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 (mailto:dasgupta.faes@gmail.com)
- Preferred method of communcation: via Slack (https://biof440.slack.com). Sign up via this link (https://join.slack.com/t/slack-5xv1054/shared_invite/zt-o6oag3f4-B6lJl1yQDmfNWeK2uCgxHQ)

Course information

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

Course description

This course will will explore the world of data visualization and Python tools to create these visualizations. Data visualization involves making choices in how to visually encode different aspects of data (or aggregated or transformed data) into aesthetics like points, lines, colors, shapes and sizes, along with creating overlays or split graphics based on some other data attribute. We will explore some principles of creating data visualizations, exploring what makes a data visualization a good or bad one. We will learn a bit of Python, enough to be able to create visualizations from data stored in a `pandas DataFrame`. We will explore three prominent Python packages (among many) to make static and dynamic graphics, each of which expresses visual encodings in a slightly different way. This exploration will be through examples of exploratory graphics, graphical representations of modeling results, and some specialized graphics for geospatial representations and bioinformatics. Given the power of the web as a medium for information exchange, we will be using Jupyter notebooks, a web-based notebook format, to write Python code, see results, and create reports; two of the packages we will explore are meant for web browsers rather than print, though we will also cover how to create and save graphics meant for print. We will culminate with a term project that will be a Jupyter notebook visually exploring a data set of your choice, preferably one applicable to your day-to-day work.

Course materials

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

Canvas (https://faes.instructure.com) site. Some readings and supplemental materials will be freely available on the web.

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 (https://jakevdp.github.io/PythonDataScienceHandbook/))
2. *Principles of Data Visualization* [PDV] by Claus O. Wilke (available online (https://serialmentor.com/dataviz/index.html))

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

Other resources: A resources page will be available on Canvas that provides links to various useful online material that can serve as reference material as you explore data visualizations using Python.

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 ingest data and manipulate it to enable good visualizations
- Appropriately use visual encodings like color, symbols, size and small multiples to express data
- Create static (for print) and dynamic (for the web) 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 (25%), homework assignments (50%) and the final project (25%).

Detailed course outline

Pre-work

- Installing Anaconda Python Distribution and learning how to start and use a Jupyter notebook
- Installing the `plotly` and `Altair` Python packages

Week 1

- Introduction to the class

- Principles of good data visualization
- An understanding of bad data visualizations

Week 2

Theme: Introduction to the tools

- A primer on Python and `pandas`
- How computer programs can create visual encodings from data
- Introducing `matplotlib` and `seaborn` for static visualizations
- Introducing `plotly` and `Altair` for dynamic visualizations

Week 3

Theme: Descriptive plots using `matplotlib/seaborn`

- Univariate and bivariate plots
- Layering components
- Basic graphs
- Grouped graphs (overlays)
- Facets (small multiples)

Week 4

Theme: Descriptive plots using `plotly/Altair`

- Univariate and bivariate plots
- Layering components
- Basic graphs
- Grouped graphs (overlays)
- Facets (small multiples)

Week 5

Theme: Domain-specific plots

- Survival analysis
    - Kaplan-Meier plots
- Geospatial data
    - Overlaying data on maps
- Bioinformatics
    - Manhattan plots
    - Heatmaps
    - Dendrograms

    - Linkage maps
    - Clustering (PCA, t-SNE and UMAP)

Week 6

Theme: How do you display analytic results.

- Producing graphs from analysis
    - Effect sizes
    - Group differences
    - Annotations
    - Networks
- Customizing plots
    - Specifying colors, fonts and themes
    - Annotating plots

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, and which parts I'll gloss over. **This is not a course to learn Python**. 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 on web notebooks 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 as much as possible
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 ask to join this channel (as it is a private 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.

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 Python 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 (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. Each submission will consist of a Jupyter notebook file (.ipynb), except for Assignment 1.

Due Dates

Homework is assigned at 10am each Monday and is due by 11:59pm the following Sunday. . No assignments will be accepted, since solutions will be published the following Monday. There will be 6 homework, and the lowest 2 scores will be dropped for final grade calculations.

Late Submission Policies

No late submissions of homework or discussion are allowed beyond the grace period. 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. Homework 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 Jupyter notebook showing data visualizations(25%)

3. Class participation (25%): Discussion topics in weeks 2-6, and participation in Slack.