MODERN DATA STRUCTURES

Columbia University
GR5072, Spring 2022
Mon, 4:10PM-6:00PM (EST)
TBD

Instructor:	Marco Morales
Email:	marco.morales@columbia.edu
Office:	509E International Affairs Building
Office Hours:	TBD, and by appointment

TA:	TBD
Email:	

I. Overview

This course is intended to provide a hands-on tour on data extraction, data cleansing, data transformation and data loading — generically known as ETL — with a heavy emphasis on the code and packages that enable each of these stages. We will start with small data and make our way to big data. Our focus will not be on Data Engineering, but on providing the tools to make data ready data for exploration, visualization, modeling, analysis, inference or prediction — which are more properly tasks for a Data Scientist.

Each week will have simple, moderate and complex examples in class, with code to follow. Students will then practice additional exercises at home.

Disclaimers:

- The course will primarily use the **R** language for instruction. (Although code examples in **Python** will be provided wherever possible.) For that reason, some familiarity with **R** — in particular with regards to the base functions — is assumed. Knowledge of specific packages and other software tools will be built throughout the course.
- The materials and topics indicated below are a provisional roadmap and may be adapted throughout the semester.

II. Course Resources

- **Textbooks:** There are no required textbooks for this course, but you will find these to be very useful in addition to the lectures and course readings:
  - Hadley Wickham and Garret Grolemund. *R for Data Science*. O'Reilly Media, Boston, MA, 2016
  - Hadley Wickham. *Advanced R*. Chapman & Hall, Boca Raton, Fl, second edition, 2019
  - Hadley Wickham. *R Packages: Organize, Test, Document, and Share Your Code*. O'Reilly Media, Boston, MA, 2015
  - Colin Gillespie and Robin Lovelace. *Efficient R Programming*. O'Reilly Media, Boston, MA, 2016
  - Bradley C. Boehmke. *Data Wrangling with R*. Springer, New York, NY, 2016
- **Course materials:** Curated materails for each week's topic — readings, videos, and on-line resources — as well as sample code and slides will be available in the course's **GitHub** repository. Starter code for in-class exercises and homework will be available in the course's **GitHub classroom**. (Please note that these are two (2) separate repositories!)
- **Software:** The course will rely heavily on **R**, **RStudio**, and **git**. Please install them before our first class.
- **Cloud services: AWS Educate** and **Databricks Community** classrooms will be available to train you to leverage data at scale. Sign up for a **GitHub** accoount if you don't have one already.
- **Communications:** A **Slack** workspace for this course will serve as the primary means of written communication before, during and after class, where students can communicate with each other and with instructors. E-mail will be reserved for official communications only.

Instructions to get access to **GitHub classroom**, **Databricks Community**, **AWS Educate classroom**, and **Slack** will be made available for registered students.

III. Course Dynamics

Synchronous Participation vs. Asynchronous Participation: This course is designed to have a combination of synchronous and asynchronous participation to enhance your learning experience. It is our strong expectation that you will participate synchronously when required so that you can benefit fully from your peers and the live instruction. That

said, it is completely understandable that your circumstances may make that very difficult, at least on some occasions. Please alert us when that is the case. On those occassions, the synchronous portions can be done asynchronously as well. Likewise, assignments and some forms of participation can also be done asynchronously.

Expectation of Regular Participation and Utilization of Course tools: We will be monitoring student participation and completion of assignments using the corresponding tools throughout the semester. We want to make sure that students are consistently engaged, and if that becomes difficult, that students alert us to their situations.

In preparation for each class:, you should have (i) consumed all the curated materials in the course's **GitHub** repository for the week; (ii) posted any questions you have on **Slack**; and (iii) submitted homework assignments for that week using **GitHub classrom**.

During each live class: we will engage in a combination of lecture and live-workshop where we will run through code examples, troubleshoot and answer questions. Starter code for in-class exercises wil be made available through **GitHub classroom** repos. You will need to **bring a laptop to class** to follow along the coding tutorials and examples.

IV. Course requirements

The grade for this course will depend on the fulfillment of three main requirements:

(i) Attendance and Class participation (10%): students are required to attend and actively participate in class exercises and discussions. Note that you will not obtain this 10% unless you attend and actively participate on every session.

(ii) Take-home exercises (60%): homework problems will be assigned on a weekly basis with few exceptions

(iii) Final Exam (30%): the final examination will require the students to generate code to perform common ETL operations operations

Late Submission Policy: All class assignments are expected to be submitted on the due date. Please note that 10% of the maximum grade will be deducted from the score for every day the assignment is submitted late.

V. Course Outline

WEEK 1 - Introduction to R

WEEK 2 - git, GitHub and R Markdown

WEEK 3 - the `tidyverse`

WEEK 4 - Functions I: their logic

WEEK 5 - Functions II: nested and complex operations

WEEK 6 - Functions III: write your own package

WEEK 7 - Functions IV: strings and dates

WEEK 8 - ACADEMIC HOLIDAY

WEEK 9 - working with APIs

WEEK 10 - working with JSON & XML

WEEK 11 - web scraping

WEEK 12 - working with SQL

WEEK 13 - working in the Cloud

WEEK 14 - distributed data processing

WEEK 15 - tbd

Statement on Academic Integrity

Columbia's intellectual community relies on academic integrity and responsibility as the cornerstone of its work. Graduate students are expected to exhibit the highest level of personal and academic honesty as they engage in scholarly discourse and research. In practical terms, you must be responsible for the full and accurate attribution of the ideas of others in all of your research papers and projects; you must be honest when taking your examinations; you must always submit your own work and not that of another student, scholar, or internet source. Graduate students are responsible for knowing and correctly utilizing referencing and bibliographical guidelines. When in doubt, consult your professor. Citation and plagiarism-prevention resources can be found at the GSAS page on Academic Integrity and Responsible Conduct of Research.

Failure to observe these rules of conduct will have serious academic consequences, up to and including dismissal from the university. If a faculty member suspects a breach of academic honesty, appropriate investigative and disciplinary action will be taken following the Dean's Discipline procedures.

Statement on Disability Accommodations

If you have been certified by Disability Services (DS) to receive accommodations, please either bring your accommodation letter from DS to your professor's office hours to confirm your accommodation needs, or ask your liaison in GSAS to consult with your professor. If you believe that you may have a disability that requires accommodation, please contact **Disability Services** at 212-854-2388 or disability@columbia.edu.

Important: To request and receive an accommodation you must be certified by DS.