Intermediate Data Analytics

DAT 301 | Fall 2026 | 3 credits

Instructor: Alex Marsella
Email: amarsella@berry.edu
Office: Green Hall 215
Office Hours: MWF 11a-12p, W 1p-4p
Meeting: MWF, 12p-12:50p
Classroom: Green 109 (Computer Lab)

Catalog Description

This course introduces students to applied descriptive and statistical analysis in a modern analysis package such as R. Students will learn how to do exploratory data analysis, data visualization, linear regression, logistic regression, and propensity score matching. Students will learn the theoretical underpinnings of statistical inference and uncertainty, how to prevent bias in statistical estimates, and the basics of causation.

Course prerequisites: DAT 101 and ((MAT 111 and 145/201) or BUS 175) or CI.

What This Course Is About

This is an applied data-analysis course. You will spend a substantial amount of time working with data in R, deciding how to approach an analytical problem, interpreting statistical output, and explaining what the evidence does and does not show. Most major applications and practicals use real public datasets. A smaller number of exercises use purpose-built synthetic data when controlling the data-generating process makes the lesson clearer.

Generative AI has changed what it means to be competent with data. Producing a few lines of R code is no longer, by itself, strong evidence that you understand an analysis. For that reason, this course will place less emphasis on memorizing syntax and more emphasis on being able to:

- formulate a useful empirical question;
- inspect, clean, reshape, and validate data;
- choose an appropriate analytical strategy;
- read, modify, debug, and verify R code;
- interpret models and statistical uncertainty correctly;
- evaluate predictive performance rather than relying only on in-sample fit;
- distinguish descriptive, predictive, and causal claims; and
- communicate and defend your analytical decisions.

You will still be expected to know the fundamentals of R without AI assistance. At the same time, you will also learn how to use AI productively and skeptically when it is permitted.

Relationship to Causal Inference

This course leads directly into Causal Inference. DAT 301 will give you the data-management, regression, modeling, and statistical interpretation skills that the next course assumes. If you like this course, you will

probably love Causal Inference.

During the final portion of this course, we will begin asking a different kind of question: when does an observed relationship tell us something about cause and effect? We will introduce counterfactual reasoning, confounding, randomized experiments, observational research design, and propensity score matching. The goal is not to cover the full causal-inference toolkit here, but to make sure you enter the next course understanding why prediction, association, and causation are different problems.

Course Learning Objectives

By the end of the semester, a successful student should be able to:

1. Import, clean, reshape, join, validate, and document datasets in R.
2. Produce informative visualizations and descriptive statistics appropriate to a substantive question.
3. Use clustering and related exploratory techniques to identify meaningful patterns in data.
4. Estimate and interpret simple and multiple linear regression models.
5. Work with categorical variables, interactions, predictions, residuals, confidence intervals, and basic model diagnostics.
6. Estimate and interpret logistic regression models and predicted probabilities.
7. Evaluate classification and prediction using train/test data, confusion matrices, and appropriate performance measures.
8. Recognize overfitting, data leakage, and the difference between in-sample model fit and out-of-sample performance.
9. Use propensity score matching to improve balance on observed pre-treatment covariates and evaluate the quality of a match.
10. Recognize confounding, selection, reverse causality, and other threats to causal interpretation.
11. Read unfamiliar code, identify errors, audit AI-generated analyses, and verify important analytical choices.
12. Produce a concise, reproducible analysis and explain why its conclusions are defensible.

Required Materials and Software

1. de Mesquita, Ethan Bueno, and Anthony Fowler. *Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis*. Princeton University Press.
2. Huntington-Klein, Nick. *The Effect: An Introduction to Research Design and Causality*. A free online version is available at `https://theeffectbook.net/`.
3. **R and RStudio Desktop.** Both are free. You are expected to have a working installation on your computer. Using the lab computers is encouraged.
4. **Berry OneDrive.** You will use your Berry OneDrive account to store and synchronize your DAT 301 project folders. This allows you to move between your own laptop and campus computers without changing file paths or losing work.

5. **Quarto.** We will introduce Quarto during the semester so that you can combine code, output, figures, and written interpretation in one reproducible document.
6. **Generative AI.** Some assignments in this course permit or encourage the use of generative AI for coding, debugging, and analytical support. You will need access to a generative AI tool on AI-ASSISTED assignments. A free tool is sufficient; you are not required to purchase a paid AI subscription. Specific rules for AI use will be stated on each assignment.

Additional readings, datasets, tutorials, and reference materials will be provided through Canvas.

R, OneDrive, and Reproducible Work

A basic skill in this course is being able to move an analysis from one computer to another without breaking it. Most substantive assignments will therefore be organized as self-contained RStudio Projects.

At the beginning of the semester, create a folder in your Berry OneDrive called `DAT301`. When an assignment is distributed as a project folder, place the entire folder inside `DAT301` and open the included `.Rproj` file before beginning your work.

Unless an assignment says otherwise:

- use **relative file paths**; do not hard-code paths such as `C:/Users/YourName/Downloads/...`;
- do not use `setwd()` in work submitted for this course;
- your code should run from top to bottom without errors;
- your analysis should not depend on objects that happen to exist in your R environment;
- do not rely on a saved `.RData` workspace to recreate your results;
- file names and data sources should be clear;
- figures and tables should be understandable without inspecting raw code; and
- you are responsible for keeping working copies of all files you submit.

If you switch between a laptop and a campus computer, save your work, close RStudio, and allow OneDrive to finish synchronizing before opening the project on the other computer.

Early in the semester, many assignments will be submitted as `.R` scripts. As the course progresses, you will begin producing reproducible reports using Quarto (`.qmd`) and rendered HTML output.

"It worked on my computer" is not the same thing as reproducible analysis.

Generative AI Policy

Assignments in this course will be marked with one of three AI rules. The rule attached to the assignment is part of the assignment instructions.

AI-FREE

You must complete the work without generative AI assistance. This includes tools such as ChatGPT, Claude, Gemini, GitHub Copilot, Posit Assistant, and similar systems that generate code, explanations, or written responses.

AI-FREE assessments are completed in person. For Coding Checks, Practical 1, and the Final Exam, I will provide a specific reference sheet through Canvas; that sheet is the only reference material permitted. Reading/Concept Quizzes are closed-resource and do not permit a reference sheet. Personal notes, previous assignments, internet searches, package documentation/help pages, and other outside resources are not permitted unless an assignment explicitly says otherwise. Ordinary RStudio features such as syntax highlighting, standard autocomplete, and error messages are allowed.

AI-FREE assignments exist because you still need an independent foundation. You should be able to inspect data, write basic R code, fit common models, and interpret standard output without outsourcing those tasks.

AI-ASSISTED

You may use generative AI for tasks such as:

- generating or revising code;
- debugging;
- explaining an error message;
- suggesting alternative models or visualizations; or
- helping you think through an analytical approach.

However, you are responsible for every line of code and every analytical claim you submit. AI-generated output is not automatically correct, and "the AI told me to do it" is not a justification for an analytical choice.

When requested, you must include a short AI audit or disclosure describing what AI helped with and how you verified the result. I may also ask you to explain any part of your submitted work.

AI-LIMITED

AI may be used only for the purposes specifically listed in the assignment. For example, you may be allowed to use AI to clarify a difficult passage in a research paper but not to write the submitted critique.

Using AI in a way that violates the assignment's stated category is an academic-integrity violation.

Attendance

Berry College policy is that attendance is mandatory. This course builds on itself, and many class meetings involve coding, data analysis, or an in-class assessment that cannot be recreated simply by reading the slides afterward.

Several assignments—including Independent Coding Checks, Reading/Concept Quizzes, Applications, and Data Analysis Practicals—will take place during class. The lowest Independent Coding Check and the lowest AI-Enabled Application are automatically dropped. If you miss one Coding Check or one Application, the zero can serve as your dropped score; these built-in drops are intended to provide flexibility for an ordinary absence, technical problem, or unusually poor performance. Missed Coding Checks and Applications therefore ordinarily cannot be made up.

Reading/Concept Quizzes and Practicals are not covered by the automatic drop policy. School-sanctioned absences are excused when you provide official documentation sufficiently in advance for us to arrange an alternative. Documented medical emergencies or other circumstances recognized by College policy may also qualify for a make-up. It is your responsibility to contact me as soon as reasonably possible.

Late Work

Unless an assignment explicitly says otherwise, work is due at the time listed in Canvas.

In-class Coding Checks, Reading/Concept Quizzes, Applications, and Practicals must be submitted by the end of the stated assessment period. Late in-class work is not accepted except in the case of an approved absence.

For out-of-class assignments, late work is generally not accepted unless you have made arrangements with me in advance or have a documented circumstance covered by College policy. Do not wait until after a deadline to tell me that you had a problem with your computer, file format, or software.

Student Engagement and Participation

For each credit hour of registered coursework, Berry College expects students to commit approximately two hours of outside engaged learning per week. Because this is a three-credit-hour course, you should expect approximately three hours of class time and at least six additional hours per week for reading, coding practice, and assignments.

The amount of time required will vary. Students who are new to R should expect some assignments to take longer early in the semester.

Assessment and Grading

There are **500 points that count toward the final grade**. More than 500 raw points are offered because the lowest Coding Check and lowest Application are automatically dropped.

Category	Points
Independent Coding Checks (6 × 12; lowest 1 dropped)	60
AI-Enabled Applications (6 × 24; lowest 1 dropped)	120
Reading/Concept Quizzes (3 × 20)	60
Research Critiques (2 × 20)	40
Data Analysis Practical 1	60
Data Analysis Practical 2	60
Final Exam	100
Total	**500**

Independent Coding Checks — 60 points

There will be six **AI-FREE** coding checks worth 12 points each. Each Coding Check is designed for 20 minutes of class time. The lowest Coding Check grade is automatically dropped, so the best five count for 60 points.

Coding Checks assess material that has already been introduced and practiced; they do not assess material being introduced for the first time later that day.

The checks will cover, respectively:

1. R foundations;

2. data wrangling;
3. visualization;
4. exploratory analysis and reproducibility;
5. regression; and
6. model validation.

These are not designed to reward memorization of obscure syntax. You will be asked to perform standard analytical tasks that an intermediate data analyst should be able to complete without outside assistance. The Canvas assignment for each Coding Check will link the specific reference sheet permitted for that assessment.

AI-Enabled Applications — 120 points

There will be six in-class applications worth 24 points each. These are AI-ASSISTED unless otherwise stated. The lowest Application grade is automatically dropped, so the best five count for 120 points.

Applications may be completed individually or with one partner. Partners will submit one analysis and ordinarily receive the same grade. Both students are responsible for understanding and being able to explain the entire submission; dividing the assignment so that each person understands only part of the analysis is not sufficient. You may work with different partners on different Applications.

You will receive a dataset and a substantive problem rather than a sequence of commands to reproduce. Applications will ask you to decide what to do, produce an analysis, interpret the result, and evaluate its limitations. **Applications cover the material listed for that day.**

The six applications focus on:

1. messy-data auditing and cleaning;
2. visualization for decision-making;
3. exploratory analysis and segmentation;
4. regression modeling and interpretation;
5. prediction and model validation; and
6. causal claims and research design.

A technically functioning analysis is necessary but not sufficient for full credit.

Reading/Concept Quizzes — 60 points

Three 20-point Reading/Concept Quizzes will be completed in class using Respondus LockDown Browser. Each quiz takes place in the first 20–30 minutes of class and is AI-FREE and closed-resource: no notes, readings, reference sheets, internet access, or other outside materials are permitted. The quizzes emphasize concepts, interpretation, and understanding of the assigned readings rather than memorization.

Research Critiques — 40 points

You will complete two short research critiques, each worth 20 points: one focused on a randomized study and one focused on an observational study.

Rather than simply summarizing the paper, you will identify the research question, the quantity the authors are trying to estimate, the source of the comparison, the most important threat to the analysis, and the evidence that most increases or decreases your confidence in the conclusions.

Unless otherwise stated, these are **AI-LIMITED**. AI may be used to help clarify difficult material in the paper, but the submitted evaluation must represent your own reasoning.

Data Analysis Practical 1 — 60 points

Practical 1 is an **AI-FREE**, in-class analysis of a dataset you have not seen before. You will be expected to inspect the data, perform appropriate analysis, produce useful output, and interpret your results under time constraints.

This practical measures your independent analytical competency. The Canvas assignment will link the reference sheet that is permitted during the practical.

Data Analysis Practical 2 — 60 points

Practical 2 is an **AI-ASSISTED**, in-class analysis of an unseen dataset.

You may use AI to help with code, debugging, or possible approaches, but your grade will depend in part on whether you use those tools critically. You will be expected to verify your analysis, compare reasonable alternatives, explain your conclusions, and complete a short audit of how AI affected your work.

This practical measures your ability to work effectively with modern analytical tools without surrendering analytical judgment.

Final Exam — 100 points

The Final Exam will be administered during the College's scheduled final-exam period. It is **AI-FREE** and will emphasize interpretation, analytical judgment, model output, prediction, propensity score matching, research design, and concepts that connect the different parts of the semester.

The Canvas assessment will link the specific reference sheet permitted during the exam. The course builds cumulatively, so although the exam will emphasize higher-level interpretation rather than memorizing syntax, earlier material remains relevant.

What Earns an A?

Getting code to run is not the same thing as demonstrating intermediate-level data-analysis skill.

Work in the A range generally does all of the following:

- uses an appropriate analytical approach;
- executes the analysis correctly;
- interprets statistical output accurately and in context;
- checks whether the analysis is credible;
- recognizes meaningful uncertainty and limitations;
- communicates the result clearly; and
- demonstrates analytical judgment rather than simply completing requested commands.

Work can be technically competent and still fall in the B or C range if it lacks interpretation, verification, or analytical judgment.

Grading Scale

Percent	Letter Grade	Points
93–100%	A	465–500
90–92.9%	A-	450–464
87–89.9%	B+	435–449
83–86.9%	B	415–434
80–82.9%	B-	400–414
77–79.9%	C+	385–399
73–76.9%	C	365–384
70–72.9%	C-	350–364
67–69.9%	D+	335–349
60–66.9%	D	300–334
0–59.9%	F	0–299

There is no automatic rounding up at the end of the semester.

Academic Integrity

Academic integrity is fundamental to the work of this course. Students are expected to submit original work and to follow the collaboration, resource-use, and AI rules stated for each assignment.

Academic dishonesty includes, but is not limited to, cheating, unauthorized collaboration, plagiarism, fabrication, submitting another person's work as your own, and using generative AI in a manner that is prohibited by an assignment.

A violation may result in a zero on the work in question and may be referred or sanctioned further under Berry College policy.

Berry College's academic-integrity information is available at:

`https://www.berry.edu/academics/student-academic-resources/academic-integrity`

Accessibility and Academic Support

Students who need academic accommodations should work with Berry College's Office of Accessibility Resources. Students are encouraged to begin the accommodation process as early as possible.

`https://www.berry.edu/academics/student-academic-resources/accessibility-resources/`

The Academic Success Center provides peer support, academic consultations, and other resources for Berry students:

`https://www.berry.edu/academics/student-academic-resources/student-success/asc`

Writing consultants are also available through the Berry College Writing Center. These resources are appropriate for data analysts too: clear writing and clear thinking are closely connected.

Tentative Course Schedule

The course is organized around a recurring sequence:

SHOW IT ⟶ HANDS-ON PRACTICE ⟶ GRADED APPLICATION

I will first introduce and demonstrate a method or analytical idea. We will then work with it together in R. After you have seen and practiced the material, you will complete an application in which you are expected to make more of the analytical decisions yourself. Short AI-free Coding Checks are layered into this sequence to make sure you are also developing independent R competency.

Because each normal class meeting is only 50 minutes, Coding Checks use approximately 20 minutes of class, Reading/Concept Quizzes use approximately 20–30 minutes, Applications use the full class period, and Practicals use the full class period.

In the schedule below, **dMF** refers to de Mesquita and Fowler, *Thinking Clearly with Data*, and **NHK** refers to Huntington-Klein, *The Effect*. Readings are due before class on the date listed.

Intro, Statistical Review, and R Foundations

Data Preparation, Visualization, Exploration, and Reproducibility

Linear and Logistic Regression, Prediction, Model Fit, and Interpretation

Causal Reasoning, Research Design, and Propensity Score Matching

Wk	Date	Reading Due	Format	Topic	Due / Assessment
1	Aug. 24	—	Orientation	Syllabus and Introduction to Data Analytics	—
1	Aug. 26	dMF Ch. 2	SHOW	Review of Statistics and R	—
1	Aug. 28	NHK Ch. 3	HANDS-ON	R Foundations and Describing Data	Ungraded Statistics Diagnostic
2	Aug. 31	—	SHOW	Data Wrangling in R	**Coding Check 1**
2	Sept. 2	—	HANDS-ON	Data Wrangling in R	—
2	Sept. 4	—	SHOW	Cleaning and Validating Data	—
3	Sept. 7	—	NO CLASS	Labor Day	—
3	Sept. 9	—	HANDS-ON	Cleaning and Wrangling in R	**Coding Check 2**
3	Sept. 11	—	APPLICATION	Cleaning and Wrangling in R	**Application 1**
4	Sept. 14	—	SHOW	Data Visualization	—
4	Sept. 16	—	HANDS-ON	Data Visualization in R	**Coding Check 3**
4	Sept. 18	—	APPLICATION	Data Visualization in R	**Application 2**
5	Sept. 21	—	SHOW	Clustering Analysis	—
5	Sept. 23	—	HANDS-ON	Clustering Analysis in R	—
5	Sept. 25	—	APPLICATION	Clustering Analysis in R	**Application 3**
6	Sept. 28	—	SHOW	Reproducible Analysis with Quarto	**Coding Check 4: Exploration & Reproducibility**
6	Sept. 30	—	HANDS-ON	Data Analysis Workshop and Practical Review	Practical preparation
6	Oct. 2	—	PRACTICAL	Independent Data Analysis	**Data Analysis Practical 1 (AI-FREE)**
7	Oct. 5	dMF Chs. 4–5	SHOW	Correlation and Linear Regression	—
7	Oct. 7	—	HANDS-ON	Linear Regression in R	—
7	Oct. 9	Quiz 1: dMF Chs. 4–5	QUIZ	Multiple Linear Regression	**Reading/Concept Quiz 1**

Wk	Date	Reading Due	Format	Topic	Due / Assessment
8	Oct. 12	—	HANDS-ON	Multiple Linear Regression in R	—
8	Oct. 14	dMF Ch. 6, pp. 94–102	SHOW	Model Fit and Statistical Inference	—
8	Oct. 16	—	HANDS-ON	Interactions and Regression Diagnostics	**Coding Check 5**
9	Oct. 19	—	NO CLASS	Fall Break	—
9	Oct. 21	dMF Ch. 6, pp. 103–111	HANDS-ON	Regression Workshop	—
9	Oct. 23	—	APPLICATION	Regression Modeling and Interpretation	**Application 4**
10	Oct. 26	—	SHOW	Logistic Regression	—
10	Oct. 28	—	HANDS-ON	Logistic Regression in R	—
10	Oct. 30	—	QUIZ	Prediction and Model Validation	**Reading/Concept Quiz 2**
11	Nov. 2	—	HANDS-ON	Prediction and Model Validation in R	—
11	Nov. 4	—	SHOW	Overfitting and Data Leakage	—
11	Nov. 6	—	HANDS-ON	Model Validation in R	**Coding Check 6**
12	Nov. 9	—	APPLICATION	Prediction and Model Validation	**Application 5**
12	Nov. 11	dMF Chs. 9–10	SHOW	Correlation and Causation Revisited	—
12	Nov. 13	—	HANDS-ON	Controlling for Confounders	Causal reasoning practice
13	Nov. 16	dMF Ch. 11 (Quiz 3 covers Chs. 9–10)	QUIZ	Randomized Experiments	**Reading/Concept Quiz 3**
13	Nov. 18	RCT paper posted on Canvas	DISCUSSION	RCT Paper	**Research Critique 1**
13	Nov. 20	—	PRACTICAL	AI-Assisted Data Analysis	**Data Analysis Practical 2 (AI-ASSISTED)**
14	Nov. 23	—	NO CLASS	Extended Thanksgiving Break	—
14	Nov. 25	—	NO CLASS	Thanksgiving Holiday	—
14	Nov. 27	—	NO CLASS	Thanksgiving Holiday	—
15	Nov. 30	NHK Ch. 14; observational-study paper on Canvas	SHOW	Propensity Score Matching	**Research Critique 2**
15	Dec. 2	—	HANDS-ON	Propensity Score Matching in R	PSM practice
15	Dec. 4	—	APPLICATION	Propensity Score Matching in R	**Application 6**
—	Dec. 7–11	—	FINAL EXAM	Final Exam	**Final Exam (AI-FREE)**

The Fall 2026 academic calendar designates November 23–24 for remote asynchronous work. DAT 301 will not hold class or assign an asynchronous class session on Monday, November 23. Thanksgiving Holiday is November 25–27; the last day of classes is December 4; and final examinations are December 7–11.

A Final Note

This course is intended to be challenging. You are not expected to know everything immediately, and getting stuck while coding is a normal part of learning data analysis. The goal is to become increasingly independent over the semester.

Ask questions when you do not understand something, use office hours, and practice enough that you can tell the difference between code that merely looks plausible and an analysis you would be willing to defend.