**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: Trần Đoàn Ngọc Duy**

*(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 3](file:///D:\Down\Đồ_án_môn_học.docx#_Toc2275)

[Chương 1: Cơ sở lí thuyết 4](file:///D:\Down\Đồ_án_môn_học.docx#_Toc28951)

[I. Thống kê là gì? 4](file:///D:\Down\Đồ_án_môn_học.docx#_Toc914)

[1/ Định nghĩa 4](file:///D:\Down\Đồ_án_môn_học.docx#_Toc10445)

[2/ Ứng dụng của thống kê 7](file:///D:\Down\Đồ_án_môn_học.docx#_Toc10567)

[II. Thực hiện thống kê trong Python 9](file:///D:\Down\Đồ_án_môn_học.docx#_Toc1220)

[1/ Giới thiệu về ngôn ngữ lập trình Python. 9](file:///D:\Down\Đồ_án_môn_học.docx#_Toc4928)

[2/ Giới thiệu về các thư viện. 9](file:///D:\Down\Đồ_án_môn_học.docx#_Toc7653)

[Chương 2: Phân tích dữ liệu 11](file:///D:\Down\Đồ_án_môn_học.docx#_Toc20692)

[I Thống kê mô tả 11](file:///D:\Down\Đồ_án_môn_học.docx#_Toc8161)

[1/ Phân tích về sự phân phối của dữ liệu 12](file:///D:\Down\Đồ_án_môn_học.docx#_Toc6626)

[2/ Phân tích xu hướng tập trung của dữ liệu 21](file:///D:\Down\Đồ_án_môn_học.docx#_Toc23124)

[3/ Phân tích mức độ phân tán của dữ liệu 23](file:///D:\Down\Đồ_án_môn_học.docx#_Toc8648)

[II Thống kê suy diễn](file:///D:\Down\Đồ_án_môn_học.docx#_Toc13438) 25

[Chương 3: Tài liệu tham khảo](file:///D:\Down\Đồ_án_môn_học.docx#_Toc30423) 34

[Chương 4: Kiểm tra đạo văn](file:///D:\Down\Đồ_án_môn_học.docx#_Toc16632) 35

# Chương 1: Cơ sở lí thuyết

## I.Thống kê là gì?

### 1/ Định Nghĩa

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

• Thống kê mô tả

• Thống kê suy diễn

**Thống kê** là thuật ngữ được sử dụng để tóm tắt một qui trình mà nhà phân tích sử dụng để mô tả đặc điểm của tập dữ liệu. Nếu tập dữ liệu mẫu liên quan tới một số liệu lớn về lượng dân số thì nhà phân tích có thể phát triển các diễn giải về dân số dựa trên kết quả thống kê từ mẫu. Phân tích thống kê liên quan đến quá trình thu thập và đánh giá dữ liệu và sau đó tóm tắt dữ liệu thành một dạng toán học.

**Thống kê** được sử dụng trong đa ngành như tâm lý học, kinh doanh, khoa học vật lí, xã hội, nhân văn.... Dữ liệu thống kê được thu thập chủ yếu bằng phương pháp thu thập mẫu. Hai loại phương pháp thống kê được sử dụng trong phân tích dữ liệu là thống kê mô tả và thống kê suy luận. Thống kê mô tả được sử dụng để đồng bộ hóa dữ liệu từ một mẫu để tính độ lệch trung bình hoặc độ lệch chuẩn. Thống kê suy luận thì được sử dụng để nhìn vào một dữ liệu được đánh giá là phần nhỏ của một con số cụ thế.

#### a) Thống kê mô tả:

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. 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.

### 2/ Ứng dụng của thống kê:

#### Ứng dụng thống kê, lý thuyết thống kê và toán thống kê:

“Thống kê ứng dụng” bao gồm thống kê mô tả và các ứng dụng của thống kê suy luận (bằng chứng cần thiết). Lý thuyết thống kê liên quan tới những lập luận logic cơ bản giải thích của phương pháp tiếp cận kết luận thống kê, cũng bao gồm toán thống kê. Toán thống kê không chỉ bao gồm các thao tác của phân phối xác suất cần thiết cho kết quả phát sinh liên quan đến các phương pháp tính toán và suy luận, nhưng còn khía cạnh khác nhau của các số liệu thống kê tính toán và thiết kế các thử nghiệm.

#### 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.

#### Tính toán thống kê:

Sự tăng nhanh và ổn định ở khả năng tính toán bắt đầu từ nửa sau thế kỷ 20 đã có một tác động đáng kể vào việc thực hành của môn khoa học thống kê. Mô hình thống kê lúc đầu gần như là của một lớp mô hình tuyến tính, nhưng khả năng tính toán, cùng với các thuật toán số học phù hợp, gây ra một lãi suất tăng trong các mô hình phi tuyến (như mạng thần kinh) cũng như tạo ra các kiểu mới, chẳng hạn như mô hình tuyến tính tổng quát và mô hình đa cấp.

Khả năng tính toán tăng cũng dẫn đến sự phổ biến ngày càng tăng của các phương pháp tính toán dựa trên chọn mẫu, chẳng hạn như xem xét hoán vị và khả năng tự hoán vị, trong khi các kỹ thuật như Gibbs lấy mẫu đã sử dụng mô hình Bayesian khả thi hơn. Các cuộc cách mạng máy tính có ảnh hưởng đến tương lai của số liệu thống kê với sự nhấn mạnh mới về “thử nghiệm” và thống kê “thực nghiệm”. Một số lượng lớn của tổng thể và đặc biệt là phần mềm thống kê taị thời điểm hiện tại.

#### Thống kê áp dụng cho toán học hay nghệ thuật:

heo truyền thống, thống kê có liên quan tới sự suy luận bản vẽ qua việc sử dụng một phương pháp bán tiêu chuẩn đã được “yêu cầu thử nghiệm” trong hầu hết các ngành khoa học. Điều này đã thay đổi việc sử dụng số liệu thống kê trong các bối cảnh không có kết luận. Những gì đã được coi là một chủ đề vô vị, thực hiện trong nhiều lĩnh vực như một mức yêu cầu, bây giờ được xem một cách nhiệt tình. Ban đầu một số người khó tính đã cười nhạo, nhưng hiện nay lại được coi là phương pháp cần thiết trong mọi lĩnh vực.

* Lý thuyết số tự nhiên n, biểu đồ phân rã của dữ liệu được tạo ra bởi một hàm phân phối có thể được chuyển đổi với các công cụ quen thuộc được sử dụng trong thống kê để cho thấy những điều cơ bản, mà sau đó có thể dẫn đến các giả thuyết.
* Phương pháp thống kê bao gồm các phương pháp dự báo được kết hợp với lý thuyết hỗn hợp và hình học nhân bản để tạo ra các tác phẩm video được đánh giá có vẻ đẹp tuyệt vời.
* Các quá trình nghệ thuật của Jackson Pollock dựa trên thí nghiệm nghệ thuật phân bố cơ bản trong tự nhiên được tiết lộ. Với sự ra đời của máy tính, các phương pháp thống kê được áp dụng để hợp thức hóa với các quá trình tự nhiên phân phối theo định hướng như vậy để thực hiện và phân tích nghệ thuật hình ảnh động.
* Phương pháp thống kê có thể được sử dụng để xác nhận trong nghệ thuật trình diễn, như trong một thẻ giả dựa trên quá trình Markov và chỉ hoạt động một thời gian nhất định, qua đó có thể dự đoán được việc sử dụng phương pháp thống kê.
* Thống kê có thể được sử dụng trong việc tạo hình nghệ thuật, như trong âm nhạc hay thống kê ngẫu nhiên phát minh bởi Lannis Xenakis, nơi âm nhạc biểu diễn rõ ràng. Mặc dù kiểu nghệ thuật không phải lúc nào cũng như mong đợi, nó diễn ra theo cách đó là được đoán trước và có hòa âm được bằng cách sử dụng thống kê.

#### 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. 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.

## II.Thống kê trong Python:

### 1/Giới thiệu về ngôn ngữ lập trình python:

**Python** là một 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/Giới thiệu về các thư viện:

#### 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, 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.

#### 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.

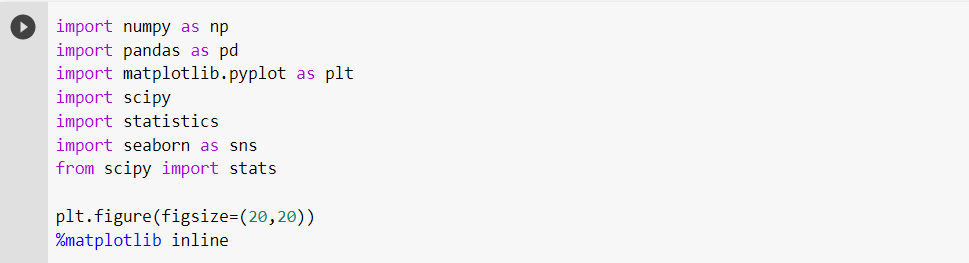
- Matplotlib có một giao diện thủ tục được đặt tên là Pylab, được thiết kế giống với MATLAB, một ngôn ngữ lập trình độc quyền được phát triển bởi MathWorks. Matplotlib cùng với NumPy có thể được coi là mã nguồn mở tương đương với MATLAB.

- 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

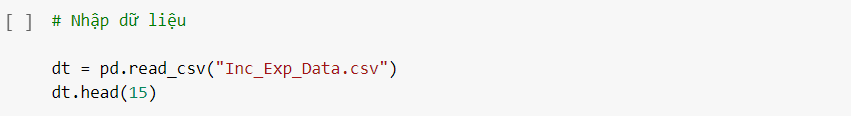
## I/ Thống kê mô tả

- Đầu tiên ta import các thư viện:

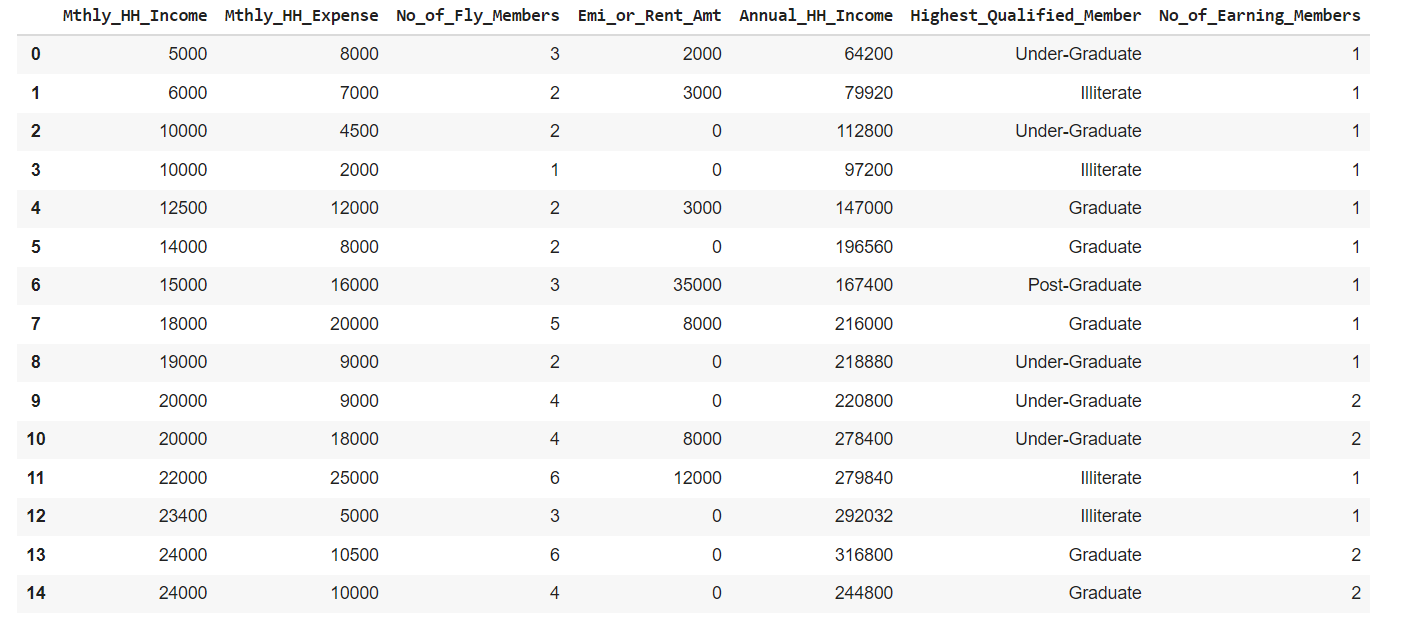


- Trong phần này, ta sử dụng tập dữ liệu Inc\_Exp\_Data.csv để phân tích thống kê mô tả.

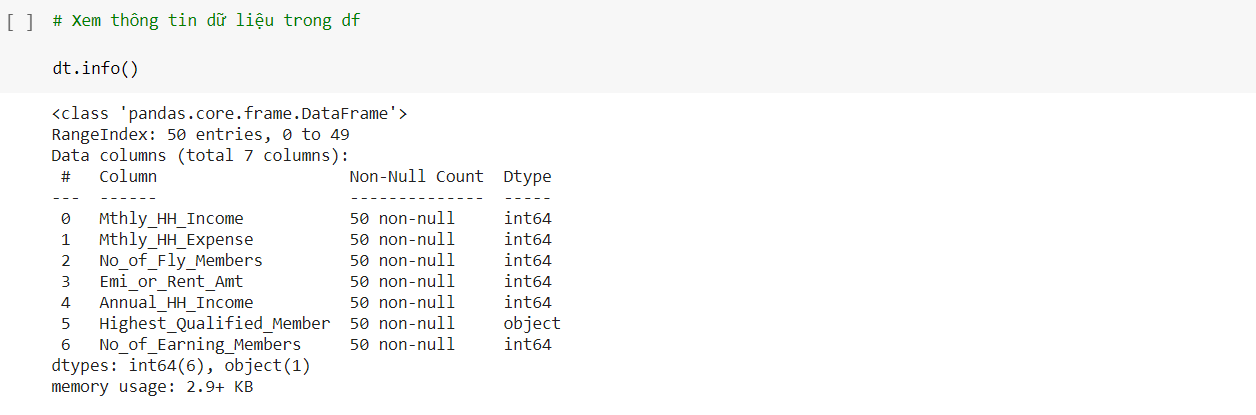
- Đầu tiên chúng ta truyền file vào bằng hàm read.csv của thư viện pandas:



-Sau khi truyền vào ta được data:



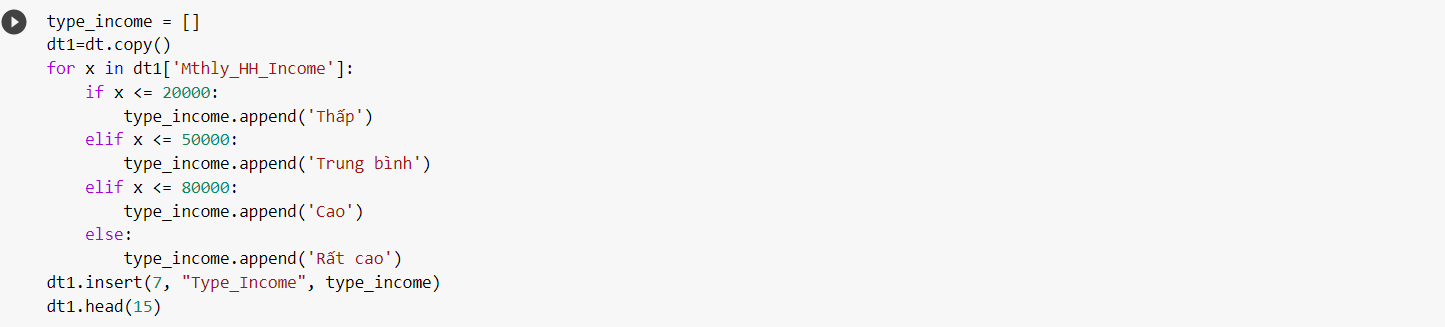
Sử dụng hàm info() để xem thông tin của data:

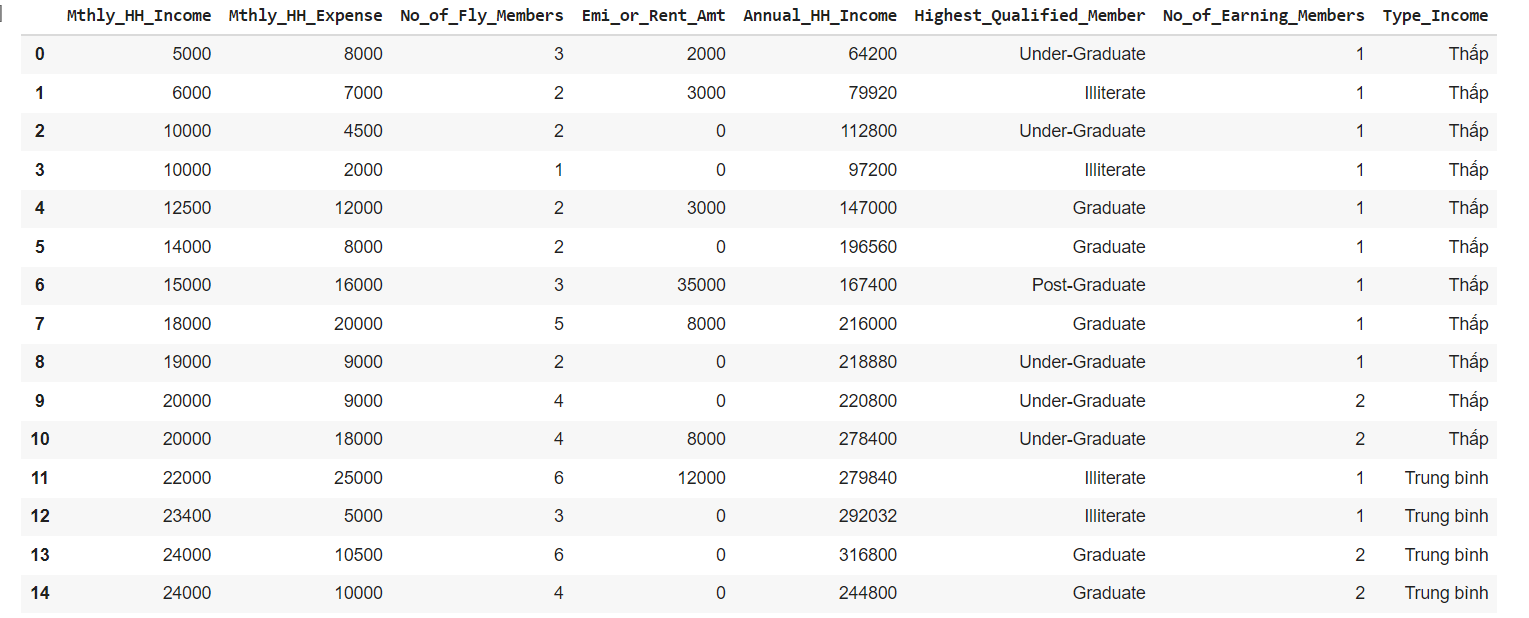


Tiếp theo ta đi vào phân tích dữ liệu.

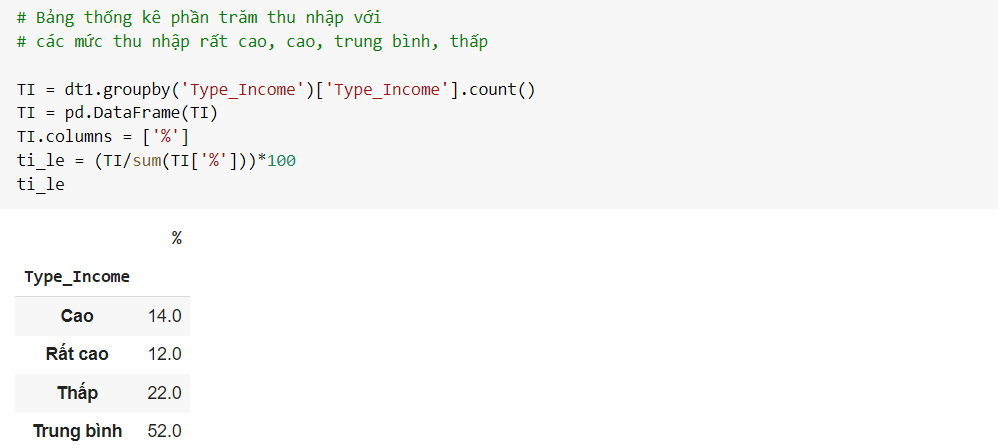
### Phân tích về sự phân phối của dữ liệu:

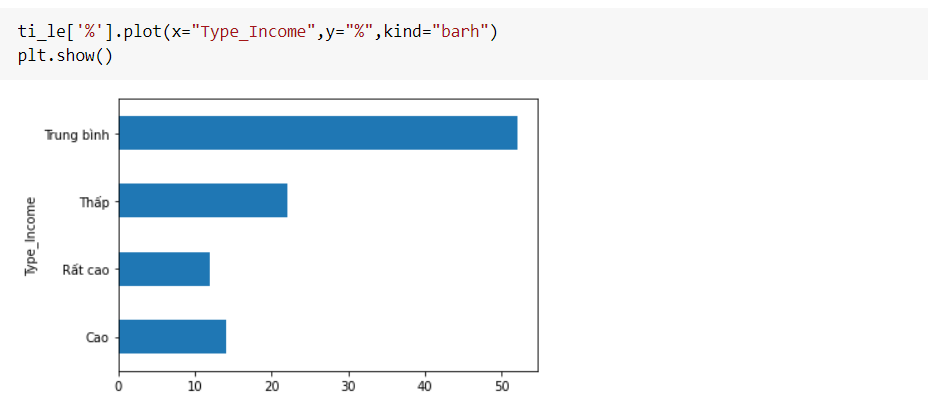
- Để thuận tiện cho việc phân tích dữ liệu, ta thêm vào một cột có tên là “Type” biểu diễn loại thu nhập. Thu nhập < 20000$ thuộc loại thấp,80000$ thuộc loại rất cao.........





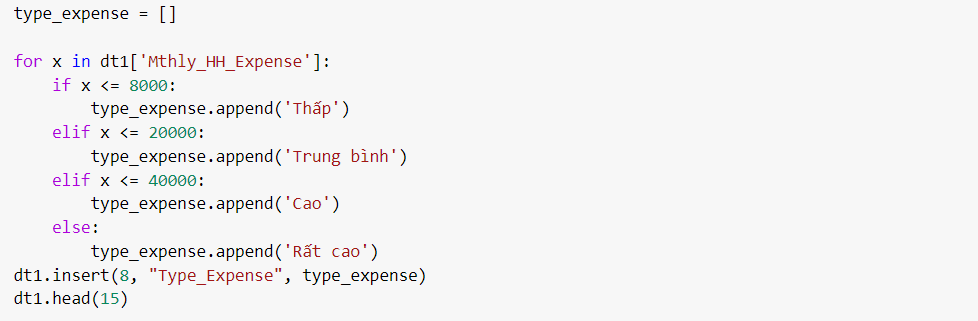
Bảng tỉ lệ và biểu đồ tương ứng:

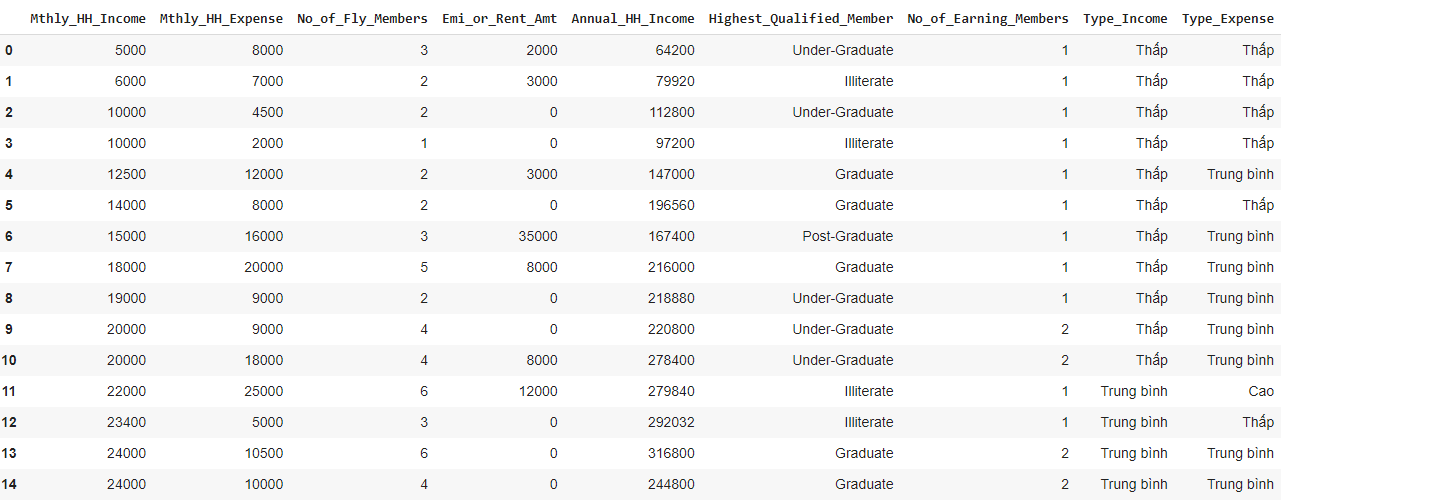




Nhận xét: tỉ lệ người có thu nhập trung bình cao nhất

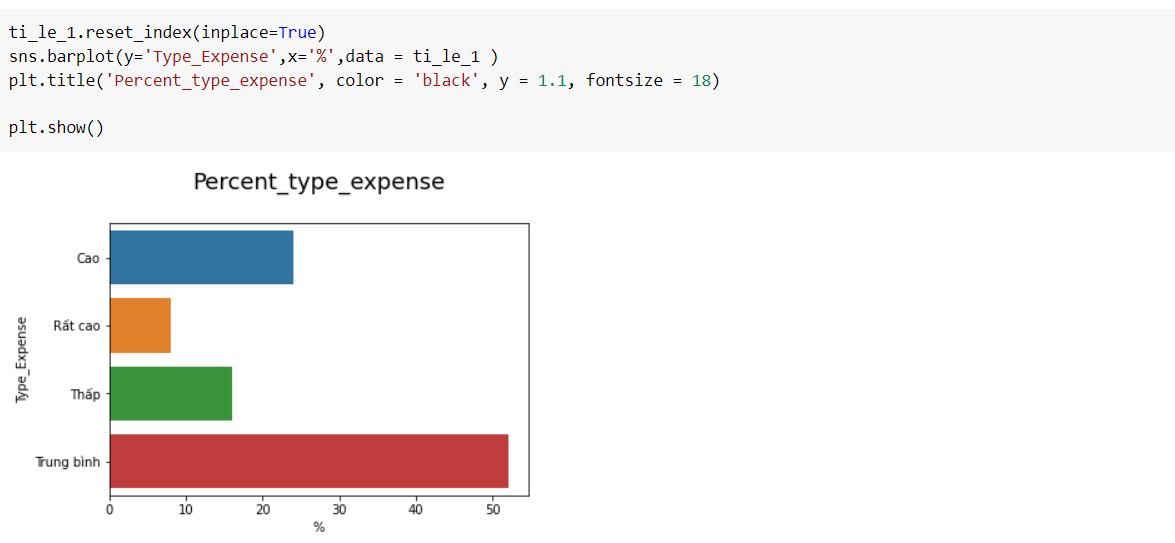
Tương tự với cột chi tiêu hằng tháng:





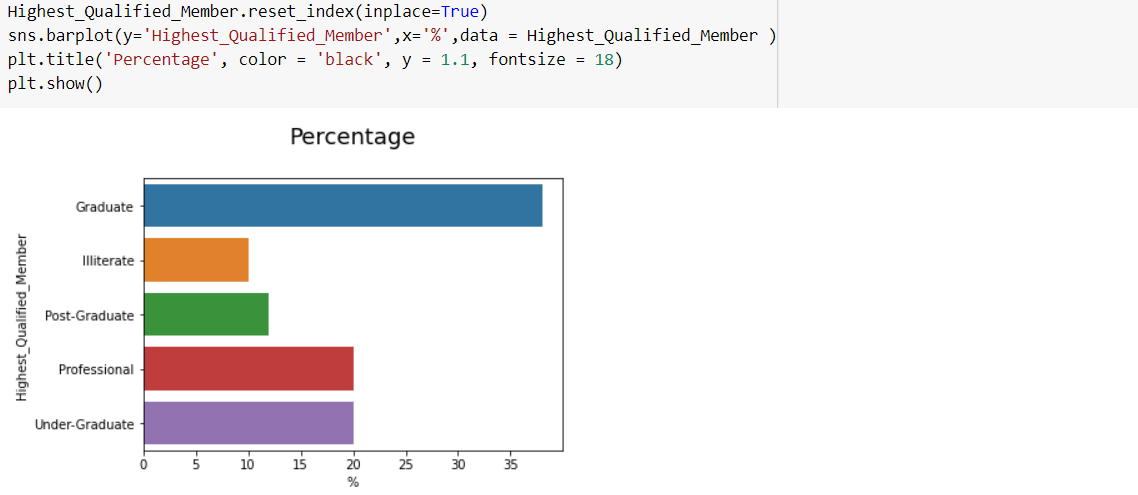
Tương tự với các cột khác:



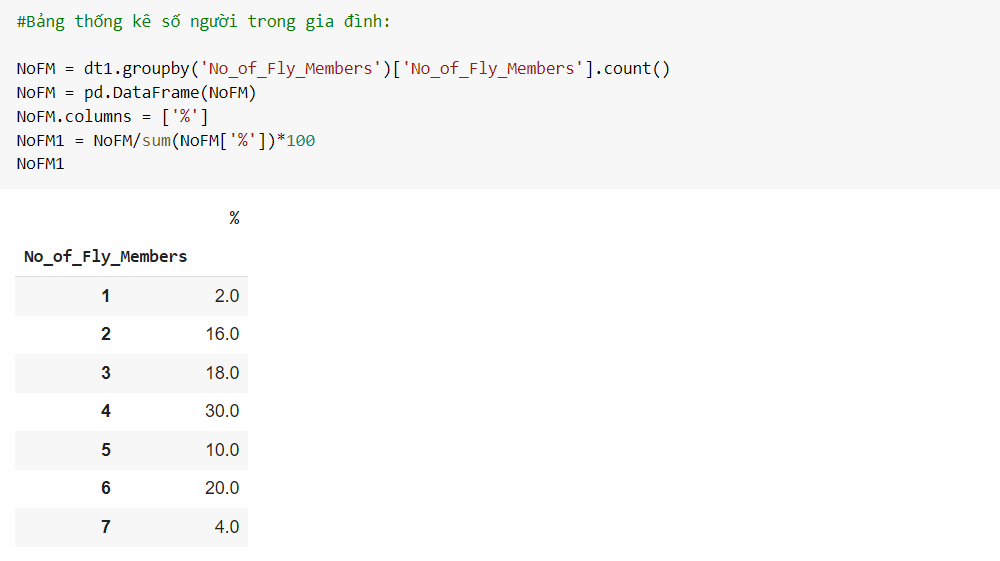


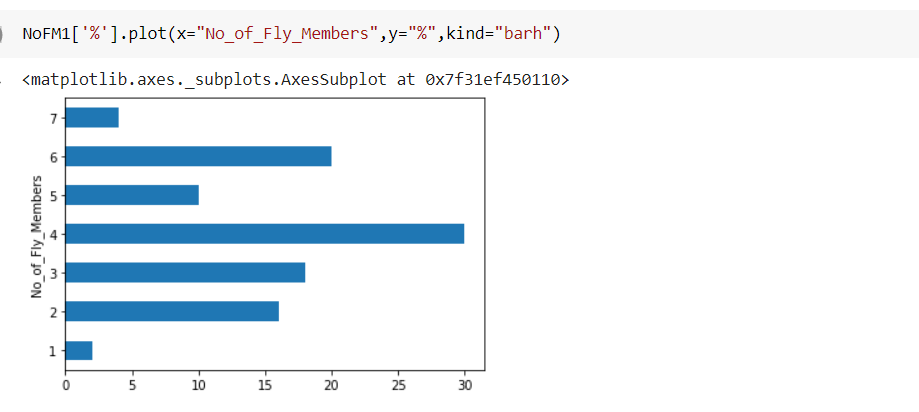
Nhận xét: Chi tiêu trung bình cao nhất



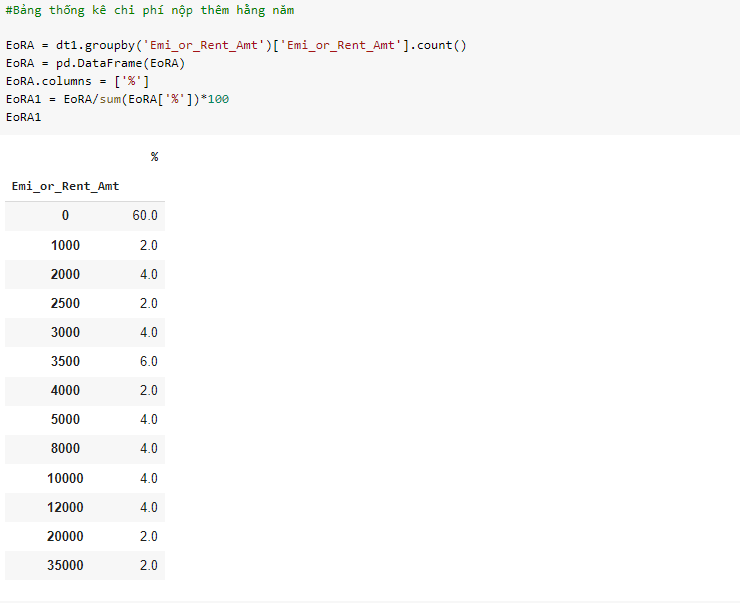


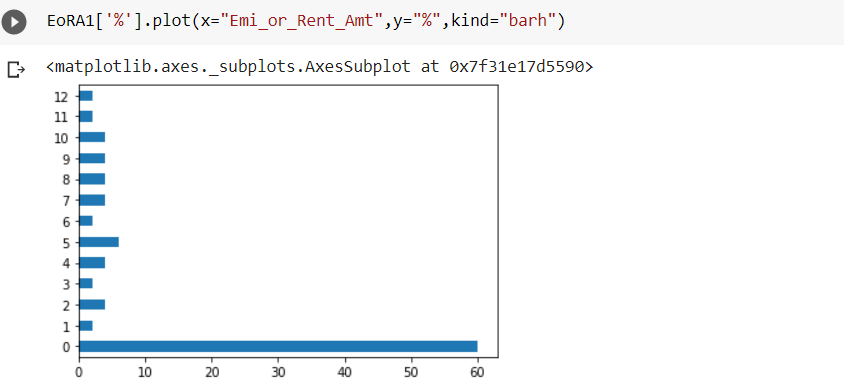
Nhận xét: trình độ cao nhất là Graduate chiếm tỉ lệ cao nhất

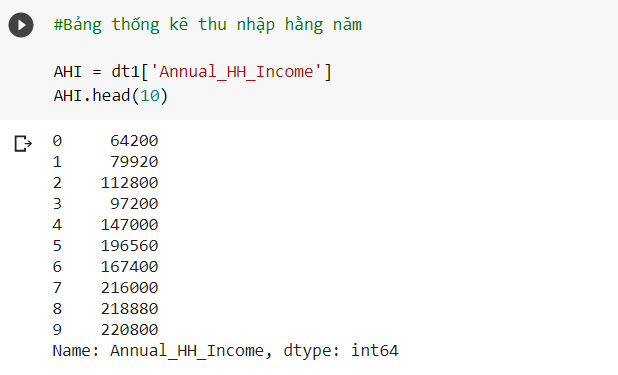


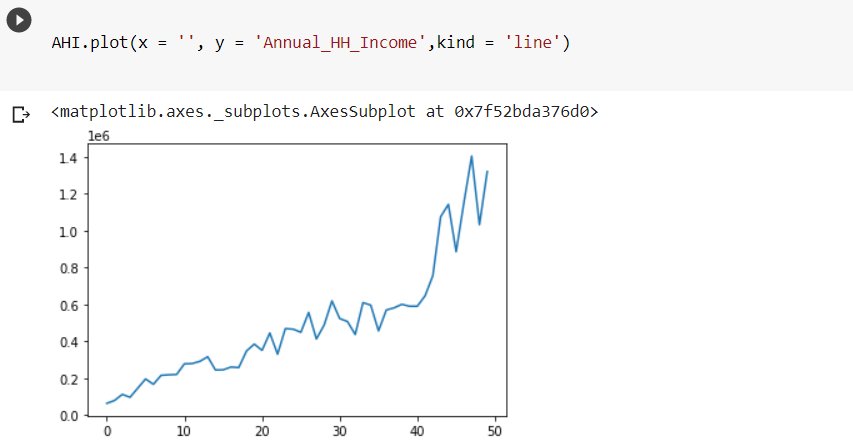


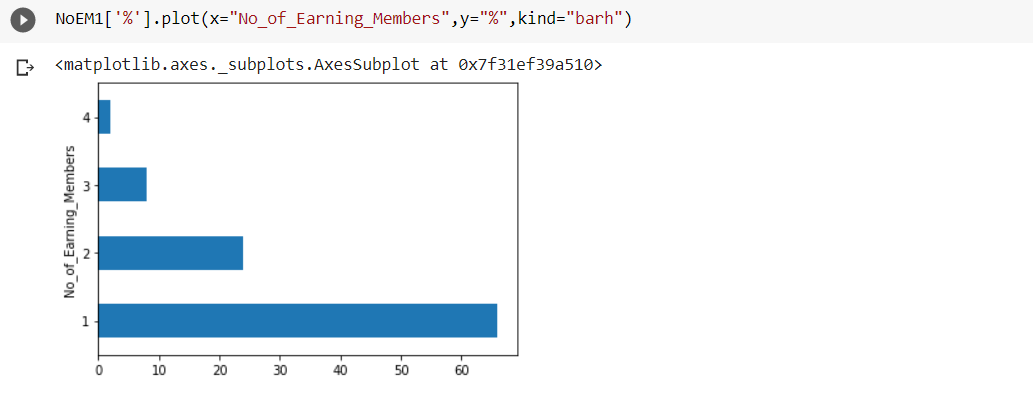
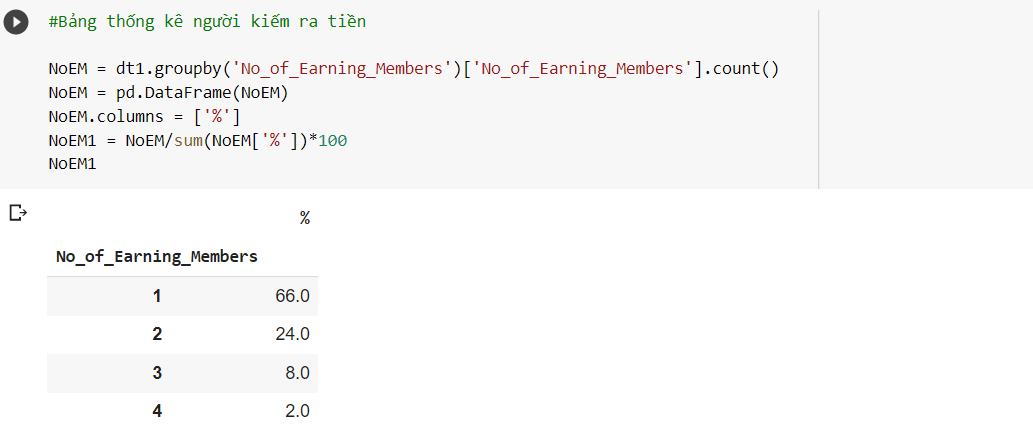
Nhận xét: Gia đình 4 người đông nhất





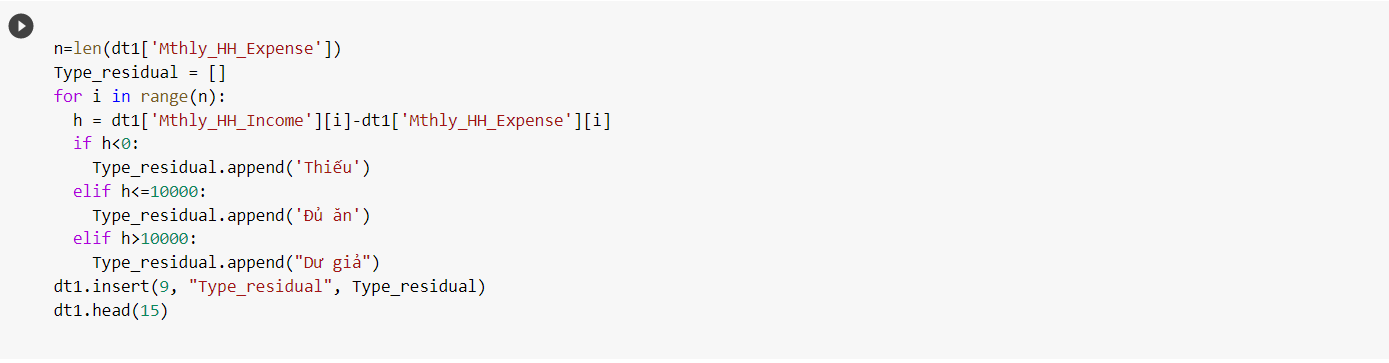


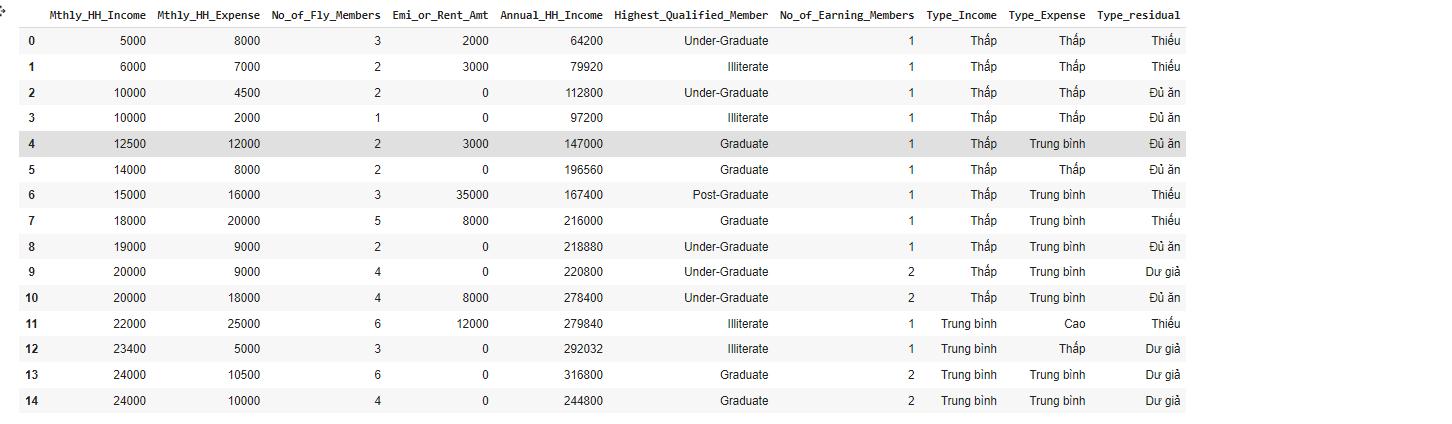




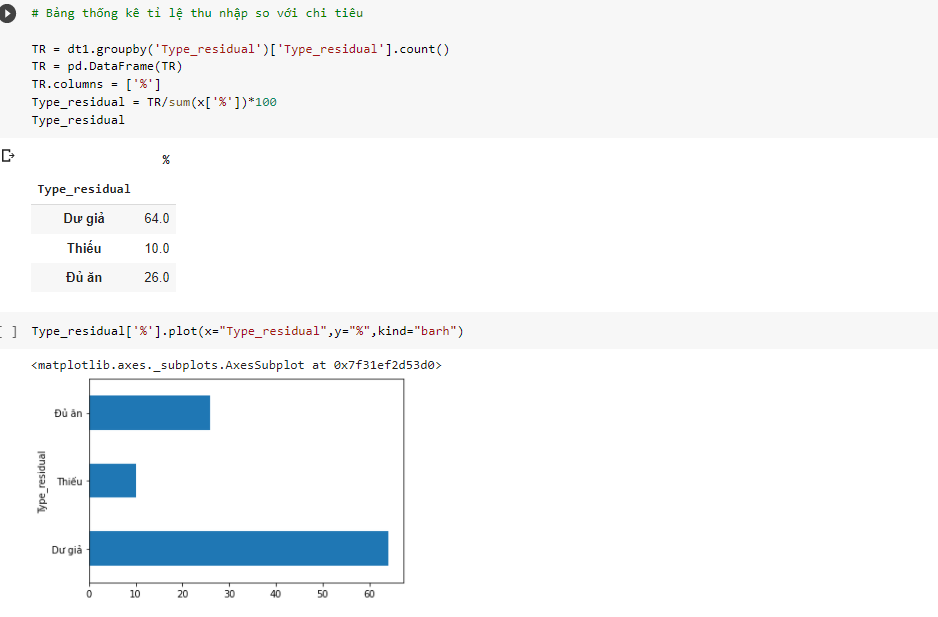
Nhận xét: Gia đình chỉ có 1 người kiếm được tiền chiếm tỉ lệ cao nhất

So sánh về thu chi hằng tháng, để xem tình trạng kinh tế gia đình đó:

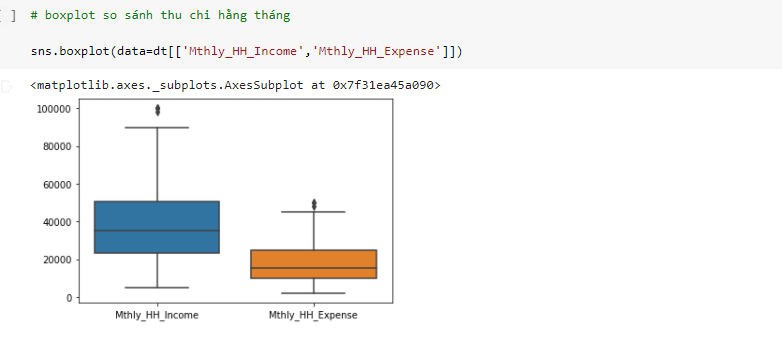




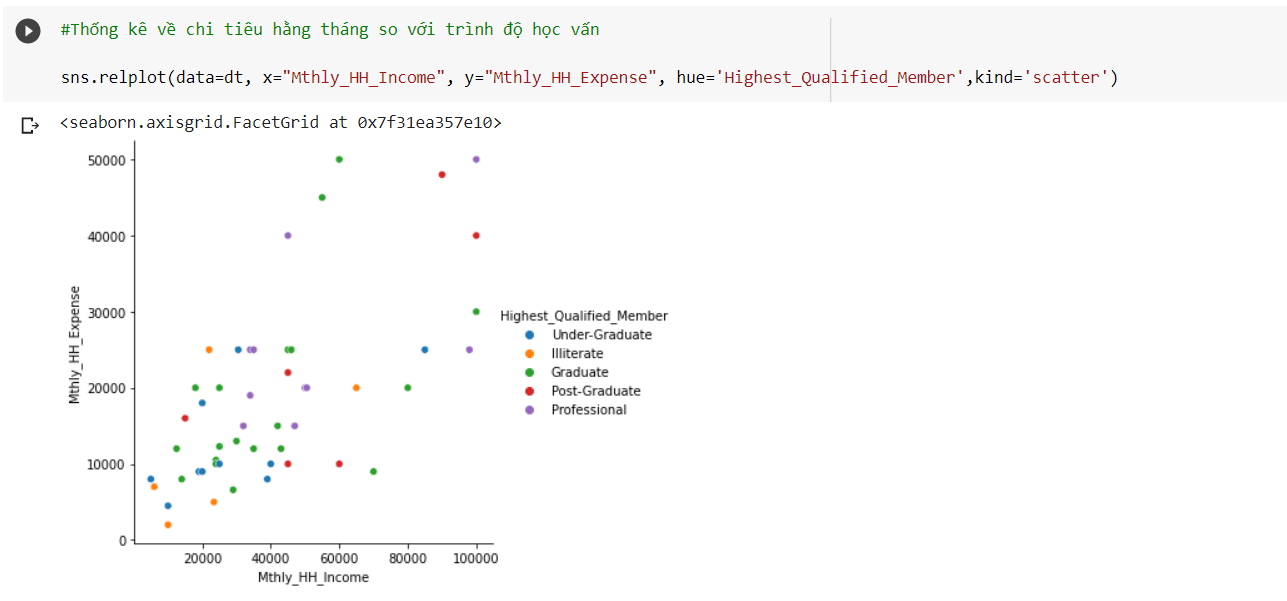
Bảng thống kê tỉ lệ và biểu đồ tương ứng:



Nhận xét: Tỉ lệ gia đình dư giả chiếm tỉ lệ cao nhất

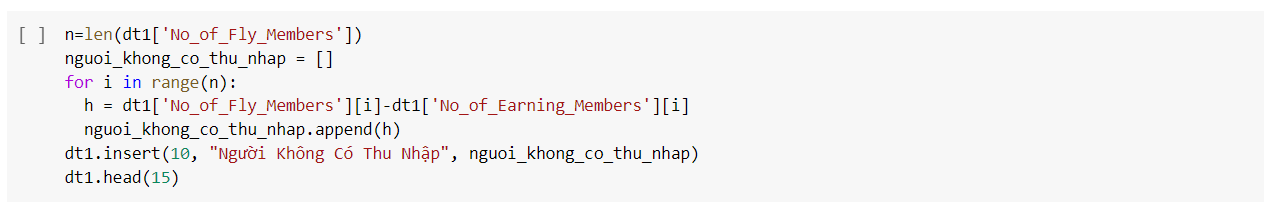


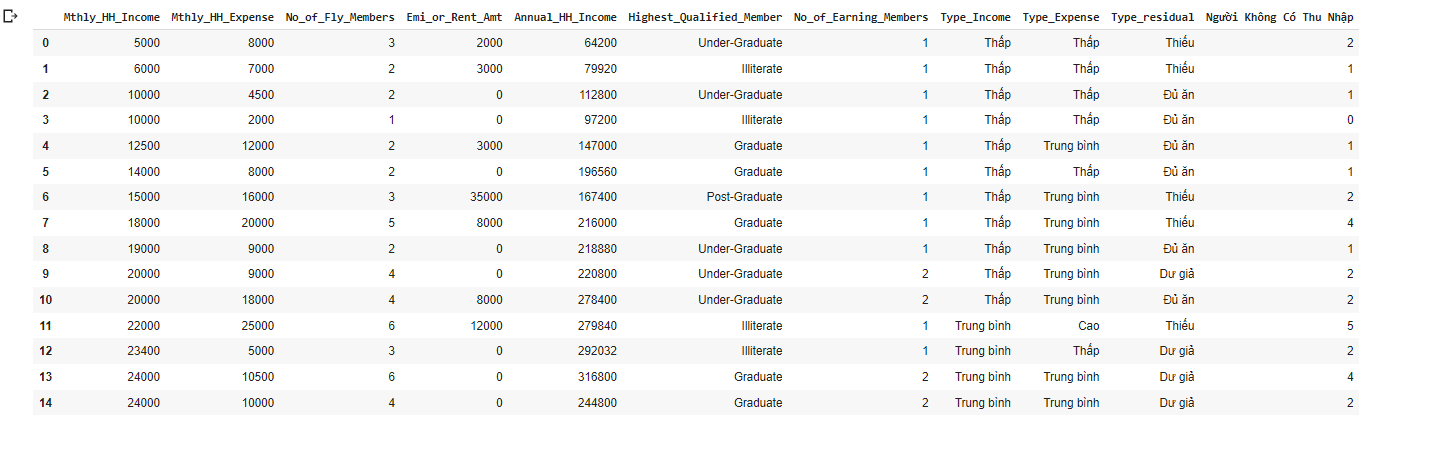
Biểu đồ chi tiêu hằng tháng so với trình độ học vấn:



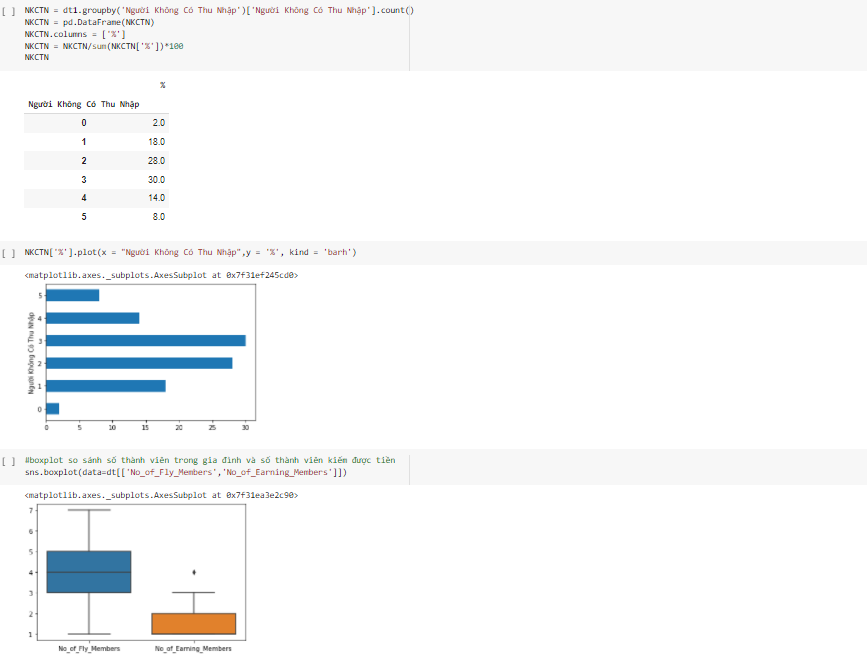
Nhận xét: Chi tiêu không phụ thuộc lắm vào trình độ học vấn

Thống kê về số người không kiếm ra tiền:





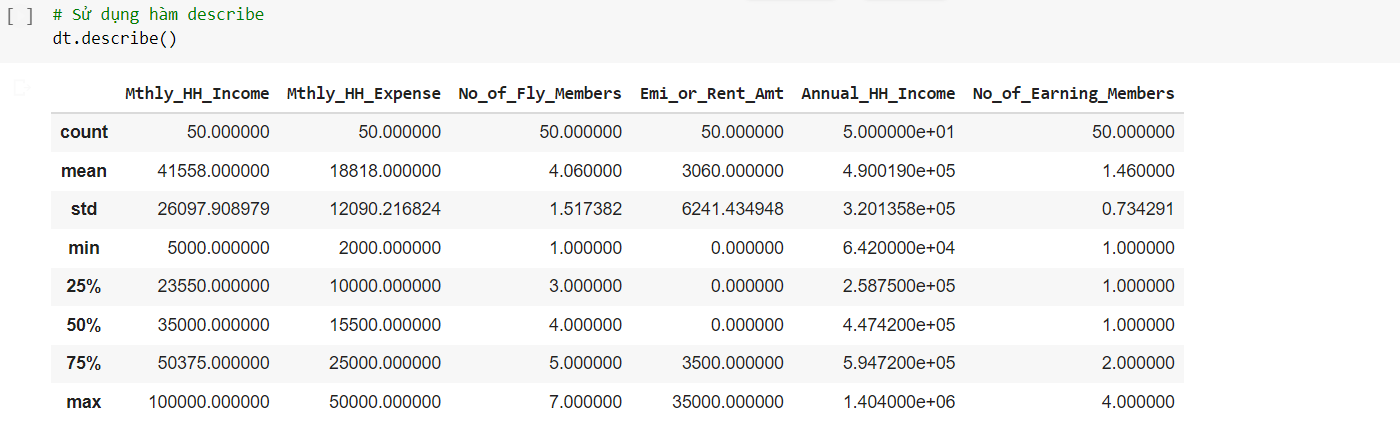
Bảng tỉ lệ và biểu đồ tương ứng:



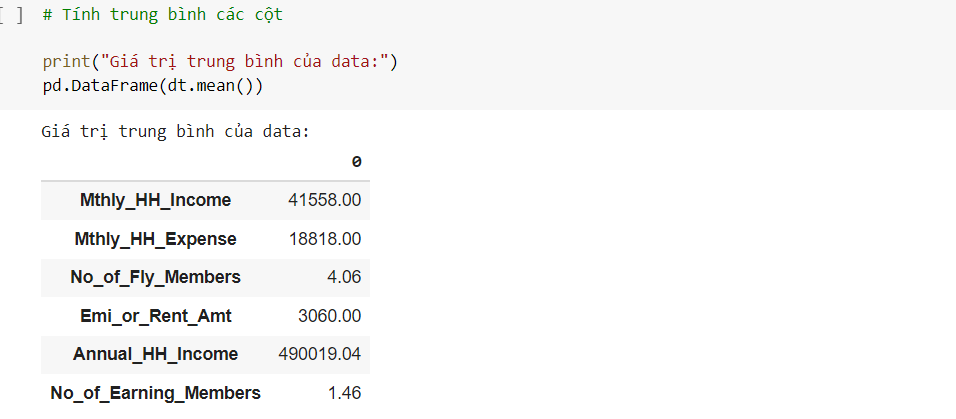
Nhận xét: Gia đình có 3 người không kiếm được tiền chiến tỉ lệ cao nhất

### 2.Xu hướng tập trung của dữ liệu:

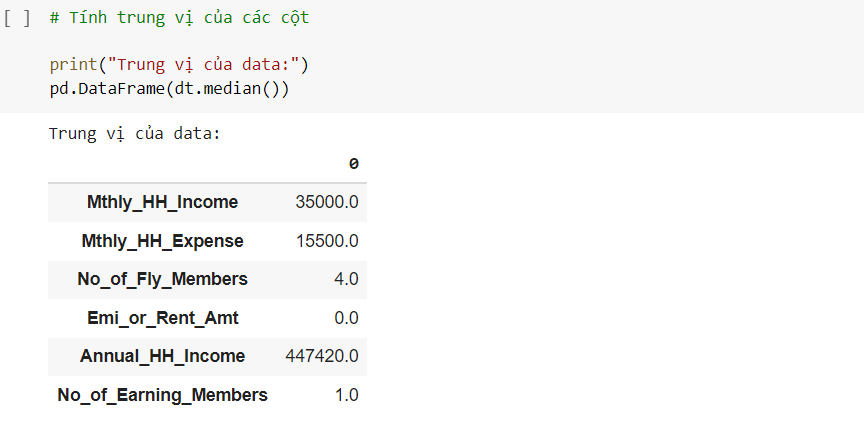
Sử dụng hàm **describe()** để có một bảng tóm tắt những thống kê cơ bản của dữ liệu. Hàm tính toán các giá trị như số lượng cá thể khảo sát, trung bình, độ lệch chuẩn, min, max, bách phân vị, …



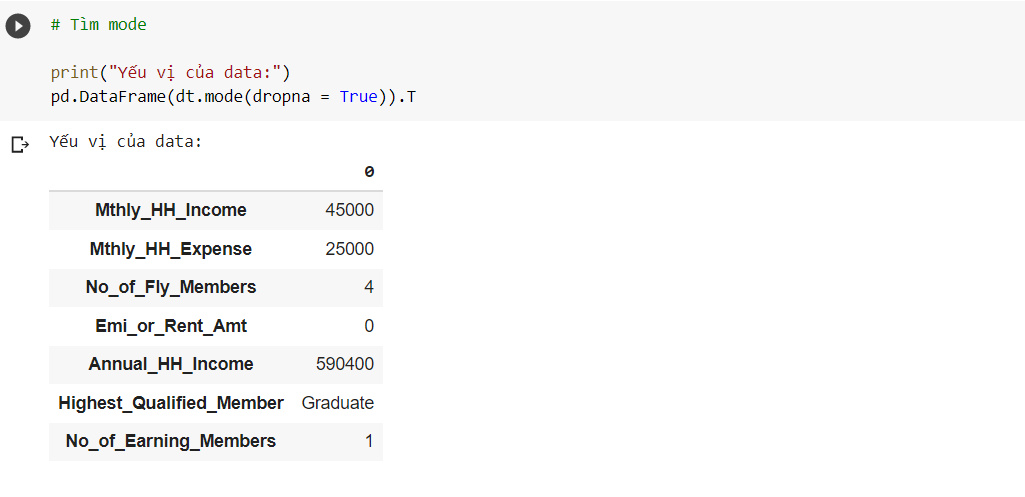
Tính trung bình:



Tính trung vị:

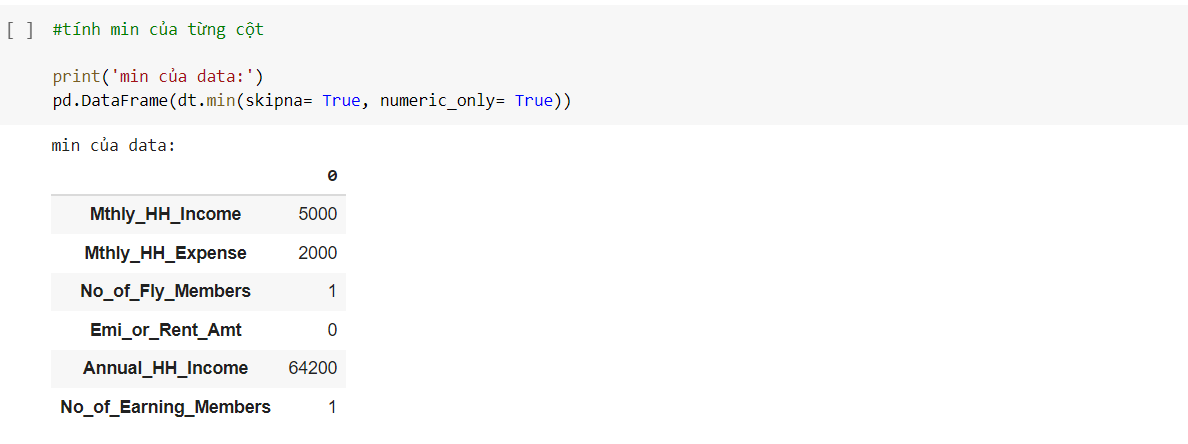


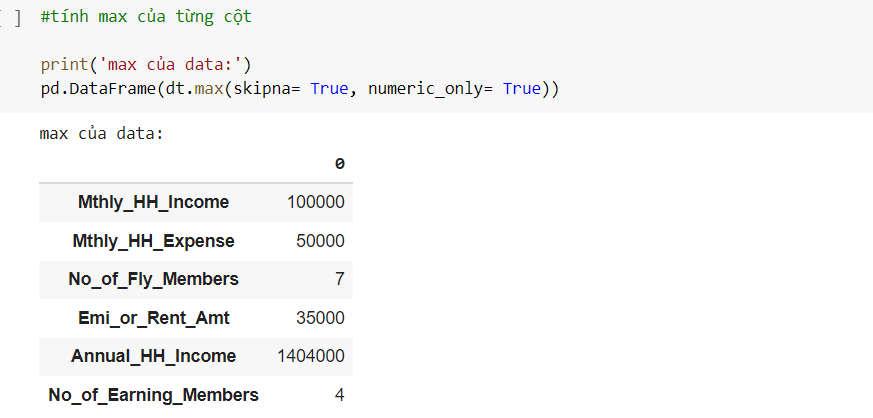
Tính yếu vị:



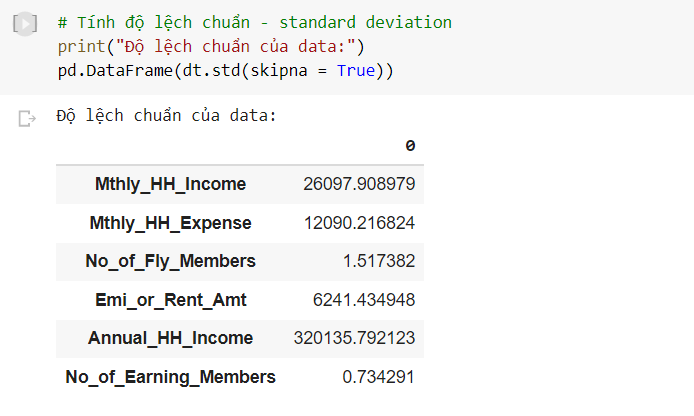
### 3.Phân tích mức độ phân tán của dữ liệu

Tính min,max:

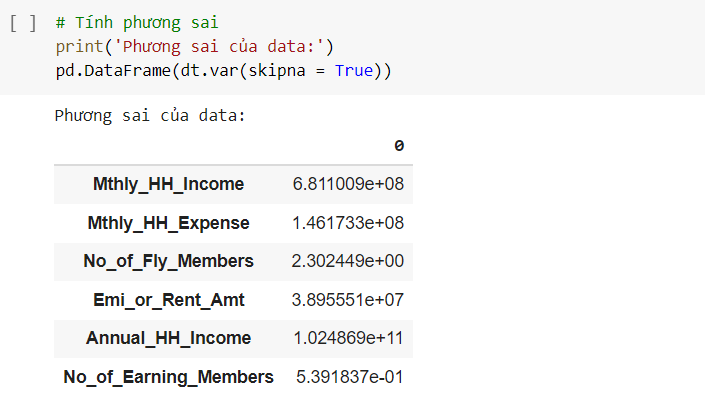




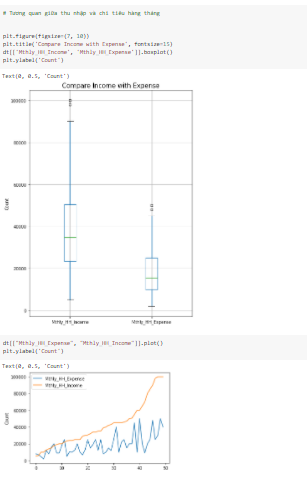
Tính độ lệch chuẩn:



Tính phương sai:

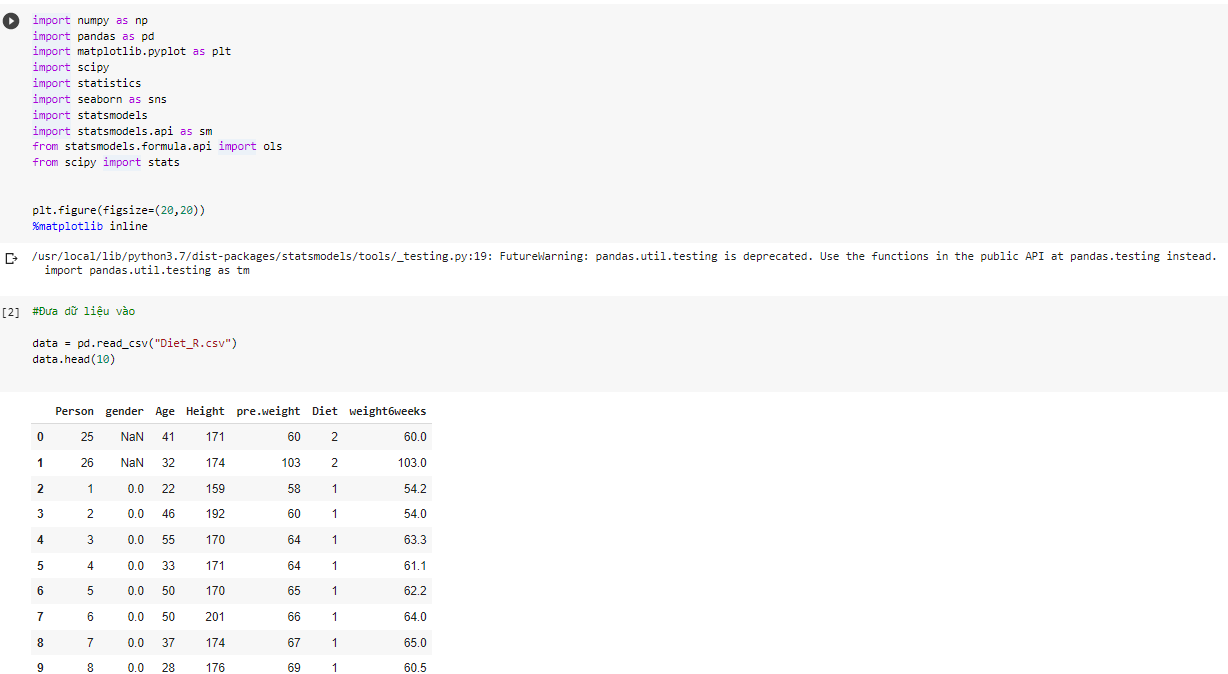


Tương quan giữa chi tiêu hằng tháng:

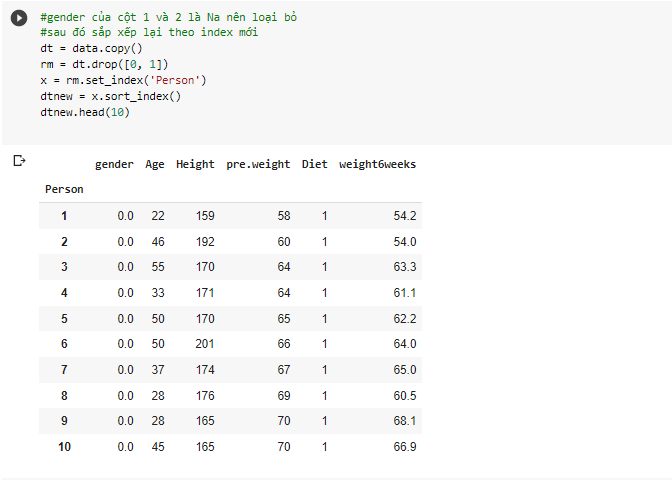


## II.Thống kê suy diễn:

Ban đầu ta import các thư viện và truyền file vào:



Chỉnh sửa data:



Các hàm cần thiết:

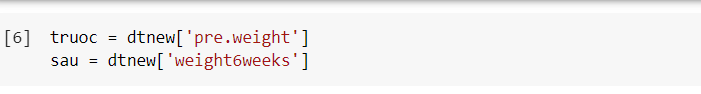


### 1.Kiểm định T Test:

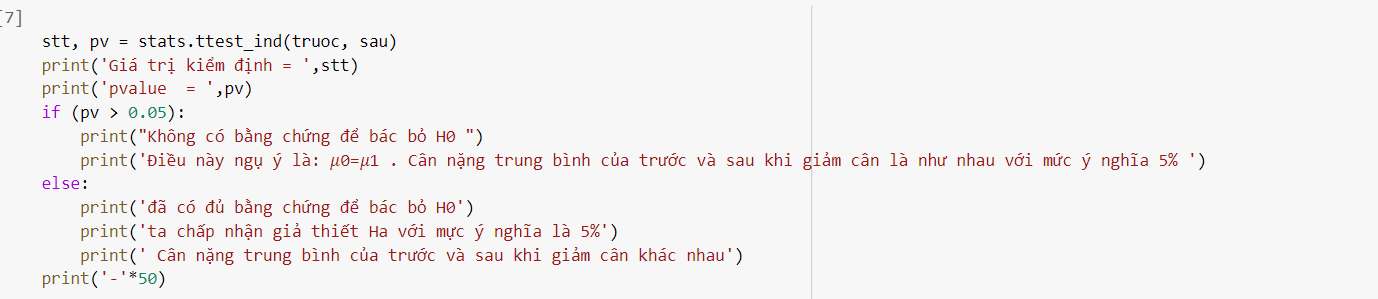
- Vậy chúng ta có giả thiết không và đối thiết như sau:

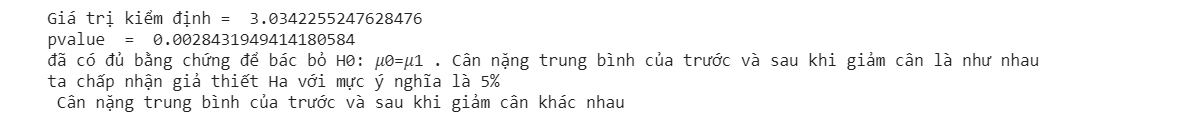
* H0: Cân nặng trung bình của mọi người trước và sau khi ăn kiêng là như nhau.
* H1: Cân nặng trung bình của mọi người trước và sau khi ăn kiêng khác nhau.

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



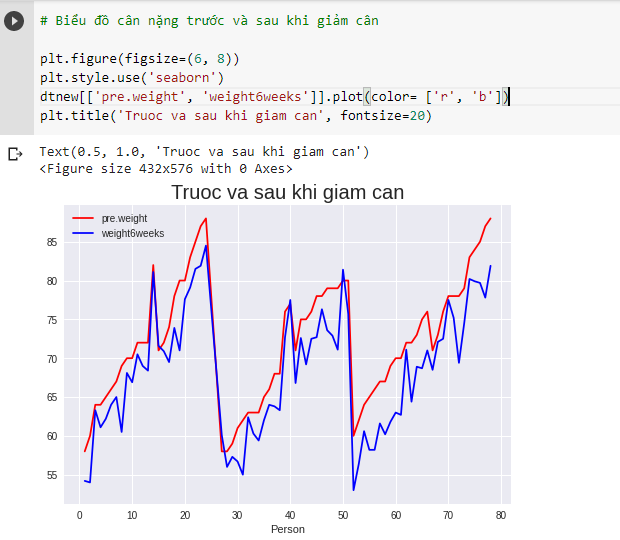
Kiểm định t test:





- 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.

Biểu đồ trước và sau khi giảm cân:



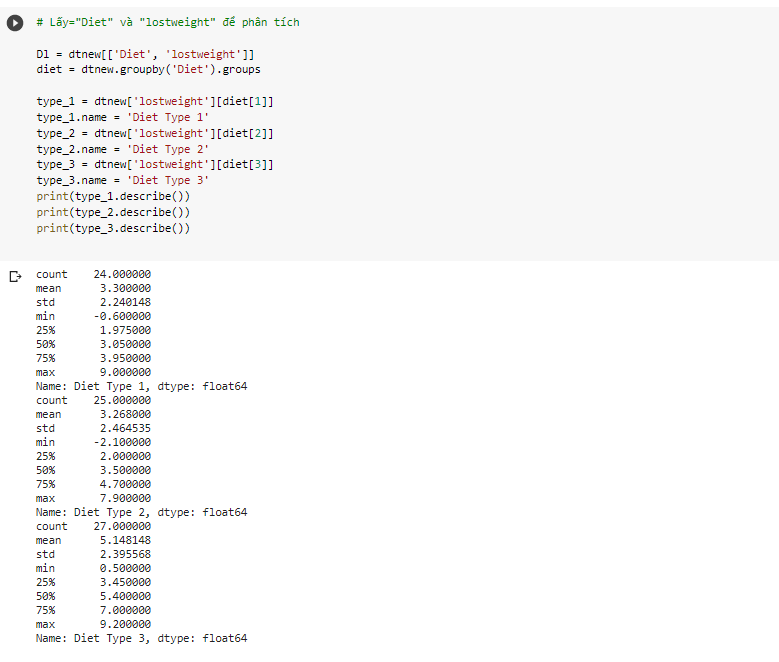
### 2.Kiểm định ANOVA

**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.

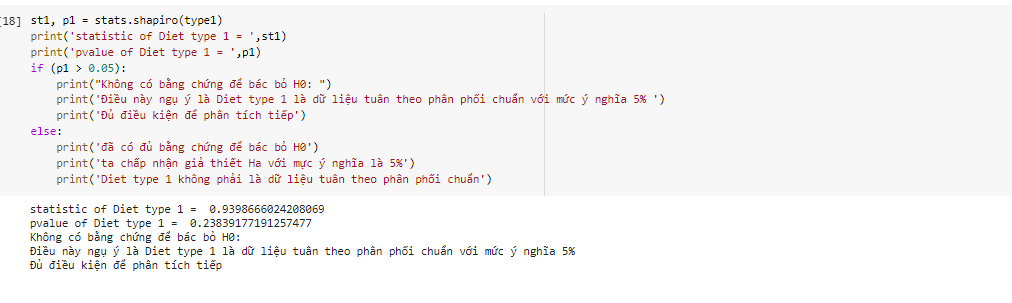
- Để kiểm chứng giả thiết chúng ta cần trích xuất dữ liệu để phân tích, phân loại dữ liệu và tiến hành phân tích ANOVA

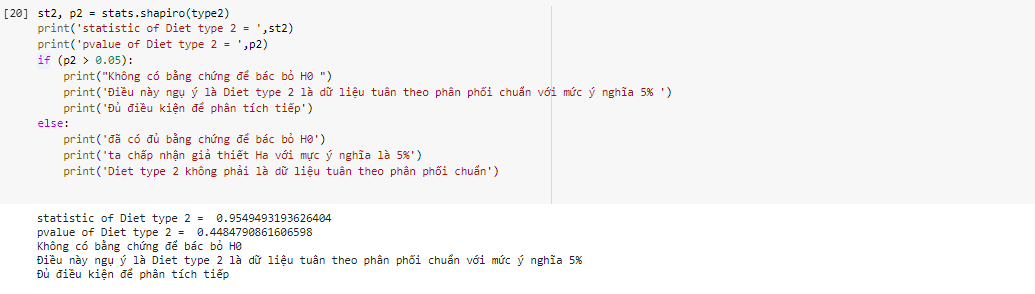


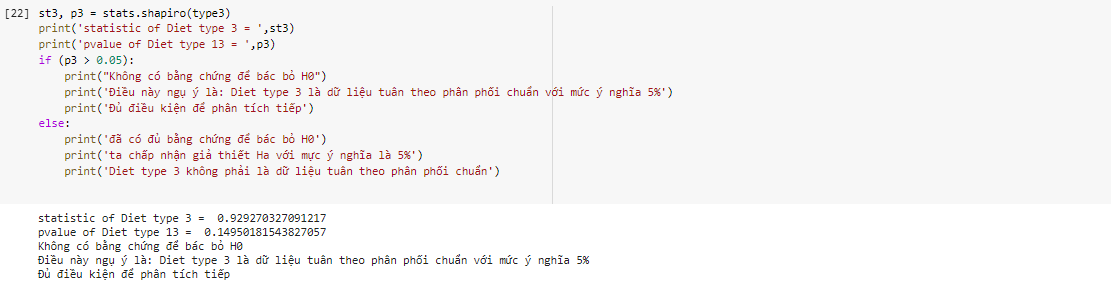
- Chú ý rằng, để thực hiện kiểm định ANOVA thì dữ liệu phải thỏa mãn các yêu cầu sau:

* 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.

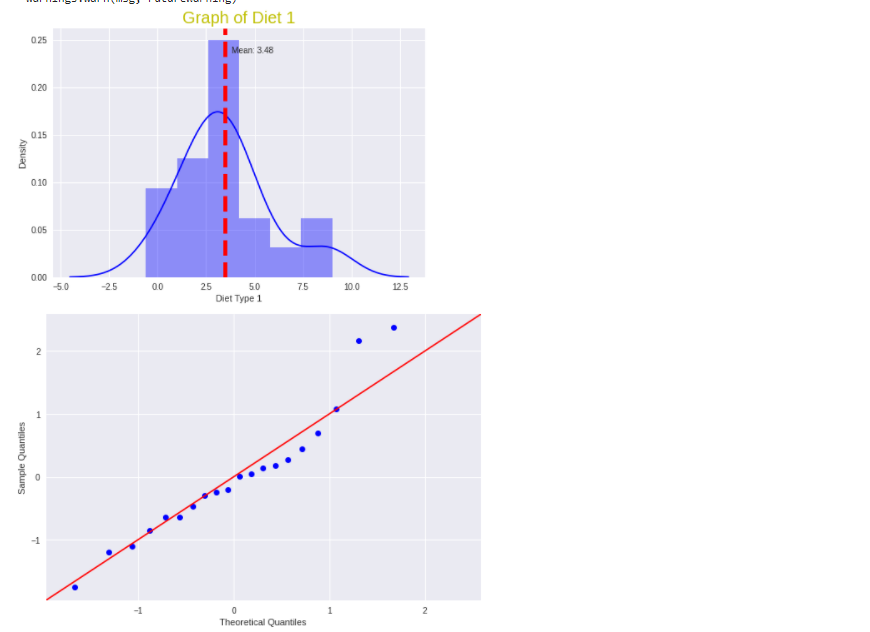
Để kiểm tra dữ liệu có tuân theo phân phối chuẩn hay không, chúng ta sử dụng kiểm định Shapiro-Wilk được tích hợp sẵn trong thư viện Scipy:

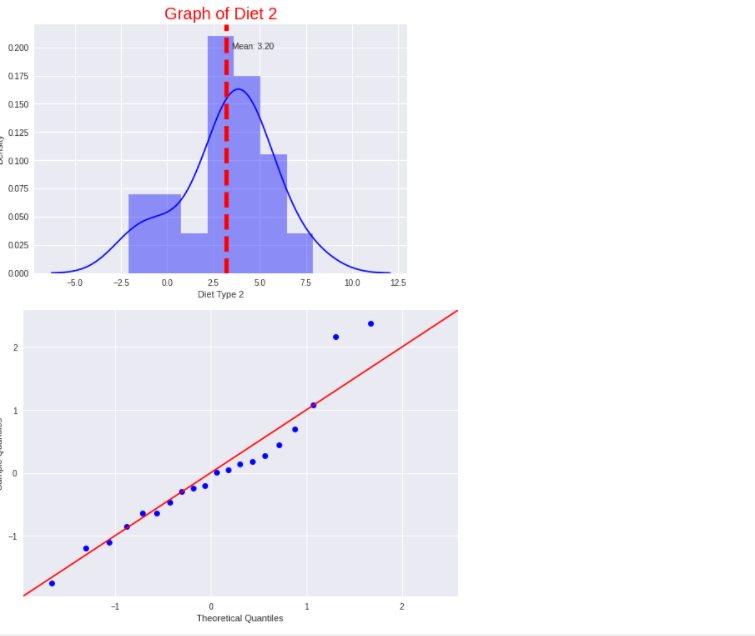


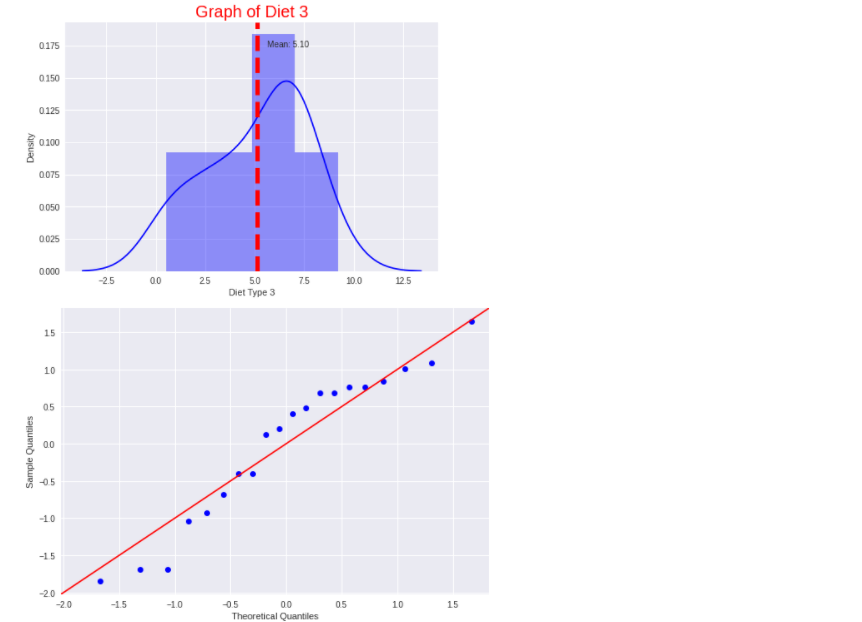




Tương ứng có các biểu đồ:







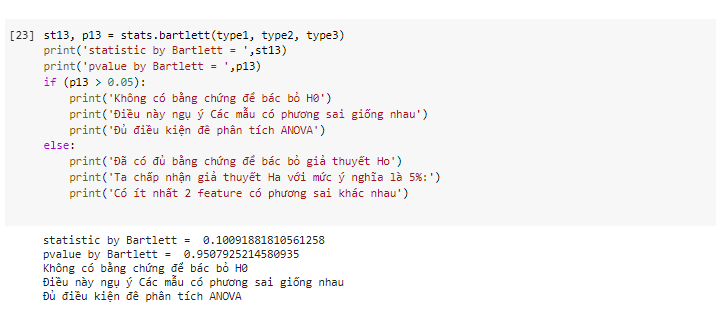
Kiểm định tính đồng nhất phương sai

- 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.

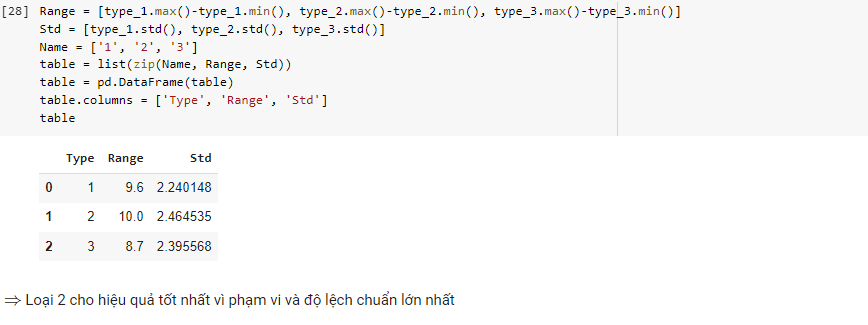


Phân tích ANOVA:



**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.



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

- Xây dựng giả thiết:

H0: Hiệu quả giảm cân của nam và nữ là như nhau.

H1: Hiệu quả giảm cân của nam và nữ là khác nhau.

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.



# 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:

