Data visualization in R (and Python)

DSC302 - Data Visualization - Syllabus Fall 2024

Marcus Birkenkrahe

July 16, 2024

Contents

1 General Course Information

- Course title: Data visualization
- Course number and section: DSC 302.01
- Meeting Times: Mon-Wed-Fri from 15:00-15:50 am
- Meeting place: Derby Science Center Computer Lab room 209
- Professor: Marcus Birkenkrahe
- Professor's Office: Derby Science Center 210
- Phone: (870) 307-7254 (office) / (501) 422-4725 (private)
- Office hours: by appointment MWF 4pm, Tue 3pm, Thu 11 am & 3 pm

- Textbooks (not mandatory):
    1. For R, Exploratory Data Analysis Using R by Ronald K. Pearson, chapters 1-4 (pp. 1-180) (CRC Press, 2018);
    2. for Python, Interactive Data Visualization with Python 2nd ed by Belorkar et al. (Packt, 2020); The Data Visualization Workshop by Döbler/Großmann (Packt, 2020).
    3. language-agnostic: Fundamentals of Data Visualization by Claus O. Wilke (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.

This course focuses on visualizing data. There are many tools and platforms out there to visualize patterns found in data and tell engaging, convincing stories. As a data scientist, you have to know what's out there, understand the principles of good visual presentation, and be able to customize plots quickly.

In this course, we look at popular platforms like Tableau and R Shiny, and at different packages within R.

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.

In this course, we use R (and some Python), which is easier to learn for students outside of computer science.

4 Student Learning Outcomes

Students who complete "Data visualization" (DSC 302) will be able to:

- Learn about, and use popular data science visualization platforms
- Understand how to visualiye exploratory data analysis results
- Apply literate programming principles to their work with Org-mode
- Use infrastructure including command line, Emacs, and GitHub
- Develop their critical thinking skills
- Know how to effectively present assignment and project results

This introduction to Exploratory Data Analysis prepares course participants for taking DSC 305, "Machine learning".

5 Course requirements

Introduction to programming (CSC 100 or CSC 115). Some knowledge of, and experience with programming and using the R language 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 system

WHEN	DESCRIPTION	IMPACT
Weekly	Programming assignments	25%
Monthly	Sprint reviews	25%
Weekly	Tests	25%
TBD	Final exam (optional)	25%

- Sprint reviews are monthly project progress reports
- Tests are open-book multiple choice exams for 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 GitHub

All course materials are available in a public GitHub repository (github.com/birkenkrahe/dviz24). Registration for students includes a free subscription to GitHub codespaces with the AI coding assistant Copilot. GitHub is the worldwide largest online platform for software development.

10 Lyon College Standard Policies

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

11 Dates and class schedule

- Summer study/preparation: Understanding data visualization (2 hours)
- Bonus: Visualizing Geospatial data (4 hours)

Week	DATACAMP ASSIGNMENT	PROJECT
1	Introduction to data science with Python	
2	Loading data in pandas	
3	Plotting data with matplotlib	
4	Different types of plots	
5		1st review
6	Introduction to Matplotlib	
7	Plotting time-series	
8	Quantitative comparisons and statistical visualizations	
9	Sharing visualizations with others	
10		2nd review
11	Introduction to Seaborn	
12	Visualizing Two Quantitative Variables	
13	Visualizing a Categorical & a Quantitative Variable	
14	Customizing Seaborn Plots	
15		3rd review
16	Final presentations	

Textbook example and topic schedule

- We will cover up to 5 chapters of this advanced introductory text.
- We emphasize general plots, Matplotlib and Seaborn (see DataCamp).
- We will try to cover more applications like geodata & animation.

Ch	Topic	Textbook "The Data Visualization Workshop"	Page	Week
1	Introduction	Introduction and setup	1-22	1-4
	Setup	Data wrangling, tools and libraries	23-27	
	Statistics	Measures of centrality and dispersion	28-34	
	NumPy	Python library for numerical computing	25-66	
	Pandas	Python library for data analysis	67-86	
		Advanced pandas operations	87-100	
2	Plots	Comparison plots: Line, bar, radar	102-115	5-8
		Relation plots: Scatter, bubble, heatmap	116-125	
		Composition plots: Pie, stacked, Venn	126-137	
		Distribution plots: Histogram, density, box	138-145	
		Geoplots: Dot, chloropleth, connection map	146-150	
3	Matplotlib	Pyplot basics	164-174	9-10
		Basic customization: text and legends	175-179	
		Basic plots	180-202	
		Layouts, images, mathematical expressions	203-225	
3	Seaborn	Simplifying visualizations using Seaborn	226-250	11-12
		Advanced plots	251-277	
4	Geospatial	Plotting geospatial data	278-327	13-14
5	Bokeh	Making things interactive with Bokeh	328-389	15-16

Page numbers follow the textbook "The Data Visualization Workshop" by Döbler and Großmann (Packt, 2020).

12 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.