CARACTERÍSTICAS CORPORALES DE HOMBRES ADULTOS Y SU RELACIÓN CON LA GRASA ACUMULADA.

INTRODUCCIÓN

El dataset en el que se basa este informe, recoge medidas de la circunferencia de varias partes del cuerpo de 252 hombres, junto con su altura, su peso, su densidad (en el agua) y su porcentaje de grasa corporal.

Este conjunto de datos se recopiló con la pretensión de poder sacar conclusiones acerca de como afecta el aumento de grasa acumulada, en la estructura corporal de los hombres.

Es de resaltar que una variedad de libros sobre salud sugieren que los lectores evalúen su salud, al menos en parte, estimando su porcentaje de grasa corporal. En Bailey (1994), por ejemplo, el lector puede estimar la grasa corporal a partir de tablas utilizando su edad y varias mediciones de pliegues cutáneos obtenidas mediante el uso de un calibrador. Otros textos proporcionan ecuaciones predictivas para la grasa corporal utilizando medidas de circunferencia corporal (por ejemplo, circunferencia abdominal) y/o mediciones de pliegues cutáneos. Véase, por ejemplo, Behnke y Wilmore (1974), pp. 66-67; Wilmore (1976), p. 247; o Katch y McArdle (1977), pp. 120-132).

Para finalizar esta introducción, resulta relevante mencionar que porcentajes de grasa corporal altos, llevan a la persona a padecer sobrepeso y obesidad, condiciones que acarrean riesgos severos para la salud como presión arterial alta (hipertensión), colesterol LDL alto, colesterol HDL bajo o niveles altos de triglicéridos (dislipidemia), diabetes tipo 2, enfermedad coronaria, derrame cerebral o apnea del sueño y problemas respiratorios entre otros, que dan lugar a una baja calidad de vida y reducción del bienestar.

Se considera que un hombre de **entre 20 y 39 años** tendrá sobrepeso si su porcentaje de grasa corporal está entre el 20 y 25% y será obeso si el porcentaje supera el 25%. Pero si su edad está **entre 40 y 59 años,** el sobrepeso se da con un 22-28% y la obesidad con cifras superiores al 28%. Y para los mas mayores, por encima de los **65 años,** el sobrepeso está entre el 25 y 30%, mientras que obesidad en porcentajes superiores al 30% de grasa corporal.

Por otro lado, la cuantía mínima de grasa corporal para la vida humana es de entorno al 3%. Es el porcentaje mínimo necesario para lograr realizar funciones vitales del organismo.

Además, es destacable mencionar que normalmente las disciplinas deportivas tienen unos estándares de porcentaje de grasa corporal asociados. Por ejemplo, en el caso del fútbol, **la media se establece en el 10%**. Un delantero como Ronaldo se sitúa por debajo de este porcentaje, lo que favorece su agilidad. No obstante, un defensa puede tener una cantidad de grasa corporal algo superior para soportar mejor los choques. Destacan sobretodo las pruebas de **resistencia extrema**, como maratones, por ser donde se encuentran los deportistas con menor porcentaje de grasa corporal, con números que se acercan al 4%. También en el culturismo o en el ciclismo. Alberto Contador, por ejemplo, en su momento de estado óptimo rondaba esta cifra.

DATASET

El conjunto de datos se llama `bodyfat.csv' y presenta 15 variables que recogen datos de 252 hombres.

Las variables recogidas son:

```
d <- read.csv('bodyfat.csv')
colnames(d)
```

```
 [1] "Density" "BodyFat" "Age"     "Weight"  "Height"  "Neck"    "Chest"
 [8] "Abdomen" "Hip"     "Thigh"   "Knee"    "Ankle"   "Biceps"  "Forearm"
[15] "Wrist"
```

i).Density: Variable cuantitativa que se refiere a la densidad del cuerpo determinada mediante un pesaje bajo el agua. Se mide en g/cm^3

ii)BodyFat: Porcentaje de grasa corporal.

iii).Age: Edad del individuo, en años.

iv).Weight: Peso del individuo, medido en libras. (lb)

v).Height: Altura en pulgadas (in)

vi).Neck: Circunferencia del Cuello. (cm)

vii). Chest: Circunferencia del Pecho. (cm)

viii).Abdomen: Circunferencia del Abdomen. (cm)

ix).Hip: Circunferencia de la cadera. (cm)

x).Thigh: Circunferencia del Muslo. (cm)

xi).Knee: Circunferencia de la Rodilla. (cm)

xii). Anckle: Circunferencia del Tobillo. (cm)

xiii).Biceps: Circunferencia del Biceps. (cm)

xiv).Forearm: Circunferencia del Antebrazo. (cm)

xv).Wrist: Circunferencia de la Muñeca. (cm)

El conjunto de datos proviene de la página web kaggle (https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset)

El dataset fue inicialmente distribuido por el Doctor A. Garth Fisher, quien dió permiso para redistribuirlo con fines no comerciales. Concretamente la persona que lo subió a la web lo obtuvo del Departamento de Matemáticas y Ciencias de la Computación de la Escuala de minas y Tecnología de Sur Dakota.

Roger W. Johnson

Department of Mathematics & Computer Science

South Dakota School of Mines & Technology

501 East St. Joseph Street

Rapid City, SD 57701

ESTUDIO INICAL DE LOS DATOS

- **Primeramente conviene visualizar un resumen de las variables** (dado que son 15 variables y puede que no se visualicen bien, lo hacemos en 3 grupos)

```
   Density         BodyFat           Age            Weight     
Min.   :0.995   Min.   : 0.00   Min.   :22.00   Min.   :118.5  
1st Qu.:1.041   1st Qu.:12.47   1st Qu.:35.75   1st Qu.:159.0  
Median :1.055   Median :19.20   Median :43.00   Median :176.5  
Mean   :1.056   Mean   :19.15   Mean   :44.88   Mean   :178.9  
3rd Qu.:1.070   3rd Qu.:25.30   3rd Qu.:54.00   3rd Qu.:197.0  
Max.   :1.109   Max.   :47.50   Max.   :81.00   Max.   :363.1  
    Height     
Min.   :64.00  
1st Qu.:68.25  
Median :70.00  
Mean   :70.31  
3rd Qu.:72.25  
Max.   :77.75  

     Neck            Chest           Abdomen            Hip       
Min.   :31.10   Min.   : 79.30   Min.   : 69.40   Min.   : 85.0  
```

```
 1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5
 Median :38.00   Median : 99.65   Median : 90.95   Median : 99.3
 Mean   :37.99   Mean   :100.82   Mean   : 92.56   Mean   : 99.9
 3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5
 Max.   :51.20   Max.   :136.20   Max.   :148.10   Max.   :147.7
     Thigh
 Min.   :47.20
 1st Qu.:56.00
 Median :59.00
 Mean   :59.41
 3rd Qu.:62.35
 Max.   :87.30
      Knee           Ankle          Biceps         Forearm          Wrist
 Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00   Min.   :15.80
 1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30   1st Qu.:17.60
 Median :38.50   Median :22.8   Median :32.05   Median :28.70   Median :18.30
 Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66   Mean   :18.23
 3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00   3rd Qu.:18.80
 Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90   Max.   :21.40
```

Resulta peculiar ver que en el resumen de la variable BodyFat, (porcentaje de grasa corporal), haya individuos con un porcentaje de grasa igual a 0, cuando en la busqueda por internet vimos que la cuantía mínima de grasa corporal para la vida humana es del 3%.

```r
sum(d$BodyFat < 3)
```

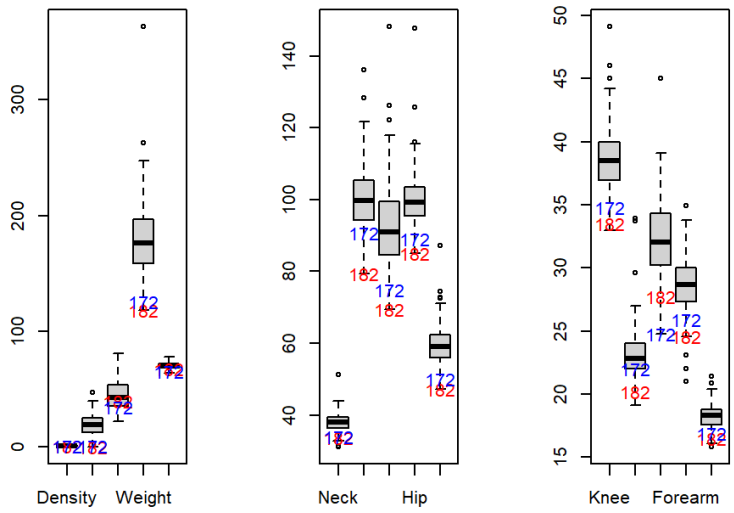
```
[1] 2
```

```
    Density BodyFat Age Weight Height Neck Chest Abdomen  Hip Thigh Knee Ankle
172  1.0983     0.7  35 125.75   65.5 34.0  90.8    75.0 89.2  50.0 34.8  22.0
182  1.1089     0.0  40 118.50   68.0 33.8  79.3    69.4 85.0  47.2 33.5  20.2
    Biceps Forearm Wrist
172   24.8    25.9  16.9
182   27.7    24.6  16.5
```

Concretamente hay dos individuos para los que el porcentaje de grasa corporal, cae por debajo del mínimo necesario para la vida. Estas observaciones atípicas son las de los individuos 172 y 182, con porcentajes de grasa corporal del 0.7 y del 0 % respectivamente. Leyendo en la propia página Kaggle, de la que proviene el conjunto de datos, resulta que muy probablemente haya una medición errónea en la densidad corporal de estos dos individuos. Por ahora solo los tendremos en cuenta, pero no los eliminamos ya que nos pueden servir en pasos posteriores.

Entre otros aspectos a resaltar del resumen de las variables, cabe mencionar que la edad de los participantes en este estudio va desde los 22 años a los 89. Esto, como se menciona en la introducción, es algo relevante a tener en cuenta, ya que conforme la edad aumenta es común aumentar también el porcentaje de grasa.

- **Visualizando las variables en gráficos de cajas y bigotes:**

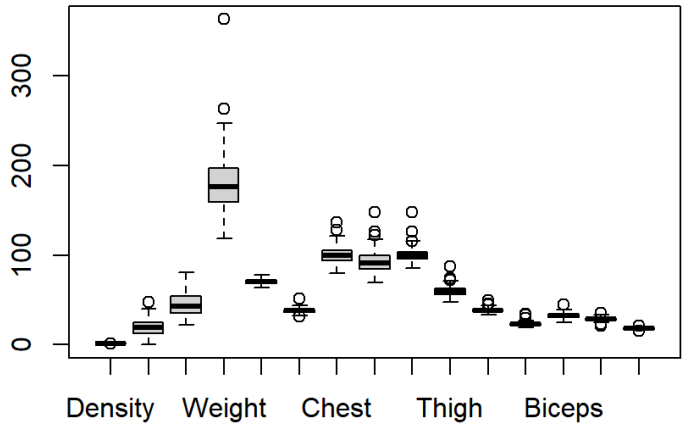


Nótese que los individuos con valores anormales del porcentaje de grasa corporal (182 y 172) no destacan realmente en el resto de variables, aunque si que es cierto que presentan valores bajos, en casi todas ellas. Esto nos indica que son individuos con un tamaño corporal pequeño, es decir no son corpulentos.

Se aprecian también valores atípicos, la gran mayoría por exceso, esto no significa que tengamos datos erróneos, sino que nos informa de la presencia de individuos con partes del cuerpo bastante grandes. En principio los atípicos al no ser muchos, no deben de suponer ningún problema para realizar las diferentes técnicas de análisis que emplearemos.

- **Escalas de las variables**

  Aunque en el apartado anterior ya se apreciaba como algunas variables podrían no pertenecer a la misma escala, conviene visualizar el gráfico de cajas y bigotes conjuntamente para todas las variables.



A simple vista para las variables más hacia la derecha, se aprecia como sí podrían estan dentro de la misma escala. Pero mirando el total del conjunto, se diferencia bastante que tienen escalas distintas.

Por ejemplo, como se muestra en el resumen del apartado anterior, la densidad corporal toma valores en un rango entre 0.995 y 1.109 g/cm3, mientras que de media la circunferencia del pecho es de 100.82 cm, o la de la muñeca es 18.23 cm

Además, si calculamos la desviación estandar de las variables, vemos como las observaciones, se desvían de forma bastante distinta respecto a sus medias, obviando que tienen escalas distintas.

```
desviaciones_estandar <- apply(d, 2, sd)
desviaciones_estandar
```

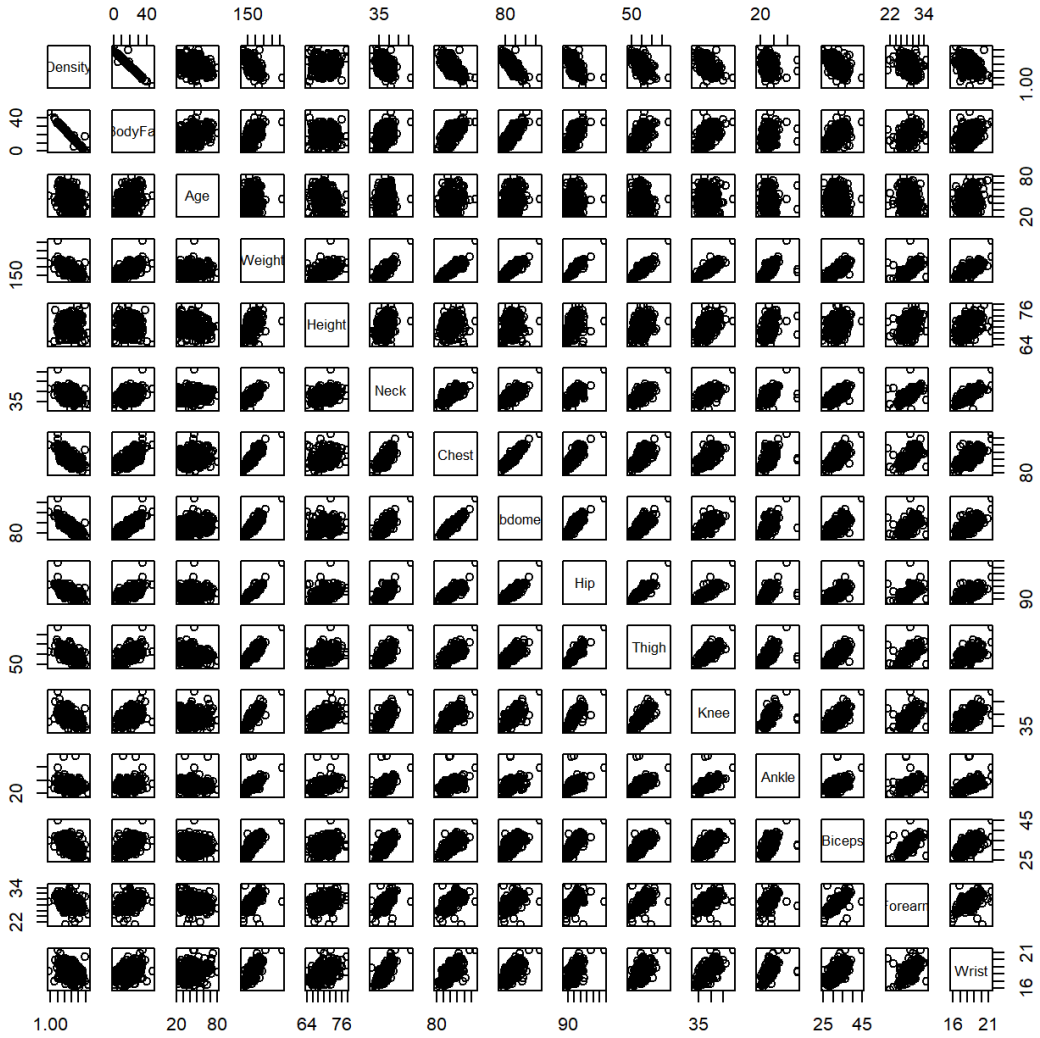
```
   Density     BodyFat         Age      Weight      Height        Neck 
0.01903143  8.36874041 12.60203972 29.38915989  2.60958292  2.43091323 
     Chest     Abdomen         Hip       Thigh        Knee       Ankle 
8.43047553 10.78307680  7.16405767  5.24995203  2.41180459  1.69489340 
    Biceps     Forearm       Wrist 
3.02127375  2.02069117  0.93358493 
```

Por tanto es conveniente estandarizarlas antes de trabajar con ellas.

La estandarización de variables es una práctica bastante común en el análisis de datos que puede mejorar la interpretación, la estabilidad numérica y la comparabilidad de los modelos estadísticos y de aprendizaje automático.

- **Relación entre variables:**

  Resulta conveniente poder conocer si las variables se encuentran bien relacionadas entre sí, o si por el contrario son más independientes.

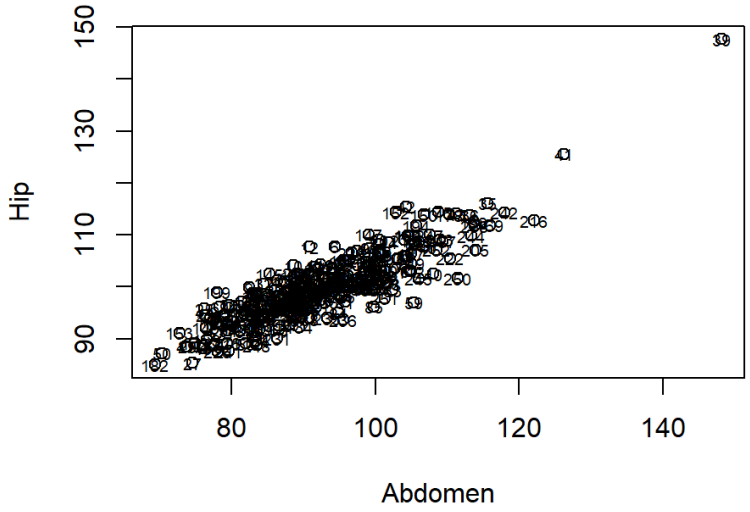


A primera vista ya vemos como la densidad y la grasa corporal son variables inversamente relacionadas con una correlación notablemente fuerte. También se aprecia cierta relación lineal entre algunas partes del cuerpo, lo que es entendible al ser partes del cuerpo de una misma persona. Sería extraño que un individuo presentase algunas partes del cuerpo anormalmente grandes frente a otras muy pequeñas.

Veamos mas de cerca para algunas variables en concreto:

a).Variables de las medidas corporales:

a1. Abdomen y Cintura

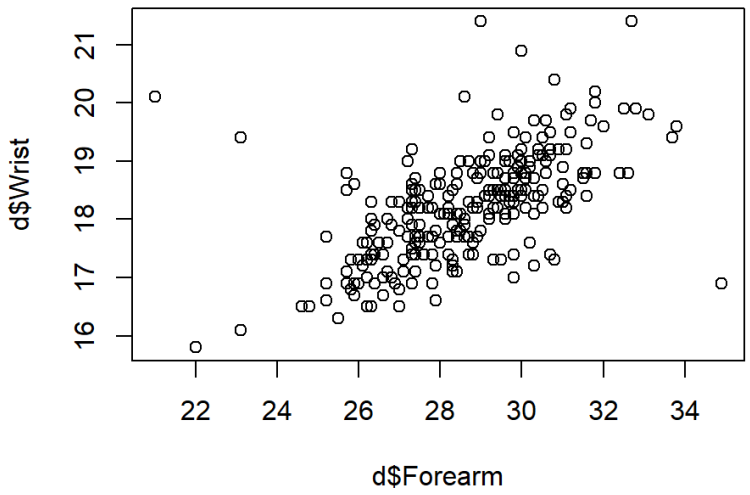



```
cor(d$Hip, d$Abdomen)
```

```
[1] 0.8740662
```

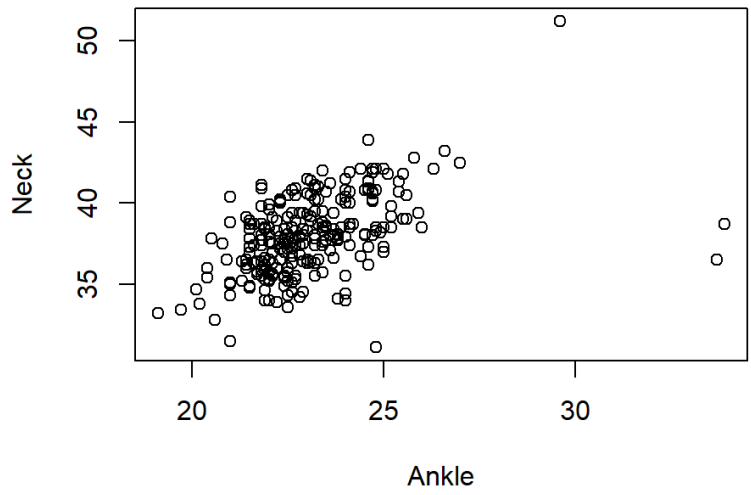
Se aprecia una fuerte correlación entre la circunferencia del abdomen y la cintura, esto queda explicado claramente al ser partes del cuerpo muy cercanas. Tendrían cuerpos muy descompensados y anormales si presentasen tamaños muy distintos de la cadera y el abdomen.

a2. Antebrazo y Muñeca



En cuanto al antebrazo y la muñeca, la relación sigue siendo notablemente visible, aunque ya no se aprecia tan directa como en el caso del abdomen y la cintura por ejemplo.

a3. Tobillo y Cuello



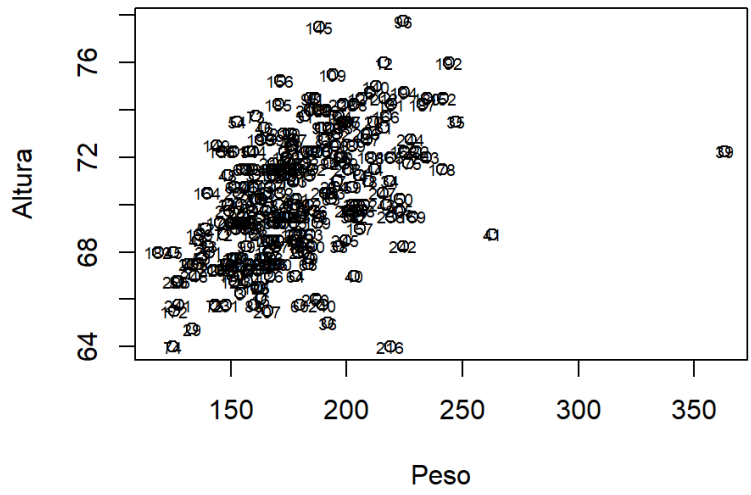

```
cor(d$Ankle, d$Neck)
```

```
[1] 0.4778924
```

En el caso del cuello y el tobillo al ser partes del cuerpo bastante distanciadas, la relación no es tan fuerte.

En general las variables que miden el tamaño de partes del cuerpo suelen estar bastante bien relacionadas. Lo cual era de esperar, ya que las partes del cuerpo de una misma persona, salvo por anomalías o casos extremos, tienden a presentan una cierta armonía. Y más aún si son partes del cuerpo pertenecientes a zonas próximas.

b).Peso y Altura:




```
cor(d$Weight, d$Height)
```

```
[1] 0.486888
```

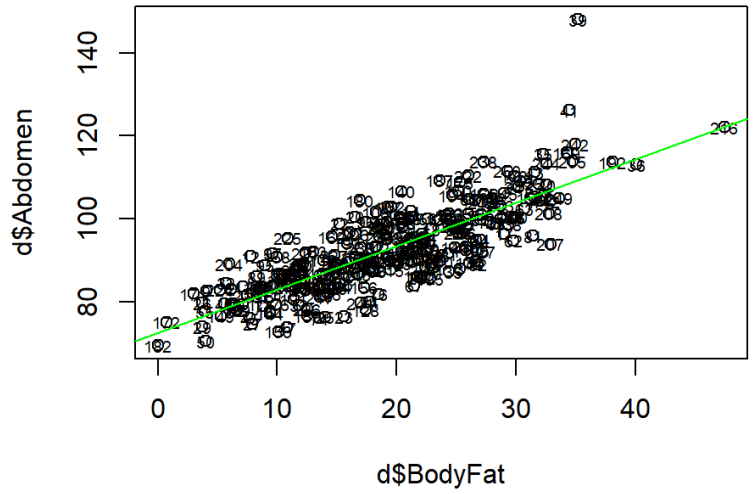
La altura y el peso parecen no guardar una relación tan significativa como la que presentan entre sí las variables que miden el tamaño de miembros del cuerpo. Podemos apreciar como un aumento en el peso no se traduce necesariamente en un aumento de altura.

Destaca sobretodo el individuo 216 por tener un mayor peso que la gran mayoría de hombres del estudio, pero siendo una persona de poca estatura.

```
    Weight Height
216    219     64
```

Mas concretamente este individuo (216) es, el más bajo y pesa más que el 75% de los demás hombres del estudio.

c).Grasa Corporal y Tamaño del Abdomen



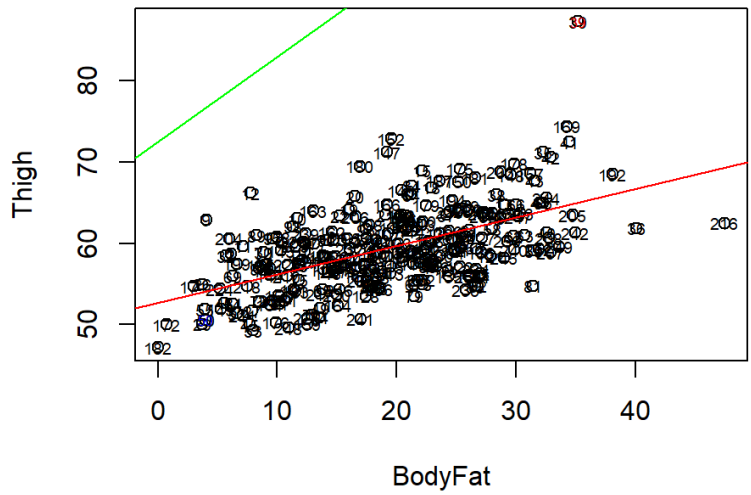

```
cor(d$BodyFat, d$Abdomen)
```

```
[1] 0.8134323
```

Esta fuerte correlación entre el aumento de la grasa corporal y el aumento de la circunferencia del abdomen, muy probablemente se deba a que es precisamente en esa parte del cuerpo donde tiende a acumularse la grasa corporal en los hombres.

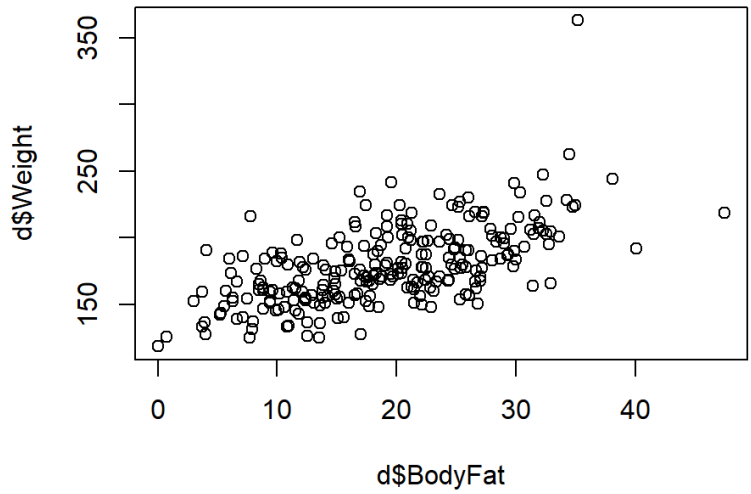
Si visualizamos la relación entre la grasa corporal y cualquier otra parte del cuerpo, por ejemplo los muslos, vemos como la relación es efectivamente más debil que con el abdomen. Indicando que se agrupa mas tejido graso en el abdomen que en los muslos.

(La línea verde representa la relación entre la grasa corporal y el tamaño abdomen, mientras que la roja la relación entre la grasa corporal y el tamaño del muslo, para así evidenciar que la primera es más fuerte que al segunda)



En ambas gráficas destaca el individuo 39 (en rojo), por tener un alto porcentaje de de grasa corporal, pero sobretodo por tener un tamaño corporal bastante por encima del resto de participantes del estudio. Y como por el contrario, obviando a los individuos 182 y 172 que eran los que tenían una medición de la grasa corporal errónea, el 50 (en azul) es uno de los que menos grasa tiene y menor tamaño corporal.

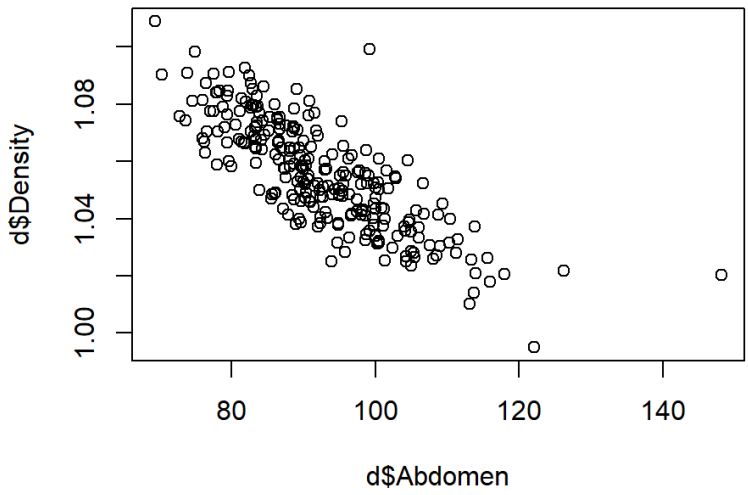
d).Grasa Corporal y Peso:



El porcentaje de grasa corporal y el peso, si que guardan relación, aunque no es tan fuerte como en otras variables. Esto puede que se explique ya que, el tejido adiposo que compone la grasa corporal no es tan pesado (0,9 kg/L) como el tejido muscular (1,06 kg/L). Por lo que un aumento de grasa corporal, sí que supondrá un aumento de peso del individuo, pero no tan significativo como un aumento de tejido muscular por ejemplo.

(fuente densidad de tejidos : https://es.wikipedia.org/wiki/Tejido_adiposo#:\%7B.uri%7D~:text=El%20tejido%20adiposo%20tiene%20una,%2C06%20kg%2FL).

e). Densidad y tamaño del abdomen

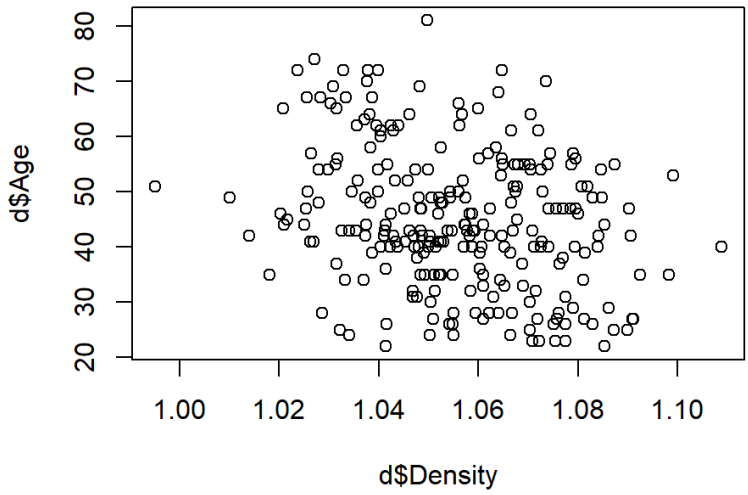


La firme relación inversa entre la desidad corporal y el tamaño del abdomen puede tener su explicación en unas características anteriormente mencionadas. En un apartado anterior se pode de manifiesto que la fuerte relación entre el tamaño del abdomen y el porcentaje de grasa corporal, se debe a que principalmente la grasa en hombres se acumule en esa zona del cuerpo. Teniendo en cuenta que el tejido adiposo, que conforma la grasa corporal, es menos pesado que el tejido muscular y óseo. Esto da lugar a que un aumento de la grasa corporal, influya proporcionalmente más en un aumento del tamaño del abdomen, que en el peso del individuo.

Siendo la densidad, una magnitud referida a la cantidad de masa en un determinado volumen (m/V), aquellos hombres con mayor porcentaje de grasa corporal presentarán un mayor volumen en proporción a su masa total, que los que preseten un menos grasa corporal.

Lo que se traduce en que, un mayor tamaño del abdomen implica más espacio ocupado en proporción al peso, y por lo tanto menor densidad.

f).Densidad Corporal y Edad



También hay variables que son bastante idependientes entre sí como es el caso de la edad con la desidad del cuerpo bajo el agua.

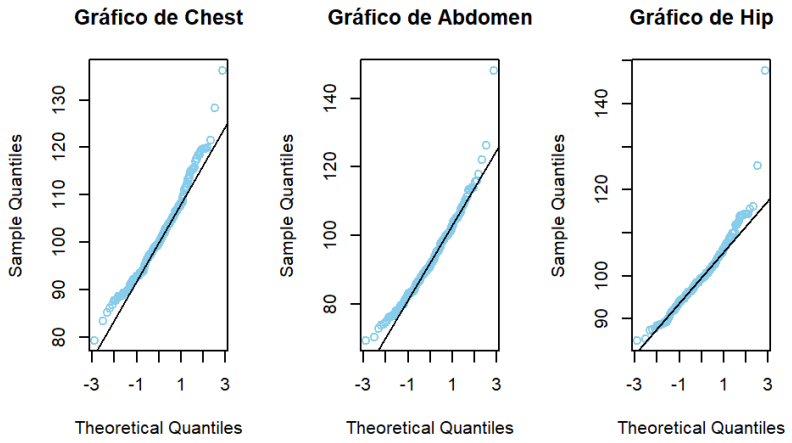
En concluisón, vemos como las variables de las partes del cuerpo, están bastante correladas, lo que puede ser una ventaja o un inconveniente. Debemos tener en cuenta que si se construye un modelo predictivo, hay que tener cuidado al usar

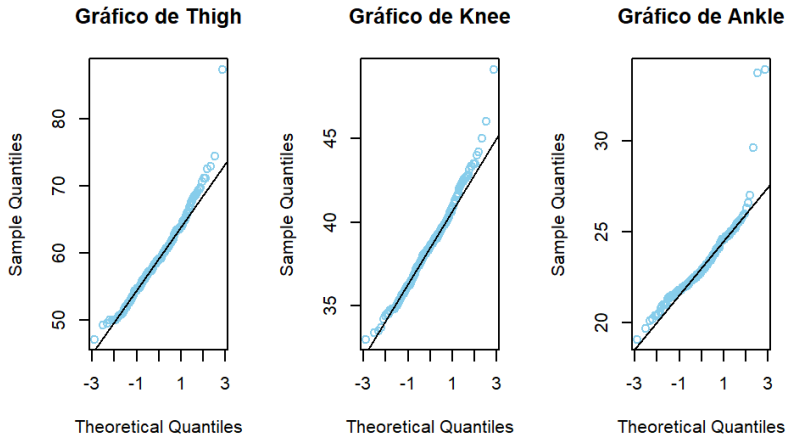
variables altamente correlacionadas, ya que pueden llevar a problemas de multicolinealidad.

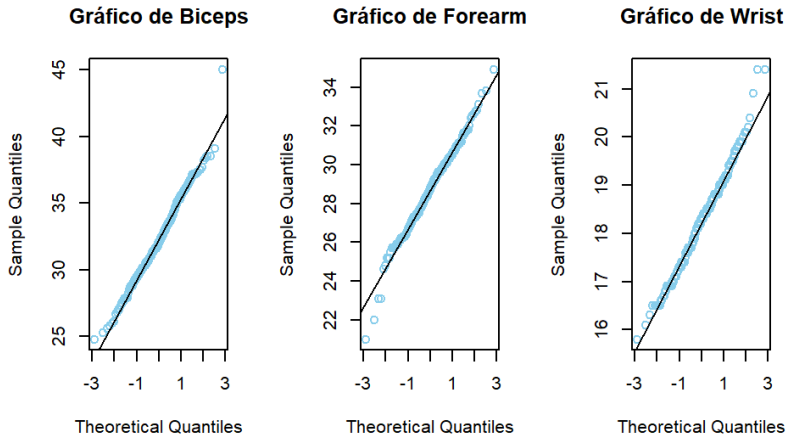
- **Distribución que siguen las variables**

Podemos visualizar las variables, con qqnorm() para intentar ver si presentan una distribución normal. Que los datos presenten o no una distribución normal, puede resultar significativo en apartados siguientes de este informe. Si los datos provienen de una distribución normal, los puntos en el gráfico Q-Q estarán aproximadamente a lo largo de la línea recta.

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.4
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.4.4     ✓ tibble    3.2.1
✓ lubridate 1.9.3     ✓ tidyr     1.3.0
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```







ANÁLISIS DE LAS COMPONENTES PRINCIPALES

El Análisis de Componentes Principales (PCA) se emplea para resumir la información contenida en muchas variables aleatorias (relacionadas) y las principales características de sus individuos, en unas pocas variables denominadas componentes principales.

Durante el estudio inicial ya vimos como había algunas variables que no solo empleaban unidades distintas, sino que sus valores se movían en rangos diferentes al resto, es decir que presentaban escalas distintas. A la hora de aplicar un PCA, para evitar que tengan más importancia las variables con escalas mayores, es conveniente que los datos se estandaricen, usando la matriz de correlaciones. De esta forma todas tienen a priori la misma importancia (varianza uno).

Además, para evitar que los individuos con valores de su porcentaje de grasa corporal inferior al mínimo vital (172 y 182), afecten a las interpretaciones de las componentes, es conveniente proseguir el estudio sin ellos.

```
data <- d[-c(172,182), ]
#Para aplicar la matriz de correlaciones el parámetro cor debe tomar valor TRUE
PCA <-princomp(data, cor = TRUE)
summary(PCA)
```

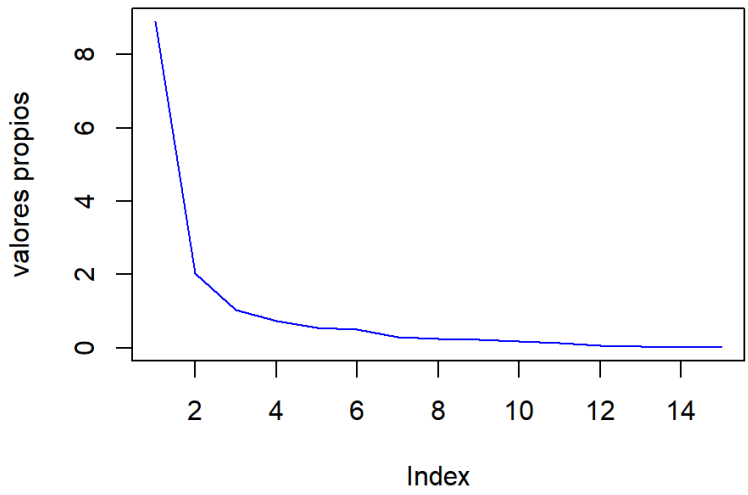
```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.9829117 1.4250073 1.02025714 0.85648885 0.74891909
Proportion of Variance 0.5931842 0.1353764 0.06939498 0.04890488 0.03739199
Cumulative Proportion  0.5931842 0.7285606 0.79795554 0.84686041 0.88425240
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.70197986 0.54660564 0.50588259 0.47113142 0.43024602
Proportion of Variance 0.03285171 0.01991852 0.01706115 0.01479765 0.01234078
Cumulative Proportion  0.91710411 0.93702263 0.95408378 0.96888143 0.98122221
                           Comp.11     Comp.12     Comp.13     Comp.14
Standard deviation     0.367220531 0.272082302 0.207743256 0.135441922
Proportion of Variance 0.008990061 0.004935252 0.002877151 0.001222968
Cumulative Proportion  0.990212267 0.995147519 0.998024670 0.999247637
                           Comp.15
Standard deviation     0.1062329662
Proportion of Variance 0.0007523629
Cumulative Proportion  1.0000000000
```

- **Selección de las Componentes Principales**

La importancia de las componentes se mide con las proporciones en tanto por uno de sus varianzas (proportion of Variance) y las proporciones acumuladas.

Para seleccionar cuantas componentes debemos tomar, podemos atender a distintos criterios. Uno de los más comunes es determinar un porcentaje mínimo de variabilidad de la información inicial, que deban abarcar las componentes, y quedarnos con número de componentes suficiente que permita alcanzar ese porcentaje mínimo. Para este caso, un porcentaje de entre el 75 y 80% parece una opción bastante apropiada. Esto sería tomar 3 componentes principales.

Otro critero que se conoce como uno de los más comunes en estadística es la Regla del codo. Esta regla establece que serán representativas las componentes hasta el primer "codo" (sin incluirlo) de la gráfica o hasta que comience la línea recta aproximada final.



Al visualizar la gráfica vemos claramente un cambio de la tendencia de la pendiente en el punto 2, y otro no tan drástico en torno al punto 3. Lo cual nos está indicando que deberíamos de quedarnos con 1 o 2 componentes.

Un último criterio que empleamos para desempatar es la regla de Kaiser, la cual nos dice que solo serán relevantes las componentes que tengan una variabilidad mayor que la variabilidad media de las variables originales. Al haber usado la matriz de correlaciones, como es equivalente a usar las variables estandarizadas, se entiende que las varianzas iniciales son 1. Entonces solo debemos coger aquellas componentes con una desviación por encima de 1, que en este caso son las 3 primeras.

- **Interpretación de las componentes principales:**

Para dar una interpretación de las componentes debemos mirar las cargas o loadings, que son los vectores propios normalizados. Estos actúan como los coeficientes de las variables iniciales en la expresión que nos permite calcular la componente principal. Por lo tanto como el valor de la componente depende en gran medida de estos coeficientes, los debemos analizar para poder darles un significado.

i).Primera componente

```
    Density     BodyFat         Age      Weight      Height        Neck
 0.22867862 -0.23436462 -0.01998217 -0.32582490 -0.12824794 -0.28364276
      Chest     Abdomen         Hip       Thigh        Knee       Ankle
-0.30477885 -0.30716876 -0.31048270 -0.29231650 -0.28897020 -0.20827228
     Biceps     Forearm       Wrist
-0.27886091 -0.22751961 -0.25261922
```

La expresión de la primera componente principal quedaría como:

Y1 = 0.223 · X1* -0.234· X2* -0.019· X3* -0.326· X4* -0.128· X5* -0.284· X6*- 0.305 · X7*- 0.307 · X8* - 0.31 · X9* -0.292 · X10* - 0.289 · X11* - 0.208 · X12* - 0.279 · X13* - 0.228 · X14*- 0.253 · X15*

En esta componente la variable Edad (Age) no se encuentra bien representada, dado que tiene un valor muy próximo a 0. También nos encontramos con que la variable Altura (Height) tiene una menor relevancia dentro de la componente que el resto de variables (valor absoluto de 0.128, frente a más de 0.200 del resto)

Destaca que las variables de circunferencia corporal (como la del abdomen, cadera, pecho, etc.) tengan una importancia similar dentro de la componente, estando mas o menos igualmente representadas y además están asociadas (negativamente) con la grasa corporal.

La densidad corporal es la única variable que presenta un valor positivo, es decir, esta componente primera es directamente proporcional al aumento de la densidad, mientras que inversamente proporcional al crecimiento del resto de variables como son el aumento de las partes del cuerpo, y el porcentaje de grasa.

Esto sugiere que esta componente es **principalmente un indicativo de las medidas corporales y la densidad y a su vez de como influye la distribución de la grasa corporal en el tamaño de partes del cuerpo.**

Entonces presentará valores pequeños cuanto más corpulento sea el indivividuo, y valores grandes cuanto menor tamaño corporal y menos porcentaje de grasa tenga (porque será más denso).

ii).Segunda Componente

```
    Density     BodyFat         Age      Weight      Height        Neck
 0.43571583 -0.43181978 -0.45715038  0.07818553  0.46272675  0.03852264
      Chest     Abdomen         Hip       Thigh        Knee       Ankle
-0.12508748 -0.21154452  0.03289539  0.09983137  0.12567210  0.24129140
     Biceps     Forearm       Wrist
 0.08940282  0.15797226  0.12072463
```

En esta segunda componente principal no están bien representadas ni el Peso (Weight), ni el tamaño del Cuello (Neck), Cadera (Hip) y Biceps. Además, el resto de variables relacionadas con la circunferencia de partes del cuerpo, no presentan una influencia excesivamente significativa.

Mientras que, las variables con los loadings más altos y por tanto fuertemente relacionadas con la componente son, la Densidad corporal, la Altura, el porcentaje de Grasa corporal y la Edad. Teniendo la Densidad y la Altura coeficientes positivos y la Grasa y la Edad, coeficientes negativos, además de encontrarse prácticamente igual representadas.

Cabe destacar que la Densidad y Grasa Corporal, están estrechamente relacionadas entre sí, como se expone en apartados anteriores, dado que un aumento de la grasa corporal supone un cambio más significativo en el volumen del cuerpo, que en el peso total de este, propiciando disminución de la densidad. O viéndolo desde otro enfoque, una alta densidad indica una mayor proporción de masa magra en comparación con la grasa corporal. Esto concuerda con que estén contrarrestándose entre sí en la componente (distinto signo).

Por otro lado, la edad y la altura son factores que influyen en la composición y estructura corporal. La edad podría representar el efecto del envejecimiento en el cuerpo, como señalábamos en la introducción. Y la altura (comúmente usada en otros campos como un indicador de la estructura corporal), podría estar relacionada con la distribución de la masa corporal, de forma que una altura mayor, pueda significar una distribución diferente de la grasa y la masa magra, dando lugar a tamaños diferentes de las distintas partes del cuerpo.

Por lo tanto, esta segunda componente principal representa **la variabilidad en la composición corporal** (más tejido muscular o graso), **y la influencia de la edad y la altura en esta composición.**

Un valor alto de la componente indicará alta densidad, gran altura y un porcentaje de grasa corporal bajo, mientras que valores bajos, lo contrario, es decir, poca densidad y alto porcentaje de grasa corporal.

iii).Tercera Componente

```
    Density     BodyFat         Age      Weight      Height        Neck
 0.23447893 -0.21230415  0.68454730 -0.02289917  0.11309620  0.21349615
      Chest     Abdomen         Hip       Thigh        Knee       Ankle
 0.04106378 -0.03565554 -0.16389951 -0.29770763  0.02067445  0.08862080
     Biceps     Forearm       Wrist
-0.02616623  0.07843375  0.48875128
```

En esta componente las circunferencias de las partes del cuerpo no son muy significativas, estando la mayoría poco representadas al tener coeficientes cercanos a 0.

También vuelve a darse en esta componente, aunque de una forma menos destacada, la relación inversa entre la densidad corporal y el porcentaje de grasa, que explicabamos para la componente anterior.

Y sobretodo tiene una altísima relevancia la variable Edad. Por lo que esta componente **es principalmente un indicativo de la edad del varón**. De la forma que tomará valores mas altos, para aquellos individuos más longevos.

- **Puntuaciones de los Individuos para las componentes principales**

Estas puntuaciones nos permiten tener una imagen de como son los individuos, segun los valores que toman en esa componente

```
S <- PCA$scores
```

Podemos analizar, algunos de los individuos con puntuaciones más extremas en las componentes.

Por ejemplo para la componente primera, el individuo con la menor puntuación es el 39.

```
which.min(S[,1])
```

```
39
39
```

Recordemos que tener una puntuación baja en esta componente estaba relacionado con tener un gran tamaño de las partes del cuerpo, y además una gran porcentaje de acumulación de grasa.

```
   Density BodyFat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee Ankle
39  1.0202    35.2  46 363.15  72.25 51.2 136.2   148.1 147.7  87.3 49.1  29.6
   Biceps Forearm Wrist
39     45      29  21.4
```

Efectivamente, viendo los valores de este individuo 39 en las variables inicales, denota un tamaño corporal por encima de la media. Por ejemplo, la cintura de media tiene 99.9 cm de circunferencia y la de este hombre son 147.7 cm, o su peso es de 363 libras que equivale a unos 160 kg.

Por el contrario, mirando al individuo con mayor puntuación en la primera componente, el 45. Se diferencia del 39, al presentar un tamaño corporal pequeño, y bajo porcentaje de grasa, llevando a su vez a una densidad mayor.

```
which.max(S[,1])
```

```
45
45
```

```
   Density BodyFat Age Weight Height Neck Chest Abdomen  Hip Thigh Knee Ankle
45  1.0814     7.7  39 125.25     68 31.5  85.1      76 88.2    50 34.7    21
   Biceps Forearm Wrist
45   26.1    23.1  16.1
```

También podemos hacer lo propio para la componente segunda:

El individuo con la máxima puntuación en la componente seguda es:

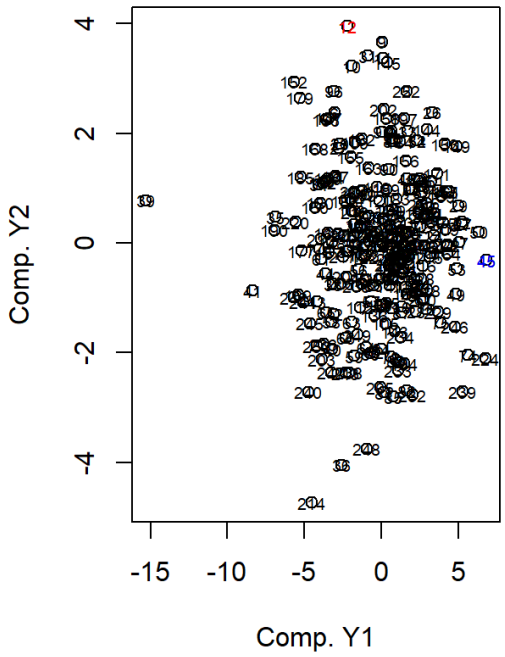
```
which.max(S[,2])
```

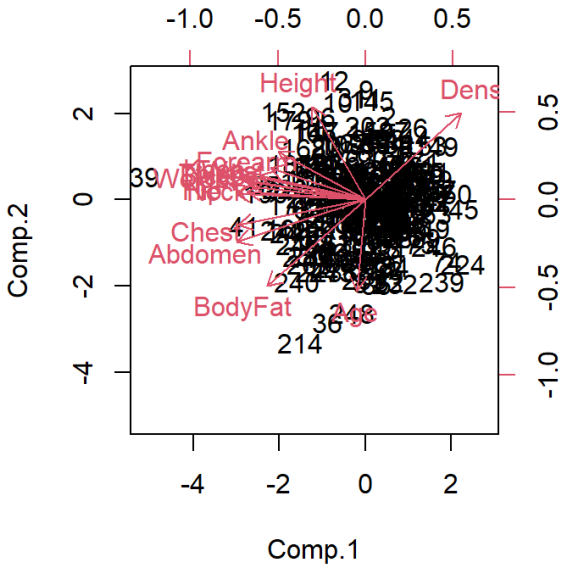
```
12
12
```

```
   Density BodyFat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee Ankle
12  1.0812     7.8  27    216     76 39.4 103.6    90.9 107.7  66.2 39.2  25.9
   Biceps Forearm Wrist
12   37.2    30.2    19
```

Tener una puntuación alta en esta segunda componente era sinónimo de una densidad alta, gran altura y bajo porcentaje de grasa corporal. Vemos como efectivamente este individuo 12 con la máxima puntuación, tiene una altura de 76 pulgadas que son 193 cm, un bajo porcentaje de grasa corporal, y además, presenta alta densidad (el 75% de los hombres del estudio tiene una densidad inferior a 1.07 g/cm3) .

Finalmente, podemos analizar los siguientes gráficos que muestran las puntuaciones de los individuos para las dos primeras componentes principales.





En el primer gráfico (izquierda), los individuos con tamaño corporal pequeño y poca acumulación de grasa se encuentran más hacia la derecha (en la componente primera), como es el caso del 45, y aquellos con mayor altura y densidad se encuentran más arriba (en la componnete segunda), como es el caso del individuo 12.

El segundo gráfico (derecha) es similar al primero, pero las variables aparecen como vectores en rojo y las puntuaciones estandarizadas como las etiquetas de los datos en negro. En este, las variables presentan vectores largos, lo que es sinónimo de que se encuentran bien representadas por las dos primeras componentes. Si tuvieran vectores cortos , es indicativo de que la información se pierde al ser proyectada, por ser casi perpendiculares, lo que sería sintoma de estar mal representadas.

Con este segundo gráfico se comprueba la interpretación hecha al gráfico anterior, al ver al individuo 45, que presentaba un tamaño corporal pequeño, en dirección contraria a los vectores de las variables que tasan el tamaño de partes del cuerpo. O al individuo 39, que se caracteriza por tener un tamaño corporal atípicamente grande, estando muy hacia la izquierda, donde confluyen la mayoría de vectores de las variables que miden el tamaño de partes del cuerpo.

Una vez tenemos las puntuaciones de todos los individuos y para todas las componentes, debemos tomar únicamente las puntuaciones de aquellas con las que elegimos quedarnos, (1,2y 3). De esta forma podríamos emplearlas, según convenga, para futuras técnicas que llevemos a cabo, en vez de utilizar las 15 variables de nuestro dataset original.

```
Puntuaciones3Comp <- data.frame(Comp1 = S[,1], Comp2 = S[,2], Comp3 = S[,3])
```

- **Saturaciones**

Para poder conocer la información total que se tiene de cada variable en las componentes debemos calcular las saturaciones al cuadrado. Estas nos indicarán cuanta información (en tanto por 1) habrá en cada componente de cada variable.

```
SAT<-cor(data,S)
```

Además para conocer el porcentaje total de cada variable entre las 3 componentes con las que nos hemos quedado, debemos calcular las comunalidades. En nuestro caso, se calculan como la suma de las saturaciones al cuadrado de las 3 componentes principales.

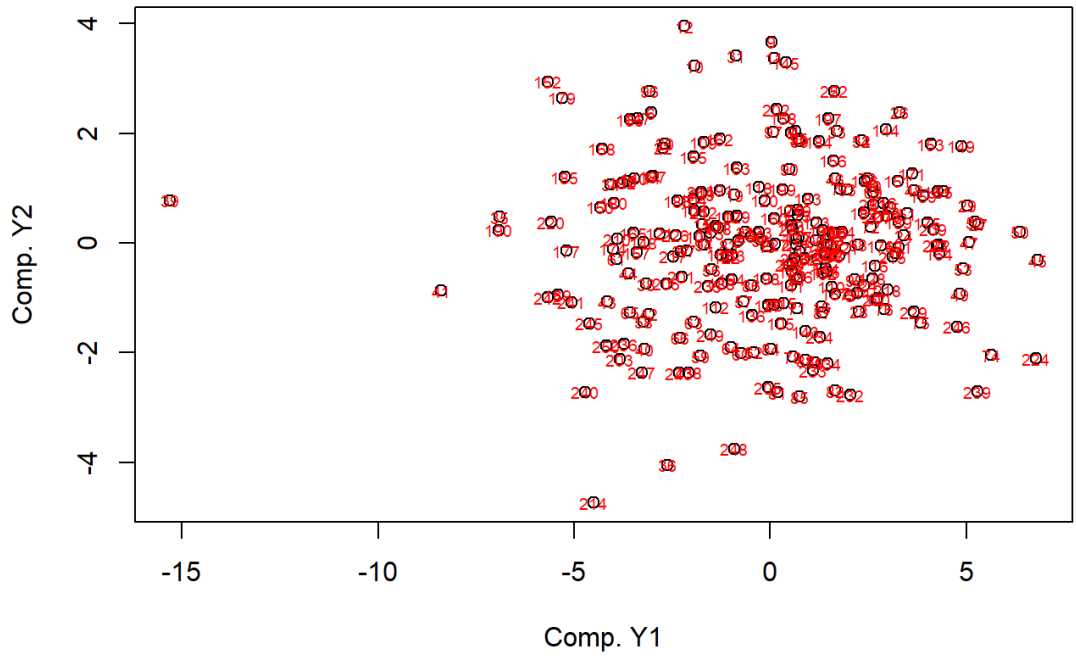
```
COMUNALIDADES
```

```
          Inf. Comp1  Inf. Comp2    Inf Comp3 Comunalidad
Density 0.465298806 0.385514637 0.0572304204   0.9080439
BodyFat 0.488725417 0.378651128 0.0469176521   0.9142942
Age     0.003552763 0.424377509 0.4877824890   0.9157128
Weight  0.944603070 0.012413291 0.0005458318   0.9575622
Height  0.146346244 0.434793869 0.0133142069   0.5944543
Neck    0.715853581 0.003013466 0.0474459751   0.7663130
Chest   0.826514458 0.031773266 0.0017552427   0.8600430
Abdomen 0.839527441 0.090873601 0.0013233460   0.9317244
Hip     0.857739928 0.002197375 0.0279624084   0.8878997
Thigh   0.760304363 0.020238029 0.0922569746   0.8727994
Knee    0.742996794 0.032070957 0.0004449254   0.7755127
Ankle   0.385961288 0.118227326 0.0081750532   0.5123637
Biceps  0.691920319 0.016230677 0.0007126916   0.7088637
Forearm 0.460594231 0.050675243 0.0064036157   0.5176731
Wrist   0.567823797 0.029595520 0.2486538023   0.8460731
```

Entre las 3 componentes, el porcentaje de información que recogen en general de cada variable es bastante alto, superando el 80%, en la mayoría de casos. Auque las variables Ankle y Forearm, son de las que peor porcentaje de información se conserva, superando levemente el 50 %.

ANÁLISIS CLÚSTER: APRENDIZAJE NO SUPERVISADO

En el gráfico que se visualizó de las puntuciones de los individuos respesto a las dos primeras componentes principales, vimos como una mayoría tendía a agruparse hacia la derecha y en torno al centro del gráfico, mientras que había otros individuos que se localizaban en posiciones más alejadas de esta agrupación principal, lo cual sugiere, que muy probablemente, podamos distinguir a los individuos de este estudio entre grupos.



Intetaremos agrupar a los individuos del dataset según la similitud de sus valores en las variables. Al no existir grupos iniciales, este tipo de técnica se denominan análisis no supervisado o aprendizaje automático. La formación de los grupos dependerá de las distancias usadas para medir las similitudes y del algoritmo de agrupación aplicado.

Existen principalmente dos algoritmos para llevar a cabo la clasificación de los individuos, K-means (análisis cluster con k-means) y h-clust (análisis cluster jerarquizado). Para el primer algoritmo es necesario indicar inicialmente el número de

grupos o clusters entre los que agrupar. Mientras que para el segundo no es necesario.

En nuestro caso, a partir del estudio anterior realizado, **no es sencillo determinar a simple vista un número concreto de grupos en el que estarían repartidos los individuos del dataset,** por lo que aplicaremos el **análisis jerarquizado por h-clust**. Con este algortimo, un dendrograma nos indicará los grupos que se van formando y nos permitirá decidir con cuántos nos quedamos.

El funcionamiento de este algoritmo se basa en calcular las distancias de cada uno de los individuos a todos los demás, e ir agrupando iteración por iteración a las observaciones que más cercanas se encuentren entre sí. Cada individuo forma un cluster inicialmente, y con que se agrupen dos observaciones, ya tendremos un nuevo cluster, que se irá ampliando con las demás observaciones o clusters similares que tengan una distancia cercana.

La idea es poder distinguir entre varias clases que se diferencien según el tipo de cuerpo del individuo y su composición/estructura.

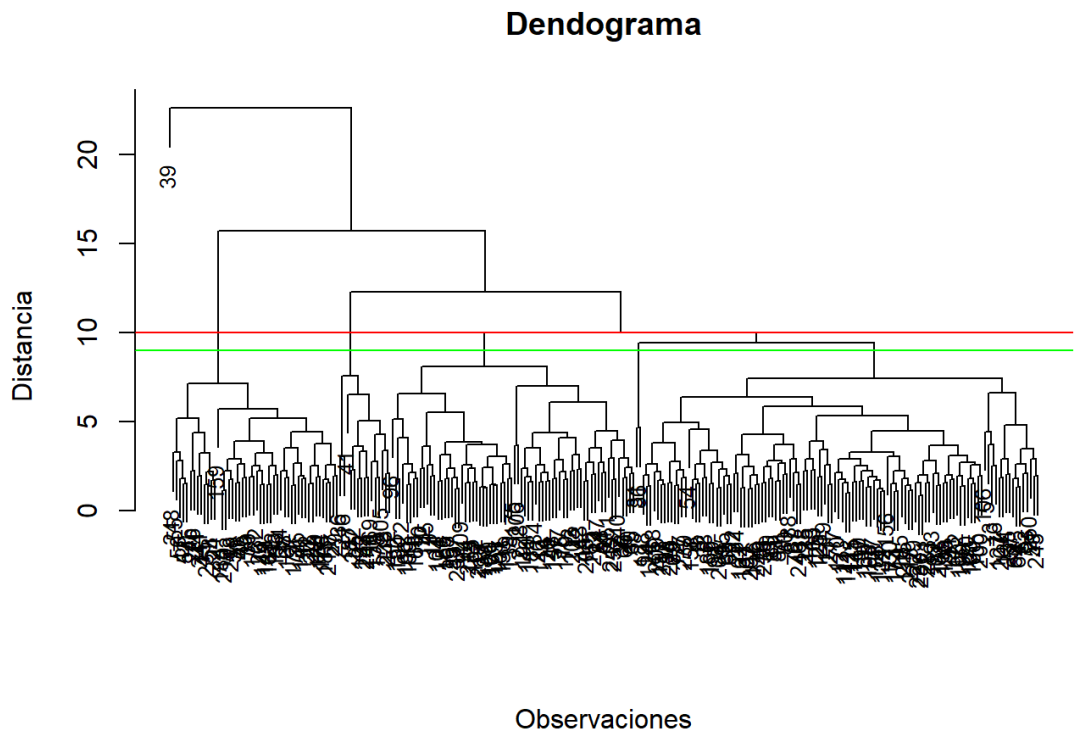
Este análisis cluster podría llevarse a cabo solo con los datos de las puntuciones de los individuos en las 3 primeras componentes pricipales que recopilamos anteriormente, sin embargo dado que el número de variables iniciales no es muy alto (15), lo haremos también con las variables originales, permitiéndonos clasificar a los individuos antendiendo al 100% de la información total. Y pudiendo comparar resultados.

1.Aplicación del análisis jerarquizado (hclust) al dataset original

```
ds <- scale(data) #Estandarizar los datos: Ya se introdujo en apartados anteriores la importancia de estandarizar
```

```
Distancias <- dist(ds, method = 'euclidean')
#Aplicamos la distancia euclídea
```

```
CA <- hclust(Distancias, method = "complete")
```



El dendograma nos muestra como se han ido formando las distintas agrupaciones, y nos permite decidir a simple vista en cuantas clases posibles podríamos dividir a los individuos del dataset.

Debido al gran número de individuos resulta dificil diferenciar a simple vista los dos individuos que primero se agruparon, pero sacándolo por medio de la matriz de distancias, vemos que se trata de los individuos 225 y 228.

Sí que está muy claro que el último en ser agrupado fue el 39, lo cual indica claramente su tamaño atípicamente grande respecto al resto de participantes.

```
M <- as.matrix(Distancias)[1:250,1:250]
min(Distancias)
```

```
[1] 0.8553559
```

```
# Encontrar la posición del valor en la matriz
posicion <- which(M == min(Distancias), arr.ind = TRUE)
# Imprimir la fila y la columna donde se encuentra el valor
print(paste("Fila:", posicion[,1], "Columna:", posicion[,2]))
```

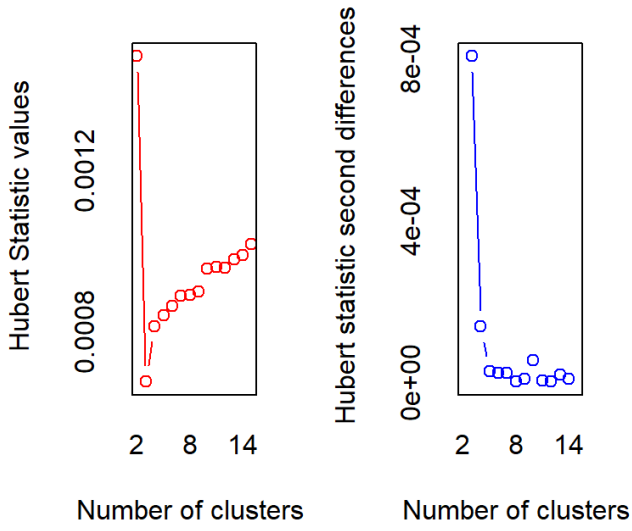
```
[1] "Fila: 228 Columna: 225" "Fila: 225 Columna: 228"
```

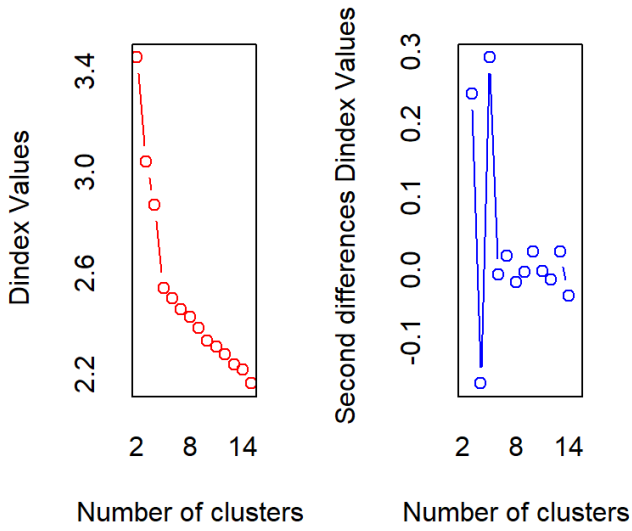
1.1Número de grupos a escoger

Mirando el dendrograma, parece una buena opción tener entre 4 y 5 grupos. Realmente no existe un número óptimo, sino que es algo que depende en gran medida de factores subjetivos, pero existen técnicas que pueden ayudarnos a decidir.

Hay índices implementados en R que nos pueden ayudar a decidirnos por un número de grupos en concreto.

```
library(NbClust)
NbClust(ds,method='complete',index='all')$Best.nc
#Escondo el resultado para evitar ocupar innecesariamente espacio
```

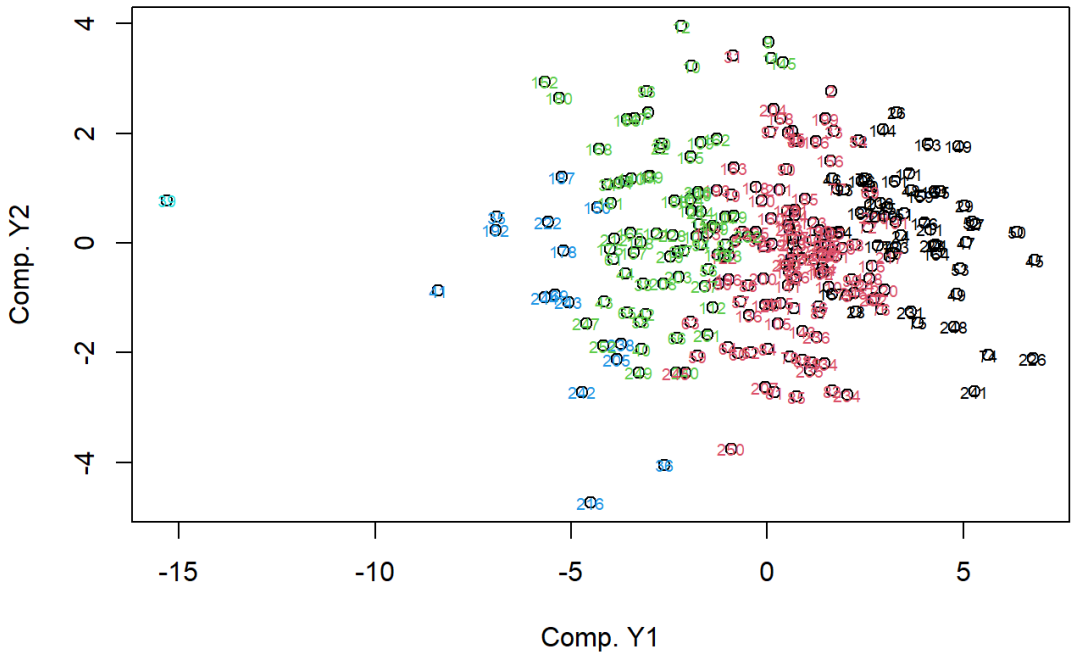




En nuestro caso, se indica que hay una mayoría de 9 índices que determinan que el mejor el número de clusters es 5, aunque también hay un segundo conjunto de 8 índices que determinan el óptimo en 2 clusters.

Agrupamos por tanto en 5 clusters.

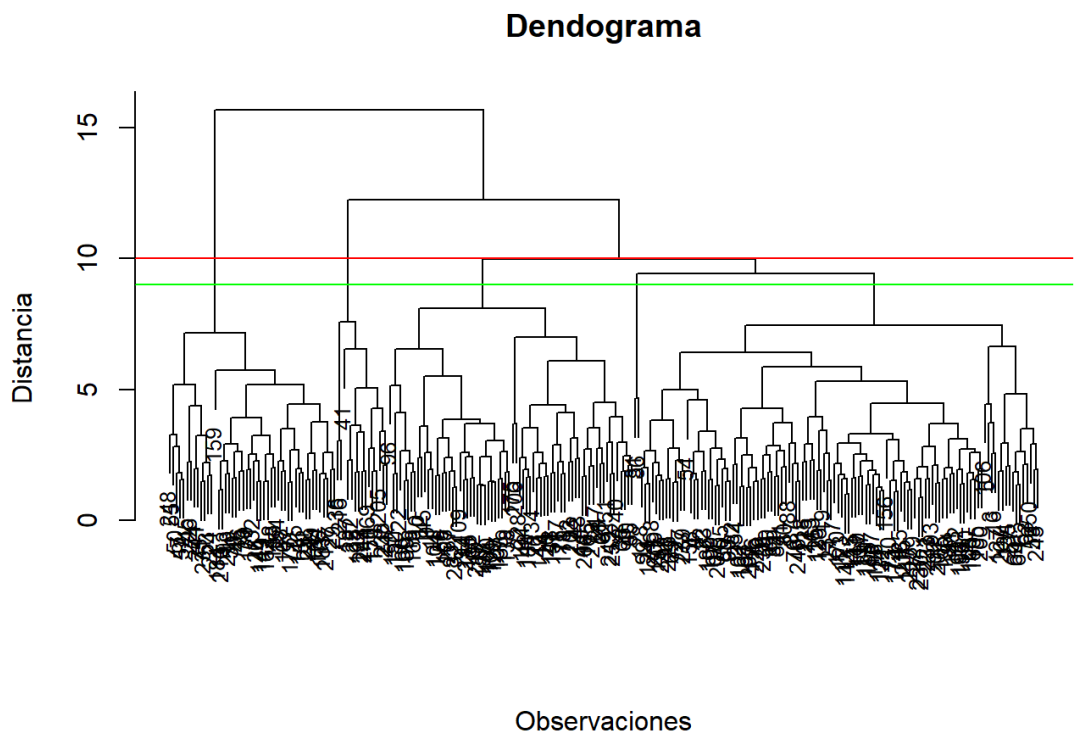
```
#Agrupando en 5 grupos
g5 <- cutree(CA, k = 5)
```



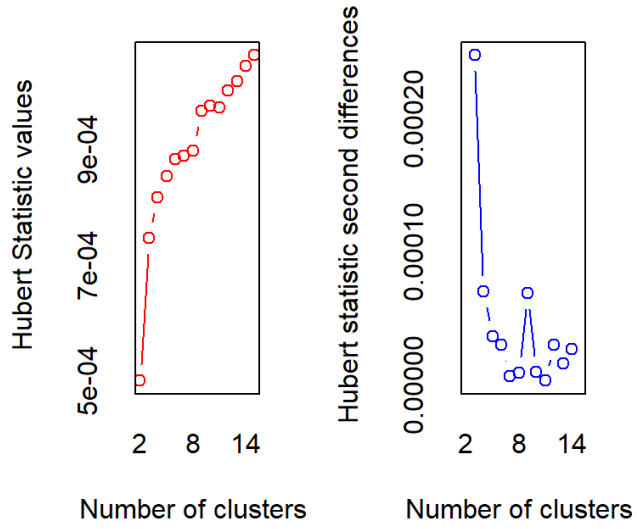
Vemos como el atípico 39, da lugar a la creación de un grupo para poder clasificarlo únicamente a él. Esto a la hora de llevar a cabo algunas técnicas como análisis discriminante, puede suponer un grave desbalanceo que afecte gravemente a las conclusiones que se obtengan. Por lo tanto debemos evaluar realmente la relevancia del atípico en el análisis, ¿Es un error de medición o un punto realmente significativo?

Realmente este indivividuo 39 no es un error de medición, sino que simplemente es alguien que presenta un tamaño corporal bastante superior a lo común. Esto es algo perfectamente posible en la realidad.

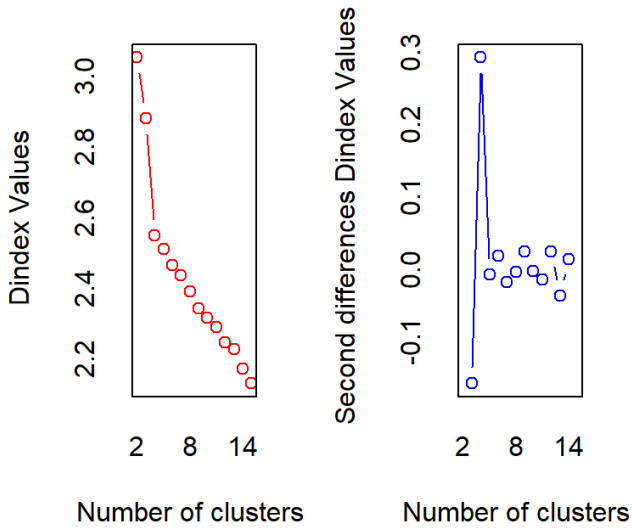
Sin embargo, este grupo de un único individuo quizás esté distorsionando los resultados, por lo que para hacer un correcto análisis cluster resulta conveniente excluirlo del dataset a la hora de clasificar. Por ello vamos a repetir el proceso sin el atípico 39.



Al eliminar el atípico 39, ya no se refleja en el dendrograma. A simple vista sigue pareciendo buena idea la clasificación entre 4 y 5 grupos.




```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.
```

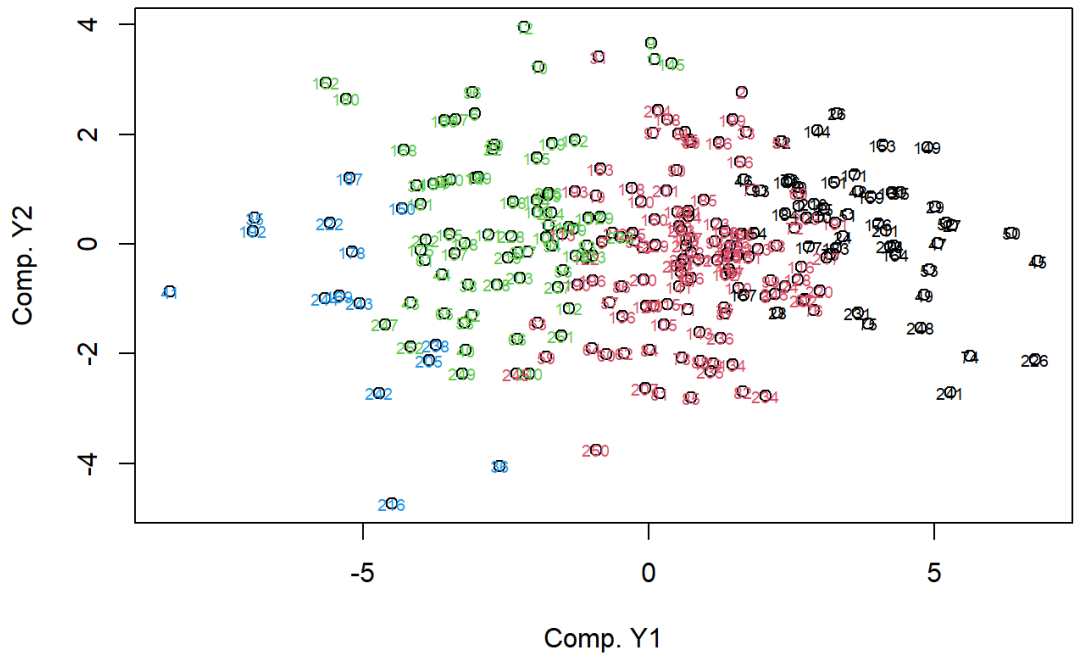



```
*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 8 proposed 4 as the best number of clusters
* 3 proposed 5 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4


*******************************************************************

                       KL      CH Hartigan     CCC    Scott      Marriot   TrCovW
Number_clusters  4.0000  2.000   4.0000  4.0000   5.0000 5.000000e+00     3.00
Value_Index     12.2447 78.632  56.7381 -7.7777 255.9707 4.392002e+26 17195.03
                  TraceW Friedman   Rubin Cindex     DB Silhouette   Duda
Number_clusters   4.0000   3.0000  4.0000 2.0000 2.0000     2.0000 4.0000
Value_Index     417.7671  30.3161 -0.3272 0.3183 1.2626     0.2186 0.8964
                PseudoT2  Beale Ratkowsky     Ball PtBiserial Frey McClain
Number_clusters   4.0000 4.0000    2.0000   3.0000     5.0000    1  2.0000
Value_Index      13.1821 1.1839    0.3412 551.5849     0.4863   NA  0.3481
                   Dunn Hubert SDindex Dindex   SDbw
Number_clusters 15.0000      0  2.0000      0 15.000
Value_Index      0.2592      0  1.3628      0  0.386
```

Sin el atípico, un mayoría de 8 índices indican 4 como el mejor número de clusters. Lo que concuerda con el resto de grupos que no estaban formados por un atípico.

Agrupando en 4 clusters:

```
g4 <- cutree(CA, k = 4)
```



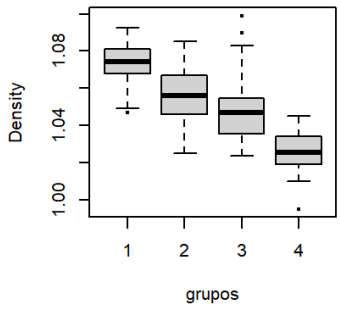
El algoritmo hclust ha hecho la misma clasificación que antes de eliminar el atípico, lo cual es lógico ya que las distancias entre los distintos individuos sigue siendo la misma a pesar de considerar un individuo menos.

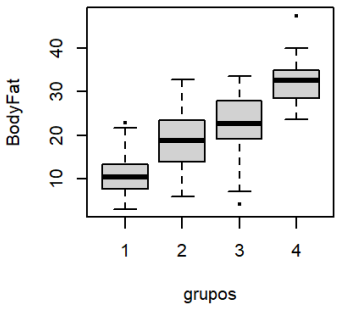
Debemos recopilar la agupación hecha como un único dataset para su posterior uso:

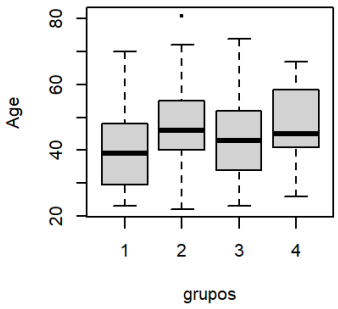
```
#Creamos dos dataset nuevos para guardar la clasificación de los datos sin el atípico 39, para evitar así elimina
dataClasified <- data.frame(data[-c(39),],g4)
dsClasified <- data.frame(ds[-c(39),],g4)
```

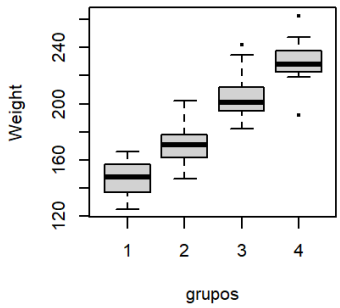
1.2 Descripción de los grupos.

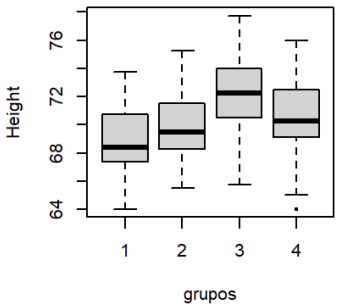
Si llevamos a cabo un breve estudio de las características de los grupos nos va a permitir describirlos.

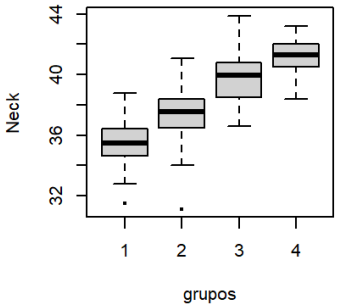

Density
1.00
1.04
1.08
1
2
3
4
grupos

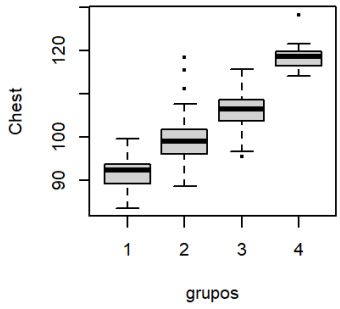

BodyFat
10
20
30
40
1
2
3
4
grupos

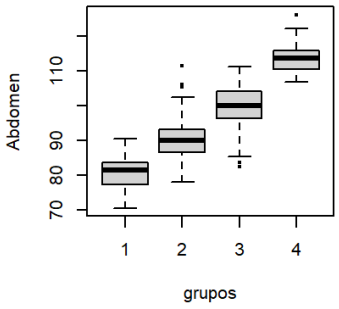

Age
20
40
60
80
1
2
3
4
grupos

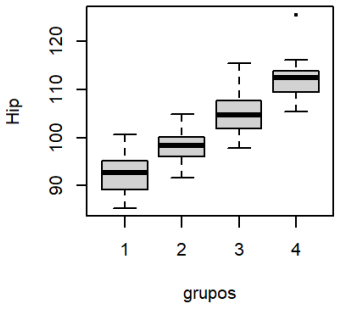

Weight
120
160
200
240
1
2
3
4
grupos

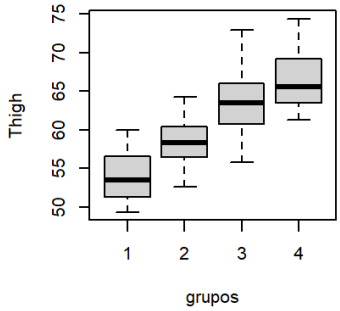

Height
64
68
72
76
1
2
3
4
grupos

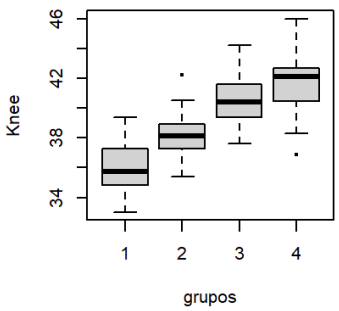

Neck
32
36
40
44
1
2
3
4
grupos

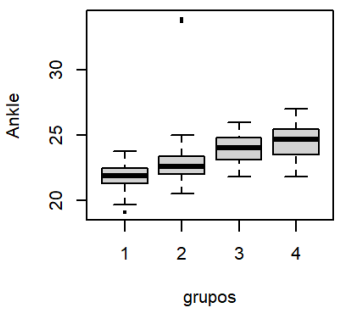

Chest
90
100
120
1
2
3
4
grupos

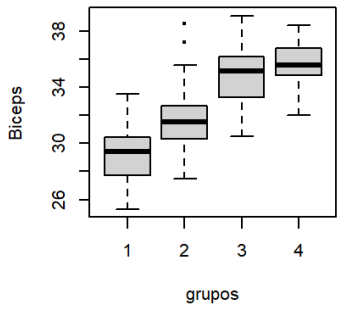

Abdomen
70
80
90
110
1
2
3
4
grupos

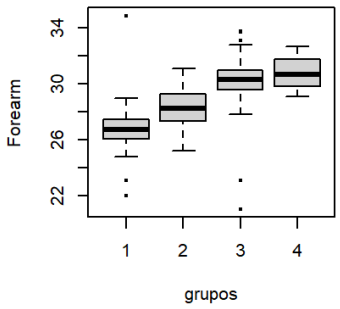

Hip
90
100
110
120
1
2
3
4
grupos

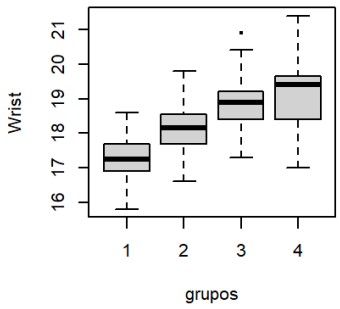

Thigh
50
55
60
65
70
75
1
2
3
4
grupos

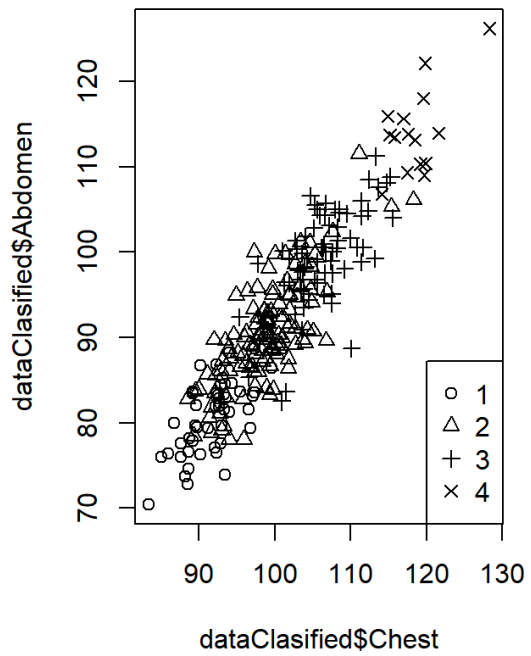

Knee
34
38
42
46
1
2
3
4
grupos

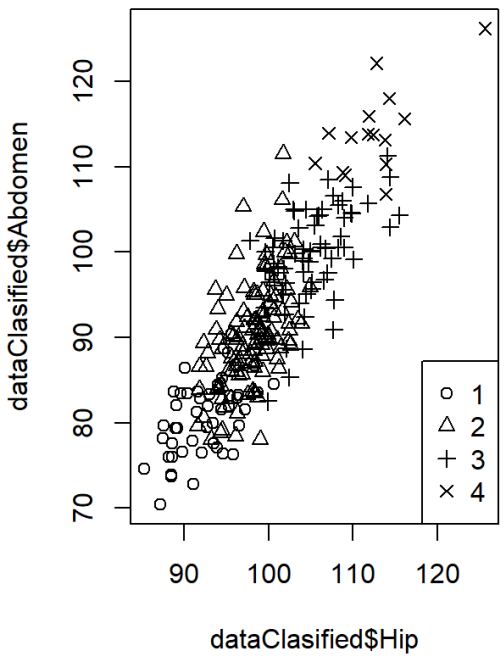

Ankle
20
25
30
1
2
3
4
grupos

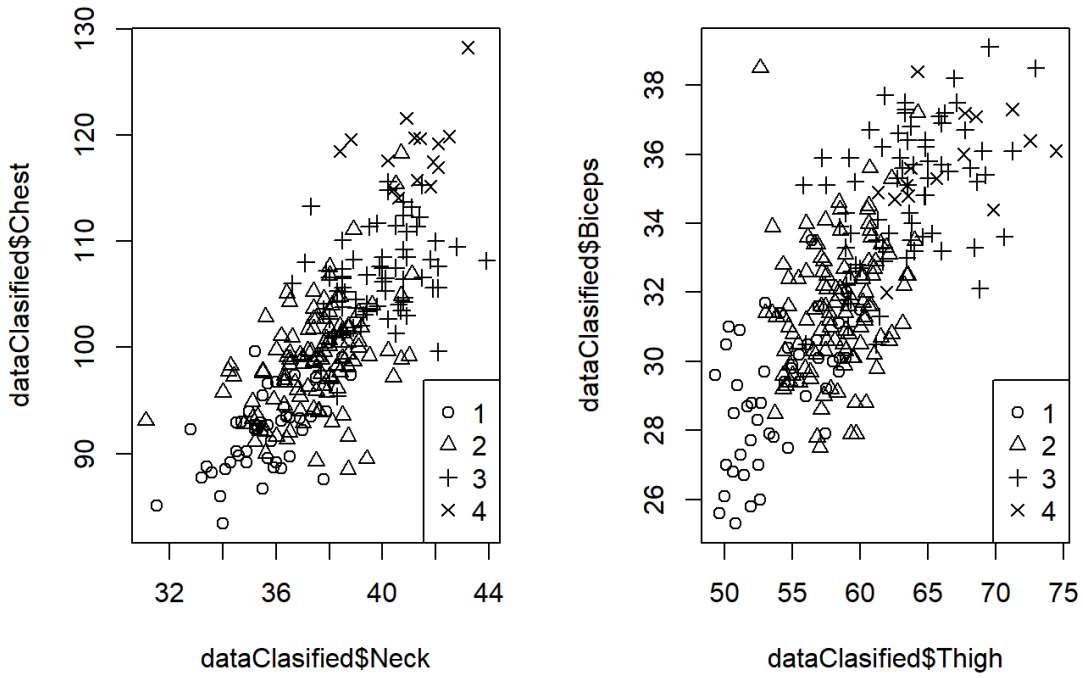

Biceps
26
30
34
38
1
2
3
4
grupos


Forearm
22
26
30
34
1
2
3
4
grupos

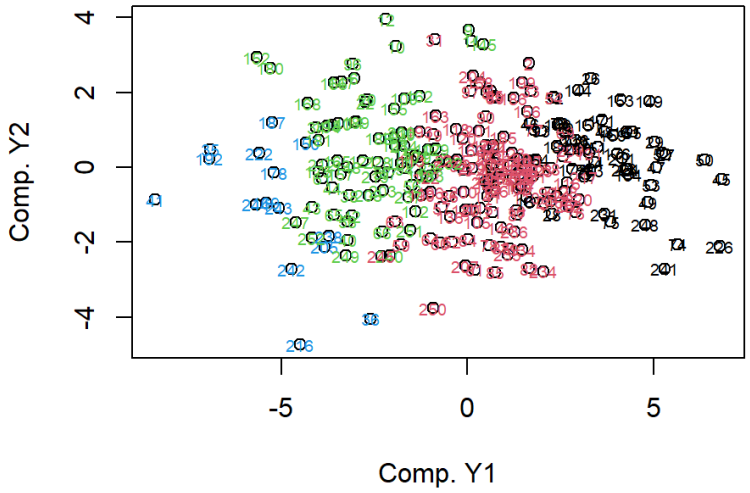

Wrist
16
17
18
19
20
21
1
2
3
4
grupos


dataClasified$Abdomen
70
80
90
100
110
120
90
100
110
120
130
dataClasified$Chest
1
2
3
4


dataClasified$Abdomen
70
80
90
100
110
120
90
100
110
120
dataClasified$Hip
1
2
3
4



Atendiendo al resumen de las variables respecto a cada grupo y a las anteriores gráficas podemos describir los grupos en los que se clasifican a los individuos de la siguiente forma:



Grupo1. (Negro): De media es el grupo con más densidad, y que menor porcentaje de grasa corporal presenta. Además sus individuos tienen el menor peso corporal. Y en cuanto al tamaño, son los individuos con menor circunferencia de las partes del cuerpo. En conclusión, son los **hombres menos corpulentos y con el menor porcentaje de grasa corporal.**

Grupo2. (Rojo):Los individuos de este grupo presentan un porcentaje de grasa corporal repartido, con valores tanto altos como bajos, pero con una mayoría de individuos con un porcentaje moderado o no muy alto , entre el 17 y el 20 %, y un peso corporal medio/bajo. El tamaño de las distintas partes del cuerpo es en general también intermedio o medio/bajo, pero por encima de los individuos del primer grupo. Se puede resumir en **individuos con un tamaño corporal medianamente pequeño (medio/bajo), y con un porcentaje de grasa mayoritariamente moderado.**

Grupo3. (Verde) En este grupo los individuos presentan un porcentaje de grasa corporal bastante repartido al igual que en el grupo 2, pero mayoritariamente algo mayor, en un rango medio/alto , en torno al 22 - 24%. Presentan un peso corporal más bien alto, junto con un tamaño corporal medio/alto. En conclusión, son **hombres con un tamaño corporal**

medianamente grande, y porcentaje de grasa corporal mediano. Destacan también por ser mayoritariamente el grupo de hombres más altos.

Grupo4. (Azul oscuro): Grupo formado por un conjunto reducido de individuos que presentan porcentajes altos de grasa, junto con un alto peso corporal. Siendo a su vez el grupo con la densidad corporal más baja en general. Destacan por tener un tamaño corporal bastante mayor que el resto de individuos. Podría resumirse como el grupo de **los hombres más corpulentos, y de mayor porcentaje de grasa corporal, además de un peso corporal pesado.**

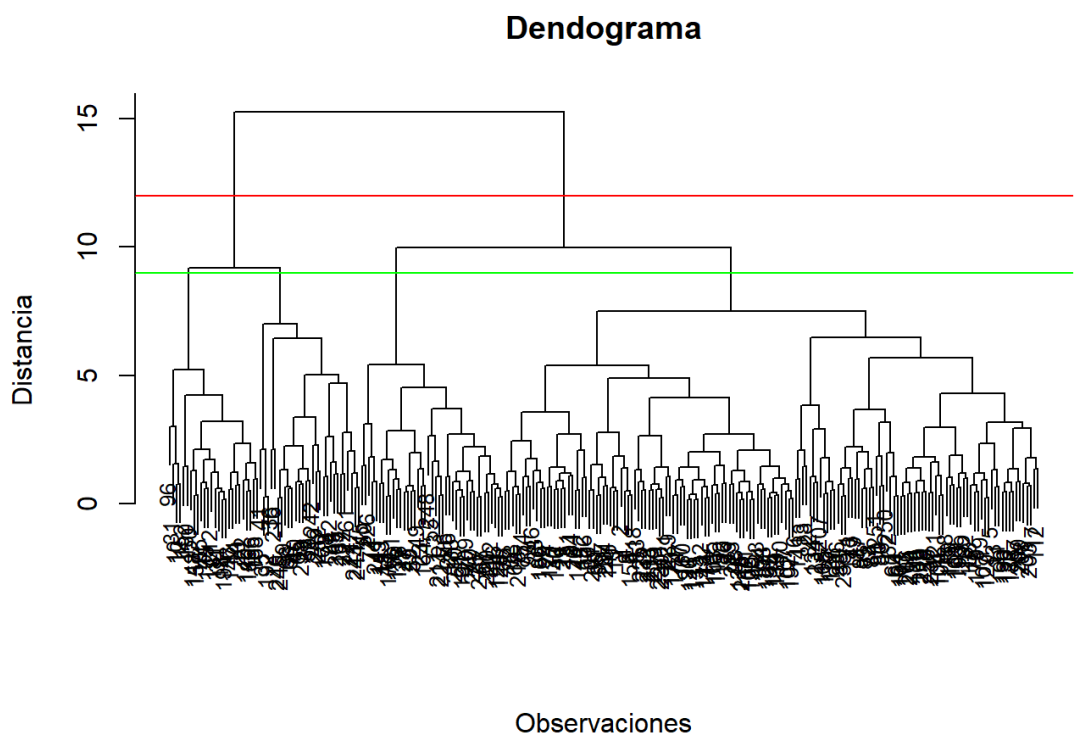
```
[1] "Individuos del grupo 1:"

[1] 48

[1] "Individuos del grupo 2"

[1] 116

[1] "Individuos del grupo 3"

[1] 70

[1] "Individuos del grupo 4"

[1] 15
```

Se da claramente un desbalanceo al haber un número total de individuos tan dispar en cada grupo. Es importante tener esto en cuenta ya que al realizar algunos análisis estadísticos puede conducir a resultados sesgados o poco confiables.

2.Aplicación del análisis jerarquizado (hclust) a partir de las puntuaciones de los individuos en las 3 componentes principales

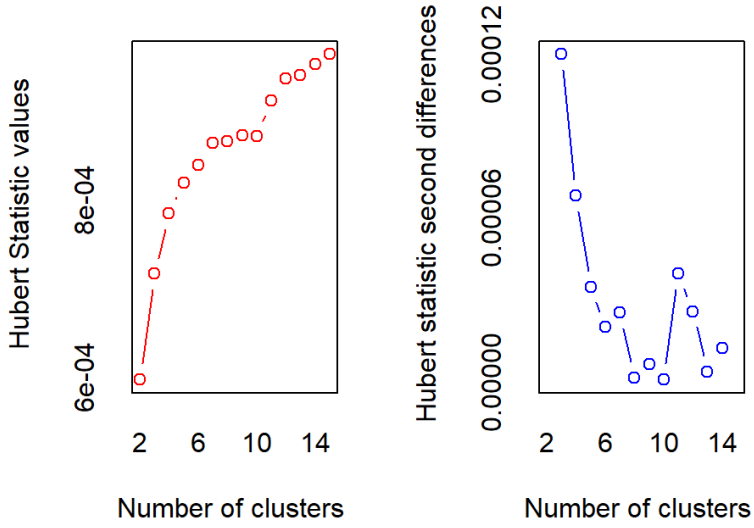
Durante el Análisis de las componentes principales, seleccionamos las 3 primeras que abarcaban un porcentaje total de la información del 80%, esto quiere decir que muy probablemente la clasificación que obtengamos en este segundo análsis cluster varíe, siendo menos precisa.

Obtenemos el siguiente dendrograma:

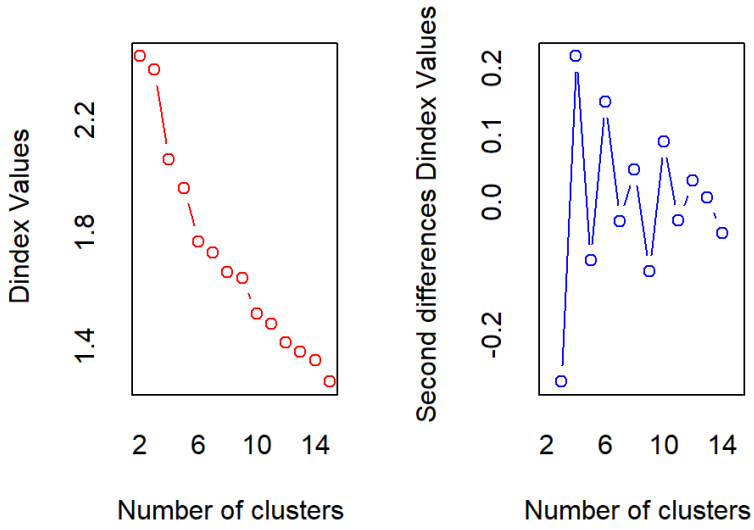


Parece que lo mejor sería tomar entre 2 y 4 grupos.

2.1Número de grupos a escoger



*** : The Hubert index is a graphical method of determining the number of clusters.
In the plot of Hubert index, we seek a significant knee that corresponds to a significant increase of the value of the measure i.e the significant peak in Hubert index second differences plot.



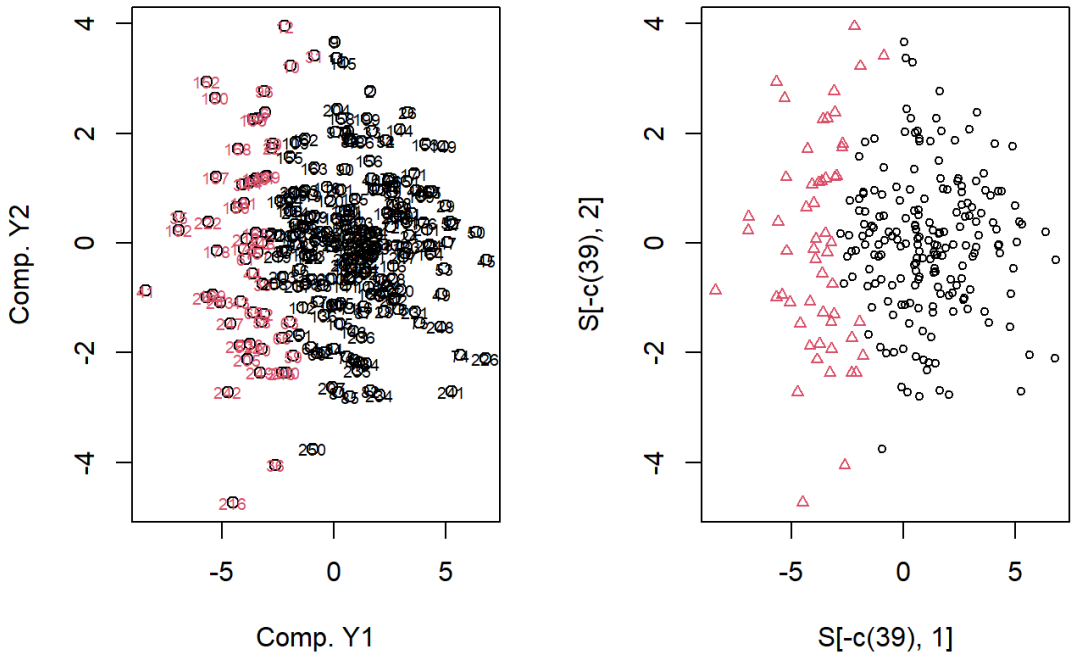
*** : The D index is a graphical method of determining the number of clusters.
In the plot of D index, we seek a significant knee (the significant peak in Dindex second differences plot) that corresponds to a significant increase of the value of the measure.

```
*******************************************************************
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 6 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 2 proposed 6 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 2 proposed 12 as the best number of clusters
* 2 proposed 15 as the best number of clusters
```

```
                    ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
                     KL       CH Hartigan     CCC    Scott   Marriot   TrCovW
Number_clusters 12.000   2.0000   3.0000  2.0000   4.0000         6      4.0
Value_Index     10.322 159.6128  67.7747 -7.2497 163.9422 457674717 219365.9
                  TraceW Friedman   Rubin Cindex     DB Silhouette   Duda
Number_clusters   4.0000  12.0000  6.0000 2.0000 3.0000     2.0000 2.0000
Value_Index     315.1699   3.1969 -0.4396 0.2223 0.7706     0.3882 1.0608
                PseudoT2   Beale Ratkowsky     Ball PtBiserial Frey McClain
Number_clusters   2.0000  2.0000   10.0000   3.0000      3.000    1  3.0000
Value_Index      -3.0962 -0.0958    0.2241 346.8593      0.538   NA  0.3022
                   Dunn Hubert SDindex Dindex    SDbw
Number_clusters 15.0000      0  3.0000      0 15.0000
Value_Index      0.0978      0  1.1507      0  0.1187
```

En este caso la mayoría de índices indican que el mejor número de clusters es 2.

```r
#Agrupando en 2 grupos
g2 <- cutree(CA2, k = 2)
```

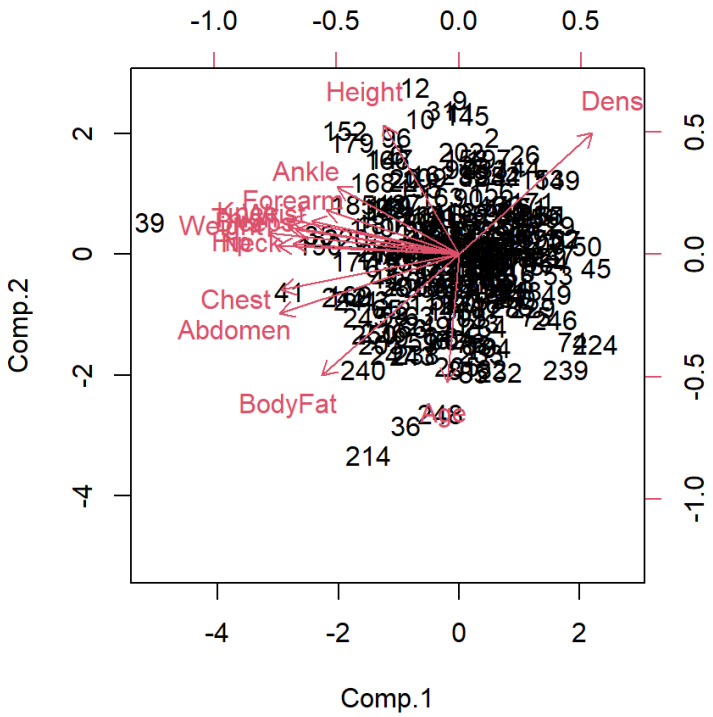


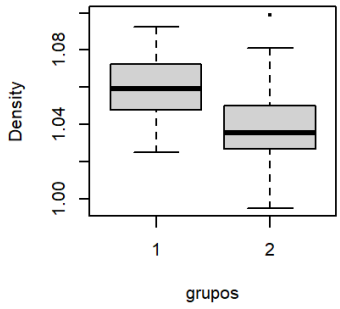
Finalmente añadimos la clasificación hecha a los dataset.

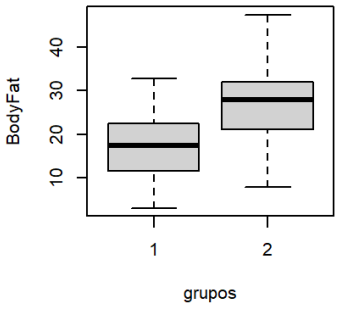
```r
#Le quito el 39 porque así lo había hecho en los dos dataset anteriores
Puntuaciones3Comp <- data.frame(Puntuaciones3Comp[-c(39),],g2)
dataClasified <- data.frame(dataClasified, g2)
dsClasified <-data.frame(dsClasified,g2)
```

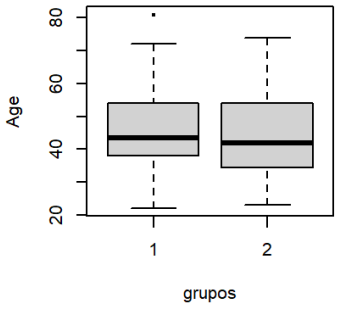
2.2 Descripción de los grupos.

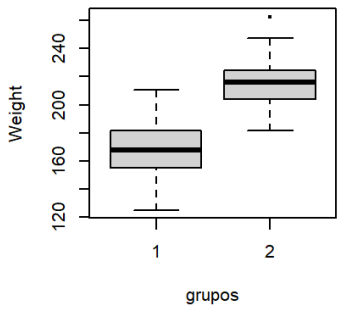
Al existir solamente dos grupos, el estudio de las características de estos se simplifica bastante.

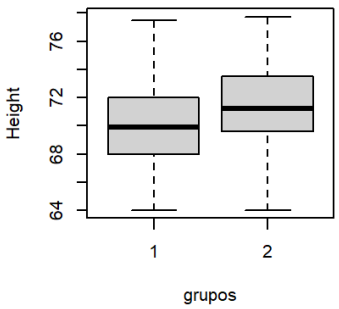

Height
Dens
Ankle
Forearm
Chest
Abdomen
BodyFat
Comp.1
Comp.2

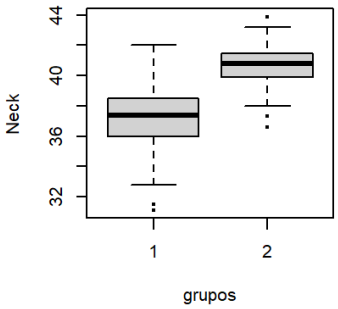

Density
grupos

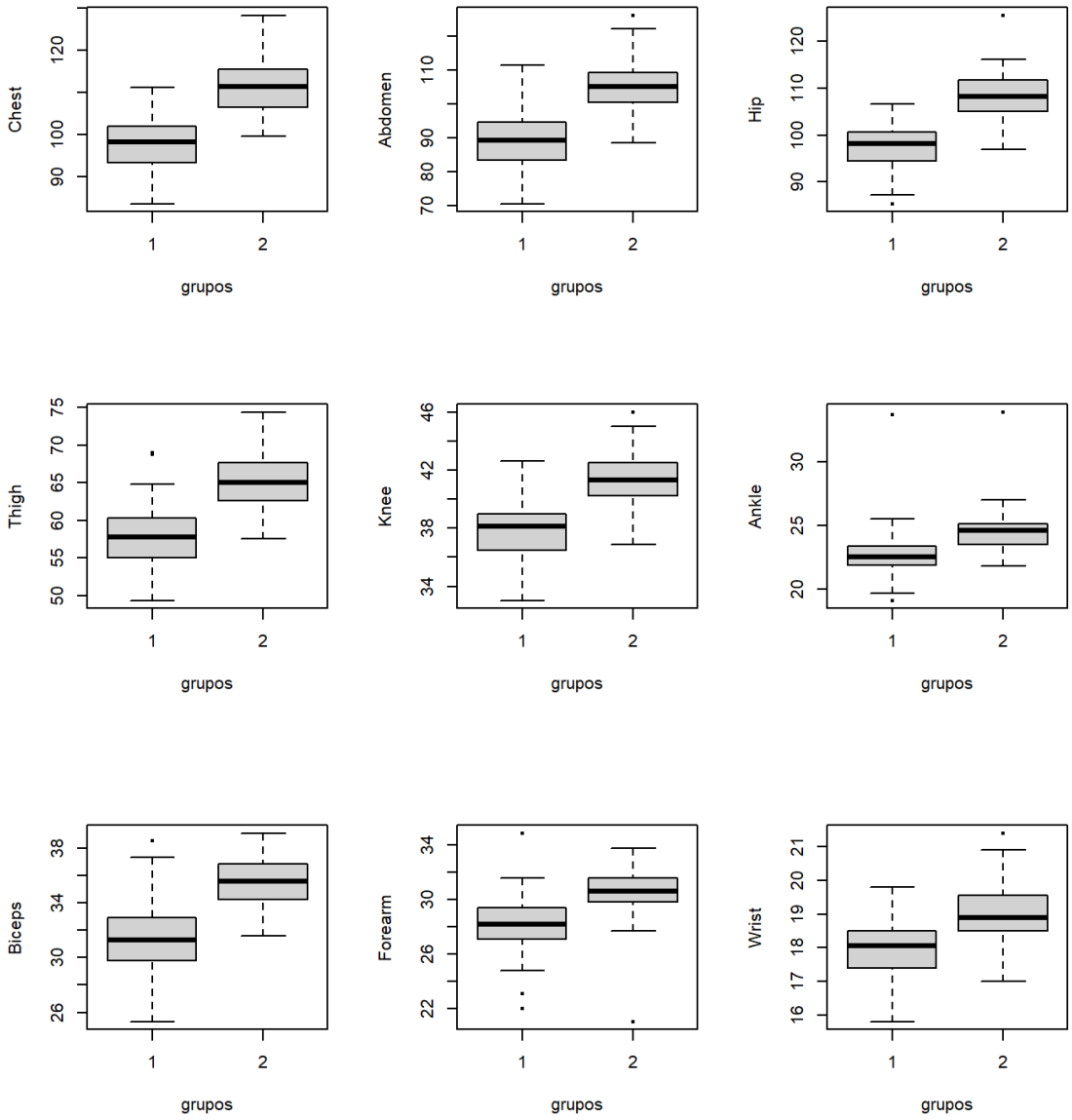

BodyFat
grupos


Age
grupos


Weight
grupos


Height
grupos


Neck
grupos



Con tan solo dos grupos, la clasificación queda algo pobre, al no ser tan precisa como la anterior hecha con 4 grupos.

En este caso tenemos un primer grupo (en color negro en la gráfica) en el que se clasifican los hombres con un porcentaje de de grasa tanto pequeño como intermedio, peso corporal entre pequeño y mediano, mayor densidad y con un tamaño corporal tanto pequeño como intermedio, es decir no excesivamente corpulentos. Y un segundo grupo que aglutina a aquellos con un mayor peso corporal, porcentaje de grasa alto, menor densidad y los más corpulentos.

GRUPO1. Los **hombres con un tamaño corporal tanto pequeño como mediano y un porcentaje graso entre bajo y medio.**

GRUPO2. Los **hombres que destacan por tener una mayor corpulencia y porcentaje de grasa que el resto.**

```
[1] "Individuos del grupo 1:"
```

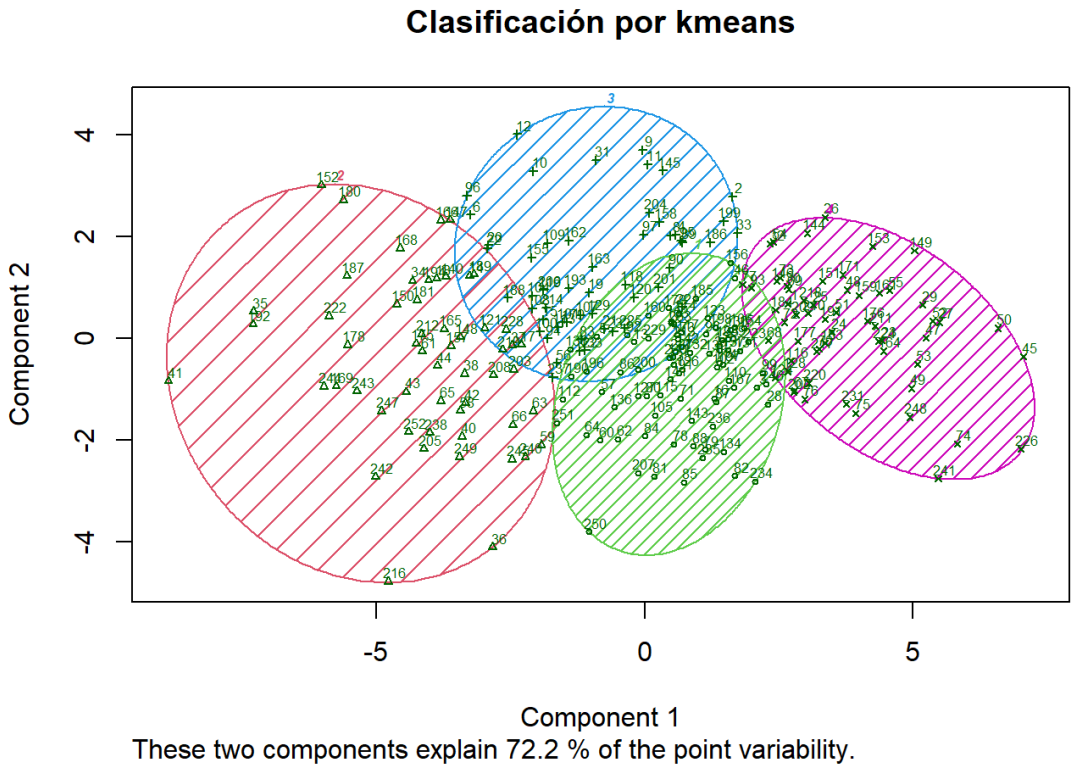
```
[1] 194
```

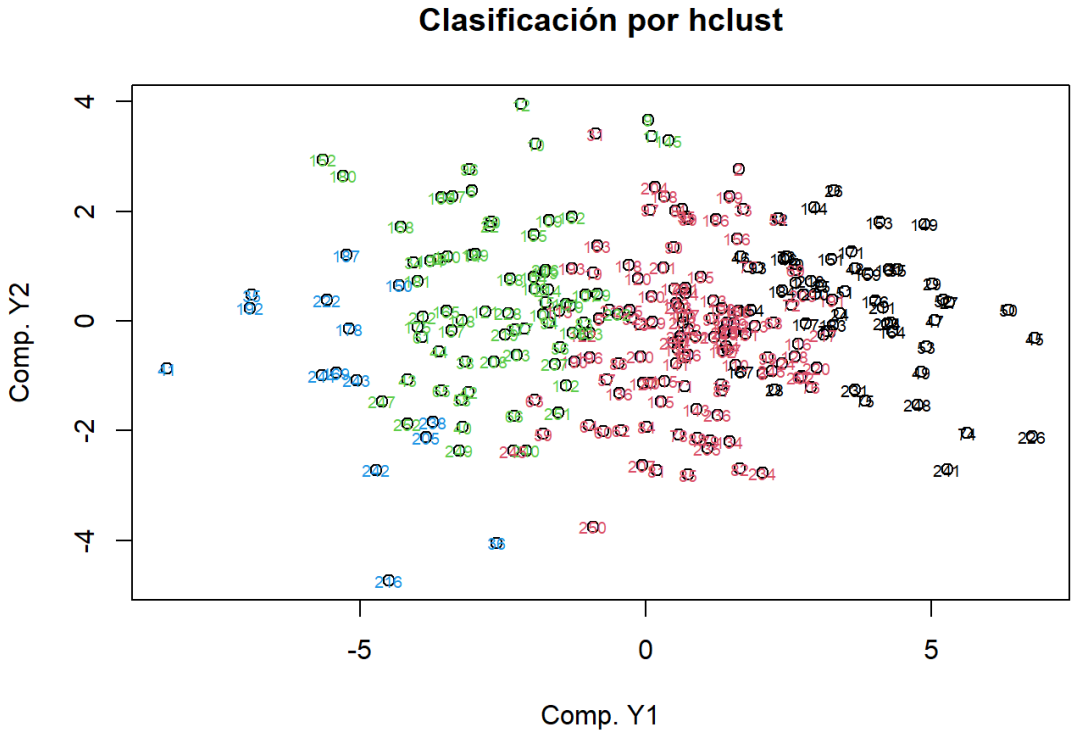
```
[1] "Individuos del grupo 2"
```

```
[1] 55
```

3.¿Cómo hubiera clasificado el algoritmo kmeans para el mismo número de grupos?

Finalemente podemos comparar la clasificación en 4 grupos hecha por hclust con la que haría Kmeans



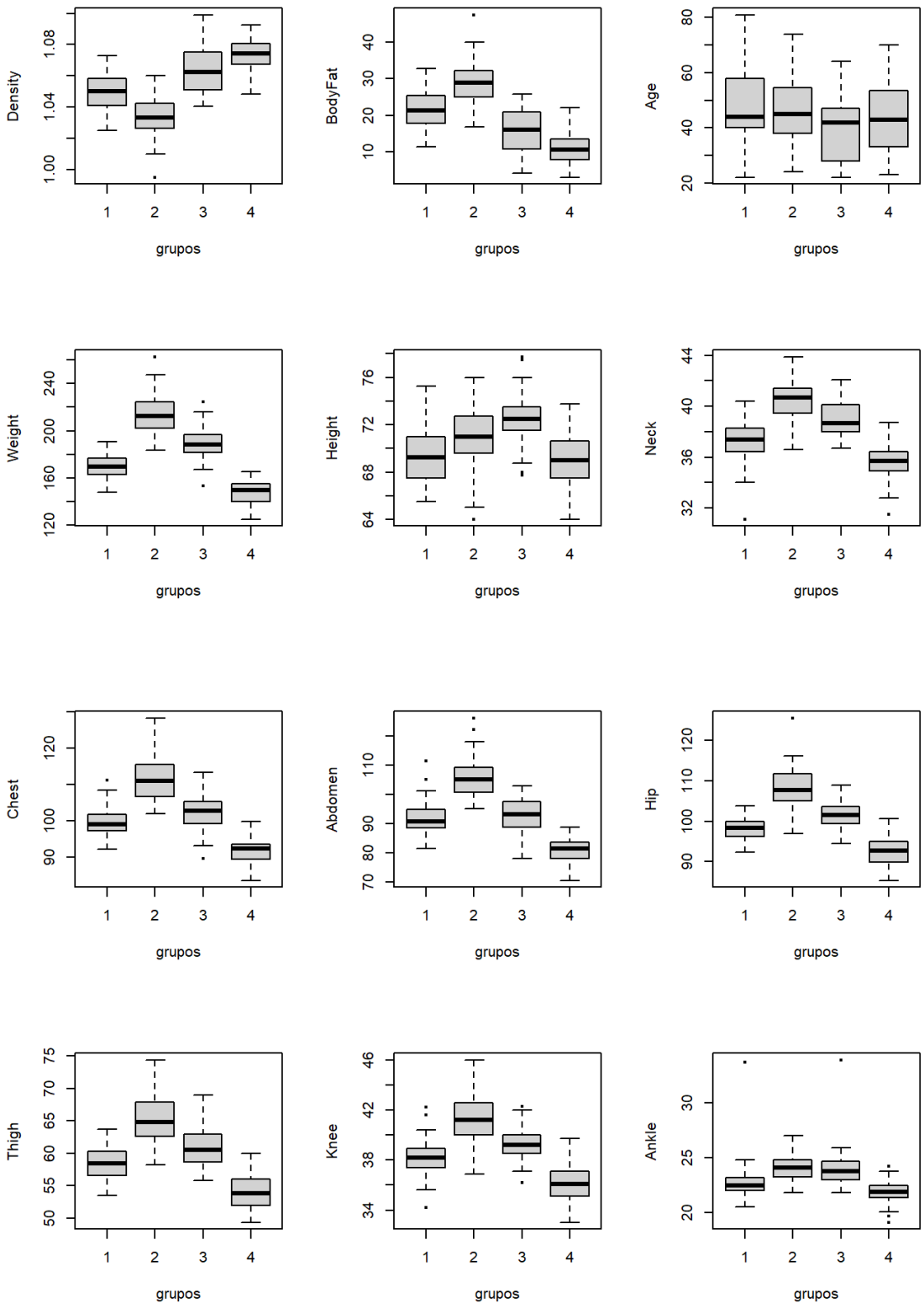


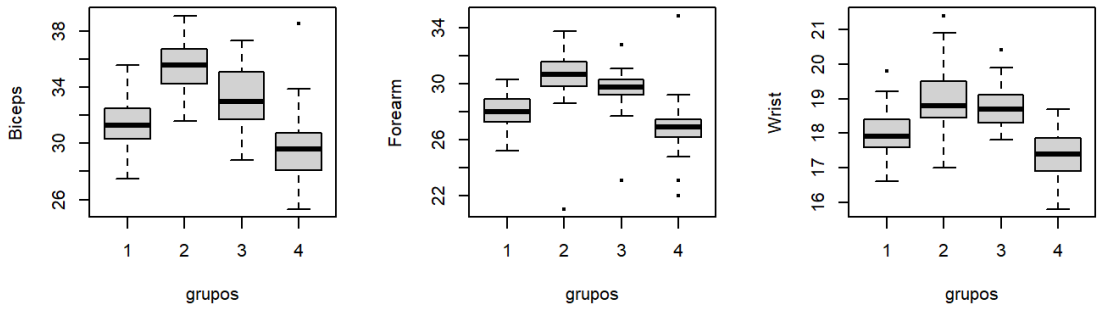
Añadimos la clasificación a los dataset

```
Y <- CA1$cluster
dataClasified <- data.frame(dataClasified, Y)
dsClasified <- data.frame(dsClasified,Y)
```

3.2 Descripción de los grupos.

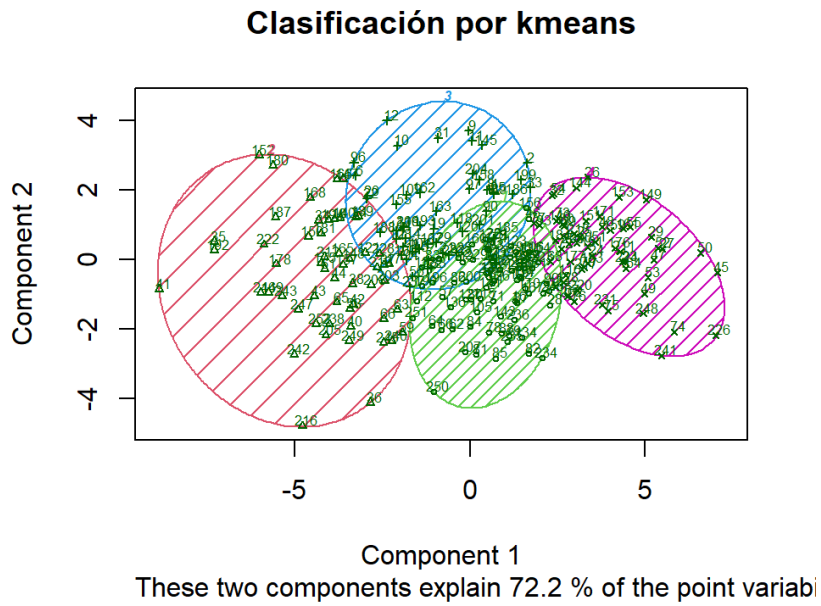
Visualizamos las variables respecto a los grupos formados por kmeans, lo que nos permite describir las caracaterísticas de estos.





La clasficación por el algortimo kmeans es distinta a la hecha anteriormente con hclust. Esto se debe a que kmeans, como primer paso determina los centroides iniciales de los k grupos al azar, agrupando a los individuos según su cercanía a estos centroides (de ahí la semilla). Lo que da lugar a que aunque los centroides se vuelvan a calcular tras cada iteración, haya una componente de aleatoriedad en este algoritmo y quizás para un conjunto de datos pequeño como este, en el que los grupos no están tan bien definidos, no sea tan preciso. Además de que puede dar lugar a una peor clasificación, dado que se vea forzado a incluir algún atípico en un grupo, teniendo que desplazar o modificar los 'límites' de otros grupos.

-Una descripción de los grupos hecha por kmeans sería:



GRUPO1. (Verde): Es un grupo en el que sus individuos presentan un **porcentaje de grasa corporal intermedio, junto con un tamaño corporal mediano y estatura entre media y baja.**

GRUPO2. (NARANJA): Se trata del grupo que engloba a los **hombres de mayor tamaño corporal, así como un porcentaje de grasa mayoritariamente más alto** que el resto de grupos, también presentan **gran estatura.**

GRUPO3. (Azul): En este grupo los individuos presentan un **porcentaje de grasa corporal medio/bajo**, con un **tamaño corporal mediano**, pese a ser los **individuos mayoritariamente más altos.** Además tienen un **peso corporal medio bajo.**

GRUPO4. (Rosa): Este grupo lo componen los **hombres de menor tamaño y peso corporal**, que además presentan el **menor porcentaje de grasa** corporal.

Para esta nueva clasificación, tenemos dos grupos que engloban a los individuos de mayor y menor tamaño corporal, estos son el 2 y el 4, cuyas características se asemejan a las de los grupos 4 y 1 del hclust respectivamente. Y además, se tiene dos cluster para aquel subconjunto de hombres con una corpulencia intermedia, pero que se diferencian mayoritariamente por su edad y altura.

Siendo estos últimos los cluster 1 y 3. Son bastante similares, presentando ambos un tamaño corporal mediano/pequeño, pero el primero de ellos destacando por componerse de indviduos más longevos y de menor estatura, mientras que el otro por tener individuos más jovenes, más altos y con un porcentaje de grasa corporal algo más bajo.

```
[1] "Individuos del grupo 1:"
```

```
[1] 81
```

```
[1] "Individuos del grupo 2"
```

```
[1] 55
```

```
[1] "Individuos del grupo 3"
```

```
[1] 54
```

```
[1] "Individuos del grupo 4"
```

```
[1] 59
```

Con kmeans sigue habiendo un problema de desbalanceo, pero este se reduce bastante.

En coclusión, con **hclust los grupos son más representativos, los individuos quedan mejor delimitados según sus atributos corporales que con kmeans** permitiéndonos definir/describir mejor las características de estos. **Pero con kmeans se da un menor balanceo lo que pueda repercutir en resultados más fiables a la hora de clasificar nuevos individuos.**

ANÁLISIS DISCRIMINANTE

Con tal de validar si los grupos obtenidos anteriormente por análisis cluster son realmente distinguibles, así como, cual de las dos algoritmos obtiene resultados más precisos (hclust o kmeans) y además descubrir qué variable es la más influyente a la hora de clasificarlos, vamos a llevar a cabo un análisis discriminante.

Aplicamos el análisis discriminante con las variables estandarizadas, ya que de esta forma tendrán valores similares, y así los coeficientes obtenidos con ellas en el análisis se podrán usar para estudiar la influencia de las variables en la clasificación.

- **¿LDA o QDA?**

Primeramente debemos intentar decidir qué método es el óptimo para nuestros datos. En función de las características de estos debemos determinar cual de los dos métodos, LDA o QDA, podría retornar mejores resultados.

Será buena opción aplicar LDA si las matrices de covarianzas de cada grupo son iguales o muy similares, mientras que la opción del QDA será más acertada si las variables usadas para clasificar son normales (multivariantes) en cada grupo.

```
library('MASS')
```

```
Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

```
library('mvnormtest')
#Comprobando para el LDA
grupo1 <- dsClasified[dsClasified$g4 == 1, 1:15]
s1 <- cov(grupo1)
grupo2 <- dsClasified[dsClasified$g4 == 2, 1:15]
s2 <- cov(grupo2)
grupo3 <- dsClasified[dsClasified$g4 == 3, 1:15]
s3 <- cov(grupo3)
grupo4 <- dsClasified[dsClasified$g4 == 4, 1:15]
s4 <- cov(grupo4)
```

**Véase, en el archivo de código aportado, como las matrices de covarianzas no coinciden, son bastante diferentes. No se adjunta la salida de las matrices al ocupar mucho espacio.

```
#Comprobación del QDA.
mshapiro.test(t(grupo1))
```

```
    Shapiro-Wilk normality test

data:  Z
W = 0.15653, p-value = 4.177e-15
```

```
mshapiro.test(t(grupo2))
```

```
    Shapiro-Wilk normality test

data:  Z
W = 0.22405, p-value < 2.2e-16
```

```
mshapiro.test(t(grupo3))
```

```
    Shapiro-Wilk normality test

data:  Z
W = 0.13549, p-value < 2.2e-16
```

```
mshapiro.test(t(grupo4))
```

```
    Shapiro-Wilk normality test

data:  Z
W = 0.41345, p-value = 7.607e-07
```

Tampoco pasan el test de normalidad para poder confirmar la aplicación del QDA.

Por ello, viendo que ninguna de las dos comprobaciones, ni la del LDA, ni la del QDA, se cumple, podemos optar por emplear ambas y quedarnos con la que mejor clasifique los datos.

`1.LDA`

La función para aplicar el LDA presenta un parámetro 'prior', que permite definir las probabilidades de perternecia a priori a cada clase. Se suele emplear cuando se conocen dichas probabilidades y son fáciles de expresar. En este caso, al estar tan

desbalanceados los grupos es preferible que no se fije ninguna probabilidad a priori, de forma que se estime automáticamente utilizando la proporción de observaciones en cada clase en los datos de entrenamiento.

Además, el análisis se lleva a cabo por validación cruzada, de forma que no tengamos que diferenciar entre dos conjuntos de test o entrenamiento, sino que es la propia función la que hace particiones de todo el conjunto de datos y va tomando por cada iteración todas las particiones menos una, para dedicarlas a entrenamiento y la que excluyó la emplea para testear. Esta es una práctica común cuando la muestra no es de gran tamaño.

- **Análisis discriminante lineal para la clasificación hecha por hclust**

```
LDACV<-lda(dsClasified[,1:15],dsClasified$g4,CV=TRUE)
matrizConfusion1<-table(dsClasified$g4,LDACV$class,dnn=c("real","predicho"))
matrizConfusion1
```

```
    predicho
real   1   2   3   4
   1  40   8   0   0
   2   7 104   4   1
   3   0   9  57   4
   4   0   0   5  10
```

```
accuracy1 <- sum(diag(matrizConfusion1)) / sum(matrizConfusion1)
accuracy1
```

```
[1] 0.8473896
```

- **Análisis discriminante lineal para la clasificación hecha por Kmeans**

```
LDACV2<-lda(dsClasified[,1:15],dsClasified$Y,CV=TRUE)
matrizConfusion2<-table(dsClasified$Y,LDACV2$class, dnn=c("real","predicho"))
matrizConfusion2
```

```
    predicho
real  1  2  3  4
   1 77  0  1  3
   2  1 53  1  0
   3  8  4 41  1
   4  4  0  0 55
```

```
accuracy2 <- sum(diag(matrizConfusion2)) / sum(matrizConfusion2)
accuracy2
```

```
[1] 0.9076305
```

`2.QDA.`

Al igual que con LDA no vamos a definir las probabilidades de perternecia a priori para cada clase. También se lleva acabo por validación cruzada.

- **Análisis discriminante cuadrático para la clasificación hecha por hclust.**

```
#QDACV<-qda(dsClasified[,1:15],dsClasified$g4,CV=TRUE)
#table(dsClasified$g4,QDACV$class)
```

No deja aplicar el QDA para la clasificación hecha por hclust, dado que hay un grupo demasiado pequeño, como es el grupo 4 de hclust, compuesto por 15 individuos.

- **Análisis discriminante cuadrático para la clasificación hecha por Kmeans.**

```
QDACV2<-qda(dsClasified[,1:15],dsClasified$Y,CV=TRUE)
matrizConfusion3 <- table(dsClasified$Y,QDACV2$class, dnn=c("real","predicho"))
```

```
matrizConfusion3
```

```
    predicho
real  1  2  3  4
   1 73  4  2  2
   2  2 48  5  0
   3 17 10 22  5
   4  9  0  4 46
```

```
accuracy3 <- sum(diag(matrizConfusion3)) / sum(matrizConfusion3)
accuracy3
```

```
[1] 0.7590361
```

3.Recopilación Resultados

Finalmente, comparamos los resultados obtenidos:

Para el LDA, la eficiencia es:

- $211/249 = 0.85$ si cojo la clasificación de **hclust**
- $226/249 = 0.9$ si cojo la clasificación de **kmeans**

Para el QDA, la eficiencia es:

- No me permite coger los datos del hclust por haber un grupo con pocos individuos, lo cual es un problema que deriva directamente del desbalanceo que sufren los datos.
- $189/249 = 0.75$ si cojo la clasificación de **kmeans.**

Para los datos con los que trabajamos, resulta mejor la aplicación de un análisis discriminante lineal, ya que se obtienen mejores resultados con este que con el cuadrático. Además de que el análisis cuadrático no permite trabajar con grupos pequeños, de ahí el error que se muestra al intentar aplicarlo para la clasificación hecha por el hclust, que tiene un grupo con solo 15 individuos.

Fijándonos solamente el el análisis discriminante lineal, aunque la diferencia no es muy significativa, la clasificación de los datos agrupados por el algortimo Kmeans resulta ser algo más eficiente que para la agrupación por hclust. Obteniédose una exactitud del 85% en la clasificación de los grupos de hclust, frente a un 90% en la clasificación de los grupos de kmeans. Esto puede que se deba a que la agrupación hecha por Kmeans presenta un menor debalanceo que a la de hclust.

- **Conocer las variables más influyentes a la hora de discriminar**

Para la agrupación que ha resultado ser la más exacta a la hora de discriminar, merece poder conocerse cuales son las variables que más influyen a la hora de distinguir si un individuo debe pertenecer a uno u otro grupo.

```
                LD1         LD2          LD3
Density -0.92871026  0.67704161 -0.121969468
BodyFat -0.26773037 -0.40215865 -0.869374127
Age      0.10913969 -0.21179829  0.004336141
Weight   0.28944144 -0.18135608  4.838927686
Height   0.14203049  0.50532471 -1.390141002
Neck     0.38827075  0.49489423 -0.269586525
Chest    0.54465420  0.15336569 -1.185218797
Abdomen -0.20921349  0.11265065 -0.513766036
Hip      0.28622390  0.50402043  0.173175749
Thigh    0.47800106 -0.23068258 -1.505302721
Knee     0.26994204 -0.53277046  0.232811100
Ankle    0.14518666  0.19473755 -0.469521147
Biceps  -0.06054826  0.01886978  0.305727727
Forearm  0.34814466  0.16088753 -0.165216332
Wrist    0.02161552  0.18386047 -0.726118482
```

Al mostrar la matriz de 'scaling' tenemos 3 componentes discriminantes. Estos son combinaciones lineales de variables predictoras que utiliza el Análisis Discriminante para discriminar entre los diferentes grupos en un conjunto de datos.

Cada componente discriminante captura una cierta cantidad de información sobre la estructura de separación entre las clases en los datos. Y se ordenan según su capacidad para discriminar entre las clases. El primer componente discriminante captura la mayor parte de la variabilidad entre las clases 87.71%, el segundo un 10.41% y el tercero 1.88%.

Entonces, atendiendo a que el primer componente discriminante es el de mayor capacidad de discriminación, las variables más relevantes son la densidad corporal, y el tamaño del pecho y muslo.

REGRESIÓN LINEAL MÚLTIPLE:

Dado que la principal motivación con la que se recopilaron los datos de este estudio, fue para sacar conclusiones acerca de la estructura corporal de los hombres, vamos a llevar a cabo un análisis de regresión lineal múltiple. De esta forma podremos obtener un modelo que nos permita encontrar una forma de predecir el tamaño de alguna parte del cuerpo atendiendo al resto de medidas.

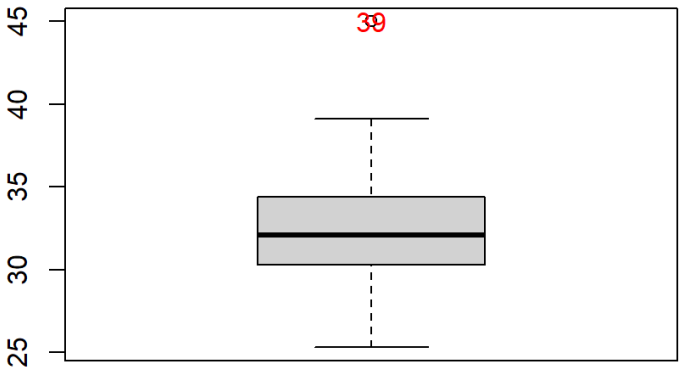
La regresión lineal la llevaremos a cabo con el dataset data, que no contempla las columnas de calsificación añadidas en apartados anteriores. Tampoco podemos aplicar los datos estandarizados ya que algunas funciones que usaremos (como boxcox()) no aceptan trabajar con una variable respuesta que presente valores negativos, como ocurre al estandarizar los datos.

1.Predicción de la variable Biceps

En el estudio incial, con la gráfica qqnorm() vimos como la variable Biceps era una de las que mejor parecía seguir una variable normal, por lo que intentaremos encontrar un modelo que prediga el tamaño del biceps del hombre en función del resto de variables.

1.1 Análisis previo de la variable respuesta

Analizaremos brevemente la normalidad de la variable Biceps, y veremos si requiere de alguna transformación en los datos.



Vemos de nuevo al individuo 39 como un atípico para el tamaño del biceps, lo cual, nos pueda suponer un problema a la hora de que los datos pasen el test de normalidad

$$H_0 : BodyFat \text{ sigue una distribución normal}$$

$$H_1 : BodyFat \text{ NO sigue una distribución normal}$$

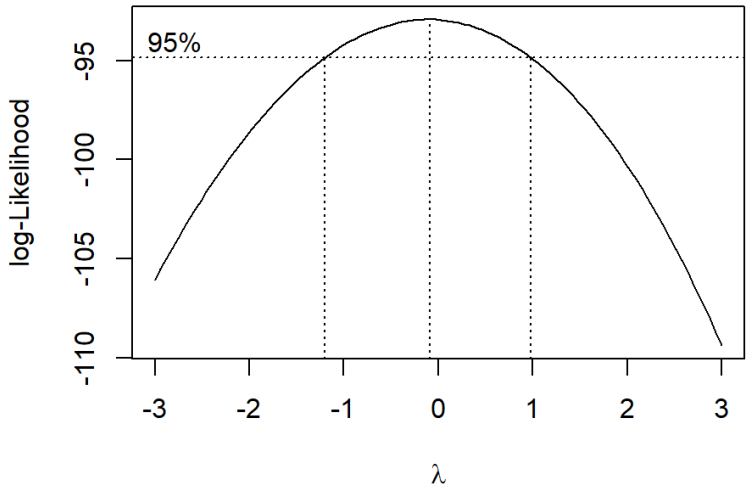
```
Shapiro-Wilk normality test
```

```
data:  Z
W = 0.9621, p-value = 3.642e-06
```

Probamos sin el atípico 39, a ver si así pasa en test de normalidad

```
    Shapiro-Wilk normality test

data:  Z
W = 0.98509, p-value = 0.01055
```

Debemos rechazar que Biceps sigua una distribución normal. Ni excluyendo al atípico parece seguir una normal. Muy probablemente esto se deba a que la variable requiera de una transformación en sus datos (por la familia Box-Cox). Para identificar la transformación más adecuada sobre la variable respuesta, usaremos la función boxcox() del paquete MASS. Indicar que la familia de transformaciones de Box-Cox consiste en elevar la variable respuesta a un exponente $\lambda > 0$, o bien tomar logaritmos neperianos si resulta $\lambda = 0$.



Se observa que la transformación más adecuada es la logarítmica. Por lo que creamos otra variable nueva, que sea la transformación de Biceps, a la que le volveremos a comprobar la normalidad. Y si pasa el test, será nuestra nueva variable respuesta.

```
data$BicepsNew <- log(data$Biceps) #Transfomación
mshapiro.test(t(data$BicepsNew))   #Comprobación de la normalidad
```

```
    Shapiro-Wilk normality test

data:  Z
W = 0.9933, p-value = 0.3249
```

Aceptamos la normalidad de la variable resultante de la transformación, ' BodyFat' , por lo que el modelo se contruye para predecir esa variable.

Antes de finalizar este primer apartado de análisis de la variable respuesta, cabe recordar que ya vimos en el estudio inicial de los datos, como algunas variables presentaban una fuerte relación lineal, lo cual nos podría resultar en problemas de multicolinealidad.

1.2 Construcción de un modelo inicial.

Primeramente, vamos a definir 2 subconjuntos, uno de entrenamiento o train, con el que construiremos el modelo y otro de test, con el que evaluaremos que tal predice nuestro modelo. De esta forma podremos evaluar qué tal se comporta con datos nuevos, teniendo así una métrica más fiable de qué tan bueno es el modelo.

```
set.seed(1234)
indices_entrenamiento <- sample(1:nrow(data), 0.8*nrow(data))
entrenamiento <- data[indices_entrenamiento,]
test <- data[-indices_entrenamiento,]
```

Construimos un modelo inicial con todas las variables, que nos pueda servir como base a partir de la cual aplicamos métodos de selección de regresores con tal de conseguir el mejor modelo posible.

```
library("rms")
```

```
Loading required package: Hmisc


Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

```
library("Hmisc")
modelo_inicial <- lm(BicepsNew ~ Density + BodyFat + Age + Weight + Height + Neck + Chest + Abdomen + Hip + Thigh
summary(modelo_inicial)
```

```
Call:
lm(formula = BicepsNew ~ Density + BodyFat + Age + Weight + Height +
    Neck + Chest + Abdomen + Hip + Thigh + Knee + Ankle + Forearm +
    Wrist, data = entrenamiento)

Residuals:
      Min        1Q    Median        3Q       Max
-0.139930 -0.028051  0.001252  0.032378  0.126131

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5460946  1.2022592   3.781 0.000210 ***
Density     -1.4687813  1.0868279  -1.351 0.178205
BodyFat     -0.0023692  0.0025229  -0.939 0.348905
Age          0.0003754  0.0004127   0.910 0.364262
Weight       0.0021606  0.0007952   2.717 0.007209 **
Height      -0.0036781  0.0023225  -1.584 0.114970
Neck         0.0042729  0.0030850   1.385 0.167711
Chest        0.0010019  0.0012941   0.774 0.439812
Abdomen     -0.0026774  0.0013805  -1.939 0.053978 .
Hip         -0.0031659  0.0017713  -1.787 0.075527 .
Thigh        0.0067236  0.0017083   3.936 0.000117 ***
Knee        -0.0033997  0.0030734  -1.106 0.270089
Ankle       -0.0001804  0.0025802  -0.070 0.944346
Forearm      0.0108996  0.0025801   4.225 3.75e-05 ***
Wrist        0.0053418  0.0068124   0.784 0.433965
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04649 on 185 degrees of freedom
Multiple R-squared:  0.7513,    Adjusted R-squared:  0.7325
F-statistic: 39.92 on 14 and 185 DF,  p-value: < 2.2e-16
```

```
vif(modelo_inicial)
```

```
  Density   BodyFat       Age    Weight    Height      Neck     Chest   Abdomen
36.655854 38.653268  2.393179 49.132964  3.314997  4.919010 10.459847 20.187212
      Hip     Thigh      Knee     Ankle   Forearm     Wrist
14.854460  7.276212  4.771260  1.903286  2.309145  3.690887
```

Vemos como el modelo presenta una bondad de ajuste relativamente buena, por encima del 0.73.

Podemos también comprobar la multicolinealidad al calcular el factor de varianza inflada, para los predictores. Este factor (VIF) mide la relación lineal de cada predictor con el resto de predictores. Valores de este factor que resulten por encima de 7 son indicativo de la existencia de multicolinealidad. Entonces, sabemos que hay variables presentando multicolinealidad.

Destacar también que hay bastantes regresores con los p-valores correspondientes a si los coeficientes son significativos, tomando un valor alto, superior a 0.10. Esto, junto con que haya variables que presentan multicolinealidad, es indicativo de que el modelo es claramente reducible.

1.3 Aplicación de los métodos de selección de regresores.

Aplicaremos los métodos de selección de regresores backward, fordward y stepwise, para intenatr optener un modelo más reducido.

a). Modelo Backward:

```
modelo_backward<- step(modelo_inicial, direction = "backward")
```

b).Modelo Forward

Para aplicar los modelos forward y stepwise, tendremos que partir de un modelo que contempla sólo la constante.

```
modelo_cte <- lm(BicepsNew ~ 1 , data = data)
```

```
modelo_forward <- step(modelo_cte, direction = "forward", scope = formula(modelo_inicial))
```

c). Modelo Stepwise

```
modelo_stepwise <- step(modelo_cte, direction = "both",scope = formula(modelo_inicial))
```

- **Comparación de los modelos**

  Se observa que el modelo obtenido por backward es distinto al obtenido por los otros dos métodos.

```
modelo_backward$coefficients
```

```
 (Intercept)       Weight       Height         Neck          Hip        Thigh
 2.953603275  0.001835021 -0.003794422  0.004687434 -0.004361102  0.006207710
     Forearm
 0.011934670
```

```
modelo_forward$coefficients
```

```
 (Intercept)       Weight      Forearm        Thigh         Neck        Chest
 3.248986683  0.001351152  0.011407284  0.004662362  0.005383144  0.002058783
     Abdomen      Density       Height
-0.003210058 -0.511864294 -0.002833185
```

```
modelo_stepwise$coefficients
```

```
 (Intercept)       Weight      Forearm        Thigh         Neck        Chest
 3.248986683  0.001351152  0.011407284  0.004662362  0.005383144  0.002058783
     Abdomen      Density       Height
-0.003210058 -0.511864294 -0.002833185
```

Debemos comparar la bondad de ajuste de los dos métodos obtenidos para así quedarnos con el mejor.

```
summary(modelo_backward)
```

```
Call:
lm(formula = BicepsNew ~ Weight + Height + Neck + Hip + Thigh +
    Forearm, data = entrenamiento)

Residuals:
      Min        1Q    Median        3Q       Max
-0.151558 -0.028739 -0.000715  0.031224  0.139112

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.9536033  0.2018426  14.633  < 2e-16 ***
Weight       0.0018350  0.0005136   3.573 0.000446 ***
Height      -0.0037944  0.0015689  -2.419 0.016510 *
Neck         0.0046874  0.0028482   1.646 0.101437
Hip         -0.0043611  0.0016674  -2.616 0.009613 **
Thigh        0.0062077  0.0014125   4.395 1.83e-05 ***
Forearm      0.0119347  0.0023167   5.152 6.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04636 on 193 degrees of freedom
Multiple R-squared:  0.742, Adjusted R-squared:  0.734
F-statistic: 92.51 on 6 and 193 DF,  p-value: < 2.2e-16
```

```
summary(modelo_forward)
```

```
Call:
lm(formula = BicepsNew ~ Weight + Forearm + Thigh + Neck + Chest +
    Abdomen + Density + Height, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.156578 -0.031328 -0.000224  0.035597  0.275481

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.2489867  0.4310035   7.538 9.63e-13 ***
Weight       0.0013512  0.0005722   2.361 0.018999 *
Forearm      0.0114073  0.0021696   5.258 3.21e-07 ***
Thigh        0.0046624  0.0013778   3.384 0.000834 ***
Neck         0.0053831  0.0025151   2.140 0.033329 *
Chest        0.0020588  0.0011832   1.740 0.083134 .
Abdomen     -0.0032101  0.0011577  -2.773 0.005993 **
Density     -0.5118643  0.3201473  -1.599 0.111166
Height      -0.0028332  0.0019804  -1.431 0.153835
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05019 on 241 degrees of freedom
Multiple R-squared:  0.7114,    Adjusted R-squared:  0.7018
F-statistic: 74.25 on 8 and 241 DF,  p-value: < 2.2e-16
```

Optaremos por quedarnos con el modelo backward ya que es el que mejor bondad de ajuste presenta.

En este modelo final, no se mejora la bondad de ajuste respecto al modelo inicial. Presenta un valor de R-cuadrado ajustado de 0.734, lo que indica que este modelo explica en un 73% la variabilidad del tamaño del biceps de un hombre. Además, no todos los regresores de este modelo final son significativos, al tener uno de ellos, Neck, un p-valor algo superior a 0.10. Esto nos indica que el modelo sigue siendo reducible.

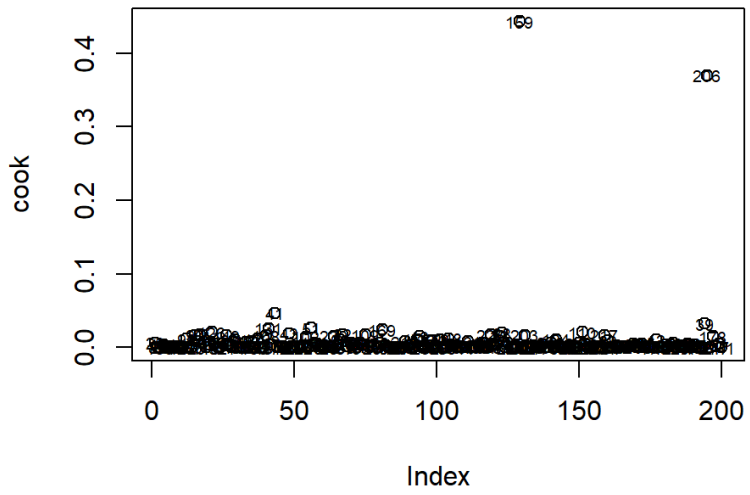
1.4 Validación del modelo

Antes de poder emplear el modelo libremente para predecir, debemos cerciorarnos de que es lo suficientemente bueno, como para considerar fiables y válidas las predicciones que se lleven a cabo con él. Se debe verificar que los residuos sigan una distribución normal y que sean independientes, que la varianza es constante en las perturbaciones aleatorias (Hipótesis de Homocedasticidad) y que no se de multicolinealidad entre variables predictoras ni existan observaciones que influyan mucho más que el resto.

a).Observaciones influyentes

Para comprobar que no hay observaciones influyentes, podemos calcular la distancia de Cook para cada observación. Como regla empírica, valores por encima de 1 indican que se trata de una observación influyente. En general, conviene representar los valores de la distancia de Cook con el fin de identificar si hay alguna observación con valores especialmente altos comparados con el resto.

```
cook <- cooks.distance(modelo_final)
plot(cook)
text(cook,cex=0.6, labels = row.names(entrenamiento))
```



b).Multicolinealidad

```
   Weight    Height      Neck       Hip     Thigh   Forearm 
20.610707  1.521175  4.215999 13.235564  5.001849  1.872244 
```

Vemos como Weight presenta gran multicolinealidad, debemos quitarla y ver que ocurre.

```
modelo_final <- lm(formula = BicepsNew ~ Height + Neck + Hip + Thigh + Forearm, data = entrenamiento)
summary(modelo_final)
```

```
Call:
lm(formula = BicepsNew ~ Height + Neck + Hip + Thigh + Forearm, 
    data = entrenamiento)
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.166857 -0.030469  0.001669  0.031106  0.156923

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3184660  0.0984736  23.544  < 2e-16 ***
Height      -0.0009548  0.0013931  -0.685    0.494
Neck         0.0108324  0.0023381   4.633 6.60e-06 ***
Hip          0.0002210  0.0010975   0.201    0.841
Thigh        0.0068554  0.0014426   4.752 3.91e-06 ***
Forearm      0.0131142  0.0023616   5.553 9.15e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04774 on 194 degrees of freedom
Multiple R-squared:  0.7249,    Adjusted R-squared:  0.7178
F-statistic: 102.3 on 5 and 194 DF,  p-value: < 2.2e-16
```

```r
vif(modelo_final)
```

```
  Height     Neck      Hip    Thigh  Forearm
1.130818 2.678763 5.406259 4.919472 1.834232
```

Al quitar Weight ya no hay multicolinealidad en el modelo. Además aprovechamos para reducir el número de variables predictoras al tener Height y Hip un p-valor bastante por encima de 0.1.

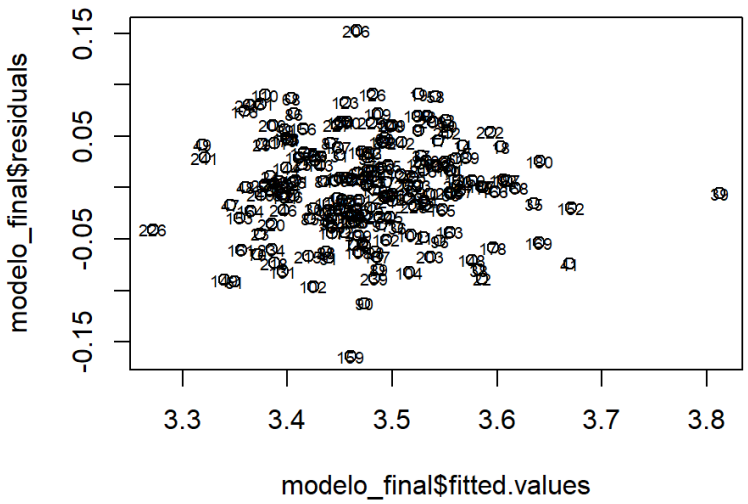
```r
modelo_final <- lm(formula = BicepsNew ~ Neck + Thigh + Forearm, data = entrenamiento)
summary(modelo_final)
```

```
Call:
lm(formula = BicepsNew ~ Neck + Thigh + Forearm, data = entrenamiento)

Residuals:
      Min        1Q    Median        3Q       Max
-0.164178 -0.029431  0.001707  0.031102  0.152730

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.267287   0.058207  38.952  < 2e-16 ***
Neck        0.010906   0.002171   5.025 1.13e-06 ***
Thigh       0.007024   0.000944   7.441 3.07e-12 ***
Forearm     0.012873   0.002326   5.534 9.94e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04756 on 196 degrees of freedom
Multiple R-squared:  0.7242,    Adjusted R-squared:   0.72
F-statistic: 171.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

El valor de R-cuadrado se acaba estableciendo superior al 72%.

c).Hipótesis de Homocedasticidad

Vemos como tiene un comportamiento aleatorio con dispersión aproximadamente constante

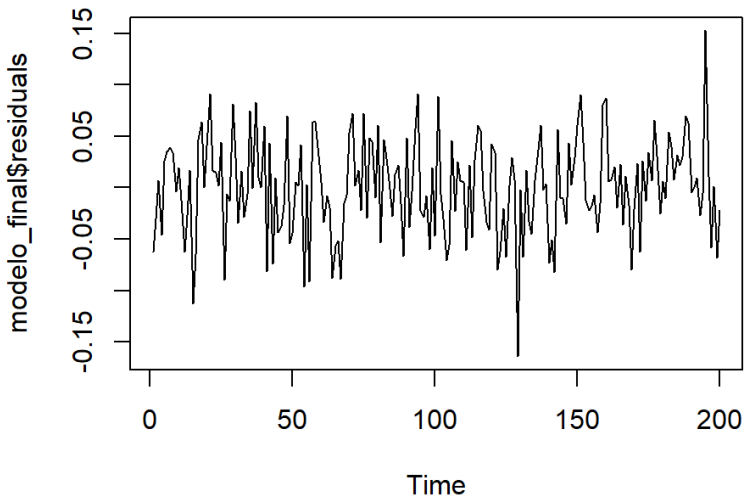


d)Hipótesis de Independencia

Analizamos la independencia con el test de Durbin-Watson y representando la serie temporal de residuos.

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


Attaching package: 'lmtest'

The following object is masked from 'package:rms':

    lrtest


    Durbin-Watson test

data:  modelo_final
DW = 1.8119, p-value = 0.174
alternative hypothesis: true autocorrelation is not 0
```

Al tener un p-valor superior a 0.10 en el test de Durbin-Watson, podemos suponer independencia de las observaciones.



e).Test de Normalidad de los residuos

```
    Shapiro-Wilk normality test

data:  modelo_final$residuals
W = 0.99379, p-value = 0.5698
```

El modelo pasa el test de normalidad de los residuos.

Se nos queda finalmente el siguiente modelo:

```
(Intercept)        Neck       Thigh     Forearm
 2.26728735  0.01090604  0.00702424  0.01287354


Call:
lm(formula = BicepsNew ~ Neck + Thigh + Forearm, data = entrenamiento)

Residuals:
      Min        1Q    Median        3Q       Max
-0.164178 -0.029431  0.001707  0.031102  0.152730

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.267287   0.058207  38.952  < 2e-16 ***
Neck        0.010906   0.002171   5.025 1.13e-06 ***
Thigh       0.007024   0.000944   7.441 3.07e-12 ***
Forearm     0.012873   0.002326   5.534 9.94e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04756 on 196 degrees of freedom
Multiple R-squared:  0.7242,    Adjusted R-squared:   0.72
F-statistic: 171.6 on 3 and 196 DF,  p-value: < 2.2e-16
```

1.5 Intento de mejora del modelo

Una vez que hemos comprobado que tenemos un modelo válido que pasa todas las comprobaciones necesarias podemos intentar probar a ajustarlo más, elevando a alguna potencia las variables predictoras.

No debemos pasarnos, por lo que entraríamos en problemas de sobreajuste.

```r
modelo_final_mejorado <- lm(formula = BicepsNew ~ I(Neck^4) + Thigh + pol(Forearm,3), data = entrenamiento)
summary(modelo_final_mejorado)
```

```
Call:
lm(formula = BicepsNew ~ I(Neck^4) + Thigh + pol(Forearm, 3),
    data = entrenamiento)

Residuals:
      Min        1Q    Median        3Q       Max
-0.125680 -0.030966  0.001398  0.035231  0.161395

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.331e+01  3.232e+00   4.120 5.62e-05 ***
I(Neck^4)                 3.845e-08  8.131e-09   4.729 4.33e-06 ***
Thigh                     6.080e-03  9.642e-04   6.306 1.89e-09 ***
pol(Forearm, 3)Forearm   -1.153e+00  3.405e-01  -3.385 0.000862 ***
pol(Forearm, 3)Forearm^2  4.206e-02  1.195e-02   3.520 0.000538 ***
pol(Forearm, 3)Forearm^3 -5.013e-04  1.392e-04  -3.602 0.000401 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0462 on 194 degrees of freedom
Multiple R-squared:  0.7424,    Adjusted R-squared:  0.7357
F-statistic: 111.8 on 5 and 194 DF,  p-value: < 2.2e-16
```

Al aplicar un polinomio de grado 3 a la variable Forearm, y elevar a potencia 4 a la variable Neck, conseguimos mejorar un poco el ajuste hasta superar el 74%.

Siendo finalmente los coeficientes del modelo mejorado:

```
            (Intercept)                I(Neck^4)                    Thigh
           1.331439e+01             3.845410e-08             6.079952e-03
  pol(Forearm, 3)Forearm pol(Forearm, 3)Forearm^2 pol(Forearm, 3)Forearm^3
           -1.152636e+00             4.206177e-02            -5.012717e-04
```

1.6 Predicciones

Ahora debemos probar a emplear el modelo para ver qué tal trabaja.

Podemos emplearlo para dar un intervalo de los valores posibles que pueden tomar los coeficientes. Cualquier modelo teórico con coeficientes incluidos en el intervalo siguiente serían también válidos teniendo en cuenta nuestros datos muestrales

```
                              2.5 %        97.5 %
(Intercept)               6.940164e+00  1.968861e+01
I(Neck^4)                 2.241766e-08  5.449054e-08
Thigh                     4.178329e-03  7.981575e-03
pol(Forearm, 3)Forearm   -1.824231e+00 -4.810406e-01
pol(Forearm, 3)Forearm^2  1.849184e-02  6.563169e-02
pol(Forearm, 3)Forearm^3 -7.757549e-04 -2.267884e-04
```

Para la predicción empleando nuevos valores, utilizaremos la función predict que permitirá obtener tanto valores de predicción como intervalos de confianza para el valor medio de la respuesta y para el valor de la respuesta utilizando la opción en el argumento interval = “confidence” e interval = “prediction”, respectivamente.

Una de las prácticas más comunes en el ámbito de la ciencia de datos es probar con los datos de test, los modelos construidos con los datos de entrnamiento. Y tasar la eficiencia de estos calculando algunas de sus métricas.

En nuestro caso calcularemos las métricas **MSE** (Error cuadrático medio) y **MAE** (Error absoluto medio)

$$MSE = E((h(X) - Y)^2).$$

$$MAE = E(|h(X) - Y|)$$

```
predicciones <- predict(modelo_final_mejorado, newdata = test, interval = "prediction", level = 0.95)
predicciones <-as.data.frame(predicciones)
```

Debemos de tener en cuenta que las predicciones son sobre la variable BicepsNew, y no sobre la variable original Biceps. Es decir las predicciones obtenidas están bajo la transformación de la familia Box-Cox, por lo que antes de calcular el MSE o MAE, se debe de deshacer dicha transformación.

```
#Deshacer la transformación
p <- exp(predicciones$fit)
MSE <- mean((test$Biceps - p)^2)
RMSE <- sqrt(MSE)
MAE <- mean(abs(test$Biceps - p))
```

```
MSE
```

```
[1] 3.930071
```

```
RMSE
```

```
[1] 1.982441
```

```
MAE
```

```
[1] 1.371291
```

El MSE al estar elevado al cuadrado, es bastante sensible a errores de predicción, y más aún si trabaja con algún atípico. Un MSE de 3.93 unidades con respecto a las observaciones reales pueda parecer un error alto para el contexto en el que nos movemos, ya que la circunferncia del biceps está más o menos en torno a un rango de 25 a 35 cm.

Pero si le hacemos la raiz cuadrada, es decir el RMSE, vemos como esta variación entre las predicciones y observaciones reales se reduce a la mitad.

Por otro lado, el MAE, indica que, en promedio, las predicciones del modelo se desvían aproximadamente en 1.37 unidades de las observaciones reales en valor absoluto.

Teniendo esto en cuenta, podemos decir que los errores no son excesivamente grandes a pesar de que la bondad de ajuste del modelo no supere el 75%.

Podemos considerarlo como un modelo relativamente bueno, que tiene la capacidad adecuada para dar una predicción cercana a la realidad.

```
summary(data$Biceps)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  25.30   30.30   32.10   32.32   34.38   45.00 
```

`-Para el individuo 182`

También podemos testear el modelo más a fondo y ver que nos retorna con los individuos que descartamos al inicio de este estudio. Por ejemplo emplearemos al individuo 182 con tal de predecir su tamaño del biceps en función del tamaño de su antebrazos, muslo, y cuello .

```
d182 <- data.frame(Forearm = d$Forearm[182], Thigh = d$Thigh[182], Neck = d$Neck[182])
```

```
#Intervalo para el valor de la respuesta
predict(modelo_final_mejorado, newdata = d182, interval = "prediction", level = 0.99)
```

```
       fit      lwr      upr
1 3.288416 3.164182 3.412649
```

```
exp(3.288416)
```

```
[1] 26.80038
```

```
d$Biceps[182]
```

```
[1] 27.7
```

Vemos como la predicción es relativamente buena para no tener el modelo una bondad de ajuste especialmente alta, se equivoca en menos de 1 cm.

2.Aplicación del Algoritmo del gradiente descendente.

Resulta interesante poder comparar los coeficientes del modelo estimados por medio del algoritmo del gradiente, con los coeficientes obtenidos por medio de la función lm(), que emplea el método de mínimos cuadrados.

El código empleado es una modificación del ofrecido en la práctica 2 de la asignatura Análisis Estadístico Multivariante.

```
#Vector de unos
x0_RLM <- c(rep(1,length(entrenamiento$BicepsNew)))
#Debemos tener en cuenta que la variable respuesta del modelo es BicepsNew
```

```
#Regresores
forearm <- entrenamiento$Forearm
thigh <- entrenamiento$Thigh
neck <-entrenamiento$Neck
chest <- entrenamiento$Chest
#Variable Respuesta
biceps <- entrenamiento$BicepsNew
```

```
n_RLM <-length(x0_RLM)
variables<-data.frame(x0_RLM,forearm,thigh,neck,chest,biceps)
k <- 4 #Número de regresores
```

```
#Metemos los datos muestrales en una matriz M (matriz de diseño)
M<-matrix(1,n_RLM,k+1)
M[,2]<-forearm
M[,3]<-thigh
M[,4]<-neck
M[,5]<-chest
#Definimos función costo usando la matriz de diseño M
h_RLM<-function(theta,x) {sum(theta*x)}
J_RLM<-function(theta,M) {0.5*sum((M %*%theta- biceps)^2)/n_RLM}
#Fijamos iteraciones, learning rate y valores iniciales
m_RLM <-50 # Número de interaciones
alfa_RLM <-0.1 # learning rate
theta_RLM <-c(1,1,1,1,1) #valores iniciales de los theta
```

```
J2_RLM<- array() #Vector con actualizaciones de la función costo en cada iteración
hv_RLM<- array() #Vector con actualizaciones de los valores ajustados en cada iteración
Z2_RLM<-matrix(NA,m_RLM+1,k+1) #Matriz con actualizaciones de los theta en cada iteración
J2_RLM[1]<-J_RLM(theta_RLM,M)
Z2_RLM[1,]<-theta_RLM
```

```
for (i in 1:m_RLM) {
hv_RLM<- M%*%theta_RLM
theta_RLM <- theta_RLM - (alfa_RLM/n_RLM) * t(M)%*%(hv_RLM-biceps)
```

```
J2_RLM[i+1]<-J_RLM(theta_RLM,M)
Z2_RLM[i+1,] <- theta_RLM
}
```

```
#Guardamos los valores de los theta y del costo en un dataframe
resultados_RLM <- data.frame(Z2_RLM, J2_RLM)
colnames(resultados_RLM) <- c("theta0", "theta1", "theta2","theta3","theta4", "costo")
```

```
A<-t(M) %*% M
B<-solve(A)
# A %*%B
thetas_exactos <-B %*% t(M) %*% biceps
```

```
[1] "Coeficientes obtenidos con el gradiente descendente:"

            [,1]
[1,] 2.275520728
[2,] 0.012346893
[3,] 0.006303383
[4,] 0.007939189
[5,] 0.001608672

[1] "Coeficientes obtenidos con el método de mínimos cuadrados: "

(Intercept)        Neck       Thigh     Forearm
 2.26728735  0.01090604  0.00702424  0.01287354
```

Se alcanzan por ambos métodos los mismos coeficientes del modelo. Aunque resulta más cómodo hacerlo solo por medio de la función lm() que solo requiere una única instrucción.

Consideraciones Finales.

Este trabajo me ha supuesto un primer contacto real no supervisado con el mundo del análisis de datos. Para llevarlo a cabo me he guiado por los archivos de la asignatura Análisis Estadístico Multivariante proporcionados por los profesores Concepción Domínguez y Jorge Navarro.

La redacción de este informe me ha permitido adquirir experiencia en la interpretación de los resultados, y también en la toma de decisiones informadas conforme a esos resultados que se van obteniendo durante el propio análisis de los datos.