

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

  

**BÁO CÁO**

**DATA PREPROCESSING**

***Môn: Phân Tích Dữ Liệu Bằng Python***

*Lớp sinh hoạt: 47K14 Lớp học phần: 46K29.1*

*SĐT Trưởng nhóm:0866730245*

|  |  |
| --- | --- |
| ***Giảng viên hướng dẫn***  **Lê Diên Tuấn** | ***Nhóm 6***  Trần Hoàng Nhật  Nguyễn Đức Việt Thái  Nguyễn Văn Phú Khang  Lê Khai Tâm  Võ Lê Anh Nhật |

*Đà Nẵng, tháng 11 năm 2023*

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và Tên** | **Phần trăm đóng góp** |
| 1 | Trần Hoàng Nhật ( Nhóm Trưởng )  Công việc: Làm, tìm hiểu lý thuyết, code, làm word, làm và quay video | 100% |
| 2 | Nguyễn Đức Việt Thái  Công việc: Quay video, trang bìa word | 80% |
| 3 | Nguyễn Văn Phú Khang  Công việc: Quay video, trang bìa word | 80% |
| 4 | Lê Khai Tâm  Công việc: Quay video, trang bìa word | 80% |
| 5 | Võ Lê Anh Nhật  Công việc: Quay video, trang bìa word | 80% |

# Lời Cam Đoan

Chúng em xin cam dự án là kết quả nghiên cứu của em dưới sự hướng dẫn của giáo viên hướng dẫn của Thầy Lê Diên Tuấn ngoài ra không có bất kì sự sao chép nào của người khác.

Dự án, nội dung báo cáo là sản phẩm mà chúng em đã nổ lực nghiên cứu trong quá trình học và tham khảo tài liệu được cung cấp từ thầy Lê Diên Tuấn. Các dữ liệu, kết quả trình bày trong báo cáo là hoàn toàn trung thực, chúng em xin chịu hoàn toàn trách nhiệm, kỷ luật bộ môn và nhà trường đề ra nếu có vấn đề nào xảy ra.

Mục Lục

[Lời Cam Đoan 3](#_Toc151303120)

[BÀI BÁO CÁO 6](#_Toc151303121)

[**1.** **Tóm Tắt:** 6](#_Toc151303122)

[**2.** **Giới Thiệu:** 6](#_Toc151303123)

[**3.** **Quy trình nghiên cứu (Framework)** 6](#_Toc151303124)

[**4.** **Cơ sở lý thuyết** 7](#_Toc151303125)

[**4.1.** **Thống kê mô tả** 7](#_Toc151303126)

[**4.1.1.** **Count** 7](#_Toc151303127)

[**4.1.2.** **Mean** 7](#_Toc151303128)

[**4.1.3.** **Std** 7](#_Toc151303129)

[**4.1.4.** **Min** 7](#_Toc151303130)

[**4.1.5.** **Phân vị thứ nhất** 7](#_Toc151303131)

[**4.1.6.** **Trung vị** 7](#_Toc151303132)

[**4.1.7.** **Phân vị thứ ba** 8](#_Toc151303133)

[**4.1.8.** **Max** 8](#_Toc151303134)

[**4.2.** **Làm sạch dữ liệu** 8](#_Toc151303135)

[**4.2.1.** **Xử lý dữ liệu thiếu (Missing Data)** 8](#_Toc151303136)

[**4.2.2.** **Nhận diện phần tử biên (Outliers)** 8](#_Toc151303137)

[**4.2.3.** **Giảm thiểu dữ liệu nhiễu (Noisy Data)** 8](#_Toc151303138)

[**4.2.4.** **Xử lý dữ liệu không nhất quán (Inconsistent Data)** 8](#_Toc151303139)

[**4.2.5.** **Giải pháp** 8](#_Toc151303140)

[**4.2.5.1.** **Giải pháp xử lý dữ liệu thiếu (Missing Data)** 8](#_Toc151303141)

[**4.2.5.2.** **Giải pháp nhận diện phần tử biên (Outliers)** 9](#_Toc151303142)

[**4.2.5.3.** **Giải pháp giảm thiểu nhiễu (Noisy Data)** 9](#_Toc151303143)

[**4.2.5.4.** **Giải pháp xử lý dữ liệu không nhất quán (Inconsistent Data)** 9](#_Toc151303144)

[**4.3.** **Biến đổi dữ liệu** 10](#_Toc151303145)

[**4.3.1.** **Chuẩn hoá (Normalization)** 10](#_Toc151303146)

[**4.3.2.** **Xây dựng thêm thuộc tính** 10](#_Toc151303147)

[**4.4.** **Rời rạc hoá dữ liệu** 10](#_Toc151303148)

[**4.4.1.** **Binning** 10](#_Toc151303149)

[**4.4.2.** **Cluster Analysis** 10](#_Toc151303150)

[**5.** **Tổng quan, giới thiệu Dataset** 11](#_Toc151303151)

[**6.** **Thực hành trên Dataset** 11](#_Toc151303152)

[**6.1.** **Mô tả Dataset** 11](#_Toc151303153)

[**6.2.** **Làm sạch dữ liệu** 13](#_Toc151303154)

[**6.3.** **Biến đổi dữ liệu** 21](#_Toc151303155)

[**6.4.** **Rời rạc hoá dữ liệu** 23](#_Toc151303156)

[**7.** **Nhận xét tổng quan bài nghiên cứu** 24](#_Toc151303157)

[**8.** **Nguồn tài liệu tham khảo** 24](#_Toc151303158)

# BÀI BÁO CÁO

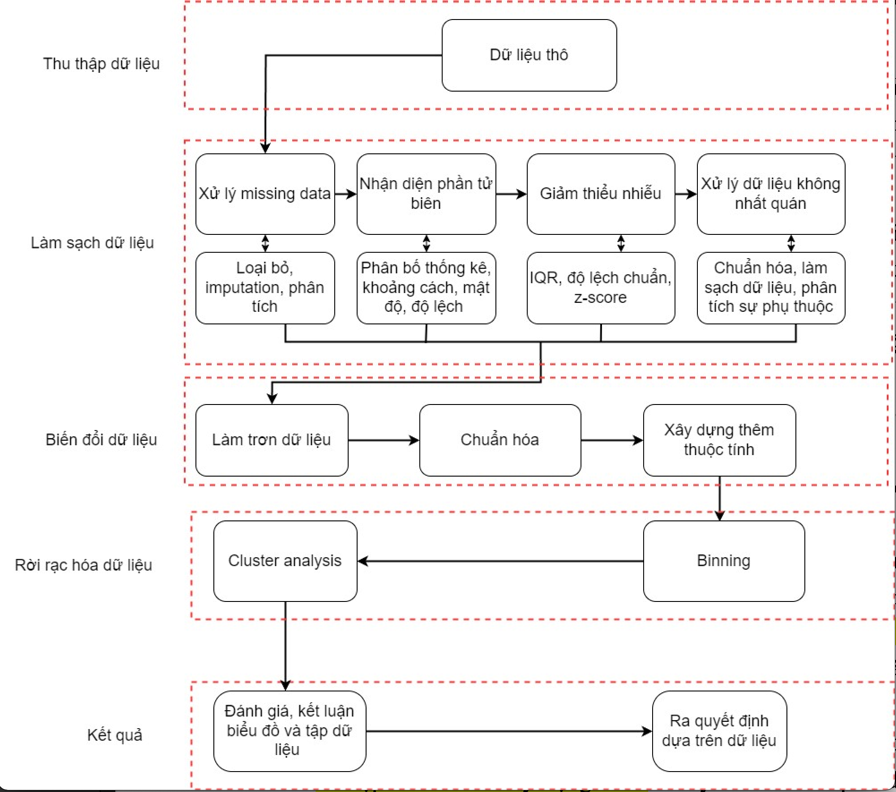
1. **Tóm Tắt:**

Trong nghiên cứu này, chúng ta tiến hành phân tích và xử lý dữ liệu thông qua một loạt các phương pháp thống kê và kỹ thuật làm sạch dữ liệu. Bằng cách sử dụng thống kê mô tả như số lượng, trung bình, độ lệch chuẩn, và các phân vị, chúng ta có thể hiểu rõ hơn về tính chất và cấu trúc của dữ liệu. Quá trình làm sạch dữ liệu bao gồm việc xử lý dữ liệu thiếu, nhận diện và giải quyết các phần tử biên, giảm thiểu nhiễu, và khắc phục sự không nhất quán. Sau đó, tiến hành biến đổi dữ liệu để chuẩn bị cho quá trình phân tích sâu hơn, bao gồm làm trơn dữ liệu, chuẩn hoá, và tạo ra các thuộc tính mới. Cuối cùng, rời rạc hoá dữ liệu là bước quan trọng giúp phân loại dữ liệu liên tục thành các nhóm rời rạc, tạo điều kiện cho việc phân tích ở các bước tiếp theo.

1. **Giới Thiệu:**

Đối với bất kỳ dự án phân tích dữ liệu nào, việc hiểu và xử lý dữ liệu đóng một vai trò quan trọng trong việc đảm bảo chất lượng và hiệu quả của kết quả nghiên cứu. Dữ liệu ban đầu thường không hoàn hảo và cần trải qua các bước làm sạch và biến đổi để phát triển các mô hình học máy và phân tích thống kê. Quy trình này không chỉ giúp loại bỏ nhiễu và sự không nhất quán mà còn giúp phát hiện 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 để cung cấp một nền tảng vững chắc cho việc phân tích dữ liệu tiếp theo. Qua đó, sẽ thấy rõ hơn về sự cần thiết của việc tiền xử lý dữ liệu, không chỉ trong lĩnh vực khoa học dữ liệu mà còn trong nhiều lĩnh vực khác.

1. **Quy trình nghiên cứu (Framework)**

****

1. **Cơ sở lý thuyết**
   1. **Thống kê mô tả**

Thống kê mô tả là bước đầu tiên và cơ bản trong quá trình phân tích dữ liệu, giúp chúng ta hiểu rõ hơn về bản chất và cấu trúc của dữ liệu thông qua các số liệu thống kê đơn giản.

* + 1. **Count**

"Count" đề cập đến việc đếm số lượng các mục không rỗng hay không null trong mỗi cột của bộ dữ liệu. Điều này cung cấp cái nhìn sơ bộ về kích thước và độ đầy đủ của dữ liệu, giúp phát hiện các cột có thể có số lượng lớn giá trị thiếu.

* + 1. **Mean**

"Mean" hay giá trị trung bình, được tính bằng cách cộng tổng tất cả các giá trị trong cột và chia cho số lượng các giá trị đó. Đây là một đại diện tính toán cho xu hướng trung tâm của một tập dữ liệu, thường được sử dụng để mô tả đặc điểm tổng quát của dữ liệu.

* + 1. **Std**

"Std" viết tắt của "Standard Deviation" hay độ lệch chuẩn, là một thước đo về mức độ phân tán của dữ liệu so với giá trị trung bình. Độ lệch chuẩn thấp cho thấy rằng các giá trị trong cột tập trung gần giá trị trung bình, trong khi độ lệch chuẩn cao cho thấy sự phân tán lớn.

* + 1. **Min**

"Min" chỉ giá trị nhỏ nhất tìm thấy trong mỗi cột. Đây là thông tin quan trọng vì nó cho thấy giới hạn dưới của dữ liệu và có thể giúp xác định các giá trị cực trị hoặc outliers.

* + 1. **Phân vị thứ nhất**

Phân vị thứ 25, hay tứ phân vị đầu tiên, là giá trị mà tại đó 25% dữ liệu nằm dưới nó. Đây là một cách đo lường sự phân phối của dữ liệu và cung cấp thông tin về sự tập trung dữ liệu ở phía dưới của phổ.

* + 1. **Trung vị**

Phân vị thứ 50, hay còn gọi là trung vị, là giá trị ở giữa khi dữ liệu đã được sắp xếp theo thứ tự. Trung vị không bị ảnh hưởng bởi các giá trị cực trị và thường được sử dụng như một đại diện tốt hơn cho xu hướng trung tâm khi dữ liệu không đối xứng.

* + 1. **Phân vị thứ ba**

Phân vị thứ 75, hay tứ phân vị thứ ba, là giá trị mà tại đó 75% dữ liệu nằm dưới nó. Điều này cung cấp cái nhìn về sự tập trung dữ liệu ở phía trên của phổ.

* + 1. **Max**

"Max" chỉ giá trị lớn nhất trong mỗi cột. Đây là thông tin cần thiết để hiểu về giới hạn trên của dữ liệu và cũng có thể dùng để phát hiện các giá trị cực trị hoặc outliers.

* 1. **Làm sạch dữ liệu**
     1. **Xử lý dữ liệu thiếu (Missing Data)**

Dữ liệu thiếu là thách thức thường gặp, có thể xảy ra do nhiều nguyên nhân như lỗi nhập liệu hoặc thiếu thông tin. Việc xử lý dữ liệu thiếu đòi hỏi việc đánh giá ngữ cảnh và mục đích sử dụng dữ liệu để áp dụng các phương pháp thích hợp như bỏ qua mục dữ liệu thiếu, điền giá trị thiếu bằng giá trị trung bình, trung vị hoặc áp dụng các thuật toán dự báo.

* + 1. **Nhận diện phần tử biên (Outliers)**

Phần tử biên là những quan sát có giá trị lệch lớn so với phần còn lại của dữ liệu, có thể là do lỗi đo lường hoặc biến động tự nhiên. Việc nhận diện và xử lý outliers là quan trọng vì chúng có th

ể làm méo các ước lượng thống kê và kết quả phân tích.

* + 1. **Giảm thiểu dữ liệu nhiễu (Noisy Data)**

Dữ liệu nhiễu bao gồm những biến động ngẫu nhiên hoặc các sai sót không hệ thống trong dữ liệu. Việc giảm thiểu nhiễu thường đòi hỏi việc áp dụng các phương pháp làm mịn dữ liệu, như bộ lọc trung bình, bộ lọc trung vị,…

* + 1. **Xử lý dữ liệu không nhất quán (Inconsistent Data)**

Dữ liệu không nhất quán có thể xuất hiện khi có sự không thống nhất trong cách ghi chép hoặc mã hóa dữ liệu. Xử lý dữ liệu không nhất quán thường yêu cầu việc phát hiện và sửa chữa các mâu thuẫn hoặc kết hợp dữ liệu từ nhiều nguồn khác nhau để đạt được sự nhất quán.

* + 1. **Giải pháp**
       1. **Giải pháp xử lý dữ liệu thiếu (Missing Data)**

Dữ liệu thiếu là tình trạng thông tin không đầy đủ, gây khó khăn trong phân tích và mô hình hóa. Các phương pháp để xử lý tình trạng này, bao gồm:

* **Loại bỏ**: Đơn giản nhất là loại bỏ những hàng hoặc cột chứa dữ liệu thiếu, phù hợp khi lượng dữ liệu thiếu không đáng kể.
* **Imputation**: Điền vào dữ liệu thiếu bằng cách sử dụng giá trị trung bình, trung vị, hoặc mốt, hoặc sử dụng các phương pháp phức tạp hơn như imputation đa biến.
* **Phân tích**: Đánh giá ảnh hưởng của dữ liệu thiếu đến kết quả phân tích để hiểu cách thức và mức độ mà nó có thể ảnh hưởng.
  + - 1. **Giải pháp nhận diện phần tử biên (Outliers)**

Phần tử biên là những quan sát lệch lạc so với phần còn lại của dữ liệu, có thể là do sai sót hoặc biến thể tự nhiên.

* **Phân bố thống kê**: Sử dụng các phép đo thống kê như Z-scores để đánh giá mức độ lệch của dữ liệu từ trung bình.
* **Khoảng cách**: Tính toán khoảng cách giữa các điểm dữ liệu và tìm kiếm những điểm cách biệt.
* **Mật độ**: Phân tích mật độ dữ liệu để xác định các vùng mật độ thấp, nơi outliers có thể tồn tại.
* **Độ lệch**: Đánh giá mức độ độ lệch của dữ liệu từ một mô hình dự đoán hoặc xu hướng cụ thể.
* **Quan sát boxplot**: Hiển thị phân phối của dữ liệu qua các quartiles và "râu" của boxplot thường được sử dụng để xác định giá trị ngoại lai. Những điểm nằm ngoài "râu" thường được coi là outliers.
  + - 1. **Giải pháp giảm thiểu nhiễu (Noisy Data)**

Dữ liệu nhiễu là những quan sát sai lệch không có mô hình nhất định, thường xuất hiện do sai sót đo lường hoặc ghi chép.

* **IQR (Interquartile Range)**: Sử dụng phạm vi giữa tứ phân vị thứ nhất và thứ ba để xác định và loại bỏ outliers.
* **Độ lệch chuẩn**: Loại bỏ dữ liệu nằm ngoài một số lượng độ lệch chuẩn nhất định từ trung bình.
* **Z-scores**: Dùng điểm Z để xác định dữ liệu có giá trị cao hoặc thấp bất thường so với trung bình.
  + - 1. **Giải pháp xử lý dữ liệu không nhất quán (Inconsistent Data)**

Dữ liệu không nhất quán xuất hiện khi có sự không đồng nhất về định dạng, kiểu dữ liệu, hoặc khi dữ liệu bị lặp lại.

* **Chuẩn hóa**: Đồng bộ hóa định dạng và kiểu dữ liệu để tất cả dữ liệu đều theo một tiêu chuẩn chung.
* **Làm sạch dữ liệu**: Áp dụng các quy trình để kiểm tra và sửa chữa những không nhất quán trong dữ liệu.
* **Phân tích sự phụ thuộc**: Xác định các mối quan hệ và sự phụ thuộc giữa các thuộc tính để phát hiện và giải quyết sự không nhất quán.
  1. **Biến đổi dữ liệu**

Biến đổi dữ liệu là một quá trình quan trọng nhằm định hình và điều chỉnh dữ liệu để nó có thể "phù hợp" tốt hơn với các thuật toán phân tích và mô hình học máy.

* + 1. **Chuẩn hoá (Normalization)**

Chuẩn hoá là quá trình điều chỉnh các biến số để chúng có cùng tỷ lệ hoặc phạm vi, làm cho quá trình phân tích trở nên dễ dàng hơn và cải thiện hiệu suất của các mô hình học máy.

* **Min-Max Scaling**: Chuyển đổi các giá trị để chúng nằm trong khoảng từ 0 đến 1.
* **Z-score Standardization**: Chuyển đổi các giá trị dựa trên độ lệch chuẩn, sao cho giá trị trung bình của thuộc tính là 0 và độ lệch chuẩn là 1.
  + 1. **Xây dựng thêm thuộc tính**

Tạo ra các thuộc tính mới từ dữ liệu hiện có, có thể thông qua các phép toán đơn giản như cộng, trừ hoặc thông qua các phép biến đổi phức tạp hơn như PCA (Phân tích thành phần chính) để nắm bắt thông tin quan trọng và giúp cải thiện hiệu suất mô hình.

* 1. **Rời rạc hoá dữ liệu**

Rời rạc hoá là quá trình chuyển đổi dữ liệu từ dạng liên tục sang dạng rời rạc, giúp đơn giản hóa phân tích và thường được sử dụng trong việc phân tích các mô hình học máy.

* + 1. **Binning**

Binning là kỹ thuật chia nhỏ dữ liệu liên tục thành các "bins" hoặc "buckets" dựa trên một số nguyên tắc nhất định, như chiều rộng bin đều hoặc bin có tần suất đều.

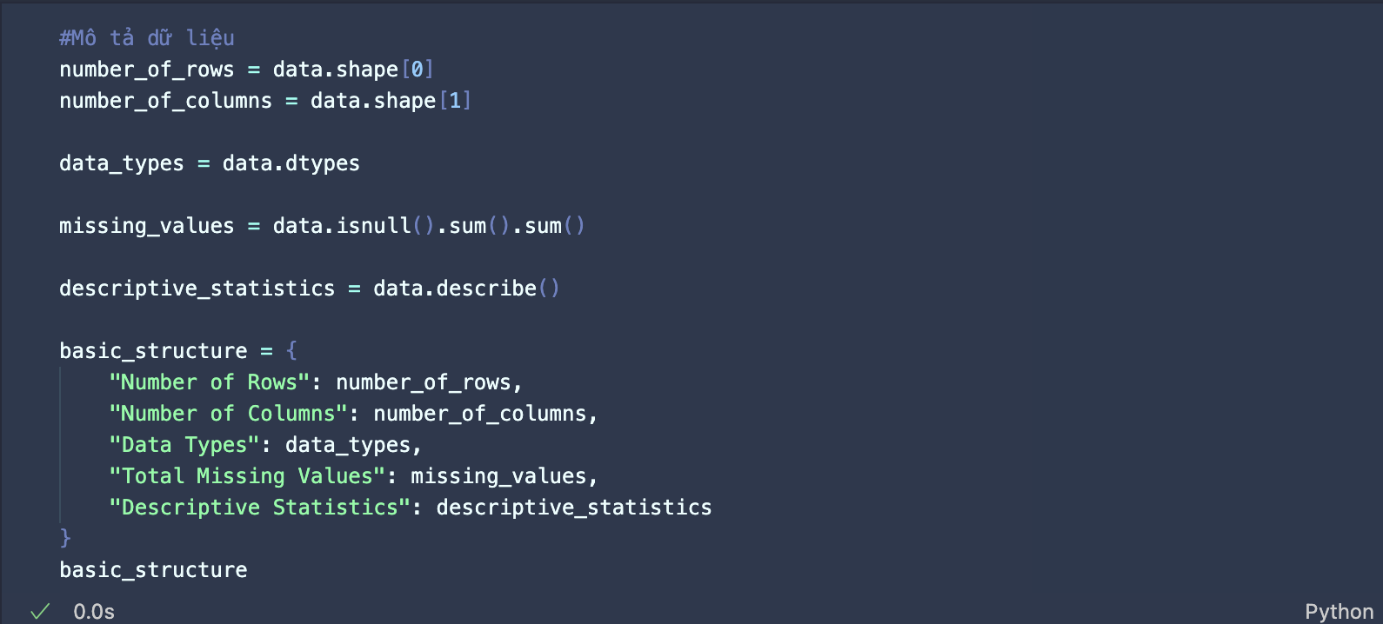
* + 1. **Cluster Analysis**

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

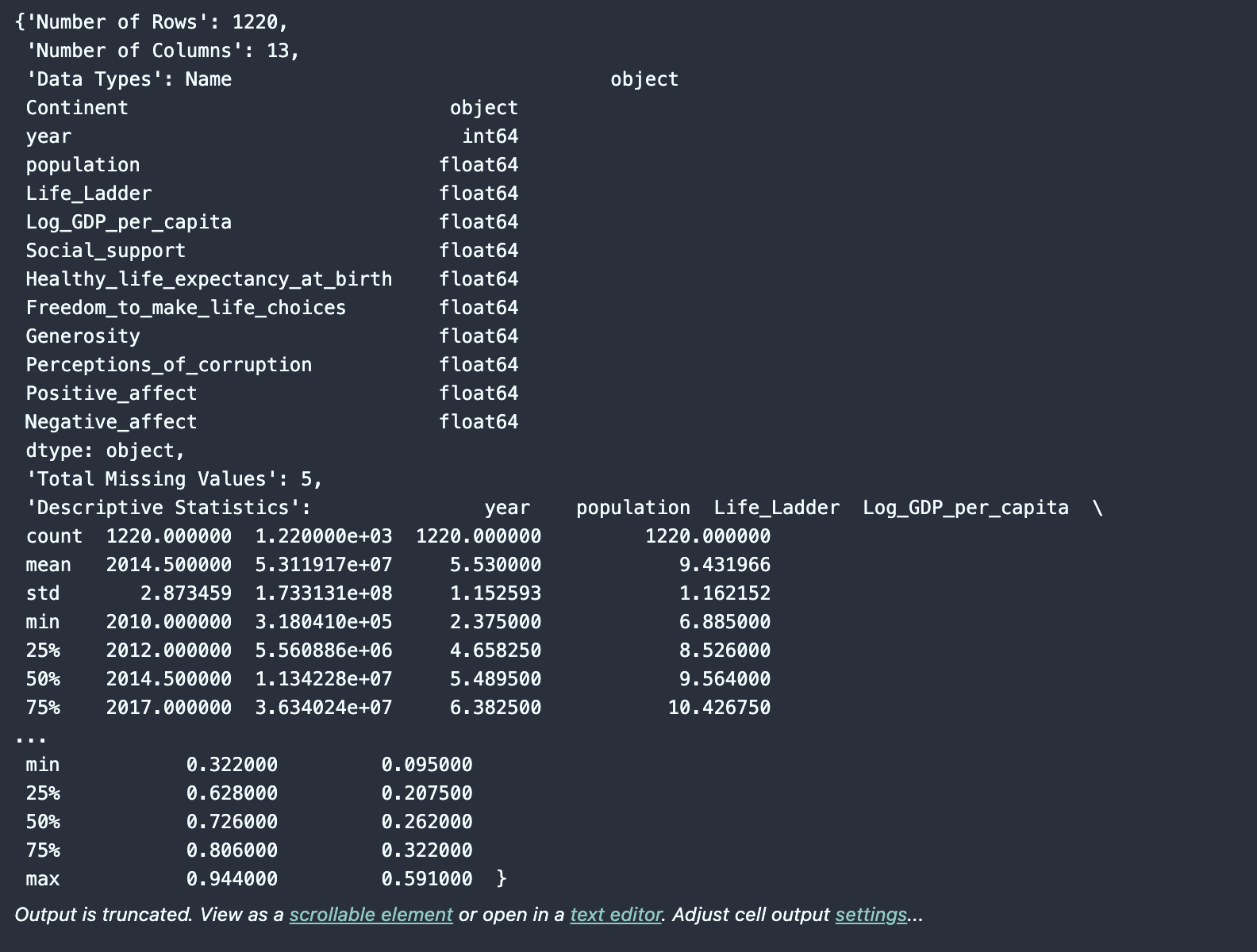
1. **Tổng quan, giới thiệu Dataset**

Bộ dữ liệu "Modified\_WH\_Report.csv" chứa thông tin từ các nước khác nhau theo từng năm, cung cấp cái nhìn toàn cảnh về các yếu tố ảnh hưởng đến mức độ hạnh phúc và phúc lợi xã hội. Nó bao gồm các chỉ số đo lường như GDP bình quân đầu người, sự hỗ trợ xã hội, tuổi thọ khỏe mạnh tại thời điểm sinh, tự do cá nhân, lòng hảo tâm, cảm nhận về mức độ tham nhũng, cùng với cảm xúc tích cực và tiêu cực. Đây là những chỉ số quan trọng giúp nghiên cứu và phân tích sâu hơn về chất lượng cuộc sống và mức độ hài lòng của người dân trong một quốc gia. Bộ dữ liệu này là nguồn thông tin quý giá cho các nhà nghiên cứu, chính trị gia, và các tổ chức quốc tế trong việc hình thành các chính sách và sáng kiến nhằm cải thiện phúc lợi xã hội.

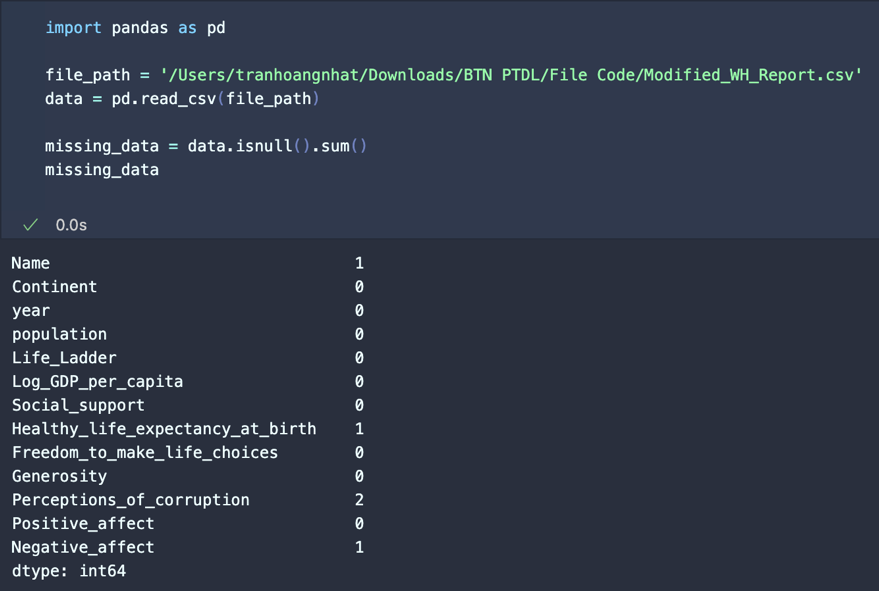
1. **Thực hành trên Dataset**
   1. **Mô tả Dataset**
      1. **Code**



* + 1. **Kết quả code**



* + 1. **Kết luận**
* **Số lượng hàng**: 1220
* **Số lượng cột**: 13
* **Kiểu dữ liệu của từng cột**:
  + Name: object (chuỗi)
  + Continent: object (chuỗi)
  + year: int64 (số nguyên)
  + population: float64 (số thực)
  + Life\_Ladder: float64 (số thực)
  + Log\_GDP\_per\_capita: float64 (số thực)
  + Social\_support: float64 (số thực)
  + Healthy\_life\_expectancy\_at\_birth: float64 (số thực)
  + Freedom\_to\_make\_life\_choices: float64 (số thực)
  + Generosity: float64 (số thực)
  + Perceptions\_of\_corruption: float64 (số thực)
  + Positive\_affect: float64 (số thực)
  + Negative\_affect: float64 (số thực)
* **Tổng số giá trị thiếu**: 5 giá trị
* **Thống kê cơ bản**:
  + year: Từ 2010 đến 2019
  + population: Trung bình là 53.119.170, với giá trị nhỏ nhất là 318.041 và lớn nhất là 1.397.715.000
  + Life\_Ladder: Trung bình là 5.53, với giá trị nhỏ nhất là 2.375 và lớn nhất là 7.858
  + Log\_GDP\_per\_capita: Trung bình là 9.432, với giá trị nhỏ nhất là 6.885 và lớn nhất là 11.648
  + Social\_support: Trung bình là 0.816, với giá trị nhỏ nhất là 0.303 và lớn nhất là 0.987
  + Healthy\_life\_expectancy\_at\_birth: Trung bình là 64.183 (có một giá trị thiếu)
  + Freedom\_to\_make\_life\_choices: Trung bình là 0.761, với giá trị nhỏ nhất là 0.304 và lớn nhất là 0.985
  + Generosity: Trung bình là 0.00246, với giá trị nhỏ nhất là -0.335 và lớn nhất là 0.698
  + Perceptions\_of\_corruption: Trung bình là 0.733 (có hai giá trị thiếu)
  + Positive\_affect: Trung bình là 0.714, với giá trị nhỏ nhất là 0.322 và lớn nhất là 0.944
  + Negative\_affect: Trung bình là 0.270, với giá trị nhỏ nhất là 0.095 và lớn nhất là 0.591 (có một giá trị thiếu)
  1. **Làm sạch dữ liệu**
     1. **Xử lý dữ liệu thiếu**
        1. **Quan sát dữ liệu thiếu**

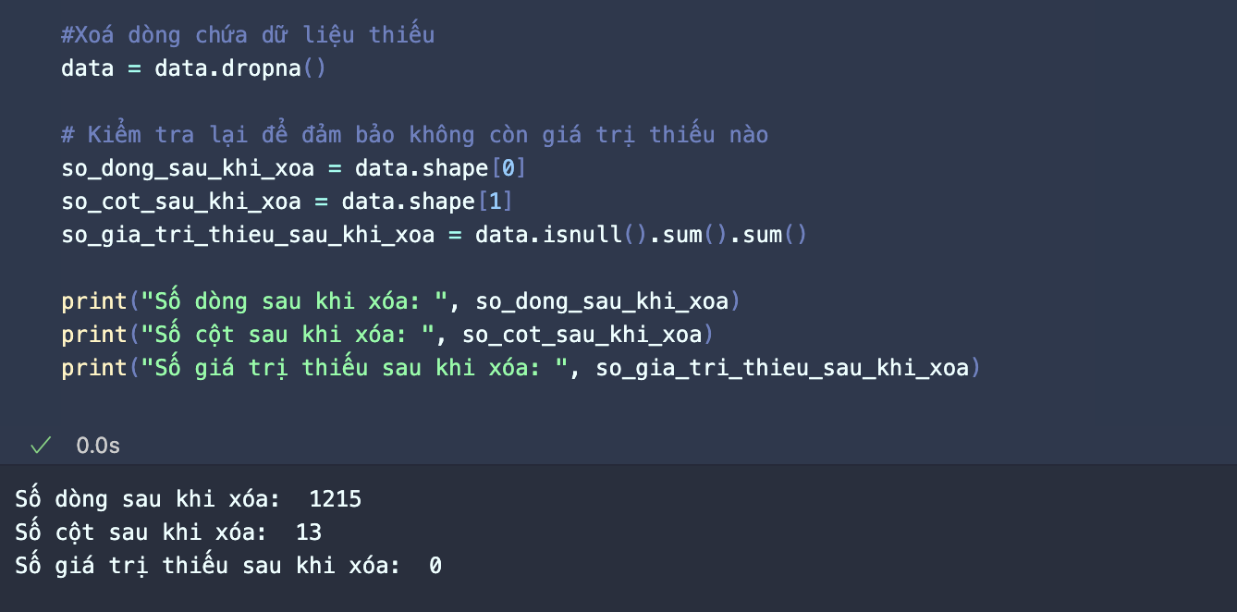
****

* + - 1. **Nhận xét**

Trong việc xử lý dữ liệu thiếu, việc quyết định giữa việc loại bỏ hoặc điền giá trị cho những dòng dữ liệu bị thiếu phụ thuộc vào ngữ cảnh và ý nghĩa của dữ liệu cũng như mục đích của việc phân tích. Trong trường hợp của bảng dữ liệu này, đây là dữ liệu thống kê qua từng năm và có thể chứa thông tin quan trọng về các chỉ số kinh tế và xã hội của quốc gia đó. Nếu các giá trị này bị thiếu, việc sử dụng các phương pháp như điền giá trị trung bình, trung vị, hoặc giá trị xuất hiện nhiều nhất có thể không phản ánh chính xác thực tế và có thể dẫn đến những hiểu lầm về xu hướng phát triển của quốc gia.

Ví dụ, việc điền giá trị GDP bình quân đầu người bằng giá trị trung bình có thể tạo ra những ước lượng không chính xác nếu có sự biến động lớn từ năm này sang năm khác. Điều này cũng tương tự với các chỉ số khác như tuổi thọ khỏe mạnh tại thời điểm sinh hay nhận thức về mức độ tham nhũng. Do đó, trong trường hợp này, việc loại bỏ các dòng dữ liệu bị thiếu là cách tiếp cận thận trọng nhất để bảo toàn tính chính xác của dữ liệu.

* + - 1. **Code**

****

* + - 1. **Kết luận**

Sau khi loại bỏ các dòng có dữ liệu thiếu, dataset giờ đây có:

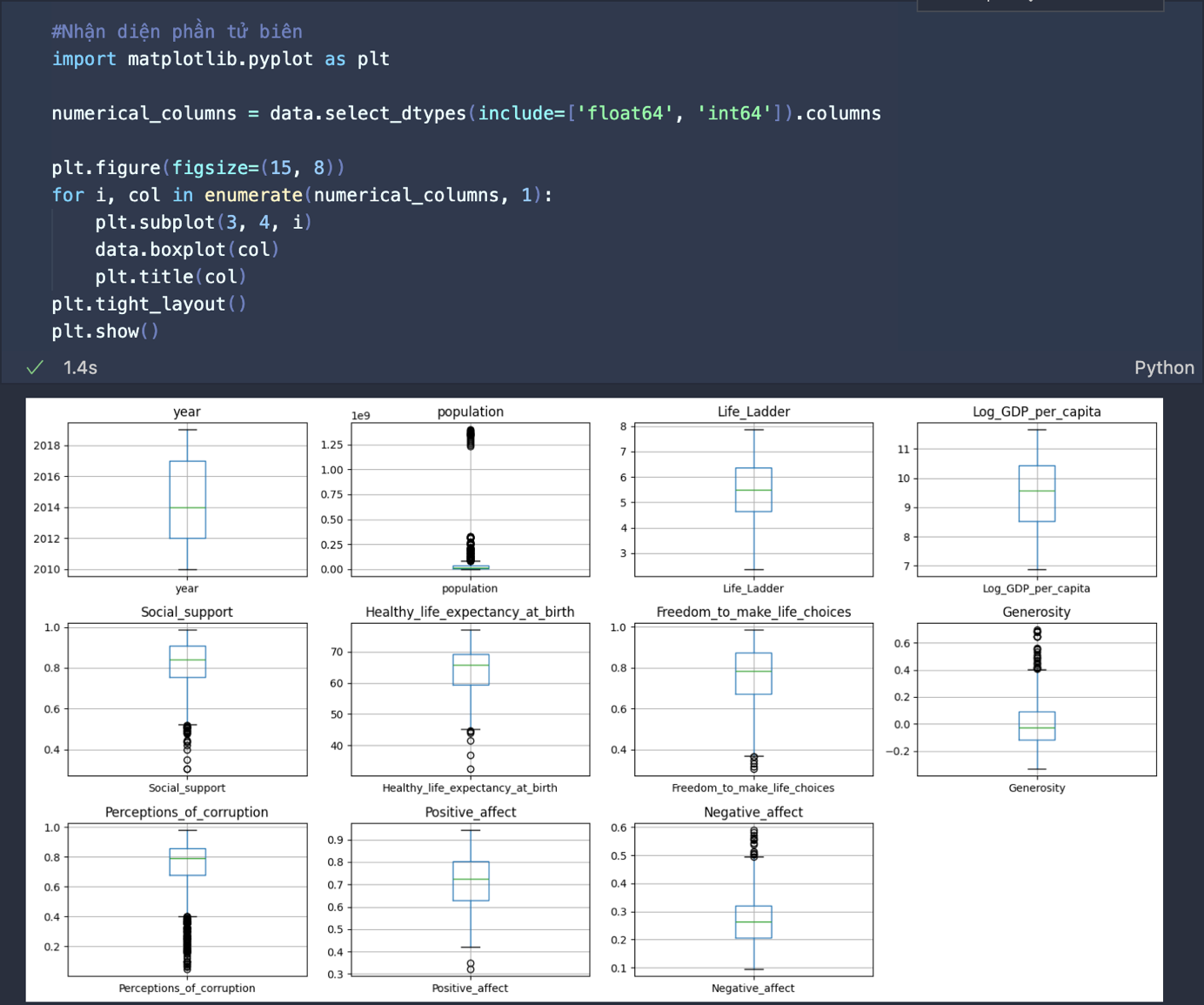
- Số lượng hàng: 1215 (giảm 5 hàng so với ban đầu)

- Số lượng cột: 13 (không thay đổi)

- Tổng số giá trị thiếu sau khi loại bỏ: 0

Dữ liệu đã được làm sạch và không còn chứa giá trị thiếu nào. Việc loại bỏ các dòng này giúp đảm bảo rằng phân tích sẽ được thực hiện trên dữ liệu đầy đủ và chính xác, phản ánh đúng thông tin thống kê qua các năm mà không bị ảnh hưởng bởi các giá trị được điền một cách tùy tiện.

* + 1. **Nhận diện phần tử biên**
       1. **Code và kết quả**



* + - 1. **Đánh giá kết quả**

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

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

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

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

• Các điểm nằm ngoài whiskers là các giá trị ngoại lệ (outliers).

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

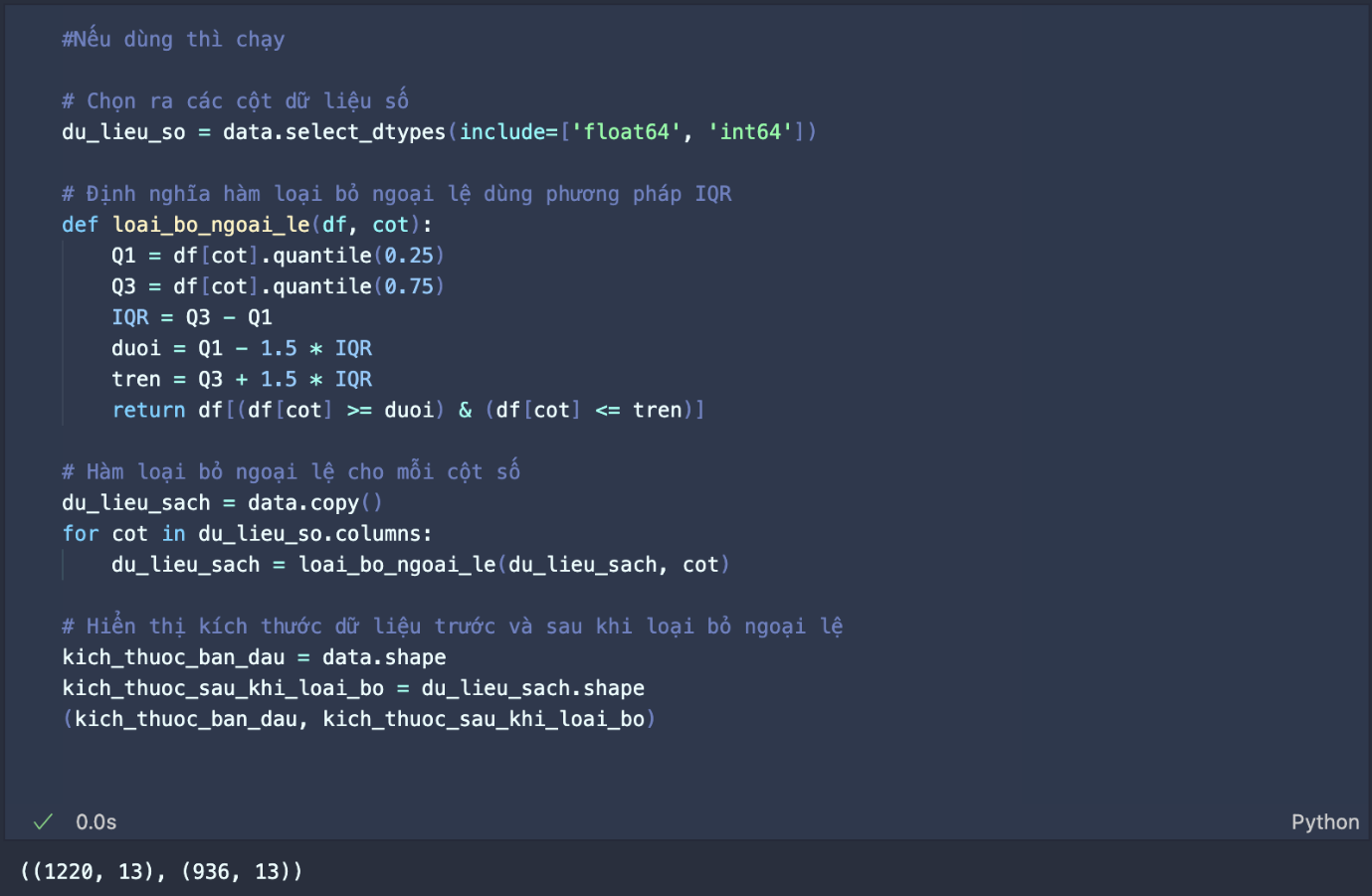
* + 1. **Giảm thiểu nhiễu**
       1. **Tổng quan:**

Nhận thấy dữ liệu bao gồm các thông số quan trọng và chính xác về dân số, kinh tế, sức khỏe, và xã hội của các quốc gia qua từng năm. Các outliers có thể phản ánh những tình huống đặc biệt hoặc những thay đổi đột ngột trong thực tế mà việc loại bỏ chúng có thể dẫn đến việc mất đi thông tin có giá trị.

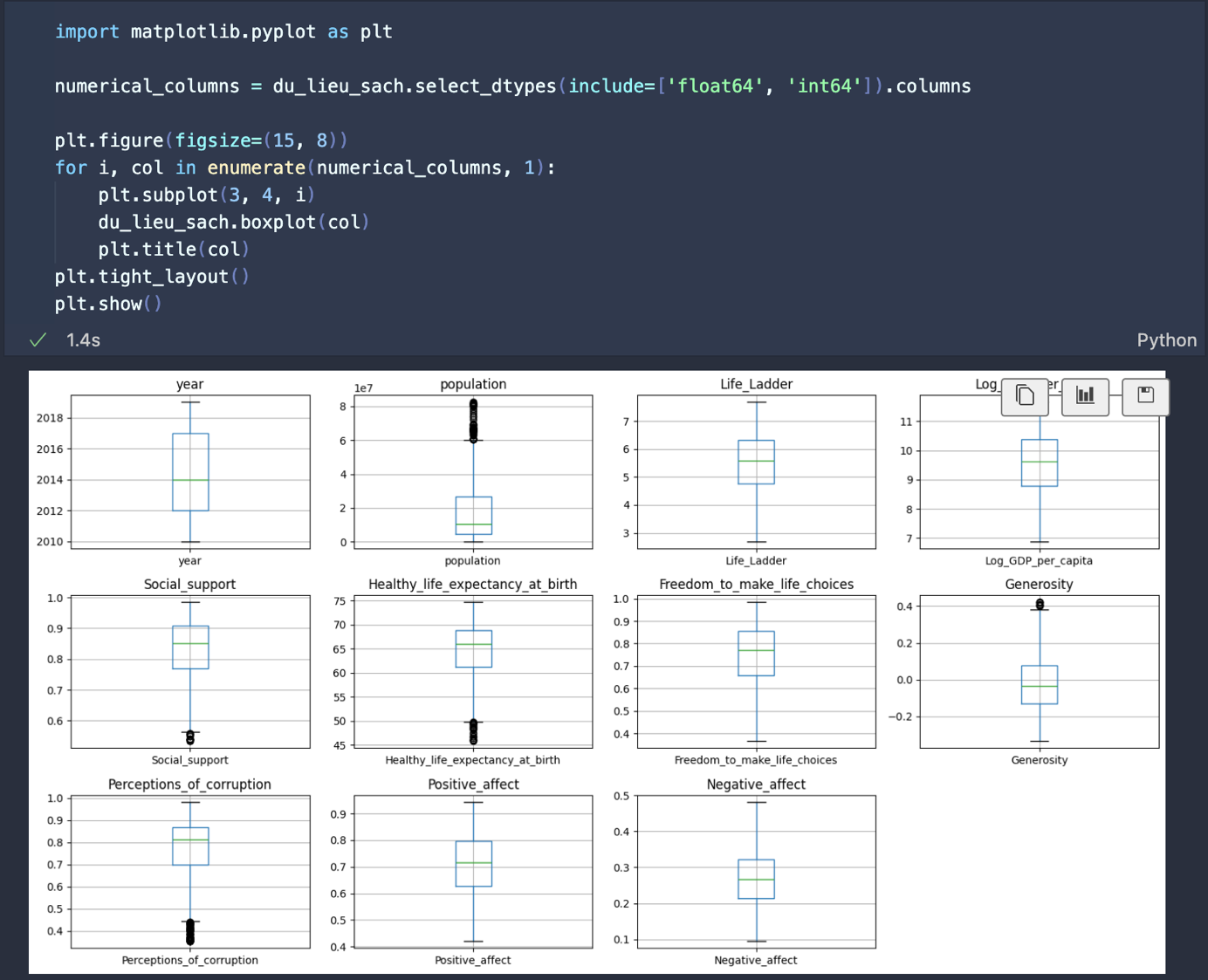
Ví dụ, một quốc gia có thể trải qua một sự kiện đặc biệt như khủng hoảng kinh tế hoặc thiên tai, dẫn đến sự thay đổi lớn trong GDP hoặc tuổi thọ. Trong khi đó, một quốc gia khác có thể có những chính sách xã hội đặc biệt thành công, dẫn đến các chỉ số hạnh phúc cao bất thường. Cả hai trường hợp này đều có thể xuất hiện như là outliers trong dữ liệu.

Loại bỏ các outliers mà không hiểu rõ nguyên nhân của chúng có thể làm méo mó những phân tích sau này và ảnh hưởng đến khả năng của chúng ta trong việc hiểu đúng về những gì đang xảy ra trong thế giới thực. Thay vào đó, những outliers này nên được phân tích kỹ lưỡng hơn để hiểu rõ nguyên nhân của chúng và cách chúng tác động đến các xu hướng tổng thể. Điều này có thể giúp chúng ta đưa ra những quyết định thông minh hơn dựa trên dữ liệu và có thể cung cấp những hiểu biết sâu sắc vào các yếu tố ảnh hưởng đến các chỉ số này.

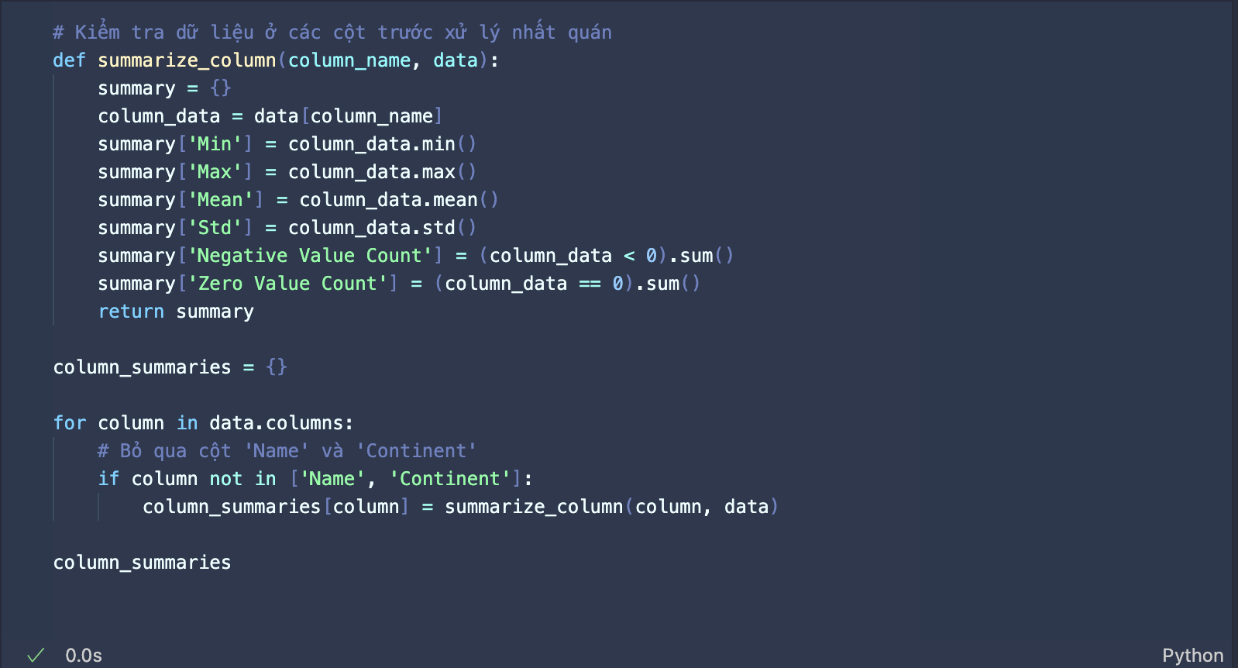
* + - 1. **Trường hợp vẫn loại outliers**

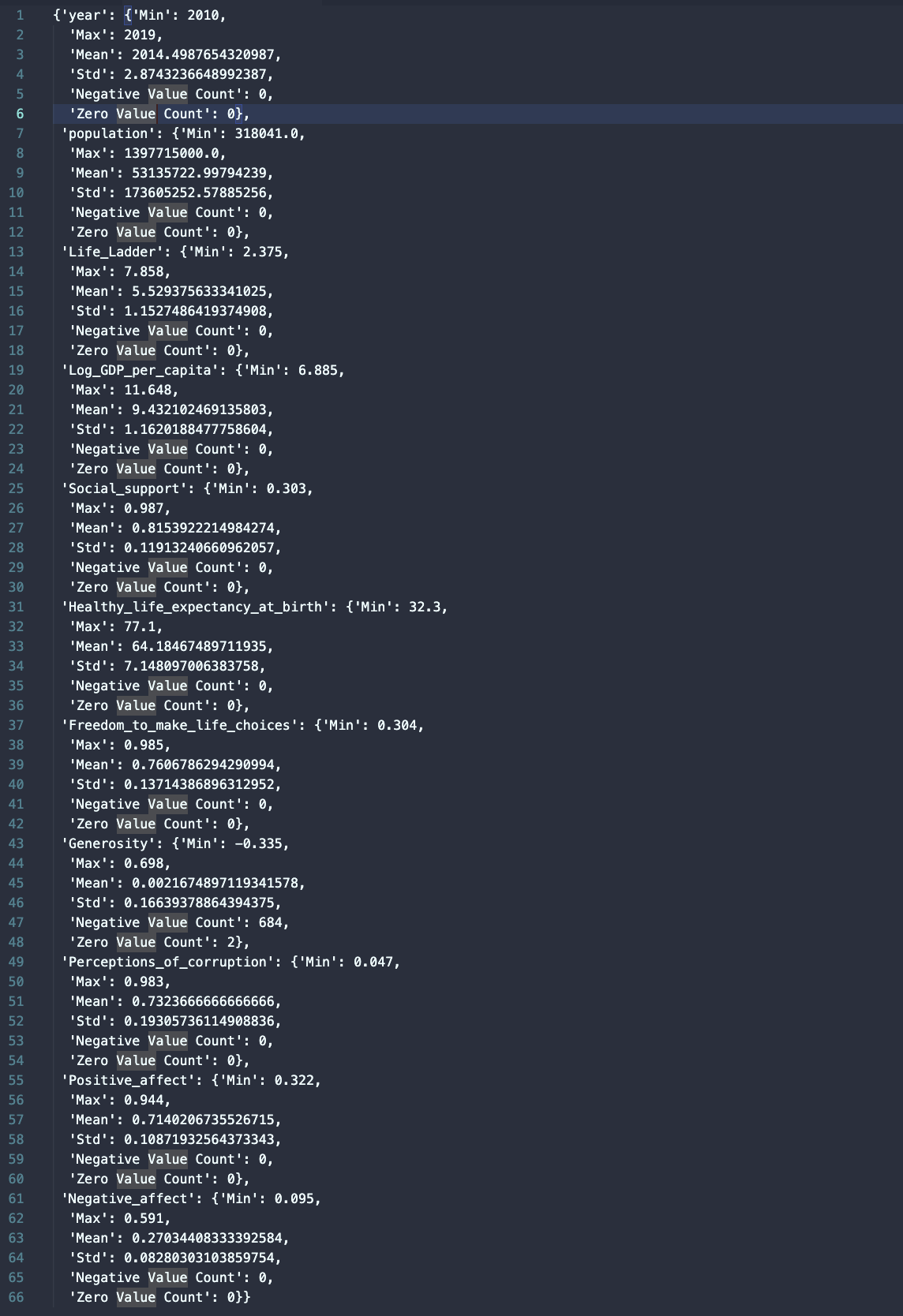
****

* + - 1. **Kết quả boxplot sau giảm nhiễu**

****

* + 1. **Xử lý dữ liệu không nhất quán**
       1. **Kiểm tra trước nhất quán dữ liệu**

****

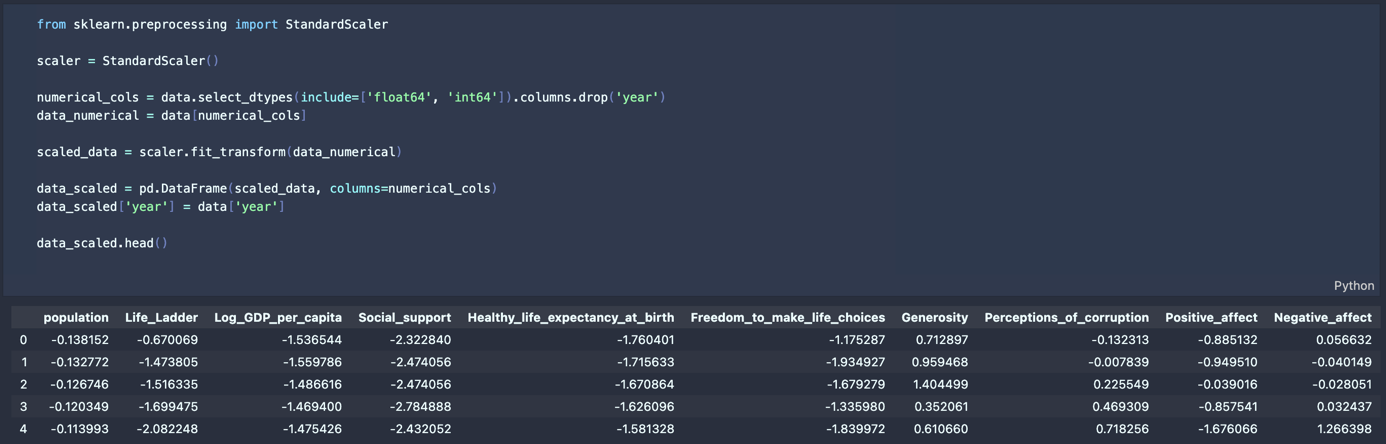
****

* + - 1. **Nhận xét**
* **Năm (year)**:

Giá trị từ 2010 đến 2019, phù hợp với một bộ dữ liệu mô tả thập kỷ đó.

Không có giá trị âm hoặc giá trị bằng không, điều này hợp lý vì cột này biểu thị năm.

* **Dân số (population)**:
  1. Khoảng từ 318.041 đến 1.397.715.000, phản ánh sự đa dạng về quy mô dân số giữa các quốc gia.
  2. Không có giá trị âm hoặc giá trị bằng không, điều này hợp lý vì dân số không thể âm hoặc không.
* **Chỉ số hạnh phúc (Life\_Ladder)**:
  1. Khoảng từ 2.375 đến 7.858, cho thấy sự biến thiên tự nhiên về mức độ hạnh phúc giữa các quốc gia.
  2. Không có giá trị âm hoặc bằng không, phù hợp vì chỉ số này thường được đo từ 1 đến 10.
* **GDP bình quân đầu người (Log\_GDP\_per\_capita)**:
  1. Khoảng từ 6.885 đến 11.648, không có giá trị âm, điều này hợp lý vì GDP không thể âm.
* **Hỗ trợ xã hội (Social\_support)**:
  + Khoảng từ 0.303 đến 0.987, không có giá trị âm hoặc bằng không, phù hợp với kỳ vọng về sự hỗ trợ xã hội.
* **Tuổi thọ khỏe mạnh tại thời điểm sinh (Healthy\_life\_expectancy\_at\_birth)**:
  + Khoảng từ 32.3 đến 77.1, không có giá trị âm hoặc bằng không, phù hợp với kỳ vọng về tuổi thọ khỏe mạnh.
* **Tự do lựa chọn cuộc sống (Freedom\_to\_make\_life\_choices)**:
  + Khoảng từ 0.304 đến 0.985, không có giá trị âm hoặc bằng không, phản ánh sự khác biệt về mức độ tự do cá nhân.
* **Hào phóng (Generosity)**:
  + Khoảng từ -0.335 đến 0.698, với 686 giá trị âm, điều này có thể phản ánh cách đo lường chỉ số này, nơi mà giá trị âm không phải lúc nào cũng chỉ ra điều tiêu cực.
* **Nhận thức về mức độ tham nhũng (Perceptions\_of\_corruption)**:
  + Khoảng từ 0.047 đến 0.983, không có giá trị âm hoặc bằng không, điều này hợp lý vì tham nhũng thường được đo từ 0 (không tham nhũng) đến 1 (cực kỳ tham nhũng).
* **Cảm xúc tích cực (Positive\_affect)** và **Cảm xúc tiêu cực (Negative\_affect)**:
  + Đối với cảm xúc tích cực: khoảng từ 0.322 đến 0.944.
  + Đối với cảm xúc tiêu cực: khoảng từ 0.095 đến 0.591.
  + Cả hai đều không có giá trị âm hoặc bằng không, phù hợp với kỳ vọng về biểu đồ cảm xúc.
    - 1. **Phương pháp xử lý**
* Dựa trên kết quả phân tích tính nhất quán và hợp lý của dữ liệu:
* Không có dấu hiệu của dữ liệu bất thường như giá trị âm hoặc bằng không ở những cột không mong đợi ngoại trừ cột 'Generosity', nơi giá trị âm có thể là một phần của cách thức đo lường và có thể không chỉ ra một vấn đề.
* Dữ liệu không có sự trùng lặp hoặc giá trị bất thường khác cần phải xử lý.
* Các cột số học có giá trị trong phạm vi hợp lý và không có bất kỳ giá trị nàocần phải được chuẩn hoá chỉ dựa trên giá trị âm hoặc bằng không.
* Như vậy, không cần thiết phải tiến hành chuẩn hoá dựa trên các kết quả phân tích trên.
  1. **Biến đổi dữ liệu**
     1. **Chuẩn hoá**
        1. **Phương pháp**
* **Chọn phương pháp Z – scores vì:**
* **Phạm vi và Độ lệch Chuẩn Khác Nhau:** Các đặc trưng trong bộ dữ liệu của bạn có phạm vi (min-max) và độ lệch chuẩn rất khác nhau, điều này có thể gây ra vấn đề trong việc so sánh hoặc kết hợp chúng trong một 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 Max-Min. Điều này đặc biệt quan trọng nếu bộ dữ liệu có ngoại lệ, mà chúng ta chưa kiểm tra.

****

* + - 1. **Kết quả**

Sau khi chuẩn hóa dữ liệu là một bảng dữ liệu mới với các giá trị đã được chuẩn hóa cho mỗi đặc trưng số (trừ 'year'), theo đó:

* **Trung bình (Mean)** của mỗi đặc trưng sau khi chuẩn hóa sẽ rơi vào khoảng 0, hoặc rất 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.

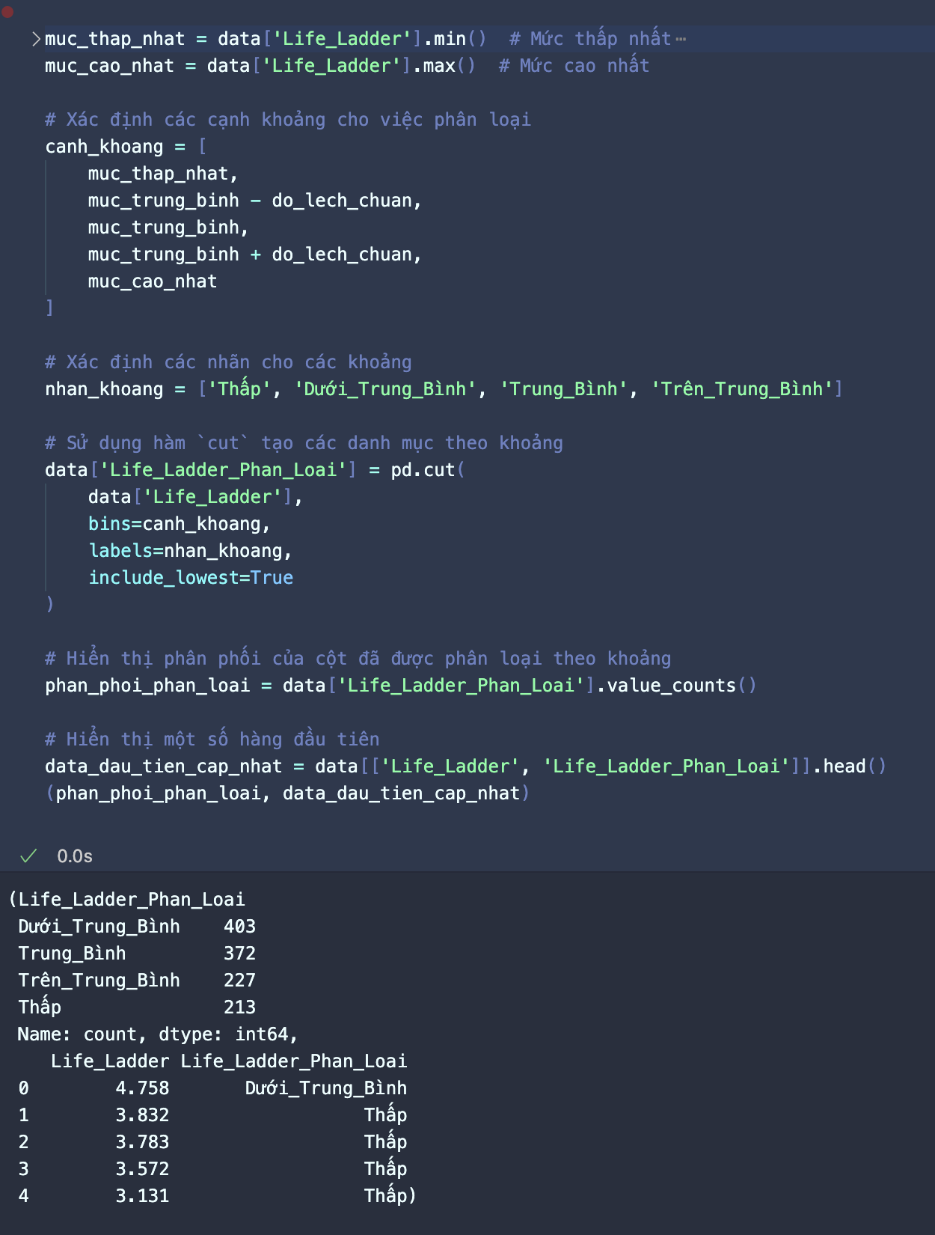
Cụ thể hơn:

* **population**: Giá trị này giờ đã được chuẩn hóa sao cho phạm vi của nó là xung quanh 0, giúp nó có thể so sánh công bằng với các đặc trưng khác.
* **Life\_Ladder**: Điểm số hạnh phúc cũng đã được điều chỉnh tương tự.
* **Log\_GDP\_per\_capita**, **Social\_support**, **Healthy\_life\_expectancy\_at\_birth**, **Freedom\_to\_make\_life\_choices**, **Generosity**, **Perceptions\_of\_corruption**, **Positive\_affect**, và **Negative\_affect**: Tất cả các đặc trưng này cũng đã được chuẩn hóa theo cùng một cách, làm cho chúng đều có trung bình ~0 và độ lệch chuẩn 1.

Đây là cách đặc trưng 'population' được chuyển từ phạm vi -0.138 (nhỏ hơn trung bình dân số) đến -0.113 (lớn hơn trung bình dân số) và đặc trưng 'Life\_Ladder' từ -0.670 (dưới trung bình) đến -2.082 (rất nhiều dưới trung bình).

Cột '**year**' không thay đổi vì nó không được chọn để chuẩn hóa; nó tiếp tục thể hiện năm mà dữ liệu được thu thập mà không bị ảnh hưởng bởi quá trình chuẩn hóa.

* + 1. **Xây dựng thêm thuộc tính**
* **Nhận xét:**Trong trường hợp của bộ dữ liệu này, không cần thiết phải xây dựng thêm thuộc tính nếu các yếu tố sau đây được đáp ứng.
* Bộ dữ liệu hiện tại đã chứa đầy đủ thông tin cần thiết cho mục tiêu phân tích hoặc mô hình học máy.
* Các thuộc tính như GDP bình quân đầu người, hỗ trợ xã hội, tuổi thọ khỏe mạnh, tự do lựa chọn, lòng hảo tâm, và cảm xúc tích cực/tiêu cực đã cung cấp một cái nhìn toàn diện về số liệu của mỗi quốc gia và khu vực.
* Việc thêm mới có thể không cần thiết và thậm chí có thể làm tăng độ phức tạp không cần thiết.
* Do đó, trong trường hợp bộ dữ liệu này đã đáp ứng đủ yêu cầu phân tích, việc giữ nguyên các thuộc tính hiện tại sẽ là quyết định hợp lý.
  1. **Rời rạc hoá dữ liệu**
* **Sử dụng phương pháp Binning cho cột Life\_Ladder**

****

* **Kết luận**
* Có tổng cộng 403 dữ liệu được phân loại vào khoảng "Dưới\_Trung\_Bình", đây là nhóm có số lượng dữ liệu nhiều nhất.
* Nhóm "Trung\_Bình" có 372 dữ liệu, xếp ở vị trí thứ hai về số lượng.
* Nhóm "Trên\_Trung\_Bình" có 227 dữ liệu.
* Nhóm "Thấp" có 213 dữ liệu, ít hơn so với các nhóm khác.

Nếu xem xét một số hàng đầu tiên của dữ liệu sau khi phân loại, ta thấy:

* Giá trị 'Life\_Ladder' đầu tiên là 4.758, được phân loại là "Dưới\_Trung\_Bình".
* Ba giá trị 'Life\_Ladder' tiếp theo là 3.832, 3.783, và 3.572, tất cả đều được phân loại vào nhóm "Thấp".
* Giá trị thứ năm là 3.131, cũng thuộc nhóm "Thấp".
* Những kết luận này phản ánh cách phân chia các giá trị 'Life\_Ladder' vào các khoảng dựa trên thống kê, và cho thấy sự phân bố của chỉ số hạnh phúc hoặc chất lượng cuộc sống trong bộ dữ liệu được xem xét.

1. **Nhận xét tổng quan bài nghiên cứu**

* Qua bài báo cáo nghiên cứu trên của nhóm, nhóm đã tiến hành một quy trình phân tích dữ liệu với đầy đủ tất cả các bước. Các bước từ làm sạch, biến đổi, cho đến rời rạc hóa dữ liệu đã được áp dụng một cách thận trọng và có chủ đích, nhằm đảm bảo rằng 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 dữ liệu mà còn đóng góp vào việc tăng cường tính khách quan và độ chính xác của kết quả nghiên cứu, từ đó tạo điều kiện để đưa ra những quyết định dựa trên dữ liệu một cách hiệu quả hơn.

1. **Nguồn tài liệu tham khảo**

* Slide của giảng viên Lê Diên Tuấn
* <https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801/full#:~:text=Data%20preprocessing%20refers%20to%20a,analysing%20massive%20building%20operational%20data>
* [Preprocessing Lecture Notes (chapter 3)](https://www3.cs.stonybrook.edu/~cse634/lecture_notes/07preprocessing.pdf)
* <https://www.studocu.com/row/document/wachemo-university/ntroduction-to-data-mining-and-data-warehousing/chapter-3-data-preprocessing/44326389>

Link video báo cáo: https://youtu.be/j1iUoIds4X0?si=lwn7URlIDFr20Ja4