QTM 151 - Introduction to Statistical Computing II

Danilo Freire

Fall 2024

E-mail: danilo.freire@emory.edu
Office Hours: Mon-Fri Afternoon
Office: 36 Eagle Row, room 480

Web: github.com/danilofreire/qtm151
Class Hours: Mon/Wed, 16:00-16:50
Classroom: XXX

Course Description

Welcome to QTM 151! This course introduces students to data analysis and statistical computing using Python and SQL. It is ideal for those with little or no programming experience who want to develop skills for data-driven decision making.

Over the semester, we will cover version control for collaborative coding, Jupyter Notebooks for reproducible research, Python programming basics, data wrangling and merging in SQL, data visualisation, and introductions to linear modelling and time series analysis.

Students will work with real-world datasets and problems, gaining practical experience in using these tools to extract insights from data. The course aims to develop both technical skills and critical thinking needed for complex data challenges. By the end, students will be ready for advanced study in quantitative methods and data science.

Learning Objectives

By the end of this course, students will be able to:

1. Perform basic operations and write functions in Python.
2. Conduct data wrangling and manipulate data using Python libraries such as Pandas.
3. Merge and manage databases using SQL.
4. Create visualisations to effectively communicate data insights.
5. Implement linear models and understand the principles of time series analysis.
6. Use Jupyter Notebooks for reproducible research.
7. Develop problem-solving skills relevant to data analysis and statistical computing.

Prerequisites

There are no prerequisites for this course. All students are welcome to join, regardless of their prior experience with programming or data analysis. Please feel free to reach out if you have any questions about the course content or your readiness to take the class.

Materials

This course is designed to be self-contained, providing all the necessary resources and materials to succeed in mastering the core concepts. However, students are encouraged to explore the following suggested books and online courses to deepen their understanding of Python and SQL.

Suggested Books

- Python for Data Analysis by Wes McKinney
- Elements of Data Science by Allen Downey
- Automate the Boring Stuff with Python by Al Sweigart
- Python for Everybody by Charles Severance
- SQL for Data Scientists by Renee M. P. Teate

Online Courses

- Coursera: Python for Everybody Specialisation
- edX: Python Basics for Data Science
- Codecademy: Learn Python
- DataCamp: Introduction to SQL
- Coursera: SQL for Data Science

Additional Resources

- Python Documentation
- Pandas Documentation
- PostgreSQL Documentation
- SQLBolt
- DataLemur for SQL interview practice
- Github Guides

Course Information

We will meet every Monday and Wednesday from 16:00 to 16:50 at XXXXX. It is important that you read the materials before class. All information about the course is available on the course's GitHub repository at https://github.com/danilofreire/qtm151. 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. Again, please check the course repository regularly to check for updates. I will also announce any changes in class and via email.

Software

We will mainly use Python in this course. Python is a free, versatile, and powerful programming language that is widely used in data science, machine learning, and scientific computing. I recommend using the Anaconda distribution as it comes with many necessary Python libraries for data analysis, such as Pandas, NumPy, and Jupyter.

You can write your Python code in any text editor, but I recommend VS Code with the Python extension. Pycharm is also well-regarded by developers. If you are feeling adventurous, you can also use Neovim with the coc-pyright plugin. That is, if you can exit the editor. :)

We will use PostgreSQL for database management. You can download PostgreSQL from the official website. Please also install pgAdmin and the VS Code extension for PostgreSQL to interact with the database.

We will also use Jupyter Notebooks in class. Jupyter itself comes pre-installed with Anaconda, but please install the Jupyter extension for VS Code as well. We will have a hands-on session to learn how to use Jupyter effectively.

To help you get started, I have prepared a series of tutorials on how to install Anaconda, Jupyter, PostgreSQL, VS Code, and open a free educational account on GitHub. Please follow these tutorials as soon as possible to ensure that you have all the necessary tools for the course.

Office Hours

I am very flexible with office hours, but it is easier to contact me via email. Feel free to send me a message any time at danilo.freire@emory.edu, and I will likely reply within a few hours. If you prefer, you can meet me in the afternoon at my office. I am in the Department of Quantitative Theory and Methods almost every weekday. My office address is in the Psychology and Interdisciplinary Sciences Building, 36 Eagle Row, room 480. If possible, please email me before coming to ensure that no two students book the same time slot.

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 honour in all academic work. The Honour 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. Any suspected case of academic misconduct will be referred to the Emory Honour Council.

Artificial Intelligence

Students have to submit ten problem sets and complete five in-class quizzes. You are allowed and encouraged to use AI to assist with your assignments. I recommend using GitHub Copilot to generate code snippets, as it is free for students and provides good suggestions and explanations. Claude, ChatGPT, and Perplexity AI are also good tools. I am available to provide support and assistance with these tools during office hours or by appointment. However, please note that any errors or omissions resulting from the use of AI tools are your responsibility. Do not rely solely on AI to complete your assignments; you must always double-check your work. Remember to cite all sources used in your problem sets and projects, including AI tools.

Special Needs and Accessibility Services

I am fully committed to providing the necessary accommodations to ensure that all students have an equal opportunity to succeed in this course. 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 should contact me with an Accommodation Letter from the DAS at the beginning of the semester, or as soon as the accommodation is granted. If you wish to do so, feel free to request an individual meeting to further discuss the specific accommodations.

English Language Learners

Emory University welcomes students from around the country and the world, and the unique perspectives international and multilingual students bring enrich the campus community. To empower multilingual learners, an array of support is available including language and culture workshops and individual appointments. For more information about English Language Learning support at Emory, please contact the ELLP Specialists at https://writingcenter.emory.edu. No student will be penalised for their command of the English language.

Assignments and Grading Policy

Problem Sets (50%). There will be ten problem sets throughout the course. These assignments are designed to reinforce concepts covered in lectures and readings, and to provide hands-on practice with statistical programming. Problem sets will include a mix of theoretical questions and practical applications. They will be assigned regularly and must be completed individually. You may discuss your work with other colleagues as long as you do not copy entire sentences, just changing a few words. If you worked with other students, please write down their names on your assignment. Please also acknowledge any sources you used in your work, including textbooks, articles, and AI resources. *Any assignment submitted after the due date/time will automatically be graded for half points.* To accommodate unexpected circumstances, your lowest assignment grade will be automatically dropped at the end of the semester. *The same applies to in-class quizzes.* Please submit all assignments in Jupyter Notebook format (`.ipynb`) via Canvas until midnight on the due date.

Class Quizzes (30%). Students will also take five in-class quizzes throughout the semester. These quizzes will be based on the lectures from the previous weeks. They will be designed to test your understanding of the material and your ability to apply the concepts to new problems. Quizzes will be open-book and open-notes, and students have 50 minutes to complete them. You are **not** allowed to use AI tools. They are individual assessments, and students are not allowed to discuss the questions with their colleagues in class.

Final Project (20%). The final project will consist of a short report, created using Jupyter and using one of the datasets shared on the course GitHub repository. Further instructions will be provided in class. The final project will be due on the last day of class.

Grading Scale

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

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

Course Outline and Suggested Readings

The lecture notes cover all the necessary material for the course, and the weekly suggested readings are recommended for those who want to deepen their understanding of the course topics. As mentioned above, the course outline is subject to change, and I will update the syllabus if needed. Please remember to check the course GitHub repository regularly.

Module 01: Introduction to Python, Jupyter, and GitHub

Wednesday, August 28:

- Syllabus and course repository: https://github.com/danilofreire/qtm151.
- Lecture 01: Welcome to QTM 151 - Introduction.
- Course Tutorials: How to Install Anaconda, Jupyter, PostgreSQL, VSCode, and Open a Free Educational Account on GitHub.

Weekly suggested readings:

- DataCamp: SQL vs Python: Which Should You Learn? (Note: *both!*)
- Cao, L. (2017). Data Science: A Comprehensive Overview. ACM Computing Surveys (CSUR), 50(3), 1-42.
- Brady, H. E. (2019). The Challenge of Big Data and Data Science. Annual Review of Political Science, 22(1), 297-323.
- Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. Information Fusion, 50, 71-91.

Monday, September 02: Labour Day (no class)

Wednesday September 04:

- Lecture 02: GitHub Review and Introduction to Jupyter Notebooks.
- **Assignment 01:** Problem Set 01.

Weekly suggested readings:

- Microsoft: Jupyter Notebooks in VSCode.
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Chapter 01: IPython: Beyond Normal Python.

- McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (3rd ed.). O'Reilly Media. Chapter 02: Python Language Basics, IPython, and Jupyter Notebooks.
- Corey Schafer: Jupyter Notebook Tutorial - Introduction, Setup, and Walkthrough. (Note: Corey Schafer is a great Python instructor on YouTube. Check out his other videos as well.)
- Markdown Guide.

Module 02: Python Data Types and Controlling Flows

Monday September 09:

- Lecture 03: Variables and Lists.

Wednesday, September 11:

- Lecture 04: Mathematical Operations, Arrays, and Random Numbers.
- **Assignment 01 due (5%).**
- **Assignment 02:** Problem Set 02.

Weekly suggested readings:

- Real Python: Python Data Types.
- Python Documentation: An Informal Introduction to Python.
- NumPy Documentation: Quickstart Tutorial.
- Programiz: Math Operations in Python.
- Matthes, E. (2019). Python Crash Course: A Hands-On, Project-Based Introduction to Programming (2nd ed.). No Starch Press. Chapter 02.
- Severance, C. (2016). Python for Everybody: Exploring Data in Python 3. CreateSpace Independent Publishing Platform. Chapters 3-11 (Note: Read only the chapters which interest you).

Monday, September 16:

- Lecture 05: Boolean Variables and If/Else Statements.

Wednesday, September 18:

- Lecture 06: While Loops and For Loops.
- **Assignment 02 due (5%).**
- **Assignment 03:** Problem Set 03.

Weekly suggested readings:

- Real Python: Conditional Statements in Python.
- Python Official Documentation: More Control Flow Tools.
- Python Official Documentation: Compound Statements.
- Real Python: Python 'while' Loops (Indefinite Iteration).
- Real Python: Python 'for' Loops (Definite Iteration).

- W3Schools: Python While Loops.
- Sweigart, A. (2019). Automate the Boring Stuff with Python: Practical Programming for Total Beginners (2nd ed.). No Starch Press. Chapter 02: Flow Control.

Module 03: Writing and Running Functions

Monday, September 23:

- Lecture 07: Applications 1: Simulation Studies.

Wednesday, September 25:

- Lecture 08: Functions and Arguments.
- **Assignment 03 due (5%).**
- **Assignment 04:** Problem Set 04.

Weekly suggested readings:

- NumPy Random Module Tutorial.
- Python Functions.
- Real Python: Defining Functions in Python.
- Python Tutorial for Beginners: Functions.

Monday, September 30:

- Lecture 09: Global and Local Variables.

Wednesday, October 02:

- Lecture 10: **Quiz 01: Application 02 - Operations over Multiple Datasets (6%)**.
- **Assignment 05:** Problem Set 05.

Friday, October 04: (exceptionally)

- **Assignment 04 due (5%).**

Weekly suggested readings:

- Programiz: Python Variable Scope (With Examples).
- NumPy Documentation: Indexing on `ndarrays`.
- Pandas Documentation: How do I Select a Subset of a `DataFrame`?.

Monday, October 07:

- Lecture 11: Subsetting Data.

Weekly suggested readings:

- McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (3rd ed.). O'Reilly Media. Chapter 05: Getting Started with Pandas.

- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Section 3.2: Data Indexing and Selection.
- Real Python: Linear Regression in Python
- Towards Data Science: Linear Regression in Python.
- Sheppard, K. (2023). Introduction to Python for Econometrics, Statistics and Data Analysis (5th ed.). University of Oxford. Chapter 21: Statistical Analysis with `statsmodels`.

Module 04: Data Manipulation with Pandas

Wednesday, October 09:

- Lecture 12: Application 03: Linear Models.
- **Assignment 05 due (5%).**
- **Assignment 06:** Problem Set 06.

Monday, October 14: Fall Break (no class)

Wednesday, October 16:

- Lecture 13: Creating and Replacing Variables.
- **Assignment 06 due (5%).**
- **Assignment 07:** Problem Set 07.

Weekly suggested readings:

- Python for Data Analysis: Data Cleaning and Preparation.
- Codedamn: How to use the Replace function in Python.

Monday, October 21:

- Lecture 14: **Quiz 2: Application 4: Random Assignment (6%)**.

Wednesday, October 23:

- Lecture 15: Aggregating Data.
- **Assignment 07 due (5%).**
- **Assignment 08:** Problem Set 08.

Weekly suggested readings:

- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Chapter 3: Data Manipulation with Pandas.
- McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (3rd ed.). O'Reilly Media. Chapter 07: Data Cleaning and Preparation.
- DataCamp: Pandas Tutorial: DataFrames in Python.
- Real Python: Pandas Tutorial: DataFrames in Python.

Module 05: Data Manipulation with SQL

Monday, October 28:

- Lecture 16: Merging Data.

Wednesday, October 30:

- Lecture 17: Introduction to SQL.
- **Assignment 08 due (5%).**
- **Assignment 09:** Problem Set 09.
- **Instructions for the Final Project:** Final Project.

Weekly suggested readings:

- Mode Analytics: SQL Tutorial.
- Real Python: SQL Databases and SQLite.
- Khan Academy: SQL Basics. (Note: Khan Academy is a great resource for learning SQL and other programming languages).
- Coursera: PostgreSQL for Everybody.
- PostgreSQL Tutorial.
- PostgreSQL Documentation: SQL Commands. (Note: For reference only).

Monday, November 04:

- Lecture 18: **Quiz 3: Application 5: Practicing Chaining (6%)**.

Wednesday, November 06:

- Lecture 19: Import SQL Data into Python.
- **Assignment 09 due (5%).**
- **Assignment 10:** Problem Set 10.

Weekly suggested readings:

- Pandas Documentation: SQL Databases.
- Real Python: Working with SQLite Databases Using Python and Pandas.
- Mode Analytics: SQL Joins.
- PostgreSQL Documentation: Joins Between Tables.

Module 06: Time Series and Panel Data

Monday, November 11:

- Lecture 20: Merging Tables in SQL.

Wednesday, November 13:

- Lecture 21: Time Series Analysis.
- **Assignment 10 due (5%).**

Weekly suggested readings:

- W3 School: SQL Joins.
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Section 3.11: Working with Time Series.
- McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (3rd ed.). O'Reilly Media. Chapter 11: Time Series.
- Pandas Documentation: Time Series / Date functionality.

Monday, November 18:

- Lecture 22: **Quiz 4: Application 6: Practice SQL Queries (6%)**.

Wednesday, November 20:

- Lecture 23: Pivot Tables.

Weekly suggested readings:

- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Section 3.9: Pivot Tables.
- Pandas Documentation: Reshaping and Pivot Tables.
- Analytics Vidhya: Create Pivot Table Using Pandas in Python.
- Real Python: How to Create Pivot Tables With pandas.

Module 07: Text Data and Advanced Plots

Monday, November 25:

- Lecture 24: **Quiz 5: Application 8: Time Data, Panel Data, and Plots (6%)**.

Wednesday, November 27: Thanksgiving Break (no class)

Monday, December 02:

- Lecture 25: Manipulating Text Data.

Wednesday, December 04:

- Lecture 26: Advanced Plots.

Weekly suggested readings:

- Real Python: Python String Formatting Best Practices.

Monday, December 09:

- **Final Project due (20%).**