DS-100: Data Speak Louder than Words

Course Description

In this course we will introduce you to three fundamental perspectives for reasoning with data: critical thinking, inferential thinking, and computational thinking. All three of these perspectives are integral to the data-driven research processes that are common in data science, thus allowing you to learn and practice how you can make and test hypotheses, and construct or deconstruct arguments that are rooted in data.

We will first use public data sets (both curated or scraped) focused on socially-relevant themes (e.g., public health, education, and environment) to model and understand real-world phenomena. We will focus on using model summarization, data visualization, and model-based simulations to interpret and communicate our understanding of these real-world phenomena as well as the potential for bringing these derived models to bear on real-world questions and applications (e.g., comparing different policies).

Particular emphasis will be placed on exposing you to and developing your appreciation for the principles underlying data mining and machine learning methods, including regression, classification and clustering, and the statistical concepts of measurement error and prediction. We will teach you critical concepts and skills in computer programming (Python), linear regression, and statistical inference. We will also delve into dilemmas surrounding data analysis such as balancing individual privacy and social utility.

This course uses a learning model where students use large language models (LLMs), commonly referred to as AI or GenAI, as learning partners for concept exploration while developing authentic data science competency through guided practice and individual verification.

Course Structure & Expectations

This course follows a flipped model. You explore concepts before class, then class time is used for verification, application, and feedback.

- **Lectures (Tue/Thu):** New concepts, worked examples, and framing for verification.
- **Discussion/Lab (Fri):** Individual verification and guided practice. Default is no-AI or limited-AI unless stated.
- **Office Hours:** Support, clarification, and help with makeups per policy.

GenAI Exploration (GAIEs) workflow ramps up after onboarding.

Hub Learning Outcomes

Social Inquiry I (SO1)

Learning Outcome #1: Students will identify and apply major concepts used in the social sciences to explain individual and collective human behavior including, for example, the workings of social groups, institutions, networks, and the role of the individual in them.

We will employ hands-on analysis of real-world datasets, including curated economic data, data scraped from digital collections, social networks, and more. In this context, the course will expose you to social and legal issues surrounding data analysis, including issues of privacy and data ownership, and will highlight the many ways in which data could be used (or misused).

In this course we will be looking at data from multiple vantage points. For example, by looking at data characterizing COVID-19 infections, hospitalizations, deaths, vaccinations, we will be able to differentiate between phenomena (e.g., correlations) identified at the macro scale (federal and state) versus those identified at the micro scale (cities and communities) and draw conclusions or make statements supported with evidence from data (e.g., impact of socioeconomic background).

We will encourage you to apply what you learn on societally-relevant case studies of your choice (e.g., case studies similar to those presented in https://www.callingbullshit.org/) by applying the tools and techniques covered in class to analyze data sets in order to support or debunk hypotheses.

Digital/Multimedia Expression (DME)

Learning Outcome #1: Students will be able to craft and deliver responsible, considered, and well-structured arguments using media and modes of expression appropriate to the situation.

We will use real data to understand relationships and patterns while also introducing critical concepts and skills in computer programming and statistical inference. In order to build your arguments, you will use multimodal data analysis and visualization in ways that are appropriate to the task at hand. This will include:

- Generation and interpretation of scatter plots, histograms, bar charts, and box plots
- Making predictions using simple regression
- Characterizing data quality and communicating associated uncertainties
- Establishing confidence in reproducible predictions
- Reaching defensible conclusions about real-world questions

These skills will be taught and evaluated through learning logs and discussion-section/in-class activities, as well as demonstrations and projects

Learning Outcome #2: Students will be able to demonstrate an understanding of the capabilities of various communication technologies and be able to use these technologies ethically and effectively.

As part of DS-100, we will introduce you to multiple forms of data visualization and presentation, including histograms, scatterplots, word clouds, heat maps, infographics, etc. Each one of these forms of communication can be particularly effective (or even misleading) in certain settings. For example, the choice of different scales (e.g., absolute vs relative change) on an axis could over or under-emphasize particular conclusions from the data.

Given the multitude of sources from which the data is collected, you will be exposed to proper ways of handling the data. For example, to preserve the privacy of individuals or communities in a large data set, and be introduced to the use of randomization techniques (blurring the data). As another example, to deal with the scale of data it may be necessary to only consider/analyze a subset of all observations. In that context, we will introduce you to various ways in which selection bias may influence conclusions you may be able to reach with implications on reproducibility.

Learning Outcome #3: Students will be able to demonstrate an understanding of the fundamentals of visual communication, such as principles governing design, time-based and interactive media, and the audio-visual representation of qualitative and quantitative data.

We will teach you how to use Python to organize and manipulate data in tables, and to visualize data effectively. Furthermore, you will be able to use computation to help your data tell a story through fundamental principles and methods of data visualization. The data used throughout this course will include longitudinal data (time series over long-time scales), geospatial data (data overlaid on apps), or both. These modalities will offer you different ways to interact with the data. For example, with time series data, you will be able to develop animations to show how phenomena or inferences may evolve over time. As another example, with geospatial data sets, you will be able to develop animations or heat maps that may project different messages/narratives based on the level of aggregation (e.g., achieved by zooming in and out).

In all of the above learning outcomes, we note that some of your work products will be in the form of multimedia reports, in which data visualization is coupled with narratives or video clips. For example, in a report on deforestation due to climate change, you may add audio or video clips to demonstrate change over time. You may also include your own narration to supplement and/or add texture to the graphs, heatmaps, etc.

Research and Information Literacy (RIL) Learning Outcomes

We will teach you critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. In discussion sections, you will work in small teams, working under the supervision of the teaching fellow to frame a question or test a hypothesis using a set of potential data sources. The key phases of that process are the exploration and identification of relevant data sets, the formulation and reformulation of the questions based on the identified data, the development of a set of data processing/analytics steps leading to an answer, and the interpretation and/or validation of the answer. To a large extent, going through these phases mirrors the six steps of the data science research process.

Books & Tools

- **Programming Environment:** Python with industry-standard libraries (numpy, pandas). We will help set up your environment so everyone can access the same tools.

Course Platforms

- **Blackboard Ultra** – Primary hub for announcements, weekly schedule, policies, and resources. The Blackboard calendar is the official student-facing schedule and authoritative if other sources conflict.
- **Gradescope** – Submit and receive feedback on learning logs, in-class work, and projects; also for regrade requests.
- **Piazza** – Questions, discussion, and project info. Use Piazza instead of emailing the teaching team.
- **TerrierGPT** – Our GenAI exploration platform (terriergpt.bu.edu) with access to OpenAI, Anthropic, Amazon, and open-source models.
  - Provided credits should be sufficient. If not, budget up to **$60** for a commercial subscription. If this would be a hardship, contact the instructor and alternatives will be arranged. *No student is penalized for inability to pay.*

Assignments & Grading

Grade Distribution

Category	Weight
GenAI Explorations	20%
Understanding Demonstrations & In-Class Activities	40%
Projects (Mini 10%, Final 30%)	40%

How Grading Works

- **GAIEs (completion-only):** Autograded for completeness as Complete or Missing. No drops apply. All GAIEs must be submitted by the last day of classes to receive a final grade. Missing one or more GAIEs after the last day of classes may result in an Incomplete, even if other coursework is strong. GAIEs are expected by the posted weekly deadlines to support preparation; late GAIEs submitted later still count for completion but do not restore preparation value for that week.
- **Understanding demonstrations & in-class activities (rubric/points):** Graded with points/rubrics. The lowest 10% of scores in this category are dropped automatically.. Late submissions are allowed with a flat -10% penalty.
- **Projects (rubric/points):** Graded on clarity, accuracy, reproducibility, ethics, and communication.
- **Practice vs. verification:** Logs are practice and preparation; demonstrations are individual verification of understanding. Accordingly, GenAI use is encouraged for logs but prohibited during demonstrations and oral verification.

Projects

Mini-project (10% of course grade)

- Team-based in small groups formed by course staff.
- Released at the beginning of Week 7, due at the end of Week 8.
- Oral verification in Week 9 (details posted in Blackboard).

Final project (30% of course grade)

- Team-based in small groups formed by course staff.
- Team deliverables (percent of final project grade):
  - Single proposal: 5%
  - Full proposal: 10%
  - Final report: 40%
  - Final code: 45%

- Individual multipliers apply to the team project score:
  - Oral interview multiplier: below expectations 0.8, meets expectations 1.0, exceeds expectations 1.1.
    - Meets: contributed meaningfully to the work and understands all aspects of the code and report.
    - Exceeds: demonstrates deep understanding.
    - Below: uncomfortable with their own work, GenAI-assisted work, or teammate work.
  - Peer review multiplier: below expectations 0.8, meets expectations 1.0 (no exceeds).
    - Peer review uses a 1-5 scale; 4+ = meets expectations.
  - Final project score = team score * oral multiplier * peer multiplier.

Timing & Late Work

- **GAIEs:** Due by the posted weekly deadlines (e.g., before Tuesday class) to support preparation. Late GAIEs may be submitted up to the last day of classes for completion, but late submissions do not restore preparation value for that week's graded activities. No drops apply.
- **Demonstrations & in-class activities:** Must be completed by the deadline. Late submissions are allowed with a flat -10% penalty. Drops apply (lowest ~10% dropped).
- **Projects (mini + final):** Deliverables may be up to 48h late (-10%); not accepted after. Oral verification is scheduled in Week 9 for the mini-project and during the final project period (details in Blackboard).
- **Excused absences:** Approved excused absences receive an automatic 7-day extension before normal late rules apply.
- **Late adds:** Students who add the course late have a 10-day grace window from their official add date before normal rules apply.
- **Completion requirement:** To receive a final grade, all GAIEs must be submitted by the last day of classes, and graded work must be submitted by the end of the instructional period; missing required components may result in an Incomplete.

How to Succeed

1. **Engage actively with learning logs.** Use GenAI tools to explore ideas, then reflect in your own words.

2. **Engage consistently in class.** Our model depends on in-class verification and discussion. Attendance isn't graded separately; evidence comes from submitted artifacts.
3. **Come prepared for Friday sections.** You'll demonstrate your understanding individually—practice concepts beforehand.
4. **Use office hours.** We welcome you for questions about concepts, projects, or GenAI strategies.
5. **Ask questions on Piazza.** It's the fastest way to get help from peers and staff.
6. **Build community.** Use ERC tutoring, writing support, and peer study groups (outside individual assessments).

Integrity & GenAI Policy

We follow BU's Academic Conduct Code and the CDS GAIA Policy, adapted for this course. Discussing ideas is encouraged; copying is not. Always cite collaborators, data, libraries, and GenAI use.

GenAI Zones

- **Green (Encouraged):** Learning logs, concept exploration, debugging help, project brainstorming.
- **Yellow (Allowed with care):** Project execution (analysis must be student-driven), writing drafts, study groups.
- **Red (Prohibited):** In-class demonstrations, oral verification interviews, or any assessment explicitly marked "individual verification."

Violations—especially GenAI use in the Red Zone—will be treated as academic misconduct.

Tone of conduct: Disagreement is welcome; disrespect is not. We aim to model the community we want for our University and industry.

Support & Accessibility

- **Disability accommodations:** Contact the Office for Disability Services (617-353-3658, access@bu.edu). Share your letter privately with the instructor.
- **Academic & well-being support:** ERC tutoring, writing support, and BU Student Wellbeing resources (see Blackboard).

Regrades

Submit requests via Gradescope within **7 days** of score release, identifying the specific error. Scores may go up, down, or remain unchanged; decisions are final.