Introduction to Data Science (DSC 105) Syllabus

DSC105 Introduction to Data Science - Syllabus - Fall 2024

Marcus Birkenkrahe

July 16, 2024

Contents

1 General Course Information

- Course title: Introduction to data science
- Course number and section: DSC 105.01
- Meeting Times: Mon-Wed-Fri from 11:00-11:50 am
- Meeting place: Derby Science Building computer lab room 209
- Professor: Marcus Birkenkrahe
- Professor's Office: Derby Science Building 210
- Phone: (870) 307-7254 (office) / (501) 422-4725 (private)
- Office hours: by appointment MWF 4pm, Tue 3pm, Thu 11 am & 3 pm

- Textbook (optional): Dive into Data Science by Bradford Tuckfield, NoStarch 2023. Recommended: Data Science from Scratch by Joel Grus (2e), O'Reilly 2019. Python Data Science Handbook by Jake Vanderplas, O'Reilly 2019.

2 Objectives

Data science is concerned with getting data to work for us, to give us its (presumed) hidden treasures. Data science has been called "the sexiest job of the 21st century". Even if you don't want to become a professonial data scientist, it's helpful to master the basic concepts if you want to succeed in today's data-driven business.

3 Target audience

The course is for anyone who is interested in becoming more data literate in their own field of interest - be it languages, theatre, biology, psychology or exercise science - and growing their personal skill stack. Visualization of data-driven insights and improved productivity when working with data and media is a concern for any professional.

4 Student Learning Outcomes

Students who complete "Introduction to data science" (DSC 105) will be able to:

- Learn computer and data science principles by playing with data
- Acquire basic programming skills using Python, R, SQL, and bash
- Apply literate programming principles by working with Org-mode
- Use infrastructure including command line, Jupyter notebooks, GitHub
- Understand the relationship of humans, machines, and data
- Develop their critical thinking skills
- Learn how to effectively use AI coding assistants like Copilot, Code Interpreter (ChatGPT)
- Understand the impact of, and can code in R, Python, SQL, bash

- Know how to effectively present assignment and project results

Students, who complete DSC 105 will have fulfilled the prerequisites for DSC 205, Introduction to advanced data science, which focuses on advanced practices in data exploration, visualization, and integration with databases and other languages through R and Python packages and APIs.

5 Course requirements

Formal prerequisites: Introduction to Programming (either CSC100 or CSC115 or CSC109, and MTH101 (College Algebra).

Otherwise no prior knowledge required. Some knowledge of, and experience with computers and/or programming is useful but not critical. Curiosity is essential. You will gain data literacy skills by taking this course. The course will prepare you for further studies in computer and data science, or in other disciplines that use modern computing, i.e. every discipline, from accounting to zoology).

6 Grading

WHEN	DESCRIPTION	IMPACT
Weekly	DataCamp assignments	25%
Monthly	Project sprint reviews	25%
Weekly	Tests	25%
TBD	Final exam (optional)	25%

- DataCamp is an online learning platform for data science
- Sprint reviews are monthly group project progress reports
- Tests are open-book multiple choice exams to be completed at home
- The final exam is optional if you want to improve your grade

7 Rubric

Component	Weight	Excellent	Good	Satisfactory	Needs Improvement	Unsatisfactory
Participation and Attendance	0%	Consistently attends and actively participates in all classes.	Attends most classes and participates in discussions.	Attends classes but participation is minimal.	Frequently absent and rarely participates.	Rarely attends classes and does not participate.
DataCamp Assignments	25%	Completes all assignments on time with high accuracy (90-100%).	Completes most assignments on time with good accuracy (80-89%).	Completes assignments but with some inaccuracies or delays (70-79%).	Frequently late or incomplete assignments with several inaccuracies (60-69%).	Rarely completes assignments and shows minimal understanding (0-59%).
Project Sprint Reviews	25%	Consistently demonstrates significant progress, excellent teamwork, and high-quality work (90-100%).	Shows good progress, effective teamwork, and good-quality work (80-89%).	Adequate progress, teamwork, and satisfactory work quality (70-79%).	Minimal progress, poor teamwork, and below-average work quality (60-69%).	Little to no progress, ineffective teamwork, and poor-quality work (0-59%).
Tests	25%	Demonstrates thorough understanding and application of concepts (90-100%).	Shows good understanding with minor errors (80-89%).	Displays basic understanding with some errors (70-79%).	Limited understanding with several errors (60-69%).	Minimal understanding and many errors (0-59%).
Final Exam (Optional)	25%	Demonstrates comprehensive understanding and application of course concepts (90-100%).	Shows strong understanding with minor errors (80-89%).	Displays adequate understanding with some errors (70-79%).	Limited understanding with several errors (60-69%).	Minimal understanding and many errors (0-59%).

8 Learning management system

- We use Lyon's Canvas installation for this course.
- The home page contains: assignments, grades, pages, people, syllabus, quizzes, Google Drive, Course evaluation and Zoom.
- The Zoom page includes cloud recordings of all past sessions.
- Recorded sessions will be deleted after the last class.

9 DataCamp

- The course includes a free subscription to the DataCamp classroom at datacamp.com for further study, and the opportunity to earn certificates. DataCamp is a popular data science online learning platform.
- We will use the DataLab workspace that comes with the DataCamp subscription will be our usual stomping to experiment with either Python or R.

10 GitHub

All course materials are available in a public GitHub repository (github.com/birkenkrahe/ds1). Registration for students includes a free subscription to GitHub Codespaces with the AI coding assistant Copilot (you need to provide proof of student status yourself). GitHub is the worldwide largest online platform for software development.

11 Lyon College Standard Policies

Online: `https://tinyurl.com/LyonPolicyOnline`, see also Class Attendance

12 Dates and class schedule

See also: 2024-25 Academic Calendar

Assignment and project schedule:

- Summer prep program: Understanding Data Science (2 hours) [Aug 19]
- Bonus: Introduction to Data Science in Python (4 hours) [Dec 6]
- We will cover 13 DataCamp courses, and 4 project sprint reviews.
- Each course carries a certificate that you can add to your resume.
- We emphasize the data science workflow and Python as a toolbox.

Week	Datacamp assignments	Project
1	Introduction to Python: Basics	
2	Introduction to Python: Lists	
3	Introduction to Python: Functions & Packages	
4		1st sprint review
5	Introduction to Python: NumPy	
6	Intermediate Python: Matplotlib	
7	Intermediate Python: Dictionaries & Pandas	
8	Intermediate Python: Logic, Control Flow & Filtering	2nd sprint review
9	Intermediate Python: Loops	
10	Intermediate Python: Case Study: Hacker Statistics	
11	Data Manipulation w/pandas: Transforming DataFrames	
12		3rd sprint review
13	Data Manipulation w/pandas: Aggregating DataFrames	
14	Data Manipulation w/pandas: Slicing/Indexing DataFrames	
15	Data Manipulation w/pandas: Creating/Visualizing DataFrames	
16		4th sprint review

Textbook example and topic schedule

- We will cover up to 4 chapters of this introductory text.
- We emphasize descriptive and prescriptive data analytics.
- We also cover aspects of data engineering and modeling.

Ch	Topic	Textbook "Dive into Data Science"	Page	Week
1	Introduction	Introduction	12-14	1
	Setting up	Setting Up the Environment	15-20	
	Exploring data	Your First Day as CEO	20-24	2
	Data tables	Displaying Data with Python	25-27	
	Summarization	Calculating Summary Statistics	28-30	3
	Subsetting	Analyzing Subsets of Data	31-34	
	Visualization	Visualizing Data with Matplotlib	35-42	4
	Correlations	Exploring Correlations	43-49	
	Visualize correlations	Creating Heat Maps	50-52	5
2	Forecasting	Predicting Customer Demand	55-56	6
	Data cleaning	Cleaning Erroneous Data	56-58	
	Plotting trends	Plotting Data to Find Trends	59-60	7
	Linear regression	Performing Linear Regression	60-69	
	Forecasting	Using Regression to Forecast Future Trends	70-72	8
		Trying More Regression Models	72-85	
3	Hypothesis testing	Reading Population Data	88-97	9
		Performing Hypothesis Testing	98-104	
		Comparing Groups in a Practical Context	105-109	10
4	A/B testing	The Need for Experimentation	111-112	11
		Running Experiments to Test New Hypotheses	113-121	
		Optimizing Frameworks	122-125	12
		Understanding Effect Sizes	126-128	
		Calculating the Significance of Data	129-131	13

- The next 7 chapters will be covered in the next course (DSC 205).
- The course is titled "Introduction to Advanced Data Science".
- This course emphasises predictive analytics, un/supervised learning.
- It also includes: Web scraping, natural language processing, SQL, R.

13 A note on using AI to write code for you or debug your code

Short summary: For students, using AI is a waste of time at best, and a crime against your ability to learn at worst. Learning never comes without pain and (temporary) desperation. AI is like a pill but one that only works

some of the time, and you'll never know when. Instead: join Lyon's Programming Student Club and experience the pain of not knowing first hand every week!

Will you be punished for using AI in my class? Not directly because nobody can tell if you used AI or not but indirectly by turning in suboptimal results, by learning less, and by having less time for other, more productive activities.

Are there any data on this? Not much on coding as such but a recent (15 July), substantive, long (59 p) paper titled "Generative AI Can Harm Learning"), based on a very carefully conducted field experiment with a large (1000) sample of high school students concluded: "Our results suggest that students attempt to use [AI] as a "crutch" during practice problem sessions, and when successful, perform worse on their own. Thus, to maintain long-term productivity, we must be cautious when deploying generative AI to ensure humans continue to learn critical skills." (Bastani et al, 2024).

References

Bastani, Hamsa and Bastani, Osbert and Sungu, Alp and Ge, Haosen and Kabakcı, Özge and Mariman, Rei, Generative AI Can Harm Learning (July 15, 2024). Available at ssrn.com.