Data Display

CMPSC 105 – Data Exploration

ALLEGHENY COLLEGE

Goals

- Review Display Types
- Review Anatomy of A Graph
- Group Activity

Display Types

- data table
- pie chart
- point or dot plot

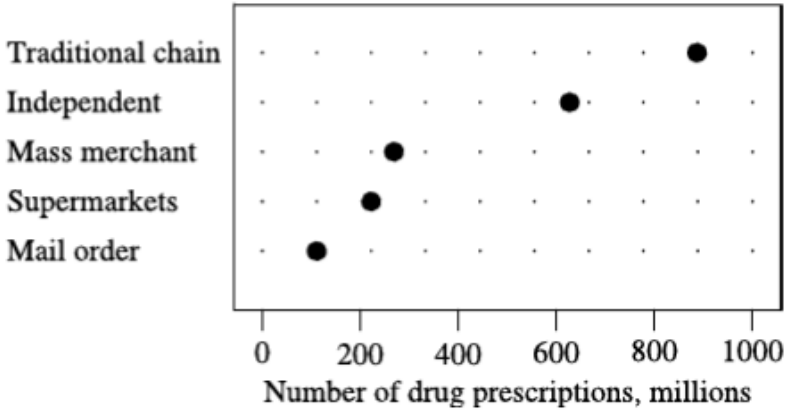
Number of drug prescriptions, millions	
Traditional chain	914
Independent	666
Mass merchant	238
Supermarkets	221
Mail order	86

Table

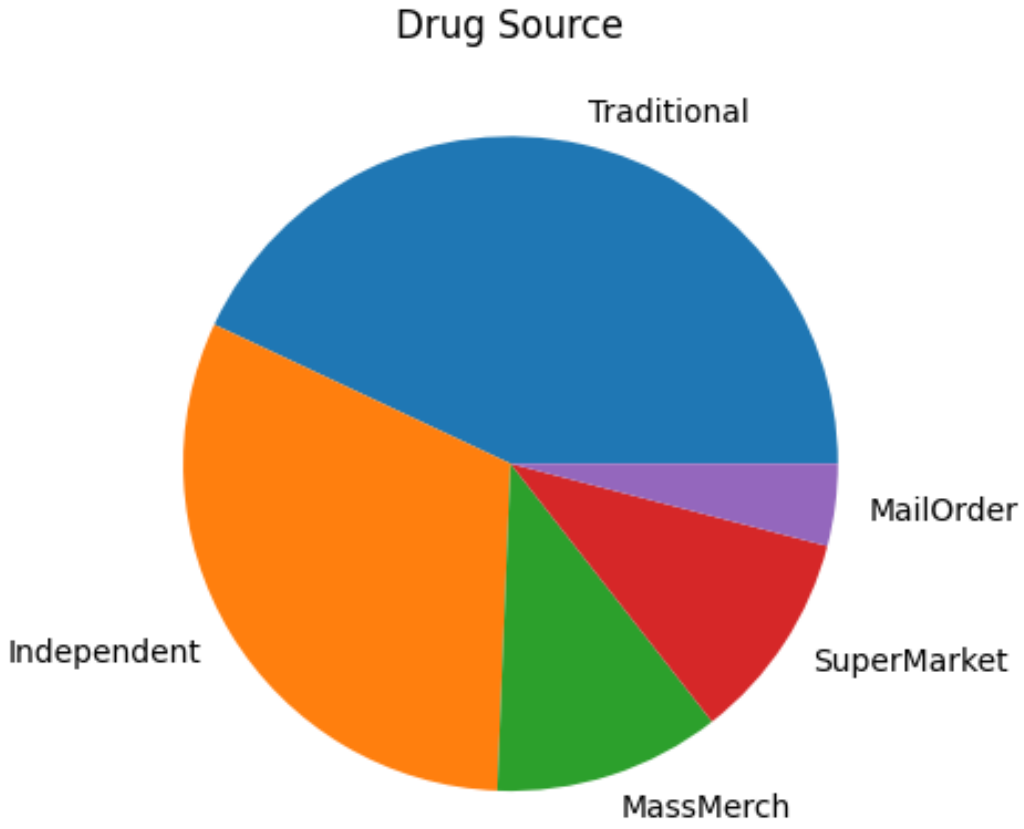
Number of drug prescriptions, millions

Pie chart

Dot plot



Pie Chart


Drug Source
Traditional
MailOrder
SuperMarket
MassMerch
Independent

Pie Chart

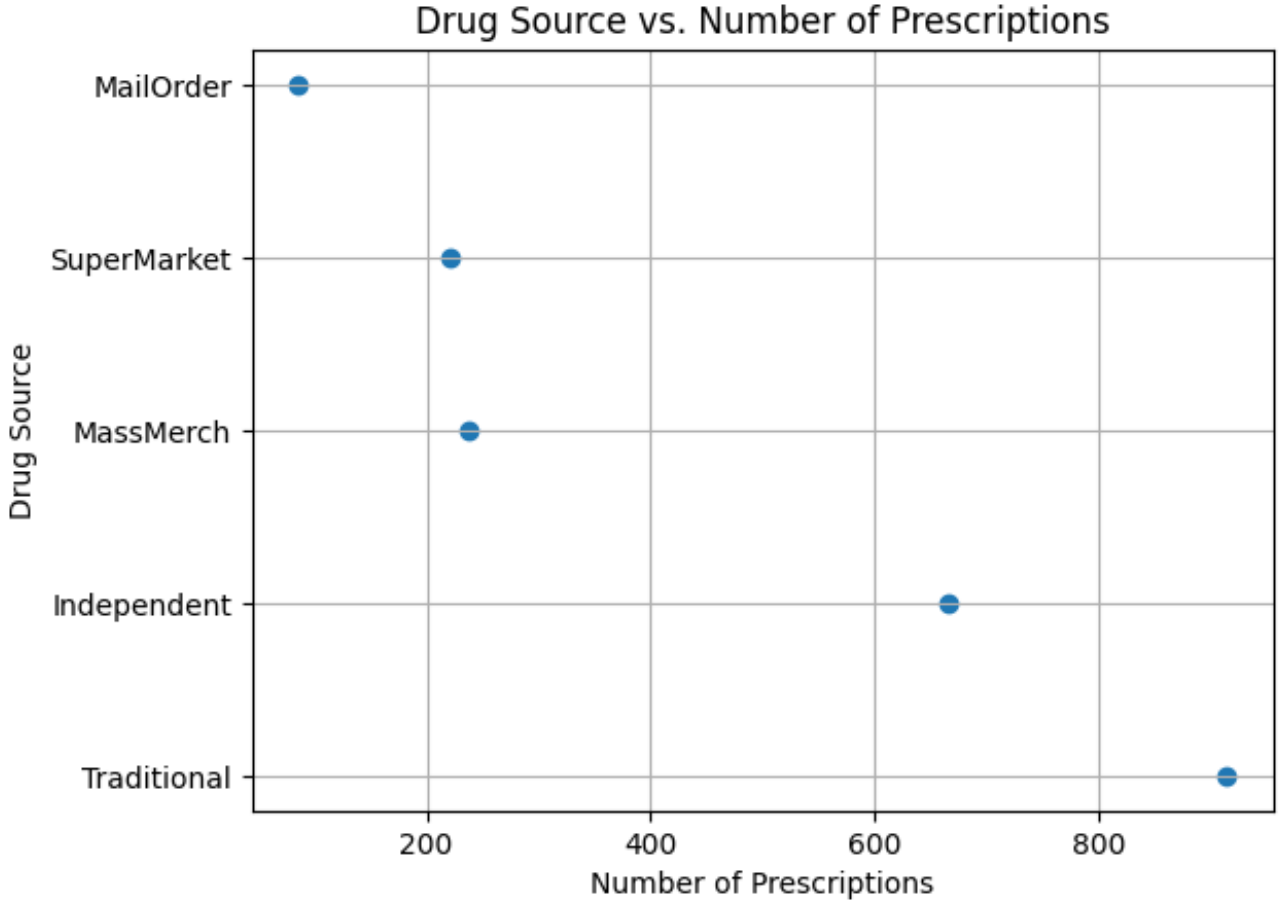
```
#| label: pie-chart

import matplotlib.pyplot as plt

# Assuming df is already created in the previous code

plt.figure()
plt.pie(df['Num Prescriptions'], labels=df['Drug Source'])
plt.title('Drug Source')
plt.show()
```

Dot Plot


Drug Source vs. Number of Prescriptions
MailOrder
SuperMarket
MassMerch
Independent
Traditional
Drug Source
200
400
600
800
Number of Prescriptions

Dot Plot

```
#| label: dot-plot

import pandas as pd
import matplotlib.pyplot as plt

# Assuming df is already created in the previous code
plt.figure()
plt.scatter(df['Num Prescriptions'], df['Drug Source'])
plt.xlabel('Number of Prescriptions')
plt.ylabel('Drug Source')
plt.title('Drug Source vs. Number of Prescriptions')
plt.grid(True)
plt.show()
```

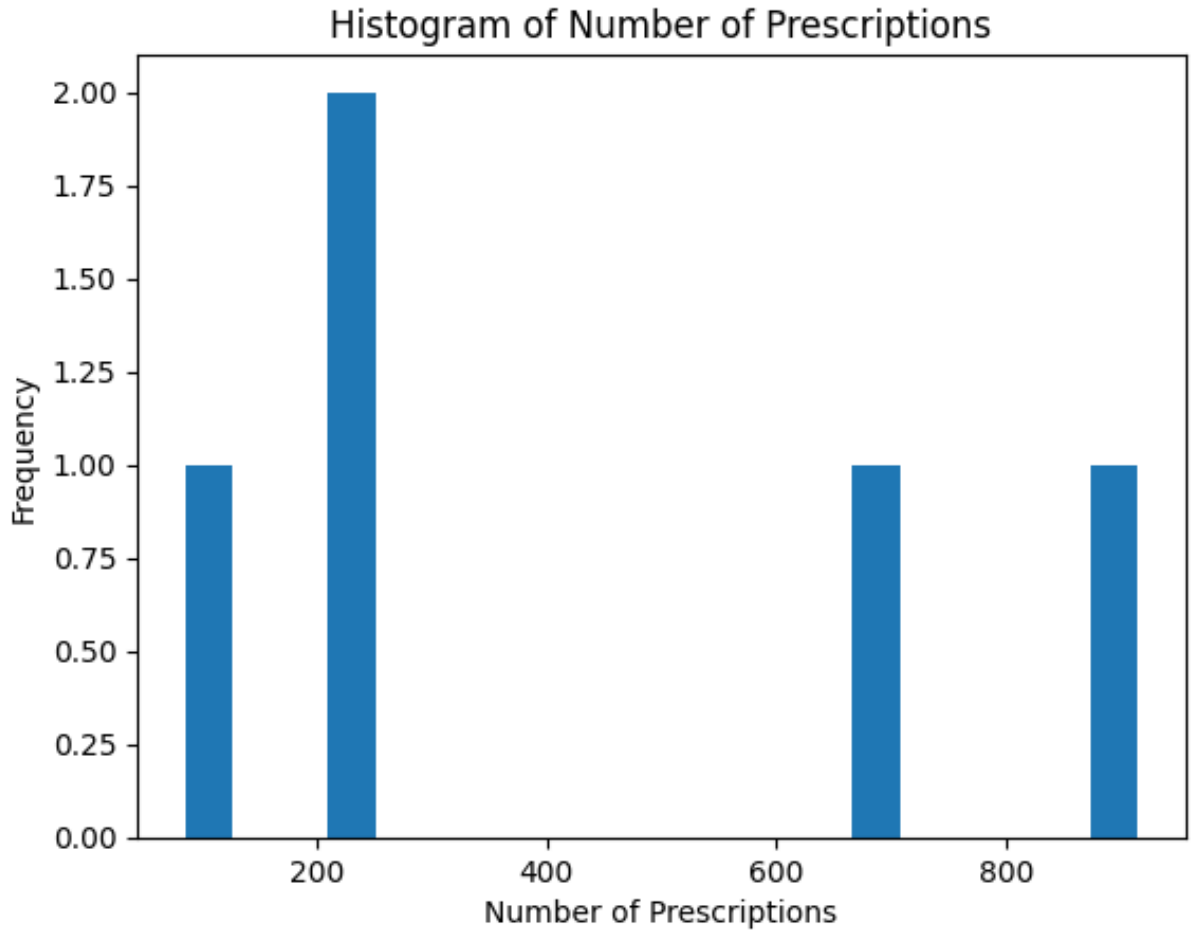
Additional Types

- histogram
- bar plot
- line plot
- scatter plot
- heat map
- box and whisker plot

Additional Types (Continue) - Which plot should I choose?

- ← occurrences (binned)
- ← processed categories
- ← suggestion of continuity
- ← looking for relationships in continuous data
- ← three variables in 2D
- ← statistics about single variable

Histogram


Histogram of Number of Prescriptions
Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
200
400
600
800
Number of Prescriptions

Histogram

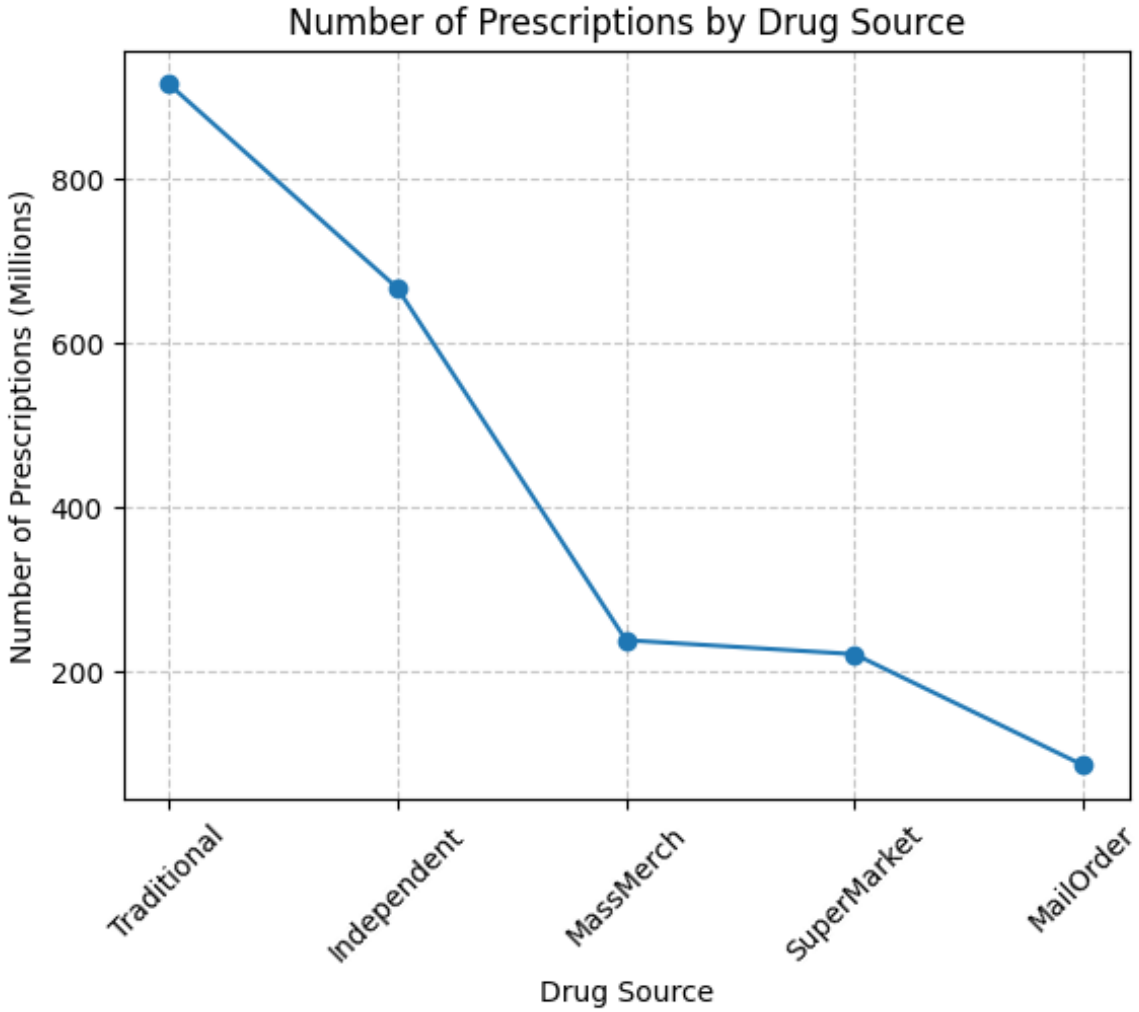
```
#| label: histogram

import matplotlib.pyplot as plt

# Assuming df is already created in the previous code

plt.figure()
plt.hist(df['Num Prescriptions'], bins=20)
plt.xlabel('Number of Prescriptions')
plt.ylabel('Frequency')
plt.title('Histogram of Number of Prescriptions')
plt.show()
```

Line Plot


Number of Prescriptions by Drug Source
Number of Prescriptions (Millions)
800
600
400
200
Traditional
Independent
MassMerch
SuperMarket
MailOrder
Drug Source

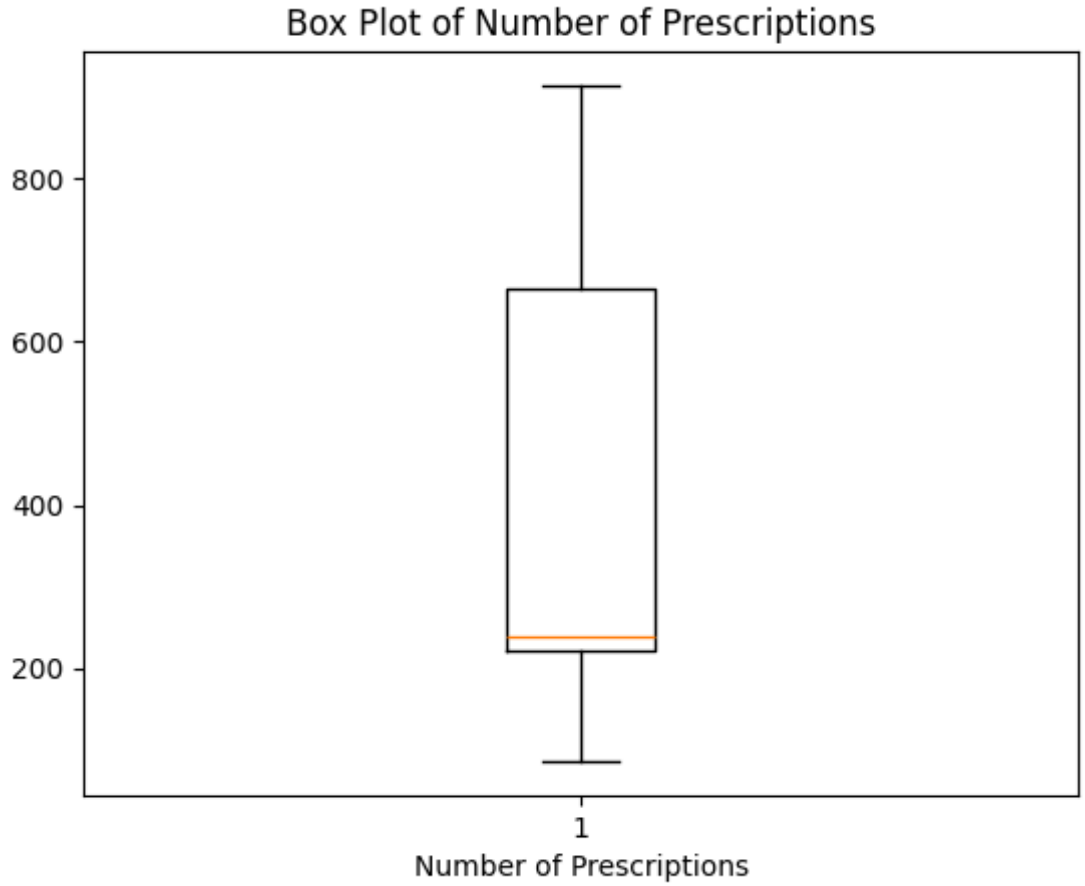
Line Plot

```
#| label: line-plot

import matplotlib.pyplot as plt

# Assuming df is already created in the previous code
plt.figure() # Adjust figure size for better visualization
plt.plot(df['Drug Source'], df['Num Prescriptions'], marker='o', linestyle='-')
plt.xlabel('Drug Source')
plt.ylabel('Number of Prescriptions (Millions)')
plt.title('Number of Prescriptions by Drug Source')
plt.grid(True, linestyle='--', alpha=0.7) # Add a grid for better readability
plt.xticks(rotation=45) # Rotate x-axis labels for better visibility
plt.show()
```

Box and Whisker Plot



Box and Whisker Plot

```
#| label: box

import matplotlib.pyplot as plt

# Assuming df is already created in the previous code
plt.figure()
plt.boxplot(df['Num Prescriptions']) # Create the boxplot
plt.xlabel('Number of Prescriptions')
plt.title('Box Plot of Number of Prescriptions')
plt.show()
```

More Example Data - Whiteboard

TABLE 2.4 Contingency Table Summarizing Counts of Cars Based on the Number of Cylinders and Ranges of Fuel Efficiency (mpg)

	Cylinders = 3	Cylinders = 4	Cylinders = 5	Cylinders = 6	Cylinders = 8	Totals
mpg (5.0–10.0)	0	0	0	0	1	1
mpg (10.0–15.0)	0	0	0	0	52	52
mpg (15.0–20.0)	2	4	0	47	45	98
mpg (20.0–25.0)	2	39	1	29	4	75
mpg (25.0–30.0)	0	70	1	4	1	76
mpg (30.0–35.0)	0	53	0	2	0	55
mpg (35.0–40.0)	0	25	1	1	0	27
mpg (40.0–45.0)	0	7	0	0	0	7
mpg (45.0–50.0)	0	1	0	0	0	1
Totals	4	199	3	83	103	392

More Example Data - Whiteboard

- ← occurrences (binned)
- ← processed categories
- ← suggestion of continuity
- ← looking for relationships in continuous data
- ← three variables in 2D
- ← statistics about single variable

Anatomy of a Graph

- legend
- markers
- marker labels
- axis labels
- axis units
- tick marks
- title
- caption
- panels

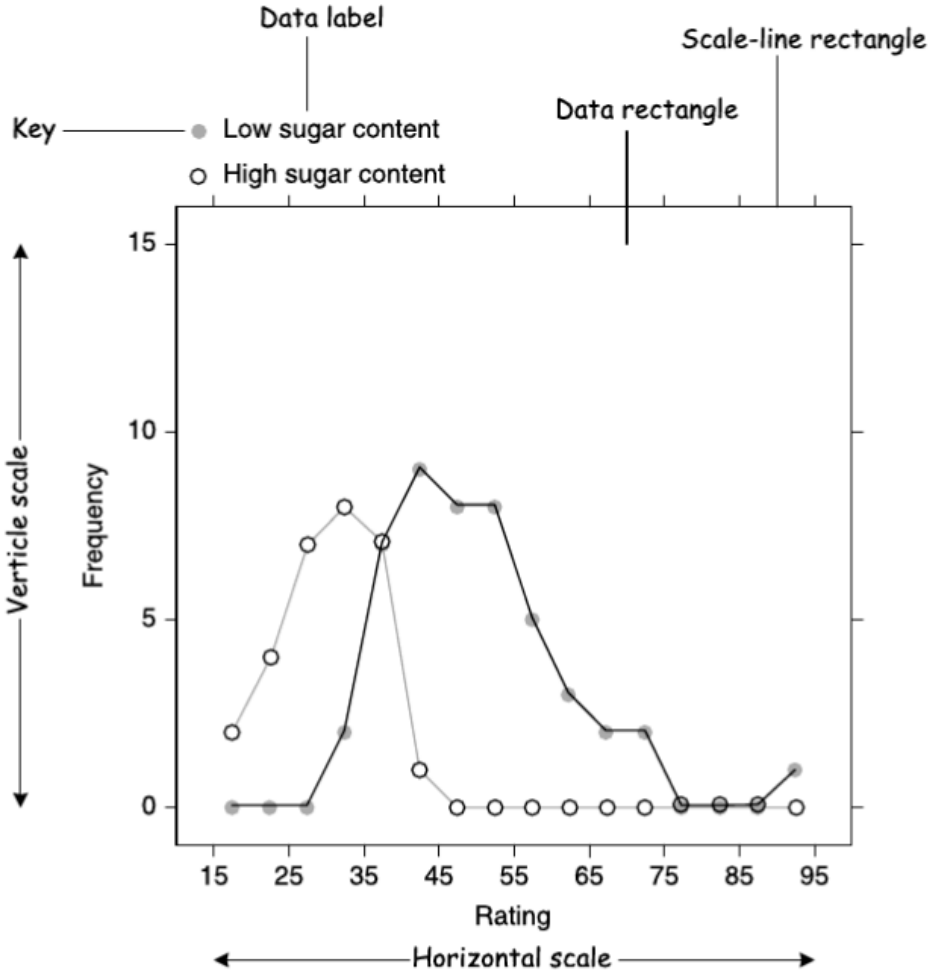


Figure 2.13 Anatomy of a graph

Group Activity

- Form pairs
- Take notes
- Interview your partner to find out about a data visualization that they recently admired
- What did the visualization make clear that was unclear before?
- What were all the salient features used to communicate information?
- Present your partner's visualization

Pandas

- pandas is a Python library for working with tabular data, similar to spreadsheets or database tables.
- It introduces three key data structures: Series (one-dimensional), DataFrame (two-dimensional tables with rows and columns) and panel (three-dimensional).
- pandas makes it easy to load, explore, clean, analyze, and visualize data using simple, readable code.

Pandas

TYPES OF DATA STRUCTUE IN PANDAS

Data Structure	Dimensions	Description
Series	1	1D labeled homogeneous array, sizeimmutable.
Data Frames	2	General 2D labeled, size-mutable tabular structure with potentially heterogeneously typed columns.
Panel	3	General 3D labeled, size-mutable array.

Matplotlib

- Matplotlib is a Python plotting library
- Produces publication-quality figures in Python in a variety of hardcopy formats and interactive environments across platforms.
- Allows you to plot your data without much extra coding

Setting Up Virtual Environment

- Create a project directory

```
mkdir projects
cd projects
```

- Create virtual environment using Python

```
python3 -m venv myenv
# see the file tree
find . -not -path '*/\.*'
```

- Activate myenv the virtual environment

```
source myenv/bin/activate  # macOS/Linux
myenv\Scripts\activate     # Windows
```

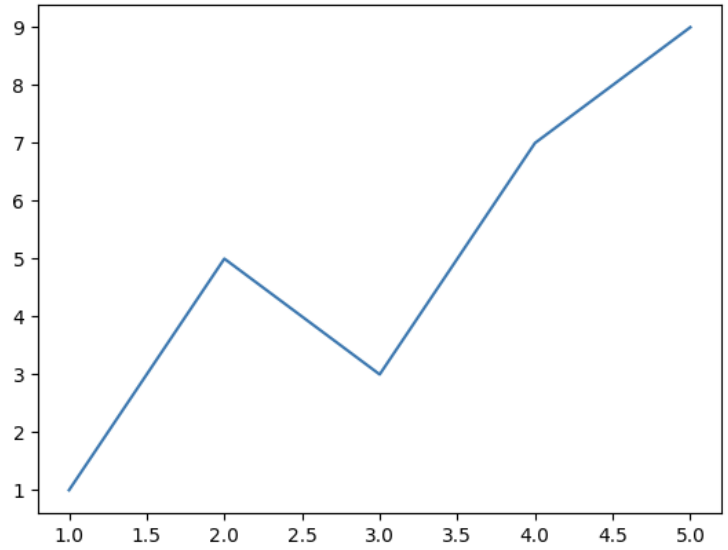
- Install Dependencies

```
pip install matplotlib
pip install numpy
```

Your First Plot

Plot some simple points

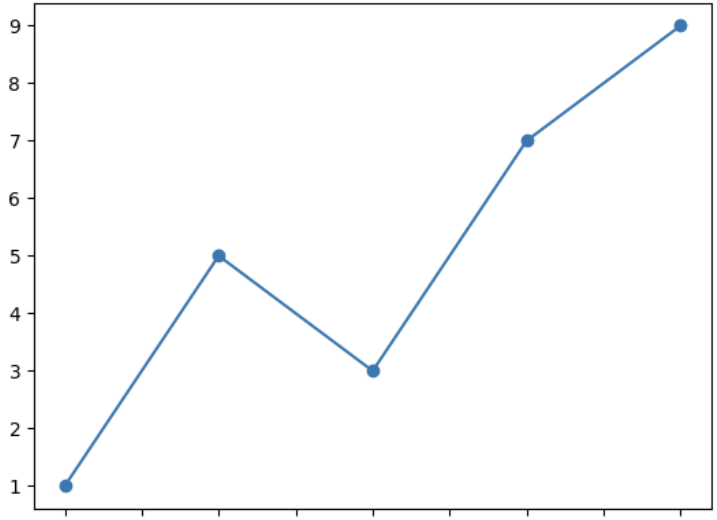
```
import matplotlib.pyplot as plt #get the library
x_num = [1,2,3,4,5] #def of x
y_num = [1,5,3,7,9] # def of y
plt.plot(x_num, y_num) # gives mem addr of obj
plt.show() # draw the plot on canvas
```



Gimme Points, Not Lines

Plot some basic numbers using points

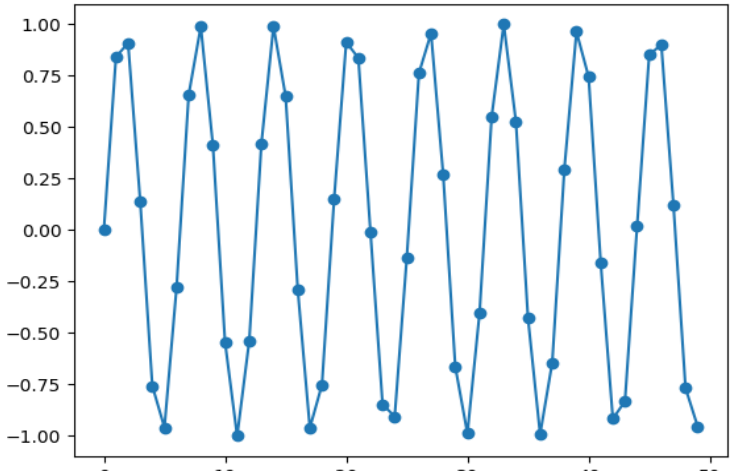
```
import matplotlib.pyplot as plt #get the library
x_num = [1,2,3,4,5] #def of x
y_num = [1,5,3,7,9] # def of y
plt.plot(x_num, y_num, marker='o')
# also including 'o', '*', 'x', and '+' as points
plt.show() # draw the plot on canvas
```



Another Amazing Example!

Plot the sin wave

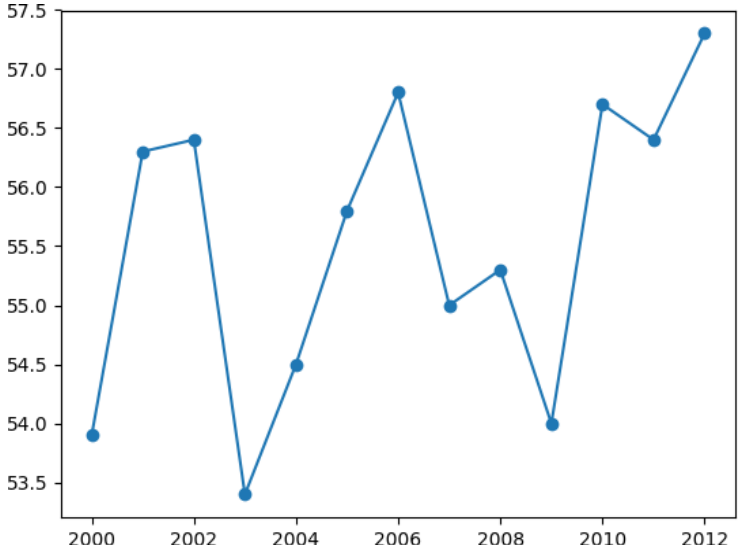
```
import matplotlib.pyplot as plt #get the library
import math
x_num = [i for i in range(50)]
y_num = [math.sin(i) for i in x_num]
plt.plot(x_num, y_num, marker='o')
# also including 'o', '*', 'x', and '+' as points
plt.show() # draw the plot on canvas
```

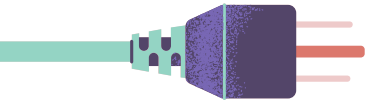
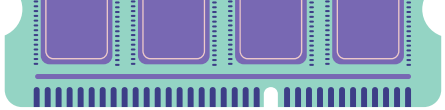


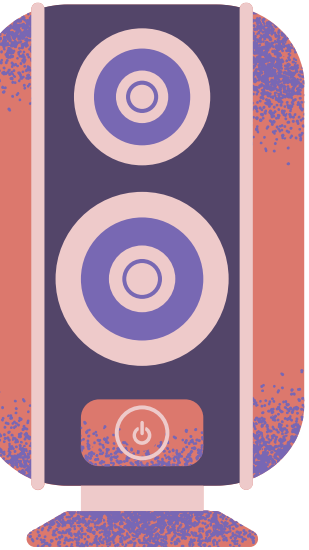
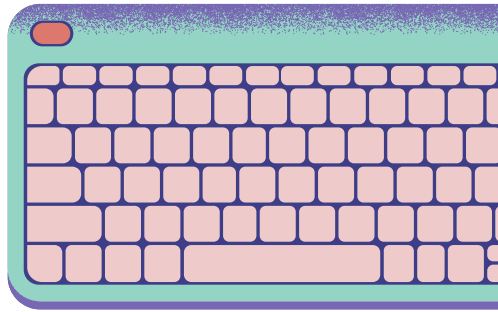
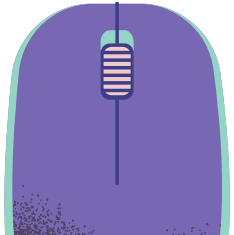
Yet, Another Amazing Example!

Plot the temperature in NYC and save the file too!

```
import matplotlib.pyplot as plt
nyc_temp = [53.9, 56.3, 56.4, 53.4, 54.5, 55.8, 56.8, 55.0, 55.3, 54.0, 56.7, 56.4, 57.3]
years = range(2000, 2013)
plt.plot(years, nyc_temp, marker='o')
# also including 'o', '*', 'x', and '+' as points
plt.savefig('mygraph.png') #save in root directory
plt.show() # draw the plot on canvas
```



THANKS