DATA/COMP 200: Data Systems for Data Analytics

Spring 2023

(Last updated January 24, 2023; syllabus is subject to change)

Instructor: Professor Eren Bilen   **Class:** Tome 121
Office: Rector North 1309   Tuesday and Friday
Email: bilene@dickinson.edu   3:00-4:15pm
Phone: 717-254-8162
Office Hours: Calendly link
     Monday 9:00-10:30am
     Thursday 10:20-11:50am
     or by appointment
QRA: Elsa Hritz, hritze@dickinson.edu
Office Hours: TBA
     TBA
Location: TBA

Class Notes and Other Required Materials

- (Required) Introduction to Data Systems: Building from Python (1st edition), by Bressoud and White, Springer, 2020. Available at bookstores, or online on Amazon.
- (Optional) Ace the Data Science Interview by Huo and Singh.
- **DataCamp** The course will make use of DataCamp.com, which is an online learning platform specifically designed for data science. You will be enrolled in our course group on the DataCamp website and should receive an e-mail to your Dickinson account with a link to register on their site. Please note:
    - You will be provided with six free months of access to all the content on Data Camp. There is an incredible selection of content available, and you are encouraged to take advantage of this free access.
    - You are not obligated to purchase any services from DataCamp. At the end of the six months, you will need to purchase a subscription if you wish to continue to use the site, unless it is made available to you through another data analytics course at Dickinson.
- Access to a computer to install and use the Anaconda distribution of Python.
- Course webpage: Github

Course Overview

Data is a crucial part of the modern world, and the ability to effectively manage and analyze it is highly sought after in many fields. This course aims to provide a comprehensive introduction to working with different types of data, including tabular data, relational databases, and hierarchical data.

The course will also introduce students to popular Python libraries that are commonly used by data scientists such as Numpy, Pandas, and Matplotlib. In addition, the course will cover the topic of web scraping, which involves using automated software (such as a "bot") to extract data from websites. This is a useful skill for collecting and analyzing data from the internet, and can be applied to a wide range of applications, including but not limited to, data journalism, market research, and more.

At the end of the course, the students will be able to

- Understand the tabular data model, the relational data model, and hierarchical data model;
- Retrieve data using Structured Query Language (SQL);
- Understand the client-server model for communication;
- Acquire data from a spectrum of external systems, ranging from structured to unstructured systems using APIs and/or web-scraping;
- Manipulate unstructured data into meaningful representations;
- Utilize the Python language as it applies to data science.

Course Policies

Attendance Policy: This course will be taught in person in Tome 121. Students are expected to attend all in-class meetings, which occur on Tuesdays and Fridays from 3-4:15pm ET. 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 exams, a take-home final exam, and homework assignments.

Exam 1 (25%):	March 19
Exam 2 (25%):	April 14
Take-home Final (25%):	by May 16, 12pm (noon)
Homework (25%):	Due dates TBA

We will not have extra credit in this course; however, I will drop your lowest homework. I expect there to be 7-8 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: We will use GitHub for homework submissions and grading. Homework assignments will be posted on course Github page as a Jupyter Notebook template file which you will insert your solutions in. You will be sent an invitation link for each assignment. After accepting the assignment, your private Github repo where you will push your files in will automatically be created. Prior to pushing your submission files to your repository, make sure to hit `Kernel > Restart & Run All` in your Jupyter Notebook, and save your output into `.ipynb` and `.md` formats with `File > Download as > Notebook (.ipynb)` and `File > Download as > Markdown (.md)`. You will push both formats. You will receive a 20% penalty if your code cannot get executed because of errors; 20% penalty if your code lacks comments and explanations. Late assignments up to 24 hours will receive a 50% penalty. Anything that is turned in after 24 hours will not be graded. Email submissions will not be accepted.

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 Python and the Anaconda Package; you will want to obtain access to a computer with Anaconda installed during the first week of classes; Anaconda package is installed in Tome 121 and various labs in Tome Hall.

Why use GitHub for homeworks? Data science projects often involve working with a team of people, and GitHub makes it easy for data scientists to collaborate on projects with their colleagues. It provides a convenient way to share code and data, track changes,

and work together to resolve issues. This is called version control. Furthermore, having a presence on GitHub can be helpful for networking and finding job opportunities. You should have a portfolio of projects that you have been working on on your personal GitHub account, which can be helpful for showcasing your work and getting feedback from the data science community. I suggest taking time to learn more about GitHub this semester. GitHub is free for personal users, so you won't have to pay any hosting fees.

Take-home Final: The course will include a final data project in lieu of a final exam that will be due by Tuesday, May 16 at 12pm (noon) ET. More information will be posted later in the semester.

Writing in the Discipline (WiD): This course is part of the WiD thread in the Data Analytics Major. Data scientists regularly report their analyses to their teams and other stakeholders. The types of writing that data scientists do takes on many different forms for a variety of audiences: graphical/visual representations, briefs, memos, reports, academic papers, code, data documentation, with audiences varying from developers, internal and external technical audiences to general audiences among others. Students completing the Data Analytics Major satisfy the WiD graduation requirement by maintaining a portfolio of assignments from the courses on the WiD Thread. This course will include one WiD assignment which will be added to your WiD portfolio, focusing specifically on:

- Communicating clearly the motivation and the goals of a web scraping project, providing a clear strategy for how the data was obtained,

- Clearly and transparently documenting and annotating data sources, data wrangling, and variable descriptions.

- Describing tangibly the ways in which the data collected could be used to uncover new findings.

The data analytics major at Dickinson achieves the Writing in the Discipline (WiD) graduation requirement via a collection of targeted assignments in several required courses. Each student stores these assignments in a WiD portfolio, and they officially achieve the WiD graduation requirement when that portfolio is assessed to be satisfactory by the senior seminar instructor. WiD assignments must include a revision phase. All students (independent of their major) are required to complete the WiD assignment of this course.

Getting Help

Office Hours: Please use this `Calendly link` to schedule an office meeting. 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 stu-

dent mentor. This semester, the QRA for our course is Elsa Hritz. She will be holding office hours during the hours on the first page of the syllabus. Location is TBA.

In addition, Elsa will host a study session 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 Spring 2023 semester, the QR Center will offer tutoring for DATA 200, 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 dropdown menu under "limit to" at the top of the scheduler and select DATA 200. 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 Database Systems, Jupyter Notebooks, GitHub

- Topic 2: File systems and paths

- Topic 3: Introduction to Numpy

- Topic 4: Introduction to Pandas

- Topic 5: Visualization with Matplotlib

- Topic 6: Tabular Model

- Topic 7: Relational Models

- Topic 8: Coding with Relational Models: SQL

- Topic 9: Hierarchical Model

- Topic 10: Introduction to APIs and Web-scraping

Important Dates for the Spring 2023 Semester

Last Day to Add/Drop or Change to/from Pass/Fail	Sunday 11:59pm, January 29
Spring Vacation	5 pm, Friday, March 10 thru 8 AM, Monday, March 20
Course Request Period for Fall 2023 Semester	Monday, March 27 thru Wednesday, March 29
Last Day to Withdraw from a Course with a "W" grade	Friday, April 21
Classes End	Friday, May 5
Reading Period Days	May 6, 7, 10, 13, 14
Final Exam Days	May 8, 9, 11, 12, 15, 16