COMP/STAT 112: Introduction to Data Science

Macalester College
Spring 2023
Section 01: TTh 8-9:30am, OLRI 241

Course Overview

Throughout this course, you will learn practical, communicative, and collaborative skills needed to successfully conduct a data science analysis. We will actively learn and reinforce concepts using R, RStudio, and R Markdown. The primary goal of this course is to get you comfortable carrying out the data science pipeline:

- **Formulate a research question**
- **Collect/scrape data**
- **Wrangle data**
- **Model**
- **Visualize**

Your Instructor

James Normington (he/him/his)

Office: Olin-Rice 230

Email: jnorming@macalester.edu

Office Hours

- TTh: 3-4pm
- Th: 12:15pm - 1:15pm
- Location: Olin-Rice 230 or by Zoom (with prior scheduling)
- By appointment – please don't hesitate to use this option!

Learning Objectives

By the end of this course, you will be able to:

- Effectively communicate the details of data and your analysis
- Collaborate on a data science analysis
- Use R, RStudio, and R Markdown to perform your analysis and output documents
- Import data and do basic cleaning steps
- Wrangle text using regular expressions
- Wrangle data using `select`, `mutate`, `filter`, `arrange`, `summarize`, and `group_by`
- Reshape data from wide to long (or vice versa)
- Join datasets
- Work with categorical variables as factors
- Understand and create effective visualizations for univariate, bivariate, multivariate, and spatial relationships
- Perform exploratory data analysis to learn about a dataset when you first encounter it
- Fit statistical models

Course Materials

The only required material in this course is access to a computer. If you do not have a functional computer, or reliable internet access, don't worry! Just fill out this form; ITS and Financial Aid will work with you to get you what you need.

Every reading will be made freely available online.

Our course notes can be found at:

https://jamesnormington.github.io/112_spring_2023/

In addition, you will need to have R and RStudio

We will install this on the first day of class, but here are the links to download/update R and RStudio

Topics Schedule

We will tentatively be covering topics as follows:

Week 1: Introductions

Weeks 2-3: Data visualization

Weeks 4-5: Data wrangling

Weeks 6: Mini-project

Week 7: Exploratory data analysis

Week 8: Midterm exam and brainstorming activity

Week 9: Data import, cleaning, and regular expressions

Week 10: Final Project work time

Week 11: Networks, interactive graphs, and R Shiny

Weeks 12-14: Final Project work time

Grading Scheme

The grading system in this course is designed to directly ***align your grades with*** our course ***learning*** objectives, and to allow space to ***make and learn from mistakes*** along the way. More specifically, grades will be based on a combination of three components:

(1) Final Project (25%)
(2) Midterm Exam (25%)
(3) Ongoing Engagement: Demonstrated engagement in course Assignments, Tidy Tuesday, and Iterative Viz (50%)

Major Assessments

Midterm

To reinforce your understanding of some practical data science concepts, we will have a midterm exam on Tuesday, March 14th. The midterm will have two parts:

- Part 1: Conceptual (20 min)
  - No notes
- Part 2: Application (40 min)
  - Open note + computer resources

If you are unable to take the midterm during the scheduled time period, please let me know at least one week in advance so I can work with you to make alternate arrangements.

Final Project.

You will work on a data analysis project starting in the middle of the semester. This project will ask you to use the skills practiced in class to explore, analyze, and interpret summaries and models fit to a dataset of your choosing. There will be multiple project checkpoints throughout the semester, allowing ample opportunities for instructor feedback. This work will culminate in a **final presentation**, submitted via recording.

There will be no final exam.

Ongoing Engagement

50% of your grade will be determined by your Ongoing Engagement, which will be measured by your submissions for the following:

(1) **Assignments**. Assignments will be continuations of in-class problems, and will be due about a week after learning the relevant material.
(2) **Tidy Tuesday**. Tidy Tuesdays are weekly data science projects put on by folks in the R/RStudio community. They will be used to further hone your abilities to form a data science question, perform data wrangling, and construct visualizations.
(3) **Iterative Viz**. Submit one visualization created from a Tidy Tuesday submission. You'll submit a visualization, I will provide feedback, you'll incorporate my comments in your next submission, and we will iterate a total of three times.

To achieve an A in the Ongoing Engagement component of STAT 112, you'll need to do *each* of the following:

(1) Submit at least eight (out of 10) Assignments in "good faith". The preceptors will offer qualitative (i.e., not a letter or numeric) grade. If a submission is incomplete or inadequately attempted, it will not count as a "good faith" submission.
(2) Earn a "3" grade on at least three Tidy Tuesday submissions.
(3) Attempt at least one Tidy Tuesday in the first five weeks.
(4) Earn a "3" grade on the last Iterative Viz submission (Iterative Viz 2)

If any of these components are missed, your final grade will be reduced by a proportionate amount.

Getting help in COMP/STAT 112

Preceptors

We have seven preceptors that will be helping out with COMP/STAT 112 this semester! They will be providing feedback on your homeworks and holding office hours.

- Sanya Bains (she/her): sbains@macalester.edu
- Avianna (Hien Anh) Bui (she/her): hbui2@macalester.edu
- Erin Franke (she/her): efranke@macalester.edu
- Ting Huang (he/him): thuang@macalester.edu
- Tanya Nangpal: tnangpal@macalester.edu
- Maria Sanchez Linare: msanche2@macalester.edu
- Samina Stack: sstack@macalester.edu

Academic Integrity

As a globally- and community-oriented institution, Macalester College expects respectful exchange of ideas. Students are expected to be familiar with the college standards on academic integrity (**website**). I encourage you to work with your classmates to discuss material and ideas for your assignments, but in order for you to receive individualized feedback on your own learning, ***all submitted work (including code!) must be written in your own words.***

When you present someone else's work as your own, you are missing a learning opportunity and cheapening the value of your education. If you are tempted to plagiarize someone's work, please consider requesting an extension instead. See below my policy on extensions.

Other Course Policies

Extensions. Life happens! There are circumstances outside of the classroom that make it difficult to focus on schoolwork. These circumstances are often entirely personal; thus, I will grant most extensions, without the need to specify a reason, when requested in advance of the corresponding deadline. My intent is that students can focus on their physical, mental, personal, and interpersonal wellness so there are fewer obstacles to learning. However, I reserve the right to deny an extension if I feel the student is abusing this policy.

Final grades are due May. 11th; any unsubmitted work, regardless of circumstances, will count as 0 unless waived by the instructor. If your circumstances make it likely that you can't submit a sizeable portion of classwork by then, you should talk with me about alternative arrangements. For deadlines on adding/dropping courses, grading options, and withdrawal, see the Spring 2023 academic calendar here.

Attendance. Attendance is expected, but not mandatory. If you are to miss class, please try to inform me in advance. Your reason for missing class does not need to be included in your message. I reserve the right to adjust the Engagement component of a student's final grade if the student makes a habit of missing class without informing me in advance, or if I suspect a student is abusing this policy.

Late Submissions. Homework submitted late, but within 7 days of the due date, will be assessed a 50% penalty. Homework submitted later than 7 days of the due date will be assessed a 100% penalty.

My Teaching Philosophy

Active learning. Statistics is an active discipline. The collection, visualization, analysis, and interpretation of data are all active steps in the statistical modeling process. Because of this, a large proportion of class time will be spent actively doing statistics problems with practical, real-world data.

Statistical thinking over statistical procedure. The rote memorization of manual statistical procedures (i) is unnecessary with modern computing, (ii) requires mathematical prerequisites beyond the scope of this course, and (iii) presents statistical techniques as disparate, non-intuitive procedures. Statistical modeling is best learned through intuitively solving real problems, and leaving the tedium of the procedure up to computers (or, for the inclined, preparation for graduate school exams!).

Diversity, equity, and inclusion. Statistics is, in a word, the study of variability. The best statistical models acknowledge and account for all sources of variability, recognizing the diverse and multivariate world we live in. There will be absolutely no discrimination based on race, ethnicity, gender identity, sexual orientation, or any other identities, life experiences, or opinions. If at any point you feel discriminated against, please either reach out to me or find other support at Macalester.

Statistical Software

Computing is an integral part of modern statistics and will thus be a main focus of our course. We will use the R programming language. RStudio, a graphical user interface (GUI) for R, will facilitate our use of R and the creation of R Markdown reports that you will use for submitting assignments. R and RStudio are *free* to use and download.

No computing background is necessary to succeed in this course. When you learn a new computing language, you will stumble at the beginning. That is normal! The key is to learn how to overcome those stumbling blocks (or coding errors). Read the error message. Double check the spelling. Consult Google. Post the code and error message on Slack.

Learning to use R/RStudio will take time and effort, but by the end of this semester you'll be leaving this course with a *very marketable and useful skill* that will serve you well in future courses, jobs, and beyond!

Advice for Success in COMP/STAT 112

1. **Actively engage with the course content.** Pay attention in lectures, actively participate during in-class activities, understanding and incorporate my feedback into your Final Project.
2. **Ask when you have questions.** Ask during lecture (interrupt me at any time; I'm serious!), come to office hours, visit the preceptors and tutors, ask your classmates, post on Slack. Our best learning comes in the vulnerable moments when we admit we don't understand.

COVID-19 Protocols and Expectations

We all deserve a safe learning environment. Please stay current with the latest COVID-19 guidelines: Shared Community Commitment (Mac Stays Safer 3.0). I will abide with these guidelines very closely.

Some key notes, and my specific expectations:

Vaccination Policy. Macalester requires all employees and students to have received a COVID-19 vaccine and booster, or to have an approved exemption in place. This is not negotiable. If you are eligible for a 2nd booster, I highly recommend it; see more information at the CDC website.

Mask Policy. There will be no requirement to wear a mask in class. However, I will wear a mask in office hours and ask that you do too. During in-class activities and Final Project working sessions, groups should feel free to come up with their own policies (so long as they fit within Macalester's and the State of Minnesota's policies) If there is a disagreement which is harming the group's ability to make progress, please see me. For more information, see the Well-Fitted Mask Policy.

Be Safe, Be Respectful. My primary concern is your well-being. If I or a classmate respectfully asks you to wear a mask, cover your sneeze, wash your hands, etc. – please do so. Your classmates may be immunocompromised, or live/work/relate with someone who is. They may also just not want to get sick! Every student has the right to feel safe attending my lectures in person. We each have varying levels of "COVID cautiousness", so please use both common sense and courtesy for others in my classroom and on campus.

Test Frequently. If you feel symptoms associated with COVID-19, **please** get tested. Please see Shared Community Commitment (Mac Stays Safer 3.0) for information on free, accessible testing.

Stay Home when Sick. If you have COVID-19, or any other contagious disease, **please** don't come to class. Email me and we will work out alternate arrangements.

Health and Wellbeing

Your health and well-being should be prioritized over my course. A student cannot effectively learn statistics if they are under physical, mental, emotional, societal, or financial stress. During class time, eat when you are hungry, drink water when you are thirsty, use the restroom, and step away/out if you are upset or need some air. Please do what is necessary so long as it does not impede your or others' ability to be mentally and emotionally present in the course.

Physical Health. Physical exercise is associated with better grades, better physical & mental health, and an improved quality of life. I highly recommend using the Leonard Center to stay physically active. Participation in intramural sports is another great way to exercise, reduce stress, and develop a sense of community on campus.

Mental Health. You can find a list of mental health resources here.

Financial Health. If you have had an unanticipated financial emergency, please consider using the Emergency Aid Program.

Beyond these resources, if you are having difficulties maintaining your well-being, please contact me. I am invested in your full learning journey, which cannot thrive unless you are OK.

The Environment You Deserve

I am committed to helping you learn and succeed in this course and in your time at Macalester. Enabling this means facilitating environments that support you in the ways that you need.

Respect. There will be *no tolerance for discrimination* based on race, ethnicity, gender identity, religion, sexual orientation, disability, and other identities or life experiences, in and out of the classroom. If you see any discrimination whatsoever, please contact me or find other support at Macalester, and be ensured the matter will be dealt with seriously.

Sensitive Topics. Statistics is used to inform policy, research, and understanding across all aspects of our society. The examples and data we use in class will reflect this. Thus, at times our work might touch on sensitive topics. I will try to give warnings for topics that I perceive may be sensitive. Please let me know if, for any reason, you feel that you are unable to participate in the planned activities due to the topic and we can discuss alternatives.

Accommodations. I am committed to creating an accessible and inclusive classroom for all students, including those with disabilities. If you are in need of any accommodations, please contact Disability Services (visit their website, email disabilityservices@macalester.edu, or call 651-696-6275) to schedule an appointment and discuss your needs. Once you've met with Disability Services, please then set a time to meet with me to discuss your accommodation plan for this course. It is important to *arrange this meeting as early in the semester as possible* (ideally within the first week), in order to ensure that your accommodations can be implemented early on. It is your responsibility to make sure you are registered with Disability Services. If you wait until later in the course, I may not be able to accommodate you retroactively.

Religious Observance. Students may wish to take part in religious observances that occur during this semester. I've done my best to schedule deadlines around major holidays, but if you have a religious observance/practice that conflicts with a scheduled class meeting or assignment deadline, *please contact me at least one week in advance* so we can discuss appropriate accommodations.

Title IX. If you or anyone you know has experienced harassment or discrimination on the basis of sex or gender, know that you are not alone. Macalester provides staff and resources to help and support you. More information is available on the Title IX website.

Please be aware that *all Macalester faculty and preceptors are mandatory reporters*, which means that if we become aware of incidents or allegations of sexual misconduct, we are required to share the matter with the Title IX Coordinator. Although we have to make that notification, you control how your case is handled, including whether or not you wish to pursue a formal complaint. If you would like to speak to someone *confidentially*, contact the Hamre Center (651-696-6275), Chaplain staff (651-696-6298), or other local and national resources listed here.

Q2 General Education Requirement. Successful completion of this course will satisfy the ***Q2 Quantitative Thinking*** general education requirement. You can find details on this requirement here.