

ĐỒ ÁN THỰC HÀNH

NHÓM 16

NHẬP MÔN KHOA HỌC DỮ LIỆU

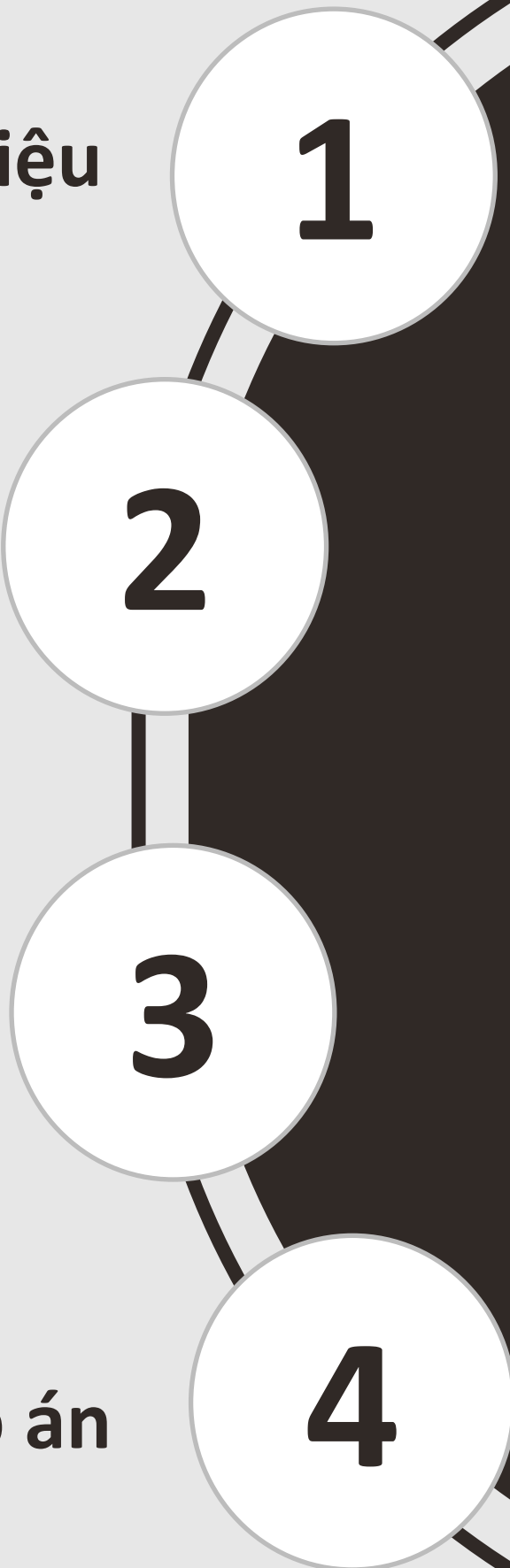
GVHD: Thầy Lê Đại Chí và Thầy Lê Nhựt Nam

DANH SÁCH THÀNH VIÊN

Nhóm 16

HỌ VÀ TÊN	MSSV
Phan Phong Lưu	20120326
Bùi Thanh Tùng	20120398
Đỗ Tấn Tài	20120408
Trần Khắc Bình	20120437

Nội dung chính

- 
- Quy trình khoa học dữ liệu
 - Tiền xử là khám phá dữ liệu
 - Đặt các câu hỏi dự trên bộ dữ liệu
 - Tổng hợp lại quá trình thực hiện đồ án

I Quy trình khoa học dữ liệu



1. Thu thập dữ liệu

Đây là Bộ dữ liệu gốc được trích xuất từ <https://platform.who.int/> về "Con người chết vì nguyên nhân gì?"

Dữ liệu được cập nhật hàng năm và hiện có dữ liệu cho đến năm 2021.

Sử dụng thư viện **Selenium** để thu thập dữ liệu.

1. Thu thập dữ liệu

Ta sẽ thu thập 14 nguyên nhân chết chủ yếu của con người trên thế giới:

- HIV/AIDS
- Malaria
- Tuberculosis
- Dengue
- Covid-19
- Brain and nervous system cancers
- Leukaemia
- Road traffic accidents
- Fires
- Drownings
- Natural disasters
- Self-inflicted injuries
- Violence
- War

1. Thu thập dữ liệu

Các bước thu thập dữ liệu:

- ❖ Dùng selenium để thu thập dữ liệu từ web
- ❖ Lưu dữ liệu đã thu thập vào file **data.csv**

- Cấu trúc file **data.csv**:

Cause of death, Age, Year, Country, Number of deaths

2

Tiền xử lý và khám phá dữ liệu



a. Các đặc điểm của dữ liệu

Đọc dữ liệu từ file `data.csv` vào dataframe tên `df`, ta có các thông tin sau:

- `df` có 262752 dòng: Mỗi dòng số lượng người mất do một nguyên nhân, ở một quốc gia trong một năm với một nhóm tuổi nhất định.

Qua kiểm tra cho thấy không có dòng nào là bị trùng lặp dữ liệu cả, nên ta không cần phải xử lý việc trùng lặp dữ liệu.

```
have_duplicated_rows=len(df.index)-len(df.index.drop_duplicates())  
assert have_duplicated_rows == False
```

✓ 0.4s

Python

a. Các đặc điểm của dữ liệu

Đọc dữ liệu từ file `data.csv` vào dataframe tên `df`, ta có các thông tin sau:

- `df` có 5 cột: Dưới đây là phần mô tả thông tin về các cột trong file "data.csv" mà nhóm đã lấy về được.
 - **Cause of death:** Tên các loại dịch bệnh.
 - **Age:** Nhóm tuổi tử vong.
 - **Year:** Năm thu thập thông tin.
 - **Country:** Quốc gia lấy thông tin.
 - **Number of deaths:** Số lượng tử vong.

a. Các đặc điểm của dữ liệu

Khoảng biểu diễn các cột dữ liệu:

❖ Cột **Cause of death**:

```
display(df['Cause of death'].unique())  
print("Số lượng các giá trị khác nhau:" ,df['Cause of death'].nunique())  
✓ 0.7s Python
```

```
array(['HIV/AIDS', 'Malaria', 'Tuberculosis', 'Dengue', 'Covid-19',  
      'Brain and nervous system cancers', 'Leukaemia',  
      'Road traffic accidents', 'Fires', 'Drownings',  
      'Natural disasters', 'Self-inflicted injuries', 'Violence', 'War'],  
      dtype=object)
```

Số lượng các giá trị khác nhau: 14

Cột **Cause of death** có 14 giá trị khác nhau

a. Các đặc điểm của dữ liệu

Khoảng biểu diễn các cột dữ liệu:

❖ Cột **Age**:

```
display(df['Age'].unique())  
print("Số lượng các giá trị khác nhau:" ,df['Age'].nunique())
```

✓ 0.7s

Python

```
array(['0', '1-4', '5-14', '15-24', '25-34', '35-54', '55-74', '75+'],  
      dtype=object)
```

Số lượng các giá trị khác nhau: 8

Cột **Age** có 8 giá trị khác nhau

a. Các đặc điểm của dữ liệu

Khoảng biểu diễn các cột dữ liệu:

❖ Cột **Country**:

```
display(df['Country'].unique())  
print("Số lượng các giá trị khác nhau:" ,df['Country'].nunique())
```

✓ 0.7s

Python

a. Các đặc điểm của dữ liệu

Khoảng biểu diễn các cột dữ liệu:

❖ Cột **Country**:

```
array(['Albania', 'Antigua and Barbuda', 'Argentina', 'Armenia',  
      'Australia', 'Austria', 'Azerbaijan', 'Bahamas', 'Bahrain',  
      'Barbados', 'Belarus', 'Belgium', 'Belize',  
      'Bosnia and Herzegovina', 'Brazil', 'Brunei Darussalam',  
      'Bulgaria', 'Cabo Verde', 'Canada', 'Chile',  
      'China, Hong Kong SAR', 'Colombia', 'Costa Rica', 'Croatia',  
      'Cuba', 'Cyprus', 'Czechia', 'Denmark', 'Dominica',  
      'Dominican Republic', 'Ecuador', 'Egypt', 'El Salvador', 'Estonia',  
      'Fiji', 'Finland', 'France', 'French Guiana', 'Georgia', 'Germany',  
      'Greece', 'Grenada', 'Guadeloupe', 'Guatemala', 'Guyana',  
      'Hungary', 'Iceland', 'Iran (Islamic Republic of)', 'Iraq',  
      'Ireland', 'Israel', 'Italy', 'Jamaica', 'Japan', 'Jordan',  
      'Kazakhstan', 'Kuwait', 'Kyrgyzstan', 'Latvia', 'Lebanon',  
      'Lithuania', 'Luxembourg', 'Maldives', 'Malta', 'Martinique',
```

```
'Lithuania', 'Luxembourg', 'Maldives', 'Malta', 'Martinique',  
'Mauritius', 'Mayotte', 'Mexico', 'Mongolia', 'Montenegro',  
'Netherlands', 'New Zealand', 'Nicaragua', 'North Macedonia',  
'Norway', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Poland',  
'Portugal', 'Puerto Rico', 'Republic of Korea',  
'Republic of Moldova', 'Romania', 'Russian Federation', 'Réunion',  
'Saint Kitts and Nevis', 'Saint Lucia',  
'Saint Vincent and the Grenadines', 'Serbia', 'Seychelles',  
'Singapore', 'Slovakia', 'Slovenia', 'South Africa', 'Spain',  
'Sri Lanka', 'Suriname', 'Sweden', 'Switzerland',  
'Syrian Arab Republic', 'Tajikistan', 'Thailand',  
'Trinidad and Tobago', 'Turkey', 'Turkmenistan', 'Ukraine',  
'United Kingdom', 'United States of America', 'Uruguay',  
'Uzbekistan', 'Venezuela'], dtype=object)
```

Cột **Country** có 113 giá trị khác nhau

Số lượng các giá trị khác nhau: 113

a. Các đặc điểm của dữ liệu

Khoảng biểu diễn các cột dữ liệu:

- ❖ Cột **Number of deaths**:

```
df['Number of deaths'].describe()
```

✓ 0.6s

Python

```
count    207240
unique     3900
top         0
freq     95353
Name: Number of deaths, dtype: object
```


a. Các đặc điểm của dữ liệu

Kiểu dữ liệu của các cột:

Ta lấy dtype (kiểu dữ liệu của mỗi phần tử) của mỗi cột trong dữ liệu và lưu kết quả vào series `col_dtypes`; series này có index là tên cột.

```
col_dtypes = pd.Series(df.dtypes, index = df.columns)
display(col_dtypes)
```

✓ 0.4s

Python

Cause of death	object
Age	object
Year	int64
Country	object
Number of deaths	object
dtype: object	

b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Cột **Year** đang có dtype là **int64**. Để có thể tiếp tục khám phá thêm về cột này, ta sẽ thực hiện bước tiền xử lý là chuyển sang dạng **datetime**.

```
df['Year'] = pd.to_datetime(df['Year'], format='%Y')
```

✓ 0.6s

Python

```
print("min_Year: ",min_Year)
```

```
print("max_Year: ",max_Year)
```

✓ 0.1s

```
min_Year: 2000
```

```
max_Year: 2020
```

b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Cột **Number of deaths** đang có dtype là **object**. Trong Pandas, kiểu dữ liệu object thường ám chỉ chuỗi, nhưng thật ra kiểu dữ liệu object có thể chứa một đối tượng bất kỳ trong Python (vì thật ra ở bên dưới kiểu dữ liệu object chứa địa chỉ). Nếu một cột trong dataframe có **dtype** là object thì có thể các phần tử trong cột này sẽ có kiểu dữ liệu khác nhau.

.

b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Ta xem chi tiết kiểu dữ liệu cột **Number of deaths**.

```
open_object_dtype(df['Number of deaths'])
```

✓ 0.1s

Python

```
{float, str}
```

b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Cột **Number of deaths** có kiểu dữ liệu **object**, ta tiến hành thay các giá trị bị thiếu thành 0 và đổi kiểu dữ liệu thành **int64**.

```
col_dtypes = pd.Series(df.dtypes, index = df.columns)
display(col_dtypes)
```

✓ 0.4s

Python

```
Cause of death      object
Age                 object
Year               datetime64[ns]
Country            object
Number of deaths    int64
dtype: object
```

b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Chia **Number of deaths** theo **Cause death** và gom nhóm theo **Country**, **Year** và **Age**.

Thay các giá trị Nan bằng giá trị 0 vì có Nan hay cũng đều không có ý nghĩa cho thống kê dữ liệu, thay thế để dễ dàng có các bước trả lời câu hỏi hơn.

b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Chia **Number of deaths** theo **Cause death** và gom nhóm theo **Country**, **Year** và **Age**.

Đếm số dòng có giá trị 0 của các cột **Country**, **Year**, **Age**.

```
Country: 0
Year    : 0
Age     : 0
```

Như vậy, chỉ có những cột nguyên nhân tử vong có chứa giá trị 0. Một dòng mà dữ liệu chỉ chứa giá trị 0 thì dòng đó không có ý nghĩa.

=> Loại bỏ những dòng này.

```
Kích thước trước khi xóa: (18984, 17)
Kích thước sau khi xóa   : (14749, 17)
```


b. Tiền xử lý dữ liệu

Vấn đề cần xử lý:

- Thêm cột tên khu vực của quốc gia.

Các khu vực (6):

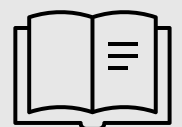
- + Europe
- + North America
- + South America
- + Asia
- + Oceania
- + Africa

Sử dụng thư viện **pycountry_convert**:

```
import pycountry_convert as pc
```

3

**Đặt các câu hỏi dựa
trên bộ dữ liệu**



a. Đưa ra các câu hỏi:

Sau khi đã khám phá dữ liệu và hiểu hơn về dữ liệu, ta thấy có một số câu hỏi có thể được trả lời bằng dữ liệu:

Câu 1: Top 3 quốc gia có số lượng tử vong vì Covid nhiều nhất?

Câu 2: Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

Câu 3: Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?

Câu 4: Diễn biến của các nguyên nhân gây tử vong theo thời gian.

Câu 5: Mối quan hệ giữa một số nguyên nhân tử vong.

Câu 6: So sánh phân phối tuổi tử vong

b. Tiền xử lý và trả lời các câu hỏi

Câu 1. Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?

- ❖ **Ý nghĩa:** với câu hỏi trên, ta biết được những quốc gia có số lượng tử vong vì Covid nhiều nhất. Tử vong nhiều như thế chứng tỏ nỗi mất mát của quốc gia rất nặng nề. Ta đánh giá mục đích có những quan tâm giúp đỡ để bù đắp phần nào về nỗi đau của họ. Bên cạnh đó, việc tử vong nhiều là do công tác phòng chống dịch bệnh Covid của họ vẫn còn nhiều khiếm khuyết, ta cần đưa ra các giải pháp hỗ trợ về mặt y tế cho các quốc gia này.

b. Tiền xử lý và trả lời các câu hỏi

Câu 1. Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?

❖ Để trả lời cho câu hỏi này, ta sẽ làm như sau:

Tính số lượng ca tử vong theo từng nhóm tuổi. Chọn ra top 3 quốc gia có tổng số ca tử vong nhiều nhất. Ta lưu kết quả vào series `num_death_covid`.

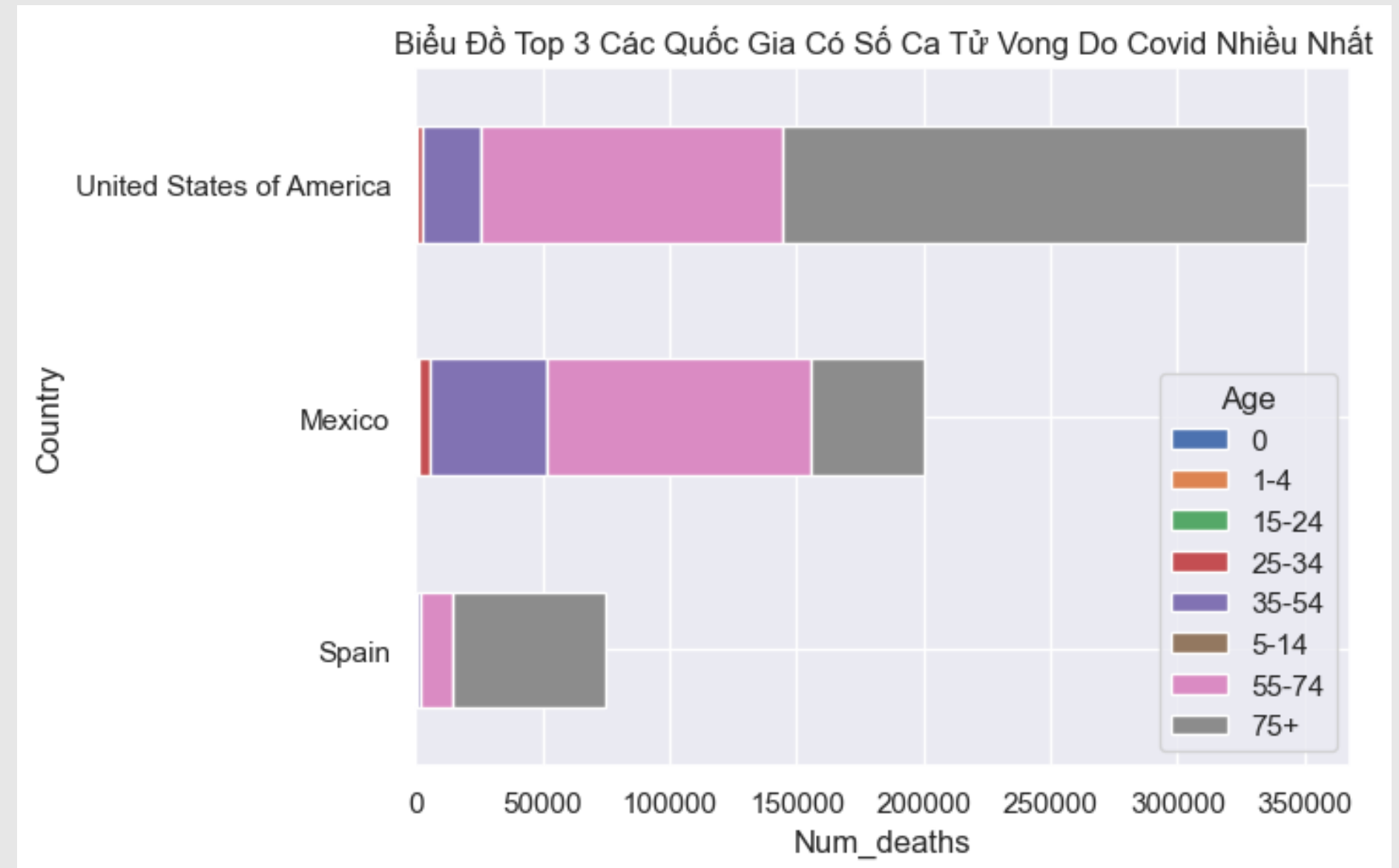
Age	0	1-4	15-24	25-34	35-54	5-14	55-74	75+
Country								
United States of America	35	19	501	2254	23043	49	118367	206559
Mexico	209	111	926	4534	45906	150	103836	44530
Spain	4	1	24	93	1442	5	12731	60539

b. Tiền xử lý và trả lời các câu hỏi

Câu 1. Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?

❖ Để trả lời cho câu hỏi này, ta sẽ làm như sau:

Từ kết quả ở trên, ta vẽ group stack bar chart, trong đó trục hoành là số ca tử vong và trục tung là quốc gia. Ta đặt tên trục hoành là **Num_deaths** và tên trục tung là **Country**.



b. Tiền xử lý và trả lời các câu hỏi

Câu 1. Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?

❖ **Từ kết quả trên, ta có một số nhận xét như sau:**

- Mỹ là quốc gia có lượng người tử vong vì covid nhiều nhất, nguyên nhân có thể giải thích là do thái độ của người dân đối với đại dịch lúc đó khá là chủ quan. Họ cho rằng đây chỉ là một căn bệnh cảm cúm thông thường, nên người dân không hề có các biện pháp phòng tránh như đeo khẩu trang, hạn chế tiếp xúc nơi đông người... Bên cạnh đó, quốc gia này không chịu điều trị cho những người mắc bệnh, từ chối hoàn toàn các ca bệnh của người dân.

b. Tiền xử lý và trả lời các câu hỏi

Câu 1. Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?

❖ **Từ kết quả trên, ta có một số nhận xét như sau:**

- Ta thấy rằng tỷ lệ tử vong đa ở 2 nhóm người, 1 là nhóm tuổi trẻ sơ sinh 0-1 tuổi, hai là nhóm người cao tuổi trên 55 tuổi, cho nên ta có thể suy đoán nhóm người dễ bị ảnh hưởng bởi căn bệnh Covid này là những người có độ miễn dịch cơ thể thấp như trẻ sơ sinh và người già. Trẻ sơ sinh khi có độ miễn dịch cơ thể còn yếu kém và nhóm người già khi đang mắc các căn bệnh trước đó như cao huyết áp hoặc tiểu đường.

b. Tiền xử lý và trả lời các câu hỏi

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

- ❖ **Ý nghĩa:** Bệnh truyền nhiễm là loại bệnh nhiễm trùng có khả năng lây lan từ người này sang người khác một cách trực tiếp hoặc gián tiếp qua môi trường trung gian (như thức ăn, đường hô hấp, dùng chung đồ dùng, máu, da, niêm mạc...) và có khả năng phát triển thành bệnh dịch. Điều tra về số ca tử vong do bệnh truyền nhiễm ở các khu vực giúp ta có thể học tập các công tác phòng chống dịch tốt từ những khu vực có ít người tử vong; nhận xét, rút ra kinh nghiệm từ những khu vực có nhiều người tử vong. Từ đó ngăn chặn các bệnh lây lan xuống mức tối thiểu.

b. Tiền xử lý và trả lời các câu hỏi

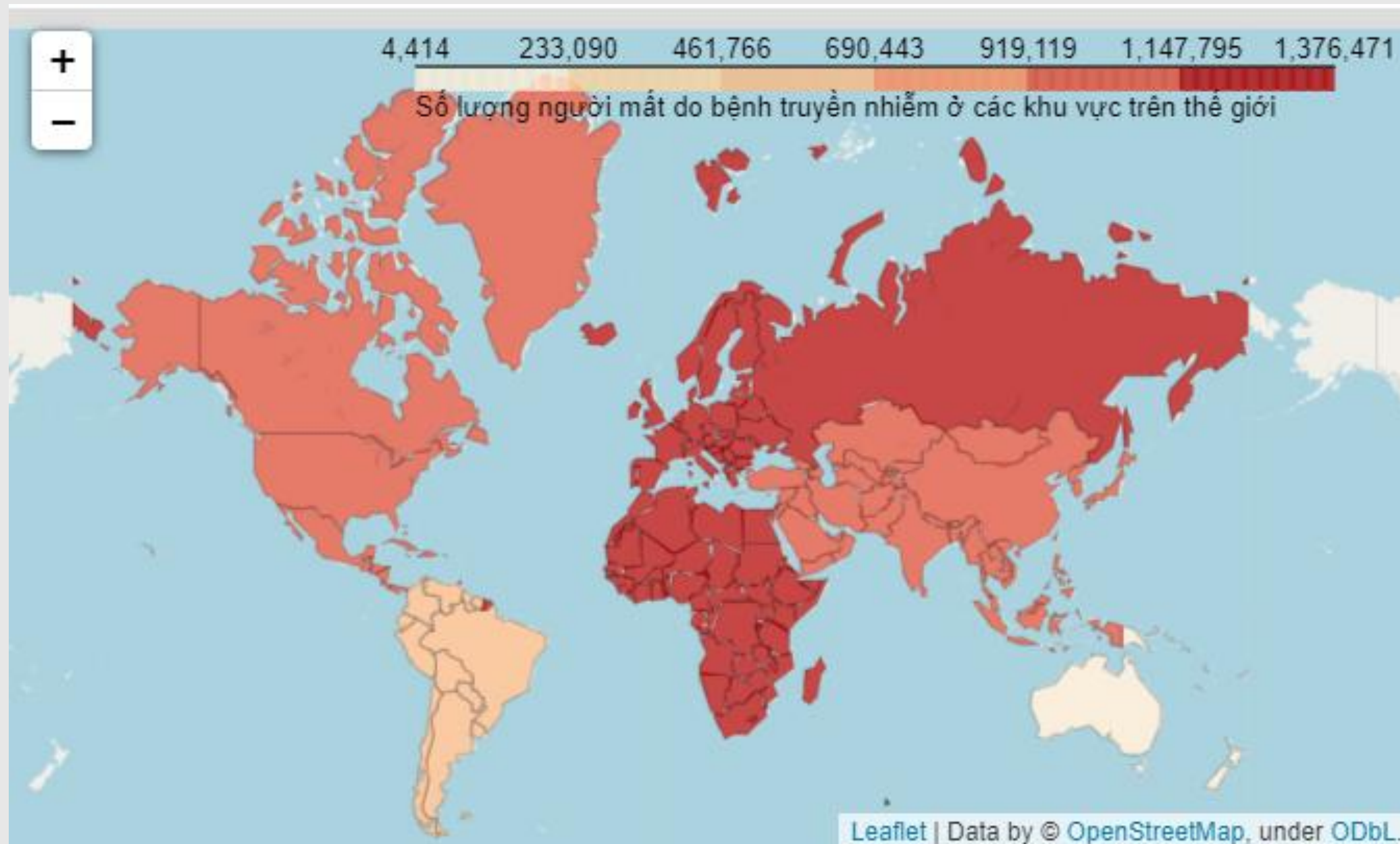
Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

❖ **Để trả lời cho câu hỏi này, ta sẽ làm như sau:**

Tính người chết do các bệnh truyền theo khu vực. Các bệnh truyền nhiễm bao gồm: HIV/AIDS, Covid-19, Dengue, Tuberculosis, Malaria.

	HIV/AIDS	Covid-19	Dengue	Tuberculosis	Malaria	Total
Continent						
Africa	319672	10	43	1047698	9048	1376471
Europe	312772	175954	42	715424	1135	1205327
North America	376820	561340	4358	92118	731	1035367
Asia	145105	5781	23558	761250	4791	940485
South America	390005	23789	13347	210102	4391	641634
Oceania	1771	898	57	1667	21	4414

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?



Trong 20 năm qua, châu Phi là khu vực có số lượng người mất do các bệnh truyền nhiễm cao nhất trên thế giới, mặc dù dữ liệu về Covid-19 ở khu vực này không được cập nhật. Nguyên nhân là do các quốc gia này có điều kiện y tế còn lạc hậu, người dân không được trang bị các kiến thức cần thiết để phòng chống các bệnh truyền nhiễm, dẫn đến dịch bệnh càng ngày càng lan rộng.

Giờ ta sẽ xét riêng từng loại bệnh truyền nhiễm ở các khu vực này.

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

Covid-19

```
round(Infectious_Causes['Covid-19']*100/Infectious_Causes['Total'],3)
```

Continent	
Africa	0.001
Europe	14.598
North America	54.217
Asia	0.615
South America	3.708
Oceania	20.344
dtype:	float64

Tại Bắc Mỹ, mặc dù dịch bệnh Covid-19 chỉ mới bắt đầu bùng phát vào 12/2019, nhưng nó lại chiếm đến 54% trên tổng số người chết do bệnh truyền nhiễm từ năm 2000-2020. Nguyên nhân có là do chính sách “miễn dịch cộng đồng” của chính phủ và sự thờ ơ phòng chống dịch của người dân. Bỏ qua Covid-19, ta thấy các quốc gia ở khu vực này đã làm tốt trong công tác phòng chống các bệnh dịch khác như Sốt rét, Lao, Sốt xuất huyết.

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

HIV/AIDS

```
round(Infectious_Causes['HIV/AIDS']*100/Infectious_Causes['Total'],3)
```

```
Continent
Africa      23.224
Europe      25.949
North America 36.395
Asia        15.429
South America 60.783
Oceania     40.122
dtype: float64
```

Từ dữ liệu trên, ta thấy HIV/AIDS là một trong những nguyên nhân hàng đầu dẫn đến tử vong ở các quốc gia trên thế giới. Chiếm tỉ lệ cao nhất là ở Nam Mỹ. Đây cũng là nguyên nhân hàng đầu dẫn đến tử vong ở châu Đại Dương.

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

Dengue

```
round(Infectious_Causes['Dengue']*100/Infectious_Causes['Total'],3)
```

Continent	
Africa	0.003
Europe	0.003
North America	0.421
Asia	2.505
South America	2.080
Oceania	1.291
dtype:	float64

Mặc dù vẫn chưa có thuốc đặc trị cho bệnh sốt xuất huyết nhưng ta thấy tỉ lệ người tử vong do mắc sốt xuất huyết trong các loại bệnh truyền nhiễm là tương đối thấp. Điều này chứng tỏ các quốc gia đã có được nhiều phương pháp để đặc hiệu để điều trị bệnh này.

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

Malaria

```
round(Infectious_Causes['Malaria']*100/Infectious_Causes['Total'],3)
```

```
Continent
Africa      0.657
Europe      0.094
North America 0.071
Asia        0.509
South America 0.684
Oceania     0.476
dtype: float64
```

Tương tự như sốt xuất huyết, các khu vực đã làm tốt việc điều trị bệnh sốt rét.

Câu 2. Số ca tử vong do các bệnh truyền nhiễm ở các khu vực?

Tuberculosis

```
round(Infectious_Causes['Tuberculosis']*100/Infectious_Causes['Total'],3)
```

```
Continent
Africa      76.115
Europe      59.355
North America  8.897
Asia        80.942
South America 32.745
Oceania     37.766
dtype: float64
```

Bệnh lao là nguyên nhân hàng đầu dẫn đến tử vong ở châu Á và châu Phi và châu Âu.

b. Tiền xử lý và trả lời các câu hỏi

Câu 3. Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?

- ❖ **Ý nghĩa:** Biết được tỉ lệ người tử vong do tự gây thương tích cao nhất, từ đó ta thực hiện nhiều giảm pháp để giảm bớt tình trạng tự tử, đặt biệt là với các nhóm tuổi top đầu mà kết quả thu được.

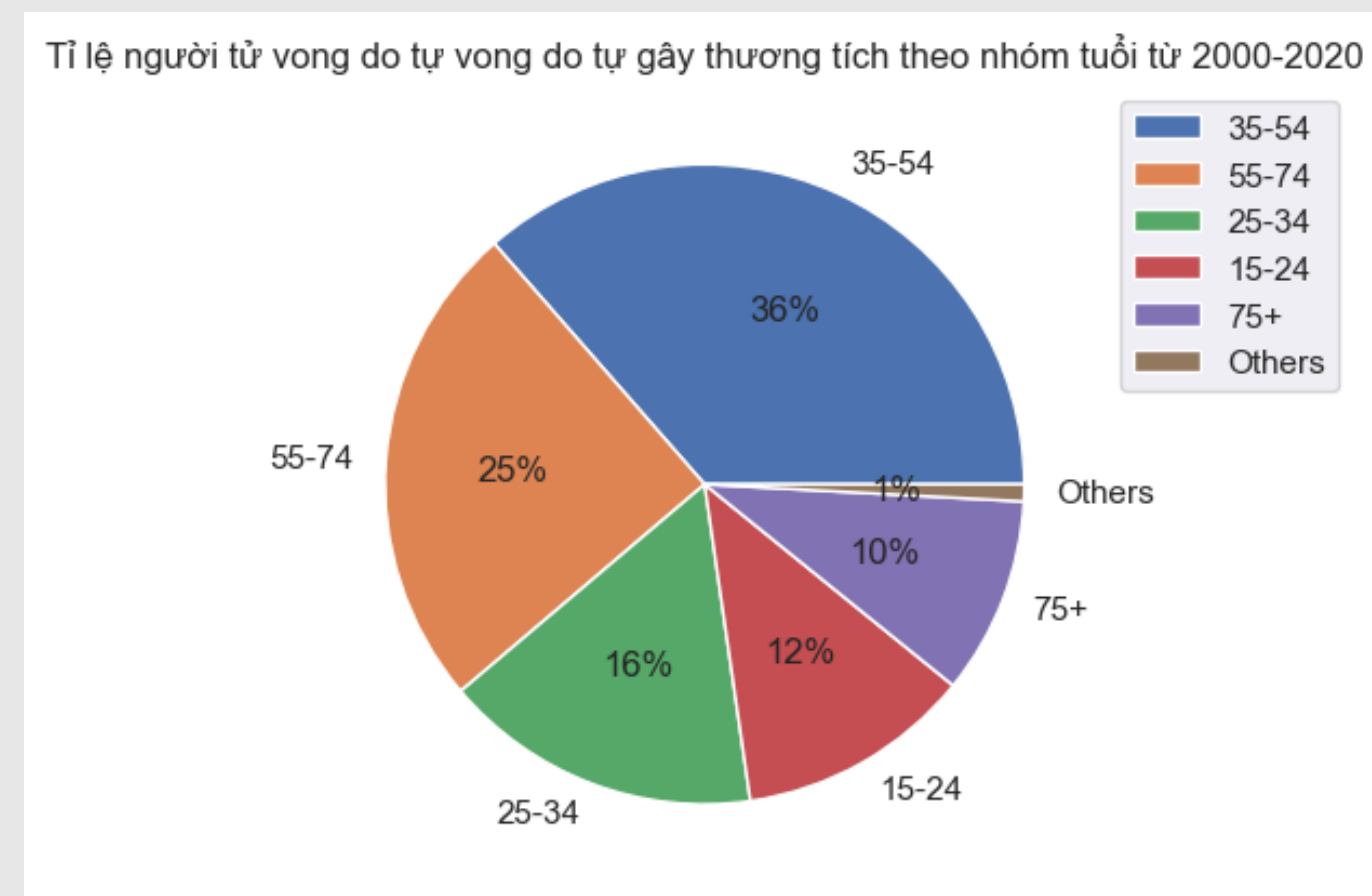
b. Tiền xử lý và trả lời các câu hỏi

Câu 3. Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?

❖ Để trả lời cho câu hỏi này, ta sẽ làm như sau:

- Gom nhóm số người mất do tự gây thương tích theo độ tuổi.
- Gộp những nhóm có tỉ lệ thấp thành Others và trực quan hóa kết quả.

Self-inflicted injuries	
Age	
35-54	1760042
55-74	1195023
25-34	777760
15-24	574349
75+	481548
Others	41120

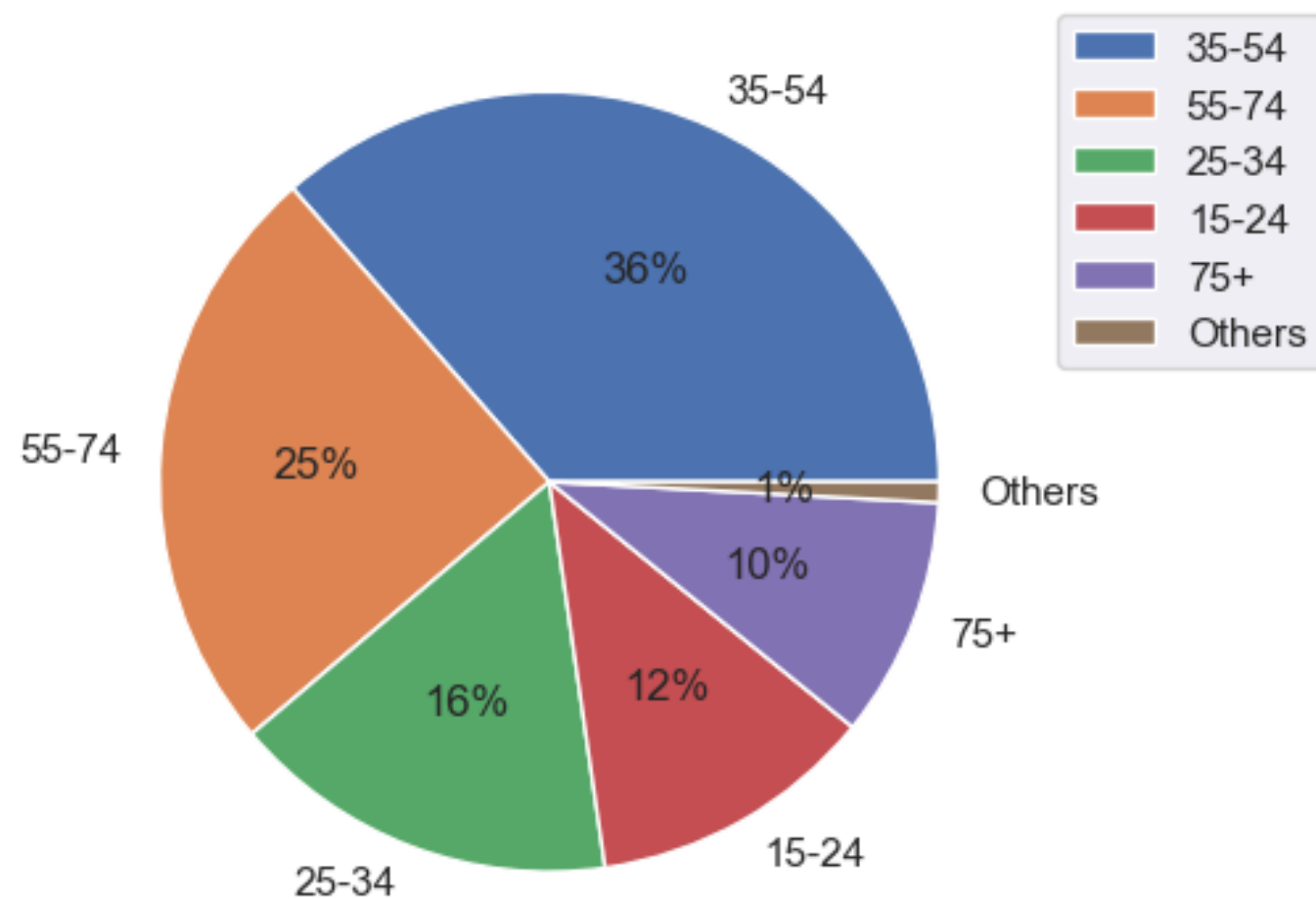


b. Tiền xử lý và trả lời các câu hỏi

Câu 3. Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?

❖ Từ kết quả trên, ta có một số nhận xét như sau:

Tỉ lệ người tử vong do tự gây thương tích theo nhóm tuổi từ 2000-2020

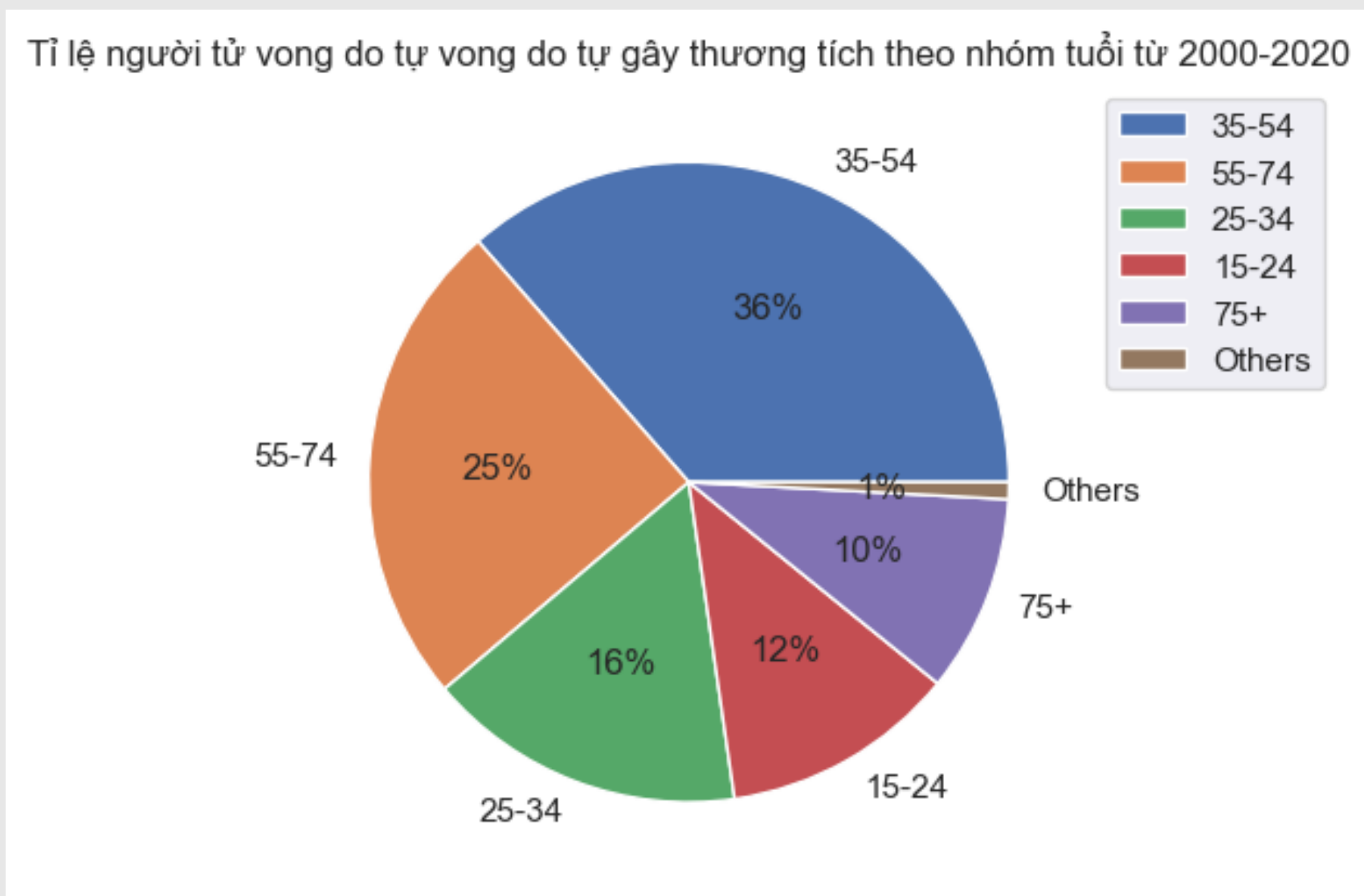


- Nhóm tuổi từ 25-75 chiếm tỉ lệ cao. Nguyên nhân chính dẫn đến tự tử có thể do áp lực công việc, cuộc sống, stress, trầm cảm.

b. Tiền xử lý và trả lời các câu hỏi

Câu 3. Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?

❖ Từ kết quả trên, ta có một số nhận xét như sau:



- Chính phủ cần tăng cường giáo dục qua các phương tiện truyền thông, đưa ra các báo cáo về vấn nạn tự tử; thực hiện các chương trình giáo dục trong giới trẻ nhằm xây dựng các kỹ năng sống cho phép họ có thể xử lý được với những căng thẳng trong cuộc sống; xác định sớm và theo dõi những người có nguy cơ tự tử, đặt biệt là những người trong nhóm tuổi từ 25-75.

b. Tiền xử lý và trả lời các câu hỏi

Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian?

- ❖ **Ý nghĩa:** Khi biết được diễn biến của các nguyên nhân gây tử vong theo thời gian sẽ giúp ta biết được nguyên nhân nào đang có xu hướng tăng, từ đó ta có thể tập trung tiền lực và nhân lực để giảm số người tử vong do nguyên nhân đó gây ra.

b. Tiền xử lý và trả lời các câu hỏi

Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian?

❖ Để trả lời cho câu hỏi này, ta sẽ làm như sau:

-Tạo dict **causes_of_death_over_time** có key là năm, value là một dict chứa số người tử vong của từng nguyên nhân trong năm tương ứng với key.

```
df1 = df[df.columns.difference(['Continent', 'Country', 'Age'])]  
df1
```


b. Tiền xử lý và trả lời các câu hỏi

Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian?

❖ Để trả lời cho câu hỏi này, ta sẽ làm như sau:

-Tạo dataframe **causes_of_death_over_time_df** có 21 dòng tương ứng với từ năm 2000 đến năm 2020 và 14 cột tương ứng với 14 nguyên nhân gây tử vong.

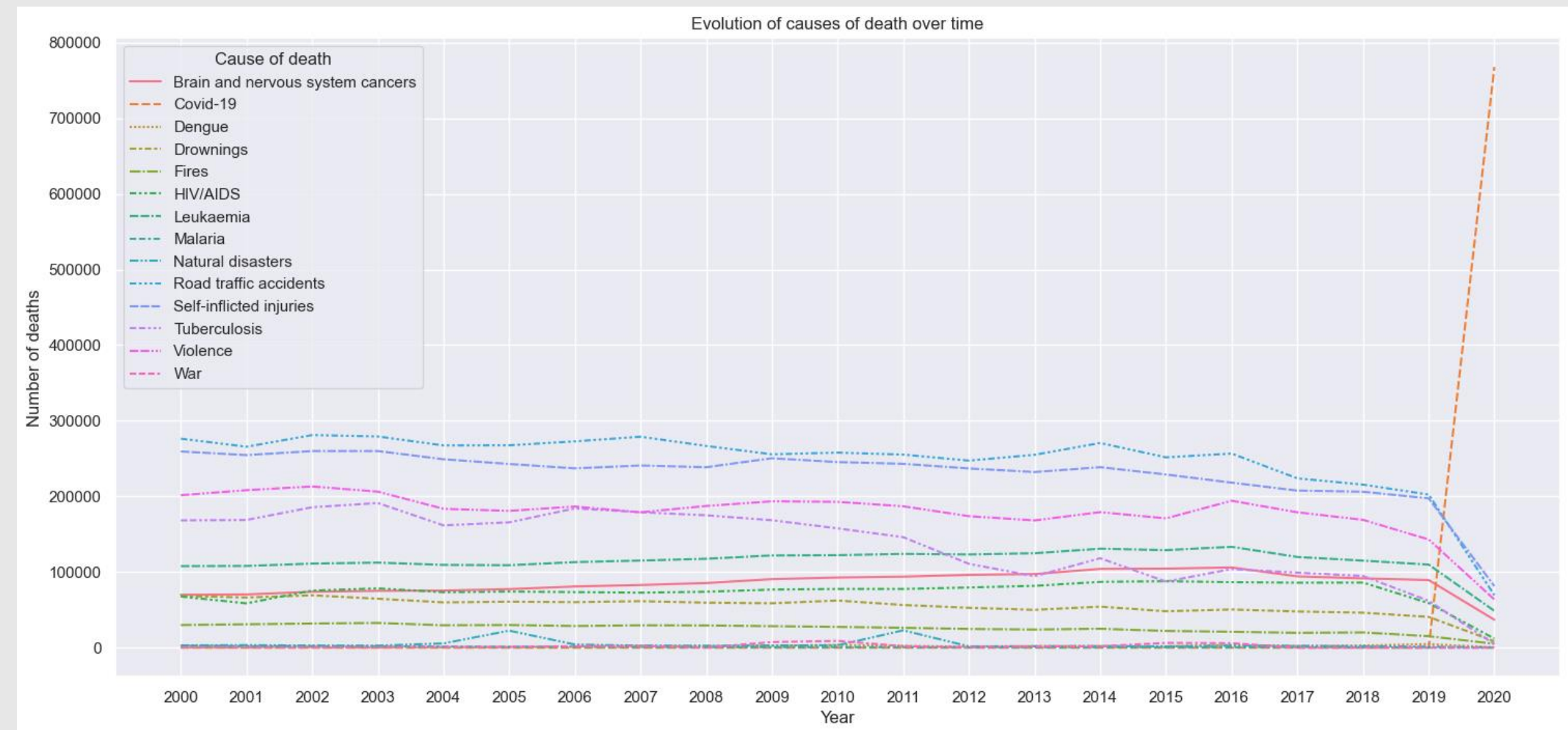
```
causes_of_death_over_time = {}  
for name, group in df1.groupby('Year'):  
    causes_of_death_over_time[name.year] = group.sum().to_dict()  
  
causes_of_death_over_time
```

b. Tiền xử lý và trả lời các câu hỏi

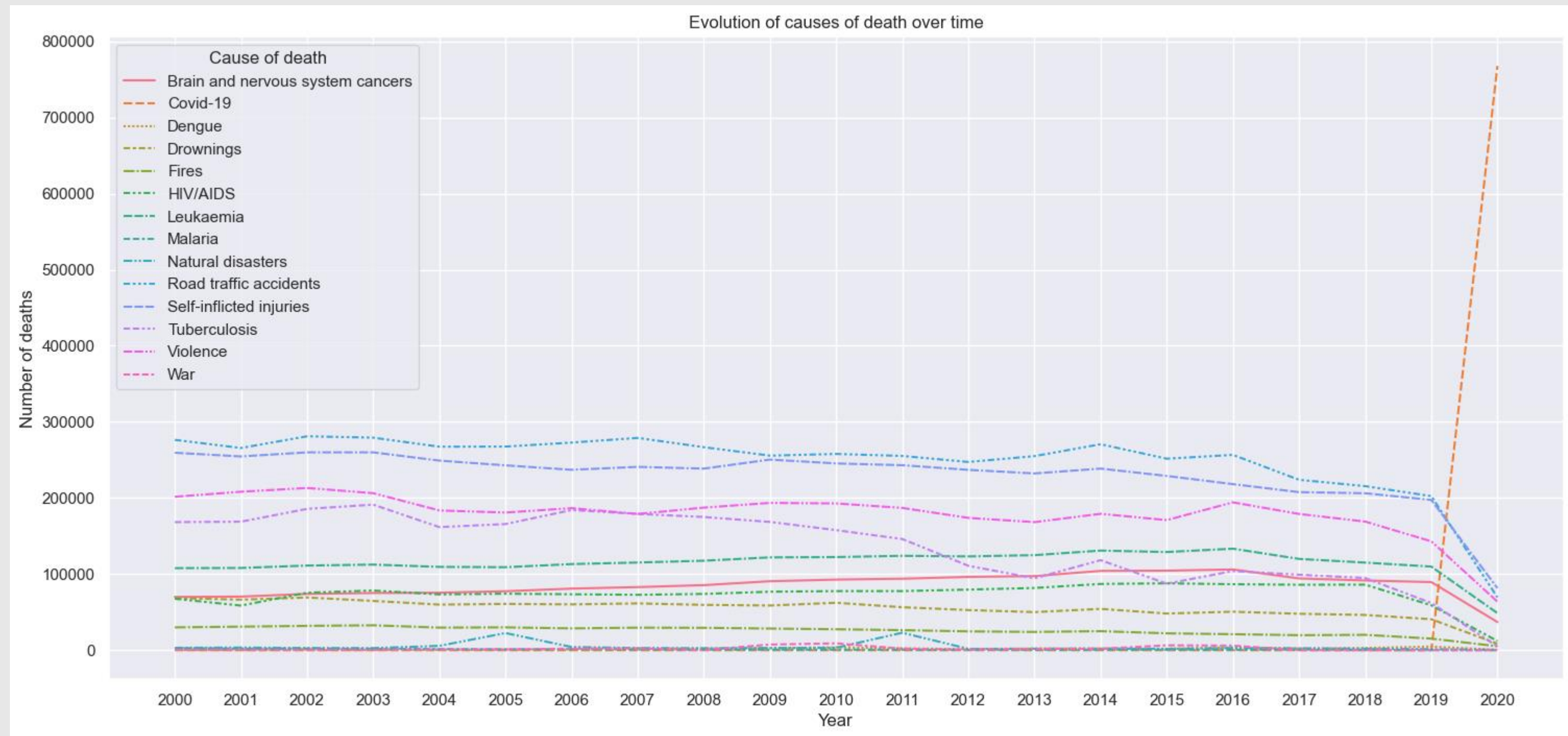
Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian?

❖ Để trả lời cho câu hỏi này, ta sẽ làm như sau:

-Vẽ đồ thị đường có 14 đường tương ứng với 14 nguyên nhân gây tử vong, trục x biểu diễn năm, trục y biểu diễn số lượng người tử vong.



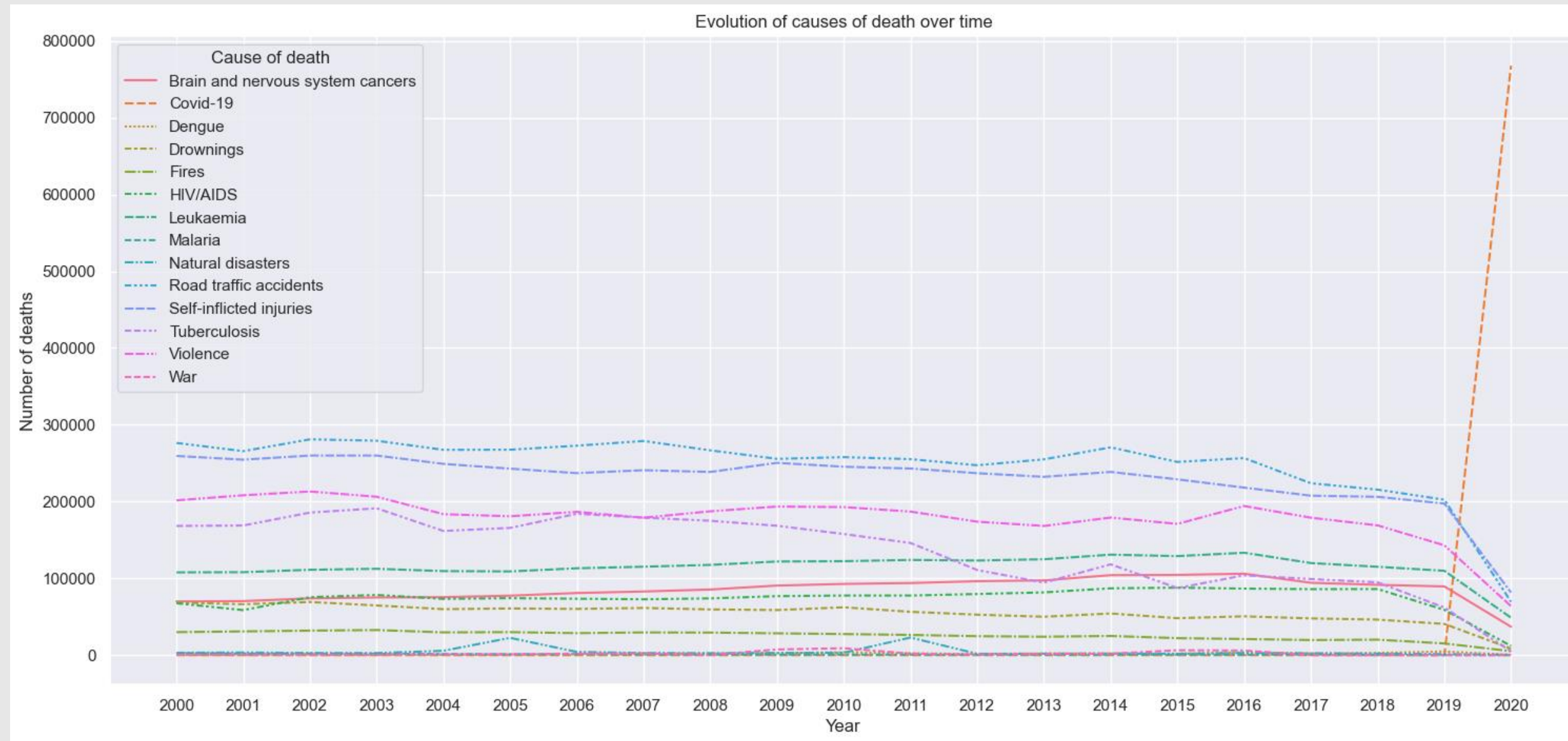
Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian



Từ đồ thị trên ta thấy được từ năm 2000 đến năm 2020:

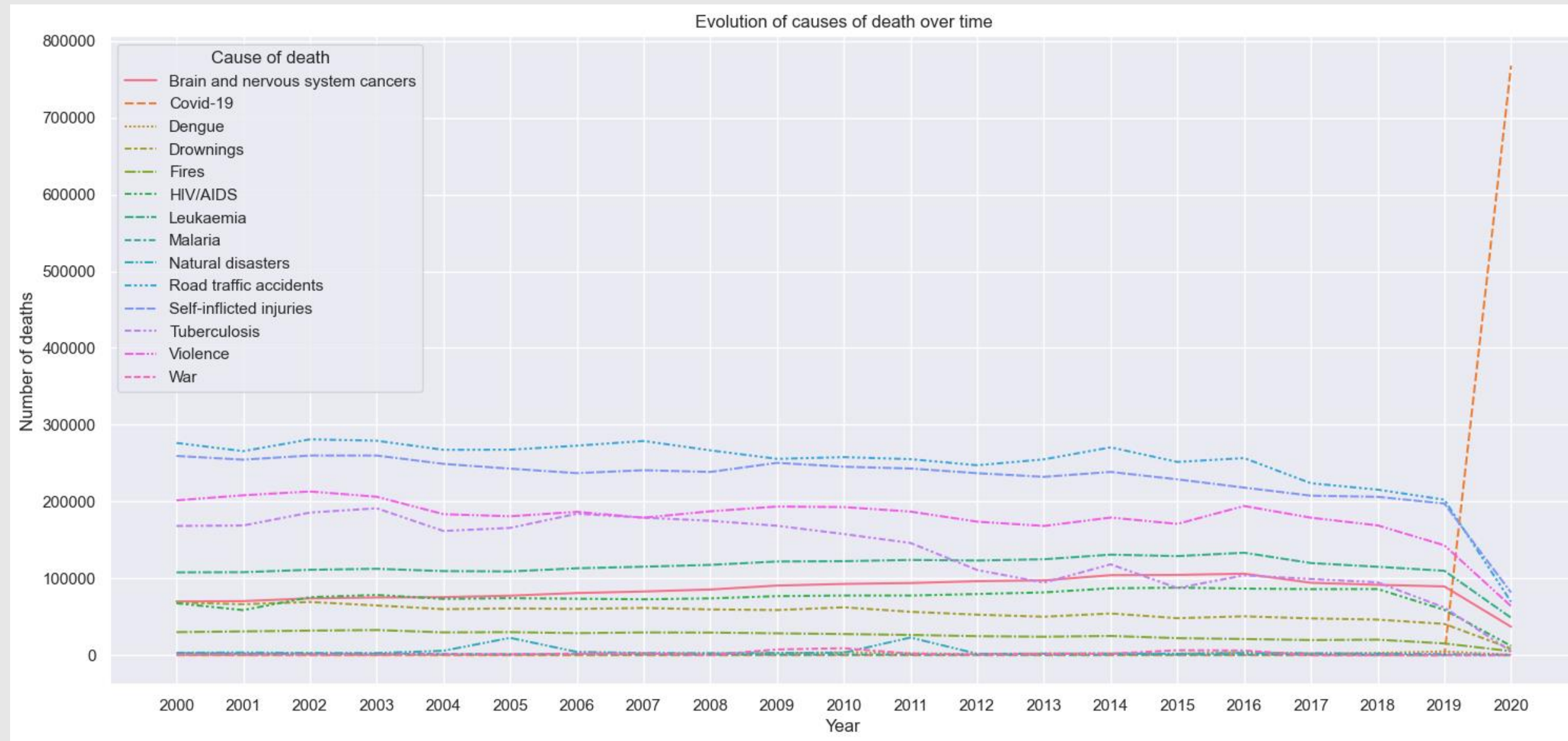
- Số người tử vong do 'Covid-19' tăng rất nhanh từ năm 2019 đến năm 2020 và là nguyên nhân gây tử vong hàng đầu trong 2 năm này.

Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian



- Số người tử vong do 13 nguyên nhân còn lại thì ổn định từ năm 2000 đến năm 2010, và đa số là có xu hướng giảm dần từ năm 2010 đến năm 2020, đặc biệt giảm một cách rõ rệt trong 2 năm là 2019 và 2020. Có thể giải thích cho việc này là vì con người đã có những biện pháp để hạn chế tử vong do 13 nguyên nhân trên gây ra. Còn trong 2 năm 2019 và 2020 giảm mạnh là vì mọi người thực hiện giãn cách xã hội để chống 'Covid-19' nên 13 nguyên nhân này không có nhiều cơ hội để gây tử vong.

Câu 4. Diễn biến của các nguyên nhân gây tử vong theo thời gian



- Những chính sách để giảm số người tử vong trong tương lai:
 - Tiếp tục thực hiện các biện pháp hiện có để làm giảm dần số người tử vong do 13 nguyên nhân còn lại.
 - Tập trung nguồn lực để hạn chế và tìm ra biện pháp chống lại 'Covid-19'.

b. Tiền xử lý và trả lời các câu hỏi

Câu 5. Mối quan hệ giữa một số nguyên nhân tử vong?

Để trả lời câu hỏi này, ta sẽ làm như sau:

- Xây dựng hàm vẽ biểu đồ phân tán theo số ca tử vong của 2 nguyên nhân của mỗi quốc gia trong từng năm.
- Những điểm dữ liệu có ít nhất một giá trị bằng 0 sẽ bị xoá.

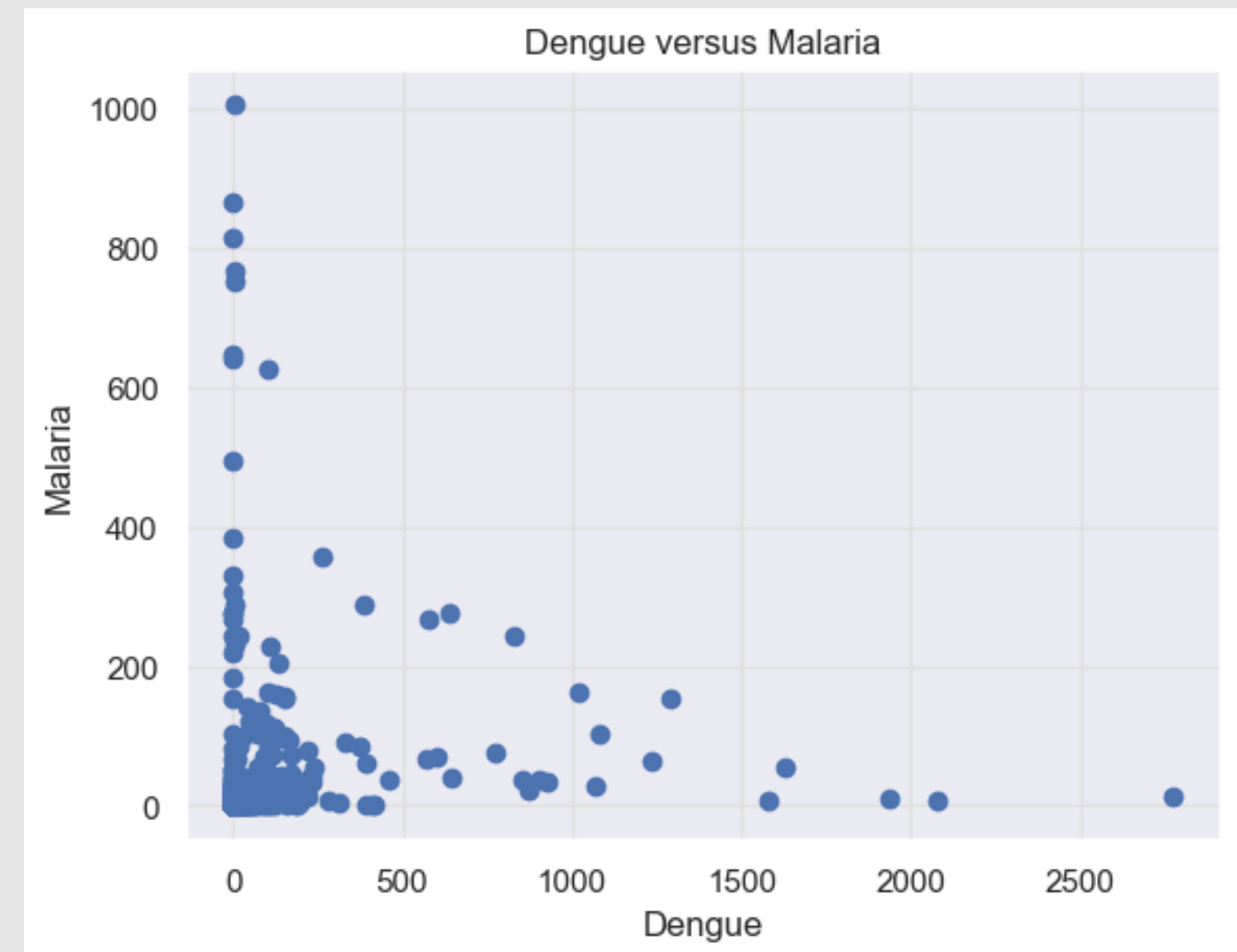
```
def pl_scatter(col1, col2):  
    pl_df = df[['Country', 'Year', col1, col2]]  
    pl_df = pl_df.groupby(['Country', 'Year']).sum().reset_index()  
    pl_df[(pl_df != 0).all(1)]  
    plt.scatter(pl_df[col1], pl_df[col2])  
    plt.title(f'{col1} versus {col2}')  
    plt.xlabel(col1)  
    plt.ylabel(col2)  
    plt.show()
```

b. Tiền xử lý và trả lời các câu hỏi

Câu 5. Mối quan hệ giữa một số nguyên nhân tử vong?

Mối liên hệ giữa bệnh sốt xuất huyết và sốt rét?

Ta biết rằng bệnh sốt xuất huyết (Dengue) và bệnh sốt rét (Malaria) đều là hai loại bệnh truyền nhiễm phổ biến ở vùng nhiệt đới và lây truyền bởi muỗi. Vậy dịch sốt xuất huyết và sốt rét có đồng hành với nhau không?

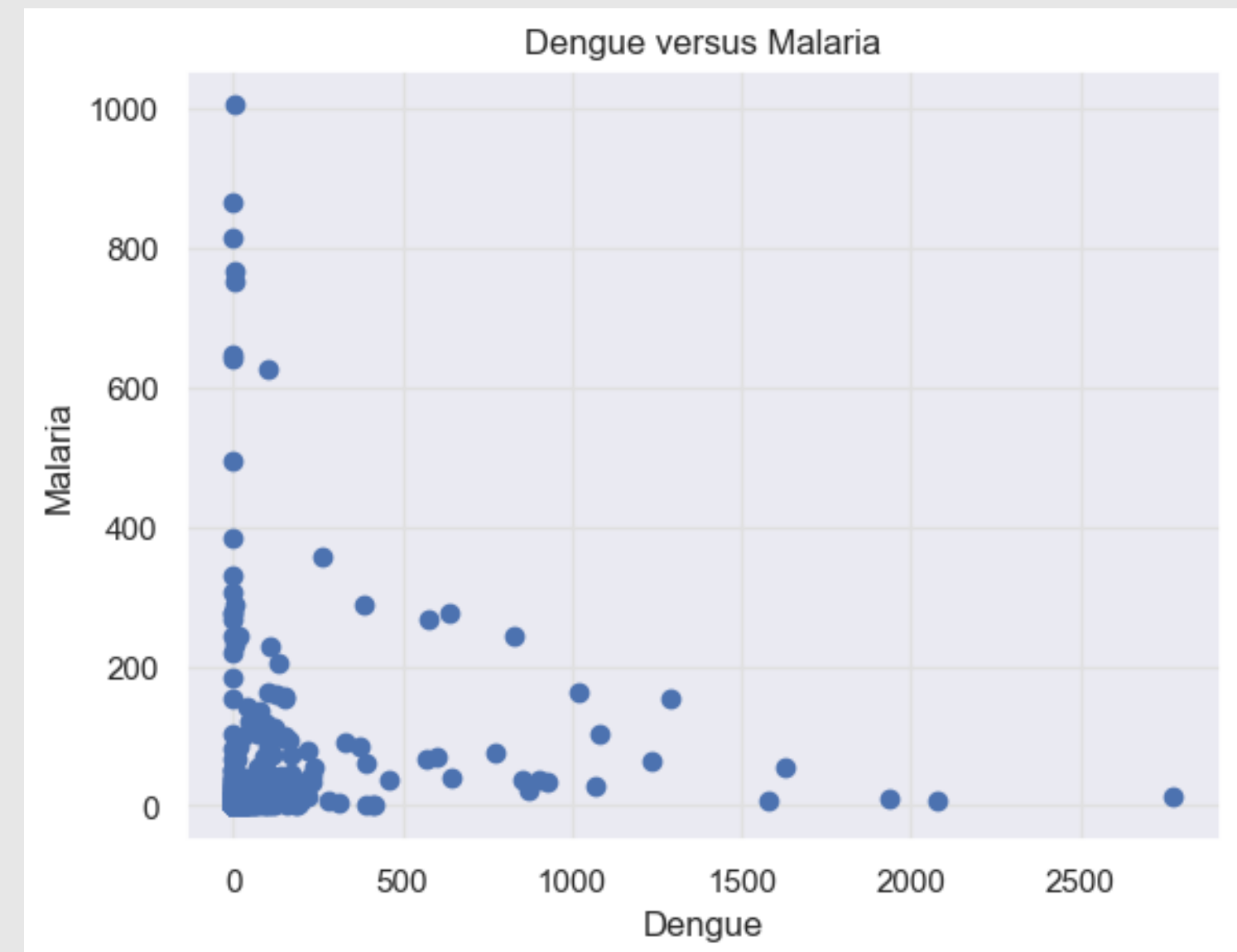


b. Tiền xử lý và trả lời các câu hỏi

Câu 5. Mối quan hệ giữa một số nguyên nhân tử vong?

Mối liên hệ giữa bệnh sốt xuất huyết và sốt rét?

→ Quan sát đồ thị, chúng ta sẽ thấy sốt rét và sốt xuất huyết thường không xuất hiện cùng với nhau, khi sốt rét tăng thì sốt xuất huyết giảm và ngược lại. Như vậy, khi một trong hai dịch bệnh này xảy ra thì chúng ta nên tập trung xử lý dịch bệnh đó, tránh dàn trải nguồn lực.

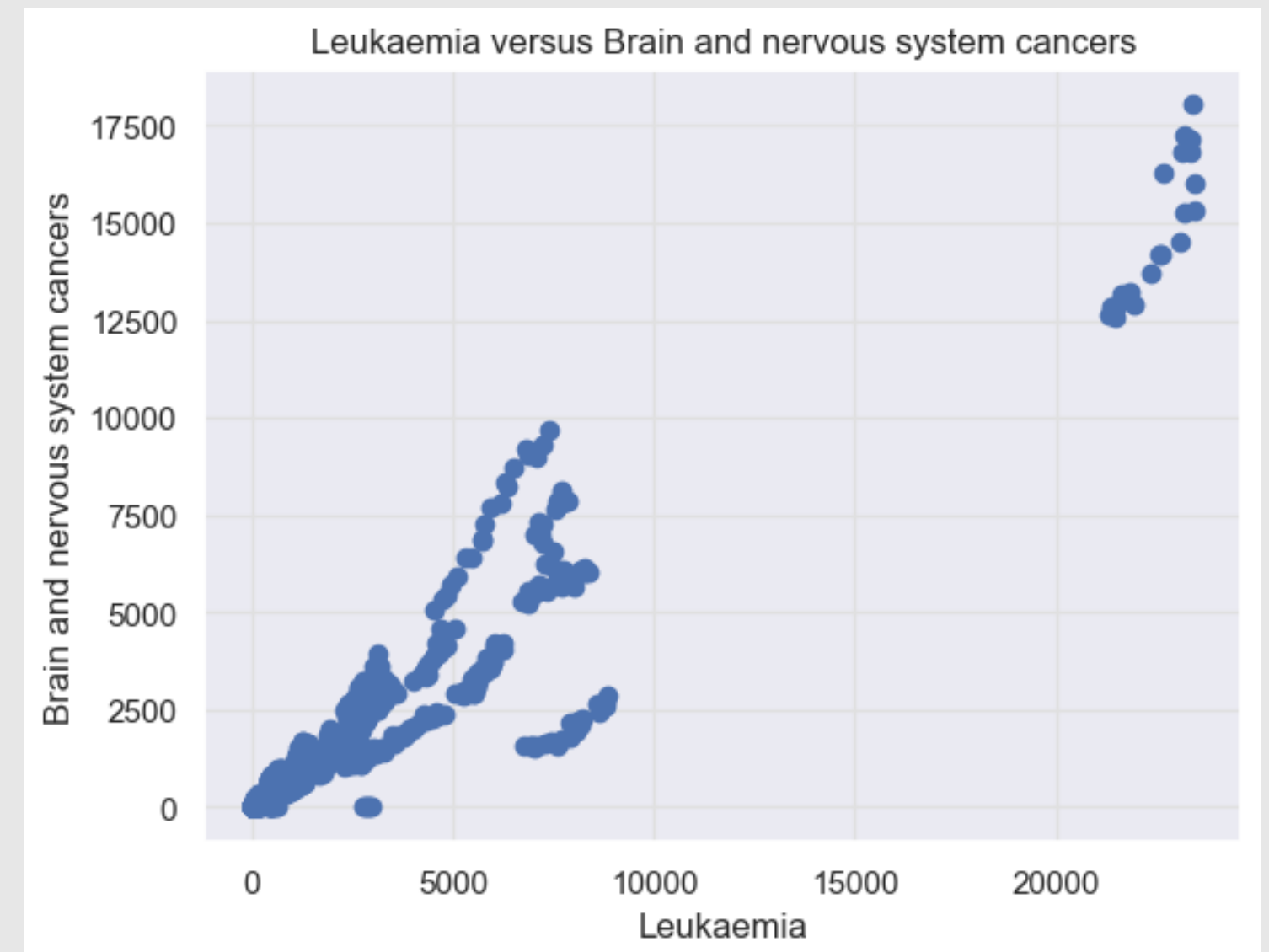


b. Tiền xử lý và trả lời các câu hỏi

Câu 5. Mối quan hệ giữa một số nguyên nhân tử vong?

Mối liên hệ giữa bệnh bạch cầu và ung thư não/hệ thần kinh?

Bệnh bạch cầu cũng là một loại ung thư và biểu đồ cho bệnh bạch cầu có xu hướng tăng khi bệnh ung thư não/hệ thần kinh tăng. Có thể thấy các bệnh ung thư nói chung có xu hướng liên quan chặt chẽ với nhau.



b. Tiền xử lý và trả lời các câu hỏi

Câu 6. So sánh phân phối tuổi tử vong?

Ý nghĩa:

- Như chúng ta đã biết, tuổi thọ (tương đương tuổi tử vong) là một chỉ số quan trọng để đánh giá chất lượng cuộc sống, chỉ số phát triển con người và mức độ tiến bộ của một quốc gia, khu vực. Khi chất lượng cuộc sống được cải thiện, hệ thống y tế tốt, tuổi thọ sẽ tăng lên và số lượng tử vong sẽ lệch về nhóm tuổi cao hơn.

- Ngoài ra, phân phối này còn cho ta thấy tỉ lệ sống sót của trẻ em (dưới 5 tuổi), là một chỉ số được Liên Hợp Quốc dùng để đánh giá tiêu chuẩn phát triển trẻ em. Chúng ta cùng tìm hiểu qua các bước sau.

b. Tiền xử lý và trả lời các câu hỏi

Câu 6. So sánh phân phối tuổi tử vong?

❖ Để trả lời câu hỏi này, ta sẽ làm như sau:

Xây dựng hàm `pl_line` dùng để tính tỉ lệ tử vong theo từng nhóm tuổi.

```
def pl_line(arg_df):  
    # lấy các giá trị cần thiết và tính tổng theo độ tuổi  
    tmp_df = arg_df.iloc[:, 3:]  
    tmp_df = tmp_df.groupby('Age').sum().reset_index()  
    # Tính tổng và tỉ lệ của các độ tuổi  
    tmp_df['Total'] = tmp_df.iloc[:, 1:].sum(axis=1)  
    S = tmp_df['Total'].sum()  
    tmp_df['Rate'] = tmp_df['Total'] / S  
    # sắp xếp tuổi theo đúng thứ tự  
    tmp_df['Age'] = tmp_df['Age'].apply(  
        lambda x: '0' + x if len(x.split('-')[0]) < 2 else x  
    )  
    tmp_df = tmp_df.sort_values(by=['Age']).reset_index()  
    # trả về tuổi và tỉ lệ  
    return tmp_df[['Age', 'Rate']]
```

b. Tiền xử lý và trả lời các câu hỏi

Câu 6. So sánh phân phối tuổi tử vong?

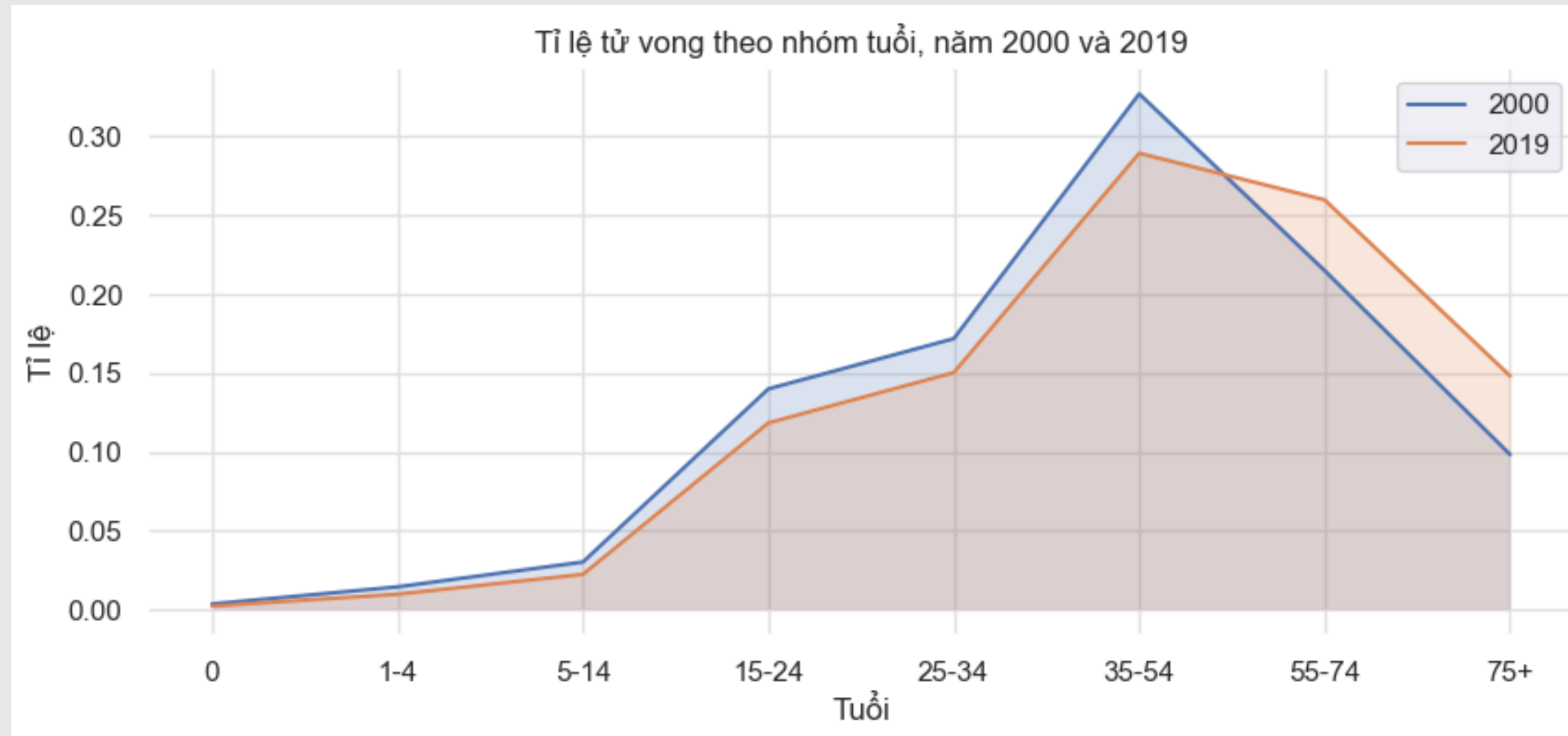
❖ Để trả lời câu hỏi này, ta sẽ làm như sau:

Query lấy dữ liệu và trực quan hóa.

```
query_2000 = '"2000" in Year'      # query lấy dữ liệu năm 2000
query_2019 = '"2019" in Year'      # query lấy dữ liệu năm 2019
query_Canada = 'Country == "Canada"' # query lấy dữ liệu Canada
query_Nicaragua = 'Country == "Nicaragua"' # query lấy dữ liệu Nicaragua
query_Af1519 = 'Continent == "Africa" and Year >= "2015-01-01" and Year <= "2019-12-31"' # query lấy dữ liệu Châu Phi từ năm 2015 - 2019
query_As1519 = 'Continent == "Asia" and Year >= "2015-01-01" and Year <= "2019-12-31"' # query lấy dữ liệu Châu Á từ năm 2015 - 2019
query_Eu1519 = 'Continent == "Europe" and Year >= "2015-01-01" and Year <= "2019-12-31"' # query lấy dữ liệu Châu Âu từ năm 2015 - 2019
```

Câu 6. So sánh phân phối tuổi tử vong

Tuổi tử vong đã thay đổi như thế nào từ năm 2000 đến 2019?

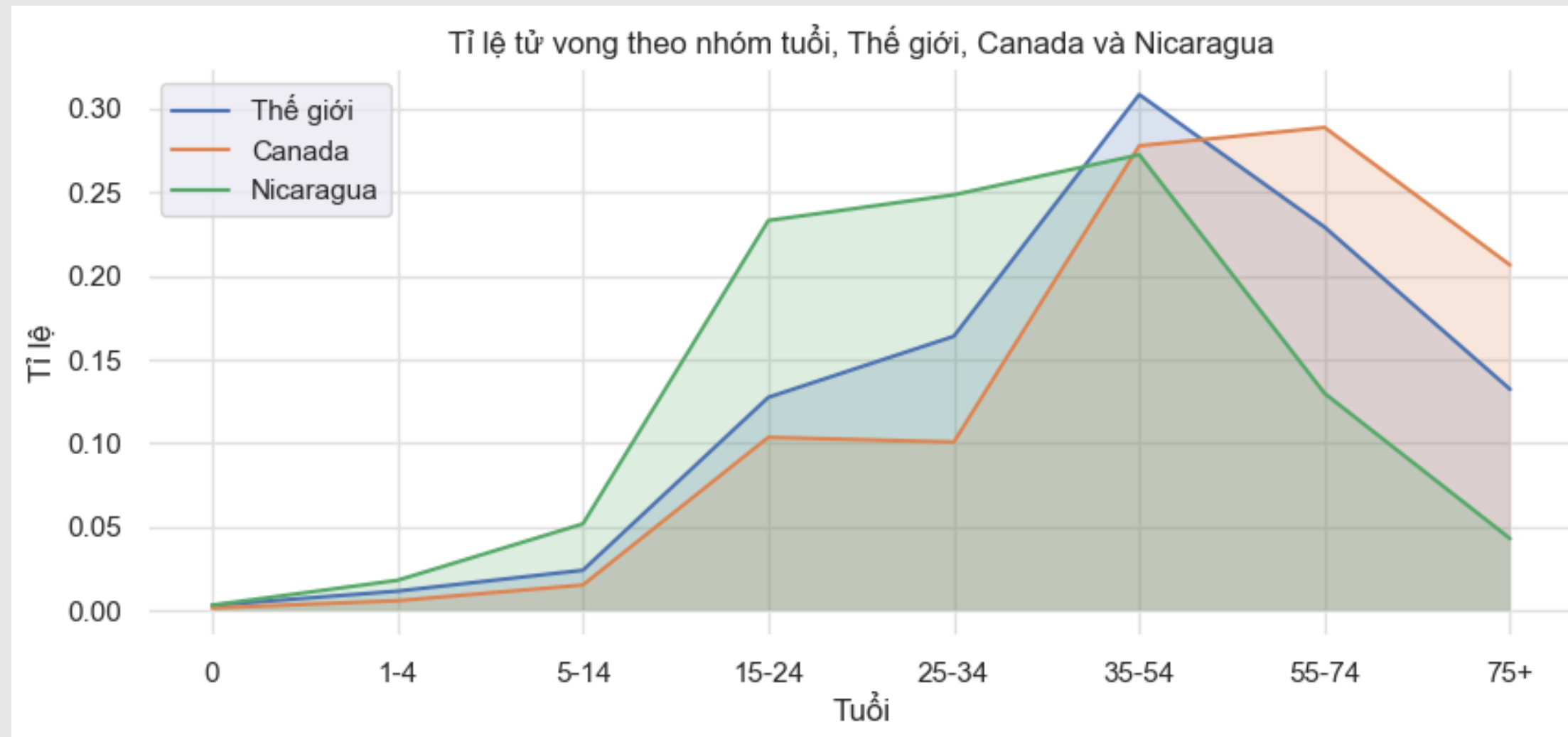


- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (2000): 1.8%
- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (2019): 1.2%

- Quan sát đồ thị có thể thấy năm 2019 so với năm 2000, tỉ lệ tử vong ở nhóm tuổi dưới 54 giảm và trên 54 tăng, đồ thị 2019 lệch qua phải. Vậy tuổi thọ đã được gia tăng và chất lượng cuộc sống được cải thiện trong 19 năm qua.

Câu 6. So sánh phân phối tuổi tử vong

Tuổi tử vong ở các nước khác nhau?

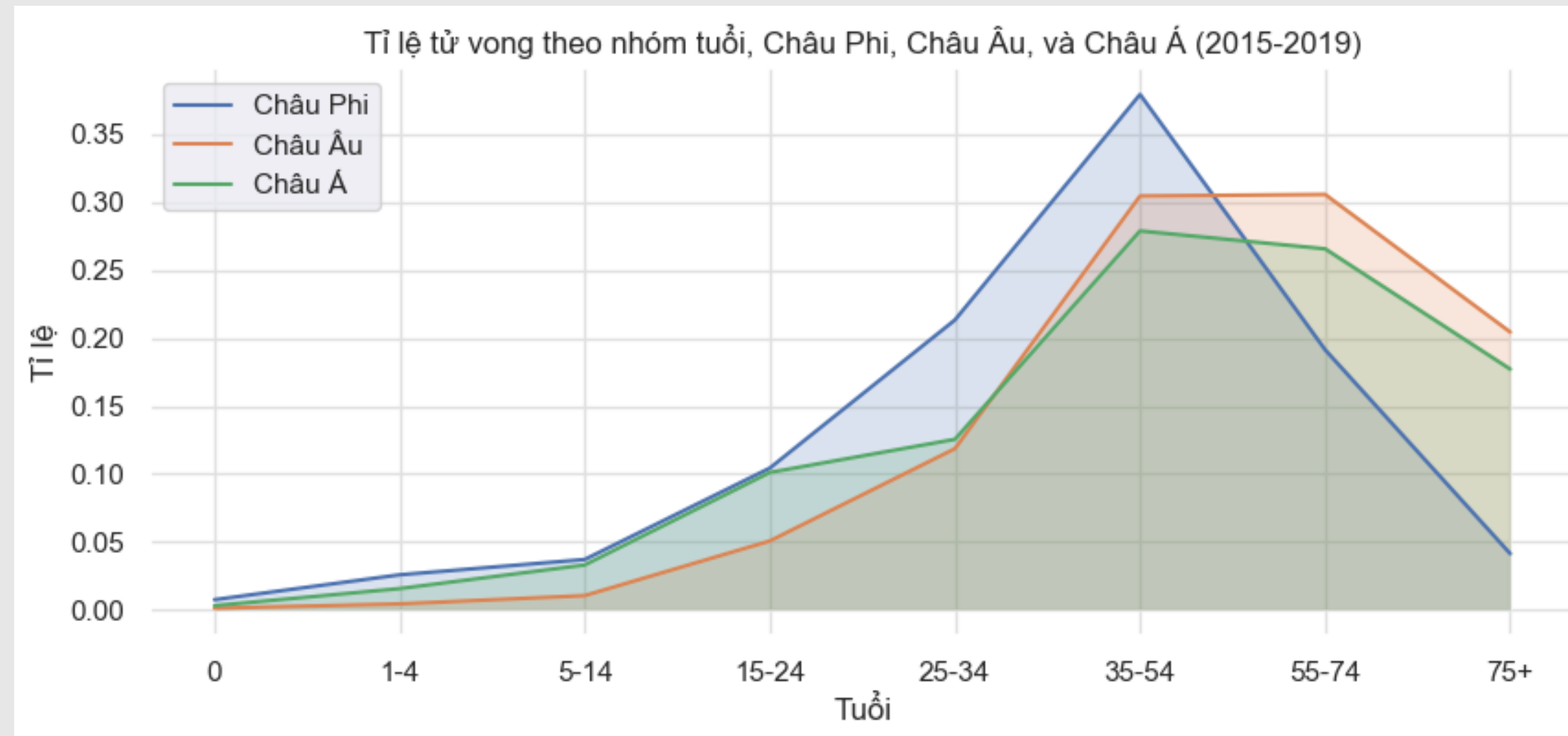


- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (Thế giới): 1.5%
- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (Canada): 0.8%
- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (Nicaragua): 2.1%

- Đồ thị của Canada lệch về bên phải so với thế giới còn Nicaragua lệch về bên trái. Vậy tuổi thọ của Nicaragua thấp, Canada cao và có thể đánh giá Canada phát triển hơn Nicaragua.

Câu 6. So sánh phân phối tuổi tử vong

Tuổi tử vong ở các châu lục khác nhau trong những năm gần đây.



- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (châu Phi): 3.3%
- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (châu Âu): 0.6%
- Tỉ lệ tử vong của trẻ em dưới 5 tuổi (châu Á): 1.9%

Nhìn vào đồ thị, có thể thấy châu Á phát triển hơn châu Phi nhưng chưa bằng châu Âu.

4

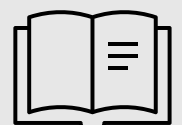
Mô hình hóa dữ liệu



Bài toán: Dự đoán số người mất do tai nạn giao thông trên thế giới.

Ý nghĩa:

- Sức khỏe: Số người tử vong trong một năm là thước đo về tình hình sức khỏe cộng đồng trong một năm đó, ảnh hưởng bởi các yếu tố như dịch bệnh, thiên tai, chiến tranh, đói nghèo,... Số lượng người tử vong thể hiện được mức độ nghiêm trọng của các tác nhân này. Dự đoán số người tử vong trong tương lai, chúng ta sẽ biết tình hình đang được cải thiện hay xấu đi từ đó đưa ra những hành động phù hợp.
- Dân số: Số lượng tử vong là một trong những yếu tố chính tác động đến dân số. Dự đoán số người tử vong cho phép chúng ta dự đoán dân số, rất cần thiết cho công việc chỉ đạo, hoạch định chính sách, lập kế hoạch.



a. Lựa chọn mô hình và những đặc trưng quan trọng cho bài toán

Lựa chọn mô hình cho bài toán:

- Nhóm lựa chọn hai mô hình để so sánh với nhau là mô hình hồi quy tuyến tính và mô hình hồi quy đa thức (bậc 2)
- Lý do chọn hai mô hình này:
 - Nhìn vào biểu đồ phân tán (xem phần dưới): Sự phân tán của các điểm có vẻ nằm trên đường thẳng hoặc cong một chút của parabol.
 - Hai mô hình này đơn giản và trực quan, thích hợp với tập dữ liệu ít của bài toán.
 - Chọn hai mô hình để so sánh, đối chiếu với nhau.

a. Lựa chọn mô hình và những đặc trưng quan trọng cho bài toán

Lựa chọn đặc trưng quan trọng cho bài toán:

- Những đặc trưng quan trọng cho bài toán mà chúng ta có thể chọn từ tập dữ liệu là **Year**
- Giải thích: Ta không chọn **Continent**, **Country**, **Age**, **Cause of death** vì chúng là những đặc trưng cố định và không thay đổi theo thời gian.

b. Tiền xử lý dữ liệu

Ta tiến hành gom nhóm dữ liệu theo cột **Year**:

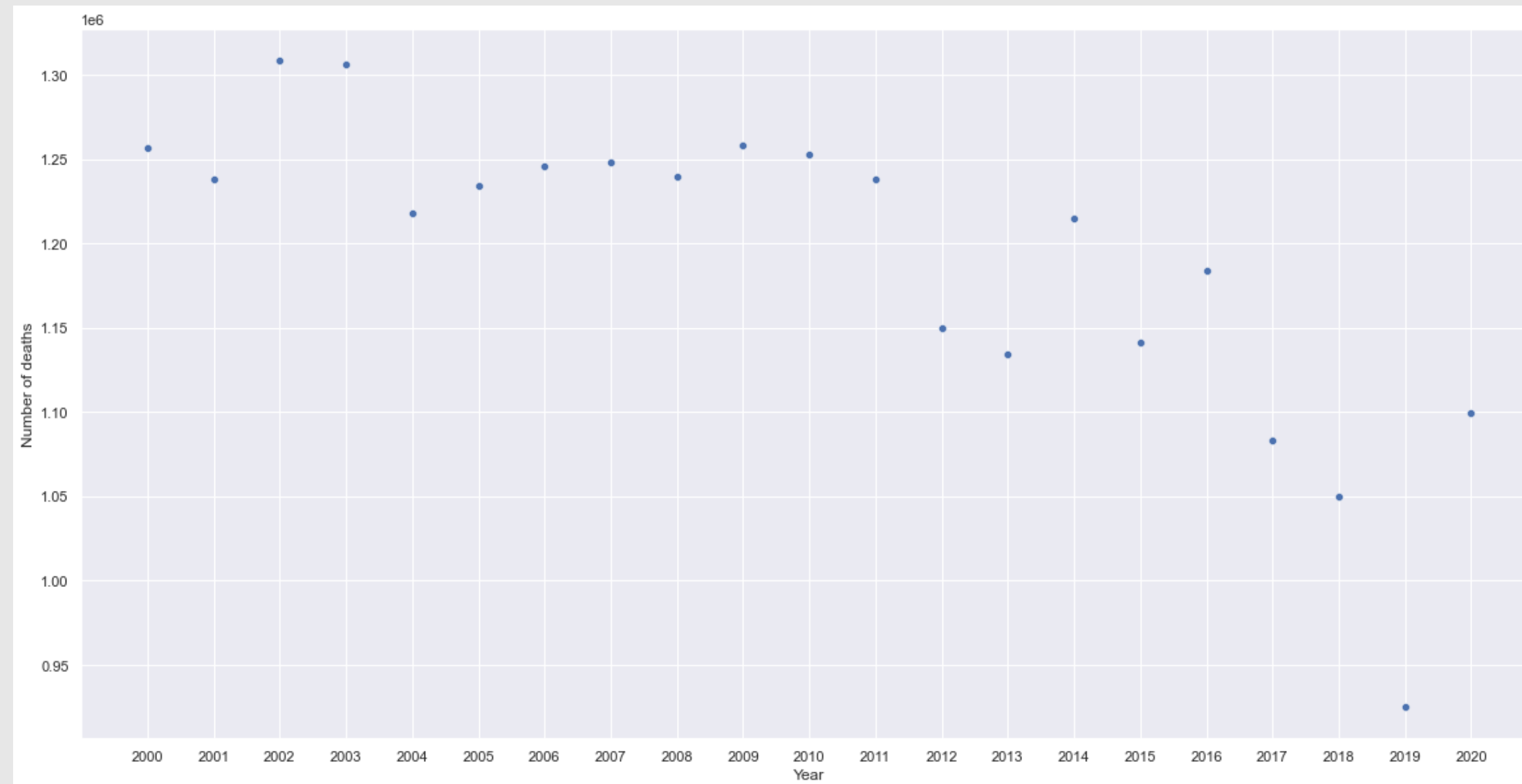
```
years, number_of_deaths = [], []
for name, group in df.drop(['Continent', 'Country', 'Age'], axis=1).groupby('Year'):
    years.append(int(name.split('-')[0]))
    number_of_deaths.append(group.drop('Year', axis=1).sum().sum())

model_df = pd.DataFrame({'Year': years, 'Number of deaths': number_of_deaths})
model_df.head()
```

	Year	Number of deaths
0	2000	1256601
1	2001	1238378
2	2002	1308353
3	2003	1306450
4	2004	1218349

c. Khám phá sự phân tán dữ liệu

Dữ liệu phân tán theo biểu đồ sau:



d. Xác thực các siêu tham số của mô hình bằng tập kiểm định (validation set)

- Vì chúng ta chỉ có duy nhất 1 đặc trưng là **Year** nên nó chính là siêu tham số duy nhất của mô hình.
- Sử dụng validation set giúp chúng ta đánh giá hiệu quả của mô hình.

```
x = np.array(model_df['Year']).reshape(-1, 1)
y = np.array(model_df['Number of deaths']).reshape(-1, 1)

x_train, x_validation, y_train, y_validation = train_test_split(x, y, train_size = 0.75)
```

e. Các mô hình hồi quy

❖ Mô hình hồi quy tuyến tính:

$$y = w_0 + w_1 * x$$

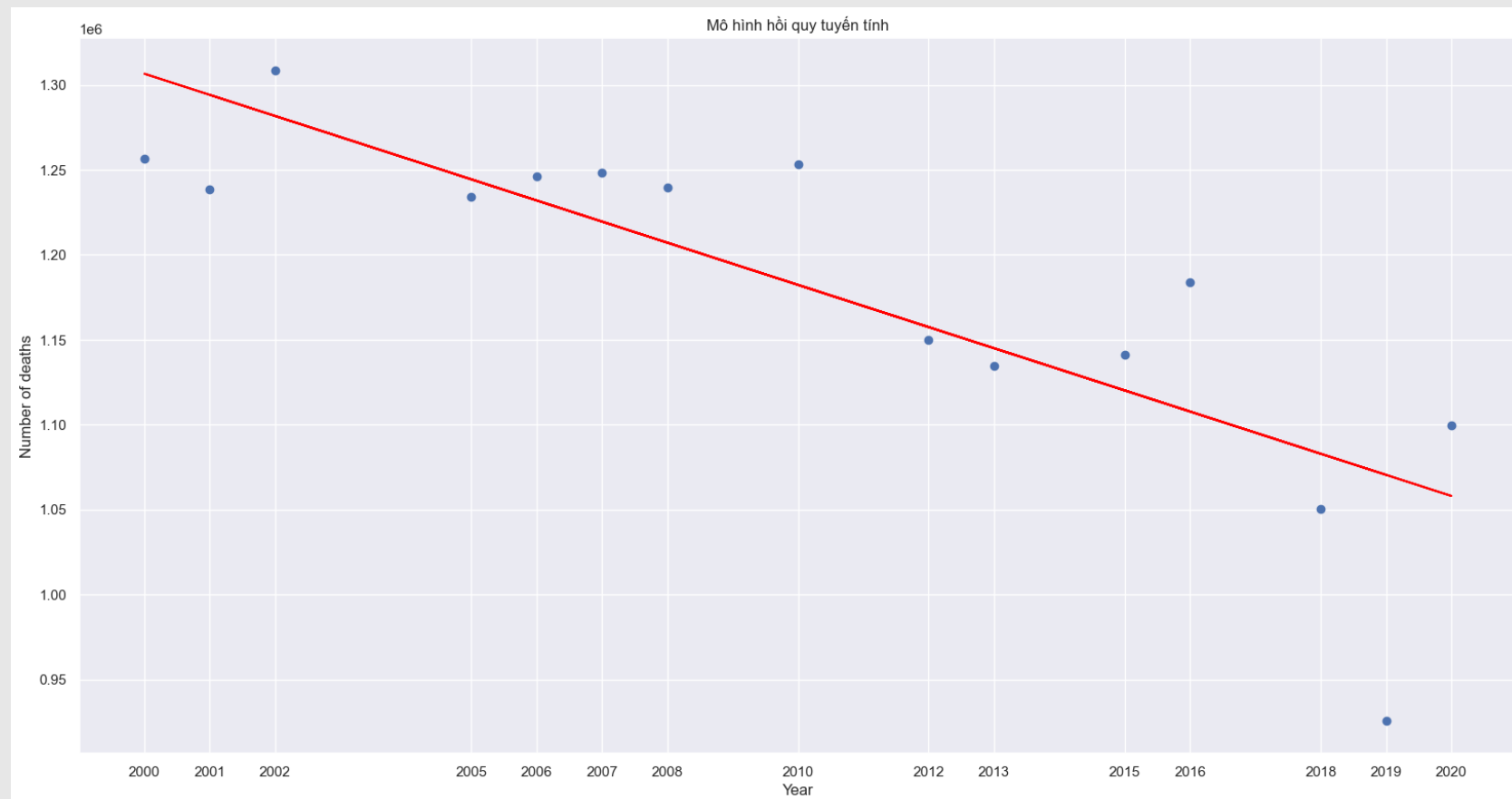
- Thực hiện mô hình:
 - Khởi tạo mô hình
 - "fit" mô hình vào dữ liệu
 - Xem giá trị của các tham số của mô hình

```
w_1: [[-12426.9939564]]  
w_0: [26160531.984891]
```

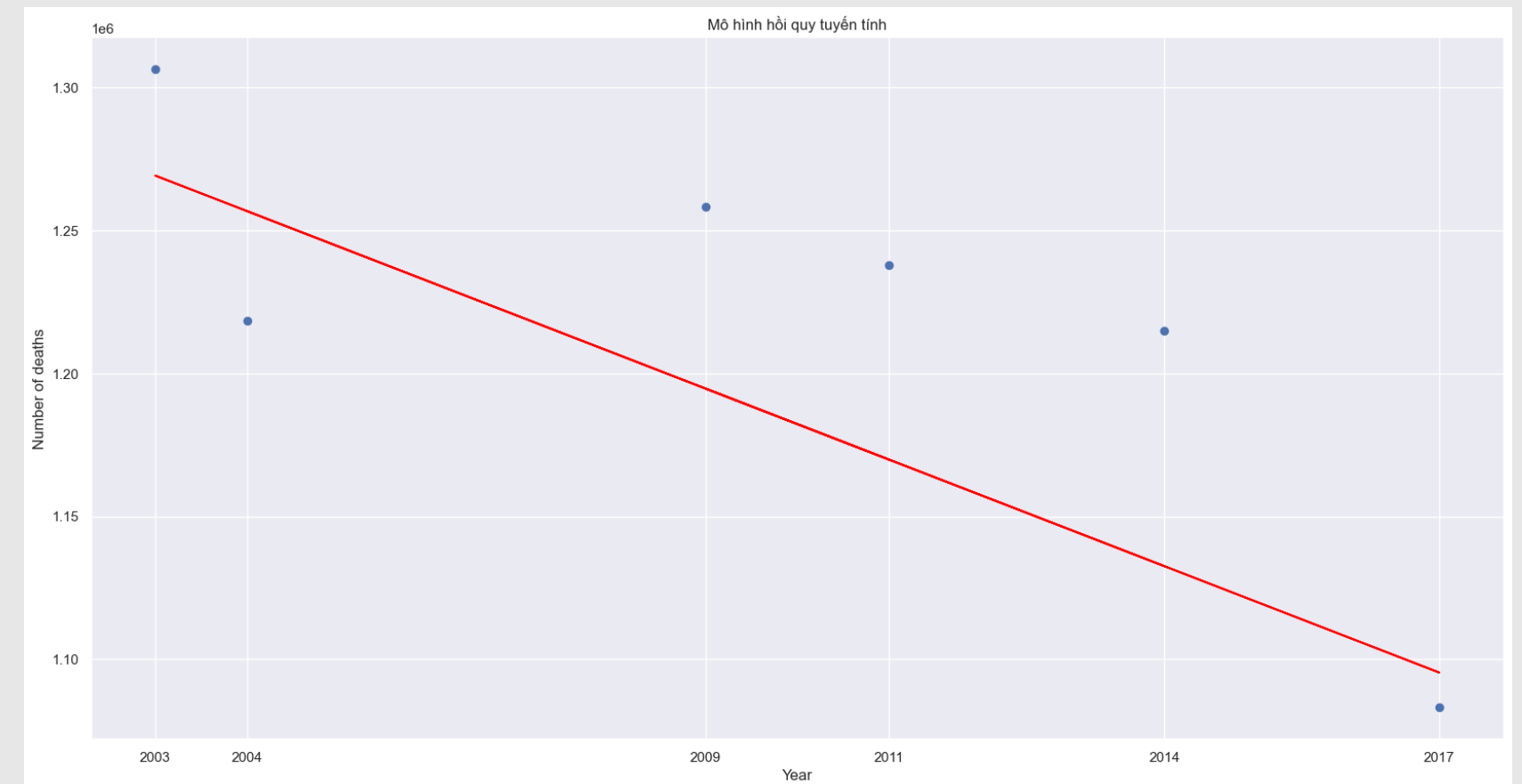
e. Các mô hình hồi quy

❖ Mô hình hồi quy tuyến tính:

- Trực quan hóa mô hình tìm được:



Trên tập train



Trên tập validation

e. Các mô hình hồi quy

❖ Mô hình hồi quy tuyến tính:

- Đánh giá hiệu suất mô hình bằng MSE và RMSE:

MSE: 3076244245.1006446

RMSE: 55463.90037763883

Hiệu suất của mô hình bằng MSE và RMSE

e. Các mô hình hồi quy

❖ Mô hình hồi quy đa thức:

$$y = w_0 + w_1 * x + w_2 * x^2$$

Ta chọn đa thức bậc 2 vì từ bước "Khám phá sự phân tán dữ liệu" ta thấy dữ liệu phân tán giống như một phần của đồ thị đa thức bậc 2. (Parabol bị úp ngược)

e. Các mô hình hồi quy

❖ Mô hình hồi quy đa thức:

- Tạo mới tập `X_train` và `X_validation` cho phù hợp với đa thức:

```
poly_x_train = poly.fit_transform(X_train)  
poly_x_validation = poly.fit_transform(X_validation)
```

e. Các mô hình hồi quy

❖ **Mô hình hồi quy đa thức:** $y = w_0 + w_1 * x + w_2 * x^2$

- Thực hiện mô hình:
 - Khởi tạo mô hình
 - "fit" mô hình vào dữ liệu
 - Xem giá trị của các tham số của mô hình

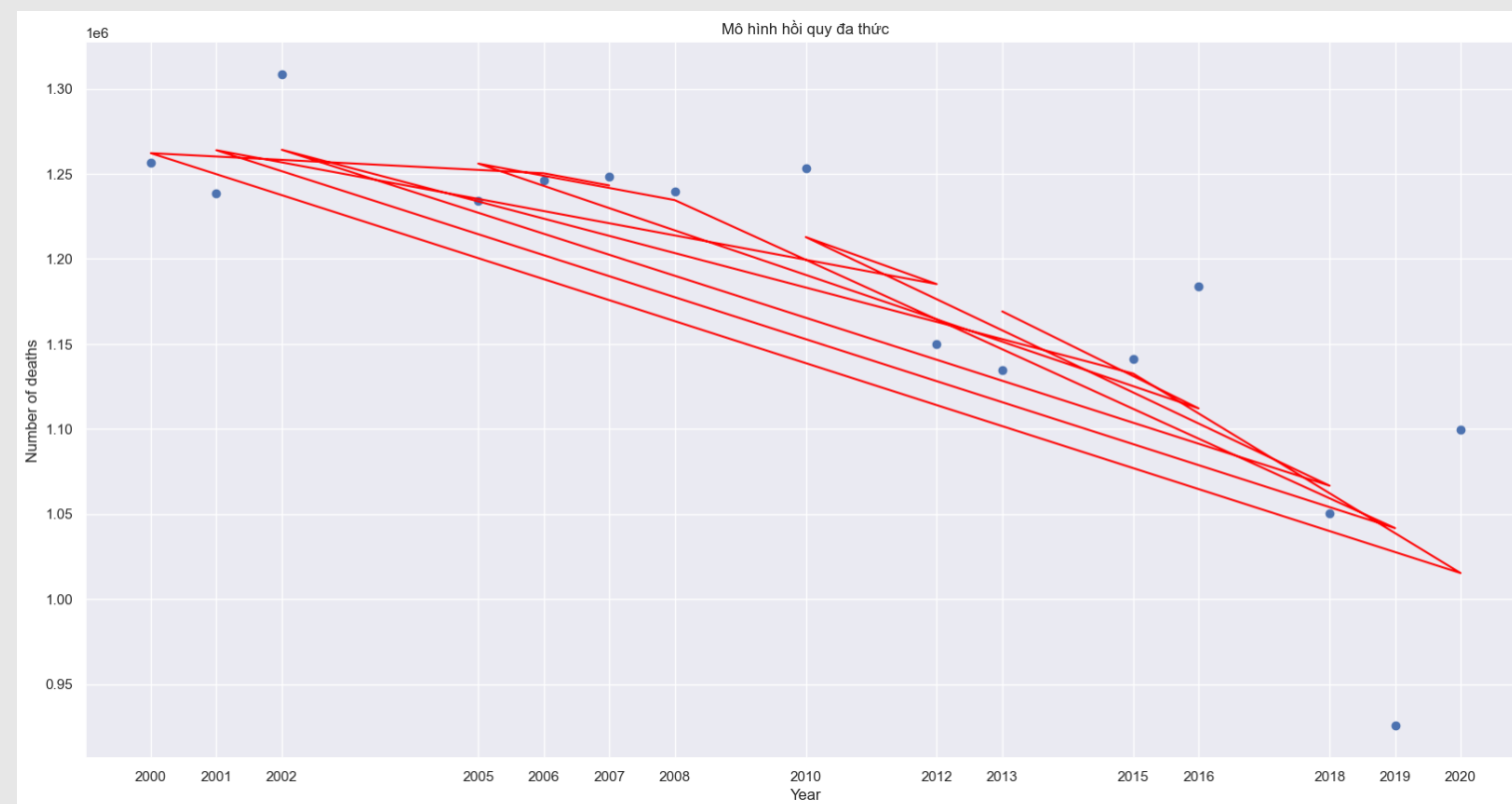
`w_1, w_2: [[2.96662953e+06 -7.41038068e+02]]`

`w_0: [-2.96784467e+09]`

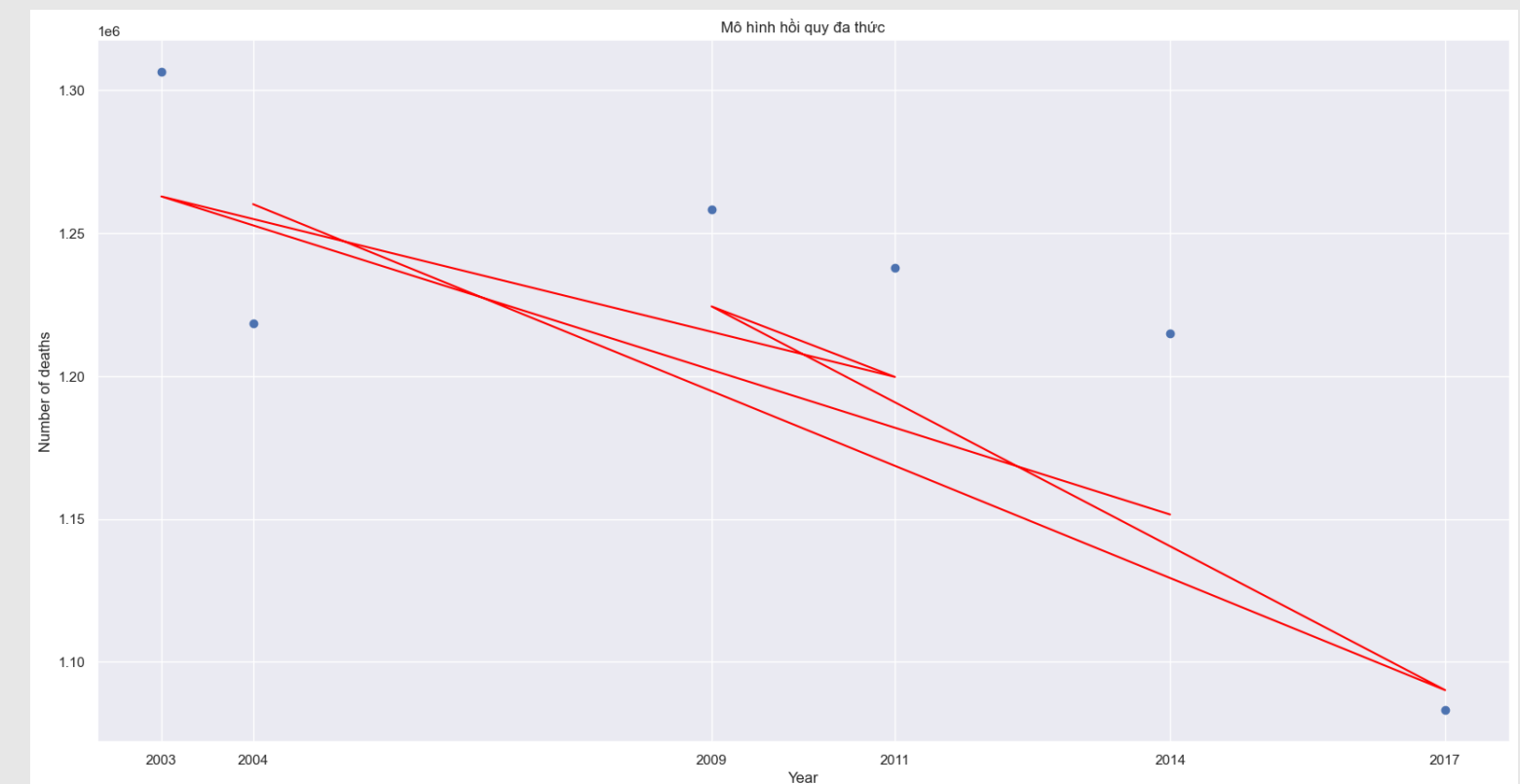
e. Các mô hình hồi quy

❖ **Mô hình hồi quy đa thức: $y = w_0 + w_1 * x + w_2 * x^2$**

- Trực quan hóa mô hình tìm được:



Trên tập train



Trên tập validation

e. Các mô hình hồi quy

❖ Mô hình hồi quy đa thức:

- Đánh giá hiệu suất mô hình bằng MSE và RMSE:

MSE: 1718225395.0259883
RMSE: 41451.48242253814

Hiệu suất của mô hình bằng MSE và RMSE

f. So sánh hai mô hình:

- **MSE** là giá trị trung bình của chênh lệch bình phương giữa tham số dự đoán và tham số quan sát được. Còn **RMSE** được tính bằng căn bậc hai của **MSE**.
- Một giá trị **MSE** (hoặc **RMSE**) càng bé thì có nghĩa là những giá trị ước lượng sẽ càng sát với giá trị thực và do đó: Một mô hình có **MSE (RMSE)** bé hơn sẽ là một mô hình tốt hơn.:

→ Vì **MSE**("Mô hình hồi quy đa thức") < **MSE** ("Mô hình hồi quy tuyến tính") nên "Mô hình hồi quy đa thức" tốt hơn "Mô hình hồi quy tuyến tính"

4

Tổng hợp lại quá trình thực hiện đồ án



Tiến trình thực hiện đồ án

# STT	Aa Tasks	☰ Người thực hiện	📅 Deadline	⚙ Status	☑ Đã merge vào main?	☰ Nội dung
1	<u>Tìm nguồn dữ liệu</u>	Tùng Tài Bình Lưu	November 6, 2022	● Done	☑	Causes of death
2	<u>Thu thập dữ liệu</u>	Tùng Bình	November 20, 2022	● Done	☑	Dùng selenium
3	<u>Tiền xử lý, khám phá dữ liệu</u>	Lưu Tài	December 1, 2022	● Done	☑	Thực hiện như các bước trong file hướng dẫn
4	<u>Đặt câu hỏi 1</u>	Tài	December 5, 2022	● Done	☑	Top 3 quốc gia có số lượng tử vong vì Covid-19 nhiều nhất?
5	<u>Đặt câu hỏi 2, 3</u>	Tùng	December 5, 2022	● Done	☑	- Số ca tử vong do các bệnh truyền nhiễm ở các khu vực? - Nhóm tuổi có tỉ lệ mất do tự gây thương tích cao nhất?
6	<u>Đặt câu hỏi 4</u>	Bình	December 5, 2022	● Done	☑	Diễn biến của các nguyên nhân gây tử vong theo thời gian
7	<u>Đặt câu hỏi 5, 6</u>	Lưu	December 5, 2022	● Done	☑	- Mối quan hệ giữa một số nguyên nhân tử vong - So sánh phân phối tuổi tử vong
8	<u>Mô hình hóa 1</u>	Bình Tài	December 13, 2022	● Done	☑	Hồi quy tuyến tính
9	<u>Mô hình hóa 2</u>	Tùng Lưu	December 13, 2022	● Done	☑	Hồi quy đa thức
10	<u>Làm slide</u>	Tài Tùng Bình Lưu	December 14, 2022	● Done	☑	
11	<u>Viết báo cáo</u>	Tùng Tài Bình Lưu	December 14, 2022	● Done	☑	

Tổng hợp lại quá trình thực hiện đồ án

STT	Họ và tên	Khó khăn mắc phải	Bài học rút ra
1	Bùi Thanh Tùng	Khó khăn trong việc xây dựng mô hình vì tập dữ liệu chúng em thu thập chỉ có 1 trường là numeric.	Nên thu thập đa dạng dữ liệu trước khi giải quyết một bài toán.
2	Đỗ Tấn Tài	Vì tập dữ liệu chỉ nói về một trường thông tin là số ca tử vong nên việc đặt câu hỏi bị khó khăn, dễ bị trùng lặp.	Lựa chọn dữ liệu có tính đa dạng hơn.
3	Trần Khắc Bình	Dữ liệu quá nhiều dòng nên việc thu thập tốn rất nhiều thời gian. Mặc khác, dữ liệu có ít dòng nên không có nhiều thứ để khác thác cho phần đặt câu hỏi và mô hình hóa	Lần sau sẽ thu thập dữ liệu có số dòng và cột phù hợp hơn.
4	Phan Phong Lưu	Khó khăn trong việc tìm câu hỏi vì một số quốc gia không có số liệu.	Tìm cách điền đầy đủ dữ liệu, thu thập dữ liệu ít bị thiếu hơn.

Tổng hợp lại quá trình thực hiện đồ án

Nếu có nhiều thời gian hơn, nhóm em sẽ:

- Cùng nhau bàn bạc, thảo luận nhiều hơn để tìm hiểu tường minh thông tin dữ liệu mà nhóm đã thu thập. Từ đó tối đa hóa việc xử lý các vấn đề phát sinh trong việc tính toán và tiền xử lý dữ liệu. Loại bỏ các dữ liệu gây nhiễu hoặc các dữ liệu không có giá trị về mặt ý nghĩa.
- Đưa ra các câu hỏi mang tính sâu sắc hơn nhằm phân tích hiệu quả giá trị mà dữ liệu đã cung cấp.
- Hoàn thiện file notebook, tuân thủ clear-coding, viết code ngắn gọn và đơn giản nhất có thể, chú thích tường minh cho từng bước xử lý.

NHÓM 16

**THANK YOU
FOR LISTENING**

