DS 6306 Doing Data Science

Course Overview and Expectations

- As a data scientist in training, you need to learn to search out answers to questions for yourself before asking your instructor. The first question I will ask when you come to me with a problem is, “What have you tried to do to answer your own question?” If you have done nothing, then I reserve the right not to answer your question. I expect that everyone in the class follow this policy. Don’t help a classmate unless the classmate can demonstrate that he or she has tried to find a solution. It’s amazing what five minutes on the Internet can turn up! Please, train yourself to try to find an answer for yourself before asking someone else. Give yourself a time limit (e.g., “I will search for 30 minutes, and if I can’t find anything, I will ask”). It’s better for you in the long run
- Watch all asynchronous videos BEFORE coming to class. Class time is for asking questions and working problems related to the asynchronous material to help solidify what you have watched.
- Read assigned material in the text BEFORE coming to class. Class time is for asking questions and working problems related to the asynchronous material to help you solidify what you have read.

Learning Objectives

Students will:

- Get a practical hands-on overview of the end-to-end data science process using industry standard tools and techniques.
- Use tools such as R, R-Studio, knitr, rmarkdown, and Github to organize and document research so that others can reproduce and/or continue your work.
- Use the principles of “tidy data” to create clean data sets from messy ones using R.
- Conduct Exploratory Data Analysis (EDA) to understand, summarize and extract insights from data sets.
- Learn and apply basic machine learning and time series modeling techniques.
- Communicate the findings of a project in a clear, concise, and scientific manner.

Textbooks and Materials

Gandrud, C. (2015), Reproducible Research With R and R-Studio. Boca Raton, FL: CRC Press.

O’Neil, C., and Schutt, R. (2014). Doing Data Science: Straight Talk From the Front Line. San Francisco, CA: O’Reilly Publishers.

Grading

Assignment/Assessment	Weight on Final Grade
Live Session Assignments:	50%
Case Studies	30%
Lecture Questions during asynchronous materials (BLTs)	15%
Live Session Attendance	5%

Assignment and Assessment Information

Live Session Assignments (50%): Data science is best learned by doing. During the live session, we will have breakout sessions that will involve solving problems using software. Some live session periods will start with questions on the assigned reading for the week. Any assignments given during live session will be due before the next live session.

Case Studies (30%): There will be two of these, each equally weighted. Deadline for submission will be determined by the live session instructor.

Videos and Questions during asynchronous material (BLTs) (15%): Several modules have questions that are asked during or at the end of a video. These questions are called “Bidirectional Learning Tools” or BLTs. Responses to these questions aid the instructor in determining what is clear and what is not clear to students, and will help the live session instructor structure assignments and activities for the live session. There is a grade for watching each unit videos as well.

Live Session Attendance (5%): This ought to be an easy grade! Just show up in the live session! I understand that work/illness/family sometimes makes class attendance impossible. Just let me know if you cannot attend a live session, and make sure you watch the recording for that live session and do any assignments for the session that you missed.

Rescheduling Exams: Life happens. Should you need to reschedule an exam, please give notice to your live session instructor at least 24 hours prior to the live session in which the exam review is discussed (Unit 7). The notice should be given via e-mail. You and your instructor will discuss the best course of action given your circumstances. Retakes of exams will not be allowed, and a missed exam cannot be made up if notification is received AFTER the exam has taken place.

Submission guidelines for assignments

- Submit solutions in problem order.
- Use an easy-to-read variable-width font with a minimum of 11-point font. (These include Arial, Helvetica, and Geneva fonts—this document is in Helvetica 11 point.).
- Relevant code and output must be included in-line at the appropriate point using Courier New (or other fixed width) font, in 10-point size. Inclusion of irrelevant code or output will be penalized.
- Any graphics must be electronically cut and pasted in-line at the appropriate point of the write- up. You can use Word to resize the graphics appropriately.
- Any mathematical notation must be provided with appropriate use of subscripts, superscripts, and symbols. Use MS Equation or another equation editor if you submit your work in Word.

Weekly Schedule

Unit	Topic	Readings	Assignments due
1	What Is Data Science? Reproducibility in Research	Gandrud, C. (2015), *Reproducible Research With R and R-Studio*. Boca Ration, FL: CRC Press, chapter 1 and chapter 2. O'Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O'Reilly Publishers, chapter 1. Driscoll, M. (2011), "What Is Data Science?" Quora.com. Donoho, D. L. (2010), "An Invitation to Reproducible Research." *Biostatistics,* 11, 3: 385–388. Woodie, A. (2015), "So You Want to Be a Data Scientist." Datanami.	Install the swirl package and complete modules 1-7.
2	Introduction to R	Required Readings Gandrud, C. (2015), *Reproducible Research With R and R-Studio.* Boca Ration, FL: CRC Press, chapters 2 and 3. Recommended Readings	Complete Modules 8 to 11 in the R Programming course of Swirl. • Complete "8: Logic" • Complete "9: Functions" • Complete "10: lapply and sapply" • Complete "11: vapply and tapply"

		Kabacoff, R. (2014), "Quick-R." Statmethods.net. The Comprehensive R Archive Network.cran.r-project.org (n.d.), "An Introduction to R."	
3	Tools for Data Science	Gandrud, C. (2015), *Reproducible Research With R and R-Studio.* Boca Ration, FL: CRC Press, chapters 4 and 5. Baggerly, K.A., and Berry, D. A. (2011), "Reproducible Research." *Amstat News.* Fomel, S., and Claerbout, J. F. (2009), "Guest Editors' Introduction: Reproducible Research." *Computing in Science & Engineering*, 11, 1: 5–7.	Complete Modules 12 to 15 in the R Programming course of Swirl. • Complete “12 : Looking at Data” • Complete “13 : Simulation” • Complete “14 :Dates and Times” • Complete “15: Graphics Basics”
4	Gathering Data for Analysis	Refsnes Data (n.d.), "XML Tutorial." W3Schools.com. Sanchez, G. (n.d.), Links to various texts and projects on string processing using R. GastonSanchez.com. "Extracting Data from XML"	None
5	Data Wrangling	Gandrud, C. (2015), *Reproducible Research With R and R-Studio.* Boca Ration, FL: CRC Press, chapter 6. Wickham, H. (2014), "Tidy Data." *Journal of Statistical Software*, 59, 10.	None

6	Exploratory Data Analysis	O’Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O’Reilly Publishers, chapter 2. O’Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O’Reilly Publishers, sample data set (on GitHub website). Anderson, C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Wired.com. Crawford, K. (2013), "The Hidden Biases in Big Data." HBR.org.	None
7	Case Study 1 Overview	Case Study Overview	None
8	Case Study 1	None	Complete Case Study 1
9	Machine Learning 1	O'Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O'Reilly Publishers, Chapter 3: pp 51 - 71, 78, pp-85-86. Recommended Readings R MLR tutorial	None
10	Machine Learning 2	O'Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O'Reilly Publishers, Chapter 3: pp 71 - 91. P. Domingos, *A few useful things to know about machine learning, Communications of the ACM*, Volume 55 Issue 10, October 2012, Pages 78-87. [see attached paper: CACM12.pdf]	None

11	Machine Learning 3	Required Readings O'Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O'Reilly Publishers, Chapter 3: Spam Filters, Naive Bayes, and Wrangling. Recommended Readings 10 more lessons learned from building machine learning Part 1 and Part 2 [see attached PDFs: 10MoreLessonsPart1.pdf and 10MoreLessonsPart2.pdf]	None
12	Modeling Financial Data 1	O'Neil, C., and Schutt, R. (2014), *Doing Data Science: Straight Talk From the Front Line.* San Francisco, CA: O'Reilly Publishers, chapter 6.	None
13	Modeling Financial Data 2	None	None
14	Case Study 2 Overview	Case Study Overview	None
15	Case Study 2	None	Complete Case Study 2

University Policies

Grading Policy: Graduate students must receive a C or better in a course in order to pass the course. If a student must retake a course, then the second grade and the first grade are averaged for the purposes of the overall GPA. Failure to maintain a GPA of 3.0 or better will result in dismissal from the program.

Incompletes will be given only in the case of extraordinary circumstances that prevent you from finishing the semester. You must have completed at least 50% of the course with a passing grade to be eligible for an incomplete.

Religious Observance: Religiously observant students wishing to be absent on holidays that require missing class should notify their professors in writing at the beginning of the semester, and should discuss with them, in advance, acceptable ways of making up any work missed because of the absence. (See University Policy No. 1.9.)

Excused Absences for University Extracurricular Activities: Students participating in an officially sanctioned, scheduled University extracurricular activity should be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to plan with the instructor prior to any missed scheduled examination or other missed assignment for making up the work. (University

Undergraduate Catalogue) **Work-related travel and meetings count as officially sanctioned activities (McGee's addition).**

Students with Disabilities

Students needing academic accommodations for a disability must first register with Disability Accommodations & Success Strategies (DASS). Students can call 214-768- 1470 or visit https://www.smu.edu/disabilities begin the process. Once registered, students should then schedule an appointment with the professor as early in the semester as possible, present a DASS Accommodation Letter, and make appropriate arrangements. Please note that accommodations are not retroactive and require advance notice to implement.

Honor Code and Academic Integrity

Students are expected to abide by the SMU Honor Code, which can be found online at http://www.smu.edu/StudentAffairs/StudentLife/StudentHandbook/HonorCode. The Honor Code prohibits academic sabotage, cheating, fabrication, facilitating academic dishonesty, and plagiarism. Definitions of all of these items can be found on the Honor Code website. The penalty for a first offense is a zero for the assignment. A second offense will result in a failing grade for the course and possible dismissal from the program.

Best Practices for Success

Attendance. Take responsibility for your commitment. Attendance means not only being there for synchronous sessions but also participating in asynchronous work.

Citizenship. You need to be actively engaged to succeed in this class. Talking on cell phones, texting, "facebooking," tweeting, or leisure web browsing are prohibited in class. These are considered to be disruptive (not to mention rude).

Integrity. A lot of the graded work occurs outside of class, so honesty and integrity are expected in your submissions. Evidence of academic dishonesty will minimally result in zeroes for all involved parties, and perhaps University-level disciplinary action. Don't risk your academic career.

Humility. Don't get lost! Ask questions in class. If something isn't clear to you, it probably isn't clear to others either. Questions may arise because your professor hasn't made a connection clear or has inadvertently left out an important point. Your question gives the professor a chance to explain more clearly. Don't be proud or shy.

Organization. Don't procrastinate! This is a technology-driven course. Count on your computer failing or your wireless connection breaking the night before a due date. Start early and give yourself a chance to succeed.

Deadlines. You will generally have a week to complete an assignment. Due dates and times will be clearly indicated. Late submissions will be penalized, but it is much better to turn in work late than not at all (or to turn in incomplete/sloppy work). Work turned in after solutions have been posted to the course website will receive no credit.

Getting help. If questions arise while doing assignments/exams, do your best to resolve these questions before the assignment is due, first by taking time to seek answers yourself, next by asking questions on the wall, and finally via e-mail to your instructor or other students. I encourage you and expect you to seek help. For questions during exams, please e-mail the live session instructor directly.

Collaboration. The formation of study groups and collaboration with your fellow students in tackling the assignments is encouraged. Working together in groups on homework is permitted, even encouraged. **However, every student should write up and complete his or her homework independently. Students who chose to turn in exactly the same work will share the grade assigned.** Talking about problems with other people does help in learning, but just copying the solutions from one another doesn't help!

Looks do matter! All assignments must be NEATLY executed and organized. You risk a zero on any assignment submitted in a sloppy manner. See submission guidelines for more detail.

This syllabus is only a guideline and is not a legal contract. The professor of record for the course has final say on any policies, due dates, etc.