BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn : Xử lý số liệu thống kê

Nhóm : 21

Danh sách thành viên :

22280042 : Đinh Xuân Khang

22280044 : Bành Đức Khánh

22280057 : Nguyễn Hồ Nam

22280059 : Võ Duy Nghĩa

22280069 : Phạm Tấn Phước

Project 3:

Hiện nay các phong trào tập thể thao đang ngày một phát triển, thu hút nhiều nhóm tuổi và giới tính. Dữ liệu bodyPerformance.csv chứa thông tin của 13,393 người tham gia tập thể thao tại Hàn Quốc, với 12 biến như sau:

- age- độ tuổi (từ 20 tới 64);
- gender- giới tính (F: nữ, M: nam);
- height_cm- chiều cao (đơn vị: cm);
- weight_kg- cân nặng (đơn vị: kg);
- body fat_%- phần trăm mỡ cơ thể (%);
- diastolic- huyết áp tâm trương (phút);
- systolic- huyết áp tâm thu (phút);
- gripForce- lực kẹp;
- sit and bend forward_cm- ngồi và gập người về phía trước;
- sit-ups counts- số lần gập bụng;
- broad jump_cm- nhảy xa (đơn vị: cm);

- class- phân lớp hiệu suất (A: tốt nhất, B,C,D).

Hãy xử lý dữ liệu này để giúp cho các chuyên gia sức khỏe biết được hiệu quả của việc tập thể dục, và các yếu tố ảnh hưởng tới hiệu quả.

Khám phá và phân tích dữ liệu.

1. Tiền xử lý dữ liệu (Data Preprocessing):
   - Tên cột: làm sạch tên cột, đưa về đúng format như "height_cm"
   - Kiểm tra dữ liệu khuyết: dữ liệu bodyPerformance.csv không có dữ liệu khuyết.
   - Huyết áp tâm trương (diastolic) và tâm thu (systolic) không thể bằng 0 và tâm trương phải nhỏ hơn tâm thu
   - Kiểm tra kiểu dữ liệu: xác định, biến đổi kiểu dữ liệu cho đúng theo từng cột. Trong dữ liệu bodyPerformance có 2 cột là kiểu factor (định tính) là "gender" và "class". Còn lại là kiểu định lượng.
   - Kiểm tra dữ liệu nhóm định tính: xem phân bổ dữ liệu nhóm định tính có bị mất cân bằng hay không?
     - gender: F: 4926, M: 8467. Dữ liệu có tỷ lệ 6.3 nam : 3.7 nữ nên chưa ở mức bị mất cân bằng nghiêm trọng.
     - class: A: 3348, B: 3347, C:3349, D: 3349. Tỷ lệ giữa các nhóm gần như cân bằng.
   - Tạo thêm biến mới: thêm cột dữ liệu mới từ các cột dữ liệu cũ từ dữ liệu bodyPerformance.csv.
     - bmi: chỉ số khối của cơ thể được tính từ biến "weight_kg" (cân nặng) và "height_cm" (chiều cao) bằng công thức:

$$\frac{Cân\ nặng\ (kg)}{(Chiều\ cao\ (m))^2}$$

     - bmi_category: phân loại bmi theo Tổ chức y tế thế giới (WHO):
       - Dưới 18.5: Gầy (Underweight)
       - 18.5 - 24.9: Bình thường (Normal)

        - 25 - 29.9: Thừa cân (Overweight)
        - 30 - 34.9: Béo phì (Obese)
    - fitness_score: Điểm sức khoẻ được tính từ biến "sit_ups_counts" (số lần gập bụng), "grip_force" (lực kẹp) và "broad_jump_cm" (nhảy bật xa):

fitness_score = sit_ups_counts*0.4 + grip_force*0.3 + broad_jump_cm*0.3

    - pulse_pressure: áp lực mạch đập được tính từ hiệu của "diastolic" (huyết áp tâm trương) và "systolic" (huyết áp tâm thu):

pulse_pressure = diastolic – systolic

    - blood_pressure_category: Phân loại theo chỉ số huyết áp từ "diastolic" và "systolic":
        - systolic < 90 & diastolic < 60: Hypotension (Huyết áp thấp)
        - systolic < 120 & diastolic < 80: Normal (Huyết áp bình thường)
        - 120 <= systolic < 130 & diastolic < 80: Elevated (Huyết áp cao)
        - 130 <= systolic < 140 hoặc 80 <= diastolic < 90: Hypertension Stage 1 (Huyết áp cao giai đoạn 1)
        - systolic >= 140 hoặc diastolic >= 90: Hypertension Stage 2 (Huyết áp cao giai đoạn 2)
        - systolic > 180 hoặc diastolic > 120: Hypertensive Crisis (Huyết áp ở mức nguy hiểm)

- Xử lý dữ liệu: xử lý các giá trị không hợp lệ theo từng cột (xuất hiện trong bảng tổng hợp trên) để tránh các giá trị outlier.
    - "diastolic" và "systolic": 2 chỉ số này không thể bằng 0 trừ trường hợp máy đo bị lỗi và diastolic luôn nhỏ hơn systolic.
    - grip_force: lực kẹp của người không thể bằng 0.
    - broad_jum_cm: con người không thể nhảy bật xa bằng 0.

2. Tóm tắt dữ liệu.

- Bảng tóm tắt chung :

	bien	gtnn	gtln	tv	tb	dlc
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	age	21	64	32	36.8	13.6
2	height_cm	125	194.	169.	169.	8.43
3	weight_kg	26.3	138.	67.4	67.4	11.9
4	body_fat	3	78.4	22.8	23.2	7.25
5	diastolic	6	126	79	78.8	10.7
6	systolic	77	201	130	130.	14.6
7	grip_force	1.6	70.5	37.9	37.0	10.6
8	sit_and_bend_forward_cm	-25	213	16.2	15.2	8.46
9	sit_ups_counts	0	80	41	39.8	14.3
10	broad_jump_cm	20	303	193	190.	39.5
11	bmi	11.1	42.9	23.5	23.6	2.94
12	fitness_score	11.4	133.	85.8	84.1	19.0
13	pulse_pressure	6	139	51	51.5	10.8

- Nhận thấy ở biến sit_and_bend_forward_cm có gtln là 213 cm , điều này khá vô lý vì theo thông tin ta tìm được trên trang www.ptdirect.com thì chỉ số tốt nhất được ghi nhận chỉ tầm ở mức trên 46,5cm và dưới 50cm . Nên ta quyết định lọc bỏ các giá trị trên 50 cm của biến này .

- Xem lại dữ liệu tóm tắt

	bien	gtnn	gtln	tv	tb	dlc
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	age	21	64	32	36.8	13.6
2	height_cm	125	194.	169.	169.	8.43
3	weight_kg	26.3	138.	67.4	67.4	11.9
4	body_fat	3	78.4	22.8	23.2	7.25
5	diastolic	6	126	79	78.8	10.7
6	systolic	77	201	130	130.	14.6
7	grip_force	1.6	70.5	37.9	37.0	10.6
8	sit_and_bend_forward_cm	-25	42	16.2	15.2	8.15
9	sit_ups_counts	0	80	41	39.8	14.3
10	broad_jump_cm	20	303	193	190.	39.5
11	bmi	11.1	42.9	23.5	23.6	2.94
12	fitness_score	11.4	133.	85.8	84.1	19.0
13	pulse_pressure	6	139	51	51.5	10.8

- Tóm tắt dữ liệu theo nhóm phân loại gender

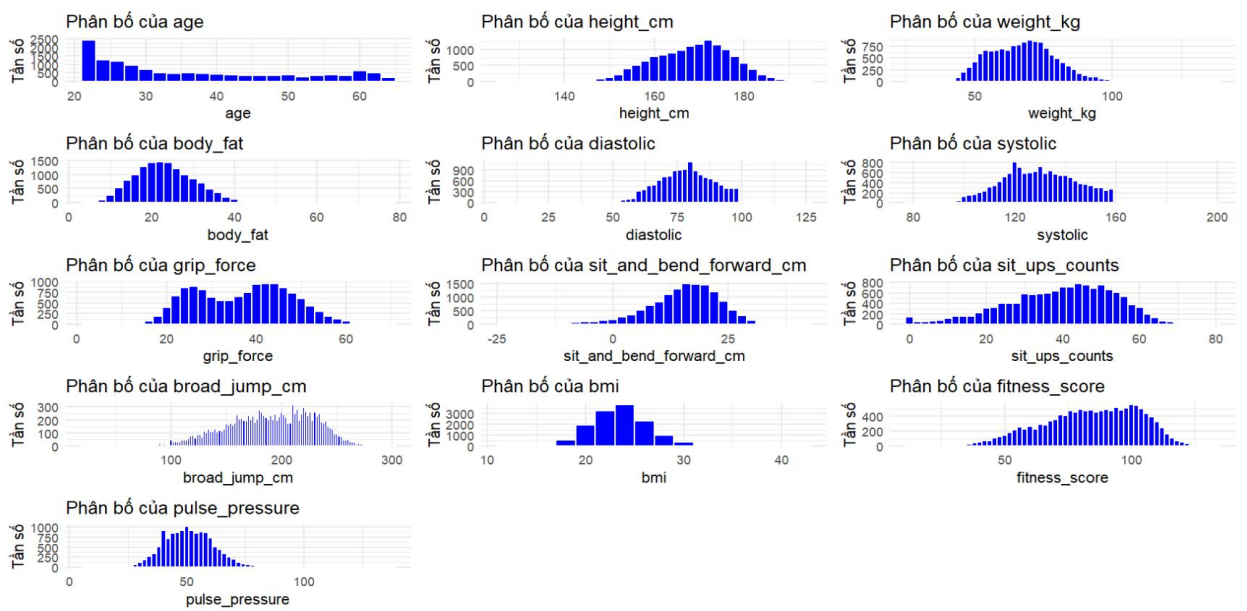
```
# A tibble: 2 × 23
  gender mean_height sd_height mean_weight sd_weight mean_body_fat sd_body_fat
  <fct>        <dbl>     <dbl>       <dbl>     <dbl>          <dbl>       <dbl>
1 F             160.      5.65        56.9      7.63           28.5        6.22
2 M             173.      5.81        73.6      9.47           20.2        5.95
# i 16 more variables: mean_diastolic <dbl>, sd_diastolic <dbl>, mean_systolic <dbl>,
#   sd_systolic <dbl>, mean_grip_force <dbl>, sd_grip_force <dbl>,
#   mean_sit_and_bend_forward <dbl>, sd_sit_and_bend_forward <dbl>, mean_sit_ups <dbl>,
#   sd_sit_ups <dbl>, mean_broad_jump <dbl>, sd_broad_jump <dbl>, mean_bmi <dbl>,
```

-Tóm tắt dữ liệu theo nhóm phân loại class

```
# A tibble: 4 × 23
  class mean_height sd_height mean_weight sd_weight mean_body_fat sd_body_fat mean_diastolic
  <fct>        <dbl>     <dbl>       <dbl>     <dbl>          <dbl>       <dbl>          <dbl>
1 A             168.      7.84        64.4      10.6           20.5        6.44           77.9
2 B             169.      8.13        66.6      10.9           22.0        6.65           78.7
3 C             169.      8.52        66.8      10.9           22.6        6.27           78.5
4 D             169.      9.11        72.0      13.9           27.7        7.50           80.1
# i 15 more variables: sd_diastolic <dbl>, mean_systolic <dbl>, sd_systolic <dbl>,
#   mean_grip_force <dbl>, sd_grip_force <dbl>, mean_sit_and_bend_forward <dbl>,
#   sd_sit_and_bend_forward <dbl>, mean_sit_ups <dbl>, sd_sit_ups <dbl>,
```

3. Vẽ biểu đồ & kiểm định giả thuyết.

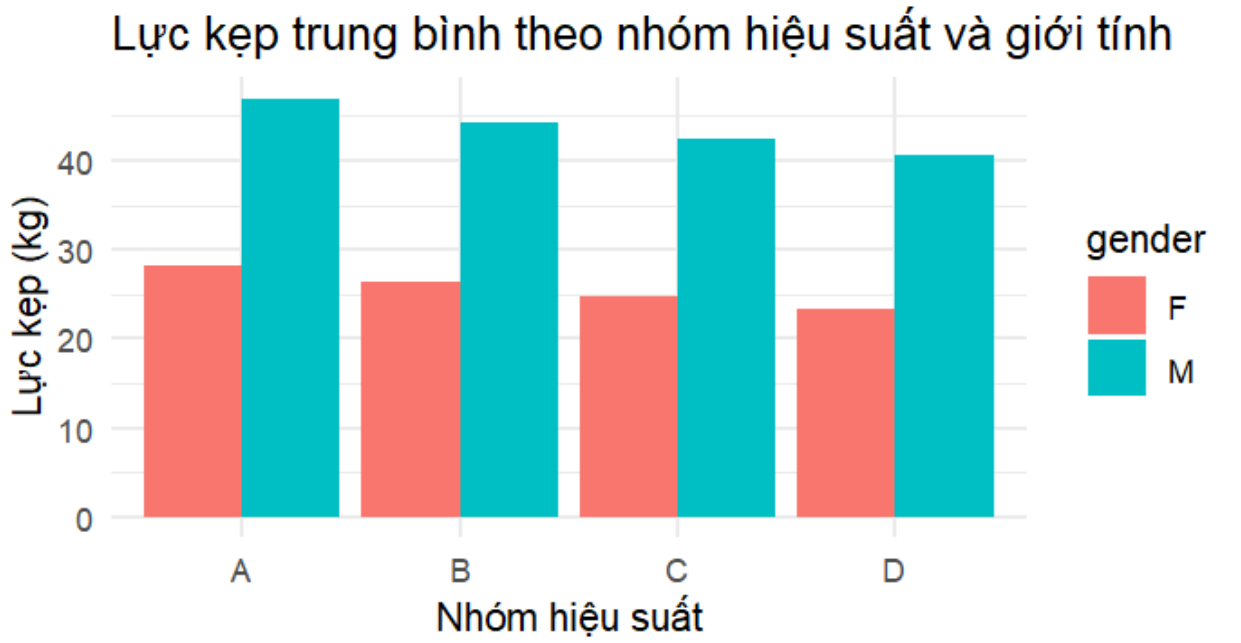
- Biểu đồ phân phối dữ liệu của các biến



3.1. Ở phần kế tiếp sẽ là các biểu đồ thể hiện những góc nhìn khác nhau về bộ dữ liệu , ở cuối mỗi biểu đồ sẽ là lời nhận xét về xu hướng của biểu đồ , sau đó là phần kiểm định A/B testings để đảm bảo phần nhận xét có mang ý nghĩa thống kê , từ đó chúng có thể trở

nên có ích cho các chuyên gia sức khỏe . Để kiểm định thì ta áp dụng permutation test với 2 nhóm và hàm aovp() với nhiều nhóm và một số phương pháp khác .

- Biểu đồ thứ 1



Nhận xét : Chỉ số lực kẹp ở nam giới nhìn chung tốt hơn nữ giới . Chỉ số này tăng đều theo phân lớp hiệu quả (class)

Để kiểm tra xem nhận xét mang ý nghĩa thống kê hay chỉ là kết quả của sự ngẫu nhiên , ta thực hiện các kiểm định sau :

H0: Không có sự khác biệt đáng kể về lực kẹp trung bình giữa nam và nữ

H1: Lực kẹp trung bình ở nữ yếu hơn nam

-> P-value = 0

Với mức ý nghĩa alpha = 0.05 , p-value < alpha nên chấp nhận H1 : Lực kẹp trung bình ở nữ yếu hơn nam

Kế tiếp ta có

H0: Không có sự khác biệt đáng kể về lực kẹp trung bình giữa các nhóm hiệu suất.

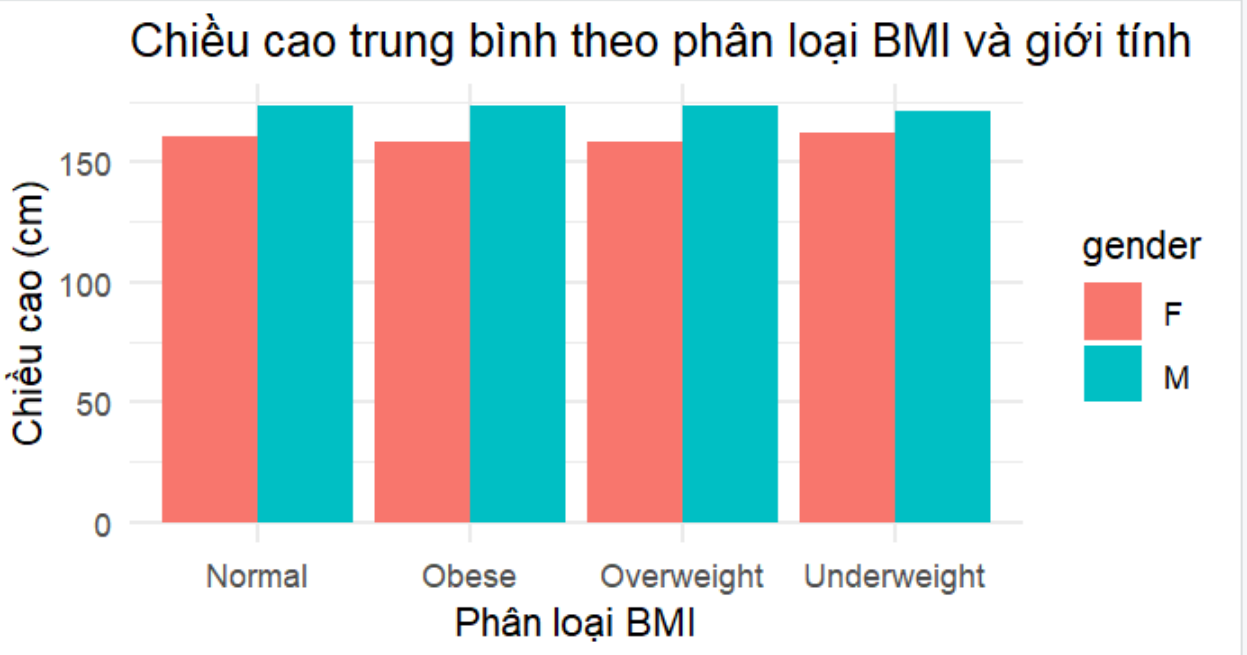
H1: Có ít nhất một sự khác biệt đáng kể về lực kẹp trung bình giữa các nhóm hiệu suất.

-> p-value = 2.2e-16

Với mức ý nghĩa alpha = 0.05 , p-val < alpha nên ta chấp nhận H1 : Có ít nhất một sự khác biệt đáng kể về lực kẹp trung bình giữa các nhóm hiệu suất.

Vậy kết luận nhận xét ban đầu là chính xác và mang ý nghĩa thống kê

-Biểu đồ thứ 2



Nhận xét : Chiều cao trung bình ở nam giới trong cả 4 phân nhóm BMI là gần như ngang nhau và tốt hơn nữ giới , trong đó nhóm Underweight có chỉ số chiều cao kém hơn một chút . Ở nữ giới thì 2 nhóm Obese và Overweight thì lại có chiều cao trung bình thấp hơn so với 2 nhóm Normal và Underweight

Để kiểm tra xem nhận xét mang ý nghĩa thống kê hay chỉ là kết quả của sự ngẫu nhiên , ta thực hiện các kiểm định sau :

H0: Không có sự khác biệt đáng kể về chiều cao trung bình giữa nam và nữ

H1: chiều cao trung bình ở nữ thấp hơn nam

->P-value = 0

Với mức ý nghĩa alpha = 0.05 , p-value < alpha nên chấp nhận H1 : chiều cao trung bình ở nữ thấp hơn nam

Kế tiếp ta có :

H0: Không có sự khác biệt về chiều cao trung bình giữa các nhóm BMI ở nam giới.

H1: Có ít nhất một nhóm BMI có chiều cao trung bình khác biệt ở nam giới.

P-val = 0.8924

Với mức ý nghĩa alpha = 0.05 ,p-val > alpha nên chấp nhận H0 : Không có sự khác biệt về chiều cao trung bình giữa các nhóm BMI ở nam giới.

Kế tiếp ta có :

H0: Không có sự khác biệt về chiều cao trung bình giữa nhóm {Obese, Overweight} và nhóm {Normal, Underweight} ở nữ giới.

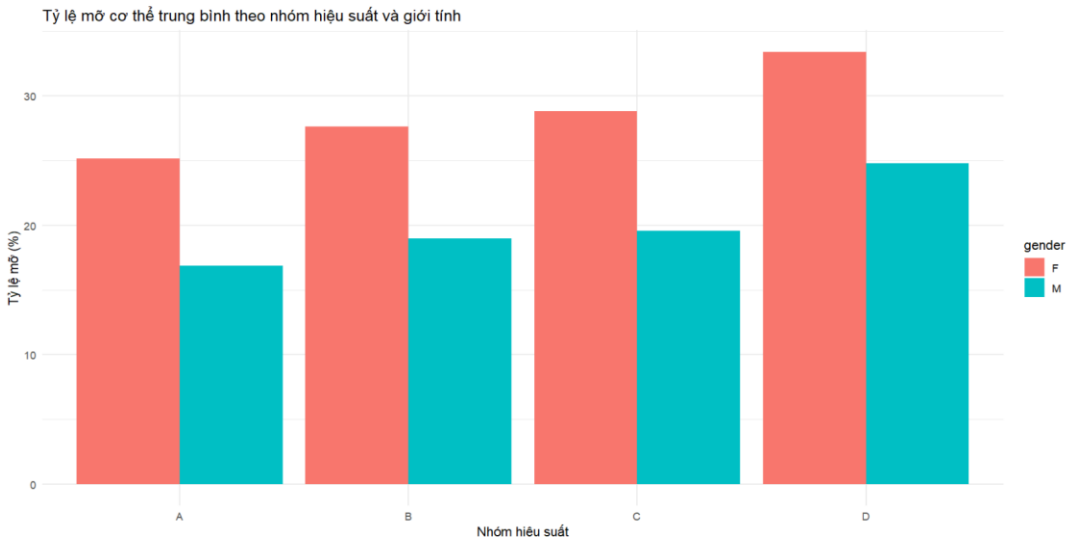
H1: Chiều cao trung bình giữa nhóm {Obese, Overweight} thấp hơn nhóm {Normal, Underweight}ở nữ giới.

p-val = 2.2e-16

Với mức ý nghĩa alpha =0.05 , p-val < alpha nên ta chấp nhận H1: Chiều cao trung bình giữa nhóm {Obese, Overweight} thấp hơn nhóm {Normal, Underweight} ở nữ giới.

Vậy kết luận nhận xét ban đầu là chính xác và mang ý nghĩa thống kê

- Biểu đồ thứ 3



Nhận xét : % Mỡ (body-fat) thấp nhất ở class A , sau đó tăng dần đều (ở cả 2 giới) tới class D là cao nhất . % mỡ ở nam giới có xu hướng thấp hơn nữ giới

Để kiểm tra xem nhận xét mang ý nghĩa thống kê hay chỉ là kết quả của sự ngẫu nhiên , ta thực hiện các kiểm định:

H0: Không có sự khác biệt về chiều tỉ lệ mỡ trung bình ở nam giới và nữ giới.

H1: Tỉ lệ mỡ trung bình ở nam giới thấp hơn nữ giới.

P-value = 0

Với mức ý nghĩa alpha = 0.05 , p-value < alpha nên chấp nhận H1 : Tỉ lệ mỡ cơ thể trung bình ở nam giới thấp hơn nữ giới

Kế tiếp ta có :

H0: Không có sự khác biệt về tỉ lệ mỡ cơ thể trung bình ở các nhóm hiệu suất.

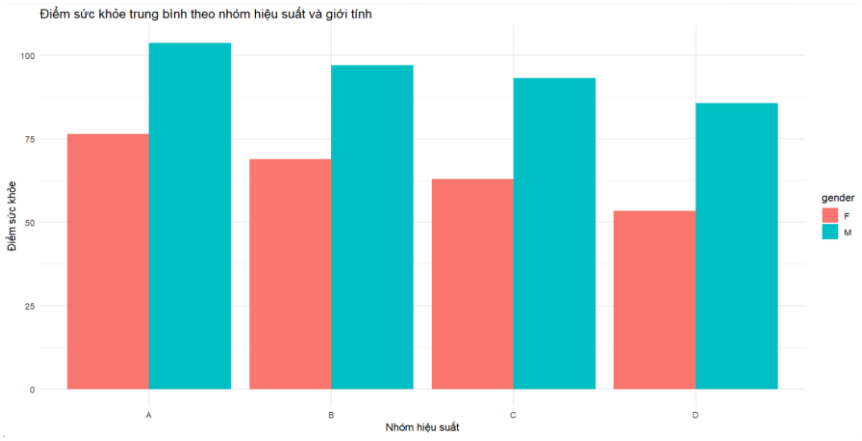
H1: Có ít nhất một nhóm hiệu suất có thỉ lệ mỡ cơ thể trung bình khác các nhóm còn lại.

p-val = 2.2e-16

Với mức ý nghĩa alpha = 0.05 , p-value < alpha nên chấp nhận H1: Có ít nhất một nhóm hiệu suất có tỉ lệ mỡ cơ thể trung bình khác các nhóm còn lại.

Vậy kết luận nhận xét ban đầu là chính xác và mang ý nghĩa thống kê

- Biểu đồ thứ 4



Nhận xét : Chỉ số điểm fitness score tốt nhất ở class A , sau đó giảm dần (ở cả 2 giới).Fitness score ở nam giới cao hơn nữ giới

Để kiểm tra xem nhận xét mang ý nghĩa thống kê hay chỉ là kết quả của sự ngẫu nhiên , ta thực hiện các kiểm định:

H0: Không có sự khác biệt về fitness score trung bình ở nam giới và nữ giới.

H1: fitness score trung bình ở nữ giới thấp hơn nam giới.

P-value = 0

Với mức ý nghĩa alpha = 0.05 , p-value < alpha nên chấp nhận H1 : fitness score trung bình ở nữ giới thấp hơn nam giới

Kế tiếp ta có :

H0: Không có sự khác biệt về fitness score trung bình ở các nhóm hiệu suất.

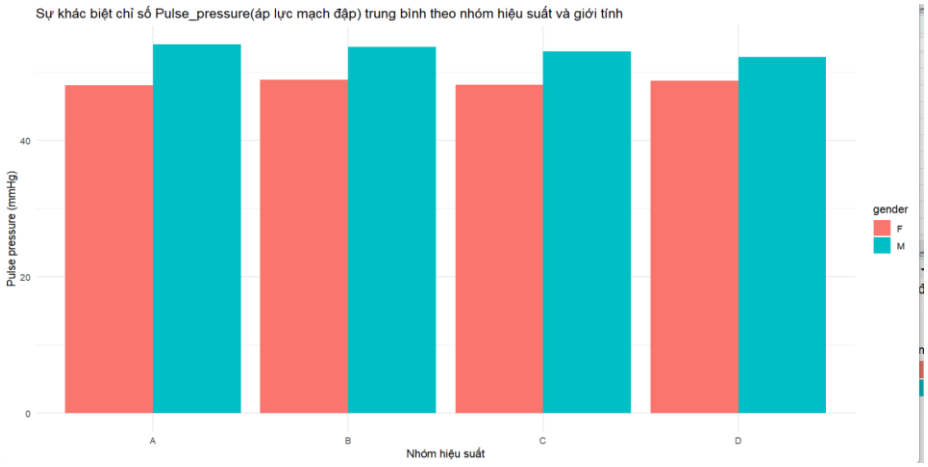
H1: Có ít nhất một nhóm hiệu suất có fitness score trung bình khác các nhóm còn lại.

p-val = 2.2e-16

Với mức ý nghĩa alpha = 0.05 , p-value < alpha nên chấp nhận H1: Có ít nhất một nhóm hiệu suất có fitness score trung bình khác các nhóm còn lại.

Vậy kết luận nhận xét ban đầu là chính xác và mang ý nghĩa thống kê

- Biểu đồ thứ 5



Nhận xét : Chỉ số pulse pressure trung bình ở nam giới giảm dần từ class A về class D và cao hơn nữ giới . Trong khi ở nữ giới lại không có xu hướng tăng giảm rõ ràng với class C thấp nhất , B cao nhất , A và D gần ngang nhau

Để kiểm tra nhận xét trên có mang ý nghĩa thống kê không , ta thực hiện các kiểm định:

H0 : Không có sự khác biệt về pulse pressure trung bình ở 2 giới tính

H1 : Pulse pressure trung bình ở nữ giới thấp hơn nam giới#p-val = 0

Với mức ý nghĩa alpha =0.05 , p-val < alpha nên chấp nhận H1

Kế tiếp ta có :

H0 : Không có sự khác biệt về pulse pressure trung bình giữa các nhóm hiệu suất của nam giới

H1 : Có ít nhất 1 sự khác biệt về pulse pressure trung bình giữa các nhóm hiệu suất của nam giới

P-val = 2.05e-08

Với mức ý nghĩa alpha = 0.05 , p-val < alpha nên chấp nhận H1

Tương tự đối với nhóm nữ :

H0 : Không có sự khác biệt về pulse pressure trung bình giữa các nhóm hiệu suất của nữ giới

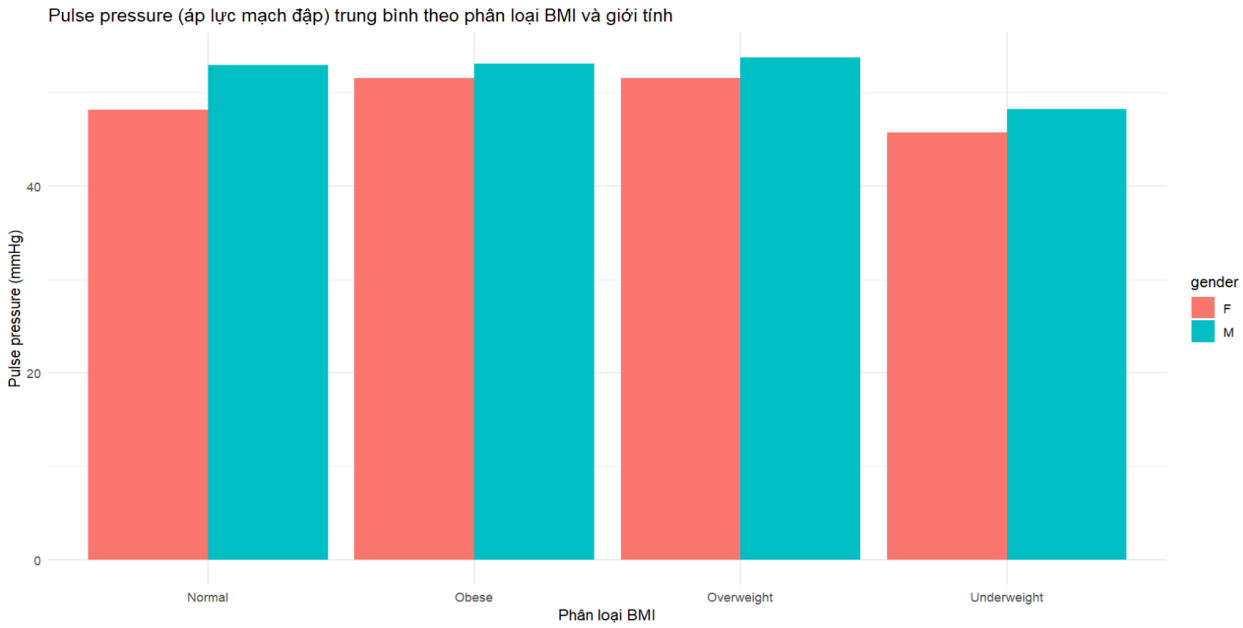
H1 : Có ít nhất 1 sự khác biệt về pulse pressure trung bình giữa các nhóm hiệu suất của nữ giới

p-val = 0.107

Với mức ý nghĩa alpha =0.05 , p-val >alpha nên chấp nhận H0

Vậy nhận xét ban đầu chưa chính xác và chỉ mang tính ngẫu nhiên vì thực tế , không có sự khác biệt về pulse pressure trung bình giữa các nhóm hiệu suất của nữ giới

- Biểu đồ thứ 6



Nhận xét : Pulse pressure trung bình ở nam giới của 3 nhóm Normal , Obese , Overweight gần như ngang nhau , của Underweight là thấp nhất, cả 4 nhóm đều cao hơn nữ giới

Ở nữ giới thì chỉ số lại khá cao trong 2 nhóm Obese và Overweight , nhóm Normal thấp hơn đôi chút và nhóm Underweight là thấp nhất

Để kiểm tra nhận xét trên , ta thực hiện các kiểm định sau :

Ta đã kiểm đinh được Pulse pressure ở nữ giới thấp hơn nam giới trong câu trước , nên giờ ta tiếp tục

H0 : Không có sự khác biệt pulse pressure trung bình ở các nhóm BMI ở nam giới

H1 : có ít nhất 1 sự khác biệt pulse pressure trung bình ở các nhóm BMI ở nam giới

P-val = 8.96e-05

Với mức ý nghĩa alpha = 0.05 , p-val < alpha nên chấp nhận H1

Kế tiếp :

H0 : Không có sự khác biệt pulse pressure trung bình ở các nhóm BMI ở nữ giới

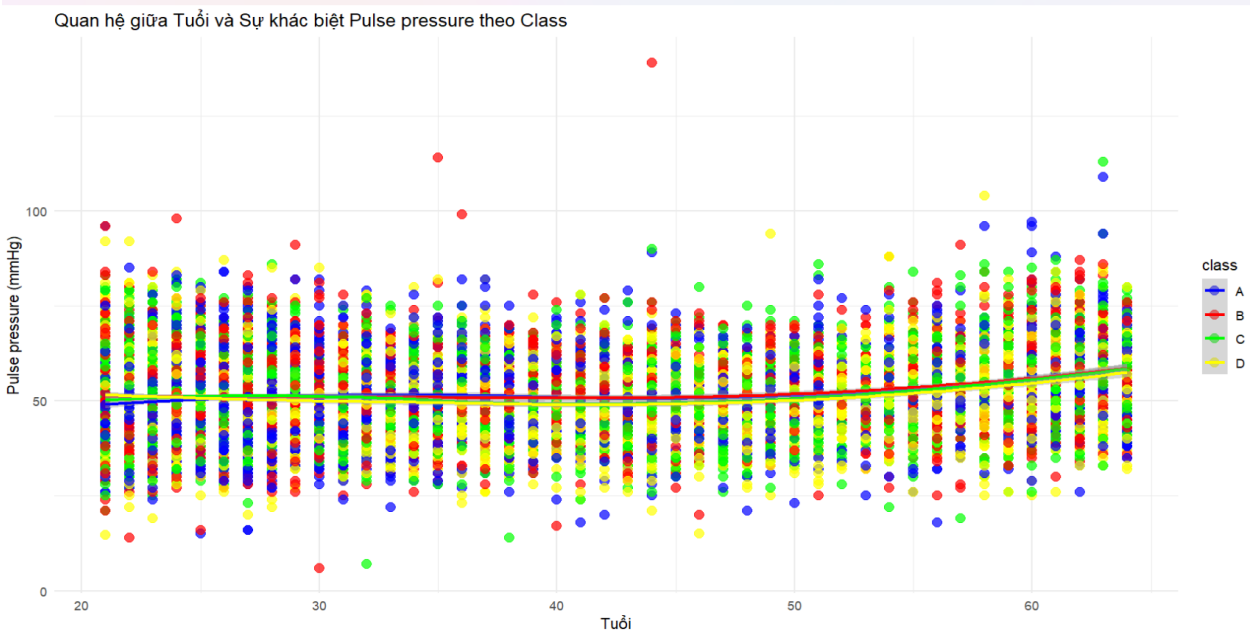
H1 : có ít nhất 1 sự khác biệt pulse pressure trung bình ở các nhóm BMI ở nữ giới

p-val = 2e-16

Với mức ý nghĩa alpha = 0.05 , p-val < alpha nên chấp nhận H1

Vậy nhận xét ban đầu có thể xem là mang ý nghĩa thống kê

- Biểu đồ thứ 7



Nhận xét : dựa vào đường cong thể hiện xu hướng tăng giảm của các chấm dữ liệu , ta thấy chỉ số áp lực mạch đập có xu hướng tăng lên khi tuổi tác càng cao

Để kiểm định nhận xét trên , ta thực hiện :

H0: Không có mối quan hệ tuyến tính giữa tuổi tác và áp lực mạch đập

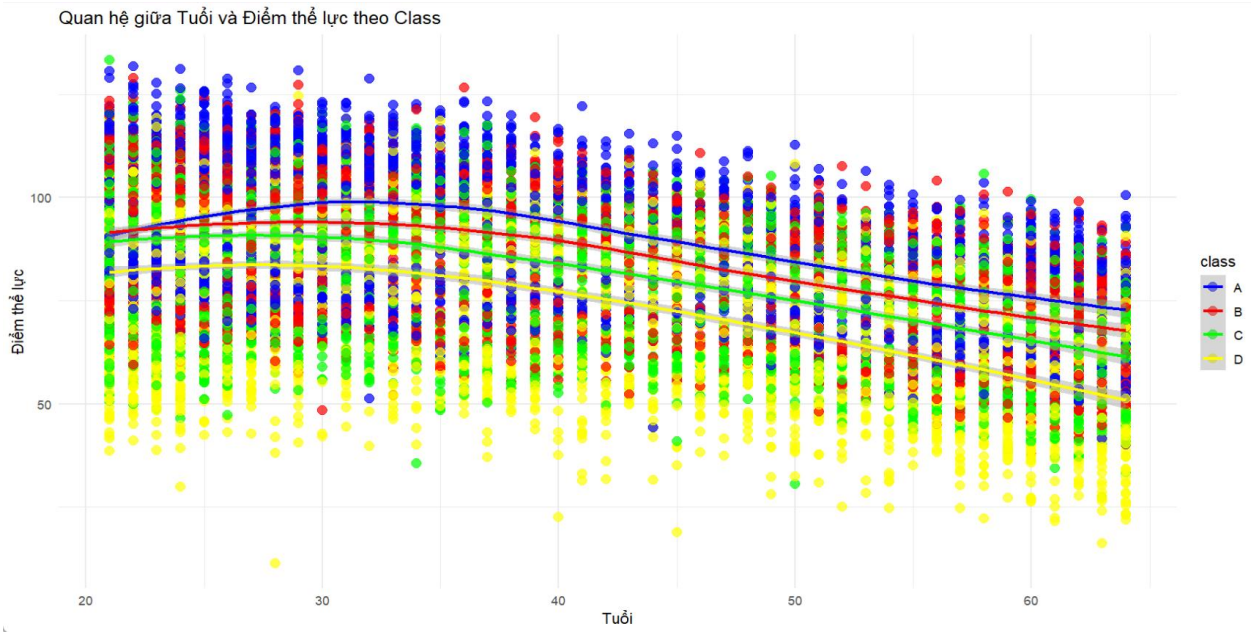
H1: Có mối quan hệ tuyến tính giữa tuổi tác và áp lực mạch đập

p-val < 2e-16

Với mức ý nghĩa alpha = 0.05 , p-val < alpha nên ta chấp nhận H1 : Có mối quan hệ tuyến tính giữa tuổi tác và áp lực mạch đập , nghĩa là tuổi tác tăng thì áp lực mạch đập cũng tăng .

Ta kết luận nhận xét ban đầu là chính xác và mang ý nghĩa thống kê

- Biểu đồ thứ 8



Nhận xét : dựa vào các đường cong thể hiện xu hướng tăng giảm của dữ liệu , ta thấy ở cả 4 class thì chỉ số fitness score cao nhất ở giai đoạn khoảng 25 tới 35 tuổi . Từ 40 tuổi trở đi thì chỉ số bắt đầu giảm dần.

Để kiểm định nhận xét trên , ta thực hiện chia ra các nhóm tuổi : "Dưới 25", "25-35", "35-40", "Trên 40" để kiểm định

H0 : Không có sự khác biệt về fitness score ở các nhóm tuổi

H1 : Có sự khác biệt về fitness score ở các nhóm tuổi

p-val <2e-16

Mức ý nghĩa alpha = 0.05

Với các p-val lần lượt :

Nhóm A : 2e-16 < alpha

Nhóm B : 2e-16 < alpha

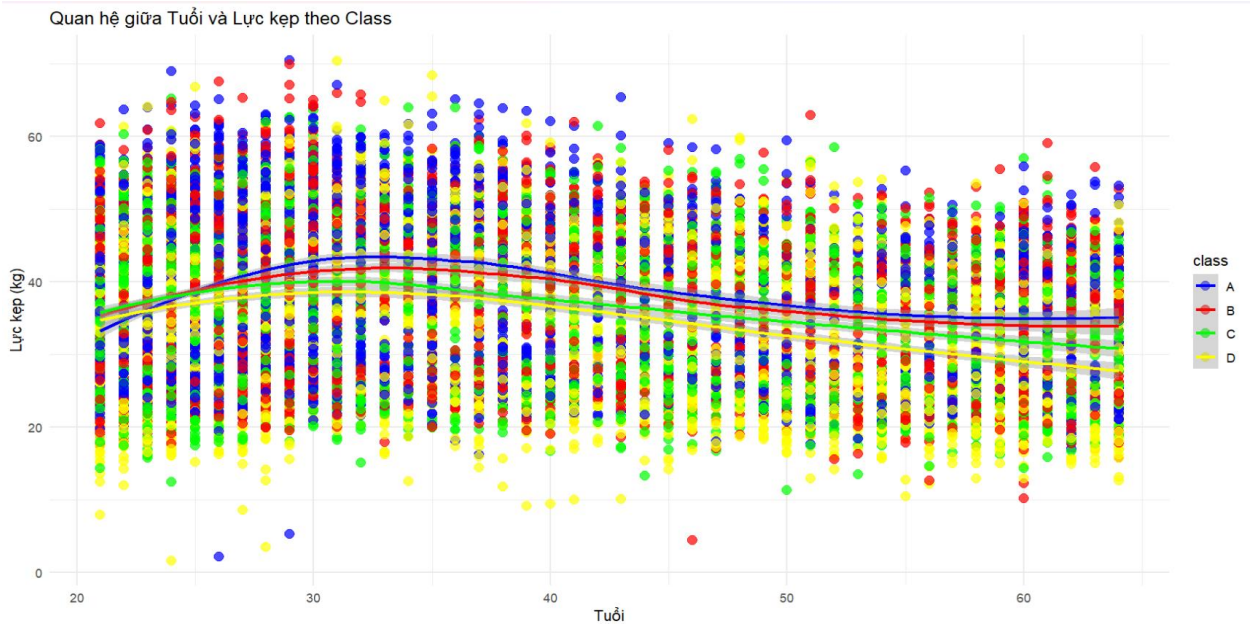
Nhóm C : 2e-16 < alpha

Nhóm D : 2e-16 < alpha

-> Ta đủ cơ sở chấp nhận H1

Vậy kết luận nhận xét ban đầu có thể tạm xem là mang ý nghĩa thống kê

- Biểu đồ thứ 9



Nhận xét : Lực kẹp ở cả 4 class đạt đỉnh ở giai đoạn 25 tới 40 tuổi , sau 40 thì chỉ số này có xu hướng giảm xuống

Để kiểm tra nhận xét trên có mang tính thống kê hay không , ta có :

H0 : Không có sự khác biệt về lực kẹp ở các nhóm tuổi

H1 : Có sự khác biệt về lực kẹp ở các nhóm tuổi

Mức ý nghĩa alpha = 0.05

Với các p-val lần lượt :

Nhóm A : 2e-16 < alpha

Nhóm B : 2e-16 < alpha

Nhóm C : 2e-16 < alpha

Nhóm D : 2e-16 < alpha

-> Ta đủ cơ sở chấp nhận H1

Vậy nhận xét ban đầu có thể tạm xem là mang ý nghĩa thống kê

- Biểu đồ thứ 10

Nhận xét : Chỉ số lực kẹp ở nam giới nhìn chung tốt hơn nữ giới ở cả 4 class . Tuy nhiên khi tuổi càng cao thì lực kẹp ở nam giới có xu hướng giảm nhưng ở nữ giới lại không thể hiện xu hướng tăng giảm rõ ràng

Để kiểm tra nhận xét trên có mang ý nghĩa thống kê không , ta thực hiện :

- Dựa vào kết quả kiểm định của plot 9 , ta đã có thể khẳng định chỉ số lực kẹp của nam tốt hơn nữ ở cả 4 class

và đồng thời đạt đỉnh ở độ tuổi 25-40 , sau đó thì đi xuống .Tuy nhiên để kiểm tra xu hướng lực kẹp theo tuổi tác ở riêng nữ giới thì ta có :

H0 : Không có sự khác biệt về lực kẹp nữ giới ở các nhóm tuổi

H1 : Có ít nhất 1 sự khác biệt về lực kẹp nữ giới ở các nhóm tuổi

Mức ý nghĩa alpha = 0.05

Với các p-val lần lượt :

Nhóm A : 2e-16 < alpha

Nhóm B : 2e-16 < alpha

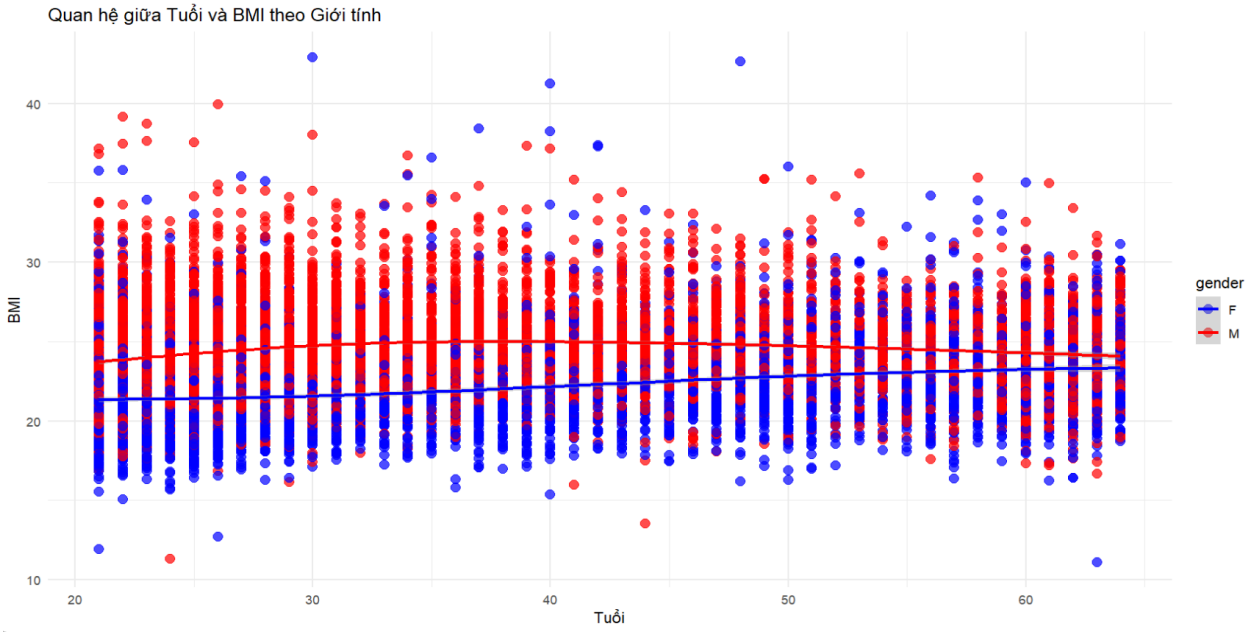
Nhóm C : 2e-16 < alpha

Nhóm D : 2e-16 < alpha

-> Ta đủ cơ sở chấp nhận H1

Vậy nhận xét ban đầu có thể tạm xem là mang ý nghĩa thống kê

- Biểu đồ thứ 11

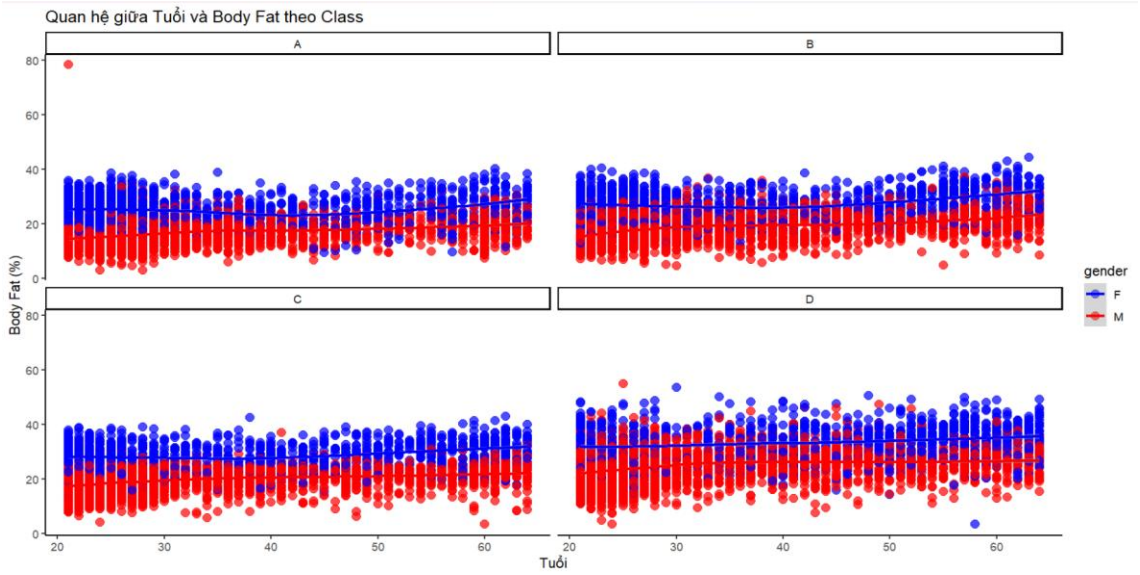


Nhận xét : Chỉ số BMI của nam giới ở mức cao trong độ tuổi từ 25 tới 40 . Sau 40 thì chỉ số bắt đầu giảm xuống

Tuy nhiên ở nữ giới thì chỉ số bmi lại thấp ở độ tuổi 20 tới trước 40 , nhưng ngoài 40 thì lại có xu hướng tăng lên , điều này khá lạ , nguyên nhân có thể do cột Female bị thiếu dữ liệu .

Biểu đồ và nhận xét trên chỉ mang tính tham khảo chứ chưa mang ý nghĩa thống kê

- Biểu đồ thứ 12



Nhận xét : % Mỡ(body_fat) ở cả 2 giới đều có xu hướng đi lên khi ở độ tuổi ngoài 40 , ở nữ giới thể hiện xu hướng này rõ hơn

Để kiểm tra nhận xét trên , ta có :

H0 : % mỡ không có sự khác biệt trên cả 2 giới ở mọi độ tuổi

H1 : % mỡ có sự khác biệt từ độ tuổi 40 trở đi trên cả 2 giới

Sau khi kiểm định t cho nữ giới và nam giới

cho alpha =0.05

Cả 2 p-val đều = 2.2e-16 < alpha nên ta chấp nhận H1

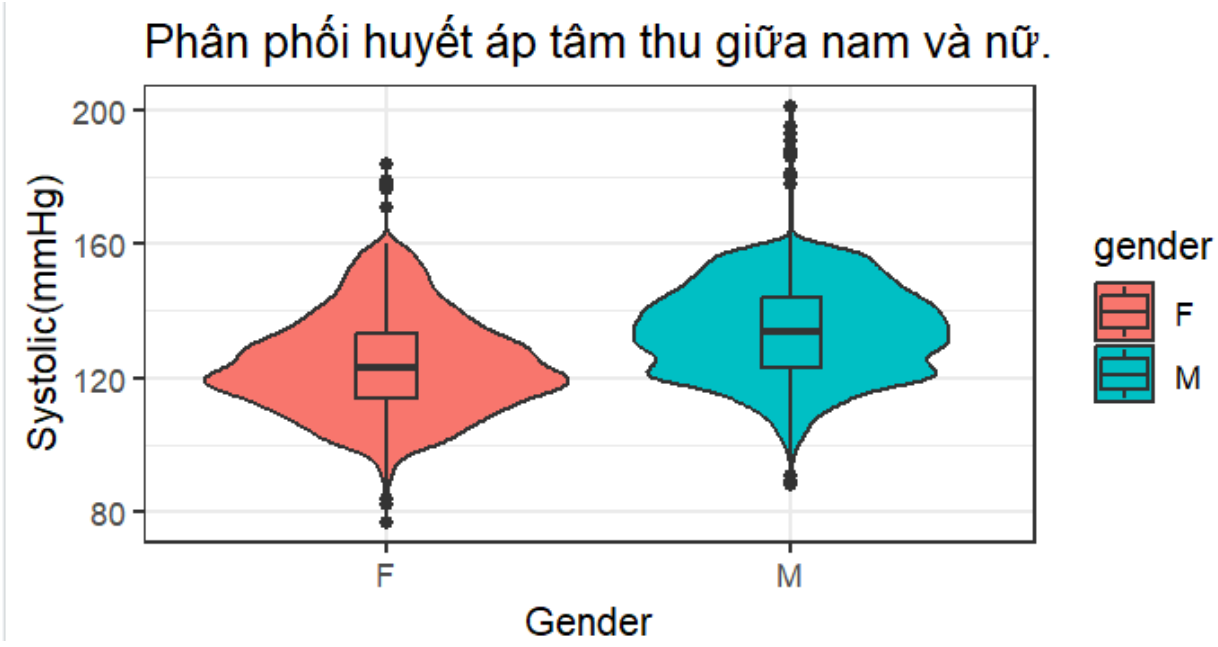
Vậy nhận xét ban đầu có mang ý nghĩa thống kê

- Biểu đồ thứ 13

Nhận xét : Ở cả 2 giới thì huyết áp tâm thu và huyết áp tâm trương đều có xu hướng tỉ lệ thuận , đồng thời 2 chỉ số này ở nam cao hơn nữ

Tách thành 2 biểu đồ nhỏ

14.1 Biểu đồ violin thể hiện phân phối huyết áp tâm thu giữa nam và nữ.



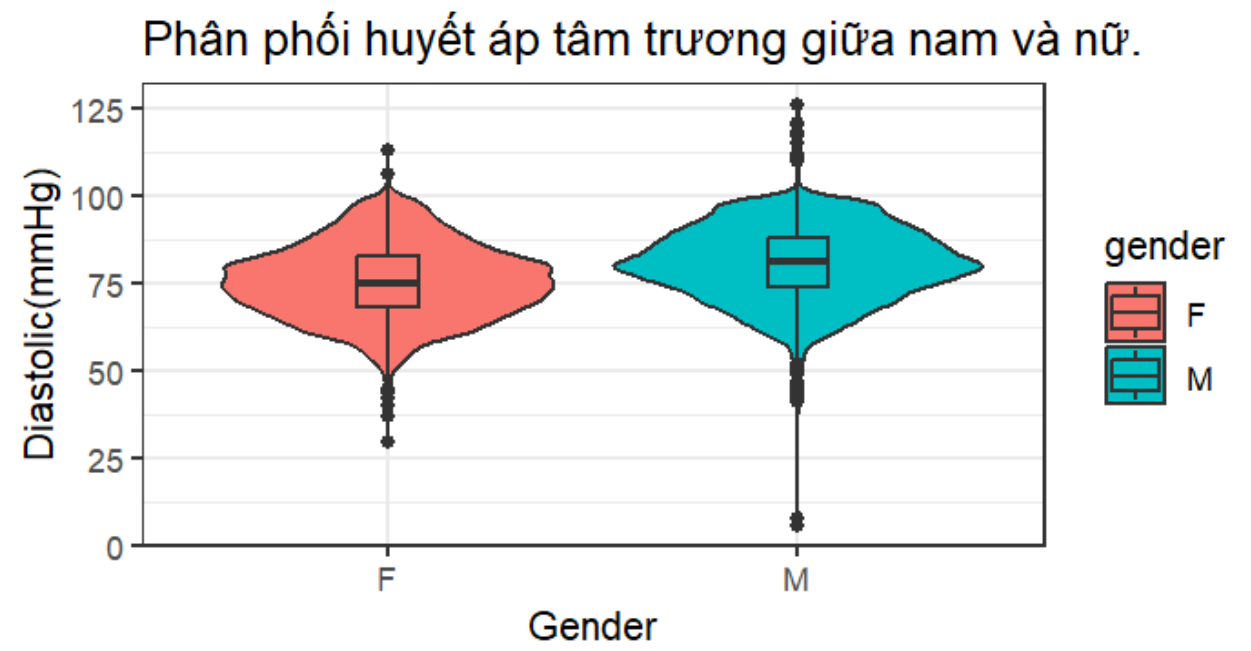
Qua biểu đồ ta thấy rõ được huyết áp tâm thu của nữ giới thấp hơn nam giới.

Ta sẽ kiểm định "Huyết áp tâm thu của nữ giới thấp hơn nam giới hay không?"

H0: Có sự giống nhau về huyết áp tâm thu giữa nam và nữ.

H1: Huyết áp tâm thu của nữ thấp hơn nam.

p-value = 0 < 0.05 -> Chấp nhận H1: Huyết áp tâm thu của nữ thấp hơn nam.



Qua biểu đồ ta thấy có vẻ huyết áp tâm trương của nữ giới thấp hơn nam giới.

Ta sẽ kiểm định "Huyết áp tâm trương của nữ giới thấp hơn nam giới hay không?"

H0: Có sự giống nhau về huyết áp tâm trương giữa nam và nữ.

H1: Huyết áp tâm trương của nữ thấp hơn nam.

p-value = 0 < 0.05 -> Chấp nhận H1: Huyết áp tâm trương của nữ thấp hơn nam.

H0 : Không có mối quan hệ tuyến tính giữa huyết áp tâm thu và huyết áp tâm trương

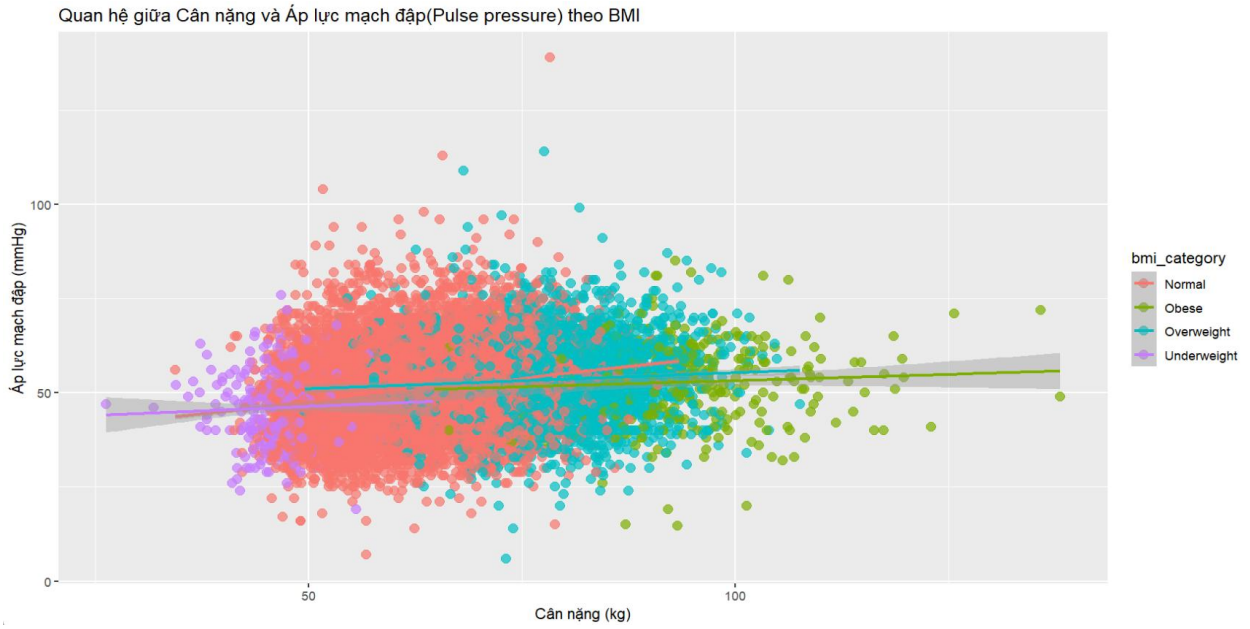
H1 : có mối quan hệ tuyến tính giữa huyết áp tâm thu và huyết áp tâm trương

Sau khi dùng cor.test() cho cả nam nữ , có

Cả 2 p-val = 2.2e-16 < alpha nên ta chấp nhận H1

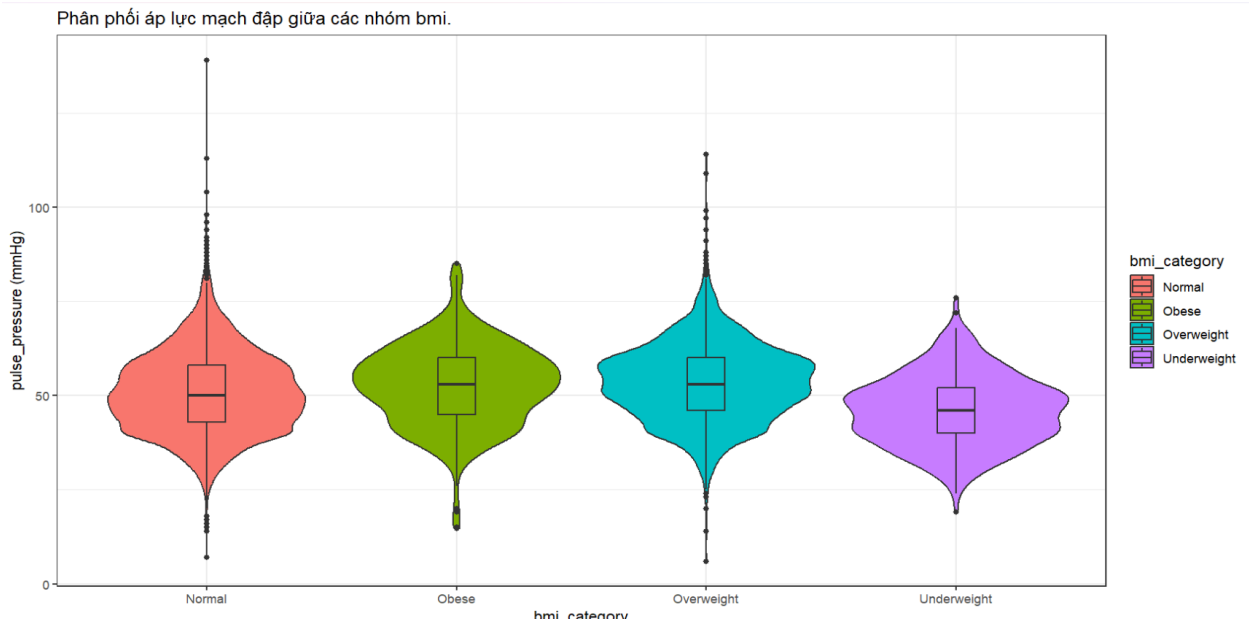
Vậy kết luận ban đầu mang ý nghĩa thống kê

- Biểu đồ thứ 14



Nhận xét : Ở cả 4 phân nhóm BMI đều không thể hiện xu hướng rõ ràng cho mối quan hệ giữa cân nặng và áp lực mạch đập . Ta không dám chắc nếu cân năng tặng thì áp lực mạch đập sẽ tăng hoặc ngược lại

Vẽ lại biểu đồ theo hướng khác



Qua biểu đồ ta thấy được sự khác biệt về áp lực mạch đập giữa các nhóm bmi

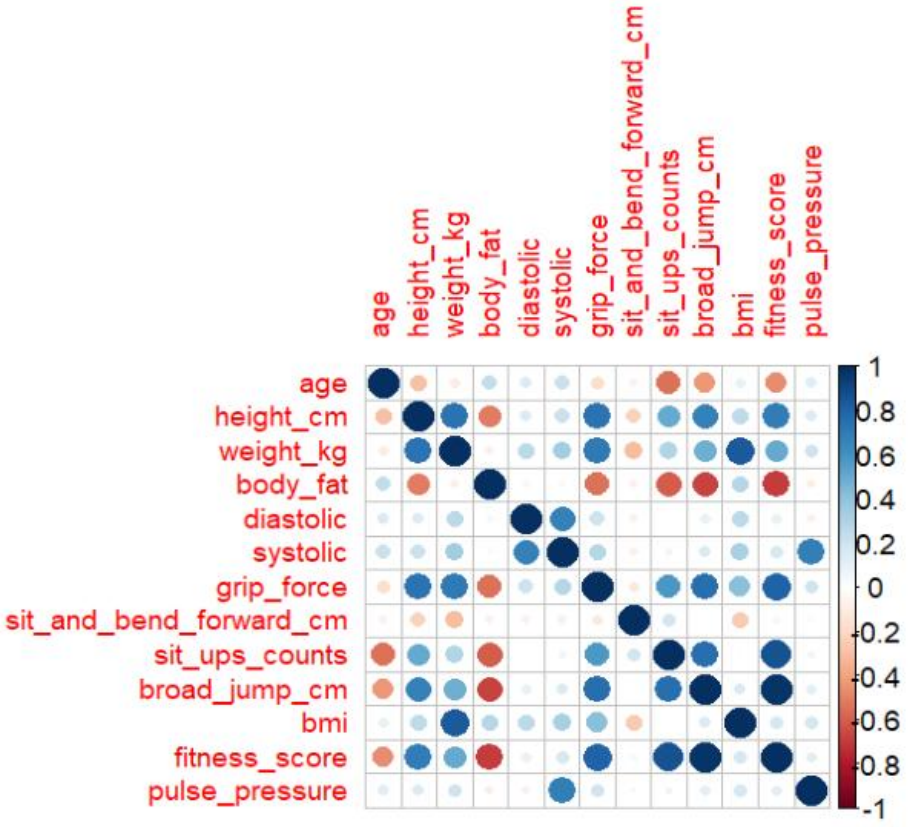
Ta sẽ kiểm định "Áp lực mạch đập giữa các nhóm bmi có sự khác biệt hay không?"

H0: Không có sự khác biệt về áp lực mạch đập giữa các nhóm bmi.

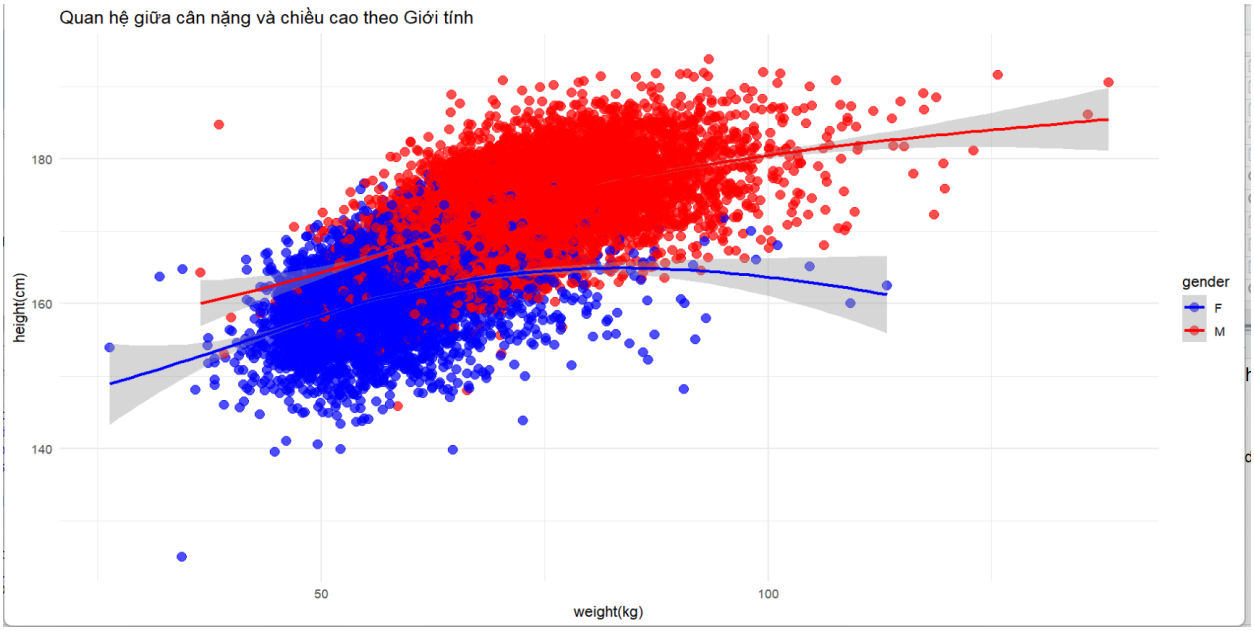
H1: Có sự khác biệt về áp lực mạch đập giữa các nhóm bmi.

p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: Có sự khác biệt về áp lực mạch đập giữa các nhóm bmi.

- Biểu đồ tương quan



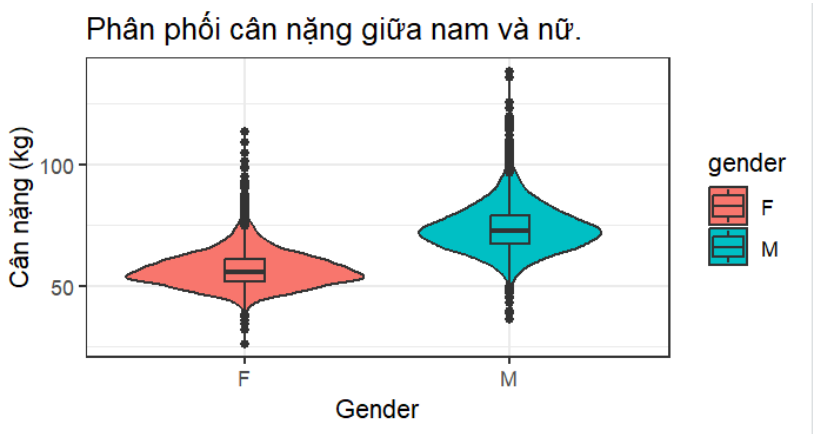
- Biểu đồ thứ 15



Nhận xét : Nhìn chung chỉ số cân nặng và chiều cao ở nam giới cao hơn nữ giới . Đường cong thể hiện xu hướng ở nữ giới

có dạng gần giống parabol , cho thấy chiều cao tăng lên theo cân nặng lúc ban đầu, sau đó dường như đạt đỉnh nhưng bắt đầu giảm ở các mức cân nặng cao hơn ,ở nam giới đường cong có xu hướng tăng lên từ từ, cho thấy chiều cao tăng theo cân nặng và có vẻ tiếp tục tăng lên ở mức cân nặng lớn hơn.

Tách ra các biểu đồ nhỏ



Qua biểu đồ ta thấy được cân nặng của nữ giới nhẹ hơn nam giới.

Ta sẽ kiểm dịnh "Nữ giới có cân nặng nhẹ hơn nam giới"

H0: cân nặng giữa nam và nữ bằng nhau.

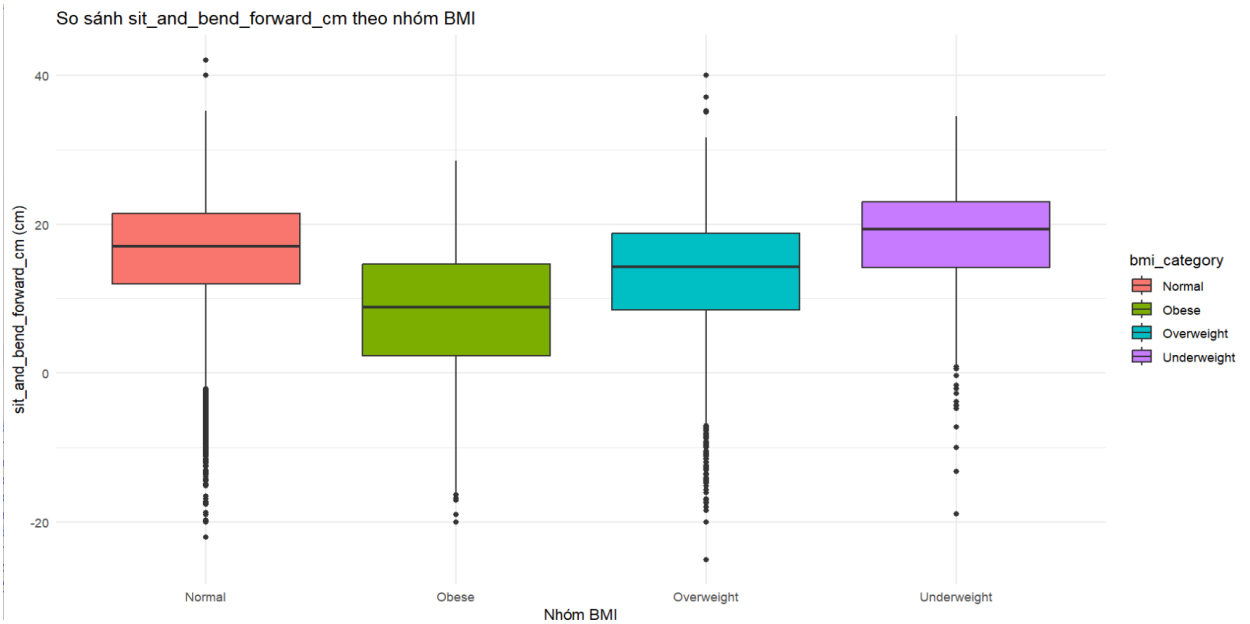
H1: Cân nặng của nữ thấp hơn nam.

p-value = 0 < 0.05 -> Chấp nhận H1: Cân nặng của nữ thấp hơn nam.

Tuy nhiên chưa tìm được cách kiểm định đường cong parabol có mang ý nghĩa thống kê không ?

Nên nhận xét ban đầu chỉ mang tính tham khảo chứ chưa có ý nghĩa thống kê

- Biểu đồ thứ 16



Nhận xét : Chỉ số sit_and_bend_forward_cm trung bình thấp nhất ở 2 nhóm Obese và Overweight . Nhóm Normal cao hơn và nhóm Underweight là cao nhất

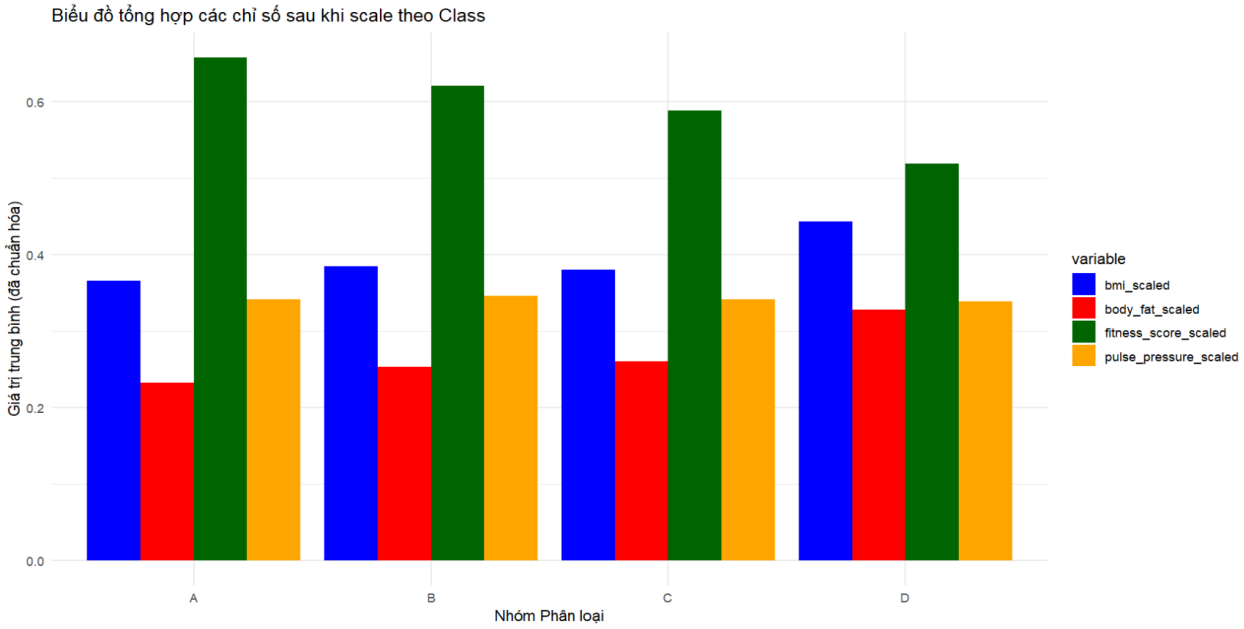
Ta sẽ kiểm định "lực kẹp giữa các nhóm bmi có giống nhau không?"

H0: sit_and_bend_forward_cm giữa các nhóm bmi không khác nhau.

H1: sit_and_bend_forward_cm giữa các nhóm bmi khác nhau.

p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: sit_and_bend_forward_cm giữa các nhóm bmi khác nhau.

-Theo các class (biểu đồ tổng hợp)



Nhận xét :

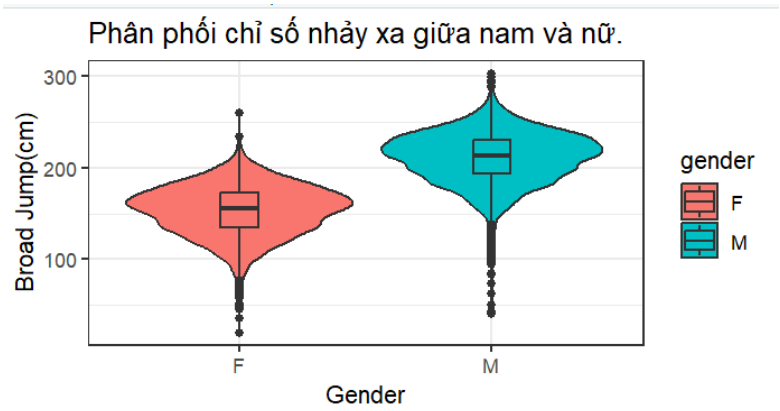
-Chỉ số BMI ở class A thấp nhất , tuy nhiên ở class B lại cao hơn class C , class D cao nhất

-Fitness score giảm dần đều theo từ class A sang D

-Body fat tăng dần đều từ class A sang D

-Pulse pressure (áp lực mạch đập) gần như ngang nhau ở cả 4 class

- Biểu đồ thứ 17



H0: Chỉ số nhảy xa giữa nam và nữ bằng nhau.

H1: Chỉ số nhảy xa của nữ yếu hơn nam.

p-value = 0 < 0.05 -> Chấp nhận H1: Chỉ số nhảy xa của nữ yếu hơn nam.

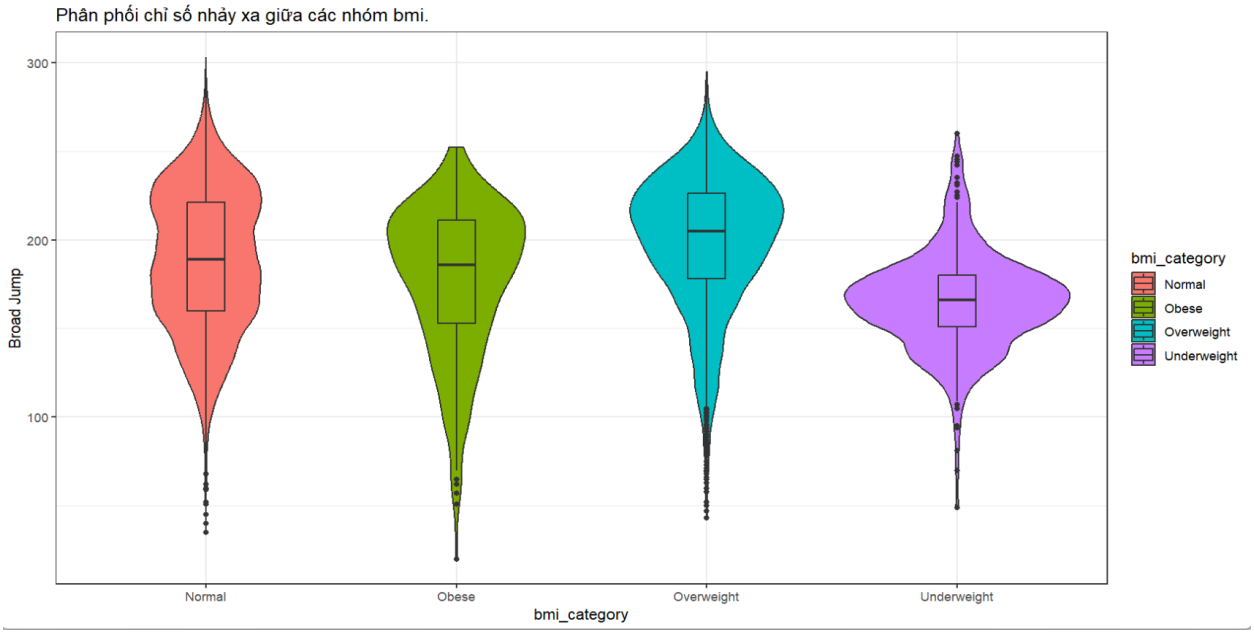
- Bổ sung : Kiểm định giả thuyết "Nhóm huyết áp phụ thuộc vào nhóm bmi hay không?"

H0: Nhóm huyết áp không phụ thuộc vào nhóm bmi. (2 nhóm độc lập nhau)

H1: Nhóm huyết áp phụ thuộc vào nhóm bmi.

- Sau khi tạo biến pivot_blood_pressure_bmi và chuyển thành ma trận , thực hiện hàm chisq.test() và có kết quả : p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: Nhóm huyết áp phụ thuộc vào nhóm BMI.

- Biểu đồ thứ 18 :

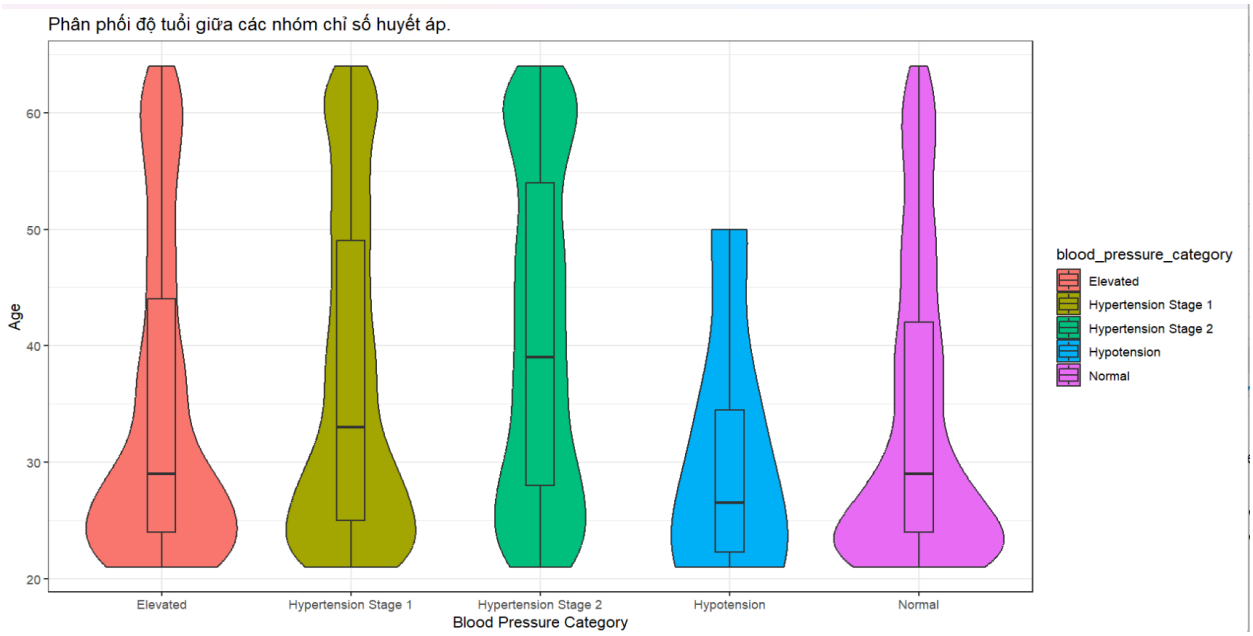


H0: Không có sự khác biệt về chỉ số nhảy bật xa giữa các nhóm bmi.

H1: Có sự khác biệt về chỉ số nhảy bật xa giữa các nhóm bmi.

p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: Có sự khác biệt về Chỉ số nhảy bật xa giữa các nhóm bmi

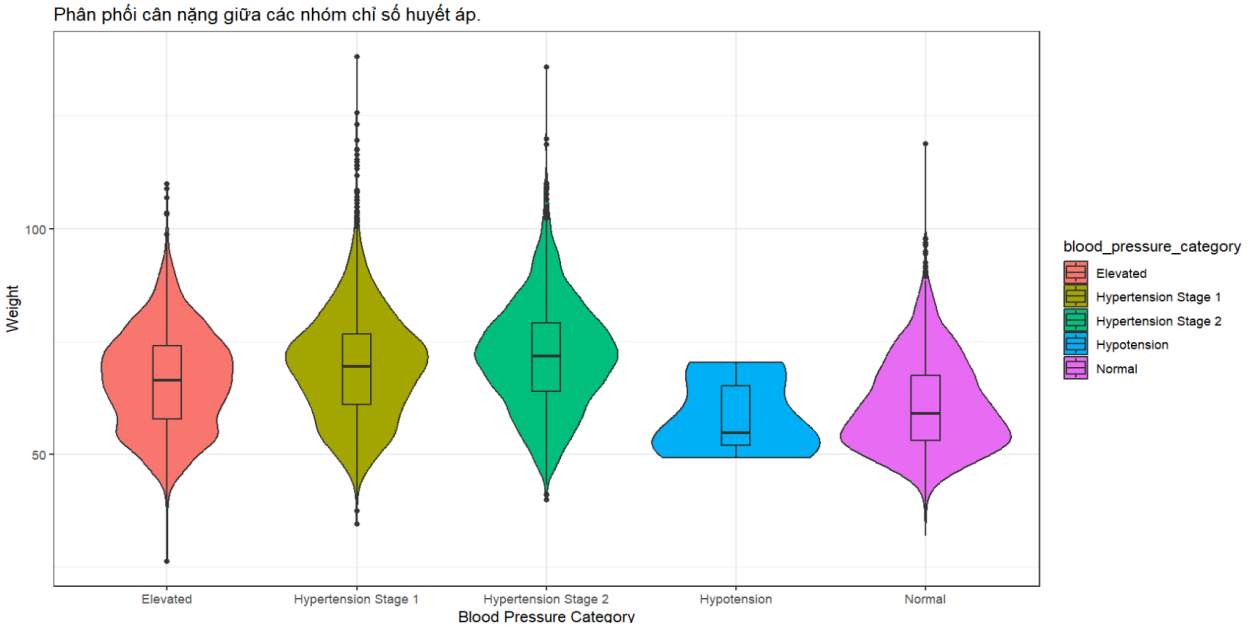
- Biểu đồ thứ 19



H0: Ở các nhóm chỉ số huyết áp thì độ tuổi không khác biệt

H1: Ở các nhóm chỉ số huyết áp thì độ tuổi có sự khác biệt

p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: Độ tuổi ở các nhóm chỉ số huyết áp có khác biệt.

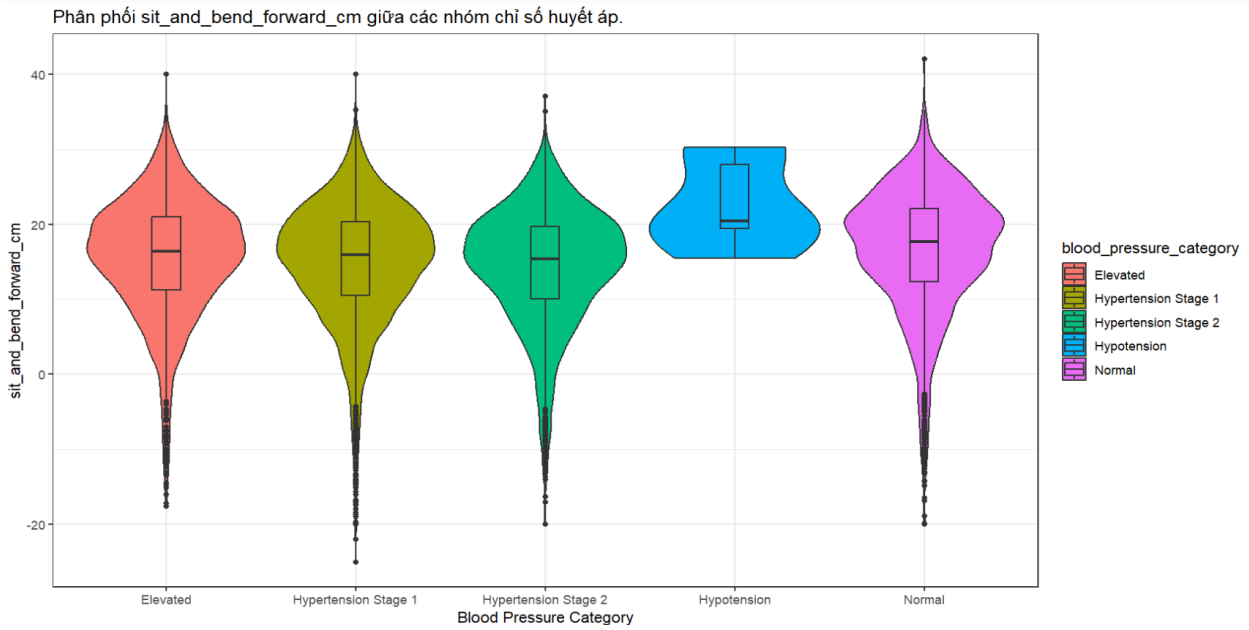


H0: Cân nặng ở các nhóm chỉ số huyết áp không khác biệt.

H1: Cân nặng ở các nhóm chỉ số huyết áp có khác biệt.

p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: Cân nặng ở các nhóm chỉ số huyết áp có khác biệt.

- Biểu đồ thứ 20



H0: sit_and_bend_forward_cm ở các nhóm chỉ số huyết áp không khác biệt.

H1: sit_and_bend_forward_cm ở các nhóm chỉ số huyết áp có khác biệt.

p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: sit_and_bend_forward_cm ở các nhóm chỉ số huyết áp có khác biệt.

- Bổ sung : Kiểm định giả thuyết "Phân lớp hiệu suất phụ thuộc vào giới tính hay không?"

H0: Phân lớp hiệu suất không phụ thuộc vào giới tính (2 nhóm độc lập nhau)

H1: Phân lớp hiệu suất phụ thuộc vào giới tính

- Tạo pivot_class_gender sau đó chuyển đổi thành ma trận và dùng hàm chisq.test() ,có p-value < 2.2e-16 < 0.05 -> Chấp nhận H1: Phân lớp hiệu suất phụ thuộc vào giới tính

4.Bootstrap

1) Chuẩn bị dữ liêu:
   Dữ liệu data_boot chứa các biến phân tích như age, heigh_cm,…, và fitness_score
2) Hàm tính toán Boostrap:
   Hàm boot_mu_fun nhận dữ liệu data và chỉ số mẫu ind, sau đó tính giá trị trung bình của dữ liệu được chọn, sau đó giá trị trung bình được trả về để làm tham số thống kê
3) Vòng lặp qua các biến số
   Với mỗi biến trong danh sách variables, thực hiện phân tích bootstrap với 1000 mẫu ngẫu nhiên (thông qua hàm boot).
   Trích xuất 3 thông số quan trọng
   - Mean: Giá trị trung bình từ bootstrap

   Bias: Độ lệch giữa giá trị trung bình bootstrap và giá trị ban đầu
   - StdError: Sai số chuẩn của các ước lượng
4) Với mỗi biến, vẽ biểu đồ histogram hiển thị phân phối của các giá trị trung bình bootstrap (out$t). Đường thẳng đứng màu xanh đánh dấu giá trị trung bình ban đầu
5) In ra bảng kết quả chứa các thông số thống kê cho từng biến và tập hợp các biểu đồ thành 1 lưới hiển thị trực quan

Bảng thống kê chứa các thông số thống kê cho từng biến

	Variable	Mean	Bias	StdError	CI_Lower	CI_Upper
1	age	36.76972	0.0012307290	0.11581607	36.54473	36.99891
2	height_cm	168.56429	0.0011707589	0.07048445	168.42956	168.70713
3	weight_kg	67.44829	-0.0001665338	0.10250553	67.25667	67.65845
4	body_fat	23.23516	-0.0012613595	0.06359111	23.10046	23.35397
5	diastolic	78.80013	-0.0014162467	0.08575023	78.62975	78.97967
6	systolic	130.26407	-0.0003961121	0.12787636	130.00365	130.50862
7	grip_force	36.97910	-0.0014185503	0.09321857	36.80357	37.15982
8	sit_and_bend_forward_cm	15.20959	0.0037828105	0.07273312	15.07471	15.35662
9	sit_ups_counts	39.78437	0.0008471178	0.12872312	39.51992	40.03921
10	broad_jump_cm	190.26827	0.0149566056	0.34375151	189.63061	190.97605
11	bmi	23.60516	-0.0001449130	0.02507862	23.55252	23.65690
12	fitness_score	84.08796	-0.0033595351	0.16517388	83.78404	84.41308
13	pulse_pressure	51.46395	-0.0010348860	0.09485866	51.27786	51.64531

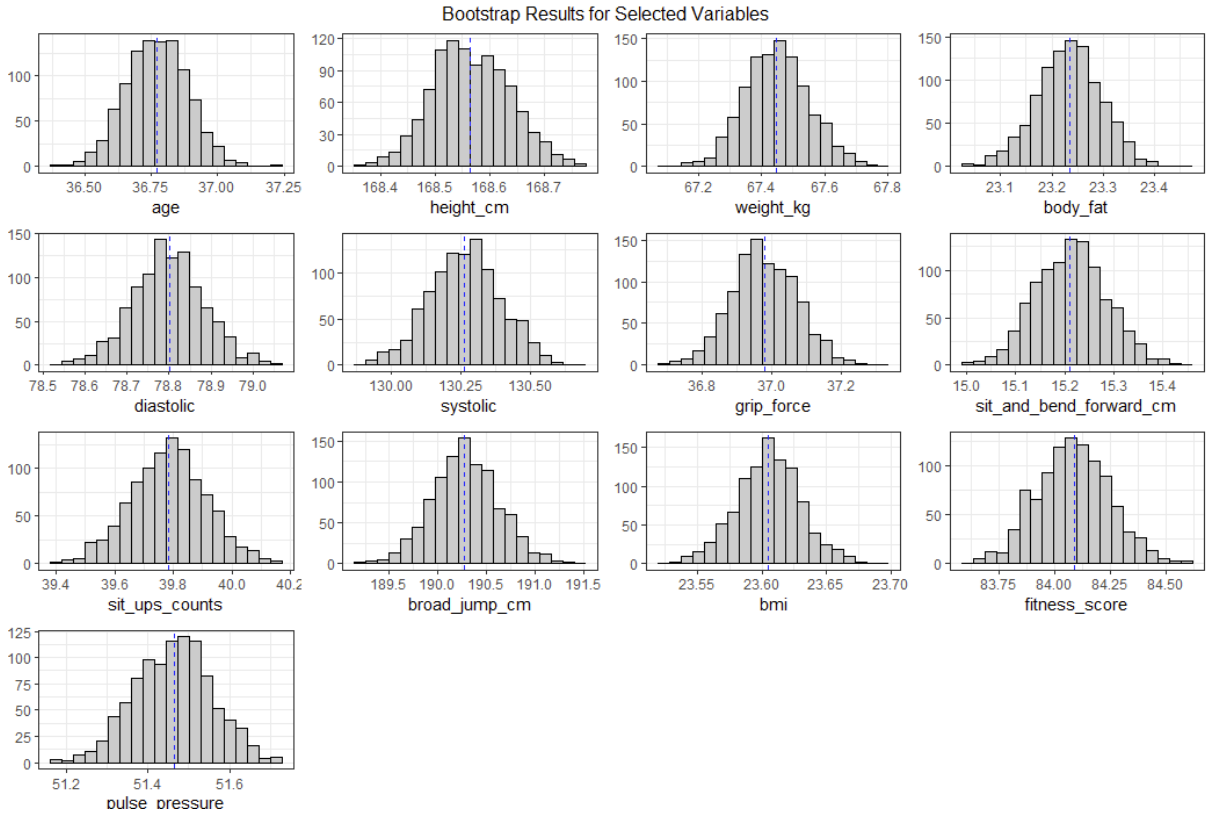
-Nhận xét: Giá trị "Bias" của hầu hết các biến là rất nhỏ (gần bằng 0), chứng tỏ mẫu

bootstrap đại diện tốt cho dữ liệu gốc.

broad_jump_cm có sai số chuẩn và khoảng tin cậy tương đối lớn, cho thấy sự biến thiên cao trong dữ liệu.

Ngược lại, các biến như age, height_cm, weight_kg, và fitness_score có khoảng tin cậy hẹp và sai số nhỏ, chứng minh sự ổn định của ước lượng.

Tập hợp các biểu đồ thành 1 lưới:



-Nhận xét:

-Phân phối các biến: Các biến như age, height_cm, weight_kg, body_fat, diastolic, systolic, grip_force, sit_and_bend_forward_cm, sit_ups_counts, broad_jump_cm, bmi, fitness_score, và pulse_pressure đều có dạng phân phối gần với phân phối chuẩn. Điều này cho thấy dữ liệu được phân phối khá đều và không có hiện tượng lệch lớn.

-Sự biến thiên: Mức độ rộng của các biểu đồ cho thấy sự biến thiên trong từng biến. Ví dụ:Biến pulse_pressure có sự phân bố khá hẹp, cho thấy dữ liệu ít biến động.Biến như sit_up_counts hoặc broad_jump_cm cũng có sự biến thiên thấp.

5. Classification

Bước 1: xử lí dữ liệu

Vì các cột data có giá trị không chênh nhau quá nhiều nên em không scale data.

Loại bỏ các cột không cần thiết :

Em thấy data có cột là bmi_catergory được tạo từ 1 cột bmi nên để nó có thể gây giảm độ chính xác của mô hình nên em bỏ nó.

Bước 2: Viết hàm đánh giá mô hình:

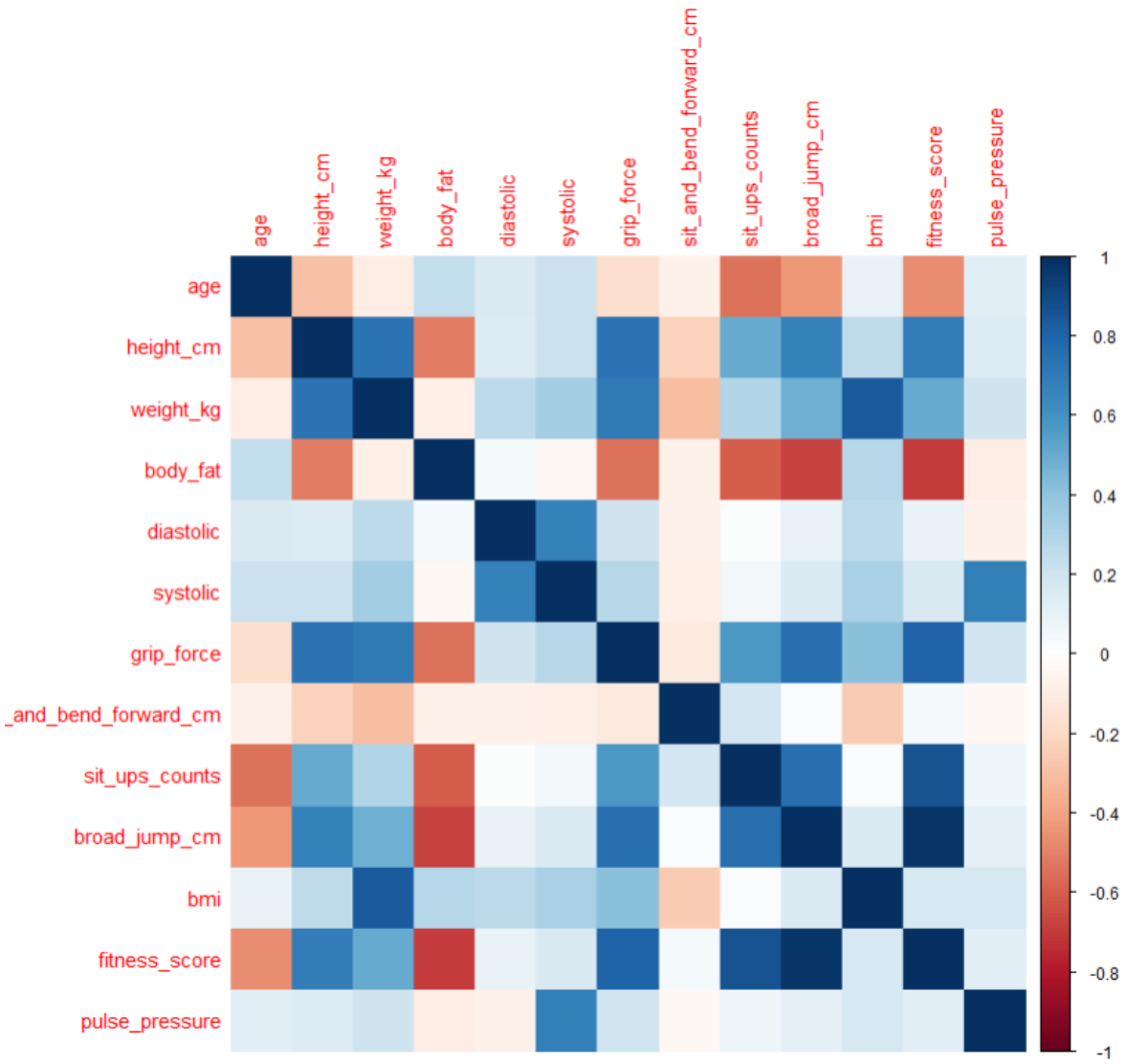
Em viết hàm đánh giá dựa trên Precision, Recall ,Kappa, Marco F1

Bước 3: Cài đặt K-fold(cv=5) để đánh giá mô hình tốt hơn.

Bước 4: chọn mô hình

- Naivebayes :

  Với mô hình này thì vì yêu cầu các biến độc lập nhau nên em cần phải vẽ đồ thị correlation



Em nhận thấy data có nhiều biến có tương quan cao với nhau nên khả năng cao là mô hình này sẽ không chính xác .

```
> conf_matrix<-table( data$class,nb_pred$class)
> eval_multi_class(conf_matrix)
$Precision
        A         B         C         D
0.5946614 0.4137566 0.4960440 0.6532775

$Recall
        A         B         C         D
0.7190675 0.3507775 0.4313491 0.7011976

$Accuracy
[1] 0.550587

$Kappa
[1] 0.4007788

$Macro_F1
[1] 0.5449593

>
```

Nhận xét mô hình

Mô hình có độ chính xác là 54.5%. Đây là một mức độ chính xác tương đối thấp, vì trong correlation matrix ta thấy được một số biến có tương quan cao với nhau nên mô hình yếu

Kappa : 0.4 cho thấy rằng mức độ tương đồng của mô hình và thực tế là khá yếu

Mô hình có Precision cao nhất ở lớp D tiếp theo là lớp A, trong khi Precision của lớp B và C còn khá thấp, điều này có thể chỉ ra rằng mô hình gặp khó khăn trong việc phân loại chính xác các đối tượng thuộc lớp B và C.

Mô hình có Recall cao nhất ở lớp A và D, nhưng lớp B và C bị bỏ sót khá nhiều đối tượng, điều này có thể cho thấy mô hình chưa đủ nhạy để nhận diện đầy đủ các đối tượng của các lớp này

54.5, điều này có nghĩa là mô hình có hiệu suất trung bình khi đánh giá tất cả các lớp, và có sự cân bằng giữa Precision và Recall, tuy nhiên vẫn khá thấp.

Dùng kflod kiểm tra thêm :

```
> print(nb_model)
Naive Bayes

13373 samples
   15 predictor
    4 classes: 'A', 'B', 'C', 'D'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 10699, 10698, 10698, 10698, 10699
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.3936290  0.1915873
   TRUE      0.4899419  0.3198798

Tuning parameter 'laplace' was held constant at a value of 0
Tuning
 parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = TRUE
 and adjust = 1.
> print(nb_model$resample)  # In ra thông tin chi tiết của từng fold
    Accuracy     Kappa usekernel laplace adjust Resample
1  0.4775617 0.3033014      TRUE       0      1    Fold1
2  0.3960359 0.1949074     FALSE       0      1    Fold1
3  0.4882243 0.3176025      TRUE       0      1    Fold2
4  0.3884112 0.1845631     FALSE       0      1    Fold2
5  0.4923364 0.3230082      TRUE       0      1    Fold3
6  0.3966355 0.1957907     FALSE       0      1    Fold3
7  0.4964486 0.3285684      TRUE       0      1    Fold4
8  0.3947664 0.1930346     FALSE       0      1    Fold4
9  0.4951384 0.3269184      TRUE       0      1    Fold5
10 0.3922962 0.1896406     FALSE       0      1    Fold5
>
```

Nhận xét tổng quan về các fold

Accuracy khá thấp , chỉ trong khoảng từ 38% đến 50% -> mô hình dự đoán kém

kappa rất thấp, chỉ tầm 19% nếu không dùng kernel đến 32% nếu có dùng kernel và nên sử dụng kernel thì mô hình có vẻ tốt hơn nhưng 31% vẫn là khá thấp nên mức độ tương đồng thực tế thấp

Sau đó em loại bỏ các cột có tương quan cao bằng hàm findCorrelation() (threshold =0.7 , em đã thử qua và thấy 0.7 là sự lựa chọn tốt nhất) có sẵn trong thư viện carret (hàm này sẽ bỏ đi 1 biến trong cặp biến có sự tương quan cao hơn ngưỡng).

Các cột được giữ lại

```
> print("Các cột giữ lại sau khi loại bỏ tương quan cao:")
[1] "Các cột giữ lại sau khi loại bỏ tương quan cao:"
> print(names(filtered_data))
[1] "age"                    "body_fat"
[3] "diastolic"              "systolic"
[5] "sit_and_bend_forward_cm" "sit_ups_counts"
[7] "bmi"                    "pulse_pressure"
>
```

```
$Precision
        A         B         C         D
0.5839136 0.4150943 0.5167087 0.7283537

$Recall
        A         B         C         D
0.7268380 0.3421053 0.4902782 0.7152695

$Accuracy
[1] 0.5686084

$Kappa
[1] 0.4248023

$Macro_F1
[1] 0.5647945

>
```

Mô hình được cải thiện rõ rệt Accuracy 57%, kappa 42%, F1 56% nhưng vẫn chưa phải là tốt

Nhận xét bằng K flod :

```
> print(nb_model)
Naive Bayes

13373 samples
    8 predictor
    4 classes: 'A', 'B', 'C', 'D'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 10699, 10698, 10699, 10697, 10699
Resampling results across tuning parameters:

  usekernel  Accuracy   Kappa
  FALSE      0.5654657  0.4206139
   TRUE      0.5782524  0.4376572

Tuning parameter 'laplace' was held constant at a value of 0
Tuning
 parameter 'adjust' was held constant at a value of 1
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were laplace = 0, usekernel = TRUE
 and adjust = 1.
> print(nb_model$resample)  # In ra thông tin chi tiết của từng fold
    Accuracy     Kappa usekernel laplace adjust Resample
1  0.5729245 0.4305571      TRUE       0      1    Fold1
2  0.5605834 0.4140973     FALSE       0      1    Fold1
3  0.5839252 0.4452287      TRUE       0      1    Fold2
4  0.5805607 0.4407450     FALSE       0      1    Fold2
5  0.5744203 0.4325463      TRUE       0      1    Fold3
6  0.5572177 0.4096224     FALSE       0      1    Fold3
7  0.5908072 0.4543854      TRUE       0      1    Fold4
8  0.5713752 0.4284834     FALSE       0      1    Fold4
9  0.5691847 0.4255684      TRUE       0      1    Fold5
10 0.5575916 0.4101214     FALSE       0      1    Fold5
>
```

Nhận xét: mô hình đã được cãi tiến, các hệ số đều cao hơn, 58% Accuracy nếu có dùng kernel và 56% nếu không dùng kernel

42% nếu không dùng kernel và 43% nếu dùng kernel cho chỉ số Kappa

--> mô hình đã được cải tiến nhưng chưa tốt

- LDA :

```
$Precision
        A         B         C         D
0.6836820 0.4581661 0.5272118 0.8223196

$Recall
        A         B         C         D
0.7325164 0.4437799 0.5650613 0.7302395

$Accuracy
[1] 0.6178868

$Kappa
[1] 0.4905091

$Macro_F1
[1] 0.6203622
```

Nhận thấy lda tốt hơn naiveBayes vì nó giả sử các feature có cùng phương sai và dùng phân phối chuẩn để tính ước lượng hàm mật độ xác suất đồng thời fj(x) --> ít bị ảnh hưởng bởi vấn đề độc lập tuyến tính hơn so với Naive Bayes

Accuracy 62% cho thấy mô hình dự đoán độ chính xác trung bình

Kappa: 49% cho thấy sự tương đồng trung bình

F1 62% : tốt hơn Naive Bayes nhưng vẫn chưa thực sự hiệu quả

```
Linear Discriminant Analysis

13373 samples
   15 predictor
    4 classes: 'A', 'B', 'C', 'D'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 10699, 10698, 10698, 10698, 10699
Resampling results:

  Accuracy  Kappa
  0.615046  0.486723

> print(lda_model$resample)  # In ra thông tin chi tiết của từng fold
   Accuracy     Kappa parameter Resample
1 0.6091997 0.4789285      none    Fold1
2 0.6100935 0.4801173      none    Fold2
3 0.6033645 0.4711560      none    Fold3
4 0.6213084 0.4950708      none    Fold4
5 0.6312640 0.5083422      none    Fold5
>
```

- Multinomial logistic

```
> eval_multi_class(conf_matrix_logistic)
$Precision
        A         B         C         D
0.6946158 0.4685162 0.5275925 0.7782490

$Recall
        A         B         C         D
0.7402869 0.4494617 0.5204906 0.7691617

$Accuracy
[1] 0.619831

$Kappa
[1] 0.4931055

$Macro_F1
[1] 0.6185441

> print(logistic_model)
Penalized Multinomial Regression

13373 samples
   15 predictor
    4 classes: 'A', 'B', 'C', 'D'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 10698, 10699, 10699, 10699, 10697
Resampling results across tuning parameters:

  decay  Accuracy   Kappa
  0e+00  0.6171360  0.4895113
  1e-04  0.6170612  0.4894115
  1e-01  0.6187065  0.4916053

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was decay = 0.1.
> print(logistic_model$resample)  # In
    Accuracy     Kappa decay Resample
1  0.6164486 0.4885959 0e+00    Fold1
2  0.6142056 0.4856056 1e-01    Fold1
3  0.6160748 0.4880976 1e-04    Fold1
4  0.6136874 0.4849112 0e+00    Fold2
5  0.6133134 0.4844124 1e-01    Fold2
6  0.6140613 0.4854095 1e-04    Fold2
7  0.6140613 0.4854152 0e+00    Fold3
8  0.6185490 0.4913982 1e-01    Fold3
9  0.6140613 0.4854152 1e-04    Fold3
10 0.6039641 0.4719490 0e+00    Fold4
11 0.6077038 0.4769362 1e-01    Fold4
12 0.6032162 0.4709518 1e-04    Fold4
13 0.6375187 0.5166852 0e+00    Fold5
14 0.6397608 0.5196743 1e-01    Fold5
15 0.6378924 0.5171835 1e-04    Fold5
>
```

#Accuracy Kappa gần giống như trên lda

Các fold có Accuracy và kappa không quá chênh lệch nhau

Sau khi train và sử dụng K-fold thì thấy mô hình phù hợp nhất cho dữ liệu là lqd và mutinomial logistic.

Decision tree :

```
$Precision
        A         B         C         D
0.5377734 0.4154037 0.6872753 0.7919371

$Recall
        A         B         C         D
0.8084280 0.5000000 0.2859707 0.6940120

$Accuracy
[1] 0.5721229

$Kappa
[1] 0.4294675

$Macro_F1
[1] 0.5895511

>
```

nhận xét mô hình Decision Trê có độ chính xác thấp cho bộ dữ liệu này

- XGboost:

```
> eval_multi_class(conf_matrix_xgb)
$Precision
        0         1         2         3
0.9090409 0.9296498 0.9710510 0.9993835

$Recall
        A         B         C         D
0.9886432 0.9207536 0.9231229 0.9706587

$Accuracy
[1] 0.9507964

$Kappa
[1] 0.9343946

$Macro_F1
[1] 0.9515374

>
```

Nhận thấy XGboost có độ chính xác rất cao, lên đến 95% Accuracy, 93% Kappa, 94% F1 cho thấy các chỉ số đều tốt

Về Precision và recall thì ở Class B có thấp hơn so với các class còn lại một chút nhưng vẫn là rất tốt.Tuy nhiên tính giải thích được của mô hình này thấp.

-Random forest :

```
Call:
 randomForest(formula = class ~ ., data = data, ntree = 100, mtry = sqrt(ncol(data) -      1), importance = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 4

        OOB estimate of  error rate: 26.85%
Confusion matrix:
     A    B    C    D class.error
A 2808  479   47   12   0.1607890
B  706 1975  523  140   0.4093900
C  277  657 2265  144   0.3224649
D   49  161  396 2734   0.1814371
>
> # Đánh giá mô hìn
> eval_multi_class(
$Precision
A B C D
1 1 1 1

$Recall
A B C D
1 1 1 1

$Accuracy
[1] 1

$Kappa
[1] 1

$Macro_F1
[1] 1

>
```

Random forest chính xác nhất (dựa trên đánh giá lần train này vì các chỉ số đều 1 hết) (nhưng có thể đây là do ngẫu nhiên hoặc mô hình bị overfit)

Kiểm tra lại bằng K-fold :

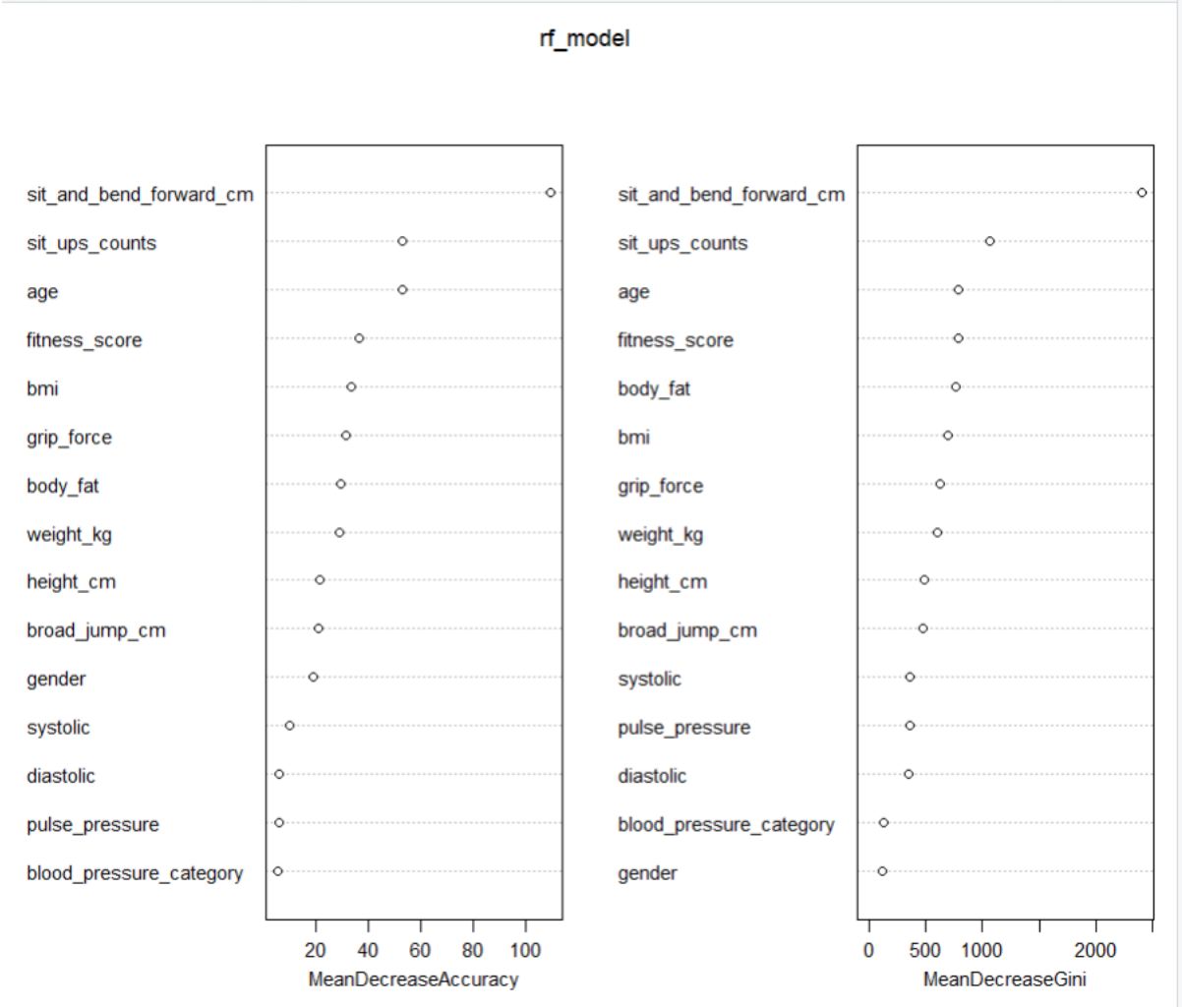
```
Random Forest

13373 samples
   15 predictor
    4 classes: 'A', 'B', 'C', 'D'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 10699, 10698, 10699, 10698, 10698
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7053018  0.6070645
  10    0.7467279  0.6622973
  18    0.7450087  0.6600041

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 10.
> print(rf_model$resample)  # In ra thông tin chi tiết của từng fold
    Accuracy     Kappa mtry Resample
1  0.7116679 0.6155564    2    Fold1
2  0.7509349 0.6679061   10    Fold1
3  0.7546746 0.6728917   18    Fold1
4  0.7050467 0.6067252    2    Fold2
5  0.7420561 0.6560698   10    Fold2
6  0.7401869 0.6535771   18    Fold2
7  0.6997008 0.5995973    2    Fold3
8  0.7341062 0.6454667   10    Fold3
9  0.7363500 0.6484571   18    Fold3
10 0.7001869 0.6002375    2    Fold4
11 0.7506542 0.6675308   10    Fold4
12 0.7446729 0.6595547   18    Fold4
13 0.7099065 0.6132062    2    Fold5
14 0.7558879 0.6745129   10    Fold5
15 0.7491589 0.6655400   18    Fold5
```

sau khi thực hiện K-fold thì thấy Accuray chỉ khoảng 74% và kappa ở tầm 66% khá tốt, thấp hơn lần trước có thể là do sự ngẫu nhiên hoặc mô hình trước bị overfit

Các chỉ số của các fold đều chênh lệch nhau không quá nhiều

Sử dụng mtry =18 hoặc =10 có vẻ là sự lựa chọn tốt.

Nhờ vào mô hình này ta cũng có được tầm quan trọng của các biến ảnh hưởng đến độ chính xác của mô hình.



- Nhận xét : Mô hình tốt nhất là random forest vì độ chính xác khá ổn và tính giải thích được của nó khá tốt.

Ngoài ra nhờ vào nó ta có thể xem xét được tầm quan trọng của các biến đối với độ chính xác mô hình.

1. **Mean Decrease Accuracy (MDA)**:
    - Thước đo này đánh giá mức độ giảm độ chính xác của mô hình nếu loại bỏ một biến cụ thể. Biến nào có giá trị MDA cao, tức là nếu loại bỏ biến đó, mô hình sẽ giảm độ chính xác đáng kể, cho thấy biến này rất quan trọng.
2. **Mean Decrease Gini (MDG)**:

- Thước đo này thể hiện mức giảm độ thuần nhất của các node trong cây quyết định khi chia tách dữ liệu theo biến cụ thể. Giá trị MDG cao cho thấy biến đó rất hữu ích để phân tách dữ liệu, và do đó đóng vai trò lớn trong dự đoán.

- Dựa vào 2 biểu đồ MDA và MDG , ta xếp tầm quan trọng của các yếu tố (biến) tới hiệu quả việc tập thể dục theo các nhóm như sau :

Nhóm yếu tố quan trọng cao nhất:

- **sit_and_bend_forward_cm**:
  - Trên cả hai thước đo, yếu tố này đều có giá trị cao nhất, vượt trội so với các biến khác. Điều này có thể do tính linh hoạt (flexibility) của cơ thể là một chỉ số quan trọng trong việc đánh giá sức khỏe và hiệu quả tập luyện.
  - Trong thực tế, khả năng gập người thường phản ánh rõ mức độ dẻo dai và khả năng vận động của cơ thể.

Nhóm yếu tố quan trọng cao:

- **sit_ups_counts**:
  - Đây là bài tập cơ bản giúp kiểm tra sức mạnh cơ bụng, một phần quan trọng trong hiệu quả tập luyện. Việc biến này có giá trị cao ở cả hai thước đo cho thấy nó là một yếu tố then chốt trong việc phân loại hiệu suất.
- **age**:
  - Tuổi tác thường liên quan chặt chẽ đến sức khỏe thể chất. Người trẻ thường có thể lực tốt hơn, trong khi tuổi cao có thể ảnh hưởng đến sức bền, sự linh hoạt và hiệu quả luyện tập.
- **fitness_score**:

- Đây có thể là biến tổng hợp hoặc điểm số phản ánh nhiều khía cạnh của thể chất. Một yếu tố tổng quát như thế này có giá trị cao là hợp lý.

Nhóm yếu tố quan trọng trung bình:

- **body_fat, bmi, grip_force, weight_kg, height_cm, broad_jump_cm**:
  - Các yếu tố này có giá trị MDA và MDG trung bình. Tuy nhiên, chúng vẫn quan trọng vì đều có liên quan trực tiếp đến khả năng vận động, sức mạnh và sức bền của cơ thể.
  - Ví dụ: **grip_force** phản ánh sức mạnh cơ tay, còn **broad_jump_cm** đánh giá sức mạnh và sự bùng nổ của cơ chân.

Nhóm yếu tố quan trọng thấp:

- **systolic, diastolic, pulse_pressure, blood_pressure_category, gender**:
  - Các yếu tố này liên quan đến huyết áp và giới tính, tuy có tác động đến sức khỏe tổng thể nhưng ít trực tiếp hơn đối với hiệu quả tập luyện cụ thể.
  - Ví dụ, huyết áp (systolic/diastolic) quan trọng hơn khi đánh giá nguy cơ bệnh tim mạch hơn là đánh giá khả năng tập luyện. Tương tự, **gender** thường chỉ ảnh hưởng gián tiếp thông qua các yếu tố khác như hormone hoặc sự khác biệt về cơ sinh học.