**ĐẠI HỌC HUẾ**

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ



🙠🙟🕮🙝🙢

**Icon

Description automatically generated**

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

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

**Lớp: K1- Khoa học dữ liệu và Trí tuệ nhân tạo**

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

**Thừa Thiên Huế, ngày 26 tháng 6 năm 2021**

**ĐẠI HỌC HUẾ**

# KHOA KỸ THUẬT VÀ CÔNG NGHỆ



🙠🙟🕮🙝🙢

**Icon

Description automatically generated**

(MẪU BÌA PHỤ)

**BÁO CÁO**

**ĐỒ ÁN**

**NĂM HỌC 2020-2021**

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

**Lớp: K1- Khoa học dữ liệu và Trí tuệ nhân tạo**

**Sinh viên thực hiện: Nguyễn Thanh Lâm**

*(ký tên và ghi rõ họ tên)*

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

**Thừa Thiên Huế, ngày 26 tháng 6 năm 2021**

MỤC LỤC

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 1](#_Toc75816413)

[KHOA KỸ THUẬT VÀ CÔNG NGHỆ 2](#_Toc75816414)

[**I.** **KHÁI QUÁT CHUNG VỀ THỐNG KÊ MÔ TẢ** 5](#_Toc75816415)

[**II.** **PHÂN TÍCH THỐNG KÊ MÔ TẢ CHO DỮ LIỆU Inc\_Exp\_Data.csv** 5](#_Toc75816416)

[**III.** **KHÁI QUÁT CHUNG VỀ THỐNG KÊ SUY DIỄN** 11](#_Toc75816417)

[**IV.** **PHẦN PHÂN TÍCH THỐNG KÊ SUY DIỄN, PHÂN TÍCH ANOVA MỘT CHIỀU CỦA TẬP TIN DIET\_R.CSV** 11](#_Toc75816418)

[**1.** **Khái quát về file Diet\_R.csv:** 11](#_Toc75816419)

[**2.** **GIẢI QUYẾT BÀI TOÁN** 12](#_Toc75816420)

[**Tiến hành tìm bảng Anova:** 13](#_Toc75816421)

[**Kiểm tra hậu nghiệm:** 13](#_Toc75816422)

[**Kiểm tra các điều kiện cho kiểm định Anova** 13](#_Toc75816423)

[Tài liệu trích dẫn: 15](#_Toc75816424)

[Kết quả kiểm tra đạo văn: 15](#_Toc75816425)

1. **KHÁI QUÁT CHUNG VỀ 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.[[1]](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_m%C3%B4_t%E1%BA%A3#cite_note-1) Có rất nhiều kỹ thuật hay được sử dụng. Có thể phân loại các kỹ thuật này như sau:

* Biểu diễn dữ liệu bằng đồ họa trong đó các đồ thị mô tả dữ liệu hoặc giúp so sánh dữ liệu;
* Biểu diễn dữ liệu thành các bảng số liệu tóm tắt về dữ liệu;
* Thống kê tóm tắt (dưới dạng các giá trị thống kê đơn nhất) mô tả dữ liệu.

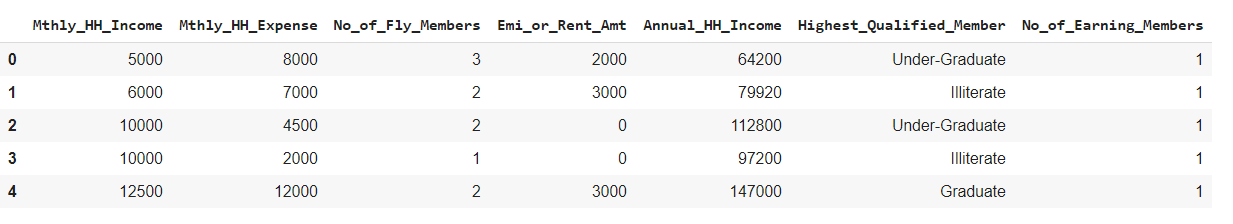
Khi tạo các trị thống kê mô tả, người ta có thể nhằm 2 mục tiêu:

1. 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).
2. 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). (1)
3. **PHÂN TÍCH THỐNG KÊ MÔ TẢ CHO DỮ LIỆU Inc\_Exp\_Data.csv**

Tập dữ liệu Inc\_Exp\_Data có 7 loại dữ liệu:

+ 4 loại dữ liệu thuộc kiểu định lượng: *Mthly\_HH\_Income, Mthly\_HH\_Expense, Emi\_or\_Rent\_Amt, Annual\_HH\_Income.*

+ 3 loại dữ liệu thuộc kiểu định tính: *Highest\_Qualified\_Member, No\_of\_Fly\_Members, No\_of\_Earning\_Members.*

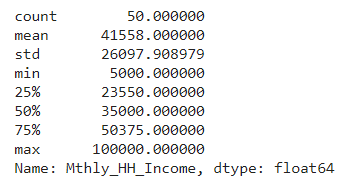


**1. Kiểu định lượng:**

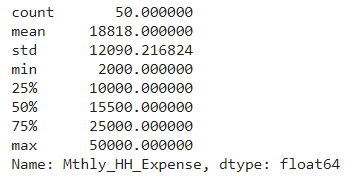
Sử dụng hàm describe() để thống kê mô tả 4 loại dữ liệu định lượng. Ta được:

|  |  |
| --- | --- |
| **count** | Số lượng mẫu |
| **mean** | Trung bình mẫu |
| **std** | Độ lệch chuẩn |
| **min** | Giá trị nhỏ nhất |
| **25%**, **50%**, **75%** | Tứ phân vị |
| **max** | Giá trị lớn nhất |

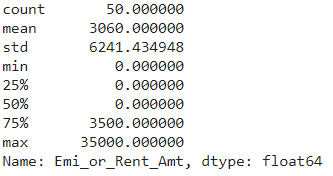
**- Mthly\_HH\_Income:**



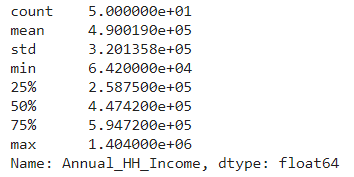
**- Mthly\_HH\_Expense:**



**- Emi\_or\_Rent\_Amt:**

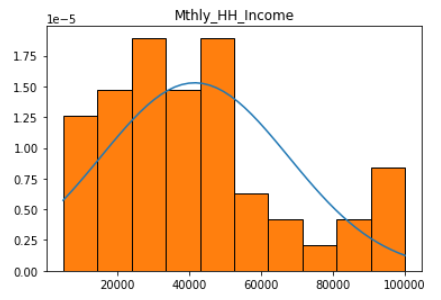


**- Annual\_HH\_Income:**

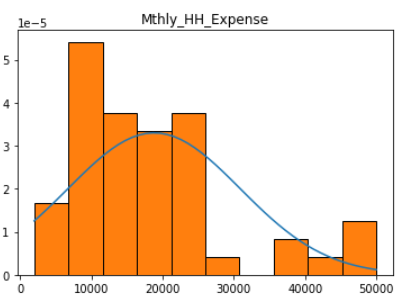


**Vẽ biểu đồ để biểu hiện rõ kiểu định lượng của 4 trường dữ liệu:**

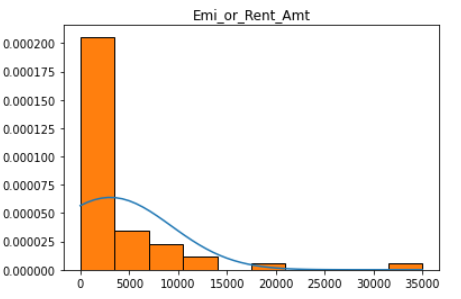
**- Mthly\_HH\_Income:**



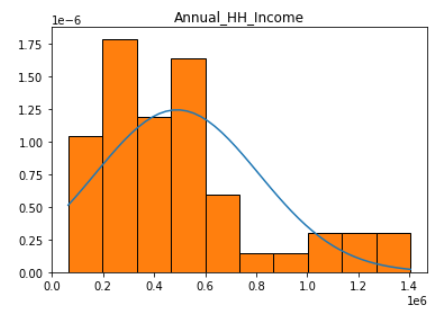
**- Mthly\_HH\_Expense**



**- Emi\_or\_Rent\_Amt**



**- Annual\_HH\_Income:**



**Nhận xét:**

**Mthly\_HH\_Income: Chủ yếu tập trung từ 20000 – 50000**

**Mthly\_HH\_Expense: Chủ yếu tập trung từ 10000 – 25000**

**Emi\_or\_Rent\_Amt: Chủ yếu tập trung từ 0 - 4000**

**Annual\_HH\_Income: Chủ yếu tập trung từ 0.2 - 0.6**

**2. Kiểu định tính:**

Sử dụng hàm describe() để thống kê mô tả 3 loại dữ liệu định tính. Ta được:

|  |  |
| --- | --- |
| count | Số lượng mẫu |
| unique | Các level khác nhau của mẫu |
| top | Level có số lượng lớn nhất |
| freq | Số lần xuất hiện của Level lớn nhất |

|  |  |
| --- | --- |
| **Highest\_Qualified\_Member**  count 50  unique 5  top Graduate  freq 19  Name: Highest\_Qualified\_Member,  dtype: object |  |
| **No\_of\_Fly\_Members**  count 50  unique 7  top 4  freq 15  Name: No\_of\_Fly\_Members,  dtype: object |  |
| **No\_of\_Earning\_Members**  count 50  unique 4  top 1  freq 33  Name: No\_of\_Earning\_Members,  dtype: object |  |

**Nhận xét:** Highest\_Qualified\_Member: Level chiếm đa số: ‘Graduate’ (19)

No\_of\_Fly\_Members: Level chiếm đa số: ‘4’ (15)

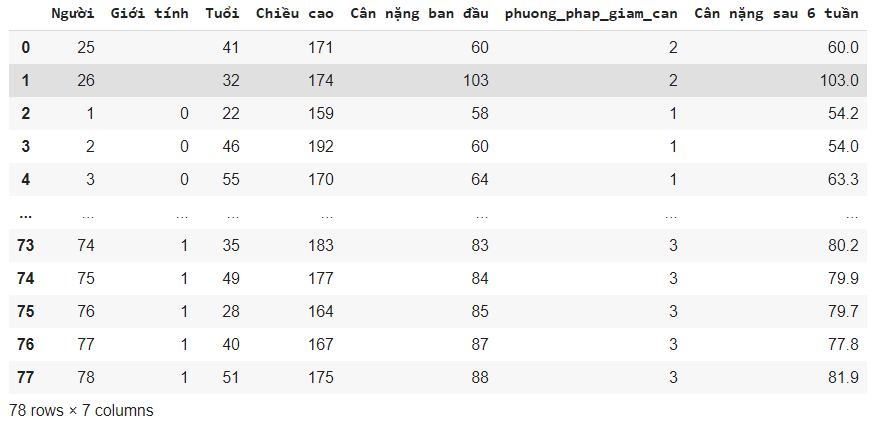
No\_of\_Earning\_Members: Level chiếm đa số: ‘1’ (33)

1. **KHÁI QUÁT CHUNG VỀ THỐNG KÊ SUY DIỄN**

**Suy luận thống kê** là quá trình suy luận suy ra các đặc điểm của một phân phối cơ bản bằng việc phân tích dữ liệu. Phân tích thống kê suy luận suy ra tính chất của tổng thể: điều này bao gồm các giả thuyết thử nghiệm và các ước tính phát sinh. Tổng thể được giả định là lớn hơn so với tạo ra các dữ liệu quan sát, nói cách khác, các dữ liệu quan sát được giả định là lấy mẫu từ một tổng thể lớn hơn.

Thống kê suy luận có thể được so sánh với số liệu [thống kê mô tả](https://vi.wikipedia.org/wiki/Th%E1%BB%91ng_k%C3%AA_m%C3%B4_t%E1%BA%A3). Thống kê mô tả chỉ quan tâm tới tính chất của dữ liệu quan sát, và không giả sử các dữ liệu đến từ dữ liệu lớn hơn. (2)

1. **PHẦN PHÂN TÍCH THỐNG KÊ SUY DIỄN, PHÂN TÍCH ANOVA MỘT CHIỀU CỦA TẬP TIN DIET\_R.CSV**
2. **Khái quát về file Diet\_R.csv:**

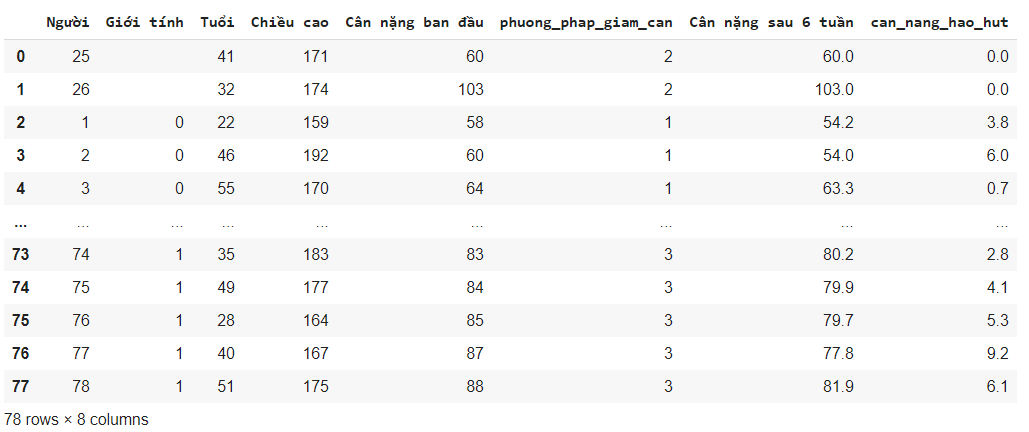


Dữ liệu trên bao gồm chỉ số cân nặng trước và sau 6 tuần khi thực hiện bằng các phương pháp giảm cân

Để giải quyết bài toán ‘Có sự khác nhau về phương pháp giảm cân’ ta tiến hành tạo thêm 1 cột **can\_nang\_hao\_hut** bằng dòng lệnh như sau :

df['can\_nang\_hao\_hut']=df['Cân nặng ban đầu'] - df['Cân nặng sau 6 tuần']

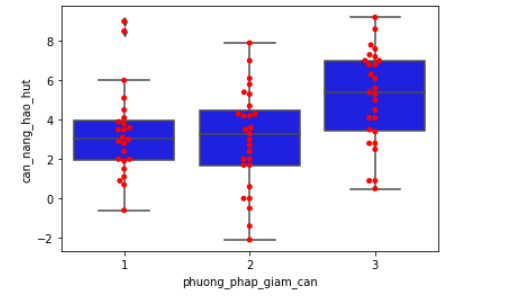
Ta được:



1. **GIẢI QUYẾT BÀI TOÁN**

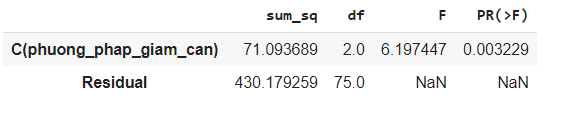
**Phát biểu giải thuyết thống kê**

* *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

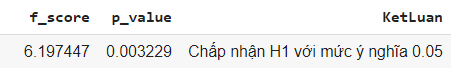


Nhìn vào đồ thị boxplot trên ta có thể nhận thấy có một chút khác biệt về hiệu quả giảm cân trung bình của phương pháp 1 và phương pháp 3, phương pháp 2 với phương pháp 3.

### **Tiến hành tìm bảng Anova:**



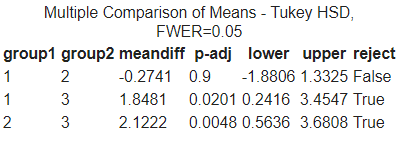
Vì p\_value bé hơn 0.05 nên ta sẽ bác bỏ giải thuyết H0.



Ta sẽ tạm chấp nhân H1: có ít nhất hai kiểu phương pháp khác nhau về hiệu quả giảm cân.

Vì chưa thể biết được sự khác biệt là ở cặp giảm cân nào nên chúng ta sẽ tiến hành kiểm tra hậu nghiệm.

### **Kiểm tra hậu nghiệm:**



*Với mức ý nghĩa là 0.05, ta nhận thấy*

Phương pháp 1 và phương pháp 2 không có sự khác biệt về hiệu quả giảm cân

Phương pháp 1 và phương pháp 3 có sự khác biệt về hiệu quả giảm cân

Phương pháp 2 và phương pháp 3 có sự khác biệt về hiệu quả giảm cân

### **Kiểm tra các điều kiện cho kiểm định Anova**

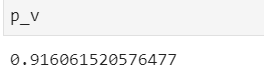
*Vẽ biểu đồ qqplot và histogram:*

**Chart, histogram

Description automatically generated**

*Bằng mắt thường ta có thể nhận thấy dữ liệu trên phân phối chuẩn, để kiểm tra chính xác hơn ta tiến hành kiểm định shapiro:*

*p\_v = stats.shapiro(model.resid)[1]*

**

Chấp nhận kiểu phân bố dữ liệu trên là phân phối chuẩn.

Tiến hành kiểm định về phương sai. Thỏa mãn các điều kiện để kiểm định ANOVA một chiều

# Tài liệu trích dẫn:

1. Thống kê mô tả. In: Wikipedia tiếng Việt [Internet]. 2021 [cited 2021 Jun 28]. Available from: https://vi.wikipedia.org/w/index.php?title=Th%E1%BB%91ng\_k%C3%AA\_m%C3%B4\_t%E1%BA%A3&oldid=64393023

2. Suy luận thống kê. In: Wikipedia tiếng Việt [Internet]. 2021 [cited 2021 Jun 28]. Available from: https://vi.wikipedia.org/w/index.php?title=Suy\_lu%E1%BA%ADn\_th%E1%BB%91ng\_k%C3%AA&oldid=64479353

# Kết quả kiểm tra đạo văn:

