**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**



🙠🙟🕮🙝🙢

****

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

**Giảng 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: Lê Đình Đăng Khoa**

*(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**

[**CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 5**](#_Toc75787114)

[**I. Thống kê 5**](#_Toc75787115)

[**1. Định nghĩa 5**](#_Toc75787116)

[**2. Phương pháp 5**](#_Toc75787117)

[**3. Các bước nghiên cứu thông kê 5**](#_Toc75787118)

[**4. Chức năng 5**](#_Toc75787119)

[**II. Thống kê trong Python 5**](#_Toc75787120)

[**1. Giới thiệu ngôn ngữ lập trình Python 5**](#_Toc75787121)

[**2. Giới thiệu các thư viện cần dùng 5**](#_Toc75787122)

[**CHƯƠNG 2: PHÂN TÍCH DỮ LIỆU 6**](#_Toc75787123)

[**I. Thống kê mô tả 7**](#_Toc75787124)

[**1. Thu thập và làm sạch dữ liệu: 7**](#_Toc75787125)

[**2. Phân tích xu hướng tập trung của trung tâm dữ liệu: 8**](#_Toc75787126)

[**3. Biện pháp phân tích sự thay đổi, phân tán của tập dữ liệu: 10**](#_Toc75787127)

[**4. Tóm tắt thống kê mô tả 13**](#_Toc75787128)

[**5. Trực quan hóa dữ liệu 14**](#_Toc75787129)

[**II. Thống kê suy diễn: 17**](#_Toc75787130)

[**1. Bỏ qua chế độ ăn kiêng và kiểm tra xem đã giảm cân chưa? 19**](#_Toc75787131)

[**2. Loại bỏ trọng lượng đã mất và yêu cầu học sinh tính toán nó bằng cách 19**](#_Toc75787132)

[**sử dụng trọng lượng trước trừ sau. 19**](#_Toc75787133)

[**3. Thống kê tóm tắt theo chế độ ăn uống 20**](#_Toc75787134)

[**4. Kiểm định ANOVA: 21**](#_Toc75787135)

[**5. Ảnh hưởng chế độ ăn kiêng và giới tính đến việc giảm cân: 25**](#_Toc75787136)

[**Tương tác 25**](#_Toc75787137)

[**6. Ancova 25**](#_Toc75787138)

[**CHƯƠNG III: TÀI LIỆU THAM KHẢO : 26**](#_Toc75787139)

[**CHƯƠNG IV: TỈ LỆ ĐẠO VĂN 26**](#_Toc75787140)

# CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1. **Thống kê**
2. **Định nghĩa**

-Thống kê là sự nghiên cứu của cá tập hợp gồm nhiều lĩnh vực khác nhau trong đó bao gồm phân tích, giải thích, trình bày và tổ chức, sắp xếp các dữ liệu cần quan sát.

-Là một dạng phân tích toán học bằng cách sử dụng các dạng mô hình, tóm tắt “định lượng” cho một tập data thực nghiệm hoặc nghiên cứu thực tế với mục đích phục vụ cho quá trình phân tích, dự đoán và quyết định mục tiêu.

-Thống kê được chia làm 2 loại: thống kê mô tả và thống kê suy luận. Với mỗi loại sẽ có mỗi đặc điểm, chức năng riêng. Khi đã xác định mục đích thông kê, chúng ta có thể lựa chọn phương pháp thực hiện, từ đó sẽ có thể đưa ra các đánh giá chính xác nhất khi thực hiện thống kê.

1. **Phương pháp**

Thống kê sẽ sử dụng 4 loại phương pháp và dự đoán để thực hiện nghiên cứu tập dữ liệu tốt nhất:

-Phương pháp thu thập và xử lý số liệu: là phương pháp sử dụng khi số liệu và dữ liệu có thể chưa đáp ứng được mục tiêu trong quá trình nghiên cứu. Vì vậy cần phải tiến hành xử lý, trình bày và tính toán. Khi đó kết quả sẽ làm khái quát đặc trưng của tổng thể.

-Điều tra chọn mẫu: sử dụng phương pháp nghiên cứu một bộ phận của tổng thể nhưng lại có thể suy luận cho hiện tượng tổng quát và bảo đảm độ tin cây.

– Nghiên cứu mối liên hệ giữa các hiện tượng: phương pháp thống kê này thống kê những mối liên hệ của các tập dữ liệu với nhau.

-Dự đoán: là phương pháp quan trọng trong mục đích thống kê. Từ 3 phương pháp ở trên có thể đưa ra những dự đoán.

1. **Các bước nghiên cứu thông kê**

Về cơ bản cần có 5 bước khi thực hiện thống kê:

-1: Lập các mục đích, kế hoạch thống kê. Bao gồm phân tích các số liệu để trả lời các câu hỏi nghiên cứu bằng việc sử dụng các thông tin.

-2: Thiết kế nghiên cứu thống kê như xóa những dữ liệu gây nhiễu

-3: Kiểm tra các nghiên cứu

-4: Kiểm tra các dữ liệu thiết lập, đề xuất các giả thuyết mới

-5: Tìm các tài liệu nghiên cứu thêm và trình bày

1. **Chức năng**

**-Thống kê mô tả:** phản ánh 1 cách khái quát đối tượng nghiên cứu

**-Thống kê suy diễn:** ước lượng đặc trưng các tổng thể từ đó dự đoán quyết định từ cơ sở thu thập trên kết quả quan sát mẫu.

1. **Thống kê trong Python**
2. **Giới thiệu ngôn ngữ lập trình Python**

**-**Python là một ngôn ngữ lập trình bậc cao cho mục đích lập trình đa năng do Guido van Rossum tạo ra và đưa vào sử dụng vào năm 1991.

**-**Python là ngôn ngữ có ưu điểm mạnh là dễ đọc, học và nhớ. Có cấu trúc rõ ràng và thuận tiện cho người lập trình.

**-**Python có tốc độ thực hiện chậm hơn nhiều so với các ngôn ngữ biên dịch vì Python là ngôn ngữ thông dịch.

1. **Giới thiệu các thư viện cần dùng**

-**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 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. Pandas 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.

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…

-**Matplotlib:** 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.

- **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. Thư viện SciPy phụ thuộc vào NumPy, cung cấp khả năng thao tác mảng N chiều thuận tiện và nhanh chóng. Lý do chính để xây dựng thư viện SciPy là nó hoạt động với các mảng NumPy. Nó cung cấp nhiều thực hành số hiệu quả và thân thiện với người dùng chẳng hạn như các quy trình để tích hợp và tối ưu hóa số. Đây là một hướng dẫn giới thiệu, bao gồm các nguyên tắc cơ bản của SciPy và mô tả cách xử lý các mô-đun khác nhau của nó.

- **NumPy**: viết tắt của Numerical Python, là một thư viện bao gồm các đối tượng mảng đa chiều và một tập hợp các quy trình để xử lý các mảng đó. Sử dụng NumPy, các phép toán logic và toán học trên mảng có thể được thực hiện. Hướng dẫn này giải thích những điều cơ bản về NumPy như kiến ​​trúc và môi trường của nó. Nó cũng thảo luận về các chức năng mảng khác nhau, các kiểu lập chỉ mục, v.v.

-**Seaborn**: là một thư viện Python mã nguồn mở, được cấp phép BSD cung cấp API cấp cao để hiển thị dữ liệu bằng ngôn ngữ lập trình Python.

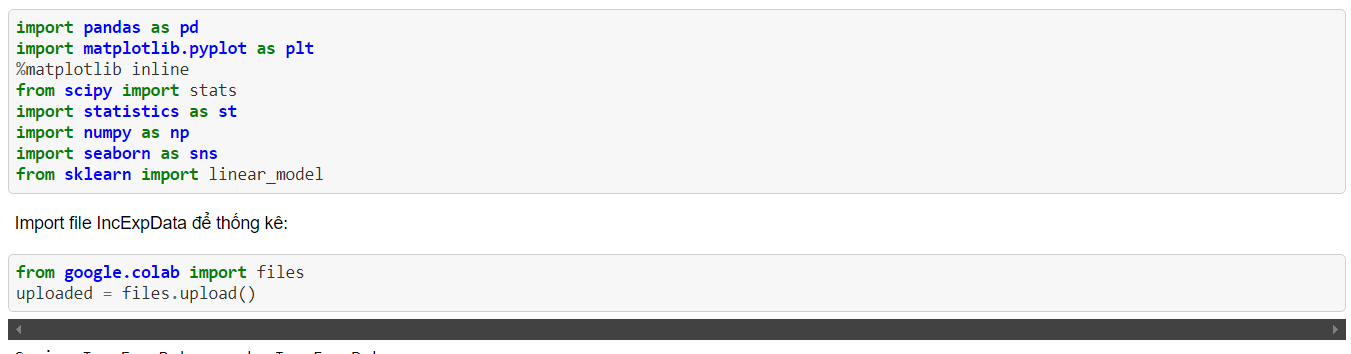
-[**Statsmodels**](https://www.statsmodels.org/stable/about.html#about-statsmodels) là một mô-đun Python cung cấp các lớp và chức năng để ước tính nhiều mô hình thống kê khác nhau, cũng như để thực hiện các bài kiểm tra thống kê và thăm dò dữ liệu thống kê. Danh sách thống kê kết quả phong phú có sẵn cho mỗi công cụ ước tính. Kết quả được kiểm tra dựa trên các gói thống kê hiện có để đảm bảo rằng chúng là chính xác. Gói được phát hành theo giấy phép BSD sửa đổi (3 điều khoản) nguồn mở

- **Pingouin** là một gói thống kê mã nguồn mở được viết bằng Python 3 và chủ yếu dựa trên Pandas và NumPy.

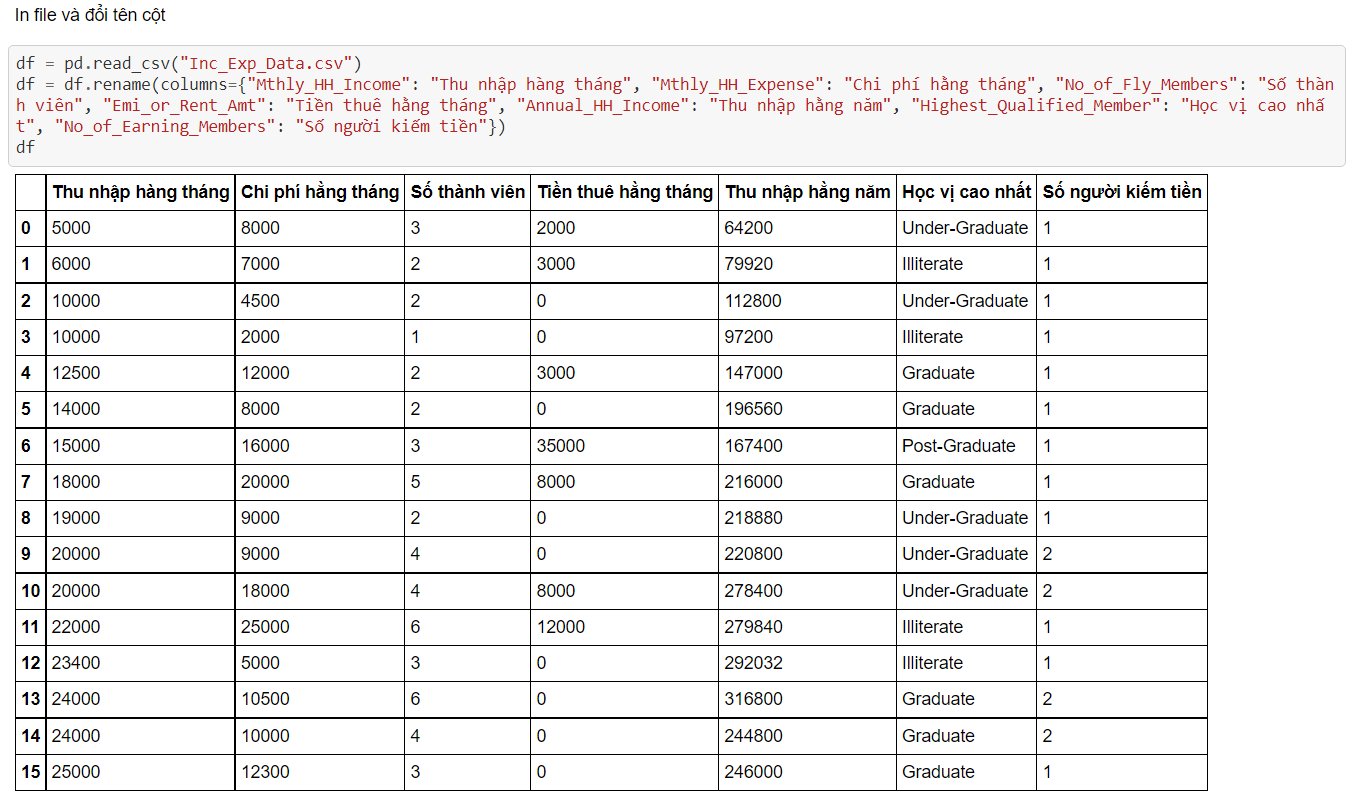
**CHƯƠNG 2: PHÂN TÍCH DỮ LIỆU**

1. **Thống kê mô tả**

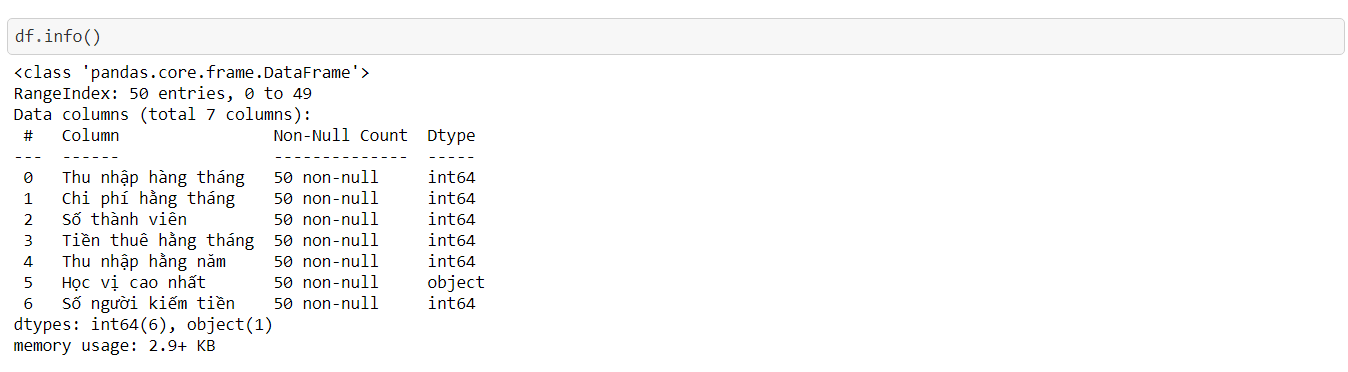
**Import file và các thư viện cần dùng:**



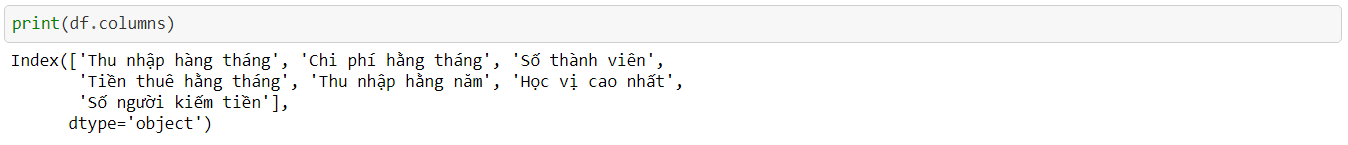
1. **Thu thập và làm sạch dữ liệu:**



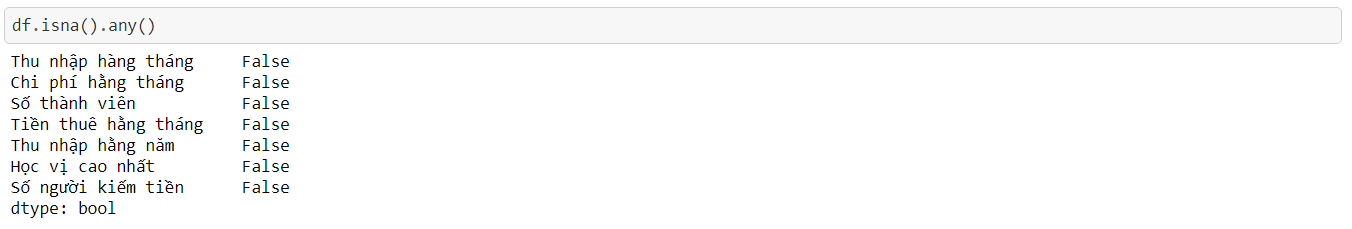
Thông tin của dataframe:



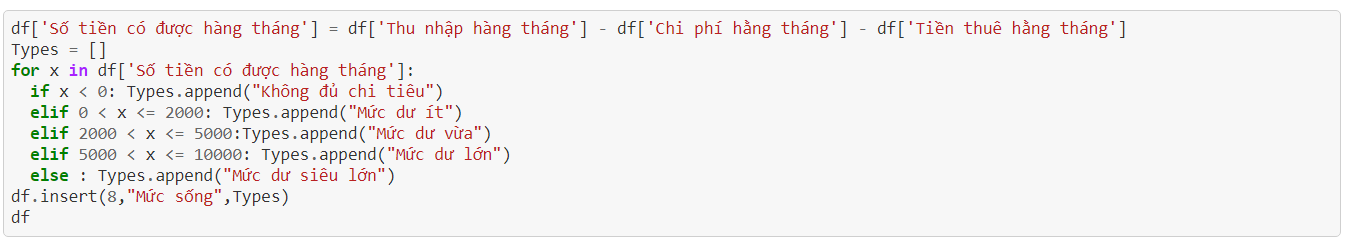
Thông tin các cột trong dataframe:

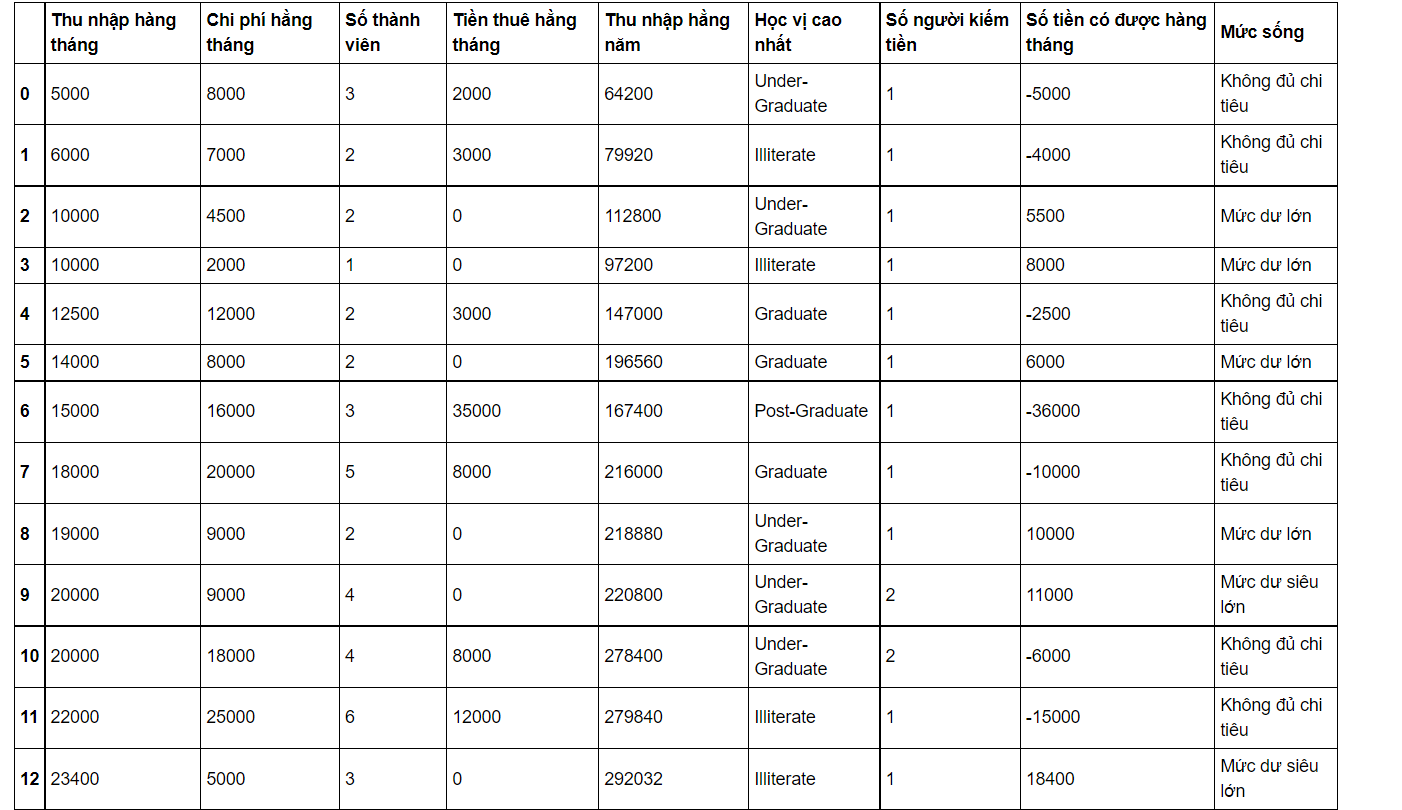


Khảo sát có tập dữ liệu rỗng không (False = không có)



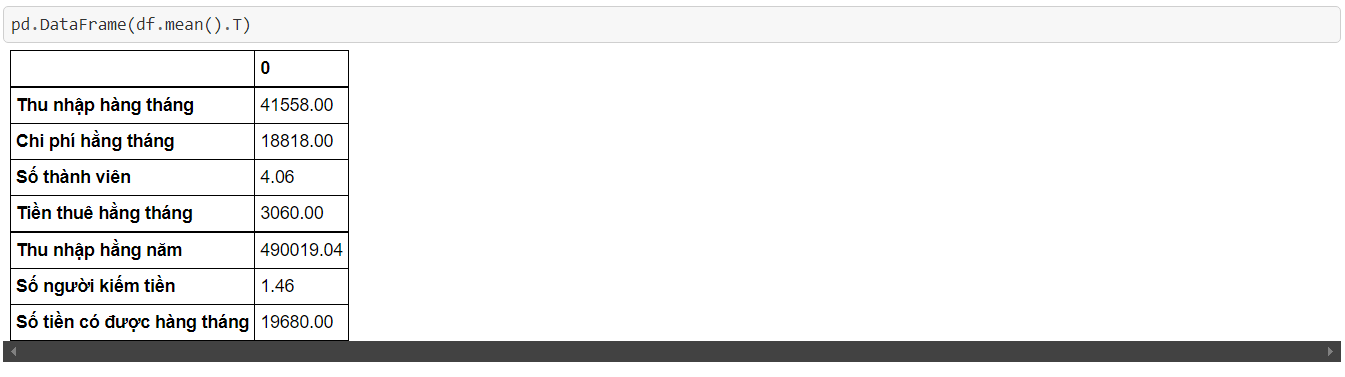
Phân chia mức sống để tối ưu trực quan dữ liệu



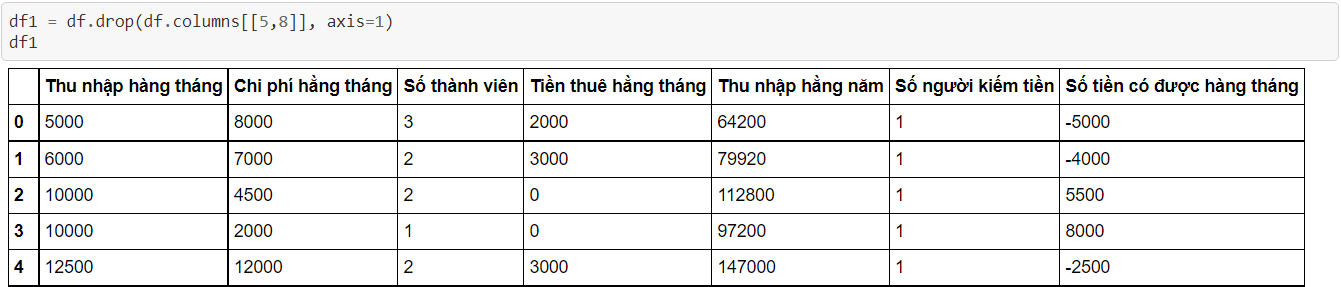


1. **Phân tích xu hướng tập trung của trung tâm dữ liệu:**

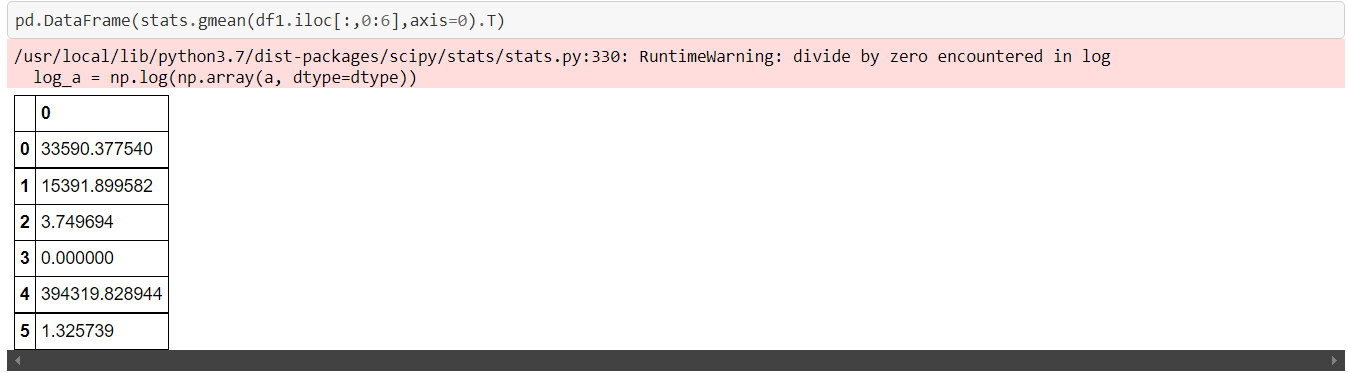
-Giá trị trung bình của tập dữ liệu (Mean):



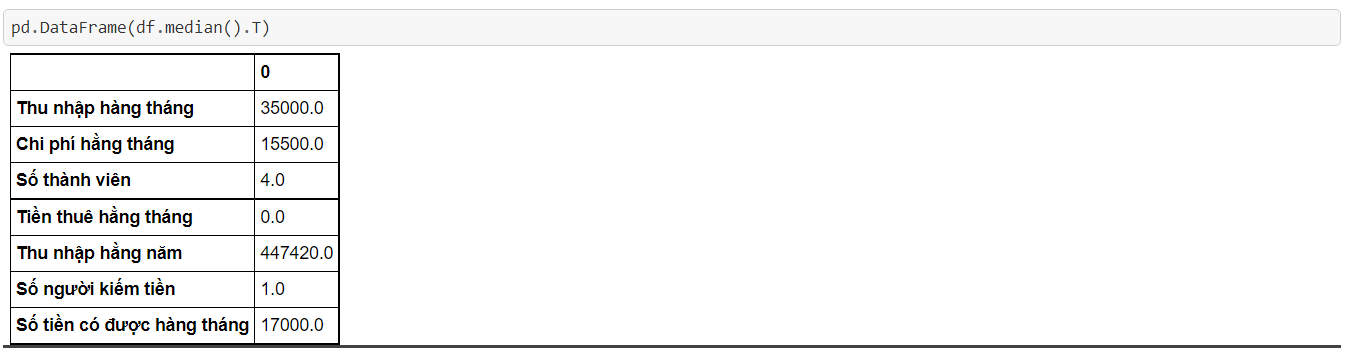
- Harmonic mean của dữ liệu cột trong Data: (để tính toán ta cần phải xóa cột học vị và mức sống): Số bình quân điều hoà giúp tìm mối quan hệ nhân hoặc chia giữa các phân số mà không phải quan tâm đến mẫu số chung. Số bình quân điều hoà thường được sử dụng trong các hoạt động trung bình như tỉ lệ



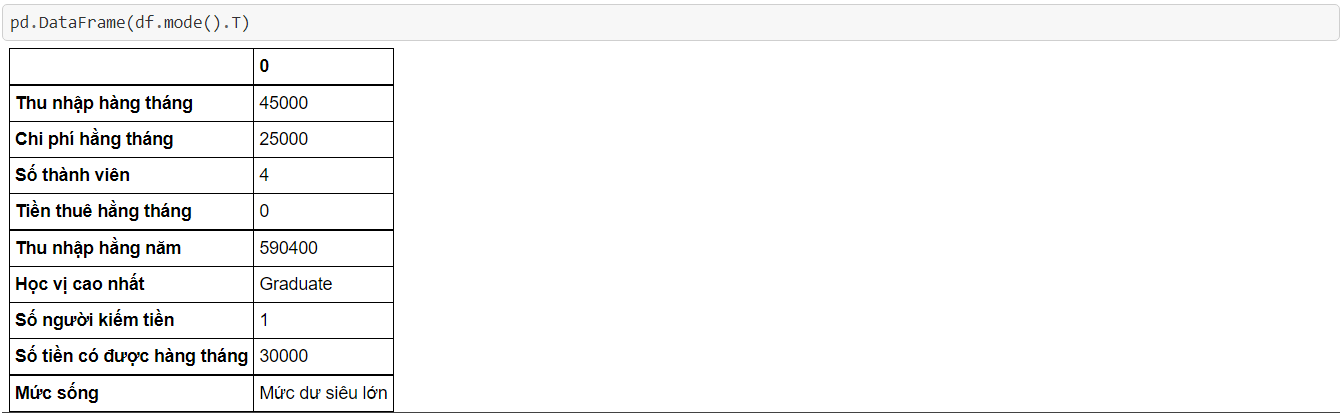
-Giá trị trung bình hình học: Số trung bình số học được tính bằng cách cộng tất cả các kết quả và chia cho tổng số của chúng



-Giá trị trung vị (median): là số ở giữa trong một danh sách các số được sắp xếp tăng dần hoặc giảm dần và có thể mô tả nhiều hơn về tập dữ liệu so với giá trị trung bình

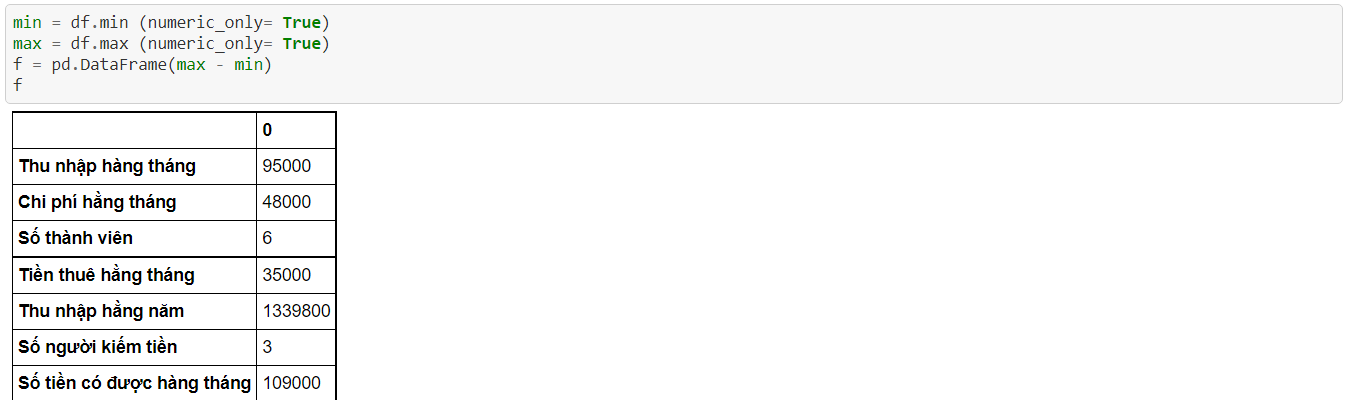


-Giá trị yếu vị (mode): của một danh sách dữ liệu hoặc một mẫu là giá trị của phần tử có số lần xuất hiện lớn nhất trong danh sách

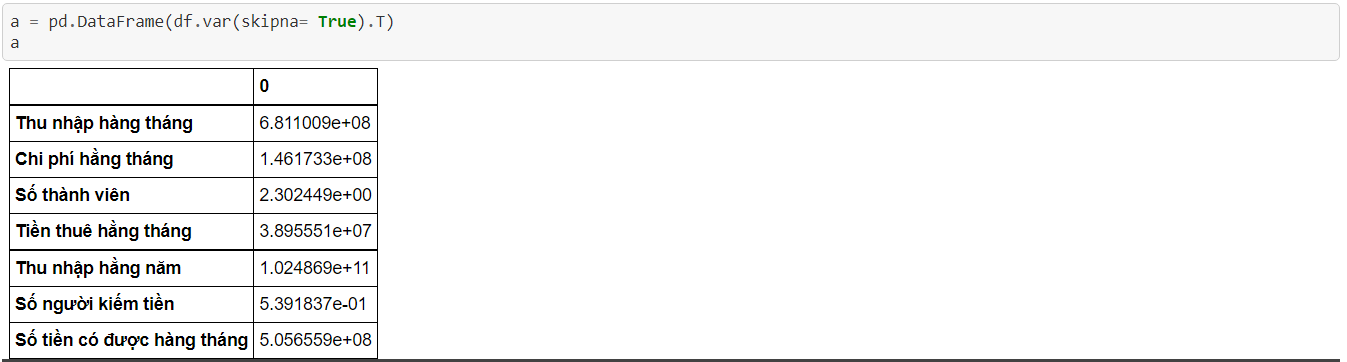


1. **Biện pháp phân tích sự thay đổi, phân tán của tập dữ liệu:**

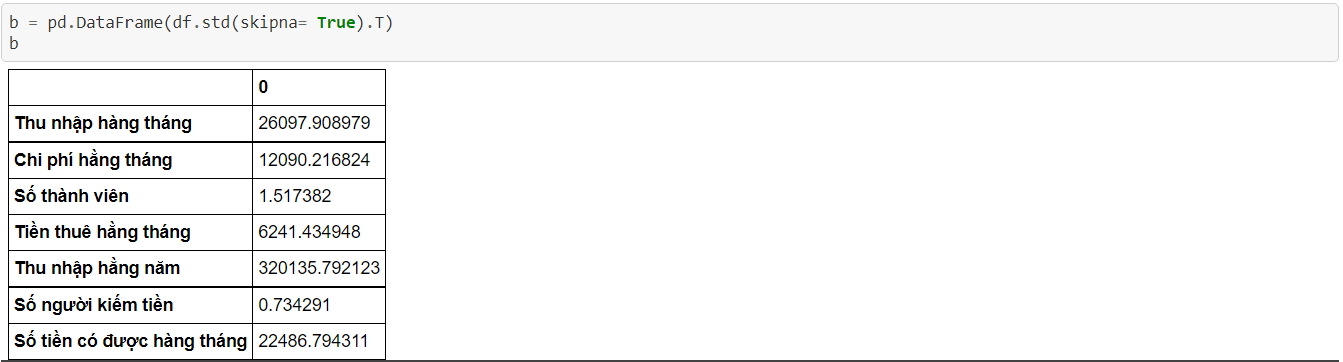
-Phạm vi (Độ biến thiên) của dữ liệu: là một đại lượng mô tả mức độ phân tán của dữ liệu. Khoảng biến thiên được tính bằng cách lấy giá trị quan sát lớn nhất trừ đi giá trị quan sát nhỏ nhất.



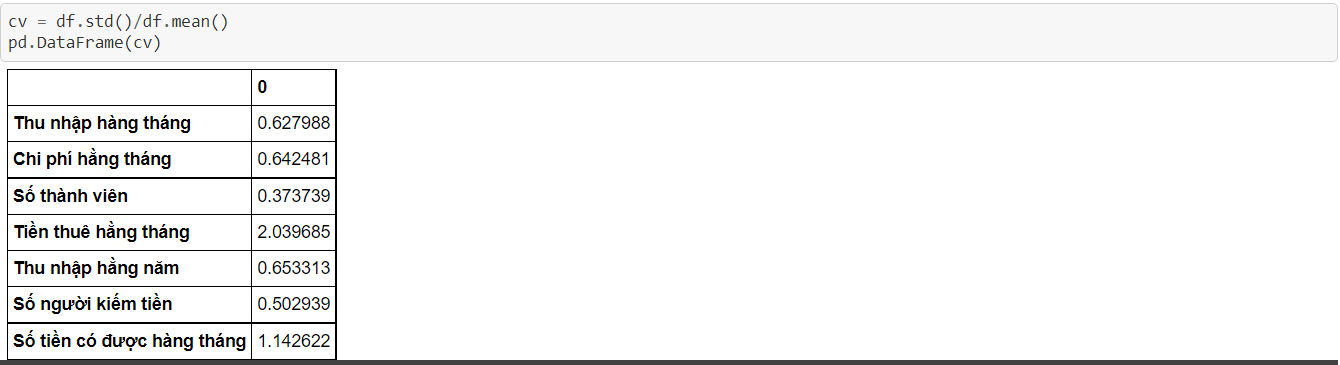
- Phương sai của dữ liệu:  là một độ đo sự [phân tán thống kê](https://vi.wikipedia.org/w/index.php?title=Ph%C3%A2n_t%C3%A1n_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1) của biến đó, nó hàm ý các giá trị của biến đó thường ở cách [giá trị kỳ vọng](https://vi.wikipedia.org/wiki/Gi%C3%A1_tr%E1%BB%8B_k%E1%BB%B3_v%E1%BB%8Dng) bao xa.



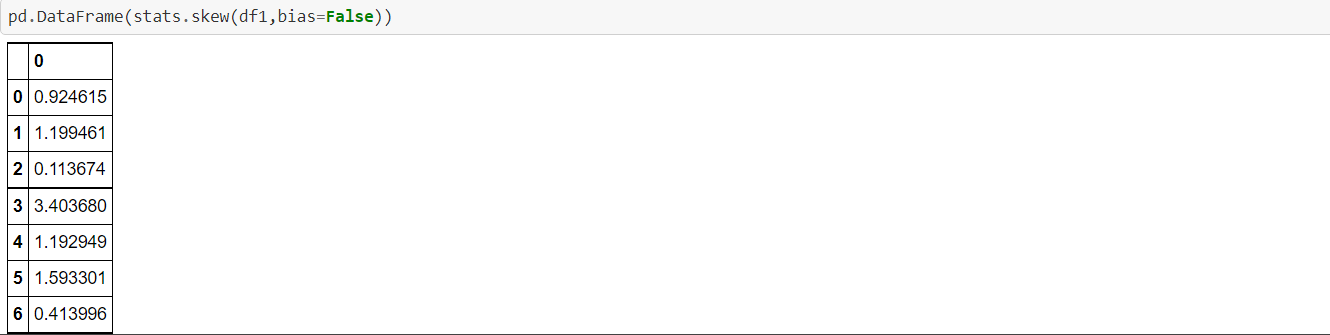
-Độ lệch chuẩn là một đại lượng [thống kê mô tả](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_m%C3%B4_t%E1%BA%A3) dùng để đo mức độ phân tán của một tập [dữ liệu](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u) đã được lập thành bảng tần số.



-Hệ số biến động là một thước đo thống kê độ phân tán của các dữ liệu trong một chuỗi dữ liệu so với giá trị trung bình.



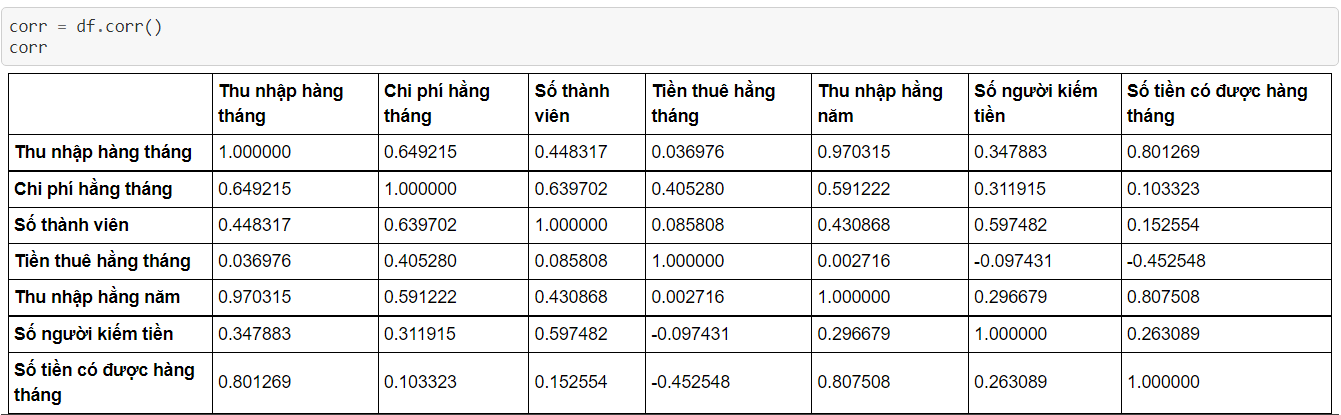
- Độ lệch mẫu đo độ không đối xứng của dữ liệu: Độ lệch (skewness) của một phân phối xác suất đo lường sự đối xứng của phân phối đó. Giá trị tuyệt đối của độ lệch càng cao thì phân phối đó càng bất đối xứng. Một phân phối đối xứng có độ lệch bằng 0.



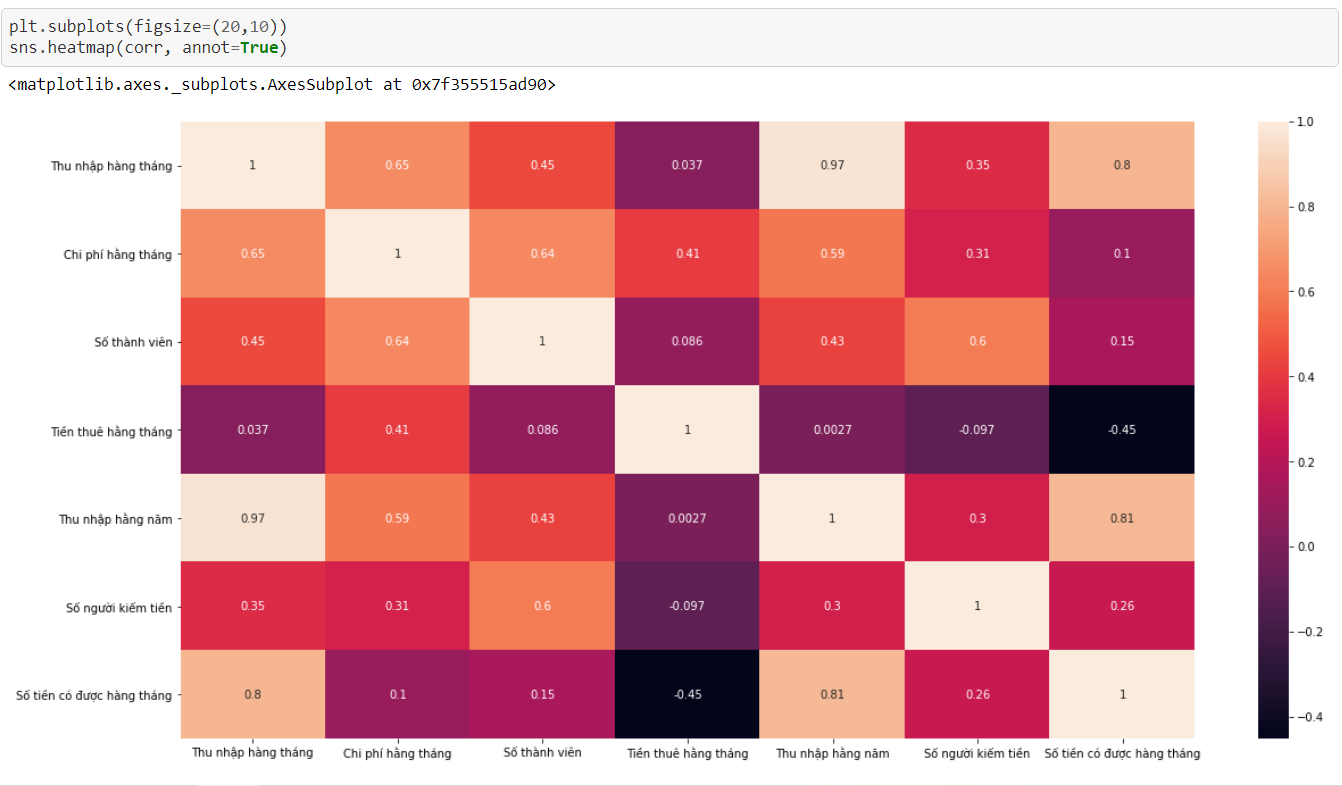
-Phần trăm (Phân vị mẫu) của dữ liệu:  là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất, thứ nhì, và thứ ba. Ba giá trị này chia một tập hợp dữ liệu (đã sắp xếp dữ liệu theo trật từ từ bé đến lớn) thành 4 phần có số lượng quan sát đều nhau.



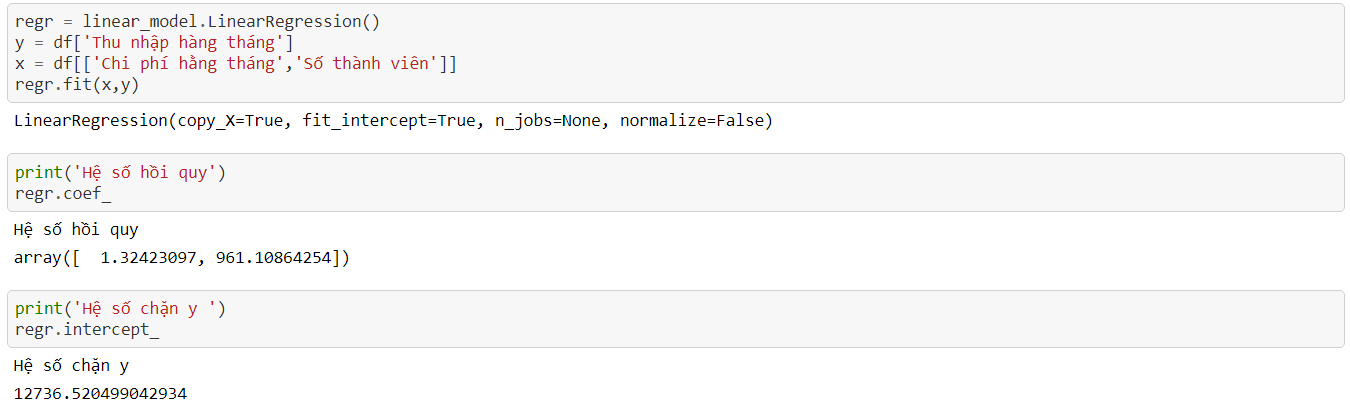
-Thước đo về mối tương quan giữa các cặp dữ liệu. Bảng thống kê mức độ phụ thuộc của từng tập dữ liệu trong data: Corr: Tính toán mối tương quan theo cặp của các cột, không bao gồm giá trị NA / null.



-Biểu đồ thể hiện mối tương quan theo cặp của các cột



-Thiết lập mô hình hồi quy tuyến tính: Phân tích hồi quy tuyến tính là một phương pháp phân tích quan hệ giữa biến phụ thuộc Y với một hay nhiều biến độc lập X. Mô hình hóa sử dụng hàm tuyến tính. Các tham số của mô hình được ước lượng từ dữ liệu. Hồi quy tuyến tính được sử dụng rộng rãi trong thực tế do tính chất đơn giản hóa của hồi quy.

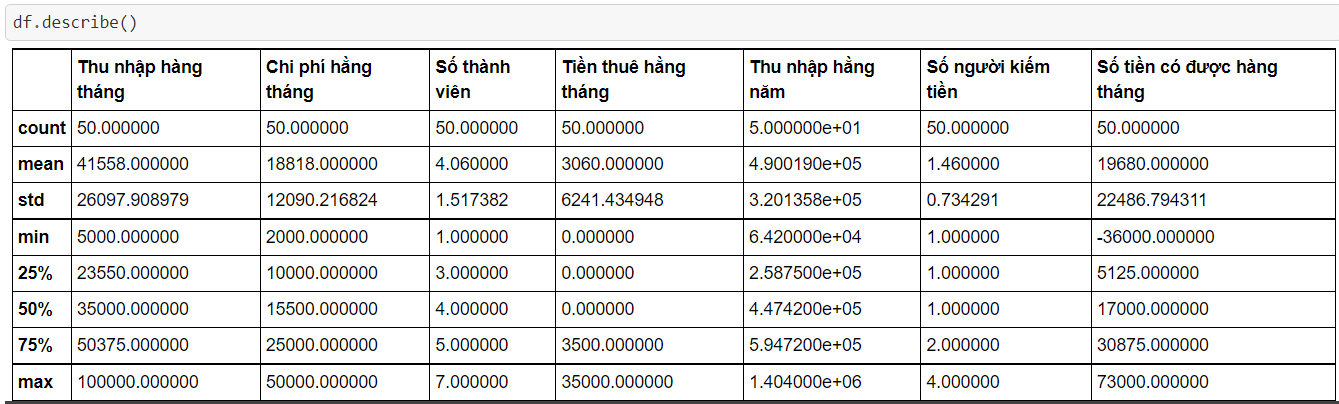


-Hệ số hồi quy: phản ánh độ dốc của đường hồi quy tuyến tính có ý nghĩa đặc biệt quan trọng trong kinh tế học, vì nó cho thấy sự thay đổi của biến phụ thuộc

-Hệ số chặn y : sử dụng quy ước chung rằng trục hoành biểu thị một biến x và trục dọc biểu thị một biến y, chặn y hoặc chặn dọc là một điểm trong đó đồ thị của hàm hoặc quan hệ giao với trục y của hệ tọa độ

1. **Tóm tắt thống kê mô tả**

**Sử dụng hàm Describe:** sử dụng để tạo thống kê mô tả tóm tắt xu hướng trung tâm, độ phân tán và hình dạng phân bố của tập dữ liệu, không bao gồm các giá trị NaN.

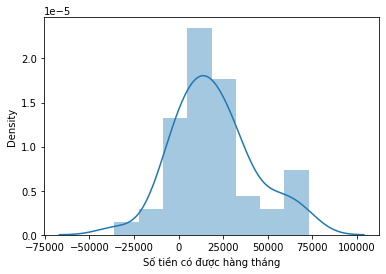


1. **Trực quan hóa dữ liệu**

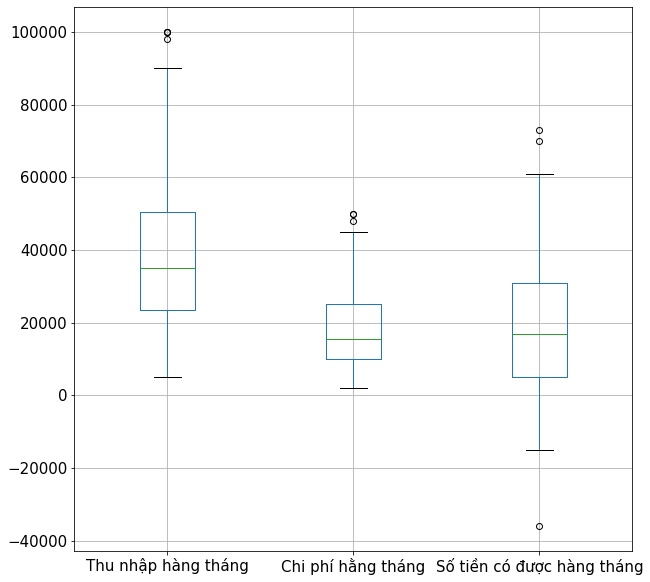
**Định nghĩa:** là kỹ thuật tạo ra những hình ảnh, biểu đồ, đồ thị để diễn tả thông điệp, thông tin đến người dùng. Trực quan hóa nghiên cứu trình bày một cách trực quan khối dữ liệu trừu tượng để tăng cường nhận thức của con người.

**Mục đích:** Mục tiêu chính của trực quan hóa dữ liệu là truyền thông tin rõ ràng và hiệu quả cho người sử dụng thông qua đồ họa được lựa chọn như bảng biểu hoặc biểu đồ.

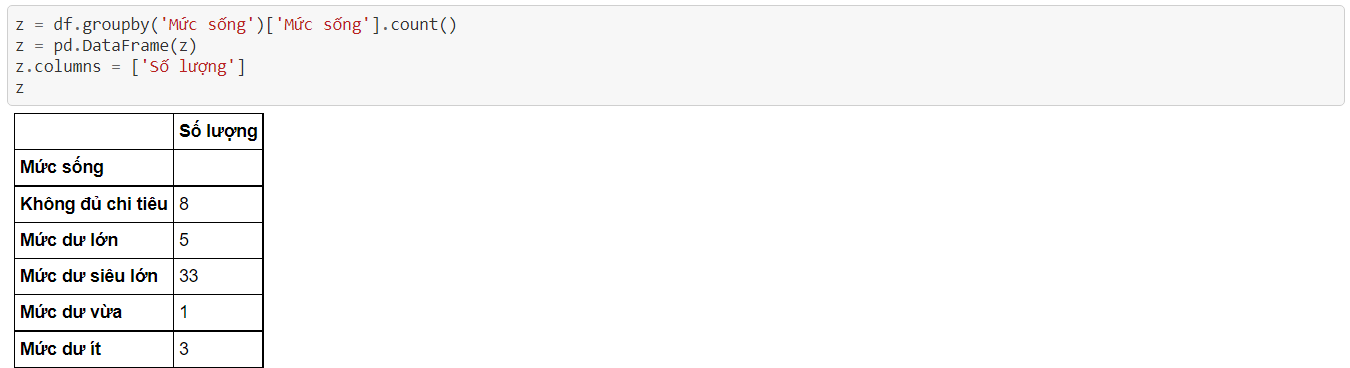
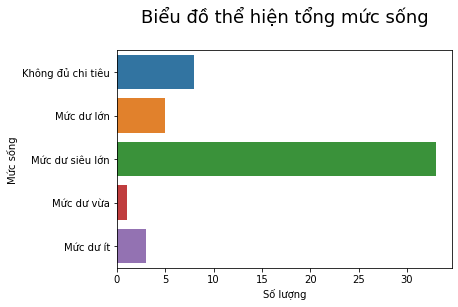
Biểu đồ thể hiện số tiền hàng tháng có được ở dữ liệu

****

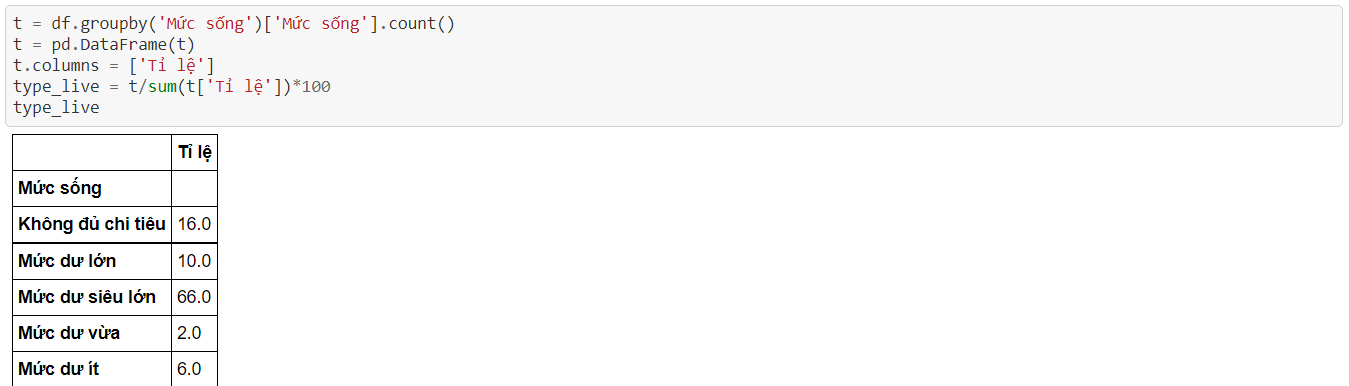
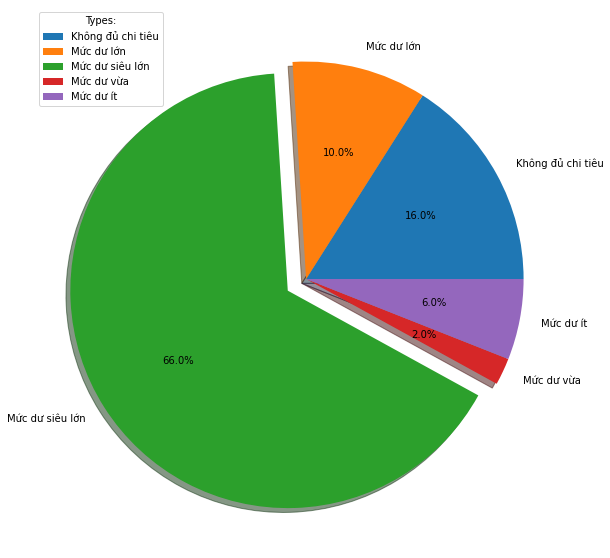
Biểu đồ hộp trình bày trực quan về thống kê mô tả của tập dữ liệu



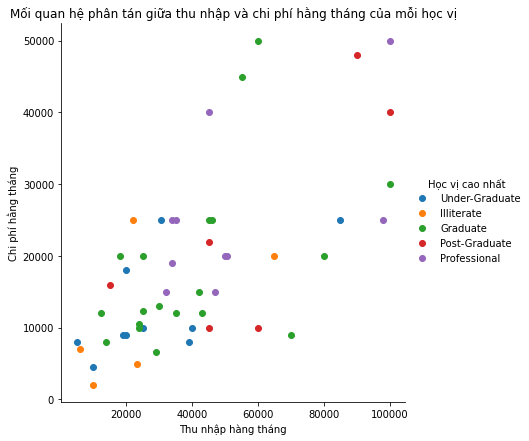
Tính số lượng phân chia mức sống trong tập dữ liệu



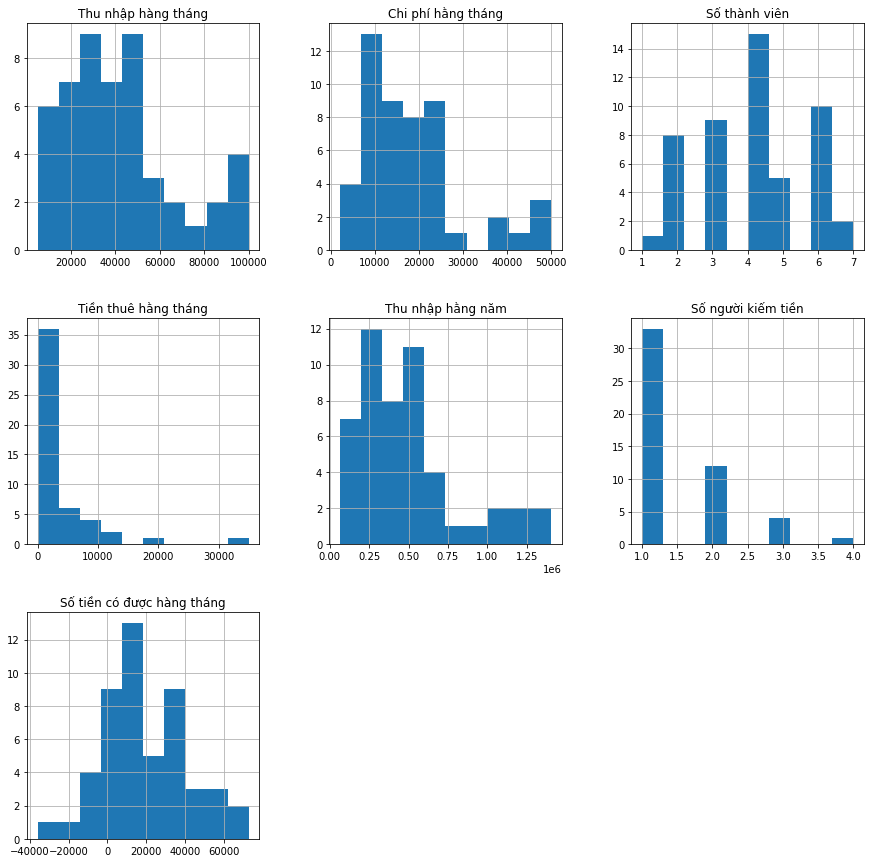
Tính tỉ lệ mức sống



Biểu đồ thể hiện sự phân tán giữa thu nhập và chi phí

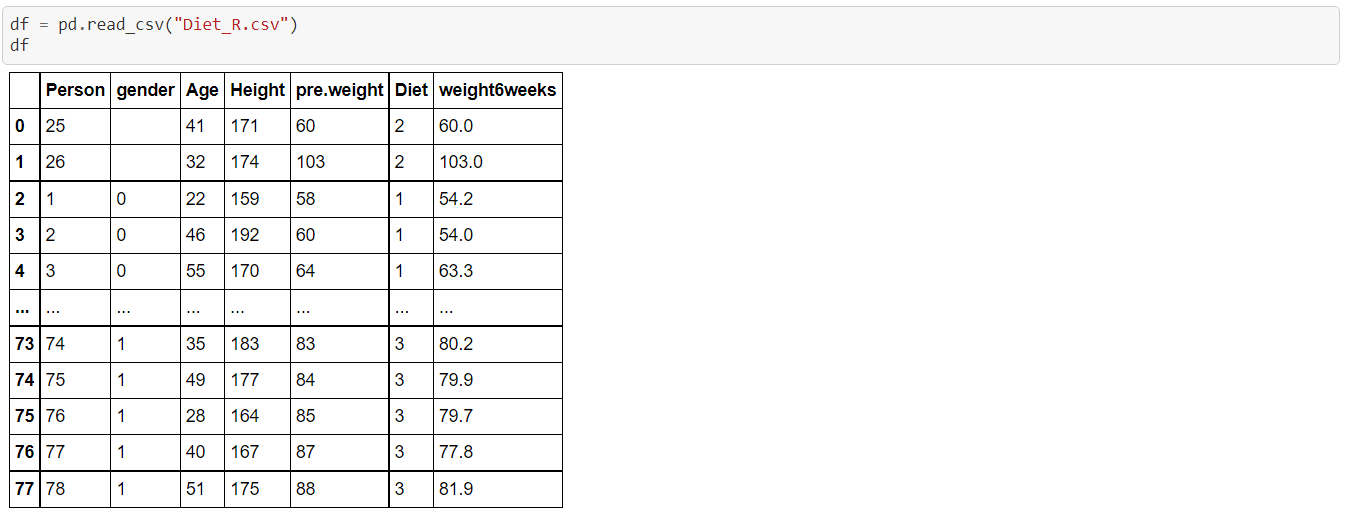


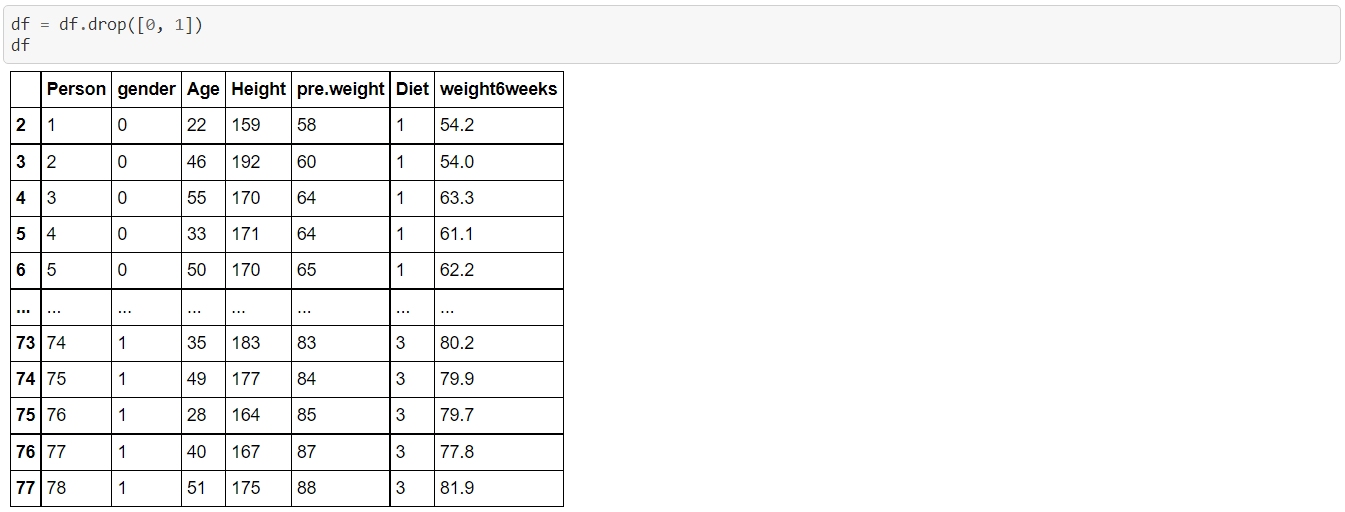
Biểu đồ histogram

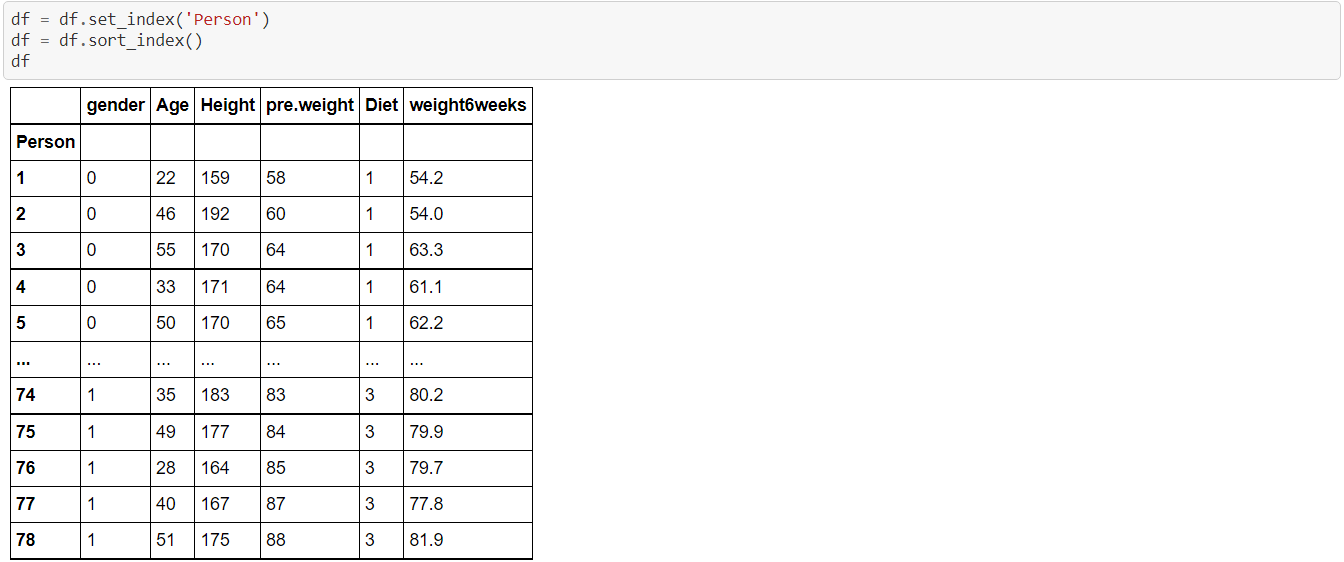


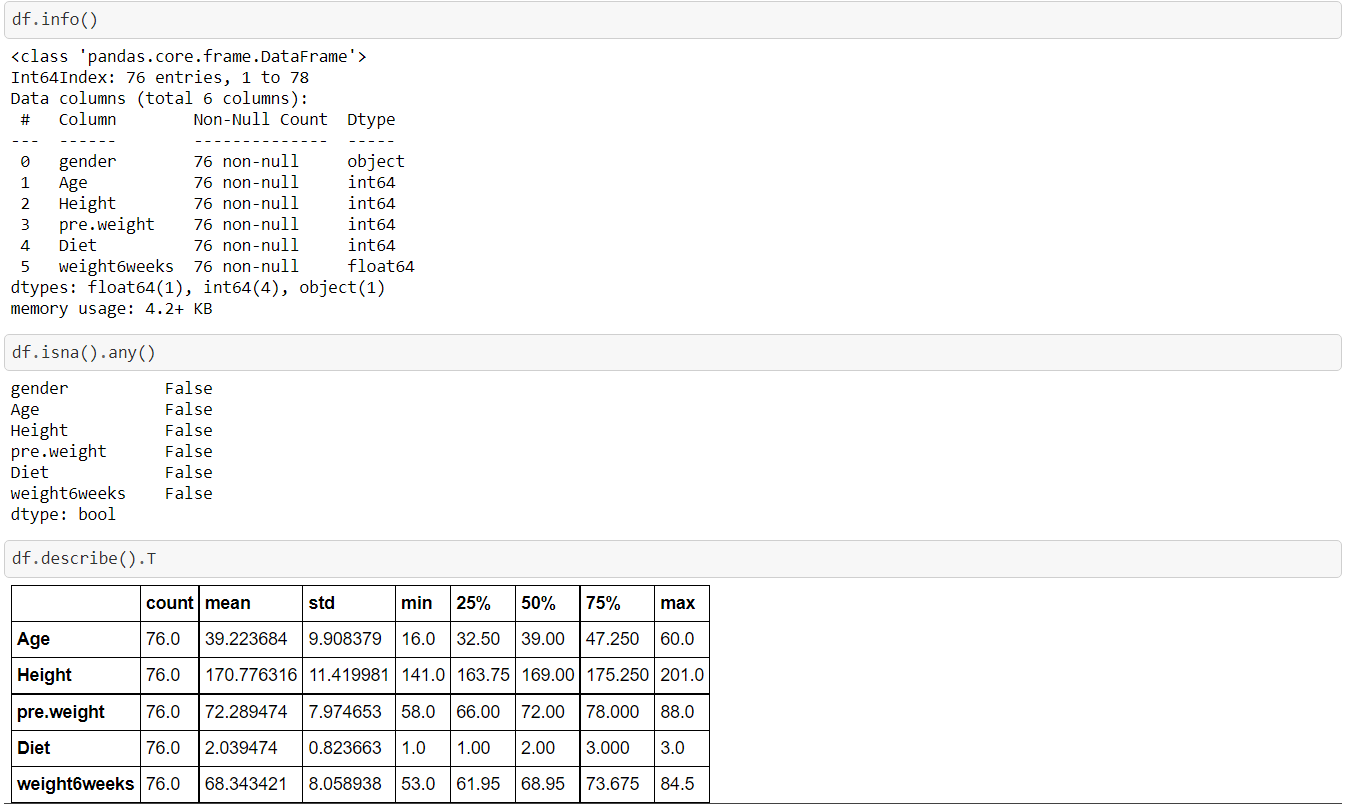
1. **Thống kê suy diễn:**

Đọc file, chỉnh sửa dữ liệu:









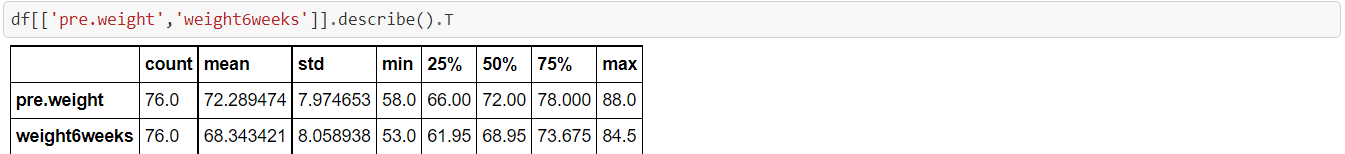
1. **Bỏ qua chế độ ăn kiêng và kiểm tra xem đã giảm cân chưa?**

Phát biểu giả thuyết:

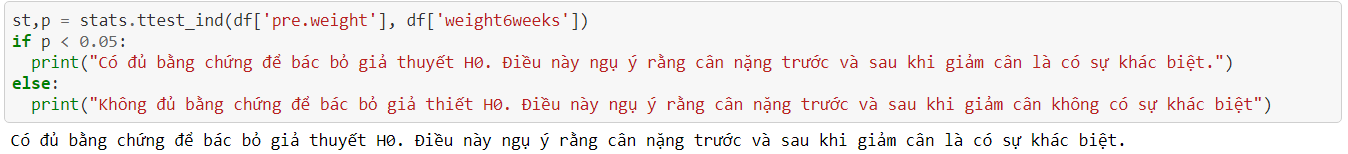
H0. Cân nặng trung bình của trước và sau khi giảm cân là như nhau

H1. Cân nặng trung bình của trước và sau khi giảm cân khác nhau

Phân tích 2 cột trọng lượng trước và sau



Sử dụng phương pháp Paired T test :

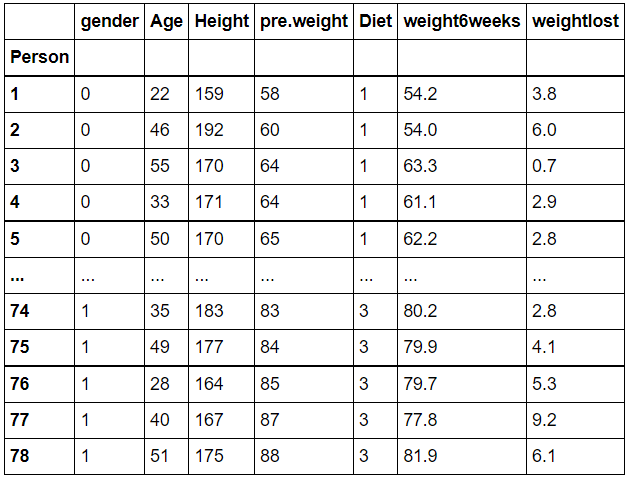


=> Cân nặng trung bình của trước và sau khi giảm cân có sự khác biệt.

1. **Loại bỏ trọng lượng đã mất và yêu cầu học sinh tính toán nó bằng cách**

**sử dụng trọng lượng trước trừ sau.**

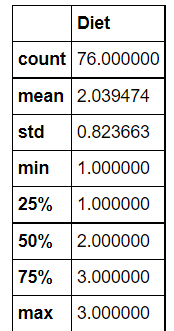
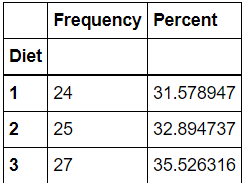




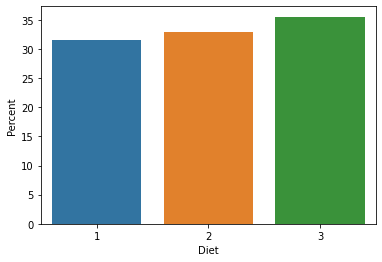
1. **Thống kê tóm tắt theo chế độ ăn uống**







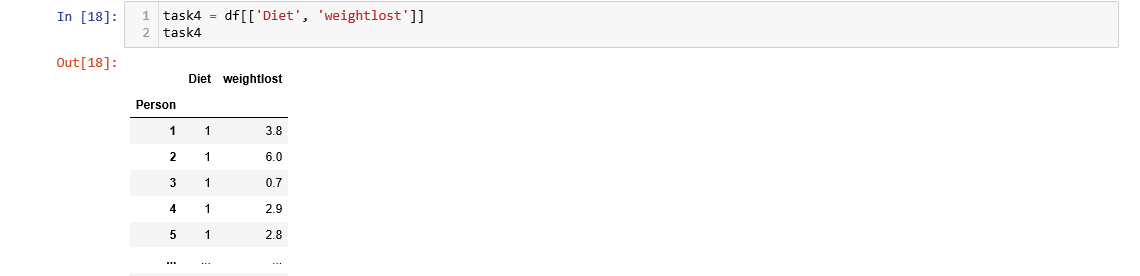




### Kiểm định ANOVA:

4.1: Phương pháp Diet hữu hiệu

Xây dựng giả thuyết:



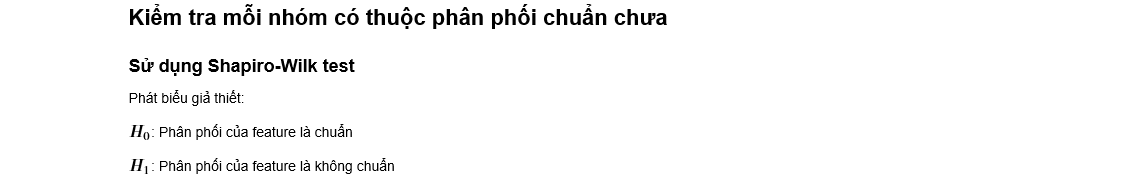
H0: u1= u2 = u3

H1: u1# u2 # u3

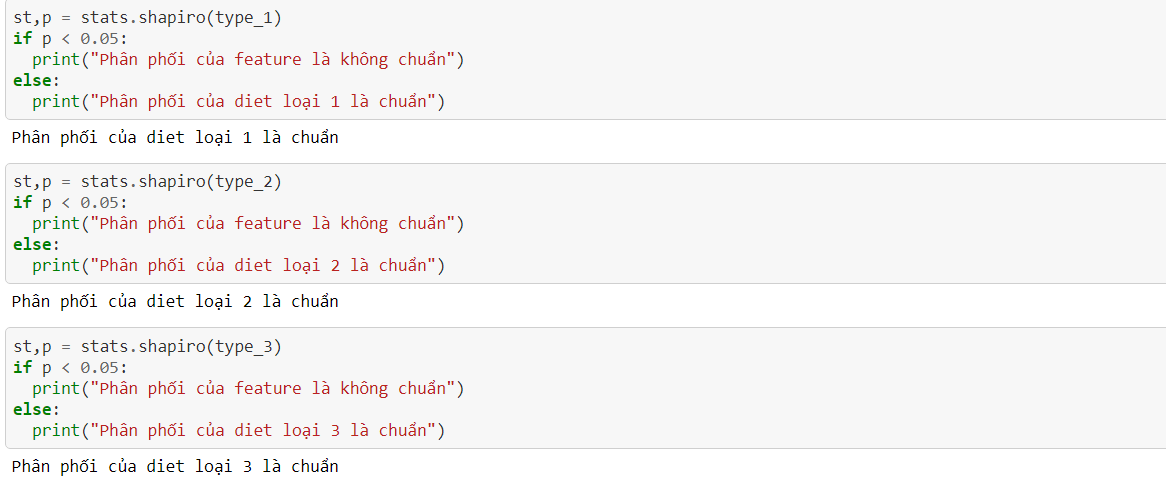
- Chú ý rằng, để thực hiện kiểm định ANOVA thì dữ liệu phải thỏa mãn các yêu cầu sau:

* 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.

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:



Tiến hành với từng phương pháp:



Kiểm định tính đồng nhất phương sai

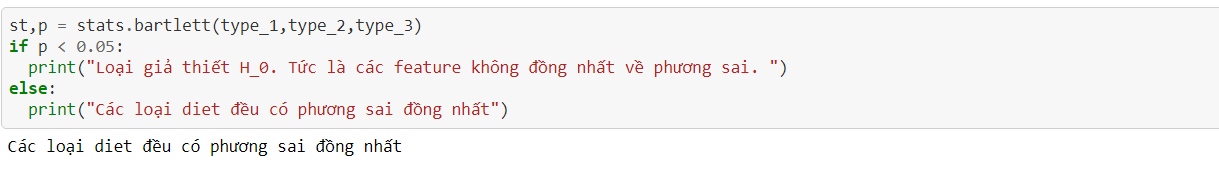
- Có nhiểu phương pháp để kiểm định tính đồng nhất phương sai như kiểm định Bartlett, kiểm định Levene, …

- Phát biểu giả thiết:

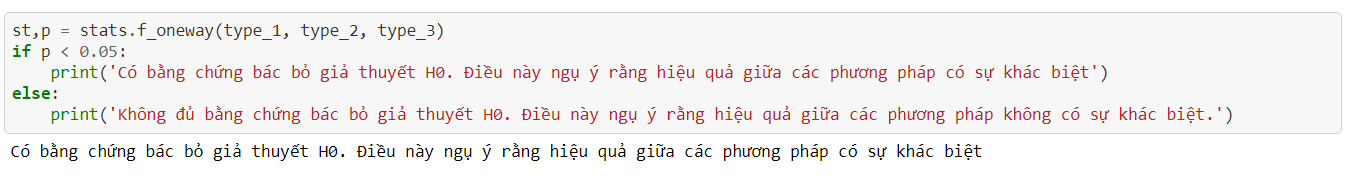
H0: Các feature đồng nhất phương sai.

H1: Các feature không đồng nhất phương sai.

Chú ý: Nếu p < 0.05 =  thì loại giả thiết. Tức là các feature không đồng nhất về phương sai.

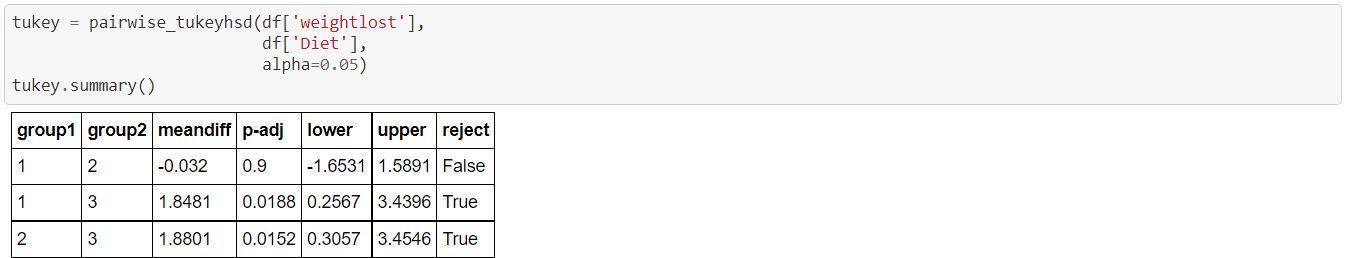


Phân tích ANOVA:



Với kết quả P value thu được, ta có đủ bằng chứng bác bỏ giả thuyết H0, điều này ngụ ý răng hiệu quả giữa các phương pháp là có sự khác biệt.

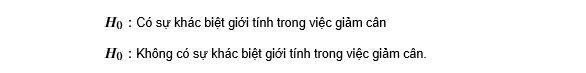
Hậu nghiệm ANOVA:



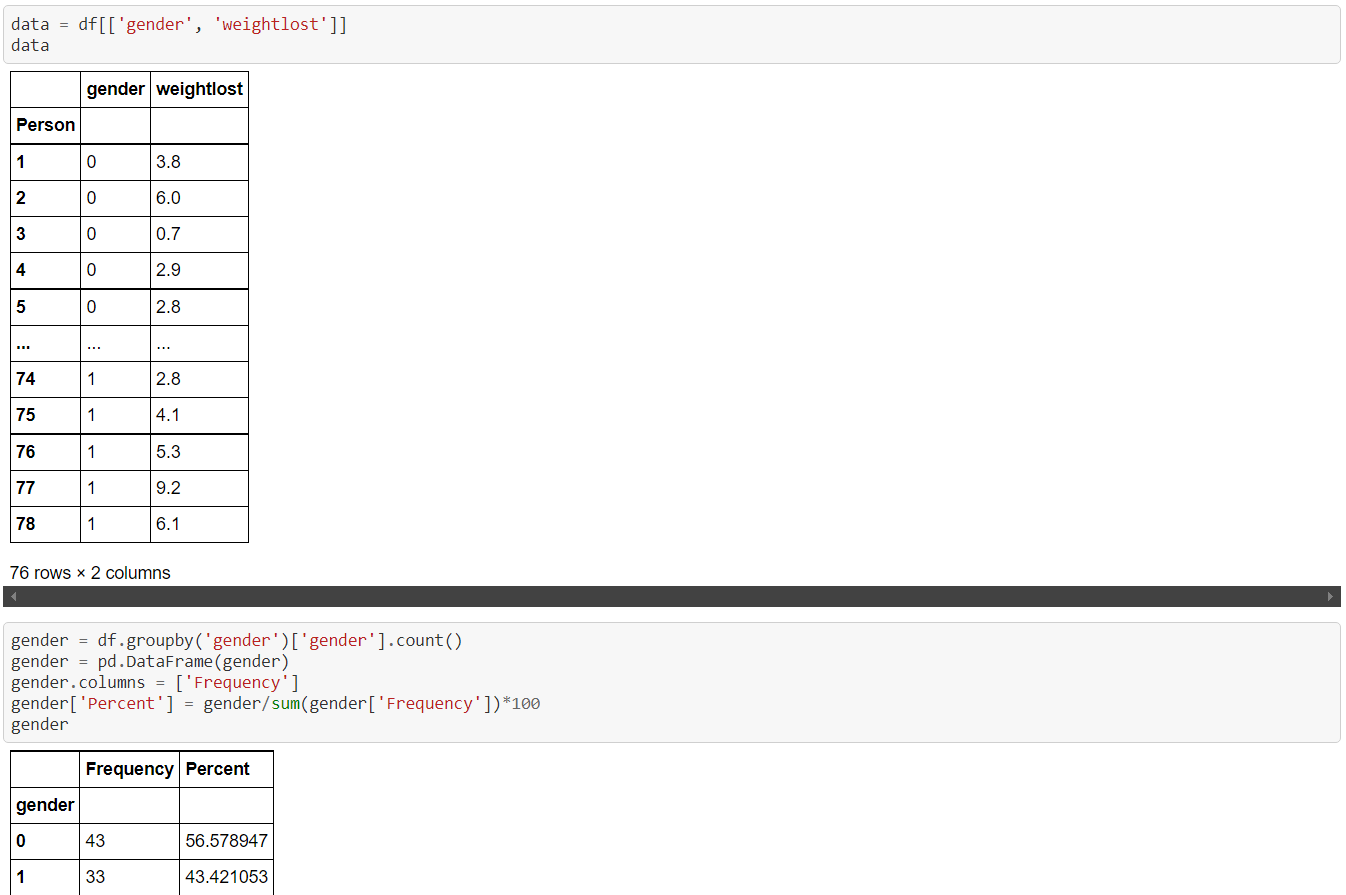
Với kết quả thu dược ta thấy, hiệu quả giữa phương pháp 1 và 2 không có nhiều khác biệt. Phương pháp 3 đem lại hiệu quả hơn 2 phương pháp còn lại

**4.2: Liệu hiệu quả giảm cân có ảnh hưởng đến giới tính không?**

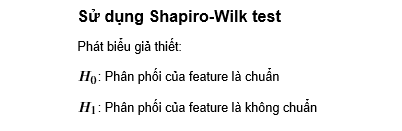
- Xây dựng giả thiết:

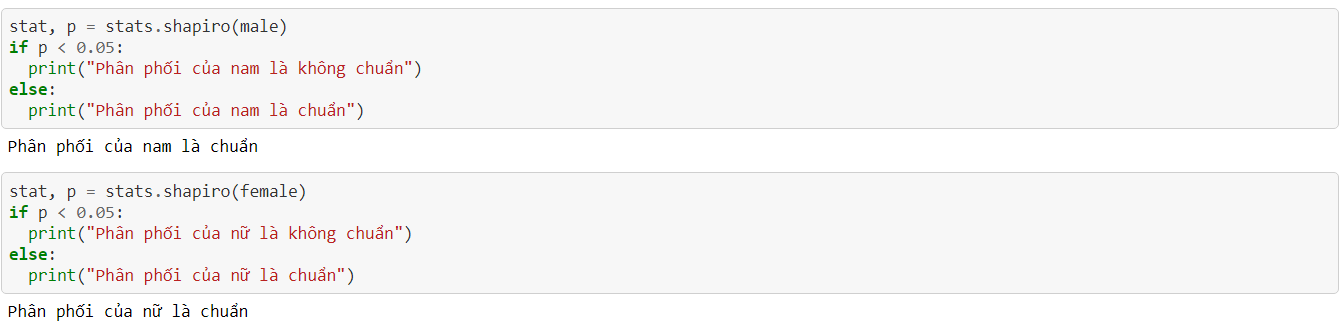


Để trả lời cho câu hỏi trên chúng ta tiếp tục thực hiện kiểm định ANOVA qua các bước trích xuất, kiểm tra, kiểm định dữ liệu và đưa ra kết luận.

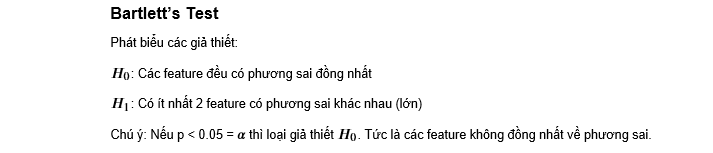


Kiểm tra mỗi nhóm thuộc phân phối chuẩn chưa:



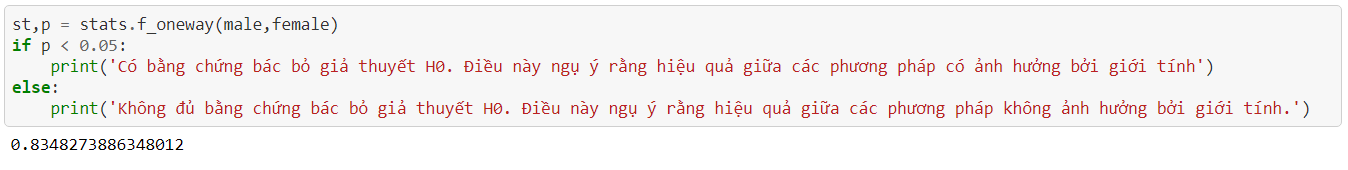


Kiểm định tính thuần nhất của phương sai:



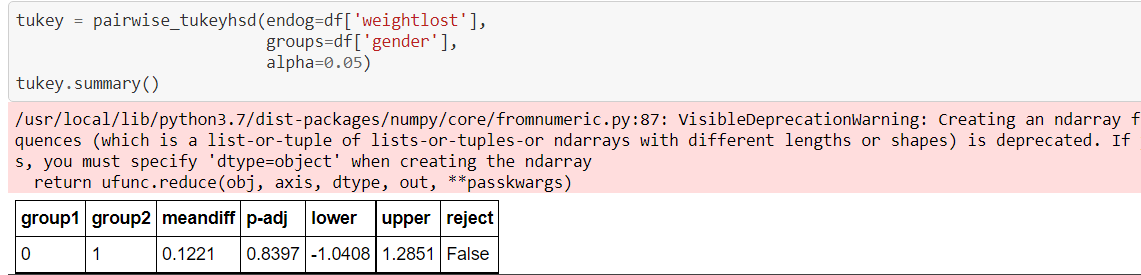


Phân tích ANOVA:

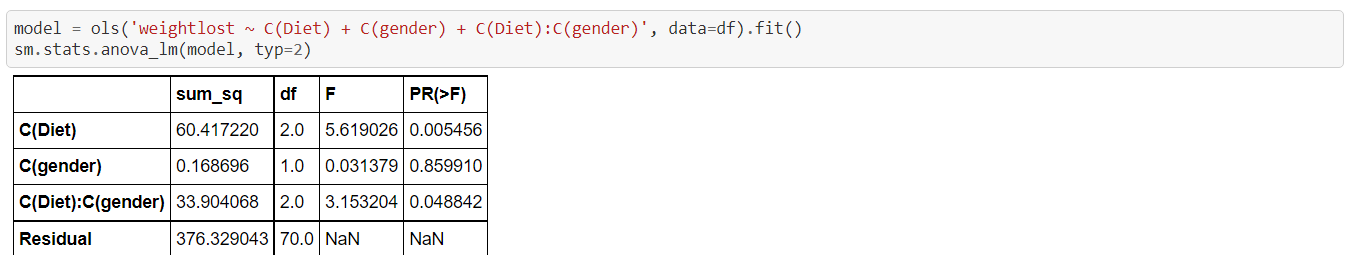


Với P value thu được ta không đủ bằng chứng bác bỏ giả thuyết H0, điều này ngụ ý răng hiệu quả giữa các phương pháp không bị ảnh hưởng bởi giới tính.

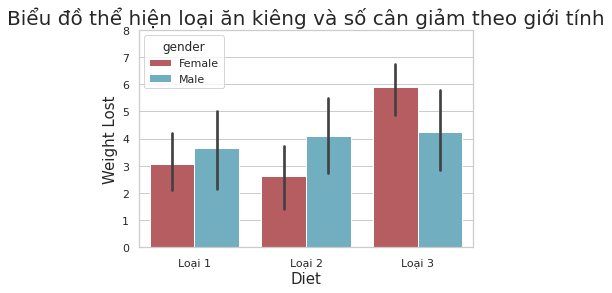
Hậu nghiệm ANOVA:



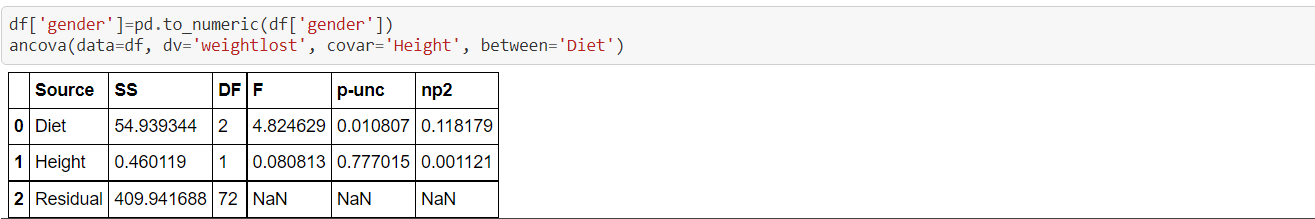
1. **Ảnh hưởng chế độ ăn kiêng và giới tính đến việc giảm cân:**



### **Tương tác**



1. **Ancova**



# CHƯƠNG III: TÀI LIỆU THAM KHẢO :

1.<https://vi.wikipedia.org/wiki/Python_(ng%C3%B4n_ng%E1%BB%AF_l%E1%BA%ADp_tr%C3%ACnh)>

2.<https://www.kaggle.com/sohier/tutorial-accessing-data-with-pandas/>

3.<https://www.kaggle.com/saurav9786/cardiogoodfitness-descriptive-statistics>

4.<https://www.datavedas.com/inferential-statistics-in-python/>

5. <https://gocnhinso.com/tim-hieu-truc-quan-hoa-du-lieu-data-visualization-la-gi/>

6. <https://vietnambiz.vn/he-so-bien-thien-coefficient-of-variation-cv-la-gi-nhung-dac-diem-can-luu-y-20191121233238319.htm>

# CHƯƠNG IV: TỈ LỆ ĐẠO VĂN

