# Syllabus - Introduction to Advanced Data Science

DSC 205 - Lyon College - Spring 2025

Marcus Birkenkrahe

January 6, 2025

## 1 General Course Information

- Meeting Times: Tuesday/Thursday, 11:00-12:15 hrs
- Meeting place: Lyon Building Computer Lab 104
- Professor: Dr. Marcus Birkenkrahe
- Office: Derby Science Building 210
- Phone: (870) 307-7254
- Office Hours: By appointment only tinyurl.com/sp25-booking
- Textbook: Book of R, Davies (NoStarch, 2016), Part II, R Programming; fasteR, Matloff (GitHub, 2024).

## 2 Standard and course policies

- **Standard Lyon College Policies** are incorporated into this syllabus and can be found at: lyon.edu/standard-course-policies.
- The **Assignments and Honor Code** and the **Attendance Policy** are incorporated into this syllabus also and can be found at: tinyurl.com/LyonPolicy.
- In addition to these rules, please read and observe my guide to Using AI to code (written in Fall 2024): tinyurl.com/Using-AI-to-code.

## 3 Objectives

This course continues the journey into data science using the functional, object-oriented statistical programming language R, begun in DSC 105, "Introduction to data science". It includes calling and writing functions, conditional and looping statements. We will also explore data science using command line UNIX tools, and several packages permitting performance improvement and database connections, like Rcpp and RSQLite. If time permits, we will look at big data frameworks like Apache Spark, and at running data science experiments in containers.

## 4 Student learning outcomes

Students who complete DSC 205, "Introduction to advanced data science", will be able to:

- Import data into R, store them, and transform them for analysis
- Visualize data as part of advanced explorative data analysis
- Understand basic predictive modeling strategies and methods
- Learn how to process data on the command-line
- Learn data science optimization with C++
- Learn how to write efficient functions
- Master the infrastructure for advanced statistical computing
- Know how to effectively present assignment results
- Be ready for advanced data science courses like data visualization (DSC 302) and machine learning (DSC 305)
- Research and present a project

## 5 Course requirements

Introductory knowledge of R as taught in DSC 105 or obtained independently by completing the DataCamp online course “Introduction to R” or “fasteR: Fast Lane to Learning R!” (chapters 1-15 only, freely available on GitHub), or Davies, The Book of R (NoStarch, 2016, Part I only). Basic R concepts are repeated and practiced at the start of the term.

## 6 Grading System

You should be able to see your current grade at any time using the Canvas gradebook for the course.

| WHEN | DESCRIPTION | IMPACT |
|---|---|---|
| Weekly | Assignments | 25% |
| Weekly | Multiple choice tests | 25% |
| Monthly | Project sprint reviews | 25% |
| TBD | Final exam (optional) | 25% |

Notes:

- **To pass:** 60% of all available points.
- **Tests:** weekly online quizzes based on classroom lectures and practice.
- **Final exam:** random selection of the known test questions. **Note:** You only have to write the final exam if you want to improve your grade at the end of the course. If the final exam result is below your final grade average up to this point, it will be ignored.

## 7 Rubric

| Component | Weight | Excellent | Good | Satisfactory | Needs Improvement | Unsatisfactory |
|---|---|---|---|---|---|---|
| Participation and Attendance | 0% | Consistently attends and actively participates in all classes. | Attends most classes and participates in discussions. | Attends classes but participation is minimal. | Frequently absent and rarely participates. | Rarely attends classes and does not participate. |
| Programming assignments | 50% | Completes all assignments on time with high accuracy (90-100%). | Completes most assignments on time with good accuracy (80-89%). | Completes assignments but with some inaccuracies or delays (70-79%). | Frequently late or incomplete assignments with several inaccuracies (60-69%). | Rarely completes assignments and shows minimal understanding (0-59%). |
| Tests | 25% | Demonstrates thorough understanding and application of concepts (90-100%). | Shows good understanding with minor errors (80-89%). | Displays basic understanding with some errors (70-79%). | Limited understanding with several errors (60-69%). | Minimal understanding and many errors (0-59%). |
| Final Exam (Optional) | 25% | Demonstrates comprehensive understanding and application of course concepts (90-100%). | Shows strong understanding with minor errors (80-89%). | Displays adequate understanding with some errors (70-79%). | Limited understanding with several errors (60-69%). | Minimal understanding and many errors (0-59%). |

::

## 8 Grading Table

| Percentage | LETTER GRADE |
|---|---|
| 100% to 89.5% | A (very good) |
| < 89.5% to 79.5% | B (good) |
| < 79.5% to 69.5% | C (satisfactory) |
| < 69.5% to 59.5% | D (passed) |
| < 59.5% to 0% | F (FAILED) |

## 9 Schedule and Workload

For **important dates**, see the 2024-2025 Academic Calendar at: catalog.lyon.edu/202425-academic-calendar.

**Workload** (estimated):

- Time in class: 48 hrs.
- Time outside of class: 42 hrs.
- Time for tests [1 hrs/test]: 14 hrs.
- Time for home assignments [2 hrs/pgm]: 28 hrs.
- Total number of hrs in term: 90.
- Weekly workload (outside of class): 5.625 (2.625)

## 10 Course Outline

For **important dates**, see the 2022-2023 Academic Calendar at: catalog.lyon.edu/202223-academic-calendar

Listed are DataCamp assignments. Besides these topics, we will review additional material, and there will be additional programming assignments losely aligned with the four parts of the course:

1. Programming in R
2. Data Processing in Shell
3. Writing Functions for data analytics
4. Optimizing data science code

Weekly schedule:

- Week 1: Intermediate R: Control flow
- Week 2: Intermediate R: Loops
- Week 3: Intermediate R: Functions
- Week 4: Intermediate R: `apply` functions
- Week 5: Intermediate R: Utilities
- Week 6: Data Processing in Shell: Downloading data
- Week 7: Data Processing in Shell: Data cleaning
- Week 8: Data Processing in Shell: Database operations
- Week 9: Data Processing in Shell: Data pipeline
- Week 10: Writing Functions: How to write a function
- Week 11: Writing Functions: All about arguments
- Week 12: Writing Functions: Return values and scope
- Week 13: Optimizing R Code with Rcpp: Writing and benchmarking
- Week 14: Optimizing R Code with Rcpp: Functions and control flow
- Week 15: Optimizing R Code with Rcpp: Vector classes
- Week 16: Project Presentations and Final Review