MTH 208a: Data Science Lab I (2022-2023 - Sem I)

INSTRUCTOR

Dootika Vats — *E-mail:* dootika@iitk.ac.in

Faculty Building 580 — *Web:* https://dvats.github.io/

COURSE DESCRIPTION

The course equips students with a fundamental computational learning base for modern data analysis. Focus will be on collecting, cleaning, and organizing data, and presenting clean insights via interactive web-based apps. Components on collaborative coding will be discussed via Google colab, cloud computing, and git. Programming languages used will be mainly R and Python.

PREREQUISITES

None.

SCHEDULE

Location: Room 301, New Core Lab (CC02)
Timings: MWF: 5:30pm - 6:30pm (with consultation from students, this might change)

STRUCTURE

This is a lab course with three lab hours per week. The style of the lab is worksheet based, so in most labs there will be a worksheet given to you that you that you have to work through.

Assignments will be given roughly once in two weeks that will be graded and evaluated.

COURSE WEBPAGE

The main resource page for the course is:

https://dvats.github.io/teaching/MTH208.html

Additionally, we will be using GitHub Classroom for lab worksheets, assignment submissions etc.

PROJECT

A significant component of the course will be a group project. More information on the project will be given in due course.

MARKS DISTRIBUTION

Grading will be done in the updated granular scheme. The following are the marks distribution

Assignments	20%
Project	20%
Mid-sem Exam	30%
Final Exam	30%

The final grade will be relative.

REFERENCES

There is no main text for this book, bu the following references may be helpful:

- Wickham, Hadley. Elegant graphics for data analysis. OReilly Media 35.211 (2009)

- Wickham, Hadley. Mastering shiny. OReilly Media, Inc., 2021.
- Bruce, Peter, Andrew Bruce, and Peter Gedeck. Practical statistics fordata scientists: 50+ essential concepts using R and Python. OReilly Media, 2020.
- VanderPlas, Jake. Python data science handbook: Essential tools forworking with data. OReilly Media, Inc., 2016.
- Boehmke, Bradley C. Data wrangling with R. Springer InternationalPublishing, 2016.
- Pineau, Joelle, et al. Improving reproducibility in machine learning research (a report from the Neurips 2019 reproducibility program). Journal of Machine Learning Research 22 (2021).

Course contents The course covers fundamentals tools of data science for statisticians. Much of the course will be structured to learning tools in R (and some in Python). Focus will also be on reproducibility and data ethics. Detailed course contents will be available on the course website.

Academic Honesty Academic integrity is essential to a positive teaching and learning environment. All students enrolled in the course are expected to complete coursework responsibilities with fairness and honesty. Failure to do so by seeking unfair advantage over others or misrepresenting someone elses work as your own, can result in disciplinary action.

> *Scholastic dishonesty means plagiarizing; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic grades, honors, awards, or professional endorsement; or altering, forging, or misusing a University academic record; or fabricating or falsifying of data, research procedures, or data analysis.*