**ĐẠI HỌC HUẾ**



# KHOA KỸ THUẬT VÀ CÔNG NGHỆ

🙠🙟🕮🙝🙢

****

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

**Giáo viên hướng dẫn: NGUYỄN ĐÌNH HOA CƯƠNG**

**Lớp: KHDL & TTNT**

|  |
| --- |
| Số phách  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày 28 tháng 6 năm 2021**

# ĐẠI HỌC HUẾ



# KHOA KỸ THUẬT VÀ CÔNG NGHỆ

🙠🙟🕮🙝🙢

****

(MẪU BÌA PHỤ)

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

**Giảng viên hướng dẫn: NGUYỄN ĐÌNH HOA CƯƠNG**

**Lớp: KHDL & TTNT**

**Sinh viên thực hiện: PHAN THẾ MINH CHÂU**

*(ký tên và ghi rõ họ tên)*

|  |
| --- |
| Số phách  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày 28 tháng 6 năm 2021**

# 

# Mục lục:

[Mục lục: 1](#_Toc75789159)

[Chương I: Cơ sở lí thuyết 2](#_Toc75789160)

[1 Thống kê là gì? 2](#_Toc75789161)

[a. Định nghĩa: 2](#_Toc75789162)

[b. Lịch sử ra đời và nguồn gốc 3](#_Toc75789163)

[c. Các ứng dụng của thống kê 3](#_Toc75789164)

[2 Thống kê bằng ngôn ngữ lập trình Python: 4](#_Toc75789165)

[a. Giới thiệu về ngôn ngữ lập trình Python. 4](#_Toc75789166)

[Chương 2: Phân tích dữ liệu 6](#_Toc75789167)

[2.1 Thống kê mô tả 6](#_Toc75789168)

[2.2 Thống kê suy diễn 11](#_Toc75789169)

[Quy trình kiểm định ANOVA: 12](#_Toc75789170)

[Chương 3: Kiểm tra đạo văn 16](#_Toc75789171)

[Chương 4: Tài liệu tham khảo 17](#_Toc75789172)

# Chương I: Cơ sở lí thuyết

## Thống kê là gì?

### a. Định nghĩa:

- **Suy luận thống kê** là quá trình suy luận suy ra các đặc điểm của một phân phối cơ bản bằng việc phân tích dữ liệu. Phân tích thống kê suy luận suy ra tính chất của tổng thể: điều này bao gồm các giả thuyết thử nghiệm và các ước tính phát sinh. Tổng thể được giả định là lớn hơn so với tạo ra các dữ liệu quan sát, nói cách khác, các dữ liệu quan sát được giả định là lấy mẫu từ một tổng thể lớn hơn.

- Thống kê được sử dụng hầu như trong tất cả các ngành khoa học như khoa học vật lý và xã hội, cũng như trong kinh doanh, nhân văn, chính phủ và sản xuất.

- Các nhà thống kê, những người làm công tác thống kê, đặc biệt quan tâm đến việc xác định cách đưa ra kết luận đáng tin cậy về các nhóm lớn và các hiện tượng chung từ các đặc điểm quan sát được của các mẫu nhỏ chỉ đại diện cho một phần nhỏ của nhóm lớn hoặc một số trường hợp hạn chế của một hiện tượng chung.

- Lợi ích của thống kê:

Thống kê sử dụng bằng chứng số để rút ra kết luận hợp lệ

Các nhà thống kê biết cách tránh những sai lầm thường gặp

- Hai lĩnh vực chính của thống kê là:

* Thống kê mô tả.
* Thống kê suy diễn

#### \*\*Thống kê mô tả

- Thống kê mô tả đề cập đến một chuyên ngành mô tả định lượng các đặc điểm quan trọng của bộ dữ liệu. Đối với mục đích mô tả các thuộc tính, nó sử dụng các số đo của xu hướng trung tâm, nghĩa là trung bình, trung vị, chế độ và các biện pháp phân tán tức là phạm vi, độ lệch chuẩn, độ lệch chuẩn và phương sai, v.v.

- Dữ liệu được nhà nghiên cứu tóm tắt, theo cách hữu ích, với sự trợ giúp của các công cụ số và đồ họa như biểu đồ, bảng và biểu đồ, để biểu diễn dữ liệu một cách chính xác. Hơn nữa, văn bản được trình bày hỗ trợ các sơ đồ, để giải thích những gì họ đại diện.

#### \*\*Thống kê suy diễn

- Thống kê suy luận là tất cả về khái quát hóa từ mẫu đến dân số, tức là kết quả phân tích mẫu có thể được suy ra cho dân số lớn hơn, từ đó mẫu được lấy. Đó là một cách thuận tiện để đưa ra kết luận về dân số khi không thể truy vấn từng thành viên trong vũ trụ. Mẫu được chọn là đại diện của toàn bộ dân số; do đó, nó nên chứa các tính năng quan trọng của dân số.

- Thống kê suy luận được sử dụng để xác định xác suất tính chất của dân số trên cơ sở tính chất của mẫu, bằng cách sử dụng lý thuyết xác suất. Các thống kê suy luận chính dựa trên các mô hình thống kê như Phân tích phương sai, kiểm tra chi bình phương, phân phối t của sinh viên, phân tích hồi quy, vv Các phương pháp thống kê suy luận:

* Ước tính các tham số
* Kiểm định giả thuyết

### b. Lịch sử ra đời và nguồn gốc

- Thời kỳ chiếm hữu nô lệ: Ghi chép các con số

- Thời kỳ Phong kiến: Phân tích, đánh giá theo thời gian, không gian

- Thời kỳ sản xuất hang hóa: Thế hiện mối quan hệ lượng chất

- Giai đoạn hiện nay: Là một trong những công cụ quản lý vĩ mô quan trọng, có vai trò cung cấp các thông tin phục vụ quản lý

### c. Các ứng dụng của thống kê

Ngày nay, thống kê đóng vai trò quan trọng trong nhiều lĩnh vực:

- Học qua máy và khai thác dữ liệu: có hai ứng dụng cho học qua máy móc và khai thác dữ liệu: quản lý dữ liệu và phân tích dữ liệu. Các công cụ thống kê cần thiết cho việc phân tích dữ liệu.

- Thống kê trong xã hội học: thống kê được áp dụng cho một loạt các môn học, bao gồm cả khoa học tự nhiên và xã hội, chính trị và kinh doanh. Thống kê tư vấn có thể giúp các tổ chức và công ty không có chuyên môn trả lời những thắc mắc.

- Một số lĩnh vực sử dụng điều tra thống kê được áp dụng rộng rãi rằng họ có chuyên môn. Những ngành này bao gồm:

* Khoa học tính toán bảo hiểm (đánh giá rủi ro trong các ngành công nghiệp bảo hiểm và tài chính)
* ứng dụng thông tin kinh tế
* thiên văn học (đánh giá thống kê của dữ liệu thiên văn)
* sinh học
* thống kê kinh doanh
* hóa học (phân tích dữ liệu hóa học)
* khai thác dữ liệu (áp dụng thống kê và nhận dạng mẫu để khám phá tri thức từ dữ liệu)
* dân số học
* kinh tế học (phân tích thống kê các số liệu kinh tế)
* thống kê năng lượng
* thống kê kỹ thuật
* khoa học nghiên cứu bệnh dịch (phân tích thống kê của bệnh)
* địa lý và hệ thống thông tin địa lý, đặc biệt trong phân tích không gian
* xử lý hình ảnh
* thống kê y tế
* thống kê về hành vi, tâm lý
* độ bền cơ khí
* thống kê xã hội

- Ngoài ra còn có các loại cụ thể của phân tích thống kê cũng đã phát triển các thuật ngữ chuyên ngành thống kê các phương pháp thống kê:

* thống kê đa biến
* phân lớp thống kê
* phân tích dữ liệu có cấu trúc (thống kê)
* mô hình phương trình cấu trúc
* phương pháp điều tra
* phân tích sự tồn tại
* thống kê trong các môn thể thao khác nhau, đặc biệt là bóng chày và bóng bầu dục.

- Thống kê áp dụng cho toán học hay nghệ thuật

## 2 Thống kê bằng ngôn ngữ lập trình Python:

### a. Giới thiệu về ngôn ngữ lập trình Python.

- Python là [ngôn ngữ lập trình](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh) bậc cao dung cho các mục đích lập trình đa năng, do [Guido van Rossum](https://vi.wikipedia.org/w/index.php?title=Guido_van_Rossum&action=edit&redlink=1) tạo ra và lần đầu ra mắt vào năm [1991](https://vi.wikipedia.org/wiki/1990). Ưu điểm của ngôn ngữ lập trình Python là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu.

#### b. Thư viện Pandas

- Pandas là một Thư viện Python mã nguồn mở cung cấp công cụ phân tích và thao tác dữ liệu hiệu suất cao bằng cách sử dụng các cấu trúc dữ liệu mạnh mẽ của nó. Cái tên Pandas có nguồn gốc từ từ Panel Data - một Kinh tế lượng từ dữ liệu Đa chiều.

- Năm 2008, nhà phát triển Wes McKinney bắt đầu phát triển gấu trúc khi cần công cụ phân tích dữ liệu hiệu suất cao, linh hoạt.

- Trước Pandas, Python chủ yếu được sử dụng để trộn và chuẩn bị dữ liệu. Nó có rất ít đóng góp vào việc phân tích dữ liệu. Gấu trúc đã giải quyết vấn đề này. Sử dụng Pandas, chúng ta có thể thực hiện năm bước điển hình trong quá trình xử lý và phân tích dữ liệu, bất kể nguồn gốc của dữ liệu - tải, chuẩn bị, thao tác, lập mô hình và phân tích.

- Python with Pandas được sử dụng trong nhiều lĩnh vực bao gồm các lĩnh vực học thuật và thương mại bao gồm tài chính, kinh tế, thống kê, phân tích,…

- Các tính năng của Pandas:

* Đối tượng DataFrame nhanh và hiệu quả với lập chỉ mục mặc định và tùy chỉnh.
* Các công cụ để tải dữ liệu vào các đối tượng dữ liệu trong bộ nhớ từ các định dạng tệp khác nhau.
* Căn chỉnh dữ liệu và tích hợp xử lý dữ liệu bị thiếu.
* Định hình lại và xoay vòng các nhóm ngày.
* Cắt dựa trên nhãn, lập chỉ mục và tập hợp con các tập dữ liệu lớn.
* Có thể xóa hoặc chèn các cột từ cấu trúc dữ liệu.
* Nhóm theo dữ liệu để tổng hợp và biến đổi.
* Kết hợp và kết hợp dữ liệu hiệu suất cao.
* Chức năng Chuỗi thời gian.

#### c. Thư viên Scipy

- SciPy, một thư viện khoa học cho Python là một thư viện mã nguồn mở, được cấp phép BSD cho toán học, khoa học và kỹ thuật.

- SciPy là một tập hợp các thuật toán toán học và các hàm tiện lợi được xây dựng dựa trên phần mở rộng NumPy của Python. Nó bổ sung sức mạnh đáng kể cho phiên Python tương tác bằng cách cung cấp cho người dùng các lệnh và lớp cấp cao để thao tác và trực quan hóa dữ liệu.

- Lợi ích bổ sung của việc dựa trên SciPy trên Python là điều này cũng tạo ra một ngôn ngữ lập trình mạnh mẽ có sẵn để sử dụng trong việc phát triển các chương trình phức tạp và các ứng dụng chuyên biệt. Các ứng dụng khoa học sử dụng SciPy được hưởng lợi từ việc phát triển các mô-đun bổ sung trong nhiều ngóc ngách của bối cảnh phần mềm bởi các nhà phát triển trên toàn thế giới. Mọi thứ từ lập trình song song đến các chương trình con và lớp cơ sở dữ liệu và web đều đã được cung cấp cho lập trình viên Python. Tất cả sức mạnh này đều có sẵn ngoài các thư viện toán học trong SciPy.

#### d. Thư viện Matplotlyb

- Matplotlib là một trong những gói Python phổ biến nhất được sử dụng để trực quan hóa dữ liệu. Nó là một thư viện đa nền tảng để tạo các đồ thị 2D từ dữ liệu trong các mảng.

- Matplotlib được viết bằng Python và sử dụng NumPy, phần mở rộng toán học số của Python. Nó cung cấp một API hướng đối tượng giúp nhúng các lô trong các ứng dụng sử dụng bộ công cụ GUI Python như PyQt, WxPythonotTkinter. Nó có thể được sử dụng trong Python và IPython shell, máy tính xách tay Jupyter và các máy chủ ứng dụng web.

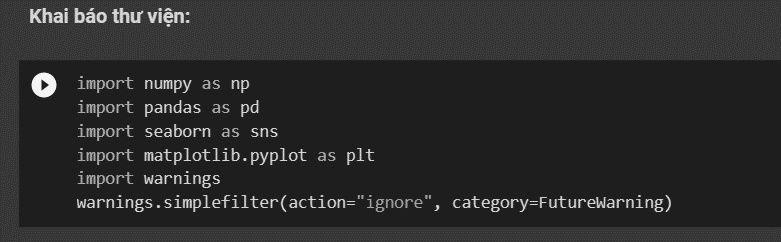
- Matplotlib có một giao diện thủ tục được đặt tên là Pylab, được thiết kế giống với MATLAB, một ngôn ngữ lập trình độc quyền được phát triển bởi MathWorks. Matplotlib cùng với NumPy có thể được coi là mã nguồn mở tương đương với MATLAB.

- Matplotlib ban đầu được viết bởi John D. Hunter vào năm 2003. Phiên bản ổn định hiện tại là 2.2.0 được phát hành vào tháng 1 năm 2018.

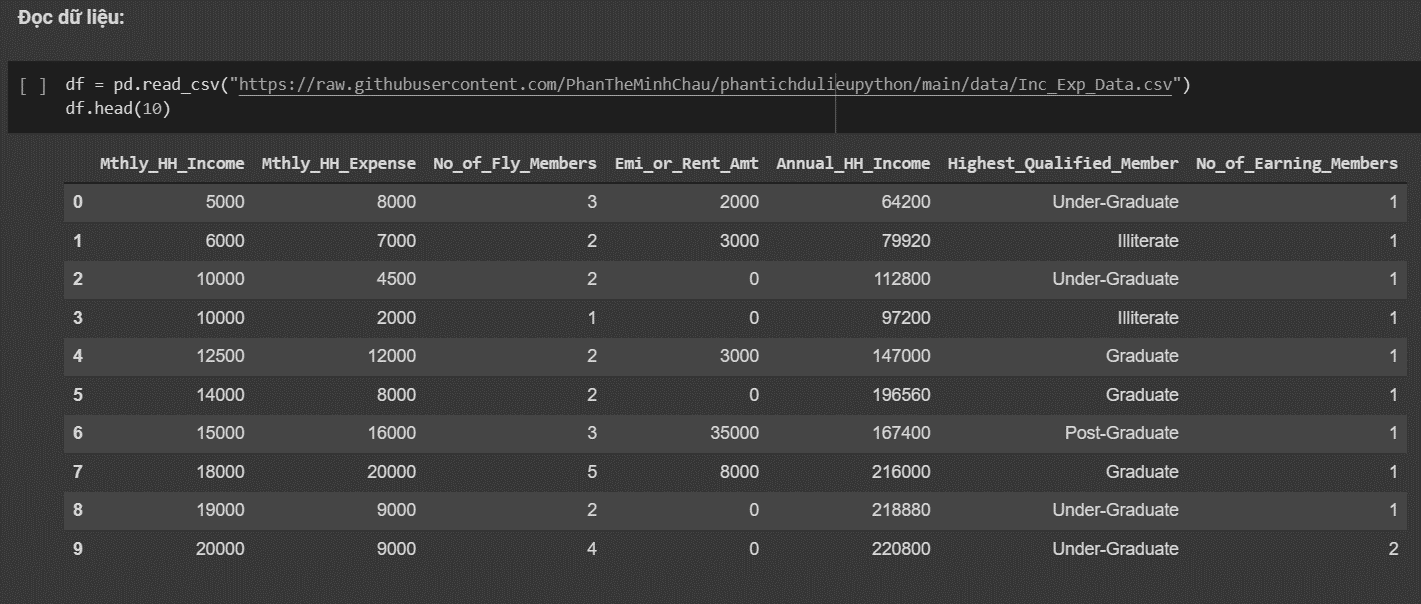
# Chương 2: Phân tích dữ liệu

## 2.1 Thống kê mô tả

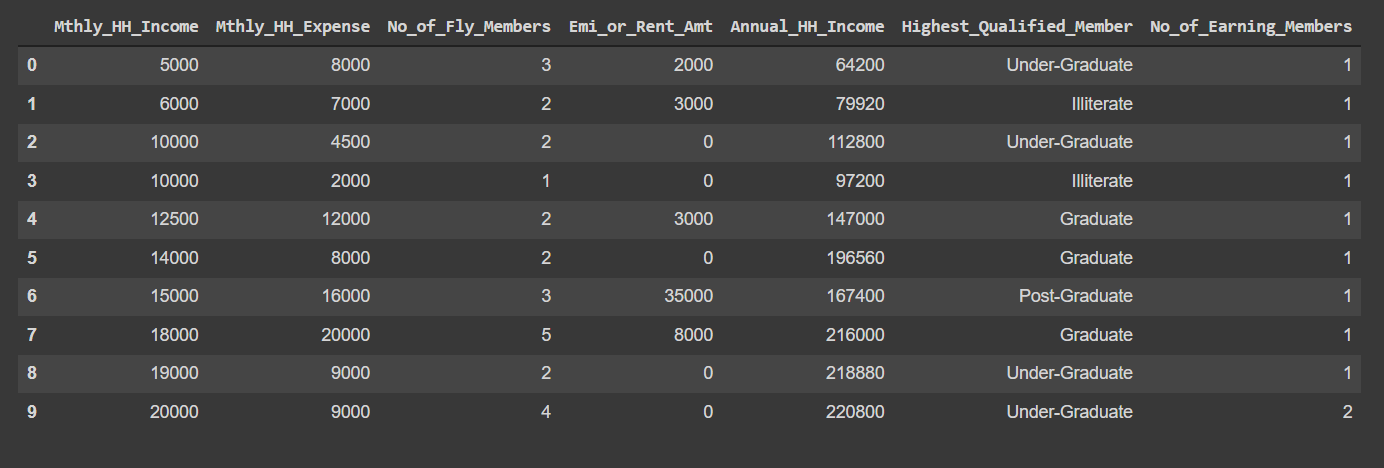
- Khai báo các thư viện cần thiết:



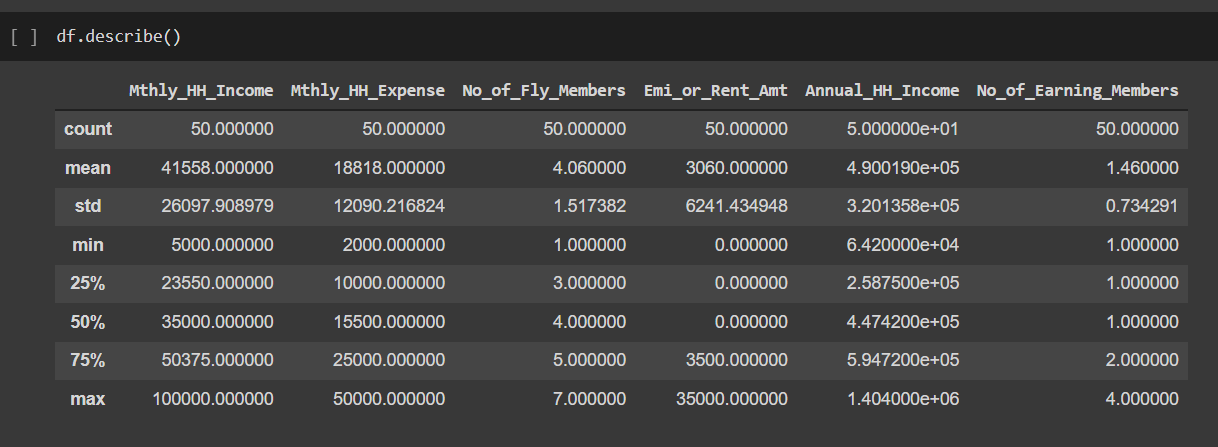
- Đầu tiên truyền file dữ liệu vào Python bằng cách dùng **read\_csv** của thư viên Pandas.



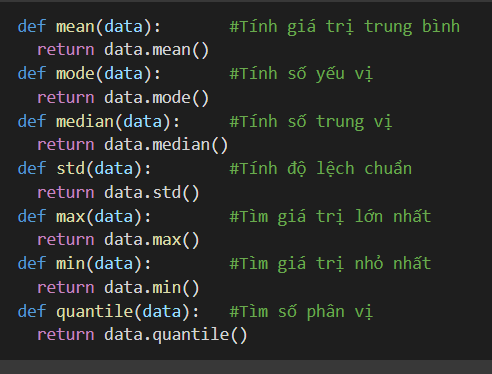
- Sau khi truyền dữ liệu vào chúng ta được bảng dữ liệu bên dưới:

**Phân tích dữ liệu:**

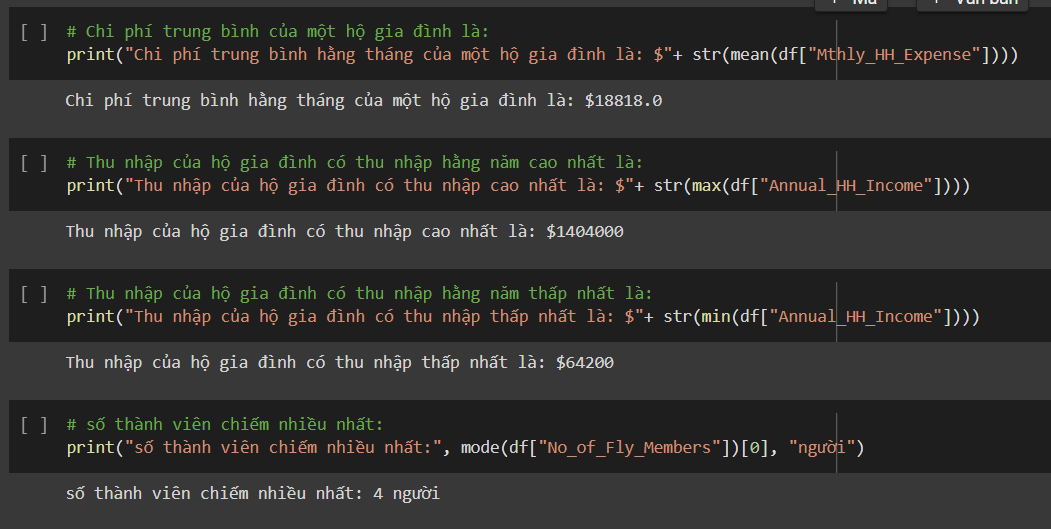
- Tính toán bao quát các số liệu cơ bản ta dùng hàm **describe()**



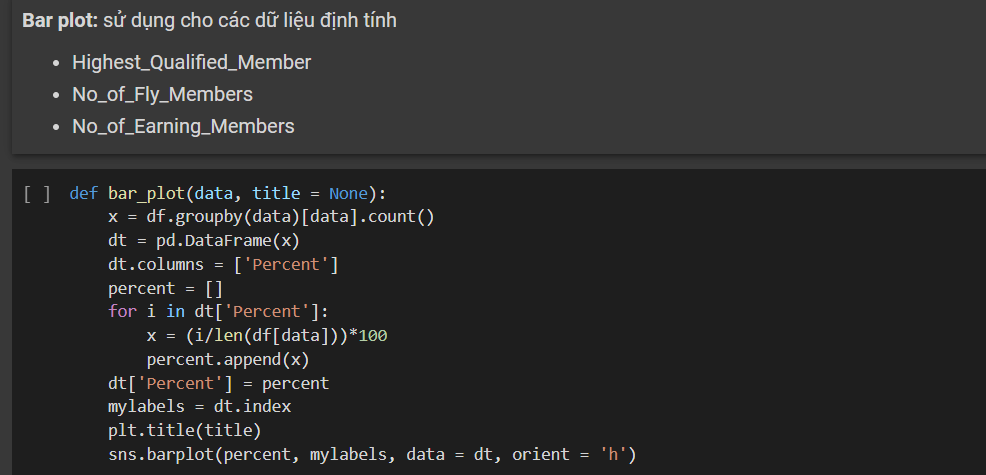
- Ta cũng có thể sử dụng các hàm để tính toán cho các phần tử độc lập:

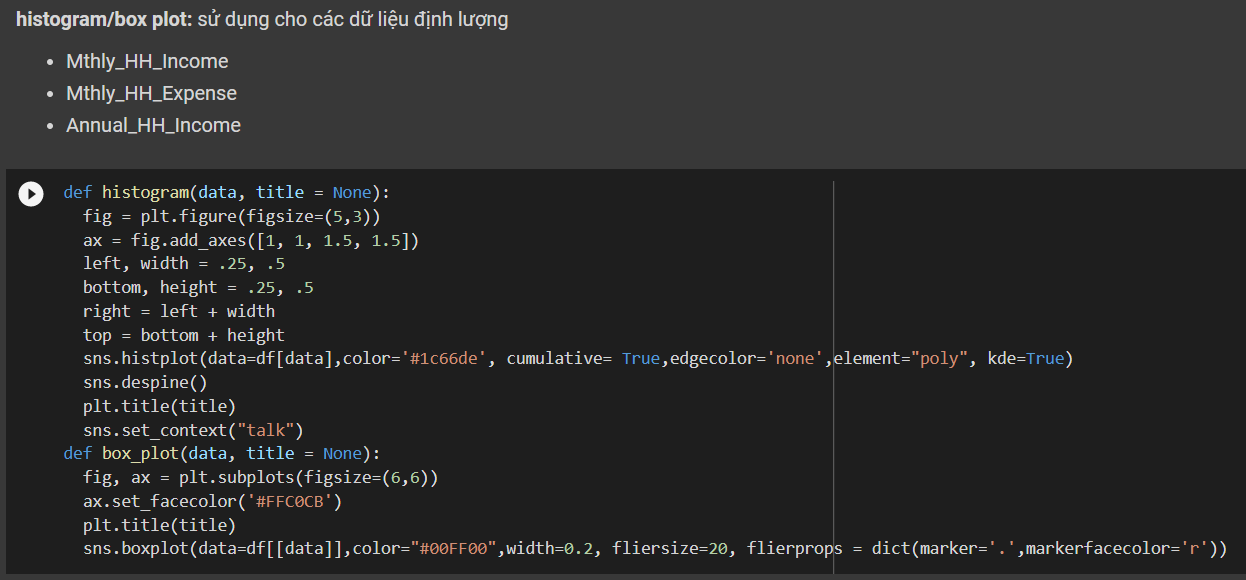


- Ví dụ:

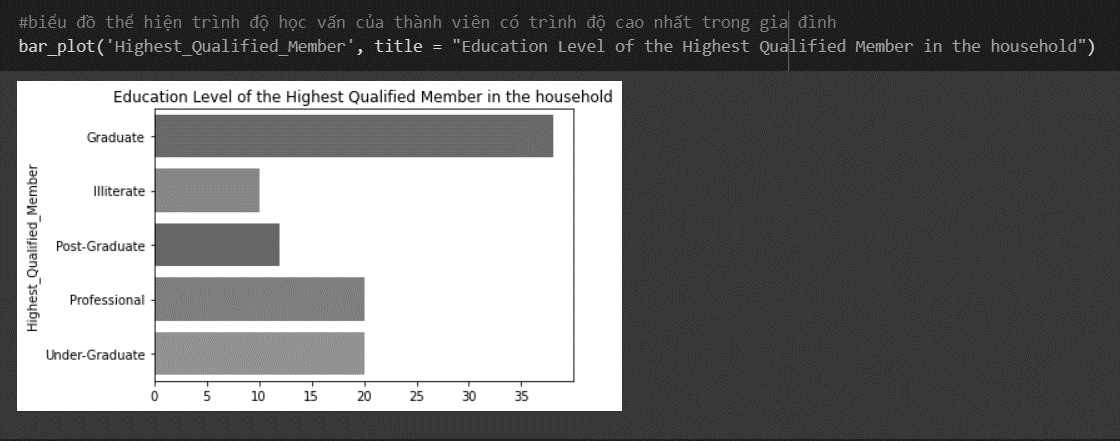


- **Sử dụng biểu đồ:**

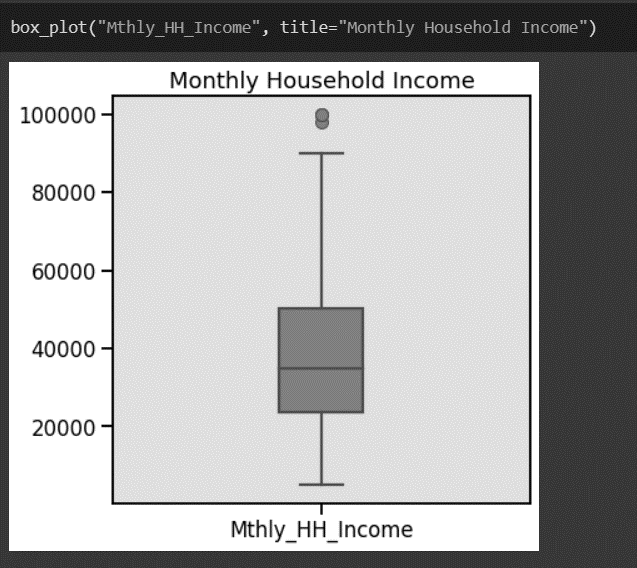
****

****

**-** kết quả:

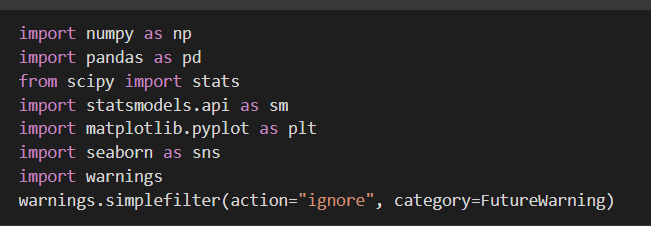




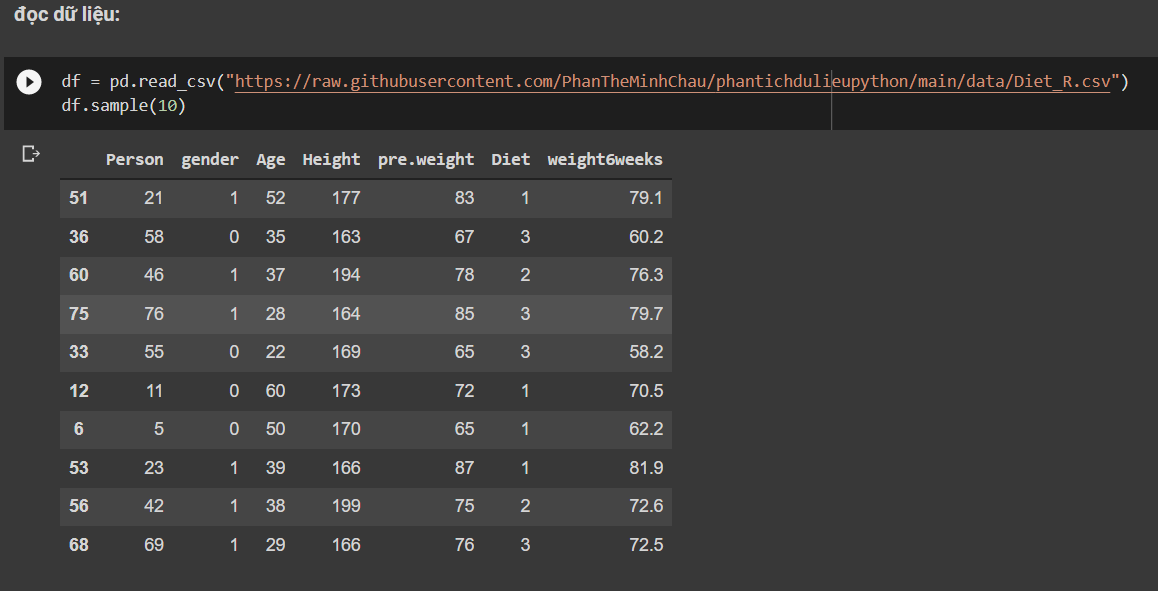


## 2.2 Thống kê suy diễn

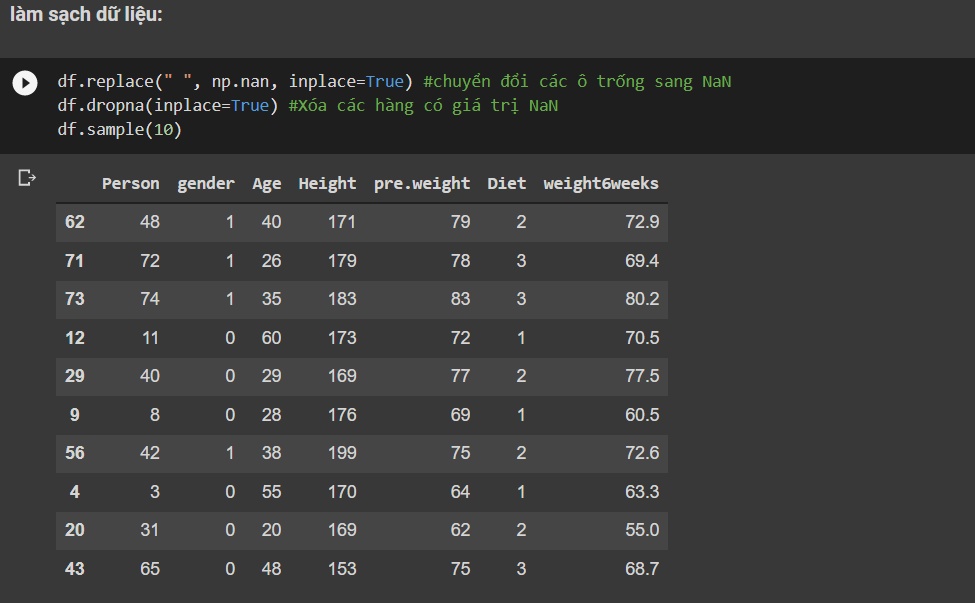
- Khai báo thư viện cần dùng:



- Giống như phần trên chúng ta cũng khai báo các thư viện và sử dụng hàm **read\_csv** để load dữ liệu vào Python.



- Làm sạch dữ liệu:

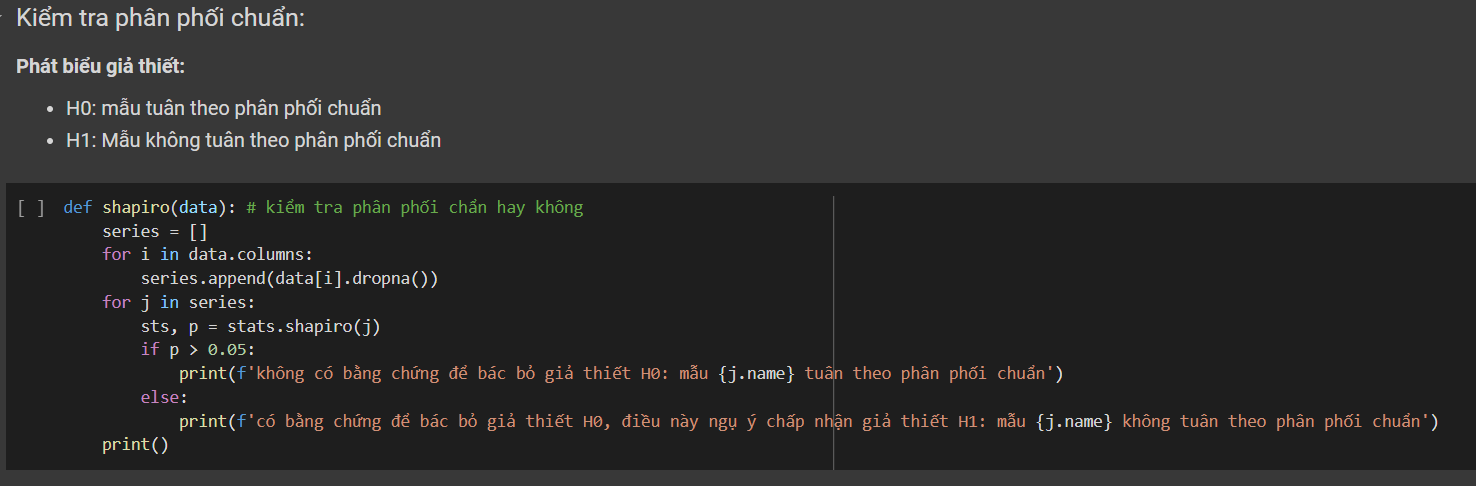


### Quy trình kiểm định ANOVA:

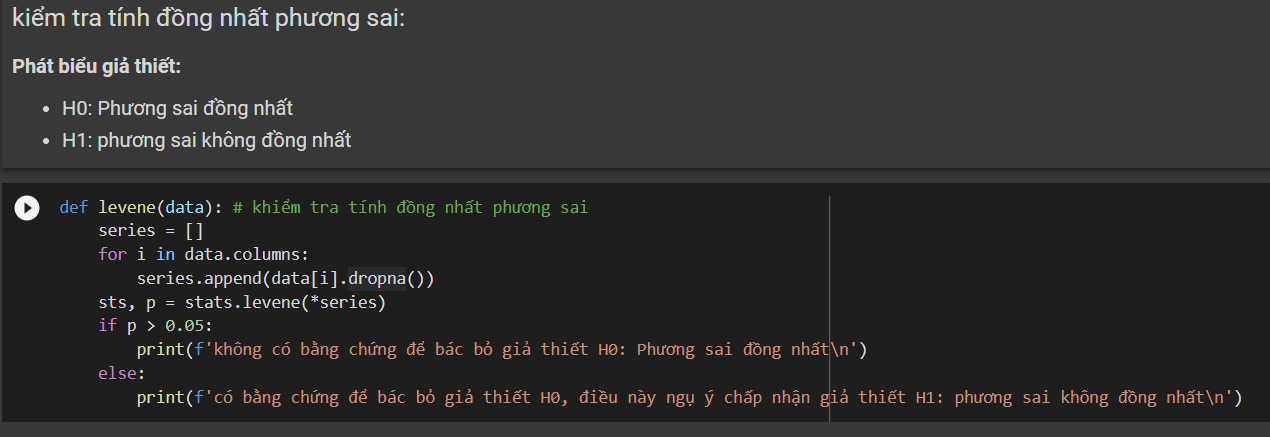
**Một số giả định** khi phân tích ANOVA:

– Các nhóm so sánh phải độc lập và được chọn một cách ngẫu nhiên.

– Các nhóm so sánh phải có phân phối chuẩn or cỡ mẫu phải đủ lớn để được xem như tiệm cận phân phối chuẩn (kiểm tra bằng shapiro hoặc qua quan sát đồ thị)



– Phương sai của các nhóm so sánh phải đồng nhất (kiểm tra bằng leneve)



Lưu ý: nếu giả định tổng thể có phân phối chuẩn với phương sai bằng nhau không đáp ứng được thì bạn có thể dùng kiểm định phi tham số Kruskal-Wallis sẽ để thay thế cho ANOVA.

***Kết quả kiểm định*** gồm hai phần:

Phần 1:

Levene test: dùng kiểm định phương sai bằng nhau hay không giữa các nhóm

Ho: “Phương sai bằng nhau”

Sig <= 0.05: bác bỏ Ho

Sig >0.05: chấp nhận Ho -> đủ điều kiện để phân tích tiếp anova

Phần 2:

ANOVA test: Kiểm định anova

Ho: “Trung bình bằng nhau”

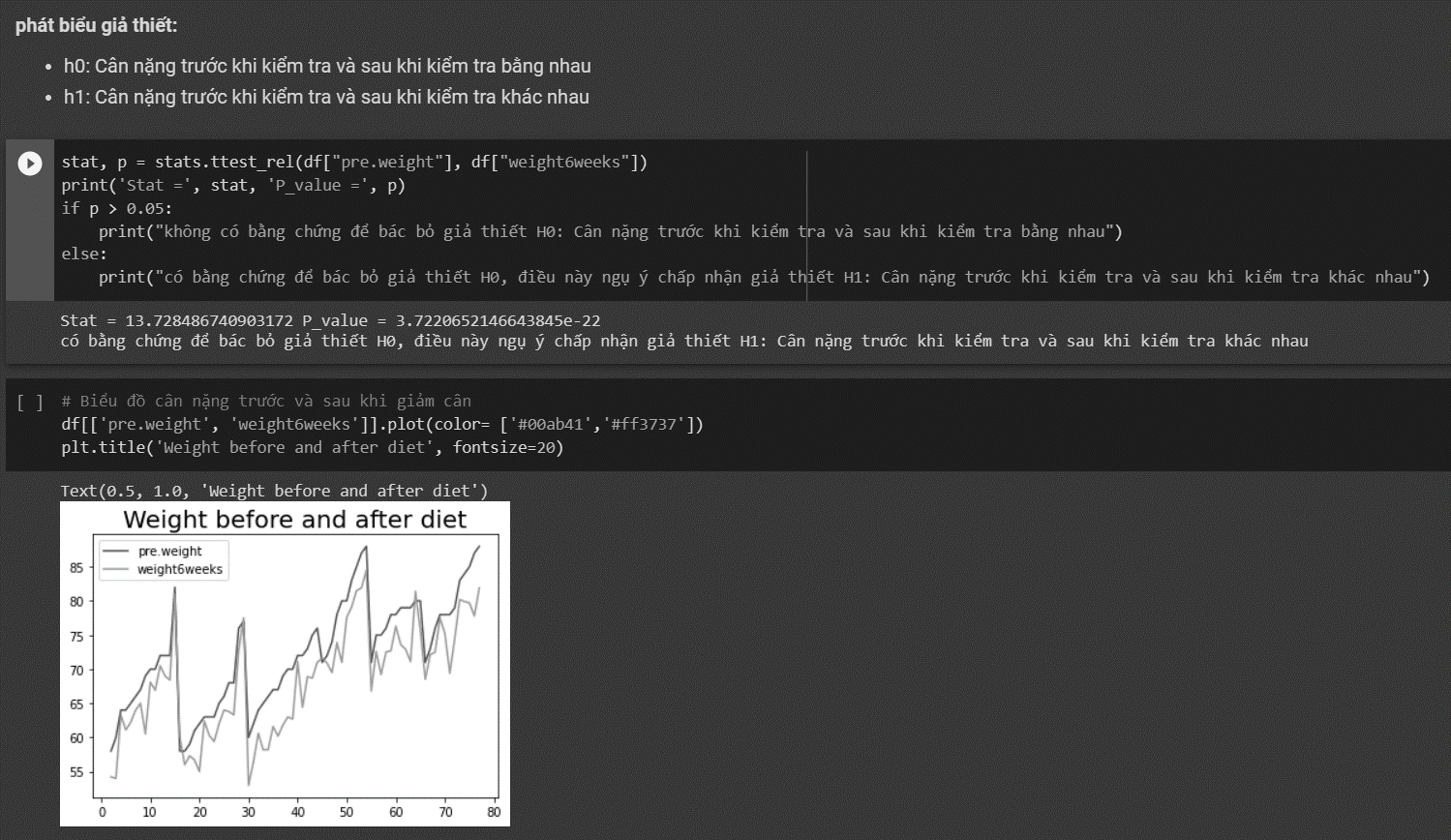
Sig <=0.05: bác bỏ Ho ->  đủ điều kiện để khẳng định có sự khác biệt giữa các nhóm đối với biến phụ thuộc

Sig >0.05: chấp nhận Ho -> chưa đủ điều kiện để khẳng định có sự khác biệt giữa các nhóm đối với biến phụ thuộc

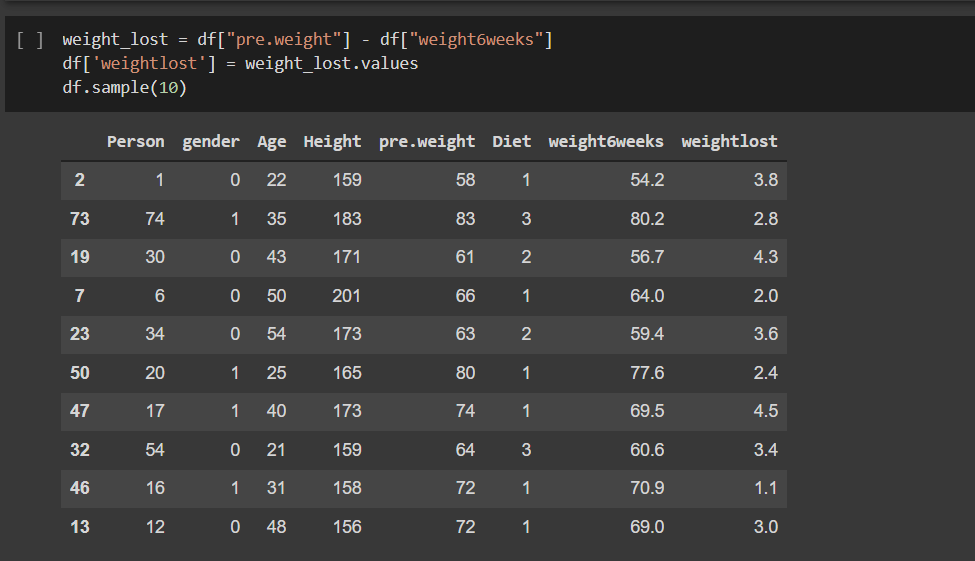
Khi có sự khác biệt thì có thể phân tích sâu hơn để tìm ra sự khác biệt như thế nào giữa các nhóm quan sát bằng các kiểm định Tukey, LSD, Bonferroni, Dunca. Kiểm định sâu anova gọi là kiểm định Post-Hoc.

**Phát biểu câu hỏi nghiên cứu (research question):**

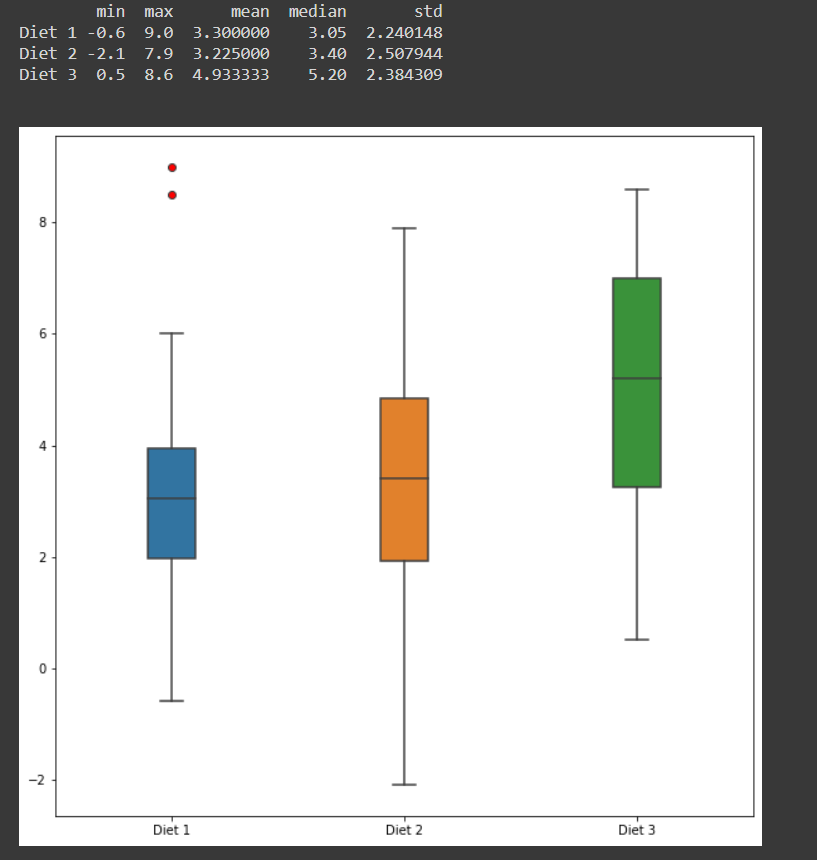
**- Ignore diet and test to see if weight has been lost (Bỏ qua chế độ ăn kiêng và kiểm tra xem đã giảm cân chưa?):**

****

- Remove weight lost and get students to calculate it using before/after weights:

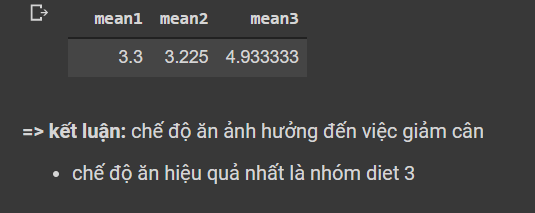


- Summary statistics by diet:



**- Kiểm tra chế độ ăn:** phân nhóm theo chế độ ăn rồi kiểm định anova;

-> kết quả:



- **sự ảnh hưởng của giới tính:**

****

# Chương 3: Kiểm tra đạo văn



# Chương 4: Tài liệu tham khảo

* <https://pandas.pydata.org/docs/>
* <https://www.scipy.org/docs.html>
* <https://matplotlib.org/>
* <https://stackoverflow.com/>