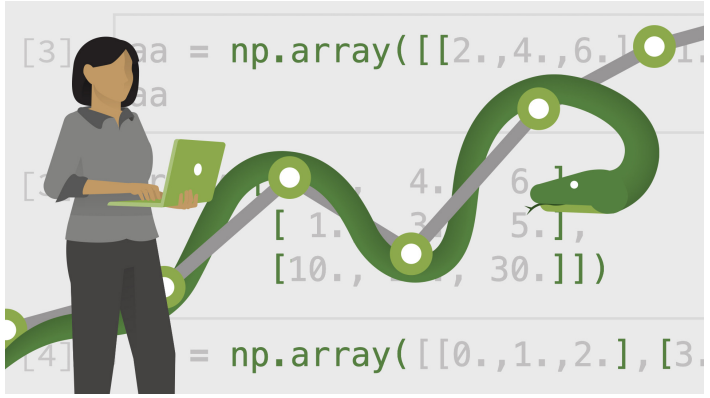
Overview of Berkeley's Data 8 class

Course description

Description:

- The UC Berkeley Foundations of Data Science course combines three perspectives: inferential thinking, computational thinking, and real-world relevance.
- Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon?
- The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks.
- It delves into social issues surrounding data analysis such as privacy and design.



Where the course is taught

1200 Berkeley students take it each semester, and many other colleges have a version of the course

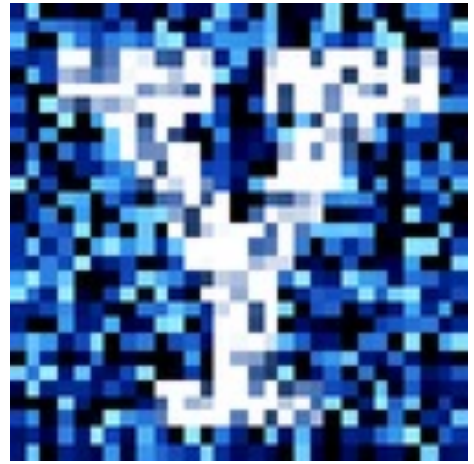
- UCSD, Cornell, Vassar, McGill, El Camino College, Anne Arundel Community College

I taught Yale's version (YData) this past spring

- Inherited the class from other professors who have been teaching it since 2019
  - Yale followed Berkeley's content very closely

I have no relationship to the creators of the original course content

- I only spoke with faculty running the course at Berkeley once



Topics covered

- Statistics
- Programming
- Data Science

- Cause and Effect
- Data Types
- Building Tables
- Census
- Charts
- Distributions
- Histograms
- Functions
- Groups
- Pivots and Joins

- Iteration
- Chance
- Sampling
- Models
- Comparing Distributions
- Decisions and Uncertainty
- A/B Testing
- Causality
- Confidence Intervals
- Center and Spread

- The Normal Distribution
- Sample means
- Design Experiments
- Correlation
- Linear Regression
- Least Squares
- Residuals
- Regression Inference
- Classification
- Classifiers

Berkeley's course calendar

Yale's course calendar

Course structure

Designed for three 50 minute meetings per week over a 13 week semester

Weekly homework assignments done in Jupyter notebooks (11 total)

Three longer "project" assignments

Weekly lab assignments

- We used these as ungraded practice exercises

We had two exams

Homework 2: Arrays and Tables

Welcome to the second homework. Please complete this notebook by filling in the cells provided.

Recommended Practice:

It is recommended (but not required) to do Practice 02 before this homework, since some functions/methods mentioned in Practice 02 might be useful for this homework.

Recommended Reading:

- Data Types
- Sequences
- Tables

1. Creating Arrays

Question 1.1. Make an array called `weird_numbers` containing the following numbers (in the given order):

1. -2
2. the sine of 1.2
3. 3
4. 5 to the power of the cosine of 1.2

Hint: `sin` and `cos` are functions in the `math` module.

```
]: # Our solution involved one extra line of code before creating
   # weird_numbers.
   ...
   weird_numbers = ...
   weird_numbers
```

Connector classes

Connector classes are courses on a particular topic that reinforce Data 8 content

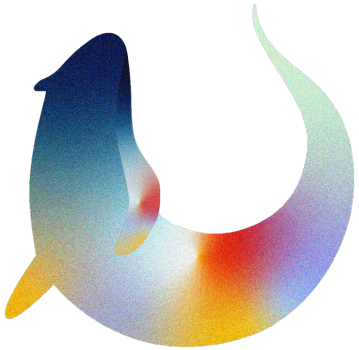
- Examples from Berkeley:
  - Data Science for Smart Cities
  - Data Science and the Mind
  - Economic Models
  - Writing Data Stories
  - etc.

I taught YData Baseball where we analyzed baseball data

Resources

There are a number of resources that are available for the class:

- Online textbook: Computational and Inferential Thinking
- Lecture material, assignments, etc., are available from Berkeley
- Auto-grading available for homework assignments (otter-grader package)
- The datascience package



The datascience package

The ***Table object*** is the main additional of the datascience package

- Easier to do data manipulation on tables compared to using pandas

Main methods of Table objects:

- Selecting columns: tb.select("column_name ")
- Filtering a subset of rows: tb.where("column_name", value)
- Aggregation: tb.group("column_name", aggregation_function)
- Visualization: tb.plot("column_x", "column_y")

Also function for visualizing data, creating maps, random sampling, etc.

Getting started with the class

Zero to Data 8 describes how to get started teaching the Data 8 class

Berkeley 2022 National Workshop on Data Science Education

- June 27-30, 2022

Try course material yourself:

- Berkeley has a GitHub repository with material they will give you access if you contact them
- YData material available on: https://ydata123.org/sp22/calendar.html

Questions?