



VINBIGDATA



Bài 8: Trực quan hóa dữ liệu với Matplotlib

AI Academy Vietnam

Nội dung bài 8

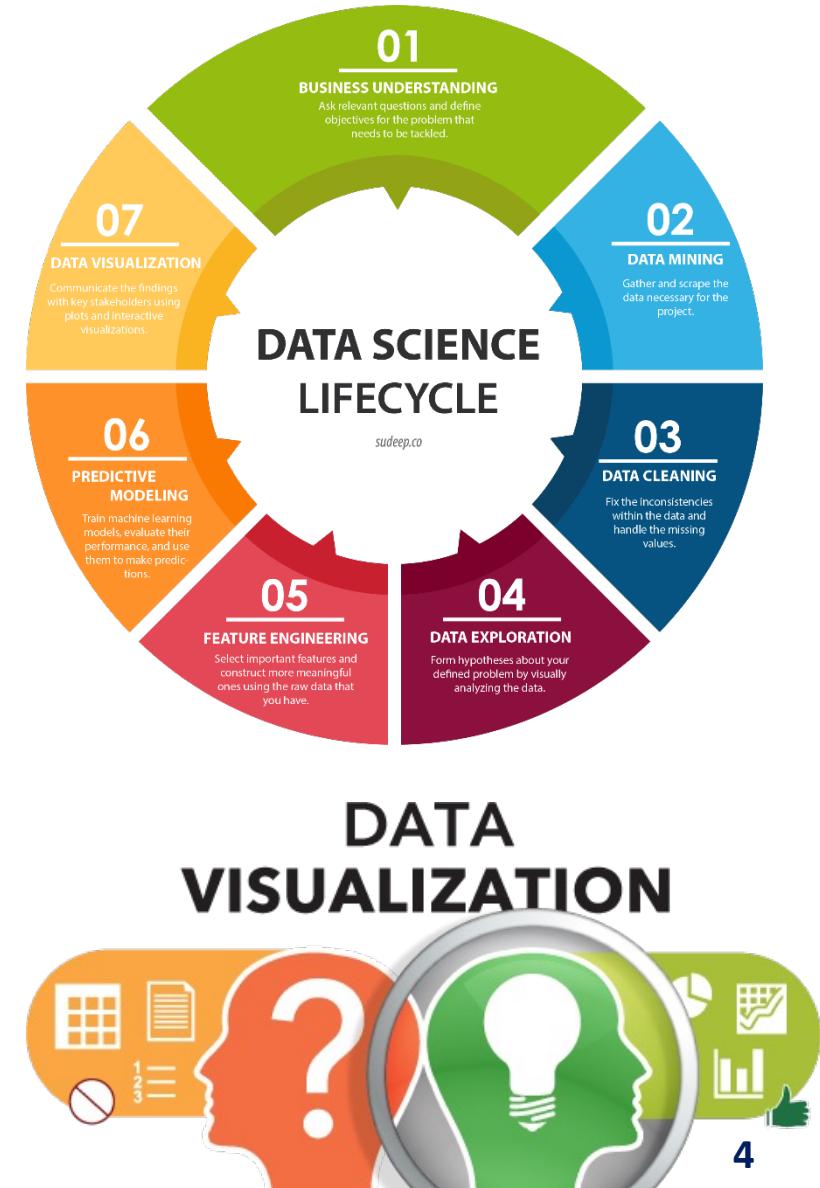
1. Tâm quan trọng của trực quan hóa dữ liệu
2. Một số lưu ý khi trực quan hóa dữ liệu
3. Một số thư viện trực quan hóa dữ liệu với Python
4. Biểu đồ Line chart
5. Biểu đồ Bar chart
6. Biểu đồ Pie chart
7. Biểu đồ Scatter plot
8. Biểu đồ Histogram plot
9. Biểu đồ Boxplot



1. Tầm quan trọng của trực quan hóa dữ liệu

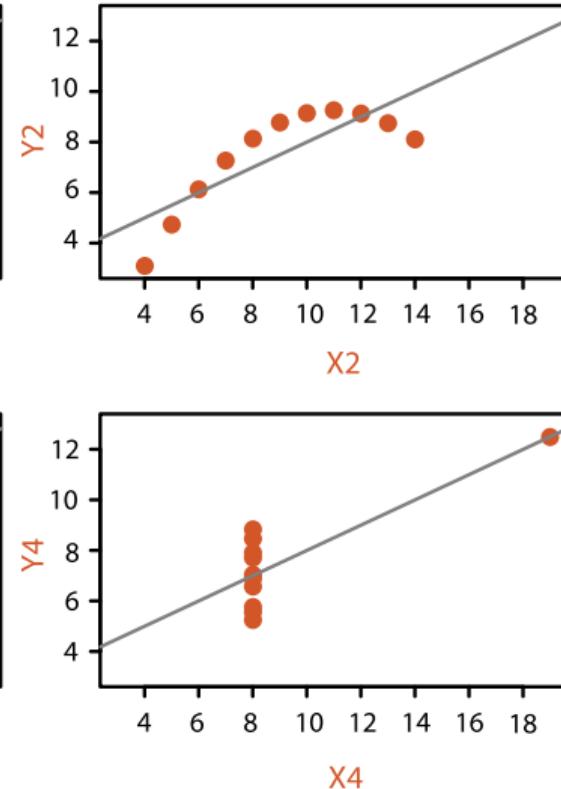
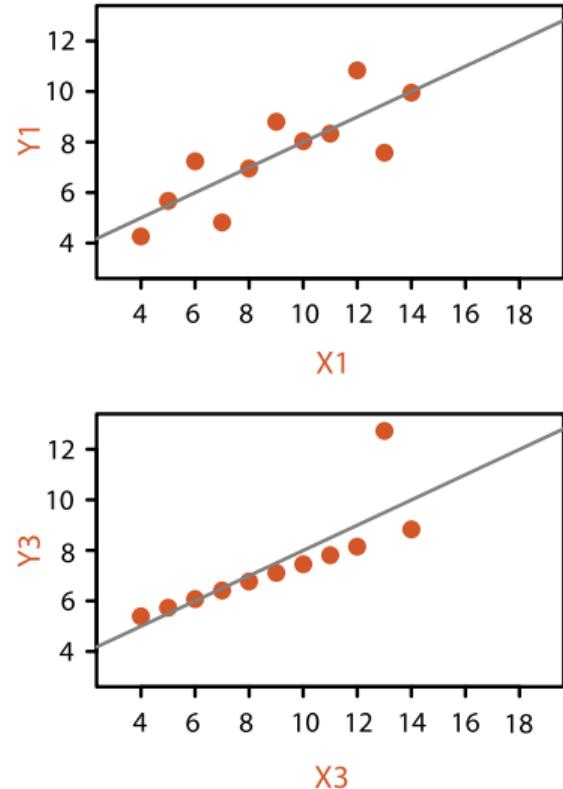
Trực quan hóa dữ liệu là gì?

- Trực quan hóa dữ liệu là việc biểu diễn đồ họa các thông tin trừu tượng nhằm 2 mục đích: Phân tích dữ liệu và truyền thông.
- Trực quan hóa dữ liệu là một công cụ mạnh mẽ để khám phá và trích rút các thông tin có giá trị (insight) từ tập dữ liệu.
- Bản chất của Trực quan hóa dữ liệu là sự trình bày dữ liệu theo định dạng hình ảnh hoặc đồ họa, từ đó truyền đạt thông tin rõ ràng và hiệu quả cho người dùng.
- Là yếu tố giao tiếp bằng hình ảnh của phân tích dữ liệu, giúp chuyển đổi dữ liệu thành thông tin và thông tin thành thông tin hữu ích.



Tầm quan trọng

| | 1 | | 2 | | 3 | | 4 | |
|-------------|-------|-------|-------|------|-------|-------|-------|-------|
| | X | Y | X | Y | X | Y | X | Y |
| | 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| | 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| | 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| | 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| | 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| | 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| | 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| | 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| | 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| | 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| | 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |
| Mean | 9.0 | 7.5 | 9.0 | 7.5 | 9.0 | 7.5 | 9.0 | 7.5 |
| Variance | 10.0 | 3.75 | 10.0 | 3.75 | 10.0 | 3.75 | 10.0 | 3.75 |
| Correlation | 0.816 | | 0.816 | | 0.816 | | 0.816 | |

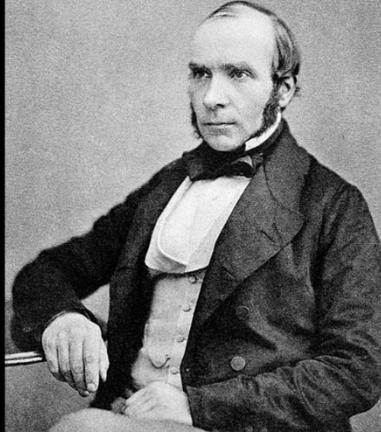


Tầm quan trọng



VINBIGDATA
VINGROUP

AS Academy
Vietnam



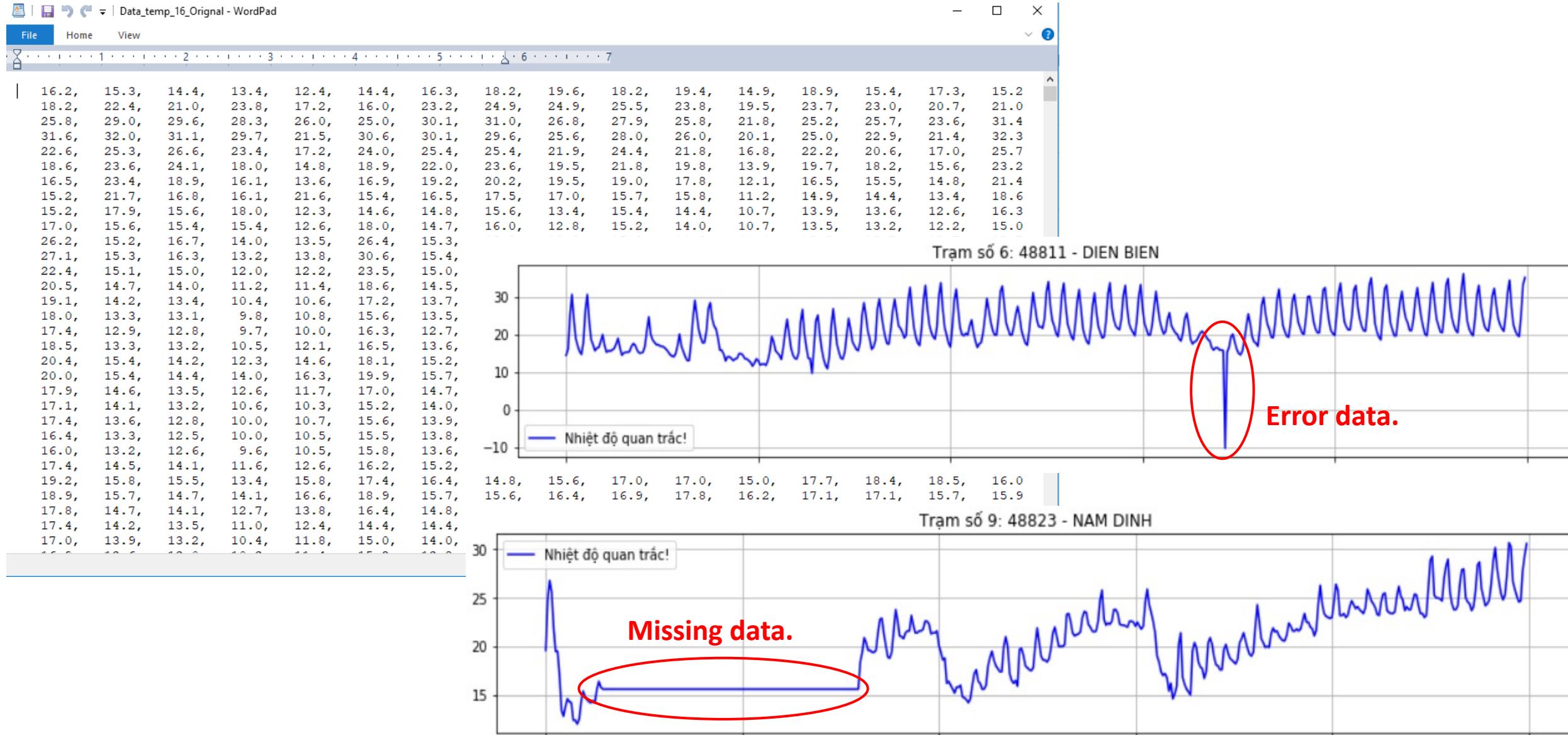
Dịch tả (London 1854), John Snow
600 người chết chỉ trong vài tuần...

Tầm quan trọng



VINBIGDATA VINGROUP

Academy Vietnam

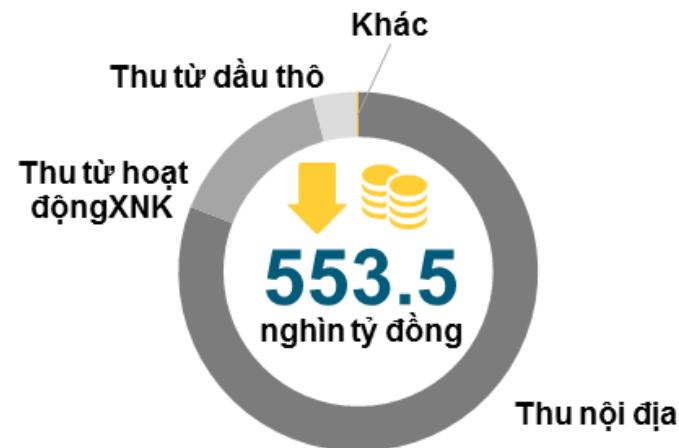




THU CHI NGÂN SÁCH NHÀ NƯỚC 5T 2019

(Từ đầu năm đến thời điểm 15/5/2019)

TỔNG THU NGÂN SÁCH NHÀ NƯỚC



Tổng thu ngân sách Nhà nước bằng
39.2% so với dự toán năm 2019

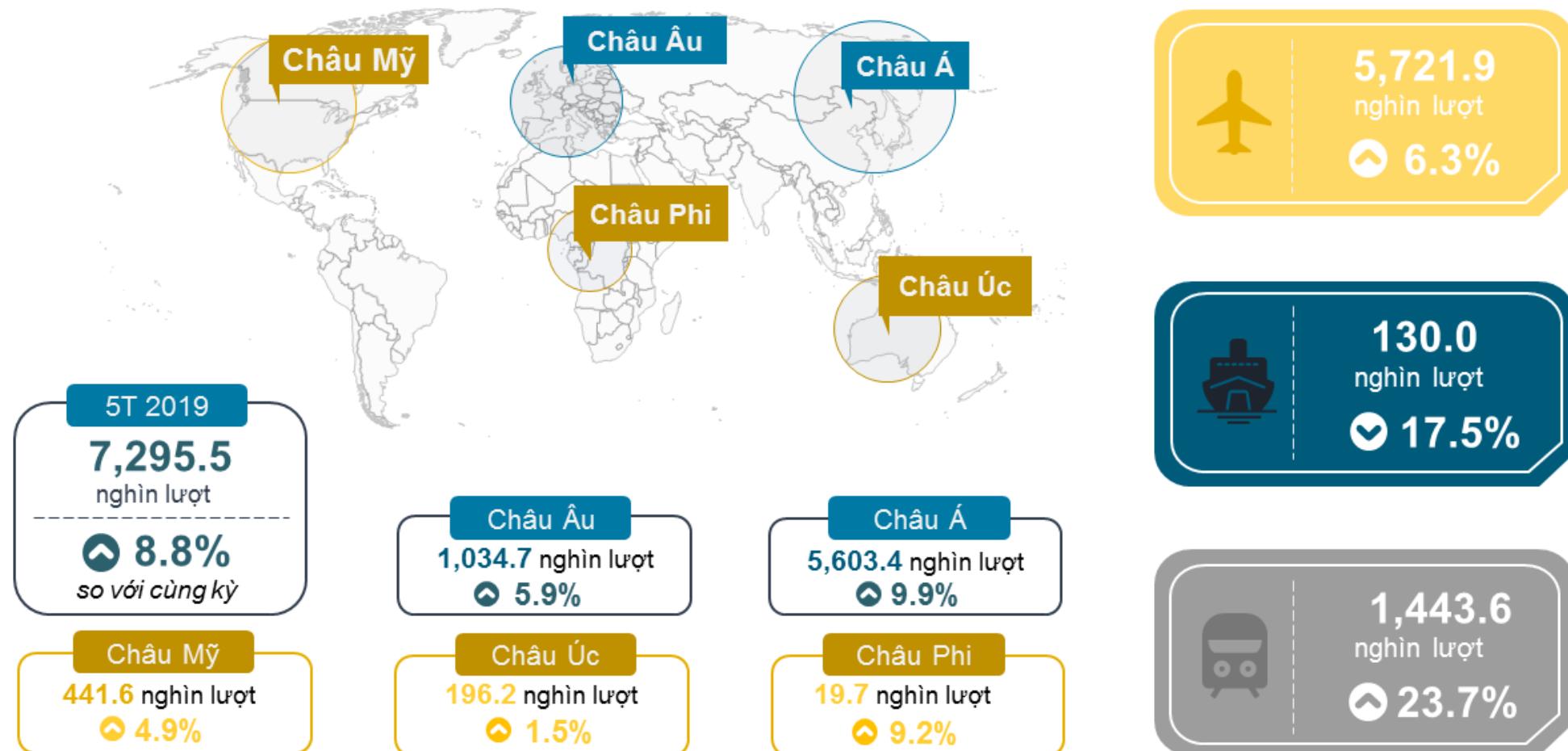
TỔNG CHI NGÂN SÁCH NHÀ NƯỚC



Tổng chi ngân sách Nhà nước bằng
29.8% so với dự toán năm 2019

Nguồn: Tổng cục Thống kê, Vietdata tổng hợp

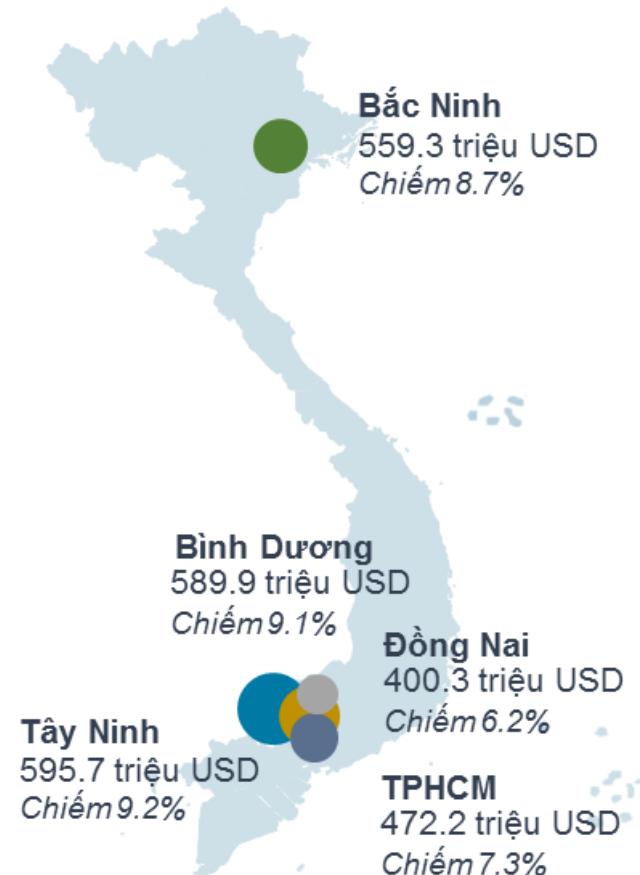
KHÁCH QUỐC TẾ ĐẾN VIỆT NAM 5T 2019



THU HÚT VỐN ĐẦU TƯ NƯỚC NGOÀI

TOP 5

Khu vực có tổng vốn đăng ký
cấp mới nhiều nhất trong 5T 2019



Tính đến 20/5, tổng vốn đầu tư
đăng ký mới và tăng thêm đạt:

9.09 tỷ USD
so với cùng kỳ năm 2018
 27.1%

Tổng vốn thực hiện trong 5T 2019

7.30 tỷ USD
so với cùng kỳ năm 2018
 7.8%

1,363 dự án được **CẤP PHÉP MỚI**

6.46 tỷ USD **Tổng vốn đăng ký**

505 dự án **ĐIỀU CHỈNH TĂNG VDT**

2.63 tỷ USD **Tổng VĐK** tăng thêm

3,160 lượt **GÓP VỐN/ MUA CỔ PHẦN**

7.65 tỷ USD **Tổng giá trị góp vốn**

Nhóm ngành thu hút vốn FDI nhiều nhất 5T 2019

(số vốn đăng ký của các dự án được cấp phép mới)



Chế biến chế tạo
4.74 tỷ USD 73.5%



Kinh doanh BĐS
742.3 triệu USD 11.5%



Ngành khác
971.2 triệu USD 15.0%

“Một bức tranh bằng cả nghìn lời nói”

XUẤT NHẬP KHẨU HÀNG HÓA

XUẤT
KHẨU

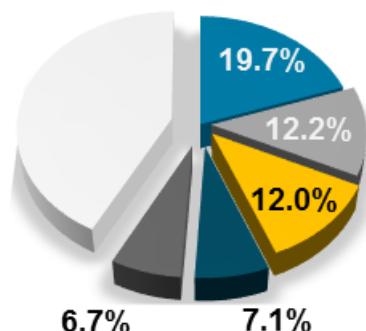
101.74
tỷ USD
↑6.7%

Tổng kim ngạch xuất khẩu
hàng hóa 5T 2019
so với cùng kỳ 2018

Thị trường xuất khẩu hàng hóa chính



Tỷ trọng một số mặt hàng XK chủ yếu



- Điện thoại/ Linh kiện
- Điện tử/ Máy tính/ Linh kiện
- Dệt, may
- Giày dép
- Máy móc/ Thiết bị/ Phụ tùng
- Khác

NHẬP
KHẨU

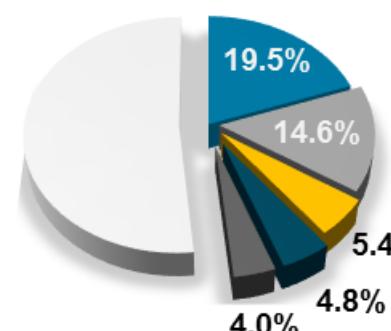
101.28
tỷ USD
↑10.3%

Tổng kim ngạch nhập khẩu
hàng hóa 5T 2019
so với cùng kỳ 2018

Thị trường nhập khẩu hàng hóa chính

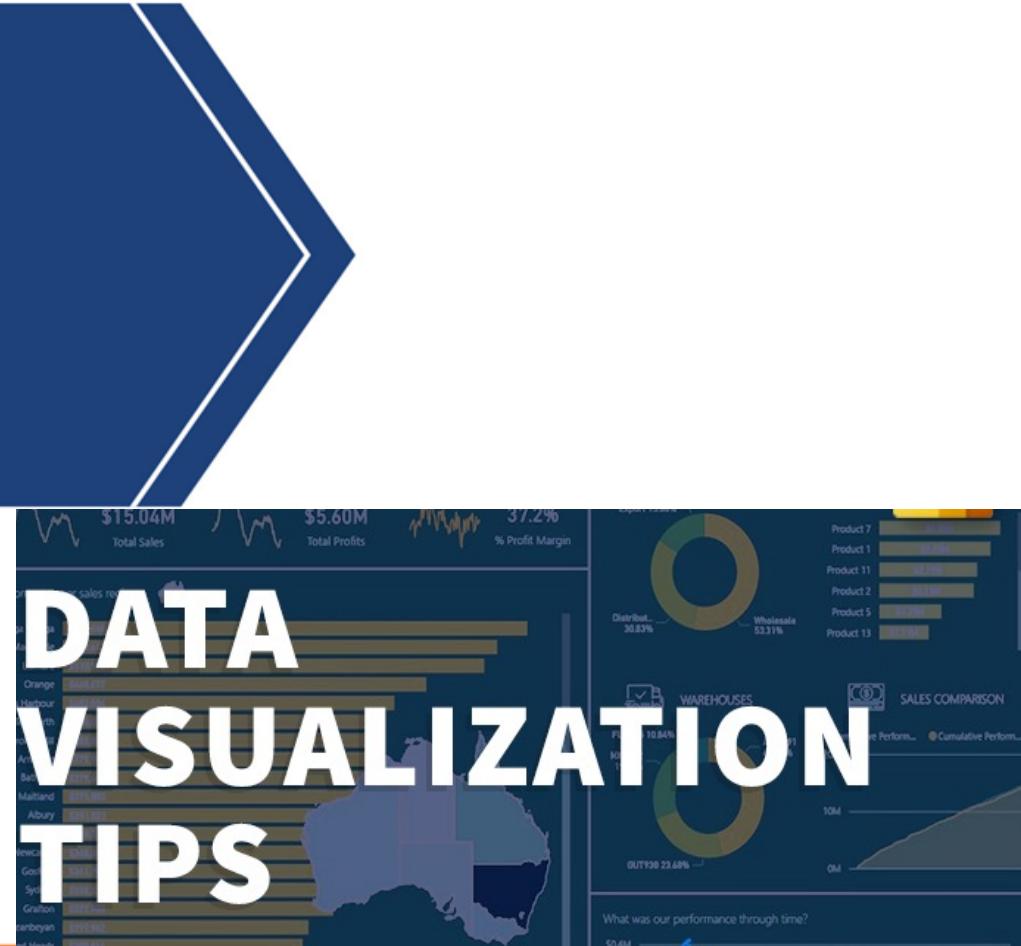


Tỷ trọng một số mặt hàng NK chủ yếu



- Điện tử/ Máy tính/ Linh kiện
- Máy móc/ Thiết bị/ Phụ tùng
- Vải
- Điện thoại/ Linh kiện
- Sắt thép
- Khác

3. Một số lưu ý khi trực quan hoá dữ liệu



<https://blog.csgsolutions.com/6-tips-for-creating-effective-data-visualizations>

Data Visualization Tips



VINBIGDATA VINGROUP

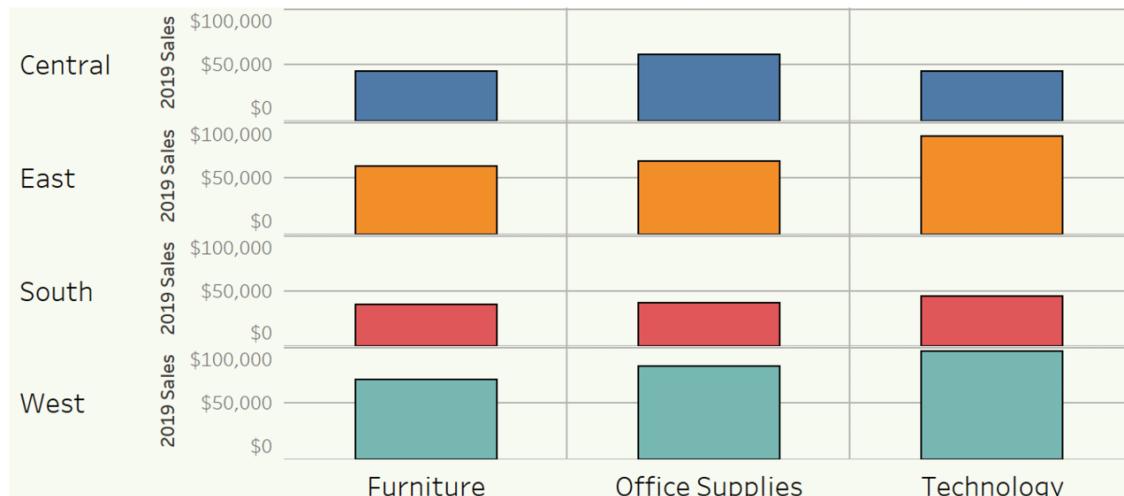
Academy
Vietnam



Data Visualization Tips

Tips 1: Xác định rõ mục đích và đối tượng cần truyền tải thông tin.

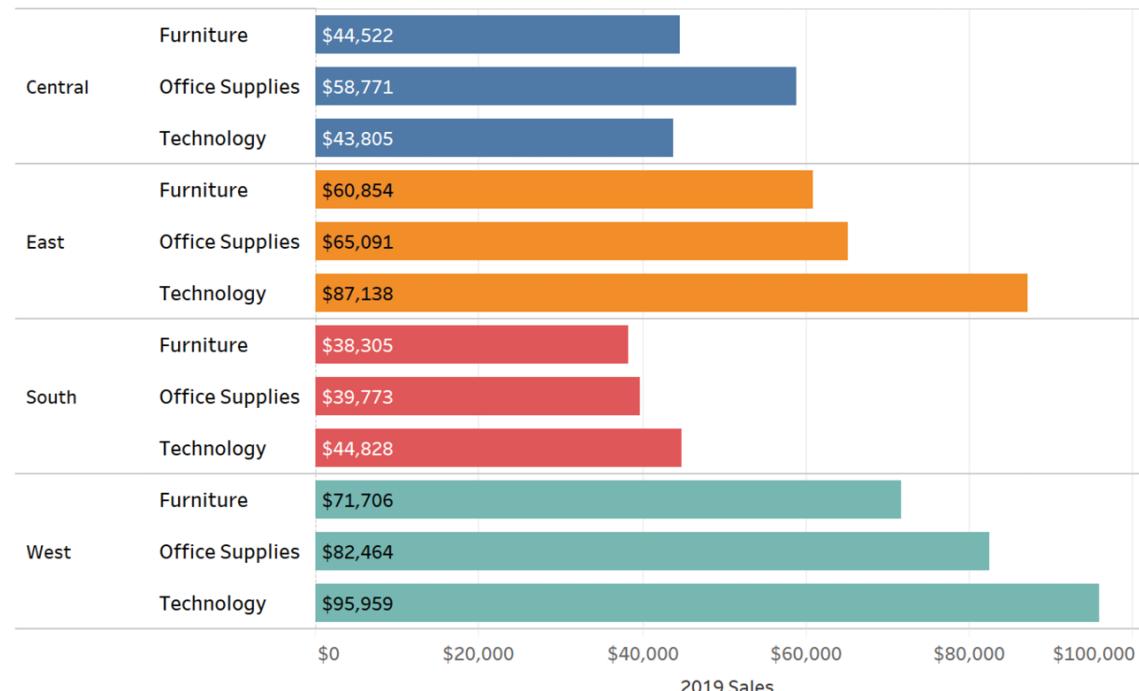
sales by region and category



| | Central | East | South | West |
|-----------------|----------|----------|----------|----------|
| Furniture | \$44,522 | \$60,854 | \$38,305 | \$71,706 |
| Office Supplies | \$58,771 | \$65,091 | \$39,773 | \$82,464 |
| Technology | \$43,805 | \$87,138 | \$44,828 | \$95,959 |

! ineffective

sales by region and category

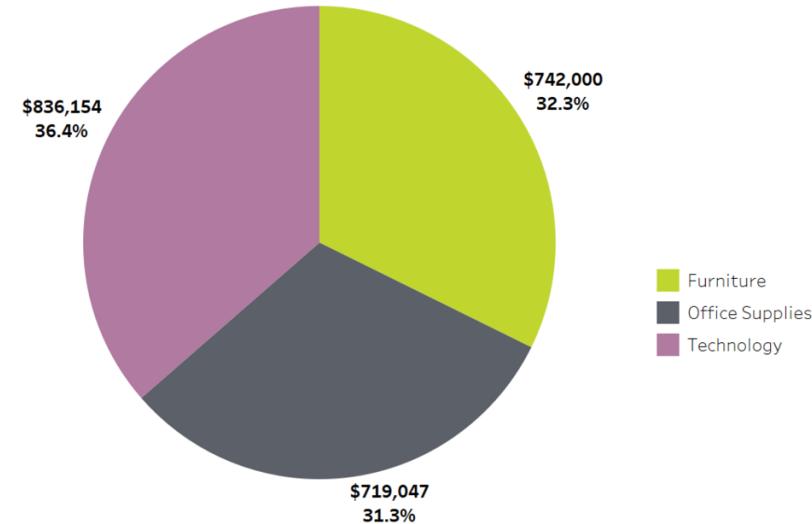


✓ effective

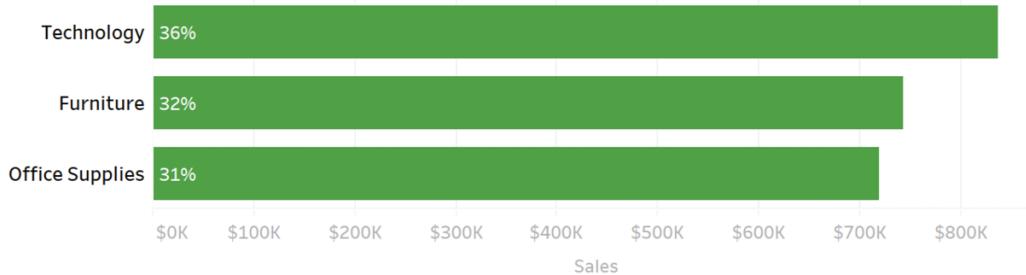
Data Visualization Tips

Tips 2: Lựa chọn loại biểu đồ phù hợp với mục đích của mình.

sales by product category



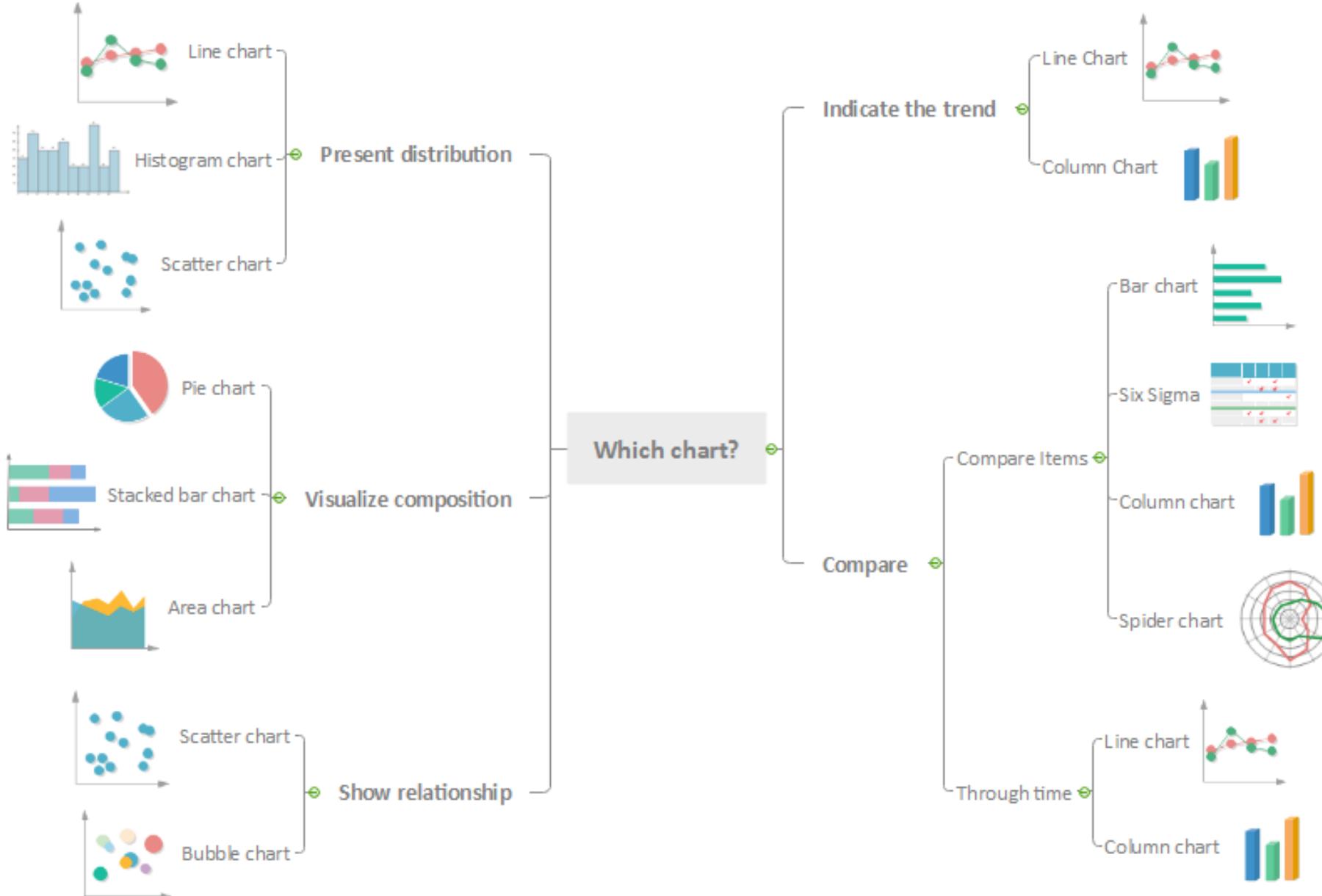
sales by product category



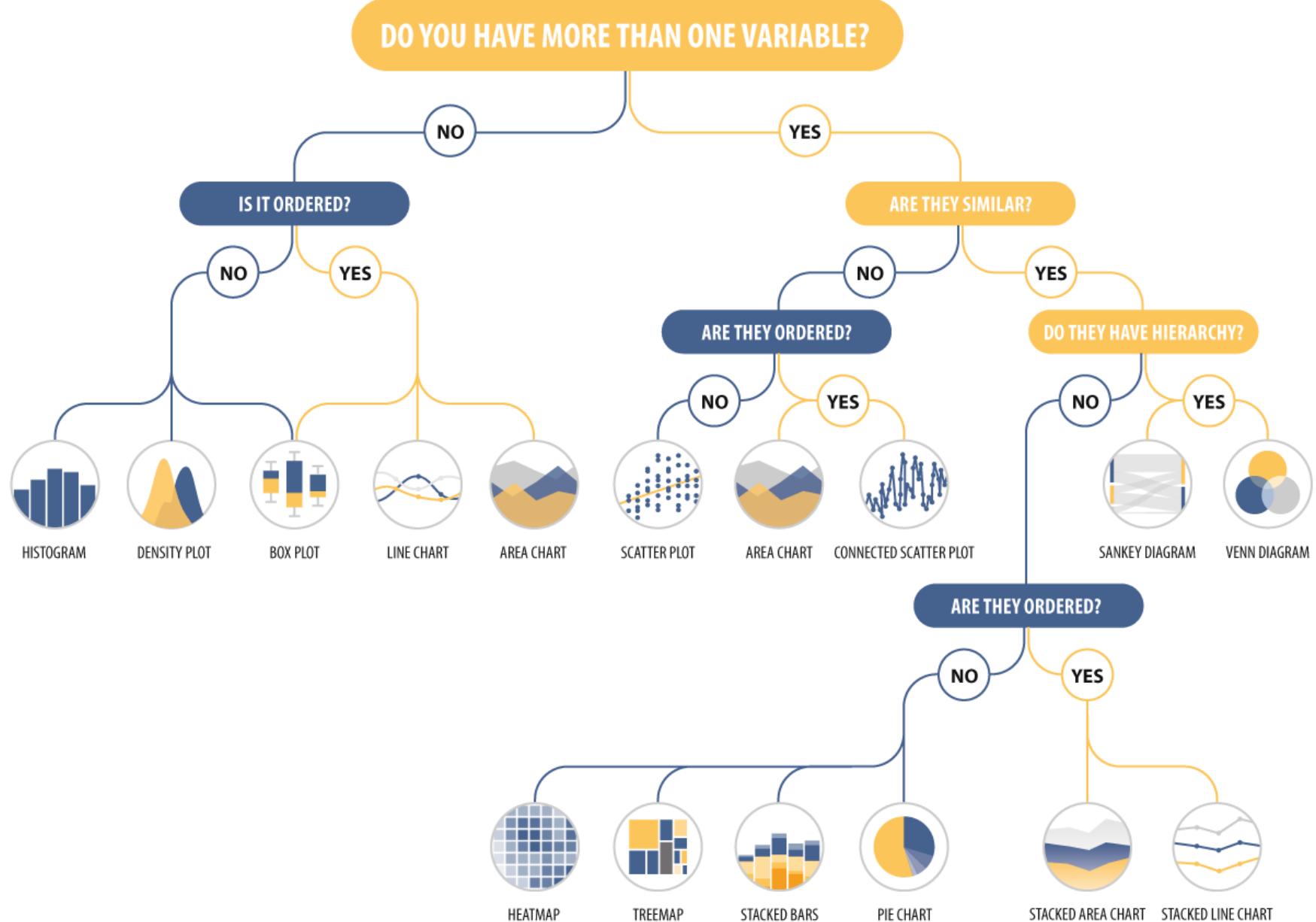
! ineffective

✓ effective

Data Visualization Tips



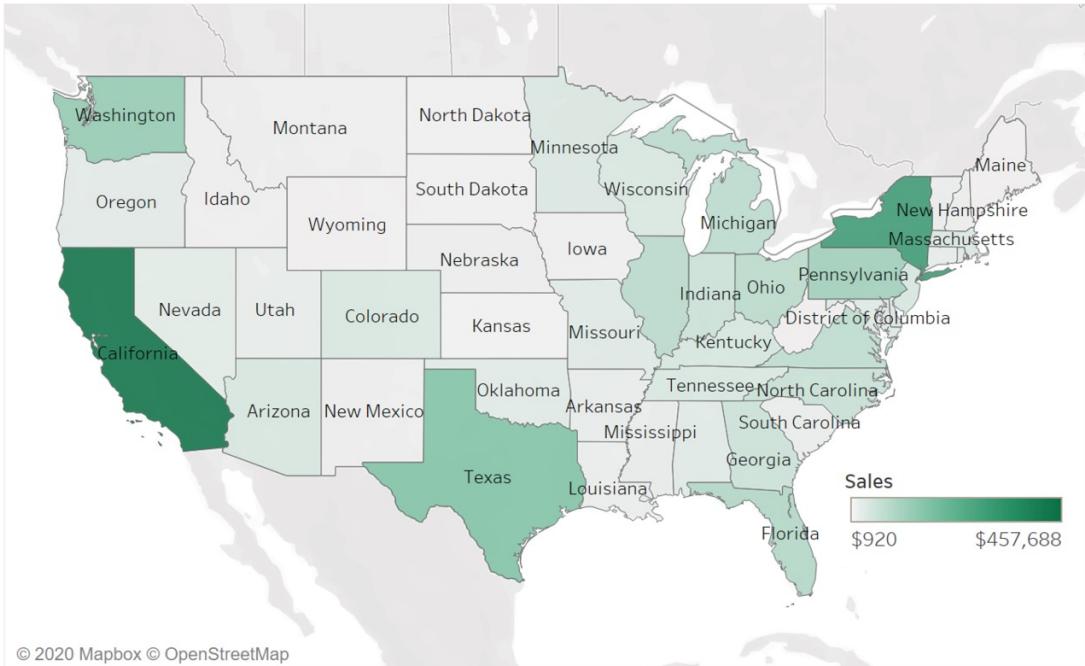
Data Visualization Tips



Data Visualization Tips

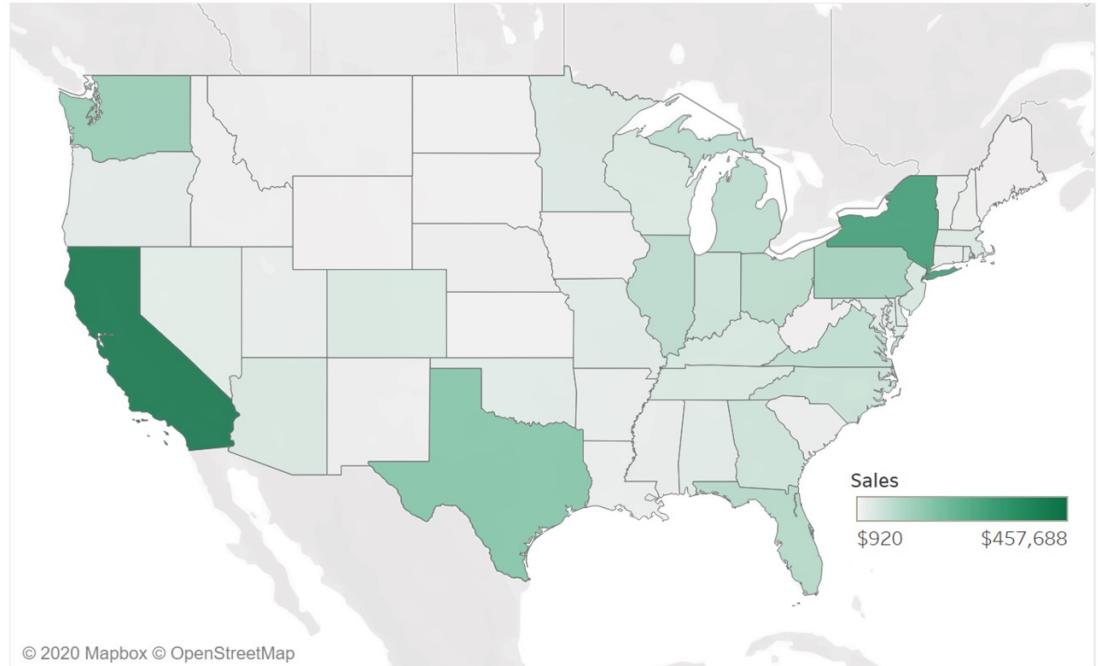
Tips 3: Đưa văn bản, nhãn vào biểu đồ hợp lý, tránh lộn xộn

total sales map



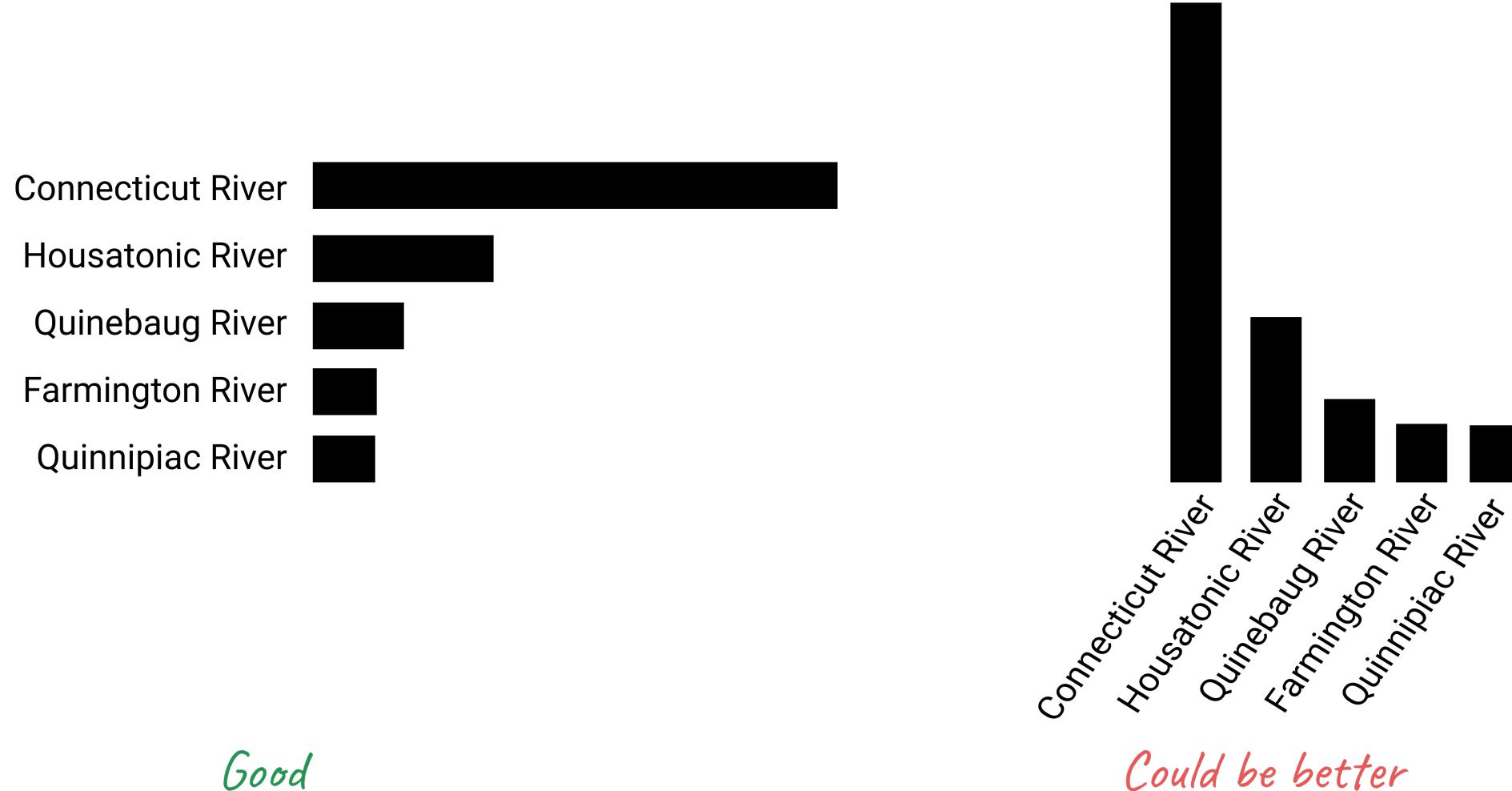
! ineffective

total sales map



✓ effective

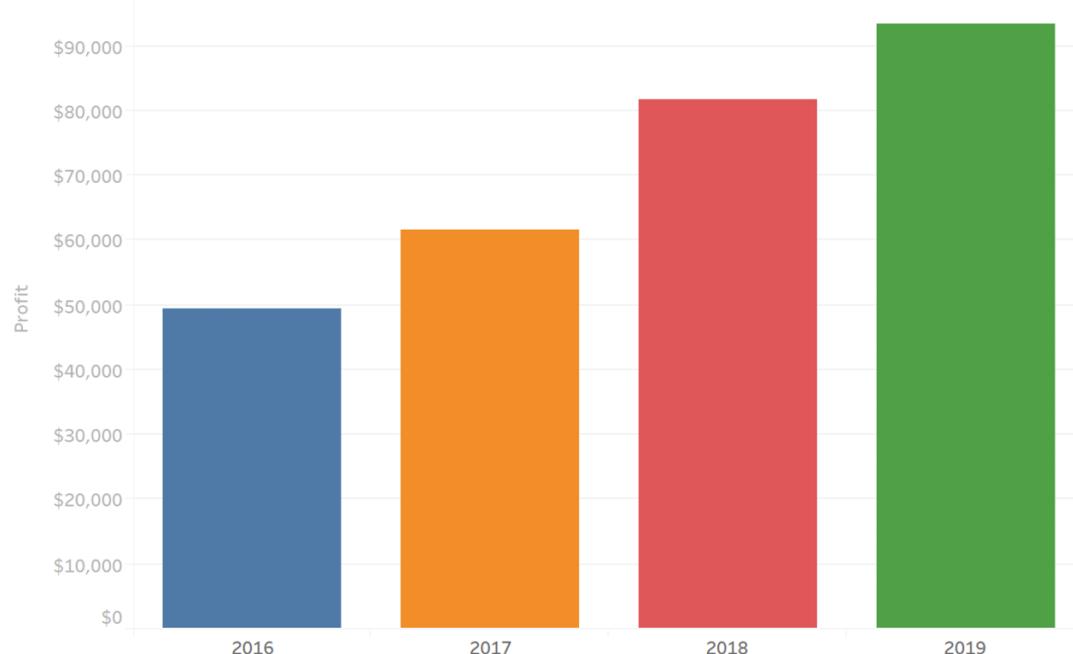
Tips 3: Đưa văn bản, nhãn vào biểu đồ hợp lý, tránh lộn xộn (T)



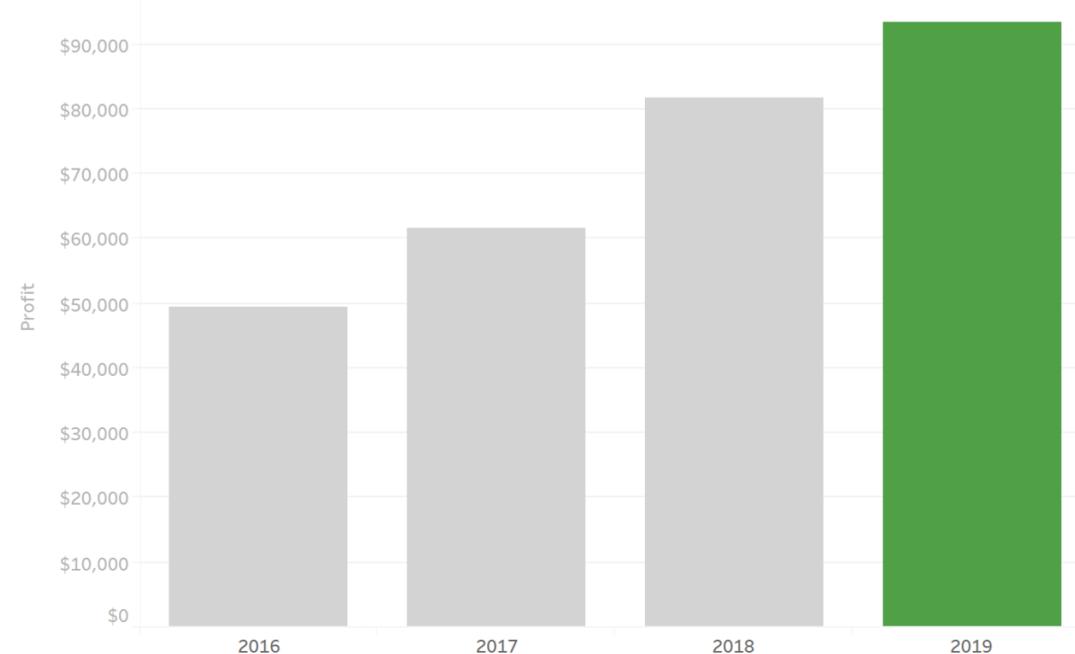
Data Visualization Tips

Tips 4: Sử dụng màu sắc hiệu quả để làm nổi bật các thông tin quan trọng.

profit by year



profit by year

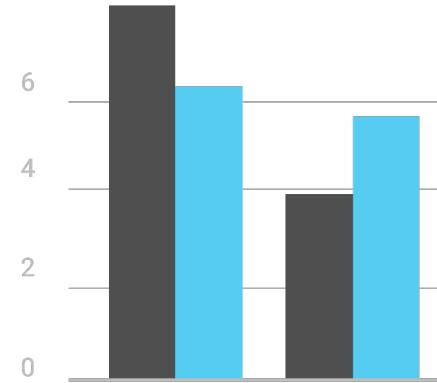


! ineffective

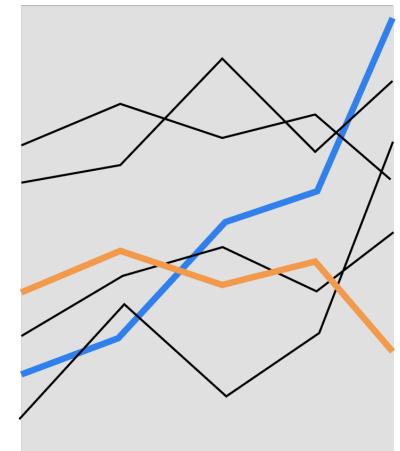
✓ effective

Data Visualization Tips

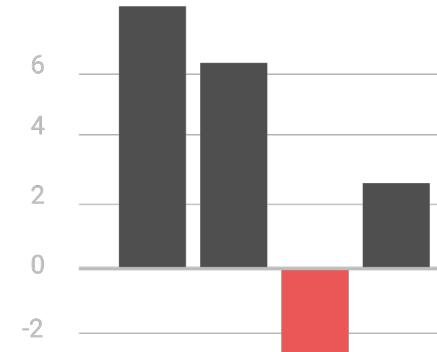
Tips 4: Sử dụng màu sắc hiệu quả để làm nổi bật các thông tin quan trọng (*t*).



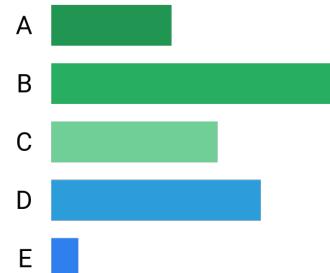
Good. Blue helps to distinguish between different data series



Good. Colored lines are distinguishable



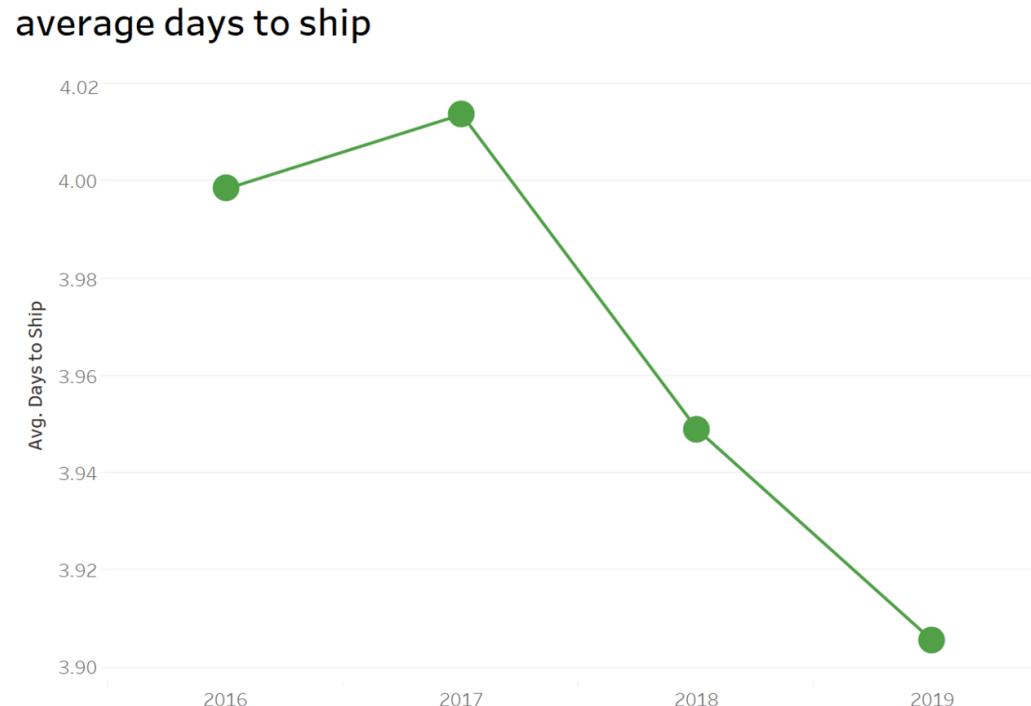
Fine. Red adds emphasis, but is not absolutely necessary



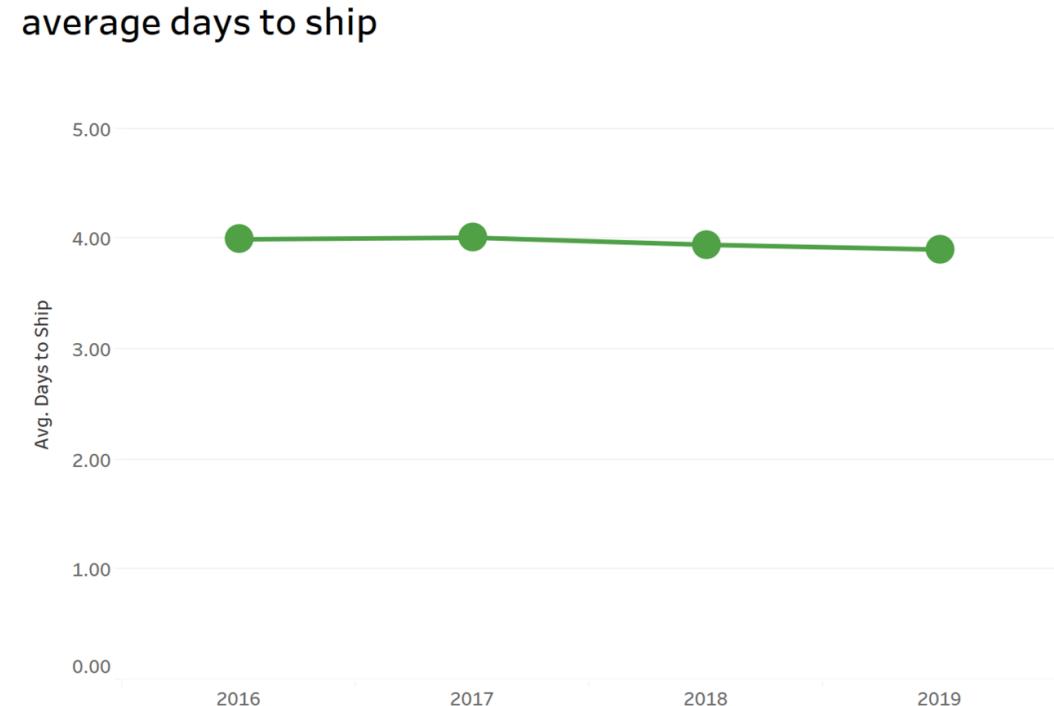
Bad. Colors are too similar, and are not needed for a separated bar chart

Data Visualization Tips

Tips 5: Tránh để người xem hình dung sai lệch dữ liệu.



! ineffective

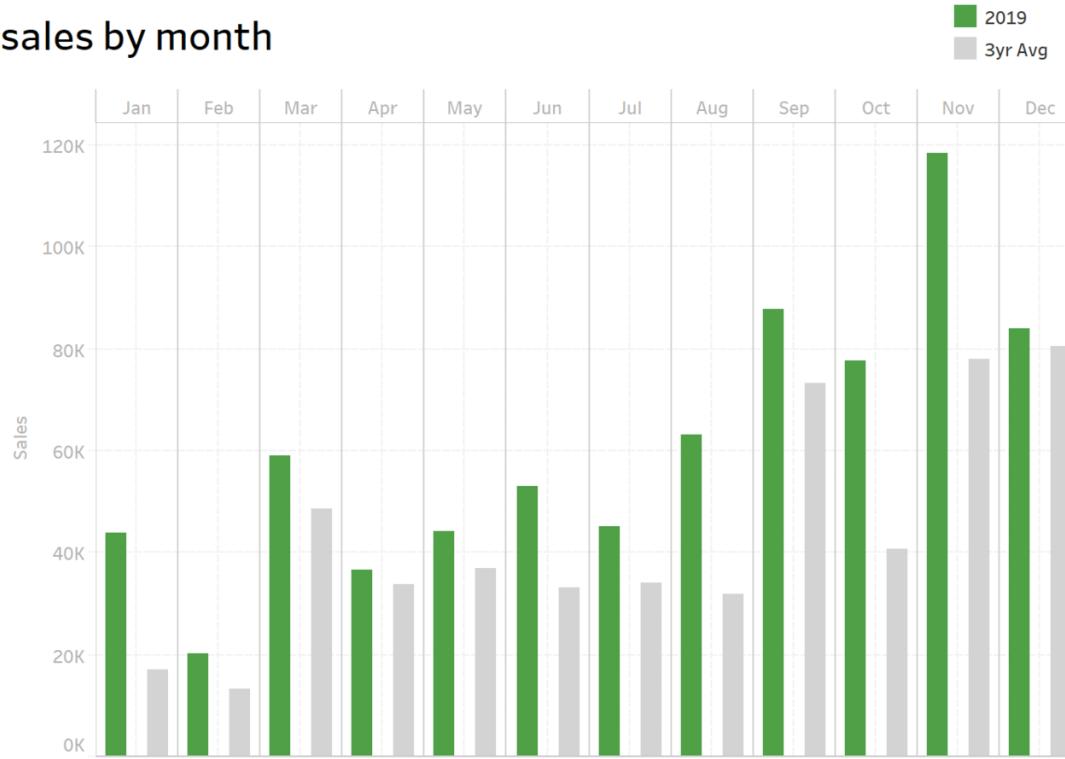


✓ effective

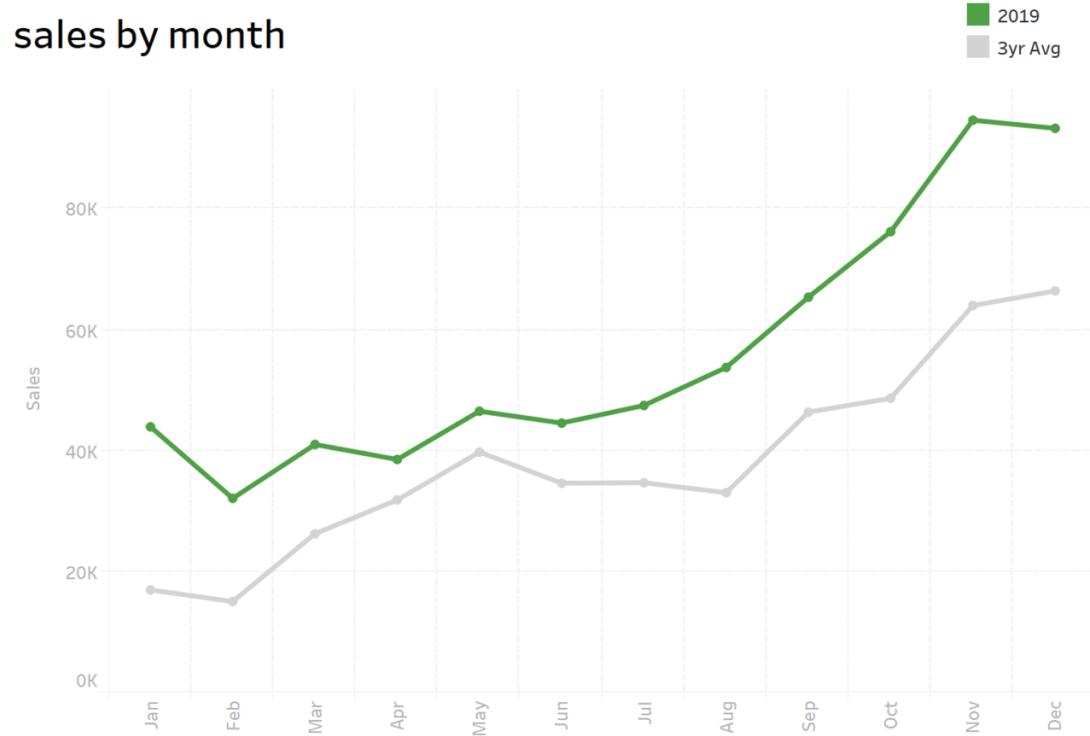
Data Visualization Tips

Tips 6: Sử dụng các biểu đồ càng đơn giản càng tốt.

sales by month



sales by month



! ineffective

✓ effective

Data Visualization Tips

Tips 7: Cân nhắc sắp xếp dữ liệu để đạt hiệu quả

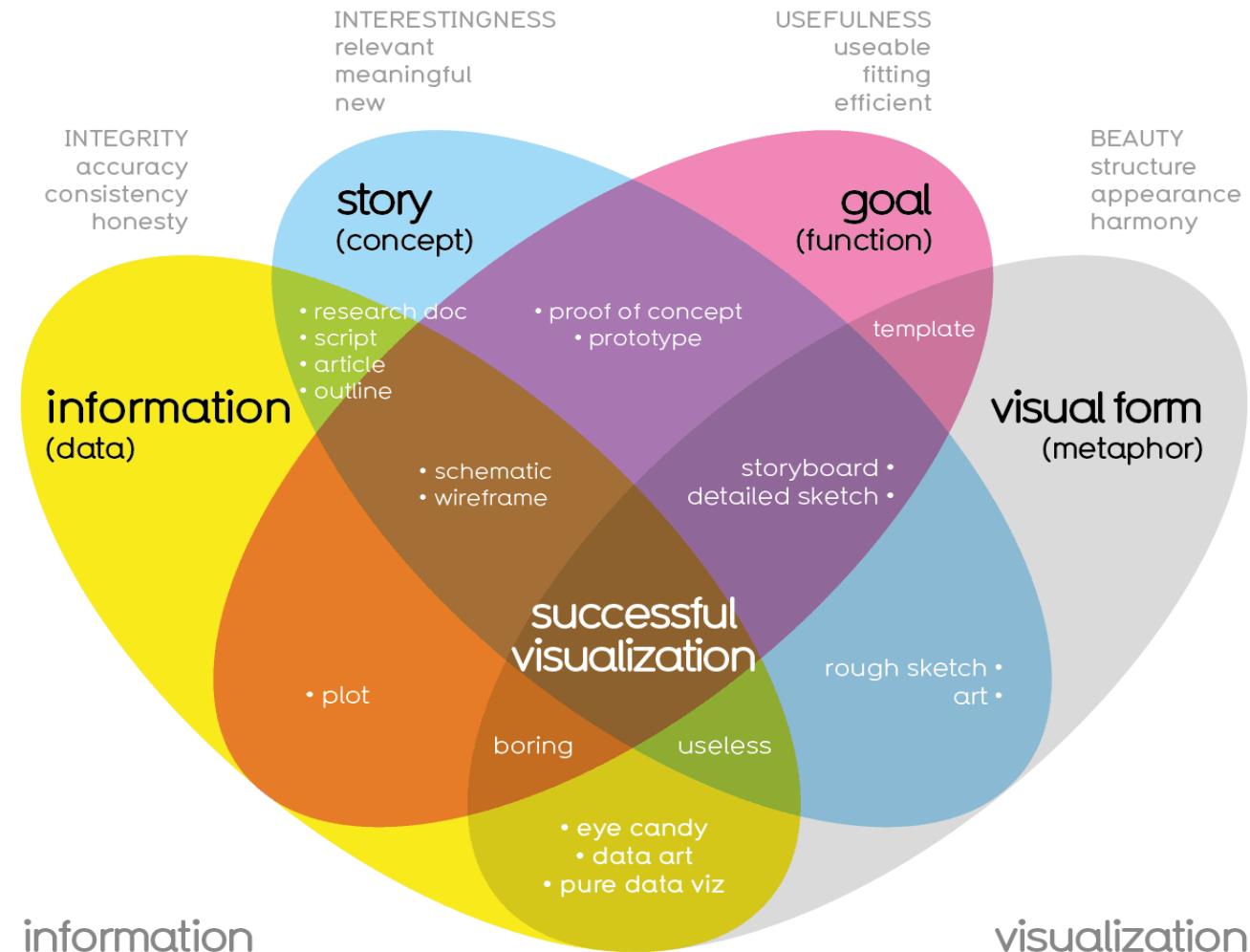


Good (Alphabetical)

Good (Values sorted)

Bad (Random order)

What Makes a Good Visualization?



3. Thư viện trực quan hóa với Python

Một số thư viện trực quan hóa

Có nhiều thư viện mạnh mẽ để trực quan hóa dữ liệu với ngôn ngữ lập trình Python.

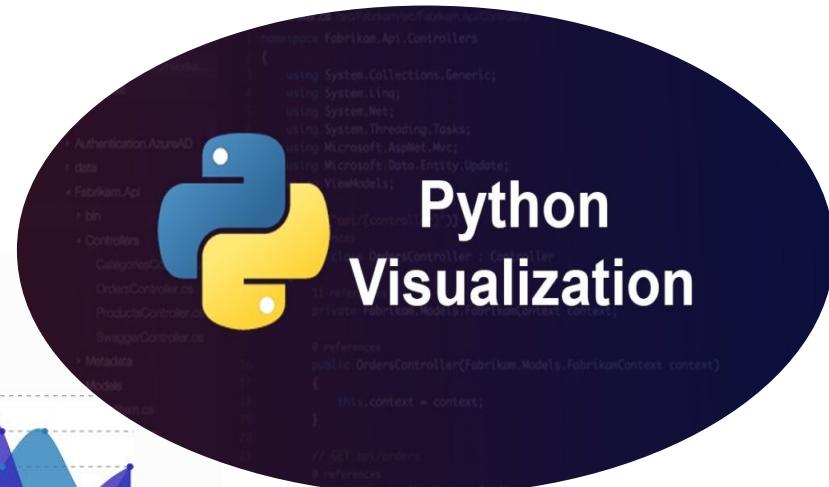
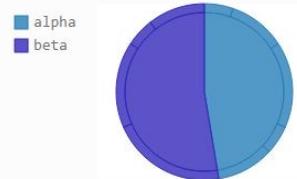
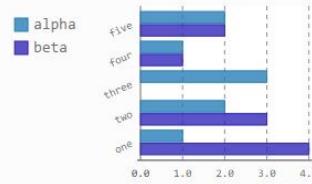
matplotlib

seaborn

plotly

Pygal

Sexy python charting



bokeh

Altair

Declarative Visualization in Python



- **Matplotlib** là thư viện dùng để vẽ đồ thị rất mạnh mẽ, có cú pháp tương tự như Matlab. Thư viện này được phát triển sớm nhất, 2003.
- Hỗ trợ nhiều loại biểu đồ, đặc biệt là các loại được sử dụng trong nghiên cứu hoặc kinh tế như biểu đồng đường, cột, tần suất (histograms), tương quan, scatterplots...
- Cấu trúc của Matplotlib gồm nhiều phần, phục vụ cho các mục đích sử dụng khác nhau. Trong đó module pyplot được sử dụng nhiều nhất, có cú pháp tương tự như Matlab.
- Matplotlib miễn phí và mã nguồn mở.

Tham khảo:

- + File: **CheatSheet-Matplotlib**
- + Link web: [**Matplotlib package!**](#)

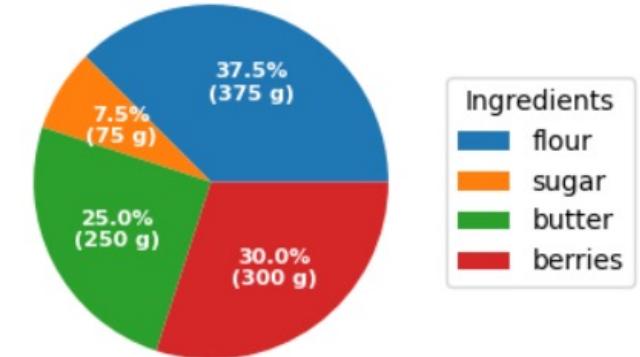
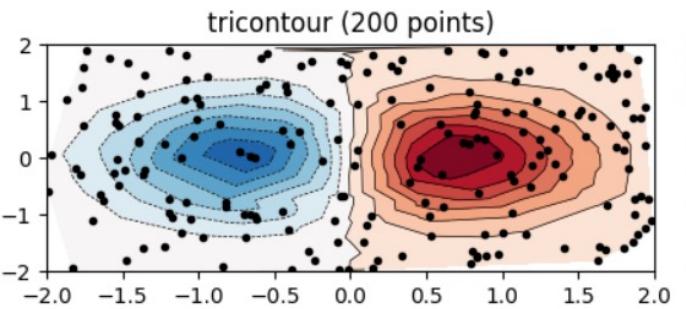
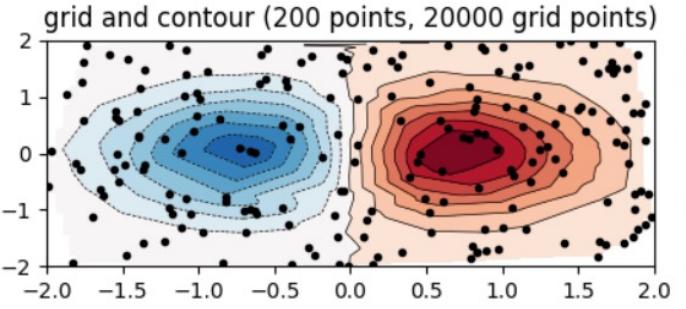
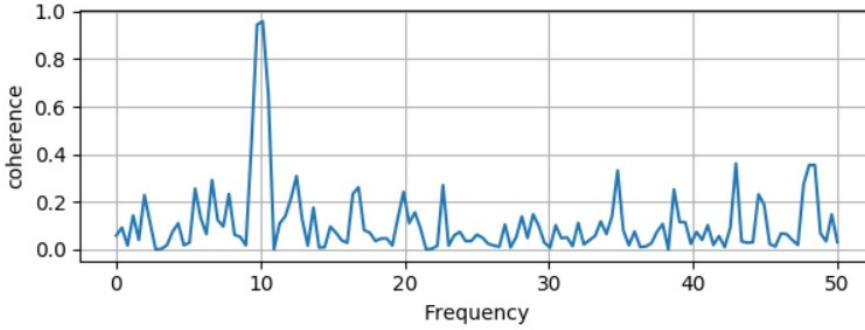
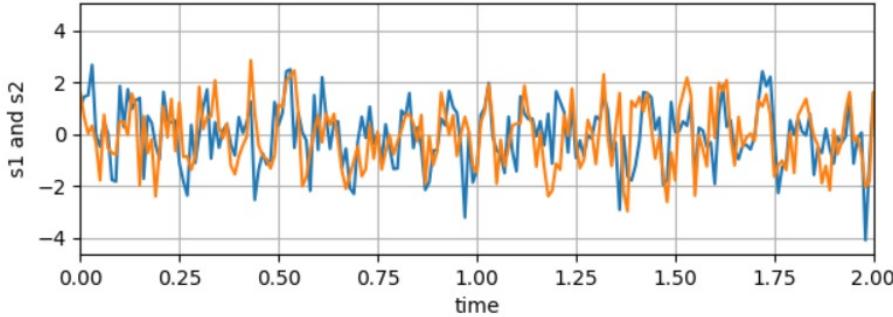


Thư viện matplotlib

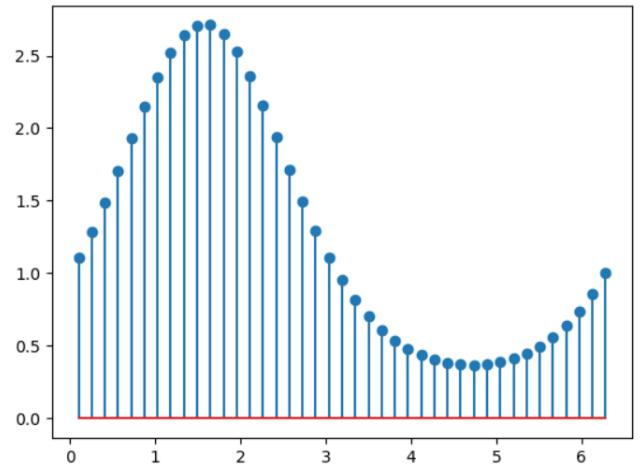
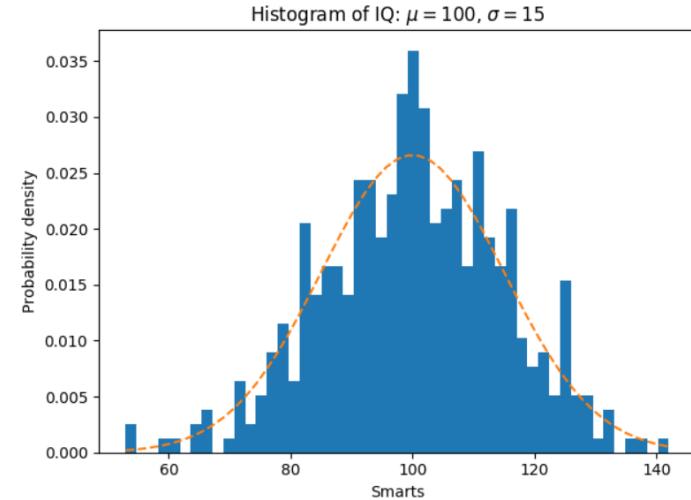
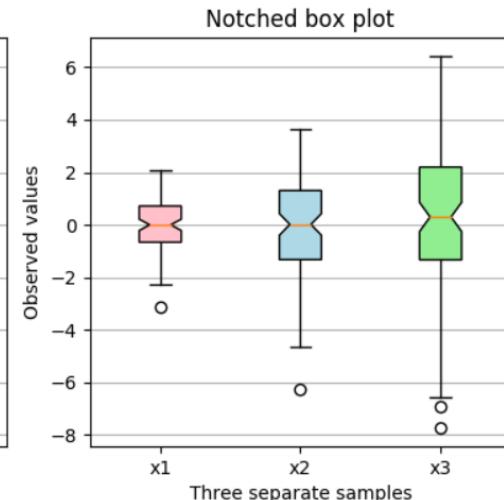
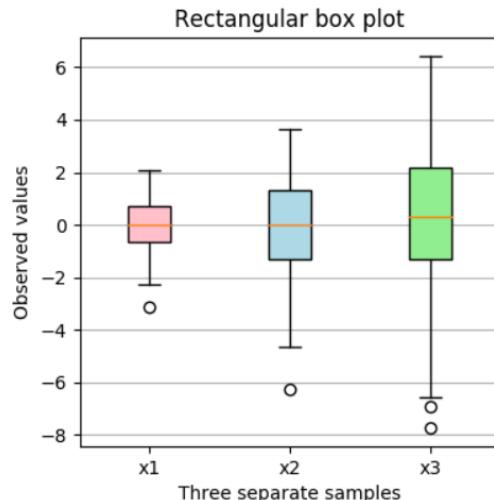


VINBIGDATA VINGROUP

Academy
Vietnam



matplotlib



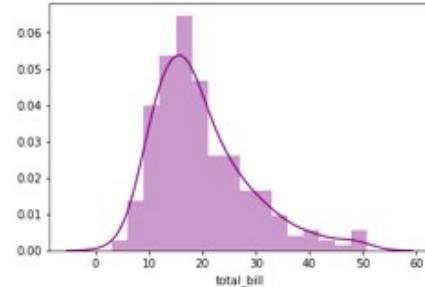
- **Seaborn** là một trong những thư viện mạnh mẽ, phổ biến trong việc trực quan hóa dữ liệu. Seaborn được coi là công cụ bổ sung cho Matplotlib (Mở rộng của Matplotlib).
- Seaborn bổ sung thêm các dạng biểu đồ mạnh mẽ, và nhiều tính năng mới, giúp cho việc trực quan hóa các biểu đồ, dữ liệu phức tạp trở nên dễ dàng hơn. Một số ưu điểm của Seaborn:
 1. Mục đích chính của Seaborn là làm cho việc trực quan hóa dữ liệu trở nên dễ dàng. Do đó, nó được xây dựng để tự động xử lý rất nhiều phép toán phức tạp ở phía sau.
 2. Seaborn hoạt động cực kỳ hiệu quả với cấu trúc dữ liệu của Pandas, đây là một thư viện Python được sử dụng rộng rãi để phân tích dữ liệu.
 3. Seaborn được xây dựng trên Matplotlib, là một thư viện trực quan hóa Python khác. Matplotlib cực kỳ linh hoạt. Seaborn cho phép chúng ta tận dụng tính linh hoạt để tránh được sự phức tạp trong quá trình xử lý dữ liệu.



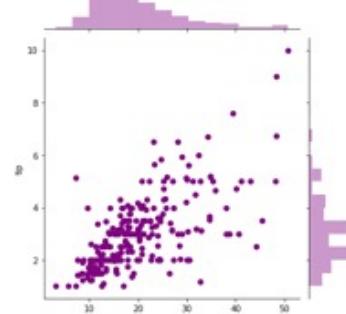
Tham khảo:

- + File: [CheatSheet-Seaborn](#)
- + Link web: [Seaborn package!](#) 31

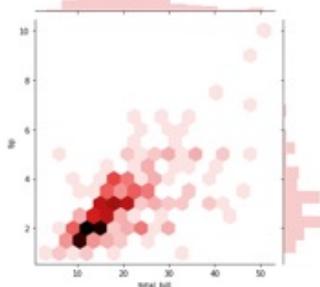
Seaborn Plots



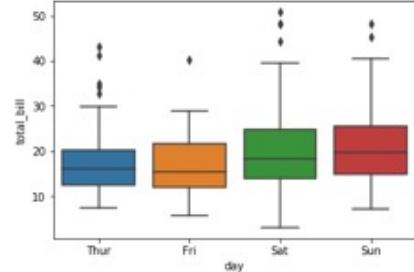
distplot



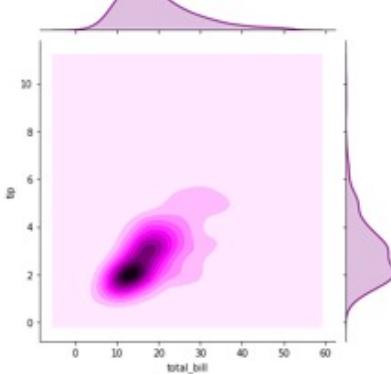
Jointplot



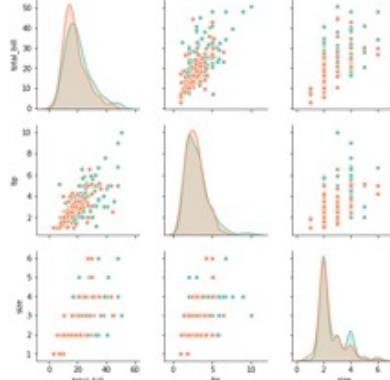
Hexplots



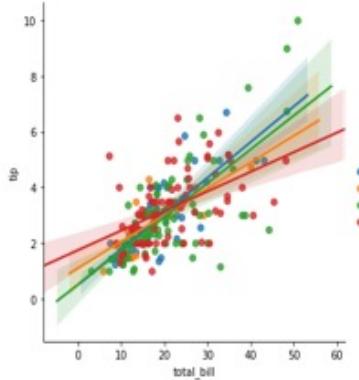
Boxplots



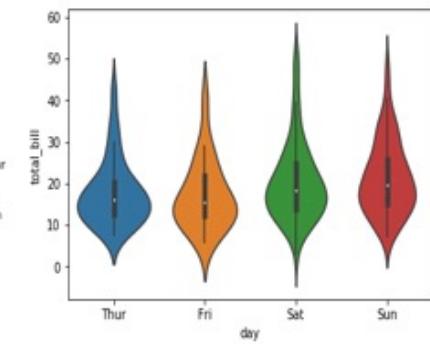
KDE Plot



Pair Plots



LM Plots



Violin Plots



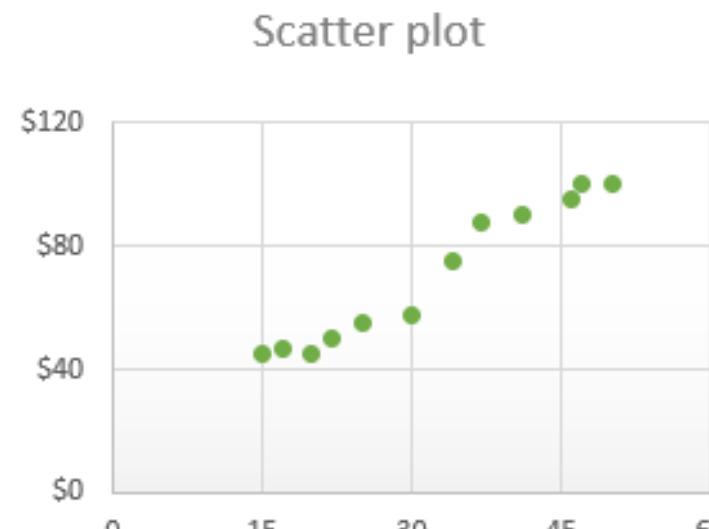
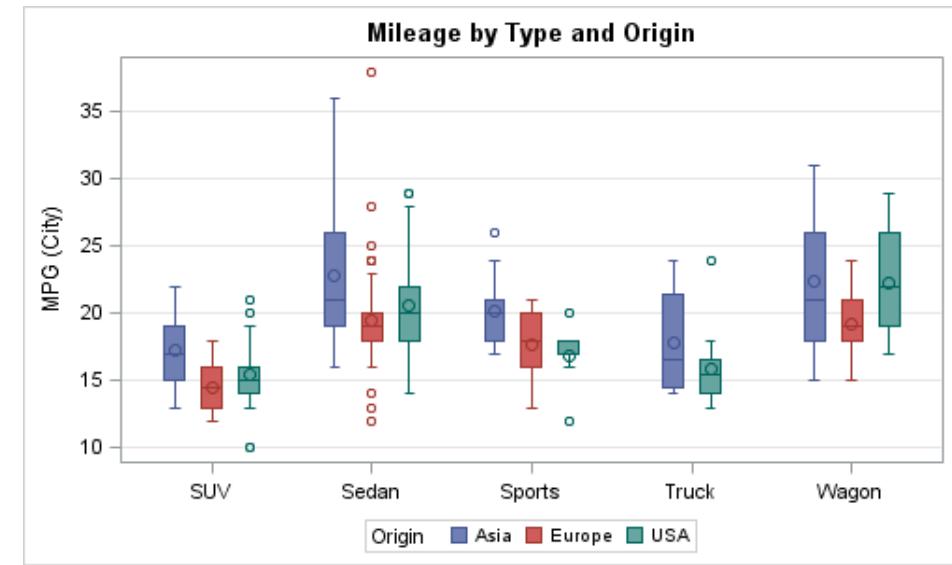
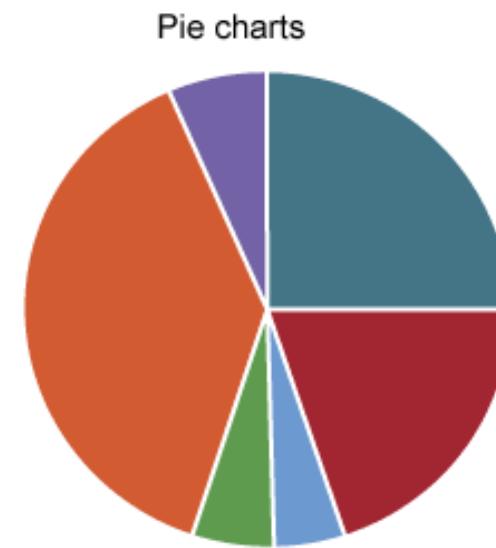
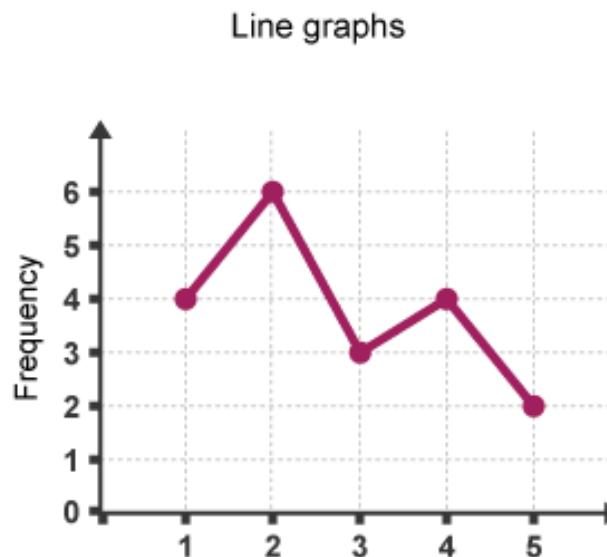
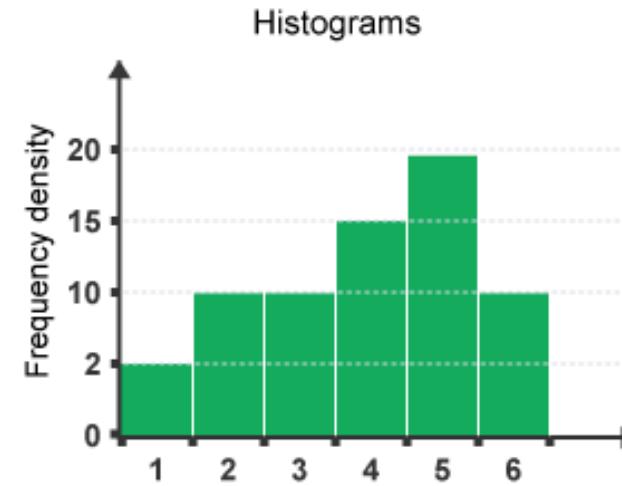
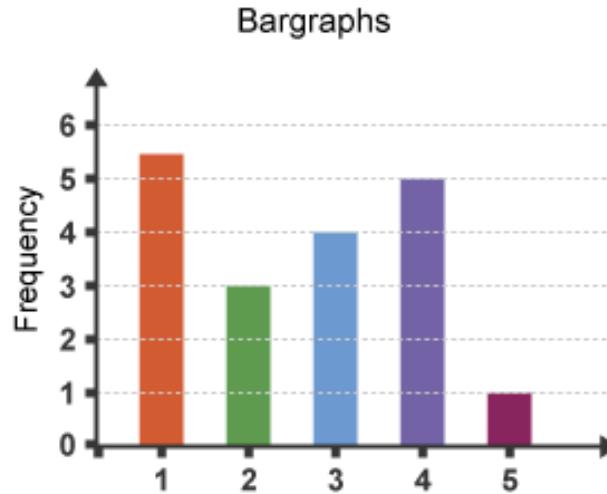
seaborn₃₂



MỘT SỐ BIỂU ĐỒ QUAN TRỌNG



Các dạng biểu đồ quan trọng

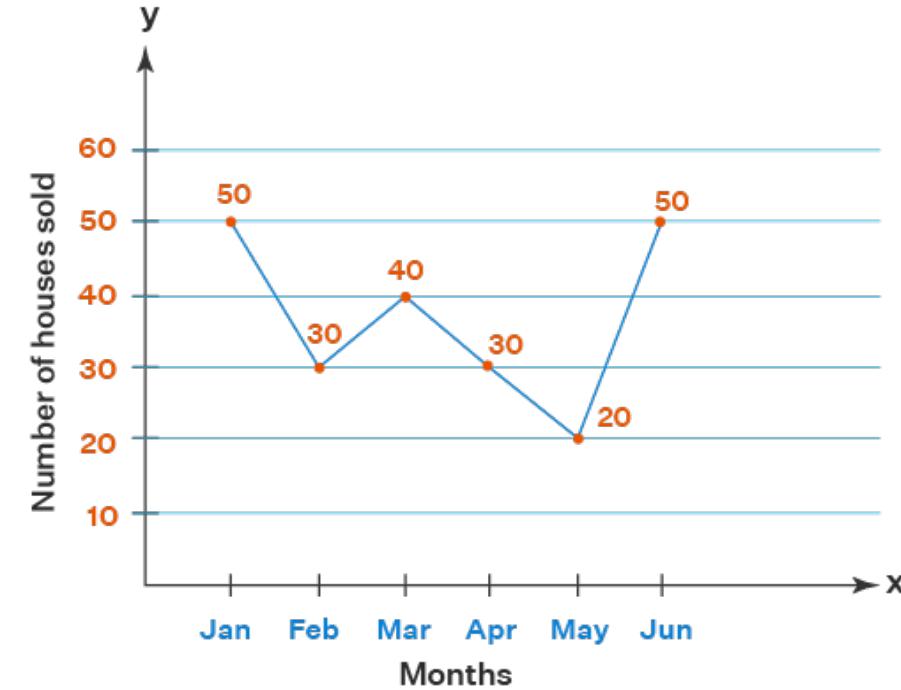
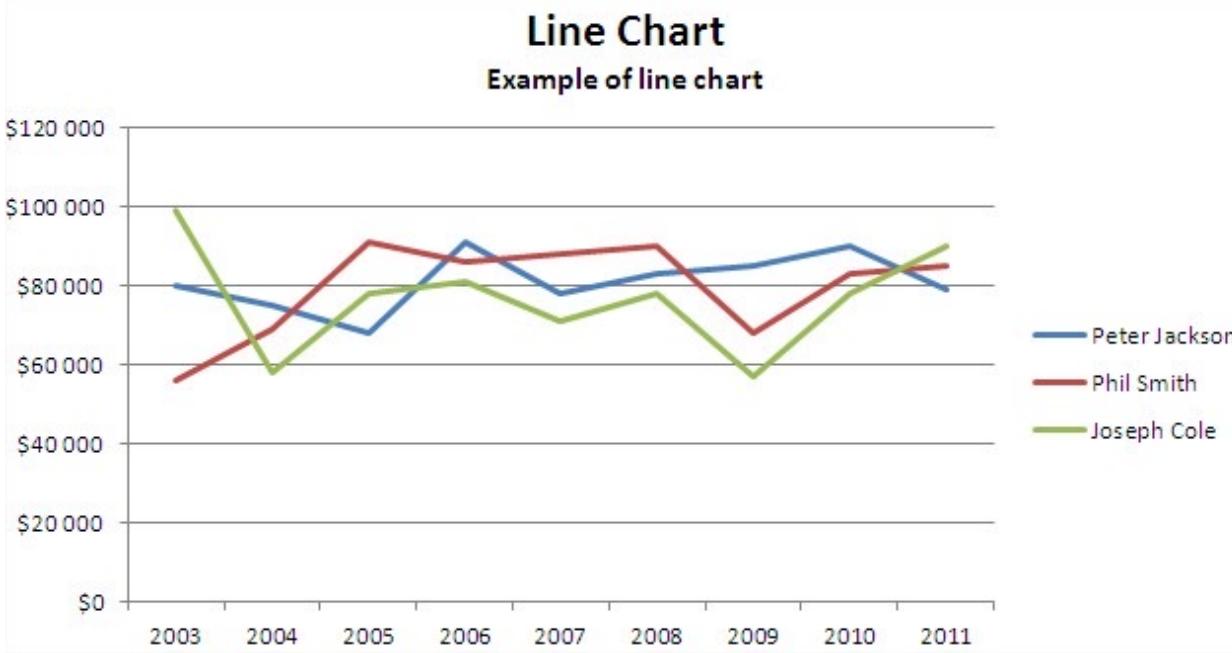




4. Biểu đồ đường (line chart)

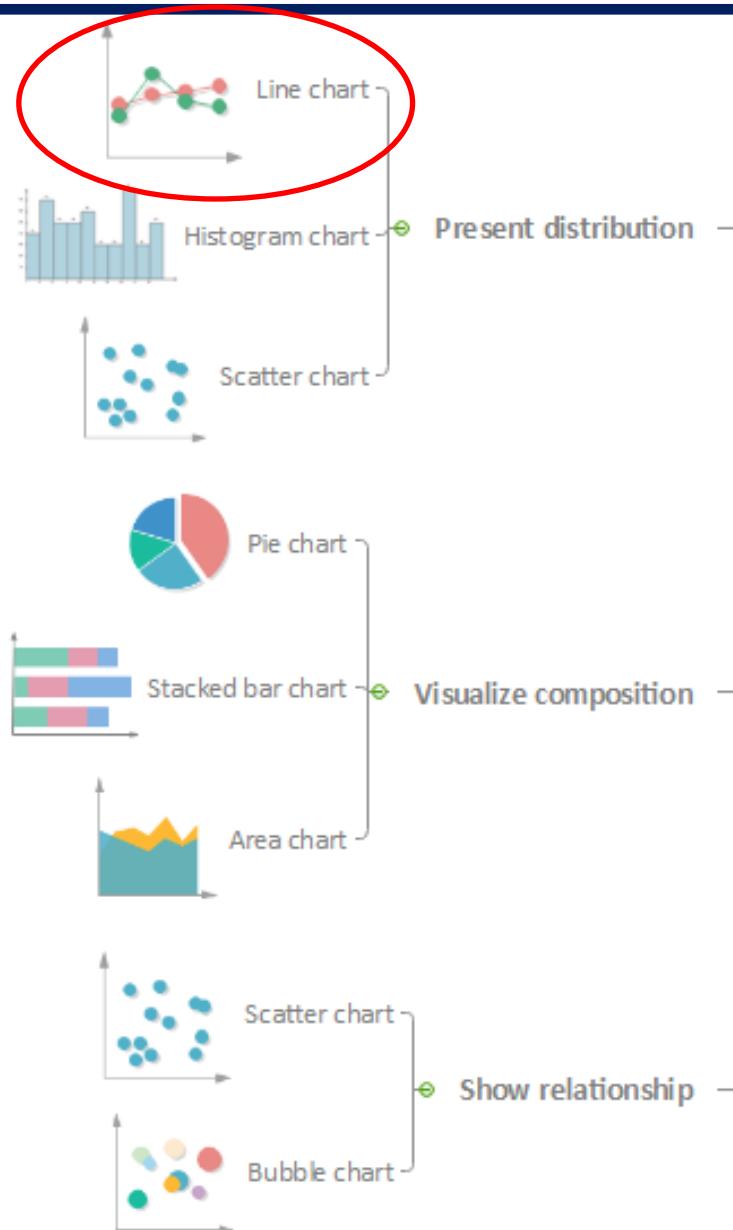


Đồ thị dạng đường (line chart)

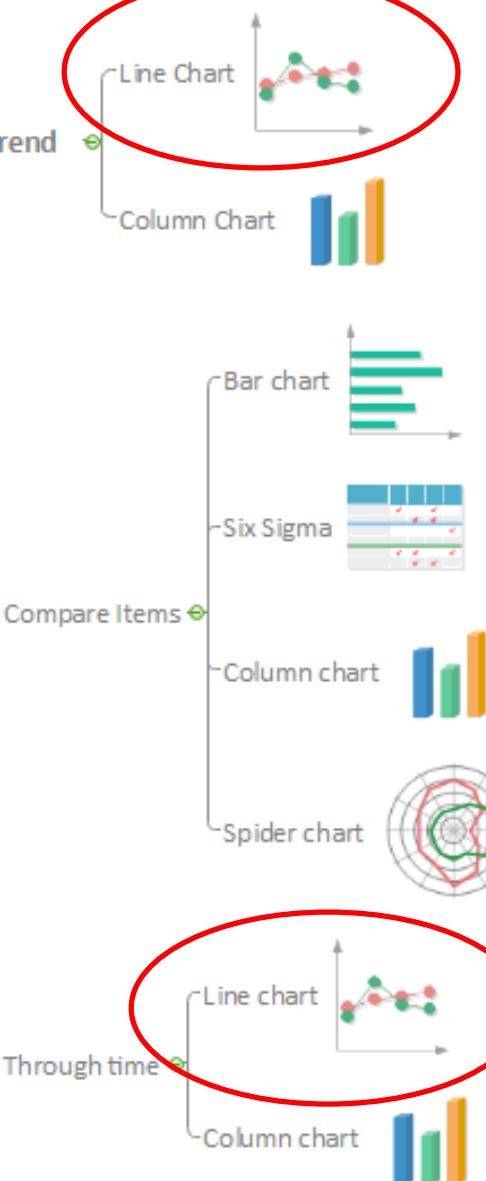


Line chart là một trong những dạng đồ thị phổ biến và hay được sử dụng trong thực tế.

- Khi muốn trình bày các dữ liệu liên mạch biểu đồ line chart là sự lựa chọn phù hợp. Các điểm trong biểu đồ đường được nối liền thành một đường, thể hiện mối quan hệ giữa các điểm đó. Thông thường các dòng dữ liệu sẽ liên quan đến các đơn vị đo lường thời gian như ngày, tháng, quý và năm.

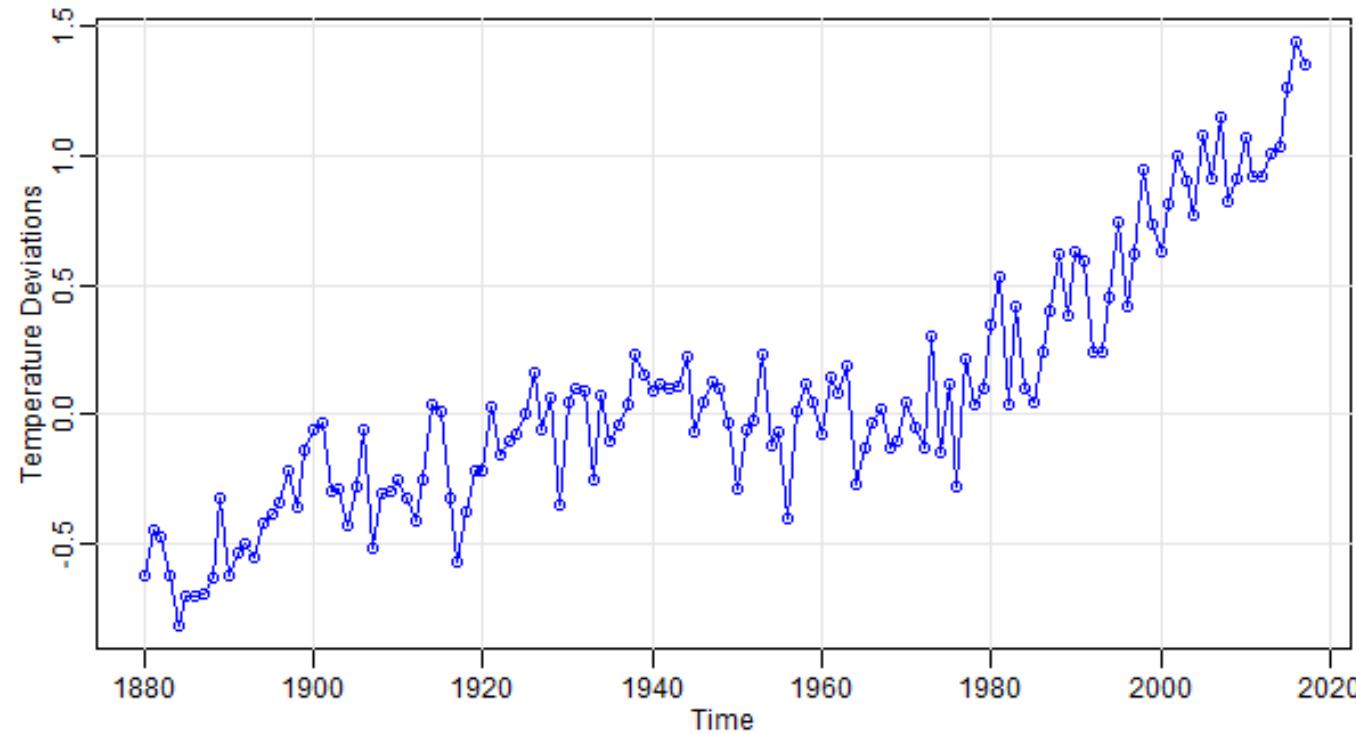


Indicate the trend

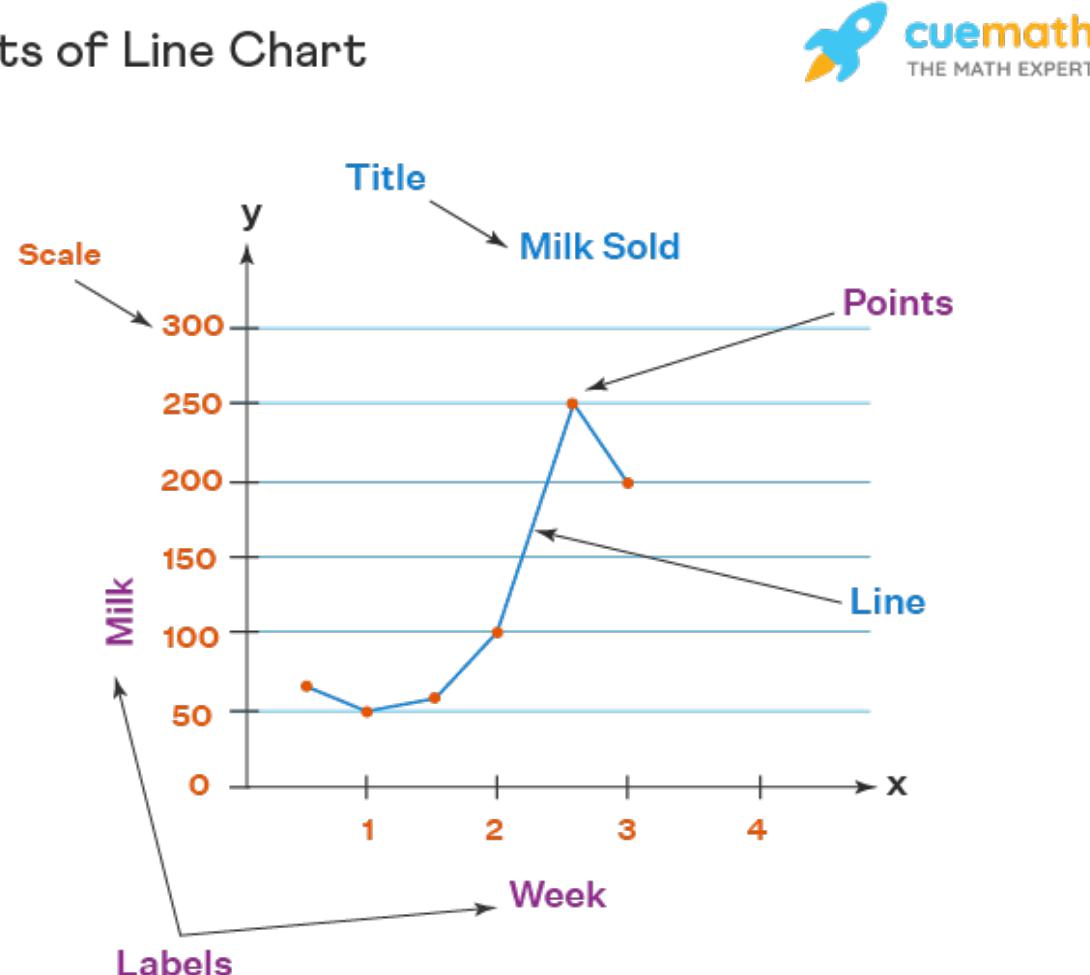


Đồ thị dạng đường (line chart)

Line chart giúp nhấn mạnh sự thay đổi trong dữ liệu của một biến vẽ trên trục x so với biến thứ 2 trên trục y.



Parts of Line Chart



Đồ thị dạng đường với Matplotlib



VINBIGDATA VINGROUP

Academy
Vietnam

Tập dữ liệu `gas_prices.csv`: Lưu trữ giá Gas của 10 nước trên thế giới trong
giai đoạn từ năm 1990 - 2008

```
1 data = pd.read_csv('Data_Visualize/gas_prices.csv')  
2 data
```

| | Year | Australia | Canada | France | Germany | Italy | Japan | Mexico | South Korea | UK | USA |
|----|------|-----------|--------|--------|---------|-------|-------|--------|-------------|------|------|
| 0 | 1990 | NaN | 1.87 | 3.63 | 2.65 | 4.59 | 3.16 | 1.00 | | 2.05 | 2.82 |
| 1 | 1991 | 1.96 | 1.92 | 3.45 | 2.90 | 4.50 | 3.46 | 1.30 | | 2.49 | 3.01 |
| 2 | 1992 | 1.89 | 1.73 | 3.56 | 3.27 | 4.53 | 3.58 | 1.50 | | 2.65 | 3.06 |
| 3 | 1993 | 1.73 | 1.57 | 3.41 | 3.07 | 3.68 | 4.16 | 1.56 | | 2.88 | 2.84 |
| 4 | 1994 | 1.84 | 1.45 | 3.59 | 3.52 | 3.70 | 4.36 | 1.48 | | 2.87 | 2.99 |
| 5 | 1995 | 1.95 | 1.53 | 4.26 | 3.96 | 4.00 | 4.43 | 1.11 | | 2.94 | 3.21 |
| 6 | 1996 | 2.12 | 1.61 | 4.41 | 3.94 | 4.39 | 3.64 | 1.25 | | 3.18 | 3.34 |
| 7 | 1997 | 2.05 | 1.62 | 4.00 | 3.53 | 4.07 | 3.26 | 1.47 | | 3.34 | 3.83 |
| 8 | 1998 | 1.63 | 1.38 | 3.87 | 3.34 | 3.84 | 2.82 | 1.49 | | 3.04 | 4.06 |
| 9 | 1999 | 1.72 | 1.52 | 3.85 | 3.42 | 3.87 | 3.27 | 1.79 | | 3.80 | 4.29 |
| 10 | 2000 | 1.94 | 1.86 | 3.80 | 3.45 | 3.77 | 3.65 | 2.01 | | 4.18 | 4.58 |

Đồ thị dạng đường với Matplotlib



VINBIGDATA VINGROUP

Academy
Vietnam

Cú pháp:

plt.plot(x, y, color, linestyle, linewidth, marker, markersize)

Trong đó:

- * X, Y – dữ liệu trục X, Y

Hàm pyplot.plot() còn có các tham số cơ bản sau:

- * Color (c): Màu của đường line
- * Linewidth (lw): Số thực - Độ rộng của đường đồ thị
- * linestyle (ls): Kiểu đường đồ thị
- * marker: Kiểu của điểm
- * markersize (ms): Số thực - Kích thước của điểm dữ liệu

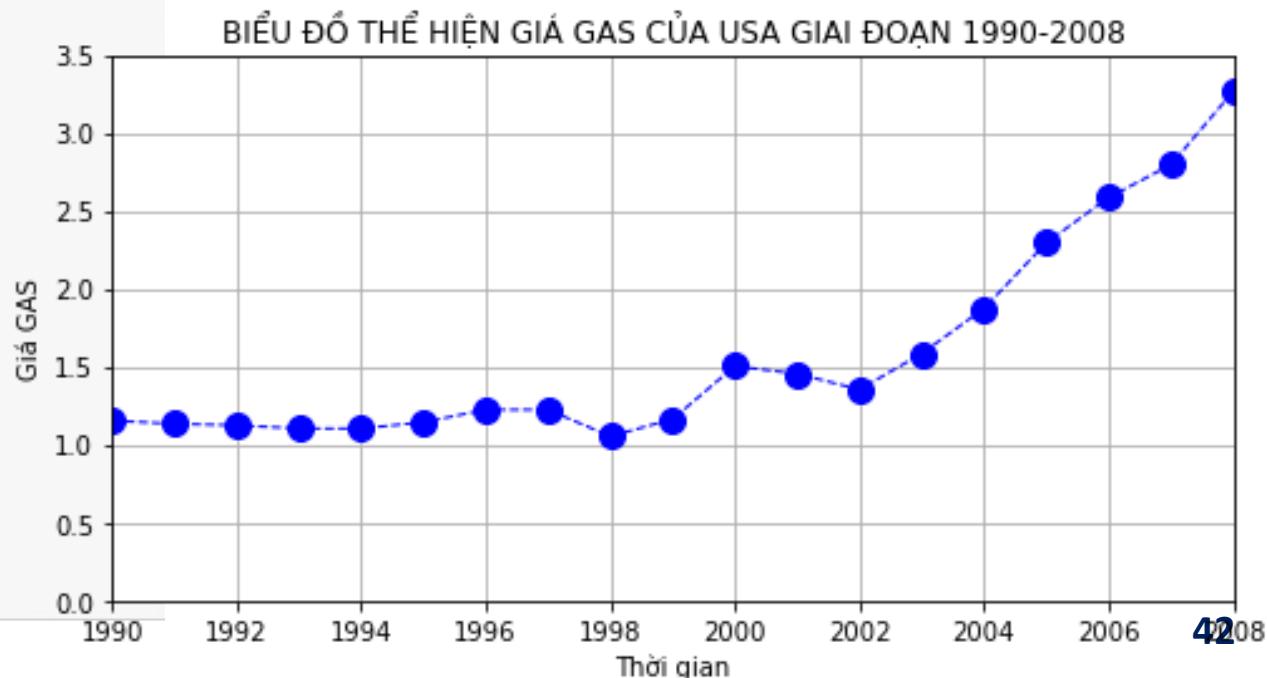
Các tham số color, marker, linestyle có thể được biểu diễn ở dạng '[color][marker][linestyle]', ví dụ: 'ro-' tương đương với color='r', marker='o', linestyle='-'.



a. Single line chart

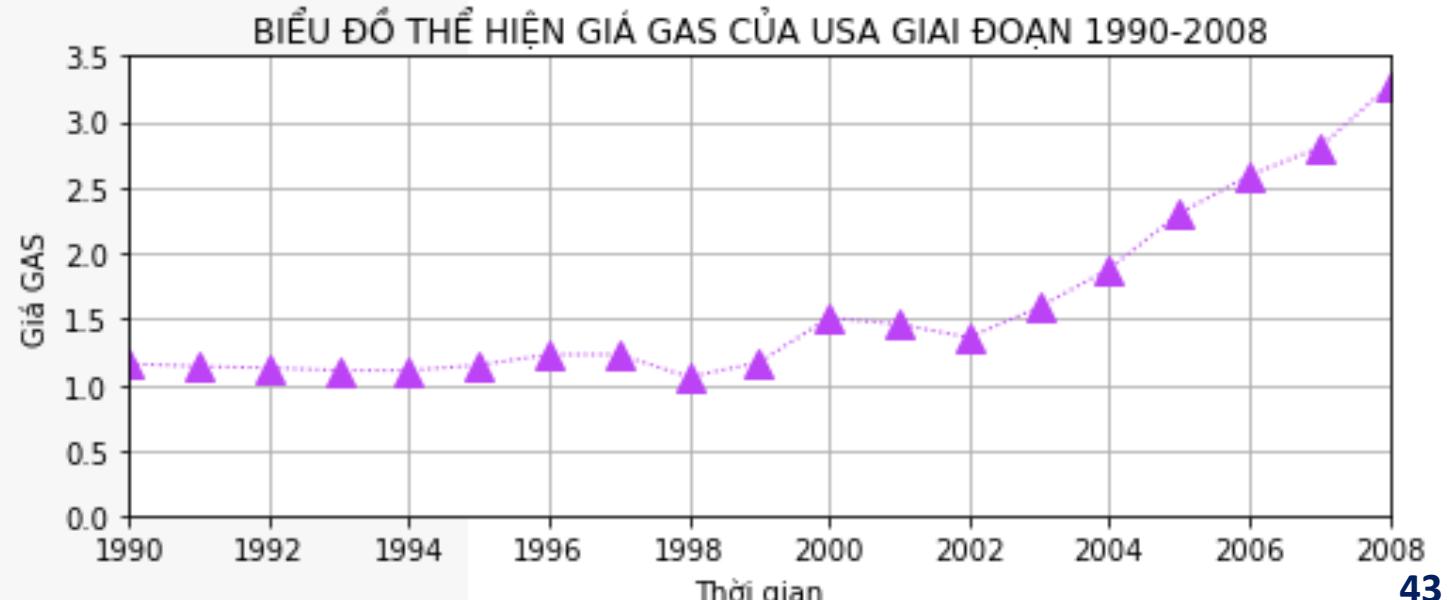
Simple line chart

```
1 plt.figure(figsize = (8,4)) #Thiết lập kích thước biểu đồ
2
3 plt.plot(x,                      #Dữ liệu trục X
4           y,                      #Dữ liệu trục Y
5           color='b',             #Màu của đường
6           linestyle='--',       #Kiểu đường
7           linewidth=1.0,        #Độ rộng của đường line
8           marker='o',           #Kiểu điểm
9           markersize = 10)      #Kích thước điểm
10
11 #Tiêu đề của đồ thị
12 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CỦA USA GIAI ĐOẠN 1990-2008')
13 #Nhãn cho trục X
14 plt.xlabel('Thời gian')
15 #Nhãn cho trục Y
16 plt.ylabel('Giá GAS')
17
18 #Setup giới hạn cho trục X:
19 plt.xlim(1990,2008)
20
21 #Setup giới hạn cho trục Y:
22 plt.ylim(0,3.5)
23
24 #Hiển thị lưới:
25 plt.grid()
26
27 plt.show()
```



Simple line chart

```
1 plt.figure(figsize = (8,3)) #Thiết lập kích thước biểu đồ
2
3 plt.plot(x,                      #Dữ liệu trục X
4           y,                      #Dữ liệu trục Y
5           c='#bc42f5',            #Màu của đường
6           ls=':',                #Kiểu đường
7           lw=1.0,                #Độ rộng của đường line
8           marker='^',            #Kiểu điểm
9           ms = 10)               #Kích thước điểm
10
11 #Tiêu đề của đồ thị
12 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CỦA USA GIAI ĐOẠN 1990-2008')
13 #Nhãn cho trục X
14 plt.xlabel('Thời gian')
15 #Nhãn cho trục Y
16 plt.ylabel('Giá GAS')
17
18 #Setup giới hạn cho trục X:
19 plt.xlim(1990,2008)
20
21 #Setup giới hạn cho trục Y:
22 plt.ylim(0,3.5)
23
24 #Hiển thị lưới:
25 plt.grid()
26
27 plt.show()
```



Simple line chart

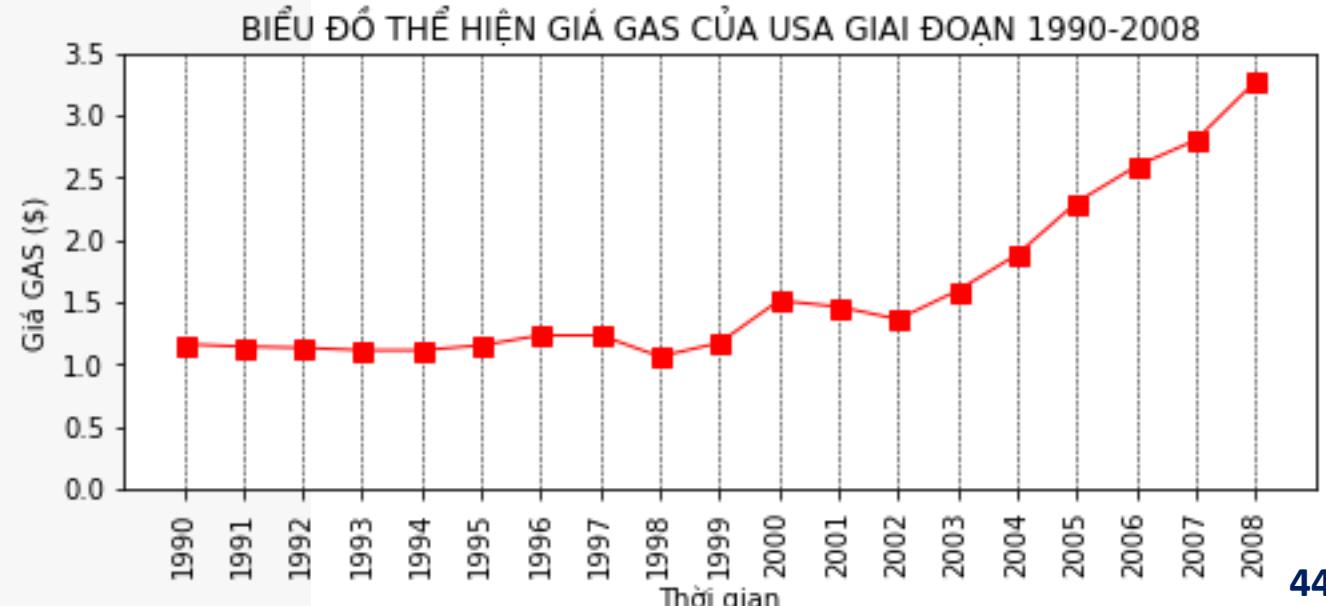


VINBIGDATA VINGROUP

Academy Vietnam

```
1 plt.figure(figsize = (8,3)) #Thiết lập kích thước biểu đồ
2
3 plt.plot(x,                      #Đữ liệu trục X
4           y,                      #Đữ liệu trục Y
5           'r-s',                  #Độ rộng của đường line
6           linewidth=1.0,          #Độ rộng của đường line
7           markersize = 7)        #Kích thước điểm
8
9 #Tiêu đề của đồ thị
10 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CỦA USA GIAI ĐOẠN 1990-2008')
11 #Nhãn cho trục X
12 plt.xlabel('Thời gian')
13 #Nhãn cho trục Y
14 plt.ylabel('Giá GAS ($)')
15 #Setup giới hạn cho trục X:
16 plt.xlim(1989,2009)
17 #Setup giới hạn cho trục Y:
18 plt.ylim(0,3.5)
19 #Setup tick cho trục X:
20 plt.xticks(x,
21             rotation=90)
22 #Thiết lập lưới:
23 plt.grid(axis='x',
24           c='black',
25           ls='--',
26           lw=0.5)
27
28 plt.show()
```

Các tham số *color*, *marker*, *linestyle* có thể
được biểu diễn ở dạng
[color][marker][linestyle],
ví dụ: 'ro-' tương đương với *color='r'*,
marker='o', *linestyle='-'*.



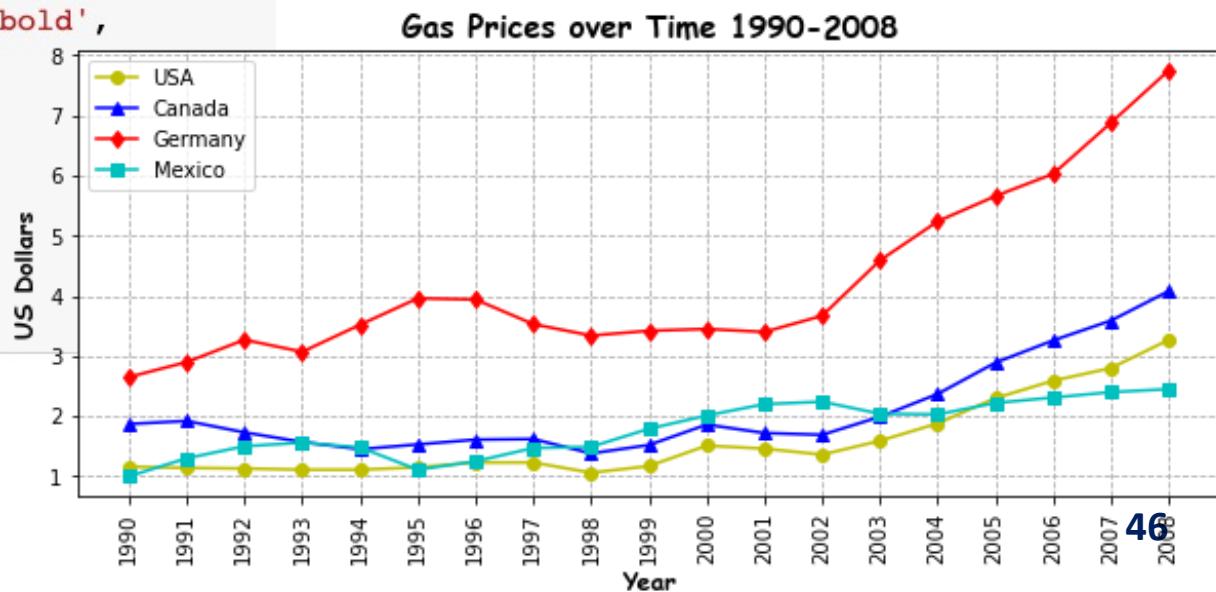


b. Multiple line chart

Multiple lines chart



```
1 plt.figure(figsize=(10,4))
2
3 #Vẽ multiple line:
4 plt.plot(x,y1,'y-o', label='USA')
5 plt.plot(x,y2,'b^-', label='Canada')
6 plt.plot(x,y3,'r-d', label='Germany')
7 plt.plot(x,y4,'c-s', label='Mexico')
8
9
10 plt.title('Gas Prices over Time 1990-2008',fontdict={'fontname':'Comic Sans MS',
11                               'fontweight':'bold',
12                               'fontsize':15})
13 plt.xlabel('Year',fontdict={'fontname':'Comic Sans MS',
14                               'fontweight':'bold',
15                               'fontsize':12})
16 plt.ylabel('US Dollars',fontdict={'fontname':'Comic Sans MS',
17                               'fontweight':'bold',
18                               'fontsize':12})
19 plt.xticks(x,rotation=90)
20 plt.grid(True,ls='--')
21
22 #Hiển thị chú thích trong biểu đồ:
23 plt.legend()
24
25 plt.show()
```



Multiple lines chart

Phương thức legend(): Hiển thị chú thích của biểu đồ Bao gồm các tham số chính:

1.loc: Xác định vị trí hiển thị của chú thích trong biểu đồ, gồm các tùy chọn sau:

- 'best' | 0
- 'upper right' | 1
- 'upper left' | 2
- 'lower left' | 3
- 'lower right' | 4
- 'right' | 5
- 'center left' | 6
- 'center right' | 7
- 'lower center' | 8
- 'upper center' | 9
- 'center' | 10

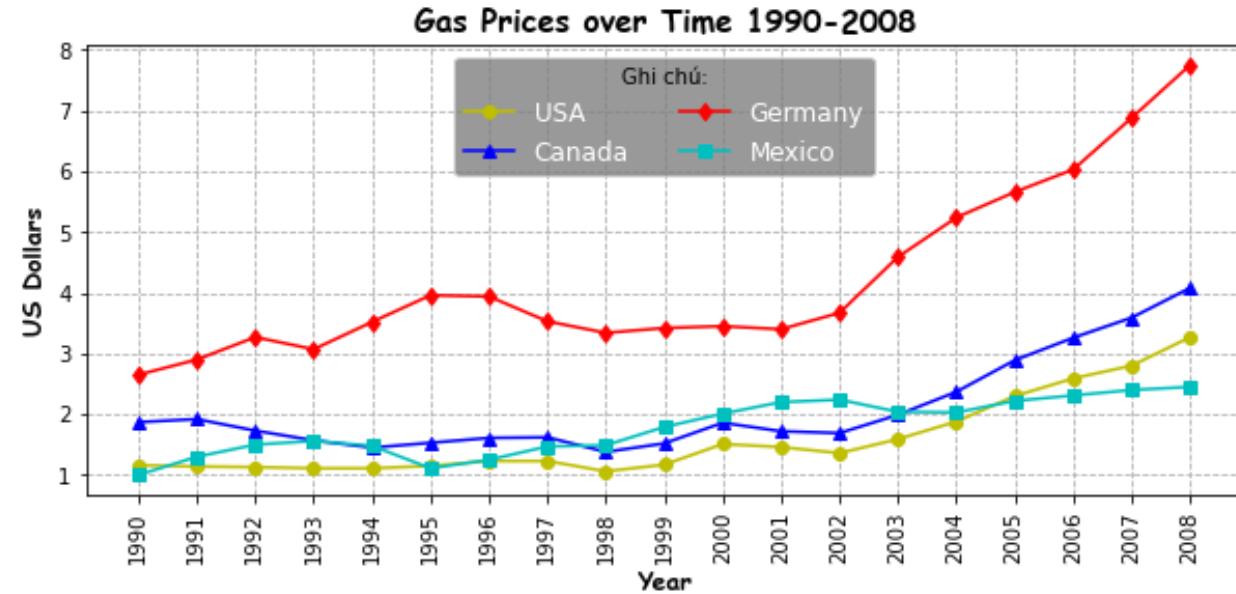
2.ncol: Số cột của chú thích (số nguyên, mặc định là 1)

3.fontsize: kích thước font chữ trong chú thích

4.labelcolor: Màu chữ trong chú thích (mặc định màu đen)

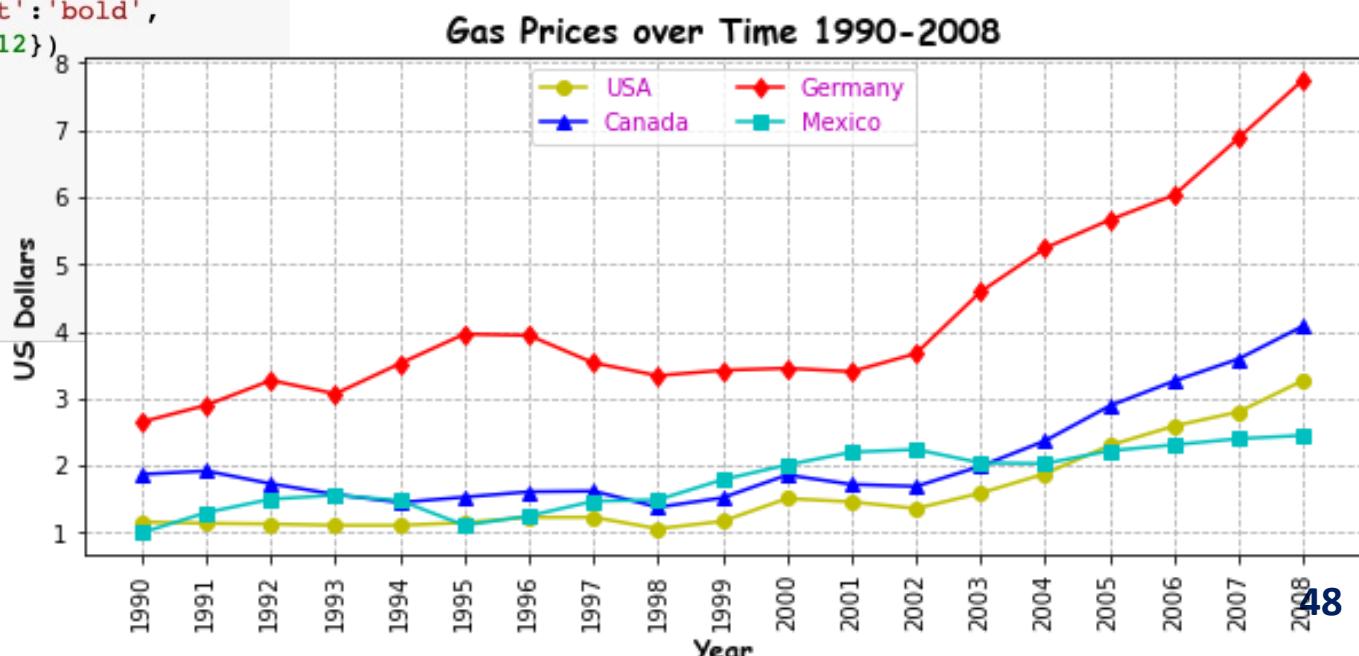
5.facecolor: Màu nền của ô chú thích (mặc định None)

6.title: Dòng tiêu đề trong chú thích



Multiple lines chart

```
1 plt.figure(figsize=(10,4))
2
3 #Vẽ multiple line:
4 plt.plot(x,y1,'y-o', label='USA')
5 plt.plot(x,y2,'b-^', label='Canada')
6 plt.plot(x,y3,'r-d', label='Germany')
7 plt.plot(x,y4,'c-s', label='Mexico')
8
9 plt.title('Gas Prices over Time 1990-2008',fontdict={'fontname':'Comic Sans MS',
10                                         'fontweight':'bold',
11                                         'fontsize':15})
12 plt.xlabel('Year',fontdict={'fontname':'Comic Sans MS',
13                           'fontweight':'bold',
14                           'fontsize':12})
15 plt.ylabel('US Dollars',fontdict={'fontname':'Comic Sans MS',
16                                   'fontweight':'bold',
17                                   'fontsize':12})
18 plt.xticks(x,rotation=90)
19 plt.grid(True,ls='--')
20 plt.legend(loc = 9, ncol=2,labelcolor='m')
21 #Lưu đồ thị:
22 plt.savefig('Save_charts/Gas',dpi=300, format='png')
23 plt.savefig('Save_charts/Gas',dpi=500,format='pdf')
24
25 plt.show()
```



Lưu biểu đồ:

plt.savefig(fname, dpi, format)

Trong đó:

1. fname: đường dẫn lưu file
2. dpi: độ phân giải của đồ thị khi lưu (số pixel điểm ảnh trên mỗi inch)
3. format: định dạng file ('png', 'pdf',...)



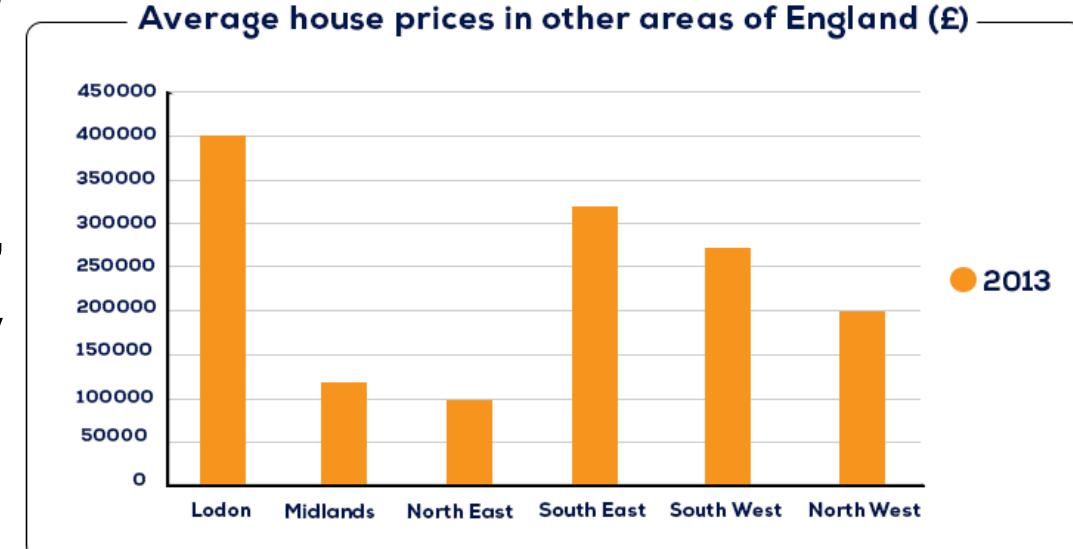
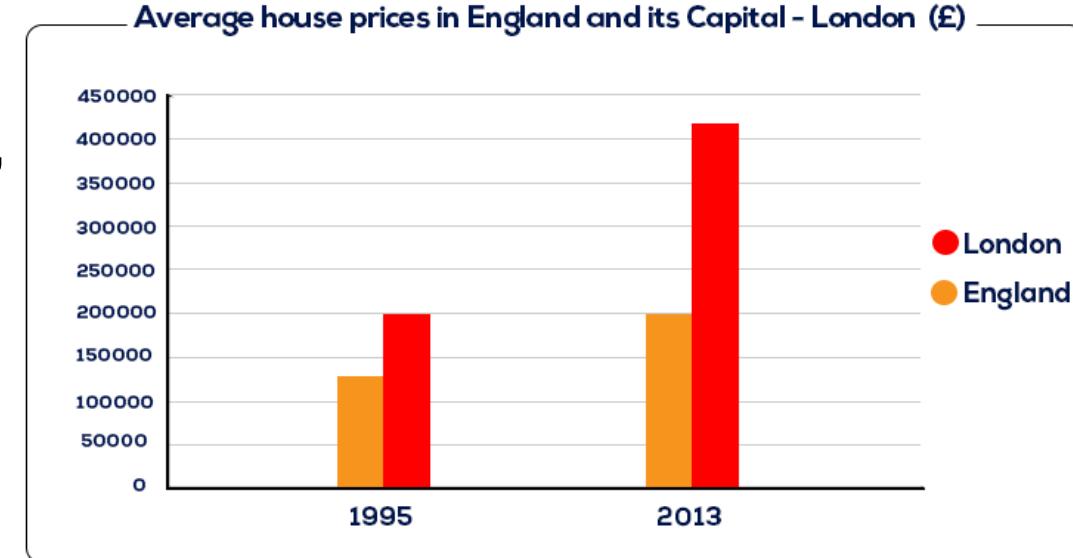
5. Biểu đồ thanh/cột (bar chart)

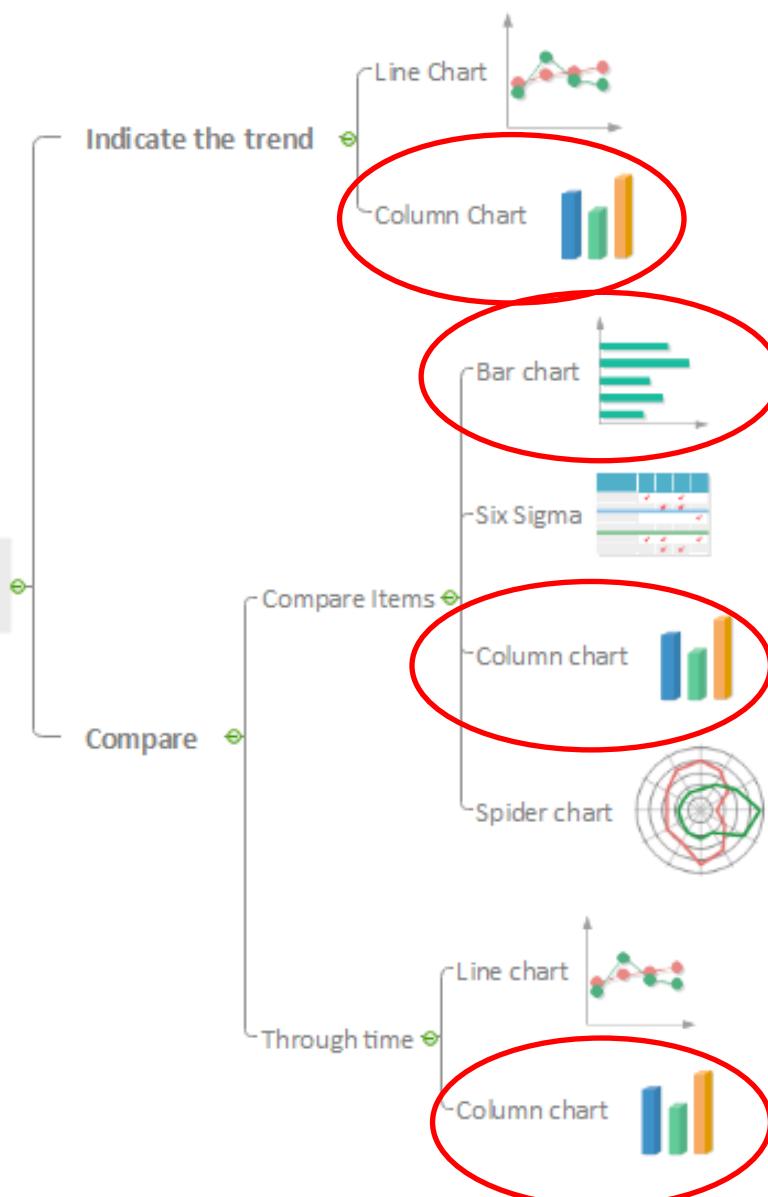
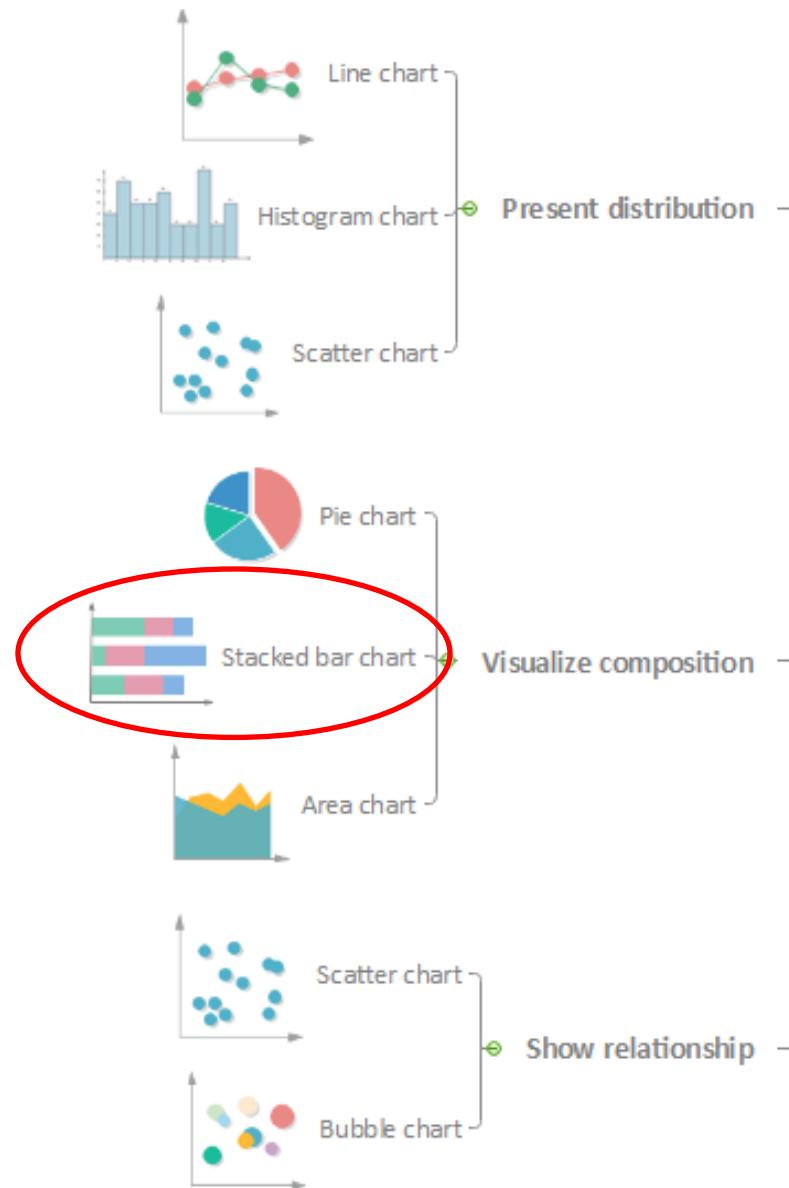
Biểu đồ Bar/Column chart

Bar chart cũng như line chart đây là một trong những dạng đồ thị phổ biến và hay được sử dụng trong thực tế.

Bar chart là một cách cụ thể để biểu diễn dữ liệu bằng cách sử dụng các thanh hình chữ nhật, trong đó chiều dài của mỗi thanh tỷ lệ với giá trị mà chúng đại diện. Nó là một cách biểu diễn đồ họa của dữ liệu bằng cách sử dụng các thanh có độ cao khác nhau.

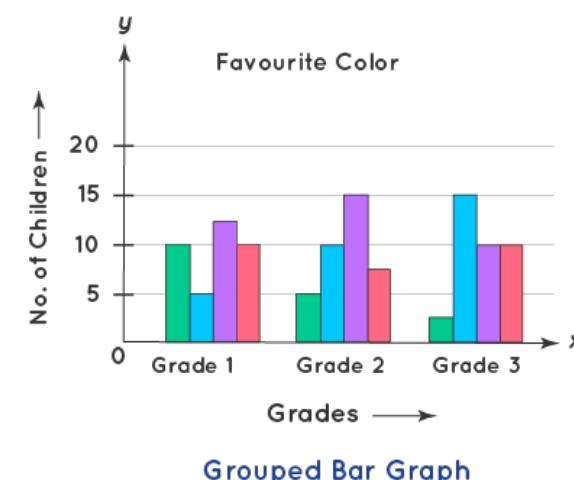
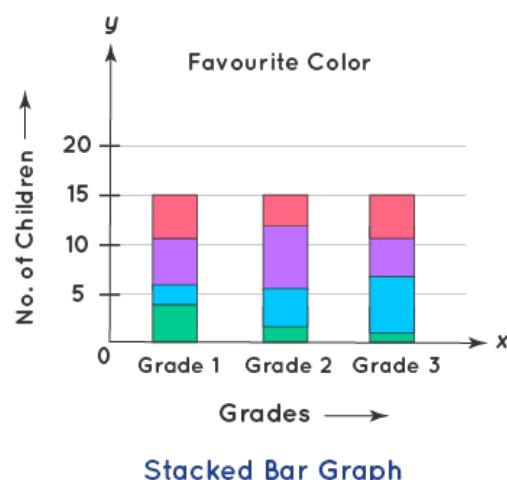
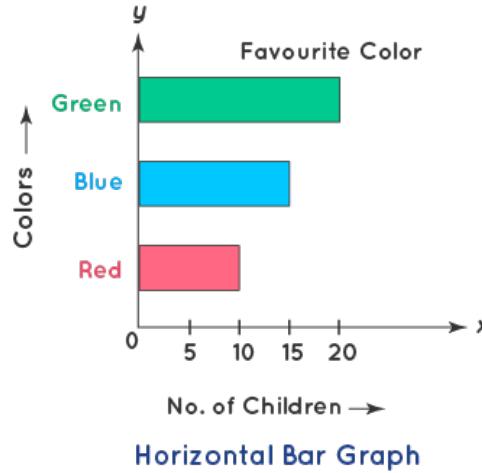
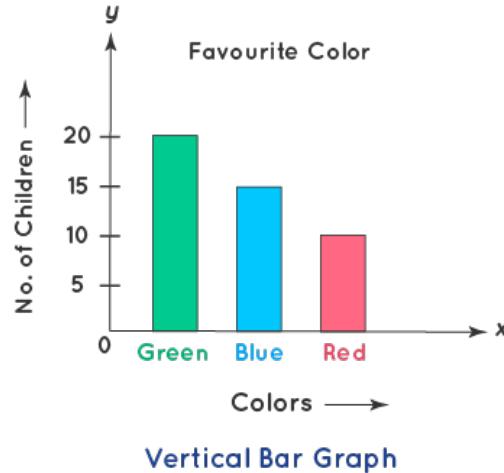
Bar chart là công cụ tuyệt vời để biểu diễn dữ liệu độc lập với nhau mà không cần theo bất kỳ thứ tự cụ thể nào khi biểu diễn.





Biểu đồ Bar/Column chart

Types of Bar Graph



Đặc điểm của Bar chart:

- Tất cả các thanh chữ nhật phải có chiều rộng bằng nhau và phải có khoảng trống giữa chúng bằng nhau.
- Các thanh hình chữ nhật có thể vẽ theo chiều ngang hoặc chiều dọc.
- Chiều cao của hình chữ nhật tương đương với giá trị của dữ liệu mà chúng đại diện.
- Các thanh hình chữ nhật phải nằm trên cùng một trực cơ sở.

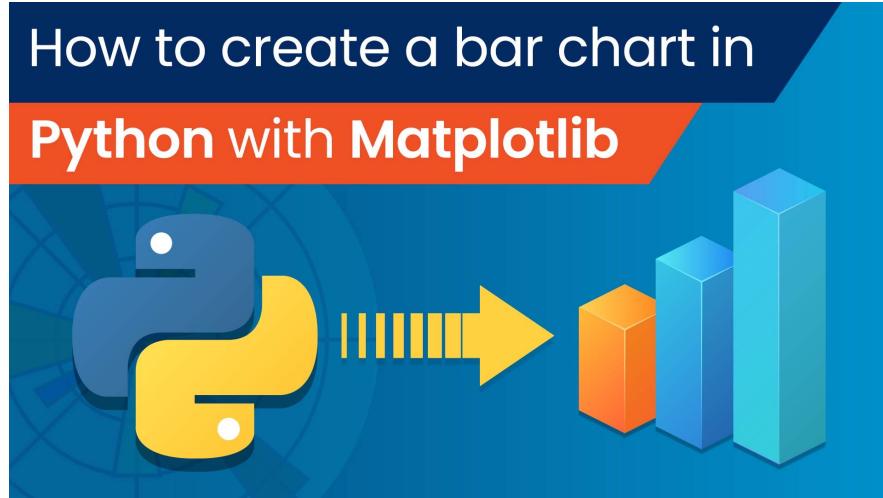
Biểu đồ Bar chart với Matplotlib



VINBIGDATA VINGROUP

Academy
Vietnam

Tập dữ liệu `gas_prices.csv`: Lưu trữ giá Gas của 10 nước trên thế giới trong
giai đoạn từ năm 1990 - 2008



```
1 data = pd.read_csv('Data_Visualize/gas_prices.csv')  
2 data
```

| | Year | Australia | Canada | France | Germany | Italy | Japan | Mexico | South Korea | UK | USA |
|----|------|-----------|--------|--------|---------|-------|-------|--------|-------------|------|------|
| 0 | 1990 | NaN | 1.87 | 3.63 | 2.65 | 4.59 | 3.16 | 1.00 | | 2.05 | 2.82 |
| 1 | 1991 | 1.96 | 1.92 | 3.45 | 2.90 | 4.50 | 3.46 | 1.30 | | 2.49 | 3.01 |
| 2 | 1992 | 1.89 | 1.73 | 3.56 | 3.27 | 4.53 | 3.58 | 1.50 | | 2.65 | 3.06 |
| 3 | 1993 | 1.73 | 1.57 | 3.41 | 3.07 | 3.68 | 4.16 | 1.56 | | 2.88 | 2.84 |
| 4 | 1994 | 1.84 | 1.45 | 3.59 | 3.52 | 3.70 | 4.36 | 1.48 | | 2.87 | 2.99 |
| 5 | 1995 | 1.95 | 1.53 | 4.26 | 3.96 | 4.00 | 4.43 | 1.11 | | 2.94 | 3.21 |
| 6 | 1996 | 2.12 | 1.61 | 4.41 | 3.94 | 4.39 | 3.64 | 1.25 | | 3.18 | 3.34 |
| 7 | 1997 | 2.05 | 1.62 | 4.00 | 3.53 | 4.07 | 3.26 | 1.47 | | 3.34 | 3.83 |
| 8 | 1998 | 1.63 | 1.38 | 3.87 | 3.34 | 3.84 | 2.82 | 1.49 | | 3.04 | 4.06 |
| 9 | 1999 | 1.72 | 1.52 | 3.85 | 3.42 | 3.87 | 3.27 | 1.79 | | 3.80 | 4.29 |
| 10 | 2000 | 1.94 | 1.86 | 3.80 | 3.45 | 3.77 | 3.65 | 2.01 | | 4.18 | 4.58 |
| | | | | | | | | | | | 1.51 |

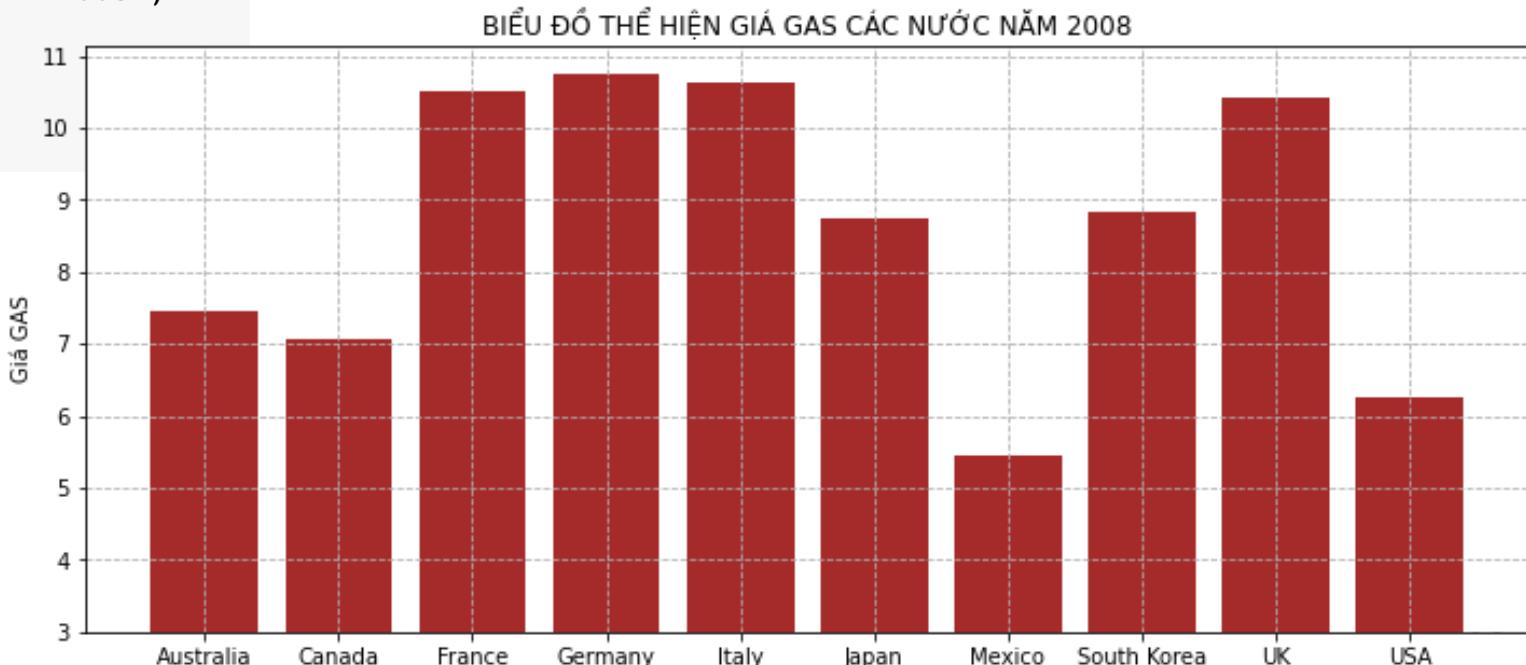


a. Vertical Bar chart

Biểu đồ cột dạng thẳng đứng

Cú pháp: plt.bar (labels, y)

```
1 plt.figure(figsize = (12,5)) #Thiết lập kích thước biểu đồ
2
3 #Vẽ biểu đồ cột:
4 plt.bar(labels,           #Nhãn của trục X
5         y_2008,          #Giá trị tương ứng với nhãn
6         color='brown',   #Màu của thanh
7         bottom=3,        #Giá trị bắt đầu của trục Y
8         width = 0.8)     #Chiều rộng của thanh
9
10 #Tiêu đề của đồ thị
11 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CÁC NƯỚC NĂM 2008')
12 plt.ylabel('Giá GAS')
13 plt.grid(ls='--')
14
15 plt.show()
```



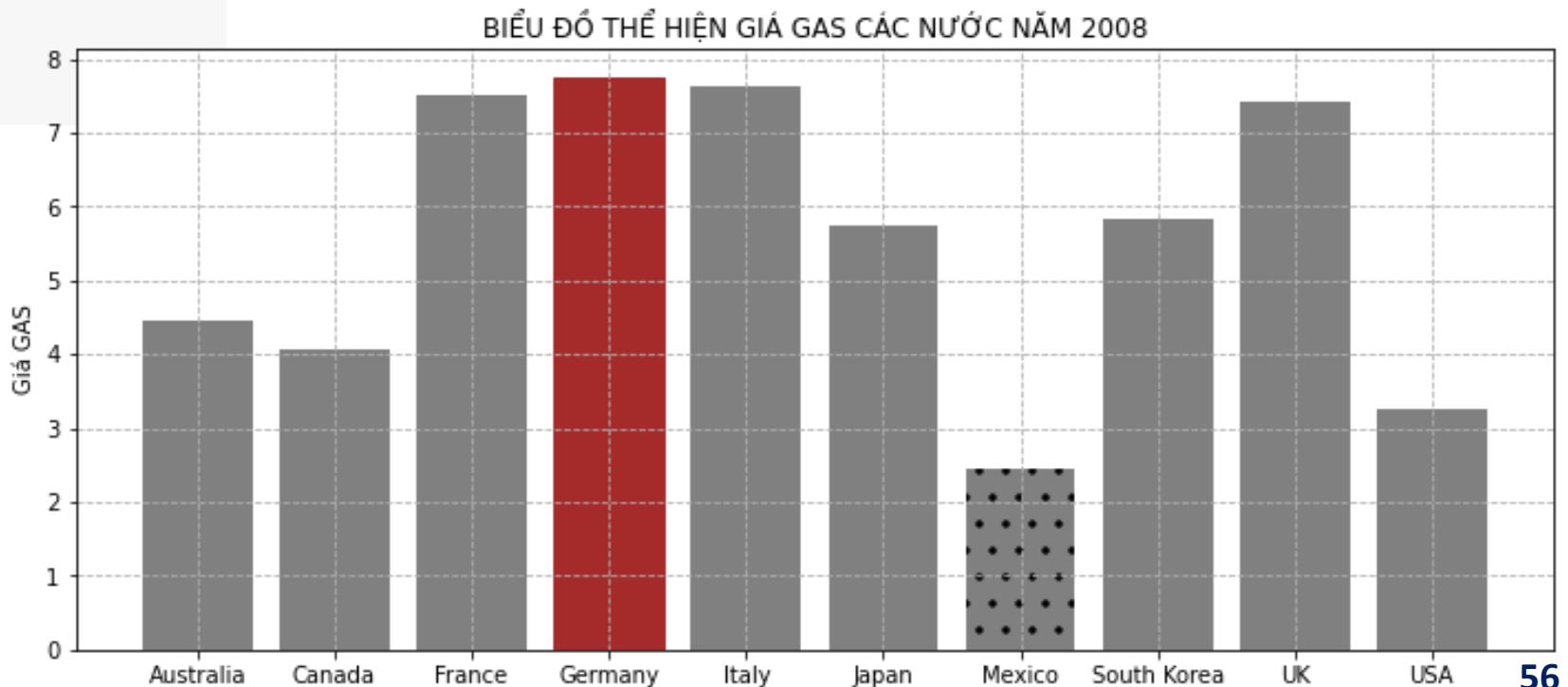
Biểu đồ cột dạng thẳng đứng



VINBIGDATA VINGROUP

Academy Vietnam

```
1 #Làm nổi bật một thanh:  
2 plt.figure(figsize = (12,5))  
3  
4 bar = plt.bar(labels,y_2008,color='gray')  
5  
6 #Thay đổi màu sắc khác, tạo hatch cho thanh  
7 bar[3].set_color('brown')  
8 bar[6].set_hatch('.')  
9  
10 #Tiêu đề của đồ thị  
11 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CÁC NƯỚC NĂM 2008')  
12 plt.ylabel('Giá GAS')  
13 plt.grid(ls='--')  
14  
15 plt.show()
```





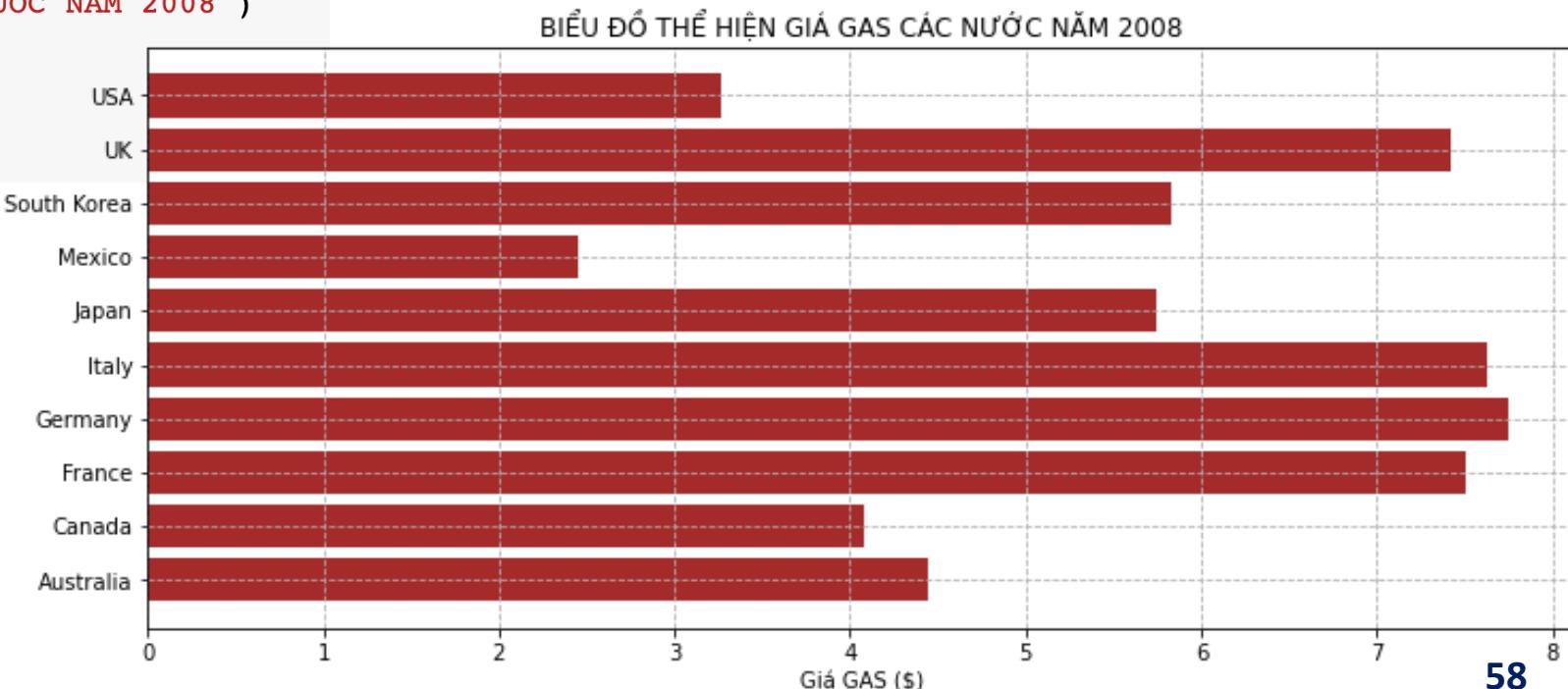
b. Horizontal Bar chart



Biểu đồ cột dạng nằm ngang

Cú pháp: plt.barh (labels, y)

```
1 plt.figure(figsize = (12,5)) #Thiết lập kích thước biểu đồ
2
3 plt.barh(labels,
4          y_2008,
5          color='brown',
6          height = 0.8) #Chiều rộng của thanh
7
8 #Tiêu đề của đồ thị
9 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CÁC NƯỚC NĂM 2008')
10 plt.xlabel('Giá GAS ($)')
11 plt.grid(ls='--')
12
13 plt.show()
```

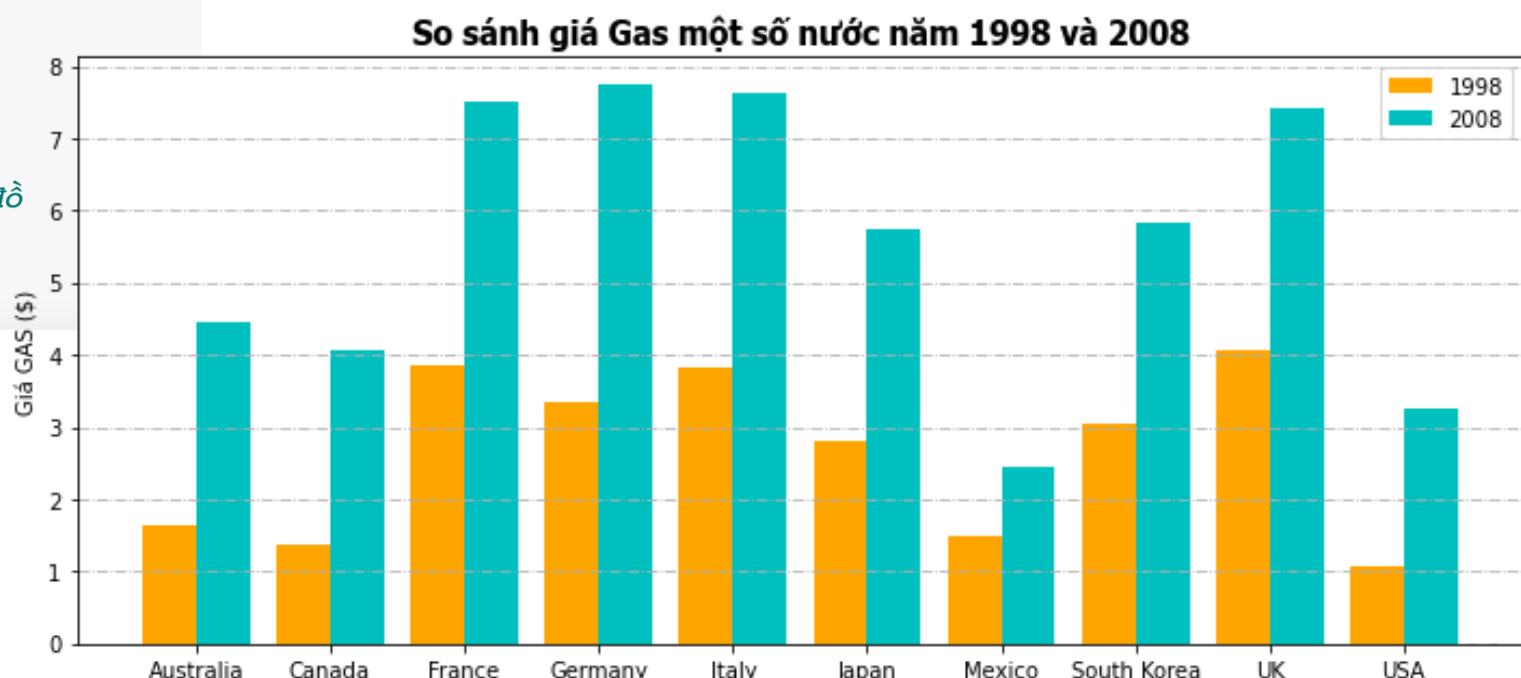




c. Grouped Bar chart

Nhiều cột trên biểu đồ

```
1 #Multiple bar:  
2 w=0.4 #Thiết lập độ rộng của thanh (Tổng < 1.0)  
3 bar1 = np.arange(len(labels))  
4 bar2 = [i+w for i in bar1]  
5  
6 plt.figure(figsize = (12,5))  
7 #Vẽ các biểu đồ cột cho từng bộ dữ liệu:  
8 plt.bar(bar1,y_1998,width=w,color='orange',label='1998')  
9 plt.bar(bar2,y_2008,width=w,color='c',label='2008')  
10  
11 plt.title('So sánh giá Gas một số nước năm 1998 và 2008',  
12     fontdict={'fontname':'Tahoma',  
13     'fontweight':'bold',  
14     'fontsize':15})  
15 plt.ylabel("Giá GAS ($)")  
16 plt.grid(axis='y',ls='-.')  
17 plt.legend()  
18  
19 #Hiển thị nhãn của trục x, căn vào giữa 2 biểu đồ  
20 plt.xticks(bar1+w/2,labels)  
21  
22 plt.show()
```



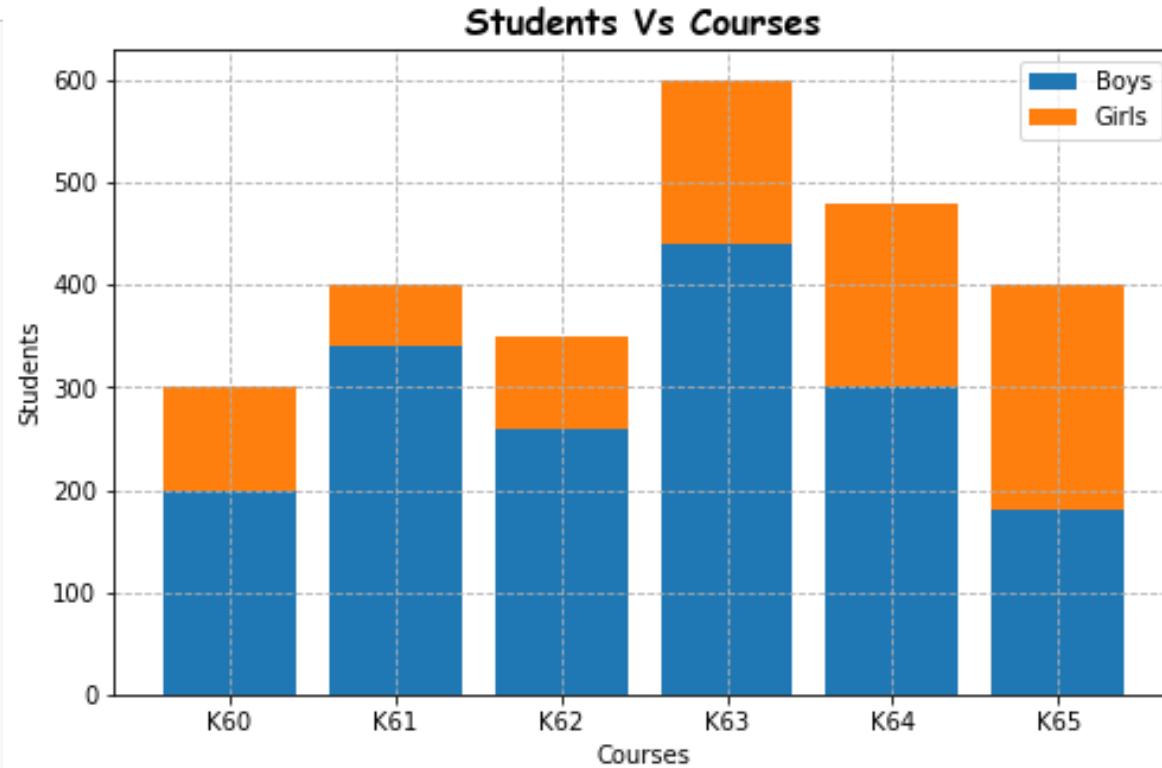


d. Stacked Bar chart

Biểu đồ cột xếp tầng

```
1 #Tạo dữ liệu: Số lượng SV Nam - Nữ theo từng khóa.  
2 labels = ['K60', 'K61', 'K62', 'K63', 'K64', 'K65']  
3 boys = [200, 340, 260, 440, 300, 180]  
4 girls = [100, 60, 90, 160, 180, 220]
```

```
1 plt.figure(figsize=(8,5))  
2  
3 #Biểu đồ cột 1 bên dưới cùng:  
4 plt.bar(labels,boys,label='Boys')  
5  
6 #Biểu đồ cột 2 xếp chồng lên biểu đồ 1:  
7 #Sử dụng: thuộc tính bottom chồng biểu đồ:  
8 plt.bar(labels,girls,bottom = boys, label='Girls')  
9  
10 plt.title('Students Vs Courses',  
11             fontdict={ 'fontname' : 'Comic Sans MS',  
12                     'fontweight' : 'bold',  
13                     'fontsize' : 15})  
14 plt.xlabel('Courses')  
15 plt.ylabel("Students")  
16 plt.grid(ls='--')  
17 plt.legend()  
18  
19 plt.show()
```



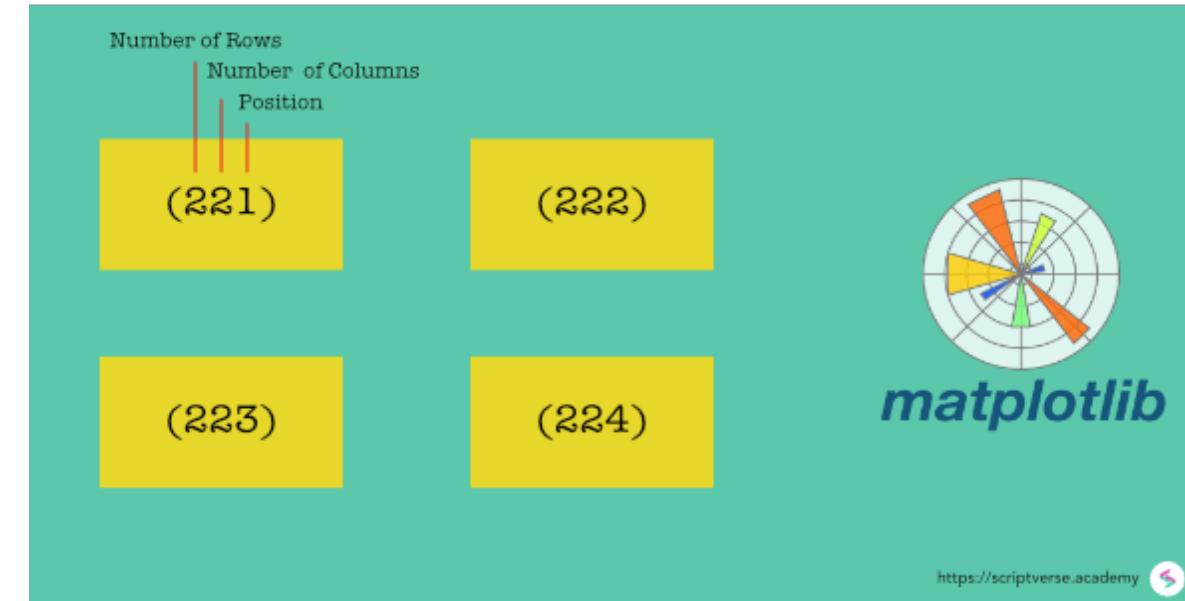
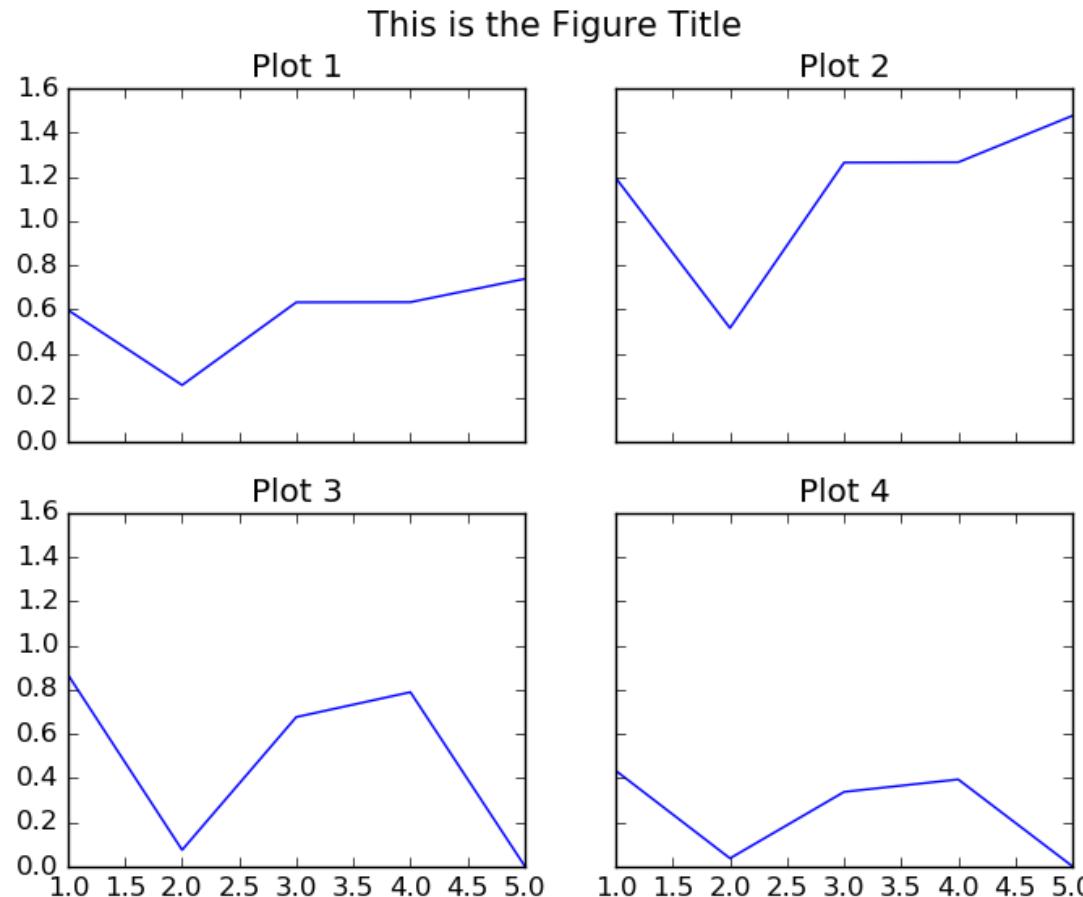


d. Hiển thị nhiều khung biểu đồ



Multiple plot

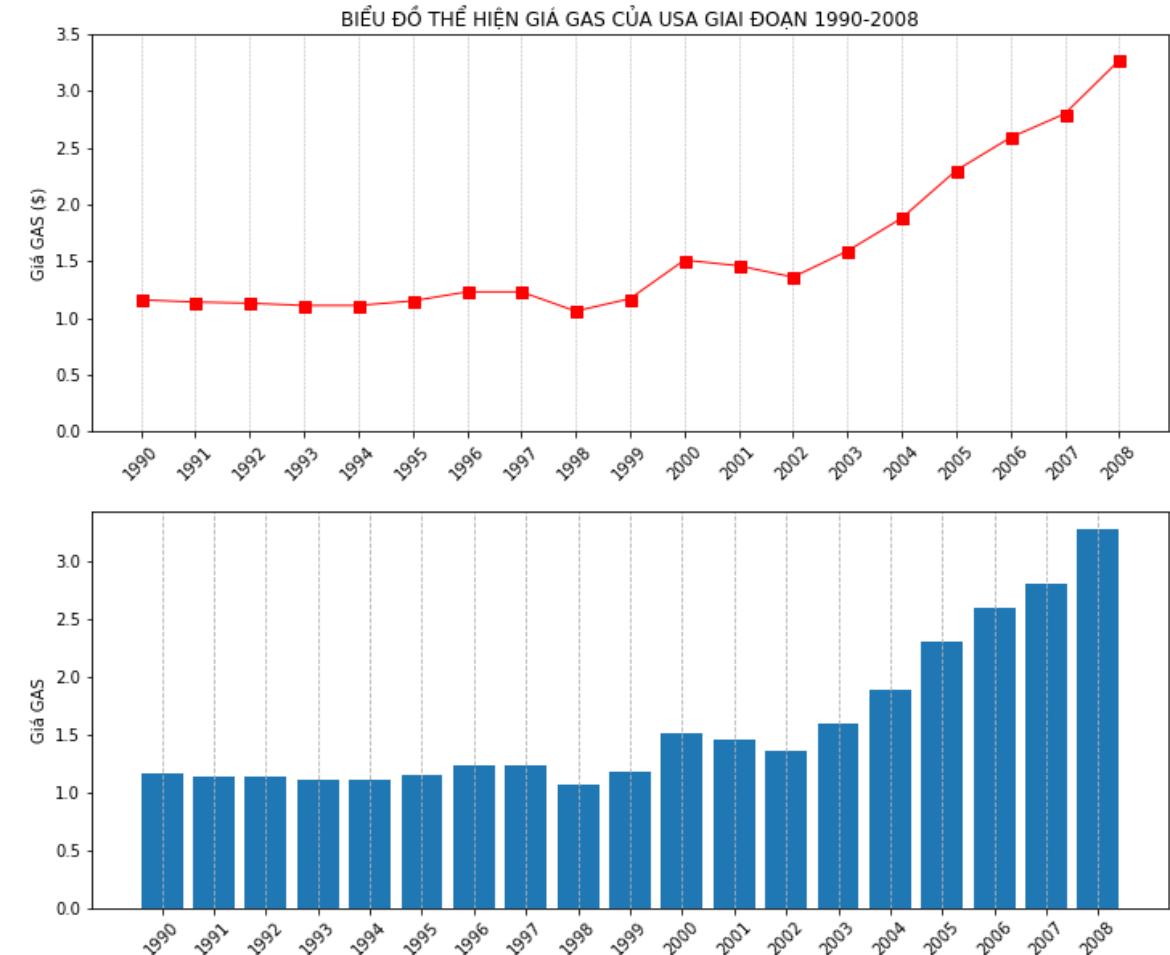
Cú pháp: plt.subplot (nrows, ncols,Position)



Multiple plot

plt.subplot (2, 1, Position)

```
1 plt.figure(figsize = (12,10)) #Thiết lập kích thước biểu đồ
2 #Vẽ biểu đồ đường trên khung 1:
3 plt.subplot(2,1,1) #Thiết lập Khung biểu đồ gồm 2 hàng 1 cột
4
5 plt.plot(x, y,'r-s',lw=1.0, ms=7) #Vẽ biểu đồ đường plot 1
6
7 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CỦA USA GIAI ĐOẠN 1990-2008')
8 plt.ylabel('Giá GAS ($)')
9 plt.ylim(0,3.5)
10 plt.xticks(x,rotation=45)
11 plt.grid(axis='x',ls='--')
12
#-----#
13 #Vẽ biểu đồ Bar trên khung 2:
14 plt.subplot(2,1,2)
15
16 plt.bar(x,y) #Vẽ biểu đồ cột plot 2
17
18 plt.ylabel('Giá GAS')
19 plt.xticks(x,rotation=45)
20 plt.grid(axis='x', ls='--')
21
22 plt.show()
```



Multiple plot

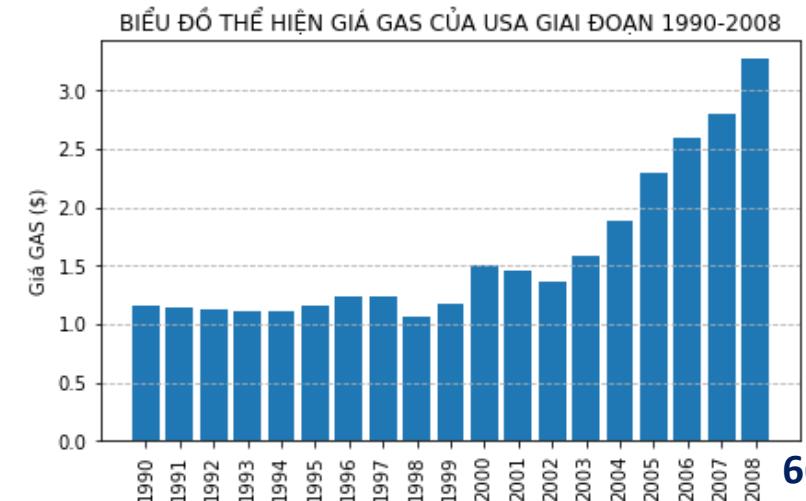
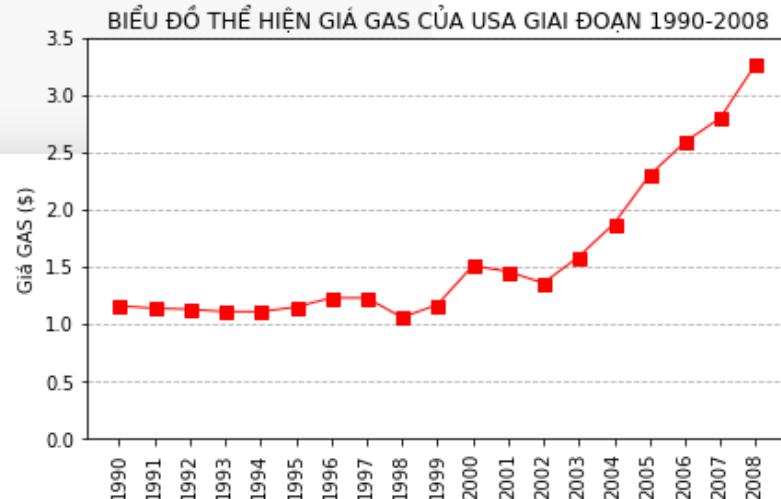


VINBIGDATA
VINGROUP

Academy
Vietnam

```
1 plt.figure(figsize = (15,4))
2 #Vẽ biểu đồ đường trên khung 1:
3 plt.subplot(1,2,1) #Thiết lập Khung biểu đồ gồm 1 hàng 2 cột
4
5 plt.plot(x, y, 'r-s', lw=1.0, ms = 7) #Vẽ biểu đồ đường plot 1
6
7 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CỦA USA GIAI ĐOẠN 1990-2008')
8 plt.ylabel('Giá GAS ($)')
9 plt.ylim(0,3.5)
10 plt.xticks(x,rotation=90)
11 plt.grid(axis='y',ls='--')
12 #
13 #Vẽ biểu đồ Bar trên khung 2:
14 plt.subplot(1,2,2)
15
16 plt.bar(x,y) #Vẽ biểu đồ cột plot 2
17
18 plt.title('BIỂU ĐỒ THỂ HIỆN GIÁ GAS CỦA USA GIAI ĐOẠN 1990-2008')
19 plt.ylabel('Giá GAS ($)')
20 plt.xticks(x,rotation=90)
21 plt.grid(axis='y', ls='--')
22
23 plt.show()
```

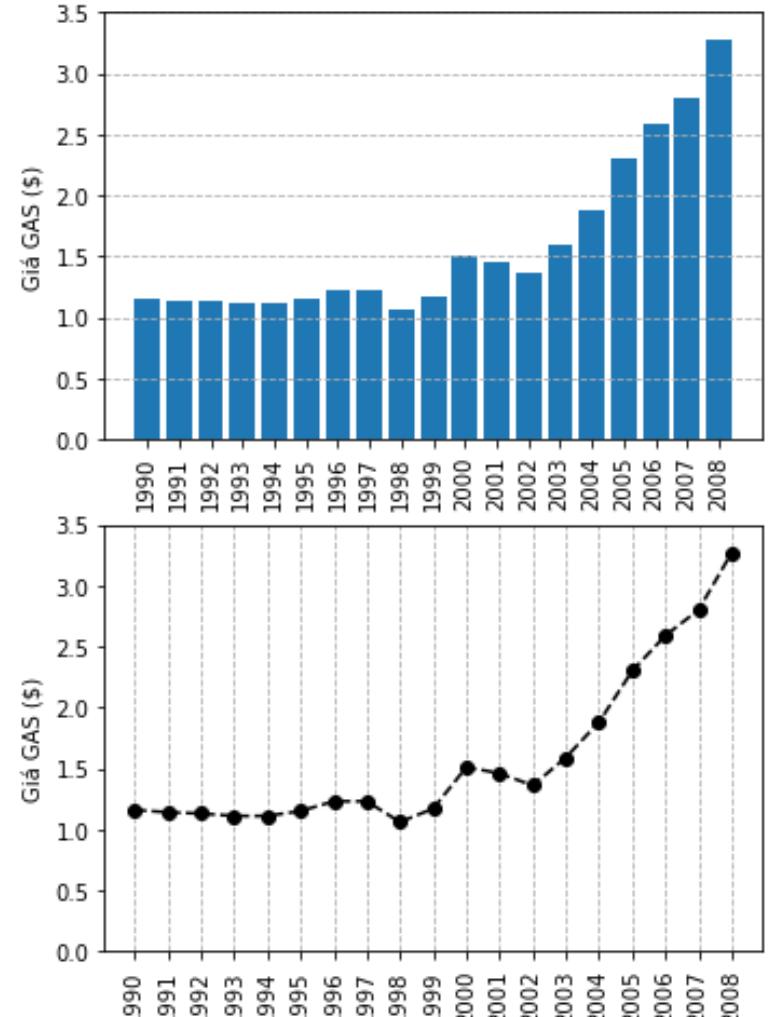
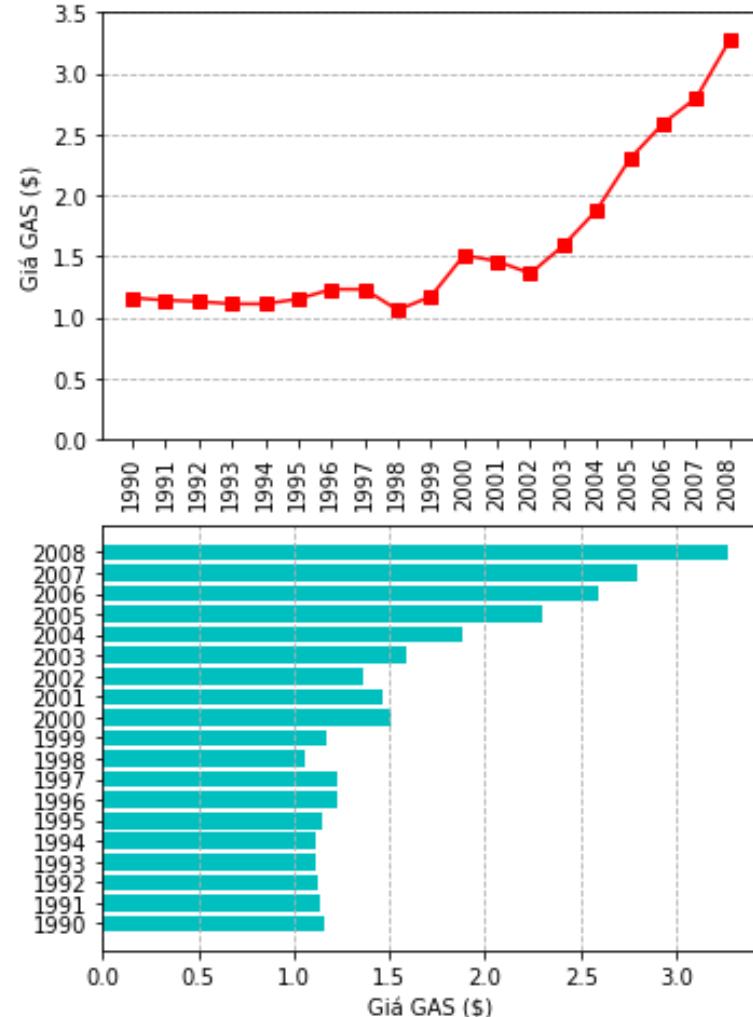
plt.subplot (1, 2, Position)



Multiple plot

```
1 plt.figure(figsize = (12,8))
2
3 #Thiết lập Khung biểu đồ gồm 2 hàng 2 cột
4 #####Vẽ biểu đồ đường trên khung 1#####
5 plt.subplot(2,2,1)
6 #Vẽ biểu đồ đường trên plot 1:
7 plt.plot(x, y, 'r-s')
8
9 plt.ylabel('Giá GAS ($)')
10 plt.ylim(0,3.5)
11 plt.xticks(x,rotation=90)
12 plt.grid(axis='y',ls='--')
13 #####Vẽ biểu đồ đường trên khung 2#####
14 plt.subplot(2,2,2)
15 #Vẽ biểu đồ cột đứng trên plot 2:
16 plt.bar(x, y)
17 plt.ylabel('Giá GAS ($)')
18 plt.ylim(0,3.5)
19 plt.xticks(x,rotation=90)
20 plt.grid(axis='y',ls='--')
21 #####Vẽ biểu đồ đường trên khung 3#####
22 plt.subplot(2,2,3)
23 #Vẽ biểu đồ cột ngang trên plot 3:
24 plt.barch(x,y,color='c')
25 plt.xlabel('Giá GAS ($)')
26 plt.yticks(x)
27 plt.grid(axis='x',ls='--')
28 #####Vẽ biểu đồ đường trên khung 4#####
29 plt.subplot(2,2,4)
30 #Vẽ biểu đồ đường trên plot 4:
31 plt.plot(x, y, 'k--o')
32 plt.ylabel('Giá GAS ($)')
33 plt.ylim(0,3.5)
34 plt.xticks(x,rotation=90)
35 plt.grid(axis='x',ls='--')
36
37 plt.show()
```

plt.subplot (2, 2, Position)



6. Biểu đồ tròn (hình bánh)

Biểu đồ tròn (Pie chart)

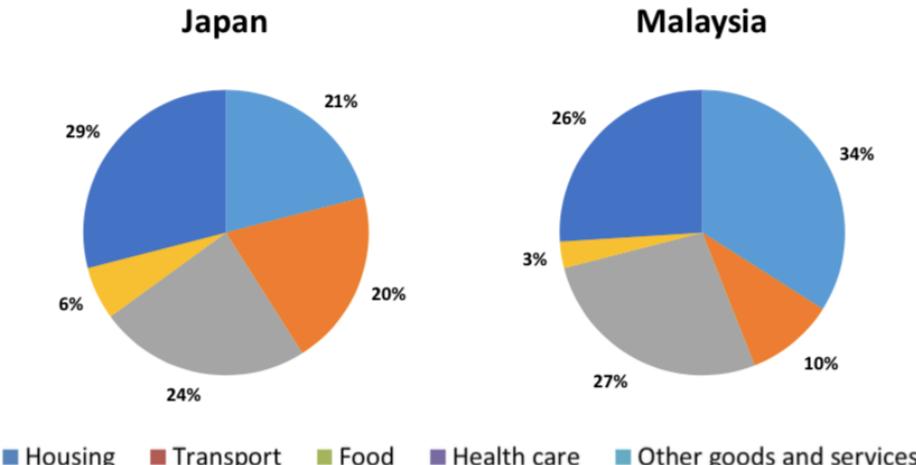
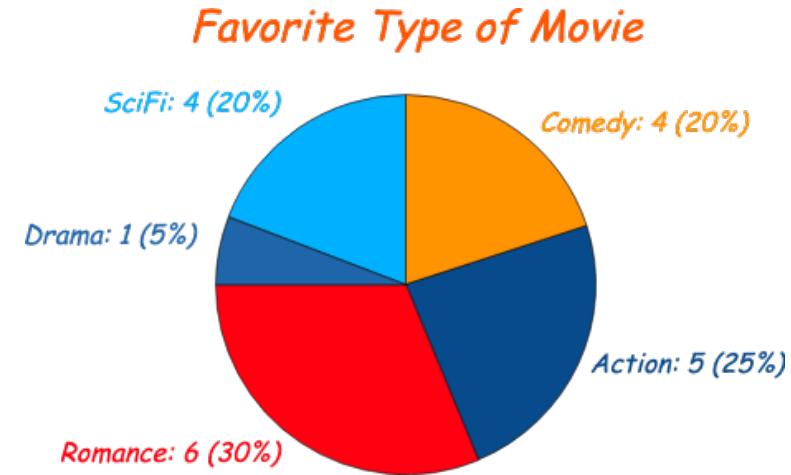


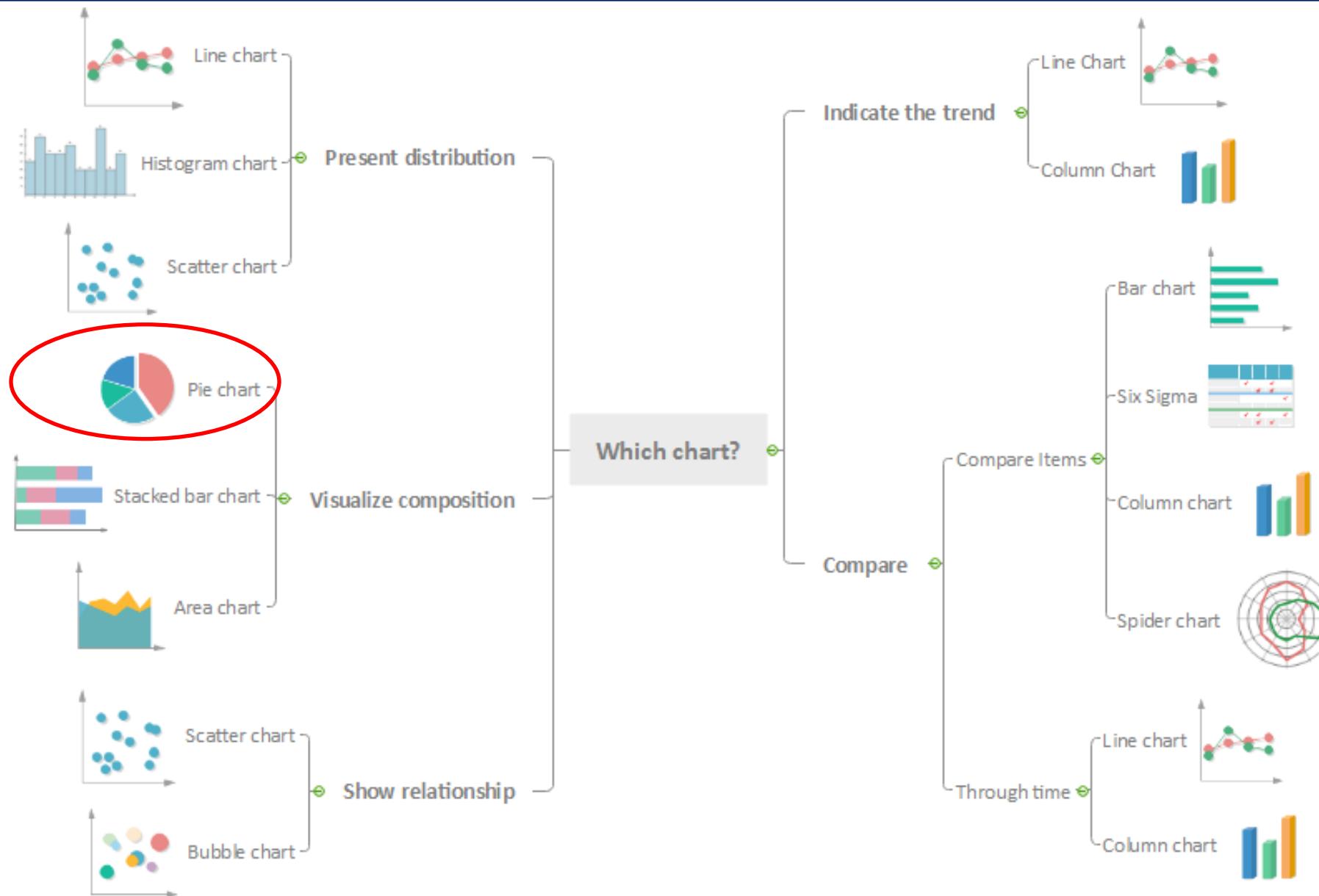
VINBIGDATA



Academy
Vietnam

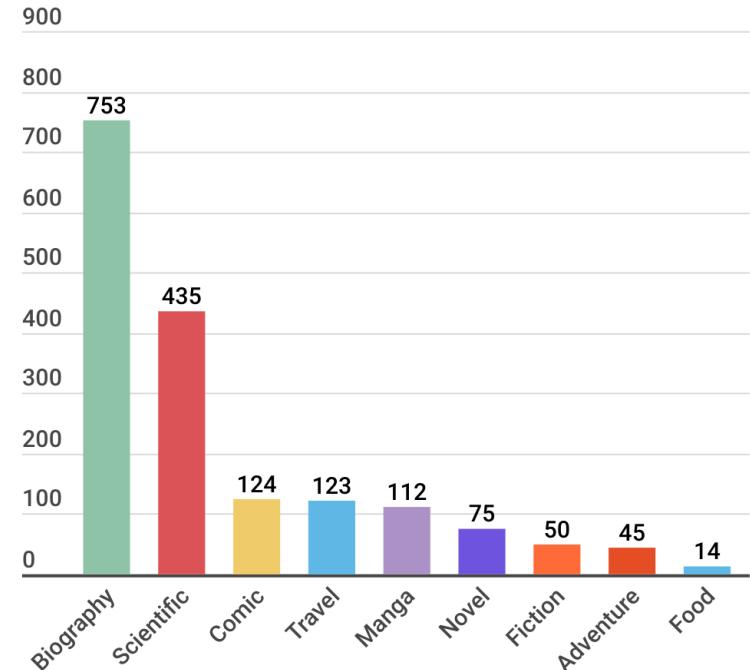
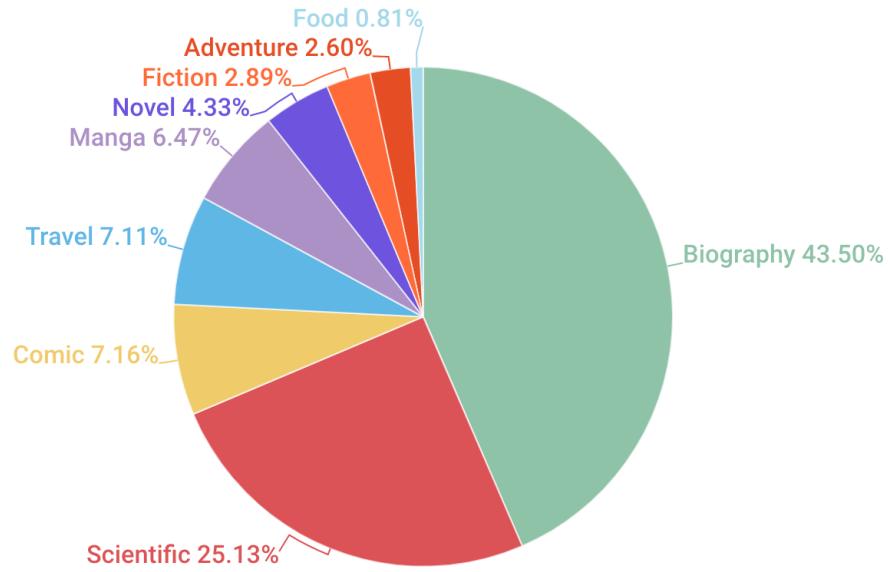
- **Biểu đồ tròn (Pie Chart)** là dạng biểu đồ hình tròn phẳng (cũng có tình huống được trình bày ở dạng 3D) dùng để so sánh giá trị phần trăm trong tổng thể.
- Các giá trị biểu diễn số liệu cho một đối tượng thông qua màu sắc riêng biệt. Đối tượng nào có màu sắc tương ứng đó và được liệt kê ở phần chú thích của biểu đồ. Phần màu càng lớn thì số liệu càng lớn và ngược lại.
- Pie Chart được sử dụng để biểu diễn tỉ lệ phần trăm của các thành phần so với tổng thể. Vì vậy, nó không được dùng để biểu diễn giá trị chính xác.





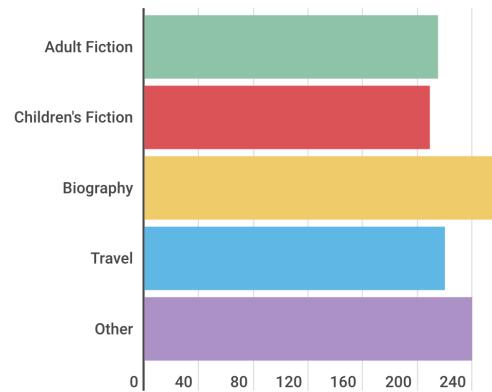
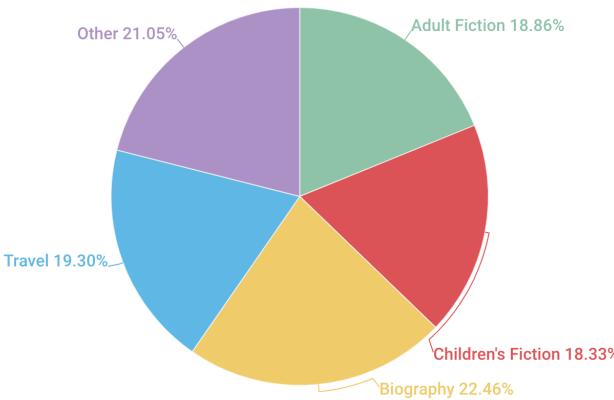
Lưu ý khi sử dụng Pie chart

- Đảm bảo tổng các thành phần là 100%:** Với các công cụ hỗ trợ thì không cần lo lắng về lỗi này vì các công cụ đã đảm bảo được sự chính xác của số liệu khi biểu diễn. Nếu vẽ Pie chart thủ công thì chúng ta cần kiểm tra lại tính đúng đắn một lần nữa.
- Chỉ dùng Pie chart khi số lượng thể loại ít hơn 6:** Việc sử dụng Pie Chart khi có quá nhiều thể loại sẽ khiến cho biểu đồ khá rối. Nếu có quá nhiều thể loại, nên xem xét một biểu đồ khác như Bar Chart.

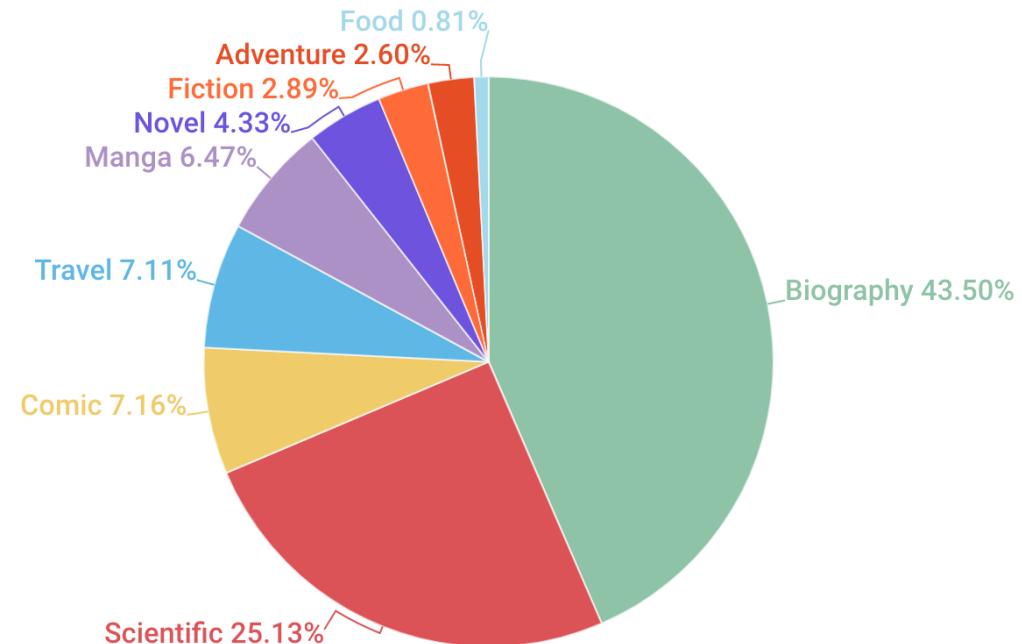


Lưu ý khi sử dụng Pie chart

3. Không dùng Pie Chart nếu tỉ lệ giữa các thể loại gần tương đương nhau: Nếu tỉ lệ giữa các thể loại là tương đương nhau thì dường như Pie Chart lúc này là vô dụng vì không thể hiện cụ thể một ý nghĩa gì. Giải pháp lúc này là xem xét một dạng biểu đồ khác như **Column Chart** hoặc **Bar Chart**.

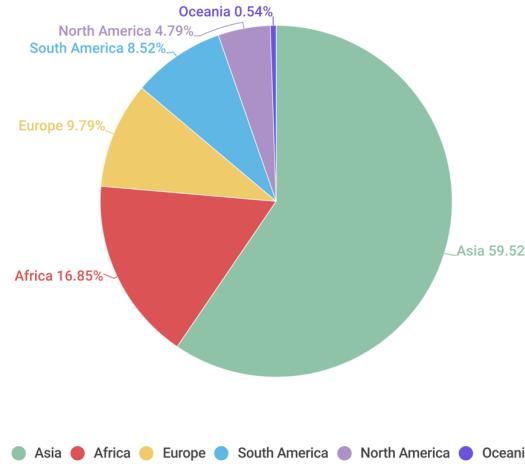


4. Nên sắp xếp giá trị các thể loại để dễ hiểu hơn: Sắp xếp lại dữ liệu giúp cho người xem nhận ra ngay thể loại có tỉ lệ cao nhất. Đồng thời với 2 thể loại gần nhau tương đương thì biết được thể loại nào có giá trị lớn hơn. Thông thường, giá trị trong Pie Chart được sắp xếp từ lớn đến nhỏ theo chiều kim đồng hồ như ví dụ bên dưới.

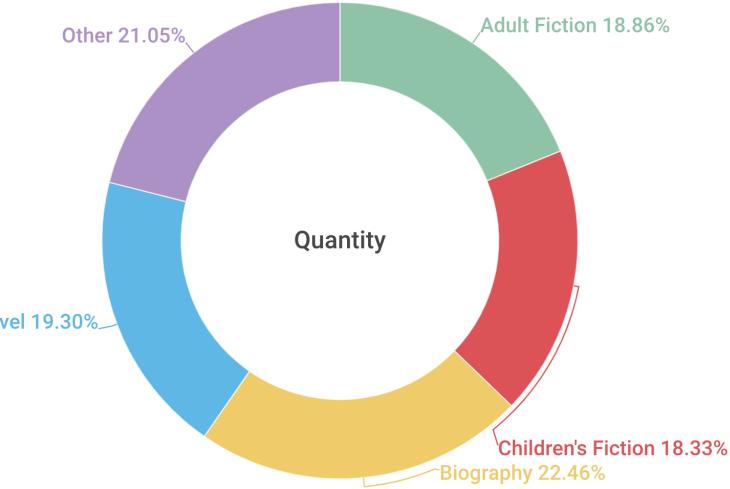


Một số dạng Pie chart

Global population by continent as of mid-2018



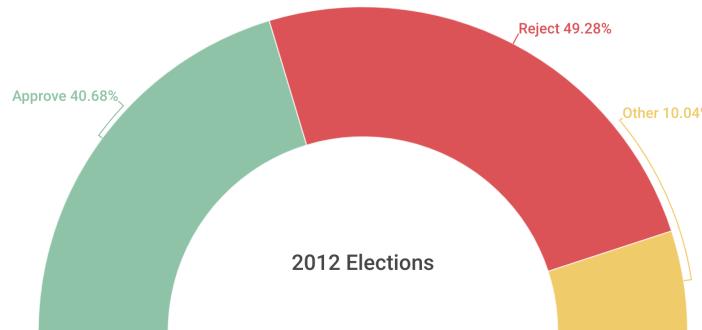
1. Pie chart



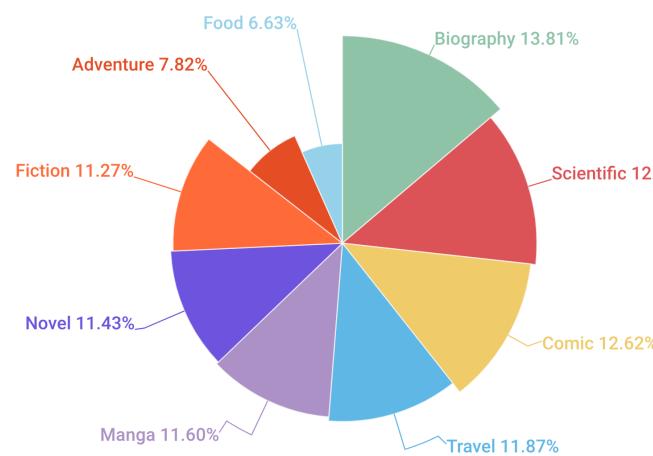
2. Donut chart



3. Stacked Donut chart



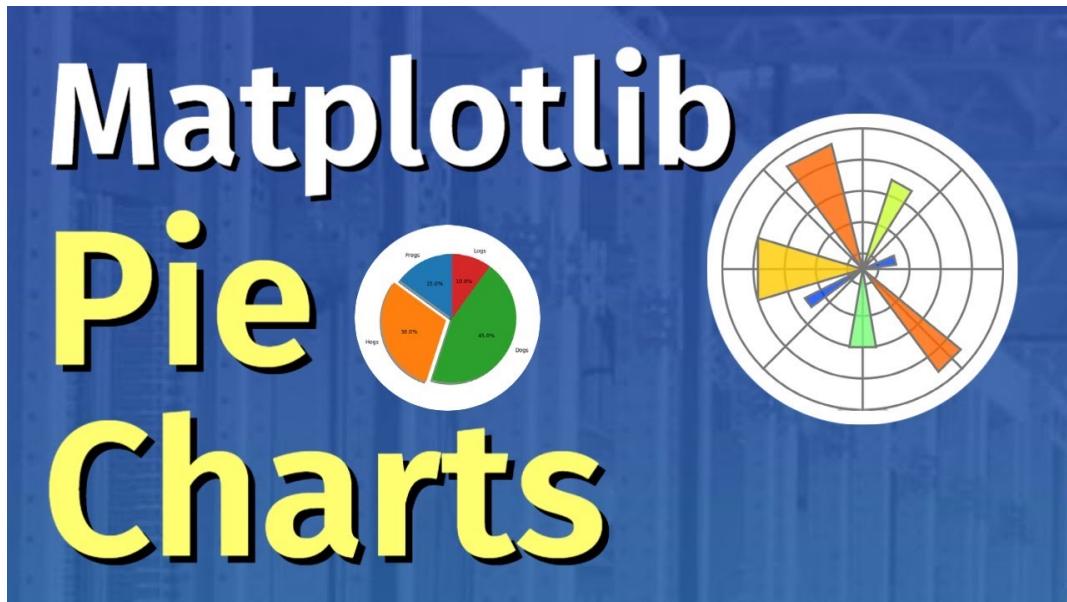
4. Semi-Circle Pie chart



5. Irregular Pie chart

Biểu đồ Bar chart với Matplotlib

Số liệu: Số lượng SV của các Khoa theo Giới tính.



| Khoa | K60 | K61 | K62 | K63 | K64 | K65 |
|------|-----|-----|-----|-----|-----|-----|
| Nam | 200 | 340 | 260 | 440 | 300 | 180 |
| Nữ | 30 | 60 | 90 | 160 | 180 | 220 |

```
1 #Tạo dữ liệu: Số lượng SV Nam - Nữ theo từng khoá.  
2 labels = ['K60','K61','K62','K63','K64','K65']  
3 boys = [200, 340, 260, 440, 300, 180]  
4 girls = [30, 60, 90, 160, 180, 220]  
5 total = list(np.array(boys) + np.array(girls))  
6  
7 sex =[ 'Nam','Nữ']  
8 sum_boy = sum(boys)  
9 sum_girl = sum(girls)
```



a. Pie chart

Biểu đồ tròn (Pie chart)

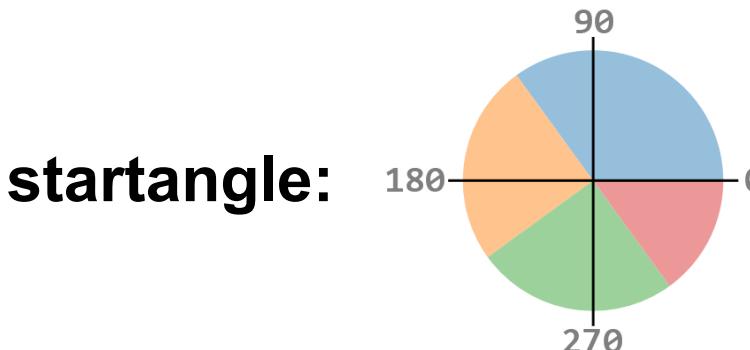


VINBIGDATA VINGROUP

Academy Vietnam

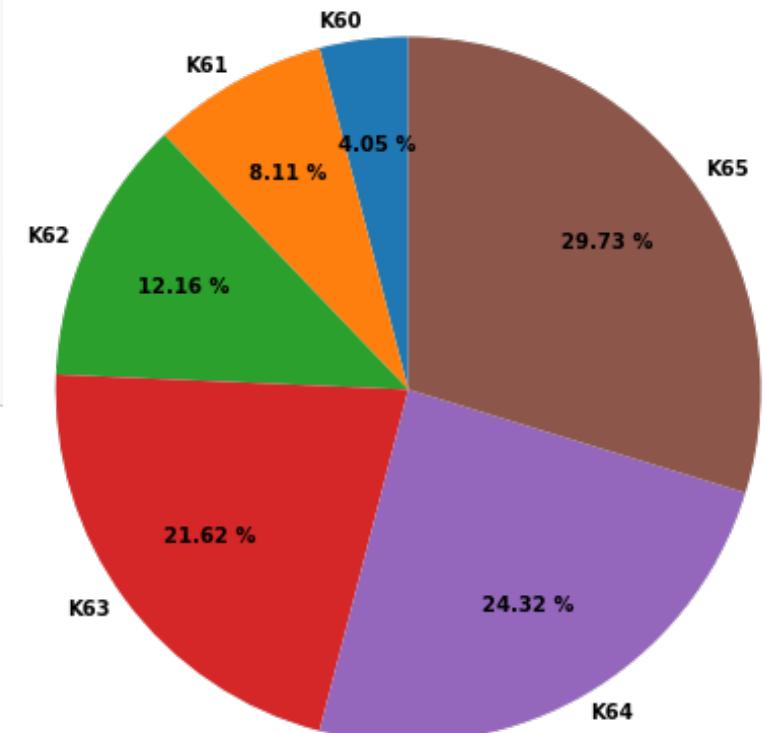
Cú pháp: plt.pie (values, labels)

```
1 plt.figure(figsize = (12,8))
2 #Vẽ biểu đồ tròn:
3 plt.pie(girls,           #Giá trị thể hiện
4          labels=labels,    #Nhãn tương ứng
5          autopct='%.2f %%', #Tính toán và hiển thị % tương ứng
6          pctdistance=0.7,   #Khoảng cách hiển thị giá trị % tới tâm.
7          startangle=90,     #Góc bắt đầu của biểu đồ
8          textprops={'color':'k','fontweight':'bold'},#thiết lập label
9          labeldistance=1.05, #Khoảng cách từ label tới biểu đồ
10         rotatelabels=False) #Label có xoay không?
11
12 plt.title('Tỷ lệ sinh viên Nữ theo từng Khoa', fontdict={'fontname':'Arial',
13                                         'fontweight':'bold',
14                                         'fontsize':18})
15 plt.show()
```



startangle:

Tỷ lệ sinh viên Nữ theo từng Khoa



Biểu đồ tròn (Pie chart)

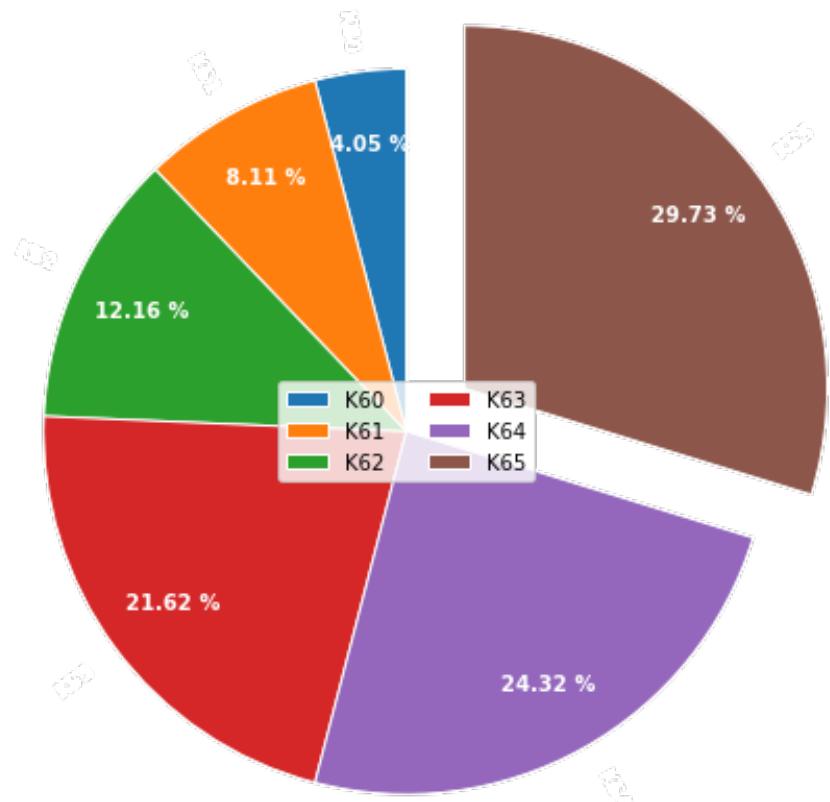


VINBIGDATA VINGROUP

Academy Vietnam

```
1 plt.figure(figsize = (12,8))
2 #Làm nổi bật một phần:
3 e = [0,0,0,0,0,0.2]
4 plt.pie(girls,
5         labels=labels,
6         autopct='%.2f %%',
7         pctdistance=0.8,
8         startangle=90,
9         labeldistance=1.05,
10        textprops={'color':'w','fontweight':'bold'},
11        rotatelabels=True,
12        wedgeprops=dict(edgecolor='w'),#Đường viền màu trắng
13        explode=e)#Làm nổi bật một phần
14
15 plt.title('Tỷ lệ sinh viên Nữ theo từng Khoa', fontdict={'fontname':'Arial',
16                                         'fontweight':'bold',
17                                         'fontsize':18})
18 plt.legend(ncol=2, loc='center')
19 plt.show()
```

Tỷ lệ sinh viên Nữ theo từng Khoa





b. Donus chart

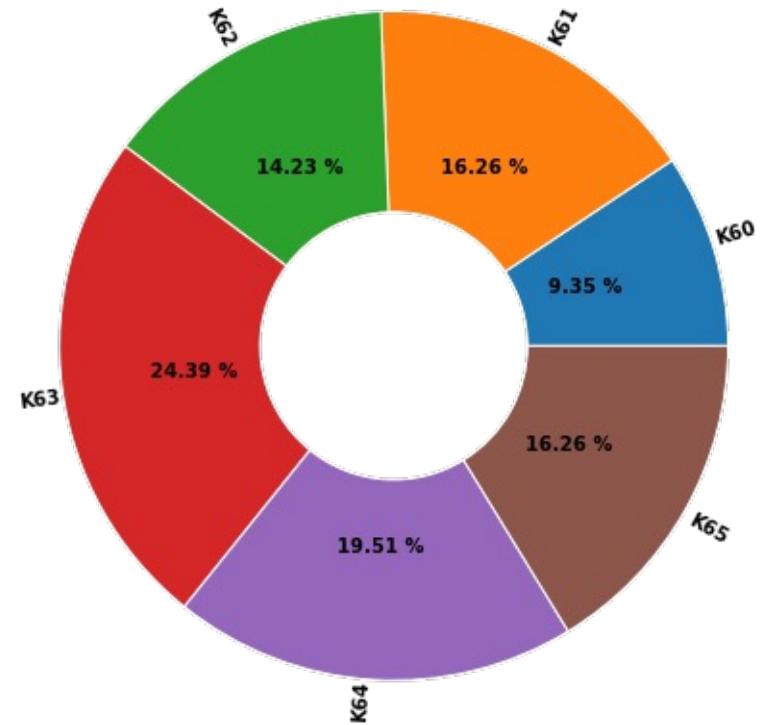


Donus chart

Cú pháp: plt.pie (values, labels)

```
1 plt.figure(figsize = (12,8))
2 #Vẽ biểu đồ:
3 plt.pie(total,
4         labels=labels,
5         textprops={'color':'k','fontweight':'bold'},
6         rotatelabels=True,
7         labeldistance=1.0,
8         wedgeprops=dict(width=0.6,edgecolor='w'), #Xác định độ rộng của Pie
9         autopct='%.2f %%',
10        pctdistance=0.6)
11
12 plt.title('TỶ LỆ SINH VIÊN CỦA TỪNG KHOÁ', fontdict={'fontname':'Tahoma',
13                                         'fontweight':'bold',
14                                         'fontsize':18})
15 plt.show()
```

TỶ LỆ SINH VIÊN CỦA TỪNG KHOÁ

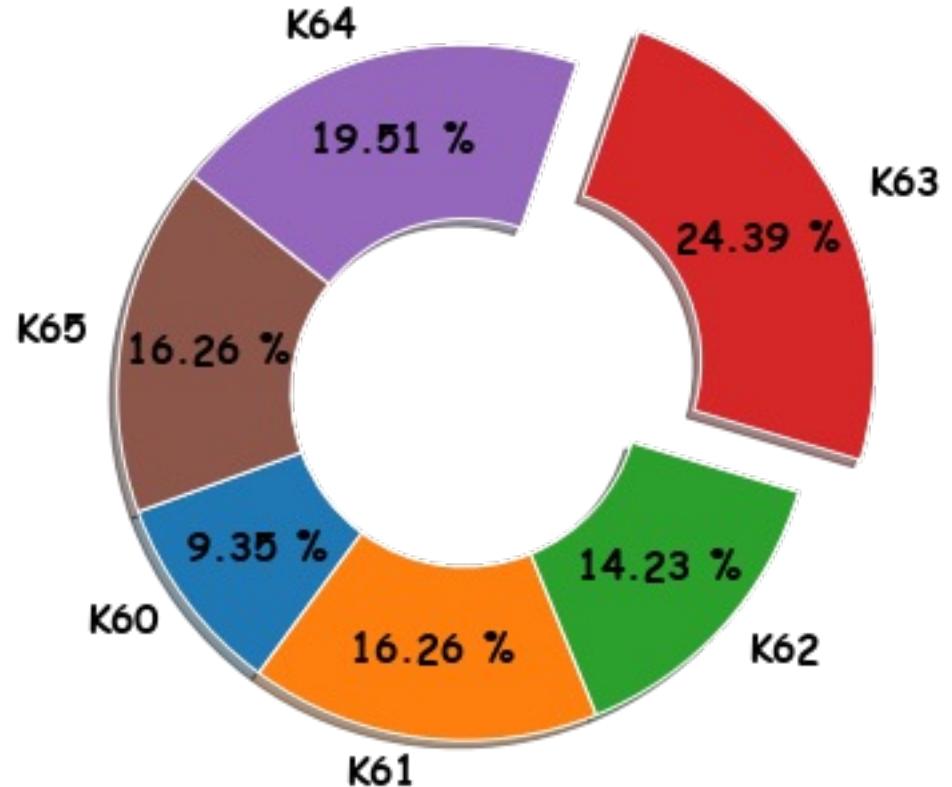




Donut chart

```
1 plt.figure(figsize = (12,6))
2
3 #Làm nổi bật một phần
4 e = [0,0,0,0.2,0,0]
5 plt.pie(total,
6         labels=labels,
7         explode=e, #Làm nổi bật biểu đồ
8         startangle=200,
9         textprops={'color':'k','fontweight':'bold',
10                 'fontname':'Comic Sans MS',
11                 'fontsize':15},
12         wedgeprops=dict(width=0.5,edgecolor='w'),
13         shadow=True, #Tạo bóng cho biểu đồ
14         autopct='%.2f %%',
15         pctdistance=0.75)
16
17 plt.title('TỶ LỆ SINH VIÊN CỦA TỪNG KHOÁ',
18             fontdict={'fontname':'Tahoma',
19                         'fontweight':'bold',
20                         'fontsize':18})
21 plt.show()
```

TỶ LỆ SINH VIÊN CỦA TỪNG KHOÁ

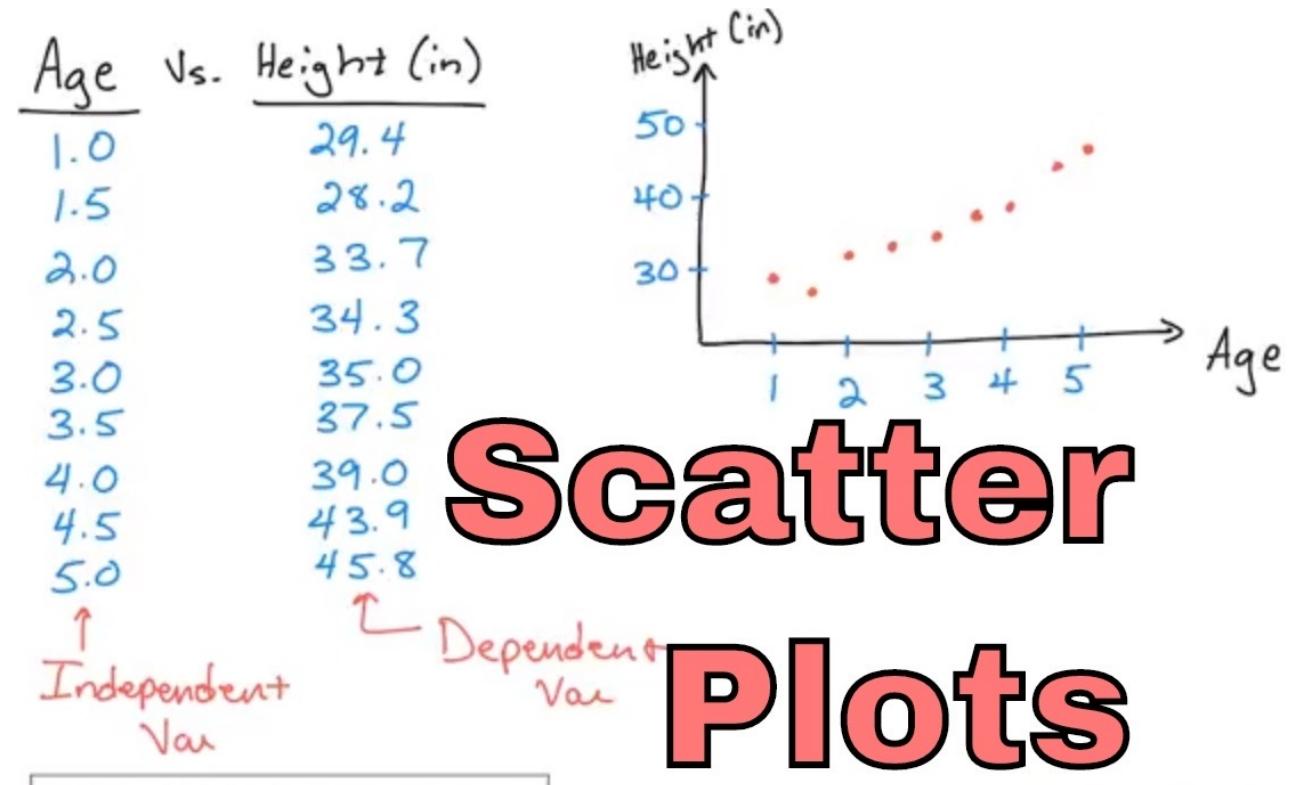


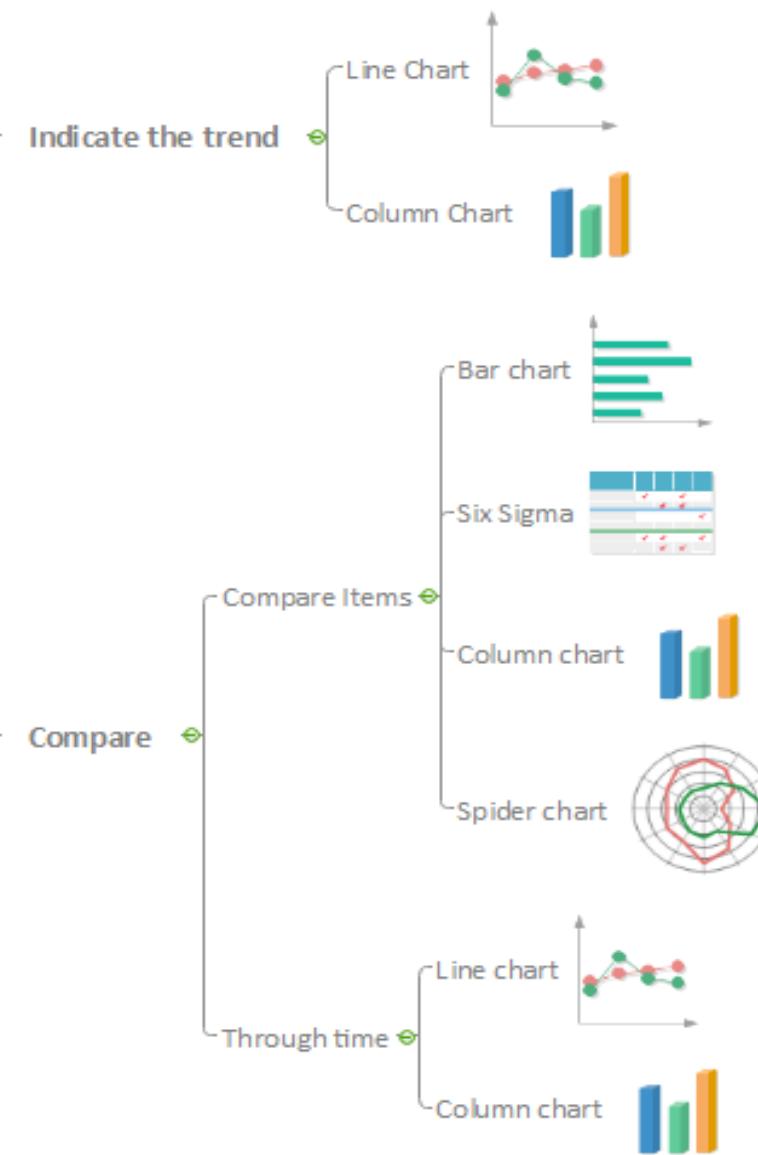
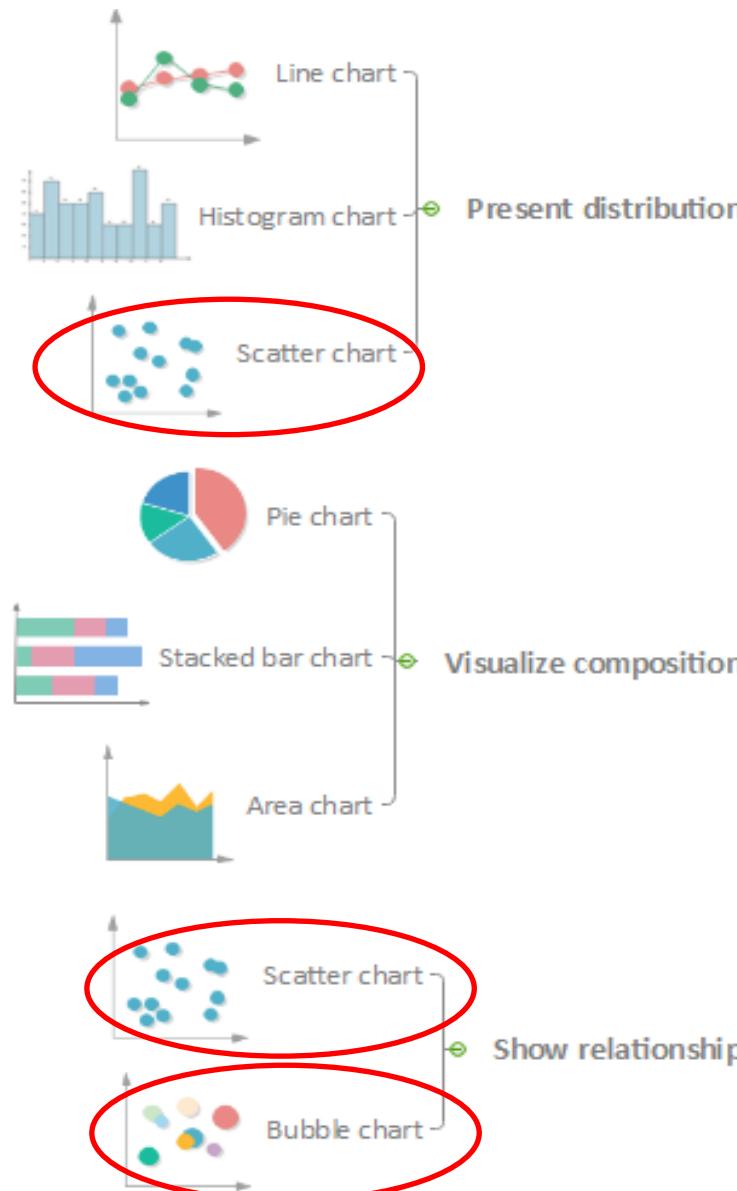


7. Biểu đồ phân tán (Scatter chart)

Biểu đồ phân tán (Scatter chart)

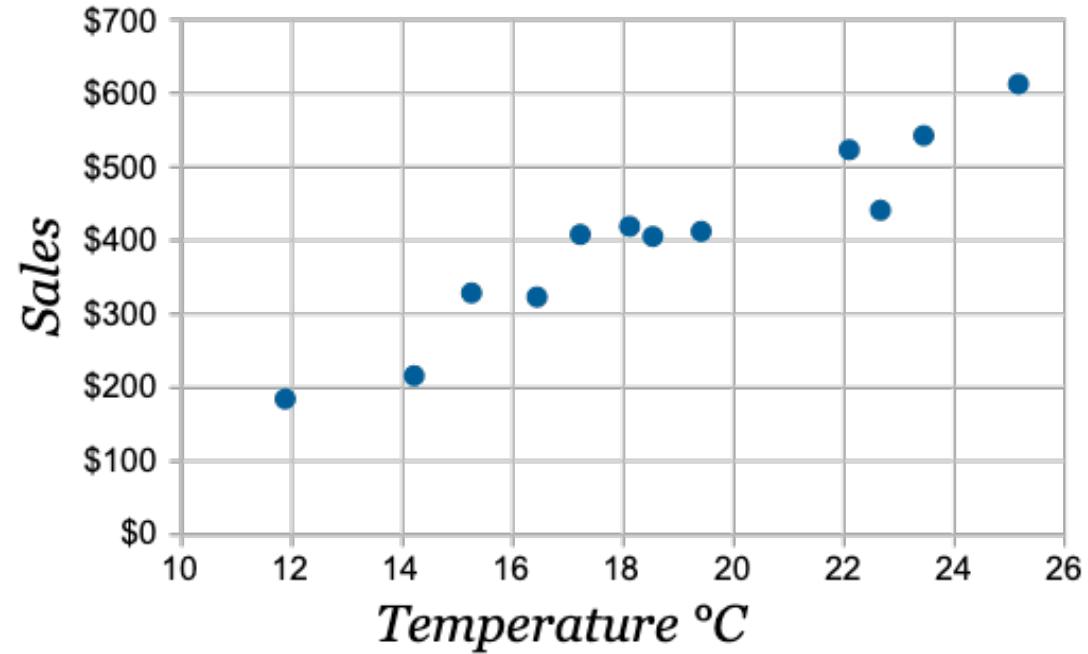
- **Biểu đồ phân tán (Scatter chart, Scatterplot, scatter graph)** là loại biểu đồ được dựng bởi các điểm theo tọa độ toán học để xác định mối tương quan giữa 2 biến.
- Đồ thị thể hiện 2 bộ dữ liệu, trục tung Y được sử dụng cho biến được dự đoán (biến phụ thuộc), trục hoành X được sử dụng cho biến dùng để dự đoán (biến độc lập).
- Scatter plot được sử dụng khi có 2 cặp dữ liệu (biến) và muốn xác định 2 biến có liên quan với nhau hay không? Liên quan nhiều hay ít và như thế nào?





Biểu đồ phân tán (Scatter chart)

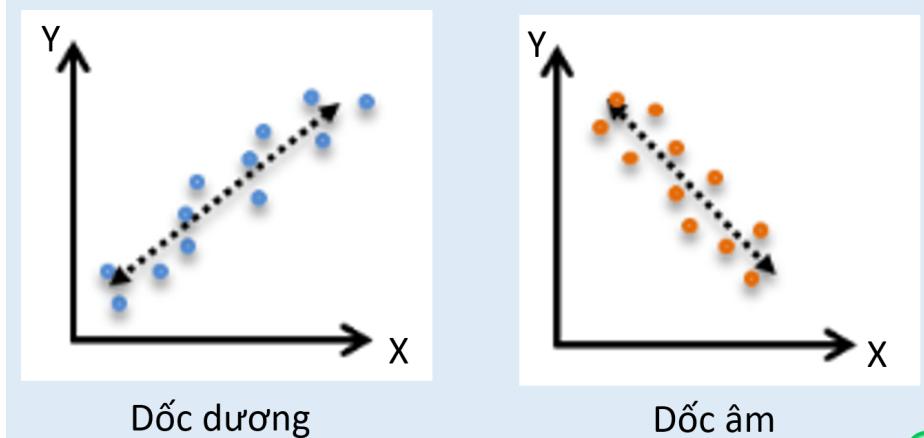
X và Y
có
tương
quan...



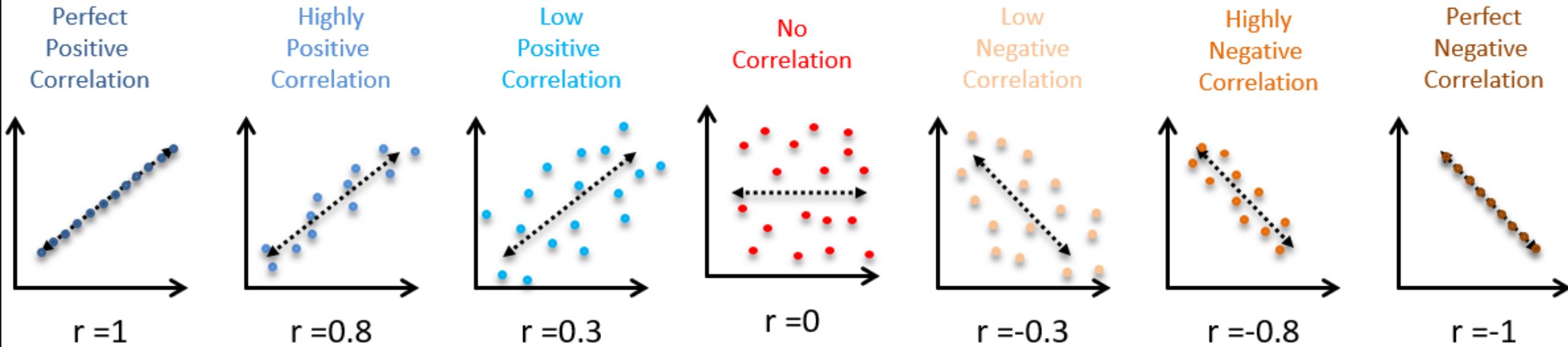
- Biểu đồ phân tán sẽ cho chúng ta thấy được mức độ tương quan giữa 2 biến

Biểu đồ phân tán (Scatter chart)

- Dựa vào hình dạng, bờ dốc và độ tập trung điểm của biểu đồ để xác định mối tương quan giữa 2 biến.

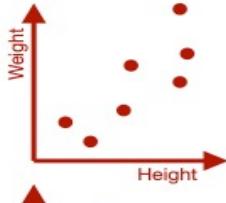


Scatter Plots & Correlation Examples



Biểu đồ phân tán (Scatter chart)

Scatter Graphs can show a relationship between two variables.



...such as people's height and weight.

...or the number of staff working in KFC and the wait time for food.

...or the distance people live from work and their best score in darts.

If the two variables have a relationship we call it correlation.

There are different types of correlation:

Positive correlation:

As one value goes up, so does the other.

Negative correlation:

As one value goes up, the other goes down.

No correlation:

There is no obvious relationship.

Correlation can be strong or weak.

If the correlation is strong, all the points will closely follow a straight line.

Strong correlation

If the correlation is weak, the points will follow the line more loosely.

Weak correlation

Scatter Graphs

We can show the correlation more clearly by drawing a Line of Best Fit.

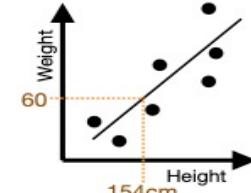
This should pass through the middle of all the points (but does not have to touch any of the points).



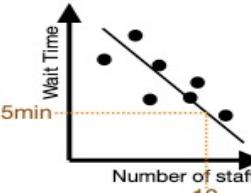
We can use the Line of Best Fit to make predictions of other results.

For example, we can estimate:

...someone's height if we know their weight is 60kg.



...or the wait time in KFC if we know they have 10 staff on today.



Sometimes you might be asked to explain the correlation in context.

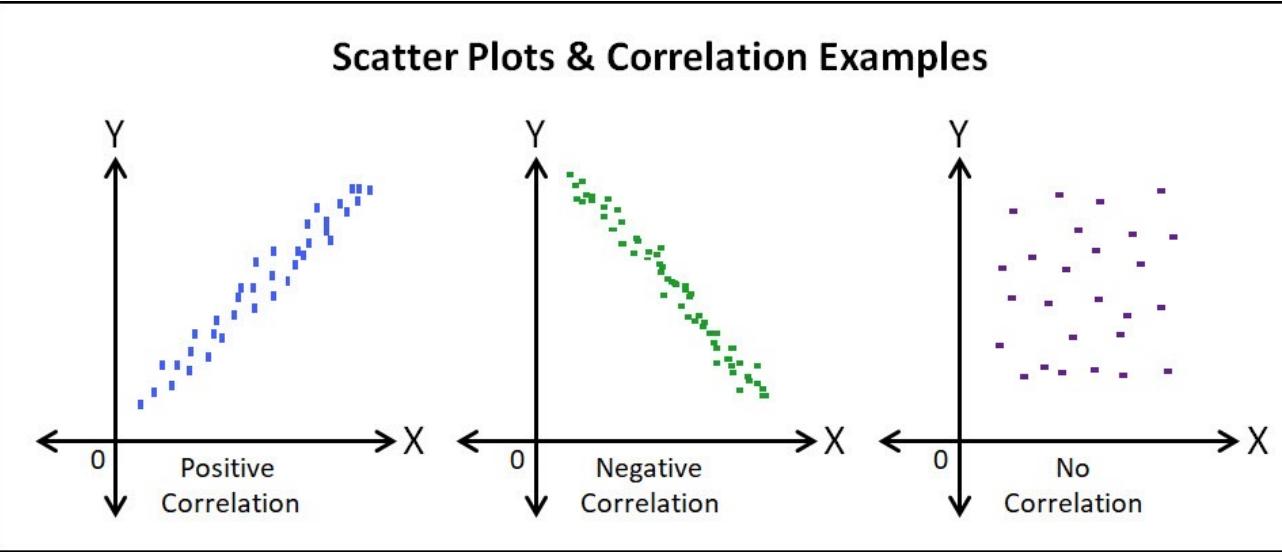
This means describing what is actually happening. eg:

"Taller people are usually heavier."

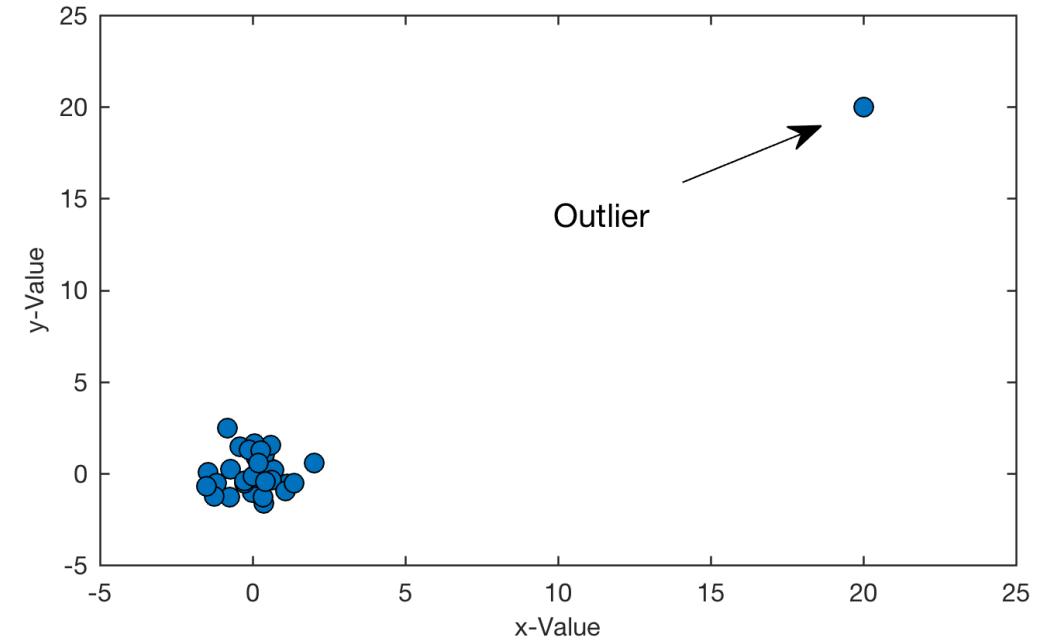
"When there are more staff working, you wait less."

"There is no relationship between how far people live from work and their darts ability."

Biểu đồ phân tán sử dụng để?



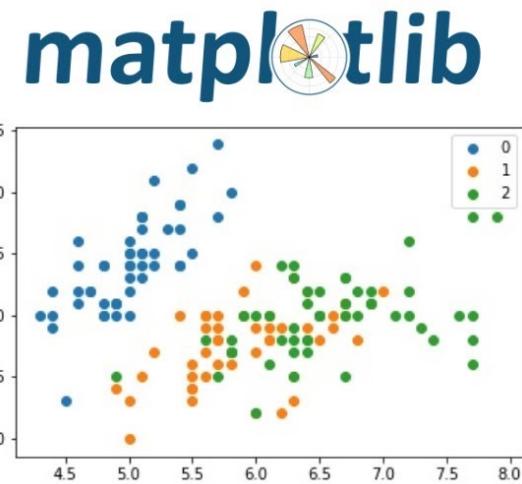
Xác định mối tương quan giữa các biến số (có hay không? Mạnh hay yếu?)



Phát hiện các điểm ngoại lai (outlier) trong tập dữ liệu

Scatter chart với Matplotlib

Tập dữ liệu **Diamonds.txt** lưu trữ trọng lượng (carat) và giá (\$) tương ứng của 50 viên kim cương.



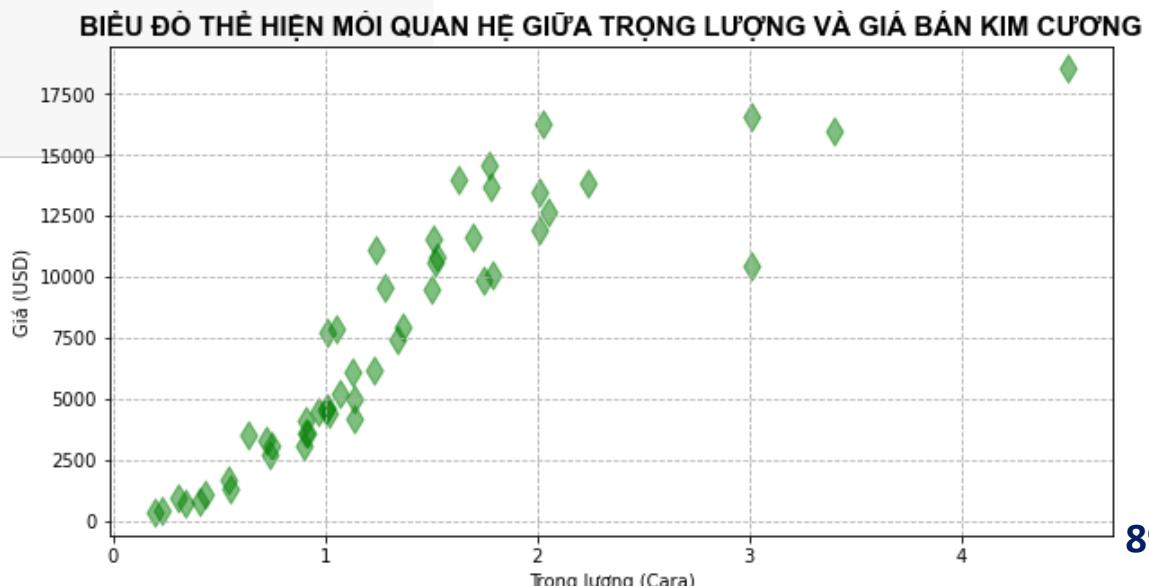
DATA VISUALIZATION PYTHON

| | Dia... |
|------|--------|
| 0.23 | 484 |
| 0.31 | 942 |
| 0.2 | 345 |
| 1.02 | 4459 |
| 1.63 | 14022 |
| 1.14 | 4212 |
| 2.01 | 11925 |
| 1.28 | 9548 |
| 1.7 | 11605 |
| 1.01 | 4642 |
| 0.64 | 3541 |
| 0.97 | 4504 |
| 1.78 | 13691 |
| 3.4 | 15964 |
| 3.01 | 10453 |
| 1.51 | 11560 |
| 1.37 | 7979 |
| 1.5 | 9533 |
| 0.54 | 1723 |
| 0.72 | 3344 |
| 1.13 | 6133 |
| 2.24 | 13827 |
| 3.01 | 16538 |
| 4.5 | 18531 |
| 0.92 | 3625 |
| 1.05 | 7879 |
| 0.55 | 1319 |
| 0.74 | 2761 |
| 0.91 | 3620 |
| 1.23 | 6165 |

Biểu đồ phân tán (Scatter chart)

Cú pháp: plt.scatter (x, y)

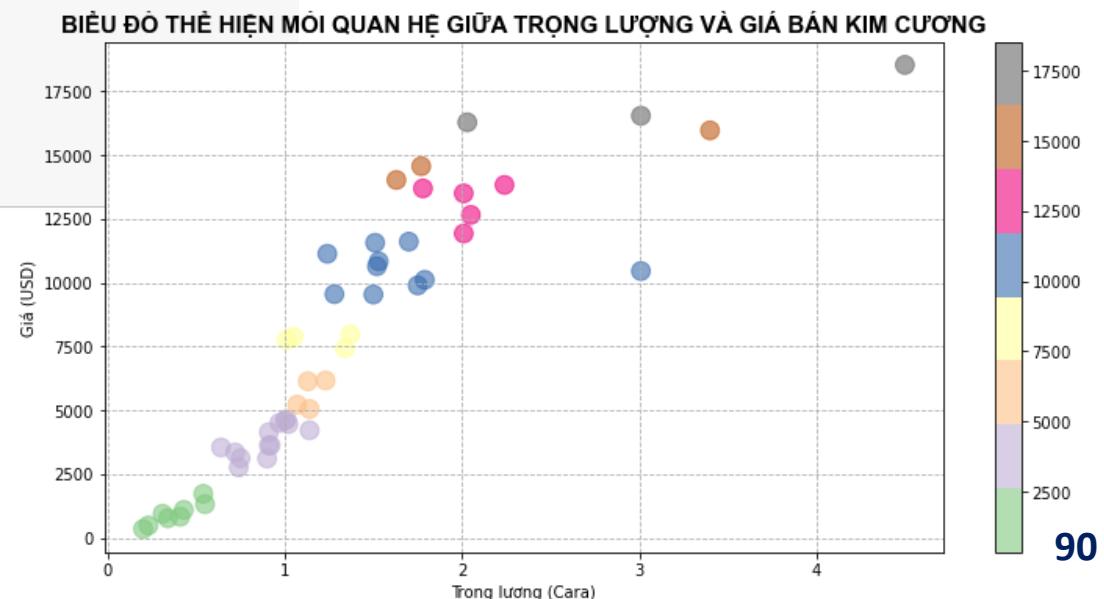
```
1 plt.figure(figsize = (10,5))
2
3 plt.scatter(weight,           #Dữ liệu trục X
4               price,          #Dữ liệu trục Y
5               c='g',            #Màu của Point
6               marker='d',       #Kiểu Point
7               s=120,             #Kích thước của Point
8               alpha=0.5)        #Độ trong suốt của Point
9
10 plt.title('BIỂU ĐỒ THỂ HIỆN MỐI QUAN HỆ GIỮA TRỌNG LƯỢNG VÀ GIÁ BÁN KIM CƯƠNG',
11            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
12 plt.grid(ls='--')
13 plt.xlabel('Trọng lượng (Cara)')
14 plt.ylabel('Giá (USD)')
15 plt.show()
```



Biểu đồ phân tán (Scatter chart)

Sử dụng colorbar:

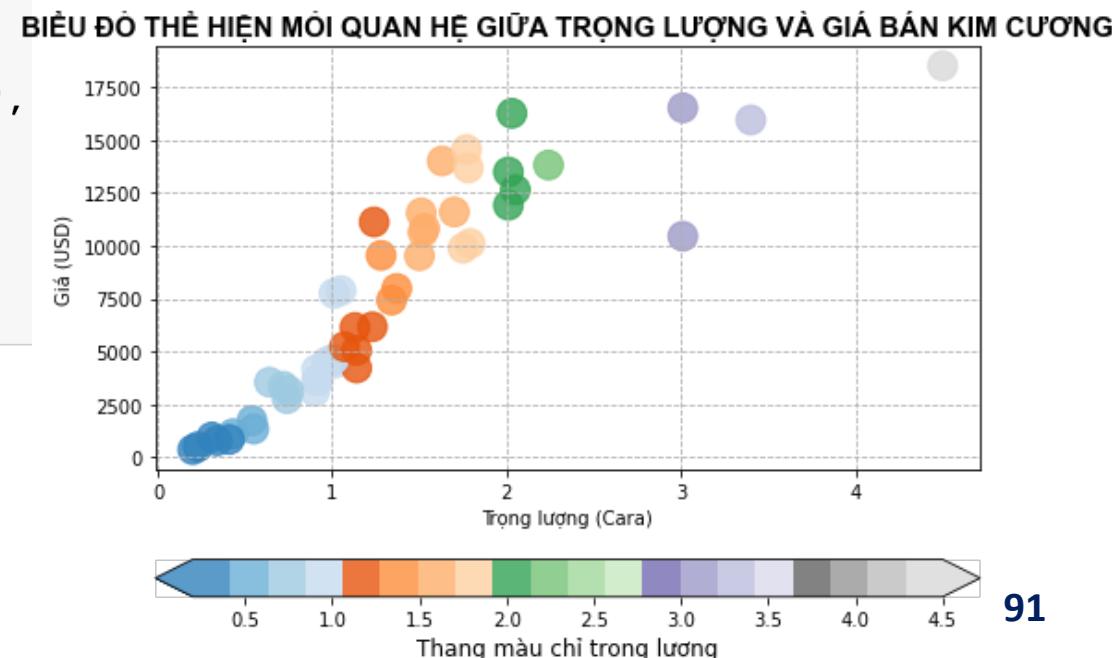
```
1 plt.figure(figsize = (12,6))
2 #Vẽ biểu đồ scatter:
3 plt.scatter(weight, price,
4             s=140,
5             alpha=0.6,
6             c=price,
7             cmap='Accent') #Sử dụng color map
8 plt.colorbar() #Hiển thị thanh color bar:
9
10
11 plt.title('BIỂU ĐỒ THỂ HIỆN MỐI QUAN HỆ GIỮA TRỌNG LƯỢNG VÀ GIÁ BÁN KIM CƯƠNG',
12            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
13 plt.grid(ls='--')
14 plt.xlabel('Trọng lượng (Cara)')
15 plt.ylabel('Giá (USD)')
16
17 plt.show()
```



Biểu đồ phân tán (Scatter chart)

Sử dụng colorbar:

```
1 plt.figure(figsize = (8,6))
2 #Vẽ biểu đồ scatter:
3 plt.scatter(weight, price,
4             s=250,
5             alpha=0.8,
6             c=weight,
7             cmap='tab20c') #Sử dụng color map
8
9 #Hiển thị và setup thanh color bar:
10 cbar = plt.colorbar(location ='bottom', #Vị trí của colorbar
11                      extend='both', #Đầu của colorbar
12                      pad=0.15) #Khoảng cách giữa thang màu và biểu đồ
13 cbar.set_label(label='Thang màu chỉ trọng lượng',size=12)
14
15 plt.title('BIỂU ĐỒ THỂ HIỆN MỐI QUAN HỆ GIỮA TRỌNG LƯỢNG VÀ GIÁ BÁN KIM CƯƠNG',
16            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
17 plt.grid(ls='--')
18 plt.xlabel('Trọng lượng (Cara)')
19 plt.ylabel('Giá (USD)')
20
21 plt.show()
```





a. Ứng dụng của Scatter chart

Đánh giá độ tương quan giữa 2 biến



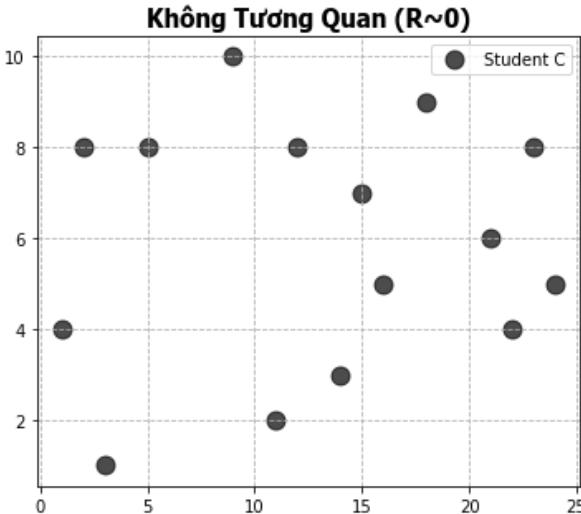
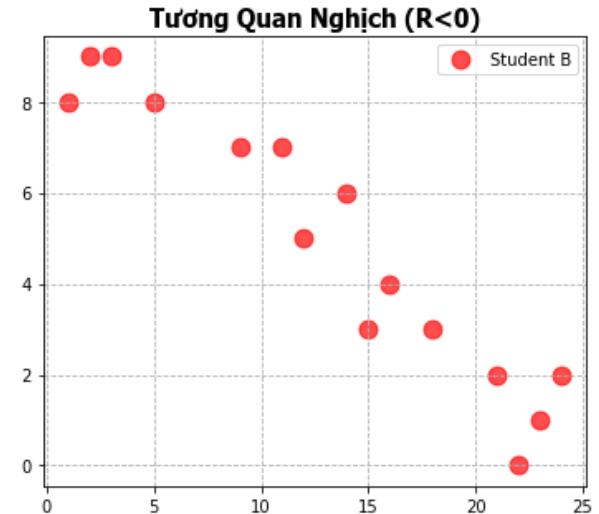
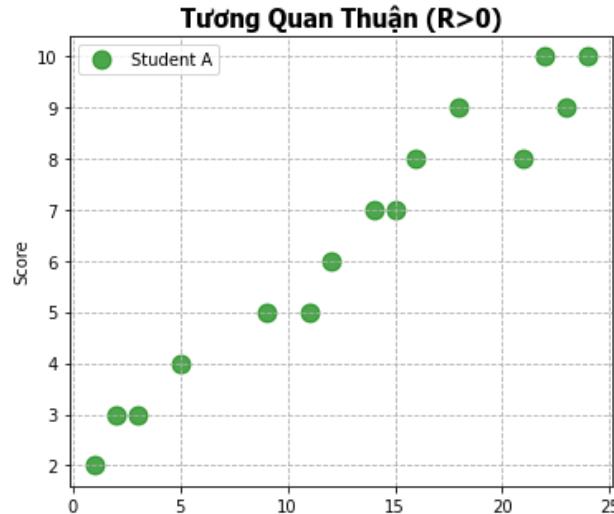
VINBIGDATA VINGROUP

Academy
Vietnam

Bảng dữ liệu thể hiện thời gian dành cho việc học và số điểm nhận được của 3 sinh viên A, B, C

| Hour | Score_A | Score_B | Score_C |
|------|---------|---------|---------|
|------|---------|---------|---------|

| | | | | |
|----|----|----|---|----|
| 0 | 1 | 2 | 8 | 4 |
| 1 | 2 | 3 | 9 | 8 |
| 2 | 3 | 3 | 9 | 1 |
| 3 | 5 | 4 | 8 | 8 |
| 4 | 9 | 5 | 7 | 10 |
| 5 | 11 | 5 | 7 | 2 |
| 6 | 12 | 6 | 5 | 8 |
| 7 | 14 | 7 | 6 | 3 |
| 8 | 15 | 7 | 3 | 7 |
| 9 | 16 | 8 | 4 | 5 |
| 10 | 18 | 9 | 3 | 9 |
| 11 | 21 | 8 | 2 | 6 |
| 12 | 22 | 10 | 0 | 4 |
| 13 | 23 | 9 | 1 | 8 |
| 14 | 24 | 10 | 2 | 5 |



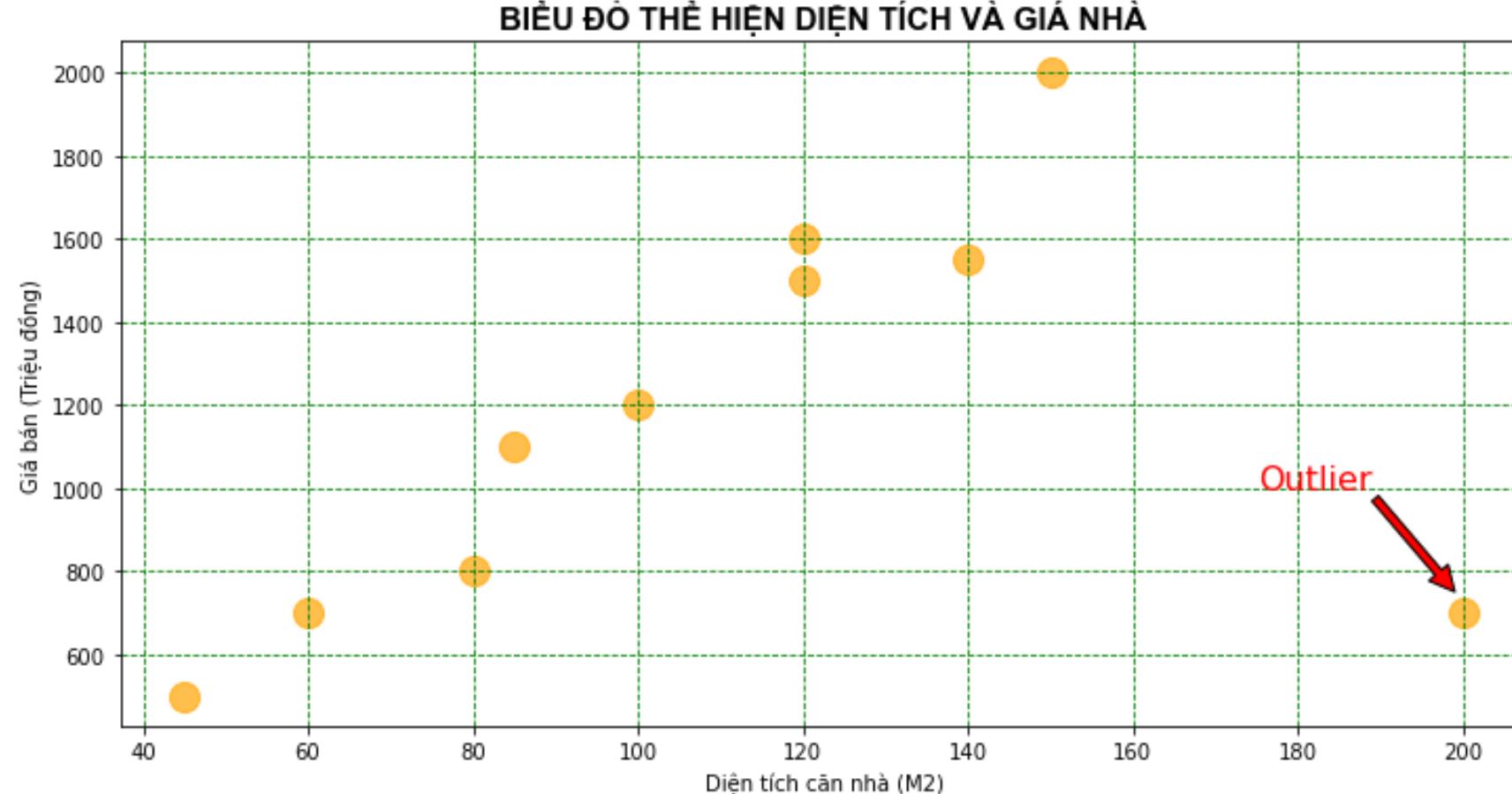
Phát hiện các giá trị ngoại lai



VINBIGDATA VINGROUP

Academy
Vietnam

```
1 #Dữ liệu bao gồm diện tích và giá của một số căn nhà:  
2 area_house = [45, 80, 120, 100, 150, 60, 200, 120, 85, 140]      #Diện tích (m2)  
3 price_house = [500, 800, 1600, 1200, 2000, 700, 700, 1500, 1100, 1550] #Giá nhà (Triệu đồng)
```



b. Mở rộng....

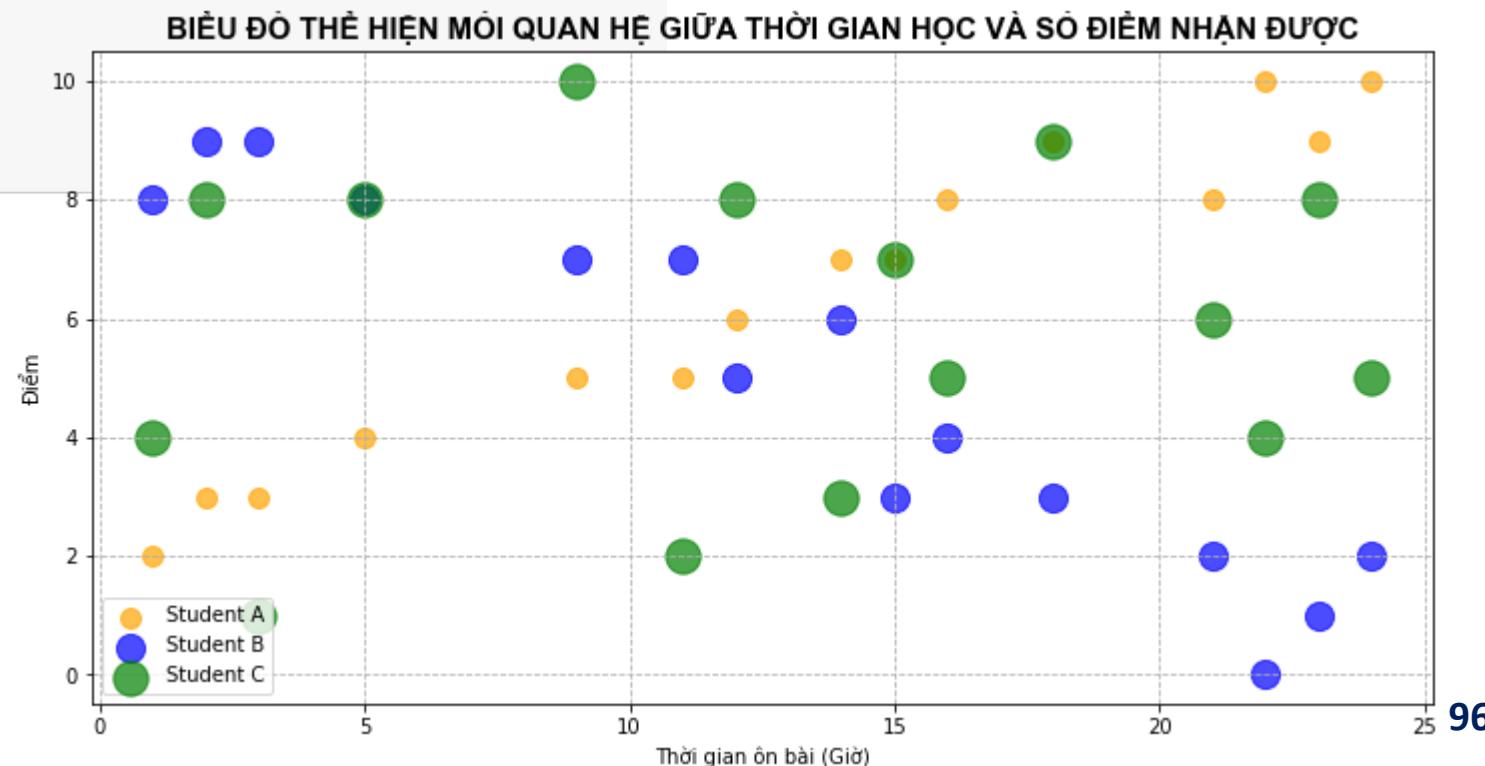
Nhiều scatter trên cùng một plot



VINBIGDATA VINGROUP

Academy
Vietnam

```
1 plt.figure(figsize = (12,6))
2
3 #Vẽ nhiều biểu đồ scatter trên một plot:
4 plt.scatter(hour, scoreA,s=100, alpha=0.7, c='orange', label='Student A')
5 plt.scatter(hour, scoreB,s=200, alpha=0.7, c='blue', label='Student B')
6 plt.scatter(hour, scoreC,s=300, alpha=0.7, c='g', label='Student C')
7
8 plt.title('BIỂU ĐỒ THỂ HIỆN MỐI QUAN HỆ GIỮA THỜI GIAN HỌC VÀ SỐ ĐIỂM NHẬN ĐƯỢC',
9             fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
10
11 plt.grid(ls='--')
12 plt.xlabel('Thời gian ôn bài (Giờ)')
13 plt.ylabel('Điểm')
14 plt.legend(loc='lower left')
15 plt.show()
```

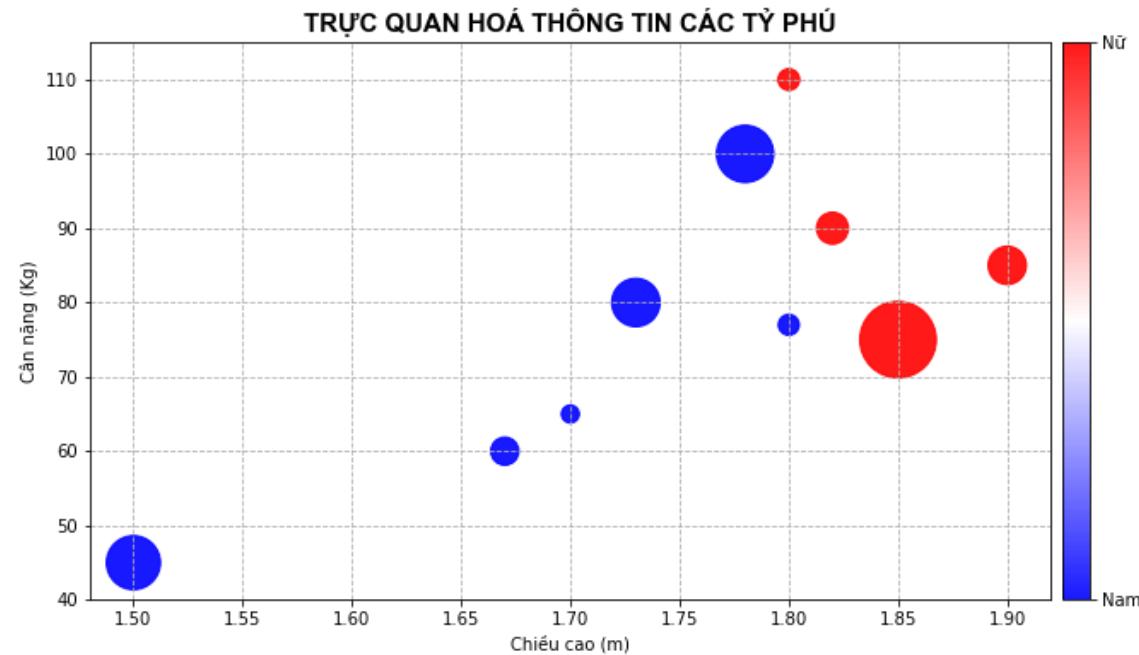


Nhiều biến trên một scatter

Mỗi một biểu đồ scatter có thể biểu diễn được tối đa 4 biến (x, y, size, color)

```
1 #Thông tin của 10 doanh nhân:  
2 weight = [ 45, 80, 110, 100, 75, 60, 85, 77, 65, 90] #Cân nặng (Kg)  
3 height = [1.50, 1.73, 1.80, 1.78, 1.85, 1.67, 1.90, 1.80, 1.70, 1.82] #Chiều cao (M)  
4 sex = [ 0, 0, 1, 0, 1, 0, 1, 0, 0, 1] #Giới tính (0:Nam - 1:Nữ)  
5 money = [1000, 800, 160, 1120, 2000, 270, 500, 150, 110, 350] # Tài sản (Triệu USD)
```

```
1 plt.figure(figsize = (12,6))  
2  
3 #Vẽ biểu đồ scatter biểu diễn thông tin của 10 doanh nhân:  
4 #Biến 1: Chiều cao - Trục X  
5 #Biến 2: Cân nặng - Trục Y  
6 #Biến 3: Giới tính - Màu sắc của Point  
7 #Biến 4: Tài sản - Kích thước của Point  
8 plt.scatter(height, weight, c=sex, s=money, alpha=0.9, cmap='bwr')  
9  
10 cbar = plt.colorbar(pad=0.01)  
11 cbar.set_ticks([0,1])  
12 cbar.set_ticklabels(["Nam", "Nữ"])  
13 plt.title('TRỰC QUAN HOÁ THÔNG TIN CÁC TỶ PHÚ',  
14 fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})  
15 plt.grid(ls='--')  
16 plt.xlabel('Chiều cao (m)')  
17 plt.ylabel('Cân nặng (Kg)')  
18 plt.ylim([40,115])  
19 plt.show()
```

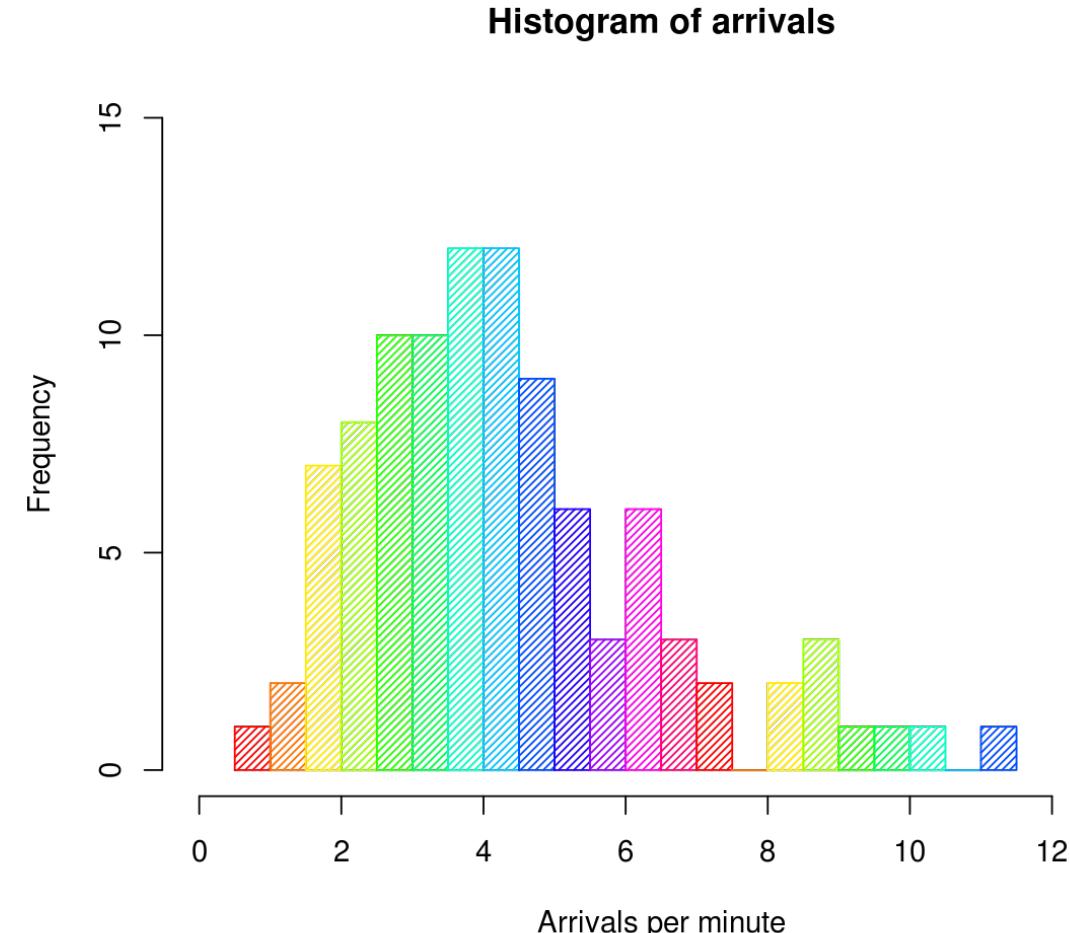




8. Biểu đồ tần suất (Histogram chart)

Biểu đồ tần suất (Histogram chart)

- **Biểu đồ Histogram chart** là một dạng biểu đồ thể hiện tần suất dạng cột. Nó mô tả dữ liệu một cách đơn giản mà không làm mất bất cứ thông tin thống kê nào của tập dữ liệu.
- Histogram cho thấy hình thái phân bố của dữ liệu. Sử dụng biểu đồ Histogram có thể trả lời cho các câu hỏi:
 - Kiểu phân bố của dữ liệu?
 - Độ rộng của dữ liệu thế nào?
 - Dữ liệu có đối xứng hay không?
 - Có dữ liệu nào nằm ngoài hay không?

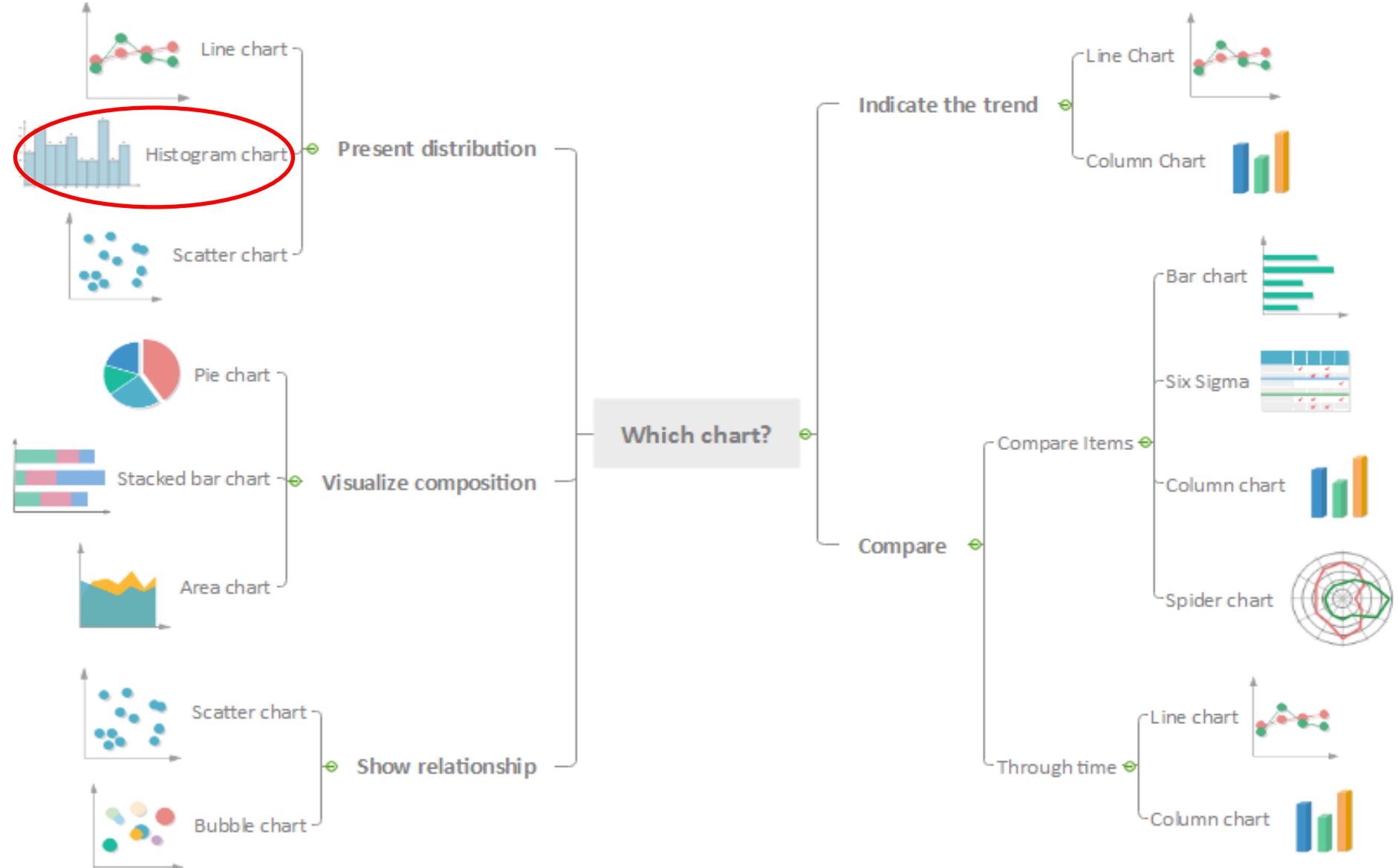


Biểu đồ tần suất (Histogram chart)



VINBIGDATA VINGROUP

Academy
Vietnam



Cách xây dựng biểu đồ Histogram



VINBIGDATA VINGROUP

Academy
Vietnam

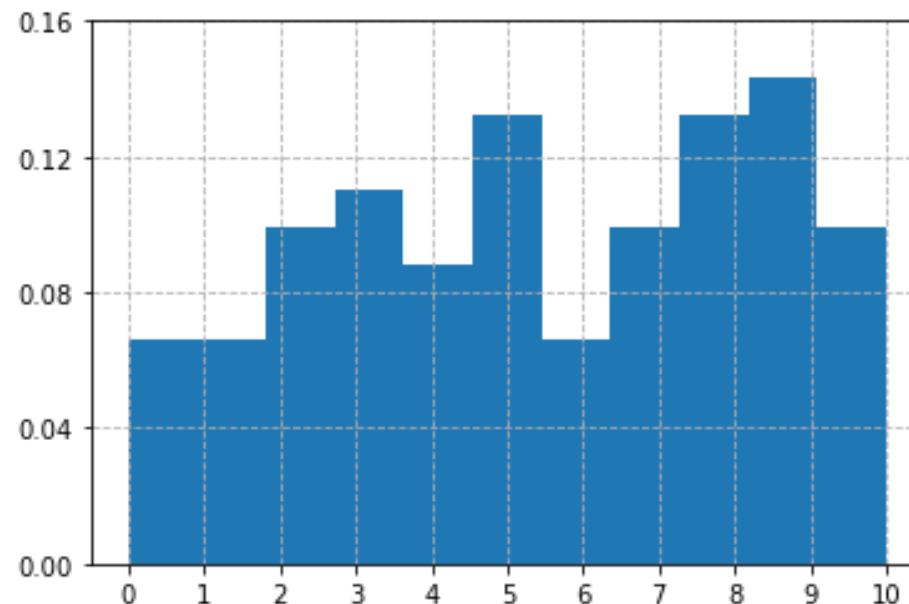
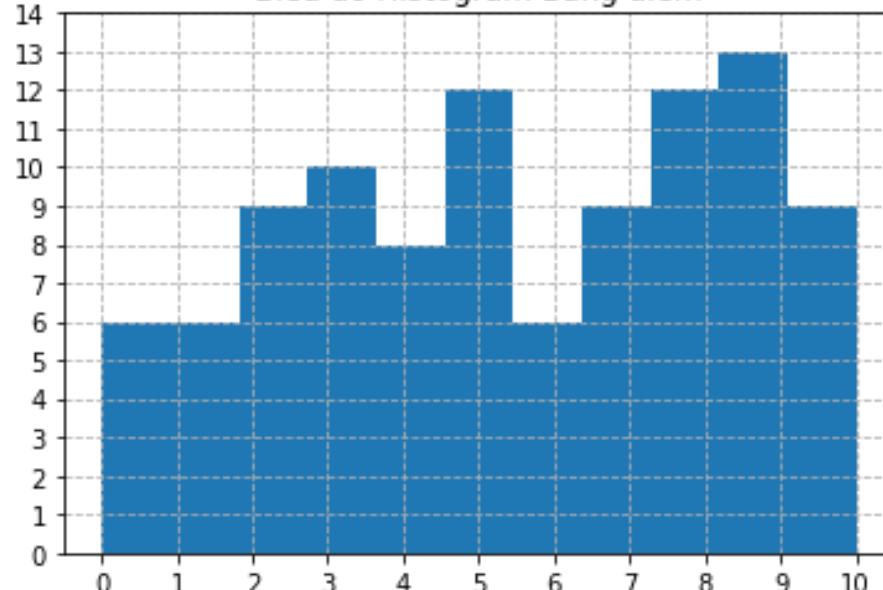
- Bảng điểm của 100 học sinh khối lớp 5

```
[ 3,  0,  5,  6,  9,  5,  3,  2,  4,  5]
[ 8,  7,  8,  0,  1,  8,  8,  8,  3,  9]
[ 4,  8,  4,  7,  2,  9,  7,  4,  7,  10]
[ 0, 10,  7,  5,  0,  5,  5,  4,  1,  8]
[10,  9,  0,  2,  3,  8,  8,  7,  7,  6]
[ 3,  9,  6,  9,  5,  5, 10,  5, 10,  5]
[ 1,  9, 10,  2,  5, 10,  3,  1,  2,  7]
[ 4,  9,  9,  2,  3,  3,  2,  6,  9,  5]
[ 9,  0,  6,  9,  4,  9, 10,  8,  4,  3]
[ 1,  2, 10,  7,  8,  8,  1,  2,  3,  6]
```

Bins = 11

| Điểm | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
|------|------|------|------|-----|------|------|------|------|------|------|------|-----|
| n | 6 | 6 | 9 | 10 | 8 | 12 | 6 | 9 | 12 | 13 | 9 | 100 |
| P(n) | 0.06 | 0.06 | 0.09 | 0.1 | 0.08 | 0.12 | 0.06 | 0.09 | 0.12 | 0.13 | 0.09 | 1 |

Biểu đồ Histogram Bảng điểm



Cách xây dựng biểu đồ Histogram



VINBIGDATA VINGROUP

Academy
Vietnam

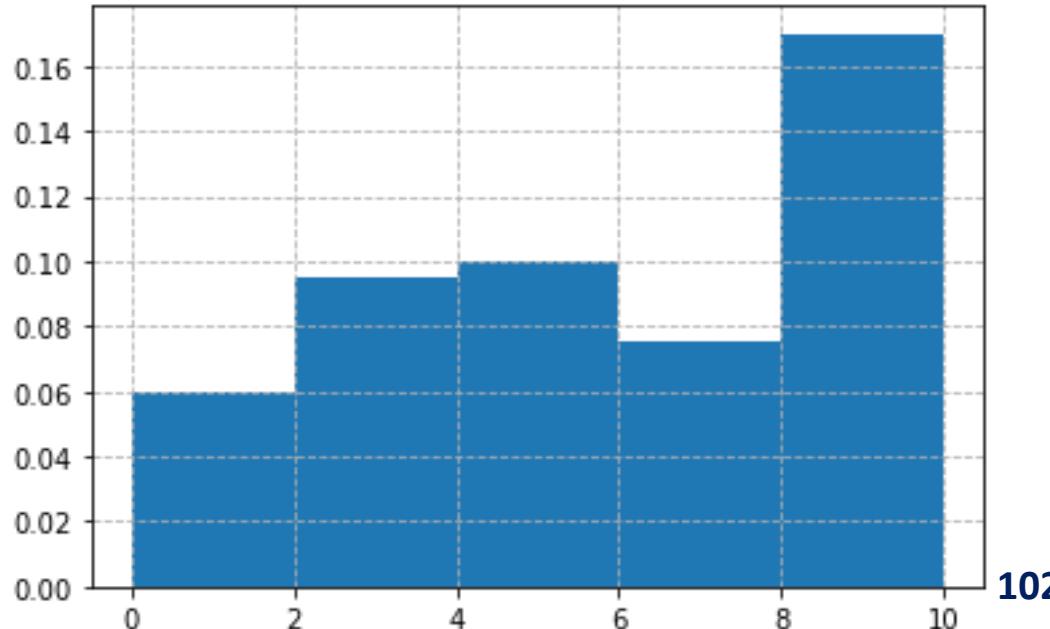
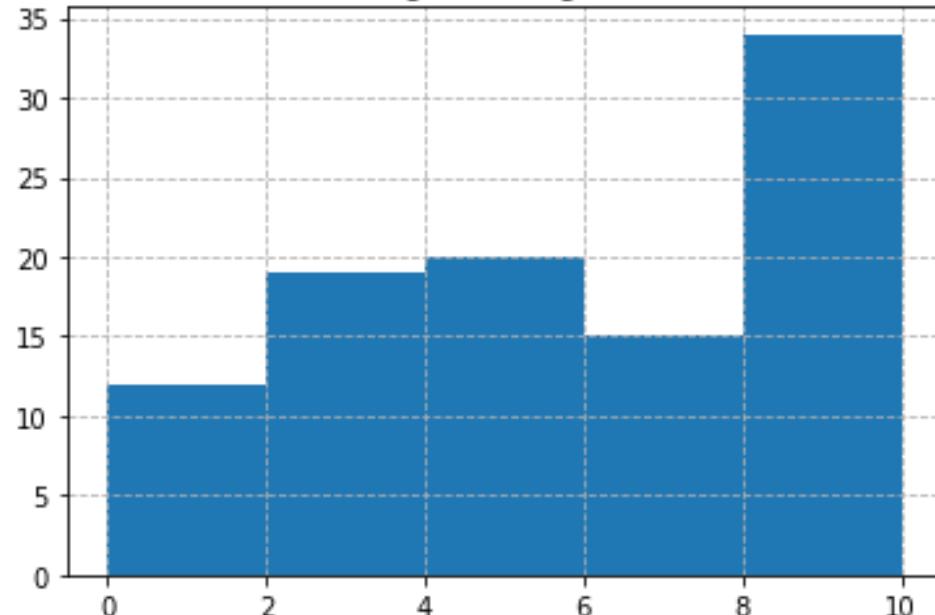
- Bảng điểm của 100 học sinh khối lớp 5

```
[ 3,  0,  5,  6,  9,  5,  3,  2,  4,  5]
[ 8,  7,  8,  0,  1,  8,  8,  8,  3,  9]
[ 4,  8,  4,  7,  2,  9,  7,  4,  7,  10]
[ 0, 10,  7,  5,  0,  5,  5,  4,  1,  8]
[10,  9,  0,  2,  3,  8,  8,  7,  7,  6]
[ 3,  9,  6,  9,  5,  5, 10,  5, 10,  5]
[ 1,  9, 10,  2,  5, 10,  3,  1,  2,  7]
[ 4,  9,  9,  2,  3,  3,  2,  6,  9,  5]
[ 9,  0,  6,  9,  4,  9, 10,  8,  4,  3]
[ 1,  2, 10,  7,  8,  8,  1,  2,  3,  6]
```

Bins = 5

| Điểm | 0-<2 | 2-<4 | 4-<6 | 6-<8 | 8-10 | |
|------|------|------|------|------|------|-----|
| n | 12 | 19 | 20 | 15 | 34 | 100 |
| P(n) | 0.12 | 0.19 | 0.20 | 0.15 | 0.34 | 1 |

Biểu đồ Histogram Bảng điểm (bins=5)



Biểu đồ tần suất (Histogram chart)



VINBIGDATA VINGROUP

Academy Vietnam

Phân phối chuẩn: Biểu đồ có dạng hình chuông. Tần suất xuất hiện nhiều nhất ở trung tâm và giảm dần về hai phía (hai phía có dạng đối xứng).

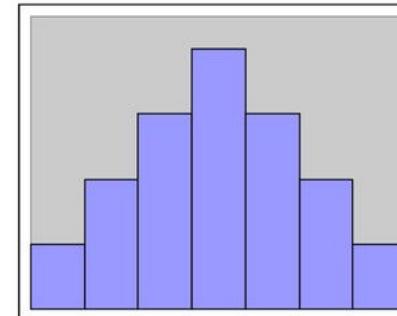
Phân phối đồng nhất là dạng phân bố mà ở đó tần suất xuất hiện của các giá trị là như nhau, không có đỉnh. Trông giống như một hình chữ nhật → phân phối hình chữ nhật.

Phân phối hai đỉnh: Biểu đồ thu được trông giống như lưng của một con lắc đà 2 biếu. Tần suất xuất hiện tại trung tâm thấp hơn các khoảng lân cận.

Phân phối lệch: là dạng phân bố không cân xứng. Giá trị trung bình của đồ thị bị lệch về bên trái hoặc bên phải tạo nên hình dáng không cân xứng.

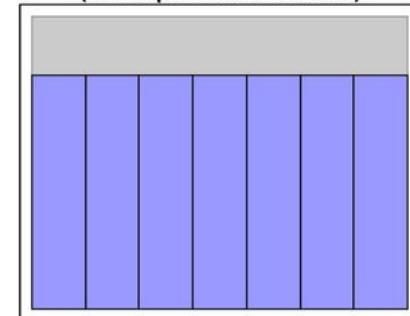
Types of Histograms

Symmetrical



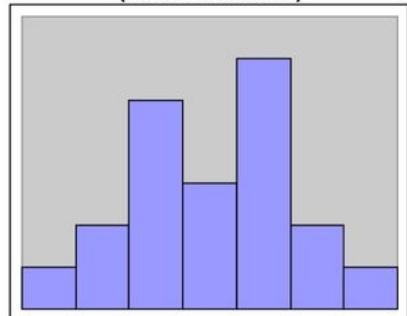
Uniform

(all frequencies the same)



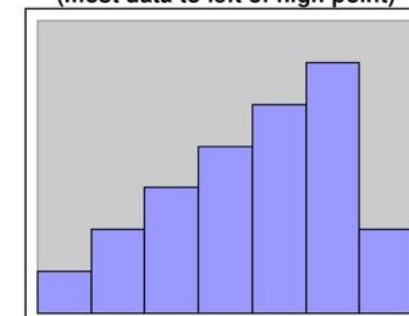
Bimodal

(no visible trend)



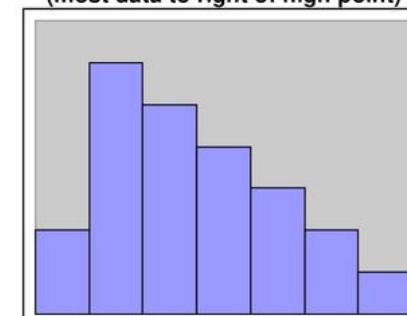
Skewed Left

(most data to left of high point)



Skewed Right

(most data to right of high point)



Histogram chart với Matplotlib

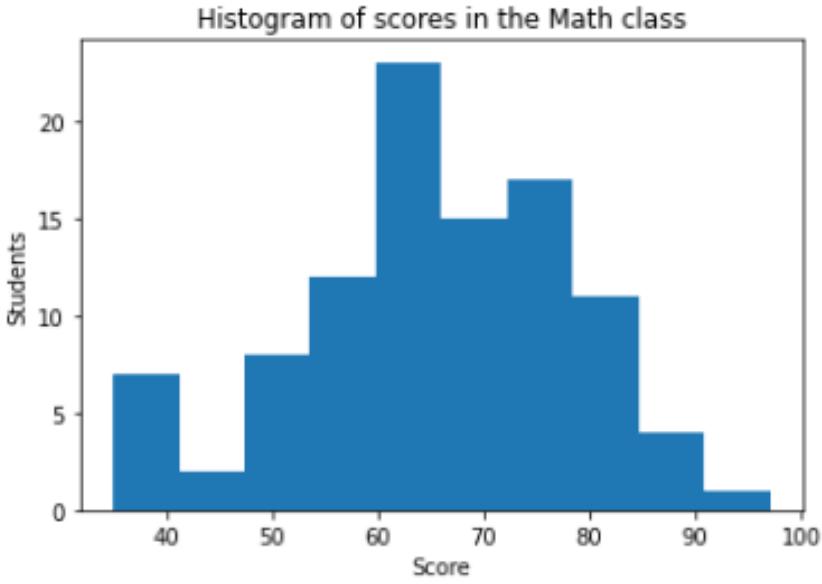


VINBIGDATA VINGROUP

Academy
Vietnam

Tập dữ liệu **Data_score.csv** lưu trữ dữ liệu 3 môn
Math, Science, History của 500 học sinh

matplotlib



Plot a Histogram

| | A | B | C |
|----|------|---------|---------|
| 1 | Math | Science | History |
| 2 | 7 | 10 | 4 |
| 3 | 4 | 9 | 3 |
| 4 | 4 | 8 | 2 |
| 5 | 4 | 8 | 3 |
| 6 | 4 | 6 | 3 |
| 7 | 6 | 8 | 1 |
| 8 | 5 | 10 | 7 |
| 9 | 5 | 10 | 5 |
| 10 | 5 | 7 | 4 |
| 11 | 6 | 8 | 3 |
| 12 | 7 | 9 | 3 |
| 13 | 6 | 8 | 2 |
| 14 | 5 | 10 | 7 |
| 15 | 5 | 9 | 6 |
| 16 | 7 | 5 | 2 |
| 17 | 8 | 10 | 3 |
| 18 | 4 | 9 | 7 |
| 19 | 5 | 7 | 2 |
| 20 | 5 | 9 | 6 |

Biểu đồ tần suất (Histogram chart)

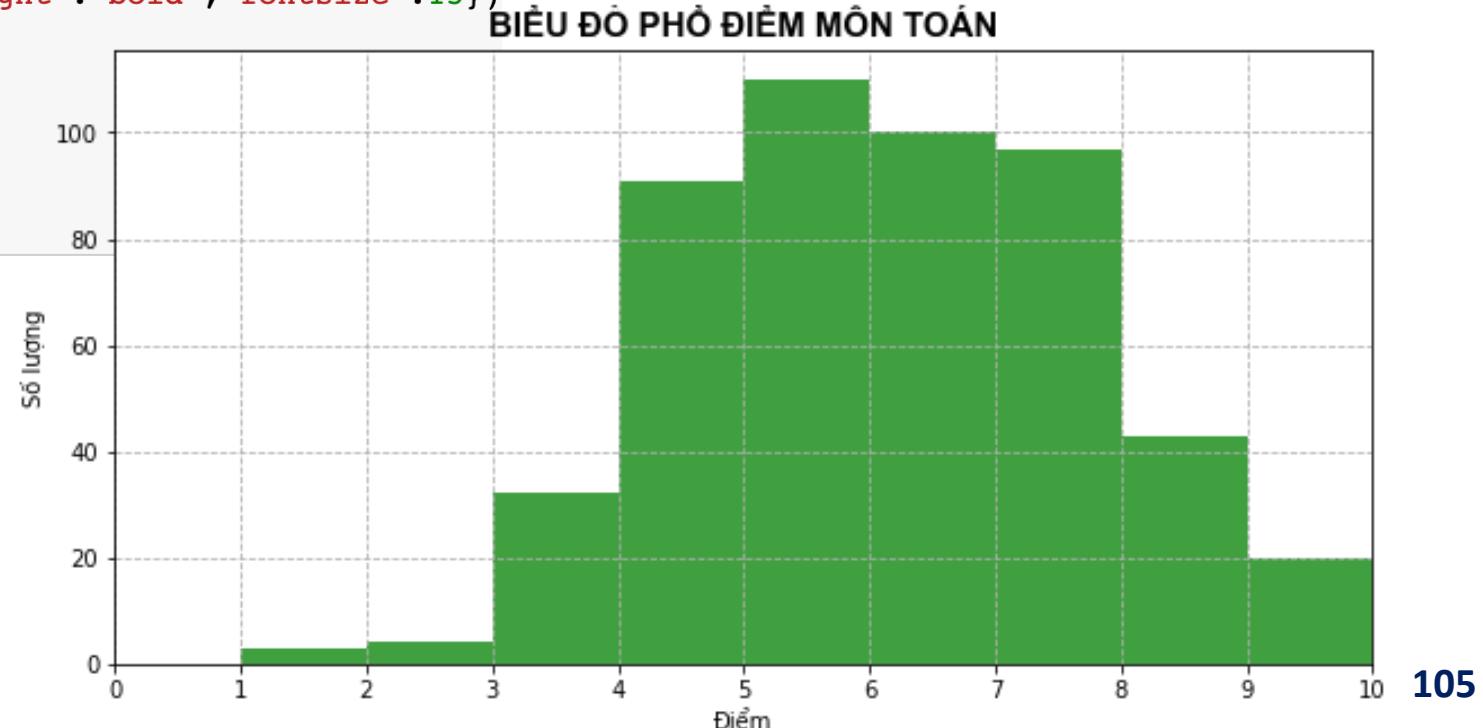


VINBIGDATA VINGROUP

Academy
Vietnam

Cú pháp: plt.hist (x)

```
1 plt.figure(figsize = (10,5))
2 #Vẽ biểu đồ Histogram:
3 plt.hist(data['Math'],      #Dữ liệu thống kê tần suất
4           color='g',        #Màu của biểu đồ
5           alpha=0.75,       #Độ trong suốt của biểu đồ
6           bins=9)          #Số lượng class muốn thống kê
7
8 plt.title('BIỂU ĐỒ PHỒ ĐIỂM MÔN TOÁN',
9            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
10 plt.grid(ls='--')
11 plt.xlabel('Điểm')
12 plt.ylabel('Số lượng')
13 plt.xlim([0,10])
14 plt.xticks([0,1,2,3,4,5,6,7,8,9,10])
15 plt.show()
```



Biểu đồ tần suất (Histogram chart)

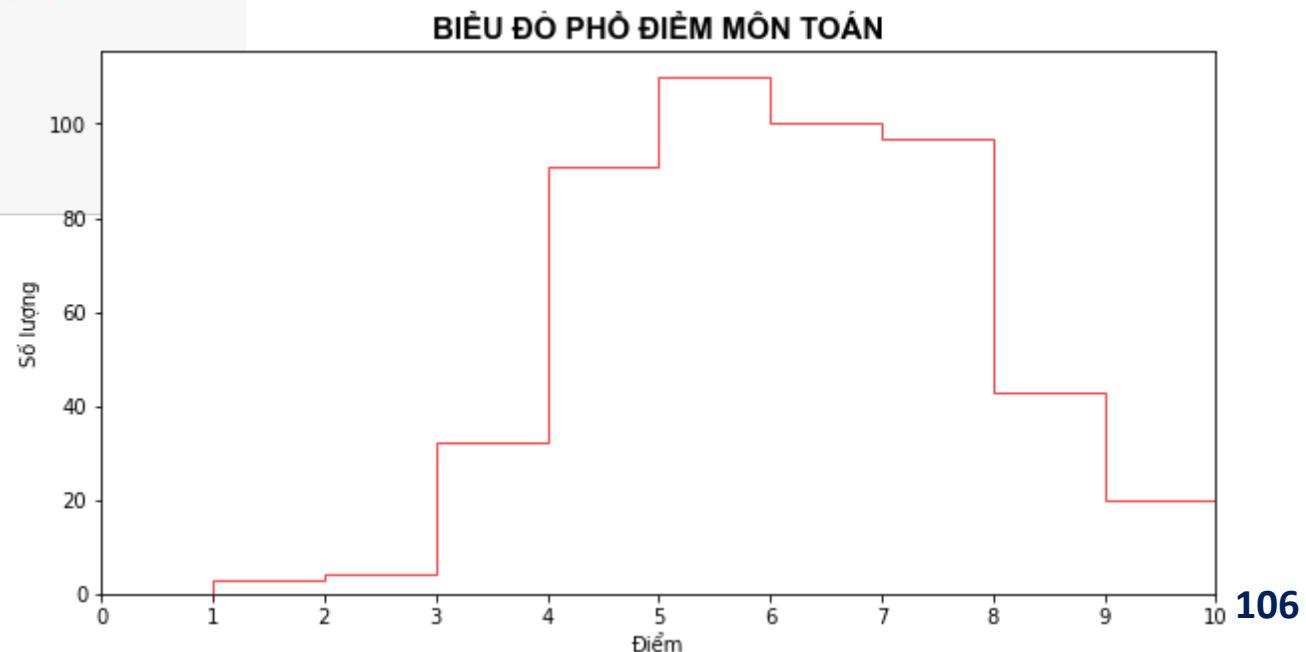


VINBIGDATA VINGROUP

Academy Vietnam

Cú pháp: plt.hist (x)

```
1 # Tùy chỉnh Histogram:  
2 plt.figure(figsize = (10,5))  
3 #Vẽ biểu đồ Histogram:  
4 plt.hist(data['Math'],  
5           alpha=0.75,  
6           bins=9,  
7           color='r',  
8           histtype='step') #Kiểu biểu diễn Histogram  
9  
10 plt.title('BIỂU ĐỒ PHỒ ĐIỂM MÔN TOÁN',  
11            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})  
12 #plt.grid(ls='--')  
13 plt.xlabel('Điểm')  
14 plt.ylabel('Số lượng')  
15 plt.xlim([0,10])  
16 plt.xticks([0,1,2,3,4,5,6,7,8,9,10])  
17 plt.show()
```



Biểu đồ tần suất (Histogram chart)

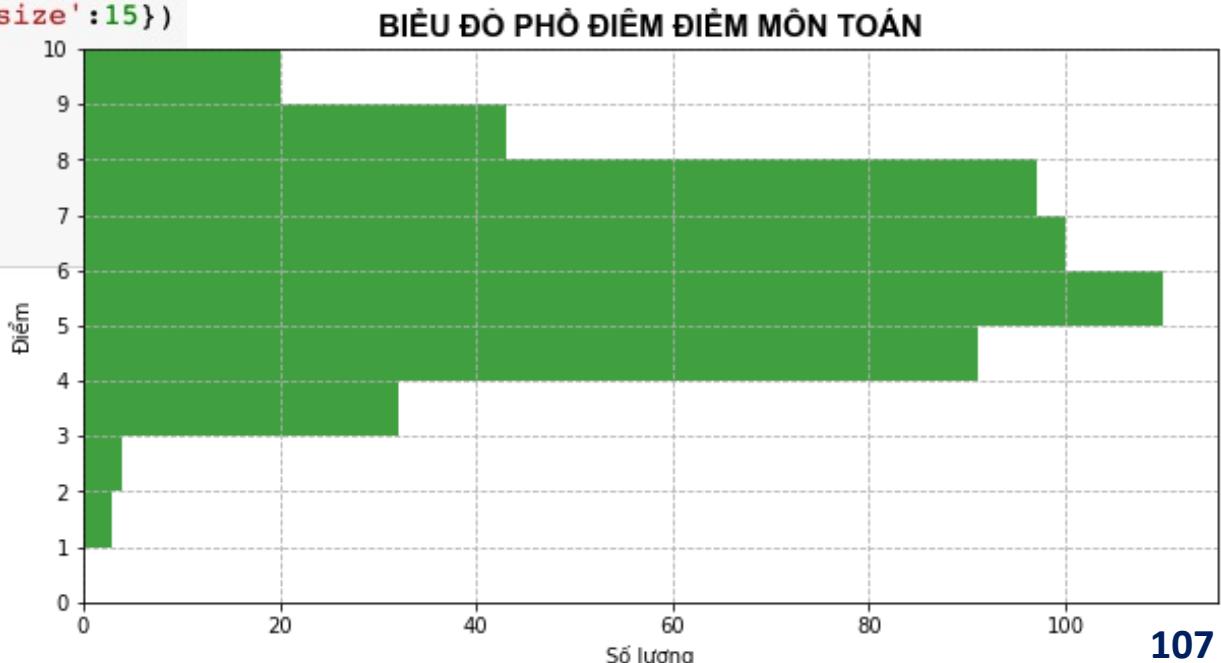


VINBIGDATA VINGROUP

Academy
Vietnam

Cú pháp: plt.hist (x)

```
1 plt.figure(figsize = (10,5))
2 #Vẽ biểu đồ Histogram:
3 plt.hist(data['Math'],
4           facecolor='g',
5           alpha=0.75,
6           bins=9,
7           orientation='horizontal') #Biểu diễn Histogram nằm ngang
8
9
10 plt.title('BIỂU ĐỒ PHỒ ĐIỂM MÔN TOÁN',
11            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
12 plt.grid(ls='--')
13 plt.ylabel('Điểm')
14 plt.xlabel('Số lượng')
15 plt.ylim([0,10])
16 plt.yticks([0,1,2,3,4,5,6,7,8,9,10])
17 plt.show()
```



Thiết lập bins

Biểu đồ tần suất (Histogram chart)



VINBIGDATA VINGROUP

Academy
Vietnam

```
1 plt.figure(figsize = (10,5))
2 #Vẽ biểu đồ Histogram:
3 t = plt.hist(data['Math'],
4               facecolor='g',
5               alpha=0.75,
6               bins=[1,5,10]) #Tách thành 2 nhóm theo ngưỡng thiết lập
7 #nhóm 1: 1--<5
8 #nhóm 2: 5--10
9
10 plt.title('BIỂU ĐỒ PHỒ ĐIỂM MÔN TOÁN',
11            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
12 plt.grid(ls='--')
13 plt.xlabel('Điểm')
14 plt.ylabel('Số lượng')
15 plt.xlim([0,10])
16 plt.xticks([0,1,2,3,4,5,6,7,8,9,10])
17 plt.show()
```



Biểu đồ tần suất (Histogram chart)



VINBIGDATA VINGROUP

Academy Vietnam

```
1 plt.figure(figsize = (10,5))
2 #Vẽ biểu đồ Histogram:
3 t = plt.hist(data['Math'],
4               facecolor='g',
5               alpha=0.75,
6               bins=[1,5,8,10]) #Tách thành 3 nhóm theo ngưỡng thiết lập
7 #nhóm 1: 1--<5
8 #nhóm 2: 5--<8
9 #Nhóm 3: 8--10
10
11 plt.title('BIỂU ĐỒ PHỒ ĐIỂM MÔN TOÁN',
12            fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})
13 plt.grid(ls='--')
14 plt.xlabel('Điểm')
15 plt.ylabel('Số lượng')
16 plt.xlim([0,10])
17 plt.xticks([0,1,2,3,4,5,6,7,8,9,10])
18 plt.show()
```





Bar chart & Histogram chart



Bar chart & Histogram chart

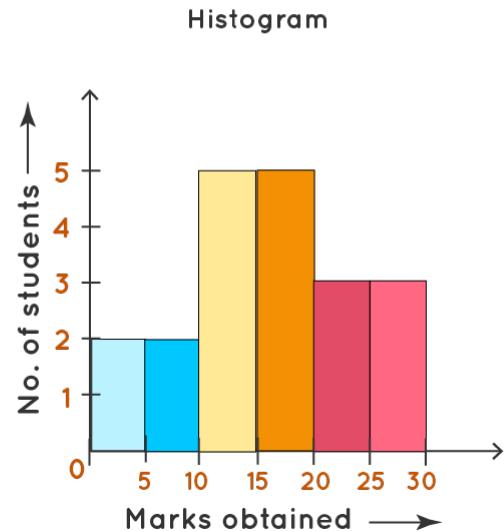
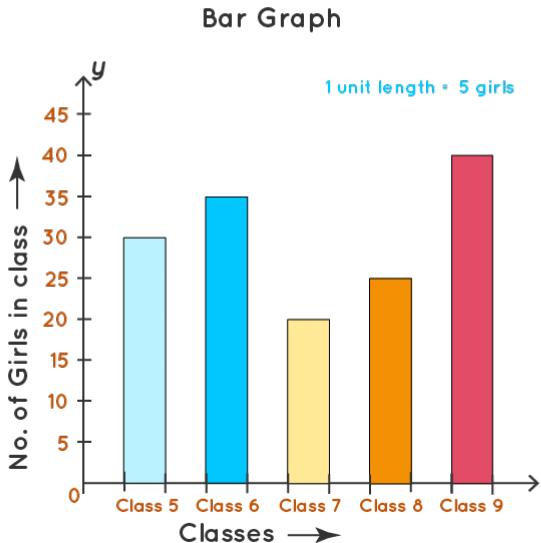


VINBIGDATA VINGROUP

Academy Vietnam

Difference Between Bar Chart and Histogram

 cuemath
THE MATH EXPERT



Bar Graph

Equal space between every two consecutive bars.

X-axis can represent anything.

Histogram

No space between two consecutive bars. They should be attached to each other.

X-axis should represent only continuous data that is in terms of numbers.

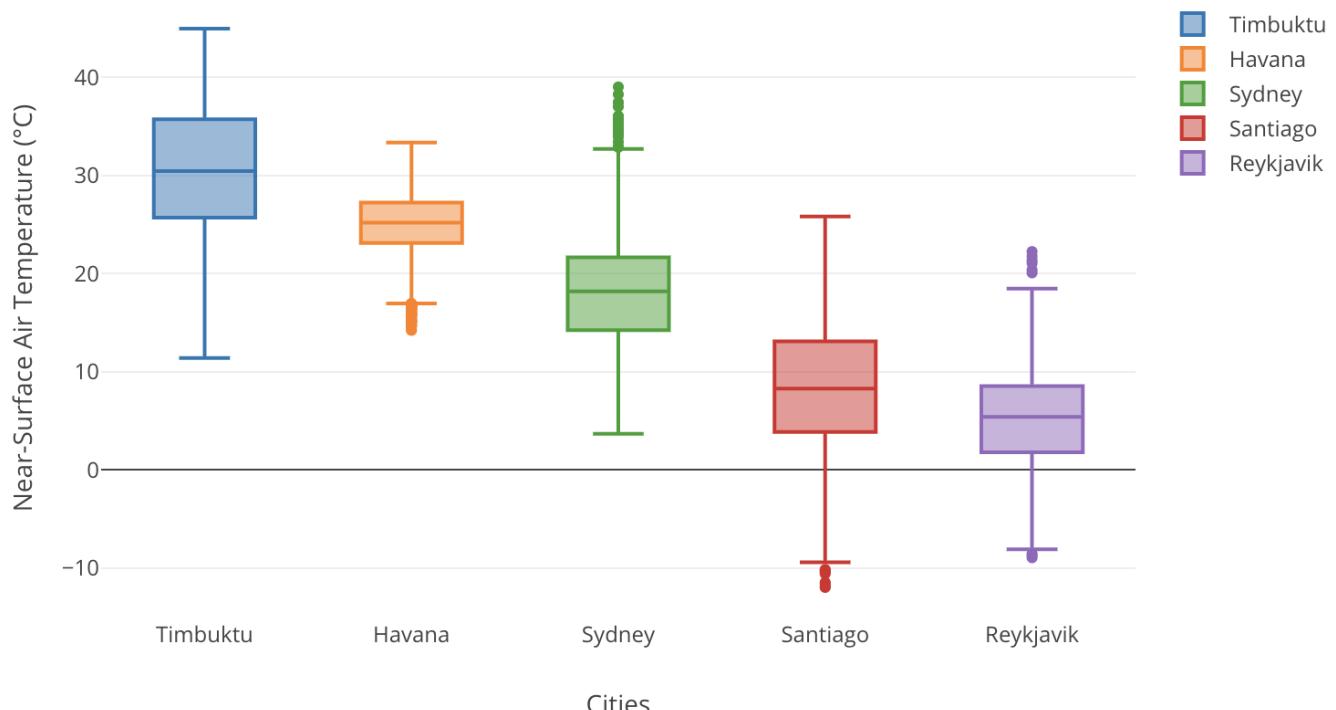
9. Biểu đồ hộp (Boxplot)

Boxplot

- **Boxplot** là một dạng biểu đồ thể hiện phân bố dữ liệu của các thuộc tính số thông qua các “*tứ phân vị*” và được giới thiệu lần đầu bởi John Tukey vào năm 1970.
- **Tứ phân vị** là một khái niệm trong thống kê dùng để mô tả sự phân bố và sự phân tán của tập dữ liệu, gồm 3 giá trị: Q1, Q2 và Q3 chia tập dữ liệu thành 4 phần bằng nhau.

Box plots

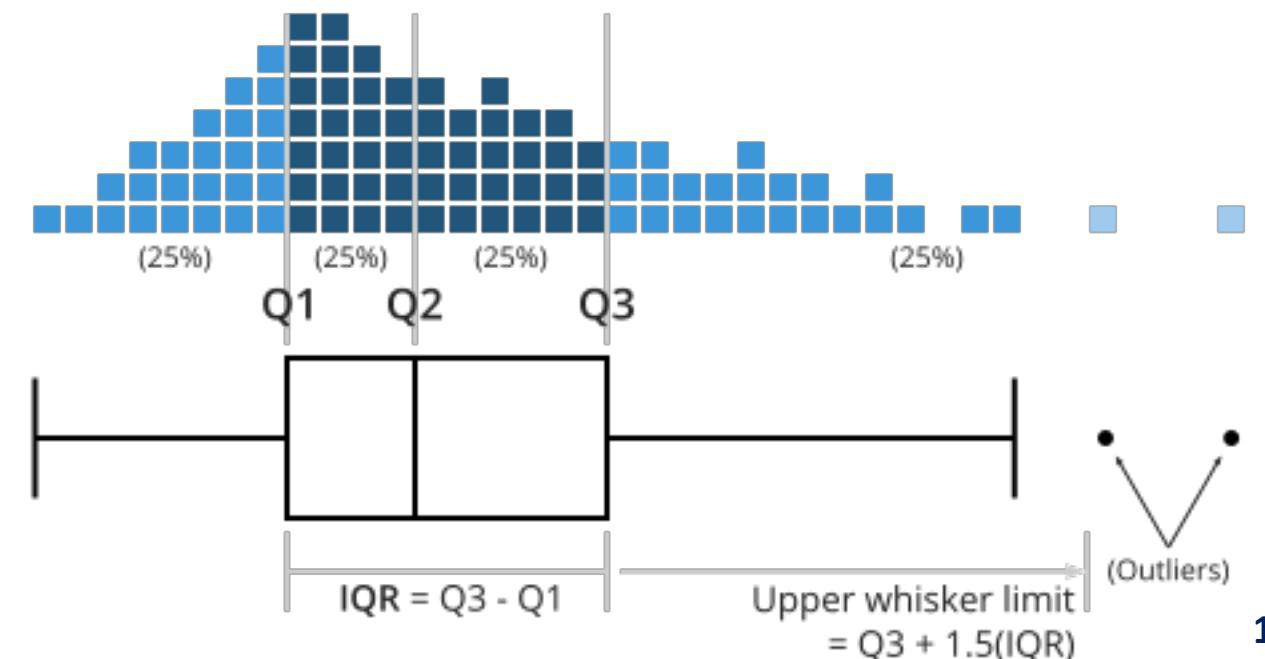
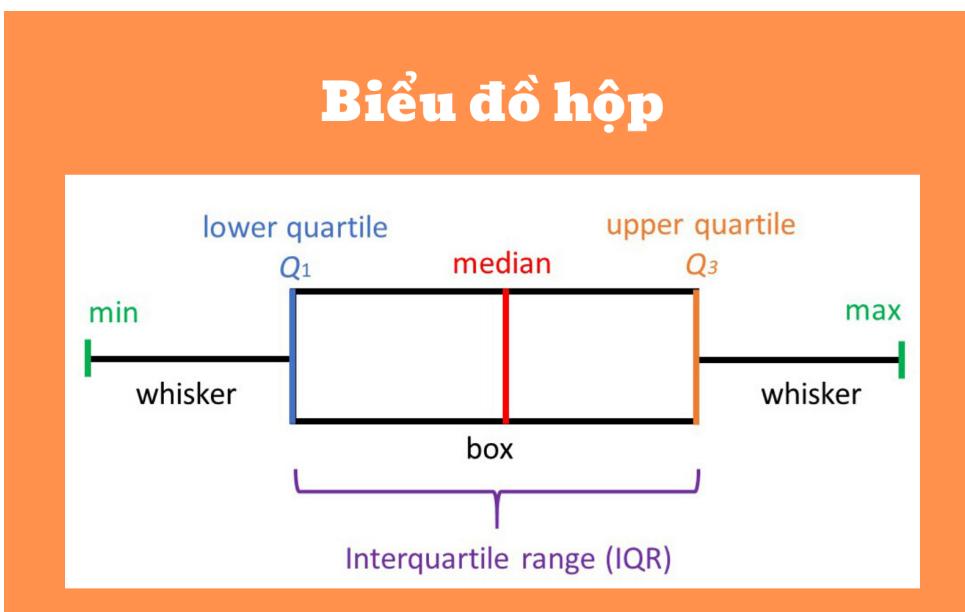
- Boxplot thể hiện các phân bố dữ liệu, nghĩa là giúp chúng ta biết được độ dàn trải của các điểm dữ liệu như thế nào, dữ liệu có đối xứng không, phân bố rộng hay hẹp, giá trị nhỏ nhất, lớn nhất và các điểm ngoại lệ.



Boxplot

Biểu đồ Boxplot thể hiện 5 thông số:

- **First quartile (Q1)**: Trung vị giữa **Median** và **phần tử nhỏ nhất** trong tập dữ liệu. Còn gọi là **25th Percentile**.
- **Median (Q2)**: Trung vị của tập dữ liệu, tức là giá trị ở phần tử giữa. (50%)
- **Third quartile (Q3)**: Trung vị giữa **Median** và **phần tử lớn nhất** trong tập dữ liệu. Còn gọi là **75th Percentile**.
- **Minimum**: Phần tử nhỏ nhất không phải ngoại lệ.
- **Maximum**: Phần tử lớn nhất không phải là ngoại lệ.

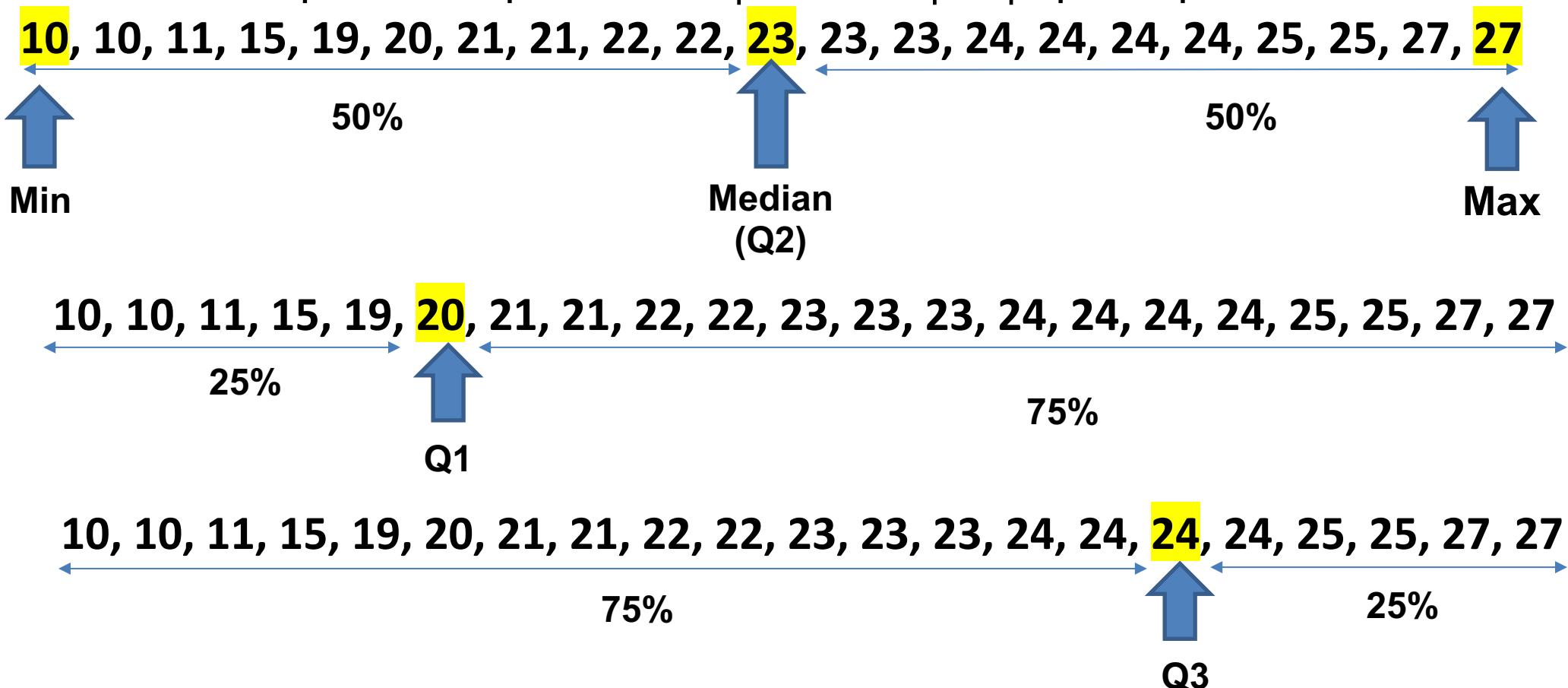


Cách xây dựng biểu đồ Boxplot

- Một nhà hàng ghi lại khoảng cách của 21 khách hàng đi từ nhà đến nhà hàng như sau:

24, 10, 23, 11, 21, 22, 23, 15, 23, 21, 20, 25, 22, 24, 24, 10, 24, 25, 27, 27, 19

- Trước tiên để tìm được các số liệu để vẽ Boxplot cần sắp xếp lại dữ liệu:



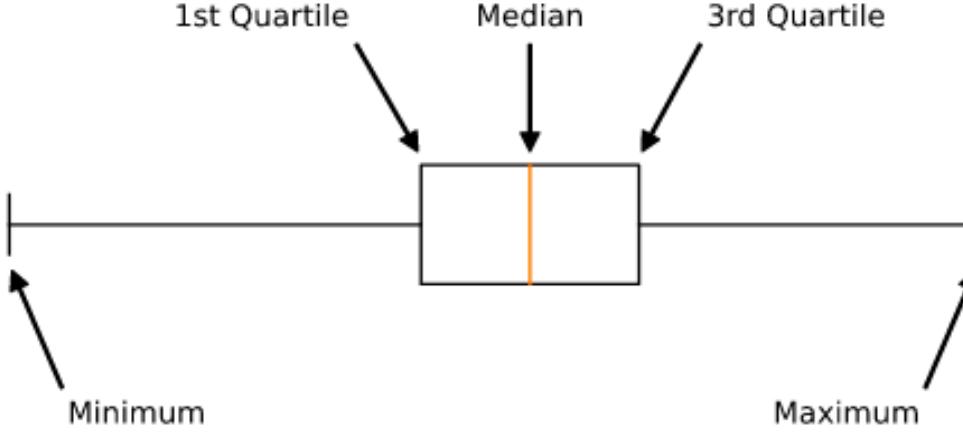
Cách xây dựng biểu đồ Boxplot



VINBIGDATA
VINGROUP

Academy
Vietnam

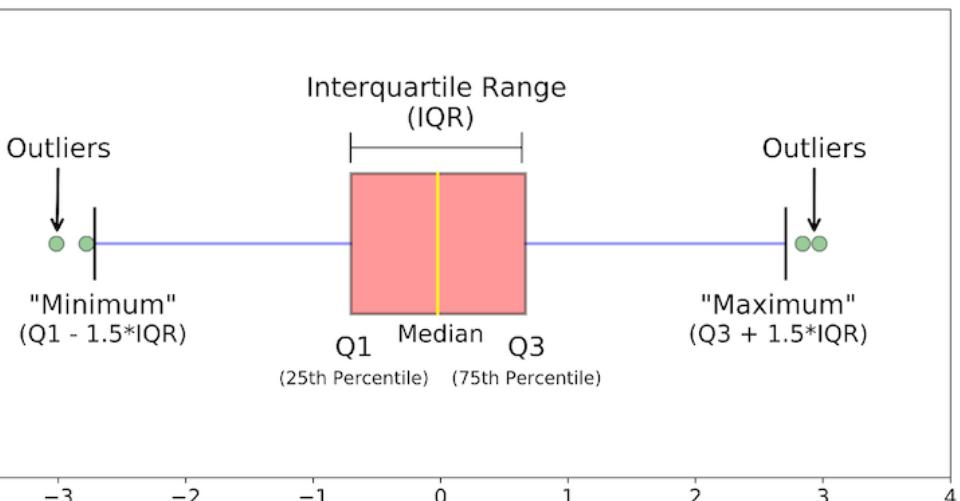
- Median = 23
- Q1 = 20
- Q3 = 24
- Min = 10
- Max = 27



Cách xây dựng biểu đồ Boxplot

- Trong thống kê, một ngoại lệ (outlier) là một điểm dữ liệu khác biệt đáng kể so với các quan sát khác. Một ngoại lệ có thể là do sự thay đổi trong phép đo hoặc là lỗi và thông thường được loại trừ khỏi tập dữ liệu bởi nó có thể gây ra vấn đề nghiêm trọng trong phân tích thống kê.
- Để tìm ngoại lệ, ta dùng thêm khái niệm **IQR**. **IQR (Interquartile Range)** là một khái niệm trong thống kê mô tả, dùng đo lường độ phân tán của dữ liệu và được tính toán bằng công thức:

$$\text{IQR} = Q_3 - Q_1$$



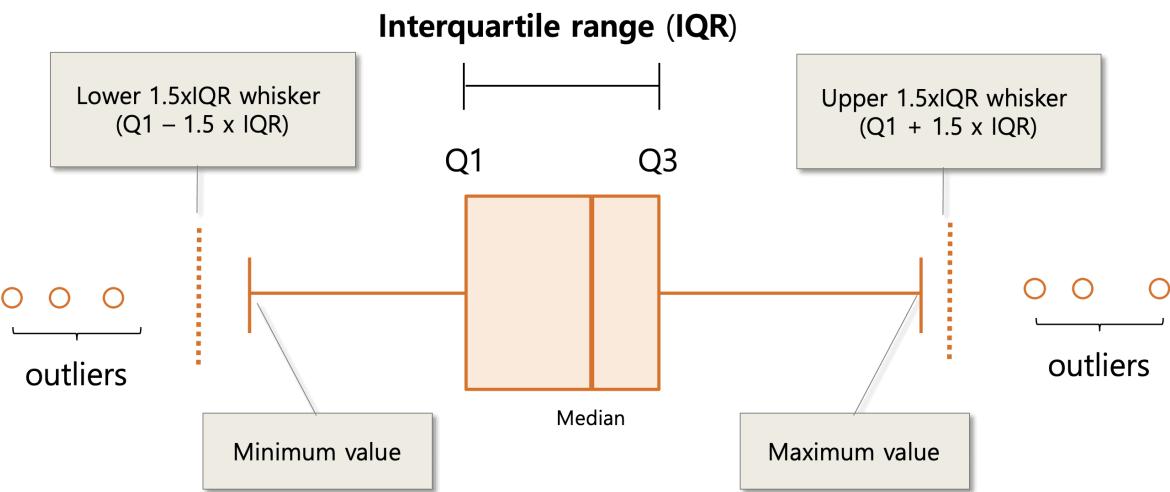
Điểm ngoại lệ sẽ là những điểm:
nhỏ hơn $< Q_1 - 1.5 * \text{IQR}$ và lớn hơn $> Q_3 + 1.5 * \text{IQR}$.

Cách xây dựng biểu đồ Boxplot

- $Q1 = 20$
- $Q3 = 24$ $\rightarrow IQR = Q3 - Q1 = 24 - 20 = 4$

Như vậy ta xác định được **Minimum** mới và **Maximum** mới như sau:

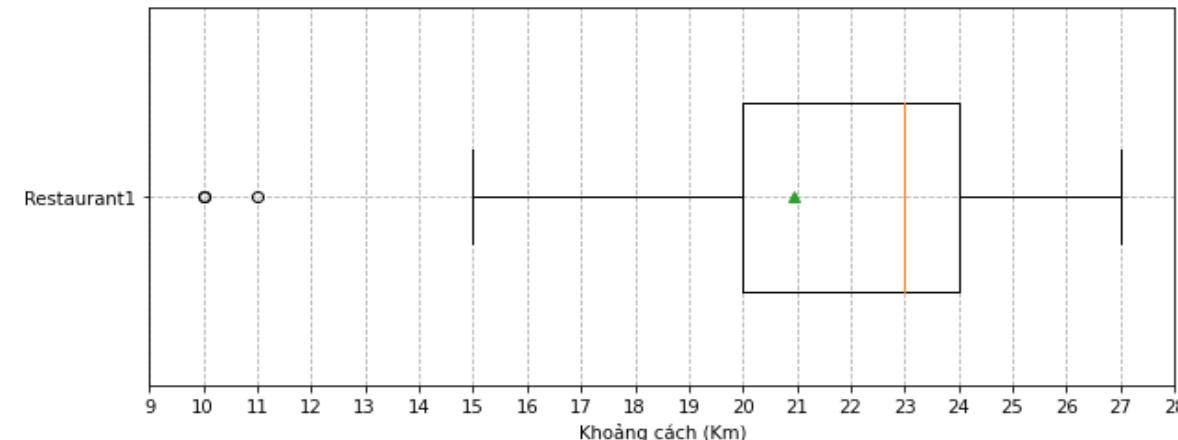
- $Q1 - 1.5 \times IQR = 20 - 1.5 \times 4 = 14$ (Trong dữ liệu giá trị nhỏ nhất ≥ 14 gần nhất là 15 $\rightarrow \text{Minimum} = 15$)
- $Q3 + 1.5 \times IQR = 24 + 1.5 \times 4 = 30$ (Trong dữ liệu giá trị lớn nhất ≤ 30 $\rightarrow \text{Maximum} = 27$)



10, 10, 11, **15**, 19, 20, 21, 21, 22, 22, 23, 23, 23, 24, 24, 24, 24, 25, 25, 27, **27**

↑
Minimum

↑
Maximum



Cách xây dựng biểu đồ Boxplot

Một nhà hàng ghi lại độ tuổi của 37 khách hàng đến nhà hàng như sau:

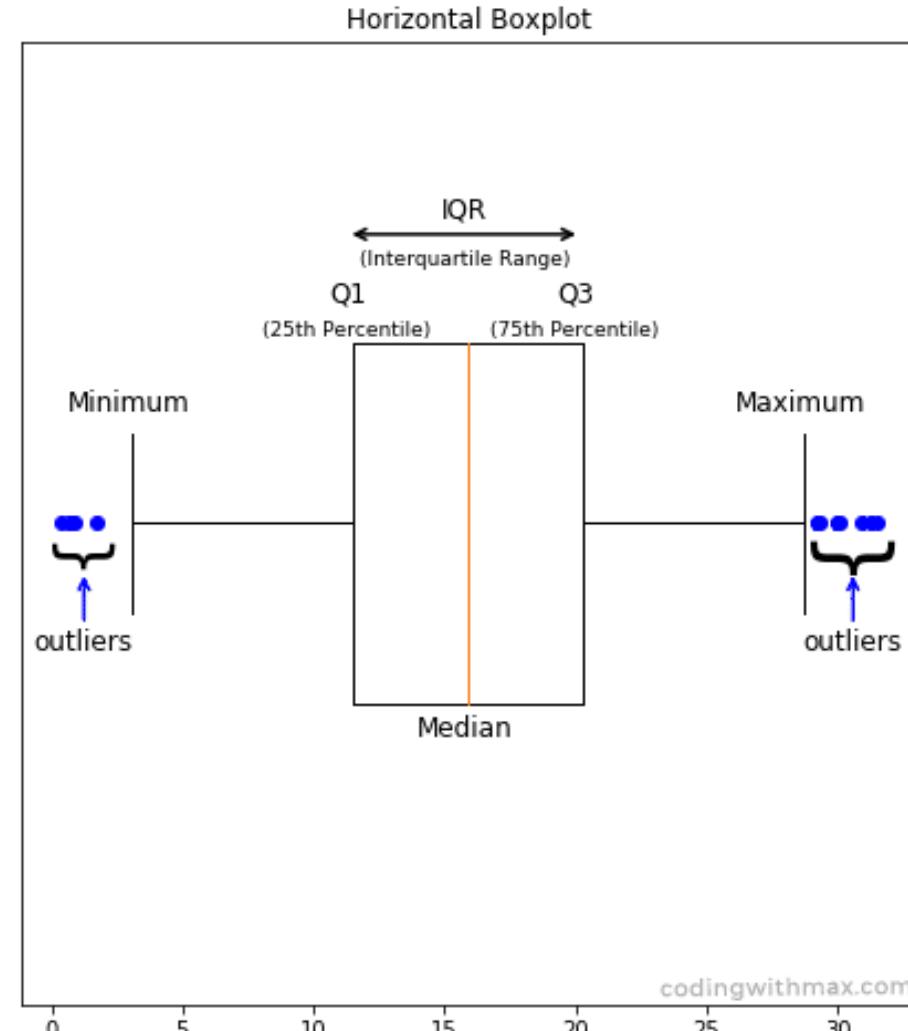
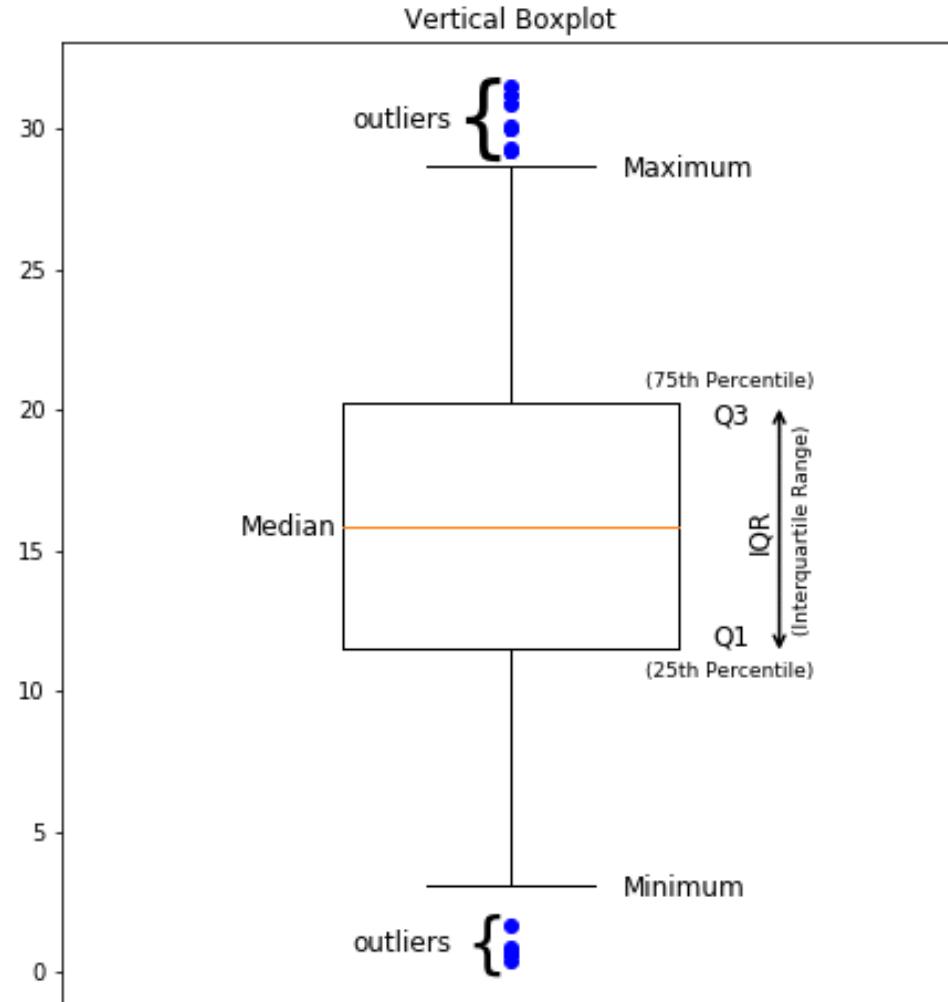
**24,32,40,26,23,31,39,21,22,23,35,23,21,20,8,22,24,24,40,24,25,27,37,25,24,28,33,29,
30,50,30,31,38,23,38,27,29**

Thực hiện Vẽ biểu đồ Boxplot cho dữ liệu trên?

BEST
PRACTICE



Các loại Biểu đồ Boxplot



Boxplot với Matplotlib

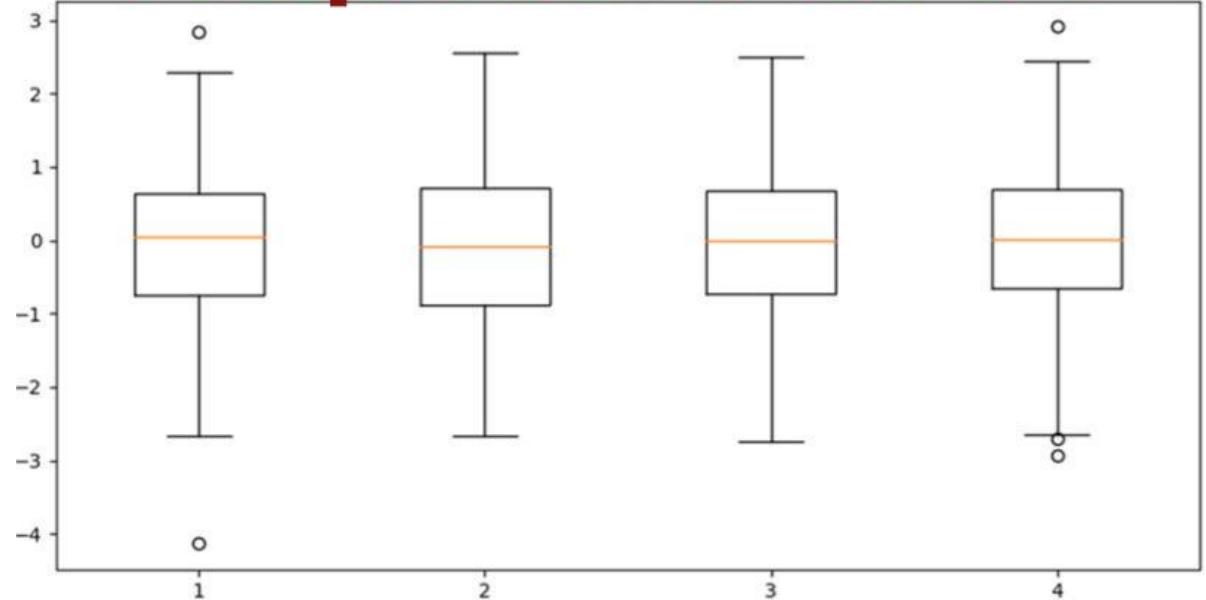


VINBIGDATA VINGROUP

Academy
Vietnam

Tập dữ liệu `Data_score.csv` lưu trữ dữ liệu 3 môn
Math, Science, History của 500 học sinh

Matplotlib BoxPlot



| | A | B | C |
|----|------|---------|---------|
| 1 | Math | Science | History |
| 2 | 7 | 10 | 4 |
| 3 | 4 | 9 | 3 |
| 4 | 4 | 8 | 2 |
| 5 | 4 | 8 | 3 |
| 6 | 4 | 6 | 3 |
| 7 | 6 | 8 | 1 |
| 8 | 5 | 10 | 7 |
| 9 | 5 | 10 | 5 |
| 10 | 5 | 7 | 4 |
| 11 | 6 | 8 | 3 |
| 12 | 7 | 9 | 3 |
| 13 | 6 | 8 | 2 |
| 14 | 5 | 10 | 7 |
| 15 | 5 | 9 | 6 |
| 16 | 7 | 5 | 2 |
| 17 | 8 | 10 | 3 |
| 18 | 4 | 9 | 7 |
| 19 | 5 | 7 | 2 |
| 20 | 5 | 9 | 6 |

Boxplot với Matplotlib

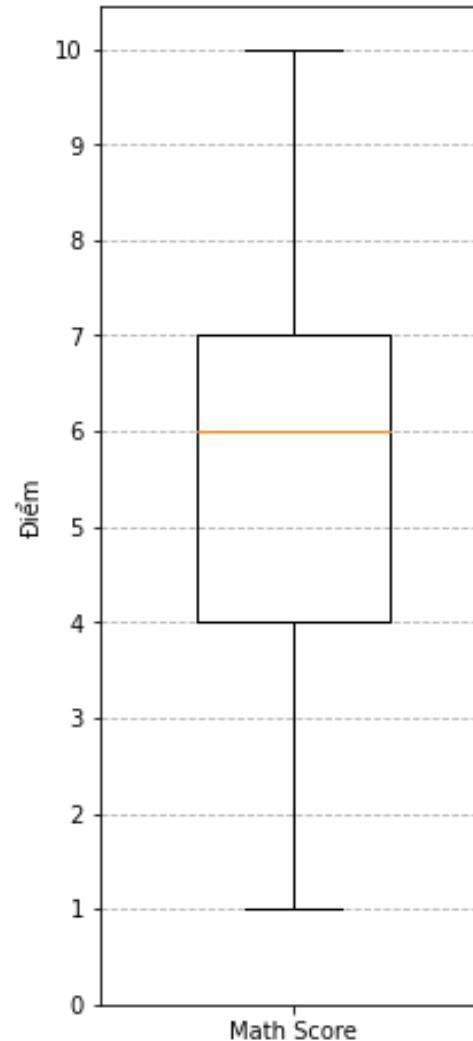


VINBIGDATA VINGROUP

Academy
Vietnam

Cú pháp: plt.boxplot (x)

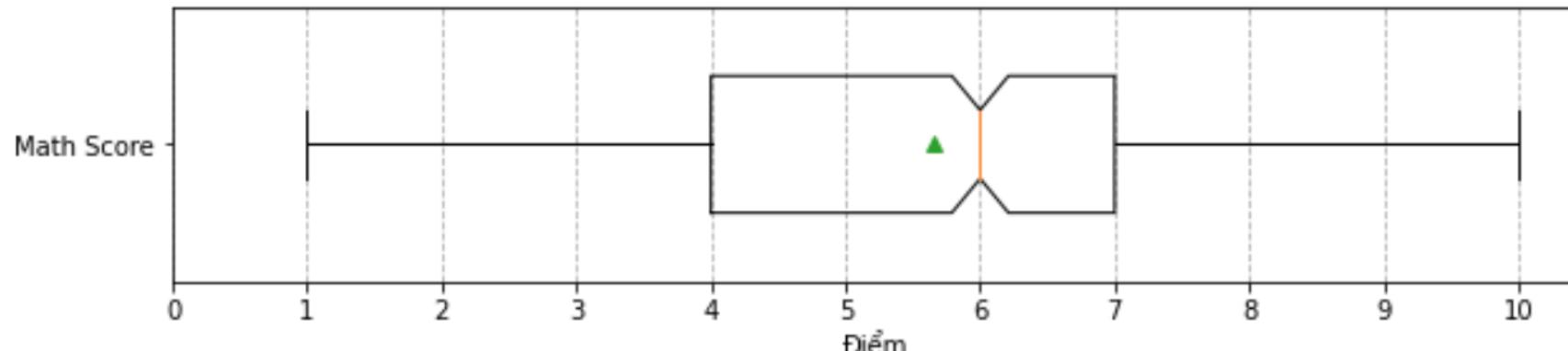
```
1 plt.figure(figsize = (3,8))
2 #Vẽ biểu đồ hộp Boxplot:
3 plt.boxplot(data[ 'Math' ],
4             widths=[0.5],
5             labels=[ 'Math Score' ]) #Nhãn của Boxplot
6
7 plt.grid( axis = 'y', ls='--' )
8 plt.ylabel('Điểm')
9 plt.yticks([0,1,2,3,4,5,6,7,8,9,10])
10 plt.show()
```



Boxplot với Matplotlib

Cú pháp: plt.boxplot (x)

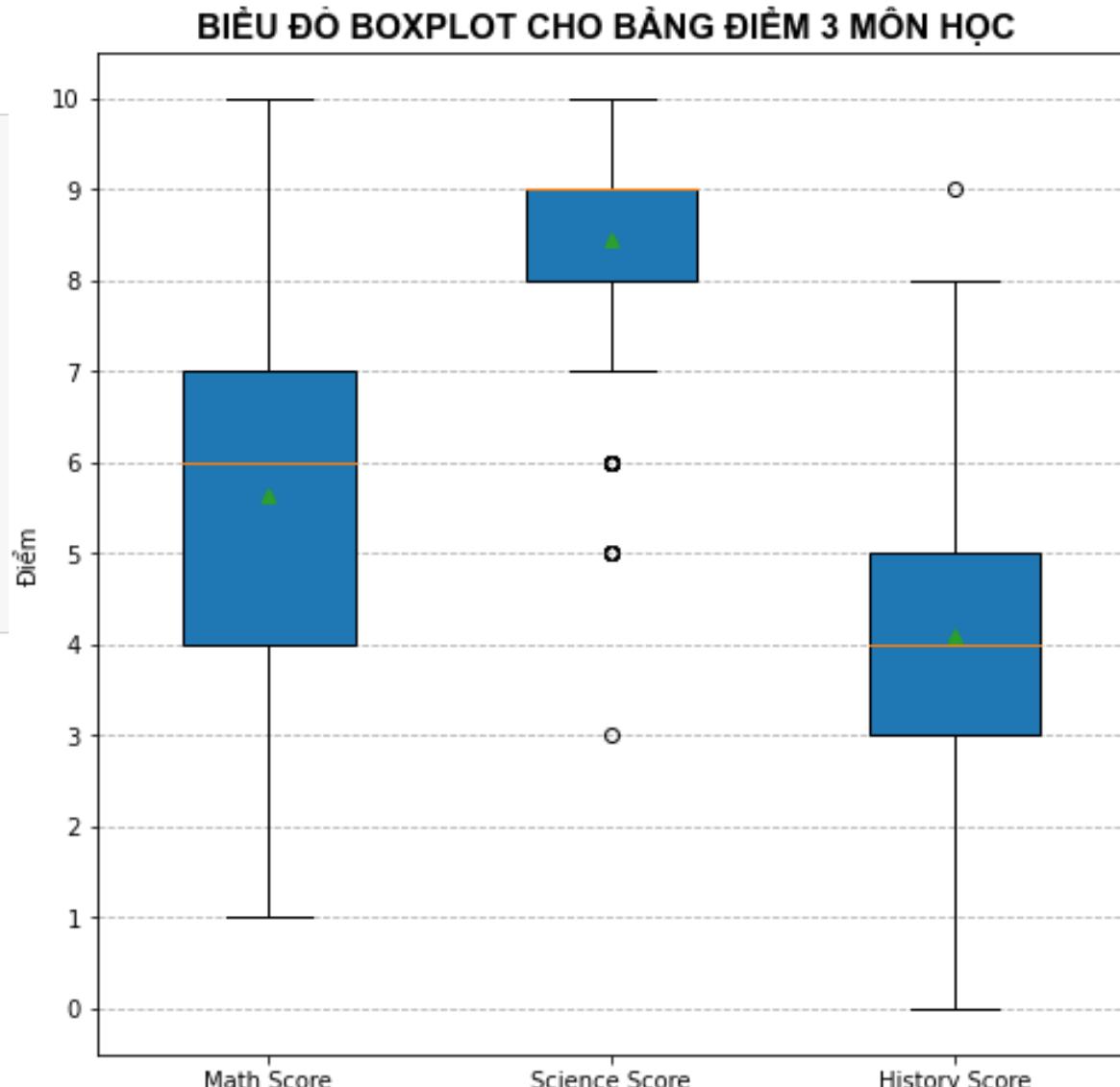
```
1 plt.figure(figsize = (10,2))
2 #Vẽ biểu đồ hộp Boxplot:
3 plt.boxplot(data['Math'],
4             labels=['Math Score'],
5             widths=[0.5],
6             notch=True,           #Tạo thóp tại vị trí Median
7             vert=False,          #Hiển thị boxplot theo chiều ngang False / chiều dọc True
8             showmeans=True)      #Hiển thị vị trí có giá trị trung bình (mean)
9
10 plt.grid( axis = 'x', ls='--')
11 plt.xlabel('Điểm')
12 plt.xticks([0,1,2,3,4,5,6,7,8,9,10])
13 plt.show()
```



Boxplot với Matplotlib

Cú pháp: plt.boxplot (x)

```
1 #Multiple boxplot trên cùng một plot:  
2 plt.figure(figsize = (8,8))  
3 plt.boxplot([data['Math'],data['Science'],data['History']],  
4             labels=['Math Score','Science Score','History Score'],  
5             widths=[0.5,0.5,0.5],  
6             showmeans=True,  
7             patch_artist=True)#Đỗ màu cho hộp  
8  
9 plt.title('BIỂU ĐỒ BOXPLOT CHO BẢNG ĐIỂM 3 MÔN HỌC ',  
10           fontdict={'fontname':'Arial','fontweight':'bold','fontsize':15})  
11 plt.grid( axis = 'y', ls='--')  
12 plt.ylabel('Điểm')  
13 plt.yticks([0,1,2,3,4,5,6,7,8,9,10])  
14  
15 plt.show()
```





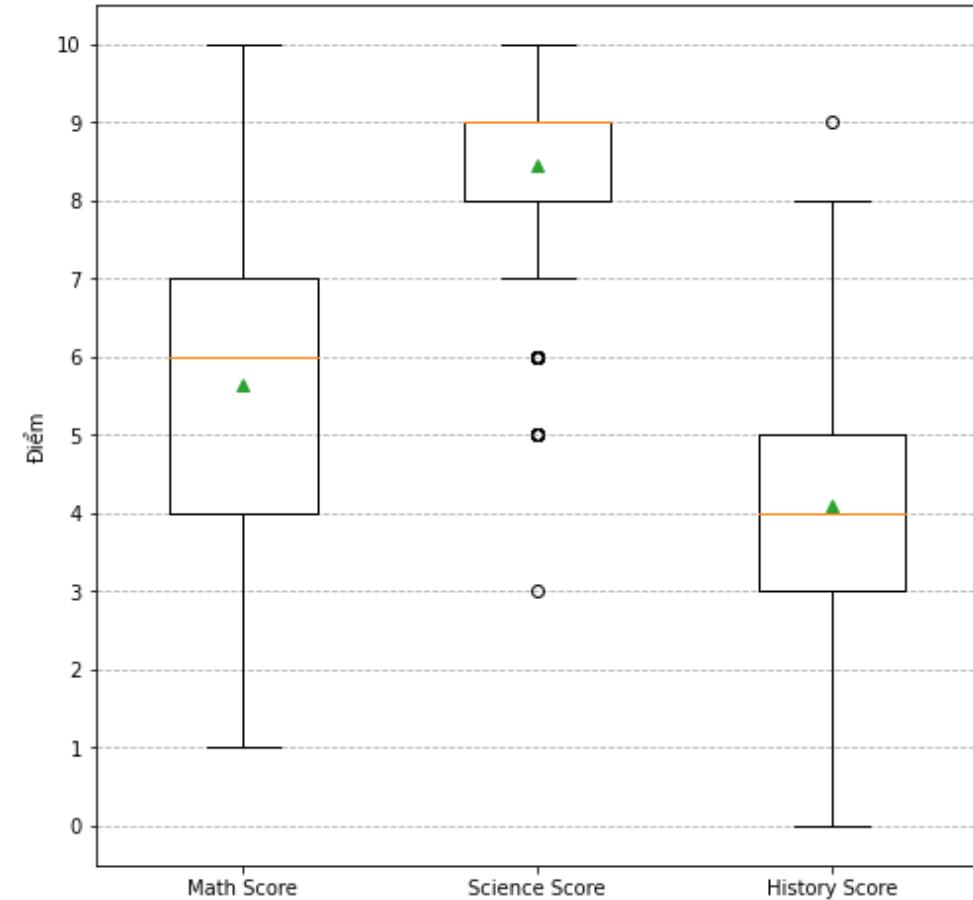
a. Boxplot sử dụng khi nào?



Boxplot sử dụng khi nào?

Khi muốn nhanh chóng có được cái nhìn tổng quan về dữ liệu. Boxplot cung cấp thông tin chung về một nhóm dữ liệu đối xứng? độ lệch, phương sai và giá trị ngoại lai? Có thể dễ dàng thấy phần lớn dữ liệu chính ở đâu.

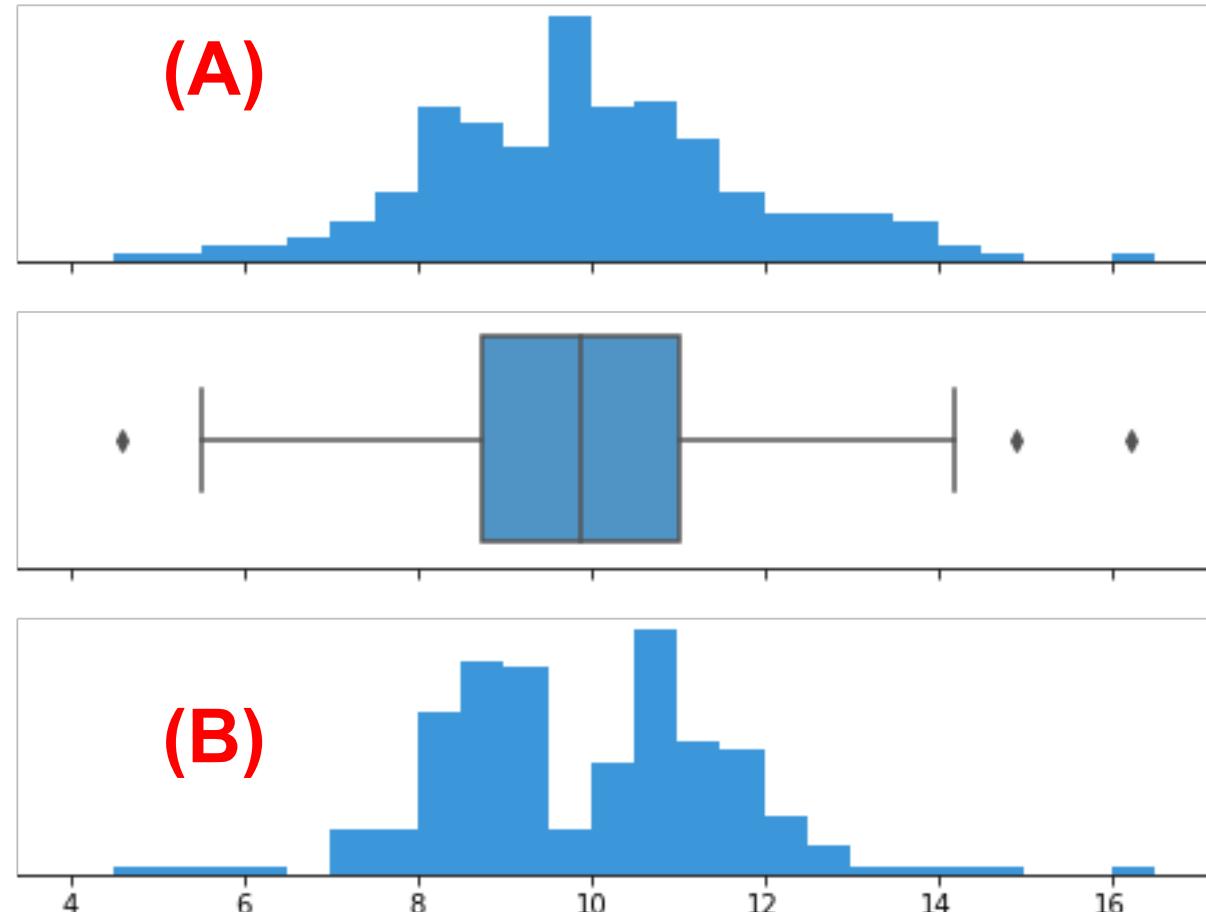
Boxplot được sử dụng tốt nhất khi cần phải thực hiện so sánh phân bố giữa các nhóm. Chúng nhỏ gọn trong việc tóm tắt dữ liệu và dễ dàng so sánh các nhóm thông qua vị trí của hộp và râu.



Boxplot sử dụng khi nào?

Hạn chế của Boxplot chính là sự đơn giản của biểu đồ, Boxplot không thể hiện được chi tiết về phân bố của dữ liệu.

Ví dụ như hình bên, Hai bộ dữ liệu (A) và (B) có phân bố dữ liệu khác nhau rất nhiều nhưng khi trực quan bằng Boxplot đều thu được một biểu đồ như nhau → Boxplot Không thể hiện được chi tiết của phân bố dữ liệu.

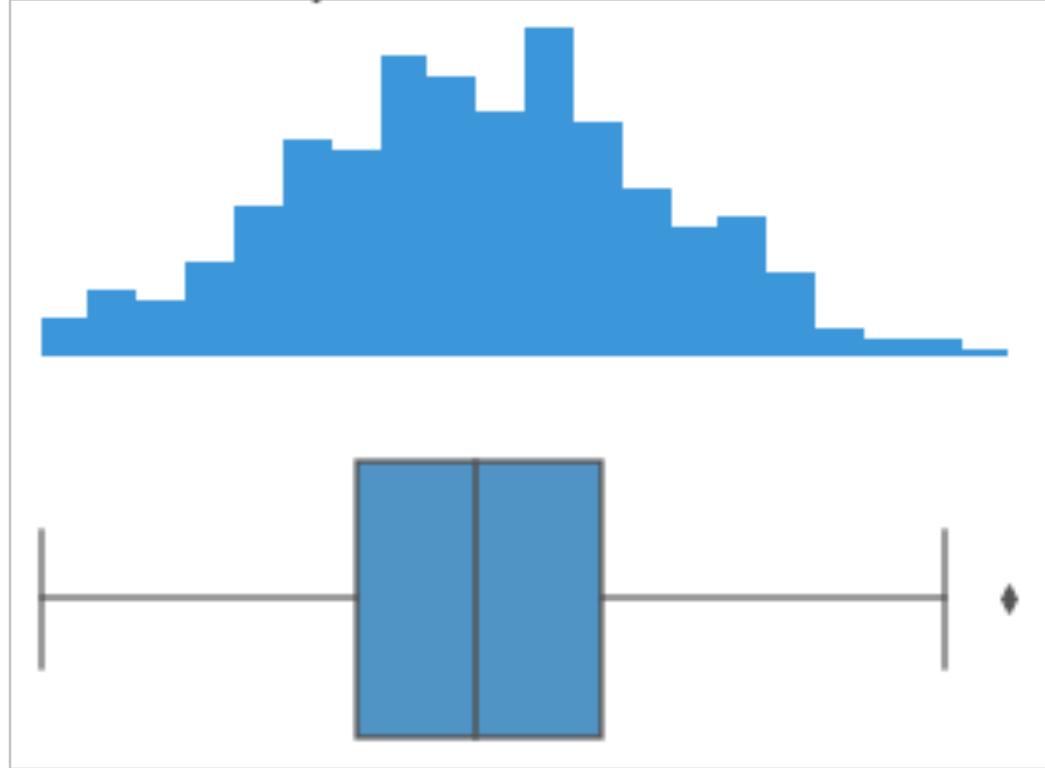




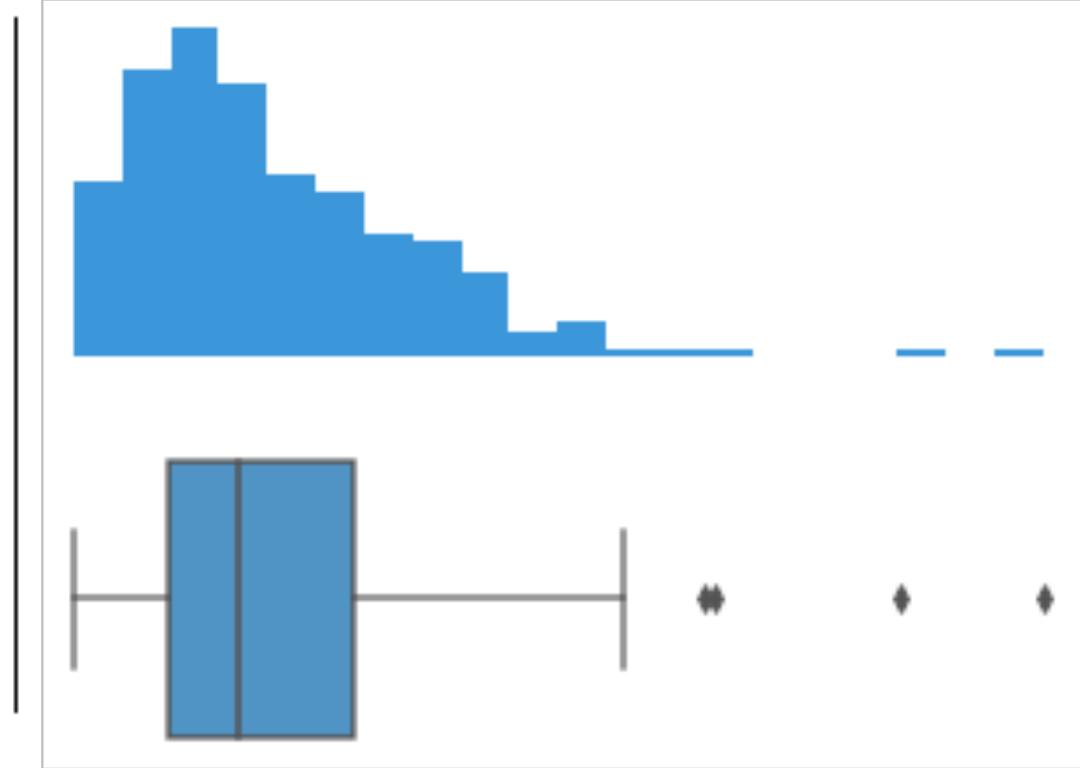
b. Boxplot & Histogram chart

Boxplot & Histogram chart

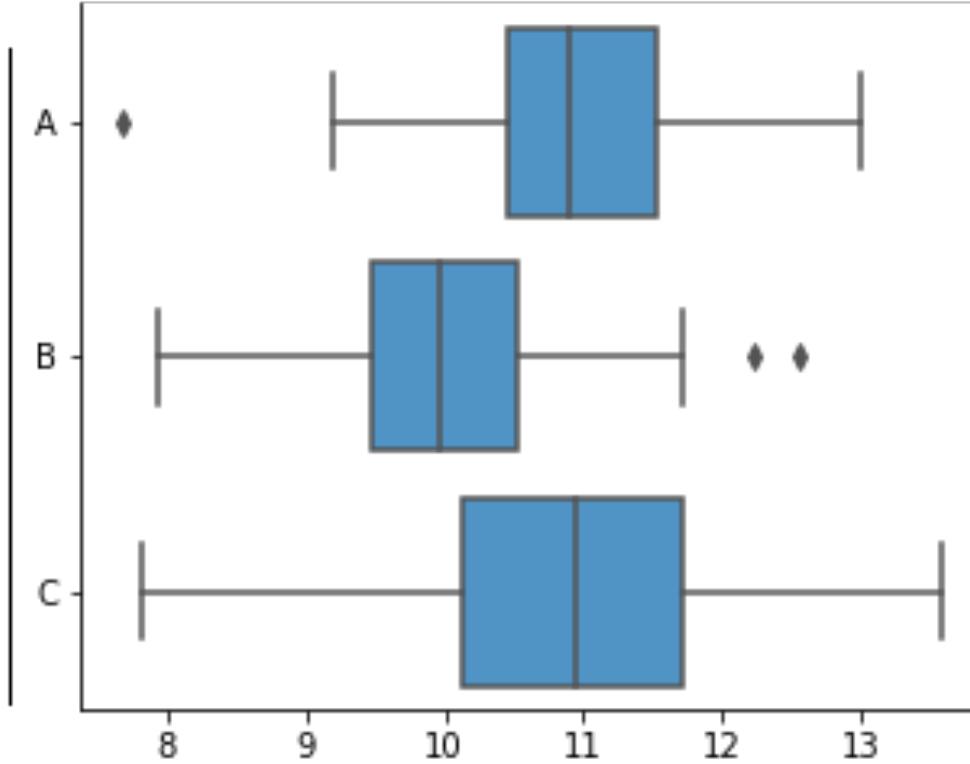
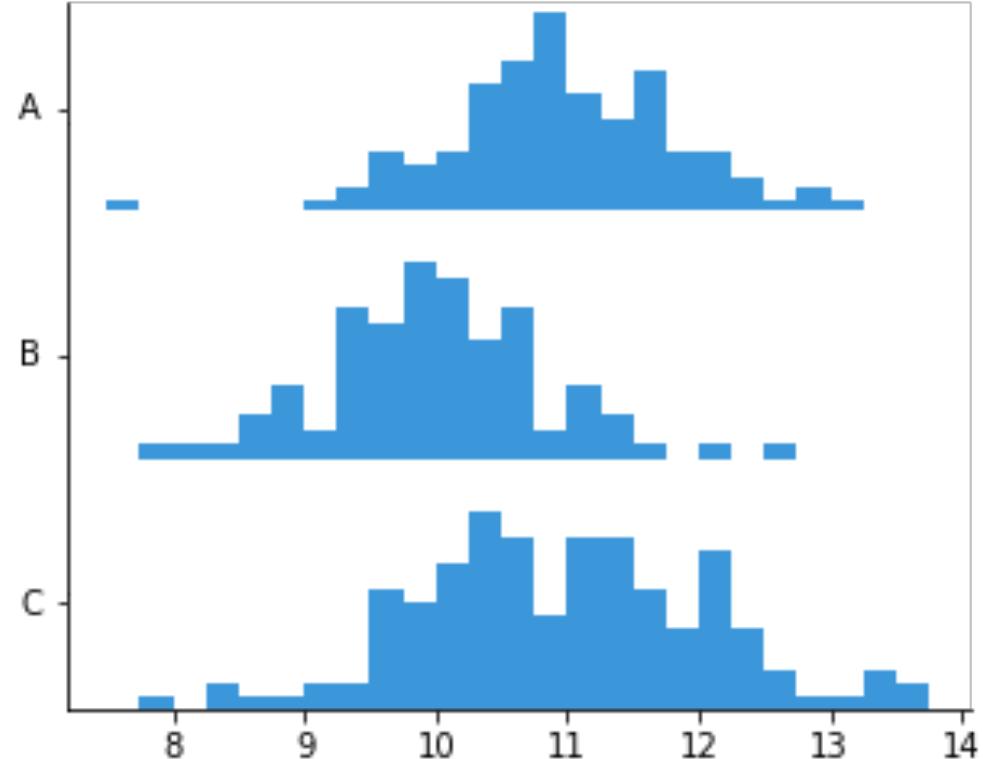
Symmetric distribution



Skewed distribution



Boxplot & Histogram chart





VINBIGDATA



VINGROUP



Academy
Vietnam

Q & A
Thank you!