Intro to Data Visualization

Data manipulation and plotting

Big data

Insights, predictions, products

https://sdtimes.com/data/big-data-go/

Data visualization

Modules

Pandas - reading data

Matplotlib - plotting data

What is Pandas ?

- *Pandas* is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

What does Pandas can do ?

Reading data

Viewing data

	gender	race/ethnicity	parental level of education	lunch	\
0	female	group C	some high school	standard	
1	female	group D	high school	standard	
2	female	group E	high school	standard	
3	male	group C	high school	standard	
4	male	group C	associate's degree	standard	
5	female	group D	associate's degree	standard	
6	male	group D	high school	standard	
7	female	group E	some high school	free/reduced	
8	female	group D	associate's degree	free/reduced	
9	female	group D	some college	standard	
10	female	group D	some high school	free/reduced	
11	female	group B	high school	free/reduced	
12	male	group C	some college	standard	
13	female	group B	some college	standard	
14	male	group C	high school	standard	
15	female	group B	some college	standard	
16	female	group E	some college	free/reduced	
17	male	group E	some college	standard	
18	female	group C	associate's degree	standard	
19	female	group E	high school	standard	
20	male	group B	some high school	free/reduced	
21	male	group E	master's degree	free/reduced	
22	male	group C	some high school	standard	
23	male	group C	some high school	standard	

```
{
  "about": "Sunlight Foundation uses cutting-edge technology and ideas to make
  "category": "Non-governmental organization (ngo)",
  "category_list": [
    {
      "id": "2235",
      "name": "Non-Governmental Organization (NGO)"
    }
  ],
  "checkins": 136,
  "description": "POSTING GUIDELINES:\n\nThe Sunlight Foundatio
ansparency and general issues related to open data. Individuals
```


json


```
id,name,released_on,price,created_at,updated_at
24,1000 Piece Jigsaw Puzzle,2012-07-03,14.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
30,360° Protractor,2012-05-03,3.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
17,7 Wonders,2012-04-21,28.75,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
13,Acoustic Guitar,2012-06-06,1025.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
15,Agricola,2012-05-22,45.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
22,Answer to Everything,2012-07-03,42.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
23,Box Kite,2012-05-19,63.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
29,CanCan Music Record,2012-05-09,2.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
12,Chocolate Pie,2012-04-12,3.14,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
9,Dog Toy Bone,2012-06-13,2.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
11,Flux Capacitor,2012-06-01,19.55,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
6,Game Console,2012-06-06,299.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
10,Heated Blanket,2012-07-19,27.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
19,Knights of Catan,2012-06-10,19.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
8,Lawn Chair,2012-05-29,34.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
21,Millennium Falcon,2012-04-10,3597200.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
14,Model Enterprise,2012-04-18,27.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
28,Model Train Rails,2012-06-30,45.0,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
3,Oak Coffee Table,2012-07-08,223.99,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
5,Oh's Cereal,2012-04-17,3.95,2012-07-09 16:50:49 UTC,2012-07-09 16:50:49 UTC
```


csv

https://sunlightfoundation.com/2014/03/11/making-json-as-simple-as-a-spreadsheet/
http://railscasts.com/episodes/362-exporting-csv-and-excel?view=asciicast

How to install Pandas

- Conda command: conda install -c anaconda pandas

- Pip command: pip install pandas

Some commands of Pandas

1. How to import Pandas.
   Command : import pandas as pd

2. Reading a csv file.
   Command : pd.read_csv(*filepath_or_buffer*)

   filepath_or_buffer (String) = Path of a file.

   Example: df = pd.read_csv("exams_100.csv")

Some commands of Pandas

3. Viewing the top n rows of a file

Command : df.head(n)

n (int) = number of rows to select. (Default is 5)

4. Viewing the bottom n rows of a file

Command : df.tail(n)

n (int) = number of rows to select. (Default is 5)

Example of reading data by pandas

```
Import pandas as pd

data = pd.read_csv('data.csv')

print(data)
```

	gender	group	level of education	test preparation course	math score	reading score	writing score
0	female	group C	some high school	none	67	65	69
1	female	group D	high school	completed	66	75	78
2	female	group E	high school	none	76	74	75
3	male	group C	high school	completed	70	76	67
4	male	group C	associate's degree	none	56	49	49

How to get data after reading it

We can look inside each column by its name

Example :

Returns : **DataFrame**

What is a DataFrame ? : a tabular data (with rows and columns).

Convert to list by mylist = list(gender)

What is Matplotlib

- Matplotlib is a Python 2D plotting library

What does Matplotlib can do ?

- Advantages of Matplotlib library
  - Large community
  - Many plot types supported
  - Easy to use with python

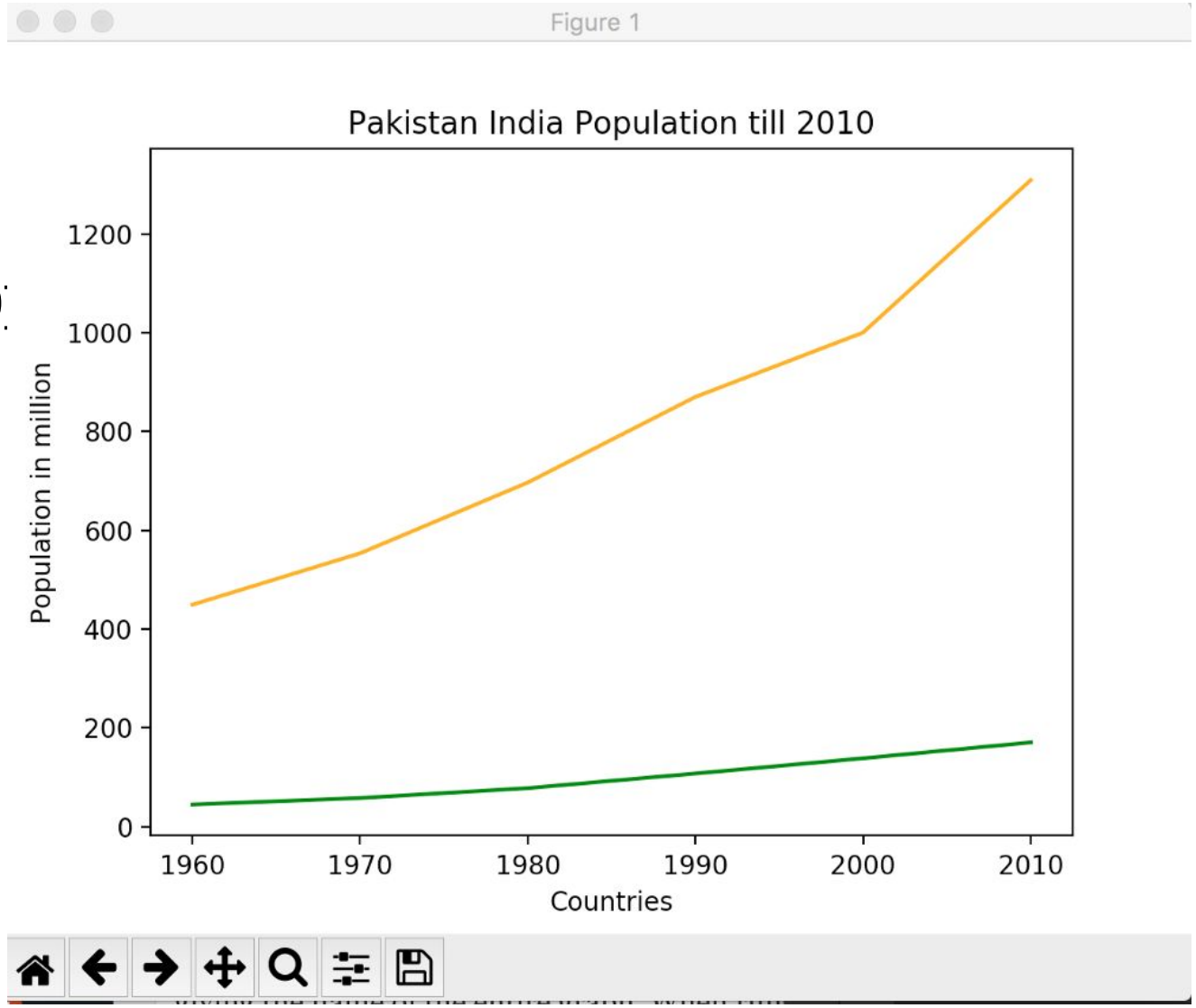
What does Matplotlib can do ?

- Some types of Matplotlib graphs
  - Line graph
  - Bar graph
  - Pie graph
  - etc

Line graph

```
import matplotlib.pyplot as plt

year = [1960, 1970, 1980, 1990, 2000, 2010]
pop_pakistan = [44.91, 58.09, 78.07, 107.7,
138.5, 170.6]
pop_india =  [449.48, 553.57, 696.783,
870.133, 1000.4, 1309.1]
plt.plot(year, pop_pakistan, color='g')
plt.plot(year, pop_india, color='orange')
plt.xlabel('Countries')
plt.ylabel('Population in million')
plt.title('Pakistan India Population till 2010')
plt.show()
```

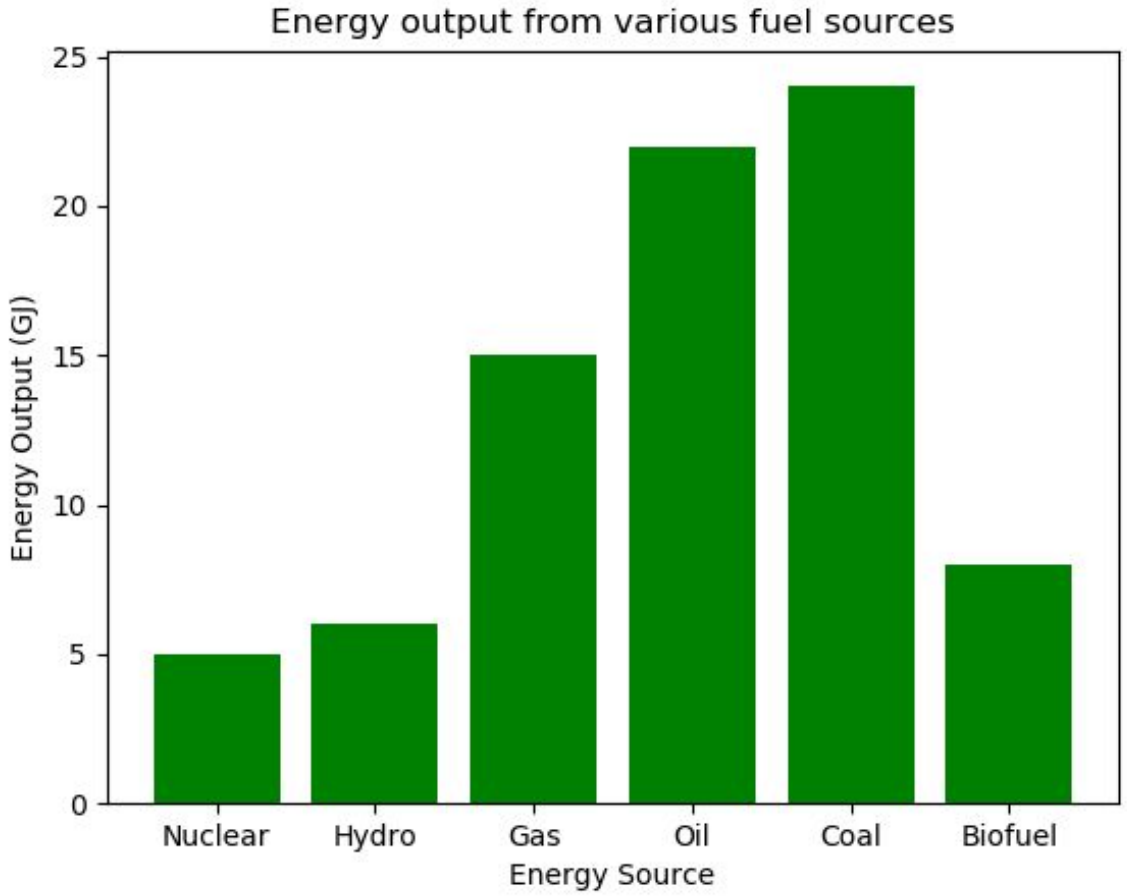


list comprehension!
What does enumerate(x) do?

Bar graph

```
import matplotlib.pyplot as plt

x = ['Nuclear', 'Hydro', 'Gas', 'Oil', 'Coal',
'Biofuel']
energy = [5, 6, 15, 22, 24, 8]
x_pos = [idx for idx, val in enumerate(x)]
plt.bar(x_pos, energy, color='green')
plt.xlabel("Energy Source")
plt.ylabel("Energy Output (GJ)")
plt.title("Energy output from various fuel
sources")
plt.xticks(x_pos, x)
plt.show()
```



mylist = [‘a’,’b’,’c’]

for item in mylist

for id, item in enumerate(mylist)

Bar graph with multiple plots

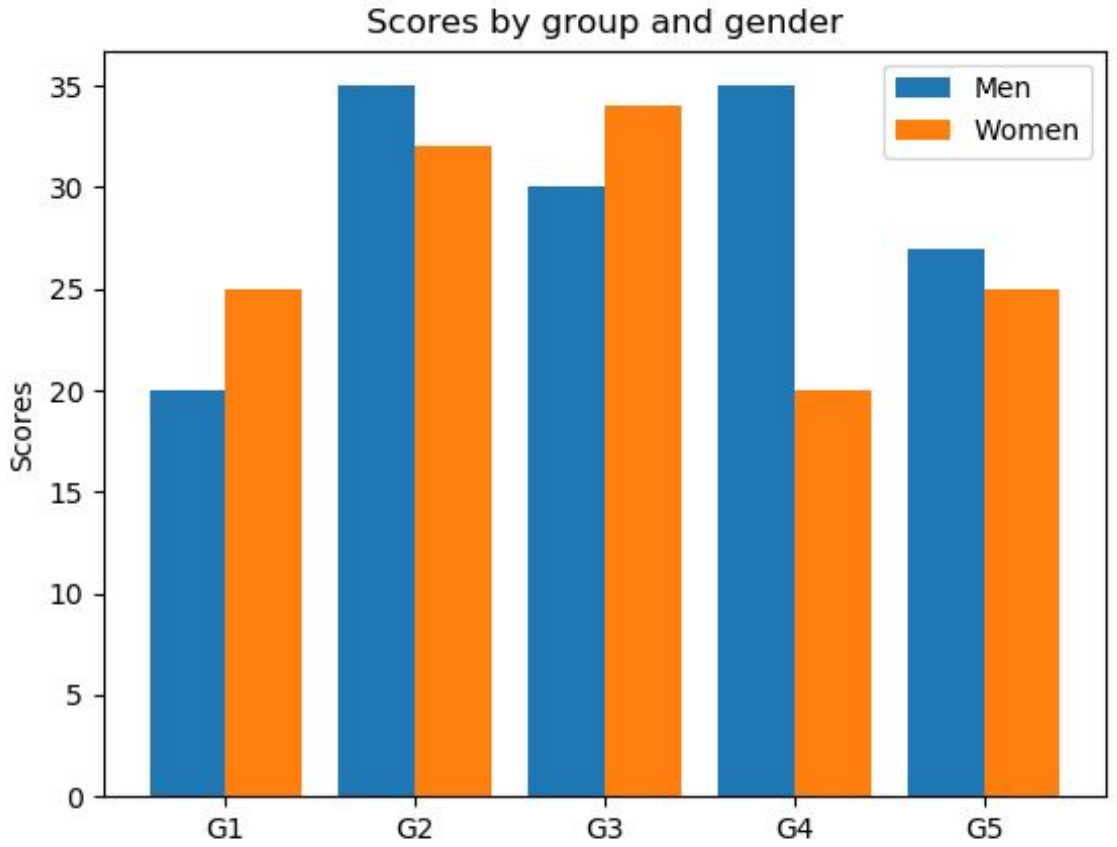
```
N = 5
men_means = (20, 35, 30, 35, 27)
women_means = (25, 32, 34, 20, 25)
width = 0.40

index_men = [idx for idx in range(N)]
index_women = [idx+0.40 for idx in range(N)]
index_number = [idx+0.40 / 2 for idx in range(N)]

plt.bar(index_men, men_means, width, label='Men')
plt.bar(index_women, women_means, width, label='Women')

plt.ylabel('Scores')
plt.title('Scores by group and gender')

plt.xticks(index_number, ('G1', 'G2', 'G3', 'G4', 'G5'))
plt.legend(loc='best')
plt.show()
```



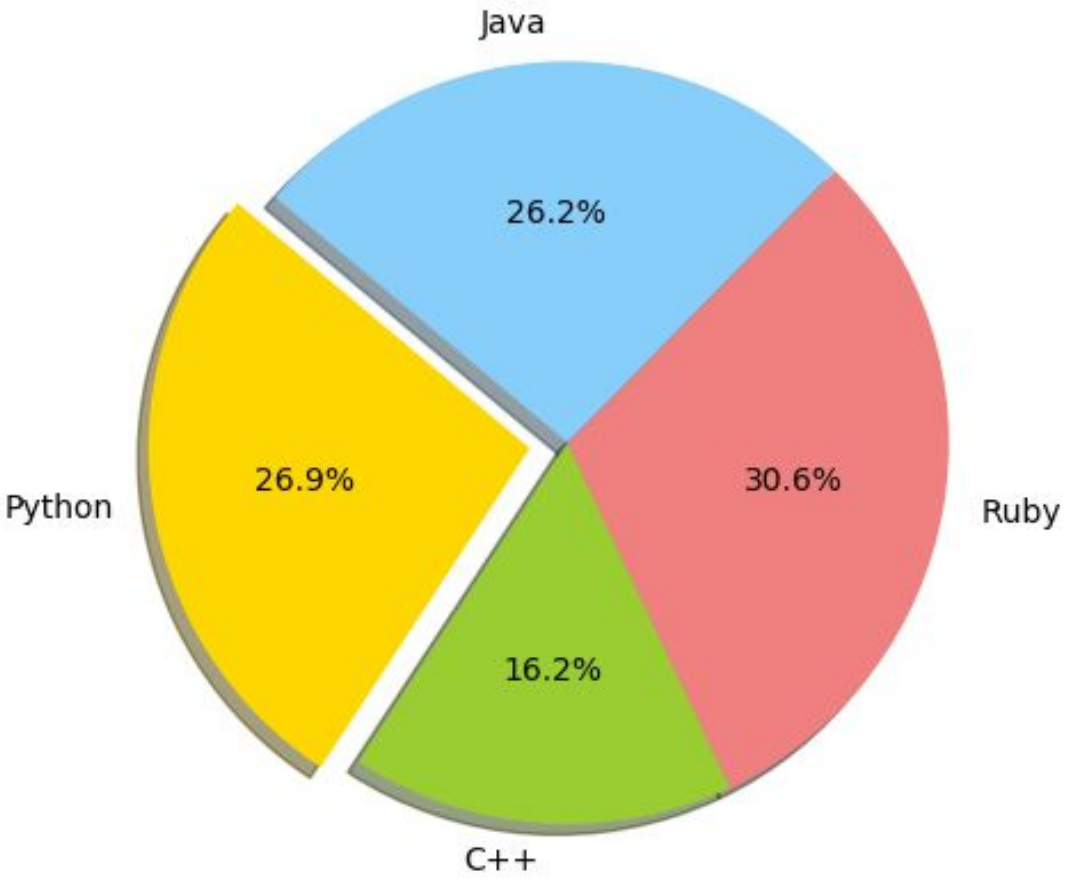
Pie graph

```
import matplotlib.pyplot as plt

# Data to plot
labels = 'Python', 'C++', 'Ruby', 'Java'
sizes = [215, 130, 245, 210]
colors = ['gold', 'yellowgreen', 'lightcoral',
'lightskyblue']
explode = (0.1, 0, 0, 0)  # explode 1st slice

# Plot
plt.pie(sizes, explode=explode,
labels=labels, colors=colors,
autopct='%1.1f%%', shadow=True,
startangle=140)

plt.show()
```



Pie graph calculation

First, put your data into a table (like above), then add up all the values to get a total:

Table: Favorite Type of Movie					
Comedy	**Action**	**Romance**	**Drama**	**SciFi**	**TOTAL**
4	5	6	1	4	**20**

Next, divide each value by the total and multiply by 100 to get a percent:

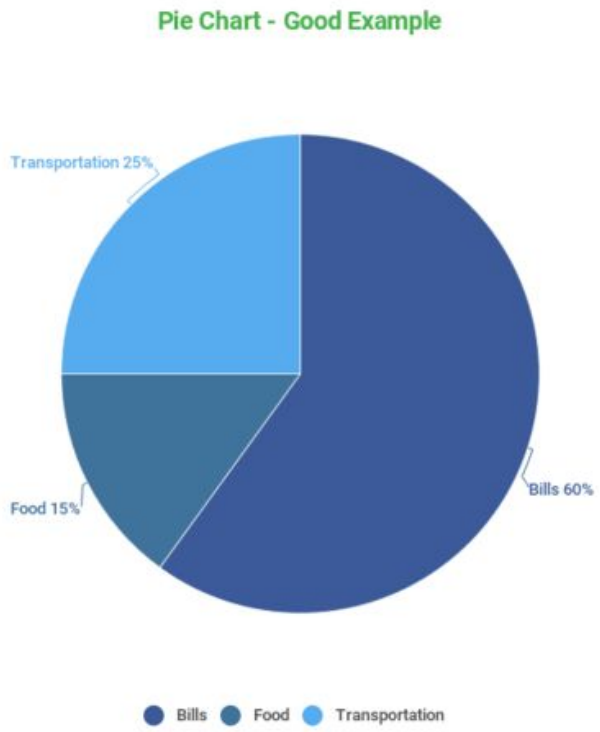
Comedy	**Action**	**Romance**	**Drama**	**SciFi**	**TOTAL**
4	5	6	1	4	**20**
4/20 = **20%**	5/20 = **25%**	6/20 = **30%**	1/20 = **5%**	4/20 = **20%**	**100%**

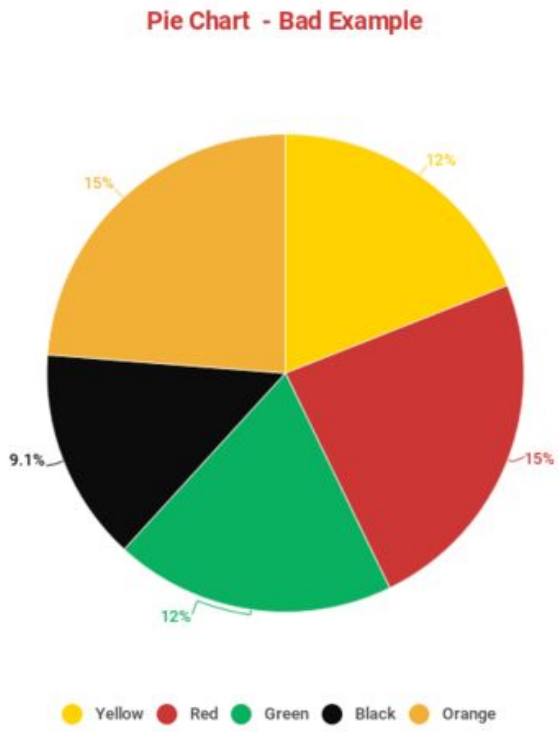
Credit: https://www.mathsisfun.com/data/pie-charts.html

Which plot type?

Bar graph: Comparison

Line graph: Trend, time series





Pie graph: proportion, clear difference, not more than 5 segments