**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: NGUYỄN TẤT DUY THÀNH**

*(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

[**KHOA KỸ THUẬT VÀ CÔNG NGHỆ 1**](#_Toc75618848)

[**ĐẠI HỌC HUẾ 2**](#_Toc75618849)

[**KHOA KỸ THUẬT VÀ CÔNG NGHỆ 2**](#_Toc75618850)

[**Mục lục 1**](#_Toc75618851)

[**Chương 1: Cơ sở lí thuyết 2**](#_Toc75618852)

[**1.1 Thống kê là gì? 2**](#_Toc75618853)

[**A/ Định nghĩa 2**](#_Toc75618854)

[**B/ Tầm quan trọng của thống kê 4**](#_Toc75618855)

[**1.2 Thực hiện thống kê trong Python như thế nào? 5**](#_Toc75618856)

[**A/ Giới thiệu về ngôn ngữ lập trình Python. 5**](#_Toc75618857)

[**B/ Giới thiệu về thư viện Pandas, Scipy, Matplotlyb. 6**](#_Toc75618858)

[**Chương 2: Phân tích dữ liệu 8**](#_Toc75618859)

[**2.1 Thống kê mô tả 8**](#_Toc75618860)

[**A/ Phân tích về sự phân phối của dữ liệu 9**](#_Toc75618861)

[**B/ Phân tích xu hướng tập trung của dữ liệu 13**](#_Toc75618862)

[**C/ Phân tích mức độ phân tán của dữ liệu 13**](#_Toc75618863)

[**2.2 Thống kê suy diễn 16**](#_Toc75618864)

[**A/ Kiểm định T 16**](#_Toc75618865)

[**B/ Kiểm định ANOVA 18**](#_Toc75618866)

[**Chương 3: Kiểm tra đạo văn 24**](#_Toc75618867)

[**Chương 4: Tài liệu tham khảo 24**](#_Toc75618868)

# Chương 1: Cơ sở lí thuyết

## Thống kê là gì?

### A/ Định nghĩa

- Thống kê là một nhánh của toán học ứng dụng liên quan đến việc thu thập, mô tả, phân tích và suy luận các kết luận từ dữ liệu định lượng. Các lý thuyết toán học đằng sau thống kê chủ yếu dựa vào phép tính vi phân và tích phân, đại số tuyến tính và lý thuyết xác suất.

- Trên thực tế, thống kê là ý tưởng chúng ta có thể tìm hiểu về các tính chất của bộ lớn của các đối tượng hoặc các sự kiện (một quần thể) bằng cách nghiên cứu các đặc điểm của một số nhỏ các đối tượng hoặc các sự kiện (tương tự mẫu ). Bởi vì trong nhiều trường hợp việc thu thập dữ liệu toàn diện về toàn bộ dân số là quá tốn kém, khó khăn hoặc không thể thực hiện được, thống kê bắt đầu với một mẫu có thể được quan sát một cách thuận tiện hoặc có thể chi trả được.

- 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.

- 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.

- 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.

#### a/ Thống kê mô tả

-Mô tả thống kê là hệ số mô tả ngắn gọn rằng tóm tắt một tập dữ liệu nhất định, đó có thể là một đại diện của toàn bộ hoặc một mẫu của một dân số. Thống kê mô tả được chia thành các thước đo về xu hướng trung tâm và các thước đo về độ biến thiên (lan truyền). Các phép đo xu hướng trung tâm bao gồm giá trị trung bình, giá trị trung bình và phương thức, trong khi các phép đo độ biến thiên bao gồm độ lệch, phương sai, các biến số tối thiểu và tối đa, [kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp) và [độ lệch](https://www.investopedia.com/terms/s/skewness.asp) .

- Thống kê mô tả được sử dụng để mô tả những đặc tính cơ bản của [dữ liệu](https://vi.wikipedia.org/wiki/D%E1%BB%AF_li%E1%BB%87u) thu thập được từ nghiên cứu thực nghiệm qua các cách thức khác nhau. Thống kê mô tả và [thống kê suy luận](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_suy_lu%E1%BA%ADn) cùng 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. Cùng với phân tích đồ họa đơn giản, chúng tạo ra nền tảng của mọi phân tích định lượng về số liệu. Để hiểu được các hiện tượng và ra quyết định đúng đắn, cần nắm được các phương pháp cơ bản của mô tả dữ liệu.[[1]](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_m%C3%B4_t%E1%BA%A3#cite_note-1) Có rất nhiều kỹ thuật hay được sử dụng. Có thể phân loại các kỹ thuật này 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.

Khi tạo các trị thống kê mô tả, người ta có thể nhằm 2 mục tiêu:

* Chọn một [trị thống kê](https://vi.wikipedia.org/w/index.php?title=Tr%E1%BB%8B_th%E1%BB%91ng_k%C3%AA&action=edit&redlink=1) để chỉ ra những đơn vị có vẻ giống nhau thực ra có thể khác nhau thế nào. Các giáo trình thống kê gọi một giải pháp đáp ứng mục tiêu này là thước đo [khuynh hướng trung tâm](https://vi.wikipedia.org/w/index.php?title=Khuynh_h%C6%B0%E1%BB%9Bng_trung_t%C3%A2m&action=edit&redlink=1).
* Chọn một trị [thống kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) khác cho thấy các đơn vị khác nhau thế nào. Loại trị thống kê này thường được gọi là một thước đo [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).

Khi tóm tắt một lượng như độ dài, cân nặng hay tuổi tác, nói chung người ta hay dùng các trị thống kê như số [trung bình cộng](https://vi.wikipedia.org/wiki/Trung_b%C3%ACnh_c%E1%BB%99ng), [trung vị](https://vi.wikipedia.org/wiki/S%E1%BB%91_trung_v%E1%BB%8B); hay trong trường hợp một phân bố đơn mốt (mode - số trung phương), người ta thường dùng [mốt](https://vi.wikipedia.org/wiki/Mode_(th%E1%BB%91ng_k%C3%AA)). Đôi khi, người ta chọn lựa những giá trị đặc thù từ [hàm phân bố tích lũy](https://vi.wikipedia.org/wiki/H%C3%A0m_ph%C3%A2n_ph%E1%BB%91i_t%C3%ADch_l%C5%A9y) gọi là các [tứ phân vị](https://vi.wikipedia.org/wiki/T%E1%BB%A9_ph%C3%A2n_v%E1%BB%8B).

Các thước đo chung nhất về mức độ phân tán của dữ liệu lượng là [phương sai](https://vi.wikipedia.org/wiki/Ph%C6%B0%C6%A1ng_sai), giá trị căn bậc 2 của nó, tức là [độ lệch chuẩn](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%99_l%E1%BB%87ch_chu%E1%BA%A9n); [khoảng](https://vi.wikipedia.org/w/index.php?title=Kho%E1%BA%A3ng_(th%E1%BB%91ng_k%C3%AA)&action=edit&redlink=1); [khoảng cách giữa các tứ phân vị](https://vi.wikipedia.org/w/index.php?title=Kho%E1%BA%A3ng_c%C3%A1ch_gi%E1%BB%AFa_c%C3%A1c_t%E1%BB%A9_ph%C3%A2n_v%E1%BB%8B&action=edit&redlink=1); và [độ lệch bình quân tuyệt đối](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BB%99_l%E1%BB%87ch_b%C3%ACnh_qu%C3%A2n_tuy%E1%BB%87t_%C4%91%E1%BB%91i&action=edit&redlink=1).

Khi thực hiện một trình diễn đồ họa để tóm tắt một bộ dữ liệu, cũng có thể áp dụng cả hai mục tiêu nói trên. Một ví dụ đơn giản về kỹ thuật đồ họa là [đồ thị phân bố](https://vi.wikipedia.org/w/index.php?title=%C4%90%E1%BB%93_th%E1%BB%8B_ph%C3%A2n_b%E1%BB%91&action=edit&redlink=1), thứ đồ thị phơi bày cả khuynh hướng trung tâm lẫn độ [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).

#### b/ Thống kê suy diễn

- Thống kê suy diễn là công cụ mà nhà thống kê sử dụng 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.

- Dựa trên kích thước mẫu và sự phân bố của dữ liệu mẫu, các nhà thống kê có thể tính toán xác suất để các thống kê đo lường xu hướng trung tâm, độ biến thiên, phân bố và mối quan hệ giữa các đặc điểm trong mẫu dữ liệu, cung cấp bức tranh chính xác về các thông số tương ứng của toàn bộ tổng thể từ đó mẫu được rút ra.

- Thống kê suy luận được sử dụng để khái quát hóa về các nhóm lớn, chẳng hạn như ước tính nhu cầu trung bình đối với một sản phẩm bằng cách khảo sát mẫu thói quen mua hàng của người tiêu dùng hoặc để cố gắng dự đoán các sự kiện trong tương lai, chẳng hạn như dự đoán lợi nhuận trong tương lai của một loại chứng khoán hoặc tài sản về lợi nhuận trong một khoảng thời gian mẫu.

### 

### B/ Tầm quan trọng 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.

- Lĩnh vực chuyên môn: Các kỹ thuật thống kê được sử dụng trong một loạt các nghiên cứu khoa học và xã hội, bao gồm: ngành sinh học, tính toán sinh học, tính toán xã hội học, hệ thống sinh học, khoa học xã hội và nghiên cứu xã hội. Những ngành này bao gồm:

* Khoa học tính toán bảo hiểm
* Ứng dụng thông tin kinh tế
* Thiên văn học
* 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

- Thống kê áp dụng cho toán học hay nghệ thuật

## 1.2 Thực hiện thống kê trong Python như thế nào?

### A/ Giới thiệu về ngôn ngữ lập trình Python.

- Python là một ngôn ngữ lập trình bậc cao cho các mục đích lập trình đa năng, do Guido van Rossum tạo ra và lần đầu ra mắt vào năm 1991.

- Python được thiết với ưu điểm là dễ đọc, dễ học và dễ nhớ. Được phát triển theo phương châm là cấu trúc rõ ràng, hình thức sáng sủa, thuận tiện cho người dùng. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu.

- Ban đầu, Python được phát triển để chạy trên nền Unix. Nhưng rồi theo thời gian, Python dần mở rộng sang mọi hệ điều hành từ MS-DOS đến Mac OS, OS/2, Windows, Linux và các hệ điều hành khác thuộc họ Unix.

### 

### B/ Giới thiệu về thư viện Pandas, Scipy, Matplotlyb.

#### a/ 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.

#### b/ Thư viên Scipy

-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. Với SciPy, một phiên Python tương tác trở thành một hệ thống đối thủ của môi trường xử lý dữ liệu và tạo mẫu hệ thống, chẳng hạn như MATLAB, IDL, Octave, R-Lab và SciLab.

-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.

#### c/ 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

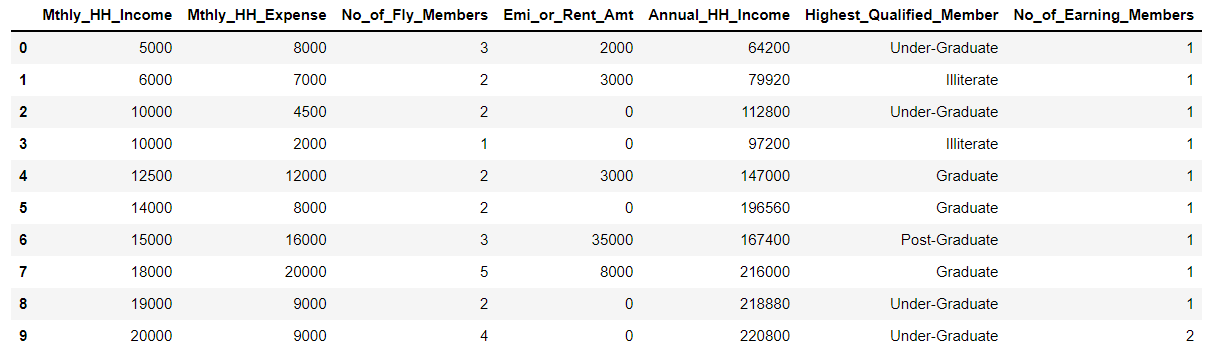
## 2.1 Thống kê mô tả

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

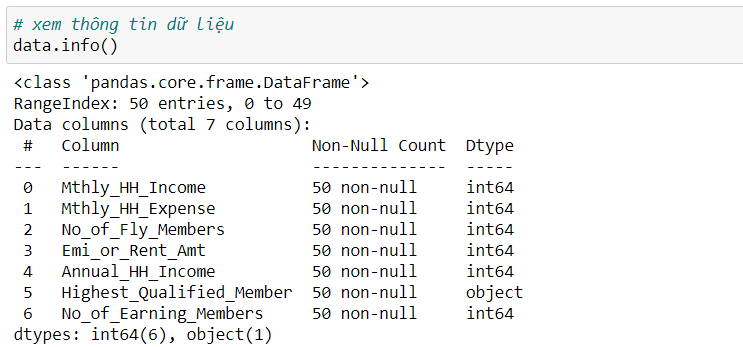
- Đầu tiên truyền file dữ liệu vào Python bằng cách dùng chức năng của **read\_csv** của thư viên Pandas.



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



- Để xem thông tin tổng quát của tập dữ liệu, ta sử dụng hàm **info()** trong Pandas. Hàm này cung cấp thông tin về các cột trong DataFrame, đếm xem trong cột có bao nhiêu giá trị khác non-available, loại dữ liệu, …



- Như vậy dữ liệu của chúng ta gồm có 7 cột được đánh số thứ từ 0-6 gồm:

* Mthly\_HH\_Income
* Mthly\_HH\_Expense
* No\_of\_Fly\_Members
* Emi\_or\_Rent\_Amt
* Annual\_HH\_Income
* Highest\_Qualified\_Member
* No\_of\_Earning\_Members

Và không có cột nào bị thiếu dữ liệu, loại dữ liệu gồm object (cột 5) và int64 (còn lại)

- Sau khi có một cái nhìn tổng quát thông tin dữ liệu chúng ta đi vào phân tích dữ liệu

### A/ Phân tích về sự phân phối của dữ liệu

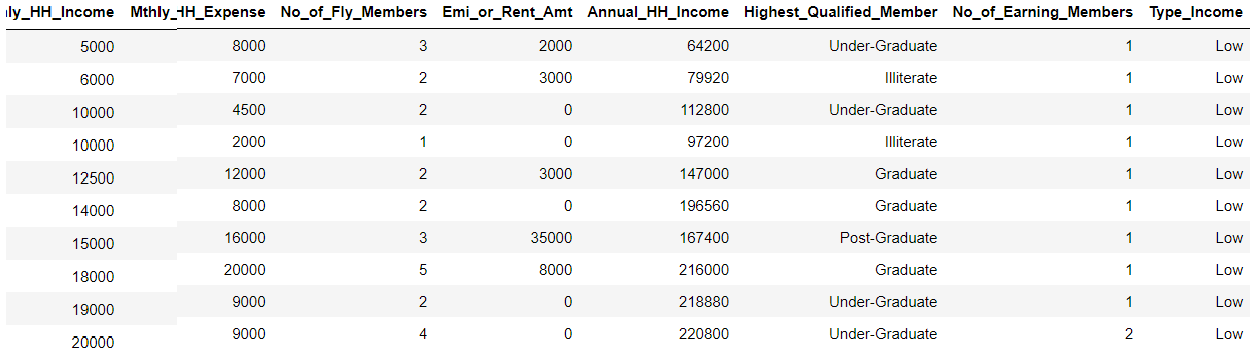
- Phân phối là bản tóm tắt tần suất của các giá trị riêng lẻ hoặc phạm vi giá trị cho một biến. Phân phối đơn giản nhất sẽ liệt kê mọi giá trị của một biến và số người có từng giá trị.

- Trong trường hợp này, chúng ta nhóm giá trị thu nhập thành các loại theo phạm vi giá trị. Ví dụ, chúng ta có thể xem xét với thu nhập dưới 20000$ thuộc loại thu nhập thấp hoặc trên 80000$ thuộc loại thu nhập khủng,...

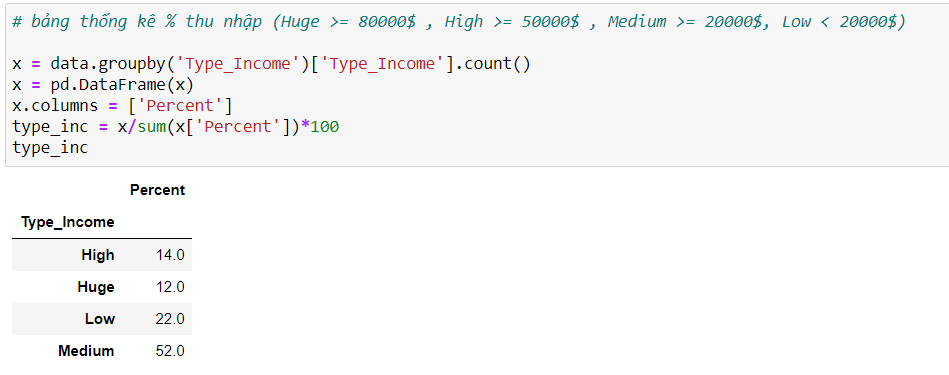
- Để thuận tiện cho việc phân tích dữ liệu, ta thêm vào một cột có tên là “Type” biểu diễn loại thu nhập.



- Chúng ta được kết quả sau đây:

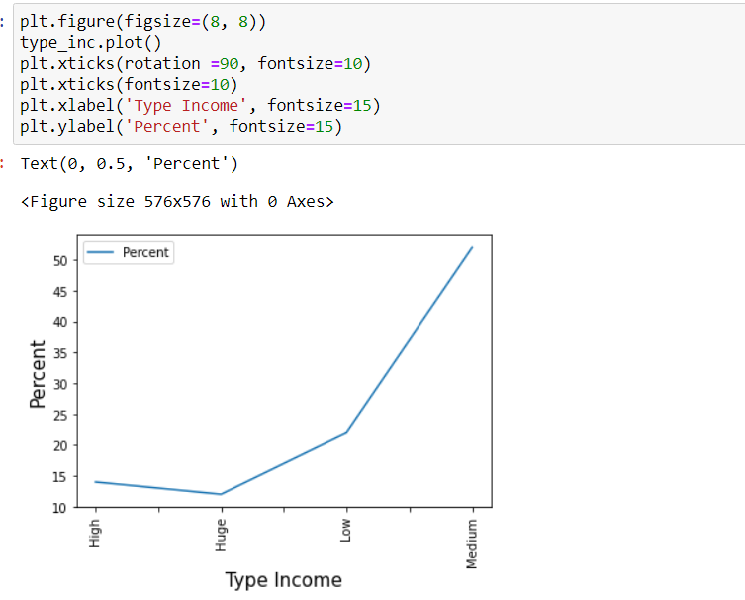


- Giả sử chúng ta chỉ quan tâm đến thu nhập hàng tháng của mỗi cá thể trong bảng dữ liệu và xem xét chúng thuộc nhóm thu nhập nào và tỉ lệ mỗi nhóm là bao nhiêu?



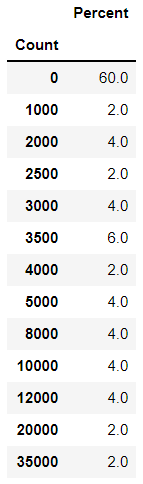
- Chúng ta đã có bảng dữ liệu đánh giá được tỉ lệ phần trăm các mức thu nhập trong dữ liệu. Nhận thấy rằng phần lớn thu nhập của người dân rơi vào khoảng thu nhập trung bình từ 20000$ - 50000$ với tỉ lệ là 52% và thấp nhất với thu nhập khủng trên 80000$ chiếm 12% tổng số dữ liệu.

- Để có một cái nhìn tốt hơn chúng ta sẽ biểu diễn bảng trên bằng đồ thị sử dụng thư viện **Matplotlyb** như sau:

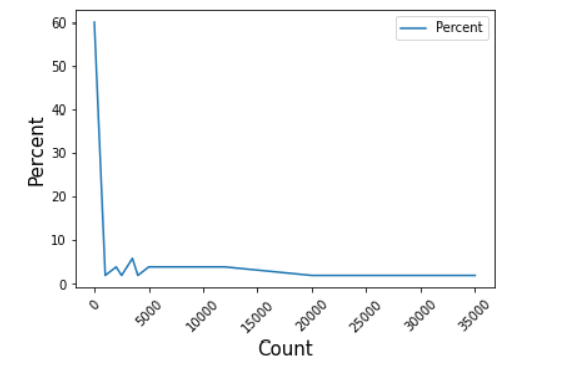


**-** Chúng ta có thể cải tiến để biểu đồ đẹp hơn bằng cách tham khảo các thư viện hỗ trợ vẽ biểu đồ như Matplotlyb, Seaborn, Altair, GGplot2 ,…

- Tiếp chúng ta phân tích cột dữ liệu “Emi\_or\_Rent\_Amt” để quan sát mức thuế người dân phải đóng như thế nào?

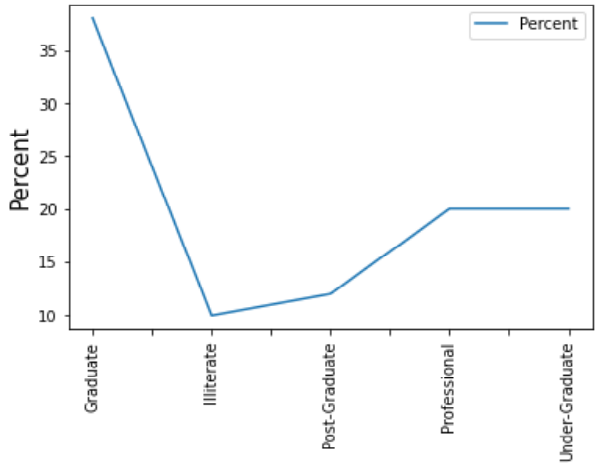
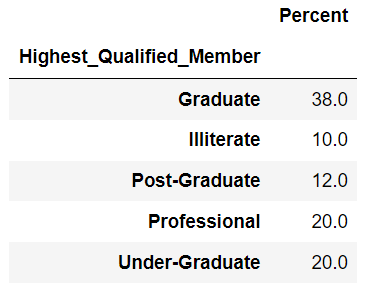


=> Có 60% trên tổng số được miễn nộp thuế, mức thuế cao nhất là 35000$ và có 2% người nộp thuế.



=> Đồ thị thể hiện tỉ lệ nộp thuế.

- Hoặc một trường dữ liệu khác có tên là “Highest\_Qualified\_Member”



### B/ Phân tích xu hướng tập trung của dữ liệu

- Xu hướng trung tâm của phân phối là ước tính về “trung tâm” của phân phối giá trị. Có ba loại ước tính chính về xu hướng trung tâm:

* Trung bình - Mean
* Trung vị - Median
* Yếu vị - Mode

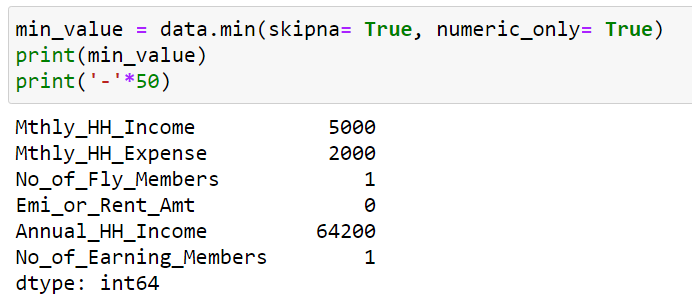
- Để tính các giá trị trung bình, trung vị, yếu vị của của dữ liệu, chúng ta sử dụng các hàm **mean()**, **median()**, **mode()** được tích hợp sẵn trong Pandas

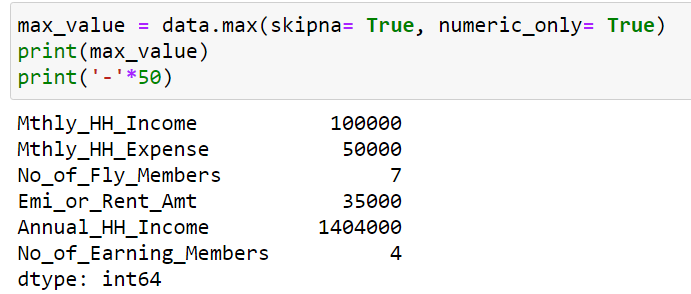
### C/ Phân tích mức độ phân tán của dữ liệu

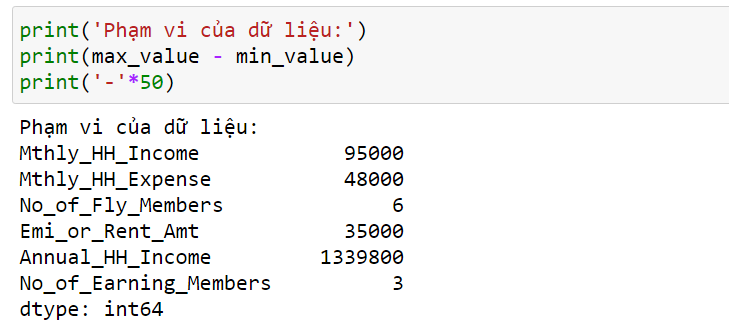
- Sự phân tán đề cập đến sự lan truyền của các giá trị xung quanh xu hướng trung tâm. Có hai thước đo phổ biến của độ phân tán, phạm vi và độ lệch chuẩn:

* Phạm vi - Range:
* Độ lệch chuẩn - Standard Deviation

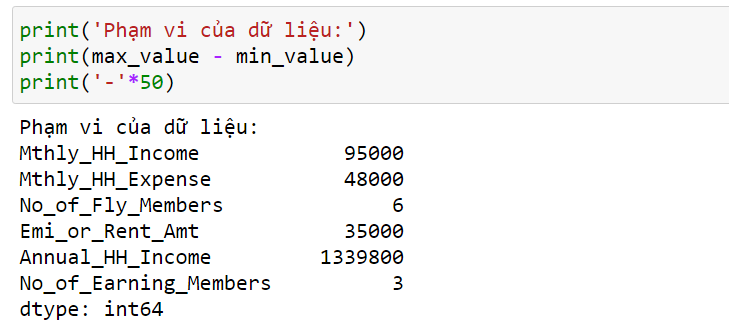
- Tính phạm vi của dữ liệu





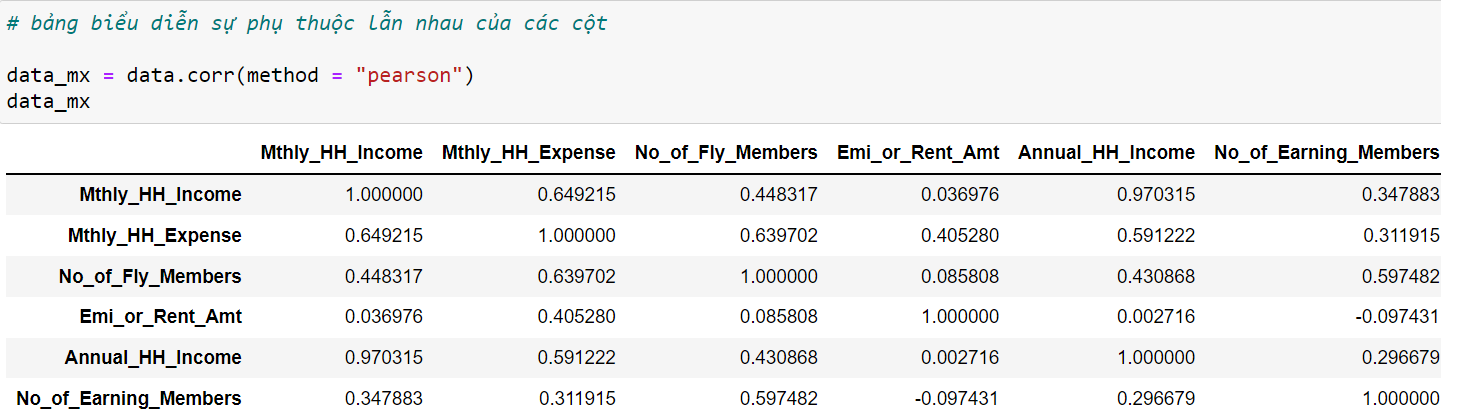


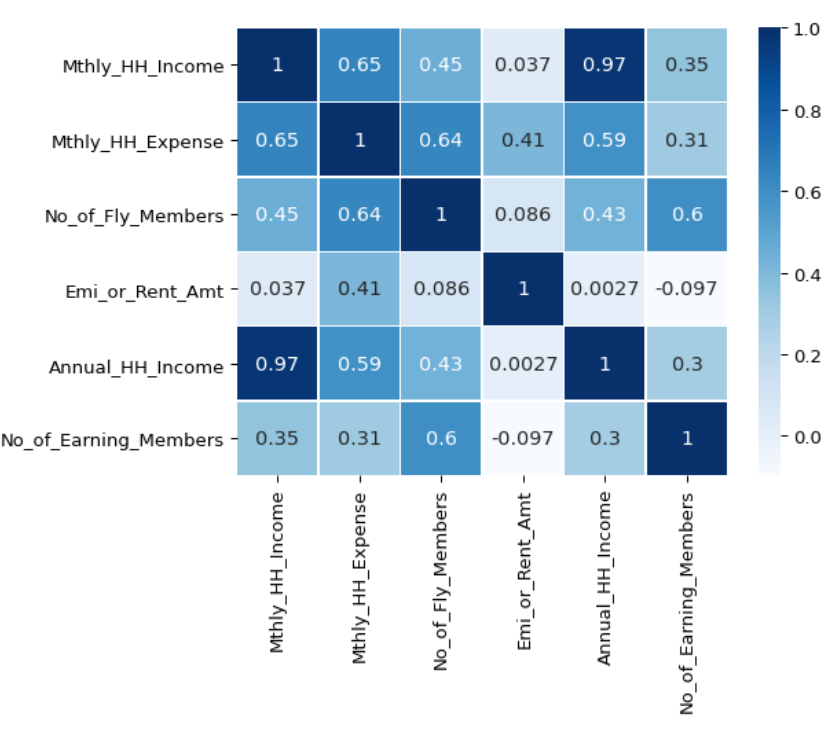
- Tính độ lệch chuẩn của dữ liệu



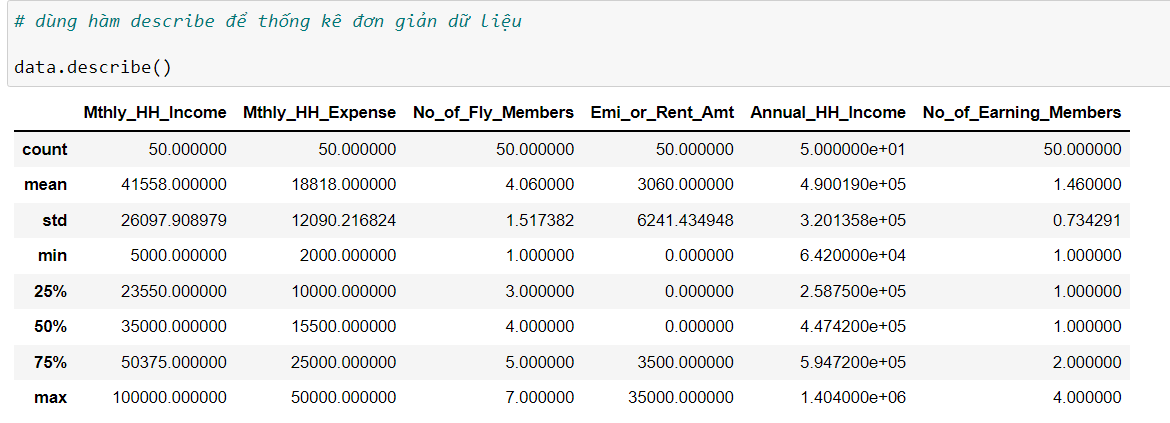
- Bảng thống kê thể hiện mức độ phụ thuộc lẫn nhau trong dữ liệu. Mức độ phụ thuộc được thể hiện bằng giá trị trong bảng nằm trong khoảng [-1, 1]. Giá trị càng lớn chứng tỏ mức độ phụ thuộc lẫn nhau là lớn và ngược lại

- Sử dụng hàm tích hợp **corr()** trong Pandas để tính toán





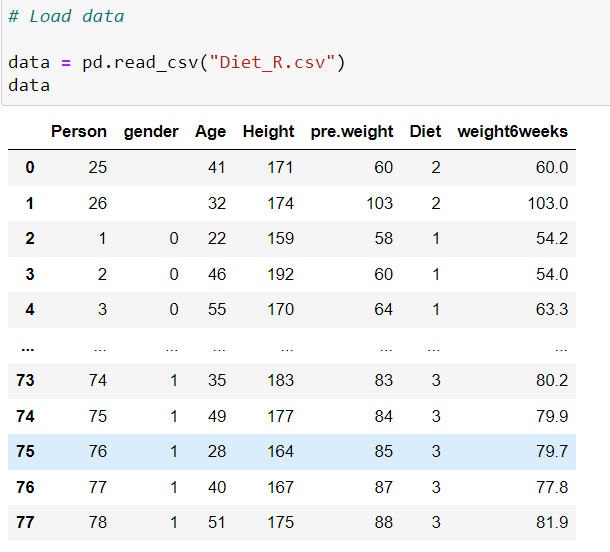
- Sử dụng hàm **describe()** để có một bảng tóm tắt những thống kê cơ bản của dữ liệu. Hàm tính toán các giá trị như số lượng cá thể khảo sát, trung bình, độ lệch chuẩn, min, max, bách phân vị, …



## 2.2 Thống kê suy diễn

- Trong phần này, chúng ta sẽ sử dụng tệp dữ liệu “Diet\_R.csv” để thực hiện thống kê suy diễn.

- Chúng ta cũng khai báo các thư viện và sử dụng hàm **read\_csv** để load dữ liệu vào Python.



### A/ Kiểm định T

- Trong phần này, chúng ta sẽ sử dụng tệp dữ liệu “Diet\_R.csv” để thực hiện thống kê suy diễn.

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

- Giả sử chúng ta cần kiểm định rằng cân nặng của mọi người trước và sau khi thực hiện ăn kiêng có hiệu quả hay không?

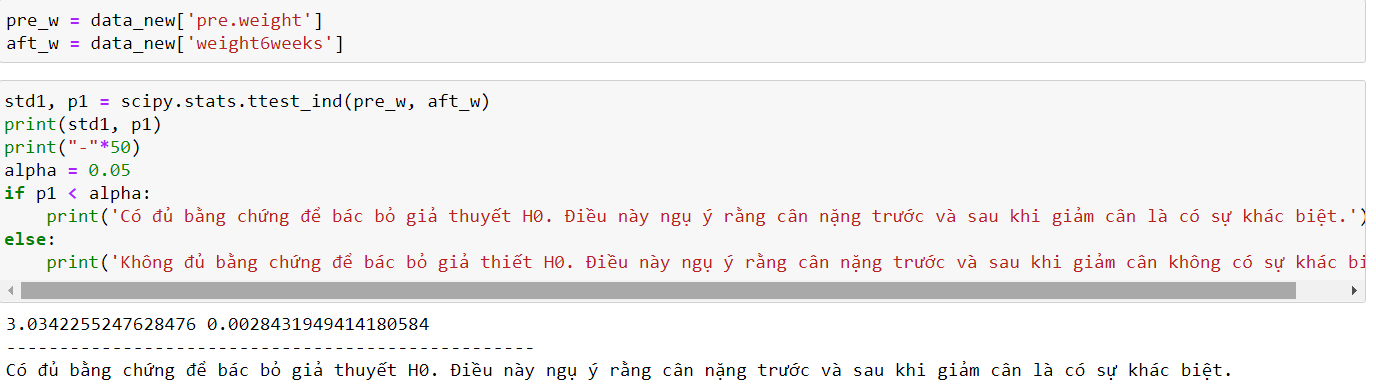
- Vậy chúng ta có giả thiết không và đối thiết như sau:

* H0: Cân nặng trung bình của mọi người trước và sau khi ăn kiêng là như nhau.
* H1: Cân nặng trung bình của mọi người trước và sau khi ăn kiêng khác nhau.

- Để kiểm chứng giả thiết đã nêu đúng hay sai chúng ta cần trích xuất dữ liệu từ DataFrame và thực hiện kiểm định t và đưa ra kết luận.

- Trích xuất dữ liệu của cột “pre.weight” và “weight6weeks”.

- Thực hiện kiểm định t bằng cách sử dụng module **ttest\_ind()** tích hợp trong thư viện scipy.



- Như vậy với mức tin cậy là 5% chúng ta có đủ tự tin để khẳng định rằng sau khi thực hiện phương pháp giảm cân là có hiệu quả.

- Để trực quan hơn chúng ta có thể vẽ biểu đồ cho dữ liệu này.



- Qua biểu đồ, chúng ta gần như toàn bộ biểu đồ “weight6weeks” (chỉ số cân nặng sau khi thực hiện giảm cân) đều nằm phía dưới “pre.weight” (chỉ số cân năng ban đầu). Điều đó chứng minh rằng phương pháp giảm cân có hiệu quả

### B/ Kiểm định ANOVA

- Sau khi chứng minh được rằng phương pháp giảm cân có hiệu quả thì có một vấn đề phát sinh đó chính là chúng ta muốn biết rằng phương pháp nào hiệu quả tốt nhất, phương pháp nào hiệu quả thấp nhất hay cả 3 phương pháp đều cho hiệu quả như nhau, liệu rằng giới tính có ảnh hưởng đến quá trình giảm cân hay không?, …

- Để trả lời những câu hỏi trên chúng ta sử dụng phương pháp phân tích phương sai - ANalysis of VAriance - ANOVA.

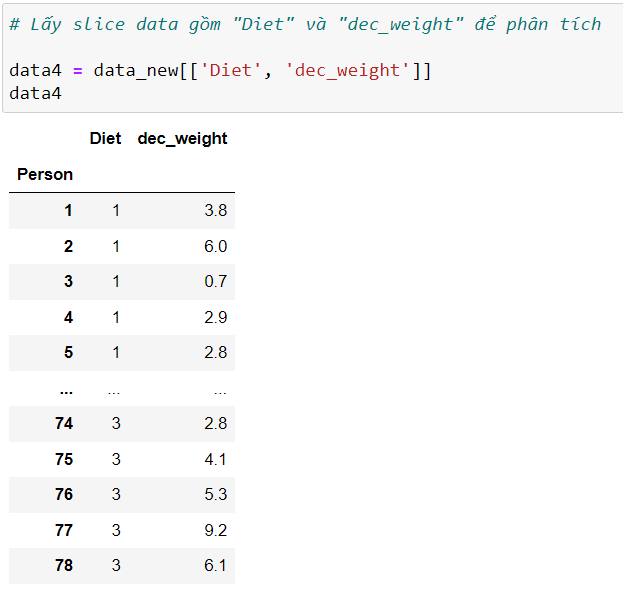
**Câu hỏi 1: Có sự khác biệt hiệu quả giữa các phương pháp giảm cân hay không?**

- Xây dựng giả thiết:

* 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 thiết chúng ta cần trích xuất dữ liệu để phân tích, phân loại dữ liệu và tiến hành phân tích ANOVA

- Trích xuất và phân loại dữ liệu:



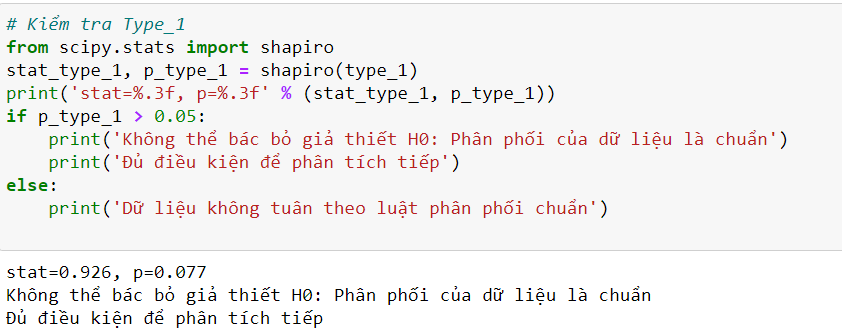
- Sử dụng hàm **groupby.groups** để phân loại dữ liệu, cụ thể ở đây chúng ta có 3 phương pháp ăn kiêng lần lượt là 1, 2, 3.

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

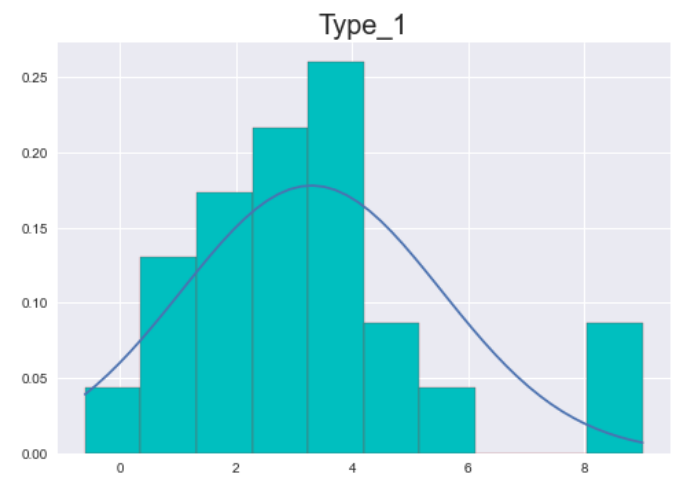
* Dữ liệu phải tuân theo phân phối chuẩn
* Dữ liệu đồng nhất về phương sai
* Các cá thể phải được lấy độc lập với nhau

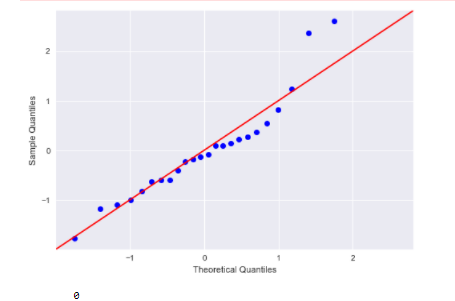
Vì vậy sau khi trích xuất dữ liệu để phân tích chúng ta cần phải tiến hành kiểm tra dữ liệu có đạt những yêu câu trên hay không.

- Để kiểm tra dữ liệu có tuân theo phân phối chuẩn hay không, chúng ta sử dụng các kiểm định như Shapiro-Wilk, kiểm định D’Agostino, kiểm định Anderson-Darling, … Và nhiều kiểm định khác để kiểm tra. Ở đây chúng ta sẽ sử dụng kiểm định Shapiro-Wilk được tích hợp sẵn trong thư viện Scipy



- Chúng ta có thể vẽ biểu đồ để quan sát rõ hơn. Dữ liệu tuân theo phân phối chuẩn sẽ có dạng **bell-shape**.





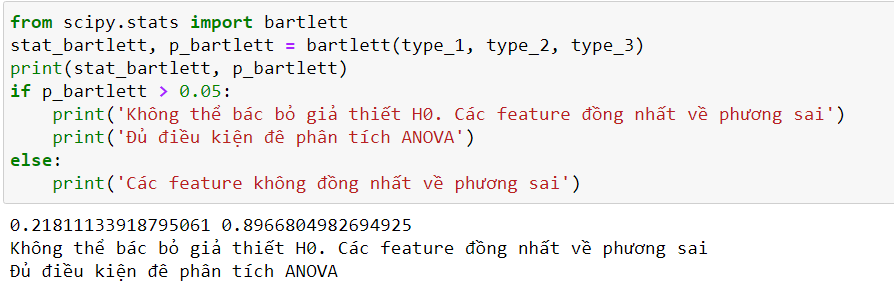
Kiểm định tính đồng nhất phương sai

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

- Phát biểu giả thiết:

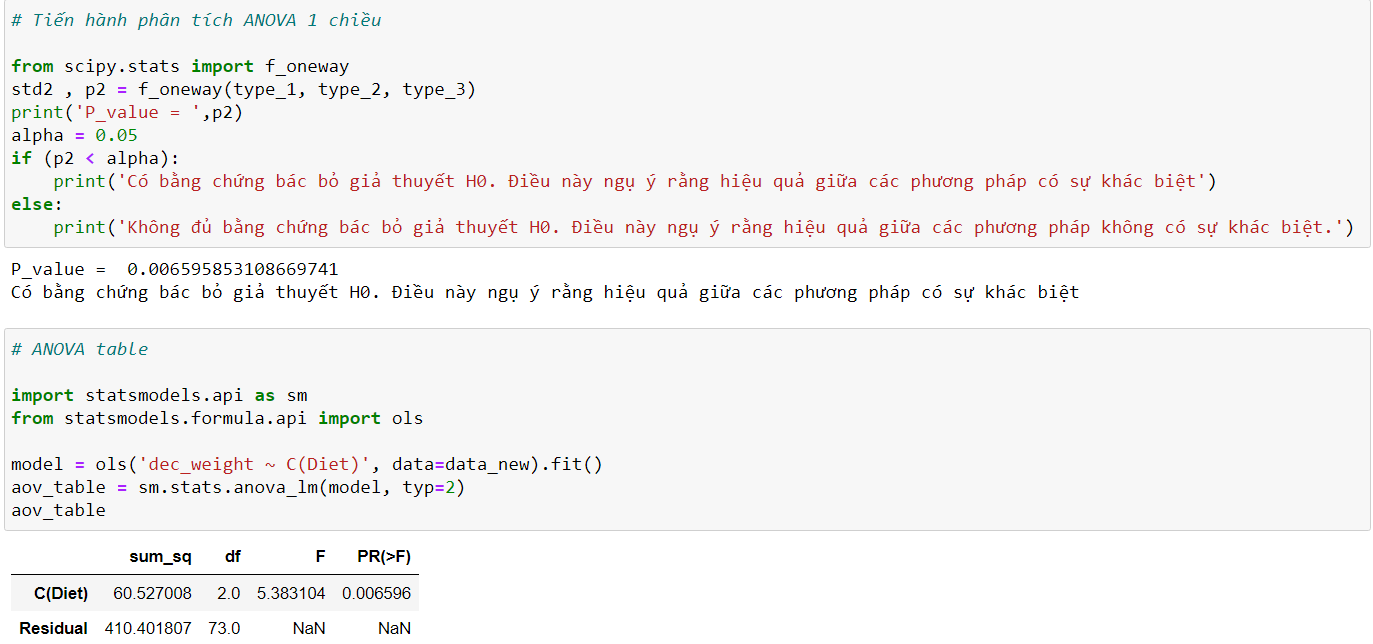
H0: Các feature đồng nhất phương sai.

H1: Các feature không đồng nhất phương sai.



Như vậy, chúng ta không đủ bằng chứng để bác bỏ giả thiết H0, điều đó đồng nghĩa rằng dữ liệu của chúng ta đủ điều kiện để thực hiện phân tích ANOVA.

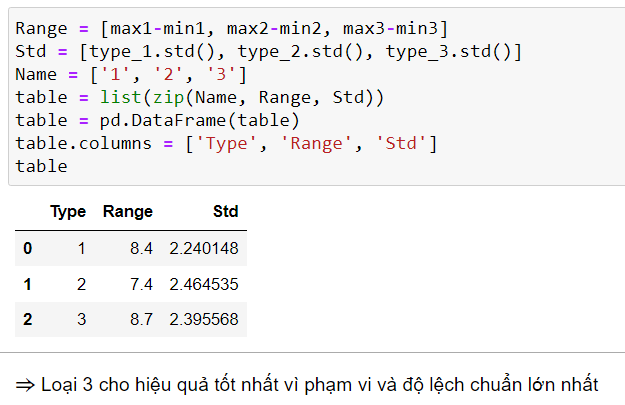
Phân tích ANOVA



Vậy giữa các feature có sự khác biệt, điều đó có nghĩa là hiệu quả giữa 3 phương pháp ăn kiêng là khác nhau.

**Câu hỏi 2: Phương pháp nào cho hiệu quả tốt nhất**

- Để trả lời được câu hỏi này chúng ta dựa vào độ phân tán của dữ liệu, tức là pham vi và độ lệch chuẩn của từng phương pháp ăn kiêng. Phương pháp nào cho độ lệch chuẩn lớn, phạm vi dữ liệu lớn tức là cho hiệu quả cao và ngược lại.



=> Kết luận: Phương pháp số 3 cho hiệu quả tốt nhất với phạm vi là 8.7 và độ lệch chuẩn 2.395568.

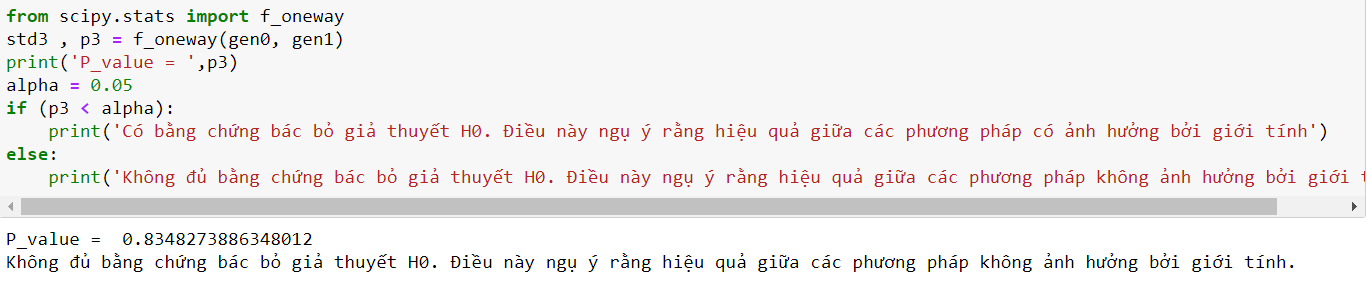
**Câu hỏi 3: Liệu hiệu quả giảm cân có ảnh hưởng đến giới tính không?**

- Xây dựng giả thiết:

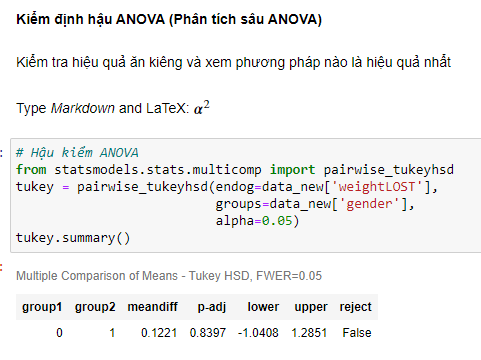
H0: Khả năng giảm cân phụ thuộc vào giới tính

H1: Khả năng giảm cân không phụ thuộc vào giới tính

- 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



Hậu kiểm ANOVA:

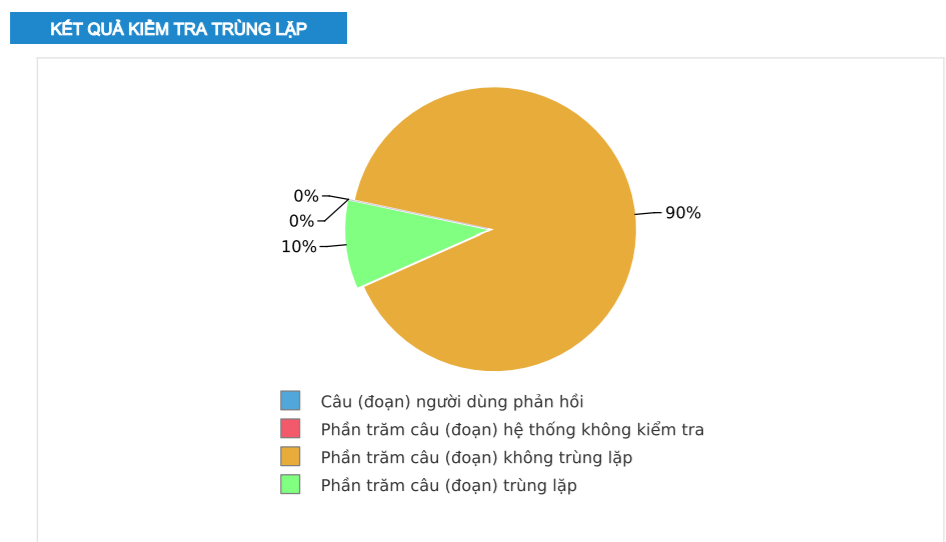


**Nhận xét:**

- **Có sự khác biệt** giữa Diet1 và Diet3 lẫn Diet2 và Diet3

- Giữa Diet1 và Diet2 không có sự khác biệt

# Chương 3: Kiểm tra đạo văn



# Chương 4: Tài liệu tham khảo

- Tham khảo cách sử dụng các thư viện và thống kê trong Python.

1. <https://pandas.pydata.org/docs/>
2. <https://www.scipy.org/docs.html>
3. <https://matplotlib.org/>
4. <https://stackoverflow.com/>
5. <https://l.facebook.com/l.php?u=https%3A%2F%2Fphantichspss.com%2Fphan-tich-sau-anova-mot-yeu-post-hoc-one-way-anova.html%3Ffbclid%3DIwAR2lOChOYMWowu6o2_sFLUX-wFLoA7S9Lna82x68OiUFzT6Po9bATv5XKcc&h=AT17VQVFp9SC9u6NqnsD3NVNCd5vJtRspI5J7fJ6W1SrMkDHfnMGgDieernf-4sgILkaKZit32lUwotvtpbLsO3QK5XX4C6XcTAt9ooh2AdqorrjHBXBeDo2DpWC40g8DcV2-g>
6. <https://www.stat.uci.edu/what-is-statistics/>
7. <https://www.investopedia.com/terms/s/statistics.asp>