

Computer Science Advanced

BÀI 12. TRỰC QUAN HÓA DỮ LIỆU VỚI MATPLOTLIB

1. Thư Viện Matplotlib

Matplotlib là một *thư viện vẽ đồ thị* trong Python, thường được sử dụng kết hợp với các thư viện **NumPy** và **Pandas**.

Sau khi cài đặt, sử dụng module **pyplot** theo cú pháp:

```
import matplotlib.pyplot as plt
```

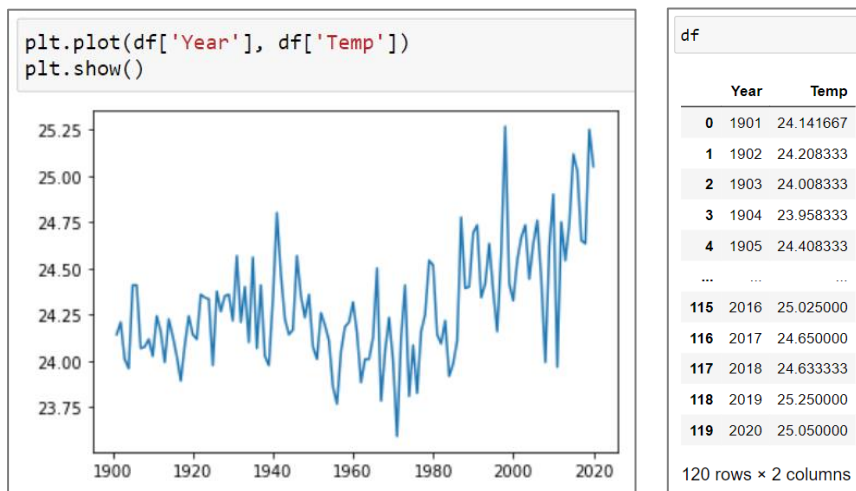
với **plt** là tên viết tắt thông dụng của module.



VÍ DỤ: VẼ BIỂU ĐỒ ĐƯỜNG THẲNG VỚI MATPLOTLIB

Nhiệt độ trung bình tại Việt Nam qua các năm.

Nguồn: [World Bank](#)



CÁC HÀM HỖ TRỢ THÔNG DỤNG

Trong ví dụ trên, hàm **plot()** được dùng để vẽ biểu đồ chính.

Ngoài ra, **Matplotlib** còn hỗ trợ các hàm tùy chỉnh biểu đồ cho từng mục đích sử dụng.

Hàm	Sử dụng
<code>plt.figure(figsize=...)</code>	Khởi tạo <i>figure</i> với kích thước theo inch. Trả về object Figure .
<code>plt.subplots(figsize=...)</code>	Khởi tạo <i>figure</i> với kích thước theo inch. Trả về object Figure và Axes .
<code>plt.xlim(<left>, <right>)</code>	Điều chỉnh giới hạn biểu đồ trên trục x (trục hoành) từ <left> đến <right>.
<code>plt.ylim(<bottom>, <top>)</code>	Điều chỉnh giới hạn biểu đồ trên trục y (trục tung) từ <bottom> đến <top>.
<code>plt.title(<label>)</code>	Đặt tên cho biểu đồ.
<code>plt.xlabel(<xlabel>)</code>	Đặt tên cho trục x.
<code>plt.ylabel(<ylabel>)</code>	Đặt tên cho trục y.
<code>plt.legend()</code>	Hiện chú thích cho biểu đồ.
<code>plt.show()</code>	Vẽ biểu đồ đã khởi tạo.

Để có thêm thông tin về các hàm trong **Matplotlib**, sử dụng hàm **help()** của Python hoặc tìm kiếm trên công cụ tìm kiếm và trang web chính thức của thư viện: <https://matplotlib.org/>.

2. Ví Dụ: Vẽ Các Biểu Đồ Thông Dụng

Dữ liệu và code bên dưới được lưu trong thư mục Visualization Samples.

Nguồn dữ liệu: [World Bank](#)

	Year	Temp
0	1901	24.141667
1	1902	24.208333
2	1903	24.008333
...
119	2020	25.050000

120 rows x 2 columns

BIỂU ĐỒ ĐƯỜNG `plt.plot(<x>, <y>, [label=...], [color=...], ...)`

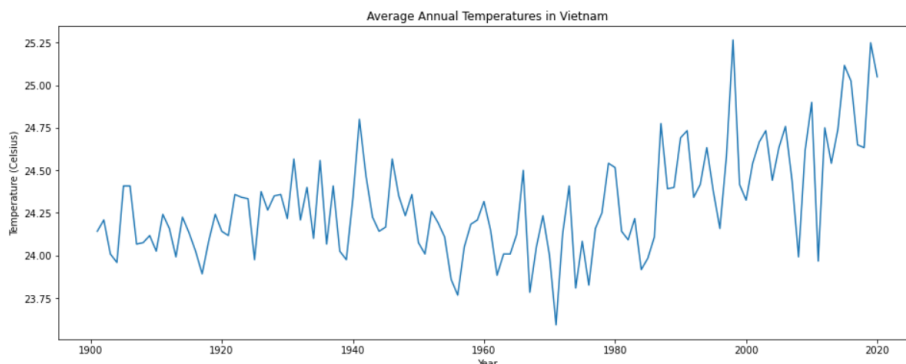
Thay đổi kích cỡ biểu đồ; đặt tên biểu đồ và các trục.

```
# initialize figure, set size in inches
plt.figure(figsize=(16, 6))

# Line plot: x=Year, y=Temp
plt.plot(df['Year'], df['Temp'])

# set title & axis labels
plt.title('Average Annual Temperatures \
in Vietnam')
plt.xlabel('Year')
plt.ylabel('Temperature (Celsius)')

# show the plot
plt.show()
```



Điều chỉnh trục x; xóa hai đường biên;
định dạng tên biểu đồ và các trục.

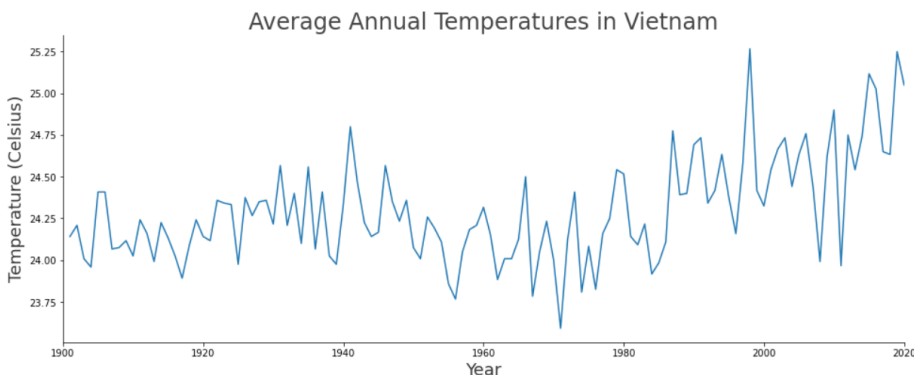
```
# initialize figure, set size in inches
fig, ax = plt.subplots(figsize=(16, 6))

# Line plot: x=Year, y=Temp
plt.plot(df['Year'], df['Temp'])

# customize axes: Limit & visibility
plt.xlim(1900, 2020)
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

# set title & axis labels
plt.title('Average Annual Temperatures in Vietnam', fontsize=24, color='#444444')
plt.xlabel('Year', fontsize=18, color='#333333')
plt.ylabel('Temperature (Celsius)', fontsize=18, color='#333333')

plt.show()
```



BIỂU ĐỒ TRÒN `plt.pie(<x>, [autopct=...], [colors=...], [startangle=...], [counterclock=...], ...)`

Định dạng phần trăm; định dạng màu; thêm đường biên giữa các vùng;
thêm bảng chú thích.

```
# initialize figure, set size in inches
fig, ax = plt.subplots(figsize=(16, 5))

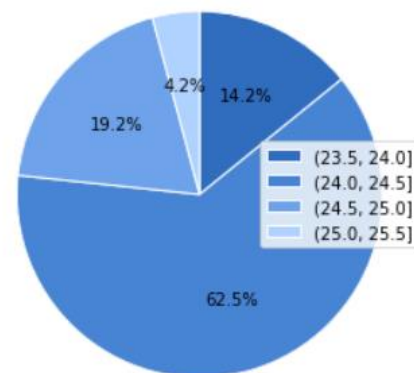
# pie chart: proportion of months in temperature ranges
colors = ['#306dbf', '#4884d4', '#6ea3eb', '#b0d2ff']
plt.pie(df['Count'], autopct='%1.1f%', startangle=90, counterclock=False,
        colors=colors, wedgeprops={'edgecolor': 'w', 'linewidth': 1})

# set title, axis labels & legends
plt.title('Proportion of Months in \nTemperature Ranges',
        fontsize=18, color='#333333', x=0.125, ha='left')
plt.legend(df.index, loc='right')

plt.show()
```

	Count
(23.5, 24.0]	17
(24.0, 24.5]	75
(24.5, 25.0]	23
(25.0, 25.5]	5

Proportion of Months in
Temperature Ranges



BIỂU ĐỒ CỘT

```
plt.bar(<x>, <height>, [width=...], [label=...], [color=...], ...)
```

Đặt tên các cột; điều chỉnh trục x; xóa hai đường biên; đặt tên biểu đồ và các trục.

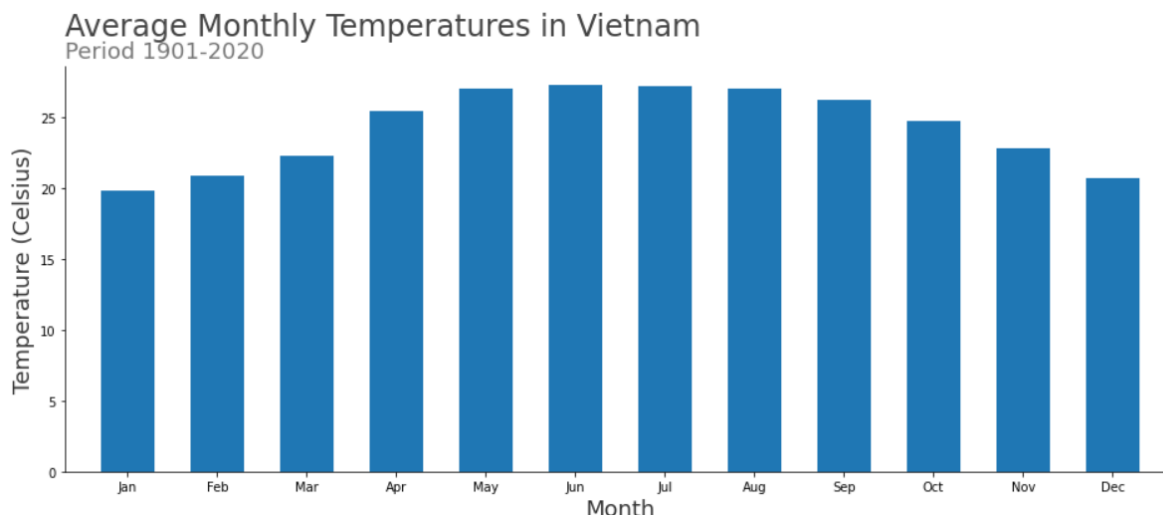
```
# initialize figure, set size in inches
fig, ax = plt.subplots(figsize=(16, 6))

# bar plot: Temp of Month
MONTHS = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']
plt.bar(df['Month'], df['Temp'], width=0.6, tick_label=MONTHS)

# customize axes: limit & visibility
plt.xlim(0.3, 12.7)
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

# set title & axis labels
plt.suptitle('Average Monthly Temperatures in Vietnam', fontsize=24, color='#444444', x=0.125, ha='left')
plt.title('Period 1901-2020', fontsize=18, color='#777777', x=0, ha='left')
plt.xlabel('Month', fontsize=18, color='#333333')
plt.ylabel('Temperature (Celsius)', fontsize=18, color='#333333')

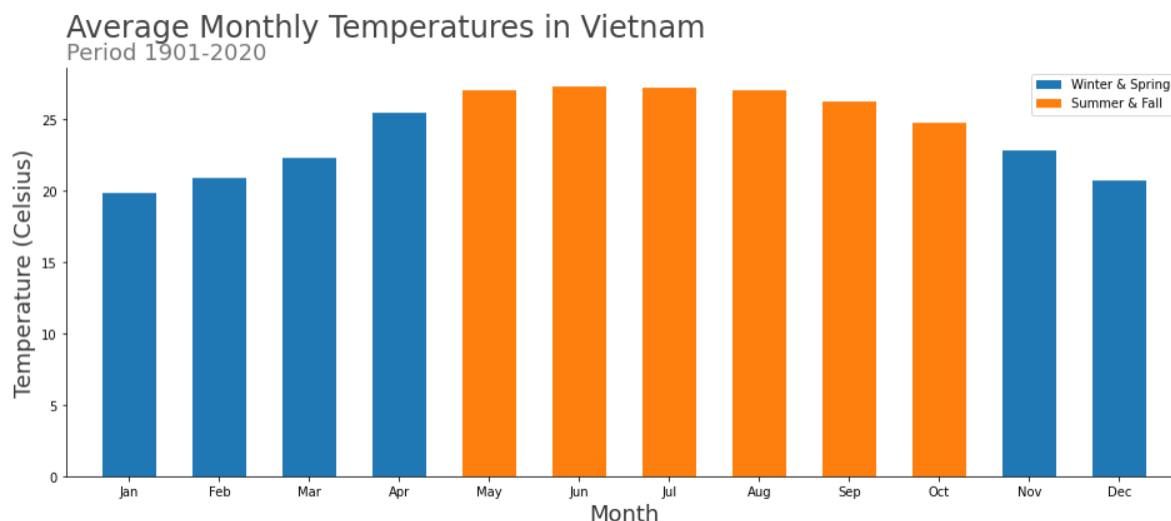
plt.show()
```



Đánh dấu
các tháng
theo mùa.

```
# ...same code as above...
# bar plot: Temp of Month
MONTHS = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec']
plt.bar(df['Month'], df['Temp'], width=0.6, tick_label=MONTHS, label='Winter & Spring')
plt.bar(df['Month'][4:10], df['Temp'][4:10], width=0.6, color='C1', label='Summer & Fall')
# ...same code as above...

plt.legend()
plt.show()
```



BIỂU ĐỒ TẦN SUẤT (HISTOGRAM) `plt.hist(<x>, [bins=...], [color=...], [ec=...], ...)`

Là biểu đồ cột thể hiện số lần xuất hiện của từng giá trị hay từng khoảng giá trị trong dữ liệu.

Thêm đường biên giữa các cột; đánh dấu trên trục x; xóa hai đường biên; đặt tên biểu đồ và các trục.

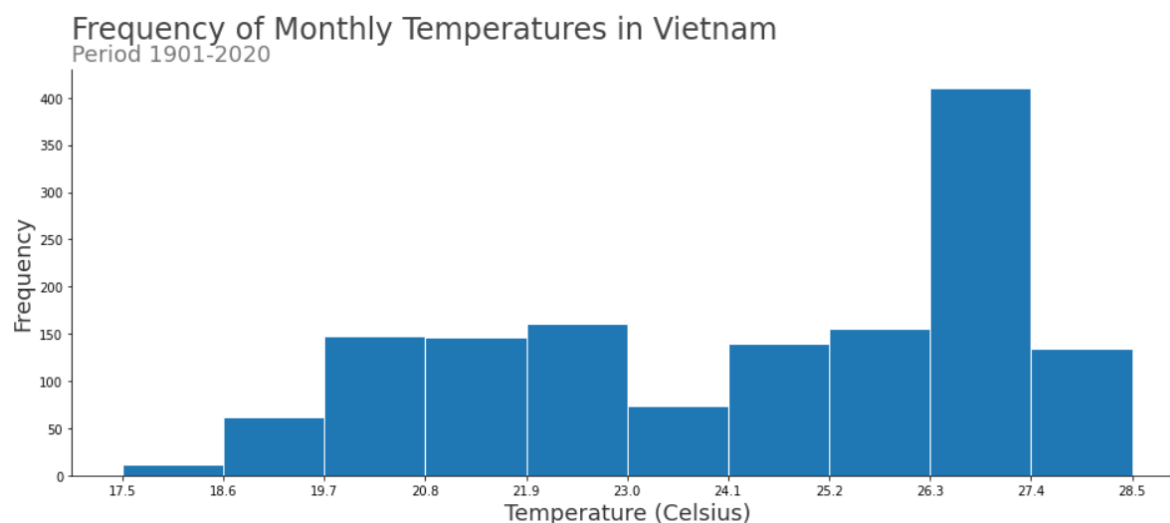
```
# initialize figure, set size in inches
fig, ax = plt.subplots(figsize=(16, 6))

# histogram: frequency of temperatures
n, bins, edges = plt.hist(df['Temp'], ec='white')
plt.xticks(bins)

# customize axes: limit & visibility
ax.spines['top'].set_visible(False)
ax.spines['right'].set_visible(False)

# set title, axis labels & legends
plt.suptitle('Frequency of Monthly Temperatures in Vietnam', fontsize=24, color='#444444', x=0.125, ha='left')
plt.title('Period 1901-2020', fontsize=18, color='#777777', x=0, ha='left')
plt.xlabel('Temperature (Celsius)', fontsize=18, color='#333333')
plt.ylabel('Frequency', fontsize=18, color='#333333')

plt.show()
```



Tăng số lượng khoảng giá trị thành 20.

```
# ...same code as above...
# histogram: frequency of temperatures
n, bins, edges = plt.hist(df['Temp'], bins=20, ec='white')
plt.xticks(bins)
# ...same code as above...

plt.show()
```

