**TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG**



**BÁO CÁO**

**DATA PPREPROCESSING**

**Môn: Phân tích dữ liệu bằng Python**

**NHÓM 6**

*Lớp học phần: MIS3041\_47K21.1*

*SĐT nhóm trưởng: 0965946401*

*Giảng viên hướng dẫn: TS. Lê Diên Tuấn*

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **Phần trăm đóng góp** |
| 1 | Đặng Phương Thảo | 100% |
| 2 | Trần Ngọc Huy | 100% |
| 3 | Hoàng Thị Thanh Nhàn | 100% |
| 4 | Nguyễn Thị Thu Hà | 100% |
| 5 | Nguyễn Thị Yến Hương | 100% |

*Đà Nẵng, ngày 6 tháng 4 năm 2024.*

**MỤC LỤC**

[**1. Giới thiệu tổng quan 1**](#_Toc163128275)

[**2. Cơ sở lý thuyết 2**](#_Toc163128276)

[**2.1. Thống kê mô tả 2**](#_Toc163128277)

[**2.1.1. Count 2**](#_Toc163128278)

[**2.1.2. Mean 2**](#_Toc163128279)

[**2.1.3. Độ lệch chuẩn 2**](#_Toc163128280)

[**2.1.4. Min 2**](#_Toc163128281)

[**2.1.5. Phân vị thứ nhất 2**](#_Toc163128282)

[**2.1.6. Phân vị thứ hai 2**](#_Toc163128283)

[**2.1.7. Phân vị thứ ba 2**](#_Toc163128284)

[**2.1.8. Max 2**](#_Toc163128285)

[**2.2. Làm sạch dữ liệu 3**](#_Toc163128286)

[**2.2.1. Xử lý dữ liệu thiếu (Missing Data) 3**](#_Toc163128287)

[**2.2.2. Nhận diện phần tử biên (Outliers) 3**](#_Toc163128288)

[**2.2.3. Giảm thiểu dữ liệu nhiễu (Noisy Data) 3**](#_Toc163128289)

[**2.2.4. Xử lý dữ liệu không nhất quán (Inconsistent Data) 3**](#_Toc163128290)

[**2.2.5. Giải pháp 3**](#_Toc163128291)

[**2.3. Biến đổi dữ liệu 5**](#_Toc163128292)

[**2.3.1. Chuẩn hóa (normalization) 6**](#_Toc163128293)

[**2.3.2. Xây dựng thêm thuộc tính 6**](#_Toc163128294)

[**2.4. Rời rạc hóa dữ liệu 6**](#_Toc163128295)

[**2.4.1. Encoding 7**](#_Toc163128296)

[**2.4.2. Binning 7**](#_Toc163128297)

[**2.4.3. Cluster Analysis 7**](#_Toc163128298)

[**3. Kết quả thực nghiệm 7**](#_Toc163128299)

[**3.1. Giới thiệu dataset 7**](#_Toc163128300)

[**3.2. Thực hành trên dataset 9**](#_Toc163128301)

[**3.2.1. Mô tả dataset 9**](#_Toc163128302)

[**3.2.2. Làm sạch dữ liệu 16**](#_Toc163128303)

[**3.2.3. Rời rạc hóa dữ liệu 31**](#_Toc163128304)

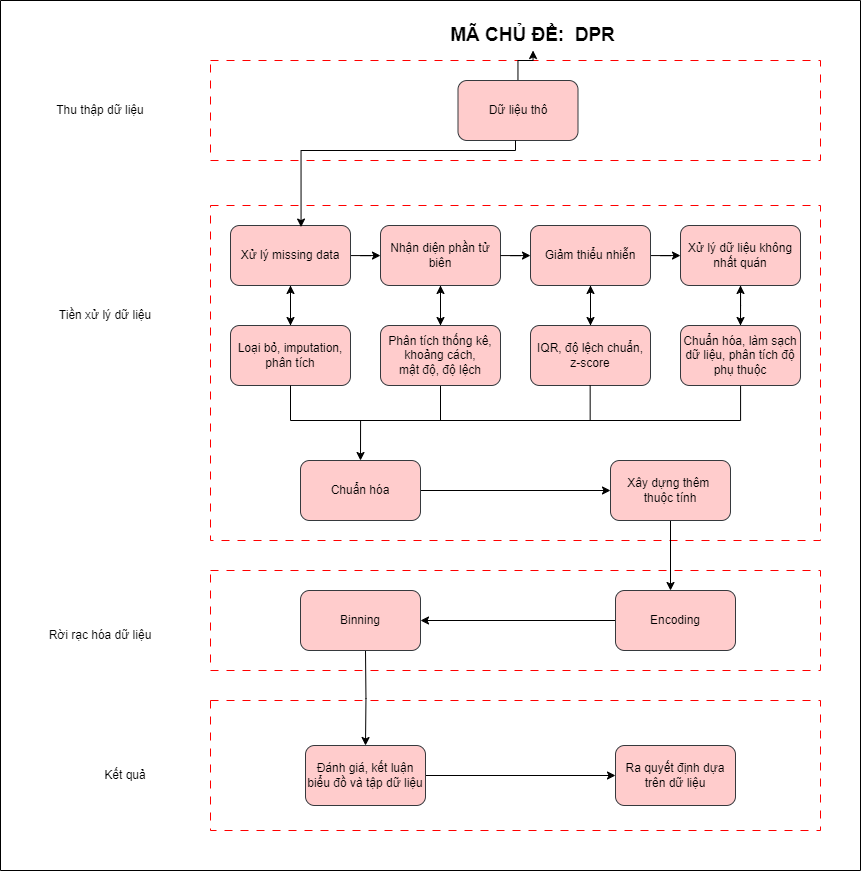
[**4. Nhận xét tổng quan về bài nghiên cứu 33**](#_Toc163128305)

[**5. Nguồn tài liệu tham khảo 35**](#_Toc163128306)

# Giới thiệu tổng quan

Trong mọi dự án phân tích dữ liệu, việc nắm bắt và xử lý dữ liệu đóng vai trò không thể phủ nhận. Đây là yếu tố đảm bảo chất lượng và hiệu quả của kết quả nghiên cứu. Dữ liệu ban đầu thường không được hoàn hảo và yêu cầu trải qua các giai đoạn làm sạch và biến đổi để phát triển mô hình học máy và phân tích thống kê. Bước này không chỉ giúp loại bỏ nhiễu và sự không nhất quán, mà còn phát hiện ra các mẫu hình và xu hướng quan trọng. Nghiên cứu này tập trung vào việc áp dụng các phương pháp thống kê mô tả và kỹ thuật làm sạch dữ liệu cơ bản, nhấn mạnh tầm quan trọng của tiền xử lý dữ liệu trong nhiều lĩnh vực, không chỉ trong lĩnh vực khoa học dữ liệu.

***Work flow***



# Cơ sở lý thuyết

## **Thống kê mô tả**

### **Count**

Count trong thống kê mô tả đề cập đến việc đếm số lượng của các mục không rỗng hoặc là không null trong mỗi cột dữ liệu. Tính count sẽ cung cấp một cái nhìn sơ bộ về kích thước, độ đầy đủ của dữ liệu. Theo đó có thể phát hiện được các cột có thể có số lượng giá trị thiếu.

### **Mean**

Mean được tính bằng cách cộng toàn bộ giá trị ở trong cột và chia cho tổng số lượng các giá trị đó. Mean hay giá trị trung bình là đại diện tính toán cho xu hướng trung tâm của một tập dữ liệu nhất định. Trong phân tích dữ liệu thường được dùng với mục đích mô tả đặc điểm tổng quát của dữ liệu.

### **Độ lệch chuẩn**

Độ lệch chuẩn (Standard deviation- Std) là thước đo độ phân tán của các giá trị trong một tập dữ liệu đã cho từ giá trị trung bình của chúng. Nó cho biết trung bình mỗi giá trị nằm bao xa so với giá trị trung bình.

### **Min**

Min là giá trị nhỏ nhất trong mỗi cột dữ liệu. Min là thông tin quan trọng bởi vì nó cho thấy được giới hạn dưới của dữ liệu, từ đó xác định được giá trị cực trị hoặc là outliers.

### **Phân vị thứ nhất**

Điểm tứ phân vị thứ nhất (Q1): Đây là điểm chia tách dữ liệu thành hai phần bằng nhau, trong đó 25% dữ liệu thấp hơn Q1 và 75% dữ liệu cao hơn Q1.

### **Phân vị thứ hai**

Điểm tứ phân vị thứ hai (Q2): Điểm tứ phân vị thứ hai tương đương với giá trị trung vị của tập dữ liệu, tức là 50% dữ liệu thấp hơn Q2 và 50% dữ liệu cao hơn Q2. Q2 cũng thường được gọi là median (trung vị).

### **Phân vị thứ ba**

Điểm tứ phân vị thứ ba (Q3): Đây là điểm chia tách dữ liệu thành hai phần bằng nhau, trong đó 75% dữ liệu thấp hơn Q3 và 25% dữ liệu cao hơn Q3.

### **Max**

Max là giá trị lớn nhất trong mỗi cột dữ liệu. Max là thông tin quan trọng để giúp hiểu rõ hơn về giới hạn trên của dữ liệu, đồng thời giúp phát hiện giá trị cực trị hoặc outliers.

## **Làm sạch dữ liệu**

### **Xử lý dữ liệu thiếu (Missing Data)**

Dữ liệu bị thiếu là dữ liệu được hiển thị như NaN, Nat, Null, N/A, v.v. xuất hiện với nhiều nguyên nhân như làm dữ liệu bị mất trong quá trình chuyển đổi,...

Thông thường, có 2 cách để xử lý missing values:

* Loại bỏ missing values
* Thay thế missing values bằng một giá trị khác
* Biến có missing values là biến số - numeric: Có thể thay thế bằng những giá trị như: 0, median, mean, v.v
* Biến có missing values là biến categorical: Có thể nhóm những trường hợp missing values vào 1 nhóm, đặt tên là Missing...)

### **Nhận diện phần tử biên (Outliers)**

Phần tử biên (Outliers) là những dữ liệu (đối tượng) không tuân theo đặc tính hành vi chung của dữ liệu (đối tượng). Giải pháp nhận diện phần tử biên dựa trên khoảng cách, mật độ, độ lệch, phân bố thống kê.

### **Giảm thiểu dữ liệu nhiễu (Noisy Data)**

Dữ liệu nhiễu (Noisy Data) là các dữ liệu không chính chính xác hoặc không thể giải thích được nên có thể làm méo mó các mẫu xu hướng, ảnh hưởng đến tính khách quan của quá trình phân tích. Nó có thể nhận diện thông qua phân tích thống kê, biểu đồ hộp hoặc trực quan hóa dữ liệu.

Để giảm thiểu dữ liệu nhiễu thường đòi hỏi áp dụng các phương pháp làm mịn. Ví dụ như bộ lọc trung bình, bộ lọc trung vị,...

### **Xử lý dữ liệu không nhất quán (Inconsistent Data)**

Dữ liệu không nhất quán có thể xuất hiện khi cách ghi chép dữ liệu không thống nhất hoặc là mã hóa dữ liệu. Quá trình xử lý dữ liệu không nhất quán thường yêu cầu sự phát hiện, sửa chữa những mâu thuẫn hoặc kết hợp dữ liệu từ nhiều nguồn khác nhau để đạt được sự nhất quán.

### **Giải pháp**

#### **Giải pháp xử lý dữ liệu thiếu (Missing Data)**

Dữ liệu bị thiếu là dữ liệu không đầy đủ, không thể thực hiện phân tích dữ liệu được. Các giải pháp xử lý dữ liệu thiếu bao gồm:

* **Loại bỏ:** Một cách đơn giản nhất là loại bỏ những giá trị thiếu nếu chúng không quan trọng hoặc chiếm tỷ lệ nhỏ trong tập dữ liệu. Nếu số lượng giá trị thiếu ít hơn 3% tổng số quan sát của một biến cụ thể, chúng ta có thể loại bỏ các hàng hoặc cột chứa dữ liệu thiếu mà không ảnh hưởng đáng kể đến kết quả phân tích.
* **Imputation:** Điền vào dữ liệu thiếu bằng các giá trị thay thế như trung bình, trung vị, mode của biến là một cách tiếp cận phổ biến. Có thể sử dụng các phương pháp phức tạp hơn như imputation đa biến để dự đoán giá trị thay thế dựa trên các biến khác, giúp giữ lại càng nhiều thông tin càng tốt trong tập dữ liệu.
* **Phân tích:** Đánh giá tác động của dữ liệu thiếu đối với kết quả phân tích là một bước quan trọng để hiểu rõ hơn về cách thức và mức độ ảnh hưởng của nó. Việc này giúp ta xác định xem liệu dữ liệu thiếu có ảnh hưởng đến kết quả phân tích và đưa ra quyết định phù hợp về cách xử lý dữ liệu thiếu.

#### **Giải pháp nhận diện phần tử biên (Outliers)**

* **Phân bố thống kê (statistical distributions):** Sử dụng các phép đo thống kê như Z-score để đánh giá mức độ lệch của dữ liệu từ giá trị trung bình. Các điểm dữ liệu có Z-score nằm ngoài một ngưỡng xác định có thể được coi là outliers.
* **Khoảng cách (distance):** Tính toán khoảng cách giữa các điểm dữ liệu và tìm kiếm các điểm có khoảng cách lớn hơn một ngưỡng xác định. Các điểm có khoảng cách lớn có thể được coi là outliers.
* **Mật độ (density):** Phân tích mật độ dữ liệu để xác định các vùng mật độ thấp, nơi outliers có thể tồn tại. Các điểm dữ liệu rơi vào các vùng mật độ thấp có thể được coi là outliers.
* **Độ lệch (deviation):** Đánh giá mức độ độ lệch của dữ liệu từ một mô hình dự đoán hoặc từ một xu hướng cụ thể. Sự lệch lớn của một điểm dữ liệu so với mô hình hoặc xu hướng có thể chỉ ra rằng nó là một outlier.
* **Quan sát boxplot:** Sử dụng biểu đồ boxplot để hiển thị phân phối của dữ liệu và xác định giá trị ngoại lai dựa trên các "râu" của boxplot. Các điểm dữ liệu nằm ngoài phạm vi của các râu của boxplot có thể được coi là outliers.

#### **Giải pháp giảm thiểu nhiễu (Noisy Data)**

Dữ liệu nhiễu là các quan sát trong tập dữ liệu mà có giá trị sai lệch đáng kể so với phần còn lại của dữ liệu. Những quan sát này thường không tuân theo mẫu hay xu hướng chung của dữ liệu và có thể làm giảm độ chính xác của các phân tích và dự đoán.

* **IQR (Interquartile Range)**: Phương pháp này sử dụng phạm vi giữa tứ phân vị thứ nhất (Q1) và thứ ba (Q3) của dữ liệu để xác định và loại bỏ outliers. Cụ thể, một quan sát được xem là một outlier nếu nó nằm ngoài phạm vi IQR nhân với một hằng số (thông thường 1.5 hoặc 3).
* **Độ lệch chuẩn**: Phương pháp này loại bỏ dữ liệu nằm ngoài một số lượng độ lệch chuẩn nhất định từ trung bình. Thông thường, các quan sát nằm ngoài 2 hoặc 3 độ lệch chuẩn từ trung bình được coi là outliers.
* **Z-scores**: Điểm Z là một độ đo thống kê cho biết mức độ mà một quan sát cụ thể khác biệt so với trung bình của dữ liệu trong các đơn vị của độ lệch chuẩn. Dùng điểm Z để xác định dữ liệu có giá trị cao hoặc thấp bất thường so với trung bình. Cụ thể, các quan sát có điểm Z lớn hơn một ngưỡng cụ thể (ví dụ: 2 hoặc 3) được coi là outliers.

#### **Giải pháp xử lý dữ liệu không nhất quán (Inconsistent Data)**

Dữ liệu không nhất quán là tình trạng trong đó dữ liệu không tuân thủ một tiêu chuẩn nhất định, thường là do sự không đồng nhất về định dạng, kiểu dữ liệu hoặc cấu trúc.

* **Chuẩn hóa dữ liệu:**
  + Xác định và thiết lập một tiêu chuẩn cho định dạng và kiểu dữ liệu mà bạn muốn sử dụng.
  + Chuyển đổi dữ liệu vào định dạng và kiểu dữ liệu chuẩn bằng cách sử dụng các công cụ như các hàm chuyển đổi, công cụ ETL (Extract, Transform, Load), hoặc script tự động.
* **Làm sạch dữ liệu:**
* Kiểm tra dữ liệu để phát hiện và sửa chữa các không nhất quán như dữ liệu lặp lại, giá trị thiếu, giá trị ngoại lai, v.v.
* Sử dụng các kỹ thuật như loại bỏ dữ liệu trùng lặp, điền giá trị thiếu bằng giá trị phù hợp, loại bỏ hoặc điều chỉnh giá trị ngoại lai.
* **Phân tích sự phụ thuộc:**
  + Phân tích cấu trúc dữ liệu để hiểu các mối quan hệ và sự phụ thuộc giữa các thuộc tính.
  + Sử dụng các phương pháp như phân tích tương quan, hồi quy, hoặc phân tích cụm để xác định mối quan hệ giữa các thuộc tính.
  + Dựa trên phân tích này, đánh giá các yếu tố gây ra sự không nhất quán và thực hiện các biện pháp để giải quyết chúng.

## **Biến đổi dữ liệu**

Biến đổi dữ liệu là quy trình quan trọng để sửa đổi và điều chỉnh dữ liệu để nó phù hợp tốt hơn với các thuật toán phân tích và mô hình học máy. Quá trình này gồm việc sửa đổi, tính toán, phân tách và kết hợp dữ liệu thô thành các mô hình dữ liệu sẵn sàng phân tích.

### **Chuẩn hóa (normalization)**

Chuẩn hoá là một quá trình quan trọng để điều chỉnh các biến số trong tập dữ liệu sao cho chúng có cùng tỷ lệ hoặc phạm vi. Mục tiêu của chuẩn hoá là làm cho dữ liệu trở nên thống nhất và dễ dàng đối xử hơn, từ đó cải thiện hiệu suất của các mô hình học máy và đảm bảo tính ổn định của kết quả.

Khi dữ liệu thu thập từ nhiều nguồn khác nhau, hoặc các biến số có phạm vi khác nhau, có thể gây ra sự không thống nhất và khó khăn trong quá trình phân tích. Chuẩn hoá giúp giảm thiểu ảnh hưởng của các biến số không cùng tỷ lệ và phạm vi, từ đó tăng cơ hội cho các mô hình học máy nhận diện các mẫu và mối quan hệ một cách chính xác hơn.

Các phương pháp chuẩn hoá phổ biến bao gồm:

* **Chuẩn hóa Min-Max:** Chuyển đổi dữ liệu thành một phạm vi cố định (thường là từ 0 đến 1) bằng cách lấy giá trị hiện tại trừ đi giá trị nhỏ nhất và chia cho phạm vi giá trị lớn nhất trừ đi giá trị nhỏ nhất.
* **Chuẩn hóa Z-score (Standardization):** Biến đổi dữ liệu sao cho có giá trị trung bình gần bằng 0 và độ lệch chuẩn gần bằng 1 bằng cách lấy giá trị hiện tại trừ đi giá trị trung bình và chia cho độ lệch chuẩn.

Bằng cách chuẩn hóa dữ liệu, chúng ta có thể cải thiện tính đồng nhất và tính linh hoạt của dữ liệu, giúp các mô hình học máy hoạt động hiệu quả hơn và dễ dàng hiểu được mối quan hệ giữa các biến số. Điều này làm tăng khả năng dự đoán và giảm thiểu sai số khi áp dụng các mô hình vào thực tế.

### **Xây dựng thêm thuộc tính**

Một phần quan trọng của quá trình tiền xử lý dữ liệu là xây dựng thêm các thuộc tính mới từ dữ liệu hiện có. Điều này có thể thực hiện thông qua các phép toán đơn giản như cộng, trừ, nhân hoặc chia các biến số hiện có, hoặc thông qua các phép biến đổi phức tạp hơn như Phân tích thành phần chính (PCA). Bằng cách này, ta có thể nắm bắt thông tin quan trọng trong dữ liệu và giúp cải thiện hiệu suất của các mô hình dự đoán.

## **Rời rạc hóa dữ liệu**

Rời rạc hóa là quá trình quan trọng trong phân tích dữ liệu. Dữ liệu được chuyển từ dạng liên tục, không liên tục sang dạng rời rạc. Nhờ vậy, việc xử lý và phân tích dữ liệu trở nên dễ dàng hơn. Thông thường, rời rạc hóa dữ liệu được ứng dụng rộng rãi trong các ứng dụng của học máy và phân tích dữ liệu. Quá trình này giúp tạo ra các biến số hoặc các nhóm dữ liệu định tính giúp thu gọn và tối ưu hóa quá trình phân tích. Đồng thời, mô hình cũng giảm bớt đi tối đa độ phức tạp cần xử lý.

### **Encoding**

Encoding of data là quá trình trình chuyển đổi dữ liệu thành các tính hiệu điện tử hoặc giá trị mà máy tính tính có thể hiểu được. Nó có thể là các giá trị nhị phân được biểu thị ở dạng 0 và 1.

### **Binning**

Binning là một kỹ thuật trong phân tích dữ liệu, trong đó chúng ta chia nhỏ dữ liệu liên tục thành các "bins" hoặc "buckets" theo một số nguyên tắc cụ thể. Các nguyên tắc này có thể bao gồm việc chia bin sao cho các khoảng có chiều rộng đồng đều, hoặc chia bin sao cho các khoảng có tần suất đồng đều. Bằng cách này, binning giúp chúng ta tổ chức và biểu diễn dữ liệu một cách cụ thể và rõ ràng hơn. Từ đó tạo điều kiện thuận lợi cho việc phân tích và trực quan hóa.

### **Cluster Analysis**

Phân tích cụm là một phương pháp sử dụng thuật toán phân cụm để nhóm các điểm dữ liệu tương tự nhau thành các cụm. Phương pháp này có thể được áp dụng để rời rạc hóa dữ liệu bằng cách gán mỗi điểm dữ liệu vào cụm tương ứng của nó.

# Kết quả thực nghiệm

## **Giới thiệu dataset**

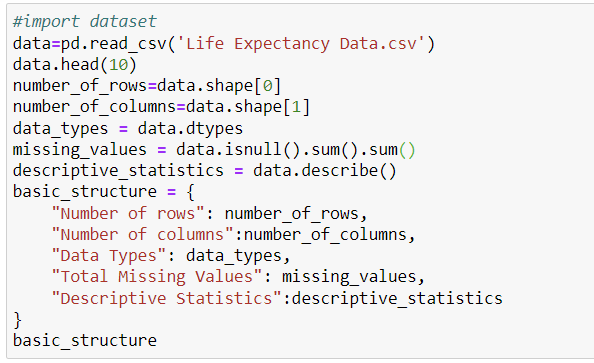
Bộ dữ liệu "*Life Expectancy Data.csv*" được tạo ra với mục đích chính là cung cấp thông tin chi tiết về các yếu tố ảnh hưởng đến tuổi thọ trung bình của các quốc gia qua từng năm khác nhau. Bao gồm các thông tin sau:

* **Country**: Tên của quốc gia.
* **Year:** Năm của dữ liệu.
* **Status:** Trạng thái phát triển của quốc gia (ví dụ: đã phát triển và đang phát triển).
* **Life expectancy:** Tuổi thọ trung bình của dân số trong quốc gia.
* **Adult Mortality:** Tỷ lệ tử vong ở người lớn (giữa 15 và 60 tuổi) trên mỗi 1000 dân trong quốc gia.
* **Infant deaths:** Số lượng trẻ sơ sinh tử vong trong quốc gia.
* **Alcohol:** Tiêu thụ cồn trên đầu người, tính bằng số lít cồn tiêu thụ trên mỗi người trên một năm.
* **Percentage expenditure:** Phần trăm chi tiêu cho sức khỏe từ tổng chi tiêucủa quốc gia.
* **Hepatitis B:** Tỷ lệ tiêm phòng viêm gan B ở trẻ em dưới 1 tuổi.
* **Measles:** Số trường hợp bệnh sởi được báo cáo trong năm.
* **BMI:** Chỉ số khối cơ thể (BMI) trung bình của người lớn trong quốc gia.
* **Under-five deaths:** Số trẻ em dưới 5 tuổi tử vong trong quốc gia.
* **Polio:** Tỷ lệ tiêm phòng bệnh bại liệt ở trẻ em dưới 1 tuổi.
* **Total expenditure:** Tổng chi tiêu cho sức khỏe trên tổng sản phẩm quốc nội của quốc gia.
* **Diphtheria:** Tỷ lệ tiêm phòng bệnh bạch hầu ở trẻ em dưới 1 tuổi.
* **HIV/AIDS:** Ước tính tỷ lệ người nhiễm HIV/AIDS trong dân số của quốc gia.
* **GDP:** Tổng sản phẩm quốc nội của quốc gia (Gross Domestic Product).
* **Population:** Dân số của quốc gia.
* **Thinness 1-19 years:** Tỷ lệ trẻ em (giai đoạn từ 10 đến 19 tuổi) có triệu chứng suy dinh dưỡng.
* **Thinness 5-9 years:** Tỷ lệ trẻ em (giai đoạn từ 5 đến 9 tuổi) có triệu chứng suy dinh dưỡng.
* **Income composition of resources:** Thành phần thụ nhập từ các nguồn khác nhau như thụ nhập lao động, thu nhập từ doanh nghiệp, và các nguồn thu nhập khác.
* **Schooling:** Số năm học trung bình của người dân trong quốc gia.

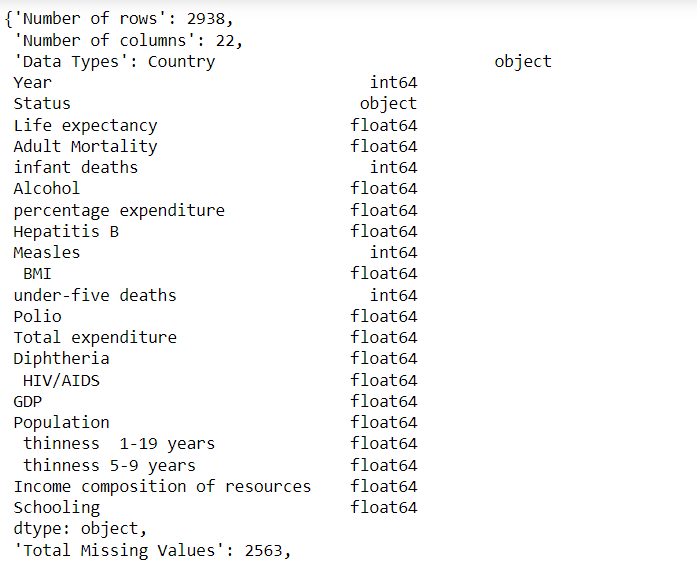
## **Thực hành trên dataset**

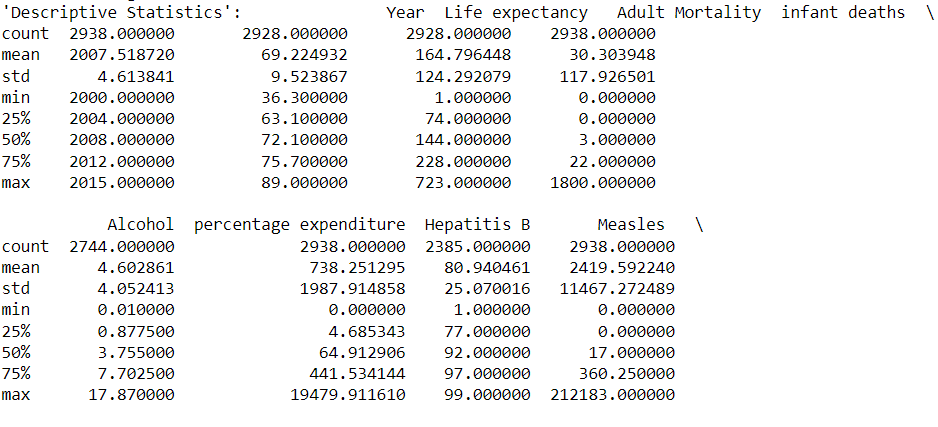
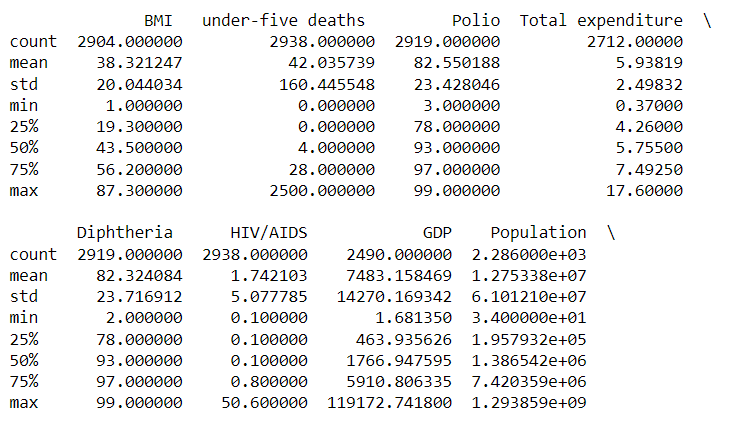
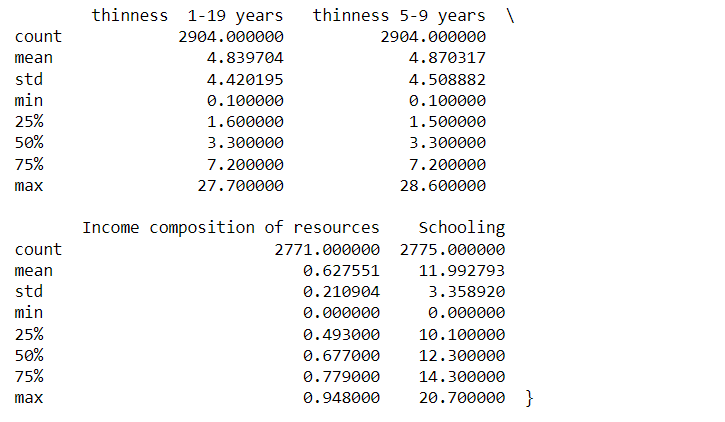
### **Mô tả dataset**

#### **Code**



#### **Kết quả**



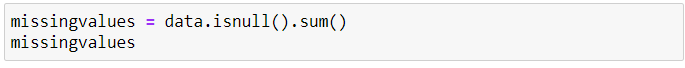
#### **Đánh giá kết quả**

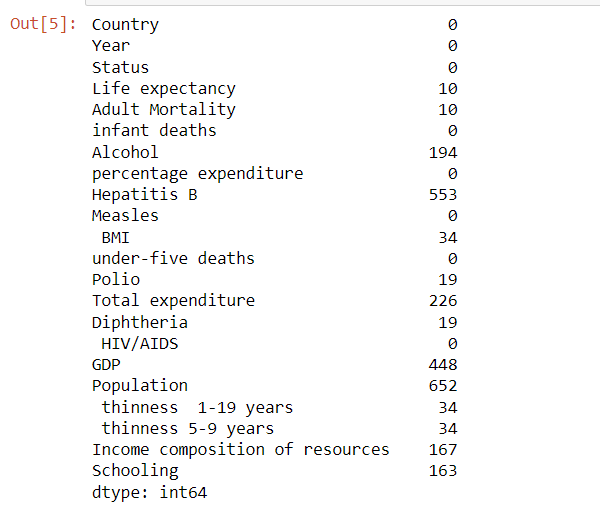
* Số lượng hàng: 2938
* Số lượng cột: 22
* Kiểu dữ liệu của từng cột:
* Year: int64
* Status: object
* Life expectancy: float64
* Adult Mortality: float64
* Infant deaths: int64
* Alcohol: float64
* Percentage expenditure: float64
* Pepatitis B: float64
* Measles: int64
* BMI: float64
* Under-five deaths: int64
* Polio: float64
* Total expenditure:                  float64
* Diphtheria:                        float64
* HIV/AIDS:                         float64
* GDP:                                float64
* Population:                         float64
* Thinness  1-19 years:              float64
* Thinness 5-9 years:                float64
* Income composition of resources:    float64
* Schooling                          float64
* Tổng số giá trị thiếu: 2563
* Thống kê cơ bản:
* Year: từ năm 2000 đến năm 2015
* Life expectancy (Tuổi thọ trung bình):
  + - Số lượng quan sát: 2928
    - Giá trị trung bình: 69.22
    - Độ lệch chuẩn: 9.52
    - Giá trị nhỏ nhất: 36.3
    - Giá trị lớn nhất: 89.0
* Adult Mortality (Tỷ lệ tử vong ở người lớn):
  + - Số lượng quan sát: 2928
    - Giá trị trung bình: 164.80
    - Độ lệch chuẩn: 124.29
    - Giá trị nhỏ nhất: 1.0
    - Giá trị lớn nhất: 723.0
* Infant deaths (Số trẻ em sơ sinh tử vong):
  + - Số lượng quan sát: 2938
    - Giá trị trung bình: 30.30
    - Độ lệch chuẩn: 117.93
    - Giá trị nhỏ nhất: 0.0
    - Giá trị lớn nhất: 1800.0
* Alcohol (Tiêu thụ cồn):
  + - Số lượng quan sát: 2744
    - Giá trị trung bình: 4.60
    - Độ lệch chuẩn: 4.05
    - Giá trị nhỏ nhất: 0.01
    - Giá trị lớn nhất: 17.87
* Percentage expenditure (Chi phí sức khỏe chiếm phần trăm):
  + - Số lượng quan sát: 2938
    - Giá trị trung bình: 738.25
    - Độ lệch chuẩn: 1987.91
    - Giá trị nhỏ nhất: 0.0
    - Giá trị lớn nhất: 19479.91
* Hepatitis B (Tỷ lệ tiêm phòng viêm gan B):
  + - Số lượng quan sát: 2385
    - Giá trị trung bình: 80.94
    - Độ lệch chuẩn: 25.07
    - Giá trị nhỏ nhất: 1.0
    - Giá trị lớn nhất: 99.0
* Measles (Số trường hợp bệnh sởi):
  + - Số lượng quan sát: 2938
    - Giá trị trung bình: 2419.59
    - Độ lệch chuẩn: 11467.27
    - Giá trị nhỏ nhất: 0.0
    - Giá trị lớn nhất: 212183.0
* BMI (Chỉ số khối cơ thể):
  + - Số lượng quan sát: 2904
    - Giá trị trung bình: 38.32
    - Độ lệch chuẩn: 20.04
    - Giá trị nhỏ nhất: 1.0
    - Giá trị lớn nhất: 87.3
* Under-five deaths (Số trẻ em dưới 5 tuổi tử vong):
  + - Số lượng quan sát: 2938
    - Giá trị trung bình: 42.04
    - Độ lệch chuẩn: 160.45
    - Giá trị nhỏ nhất: 0.0
    - Giá trị lớn nhất: 2500.0
* Polio (Tỷ lệ tiêm phòng bệnh bại liệt):
  + - Số lượng quan sát: 2919
    - Giá trị trung bình: 82.55
    - Độ lệch chuẩn: 23.43
    - Giá trị nhỏ nhất: 3.0
    - Giá trị lớn nhất: 99.0
* Total expenditure (Tổng chi tiêu cho sức khỏe):
  + - Số lượng quan sát: 2712
    - Giá trị trung bình: 5.94
    - Độ lệch chuẩn: 2.50
    - Giá trị nhỏ nhất: 0.37
    - Giá trị lớn nhất: 17.6
* Diphtheria (Tỷ lệ tiêm phòng bệnh bạch hầu):
  + - Số lượng quan sát: 2919
    - Giá trị trung bình: 82.32
    - Độ lệch chuẩn: 23.72
    - Giá trị nhỏ nhất: 2.0
    - Giá trị lớn nhất: 99.0
* HIV/AIDS (Tỷ lệ người nhiễm HIV/AIDS):
  + - Số lượng quan sát: 2938
    - Giá trị trung bình: 1.74
    - Độ lệch chuẩn: 5.08
    - Giá trị nhỏ nhất: 0.1
    - Giá trị lớn nhất: 50.6
* GDP (Tổng sản phẩm quốc nội):
  + - Số lượng quan sát: 2490
    - Giá trị trung bình: 7483.16
    - Độ lệch chuẩn: 14270.17
    - Giá trị nhỏ nhất: 1.68
    - Giá trị lớn nhất: 119172.74
* Population (Dân số):
  + - Số lượng quan sát: 2.286000e+03
    - Giá trị trung bình: 1.275338e+07
    - Độ lệch chuẩn:6.101210e+07
    - Giá trị nhỏ nhất:3.400000e+01
    - Giá trị lớn nhất: 1.293859e+09
* Thinness 1-19 years (Tỷ lệ trẻ em suy dinh dưỡng từ 1 đến 19 tuổi):
  + - Số lượng quan sát: 2904
    - Giá trị trung bình: 4.83
    - Độ lệch chuẩn: 4.42
    - Giá trị nhỏ nhất: 0.1
    - Giá trị lớn nhất: 27.7
* Thinness 5-9 years (Tỷ lệ trẻ em suy dinh dưỡng từ 5 đến 9 tuổi):
  + - Số lượng quan sát: 2904
    - Giá trị trung bình: 4.87
    - Độ lệch chuẩn: 4.51
    - Giá trị nhỏ nhất: 0.1
    - Giá trị lớn nhất: 28.6
* Income composition of resources (Thành phần thu nhập từ nguồn tài nguyên):
  + - Số lượng quan sát: 2771
    - Giá trị trung bình: 0.63
    - Độ lệch chuẩn: 0.21
    - Giá trị nhỏ nhất : 0.0
    - Giá trị lớn nhất :0.95
* Schooling (Năm học):
  + - Số lượng quan sát: 2775
    - Giá trị trung bình: 11.99
    - Độ lệch chuẩn: 3.36
    - Giá trị nhỏ nhất: 0.0
    - Giá trị lớn nhất: 20.7

### **Làm sạch dữ liệu**

#### **Xử lý dữ liệu thiếu (Missing Data)**

#### **Quan sát dữ liệu bị thiếu**

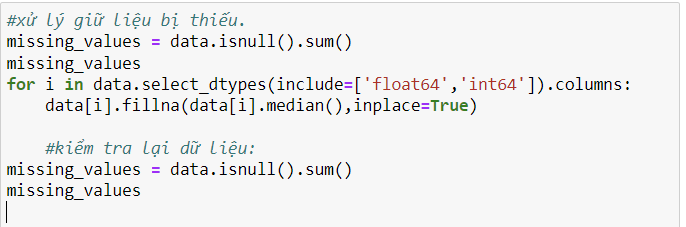




Nhận xét:

* Trong việc xử lý các dữ liệu thiếu (missing data), việc quyết định loại bỏ hoặc điền vào giá trị cho các dòng dữ liệu bị thiếu phụ thuộc vào ngữ cảnh, ý nghĩa của dữ liệu và mục đích của việc phân tích. Trong trường hợp này, dataset là dữ liệu thống kê của các quốc gia qua từng năm. Việc xử lý dữ liệu bị thiếu sẽ giúp cung cấp đủ thông tin cho mô hình để triển khai đào tạo, dự đoán một cách hiệu quả hơn.
* Nếu như dữ liệu bị thiếu không được xử lý có thể dẫn đến việc thiếu sót thông tin về một nhóm hoặc một biến quan trọng có ảnh hưởng trực tiếp lên biến mục tiêu “Life expectancy”, làm cho việc phân tích trở nên không công bằng, không có tính khách quan.
* Bằng cách xử lý dữ liệu bị thiếu một cách chính xác, ta đảm bảo rằng việc phân tích được thực hiện trên toàn bộ dữ liệu và không bị thiên vị.

#### **Code**



#### **Kết quả**



#### **Đánh giá kết quả**

Sau khi xử lý dữ liệu bị thiếu (missing data) bây giờ có:

* Số lượng hàng: 2938 (không đổi)
* Số lượng cột: 22 (không đổi)

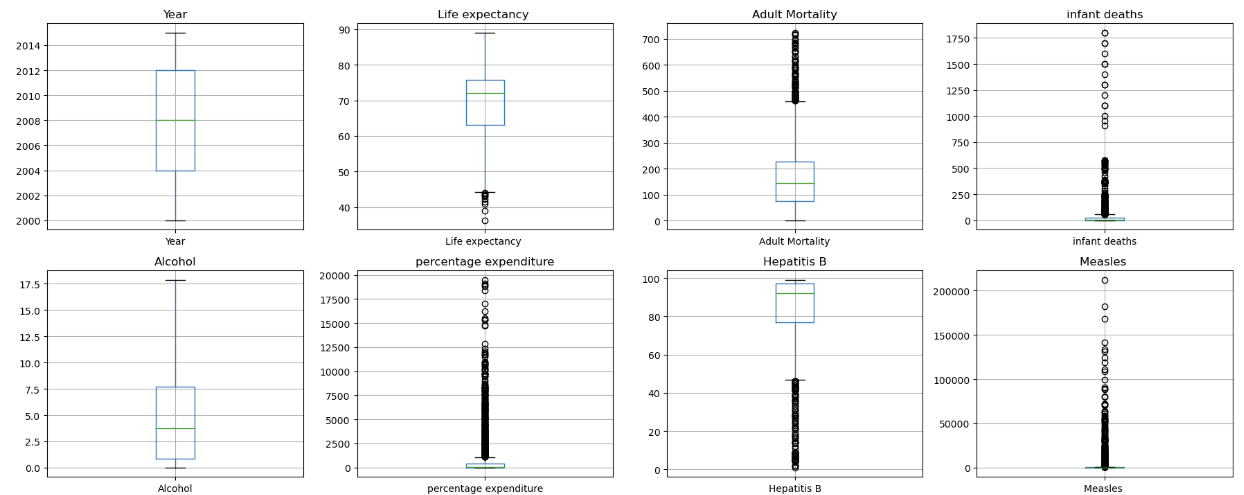
Dữ liệu bị thiếu đã được thay thế bằng giá trị trung vị. Giá trị trung vị sẽ ít bị ảnh hưởng bởi các giá trị ngoại lệ (outliers) hơn so với giá trị trung bình. Đặc biệt, đối với dữ liệu có phân phối không đối xứng như dataset Life Expectancy Data thì giá trị trung vị sẽ phản ánh trung tâm của dữ liệu tốt hơn. Theo đó, quá trình phân tích sẽ được thực hiện trên dữ liệu một cách đầy đủ và có độ chính xác cao.

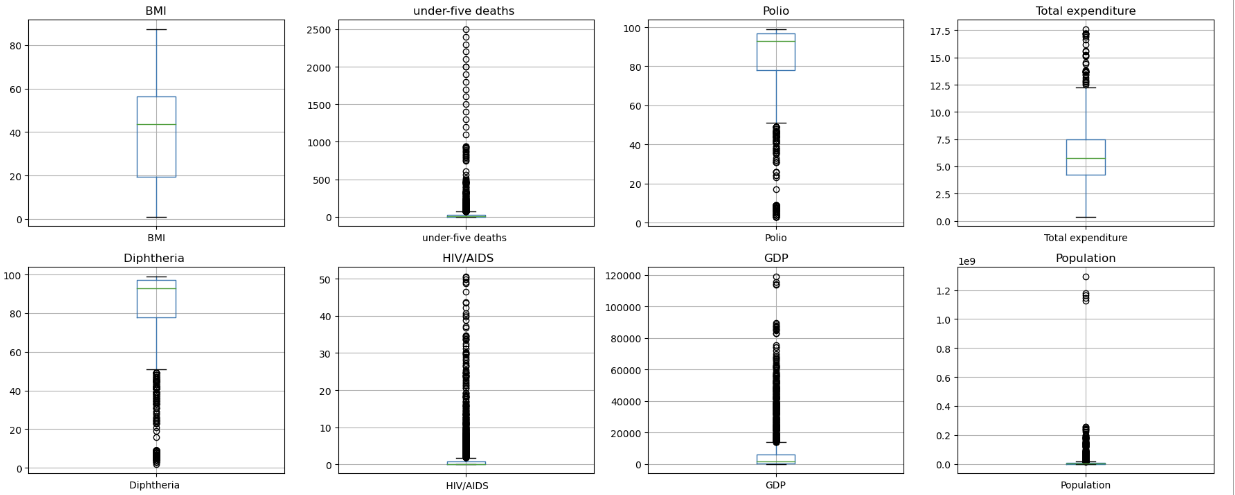
#### **Nhận diện phần tử biên**

#### **Code**

****

#### **Kết quả**

****

****

****

#### **Đánh giá kết quả**

Biểu đồ boxplot cho mỗi cột số. Trong mỗi biểu đồ:

Dòng ngang giữa hộp là giá trị trung vị của dữ liệu

Các cạnh của hộp thể hiện phạm vi Q1 -> Q3, tức là từ phân vị 25% -> 75% của dữ liệu.

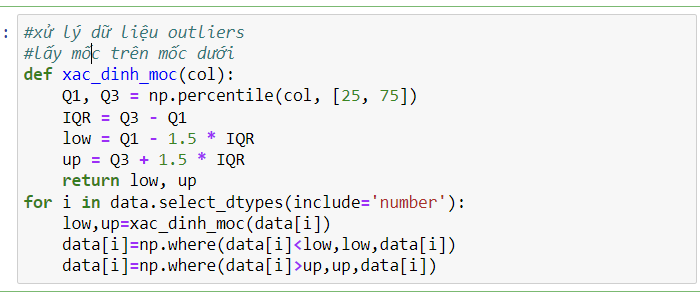
Các đường ngang (whiskers) mở rộng từ hộp thể hiện phạm vi của dữ liệu, ngoại trừ các outliers.

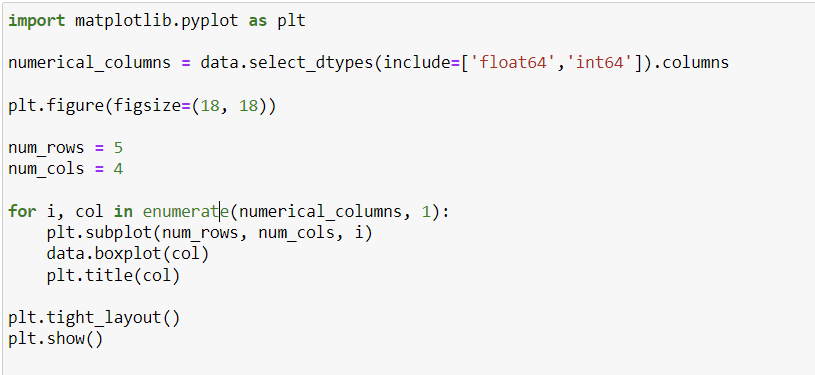
Các điểm nằm ngoài whiskers là giá trị ngoại lệ.

Các outliers này có thể là do lỗi đo lường, có thể là sự sai sót trong dữ liệu hoặc nó thực sự là chỉ số phản ánh sự biến thiên lớn trong dữ liệu.

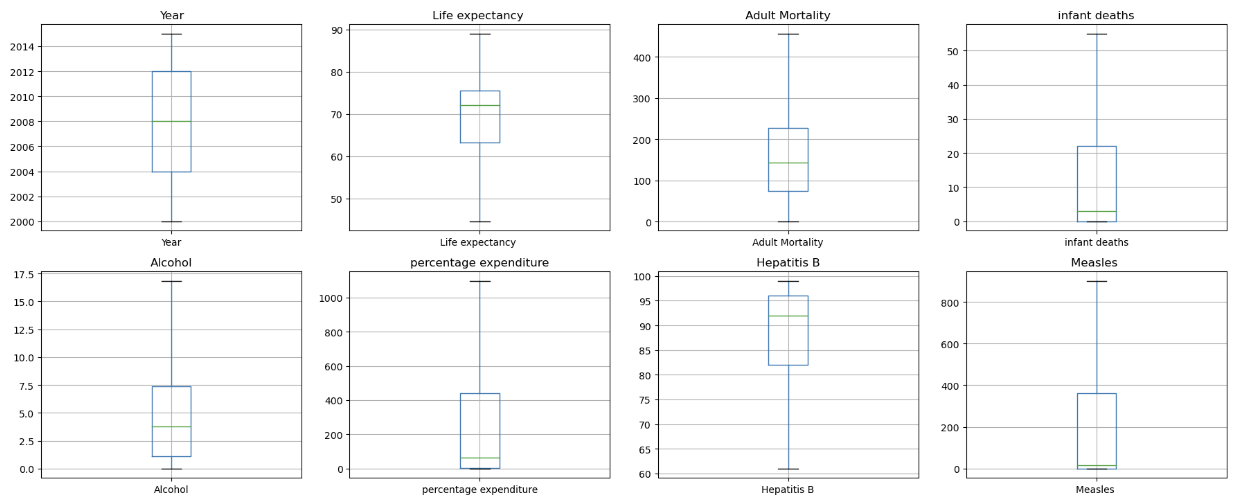
#### **Giảm thiểu nhiễu**

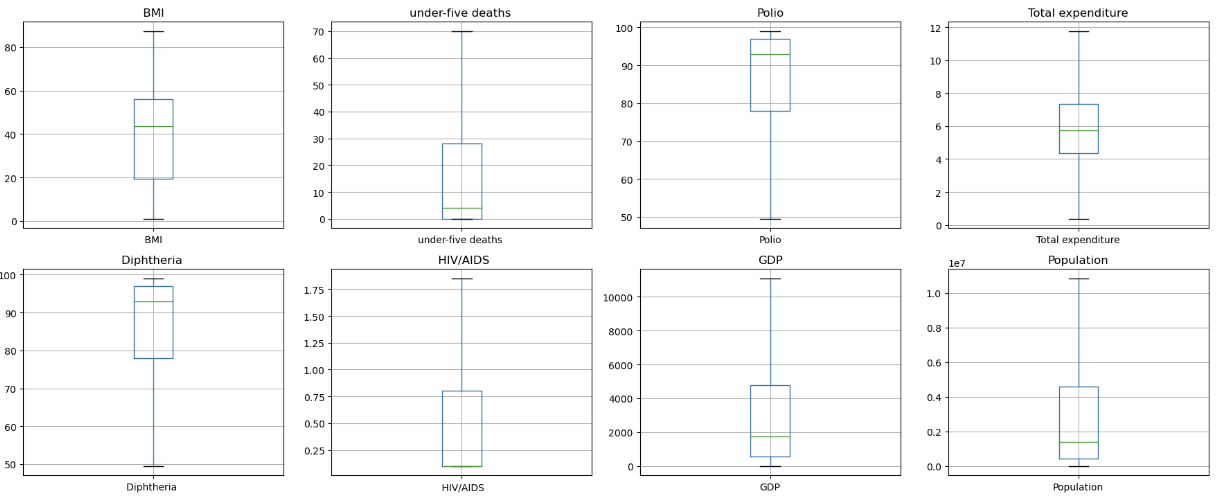
#### **Code**

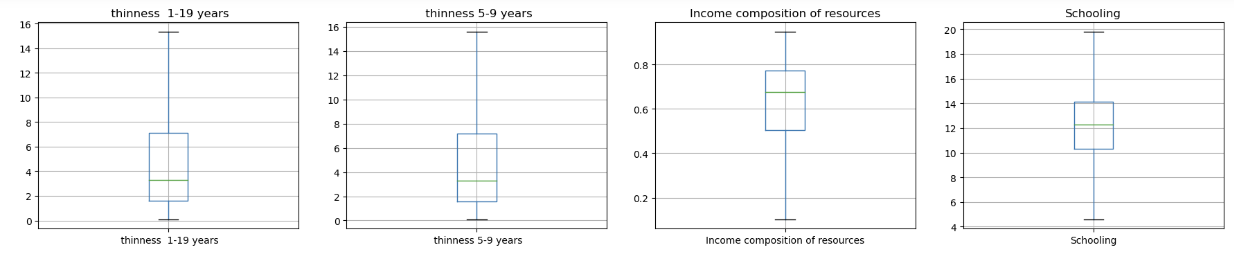
****

****

#### **Kết quả**

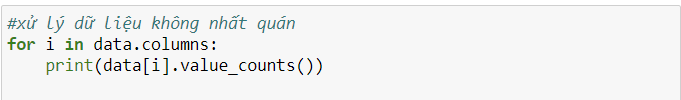




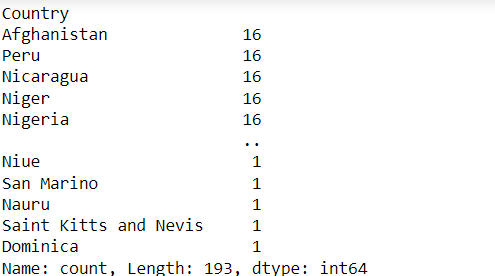


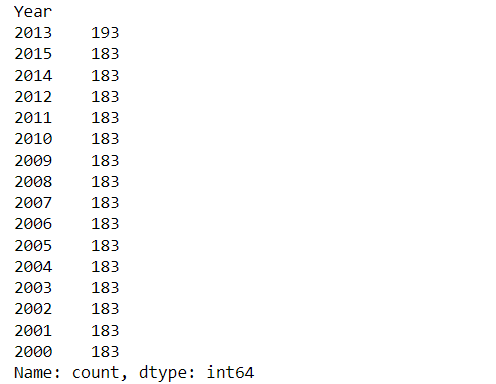
#### **Xử lý dữ liệu không nhất quán**

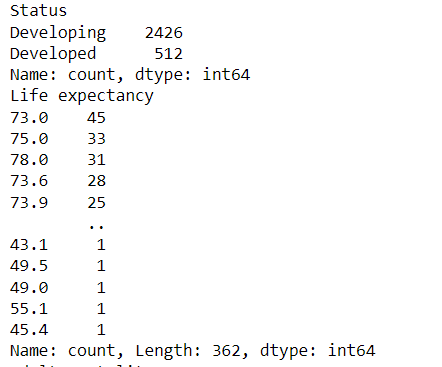
#### **Code**

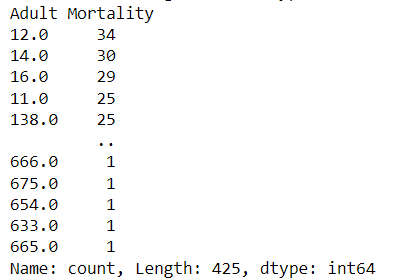
****

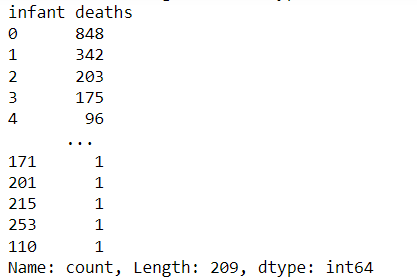
#### **Kết quả**

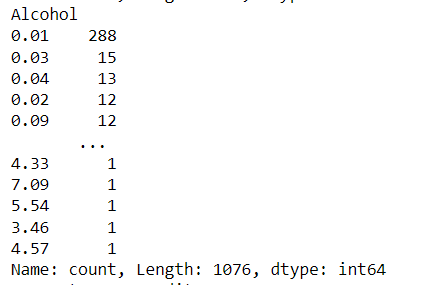
****

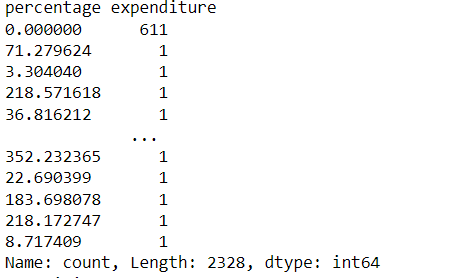
****

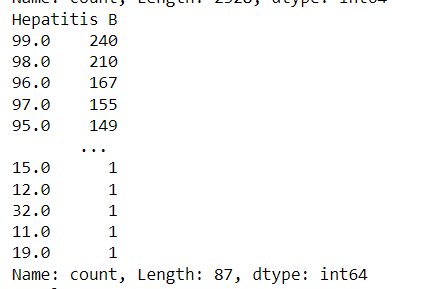
****

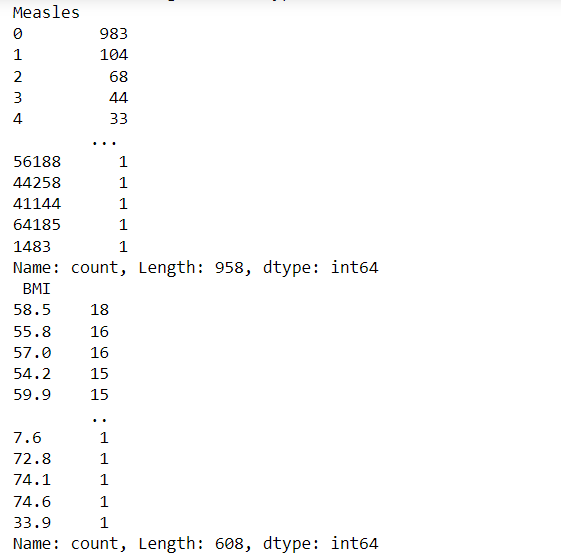
****

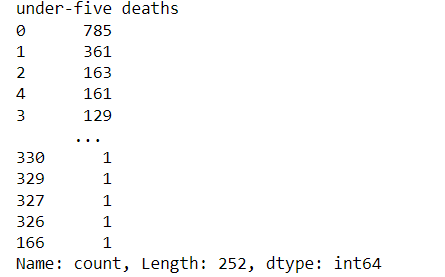
****

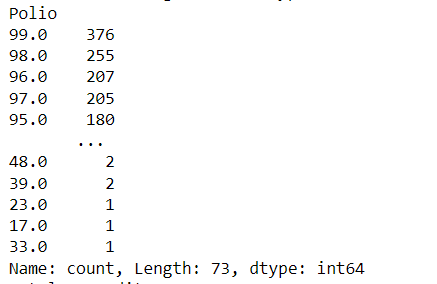
****

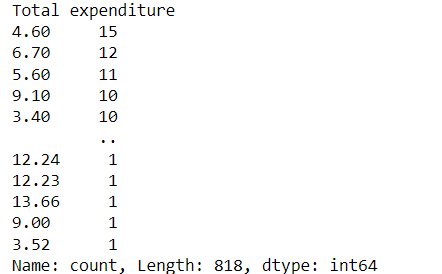
****

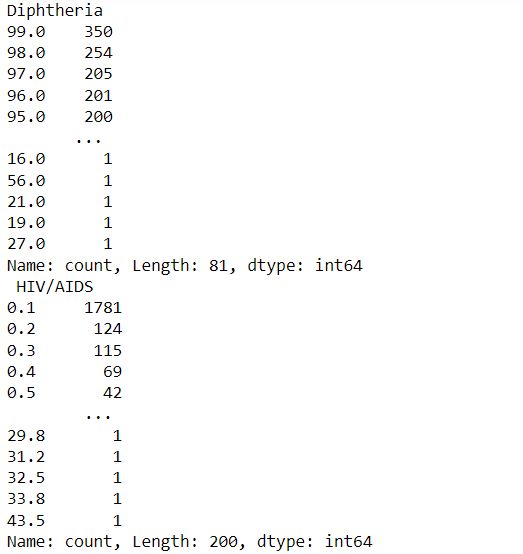
****

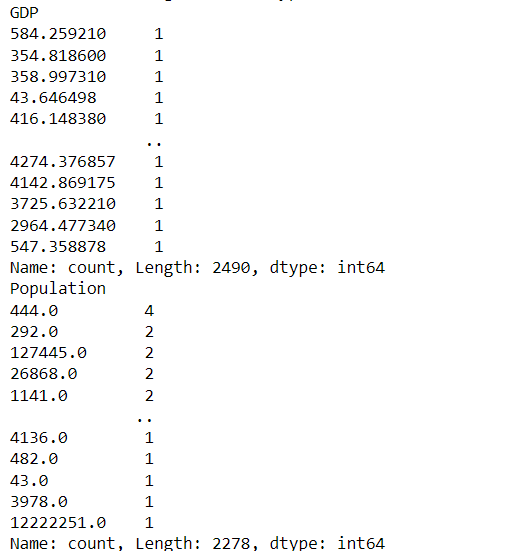


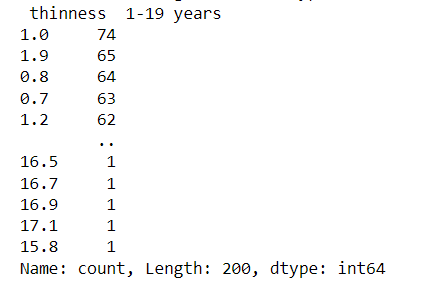


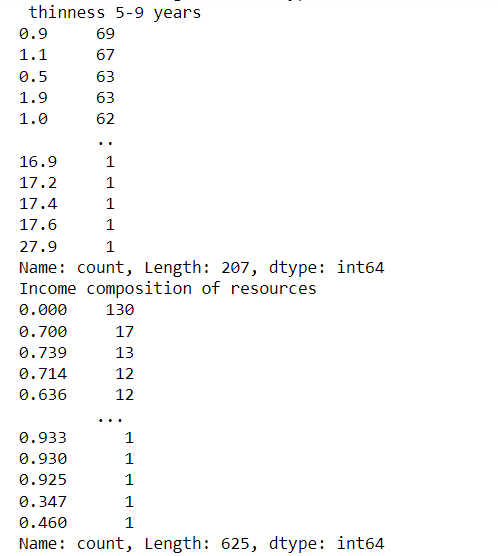


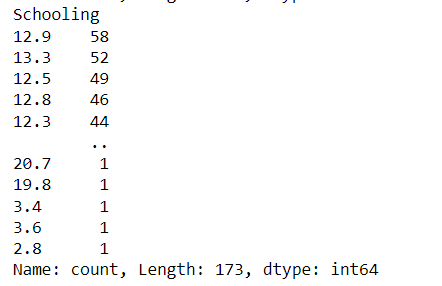












#### **Đánh giá kết quả**

Dataset trên không chứa dữ liệu không nhất quán. Các dữ liệu tuân thủ theo một tiêu chuẩn nhất định, có sự đồng nhất về định dạng, kiểu dữ liệu, cấu trúc.

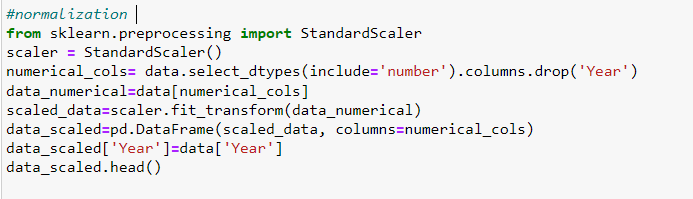
#### **Chuẩn hóa**

#### **Phương pháp**

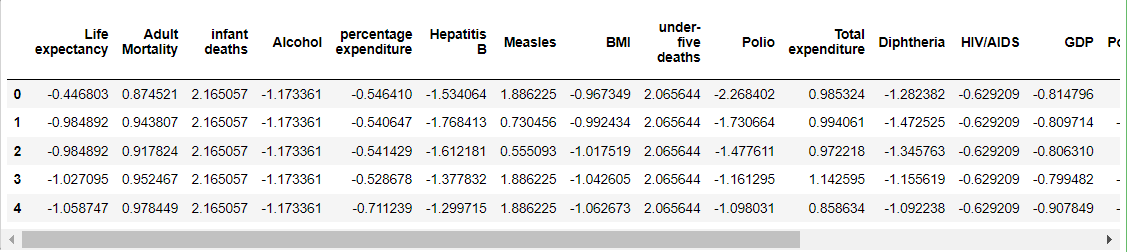
Chọn phương pháp Z - score vì:

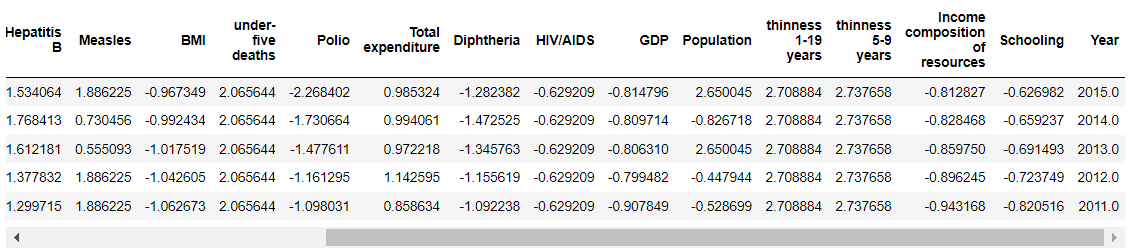
* Phạm vi và độ lệch chuẩn khác nhau: Các đặc trưng trong bộ dữ liệu trên có phạm vi min - max và độ lệch chuẩn khác nhau. Vậy nên điều này có thể gây ra những vấn đề trong việc so sánh hoặc kết hợp chúng trong mô hình học máy.
* Ít nhạy cảm với ngoại lệ: Chuẩn hóa z - score ít nhạy cảm với sự xuất hiện của ngoại lệ hơn so với chuẩn hóa min - max. Điều này đặc biệt quan trọng nếu bộ dữ liệu ngoại lệ mà chúng ta chưa kiểm tra.

#### **Code**

****

#### **Kết quả**

****

****

#### **Đánh giá kết quả**

Sau khi chuẩn hóa dữ liệu là một bảng dữ liệu mới với những giá trị được chuẩn hóa cho mỗi đặc trưng số, trừ cột ‘Year’, theo đó:

* Trung bình (Mean) của mỗi đặc trưng sau khi thực hiện chuẩn hóa sẽ rơi vào khoảng 0, hoặc gần với 0 nếu không phải chính xác bằng 0 do làm tròn số.
* Độ lệch chuẩn (Standard Deviation) của mỗi đặc trưng sau khi chuẩn hóa sẽ là 1.
* Cột year không thay đổi vì nó không được lựa chọn để chuẩn hóa, nó thể hiện năm mà dữ liệu được thu thập, không bị bị ảnh hưởng bởi quá trình chuẩn hóa.

#### **Xây dựng thêm thuộc tính**

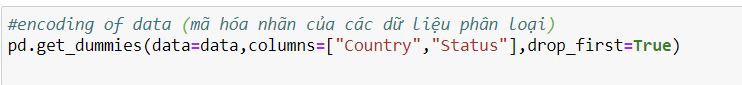
Trong dataset trên, không cần thiết phải xây dựng thêm thuộc tính bởi vì các yếu tố sau đã được đáp ứng:

* Bộ dữ liệu Life Expectancy Data hiện tại tại đã chứa đầy đủ thông tin cần thiết cho mục tiêu chính là phân tích tuổi thọ “Life Expectancy” hoặc mô hình học máy.
* Các thuộc tính trong dataset đã cung cấp một cái nhìn toàn diện về số liệu tuổi thọ trung bình của mỗi quốc gia qua từng năm.
* Việc thêm mới thuộc tính có thể không cần thiết hoặc thậm chí có thể tăng độ phức tạp của dữ liệu.
* Vậy nên, dữ liệu này đã đáp ứng đủ yêu cầu phân tích, việc giữ nguyên toàn bộ thuộc tính hiện tại là một quyết định hợp lý.

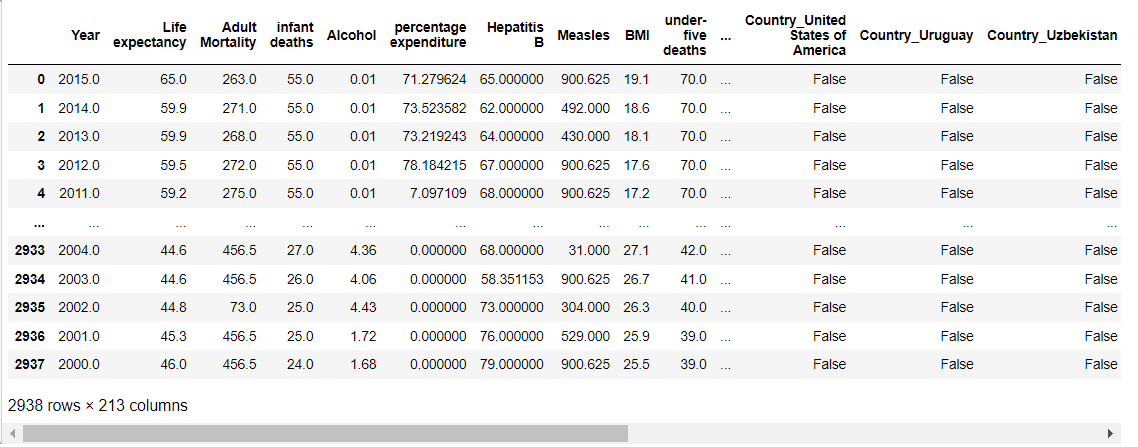
### **Rời rạc hóa dữ liệu**

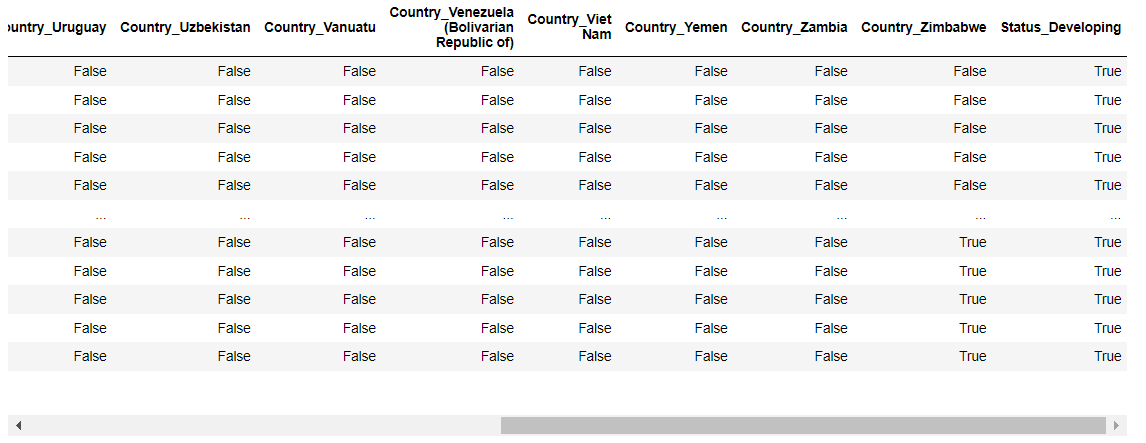
#### **Encoding**

#### **Code**

****

#### **Kết quả**

****

****

#### **Đánh giá kết luận**

One hot encoding thực hiện bằng lệnh get\_dummies. Cột được mã hóa đầu tiên là “country”.

Trong cơ sở dữ liệu, cột “Country” có nhiều giá trị có thể có như “VietNam”, “Yemen”, “Vanuatu”, v.v., xem xét các cột được mã hóa , đối với mỗi giá trị có thể, một cột mới được tạo trong cơ sở dữ liệu, với tên cột được thêm vào dưới dạng tiền tố.

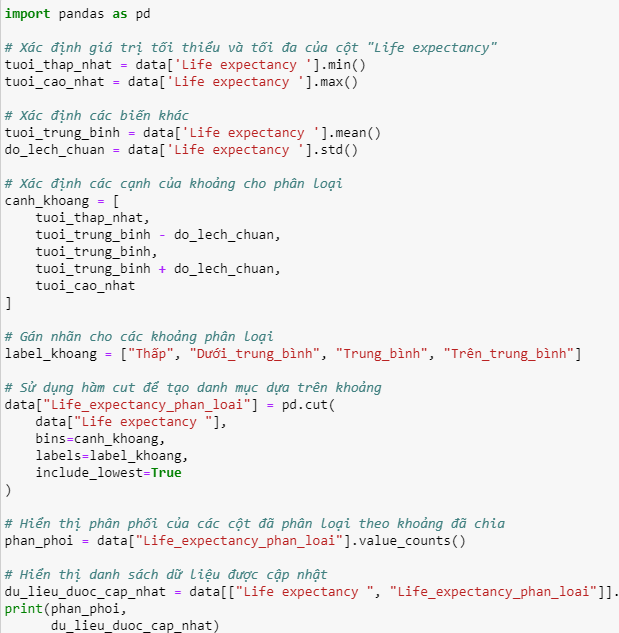
Do đó, ta có các cột được gọi là “Country-VietNam”, “Country-Yemen”, v.v.

Cột thứ hai trong cơ sở dữ liệu mới được tạo là status cho biết tên các trạng thái country đang phát triển (developing)

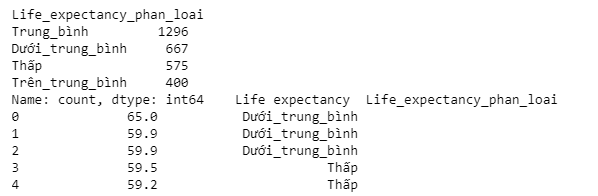
#### **Binning**

#### **Code**

Sử dụng phương pháp Binning cho cột Life expectancy



#### **Kết quả**



#### **Đánh giá kết quả**

* Có tổng 1296 dữ liệu được phân loại vào khoảng “Trung\_bình”, đây là nhóm có số lượng dữ liệu nhiều nhất.
* Nhóm “Dưới\_trung\_bình” có 667 dữ liệu, xếp ở vị trí thứ hai về số lượng.
* Nhóm “Thấp” có 575 dữ liệu.
* Nhóm “Trên\_trung\_bình” có 400 dữ liệu, ít hơn so với những nhóm khác.
* Nếu xem xét một số hàng đầu tiên của dữ liệu sau khi phân loại, ta thấy:
* Ba dữ liệu “Life\_expecttancy” đầu tiên với giá trị lần lượt là 65.0; 59,9 và 59,9. Tất cả đều được phân loại là “Dưới\_trung\_bình”
* Hai dữ liệu tiếp theo với giá trị bằng 59.5 và 59.2 được phân loại vào nhóm “Thấp”
* Từ những kết quả này, ta có thể nhận thấy rằng phân phối của chỉ số "Life expectancy" trong bộ dữ liệu phản ánh sự đa dạng về mức độ tuổi thọ hoặc chất lượng cuộc sống của các cộng đồng được nghiên cứu.

# Nhận xét tổng quan về bài nghiên cứu

Với bài báo cáo nghiên cứu trên, nhóm đã thực hiện đúng đầy đủ quy trình các bước tiền xử lý dữ liệu. Từ các bước làm sạch, biến đổi, cho đến rời rạc hóa dữ liệu đã được nhóm áp dụng một cách thận trọng, có chủ đích rõ ràng. Theo đó đảm bảo dữ liệu đầu vào cho các mô hình phân tích sau này là sạch và đáng tin cậy.

Cách tiếp cận này không chỉ cải thiện chất lượng của dữ liệu mà còn góp phần tăng cường tính khách quan, độ chính xác của kết quả nghiên cứu tuổi thọ trung bình các quốc gia. Theo đó sẽ tạo điều kiện thuận lợi, đưa ra các quyết định dựa trên dataset một cách hiệu quả hơn.

# Nguồn tài liệu tham khảo

Slide của giảng viên Lê Diên Tuấn

Preprocessing Lecture Notes (Chapter 3)

# Link video báo cáo của nhóm:

# https://www.youtube.com/watch?v=FpTzS7G-Ctk