**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**ĐẠI HỌC**

**TRƯỜNG CÔNG NGHỆ VÀ**

ĐỒ ÁN MÔN HỌC

**ĐỀ TÀI:**

**PHÂN TÍCH VÀ TRỰC QUAN HOÁ DỮ LIỆU TRONG NGÀNH BẢO HIỂM Y TẾ TẠI HOA KỲ**

**Học phần**: Trực Quan Hoá Dữ Liệu

**Nhóm Sinh Viên**:

**Chuyên Ngành**: Trí tuệ nhân tạo và Khoa Học Dữ Liệu

**K**

**Giảng Viên**: TS. Nguyễn An Tế

*T*

# MỤC LỤC

[MỤC LỤC 1](#_bookmark0)

[LỜI MỞ ĐẦU 3](#_bookmark1)

[CHƯƠNG I. TỔNG QUAN ĐỀ TÀI 4](#_bookmark2)

1. [Sơ lược về đề tài 4](#_bookmark3)
2. [Mục tiêu nghiên cứu 4](#_bookmark4)
3. [Phương pháp nghiên cứu 4](#_bookmark5)
4. [Tài nguyên sử dụng 4](#_bookmark6)

[CHƯƠNG II. TỔNG QUAN BỘ DỮ LIỆU BAN ĐẦU 5](#_bookmark7)

1. [Sơ lược về bộ dữ liệu 5](#_bookmark8)
2. [Mô tả thuộc tính của bộ dữ liệu 5](#_bookmark9)
3. [Mô tả tổng quan bộ dữ liệu ban đầu 6](#_bookmark11)
   1. *[Thông tin cơ bản 6](#_bookmark12)*
   2. *[Kiểm tra dữ liệu 6](#_bookmark14)*
   3. *[Trực quan hóa dữ liệu 8](#_bookmark17)*
   4. *[Kiểm tra sự tương quan](#_bookmark20)**[9](#_bookmark20)*

[CHƯƠNG III. CHỈNH DẠNG VÀ Ý NGHĨA CỦA BỘ DỮ LIỆU 11](#_bookmark22)

1. [Thay đổi ý nghĩa ban đầu của bộ dữ liệu 11](#_bookmark23)
2. [Chỉnh dạng cho bộ dữ liệu ban đầu 11](#_bookmark24)
   1. *[Thêm cột 11](#_bookmark25)*
   2. *[Đồng nhất kiểu dữ liệu các biến phân loại 14](#_bookmark27)*
   3. *[Thay đổi công thức tính 15](#_bookmark28)*
3. [Kiểm tra hiệu quả thay đổi biến 16](#_bookmark30)
4. [Kiểm tra Overfitting bằng Cross-Validation và Learning Curve 18](#_bookmark33)

[CHƯƠNG IV. TIỀN XỬ LÝ DỮ LIỆU 20](#_bookmark36)

1. [Mô tả tổng quan bộ dữ liệu sau thay đổi 20](#_bookmark37)
   1. *[Thông tin cơ bản 20](#_bookmark38)*
   2. *[Kiểm tra dữ liệu 20](#_bookmark40)*
   3. *[Trực quan hóa dữ liệu 22](#_bookmark45)*
   4. *[Kiểm tra sự tương quan](#_bookmark50)**[26](#_bookmark50)*
2. [Xử lý giá trị Outliers 35](#_bookmark58)

[CHƯƠNG V. KIỂM ĐỊNH 39](#_bookmark63)

1. [Phân tích hai biến (Bivariate Analysis) 39](#_bookmark64)
   1. *[Yếu tố Nhân khẩu học (Demographics) 39](#_bookmark65)*
   2. *[Yếu tố Địa lý (Geography) 41](#_bookmark67)*
2. [Phân tích đa biến (Multivariate Analysis) 44](#_bookmark70)

[CHƯƠNG VI. BIỂU DIỄN TRỰC QUAN DỮ LIỆU NGHIÊN CỨU 61](#_bookmark78)

1. [Xu hướng/Time series 61](#_bookmark79)
2. [Mật độ/Khu vực 65](#_bookmark86)
   1. *[Phân tích mật độ sử dụng và đóng bảo hiểm theo từng tiểu bang 66](#_bookmark87)*
   2. *[Phân tích độ tuổi theo từng khu vực 68](#_bookmark91)*
   3. *[Phân tích độ tuổi trung bình theo từng tiểu bang 69](#_bookmark93)*
   4. *[Phân tích số tiền đóng bảo hiểm theo từng bang trong 5 khu vực ở Hoa Kỳ](#_bookmark95)**[70](#_bookmark95)*
   5. *[Phân tích Đóng góp chi phí BHYT theo khu vực 71](#_bookmark98)*
   6. *[Tổng chi phí BHYT theo số lượng trẻ em trong gia đình ở mỗi khu vực 72](#_bookmark100)*
   7. *[Kết luận 73](#_bookmark102)*
3. [Sức khỏe & Gia đình 73](#_bookmark103)
   1. *[Tương quan giữa Tuổi, Chỉ số khối cơ thể và Chi phí BHYT 74](#_bookmark104)*
   2. *[Sự phân phối của Tổng chi phí BHYT dựa trên Tình trạng việc làm và Thói quen](#_bookmark106)* [hút thuốc của KH 75](#_bookmark106)
   3. *[Phân bố Dân số theo Nhóm tuổi và Tình trạng hôn nhân 75](#_bookmark108)*
   4. *[Kết luận 76](#_bookmark110)*
4. [Phân tích tổng hợp 77](#_bookmark111)

[CHƯƠNG VII. KẾT LUẬN 80](#_bookmark114)

[DANH MỤC BẢNG BIỂU 81](#_bookmark115)

[DANH MỤC HÌNH ẢNH 81](#_bookmark116)

[PHỤ LỤC 84](#_bookmark117)

1. [Link nguồn 84](#_bookmark118)
2. [Bảng phân công 84](#_bookmark119)

[TÀI LIỆU THAM KHẢO 85](#_bookmark120)

# LỜI MỞ ĐẦU

Trong thời đại ngày nay sự phát triển nhanh chóng của công nghệ thông tin và trí tuệ nhân tạo đã mở ra những cánh cửa mới cho việc ứng dụng dữ liệu trong nhiều lĩnh vực khác nhau, đặc biệt trong lĩnh vực Bảo hiểm Y tế (BHYT). Trong bối cảnh này, việc sử dụng phân tích trực quan dữ liệu trở thành một công cụ quan trọng, mang lại cái nhìn sâu sắc và toàn diện về hệ thống bảo hiểm sức khoẻ, phương pháp này không chỉ giúp hiểu rõ hơn về các xu hướng và biến động trong lĩnh vực BHYT mà còn đóng vai trò quan trọng trong việc đưa ra quyết định chiến lược và cải thiện chất lượng dịch vụ.

Mặc dù có nhiều tiềm năng, hiện vẫn không nhiều nghiên cứu tập trung vào những ảnh hưởng cụ thể của phân tích trực quan dữ liệu trong lĩnh vực này. Để đáp ứng thách thức này, nghiên cứu phân tích trực quan dữ liệu trong lĩnh vực BHYT này được tiến hành, với mong muốn làm rõ tầm quan trọng và xu hướng tương lai mà phân tích trực quan có thể mang lại cho lĩnh vực BHYT.

**Từ khoá**: Bảo hiểm Y tế (BHYT), Phân tích dữ liệu, Trực quan dữ liệu.

# CHƯƠNG I. TỔNG QUAN ĐỀ TÀI

## Sơ lược về đề tài

Trong bối cảnh số hóa đang biến đổi nhanh chóng hiện nay, ngành bảo hiểm đang gặp khó khăn trong việc thích nghi và hưởng lợi từ công nghệ mới so với các ngành khác, thậm chí trong lĩnh vực Tài chính - Bảo hiểm - Ngân hàng. Các bộ quy tắc đánh giá rủi ro cực kỳ phức tạp khác nhau trong các dòng sản phẩm, môi trường non-KYC (Know Your Customer) thiếu sự tập trung thông tin và mối quan hệ phức tạp với khách hàng (KH) trong quy trình đánh giá rủi ro truyền thống, nơi đôi khi sự tập trung vào KH chạy ngược lại với lợi nhuận kinh doanh cũng là thách thức đặc biệt mà ngành đang phải đối mặt.

Bộ dữ liệu gốc của bài toán này có thể là một bước đệm cơ sở đơn giản nhưng rất thực tiễn trong việc đánh giá tiềm năng và rủi ro Bảo hiểm Y tế (BHYT), tìm hiểu sự tương tác giữa các thuộc tính của KH mua bảo hiểm và sự ảnh hưởng đến chi phí bảo hiểm tại Hoa Kỳ.

## Mục tiêu nghiên cứu

Phân tích bộ dữ liệu trong lĩnh vực bảo hiểm có thể giúp các doanh nghiệp đánh giá rủi ro trong lĩnh vực bảo hiểm sức khỏe, hoặc phân tích sự ảnh hưởng của các yếu tố liên quan đến chi phí bảo hiểm, từ đó, đề xuất những chiến lược hiệu quả nhằm tối ưu hóa các quy trình và tạo ra các chính sách linh hoạt, đạt hiệu suất cao.

Ngoài ra, việc trực quan hóa dữ liệu từ việc phân tích máy móc cũng giúp dễ dàng đưa ra sự so sánh trực quan, tương tác, nhận biết những rủi ro và khả năng thuộc về xu hướng tương lai. Qua đó có thể hiểu rõ hơn về phân phối và các thông tin ẩn sau những dữ liệu khô khan, nhằm tạo tiền đề để đưa ra những quyết định mang tính chiến lược.

## Phương pháp nghiên cứu

Nhóm sử dụng nhiều phương pháp khác nhau linh hoạt theo nội dung nghiên cứu:

* + Chương 3: Mô hình Linear Regression và so sánh R2.
  + Chương 4: Phương pháp Univariate Analysis, Phương pháp kiểm tra giá trị nhiễu bằng Quy tắc 3-sigma và Khoảng tứ phân vị (Interquartile Range).
  + Chương 5: Phương pháp Bivariate Analysis và Multivariate Analysis, sử dụng kiểm định Z-test, Chi-square, One-way ANOVA.
  + Chương 6: Các loại biểu đồ đặc thù nhằm mục đích phân tích chi tiết theo từng nội dung.

## Tài nguyên sử dụng

* + Ngôn ngữ Python: sử dụng để phân tích và biểu diễn trực quan.
  + Bộ dữ liệu “US Health Insurance Dataset” từ Kaggle: *[kaggle.com/datasets](https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data)*

# CHƯƠNG II. TỔNG QUAN BỘ DỮ LIỆU BAN ĐẦU

**Mã nguồn**: [1-Preprocessing.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1rLvOZQ1JPHyaDAHJ0hqmuG4JBgYaXZYK)

## Sơ lược về bộ dữ liệu

**US Health Insurance EDA** được đăng tải trên Kaggle vào ngày 16/02/2020, bởi Anirban Datta – chuyên gia phân tích dữ liệu trong lĩnh vực bảo hiểm. Với mục đích thực hiện phân tích khám phá, bộ dữ liệu này tập trung vào các yếu tố gây ảnh hưởng đến chi phí Bảo hiểm Y tế (BHYT) tại Hoa Kỳ.

## Mô tả thuộc tính của bộ dữ liệu

|  |  |  |
| --- | --- | --- |
| **TÊN THUỘC TÍNH** | **MÔ TẢ** | **CHÚ THÍCH** |
| age | Độ tuổi của KH tham gia BHYT | Độ tuổi được thu thập giao động từ 18 đến 64. |
| sex | Giới tính của KH tham gia BHYT | Với **male** là nam giới và  **female** là nữ giới. |
| bmi | Chỉ số khối cơ thể của KH tham gia BHYT | * Sử dụng tỷ lệ đo khách quan giữa chiều cao và cân nặng (kg/m2) để xác định tình trạng cơ thể. * Trạng thái tiêu chuẩn: 18,5 < BMI < 24,9 |
| children | Số lượng con cái của KH tham gia BHYT |  |
| smoker | Tình trạng KH có hút thuốc hay không | Với **yes** là có hút thuốc và **no**  là không hút thuốc. |
| region | Khu vực cư trú của KH tham gia BHYT | Bao gồm 4 khu vực chính của Hoa Kỳ, lần lượt là:   * **northeast**: Đông Bắc * **northwest**: Đông Nam * **southeast**: Tây Nam * **southwest**: Tây Bắc |

|  |  |  |
| --- | --- | --- |
| chargers | Chi phí y tế cá nhân do BHYT thanh toán |  |

*Bảng 1: Bảng giải thích thuộc tính của bộ dữ liệu gốc*

## Mô tả tổng quan bộ dữ liệu ban đầu

### Thông tin cơ bản

Để hiểu rõ về bộ dữ liệu, nhóm cần biết những thông tin cơ bản như số lượng quan sát (dòng), số lượng biến (cột), và các thông tin mô tả về các biến trong bộ dữ liệu. Điều này giúp xác định quy mô của dự án và tạo nền tảng cho các quyết định về phân tích và xử lý dữ liệu.



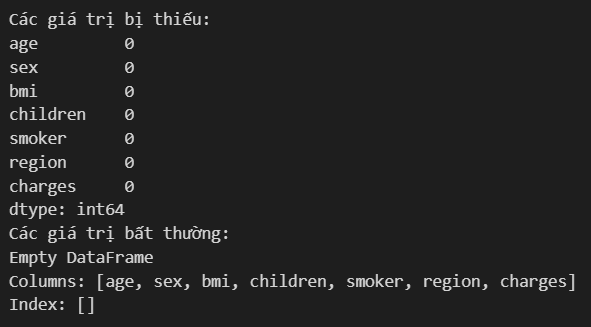
*Hình 1: Mô tả thông tin cơ bản của bộ dữ liệu ban đầu*

Từ kết quả ở trên, nhóm đã thu được thông tin rằng: bộ dữ liệu chứa tổng cộng 1.338 quan sát và 7 cột thuộc tính. Số lượng quan sát cho biết mức độ lớn của dữ liệu, trong khi số lượng cột cho biết có bao nhiêu biến được quan tâm. Sau khi hiểu rõ về quy mô này, nhóm sẽ tiếp tục thực hiện phân tích và xử lý dữ liệu.

### Kiểm tra dữ liệu

Đây là bước quan trọng trong quá trình tiền xử lý dữ liệu, giúp phát hiện và xử lý giá trị trống (missing values) cũng như các giá trị bất thường có thể ảnh hưởng đến kết quả của phân tích dữ liệu. Quá trình này đảm bảo tính nhất quán và độ tin cậy của dữ liệu trước khi tiến hành các bước phân tích tiếp theo.

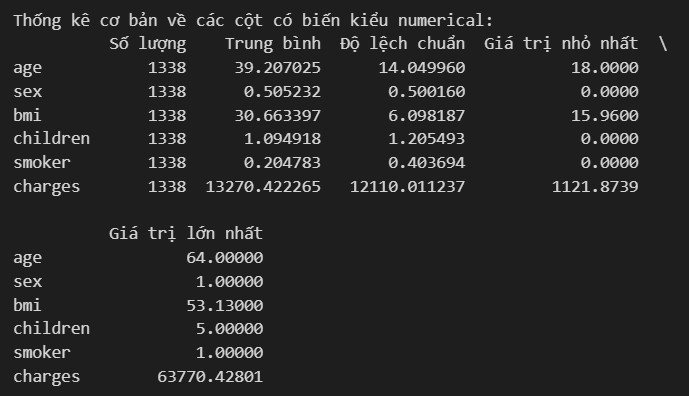
Trong bộ dữ liệu, mọi biến đều được quy định phải có giá trị dương. Do đó, bất kỳ giá trị nào xuất hiện dưới dạng số âm đều được coi là bất thường.



*Hình 2: Quan sát giá còn thiếu và giá trị bất thường của bộ dữ liệu ban đầu*

Nhìn chung, bộ dữ liệu này được xử lý khá tốt vì không có giá trị trống hoặc giá trị bất thường nào. Điều này giúp giảm bớt công đoạn tiền xử lý và làm cho việc phân tích dữ liệu trở nên dễ dàng hơn. Tuy nhiên, nhóm vẫn cần phải thực hiện các bước kiểm tra và phân tích dữ liệu cơ bản để đảm bảo chất lượng và tính chính xác của bộ dữ liệu.

Để hiểu rõ hơn về phân phối và biến đổi của dữ liệu, nhóm cần tiến hành phân tích thống kê tổng quát, bao gồm việc xác định giá trị trung bình và độ lệch chuẩn của dữ liệu.

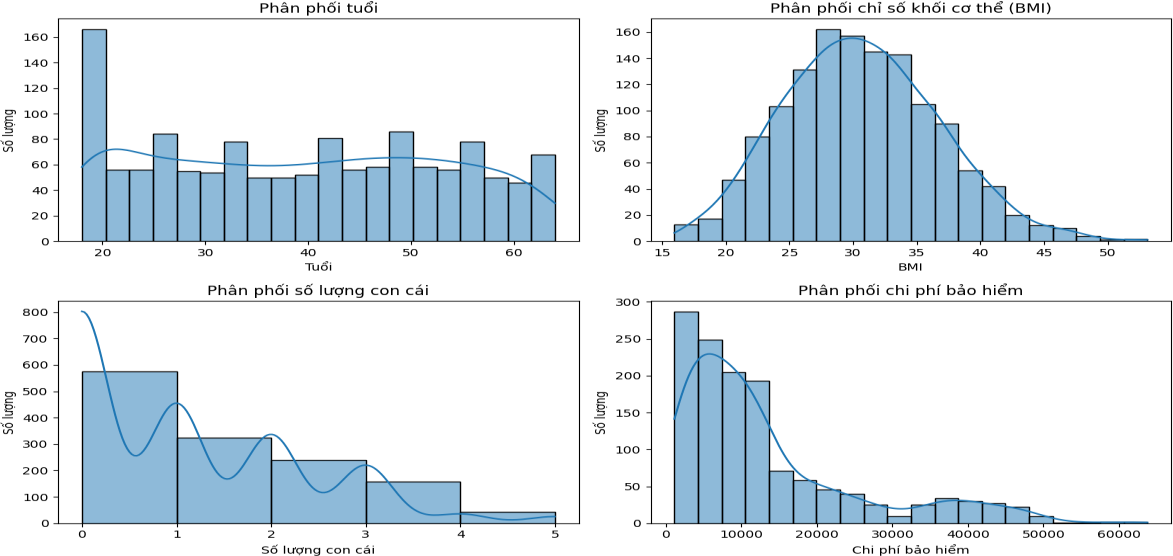


*Hình 3: Bảng thống kê tổng quát của bộ dữ liệu ban đầu*

Dựa trên kết quả thống kê trên, nhóm nhận thấy một số điểm quan trọng. Độ tuổi (age) có giá trị trung bình là khoảng 39 tuổi, với độ lệch chuẩn tương đối thấp, cho thấy sự phân phối gần với giá trị trung bình. Chỉ số khối cơ thể (BMI) có giá trị trung bình là khoảng 30, với độ lệch chuẩn cũng tương đối nhỏ. Số lượng con cái (children) trung bình thấp, và các khoản chi phí bảo hiểm (charges) có phạm vi biến đổi lớn, từ 1.121 đến 63.770 USD. Thông qua những thống kê cơ bản này, nhóm có cái nhìn tổng quan về sự biến đổi và tính chất cơ bản của các biến số học trong bộ dữ liệu.

### Trực quan hóa dữ liệu

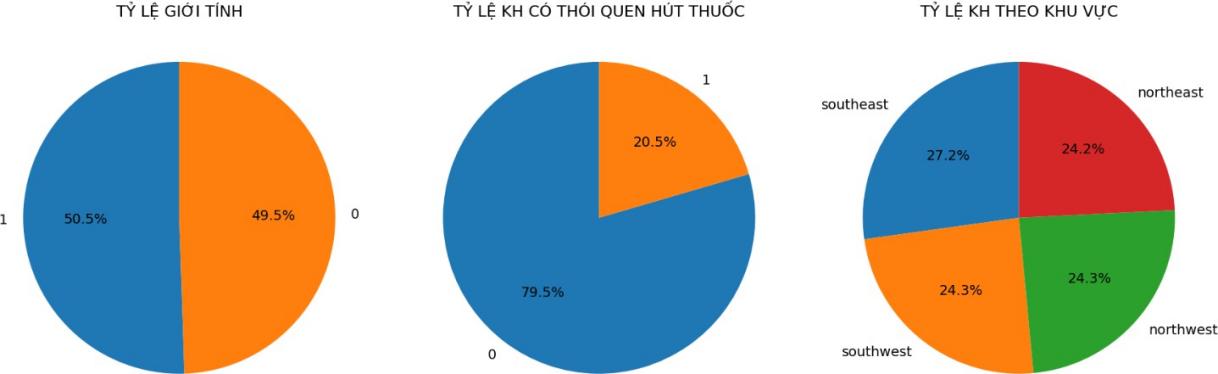
* + - Vẽ biểu đồ phân phối cho cột dữ liệu số (**age**, **bmi**, **children** và **charges**):



*Hình 4: Các biểu đồ phân phối cho cột dữ liệu số của dữ liệu gốc*

Nhận xét về dữ liệu:

* **age**: Phân bố tuổi của KH tham gia BHYT khá đều, với số lượng KH cao nhất ở độ tuổi 18 và 19, và thấp nhất ở độ tuổi 64. Điều này có thể do nhu cầu và khả năng thanh toán BHYT của các đối tượng khác nhau.
* **bmi**: Phân bố BMI của KH tham gia BHYT có hình dạng xấp xỉ chuẩn, với giá trị trung bình khoảng 30. Điều này có thể do đa số KH có BMI cao hơn mức tiêu chuẩn, gây ra các vấn đề sức khỏe và chi phí y tế cao hơn.
* **children**: Phân bố số lượng con cái của KH tham gia BHYT có xu hướng giảm dần, với số lượng KH không có con chiếm đa số, và số lượng KH có 5 con chiếm ít nhất. Điều này có thể do ảnh hưởng của các yếu tố như kinh tế, văn hóa, giáo dục, hay chính sách dân số.
* **charges**: Phân bố chi phí BHYT của KH có hình dạng lệch phải, với giá trị trung bình khoảng 13.000 USD. Điều này có thể do có một số KH có chi phí BHYT rất cao, gây ra sự chênh lệch so với đa số KH có chi phí BHYT thấp hơn.
  + - Vẽ biểu đồ tỷ lệ cho cột dữ liệu phân loại (**sex**, **smoker** và **region**):



*Hình 5: Các biểu đồ tròn cho cột dữ liệu phân loại của dữ liệu gốc*

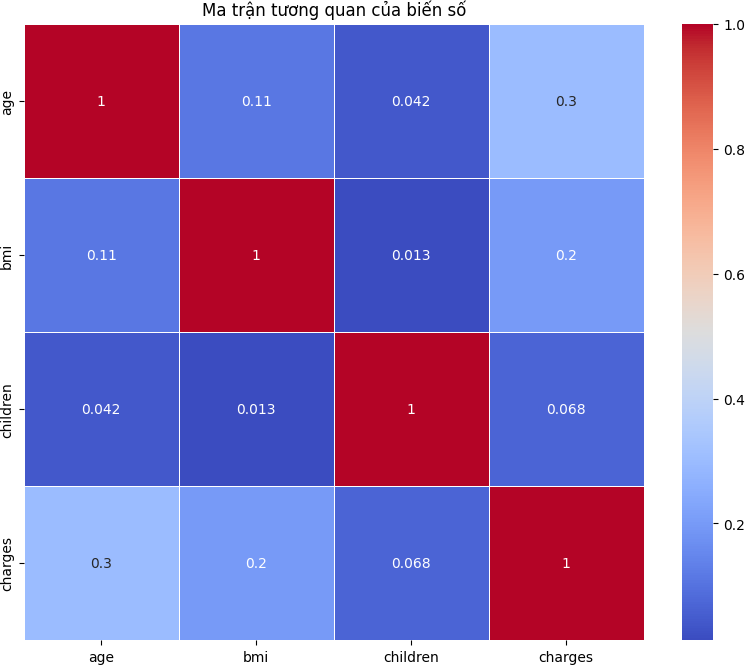
Nhận xét về dữ liệu:

* **sex**: Tỷ lệ giới tính của KH tham gia BHYT khá cân bằng, với 50,5% là nam giới và 49,5% là nữ giới. Điều này có thể do không có sự phân biệt giới tính trong việc cung cấp và sử dụng BHYT tại Hoa Kỳ.
* **smoker**: Tỷ lệ KH có hút thuốc là rất thấp, chỉ chiếm 20,5%, trong khi tỷ lệ KH không hút thuốc là 79,5%. Điều này có thể do nhận thức về tác hại của thuốc lá đối với sức khỏe và chi phí y tế, cũng như các chính sách hạn chế hút thuốc tại nhiều nơi công cộng.
* **region**: Tỷ lệ KH tham gia BHYT theo khu vực cũng khá đồng đều, với Southeast chiếm nhiều nhất (27,2%), tiếp theo là Northwest (24,3%), Southwest (24,3%) và Northeast (24,2%). Điều này có thể do sự phân bố dân số và mức độ phát triển kinh tế của các khu vực.

### Kiểm tra sự tương quan

* + - Vẽ biểu đồ thể hiện sự tương quan giữa các biến số trong bộ dữ liệu (**age**,

**bmi**, **children** và **charges**)



*Hình 6: Ma trận tương quan của biến số có trong bộ dữ liệu gốc*

Nhận xét chung về dữ liệu:

Biểu đồ cho thấy mối tương quan giữa các biến số **age**, **bmi**, **children** và **charges** là không cao. Nghĩa là, không có biến số nào hoàn toàn độc lập hay quyết định chi phí BHYT. Từ đó khó để có thể đưa ra được kết luận chung nhất cho dữ liệu, chính vì vậy mà nhóm đã thực hiện thay đổi ý nghĩa của bộ dữ liệu như phần trên. Tiếp theo, nhóm sẽ phân tích tổng quan bộ dữ liệu sau khi thay đổi.

# CHƯƠNG III. CHỈNH DẠNG VÀ Ý NGHĨA CỦA BỘ DỮ LIỆU

**Mã nguồn:** [1-Preprocessing.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1rLvOZQ1JPHyaDAHJ0hqmuG4JBgYaXZYK)

## Thay đổi ý nghĩa ban đầu của bộ dữ liệu

* + **Mục tiêu ban đầu**: Phân tích các yếu tố ảnh hưởng đến chi phí mà BHYT phải chi trả cho các khoản y tế cá nhân của khách hàng (KH).
  + **Mục tiêu mới**: Phân tích các yếu tố ảnh hưởng đến chi phí mà KH phải chi trả cho BHYT cá nhân.
  + **Lý do**: Thị trường BHYT tại Hoa Kỳ phức tạp và đa dạng. Việc phân tích xu hướng sử dụng BHYT sẽ giúp hiểu rõ hơn về mức phí y tế, nhu cầu của người dân, đối tượng có khả năng thanh toán và nguyên nhân khiến KH sẵn sàng trả mức phí cao hơn tại mỗi khu vực.
  + **Lợi ích**: Bài phân tích phần nào sẽ giúp các công ty bảo hiểm cung cấp được các giải pháp phù hợp tùy theo môi trường địa phương, tối ưu hóa chiến lược kinh doanh, cải thiện hiệu suất kinh doanh và đảm bảo lợi ích cho cả KH và công ty.

## Chỉnh dạng cho bộ dữ liệu ban đầu

### Thêm cột

|  |  |  |
| --- | --- | --- |
| **TÊN CỘT** | **Ý NGHĨA/ GIẢI THÍCH** | **CÁC XÁC ĐỊNH GIÁ TRỊ** |
| **CusID**  *(Mã KH)* | Định danh duy nhất cho mỗi quan sát, giúp dễ dàng theo dõi và quản lý dữ liệu. |  |
| **state**  *(Tiểu Bang)* | Chỉ rõ tác động của vị trí địa lý đến chi phí BHYT cá nhân. |  |
|  | Điều này rất quan trọng vì giá trị và mức độ tiếp cận dịch vụ y tế có sự biến đổi khác nhau ở các tiểu bang tại Hoa Kỳ. |
|  | Điều này giúp nhóm phát hiện xu hướng và khác biệt về chi phí bảo hiểm cũng như chăm sóc sức khỏe giữa khu vực thành thị và nông thôn. Từ đó cải thiện định hướng và đa dạng hóa sản phẩm bảo hiểm để phản ánh tốt hơn về nhu cầu của KH ở từng vùng. |

|  |  |  |
| --- | --- | --- |
| **married**  *(Tình trạng hôn nhân của KH: 1 – Đã kết hôn*  *hoặc 0 – Còn độc thân )* | Mục tiêu của việc thêm cột:  Liệu người đã kết hôn có những yếu tố nào khác ảnh hưởng đến chi phí bảo hiểm so với người độc thân hay không? | Lưu ý: Với những KH chưa kết hôn thì có số con trong gia đình (children) là 0 |
|  | Ngoài ra, thông tin này có thể hỗ trợ trong việc thiết kế các sản phẩm BHYT phù hợp với nhu cầu cụ thể của cả hai nhóm tình trạng hôn nhân. |  |
| **employ**  *(Tình trạng việc làm của KH: 1 - Có hoặc 0 - Không)* | Mục tiêu của việc thêm cột:  KH có công việc ổn định có chi phí bảo hiểm khác biệt như thế nào so với KH không có việc  làm? | Theo Bộ Thống kê Lao động Hoa Kỳ, việc KH có làm mới (hoặc đóng lại) BHYT hay không, có liên quan mật thiết đến tình trạng thất nghiệp. |
|  | Đồng thời, thông tin này cũng cũng cung cấp thêm về tác động của việc làm đến mức độ quan tâm của KH đối với BHYT trong việc duy trì sức khỏe lao động. | Xét trên trường hợp cụ thể, một KH 40 tuổi - độ tuổi trung niên, đã thực hiện làm mới BHYT 17 lần. Tuy nhiên, so với mức trung bình ~26 lần làm mới BHYT ở độ tuổi này, khả năng cao người này không thể tiếp tục đóng BHYT do thất nghiệp. |
|  |  | * Lưu ý, với KH dưới 24 và đang tham gia học tập tại các Trường Đại Học, Cao Đẳng thì vẫn được xem là KH có việc làm. |
| **most\_recent\_ insur\_pur**  *(Ngày mua bảo hiểm gần nhất của KH)* | Cột này sẽ ghi lại ngày mà KH mua bảo hiểm gần nhất.  Thông tin này sẽ hỗ trợ việc phân tích liệu thời điểm mua bảo hiểm có ảnh hưởng đến chi phí bảo hiểm của KH hay không, và xem xét sự tương quan giữa thời gian mua bảo hiểm và tình trạng | Tại Hoa Kỳ, có hai nhóm người chia thành hai mức độ quan tâm đối với BHYT:   * Mức độ quan tâm thấp: được xác định dựa trên mức độ tuổi và mốc thời gian BHYT |

|  |  |  |
| --- | --- | --- |
|  | sức khỏe của KH nhằm hiểu rõ về nhu cầu của họ. | được phổ biến với người dân Hoa Kỳ (2005). |
|  | * Nhóm người già (50-64 tuổi) có số năm giao động từ 2005 đến 2010. * Trung niên (30-49 tuổi) giao động từ 2008 đến 2017. * Vị thành niên (18-29 tuổi) giao động từ 2018 đến 2020. |
|  | * Mức độ quan tâm cao: Tương tự với cách xác định như vậy nhưng giới hạn trên nằm phải ở năm 2022 và giới hạn dưới là 2017 với nhóm người có mức độ quan tâm cao. |
|  | * Nhóm người già (50-64 tuổi) có số năm giao động từ 2017 đến 2022. * Trung niên (30-49 tuổi) giao động từ 2018 đến 2022. * Vị thành niên (18-29 tuổi) giao động từ 2021 đến 2022. |
|  | * Lưu ý rằng tại các khu vực trung tâm, hơn 80% dân số tham gia BHYT. |
| **insur\_renewals** *(Số lần đóng lại BHYT của KH)* | Việc đánh giá mối liên hệ giữa tần suất gia hạn bảo hiểm và chi phí BHYT sẽ giúp xác định được mức độ quan tâm của KH đối với BHYT.  Điều này có thể được phân tích theo từng bang ở Hoa Kỳ, độ | Dựa trên các khảo sát của ContactEngine và Embroker, việc làm mới BHYT được xác định theo các yếu tố sau:   * Nếu KH hút thuốc, có BMI thấp hơn mức trung bình và quan tâm về y tế thì số lần |

|  |  |  |
| --- | --- | --- |
|  | tuổi, hoặc bất kỳ yếu tố nào khác có thể ảnh hưởng đến chi tiêu y tế cá nhân của KH. Từ đó, hỗ trợ cho việc tối ưu hóa kế hoạch bảo hiểm và quản lý chi phí, mang lại lợi ích cho cả KH và công ty bảo hiểm. | làm mới bảo hiểm bằng 85% số lần của độ tuổi của họ (*nếu KH 30 tuổi thì số lần làm mới BHYT của KH là*  *~26 lần*).   * Nếu KH là nữ và có ít nhất 2 con trở lên, và quan tâm về y tế thì số lần làm mới BHYT bằng 72%. * Nếu KH ở trong khu vực trung tâm và quan tâm về y tế thì số lần làm mới BHYT bằng 68%. * Nếu KH không quan tâm đến y tế, không hút thuốc và có BMI trung bình thì số lần làm mới bảo hiểm bằng 12%. * Nếu KH không quan tâm đến y tế và có nhiều con, số lần làm mới bảo hiểm bằng 27% số lần của độ tuổi của họ. |

*Bảng 2: Bảng giải thích ý nghĩa thêm thuộc tính cho bộ dữ liệu*

### Đồng nhất kiểu dữ liệu các biến phân loại

* + **Nhóm thực hiện một số thay đổi**:
    - Tên cột **children** đổi thành **num\_of\_children** nhằm hiểu rõ hơn về ý nghĩa của biến này.
    - Tên cột **sex** đổi thành is\_**male**, đồng thời thay thế giá trị *male* và *female* của cột lần lượt sang giá trị *1* và *0*.
    - Cột **smoker** được thay thế giá trị *yes* và *no* của lần lượt thành *1* và *0*.
  + **Ý nghĩa**:
    - Việc thay đổi trên giúp công đoạn biểu diễn dữ liệu dễ dàng hơn cho mô hình hóa, tương quan và trực quan hóa dữ liệu. Những biến số được số hóa này hỗ trợ tích hợp với thuật toán và phân tích dữ liệu, đồng thời giúp hiểu rõ mối quan hệ giữa giới tính, tình trạng hút thuốc và các biến khác trong tập dữ liệu một cách đồng nhất.

### Thay đổi công thức tính

Với mục tiêu mới của bộ dữ liệu, biến **charges** giờ đây biểu diễn cho tổng số tiền đã thanh toán cho BHYT của KH. Nhóm thiết lập một số điều kiện sau khi thực hiện tính toán cho cột này:

* + Mức chi phí bảo hiểm trung bình hằng năm ở các bang tại Hoa Kỳ có sự khác biệt rõ rệt, thể hiện qua bảng như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| **TIỂU BANG** | **CHI PHI TRUNG BÌNH USD/NĂM** | **TIỂU BANG** | **CHI PHI TRUNG BÌNH USD/NĂM** |
| Nam Dakota | 9.972 | Illinois | 6.672 |
| Wyoming | 9.168 | Delaware | 6.660 |
| Vermont | 9.120 | New Jersey | 6.444 |
| Louisiana | 8.736 | Massachusetts | 6.420 |
| Alaska | 8.580 | Kansas | 6.408 |
| New York | 8.556 | Lowa | 6.396 |
| Nebraska | 8.220 | North Dakota | 6.288 |
| Oklahoma | 7.620 | Idaho | 6.192 |
| Bắc Carolina | 7.608 | Wisconsin | 6.168 |
| Missouri | 7.440 | Virginia | 6.144 |
| Florida | 7.020 | Mississippi | 6.132 |
| Alabama | 6.948 | California | 6.444 |
| Nevada | 6.936 | Tennessee | 6.096 |
| Arizona | 6.924 | Pennsylvania | 5.976 |
| Texas | 6.900 | Hawaii | 5.880 |
| Connecticut | 6.768 | Ohio | 5.880 |
| Utah | 6.756 | New Mexico | 5.760 |
| Montana | 5.748 | Michigan | 4.920 |
| Kentucky | 5.736 | Colorado | 4.908 |

|  |  |  |  |
| --- | --- | --- | --- |
| Oregon | 5.700 | Minnesota | 4.668 |
| Maine | 5.580 | Maryland | 4.380 |
| Washington | 5.316 | New Hampshire | 4.320 |
| South Carolina | 5.232 | Georgia | 3.708 |
| Indiana | 5.196 | Arkansas | 5.028 |
| Rhode Island | 4.956 | *Các bang còn lại* | *6.342* |

*Bảng 3: Bảng thể hiện chi phí bảo hiểm trung bình hằng năm theo bang ở Hoa Kỳ [12]*

Tuy nhiên, theo nghiên cứu của Cleveland Clinic, KH có lối sống hút thuốc thì mức phí phải đóng cho BHYT cao hơn những KH không có thói quen sử dụng thuốc lá. Để xác định được loại mức phí tăng cường đó, nhóm thực hiện lấy:

Trung bình mức phí BHYT + Khoảng chênh lệch

thường niên theo bang

theo sai số của nó

* **Xác định công thức tính giá trị cho cột charges**:

Số lần làm mới BHYT

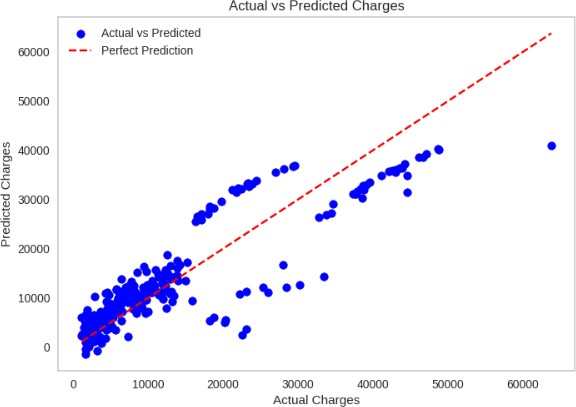
Mức phí BHYT trung bình Phí tăng cường

× thường niên theo bang + (Nếu KH có hút thuốc)

## Kiểm tra hiệu quả thay đổi biến

Ở phần này, nhóm sử dụng mô hình hồi quy tuyến tính để thực hiện phân tích việc thay đổi những biến này sự gia tăng hiệu quả như thế nào.

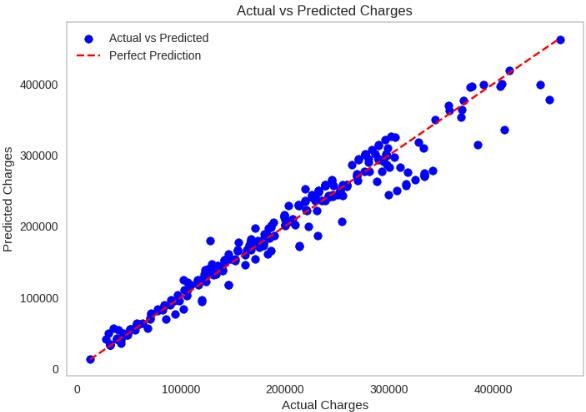
* + **Mục tiêu**: Kiểm tra việc thay đổi các biến độc lập trong bộ dữ liệu có làm tăng hiệu quả của mô hình hồi quy tuyến tính hay không.
  + **Phương pháp**: So sánh hệ số R2 và biểu đồ giữa giá trị thực tế và giá trị dự đoán của mô hình, đối với hai bộ dữ liệu: bộ dữ liệu ban đầu và bộ dữ liệu sau khi thêm các biến mới như **married**, **employ**, **insur\_renewals**.
  + **Kết quả**:
    - Với bộ dữ liệu ban đầu:



*Hình 7: Biểu đồ phân tán giữa giá trị thực tế và dự đoán của bộ dữ liệu ban đầu*

Với hệ số xác định 𝑹𝟐 = 𝟎, 𝟕𝟖𝟏𝟏 nghĩa là sau khi thực hiện hồi quy tuyến tính, mô hình giải thích được khoảng 𝟕𝟖, 𝟏𝟏% sự biến thiên của biến **charges**. Ngoài ra, từ biểu đồ cho thấy dữ liệu xuất hiện nhiều giá trị nằm xa đường hoàn hảo, tức sai số giữa dự đoán và giá trị thực tế cao.

* + **Với bộ dữ liệu sau thay đổi**:



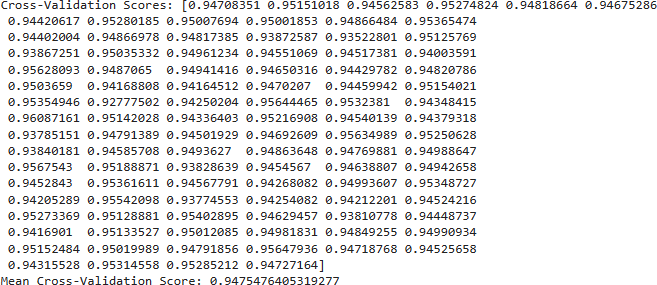
*Hình 8: Biểu đồ phân tán giữa giá trị thực tế và dự đoán của bộ dữ liệu mới*

Hệ số xác định sau thay đổi, 𝑹𝟐 = 𝟎, 𝟗𝟓𝟗𝟗, sau khi thực hiện thay đổi bộ dữ liệu, mô hình đã giải thích được 𝟗𝟓, 𝟗𝟗% sự biến biến thiên của **charges**. Bên cạnh đó, biểu đồ cho thấy ít điểm nằm xa đường hoàn hảo hơn với bộ dữ liệu ban đầu, tức sai số nhỏ hơn ban đầu.

Sau khi thêm biến thuộc tính, sự thay đổi về biến charges hiệu quả hơn so với bộ dữ liệu ban đầu. Điều này có thể do các biến mới như **married**, **employ**, **insur\_renewals** đã phần nào ảnh hưởng đến biến **charges** và làm tăng độ chính xác của mô hình hồi quy tuyến tính.

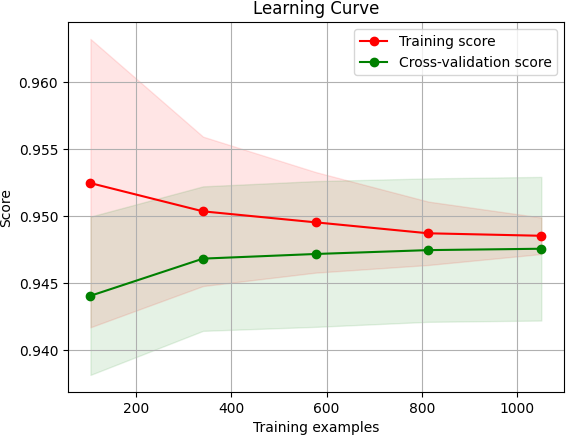
## Kiểm tra Overfitting bằng Cross-Validation và Learning Curve

* + **Cross-Validation**: là phương pháp chia dữ liệu thành nhiều phần nhỏ, mỗi phần sẽ được dùng làm tập kiểm tra một lần, còn các phần còn lại sẽ làm tập huấn luyện. Kết quả của Cross-Validation là trung bình của các điểm số trên các phần kiểm tra. Điểm số cao cho thấy mô hình có khả năng tổng quát hóa tốt trên dữ liệu mới.



*Hình 9: Kết quả kiểm tra Cross-Validation trên bộ dữ liệu mới*

* + **Learning Curve**: là biểu đồ thể hiện sự thay đổi của điểm số trên tập huấn luyện và tập kiểm tra khi kích thước tập huấn luyện tăng. Nếu hai đường biểu diễn gần nhau và cao, cho thấy mô hình không bị Overfitting. Nếu hai đường biểu diễn xa nhau và thấp, cho thấy mô hình bị Overfitting.



*Hình 10: Kết quả kiểm tra Learning Curve trên bộ dữ liệu mới*

Nhận xét:

Mô hình sau khi thêm biến cho bộ dữ liệu có điểm số Cross-Validation trung bình là 0,9585 - rất cao so với mô hình ban đầu. Quan sát biểu đồ, nhóm nhận thấy đường biểu diễn sai số huấn luyện (Training Error) và sai số kiểm tra (Validation Error) có xu hướng hội tụ khi kích thước tập huấn luyện tăng lên, điều đó có nghĩa là mô hình có khả năng khái quát hóa tốt trên cả tập huấn luyện và tập kiểm tra.

Ngoài ra, sai số huấn luyện đều ở mức thấp, khoảng 0.02. Chứng tỏ, mô hình có độ chính xác cao, không bị tình trạng Under hay Over Fitting. Có thể nói, mô hình có độ nhất quán cao và không bị ảnh hưởng nhiều bởi các yếu tố nhiễu hay biến động của dữ liệu.

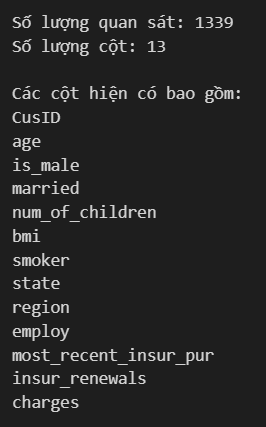
# CHƯƠNG IV. TIỀN XỬ LÝ DỮ LIỆU

**Mã nguồn**: [1-Preprocessing.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1rLvOZQ1JPHyaDAHJ0hqmuG4JBgYaXZYK)

## Mô tả tổng quan bộ dữ liệu sau thay đổi

### Thông tin cơ bản

* + - Đầu tiên, nhóm thực hiện quan sát những thông tin cơ bản sau sửa đổi như số lượng quan sát (dòng), số lượng biến (cột), và các thông tin mô tả về các biến trong bộ dữ liệu.
    - Sau thay đổi, bộ dữ liệu đã có thêm 6 biến thuộc tính, bao gồm: **CusID**, **married**, **state**, **employ**, **most\_recent\_insur\_pur** và **insur\_renewals**.



*Hình 11: Mô tả thông tin cơ bản của bộ dữ liệu sau thay đổi*

### Kiểm tra dữ liệu

* + - Tương tự như bộ dữ liệu ban đầu, nhóm cần thực hiện tìm và xử lý giá trị trống (missing values) cũng như các giá trị không hợp lệ (invalid values) có thể ảnh hưởng đến kết quả của phân tích dữ liệu.



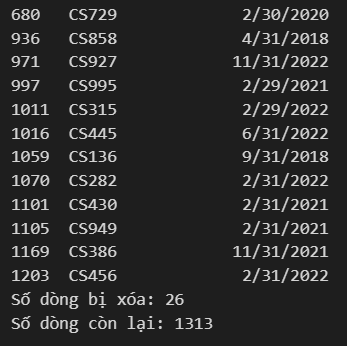
*Hình 12: Quan sát giá còn thiếu và giá trị bất thường của bộ dữ liệu sau thay đổi*

* + - Từ kết quả trên nhóm thấy được rằng cột **insur\_renewals** và **charges** có giá trị bị thiếu do hai lý do chính:
      * Đầu tiên, nhiều bang tại Hoa Kỳ không công bố mức bảo hiểm trung bình hàng năm, do đó không thể tính được giá trị cho cột **insur\_renewals**.
      * Thứ hai, cột charges được tính toán dựa vào số lần làm mới bảo hiểm (insur\_renewals). Khi **insur\_renewals** không có số liệu, điều này dẫn đến việc biến **charges** cũng không có số liệu. Vì vậy, sự thiếu hụt dữ liệu trong cả hai cột này là kết quả của sự thiếu thông tin từ các bang cũng như sự phụ thuộc lẫn nhau giữa hai cột dữ liệu này.
    - Để xử lý giá trị bị thiếu ở hai cột **insur\_renewals** và **charges**, nhóm sẽ xóa chúng ra khỏi bộ dữ liệu.



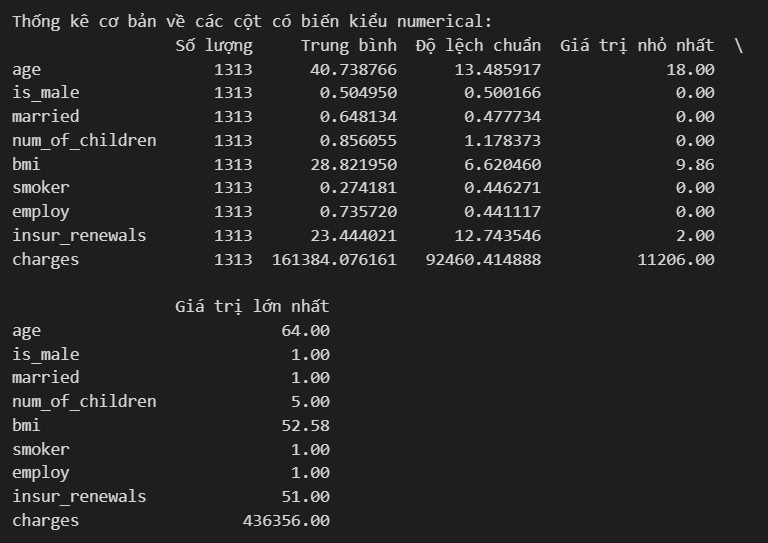
*Hình 13: Quan sát số giá trị trống có trong bộ dữ liệu sau thay đổi*

* + - Xử lý giá trị không hợp lệ của thuộc tính thời gian **most\_recent\_insur\_pur**, trong quá trình khởi tạo giá trị cho cột này, nhóm đã sử dụng hàm **RANDBETWEEN()** của Excel, chính vì thế mà không thể tránh khỏi việc xuất hiện của những biến thời gian bất thường như Tháng 2 có ngày 30 hay Tháng 4 có ngày 31. Chính vì vậy, nhóm cần thực hiện xóa những dòng dữ liệu không phù hợp như vậy ra khỏi bộ dữ liệu.



*Hình 14: Quan sát số giá trị bất thường có trong bộ dữ liệu sau thay đổi*

* + - Tiếp đến, nhóm sẽ thực hiện tìm ra phân phối và sự biến đổi của dữ liệu thông qua phân tích thống kê tổng quát, bao gồm việc xác định giá trị trung bình và độ lệch chuẩn của dữ liệu. Kết quả mà nhóm thu được như sau:

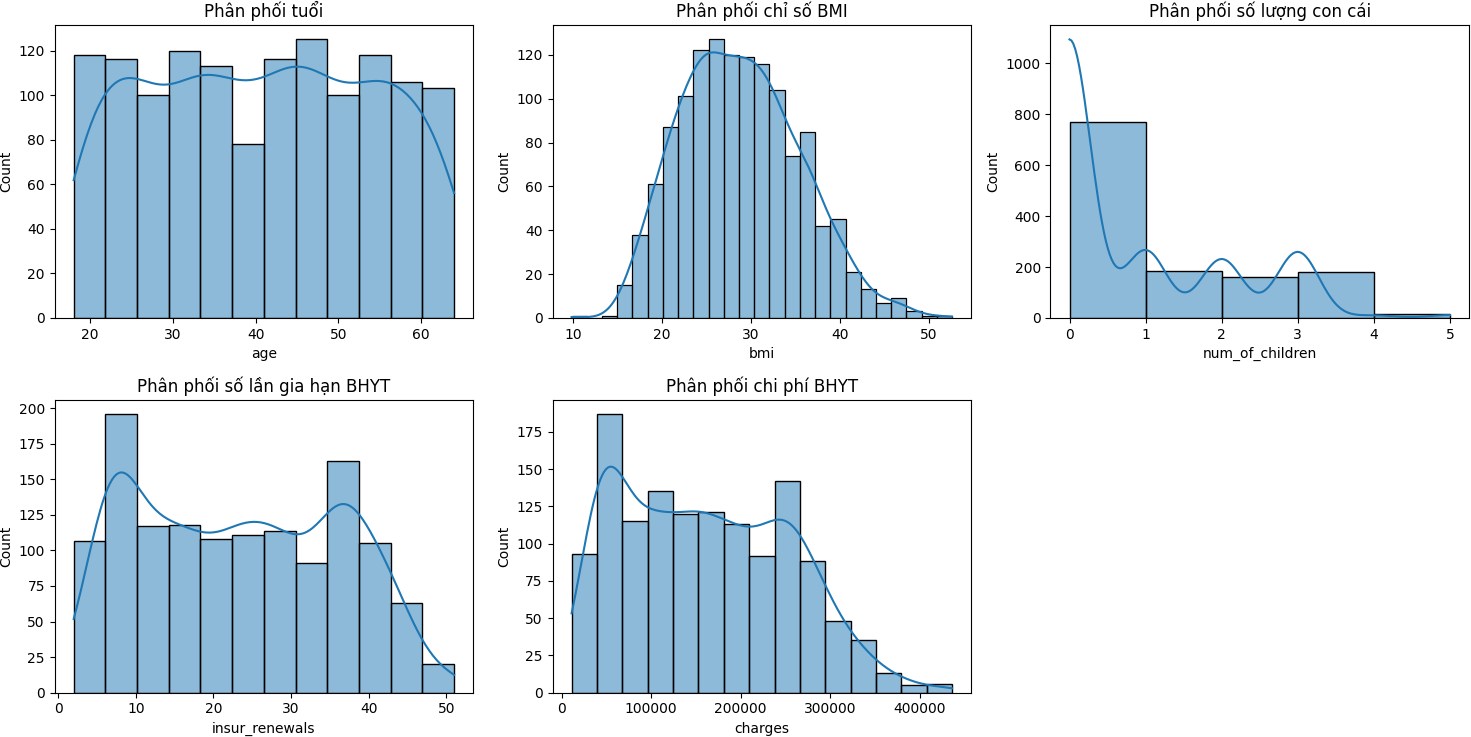


*Hình 15: Bảng thống kê tổng quát của bộ dữ liệu sau thay đổi*

* + - Nhìn chung nhóm thấy được rằng, với bộ dữ liệu mới đã có nhiều bước thay đổi tích cực trong việc phân tích về mặt ý nghĩa như sau:
      * Trong bộ dữ liệu mới, biến **num\_of\_children** (hay **children**) cho thấy sự biến động rất lớn với giá trị trung bình là 0,84, nhưng dao động từ 0 đến 5. Điều này cho thấy sự đa dạng về tình trạng có con cái hay không trong quá trình tính toán chi phí bảo hiểm.
      * Biến charges ở bộ này có giá trị trung bình cao hơn (160.747,48) và độ lệch chuẩn lớn (92.730,57) so với biến tương tự trong bộ dữ liệu ban đầu (13.270,42 và 12.110,01), cho thấy sự biến động đáng kể trong chi phí BHYT. Sự biến động lớn này có thể chứa nhiều thông tin ý nghĩa về sự ảnh hưởng của các yếu tố khác nhau đối với chi phí bảo hiểm.
      * Ngoài ra với các biến phụ mới như **married**, **employ**, **state**, **most\_recent\_insur\_pur** và **insur\_renewals** đã cung cấp thêm cái nhìn nhiều chiều đối với việc chi trả BHYT của khách hàng (KH) cho bộ dữ liệu.

### Trực quan hóa dữ liệu

* + - Vẽ biểu đồ phân phối cho cột dữ liệu số (**age**, **bmi**, **num\_of\_children**, **insur\_renewals** và **charges**):

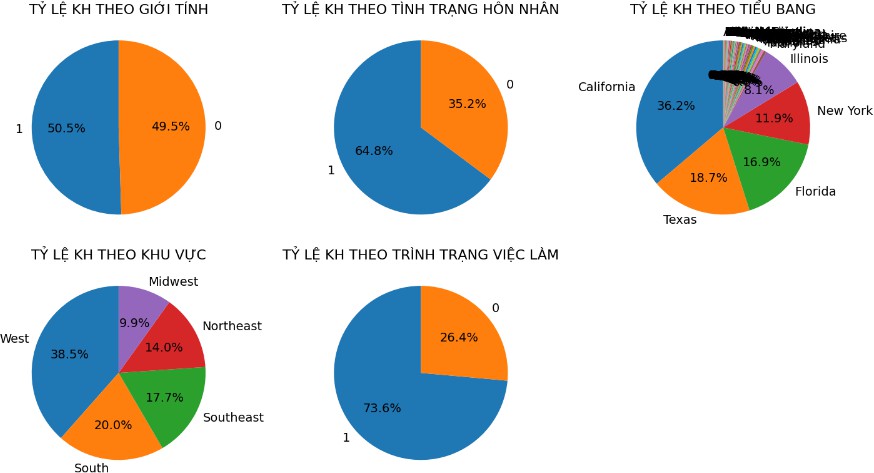


*Hình 16: Biểu đồ phân phối cho cột dữ liệu số của bộ dữ liệu mới*

Nhận xét về sự thay đổi của dữ liệu:

* **age** và **bmi**: Nhìn chung, ở hai biến này cũng không có nhiều sự thay đổi bởi đây chính là một trong những yếu tố quan trọng để KH thực hiện chi tiền cho BHYT.
* **num\_of\_children**: Sau khi thay đổi, biểu đồ phân phối đã có sự khác biệt rõ rệt. Số lượng con cái của KH tham gia BHYT có xu hướng tăng lên, với đỉnh cao nhất là 5 con và thấp nhất là 0 con. Điều này cho thấy số lượng con cái là một yếu tố ảnh hưởng đến chi phí BHYT của KH, vì càng có nhiều con thì càng cần nhiều chi phí y tế cho gia đình.
* **charges**: Sau khi thay đổi ý nghĩa ban đầu của bộ dữ liệu, nhóm nhận thấy số liệu đã tăng rõ rệt, cho thấy được tổng số tiền KH đóng cho BHYT lớn hơn nhiều so với BHYT chi trả cho y tế cá nhân của KH.
* **insur\_renewals**: Số lần đóng lại BHYT của KH có phân bố khá đều, từ khoảng 3 đến 55 lần, cao nhất là khoảng 12 lần và thấp nhất là 3 lần. Điều này cho thấy số lần đóng lại BHYT của KH phản ánh chính xác mức độ quan tâm của họ đối với BHYT, cũng như tình trạng sức khỏe của họ.
  + - Vẽ biểu đồ tỷ lệ cho cột dữ liệu phân loại (**is\_male**, **married**, **state**, **region** và

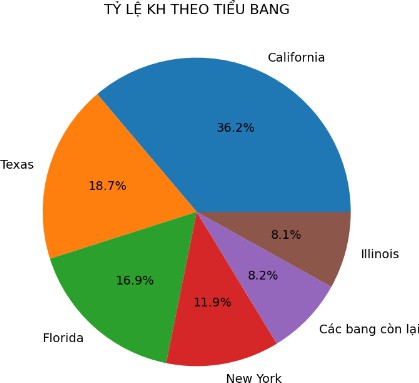
**employ**):



*Hình 17: Biểu đồ tròn cho cột dữ liệu phân loại của bộ dữ liệu mới*

Nhận xét biểu đồ:

* Đối với biểu đồ Tỷ lệ tiểu bang, nhóm thấy được rằng số liệu chỉ tập trung phần đa vào 5 bang trung tâm của Hoa Kỳ bao gồm: Florida, California, Texas, New York và Illinois, các bang còn lại vì số liệu quá ít khiến cho biểu đồ bị “**đen đặc**” lại.
* Để xử lý điều đó, nhóm thực hiện gom nhóm các tiểu bang khác ngoài các bang trung tâm thành một cụm và đặt tên cho giá trị là “**Các bang còn lại**” để biểu đồ trở nên rõ ràng và dễ nhìn hơn.



*Hình 18: Biểu đồ tròn thể hiện tỷ lệ KH theo tiểu bang*

Nhận xét về sự thay đổi của dữ liệu:

* Biểu đồ tỷ lệ của 2 biến, **sex** (hoặc **is\_male**) và **region** trước khi thay đổi, cho thấy sự cân bằng giới tính trong việc tham gia BHYT và phân phối theo khu vực. Tuy nhiên, với biểu đồ ban đầu, khó để nhận biết sự khác biệt giữa giới tính theo khu vực.
* Sau khi thay đổi, biểu đồ cho thấy tỷ lệ nam giới và nữ giới tham gia BHYT theo khu vực dễ dàng hơn. Nam giới tham gia BHYT nhiều hơn nữ giới ở các khu vực Đông Nam, Tây Nam và Tây Bắc, trong khi nữ nhiều hơn nam ở khu vực Đông Bắc.
* Biểu đồ tỷ lệ của 3 biến mới là **married**, **state** và **employ** cho thấy sự ảnh hưởng của tình trạng hôn nhân, tiểu bang và tình trạng việc làm đối với việc tham gia BHYT. Người đã kết hôn tham gia nhiều hơn so với người chưa kết hôn, và phân phối theo tiểu bang và tình trạng việc làm cũng đa dạng. California có tỷ lệ cao nhất và người có việc làm tham gia nhiều hơn so với người thất nghiệp hoặc nghỉ hưu.
  + - Vẽ biểu đồ phân phối cho cột dữ liệu thời gian (**most\_recent\_insur\_pur**)



*Hình 19: Biểu đồ phân phối cho cột dữ liệu thời gian của bộ dữ liệu mới*

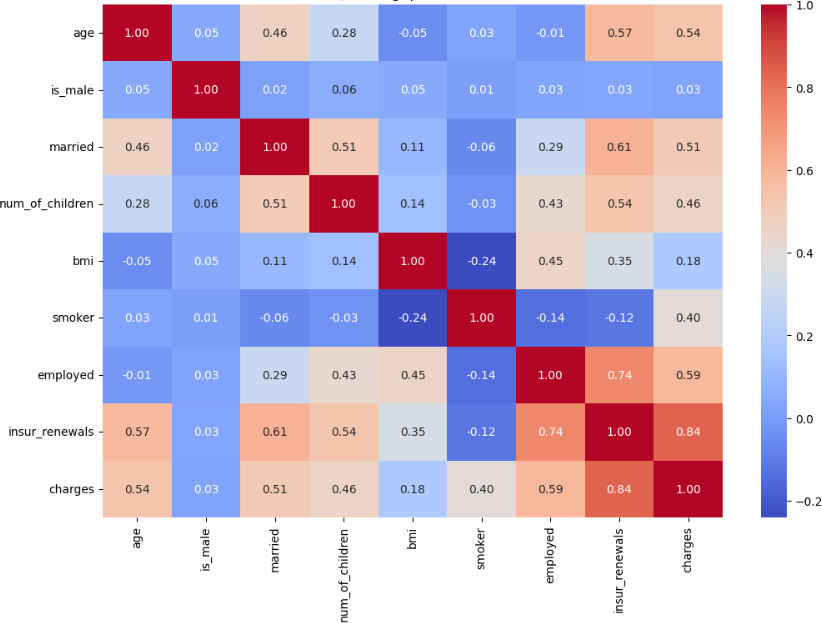
Nhận xét về dữ liệu:

Từ biểu đồ trên, nhóm nhìn nhận được xu hướng quan tâm BHYT của KH được phân bổ làm hai nhóm rõ rệt. Với mức độ quan tâm cao chiếm ưu thế (đóng từ năm

2018 trở đi) và số ít KH có mức quan tâm thấp (lần cuối tham gia bảo hiểm là dưới năm 2016).

### Kiểm tra sự tương quan

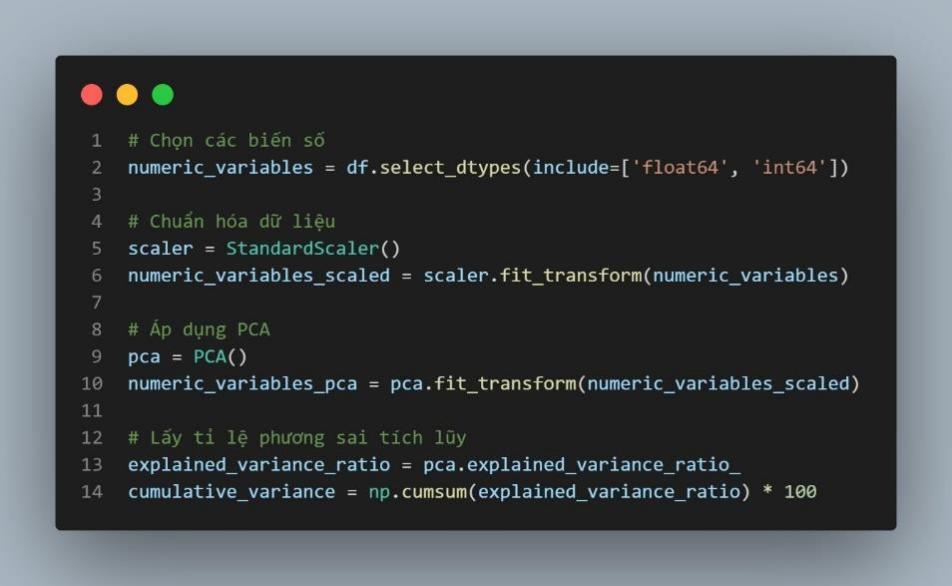
* + - Vẽ biểu đồ thể hiện sự tương quan giữa các biến có trong bộ dữ liệu:

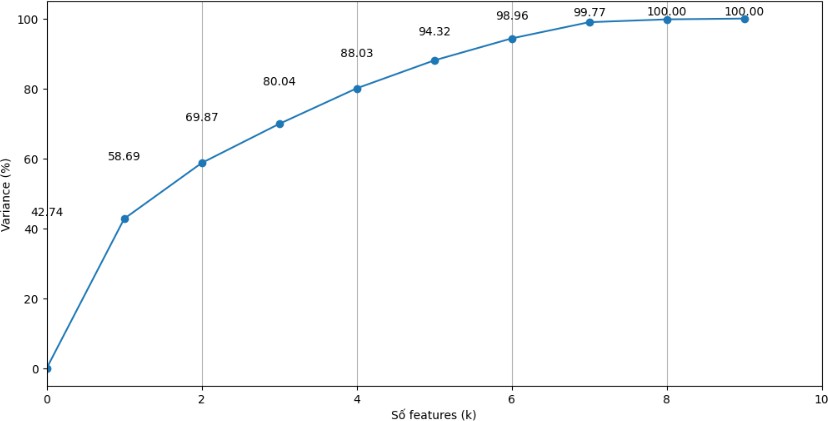


*Hình 20: Ma trận tương quan của biến số có trong bộ dữ liệu sau thay đổi*

Nhận xét chung về dữ liệu:

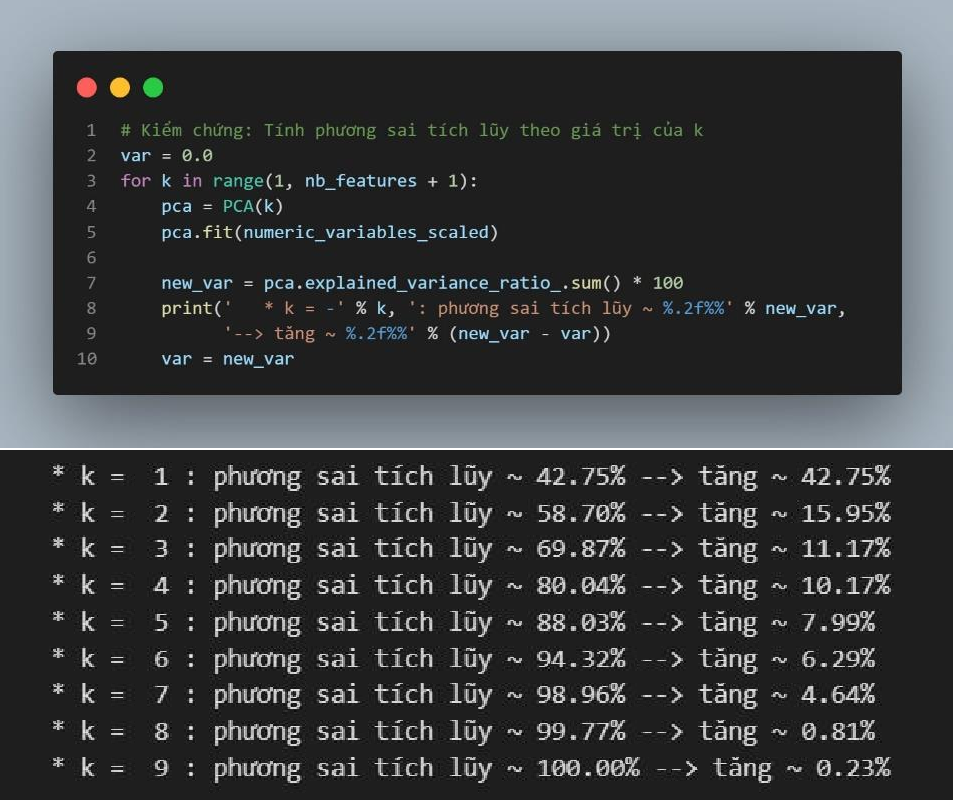
* Từ biểu đồ trên cho thấy độ tương quan giữa các biến có mối quan hệ khá chặt chẽ với nhau. Nghĩa là, giờ đây các biến đã có thể phần nào giải thích cho quyết định chi trả cho BHYT của KH.
* Tuy nhiên nhìn từ ma trận tương quan, nhóm nhận thấy có một số biến thuộc tính có độ tương quan thấp với biến mục tiêu mà nhóm nghiên cứu (Ví dụ: **is\_male**, **bmi** hay thậm chí là biến **smoker** chỉ có độ tương quan cao duy nhất đối với biến mục tiêu là **charges**, còn ở các biến khác thì rất thấp hay có thể nói là độc lập với nhau).
* Do đó, để có thể biết được bộ dữ liệu có đang bị dư thừa Feature hay không, nhóm thực hiện sử dụng phương pháp **PCA** cùng với biểu đồ đường để quan sát phương sai tích lũy theo số Feature của bộ dữ liệu:





*Hình 21: Đồ thị biểu diễn % Phương sai tích lũy theo số Features (k)*

* Tiếp tục, ta thực hiện kiểm chứng bằng cách tính lại phương sai tích lũy theo giá trị của Feature (k):



*Hình 22: Phần trăm Phương sai tích lũy theo số Features (k)*

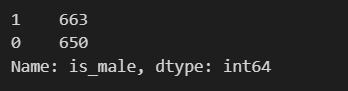
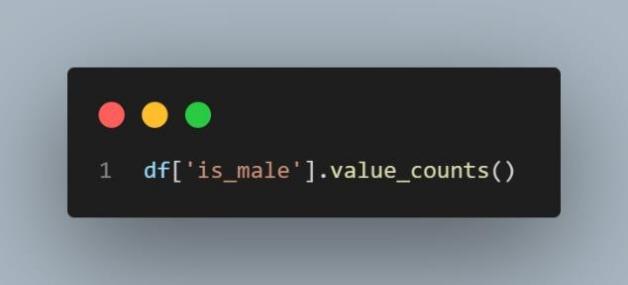
Nhận xét:

* Dựa trên biểu đồ và bảng phương sai tích lũy, nhóm có thể thấy rằng việc tăng số lượng biến thuộc tính từ 1 lên 7 đã giúp giải thích gần như toàn bộ sự biến đổi trong dữ liệu (với phương sai tích lũy tăng từ 42,75% lên 98,96%).
* Tuy nhiên, việc tăng thêm số lượng biến thuộc tính từ 7 lên 9 chỉ làm tăng phương sai tích lũy thêm 1,04%, từ 98,96% lên 100%. Điều này cho thấy việc thêm biến thuộc tính sau k = 7 không cung cấp nhiều thông tin hơn.
* Vì vậy, ta có thể chọn k = 7 làm số lượng biến thuộc tính phù hợp để giữ lại hầu hết thông tin mà không làm tăng đáng kể độ phức tạp của mô hình. Đồng nghĩa với việc, nhóm sẽ loại đi 2 biến có mức tương quan thấp nhất với biến mục tiêu là: **is\_male** và **bmi**. Tuy nhiên, để có thể loại bỏ được hai biến trên thì ta phải đưa yêu cầu cụ thể nào đó để xác định được mức độ chấp nhận của việc mất mát thông tin.
* Nhằm tránh việc mất mát thông tin, nhóm thực hiện một số kiểm định sau để xác định được có nên loại bỏ hai biến **is\_male** và **bmi** hay không:
  + - **Kiểm định giả thuyết 1**: Nam và Nữ sử dụng Bảo hiểm Y tế (BHYT) như nhau.

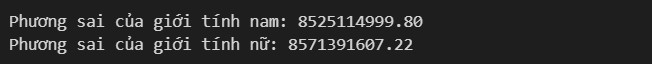
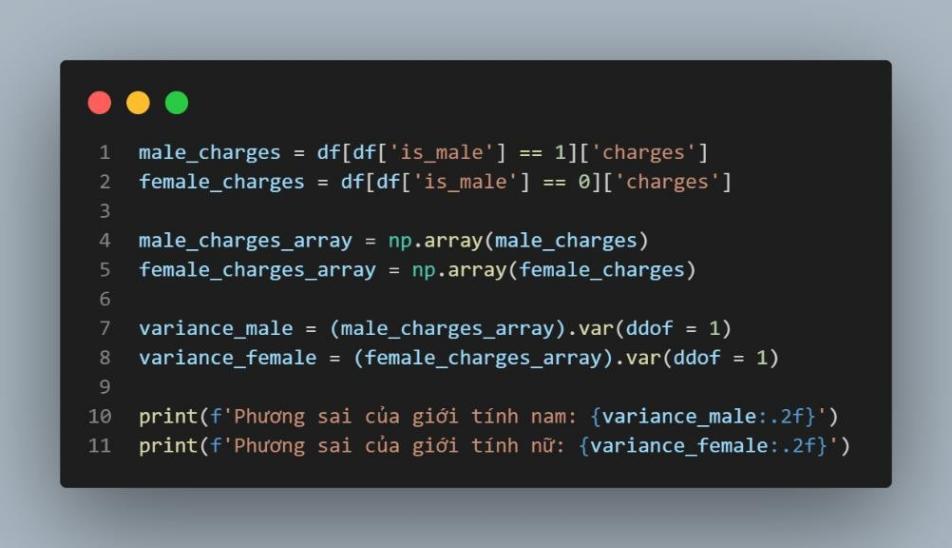
Đầu tiên, nhóm xây dựng H0 và H1 tương ứng với giả thuyết 1 và chọn độ tin cậy cho bài toán là 95%, tương đương với mức ý nghĩa α = 0,05.

{H0: μMale(charges) = μFemale(charges) H1: μMale(charges) ≠ μFemale(charges)

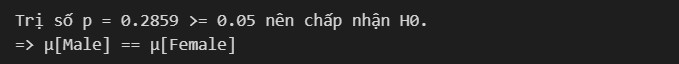
Khi xem xét số liệu về tổng tiền chi trả cho BHYT được phân loại theo giới tính, nhóm nhận thấy được số lượng lấy mẫu khá tương đồng nhau.



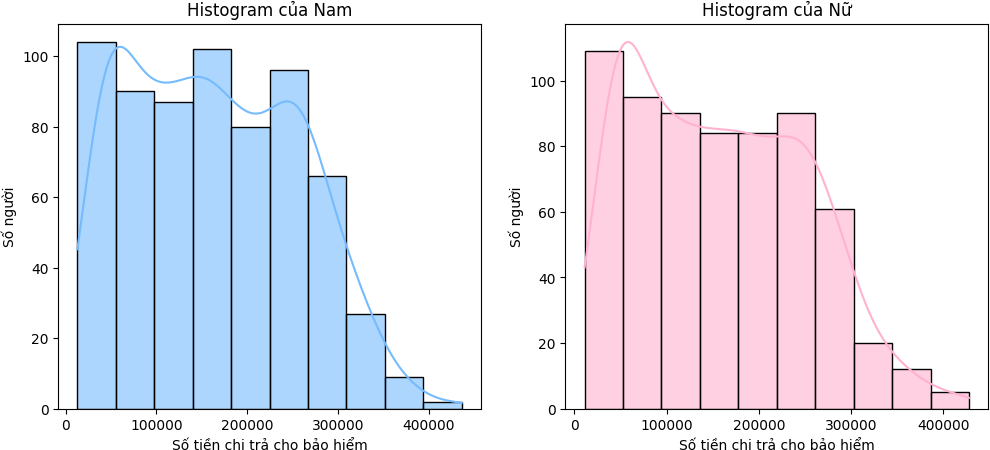
Đây là bài toán kiểm định hai mẫu độc lập với biến Categorical là **is\_male** và biến Numerical là **charges**. Nhóm so sánh phương sai hai biến và thấy được rằng Var(Male) và Var(Female) không giống nhau.



Nhóm quyết định sử dụng Phương pháp kiểm định dựa trên trị số 𝑝. Với số lượng mẫu khá lớn, nhóm áp dụng hàm ztest\_ind() để tìm trị thống kê 𝑧 và trị số p. Từ đó, nhóm tiếp tục so sánh p và α để rút ra kết luận.



Khi áp dụng Phương pháp sử dụng p-value với mức ý nghĩa 5%, vì trị số p lớn hơn α nên nhóm chấp nhận giả thuyết H0. Hoặc hiểu rằng, trung bình số tiền nam giới chi trả cho bảo hiểm cũng tương đồng với số tiền của nữ.



*Hình 23: So sánh hai Biểu đồ Histogram thể hiện số tiền chi trả bảo hiểm theo số người của mỗi giới tính*

Nhận xét biểu đồ:

Đối với biểu đồ Histogram, nhóm thấy được với số lượng mẫu nam nhiều hơn mẫu là nữ, tổng khoảng tiền nam giới chi trả cho bảo hiểm cũng chênh lệch hơn nữ giới khá nhỏ.

* Vậy từ kiểm định và biểu đồ trực quan, nhóm đã có cơ sở để loại bỏ biến **is\_male**

ra bộ dữ liệu để tiết kiệm thời gian nghiên cứu.

* + - **Kiểm định giả thuyết 2**: Chỉ số sức khỏe (BMI) không làm ảnh hưởng đến việc làm mới bảo hiểm.

Thay vì kiểm định chỉ số BMI có ảnh hưởng gì đến tổng chi phí mà KH phải chi trả cho BHYT, thì nhóm sẽ kiểm định chỉ số này có ảnh hưởng gì việc làm mới (đóng lại) BHYT. Bởi, việc làm mới bảo hiểm phản ánh trực tiếp về nhu cầu của KH, ngoài ra đây là biến có độ tương quan cao với biến chi phí (**charges**) nên sẽ mang lại nhiều thông tin hơn, tránh việc mất mát thông tin không đáng có.

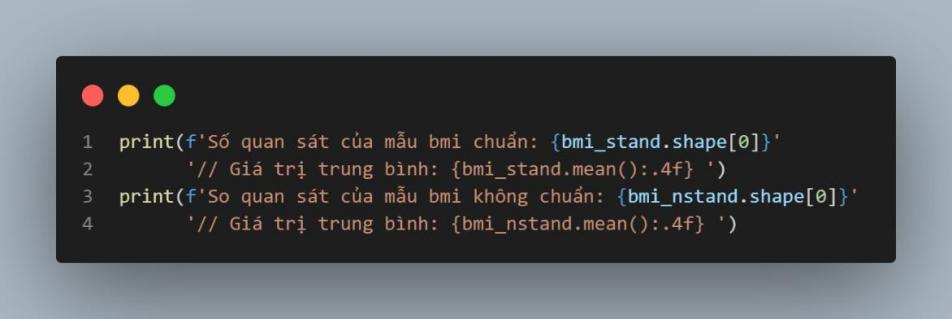
Để có thể kiểm tra giả thuyết trên, nhóm sẽ tiến hành lấy trung bình số lần làm mới bảo hiểm của nhóm người có chỉ số BMI đạt chuẩn so với nhóm người có chỉ số BMI cao hơn hoặc dưới tiêu chuẩn (mức chỉ số BMI tiêu chuẩn là từ 18,5 đến 24,9).

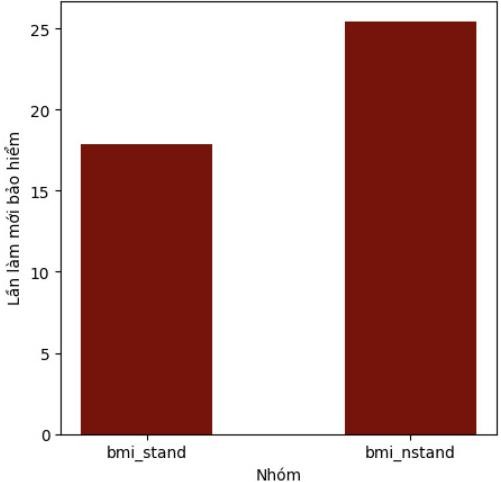
Nhóm bắt đầu xây dựng giả thuyết H0 và giả thuyết H1 theo ý tưởng trên, với mức ý nghĩa được chọn ở đây là *α* = 5% và độ tin cậy là 95%:

{H0: μbmi\_standinsur\_renewals = μbmi\_nstandinsur\_renewals H1: μbmi\_standinsur\_renewals ≠ μbmi\_nstandinsur\_renewals

Đầu tiên, nhóm xem xét sơ qua về các đại lượng thống kê (độ lệch chuẩn và giá trị trung bình 𝜇), biểu đồ phân phối và số lượng phần tử để quan sát rõ hơn về của hai nhóm được đưa lên kiểm định.

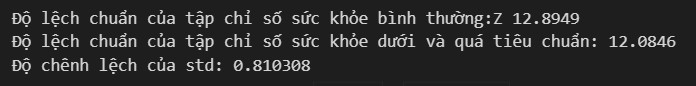
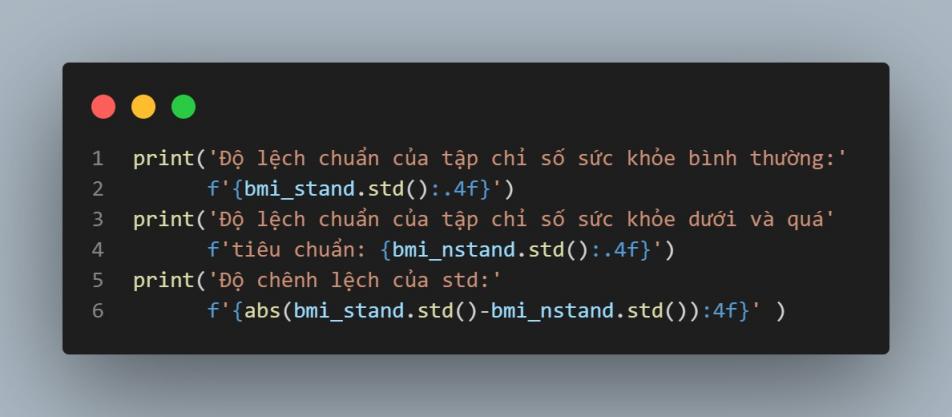
* Giá trị trung bình và số lượng quan sát của hai nhóm:

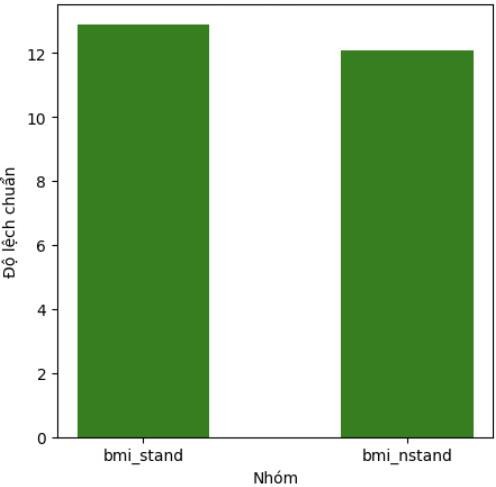




*Hình 24: Biểu đồ thanh thể hiện giá trị trung bình lần làm mới bảo hiểm của hai nhóm*

* Giá trị độ lệch chuẩn của hai biến quan sát:





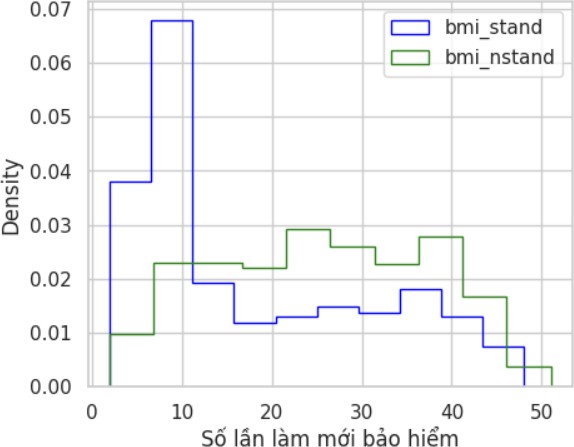
*Hình 25: Biểu đồ thanh thể hiện độ lệch chuẩn của hai nhóm*

* + Số phần tử của cả hai biến là khá lớn khi số phần tử của mẫu **bmi** không chuẩn gần như gấp 3 lần số lượng phần tử của mẫu **bmi** chuẩn. Và giá trị trung bình μ cũng cho thấy sự chênh lệch khá lớn.
  + Độ biến thiên của hai biến là tương đối giống nhau, mức độ chênh lệch

giữa hai giá trị độ lệch chuẩn chỉ ở mức 0,81.

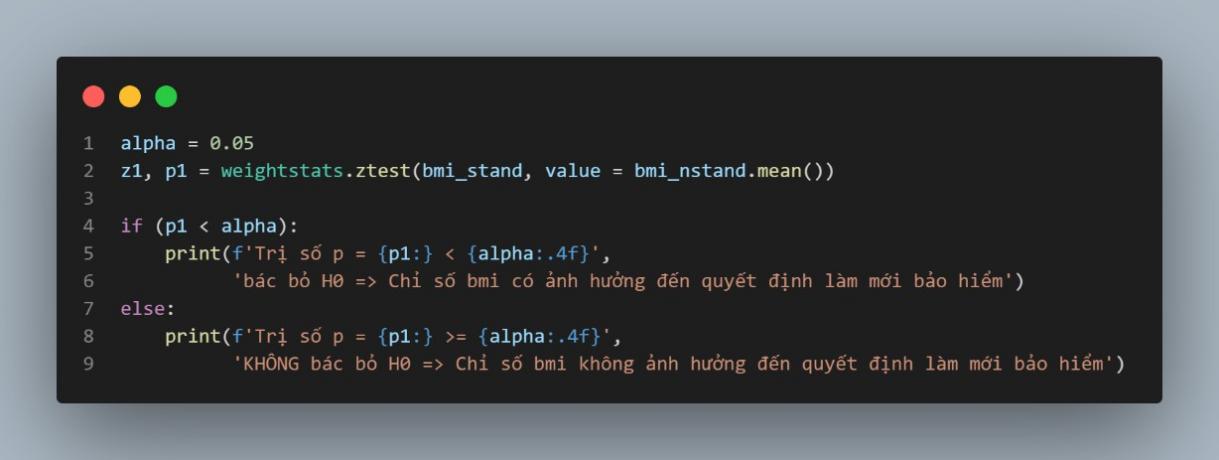
* Tiến hành vẽ biểu đồ Histogram để quan sát phân phối của hai biến:





*Hình 26: Biểu đồ phân phối xác suất số lần làm mới bảo hiểm của hai biến*

* + Thông qua biểu đồ phân phối của hai biến, nhóm thấy được rằng nhóm người có chỉ số BMI không chuẩn có xu hướng làm mới BHYT khá đều ở từng mức độ, trong khi đó nhóm người có chỉ số BMI chuẩn lại hầu hết chỉ tập trung làm mới bảo hiểm từ 1 đến 10 lần, sau đó họ không còn có xu hướng làm mới bảo hiểm thêm nữa.
  + Vậy, nhóm nhận thấy được rất rõ ràng chỉ số BMI có thể ảnh hưởng đến quyết định của KH trong việc lựa chọn làm mới BHYT của mình.
* Để làm rõ hơn nhận định này nhóm sẽ quyết định thực hiện việc kiểm định giả thuyết H0 đã được xác định ở trên. Do số lượng quan sát của mẫu lớn (>30 quan sát), nhóm quyết định thực hiện kiểm định bằng Z-test.



Sử dụng hàm ztest\_ind() của thư viện statsmodels với mức ý nghĩa là 5%, nhóm thu được kết quả của kiểm định cho thấy rằng trị số *p* (p-value) nhỏ

hơn rất nhiều so với mức ý nghĩa 5% tức là tỉ lệ mắc sai lầm khi bác bỏ H0 là rất thấp.

* Vậy từ kiểm định và biểu đồ trực quan, nhóm nhóm vẫn chưa có cơ sở để loại bỏ biến **bmi** ra khỏi bộ dữ liệu.

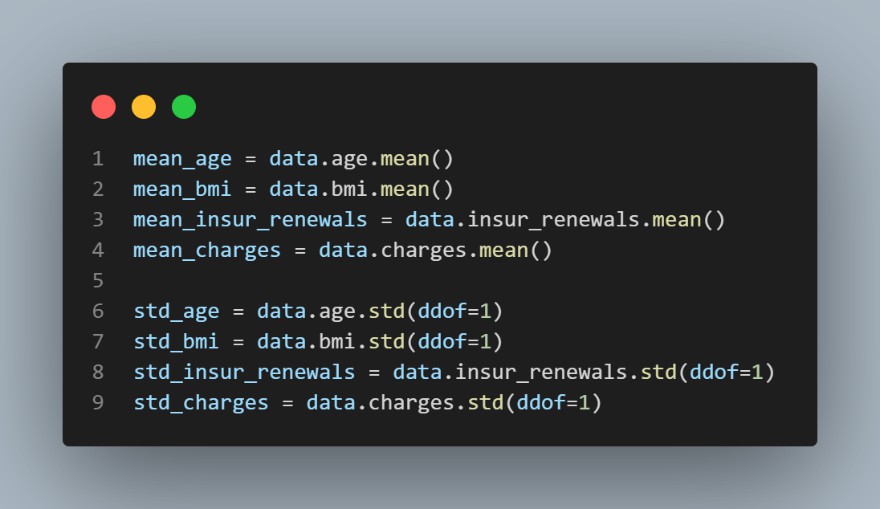
## Xử lý giá trị Outliers

Đầu tiên, nhóm tìm ra các cột chứa giá trị định lượng có nguy cơ xuất hiện giá trị Outliers nhất.

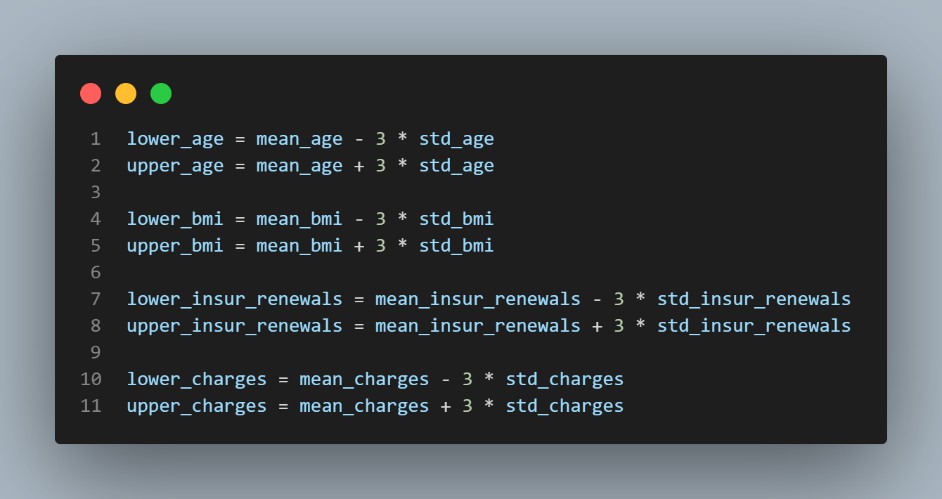


Nhóm sử dụng **quy tắc 3-sigma** để tìm giá trị Outliers của từng cột chứa giá trị định lượng:

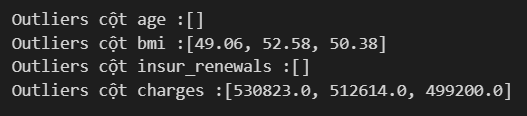
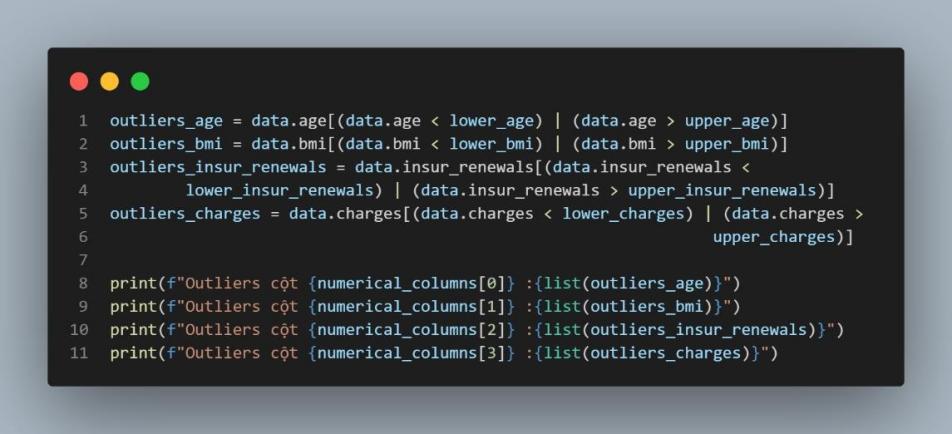
* Tính các giá trị trung bình (Mean) và độ lệch chuẩn (Standard Deviation) của các cột định lượng (**age**, **bmi**, **insur\_renewals** và **charges**) trong DataFrame df\_cleaned. Độ lệch chuẩn được tính với ddof = 1, nghĩa là được hiệu chỉnh bằng cách chia cho (n - 1) thay vì chỉ chia cho n để tính độ lệch chuẩn mẫu.



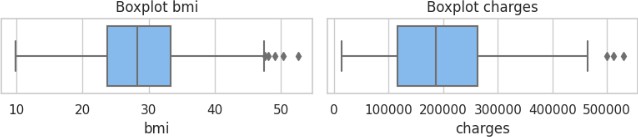
* Sau đó, nhóm tính giới hạn dưới (Lower) và giới hạn trên (Upper) cho từng cột dựa trên kỹ thuật 3-sigma từ giá trị trung bình tương ứng. Điều này được sử dụng để xác định khoảng giá trị bình thường dự kiến trong phạm vi 3 độ lệch chuẩn.



* + Tiếp theo, đoạn mã tạo các danh sách (Series) riêng biệt cho các giá trị ngoại lệ (Outliers) trong mỗi cột. Điều này được thực hiện bằng cách so sánh giá trị của mỗi phần tử trong cột với giới hạn dưới và giới hạn trên tương ứng cho cột đó. Nếu giá trị nằm ngoài khoảng này, nó sẽ được coi là ngoại lệ và được thêm vào danh sách giá trị Outliers tương ứng.

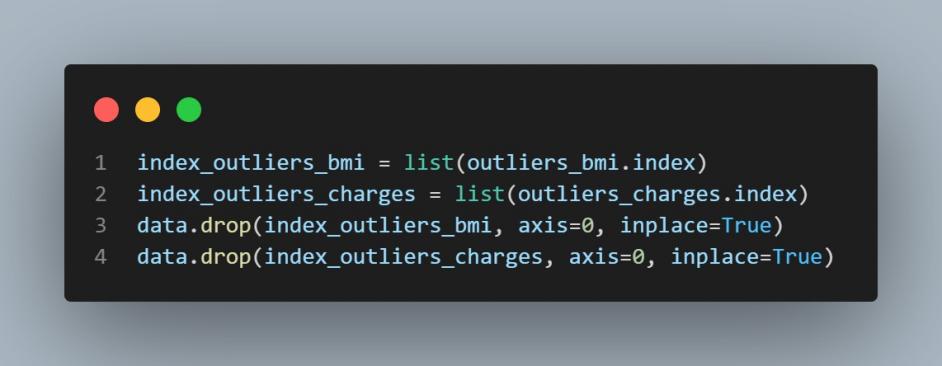


* + Dựa vào kết quả nhóm có thể thấy được cột **bmi** và **charges** chứa các giá trị Outliers. Để trực quan hơn, nhóm sẽ sử dụng biểu đồ boxplot để quan sát những giá trị đó.

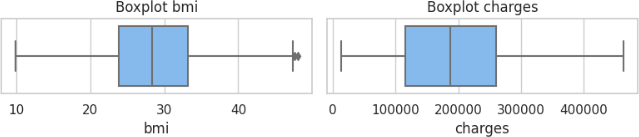


*Hình 27: Biểu đồ hộp quan sát giá trị Outliers của biến* ***bmi*** *và* ***charges***

* + Tiếp theo, nhóm thực hiện loại bỏ giá trị Outlier ra khỏi Dataset.



* + Biểu đồ boxplot sau khi loại bỏ giá trị Outliers:



*Hình 28: Biểu đồ hộp thể hiện Outliers cho* ***bmi*** *và* ***charges*** *sau khi áp dụng 3-sigma*

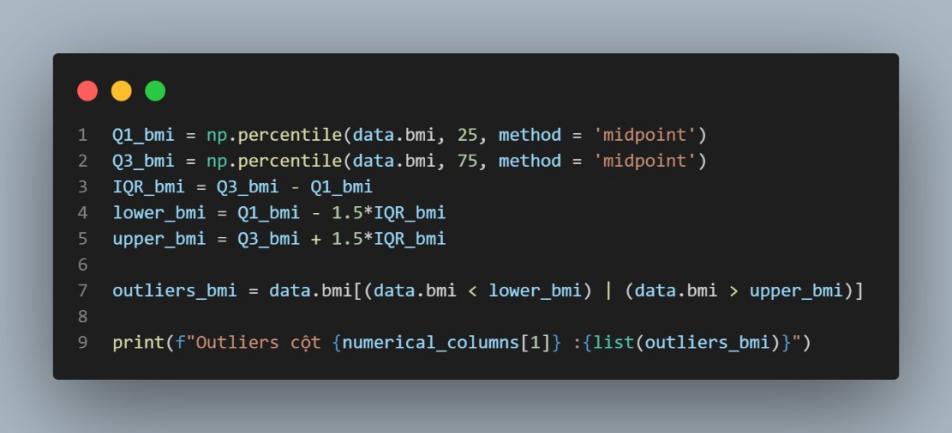
* + Quan sát biểu đồ Boxplot sau khi loại bỏ Outliers nhưng cột **bmi** vẫn còn 1 vài giá trị Outliers.

Nhóm tiếp tục sử dụng phương pháp **Khoảng tứ phân vị (Interquartile Range - IQR)**

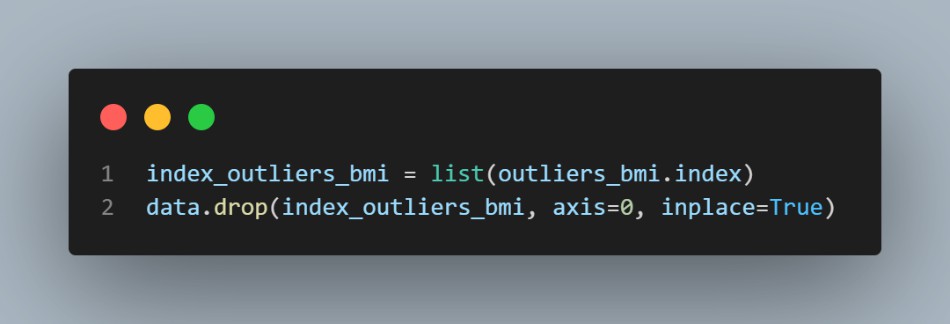
để xử lý tiếp những Outliers còn sót lại cho cột **bmi**.

* + Đầu tiên, đoạn mã sử dụng thư viện NumPy để tính các giá trị cần thiết để xác định giá trị ngoại lệ dựa trên phân vị (Percentiles), cụ thể:
    - **Q1\_bmi:** là giá trị của **bmi** ở phân vị 25% (Q1), tức là giá trị mà 25% dữ liệu **bmi** nhỏ hơn và 75% lớn hơn.
    - **Q3\_bmi:** là giá trị của **bmi** ở phân vị 75% (Q3), tức là giá trị mà 75% dữ liệu **bmi** nhỏ hơn và 25% lớn hơn.
    - **IQR\_bmi:** là khoảng cách giữa Q3 và Q1, được gọi là phạm vi tương quan (Interquartile Range).
  + Sau đó, xây dựng đoạn mã tính giới hạn dưới (lower\_bmi) và giới hạn trên (upper\_bmi) dựa trên phạm vi tương quan (IQR). Cụ thể, giới hạn dưới là Q1 trừ

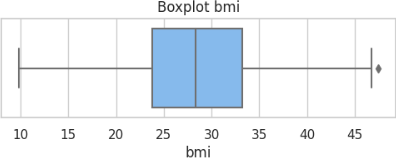
1.5 lần IQR, và giới hạn trên là Q3 cộng 1.5 lần IQR. Các giá trị nằm ngoài khoảng này sẽ được xem xét là giá trị Outliers.



* + Cuối cùng, mã code sử dụng phương thức drop để loại bỏ các hàng (axis=0) có chỉ mục nằm trong danh sách index\_Outliers\_bmi khỏi DataFrame. Điều này dẫn đến việc loại bỏ các giá trị ngoại lệ khỏi dữ liệu.

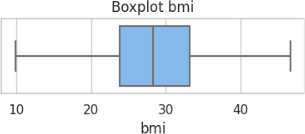


* + Biểu đồ Boxplot sau khi loại bỏ giá trị Outlier bằng phương pháp Khoảng trải giữa (Interquartile range - IQR):



*Hình 29: Biểu đồ hộp thể hiện Outliers cho* ***bmi*** *khi sử dụng IQR*

Như trên biểu đồ, nhóm nhận thấy thì các giá trị Outlier của cột định lượng **bmi** chỉ còn duy nhất một giá trị Outliers, nhóm tiếp tục sử dụng thêm một lần phương pháp **Khoảng tứ phân vị** để loại bỏ hoàn toàn giá trị Outliers ra khỏi bộ dữ liệu.



*Hình 30: Biểu đồ hộp thể hiện Outliers cho* ***bmi*** *khi sử dụng thêm một lần phương pháp IQR*

# CHƯƠNG V. KIỂM ĐỊNH

**Mã nguồn**: [2-Hypothesis\_Testing.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1xu-TmsNxA7V_0IkEciAexE0TGljkNm5G)

## Phân tích hai biến (Bivariate Analysis)

### Yếu tố Nhân khẩu học (Demographics)

* + - **Kiểm định giả thuyết 3**: Số lần trung bình đóng lại BHYT của nhóm cao tuổi (từ 50 tuổi trở lên) không ít hơn 39 lần.

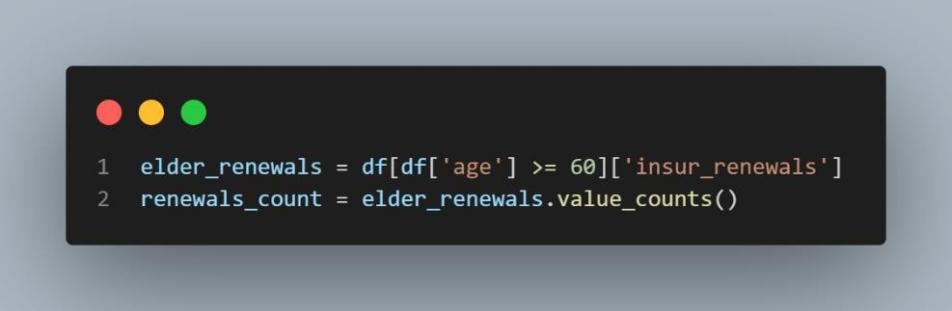
Tương tự với bài toán 1, nhóm xây dựng giả thuyết H0 và H1 tương ứng và chọn mức ý nghĩa α = 0,05.

{H0: 39 ≤ μElder(renewals) H1: 39 > μElder(renewals)

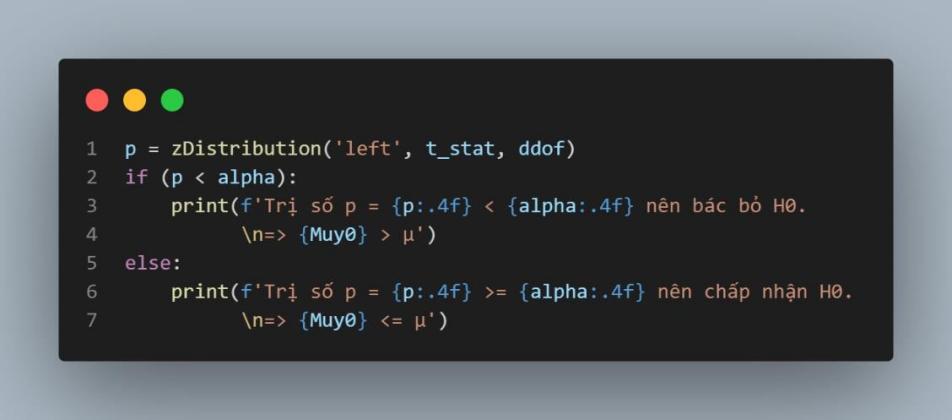
Đặt dấu ≤ đúng chiều ở H0 để biết đây là kiểm định bên trái hay bên phải. Bởi vì, hằng số 39 nằm ở phía bên trái của dấu bé bằng, vậy đây sẽ là kiểm định left-tailed. Nhưng trước hết, nhóm sẽ xây dựng hàm kiểm định một phía của phân phối Z.



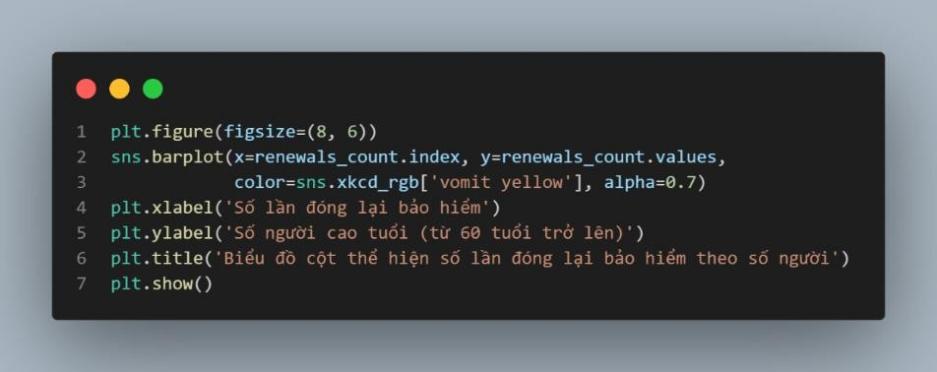
Nhóm tạo hàm UDF 𝐳𝐃𝐢𝐬𝐭𝐫𝐢𝐛𝐮𝐭𝐢𝐨𝐧(), trong đó nhóm sẽ áp dụng hàm cdf() trong thư viện scipy. stats để tìm trị số 𝑝 và so sánh với α để rút ra kết luận kiểm định.

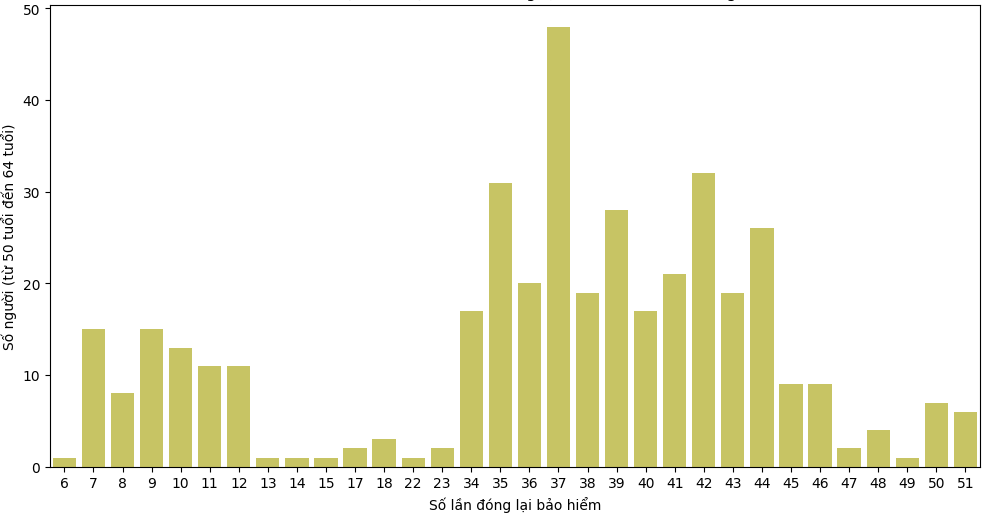


Tiếp theo, nhóm sẽ áp dụng công thức để tìm trị thống kê 𝑧.



Áp dụng hàm UDF vừa tạo, nhóm tìm được p-value của kiểm định này rất nhỏ. Với mức ý nghĩa 5% nên 𝛼 lớn hơn trị số *p*, nhóm quyết định bác bỏ giả thuyết H0 chấp nhận H1. Vậy, nhóm KH có độ tuổi từ 50 đến 64 trung bình sẽ đóng lại BHYT ít hơn 39 lần. Tiếp theo, nhóm vẽ bar chart để tìm hiểu về số lần đóng lại BHYT của nhóm người cao tuổi này.





*Hình 31: Biểu đồ cột thể hiện số lần làm mới bảo hiểm theo nhóm KH cao tuổi (hơn 50 tuổi)*

Nhận xét biểu đồ:

* + - * Nhìn chung, sau quan sát nhóm thấy được rằng số lần đóng bảo hiểm của nhóm khách hàng (KH) cao tuổi có thể chia thành 2 loại:
        + Đầu tiên là nhóm KH chỉ đóng bảo hiểm khi cần thiết, có tổng số lần làm mới bảo hiểm từ 23 lần trở xuống.
        + Thứ hai, là nhóm đóng theo đợt để được hưởng thêm ưu đãi, số lần làm mới bảo hiểm của nhóm này sẽ từ 34 đến 51 lần.
      * Chi tiết hơn:
        + Số lượng KH có số lần đóng lại bảo nhiều hơn 37 lần chiếm đa phần trong tổng số.
        + Lượng KH đóng bảo hiểm chỉ khi cần thiết tương đối thấp khi so với xu hướng này (khoảng 84 người).
* Từ bài toán kiểm định giả thuyết này, nhóm nhìn nhận được xu hướng những người càng cao tuổi thì càng có khả cao làm mới (đóng lại) BHYT nhiều hơn. Nếu vậy, các tổ chức phát hành bảo hiểm có thể hướng tới đối tượng KH là người cao tuổi để tăng doanh số cho doanh nghiệp.

### Yếu tố Địa lý (Geography)

* + - **Kiểm định giả thuyết 4:** Các bang ở các vùng khác nhau có xu hướng chi tiền cho BHYT khác nhau.

Các kiểm định giả thuyết ở phần trên cho thấy rõ về ảnh hưởng của các nhân tố nhân khẩu học tác động vô cùng lớn đến quyết định chi tiền và làm mới BHYT. Vậy liệu yếu tố địa lý có tác động như thế nào đến những quyết định này?

Nhóm đặt ra giả thuyết về sự khác nhau giữa các vùng của nước Mỹ trong việc chi tiền cho BHYT. Các vùng địa lý ở Mỹ được chia thành 5 khu vực khác nhau nên nhóm sẽ tiến hành kiểm định ANOVA một chiều để kiểm định giả thuyết H0 cả 5 giá trị trung bình của 5 khu vực là như nhau.

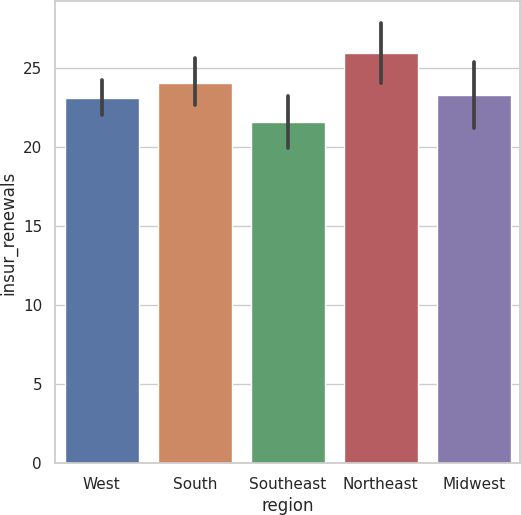
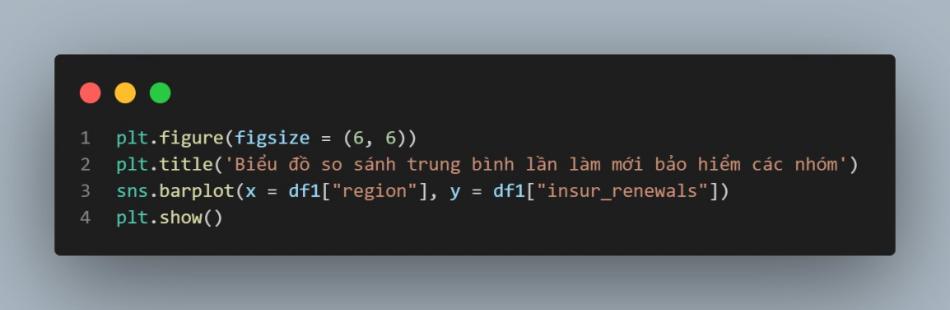
* **Xây dựng cặp giả thuyết**:

H0: μWest = μMidwest = μNortheast = μSoutheast = μSouth

{ H : Có ít nhất một giá trị μ khác biệt

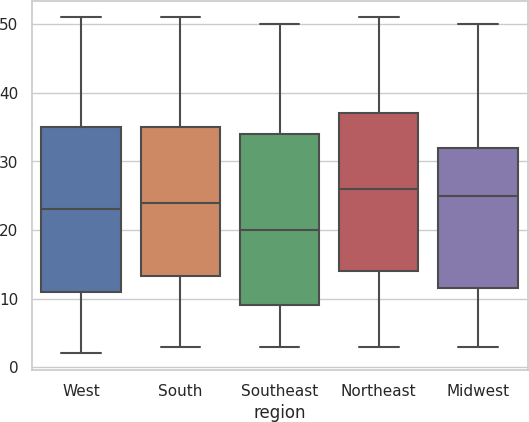
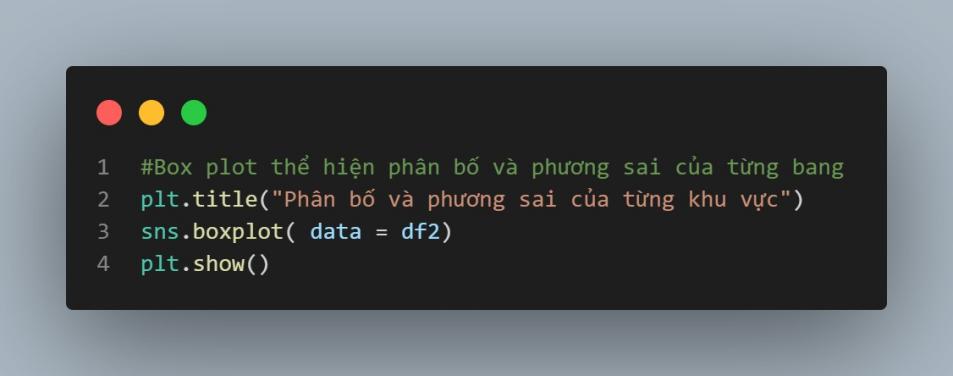
1

Để quan sát tốt hơn về các khu vực được kiểm định nhóm tiến hành trực quan hóa dữ liệu các chỉ số thống kê và phân phối của các vùng bằng biểu đồ thanh và biểu đồ hộp.



*Hình 32: Biểu đô cột so sánh trung bình lần làm mới bảo hiểm theo các nhóm khu vực*

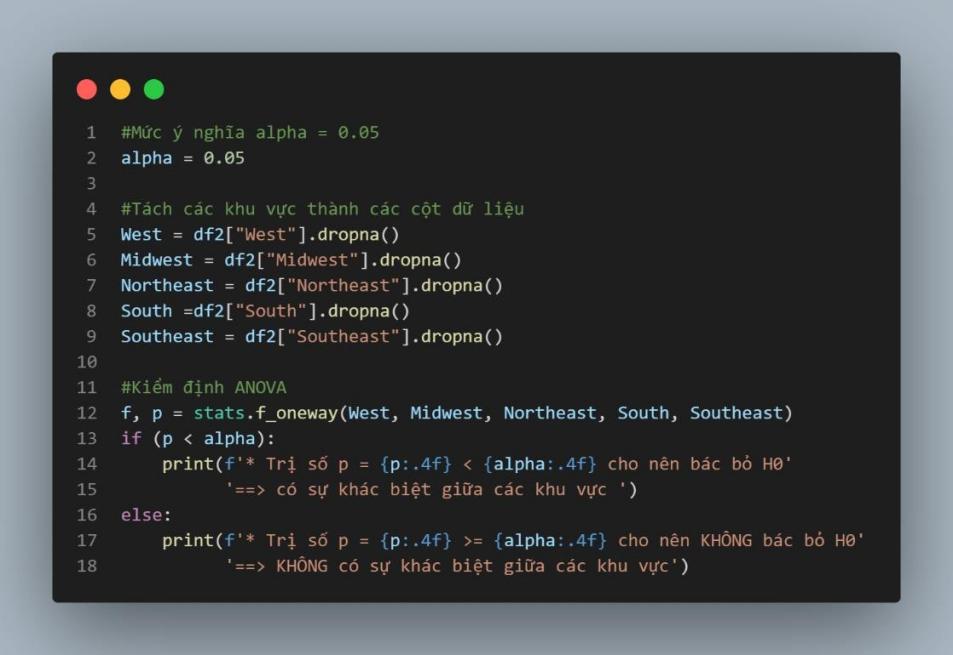
Biểu đồ trung bình lần làm mới bảo hiểm cho thấy các giá trị trung bình ở các vùng Southeast, Northeast chênh lệch tương đối lớn. Trong khi đó các giá trị của ba vùng West, Midwest và South gần như không có sự chênh lệch đặc biệt là West và Midwest.



*Hình 33: Biểu đồ hộp thể hiện phân bố và phương sai theo từng khu vực*

Biểu đồ hộp cho nhóm thấy được độ phân tán giá trị của từng khu vực không phải là quá lớn. Tuy nhiên độ phân tán của các giá trị trung bình các khu vực là tương đối cao, đặc biệt là của vùng Southeast và Northeast so với các vùng còn lại. Điều này cho thấy rất có khả năng giá trị trung bình lần làm mới bảo hiểm của 2 vùng này khác biệt so với các vùng khác.

Như vậy dựa vào hình ảnh biểu đồ nhóm có thể nhận thấy rằng các vùng khác nhau có ảnh hưởng rất lớn đến quyết định làm mới bảo hiểm của KH. Để khẳng định nhận định này nhóm sẽ tiến hành thực hiện kiểm định ANOVA một chiều bằng hàm f\_oneway của thư viện scipy.stats.



Với mức ý nghĩa alpha = 5% kết quả kiểm định cho thấy giá trị p-value = 0,97% thấp hơn nhiều so với mức ý nghĩa alpha. Điều này có nghĩa là xác xuất mắc sai lầm khi bác bỏ giả thuyết H0 là rất thấp.

Kết luận: Bác bỏ giả thuyết H0 và chấp nhận giả thuyết H1: Có ít nhất một giá trị μ

khác biệt so với phần còn lại.

Từ kết luận rút ra được từ kiểm định trên. Nhóm đề xuất cho các doanh nghiệp trong ngành BHYT nên nghiên cứu thị trường theo từng khu vực địa lý, đặc biệt là các vùng Southeast và Northeast. Từ đó, các công ty nên triển khai các chiến dịch marketing phù hợp đối với các vùng riêng biệt để tăng hiệu quả kinh doanh một cách tối đa.

## Phân tích đa biến (Multivariate Analysis)

Tiếp theo, nhóm sẽ kiểm định giả thuyết với số lượng mẫu nhiều hơn 2 biến về chủ đề thói quen và xu hướng sống (lối sống) của KH.

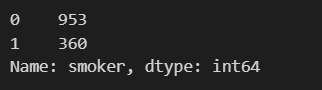
Để chắc chắn về việc loại bỏ biến **is\_male** là đúng đắn, nhóm một lần nữa thực hiện xây dựng thêm một kiểm định để tìm hiểu rằng: giới tính có mang lại bất kỳ giá trị thông tin nào cho bộ dữ liệu hay không:

* **Kiểm định giả thuyết 5**: Số tiền mà người nghiện thuốc lá chi trả cho BHYT so với người bình thường có giống nhau không? Đồng thời, kiểm định xem thói quen hút thuốc có liên quan đến giới tính của KH hay không?

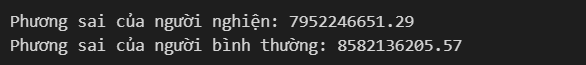
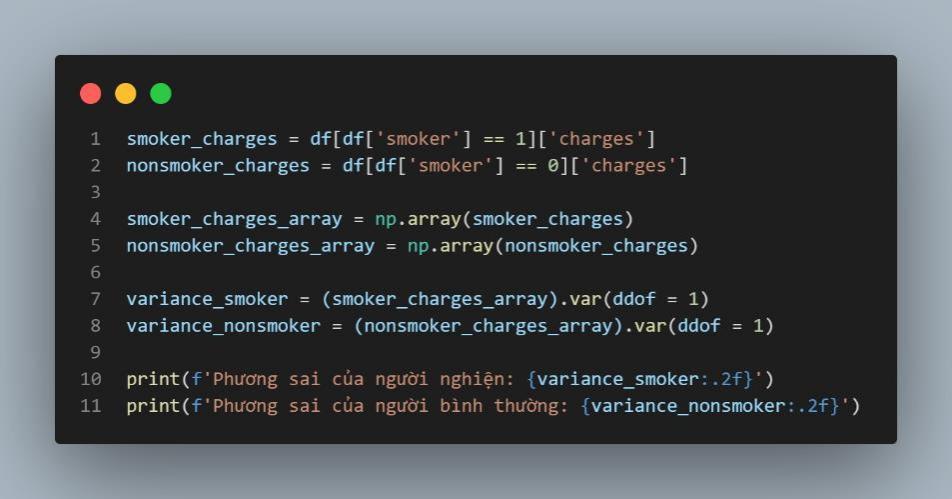
Nhóm xây dựng 2 bài toán kiểm định tương ứng với H0 và H1 cho từng giả thuyết cần kiểm định của bài toán 5 này và chọn độ tin cậy cho bài toán là 95%. Chi tiết:

* 1. : {H0: μSmoker(charges) = μnon−Smoker(charges) H1: μSmoker(charges) ≠ μnon−Smoker(charges)
  2. : {H0: "smoker" và "is\_male" là hai biến độc lập H1: "smoker" và "is\_male" là hai biến phụ thuộc

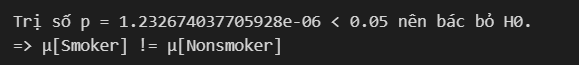
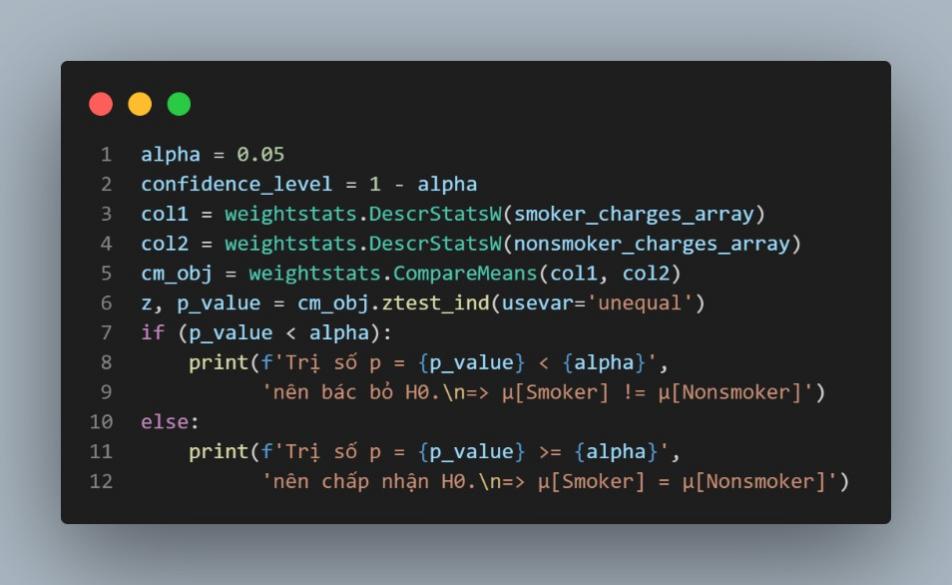
Đầu tiên, nhóm xây dựng đoạn mã để nhìn tổng quan số lượng người sử dụng thuốc lá, kết quả trả về số lượng mẫu khá chênh lệch khi số người không hút thuốc gấp khoảng 2,65 lần với số người hút thuốc.



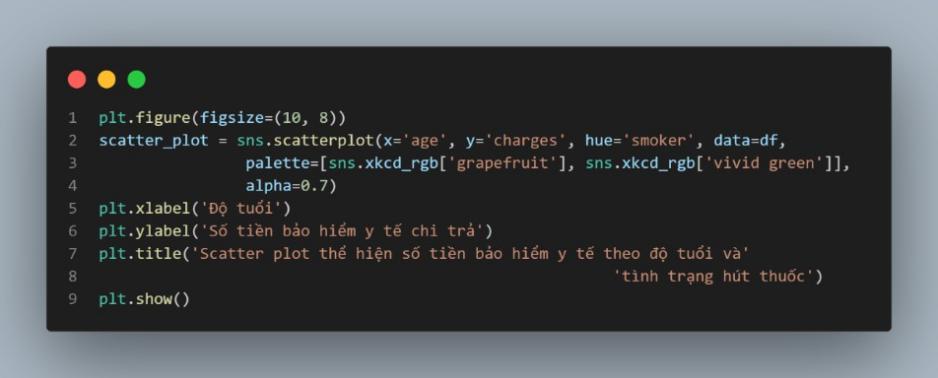
Với vấn đề đầu tiên, việc sử dụng hàm tìm trị thống kê để thực hiện cho việc kiểm định, nhóm cần so sánh phương sai của hai biến để sử dụng làm tham số trong hàm. Nhóm tính toán và thấy được rằng Var(Smoker) không giống với Var(Non-Smoker).

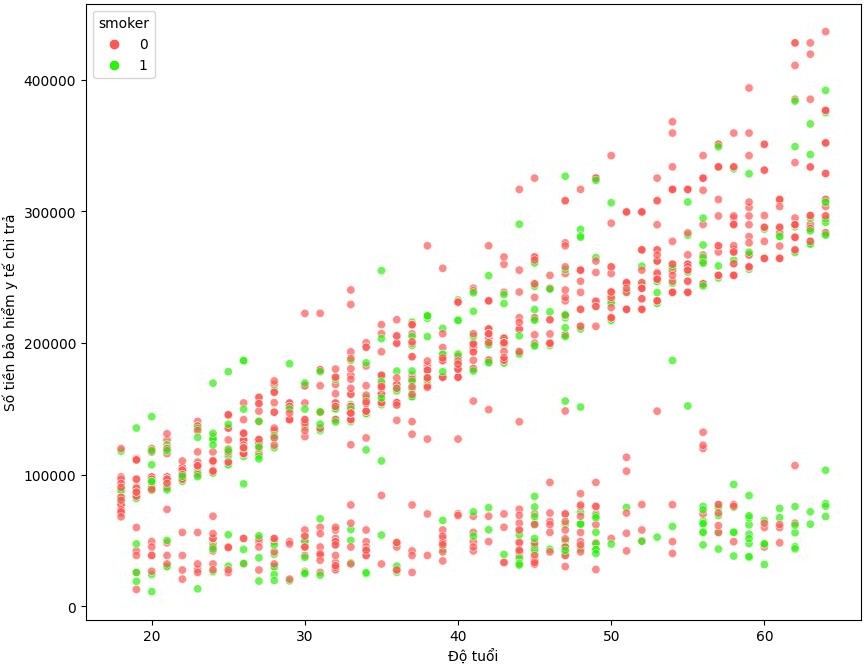


Với phương sai của hai biến không bằng nhau, nhóm sử dụng Phương pháp kiểm định dựa trên trị số 𝑝 bằng hàm ztest\_ind() với tham số usevar là “unequal”. Từ đó, nhóm tìm được trị thống kê 𝑧 và trị số 𝑝 để đi đến kết luận.



Với mức ý nghĩa 5%, nhóm thấy trị số p lớn hơn α vậy nên chấp nhận giả thuyết H0 và kết luận trung bình số tiền chi trả cho bảo hiểm của nhóm người có lối sống nghiện thuốc lá sẽ khác biệt với số tiền của người không nghiện thuốc lá. Để trực quan hơn, nhóm biểu diễn biểu đồ phân tán để dễ tìm ra xu hướng tổng quan từ giả thuyết này.





*Hình 34: Biểu đồ phân tán thể hiện số tiền chi trả bảo hiểm theo độ tuổi và lối sống hút thuốc của mỗi người*

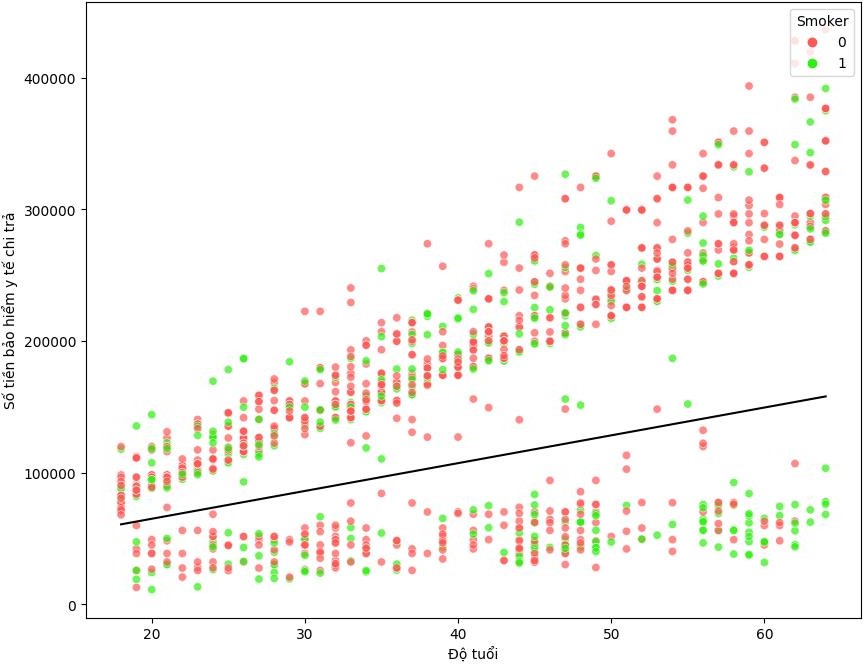
Nhận xét biểu đồ:

Tổng quát biểu đồ phân tán, nhóm thấy được số lượng chấm đỏ chiếm số lượng nhiều hơn số lượng chấm xanh lá rất nhiều. Nhóm nhận thấy điều này là dễ hiểu bởi khi chạy hàm kiểm tra số lượng người hút thuốc phía trên, số người không hút thuốc (đại diện bởi màu đỏ) nhiều hơn gần khoảng 2,7 lần so với đại diện chấm màu xanh.

Vậy từ đó, tổng tiền BHYT trả cho người không nghiện thuốc lá chắc chắn sẽ nhiều hơn tổng tiền của nhóm người nghiện. Biểu đồ có thể chia thành 2 phần riêng để phân tích, nhóm vẽ đường tuyến tính màu đen với:

* Phần phía dưới của đường tuyến tính là phần trung bình thấp mà số tiền bảo hiểm chi trả, gọi là phần (a).
* Phần còn lại nằm phía trên của đường tuyến tính, gọi là phần (b).





*Hình 35: Hình 33 kết hợp đường tuyến tính phân chia vùng diện tích*

Từ biểu đồ, nhóm thực hiện khai thác một số thông tin nổi bật và thấy được rằng:

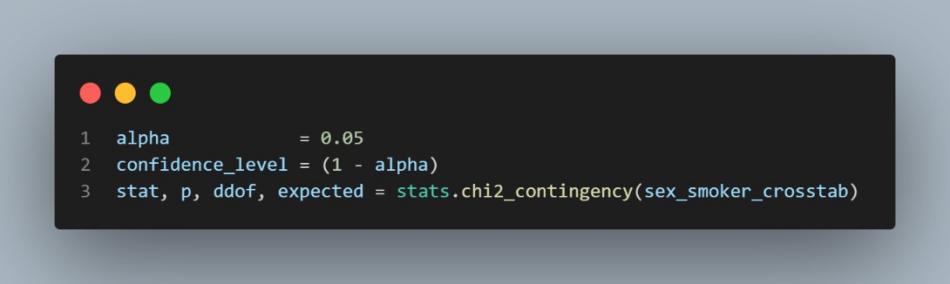
* Phần (a) có tỷ lệ người hút thuốc chiếm cao hơn so với người không hút thuốc, nhất là ở phần diện tích nằm trong khoảng độ tuổi từ gần 50 đến trên 60 tuổi.
* Còn phần (b), tỷ lệ người không hút thuốc chiếm nhiều hơn. Nhưng số tiền BHYT nhiều nhất được nhóm người hút thuốc chi trả sẽ lớn hơn so với không hút thuốc ở khoảng tuổi từ gần 20 đến 50 tuổi. Còn khoảng từ 50 trở lên thì xu hướng số tiền này thuộc về nhóm người không hút thuốc.
* Khi nhìn theo khuynh hướng độ tuổi, những người càng lớn tuổi thì có xu hướng trả tiền bảo hiểm càng nhiều hơn, bất chấp cho việc người đó có hút thuốc hay không.

Vậy nhóm có thể kết luận rằng lối sống hút thuốc và tuổi tác cũng là yếu tố ảnh hưởng đến chi phí bảo hiểm chăm sóc sức khỏe; bài toán (1) coi như hoàn tất, nhóm tiếp tục kiểm định giả thuyết cho bài toán (2) để xem yếu tố giới tính có ảnh hưởng lên lối sống hút thuốc này hay không.

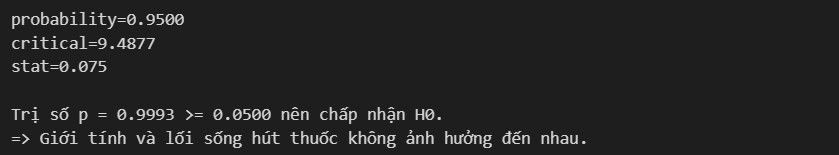
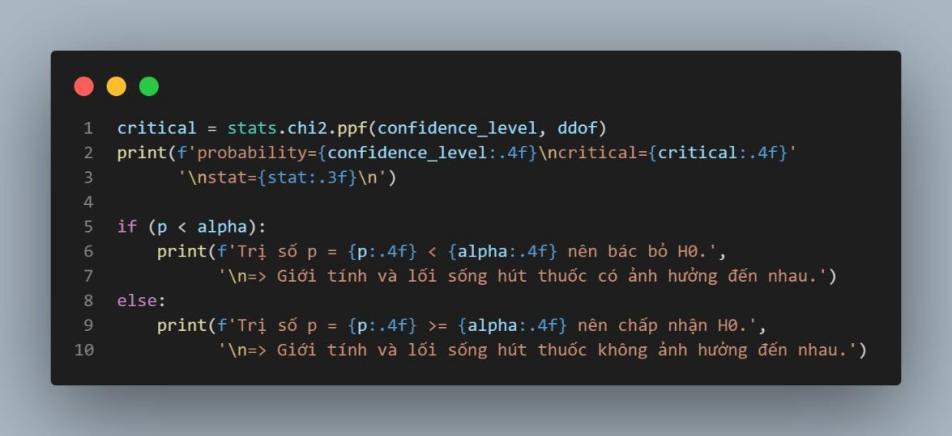
Đối với vấn đề thứ hai, nhóm xác định đây là bài toán kiểm định tương quan đối với hai mẫu là biến phân loại. Vậy nhóm sẽ sử dụng Chi-square testing để thực hiện, đầu tiên, nhóm lập bảng Crosstab để nhìn tổng quan số lượng giới tính phân chia theo lối sống hút thuốc.



Nhóm nhận thấy sau khi phân mẫu thu thập theo bốn nhóm nhỏ, số lượng của mỗi nhóm có sự phân chia chủ yếu theo xu hướng hút thuốc bởi số lượng người, hoặc nam hoặc nữ, theo lối sống không hút thuốc sẽ nhiều hơn gần gấp ba lần so với nhóm người cùng giới tính đó nhưng theo lối sống hút thuốc.

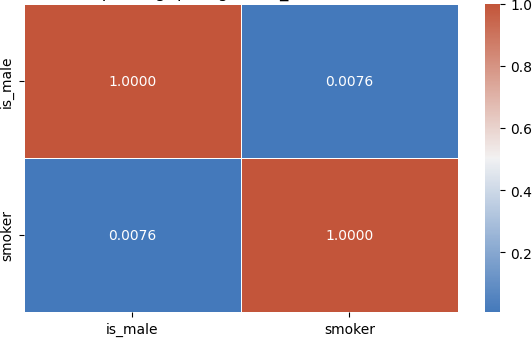


Nhóm sử dụng Phương pháp kiểm định dựa trên trị số p bằng hàm chi2\_ppf() với tham số ddof được tính từ hàm chi2\_contigency() trong thư viện scipy. stats.



Với độ tin cậy được xác định từ trước là 95%, nhóm nhận thấy p-value lớn hơn 𝛼 vậy nên nhóm sẽ chấp nhận giả thuyết H0 và kết luận rằng giới tính và lối sống hút thuốc không ảnh hưởng đến nhau. Nhóm muốn kiểm tra lại kết quả vừa

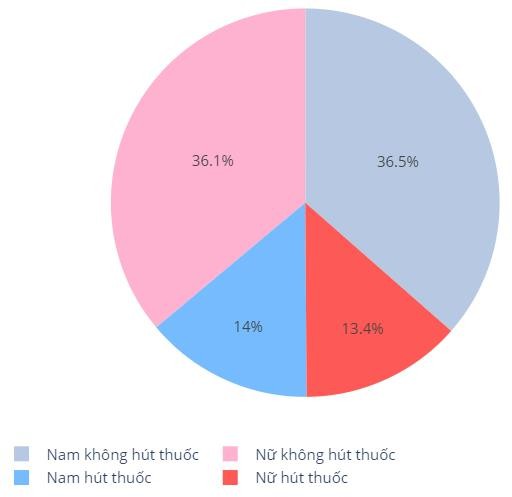
đạt được nên đã sử dụng Heatmap để tìm sự tương quan giữa hai mẫu này và kết quả trả về khá hợp lý.



*Hình 36: Ma trận tương quan giữa giới tính và lối sống hút thuốc*

Dựa vào Heatmap Correlation, hai biến **is\_male** và **smoker** không có sự tương quan cao bởi chỉ số giữa chúng khi tính bằng biểu đồ chỉ là 0,0076. Nhóm sẽ tiếp tục vẽ Pie Chart để trực quan bảng số liệu crosstab trước đó, và để tương tác được, nhóm sẽ thử biểu diễn bằng thư viện plotly. express và kiểm tra lại số liệu có chính xác không.





*Hình 37: Biểu đồ tròn thể hiện tỷ lệ người theo lối sống hút thuốc và giới tính*

Kết quả trả về chính xác như bảng chéo (Crosstab) trước đó khi số lượng nữ giới không có lối sống hút thuốc là 474, tương tự với 3 nhóm còn lại.

Từ những thông tin mà nhóm vừa tổng hợp được, cho thấy biến **is\_male** không mang lại bất kỳ giá trị thông tin nào, hay việc loại bỏ biến không làm ảnh hưởng đến bất cứ biến nào trong bộ dữ liệu.

* **Kiểm định giả thuyết 6**: Số lượng KH có thói quen hút thuộc ở độ tuổi vị thành niên (từ 18 tuổi đến 29 tuổi) sử dụng BHYT bằng nhóm KH ở độ tuổi trung niên (từ 30 tuổi đến 49 tuổi).

Bài toán này khá giống với một phần nhỏ của vấn đề 1 thuộc bài toán kiểm định số 5, nhóm có thể chỉ nhìn vào biểu đồ phân phối (Scatter Plot) mà đưa ra những kết luận sơ bộ đối với giả thuyết này. Nhưng ở đây, nhóm muốn đi sâu vào giả thuyết của hai nhóm người này và xây dựng biểu đồ trực quan để nhìn được rõ hơn về tổng số tiền mà hai nhóm đã chi trả cho bảo hiểm.

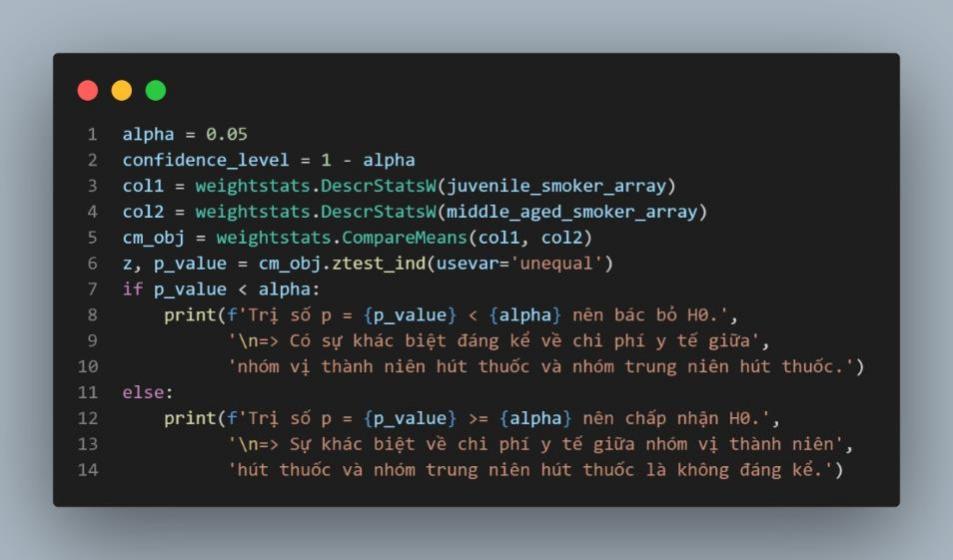
Vậy, nhóm sẽ xây dựng giả thuyết kiểm định H0 và H1 tương ứng với từng giả thuyết đối với bài toán 6 và chọn độ tin cậy cho bài toán là 95%. Chi tiết:

* + H0: Sự khác biệt về **chi phí đóng bảo hiểm** giữa **nhóm thanh thiếu niên sử dụng thuốc lá** và **nhóm trung niên sử dụng thuốc lá** là không đáng kể.
  + H1: Có sự khác biệt đáng kể về **chi phí đóng bảo hiểm** giữa **nhóm thanh thiếu niên sử dụng thuốc lá** và **nhóm trung niên sử dụng thuốc lá**.

Đầu tiên, nhóm xác định đây là dạng kiểm định giả thuyết sử dụng phương pháp p-value, vậy nhóm sẽ kiểm tra phương sai của hai nhóm người, cụ thể:



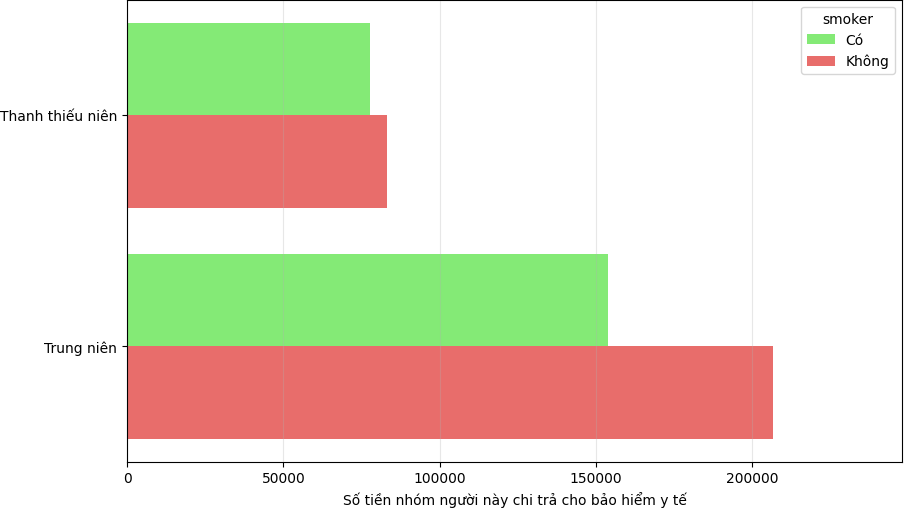
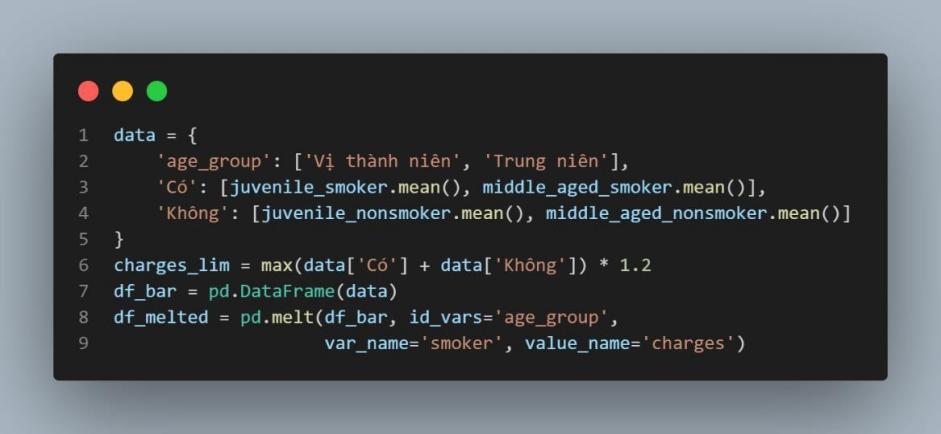
Tiếp theo, nhóm sử dụng hàm ztest\_ind() để tìm trị thống kê 𝑧 và p-value nhằm đi đến kết luận. Như nhóm đã so sánh, hai phương sai của nhóm này khác nhau, vậy Parameter usevar sẽ được để là "unequal" và chạy code.



Từ đoạn mã, vì p-value thấp hơn mức ý nghĩa được xác định từ trước, nhóm kết luận trung bình chi phí BHYT chi trả bởi nhóm thanh thiếu niên có hút thuốc sẽ có sự khác biệt đáng kể so với số tiền cũng được chi trả nhưng bởi nhóm người trung niên có sử dụng thuốc lá.

Để nhóm thuận tiện cho việc quan sát, nhóm sẽ vẽ Stacked Horizontal Column Chart để nhìn chi phí mà hai nhóm độ tuổi này chi trả cho BHYT, và sẽ

xác định hơn nếu mỗi nhóm độ tuổi này được băm nhỏ thành hai nhóm có và không sử dụng thuốc lá.



*Hình 38: Biểu đồ cột thể hiện số tiền chi trả bảo hiểm của nhóm độ tuổi và lối sống hút thuốc*

Từ biểu đồ thanh ngang, nhóm nhìn nhận được một số thông tin nổi bật đối với bộ dữ liệu thu thập được, như sau:

* Tỷ lệ chi trả BHYT của nhóm người không hút thuốc luôn cao hơn đối với nhóm người có hút thuốc mặc cho lối sống đó có ở bất kỳ độ tuổi nào.

*Điều nào có thể giải thích là do số mẫu lấy trong bộ dữ liệu của nhóm người không hút thuốc lớn hơn nhiều so với nhóm người có hút thuốc, lớn hơn gần gấp 3 lần, vậy nên tổng số tiền chi trả cho bảo hiểm cũng sẽ bị ảnh hưởng.*

* Số tiền chi trả cho bảo hiểm của cả hai nhóm người hút thuốc và không hút thuốc đều sẽ tăng dần theo độ tuổi.

*Điều này có thể lý giải do việc hút thuốc ở độ tuổi vị thành niên (từ 18 tuổi đến 29 tuổi) không phổ biến bằng ở độ tuổi trung niên (từ trên 30 tuổi đến 49 tuổi). Vậy nên việc lấy mẫu thì khoảng tuổi nhỏ hơn này cũng sẽ ít có trường hợp hơn, từ đó, số tiền mà nhóm tuổi nhỏ hơn này chi trả cho bảo hiểm cũng sẽ ít hơn theo.*

* Ngoài ra, nhóm đã biết được số mẫu của nhóm người không hút thuốc sẽ lớn hơn gần gấp 3 lần so với nhóm người hút thuốc, tuy vậy khi nhìn vào Bar chart, nhóm nhận thấy thanh biểu đồ của nhóm người không hút thuốc chi trả cho bảo hiểm chỉ dài hơn một chút so với nhóm hút thuốc. Vậy có thể nhìn được xu hướng, nếu mà nhóm người đó có xu hướng sử dụng thuốc lá, vậy khả năng nhóm người đó sẽ chi trả cho BHYT nhiều chi phí hơn so với nhóm người không có xu hướng hút thuốc.

Vậy các công ty bảo hiểm có thể đặc biệt quan tâm đến nhóm người có xu hướng sử dụng thuốc lá hoặc nhóm người có độ tuổi cao để giới thiệu về các loại bảo hiểm, bởi từ số liệu, số tiền mà hai nhóm này chi trả cho bảo hiểm sẽ có xu hướng tăng đối với những thông tin mà nhóm tìm được qua bài toán kiểm định này.

* **Kiểm định giả thuyết 7**: Yếu tố hôn nhân gia đình có làm ảnh hưởng đến việc chi trả cho BHYT hay không?

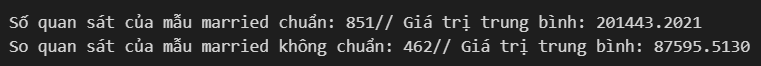
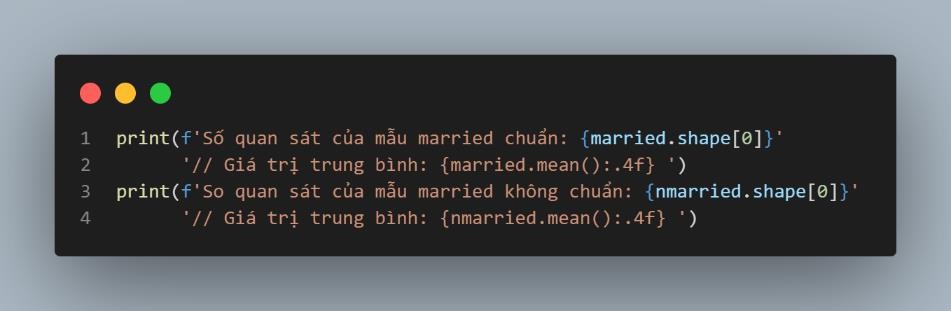
Để kiểm tra giả thuyết này, nhóm sẽ lấy trung bình giá trị **charges** (tổng số tiền đã chi trả cho BHYT) theo 2 nhóm đã kết hôn và chưa kết hôn.

Bắt đầu thiết lập cặp giả thuyết H0 và H1 theo ý tưởng trên. Chọn mức ý nghĩa là 5%.

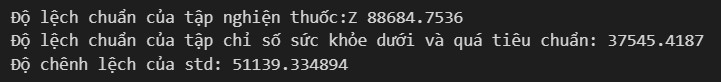
{H1: μmarriedcharges = μnmarriedcharges H0: μmarriedcharges ≠ μnmarriedcharges

Nhóm tiến hành tính toán các số liệu thống kê của 2 nhóm người và trực quan hóa bằng biểu đồ thanh:

* Tính giá trị trung bình và số quan sát của hai mẫu:

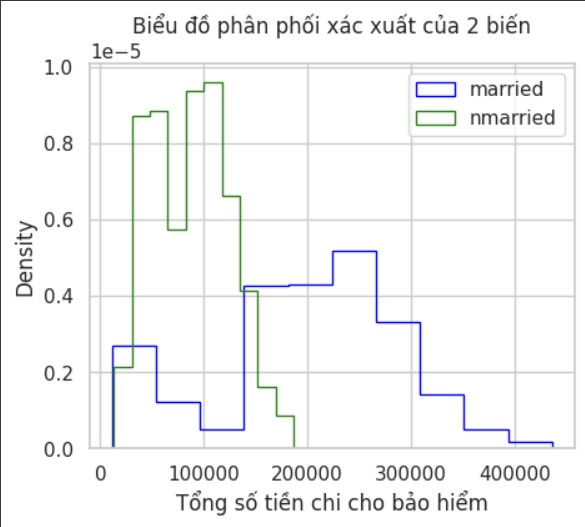


* Tính giá trị độ lệch chuẩn của 2 nhóm:



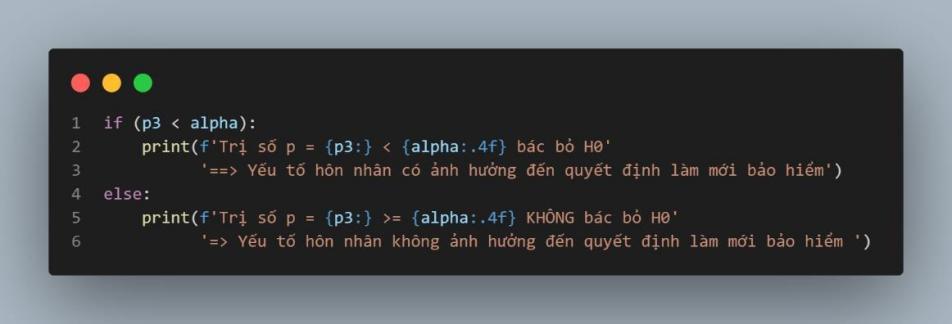
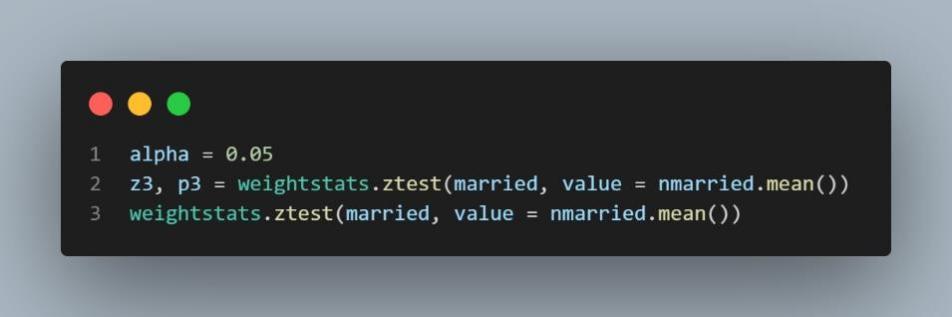
* + Nhóm nhận thấy giá trị trung bình của hai nhóm có chênh lệch vô cùng đáng kể. Điều này chứng tỏ rằng ảnh hưởng của yếu tố gia đình đến việc chi trả bảo hiểm là vô cùng cao.
  + Tương tự với độ lệch chuẩn của hai nhóm cũng cho thấy rõ mức độ phân tán của nhóm có gia đình lớn hơn rất nhiều so với nhóm không có gia đình. Mức độ chênh lệch của nhóm married là hơn gấp đôi so với nhóm nmarried.
* Để quan sát rõ hơn về độ phân tán của hai nhóm. Nhóm quyết định sẽ trực quan hóa phân phối của hai nhóm bằng biểu đồ Histogram:





*Hình 39: Biểu đồ phân phối xác suất tổng chi phí trả cho bảo hiểm của hai biến*

* + Có thể thấy từ biểu đồ Historam, nhóm người chưa lập gia đình có xu hướng chi trả số tiền ở mức từ 0 đến 10.000. Trong khi đó, nhóm đã lập gia đình chi cho bảo hiểm ở mức hoàn toàn người lại từ mức 15.000 cho đến 30.000.
  + Mức độ phân tán của các giá trị trong nhóm có gia đình có thể dễ dàng nhận ra phân bố vô cùng rộng. Trong khi đó mức độ phân tán của những người chưa có gia đình hầu hết chỉ tập trung trong khoảng từ 0 đến chỉ 15.000.
* Như vậy, qua các yếu tố thống kê được quan sát qua hình ảnh biểu đồ trực quan, nhóm dự đoán chắc rằng yếu tố gia đình có làm ảnh hưởng đến việc chi tiêu cho BHYT. Để khẳng định dự đoán này, nhóm tiến hành thực hiện kiểm định Z (do quan sát của tập mẫu >30) cho cặp giả thuyết đã xác định ở trên.



* + Sử dụng hàm ztest() của thư viện statsmodels với mức ý nghĩa là 5%, nhóm thu được kết quả của kiểm định cho thấy rằng trị số p (p-value) nhỏ hơn rất nhiều so với mức ý nghĩa 5% tức là tỉ lệ mắc sai lầm khi bác bỏ H0 là rất thấp.

Kết luận: Bác bỏ giả thuyết H0 và chấp nhận giả thuyết H1: Yếu tố hôn nhân có làm ảnh hưởng đến việc chi trả cho BHYT.

* **Kiểm định giả thuyết 8:** KH có thói quen hút thuốc có xu hướng làm mới BHYT nhiều hơn những KH không có thói quen hút thuốc.

Từ bài toán kiểm định số 6, nhóm trực quan được biểu đồ thanh thể hiện và nhận xét được rằng “*xu hướng nhóm người sử dụng thuốc lá sẽ chi trả cho BHYT nhiều chi phí hơn so với nhóm người không có xu hướng hút thuốc*”.

Tuy nhiên, khi nhóm nhìn vào ma trận tương quan ở trên, xu hướng của chi phí BHYT tăng cũng kéo theo số lần làm mới BHYT tăng, từ đó, có thể kết luận rằng, nếu nhóm người hút thuốc trả nhiều tiền hơn cho BHYT cũng sẽ đóng tiền cho BHYT nhiều hơn.

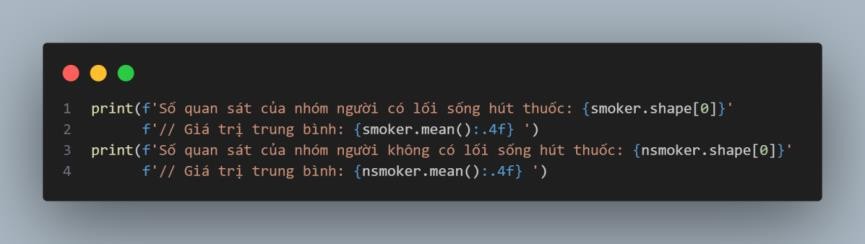
Nhưng chỉ dựa vào ma trận và các kết luận gián tiếp thì sẽ không thể đảm bảo cơ sở về mặt thống kê. Vậy kiểm định số 8 này sẽ giúp nhóm đi trả lời cho câu hỏi: **Liệu những người hút thuốc, vì lo lắng cho sức khỏe của mình, sẽ có xu hướng làm mới BHYT nhiều hơn hay không?**

Và để trả lời cho câu hỏi đó nhóm đã quyết định xây dựng giả thuyết H0 và giả thuyết H0 như câu hỏi trên, với mức ý nghĩa được chọn ở đây là α = 5% và độ tin cậy là 95%:

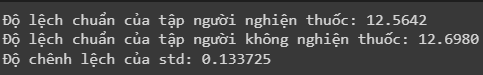
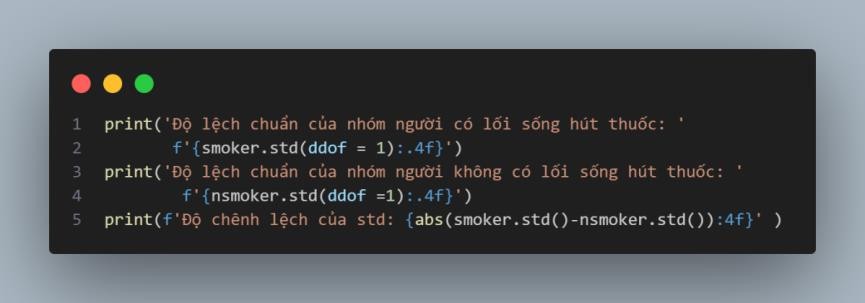
{H0: μsmokerinsur\_renewals ≤ μnsmokerinsur\_renewals H1: μsmokerinsur\_renewals > μnsmokerinsur\_renewals

Nhóm tiến hành tính toán các số liệu thống kê của hai nhóm người:

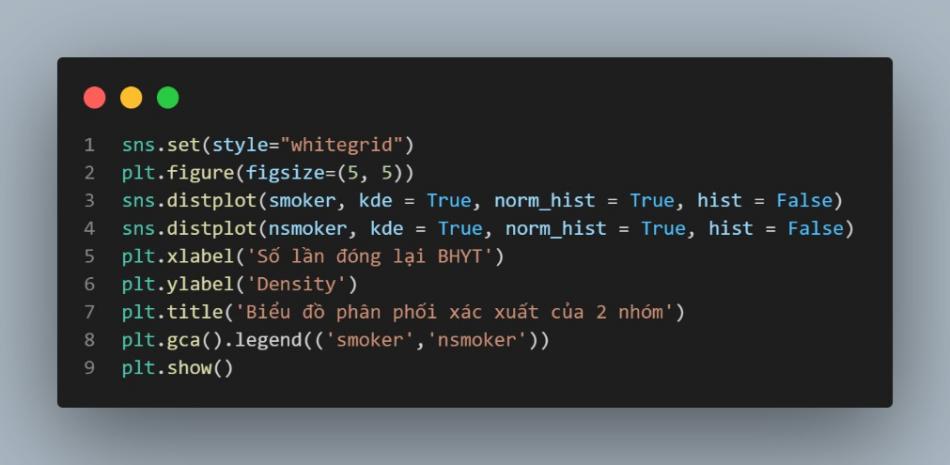
* Tính toán giá trị trung bình và số quan sát của 2 nhóm:

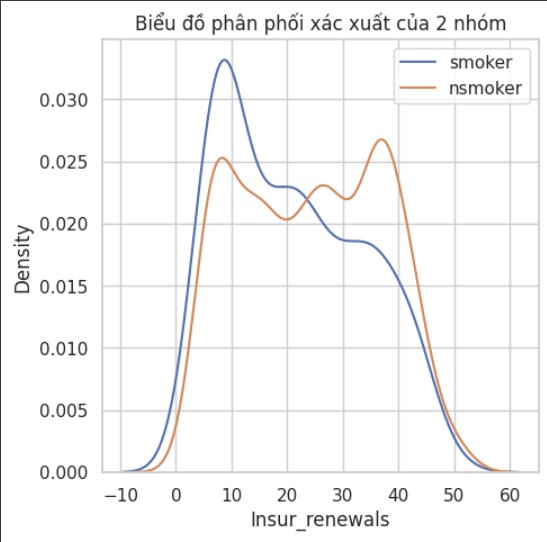


* Tính toán giá trị độ lệch chuẩn của hai nhóm:



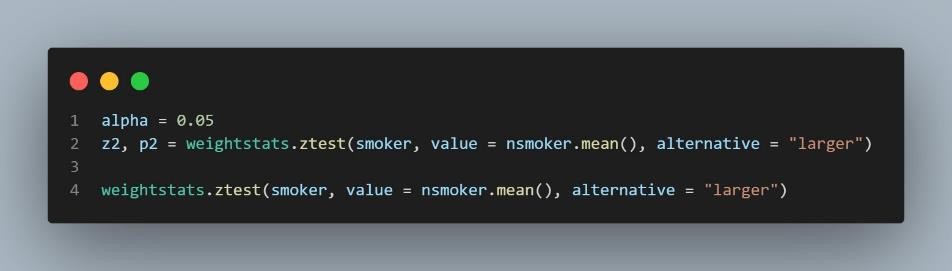
* + Về giá trị trung bình, nhóm có thể thấy rằng chênh lệch của cả hai nhóm là tương đối lớn. Điều này sẽ ảnh hưởng rất lớn đến kết quả kiểm định của giả thuyết.
  + Độ lệch chuẩn của cả hai nhóm vô cùng cao so với giá trị trung bình chứng tỏ cả hai nhóm đều có giá trị phân tán mạnh so với giá trị trung bình. Ngoài ra, độ chênh lệch của 2 độ lệch chuẩn khi tính toán cho ra giá trị chỉ khoảng 0,1337, rất nhỏ so với độ lớn của 2 giá trị độ lệch chuẩn
* Nhóm thực hiện vẽ biểu đồ Distribution để trực quan hóa phân phối của hai nhóm người:





*Hình 40: Biểu đồ đường thể hiện phân phối xác suất theo số lần làm mới BHYT của hai nhóm*

* + Nhóm nhận thấy phân phối của hai nhóm về việc nghiện thuốc có những điểm khá tương đồng với phân phối của hai nhóm về chỉ số cơ thể (BMI) đã quan sát ở kiểm định trước đó.
  + Phân phối của nhóm người không có lối sống hút thuốc phân phối khá đồng đều, trong khi đó phân phối của nhóm người hút thuốc tập trung cao ở mức từ 0 đến 10 lần và xu hướng làm mới ở những mốc cao hơn là thấp hơn hẳn.
* Điều này chứng tỏ việc nghiện hút thuốc sẽ có ảnh hưởng đến việc làm mới bảo hiểm nhiều hay ít hơn so với lối sống không hút thuốc.
  + Nhóm tiến hành kiểm định giả thuyết đã thiết lập ở trên để biết được chính xác việc ảnh hưởng này như thế nào. Số lượng quan sát của mẫu ở đây là lớn (>30 quan sát) cho nên nhóm sẽ tiến hành thực hiện kiểm định bằng kiểm định Z.



* + - Sử dụng hàm ztest() của thư viện statsmodels với mức ý nghĩa là 5%, nhóm thu được kết quả của kiểm định cho thấy rằng trị số p (p-value) lớn hơn rất nhiều so với mức ý nghĩa 5% tức là tỷ lệ mắc sai lầm khi bác bỏ H0 là tương đối cao.

Kết luận: Không có cơ sở để bác bỏ giả thuyết H0: KH có thói quen hút thuốc thường làm mới bảo hiểm ít hơn KH không có thói quen hút thuốc.

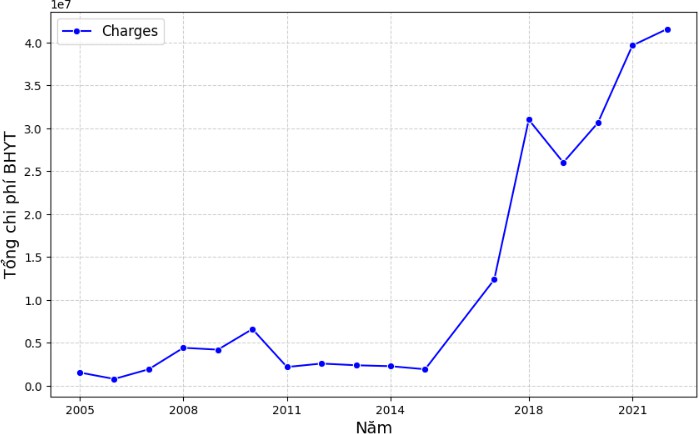
# CHƯƠNG VI. BIỂU DIỄN TRỰC QUAN DỮ LIỆU NGHIÊN CỨU

**Mã nguồn**: [3-Data\_Visualization.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1f4R2uOlaoF2YoB9GOKs70BOuxYu8QlCu)

## Xu hướng/Time series

* + Phân tích biến thời gian có trong bộ dữ liệu là yếu tố quan trọng để hiểu sâu hơn về nhu cầu của khách hàng (KH) đối với Bảo hiểm Y tế (BHYT), cũng như các xu hướng thay đổi theo thời gian. Trong trường hợp này, biến **most\_recent\_insur\_pur** – thời điểm gần nhất khi KH mua bảo hiểm, được chọn làm biến chính trong phân tích chuỗi thời gian.
  + **Mục tiêu chính**: Xác định và theo dõi sự thay đổi trong nhu cầu mua bảo hiểm qua thời gian. Qua đó phát hiện ra các xu hướng, đánh giá sự biến động, và đặc biệt là tìm hiểu về các yếu tố có thể ảnh hưởng đến quyết định mua bảo hiểm sức khỏe của KH.

1. **Phân tích Tổng số tiền chi trả cho BHYT qua từng năm**

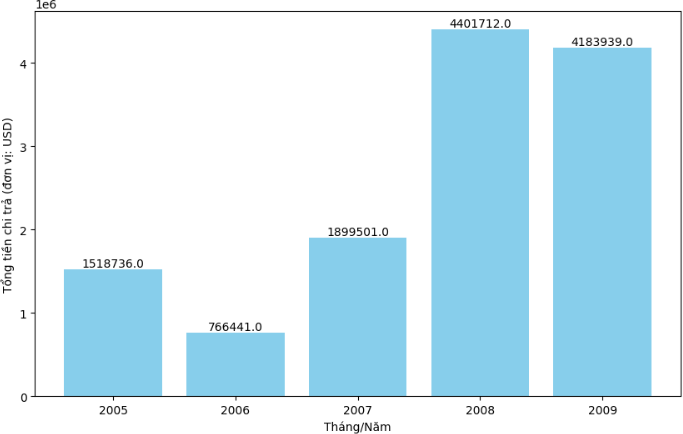


*Hình 41: Biểu đồ Line Plot thể hiện tổng tiền BHYT mà KH đã chi trả từ năm 2005 - 2022*

* + Từ biểu đồ cho biết, nhu cầu về BHYT tại Hoa Kỳ tăng mạnh sau năm 2015. Do ảnh hưởng mạnh của đạo luật Obamacare (Đạo luật Chăm sóc Y tế Giá rẻ - ACA)

[1] bắt buộc người dân Hoa Kỳ phải tham gia BHYT, khiến cho tổng chi phí chi trả cho BHYT từ dưới 0,5 triệu USD đã vượt ngưỡng 30 triệu USD vào năm 2018 (tức tăng trưởng ~500% trong vòng 3 năm). Tuy nhiên, từ ngày 1 tháng 1 năm 2019, việc có BHYT không còn bắt buộc ở cấp độ liên bang [1] khiến tổng chi phí chi trả cho BHYT sụt nhẹ xuống dưới ngưỡng 30 triệu USD, nhưng vẫn cao hơn gấp 3,57 lần ở hầu hết ở các năm từ 2015 đổ lại.

* + Để thấy rõ sự thay đổi đó nhóm tiếp tục quan sát một số biểu đồ sau:
    - Năm năm có tổng số tiền chi trả cho BHYT thấp nhất:



*Hình 42: Biểu đồ thanh thể hiện 5 năm có tổng tiền chi trả cho BHYT thấp nhất*

* + - * Khi BHYT mới được giới thiệu, tổng số tiền mà người dân Hoa Kỳ phải chi trả cho BHYT khá thấp, phản ánh sự phổ biến ban đầu, mức sống và nhu cầu của người dân. Cụ thể, trong một năm, người dân chỉ phải chi trả khoảng 4.024 USD cho BHYT (vào trước những năm 2005) [2], một mức phí khá thấp so với mức phí trung bình hiện tại là 7.739 USD mỗi năm [2], tương đương với mức tăng gấp 1,92 lần.
    - Năm năm có tổng số tiền chi trả cho BHYT cao nhất:

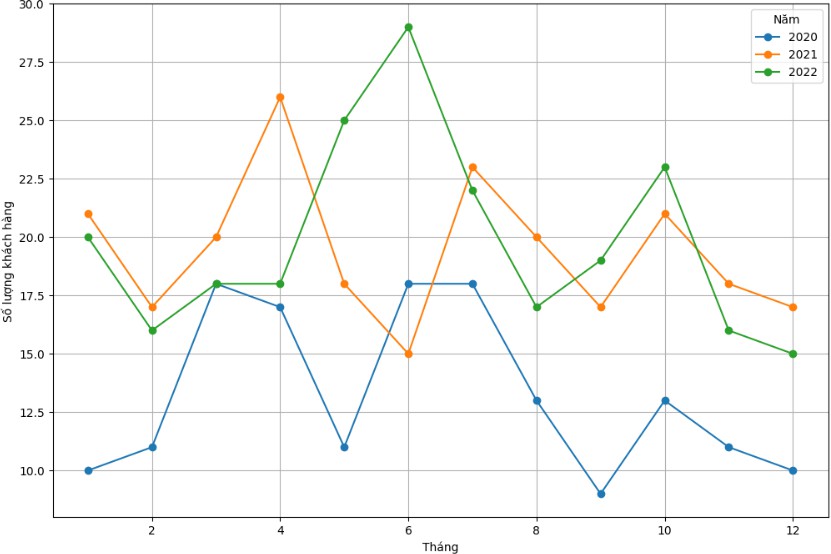


*Hình 43: Biểu đồ thanh thể hiện 5 năm có tổng tiền chi trả cho BHYT cao nhất*

* + - * Sự sụt giảm mạnh mẽ trong tổng số tiền chi trả cho BHYT vào năm 2019 có thể được giải thích bằng việc BHYT không còn bắt buộc và tác động của đại dịch COVID-19. Đặc biệt, tình trạng mất việc và suy thoái kinh tế nghiêm trọng đã khiến tổng số tiền chi trả cho BHYT giảm xuống còn khoảng 26 triệu USD. Tuy nhiên, nhờ sự hỗ trợ của chính phủ, các chương trình như Medicaid, Chương trình Bảo hiểm Y tế Trẻ em (CHIP) và Đạo

luật Chăm sóc Y tế Giá rẻ (ACA) đã giúp những người mất thu nhập đột ngột có thể tiếp cận bảo hiểm miễn phí hoặc với chi phí thấp [3]. Điều này đã góp phần đáng kể vào việc phục hồi chi phí BHYT trong năm 2020.

* + - * Ngoài ra, nhờ các chính sách của Hoa Kỳ nhằm tăng cường tiếp cận dịch vụ chăm sóc sức khỏe thông qua nhiều biện pháp và chương trình khác nhau trong đại dịch, nhận thức về tầm quan trọng của việc tham gia BHYT của người dân đã tăng mạnh. Điều này được chứng minh bằng việc vào năm 2022, tổng số tiền chi trả cho BHYT đã đạt mức cao nhất, với hơn 41 triệu USD.
      * Tiếp tục quan sát biểu đồ sau để hiểu rõ hơn về sự tăng trưởng cụ thể của tổng chi phí BHYT sau đại dịch, từ năm 2020 đến năm 2022:
    - Sự gia tăng số lượng KH tham gia BHYT sau năm 2019:



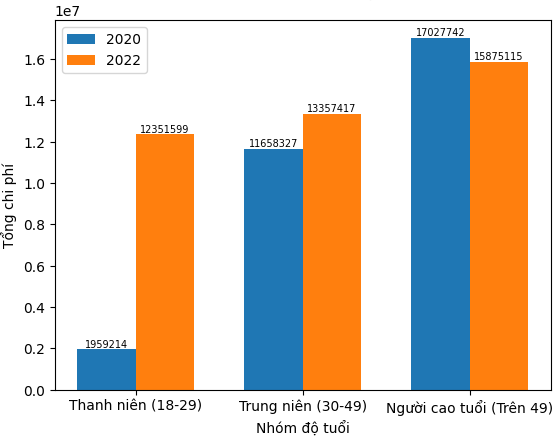
*Hình 44: Biểu đồ Line Plot thể hiện sự gia tăng về số lượng KH đăng ký BHYT theo từng tháng trong 3 năm (2020, 2021, 2022)*

* + - * Dựa vào biểu đồ, nhóm có thể nhận thấy rằng số lượng KH tham gia BHYT đã tăng một cách đồng đều trong suốt 3 năm (2020, 2021, 2022). Đặc biệt, vào các tháng 4, 6 và 10, số lượng KH tăng đột biến, trong khi đó, số lượng lại giảm đáng kể vào các tháng còn lại, nhất là tháng 5 và tháng 9.
      * Để giải thích cho điều này, thì đây là khoảng Thời gian Đăng ký Đặc biệt

- SEP tại Hoa Kỳ, cho phép các cá nhân mua hoặc thay đổi các chương trình BHYT mà không cần phải xét điều kiện, chẳng hạn như mất việc hoặc chuyển nhà [4]. Trong những giai đoạn này, nhu cầu về BHYT sẽ tăng mạnh.

1. **Phân tích Tổng số tiền chi trả cho BHYT theo nhóm độ tuổi sau năm 2019**

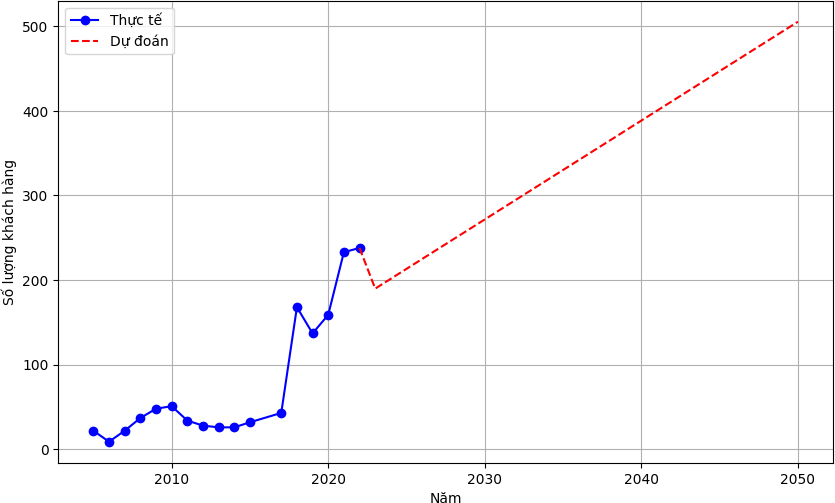
* Phân tích Tổng chi tiêu cho BHYT theo từng độ tuổi qua các năm cung cấp một cái nhìn sâu hơn về cấu trúc chi phí so với việc chỉ xem xét Tổng chi tiêu hàng năm của KH. Từ việc xác định rõ về cách thức chi tiêu cho BHYT trong các nhóm tuổi khác nhau, mà nhóm có thể phát hiện ra các xu hướng hoặc mô hình chi tiêu cụ thể. Điều này không chỉ giúp chính phủ hoặc doanh nghiệp hiểu rõ hơn về cấu trúc chi phí, mà còn hỗ trợ lập kế hoạch tài chính và đưa ra các quyết định tốt hơn trong việc quản lý và điều hành hệ thống BHYT trong tương lai.
* Như đã đề cập từ **Chương III**, nhóm đã thực hiện phân nhóm độ tuổi của KH thành ba nhóm chính, bao gồm:
  + Nhóm KH cao tuổi: Đối với những KH > 49 tuổi;
  + Nhóm KH ở độ tuổi trung niên: Đối với những KH dưới 50 và trên 29 tuổi;
  + Nhóm KH ở độ tuổi thanh niên: Đối với những KH từ 18 đến 29 tuổi.



*Hình 45: Biểu đồ thanh thể hiện tổng chi phí chi trả cho BHYT của 3 nhóm độ tuổi trong 2 năm (2020, 2022)*

* Dựa vào biểu đồ, nhóm có thể thấy rõ tác động của đại dịch COVID-19 đối với KH trong nhóm người ở Độ tuổi Lao động (18 - 49) [5] về sự khó khăn trong việc thanh toán BHYT. Trong khi đó, nhóm người Cao tuổi lại có thể tiếp cận các chương trình trợ cấp y tế hoặc các chính sách hỗ trợ khác từ chính phủ [3], nên chi phí BHYT ở nhóm này có sự gia tăng mạnh.
* Đầu năm 2022, nhờ việc gia tăng quyền lợi cho BHYT, cùng với sự tăng cường khả năng tiếp cận bảo hiểm miễn phí hoặc với chi phí thấp tại Hoa Kỳ [3], tổng số tiền BHYT của nhóm Độ tuổi Thanh niên đã tăng lên ~530% (tăng hơn 10 triệu USD trong vòng 2 năm). Nhóm KH ở Độ tuổi Trung niên cũng có sự tăng trưởng nhẹ, khoảng 14% so với năm 2020.
* Tuy nhiên, đối với nhóm người Cao tuổi, mức chi phí BHYT đã được chính phủ điều chỉnh cho phù hợp với tình hình kinh tế hậu khủng hoảng [6], do đó có sự giảm nhẹ từ hơn 17 triệu USD xuống còn khoảng 15,9 triệu USD.
* Dựa trên những dữ liệu được cung cấp ở phần *a* và *b*, nhóm có thể thấy rằng xu hướng tăng trưởng của số lượng KH tham gia BHYT đã và đang diễn ra một cách tích cực. Đặc biệt, với sự gia tăng mạnh mẽ trong nhóm Độ tuổi Thanh niên và Trung niên, cùng với sự điều chỉnh hợp lý của chính phủ đối với nhóm người Cao tuổi, nhóm có thể kỳ vọng sự gia tăng mạnh về số lượng KH tham gia BHYT trong tương lai. Cụ thể, dựa vào những xu hướng này, nhóm có thể đưa ra dự đoán về sự gia tăng số lượng KH tham gia BHYT đến năm 2050 như sau:
  + Dự đoán sự gia tăng số lượng KH tham gia BHYT đến năm 2050:

Để có thể dự đoán số liệu từ năm 2023 đến 2050, nhóm đã áp dụng mô hình Linear Regression. Thông qua mô hình, thuật toán đã tạo ra một dãy số từ 2023 đến 2050 và dự đoán số lượng KH tương ứng cho các năm trong tương lai. Phương pháp này giúp dự đoán xu hướng tăng trưởng của số lượng KH một cách chính xác và hiệu quả.



*Hình 46: Biểu đồ Line Plot thể hiện Số lượng KH tham gia bảo hiểm từ 2005 đến 2022 và dự đoán tăng trưởng đến năm 2050*

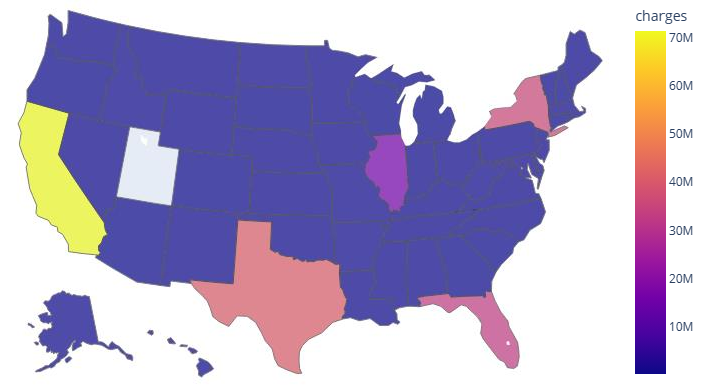
## Mật độ/Khu vực

* + Mật độ dân số và khu vực địa lý có sự ảnh hưởng mật thiết với sự phân bố của dịch vụ y tế. Các khu vực có mật độ dân số cao thường có nhu cầu sử dụng dịch vụ y tế cao hơn, và do đó, chi phí BHYT có thể khác nhau
  + Một số khu vực có đặc điểm dân số đặc biệt, chẳng hạn như một cộng đồng có nhiều người cao tuổi hoặc một khu vực với nhu cầu chăm sóc y tế đặc biệt. Điều này có

thể ảnh hưởng đến chi phí BHYT và yêu cầu phát triển chiến lược quản lý riêng cho khu vực đó.

* + Với mục tiêu phân tích về mật độ và khu vực trong bộ dữ liệu để nhóm có thể hiểu sâu hơn về hành vi sử dụng BHYT cũng như xu hướng sử dụng BHYT của thị trường Mỹ theo từng khu vực và mật độ khác nhau. Chính vì vậy 2 biến **state** (bang) và **region** (khu vực) sẽ là 2 biến được phân tích chính cho những biểu đồ sau.
  + **Mục tiêu**: Các công ty bảo hiểm có thể sử dụng thông tin về mật độ và khu vực để đưa ra chính sách và chiến lược quản lý chi phí hiệu quả. Điều này có thể bao gồm việc điều chỉnh giá cả theo khu vực, cung cấp gói bảo hiểm đặc biệt cho những khu vực có yêu cầu đặc biệt, và phát triển chiến lược tiếp cận đặc thù.

### Phân tích mật độ sử dụng và đóng bảo hiểm theo từng tiểu bang



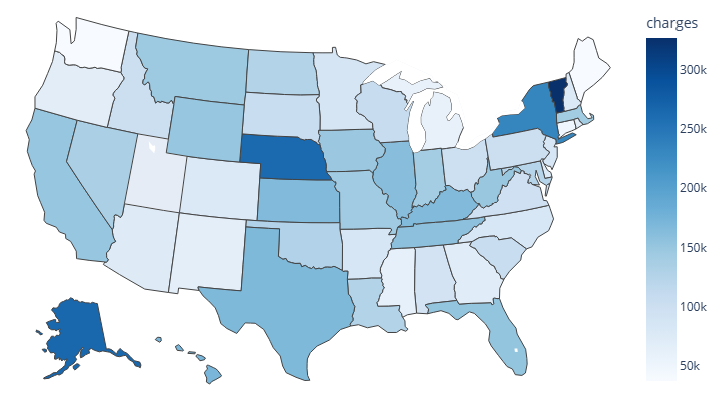
*Hình 47: Biểu đồ Choropleth thể hiện mật độ sử dụng bảo hiểm theo từng bang*

* + Biểu đồ trên cho nhóm thấy được sự chênh lệch giữa tổng mức đóng BHYT của từng bang theo gam màu (ngoài trừ Utah do bộ dữ liệu thiếu thông tin thu thập về tiểu bang này). Trong bộ dữ liệu này thống kê những bang có tổng mức đóng BHYT theo thang đo từ $10M - $70M, nhóm dễ dàng nhận thấy được những tiểu bang có mật độ dân số cao như California, Texas, Florida, New York cũng kéo theo tổng mức đóng BHYT sẽ cao.
  + Và để biết được cụ thể số lượng KH trong từng tiểu bang có mật độ dân số cao trên cách biệt như thế nào với các bang còn lại nhóm thực hiện biểu diễn với biểu đồ sau:



*Hình 48: Biểu đồ thanh ngang thể hiện 10 bang có số lượng KH cao nhất*

* + Từ biểu đồ, nhóm thấy được sự chênh lệch đáng kể giữa những tiểu bang có mật độ dân số cao so với những bang còn lại. Nếu để so sánh có thể lấy bang có số lượng KH cao nhất là California gấp hơn 100 lần so với các bang còn lại ngoài Top 5.



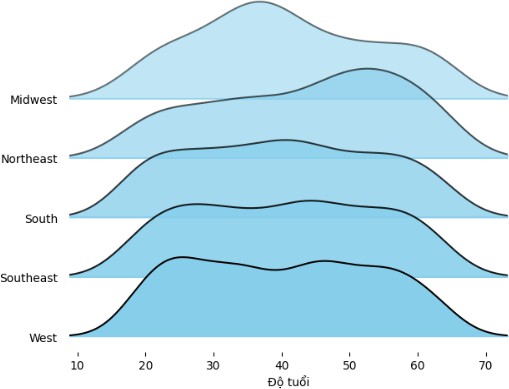
*Hình 49: Biểu đồ Choropleth thể hiện mật độ sử dụng bảo hiểm trung bình theo từng bang*

* + Nhưng điều cần quan tâm là mức đóng BHYT trung bình theo từng tiểu bang từ đó so sánh với với tổng thể, nhóm thấy rằng không phải dân số đông thì mức đóng bảo hiểm trung bình cũng cao theo hoàn toàn. Như hình dưới mức đóng BHYT trung bình cao nhất là những tiểu bang như Vermont, Alaska, New York.
    - Từ đó nhóm thấy được rằng có sự khác biệt giữa mức đóng BHYT theo từng tiểu bang và mật độ dân số chỉ thể hiện một phần trong đó.

Chiến lược tối ưu BHYT:

Đề xuất điều chỉnh mức đóng BHYT theo từng tiểu bang (do mức đóng BHYT không chỉ khác nhau theo gói mà còn khác nhau theo tiểu bang [11]) để không gây mất cân bằng giữa những tiểu bang thưa dân cư.

### Phân tích độ tuổi theo từng khu vực



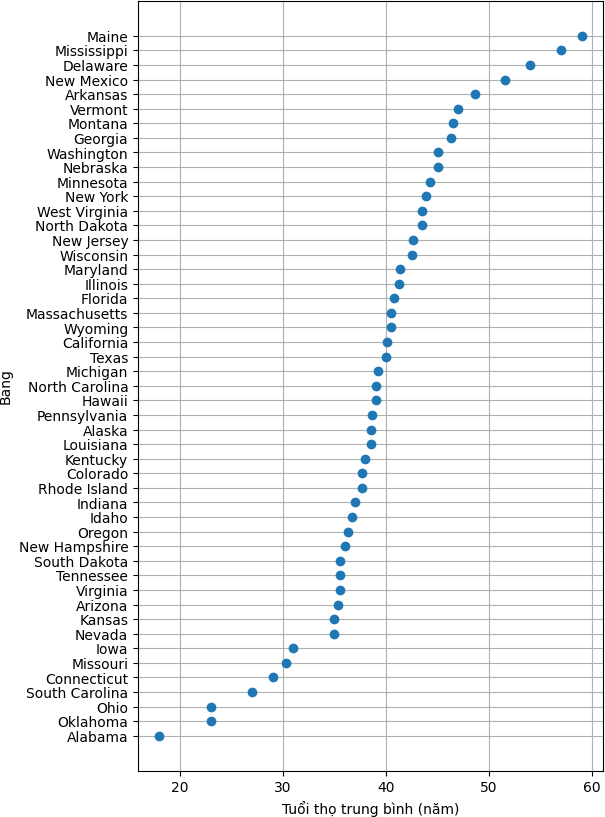
*Hình 50: Biểu đồ đường thể hiện phân phối tuổi theo từng khu vực*

* + Biểu đồ trên cho nhóm 5 phân phối về độ tuổi của 5 khu vực khác nhau. Với biểu đồ này nhóm có thể so sánh một cách tổng quát về dân số theo độ tuổi của từng khu vực. Cụ thể hơn thấy khu vực Midwest có đỉnh nằm trong vùng 30-40 tuổi trong khi khu vực Northeast lại có đỉnh nằm trong vùng 50-60 tuổi thể hiện thấy sự phân hóa trong khung tuổi theo từng khu vực.
  + Nhóm quan sát thấy trong biểu đồ là các khu vực ở miền Bắc Hoa Kỳ có xu hướng có độ tuổi trung bình cao hơn các khu vực ở miền Nam Hoa Kỳ. Điều này có thể là do khí hậu ôn hòa hơn ở miền Bắc Hoa Kỳ, khiến người dân có tuổi thọ cao hơn.

Đề xuất:

Do BHYT là loại hình dịch vụ được cung cấp theo gói nên hiểu được độ tuổi chính yếu của từng khu vực để cung cấp gói hợp lý (trong trường hợp này là gói theo độ tuổi) là điều vô cùng cấp thiết.

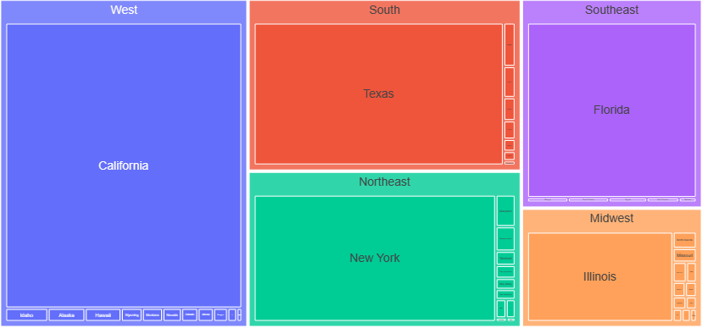
### Phân tích độ tuổi trung bình theo từng tiểu bang



*Hình 51: Biểu đồ phân tán thể hiện độ tuổi trung bình theo từng bang*

* + Biểu đồ thể hiện độ tuổi trung bình theo từng bang ở Hoa Kỳ. Biểu đồ có 49 điểm dữ liệu, một điểm dữ liệu cho mỗi bang. Các điểm dữ liệu được sắp xếp theo thứ tự độ tuổi trung bình, từ thấp nhất đến cao nhất.
  + Từ biểu đồ trên nhóm thấy được độ tuổi trung bình của các bang tập trung trong khoảng 35-45 tuổi. Và những tiểu bang có dân số già (trên 50) chỉ gồm 5 bang là Maine, Mississippi, Delaware, New Mexico và Arkansas.
  + Nhóm nhận thấy rằng các bang ở miền đông bắc Hoa Kỳ có xu hướng có độ tuổi trung bình cao. Điều này có thể là do các bang ở miền Đông Bắc là vùng giàu có nhất của Hoa Kỳ nên mức sống của dân số ở các bang này cũng cao hơn từ đó có tuổi thọ cao hơn các miền khác.

### Phân tích số tiền đóng bảo hiểm theo từng bang trong 5 khu vực ở Hoa Kỳ



*Hình 52: Biểu đồ Treemap thể hiện tổng tiền đóng bảo hiểm theo khu vực và bang*

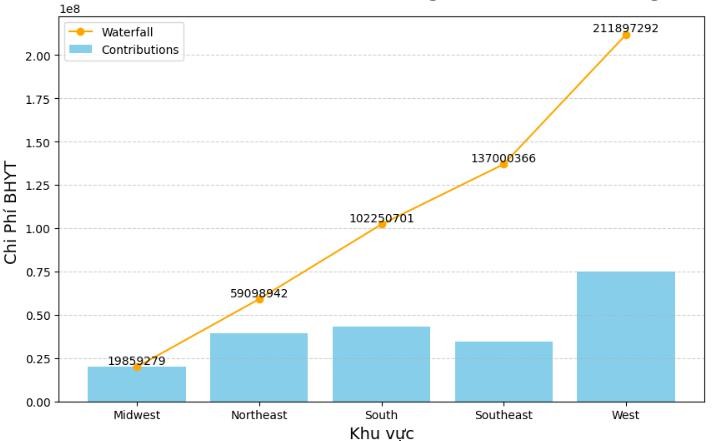
* + Biểu đồ Treemaps tương tác cho thấy tỷ lệ đóng góp BHYT của từng khu vực ở Hoa Kỳ và tiểu bang có số lượng đóng góp cao nhất trong mỗi khu vực cũng được thể hiện trong biểu đồ.



*Hình 53: Biểu đồ Treemap thể hiện tổng tiền đóng bảo hiểm theo khu vực Phía Tây*

* + Đứng đầu bảng trong khu vực là West với California chiếm đa trọng số ($71,5M so với bang đứng sau là Idaho với $607k). Do khu vực Tây Hoa Kỳ có tiểu bang California là một khu vực kinh tế quan trọng với nhiều ngành công nghiệp lớn mang lại lợi nhuận cao cũng như sự đa dạng về thời tiết làm cho người dân nơi đây rất có nhu cầu cũng như khả năng chi trả cho BHYT.
  + Với các khu vực còn lại cũng có trọng số giữa bang đứng đầu và các bang còn lại là rất chênh lệch (So với khu vực Midwest có ít sự chênh lệch nhất thì bang đứng đầu là Illinois cũng chiếm ~86% tổng BHYT toàn Midwest).

### Phân tích Đóng góp chi phí BHYT theo khu vực



*Hình 54: Biểu đồ thác nước thể hiện tổng chi phí BHYT theo khu vực*

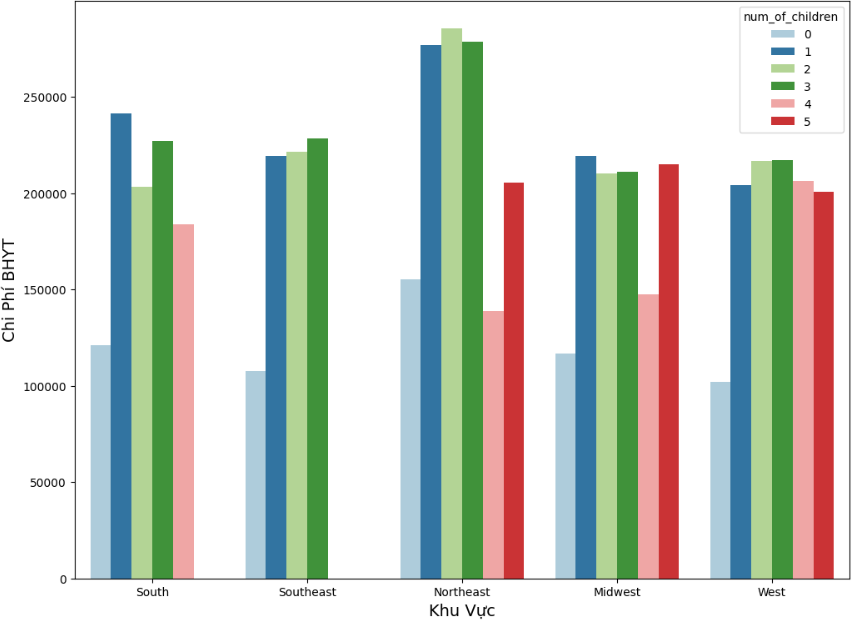
Phân tích Từng khu vực:

* **Khu vực Phía Tây**:
  + Đóng góp cao nhất: Khu vực West có đóng góp lớn nhất, nổi bật hơn so với các khu vực khác.
  + Giả thuyết: Điều này có thể chỉ ra chi phí y tế cao ở khu vực này, có thể xuất phát từ chi phí dịch vụ y tế và chi phí sống cao. Dân số có thu nhập cao cũng có thể đóng góp vào mức đóng góp cao, do họ có khả năng sử dụng và đòi hỏi dịch vụ y tế cao hơn.
* **Các khu vực Phía Đông Bắc, Nam và Đông Nam**:
  + Đóng góp trung bình: Các khu vực này có mức đóng góp tương đối cân bằng.
  + Lý do: Có thể phản ánh sự cân bằng về dân số, chi phí y tế và mô hình sử dụng dịch vụ y tế trong ba khu vực.
* **Khu vực Phía Trung Tây**:
  + Đóng góp thấp nhất: Khu vực Trung Tây có đóng góp thấp nhất.
  + Nguyên nhân có thể bao gồm: Số lượng người dùng dịch vụ y tế thấp. Chi phí y tế thấp hơn so với các khu vực khác. Mô hình bảo hiểm khác nhau. Chi phí sống thấp có thể làm giảm mức đóng góp, ảnh hưởng đến cả chi phí y tế và khả năng thanh toán cho bảo hiểm.

Phân tích Chiến lược:

* Tối ưu hóa chi phí BHYT: Với sự chênh lệch giữa các khu vực, có cơ hội để tối ưu hóa chi phí BHYT thông qua nghiên cứu và áp dụng chiến lược hiệu quả cho từng khu vực.
* Chiến lược Chăm sóc y tế đặc biệt: Chú ý đến khu vực West có thể đòi hỏi các chiến lược đặc biệt để quản lý chi phí và cung cấp chăm sóc y tế hiệu quả cho dân số địa phương.
* Đối mặt với điều kiện đặc thù của từng khu vực: Hiểu rõ điều kiện kinh tế và xã hội của từng khu vực giúp phát triển các chiến lược linh hoạt và phù hợp.
* Phân tích này giúp hiểu rõ nguồn gốc của sự chênh lệch trong chi phí BHYT và cung cấp cơ sở cho việc xây dựng chiến lược quản lý sức khỏe và BHYT có hiệu quả cho mỗi khu vực.

### Tổng chi phí BHYT theo số lượng trẻ em trong gia đình ở mỗi khu vực



*Hình 55: Biểu đồ cột thể hiện số tiền chi trả BHYT theo từng khu vực*

* **Mức chi phí y tế và nhóm tuổi trẻ:**
  + Mức chi phí y tế cao nhất thường xuất hiện ở các khu vực có số lượng trẻ em từ 1-3. Điều này cho thấy nhóm tuổi trẻ này đóng góp lớn vào chi phí y tế, có thể do nhu cầu cao về dịch vụ y tế cho trẻ em.
* **Đặc điểm nổi bật của khu vực Phía Đông Bắc:**
  + Khu vực Phía Đông Bắc nổi bật với mức chi phí y tế cao nhất trong nhóm trẻ em từ 1-3. Điều này có thể xuất phát từ các yếu tố đặc biệt như chất lượng dịch vụ y tế, cơ sở hạ tầng, hoặc yếu tố văn hóa.
* **Thiếu hụt số lượng trẻ em ở khu vực Phía Đông Nam và Phía Nam:**
  + Sự thiếu hụt số lượng trẻ em từ 4 ở Khu vực Southeast và trẻ em từ 5 ở Khu vực South có thể gợi ý đến xu hướng dân số đặc biệt trong các khu vực này. Điều này có thể ảnh hưởng đến cả nhu cầu sử dụng dịch vụ y tế và chi phí tương ứng.
* **Cơ hội tối ưu hóa chi phí BHYT:**
  + Dựa trên insight về mối quan hệ giữa số lượng trẻ em và chi phí y tế, có cơ hội để tối ưu hóa chi phí BHYT. Các biện pháp chăm sóc sức khỏe đặc biệt cho nhóm tuổi trẻ có thể là chìa khóa để giảm chi phí và cung cấp dịch vụ hiệu quả.
  + Qua biểu đồ và nhận định trên, nhóm có thể phát triển chiến lược quản lý chi phí và cung ứng dịch vụ y tế phù hợp với đặc thù của từng khu vực và nhóm tuổi, tạo ra hệ thống BHYT hiệu quả và phản ánh đúng nhu cầu cộng đồng.

### Kết luận

* Mật độ dân số ảnh hưởng đến khả năng tiếp cận và chi phí bảo hiểm ở Hoa Kỳ. Các khu vực có mật độ dân số cao thường có nhiều lựa chọn bảo hiểm hơn và chi phí thấp hơn so với các khu vực có mật độ dân số thấp.
* Các khu vực có mật độ dân số cao thường có nhiều lựa chọn bảo hiểm hơn. Điều này là do các khu vực này có nhiều người mua bảo hiểm tiềm năng, điều này khiến các công ty bảo hiểm có động lực cung cấp nhiều lựa chọn hơn.
* Ngoài ra, các khu vực có mật độ dân số cao thường có nhiều nguồn lực hơn để hỗ trợ những người không có bảo hiểm. Điều này có thể bao gồm các chương trình Medicaid và Medicare, cũng như các chương trình trợ cấp của tiểu bang và địa phương [11].

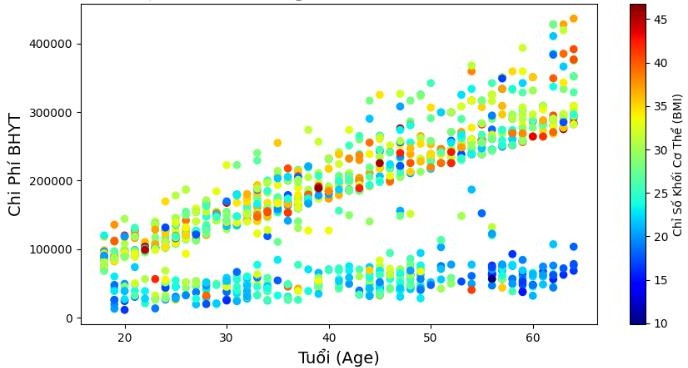
## Sức khỏe & Gia đình

* + Trong phân tích BHYT, **Sức khỏe** là yếu tố mang lại nhiều góc nhìn sâu sắc, đồng thời đóng vai trò quyết định trong nhiều khía cạnh của ngành bảo hiểm và chăm sóc sức khỏe, người có sức khỏe tốt thường có mức độ quan tâm BHYT thấp hơn so với những người có vấn đề về sức khỏe.
  + Bên cạnh đó, các yếu tố khác như: **Tình trạng hôn nhân**, **Việc làm** đều có tác động quan trọng đến nhu cầu và hành vi mua BHYT.

*Những người đã kết hôn và có con thường có lối sống lành mạnh hơn [12], giảm thiểu các hành vi có hại cho sức khỏe, từ đó giảm nguy cơ mắc bệnh và nhu cầu sử dụng BHYT. Hay, việc làm ổn định cung cấp thu nhập, tạo điều kiện thuận lợi cho việc mua bảo hiểm [13].*

* + **Mục tiêu**: Xác định nhóm đối tượng mục tiêu cụ thể, từ đó phát triển các gói bảo hiểm phù hợp với nhu cầu và khả năng tài chính của từng nhóm.

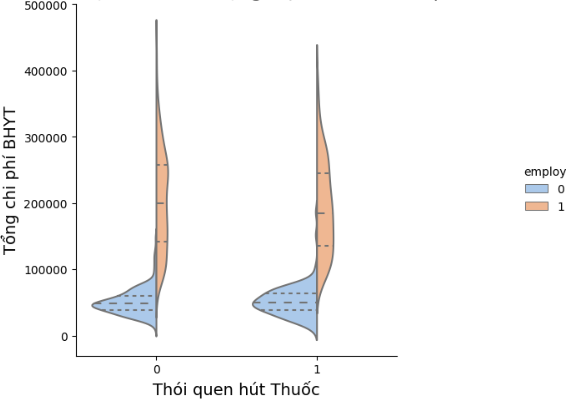
### Tương quan giữa Tuổi, Chỉ số khối cơ thể và Chi phí BHYT



*Hình 56: Biểu đồ phân tán thể hiện mối quan hệ giữa độ tuổi, tổng chi phí BHYT và chỉ số khối cơ thể (BMI)*

* + **Nhóm người có chi phí BHYT cao:**
    - Người có tuổi cao và chỉ số BMI ở mức 25-45 có chi phí BHYT cao. Điều này có thể chỉ ra nhu cầu y tế tăng cao trong nhóm này, có thể liên quan đến vấn đề sức khỏe liên quan đến tuổi tác và BMI cao.
  + **Nhóm người có chi phí BHYT thấp:**
    - Người có tuổi cao và chỉ số BMI dưới 20 có chi phí BHYT thấp. Điều này có thể phản ánh sự giảm rủi ro hoặc các yếu tố tích cực về sức khỏe trong nhóm này.
  + **Chênh lệch giữa các nhóm:**
    - Chênh lệch chi phí BHYT giữa các nhóm có thể xuất phát từ sự biến động lớn trong nhu cầu y tế và sức khỏe của các đối tượng khác nhau.
  + **Chiến lược quản lý:**
    - Đối tượng cần chú ý và quản lý chi phí có thể là nhóm người cao tuổi với BMI trong khoảng 25-45. Cần xem xét chiến lược quản lý và chăm sóc y tế đặc biệt cho nhóm này.

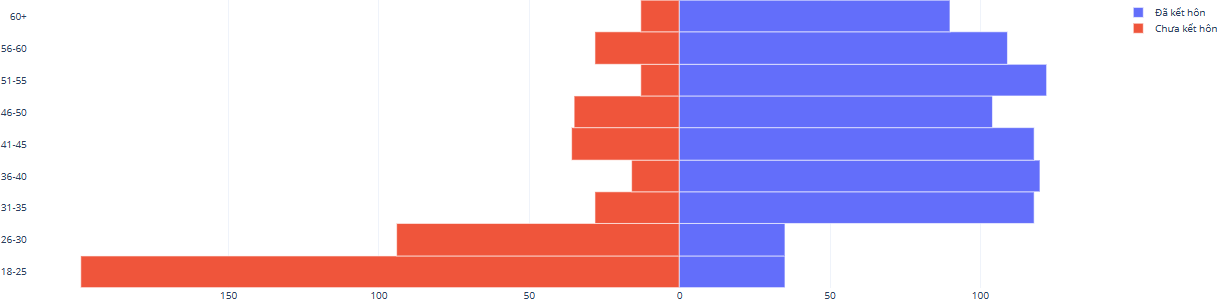
### Sự phân phối của Tổng chi phí BHYT dựa trên Tình trạng việc làm và Thói quen hút thuốc của KH



*Hình 57: Biểu đồ Violin thể hiện Phân phối của Tổng chi phí BHYT dựa trên Tình trạng việc làm và Thói quen hút thuốc của KH*

* Từ biểu đồ, nhóm xác định được có sự khác biệt rõ rệt về chi phí BHYT giữa nhóm KH có hút thuốc và nhóm không hút thuốc. Nhóm hút thuốc có chi phí BHYT cao hơn nhiều so với nhóm không hút thuốc, cho dù có việc làm hay không.
* Ngoài ra, ở biểu đồ này cũng làm rõ hơn cho **Giả thuyết số 8** được đặt ra ở **Chương V**. Để giải thích cho việc: nhóm KH Có thói quen hút thuốc thường có mức quan tâm thấp đối với BHYT, là do chính sách “Tobacco Rating” của đạo luật ACA, cho phép các công ty bảo hiểm tính phí cao hơn 50% (hoặc gấp 1.5 lần) đối với nhóm KH này [14].

### Phân bố Dân số theo Nhóm tuổi và Tình trạng hôn nhân



*Hình 58: Biểu đồ cột đối xứng thể hiện tháp dân số*

* Tiếp tục phân tích biểu đồ, ta có thể thấy một số điểm đáng chú ý sau:
  + Nhóm tuổi từ 26-35: Đây là giai đoạn mà nhiều người bắt đầu xây dựng gia đình và sự nghiệp. Từ biểu đồ, nhóm thấy rằng số lượng người đã kết hôn

trong nhóm tuổi này tăng đáng kể so với nhóm tuổi 18-25. Điều này cho thấy xu hướng kết hôn ở độ tuổi này khá cao.

* + Nhóm tuổi từ 41 trở đi: Đối với nhóm tuổi này, số lượng người đã kết hôn và chưa kết hôn có sự chênh lệch rõ ràng. Số lượng người chưa kết hôn giảm dần theo thời gian, trong khi số lượng người đã kết hôn vẫn duy trì ổn định. Điều này cho thấy rằng sau tuổi 40, tỷ lệ kết hôn giảm đáng kể.
  + Nhóm tuổi từ 51 trở đi: Số lượng người đã kết hôn trong nhóm tuổi này chiếm đa số. Điều này có thể cho thấy rằng hầu hết mọi người thường đã kết hôn trước tuổi 50.

Nhìn chung, biểu đồ thể hiện rõ ràng xu hướng kết hôn theo độ tuổi và giúp ta hiểu hơn về cấu trúc dân số theo tình trạng hôn nhân. Từ đó nắm bắt được xu hướng thị trường và tâm lý người tiêu dùng có thể thay đổi theo thời gian, hỗ trợ doanh nghiệp cung cấp dịch vụ bảo hiểm đưa ra được sản phẩm phù hợp với mọi phân khúc KH.

### Kết luận

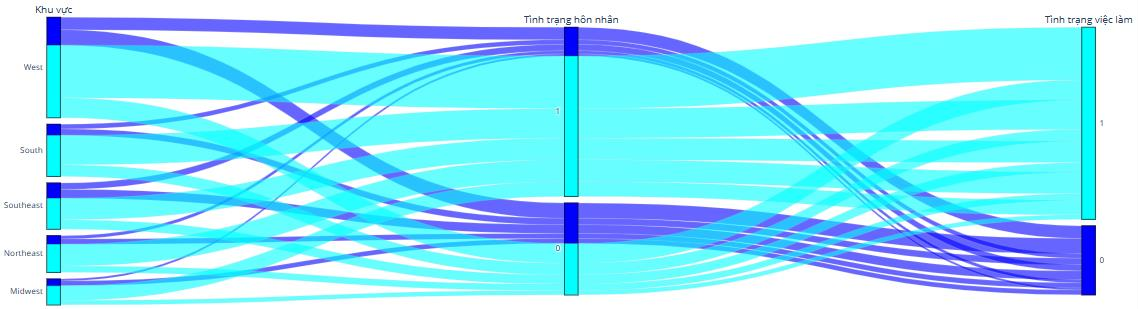
Dựa trên phân tích chi tiết từ biểu đồ của các mối quan hệ giữa sức khỏe, tuổi tác, tình trạng hôn nhân và thói quen hút thuốc của KH, nhóm có thể đưa ra một số kết luận và chiến lược quản lý BHYT như sau:

* + **Quản lý chi phí và chăm sóc y tế đặc biệt cho nhóm tuổi 25-45 và BMI 25- 45:**
    - Nhóm người có tuổi từ 25-45 và chỉ số BMI từ 25-45 đóng góp nhiều vào chi phí BHYT. Cần xem xét và phát triển chính sách quản lý chi phí và các chương trình chăm sóc y tế đặc biệt để đáp ứng nhu cầu của nhóm này.
  + **Tăng cường chiến lược thúc đẩy bảo hiểm và quản lý chi phí cho nhóm KH có thói quen hút thuốc:**
    - Nhóm KH có thói quen hút thuốc hiện đang có chi phí BHYT thấp hơn so với nhóm không hút thuốc. Có thể cần tăng cường chiến lược thúc đẩy mức độ quan tâm đối với bảo hiểm và quản lý chi phí cho nhóm KH này, như việc cân nhắc giảm phí thuốc là từ “Tobacco Rating” hoặc các chính sách tương tự khác.
  + **Tập trung nhóm đối tượng:**
    - Tăng cường dịch vụ tư vấn và hỗ trợ cho KH, giúp họ hiểu rõ hơn về lợi ích của việc sử dụng BHYT. Thực hiện ở nhóm KH có tuổi từ 26-35 và 36 trở đi, đặc biệt là những người đã kết hôn, vì họ có nhu cầu về BHYT cao hơn.
    - Đối với chính phủ từng bang, cần đưa ra các chương trình nhân đạo: ưu đãi hoặc tạo cơ hội dành riêng cho đối tượng KH kết hôn sớm ở độ tuổi dưới 26 trở đi để thu hút họ sử dụng BHYT.

## Phân tích tổng hợp

Dựa trên một số nghiên cứu khoa học gần đây về nhu cầu BHYT tại Hoa Kỳ, nhóm nhận thấy rằng tình trạng việc làm, giới tính, khu vực cư trú và tình trạng hôn nhân đều có ảnh hưởng mạnh mẽ đến nhu cầu BHYT của người dân, cụ thể:

* + Theo Viện Quốc gia về An toàn và Sức khỏe Nghề nghiệp (NIOSH) đã phát hiện ra rằng việc tiếp cận chăm sóc sức khỏe và các nhu cầu liên quan đến sức khỏe tại Hoa Kỳ là khác nhau dựa trên tình trạng việc làm [7].
  + Khu vực cư trú có ảnh hưởng mạnh mẽ đến chi tiêu y tế và việc phân bổ tài chính chăm sóc sức khỏe [8].
  + Tình trạng hôn nhân có ảnh hưởng sâu sắc đến việc sử dụng dịch vụ chăm sóc sức khỏe tại Hoa Kỳ. Cụ thể, hôn nhân là một trong những yếu tố giúp cho người dân nơi đây hưởng lợi từ chương trình Medicare [9].
* Việc tổng hợp và phân tích các yếu tố từ góc nhìn trên không chỉ giúp hiểu rõ hơn về cách phân bổ nhu cầu BHYT, mà còn cho phép mở ra nhiều phương hướng nghiên cứu mới nhằm cải thiện hệ thống BHYT. Điều này sẽ giúp hệ thống trở nên công bằng và hiệu quả hơn.



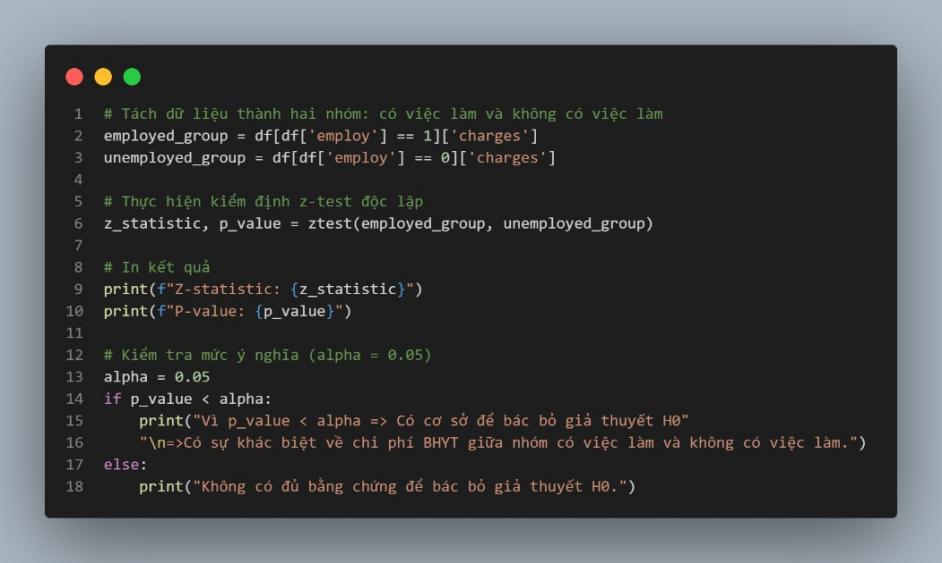
*Hình 59: Biểu đồ Parallel Categories Plot thể hiện tỷ lệ KH vẫn đang đi làm khi tham gia BHYT theo các thuộc tính phân loại (hôn nhân và khu vực)*

Từ thông tin được cung cấp ở biểu đồ trên, nhóm nhận thấy rằng:

* + **Hôn nhân**: KH đã kết hôn chiếm tỷ lệ lớn nhất (64,8%), trong khi KH chưa kết hôn chiếm tỷ lệ thấp hơn (35,2%). Điều này có thể do KH đã kết hôn thường có nhiều trách nhiệm gia đình hơn và cần BHYT hơn.
  + **Khu vực**: KH ở khu vực phía Tây chiếm tỷ lệ lớn nhất (38,5%), trong khi KH ở khu vực phía Trung Tây chiếm tỷ lệ thấp nhất (9,9%). Điều này có thể phản ánh sự khác biệt về cơ hội việc làm và mức sống giữa các khu vực.
  + **Tình trạng việc làm**: Tỷ lệ KH đang đi làm khi tham gia BHYT là 73,6%, trong khi tỷ lệ KH không đi làm là 26,4%. Điều này cho thấy hầu hết KH tham gia BHYT đều đang có việc làm.
  + **Dự đoán sơ bộ**: Dựa trên xu hướng hiện tại, nhóm có thể dự đoán rằng tỷ lệ KH vẫn đang đi làm khi tham gia BHYT sẽ tiếp tục tăng trong tương lai, đặc biệt là

trong nhóm KH đã kết hôn và KH ở khu vực phía Tây. Tuy nhiên, cần có thêm nghiên cứu để xác nhận và hiểu rõ hơn về những xu hướng này.

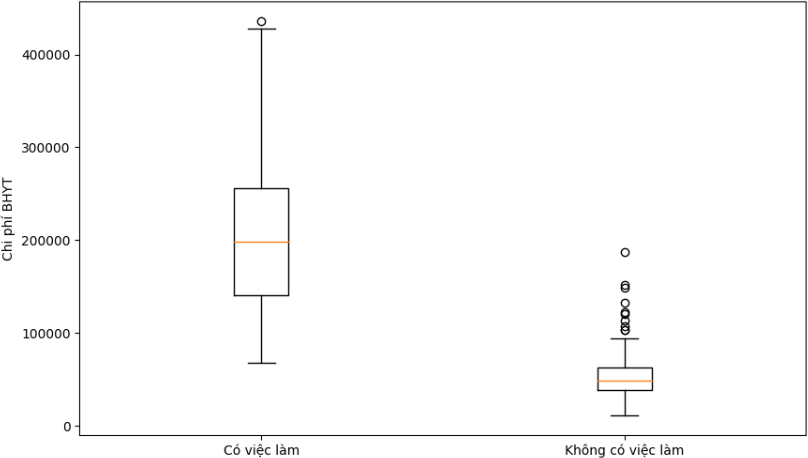
* **Kiểm định giả thuyết:** Có sự khác biệt về chi phí BHYT giữa nhóm có việc làm và không có việc làm (Với độ tin cậy cho bài toán là 95%).
  + μ1: Trung bình chi phí BHYT của nhóm có việc làm.
  + μ2: Trung bình chi phí BHYT của nhóm không có việc làm.
  + H0: μ1 = μ2 (Không có sự khác biệt về chi phí BHYT giữa hai nhóm)
  + H1: μ1 ≠ μ2 (Có sự khác biệt về chi phí BHYT giữa hai nhóm)
* Phương pháp: Bằng cách sử dụng **z-test** để thực hiện để kiểm tra giả thuyết H0. Nếu giá trị p-value < 𝛼, nhóm đã có cơ sở để bác bỏ giả thuyết H0 và chấp nhận giả thuyết H1, tức là có sự khác biệt đáng kể về chi phí BHYT giữa nhóm có việc làm và nhóm không có việc làm.



* Kết quả:



* Biểu diễn trực quan kết quả:



*Hình 60: Biểu đồ Box Plot thể hiện phân phối chi phí BHYT giữa nhóm KH có việc làm và không có việc làm*

Việc sử dụng Box Plot để biểu diễn trực quan cho kiểm định trên vì nó có thể cung cấp một số thông tin sau để chứng minh sự khác biệt về tổng chi phí BHYT của hai nhóm KH:

+ **Vị trí của Median**: Cách xa nhau, có thể nói rằng có sự khác biệt về chi phí BHYT giữa hai nhóm KH.

+ **Độ rộng của IQR**: Là khác nhau, chỉ ra sự khác biệt về độ phân tán của chi phí BHYT giữa hai nhóm.

+ **Độ dài của Whiskers**: Là khác nhau, cho biết sự khác biệt về phạm vi của chi phí BHYT giữa hai nhóm.

+ **Số lượng & Vị trí Outliers**: Là khác nhau, có thể thấy được sự khác biệt về các giá trị bất thường của chi phí BHYT giữa hai nhóm.

# CHƯƠNG VII. KẾT LUẬN

Sau khi thực hiện bài nghiên cứu: “**Phân tích & Trực quan hóa dữ liệu trong ngành Bảo hiểm Y tế tại Hoa Kỳ**”, nhóm đã tiến hành phân tích và trực quan hóa dữ liệu để đưa ra một cái nhìn tổng quan, nhằm đánh giá rủi ro trong lĩnh vực bảo hiểm sức khỏe và phân tích sự ảnh hưởng của các yếu tố trong bộ dữ liệu đối với chi phí bảo hiểm.

**Đặc điểm nổi bật trong nghiên cứu**: Nhóm đã áp dụng đa dạng các phương pháp phân tích dữ liệu và nguyên tắc thiết kế biểu đồ sao cho hiệu quả nhất. Đồng thời, sự linh hoạt trong việc sử dụng nhiều loại biểu đồ khác nhau đã giúp tăng cường khả năng truyền đạt thông tin và làm cho dữ liệu dễ tiếp nhận hơn. Không những thế, đây còn là cơ sở để hỗ trợ quá trình ra quyết định trong lĩnh vực bảo hiểm. Ngoài ra, sự đa dạng trong việc sử dụng màu sắc cũng được nhóm chú trọng, thông qua việc áp dụng màu sắc theo các thang đo có chủ đích đã giúp người dùng phân biệt các nhóm chủ đề một cách dễ dàng hơn.

**Đánh giá chất lượng biểu đồ:**

* **Rõ ràng, dễ hiểu**: Biểu đồ nên có tiêu đề rõ ràng, các nhãn trục phù hợp, và có chú thích để người xem có thể hiểu được biểu đồ mà không cần thông tin bổ sung.
* **Không gây nhầm lẫn**: Màu sắc và hình dạng được sử dụng một cách cẩn thận để tránh gây nhầm lẫn.
* **Không thừa thông tin & Tính liên kết**: Biểu đồ không chứa quá nhiều thông tin đến mức gây rối, mỗi biểu đồ chỉ tập trung vào một hoặc một số ít thông điệp chính. Ngoài ra, tất cả các biểu đồ trong phần này đều được liên kết với nhau chặt chẽ về mặt nội dung.
* **Đúng tỷ lệ**: Trục của biểu đồ được vẽ đúng tỷ lệ để không gây hiểu lầm về dữ liệu.

Tuy nhiên, bài nghiên cứu vẫn còn một số hạn chế cần được cải thiện. Đầu tiên là khả năng tương tác trong các biểu đồ dữ liệu, tiếp đến là tối ưu hóa hiệu suất của các biểu đồ. Do nhóm thực hiện là sinh viên và chưa được trang bị đầy đủ kiến thức trong lĩnh vực bảo hiểm, khi thực hiện đề tài, nhóm chúng em đã phải dựa vào nhiều tài liệu tham khảo từ bên ngoài để đưa ra được những insights và giải pháp cần thiết cho vấn đề, do đó, chúng em không thể chắc chắn hoàn toàn về tính chính xác của những đề xuất được đưa ra trong các bài nghiên cứu đó. Để nâng cao chất lượng của bài nghiên cứu, nhóm cần phân tích sâu hơn về các mô hình dự đoán và đánh giá ảnh hưởng của các yếu tố khác nhau đối với chi phí BHYT.

# DANH MỤC BẢNG BIỂU

[Bảng 1: Bảng giải thích thuộc tính của bộ dữ liệu gốc 6](#_bookmark10)

[Bảng 2: Bảng giải thích ý nghĩa thêm thuộc tính cho bộ dữ liệu 14](#_bookmark26)

[Bảng 3: Bảng thể hiện chi phí bảo hiểm trung bình hằng năm theo bang ở Hoa Kỳ [12] 16](#_bookmark29)

# DANH MỤC HÌNH ẢNH

[Hình 1: Mô tả thông tin cơ bản của bộ dữ liệu ban đầu 6](#_bookmark13)

[Hình 2: Quan sát giá còn thiếu và giá trị bất thường của bộ dữ liệu ban đầu 7](#_bookmark15)

[Hình 3: Bảng thống kê tổng quát của bộ dữ liệu ban đầu 7](#_bookmark16)

[Hình 4: Các biểu đồ phân phối cho cột dữ liệu số của dữ liệu gốc 8](#_bookmark18)

[Hình 5: Các biểu đồ tròn cho cột dữ liệu phân loại của dữ liệu gốc 9](#_bookmark19)

[Hình 6: Ma trận tương quan của biến số có trong bộ dữ liệu gốc 10](#_bookmark21)

[Hình 7: Biểu đồ phân tán giữa giá trị thực tế và dự đoán của bộ dữ liệu ban đầu 17](#_bookmark31)

[Hình 8: Biểu đồ phân tán giữa giá trị thực tế và dự đoán của bộ dữ liệu mới 17](#_bookmark32)

[Hình 9: Kết quả kiểm tra Cross-Validation trên bộ dữ liệu mới 18](#_bookmark34)

[Hình 10: Kết quả kiểm tra Learning Curve trên bộ dữ liệu mới 19](#_bookmark35)

[Hình 11: Mô tả thông tin cơ bản của bộ dữ liệu sau thay đổi 20](#_bookmark39)

[Hình 12: Quan sát giá còn thiếu và giá trị bất thường của bộ dữ liệu sau thay đổi 21](#_bookmark41)

[Hình 13: Quan sát số giá trị trống có trong bộ dữ liệu sau thay đổi 21](#_bookmark42)

[Hình 14: Quan sát số giá trị bất thường có trong bộ dữ liệu sau thay đổi 21](#_bookmark43)

[Hình 15: Bảng thống kê tổng quát của bộ dữ liệu sau thay đổi 22](#_bookmark44)

[Hình 16: Biểu đồ phân phối cho cột dữ liệu số của bộ dữ liệu mới 23](#_bookmark46)

[Hình 17: Biểu đồ tròn cho cột dữ liệu phân loại của bộ dữ liệu mới 24](#_bookmark47)

[Hình 18: Biểu đồ tròn thể hiện tỷ lệ KH theo tiểu bang 25](#_bookmark48)

[Hình 19: Biểu đồ phân phối cho cột dữ liệu thời gian của bộ dữ liệu mới 25](#_bookmark49)

[Hình 20: Ma trận tương quan của biến số có trong bộ dữ liệu sau thay đổi 26](#_bookmark51)

[Hình 21: Đồ thị biểu diễn % Phương sai tích lũy theo số Features (k) 27](#_bookmark52)

[Hình 22: Phần trăm Phương sai tích lũy theo số Features (k) 28](#_bookmark53)

[Hình 23: So sánh hai Biểu đồ Histogram thể hiện số tiền chi trả bảo hiểm theo số người](#_bookmark54) [của mỗi giới tính 30](#_bookmark54)

[Hình 24: Biểu đồ thanh thể hiện giá trị trung bình lần làm mới bảo hiểm của hai nhóm . 32](#_bookmark55)

[Hình 25: Biểu đồ thanh thể hiện độ lệch chuẩn của hai nhóm 33](#_bookmark56)

[Hình 26: Biểu đồ phân phối xác suất số lần làm mới bảo hiểm của hai biến 34](#_bookmark57)

[Hình 27: Biểu đồ hộp quan sát giá trị Outliers của biến bmi và charges 36](#_bookmark59)

[Hình 28: Biểu đồ hộp thể hiện Outliers cho bmi và charges sau khi áp dụng 3-sigma 37](#_bookmark60)

[Hình 29: Biểu đồ hộp thể hiện Outliers cho bmi khi sử dụng IQR 38](#_bookmark61)

[Hình 30: Biểu đồ hộp thể hiện Outliers cho bmi khi sử dụng thêm một lần phương pháp](#_bookmark62) [IQR 38](#_bookmark62)

[Hình 31: Biểu đồ cột thể hiện số lần làm mới bảo hiểm theo nhóm KH cao tuổi (hơn 50](#_bookmark66) [tuổi) 41](#_bookmark66)

[Hình 32: Biểu đô cột so sánh trung bình lần làm mới bảo hiểm theo các nhóm khu vực. 42](#_bookmark68) [Hình 33: Biểu đồ hộp thể hiện phân bố và phương sai theo từng khu vực 43](#_bookmark69)

[Hình 34: Biểu đồ phân tán thể hiện số tiền chi trả bảo hiểm theo độ tuổi và lối sống hút](#_bookmark71) [thuốc của mỗi người 47](#_bookmark71)

[Hình 35: Hình 33 kết hợp đường tuyến tính phân chia vùng diện tích 48](#_bookmark72)

[Hình 36: Ma trận tương quan giữa giới tính và lối sống hút thuốc 50](#_bookmark73)

[Hình 37: Biểu đồ tròn thể hiện tỷ lệ người theo lối sống hút thuốc và giới tính 51](#_bookmark74)

[Hình 38: Biểu đồ cột thể hiện số tiền chi trả bảo hiểm của nhóm độ tuổi và lối sống hút](#_bookmark75) [thuốc 53](#_bookmark75)

[Hình 39: Biểu đồ phân phối xác suất tổng chi phí trả cho bảo hiểm của hai biến 56](#_bookmark76)

[Hình 40: Biểu đồ đường thể hiện phân phối xác suất theo số lần làm mới BHYT của hai](#_bookmark77) [nhóm 59](#_bookmark77)

[Hình 41: Biểu đồ Line Plot thể hiện tổng tiền BHYT mà KH đã chi trả từ năm 2005 - 2022](#_bookmark80)

[. 61](#_bookmark80)

[Hình 42: Biểu đồ thanh thể hiện 5 năm có tổng tiền chi trả cho BHYT thấp nhất 62](#_bookmark81)

[Hình 43: Biểu đồ thanh thể hiện 5 năm có tổng tiền chi trả cho BHYT cao nhất 62](#_bookmark82)

[Hình 44: Biểu đồ Line Plot thể hiện sự gia tăng về số lượng KH đăng ký BHYT theo từng](#_bookmark83) [tháng trong 3 năm (2020, 2021, 2022) 63](#_bookmark83)

[Hình 45: Biểu đồ thanh thể hiện tổng chi phí chi trả cho BHYT của 3 nhóm độ tuổi trong](#_bookmark84) [2 năm (2020, 2022) 64](#_bookmark84)

[Hình 46: Biểu đồ Line Plot thể hiện Số lượng KH tham gia bảo hiểm từ 2005 đến 2022 và](#_bookmark85) [dự đoán tăng trưởng đến năm 2050 65](#_bookmark85)

[Hình 47: Biểu đồ Choropleth thể hiện mật độ sử dụng bảo hiểm theo từng bang 66](#_bookmark88)

[Hình 48: Biểu đồ thanh ngang thể hiện 10 bang có số lượng KH cao nhất 67](#_bookmark89)

[Hình 49: Biểu đồ Choropleth thể hiện mật độ sử dụng bảo hiểm trung bình theo từng bang](#_bookmark90)

[. 67](#_bookmark90)

[Hình 50: Biểu đồ đường thể hiện phân phối tuổi theo từng khu vực 68](#_bookmark92)

[Hình 51: Biểu đồ phân tán thể hiện độ tuổi trung bình theo từng bang 69](#_bookmark94)

[Hình 52: Biểu đồ Treemap thể hiện tổng tiền đóng bảo hiểm theo khu vực và bang 70](#_bookmark96)

[Hình 53: Biểu đồ Treemap thể hiện tổng tiền đóng bảo hiểm theo khu vực Phía Tây 70](#_bookmark97)

[Hình 54: Biểu đồ thác nước thể hiện tổng chi phí BHYT theo khu vực 71](#_bookmark99)

[Hình 55: Biểu đồ cột thể hiện số tiền chi trả BHYT theo từng khu vực 72](#_bookmark101)

[Hình 56: Biểu đồ phân tán thể hiện mối quan hệ giữa độ tuổi, tổng chi phí BHYT và chỉ số](#_bookmark105) [khối cơ thể (BMI) 74](#_bookmark105)

[Hình 57: Biểu đồ Violin thể hiện Phân phối của Tổng chi phí BHYT dựa trên Tình trạng](#_bookmark107) [việc làm và Thói quen hút thuốc của KH 75](#_bookmark107)

[Hình 58: Biểu đồ cột đối xứng thể hiện tháp dân số 75](#_bookmark109)

[Hình 59: Biểu đồ Parallel Categories Plot thể hiện tỷ lệ KH vẫn đang đi làm khi tham gia](#_bookmark112) [BHYT theo các thuộc tính phân loại (hôn nhân và khu vực) 77](#_bookmark112)

[Hình 60: Biểu đồ Box Plot thể hiện phân phối chi phí BHYT giữa nhóm KH có việc làm](#_bookmark113) [và không có việc làm 79](#_bookmark113)

# PHỤ LỤC

## Link nguồn

Link Github: *[Neyung/DV: PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU TRONG](https://github.com/Neyung/DV) [NGÀNH BẢO HIỂM Y TẾ TẠI HOA KỲ (github.com)](https://github.com/Neyung/DV)*

Mã Nguồn:

* + **Tiền xử lý**: *[1-Preprocessing.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1rLvOZQ1JPHyaDAHJ0hqmuG4JBgYaXZYK)*
  + **Kiểm định**: *[2-Hypothesis\_Testing.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1xu-TmsNxA7V_0IkEciAexE0TGljkNm5G)*
  + **Biểu diễn trực quan**: *[3-Data\_Visualization.ipynb - Colaboratory (google.com)](https://colab.research.google.com/drive/1f4R2uOlaoF2YoB9GOKs70BOuxYu8QlCu)*

## Bảng phân công

|  |  |  |
| --- | --- | --- |
| **THÀNH VIÊN** | **PHÂN CÔNG** | **ĐÁNH GIÁ** |
| Nguyễn Ngọc Nam | Thay đổi ý nghĩa bộ dữ liệu, Tiền xử lý, Biểu diễn trực quan, Nhận xét biểu đồ (Khu vực, Sức khỏe và Gia đình) | 100% |
| Trần Phạm Hải Nam | PCA, Xây dưng kiểm định, Phân tích kiểm định, Tổng hợp nội dung (Văn bản), Viết báo cáo tổng hợp | 100% |
| Lý Minh Nguyên | PCA, Xây dưng kiểm định, Phân tích kiểm định, Viết nội dung báo cáo | 100% |
| Võ Minh Nguyên | Tiền xử lý, Kiểm tra hiệu quả thay đổi biến, Biểu diễn trực quan, Nhận xét biểu đồ (Time Series & Phân tích tổng hợp), Sửa nội dung báo cáo | 100% |
| Phan Đình Nhân | Kiểm tra Overfitting, Biểu diễn trực quan, Nhận xét biểu đồ (Mật độ & Khu vực), Tổng hợp nội dung (Mã nguồn) | 100% |

# TÀI LIỆU THAM KHẢO

[0] Giáo trình Biểu diễn Trực quan Dữ liệu, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

[0] Giáo trình Máy học, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

[0] Giáo trình Lập trình Phân tích Dữ liệu, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.

1. Giáo trình Khai phá Dữ liệu, TS. Nguyễn An Tế, Khoa Công nghệ Thông tin trong Kinh doanh, Trường Công nghệ và Thiết kế - Đại học UEH, 2023.
2. [Is Health Insurance Mandatory? (investopedia.com)](https://www.investopedia.com/is-health-insurance-mandatory-4773106)
3. [U.S.: single coverage health premiums 2000-2021 | Statista](https://www.statista.com/statistics/654617/health-premiums-for-single-employee-coverage-us/)
4. [Policies To Improve Health Insurance Coverage as America Recovers From COVID-](https://www.americanprogress.org/article/policies-improve-health-insurance-coverage-america-recovers-covid-19/) [19 - Center for American Progress](https://www.americanprogress.org/article/policies-improve-health-insurance-coverage-america-recovers-covid-19/)
5. [What is a Special Enrollment Period? - HealthSherpa Blog](https://blog.healthsherpa.com/what-is-special-enrollment-period-sep#%3A~%3Atext%3DIf%20you%20miss%20your%20SEP%2C%20you%E2%80%99ll%20have%20to%2Cand%20runs%20through%20January%2015th%2C%20in%20most%20states)
6. [Study shows how health insurance coverage changed during the COVID-19 pandemic](https://www.news-medical.net/news/20210909/Study-shows-how-health-insurance-coverage-changed-during-the-COVID-19-pandemic.aspx) [(news-medical.net)](https://www.news-medical.net/news/20210909/Study-shows-how-health-insurance-coverage-changed-during-the-COVID-19-pandemic.aspx)
7. [The Inflation Reduction Act will cut health care costs for some patients. But we need](https://www.aamc.org/news/inflation-reduction-act-will-cut-health-care-costs-some-patients-we-need-do-more) [to do more | AAMC](https://www.aamc.org/news/inflation-reduction-act-will-cut-health-care-costs-some-patients-we-need-do-more)
8. [NIOSH Study Examines Relationship between Employment Status, Healthcare Access,](https://www.cdc.gov/niosh/updates/upd-11-18-21.html) [and Health Outcomes (cdc.gov)](https://www.cdc.gov/niosh/updates/upd-11-18-21.html)
9. [Ethics and geographical equity in health care | Journal of Medical Ethics (bmj.com)](https://jme.bmj.com/content/27/4/256)
10. [(43) The impact of marital status on health care utilization among Medicare](https://www.academia.edu/81807324/The_impact_of_marital_status_on_health_care_utilization_among_Medicare_beneficiaries) [beneficiaries | Kiran Raj Pandey - Academia.edu](https://www.academia.edu/81807324/The_impact_of_marital_status_on_health_care_utilization_among_Medicare_beneficiaries)
11. [Health Insurance Coverage of the Total Population | KFF](https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D)
12. [Y tế ở Mỹ: Hệ thống y tế, bảo hiểm & chi phí cần biết (newoceanimmi.com)](https://newoceanimmi.com/y-te-o-my/)
13. [Family Status and Health Behaviors: Social Control as a Dimension of Social](https://www.jstor.org/stable/2136848) [Integration on JSTOR](https://www.jstor.org/stable/2136848)
14. [Socioeconomic Factors | CDC](https://www.cdc.gov/dhdsp/health_equity/socioeconomic.htm)
15. [What You Need to Know About Smoking and Health Insurance (healthmarkets.com)](https://www.healthmarkets.com/resources/health-insurance/smoking-and-health-insurance/)