


Course syllabus

DA5020 – Collecting, Storing, and Analyzing Data

Instructor	Ben Tasker
Email	b.tasker@northeastern.edu
Messaging & Virtual Office	Use Teams




Required Textbooks

- Grolemund, G., Wickham, H. **R for Data Science: Visualize, Transform, Tidy and Import Data**. O'Reilly
(available online at <http://r4ds.had.co.nz/>  [\(http://r4ds.had.co.nz/\)](http://r4ds.had.co.nz/) or in print on Amazon).
- Schedlbauer, M., Jain, Y. Durant, K. **Data Collection, Integration and Analysis** (<https://northeastern.instructure.com/courses/157427/files/22232165?wrap=1>). Unpublished manuscript attached.
- R in Action by Robert I. Kabacoff **R_in_Action.pdf** (<https://northeastern.instructure.com/courses/157427/files/22232195?wrap=1>).
- Introduction to Statistical Learning V.2: **ISLRv2.pdf** (<https://northeastern.instructure.com/courses/157427/files/22232194?wrap=1>).

Optional Textbooks

- R Quick Reference Info** (<https://northeastern.instructure.com/courses/157427/files/22232217?wrap=1>)

Tools

- R Base**  [\(https://www.r-project.org/\)](https://www.r-project.org/) / base installation of R; required for R Studio Desktop
- R Studio Desktop**  [\(https://rstudio.com/\)](https://rstudio.com/) / R programming environment for Windows, MacOS, and Linux
- R Studio Cloud**  [\(https://rstudio.cloud/\)](https://rstudio.cloud/) / cloud version running in Linux VM for most browsers

Course Prerequisites

An undergraduate course in statistics is required. Exposure to modern programming languages is helpful.

Course Description

In this course students will learn how to build large-scale information repositories of different types of information objects so that they can be selected, retrieved, and transformed for analytics and discovery, including statistical analysis. Students will become knowledgeable about traditional approaches to data storage and how they can be applied alongside modern approaches that use non-relational or NoSQL architectures. Through case studies, readings on background theory, and hands-on experimentation, students will have the opportunity to learn how to select, plan, and implement storage, search and retrieval components of large-scale structured and unstructured information repositories. In particular, students will be conversant in the tools and techniques used to assess and recommend efficient and effective large-scale information storage and retrieval components that provide data scientists with properly structured, accurate, and reliable access to information needed for investigation.

Course Outcomes

After completing this course, students will be able to:

- Classify information and data storage approaches based on object type and retrieval requirements
- Select an appropriate information storage structured depending on object type and analysis goals
- Plan an information repository for data analysis and discovery
- Collect data from a variety of sources using R
- Clean and transform data into effective storage structures in R
- Discuss how NoSQL databases can be applied to store and retrieve unstructured data
- Complete simple implementations of structured and unstructured data repositories using R
- Use SQL to store and retrieve data from relational databases
- Transform data objects into representations that can be transferred to data analysis platforms
- Distinguish between storage needs for statistical and non-statistical analysis of data
- Outline tiered information architectures for efficient data retrieval and search
- Apply knowledge from case studies to select, plan, and implement information repositories
- Analyze data through exploratory visualization
- Identify outliers and transform data for predictive analytics
- Construct simple predictive analytics models using time series forecasting, multiple linear regression, logistic regression, and k-NN
- Evaluate and tune predictive models
- Work within a reproducible analytics pipeline that follows CRISP-DM

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

- Completion of eleven short assignments involving scripting in R, data acquisition, database storage and retrieval, and model construction.
- Completion of three substantial practicums. **A minimum average of 70% on the practicums is required to pass the course regardless of other grades achieved in the course.**

Each practicum group must also submit a 10-15 minute video reviewing the assignment, methodology, analysis, and results for each rubric requirement. Based upon the rubric, each group will state how many points they believed they earned for each rubric element and why. If the evaluator agrees with the groups self grading, that is what the group will earn for that rubric requirement. The evaluator reserves the right to grade higher or lower based upon the factors outlined above. Remember that it is difficult to earn a 100% as it would mean that your analysis is perfect. All group members must participate in the video, and each member must explain at least 1 question. Name the video file DA5020.PX.GroupX.mp4, where X is your practicum number/group number.

- Completion of an online final exam. **A minimum grade of 60% on the final exam is required to pass the course regardless of other grades achieved in the course.**

Each assessment contributes toward your overall grade as follows:

- Assignments - 30%
- Practicums - 60%
- Final Exam - 10%

Course Methodology

Each week, students are expected to:

1. Review the week's learning objectives
2. Complete all assigned readings
3. Complete all lessons for the week
4. Participate in online discussions and collaborative code walks
5. Complete and submit all assignments and assessments by the due date

Participation/Discussion Board

Interaction occurs primarily through Teams and the discussion area in Canvas. The discussion area in Canvas is for an assigned discussion. Teams is for any other questions.

Each week, students are expected to:

- Post their questions in Teams.
- Respond or comment on other students' posts as appropriate.
- Join the online interactive recitation sessions (optional). They will be recorded and reviewing the recording is required weekly

Communication

We are here to help you learn. You should start by posting in Teams. Please avoid contacting the instructor or a particular TA directly, as you will get a much faster response when all of the instructional staff have a chance to see and answer your question. In an online course, you can't look around the classroom and see that other people are also confused, and posting your question will help your classmates realize that they are not the only ones who feel uncertain.

Submission of Work

All work for the course is expected to be completed by the due date and time and must be submitted in the Assignments folder on Canvas. **No late submissions past three days or emailed submissions or are accepted.** Late submissions past the due date are subject to a 10% reduction in grade for each day late with specific rules stated for each graded assessment item.

In the Assignments or Submissions folder, click on the correct Assignment link to view an assignment. Attach your files or documents along with explanatory comments and click Submit to turn them in.

Once an assignment has been graded, students will be able to view the grade and feedback by clicking on My Grades. Assignments are to be done in a professional manner at a graduate level -- points will be deducted for any work that is not at an appropriate level of quality. All code must properly run from start to end and all libraries, packages, files must be installed and/or loaded as part of the submitted code.

Bonus points will not be awarded for late submissions

Online Live Conversations ("Class"/"Lecture")

To enhance the learning experience, there is a live, online recitation each week. The time is announced in Canvas. While participation is optional, it is encouraged. The sessions are generally recorded for later playback. You may not share the link to the class with anyone outside the class.

You may not post any lewd or objectionable messages, use any objectionable or lewd screen names, change your background to a lewd or objectionable image, upload or share any files, or in any other way disrupt the virtual class. Any violation of these rules is grounds for a report to OSCCR, reduction in final course grade, or dismissal from the course.

Recommendations or Letters of Reference

Should you ask me to provide a recommendation or reference for co-op, doctoral program application, or for some other purpose, please ensure that I know you. That means, come to office hours frequently, talk to me about your work or research, show me what you can do. I cannot write recommendations for students who I do not know well enough to judge their ability or potential.

Social Media Connections

I do not connect with students on social media, such as Facebook. I do consider connections on LinkedIn

Grading

The course requires a 70% overall score to pass the course with the additional requirement that the average grade for the practicums also requires at least 70% and at least 60% for the final exam to pass the course. In other words, you cannot pass the course without getting passing grades on the practicums and the final. If you get a 50% on the three practicums and get a 100% on everything else you will not pass. Of course, getting a 100% on the project and a 0% on everything else also means you will not pass as you did not meet the minimum passing grade for the course.

Grades are a reflection of your performance in class, and they cannot be adjusted because you want or need a higher grade, or because you came close to a cutoff.

95% and above	A
90% - 94.99%	A-
87% - 89.99%	B+
84% - 86.99%	B
80% - 83.99%	B-
77% - 79.99%	C+
73% - 76.99%	C
70% - 72.99%	C-
Less than 70%	F

Academic Integrity Policy

The University views academic dishonesty as one of the most serious offenses that a student can commit while in college and imposes appropriate punitive sanctions on violators. Here are some examples of academic dishonesty. While this is not an all-inclusive list, we hope this will help you to understand some of the things instructors look for. The following is an excerpt from the University's policy on academic integrity; the complete policy is available in the Student Handbook.








- *Cheating* – intentionally using or attempting to use unauthorized materials, information or study aids in an academic exercise; this includes submitting work of another student or work prepared for another course
- *Fabrication* – intentional and unauthorized falsification, misrepresentation, or invention of any data, or citation in an academic exercise
- *Plagiarism* – intentionally representing the words, ideas, or data of another as one's own in any academic exercise without providing proper citation, including code fragments from websites such as stackoverflow
- *Unauthorized collaboration* – instances when students submit individual academic works that are substantially similar to one another; while several students may have the same source material, the analysis, interpretation, and reporting of the data must be each individual's independent work.

- *Participation in academically dishonest activities* – any action taken by a student with the intent of gaining an unfair advantage
- *Facilitating academic dishonesty* – intentionally or knowingly helping or attempting to violate any provision of this policy
- *Impersonation* – working on behalf of another students or allowing someone else to represent a student online, in discussion groups, for presentation, in any communication with the instructor, or in exams
- *Multiple Submissions* – submitting the same or substantially the same work in two courses

Any incident of academic misconduct will result in a 0 for the graded item, a report to OSCCR, and a full two-letter reduction in the final course grade, except in cases of impersonation or fabrication. In those two cases the student will receive a report to OSCCR with a recommendation for dismissal from the University and an automatic failing grade of F.

Course summary:

Date	Details	Due
Sun, 10 Sep 2023	 Assignment 1 (https://northeastern.instructure.com/courses/157427/assignments/1893321)	due by 23:59
Sun, 17 Sep 2023	 Assignment 2 (https://northeastern.instructure.com/courses/157427/assignments/1893324)	due by 23:59
Sun, 24 Sep 2023	 Assignment 3 (https://northeastern.instructure.com/courses/157427/assignments/1893325)	due by 23:59
Sun, 1 Oct 2023	 Assignment 4 (https://northeastern.instructure.com/courses/157427/assignments/1893326)	due by 23:59
Sun, 8 Oct 2023	 Practicum 1 (https://northeastern.instructure.com/courses/157427/assignments/1893331)	due by 23:59
Sun, 15 Oct 2023	 Assignment 5 (https://northeastern.instructure.com/courses/157427/assignments/1893327)	due by 23:59
Sun, 22 Oct 2023	 Assignment 6 (https://northeastern.instructure.com/courses/157427/assignments/1893328)	due by 23:59
Sun, 29 Oct 2023	 Assignment 7 (https://northeastern.instructure.com/courses/157427/assignments/1893329)	due by 23:59

Date	Details	Due
Sun, 5 Nov 2023	 <u>Practicum 2 — Hospital Nursing Intervention Pilot Program</u> (https://northeastern.instructure.com/courses/157427/assignments/1893333)	due by 23:59
Sun, 12 Nov 2023	 <u>Assignment 8</u> (https://northeastern.instructure.com/courses/157427/assignments/1893320)	due by 23:59
Sun, 19 Nov 2023	 <u>Assignment 9</u> (https://northeastern.instructure.com/courses/157427/assignments/1893330)	due by 23:59
Sun, 26 Nov 2023	 <u>Assignment - Bonus</u> (https://northeastern.instructure.com/courses/157427/assignments/1893323)	due by 23:59
Sun, 3 Dec 2023	 <u>Assignment 10</u> (https://northeastern.instructure.com/courses/157427/assignments/1893322)	due by 23:59
Sun, 10 Dec 2023	 <u>Practicum 3</u> (https://northeastern.instructure.com/courses/157427/assignments/1893334)	due by 23:59
Sat, 16 Dec 2023	 <u>Final Exam</u> (https://northeastern.instructure.com/courses/157427/assignments/1893319)	due by 23:59