DATA/MATH/COMP 180: Introduction to Data Science

Fall 2022, Section 1

(Last updated December 16, 2022; syllabus is subject to change)

Instructor: Professor Eren Bilen
Office: Rector North 1309
Email: bilene@dickinson.edu
Phone: 717-254-8162
Office Hours: Monday 4:30-5:30pm
Tuesday 4:00-5:00pm
Thursday 1:00-2:00pm
or by appointment
QRA: Ashley Doan, doana@dickinson.edu
Office Hours: Tuesday 6:30:8:30pm
Location: Rector North 1311

Class: Tome 120
Tuesday and Thursday
9:00-10:15am

Class Notes and Other Required Materials

- MATH 180: Introduction to Data Science Course Packet by Jeff Forrester, available at the Dickinson College bookstore (required)
- Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (optional)
- Reasoning with Data by Jeff Stanton (optional)
- Access to a computer to install and use R
- Course webpage: Github

Course Goals

The ability to work with and derive information from ever-increasing amounts of data will be one of the important stories of the 21st century. New analytical techniques coupled with rapidly advancing computational power continues to change way data is collected, organized, analyzed, and understood. A facility with data science techniques allows the student to bring this exciting new toolkit to bear helping to mine information from almost every area of human interest. DATA 180 provides an introduction to the core ideas of data science. Topics include data visualization, data wrangling, statistical measures of center, spread, and position, and supervised and unsupervised statistical/machine learning. Upon successful completion of the course a student will be able to:

- Organize, manipulate, and transform data using R,
- Use Github and RMarkdown to create reproducible reports and maintain a repository for version control,
- Analyze and interpret data using visualization techniques and statistical summaries,

- Employ supervised and unsupervised machine learning techniques for predictive modeling,
- Identify internal structure in data organize, manipulate, and transform data in a statistical programming environment,
- Comprehend and create basic numerical and/or logical arguments.

We will make extensive use of the R and R-Studio to generate graphical and numerical representations of data, and apply basic machine learning techniques while we interpret the results. R is a fun and useful computational tool as well as an immediate resume builder!

Course Policies

Attendance Policy: This course will be taught in person in Tome 120. Students are expected to attend all in-class meetings, which occur on Tuesdays and Thursdays from 9:00-10:30am EDT. While I will not take formal attendance, it is important for you to attend the class meetings and take notes. If you will be unable to attend a class meeting for any health-related issues or other emergencies, please contact me beforehand so that arrangements can be made.

Use of Laptops, Tablets, and Phones: Laptops and tablets are permitted for note-taking during this course. In exchange for trusting you to use these devices, I ask that you not use them as distractions. I maintain the right to change this policy for individual students or for everyone if these tools become a problem during class. Phones are not permitted and should be put away in silent mode.

Grading: Your course grade is based on two closed-book midterms, a take-home final exam, and homework assignments.

Midterm 1 (20%):	October 13
Midterm 2 (20%):	December 1
Take-home Final (20%):	by December 12, 2pm
Homework (40%):	Due dates TBA

While I will not be giving extra credit in this course, I will drop your lowest homework. I expect there to be 8-10 total assignments (depending on course pacing). Occasionally, an assignment may be weighted to count as two assignments (because of the complexity or length), this will be clearly indicated when it is assigned.

The following scale will be used to determine your final grade:

Score	Letter	GPA	Score	Letter	GPA
$93 \geq x$	A	4.0	$73 \leq x < 77$	C	2.0
$90 \leq x < 93$	A-	3.7	$70 \leq x < 73$	C-	1.7
$87 \leq x < 90$	B+	3.3	$67 \leq x < 70$	D+	1.3
$83 \leq x < 87$	B	3.0	$63 \leq x < 67$	D	1.0
$80 \leq x < 83$	B-	2.7	$60 \leq x < 63$	D-	0.7
$77 \leq x < 80$	C+	2.3	$x < 60$	F	0.0

Make-up Exams: There will be no make-up exams unless a student must be away from campus on university business or due to an emergency. The student must provide documentation. If an emergency arises, you must inform me as soon as possible. Once you provide me an official documentation related to the emergency/university business, you may schedule a make-up exam. Warning! It is absolutely essential to provide me documentation. You will receive 0 if you are unable to get an official documentation. Therefore, you should definitely not skip a test if your situation cannot produce documentation.

Homework: Homework assignments will be posted on course Github page as an R-Markdown file template on which you will insert your solutions. Due dates will be provided for each assignment. You will turn in your assignments as an R-Markdown file via a pull request from your private GitHub.com repository which is a clone of the class master repository. (You will need to set up a GitHub account if you do not already have one.) You will be sent an invitation link for each assignment. After accepting the assignment, your private repo where you will push your files will automatically be created. Prior to pushing your submission files to your repository, make sure to hit `Knit` on R-Studio, and include the `.Rmd` file in your commit. Make sure your code executes with no issues. You will receive a 20% penalty if any part of your code cannot get executed because of errors. Email submissions will not be accepted. Late assignments will not be graded.

You are encouraged to work in teams, but your submissions must be individual. It is important that you must understand and be able to explain every part of the code you are submitting. I do not want to see a bunch of copies of identical code. I do want to see each of you learning how to code these problems so that you could do it on your own. Homework assignments will require the use of R and R-Studio; you will want to obtain access to a computer with R-Studio installed during the first week of classes; R is installed in Tome 121 and various labs in Tome Hall.

Take-home Final: The course will include a final written data science assessment in lieu of a final exam that will be due Monday, December 12 at 2:00 pm EDT. Similar to many take-home data scientist interviews, you will have a fixed duration e.g., 24-48 hours to prepare your analysis. You are allowed to refer to your notes, or any online resources, help files, docs. More information will be posted later in the semester.

Getting Help

Office Hours: I will be holding three hours of office hours each week. Please see the first of page of the syllabus for my hours. I am also available by appointment. If there is a conflict and you are unable to make it to any of my hours, please feel free to send me an email. My availability outside office hours is not guaranteed, however I devote my attention fully to you during my office hours. Therefore, I highly encourage you to come to my office hours and ask questions.

Quantitative Reasoning Associate: This semester, we are fortunate to have a Quantitative Reasoning Associate (QRA) working with us. A QRA is a fellow student who completed this course in the past and will be helping us as a course facilitator and student mentor. This semester, the QRA for our course is Ashley Doan. She will be holding office hours as posted on the first page of the syllabus, including its location.

Additionally, Ashley will host study sessions before each exam, which will be announced closer to exams.

Quantitative Reasoning Center

Dickinson College provides additional support for students taking courses with quantitative content across the curriculum through the Quantitative Reasoning (QR) Center. For the fall 2021 semester, the QR Center will offer tutoring for DATA 180, in addition to general quantitative support. You are strongly encouraged to make an appointment with them. `Click here` to access the QR Center webpage.

Please visit `dickinson.mywconline.com` to make an appointment. Then, access the drop-down menu under "limit to" at the top of the scheduler and select DATA 180. This will restrict the tutor list and schedule to only those tutors approved for this course. When you make your appointment, please also paste or upload your assignment and any work that you have done.

Other Important Information

Referencing the Work of Others: When submitting your work, you must follow common-sense ground rules. External sources may only be used to improve your own understanding of the material. When you write your solutions, you should do it on your own without the direct help of any external sources, and certainly should not write down anything that you do not understand. If you do use external references, please be sure to cite them. Failure to cite references will be treated as academic dishonesty.

Respect for Intellectual Property: It is important that you be aware of and respect the intellectual property rights of others. Unless explicitly stated otherwise, all materials available on the Internet, in libraries, and elsewhere are considered intellectual property and can only be used with the permission of the owner. Specifically, with regards to this class, you should not share any of the course materials, including homework answer keys, with others, even after the completion of the course.

Statement on Disabilities: Dickinson values diverse types of learners and is committed to ensuring that each student is afforded equitable access to participate in all learning experiences. If you have (or think you may have) a learning difference or a disability – including a mental health, medical, or physical impairment – that would hinder your access to learning or demonstrating knowledge in this class, please contact Access and Disability Services (ADS). They will confidentially explain the accommodation request process and the type of documentation that Dean and Director Marni Jones will need to determine your eligibility for reasonable accommodations. To learn more about available supports, go to `www.dickinson.edu/ADS`, email `access@dickinson.edu`, call (717) 245-1734, or go to the ADS office in Room 005 of Old West, Lower Level (aka "the OWLL").

If you have already been granted accommodations at Dickinson, please follow the guidance at `www.dickinson.edu/AccessPlan` for disclosing the accommodations for which you are eligible and scheduling a meeting with me as soon as possible so that we can discuss your accommodations and finalize your Access Plan. If test proctoring will be needed from ADS, remember that we will need to complete your Access Plan in time to give them at least one week's advance notice.

SOAR: Academic Success Support: Students can find a wealth of strategic guidance by going to www.dickinson.edu/SOAR. This website for SOAR (Strategies, Organization, and Achievement Resources) includes apps, tips, and other resources related to time management, study skills, memory strategies, note-taking, test-taking, and more. You will also find information aimed to helps students "SOAR Through Academic Challenges," as well as a schedule of academic success workshops offered through Academic Advising. If you would like to request one-on-one assistance with developing a strategy for a manageable and academically successful semester, email SOAR@dickinson.edu.

Course Outline: Below is a list of topics to be covered in this course. There may be adjustments on the list during the semester depending on progress. Any adjustments will be announced and this syllabus will be updated.

- Topic 1: Introduction to Data Science, R intro
- Topic 2: Data and Variables
- Topic 4: Introduction to Data Wrangling: Tidyverse
- Topic 5: Visualization in R
- Topic 6: Unsupervised Learning: Cluster Analysis
- Topic 7: Introduction to Text Mining
- Topic 8: Introduction to Supervised (Machine) Learning

Important Dates for the Fall 2022 Semester

Last Day to Add/Drop or Change to/from Pass/Fail	Friday, September 2
Mid-Term Pause	5 pm, Friday, October 14 thru 8 AM, Wednesday, October 19
Course Request Period for Spring 2022 Semester	Monday, October 31 thru Wednesday, November 2
Thanksgiving Vacation	5PM, Tuesday, November 22 thru 8 AM, Monday, November 28
Last Day to Withdraw from a Course with a "W" grade	Tuesday, November 22
Classes End	Friday, December 9
Reading Period Days	December 10, 11

Semester Schedule

Date	Day	Topic	Pages in Notes	Homework Due
Week 1				
Aug 30	T	Ch1-2: Introduction to Data Science	1-6	
Sep 1	Tr	Ch3: Variables and Data	7-17	
Week 2				
Sep 6	T	ChX: Introduction to R and RMarkdown		
Sep 8	Tr	Ch3: Visualization	30-42	#1
Week 3				
Sep 13	T	Ch3: Visualization	43-57	
Sep 15	Tr	Ch5: Data Transformations	58-64	#2
Week 4				
Sep 20	T	Ch6: Introduction to Data Wrangling	65-73	
Sep 22	Tr	Ch6: Introduction to Data Wrangling	74-78	#3
Week 5				
Sep 27	T	Ch6: Introduction to Data Wrangling	78-88	
Sep 29	Tr	Ch7: Unsupervised Learning	88-96	#4
Week 6				
Oct 4	T	Ch7: Dissimilarity Measurements	97-106	
Oct 6	Tr	Ch7: Inter-group Proximity Measures	106-117	#5
Week 7				
Oct 11	T	Ch7: Dendograms	118-126	
Oct 13	Tr	**Midterm #1**		
Week 8				
Oct 18	T	*Fall Break: No Class*		
Oct 20	Tr	Ch7: Standardizing Variables	127-136	#7
Week 9				
Oct 25	T	Ch7: K-means Clustering	137-144	
Oct 27	Tr	Ch7: Between Group Sum of Squares	145-151	#8
Week 10				
Nov 1	T	Ch7: Similarity Measures for Binary Data	152-163	
Nov 3	Tr	ChX: Text Analysis		#9
Week 11				
Nov 8	T	ChX: Text Analysis		
Nov 10	Tr	Ch8: Supervised Learning: Classification/Pred.	185-193	#10
Week 12				
Nov 15	T	Ch10: Supervised Learning: Linear Regression	193-199	
Nov 17	Tr	Ch10: Supervised Learning: Decision Tree	200-211	#11
Week 13				
Nov 22	T	Ch10: Supervised Learning: Decision Tree, app	212-219	
Nov 24	Tr	*Thanksgiving Break: No Class*		
Week 14				
Nov 29	T	Ch10: Supervised Learning: knn reg	220-227	
Dec 1	Tr	**Midterm #2**	228-235	
Week 15				
Dec 6	T	Ch10: Probability: Logistic Regression		
Dec 8	Tr	Ch10: Probability: knn classification		#12
Dec 12	M	**Final Project, due by 5:00pm**		