George Mason University

STAT 463 – Introduction to Exploratory Data Analysis

Fall 2023

Instructor: David Kepplinger

Version 3 (08/23/2023)

Administrative

Course dates: Tuesday, Aug 22 – Thursday, Nov 30 (final exam period: Dec 12);
no class on Tuesday, Oct 10 (Fall Break) and Thursday, Nov 23 (Thanksgiving Day)

Instructor: Dr. David Kepplinger (he/him/his)

Email: dkepplin@gmu.edu

Office: Room 1711, Nguyen Engineering Building (ENGR)

Office hours:
Tuesday, 3 – 4 P.M. in-person in ENGR 1711
Thursday, 3 – 4 P.M. online via Zoom (link available on Blackboard)
Additional office hours by appointment

GTA: Siqi Wei (he/him/his)

Office hours: Wednesday, 1 – 2 P.M. online via Zoom (link available on Blackboard)

Blackboard course page: https://mymasonportal.gmu.edu/ultra/courses/_494943_1/

Class time:
Tue, 1:30 – 2:45 P.M. in-person in Innovation Hall Room 137
Thur, 1:30 – 2:45 P.M. synchronous online via Zoom (link posted on Blackboard)

The class is scheduled as hybrid between face-to-face on-campus meetings (Tuesday classes) and synchronous online meetings via Zoom (Thursday classes). All learners taking courses with a face-to-face component are required to follow the university's public health and safety precautions and procedures outlined on the university Safe Return to Campus webpage (https://www2.gmu.edu/safe-return-campus). If the campus closes, or if a class meeting needs to be canceled or adjusted due to weather or other concern, learners should check Blackboard for updates on how to continue learning and for information about any changes to events or assignments.

Communications The Blackboard site for this course is the primary channel of communication. Please check the Blackboard course regularly for updates! Information posted on the Blackboard site includes

- announcements,
- lecture notes,
- homework assignments and exams,
- changes to the posted office hours,
- handouts and readings.

Any question related to concepts and topics should be asked on Harmonize, accessible through Blackboard (under *Syllabus > Course Q&A)*. Questions will be visible to all registered students, and everyone is expected to actively participate in answering questions posted by peers. Active participation in answering questions is part of the participation grade.

E-mail communication is restricted to questions relating to sensitive and confidential information (such as grade concerns, personal circumstances requiring specific accommodations, etc.).

- E-mails will be returned within 2 business days and may not be returned on weekends/holidays.
- When you send an e-mail to me, please put `STAT 463` at the beginning of the subject line.
- E-mails related to this course must be sent and received via your Mason e-mail account. **E-mails sent from other e-mail accounts may not be answered.** (This is a university policy and part of your guaranteed rights under FERPA.)
- E-mails with questions that should be posted to Harmonize may not be answered.

Should you have concerns that you may not be able to fully participate or engage in any of the activities listed below, please do not hesitate to contact me either by e-mail or speak to me in person during office hours or after class. We can discuss alternative arrangements that suit your needs.

Course requirements

Prerequisites: STAT 350, 354, 360, or BUS 310.

Required readings: The required textbooks for this course are

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R.* 2nd Edition. Springer. This book is available online for free.
- Wickham, H., Çetinkaya-Rundel, M., Grolemund, G. (2023). *R for Data Science.* 2nd Edition. O'Reilly. This book is available online for free.
- Lander, J. (2017). *R for Everyone: Advanced Analytics and Graphics.* 2nd Edition. Pearson. This book is available online through the GMU Library.

A number of relevant articles will be posted in Blackboard as different topics are discussed.

Hardware requirements: We will frequently be using laptop computers for in-class activities. Please be respectful of your peers and your instructor and do not engage in activities that are unrelated to the class.

Software requirements: This class will use R (version 4.2 or higher; available from https://cran.r-project.org/) and the RStudio IDE (version 2023.06.0 or newer; available from https://posit.co/products/open-source/rstudio/). Assignments and in-class activities will use interactive tutorials powered by Posit Connect (formerly known as RStudio Connect). To access these assignments a recent web browser with Javascript support fully enabled is required. For all assignments the complete code must be submitted for reproducibility.

Activities and assignments in this course will regularly use web-conferencing software (Zoom). In addition to the requirements above, students are required to have a device with a functional camera and microphone. In an emergency, students can connect through a telephone call, but video connection is the expected norm.

Course description

Learning objectives: After successfully completing STAT 463, you will be able to:

- Curate and access data from various sources.
- Wrangle data for the data analysis workflow.
- Compute basic descriptive statistics and perform comparisons to explore quantitative data.

- Use and interpret appropriate statistical models to find patterns in data.
- Visualize data to gain insights about data, the underlying phenomena, and to communicate selected results with a defined audience.

Main topics: You can expect the following topics to be covered in some detail.

- Introduction to the programming language R and the “tidyverse”.
- Visualizing data via ggplot2 in R.
- Simple and multivariate linear regression.
- Variable selection in regression models.
- Cross-validation for variable selection.
- Regression/classification trees and random forests.
- Generalized Linear Models, particularly the logistic regression model.
- Principal component analysis.
- Regularized estimation in regression models (LASSO, EN, and Ridge regression).
- Cluster analysis.

Assessments and grading

Your grade in this course will be based on weekly homework assignments of various types, in-class quizzes, an in-class midterm, a take-home final exam, and participation. The number of quizzes and homework assignments, and their relative grading emphasis may be adjusted. The instructor reserves the right to change these percentages if needed.

Assignment	Tentative due date	Weight
Homework assignments	weekly	25%
Quizzes	throughout the term	20%
Midterm	Oct 12	20%
Final exam	Dec 10 – Dec 12	30%
In-class participation	weekly	5%
Total		100%

Written and oral communication are an integral part of any statistical work, and as such, grammar, style, and spelling are part of grading rubrics applied to all deliverables. You are strongly encouraged to use the resources and tutoring offered by the writing center (https://writingcenter.gmu.edu).

All assignments in this course are designated as individual assignments, which are to be undertaken independently. You may discuss your ideas with others but everything you turn in must be your own work. You may not share analyses, graphs, code, and other materials. You are responsible for making sure that there is no reason to doubt that the work you hand in is your own. The following types of collaboration on individual assignments are not honor code violations:

- Working on assignments with someone who is at roughly the same stage of progress as you, provided both learners contribute in roughly equal quantity and quality (in particular, thinking) on whatever problem or problem parts they collaborate. This type of collaboration is actually encouraged!
- A moderate amount of asking, "How do I do this in R?" However, as you gain enough familiarity, you should get in the habit of using online help and trying logical possibilities, then asking for help only if these do not succeed after a reasonable try.
- Using programming code found on the internet or in software libraries, if using proper attribution (clearly identifying and citing all code snippets which are not your intellectual product).

The following types of collaboration on assignments **are** honor code violations:

- Working together with one learner the doer and one the follower.
- Any type of copying. In particular, splitting up a problem so that different learners do different parts is not authorized collaboration on homework. This also includes copying code from the internet without properly identifying the source.

Attendance: Attendance is expected and in-class participation is part of your final grade. If you miss class, please get notes from your peers. You are responsible for material covered in class and announcements made during class.

Participation: Success in this course requires active participation in in-class activities and discussions, for which you will need to prepare in advance for each class period. Accordingly, you are expected to prepare for class period by

- reading the corresponding sections of textbooks or research articles to be covered in class,
- reviewing class materials posted in Blackboard to be covered in class,
- familiarizing yourself with the use of the covered methods and techniques in R.

Homework assignments: There will be weekly homework assignments throughout the term which will vary in length and content. Some involve in-class activities and continuing problems started in class, others involve solving exercises related to the material covered in class. **Only 11 of your homework assignments will be graded.** It is your responsibility to not submit a homework assignment if you do not want it to be graded. Once the homework is submitted, you cannot withdraw that homework and it will count towards your final grade. Once you submitted 11 homework assignments over the course of the semester, any following homework assignments will not be graded or considered for the final grade. You will typically have 7 days to complete each homework. Unless posted otherwise on Blackboard, all homework assignments are due Fridays, 11:59 P.M. Specific due dates and submission instructions will be posted in Blackboard. Late submissions will be penalized by reducing the total number of points possible by 20% of the original total number of points for each day late. For example, if a homework assignment is worth a total of 10 points, it will only be worth 8 points when submitted within the first 24 hours after the due date, 6 points when submitted during the second 24 hours after the due date, and so on. Submissions will not be accepted more than 4 days past the due date.

Quizzes: There will be 4 in-class quizzes spread across the semester. For each quiz you will have about 25 minutes. Quizzes will cover the materials present up to and including the previous lecture, with an emphasis on the most recent topics.

Midterm: There will be an in-class midterm exam. The midterm will cover all materials presented up to and including the previous lecture.

Final exam: The final exam will be an open-book, take-home exam, covering all materials presented over the semester. Detailed instructions and policies will be posted on Blackboard. Late submissions of the final exam will not be accepted.

Regrading policies: You have at most one week after a score is posted for an assignment to appeal the score. If you want parts of an assignment remarked, send an email specifying the question/part and the reason for requesting a review of grading. If you do not write a notification of any issues with your score within that time, then the posted score stands (whether or not it is correct).

Policies and Classroom Climate

During classes and online you are encouraged to discuss and share ideas with your classmates (see above how this relates to the honor code). To facilitate a respectful and inclusive classroom climate, be open to explore and challenge each other's ideas without criticizing individuals. Diversity is a source of creativity and innovation and I ask that learners appreciate diverse perspectives, that they listen respectfully and let everyone speak. If you have concerns about the dynamics or classroom climate, please do not hesitate to bring them to my attention.

The School of Computing seeks to create a learning environment that fosters respect for people across identities. We welcome and value individuals and their differences, including gender expression and identity, race, economic status, sex, sexuality, ethnicity, national origin, first language, religion, age and ability. We encourage all members of the learning environment to engage with the material personally, but to also be open to exploring and learning from experiences different than their own.

Gender identity and pronoun use: If you wish, please share your name and gender pronouns with me and how best to address you in-person or via email. I use he/him/his for myself and you may address me as “David”, “Prof. Kepplinger” or “Dr. Kepplinger” in email and verbally.

Individual accommodations: Disability Services at George Mason University is committed to providing equitable access to learning opportunities for all learners by upholding the laws that ensure equal treatment of people with disabilities. If you are seeking accommodations for this class, please first visit http://ds.gmu.edu for detailed information about the Disability Services registration process. Then please discuss your approved accommodations with me. Disability Services is located in Student Union Building I (SUB I), Suite 2500. Email: ods@gmu.edu | Phone: (703) 993-2474.

Class etiquette: Class will start on time at 1:30 PṀand end on time at 2:45 PṀ. Although situations may arise making it impossible for you to arrive on time and/or requiring you to leave early, please remember that late arrivals and early departures can be quite disruptive to your classmates. So, please make arriving to class late or leaving early an exception, not a habit. This applies to the in-person lecture on Tuesday as well as the online lecture on Thursday! **Regular attendance for the full period of each class is very important for this course!**

- Mute your phones during class, and keep them stowed away.
- You may eat during class, as long as it is done discreetly, quietly, and odorless.
- Immediately before or after class is not a good time to ask lengthy questions. Please come to office hours (or make an appointment) instead. Questions during class are welcomed and encouraged.

Office hour policy: Please prepare specific questions to ask during office hour. You may also just sit in and listen to others’ questions. You can always just show up during office hours

Netiquette: We will often communicate via discussion forums other forms of online communication. To facilitate effective communication via these channels, please adhere to the following:

- *Be relevant and concise:* When posting a message to an online discussion, stick to the topic, make sure that you send enough information, and be concise.
- *Use accurate topic titles:* Each posting should include a topic title (a subject line) that lets the recipient know the posting's content. This allows others to scan their online messages, read the more important messages first, and keep organized.
- *Read before posting:* Read posted questions/answers before asking a new question to avoid repeating points already made, asking questions already answered, or bringing up points that have already been argued and either accepted, rejected, or exhausted. In addition, by "replying" to messages instead of starting a new message, a thread of communication can be kept going.
- *Be polite:* Avoid inflammatory messages and language. Do not send a message that ridicules someone else. Also, be careful when using humor or sarcasm, as most of it gets lost in the medium.
- *Review messages before submitting:* Think before you "speak" electronically. For the most part, electronic communication is a non-visual form of communication; therefore, people are unable to rely on facial expressions, tone of voice, or body language to interpret electronic messages. Misunderstandings can easily occur because of these factors.

Notice of mandatory reporting of sexual or interpersonal misconduct: As a faculty member, I am designated as a "Non-Confidential Employee," and must report all disclosures of sexual assault, sexual harassment, interpersonal violence, stalking, sexual exploitation, complicity, and retaliation to Mason's Title IX Coordinator per University Policy 1202. If you wish to speak with someone confidentially, please contact one of Mason's confidential resources, such as Student Support and Advocacy Center (SSAC) at 703-380-1434 or Counseling and Psychological Services (CAPS) at 703-993-2380. You may also seek assistance or support measures from Mason's Title IX Coordinator by calling 703-993-8730, or emailing titleix@gmu.edu.

Honor Code: The integrity of the University community is affected by the individual choices made by each of us. Mason has an Honor Code with clear guidelines regarding academic integrity; you are responsible to know your requirements for this course. All violations of these rules will be referred to the Honor Committee; I take the Honor Code seriously and so should you. No grade is important enough to justify academic misconduct. If you have any questions concerning the Honor Code and how it relates to this particular course, please contact me.

Some kinds of participation in online study sites violate the Mason Honor code: these include uploading of any of the course materials or exams; and uploading any of your own answers or finished work. Always consult your syllabus and me before using these sites.

Privacy: Your privacy is governed by the Family Educational Rights and Privacy Act (FERPA) and is an essential aspect of this course. You must use your MasonLive email account to receive important University information, including communications related to this class. I will not respond to messages sent from or send messages to a non-Mason email address.

All course materials posted to Blackboard or other course sites are private to this class; by federal law, any materials that identify specific learners (via their name, voice, or image) must not be shared with anyone not enrolled in this class.

- Videorecordings — whether made by me or learners — of class meetings that include audio, visual, or textual information from other learners are private and must not be shared outside the class.
- Live video conference meetings (e.g., Zoom) that include audio, textual, or visual information from other learners must be viewed privately and not shared with others in your household or recorded and shared outside the class.

Copyrights of course material: This course gives you access to presentations, handouts, and copyrighted material and articles. Please treat them accordingly. All material other than copyrighted material should be regarded as authored materials, which if used or referred to must be fully credited through reference to me, the course, and date. If used beyond citation, my permission is required.

Version History

Version 1 (08/15/2023) Initial version.

Version 2 (8/16/2023) Updated the location of the Harmonize Q&A.

Version 3 (08/23/2023) Add information on GTA office hours.