BÀI TẬP THỰC HÀNH
MÔN TRÍ TUỆ NHÂN TẠO
RÚT TRÍCH DỮ LIỆU VÀ TIỀN XỬ LÝ DỮ LIỆU

A. Làm quen pandas

1. https://www.w3schools.com/python/pandas/default.asp
2. https://pandas.pydata.org/docs/getting_started/index.html - getting-started

B. Làm quen matplotlib

1. https://www.w3schools.com/python/matplotlib_intro.asp
2. https://matplotlib.org/

Bài tập 1: Xử lý dữ liệu cơ bản:

- *Bước 1:* Cho file csv có dữ liệu gồm:

```
name,age,state,point
Alice,24,NY,64
Bob,42,CA,92
Charlie,18,CA,70
Dave,68,TX,70
Ellen,24,CA,88
Frank,30,NY,57
Alice,24,NY,64
Bob,42,CA,92
Charlie,18,CA,70
Dave,68,TX,70
Ellen,24,CA,88
Frank,30,NY,57
```

- *Bước 2:* Dùng pandas đọc dữ liệu từ file csv, hiển thị 5 dòng đầu, hiển thị 5 dòng cuối.
- *Bước 3:* Hiển thị toàn bộ tên các cột của data

```
Index(['name', 'age', 'state', 'point'], dtype='object')
```

- *Bước 4:* Hiển thị số dòng, số cột của data đọc được.

```
      name  age state  point
0    Alice   24    NY     64
1      Bob   42    CA     92
2  Charlie   18    CA     70
3     Dave   68    TX     70
4    Ellen   24    CA     88
5    Frank   30    NY     57
6    Alice   24    NY     64
7      Bob   42    CA     92
8  Charlie   18    CA     70
9     Dave   68    TX     70
10   Ellen   24    CA     88
11   Frank   30    NY     57
(12, 4)
```

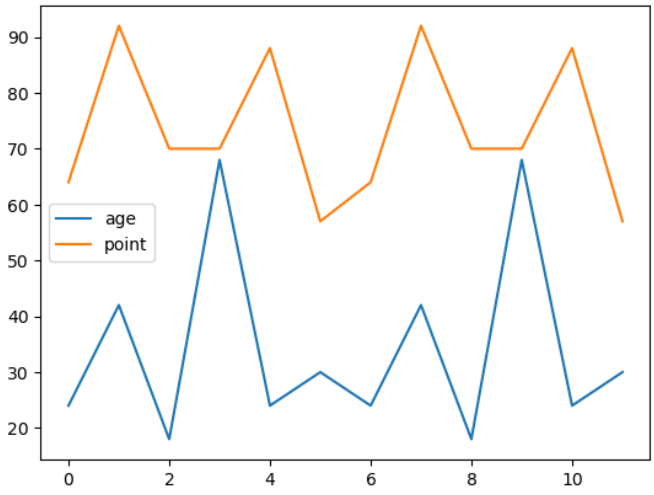
- *Bước 5:* Hiển thị toàn bộ thông tin mô tả: min, max, count, mean, .. của các cột là số trong data đã đọc được.

	age	point
count	12.000000	12.000000
mean	34.333333	73.500000
std	17.536110	13.069743
min	18.000000	57.000000
25%	24.000000	64.000000
50%	27.000000	70.000000
75%	42.000000	88.000000
max	68.000000	92.000000

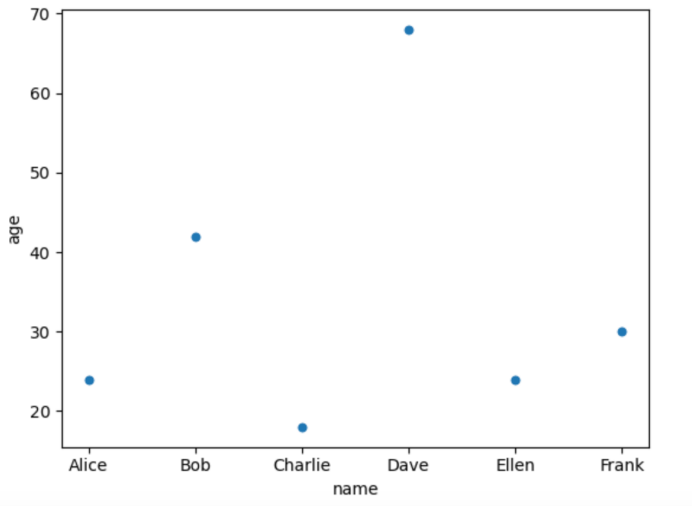
Hiển thị kiểu dữ liệu từng cột trong dataframe.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12 entries, 0 to 11
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   name    12 non-null     object
 1   age     12 non-null     int64 
 2   state   12 non-null     object
 3   point   12 non-null     int64 
dtypes: int64(2), object(2)
memory usage: 512.0+ bytes
```

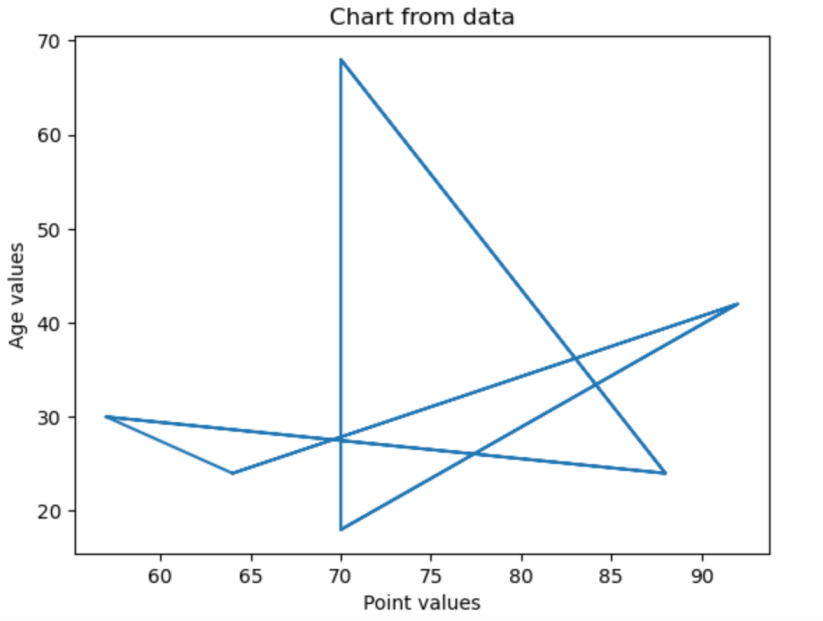
- *Bước 6:* Hiển thị các dữ liệu trùng lắp trong data của cột 'age'
- *Bước 7:* Hiển thị lại data cột 'age' đã bỏ trùng lắp, hiển thị số dòng còn lại
- *Bước 8:* Hiển thị toàn bộ data ban đầu bằng plot từ pandas.



- *Bước 9:* Hiển thị data name và age theo dạng đồ thị:



Tìm hiểu hiển thị trên matplotlib.



- *Bước 10:* Hiển thị data trên nhiều subplot của matplotlib:

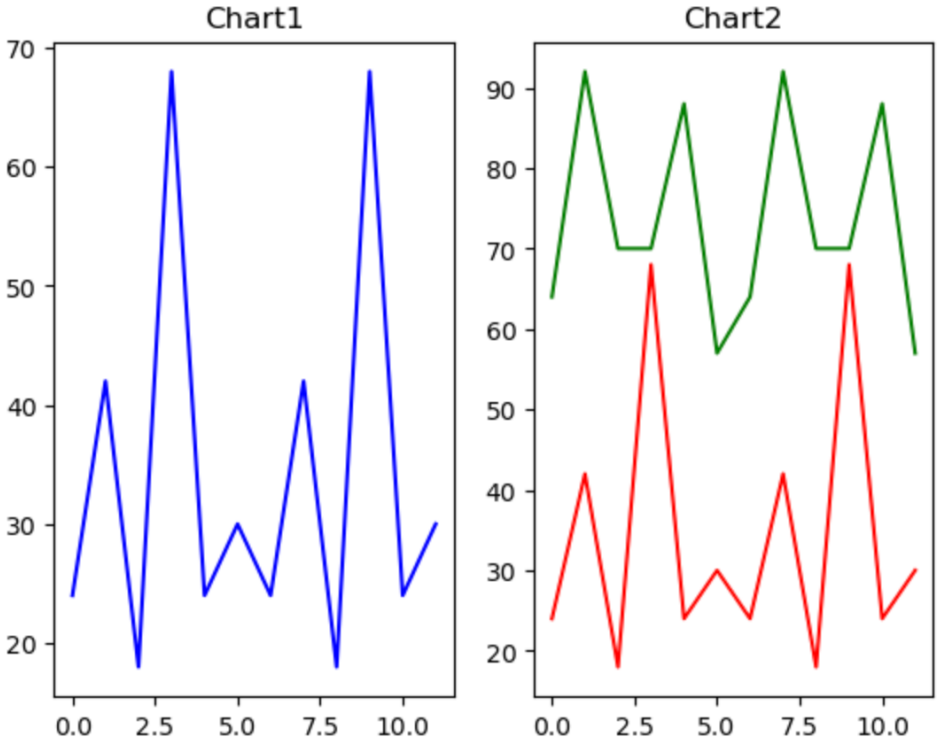
+ Tạo DataFrame chỉ gồm cột name và age:

	name	age
0	Alice	24
1	Bob	42
2	Charlie	18
3	Dave	68
4	Ellen	24
5	Frank	30
6	Alice	24
7	Bob	42
8	Charlie	18
9	Dave	68
10	Ellen	24
11	Frank	30

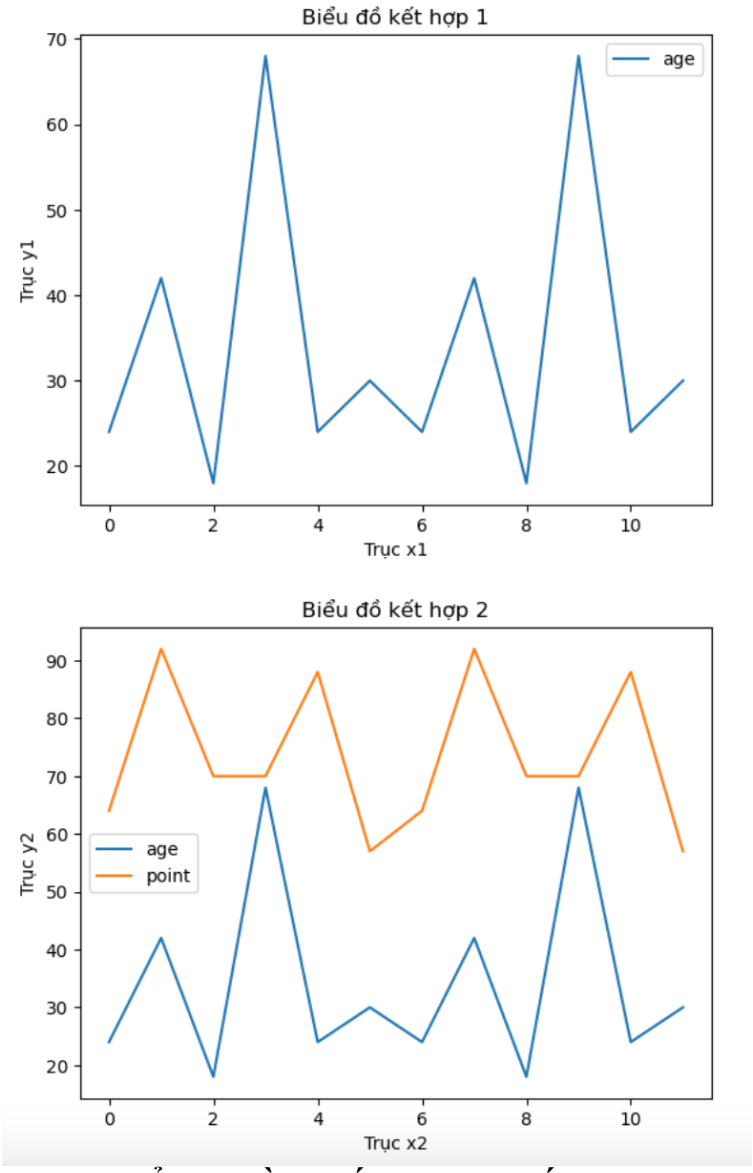
+ Tạo DataFrame chỉ gồm cột age và point:

	age	point
0	24	64
1	42	92
2	18	70
3	68	70
4	24	88
5	30	57
6	24	64
7	42	92
8	18	70
9	68	70
10	24	88
11	30	57

+ Hiển thị plot từ DataFrame lên subplot (chung trong 1 đồ thị)



- *Bước 11*: Tiến hành bổ sung các giá trị name lên cho subplot[0], bổ sung tên các subplot tương ứng. Đổi màu các đường line và sau đó dùng biểu đồ thể hiện điểm (scatter) để hiển thị lại.
- *Bước 12*: Tạo 1 màn hình xuất cho biểu đồ gồm các thể hiện như sau:



Bài tập 2: Xử lý data csv và hiển thị đồ thị số ca mới và số ca tử vong của Covid-19 (dữ liệu từ WHO)

- *Bước 1:* Chuẩn bị dữ liệu csv và thư viện:

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

```
WHO_covid_data = pd.read_csv('WHO-COVID-19-global.csv')
# print(WHO_covid_data)
headers = list(WHO_covid_data)
print(headers)
```

Quan sát kết quả thu được.

- *Bước 2:* Hiển thị dữ liệu tương ứng với từng loại dữ liệu trong toàn tập dữ liệu lớn.

```
every_nth_days = 1

country_df  = WHO_covid_data[WHO_covid_data[' Country'] == 'Brazil']
info = ['cases']
for info_type in info:
    print(info_type)
    brazil_df = pd.DataFrame(country_df)
    #print(d)
    print(brazil_df.head(10))
    brazil_df.plot()
    plt.show()
```

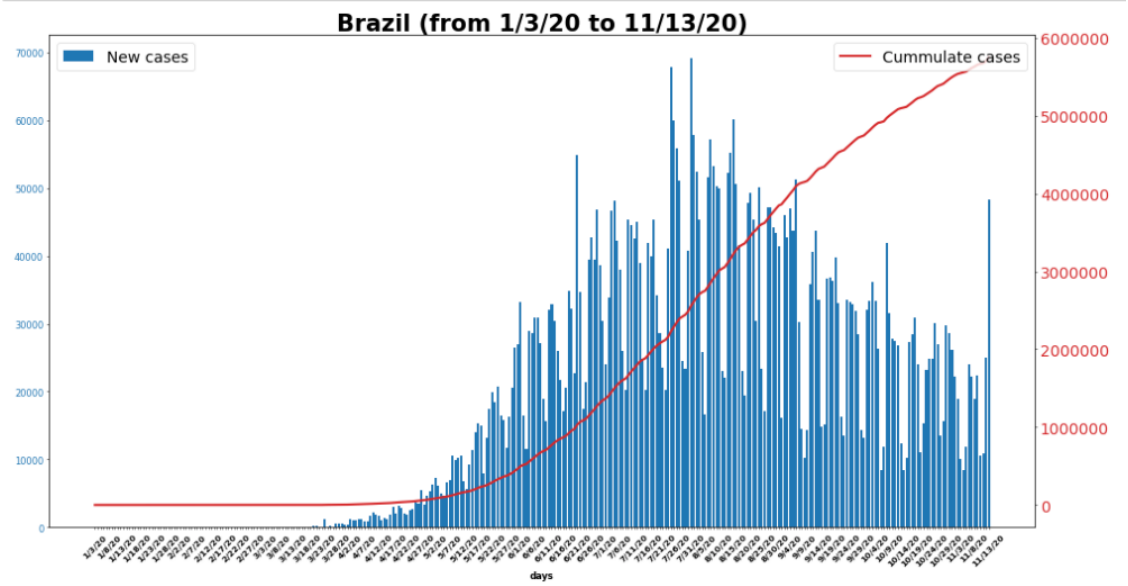
Quan sát kết quả thu được.

- *Bước 3:* Hiển thị dữ liệu tương ứng trong dãy thời gian qui định (từ date[0] đến date[len(date) – 1] trong dữ liệu:

```
every_nth_days = 1

country_df  = WHO_covid_data[WHO_covid_data[' Country'] == 'Brazil']
info = ['cases']
for info_type in info:
    fig, ax1 = plt.subplots(figsize=(20,10), dpi= 50)
    color = 'tab:blue'
    ax1.set_xlabel('days',fontsize='large', fontweight='bold')
    ax1.bar(country_df['Date_reported'][::every_nth_days],country_df[' New_{}'.format(info_type)][::every_nth_days],col
    ax1.tick_params(axis='y', labelcolor=color,labelsize='large')

    for idx, t in enumerate(ax1.get_xticklabels()):
        t.set_fontsize(11)
        t.set_fontweight('bold')
        if (idx % 5) != 0:
            t.set_visible(False)

    plt.ticklabel_format(style='plain', axis='y')
    plt.xticks(rotation=90)
    dates = country_df['Date_reported'].unique()
    start_date = dates[0]
    end_date = dates[len(dates)-1]

    plt.title('{} (from {} to {})'.format('Brazil',start_date,end_date),fontsize=30, fontweight='bold')
    fig.tight_layout()
    ax1.legend(loc='upper left',fontsize=20)
    plt.setp(ax1.xaxis.get_majorticklabels(), rotation=45)

    plt.show()
```

Quan sát kết quả thu được.

- *Bước 4:* bổ sung thêm giá trị tích lũy trên chart vừa vẽ ở bước 3 sao cho kết quả thu được: (dùng twinx để bổ sung chia đôi trục x để hiển thị giá trị cummulate)

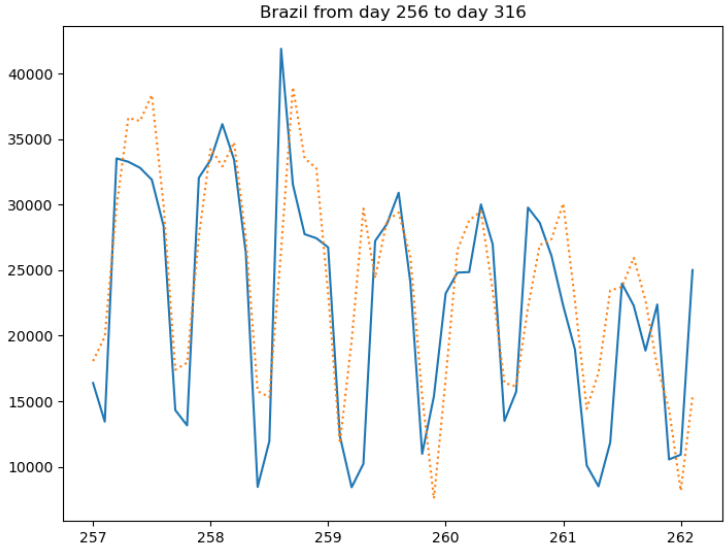


Bài tập 3: Dự đoán dựa trên DataFrame:

- *Bước 1:* Sử dụng biến đổi dữ liệu: (7 ngày liên tiếp -> dự đoán cho ngày 8) Hoặc sử dụng group ngày tùy biến.
- *Bước 2:* Sử dụng Linear Regression để dự đoán từ ngày 256 đến 316:

```
from sklearn.linear_model import LinearRegression
regr = LinearRegression()

regr.fit(X_train, y_train)
y_pred = regr.predict(X_test)
```

- *Bước 3:* Hiển thị dữ liệu dự đoán so với dữ liệu gốc:



Bài tập 4: Tiến hành thực hiện dự đoán với các nước còn lại trong csv và hiển thị dữ liệu tương tự bước 3 của bài tập 3 nhưng có bổ sung tên cột, tên các đường line cho biểu đồ.

Bài tập 5: Tiến hành đọc file "Salary_dataset.csv" hãy thực hiện các yêu cầu sau:

- Loại bỏ các giá trị rỗng.
- Loại bỏ các giá trị trùng lắp.

- Cho biết số records còn lại.
- Hiển thị dữ liệu cột “salary” trên biểu đồ.
- Hiển thị dữ liệu cả 2 cột trên cùng 1 biểu đồ.
- Dùng 2/3 dữ liệu đi train và 1/3 dữ liệu đi kiểm thử cho việc dự đoán “salary” dựa trên “YearsExperience”.

--Hết--