ĐỒ ÁN THỰC HÀNH

NHÓM 6

LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

Chủ đề: 120 years of Olympic history: athletes and results

GVHD: Thầy Lê Đại Chí và Thầy Lê Nhựt Nam

DANH SÁCH THÀNH VIÊN

Nhóm 6

HỌ VÀ TÊN	MSSV
Bùi Thanh Tùng	20120398
Đỗ Tấn Tài	20120408
Trần Khắc Bình	20120437
Lê Văn Hùng	20120485

Nội dung chính

1. Thu thập dữ liệu
2. Khám phá dữ liệu
3. Đưa ra các câu hỏi có ý nghĩa cần trả lời
4. Tiền xử lý và phân tích dữ liệu để trả lời cho từng câu hỏi
5. Tổng hợp lại quá trình thực hiện đồ án

1 Thu thập dữ liệu

Đây là bộ dữ liệu lịch sử về Thế vận hội Olympic hiện đại, bao gồm tất cả các Thế vận hội từ Athens 1896 đến Rio 2016

Athens 1896

Rio 2016

- Thế vận hội Mùa đông và Mùa hè được tổ chức trong cùng một năm cho đến năm 1992.
- Sau đó, họ xen kẽ chúng (2 năm/lần): Thế vận hội Mùa đông diễn ra năm 1994, sau đó là Mùa hè năm 1996, rồi Mùa đông năm 1998, …
- Một sai lầm phổ biến mà mọi người mắc phải khi phân tích dữ liệu này là cho rằng Thế vận hội Mùa hè và Mùa đông luôn luôn được sắp xếp xen kẽ.

- **Link dataset**: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results
- **License**: CC0: Public Domain (không có bản quyền, cho phép mọi người có thể sao chép, sửa đổi, phân phối hay ngay cả sử dụng với mục đích thương mại mà không cần phải xin phép).

Cách thu thập dữ liệu

- Tác giả đã sử dụng R để cào dữ liệu từ website www.sports-reference.com vào tháng 5/2018.
- Dữ liệu của website là kết quả 1 cuộc nghiên cứu quy mô lớn của 1 nhóm những người đam mê lịch sử Thế vận hội và những người tự xưng mình là 'nhà thống kê'.

2 Khám phá dữ liệu

a. Thông tin dữ liệu

Dữ liệu thu thập gồm 2 file:

- athlete_events.csv: cho biết những thông tin về các vận động viên đã tham gia Olympics.
- noc_regions.csv: cho biết mỗi mã NOC (National Olympic Committees - Đoàn đại biểu Uỷ ban Olympic quốc gia) sẽ tương ứng với vùng lãnh thổ nào và những ghi chú (nếu có).

Để thuận lợi cho việc thao tác với dữ liệu, ta sẽ tiến hành gộp 2 dataframe thông qua cột chung là NOC và lưu vào biến ethlete_df.

b. Đặc điểm dữ liệu

Số dòng và số cột và lần lượt lưu vào 2 biến là num_rows và num_cols

```
num_rows, num_cols = ethlete_df.shape
print("Num rows: ", num_rows)
print("Num cols: ", num_cols)
```

✓ 0.6s

```
Num rows:  271116
Num cols:  17
```

Ý nghĩa mỗi dòng: Theo mô tả dữ liệu của trang web www.sports-reference.com và theo quan sát sơ bộ về dữ liệu thì một dòng cho biết thông tin của một vận động viên tham gia Thế vận hội Mùa đông và Mùa hè.

b. Đặc điểm dữ liệu

Khi kiểm tra dữ liệu, ta nhận thấy có dòng bị trùng lặp, do đó ta cần loại bỏ những dòng bị trùng lặp đó đi.

```
have_duplicated_rows = ethlete_df.duplicated().sum() > 0
have_duplicated_rows
```

✓ 0.3s

```
True
```

```
ethlete_df.drop_duplicates(inplace = True)
ethlete_df.duplicated().sum()
```

b. Đặc điểm dữ liệu

Dưới đây là phần mô tả của trang web www.sports-reference.com về các cột trong file "ethlete_event.csv":

- **ID**: Chỉ số duy nhất cho mỗi vận động viên.
- **Name**: Tên của vận động viên.
- **Sex**: Giới tính của vận động viên.
- **Age**: Tuổi của vận động viên.
- **Weight**: Cân nặng của vận động viên.
- **Team**: Tên đội của vận động viên.
- **NOC**: Ủy ban Olympic quốc gia.
- **Games**: Tên của Thế vận hội.
- **Year**: Năm tổ chức Thế vận hội.
- **Season**: Mùa tổ chức Thế vận hội.
- **City**: Thành phố tổ chức Thế vận hội.
- **Sport**: Môn thi đấu của vận động viên.
- **Event** Nội dung thi đấu vận động viên.
- **Medal**: Huy chương vận động viên đạt được.
- **region**: Quốc tịch của vận động viên.
- **notes**: Ghi chú.

c. Xử lý dữ liệu

- Ta tính tỉ lệ phần trăm các giá trị bị thiếu trong các cột để chọn ra những cột có tỉ lệ giá trị bị thiếu lớn và loại bỏ.
- Ta thấy cột **notes** có tỉ lệ giá trị thiếu rất lớn và cột này cũng thật sự không cần thiết trong việc phân tích dữ liệu nên ta sẽ bỏ nó.

```
'Medal': 0.8481941309255079,
'region': 0.0013647294884846339,
'notes': 0.976320091457531}
```

```
ethlete_df.drop('notes', axis = 1, inplace = True)
num_cols = ethlete_df.shape[1]
print("Num cols: ", num_cols)
ethlete_df.columns
```

c. Xử lý dữ liệu

Dưới đây là kiểu dữ liệu của mỗi cột trong ethlete_df :

```
ethlete_df.dtypes
✓ 0.8s

ID          int64
Name       object
Sex        object
Age       float64
Height    float64
Weight    float64
Team       object
NOC        object
Games      object
Year        int64
Season     object
City       object
Sport      object
Event      object
Medal      object
region     object

dtype: object
```

Có vẻ như các cột đều có kiểu dữ liệu phù hợp, không cần phải xử lý.

Sự phân bố giá trị các cột **numeric**:

- Hiện tại, ta đang có 5 cột có vẻ thuộc nhóm numeric là: "ID", "Age", "Height", "Weight", "Year". Tuy nhiên, cột "ID" thật ra lại là dạng **categorical** (vì ID là mã được đánh dưới dạng số). Như vậy, chỉ có 4 cột numeric là "Age", "Height","Year" và "Weight".
- Với mỗi cột **numeric** ta sẽ tính tỉ lệ % giá trị thiếu (từ 0 đến 100), min, max.

	Age	Height	Year	Weight
missing_ratio	3.5	21.8	0.0	22.8
min	10.0	127.0	1896.0	25.0
max	97.0	226.0	2016.0	214.0

- Nhìn vào độ tuổi lớn nhất và nhỏ nhất, ta nhận thấy có một vài vận động viên rất 'đặc biệt'.
- Có những vận động viên có thể đạt đến chiều cao và cân nặng kỷ lục !
- Nhìn vào min, max của Year ta có thể thấy dữ liệu được thu thập từ năm 1896 đến 2016, đúng như mô tả ban đầu.

Sự phân bố giá trị các cột **categorical**:

- Có 12 cột **categorical** là: "ID", "Name", "Sex", "Team", "NOC", "Games", "Season", "City", "Sport", "Event", "Medal", "region".
- Với mỗi cột **categorical**, ta tính tỉ lệ % giá trị thiếu (từ 0 đến 100), số lượng giá trị khác nhau (không xét giá trị thiếu), list các giá trị khác nhau (không xét giá trị thiếu).

	ID	Name	Sex	Team	NOC	Games	Season	City	Sport	Event	Medal	region
missing_ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	85.3	0.1
num_diff_vals	135571	134732	2	1184	230	51	2	42	66	765	3	205
diff_vals	[106296, 115354, 119591, 129196, 89187, 119590...	[Heikki Ilmari Savolainen, Joseph "Josy" Stoff...	[M, F]	[United States, France, Great Britain, Italy, ...	[USA, FRA, GBR, ITA, GER, CAN, JPN, SWE, AUS, ...	[2000 Summer, 1996 Summer, 2016 Summer, 2008 S...	[Summer, Winter]	[London, Athina, Sydney, Atlanta, Rio de Janei...	[Athletics, Gymnastics, Swimming, Shooting, Cy...	[Football Men's Football, Ice Hockey Men's Ice...	[Gold, Bronze, Silver]	[USA, Germany, France, UK, Russia, Italy, Cana...

➔Có vẻ như không có gì bất thường.

3

Đưa ra các câu hỏi có ý nghĩa cần trả lời

a. Các câu hỏi có ý nghĩa

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:

- Top 5 vận động viên tham gia nhiều năm Thế vận hội nhất?
- Top 5 vận động viên đạt nhiều huy chương nhất?
- Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?
- Thành tích của Thể thao Việt Nam trong lịch sử tham dự Olympic?
- Trong một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?
- Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?
- Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội?

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 1. Top 5 vận động viên tham gia nhiều năm Thế vận hội nhất?

- **Ý nghĩa**: với câu hỏi trên, ta biết được những vận động viên xuất sắc nhất được quốc gia mình nhiều năm cử đi đại diện tham gia Thế vận hội. Có thể, ban tổ chức sẽ có một giải thưởng là "Tuyên dương những vận động viên có số lần tham gia nhiều nhất lịch sử" và họ sẽ được nhiều người đời sau biết đến.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 2. Top 5 vận động viên đạt nhiều huy chương nhất?

- **Ý nghĩa**: với câu hỏi trên, ta biết được những vận động viên xuất sắc nhất đã giành được nhiều huy chương về cho quốc gia mình. Đó là những vận động viên đã làm rạng danh đất nước, xứng đáng trở thành huyền thoại của đất nước đó nói riêng và của các kỳ Olympics nói chung.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

- **Ý nghĩa**: Với câu hỏi trên, ta có thể biết được những quốc gia có thành tích tốt nhất cho đến năm 2016. Những quốc gia đó đã cho thấy khả năng thể chất tuyệt vời, khẳng định được sức mạnh và vị thế thể thao của mình trên sân chơi quốc tế.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 4. Thành tích của Thể thao Việt Nam trong lịch sử tham dự Olympic?

- **Ý nghĩa**: Với câu hỏi trên, ta biết được thành tích của thể thao Việt Nam ở sân chơi Olympic để có thể tự hào và từ đó phấn đấu để đạt được nhiều huy chương hơn, xứng danh đất nước Con Rồng cháu Tiên.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 5. Trong một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?

Ý nghĩa: Với câu hỏi trên, ta biết được kỷ lục về số nội dung thi và số huy chương đạt được của một vận động viên trong một thế vận hội. Những kỷ lục trên nhằm vinh danh những vận động viên đã cống hiến hết mình cho thể thao, là tấm gương sáng để các thế hệ vận động viên sau này noi theo và cố gắng để phá vỡ kỷ lục.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

Ý nghĩa: Lợi thế chủ nhà là khi một quốc gia đăng cai có nhiều điều kiện thuận lợi hơn so với các nước khác (do nhiều yếu tố như cổ động viên, sự thích nghi khí hậu, không cần di chuyển xa, luật lệ nước chủ nhà...). Với lợi thế đó, ta sẽ xem liệu nước chủ nhà ở các kì Olympic có tận dụng được hay không.

b. Ý nghĩa các câu hỏi

Mỗi câu hỏi được đặt ra đều có các ý nghĩa giúp ta hiểu rõ hơn về dữ liệu đã được thu thập:

Câu 7. Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội?

Ý nghĩa: Với câu hỏi trên, nếu trả lời được sẽ giúp ích cho việc dự đoán số huy chương đạt được của mỗi quốc gia trước mỗi kỳ thế vận hội.

4 Tiền xử lý và phân tích dữ liệu để trả lời cho từng câu hỏi

Tiền xử lý

Để dễ dàng hơn trong việc tính số lượng huy chương, ta sẽ tạo thêm cột isWon để lưu giá trị xem vận động viên có giành được huy chương hay không, nếu có thì isWon sẽ có giá trị là 1, ngược lại là 0.

```
ethlete_df['isWon'] = np.where(ethlete_df['Medal'].isna(), 0, 1)
ethlete_df['isWon'] = ethlete_df['isWon'].astype(bool)
```

1. Top 5 vận động viên tham gia nhiều năm Thế vận hội nhất?

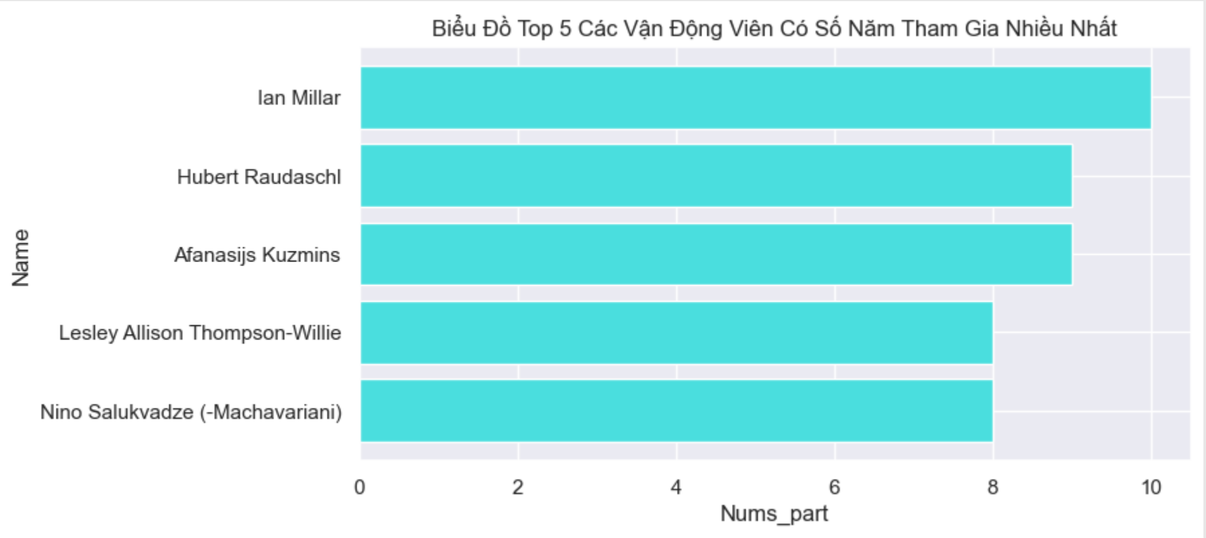
Để trả lời cho câu hỏi này, ta sẽ làm như sau:

Tính số lượng năm Thế vận hội mà mỗi vận động viên tham gia, chọn ra top 5 người có số lượng tham gia nhiều nhất. Ta lưu kết quả vào series num_year_participate, trong đó index là tên vận động viên.

	Name	0
50684	Ian Millar	10
50177	Hubert Raudaschl	9
1465	Afanasijs Kuzmins	9
74851	Lesley Allison Thompson-Willie	8
94058	Nino Salukvadze (-Machavariani)	8

1. Top 5 vận động viên tham gia nhiều năm Thế vận hội nhất?

Từ kết quả ở trên, ta vẽ đồ thị dạng cột, trong đó trục hoành là tên và trục tung là số lượng năm tham gia.Ta đặt tên trục hoành là "Year" và tên trục tung là Nums_part.



2. Top 5 vận động viên đạt nhiều huy chương nhất?

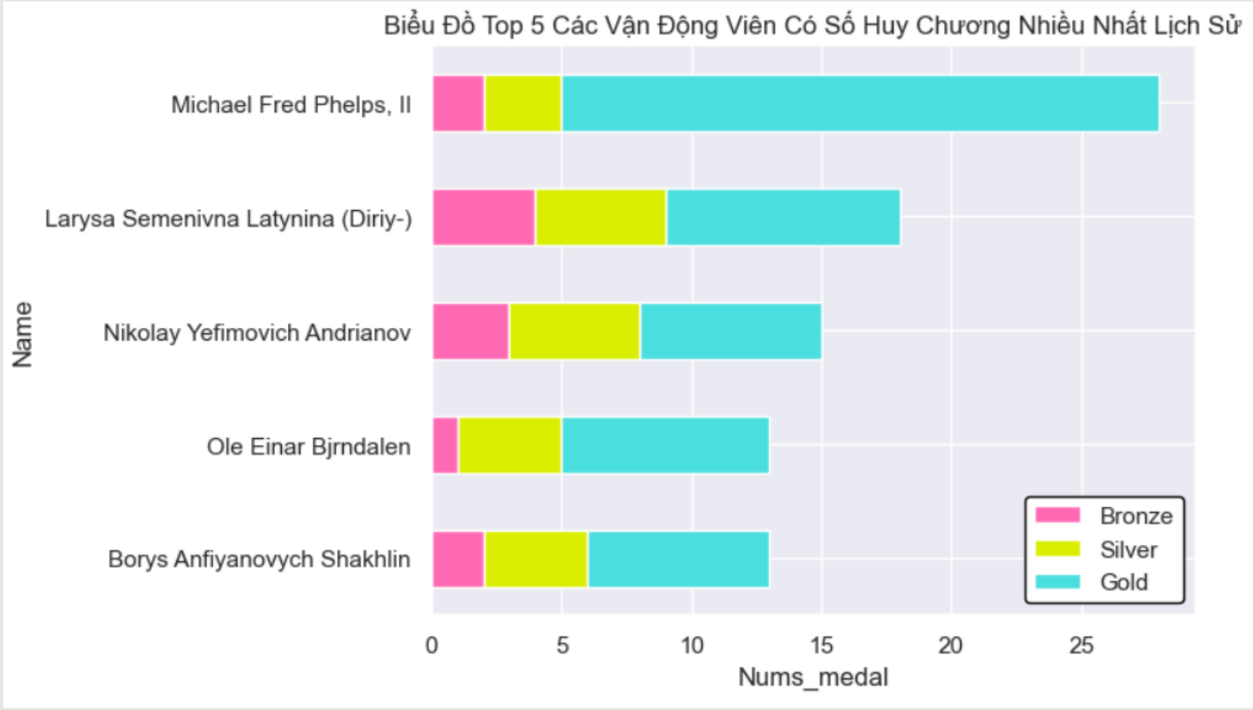
Để trả lời cho câu hỏi này, ta sẽ làm như sau:

Tính số lượng huy chương theo từng loại Gold, Silver, Bronze của các vận động viên. Chọn ra top 5 người có tổng số huy chương nhiều nhất. Ta lưu kết quả vào series num_medal_reward.

Name			
Michael Fred Phelps, II	2	3	23
Larysa Semenivna Latynina (Diriy-)	4	5	9
Nikolay Yefimovich Andrianov	3	5	7
Ole Einar Bjrndalen	1	4	8
Borys Anfiyanovych Shakhlin	2	4	7

2. Top 5 vận động viên đạt nhiều huy chương nhất?

Từ kết quả ở trên, ta vẽ group stack bar chart, trong đó trục hoành là tên và trục tung là số lượng huy chương. Ta đặt tên trục hoành là Name và tên trục tung là Nums_medal.



3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

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

- Ta sẽ tìm top 3 quốc gia có tổng số huy chương cao nhất và biểu diễn số lượng huy chương của các quốc gia này giành được qua các năm (Olympic mùa hè).

- Tuy nhiên, một vấn đề cần xử lý là đối với những môn thể thao đồng đội, tuy huy chương được trao cho từng thành viên trong đội, nhưng khi tính tổng số huy chương của quốc gia ta chỉ tính 1 huy chương cho một nội dung thi đấu.

3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

- Đầu tiên ta sẽ tìm tên những môn thể thao đồng đội, nó sẽ là Event mà trong một năm có nhiều bộ huy chương được trao.

```
array(["Polo Men's Polo", "Rowing Men's Double Sculls",
       "Football Men's Football", "Basketball Men's Basketball",
       "Cycling Men's Madison", 'Sailing Mixed Multihull',
       "Hockey Men's Hockey", "Rugby Men's Rugby",
       "Swimming Men's 4 x 200 metres Freestyle Relay",
       "Cycling Men's Tandem Sprint, 2,000 metres",
       "Athletics Women's 4 x 100 metres Relay",
```

Một số môn thể thao minh họa trong số kết quả tìm được

3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

- Trong quá trình tìm, ta nhận thấy có 1 vài môn thể thao không phải là môn thể thao đồng đội nhưng vẫn được thêm vào . Nguyên nhân là do có tồn tại trường hợp trao nhiều hơn 1 huy chương vàng khi các vận động viên có cùng điểm số/ thời gian. Ta sẽ loại bỏ những môn thể thao đó đi.

```
remove_sports = ["Gymnastics Women's Balance Beam", "Gymnastics Men's Horizontal
                 "Swimming Women's 100 metres Freestyle", "Swimming Men's 50 metr
team_sports = list(set(team_sports) - set(remove_sports))
```

3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

- Ta tạo thêm cột isTeamSport để ghi nhận việc Event đó có phải là môn thể thao đồng đội hay không.

```
ethlete_df['isTeamSport'] = np.where(ethlete_df['Event']
                                     .map(lambda x: x in team_sports), 1, 0)
ethlete_df['isTeamSport'] = ethlete_df['isTeamSport'].astype(bool)
```

3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

Tiếp theo, ta sẽ tính số huy chương mỗi quốc gia cho các môn thể thao thuộc từng loại: cá nhân và đồng đội. Sau khi tính, ta gộp 2 dataframe lại với nhau để ra số huy chương thực tế của mỗi quốc gia.

Medal	Gold	Silver	Bronze	Total
region				
USA	820	728	649	2197
Russia	494	432	442	1368
Germany	332	385	413	1130
UK	233	264	261	758
France	200	230	247	677
...	...	...	...	...
Kosovo	1	0	0	1
Montenegro	0	1	0	1
Macedonia	0	0	1	1
Monaco	0	0	1	1
Gabon	0	1	0	1

Bây giờ ta sẽ tìm top 3 quốc gia đạt được nhiều huy chương nhất các kỳ Olympics mùa hè và phân tích 1 chút về thành tích của 3 quốc gia đó qua các giai đoạn :

```
top3_countries = list(Total_Medal.head(3).index)
top3_countries
```

```
['USA', 'Russia', 'Germany']
```

Không quá bất ngờ khi top 3 dẫn đầu số huy chương Olympics đều là những cường quốc hàng đầu thế giới (Mỹ, Nga, Đức). Sau đây là thống kê số huy chương mà 3 quốc gia đó đã đạt được.

3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

region	USA	Russia	Germany
Year			
1896	20.0	NaN	32.0
1900	63.0	0.0	45.0
1904	394.0	NaN	16.0
1906	24.0	NaN	30.0
1908	65.0	3.0	21.0
1912	107.0	14.0	53.0
1920	194.0	NaN	NaN
1924	182.0	0.0	NaN
1928	88.0	NaN	77.0
1932	189.0	NaN	43.0
1936	96.0	NaN	224.0
1948	152.0	NaN	NaN
1952	134.0	117.0	40.0
1956	123.0	169.0	52.0
1960	125.0	169.0	89.0
1964	169.0	174.0	116.0
1968	166.0	192.0	103.0
1972	171.0	214.0	253.0
1976	164.0	286.0	273.0
1980	NaN	442.0	264.0
1984	352.0	NaN	158.0
1988	207.0	300.0	296.0
1992	224.0	220.0	198.0
1996	259.0	115.0	124.0
2000	242.0	187.0	118.0
2004	263.0	189.0	149.0
2008	317.0	142.0	99.0
2012	248.0	140.0	94.0
2016	264.0	115.0	159.0

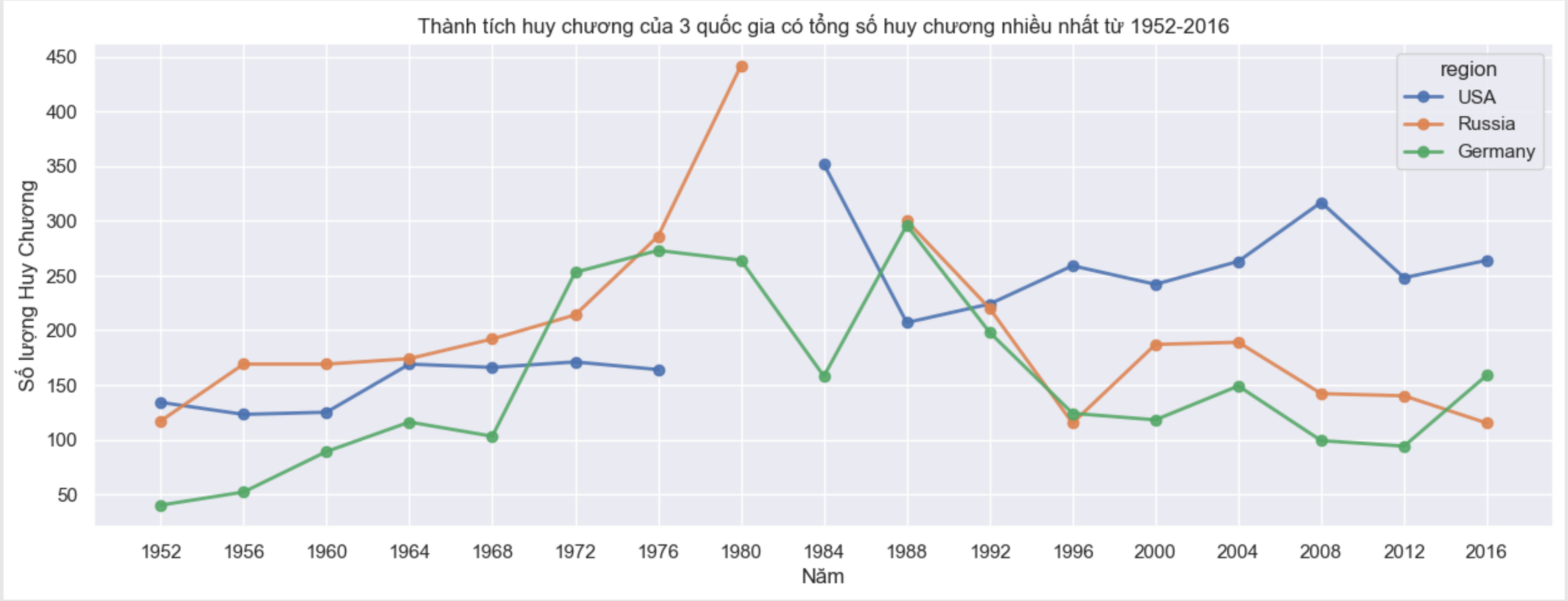
Nhìn vào bảng kết quả, ta thấy dữ liệu của nước Nga bị thiếu rất nhiều trong khoảng thời gian từ kỳ Thế vận hội năm 1896 đến 1948 nhưng đây không phải là lỗi thiếu dữ liệu ! Và đây là lí do:

- Nga lần đầu tham dự Thế vận hội vào năm năm 1900 cũng như gửi vận động viên (VĐV) tới các kỳ Mùa hè 1908 và 1912. Sau Cách mạng Nga (1917) và tiếp đó là sự thành lập Liên bang Xô Viết năm 1922, sự góp mặt của các VĐV Nga bị gián đoạn cho tới 1952.

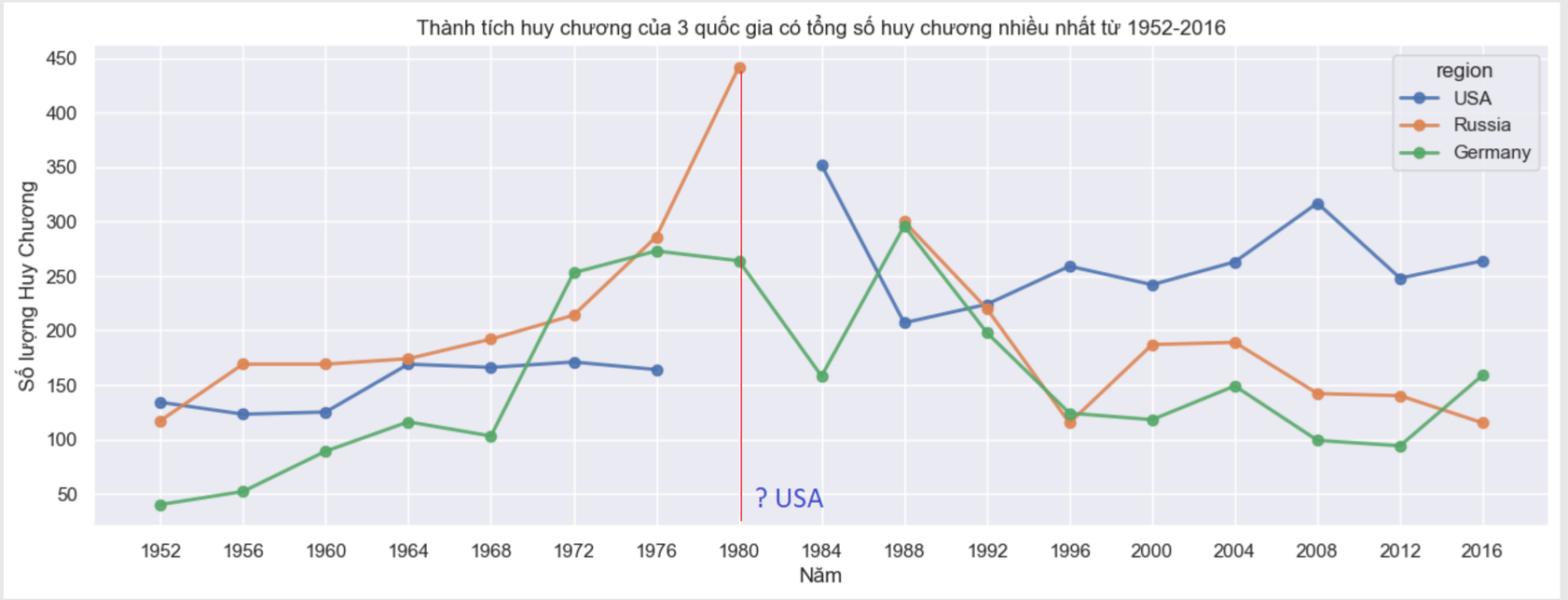
region	USA	Russia	Germany
Year			
1896	20.0	NaN	32.0
1900	63.0	0.0	45.0
1904	394.0	NaN	16.0
1906	24.0	NaN	30.0
1908	65.0	3.0	21.0
1912	107.0	14.0	53.0
1920	194.0	NaN	NaN
1924	182.0	0.0	NaN
1928	88.0	NaN	77.0
1932	189.0	NaN	43.0
1936	96.0	NaN	224.0
1948	152.0	NaN	NaN

➔Vì vậy ta sẽ chỉ phân tích dữ liệu từ kỳ Thế vận hội năm 1952 trở về sau.

3. Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè?

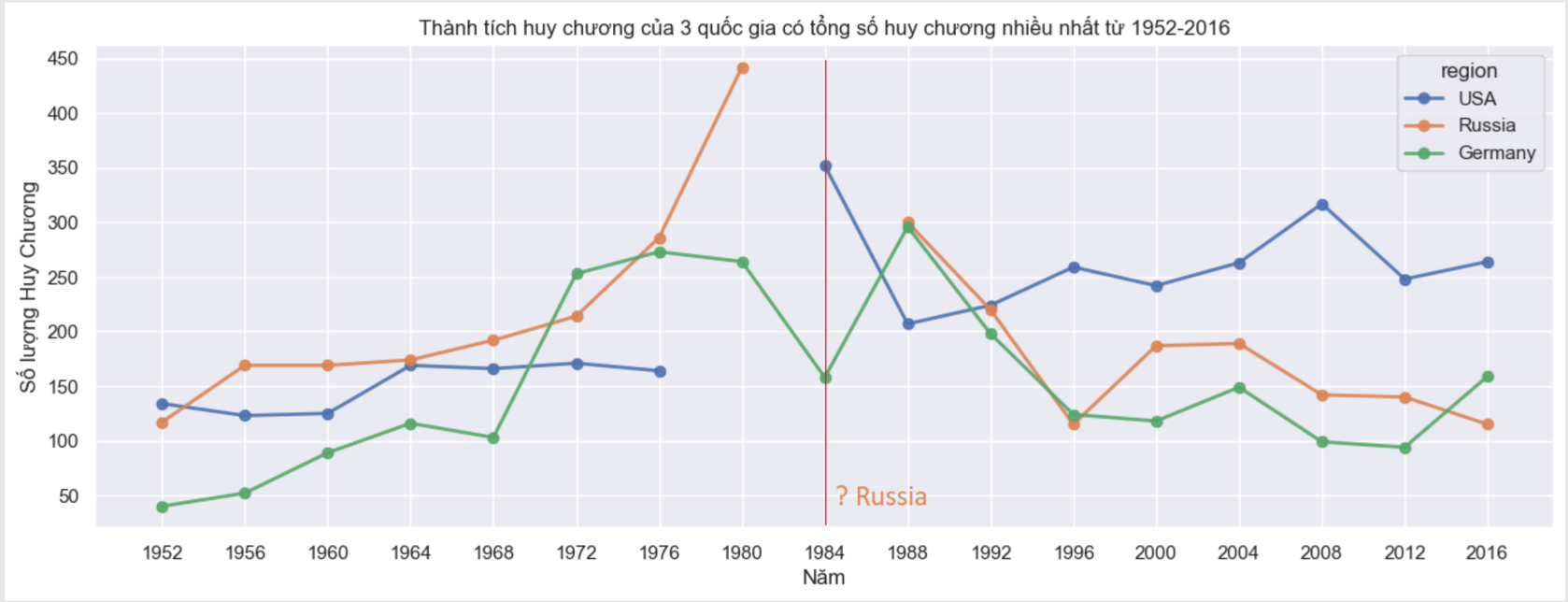


Biểu đồ thành tích huy chương của top 3 quốc gia từ 1952-2016



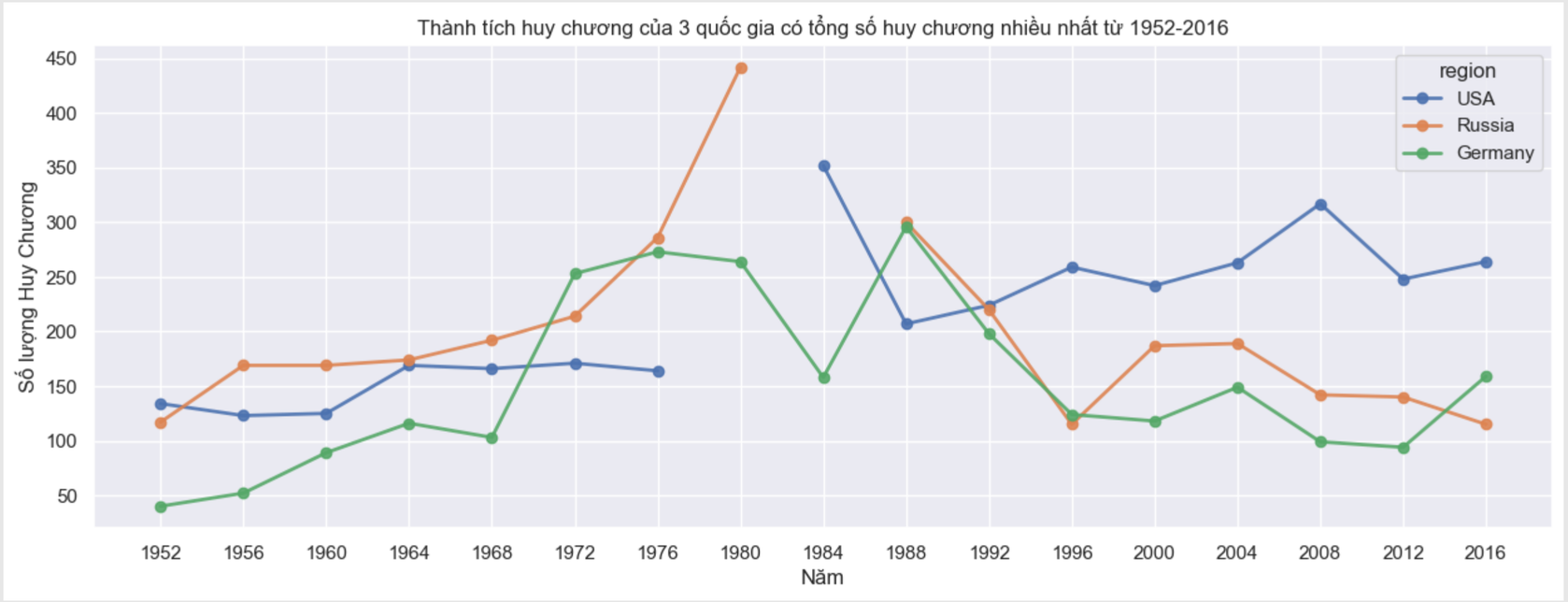
Điểm giá trị bị thiếu của Mỹ vào năm 1980 không phải là lỗi thiếu dữ liệu!

- Năm 1980, Hoa Kỳ kêu gọi tẩy chay Thế vận hội Olympic mùa hè ở Moscow để phản đối cuộc xâm lược Afghanistan của Liên Xô vào cuối năm 1979, có 64 quốc gia khác hưởng ứng cuộc tẩy chay này. Đây là lần đầu tiên và duy nhất mà Mỹ tẩy chay Olympics.
- Thế vận hội 1980 là Kỳ Thế vận hội đầu tiên mà Việt Nam tham dự với tư cách là một quốc gia thống nhất.



Điểm giá trị bị thiếu của Nga vào năm 1984 cũng không có lỗi!

- 08/05/1984, Liên Xô tuyên bố tẩy chay Olympics 1984. Việc tẩy chay Thế vận hội Mùa hè 1984 ở Los Angeles diễn ra sau 4 năm kể từ cuộc tẩy chay Thế vận hội Mùa hè 1980 ở Moscow do Hoa Kỳ lãnh đạo.
- Nguyên nhân chính được các quan chức Moscow đưa ra là vận động viên Liên Xô có thể sẽ không được an toàn trước các cuộc biểu tình và tấn công thù địch do phía Mỹ hậu thuẫn. Sau khi Liên Xô đưa ra tuyên bố "tẩy chay" Olympic 1984, 13 nước có quan hệ thân thiết với quốc gia xã hội chủ nghĩa lớn nhất thế giới cũng đưa ra các thông báo tương tự và từ chối tham dự.



Với các mốc thời gian còn lại:

- Từ những năm 1952 đến 1976, số lượng huy chương giành được của các quốc gia Mỹ, Nga, Đức gần như tăng dần qua các năm.
- Từ những năm 1996 đến 2016, Mỹ luôn dẫn đầu về thổng số lượng huy chương trong một Kỳ thế vận hội, trong khi Nga thì có xu hướng giảm, Đức thì không ổn định.

4. Thành tích của thể thao Việt Nam trong lịch sử tham dự Olympic?

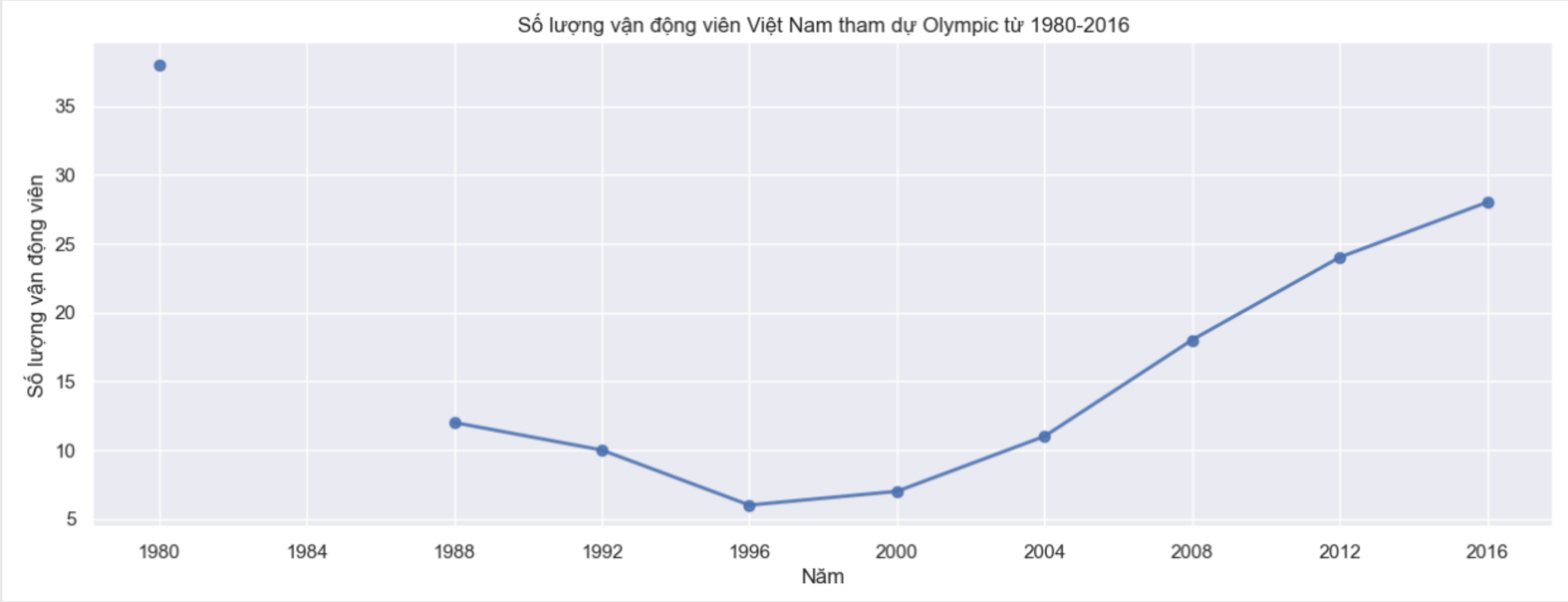
Để trả lời cho câu hỏi này, ta sẽ làm như sau:

- Đầu tiên, ta kiểm tra Việt Nam có từng đạt được huy chương môn đồng đội nào không. Sau khi nhận thấy không có, ta có thể thực hiện tính toán như bình thường.

4. Thành tích của thể thao Việt Nam trong lịch sử tham dự Olympic?

- Tiếp theo, ta tạo Dataframe chứa dữ liệu gồm số vận động viên Việt Nam tham dự, số huy chương đạt được qua các năm. Vì Việt Nam lần đầu tham dự Thế vận hội với tư cách là một quốc gia thống nhất là vào năm 1980, do đó ta sẽ chỉ phân tích dữ liệu từ năm 1980.

Year	Season	number of athletes	Gold	Silver	Bronze
1980	Summer	38.0	NaN	NaN	NaN
1984	Summer	NaN	NaN	NaN	NaN
1988	Summer	12.0	NaN	NaN	NaN
1992	Summer	10.0	NaN	NaN	NaN
1996	Summer	6.0	NaN	NaN	NaN
2000	Summer	7.0	NaN	1.0	NaN
2004	Summer	11.0	NaN	NaN	NaN
2008	Summer	18.0	NaN	1.0	NaN
2012	Summer	24.0	NaN	NaN	NaN
2016	Summer	28.0	1.0	1.0	NaN



Biểu đồ thống kê cho số lượng vận động viên Việt Nam từ 1980-2016

Nhận xét:

- Số lượng vận động viên Việt Nam giảm từ 1988 đến 1996, sau đó lại tăng liên tục từ năm 1996 đến 2016.
- Việt Nam là quốc gia không có mùa đông băng tuyết, vận động viên sẽ không có đủ điều kiện luyện tập, thường phải luyện tập ở nước ngoài. Vì những điều kiện khó khăn đó cho nên từ 1952 đến 2016 Việt Nam chưa từng tham dự Olympic mùa Đông.
- Vì lý do kinh tế và chính trị, Việt Nam không tham dự kỳ Olympic mùa Hè 1984.

Ta tiếp tục quan sát biểu đồ về số huy chương mà các vận động viên Việt Nam đã đạt được:

Thông tin về các vận động viên Việt Nam đã giành được huy chương:

Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport
Hong Anh Tun	M	23.0	152.0	56.0	Vietnam	VIE	2008 Summer	2008	Summer	Beijing	Weightlifting
Hong Xun Vinh	M	41.0	175.0	75.0	Vietnam	VIE	2016 Summer	2016	Summer	Rio de Janeiro	Shooting
Hong Xun Vinh	M	41.0	175.0	75.0	Vietnam	VIE	2016 Summer	2016	Summer	Rio de Janeiro	Shooting
Trn Hiu Ngn	F	26.0	157.0	47.0	Vietnam	VIE	2000 Summer	2000	Summer	Sydney	Taekwondo

Đánh giá dữ liệu:

Gần 4 thập kỷ trôi qua từ lần đầu tiên góp mặt tại các kỳ Olympic, thể thao Việt Nam mới có 4 VĐV đoạt huy chương, trong đó xạ thủ bắn súng Hoàng Xuân Vinh là người duy nhất sở hữu HCV. Sau đây là lịch sử các kỳ Olympic của Đoàn thể thao Việt Nam:

- Olympic Moscow (Nga) 1980: Đoàn TTVN tham dự với 38 VĐV, không giành huy chương.
- Olympic Seoul (Hàn Quốc) 1988: Tham dự với 12 VĐV, không giành huy chương.
- Olympic Barcelona (Tây Ban Nha) 1992: Tham dự với 10 VĐV, không giành huy chương.

- Olympic Atlanta (Mỹ) 1996: Tham dự với 6 VĐV, không giành huy chương.
- Olympic Sydney (Australia) 2000: Tham dự với 7 VĐV, lần đầu tiên có tên trên bảng tổng sắp huy chương, với tấm HCB của nữ võ sĩ taekwondo Trần Hiếu Ngân ở hạng cân 57kg.
- Olympic Athens (Hy Lạp) 2004: Tham dự với 11 VĐV, không giành huy chương.
- Olympic Bắc Kinh (Trung Quốc) 2008: Tham dự với 18 VĐV, giành 1 HCB nhờ công của lực sỹ cử tạ Hoàng Anh Tuấn (hạng 56 kg nam).
- Olympic London (Anh) 2012: Tham dự với 24 VĐV, giành 1 HCĐ sau khi Trần Lê Quốc Toàn được đôn từ hạng tư nội dung 56kg lên vị trí thứ ba, thế chỗ VĐV của Azerbajan bị tước huy chương vì doping.
- Olympic Rio (Brazil) 2016: Tham dự với 28 VĐV, trong đó xạ thủ bắn súng Hoàng Xuân Vinh xuất sắc giành 1 HCV nội dung 10m súng ngắn hơi và 1 HCB nội dung 50m súng ngắn.

5. Trong cùng một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?

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

- Đầu tiên, tạo dict **games_id** chứa thông tin của các vận động viên theo từng thế vận hội.
  - **key** : tên thế vận hội.
  - **value**: danh sách thông tin của vận động viên tham gia thế vận hội đó, trong đó thông tin của một vận động viên gồm: 'ID', 'Name', 'Num of attend', 'Num of medal'

5. Trong cùng một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?


```
{'1896 Summer': [{'ID': 1724,
   'Name': 'Aristidis Akratopoulos',
   'Num of attend': 2,
   'Num of medal': 0},
  {'ID': 1725,
   'Name': 'Konstantinos "Kostas" Akratopoulos',
   'Num of attend': 2,
   'Num of medal': 0},
  {'ID': 4113,
   'Name': 'Anastasios Andreou',
   'Num of attend': 1,
   'Num of medal': 0},
  {'ID': 4116,
   'Name': 'Ioannis Andreou',
   'Num of attend': 1,
   'Num of medal': 1},
```

```
  {'ID': 4189,
   'Name': 'Nikolaos Andriakopoulos',
   'Num of attend': 1,
   'Num of medal': 1},
  {'ID': 4431,
   'Name': 'Georgios Anninos',
   'Num of attend': 1,
   'Num of medal': 0},
  {'ID': 4493,
...
  {'ID': 11194,
   'Name': 'Mohamed Ali Bhar',
   'Num of attend': 1,
   'Num of medal': 0},
  ...]}
```

Tiếp theo, ta tạo 2 dict **games_maxAttend** và **games_maxMedal** :

- **games_maxAttend**: **key** là tên của thế vận hội, **value** là số nội dung tham gia của vận động viên có số nội dung tham gia nhiều nhất thế vận hội năm đó.

```
{'1896 Summer': 12, '1900 Summer': 8, '1904 Summer': 10, '1906 Summer': 12, '1908
Summer': 7, '1912 Summer': 12, '1920 Summer': 15, '1924 Summer': 9, '1924 Winter':
6, '1928 Summer': 7, '1928 Winter': 4, '1932 Summer': 9, '1932 Winter': 4, '1936
Summer': 8, '1936 Winter': 5, '1948 Summer': 8, '1948 Winter': 6, '1952 Summer': 8,
'1952 Winter': 4, '1956 Summer': 8, '1956 Winter': 4, '1960 Summer': 8, '1960
Winter': 4, '1964 Summer': 8, '1964 Winter': 4, '1968 Summer': 8, '1968 Winter': 5,
'1972 Summer': 8, '1972 Winter': 4, '1976 Summer': 8, '1976 Winter': 5, '1980
Summer': 8, '1980 Winter': 5, '1984 Summer': 8, '1984 Winter': 5, '1988 Summer':
10, '1988 Winter': 5, '1992 Summer': 8, '1992 Winter': 5, '1994 Winter': 5, '1996
Summer': 8, '1998 Winter': 5, '2000 Summer': 8, '2002 Winter': 6, '2004 Summer': 8,
'2006 Winter': 5, '2008 Summer': 8, '2010 Winter': 6, '2012 Summer': 8, '2014
Winter': 6, '2016 Summer': 8}
```

Tiếp theo, ta tạo 2 dict **games_maxAttend** và **games_maxMedal** :

- **games_maxMedal**: **key** là tên của thế vận hội, **value** là số huy chương của vận động viên có số huy chương nhiều nhất thế vận hội năm đó.

```
{'1896 Summer': 6, '1900 Summer': 5, '1904 Summer': 6, '1906 Summer': 5, '1908
Summer': 3, '1912 Summer': 5, '1920 Summer': 7, '1924 Summer': 6, '1924 Winter': 5,
'1928 Summer': 4, '1928 Winter': 3, '1932 Summer': 4, '1932 Winter': 2, '1936
Summer': 6, '1936 Winter': 4, '1948 Summer': 5, '1948 Winter': 3, '1952 Summer': 7,
'1952 Winter': 3, '1956 Summer': 6, '1956 Winter': 4, '1960 Summer': 7, '1960
Winter': 3, '1964 Summer': 6, '1964 Winter': 4, '1968 Summer': 7, '1968 Winter': 3,
'1972 Summer': 7, '1972 Winter': 3, '1976 Summer': 7, '1976 Winter': 4, '1980
Summer': 8, '1980 Winter': 5, '1984 Summer': 6, '1984 Winter': 4, '1988 Summer': 7,
'1988 Winter': 3, '1992 Summer': 6, '1992 Winter': 5, '1994 Winter': 5, '1996
Summer': 6, '1998 Winter': 5, '2000 Summer': 6, '2002 Winter': 4, '2004 Summer': 8,
'2006 Winter': 5, '2008 Summer': 8, '2010 Winter': 5, '2012 Summer': 6, '2014
Winter': 5, '2016 Summer': 6}
```

5. Trong cùng một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?

Từ 2 dict đã tạo, ta lấy ra số nội dung nhiều nhất và số huy chương nhiều nhất của 1 vận động viên trong cùng một thế vận hội cùng với danh sách tên các thế vận hội đó:

```
['1920 Summer']
['1980 Summer', '2004 Summer', '2008 Summer']
```

```
num_most_attend:  15
num_most_medal:  8
```

5. Trong cùng một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?

Cuối cùng, ta lưu lại thông tin các vận động viên đã tìm vào dict và chuyển sang dạng bảng để trực quan hơn:

	ID	Name	Thế vận hội	Số nội dung tham gia	Số huy chương đạt được
0	68189	Willis Augustus Lee, Jr.	1920 Summer	15	7

VĐV tham gia nhiều nội dung thi nhất

5. Trong cùng một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất?

Cuối cùng, ta lưu lại thông tin các vận động viên đã tìm vào dict và chuyển sang dạng bảng để trực quan hơn:

	ID	Name	Thế vận hội	Số nội dung tham gia	Số huy chương đạt được
0	28790	Aleksandr Nikolayevich Dityatin	1980 Summer	8	8
1	94406	Michael Fred Phelps, II	2004 Summer	8	8
2	94406	Michael Fred Phelps, II	2008 Summer	8	8

VĐV giành được nhiều huy chương nhất

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

Để trả lời cho câu hỏi này, ta sẽ phân tích bằng cách so sánh số huy chương mà nước chủ nhà đạt được tại 3 thời điểm:

- Tại thế vận hội mà nước chủ nhà tham gia gần nhất trước khi đăng cai thế vận hội đang xét.
- Tại thế vận hội mà nước chủ nhà đăng cai.
- Tại thế vận hội mà nước chủ nhà tham gia gần nhất sau khi đăng cai thế vận hội đang xét.

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

	Year	City
0	1994	Lillehammer
1	1996	Atlanta
2	1998	Nagano
3	2000	Sydney
4	2002	Salt Lake City
5	2004	Athina
6	2006	Torino
7	2008	Beijing
8	2010	Vancouver
9	2012	London
10	2014	Sochi
11	2016	Rio de Janeiro

- Trước tiên ta sẽ tìm những thành phố đã đăng cai Olympic từ năm 1994:

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

	Year	City
0	1994	Lillehammer
1	1996	Atlanta
2	1998	Nagano
3	2000	Sydney
4	2002	Salt Lake City
5	2004	Athina
6	2006	Torino
7	2008	Beijing
8	2010	Vancouver
9	2012	London
10	2014	Sochi
11	2016	Rio de Janeiro

- Ngoài tên những thành phố nổi tiếng đã quá quen thuộc, ta thấy có những cái tên lạ như Athina hay Torino. Đây không phải những thành phố lạ lẫm đâu. Đó chính là 2 thành phố Athens ở Hy Lạp và Turin ở Ý (Athina là Athens phát âm trong tiếng Hy Lạp, còn Torino là phát âm của Turin trong tiếng Ý). Ta có thể sửa lại chúng theo tiếng Anh trước khi tiếp tục phân tích.

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

- Tiếp theo, ta sẽ dựa trên tên những thành phố này để đổi sang tên quốc gia tương ứng.

	Year	Host_Country
0	1994	Norway
1	1996	United States
2	1998	Japan
3	2000	Australia
4	2002	United States
5	2004	Greece
6	2006	Italy
7	2008	China
8	2010	Canada
9	2012	Great Britain
10	2014	Russia
11	2016	Brazil

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

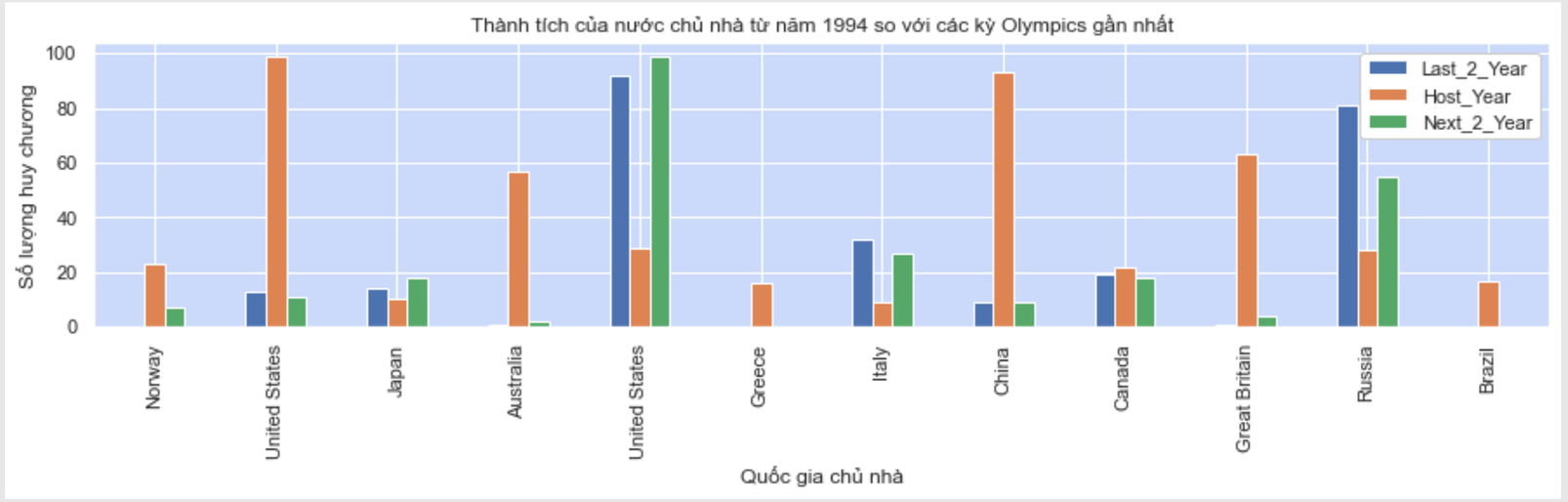
- Sau đó, ta sẽ tính số huy chương cá nhân và đồng đội rồi gộp lại để ra số huy chương mà các nước đã đạt được gần đây.

Team	Norway	United States	Japan	Australia	United States	Greece	Italy	China	Canada	Great Britain	Russia	Brazil
Year												
1994	23.0	13.0	5.0	1.0	13.0	0.0	17.0	3.0	12.0	2.0	19.0	0.0
1996	7.0	99.0	14.0	40.0	99.0	8.0	35.0	44.0	21.0	15.0	63.0	13.0
1998	22.0	11.0	10.0	1.0	11.0	0.0	9.0	8.0	14.0	1.0	14.0	0.0
2000	10.0	92.0	18.0	57.0	92.0	13.0	34.0	51.0	14.0	27.0	89.0	9.0
2002	24.0	29.0	2.0	2.0	29.0	0.0	12.0	7.0	16.0	2.0	11.0	0.0
2004	6.0	99.0	37.0	50.0	99.0	16.0	32.0	57.0	12.0	29.0	90.0	8.0
2006	19.0	23.0	1.0	2.0	23.0	0.0	9.0	9.0	23.0	1.0	19.0	0.0
2008	9.0	106.0	25.0	46.0	106.0	4.0	27.0	93.0	19.0	48.0	72.0	14.0
2010	23.0	34.0	5.0	3.0	34.0	0.0	5.0	9.0	22.0	1.0	13.0	0.0
2012	4.0	98.0	37.0	35.0	98.0	2.0	28.0	84.0	18.0	63.0	81.0	15.0
2014	26.0	23.0	8.0	3.0	23.0	0.0	8.0	9.0	23.0	4.0	28.0	0.0
2016	4.0	117.0	41.0	29.0	117.0	6.0	27.0	68.0	22.0	67.0	55.0	17.0

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

- Sau đó lấy ra số huy chương mà các nước chủ nhà đạt được trong năm đăng cai cũng như các kỳ Olympic 2 năm trước và sau đó

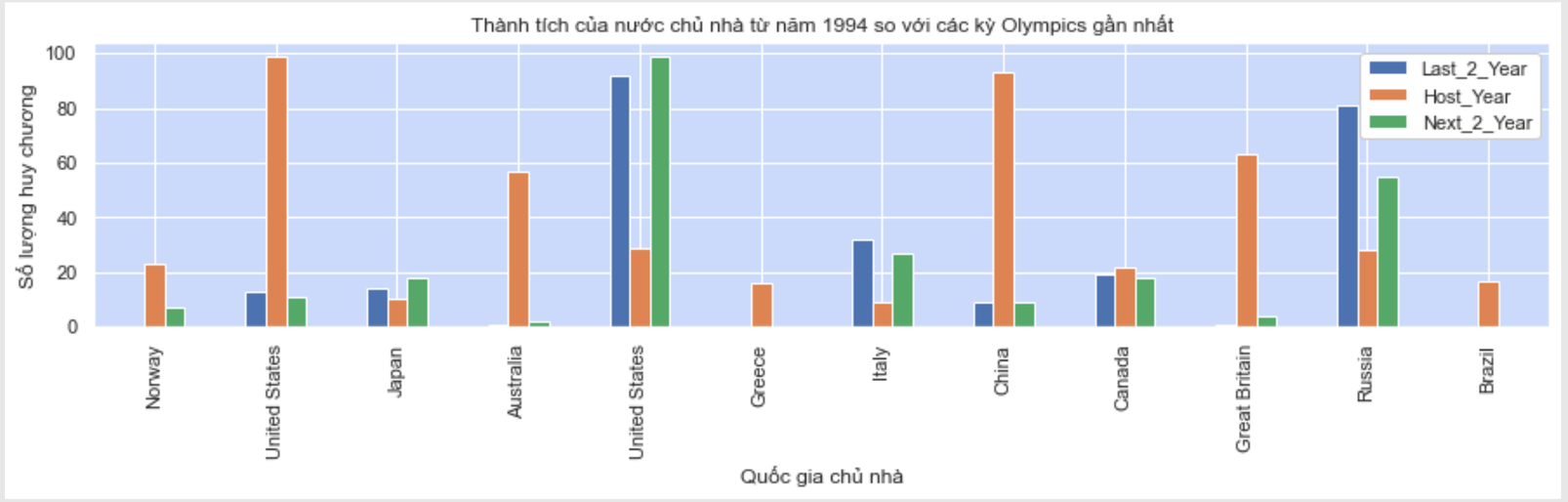
Team	Year	Last_2_Year	Host_Year	Next_2_Year
Norway	1994	NaN	23.0	7.0
United States	1996	13.0	99.0	11.0
Japan	1998	14.0	10.0	18.0
Australia	2000	1.0	57.0	2.0
United States	2002	92.0	29.0	99.0
Greece	2004	0.0	16.0	0.0
Italy	2006	32.0	9.0	27.0
China	2008	9.0	93.0	9.0
Canada	2010	19.0	22.0	18.0
Great Britain	2012	1.0	63.0	4.0
Russia	2014	81.0	28.0	55.0
Brazil	2016	0.0	17.0	NaN



Biểu đồ thống kê cho số huy chương

Nhận xét:

- Nhìn vào thống kê trên, ta có thể thấy được rằng đa số các nước chủ nhà đều tận dụng rất tốt lợi thế chủ nhà và có được số huy chương xấp xỉ, thậm chí còn vượt trội rất nhiều hơn so với 2 kỳ Olympic gần nhất (hơi đáng ngờ thật, không biết do tận dụng tốt thật hay có sự thiên vị gì ở đây không).
- Tuy vậy, vẫn có 1 vài trường hợp đặc biệt như Mỹ (Olympic 2002) (mặc dù Mỹ đã làm rất tốt ở kỳ chủ nhà Olympic 6 năm trước đó), Nga (Olympic 2014) và Ý (2006) có số huy chương thấp hơn rất nhiều.



Biểu đồ thống kê cho số huy chương

Nhận xét:

- Điều này cũng khá dễ hiểu vì những kỳ Olympic này đều diễn ra vào mùa đông. So với các kỳ Olympic mùa hè, các môn thể thao mùa đông thường khắc nghiệt hơn, đòi hỏi vận động viên·phải đã quen với môi trường khí hậu (không phù hợp với các nước nhiệt đới và cận nhiệt đới) cũng như khán giả rất khó để theo dõi, quan sát. Bên cạnh đó, số tiền mà nước chủ nhà bỏ ra để đầu tư cho các môn thể thao mùa đông là hàng triệu đô (gấp nhiều lần so với mùa hè). Vì vậy, số nước có thể tham gia cũng như số môn thể thao và số huy chương sẽ hạn chế đi đáng kể. Đó cũng là lý do tại sao người ta lại ưa chuộng và quan tâm đến các Thế vận hội mùa hè hơn rất nhiều so với các kỳ Olympic mùa đông.

6. Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai tổ chức liệu có lợi thế chủ nhà hay không ?

- Để kiểm tra điều này, ta có thể xem qua thống kê số môn thể thao mùa đông so với mùa hè cũng như số vận động viên tham gia:

```
Summer Olympics:
Sports:  52
Athletes:  116776
Winter Olympics:
Sports:  17
Athletes:  18958
```

Thật là 1 sự áp đảo toàn diện của các kỳ Olympics mùa hè!!!

7. Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội?

Để trả lời cho câu hỏi trên ta làm như sau:

- Tính số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội.

	Olympics	Countrys	IsWons	Number of events	Number of medals
0	1896 Summer	Australia	[False, True, True, False, True]	5	3
1	1896 Summer	Austria	[True, True, False, True, False, True, False, ...	8	5
2	1896 Summer	Denmark	[True, True, False, True, False, False, False,...	15	6
3	1896 Summer	France	[True, False, False, False, True, False, True,...	26	11
4	1896 Summer	Germany	[False, False, False, False, False, False, Tru...	60	13
...	...	...	...	...	...
3781	2016 Summer	Virgin Islands, British	[False, False, False, False]	4	0
3782	2016 Summer	Virgin Islands, US	[False, False, False, False, False, False, False]	7	0
3783	2016 Summer	Yemen	[False, False, False]	3	0
3784	2016 Summer	Zambia	[False, False, False, False, False, False, False]	7	0
3785	2016 Summer	Zimbabwe	[False, False, False, False, False, False, Fal...	15	0

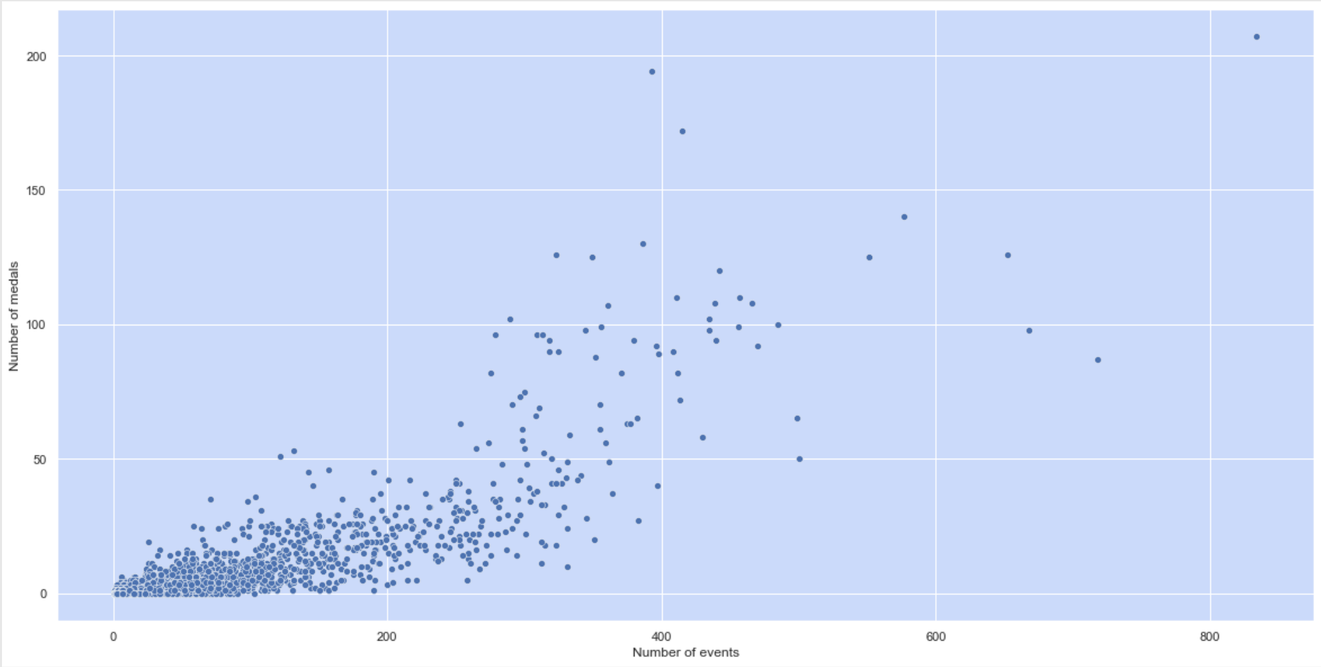
3786 rows × 5 columns

Số nội dung và huy chương đạt được của mỗi quốc gia

7. Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội?

Để trả lời cho câu hỏi trên ta làm như sau:

- Trực quan hóa số nội dung tham gia và số huy chương đạt được.

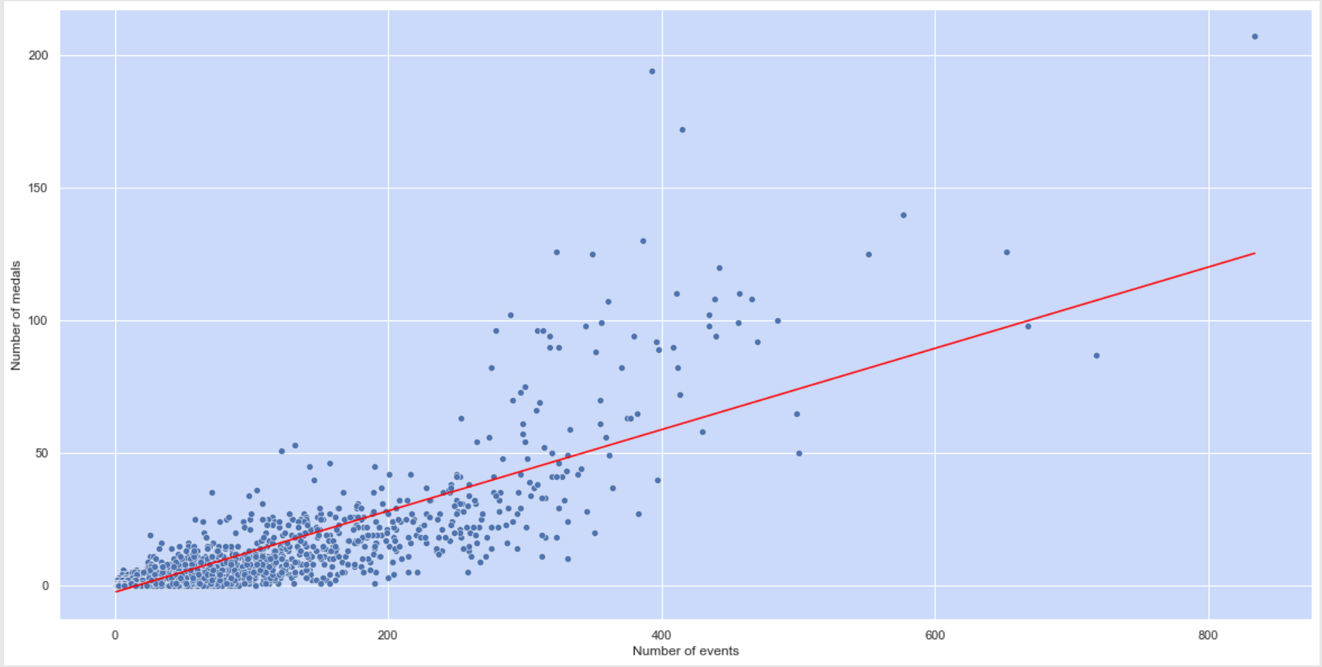


7. Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội?

Để trả lời cho câu hỏi trên ta làm như sau:

➔Từ đồ thị ở trên ta thấy mô hình hồi quy phù hợp là tuyến tính.

- Sử dụng mô hình hồi quy để tìm ra mối quan hệ giữa 2 yếu tố trên.



7. Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội?

Để trả lời cho câu hỏi trên ta làm như sau:

- Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội là mối quan hệ tuyến tính, tuân theo công thức sau đây:

```
print(f'[Number of medals] = ({slope})*[Number of events] + ({intercept})')
```

```
[Number of medals] = (0.1532501045391683)*[Number of events] + (-2.558322659649371)
```

5

Tổng hợp lại quá trình thực hiện đồ án

Tiến trình thực hiện đồ án

STT	Tasks	Người thực ...	Deadline	Status	Merge vào main?	Nội dung
1	Tìm nguồn dữ liệu, đề tài	Tùng Tài Bình Hùng	November 13, 2022	Done	☑	Chọn đề tài olympic
2	Tiền xử lý dữ liệu	Hùng Tài Bình Tùng	December 1, 2022	Done	☑	Thực hiện như các bước trong file hướng dẫn
3	Đặt câu hỏi 1,2	Tài	December 10, 2022	Done	☑	- Top 5 VDV tham gia nhiều năm thế vận hội nhất? - Top 5 VDV đạt nhiều huy chương nhất?
4	Đặt câu hỏi 3,4	Tùng	December 10, 2022	Done	☑	- Quốc gia có thành tích tốt nhất qua các kỳ Olympic mùa hè? - Thành tích của Thể thao Việt Nam trong lịch sử tham dự Olympic?
5	Đặt câu hỏi 5,7	Bình	December 10, 2022	Done	☑	- Trong một thế vận hội: ai là người tham gia thi nhiều nội dung nhất, ai là người giành nhiều huy chương nhất? - Mối quan hệ giữa số nội dung tham gia và số huy chương đạt được của mỗi quốc gia qua các kỳ thế vận hội.
6	Đặt câu hỏi 6	Hùng	December 10, 2022	Done	☑	Trong những năm gần đây (từ khi Olympic mùa hè và mùa đông bắt đầu xen kẽ nhau), mỗi quốc gia đăng cai liệu có lợi thế chủ nhà hay không ?
7	Làm slide	Tài Tùng Bình Hùng	December 14, 2022	Done	☐	
8	Viết báo cáo	Tùng Tài Bình Hùng	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	Việc không hiểu rõ về thông tin dữ liệu khiến em trả lời sai câu hỏi mà em đặt ra.	Tìm hiểu kỹ về dữ liệu trước khi tiến hành phân tích.
2	Đỗ Tấn Tài	- Nội dung trình bày của nhóm khá dài khiến việc làm slide mất nhiều thời gian hơn dự kiến. - Đặc điểm của dữ liệu có một số điểm đặc biệt khiến việc tiền xử lý gặp một số trục trặc nhỏ.	- Nên phân bố thời gian làm slide hợp lý để tránh việc thiếu thời gian, khiến slide chưa được hoàn thiện tốt nhất. - Nên tìm hiểu kỹ trước thông tin dữ liệu trước khi tiến hành xử lý.
3	Trần Khắc Bình	Phân tích chưa kỹ dữ liệu dẫn đến hiểu sai về dữ liệu (tính sai số huy chương của quốc gia do các môn thể thao đồng đội chỉ tính 1 huy chương cho tập thể). Dữ liệu có nhiều cột nên việc tiền xử lý hơn khó khăn.	Dành nhiều thời gian để phân tích dữ liệu.
4	Lê Văn Hùng	- Nội dung dài nên phải cắt bớt câu hỏi để đủ thời gian trình bày của nhóm và làm slide khá mất thời gian - Việc phân tích dữ liệu khá tốn thời gian vì phải tìm những bài báo nước ngoài và đã cũ để tìm kiếm đúng thông tin	- Phải ước lượng trước độ dài của đáp án định làm cũng như các câu hỏi khác để chọn lọc câu hỏi phù hợp độ dài nội dung. - Trau dồi thêm ngôn ngữ để có thể đọc hiểu báo nhanh 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 6

THANK YOU
FOR WATCHING