**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM**

Logo, company name

Description automatically generated**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN: NHẬP MÔN LẬP TRÌNH PYTHON CHO PHÂN TÍCH**

**PHÂN TÍCH CÁC YẾU TỐ NGUY CƠ DẪN ĐẾN BỆNH TIM MẠCH**

**GVHD: Quách Đình Hoàng**

**Nhóm 22:**

|  |  |
| --- | --- |
| 20133108 | Trần Tuyên Trường |
| 20133057 | Lê Vũ Thế Khôi |
| 20133075 | Võ Hoàng Nguyên |
| 20133008 | Hoàng Uyên |

***TP. Hồ Chí Minh, tháng 06 năm 2022***

**NHẬN XÉT**

**Giáo viên hướng dẫn**

(Ký và ghi rõ họ tên)

**Quách Đình Hoàng**

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc106214440)

[NỘI DUNG 2](#_Toc106214441)

[1. Tóm tắt 2](#_Toc106214442)

[2. Giới thiệu 2](#_Toc106214443)

[3. Dữ liệu 3](#_Toc106214444)

[4. Trực quan hóa dữ liệu 8](#_Toc106214445)

[5. Mô hình hóa dữ liệu 14](#_Toc106214446)

[a. Thống kê mô tả 14](#_Toc106214447)

[b. Thống kê suy diễn 14](#_Toc106214448)

[6. Thực nghiệm, kết quả, và thảo luận 15](#_Toc106214449)

[7. Kết luận 18](#_Toc106214450)

[8. Phụ lục 18](#_Toc106214451)

[9. Đóng góp 19](#_Toc106214452)

[10. Tham khảo 19](#_Toc106214453)

# **LỜI CẢM ƠN**

Để hoàn thành tốt đề tài và bài báo cáo này, chúng em xin gửi lời cảm ơn chân thành đến giảng viên Quách Đình Hoàng, người đã trực tiếp hỗ trợ chúng em trong suốt quá trình học. Chúng em cảm ơn thầy đã đưa ra những lời khuyên từ kinh nghiệm thực tiễn của mình để giúp đỡ và định hướng cho chúng em đi đúng với yêu cầu của đề tài đã chọn, luôn giải đáp thắc mắc và đưa ra những góp ý, chỉnh sửa kịp thời giúp chúng em khắc phục nhược điểm và hoàn thành đề tài đúng tiến độ đề ra.

Chúng em cũng xin gửi lời cảm ơn chân thành các quý thầy cô trong khoa Công Nghệ Thông tin và các anh chị khoá trên đã tận tình hướng dẫn, truyền đạt những kiến thức cần thiết giúp chúng em có nền tảng để làm nên đề tài này, đã tạo điều kiện để chúng em có thể tìm hiểu và thực hiện tốt đề tài. Cùng với đó, chúng em xin được gửi cảm ơn đến các bạn cùng khóa đã cung cấp nhiều thông tin và kiến thức hữu ích giúp chúng em có thể hoàn thiện hơn đề tài của mình.

Đề tài và bài báo cáo được chúng em thực hiện trong khoảng thời gian ngắn, còn gặp nhiều hạn chế trong kiến thức và kỹ thuật cũng như kinh nghiệm trong việc thực hiện một đề tài viết chương trình. Do đó, chương trình còn nhiều thiếu sót là điều không thể tránh khỏi. Vậy nên chúng em rất mong nhận được những ý kiến đóng góp quý báu của các quý thầy để nhóm chúng em có thể hoàn thiện hơn về kiến thức, tư duy và rèn luyện kỹ năng để có thể làm tốt hơn nữa trong những lần sau.

Chúng em xin chân thành cảm ơn. Cuối lời, chúng em kính chúc thầy luôn dồi dào sức khỏe và thành công hơn nữa trong sự nghiệp trồng người.

Nhóm chúng em xin chân thành cảm ơn!

# **NỘI DUNG**

**1. Tóm tắt**

- Động lực: Bệnh tim mạch là các tình trạng liên quan đến sức khỏe của trái tim, sự hoạt động của các mạch máu gây suy yếu khả năng làm việc của tim. Các bệnh tim mạch bao gồm: các bệnh mạch máu như bệnh động mạch vành, bệnh cơ tim, loạn nhịp tim và suy tim. Bệnh tim mạch có thể xảy ra ở mọi độ tuổi, giới tính, nghề nghiệp. Bệnh không thể chữa khỏi hoàn toàn, đòi hỏi sự điều trị và theo dõi cẩn thận (thậm chí là suốt đời), tốn kém nhiều chi phí. Theo Tổ chức Y tế Thế giới, hàng năm có đến hơn 17 triệu người tử vong vì bệnh tim mạch. Con số này cho thấy mức độ phổ biến của các bệnh lý về tim là rất cao. Do đó, nhóm chúng em đã đi đến thống nhất sẽ tìm hiểu về các nguyên nhân dẫn đến việc bệnh tim ngày càng xuất hiện nhiều trong đời sống con người qua chủ đề: "Phân tích các yếu tố nguy cơ dẫn đến bệnh tim mạch".

- Phương pháp được nhóm chúng em áp dụng để phân tích chủ đề là: thống kê mô tả và thống kê suy diễn.

- Kết quả sơ bộ: tìm ra được 1 số nguyên nhân chính gây ra bệnh tim mạch ở con người hiện nay.

## **2. Giới thiệu**

Do nhiều nguyên nhân khác nhau mà tỷ lệ xuất hiện bệnh tim mạch ngày càng cao trong đời sống. Điều này gây hoang mang cho rất nhiều người dân, họ không biết rằng liệu các yếu tố nào trong cuộc sống hằng ngày có thể làm cho họ mắc phải căn bệnh vừa tổn thất về mặt vật chất lẫn tinh thần này. Từ đó, nhóm chúng em đã đặt ra những câu hỏi sau đây để phân tích và trả lời một cách sơ lược về các yếu tố là nguyên nhân gây ra bệnh về tim mạch.

    - Một câu hỏi được rất nhiều người đặt ra đó là liệu người ở độ tuổi nào sẽ có nguy cơ mắc bệnh tim cao hơn. Có phải người trẻ thì sẽ không mác phải căn bệnh nguy hiểm này không? Từ những vấn đề trên, nhóm chúng em quyết định sẽ nghiên cứu rằng: "Người lớn tuổi có khả năng mắc bệnh tim mạch cao hơn người trẻ tuổi không?" Biến giải thích được sử dụng cho câu hỏi này là: Age và AgeCategory, biến kết quả là HeartDisease.

    - Ở một khía cạnh khác, chỉ số khối cơ thể (BMI) được cho là một nguyên nhân gây ra bệnh tim mạch ở con người ngày nay, nhưng BMI cao hay thấp, như thế nào là vừa đủ để tránh khỏi nguy cơ mắc bệnh tim. Câu hỏi đặt ra là: "Người có chỉ số khối cơ thể (BMI) cao có khả năng mắc bệnh tim cao hơn không?" với biến giải thích là BMI và biến kết quả là HeartDisease.

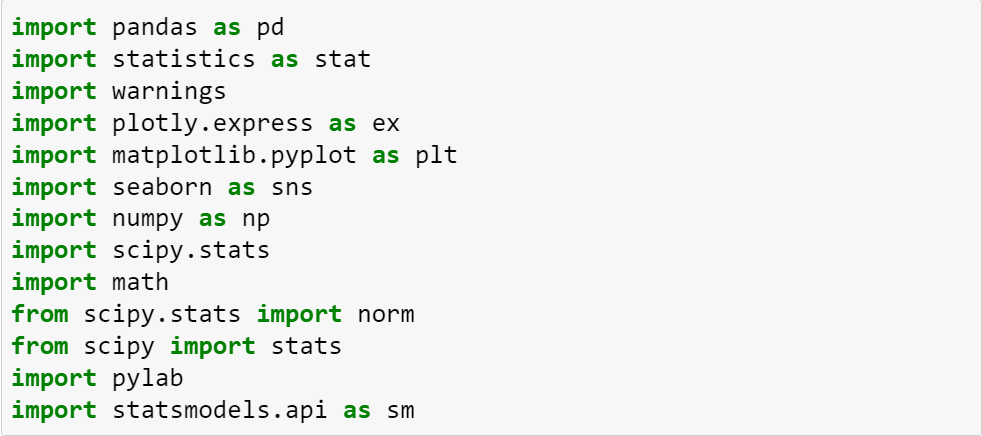
    - Rượu bia và thuốc lá là những thú vui được rất nhiều người sử dụng trong cuộc sống hiện nay, nhưng ngoài là nguyên nhân gây ra các bệnh về gan, phổi như chúng ta thường nghe thì liệu chúng có là nguy cơ dẫn tới bệnh về tim mạch không và cái nào có nguy cơ cao hơn? Với 2 biến giải thích là Smoking và AlcoholDrinking, biến kết quả HeartDisease, nhóm chúng em sẽ phân tích câu hỏi: "Người hút thuốc là có khả năng mắc bệnh tim cao hơn so với người nghiện rượu không?"

1. **Dữ liệu**

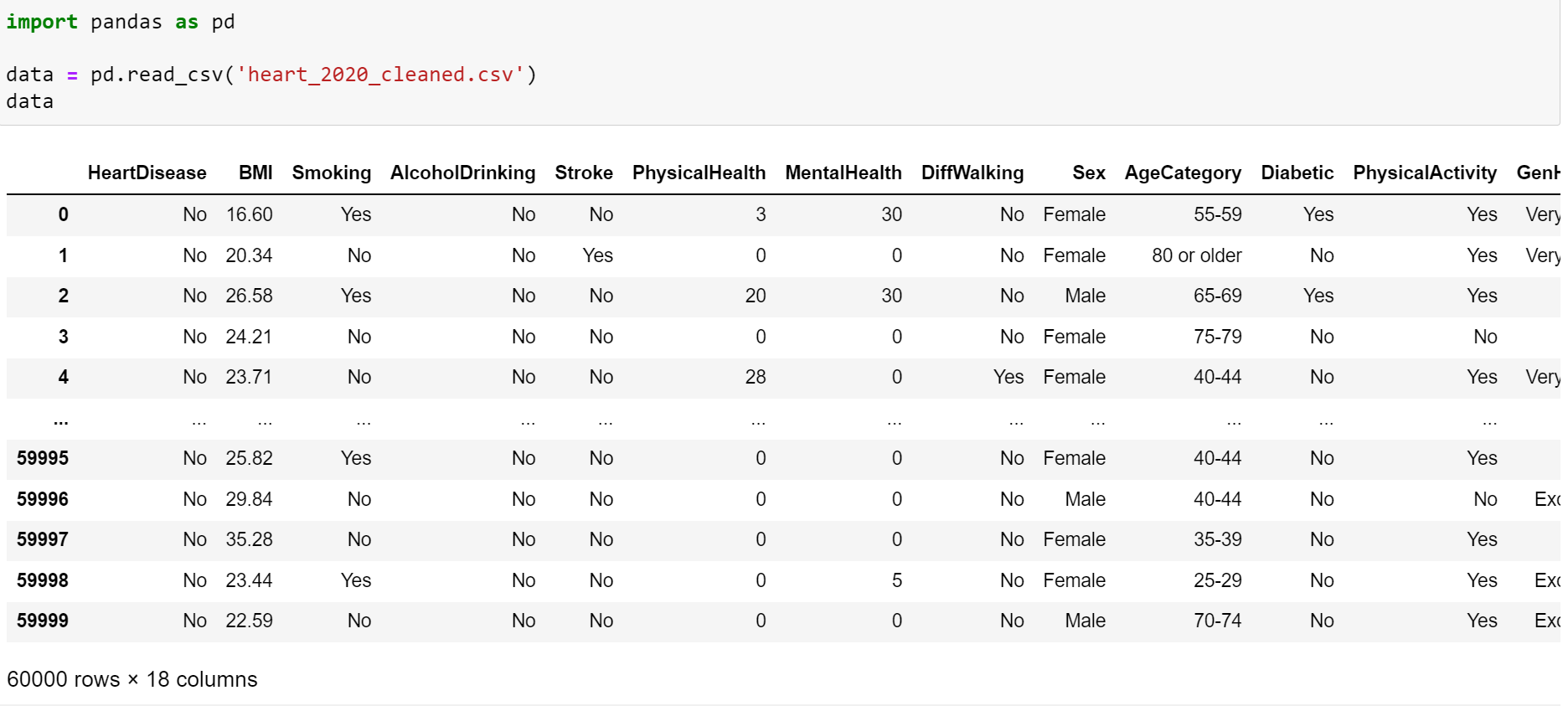
* Tên đề tài: Phân tích các yếu tố nguyên nhân dẫn đến bệnh tim mạch.
* Tập dữ liệu được lấy từ trang:

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

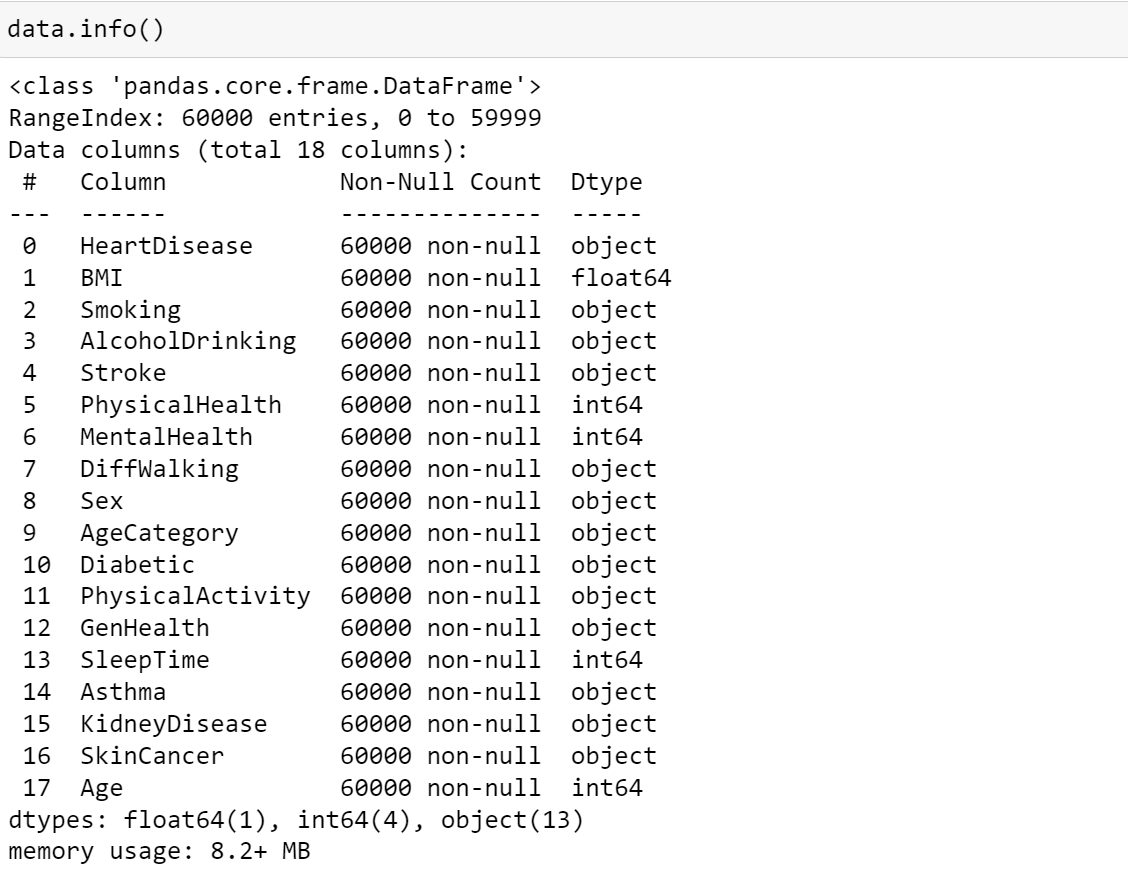
* Ban đầu, bộ dữ liệu đến từ CDC và là một phần chính của Hệ thống giám sát yếu tố rủi ro hành vi (BRFSS), tiến hành các cuộc khảo sát qua điện thoại hàng năm để thu thập dữ liệu về tình trạng sức khỏe của cư dân Hoa Kỳ. Được thành lập vào năm 1984 với 15 tiểu bang, BRFSS hiện đang thu thập dữ liệu ở tất cả 50 tiểu bang cũng như Quận Columbia và ba vùng lãnh thổ của Hoa Kỳ. BRFSS hoàn thành hơn 400.000 cuộc phỏng vấn người lớn mỗi năm, khiến nó trở thành hệ thống khảo sát sức khỏe được tiến hành liên tục lớn nhất trên thế giới.
* Bộ dữ liệu này được sử dụng để dự đoán liệu một người có khả năng mắc bệnh tim hay không dựa trên các thông số đầu vào như giới tính, tuổi tác, hút thuốc, bia rượu. Mỗi hàng trong dữ liệu cung cấp thông tin liên quan về bệnh nhân.
* Bộ dữ liệu gồm: 18 cột và 60.000 dòng.
* Biến phân loại (định tính): HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, AgeCategory, Diabetic, PhysicalActivity, GenHealth, Asthma, KidneyDisease, SkinCancer
* Biến số (định lượng): BMI, PhysicalHealth, MentalHealth, SleepTime, Age
* Biến kết quả (phản hồi, Y): HeartDisease
* Biến dự đoán (giải thích, X): BMI, Smoking, AlcoholDrinking, Stroke, PhysicalHealth, MentalHealth, DiffWalking, Sex, AgeCategory, Diabetic, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer, Age.
* Một số thư viện nhóm đã dùng:

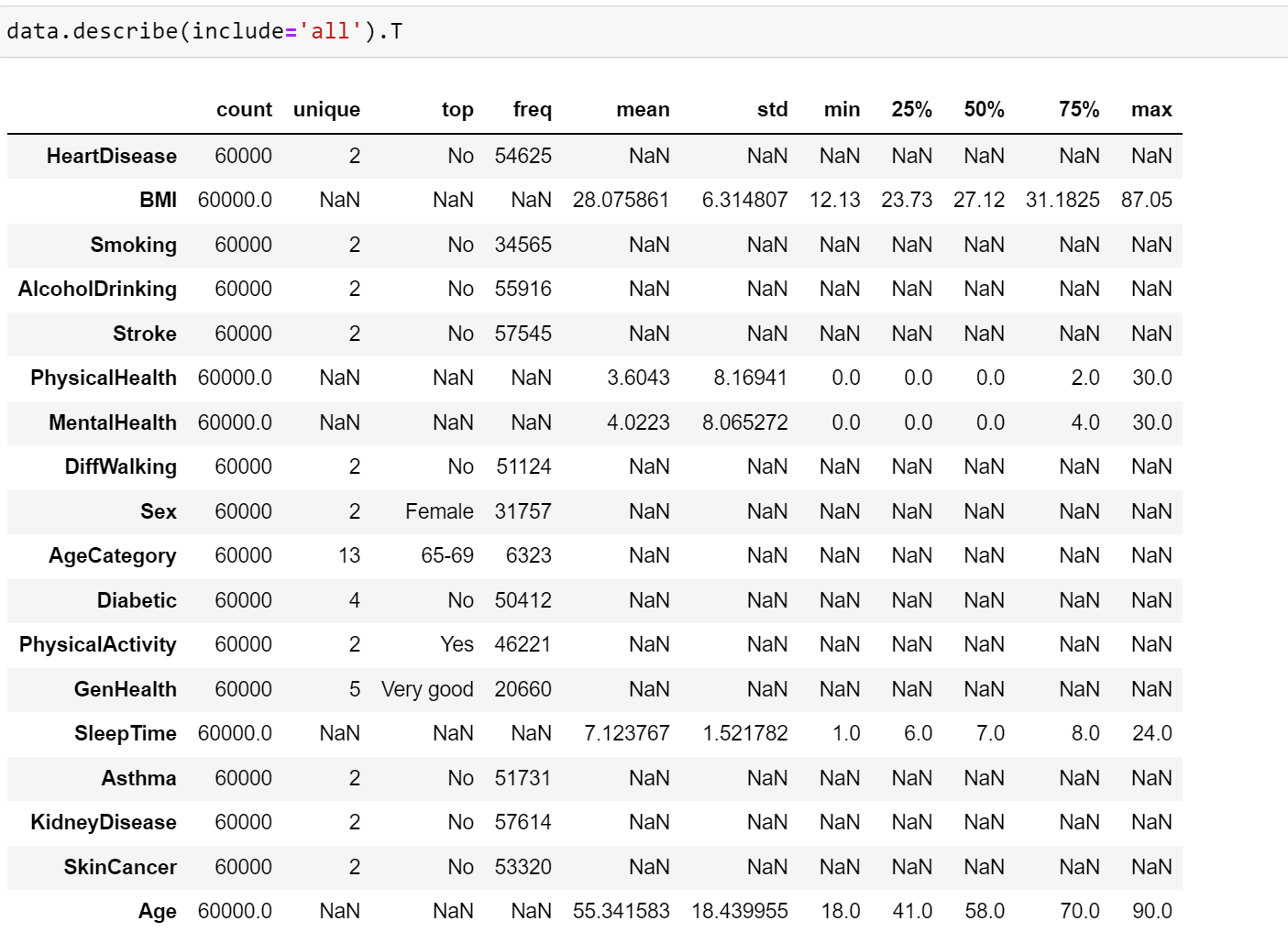


* Đọc dữ liệu từ file csv:

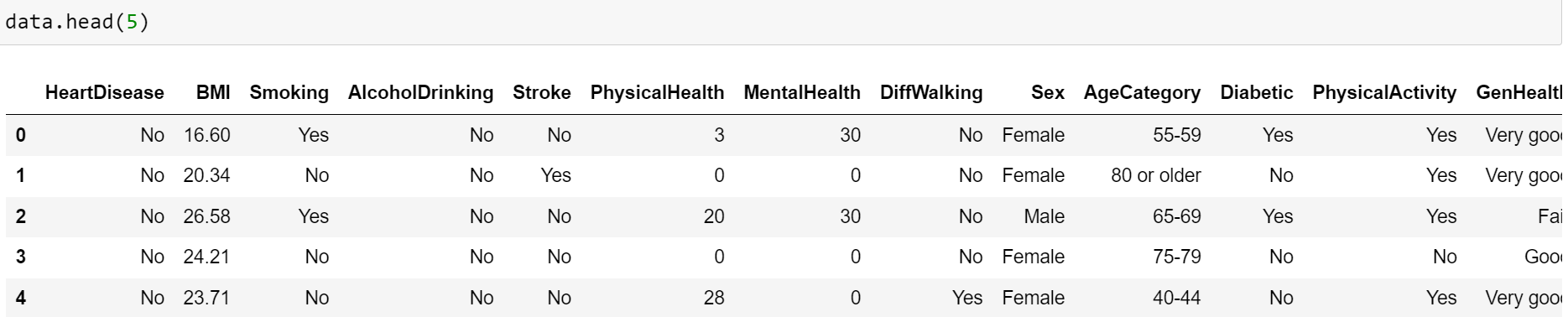


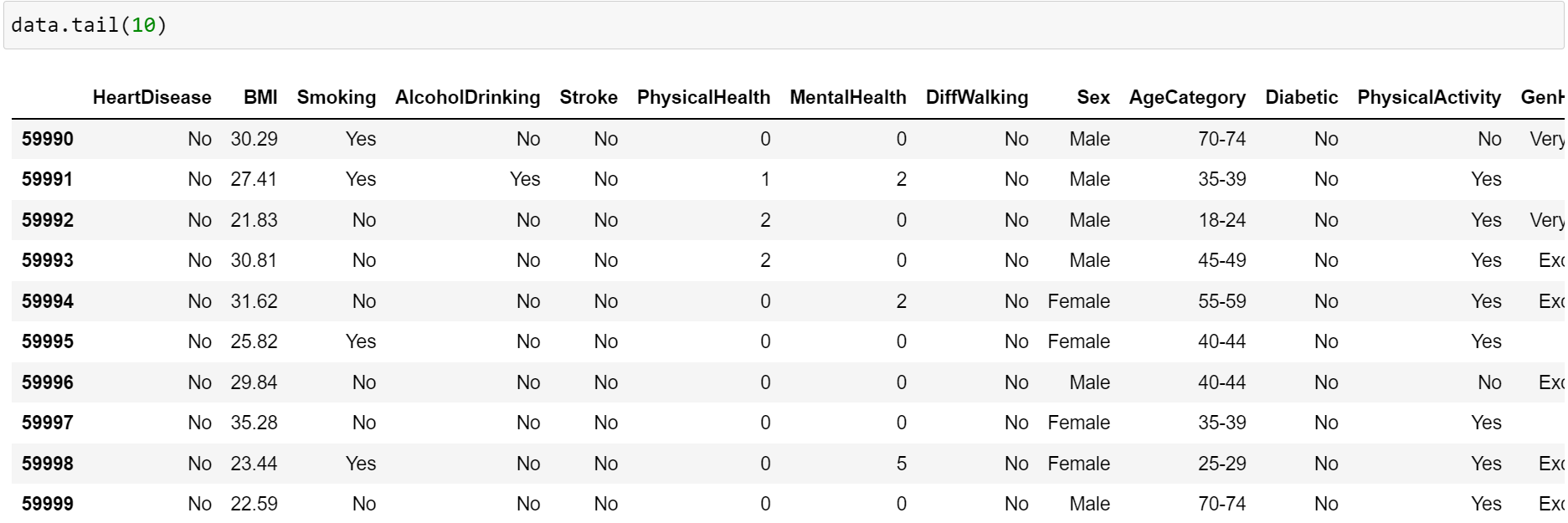
* Thông tin tổng quát về tập dữ liệu:

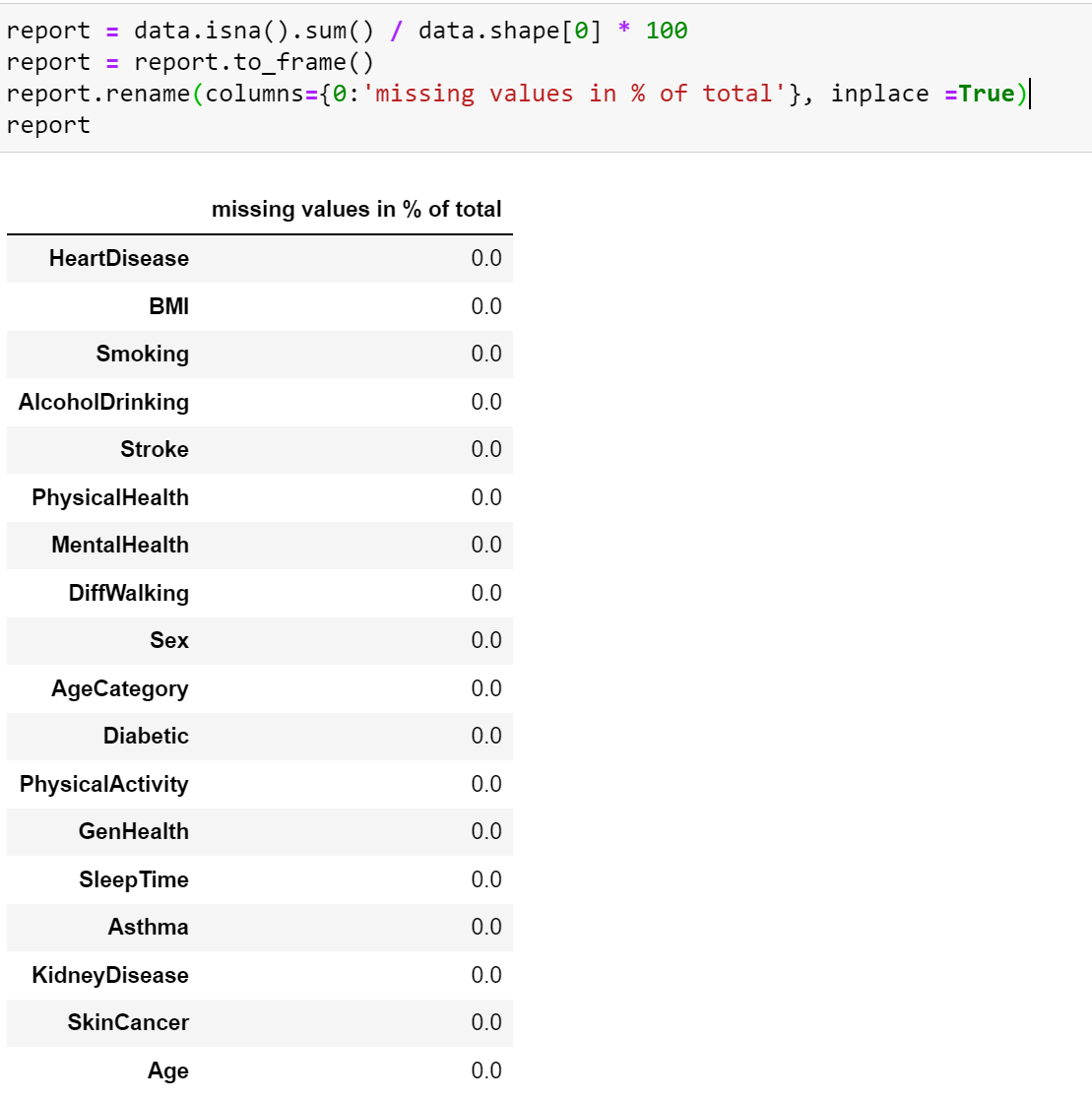
****

****

* Một số dòng đầu và cuối của tập dữ liệu:

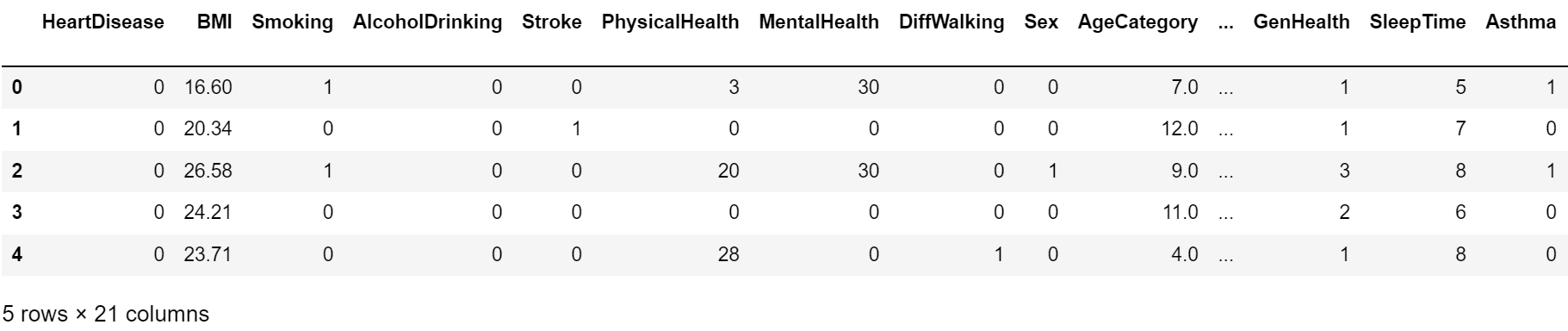


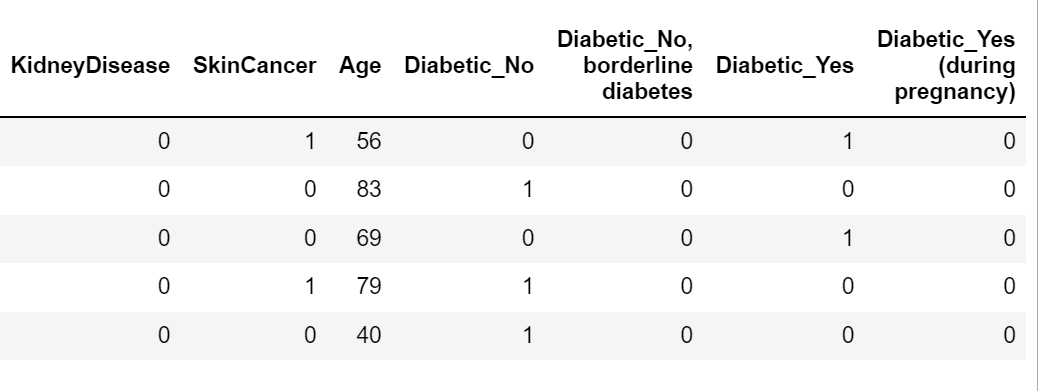


****

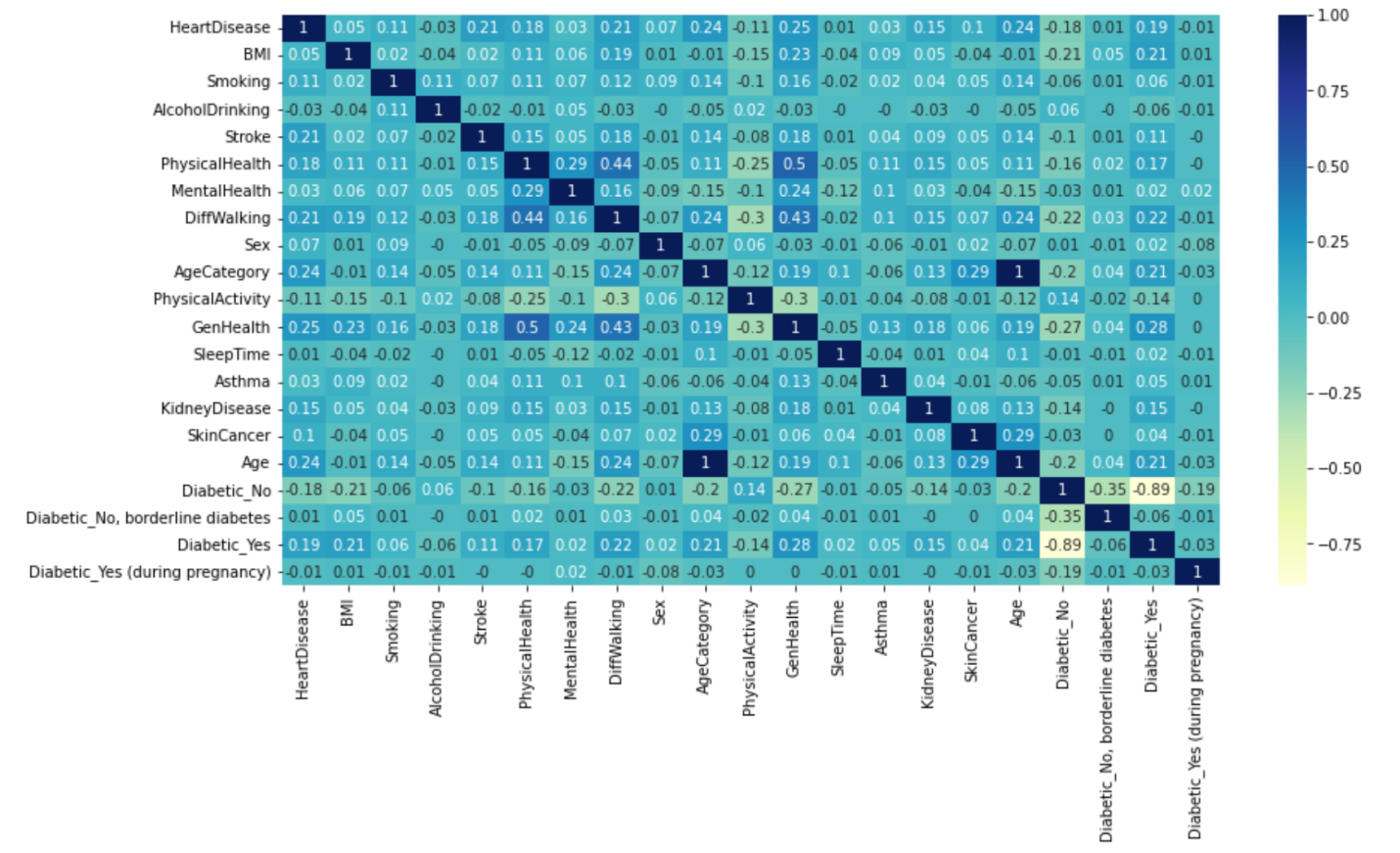
1. **Trực quan hóa dữ liệu**

* Mã hóa các cột đối số thành số và in ra 5 dòng đầu của tập dữ liệu:

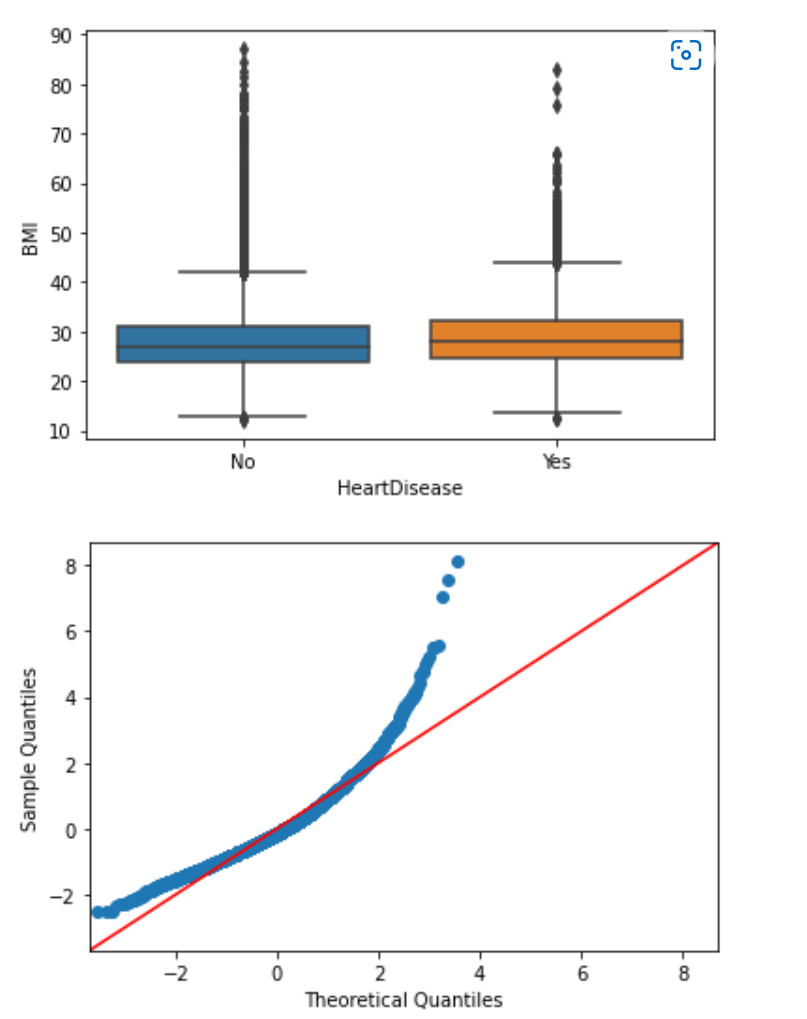
****

****

* Biểu đồ biểu thị độ tương quan giữa các thuộc tính với nhau

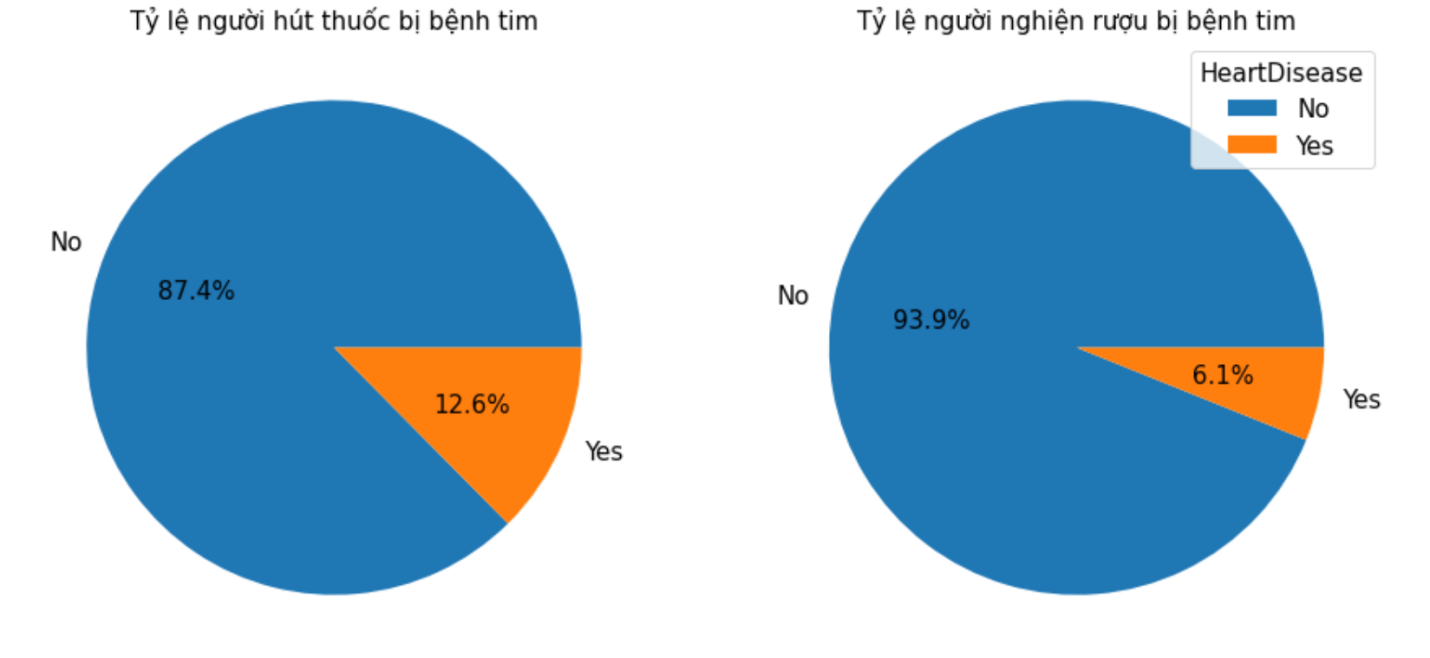


* Một vài biểu đồ trực quan cho thấy các thuộc tính ảnh hưởng thế nào đến bệnh tim mạch:
* BMI



**Nhận xét:**

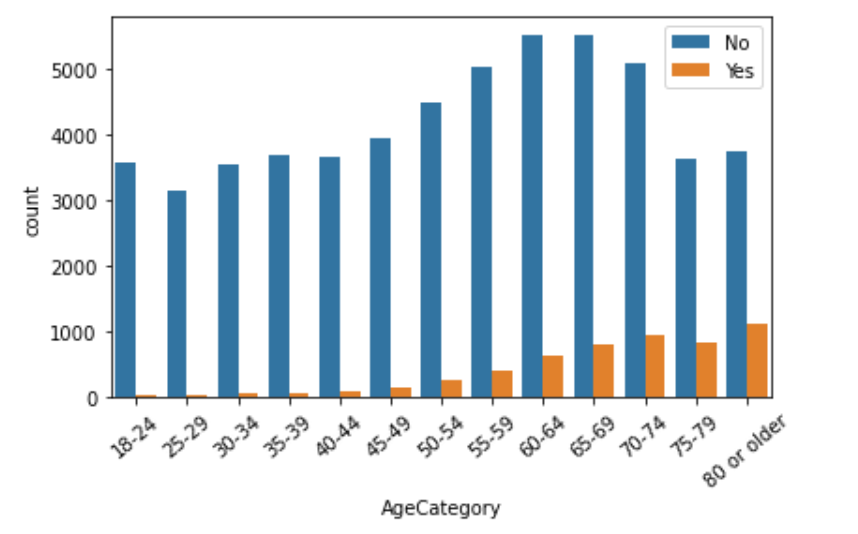
* Người bị bệnh tim có trung bình và trung vị chỉ số khối cơ thể (BMI) cao hơn người không bị bệnh tim nhưng không nhiều.
* Dựa vào biểu đồ hộp ta có thể thấy 50% số người bị bệnh tim có số BMI khoảng 25 - 32kg/m2.
* Dựa vào biểu đồ QQplot ta có thể thấy những trường hợp bị bệnh tim khi có BMI quá cao bị sai lệch khá nhiều còn lại thì sai lệch khá ít so với đường thẳng QQPlot.
* Smoking và AlcoholDrinking

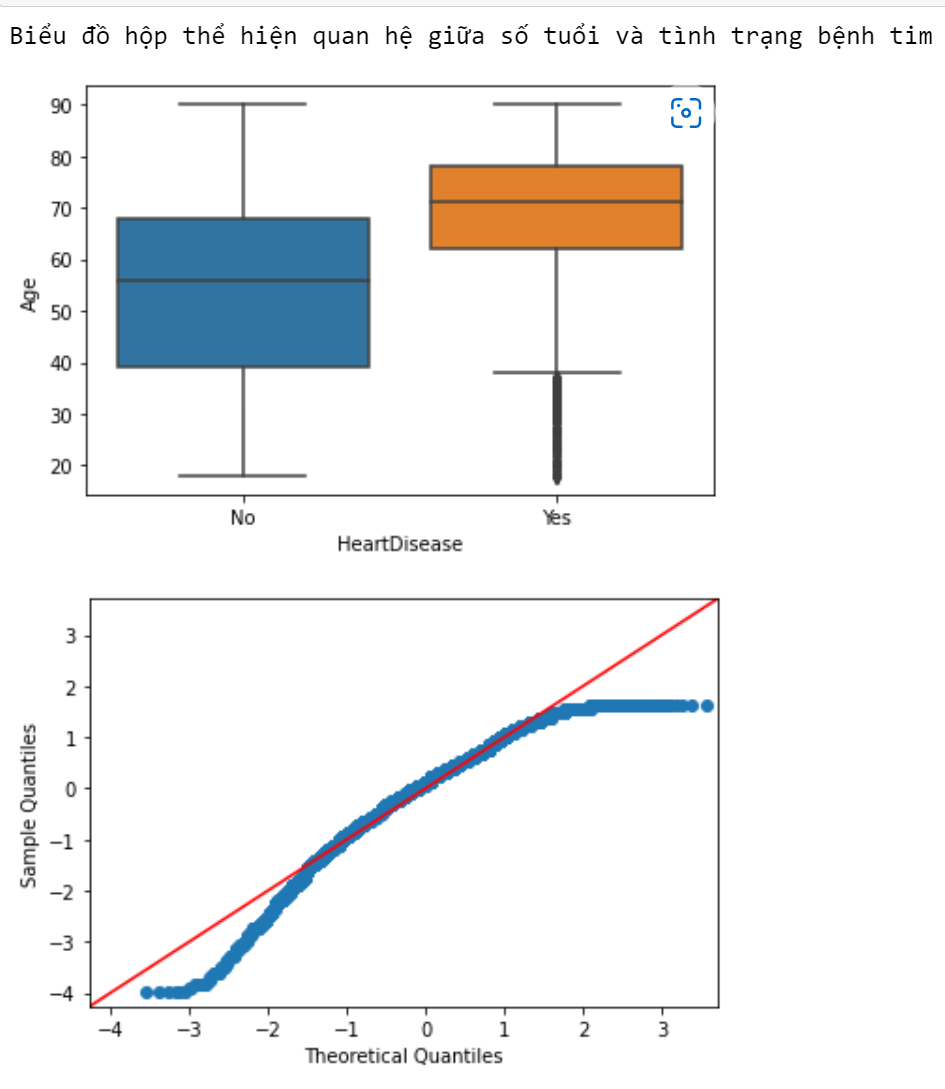


**Nhận xét:**

Từ biểu đồ trên ta có thể thấy tỷ lệ những người hút thuốc lá có nguy cơ bị bệnh tim cao hơn những người nghiện rượu.

* AgeCategory và Age





**Nhận xét:**

* Tỷ lệ người lớn tuổi (>= 60 tuổi) có nguy cơ mắc bệnh tim cao hơn hẳn so với người trẻ tuổi (< 60 tuổi) là hơn 10%.
* Trung bình độ tuổi của những người bị bệnh tim lớn hơn hẳn so với những người không bị bệnh tim.
* Dựa vào biểu đồ hộp ta có thể thấy có 50% số người bị bệnh tim nằm trong khoảng 63 - 79 tuổi và rất ít người dưới 40 tuổi bị bệnh tim (giá trị ngoại lai).
* Biểu đồ QQplot ta có thể thấy những trường hợp bị bệnh tim khi ở độ tuổi quá thấp hoặc quá cao bị sai lệch khá nhiều nhưng độ tuổi còn lại thì gần như không sai lệch so với đường thẳng QQplot.

1. **Mô hình hóa dữ liệu**
2. **Thống kê mô tả**

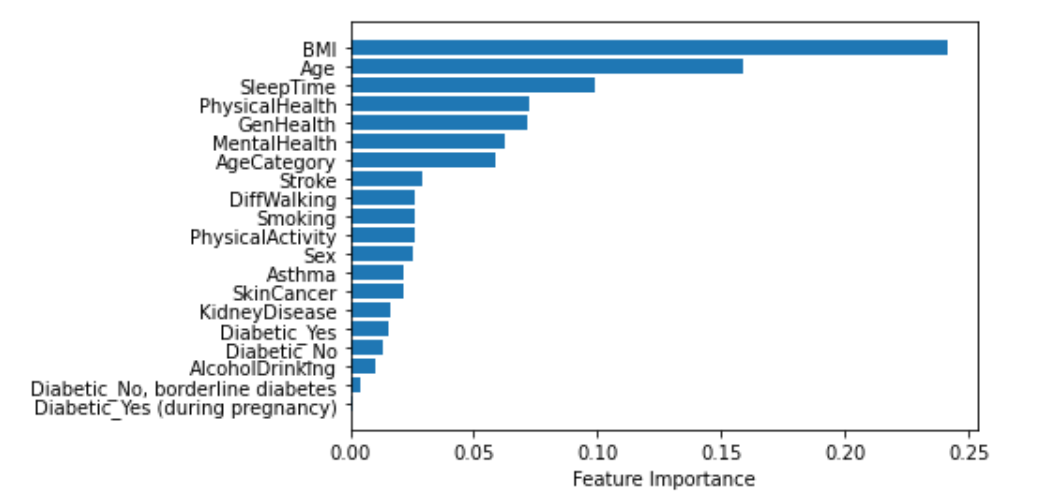
* Thống kê mô tả là phương pháp dùng để giải thích, tóm tắt hoặc mô tả các đặc tính của một mẫu nghiên cứu dưới dạng số liệu hay các biểu đồ trực quan.
* Để tóm tắt và mô tả tập dữ liệu, nhóm chúng em đã sử dụng:

+ Các công cụ số là mean, mode, median, độ lệch chuẩn, max, min và khoảng tứ phân vị.

+ Các công cụ trực quan là các biểu đồ cột, biểu đồ 2 cột, biểu đồ tròn, biểu đồ hộp, biểu đồ Q-Q.

1. **Thống kê suy diễn**

* Thống kê suy diễn là phương pháp dùng để suy luận và rút ra kết luận về đặc tính của một quần thể trên cơ sở phân tích và quan sát mẫu.
* Cụ thể, nhóm chúng em đã sử dụng kiểm định giả thuyết là kiểm định t-test. Và cụ thể hơn thì nhóm em sử dụng kiểm định two-sample t-test để kiểm định sự khác biệt của giá trị trung bình giữa 2 mẫu.

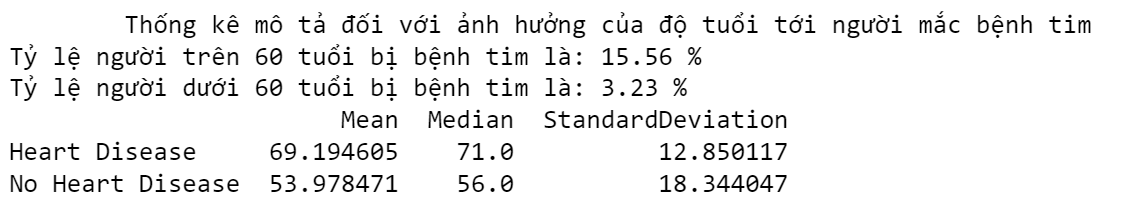


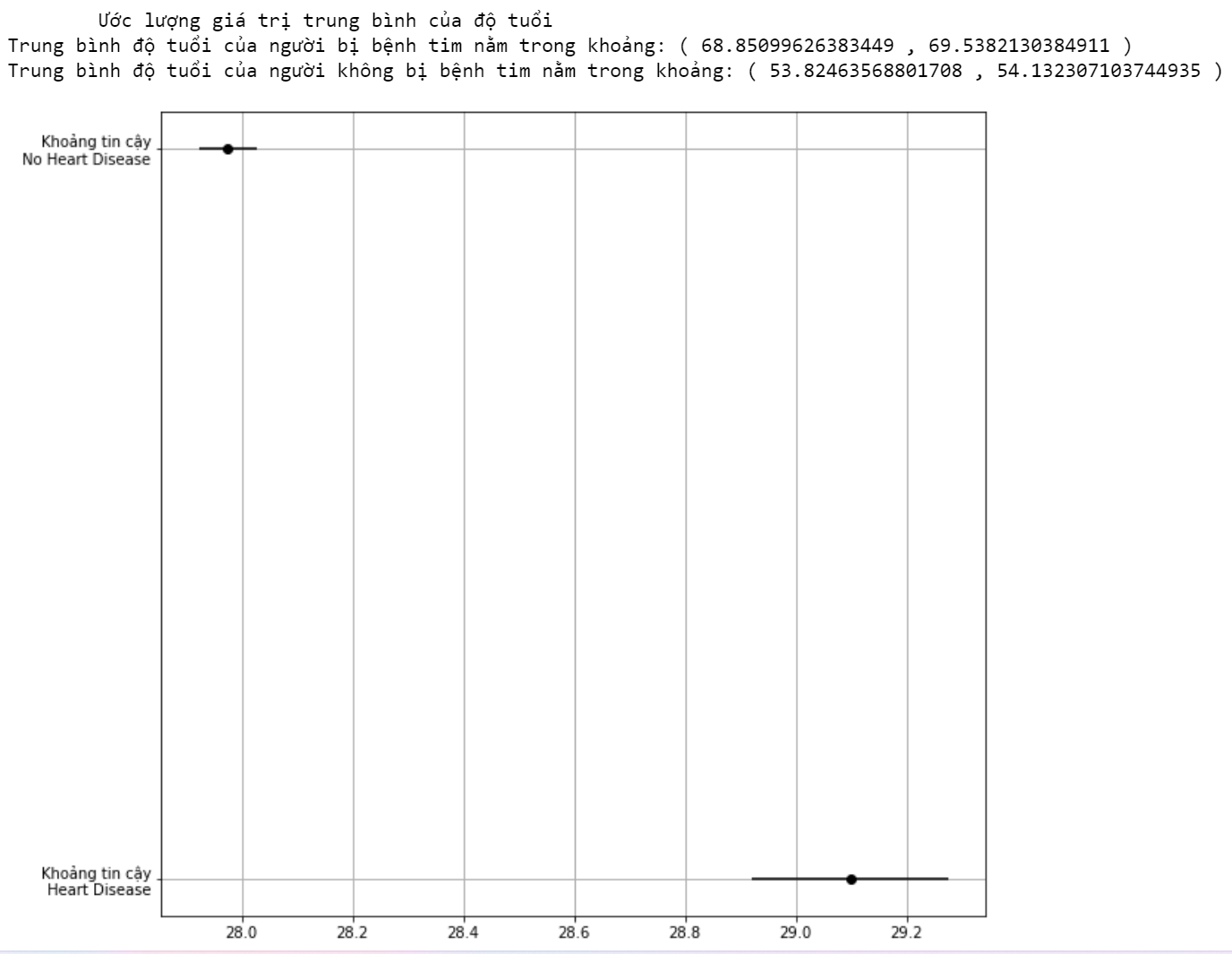
Khi đo mức độ quan trọng của từng thuộc tính đối với nguy cơ mắc bệnh tim, ta thấy BMI là yếu tố có ảnh hưởng đến nguy cơ bị bệnh tim nhiều nhất trong số các thuộc tính.

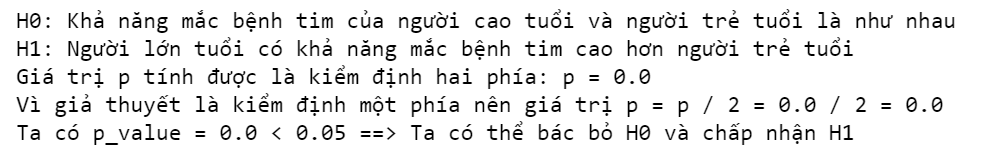
1. **Thực nghiệm, kết quả, và thảo luận**

*Câu hỏi nghiên cứu:*

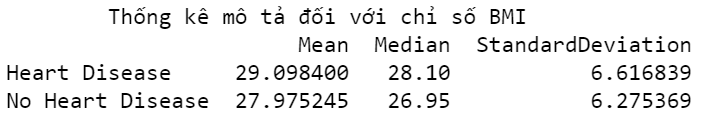
***- H1: Người lớn tuổi có khả năng mắc bệnh tim cao hơn người trẻ tuổi.***

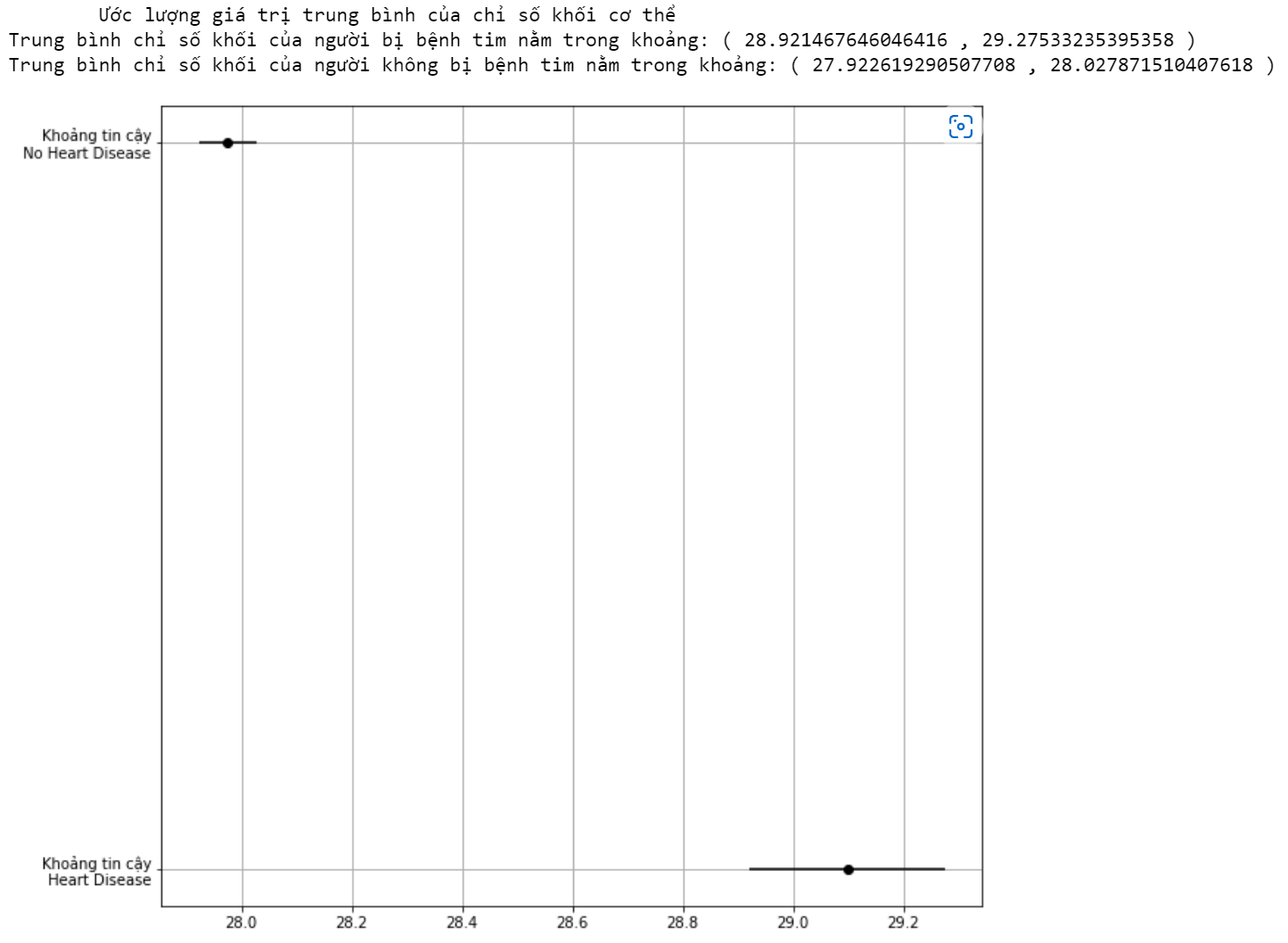


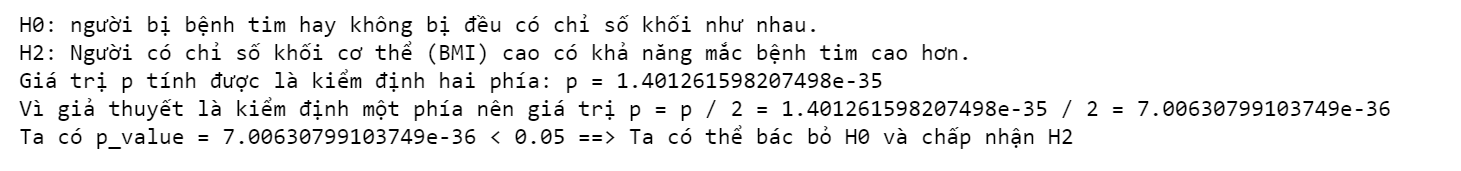




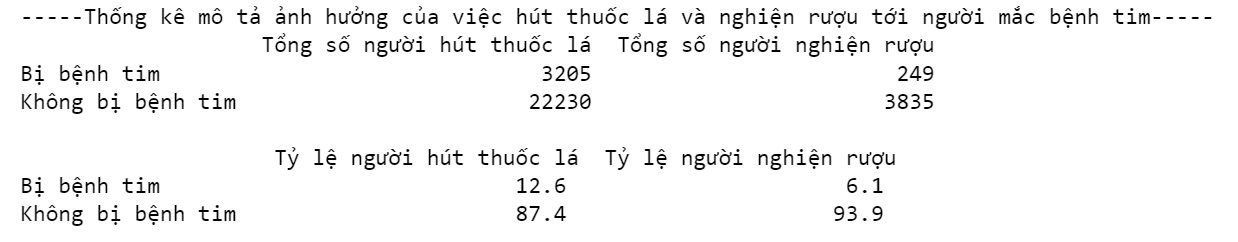
***- H2: Người có chỉ số khối cơ thể (BMI) cao có khả năng mắc bệnh tim cao hơn.***

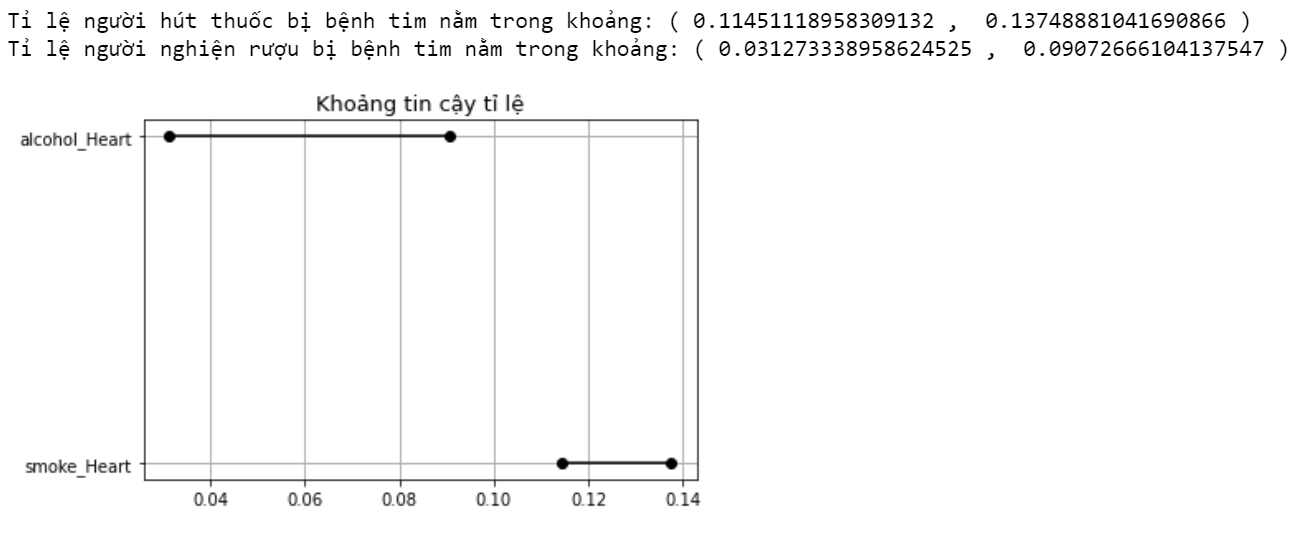
******

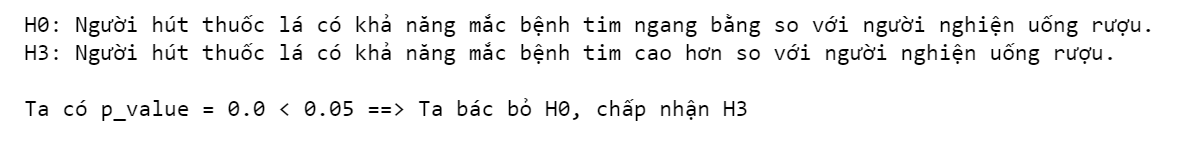
******



***- H3: Người hút thuốc lá có khả năng mắc bệnh tim cao hơn so với người nghiện uống rượu.***

******

******



1. **Kết luận**

- Sau khi phân tích và kiểm định, chúng em đã có bằng chứng chỉ ra rằng những người có độ tuổi càng cao thì sẽ có tỉ lệ mắc bệnh tim cao hơn những người có độ tuổi thấp hơn. Thêm vào đó, những người có chỉ số khối cơ thể (BMI) cao sẽ có nguy cơ bị bệnh tim cao hơn những người có chỉ số BMI thấp.

- Do nghiên cứu chúng em thực hiện chỉ là nghiên cứu quan sát nên chỉ có thể kết luận sự khác biệt chứ không thể đưa ra kết luận chính xác, vì còn phụ thuộc nhiều vào các biến ẩn khác.

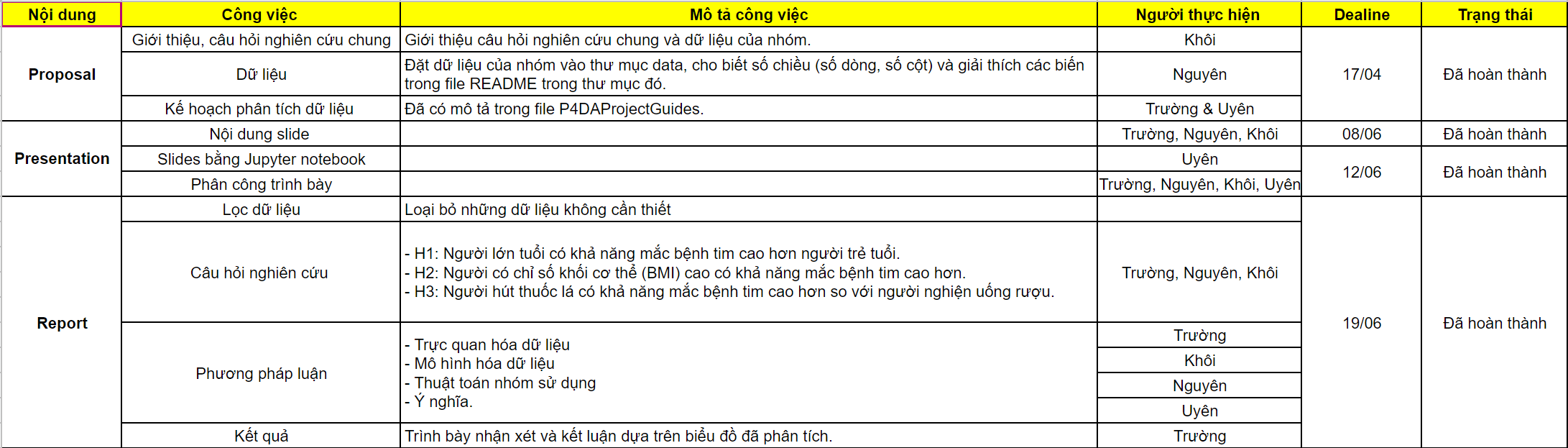
- Nhóm chúng em chọn tập dữ liệu trên kaggle và tập dữ liệu này được lấy từ khảo sát trực tuyến nên những người tham gia có thể bị ảnh hưởng bởi nhiều yếu tố chủ quan làm tập dữ liệu có nhiều sai sót.

- Sự hiểu biết và thời gian có hạn nên bài báo cáo vẫn còn nhiều lỗi. Nếu nhóm em có nhiều thời gian hơn thì nhóm em sẽ thay đổi phương pháp chọn mẫu để có kết quả khách quan hơn và tìm ra thêm nhiều biến ẩn ảnh hưởng đến bệnh tim để kết quả cuối cùng có thể chính xác hơn.

1. **Phụ lục**

*Không có*

1. **Đóng góp**

****

1. **Tham khảo**
   * Quách Đình Hoàng (2022). *Các bài giảng môn Nhập môn lập trình Python cho phân tích*.
   * Kaggle (2022), *Personal Key Indicators of Heart Disease.* Retrieved from:

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

* + Kaggle (2022), *Classification of Heart Disease Personal Indictors.* Retrieved from:

https://www.kaggle.com/code/catherinevj/classification-of-heart-disease-personal-indictors