BIOF 440: Data Visualization using Python

Number of credits : 1

Spring 2021 Term B

Syllabus

Instructor

Abhijit Dasgupta, PhD

Contact information:

- E-mail: dasgupta.faes@gmail.com
- Preferred method of communcation: via Slack.

Course information

Prerequisites, if any: None, though some knowledge and practice of Python may be useful

Course description

This course will demonstrate and practice the use of Python in creating and presenting data visualizations. After a short introduction to Python tools and packages for data science, especially the PyData stack, we will look at good principles for data visualization, examples of good and bad visualizations, and the use of matplotlib/seaborn/pandas to create static publication-quality graphs. We will also explore modern web-based interactive graphics using plotly/bokeh/Altair. We will explore ways in which bioinformatics data can be presented using static and dynamic visualizations. Finally, we will use Jupyter Notebooks to develop web pages for presenting data visualizations as self-explanatory storyboards.

Course materials

All course materials (lectures, videos, homework, discussions) will be available on the class Canvas site.

Learning Materials

Required and Recommended Texts: There are no required texts for this class. However, the following texts, freely available online, will be used for reference:

1. *Python Data Science Handbook* [PDSH] by Jake VanderPlas (available online)
2. *Principles of Data Visualization* [PDV] by Claus O. Wilke (available online)

Required Journal Articles: There are no required journal articles for this class

Course Goals

When you complete the course successfully, you will be able to:

- Understand principles of good data visualization; avoid poor or inappropriate data visualization
- Use Python to create data visualizations, and manipulate data to enable good visualizations
- Understand and appropriately use color, symbols and small multiples
- Create static and dynamic data visualizations
- Use the web as a presentation medium

Structure of the course

This course will run for 7 weeks. Of these, there will be instructional material, including videos, lectures, slides, discussion, tutorials and homework, for 6 of the weeks. The seventh week will be dedicated to a culminating project that will be submitted by the end of the seventh week. Your grade will be determined by class participation, i.e., discussions & Slack participation (30%), homework assignments (50%) and the final project (20%).

Detailed course outline

Week 1

- Introduction to the class
- Principles of good data visualization
- An understanding of bad data visualizations
- Introduction to Python, Spyder and JupyterLab/Notebook

Week 2

Theme: Descriptive plots

- Introduction to matplotlib and seaborn
- Layering components
- Basic graphs
- Grouped graphs
- Facets (small multiples)
- Thematic personalization
- Easy animation

Week 3

Theme: Analytic plots

- Using Python for analysis

- o Data manipulation
- o Statistics
- o Hypothesis tests
- Producing graphs from analysis
  - o Effect sizes
  - o Group differences
  - o Annotations
  - o Networks
  - o Maps

Week 4

Theme: Python for Bioinformatics

- Visualizations for Bioinformatics
  - o Manhattan plots
  - o Heatmaps
  - o Dendrograms
  - o Phylogenetic trees
  - o Linkage maps
- Visualizations for RNA-seq and NGS analyses

Week 5

Theme: Dynamic visualization

- Introduction to packages for dynamic visualizations
  - o plotly, bokeh, Altair
- Interactive heatmaps
- 3-D visualization

Week 6

Theme: Presenting your graphs

- Jupyter notebooks
- Leveraging R Markdown as needed
- Dash/Streamlit for dashboards (?)

Week 7

Class presentations and discussion

The Learning Process

I believe in teaching practical methods for using Python as a tool in achieving informative data-driven visualizations. As such, this course is opinionated, in that I make certain choices of what parts of Python to teach to make things most accessible and

useful. The course will be a mixture of didactic lessons, interactive tutorials and exercises, culminating in a final project that brings different aspects of the course together into a single dashboard.

Python is a tool to be used, not studied, and so I promote active learning by doing in order to become familiar with Python, its advantages and disadvantages, and using Python regularly through the course to learn its capabilities to visualize data. Students will be expected to create visualizations and dashboards to show their data story from the first day, thus learning how to apply their learning to their own workflows and work environments.

Methods for students to achieve success

1. Practice programming and coding with Python
2. See high quality online examples provided by members of the Python community and learn
3. Participate in class discussions on Slack
4. Determine a target visualization they would like to create for presentation to their labs and work towards creating that.

Time commitment Daily practice for even 30 minutes is good, but for particular class work I don't expect more than a couple of hours a week.

This course should take around 4-6 hours of time weekly, depending on the week.

Communication

This class will communicate primarily via Slack. You will see a channel `#spring2021-b`. Please join this channel. Please use Slack for broadcasting messages, answering questions and the like. When you ask a question, please ask it under the `#general` or `#spring2021-b` channels, so others can learn as well. I should respond within 24 hours.

There will be office hours via Zoom on Thursdays, 7:00-8:30 pm

The Canvas Discussion forum will be used for guided class discussions.

Etiquette

The most important thing is to be polite, considerate and empathetic in all communications and discussions. There are different levels of knowledge about R in this class, and so some questions may appear trivial to some but are essential for others. Be kind, and if you can help a classmate, do so with grace and civility. The class learns best if we all help and support each other.

Policies

Academic Policies

This course adheres to all FAES policies described in the academic catalog and student handbook, including the Academic Integrity policy listed on page 11 of the academic

catalog and student handbook. Be certain that you are knowledgeable about all of the policies listed in this syllabus, in the academic catalog and student handbook, and on the FAES website. As a student in this program, you are bound by those policies.

Copyright

All course materials are the property of FAES and are to be used for the student's individual academic purpose only. Any dissemination, copying, reproducing, modification, displaying, or transmitting of any course material for any other purpose is prohibited, will be considered misconduct, and may be cause for disciplinary action. In addition, encouraging academic dishonesty by distributing information about course materials or assignments which would give an unfair advantage to others may violate the FAES Academic Integrity policy. Course materials may not be exchanged or distributed for commercial purposes, for compensation, or for any purpose other than use by students enrolled in the course. Distributions of course materials may be subject to disciplinary action.

Guidelines for Disability Accommodations

FAES is committed to providing reasonable and appropriate accommodations to students with disabilities. Students with documented disabilities should contact Dr. Mindy Maris, Assistant Dean of Academic Programs.

Dropping the Course

Students are responsible for understanding FAES policies, procedures, and deadlines regarding dropping or withdrawing from the course or switching to audit status.

Harassment

FAES adheres to the NIH's harassment policies, which can be found at the following link: https://hr.nih.gov/working-nih/civil/statement-workplace-harassment Faculty and students in FAES courses are responsible for being familiar with the NIH's harassment policies and adhering to them.

Attendance

It is in your best interest to use, utilize, question and understand all the instructional material provided, and to submit questions and homework in a timely manner. Since this course is completely asynchronous, there is no attendance required at particular times.

Participation

Participation will be judged through the assigned discussions as well as through activity on Slack.

Assignment Submission

Assignment submission is through Canvas.

Due Dates

Homework is assigned at 10am each Monday and is due by 11:59pm the following Sunday.

Late Submission Policies

No late submissions of homework or discussion are allowed. However, for homework, I will only use the top 4 scores for your grade, so you will have the option of not submitting or doing poorly on 2 of them.

Step-by-Step Guidelines for Submitting Assignments:

The guidelines for submitting assignments will be posted as a screencast during the first week of class.

Expectations for instructor's feedback on assignments:

We will get your assignment grades and feedback to you within a week of submission.

Major Assignments

Grades will be based on the following requirements:

1. Homeworks for each week are due Sunday at 11:59pm (50%)
   - No late homeworks
   - We'll have 6 homeworks, I'll score the top 4 for grade
2. Final project: A Python-based webpage or dashboard showing data visualizations(30%)
3. Class participation (20%): Discussion topics in weeks 2-6