**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: LÊ VĂN TÂN**

*(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](#_Toc76684345)

[ĐẠI HỌC HUẾ 2](#_Toc76684346)

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 2](#_Toc76684347)

[Mục lục 1](#_Toc76684348)

[Chương I: CƠ SỞ LÝ THUYẾT 2](#_Toc76684349)

[I.1 Thống kê là gì? 2](#_Toc76684350)

[1.1 Định nghĩa 2](#_Toc76684351)

[1.2 Thống kê được sử dụng trong Python như thế nào? 2](#_Toc76684352)

[1.2.1 Ngôn ngữ lập trình Python 2](#_Toc76684353)

[1.2.1 Giới thiệu về các thư viện 3](#_Toc76684354)

[Chương 2: Phân tích dữ liệu với các tập dữ liệu 4](#_Toc76684355)

[2.1 Thống kê mô tả 4](#_Toc76684356)

[2.2 Thống kê suy diễn 12](#_Toc76684357)

[TÀI LIỆU THAM KHẢO 17](#_Toc76684358)

[KIỂM TRA ĐẠO VĂN 17](#_Toc76684359)

# Chương I: CƠ SỞ LÝ THUYẾT

## I.1 Thống kê là gì?

### 1.1 Định nghĩa

- Thông kê là nghiên cứu của tập hợp nhiều lĩnh vực khác nhau, bao gồm phân tích, giải thích , trình bày và tổ chức dữ liệu.

- 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.

- 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.

- Hai phương pháp chính của thống kê là:

* Thống kê mô tả: là các phương pháp liên quan đến việc thu thập số liệu, tóm tắt, trình bày, tính toán và mô tả các đặc trưng khác nhau để phản ánh một cách tổng quát đối tượng nghiên cứu.
* Thống kê suy luận: là bao gồm các phương pháp ước lượng các đặc trưng của tổng thể, phân tích mối liên hệ giữa các hiện tượng nghiên cứu, dự đoán hoặc ra quyết định trên cơ sở thu thập thông tin từ kết quả quan sát mẫu.

## 1.2 Thống kê được sử dụng trong Python như thế nào?

### 1.2.1 Ngôn ngữ lập trình Python

- Khi nhắc đến một ngôn ngữ lập trình phổ biến thì nhiều người sẽ nghĩ ngay đến Python.Đây là ngôn ngữ lập trình thiên hướng đến các đối tượng bậc cao.Python được sử dụng rất nhiều trong quá trình phát triển ứng dụng và website khác nhau.Trong một dự án mở của mình ông Guido van Rossum đã tạo ngôn ngữ lập trình thông minh này.Python được biết đến là ngôn ngữ lập trình rất thích hợp cho những người mới muốn học lập trình.Bởi vì nó có cú pháp rất đơn giản nhưng rõ ràng,hình thức trình bày không cầu kỳ.Ngôn ngữ lập trình Python có khả năng di động linh hoạt tương thích với rất nhiều hệ điều hành.Python là ngôn ngữ lập trình phù hợp,dễ hiểu để đưa vào các chương trình dạy học ở các trường Đại học.Nhưng thậm chí dạy Python cho cấp THPT cũng không phải là quá khó khăn.Python có khả năng phân tích dữ liệu do nó sở hữu một thư viện chuẩn lớn và rất đa dạng như Numpy,SciPy,Panda…Vì vậy Python được xem như là một công cụ tính toán và phân tích trong rất nhiều lĩnh vực phải kể đến như Data Analytics,Computing Science,Machine Learing.Và không dễ để tìm được một ngôn ngữ lập trình nào khác có thể làm tốt việc này hơn Python.Điểm mạnh và đặc biệt ở đây nằm ở các thư viện mà Python sở hữu.

- 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.

### 1.2.1 Giới thiệu về các thư viện

#### Thư viện Matplotlyb

- Để thực hiện các suy luận thống kê cần thiết, cần phải trực quan hóa dữ liệu của bạn và Matplotlib là một trong những giải pháp như vậy cho người dùng Python. Nó là một thư viện vẽ đồ thị rất mạnh mẽ hữu ích cho những người làm việc với Python và NumPy. Module được sử dụng nhiều nhất của Matplotib là Pyplot cung cấp giao diện như MATLAB nhưng thay vào đó, nó sử dụng Python và nó là nguồn mở.

- Matplotlib ban đầu được viết bởi John D. Hunter vào năm 2003.

- 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.

#### b/ Thư viện Pandas

- Thư viện pandas trong python là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python. Thư viện này được sử dụng rộng rãi trong cả nghiên cứu lẫn phát triển các ứng dụng về khoa học dữ liệu. Thư viện này sử dụng một cấu trúc dữ liệu riêng là Dataframe. Pandas cung cấp rất nhiều chức năng xử lý và làm việc trên cấu trúc dữ liệu này. Chính sự linh hoạt và hiệu quả đã khiến cho pandas được sử dụng rộng rãi.

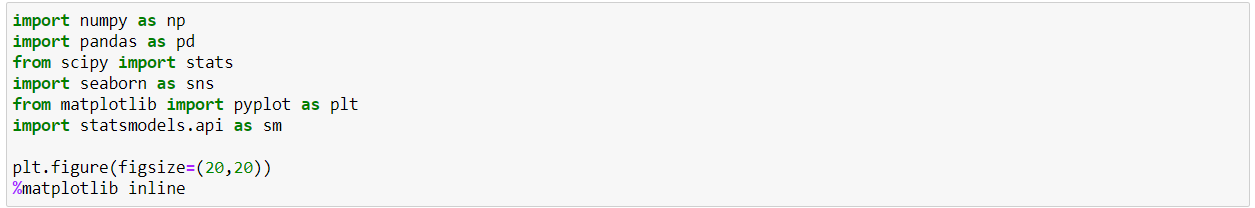
- Pandas là một thư viện mã nguồn mở được xây dựng dựa trên NumPy, sử dụng thao tác và phân tích dữ liệu, được thiết kế để cho phép bạn làm việc với dữ liệu được gắn nhãn hoặc quan hệ theo cách trực quan hơn

* Có thể xử lý tập dữ liệu khác nhau về định dạng: chuỗi thời gian, bảng không đồng nhất, ma trận dữ liệu
* Khả năng import dữ liệu từ nhiều nguồn khác nhau như CSV, DB/SQL
* Có thể xử lý vô số phép toán cho tập dữ liệu: subsetting, slicing, filtering, merging, groupBy, re-ordering, and re-shaping,..
* Xử lý dữ liệu mất mát theo ý người dùng mong muốn: bỏ qua hoặc chuyển sang 0
* Xử lý, phân tích dữ liệu tốt như mô hình hoá và thống kê
* Tích hợp tốt với các thư viện khác của python
* Cung cấp hiệu suất tốt

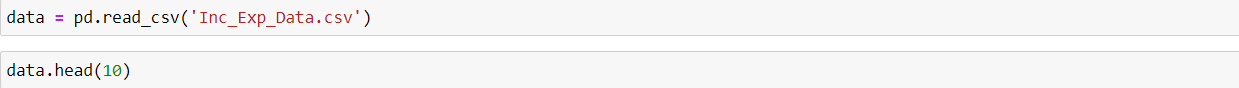
# Chương 2: Phân tích dữ liệu với các tập dữ liệu

## 2.1 Thống kê mô tả

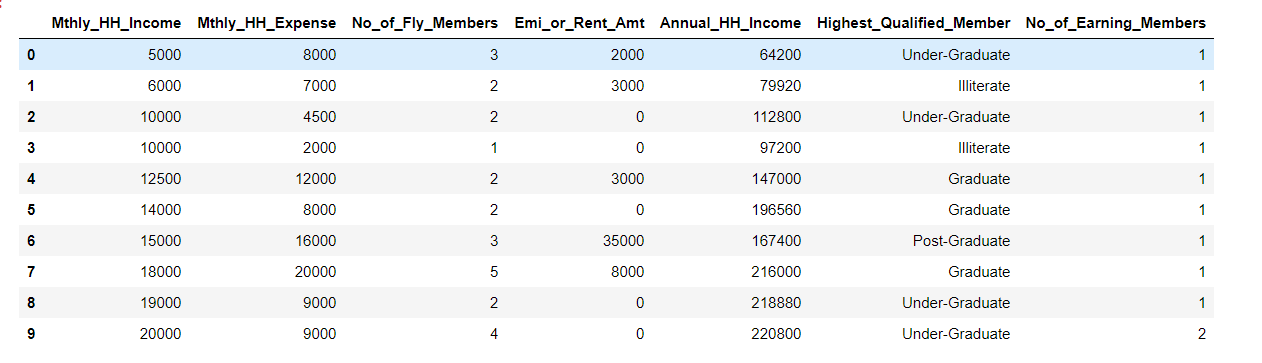
- Bước đầu tiên chúng ta import các thư viện cần thiết vào :



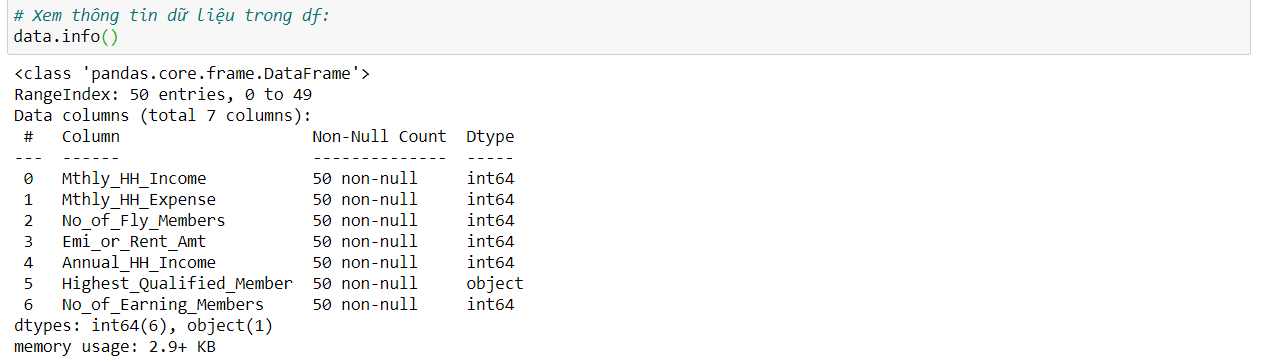
-Chúng ta sử dụng tập dữ liệu Inc\_Exp\_Data.csv nên chúng ta sẽ đọc tập dữ liệu vào để tiến hành thống kê mô tả.



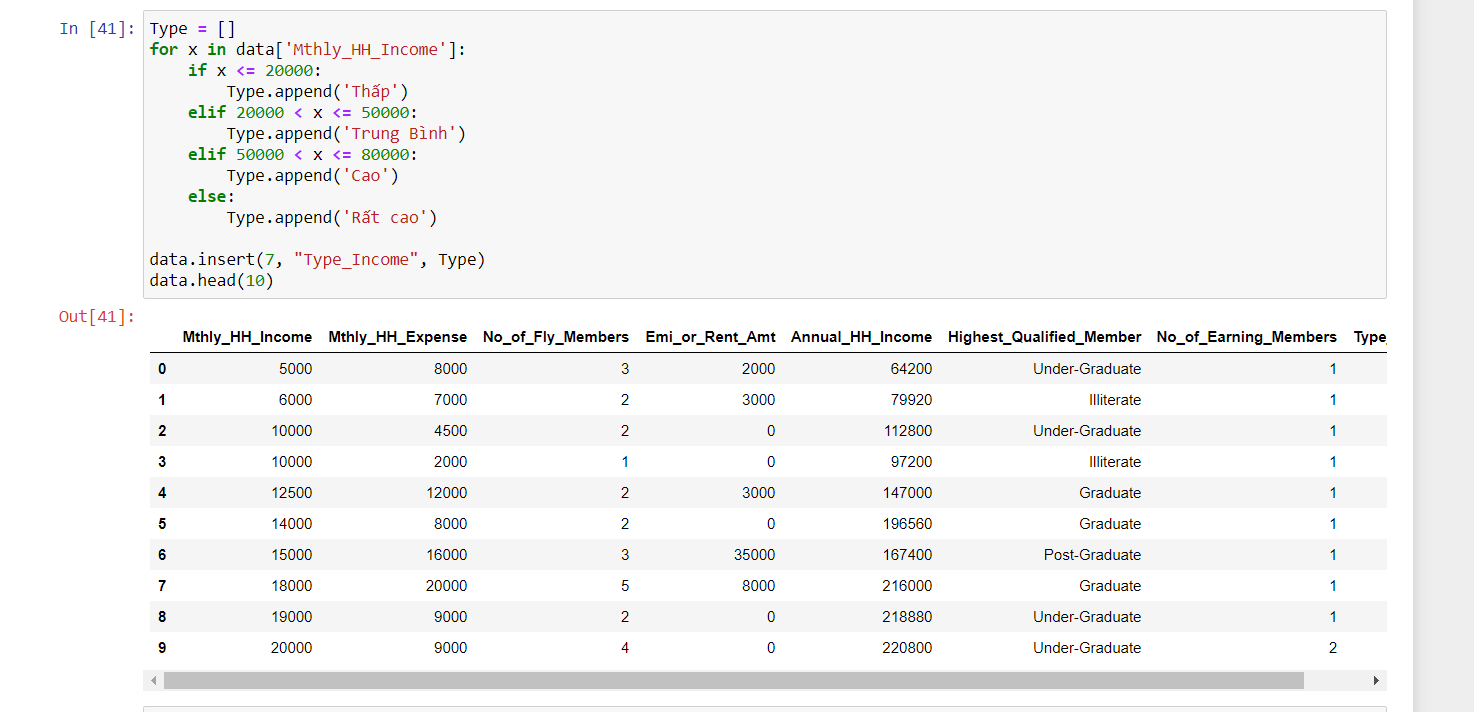
Sau khi truyền vào thì tập dữ liệu sẽ được như sau:



Ta sử dụng hàm info() để kiểm tra các thông tin của dữ liệu:



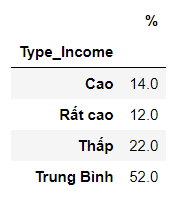
2.1.1 Phân tích về sự phân phối của dữ liệu

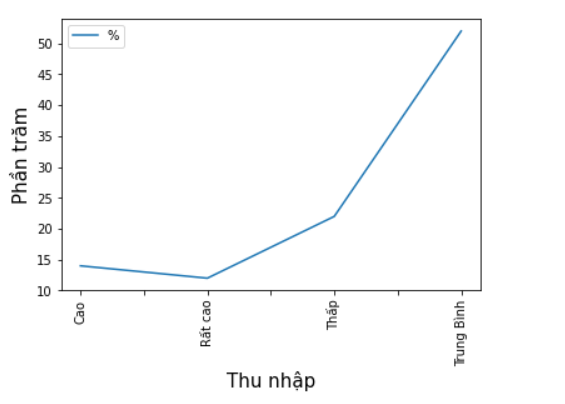


Đây là mẫu lọc các loại thu nhập:

* <= 20000 : Có thu nhập thấp
* 20000 < x <= 50000 : Có thu nhập trung bình
* 50000 < x <= 80000 : Có thu nhập cao
* 80000 < x : Có thu nhập rất cao

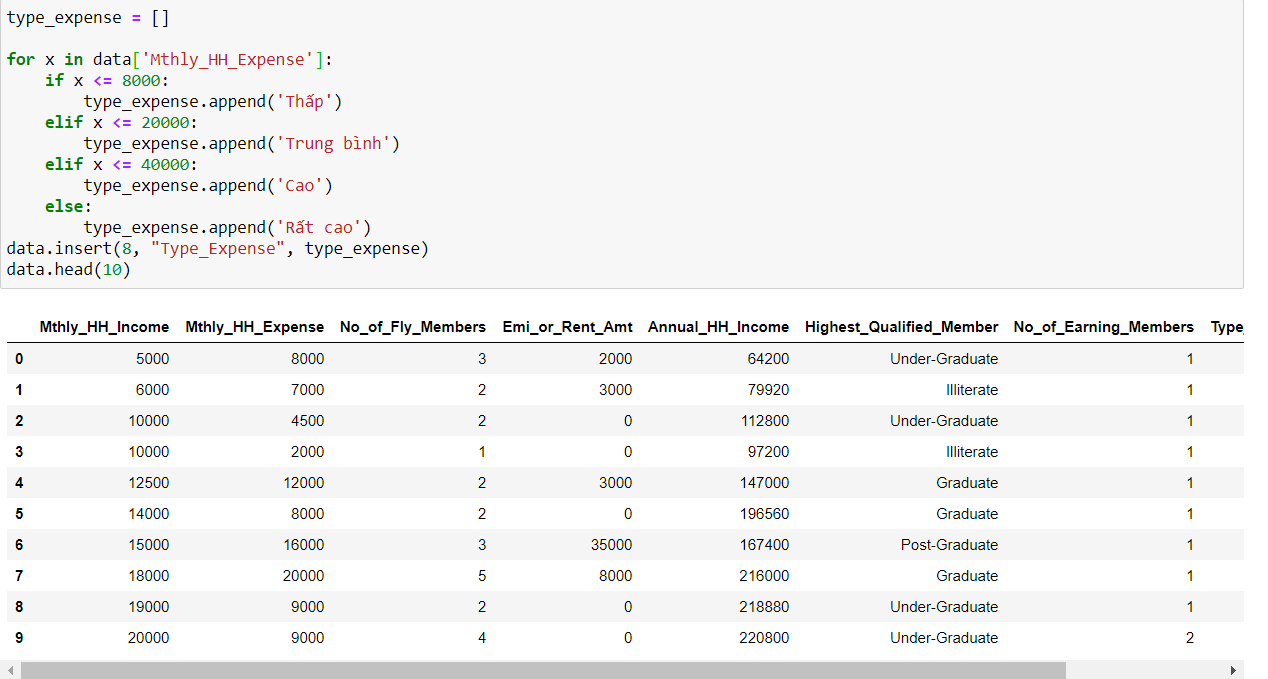
Nhìn bảng tỉ lệ và biểu đồ chúng ta sẽ thấy rõ hơn về điều này

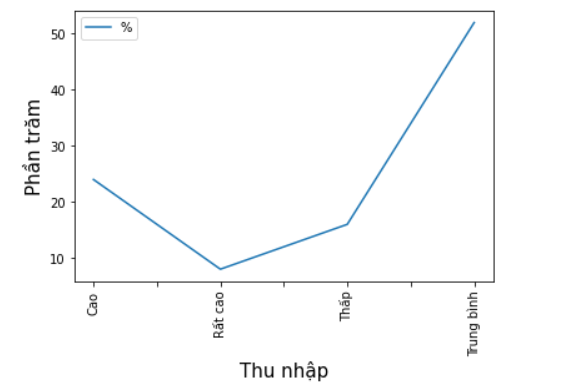
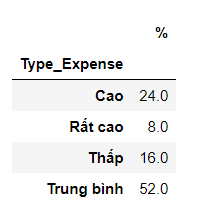




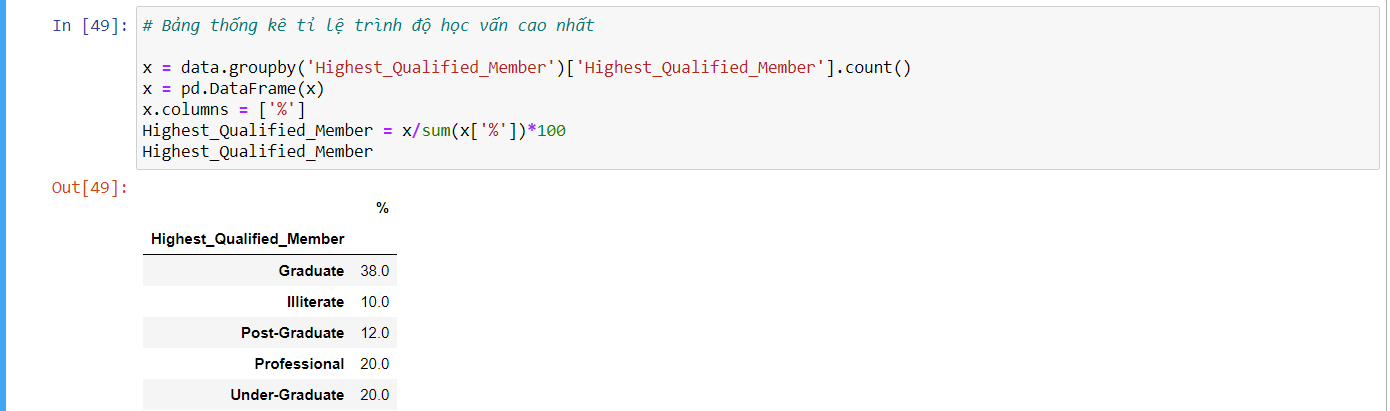
**Nhìn vào các số liệu thì ta thấy được thu nhập trung bình chiếm tỉ lệ lớn nhất**

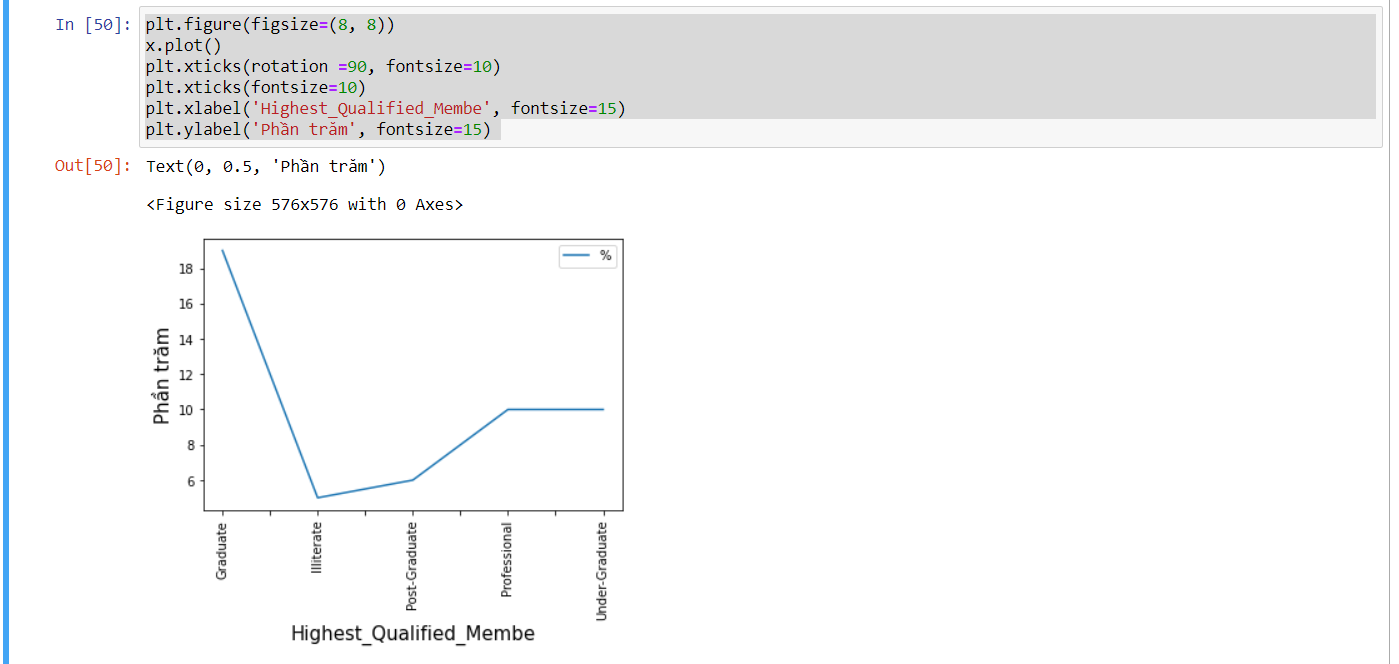
Đây là cột chi tiêu hàng tháng chúng ta làm giống như trên sẽ có được dữ liệu như sau:





**Nhìn vào các số liệu thì ta thấy được chi tiêu của người có mức thu nhập trung bình cao nhất**

****

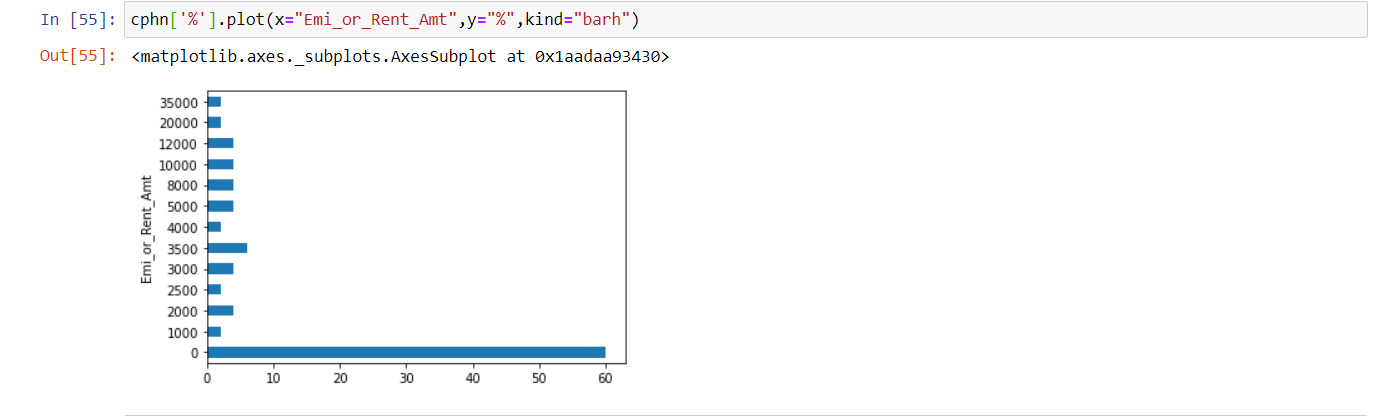
****

**Nhìn vào các số liệu thì ta thấy được người đã tốt nghiệp chiếm tỉ lệ lớn nhất**

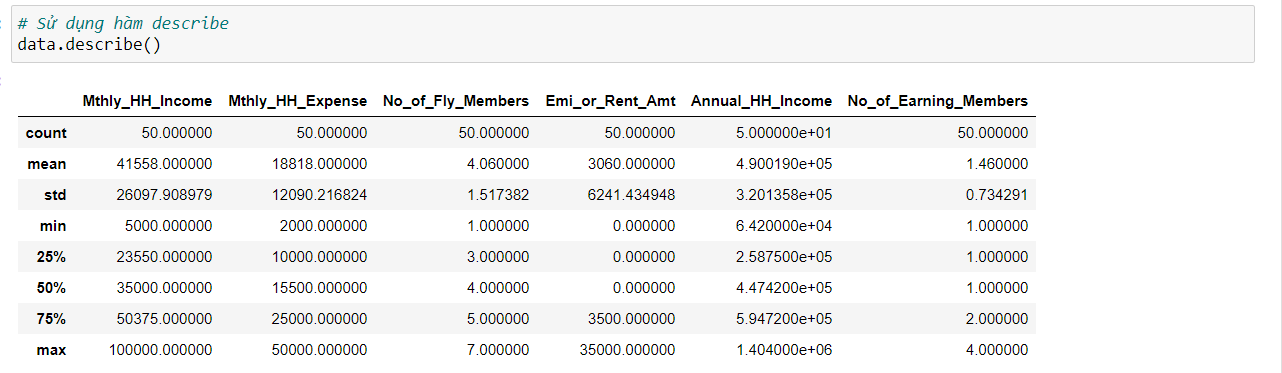
****

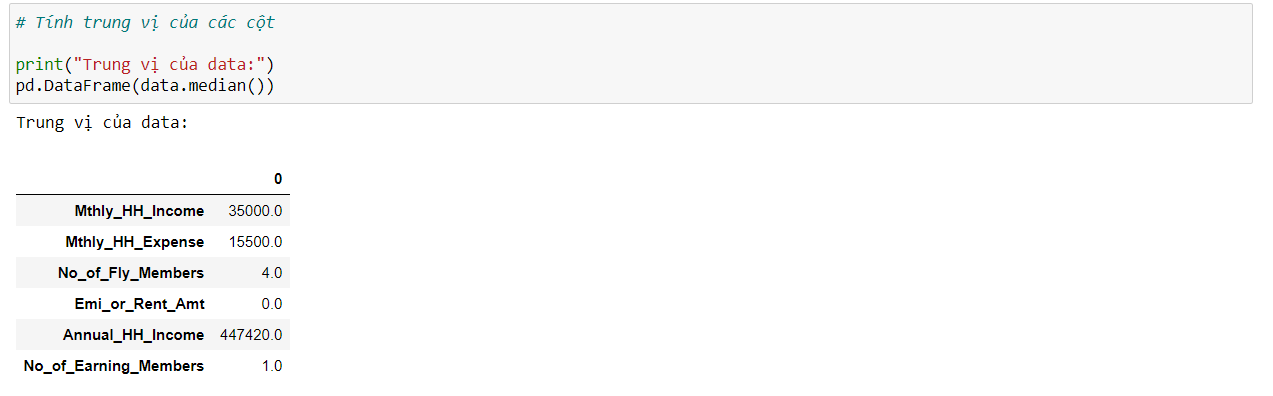
**Nhìn vào các số liệu thì ta thấy được gia đình 4 thành viên chiếm tỉ lệ lớn nhất**

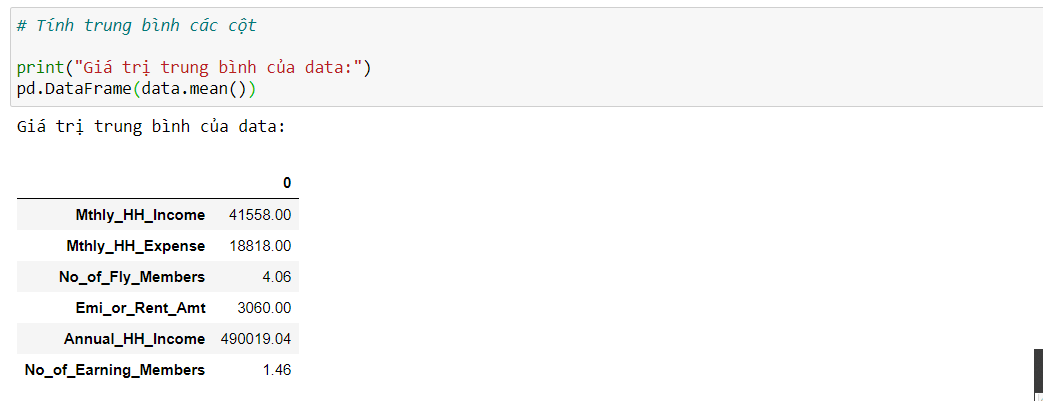
****

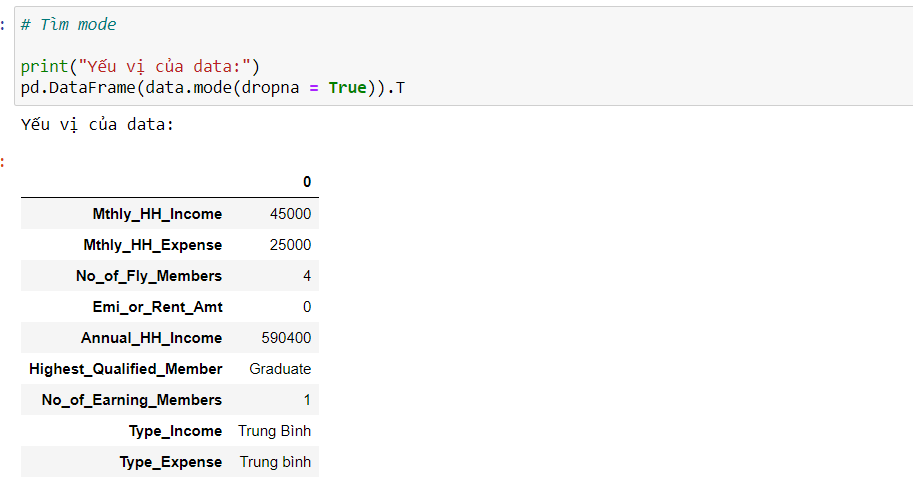
****

**Xu hướng tập trung của các dữ liệu:**

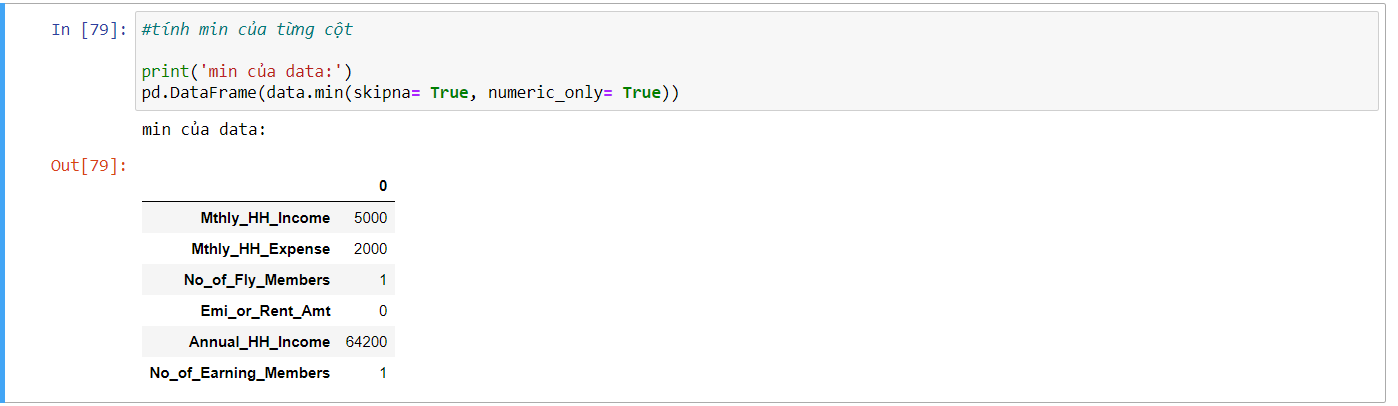
****

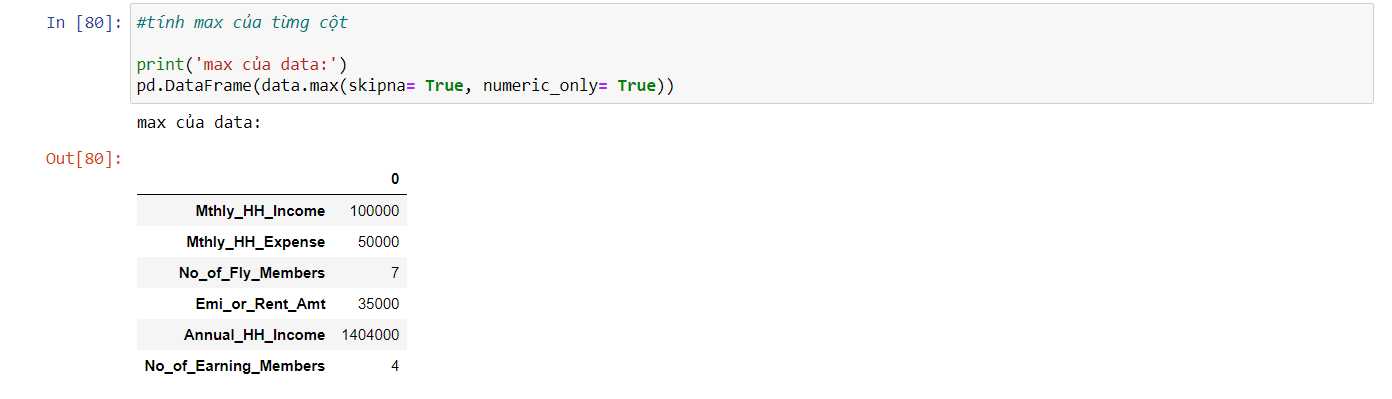
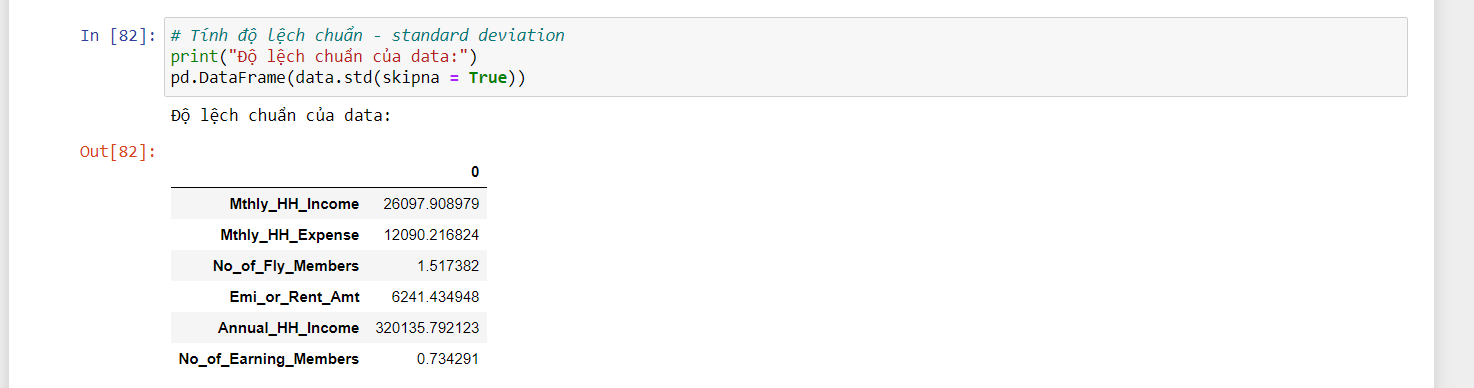
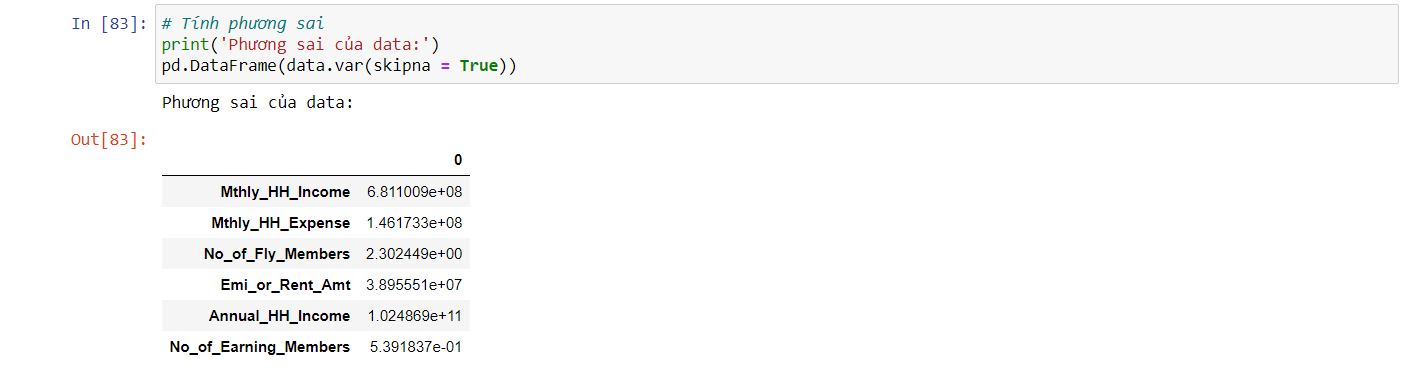
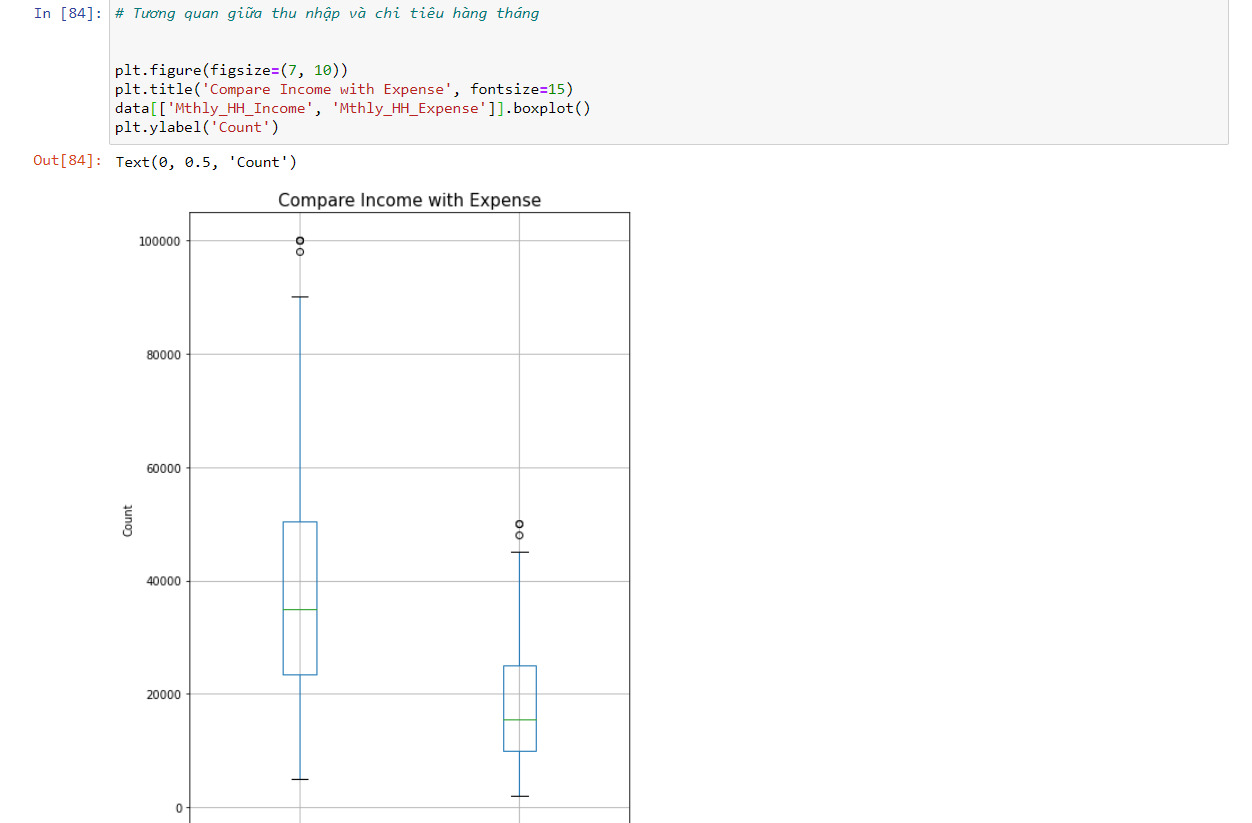
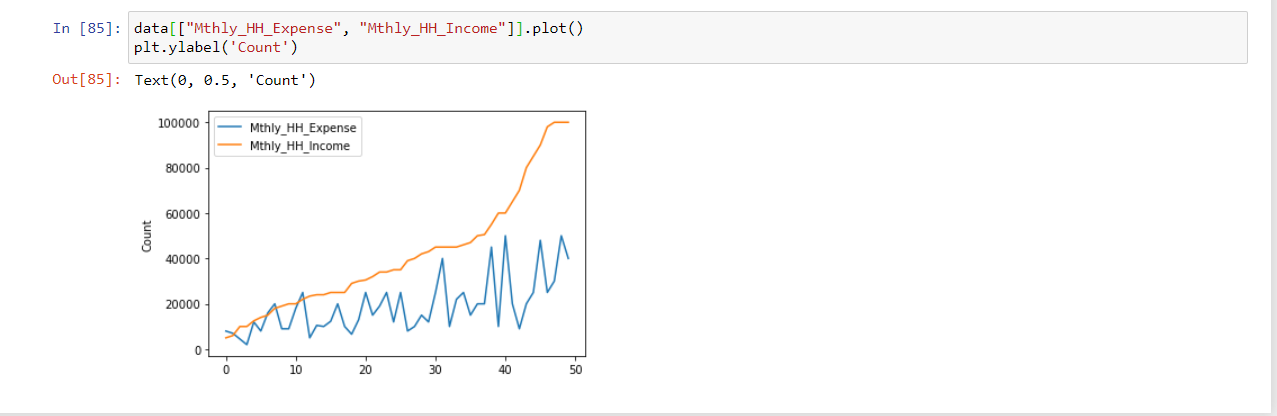
****

****

****

**Độ phân tán của dữ liệu:**

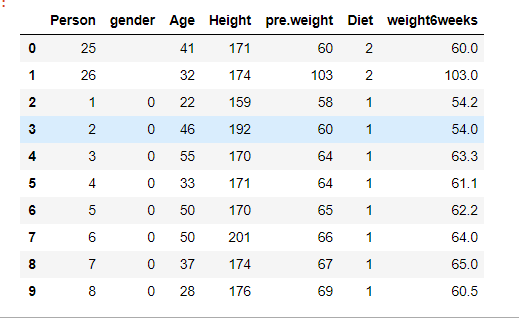
****

**** ****   

## 2.2 Thống kê suy diễn

Dữ liệu từ file Diet\_R.csv

- Đây là dữ liệu gồm các chỉ số cân nặng của nhiều người trước và sau khi đã thực hiện các phương pháp giảm cân.

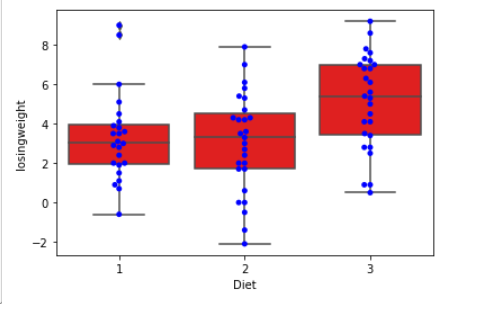


* Lúc này chúng ta đi giải quyết 2 bài toán như sau:

**Bài toán 1: Có sự khác nhau về hiệu quả của các phương pháp giảm cân ?**

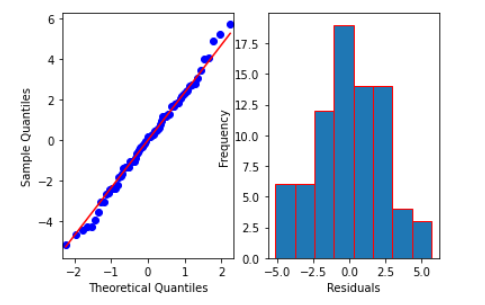
Phát biểu giả thuyết thống kê như sau:

* *H0*: *không có sự khác biệt về hiệu quả giảm cân giữa 3 phương pháp*
* *H1*: *có ít nhất hai kiểu phương pháp khác nhau về hiệu quả giảm cân*



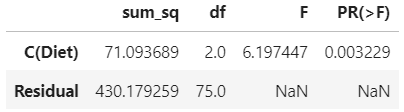
Kết luận : Nhìn vào biểu đồ boxplot thì ta có thể nhận thấy có sự khác biệt tương đối giữa các phương pháp giảm cân.

***Tiến hành kiểm tra các điều kiện cho kiểm định Anova***

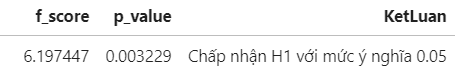
******

Kết luận: 2 biểu đồ qqplot và histogram có thể cho ta thấy dữ liệu trên là dữ liệu phân phối chuẩn.

***Tiến hành kiểm định Anova***

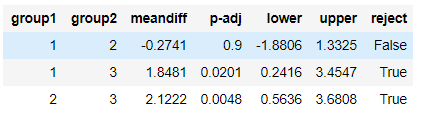


Kết luận : Vì p\_value bé hơn 0.05 nên ta sẽ bác bỏ giả thuyết *H0*



Kết luận : Có thể có ít nhất 2 phương pháp khác nhau về hiệu quả giảm cân nên ta sẽ chấp nhận giả thuyết *H1*

*Chúng ta sẽ tiến hành kiểm tra hậu nghiệm để tìm ra cặp giảm cân nào có sự khác nhau về hiệu quả.*

**

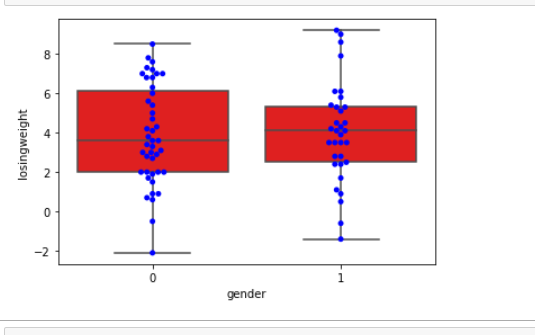
Kết luận :

* Phương pháp 1 và phương pháp 2 không có sự khác nhau về hiệu quả ở phương pháp giảm cân.
* Phương pháp 1 và phương pháp 3 có sự khác nhau về hiệu quả ở phương pháp giảm cân.
* Phương pháp 2 và phương pháp 3 có sự khác nhau về hiệu quả ở phương pháp giảm cân.

**Bài toán 2: Giới tính có thật sự ảnh hưởng đến hiệu quả giảm cân?**

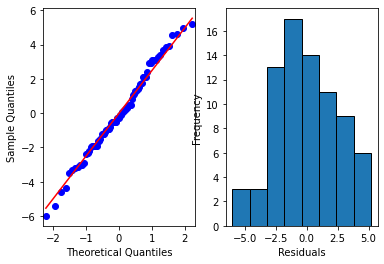
Phát biểu giả thuyết thống kê như sau:

* *H0*: *không có sự khác biệt về hiệu quả giảm cân đối với giới tính khác nhau*
* *H1*: *Giới tính có thể hiện sự khác nhau về giảm cân*



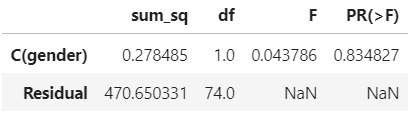
Kết luận: Qua hình vẽ thì không thấy sự khác nhau nhiều giữa hiệu quả giảm cân của 2 giới tính.

***Tiến hành kiểm tra các điều kiện cho kiểm định Anova***

****

Kết luận: 2 biểu đồ qqplot và histogram có thể cho ta thấy dữ liệu trên là dữ liệu phân phối chuẩn.

***Tiến hành kiểm định Anova***



Kết luận : Vì p\_value lớn hơn 0.05 nên ta sẽ chấp nhận giả thuyết *H0*

Vì dữ liệu không đủ nên ta không cần tiến hành kiểm tra hậu nghiệm.

# TÀI LIỆU THAM KHẢO

(Wkipedia, 2021) (giaiphapseo, n.d.) (thuvienpython, n.d.) (wikipedia, Beautiful\_Soup, 2021) (wikipedia, HMTL, 2021)

# KIỂM TRA ĐẠO VĂN

