90%
B	83<87%
B-	80<83%
C+	77<80%
C	73<77%
C-	70<73%
D	60<70%
F	0<60%

5 Course Schedule

Below is an overview of course topics. Topics and the schedule are subject to change at the instructor's discretion. Topics may be cut if more time is needed to learn the most central topics. It is unlikely that additional topics will be added. The official course schedule will be posted on the course website, including weekly required reading and optional reading.

Week 1: Introduction to R

- Introduction to R and R data structures
- Execute R commands, understand R objects and data structures, use R functions
- Introduce atomic vectors, lists, and functions for investigating objects (e.g., length, type, str)

Week 2: Investigating data patterns in Base R

- Data investigation and manipulation using Base R
- Investigate R object type and structure, isolate elements using Base R subset operators and the `subset()` function, create new variables in Base R

Week 3: Enter the Tidyverse Part I: Pipes & Dplyr

- Data investigation and manipulation using tidyverse
- Select, filter, and sort data using `tidyverse` functions, chain functions together using pipes (`%>%`)

Week 4: Enter the Tidyverse Part II: variable creation

- Create new variables using `mutate()`
- Create new variables conditionally using `if_else()`, `recode()`, and `case_when()`

Week 5: Processing across rows

- Calculate aggregate statistics from multiple rows of data
- Group rows of data using `group_by()`, create aggregate statistics using `summarize()`

Week 6: Attributes and class

- Understand the class and attributes of R objects
- Investigate R object class and attributes, work with factor variables, label variables and values of a dataframe using the `labelled` package

Week 7: Strings and dates

- Work with strings and date/datetime objects
- Understand string basics, manipulate strings using `stringr` functions, work with dates and times using the `lubridate` package

Week 8: Create plots w/ ggplot

- Understand the layered grammar of graphics for visualizing data with ggplot2
- Make plots with the `ggplot` function (e.g., bar plots, scatter plots)
- Customize plots through color palettes, labels, legends, etc.

Week 9: Tidy data

- Understand tidy data structure and reshaping data
- Define tidy data and how to reshape untidy data into tidy form, reshape data from wide to long using `pivot_longer()`, reshape data from long to wide using `pivot_wider()`, handle missing values during reshaping

Week 10: Joining data

- Combine data from multiple datasets using joins
- Merge datasets using mutating joins, check the quality of merge using filtering joins, append datasets by stacking rows