STAA57H3 Introduction to Data Science

Winter 2019

Department of Computer & Mathematical Sciences
University of Toronto Scarborough

Lectures: WE 9-11 @ MW170
FR 9-11 @ SW143
Instructor: Sotirios (Sam) Damouras (sdamouras@utsc.utoronto.ca)
Office Hours: MO 11-13 | WE 11-13 | FR 13-15 @ IC456
TAs: Xichen (Sherry) He (xichen.he@mail.utoronto.ca)
Jiali Pan (jiali.pan@mail.utoronto.ca)
Xincheng Zhang (xincheng.zhang@mail.utoronto.ca)

About the course

Welcome to STAA57! This course provides an overview of *Data Science*, i.e. the study of extracting knowledge from data. We will examine the fundamental concepts, principles, and methods of Data Science, using computing as our primary tool. In particular, we will be using the R language and RStudio environment for analyzing data and creating reports. The course content is divided along three themes[1]:

- Data Management & Visualization
- Statistical Inference
- Predictive Methods (Machine Learning)

By the end of this course, you should be able to carry out a statistical investigation, including: a) formulating relevant questions, b) selecting appropriate methods to address them, c) processing and analyzing data, d) and communicating the results.

Textbooks

We will cover part of the following free textbooks:

- R for Data Science, by Garrett Grolemund & Hadley Wickham
- Introductory Statistics with Randomization and Simulation, by David M Diez, Christopher D Barr, and Mine Çetinkaya-Rundel
- An Introduction to Statistical Learning, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Marking Scheme

Worksheets	completed in class	15% (best 18/22)
Midterm Exam	∼week 7 (TBD)	25%
Course Project	throught semester	20%
Final Exam	April (TBD)	40%

[1] Detailed list of topics provided at the end.

Homepage

Available through Quercus (https://q.utoronto.ca/)

Worksheets

Each 2hr class session will consist of two parts: the first part (1hr) will be a traditional lecture, and the second part (1hr) will be a facilitated practice session where students will have to complete worksheets. During the practice session, the instructor and TAs will be there to provide help: they can go over the material, clarify worksheet questions, and offer suggestions, but they will not provide or verify answers. You are allowed, and even encouraged, to discuss the worksheets with your peers, but you must write and submit your own answers! There will be 22 worksheets in total, which will be graded for correctness and completeness. Your best 18 out of 22 worksheet marks will count for 15% of the final marks. This means that you can miss up to 4 worksheets without penalty. If you miss several worksheets for a valid reason (e.g. medical reason) and with appropriate documentation, I will make an adjustment.

Course Project

The project will be completed by teams of 3-4 students. The goal is to conduct an open-ended statistical inquiry on a topic of interest, using available data and the tools you learned in class. There will be three milestones to the project, distributed throughout the semester:

Milestone	Rel. Weight	Details
Initial Proposal	20%	Describe research questions & data used to answer them
Draft Report	20%	Preliminary analysis results
Final Report	60%	Present final results & conclusions

Each milestone has a deliverable: the first two are meant to give you feedback, while the last one is for evaluating your work. More details on the project will be provided on the course webpage.

Midterm Policy

If you miss the midterm for medical reasons, you must submit a UofT medical certificate (available here) within one calendar week of the missed test; other medical certificates or notes will NOT be accepted. Medical certificates must be signed by an Ontario-registered MD, with registration number and phone number. The doctor must specifically indicate that there was a disabling health problem on the day of the test. The doctor should be contactable by us for verification. Each student with accepted documentation will be required to take a rigorous make-up test. If documentation is not provided or accepted, your missed midterm grade will be zero.

Accessibility

Students with diverse learning styles and needs are welcome in this course. In particular, if you have a disability/health consideration that may require accommodations, please feel free to approach me and/or the AccessAbility Services Office as soon as possible. I will work with you and AccessAbility Services to ensure you can achieve your learning goals in this course (enquiries are confidential). The UTSC AccessAbility Services staff (located in S302) are available by appointment to assess specific needs, provide referrals and arrange appropriate accommodations, at (416) 287-7560 or mailto:ability@utsc.utoronto.ca.

Wellness

University life and academic studies can be stressful, so I encourage you to take good care of yourself. Do your best to maintain a healthy lifestyle throughout the semester by eating well, exercising, socializing, getting enough sleep and taking time to relax. This will help you achieve your goals and cope with stress.

If you, or anyone you know, experiences severe academic stress, difficult life events, or feelings of anxiety or depression, I strongly encourage you to seek support. Consider reaching out to a friend, family, or faculty member that you trust sooner rather than later. Do not hesitate, because learning to ask for help is an important lesson in itself. And keep in mind that the University's Health and Wellness Centre (https://www.utsc.utoronto.ca/hwc/counselling-supports-and-services) is always available to you for counseling and support.

Academic Integrity

Academic integrity is fundamental to learning and scholarship at the University of Toronto. Participating honestly, respectfully, responsibly, and fairly in this academic community ensures that the U of T degree that you earn will be valued as a true indication of your individual academic achievement, and will continue to receive the respect and recognition it deserves.

Familiarize yourself with the University of Toronto's Code of Behaviour on Academic Matters ((http://www.governingcouncil.utoronto.ca/policies/behaveac.htm). It is the rule book for academic behaviour at the U of T, and you are expected to know the rules. Potential offences include, but are not limited to:

In papers and assignments:

- Using someone else's ideas or words without appropriate acknowledgement.
- Copying material word-for-word from a source (including lecture and study group notes) and not placing the words within quotation marks.
- Submitting your own work in more than one course without the permission of the instructor.
- Making up sources or facts.
- Including references to sources that you did not use.
- Obtaining or providing unauthorized assistance on any assignment including:
  - working in groups on assignments that are supposed to be individual work;
  - having someone rewrite or add material to your work while "editing".
- Lending your work to a classmate who submits it as his/her own without your permission.

On tests and exams:

- Using or possessing any unauthorized aid, including a cell phone.
- Looking at someone else's answers
- Letting someone else look at your answers.
- Misrepresenting your identity.
- Submitting an altered test for re-grading.

Misrepresentation:

- Falsifying or altering any documentation required by the University, including doctor's notes.
- Falsifying institutional documents or grades.

Lecture Schedule

Theme	Lec #	Topic
Data Mgmt & Vis	1	Rstudio & Markdown
	2	R Language Basics
	3	Data Tables & Subsetting
	4	Data Transformations & Summaries
	5	Combining Data
	6	Data Visualization
	7	Working with Text
	8	Hierarchical Data & Web Scraping
Statistical Inference	9	Statistical Sampling
	10	Estimation & Bootstrap
	11	Hypothesis Testing
	12	More Hypothesis Testing
	13	Multi-group Comparisons
	14	Linear Regression
	15	Causation & Experiments
	16	Multiple Regression
Prediction	17	Intro to Classification
	18	Multivariate Classification
	19	Overfitting
	20	Multiple Classes & Ensemble Methods
	21	Nonlinear Regression
	22	Model Selection
	23	(Work on Project)
	24	(Project Presentations)