Master of Science in Data Science

DS 6306: Doing Data Science Course Syllabus

Welcome to Doing Data Science

Welcome to Doing Data Science! In this course we will first be introduced to a version of the Data Science Pipeline: Import, Tidy, Transform, Visualize, Model and Communicate and then we will spend each unit focused on a particular piece of the pipeline. We will be introduced to the concept of "tidy" data and will spend the first few units of the course covering how import data from various sources and how to "tidy" various types and forms of data that we have imported. We will then cover various models aimed at both classification and regression and cover concepts such as cross validation and the curse of dimensionality. We conclude the course with a unit on deploying models and analysis as well as how to apply models to large data sets by leveraging AWS EC2 and S3. The goal of the course is to provide the student the ability to tackle a data science problem from start to finish. In addition, the course is full of interviews with industry professionals in order to provide insight into what is in demand in industry.

Course Designers

Dr. Sadler has a BS in Mathematics from Texas Tech University and graduated with a PhD in statistics from SMU in 2014. He has taught statistics and data science at the undergraduate and graduate levels at SMU since 2014. He has over 20 years of teaching experience in mathematics and statistics at the university level. In industry, Dr. Sadler worked as a statistician/java programmer at Motorola and has been (and is currently) consulting on projects in industry with clients such as Motorola, DISD (Dallas Independent School District), KadAfrica (Ugandan nonprofit), and RNDC (Republic National Distributing Company).

Course Student Learning Outcomes

Learning outcomes, or learning goals, are what you are able to do as a result of the videos, readings, instruction, course assignments, and other activities that you participate in and complete during this course. The primary learning outcomes of this course are:

1. Get a practical hands-on overview of the end-to-end data science process using industry standard tools and techniques.
2. 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.
3. Use the principles of "tidy data" to create clean data sets from messy ones using R.
4. Conduct Exploratory Data Analysis (EDA) to understand, summarize and extract insights from data sets.
5. Learn and apply basic machine learning and time series modeling techniques.
6. Learn how to deploy models with GitHub Pages and ShinyApps.io
7. Gain the ability to analyze very large data sets with AWS EC2 and S3.
8. Communicate the findings of a project in a clear, concise, and scientific manner.

This course supports, through its various synchronous, asynchronous, and other activities, broad general learning outcomes that are supported by the Master of Science in Data Science program, including:

1. An ability to design and conduct experiments that yield relevant and reproducible data.
2. An ability to manage and clean data sets.
3. An ability to apply knowledge of data analytics to explore and identify relevant information contained

within a data set.
4. An ability to function on teams using data science tools and technologies.
5. An ability to identify, formulate, and solve data science problems based on a fundamental understanding of concepts of data science.
6. An ability to communicate effectively both in oral and written form.
7. Knowledge of the broad foundational data science education necessary to understand the impact of data science solutions in a global, economic, environmental, and social context.
8. Knowledge of contemporary issues in data science.
9. An ability to use the techniques, skills, and modern data science tools necessary for data science practice.

Course Instruction Using Synchronous and Asynchronous Sessions

The course uses a combination of synchronous class sessions and asynchronous material and activities to teach students the course material and guide them through the learning process. Synchronous class sessions occur once per week during the course of the term. These sessions consist of lectures, discussions, problem solving, in-class assignments, and quizzes based on the asynchronous material, including the course video lectures, assigned activities and work, and any readings assigned. It is expected that all asynchronous material will be completed (e.g., videos viewed, assigned readings read, and assigned work completed and turned in) prior to the synchronous session associated with that material.

Course Prerequisite

A student taking MSDS 6306 must be enrolled in the Master of Science in Data Science program at SMU.

Course Textbooks

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.

Wickham, H, and Grolemund, Garrett (2017), *R for Data Science*. Sebastopol, CA: O'Reilly Media Inc.

You should be able to find the Gandrud and O'Neil books on Amazon for a very low price and a pdf copy of the Wickham text is available (endorsed by Hadley Wickham) at: https://r4ds.had.co.nz/

Technology Requirements

DS 6306 is a course taught online with both synchronous and asynchronous portions requiring the transfer of video. Students are expected to have access to a computer with reliable, high-speed internet access. Students are expected to have access to a computer with a web camera with the computer capable of running the required software to access the learning management system, read online documents, watch course videos, and participate in the synchronous classes (including being on camera). Students are also expected to have access to a reliable phone connection in order to participate in the synchronous classes.

DS 6306 course utilizes R to teach the course material. A local copy of R and RStudio is assumed.

All students enrolled in SMU have an SMU email account. Notifications from the learning management system and from the course instructor utilize your SMU email account. Students are assumed to check this email regularly.

Course Access

This course is accessible to registered students in the SMU MSDS program only. Course asynchronous material, course information, and course communications occur through the 2DS learning management system. Access to the 2DS learning management system is available at https://2ds.datascience.smu.edu/.

Students who experience technical issues with the learning management system or the Zoom classroom should contact technical support as described below.

Students will have access to only those courses and course sections in which they are currently enrolled or have been enrolled in previous terms. Access to other sections is at the discretion of the section instructor. Access to recordings of synchronous sections where the student did not participate or was not an enrolled student is prohibited to protect the privacy of the students who do attend and participate.

Communication and Technical Support

Direct communications with the instructor should be made in the manner indicated by the instructor. General questions and questions that are relevant to multiple students—that is, questions that are not specific to an individual and involve that individual's private information—should be posted on the course wall.

Technical support for the learning management system and the online classroom may be reached 24 hours a day, seven days a week via:

- *Chat Support*: Click "Live Support" in the lower right-hand corner of the 2DS screen after logging into the system to chat with a technical support representative. Chat support generally responds and engages in five minutes or less.
- *Phone*: Students should call 1-844-768-5637 (toll free) to speak with a technical support representative.
- *Email*: studentsupport@datascience.smu.edu to initiate a support request with a technical support representative.

For other questions or concerns, please contact the appropriate SMU department for your questions or concerns or send email to datascience@smu.edu.

It is the student's responsibility to ensure that all communications are received or acted upon.

Course Procedures and Policies

This course has a number of policies and procedures that students should understand and follow if appropriate. The following sections present the general course policies and procedures that students must follow. Additional policies and procedures may be given by the instructor. Please discuss as early in the term as possible with the instructor any questions or concerns that you may have regarding the course procedures and policies as defined herein or any additions made by the instructor to the course procedures and policies.

Course Grading Policy

This course consists of a number of assignments and projects that are to be completed throughout the term. Every submitted assignment is graded on a scale of 0 – 100 and contributes to the cumulative percentage for the course. Individual percentage breakdowns for each type of assignment are below. Questions regarding the grading of any assignments should be directed to the course instructor as soon as possible and in accordance with any regrading policy instituted by the instructor. This course is not graded on a curve. The required cumulative percentage needed to earn each letter grade is given in Table 1.

Table 1: Cumulative Percentage Required to Reach Each Letter Grade

Cumulative Percentage	Earned Grade
[100 – 93]	A
(93 – 90]	A-
(90 – 88]	B+
(88 – 83]	B
(83 – 80]	B-
(80 – 78]	C+
(78 – 73]	C
(73 – 70]	C-
(70 – 60]	D
< 60	F

The cumulative percentage for the course is determined by the course assignment components with their corresponding percentages defined in Table 2.

Table 2: Grade Components and Weightings of the Cumulative Percentage

Percentage of Cumulative Percentage	Component
15%	Live Session Attendance and Participation
15%	Asynchronous Video Response Questions and Discussions *(Must be completed before live session in the week assigned.)*
20%	For Live Session Assignments
5%	Project 1 EDA and Initial Meeting
20%	Project 1 Final Documentation and Presentation
5%	Project 2 EDA and Initial Meeting
20%	Project 2 Final Documentation and Presentation

Live Session Attendance and Participation (15%): It is CRITICAL to attend Live Session. This is where much of the learning is accomplished and cured into longer term memory. This is where the pieces of the puzzle are fit together in order to see the big picture. Everyone understands that life will happen and some students will need to miss occasionally. It is hard to put a number on these types of occurrences so keep your professor advised any issues that may come up that prevent you from attending live session. That said, on average, a student should not need to miss more than 2 live sessions and students who miss 2 or more live sessions will be at a significant academic disadvantage. Furthermore, when a student misses a live session, it hurts the entire cohorts experience. Live session is an amazing team learning experience where students connect the dots and learn from each other as well as the professor. If a student or students is missing, that detracts from the entire cohorts learning experience. Finally, reasons for missing a live session should be unavoidable and unexpected. A business flight is usually not a valid excuse as this usually could have been scheduled in advance. Business travel is usually not a valid excuse as this is one of the major advantages of the online delivery and have classes at night. Even many family events are not valid excuses (as much as I hate to say it) as those types of decisions should be made before enrolling in the program. Everyone is making sacrifices to advance their knowledge and missing live session will impact their learning (even if it is for a valid reason) and the learning of the cohort. Again, life will come up during the semester, (family, business, friends, emergencies) and many of these of course will warrant missing live session. The point is to communicate with your professor early if possible and work the best path forward.

Asynchronous Video Response Questions and Discussions (15%): Throughout the videos, there are various concept check questions to make sure the student understands the material before moving on to learn new material. These questions are often in the form of multiple choice or matching questions, and most are gated, which means the student must get the answer correct before moving on (but don't worry, you have an unlimited number of chances to get the questions right!). These questions may also be discussion questions in which the student will respond to a prompt and then be able to see all other students' responses after they submit their response. At that point, it is our hope that a discussion will ensue. The student has the option to keep the conversation going by responding to their peers' responses. The instructor will be checking for participation in these discussions and may even participate in the discussion themselves. Given the fact that most of the material in this course builds on the material presented before it, participation in the concept check questions and the discussions must be completed during the week they are assigned.

For Live Session Assignments (20%): This is really where a lot of the "doing" is done. It is critical to practice using, applying, and interpreting the methods and models presented in this course, and these assignments are a big part of that practice. Each week you will have a list of assignments to complete before the live session. These assignments will be completed and presented in a PowerPoint deck and submitted to the course website. They will be given a completion grade based on the thoroughness of the student's responses. In addition, the student will present their work to their peers in a breakout session, and the instructor will answer questions, facilitate discussion, and present their solution/approach to the assignment. To get full credit each week, the PowerPoint deck must be completed and submitted to the class website (no later than 24 hours before live session, verified by time stamp). This is to allow your professor to customize the Live Session to the student's particular needs and questions. In addition, each student will be presenting some or all of their work on the For Live Session Assignment in Live Session Break Outs! See Table 3 below for due dates.

Project 1 EDA and Peer Review (5%): You will have two projects in the semester and for each project you will present an exploratory data analysis and meet with the professor to discuss any issues / challenges / successes or ideas that you have with respect to the project. The grade is based on the presentation and preparation of the EDA and attendance to the meeting. In addition, you will provide a peer review of your partner up to that point. Details are in the Peer Review Rubric.

Project 1 Final Documentation and Presentation (20%): A week after the initial meeting you will have present your project live to your professor and / or record your presentation on YouTube (your professor will give you more details). You will also turn in your PowerPoint deck and any other materials required by your professor.

Project 2 EDA (5%): You will have two projects in the semester and for each project you will present an exploratory data analysis and meet with the professor to discuss any issues / challenges / successes or ideas that you have with respect to the project. The grade is based on the presentation and preparation of the EDA and attendance to the meeting.

Project 2 Final Documentation and Presentation (20%): A week after the initial meeting you will have present your project live to your professor and / or record your presentation on YouTube (your professor will give you more details). You will also turn in your PowerPoint deck and any other materials required by your professor.

A course grade of *Incomplete* (I) will be given only in the case of extraordinary circumstances that prevent the student from finishing the semester. Students must have completed at least 50% of the course with a passing grade to be eligible for an *Incomplete* grade.

Course Synchronous Session Schedule

Table 3: Course Schedule for Each Week of the Course

Week/Unit	Topic	Deliverable	Reading H = Hadley G = Gandrud O = O'Neil
1	**Git-ing Sta-R-ted!**	**For Live Session Assignment 1 due 24 Hours Before LS1**	**G: Chapter 1,2** **Suggested: G: Chapter 3** **O: Chapter 1**
2	**Visualization | ggplot2, plotly**	**For Live Session Assignment 2 due 24 Hours Before LS2**	**H: Chapter 1**
3	**Wrangle 1: Data Transformation, Factors and Exploratory Data Analysis (EDA)| dpylr**	**For Live Session Assignment 3 due 24 Hours Before LS3**	**H: Chapter 3,5**
4	**Wrangle 2: Tidy Data and Advanced Data Import (JSON, XML & APIs) | tidyverse, jsonlite, xml2**	**For Live Session Assignment 4 due 24 Hours Before LS4**	**H: Chapter 9** **G: 6.3.4** **Suggested(G: All of Chapter 6)**
5	**Wrangle 3: Relational Data and Regular Expressions | dplyr and stringr**	**For Live Session Assignment 5 due 24 Hours Before LS5**	**H: Chapter 10, 11**
6	**Machine Learning for Classification 1: KNN**	**For Live Session Assignment 6 due 24 Hours Before LS6**	**O: Pages 51 – 53 & Pages 71-81**
7	**Machine Learning for Classification 2: Naive Bayes**	**For Live Session Assignment 7 due 52 Hours Before LS7**	**O: Pages 98 - 105**
8	**Project Work Week**	**EDA for Project 1 and Partner Assessment**	**None**
9	**Project Work Week**	**Final Documentation and Presentation of Project 1**	**None**
10	**Machine Learning for Numerical Responses: Linear Regression**	**For Live Session Assignment 10 due 24 Hours Before LS10**	**O: 55 - 71**
11	**Time Series Analysis: Holt-Winters Smoothing**	**For Live Session Assignment 11 due 24 Hours Before LS11**	**O: Chapter 6**
12	**Deployment: Webpages and RShiny**	**For Live Session Assignment 12 due 24 Hours Before LS12**	**None**
13	**AWS / Big Data**	**For Live Session Assignment 13 due 24 Hours Before LS13**	**None**
14	**Project Work Week**	**EDA for Project 2**	**O: Chapter 16**
15	**Project Work Week**	**Final Documentation and Presentation of Project 2**	**None**

Grade Grievance Policy

Students are responsible for saving all graded materials as evidence in case of a discrepancy with the assigned grades. Students are responsible for ensuring that all grades are correctly reflected on the grade store. Any identified discrepancies should be brought to the attention of the instructor as soon as the discrepancy is found. Refer to the university catalogue for the university policy and process for grade grievances.

Assignment and Collaboration Policy

Data science is an inherently collaborative subject, and learning often occurs best when subjects are taught both to and from peers. Collaboration is expected to occur both in learning the course material and in performing the course work. However, each student must hand in their own work performed by themselves unless explicitly allowed by written directions given by the instructor. Collaboration means helping one another learn the material. Collaboration does not mean copying answers from one another. A good process is to ask questions and have discussions in groups and to always write up answers alone.

Assignment submissions that contain substantially the same answers shall receive a grade of zero on the first instance and a course grade of F upon a second instance. In order to mitigate potential issues and questions of similarity, peers with whom a student collaborates should be clearly identified by that student in their submissions.

Scholarly Expectations

Work submitted at the graduate level is expected to demonstrate critical and creative thinking skills and be of significantly higher quality than work produced at the undergraduate level. To achieve this expectation, all students are responsible for giving and receiving peer feedback of their work. Students are also expected to resolve technical issues, be active problem solvers, and embrace challenges as positive learning opportunities. Data science professionals must be able to teach themselves and teach others to fill in any gaps in their knowledge or to find a way of learning new material that is most conducive to their learning style. Data science professionals must also be able to work cooperatively and collaboratively with others—skills that students are expected to practice in this course. Students are expected to ask questions and ask for help when they need it and to offer help when others are in need.

Absent questions or requests for assistance, instructors must assume that students understand the material being covered and are able to complete the assignments. It is primarily through your questions that the instructor learns where the students are struggling to understand and on which topics more time needs to be spent for the students' benefit.

Timeliness

Because a 15-week term goes by quickly, assignments must be submitted by the designated due dates. Full credit cannot be earned by late or incomplete assignments. Assignments may lose up to 20% of their possible value each day late if submitted after the posted due date/time (e.g., assignments can lose all of their value at 5 days past due). When a project incorporates peer review, it is imperative that all projects be available at the beginning of the review period and that reviews are completed by the end of the review period so that others may incorporate feedback into project revisions. You will have plenty of notification and time to complete course assignments. If you know you are going to be out of town, involved in a special event/project, or unable to access a computer, please plan ahead. Also ensure that you have a backup plan ready in the event you lose power, internet access, or your available technology.

Virtual Office Hour

Your professor will keep a virtual office hour (on Zoom) once a week for one hour to answer any questions that may come up. Your professor will indicate the specific time and day of the office hour in live session and will be available for the entire hour (whether there are students present or not.).

Time Commitment

As a technical graduate level course, it is expected that students will spend between three and four hours beyond course instruction for each hour spent in instruction. MSDS courses are designed to have approximately three hours of course instruction, or contact hours, per week of the course. Therefore, it is expected that students will spend between 12 and 15 hours per week on this course.

Attendance Policy

Attendance and on-camera participation at the weekly synchronous sessions in this course are mandatory. Students with more than three (3) unexcused absences will receive a final grade of F for this course. It is the student's responsibility to notify the instructor if a synchronous session will be missed for either an excused or unexcused reason at least 24 hours, or as soon as reasonably possible, prior to the synchronous session.

Drop Policy

Refer to the university drop policy for a complete description of the drop and withdrawal policies for this course.

Campus Concealed Carry

Concealed handguns are prohibited on the Southern Methodist University campus. Pursuant to section 30.06, Penal Code (Trespass by License Holder with a Concealed Handgun), a person licensed under subchapter H, Chapter 411, Government Code (Handgun License Law), may not enter SMU property with a concealed handgun. Report violations to the Southern Methodist University Police Department by dialing 9-1-1 or 214-768-3388 (non-emergency) or 214-768-3333 (emergency).

Americans With Disabilities Act

Disability Accommodations: Students needing academic accommodations for a disability must first be registered with Disability Accommodations & Success Strategies (DASS) to verify the disability and to establish eligibility for accommodations. Students may call 214-768-1470 or visit http://www.smu.edu/alec/dass to begin the process. Once registered, students should then schedule an appointment with the professor to make appropriate arrangements. (See University Policy No. 2.4.)

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.) Failure to notify your professor prior to your absence will result in an unexcused absence and possibly a grade of zero for any assignments.

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 make arrangements with the instructor prior to any missed scheduled examination or other missed assignment for making up the work.

Academic Integrity

It is the philosophy of Southern Methodist University that academic dishonesty is a completely unacceptable mode of conduct and will not be tolerated in any form. All persons involved in academic dishonesty will be disciplined in accordance with University regulations and procedures. Discipline may include suspension or expulsion from the University.

Scholastic dishonesty includes but is not limited to cheating, plagiarism, collusion, the submission for credit of any work or materials that are attributable in whole or in part to another person, taking an examination for another person, any act designed to give unfair advantage to a student, or the attempt to commit such acts.

Example of academic dishonesty: In this course, students who have taken the course before or students from past cohorts may have the answers to some homework problems. It is considered academically dishonest to share solutions with anyone who is currently taking the course before the instructor posts the solutions for those students. It is also academically dishonest to accept solutions before a student's instructor makes them available. This falls under the category of presenting someone else's work as your own and is not only a serious violation of the SMU Honor Code but severely detrimental to the student's understanding of the material. In general, if it feels the slightest bit wrong, it probably is. The safest thing to do is to consult your instructor with any questions before action is taken.

Students caught being academically dishonest shall receive a grade of F for this course and will be referred to the SMU Honor Council for a hearing and possible sanctions including a 3-year mark on the student's transcript or expulsion. On a more positive note, our overwhelmingly main goal is to facilitate and foster each student's educational experience in order to enable them to achieve their academic and professional goals as a data scientist. Furthermore, these measures are aimed at "protecting your degree" in order to ensure that those with a Master of Science in Data Science from SMU have the utmost respect throughout academical and industry. This is our passion, and it is an amazing experience when everyone is working together and working hard. Let's get to it!

University Honor Code

When you signed your letter of intent to enroll in the MSDS program, you initialed the following statement:

> "I have read and agree to abide by the SMU Honor Code available online at: https://www.smu.edu/StudentAffairs/StudentLife/StudentHandbook/HonorCode"

The Honor Code is taken seriously at all levels within the university. Students who are found to have violated the honor code will be disciplined, which often includes expulsion from the university.

Plagiarism

Plagiarism is the "practice of taking someone else's work or ideas and passing them off as one's own" (this definition is from Google Dictionary). An example of plagiarism is as follows:

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.

The following is NOT plagiarism:

"A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables" (https://www.easycalculation.com/statistics/learn-regression.php).

The difference is in the punctuation and the attribution. Note that one can self-plagiarize. If you are using something that you wrote (e.g., a blog or a previously published article), please reference yourself.

DO NOT PLAGIARIZE. If you have any question as to what is and what is not plagiarism, ask your instructor. As a general rule, always use your own words and cite your source.

The consequence for being caught plagiarizing is to earn at least a zero on the identified assignment and may include earning a course grade of F and a referral to the SMU Honor Council for your Honor Code violation.

Best Practices for Success in the Course

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. I consider these to be a disruption (not to mention rude).

Integrity. A lot of the graded work occurs outside of class, so I expect honesty and integrity in what you submit for evaluation. Evidence of academic dishonesty will minimally result in zeros for all involved parties and perhaps University-level disciplinary action. Don't risk your 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 I haven't made a connection clear or have inadvertently left out an important point. Your question gives me 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 email to your instructor or other students. I encourage you and expect you to seek help. For questions during exams, please email the live session instructor directly.

Collaboration. I encourage the formation of study groups and collaboration with your fellow students in tackling the assignments. Working together in groups on homework is permitted, even encouraged. However, every student should write up and complete their homework independently. 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.

Have fun! Learning is meant to be a fun activity. While it can be difficult, time consuming, frustrating, and sometimes disappointing, always seek to find the fun in what you are doing and learning. The gratification from learning complex concepts and applying them to solve hard problems is what we are all striving to achieve. Having fun while we are learning and teaching others just makes the learning easier and friendships better.