AGENDA - INTRODUCTION TO DATA SCIENCE

DSC 105 Data Structures Lyon College Fall 2025

Marcus Birkenkrahe

September 15, 2025

Contents

1 Week 4: Expo & More R (Sep 15, 17, 19)

☐ Career expo - why you should go

☐ 1st Sprint Review (project proposals) graded with feedback

☐ 2nd Sprint Review deliverable is a literature review (Oct 10)

☐ This week: Continue the DataCamp lesson review

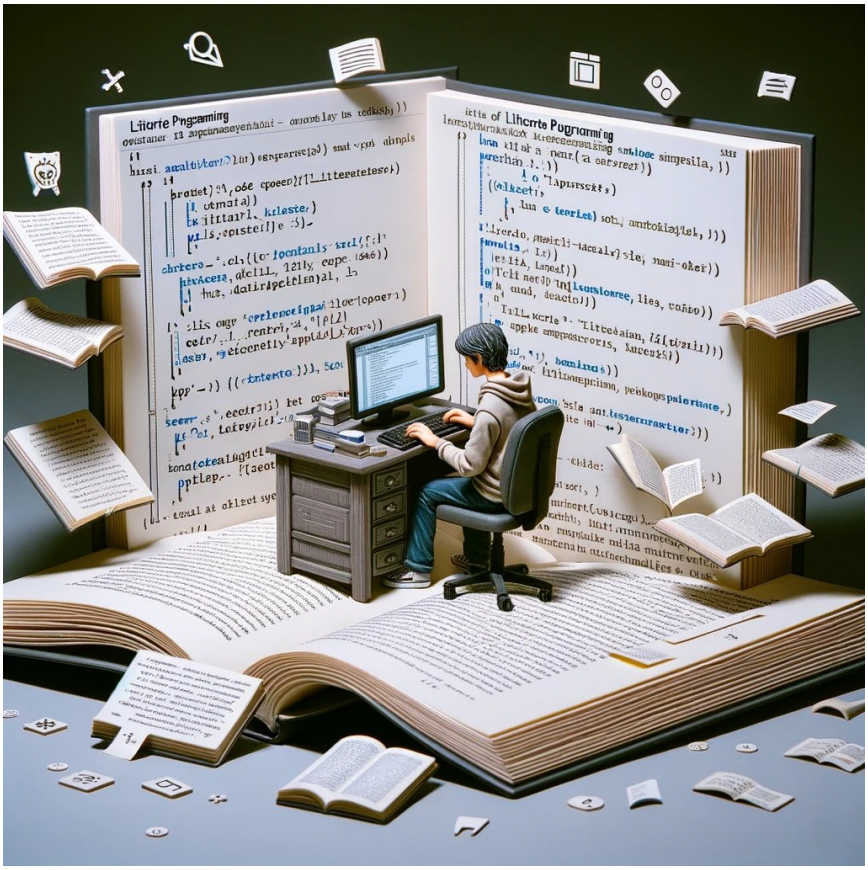
☐ Last lesson of the "Understanding data science" course (due Fri)

1.1 Why (and how) should you attend the Lyon Career Expo?

- Explore opportunities for future career paths.
- Build connections with local employers.
- Show initiative. Dress the part. Bring your resume.
- See our new career center director, Cassidy Mitchell

Careers often unfold in unpredictable ways. Some things are in your control: Being diligent, curious, responsible, act honorably, listen more than you speak. Those qualities will carry you far, no matter your career path.

1.2 Projects - update and next steps



- These were very good beginner proposals! Best ever in my view!
- I took away a point if you violated the constraints.
- Completeness and compliance is more important than creativity.
- Getting something (anything!) done is more important than ambition.
- Most of you are biting off more than you can chew.
- Make the problem as small as you can, get results, then expand.
- Remember to stick to what you already know and expand from there.
- Don't do what I do but do what I tell you and don't do myself.

- Next: Move on from "What" to "How". Deliverable: **Literature Review.**
- Did I use AI to prepare my feedback? For 3/6 feedback reports.

  1. I read your proposal twice, taking notes along the way.
  2. I fed your proposal and my assignment text into ChatGPT.
  3. I asked for a section-by-section critique draft.
  4. I copied the draft and edited it for content.
  5. I think the feedback is now more structured than before.

  I'm going to demonstrate to you how much more time you will have to spend when you opt to enroll stupid generative AI for a creative task: Now I need to re-validate all proposals to be sure of you.
- Did you use generative AI tools to prepare the proposal? How?

1.3 Review questions on your projects

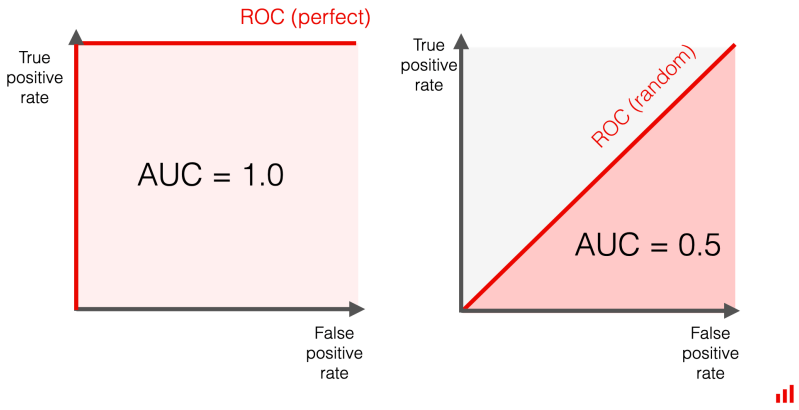
I promised that you would get an opportunity to check your understanding of your own work especially if you had AI assistance. But even if you did not, you need to be able to answer with confidence and skill to your own creative work.

1. **Fixture congestion** (Diego, Matheus, Frederico)

   What is AUC/ROC?

AUC/ROC is an evaluation method for binary classification model:

- ROC (Receiver Operating Characteristic curve) plots true positive rate against false positive rate. Shows how well the model separates two classes (e.g. injury vs. no injury)
- AUC (Area Under the Curve) is a single number that summarizes the ROC curve, ranging from 0.5 (no better than random guessing) to 1.0 (perfect classification).



2. **Titanic survivors** (Olivia, Ava)

   You mention that you "want to analyze demographics from the surviving Titanic passengers to see if factors like race, age, gender, and economic and social status influenced who survived."

   What are the data contained in the referenced Kaggle dataset? Where do the data originate from?

   The dataset contains 12 features:

   (a) Passenger ID
   (b) Survived
   (c) Passenger class
   (d) Name
   (e) Sex

(f) Age
(g) Number of siblings aboard
(h) Number of parents/children
(i) Ticket number
(j) Passenger fare
(k) Cabin number
(l) Port of embarkation

Most notably, **race** is not captured, and cannot be inferred either. Economic and social status could be inferred but "status" is a vague category (hard to measure - means different things to different people in different places).

The original source seems to be the Titanic passenger manifest but I did not find this information on Kaggle who altered the dataset. This is not the raw historical dataset.

3. **Morse code** (Matthew)

   This is clearly not AI generated since the proposal misses idiosyncratically out on several assignment categories. An imperfect but intelligent proposal is almost always a clear sign of human creativity.

   Still, for good measure: You write that using Scrum (not an acronym) is used so that you "are encouraged as aspiring Data Scientists to continually evaluate and learn through our experiences with attention to our personal skill stack and problem-solving abilities as they relate to the task at hand."

   A mouthful! Very typical for AI by the way. Question: What is really the core of Scrum for agile project management?

   The core of the Scrum agile management method is to manage complex work through short, iterative cycles and continuous feedback. To do this Scrum employs a set of roles (like "product owner"), artifacts (like "sprint backlog", items selected for current sprint), and events (like "sprint review").

4. **Retail data** (Prabhat, Riya, Avash)

Some formulations are typically AI-vague, and the unstructured references could give AI away - but the proposal reads personal.

Question: In your comments, you write that "the chest dataset" (which you're not using) "would solve a meaningful problem in data science."

Which problem are you referring to here?

5. **Time management** (Surendra, Saksham, Jenish)

   What are the main findings and recommendations of the two papers that you cited? What data were used by the authors?

   Answer:

   (a) Wilson et al. (2021) found correlations between students' time management skills and academic success, based on 140 students at the Austraila Defence Force Academy. Recommends to implement time management training and further research.

   (b) Terzi et al. (2024) found that students who manage their time experienced an enhanced quality of life (because they now had more time for fun aka leisure activities). Based on a sample of 213 Turkish students, 18-35 years. Recommends training.

6. **Fentanyl in mice** (Levi)

   Lots of additional documentation here - curious to hear if you or any of your collaborators benefitted from AI assistance.

   You could show your skill level by explaining the difference between a "Type 1 or Type 2 error" (mentioned in "Constraints").

   - **Type I error** (false positive): You conclude that there is an effect when there really isn't one because you found something (positive) that isn't there (false). Example 1: You conclude that DOI (some amphetamine?) does mitigate withdrawal symptoms in mice when it does actually not do that. Example 2: When predicting for spam, false positives are messages marked as spam (what you are looking for) that are not spam.

- **Type II error** (false negative): You fail to detect an effect that actually exists because you didn't find something (negative) that is there (false). Example 1: You fail to detect a real effect of DOI on withdrawal symptoms. Example 2: You miss spam messages.