Proyecto Multivariante

Javier García Fernández

1. Introducción

Este proyecto consiste en un análisis estadístico aplicando las técnicas vistas en clase a un fichero de datos previamente desconocido. La dirección de donde se obtuvo ndefineddicho fichero es https://data.world/achou/nba-draft-combine-measu

En este caso, dicho fichero trata sobre la NBA, en concreto sobre las medidas físicas de los jugadores. Se intentará llegar a conclusiones gracias a estas medidas, como qué cualidades son las que hacen que un jugador tenga mayor probabilidad de ser escogido.

Aclarar que las alturas han sido medidas en pulgadas, para pasarlas a metros habría que dividir el valor en pulgadas entre 39.37. Los pesos se encuentran en libras, para pasarlo a kilogramos habría que dividir el valor en libras entre 2.2046.

También es interesante conocer qué significan las variables:

- **X:** es el índice por el que se ordenan a los jugadores, como el que proporciona R pero este empieza por 0 en vez de por 1.
- **Player:** esta variable guarda el nombre de los jugadores.
- **Draft.pick:** esto simboliza el proceso de selección de nuevos jugadores por parte de los equipos cada año. Cuanto menor sea el valor, antes habrá sido elegido, siendo el primero el que tenga valor 1 y así sucesivamente. Si algún individuo no tiene ningún valor, significa que no fue elegido.
- **Height..No.Shoes.:** representa la altura del jugador sin zapatos.
- **Height..With.Shoes.:** representa la altura del jugador con zapatos.
- **Wingspan:** envergadura (distancia desde la punta de los dedos de una mano hasta la punta de los dedos de la otra mano, estando ambos brazos extendidos horizontalmente).
- **Standing.reach:** esta es la altura máxima con los brazos extendidos hacia arriba sin saltar.
- **Vertical..Max.:** es la altura máxima del salto.

- **Vertical..Max.Reach.:** es la altura máxima de salto sumada a la altura del jugador.
- **Vertical..No.Step.:** es la altura máxima de salto sin salto previo.
- **Vertical..No.Step.Reach.:** es la altura máxima de salto sumada a la altura del jugador pero sin salto previo.
- **Weight:** variable que representa el peso de cada individuo.
- **Body.Fat:** cantidad de grasa corporal de cada jugador.
- **Hand..Length.:** mide el largo de la mano de cada jugador.
- **Hand..Width.:** mide la anchura de la mano de cada jugador.
- **Bench:** número de repeticiones que un jugador puede hacer en el press de banca, levantando 84 kilogramos.
- **Agility:** mide la agilidad de cada jugador al realizar cierta prueba. El valor proporcionado es el tiempo en dicha prueba.
- **Sprint:** mide el tiempo que tarda un jugador en correr durante poco tiempo una corta distancia.

a) Análisis exploratorio

Podemos ver que existen muchos datos faltantes en ciertas columnas, posiblemente porque dichos datos son difíciles de obtener, como por ejemplo la longitud y anchura de la mano. También ocurre que ciertos individuos tienen muchos datos faltantes, posiblemente porque no quieren hacerlos públicos. Teniendo en cuenta todo esto, se procederá a tratar los datos.

```
summary(datos_nba)
```

```
# con esta libreria se hace un resumen mas preciso
library(Hmisc)
describe(datos_nba)
```

Podemos destacar ciertos aspectos, como por ejemplo que contamos con datos del 2009 hasta el 2017. También que el número máximo que se permite escoger en un draft es de 60, por lo que no existe ningún año donde se hayan escogido 61 o más. También podemos ver que de media los jugadores miden 77.61 y pesan de media 214.8 y demás.

Debido a la gran variedad de variables y los datos faltantes, hay que considerar reducir el abanico de posibilidades y tratar los datos con cautela. Primero se va a proceder a contar los datos faltantes para ver si tienen alguna relación o es que no se pudo obtener información en el momento de su recogida.

b) Limpieza de los datos

```
DATOS FALTANTES DE CADA VARAIBLE

[1] "X =  0"

[1] "Player =  0"

[1] "Year =  0"

[1] "Draft.pick =  133"

[1] "Height..No.Shoes. =  0"

[1] "Height..With.Shoes. =  1"

[1] "Wingspan =  0"

[1] "Standing.reach =  0"

[1] "Vertical..Max. =  67"

[1] "Vertical..Max.Reach. =  67"

[1] "Vertical..No.Step. =  67"

[1] "Vertical..No.Step.Reach. =  67"

[1] "Weight =  1"

[1] "Body.Fat =  3"

[1] "Hand..Length. =  47"

[1] "Hand..Width. =  49"

[1] "Bench =  233"
```

```
[1] "Agility =  73"
```

```
[1] "Sprint =  71"
```

Podemos sacar varias conclusiones a partir de estos datos faltantes:

Que en Draft.pick falten tantos datos es normal, ya que en el contexto de la NBA, en cada draft se escogen a un número de jugadores, pero no a todos, por lo que los Na representan aquellos que estaban en el conjunto de jugadores posibles de elegir pero que finalmente no fueron escogidos.

Si continuamos, que de unos 500 individuos falte el dato de solo 1 en Height..With.Shoes. es algo poco relevante, podemos seguir usando esa columna rellenando al individuo que le falta ese dato con el valor medio de las suelas de zapato, lo cual se puede hacer restando a la altura de cada individuo con zapato su altura sin él, hallar la media resultante y sumar dicho valor al individuo.

En Weight solo hay un dato faltante, podemos proceder igual que antes y rellenarlo con la media. Para la grasa podemos afinar un poco más. Hay estudios que parecen apuntar a que para un jugador de baloncesto la cantidad de grasa corporal ronda el 10.5%. Este valor de la grasa corporal varía mucho, pero este método es mejor que usar la media como antes, ya que la grasa corporal depende del peso.

```
[1]  96 427 505
```

```
[1] 207
```

```
[1] 93.89458
```

```
[1] 9.858931
```

```
[1] 214.8333
```

```
[1] 97.44776
```

```
[1] 10.23201
```

```
[1] 238
```

```
[1] 107.9561
```

```
[1] 11.33539
```

Algo más interesante ocurre con las columnas "Vertical..." (4), en todas falta el mismo número de individuos con datos, podemos comprobar si esto es casualidad o son el mismo individuo en cada caso con una función.

```
quien_na <- function(datos, column) {
  indices_na <- which(is.na(datos[[column]]))
  return(indices_na)
}
```

Vemos que entre con y sin salto coinciden, es decir, entre las parejas que consideran la altura con y sin carrerilla respectivamente. Esto parece indicar que, en el caso de que a un individuo le falte algún dato en una de estas 4 variables, le faltará en todas ellas.

Pero entre ellos hay algunos que sí y otros que no, por lo que debe haber algún individuo distinto que sí tenga un par y le falte el otro, es decir, que rompa con lo mencionado anteriormente y tenga 2 variables sin datos en lugar de las 4.

```
[1] 358
```

```
[1] "Rakeem Christmas"
```

Sin embargo, si hay 1 individuo que tiene un par de variables, debe haber otro que tenga el otro par de variables para que el número de datos faltantes coincida en las 4 columnas.

```
[1] 87
```

```
[1] "Stanley Robinson"
```

Llegados a este punto, si ordenamos las columnas, por ejemplo, "Vertical..Max. de mayor a menor, podemos fijarnos que llega un punto en el que practicamente todos los individuos restantes no tienen datos. Uno podría pensar que esto se debe a que estos jugadores son malos o por lo menos los peores entre los que se encuentran, y saltan tan poco que no se les ha tomado ninguna medida de estas variables.

Posteriormente, y cuando se haga un análisis más serio, se podrá comprobar si efectivamente el motivo era este o no tiene nada que ver.

Procedemos igual que antes para ver si existen jugadores en común a quienes les falten ambos datos referidos a la mano.

```
[1] "Hand..Length. =  47"
```

```
[1] "Hand..Width. =  49"
```

Podemos ver que los jugadores que no tienen registrados los datos de la longitud de la mano tampoco lo tienen de la anchura de la mano, pero existen 2 individuos que sí tienen la longitud de la mano pero no el ancho. Para estos dos individuos, como solo son dos, se van a rellenar dichos datos faltantes con la media de la anchura de las manos.

```
[1] 350 358
```

Para la variable Bench, el número de datos faltantes es elevado, pues los Na en este caso representan a aquellos jugadores que no han tomado esta prueba. Siguiendo la línea de razonamiento anterior, es posible que se deba a que solo aquellos más prometedores son sometidos a todas las pruebas.

Con Agility y Sprint sucede como con las variables referidas a las manos, vamos a ver si coinciden los individuos.

Coinciden bastantes, pero no todos, vamos a ver quiénes son los que le faltan los datos de Agility pero tienen los de Sprint y viceversa, respectivamente. Para que queden con el par de columnas rellenas, podemos añadir otra vez la media en cada una.

```
[1] 102 103 137 194 374
```

```
[1] 328 387 408
```

Se puede observar que, por ejemplo, la altura sin zapatos y con zapatos tienen una correlación positiva prácticamente perfecta, lo cual no es de extrañar. Podemos eliminar la variable Height..With.Shoes. pues es prácticamente la misma que sin zapatos y encima el grosor de la suela del zapato puede variar constantemente.

También podemos eliminar la variable Vertical..Max.Reach. pues es la suma de Vertical..Max. y Standing.reach, y de igual modo ocurre con la variable Vertical..No.Step.Reach.

Llegados a este punto, tenemos dos opciones: quedarnos con aquellas variables que no tienen ningún dato faltante, lo que daría como resultado 5 variables para analizar, o quitar los individuos que tienen datos faltantes.

Esta segunda opción puede ser interesante, pues si nos fijamos en la tabla, y gracias a los pasos previos, normalmente cuando falta algún dato de una variable, suele faltar la de la otra variable relacionada. Por ejemplo, con Vertical..Max. y Vertical..No.Step., o con las medidas

de las manos, es decir que se estarían quitando individuos que tienen varios datos faltantes, no distintos, lo cual no es tan malo.

Finalmente, se ha decidido quitar la variable Bench debido a la cantidad de Na's que posee y quitar aquellos individuos con Na's pues suelen tener datos faltantes en más de una variable. Este es un término medio entre eliminar todos los individuos con Na's y todas las variables con Na's.

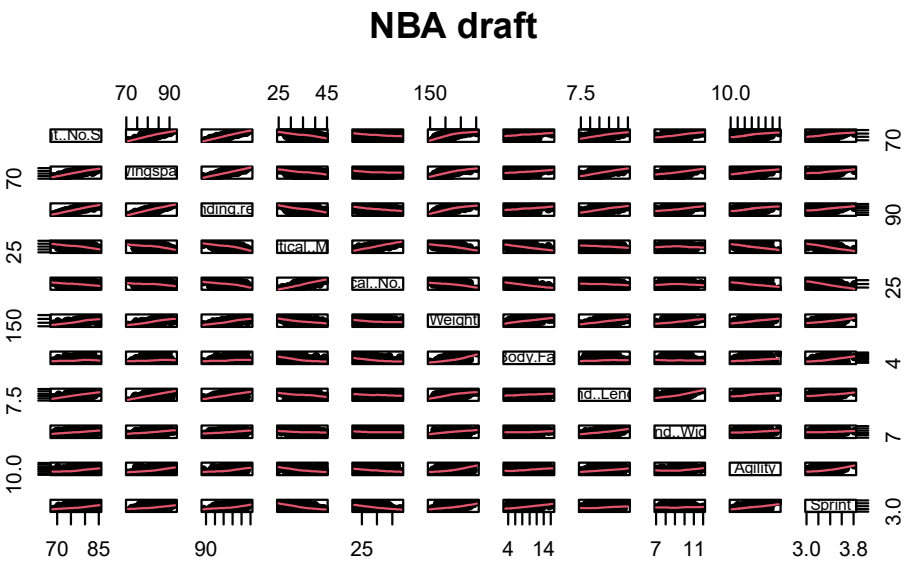
Obtenemos un total de 403 individuos con casi todas las variables (11), podemos proceder a hacer el PCA.

2. Análisis de componentes principales

Empezamos por analizar las componentes principales, ya que puede ayudarnos a seleccionar aquellas variables más representativas.

a) Estudio inicial de los datos

Puesto que ya hemos hecho un análisis previo, en este caso nos centraremos en ver otros aspectos que no se han tratado anteriormente, teniendo en cuenta además los datos de los que disponemos ahora.



Se puede observar que, por ejemplo, existe una correlación lineal positiva bastante fuerte entre las tres primeras variables. También ocurre algo parecido entre los dos tipos de saltos. Algo interesante también y que tiene sentido es que las variables de las manos parecen ser independientes con el resto de variables, menos entre ellas que están un poco correlacionadas. Por último, cabe destacar también una pequeña correlación negativa entre las variables de salto y las de agilidad y sprint.

Podemos hacer también histogramas para estudiar la normalidad de las variables.

```
for (i in 2:12) {
  hist(datos_all[[i]], probability = TRUE,
       xlab = names(datos_all)[i],
       main = paste(names(datos_all)[i]))

  curve(dnorm(x, mean(datos_all[[i]]),
              sd(datos_all[[i]])),
        add = TRUE, col = "red")
}
```

Parecen seguir todos una distribución normal, estamos en condiciones de comenzar el análisis PCA.

b) Análisis de componentes principales

Con 6 componentes se captura el 90% de la información, por lo que se va a proceder a ver qué significan las dos primeras componentes más representativas. Tener en cuenta que un valor alto positivo de sprint o agilidad no es bueno, quiere decir que el jugador es lento, algo no deseable por lo general.

La pregunta es, ¿Quién influye en Y1? ¿Qué es? ¿A qué jugadores representa?

Viendo los coeficientes, podemos decir que los que más influirán en orden descendente serán: Standing.reach, Height..No.Shoes., Wingspan y Weight. En cuanto al resto, podemos ver que todas son positivas, excepto las 2 variables referidas al salto y todas con coeficientes a tener en cuenta, parece que no se puede reducir más en cuanto a variables se refiere.

Según el orden de las variables, al ejecutar L[,1], tenemos los siguientes coeficientes:

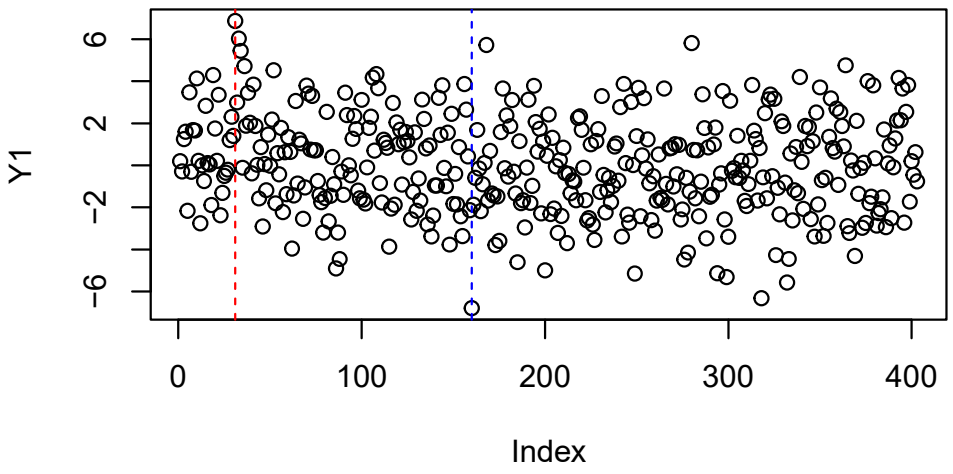
$$\begin{aligned} Y_1 = {} & 0.368 \cdot X_1^* + 0.367 \cdot X_2^* + 0.385 \cdot X_3^* - 0.290 \cdot X_4^* \\ & - 0.223 \cdot X_5^* + 0.361 \cdot X_6^* + 0.213 \cdot X_7^* + 0.297 \cdot X_8^* \\ & + 0.195 \cdot X_9^* + 0.263 \cdot X_{10}^* + 0.271 \cdot X_{11}^* \end{aligned}$$

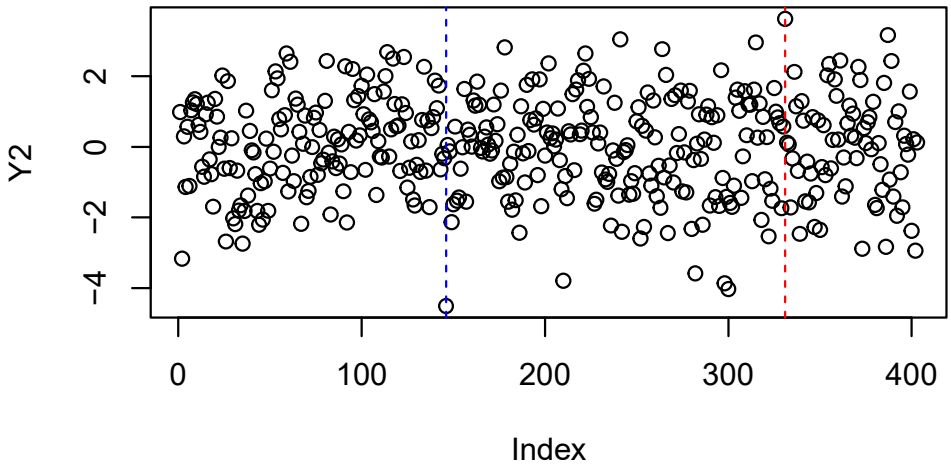
¿Quién influye en Y2? ¿Qué es? ¿A qué jugadores representa?

Esta es prácticamente lo contrario a lo anterior, parece indicar aquellos jugadores que saltan mucho, además de que tienen las variables referidas a la grasa, el sprint y la agilidad negativas, lo cual es bueno también, porque cuanta menos grasa mejor y también quiere decir que son veloces y ágiles. Lo único malo es que, a cambio de la agilidad, no tienen peso pero compensan con el salto, por lo que esta componente parece indicar jugadores veloces y con buen salto, pero no grandes en proporciones corporales.

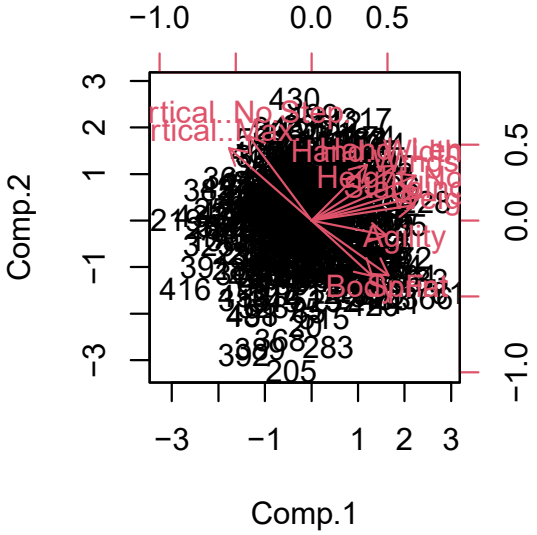
Podemos comprobar qué atleta es el máximo y mínimo de cada componente principal.

Se corresponden en 31 y 160 y el 331 y 146, respectivamente. Si quisiéramos verlo gráficamente.





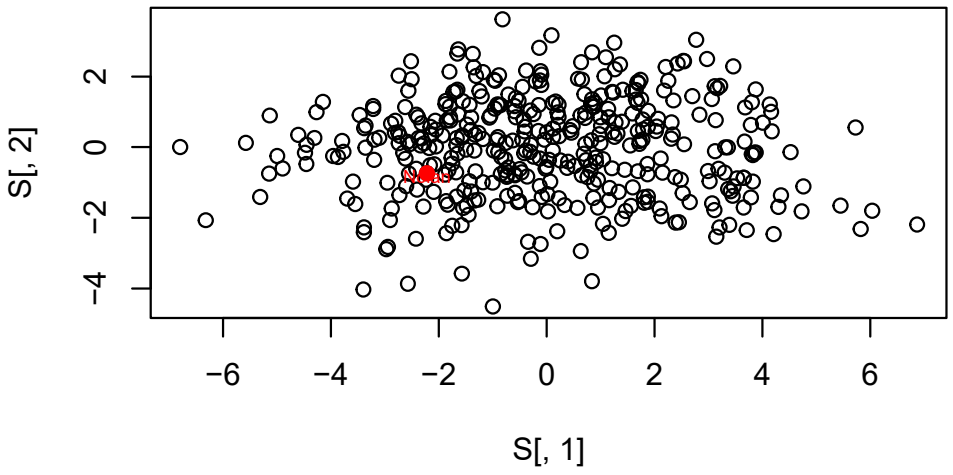
También podemos hacer un gráfico de las componentes principales, en este caso de la primera y segunda, que son las más representativas.



A la vista del gráfico anterior, podemos decir que los jugadores que se encuentran arriba a la

izquierda serán buenos saltadores y que corren y tienen poca grasa corporal, aunque no son muy altos ni de grandes dimensiones, por lo que son gente ágil. Por otro lado, aquellos que se encuentren arriba a la derecha serán más equilibrados en cuanto a salto y velocidad, pero altos y con envergadura. Los peores jugadores se encontrarán abajo a la izquierda, pues son los más delgados y con menores proporciones físicas, además de que no saltan ni corren demasiado. Los que se encuentran debajo a la derecha tampoco son demasiado buenos, pero por lo menos son algo más fuertes que los anteriores.

Si quisiéramos buscar a un jugador en específico, por ejemplo a Nolan Smith. Veremos que es normalillo en general.



Si recordamos la variable Draft.pick, esta nos decía si un jugador había sido seleccionado o no, y de haber sido seleccionado en qué orden. Tiene sentido pensar que aquellos atletas con peores aptitudes físicas serán los no seleccionados (Na), o su valor de Draft.pick será elevado (recordemos que el máximo era 60). Vamos a hacer algunas inferencias para ver si esto es cierto.

Puesto que habíamos dicho que los peores jugadores se encuentran abajo a la izquierda del biplot, podemos coger varios individuos de esta región para confirmar nuestra sospecha.

```
392 = NA
```

```
391 = NA
```

```
389 = NA

368 = 51

205 = NA
```

Vemos que casi todos son Na menos el 368, pero tienen el índice de selección elevado.

Siguiendo esta linea de razonamiento, los que están arriba deberían haber sido seleccionados y con un valor relativamente pequeño.

```
360 = 41

195 = 55

239 = 4

199 = 7

346 = 21
```

Vemos que todos han sido elegidos, un par con valores elevados. Esto quizás se deba a que son jugadores demasiado extremos (mucha fuerza pero poco salto y/o viceversa).

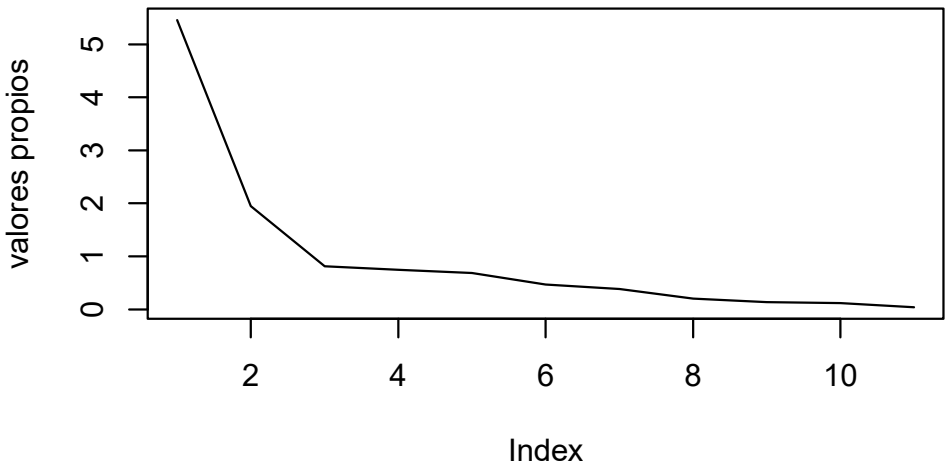
c) Número de componentes

Desde un principio parecía interesante quedarnos con 6 componentes principales ya que reduce el número de variables a la mitad y aporta un 90% de información, pero no hemos mirado otro posible número de componentes. Para ello podemos usar diversas reglas.

- Regla de Rao: solo serán relevantes las componentes que tengan una variabilidad mayor que la variabilidad mínima de las variables originales. En la matriz de correlaciones, esto equivale a que su desviación estándar sea mayor de 1. En este caso obtenemos 2 componentes.

```
   Comp.1    Comp.2    Comp.3
2.3362380 1.3951563 0.9007807
```

- Regla del codo: consiste en mirar en qué punto hay un cambio drástico de pérdida de información y después se estabiliza (con cierta pendiente). Al punto resultante habría que quitarle 1, en este caso sale $3 - 1 = 2$, es decir, 2 componentes principales.



- Prueba de esfericidad: establecemos como hipótesis nula que las variables no están correlacionadas y tienen igual varianza, y como hipótesis alternativa que las variables sí están correlacionadas. Supondremos m = 2, pues parece que 2 componentes serán lo más prometedor.

```
[1] 0
```

Como el p-valor es menor de 0.05, no podemos afirmar que las variables no están relacionadas, por lo que parece que las variables están correlacionadas entre sí. Si aumentamos el número de componentes a 6, también se obtiene 0, lo que deja entre ver que están correlacionadas.

```
[1] 0
```

Por tanto, podríamos trabajar con las dos primeras componentes principales, de las cuales la primera componente usa a todas las variables. Ya que no tenemos demasiadas variables, podemos considerar trabajar con todas ellas y reducir la dimensionalidad del momento en cada caso.

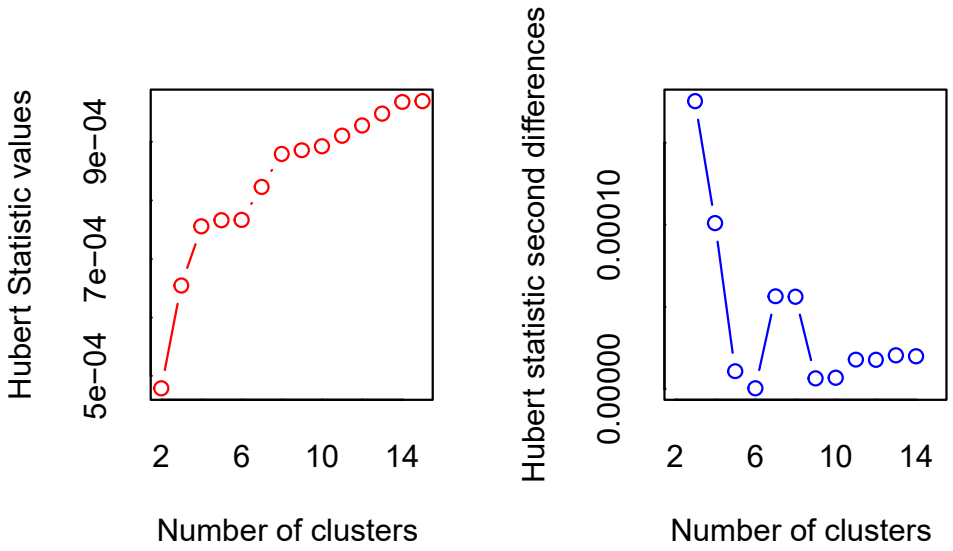
3. Análisis clúster

Puesto que no sabemos los grupos en los que se pueden agrupar los individuos, usaremos las técnicas no supervisadas para intentar agruparlos en el número de grupos más prometedor.

a) Número de grupos

Puesto que son bastantes individuos y en términos de deportes no hay una división clara sino que depende de que se quiera estudiar, el número de grupos puede ser muy arbitrario. En un principio, podría ser sensato considerar 3 grupos para clasificar a aquellos que son buenos, los no tan buenos y los peores que es posible que no hayan sido elegidos en el draft.

Sin embargo, puesto que existen ciertas técnicas que nos pueden ayudar a elegir cuantos grupos escoger, las usaremos como punto de partida, teniendo en cuenta también nuestros intereses.



*** : 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.