# MTH 208: Data Science Lab I (2024-2025 - 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 first on logical coding skills, and simulating experiments. Further we will focuse on collecting, cleaning, and organizing data, and presenting clean insights via interactive web-based apps. We will also study code-benchmarking, and how to integrate C++ codes in R. Programming languages used will be mainly R and a little of Python.

**Prerequisites**

None.

**Schedule**

Location: Room 204H, Diamond Jubilee Building
Timings: T: 9:00am - 10:30am, Th: 2:00pm - 3:30pm

**Structure**

This is a lab course with 2 lab classes, each of length 1.5 hours. 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. The goal is to complete the whole worksheet in the 1.5 hours of the lab.

The worksheet will be mostly self-sufficient, but often I will explain solutions, concepts, and keys steps. The goal is to not take anything home to do.

**Course Webpage and GitHub**

The main resource page for the course is the course on HelloIITK. Additionally, we will be using GitHub Classroom for lab worksheets, assignment submissions etc.

Make sure you have a GitHub account before the first day of class.

**Project**

A significant component of the course will be a group project. Below will be the following components of the Project:

1. Obtain data by scraping
2. Ask questions about the data
3. Do exploratory data analysis
4. Make a Shiny app on the data analysis
5. Make a presentation
6. Write a report

More information on the project will be given in due course.

**Attendance**

A minimum of 75% attendance is required for you to be allowed to take the final exam. If you do not have 75% attendance by the end of the final week of classes, then you will deregistered from the course.

**Marks Distribution**

The following are the marks distribution

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

The final grade will be relative.

Repeat: In order to take the Final Exam, you must have 75% attendance in the course.

## Academic Honesty

IIT Kanpur is committed to stop the use of unfair means in academic activities. The following policy will be followed in this course:

Use of unfair means in course-work:

If a student cheats or uses any unfair means in the course:

1. The student will get an $F$ in the course.
2. The student will not be allowed to drop the course.
3. The name of the student will be made public to the students of the course.
4. The matter will be sent to the repespective departments' ethics/student affairs' committee.
5. A complaint will be sent to SSAC for disciplinary action.
6. An intimation will be sent to DoAA to be kept in the student's record.

What constitutes cheating:

1. Copying code from each other during assignments/exams
2. Copying any part of the code for GPT systems
3. Using plagiarised language in project reports (all group members will be failed)
4. Unapproved accessing of the internet during assignment/exams
5. If you are found talking even for a brief moment in an assignment/exam, you will be asked to leave immediately

## References

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

- Student-sourced book to prepare for this course: https://dvats.github.io/BasicRProblems/
- Wickham, Hadley. "Elegant graphics for data analysis." O'Reilly Media 35.211 (2009)
- Wickham, Hadley. "Mastering shiny". O'Reilly Media, Inc., 2021.
- Bruce, Peter, Andrew Bruce, and Peter Gedeck. "Practical statistics fordata scientists: 50+ essential concepts using R and Python". O'Reilly Media, 2020.
- VanderPlas, Jake. "Python data science handbook: Essential tools forworking with data." O'Reilly 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).