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: Recorder:

Presenter: Reflector:

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:

1. Identify in the source code which line numbers:

a) generated the data?

b) set the axes properties?

c) displayed the window?

d) plotted the actual data?

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

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

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

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

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

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

8. Which three Python libraries are used in Model 1? Quickly search the Internet and find their websites. Write a one-sentence description about each library.

Model 2 Histograms

Recall that 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.

```
import matplotlib.pyplot as plt
import random

def model_two(npts):
    numbers = []
    for _ in range(npts):
        numbers.append(random.random())
    plt.hist(numbers)
    plt.show()

if __name__ == "__main__":
    model_two(100)
```

Questions (10 min) Start time:

9. Based on the Python code:

a) What is the range of values generated by the `random` function?

b) How many random values are generated?

10. Based on the figure plotted:

a) How many bars are displayed?

b) What is the width of each bar?

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

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

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

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

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

Model 3 CSV Data

"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:

15. In the example `data.csv` file above:

a) In what way is the first line different?

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

16. Compare `data.csv` with the program output:

a) Are quote marks included in data.csv? In the program output?

b) What is the purpose of the quote marks?

17. In the Python code above:

a) Which line of code reads the first line of the file?

b) What type of data does the variable `row` contain?

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 "Scorecard Data 7 MB CSV" from https://collegescorecard.ed.gov/data/ (listed halfway down under "Featured Downloads"). Open the CSV file in Excel or a similar program, and skim its contents.

a) How many rows does it have?

b) How many columns does it have?

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

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

b) Which school has the most students enrolled?

c) Do all rows have an integer value for UGDS?

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?

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

c) What code is necessary to read the entire column into a list? (Note: Column CH in Excel is `row[85]` in Python.)

d) By default, data from text/csv files are read as strings. Write the code to convert the `row[85]` values to integers. Be sure not include the `"NULL"` values in the final list.

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

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?

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