

ĐẠI HỌC KINH TẾ TP. HỒ CHÍ MINH



College of  
Technology and Design

SCHOOL OF BUSINESS INFORMATION TECHNOLOGY

# BÁO CÁO



**Nhóm sinh viên:** Nguyễn Đôn Đức, Đặng Thị Thu Hiền, Bùi Tiến Hiếu, Đỗ Thanh Hoa, Nguyễn Huy Hoàng

**Môn học:** Biểu diễn trực quan dữ liệu

**Giảng viên hướng dẫn:** Nguyễn An Tế

**Đề tài:** Phân tích các yếu tố nhân khẩu học, tình trạng sức khỏe và hiệu suất tập luyện của người tập gym

Hồ Chí Minh, ngày 08 tháng 12 năm 2024

## THÀNH VIÊN NHÓM

STT	Họ và tên	MSSV
1	Nguyễn Đôn Đức	31221024296
2	Đặng Thị Thu Hiền	31221025556
3	Bùi Tiến Hiếu	31221026291
4	Đỗ Thanh Hoa	31211025193
5	Nguyễn Huy Hoàng	31221023992

## LỜI NÓI ĐẦU

Kính gửi quý thầy cô và các quý độc giả, đây là bài báo cáo thuộc đồ án cuối kỳ môn Biểu diễn trực quan dữ liệu. Cấu trúc bài báo cáo chia làm 5 chương,

*Chương 1:* Tổng quan đề tài của nhóm, mục tiêu mà nhóm muốn đạt được cũng như cách thức mà nhóm sẽ thực hiện để đạt được mục tiêu,

*Chương 2:* Giới thiệu về bộ dữ liệu Gym Members Exercise Dataset,

*Chương 3:* Quy trình mà nhóm sẽ tiến xử lý dữ liệu để có thể phân tích và biểu diễn,

*Chương 4:* Phân tích và biểu diễn trực quan kết quả mà nhóm thu hoạch được,

*Chương 5:* Thảo luận về kết quả đạt được, hạn chế cũng như đề xuất hướng phát triển.

Vì gặp giới hạn về thời gian và kiến thức nên bài báo cáo không thể tránh khỏi những sai sót. Nhóm tác giả rất mong nhận được sự đóng góp ý kiến từ quý thầy cô và các quý độc giả để đồ án được hoàn thiện hơn.

Nhóm tác giả cũng xin chân thành cảm ơn Ts. Nguyễn An Tế đã tận tình giảng dạy, hướng dẫn để nhóm hoàn thành đồ án này.

Trân trọng,

Nhóm 03

## MỤC LỤC

### DANH MỤC HÌNH ẢNH

CHƯƠNG I: TỔNG QUAN ĐỀ TÀI	1
1.1. Tổng quan đề tài	1
1.2. Mục tiêu nghiên cứu	2
1.3. Phương pháp nghiên cứu	3
CHƯƠNG II: TỔNG QUAN BỘ DỮ LIỆU	5
2.1. Sơ lược về dữ liệu	5
2.2. Mô tả thuộc tính	5
CHƯƠNG III: TIỀN XỬ LÝ DỮ LIỆU	7
3.1. Kiểm tra giá trị thiếu	7
3.2. Kiểm tra giá trị trùng lặp	7
3.4. Mã hóa dữ liệu	8
CHƯƠNG IV: PHÂN TÍCH VÀ BIỂU DIỄN TRỰC QUAN	9
4.1. Phân tích đơn biến	9
4.2. Phân tích đa biến	19
4.2.1. Phân tích tương quan sức khỏe và thể trạng của người tập	19
4.2.2. Phân tích tương quan về thói quen tập luyện của người tập	21
4.2.3. Phân tích quá trình tập luyện của người tập	25
4.4. Khuyến nghị và đề xuất	36
CHƯƠNG V: KẾT LUẬN	38
5.1. Kết quả đạt được	38
5.2. Hướng phát triển	38
TÀI LIỆU THAM KHẢO	
ĐÁNH GIÁ CÔNG VIỆC	

## DANH MỤC HÌNH ẢNH

Hình 3.1: Kiểm tra giá trị thiếu	6
Hình 3.2: Kiểm tra giá trị trùng lặp	7
Hình 4.1: Phân phối tuổi giữa nam và nữ	9
Hình 4.2: Phân phối của cân nặng	10
Hình 4.3: Phân phối của chiều cao	10
Hình 4.4: Biểu đồ hộp của nhịp tim tối đa, trung bình khi tập và ở trạng thái nghỉ	12
Hình 4.5: Phân phối của thời gian tập	13
Hình 4.6: Phân phối của calories	14
Hình 4.7: Biểu đồ thanh của loại hình thể thao	15
Hình 4.8: Phân phối của tỷ lệ chất béo	16
Hình 4.9: Biểu đồ thanh của số ngày tập trong tuần	17
Hình 4.10: Biểu đồ thanh của nhóm BMI	18
Hình 4.11: Phân phối nhóm BMI theo độ tuổi	19
Hình 4.12: Tương quan giữa % mỡ và cân nặng theo nhóm chỉ số BMI	20
Hình 4.13: Ma trận tương quan của các biến liên quan đến thói quen tập luyện	21
Hình 4.15: Phân bố thời gian tập luyện theo mức độ kinh nghiệm	22
Hình 4.16: Thời lượng buổi tập theo số lượng buổi tập	23
Hình 4.17: Tần suất đi tập giữa các mức kinh nghiệm tập	24
Hình 4.18: Phân phối nhịp tim trung bình theo loại bài tập	25
Hình 4.19: Biểu đồ tương quan giữa BMI và các chỉ số về nhịp tim	26
Hình 4.20: Kết quả Kiểm định Pearson giữa Max_BPM và BMI	27
Hình 4.21: Kết quả Kiểm định Pearson giữa Resting_BPM và BMI	27
Hình 4.22: Kết quả Kiểm định Pearson giữa Avg_BPM và BMI	28
Hình 4.23: Tương quan giữa lượng Calories đốt và các biến liên quan đến nhịp tim	28
Hình 4.24: Kết quả Kiểm định Pearson giữa Max_BPM và Calories_Burned	29

Hình 4.25: Kết quả Kiểm định Pearson giữa Resting_BPM và Calories_Burned	29
Hình 4.26: Kết quả Kiểm định Pearson giữa Avg_BPM và Calories_Burned	30
Hình 4.27: Tỷ lệ mỡ cơ thể theo tần suất tập luyện	31
Hình 4.28: Kết quả Kiểm định Shapiro-Wilk và Levene về phân phối và phương sai Fat Percentage giữa các nhóm	32
Hình 4.29: Kết quả Kiểm định Kruskal-Wallis về sự khác biệt về lượng % mỡ cơ thể giữa các nhóm người có tần suất tập luyện khác nhau	32
Hình 4.30: Tương quan giữa thời gian tập luyện và lượng calo đốt cháy	32
Hình 4.31: Biểu đồ hộp thể hiện lượng Calories tiêu hao theo các loại bài tập	33
Hình 4.32: Kết quả kiểm định Shapiro-Wilk và Levene về lượng Calories giữa các loại bài tập	34
Hình 4.33: Kết quả Kiểm định Kruskal-Wallis về lượng Calories giữa các loại bài tập.	34

## **DANH MỤC BẢNG BIỂU**

Bảng 1: Mô tả các thuộc tính của tập dữ liệu

6

## CHƯƠNG I: TỔNG QUAN ĐỀ TÀI

### 1.1. Tổng quan đề tài

Hàng ngàn năm nay, sức khỏe vẫn luôn được coi là ưu tiên hàng đầu của con người. Trong thời đại hiện nay, sức khỏe càng nhận được nhiều sự quan tâm hơn bao giờ hết. Khi đối diện với nhịp sống hối hả, môi trường ô nhiễm và áp lực cuộc sống đang ngày càng gia tăng thì con người cần một sức khỏe tốt. Việc duy trì được một sức khỏe tốt không chỉ giúp ta nâng cao được chất lượng cuộc sống mà còn cải thiện sức khỏe tinh thần, cải thiện ngoại hình, gia tăng năng suất lao động, tăng sức đề kháng từ đó giảm thiểu bệnh tật và nâng cao tuổi thọ.

Tuy vậy, trong bối cảnh tại Việt Nam, không gian sống tại các thành phố thường hạn hẹp, tập luyện tại nhà và không gian công cộng thì thiếu thốn các dụng cụ phục vụ việc tập luyện nên phòng gym đã liên tục mọc lên và trở thành một lựa chọn tối ưu cho việc rèn luyện thể chất. Tập gym không còn là một xu hướng mà dần trở nên phổ biến và là một phần không thể thiếu với nhiều người. Tập gym giúp cải thiện sức khỏe toàn diện, không chỉ tăng cường sức mạnh về cơ bắp mà còn cải thiện hệ tim mạch, kích thích việc tiết hoóc-môn hỗ trợ cân bằng tâm lý và giảm căng thẳng, cải thiện ngoại hình và nâng cao sự tự tin cho người tập. Tuy nhiên, nếu tập luyện không đúng cách và không phù hợp với thể trạng của bản thân có thể gây ra những hậu quả khôn lường như tổn thương hệ cơ xương khớp, tổn hại đến tim, rối loạn hoóc-môn, kiệt sức hay thậm chí có những trường hợp dẫn đến tử vong. Tiền mất mà tật mang.

Từ những yếu tố kể trên đã đặt ra yêu cầu về phân tích và khai thác dữ liệu về thói quen tập luyện để nâng cao trải nghiệm và bảo vệ người tập. Bộ dữ liệu “gym\_members\_exercise\_tracking.csv” cung cấp cho ta một nguồn thông tin phong phú về các hoạt động thể chất của người dùng và các thông tin nhân khẩu học khác. Phân tích dữ liệu giúp ta khám phá các xu hướng tập luyện, đánh giá được mức độ gắn kết từ đó tạo ra được những phương pháp, những thông số nhằm hỗ trợ việc luyện tập của khách hàng. Tuy nhiên, dữ liệu dưới dạng thô thường khó tiếp cận bởi nếu nhìn vào dữ liệu thô sẽ khó thấy được tri thức tiềm ẩn trong lĩnh vực của tập dữ liệu và không thể sử dụng trực tiếp nếu người sử dụng không có chuyên môn về dữ liệu. Do đó, trực quan hóa dữ liệu đóng



vai trò quan trọng trong việc chuyển đổi các thông tin phức tạp và khó hiểu thành dạng hình ảnh và những biểu đồ trực quan, dễ hiểu giúp hỗ trợ các phòng gym trong việc xây dựng chiến lược tập luyện cho khách hàng, tối ưu hóa trải nghiệm của khách hàng bằng việc tạo niềm tin cho họ, giúp họ hiểu và tuân thủ những chiến lược đã được đề ra nhằm cải thiện hiệu suất tập luyện.

## **1.2.Mục tiêu nghiên cứu**

Mục tiêu chính của nghiên cứu này là ứng dụng phân tích và trực quan hóa dữ liệu để khai thác dữ liệu từ bộ dữ liệu “gym\_members\_exercise\_tracking.csv” nhằm tìm hiểu sâu hơn về thói quen tập luyện của các thành viên phòng gym. Thông qua đó ta hướng tới việc đưa ra những thông tin hữu ích giúp hỗ trợ việc cải thiện hiệu quả tập luyện của khách hàng tại các phòng gym. Với các mục tiêu cụ thể như sau:

Thứ nhất, nghiên cứu tập trung vào việc phân tích đặc trưng của dữ liệu. Ta khám phá phân phối và các đặc điểm thống kê của các biến số, xác định được các mối tương quan quan trọng trong dữ liệu.

Thứ hai, nghiên cứu đi sâu vào việc phân tích các nhóm đối tượng. Nhóm quyết định phân nhóm đối tượng thành nhóm khách hàng có hiệu quả tập luyện chưa tốt và nhóm có hiệu quả tập luyện cao. Với nhóm có hiệu quả tập luyện chưa tốt ta tìm hiểu các đặc điểm của nhóm này như nhân khẩu học, thói quen tập luyện, thể trạng .... từ đó tìm hiểu nguyên nhân dẫn đến việc tập chưa hiệu quả. Với nhóm có hiệu quả tập luyện cao, ta nghiên cứu các hành vi và đặc điểm nổi bật của họ nhằm xác định được các yếu tố cốt lõi dẫn đến hiệu quả vượt trội của họ. Sau khi phát hiện ra các yếu tố, ta tiến hành rút ra bài học và các khuyến nghị phù hợp để áp dụng cho nhóm có hiệu quả tập luyện chưa tốt.

Thứ ba, nghiên cứu xây dựng các kế hoạch cải thiện hiệu quả tập luyện thông qua việc kết hợp thông tin và hiểu biết từ hai nhóm khách hàng. Hướng tới việc xây dựng kế hoạch tập luyện cá nhân hóa cho các thành viên.

Thứ tư, tối ưu khả năng truyền đạt thông tin thông qua trực quan hóa dữ liệu. Thiết kế các biểu đồ, dashboard thân thiện với người xem, giúp truyền tải thông tin một cách dễ hiểu, nhanh và hữu ích cho cả phòng gym và khách hàng.

### 1.3. Phương pháp nghiên cứu

Nghiên cứu này sử dụng tập dữ liệu “gym\_members\_exercise\_tracking.csv” từ Kaggle mục đích để hiểu rõ hơn về thói quen tập luyện của người tập gym từ đó rút ra những khuyến nghị dành cho họ. Quy trình được thực hiện như sau:

Đầu tiên, nhóm tiến hành tiền xử lý dữ liệu nhằm cải thiện chất lượng dữ liệu, đảm bảo tính nhất quán, chính xác và đầy đủ, từ đó giúp cho việc phân tích và khai thác dữ liệu hiệu quả hơn. Trong bài nghiên cứu này nhóm đã kiểm tra các giá trị thiếu, giá trị trùng lặp, giá trị nhiễu và làm sạch chúng. (Nhóm tiến hành chuyển đổi dữ liệu, mã hóa dữ liệu của một số biến để phù hợp cho phân phân tích.)

Tiếp theo, nhóm tiến hành phân tích đơn biến để khám phá các đặc trưng riêng lẻ của từng biến trong tập dữ liệu. Nhóm đã tìm hiểu về phân phối của tuổi, cân nặng, chiều cao, nhịp tim, thời gian tập luyện, lượng calories đốt, loại hình tập luyện, tỷ lệ mỡ trong cơ thể, lượng nước uống, tần suất tập luyện và chỉ số BMI từ đó rút ra những nhận định và đặt ra những câu hỏi cho phân phân tích đa biến.

Từ những câu hỏi trong phân phân tích đơn biến nhóm tiến hành phân tích đa biến để tìm ra những mối quan hệ có ý nghĩa giữa các biến khác nhau. Nhóm tập trung tìm kiếm mối tương quan về các nhóm liên quan đến sức khỏe và thể trạng của người tập, thói quen tập luyện của họ và quá trình họ tập luyện.

Sau khi phân tích, nhóm sử dụng các kiểm định thống kê để xác định ý nghĩa của các mối quan hệ quan sát được. Trong phần này nhóm sử dụng kiểm định tương quan Pearson, kiểm định phương sai đồng nhất Levene, kiểm định Shapiro (kiểm định mẫu dữ liệu có tuân theo phân phối chuẩn hay không). Nhóm sử dụng kiểm định Anova khi dữ liệu thỏa cả Levene và Shapiro, trong khi thực hiện nhận thấy dữ liệu không thỏa kiểm định Shapiro nên nhóm tiến hành kiểm định Kruskal-Wallis để thay thế cho kiểm định Anova.

Cuối cùng, nhóm tiến hành tổng hợp tri thức từ kết quả của phân tích đơn biến, phân tích đa biến và các kiểm định để đưa ra bức tranh tổng thể về thói quen tập luyện của các thành viên phòng gym và các khuyến nghị dành cho từng nhóm người cụ thể.

## CHƯƠNG II: TỔNG QUAN BỘ DỮ LIỆU

### 2.1. Sơ lược về dữ liệu

Bộ dữ liệu `gym_members_exercise_tracking.csv` là một bộ dữ liệu được lấy từ Kaggle, mô tả cung cấp cái nhìn chi tiết về thói quen tập luyện, các chỉ số thể chất, và thông tin liên quan đến các thành viên phòng gym. Bộ dữ liệu bao gồm 973 dòng thông tin và 15 cột thuộc tính, trong đó chứa các chỉ số liên quan đến sức khỏe, thể trạng cũng như hiệu suất tập luyện của người tập.

Đường dẫn đến bộ dữ liệu: [Link](#)

### 2.2. Mô tả thuộc tính

STT	Tên thuộc tính	Mô tả	Chú thích
1	Age	Tuổi của thành viên phòng gym	
2	Gender	Giới tính của thành viên phòng gym	“Male”: Nam “Female”: Nữ
3	Weight (kg)	Cân nặng của thành viên, tính bằng kilogram	
4	Height (m)	Chiều cao của thành viên, tính bằng mét.	
5	Max_BPM	Nhịp tim tối đa (số nhịp/phút) trong các lần tập	
6	Avg_BPM	Nhịp tim trung bình trong các lần tập luyện	
7	Resting_BPM	Nhịp tim khi nghỉ ngơi trước khi tập luyện	
8	Session_Duration (hours)	Thời gian của mỗi lần tập, tính bằng giờ	

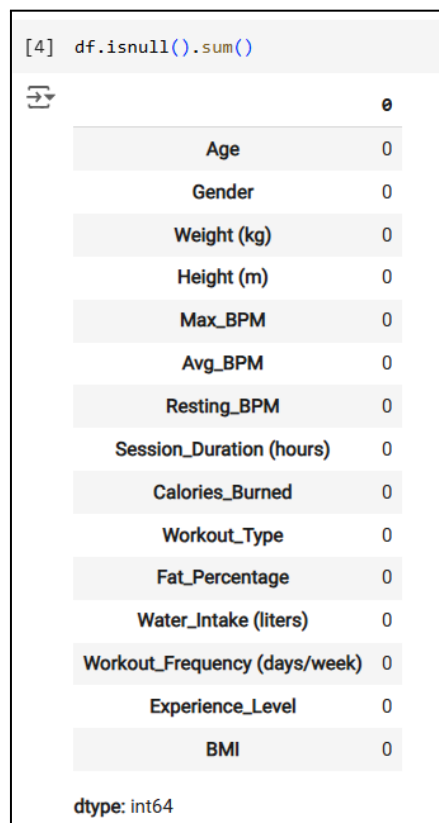
STT	Tên thuộc tính	Mô tả	Chú thích
9	Calories_Burned	Tổng lượng calo tiêu hao trong mỗi lần tập	
10	Workout_Type	Loại hình tập luyện thực hiện	“Cardio” “HIIT” “Strength” “Yoga”
11	Fat_Percentage	Tỷ lệ mỡ cơ thể của thành viên	
12	Water_Intake (liters)	Lượng nước tiêu thụ hàng ngày trong các lần tập, tính bằng lít	
13	Workout_Frequency (days/week)	Số ngày tập luyện trong một tuần	
14	Experience_Level	Mức độ kinh nghiệm	“1”: Người mới “2”: Đã có kinh nghiệm nhất định “3”: Chuyên gia
15	BMI	Chỉ số khối của cơ thể	

*Bảng 1: Mô tả các thuộc tính của tập dữ liệu*

## CHƯƠNG III: TIỀN XỬ LÝ DỮ LIỆU

### 3.1. Kiểm tra giá trị thiếu

Trước tiên thì nhóm sẽ kiểm tra giá trị thiếu, lý do cho việc này bởi nếu tồn tại giá trị thiếu nó sẽ gây ảnh hưởng đến chất lượng của bộ dữ liệu, tác động đến các nhận định, đánh giá nhóm đưa ra từ biểu diễn trực quan, đồng thời, giá trị thiếu sẽ gây cản trở cho quá trình tính toán các trị thống kê kiểm định, hệ số tương quan cũng như các chỉ số liên quan. Do đó ở đây, nhóm sẽ kiểm tra giá trị thiếu của tập dữ liệu. Sử dụng phương thức `df.isna().null()` để kiểm tra giá trị thiếu trong tập dữ liệu, nhóm nhận thấy rằng tập dữ liệu không có giá trị bị thiếu. Vậy, vấn đề về giá trị thiếu đã được giải quyết do không tồn tại giá trị thiếu trong tập dữ liệu này.



```
[4] df.isnull().sum()
```

	0
Age	0
Gender	0
Weight (kg)	0
Height (m)	0
Max_BPM	0
Avg_BPM	0
Resting_BPM	0
Session_Duration (hours)	0
Calories_Burned	0
Workout_Type	0
Fat_Percentage	0
Water_Intake (liters)	0
Workout_Frequency (days/week)	0
Experience_Level	0
BMI	0

dtype: int64

Hình 3.1: Kiểm tra giá trị thiếu

### 3.2. Kiểm tra giá trị trùng lặp

Yếu tố thứ hai mà ta cần xem xét đó là giá trị trùng lặp. Nếu ta không xử lý giá trị trùng lặp, đánh giá của ta sẽ phần nào được coi là thiên vị, bởi nó sẽ tăng số lượng của một đối

tượng nào đó lên. Ngoài ra thì giá trị trùng lặp cũng làm cho các trị thống kê và chỉ số khác trở nên kém chính xác hơn. Sử dụng phương thức `df.duplicated().sum()` để kiểm tra các giá trị trùng lặp, nhóm cũng nhận thấy rằng dữ liệu không bị trùng lặp ở bất kỳ dòng nào. Như vậy, vấn đề về giá trị trùng lặp đã được giải quyết.

```
df.duplicated().sum()  
0
```

*Hình 3.2. Kiểm tra giá trị trùng lặp*

## CHƯƠNG IV: PHÂN TÍCH VÀ BIỂU DIỄN TRỰC QUAN

Trước khi đi sâu và phân tích và biểu diễn trực quan dữ liệu cũng như xây dựng các giả thuyết liên quan, có một số lưu ý nhóm cần đọc giả lưu ý,

Một, mức ý nghĩa nhóm thống nhất đó là  $\alpha = 0.05$ .

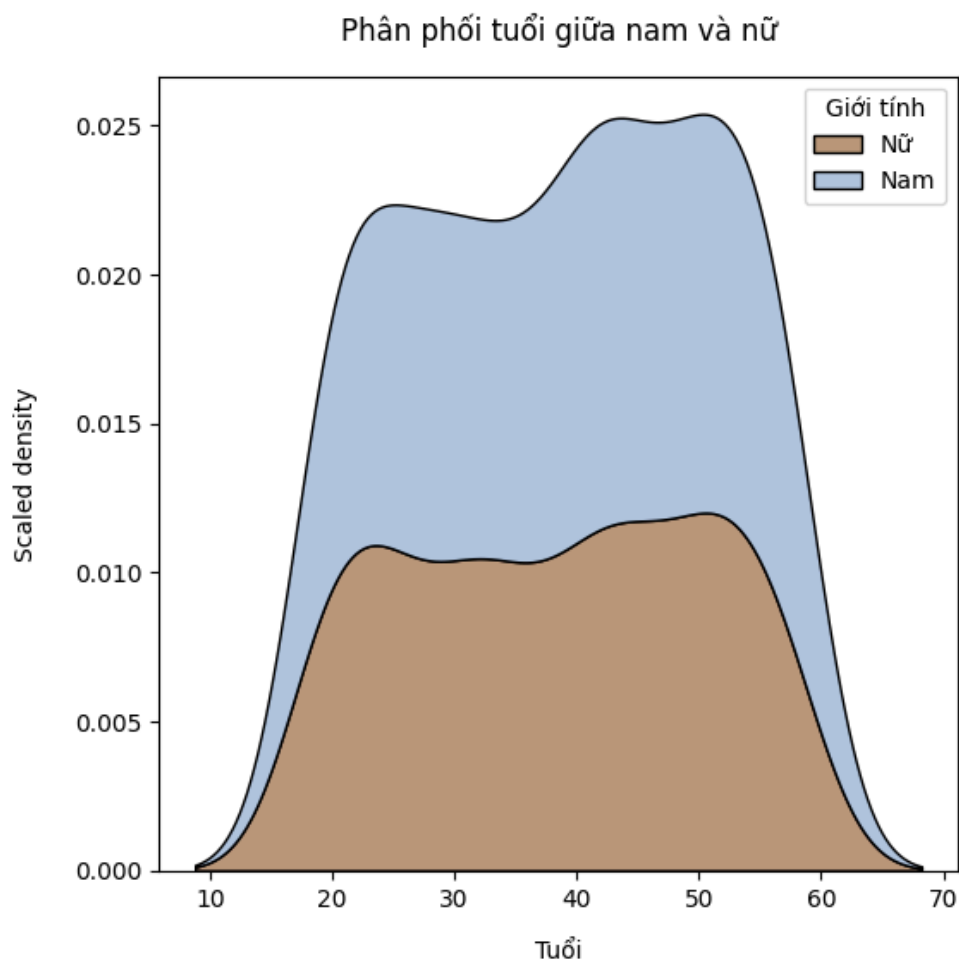
Hai, đối với các loại kiểm định như ANOVA, ta cần phải thỏa mãn các điều kiện, (i) các nhóm xem xét phải độc lập với nhau, (ii) phương sai phải đồng nhất giữa các nhóm và (iii) các nhóm phải có phân phối chuẩn. Để thỏa các điều kiện ANOVA đặt ra, nhóm sẽ sử dụng thêm các kiểm định khác như Kiểm định Shapiro-Wilk, Kiểm định Levene và Kiểm định Kruskal-Wallis.

### 4.1. Phân tích đơn biến

Trước tiên thì ta hãy cùng xem xét những người tập gym sẽ nằm trong những độ tuổi nào, với tập dữ liệu mà nhóm có được, nhóm tuổi nằm trong khoảng từ 18 đến 56 tuổi. Ở đây ta sẽ tiến hành phân tuổi thành ba nhóm, thứ nhất đó là nhóm thanh niên (18-30), thứ hai là nhóm trung niên (31-50) và nhóm cao tuổi ( $> 50$ ).

Sau khi phân nhóm thì nhóm tiến hành xem xét nhóm tuổi nào có số lượng nhiều nhất, điều đó được thực hiện bằng cách lấy mode của các nhóm này. Kết quả trả về đó là nhóm trung niên, như vậy thì giữa ba nhóm thì nhóm trung niên chiếm nhiều nhất. Có lẽ lý do cho nhóm trung niên phổ biến nhất trong những người tập gym đó là những người thuộc nhóm tuổi này đã có công việc ổn định, thu nhập ổn định. Họ quan tâm hơn đến sức khỏe và thể trạng của mình hơn và phải chăng điều đó đã thúc đẩy họ đến với gym.

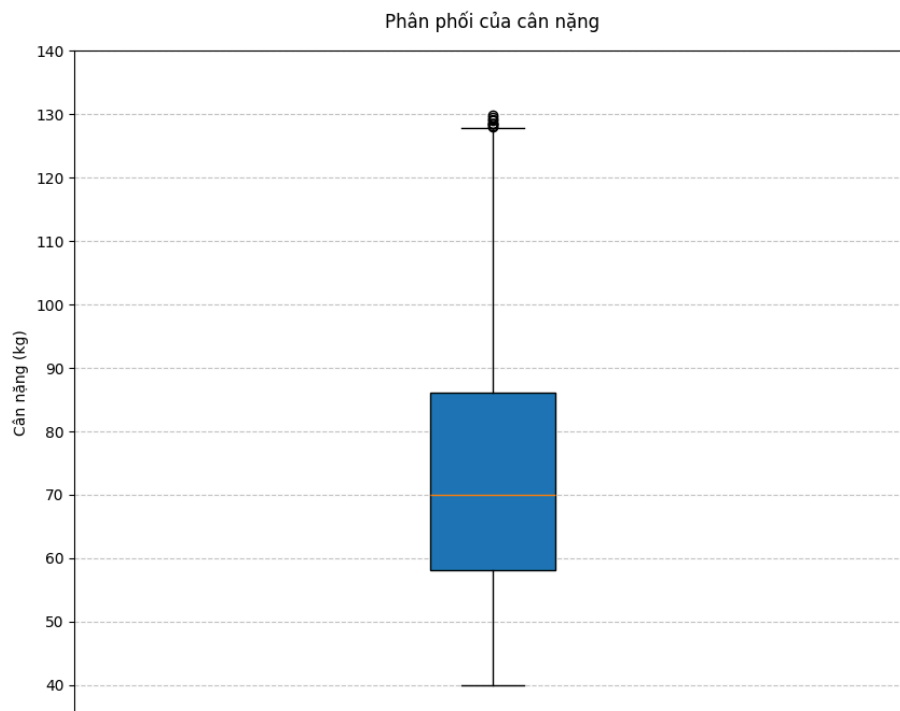
Tiếp đến, ta sẽ coi đến tuổi giữa nam và nữ, liệu tuổi giữa nam và nữ có gì khác biệt. Nam nhiều hơn nữ hay nữ nhiều hơn nam hoặc cả hai nhóm là như nhau. Quan sát biểu đồ dưới thì ta biết được rằng, nam nhiều hơn nữa thậm chí có phần chiếm ưu thế hơn so với nữ khi xét đến tập gym. Tuy biểu đồ này ban đầu khó giải thích nhưng nó đủ để cho ta thấy và nhấn mạnh rằng nam nhiều hơn nữ ở các nhóm tuổi khi xét đến tập gym.



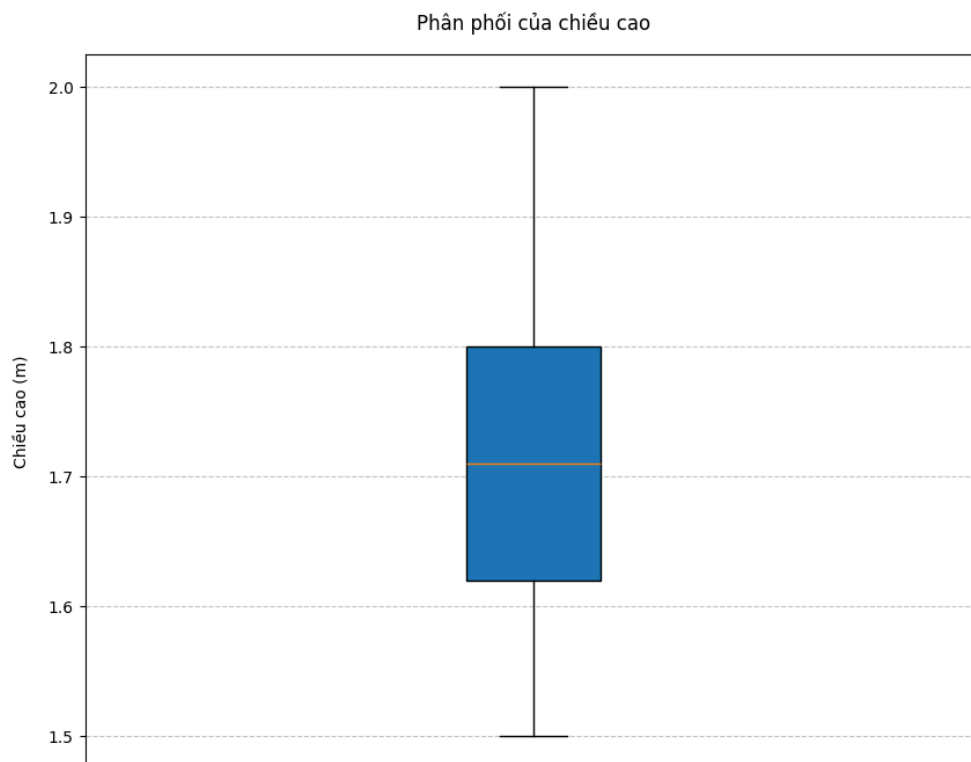
*Hình 4.1: Phân phối tuổi giữa nam và nữ*

Để đi sâu hơn nữa, nhóm sẽ tiến hành khám phá về cân nặng cũng như chiều cao của những người tập gym,





Hình 4.2: Phân phối của cân nặng

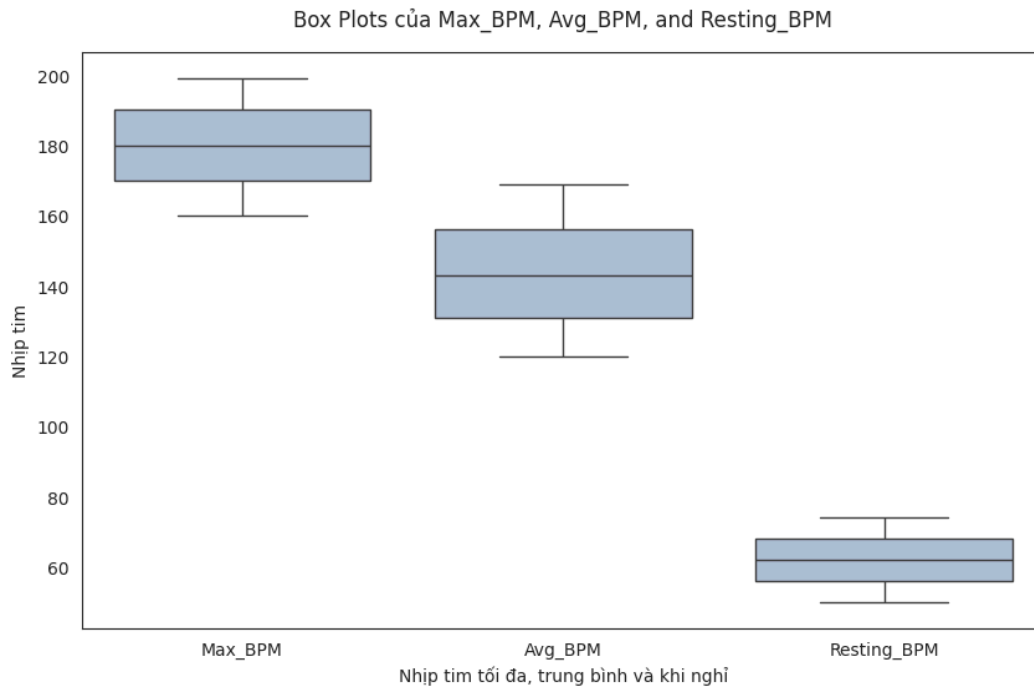


Hình 4.3: Phân phối của chiều cao

Cân nặng của những người tập gym trải dài từ 40kg đến khoảng 120kg, với biểu đồ boxplot trên thì có thể thấy tồn tại một số giá trị ngoại lai nhưng khi kiểm tra lại bằng phương pháp 3-sigma thì không tồn tại điểm ngoại lai nào nên có thể nói biến cân nặng không tồn tại điểm ngoại lai, phần râu bắt đầu từ Q3 có phần dài hơn râu ở Q1 trở xuống nên có thể thấy cân nặng phân tán hơn ở khoảng (Q3, Q4). Có lẽ vì thế mà ảnh hưởng đến trung bình, trung bình của cân nặng xấp xỉ 73,85 (kg).

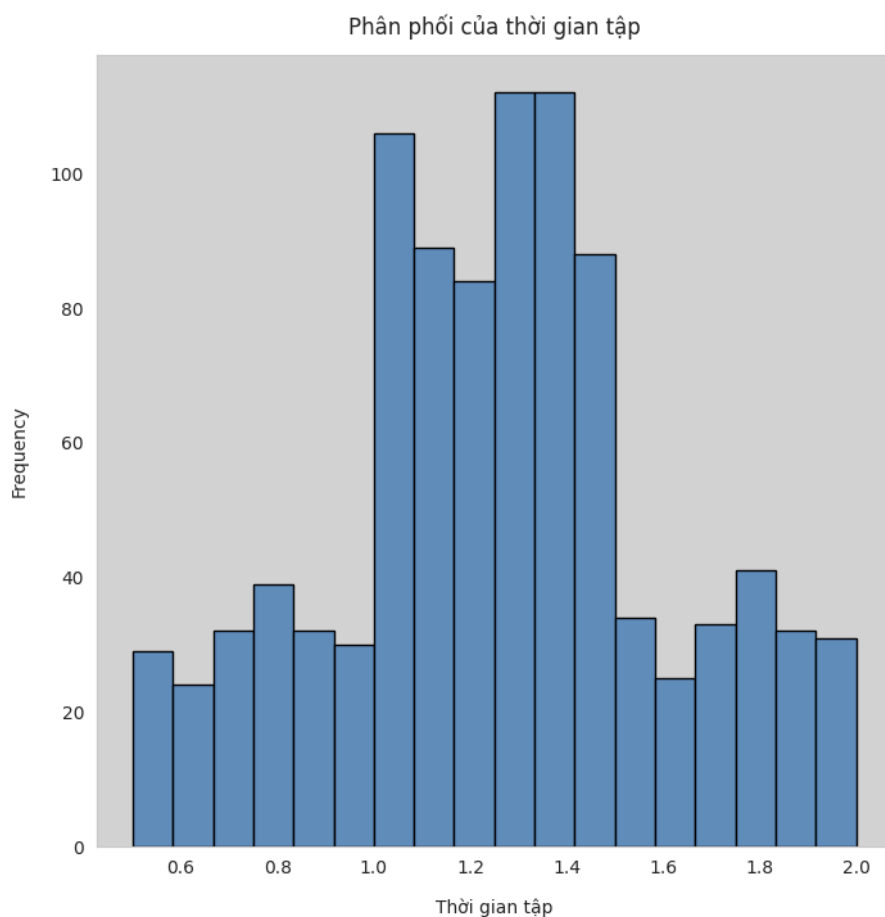
Chiều cao dựa trên boxplot không có các giá trị ngoại lai, hai râu nhìn chung không chênh lệch quá lớn nhưng vẫn có sự khác nhau đáng kể, median dường như ở vị trí "cân bằng" hơn. Xét trên tổng thể, chiều cao vào khoảng từ 1.5m cho tới 2.0m, chiều cao trung bình mà nhóm đo lường được là 1.722m. Độ lệch chuẩn 0.12787, không quá lớn cho thấy chiều cao phân tán không quá rộng.

Khi nhắc tới tập gym cũng như các vấn đề sức khỏe khác, thì một trong những yếu tố sẽ được quan tâm đó là nhịp tim, ở đây ta sẽ vẽ boxplot của nhịp tim tối đa khi tập (Max\_BPM), nhịp tim trung bình khi tập (Avg\_BPM) cũng như nhịp tim ở trạng thái nghỉ trước khi nghỉ (Resting\_BPM). Xét về độ phân tán thì nhịp tim trước khi nghỉ là hẹp hơn so với hai biến còn lại. Điều có thể được diễn giải dễ dàng bởi nhịp tim con người ở trạng thái nghỉ ngơi ít biến động hơn so với khi tập do đó mà nhịp tim nghỉ ngơi ít biến động hơn. Ngoài ra, khi tập thể dục thì cơ thể sẽ cần nhiều oxy và năng lượng hơn, lúc đó tim sẽ bắt đầu đập nhanh hơn để bơm máu và oxy đến các cơ vì thế mà khi tập nhịp tim sẽ tăng cao, vậy các yếu tố như cường độ tập, thời gian tập,... dường như có tác động đến việc này. Ta có thể tìm hiểu xem liệu cường độ tập tăng lên, thời gian tập tăng lên thì nhịp tim sẽ như thế nào, lượng calories biến đổi ra sao hòng trích ra được những hiểu biết về sự khác nhau trong nhịp tim giữa các nhóm người cũng như lượng calories được đốt cháy như thế nào.



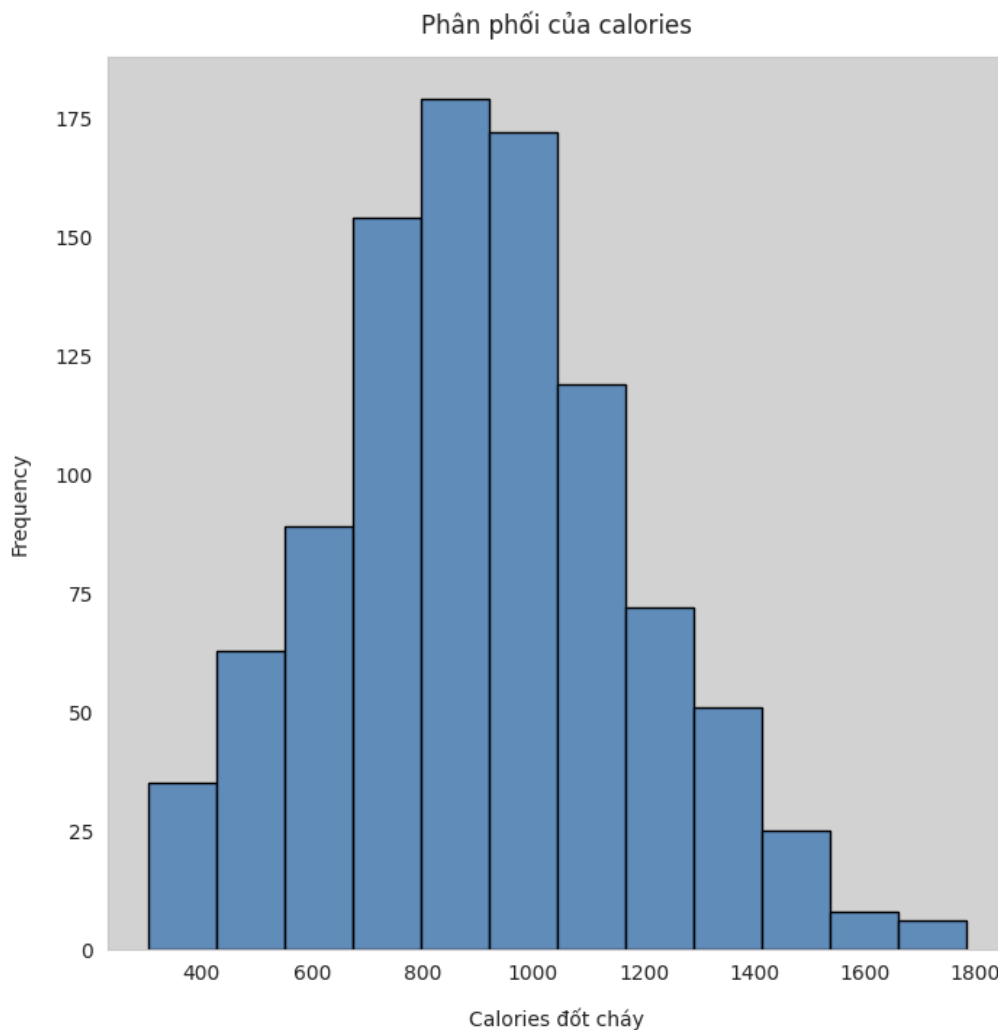
*Hình 4.4: Biểu đồ hộp của nhịp tim tối đa, trung bình khi tập và ở trạng thái nghỉ*

Một mối quan tâm được gợi lên cho chúng ta, đây là mối quan tâm về thời gian tập và lượng calories đốt cháy. Trước nhất ta hãy thử khám phá thời gian tập, với biểu đồ histogram của thời gian tập như hình dưới,



*Hình 4.5: Phân phối của thời gian tập*

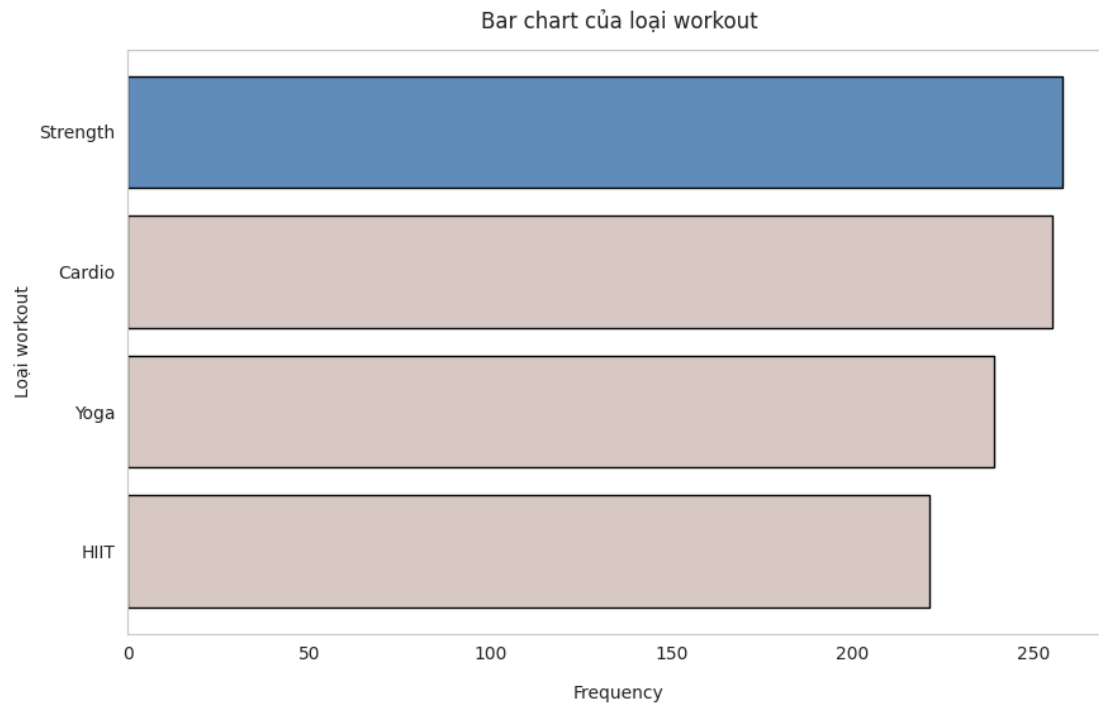
Từ biểu đồ trên ta có thể thấy rằng phân phối của thời gian tập dường như không bị nghiêng về một phía nào, suy ra rằng phân phối không bị lệch giúp tăng tính chính xác của các ước lượng liên quan đến thời gian tập. Tương tự với lượng calories đốt cháy khi tập, ta cũng vẽ biểu đồ histogram,



*Hình 4.6: Phân phối của calories*

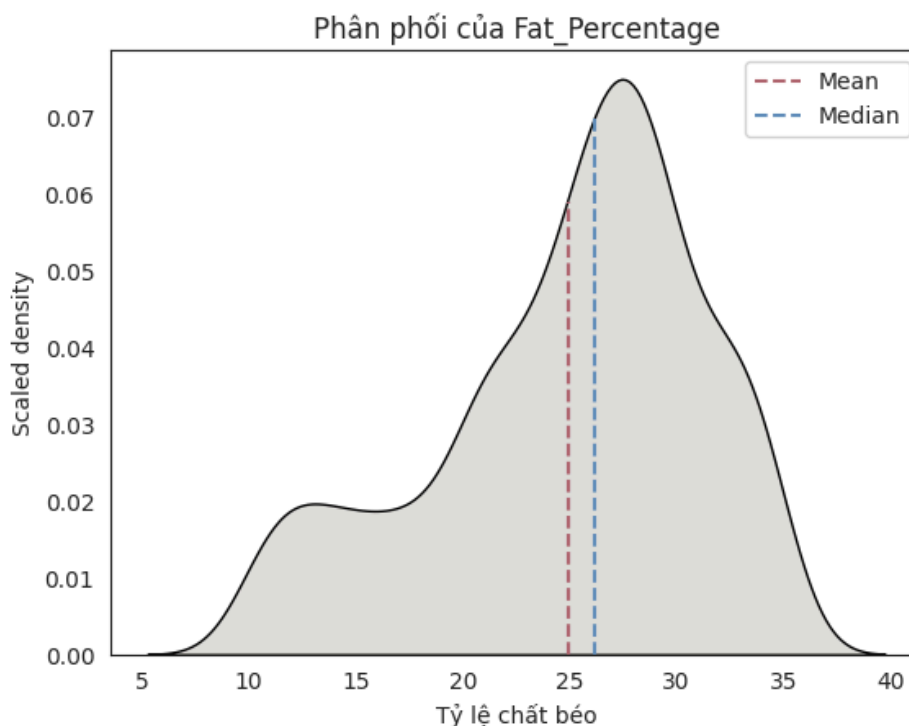
Với kết quả từ biểu đồ histogram thì dường như phân phối của calories xấp xỉ phân phối chuẩn, tuy vậy có vẻ như bị lệch. Để rõ hơn thì nhóm đã tính mean và median của biến này. Thu được 893.0 cho mean và 905.42 cho median đã củng cố cho lập luận phân phối bị lệch ban đầu. Lượng calories đốt cháy tập trung từ khoảng 800 cho tới 1000 đơn vị.

Trong bốn loại hình thể thao Strength, Cardio, Yoga và HIIT thì Strength là loại hình được chọn nhiều nhất tuy vậy Cardio cũng không kém cạnh khi có chênh lệch rất nhỏ so với Strength. Điều này cho thấy rằng người tập đang có xu hướng lựa chọn Strength và Cardio là loại hình tập luyện để nâng cao sức khỏe của họ.



*Hình 4.7: Biểu đồ thanh của loại hình thể thao*

Phân phối của tỷ lệ chất béo khá lệch, rõ hơn nữa ta hãy nhìn vào đường mean và median. Mean bé hơn median cho thấy một phân phối lệch trái. Phân phối lệch khiến đánh giá của ta bị thiên vị nhưng sẽ khá thú vị nếu ta những người có tỷ lệ chất béo ở khoảng đuôi trái này có gì đặc biệt trong gym.



Hình 4.8: Phân phối của tỷ lệ chất béo

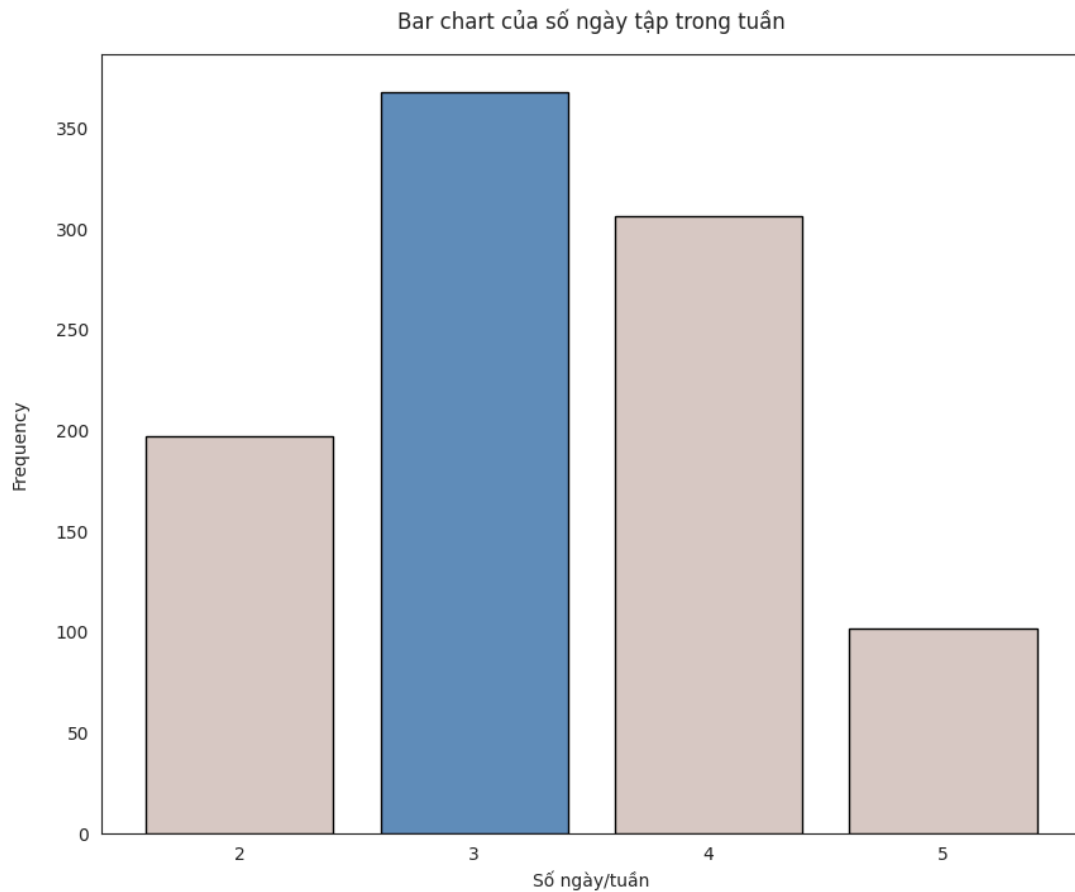
Nghiên cứu kế hoạch tập gym thì ta cũng cần quan tâm đến lượng nước tiêu thụ cũng như trình độ của người tập. Nhóm đã khám phá thấy được lượng nước tiêu thụ rơi vào khoảng 24.98 lít, độ lệch chuẩn vào khoảng 6.26. Trong ba trình độ từ 1 (người mới tập) cho tới 3 (chuyên gia) thì số đông nằm ở mức 2. Với mức này có thể đó là mức phổ thông, mức mà nhiều người đều đạt được.

Số đông những người tập gym thường tập 3 ngày/tuần. Để có thể chính xác hơn, ta sẽ tiến hành kiểm thử xem liệu có phải chẳng trung bình số buổi tập mỗi tuần của những người tập gym là 3. Với mức ý nghĩa  $\alpha = 5\%$ , ta phát biểu cặp giả thuyết  $H_0$  và  $H_1$  như sau,

$$H_0: \mu = 3$$

$$H_1: \mu \neq 3.$$

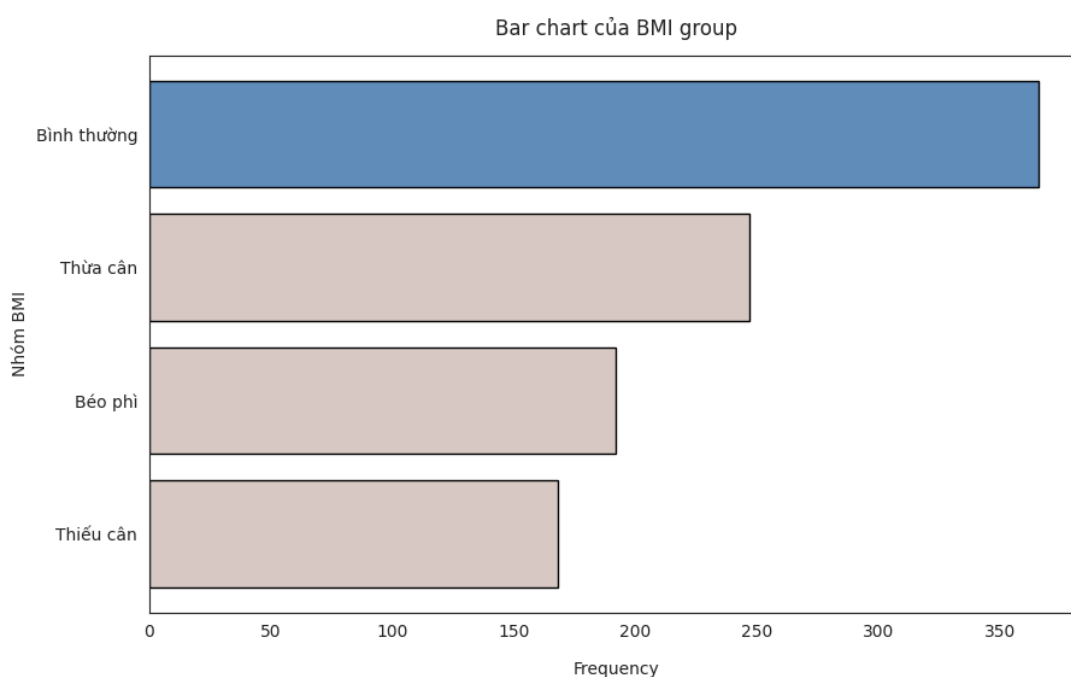
Ta thu được p-value  $0.00 < 0.05$  nên ta bác bỏ  $H_0$  vì vậy mà trung bình số buổi tập mỗi tuần của những người tập gym là khác 3.



*Hình 4.9: Biểu đồ thanh của số ngày tập trong tuần*

Để có thể phân tích tình trạng sức khỏe của những người tập gym thì chỉ số khối cơ thể (BMI) là một yếu tố không thể thiếu. Ở đây nhóm sẽ rời rạc hóa biến BMI, phân thành bốn nhóm, thứ nhất đó là những người thiếu cân ( $BMI < 18.5$ ), thứ hai đó là những người bình thường ( $18.5 \leq BMI < 24.9$ ), thứ ba là những người thừa cân ( $24.9 \leq BMI < 30$ ) và thứ tư đó là những người béo phì ( $BMI \geq 30$ ).





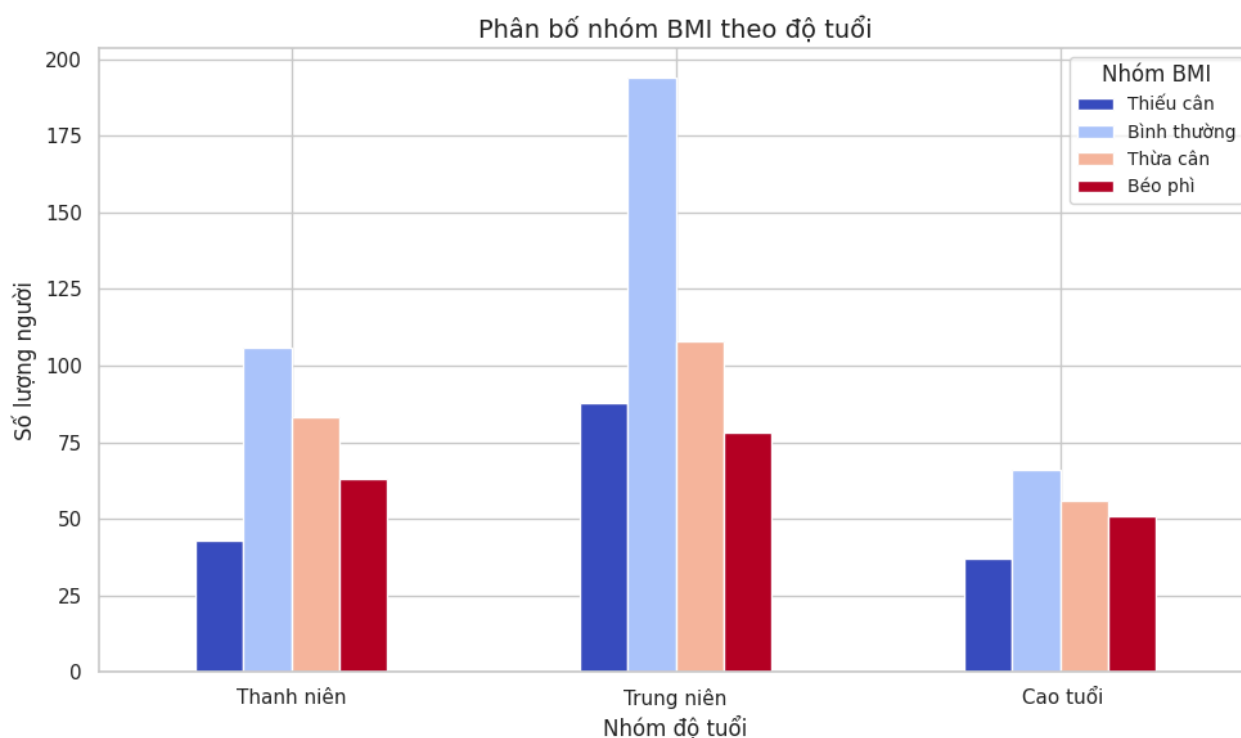
*Hình 4.10: Biểu đồ thanh của nhóm BMI*

Nhóm người đến tập gym nhiều nhất đó là những người có chỉ số khối cơ thể ở mức bình thường. Những người này đến tập gym có lẽ phần vì muốn duy trì vóc dáng cũng như cải thiện sức khỏe. Nhóm thừa cân và béo phì mục đích của họ đến với gym là vì muốn giảm trọng lượng cơ thể, nhóm thừa cân phải chăng đó là vì muốn tăng sự trao đổi chất để tăng khả năng hấp thụ chất.

## **4.2. Phân tích đa biến**

### **4.2.1. Phân tích tương quan sức khỏe và thể trạng của người tập**

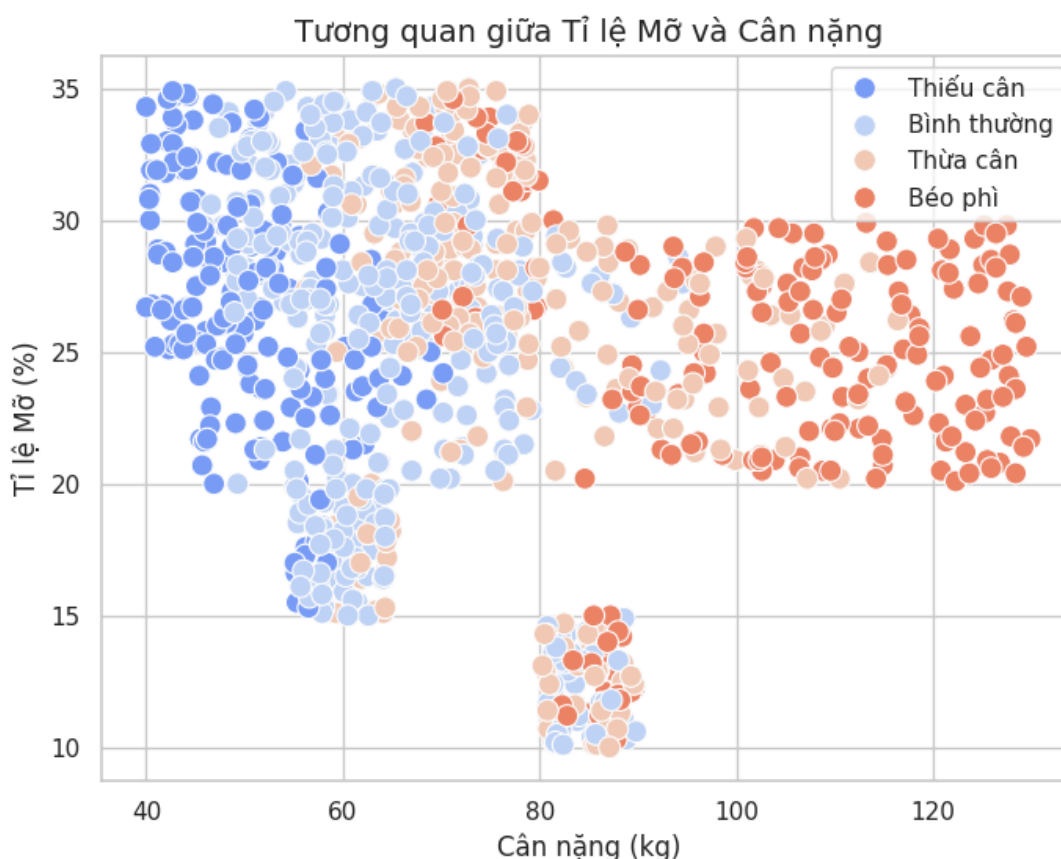
Ta sẽ tiến hành tiến hành phân tích các chỉ số sức khỏe cũng như thể trạng của người tập thông qua tuổi, chỉ số khối cơ thể, tỷ lệ chất béo và cân nặng của họ. Bước đầu, ta hãy thử khám phá BMI giữa các nhóm tuổi,



*Hình 4.11: Phân phối nhóm BMI theo độ tuổi*

Dựa vào biểu đồ Hình 4.11, xét nhóm tuổi thanh niên từ 18 đến 30 tuổi, tuổi đời trẻ đồng thời đang trong giai đoạn phát triển nên nhóm chỉ số BMI bình thường có số lượng người đông nhất, tuy nhiên xếp sau đó là nhóm người Thừa cân và Béo phì và cuối cùng là Thiếu cân. Điều này khá đáng chú ý bởi vì khi càng lớn tuổi, càng cần phải kiểm soát dinh dưỡng để tránh các bệnh liên quan đến tim mạch, nhiễm mỡ...

Để xem xét kĩ hơn về tình trạng cơ thể của từng nhóm BMI, nhóm tiến hành vẽ biểu đồ theo tỉ lệ mỡ và cân nặng giữa các nhóm chỉ số BMI:

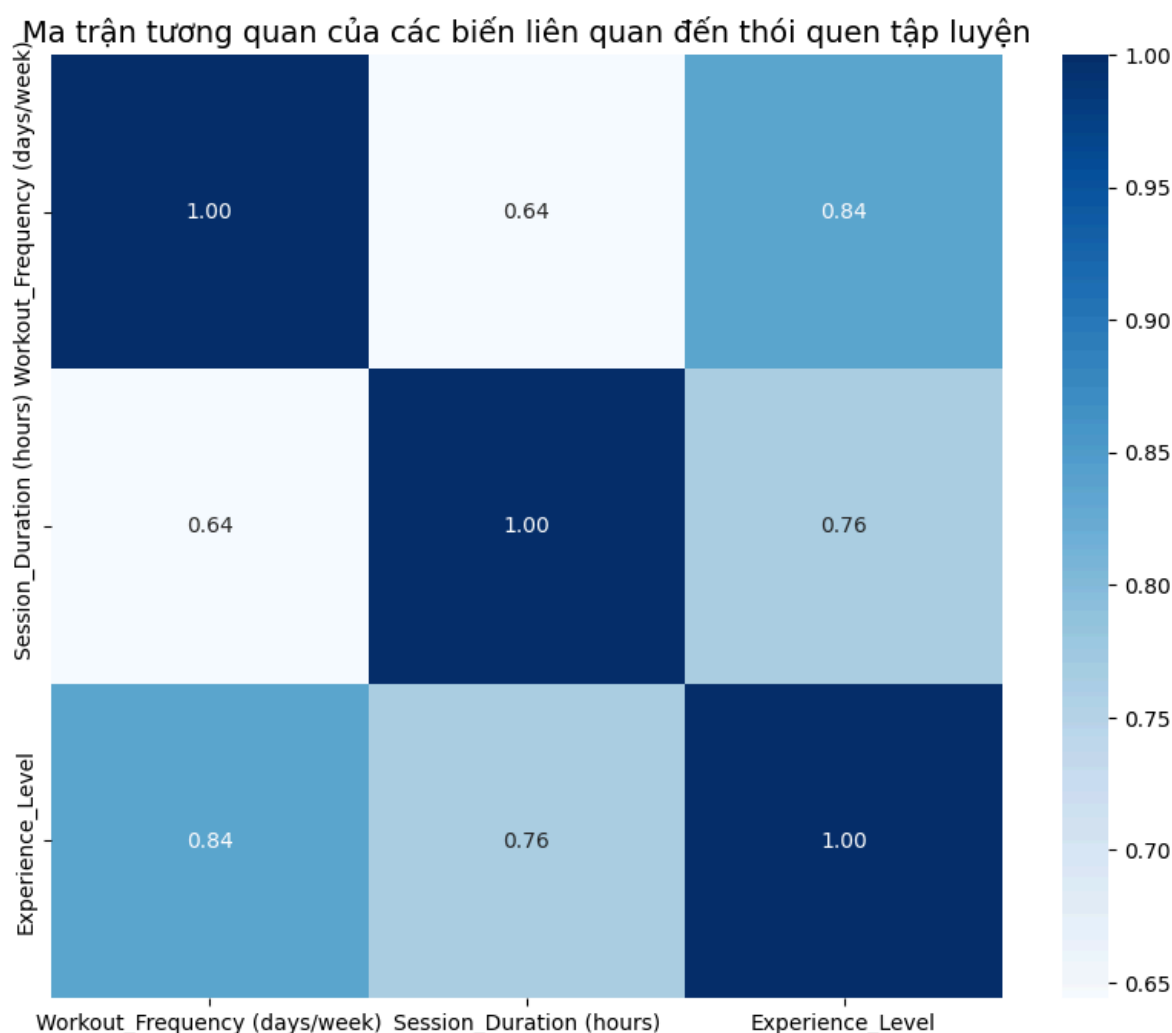


Hình 4.12: Tương quan giữa % mỡ và cân nặng theo nhóm chỉ số BMI

Qua biểu đồ Hình 4.12, ta có thể thấy sự phân bố của các điểm dữ liệu khi biểu diễn theo nhóm BMI. Dễ nhận thấy rằng nhóm người từ thừa cân đến béo phì phân bố chủ yếu từ 70kg trở lên nhưng tỷ lệ mỡ cơ thể lại thấp hơn. Mâu thuẫn trên xuất phát từ công thức tính tỷ lệ mỡ cơ thể, khi ta so sánh 25% mỡ của một người có cân nặng 60kg với một người cũng có 25% mỡ nhưng cân nặng lên đến 100kg. Ngoài ra, khi tính cân nặng còn bao gồm về khối lượng cơ, xương và nước trong cơ thể, do đó mối tương quan giữa tỷ lệ % mỡ và cân nặng không phản ánh đúng thể trạng thực tế của một người.

#### 4.2.2. Phân tích tương quan về thói quen tập luyện của người tập

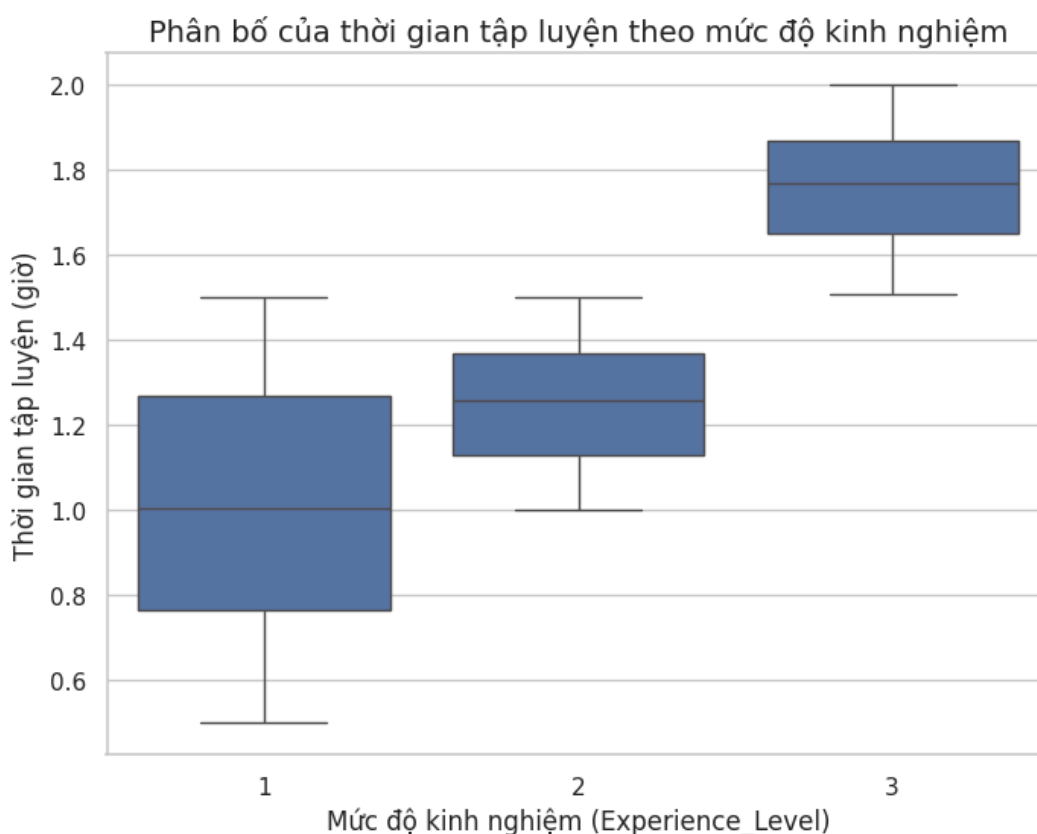
Để phân tích và hiểu rõ hơn về thói quen tập luyện của các thành viên trong phòng gym thì nhóm lựa chọn các biến liên quan đến thói quen tập luyện như Workout\_Frequency (days/week), Session\_Duration (hours) và Experience\_Level



Hình 4.13: Ma trận tương quan của các biến liên quan đến thói quen tập luyện

Kết quả từ Hình 4.14 trên cho thấy có mối tương quan tuyến tính thuận, mạnh giữa thời gian tập luyện và tần suất tập luyện (hệ số tương quan = 0.76), cho thấy khi tần suất tăng thì thời gian mỗi buổi tập cũng tăng. Hệ số tương quan 0.84 giữa thời gian tập luyện và mức độ kinh nghiệm cho thấy những người có kinh nghiệm cao thường dành nhiều thời gian hơn cho việc tập luyện. Tương tự, hệ số tương quan 0.76 giữa tần suất và kinh nghiệm chỉ ra rằng người có kinh nghiệm thường tập luyện thường xuyên hơn. Các hệ số tương quan đều từ 0.64 đến 1.0, cho thấy mối tương quan chặt chẽ giữa các yếu tố này.

Để hiểu rõ hơn về thói quen tập luyện, việc đánh giá sự khác biệt về thời lượng buổi tập của người tập giữa các trình độ kinh nghiệm khác nhau là rất cần thiết, nhằm xác định liệu mức độ kinh nghiệm có ảnh hưởng đến thời gian mà mỗi người dành cho việc tập luyện hay không.

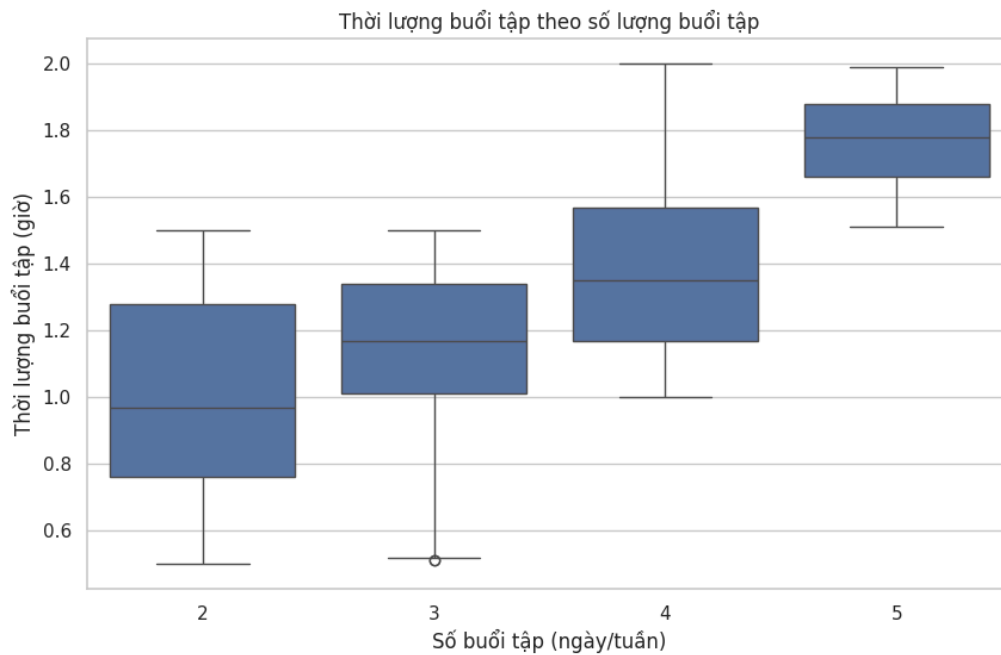


*Hình 4.15: Phân bố thời gian tập luyện theo mức độ kinh nghiệm*

Nhóm có mức độ kinh nghiệm thấp nhất có thời gian tập luyện trung bình từ 1 đến 1.4 giờ với median khoảng 1 giờ và có sự phân bố rộng nhất, cho thấy có sự đa dạng lớn trong thời gian tập luyện của họ. Nhóm kinh nghiệm với Experience\_Level = 2 có thời gian từ 1 đến khoảng 1.6 giờ. Nhóm có kinh nghiệm cao nhất có thời gian tập luyện từ 1.6 đến 2 giờ cao hơn đáng kể so với hai nhóm còn lại. Nhìn chung, thời gian tập luyện tăng dần theo mức độ kinh nghiệm, với nhóm có kinh nghiệm cao hơn thường dành nhiều thời gian và ổn định hơn trong hành vi tập luyện.

Biểu đồ trên cho thấy thời gian tập luyện phân bố theo mức độ kinh nghiệm và gợi ý rằng mức độ tham gia tập luyện có thể ảnh hưởng đến các chỉ số sức khỏe khác như lượng mỡ, lượng calories đốt cháy,... Điều này đặt ra câu hỏi liệu tăng tần suất tập luyện có tác động đến các yếu tố này hay không.

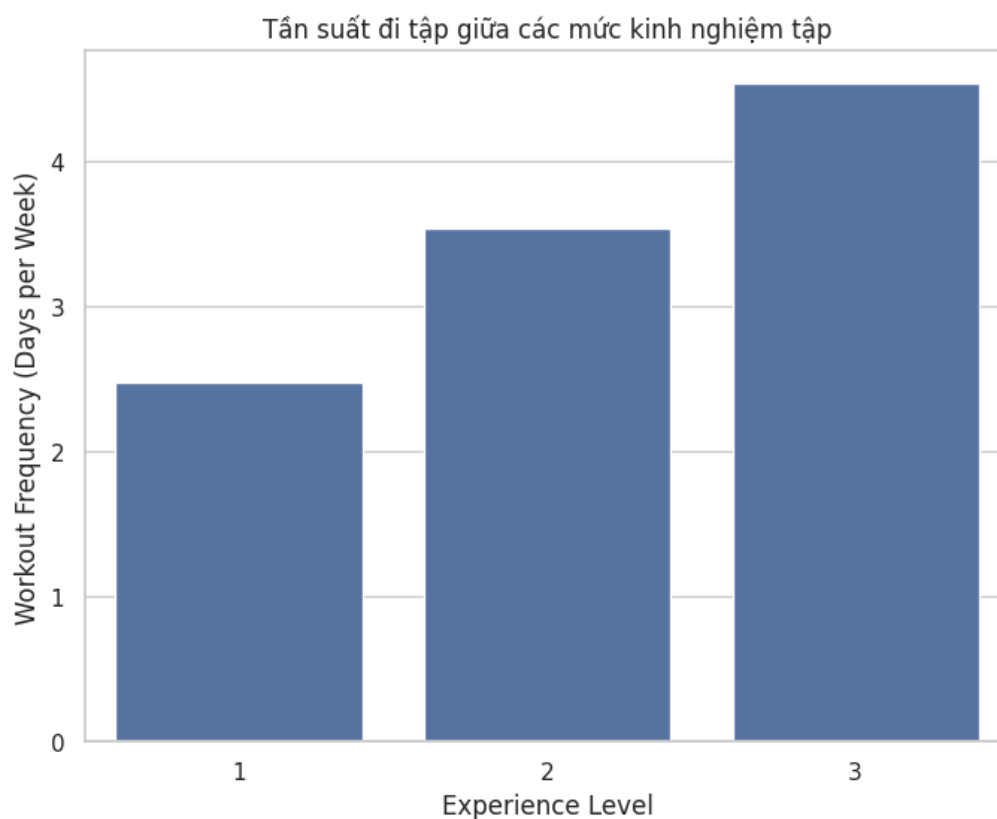
Nhóm sẽ đánh giá sự khác biệt về thời lượng buổi tập của người tập với số lượng buổi tập:



*Hình 4.16: Thời lượng buổi tập theo số lượng buổi tập*

Dựa vào biểu đồ Hình 4.16, ta cũng nhận thấy rằng, nhóm người tập luyện từ 4-5 buổi trở lên cũng có xu hướng về thời gian tập mỗi buổi tập lâu hơn. Các nhóm còn lại có số buổi tập ít hơn nhưng số lượng lại chiếm khá nhiều đi kèm với đó thời lượng mỗi buổi tập ngắn hơn khá nhiều so với nhóm 4-5.

Nhóm tiếp tục tìm hiểu xem từ tần suất tập và thời lượng tập như vậy thì những nhóm người có kinh nghiệm như thế nào sẽ có những thói quen tập luyện gì.



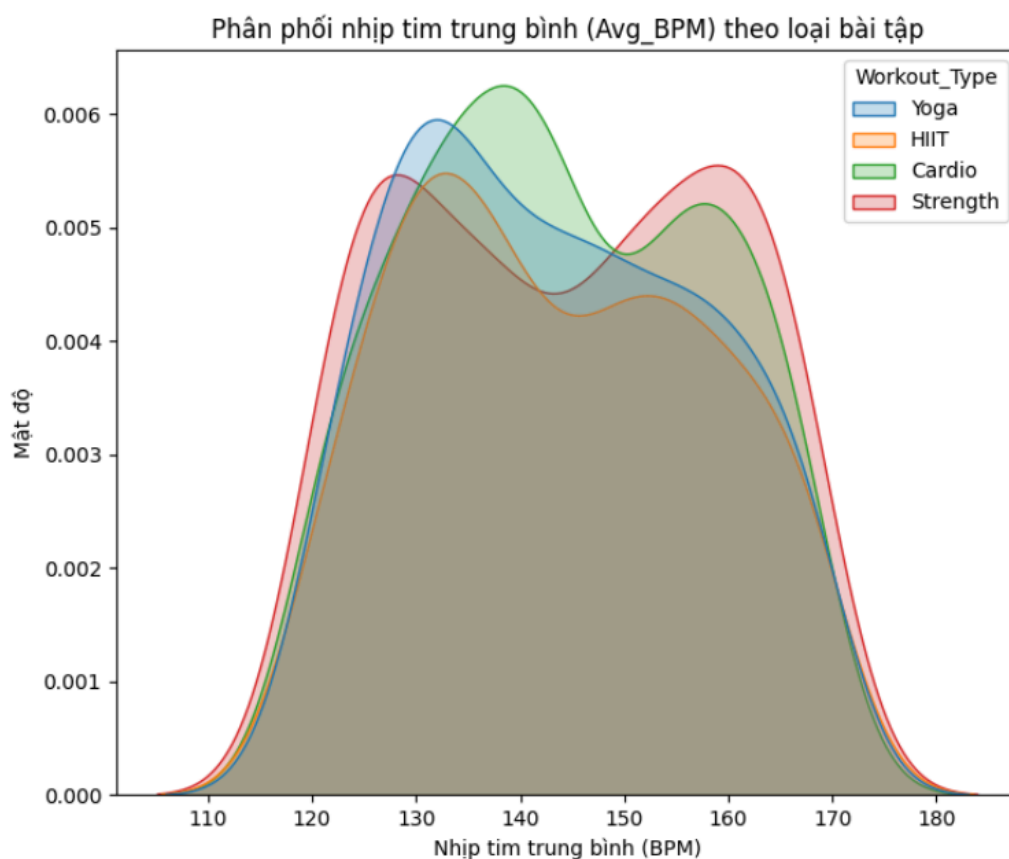
*Hình 4.17: Tần suất đi tập giữa các mức kinh nghiệm tập*

Biểu đồ Hình 4.17 cho thấy có sự chênh lệch giữa các nhóm người có kinh nghiệm khi nhắc đến tần suất đi tập luyện. Người có kinh nghiệm luyện tập càng cao thì tần suất đi tập luyện cũng nhiều hơn. Liên hệ với biểu đồ Hình 4.16, như vậy ta có thể hiểu rằng những người có kinh nghiệm cao, họ có tần suất tập luyện từ 4-5 buổi và mỗi buổi tập của họ có thời lượng từ 1.5-2 giờ mỗi buổi.

#### **4.2.3. Phân tích quá trình tập luyện của người tập**

Để đánh giá hiệu suất quá trình tập luyện của người tập, nhóm thực hiện đánh giá qua các tiêu chí sau,

Một, về nhịp tim của người tập,

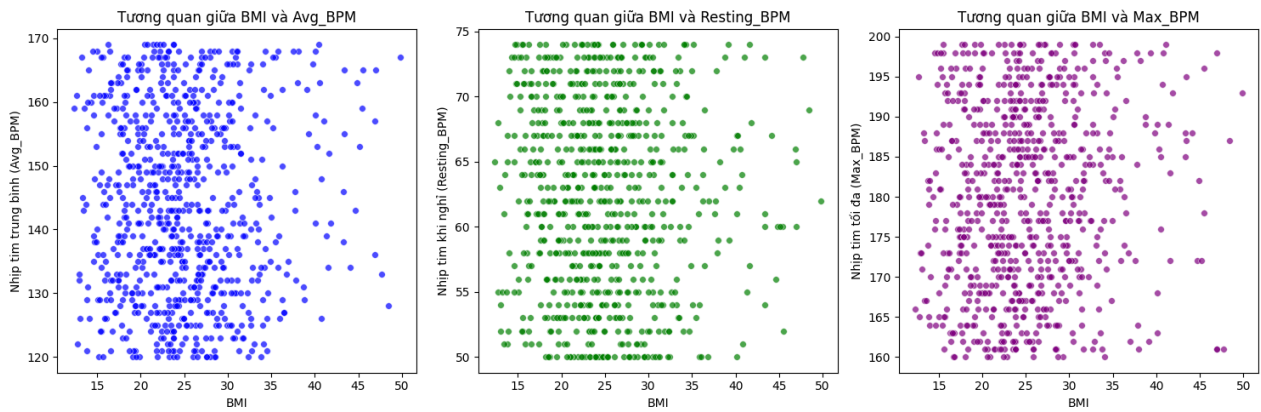


*Hình 4.18: Phân phối nhịp tim trung bình theo loại bài tập*

Khi phân tích nhịp tim của người tập qua các dạng bài tập, ta nhận thấy rằng Strength là bài tập có phân phối trải rộng nhất và cũng có sự chênh lệch dần về 2 đuôi. Điều này có thể do các bài tập Strength thiên về sự kết hợp chung giữa các nhóm cơ khi phát lực sẽ làm nhịp tim tăng mạnh, tuy nhiên quãng nghỉ giữa các bài tập Strength thường dài hơn. Các bài tập Cardio, HIIT và Yoga khác thường sẽ tập liên tục các bài không dừng nghỉ trong suốt khoảng thời gian, do đó nhịp tim sẽ duy trì ở ngưỡng trung bình đến cao, ngoài ra ta cũng thấy rằng dáng điệu đồ thị phân phối các bài tập này có sự tương đồng nhất định.

Về nhịp tim theo thể trạng của người tập,





Hình 4.19: Biểu đồ tương quan giữa BMI và các chỉ số về nhịp tim

Biểu đồ tương quan giữa BMI và các chỉ số về nhịp tim đã mô tả sự tương quan giữa chỉ số khối cơ thể (BMI) và 3 thuộc tính về nhịp tim lần lượt là nhịp tim trung bình (Avg\_BPM), nhịp tim khi nghỉ ngơi (Resting\_BPM) và nhịp tim cao nhất ghi nhận (Max\_BPM). Thông qua việc quan sát biểu đồ ta có thể thấy rằng các điểm trên biểu đồ có sự phân tán, ta tiến hành kiểm định hệ số tương quan để xác định xem liệu có tồn tại mối liên hệ có ý nghĩa thống kê nào giữa các biến này hay không.

Kiểm định giả thuyết về sự tương quan giữa chỉ số khối cơ thể (BMI) so với các biến liên quan đến nhịp tim, sử dụng kiểm định hệ số tương quan Pearson, ta thực hiện kiểm định cho từng cặp biến với các giả thuyết như sau:

*Thiết lập cặp giả thuyết với chỉ số khối cơ thể (BMI) và nhịp tim cao nhất ghi nhận (Max\_BPM). Chọn mức ý nghĩa là 5%.*

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa chỉ số khối cơ thể (BMI) và nhịp tim cao nhất ghi nhận (Max\_BPM).

$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq$  0: Có mối tương quan tuyến tính giữa chỉ số khối cơ thể (BMI) và nhịp tim cao nhất ghi nhận (Max\_BPM).

Hệ số tương quan Pearson giữa Max\_BPM và BMI: 0.0671  
 Giá trị p: 0.0364  
 Có sự tương quan có ý nghĩa giữa Max\_BPM và BMI.

*Hình 4.20: Kết quả Kiểm định Pearson giữa Max\_BPM và BMI*

Nhóm thu được kết quả với  $p\text{-value} = 0.0364$ . Vì  $p\text{-value} = 0.0364 < 0.05$  nên ta bác bỏ giả thuyết  $H_0$  và chấp nhận giả thuyết  $H_1$ . Vì vậy ta kết luận rằng có sự tương quan có ý nghĩa thống kê giữa Max\_BPM và BMI. Tuy nhiên, hệ số tương quan  $\rho = 0.0671$  khá nhỏ, cho thấy được mối tương quan này yếu.

*Thiết lập cặp giả thuyết với chỉ số khối cơ thể (BMI) và nhịp tim lúc nghỉ (Resting\_BPM) với mức ý nghĩa là 5%.*

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa chỉ số khối cơ thể (BMI) và nhịp tim lúc nghỉ (Resting\_BPM).

$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq 0$ : Có mối tương quan tuyến tính giữa chỉ số khối cơ thể (BMI) và nhịp tim lúc nghỉ (Resting\_BPM).

Hệ số tương quan Pearson giữa Resting_BPM và BMI: -0.0325 Giá trị p: 0.3106 Không có sự tương quan có ý nghĩa giữa Resting_BPM và BMI.
--

*Hình 4.21: Kết quả Kiểm định Pearson giữa Resting\_BPM và BMI*

Kết quả kiểm định cho thấy  $p\text{-value} = 0.3106$ . Vì  $p\text{-value} = 0.3106 > 0.05$  nên ta không bác bỏ giả thuyết  $H_0$ . Như vậy, không có đủ bằng chứng để kết luận rằng có mối tương quan tuyến tính giữa Resting\_BPM và BMI.

*Thiết lập cặp giả thuyết với chỉ số khối cơ thể (BMI) và nhịp tim trung bình (Avg\_BPM) với mức ý nghĩa là 5%.*

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa chỉ số khối cơ thể (BMI) và nhịp tim trung bình (Avg\_BPM).

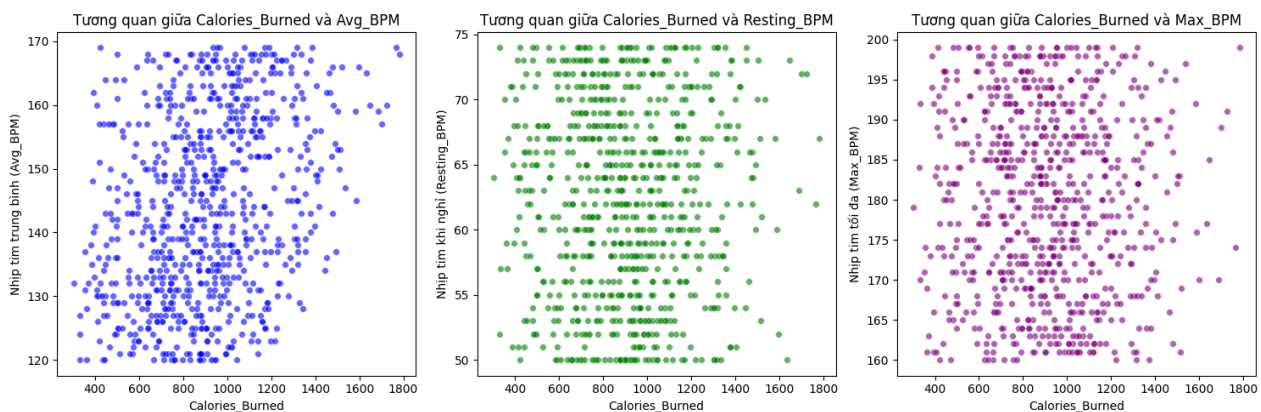
$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq 0$ : Có mối tương quan tuyến tính giữa chỉ số khối cơ thể (BMI) và nhịp tim trung bình (Avg\_BPM).

Hệ số tương quan Pearson giữa Avg_BPM và BMI: 0.0216 Giá trị p: 0.5009 Không có sự tương quan có ý nghĩa giữa Avg_BPM và BMI.
---

Hình 4.22: Kết quả Kiểm định Pearson giữa Avg\_BPM và BMI

Ta thu được kết quả kiểm định với hệ số tương quan Pearson giữa Avg\_BPM và BMI là  $\rho = 0.0216$  với  $p\text{-value} = 0.5009$ . Vì  $p\text{-value} = 0.5009 > 0.05$  nên ta không bác bỏ giả thuyết  $H_0$ . Do đó, không có đủ bằng chứng để kết luận có mối tương quan tuyến tính giữa Avg\_BPM và BMI.

Hai, phân tích sự tương quan giữa lượng Calories đốt và nhịp tim khi tập luyện,



Hình 4.23: Tương quan giữa lượng Calories đốt và các biến liên quan đến nhịp tim

Để đánh giá mối quan hệ giữa lượng Calories đốt cháy và các chỉ số nhịp tim (nhịp tim trung bình Avg\_BPM, nhịp tim khi nghỉ ngơi Resting\_BPM, và nhịp tim cao nhất ghi nhận Max\_BPM), chúng ta đã phân tích biểu đồ mô tả sự tương quan giữa chúng. Nhận thấy sự phân tán của các điểm trên biểu đồ, ta tiến hành kiểm định hệ số tương quan để xác định xem liệu có tồn tại mối liên hệ có ý nghĩa thống kê nào giữa các biến này hay không.

Kiểm định giả thuyết về sự tương quan giữa lượng Calories đốt (Calories\_Burned) so với các biến liên quan đến nhịp tim, sử dụng kiểm định hệ số tương quan Pearson, ta thực hiện kiểm định cho từng cặp biến với các giả thuyết như sau:

*Thiết lập cặp giả thuyết với lượng Calories đốt (Calories\_Burned) và nhịp tim cao nhất ghi nhận (Max\_BPM) với mức ý nghĩa 5%.*

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa lượng Calories đốt cháy (Calories\_Burned) và nhịp tim cao nhất ghi nhận (Max\_BPM).

$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq$  0: Có mối tương quan tuyến tính giữa lượng

Calories đốt cháy (Calories\_Burned) và nhịp tim cao nhất ghi nhận (Max\_BPM).

Hệ số tương quan Pearson giữa Max_BPM và Calories_Burned: 0.0021 Giá trị p: 0.9481 Không có sự tương quan có ý nghĩa giữa Max_BPM và Calories_Burned.
---

*Hình 4.24: Kết quả Kiểm định Pearson giữa Max\_BPM và Calories\_Burned*

Nhóm đạt được kết quả như sau, với p-value = 0.9481. Vì p-value = 0.9481 > 0.05 nên ta không bác bỏ giả thuyết  $H_0$ . Do đó, không có đủ bằng chứng để kết luận có mối tương quan tuyến tính giữa Max\_BPM và Calories\_Burned.

*Thiết lập cặp giả thuyết với lượng Calories đốt cháy (Calories\_Burned) và nhịp tim lúc nghỉ (Resting\_BPM) với mức ý nghĩa là 5%.*

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa lượng Calories đốt cháy (Calories\_Burned) và nhịp tim lúc nghỉ (Resting\_BPM).

$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq$  0: Có mối tương quan tuyến tính giữa lượng Calories đốt cháy (Calories\_Burned) và nhịp tim lúc nghỉ (Resting\_BPM).

Hệ số tương quan Pearson giữa Resting_BPM và Calories_Burned: 0.0165 Giá trị p: 0.6068 Không có sự tương quan có ý nghĩa giữa Resting_BPM và Calories_Burned.
---

*Hình 4.25: Kết quả Kiểm định Pearson giữa Resting\_BPM và Calories\_Burned*

Kết quả kiểm định cho thấy rằng hệ số tương quan Pearson giữa Resting\_BPM và Calories\_Burned là  $\rho = 0.0165$  với p-value = 0.6068. Vì p-value = 0.6068 > 0.05 nên ta không bác bỏ giả thuyết  $H_0$ . Do vậy, không có đủ bằng chứng để kết luận rằng có mối tương quan tuyến tính giữa Resting\_BPM và Calories\_Burned.

*Thiết lập cặp giả thuyết với lượng Calories đốt cháy (Calories\_Burned) và nhịp tim trung bình (Avg\_BPM) với mức ý nghĩa là 5%.*

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa lượng Calories đốt cháy (Calories\_Burned) và nhịp tim trung bình (Avg\_BPM).

$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq$  0: Có mối tương quan tuyến tính giữa lượng

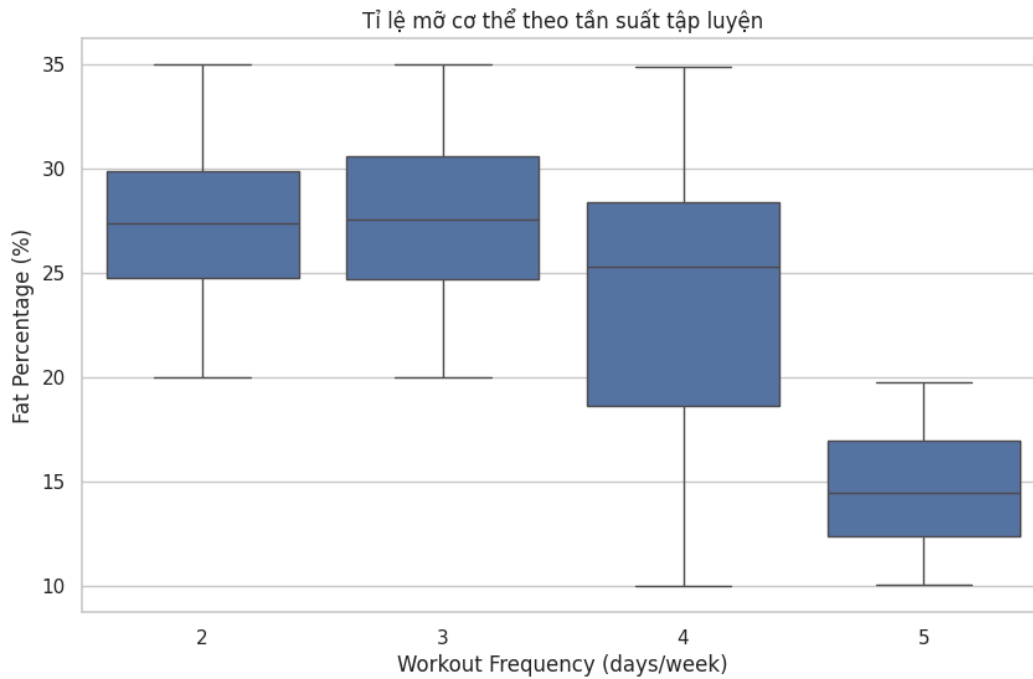
Calories đốt cháy (Calories\_Burned) và nhịp tim trung bình (Avg\_BPM).

Hệ số tương quan Pearson giữa Avg_BPM và Calories_Burned: 0.3397 Giá trị p: 0.0000 Có sự tương quan có ý nghĩa giữa Avg_BPM và Calories_Burned.
---

*Hình 4.26: Kết quả Kiểm định Pearson giữa Avg\_BPM và Calories\_Burned*

Kiểm định trả về p-value = 0.0000. Vì p-value = 0.0000 < 0.05 nên ta bác bỏ giả thuyết  $H_0$  và chấp nhận giả thuyết  $H_1$ . Do đó, ta kết luận rằng có sự tương quan có ý nghĩa thống kê giữa Avg\_BPM và Calories\_Burned.

Để giải thích tại sao Avg\_BPM lại có sự tương quan với Calories\_Burned như vậy, đầu tiên ta sẽ nói đến thuộc tính Resting\_BPM, vì là nhịp tim ghi nhận lúc nghỉ ngơi nên đương nhiên quá trình này sẽ không tiêu thụ Calories. Đến với Max\_BPM, nhịp tim tối đa được ghi nhận không phản ánh toàn bộ quá trình tập luyện của người tập, do đó ta cũng hiểu được tại sao 2 biến này lại không có tương quan với Calories\_Burned. Còn biến Avg\_BPM, nhịp tim trung bình phản ánh toàn bộ các dao động về chỉ số nhịp tim trong toàn thời lượng buổi tập, kể cả chỉ số khi nghỉ ngơi và nhịp tim tối đa cũng được phản ánh trong Avg\_BPM, do đó Avg\_BPM cũng phản ánh tốt sự tương quan với lượng Calories đốt trong buổi tập.



Hình 4.27: Tỉ lệ mỡ cơ thể theo tần suất tập luyện

Nhóm người thường xuyên tập luyện chăm chỉ từ 4 đến 5 buổi 1 tuần duy trì được lượng mỡ thấp so với các nhóm còn lại, nhóm sẽ tiến hành tìm hiểu xem liệu lượng mỡ trung bình giữa các nhóm có thời gian tập từ hai đến năm ngày trong một tuần có khác nhau hay không, ta có cặp giả thuyết sau,

$$H_0: \mu_{Fat\ Percentage, 2} = \mu_{Fat\ Percentage, 3} = \mu_{Fat\ Percentage, 4} = \mu_{Fat\ Percentage, 5}$$

$H_1$ : Trung bình Fat Percentage có sự khác biệt đáng kể giữa ít nhất hai nhóm Workout Frequency. Tức là tần suất tập luyện có ảnh hưởng đến lượng mỡ cơ thể.

Trước tiên, ta cần kiểm tra các giả thuyết về phân phối chuẩn và phương sai thuần nhất của Fat Percentage giữa các nhóm,

```

Nhóm 4: p-value = 0.0000
Dữ liệu của nhóm 4 (days/week) không tuân theo phân phối chuẩn.
Nhóm 3: p-value = 0.0000
Dữ liệu của nhóm 3 (days/week) không tuân theo phân phối chuẩn.
Nhóm 5: p-value = 0.0011
Dữ liệu của nhóm 5 (days/week) không tuân theo phân phối chuẩn.
Nhóm 2: p-value = 0.0011
Dữ liệu của nhóm 2 (days/week) không tuân theo phân phối chuẩn.
-----
Kiểm định Levene: p-value = 0.0000
Phương sai giữa các nhóm không đồng nhất.
  
```

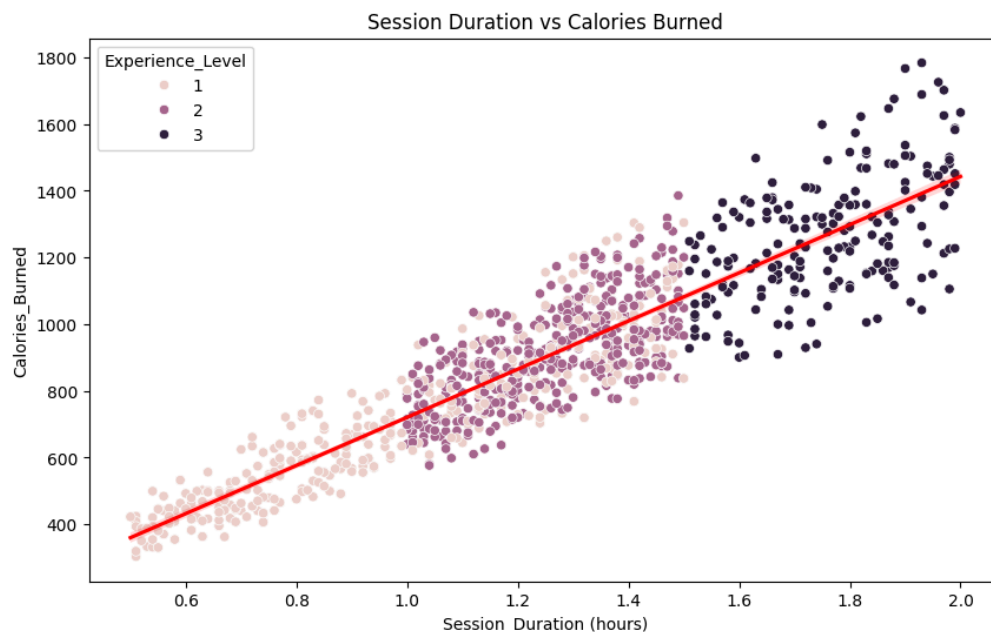
Hình 4.28: Kết quả Kiểm định Shapiro-Wilk và Levene về phân phối và phương sai Fat Percentage giữa các nhóm

Dựa vào kết quả kiểm định, các p-value đều nhỏ hơn giá trị  $\alpha = 0.05$ , khi đó ta không thể áp dụng Kiểm định ANOVA ở giả thuyết này. Thay vào đó sẽ sử dụng Kiểm định phi tham số Kruskal-Wallis, ta thu được p-value =  $0.000 < 0.05$ . Kết quả kiểm định cho thấy ta có đủ bằng chứng để bác bỏ  $H_0$ . Do đó, có sự khác biệt đáng kể về lượng mỡ cơ thể giữa các nhóm người có tần suất tập luyện khác nhau.

Kruskal-Wallis H Test statistic: 282.66, p-value: 0.0000  
Có sự khác biệt đáng kể về Fat Percentage giữa các nhóm Workout Frequency.

Hình 4.29: Kết quả Kiểm định Kruskal-Wallis về sự khác biệt về lượng % mỡ cơ thể giữa các nhóm người có tần suất tập luyện khác nhau

Ba, lượng Calories đốt cháy trong quá trình tập luyện:



Hình 4.30: Tương quan giữa thời gian tập luyện và lượng calo đốt cháy

Ta dường như thấy sự tương quan giữa thời lượng tập luyện và lượng calo đốt cháy. Đồng thời, có sự phân hóa rõ ràng giữa những người tập có kinh nghiệm so với nhóm còn lại. Để củng cố cho lập luận này, nhóm sẽ tiến hành kiểm định mối tương quan giữa thời lượng tập luyện đến lượng calo đốt cháy của buổi tập, cặp giả thuyết  $H_0$  và  $H_1$  được phát biểu

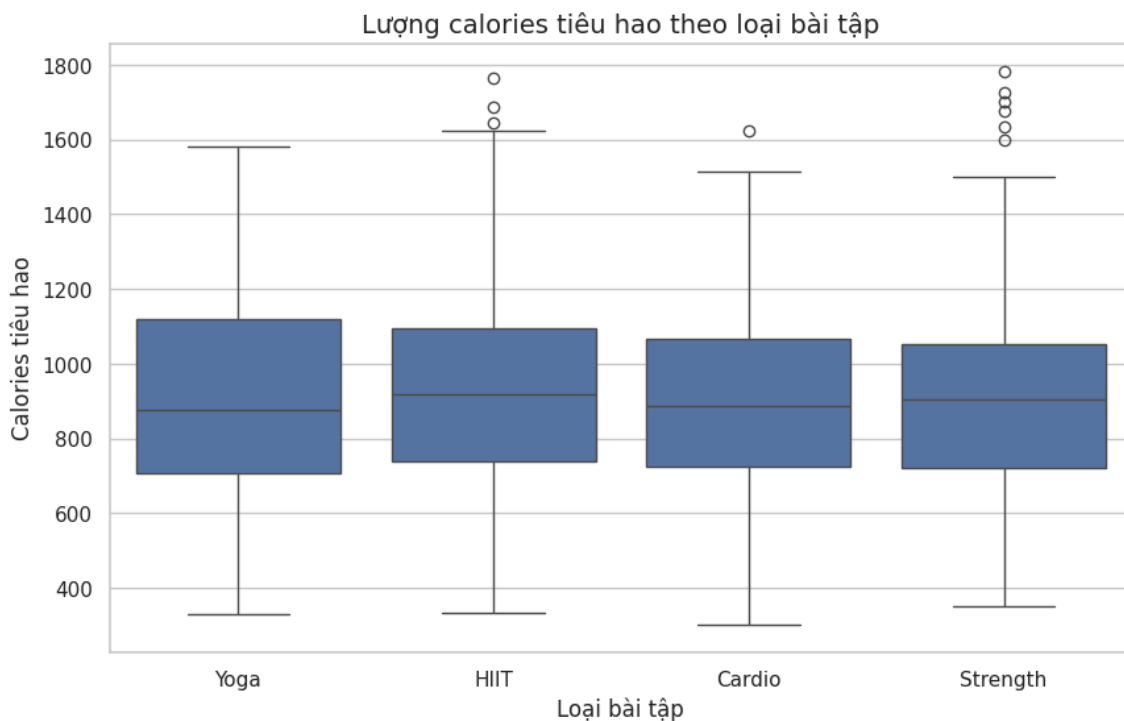


như sau,

$H_0$ : Hệ số tương quan Pearson ( $\rho$ ) = 0: Không có mối tương quan tuyến tính giữa thời lượng tập luyện và lượng calo đốt cháy.

$H_1$ : Hệ số tương quan Pearson ( $\rho$ )  $\neq$  0: Có mối tương quan tuyến tính giữa thời lượng tập luyện và lượng calo đốt cháy.

Nhóm thu được kết quả  $p\text{-value} = 0.0000 < 0.05$ , do vậy ta sẽ bác bỏ  $H_0$ . Như vậy, có mối tương quan tuyến tính đáng kể giữa thời lượng tập luyện đến lượng calo đốt cháy của buổi tập



Hình 4.31: Biểu đồ hộp thể hiện lượng Calories tiêu hao theo các loại bài tập

Qua biểu đồ Hình..., là biểu đồ boxplot thể hiện sự phân phối giữa lượng Calories tiêu hao khi so sánh các loại bài tập khác nhau. Ta có thể nhận xét rằng giữa các bài tập thì không có sự khác biệt về lượng Calories đốt cháy. Để có đủ bằng chứng thống kê hơn, nhóm thiết lập cặp giả thuyết để kiểm định như sau,

$H_0$ : Trung bình lượng calories tiêu hao giữa các loại bài tập là giống nhau.

$H_1$ : Có ít nhất một loại bài tập có lượng calories tiêu hao trung bình khác biệt.



Trước tiên, kiểm tra các điều kiện về phân phối chuẩn và phương sai thuần nhất khi xây dựng giả thuyết này.

```
Nhóm Yoga: p-value = 0.0112
Dữ liệu của nhóm Yoga không tuân theo phân phối chuẩn.
Nhóm HIIT: p-value = 0.0375
Dữ liệu của nhóm HIIT không tuân theo phân phối chuẩn.
Nhóm Cardio: p-value = 0.0511
Dữ liệu của nhóm Cardio tuân theo phân phối chuẩn.
Nhóm Strength: p-value = 0.0012
Dữ liệu của nhóm Strength không tuân theo phân phối chuẩn.
-----
Kiểm định Levene: p-value = 0.7890
Phương sai giữa các nhóm đồng nhất.
```

Hình 4.32: Kết quả kiểm định Shapiro-Wilk và Levene về lượng Calories giữa các loại bài tập

Kết quả kiểm định Shapiro-Wilk cho thấy, chỉ có dữ liệu Calories của nhóm bài tập Cardio là có phân phối chuẩn. Ta tiếp tục kiểm định về phương sai thuần nhất giữa các nhóm. Kết quả kiểm định Leneve cho thấy phương sai giữa các nhóm đồng nhất.

Vì các điều kiện cần của giả thuyết này chỉ thỏa mãn phương sai thuần nhất, không đủ để sử dụng kiểm định ANOVA, nhóm chọn sử dụng Kiểm định phi tham số Kruskal-Wallis để thay thế:

```
Kruskal-Wallis: p-value = 0.6860
Không có sự khác biệt đáng kể về Calories Burned giữa các nhóm Workout_Type.
```

Hình 4.33: Kết quả Kiểm định Kruskal-Wallis về lượng Calories giữa các loại bài tập.

Kết quả Kiểm định Kruskal-Wallis cho thấy ta không có đủ bằng chứng để bác bỏ  $H_0$ , do đó lượng Calories tiêu hao giữa các loại bài tập là giống nhau.

#### 4.4. Khuyến nghị và đề xuất

Từ các thông tin có được từ các phân tích biểu diễn ở trên, ta thu được một số kết luận dưới đây:

Thứ nhất, về sức khỏe và thể trạng của người tập. Chỉ số BMI đo lường thể trạng giữa các

nhóm tuổi cho thấy, chiếm phần lớn là những người có mức BMI Bình thường, tuy nhiên đáng lưu tâm là nhóm Thừa cân và Béo Phì là các nhóm chiếm cao thứ 2 và 3 sau đó mới đến nhóm Thiếu cân. Chỉ số lượng phần trăm mỡ trong cơ thể tuy không phản ánh đúng cân nặng của người tập, nhưng lại có hiệu quả khi dùng để đánh giá về hiệu quả tập luyện của người tập. Điều này cho thấy nhóm này cần các chương trình hỗ trợ thúc đẩy động lực tập luyện.

Thứ hai, về thói quen tập luyện của người tập. Nhóm tập luyện từ 4-5 buổi/tuần có thời gian tập trung bình dài hơn, điều này cho thấy họ có động lực cao hơn, đồng thời tỷ lệ mỡ cơ thể thấp hơn phần nào thể hiện nhận thức của họ về sức khỏe của chính mình. Ngược lại, nhóm dưới 3 buổi/tuần có thời gian buổi tập ngắn, điều này có thể phản ánh sự thiếu tập trung hoặc thời gian eo hẹp, không thể duy trì đều đặn.

Thứ ba, về quá trình tập luyện của người tập. Việc lựa chọn loại bài tập không quá ảnh hưởng đến mục tiêu sức khỏe của người tập khi chúng có lượng Calories đốt là như nhau, tuy nhiên cần chú ý thêm các tiêu chí phụ như tăng cơ sẽ tập các bài về Strength, tập Cardio và HIIT đều là các dạng bài tập giúp tăng sức bền và đẩy nhịp tim lên cao, các bài tập Yoga thiên về kéo giãn thân người và các cơ xương khớp. Lượng Calories đốt cháy có tương quan chặt chẽ với thời lượng buổi tập, có thể hiểu rằng khi thời lượng luyện tập càng lâu, cơ thể tăng dần nhịp tim và sử dụng nhiều năng lượng hơn, hiệu suất buổi tập được tăng lên và đốt cháy được nhiều Calories hơn. Ngược lại với những buổi tập có thời lượng ngắn, khi đó cơ thể chưa đủ thời gian để thích ứng thì đã dừng lại nên hiệu quả tập luyện cũng kém đi.

Từ những kết luận này, nhóm đúc kết được các khuyến nghị về việc xây dựng kế hoạch tập luyện cho một số nhóm người,

Một, nếu là nhóm người trung niên bận rộn công việc hoặc gia đình, ta cần xây dựng kế hoạch tập phù hợp như kết hợp tập các nhóm cơ có cùng chức năng vào một buổi và phân bổ các bài tập phù hợp để họ có hiệu suất tập luyện tốt hơn kể cả khi số buổi tập dưới trung bình.

Hai, với nhóm người trung niên có sức khỏe bắt đầu có dấu hiệu giảm sút, ta cần ưu tiên xây dựng lộ trình phù hợp để cải thiện sức khỏe tim mạch, cơ xương khớp, giảm mỡ...

như các bài tập Yoga, Cardio.

Ba, nhóm người có chỉ số BMI thấp Thiếu cân cần tập trung về dinh dưỡng và xây dựng cơ bắp bằng cách tăng cường tập luyện các bài Strength để kích hoạt sự phát triển cơ bắp toàn diện, tuy nhiên dinh dưỡng cần chú ý về lượng mỡ hấp thu vì nhóm này rất dễ tăng cơ bắp nhưng theo đó là lượng mỡ cũng tăng theo rất nhanh.

Bốn, nhóm người Thừa cân và Béo phì nên được theo dõi về dinh dưỡng sát sao, đồng thời khi xây dựng chế độ tập luyện cho nhóm này phải kết hợp được các bài tập tăng cơ và giảm mỡ để cơ thể được cân đối hơn, giảm các nguy cơ mắc bệnh tim mạch, mỡ máu...

Đây là những khuyến nghị dựa trên tập dữ liệu đã phân tích, khi sử dụng các khuyến nghị này các Huấn luyện viên và trung tâm thể hình cần phải cân nhắc thể trạng thực tế của người tập để xây dựng kế hoạch luyện tập và chế độ dinh dưỡng phù hợp.

## **CHƯƠNG V: KẾT LUẬN**

### **5.1. Thảo luận**

Qua báo cáo trên, nhóm đã thành công phân tích và trực quan hóa dữ liệu đồng thời xây dựng và kiểm định các giả thuyết về các đặc trưng sức khỏe, thể trạng, thói quen và hiệu suất tập luyện của những người tập gym, kết quả của bài báo cáo sẽ giúp các Trung tâm Thể hình và các Huấn luyện viên xây dựng được kế hoạch luyện tập phù hợp với người tập và giúp họ có được kết quả phù hợp với mục tiêu phát triển bản thân, tăng cường sức khỏe.

Tuy nhiên, bài báo cáo vẫn còn nhiều hạn chế. Thứ nhất, bộ dữ liệu hiện tại bao gồm thông tin từ 973 người, kích cỡ mẫu vẫn chưa đủ lớn nên bị giới hạn về khả năng khái quát hóa cho các nhóm đối tượng. Thứ hai, về các thuộc tính của dữ liệu, tập dữ liệu thiếu các thuộc tính liên quan đến chế độ dinh dưỡng của người tập hoặc thói quen sinh hoạt hàng ngày như giấc ngủ, dẫn đến việc phân tích các yếu tố liên quan đến thể trạng của người tập vẫn còn chưa được tổng quát. Thứ ba, nghiên cứu này chưa đi sâu vào việc phân tích tác động dài hạn (ảnh hưởng của thời gian) của các thói quen tập luyện đến sức khỏe hoặc thành tích thể chất của người tập.

### **5.2. Hướng phát triển**

Dựa vào những điểm hạn chế trên, nhóm mong muốn nghiên cứu trong tương lai có thể được mở rộng quy mô dữ liệu bằng cách thu thập thêm thông tin từ nhiều đối tượng và khu vực hơn, tăng tính chính xác của kết quả phân tích hơn. Ngoài ra, việc tích hợp thêm các chỉ số về dinh dưỡng, giấc ngủ, hoặc tâm lý sẽ giúp phân tích toàn diện hơn về hành vi và sức khỏe của người tập gym. Cuối cùng, thêm yếu tố thời gian trong việc theo dõi người tập trong dài hạn và ngắn hạn sẽ giúp đánh giá rõ hơn tác động của các thói quen tập luyện đến sức khỏe, từ đó đưa ra những khuyến nghị chính xác và có giá trị ứng dụng cao hơn.

## TÀI LIỆU THAM KHẢO

- [1] Wilke, C. O. (n.d.). *Fundamentals of Data Visualization*. Claus O. Wilke. Retrieved December 7, 2024, from <https://clauswilke.com/dataviz/>