Visualizing Data

As a general-purpose programming language, Python is incredibly useful for analyzing data and visualizing results. This activity is a first look at `matplotlib`, one of the most widely used 2D plotting libraries.

Manager:	Ethan Trokie	Recorder:	Ellyn Cho
Presenter:	Nick Gulson	Reflector:	Pranav Varanasi

Content Learning Objectives

After completing this activity, students should be able to:

- Explain the basic structure of code for plotting a mathematical function.
- Analyze visually the behavior of the Python random number generator.
- Read data from a CSV file and generate histograms of various columns.

Process Skill Goals

During the activity, students should make progress toward:

- Navigating the documentation for a third-party library. (Information Processing)

Model 1 Simple Plot

When analyzing data, it's helpful to create charts, plots, and other visualizations. Doing so allows you to see important numerical relationships. Enter the following code into a Python Editor, and run the program.

```
1  import matplotlib.pyplot as plt
2  import numpy as np
3
4  def model_one():
5      x = np.arange(0.0, 2.0, .01)
6      y = np.sin(2 * np.pi * x)
7      plt.plot(x, y)
8      plt.xlabel('time (s)')
9      plt.ylabel('volts (mV)')
10     plt.show()
11
12 if __name__ == "__main__":
13     model_one()
```

Questions (15 min)

Start time: 4:13

1. Identify in the source code which line numbers:

   a) generated the data? 5, 6

   b) set the axes properties? 8, 9

   c) displayed the window? 10

   d) plotted the actual data? 7

2. Describe in your own words what is being plotted.

The program plots a sine wave. Since the x values range from 0 to 2 seconds, the graph shows two complete cycles of the sine function. The horizontal axis represents time(s), and the vertical axis represents voltage(mV).

3. Modify the code to plot only one cycle of the sine wave (instead of two). Write the edited line of code below.

x = np.arange(0.0, 1.0, .01)

4. Change the third argument of `np.arange` from 0.01 to 0.15. What is the result?

The spacing between x values increases (fewer points). The plot will still show two cycles, but the sine wave will look jagged instead of smooth because there are fewer data points to draw the curve.

5. Add `"o"` as a third argument to the `plot` function. What is the result?

Data points are plotted as dots

6. How does the third parameter of `np.arange` affect how the plot looks?

It affects the step size. A smaller step means values are closer together, so the plot looks smooth. A larger step means values are farther apart, so the plot looks rougher and jagged, with fewer points.

7. How would you modify the code to plot the function $y = x^2 - 1$ instead? Show the results from -2 to +2.

```
x = np.arange(-2.0, 2.0, 0.01)
y = x**2 - 1
```

8. Which two Python libraries are used in Model 1? Quickly search the Internet and find their websites. Write a one-sentence description about each library. (Can you identify the third library?)

Numpy
Matplotlib - pyplot

Model 2 Histograms

You can generate a sequence of numbers using the `random` module. Merge the code below into your program from Model 1. Run the program, and view the output.

```
1  import matplotlib.pyplot as plt
2  import random
3
4  def model_two(npts):
5      numbers = []
6      for _ in range(npts):
7          numbers.append(random.random())
8      plt.hist(numbers)
9      plt.show()
10
11 if __name__ == "__main__":
12     model_two(100)
```

Questions (10 min)

Start time: 4:20

9. Based on the Python code:

a) What is the range of values generated by the `random` function? [0.0, 1.0)

b) How many random values are generated? 100

10. Based on the figure plotted:

a) How many bars are displayed? 10

b) What is the width of each bar? 0.1

c) What is the sum of the heights of the bars? 100

11. Based on your answers above, what are appropriate labels for the x and y axes?

x-axis: value
y-axis: frequency

12. Increase the argument of `model_two` to 1000, 10000, and 100000. Describe how the output plot changes when you run the program.

When the number of generated points increase, the bars become more balanced in height.

13. Add the number 50 as second argument to the `hist` function. What is the meaning of the result?

This changes the number of bins (bars) to 50. The histogram becomes more detailed, showing finer granularity of the distribution.

14. In general, describe what the `hist` function does with the list of random numbers to create this type of plot.

The hist function converts a list of numbers into a frequency distribution plotted as bars.

Model 3 CSV Data

Recall that "Comma Separated Values" is a common file format when exporting data from spreadsheets and databases. Each line of the file is a row, and each column is separated by a comma. Cells that contain commas are wrapped in quote marks.

data.csv file contents:

```
Name,Location,URL,Students
Westminster College,"Salt Lake City, UT",westminstercollege.edu,2135
Muhlenberg College,"Allentown, PA",muhlenberg.edu,2330
University of Maine,"Orono, ME",umaine.edu,8677
James Madison University,"Harrisonburg, VA",jmu.edu,19019
Michigan State University,"East Lansing, MI",msu.edu,38853
```

Python includes a csv module (https://docs.python.org/3/library/csv.html) that makes it easy to read and write CSV files.

```
import csv

infile = open("data.csv")
data = csv.reader(infile)
names = next(data)  # column names
for row in data:
    print(row[1])   # 2nd column
```

Program output:

```
Salt Lake City, UT
Allentown, PA
Orono, ME
Harrisonburg, VA
East Lansing, MI
```

Questions (20 min)

Start time: 4:28

15. In the example data.csv file above:

a) In what way is the first line different? It contains the column headers

b) How many rows of data are there? 5 How many columns? 4

16. Compare data.csv with the program output:

a) Are quote marks included in data.csv? yes In the program output? no

b) What is the purpose of the quote marks? To group values that contain commas so that the pr

17. In the Python code above:

a) Which line of code reads the first line of the file? names = next(data)

b) What type of data does the variable row contain? a list of strings

In 2013, the U.S. Department of Education released the "College Scorecard" website to help students and families compare institutions of higher education. The Scorecard data includes information like average cost of attendance, graduation and retention rates, student body demographics, etc.

18. Download the "Most Recent Institution-Level Data" from https://collegescorecard.ed.gov/data/ (listed halfway down on the right). Open the CSV file in Excel or a similar program, and skim its contents.

a) How many rows does it have? 6429

b) How many columns does it have? 3306

19. Column KE is named UGDS, which means "Enrollment of undergraduate certificate / degree-seeking students".

a) What is the range of values in this column? 0.0 to 156755.0

b) Which school has the most students enrolled? Southern New Hampshire University

c) Do all rows have an integer value for UGDS? Yes (all non-missing UGDS values are i

20. Based on the code in Model 2 and Model 3, write a program that plots a histogram of the UGDS column. Complete the following steps to consider each part of the program.

a) What two import statements will you need at the top?

```
import matplotlib.pyplot as plt
import csv
```

b) What three statements prepare the csv file for reading?

```
infile = open("Most-Recent-Cohorts-Institution_05192025.csv")
data = csv.reader(infile)
header = next(data)
```

c) What code is necessary to read the entire column into a list?

```
ugds_raw = []
for row in data:
    ugds_raw.append(row[290])
```

d) By default, data from text/csv files are read as strings. Write the code to convert the

row[290] values to integers. Be sure not include the "NULL" values in the final list.

```
ugds_list = []
for val in ugds_raw:
    try:
        ugds_list.append(int(float(val)))
    except ValueError:
        continue
```

e) Write the last two lines that plot and show the histogram.

```
plt.hist(ugds_list, bins=50)
plt.show()
```

21. Run the program, and compare your results with another team's. What does the histogram tell you about undergraduate enrollments in the United States?

Most undergraduate enrollments are small, often under 5,000 students.(the distribution of undergraduate enrollments is highly right-skewed)

22. What other questions could you ask about these data? How would you answer them using histograms, line charts, and scatter plots?

How does enrollment differ between public and private institutions?: Use two histograms to compare
How has enrollment changed over time?: Use a line chart plotting year vs. enrollment
Is there a relationship between tuition and enrollment?: Use a scatter plot with tuition on the x-axis and UGDS on the y-axis