MATH 49995/57091: Introduction to Data Science
Spring 2020

Course information

- Sections: MATH-49995-001 and MATH-57091-001
- Lecture times: Mondays and Wednesdays 2:15PM–3:30PM
- Location: MSB 158

Instructor

- Tsung-Heng Tsai
- E-mail: ttsai1@kent.edu
- Office: MSB 372
- Office hours: Mondays and Wednesdays 3:30PM–5PM or by appointment

Course objectives

This course offers a gentle introduction to the field of data science. The goal of this course is to teach students how to answer questions with data. We will cover topics of data wrangling, exploratory data analysis, statistical inference and modeling, and statistical learning. We will also teach the necessary skills to gather, organize, explore and analyze data, and to develop data products to facilitate effective communication and reproducible research. The programming language R will be used extensively. The course will rely on the existing implementations of statistical methods in many cases, but some programming efforts will also be required.

Textbooks

There is no required textbook but course notes will be provided throughout the course. Useful references are:

1. *Introduction to Data Science: Data Analysis and Prediction Algorithms with R*, Rafael A. Irizarry, CRC Press, 1st Edition, 2019
2. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, Hadley Wickham and Garrett Grolemund, O'Reilly Media, 1st Edition, 2017

Further reading will be recommended to support weekly class material.

Course prerequisites

You should have completed MATH 20011 with a grade of C or better. If you are enrolled in 57091, you must have graduate standing in Mathematics. Students who do not have the proper prerequisites risk being deregistered from the class. Please contact instructor if you would like to take the course, but do not satisfy the prerequisite.

Grading

Grades will be calculated as follows:

- 6 homeworks: 30% (the lowest grade dropped)

- Midterm exam: 30%
- Final project: 40%

The final letter grades will follow the usual scale:

- 90–100 = A-range (i.e., A or A-)
- 80–89 = B-range (i.e., B+, B or B-)
- 70–79 = C-range (i.e., C+, C or C-)
- 60–69 = D
- 0–59 = F

The cutoffs for "+" and "-" will be determined at the end of the semester, at the discretion of the instructor. This scale is subject to change at the discretion of the instructor.

Homework

There will be six homework assignments. Each homework is due before class on the day it is listed, and should be completed in R Markdown format (with Rmd extension). An Rmd file contains a combination of content with simple text and R code chunks. Both the R Markdown source file and the resulting PDF output (generated by calling "Knit to PDF") must be turned in through Blackboard.

Students may discuss and collaborate with friends, but your submitted work must be your own. Sharing of solutions will not be tolerated and will be considered cheating.

No late work will be accepted. Extensions may be given individually if requested at least 48 hours in advance of the due date with a reasonable justification. The lowest homework grade will be dropped.

Exam

There will be an in-class midterm exam that counts for 30% of your grade. No collaboration with peers is allowed.

Final project

There will be a final project that counts for 40% of your grade. More details will be posted and discussed later in the semester.

Registration information

The official registration deadline for this course is Jan 19, 2020. University policy requires all students to be officially registered in each class they are attending. Students who are not officially registered for a course by published deadlines should not be attending classes and will not receive credit or a grade for the course. Each student must confirm enrollment by checking his/her class schedule (using Student Tools in FlashLine) prior to the deadline indicated. Registration errors must be corrected prior to the deadline.

The course withdrawal deadline is March 22, 2020. Other important Registrar dates can be found at http://www.kent.edu/registrar/registrar-dates-term.

Academic integrity

University policy 3-01.8 deals with the problem of academic dishonesty, cheating, and plagiarism. None of these will be tolerated in this class. The sanctions provided in this policy will be used to deal with any violations. If you have any questions, please read the policy at http://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism and/or ask.

Accommodations for students with disabilities

University policy 3-01.3 requires that students with disabilities be provided reasonable accommodations to ensure their equal access to course content. If you have a documented disability and require accommodations, please contact the instructor at the beginning of the semester to make arrangements for necessary classroom adjustments. Please note, you must first verify your eligibility for these through Student Accessibility Services (contact 330-672-3391 or visit www.kent.edu/sas for more information on registration procedures).

Schedule

This schedule is tentative and subject to change.

- Mon Jan 13. Introduction to data science and the course
- Wed Jan 15. R, RStudio, R Markdown
- Mon Jan 20. *No class - Martin Luther King Jr. Day*
- Wed Jan 22. Basic R
- Mon Jan 27. Basic R
- Wed Jan 29. Basic R
- Mon Feb 3. Layered grammar of graphics
- Wed Feb 5. Data visualization with ggplot2
- Mon Feb 10. Principles of data visualization
- Wed Feb 12. Data transformation
- Mon Feb 17. Tidy data
- Wed Feb 19. Data importing
- Mon Feb 24. Data wrangling
- Wed Feb 26. Exploratory data analysis
- Mon Mar 2. Statistical inference
- Wed Mar 4. Resampling techniques
- Mon Mar 9. Review
- Wed Mar 11. Midterm exam (in class)
- Mon Mar 16. Statistical modeling
- Wed Mar 18. Advanced modeling
- Mon Mar 23. *No class - Spring Break*
- Wed Mar 25. *No class - Spring Break*
- Mon Mar 30. Version control, git, and GitHub
- Wed Apr 1. Building R Shiny apps
- Mon Apr 6. Statistical prediction
- Wed Apr 8. Model evaluation and cross validation
- Mon Apr 13. Dimensionality reduction
- Wed Apr 15. Clustering
- Mon Apr 20. Databases
- Wed Apr 22. R package
- Mon Apr 27. Student presentations on final projects