

CS49000-VIZ - Fall 2020

Introduction to Data Visualization

Visualization

Libraries

Lecture 2

August 27, 2020

Today

- Announcements
- Pandas for Data Handling
- Matplotlib.pyplot vs. Bokeh
- First Assignment



Data structures & Data Analysis

- works harmoniously with numpy
- imports csv files into DataFrame objects
- Slice, filter, count, select, replace, dump, group...



- Excellent tutorials and user guide
- Read the doc if unfamiliar
- Very helpful for projects that involve tabular data



Online doc

Matplotlib vs Bokeh

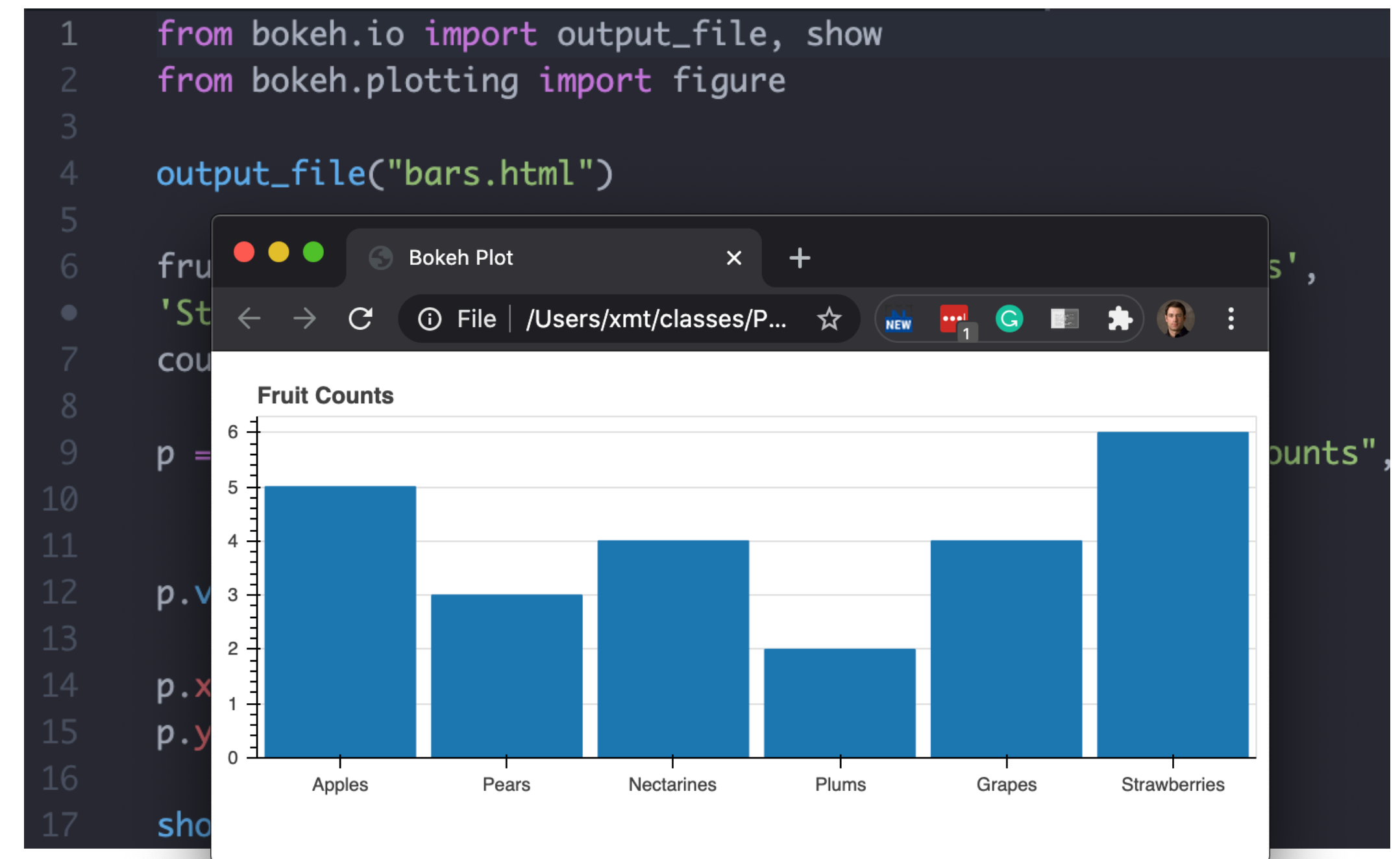
- Bar Chart

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 fruits = ['Apples', 'Pears', 'Nectarines', 'Plums', 'Grapes',
5 • 'Strawberries']
6 counts = [5, 3, 4, 2, 4, 6]
7
8 fig, axs = plt.subplots()
9 axs.bar(fruits, counts, width=0.9)
10 plt.show()
```

```
1 from bokeh.io import output_file, show
2 from bokeh.plotting import figure
3
4 output_file("bars.html")
5
6 fruits = ['Apples', 'Pears', 'Nectarines', 'Plums', 'Grapes',
7 • 'Strawberries']
8 counts = [5, 3, 4, 2, 4, 6]
9
10 p = figure(x_range=fruits, plot_height=250, title="Fruit Counts",
11           toolbar_location=None, tools="")
12
13 p.vbar(x=fruits, top=counts, width=0.9)
14
15 p.xgrid.grid_line_color = None
16 p.y_range.start = 0
17
18 show(p)
```

Matplotlib vs Bokeh

- Bar Chart



Matplotlib vs Bokeh

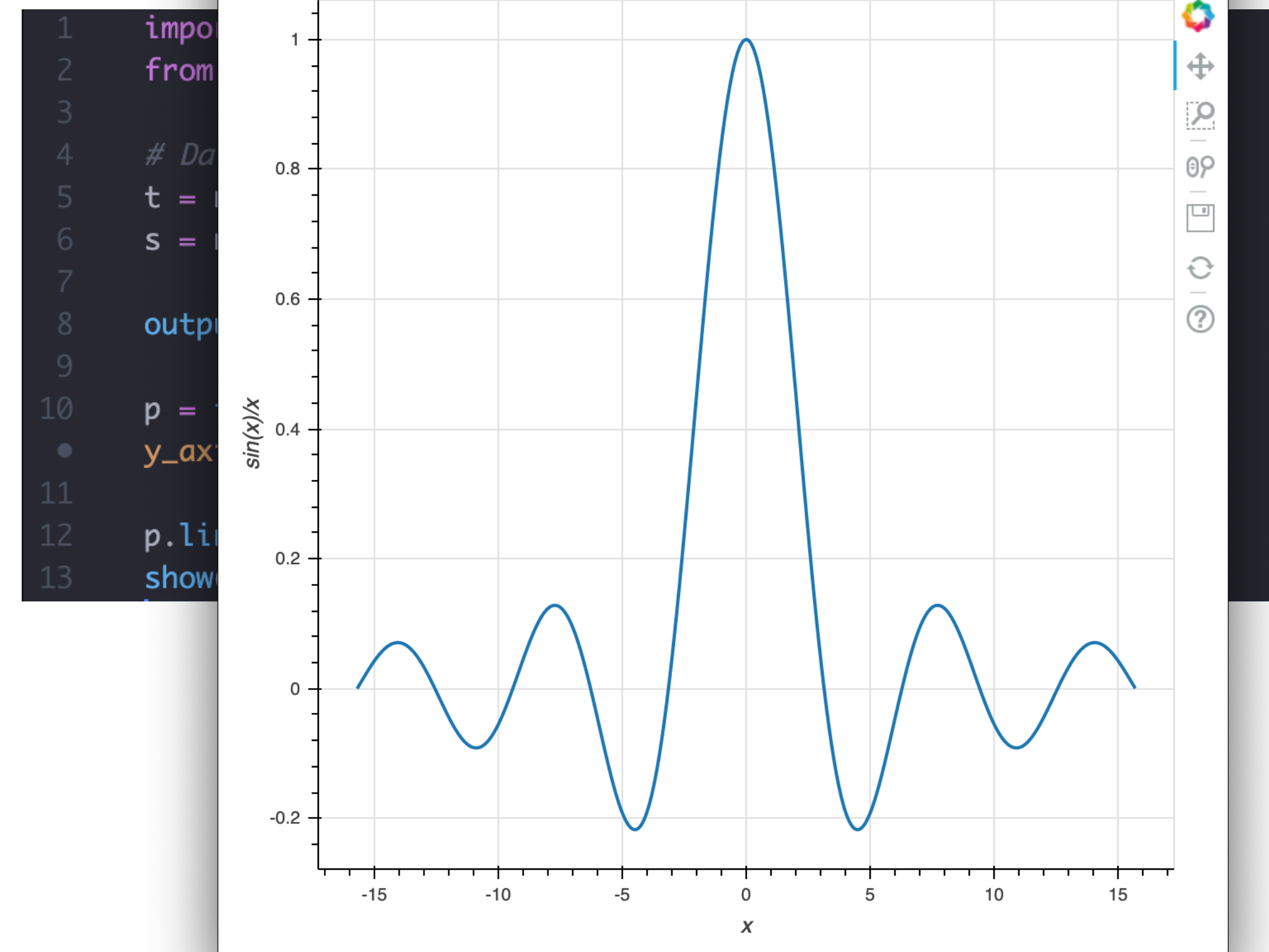
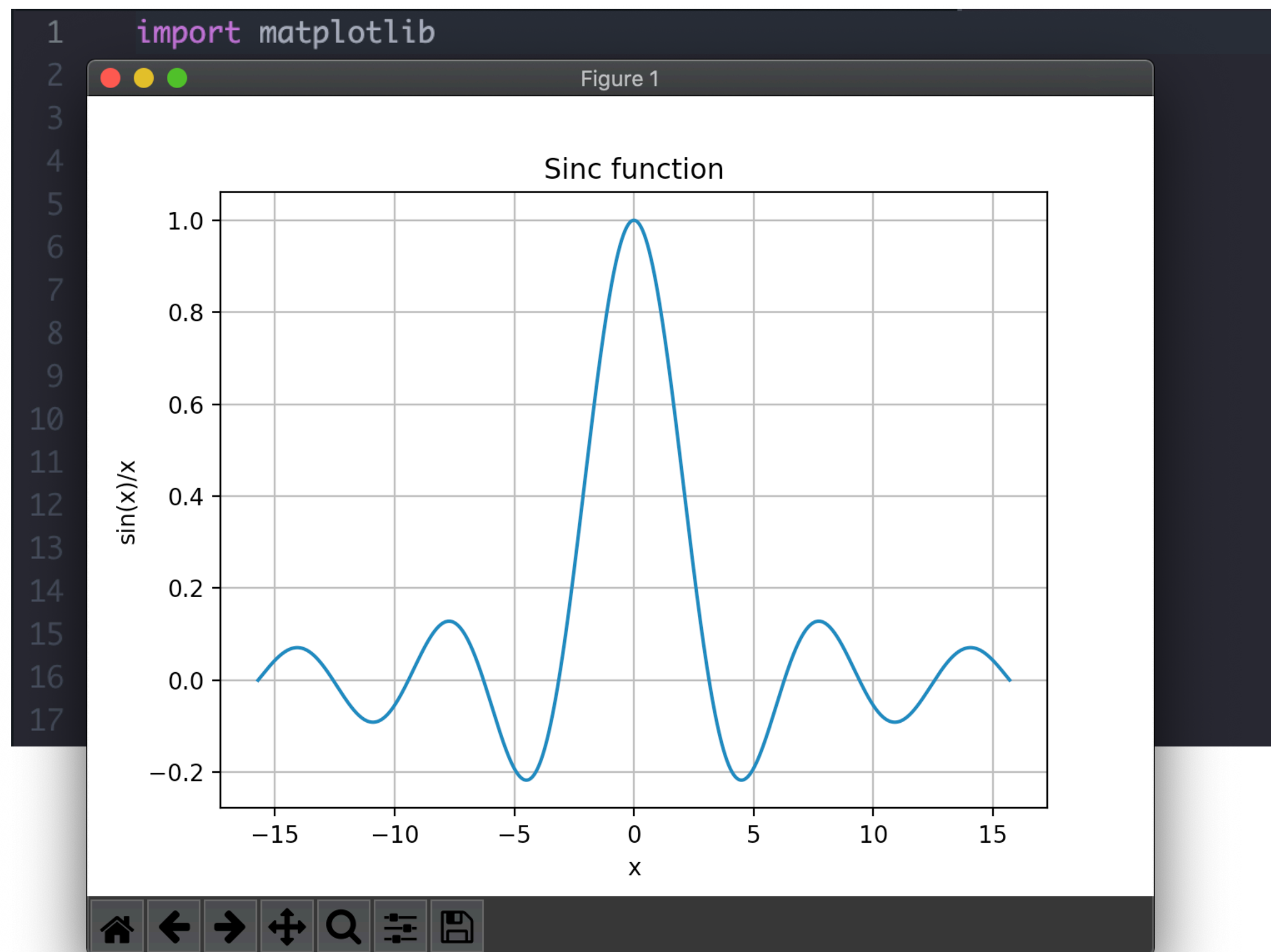
- Line Chart

```
1 import matplotlib
2 import matplotlib.pyplot as plt
3 import numpy as np
4
5 # Data for plotting
6 t = np.arange(-5.0*np.pi, 5.0*np.pi, 0.01)
7 s = np.sin(t)/t
8
9 fig, ax = plt.subplots()
10 ax.plot(t, s)
11
12 ax.set(xlabel='x', ylabel='sin(x)/x',
13       title='Sinc function')
14 ax.grid()
15
16 fig.savefig("sinc.png")
17 plt.show()
```

```
1 import numpy as np
2 from bokeh.plotting import figure, output_file, show
3
4 # Data for plotting
5 t = np.arange(-5.0*np.pi, 5.0*np.pi, 0.01)
6 s = np.sin(t)/t
7
8 output_file('sinc.html')
9
10 p = figure(title = 'Sinc function', x_axis_label = 'x',
11           • y_axis_label = 'sin(x)/x')
12 p.line(t, s, line_width=2)
13 show(p)
```


Matplotlib vs Bokeh

- Line Chart



Matplotlib vs Bokeh

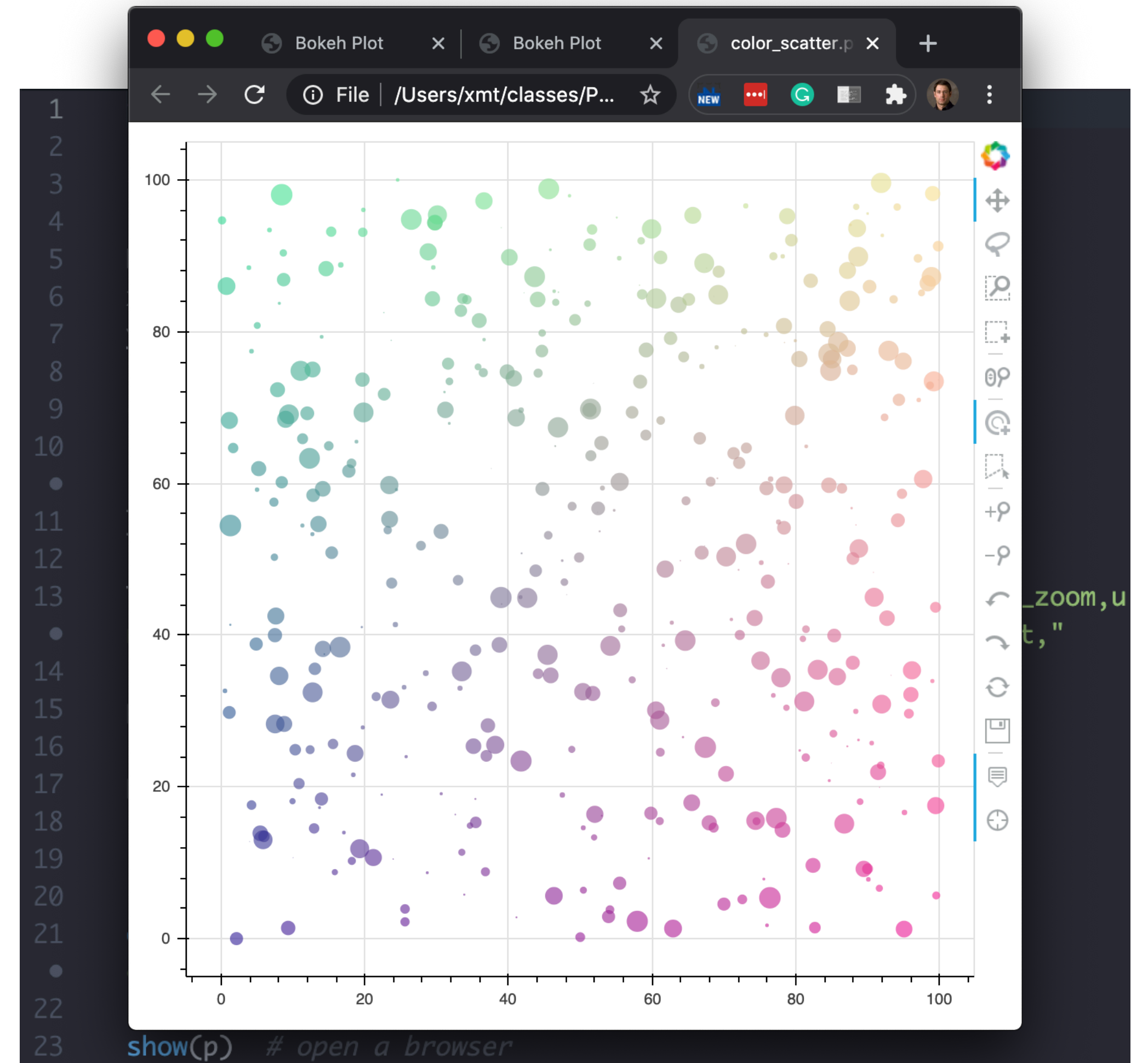
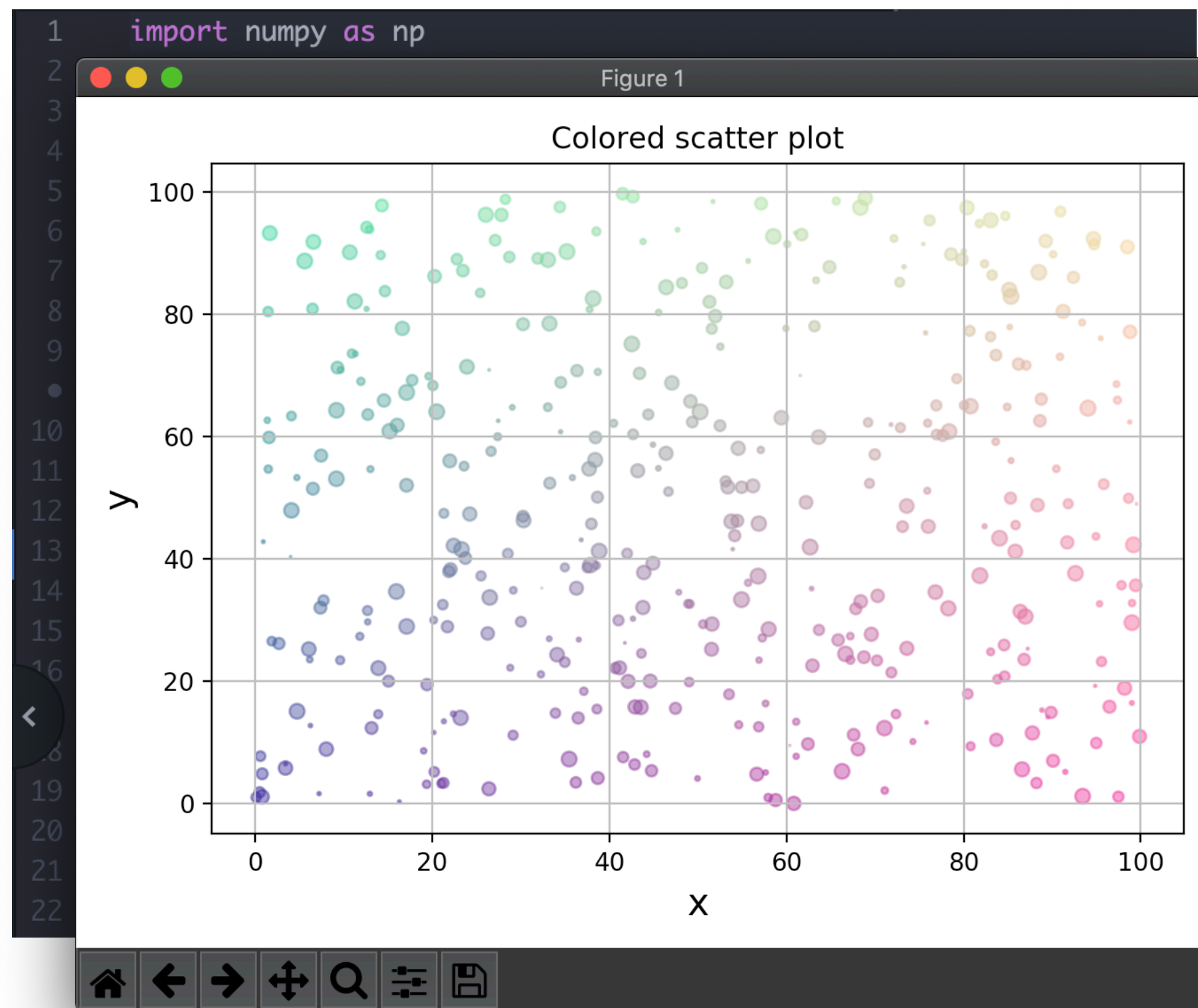
- Scatter plot

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 N = 400
5 x = np.random.random(size=N) * 100
6 y = np.random.random(size=N) * 100
7 radii = np.random.random(size=N) * 1.5
8 colors = [
9     "%02x%02x%02x" % (int(r), int(g), 150) for r, g in
10     zip(50+2*x, 30+2*y)
11 ]
12 fig, ax = plt.subplots()
13 ax.scatter(x, y, c=colors, s=25*radii, alpha=0.5)
14
15 ax.set_xlabel('x', fontsize=15)
16 ax.set_ylabel('y', fontsize=15)
17 ax.set_title('Colored scatter plot')
18
19 ax.grid(True)
20 fig.tight_layout()
21
22 plt.show()
```

```
1 import numpy as np
2
3 from bokeh.plotting import figure, output_file, show
4
5 N = 400
6 x = np.random.random(size=N) * 100
7 y = np.random.random(size=N) * 100
8 radii = np.random.random(size=N) * 1.5
9 colors = [
10     "%02x%02x%02x" % (int(r), int(g), 150) for r, g in
11     zip(50+2*x, 30+2*y)
12 ]
13 TOOLS="hover,crosshair,pan,wheel_zoom,zoom_in,zoom_out,box_zoom,undo,redo,reset,tap,save,box_select,poly_select,lasso_select,"
14
15 p = figure(tools=TOOLS)
16
17 p.scatter(x, y, radius=radii,
18           fill_color=colors, fill_alpha=0.6,
19           line_color=None)
20
21 output_file("color_scatter.html", title="color_scatter.py
22           example")
23
24 show(p) # open a browser
```


Matplotlib vs Bokeh

- Scatter plot



First Project

due 09/10

HOME SYLLABUS SCHEDULE **PROJECTS** SOFTWARE RESOURCES

CS49000-VIZ - Fall 2020

Programming Assignment 1: Basic Visualization Techniques

Key Dates

Handed out: August 27, 2020
Due date: **September 10, 2020** (before 11:59 PM)

Objectives

This first programming assignment will give you a quick overview of various basic visualization techniques while allowing you to become more familiar with a visualization library. You will work with a tabular dataset and visualize it in different ways. Since this is the first assignment, your visualizations will be rather simple and no interactivity is required.

Context

The dataset for this project is the classical *old cars* dataset, which lists several characteristics of various cars built between 1970 and 1982. This is a typical tabular dataset where the items (or keys) correspond to the individual car models and the attributes correspond to the available characteristics. This type of dataset is non-trivial to visualize because several attributes (in this case 6) need to be represented simultaneously. Here, we will mostly circumvent this challenge by considering different subsets of the data.

Practically, the dataset indicates the number of cylinders, the engine volume, the vehicle weight, the year, the geographic origin, the horsepower, and the gas mileage of 398 different models. The dataset is provided as a csv file.

Tasks

Task 1: Grouped Bar Chart