Trực quan hóa dữ liệu với Matplotlib

Giảng viên: TS. **Nguyễn Văn Quyết**

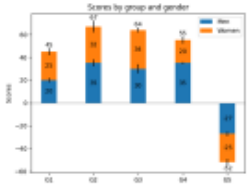
Nội dung

- Giới thiệu về Matplotlib
- Vẽ biểu đồ Line
- Vẽ biểu đồ Bar
- Vẽ biểu đồ Histogram và Density
- Vẽ biểu đồ Scatter

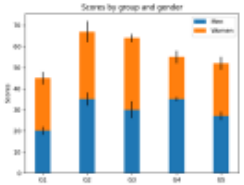
1. Giới thiệu về Matplotlib

- Matplotlib (Mathematical plotting library) là một thư viện đa nền tảng cung cấp các công cụ nhằm trực quan hóa dữ liệu.

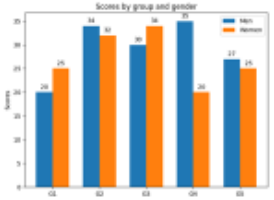
Lines, bars and markers



Bar Label Demo



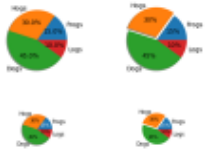
Stacked bar chart



Grouped bar chart with labels

Pie and polar charts

Basic pie chart



Pie Demo2

Bar of pie

Khai báo thư viện matplotlib

- Matplotlib có một quy tắc khai báo thư viện như sau:

```
import matplotlib as mpl
import matplotlib.pyplot as pl
```

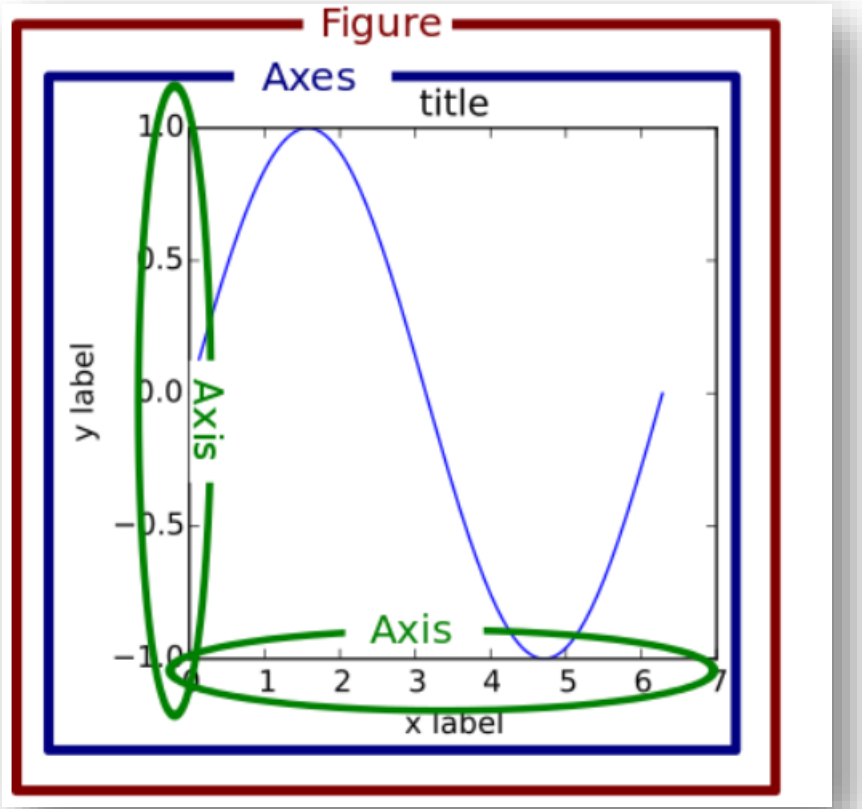
Một số khái niệm chung

- **Figure**: Một cửa sổ chứa tất cả những gì bạn sẽ vẽ trên đó.
- **Axes**: Thành phần chính của một figure là các axes (những khung nhỏ hơn để vẽ hình lên đó). Một figure có thể chứa một hoặc nhiều axes.
- **Axis**: các trục của hình vẽ là các đối tượng đảm nhiệm việc tạo các giới hạn biểu đồ.

Figure (Hình vẽ)

- Figure là lớp nằm trong module matplotlib figure. Nó là nơi chứa đựng tất cả các thành phần khác.
- Đối tượng figure được khởi tạo bằng hàm figure() trong pyplot module.

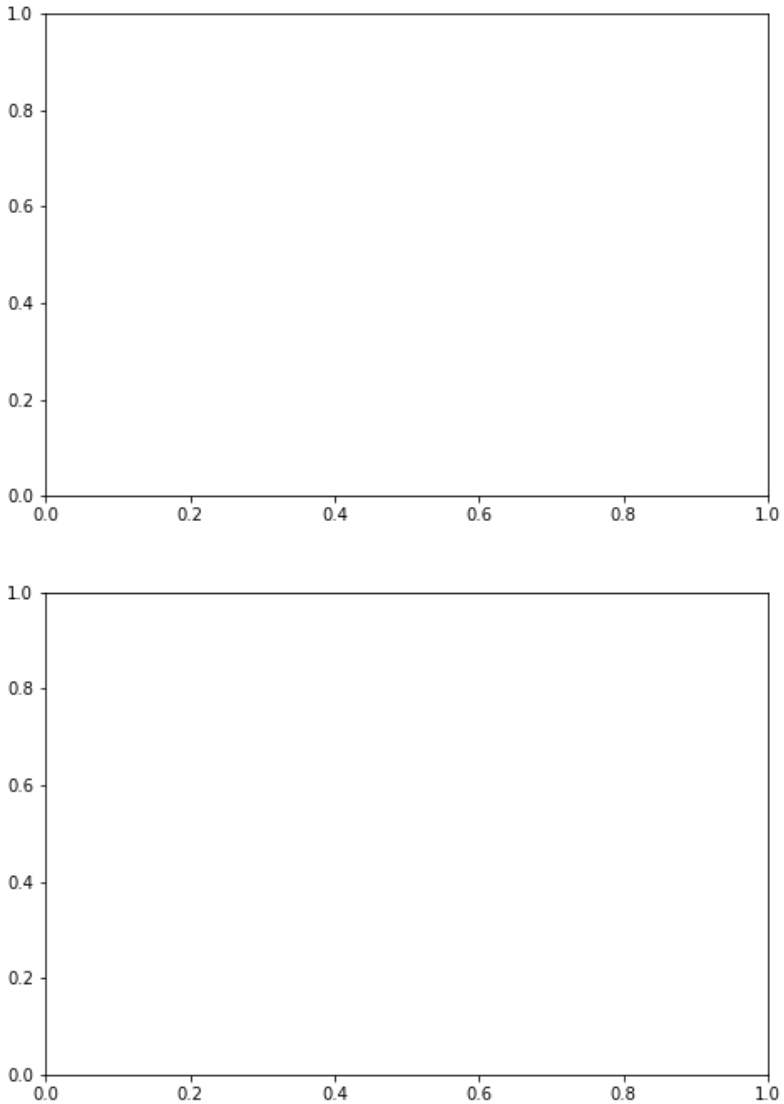
```
In [7]: fig=plt.figure()
```



Axes

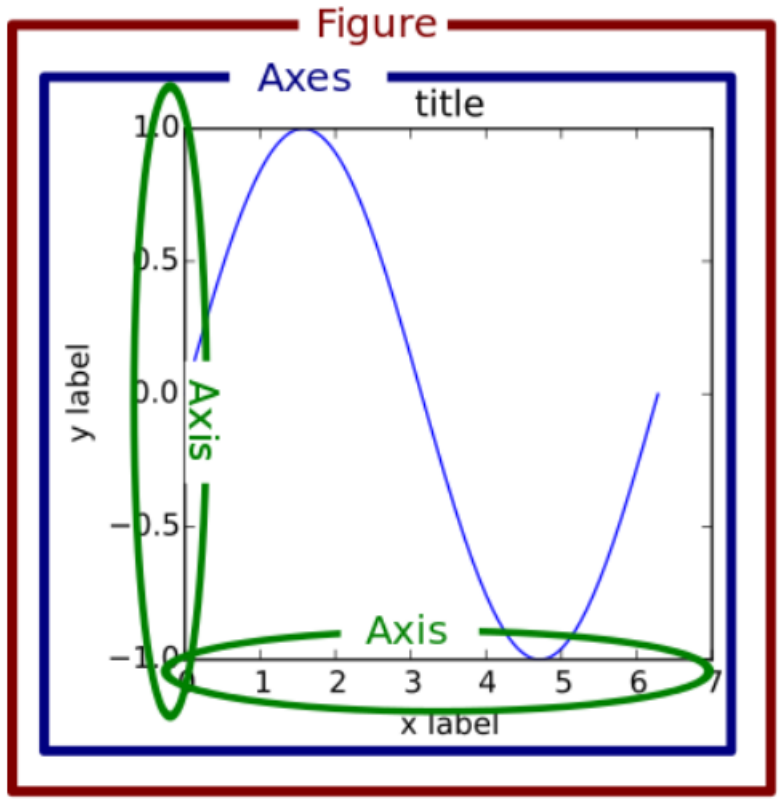
- Axes là một khu vực hình ảnh nằm trong figure. Một đối tượng figure có thể chứa nhiều đối tượng axes nhưng một đối tượng axes chỉ có thể thuộc về một đối tượng figure.
- Đối tượng Axes được thêm vào figure bằng hàm add_axes() nhận vào là một list gồm 4 giá trị **[left, bottom, width, height]** của axes định tạo.

```
fig = pl.figure()
fig.add_axes([0,0,1,1])
fig.add_axes([0,1.2,1,1])
```



Axis (Trục tọa độ)

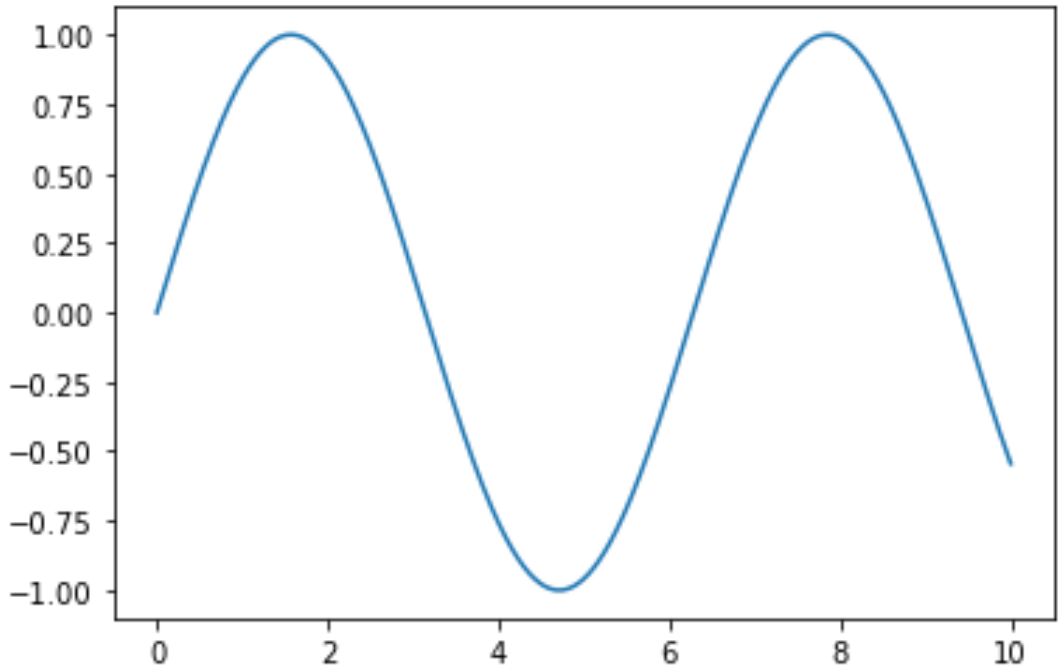
- Là thành phần nằm trong Axes có khả năng nhận title và ticks label



Biểu đồ Line

- Biểu đồ Line hay biểu đồ đường (thẳng hoặc cong) là loại biểu đồ được vẽ bằng cách nối các điểm trên hình vẽ với nhau.
- Để vẽ biểu đồ line, ta dùng hàm **plot()** với hai tham số nhận vào:
  - x: là một dãy các hoành độ của các điểm.
  - y: là một dãy các tung độ của các điểm.

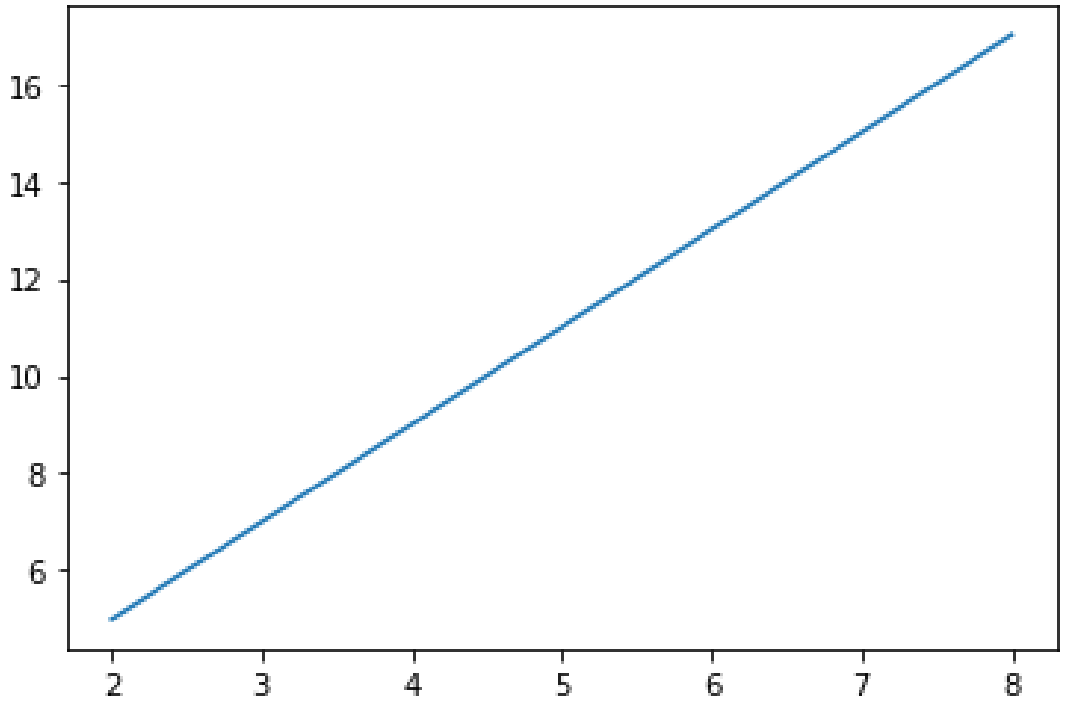
```
x = np.linspace(0,10,1000)
plt.plot(x,np.sin(x))
```



Biểu đồ Line

- Vẽ đường thẳng y=2x+1

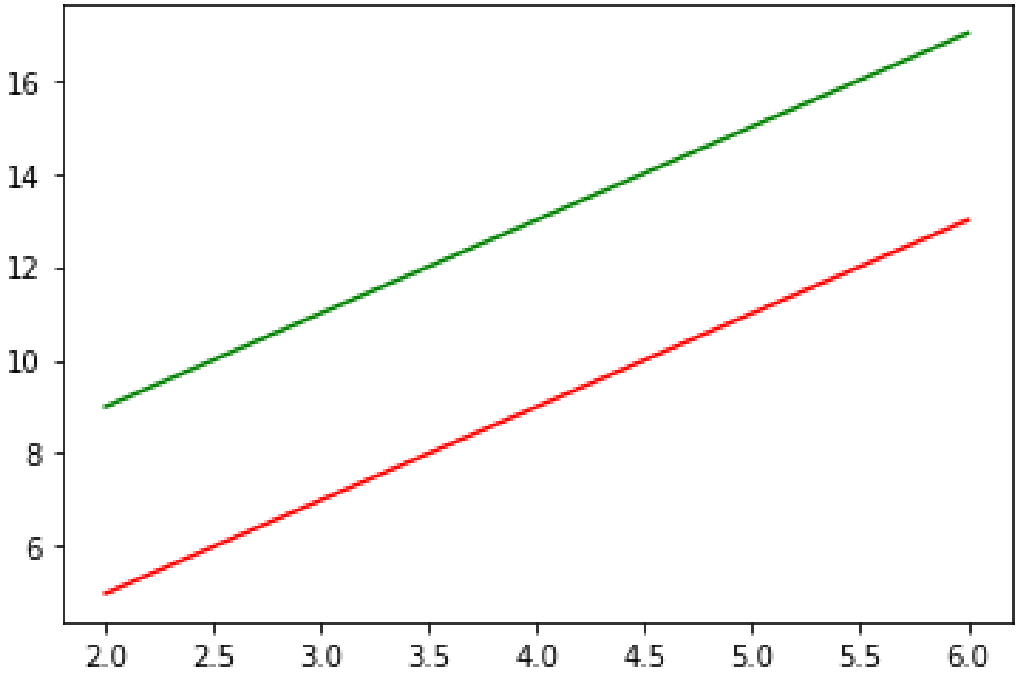
```
x = np.random.randint(-10,10,2)
y = 2*x + 1
plt.plot(x,y)
```



Biểu đồ Line

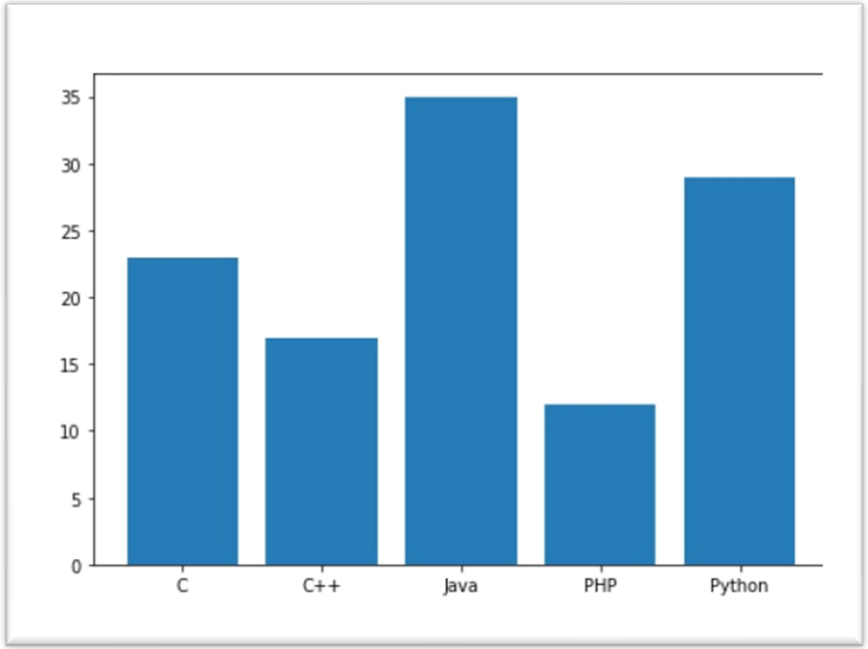
- Tùy biến màu sắc của line sử dụng thuộc tính color

```
x = np.random.randint(0,10,2)
plt.plot(x,2*x+1,color='red')
plt.plot(x,2*x+5,color='green')
```



3. Vẽ biểu đồ Bar

- Biểu đồ bar là biểu đồ thường được dùng để so sánh được vẽ bằng hàm bar với hai tham số truyền vào là x, height trong đó:
  - x: là dãy truyền vào các tọa độ của các thanh bar.
  - height: chiều cao tương ứng của các thanh bar.

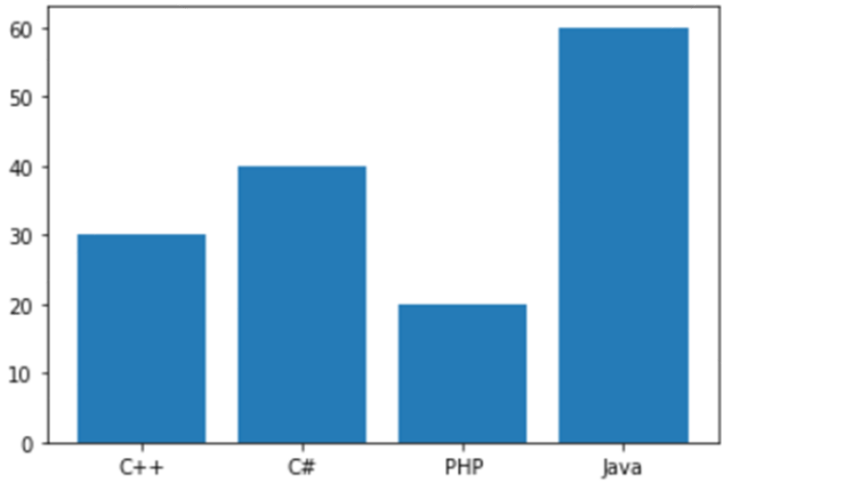


Ví dụ 1: Vẽ biểu đồ bar

- Vẽ biểu đồ bar có các giá trị như sau:

	A	B	C
1	Language	Percent	
2	C++	30	
3	C#	40	
4	PHP	20	
5	Java	60	
6			

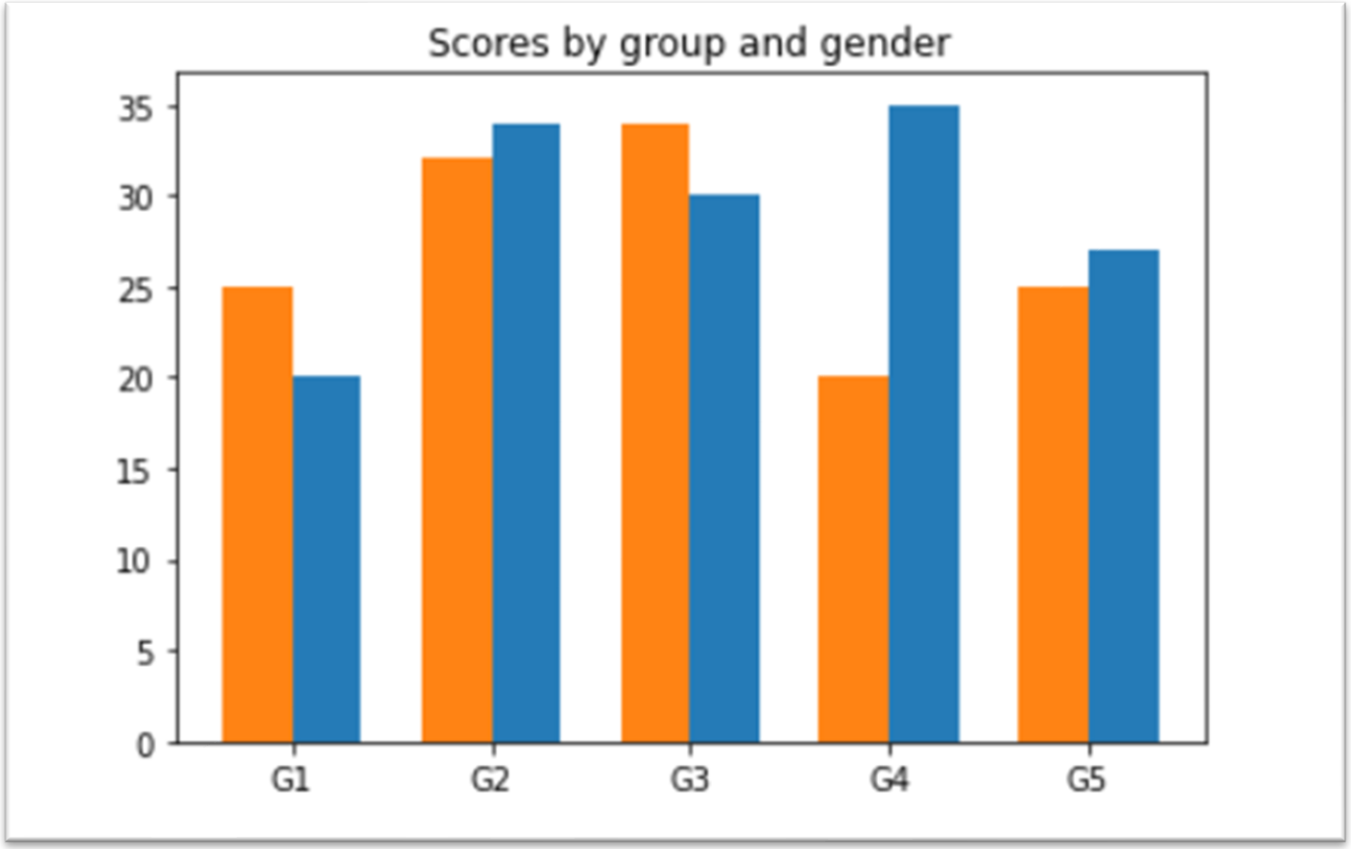
```
In [14]: plt.bar(["C++","C#","PHP","Java"],[30,40,20,60])
Out[14]: <BarContainer object of 4 artists>
```



Ví dụ 2: Vẽ biểu đồ nhóm bar

- Vẽ biểu đồ bar có các giá trị cho trong bảng sau:

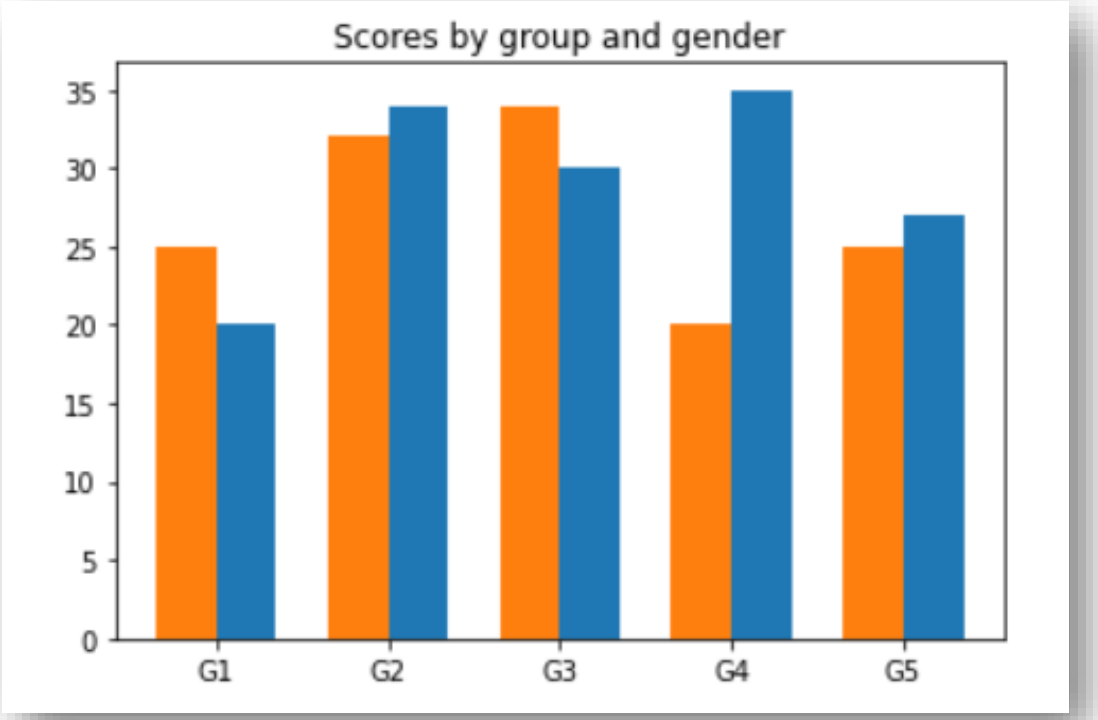
Scores by group and gender		
Group	Men	Women
G1	20	25
G2	34	32
G3	30	34
G4	35	20
G5	27	25



Ví dụ 2: Vẽ biểu đồ nhóm bar

- Ta sẽ dùng pyplot để vẽ biểu đồ cho bài toán này như sau:

```
In [25]: labels=['G1','G2','G3','G4','G5']
         men=[20,34,30,35,27]
         women=[25,32,34,20,25]
         width=0.35 #width of a bar
         X=np.array([1,2,3,4,5])
         plt.bar(X+width/2,men,width,label="Men")
         plt.bar(X-width/2,women,width, label="Women")
         plt.title("Scores by group and gender")
         plt.xticks(X,labels)
         pass
```



Bài tập: Vẽ biểu đồ nhóm bar

- Cho danh sách sinh viên các lớp lưu trong tệp CSV: **TK16-Students.csv**

STT	Mã SV	Họ và tên	Giới tính	Ngày sinh	Quê quán	Mã lớp
1	10118350	Nguyễn Đức Tuấn Anh	Nam	01-03-2000	Khoái Châu - Hưng Yên	101181
2	10118002	Nguyễn Tuấn Anh	Nam	07-08-2000	Tân Dân-Khoái Châu-Hưng Yên	101181
3	10118332	Nguyễn Tuấn Anh	nam	12-12-2000	nhân hòa mỹ hào hưng yên	101181
4	10118354	Nguyễn Văn Chiến	Nam	16-04-2000	Việt Cường - Yên Mỹ - Hưng Yên	101181
5	10118344	Nguyễn Hữu Chung	Nam	10-09-2000	Ứng Hòa-Hà Nội	101182
6	10118358	Nguyễn Trọng Dũng	Nam	23-01-2000	Ứng Hòa-Hà Nội	101182
7	10118359	Vũ Chung Dũng	Nam	20-02-2000	Hoàn Long - Yên Mỹ - Hưng Yên	101182
8	10118361	Ngô Thị Dương	Nữ	13-10-2000	Tiên Lữ - Hưng Yên	101183
9	10118365	Bùi Thành Đạt	Nam	21-03-2000	Yên Thủy-Hòa Bình	101183

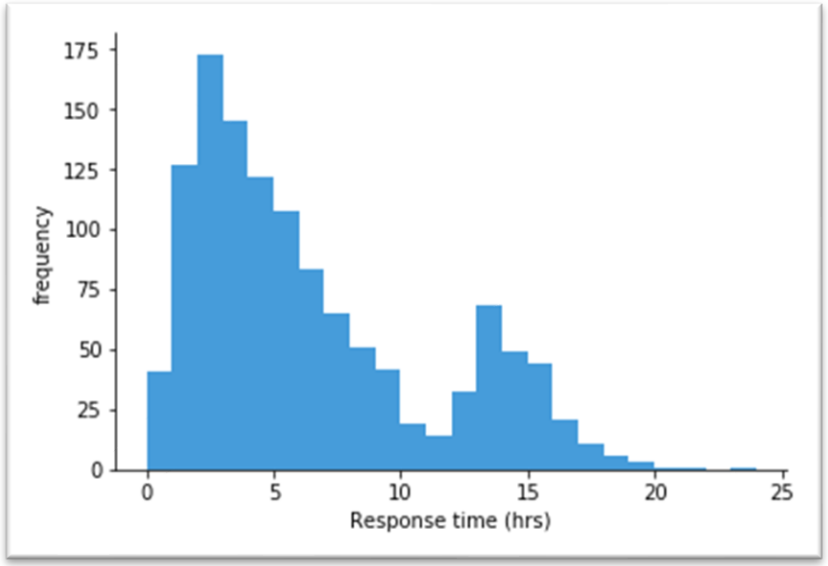
- **Yêu cầu:**

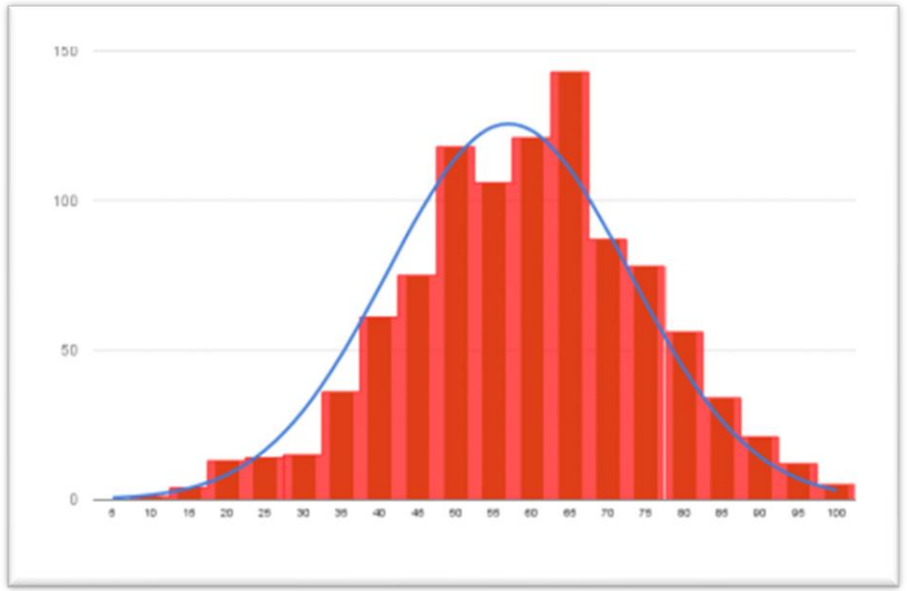
1) Vẽ biểu đồ Bar thống kê sĩ số của từng lớp (8 lớp, 223 sinh viên)

2) Vẽ biểu đồ Bar thống kê sĩ số theo nhóm nam, nữ của từng lớp

Vẽ biểu đồ Histogram và Density

- Histogram là biểu đồ dùng để biểu diễn sự phân phối của dữ liệu số.





Ví dụ: Vẽ biểu đồ histogram với bảng dữ liệu như sau:

- Dùng biểu đồ histogram để biểu diễn số phim ra mắt hàng năm của Netflix: Tải file tại đây.

```
In [6]: data=pd.read_csv("netflix_titles.csv")
        data
```

Out[6]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

Bước 1: Lấy dữ liệu bằng pandas

- Ta lấy dữ liệu bằng pandas lấy về trường “release_year”.

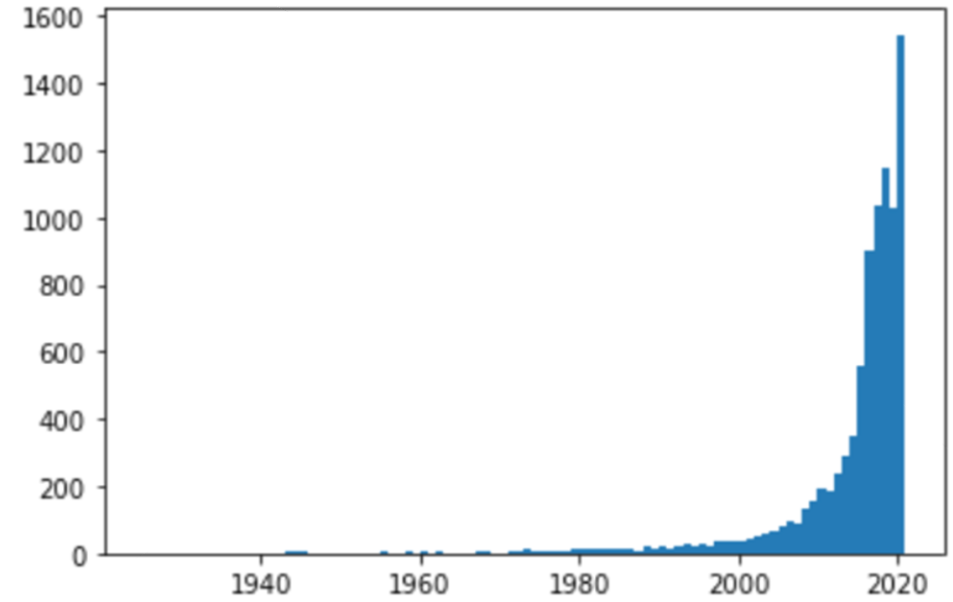
```
In [8]: data=pd.read_csv("netflix_titles.csv")
        data=data["release_year"].values
        data

Out[8]: array([2020, 2021, 2021, ..., 2009, 2006, 2015], dtype=int64)
```

Bước 2: Vẽ biểu đồ

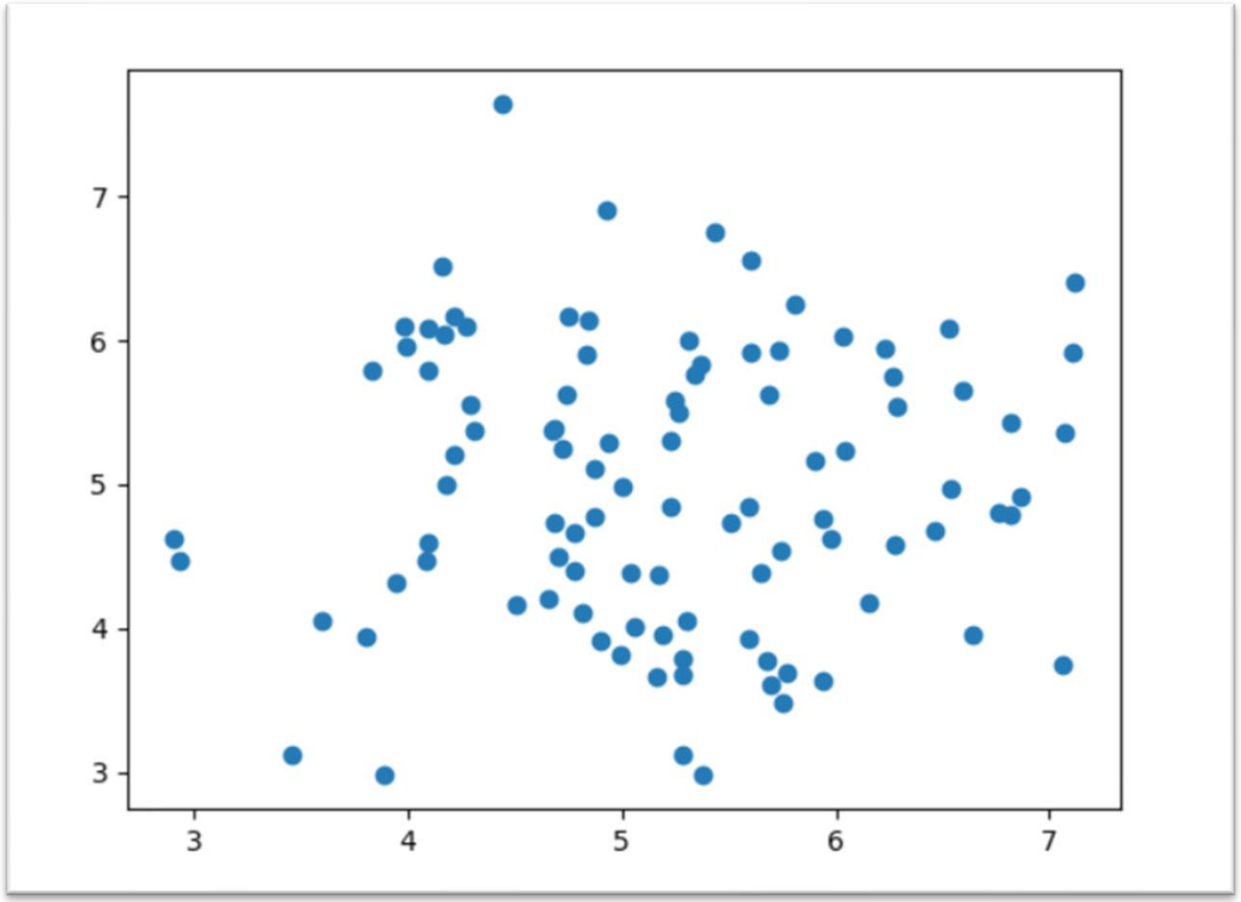
- Ta vẽ biểu đồ như sau:

```
plt.hist(data, bins=np.arange(min(data),max(data)+1,1))
pass
```



5. Vẽ biểu đồ Scatter

- Biểu đồ Scatter là biểu diễn các giá trị trong một tập dữ liệu bằng một điểm.



5. Vẽ biểu đồ Scatter

- Ví dụ: Biểu diễn dữ liệu tương quan giữa tuổi và tốc độ của xe hơi
  - Ta lấy 2 mảng ngẫu nhiên trong đó x biểu diễn tuổi, y biểu diễn tốc độ trung bình của xe.

```
In [8]: #day one, the age and speed of 13 cars:
        x = np.array([5,7,8,7,2,17,2,9,4,11,12,9,6])
        y = np.array([99,86,87,88,111,86,103,87,94,78,77,85,86])
        plt.scatter(x, y)

        #day two, the age and speed of 15 cars:
        x = np.array([2,2,8,1,15,8,12,9,7,3,11,4,7,14,12])
        y = np.array([100,105,84,105,90,99,90,95,94,100,79,112,91,80,85])
        plt.scatter(x, y)
        pass
```