**ĐẠI HỌC HUẾ**

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ



🙠🙟🕮🙝🙢

****

**BÁO CÁO**

**PHÂN TÍCH DỮ LIỆU VỚI PYTHON**

**NĂM HỌC 2020-2021**

**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: KHDL & TTNT K2**

|  |
| --- |
| Số phách  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày …tháng…năm.....**

**ĐẠI HỌC HUẾ**

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ



🙠🙟🕮🙝🙢

****

(MẪU BÌA PHỤ)

**BÁO CÁO**

**ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN**

**NĂM HỌC 2020-2021**

**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: KHDL & TTNT K2**

**Sinh viên thực hiện: Đặng Thái Nhật Minh**

*(ký tên và ghi rõ họ tên)*

|  |
| --- |
| Số phách  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày …tháng…năm.....**

Mục lục

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 1](#_Toc106013655)

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 2](#_Toc106013656)

[1. Phân tích thống kê mô tả trên tập dữ liệu Inc\_Exp\_Data.csv 4](#_Toc106013657)

[1.1. Xác định thư viện cần dung cho quá trình thống kê 5](#_Toc106013658)

[1.2. Phân tích dữ liệu định lượng 5](#_Toc106013659)

[1.3. Trực quan hóa dữ liệu định lượng 5](#_Toc106013660)

[1.4. Dữ liệu định tính 5](#_Toc106013661)

[1.5. Trực quan hóa dữ liệu định tính 5](#_Toc106013662)

[2. Quy trình kiểm định ANOVA 5](#_Toc106013663)

[3. Phân tích thống kê suy diễn trên tập dữ liệu Diet\_R.csv 5](#_Toc106013664)

[3.1. Tập dữ liệu Diet\_R.csv chứa thông tin về 78 người sử dụng một trong ba chế độ ăn kiêng. 5](#_Toc106013665)

[3.2. Chuẩn bị 5](#_Toc106013666)

[3.3. Paired t-test 5](#_Toc106013667)

[3.4. Summarry statistics 5](#_Toc106013668)

[3.5. One way ANOVA 5](#_Toc106013669)

[3.5.1. Research question: Which diet was best for losing weight? (Chế độ ăn kiêng nào là tốt nhất để giảm cân?) 5](#_Toc106013670)

[3.5.2. Research question: Are there gender differences for weight lost? (Khả năng giảm cân có phụ thuộc vào giới tính hay không?) 5](#_Toc106013671)

[4. Kiểm tra đạo văn 5](#_Toc106013672)

# Phân tích thống kê mô tả trên tập dữ liệu Inc\_Exp\_Data.csv

## Xác định thư viện cần dung cho quá trình thống kê

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

#Đọc dữ liệu

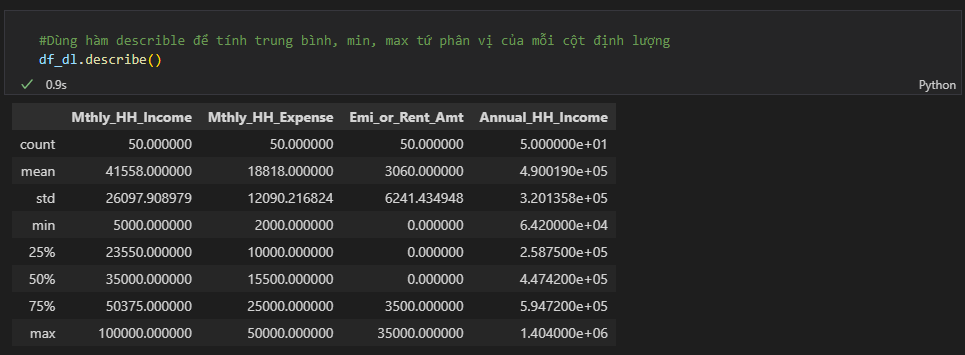
df = pd.read\_csv("Inc\_Exp\_Data.csv")

## Phân tích dữ liệu định lượng

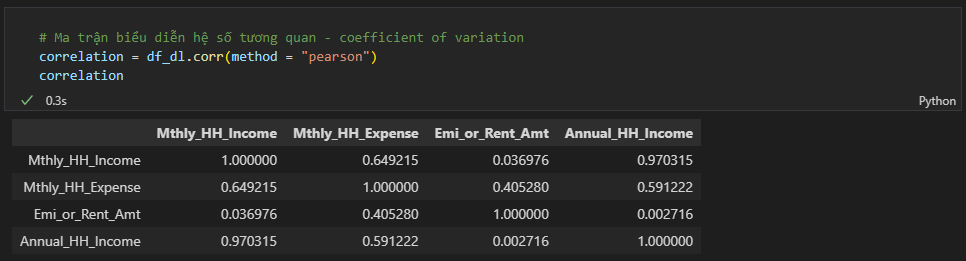
* Các cột dữ liệu định lượng:
  + Mthly\_HH\_Income
  + Mthly\_HH\_Expense
  + Emi\_or\_Rent\_Amt
  + Annual\_HH\_Income
* Xác định và tạo Dataframes gồm các cột định lượng:

df\_dl = pd.DataFrame(columns = ['Mthly\_HH\_Income', 'Mthly\_HH\_Expense', 'Emi\_or\_Rent\_Amt', 'Annual\_HH\_Income', ], data = df)

* Dùng hàm describle để tính trung bình, min, max tứ phân vị của mỗi cột định lượng

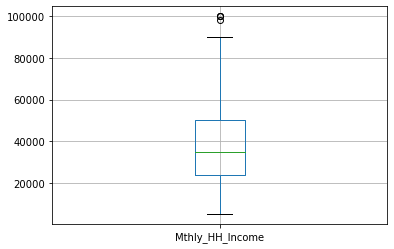


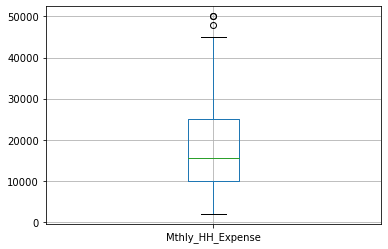
* Xác định hệ số tương quan giữa các cột định lượng

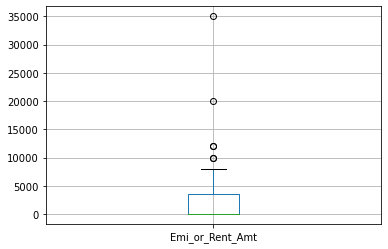


## Trực quan hóa dữ liệu định lượng

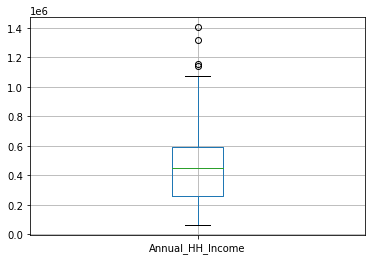
* Vẽ biểu đồ boxplot
  + #Vẽ biểu đồ boxplot cho Mthly\_HH\_Income.

df\_dl.boxplot(['Mthly\_HH\_Income'])

* + #Vẽ biểu đồ boxplot cho Mthly\_HH\_Expense.df\_dl.boxplot(['Mthly\_HH\_Expense'])
  + #Vẽ biểu đồ boxplot cho Emi\_or\_Rent\_Amt.

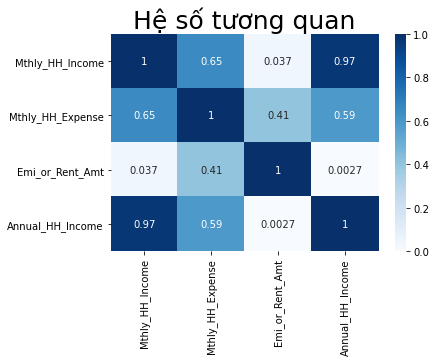
df\_dl.boxplot(['Emi\_or\_Rent\_Amt'])

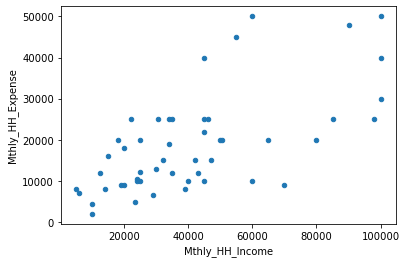
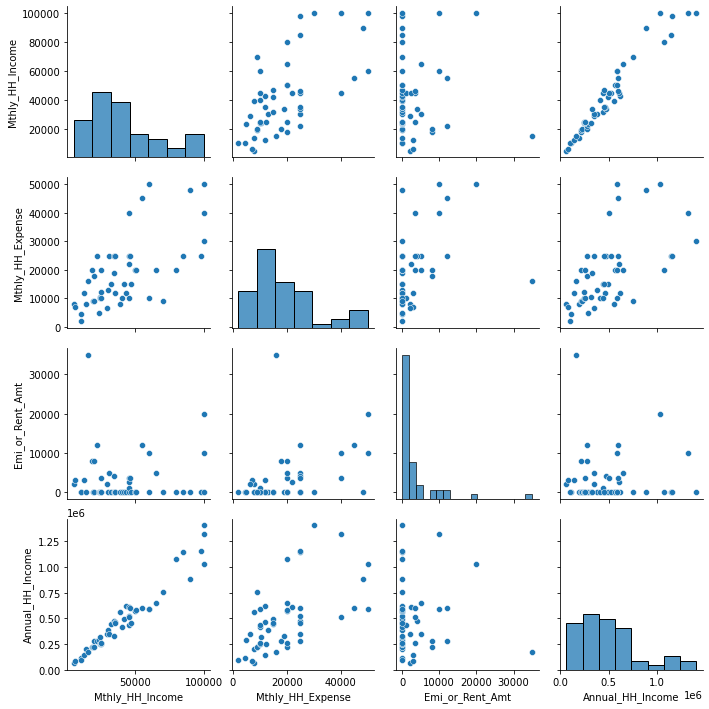
* + #Vẽ biểu đồ boxplot cho Annual\_HH\_Income.

df\_dl.boxplot(['Annual\_HH\_Income'])

* Biểu đồ nhiệt (Heat map)
  + plt.title("Hệ số tương quan", fontsize = 25)

sns.heatmap(correlation, annot = True, cmap = "Blues",vmin = 0, vmax = 1)



* Biểu đồ scatterplot
  + df\_dl.plot.scatter(x = "Mthly\_HH\_Income", y = "Mthly\_HH\_Expense")
* Biểu đồ cặp
  + sns.pairplot(df\_dl)

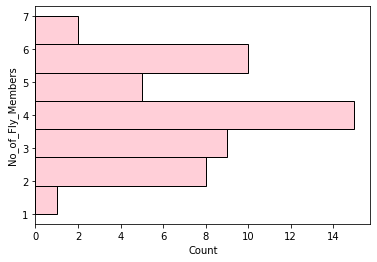
## Dữ liệu định tính

* Các cột dữ liệu định tính
  + No\_of\_Fly\_Members
  + No\_of\_Earning\_Members
* Tạo dataframes gồm các cột dữ liệu định tính
  + df\_dt = pd.DataFrame(columns = ["No\_of\_Earning\_Members", '', 'No\_of\_Fly\_Members', 'Highest\_Qualified\_Member', ], data = df)

## Trực quan hóa dữ liệu định tính

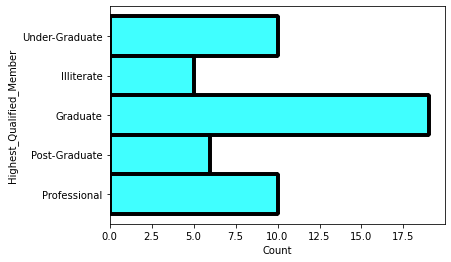
* No\_of\_Fly\_Members'  
  NFM = df\_dt['No\_of\_Fly\_Members']

sns.histplot(data = NFM, color = 'pink', y = NFM)



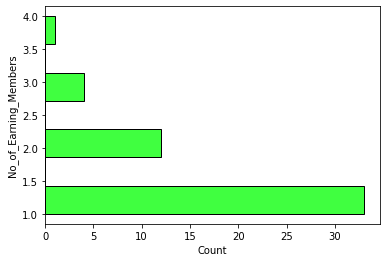
* Highest\_Qualified\_Member  
  HQM = df\_dt['Highest\_Qualified\_Member']

sns.histplot(data = HQM, color = 'cyan', y = HQM)



* No\_of\_Earning\_Members  
  NEM = df\_dt['No\_of\_Earning\_Members']

sns.histplot(data = NEM, color = 'lime', y = NEM)



# Quy trình kiểm định ANOVA

* B1: Đặt ra câu hỏi để nghiên cứu
* B2: Kiểm tra mẫu có tuân theo phân phối chuẩn hay không (Sử dụng kiểm điểm shapiro).
* B3: Kiểm tra tính thuần nhất phương sai (Sử dụng kiểm định bartlett).
* B4: Thực hiện kiểm định ANOVA

# Phân tích thống kê suy diễn trên tập dữ liệu Diet\_R.csv

## Tập dữ liệu Diet\_R.csv chứa thông tin về 78 người sử dụng một trong ba chế độ ăn kiêng.

|  |  |  |
| --- | --- | --- |
| Tên biến | Biến | Kiểu dữ liệu |
| Person | Số người tham gia |  |
| Gender | 1 = nam, 0 = nữ | Nhị phân (binary) |
| Age | Tuổi (năm) | Tỷ lệ (scale) |
| Height | Chiều cao(cm) | Tỷ lệ (scale) |
| Preweight | Cân nặng trước khi ăn kiêng (kg) | Tỷ lệ (scale) |
| Diet | Chế độ ăn kiêng | Nhị phân (binary) |
| Weight10weeks | Cân nặng sau 10 tuần (kg) | Tỷ lệ (scale) |
| weightLOST | Cân nặng đã giảm sau 10 tuần (kg) | Tỷ lệ (scale) |

## Chuẩn bị

* Xác định thư viện thống kê:
  + import numpy as np
  + import pandas as pd
  + import matplotlib.pyplot as plt
  + import seaborn as sns
  + from scipy import stats as st
  + import statsmodels.api as sm
* #Đọc dữ liệu

df = pd.read\_csv("Inc\_Exp\_Data.csv")

* Ta nhận thấy người thứ 25 và 26 bị thiếu mất dữ liệu. Ta tiến hành xóa đi người 25, 26 và sắp xếp lại thứ tự.
  + # Xóa những người bị thiếu dữ liệu (25, 26)

df.drop(df.loc[df.Person == 25].index, inplace = True)

df.drop(df.loc[df.Person == 26].index, inplace = True)

* + # Sắp xếp lại số thứ tự

serial = []

print(len(df))

for i in range (1, len(df) + 1):

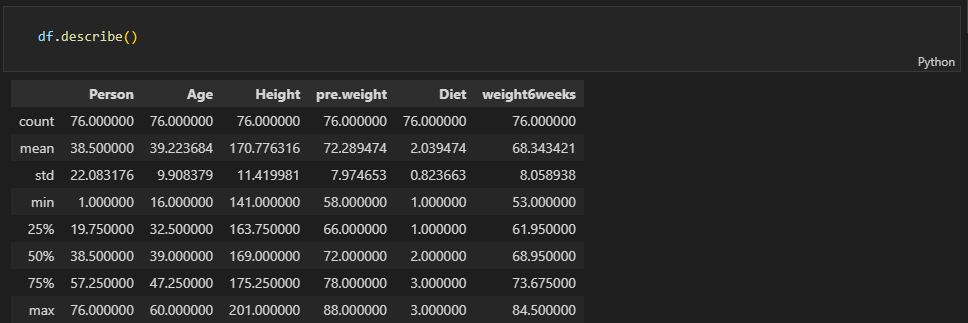
serial.append(i)

df["Person"] = serial

df\_new = df.set\_index("Person").sort\_index()

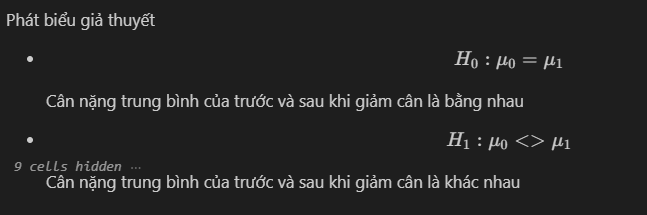
df\_new

* + Dùng hàm describe() để xác định tứ phân vị, min, max, trung bình của dữ liệu



## Paired t-test

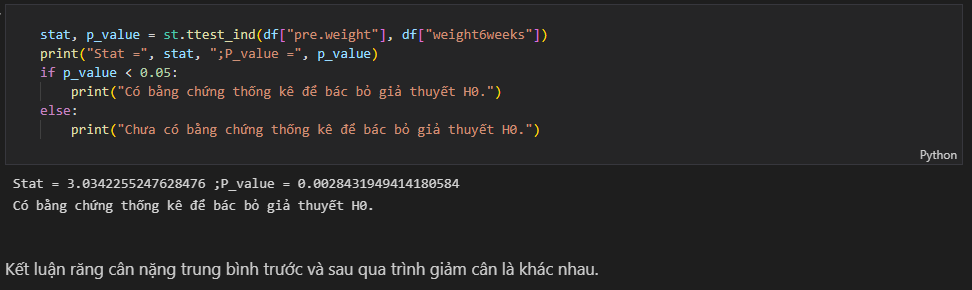
* Research question: Ignore diet and test to see if weight has been lost (Bỏ qua chế độ ăn kiêng và kiểm tra xem đã giảm cân chưa)
* Phát biểu giả thuyết

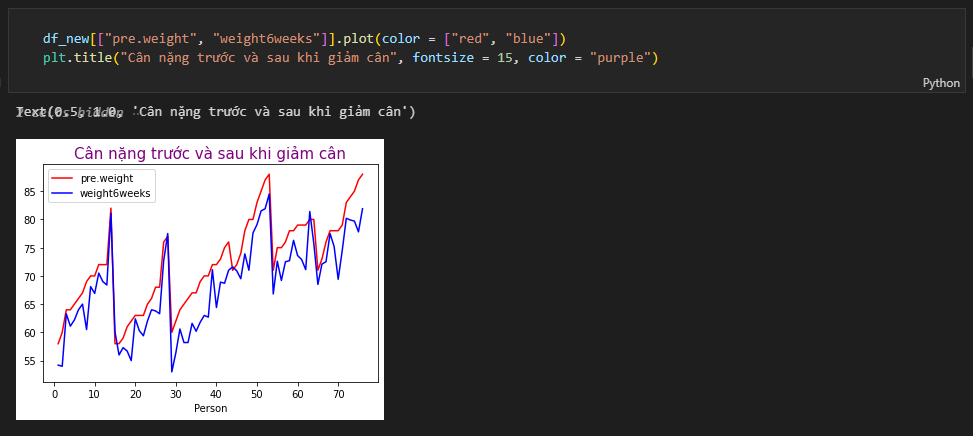


* Phân tích và kết luận:

Sự thay đổi cân nặng trước và sau khi áp dụng 3 phương pháp giảm cân

Kiểm định: Ta sử dụng đoạn Kiểm định t\_test

******

* Ta sử dụng seaborn để tạo một biểu đồ trực quan thể hiện về sự thay đổi, khác nhau về cân nặng trước và sau khi áp dụng 3 phương pháp giảm cân

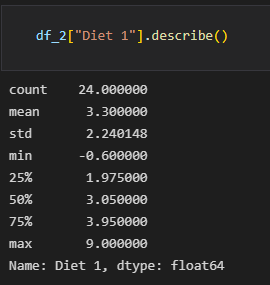
## Summarry statistics

* Research question: Summarry statistics by diet (Thống kê tóm tắt theo chế độ ăn kiêng).
* Tạo dataframes

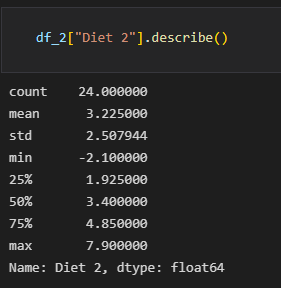


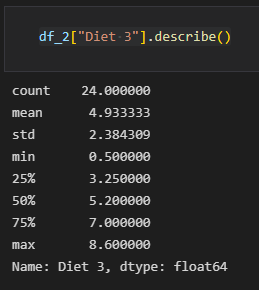


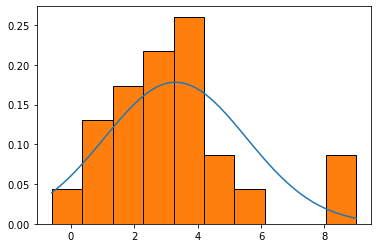
* Tính tứ phân vị, trung bình, min, max của từng phương pháp giảm cân:
  + Diet 1



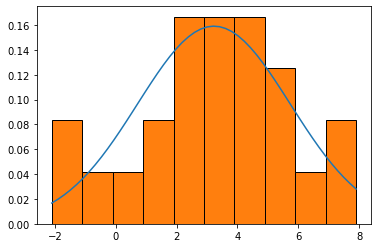
* + Diet 2

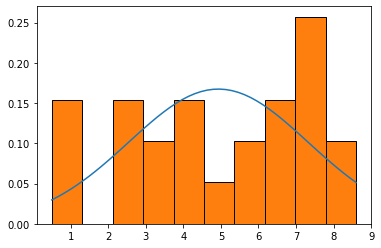


* + Diet 3
  + 
* Vẽ biểu đồ hist
  + Diet 1



* + Diet 2

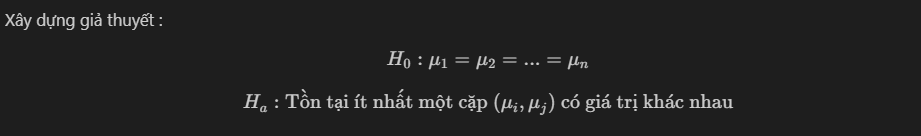


* + Diet 3  
    

## One way ANOVA

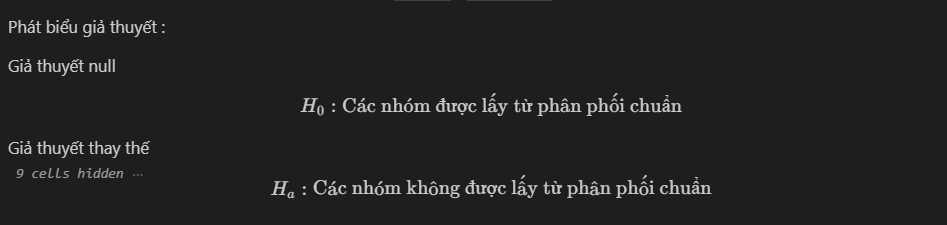
### Research question: Which diet was best for losing weight? (Chế độ ăn kiêng nào là tốt nhất để giảm cân?)

* Xây dựng giả thuyết:

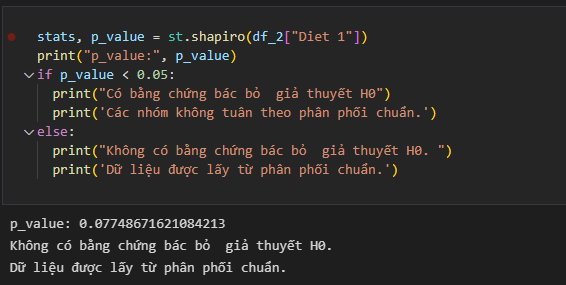


* Bắt đầu tiến hành kiểm định các nhóm có lấy từ phân phối chuẩn

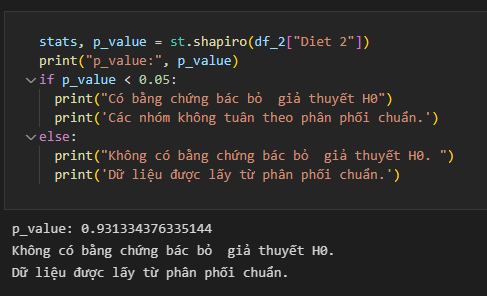
Ta dùng kiểm định shapiro.



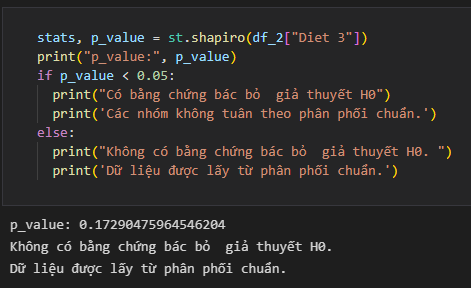
* + Diet 1



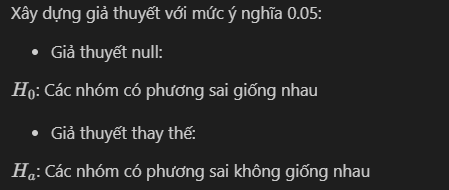
* + Diet 2



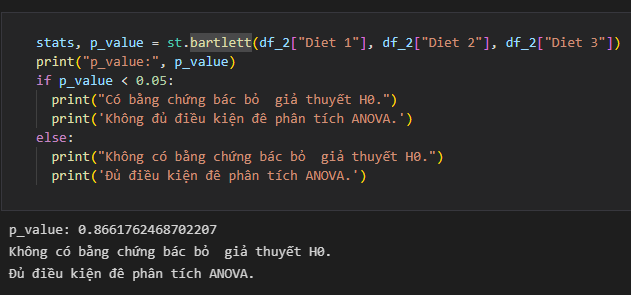
* + Diet 3



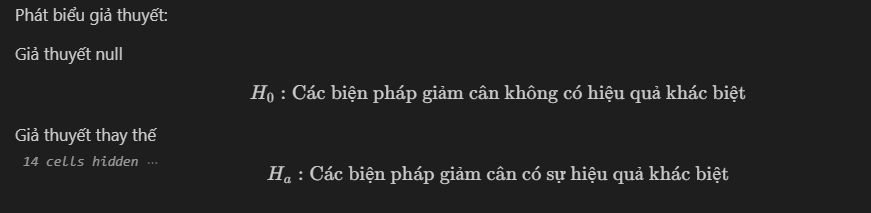
* Tiến hành kiểm định phương sai của các nhóm có sự khác biệt.

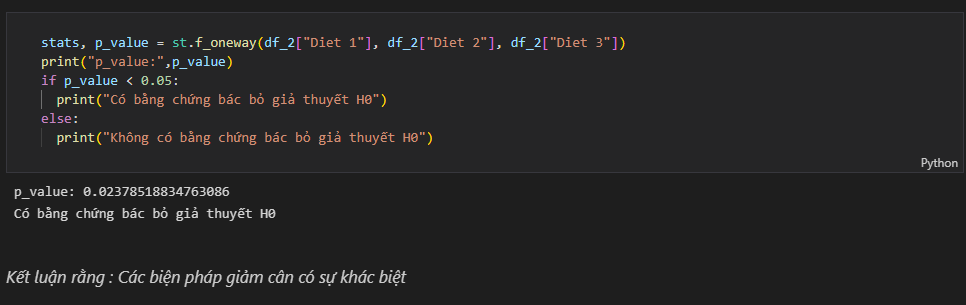


* Ta dùng kiểm định bartlett

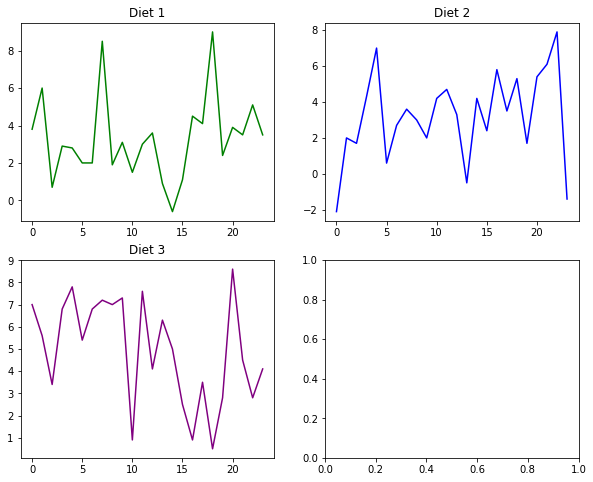


* Ta tiến hành phân tích ANOVA 1 chiều với dữ liệu vì các nhóm đều độc lập, được lấy từ phân phối chuẩn và giống về phương sai:

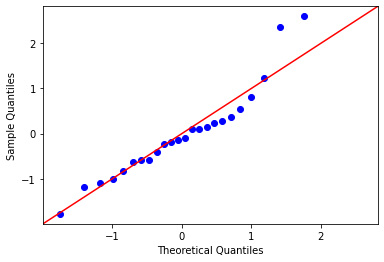




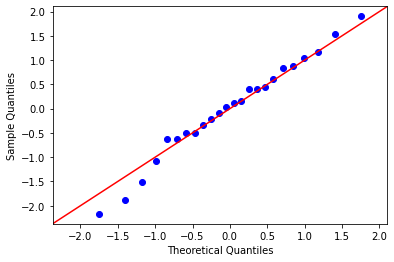
* + - Kết luận rằng các biện pháp giảm cân có sự khác biệt.
* Ta thể hiện bằng biểu đồ:



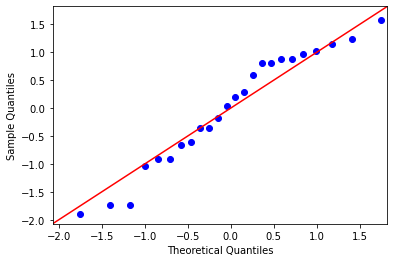
* Vẽ biểu đồ qqplot theo thứ tự từ Diet1, Diet2, Diet 3:
  + Diet 1



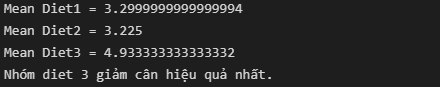
* + Diet 2



* + Diet 3

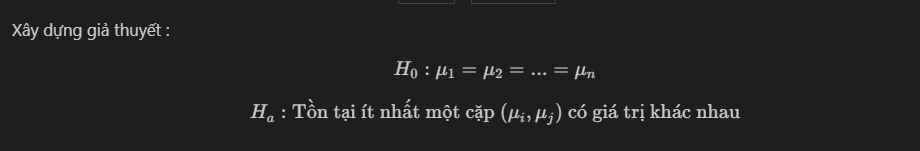


* So sánh mean của 3 phương pháp, chọn ra phương pháp giảm cân hiệu quả nhất. Ta được kết quả:

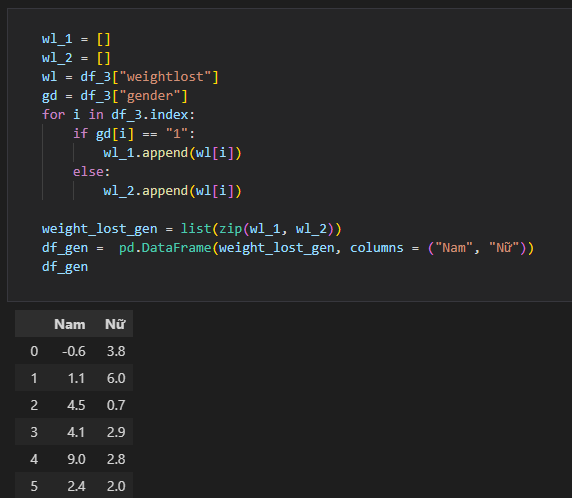


### Research question: Are there gender differences for weight lost? (Khả năng giảm cân có phụ thuộc vào giới tính hay không?)

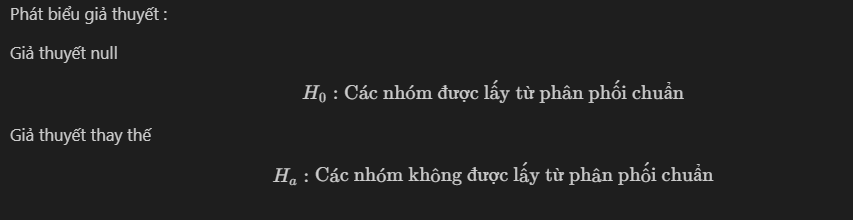
* Xây dựng giả thuyết:



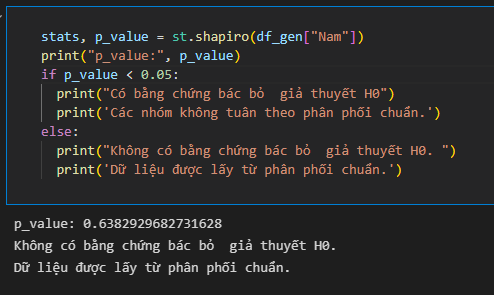
* Tạo dataframes phù hợp với yêu cầu bài:



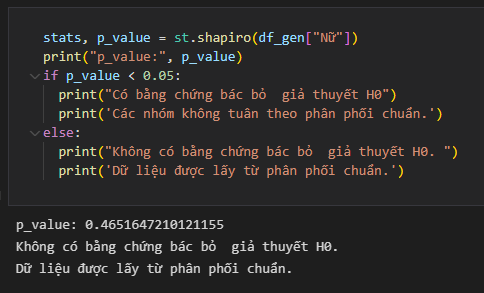
* Bắt đầu tiến hành kiểm định các nhóm có lấy từ phân phối chuẩn



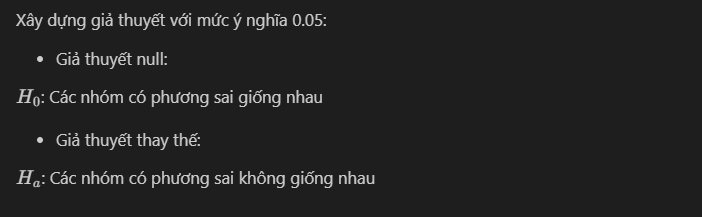
* Ta dùng kiểm định shapiro:
  + Nam



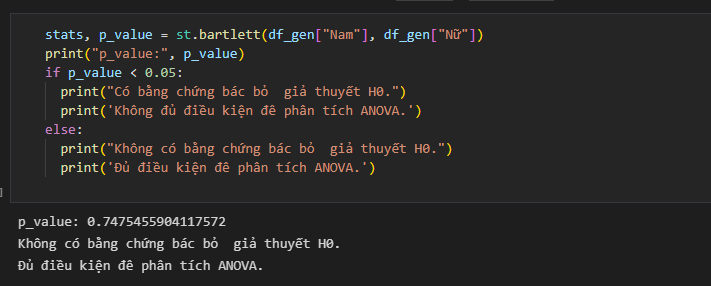
* + Nữ



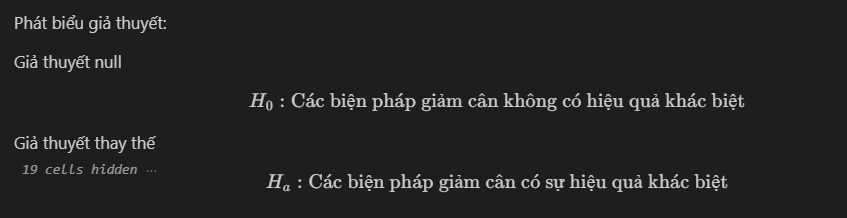
* Tiến hành kiểm định phương sai của các nhóm có sự khác biệt

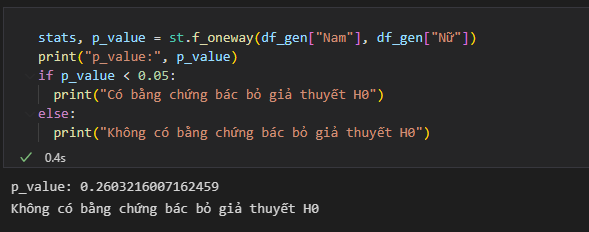


* Ta sử dụng kiểm định bartlett

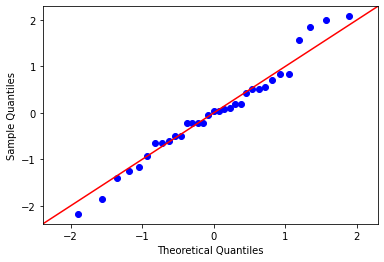


* Ta tiến hành phân tích ANOVA 1 chiều với dữ liệu vì các nhóm đều độc lập, được lấy từ phân phối chuẩn và giống về phương sai.

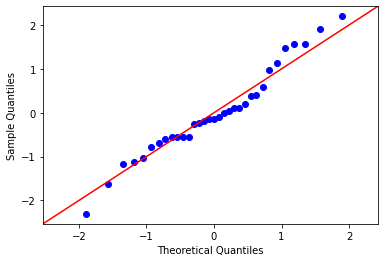




* + - Kết luận rằng : Sự giảm cân không phụ thuộc vào giới tính
* Vẽ biểu đồ qqplot cho Nam và Nữ:
  + Nam



* + Nữ



# Kiểm tra đạo văn

Graphical user interface, text, application

Description automatically generated