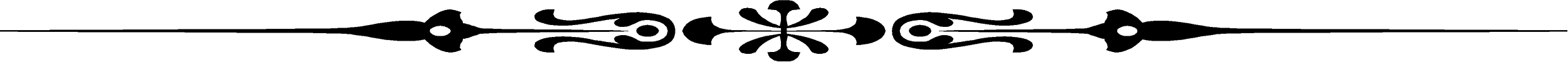
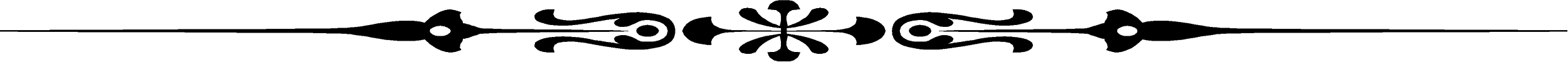
**ĐẠI HỌC HUẾ**



KHOA KỸ THUẬT VÀ CÔNG NGHỆ

🙠🙟🕮🙝🙢

****

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2021-2022**

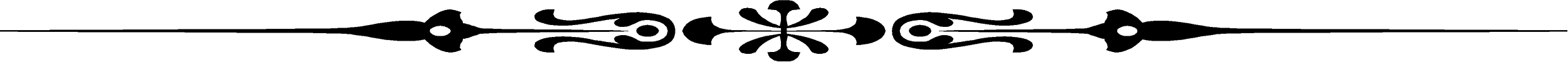
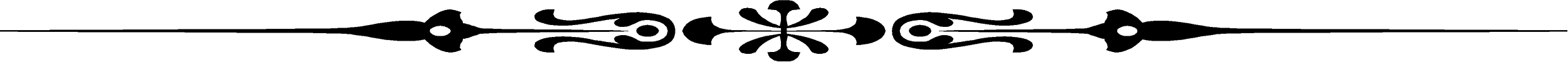
**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: KHDL&TTNT**

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

**Thừa Thiên Huế, ngày 30 tháng 5 năm 2022**

**ĐẠI HỌC HUẾ**



KHOA KỸ THUẬT VÀ CÔNG NGHỆ

🙠🙟🕮🙝🙢

****

(MẪU BÌA PHỤ)

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2021-2022**

**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: KHDL&TTNT**

**Sinh viên thực hiện: Hoàng Nữ Thu Phương**

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

**Thừa Thiên Huế, ngày 30 tháng 5 năm 2022**

ĐẠI HỌC HUẾ

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

**PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN**

**Học kỳ II, năm học 2021 - 2022**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT1:**  Bằng số: .........................................  Bằng chữ: ....................................... | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT2:**  Bằng số: .........................................  Bằng chữ: ....................................... |

Điểm kết luận: ...........................................................................................................

Bằng số:.....................................................................................................................

Bằng chữ: .................................................................................................................

Thừa Thiên Huế, ngày tháng năm 2022

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1**  (Ký và ghi rõ họ và tên) | **Cán bộ chấm thi 2**  (Ký và ghi rõ họ và tên) |

Mục lục

[Chương 1: Cơ sở lí thuyết 5](#_Toc105094824)

[1.Thống kê là gì? 5](#_Toc105094825)

[1.1. Định Nghĩa 5](#_Toc105094826)

[2. Thống kê trong Python: 5](#_Toc105094827)

[2.1. Giới thiệu về ngôn ngữ lập trình python: 5](#_Toc105094828)

[2.2. Giới thiệu các thư viện: 5](#_Toc105094829)

[Chương 2. Phân tích dữ liệu 7](#_Toc105094830)

[1. Thống kê mô tả 7](#_Toc105094831)

[1.1. Tính toán các thông số thống kê của tập dữ liệu định lượng: 9](#_Toc105094832)

[1.2. Tính độ tương quan giữa các cột dữ liệu định lượng 9](#_Toc105094833)

[1.3. Phân tích thống kê mô tả với tập dữ liệu “Mthly\_HH\_Income Variable” - Thu nhập theo hộ gia đình hàng tháng 11](#_Toc105094834)

[1.4. Phân tích thống kê mô tả với tập dữ liệu “Mthly\_HH\_Expense Variable” - Chi tiêu hộ gia đình hàng tháng 13](#_Toc105094835)

[1.5. Phân tích thống kê mô tả với tập dữ liệu “Highest\_Qualified\_Member Variable” - Trình độ học vấn cao nhất 15](#_Toc105094836)

[1.6. Phân tích Mthly\_HH\_Income và Mthly\_HH\_Expense 17](#_Toc105094837)

[1.7. Phân tích mức thu nhập với trình độ học vấn 20](#_Toc105094838)

[1.8. Phân tích mức độ phân tán của dữ liệu 21](#_Toc105094839)

[2.Thống kê suy diễn: 21](#_Toc105094840)

[2.1. Paired t-test 24](#_Toc105094841)

[2.2. Compute variable 26](#_Toc105094842)

[2.3. Summarry statistics 27](#_Toc105094843)

[2.4. One\_way ANOVA 2](#_Toc105094843)9

[2.5. Interactions 38](#_Toc105094843)

[Chương 3: Tài liệu tham khảo 39](#_Toc105094844)

[Chương 4: Kiểm tra đạo văn 39](#_Toc105094844)

# Chương 1: Cơ sở lí thuyết

## 1.Thống kê là gì?

### 1.1. Định Nghĩa

**Thống kê(Statistics)** là một dạng phân tích toán học sử dụng mô hình biểu diễn và tóm tắt định lượng cho tập dữ liệu nhất định. Thống kê nghiên cứu các phương pháp để thu thập, phân tích và rút ra kết luận từ tập dữ liệu. Một số biện pháp thống kê bao gồm:

#### Thống kê mô tả:

Thống kê mô tả được dùng để mô tả những đặc tính cơ bản của tập dữ liệu đã được thu thập từ các cách thức khác nhau. Thống kê mô tả và [thống kê suy](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_suy_lu%E1%BA%ADn) diễn cung cấp những tóm tắt đơn giản về [mẫu](https://vi.wikipedia.org/wiki/M%E1%BA%ABu_(th%E1%BB%91ng_k%C3%AA)) và các thước đo. Có rất nhiều kỹ thuật hay được sử dụng như sau:

•Biểu diễn dữ liệu bằng đồ họa trong đó các đồ thị mô tả dữ liệu hoặc giúp so sánh dữ liệu;

•Biểu diễn dữ liệu thành các bảng số liệu tóm tắt về dữ liệu;

•Thống kê tóm tắt (dưới dạng các giá trị thống kê đơn nhất) mô tả dữ liệu.

#### Thống kê suy diễn:

- Thống kê suy diễn là công cụ để rút ra kết luận về các đặc điểm của một quần thể từ các đặc điểm của mẫu và quyết định mức độ chắc chắn của chúng về độ tin cậy của các kết luận đó.

## 2. Thống kê trong Python:

### 2.1. Giới thiệu về ngôn ngữ lập trình python:

**Python** là ngôn ngữ lập trình bậc cao phục vụ cho những mục đích lập trình đa năng, do Guido van Rossum tạo ra và ra mắt vào năm 1991.

Python có ưu điểm mạnh là dễ đọc, dễ học và dễ nhớ. Python là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rất rõ ràng, rất thuận tiện cho người mới học lập trình.

Cấu trúc của Python còn có thể cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu.

### 2.2. Giới thiệu các thư viện:

#### Thư viện Pandas

- Pandas là một Thư viện Python mã nguồn mở, cung cấp công cụ phân tích và thao tác dữ liệu với hiệu suất cao bằng cách sử dụng các cấu trúc dữ liệu của nó.

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

- Các tính năng của Pandas:

* Cung cấp các công cụ để tải dữ liệu vào các đối tượng dữ liệu trong bộ nhớ từ các định dạng tệp khác nhau.
* Giúp căn chỉnh dữ liệu và tích hợp xử lý dữ liệu bị thiếu.
* 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.
* .....

#### Thư viên Scipy

- SciPy là một thư viện mã nguồn mở cảu Python cho toán học, khoa học và kỹ thuật.

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

#### Thư viện Matplotlyb

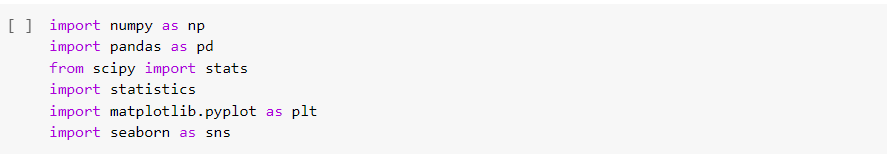
- Matplotlib là một trong những gói Python phổ biến nhất để trực quan hóa dữ liệu. Nó là một thư viện đa nền tảng để tạo các đồ thị 2D từ dữ liệu thu thập được trong các mảng.

- Matplotlib được viết bằng Python và sử dụng NumPy. 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 được dùng trong Python và IPython shell, Jupyter và các máy chủ ứng dụng web.

# Chương 2. Phân tích dữ liệu

## 1. Thống kê mô tả

- Đầu tiên ta import các thư viện:

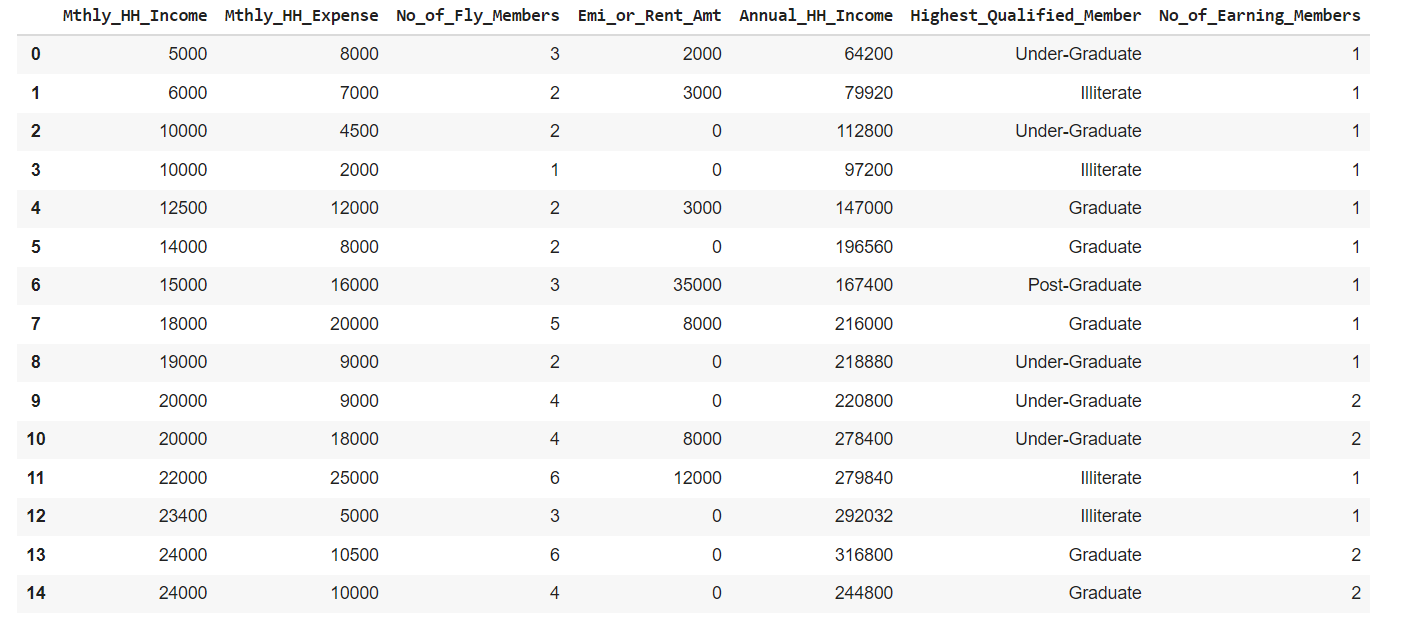


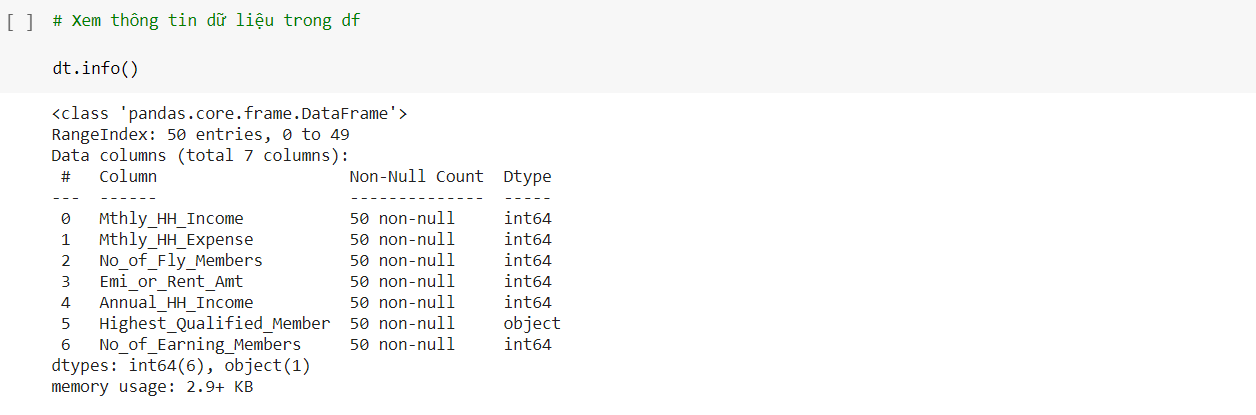
- Trong phần này, ta sử dụng tập dữ liệu Inc\_Exp\_Data.csv để phân tích thống kê mô tả.

- Đầu tiên chúng ta truyền file vào bằng hàm read.csv của thư viện pandas:

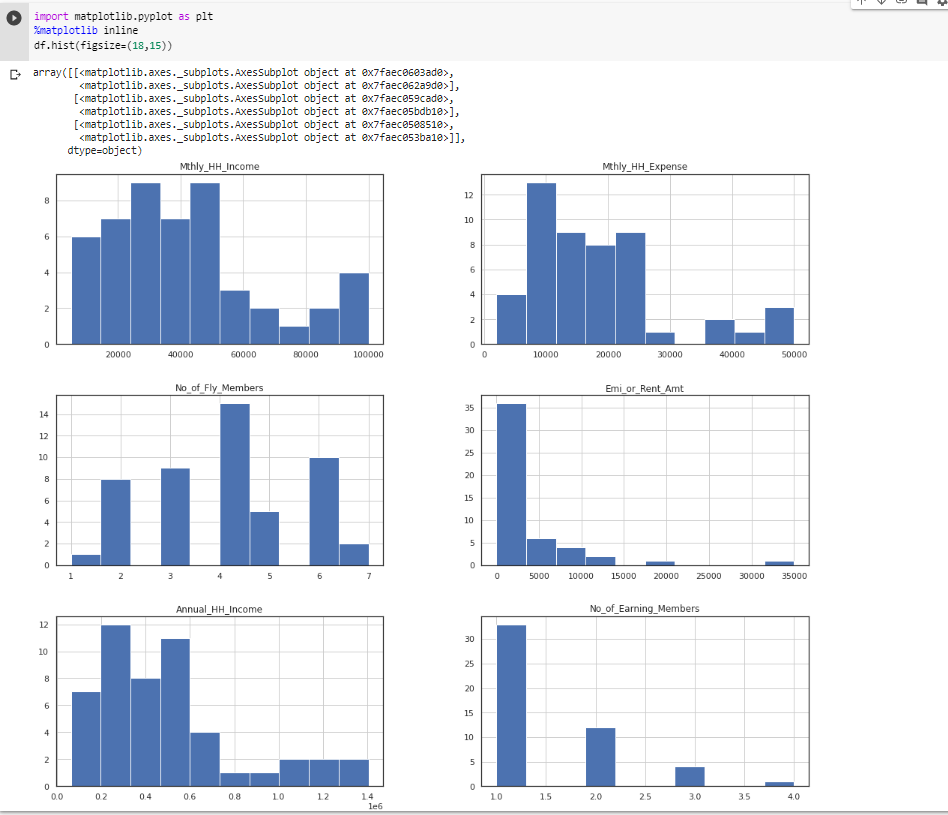


-Sau khi truyền vào ta được data:



Sử dụng hàm info() để xem thông tin của data: 

Sử dụng thư viện matplotlib để xem tổng quan dữ liệu của data:

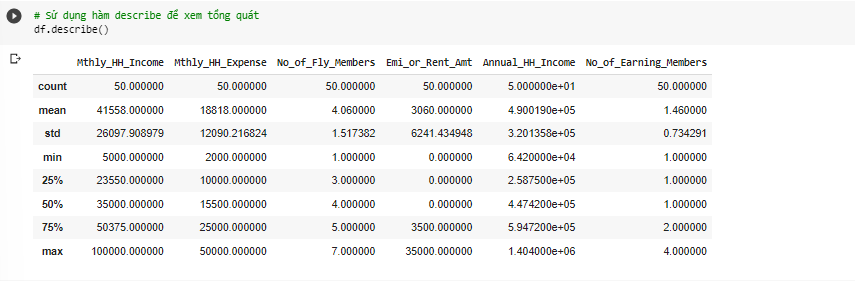


Tiến hành xuất kích thước của tập dữ liệu và xóa bỏ giá trị rỗng (nếu có) 

Nhận xét: Tập dữ liệu ban đầu không có giá trị rỗng (null/NA)

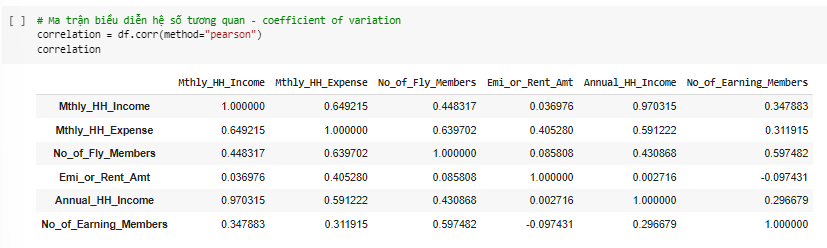
Tiếp theo ta đi vào phân tích dữ liệu.

### Tính toán các thông số thống kê của tập dữ liệu định lượng:



### 1.2. Tính độ tương quan giữa các cột dữ liệu định lượng

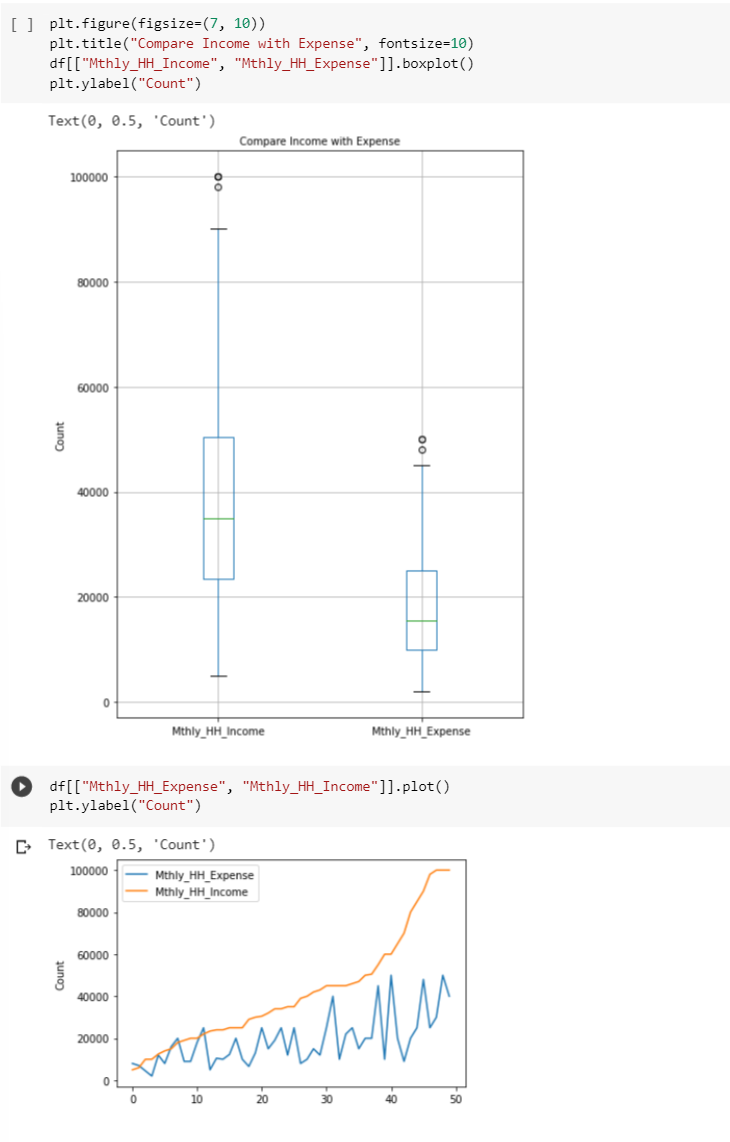
- Đầu tiên ta biểu diễn ma trận hệ số tương quan



- Mô tả bằng Heat map



- Biểu điễn tương quan giữa thu nhập và chi tiêu hàng tháng

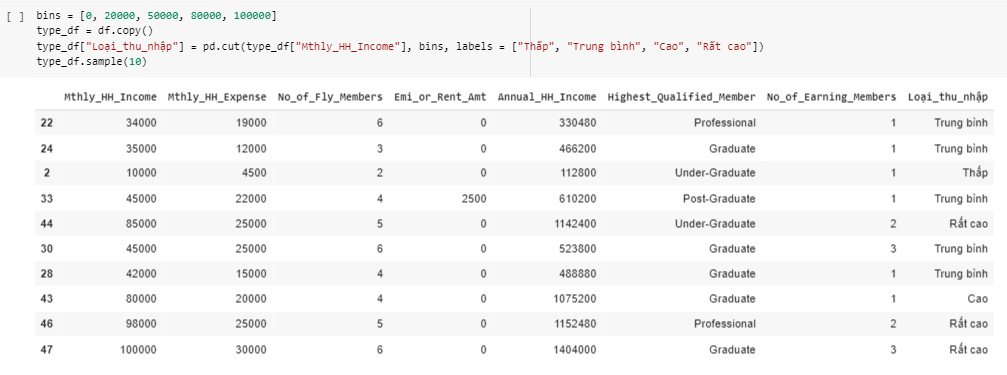


### 1.3. Phân tích thống kê mô tả với tập dữ liệu “Mthly\_HH\_Income Variable” - Thu nhập theo hộ gia đình hàng tháng



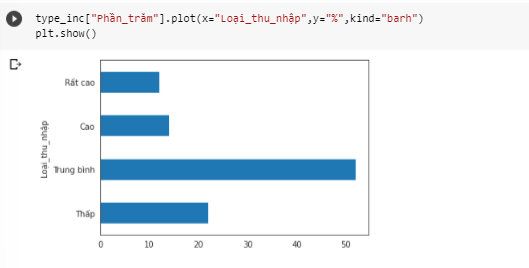
- Để thuận tiện cho việc phân tích dữ liệu, ta thêm vào một cột có tên là Loại\_thu\_nhập với các phân loại như sau:

* Rất cao >= 80000$
* Cao >= 50000$
* Trung bình >= 20000$
* Thấp < 20000$



Bảng tỉ lệ và biểu đồ tương ứng:







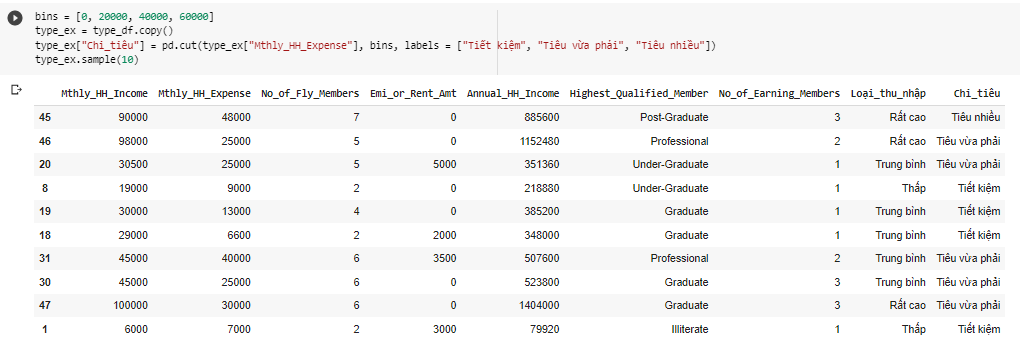
Nhận xét: tỉ lệ người có thu nhập trung bình cao nhất.

### 1.4. Phân tích thống kê mô tả với tập dữ liệu “Mthly\_HH\_Expense Variable” - Chi tiêu hộ gia đình hàng tháng

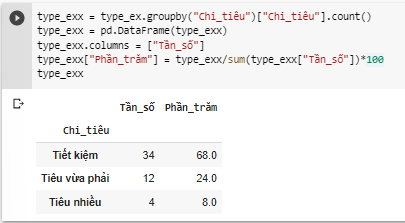


- Để thuận tiện cho việc phân tích dữ liệu, ta thêm vào một cột có tên là Chi\_tiêu với các phân loại như sau:

* Tiêu nhiều >= 40000$
* Tiêu vừa phải >= 20000$
* Tiết kiệm < 20000$



Bảng tỉ lệ và biểu đồ tương ứng:



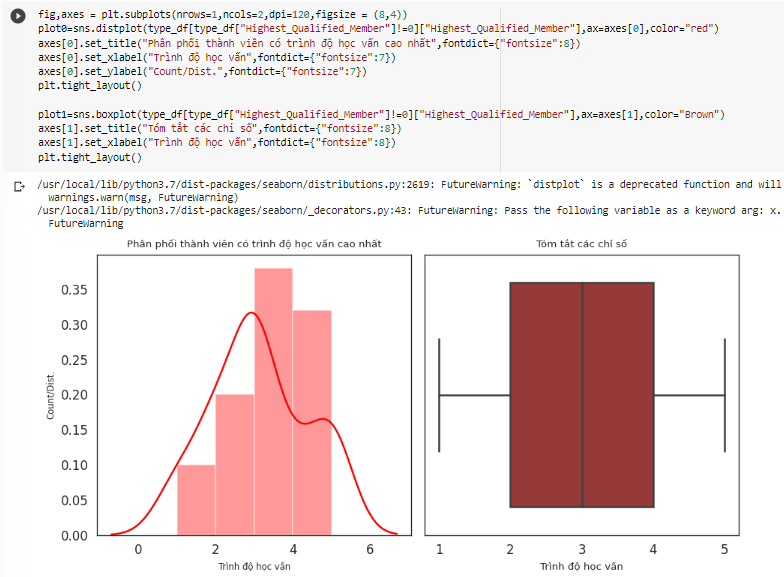


Nhận xét: Chi tiêu tiết kiệm chiếm tỉ lệ cao nhất

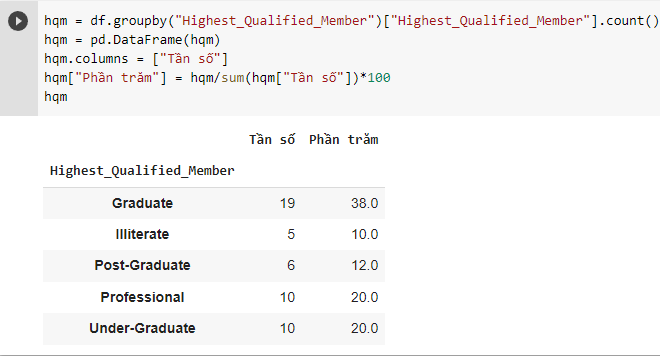
### 1.5. Phân tích thống kê mô tả với tập dữ liệu “Highest\_Qualified\_Member Variable” - Trình độ học vấn cao nhất

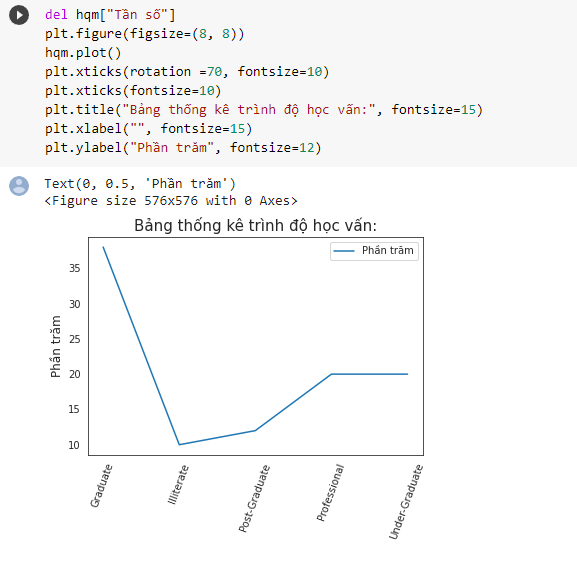
-Đầu tiên để có thế tiến hành thống kê ta phải chuyển đổi các giá trị về dạng số tưởng ứng như sau:

* Illiterate = 1
* Under-Graduate = 2
* Graduate = 3
* Post-Graduate = 4
* Professional = 5

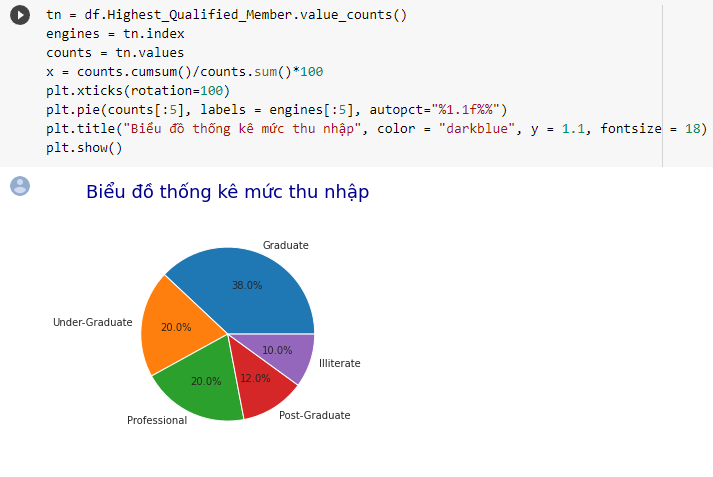


Bảng tỉ lệ và biểu đồ tương ứng





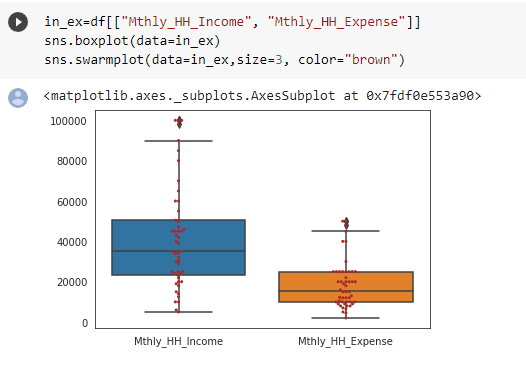
Biểu đồ thống kê mức thu nhập phụ thuộc vào trình độ học vấn



Nhận xét: - Trình độ Graduate chiếm tỉ lệ kiếm được thu nhập cao nhất

- Trình độ Illiterate chiếm tỉ lệ kiếm được thu nhập ít nhất

### 1.6. Phân tích Mthly\_HH\_Income và Mthly\_HH\_Expense

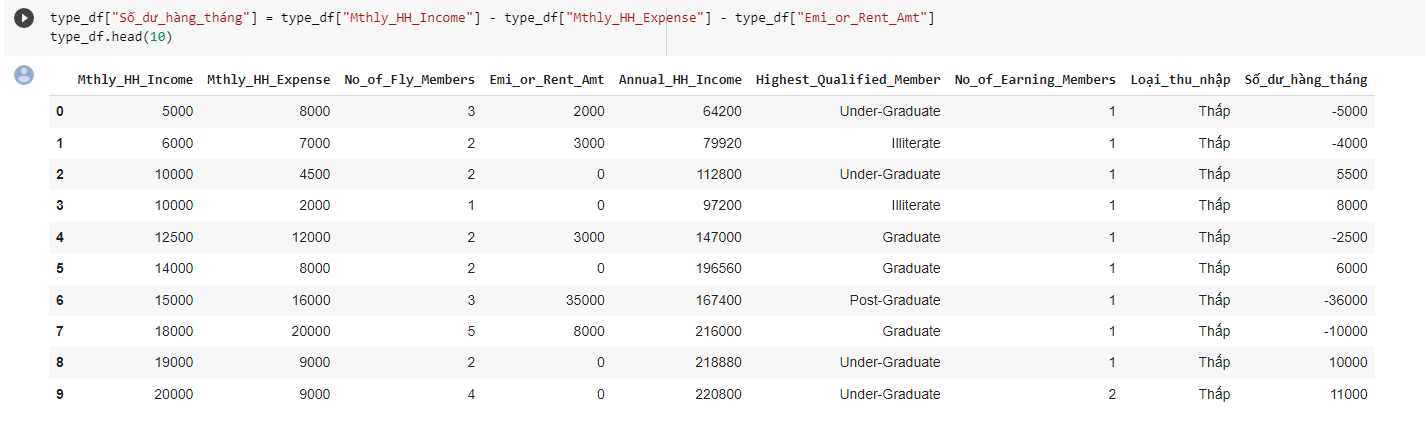


Tính số dư hang tháng để tiến hành phần tích.về thu chi và để xem xét tình trạng kinh tế của gia đình đó.

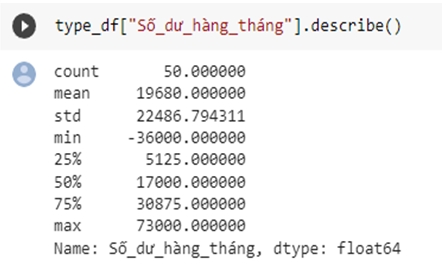
* **Tính số dư hàng tháng**

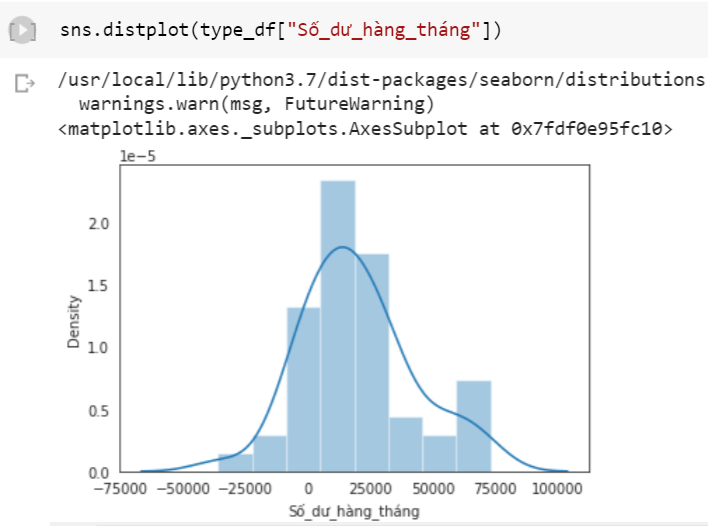
Số dư = thu nhập hàng tháng - chi tiêu hàng tháng - thuế hàng tháng

= (Mthly\_HH\_Income) - (Mthly\_HH\_Expense) - (Emi\_or\_Rent\_Amt)

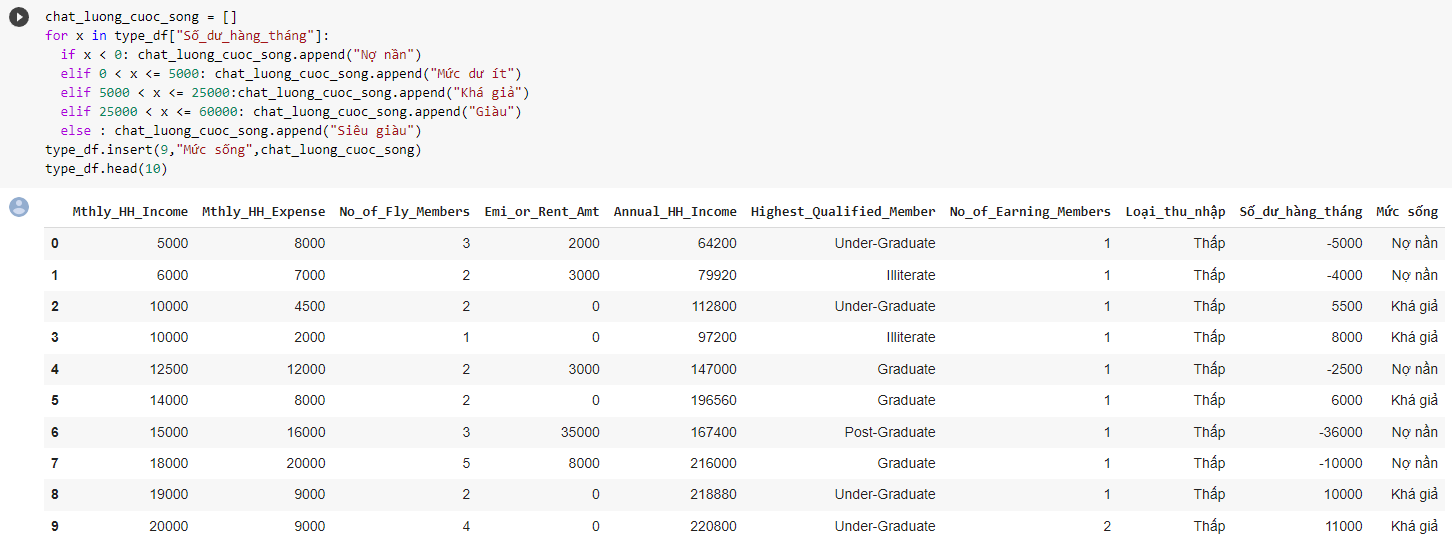


Các thông số của dữ liệu:



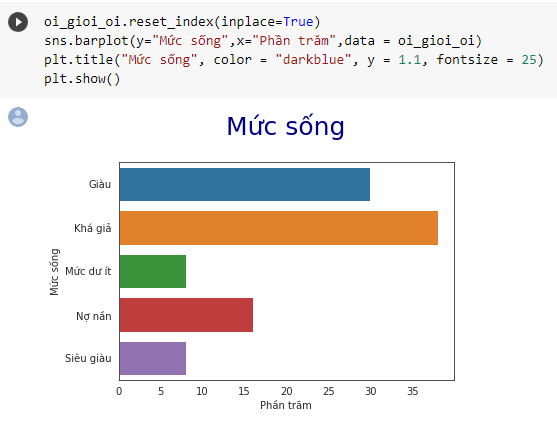


Xây dựng hàm chat\_luong\_cuoc\_song để chia dữ liệu số dư ra làm các nhóm và thêm vào dataframe cột “Mức sống” phản ánh tình trạng kinh tế của từng hộ gia đình.

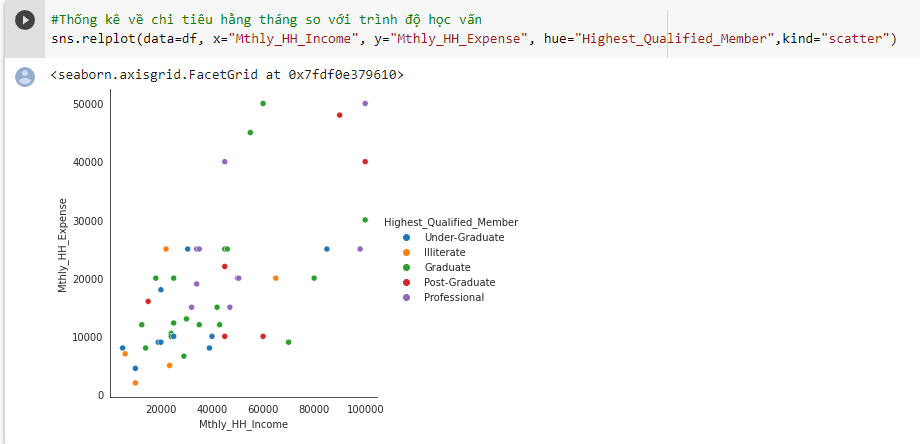


Bảng tần số và biểu đồ minh họa

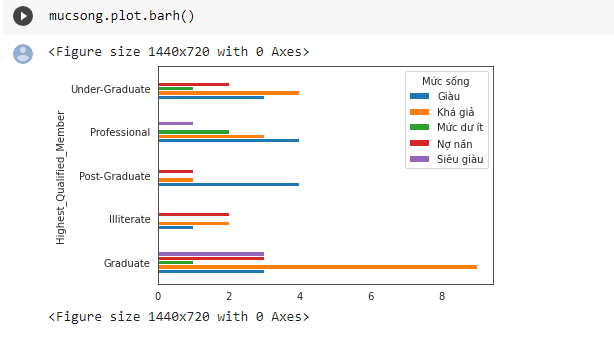




### 1.7. Phân tích mức thu nhập với trình độ học vấn





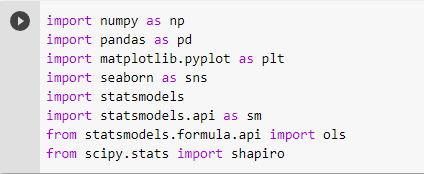


### 1.8. Phân tích mức độ phân tán của dữ liệu

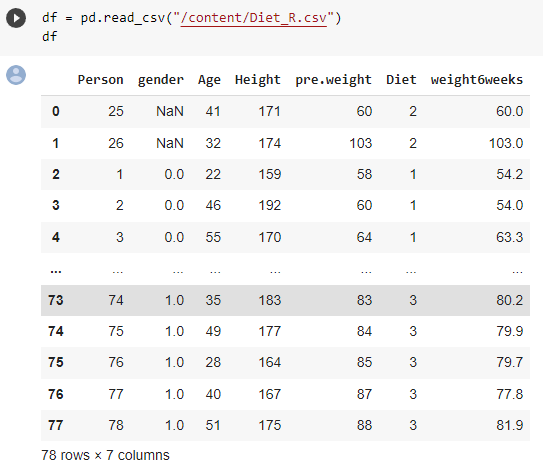
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |

## 2.Thống kê suy diễn:

Các thư viện cần thiết

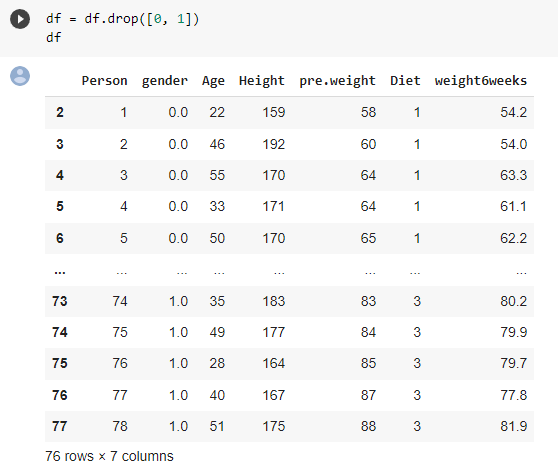


Đọc dữ liệu:

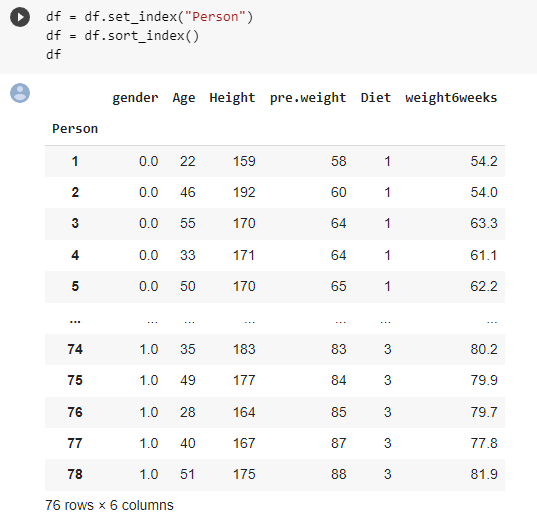




Vì Person 25 và 26 thiếu dữ liệu ở cột gender -> loại bỏ để tiến hành thống kê suy diễn



Sắp xếp lại thứ tự



### 2.1. Paired t-test

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"

**Phát biểu giả thuyết:**

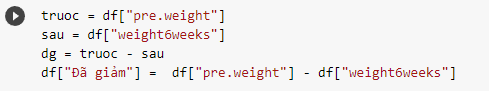
H0 : μ0=μ1.

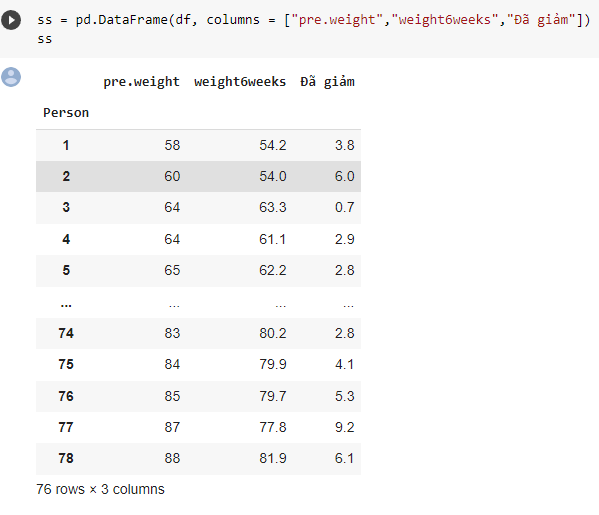
Cân nặng trung bình của trước và sau sáu tuần là như nhau (Không giảm được cân)

Ha :μ0≠μ1.

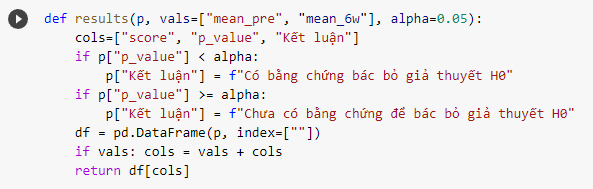
Cân nặng trung bình của trước và sau sáu tuần có sự khác nhau

-Tạo thêm một cột dữ liệu “Đã giảm” để thống kê cân nặng giảm được sau 6 tuần.



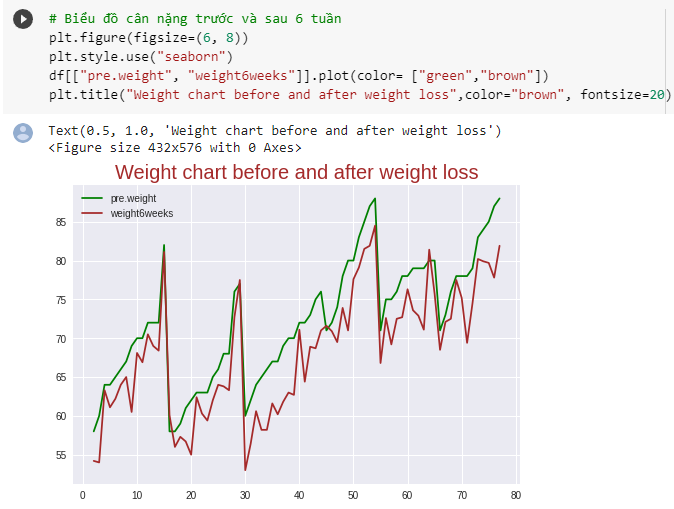


Xây dựng hàm để tiến hành kiểm định t





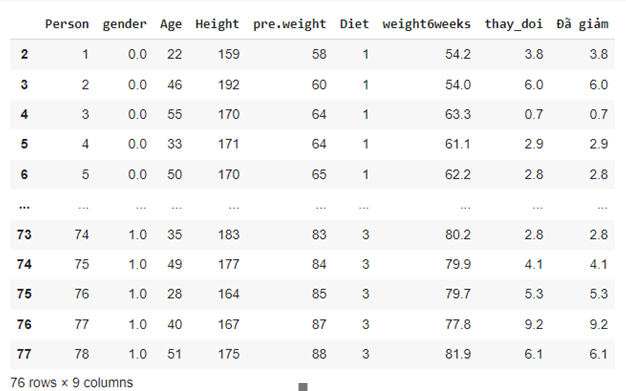
**Kết luận:** Cân nặng trước và sau 6 tuần có sự khác nhau.



### 2.2. Compute variable

Research question: **Remove weight lost and get students to calculate it using before/after weights -** "Loại bỏ trọng lượng đã giảm và hãy tính toán nó bằng cách sử dụng cân nặng trước và sau"

-Hiển thị dữ liệu

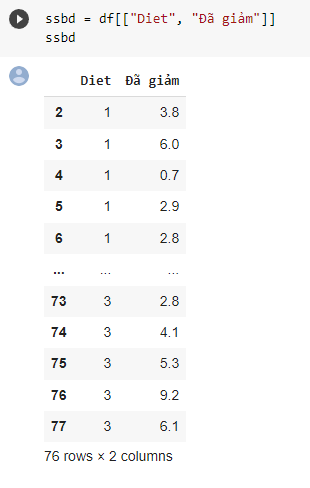


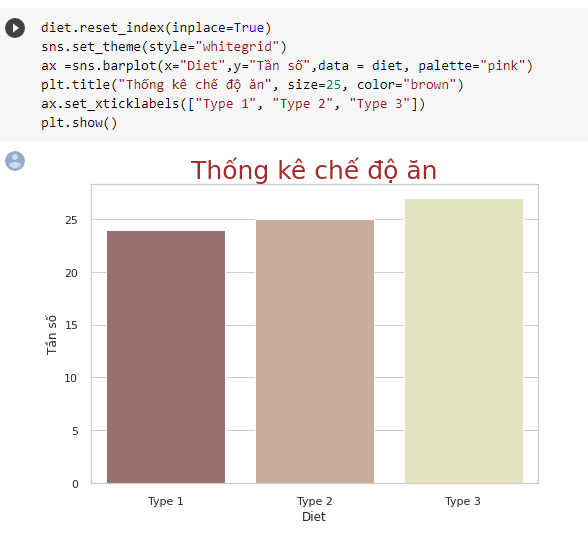
-Đồ thị minh họa



### 2.3. Summarry statistics

Research question:  **Summarry statistics by diet** - "Thống kê tóm tắt theo chế độ ăn kiêng"



Biểu đồ thống kê chế độ ăn:



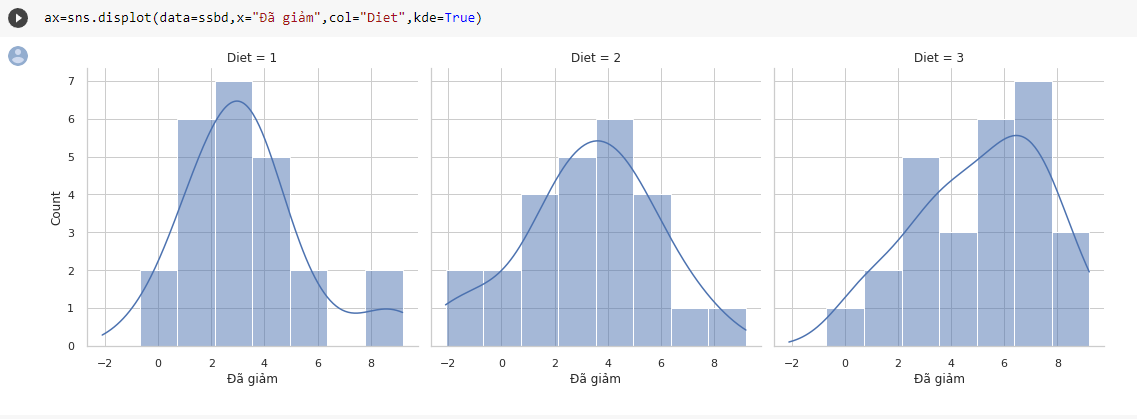
**2.4.One\_way ANOVA**

Research question 1:  **Which diet was best for losing weight? -** "Chế độ ăn kiêng nào là tốt nhất để giảm cân?"

**Xây dựng giả thuyết:**

* H0: Các phương pháp giảm cân cho hiệu quả là như nhau.
* H1: Các phương pháp giảm cân cho hiệu quả khác nhau.

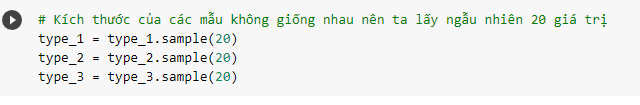
- Để kiểm chứng giả thuyết chúng ta cần trích xuất dữ liệu để phân tích, phân loại dữ liệu và tiến hành phân tích ANOVA

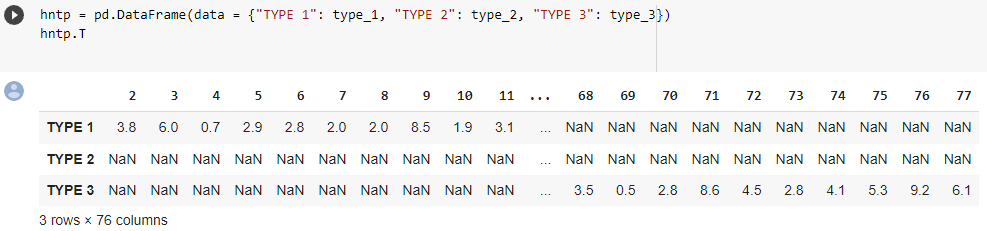


- Chú ý rằng, để thực hiện kiểm định ANOVA thì dữ liệu phải thỏa mãn các yêu cầu sau:

* Dữ liệu phải tuân theo phân phối chuẩn.
* Dữ liệu đồng nhất về phương sai .
* Các cá thể phải được lấy độc lập với nhau.

-Vì kích thước các mẫu không giống nhau nên ta lấy lại mẫu



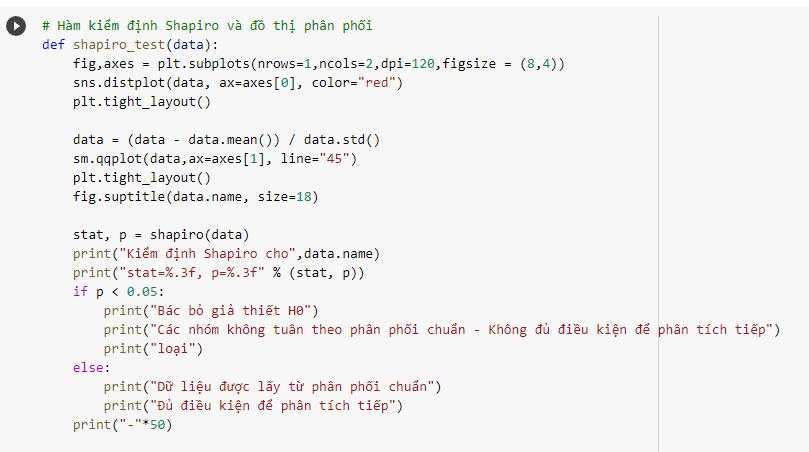


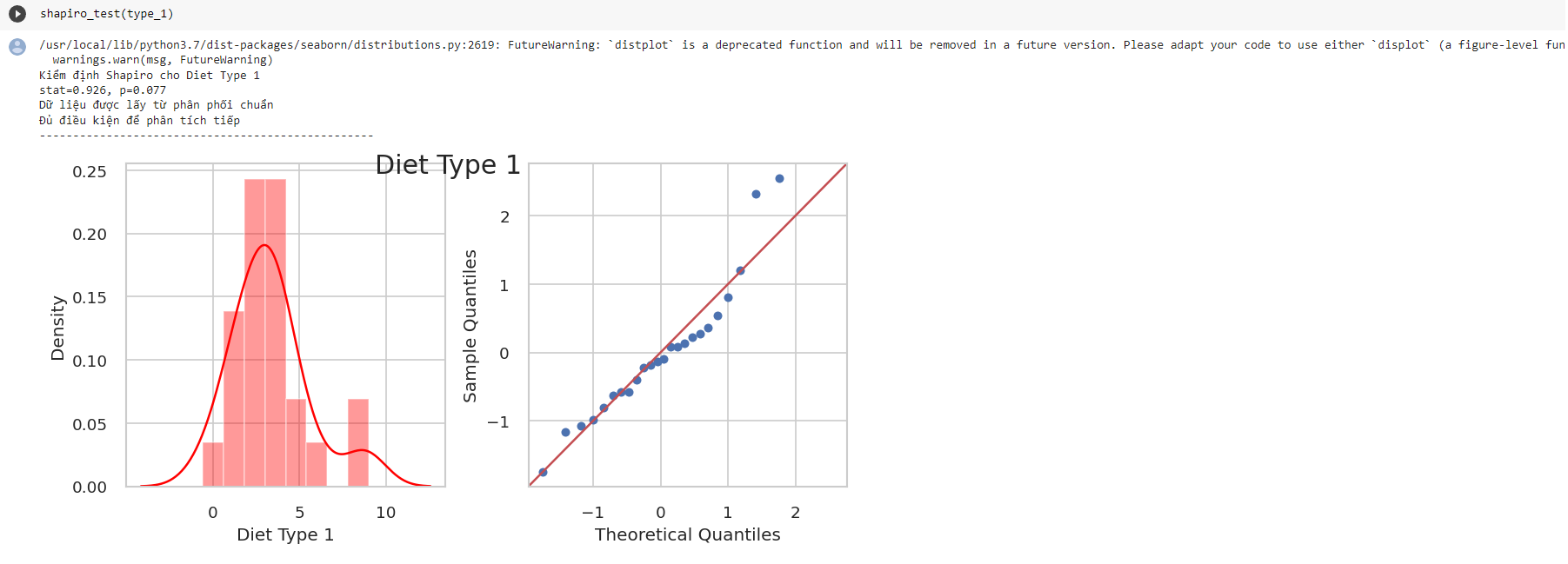
- Để kiểm tra dữ liệu có tuân theo phân phối chuẩn hay không, chúng ta sử dụng kiểm định Shapiro-Wilk được tích hợp sẵn trong thư viện Scipy:

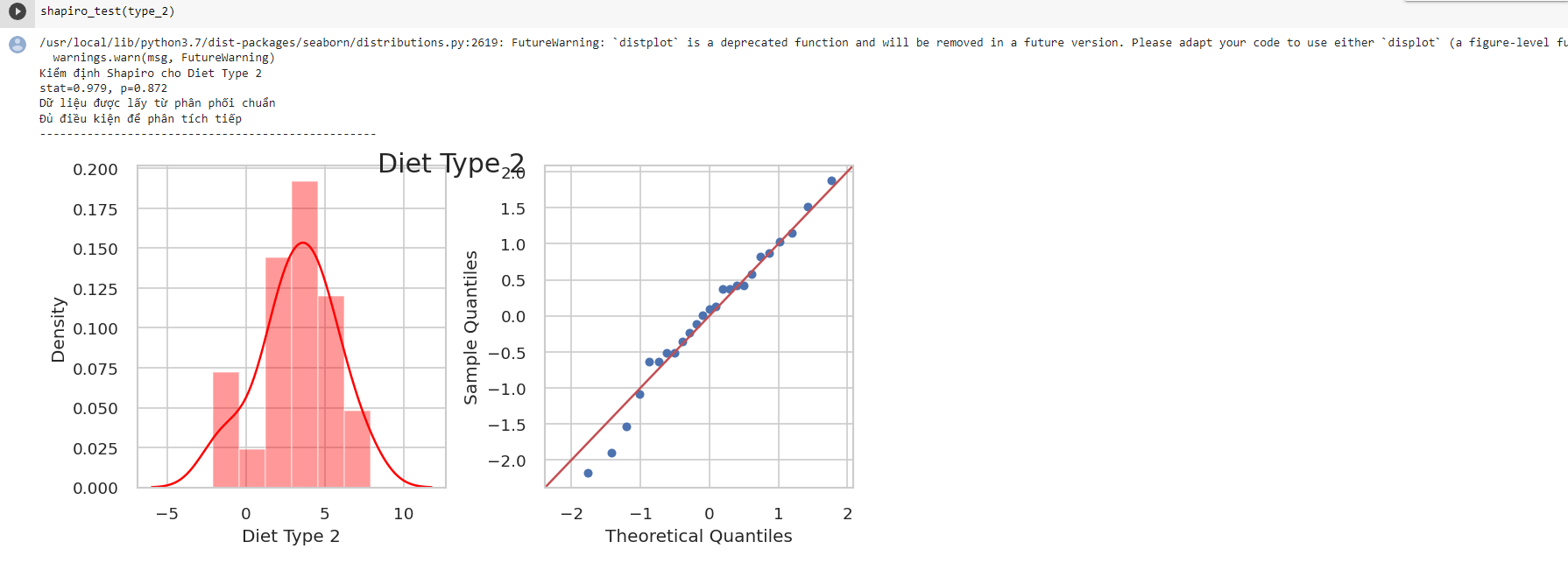
**Phát biểu giả thuyết :**

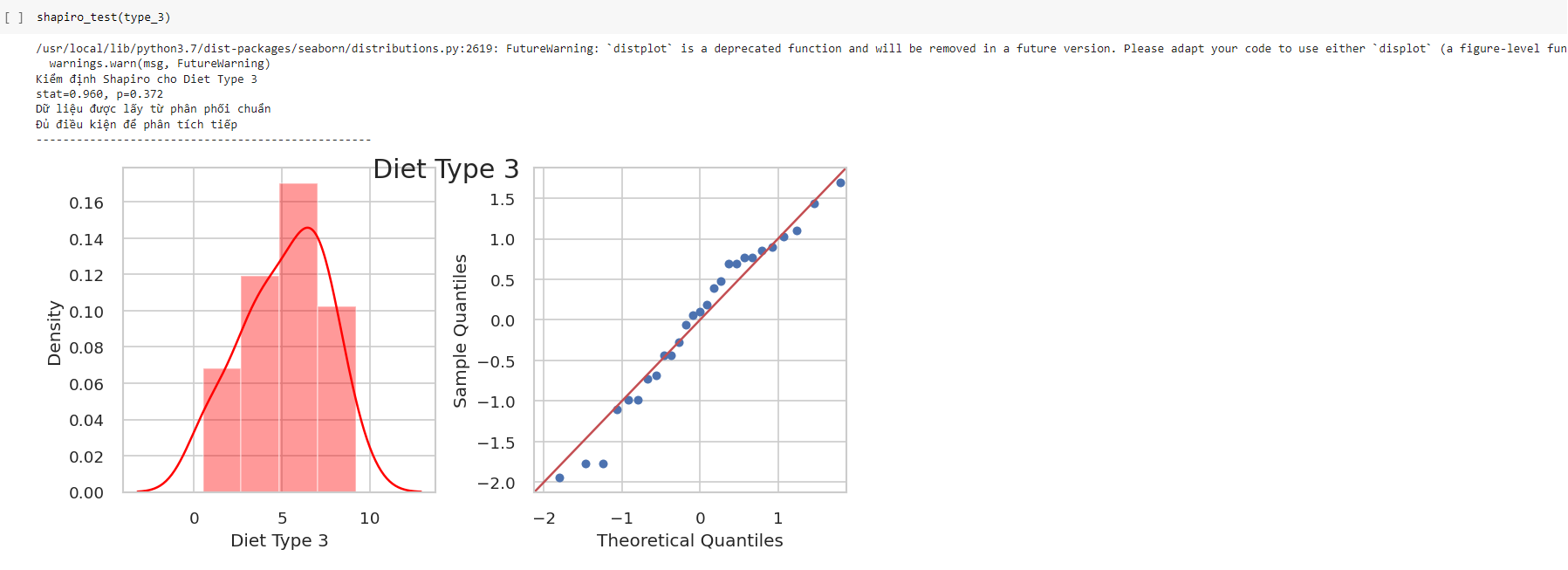
H0: Các nhóm được lấy từ phân phối chuẩn

Ha: Các nhóm không được lấy từ phân phối chuẩn









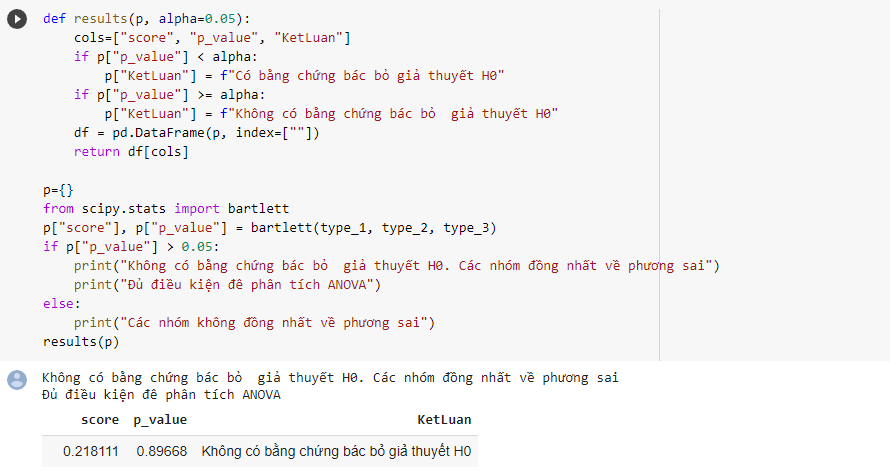
**Kết luận:** Cả ba nhóm dữ liệu đều được lấy từ phân phối chuẩn

-Bắt đầu tiến hành kiểm định tính thuần nhất của phương sai - Sử dụng Bartlett"s Test

**Phát biểu giả thuyết :**

H0: Các nhóm có phương sai giống nhau

Ha: Các nhóm có phương sai không giống nhau



- Kiểm định ANOVA one way

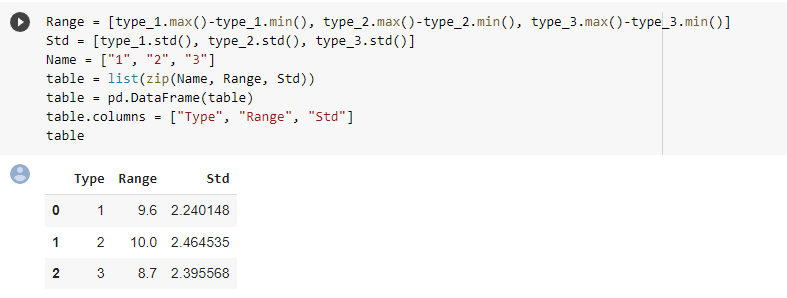
**Phát biểu giả thuyết:**

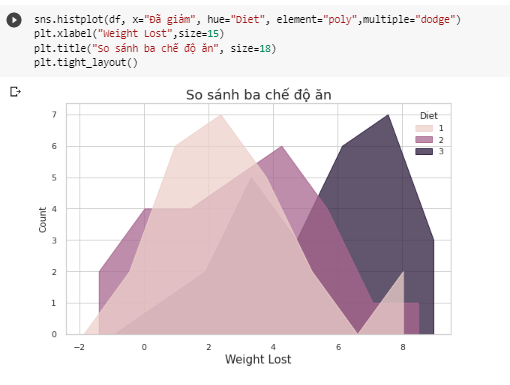
H0:Các chế độ ăn có hiệu quả giảm cân giống nhau

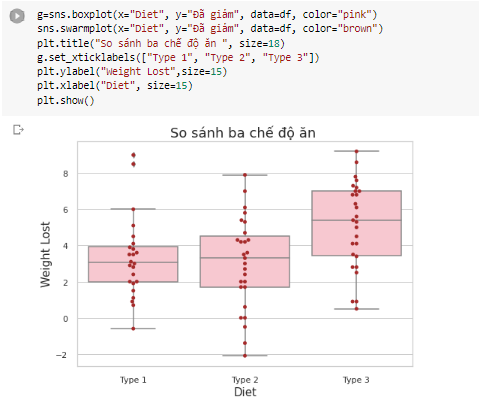
Ha:Các chế độ ăn có hiệu quả giảm cân không giống nhau



-Vậy chế độ ăn nào là tốt nhất để giảm cân?







**Kết luận:**

- Nhìn bảng thống kê và hai biểu đồ ở trên, ta dễ dàng nhận thấy chế độ ăn Diet 3 có khoảng tin cậy lớn nhất, giá trị trung bình lớn nhất và độ lệch chuẩn tương đối nhỏ nên chế độ ăn thứ 3 có hiệu quả giảm cân tốt nhất.

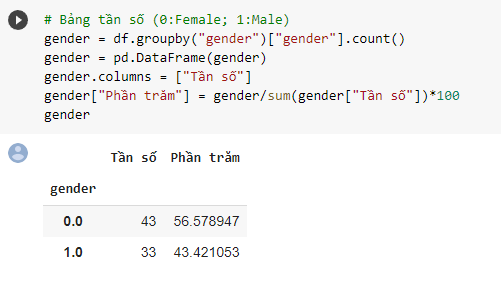
**Research question 2: Are there gender differences for weight lost?**

**"Khả năng giảm cân có phụ thuộc vào giới tính hay không?"**

Xây dựng giả thuyết:

H0: μ1=μ2

Ha: μ1≠μ2

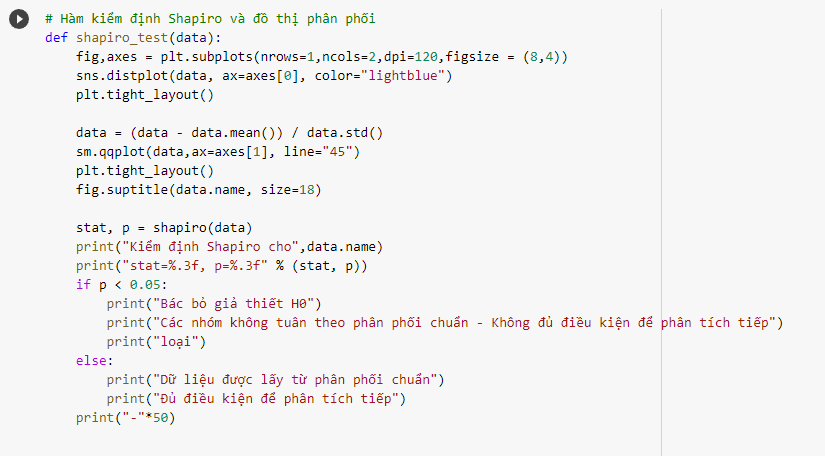


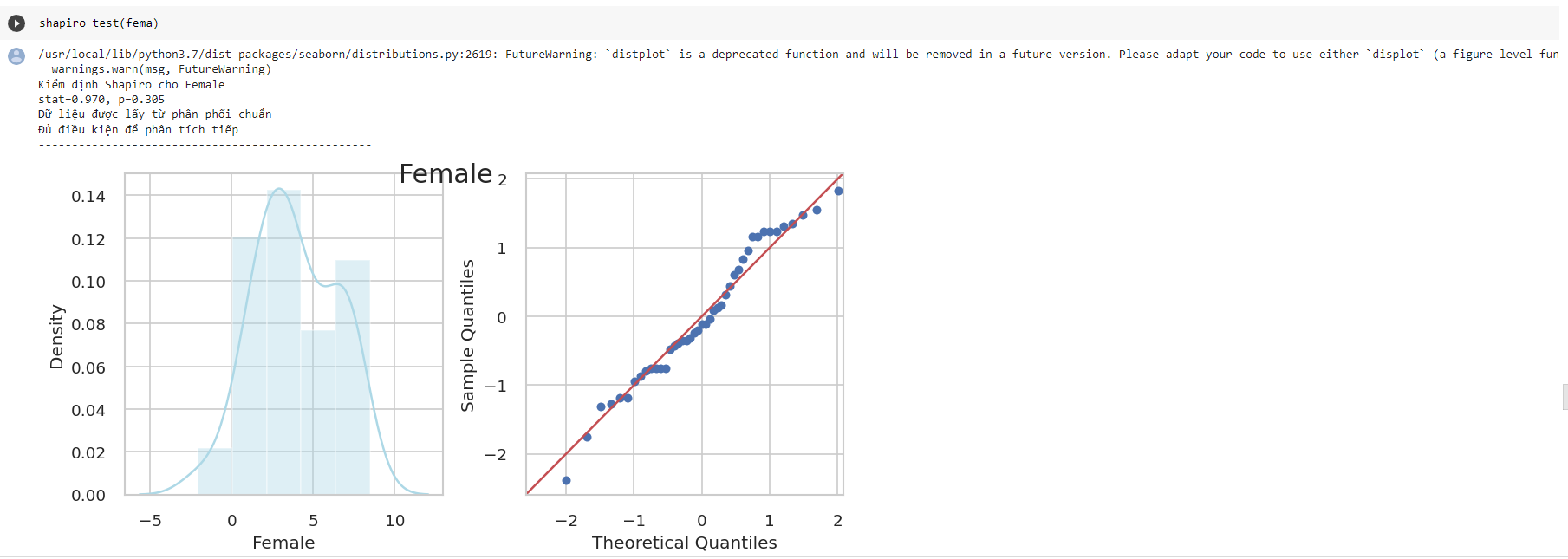
-Kiểm tra xem có được lấy từ phân phối chuẩn

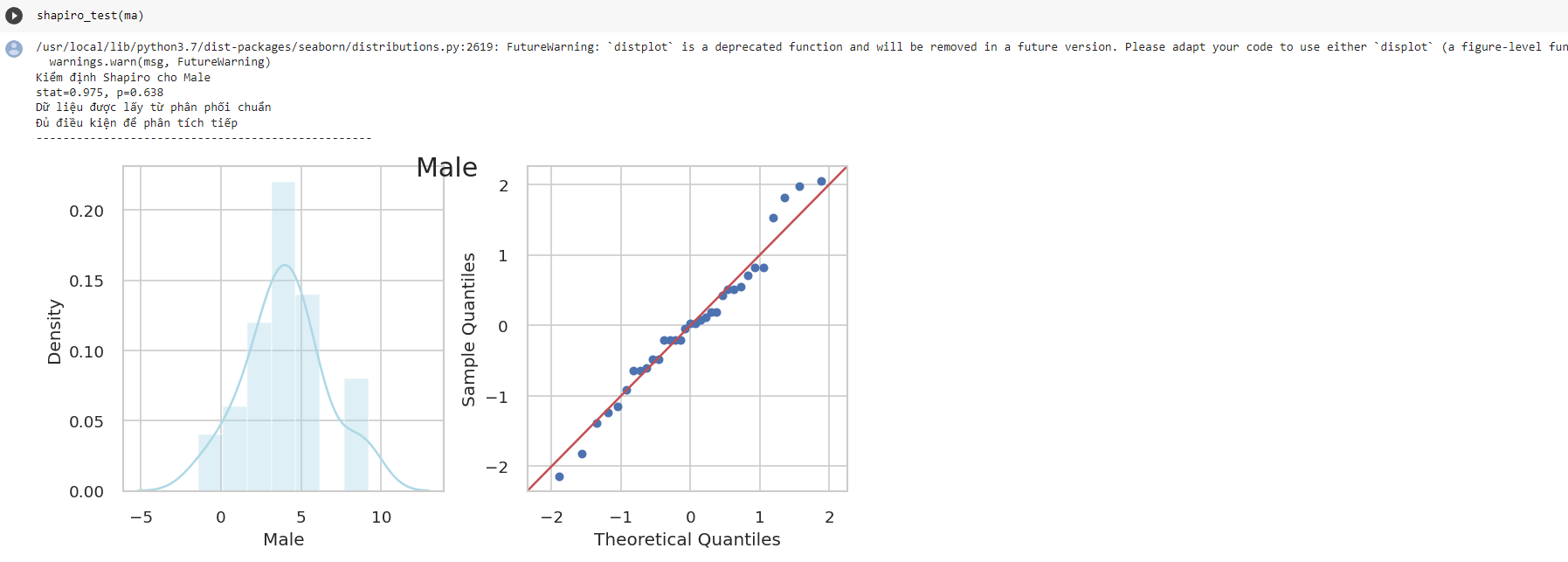
**Phát biểu giả thuyết :**

H0: Các nhóm được lấy từ phân phối chuẩn

Ha: Các nhóm không được lấy từ phân phối chuẩn







**Kết luận:**

Cả hai nhóm dữ liệu đều tuân theo phân phối chuẩn

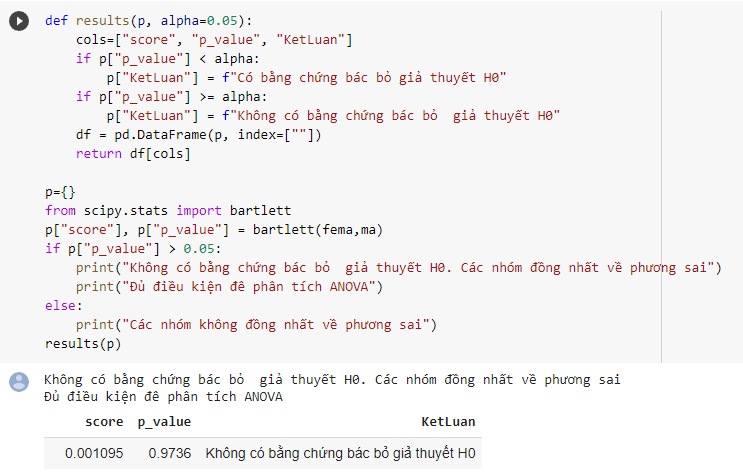
-Bắt đầu tiến hành kiểm định tính thuần nhất của phương sai - Sử dụng Bartlett"s Test

**Phát biểu giả thuyết :**

H0:Các nhóm có phương sai giống nhau

Ha:Các nhóm có phương sai không giống nhau

Xây dựng hàm để kiểm định phương sai

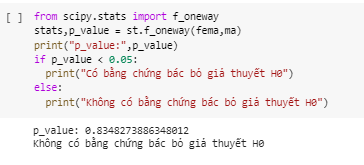


**-Kiểm định ANOVA one way**

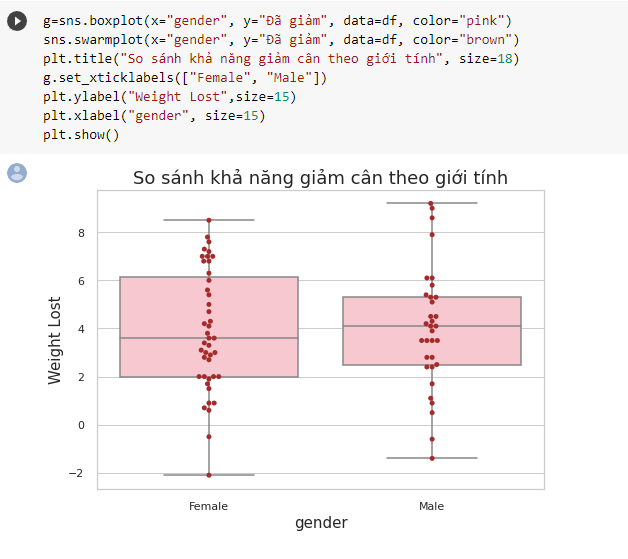
Phát biểu giả thuyết :

H0:Khả năng giảm cân không phụ thuộc vào giới tính

Ha:Khả năng giảm cân phụ thuộc vào giới tính







**Kết luận:**

Khả năng giảm cân không phụ thuộc vào giới tính

* 1. **Interactions**

Research question: Means plot of weight lost by diet and gender

"Biểu đồ giảm cân theo chế độ ăn uống và giới tính"





# Chương 3: Tài liệu tham khảo

<https://pandas.pydata.org/docs/>

<https://www.scipy.org/docs.html>

<https://matplotlib.org/>

<https://stackoverflow.com/>

<https://vietnambiz.vn/thong-ke-statistics-la-gi-phan-loai-thong-ke-2020061716451094.htm>

<https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_m%C3%B4_t%E1%BA%A3>

<https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA#%E1%BB%A8ng_d%E1%BB%A5ng>

<https://en.wikipedia.org/wiki/Python_(programming_language)>

**Chương 4: Kiểm tra đạo văn**

