

DATA1030 Fall25 S01 Hands-on Data Science

Staff

Instructor: Andras Zsom (reach by email at andras_zsom@brown.edu (mailto:andras_zsom@brown.edu))

HTA: Asher Labovich (dsi1030headtas@lists.brown.edu (<mailto:dsi1030headtas@lists.brown.edu>))

TAs: Sarah Bao, Junhui Huang, John Farrell, Penggao Gu, Shiyu Liu, Yicen Ye, Zhaocheng (Harry) Yang, Huaxing Zeng, Devraj Raghuvanshi (dsi1030tas@lists.brown.edu (<mailto:dsi1030tas@lists.brown.edu>))


Course schedule and location

We meet Mondays and Wednesdays at 3pm to 4:20pm in Friedman Hall (90 George street), Room 102. This is a lecture room for 100 students so it might get crowded during the shopping period. The first lecture is on September 3rd. All lectures will be recorded and the recordings posted shortly after the lecture ends in the Media Library section of the course's canvas site (link on the left).

Important links

Course forum: [Ed Discussion](https://edstem.org/us/courses/83381/discussion)  (<https://edstem.org/us/courses/83381/discussion>) (also on the left)

Course github repository: [DATA1030-Fall2025](https://github.com/BrownDSI/data1030-fall2025)  (<https://github.com/BrownDSI/data1030-fall2025>)

All assignments need to be submitted to [Gradescope](https://www.gradescope.com/courses/1089585)  (<https://www.gradescope.com/courses/1089585>) (also on the left)

Course description

By the end of the course, you'll be able to perform all aspects of the data science pipeline: data cleaning, handling missing data, exploratory data analysis, visualization, feature engineering, machine learning modeling, interpretation, presentation in the context of real-world datasets. Fundamental considerations for data analysis are emphasized (the bias-variance tradeoff, training, validation, testing). Classical models and techniques for classification and regression are included (linear regression and logistic regression, support vector machines, decision trees, ensemble methods). We use the Python data science ecosystem (pandas, matplotlib, scikit-learn, XGBoost,

SHAP mostly).

Learning goals

Students will be able to complete data science projects from the initial question to final presentation. Students will be able to clean the data, explore the data visually, apply regression and classification models, discuss the advantages and disadvantages of particular techniques, and interpret and present their findings.

Course structure and grading

Assessment in DATA1030 is based on weekly homework assignments (45% weight), one individual project (40% weight), class attendance based on quiz completion (5%) and a final exam (10% weight). Your final grade is calculated using the following equations:

homework grade = (your points) / (max points)

individual project grade = (your points) / (max points)

class attendance grade = $\min(1, (\text{your points})/(\text{max_points}) + 0.1)$ -- you can miss 10% of lectures with no penalty

exam grade = (your points) / (max points)

final grade = $0.45 * (\text{homework grade}) + 0.4 * (\text{individual project grade}) + 0.05 * (\text{class attendance grade}) + 0.1 * (\text{exam grade})$

If your final grade is above 0.9, you'll receive an A. Thresholds for B and C will likely be at 0.8 and 0.7, respectively.

The weekly homework assignments will be a mix of coding problems and short essay questions. The goal of the homework assignments is to practice the material we covered during class but also to introduce new concepts and techniques we couldn't cover in class.

The project will entail building a machine learning pipeline from scratch and apply the ideas developed in the course to a dataset of your choosing. It is highly recommended to select a dataset on a topic you enjoy and are interested in. There will be a midterm presentation around mid October and a final report and final presentation early December. The TAs will serve as mentors and will guide you through the project.

The final exam is in-person and it will take place during the last lecture of the term (tentatively

on December 3rd or 8th). The final presentations will take place after the final exam (during the weeks of December 8th or 15th, tentatively). More accurate dates will be announced by mid-November the latest.

The grades will be closed and submitted latest by the end of the day on December 19th.

Course policies

If your course participation is disrupted for a short term (no more than a week or two) due to illness, grief, any other special circumstance, official documentation such as a doctor's note or a Dean's note is required. Unless your circumstance prohibits you from doing so, documentation must be provided to the instructor within 24 hours of being granted. Long term accommodations that might span multiple weeks must be requested from [SAS \(https://studentaccessibility.brown.edu/\)](https://studentaccessibility.brown.edu/).

The course policies outlined below apply under regular circumstances.

The weekly homeworks need to be submitted on gradescope by the deadline. Late submission on gradescope is possible no later than three days after the original deadline. Once the late deadline passes (i.e., the gradescope submission is closed), we do not accept submissions anymore.

You have a total of 6 late days (144 hours) for the term. If your total late submission time exceeds 144 hours, you will lose 10% of the points per late day for the assignment.

Due dates and deadlines for the final project will be released with at least three weeks of lead time. There is no late submission policy for the final project. Those deadlines need to be met.

Schedule (preliminary, subject to change)

Lecture 1: Course and ML intro

Lecture 2: Working with data

Lecture 3: Working with data continued

Lecture 4: Data visualization

Lecture 5: Splitting iid data

Lecture 6: Preprocessing

Lecture 7: Linear models

Lecture 8: Regularization

Lecture 9: Cross-validation and interpretability

Lecture 10: Time series splitting

Lecture 11: Group-based splitting

Lecture 12: Evaluation metrics, part 1

Lecture 13: Evaluation metrics, part 2

Lecture 14: Non-linear ML models, part 1

Lecture 15: Non-linear ML models, part 2

Lecture 16: Non-linear ML models, part 3

Lecture 17: Imbalanced classification problems

Lecture 18: Missing values in ML, part 1

Lecture 19: Missing values in ML, part 2

Lecture 20: Interpretability, part 1

Lecture 21: Interpretability, part 2

Lecture 22: Notes on deployment

Lecture 23: Course review

Lecture 24: Final exam

Academic Integrity

Plagiarism and cheating are serious offenses and are more harmful to you, the student, than to the university. Please refer to the [Brown University Academic and Student Conduct Codes](https://graduateschool.brown.edu/academics-research/rules-regulations/academic-code) (<https://graduateschool.brown.edu/academics-research/rules-regulations/academic-code>) for details regarding Brown University's policy on academic integrity and penalties for violating the academic code.

"Academic achievement is evaluated on the basis of work that a student produces independently. A student who obtains credit for work, words, or ideas which are not the products of his or her own

effort is dishonest."

Misunderstanding the Code will not be accepted as an excuse for dishonest work. If a student has questions on any aspect of the Academic Code as it relates in a particular course or as it may be interpreted in practice, they should consult the instructor in the course or one of the deans of the Graduate School so as to avoid the serious charge of academic dishonesty.

We encourage you to discuss homework assignments with other students in the class. You may work out solutions together on whiteboards or other media, but you are not allowed to take away any written notes, diagrams, or code from joint work sessions. Emails, text messages, and other forms of virtual communication also constitute "notes" and should not be used for discussing problems. The issue is that direct recording of other people's ideas makes it very difficult for you to express the answer independently. People have a tendency to write down what was shared in an undigested form.

If we find evidence of cheating or plagiarism, all students involved can decide between two options:

- all students involved receive 0 points for the assignment,
- if the student(s) feel they are unjustly accused, we bring the case to the Academic Standing Committee where it is on the instructor to provide proof of academic code violation. If the instructor can convince the committee, the student(s) will receive 0 points for the assignment and a permanent record will be entered in the student's internal academic folder. Otherwise, the case is dropped and no points will be deducted.

Repeat offenders or students involved in serious offenses might receive a lower final grade (B instead of an A, or C instead of a B) or an NC (no credit).

A note on GenAI tools

Code and text completion tools like github copilot, chatGPT, and Gemini became widely known. Think of these tools as advanced versions of stackoverflow and wikipedia. While we cannot ban the usage of these tools because there is no way to meaningfully enforce such a ban, we advise you to not use them while you are learning. Code and text completion tools are not perfect and you won't be able to recognize the non-perfect solutions unless you understand the material first. Relying too much on code completion tools also comes with a danger: most data science job interviews have live coding and technical components with no tools allowed. If you can't code without copilot, ChatGPT, and stackoverflow, you might find it difficult to succeed in job interviews.

If a student decides to use ChatGPT or any other AI tool for course assignments, they must

acknowledge and thoroughly document their use of the tool. The student must: 1) cite the tool used (e.g. include the link to the chat), 2) include an explanation of how the tool was used for the assignment, and 3) document the student's own contribution vs. the contribution of the tool. All assignments will be graded based on the student's original ideas – students risk losing credit if the documentation provided is insufficient to determine the student's original contributions.

Accessibility and inclusion

Brown University is committed to full inclusion of all students. Please inform me early in the term if you may require accommodations or modification of any of course procedures. You may speak with me after class, during office hours, or by appointment. If you need accommodations around online learning or in classroom accommodations, please be sure to reach out to [Student Accessibility Services \(SAS\)](https://studentaccessibility.brown.edu/) (<https://studentaccessibility.brown.edu/>) for their assistance (seas@brown.edu, 401-863-9588). Undergraduates in need of academic advice or support can [contact an academic dean in the College](https://www.brown.edu/academics/college/academicadvisinghours) (<https://www.brown.edu/academics/college/academicadvisinghours>) by emailing college@brown.edu (<mailto:college@brown.edu>). Master's students may contact Dean Perkins (Assistant Dean of Student Affairs, janae_perkins@brown.edu (mailto:janae_perkins@brown.edu)). PhD students may contact one of the deans in the Graduate School by emailing graduate_school@brown.edu (mailto:graduate_school@brown.edu).