

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**  
**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN THỰC HÀNH**

**Học phần: ỨNG DỤNG PHÂN TÍCH DỮ LIỆU**  
**THÔNG MINH**

Mã lớp: 20\_21

Sinh viên thực hiện: Lê Hoàng Huy - 20120105

*TP.HCM, tháng 12 năm 20223*

## Mục lục

1. Data Preprocessing .....	1
1.1. Thu thập dữ liệu .....	1
1.2. Làm sạch dữ liệu .....	2
1.2.1. Dữ liệu thiếu và dữ liệu bị trùng lặp .....	2
1.2.2. Nhập dữ liệu.....	2
1.2.3. Phân bố dữ liệu .....	4
2. Phân tích và trực quan hóa dữ liệu .....	5
2.1. Sự phân phối tài sản .....	5
2.2. Nguồn gốc tài sản.....	9
2.3. Phân tích theo địa lý.....	11
2.4. Phân tích theo không gian .....	16
3. Tài liệu tham khảo .....	18

## Danh mục hình

2. 1. Tổng quan về sự tài sản các tỷ phú .....	6
2. 2. Phân hóa giữa tỷ phú tự thân và được thừa kế .....	6
2. 3. Phân bố theo giới tính .....	7
2. 4. Biểu đồ phân bố theo độ tuổi .....	7
2. 5. Biểu đồ phân bố theo nhóm tuổi .....	8
2. 6. Top 10 quốc gia có nhiều tỷ phú nhất .....	9
2. 7. Số lượng tỷ phú theo tháng sinh.....	9
2. 8. Số lượng tỷ phú ở mỗi ngành.....	10
2. 9. Số lượng tỷ phú ở các lĩnh vực phổ biến .....	11
2. 10. Kết quả về các nước có nhiều tỷ phú nhất .....	11
2. 11. Số lượng tỷ phú ở mỗi tiểu bang .....	12
2. 12. Mối tương quan giữa GDP và số lượng tỷ phú .....	13
2. 13. Tương quan CPI và số lượng tỷ phú .....	14
2. 14. Tương quan giữa tỷ lệ đăng ký giáo dục và số lượng tỷ phú.....	14
2. 15. Quan hệ giữa HDI và top10 tỷ phú .....	15
2. 16. Mức thuế trung bình ở các nước nhiều tỷ phú .....	16
2. 17. Tương quan dân số và số lượng tỷ phú .....	16
2. 18. Vị trí theo khu vực .....	18

## 1. Data Preprocessing

### 1.1. Thu thập dữ liệu

- Trong đồ án này, chúng ta thu thập và phân tích danh mục tỷ phú trên thế giới năm 2023 được lấy từ Kaggle.
- Về dataset: dữ liệu là gồm 1 file csv có 2640 dòng và 35 cột ghi nhận thông tin về các tỷ phú hàng đầu trên thế giới và các chỉ số kinh tế, xã hội, và tài chính liên quan đến đất nước của họ.
- Mô tả các cột trong file csv:

Tên cột	Nội dung
<b>Rank</b>	Xếp hạng các người giàu trong danh sách
<b>FinalWorth</b>	Giá trị tài sản của họ (tính bằng USD)
<b>Category</b>	Lĩnh vực hoạt động kinh doanh
<b>PersonName</b>	Tên đầy đủ
<b>Age</b>	Tuổi (tính tới ngày 4/4/2023)
<b>Country</b>	Quốc gia mà họ đang sinh sống
<b>City</b>	Thành phố nơi họ sống
<b>Source</b>	Các nguồn thu nhập chính của họ
<b>Industries</b>	Ngành công nghiệp mà họ liên quan đến
<b>CountryOfCitizenship</b>	Quốc tịch
<b>Organization</b>	Tên tổ chức mà họ liên quan đến (nếu có)
<b>SelfMade</b>	Trạng thái tự tạo ra tài sản (True hoặc False)
<b>Status</b>	Tình trạng hôn nhân (D là độc thân, U là ly hôn)
<b>Gender</b>	Giới tính (M là nam, F là nữ)
<b>BirthDate</b>	Ngày tháng năm sinh của họ
<b>Lastname</b>	Họ
<b>Firstname</b>	Tên
<b>Title</b>	Chức vị trong tổ chức (nếu có)

<b>Date</b>	Ngày tháng năm mà dữ liệu cập nhật
<b>State</b>	Tiểu bang hoặc khu vực họ đang sống
<b>ResidenceStateRegion</b>	Khu vực hoặc vùng họ đang sống
<b>BirthYear</b>	Năm sinh
<b>BirthMonth</b>	Tháng sinh
<b>BirthDay</b>	Ngày sinh
<b>CPI_country</b>	Chỉ số giá trị tiêu dùng quốc gia (nơi họ sống)
<b>CPI_change_country</b>	Thay đổi chỉ số giá trị tiêu dùng quốc gia (nơi họ sống)
<b>GDP_country</b>	GDP của quốc gia (nơi họ sống)
<b>Gross_tertiary ...</b>	Tỷ lệ đăng ký giáo dục đại học (nơi họ sống)
<b>Gross_primary ...</b>	Tỷ lệ đăng ký giáo dục tiểu học (nơi họ sống)
<b>Life_expectancy_country</b>	Tuổi thọ trung bình (nơi họ sống)
<b>Tax_revenue ...</b>	Thuế thu tại quốc gia (nơi họ sống)
<b>Total_tax_rate_country</b>	Tổng thuế (nơi họ sống)
<b>Population_country</b>	Dân số của quốc gia (nơi họ sống)
<b>Latitude_country</b>	Vĩ độ của quốc gia (nơi họ sống)
<b>Longtitude_country</b>	Kinh độ của quốc gia (nơi họ sống)

## 1.2. Làm sạch dữ liệu

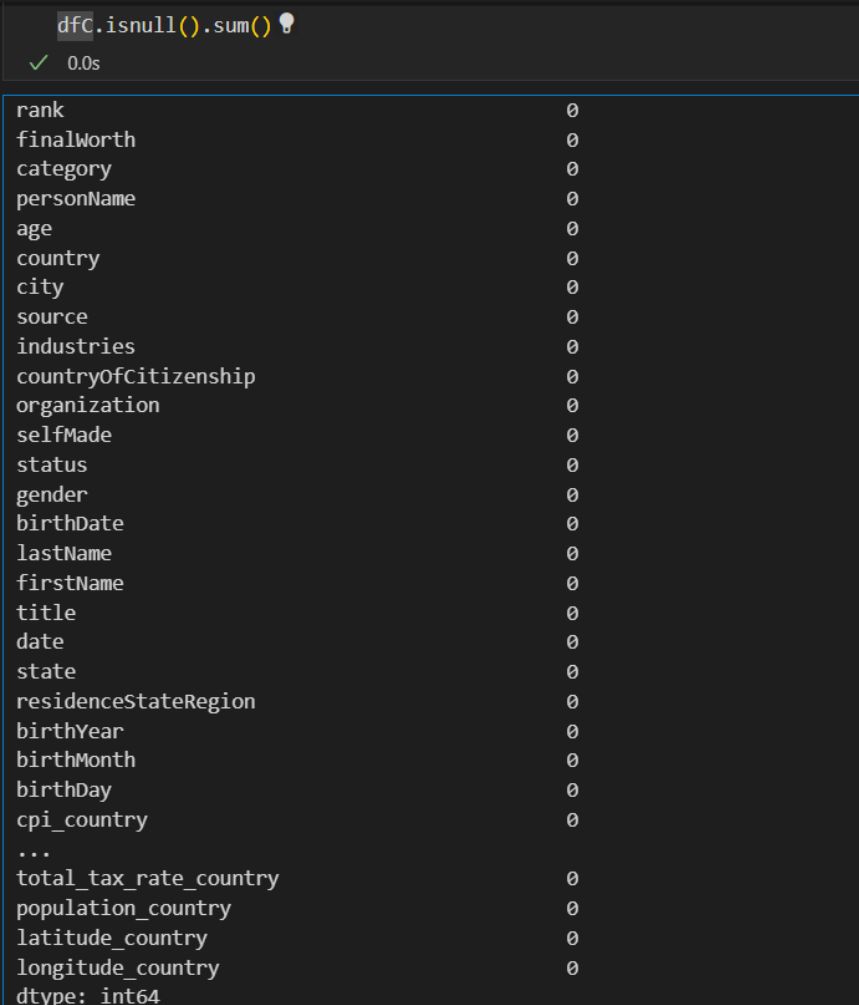
### 1.2.1. Dữ liệu thiếu và dữ liệu bị trùng lặp

- Ở các cột, ngoại trừ các cột có thể để trống thì dữ liệu bị thiếu không đáng kể.
- Tương tự như vậy, ta có thể thấy không có dữ liệu nào bị trùng.
- Do đó ta sẽ bỏ qua bước này.

### 1.2.2. Nhập dữ liệu

- Đối với cột 'country' và 'city', các giá trị còn thiếu được điền bằng mode (giá trị xuất hiện nhiều nhất) của từng cột tương ứng.
- Đối với 'age', các giá trị còn thiếu được điền bằng trung bình của cột 'age'.
- Đối với 'title', 'organization', 'residenceStateRegion', và 'state', giá trị còn thiếu

- được điền bằng chuỗi 'Not Specified'.
- Các giá trị còn thiếu trong cột 'birthDate' được điền bằng giá trị phổ biến nhất (mode) trong cột 'birthDate'.
  - Một số tên đầu tiên được ánh xạ với họ tương ứng bằng cách sử dụng một từ điển được định nghĩa trước (**name\_mapping**). Ví dụ, nếu họ là 'Tahir', tên đầu tiên sẽ được đặt là 'Muhammad'.
  - Đối với 'birthYear', 'birthMonth', và 'birthDay', các giá trị còn thiếu được điền bằng trung bình, mode, và mode của từng cột tương ứng.
  - Đối với các cột khác như 'cpi\_country', 'cpi\_change\_country', v.v., giá trị còn thiếu được điền bằng trung bình hoặc trung vị của từng cột tương ứng.
  - Sau khi thực hiện các công đoạn trên ta thu được kết quả sau:



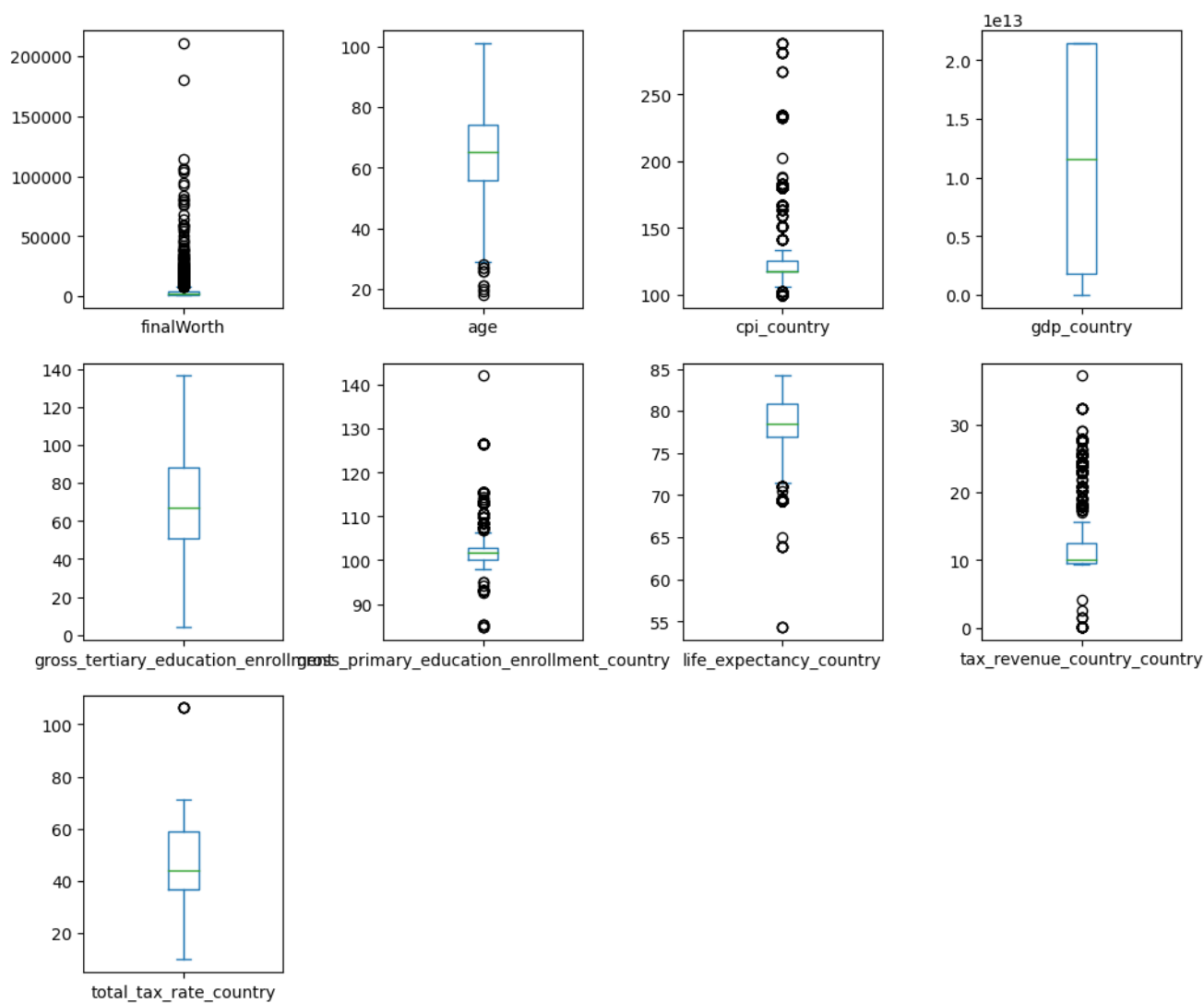
```
dfC.isnull().sum()
```

rank	0
finalWorth	0
category	0
personName	0
age	0
country	0
city	0
source	0
industries	0
countryOfCitizenship	0
organization	0
selfMade	0
status	0
gender	0
birthDate	0
lastName	0
firstName	0
title	0
date	0
state	0
residenceStateRegion	0
birthYear	0
birthMonth	0
birthDay	0
cpi_country	0
...	
total_tax_rate_country	0
population_country	0
latitude_country	0
longitude_country	0
dtype: int64	

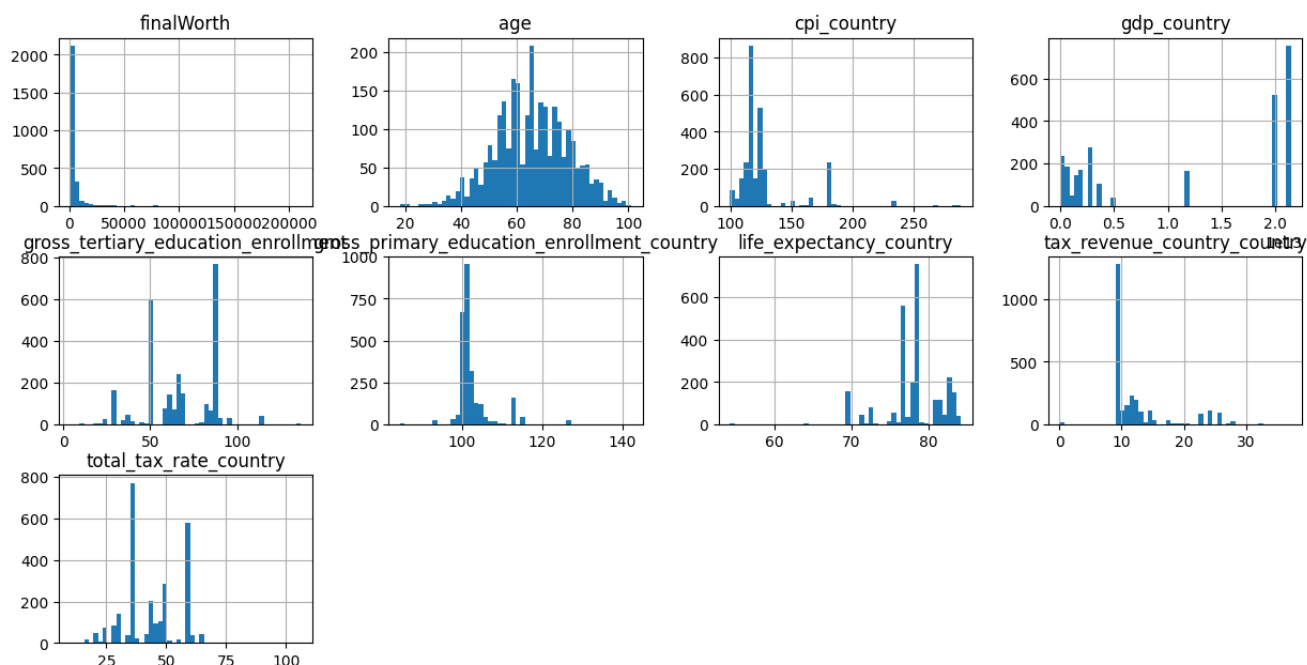
### 1. 1. Kết quả kiểm tra giá trị null

### 1.2.3. Phân bố dữ liệu

- Ở bước này chúng ta chỉ kiểm tra sự phân bố của các dữ liệu dạng số.
- Lần lượt vẽ các biểu đồ dạng boxplot và biểu đồ histogram để xem xét sự phân bố dữ liệu của các cột có dạng số.
- Ta thu được 2 các kết quả sau:



1. 2. Biểu đồ boxplot của các cột dạng số



### 1. 3. Biểu đồ histogram cho các cột dạng số

- Dữ liệu phân bố một cách hợp lí, không có vấn đề cần khắc phục nên ta sẽ coi như hoàn thành bước này.

## 2. Phân tích và trực quan hóa dữ liệu

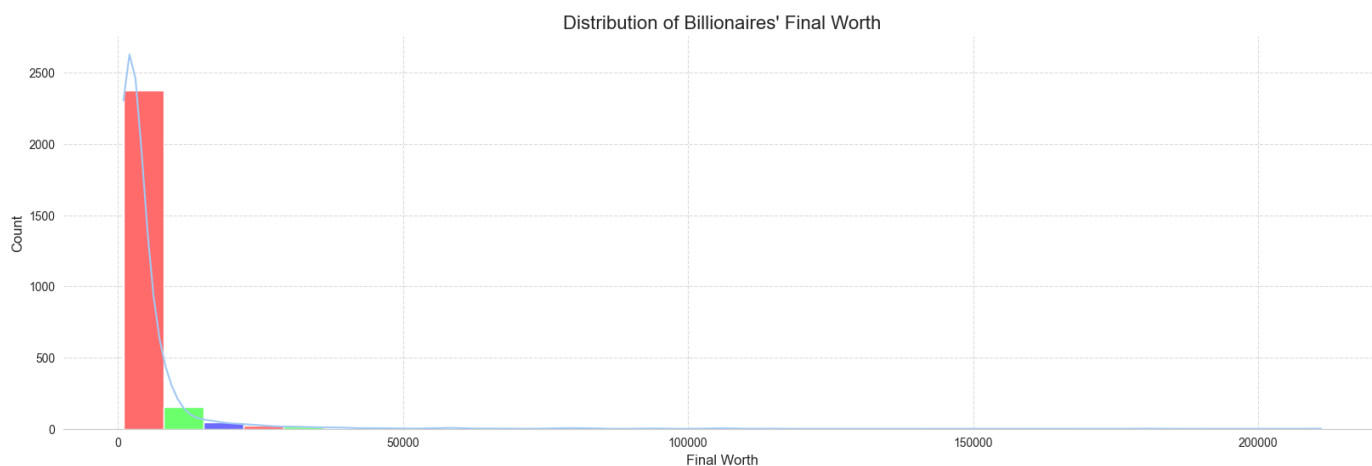
### 2.1. Sự phân phối tài sản

#### Các vấn đề sẽ phân tích

- Cái nhìn tổng quan.
- Có bao nhiêu tỷ phú tự thân và được thừa kế tài sản?
- Có bất kì xu hướng hoặc mô hình tích lũy tài sản dựa trên giới tính không?
- Phân bố theo độ tuổi.
- Phân bố theo quốc gia.
- Phân bố theo tháng sinh.

- Tổng quan về tài sản của các tỷ phú

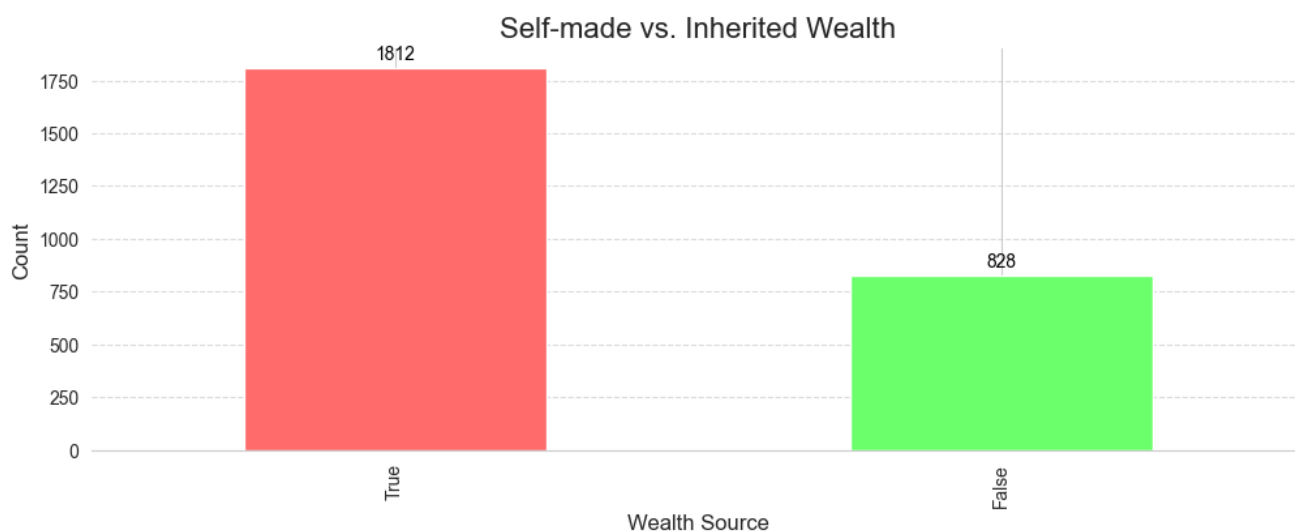




### 2. 1. Tổng quan về sự tài sản các tỷ phú

- Sự phân hóa giữa tỷ phú tự thân và tỷ phú được thừa kế

Ở đây chúng ta chỉ cần vẽ 2 biểu đồ cột để đếm số lượng tỷ phú tự thân và được thừa kế.



### 2. 2. Phân hóa giữa tỷ phú tự thân và được thừa kế

⇒ Nhận thấy trong tổng số 2640 người trong danh sách có tới 1812 người (68.6%) người là tự thân trở thành tỷ phú.

- Sự phân bố theo giới tính

Tiến hành vẽ violin plot để tìm hiểu về sự phân bố theo 2 giới tính nam và nữ.



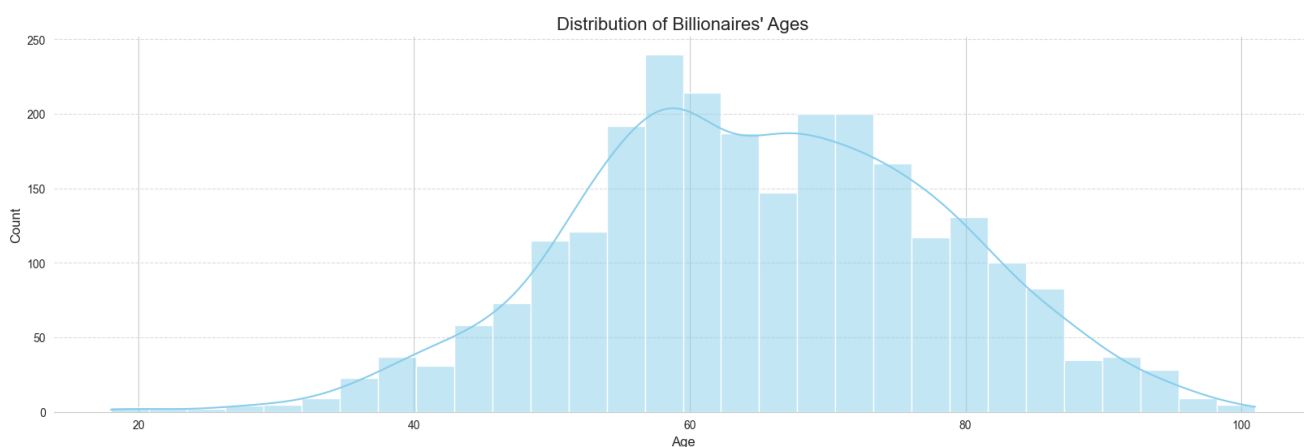
### 2. 3. Phân bố theo giới tính

⇒ Ở nữ có sự tích lũy tài sản nhiều hơn từ 0 đến dưới 100000. Tuy nhiên khi tài sản lớn hơn thì phần lớn tỷ phú có giới tính nam.

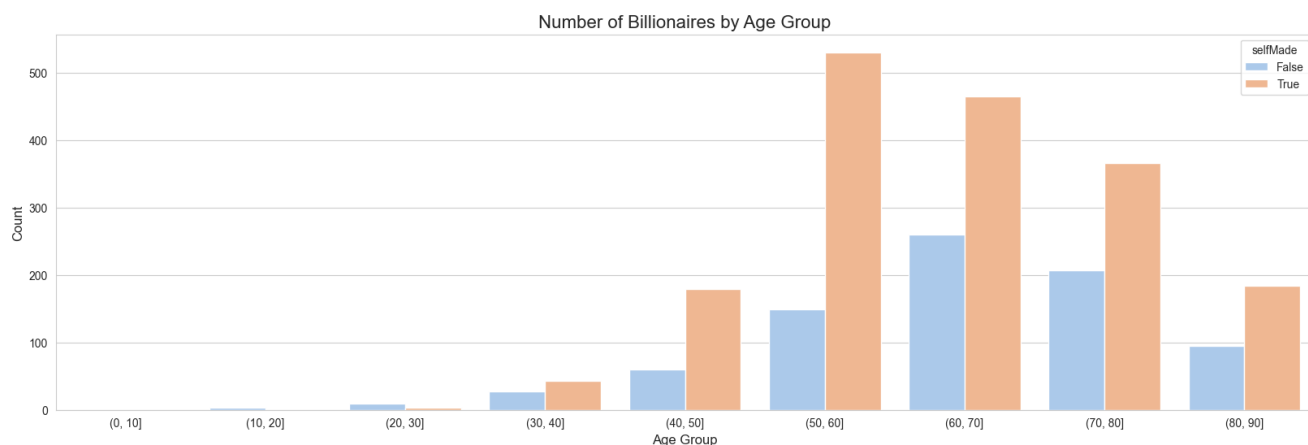
- Phân bố theo độ tuổi

Theo độ tuổi chúng ta sẽ tiến hành phân tích theo 2 hướng: theo từng độ tuổi và theo nhóm tuổi.

Ở độ tuổi chúng ta sẽ vẽ biểu đồ cột kết hợp với đường từ, còn ở theo độ tuổi chúng ta sẽ vẽ 2 cột theo mỗi nhóm tuổi để rút ra kết luận về sự tương quan về độ tuổi và vấn đề tài sản được thừa kế hay không.



### 2. 4. Biểu đồ phân bố theo độ tuổi



### 2. 5. Biểu đồ phân bố theo nhóm tuổi

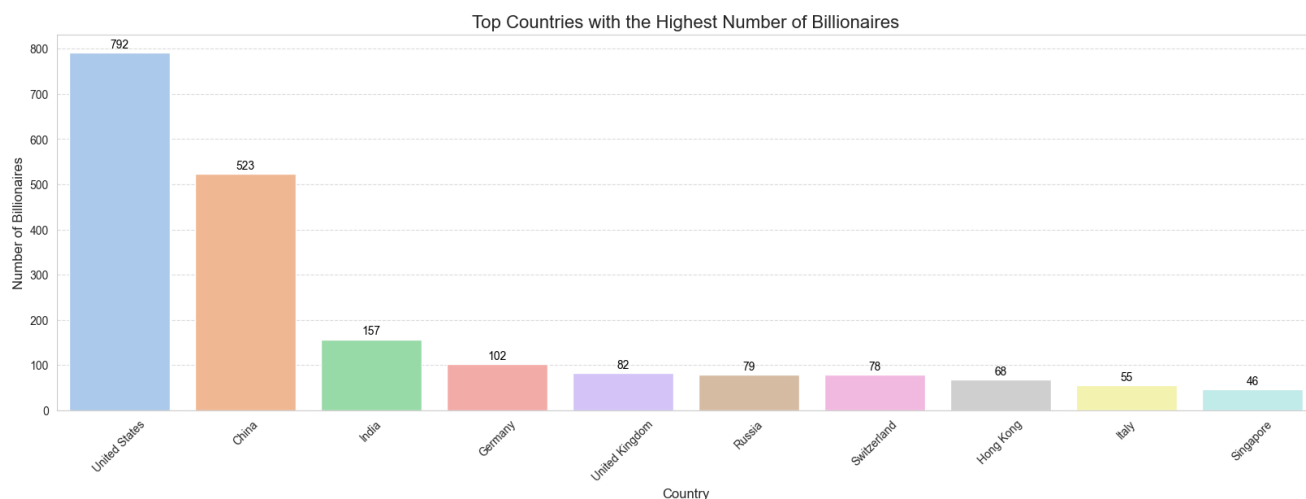
⇒ Chúng ta có thể suy ra từ biểu đồ đầu tiên, rằng những người trên 80 tuổi có mức tài sản trung bình cao hơn, nhưng các nhóm tuổi khác có mức tài sản trung bình gần như tương tự, trong đó nhóm tuổi 40-50 có mức tài sản trung bình thấp nhất.

Điều thú vị là, từ biểu đồ thứ hai, chúng ta có thể thấy rằng những người trên 90 tuổi có ít cổ phần gia đình hơn và những người trẻ tuổi không có tài sản chung của gia đình.

Những người dưới 40 tuổi cũng có số cổ phần gia đình ít hơn, trong khi những người ở độ tuổi từ 50-90 có số cổ phần gia đình cao nhất.

- Các quốc gia có nhiều tỷ phú nhất

Vì số lượng các quốc gia quá lớn, đồng thời tồn tại rất nhiều nước không có tỷ phú nào trong danh sách. Do đó chúng ta chỉ tiến hành xem xét top 10 quốc gia có số lượng tỷ phú nhiều nhất.

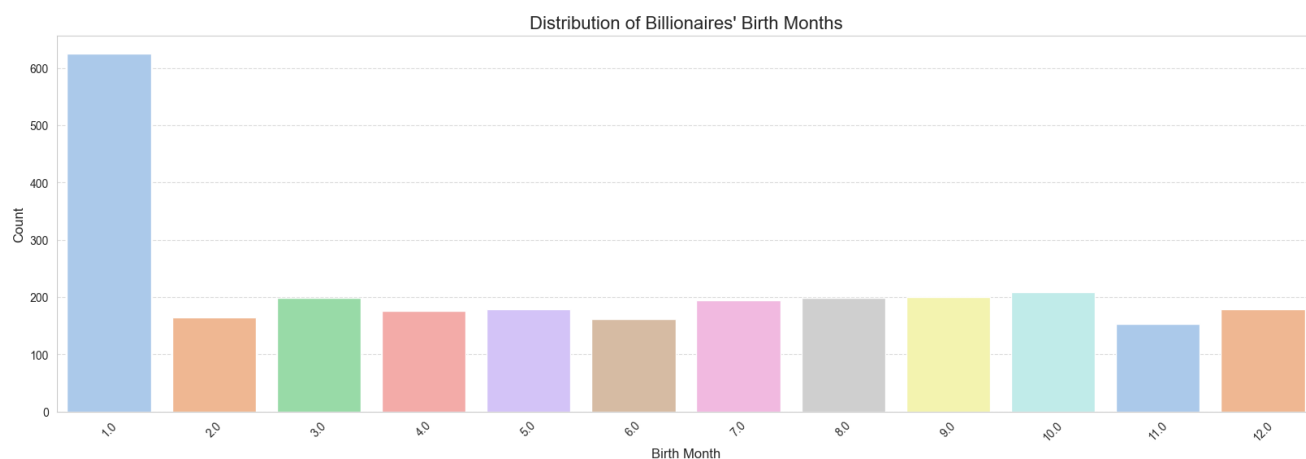


### 2. 6. Top 10 quốc gia có nhiều tỷ phú nhất

⇒ Cùng nằm trong top 10, nhưng Mỹ và Trung Quốc có số lượng tỷ phú vượt trội so với các nước khác.

- Phân bố theo tháng sinh

Đơn giản chúng ta chỉ cần vẽ biểu đồ cột cho 12 tháng trong năm và đếm số lượng tỷ phú theo mỗi tháng.



### 2. 7. Số lượng tỷ phú theo tháng sinh

⇒ Từ biểu đồ trên ta thấy được rằng, số lượng tỷ phú ở sinh ra phần lớn ở tháng 1 (hơn 600 người) trong khi các tháng khác số lượng tỷ phú chưa tới 200 người.

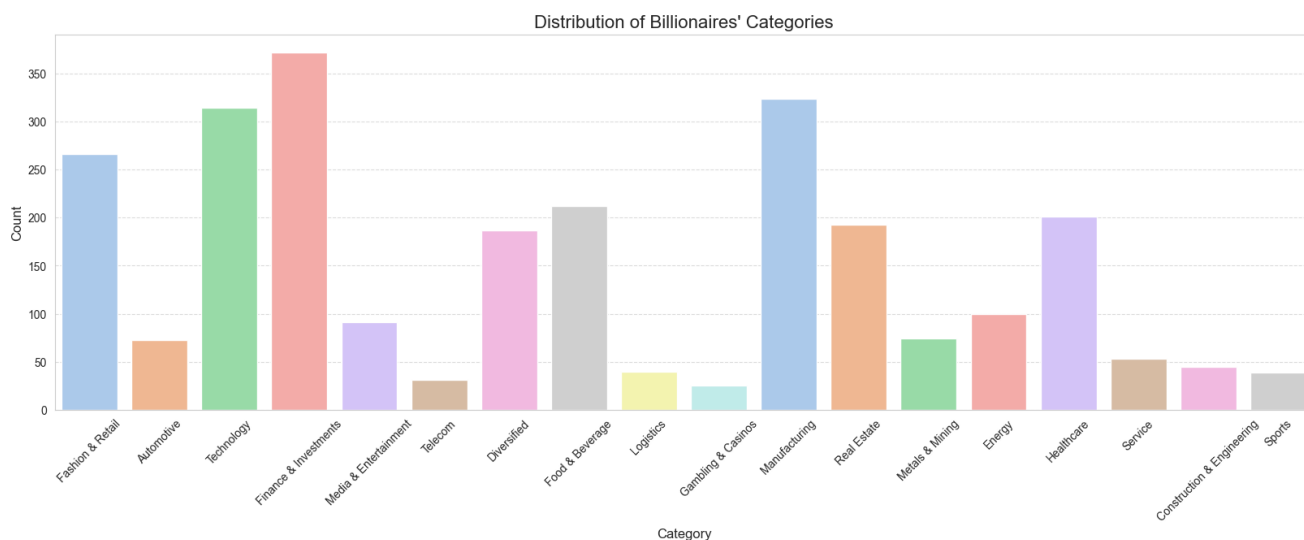
## 2.2. Nguồn gốc tài sản

### Các vấn đề sẽ phân tích

- Các ngành chiếm ưu thế về số lượng tỷ phú đại diện.
- Tài sản chủ yếu của họ trong lĩnh vực gì?

- Các ngành có nhiều tỷ phú nhất so với các ngành còn lại

Tiến hành vẽ biểu đồ cột về số lượng tỷ phú cho tất cả các ngành từ đó quan sát và rút ra nhận xét cơ bản về số lượng tỷ phú ở mỗi ngành.

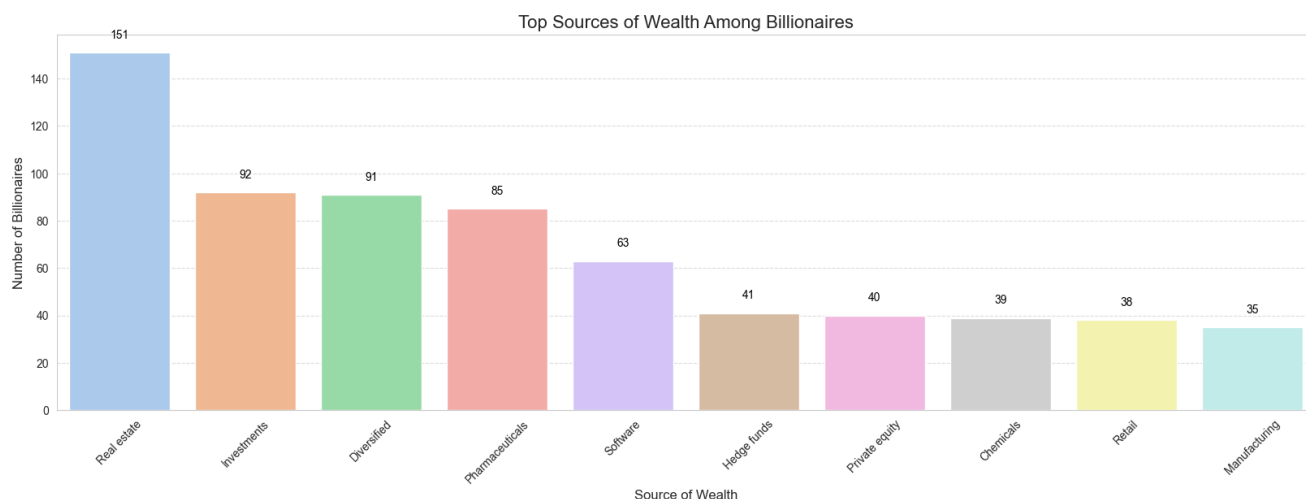


### 2. 8. Số lượng tỷ phú ở mỗi ngành

⇒ Các ngành như tài chính, công nghệ, chế tác và thời trang có ưu thế về số lượng tỷ phú (hơn 250 người) so với các ngành còn lại.

- Từ các ngành nghề trên chúng ta sẽ tiếp tục tìm hiểu về lĩnh vực chủ yếu của các tỷ phú

Tiến hành vẽ biểu đồ tương tự như trên để tìm hiểu số lượng tỷ phú ở mỗi lĩnh vực.



## 2. 9. Số lượng tỷ phú ở các lĩnh vực phổ biến

⇒ Dựa theo tìm hiểu về số lượng tỷ phú ở các ngành ta có thể thấy được rằng, Finance & Investment có số lượng tỷ phú dẫn đầu. Do đó phần nào ta cũng hiểu được vì sao phần lớn tài sản của họ đến từ Real estate và Investment (151 người) trong khi các lĩnh vực khác chưa tới 100 người.

## 2.3. Phân tích theo địa lý

Đầu tiên chúng ta sẽ đếm số lượng các tỷ phú ở các nước.

	country	count
0	United States	792
1	China	523
2	India	157
3	Germany	102
4	United Kingdom	82

## 2. 10. Kết quả về các nước có nhiều tỷ phú nhất

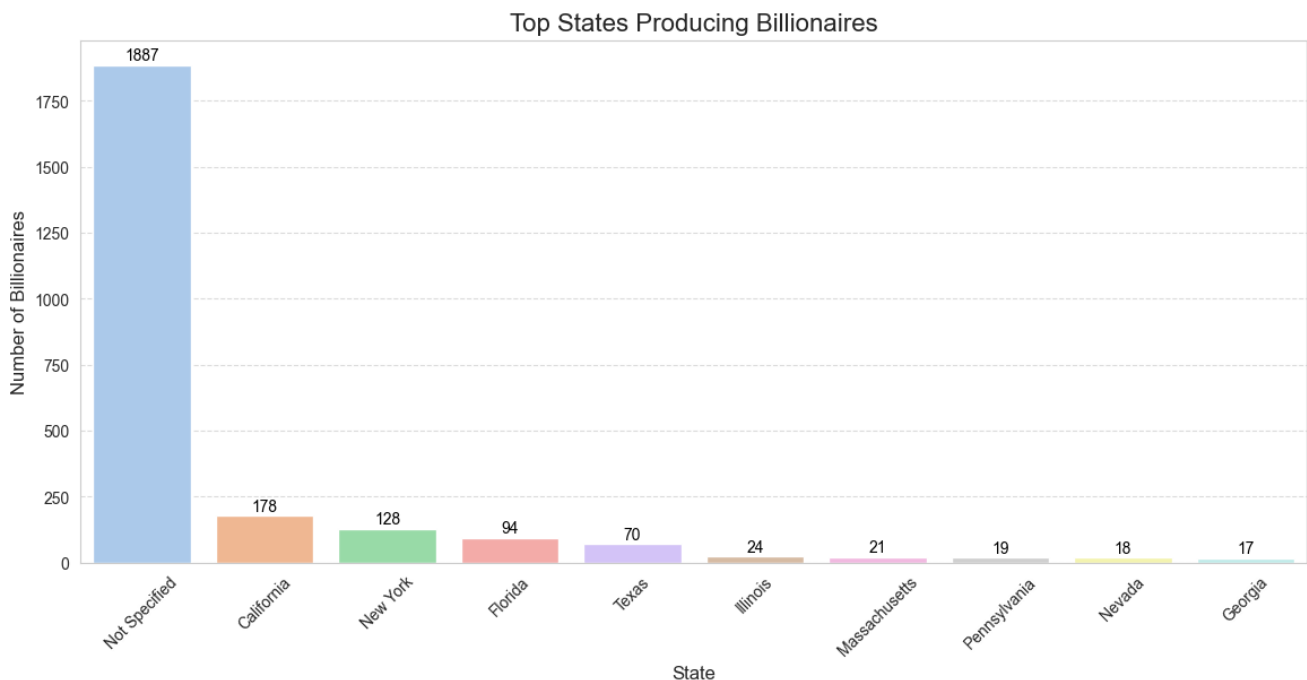
Từ kết quả trên ta thấy được rằng, nước Mỹ có nhiều tỷ phú nhất. Do đó ta sẽ tập trung chủ yếu phân tích vào đất nước này, đồng thời cũng tiến hành so sánh với các nước còn lại trong bảng xếp hạng.

### Các vấn đề sẽ phân tích

- Có tồn tại khu vực hoặc tiểu bang nào chiếm phần lớn tỷ phú không?
- Tổng GDP của quốc gia liên quan thế nào đến số lượng tỷ phú?
- Mối tương quan giữa CPI và số lượng tỷ phú là gì?
- Tỷ lệ đăng ký giáo dục và số lượng tỷ phú.
- Mối tương quan giữa tỷ phú và chỉ số HDI của mỗi quốc gia.
- Mức thuế trung bình ở các quốc gia nhiều tỷ phú.
- Dân số và số lượng tỷ phú ở mỗi quốc gia.

- Số lượng tỷ phú ở mỗi tiểu bang (nước Mỹ)

Nước Mỹ có tới 50 tiểu bang, vì vậy sẽ gây khó khăn trong việc trực quan hóa tất cả. Do đó ở đây chúng ta chỉ vẽ biểu đồ cột cho top 10 tiểu bang ở nước Mỹ.



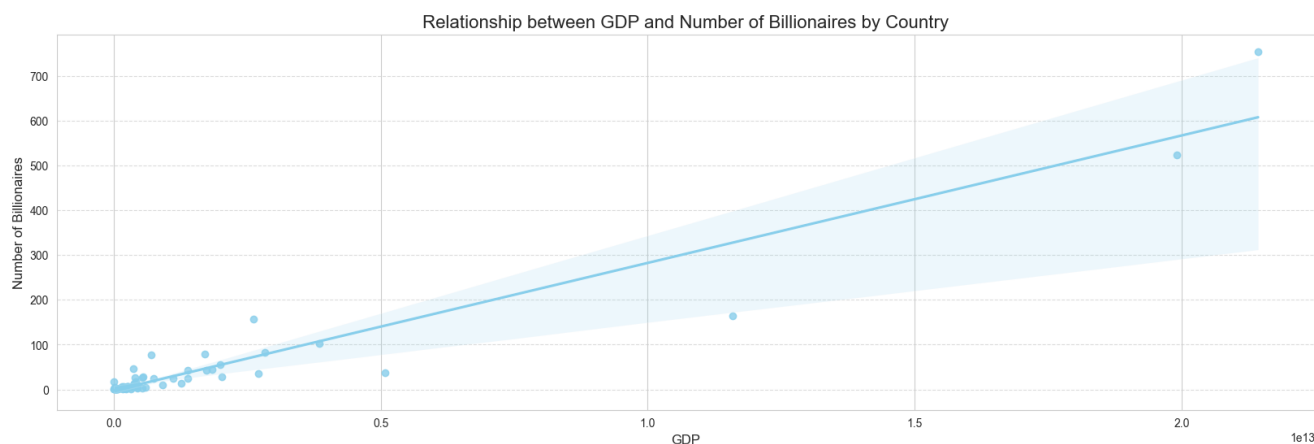
### 2. 11. Số lượng tỷ phú ở mỗi tiểu bang

⇒ California và New York có số lượng nhiều hơn so với các tiểu bang khác (hơn 100 người).

- Phân tích GDP và số lượng tỷ phú ở các quốc gia

Dùng một biểu đồ phân tán kết hợp với một đường hồi quy từ đó rút ra tương quan

giữa GDP và số lượng tỷ phú ở các quốc gia.



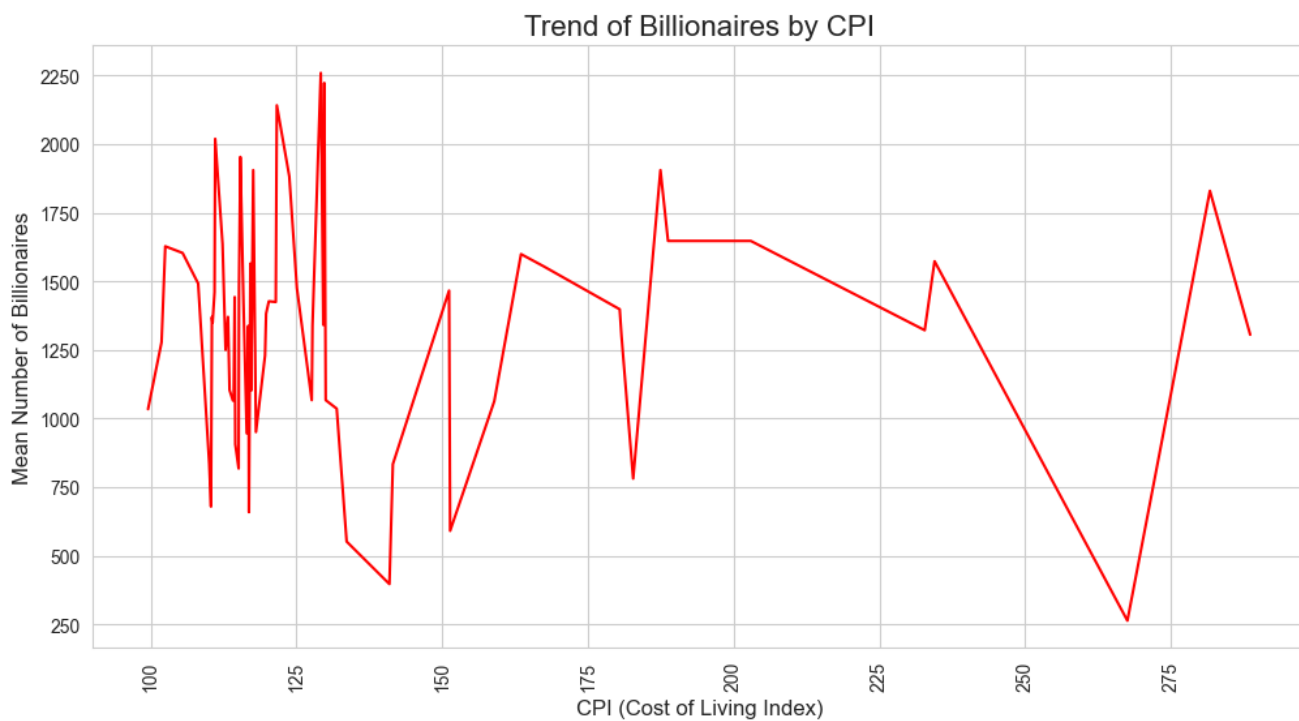
## 2. 12. Mối tương quan giữa GDP và số lượng tỷ phú

⇒ Biểu đồ trên mô tả mối quan hệ giữa GDP và số lượng tỷ phú theo từng quốc gia. Đường hồi quy tăng dần, cho thấy sự tương quan tích cực giữa hai biến số này. Các quốc gia có GDP lớn thường có số lượng tỷ phú cao. Tuy nhiên, cũng cần chú ý đến các trường hợp ngoại lệ và yếu tố bổ sung để có cái nhìn toàn diện về mối quan hệ này.

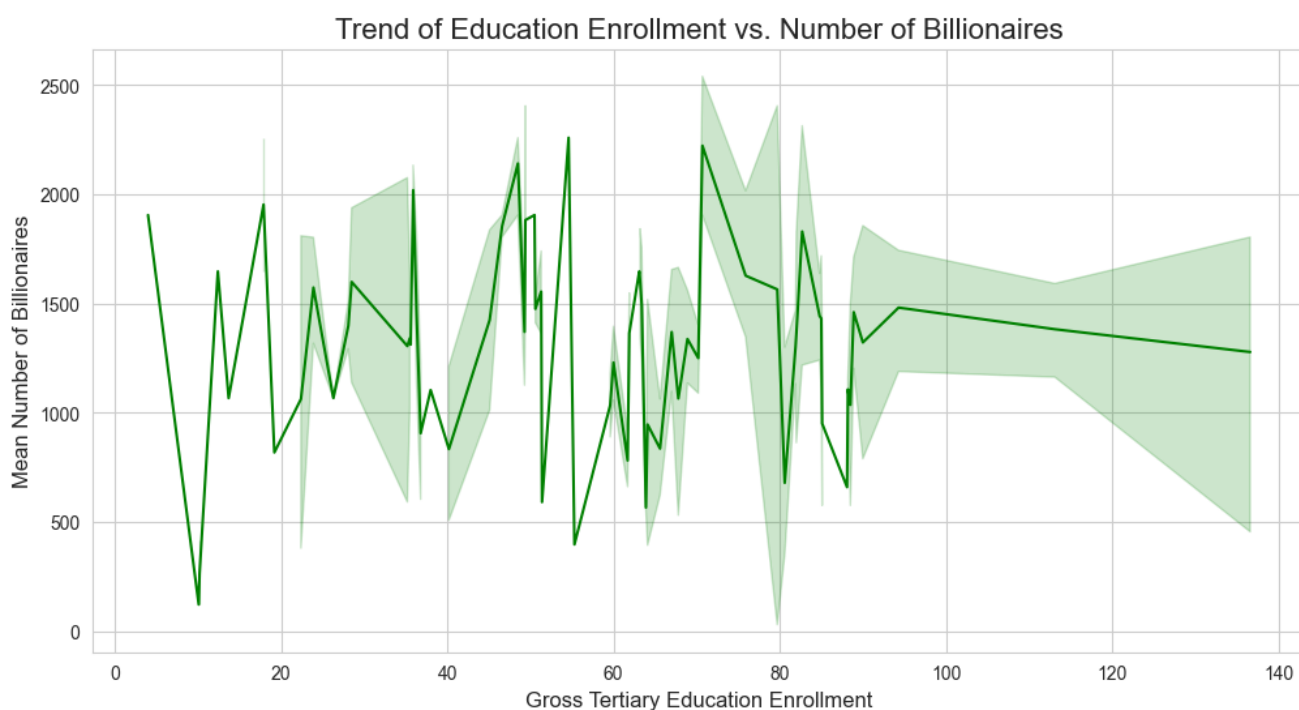
- Phân tích CPI, tỷ lệ đăng ký giáo dục và số lượng tỷ phú

Ở phần này chúng ta sẽ tiến hành vẽ lần lượt hai biểu đồ đường giữa hai thông số trên so với số lượng tỷ phú từ đó ta rút ra nhận xét.





2. 13. Tương quan CPI và số lượng tỷ phú



2. 14. Tương quan giữa tỷ lệ đăng ký giáo dục và số lượng tỷ phú

⇒ Nhận thấy hai chỉ số trên đều thay đổi không ổn định theo thời gian. Tuy nhiên có thể

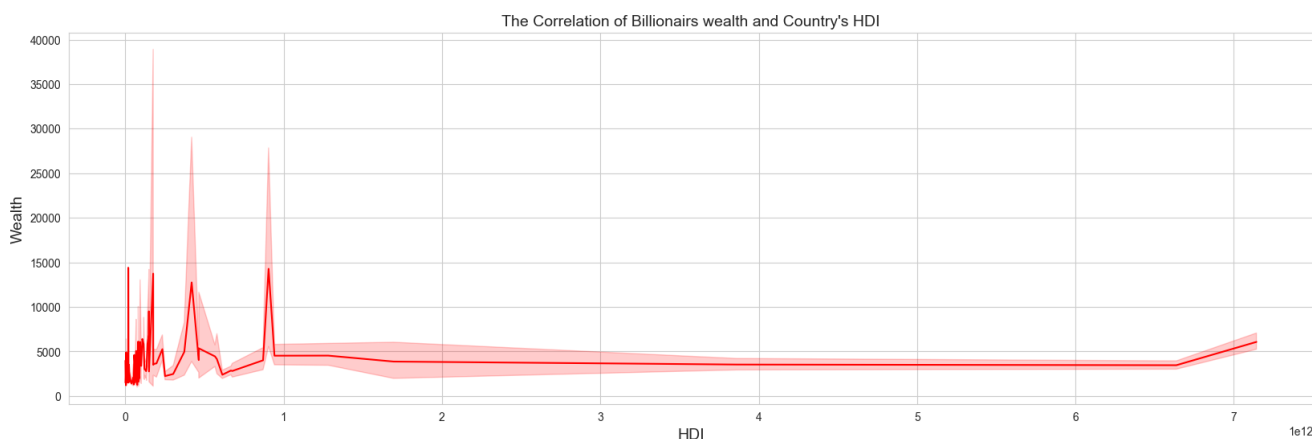
thấy rằng CPI không ảnh hưởng nhiều đến số lượng tỷ phú. Mặc khác tỷ lệ đăng ký giáo dục dù không ổn định, không tang mạnh nhưng vẫn giữ ở mức cao. Từ đó cho thấy giáo dục có ảnh hưởng lớn đến số lượng tỷ phú ở mỗi quốc gia.

- Mối tương quan của các tỷ phú hàng đầu với chỉ số HDI của mỗi quốc gia

Tại đây chúng ta sẽ tìm hiểu về một chỉ số mới là HDI ( chỉ số phát triển con người) dựa trên dữ liệu từ GDP, tỷ lệ đăng ký giáo dục tiểu học và tuổi thọ trung bình tại cùng quốc gia.

Đầu tiên ta tính HDI bằng cách tính trung bình công của 3 chỉ số trên.

Tiếp theo ta vẽ biểu đồ đường cho chỉ số HDI vừa tính được với top 10 tỷ phú hàng đầu trên thế giới.



2. 15. Quan hệ giữa HDI và top10 tỷ phú

⇒ Biểu đồ đường vừa tạo ra để theo dõi mối quan hệ giữa chỉ số Phát triển Nhân loại (HDI) và tổng tài sản của các tỷ phú không cho thấy sự tương quan rõ ràng giữa hai biến số này. Dường như, mức độ phát triển kinh tế và xã hội của một quốc gia, như đo lường bằng HDI, không tương đồng với sự giàu có của các tỷ phú. Có thể xuất phát từ nhiều yếu tố khác nhau như chính sách kinh tế, môi trường kinh doanh, hoặc các biến động trong thị trường tài chính. Điều này cho thấy rằng, trong việc hiểu về sự phát triển kinh tế và sự giàu có cá nhân, cần phải xem xét một loạt các yếu tố và không chỉ dựa vào chỉ số HDI một cách đơn giản.

- Cuối cùng ở phần này chúng ta sẽ tính thuế trung bình ở các nước có nhiều tỷ phú, sau đó quan sát mối tương quan giữa dân số so với số lượng tỷ phú.

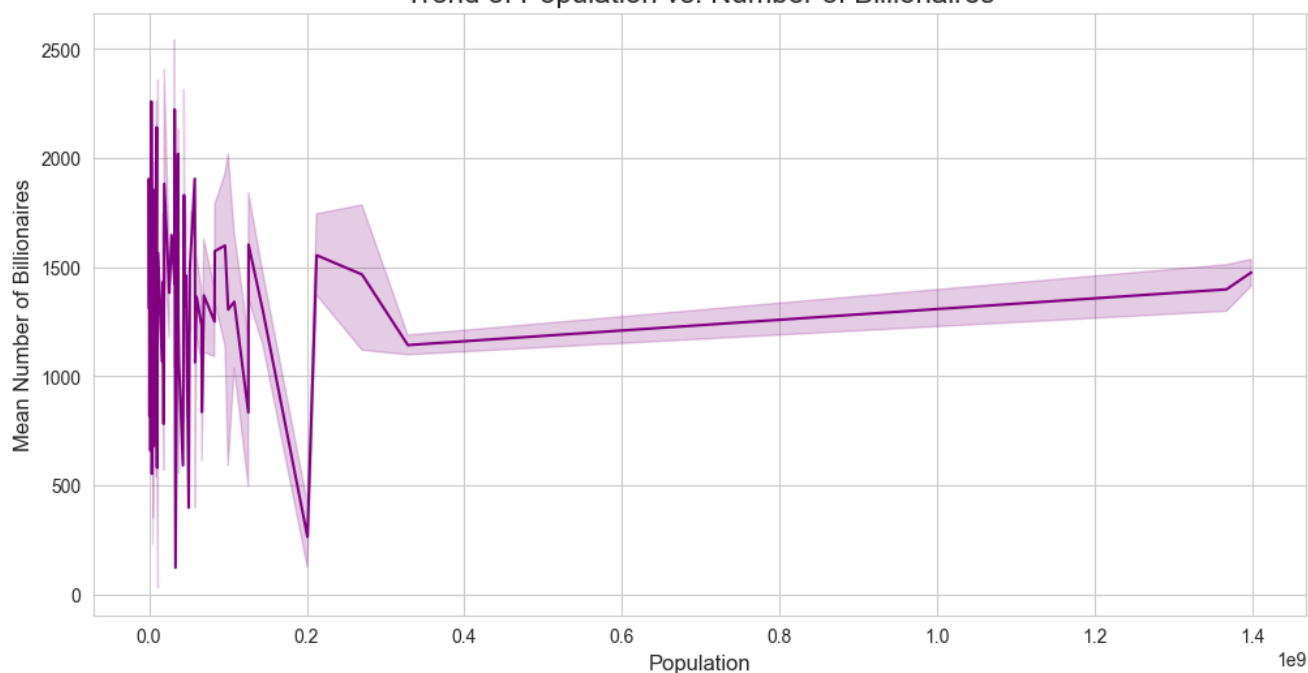
Để thực hiện phần này đầu tiên ta tính được thuế trung bình như sau:

**Average Total Tax Rate in Countries with the Most Billionaires: 44.71**

## 2. 16. Mức thuế trung bình ở các nước nhiều tỷ phú

Tiến hành vẽ biểu đồ đường để tìm quan hệ giữa dân số là số lượng tỷ phú.

Trend of Population vs. Number of Billionaires



## 2. 17. Tương quan dân số và số lượng tỷ phú

⇒ Khi dân số của một quốc gia tăng lên, số lượng tỷ phú có xu hướng tăng nhẹ. Tuy nhiên, mối tương quan tương đối yếu, cho thấy rằng riêng dân số có thể không phải là yếu tố dự báo mạnh mẽ về số lượng tỷ phú ở một quốc gia. Các yếu tố khác có thể góp phần vào mối quan hệ này.

## 2.4. Phân tích theo không gian

**Các vấn đề sẽ phân tích**

- Trực quan hóa vị trí các tỷ phú trên bản đồ thế giới.
- Sự phân bố của họ theo kinh độ và vĩ độ.
- Cuối cùng trả lời câu hỏi họ chiếm bao nhiêu phần trăm tài sản thế giới?

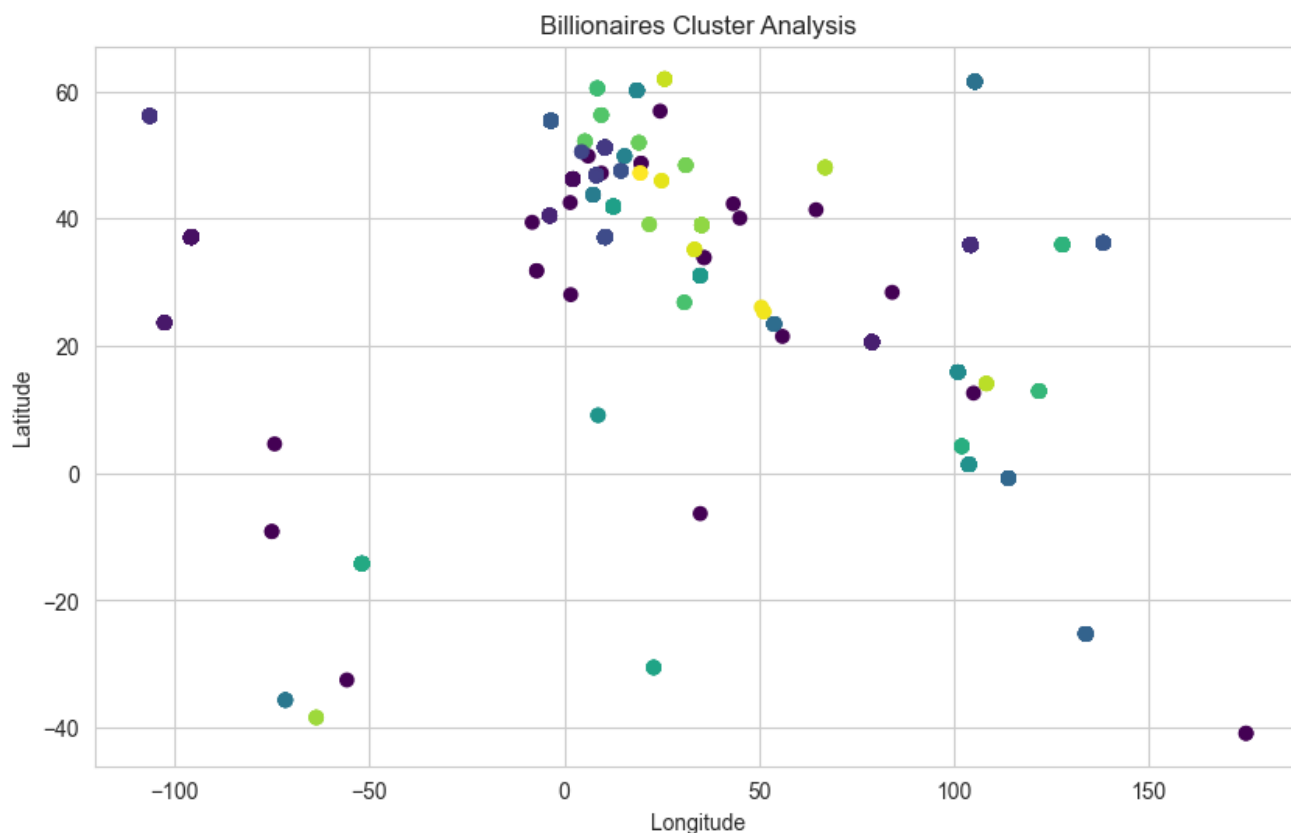
- Trực quan hóa vị trí các tỷ phú

Ở phần này, chúng ta sử dụng thư viện Folium để tạo một bản đồ thế giới và thêm các điểm đánh dấu tại vị trí địa lý của các quốc gia từ DataFrame `dfC`. Sử dụng Marker Cluster để nhóm các điểm đánh dấu gần nhau, làm cho bản đồ trở nên gọn gàng và dễ quản lý. Các điểm đánh dấu được đặt tại vị trí có tọa độ latitude và longitude, giúp hiển thị sự phân bố địa lý của các quốc gia trên bản đồ.

**(Xem kết quả trong file notebook)**

- Vị trí các tỷ phú theo khu vực (kinh độ, vĩ độ)

Sử dụng thuật toán DBSCAN để phân cụm quốc gia dựa trên tọa độ địa lý. Sau đó, vẽ đồ thị scatter để thể hiện sự phân bố của các cụm trên bản đồ.



2. 18. Vị trí theo khu vực

⇒ Các tỷ phú tập trung trong khoảng từ vĩ độ 20-60 và từ kinh độ 0-50.

- Tài sản các tỷ phú trong danh sách so với tài sản thế giới

Biết rằng tổng tài sản cả thế giới là 514.017 tỷ USD (số liệu năm 2023). Do đó sau khi tính ta thu được kết quả là: **0.0023747852697478876**

Có thể thấy được các tỷ phú tuy sở hữu tài sản rất lớn nhưng chỉ chiếm vồn vẹn 0.002% tổng tài sản thế giới.

### 3. Tài liệu tham khảo

Trong quá trình thực hiện đồ án, em có thảo khảo cũng như nghiên cứu các nguồn online. Do quá nhiều lần tra cứu không thể liệt kê hết ở đây, em chỉ trình bày các nguồn tham khảo chính sau:

- [Matplotlib Cheatsheet](#)

- [Matplotlib Documentation](#)
- [Seaborn Gallery](#)
- [Stack Overflow](#)
- [Geeks for Geeks](#)