STA 309: Building, Managing and Exploring Data Sets in Analytics

Spring 2022

Table of Contents

Bulletin Description

Techniques for constructing, downloading, cleaning, combining, extracting and manipulating data sets to prepare them for statistical analysis and visualization. Emphasis on programming languages used in data analytics and structured query language.

Back to Table of Contents

Prerequisites

STA/POL 308; STA 363 or ISA 291 or POL 306; MTH 133 or ISA/STA 250.

Section and Instructor Information

Section A: Mon, Wed, Fri 1:15 pm - 2:10 pm

- Instructor: **Dr. John Bailer**
- Email: *baileraj@miamioh.edu*
- Office hours: 10:00-11:15 MWF (Zoom) - NOT 28 Jan, 11 Feb, 18 Feb
  - or by appointment
- Twitter: @john_bailer / @statsandstories
- Home Page: http://www.users.miamioh.edu/baileraj/
- GitHub: https://github.com/baileraj

Section B: Mon, Wed, Fri 2:50 pm - 3:45pm

- Instructor: **Dr. Tom Fisher**
- Email: *fishert4@miamioh.edu*
- Office hours: MW 3:50-4:20 (in person, after class)
  TuTh 2:00-3:40 (via Zoom, [schedule appointment](#))
- Home Page: https://tjfisher19.github.io/

*** Note: drinks/food are NOT permitted in the SCALE-UP classrooms in Upham Hall.

Detailed Class Description

In many ways, STA 309, is the culminating course covering statistical methods for the Bachelor of Arts in Data Analytics. This course is the final course in the "core" of the degree. Upon successful completion of STA 309, should should be able to

- Write code to extract, combine, transform and summarize data from multiple sources.
- Develop code to clean and organize unstructured data into a structured format.
- Manage, reorganize and reshape data for multiple types of analysis and visualization.
- Work with data from different sources (*e.g.*, unstructured data from websites)
- Implement the practice of key elements of analytics practice introduced in STA 308 and 363.
  - Code sharing and version control.
  - Reproducible analysis.

Anticipated FAQ

Is STA 309 simply a continuation of STA 308?

The short answer is NO. STA 309 is not simply a continuation of STA 308. However, we will be using most of the tools covered in that class. In this course we will build off the knowledge you have gained in all the prerequisite courses. We will be using material from STA 308, STA 363 and MTH 133; among others. While the nuts and bolts of coding and scripting was the focus of STA 308, in STA 309 we focus on building and investigating a data set that is a key in an analytics investigation.

Why is MTH 133 a prerequisite for this course?

MTH 133 (or equivalent) is a requirement for this course because it is important you have a full understanding of elementary relational algebra. In less technical terms, the ideas of *intersection* and *union* are imperative when it comes to working with databases and multiple datasets. These abstract concepts from this mathematics course are fundamental to working with multiple datasets.

What language will be used in STA 309?

We will primarily be working with R in this course. We may use python as well but the course will primarily be taught using R.

Back to

Class Topics

The class is being structured into **three primary modules** or units

<u>Foundations of Data Visualization</u> (5+ weeks)

- *Review* of STA 308 materials (*i.e.*, data input & processing, `dplyr` verbs).
- Working in the `ggplot2` framework.
- Principles of Data Visualization (*e.g.*, Gestalt principles, Tufte & Cleveland principles/recommendations).
- Building *advanced* static visualizations (*e.g.*, choropleth maps, multiple panel plots [dashboards]).
- Brief introduction to interactive visualizations (*plotly* and animation).
- Understanding the need to reshape data for visualization.

<u>Advanced Data Wrangling</u> (3-4 weeks)

- Pivoting and reshaping data.
- Extracting variables and data from strings of text.
  - Rudimentary text analysis and visualization.
- Advanced data formats - databases, API calls.
- How to structure data for predictive modeling applications.

<u>Data and Methods for Analytics Applications</u> (3-4 weeks)

- Introduction to data-concepts for predictive modeling.
  - Testing and validation vs training and model building.
  - Cross validation algorithms: LOOCV, *k*-fold CV, repeated CV.
- Predictive modeling techniques.
  - Linear modeling based methods.
  - Tree based methods.
- The data analytics story: from data input, to exploratory visualization, modeling, and story-telling.

Caution / Advice:

- Many of the concepts in this course will build up over a semester. If you do not understand a concept early in the semester, then you will find you have a poor foundation for learning more complicated concepts.
- Keep current and do not fall behind.
  - If you find yourself struggling, seek help!
- Programming / analysis tasks usually take longer than you estimate (similar to home improvement projects). If you think it will take 1 hour to complete a project, plan on 5 (may be an underestimate).

Back to

Class Format

Expect this class to be very interactive and require regular student participation. You learn coding by doing it! The code and material will both review and build upon itself. Throughout the semester you can expect examples and problems where we need to revisit tools from earlier in the semester (or even STA 308 or 363).

Back to

Quizzes, In-class activities and Small Group Activities

Expect regular in-class assignments, short quizzes and short-turnaround homeworks in this course. We will use these assignments/quizzes to assess the classes learning and adjust the curriculum, or its pace, if necessary. These assignments will be such that each individual assignment is low stakes, but the accumulation will have a meaningful impact on your grade. In many of these activities/assignments you often will be encouraged to work together in groups as, not only will it help you learn from each other, it is reminiscent of how modern analytics is performed in industry.

Back to

Homeworks

More comprehensive homeworks will be assigned throughout the semester (anticipating *upwards of 6* assignments) as an effort to assess your overall comprehension of the material. The first assignment will be given in the first few weeks and there is an ungraded

Homework #0. Graded homework will be due about every week or two weeks. Late homework will **not** be accepted unless granted **prior** permission from the instructor.

Back to

Mid Semester & Final Evaluations

Two mid-semester evaluations will be given along with a Final evaluation. You can think of these as corresponding to exams but the format may not necessarily be in the form of an "exam." This course is meant to be very applied, so the format is to be determined but should be expected to be a mix of traditional exam with a take-home portion, perhaps even in the form of a project.

Back to

Class Materials

You will need access to a computer to complete this course. Any of the major operating systems (Windows, Mac or Linux) is acceptable. You will need a computer in nearly every class meeting to experiment with code in class. A small notebook may be useful to take notes while attending class.

Back to

Course Textbook

There is no required textbook for this course but we will be using material from several sources. The *R for Data Science* book by Wickham and Grolemund and the *R Graphics Cookbook* by Chen may be useful online resources

Back to

Backup

No allowance will be made for computer or network failures. Be sure to back up your work frequently, and do not wait until the last minute to complete the assignment. To back up a file, save it first to one computer drive such as your UDS / M drive / Google Drive or

the hard drive of your home computer, and then save it again to a portable drive. Do not trust that a thumb drive is sufficient. Email to yourself is a simple option. Use Google Drive or campus drives.

Software

In this course we will primarily be utilizing R and RStudio. You can download both pieces of software for free from:

- http://cran.r-project.org/   [R language - select local download mirror]

- http://www.rstudio.com/   [Integrated Development Environment - great choice for using R!]

Countless resources are available for help in using R - Google "R Tutorial" and check out Youtube for examples.

We may also be using Python and may use other tools as well. You can download the Anaconda Navigator that includes Python and IDEs such as Spyder from https://www.anaconda.com/ .

We recommend you complete Homework #0 as soon as possible.

Back to Table of Contents

Discussion Boards

The discussion boards on Canvas are available for you to post questions, discussions and to help your fellow classmates. Please note that the posting of solutions is in violation of the academic integrity code. Those postings will be deleted and students may be prosecuted for a violation of academic integrity standards. The boards are activated as a way to communicate with your classmates.

Back to Table of Contents

Notable Dates on Calendar

Date(s)	Topic
Mon, Jan 24	Classes begin

Thur, Feb 10	Last day to drop without grade
Late February	First midterm evaluation (date & structure TBD)
Fri, Mar 18	Midterm grades available on Banner
Mar 21 - 25	**Spring Break**
Mon, April 4	Last day to drop with a W
Early-to-mid April	Second midterm evaluation (date & structure TBD)
Fri, May 6	Last day of classes
May 09 - 13	Final Exam week (Assessment TBD)

Back to

Student & Instructor Expectations:

You, as a student, should try your best to:

- Attend every class.

- Check the Canvas class site at least once per day during the workweek.

- Check your Miami email every 24 hours (more often if you have sent an email to the instructor or a classmate)

- Read Announcements that are sent via Canvas. Check Announcements on Canvas before emailing your instructor, as the information may have already been posted.

- Post questions you have about the course in the Discussion Forums. If you can answer another student's question, without violating the academic integrity policy, please feel free to do so.

- Actively participate in all class sessions, discussions, assignments, and other activities required in this course.

- Post your ideas and assignments; react to others' ideas, and exchange thoughts with fellow students and the instructor.

- Complete any readings in a timely manner so you can get help if needed.

- Submit assignments at the very least by their due date/time. Start assignments early as a way to guarantee this will occur.

- Spend at least 6-9 hours a week outside class, studying, reading from supplemental sources, and completing assignments.

- Submit work that demonstrates a clear understanding of the material.

- Ask questions if you don't understand! Ask questions in class AND in office hours.

- Keep an open mind regarding the material and the opinions of others.

- Notify the instructor, in a timely manner (within 24 hours), if you have any problems.

The instructor will try his/her best to:

- Check the Canvas class site every day to monitor progress.

- Check email at least twice a day, Monday through Friday.

- Respond to all emails within 24 hours (except on weekends).

- Post all grades within 7 days of assignment due dates, except for homeworks which may require up to 2 weeks.

- Make every effort to meet with students who request a meeting.

Back to

Academic Integrity

[Academic Integrity](#) is at the heart of the mission and values of Miami University and is an expectation of all students. Maintaining academic integrity is a reflection of your character and underpins your learning and understanding of the course material.

According to the Miami University Student Handbook, "Academic dishonesty is defined as **any activity that compromises the academic integrity of the institution** or **subverts the educational process**."

Examples of academic dishonesty include but are not limited to:

- Referencing outside sources or course notes while taking an exam.

- Referencing non-permitted websites while taking a quiz

- Collaborating with another person in any way while taking a quiz or exam

- Copying someone else's solutions to a problem and submitting it as your own work

- Posting class material and assignments to sites not affiliated with Miami University

- Plagiarism - Using another person's ideas without citation and calling them your own. This can occur if you search online for coding help, use code that is found in your solution AND you don't cite the web page / source.

All Miami University policies concerning academic integrity apply to this course. See the [Student Guide to Academic Integrity](#) for details.
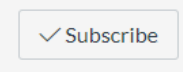
Suspected Dishonesty: Any suspected instances of academic dishonesty will be handled under Miami University's [Academic Integrity policy](#). It is a student's responsibility to read this policy. Lack of knowledge or understanding of the appropriate academic conduct is not an excuse for committing academic dishonesty.

Students who are found responsible for committing academic dishonesty will receive a sanction that ranges from a zero on the assignment to an F in the course, which could contain the AD transcript notation. Students who are found responsible for committing two acts of dishonesty (academic or Code of Student Conduct section 102 (Dishonesty)) automatically will be suspended from Miami University.

Students should act honestly in the completion of all course assignments, and ask the instructor for clarification on any assignment guidelines that are not clear.

Back to

Communication Guidelines & Netiquette

Announcements: Please subscribe to the Course Announcements so you can be automatically notified about important information

through your email. To subscribe for auto-email notifications, click on the subscribe button. , it will turn green with

"subscribed", .

Contacting Your Instructor: Email is the best way to reach your instructor. However, if you would prefer to have a conversation, you may speak to your instructor during office hours or send your instructor an email to set up a time to meet. Given today's ease of communication, it is your responsibility to contact your instructor within 24 hours if you are having any problems. While instructors will attempt to answer emails quickly, it may take up to 24 hours (or longer on weekends) for you to receive a response. Many faculty receive over 50 emails a day. We ask that you put "STA 309: *summary of your email*" in the subject line of your email to help sort; we will try to do the same with announcements and emails.

Communication Guidelines: Email is the official mode of communication for the University. You are responsible for any communication that is sent to your Miami email account, so please be sure to check your account frequently. If you have your email forwarded to your preferred account, please be sure you have enough server space for your Miami emails and please check to see that your server will accept Miami emails.

You will be asked to communicate with other students in this class via email. You may access email tools on the Canvas class site. All group members are responsible for initiating contact with one another as soon as the assignment is opened. Failing to make contact with your group members is not an excuse to miss completing an assignment.

Netiquette:* Diversity has many manifestations, including diversity of thought, opinion, and values. We encourage **all learners to be polite** and **respectful** of that diversity and to refrain from inappropriate or offensive commentary. If inappropriate or offensive content is either emailed or posted on the class site, we may recommend college disciplinary action. Students guilty of academic

misconduct, either directly or indirectly through participation or assistance, are subject to disciplinary action through the regular procedures of Miami University. Learners as well as faculty should be guided by common sense and basic etiquette. Criticism should be presented in a positive light. The following are good guidelines to follow:

- Be professional in your correspondence with the instructors and with other students.

- Never post harassing, threatening, or embarrassing comments.

- Never post content that is harmful, abusive; racially, ethnically, or religiously offensive; vulgar; sexually explicit; or otherwise potentially offensive.

- Never post, transmit, promote, or distribute content that is known to be illegal.

- If you disagree with someone, respectfully respond to the subject, not the person.

- Treat others as you would like to be treated.

- It can be difficult to communicate emotion in email, so use emoticons such as 😊 or ;-) when appropriate.

- Be timely in your communication with others.

- Be as brief and succinct as possible.

- Include a descriptive subject line in all emails.

- Use proper spelling and grammar.

- Cite sources appropriately including URLs of sites you use for code inspiration.

Remember that "tone" can usually be detected accurately in verbal communication, but often can be misunderstood in electronic communication. Because of this phenomenon, we encourage you to err on the side of politeness.

*adapted from Regents Online Campus Collaborative

For more information on netiquette check Netiquette Rules for Electronic Communications.

Back to Table of Contents

ADA & Students with Disabilities

Miami University is committed to ensuring equal access to students with disabilities. Miami's Office of Student Disability Services (SDS) assists students with determining eligibility for services and accommodation planning. Miami's AccessMU provides resources and guidance toward equal opportunity for all individuals. Refer to Miami University's [Accessible Technology Policy](#) for definitions and additional information.

Students who are entitled to disability-related academic adjustments, auxiliary aids, etc., must register with SDS to receive accommodations in university courses. Please understand that formal communication from SDS must be presented prior to the coordination of accommodations for this course. For more information, see [Student Disability Services](#) and/or [the Rinella Learning Center](#). Students may also contact SDS at (513) 529-1541 or via email at sds@miamioh.edu.

If you have a disability, please contact me, and I will be glad to make any necessary accommodations.

Back to [Table of Contents](#)

Diversity & Discrimination

All Miami University policies concerning diversity and equal opportunity will be upheld in this class.

> Miami University is a community dedicated to intellectual engagement. Our campuses consist of students, faculty, and staff from a variety of backgrounds and cultures. By living, working, studying, and teaching, we bring our unique viewpoints and life experiences together for the benefit of all. This inclusive learning environment, based upon an atmosphere of mutual respect and positive engagement, invites all campus citizens to explore how they think about knowledge, about themselves, and about how they see themselves in relation to others. Our intellectual and social development and daily educational interactions, whether co-curricular or classroom related, are greatly enriched by our acceptance of one another as members of the Miami University community. Through valuing our own diversity, and the diversity of others, we seek to learn from one another, foster a sense of shared experience, and commit to making the university the intellectual home for us all.

Please see the [General Bulletin](#) for more information.

Miami University is committed to providing equal opportunity and an educational and work environment free from discrimination on the basis of sex, race, color, religion, national origin, disability, age, sexual orientation, gender identity, military status, or veteran status. Miami shall adhere to all applicable state and federal equal opportunity/affirmative action statutes and regulations.

Please see the Miami University policies regarding discrimination and harassment for more information.

Back to Table of Contents

COVID-19 Policies and Considerations

Students, faculty or staff should **not come to campus when ill** or under orders from the Butler County General Health District to isolate because of a diagnosis or quarantine because of close contact with someone who tested positive for COVID-19. However, students are ultimately responsible for material covered in class, regardless of whether the student is absent or present. Instructors are not expected to create a facsimile of in-class instruction, but should identify reasonable ways to accommodate the absence and may wish to make some or all of their office hours remotely accessible to assist in this accommodation. If your absence is of significant duration or severity, as your instructor I will advise you about other options that might be available including assigning an incomplete grade or requesting a medical withdrawal.

Face masks are required during **all class meetings** to promote the health and safety of all university members. There may be university approved exceptions to this requirement. Students who cannot wear a facial covering due to medical or disability-related reasons should contact the Miller Center for Student Disability Services at sds@miamioh.edu or Regional Student Disability Services at regionalsds@miamioh.edu.

If a student comes to class without a face mask or refuses to maintain physical distancing, I will first ask the student to comply (e.g. put on a face mask). If the student refuses, I will ask the student to leave the classroom and inform the student that the class will not proceed until the student either complies or leaves. If the student continues to refuse, I will dismiss the class and immediately report the student to the Office of Community Standards

Final Grade Calculation

At the conclusion of the semester, final grades will be compiled using the following point distribution:

Item	Contribution to Final Grade

In-class Assignments, Quizzes & Small-scale Homeworks
• Expect regularly
  (at least once a week but should expect more)
• Quizzes would be short
• An in-class assignment may be interactive with peers
• Small homeworks with quick turnarounds
  ○ *e.g.*, due before the next class
	10%

Comprehensive Homeworks • More in-depth application of covered material • Some may be challenging	40%
Mid Semester Evaluations	30%
Final Evaluation	20%

Grading Scale

A+ = [97, 100]	A = [93, 97)	A- = [90, 93)
B+ = [87, 90)	B = [83, 87)	B- = [80, 83)
C+ = [77, 80)	C = [73, 77)	C- = [70, 73)

D+ = [67, 70)   D = [63, 67)   D- = [60, 63)

F = [0, 60)

Back to