ĐẠI HỌC BÁCH KHOA HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
——————— * ———————

BÁO CÁO CUỐI KỲ
MÔN: PHÂN TÍCH NGHIỆP VỤ THÔNG MINH

Project 8. Phân tích tuổi thọ

Nhóm sinh viên thực hiện: Nhóm 10

STT	MSSV	Họ tên
1	20194185	Trịnh Đức Tiệp
2	20194139	Trần Văn Phúc
3	20194102	Trần Thành Long

Mã lớp: 141350
GVHD: TS. Nguyễn Bình Minh

Hà Nội, tháng 7 năm 2023

MỤC LỤC

1. Phân công công việc các thành viên trong nhóm

Bảng phân chia công việc của các thành viên nhóm 10:

MSSV	Họ và tên	Công việc
20194102	Trần Thành Long	+ Cài đặt cụm bigdata xử lý bài toán; + Tiền xử lý dữ liệu; + Làm báo cáo, slide thuyết trình
20194185	Trịnh Đức Tiệp	**+ CH2.** Phân tích GDP của khu vực có tuổi thọ cao nhất, thấp nhất; **+ CH4.** Phân tích xu hướng GDP và tuổi thọ
20194139	Trần Văn Phúc	**+ CH1.** Phân tích mối tương quan giữa GDP và tuổi thọ; + **CH3.** Phân tích các vấn đề tiềm ẩn ở những quốc gia có tuổi thọ thấp

2. Giới thiệu bài toán

2.1. Mô tả bài toán

a) Giới thiệu bài toán

Tuổi thọ là một chỉ số quan trọng để đo lường sức khỏe và chất lượng cuộc sống của một quốc gia hoặc khu vực cụ thể. Chỉ số này không chỉ phản ánh mức độ y tế trong quốc gia, mà còn bị ảnh hưởng bởi các yếu tố môi trường, bối cảnh kinh tế và chính trị, cùng với các xu hướng xã hội.

Trong dự án phân tích kinh doanh này, nhóm em sẽ tập trung vào việc phân tích mối tương quan giữa tổng sản phẩm quốc nội (GDP) bình quân đầu người và tuổi thọ. Mục tiêu là tìm hiểu liệu có mối liên hệ rõ ràng giữa mức độ phát triển kinh tế và tuổi thọ, và cũng khám phá ra những hiểu biết có ý nghĩa từ dữ liệu.

b) Mục tiêu bài toán

Mục tiêu của dự án này là thực hiện phân tích sâu hơn về mối quan hệ giữa GDP bình quân đầu người và tuổi thọ, và khám phá các xu hướng và mô hình trong quá trình phát triển của các quốc gia và khu vực khác nhau. Nhóm sẽ sử dụng các tập dữ liệu đã cho để trực quan hóa dữ liệu bằng các biểu đồ thích hợp và phát triển những hiểu biết có ý nghĩa.

c) Câu hỏi định hướng

Trong quá trình thực hiện dự án này, nhóm em sẽ tìm hiểu và thực hiện trả lời cho các câu hỏi định hướng theo hướng dẫn của đề tài như sau:

- Đối với mỗi đơn vị địa lý, có mối tương quan rõ ràng giữa GDP bình quân đầu người và tuổi thọ không?

- Đâu là đơn vị địa lý có tuổi thọ cao nhất và thấp nhất? GDP của họ thì sao?

- Những vấn đề tiềm ẩn nào khác có thể xảy ra ở các đơn vị địa lý có tuổi thọ thấp hơn?

- Nhìn chung, tuổi thọ trong thế giới hiện đại có tăng lên không? Còn GDP?

d) Phương pháp thực hiện và kết quả dự kiến

Nhóm sẽ sử dụng phương pháp phân tích dữ liệu và trực quan hóa để thực hiện dự án này. Bằng cách sử dụng các công cụ và kỹ thuật phân tích dữ liệu của pyspark, nhóm sẽ tiến hành khám phá dữ liệu, xây dựng các biểu đồ và hình dung dữ liệu để hiểu rõ hơn về mối tương quan giữa GDP và tuổi thọ cùng những thuộc tính khác.

Kết quả của nghiên cứu này sẽ cung cấp cái nhìn tổng quan về mối tương quan giữa GDP bình quân đầu người và tuổi thọ, và giúp chúng ta hiểu rõ hơn về xu hướng và mô hình phát triển trong từng quốc gia và khu vực. Các hiểu biết có ý nghĩa từ nghiên cứu sẽ hỗ trợ quyết định và chính sách liên quan đến sức khỏe và kinh tế-xã hội, và góp phần vào phát triển bền vững và cải thiện chất lượng cuộc sống của các quốc gia và khu vực trên toàn thế giới.

2.2. Mô tả tập dữ liệu

Tập dữ liệu bao gồm 2.939 bản ghi.

Tập dữ liệu từ link chứa thông tin về tuổi thọ và các yếu tố liên quan đến tuổi thọ từ nhiều quốc gia và khu vực khác nhau. Dữ liệu được thu thập bởi Tổ chức Y tế Thế giới (WHO) và bao gồm các thông tin sau:

STT	Tên trường	Kiểu dữ liệu	Ý nghĩa
1	Country	string	Tên của quốc gia hoặc khu vực
2	Year	int	Năm ghi nhận dữ liệu
3	Status	string	Trạng thái phát triển của quốc gia (Developed hoặc Developing)
4	Life expectancy	double	Tuổi thọ trung bình (tính bằng số năm)
5	Adult Mortality	int	Tỷ lệ tử vong người trưởng thành (tính bằng số lượng người trưởng thành mất mạng trên 1000 người sống)
6	Infant deaths	int	Số trẻ em chết khi còn nhỏ (dưới 1 tuổi)
7	Alcohol	double	Tiêu thụ rượu trong quốc gia (lít mỗi người trong năm)
8	Percentage expenditure	double	Tỷ lệ chi tiêu y tế trên tổng chi tiêu (phần trăm của GDP)
9	Hepatitis B	int	Tỷ lệ tiêm chủng viêm gan B ở trẻ em (dưới 1 tuổi).
10	Measles	int	Số trẻ em mắc bệnh sởi (tỷ lệ trên 1000 người sống).
11	BMI	double	Chỉ số khối cơ thể trung bình của người trưởng thành (BMI).
12	Under-five deaths	int	Số trẻ em chết dưới 5 tuổi.
13	Polio	int	Tỷ lệ tiêm chủng bại liệt ở trẻ em (dưới 5 tuổi).

14	Total expenditure	double	Tổng chi tiêu y tế (phần trăm của GDP).
15	Diphtheria	int	Tỷ lệ tiêm chủng bạch hầu ở trẻ em (dưới 5 tuổi).
16	HIV/AIDS	double	Tỷ lệ người nhiễm HIV/AIDS (tỷ lệ trong số người trưởng thành, từ 15 đến 49 tuổi).
17	GDP	double	Tổng sản phẩm quốc nội bình quân đầu người
18	Population	double	Dân số của quốc gia hoặc khu vực
19	Thinness 1-19 years	double	Tỷ lệ suy dinh dưỡng ở trẻ em từ 1 đến 19 tuổi (tỷ lệ phần trăm)
20	Thinness 5-9 years	double	Tỷ lệ suy dinh dưỡng ở trẻ em từ 5 đến 9 tuổi (tỷ lệ phần trăm)
21	Income composition of resources	double	Chỉ số tỷ lệ thu nhập từ các nguồn (các giá trị từ 0 đến 1)
22	Schooling	double	Số năm học trung bình

Tập dữ liệu này cho phép phân tích mối tương quan giữa các yếu tố trên và tuổi thọ, và cung cấp thông tin về sự phát triển và sức khỏe của các quốc gia và khu vực trên thế giới.

2.3. Công cụ thực hiện

Nhóm thực hiện triển khai mô hình trên môi trường docker. Công nghệ sử dụng: docker, hadoop, spark. IDE được sử dụng là Jupyter.

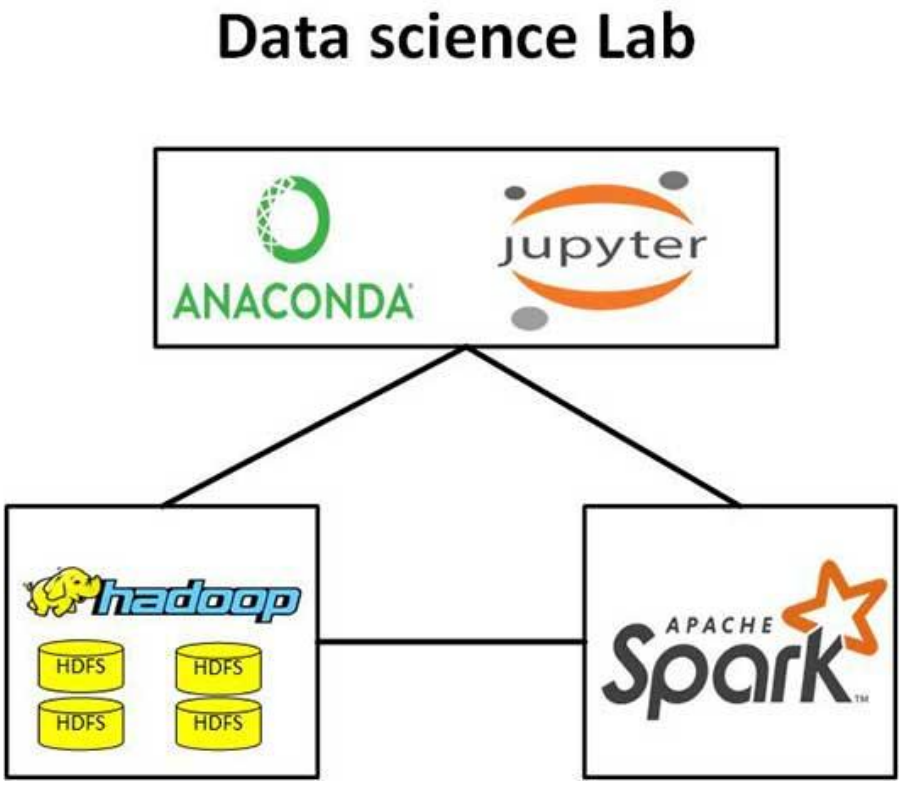
3. Hướng giải quyết bài toán và kết quả

3.1. Cài đặt cụm bigdata xử lý bài toán

Nhóm thực hiện cài đặt cụm node xử lý bài toán trên môi trường docker. Cụm hadoop bao gồm 1 namenode và 2 datanode, cụm spark gồm 1 master và 2 slave. Cụ thể, các node chính trong cụm bao gồm:

STT	Tên container	Vai trò
1	namenode	Name node của cụm hadoop
2	datanode1, datanode2	Hai datanode của cụm hadoop
3	spark-master	Master của cụm spark
4	spark-slave-1, spark-slave-2	Hai slave của cụm spark
5	Jupyter	Node chứa IDE jupyter để lập trình python

Kiến trúc của cụm:



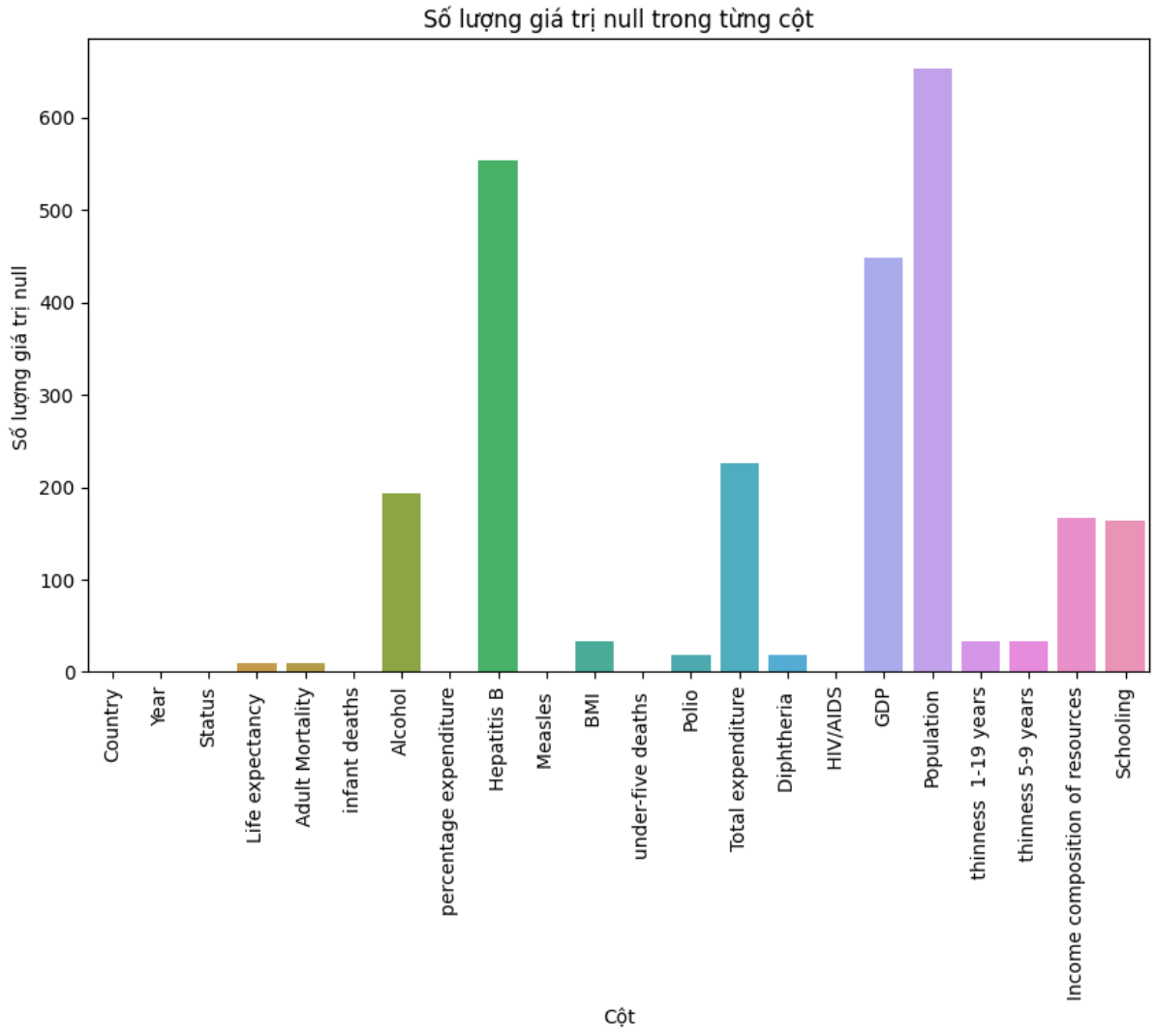
Cấu hình các thông số, port của từng node:

STT	Tên container	Cổng	Thông số cấu hình khác
1	namenode	50070, 8020	
2	datanode1, datanode2	50075, 50076	
3	spark-master	8080	
4	spark-slave-1, spark-slave-2	8081, 8082	CPU = 2 cores RAM = 4GB
5	Jupyter	8888	

3.2. Tiền xử lý dữ liệu

Dữ liệu được cung cấp là một tập dữ liệu có kích thước lớn, nhiều trường dữ liệu và nhiều kiểu dữ liệu phức tạp. Dữ liệu chưa được ở dạng tinh khiết vì còn nhiều trường dữ liệu bị thiếu dữ liệu (mang giá trị None) dẫn tới nhiều kết quả không mong muốn.

Biểu đồ số lượng giá trị bị thiếu của tập dữ liệu:



Nhóm sẽ thực hiện xử lý trường hợp này như sau:

+ Tính giá trị trung bình của từng cột.
+ Điền giá trị trung bình cột thay cho ô có giá trị None tương ứng.

Như vậy, mỗi giá trị None sẽ được gán bằng trung bình của cột hiện tại.

3.3. Phân tích mối tương quan giữa GDP và Tuổi thọ

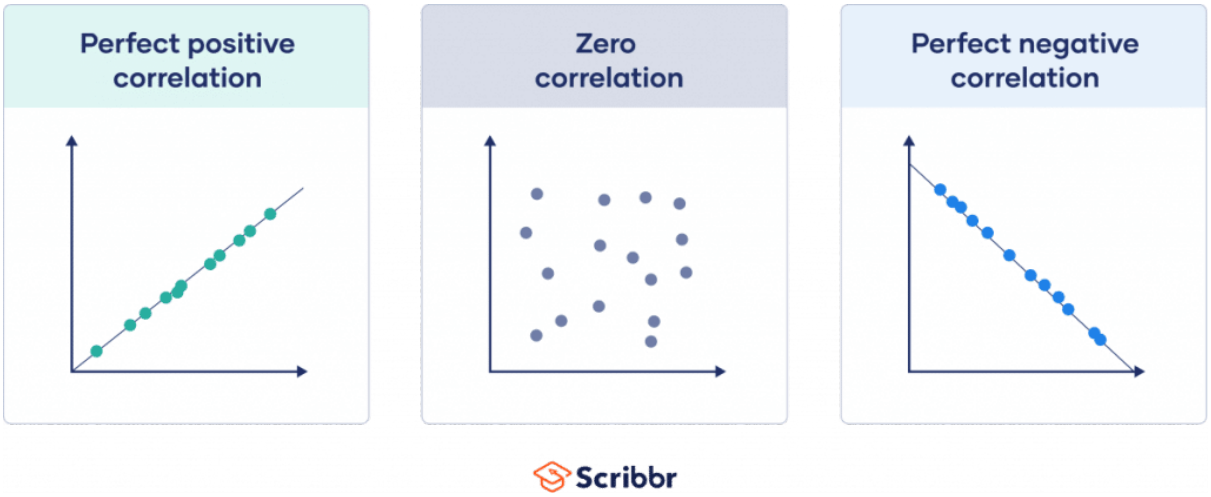
a) Hướng giải quyết

Mối tương quan giữa hai tập dữ liệu là một đo lường thống kê về mức độ tương đồng hoặc tương phản giữa chúng. Nó đo lường mức độ biến đổi đồng thời của hai tập dữ liệu theo một quy tắc nhất định.

Mối tương quan thường được biểu diễn bằng hệ số tương quan, được ký hiệu bằng "r". Công thức tính hệ số tương quan:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Giá trị của hệ số tương quan nằm trong khoảng từ -1 đến 1. Một giá trị gần -1 cho thấy mối tương quan âm mạnh, trong đó hai tập dữ liệu có xu hướng thay đổi ngược chiều (nghịch biến). Một giá trị gần 1 cho thấy mối tương quan dương mạnh, trong đó hai tập dữ liệu có xu hướng thay đổi cùng chiều (đồng biến). Một giá trị gần 0 cho thấy mối tương quan yếu hoặc không có mối tương quan tuyến tính.



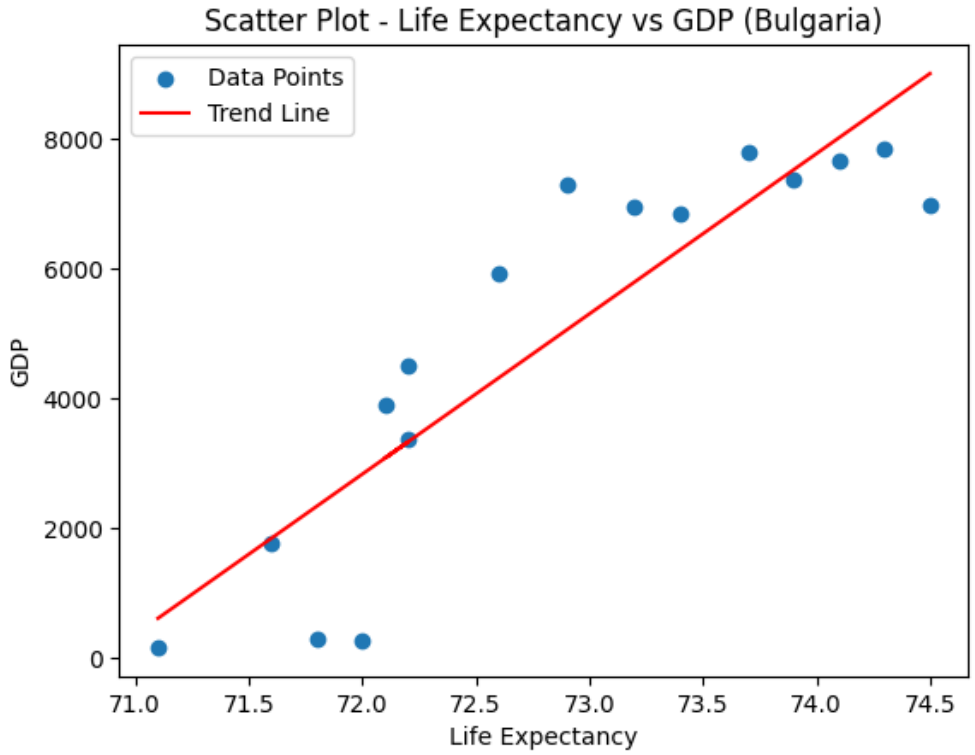
Để trả lời cho câu hỏi "*Đối với mỗi đơn vị địa lý, liệu có mối tương quan rõ ràng giữa GDP bình quân đầu người và tuổi thọ không?*", nhóm sẽ thực hiện tính toán hệ số tương quan giữa tập GDP và tập Life Expectation của mỗi quốc gia (từ năm 2000 đến năm 2015). Sau đó vẽ biểu đồ trực quan hóa kết quả tính toán được là danh sách các giá trị tương quan của các quốc gia trên thế giới.

Tiếp theo nhóm sẽ chọn ra 3 quốc gia tiêu biểu cho từng nhóm của hệ số tương quan, vẽ biểu đồ quan hệ giữa Life expectation và GDP để thấy rõ sự tương quan giữa hai tiêu chí này.

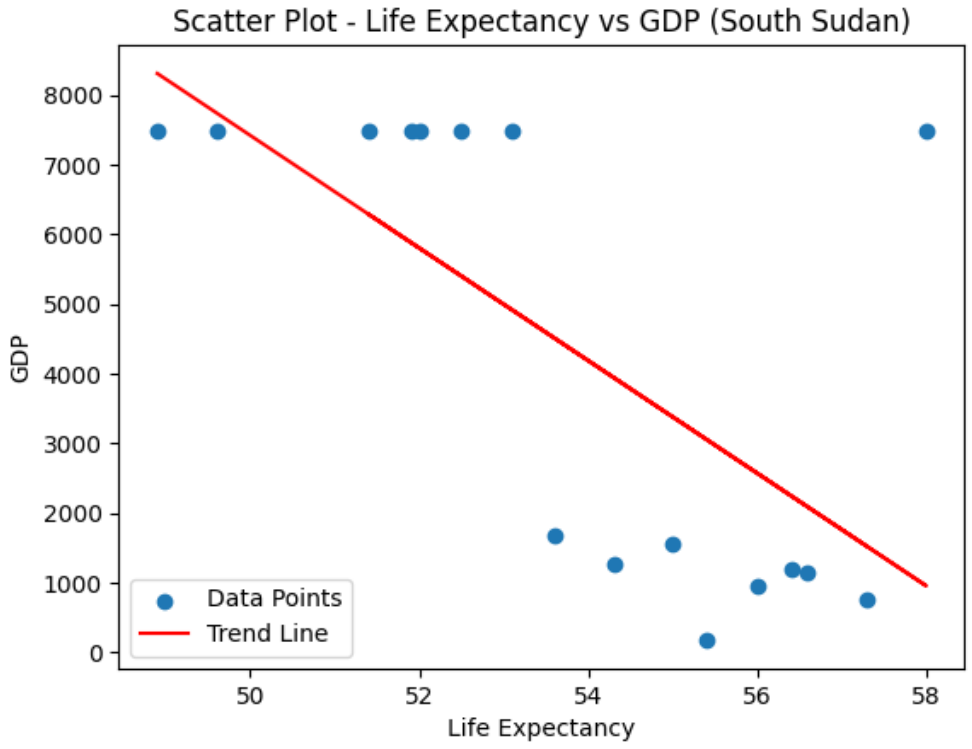
b) Phân tích kết quả và nhận xét

Dưới đây là ba quốc gia đại diện cho ba giá trị: độ tương quan dương mạnh, độ tương quan âm mạnh và ít hoặc không tương quan.

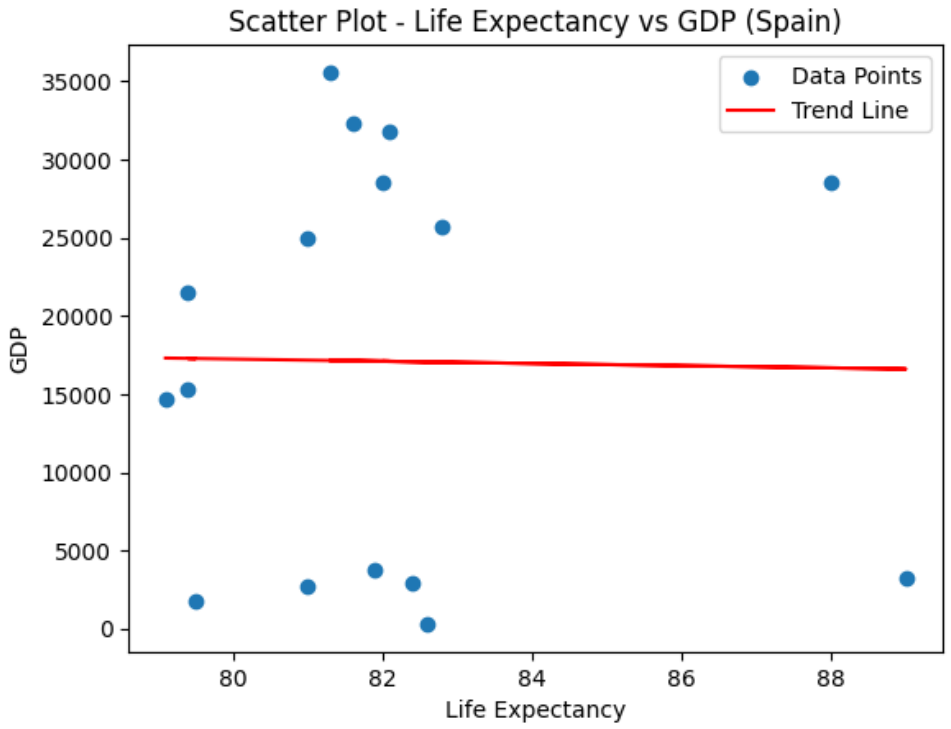
- Quốc gia có độ tương quan dương lớn: Bulgaria (0.89)



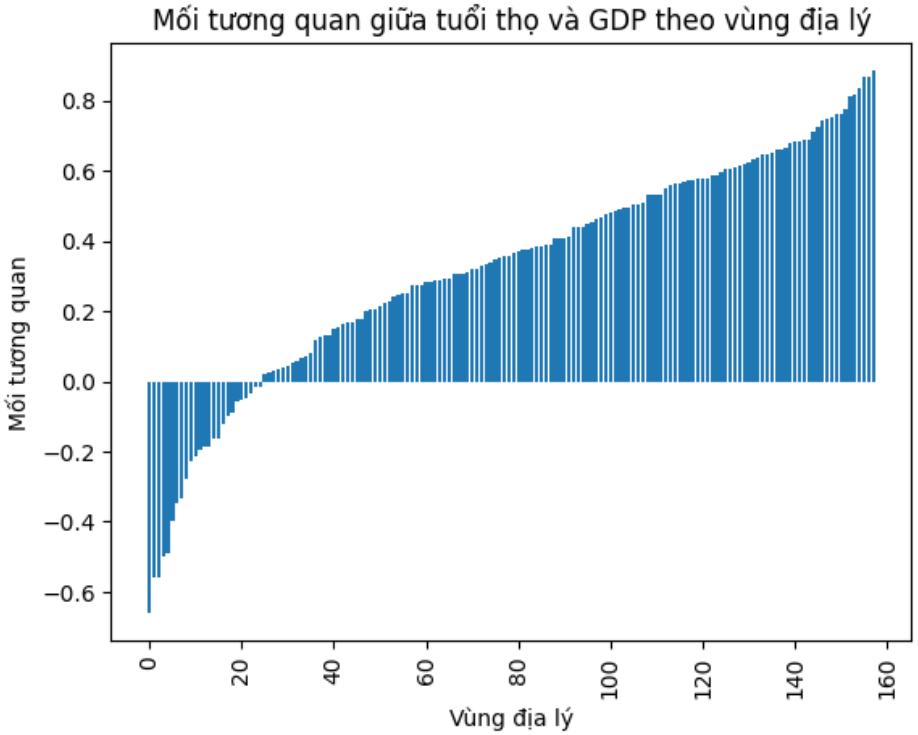
- Quốc gia có độ tương quan âm lớn: South Sudan (-0.66)



- Quốc gia gần như không cho độ tương quan giữa GDP và tuổi thọ: Tây Ban Nha (0.015)



Dưới đây là biểu đồ thể hiện giá trị của độ tương quan giữa tuổi thọ và GDP theo tất cả quốc gia trên thế giới (tính từ năm 2000 đến năm 2015).



Nhìn vào biểu đồ ta có thể thấy giá trị của hệ số tương quan giữa 2 tiêu chí này của các quốc gia trên thế giới có sự khác biệt nhau: Có những quốc

gia có độ tương quan dương khá lớn, có những quốc gia có độ tương quan xấp xỉ bằng 0 (không tương quan với nhau) và cũng có những quốc gia có độ tương quan âm.

Tuy nhiên biểu đồ cũng cho thấy phần lớn các quốc gia đều có hệ số tương quan dương (133 quốc gia trên tổng số 158 quốc gia - chiếm 84.18%). Giá trị trung bình của độ tương quan của các quốc gia trên toàn thế giới là 0.32. Điều này cho thấy phần lớn các quốc gia là có sự tương quan dương giữa GDP và tuổi thọ, nghĩa là khi GDP tăng thì tuổi thọ cũng có xu hướng tăng theo. Tuy nhiên sự tương quan này là không quá lớn, điều này cũng thể hiện rằng mối tương quan giữa GDP và tuổi thọ là có đáng kể nhưng không phải là tương quan quá mạnh, quá hoàn toàn với nhau, mà còn có những chỉ số khác có thể gây ảnh hưởng tới mối tương quan này.

Như vậy, phân tích trên đã chỉ ra rằng chỉ số GDP có ảnh hưởng một phần đáng kể tới tuổi thọ của một quốc gia trên thế giới. Vì thế để nâng cao đời sống cũng như tuổi thọ của người dân, việc phát triển kinh tế để cải thiện chất lượng sống, nâng cao chỉ số y tế giáo dục cho người dân là một biện pháp đóng vai trò rất quan trọng mà chính phủ các nước luôn tập trung hàng đầu.

3.4. Phân tích GDP của khu vực có tuổi thọ cao nhất, thấp nhất

a) Hướng giải quyết

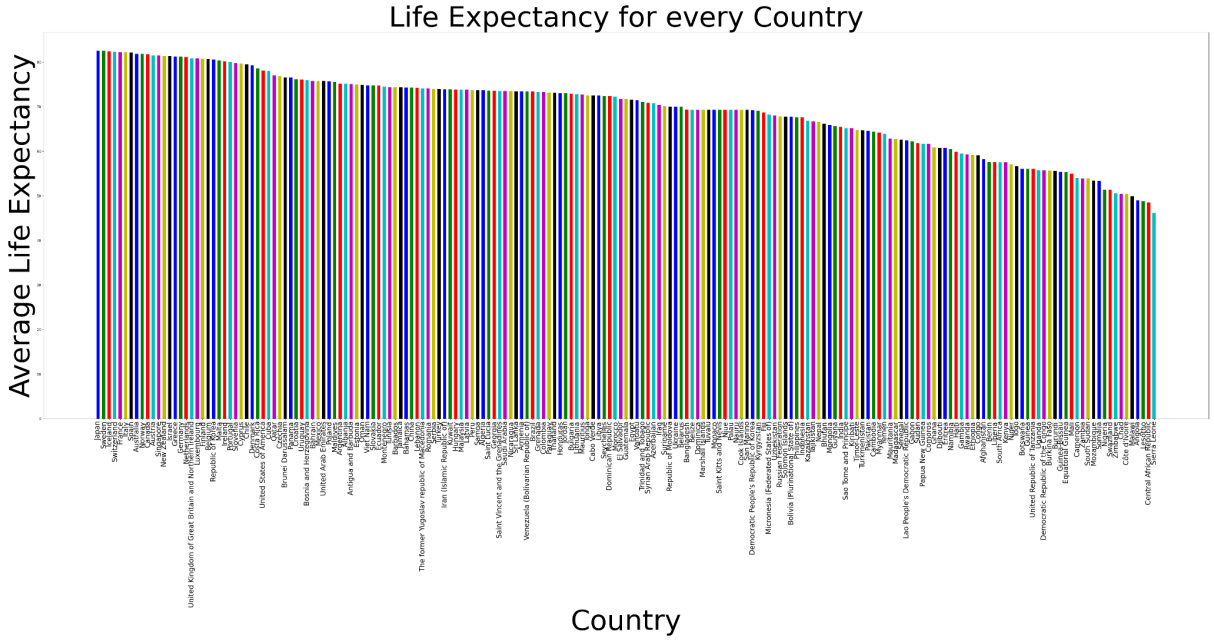
Mỗi quốc gia trong tập dữ liệu được thống kê một danh sách lịch sử các features từ năm 2000 đến năm 2015, trong đó có tuổi thọ và GDP. Để xác định xem quốc gia nào có tuổi thọ cao nhất, thấp nhất cũng như phân tích GDP của chúng, ứng với mỗi quốc gia, nhóm em sẽ tính toán giá trị trung bình về tuổi và GDP trong 15 năm lịch sử để đại diện cho quốc gia đó.

Giá trị min, max và GDP được phân tích cũng chính là giá trị trung bình được tính cho mỗi quốc gia trong lịch sử 15 năm.

Sau khi tính toán giá trị trung bình, các quốc gia được sắp xếp và lựa chọn ra top 5 các quốc gia có tuổi thọ cao nhất và top 5 các quốc gia có tuổi thọ thấp nhất. Giá trị GDP cũng được xác định cùng các quốc gia ấy để có thể dễ dàng so sánh được sự khác biệt giữa hai nhóm khu vực có tuổi thọ chênh lệch nhau.

b) Phân tích kết quả và nhận xét

Biểu đồ thể hiện tuổi thọ trung bình của tất cả các quốc gia trên thế giới từ năm 2000 đến 2015 ở dưới:



Tuổi thọ trung bình toàn thế giới từ năm 2000 đến năm 2015 là 69.22 tuổi.

Dựa vào biểu đồ, thống kê ra top 5 quốc gia có tuổi thọ trung bình cao nhất thế giới (từ năm 2000 đến năm 2015):

Country	Average Life Expectancy	Average GDP
Japan	82.5375	24892.544784375
Sweden	82.51875	29334.990639375
Iceland	82.44375000000001	30159.5029075
Switzerland	82.33125	57362.874601250005
France	82.21875	26465.551380625

Top 5 quốc gia có tuổi thọ trung bình thấp nhất thế giới (từ năm 2000 đến năm 2015):

Country	Average Life Expectancy	Average GDP
Sierra Leone	46.1125	271.50556128125004
Central African Republic	48.51250000000001	363.05590484999993
Lesotho	48.78124999999999	794.5230103125001
Angola	49.01875	1975.1430451187498
Malawi	49.893750000000004	237.50404201875

Nhật Bản là quốc gia có tuổi thọ trung bình cao nhất thế giới - 82.54 tuổi, Sierra Leone là quốc gia có tuổi thọ trung bình thấp nhất thế giới - 46.11 tuổi.

Có thể thấy sự khác biệt rất lớn giữa top 5 quốc gia có tuổi thọ trung bình cao nhất thế giới và top 5 quốc gia có tuổi thọ trung bình thấp nhất thế giới. Phần lớn các quốc gia có tuổi thọ cao trên thế giới là Nhật Bản và các quốc gia châu Âu như Thụy Điển, Iceland, Thụy Sĩ, Pháp, … Đây đều là các quốc gia phát triển, có GDP cao top đầu thế giới. Còn các quốc gia có tuổi thọ trung bình ở mức thấp nhất thế giới đều là các quốc gia đến từ châu Phi - một châu lục khắc nghiệt với trình độ phát triển kinh tế ở mức kém, đi kèm với chỉ số GDP thấp so với thế giới.

So sánh giữa Nhật Bản và Sierra Leone có thể thấy GDP của Nhật Bản cao gấp khoảng 100 lần so với GDP của Sierra Leone.

Tóm lại có sự chênh lệch lớn giữa trình độ phát triển kinh tế giữa các nhóm quốc gia có tuổi thọ cao và các nhóm quốc gia có tuổi thọ thấp. Trong khi các quốc gia có tuổi thọ cao là các quốc gia có nền kinh tế phát triển hàng đầu thế giới với GDP rất cao thì các quốc gia có tuổi thọ thấp lại là các quốc gia có nền kinh tế kém phát triển và GDP cũng ở mức thấp.

3.5. Phân tích các vấn đề tiềm ẩn ở những khu vực có tuổi thọ thấp

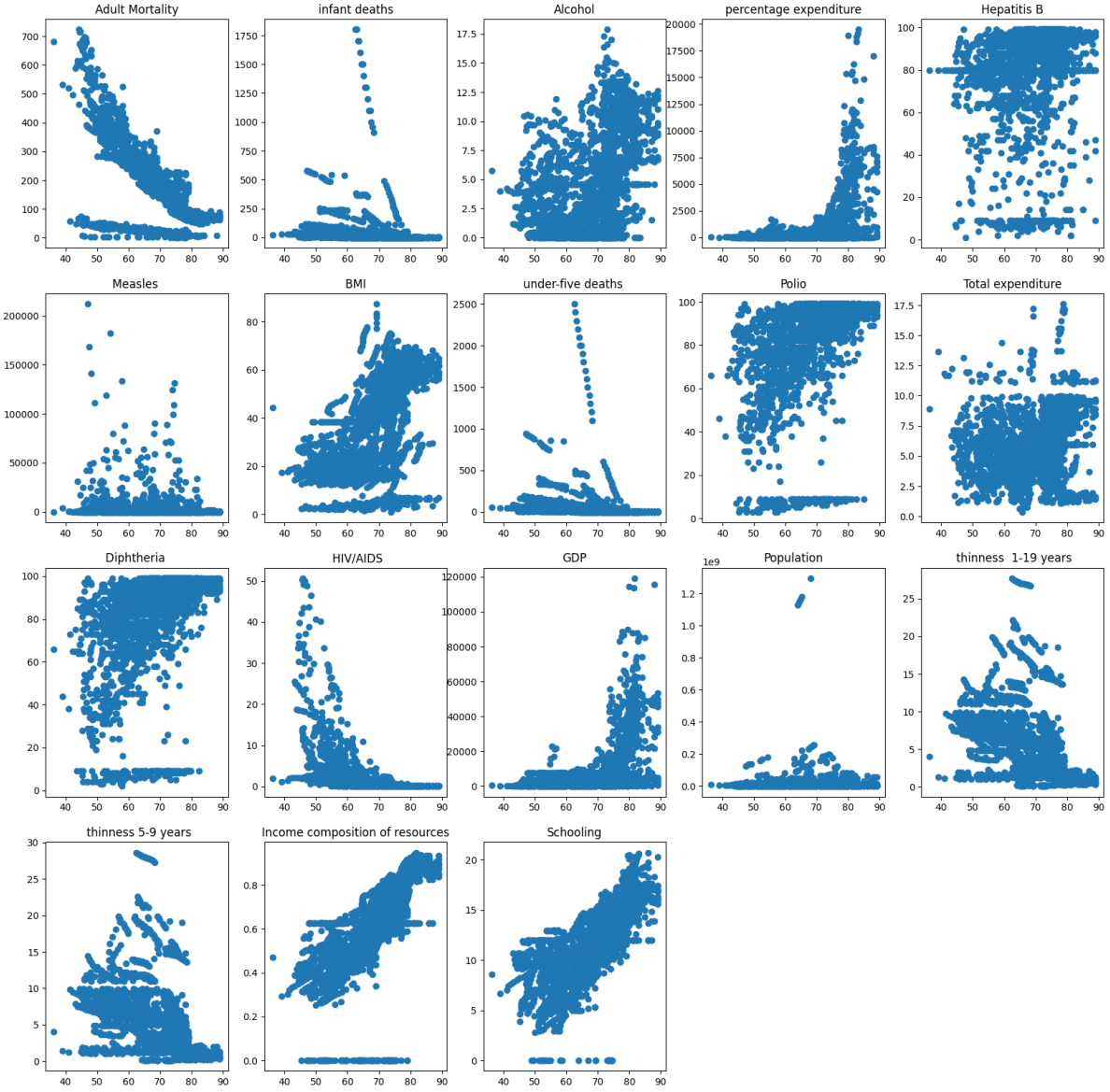
a) Hướng giải quyết

Để trả lời cho câu hỏi *"Những vấn đề tiềm ẩn nào khác có thể xảy ra ở các đơn vị địa lý có tuổi thọ thấp hơn?"*, nhóm sẽ thực hiện trực quan hóa các biểu đồ thể hiện mối quan hệ tương quan giữa tuổi thọ và các trường dữ liệu khác ở dạng biểu đồ Scatter.

Sau đó, nhóm sẽ lựa chọn ra các biểu đồ cho thấy có sự tương quan mạnh mẽ và rõ ràng đối với tuổi thọ. Sử dụng các biểu đồ này để phân tích các giá trị của dữ liệu khi giá trị của tuổi thọ ở mức thấp hơn. Cuối cùng là rút ra những nhận xét và kết luận cho các vấn đề tiềm ẩn có thể xảy ra ở các khu vực, quốc gia có tuổi thọ thấp.

b) Phân tích kết quả và nhận xét

Kết quả xây dựng biểu đồ Scatter thể hiện mối tương quan giữa trường tuổi thọ và các trường dữ liệu khác của tập dữ liệu:

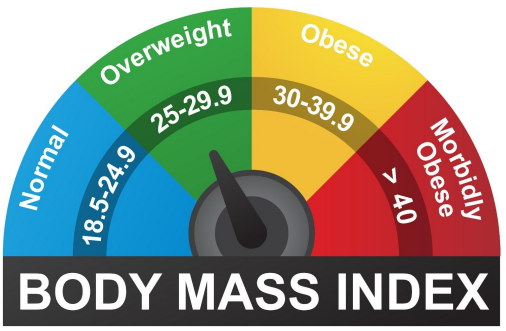


Dựa vào 18 biểu đồ tương quan giữa tuổi và các trường dữ liệu khác, ta lựa chọn ra một số trường dữ liệu có độ tương quan khá rõ ràng với tuổi thọ, đó là: Income composition of resources (Chỉ số phân phối thu nhập từ các nguồn tài nguyên - tương hỗ dương), Adult Mortality (tỷ lệ tử vong ở người trưởng thành - tương hỗ âm), BMI (chỉ số khối cơ thể trung bình - tương hỗ dương), Schooling (Số năm học trung bình - tương hỗ dương). Sau đây ta sẽ phân tích để hiểu rõ hơn 4 chỉ số này:

Chỉ số "Income composition of resources" (ICR) là một chỉ số đo lường mức độ phân phối thu nhập từ các nguồn tài nguyên trong một quốc gia hoặc khu vực cụ thể. Chỉ số này thường được sử dụng để đánh giá mức độ phát triển kinh tế và mức sống của một quốc gia. ICR đo lường tỷ lệ phần trăm của thu nhập đến từ các nguồn tài nguyên khác nhau, chẳng hạn như thu nhập lao động, thu nhập từ vốn đầu tư, thu nhập từ tài sản, và các nguồn thu nhập khác. Chỉ số ICR có thể giúp ta hiểu được cách mà nguồn thu nhập được phân phối trong xã hội và mức độ chia sẻ lợi ích từ sự phát triển kinh tế. Một ICR cao cho thấy sự phân phối thu nhập tương đối công bằng và mức sống cao hơn trong quốc gia đó, trong khi một ICR thấp có thể chỉ ra một sự chênh lệch lớn về thu nhập và mức sống giữa các tầng lớp xã hội. ICR là một trong những chỉ số kinh tế quan trọng trong việc đánh giá sự phát triển và mức độ bình đẳng của một quốc gia.

Chỉ số Adult Mortality thể hiện xác suất để một người 15 tuổi chết trước khi tròn 60 tuổi. Chỉ số được tính bằng tỉ lệ tử vong trong độ tuổi từ 15 đến 60 tuổi trên 1.000 dân số mỗi năm. Tỉ lệ này càng cao thì càng biểu hiện một xã hội nguy hiểm và bất ổn của quốc gia đang xét.

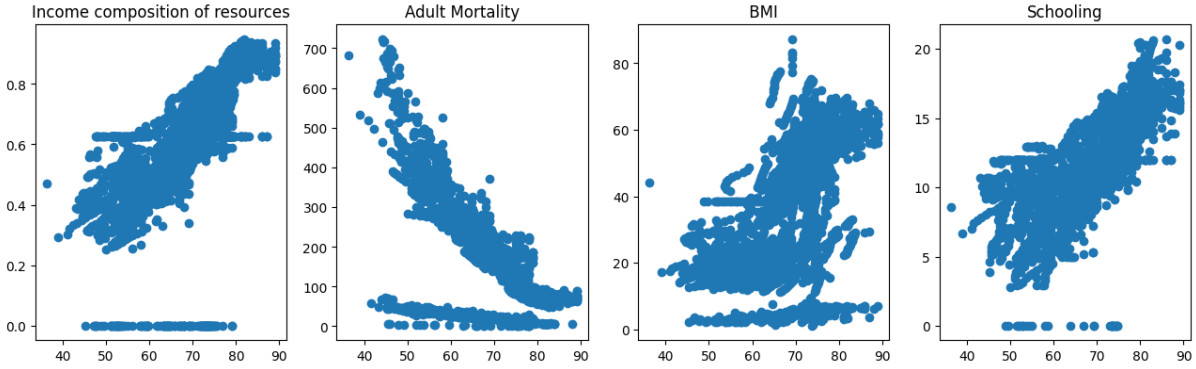
Chỉ số khối cơ thể, cũng gọi là chỉ số thể trọng, viết tắt là BMI theo tên tiếng Anh body mass index, là một cách nhận định cơ thể của một người là gầy hay béo bằng một chỉ số. Chỉ số này do nhà khoa học người Bỉ Adolphe Quetelet đưa ra năm 1832. Chỉ số



khối cơ thể của một người tính bằng trọng lượng (kg) chia cho bình phương chiều cao (m). Chỉ số này có thể giúp xác định một người bị béo phì hay bị suy dinh dưỡng một cách khoa học căn cứ trên số liệu về hình dáng, chiều cao và cân nặng cơ thể. BMI < 18.5 thể hiện một người đang bị thiếu cân, BMI từ 18.5 - 22.9 thể hiện một thể trạng bình thường, BMI từ 23.0 đến 24.9 thể hiện một người có dấu hiệu thừa cân, BMI từ 25.0 trở lên thể hiện một người đang có thể trạng là béo.

Chỉ số Schooling tính bởi số năm học trung bình của dân số một quốc gia. Chỉ số này thể hiện trình độ dân trí cũng như điều kiện kinh tế xã hội của quốc gia đó. Chỉ số Schooling càng nhỏ cho thấy quốc gia thiếu thốn về điều kiện kinh tế xã hội khiến người dân không có đầy đủ cơ hội và điều kiện để đi học.

Sau đây ta sẽ tập trung phân tích 4 biểu đồ tương ứng với 4 chỉ số:



Nhìn 4 biểu đồ và qua việc phân tích các chỉ số ở trên, ta rút ra được một số nhận xét. Ở nhóm các quốc gia có tuổi thọ thấp (đi về phía bên trái của mỗi biểu đồ đang xét), thì chỉ số Adult Mortality có xu hướng tăng lên còn các chỉ số BMI, Income composition of resources, Schooling lại có xu hướng giảm đi. Điều đó giúp ta đưa ra nhận xét: ở nhóm các quốc gia có tuổi thọ thấp có một số vấn đề tiềm ẩn có nguy cơ xảy ra như sau:

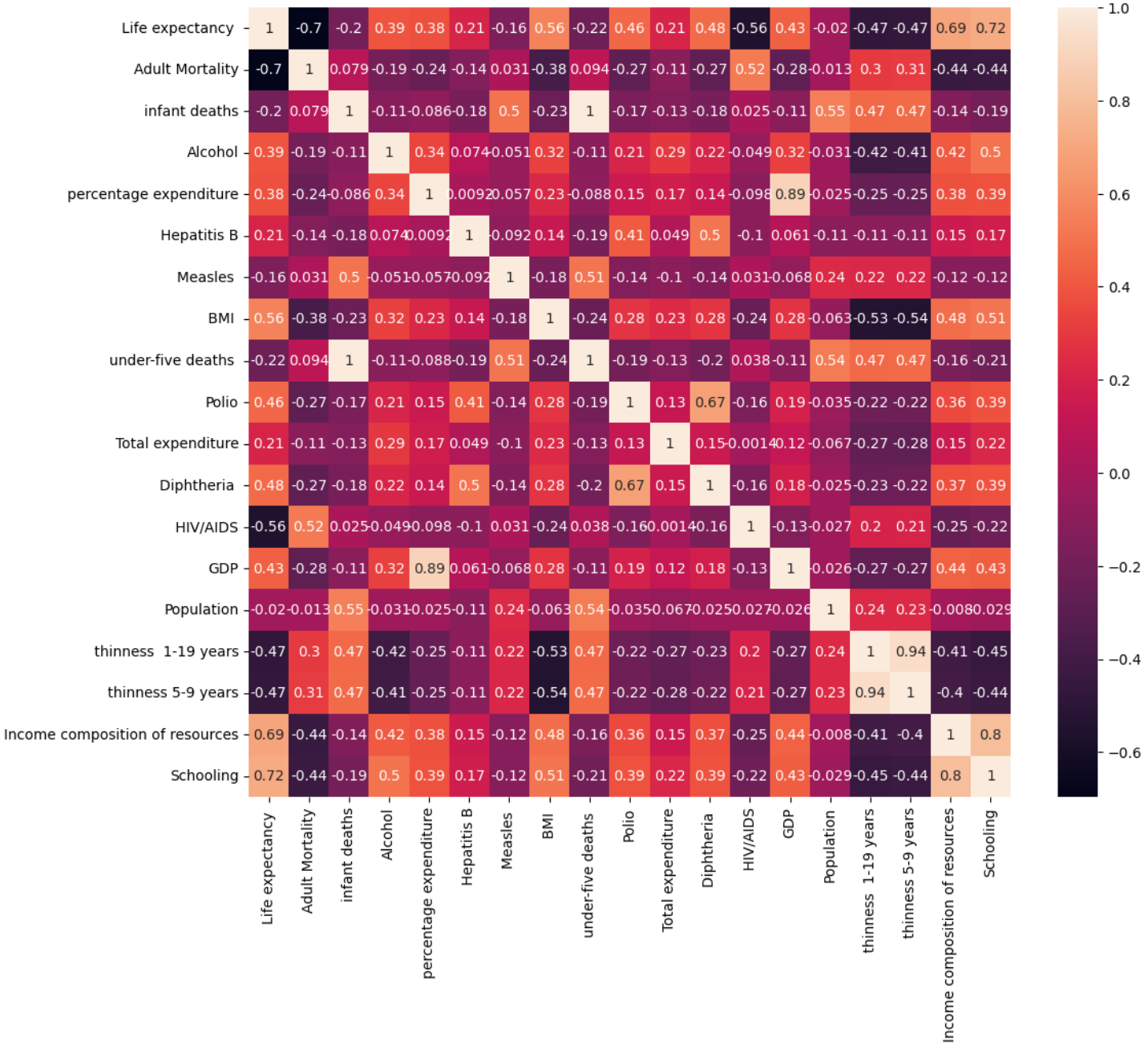
1. Tỷ lệ tử vong ở người trưởng thành là cao hơn so với nhóm quốc gia có tuổi thọ cao. Nhóm quốc gia này có thể có những bất ổn về tình hình chính trị, y tế gây ảnh hưởng tới sức khỏe và tính mạng của người dân.

2. Chỉ số BMI, Income composition of resources và Schooling có xu hướng giảm đi. Người dân của nhóm các quốc gia này có thể trạng không tốt bằng người dân ở các quốc gia có tuổi thọ cao, thể hiện qua tỷ lệ người gầy, thiếu cân là cao hơn. Mức độ phát triển kinh tế và mức sống của nhóm quốc gia này cũng là không cao, phân phối thu nhập của các quốc gia này chưa công bằng và có sự chênh lệch lớn về thu nhập và mức sống giữa các tầng lớp xã hội. Nhóm quốc gia này cũng phần nào cho thấy sự thiếu thốn về điều kiện kinh tế xã hội khiến người dân không có đầy đủ cơ hội và điều kiện để đi học dẫn tới trình độ dân trí và phát triển con người cũng là không cao.

3.6. Phân tích xu hướng GDP và tuổi thọ thế giới

a) Hướng giải quyết

Nhóm thiết lập ma trận tương quan giữa hai trường dữ liệu bất kỳ xuất hiện trong tập dữ liệu. Thu được biểu đồ ma trận độ tương quan như hình dưới đây:



Đối với xu hướng phát triển tuổi thọ, nhóm sẽ thực hiện huấn luyện mô hình học máy để đưa ra dự đoán về tuổi thọ trung bình của thế giới trong năm tiếp theo: 2016. Để thực hiện huấn luyện được mô hình này, nhóm sẽ thiết lập ra một tập dữ liệu dùng cho bài toán học có giám sát. Tập dữ liệu này cần chứa những trường dữ liệu có độ tương quan là lớn đối với tuổi thọ để giúp tăng độ chính xác cho mô hình.

Dựa vào ma trận tương quan ở trên, nhóm sẽ chọn các trường làm dữ liệu huấn luyện cho mô hình học máy dự đoán tuổi thọ trung bình thế giới bao gồm: Adult Mortality, BMI, Income composition of resources, Schooling, Year. Mô hình được lựa chọn để huấn luyện là: Gradient Boosted Trees Regression

Đối với xu hướng phát triển của GDP bình quân thế giới, nhóm sẽ vẽ biểu đồ đường biểu diễn GDP từ năm 2000 đến năm 2015 rồi phân tích biểu đồ để rút ra nhận xét về xu hướng GDP.

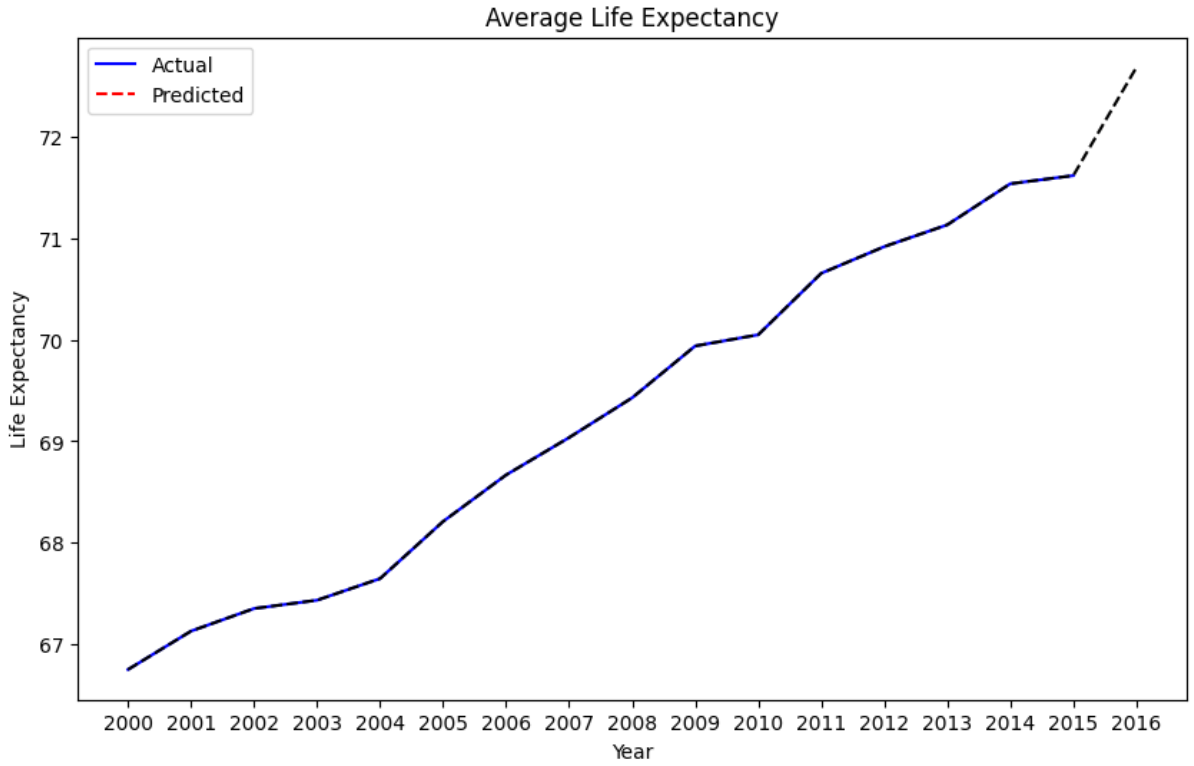
b) Phân tích kết quả và nhận xét

Huấn luyện trên mô hình Gradient Boosted Trees Regression cho dự đoán tuổi thọ cho độ chính xác trên tập test là:

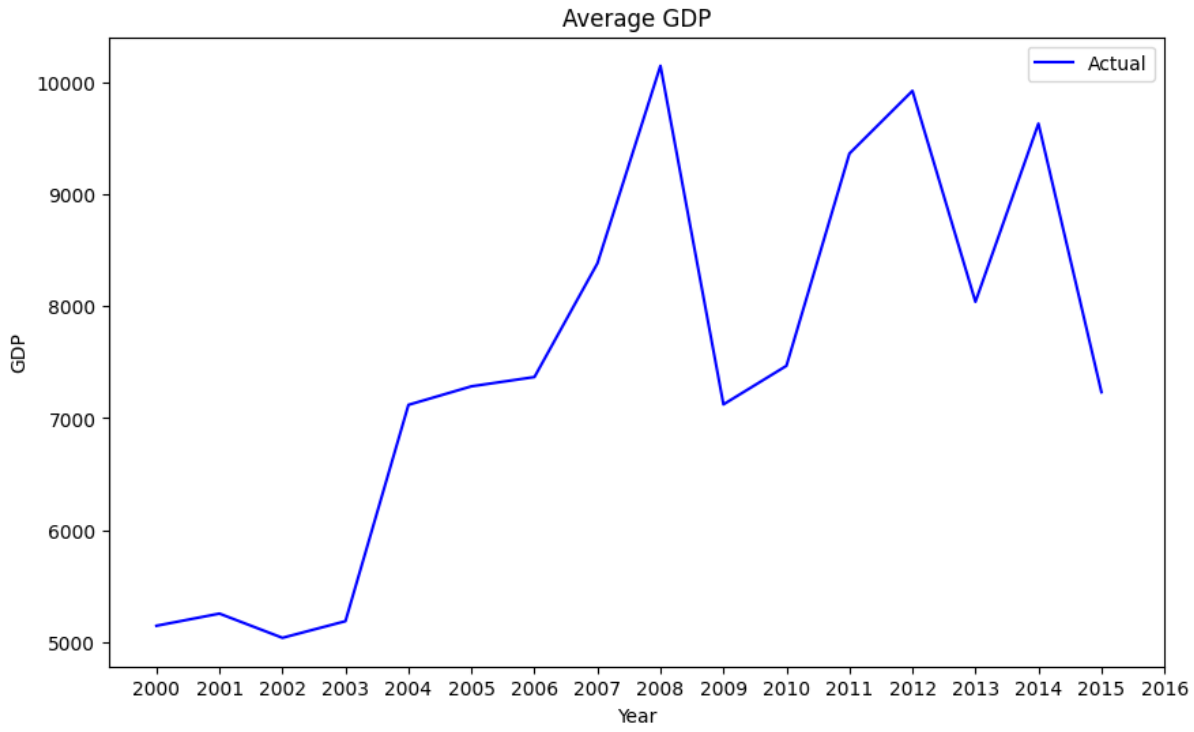
RMSE: 2.5461731497163553

Adult Mortality	BMI	Income composition of resources	Schooling	Year	prediction
152.86338797814207	42.70569668367733	0.6882322986118916	12.876108895781035	2016	72.68131784604888

Dự đoán tuổi thọ trung bình thế giới năm 2016 cho kết quả: 72.68 tuổi với đồ thị trực quan hóa độ tuổi trung bình toàn thế giới như biểu đồ dưới:.



Biểu đồ giá trị trung bình GDP thế giới từ năm 2000 đến năm 2015:



Dựa theo 2 đồ thị trên, ta rút ra nhận xét:

+ Tuổi thọ trung bình của thế giới có xu hướng tăng tu mạnh qua các năm

+ GDP trung bình của thế giới có xu hướng tăng nhưng không ổn định

5. Kết luận

Dựa trên phân tích và khám phá dữ liệu tuổi thọ, nhóm em đã rút ra các kết luận quan trọng về mối tương quan giữa tuổi thọ và các yếu tố khác, cũng như xu hướng chung của tuổi thọ và GDP trong thế giới hiện đại như sau:

- Có mối tương quan không quá mạnh giữa GDP bình quân đầu người và tuổi thọ. Các đơn vị địa lý có GDP cao hơn thường có tuổi thọ cao hơn, trong khi các đơn vị địa lý có GDP thấp hơn có xu hướng có tuổi thọ thấp hơn. Điều này chỉ ra rằng mức độ phát triển kinh tế có tác động mạnh mẽ đến tuổi thọ trong một quốc gia hoặc khu vực.

- Nhóm em đã xác định được các quốc gia có tuổi thọ cao nhất và GDP tương ứng. Điều này có thể cung cấp thông tin quan trọng về các thành công và yếu tố đóng góp vào tuổi thọ cao. Ngược lại, nhóm em cũng đã xác định được nhóm các quốc gia có tuổi thọ thấp nhất và GDP tương ứng. Điều này có thể chỉ ra các vấn đề tiềm ẩn khác mà các đơn vị địa lý này đang phải đối mặt, bao gồm các vấn đề về y tế, môi trường, kinh tế và xã hội. Nhóm em cũng đã phân tích và so sánh được sự đối lập giữa hai nhóm quốc gia này.

- Từ quan sát tổng thể, nhóm em nhận thấy xu hướng tăng lên của tuổi thọ và GDP trong thế giới hiện đại. Điều này có thể được liên kết với sự phát triển y tế, cải thiện điều kiện sống và những tiến bộ trong lĩnh vực y học. Tuy nhiên, cần phải đánh giá kỹ hơn và có những nhận định chính xác hơn của các chuyên gia cũng như làm mới, cập nhật tập dữ liệu để hiểu rõ hơn về các yếu tố ảnh hưởng và xu hướng này.

Tóm lại, dự án phân tích tuổi thọ đã mang lại những hiểu biết quan trọng về mối tương quan giữa tuổi thọ và GDP, xác định các đơn vị địa lý có tuổi thọ cao nhất và thấp nhất, cùng nhận thức về những vấn đề tiềm ẩn có thể xảy ra ở các đơn vị địa lý có tuổi thọ thấp hơn, và nhìn chung, nhận thấy sự gia tăng tuổi thọ trong thế giới hiện đại. Những kết luận này cung cấp cơ sở cho việc đưa ra quyết định và hướng đi trong các chính sách y tế, kinh tế và xã hội nhằm cải thiện tuổi thọ và chất lượng cuộc sống của cộng đồng.

6. Link mã nguồn

Nhóm sẽ sử dụng google drive làm kho lưu trữ mã nguồn. Link mã nguồn tham khảo của nhóm ở đây:

https://drive.google.com/drive/folders/1hd9dAt9a3NEBrfOfALcIdAun4S2KFR1G?usp=sharing

TÀI LIỆU THAM KHẢO

[1]. Hỗ trợ cài đặt cụm big data, Jupyter Apache Spark Standalone Cluster avec Docker.
[2]. Tập dữ liệu phân tích, Life Expectancy (WHO).