**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: Đặng Văn Thắng**

*(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/ Chức năng 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. Thống kê được chia làm 2 loại:

• Thống kê mô tả (Descriptive Statistics).

• Thống kê suy diễn (Inferential Statistics).

**Thống kê** là một cách để hiểu dữ liệu được thu thập về con người chúng ta và thế giới xung quanh. Tất cả dữ liệu sẽ là vô nghĩa nếu chúng ta không có cách giải thích, diễn giải, đây là lí do **Thống kê** xuất hiện.

**Thống kê** được sử dụng trong đa ngành như tâm lý học, kinh doanh, khoa học vật lí, xã hội, nhân văn.... **Thống kê** sẽ là nền tảng của những kết quả phân tích, dự báo có độ chính xác cao nếu sử dụng các quy trình một cách hợp lí, đảm bảo tuân theo các phương pháp thích hợp. Thống kê được vận hành dưới một quy trình phức tạp không hề ngắn gọn, nó bắt đầu từ thiết kế mô hình nghiên cứu, lựa chọn và đo lường các biến, áp dụng kỹ thuật lấy mẫu và xác định cỡ mẫu, làm sạch dữ liệu và xác định loại phân tích thích hợp, …

#### 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/ Chức năng của thống kê:

* Thống kê giúp cung cấp sự hiểu biết tốt hơn và mô tả chính xác về 1 hiện tượng, giải thích mối liên hệ của các đối tượng có trong nghiên cứu.
* Thống kê giúp lập kế hoạch phù hợp và hiệu quả của một cuộc điều tra, thống kê trong bất kỳ lĩnh vực nghiên cứu.
* Thống kê giúp thu thập dữ liệu định lượng một cách thích hợp.
* Thống kê giúp trình bày dữ liệu phức tạo dưới dạng bảng, sơ đồ và đồ họa phù hợp để dễ hiểu và rõ ràng về dữ liệu.
* Số liệu thống kê giúp hiểu được bản chất và mô hình biến đổi của một hiện tượng để có thể hỗ trợ đưa ra dự báo trong tương lai.
* Thống kêu giúp rút ra các suy luận, kết luận hợ lý thông qua việc sử dụng các định lý, phương pháp và kỹ thuật được sử dụng trong thu thập và phân loại dữ liệu, phân tích và giải thích dữ liệu, đánh giá, tính toán độ tin cậy.
* Các công cụ thống ke có thể so sáng các bộ dữ liệu với nhau ( datasets) để xem chúng giống nhau, mức độ tương thích thế nào, và đặc điểm của dữ liệu ra sao.

#### 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.

Thế mạnh của ngôn ngữ lập trình Python trong phân tích dữ liệu có nhiều thư viện trong Python cho khoa học và tính toán số liệu, như SciPy và NumPy, được sử dụng cho những mục đích chung chung trong tính toán. Và, có những thư viện cụ thể như: EarthPy cho khoa học trái đất, AstroPy cho Thiên văn học,... Ngoài ra, Python còn được sử dụng nhiều trong machine learning, khai thác dữ liệu và deep learning.

### 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.

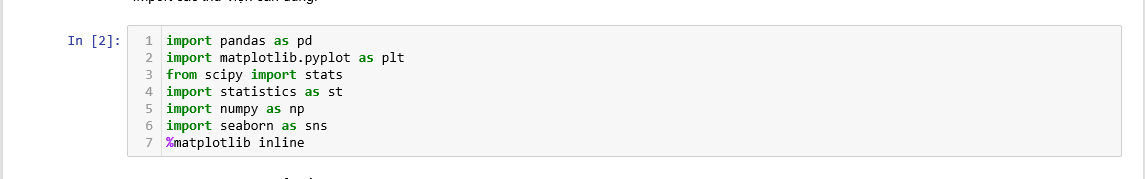
- Matplotlib có một giao diện thủ tục được đặt tên là Pylab, được thiết kế giống với MATLAB, một ngôn ngữ lập trình độc quyền được phát triển bởi MathWorks. Matplotlib cùng với NumPy có thể được coi là mã nguồn mở tương đương với MATLAB.

- 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:

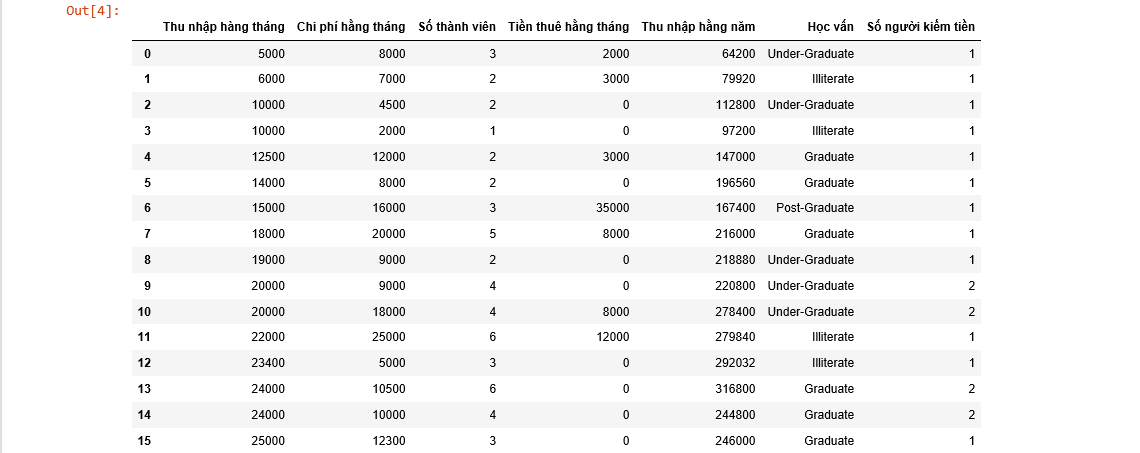


- Trong phần này, ta sử dụng tập dữ liệu Inc\_Exp\_Data.csv để phân tích thống kê mô tả.

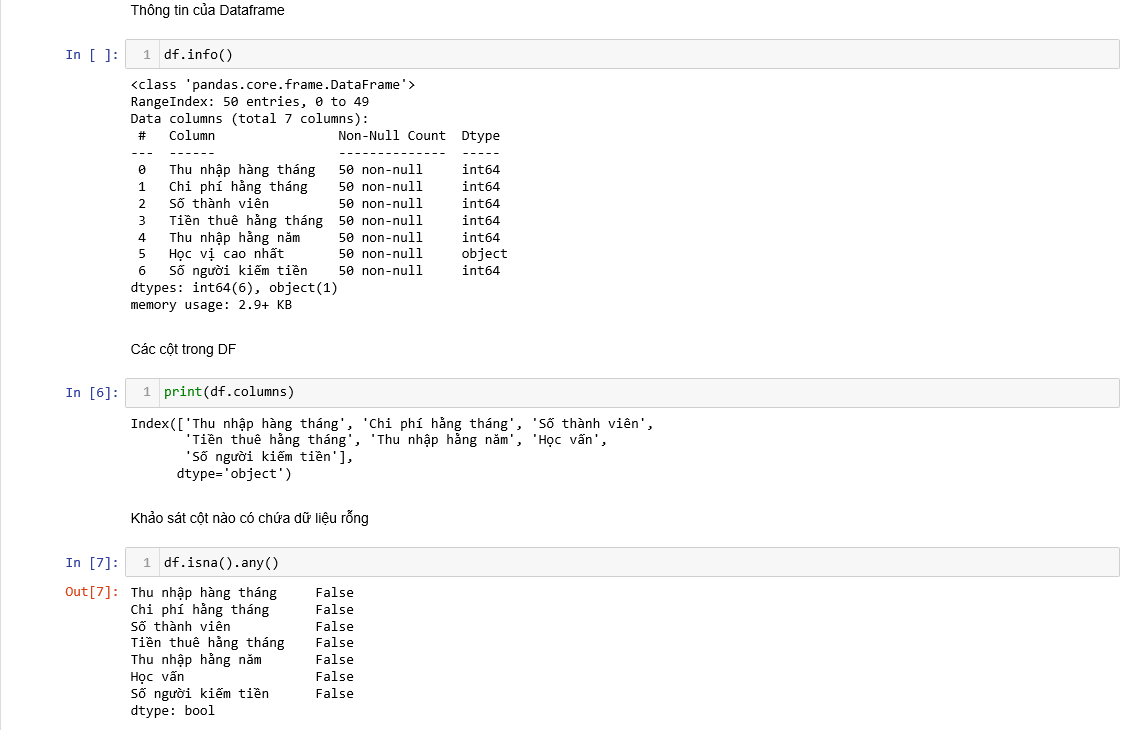
- Đầu tiên chúng ta truyền file vào bằng hàm read.csv của thư viện pandas:

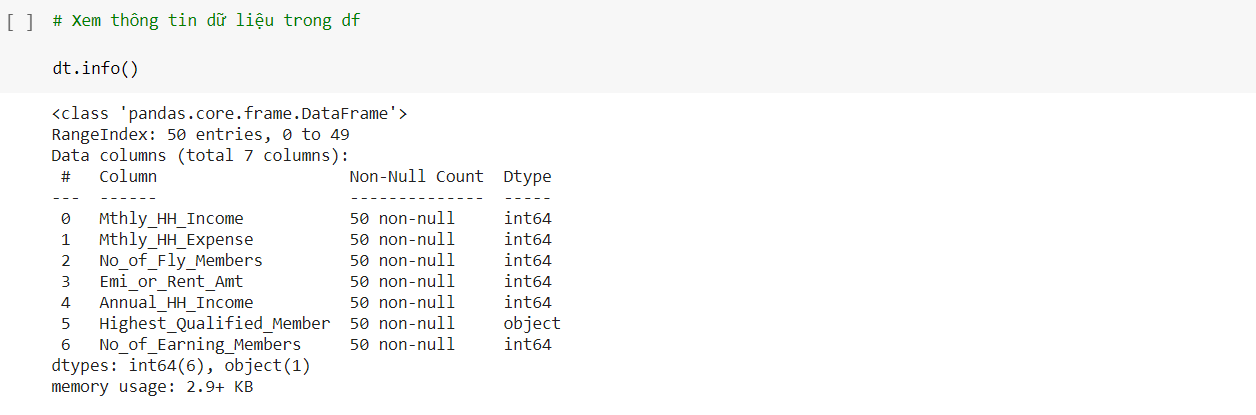


-Sau khi truyền vào ta được data:



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

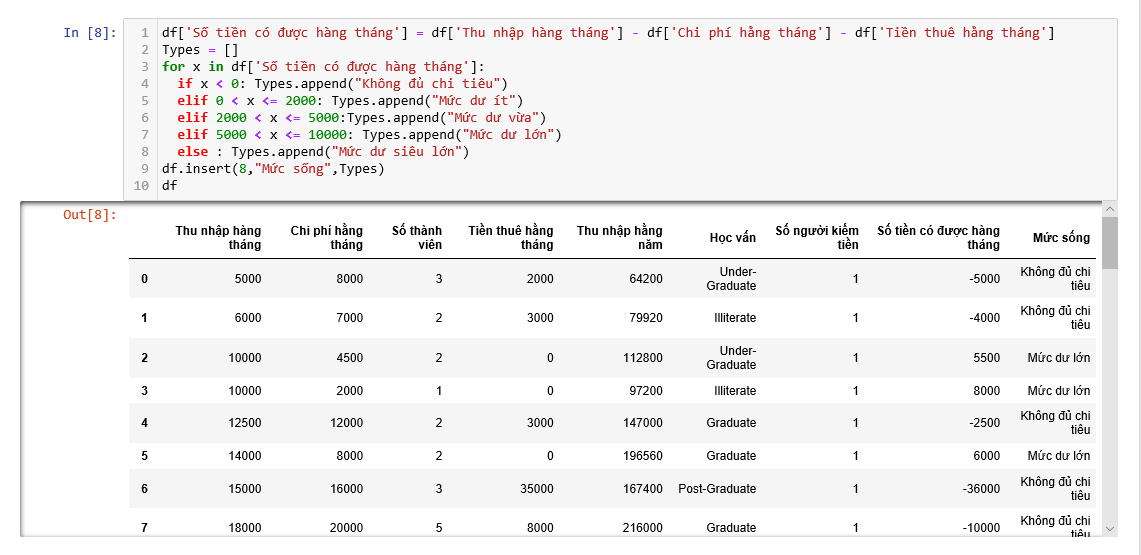




Tiếp theo ta đi vào phân tích dữ liệu.

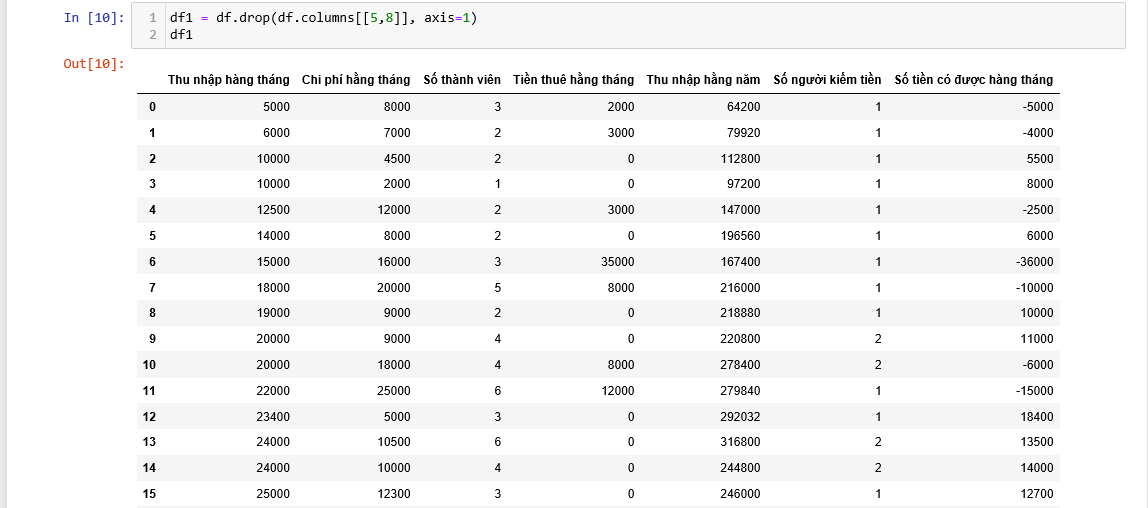
### Phân tích về sự phân phối của dữ liệu:

- 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 “Mức dư ít”, 80000$ thuộc loại “Mức dư lớn”.........

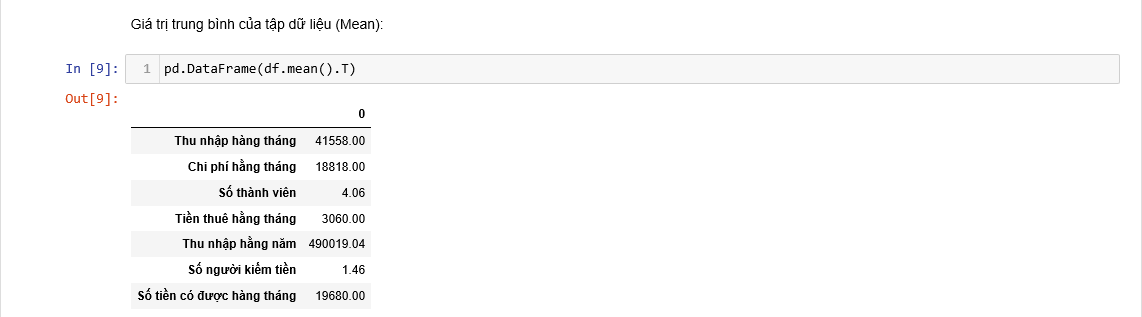


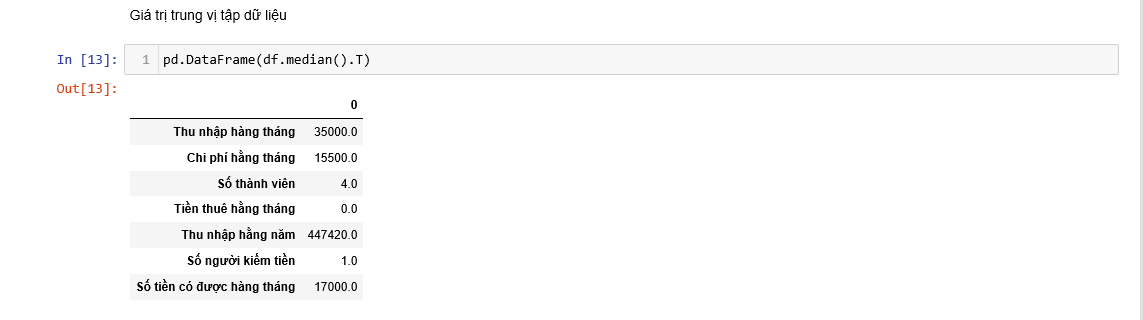
**2: Biện pháp phân tích xu hướng tập trung của trung tâm dữ liệu:**

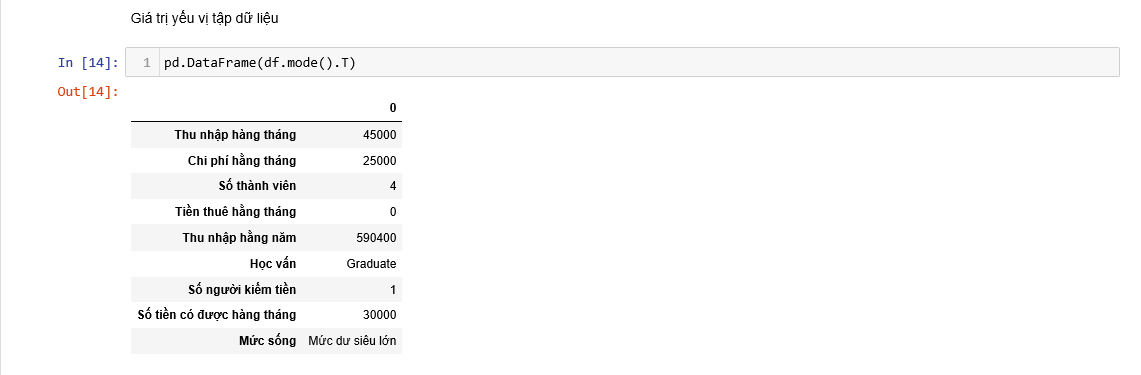
Lược bỏ dữ liệu định tính để dễ dàng tính toán.



Thực hiện các phép tính phân tích trung tâm:







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

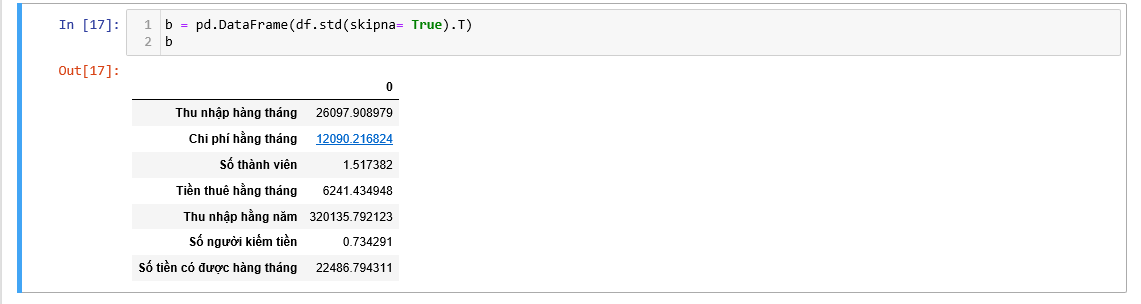
Phạm vi của dữ liệu:



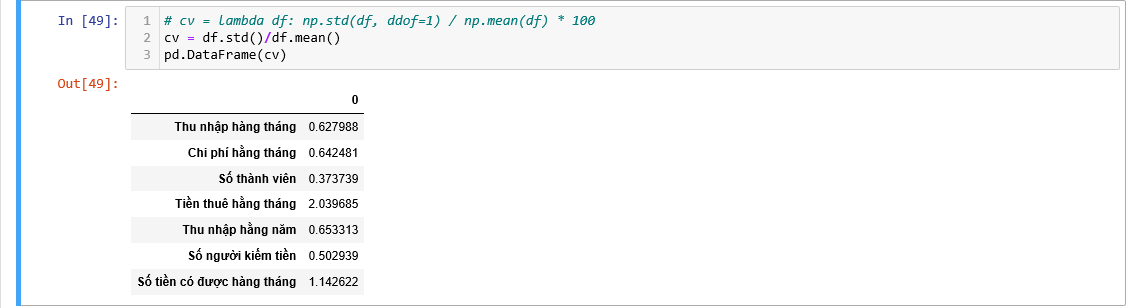
Phương sai của dữ liệu



Độ lệch chuẩn dữ liệu



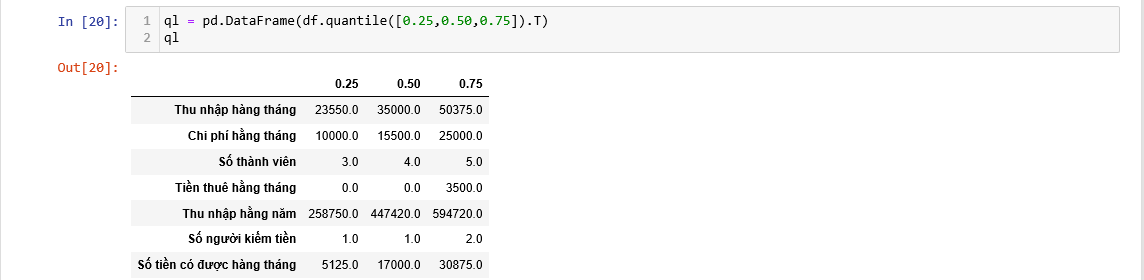
Hệ số biến động của dữ liệu



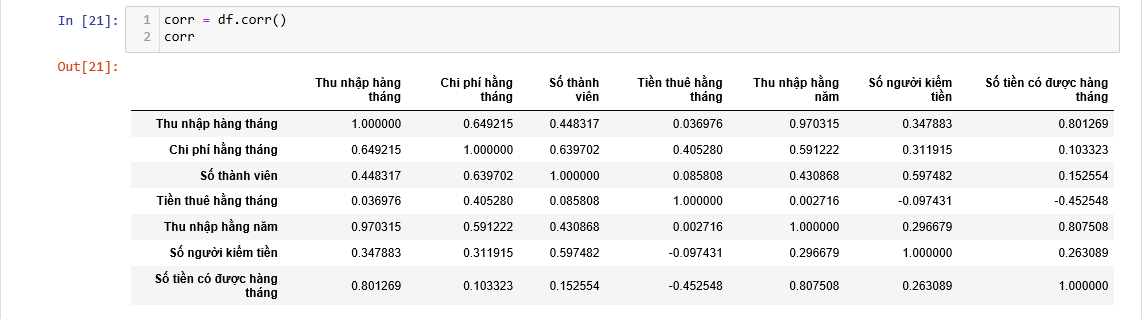
Độ lệch mẫu đo độ không đối xứng của dữ liệu



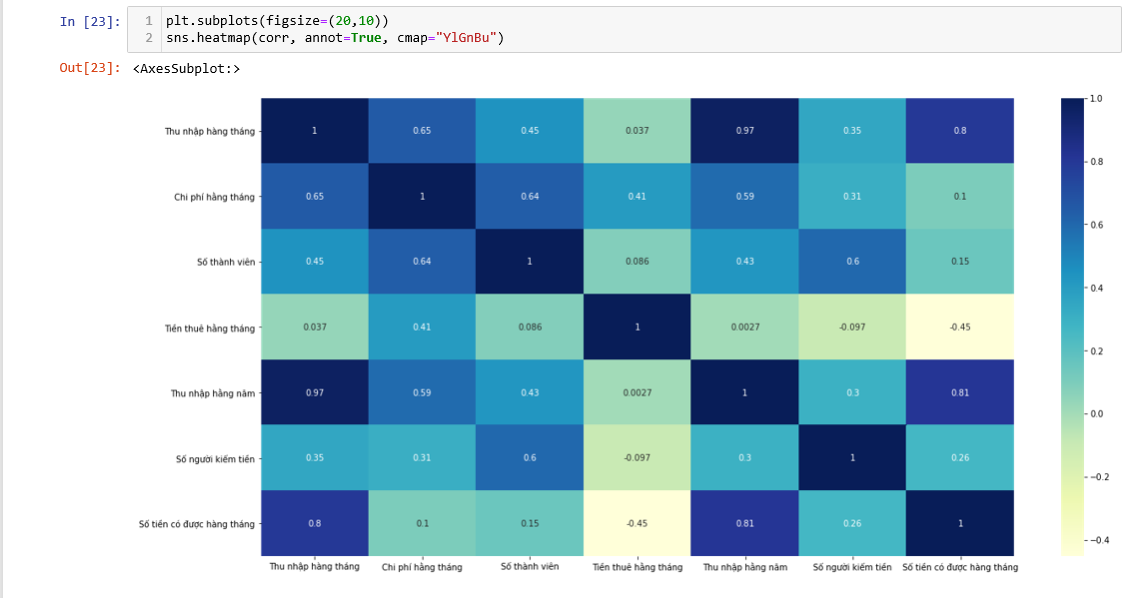
Phần trăm (Phân vị mẫu) của dữ liệu



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



Biểu đồ thống kê mức độ phụ thuộc của từng tập dữ liệu trong data



**4: Tóm tắt thống kê mô tả:**

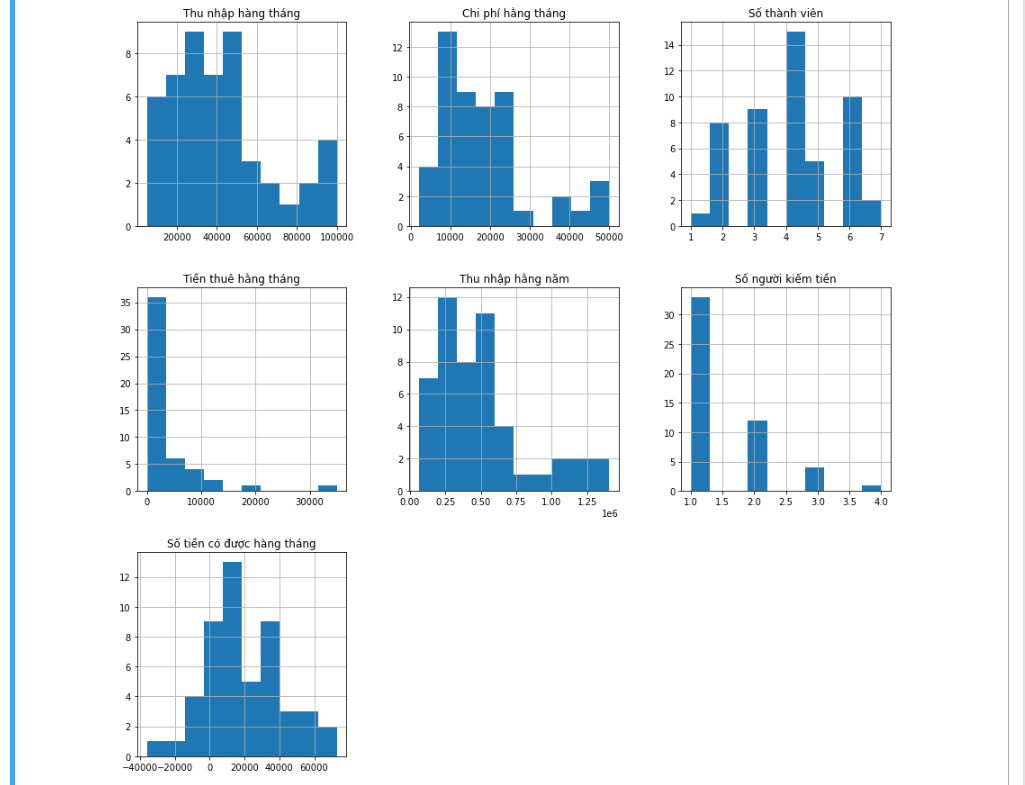
Bảng tóm tắt thống kê mô tả



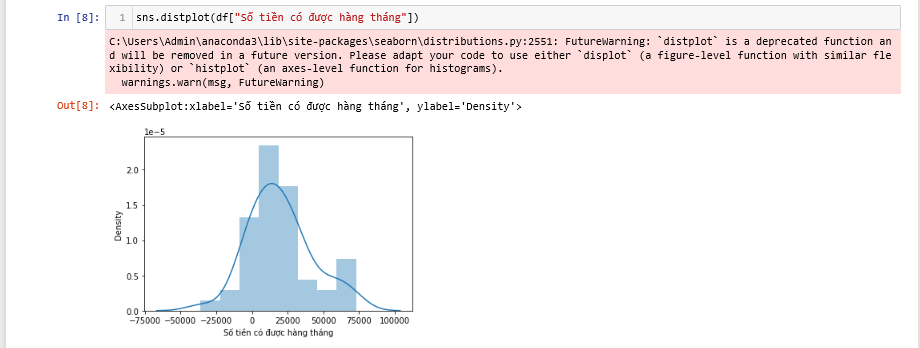
**5: Trực quan hóa dữ liệu:**

Biểu đồ thế hiện các cột số liệu:

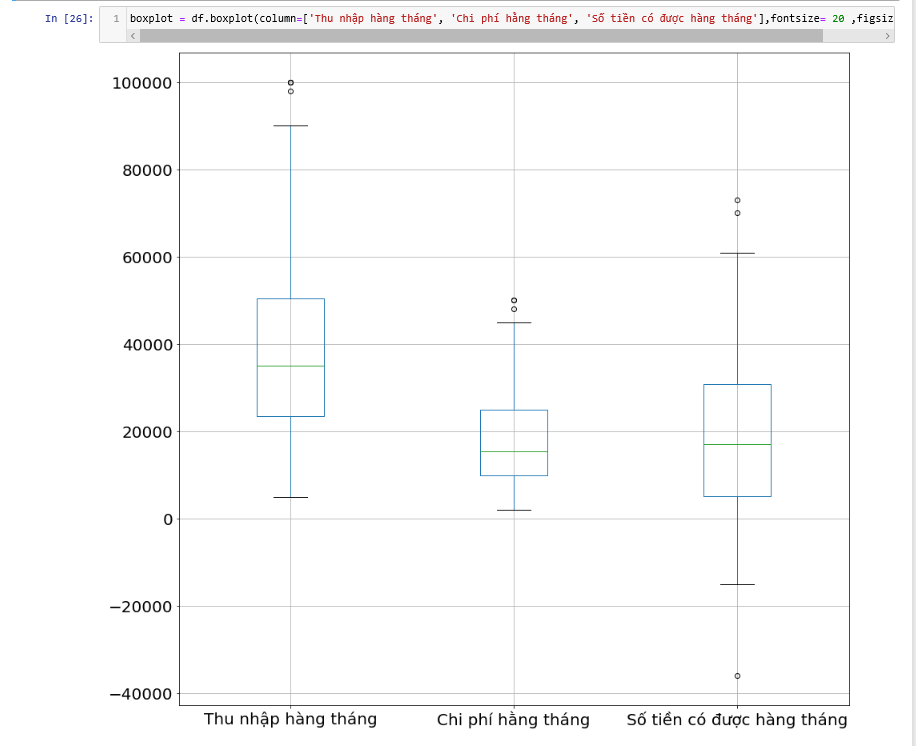




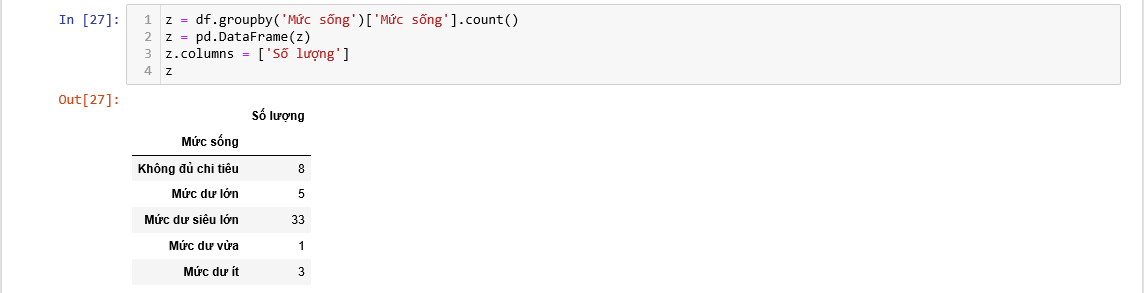
Biểu đồ cột thể hiện số tiền có được hàng tháng của tập dữ liệu



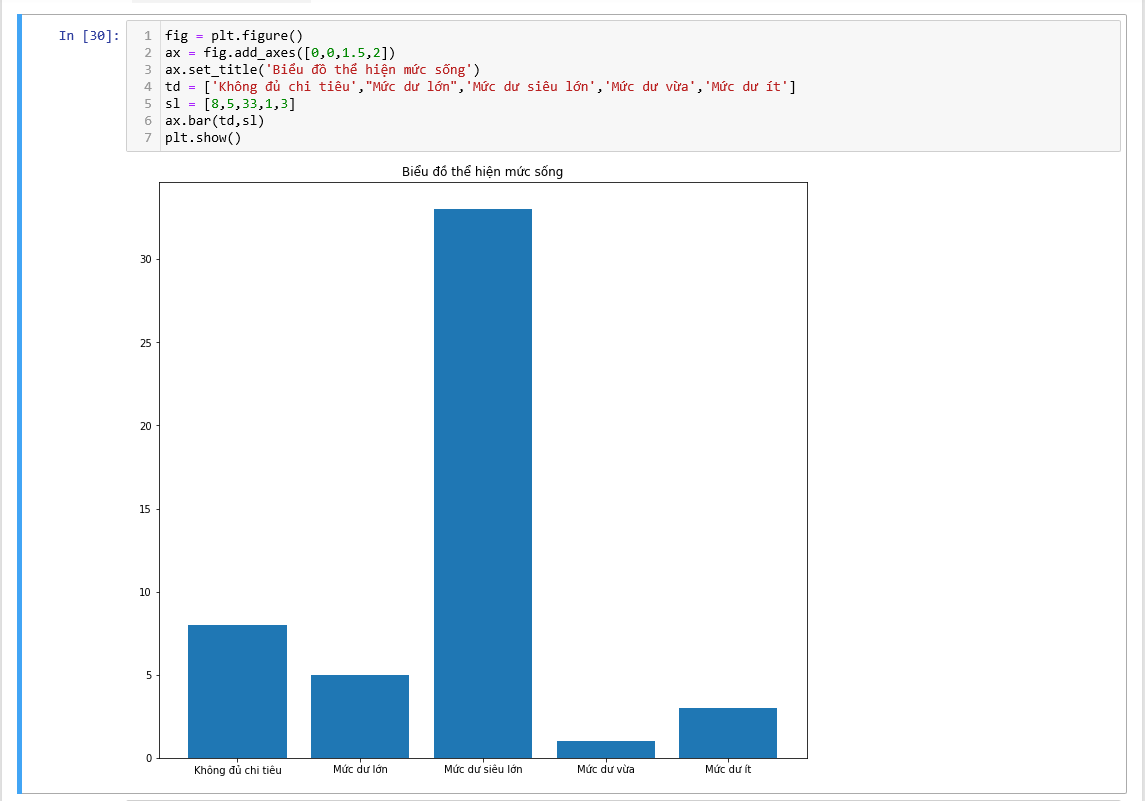
Biểu đồ trình bày trực quan về thống kê mô tả của tập dữ liệu



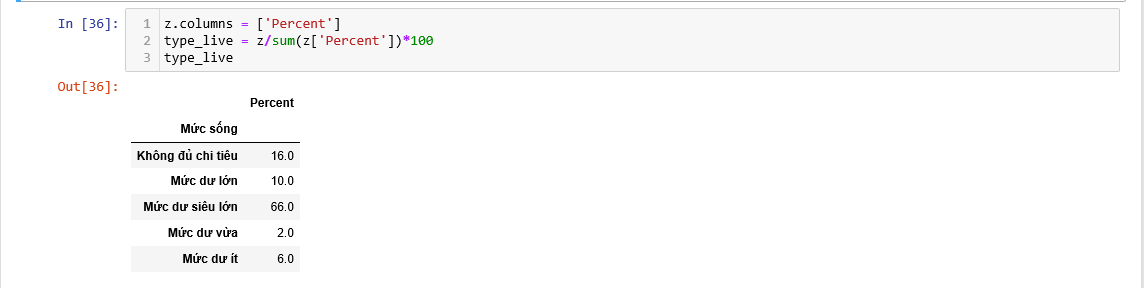
Tỉ lệ mức sống trong tập dữ liệu:



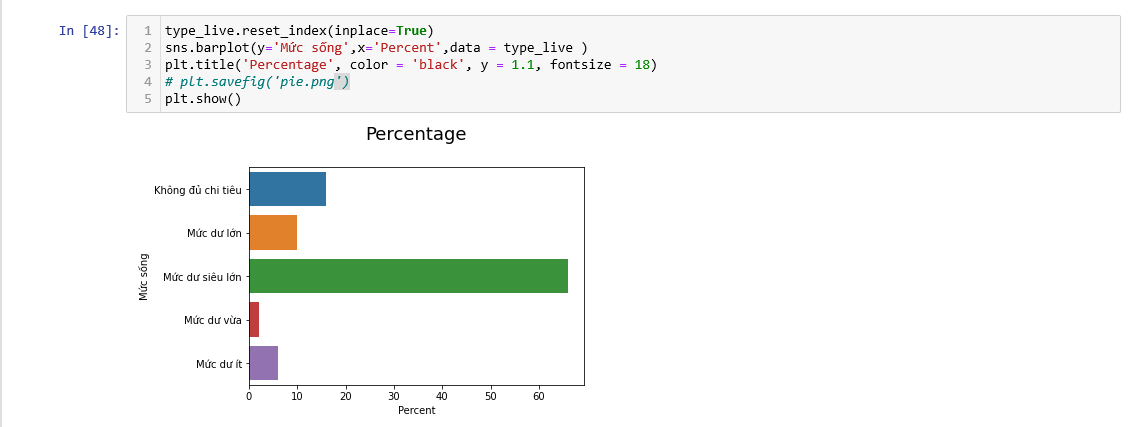
Biểu đồ thế hiện mức sống:



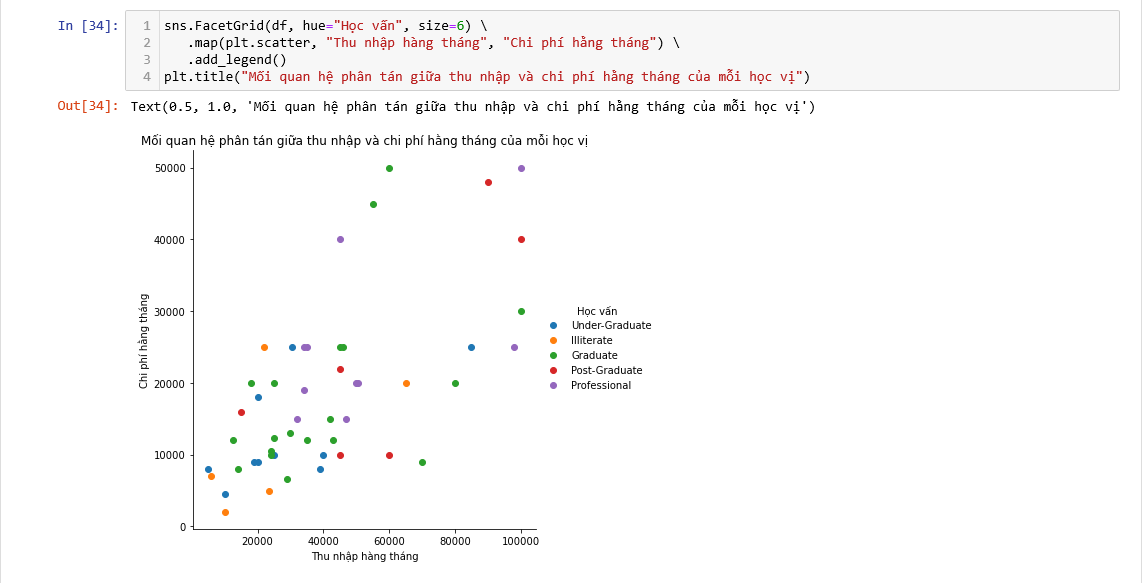
Thể hiện phần trăm:



Biểu đồ phần trăm:

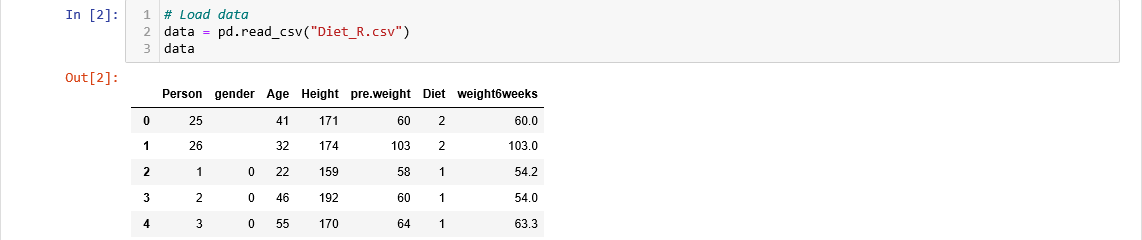


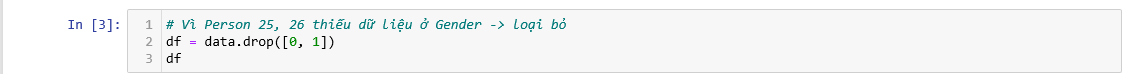
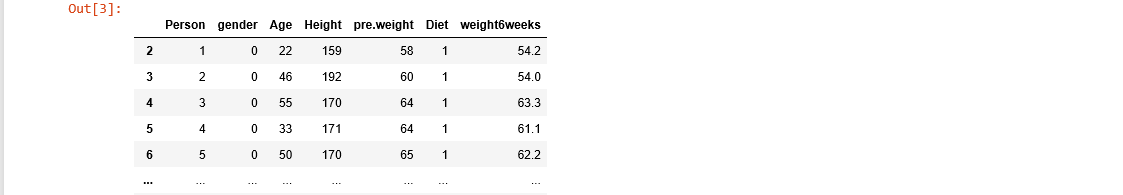
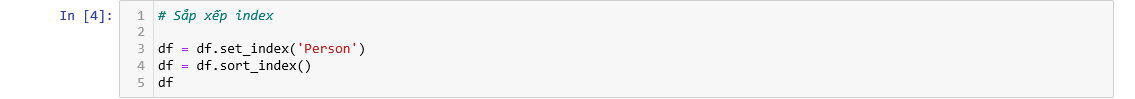
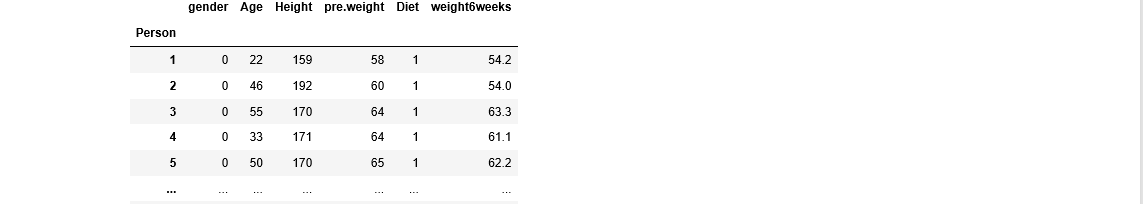
Biểu đồ thể hiện sự phân tán giữa thu nhập và chi phí của từng nhóm học vấn:



## II.Thống kê suy diễn:

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

Đọc và chỉnh sửa Data: 

### 1.Kiểm định T Test:

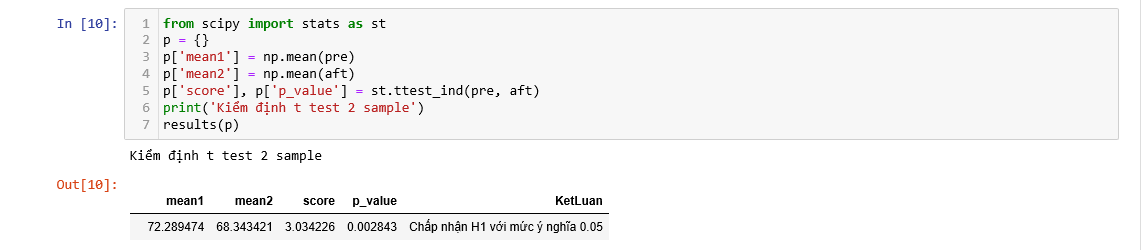
- Vậy chúng ta có giả thiết không và đối thiết như sau:

* 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”.

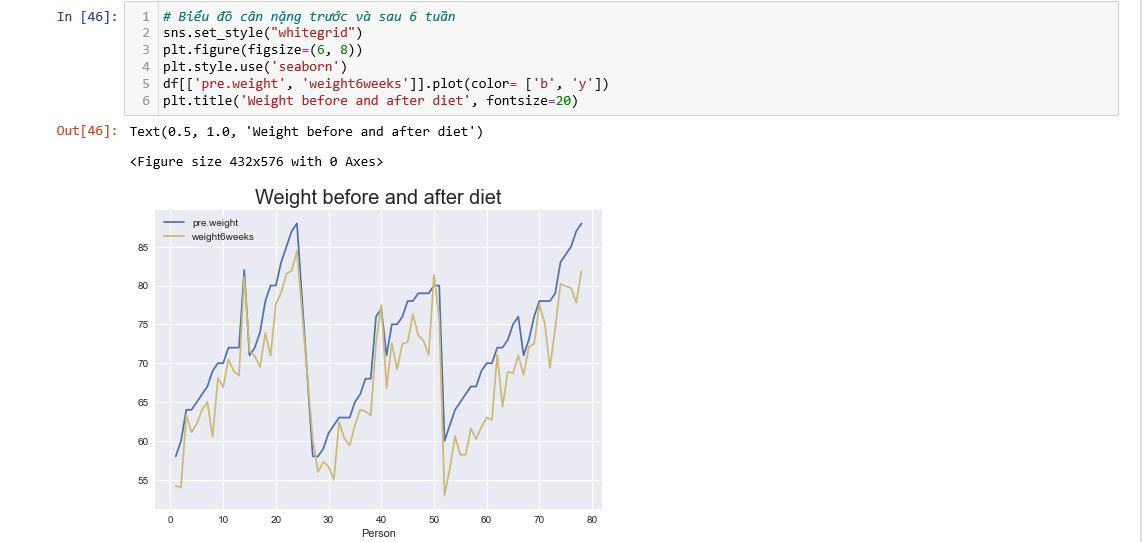


Kiểm định t test:

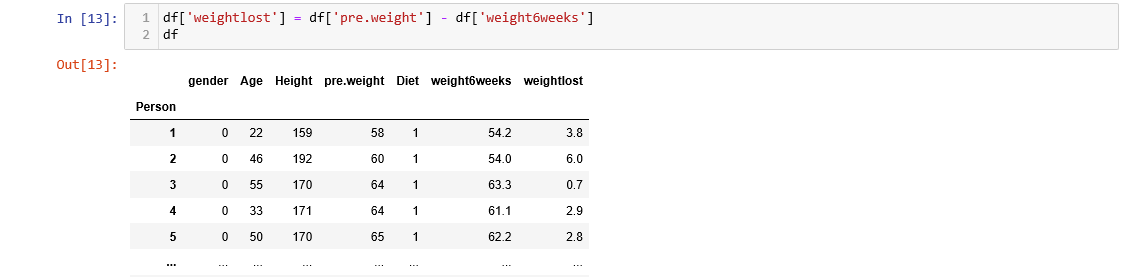


- Như vậy với mức tin cậy là 5% chúng ta có đủ tự tin để khẳng định rằng sau khi thực hiện phương pháp giảm cân là có hiệu quả.

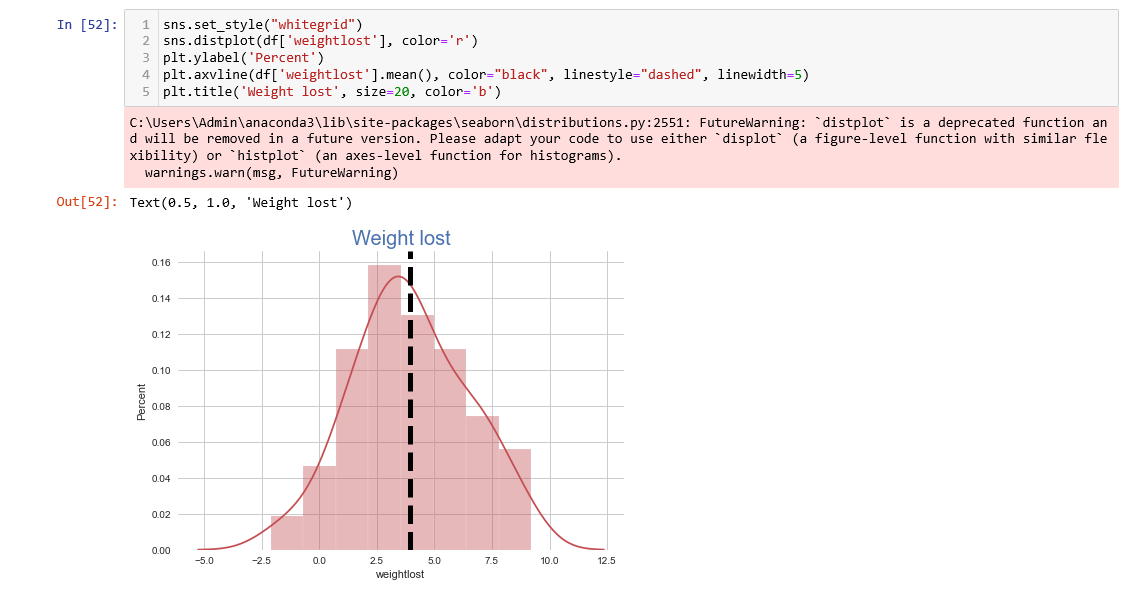
Biểu đồ trước và sau khi giảm cân:



### 2.Thực hiện tính toán Weight lost:



Biểu đồ trực quan:



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

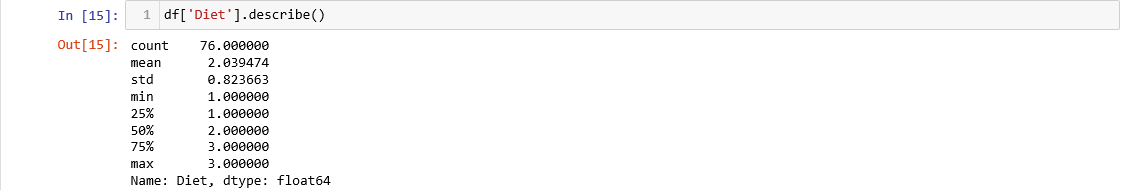
- Xây dựng giả thiết:

* 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

### 3.Thống kê tóm tắt:

Tóm tắt thống kê chế độ ăn:

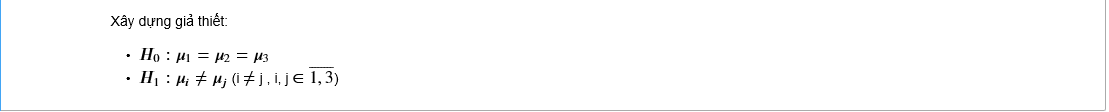


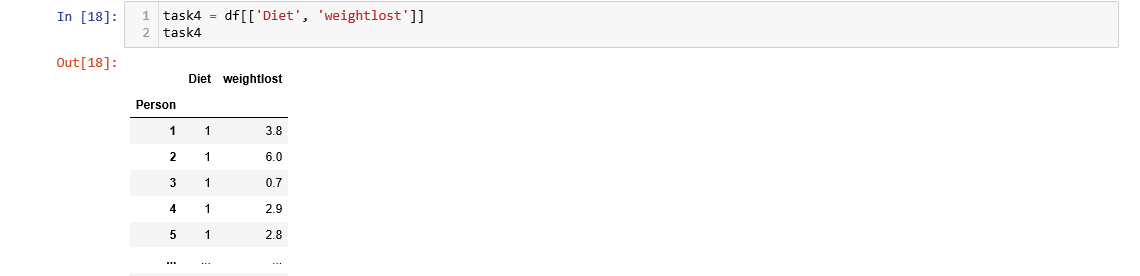




### 4.Kiểm định ANOVA:

4.1: Phương pháp Diet hữu hiệu:

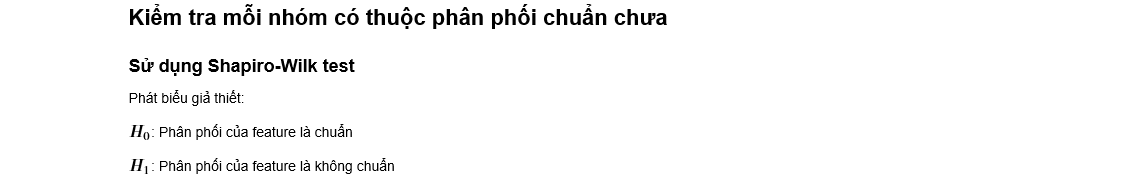


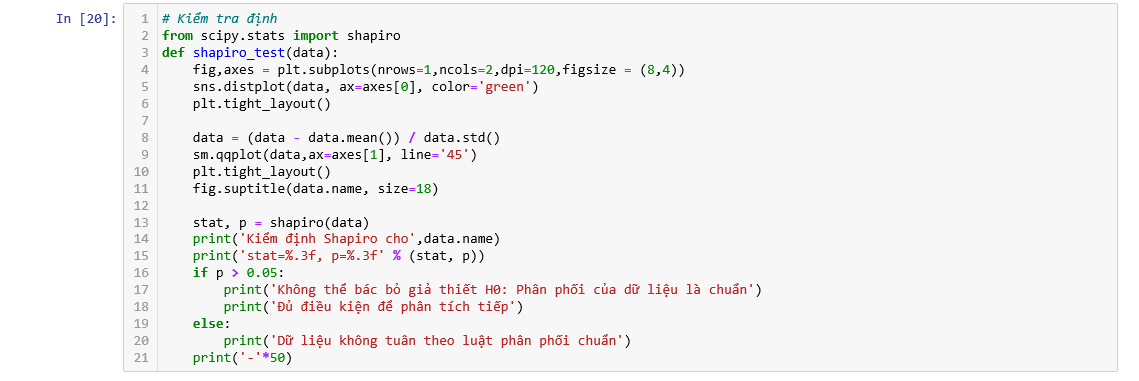


- Chú ý rằng, để thực hiện kiểm định ANOVA thì dữ liệu phải thỏa mãn các yêu cầu sau:

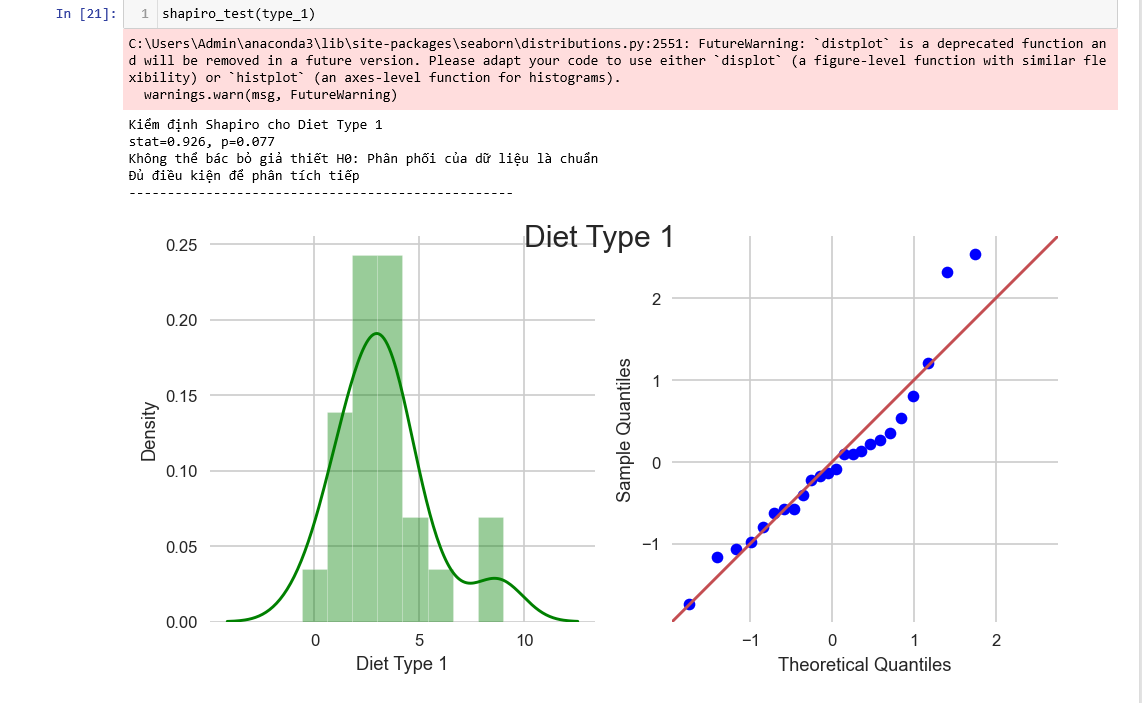
* Dữ liệu phải tuân theo phân phối chuẩn.
* Dữ liệu đồng nhất về phương sai .
* Các cá thể phải được lấy độc lập với nhau.

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:



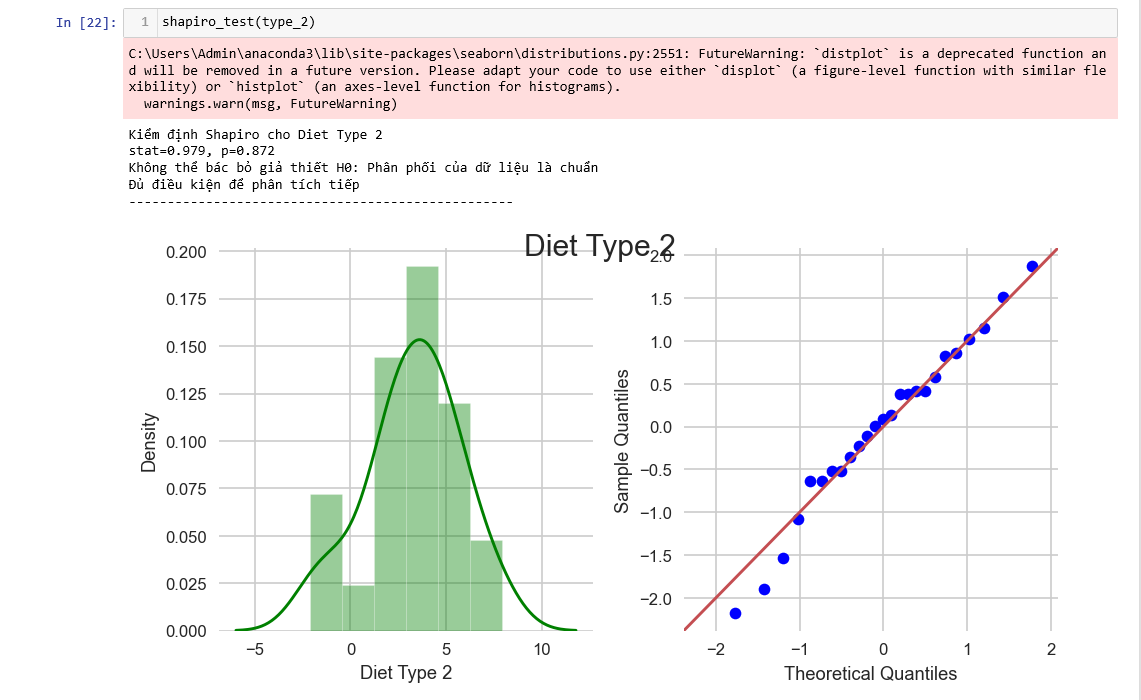
Kiểm định tính đồng nhất phương sai

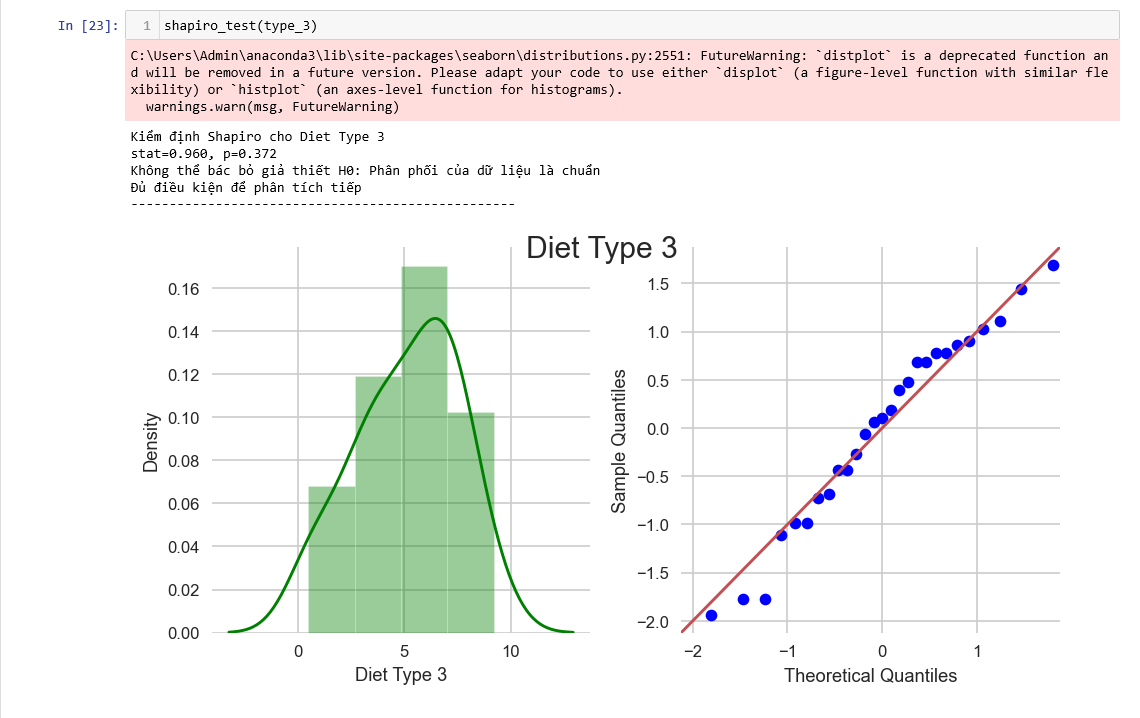
- Có nhiểu phương pháp để kiểm định tính đồng nhất phương sai như kiểm định Bartlett, kiểm định Levene, …

- Phát biểu giả thiết:

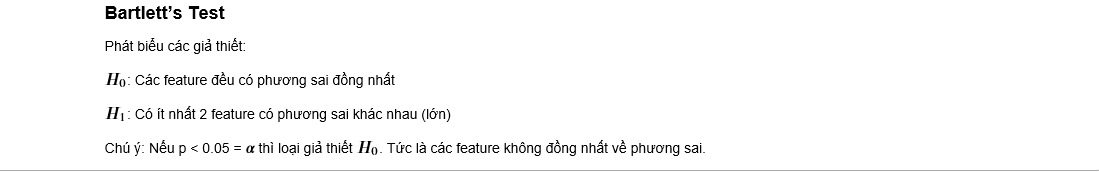
H0: Các feature đồng nhất phương sai.

H1: Các feature không đồng nhất phương sai.

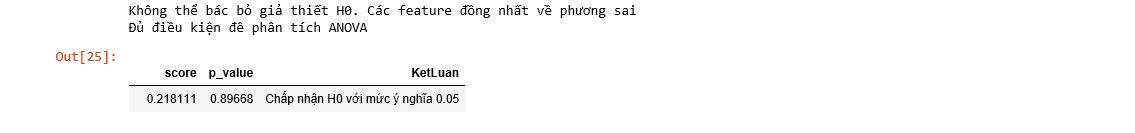




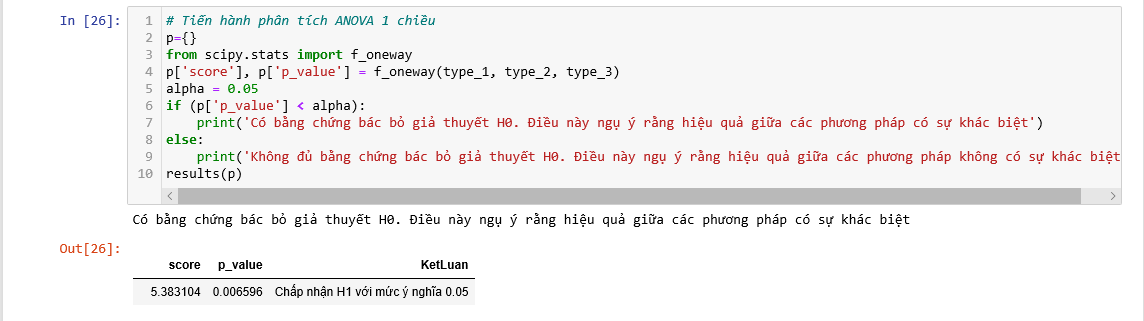
Kiểm định tính thuần nhất phương sai bằng phương pháp Bartlett:





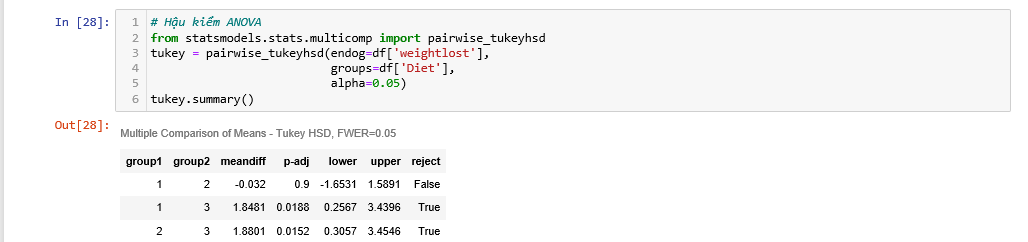


Phân tích ANOVA:



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.

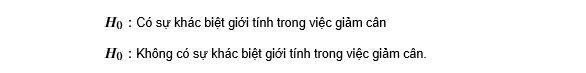
Hậu nghiệm ANOVA:



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:

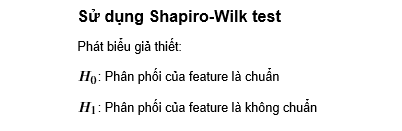


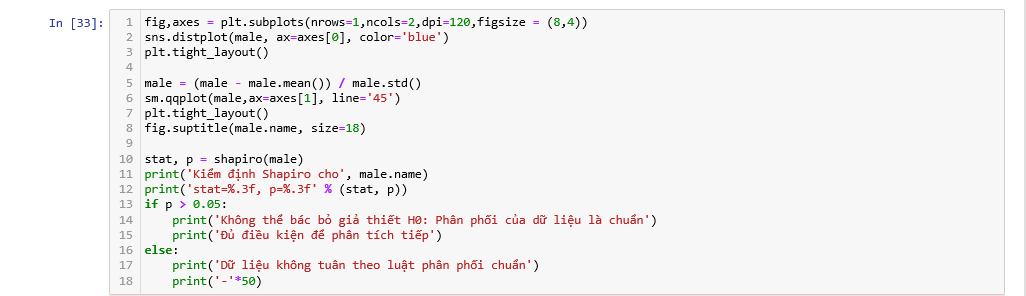
Để trả lời cho câu hỏi trên chúng ta tiếp tục thực hiện kiểm định ANOVA qua các bước trích xuất, kiểm tra, kiểm định dữ liệu và đưa ra kết luận.

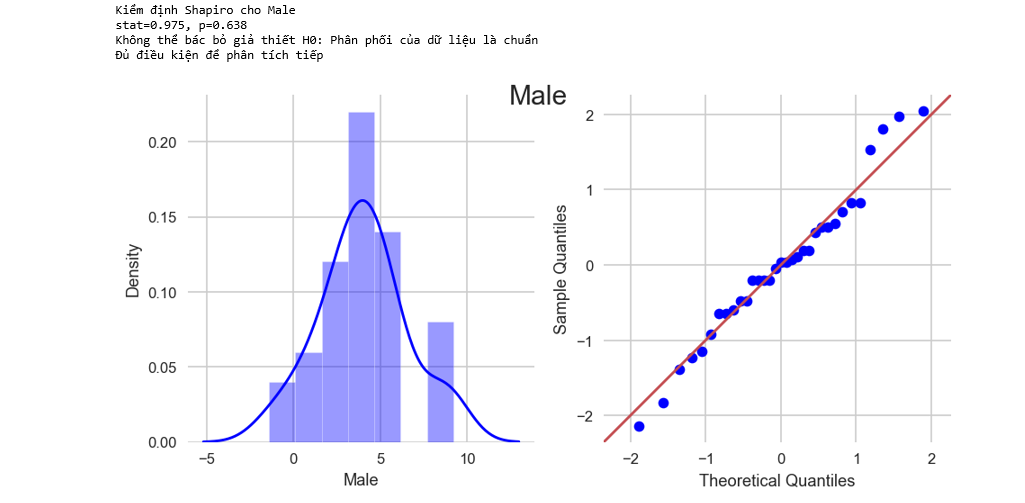


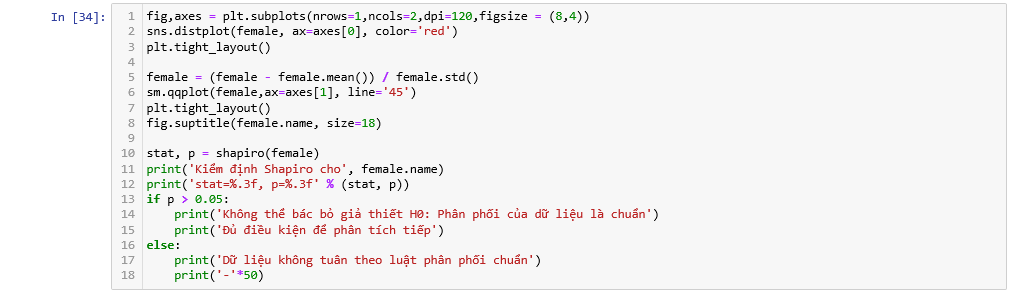


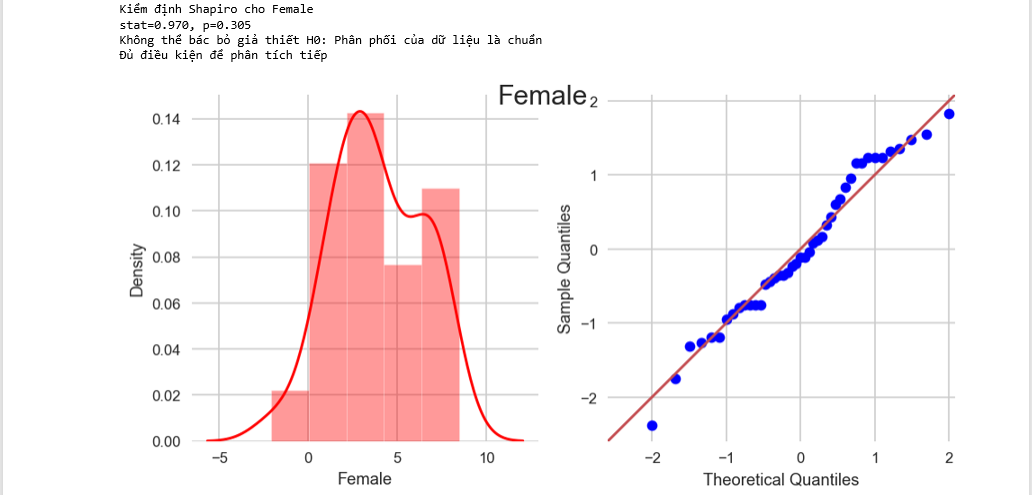
Kiểm tra mỗi nhóm thuộc phân phối chuẩn chưa:



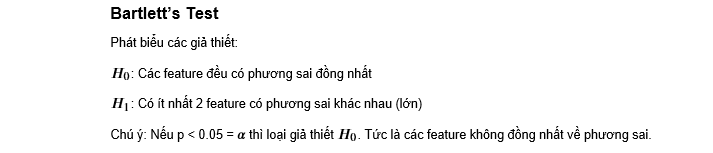


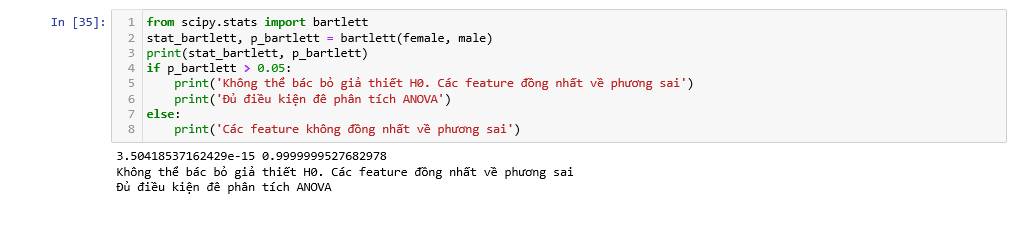




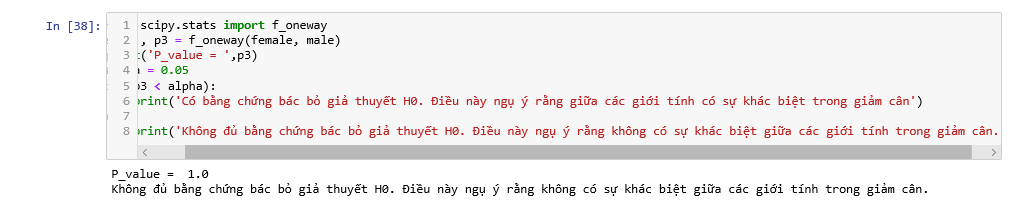


Kiểm định tính thuần nhất của phương sai:





Phân tích ANOVA:



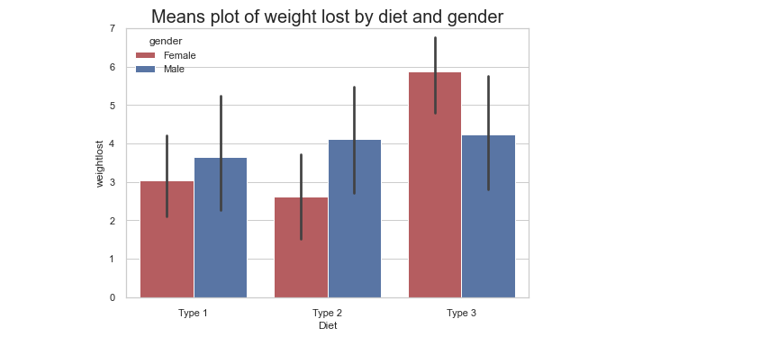
Với P value thu được ta không đủ bằng chứng bác bỏ giả thuyết H0, điều này ngụ ý răng hiệu quả giữa các phương pháp không bị ảnh hưởng bởi giới tính.

Hậu nghiệm ANOVA:

Do không đủ bằng chứng bác bỏ giả thuyết H0 nên t bỏ qua bước hậu nghiệm.

**5.Tương tác giữa các số liệu:**

Biểu đồ ý nghĩa việc giảm cân với giới tính và phương pháp:



# Chương 3: Tài liệu tham khảo

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

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

<https://matplotlib.org/>

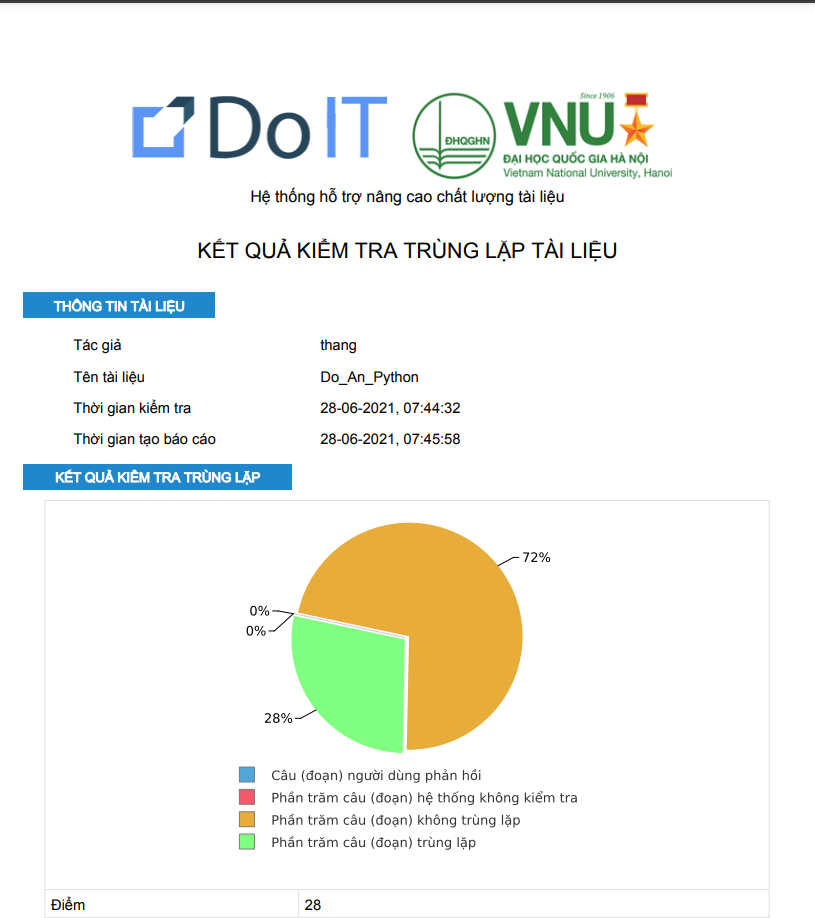
<https://stackoverflow.com/>

<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:



http://doit.uet.vnu.edu.vn/documents/do-an-python/report/reportdo-an-python.docx.pdf.