DS 5610: Exploratory Data Analysis

Data Science Institute

Michael E. Shepherd, Ph.D.

Vanderbilt University

Course Description

This course will teach students how to explore, summarize, and graph data (big and small). Topics include principles of perception, how to display data, scatterplots, histograms, boxplots, bar charts, dynamite plots, proper data summaries, dimensionality reduction, multidimensional scaling, and unsupervised clustering algorithms, such as principal component analysis, k-means clustering, and nearest neighbor algorithms.

Location and Time

- Class: Tuesdays and Thursdays from 3:15-4:30pm. Sony Building 2001-A.
- Professor Office Hours:
    - In-person: Tuesdays from 10:00am-11:00am; 331 Calhoun Hall
    - Virtually: Wednesdays and Fridays 10:00am-11:30am/by appointment (Zoom)

Instructors

Position	Person	Email	Office Hours
Professor	Michael Shepherd	Michael.e.shepherd@vanderbilt.edu	T 10:00-11:00am
TA	Peter Busienei	Peter.k.busienei@vanderbilt.edu	MW 10:00-11:00am
TA	Tonnar Castellano	Tonnar.castellano@vanderbilt.edu	W 6:00-7:00pm; F 1-2:00pm
TA	Ty Painter	Ty.d.painter@vanderbilt.edu	TH 4:30-6:30pm
TA	Hongyu Dai	hongyu.dai@vanderbilt.edu	F 10:00am-12:00pm

COVID-19 Protocols

Per CDC recommendations, Vanderbilt requirements, AND the realities of our inability to social distance in the classroom, **MASKS ARE REQUIRED** to attend this class. We cannot stay six feet apart in the classroom simply as a matter of fact. As a result, during the class session we will wear masks until the guidance from the university changes.

Textbook and Software

- This course requires *R for Data Science* by Hadley Wickham & Garret Grolemund.
- Many additional coding resources will be posted each week and are available online
- Readings beyond the text will be posted within modules on the course website.
- Computing: we use R and R Studio (and only those programs) for data analyses in this course. We will also use Github throughout the semester.

Learning Goals and Assessment

Course Goal:
Students will learn how to organize, visualize, analyze, and interpret data & statistical information using the R programming language.

Learning objectives (LOs):

- LO1 Learn how to keep files both logically and remotely organized while using GitHub
- LO2 Gain proficiency in data wrangling, visualization, analysis in R
- LO3 Increase data literacy, understanding data types and structures, and data communication
- LO4 Learn when do use the appropriate graphic, algorithm, or other EDA tool to summarize data and report trends.

Assessment:
This class is divided into three parts:

Part 1 (see LOs 1-2) will build students' skill-sets to be able to read data into R. Students will then learn how to manipulate, clean and understand this data and save their changes to an assignment or data via GitHub. Students will then learn how to visualize data using a wide variety of graphical concepts and tools with a careful eye towards graphic foundations and how to avoid misleading graphics.

Part 2 (see LOs 3-4) will build on part 1 and focus on uncovering relevant patterns in the data using a variety of modeling approaches. Students will begin, through practice, to defend their choices in terms of graphical presentation and data summarization. This will increase their data literacy. In addition, students will learn how to work with different "types" of data, such as network and text data.

Part 3 (synthesis of all LOs) will introduce students to advance R skills and puts the parts together and has students explore different data from start to finish, with an emphasis on learning how to communicate their analyses to an audience. Students will also be introduced to advanced topics through a brief machine learning introduction.

Grading

Final grades will be assigned based on the following weights:

- 30% Assignments and Attendance /Participation
- 30% Mid-term exam
- 40% Final Project

Assignment and Participation 30%

Attendance & Participation: Throughout the class, I keep a folder of notes on who is attending class and asking questions, participating in quick fires, prepared to ask questions to guests, and generally showing high vs. low engagement in the course.

Homework Grades: will work along these lines: you will either be graded as having completed an assignment (10) or not (0). If you have not completed an assignment, but made a best faith effort to accurately complete it, you will be asked to complete it within one week of the due date. If you have not at all tried to complete the assignment by the time of the due date, you will simply be given a 0 = incomplete.

- ***Homework is always due in the mornings before class (9:30 AM CST) on Thursdays unless otherwise noted. We will try hard to be consistent about this.***

Other Assignments: Throughout the class we will have surveys, quizzes, and other quick activities that I will grade.

Mid-term exam (30%) and Final Project (40%)

- Your final exam and projects will be given a letter grade (A, B, C, F).

  Grade Feedback and Grade Disputes

  Questions: if you want feedback only you can simply make an appointment with the TA or instructor. Feedback means you're there to listen to learn how to improve.

  Disputes: I only accept grade reconsideration requests in writing. You must email me, in writing, to dispute your grade after a 24-hour cooling-off period (yes, you must wait 24 full hours after the grade has been distributed to send this request. I promise this is going the helpful for both of us). You must include a justification as to why you are disputing your grade. In some cases, I will schedule a meeting with you, the director of graduate studies, and myself to discuss the issue.

  - ***Please keep in mind, grade disputes can lead to grade increases or decreases.***

Subject to Change

- The syllabus is a guide. All assignments and grading decisions are subject to change at professor's discretion. All changes will be announced in class and via email. Grades will be made available on the course website.

Course Schedule

SCHEDULE SUBJECT TO CHANGE. CHANGES WILL BE ANNOUNCED IN CLASS AND POSTED ONLINE

Part	Topic	Class Week	Date
1	EDA in data science & course set up	1	08/26
1	Summarizing Data & Foundations of Graphics	2	09/02
1	ggplot2 fundamentals + the scatterplot	3	09/09
1	Visualize: distributions and correlations	4	09/16
1	Visualize: distribution and correlations II	5	09/23
1	Visualize: rankings and other patterns	6	09/30
1	Visualize: How to “see” time in data _ midterm review	7	10/07
	Midterm	8	10/12
2	Introduction to Machine learning + Introduction to Final Projects	9	10/21
2	Data Types I: Text-as-data	10	10/28
2	Data Types II: Spatial Data and Maps	11	11/4
3	ML: Unsupervised learning, k-means	12	11/11
3	ML: Dimension reduction, principal components analysis	13	11/18
	Holiday Break	14	11/25
3	Wrap-up	15	12/02
	Finals Week	16	

Communication

Students will be invited to a course slack channel. Questions related to course logistics, content, homework, quizzes, or the final project should be posted in the slack channel. Individual questions should be sent to the instructor and/or TA by direct slack message.

Collaborative Learning

Students are encouraged to work together on homework assignments. Unless specifically noted in the instructions, students should not collaborate on quizzes or work otherwise noted as 'individual work.' Students that violate the collaborative-work policy on a quiz will fail the quiz in question and forfeit the opportunity to retake or resubmit.

Inclusivity Policy

This class respects and welcomes students of all backgrounds, identities, and abilities. If there are circumstances that make our learning environment and activities difficult, if you have medical information that you need to share with me, or if you need specific arrangements in case the building needs to be evacuated, please let me know. I am committed to creating an effective learning environment for all students, but I can only do so if you discuss your needs with me as early as possible. I promise to maintain the confidentiality of these discussions. If appropriate, also contact Student Access Services to get more information about specific accommodations.

Mental Health & Wellness

If you are experiencing undue personal and/or academic stress during the semester that may be interfering with your ability to perform academically, Vanderbilt's Student Care Network offers a range of services to assist and support you. I am available to speak with you about stresses related to your work in my course, and I can assist you in connecting with the Student Care Network. The Office of Student Care Coordination (OSCC) is the central and first point of contact to help students navigate and connect to appropriate resources on and off-campus, develop a plan of action, and provide ongoing support. You can schedule an appointment with the OSCC at https://www.vanderbilt.edu/carecoordination/ or call 615-343-WELL.

The Student Care Network also offers drop-in services on campus on a regular basis. You can find a calendar of services https://www.vanderbilt.edu/studentcarenetwork/satellite-services/.

If you or someone you know needs to speak with a professional counselor immediately, the University Counseling Center offers Crisis Care Counseling during the summer and academic year. Students may come directly to the UCC and be seen by the clinician on call, or may call the UCC at (615) 322-2571 to speak with a clinician. You can find additional information at https://www.vanderbilt.edu//ucc/.

Sexual Misconduct

Vanderbilt is committed to providing a community built on trust and mutual respect, where all can feel secure and free from harassment. Sexual misconduct including sexual violence, sexual harassment, intimate partner violence, and stalking, violates a person's rights, dignity and integrity and is contrary to our community principles and the mission of the college. The University is committed to fostering a community that promotes prompt reporting of sexual misconduct and timely and fair resolution of sexual misconduct reports. Creating a safe, respectful, and inclusive environment is the responsibility of everyone at Vanderbilt.

We encourage all members of our campus community to seek support from the Project Safe Center; 615-322-7233. We encourage community members to report all incidents of sexual harassment and sexual misconduct directly to the Title IX Coordinator (615-322-4705). Staff in these departments will assist in eliminating the misconduct, preventing its recurrence, and addressing its effects.

Mandatory Reporting

All faculty, many staff, and some students are "mandatory reporters" who are legally obligated to report any allegations of sexual misconduct (assault, harassment, dating violence, domestic violence, stalking and child abuse) and any suspected discrimination (about age, race, color, creed, religion, ancestry, national or ethnic origin, sex/gender, sexual orientation, disability, genetic information, military status, familial status or other protected categories under local, state or federal law) to Vanderbilt's Title IX Coordinator (615-343-9004).

This means that students who discuss such things with their peers and faculty do not have confidentiality. Students should be aware of that fact so they, both have choice about reporting, and options for other, confidential resources on campus. Title IX calls on the University to address the "impact" of sexual harassment and violence, so there are no time or geographical exclusions to what must be reported. Your reporting obligation extends to incidents that occur or occurred off-campus and to those that occurred prior to a person's affiliation with the University. Also, your reporting obligation applies in all situations, not just the classroom or in connection with a course. The only exclusion is a confidential support group setting, such as at the University Counseling Center or the Center for Student Wellbeing.

If you have any questions about the scope of your obligation, please contact the University Title IX Coordinator or the Director of the Project Safe Center.

Safety

The safety of students, faculty, and staff at Vanderbilt University is of the utmost importance. As a Vanderbilt student, you are automatically enrolled in AlertVU, which is used in emergencies which pose an imminent threat to the community. If you need to contact the Vanderbilt Police in an emergency, call 911 from any campus phone or (615) 421-1911 from any other phone. Additional information about emergency preparedness is available online.