QTM 350: Data Science Computing

Quantitative Theory and Methods Department, Emory University

Professor: Davi Moreira

January 16, 2024

Course Description

This course equips students with computing skills and knowledge for data science applications. Students will gain knowledge foundations and hands-on experience with technologies such as Version Control, Project Collaboration, Data Structures and Algorithms, Database Designs, Database Management, and Cloud Computing. Prospective data scientists, statisticians, and other quantitative professionals will learn how to efficiently utilize database structures and foundational cloud services for data science by completing creative projects.

Course Website: https://davi-moreira.github.io/2024S_dsc_emory_qtm_350

Instructor and TAs

Instructor: Professor Davi Moreira

- Email: davi.moreira@emory.edu
- Office hours: Tuesdays, 9:15am - 10:15am, or by appointment
  - Zoom link in your Course Canvas Page.

Teaching Assistant: Michael Cao

- Email: michael.cao@emory.edu
- Office Hours: Mondays, 6-7 pm
  - Zoom link in your Course Canvas Page.

Learning Outcomes

By the end of this course, students will be able to: 1. Demonstrate proficiency in data science project collaboration and version control. 2. Utilize advanced data storage, manipulation, and querying. 3. High-level understanding of data structures, algorithms, and cloud computing concepts. 4. Critically navigate the emergent trends in data science computing.

Objectives

- **Conceptual Understanding**: To provide students with a foundational grasp of data structures, algorithms, and data modeling techniques pertinent to SQL and NoSQL databases and data analytics.
- **Technical Proficiency**: To equip students with practical skills in version control, Python programming, and database management using SQL and MySQL, enabling them to execute data manipulation and analysis tasks proficiently.
- **Cloud Computing Literacy**: To introduce students to the fundamentals of cloud computing, focusing on models, components, and security considerations, and offer hands-on experience.
- **Critical Integrated Learning**: To offer a holistic educational experience that combines theoretical learning with practice, ensuring students can apply their knowledge to real-world projects and foster an awareness of emerging trends in the data science computing landscape.

Course References

- **Computing Skills for Biologists**: A toolbox with basic computational skills necessary for the course.
- **Elements of Data Science**: a digital textbook by Allen Downey written in the form of Jupyter notebooks. It provides an introduction to data science in Python for students with limited programming experience.
- **Think Python**: An introduction to programming using Python.
- **Applied Computing**: Applied Computing is an online textbook. It provides an introduction to spreadsheets and SQL. To view the book, students need to register using the course name.
- **SQL & NoSQL Databases**: Models, Languages, Consistency Options, and Architectures for Big Data Management: Explores relational (SQL) and non-relational (NoSQL) databases. Covers database management, modeling, languages, consistency, architecture, and more.
- **Pro Git Book**: A comprehensive resource for learning Git, covering everything from the basics to advanced topics by Scott Chacon and Ben Straub.

- **Overview of Cloud Computing**: An introductory textbook on cloud computing by Michael Wufka and Massimo Canonico.

Additional References

- **Google Cloud Computing Foundations**: Google Cloud experts develop and maintain the Computing Foundations reference material specifically for university courses such as this to ensure it keeps pace with cloud innovation and to prepare students seeking to launch or pivot to careers in a cloud-first world.
- **Big Data: Principles and best practices of scalable real-time data systems:**: It describes a scalable, easy to understand approach to big data systems that can be built and run by a small team.
- **Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development**: This book explains how Git works and how to use it effectively. By Jon Loeliger and Matthew McCullough.
- **GitHub Learning Lab**: Offers a variety of exercises to get hands-on experience using Git and GitHub.
- **AWS Academy Cloud Foundations (login required)**: AWS experts develop and maintain the Cloud Foundations reference material to ensure it keeps pace with cloud innovation and prepares students for real-world, industry challenges. This digital textbook with accompanying labs comes in the form of a Canvas page, separate from the class Canvas page.

Assessment

Final grades will be based on:

Assignment	Percentage
Lecture Quizzes	10%
Problem Sets	40%
Final Project Proposal	5%
Final Project Submission	25%
Final Project Presentation	10%
Peer Review	10%

Lecture Quizzes

Each lecture will be accompanied by a set of questions, available on the course's Canvas page and to be completed individually. Students may complete these quizzes either during or after class, but must submit them by 11:59 p.m. on the lecture day. To accommodate the learning

process, the lowest two quiz scores will be excluded from the final grade calculation. While individual submission is mandatory, collaborative discussions are encouraged. Please note that no extensions for quiz submissions will be granted under any circumstances, ensuring fairness and consistency in assessment for all students.

Problem Sets

Problem sets aligned with each topic will be assigned to solidify and apply the concepts covered. These sets are to be collaboratively developed in groups of up to three members, emphasizing the importance of code collaboration. Consequently, individual submissions will not be accepted. The primary objective of these problem sets is to provide a practical application of the content discussed during the semester. Assignments will be distributed via Canvas and/or GitHub, and may be formatted as either a Jupyter Notebook (`.ipynb`) or Quarto documents (`.qmd`). Groups will be required to submit the complete source code of their assignments (`.ipynb` or `.qmd`). Each problem set will be meticulously evaluated, with grading based on both the accuracy and the overall quality of the work submitted. For instance, you must guarantee:

- **All code must run**;
- **Each problem set material will have its own GitHub repository**;
- **Readable Solutions**: To facilitate effective evaluation and comprehension of the coding assignments, students must adhere to the following standards for code readability:
- 1. **Comprehensive Commenting**: All code must include thorough comments. These comments are essential as they allow the Professor and Teaching Assistants to understand the purpose and functionality of the code solely through these annotations. It is crucial that the comments are clear and concise, providing insight into the logic and purpose behind each segment of code.
- 2. **Structured Code Segmentation**: Solutions should be methodically organized into distinct code chunks within Jupyter or R Markdown notebooks. For clarity on this format, refer to examples provided in class or consult with the Professor or Teaching Assistants.
- 3. **Detailed Documentation of Functions**: Every function defined by a student must be accompanied by a docstring. This documentation should clearly explain the function's purpose, describe each input argument, and outline what the function returns.

Final Project

At the heart of Data Science lies the critical skill of managing data effectively. This encompasses the ability to collect, organize, retrieve, filter, sort, and prepare data, forming the foundational steps before embarking on any comprehensive analysis. This project is your opportunity to demonstrate proficiency in applying the concepts and techniques learned throughout the course. In groups up to three members, you will be expected to showcase your competence in handling and preparing data for insightful analysis. Through this project, you will apply the principles and methodologies covered in the course.

Final Project Proposal

Each group is required to submit a concise, yet comprehensive, two-page proposal outlining their intended project. This proposal is a crucial step in ensuring the viability and academic rigor of your planned research. Prior to its development, it is mandatory for each group to arrange a 15-minute consultation meeting with me. This meeting is intended to provide guidance, clarify objectives, and ensure alignment with course goals. To make the most of this consultation, groups must submit a draft of their project ideas in advance of the meeting. Please note that these consultation meetings will be scheduled during Week 10. It is imperative to adhere to this timeline to ensure timely feedback and guidance on your proposals.

Final Project Submission

Students are provided with two distinct opportunities to submit their final projects. The first submission is madatory for all groups and must be completed before the start of the presentation week. This phase allows for a preliminary assessment of your project. The subsequent submission is optional and should occur by the course's last date. This opportunity is particularly beneficial for incorporating improvements based on peer review feedback received after the first submission. The final deliverable should be a comprehensive GitHub repository and a `.zip` file encompassing the following components:

- **Raw Source Data**: Include the raw data utilized for your project. If the data file size exceeds GitHub's limitations, it should be exclusively contained within the `.zip` file.
- **Detailed README**: Your repository should contain a `README` file providing an in-depth description of each included file, covering:
  - **Inputs**: Detail the inputs to each file, such as raw data or files containing credentials for API access.
  - **Functionality**: Clearly describe the major transformations and operations performed by the file.
  - **Outputs**: Enumerate any outputs generated by the file (e.g. cleaned datasets).

- **Data Transformation and Organization Code**: Submit all code files that are used to transform and organize your data.
- **Data Query for Analysis**: Provide the set of code files that query your dataset, rendering it suitable for addressing your chosen data analysis question.

Further details and specific requirements will be shared before Week 10 of the course.

Final Project Presentation

During the final week of the course, each group will have a 10 to 15-minute time slot to present their project. This presentation is a vital component of your project, providing an opportunity to showcase the depth of your analysis, the insights gained, and the skills acquired throughout the course. The scheduling of presentations will be randomly defined. Specific dates and the order of presentations for each group will be assigned and communicated in advance.

Further details and specific requirements will be shared before Week 13 of the course.

Peer Review

In the week of presentations, your active participation in the peer review process will be essential. Each student is expected to engage constructively in evaluating and providing feedback on their colleagues' projects. You will be asked to thoughtfully comment on various aspects of the projects, including methodology, data structure, presentation features, and the overall effectiveness of the conveyed message. This exercise is not only a crucial part of your learning experience but also an opportunity to contribute to the academic development of your peers.

Grading

Each student's final grade will be based on the following after rounding up to the nearest point:

Grade	Range
A	91% – 100%
A-	86% – 90%
B+	81% – 85%
B	76% – 80%
B-	71% – 75%
C	66% – 70%
D	60% – 65%
F	$< 60\%$

AI policy

I encourage you to use AI tools you believe will enhance your individual or group performance. Learning to use AI is a valuable and emerging skill, and I am available to provide support and assistance with these tools during office hours or by appointment.

Be aware of the following guidelines:

- You are not allowed to use AI tools during the exams.
- Providing low-effort prompts will result in low-quality outputs. You must refine your prompts to achieve desirable outcomes. Use the course knowledge for that!
- Do not blindly trust the information provided by the output. If the output contains a number, index, analysis, conclusion, or fact, assume it is incorrect and check its veracity. Any errors or omissions resulting from using the AI tool will be your responsibility. Remember, the AI tool works better for topics that you already understand.
- While AI is a tool, you must acknowledge its use. Always cite! Include a paragraph or note at the end of any document to mention that you used AI on its development.

Academic Integrity

Upon every individual who is a part of Emory University falls the responsibility for maintaining in the life of Emory a standard of unimpeachable honor in all academic work. The Honor Code of Emory College is based on the fundamental assumption that every loyal person of the University not only will conduct his or her own life according to the dictates of the highest honor, but will also refuse to tolerate in others action which would sully the good name of the institution. Academic misconduct is an offense generally defined as any action or inaction which is offensive to the integrity and honesty of the members of the academic community. **The typical sanction for a violation of the Emory Honor Code is an F in the course. Any suspected case of academic misconduct will be referred to the Emory Honor Council.**

Communication

- Check the Course Website and Canvas Page regularly to keep yourself informed with up-to-date information about the course. Also, be sure to check the course syllabus before asking any questions about the course schedule/policies.
- If you cannot attend the office hours due to conflicts with other course schedule or attending the university-sanctioned events (proof required), email the instructor at least two days in advance to set up an appointment. Note that each appointment will be 15-minutes long, and it may be done in a small group or individually. No appointments will be allowed nearing the exam dates.

- When attending virtual office hours, make sure you are in a private setting with a little to no background noise. The use of headphones is strongly encouraged. This is especially true when you are discussing private matters with the instructor.
- Do not use email for asking content-related questions, and do not use Canvas messages.
- Do not email me your private stories. Keep your email brief, and you will receive a response from me within 48 hours, except for the weekends. Similarly, if you receive an email from me, you are also expected to respond within 48 hours. Set up an individual appointment to discuss such things.
- Finally, if you are experiencing situations that negatively impact your overall student life, you should immediately contact the Office of Undergraduate Education.

Regarding absences

- If you miss a lecture for any reasons, understand that you are still responsible for the missed course materials. First, review the missed materials, then you may attend the instructor office hours to ask specific questions.
- Attendance is not monitored in lecture except on the exam dates.
- Emory College of Arts and Sciences policy states, "A student who fails to take any required midterm or final examination at the scheduled time may not make up the examination without written permission from a dean in the Office for Undergraduate Education. Permission will be granted only for illness or other compelling reasons, such as participation in scheduled events off-campus as an official representative of the University.

Access and Disability Resources

Students with medical/health conditions that might impact academic success should visit the Department of Accessibility Services (DAS) to determine eligibility for appropriate accommodations. Students who receive accommodations must contact the instructor with an Accommodation Letter from the DAS at the beginning of the semester, or as soon as the accommodation is granted. If you have DAS accommodations, you must inform the instructor after confirming that your accommodation letter is available in the DAS web portal. The instructor will respond to your email confirming which accommodations you will receive for this class. If you wish to do so, you may request an individual meeting to further discuss the specific accommodations.

Subject to Change Policy

While I will try to adhere to the course schedule as much as possible, I also want to adapt to your learning pace and style. The syllabus and course plan may change in the semester.

Schedule

Week	Topic	Title	Date
Week 01	Topic 01	Introduction, Syllabus and Set up	Jan 18
Week 02	Topic 02	Version Control	Jan 23
Week 02	Topic 02	Version Control	Jan 25
Week 03	Topic 03	Python essentials	Jan 30
Week 03	Topic 03	Python essentials	Feb 1
Week 04	Topic 03	Python essentials	Feb 6
Week 04	Topic 03	Python essentials	Feb 8
Week 05	Topic 04	API Interactions and Web Scraping	Feb 13
Week 05	Topic 04	API Interactions and Web Scraping	Feb 15
Week 06	Topic 05	Overview of Data Structure and Algorithms	Feb 20
Week 06	Topic 05	Overview of Data Structure and Algorithms	Feb 22
Week 07	Topic 06	SQL Operations and Operators	Feb 27
Week 07	Topic 06	SQL Operations and Operators	Feb 29
Week 08	Topic 06	SQL Operations and Operators	Mar 5
Week 08	Topic 06	SQL Operations and Operators	Mar 7
Week 09	-	Spring Break (no classes, no office hours)	Mar 12
Week 09	-	Spring Break (no classes, no office hours)	Mar 14
Week 10	Topic 07	NoSQL	Mar 19
Week 10	Topic 07	NoSQL	Mar 21
Week 11	Topic 08	Database Design, Structures and Management	Mar 26
Week 11	Topic 08	Database Design, Structures and Management	Mar 28
Week 12	Topic 09	Create, Populate and Manipulate databases	Apr 2
Week 12	Topic 09	Create, Populate and Manipulate databases	Apr 4
Week 13	Topic 10	Cloud Computing and Data Bases	Apr 9
Week 13	Topic 10	Cloud Computing and Data Bases	Apr 11
Week 14	Topic 10	Cloud Computing and Data Bases	Apr 16
Week 14	Topic 10	Cloud Computing and Data Bases	Apr 18
Week 15	Topic 11	Final Project	Apr 23
Week 15	Topic 11	Final Project	Apr 25