Es un sistema de computación distribuida en tiempo real orientado a procesar flujos constantes de datos, por ejemplo, datos que provienen de Twitter, pudiendo realizar estudios sobre *"trending topics"* al momento.

4. **Lenguaje R.**
   Es un lenguaje y un entorno de software frecuentemente usado para el cálculo estadístico y la visualización de gráficos. Es utilizado para la minería de datos, la investigación bioinformática y las matemáticas financieras.
   R se asemeja más a un lenguaje matemático más que a un lenguaje de programación, por lo que puede ser un inconveniente para los programadores para realizar análisis de Big Data. Su punto fuerte es el gran número de librerías creadas por la comunidad entre otras herramientas.

5. **Python**
   Es un lenguaje avanzado cuya ventaja a otros lenguajes es su uso relativamente fácil para usuarios que no están familiarizados con la programación, pero que necesitan trabajar con análisis de datos.
   También dispone de una gran comunidad detrás de este lenguaje que proporcionan un gran número de librerías, haciendo de Python un lenguaje muy eficiente para realizar Big Data.

3. DESARROLLO

3.1 PREPARACIÓN DE LOS DATOS

3.1.1 Clasificación entre: ALIVE, DEATH, NO-DATA y MD/GR

En esta fase, se ha realizado una clasificación de los datos dados según 4 resultados finales:

- Fallecidos o con discapacidades severas (SD-D).
- Con discapacidad moderada o buena recuperación (MR-GR).
- Vivos (pero sin resultados finales).
- Sin datos.

Para este procesado se han tenido en cuenta principalmente las siguientes variables:

- EO_Outcome
- EO_Symptoms
- TH_Outcome
- TH_Symptoms
- GOS5
- GOS8

Cuando las variables de GOS5 y GOS8 tienen datos

Si las filas ya contenían datos en las columnas de GOS5 y GOS8, directamente se han clasificado -según estas variables-. De lo contrario, se ha tenido que analizar las otras variables.

```
head(datos.modelo[,c(17,18,27,28,29,30)])

##   EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8
## 1          4           1         NA          NA <NA> <NA>
## 2          4           3         NA          NA <NA>  MD+
## 3          4           2         NA          NA  SD* <NA>
## 4          4           2         NA          NA <NA>  GR+
## 5          4           1         NA          NA <NA> <NA>
## 6          4           2         NA          NA <NA>  SD-
```

Cuando las variables de GOS5 y GOS8 no tienen datos

Si las variables de *"outcome"* contenían el valor de 1 (fallecimiento) o las variables de *"Symptoms"* contenían el valor de 6, directamente esas filas del conjunto de datos pasaban a clasificarse como fallecidos.

```
##    EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8 TH_Cause
## 22          1           6         NA          NA <NA> <NA>       NA
## 38          1           6         NA          NA <NA> <NA>       NA
## 50          1           6         NA          NA <NA> <NA>       NA
## 55          1           6         NA          NA <NA> <NA>       NA
## 61          1           6         NA          NA <NA> <NA>       NA
## 85          1           6         NA          NA <NA> <NA>       NA
```

Si las variables de *"outcome"* contenían el valor de 4 (alta) y las de "*Symptoms"* el valor de 1, entonces se han clasificado como "Vivos (pero sin resultados finales)".

```
##    EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8 TH_Cause
## 1           4           1         NA          NA <NA> <NA>       NA
## 5           4           1         NA          NA <NA> <NA>       NA
## 18          4           1         NA          NA <NA> <NA>       NA
## 20          4           1         NA          NA <NA> <NA>       NA
## 36          4           1         NA          NA <NA> <NA>       NA
## 51          4           1         NA          NA <NA> <NA>       NA
```

Se clasificarán como "Sin datos" todas aquellas filas que no contengan valores ni en las columnas de "*Symptoms"*. Se tienen en cuenta los transferidos a otros hospitales.

```
##      EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8 TH_Cause
## 384          NA          NA         NA          NA <NA> <NA>       NA
## 417          NA          NA         NA          NA <NA> <NA>       NA
## 985          NA          NA         NA          NA <NA> <NA>       NA
## 997          NA          NA         NA          NA <NA> <NA>       NA
## 2270         NA          NA         NA          NA <NA> <NA>       NA
## 2292         NA          NA         NA          NA <NA> <NA>       NA
```

Si los "*Symptoms*" son de 4 o de 5 (Discapacidad Severa), entonces se clasificarán como "Fallecidos o con discapacidades severas".

```
##     EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8 TH_Cause
## 160          5           5         NA          NA <NA> <NA>       NA
## 241          5           5         NA          NA <NA> <NA>       NA
## 317          5           5         NA          NA <NA> <NA>       NA
## 336          5           5         NA          NA <NA> <NA>       NA
## 357          5           5         NA          NA <NA> <NA>       NA
## 573          5           5         NA          NA <NA> <NA>       NA
```

Así mismo, si las variables de *"outcome"* contenían el valor de 4 y las de *"Symptoms"* el valor de 9, significa que el paciente ha sido dado de alta, pero no se tiene ningún dato sobre el estado final, por lo tanto, se han incluido en la clasificación de "Vivos (pero sin resultados finales)".

```
##      EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8 TH_Cause
## 409          4           9         NA          NA <NA> <NA>       NA
## 1000         4           9         NA          NA <NA> <NA>       NA
## 4859         4           9         NA          NA <NA> <NA>       NA
```

Se han visto 3 elementos de "NODATA", cuyos pacientes obtienen un estado de *"Symptoms"* de 4, por lo que se envía a estado de fallecido, son datos anómalos.

```
##      EO_Outcome EO_Symptoms TH_Outcome TH_Symptoms GOS5 GOS8 TH_Cause
## 52            2           4         NA           4 <NA> <NA>        3
## 699           2           4         NA           4 <NA> <NA>       NA
## 3025          2           4         NA           4 <NA> <NA>        1
```

DATOS FINALES:

- Fallecidos o con discapacidades severas*:* 3559
- Con discapacidad moderada o buena recuperación*:* 5997
- Vivos (pero sin resultados finales)*:* 127
- Sin datos*:* 86
- Con NA*:* 239

3.1.2 Clasificación entre: ESCANEADOS y NO ESCANEADOS

En primer lugar, se han encontrado ciertos datos anómalos, en los que aparecen datos escaneados (1) y no tienen los datos del escáner, entonces deberíamos ponerlo como no escaneado (2).

```
##      EO_Head.CT.scan EO_1.or.more.PH EO_Subarachnoid.bleed
## 201                1              NA                    NA
## 314                1              NA                    NA
## 1277               1              NA                    NA
## 3234               1              NA                    NA
## 3687               1              NA                    NA
## 4256               1              NA                    NA
##      EO_Obliteration.3rdVorBC EO_Midline.shift..5mm EO_Non.evac.haem
## 201                        NA                    NA               NA
## 314                        NA                    NA               NA
## 1277                       NA                    NA               NA
## 3234                       NA                    NA               NA
## 3687                       NA                    NA               NA
## 4256                       NA                    NA               NA
##      EO_Evac.haem
## 201            NA
## 314            NA
## 1277           NA
## 3234           NA
```

```
## 3687           NA
## 4256           NA
```

A continuación, se van a clasificar los datos como:

- Escaneados
- No escaneados
- En analisis

Si el *"Outcome"* es 2 (el paciente se ha transferido a otro hospital), se ha escaneado en dicho hospital ("TH_SCAN") y no se tiene ninguna información en los escáneres, se clasificarán como "En análisis".

```
##     EO_Outcome TH_Head.CT.scan TH_1.or.more.PH TH_Subarachnoid.bleed
## 52           2            <NA>              NA                    NA
## 128          2            <NA>              NA                    NA
## 135          2            <NA>              NA                    NA
## 188          2            <NA>              NA                    NA
## 193          2            <NA>              NA                    NA
## 207          2            <NA>              NA                    NA
```

Sobre el dataset **NO ESCANEADO**: Si el *"Outcome"* es 2 (el paciente se ha transferido a otro hospital) y no se ha realizado ningún escáner, pero si contiene datos en el escáner, entonces se clasificará como "Escaneado".

```
##      EO_Outcome TH_Head.CT.scan TH_1.or.more.PH TH_Subarachnoid.bleed
## 201           2               1               2                     2
## 217           2               1               2                     1
## 257           2               1               1                     2
## 314           2               1               1                     2
## 318           2               1               2                     2
## 1184          2               1               2                     2
```

Sobre el dataset **NO ESCANEADO**: Nos hemos dado cuenta que existen datos anómalos, que contienen varios escáneres, pero, sin embargo, no se indica como escaneado, son los registros: 2628,3276,3279,8469,8655, etc. (En total son 12)

```
##      EO_Head.CT.scan EO_1.or.more.PH EO_Subarachnoid.bleed
## 2628               2               2                     2
## 3276               2               2                     2
## 3279               2               2                     2
## 3720               2               2                     2
```

```
## 7286               2                2                2
## 8469               2                2                2
```

Sobre el dataset **EN ANALISIS**: Nos hemos dado cuenta de que existen datos anómalos. Para las variables de los pacientes que se han transferido a otro hospital (TH), existen variables de escáner *("TH_Head.CT.scan")* que se encuentran vacías, junto con el resto de variables del escáner en particular. Por lo tanto, se ha asignado el valor de 2 a la variable de escáner *("TH_Head.CT.scan")* y se han incluido en los escaneados, puesto que en todos ellos, en la variable *"EO_Head.CT.scan"* sí que existe un valor de 1 (escaneados) y no se han encontrado más anomalías en dichos datos.

```
##      EO_Head.CT.scan EO_1.or.more.PH EO_Outcome TH_Head.CT.scan
## 681                1               2          2            <NA>
## 1639               1               2          2            <NA>
## 5743               1               2          2            <NA>
## 8434               1               2          2            <NA>
## 8972               1               2          2            <NA>
##      TH_1.or.more.PH TH_Subarachnoid.bleed
## 681               NA                    NA
## 1639              NA                    NA
## 5743              NA                    NA
## 8434              NA                    NA
## 8972              NA                    NA
```

Sobre el dataset **ESCANEADO**: Se van a eliminar todas las filas que no tengan información en el *"TH_Major.EC.injury"* y en el *"EO_Major.EC.injury"*.

```
##     EO_Outcome TH_Major.EC.injury
## 76           2                 NA
## 90           2                 NA
## 315          2                 NA
## 361          2                 NA
## 510          2                 NA
## 565          2                 NA
```

Sobre el dataset **ESCANEADO**: Comprobamos que las variables: *"EO_Cause"* y *"EO_Symptoms"*, no contengan valores nulos.

```
##     EO_Cause EO_Major.EC.injury
## 177        3                 2
## 211       NA                 1
## 242       NA                 2
## 255        2                 2
```

```
## 293        NA                   2
## 321        NA                   1
```

Sobre el dataset **ESCANEADO**: Comprobamos que la variable: *"EO_Outcome"* no se encuentre nula. (En total son 2 registros).

```
##      EO_Cause EO_Outcome
## 9036        3         NA
## 9333        3         NA
```

Sobre el dataset **ESCANEADO**: Comprobamos que existe un valor anómalo (que se sale del rango) en un registro en la columna de *"EO_Major.EC.Injury"*. Este valor lo cambiaremos a positivo -> 1.

```
##      EO_Cause EO_Major.EC.injury
## 3862        2                -1
```

DATOS FINALES:

- Vivos y escaneados: 4157
- Vivos y no escaneados: 1535
- Vivos en análisis: 305
- Fallecidos y escaneados: 2829
- Fallecidos y no escaneados: 439
- Fallecidos en análisis: 291

3.1.3 Eliminación y centralización de variables

Se va a centralizar las variables de *"PUPIL_REACT_LEFT"* y *"PUPIL_REACT_RIGHT".*

```
##   PUPIL_REACT_LEFT PUPIL_REACT_RIGHT ESTADOESCANER
## 1                1                 1      SCANEADO
## 2                1                 1      SCANEADO
## 3                1                 1      SCANEADO
## 4                1                 1      SCANEADO
## 5                1                 1      SCANEADO
## 6                1                 1      SCANEADO
```

- Both reactive: 5662
- No response unilateral: 497
- No response: 634

- Unable to assess: 193

Ahora vamos a ver si podemos prescindir o aunar las variables de *"EO_Cause"* y *"TH_Cause"*. Para ello veremos en qué caso, ambas variables difieren:

```
##      EO_Cause TH_Cause
## 2307        1        3
## 2813        3        1
## 3285        2        3
## 4021        3        1
```

Como se puede observar, podríamos prescindir de la variable *"TH_Cause"*, puesto que recoge la misma información que *"EO_Cause"*.

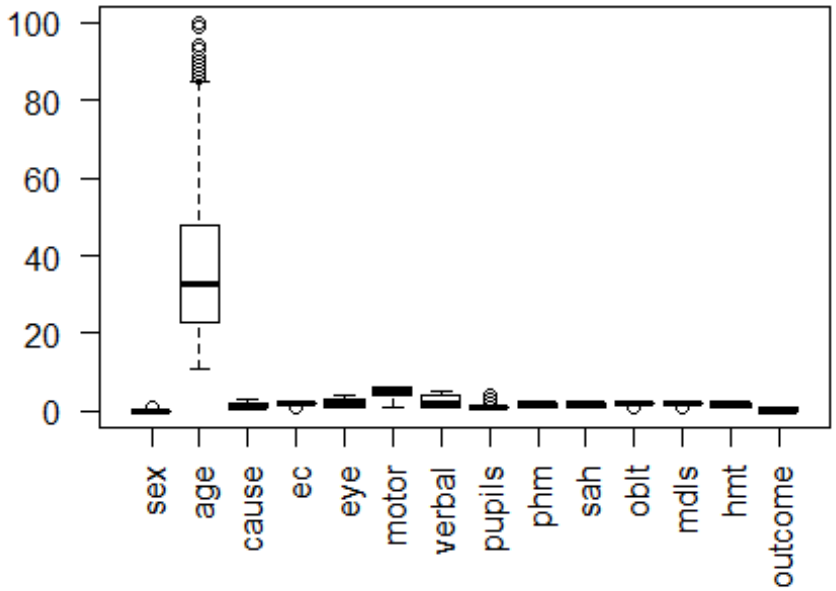
A continuación, vamos a aunar todas las variables del escáner. Si un paciente ha sido transferido a otro hospital y se han realizado los escáneres en dicho hospital, entonces, se mantendrán los últimos valores del escáner. En caso contrario, se usarán los primeros resultados de escáner obtenidos en el primer. Además, eliminaremos todas las variables que no se utilicen.

```
##   sex age cause ec eye motor verbal pupils phm sah oblt mdls hmt outco
me
## 1   0  11     1  1   1     5      1      1   2   2    2    2   2    MD
GR
## 2   0  14     1  2   1     2      1      1   1   2    2    2   1
D
## 3   0  14     1  2   2     5      1      1   2   2    2    2   1
D
## 4   0  14     1  2   2     5      2      1   2   2    2    2   2    MD
GR
## 5   0  14     3  2   4     6      4      1   2   1    2    2   2    MD
GR
## 6   0  15     1  2   1     5      1      1   2   2    2    2   2
D
```

3.2 PRE-PROCESADO DE LOS DATOS

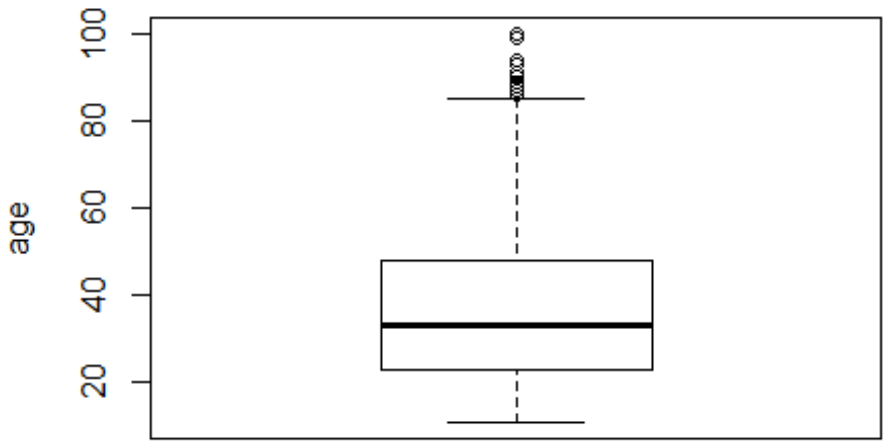
3.2.1 Búsqueda de OUTLIERS (datos anómalos)

En primer lugar, se ha realizado un diagrama para cada una de las variables del conjunto de datos.



En este gráfico, nos hemos dado cuenta que la variable que contiene un gran numero datos anómalos es *"age"*.

A continuación, pasamos a estudiar a fondo los motivos que producen que esta variable tenga datos anómalos y comprobaremos si es necesario o no la eliminación de dichos datos.



Estas son las edades anómalas:

```
##  [1]  86  86  86  86  86  86  87  87  87  87  88  88  88  88  89  89
90
## [18]  91  91  91  93  93  94  86  86  86  86  86  86  86  87  87  87
87
## [35]  87  87  87  88  88  88  88  89  89  89  89  89  89  89  90  90
91
## [52]  91  93  94  99 100
```

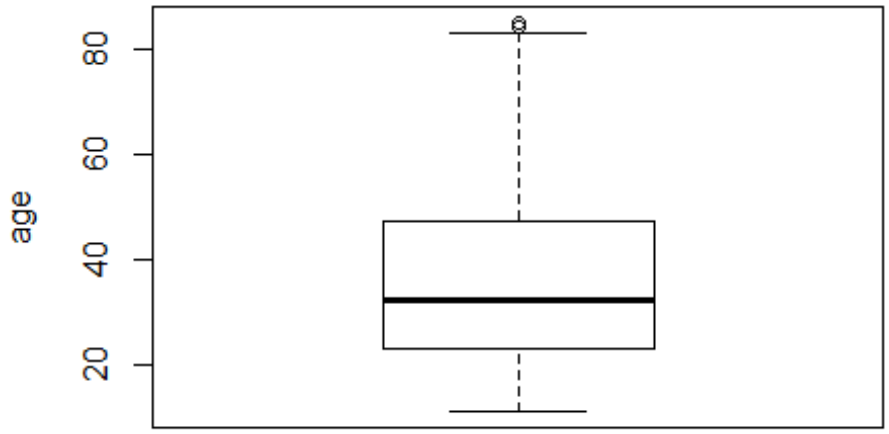
Antes de proceder a eliminar los datos anómalos, vamos a ver la correlación existente con la variable *"Outcome"*: 0.2598931

También vamos a observar como es la media y la mediana:

- Media: 37.02119
- Mediana: 33

A continuación, vamos a proceder con la eliminación relativa de los datos anomalos para ver cómo afecta al conjunto de los datos. Para ello hemos creado una función que nos

elimine directamente los datos anómalos, es decir, todos aquellos datos que no se encuentren en el rango Q1-1.5·RIC o superiores a Q3+1.5·RIC se eliminaran. Siendo RIC el rango intercuartil (Q1-Q3)



La correlación obtenida posteriormente a la eliminación de los datos anómalos con la variable de *"Outcome"* es: 0.2452825. Vemos que la correlación ha empeorado un poco. De todas formas, la correlación entre la edad y el *"outcome"* es bastante débil.

También vamos a observar como es la media y la mediana:

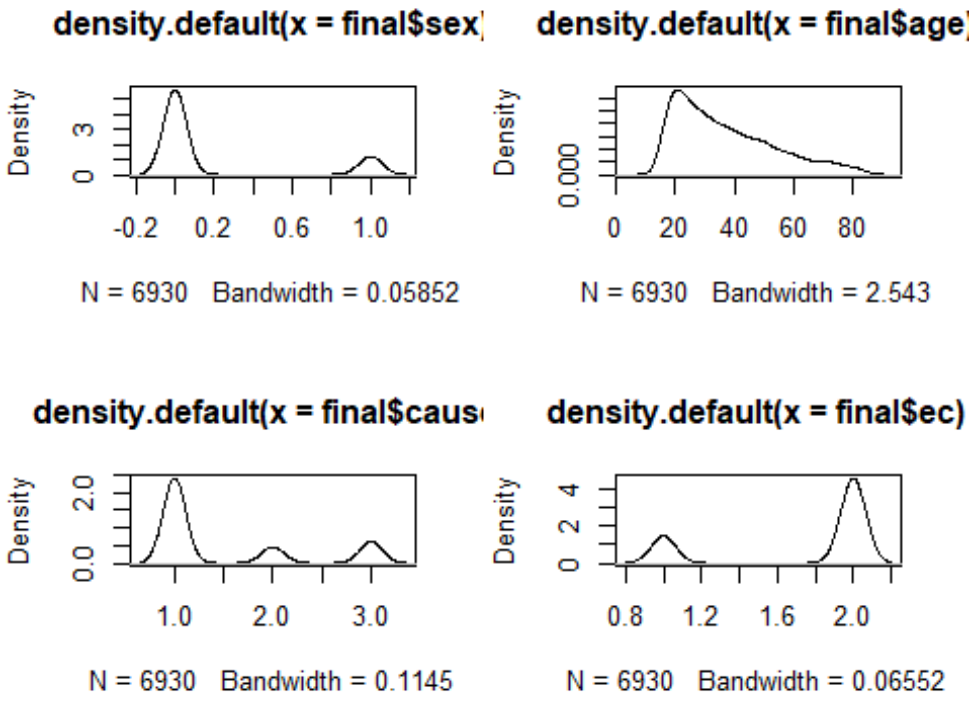
- Media: 36.60303
- Mediana: 32

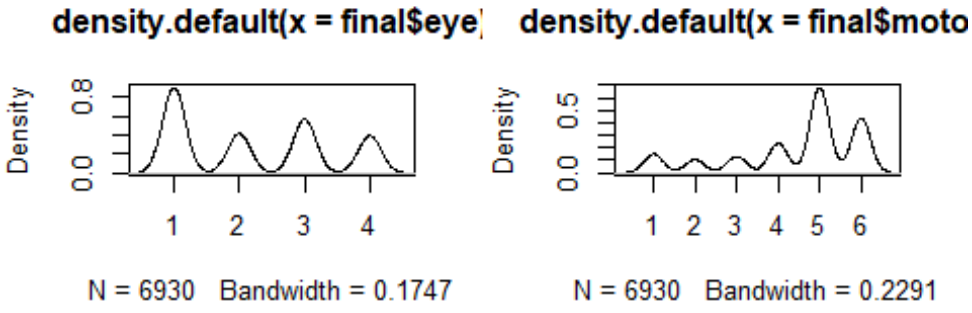
Como se puede comprobar, la eliminación de los "*outliers"* no ha afectado demasiado a las variables estadísticos por lo que no existe motivo para su eliminación.

Por otro lado, es necesario destacar que estos pacientes cuya edad es anómala (estadísticamente), en la naturaleza tampoco se consideran pacientes anómalos, ya que se encuentran en un rango de edades en las que sufrir un traumatismo craneocefálico es totalmente posible.

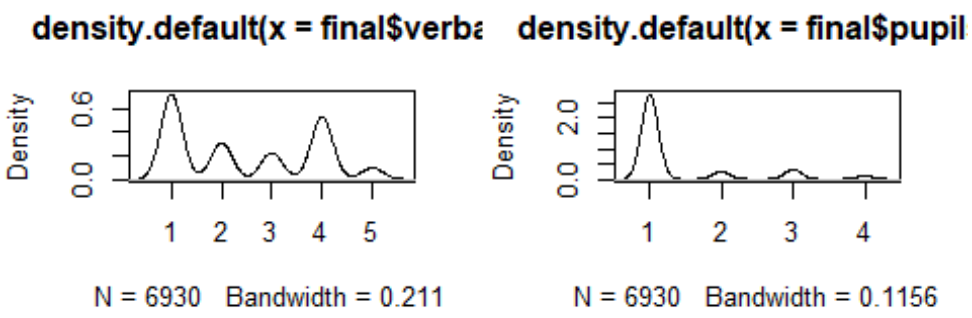
3.2.2 Análisis de normalidad

En primer lugar, visualizaremos la densidad de nuestras variables (individualmente), con el objetivo de observar a simple vista si cumplen o no con una distribución normal.

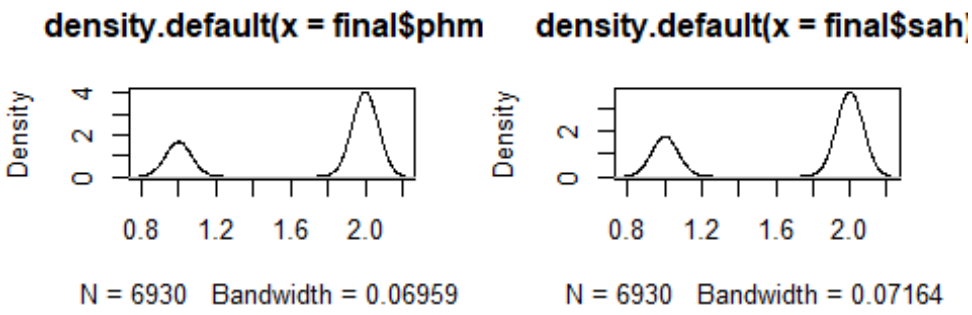



density.default(x = final$eye
Density
0.0
0.8
1
2
3
4
N = 6930 Bandwidth = 0.1747

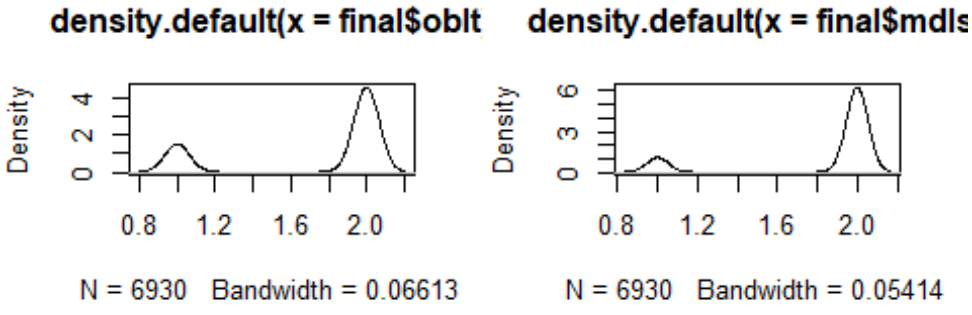
density.default(x = final$moto
Density
0.0
0.5
1
2
3
4
5
6
N = 6930 Bandwidth = 0.2291


density.default(x = final$verba
Density
0.0
0.6
1
2
3
4
5
N = 6930 Bandwidth = 0.211

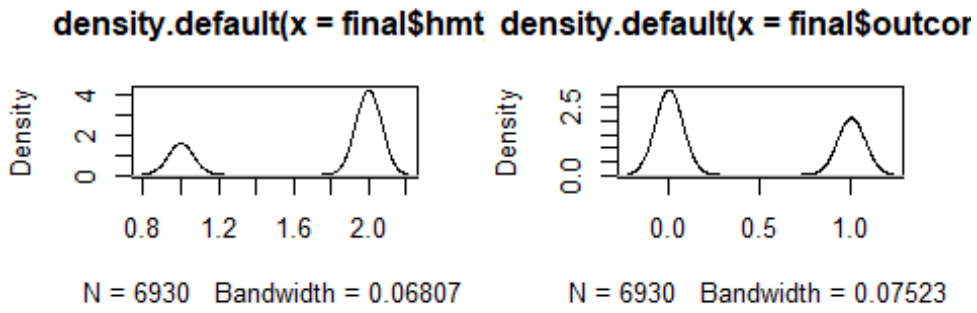
density.default(x = final$pupil
Density
0.0
2.0
1
2
3
4
N = 6930 Bandwidth = 0.1156


density.default(x = final$phm
Density
0
2
4
0.8
1.2
1.6
2.0
N = 6930 Bandwidth = 0.06959

density.default(x = final$sah
Density
0
2
0.8
1.2
1.6
2.0
N = 6930 Bandwidth = 0.07164


density.default(x = final$oblt
Density
0
2
4
0.8
1.2
1.6
2.0
N = 6930 Bandwidth = 0.06613

density.default(x = final$mdls
Density
0
3
6
0.8
1.2
1.6
2.0
N = 6930 Bandwidth = 0.05414



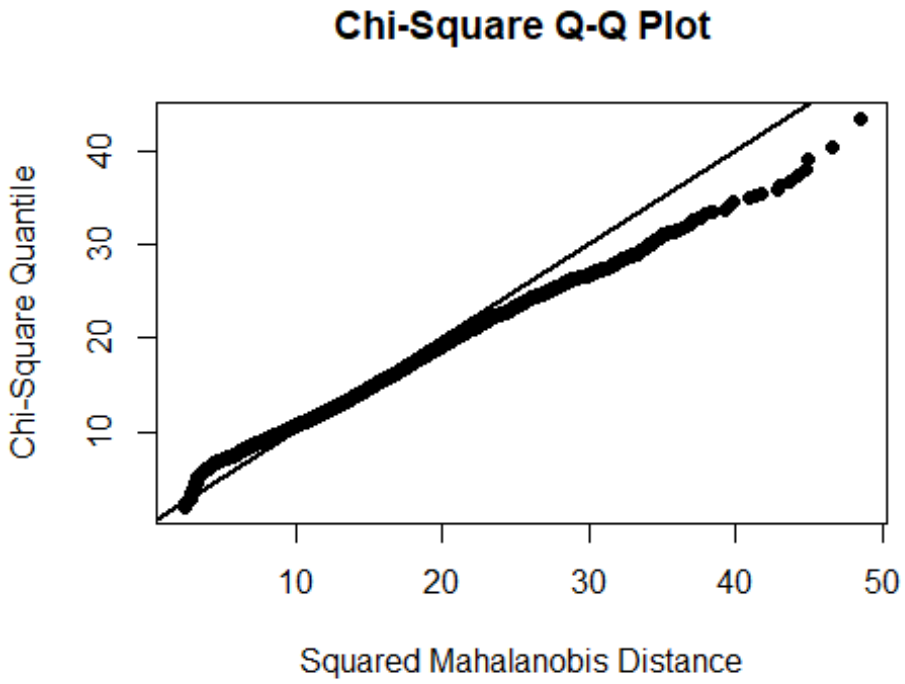
Como podemos comprobar, al tratarse de variables discretas (excepto la variable de edad -*age*-), no lograremos conseguir una distribución normal de forma individual.

Otro aspecto a tener en cuenta es que para que un conjunto de datos (teniendo en cuenta todas las variables) posea una distribución normal, es necesario que todas las variables verifiquen normalidad univariante, ya que es una condición necesaria (aunque no suficiente). Por lo tanto, rechazamos la hipótesis de normalidad del conjunto de datos.

Aun así, comprobaremos los resultados obtenidos mediante el Test de normalidad de Mardia:

```
## [1] 2
    ## Mardia's test for class 1
    ## mard1= 34629.84
    ## pvalue for m3= 0
    ## mard2= 74.78288
    ## p-value for m4= 0
    ## There is not statistical evidence for normality in class 1
    ## Mardia's test for class 2
    ## mard1= 6201.334
    ## pvalue for m3= 0
    ## mard2= -7.620724
    ## p-value for m4= 2.531308e-14
    ## There is not statistical evidence for normality in class 2
```

También vamos a utilizar el test de Henze-Zirkler:

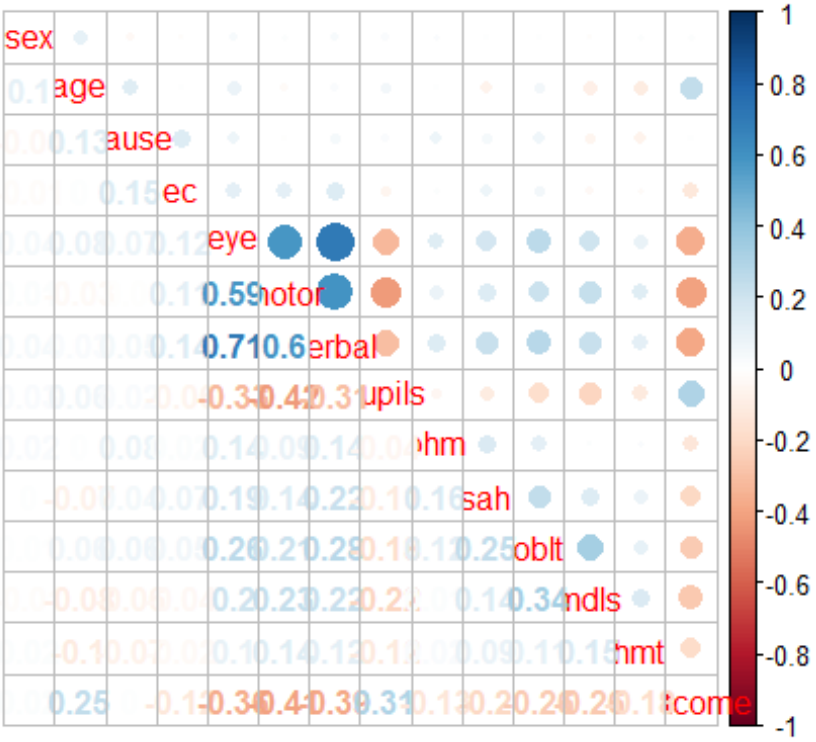



```
##             Henze-Zirkler test for Multivariate Normality
##
##   data : final
##
##   HZ               : 15.38209
##   p-value          : 0
##
##   Result   : Data are not multivariate normal (sig.level = 0.05)
```

Como se puede comprobar, al ser el p-value menor de 0.05 en ambos test, los datos no se ajustan a una distribución normal.

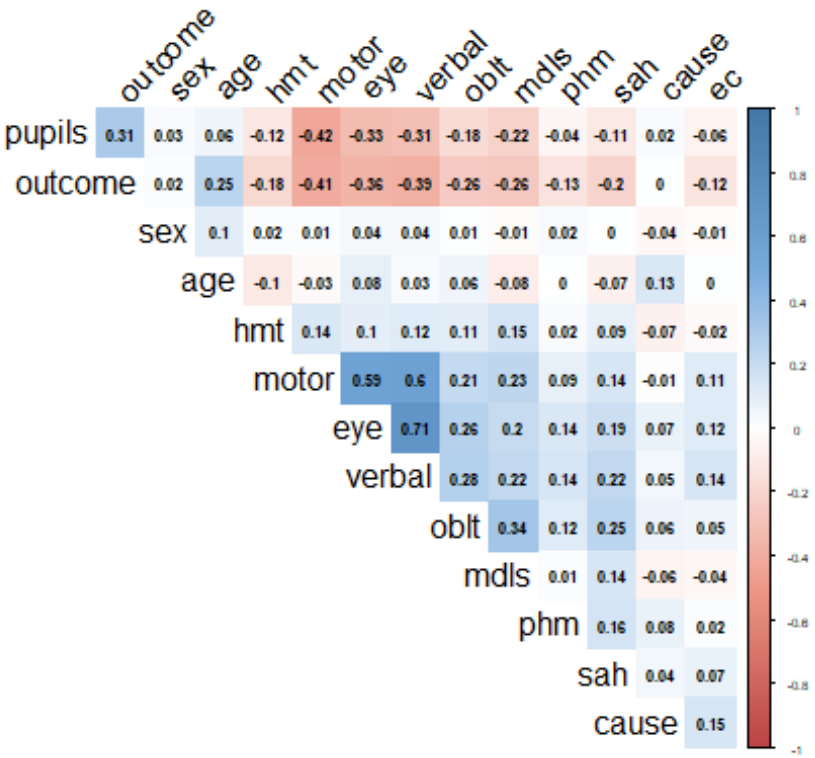
3.2.3 Estudio de correlación

Para el estudio de la correlación, utilizaremos el **coeficiente de correlación de Pearson (R)**. Mediante el siguiente gráfico, vamos a observar las relaciones que tienen los pares de variables entre sí.

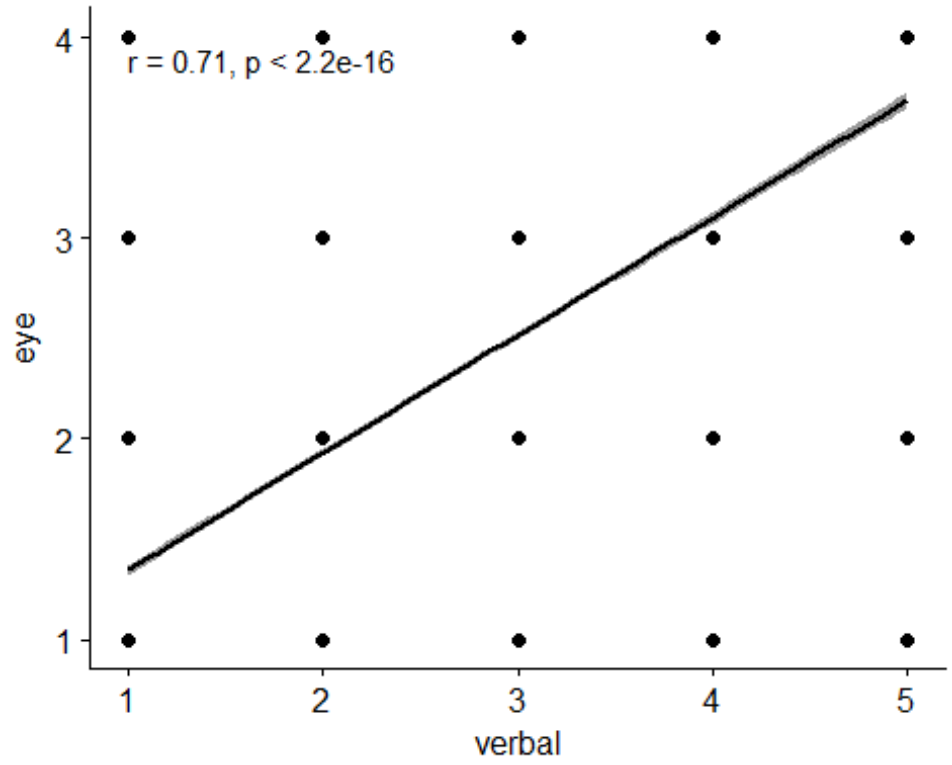


En este grafico podemos observar como por ejemplo las variables de *"motor"*, *"verbal"* y *"eye"* tienen bastante relación y dependencia entre sí. Sin embargo, hay algo que no nos cuadra y es que no existe una gran dependencia entre la variable "age" y la variable de "*outcome*", aspecto que podría ser más sustancial en la naturaleza.

Teniendo en cuenta los valores de la variable *"outcome"* (1 fallece y 0 vive), la correlación negativa de las variables del test de Glasgow (eye, motor, verbal) tiene sentido, puesto que en general, cuanto mayor sea el valor de estas variables, mejor pronóstico de vida hay. La variable de "*pupils*" es, al contrario, cuanto mayores sean sus valores, más probable es el pronóstico de fallecimiento.

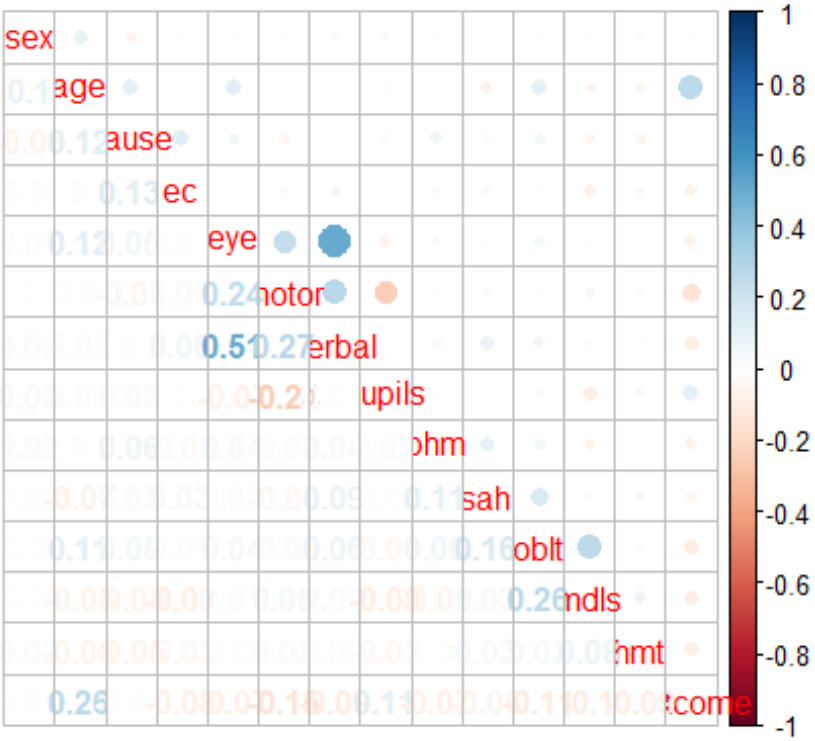


Como se ha podido apreciar en las 2 graficas anteriores, existe una gran correlación entre las variables *"motor"*, *"eye"* y *"verbal"*.



En la gráfica anterior podemos volver a comprobar que existe una gran relación lineal positiva entre las variables más correlacionadas que son "*verbal*" y "*eye*".

A continuación, se muestra la matriz de correlaciones parciales.



Con la matriz de correlaciones parciales, obtendremos las correlaciones parciales que existe entre los pares de variables eliminando el efecto de las restantes. Vemos que las correlaciones fuertes se encuentran entre los mismos pares de variables que en la matriz de correlación total.

A continuación, a modo de información, se muestra un listado en orden descendente con las mayores correlaciones existentes:

```
##     First.Variable Second.Variable Correlation
## 89             eye          verbal   0.7088056
## 90           motor          verbal   0.5989863
## 75             eye           motor   0.5863534
## 104          motor          pupils  -0.4224881
## 188          motor         outcome  -0.4098411
## 189         verbal         outcome  -0.3893489
## 187            eye         outcome  -0.3618281
## 165           oblt            mdls   0.3377311
## 103            eye          pupils  -0.3254853
## 190         pupils         outcome   0.3099949
## 105         verbal          pupils  -0.3073014
## 147         verbal            oblt   0.2765892
## 194           mdls         outcome  -0.2634093
## 145            eye            oblt   0.2609820
## 193           oblt         outcome  -0.2570853
```

Las correlaciones entre las variables y la clase ordenadas en orden descendente son las siguientes:

```
##     First.Variable Second.Variable Correlation
## 188          motor         outcome -0.40984112
## 189         verbal         outcome -0.38934889
## 187            eye         outcome -0.36182805
## 190         pupils         outcome  0.30999486
## 194           mdls         outcome -0.26340926
## 193           oblt         outcome -0.25708526
## 184            age         outcome  0.24528254
## 192            sah         outcome -0.20016386
## 195            hmt         outcome -0.18246423
## 191            phm         outcome -0.13254914
## 186             ec         outcome -0.12180868
## 183            sex         outcome  0.01988148
## 185          cause         outcome  0.00362152
## 196        outcome         outcome  0.00000000
```

Por consiguiente, consideramos que, aunque exista una correlación importante entre las variables "*eye*", "*verbal*" y "*motor*", no es lo suficientemente fuerte como para concluir que estas variables contienen la misma información y sea necesario la eliminación de algunas de ellas. Por lo que no se procede a descartar ninguna de estas variables en estudios posteriores.

3.2.4 Selección de variables con más importancia

3.2.4.1 Uso de "Random Forest"

En este apartado, se buscará obtener un listado con las variables más importantes, usando el algoritmo de *"Random Forest"* para posteriormente tener en cuenta posibles descartes de variables en estudios posteriores.

La idea que existe detrás de los *"Random Forest"* es generar un número importante de árboles, entrenarlos y calcular el promedio de su salida.

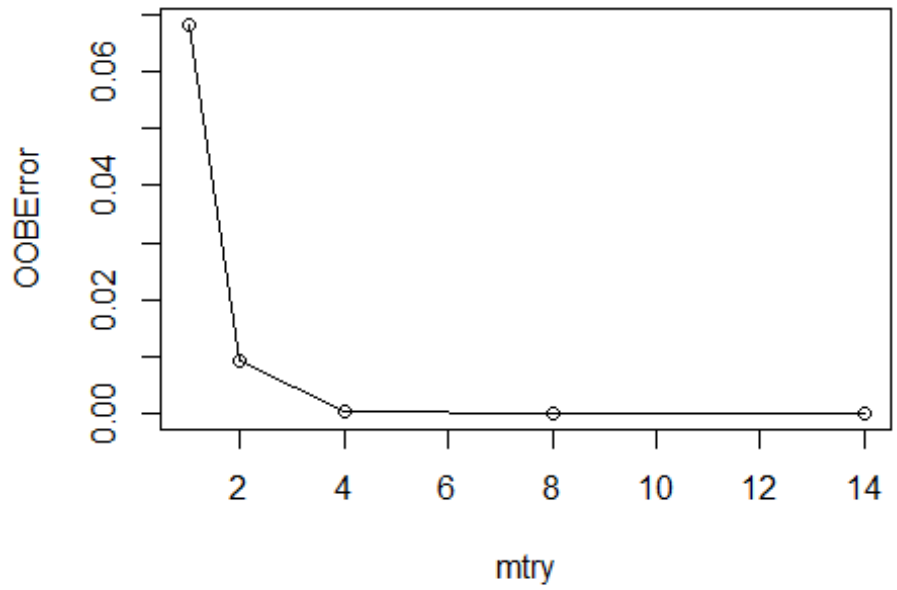
En cada iteración del algoritmo de *"Random Forest"* se genera un error conocido como **OOB**, este error ira aumentando o disminuyendo en cada iteración y por cada variable que se incluya en el algoritmo.

En cada paso (nodo) se recalcula el conjunto de **"m"** predictores permitidos. Lo más típico es elegir la raíz cuadrada del número total de variables. En nuestro caso, contamos con un total de 13 variables, por lo que se escogerían 4 variables (redondeando hacia arriba en caso

de no ser un número entero) en el caso de **árboles de clasificación** y **m=p/3** en el caso de **árboles de regresión**. Siendo **"p"** el número de variables.

Aun así, es necesario calcular la variable **"mtry"**, puesto que es el único parámetro ajustable al cual los bosques aleatorios son algo sensibles. El "mtry" es el número de variables aleatorias utilizadas en cada árbol. La reduccion del "mtry" reduce tanto la correlación como la fuerza, aumentando ambas en caso contrario.

En algún punto intermedio hay un rango "óptimo" de "mtry", generalmente bastante ancho.Usando la tasa de error de OOB, se puede encontrar rápidamente un valor de mínimo en el rango.




```
##    mtry     OOBError
## 1     1 6.830054e-02
## 2     2 9.155807e-03
## 4     4 2.020574e-04
## 8     8 8.633286e-07
## 14   14 1.526225e-07
```

Como se puede comprobar, el error OOB, se estabiliza, indicando cuantas particiones se deben realizar para obtener los mejores resultados En este caso, con **4 variables sería suficiente** (puesto que es el numero donde se estabiliza el error OOB).

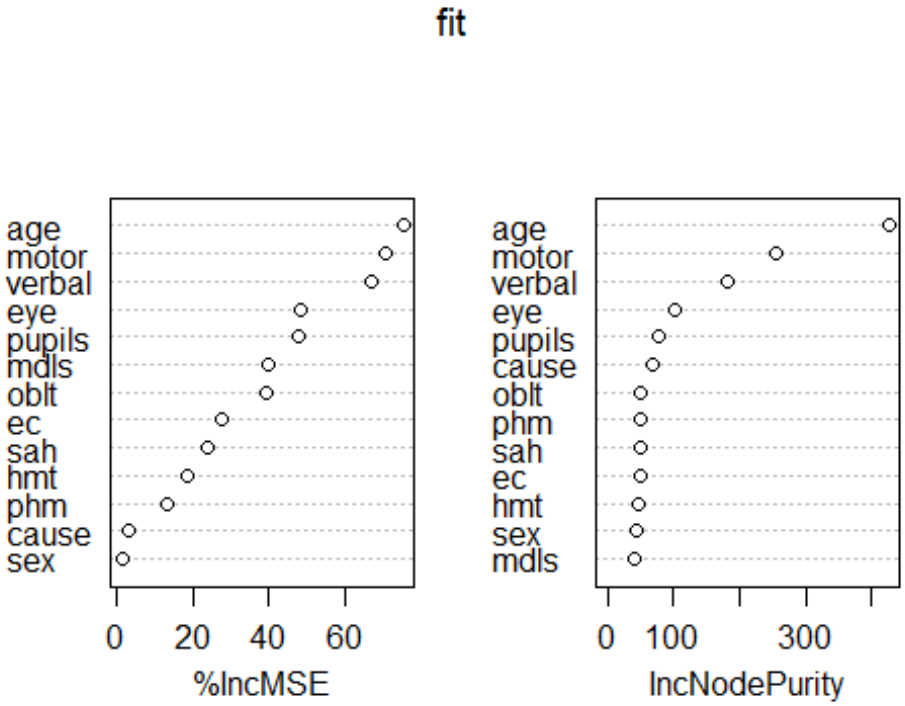
Las variables más importantes utilizando el "mtry" son las siguientes:

```
##          X.IncMSE IncNodePurity
## sex      1.722602      43.18384
## cause    3.414805      67.70075
## phm     13.530135      49.85977
## hmt     18.625224      45.77766
## sah     23.916382      49.68159
## ec      27.739004      49.00917
## oblt    39.137262      50.54574
## mdls    39.767808      40.05219
## pupils 47.516489      76.13479
## eye     48.270418     101.46189
## verbal 66.382751     181.83672
## motor  70.147994     255.89314
## age     74.945217     426.70467
```

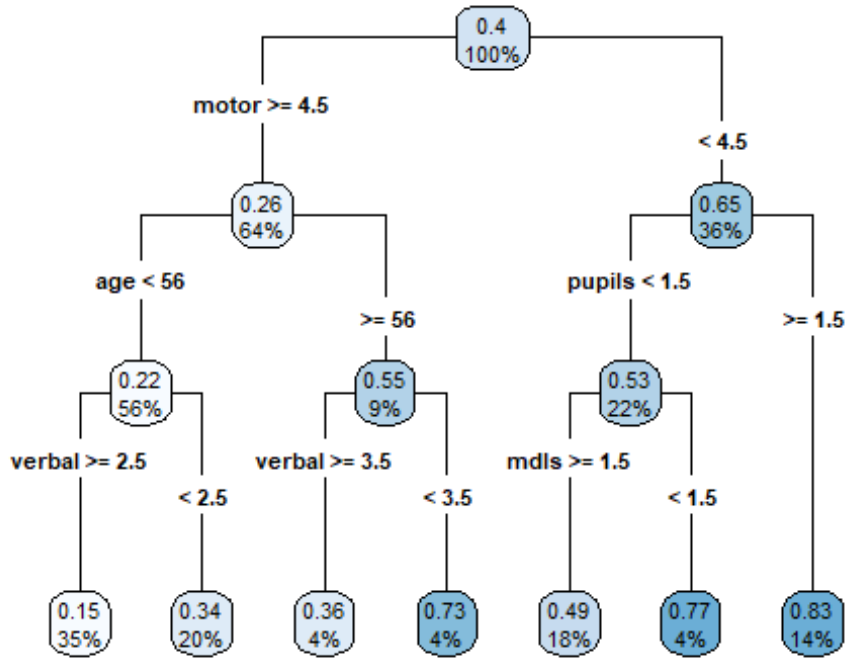
La variable **"IncNodePurity"** se la conoce también como la media de decrecimiento de de Gini. El índice de Gini es una "medida de desorden" en este caso *"IncNodePurity"* tiene el siguiente sentido, a mayor medida, mayor importancia en los modelos creados, puesto que valores próximos a 0 implican un mayor desorden. Por tanto, si computamos la media del "decrecimiento" del índice de Gini cuanto mayor sea esta medida, mas variabilidad aporta a la variable dependiente.

Por otro lado, la variable **"IncMSE"** es la media de decrecimiento en la precisión, y es también un indicador sobre la importancia de las variables en el modelo.

El siguiente grafico representa la importancia de las variables según su media y los valores de *"Random Forest"* mostrados anteriormente:



A continuación, se va a utilizar un árbol de clasificación, que nos mostrara la importancia de las variables según este algoritmo de clasificación.




```
## n= 6930
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 6930 1664.39500 0.4008658
##    2) motor>=4.5 4458  864.83090 0.2633468
##      4) age< 56.5 3861  661.19090 0.2193732
##        8) verbal>=2.5 2445  306.28790 0.1468303 *
##        9) verbal< 2.5 1416  319.81920 0.3446328 *
##      5) age>=56.5 597  147.88940 0.5477387
##       10) verbal>=3.5 296   68.32095 0.3614865 *
##       11) verbal< 3.5 301   59.20266 0.7308970 *
##    3) motor< 4.5 2472  563.21680 0.6488673
##      6) pupils< 1.5 1504  374.27060 0.5339096
##       12) mdls>=1.5 1260  314.79680 0.4873016 *
##       13) mdls< 1.5 244   42.60246 0.7745902 *
##      7) pupils>=1.5 968  138.18900 0.8274793 *
```

Si recordamos, un resultado en el *"outcome"* de 0 eran aquellos pacientes que a los 6 meses habían vivido mientras que un resultado de 1, significaba que los pacientes fallecian. Teniendo en cuenta este dato, podemos observar que los nodos del árbol son aquellas

variables que el algoritmo considera más relevantes y que hacen que un paciente viva o fallezca.

La interpretación que se da al árbol es la siguiente: Cada nodo contiene el porcentaje de información que contiene además de la media de la variable *"outcome"* en cada particion. Por ejemplo, la media de *"outcome"* es de 0.4, que coincide con el 0.4 del nodo raiz. Sin embargo, cuando la variable *"motor"* es mayor de 4.5, entonces el número de datos se reduce al 64% y la media de *"outcome"* se vuelve a 0.26, significando para este caso que la mayoria de los pacientes viven, puesto que se aproxima a 0.

Como conclusiones, utilizaremos las variables que se han considerado como más importantes en el algoritmo del árbol de clasificación y son las siguientes: *"motor"*, *"age"*, *"pupils"*, *"verbal"* y *"mdls"*.

3.2.4.2 Uso del método de regresión paso a paso (Stepwise Regression)

Este método es uno de los que se utilizan en la selección algorítmica del modelo. Se utiliza para identificar aquellas variables que se deberán integrar o no en los modelos a estudiar.

*La lógica subyacente de este algoritmo consiste en conservar las variables independientes que contienen información relevante y a la vez prescindir de aquellas que resulten redundantes respecto de las que quedaron en el modelo.

```
##
## Call:
## glm(formula = outcome ~ ., family = binomial, data = final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7765  -0.7606  -0.3985   0.7762   2.6408
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.054157   0.310487  16.278  < 2e-16 ***
## sex          0.041738   0.080207   0.520    0.603
## age          0.041875   0.001976  21.187  < 2e-16 ***
## cause       -0.024701   0.040518  -0.610    0.542
## ec          -0.437802   0.069418  -6.307 2.85e-10 ***
## eye         -0.237643   0.038838  -6.119 9.43e-10 ***
## motor       -0.289910   0.025841 -11.219  < 2e-16 ***
## verbal      -0.237017   0.032541  -7.284 3.25e-13 ***
## pupils       0.378928   0.043887   8.634  < 2e-16 ***
## phm         -0.394038   0.065282  -6.036 1.58e-09 ***
## sah         -0.279699   0.065385  -4.278 1.89e-05 ***
```

```
## oblt        -0.638613   0.073955  -8.635  < 2e-16 ***
## mdls        -0.709317   0.091718  -7.734 1.04e-14 ***
## hmt         -0.488884   0.067500  -7.243 4.40e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9332.8  on 6929  degrees of freedom
## Residual deviance: 6751.1  on 6916  degrees of freedom
## AIC: 6779.1
##
## Number of Fisher Scoring iterations: 4
```

Como podemos comprobar a simple vista, todas las variables son estadísticamente significante excepto *"age"* y *"cause"*, cuyo p-valor es mayor a 0.05.

A continuación, utilizamos el algoritmo de regresión paso a paso:

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## outcome ~ sex + age + cause + ec + eye + motor + verbal + pupils +
##     phm + sah + oblt + mdls + hmt
##
## Final Model:
## outcome ~ age + ec + eye + motor + verbal + pupils + phm + sah +
##     oblt + mdls + hmt
##
##
##      Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                           6916   6751.097 6779.097
## 2   - sex  1 0.2704474      6917   6751.368 6777.368
## 3 - cause  1 0.4047151      6918   6751.772 6775.772
```

Una vez más podemos comprobar que las variables de *"cause"* y *"sex"* son las que se descartan usando este algoritmo.

3.2.4.3 Análisis de PCA

En primer lugar, antes de proceder con el análisis de componentes principales, vamos a tener en cuenta la matriz de correlaciones, puesto que un PCA tiene sentido si existen altas correlaciones entre las variables, ya que como se ha comentado con anterioridad, esto es indicativo de que existe información redundante y, por tanto, pocos factores explicaran gran parte de la variabilidad total.

Como ya vimos con las matrices de correlaciones solo obtuvimos correlaciones medianamente fuertes entre las variables de *"motor"*, *"eye"* y *"verbal"*, pero la correlación no era significativa por lo que no se descartó ninguna variable.

Un problema en el análisis de datos multivariante es la reducción de la dimensionalidad: es decir, si se puede conseguir con precisión los valores de las variables (p) con un pequeño subconjunto de ellas (r<p), habremos conseguido reducir la dimensión a costa de una pequeña perdida de información.

El análisis de componentes principales tiene este objetivo. Dada n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un conjunto menor de variables (construidas como combinaciones lineales de las originales).

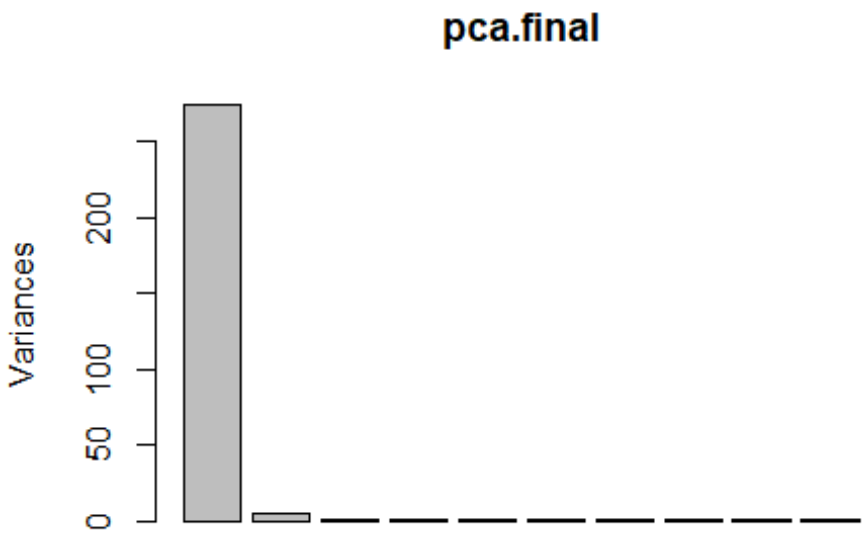
El primer paso en el análisis de componentes principales consiste en la obtención de los valores y vectores propios de la matriz de covarianzas muestral o de la matriz de coeficientes de correlación que se obtienen a partir de los datos.

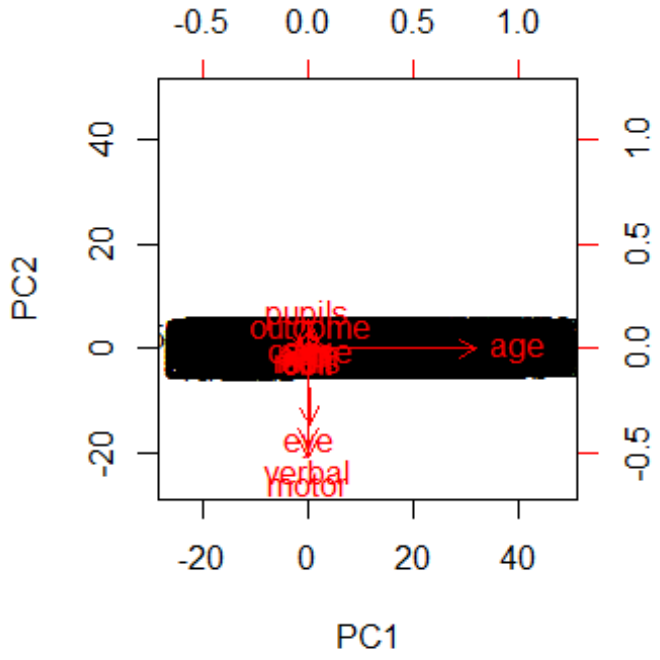
Debemos saber que el análisis de componentes principales utiliza la versión normalizada de los predictores originales. Estas variables pueden encontrarse en distintas escalas (kilómetros, litros, euros, etc.) y por lo tanto, las varianzas también tendrán varias escalas.

Realizar el PCA con variables no normalizadas dará lugar a que haya cargas bastante grandes para variables con una varianza alta y a su vez, esto llevará a la dependencia de una componente principal con la variable con la varianza más alta. Esto no es deseable. Por lo que se llevara a cabo una normalización de las variables. Al normalizar las variables, la distribución de la variabilidad entre las componentes parece más racional.

Veamos qué ocurre si utilizamos la **matriz de covarianza**, sin haber normalizado las variables:

```
## Importance of components:
##                             PC1    PC2     PC3     PC4     PC5    PC6
## Standard deviation     16.5659 2.0860 0.96050 0.76858 0.68086 0.6514
## Proportion of Variance  0.9713 0.0154 0.00327 0.00209 0.00164 0.0015
## Cumulative Proportion   0.9713 0.9867 0.98999 0.99208 0.99373 0.9952
##                            PC7    PC8     PC9    PC10    PC11   PC12
## Standard deviation     0.51378 0.4441 0.42801 0.41667 0.40227 0.3766
## Proportion of Variance 0.00093 0.0007 0.00065 0.00061 0.00057 0.0005
## Cumulative Proportion  0.99616 0.9969 0.99751 0.99812 0.99870 0.9992
##                           PC13    PC14
## Standard deviation     0.37091 0.29861
## Proportion of Variance 0.00049 0.00032
## Cumulative Proportion  0.99968 1.00000
```

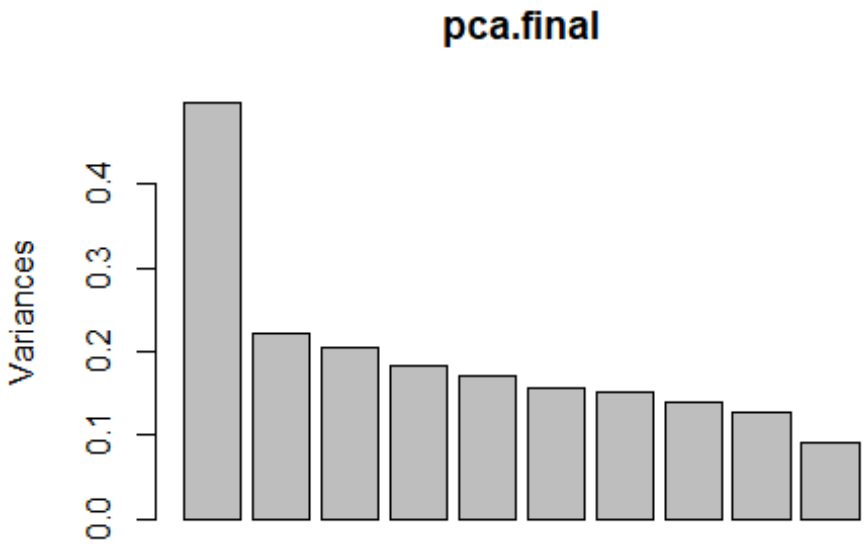

pca.final
Variances
0
50
100
200

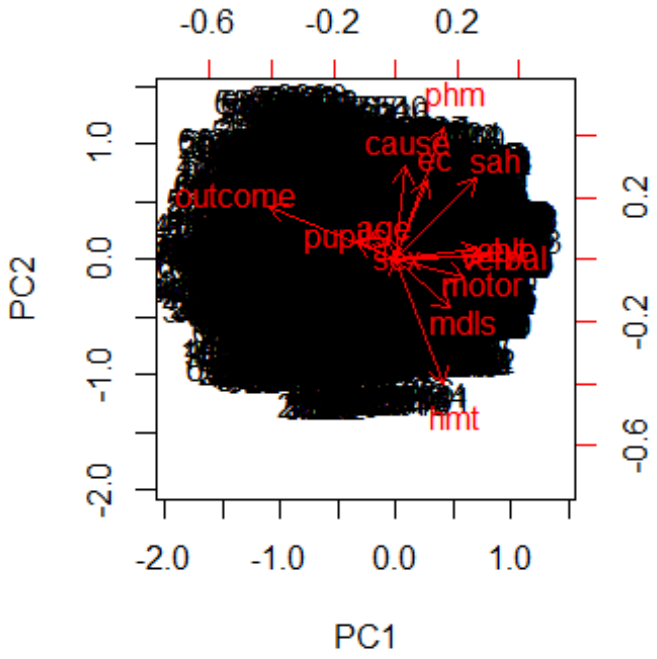

-0.5
0.0
0.5
1.0
PC2
40
20
0
-20
pupils
outcome
age
eye
verbal
motor
-20
0
20
40
PC1

Como se puede comprobar en la gráfica anterior, al no haber escalado las variables, la primera componente principal (PC1) está dominada por la variable *"age"*, mientras que la segunda componente principal está dominada por las variables: *"eye"*, *"motor"* y *"verbal"*.

Ahora vamos a utilizar la **matriz de covarianza**, habiendo normalizado todas las variables.

```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     0.7051 0.4706 0.45354 0.42754 0.41448 0.39605
## Proportion of Variance 0.2351 0.1047 0.09728 0.08644 0.08124 0.07418
## Cumulative Proportion  0.2351 0.3398 0.43709 0.52353 0.60477 0.67895
##                            PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.39038 0.37255 0.35681 0.29999 0.23364 0.20575
## Proportion of Variance 0.07207 0.06564 0.06021 0.04256 0.02582 0.02002
## Cumulative Proportion  0.75101 0.81665 0.87686 0.91941 0.94523 0.96525
##                           PC13    PC14
## Standard deviation     0.19667 0.18658
## Proportion of Variance 0.01829 0.01646
## Cumulative Proportion  0.98354 1.00000
```





Como se puede comprobar en la gráfica anterior, al normalizar las variables, vemos que el peso de estas se distribuye de forma más uniforme entre las 2 componentes principales.

Para elegir nuestras componentes principales, podremos utilizar dos métodos:

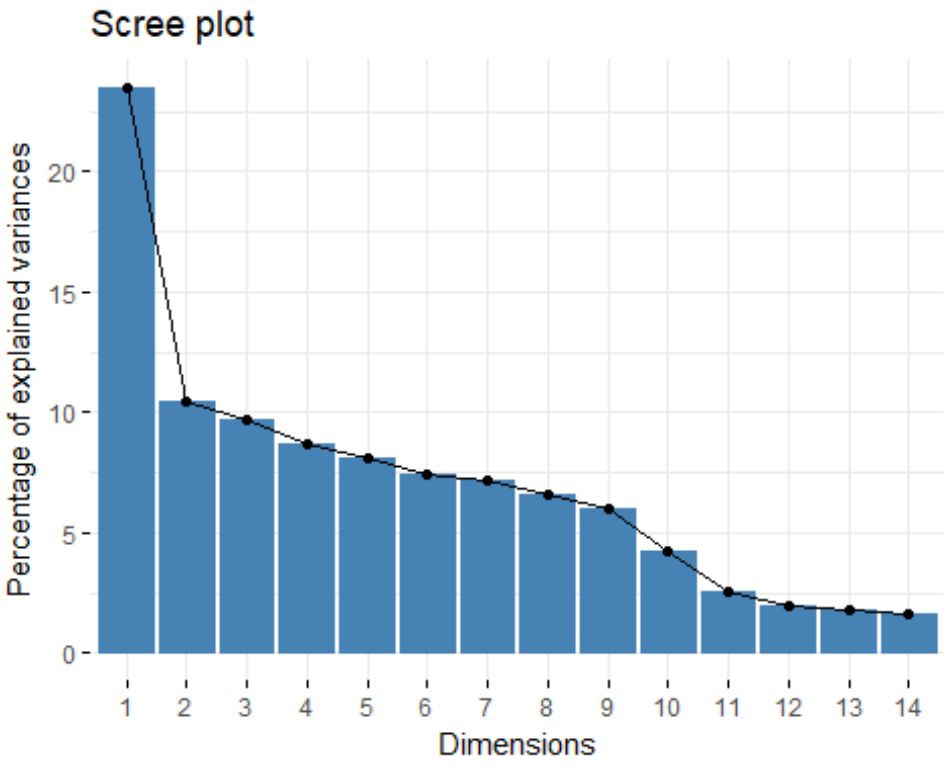
- Por un lado, podemos utilizar el **criterio de Kaiser**, que consiste en conservar aquellos factores cuya desviación estándar al cuadrado asociada sea mayor que 1.

```
##  [1] 0.49713279 0.22143722 0.20570161 0.18278879 0.17178962 0.1568
5861
##  [7] 0.15239376 0.13879386 0.12731397 0.08999118 0.05458915 0.0423
3462
## [13] 0.03867736 0.03481073
```

  Como se puede comprobar, utilizando este criterio, podríamos quedarnos con los componentes PC1,PC2,PC3,PC4 y PC5.

- Otra forma para saber cuántos componentes tener en cuenta es mantener el número de componentes necesarios para explicar al menos un porcentaje del total de la varianza. Por ejemplo, es importante **explicar al menos un 80%** de la varianza.

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.49713279        23.509395                    23.50940
## Dim.2  0.22143722        10.471760                    33.98115
## Dim.3  0.20570161         9.727623                    43.70878
## Dim.4  0.18278879         8.644076                    52.35285
## Dim.5  0.17178962         8.123926                    60.47678
## Dim.6  0.15685861         7.417839                    67.89462
## Dim.7  0.15239376         7.206697                    75.10132
## Dim.8  0.13879386         6.563557                    81.66487
## Dim.9  0.12731397         6.020674                    87.68555
## Dim.10 0.08999118         4.255680                    91.94123
## Dim.11 0.05458915         2.581520                    94.52275
## Dim.12 0.04233462         2.002003                    96.52475
## Dim.13 0.03867736         1.829051                    98.35380
## Dim.14 0.03481073         1.646198                   100.00000
```



Según este criterio, deberíamos quedarnos con los primeros componentes principales: PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8 y PC9.

A continuación, podremos ver la carga de cada variable respecto a las componentes principales.

```
##                PC1          PC2         PC3         PC4         PC5
## sex      0.00754185  0.004825702 -0.05061100  0.09673923 -0.08169835
## age     -0.03305719  0.089077063  0.02955650 -0.01094135 -0.09084363
## cause    0.03634525  0.376804210  0.18316340 -0.09565336  0.12960828
## ec       0.12719737  0.321982271  0.52579134 -0.25862893  0.53735713
## eye      0.37584853  0.019196020  0.23805982  0.11908233 -0.21180882
## motor    0.27860212 -0.072483884  0.17979746  0.09522108 -0.10943985
## verbal   0.35002340  0.010691932  0.18829500  0.08182141 -0.15454678
## pupils  -0.15858549  0.076136117 -0.06963948 -0.02843065  0.05073360
## phm      0.19546358  0.532494470 -0.37457091  0.67612986  0.14592864
## sah      0.32519856  0.328128704 -0.48960240 -0.55085227  0.09217214
## oblt     0.34248291  0.040645898 -0.19093528 -0.28120171 -0.30335822
## mdls     0.22183247 -0.197257325 -0.10251020 -0.10980630 -0.19521708
## hmt      0.19633261 -0.505774550 -0.31239966  0.05260687  0.65098174
## outcome -0.51555059  0.217405899 -0.18078512 -0.17100730 -0.11188899
##                PC6          PC7         PC8          PC9         PC10
## sex      0.468631120 -0.545637298  0.56848224 -0.362100975  0.05449783
## age      0.213465260 -0.004555270 -0.05438566  0.015539408 -0.06192319
## cause    0.266773385  0.351493551 -0.30349198 -0.693069485  0.13967486
## ec       0.075208510 -0.005328975  0.31834880  0.361266429  0.09920550
## eye      0.216580398 -0.199155308 -0.32508940  0.111568280 -0.05697973
## motor    0.059764349 -0.130958699 -0.16923921  0.085139555  0.03200627
## verbal   0.158055809 -0.182516720 -0.24157401  0.110972727 -0.03748816
## pupils   0.012761929  0.016462188  0.05708582 -0.041706107 -0.07740804
## phm     -0.007730786  0.116800019  0.08439404  0.173756733  0.08844781
## sah     -0.230815829 -0.386705035 -0.15835007 -0.053806287  0.02958798
## oblt     0.305400757  0.515894927  0.33718561  0.134660703 -0.41366626
## mdls     0.047519909  0.225944825  0.14468847  0.082413000  0.86550643
## hmt      0.374508857  0.059210871 -0.18071620  0.002877392 -0.06071833
## outcome  0.541525055 -0.066325752 -0.29153155  0.407445568  0.13989249
##                PC11         PC12         PC13         PC14
## sex     -0.046320481  0.051178890 -0.043546584  0.008349846
## age      0.064650181 -0.794933844  0.526023857 -0.127687713
## cause   -0.058199593  0.082271746 -0.018041653  0.006485046
## ec       0.007524695 -0.018597326 -0.011587465  0.011292569
## eye      0.283758366 -0.264686334 -0.509573504  0.352793604
## motor   -0.444932122  0.256671073  0.528239380  0.513066463
## verbal   0.222888103  0.350653837  0.224483179 -0.676707194
## pupils   0.791456558  0.250663510  0.353092549  0.368086924
## phm     -0.024586186 -0.001250305  0.003211589  0.006193667
## sah     -0.020857062 -0.040982217  0.022648597  0.025366341
## oblt    -0.045305471  0.083331468 -0.030150696  0.035015405
## mdls     0.134062896 -0.046673682  0.029152475  0.008125267
## hmt      0.024565131 -0.026644622  0.008717377  0.001351517
## outcome -0.118307105  0.158441132 -0.069394421  0.018965211
```

Como conclusiones teniendo en cuenta el PCA y las matrices de correlaciones, no se puede descartar ninguna variable por los siguientes motivos:

- Las correlaciones entre las variables *"eye"*, *"motor"* y *"verbal"* no son lo suficientemente fuertes como para considerar que existe información redundante. El resto de pares de variables tienen una correlación poco significativa.
- Los criterios utilizados para elegir las componentes principales nos han indicado que se necesitan al menos 5 componentes principales usando el criterio de Kaiser y 9 utilizando el criterio del 80% de la proporción de la varianza. Teniendo en cuenta que poseemos 14 variables, la reducción no es significativa y se perdería interpretabilidad.

4. RESULTADOS

Sección pendiente de desarrollo

5. CONCLUSIONES

Sección pendiente de desarrollo

6. LÍNEAS FUTURAS

Sección pendiente de desarrollo

7. BIBLIOGRAFÍA

[Manual abreviado de Análisis Estadístico Multivariante. Jesús Montanero Fernández] http://matematicas.unex.es Recuperado el 28 de marzo de 2018 de: https://ignsl.es/historia-del-big-data/

[Análisis Multivariente, usando R. José Carlos Vega Vilca] http://cicia.uprrp.edu Recuperado el 28 de marzo de 2018 de: http://cicia.uprrp.edu/publicaciones/Papers/ManualESTA5503.pdf