UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250 Foundations of Data Science

Module 6: Descriptive Statistics and Visualization

Course Plan

Module Titles
Module 1 – Introduction to Data Science
Module 2 – Introduction to Python
Module 3 – NumPy
Module 4 – Pandas
Module 5 – Data Collection and Cleaning
Current Focus: Module 6 – Descriptive Statistics and Visualization
Module 7 – Workshop (No Content)
Module 8 – Time Series
Module 9 – Introduction to Regression and Classification
Module 10 – Databases and SQL
Module 11 – Data Privacy and Security
Module 12 – Term Project Presentations (no content)

Learning Outcomes for this Module

- Review the main concepts of descriptive statistics, including mean, median and standard deviation
- Review correlation
- Explain the role and power of data visualization
- Learn from some classic examples
- Pick up some pointers for designing effective visualizations
- Describe some of the new technologies behind rich, interactive Web visualizations
- Use matplotlib to create data plots and charts

Topics for this Module

- **6.1** Descriptive Statistics
- **6.2** Correlation
- **6.3** Data Visualization
- **6.4** Matplotlib
- **6.5** Resources and Homework

Module 6 – Section 1

Descriptive Statistics

Types of Data

- Numerical (Quantitative)
  - Discrete:
    - Measured quantities
    - Results of experiments
    - Numerical values obtained by counting
  - Continuous:
    - Value obtained by measuring (e.g. height of all students)
    - All values in a given interval of numbers (e.g. federal spending)
- Categorical (Qualitative)
  - Ordinal:
    - Natural ordering (e.g. "hot", "medium", "cold")
  - Nominal:
    - Any categorical data that doesn't have an order (e.g. "blue", "red", "green")
- Other e.g. Text, Video, Binary

Data Distributions

- The probability distribution function of a ***categorical*** random variable is a list of probabilities associated with each of its possible values
  - How probable is each one of the values?

- A ***continuous*** random variable is one which takes an infinite number of possible values; for example: the speed of automobiles on a highway at any one location
  - A continuous random variable is not defined at any specific value.
  - Which values appear more likely than others, how much are they spread etc.

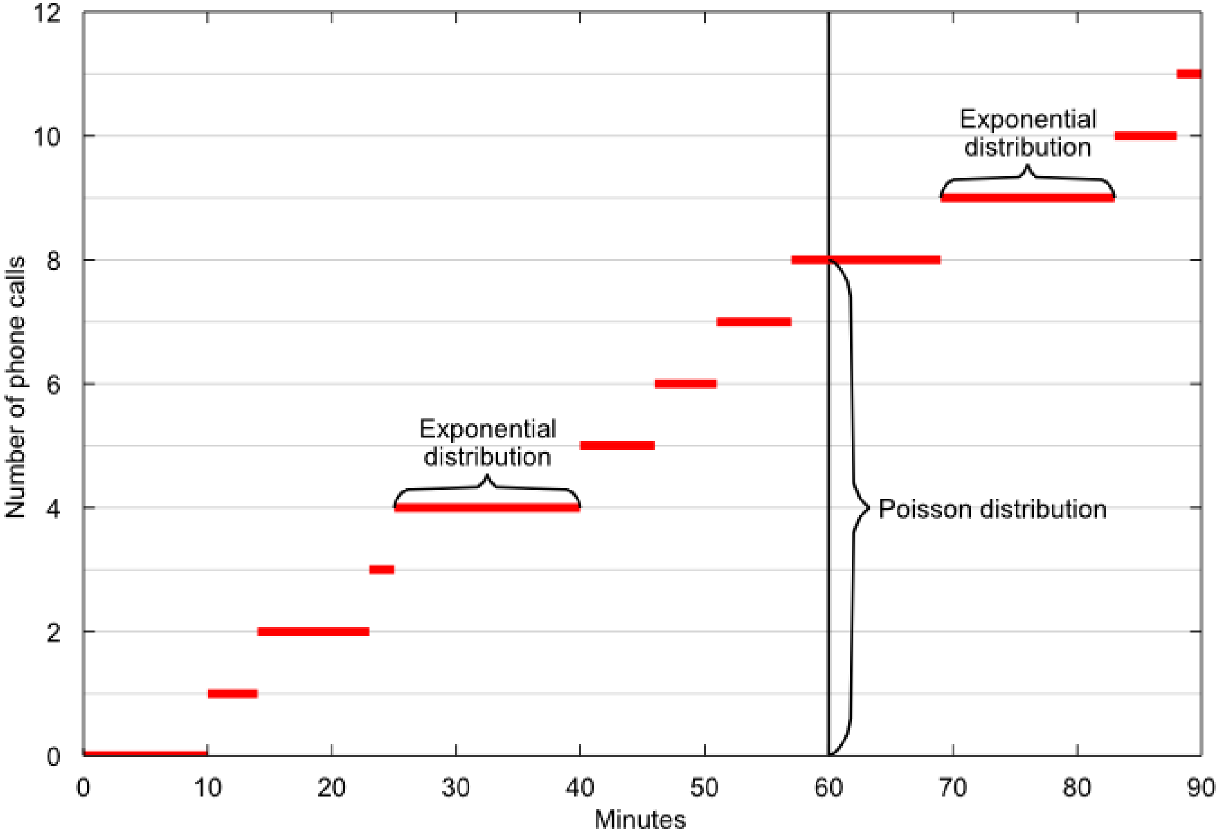
Standard Distributions

- The primary statistical modelling technique is to assume that one of the standard distributions will be a reasonable model
- There are many standard distributions that model real-world situations well because their mathematical form was derived from scientific study of physical processes

A Few Standard Distributions

- **Uniform**: Several outcomes are all equally likely e.g. which number comes up on a thrown die
- **Normal** (aka. Gaussian): The probability density falls off as the square of the distance from the mean e.g. heights of women
- **Poisson**: Rates of sparse events in space or time e.g. lightening strikes, machine failures
- **Exponential**: “Memoryless” decay e.g. number of light bulbs still working at a point in time

Poisson vs Exponential distribution



Descriptive Statistics for Continuous Data

- Measures of ***location*** (or ***center***), where the values are mostly “located”:
  - Mean
  - Median
  - Percentile
  - Geometric mean
  - Trimmed mean etc.
- Measures of ***spread***, how spread the data are:
  - Standard deviation
  - Variance
  - Range
  - Interquartile range etc.

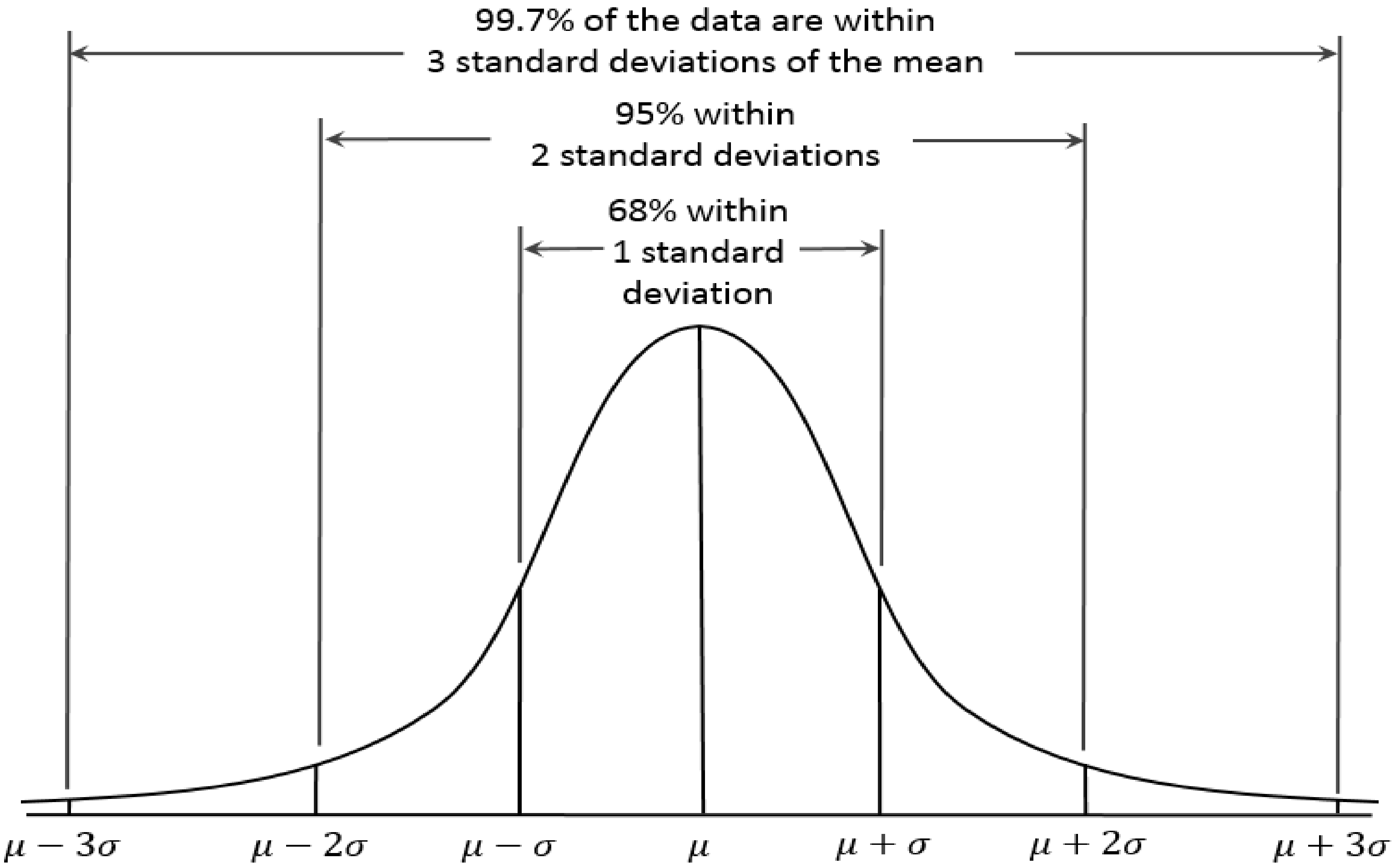
Summary Statistics

- Mean or expected value is the probability weighted average of all values
- In order words, it is a weighted average of the random variable by the likeliness of its outcomes
- The calculations are different for categorical and continuous random variables
- Categorical – calculated as a weighted sum
- Continuous – calculated as a weighted integral over its domain
- Median is the center value
- Median is less affected by outliers whereas mean is not

Summary Statistics (cont'd)

- Variance is the expectation of the squared deviation of a random variable from its mean
- Calculated similarly as mean with the weighted average of the random variable offset by its mean and the whole quantity squared
- Standard deviation is the square root of variance
- Range is the difference of the maximum and the minimum value
- Interquartile range is equal to the difference between the 75th and 25th percentiles (IQR = $Q_3 – Q_1$)
- Quartiles divide a rank ordered data set into 4 equal parts

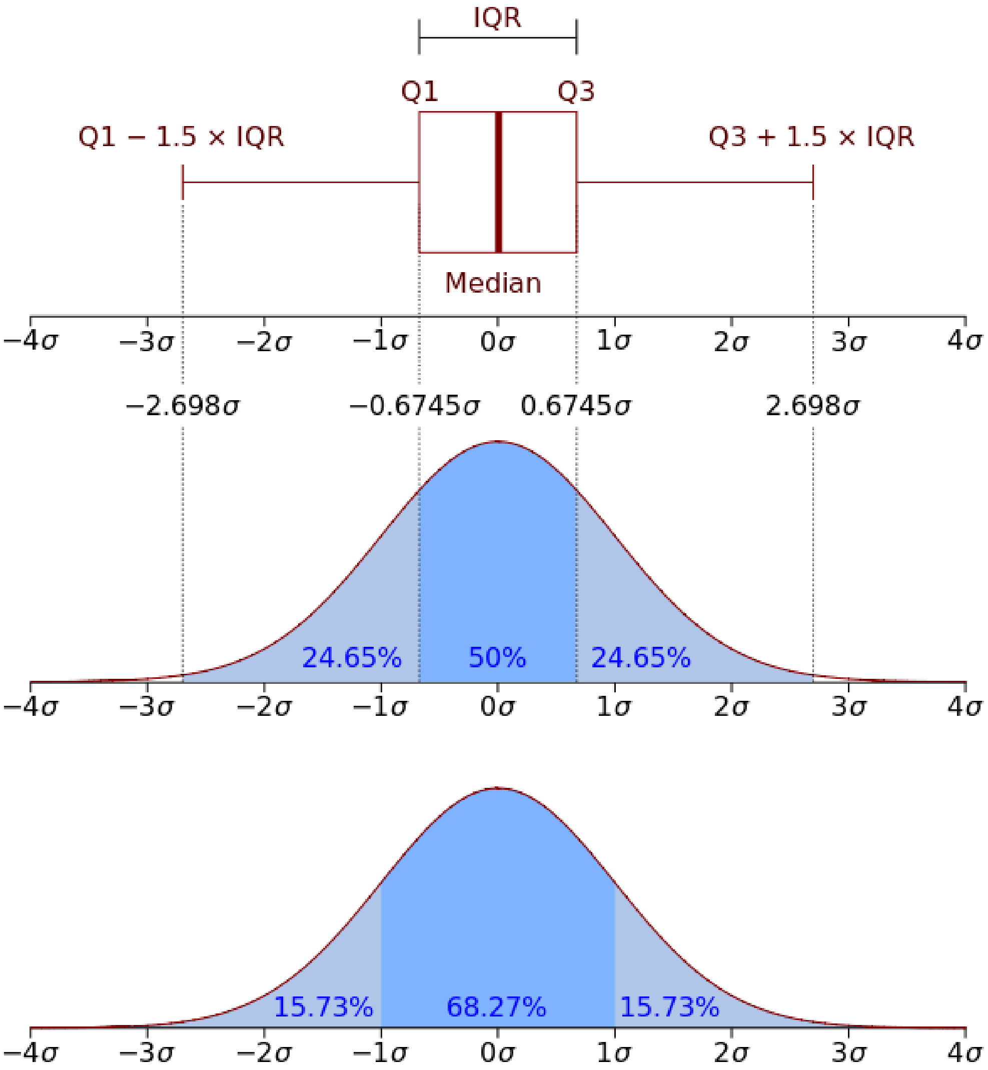
Measures of Distribution



99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Source

Measures of Distribution (cont'd)



UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Source

Measures of Distribution (cont'd)

- **Variance**: a measure of spread around the mean

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

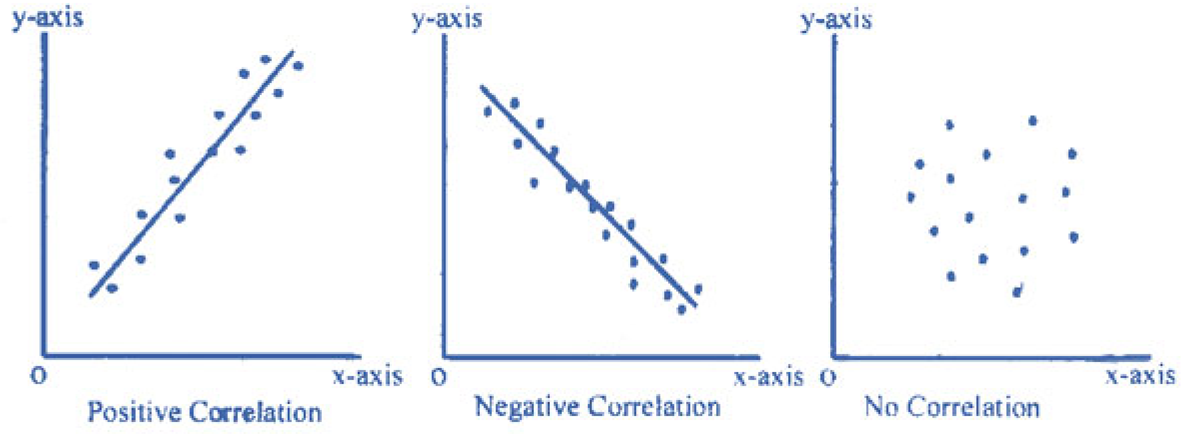
- **Standard deviation:** square root of variance
- **Range**: max-min
- **Interquartile range** (IQR): Q_3-Q_1
- **Coefficient of variation** (CV): $cv = \frac{s}{\bar{x}}$
  - CV is a measure of spread that describes the amount of variability relative to the mean
  - CV standardizes variability making it comparable across samples with different arithmetic means

Module 6 – Section 2

Correlation

Covariance and Correlation

- Two or more variables can co-vary:
  - High values of one are associated with high values of the other (positive correlation)
  - Or, vice versa (negative correlation)



Source

Covariance vs Correlation

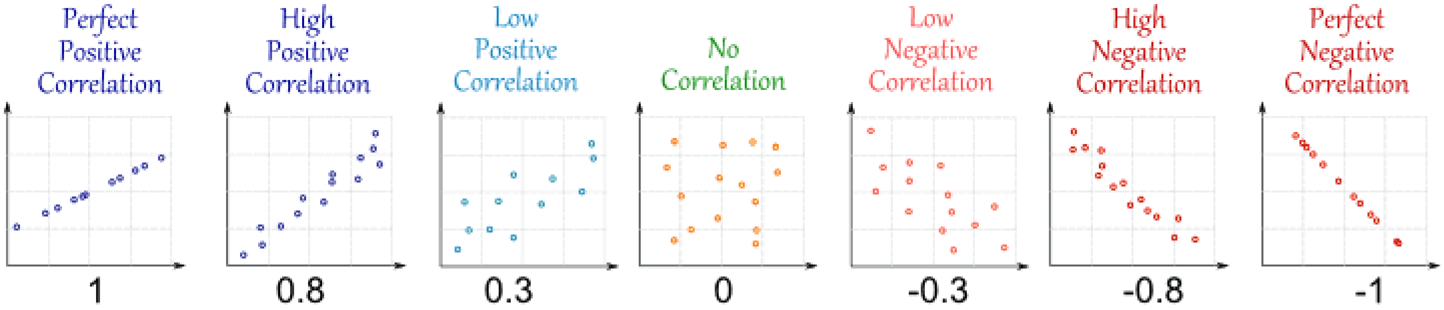
- Covariance measures the extent to which two (or more) variables move in tandem over time or tend to be both high or both low at the same time (co-vary)

- Correlation coefficient is a unit-less measure of covariance

- Correlation is a fundamental pattern matching technique

Correlation Coefficient (r)

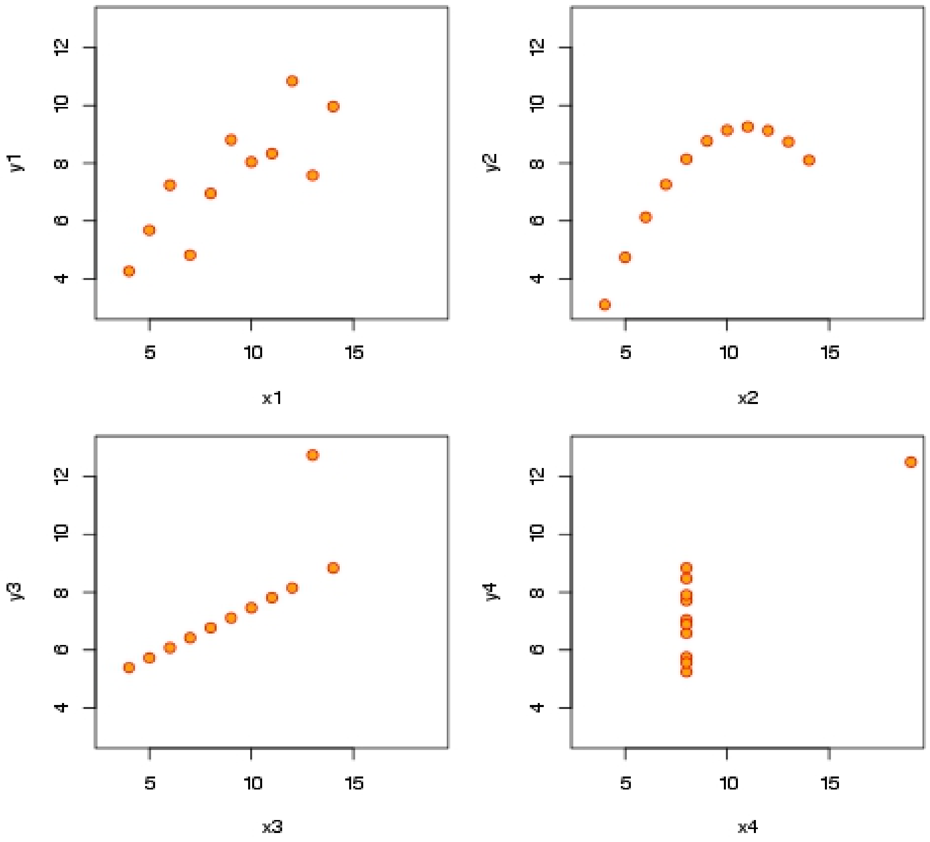
- -1: Perfectly negatively correlated
- < 0: Negative correlation
- Close to 0: Uncorrelated
- > 0: Positively correlated
- 1: Perfectly positively correlated



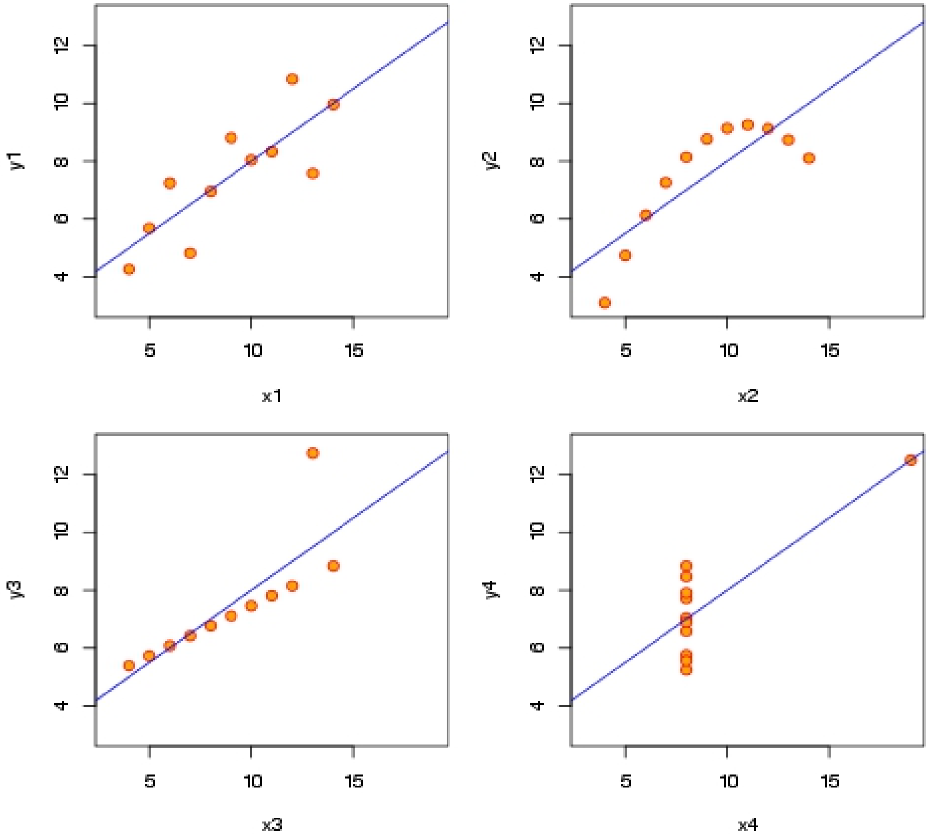
Correlation - Uses

- Might mean there is an underlying cause and effect to investigate further
- Can be used to reduce market risk (through negative correlation of returns)
- Through regression becomes a predictive tool

Quiz: Highest Correlation?



Anscombe's Quartet: r=0.816



Module 6 – Section 3

Data Visualization

The Role of Data Visualization

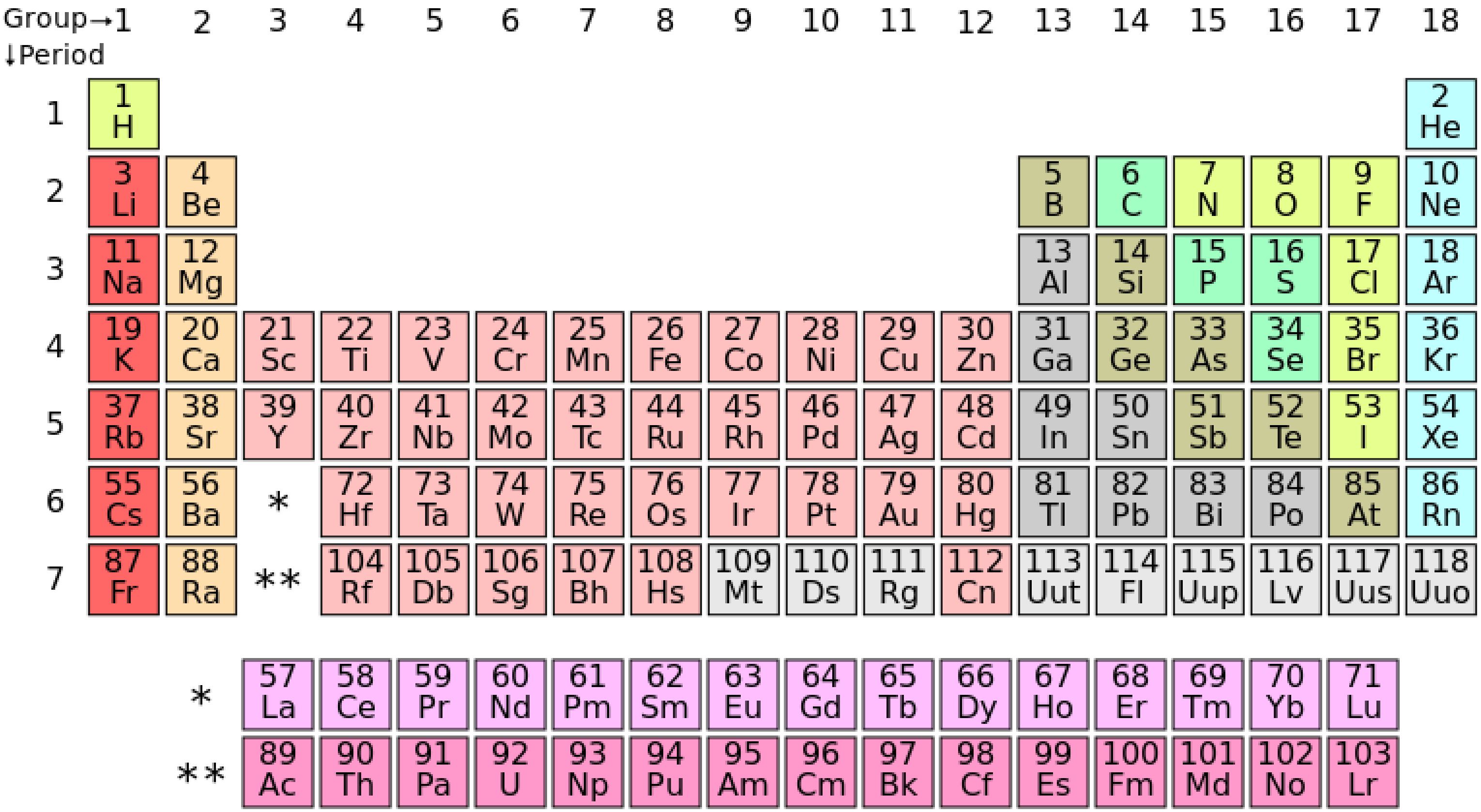
- Video by Nature Research "Science is Beautiful" about the role that visualization plays in the scientific process:

- Visualizations mentioned in the video:
  - Florence Nightingale's Rose Diagram and the story behind it
  - Circle of Life: biological data visualization
  - Perpetual Ocean, visualization of the ocean by NASA

A Few of the World’s Most Influential Visualizations

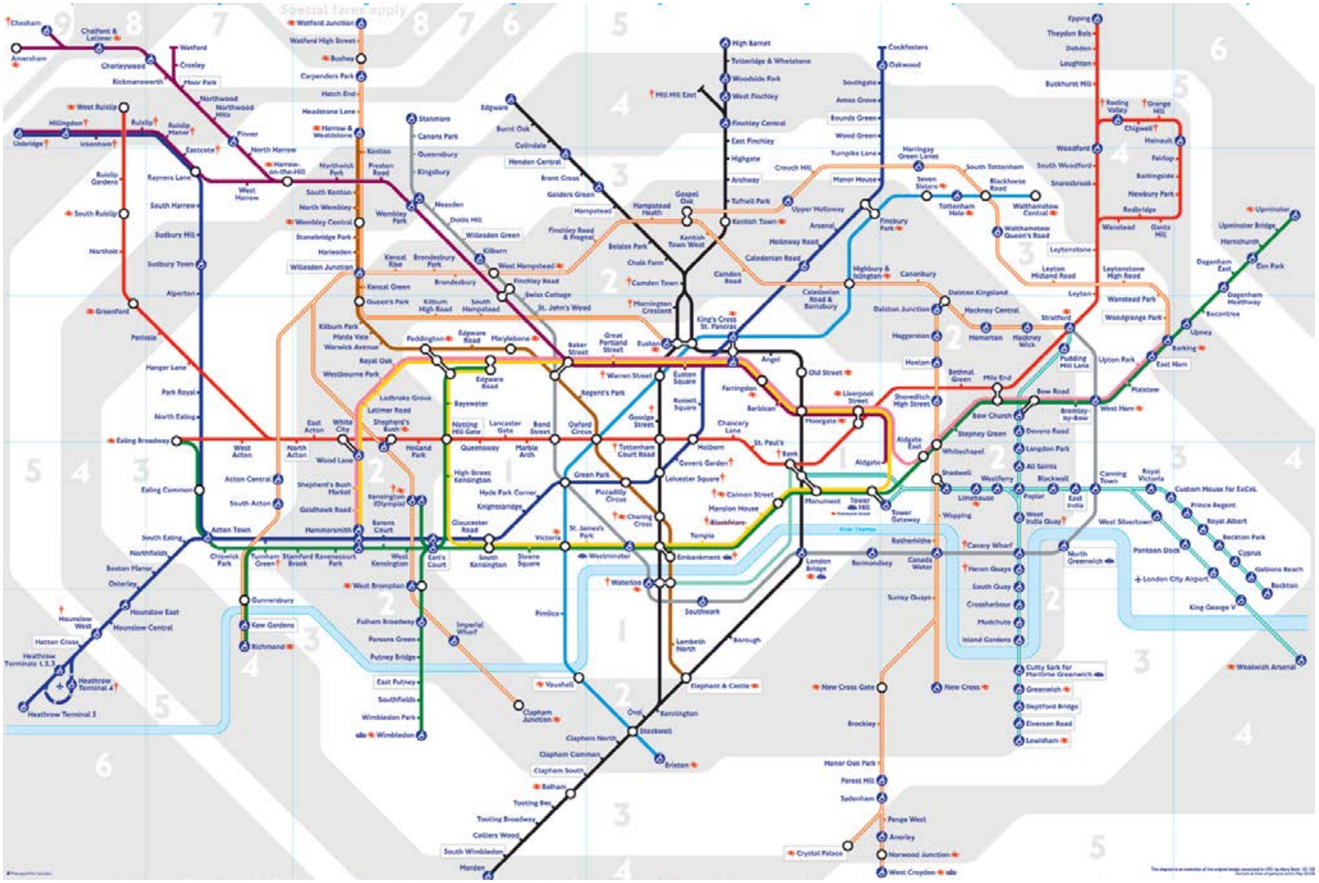
- Snow’s Cholera Map:
- Minard’s March on Moscow:
- Priestley’s World History:

Classics: The Periodic Table

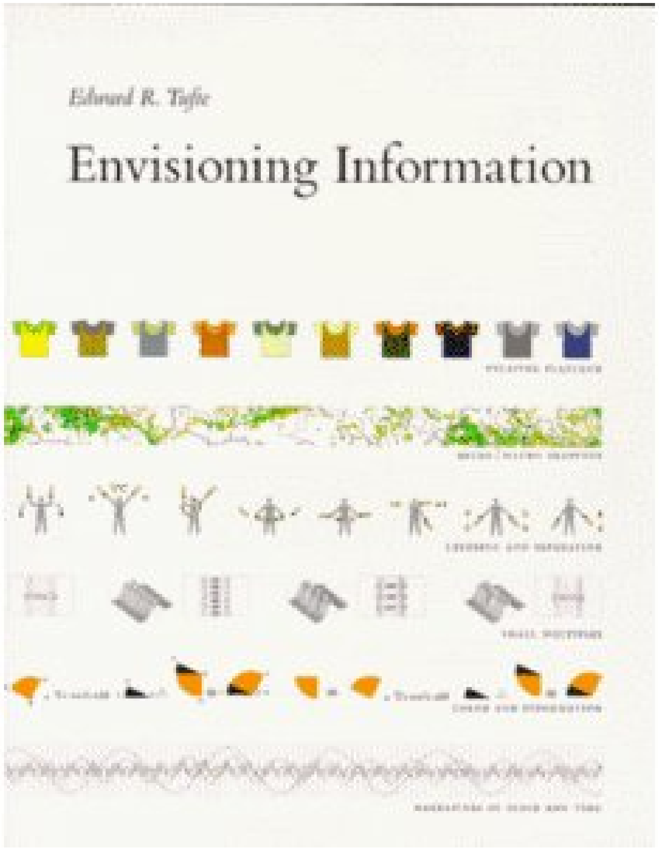


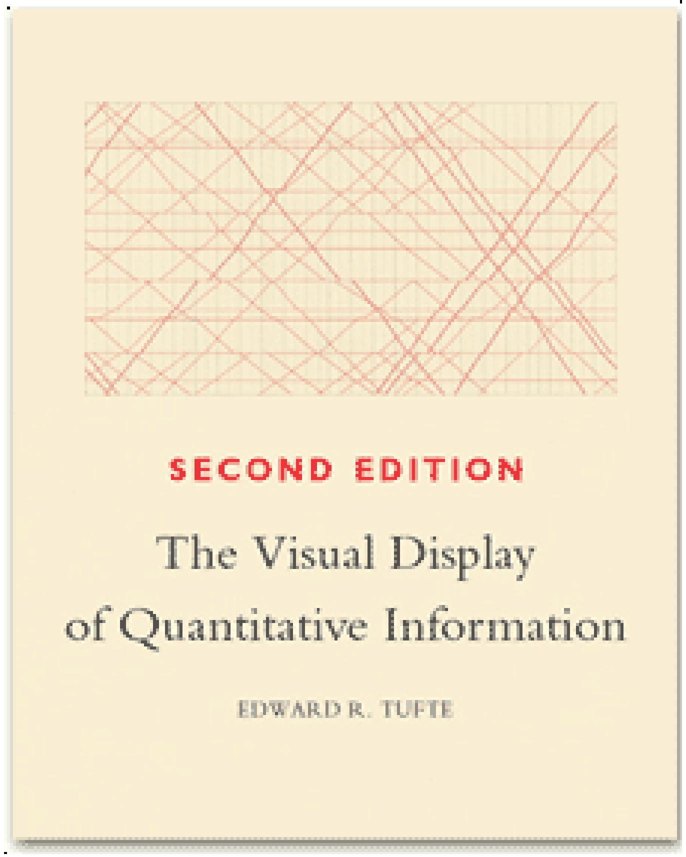
Group→ ↓Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo
		*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu	
		**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr	

Classics: The London Tube Map



Classics: Tufte’s Books





Some Useful Visualization & Infographic Sites

- visualisingdata.com
- infosthetics.com
- visualcomplexity.com
- flowingdata.com
- perceptualedge.com
- visual.ly
- visualdatahub.wordpress.com/
- Visualization techniques: www.visual-literacy.org/periodic_table/periodic_table.html
- Visualizing algorithms: bost.ocks.org/mike/algorithms/

Some More Excellent Videos on Visualization

- Jer Thorp (NY Times visualizations):
- Hans Rosling (Classic Ted presentation):
- Sarah Bird (Bokeh):

Visualization in Python

- Matplotlib:
- Bokeh:
- ggplot2 port to Python:
- Vispy GPU-powered scientific visualization:
- Google visualization API:

Visualization in Python (cont'd)

- Seaborn:
- Graph visualization: igraph.org/ and networkx.github.io
- Spyre:
- Holoviews:

Visualization in Python (cont’d)

- Altair:

JavaScript Libraries for Visualization

- d3.js:
- three.js: Jenga Example
- WebGL:
  - chromeexperiments.com/webgl/
  - developer.mozilla.org/en-US/docs/Web/WebGL

Module 6 – Section 4

Matplotlib

Matplotlib

- A general-purpose 2D plotting package in Python
- Makes use of NumPy to maintain good performance with large data sets
- Two APIs
  - MATLAB style
  - Object-Oriented

Key Benefits

- Easy & fast to use
- Attractive, publication-quality plots
- Excellent TeX formatting support

Making Fake Data

- Time Series Example

```
x = pd.period_range(‘2010-01-01’, periods=60, freq=“M”)
y = np.random.randn(len(x)).cumsum()
```

- Simple Function Example

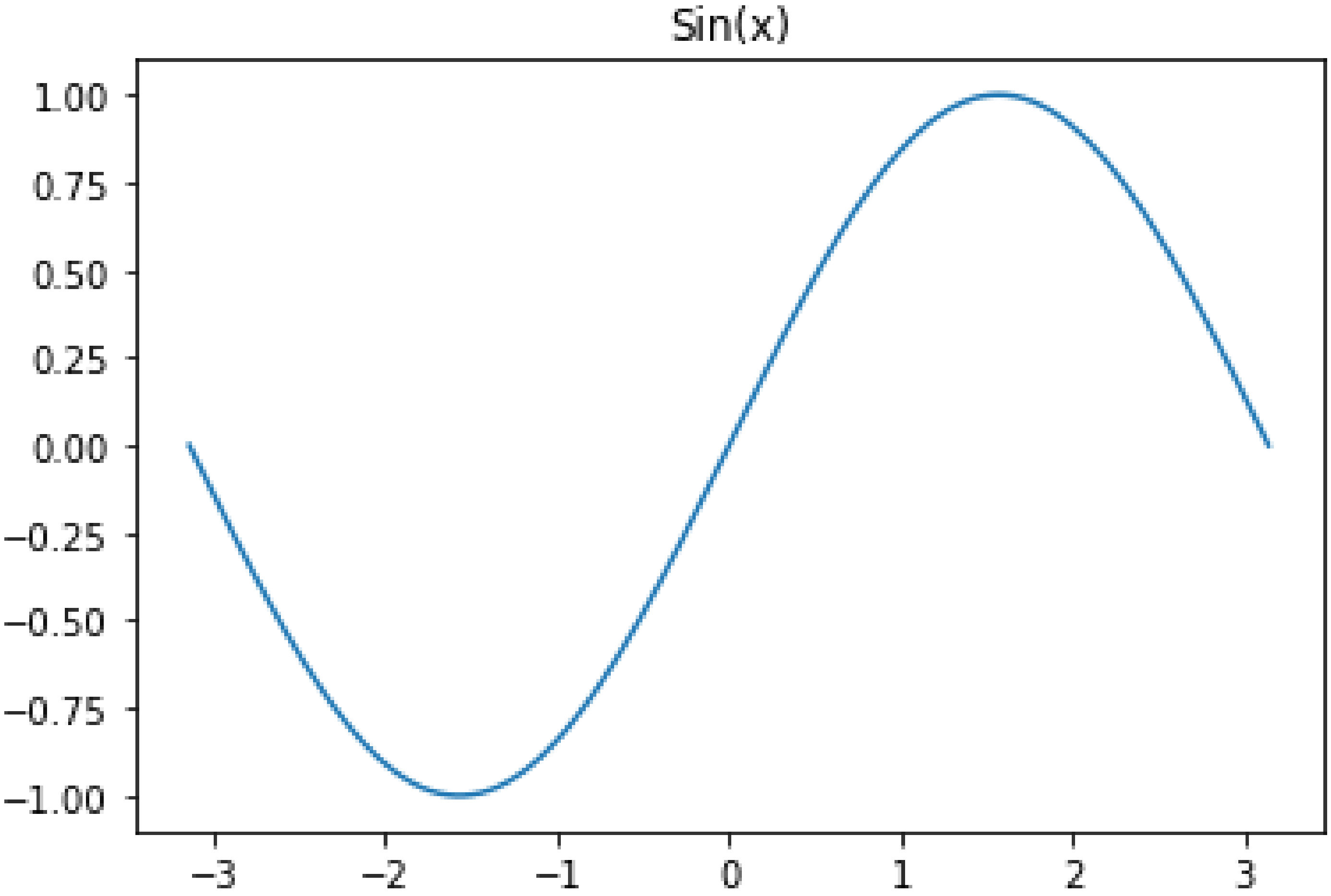
```
x = np.linspace(0, 1, 256, endpoint=True)
y = np.sin(x)
```

- 2-Variable Scatter Example

```
x = np.random.normal(0, 1, 100)
y = np.random.normal(0, 1, 100)
```

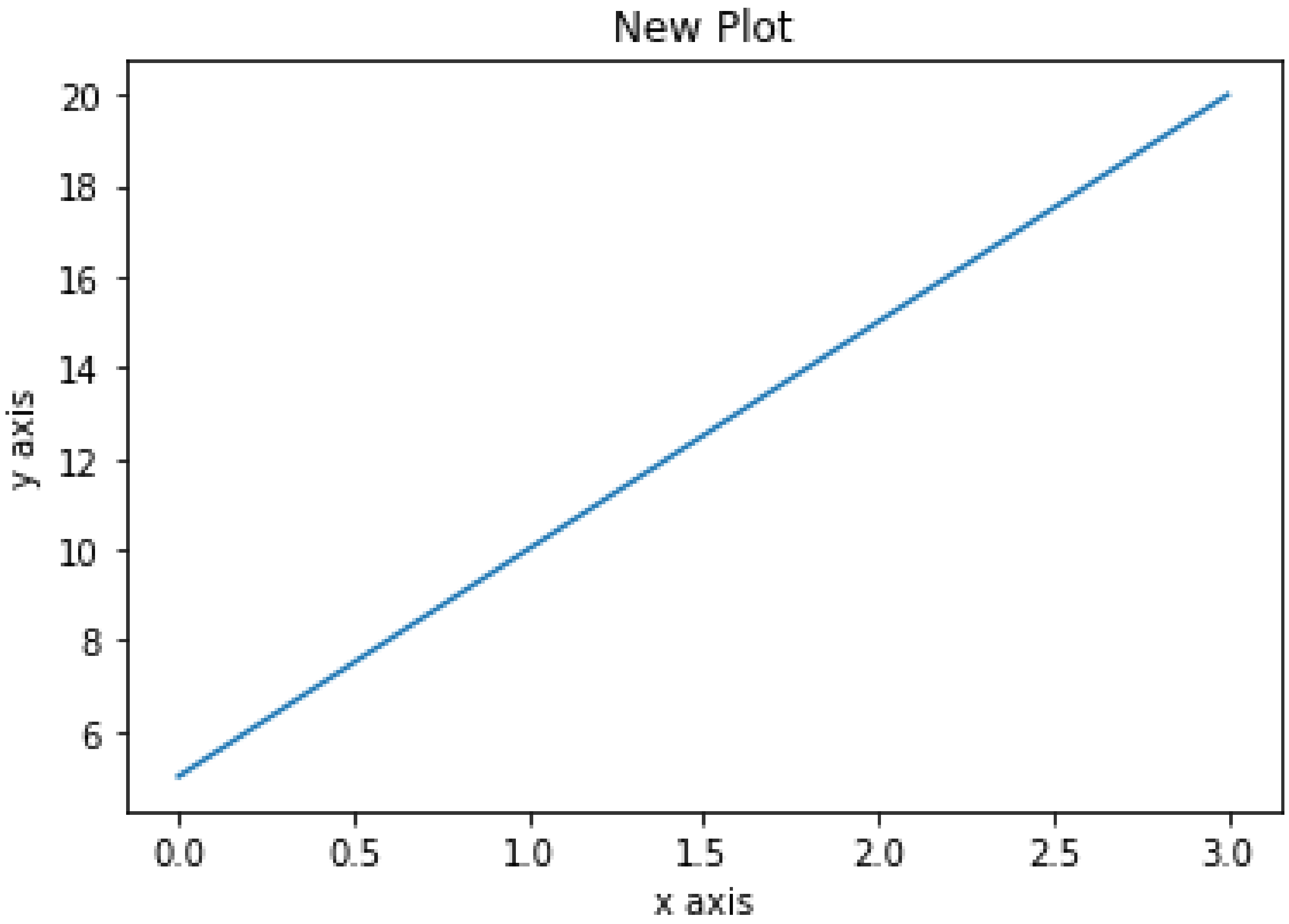
Basic Plots

```
import numpy as np
import matplotlib.pyplot as plt

X = np.linspace(-np.pi, np.pi, 256, endpoint=True)
S = np.sin(X)

plt.plot(X, S)

plt.title("Sin(x)")

plt.show()
```


Sin(x)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−3
−2
−1
0
1
2
3

Basic Plots (cont’d)

```
import matplotlib.pyplot as plt
plt.plot([5,10,15,20])
plt.title(“New Plot”)
plt.xlabel(“x axis”)
plt.ylabel(“y axis”)
plt.show()
```


New Plot
y axis
20
18
16
14
12
10
8
6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x axis

Basic Plots (cont’d)

```
import matplotlib.patches as mpatches
import matplotlib.pyplot as plt
plt.plot([10,15,20,25,30,35,40], [6,9,13,18,25,35,48])
plt.axis([5, 45, 1, 60])
plt.title("Figure #1")
plt.xlabel("X")
plt.ylabel("Y")
plt.grid(True)
blue_patch = mpatches.Patch(color="blue",
label="Data")
plt.legend(handles=[blue_patch])
plt.show()
```

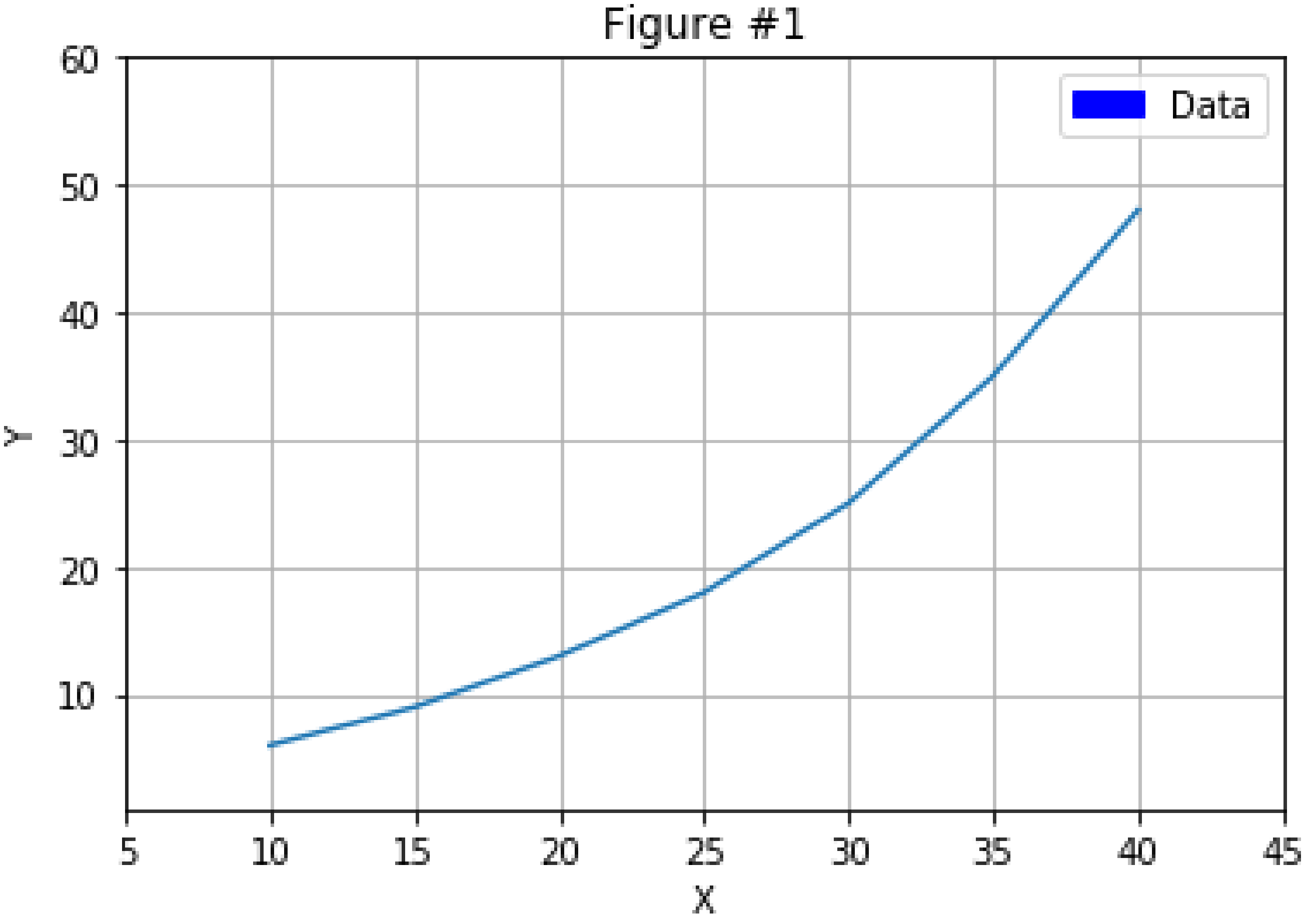
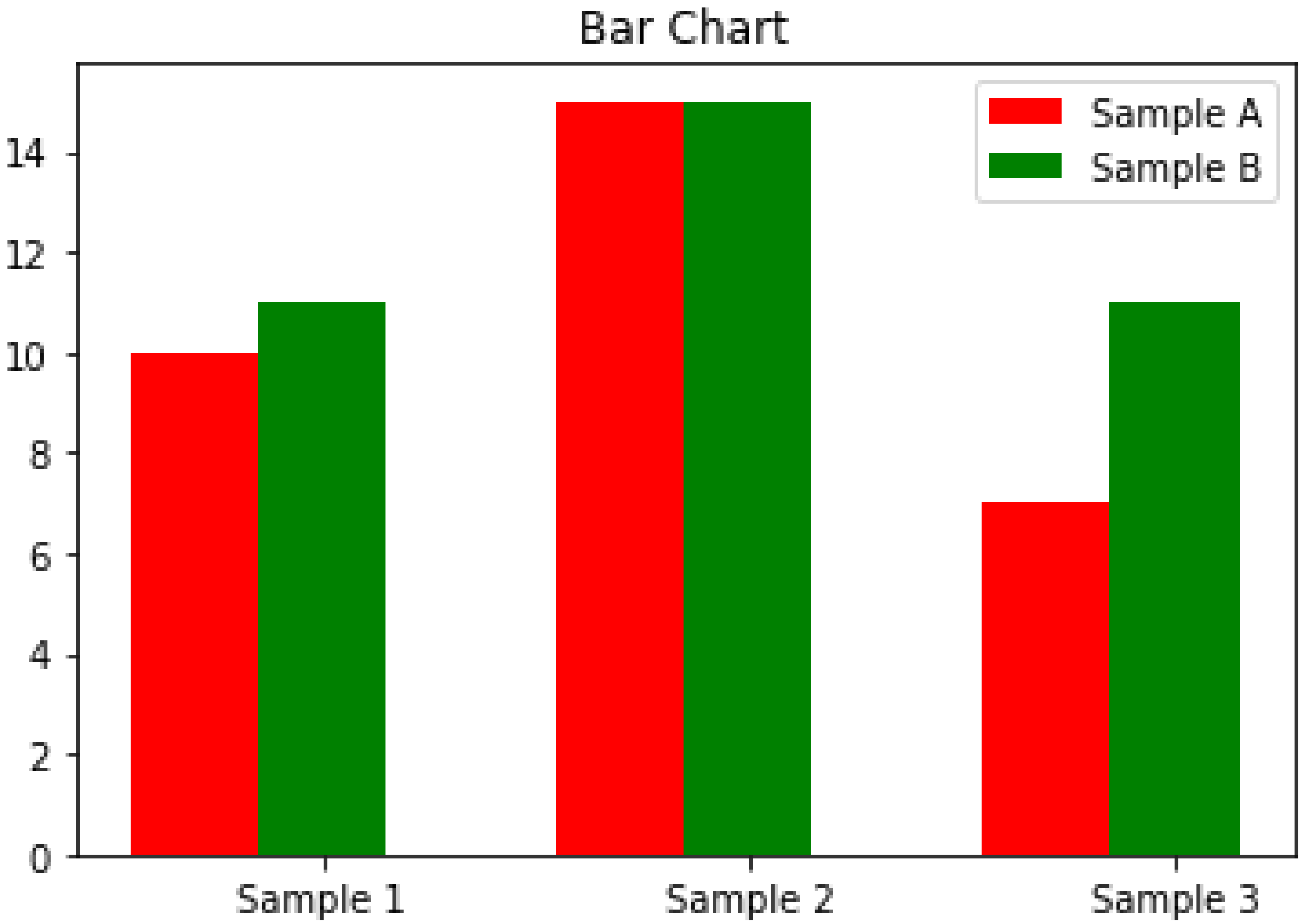


Figure #1
Data
Y
X
60
50
40
30
20
10
5
10
15
20
25
30
35
40
45

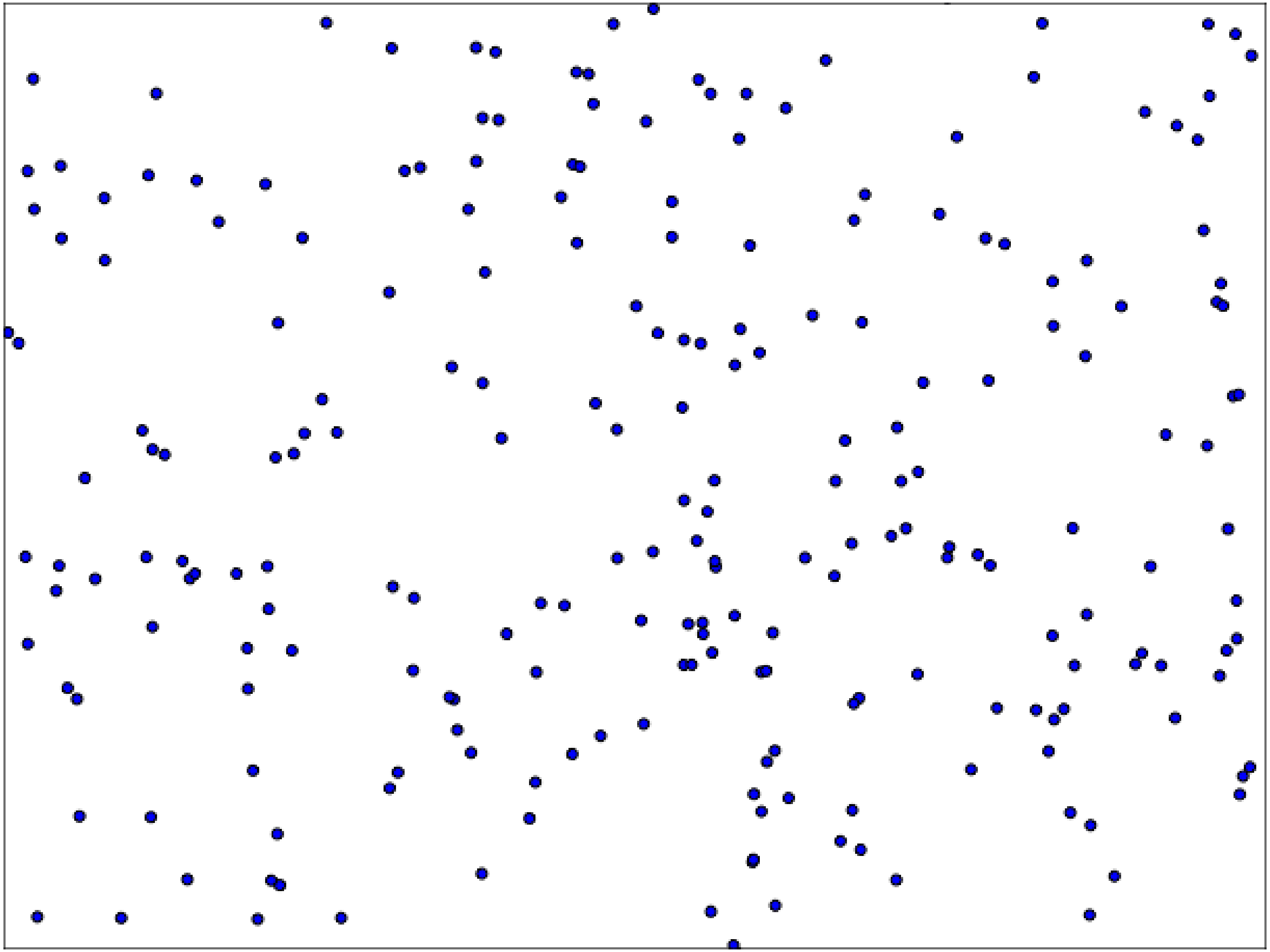
Bar Chart

```
import numpy as np
import matplotlib.pyplot as plt

N = 3
a = (10, 15, 7)
b = (11, 15, 11)

ind = np.arange(N)
width = 0.3
fig, ax = plt.subplots()

rects1 = ax.bar(ind, a, width, color='red')
rects2 = ax.bar(ind+width, b, width, color='green')

# add some text for labels, title and axes ticks
ax.set_title('Bar Chart')
ax.set_xticks(ind + width)
ax.set_xticklabels(('Sample 1', 'Sample 2', 'Sample 3'))

ax.legend((rects1[0], rects2[0]), ('Sample A', 'Sample B'))

plt.show()
```


Bar Chart
Sample A
Sample B
14
12
10
8
6
4
2
0
Sample 1
Sample 2
Sample 3

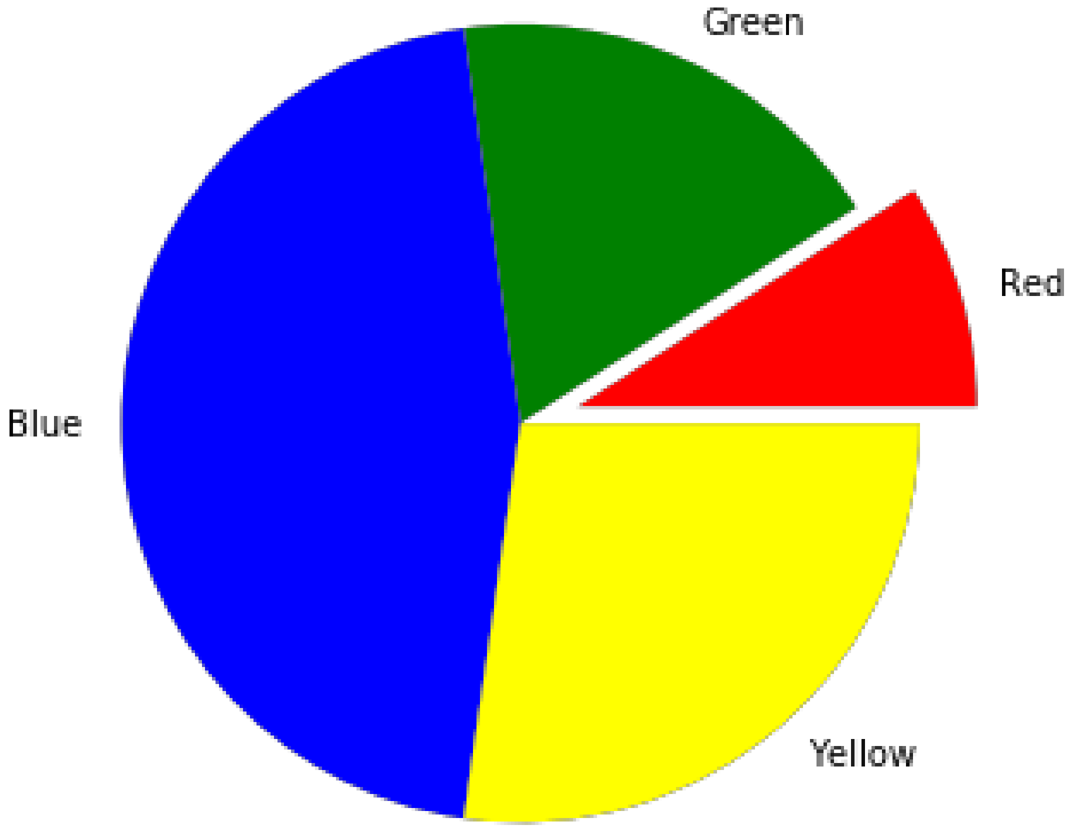
Scatter Plot

```
import numpy as np
import matplotlib.pyplot as plt

n = 500
X = np.random.normal(0, 1, n)
Y = np.random.normal(0, 1, n)

plt.axes([-1, -1, 1, 1])
plt.scatter(X, Y)

plt.xlim(-1, 1)
plt.xticks(())
plt.ylim(-1, 1)
plt.yticks(())

plt.show()
```



Pie Chart

```
import numpy as np
import matplotlib.pyplot as plt

N = 5
data = [1, 2, 5, 3]
plt.axes([0, 0, 0.9, 0.9])

plt.pie(data, explode = (0.15, 0, 0, 0), labels =
("Red", "Green", "Blue", ”Yellow"), colors = ["red",
"green", "blue", ”yellow"])
plt.axis('equal')
plt.xticks()
plt.yticks()

plt.show()
```


Green
Red
Blue
Yellow

Common Settings

- Colour, Linewidth

```
plt.plot(X, color="blue", linewidth=3, linestyle="-")
```

- Axis Range

```
plt.xlim(X.min() * 1.1, X.max() * 1.1)
plt.ylim(S.min() * 1.1, S.max() * 1.1)
```

- Tick Marks

```
plt.xticks([-2, -1, 0, 1], [-2, -1, 0, 1])
plt.yticks([-2, -1, 0, 1], [-2, -1, 0, 1])
```

Common Settings (cont’d)

- Spines

```
ax.spines['left'].set_position('center')
ax.spines['bottom'].set_position('center')
ax.spines['left'].set_smart_bounds(True)
ax.spines['bottom'].set_smart_bounds(True)
ax.xaxis.set_ticks_position('bottom')
ax.yaxis.set_ticks_position('left’)
```

- Legend

```
blue_patch = mpatches.Patch(color="blue",
label="Data")
plt.legend(handles=[blue_patch])
```

Saving Figures

- Figures can be saved
  - figure_name.savefig(“file_name.file_type”)
    - .png
    - .jpg
    - .pdf
    - etc.

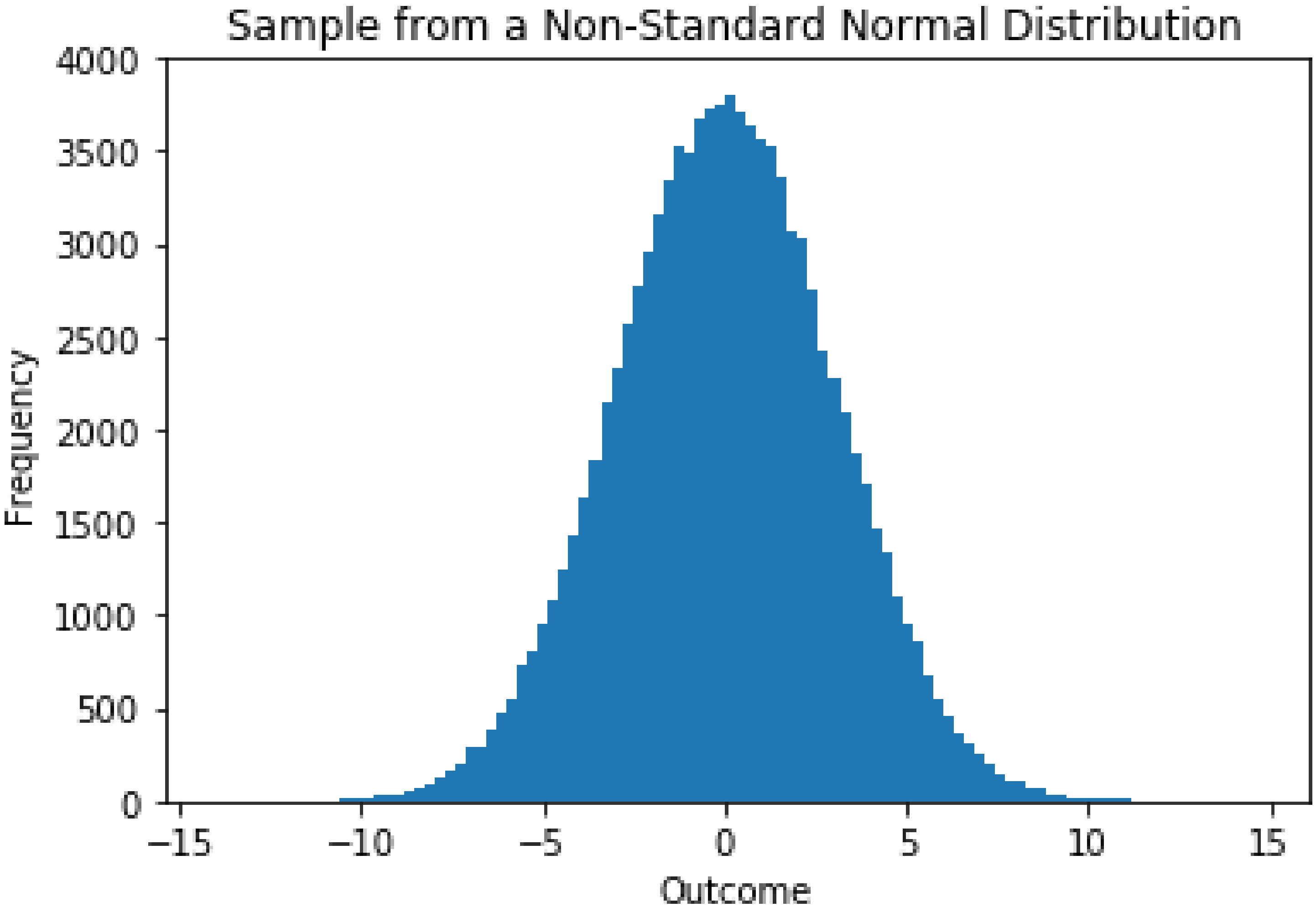
- Example:

```
plt.savefig(”figure1.pdf") (Exports to a PDF)
```

Histograms

- Example:

```
import matplotlib.pyplot as plt
import numpy as np
mean = 0
stddev = 3
x = mean + stddev * np.random.randn(100000)
n, bins, patches = plt.hist(x, 100, normed=1)
plt.title("Non-Standard Normal Distribution")
plt.xlabel("Outcome")
plt.ylabel("Frequency")
plt.show()
```


Sample from a Non-Standard Normal Distribution
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
−15
−10
−5
0
5
10
15
Outcome

Non-Linear Axes

- Non-linear axes are used for many purposes:
  - Data with a very large range of dependent variables
    - plt.yscale(“log”)
  - Historical financial data transformed so that a linear distance along the y axis always represents an equal percentage return
    - plt.yscale(“log”)
  - Representing log-odds of an occurrence
    - plt.yscale(“logit”)
  - Other scales are available for both axis

Text Formatting

- TeX
  - A typesetting system with the goal of appearing identically across all systems
  - Commonly used to present mathematical formulae
  - Uses $...$ syntax

Text Formatting (cont’d)

```
import matplotlib.patches as mpatches
import matplotlib.pyplot as plt
plt.plot([1, 2, 3, 4, 5], [37, 27, 15, 1, -15])
plt.axis([0.9, 5.1, -20, 40])
plt.title("Figure X")
plt.xlabel("X Label")
plt.ylabel("Y Label")
plt.grid(True)
blue_patch = mpatches.Patch(color="blue", label="$y = -x^2 - 7x + 45$")
plt.legend(handles=[blue_patch])
plt.savefig("pic1.pdf")
plt.show()
```

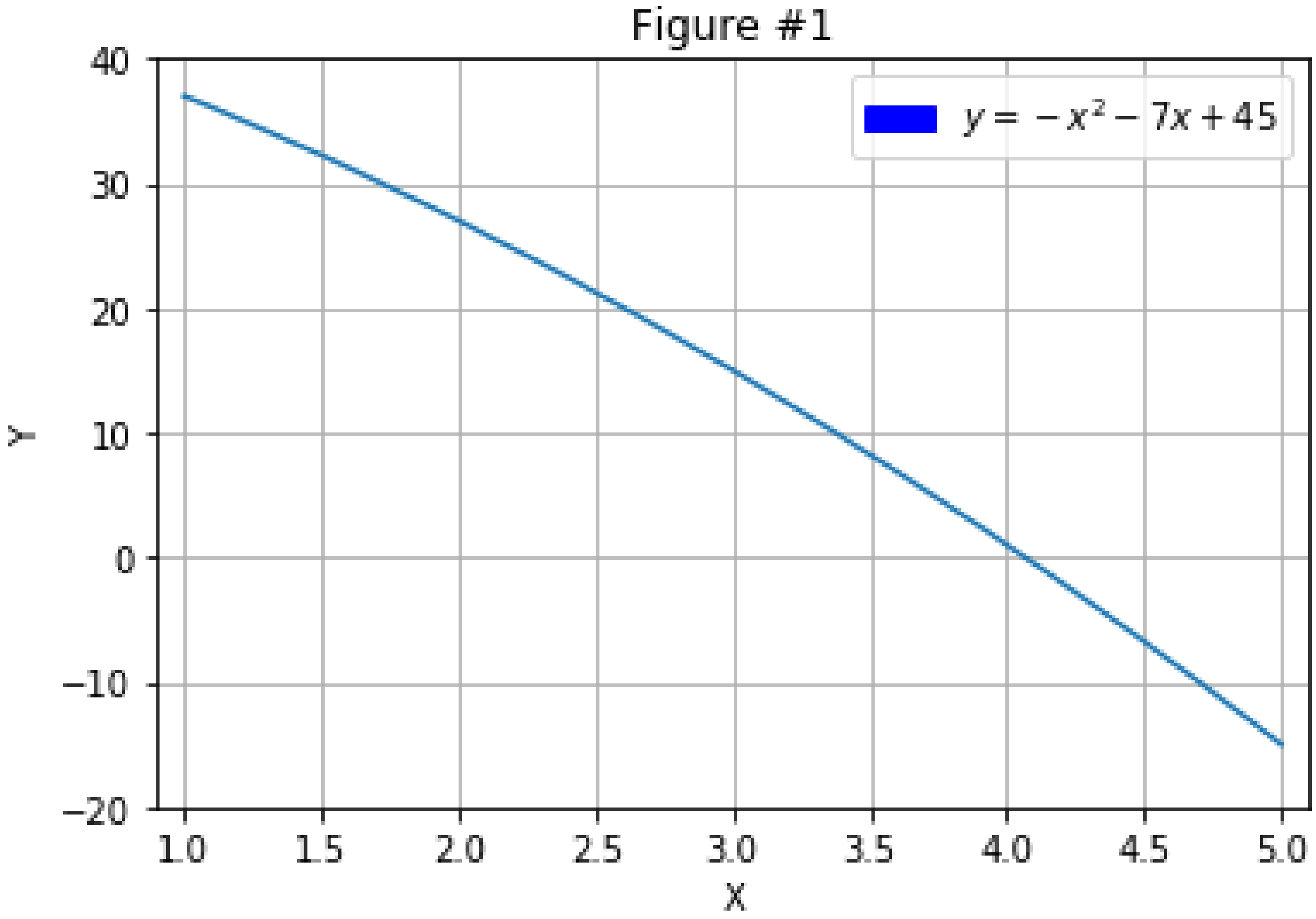


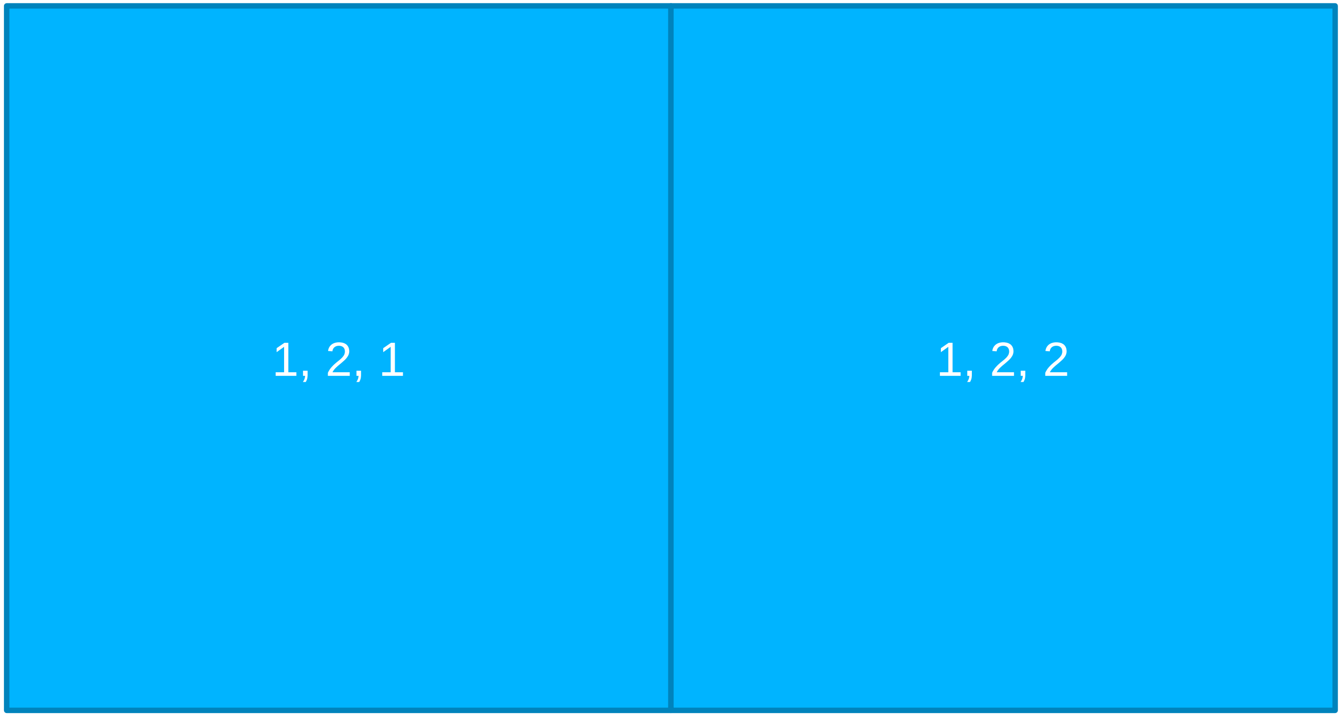
Figure #1
🟦 $y = -x^2 - 7x + 45$
Y
40
30
20
10
0
−10
−20
X
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Figures

- A figure is a container for multiple diagrams called ***subplots***

Subplots

- add_subplot(row, column, num)



Module 6 – Section 5

Resources and Homework

Resources

- Edward Tufte. Envisioning Information.
- Colin Ware. Information Visualization: Perception for Design.
- Steele and Iliinsky. Beautiful Visualization.
- Riccardo Mazza. Introduction to Information Visualization.
- Winston Chang. R Graphics Cookbook.
- Garr Reynolds. Presentation Zen.

Resources (cont'd)

- Scott Murray. Interactive Data Visualization.
- Kostiantyn Kucher. Python Data Visualization.
- PyLatex:
- colah.github.io/posts/2015-09-Visual-Information/
- michaelnielsen.org/reinventing_explanation/
- www.tableau.com/learn/whitepapers/tableau-visual-guidebook

Next Class

- Workshop

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

Any questions?

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies