Estudio de Ratones de Laboratorio con Síndrome de Down y las Variables Moleculares que Influyen en su Aprendizaje.

Isabelle Archer, Nadal Bardisa Quintero, David Gilsanz Domínguez, Haoxiang Liu, Chenyao Lin

2025-06-08

Grupo: A2-20

Introducción

Este trabajo aborda el análisis, entendimiento y desarrollo de conclusiones sobre los datos presentes en el fichero *Data_Cortex_Nuclear.xlsx*. El objetivo del estudio es encontar y modelar las relaciones entre variables moleculares y experimentales en ratones de laboratorio para identificar patrones asociados a su genotipo, tratamiento y comportamiento. Con este estudio se puede identificar las proteínas criticales al aprendizaje de un ratón con síndrome de Down.

Se eligió la base de datos siguiendo algunas condiciones: se recomendaba tener 200 observaciones hasta 3000 observaciones, al menos 20 variables numéricas (discretas o continuas) o categóricas ordinales, y al menos una variable categórica nominal. La base de datos elegida se trata de ratones con síndrome de Down y el efecto de cierto psicofármaco en el desarrollo de su capacidad de aprendizaje. Hay 15 observaciones por ratón, contando con 72 sujetos distintos con representación equitativa en cuanto a recepción de fármaco o placebo y a su vez la presencia o falta del genotipo, para un total de 1080 observaciones. Contiene 82 variables en total: 77 variables numéricas continuas con la medida de las proteínas en los cerebros de los ratones, 3 variables binarias que representan el genotipo, el tratamiento, y el comportamiento, 1 variable categórica nominal que representa el conjunto de las variables binarias, y finalmente una variable identificador del ratón y el número de repetición de medida.

Análisis Exploratorio

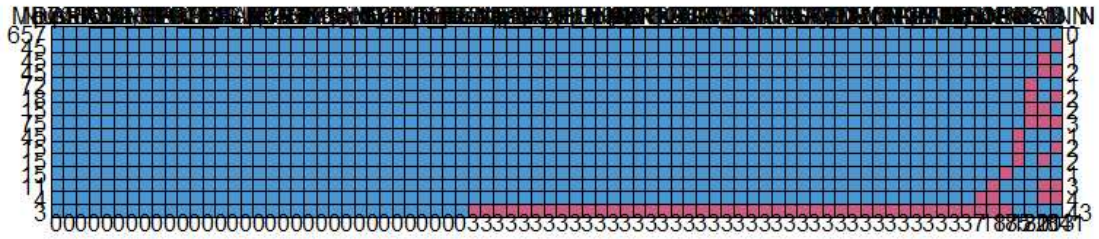
Se ha realizado una exploración de los datos a nivel visual y se detectaron y modificaron las variables que requieran una recodificación. Así pues se renombraron las variables *Genotype*, *Treatment*, *Behavior* y *Class* a *Genotipo*, *Tratamiento*, *Comportamiento* y *Clase*. Después, se transformaron las variables *Genotipo*, *Tratamiento* y *Comportamiento* a su formato binario numérico (convirtiendo los valores de estas variables a un formato

entendible para las herramientas que se utilizan), indicando con valor 0 el valor de *Genotipo* “control” y 1 el valor “Ts65Dn”, con 0 el valor de *Tratamiento* “Saline” y con 1 el valor “Memantine” y con 0 el valor de *Comportamiento* “S/C” y con 1 el valor “C/S”.

Una vez realizadas las recodificaciones necesarias, se resumieron estadísticamente las variables numéricas, excluyendo las variables categóricas renombradas y recodificadas previamente. Se calculó la desviación típica, la media y el coeficiente de variación de cada variable para tener una primera idea de su dispersión relativa. Esto permitio detectar posibles inconsistencias o valores con alta variabilidad.

Se analizó la presencia de valores faltantes en cada variable. Se calculó tanto el número absoluto como el porcentaje de valores ausentes por columna, y se construyó una tabla resumen. A partir de ella, se eliminaron del conjunto de datos aquellas variables que presentan un porcentaje igual o superior al 20% de valores faltantes. Se recalcularon las métricas para comprobar el estado actualizado de los datos. Posteriormente, se realizó el mismo análisis pero a nivel de fila, identificando los individuos con un porcentaje elevado de datos faltantes. Se mostró un resumen estadístico del número de valores ausentes por individuo, así como un gráfico de barras con la distribución porcentual de los casos. Aunque se identificaron las filas con más de un 20% de datos ausentes, en este bloque no se eliminaron.

Para estudiar la estructura de los valores ausentes, se utilizó la función **md.pattern** del paquete **mice**, y se extrajeron únicamente las columnas con valores faltantes. Sobre estas variables, se aplicó un proceso de imputación múltiple utilizando el método por defecto del paquete, generando cinco datasets imputados a partir de una semilla fija para garantizar la reproducibilidad. Se extrajeron gráficos tipo stripplot para evaluar visualmente los resultados de la imputación en distintas variables seleccionadas, y se completaron los valores faltantes del conjunto original con los imputados de la segunda iteración. Finalmente, se compararon visualmente los valores antes y después de la imputación mediante diagramas de caja, y el nuevo conjunto de datos limpio se guardó en un archivo Excel para su uso posterior. Para ver los resultados de esta parte, ver el **anexo 1** Análisis exploratorio.



Al final del pre-proceso y la limpieza de la base de datos, se obtuvo una base de datos sin valores faltantes y con las variables recodificadas adecuadamente para el análisis posterior.

Análisis PCA

En este apartado se va realizar un PCA sobre los datos para reducir la gran cantidad de variables sobre proteínas que hay e intentar relacionar las dimensiones con las variables de proteínas. Además, ver como afectan el tipo de los ratones y comparar las clases de ratones. Se abre el fichero creado en el apartado anterior para trabajar sobre la base de datos limpia.

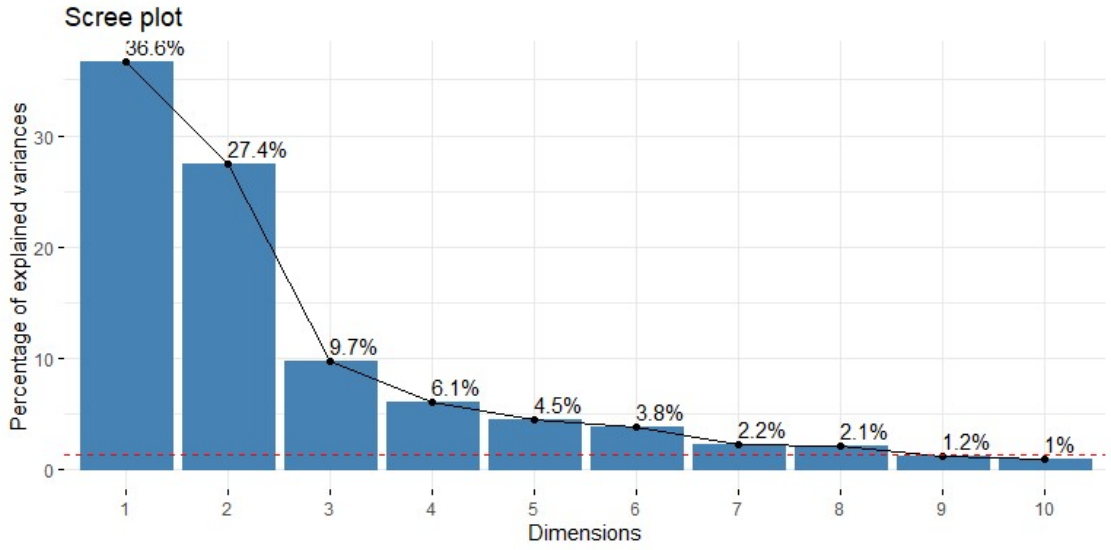
Se crea una tabla descripción que contenga los tipos de todas las variables para poder elegir variables por su tipo en el análisis. Se muestra las tres primeras filas.

Se centran los datos numéricos de las proteínas con la función **scale** para poder trabajar con ellos.

Para hacer el PCA habían dudas de si hacerlo escalando los datos o no. Se hizo el análisis paralelamente con los dos PCA y salían todos los resultados muy similares. Al final se centró el análisis en el PCA sin escalar ya que todos los valores de expresiones de proteínas parecen haberse medido en la misma unidad y no hay que restarle importancia a las proteínas con valores más altos. En el **anexo 2** se puede ver los resultados del PCA con escalado, pero a continuación todos los resultados son del PCA sin escalar los datos.

Para crear el PCA primero se creó uno para graficar cuál es el número óptimo de dimensiones. Se crea el PCA auxiliar con la función **PCA** y hace el scree plot con la función **fviz_eig**. Según este gráfico decidió coger 4 dimensiones para el PCA basando en el criterio del codo, entonces se crea el nuevo PCA con número de compontentes 4. Además, en el PCA se añaden como variables suplementarias las variables categóricas *mouseID*, *Clase* y las tres variables binarias que indican el tipo de ratón (*Genotipo*,

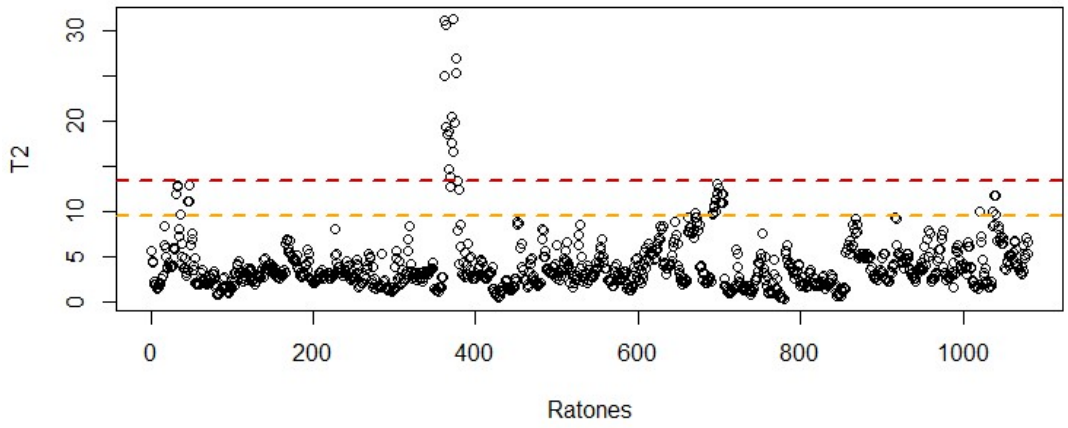
Comportamiento y *Tratamiento*). Para que sea un PCA sin escalado se específica la variable **scale.unit** a FALSE.



	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.2448417	36.626152	36.62615
Dim.2	1.6810050	27.426764	64.05292
Dim.3	0.5967225	9.735941	73.78886
Dim.4	0.3742295	6.105815	79.89467
Dim.5	0.2748475	4.484328	84.37900
Dim.6	0.2332461	3.805573	88.18457

Anómalos

Se hace un gráfico T2 de Hotelling para visualizar los anómalos. Se guarda en una variable el valor de la T2 y se calcula también los límites F95 y F99. Se crea un gráfico de la T2 con plot y se incluyen dos lineas que hacen referencia a los límites F95 y F99 con la función abline.



Para identificar a los anómalos que pueden entorpecer el análisis se establece el criterio de borrar todos aquellos que sobrepasan por 2 el límite F95. Se muestran las observaciones que sus T2 sobrepasan por dos el valor de F95 y se observa que la mayoría de estas se centran en una zona que va desde la observación 361 a la 376 aproximadamente. Tras revisar las observaciones se ve que eran todas observaciones de un mismo ratón, así que posteriormente se decide borrarlas ya que parecía un error de medición aislado.

PCA sin anómalos

Ahora se procede a quitar todas las observaciones anómalas (observaciones cuyo T2 sobrepasen el umbral F95*2) y a crear el PCA sin escalado habiendo quitado ya los datos anómalos. Para su creación se vuelve a hacer primero el PCA auxiliar para ver si el scree plot cambia de alguna forma. Viendo que no ha cambiado demasiado se vuelven a coger cuatro dimensiones y se crea el PCA sin escalado definitivo. Agregando otra vez las variables categóricas y binarias como suplementarias.

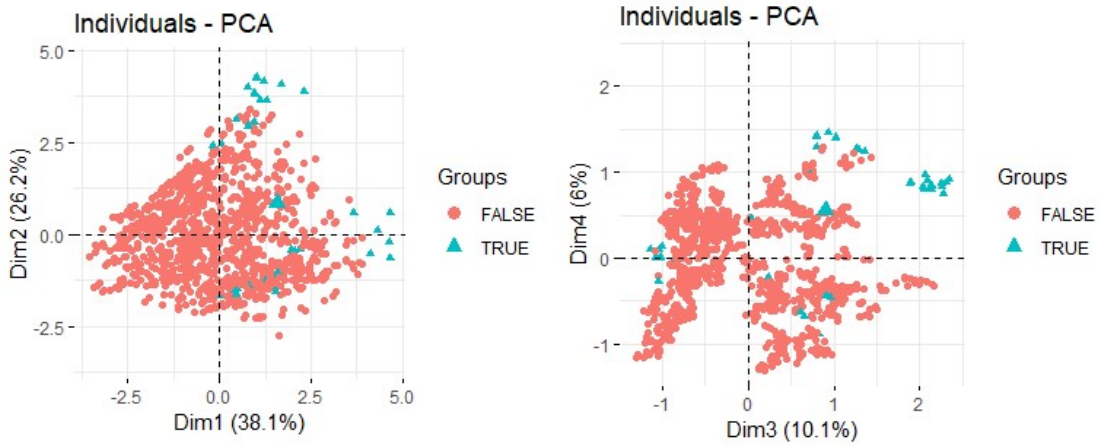
Anómalos sin T2 > F95*2

Para ver el cambio en las observaciones anómalas se vuelve a graficar la T2 de Hotelling recalculando toda la T2 y los límites con el PCA definitivo. Se observa que ya no hay observaciones anómalas que puedan sesgar el modelo bajo el criterio establecido. Solo hay un par que sobrepasan un poco el límite F99, pero no parecen lo suficientemente graves como para borrarlas.

Gráficos individuos

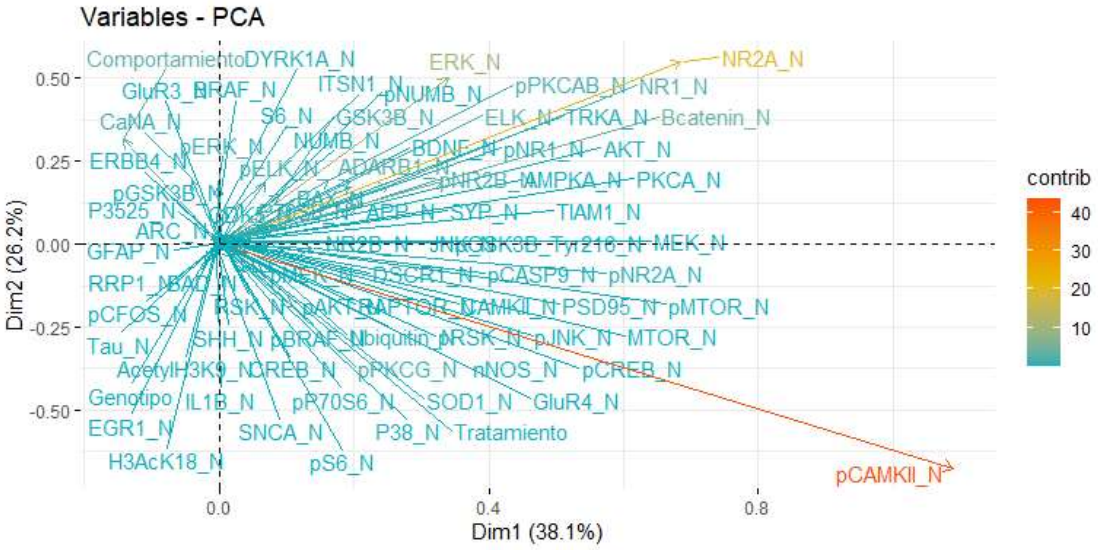
Para asegurarse de que el PCA no está sesgado por algún dato anómalo se decide hacer gráficos de individuos separando el grupo de observaciones anómalas incluyendo

factor(T2 > F95) en habillage. De esta forma se verán dos grupos diferenciados de individuos, uno de observaciones anómalas y otro de observaciones típicas.



En el gráfico se puede diferenciar entre los valores normales (FALSE) y los atípicos (TRUE). Se observa que justamente los ratones con valores anómalos (T2 mayor que F95) se van hacia los extremos sobre todo en las dimensiones 1 y 2, pero no excesivamente como para sesgar el modelo completamente por lo que se conservan estos valores anómalos moderados.

Una vez ya se ha hecho el PCA definitivo habiendo tratado los datos anómalos se procede a su análisis mediante gráficos. En este apartado se verán las variables que más contribuyen en las dimensiones del PCA con el gráfico de variables usando la función **fviz_pca_var** y coloreando por las variables por su contribución a las dimensiones. Se crean dos gráficos de variables, uno para las dimensiones 1 y 2 (axes = c(1, 2)) y otro para las dimensiones 3 y 4 (axes = c(3, 4))



En los gráficos de variables se observa que la mayoría de variables tienen poca contribución, pero hay 2 o 3 que tienen mucha. Destacando la variable *pCAMKII_N* con las dimensiones 1 y 2 y la variable *pKCG_N* con las dimensiones 3 y 4. También se puede ver que la variable categórica o binaria con más contribución es *Comportamiento* en la dimensión 4. En el **anexo 3** se puede ver los gráficos de individuos por cada variable binaria.

En conclusión, las proteínas *pCAMKII_N*, *pKCG_N* y la variable *Comportamiento* son las variables importantes de la base de datos para distinguir los ratones.

Análisis Clustering

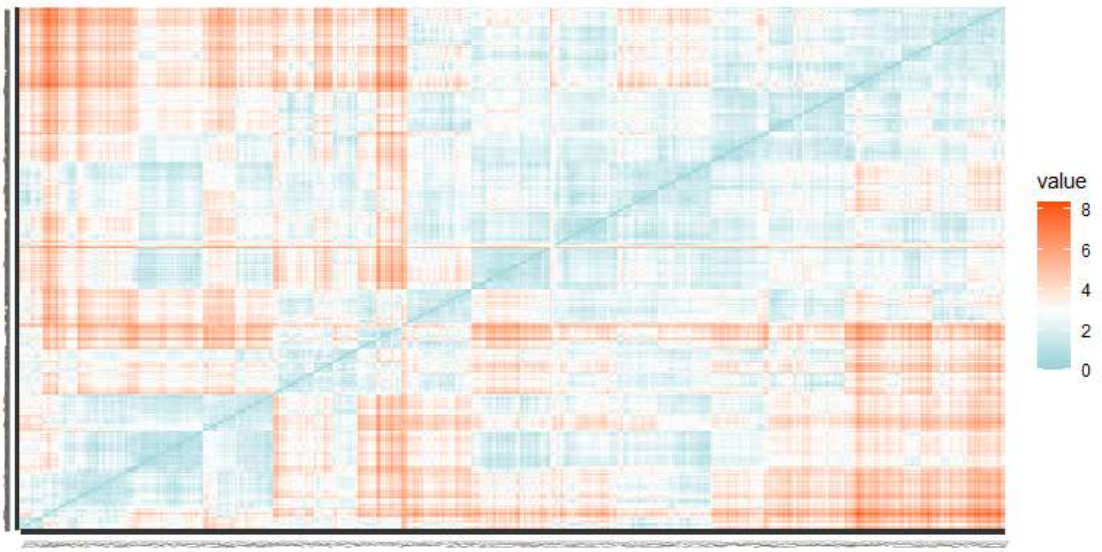
El objetivo es realizar un análisis de agrupamiento de ratones con niveles de expresión de proteínas similares.

Selección de variables a utilizar y preparación de datos

Para ello, se han seleccionado las variables de proteínas para realizar el clustering, que son todas las variables numéricas.

Medida de distancia

Se utilizará la distancia euclídea como medida de distancia, porque se desea agrupar ratones con valores de expresión de proteínas similares y no con perfiles similares de proteínas. No se ha realizado ni un centrado ni un escalado de los datos porque las variables están en las mismas unidades.



A lo largo de la diagonal del mapa de calor se identifican bloques definidos de color azul, lo que indica la presencia de grupos de observaciones cercanas. Además, la separación entre estos bloques mediante zonas de color rojo refuerza la hipótesis de que los grupos están bien diferenciados. Los bloques mejor definidos son: un grupo grande con aproximadamente la mitad de las observaciones y dos grupos más pequeños cada uno con aproximadamente un cuarto de las observaciones.

Tendencia de agrupamiento

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.8965  0.9050  0.9097  0.9096  0.9141  0.9224
```

El valor del estadístico de Hopkins está alrededor del 0.9, lo que indica que los datos tienen una tendencia a agruparse al estar cerca del 1.

Habiendo observado el mapa de calor de distancias y el estadístico de Hopkins, se puede suponer que hay clústers. Por lo tanto, se procedió a probar varios métodos de clustering para encontrar el más adecuado. Se combinó el análisis del coeficiente de Silhouette (maximizando) con la variabilidad intra-cluster (minimizando) para elegir el número de grupos con el que se realizará el clustering. Para ver todos los métodos probados, se

puede ver el **anexo 4** para los métodos jerárquicos y el **anexo 5** para los métodos de partición.

Selección del método de clustering

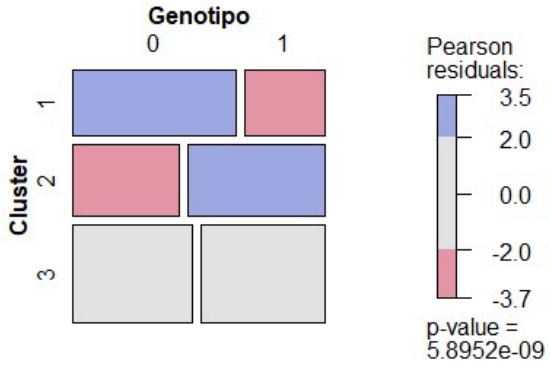
Observando los coeficientes de Silhouette de los resultados del **anexo 5**, se elige el método de k-means con 3 clústers ya es el que tiene mayor Silhouette media y menor cantidad de observaciones con coeficientes negativos, es decir, mal clasificados.

Estudio de los clústers

Se añadirá la variable "cluster" a los datos utilizados para estudiar los clústers

Se van a realizar gráficos de mosaico para analizar como se reparten las observaciones en las combinaciones de las tres variables binarias (*Genotipo*, *Tratamiento* y *Comportamiento*), también de la variable *Clase*, respecto de los tres clústers. El tamaño de cada bloque es proporcional a la frecuencia observada y su color refleja el grado de desviación respecto a lo que cabría esperar si ambas variables fueran independientes. Los residuos de Pearson, codificados en una escala de colores, indican la dirección y la intensidad de esa desviación: los tonos azules señalan un exceso de casos frente al modelo de independencia, mientras que los rojos indican un déficit. Cuanto más intenso es el color, mayor es la discrepancia. Por último, el p-valor de la prueba de bondad de ajuste resume en un solo indicador la solidez de esta asociación global; si el valor es pequeño (inferior a 0.05) se descarta la hipótesis nula de independencia y se confirma la existencia de una asociación estadísticamente significativa entre ambas variables.

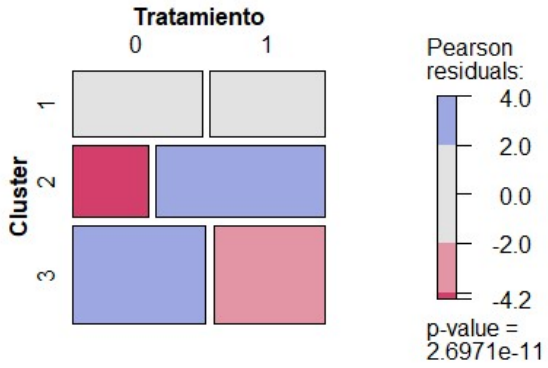
Análisis con Genotipo



Dado que el valor p de la prueba de bondad de ajuste es menor a 0.05, se concluye que existe una relación significativa entre el *Cluster* y el *Genotipo*, por lo que se rechaza la hipótesis de independencia. Los clústers 1 y 2 no presentan desviaciones apreciables en la proporción de ratones trisómicos y ratones neurotípicos ya que para ambos genotipos los

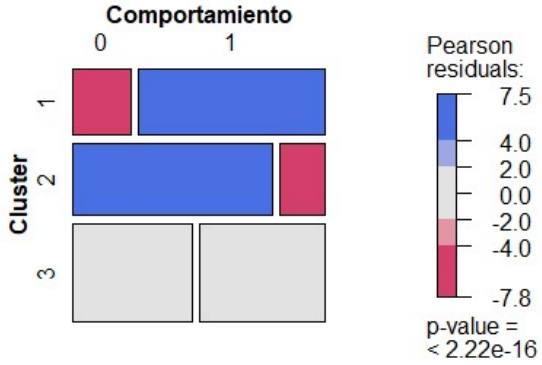
residuos de Pearson se mantienen cercanos a cero. Por otra parte, en el clúster 3 el residuo de Pearson para ratones neurotípicos (Genotipo = 0) se sitúa entre +2 y +3, lo que refleja que hay ligeramente más ratones neurotípicos que con trisomía.

Análisis con Tratamiento



El p-valor de la prueba de bondad de ajuste resulta ser inferior a 0.05, por lo que se rechaza la hipótesis de independencia entre *Cluster* y *Tratamiento*. En el clúster 1 se observa un residuo de Pearson entre +2 y +4 para los ratones que no tomaron la medicina (Tratamiento = 0), lo que indica una mayor presencia de estos ratones en comparación con los que sí que recibieron el tratamiento. En el clúster 2 ocurre lo opuesto, el residuo para ratones no tratados es de aproximadamente –4, señalando que hay déficit notable de ratones sin tratar. Por último, en el clúster 3 ambos bloques tienen residuos cercanos a 0, lo que sugiere que la proporción de ratones tratados y no tratados en ese grupo se ajusta a lo previsto bajo independencia.

Análisis con Comportamiento



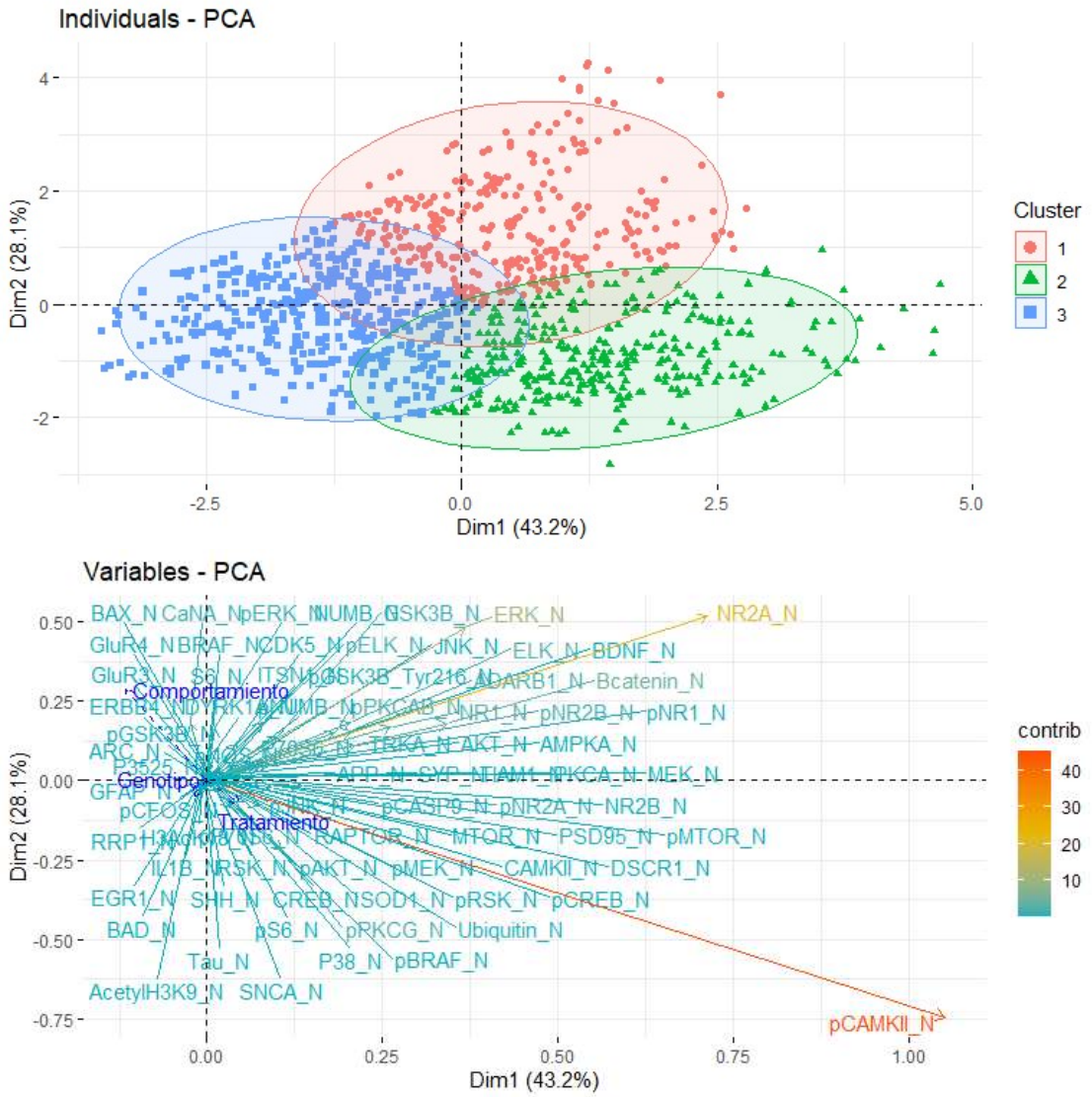
Como el valor p de la prueba de bondad de ajuste es menor a 0.05, se descarta la hipótesis de que *Cluster* y *Comportamiento* sean independientes. En el clúster 1, tanto los ratones estimulados como los no estimulados presentan frecuencias observadas muy cercanas a las esperadas, lo que sugiere una distribución coherente con el supuesto de independencia. En el clúster 2, el bloque correspondiente a ratones sin estimulación (Comportamiento = 0) presenta un residuo de Pearson superior a +4, señalando un exceso muy pronunciado de ratones con ese comportamiento en comparación con lo esperado. Por último, en el clúster 3 sucede lo inverso. El bloque de ratones con estimulación (Comportamiento = 1) arroja un residuo por encima de +4, mostrando un exceso muy significativo de ratones estimulados a aprender.

En el **anexo 6**, se ha realizado el mismo proceso con la variable *Clase* y los resultados son los mismos. En conclusión, se pueden caraterizar los clústers: -El clúster 1 es un grupo bastante homogéneo pero destacan los ratones que no han tomado medicina, sin importar tanto ni el genotipo ni el comportamiento. -El clúster 2 está formado mayoritariamente por ratones que no han recibido estimulación para aprender, especialmente los que han tomado la medicina, sin importar el genotipo. -El clúster 3 destacan los ratones que han recibido estimulación para aprender, especificamente los que son neurotípicos.

Análisis con proteínas

Se utilizará el PCA estudiado anteriormente, pero añadiéndole como variable suplementaria “Cluster”, para estudiar los clústers sobre las expresiones de las proteínas. Se generará la tabla auxiliar de variables incluyendo su tipo.

Se generarán los scores plots para observar los clústers y también los gráficos de variables para las proteínas. Para sacar las conclusiones se compararán los dos tipos de gráficos.



Se observa que el clúster 2 se proyecta en la misma dirección que la proteína *pCAMKII_N* y el clúster 3 se proyecta en la misma dirección que las proteínas *NR2A_N* y *ERK_N*. Por otra parte, el clúster 1 se encuentra justo en la dirección contraria de los otros dos clústers y de las tres proteínas más significativas, principalmente en dirección contrario del clúster 3 y de las proteínas *NR2A_N* y *ERK_N*. El clúster 1 destaca por ratones sin medicina y está en la dirección contraria a las proteínas significativas, por lo que se puede deducir que estas proteínas se expresan debido a la medicina administrada. Además, el clúster 3 es el de los ratones neurotípicos y con estimulación, por lo que las proteínas *NR2A_N* y *ERK_N* podrían estar relacionadas con la estimulación, además de con la medicina. Por otra parte, el clúster 2 es el de los ratones sin estimulación y con medicina, por lo que se podría deducir que la proteína *pCAMKII_N* está relacionada con la falta de estimulación a parte de con la medicina.

No se observa una separación clara de los clústers ni en la componente 3 ni en la 4, por lo que no se pueden sacar conclusiones.

En conclusión, las proteínas *NR2A_N* y *ERK_N* están relacionadas con la medicina y la estimulación, mientras que la proteína *pCAMKII_N* está relacionada con la medicina y la falta de estimulación.

Análisis PLS-DA

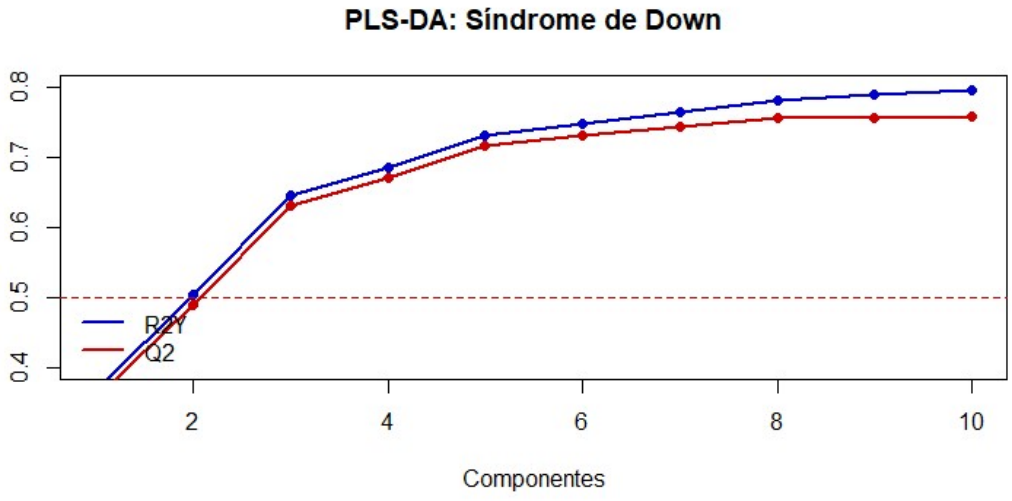
```
## Loading required package: viridisLite

## Warning: package 'caret' was built under R version 4.4.3

## Loading required package: lattice
```

Se ha aplicado un modelo PLS-DA (Partial Least Squares Discriminant Analysis) para clasificar a los ratones según su genotipo (síndrome de Down o Control) a partir de variables de expresión. Esta técnica permite detectar estructuras latentes que explican la separación entre grupos y evaluar su capacidad predictiva mediante validación cruzada.

Los resultados del modelo (ver Anexo) muestran que se explica el 80% de la variabilidad de la variable Genotipo (R2Y = 0.80) y que la capacidad predictiva alcanza un valor elevado (Q2 = 0.76). El error medio de estimación es bajo (RMSEE = 0.23), y el test de permutación indica que el modelo es estadísticamente significativo (pR2Y = 0.05, pQ2 = 0.05). El modelo estima que el número óptimo de componentes es en un principio 8, aunque luego se corroborará con los gráficos.

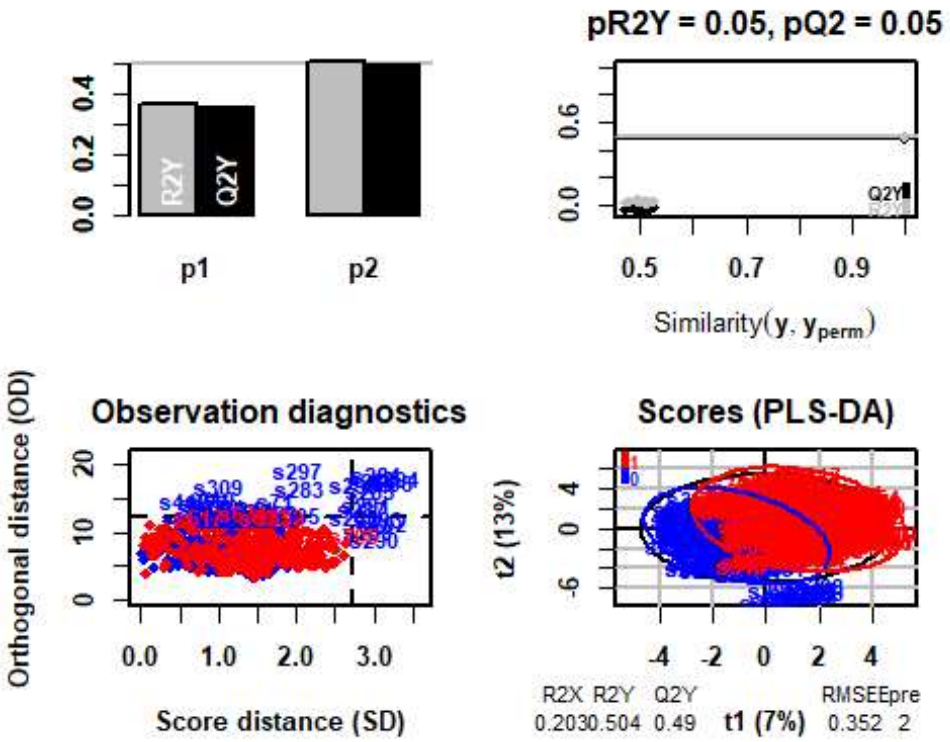
Es por ellos que el siguiente gráfico, que representa la evolución de R2Y y Q2 en función del número de componentes, se observa que ambos valores aumentan hasta estabilizarse en el componente 10, lo que justifica su selección como número óptimo:

```
## PLS-DA
## 864 samples x 75 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total    0.776    0.795   0.758 0.228  10   0 0.05 0.05
```



Para facilitar la interpretación visual, se ajustó de nuevo el modelo utilizando únicamente dos componentes. Esta decisión se basa en que dos componentes son suficientes para representar los datos en un espacio bidimensional, lo cual permite analizar gráficamente la separación entre grupos. A continuación, se muestra el código que genera los gráficos de diagnóstico automáticos:

```
## PLS-DA
## 864 samples x 75 variables and 1 response
## standard scaling of predictors and response(s)
##       R2X(cum) R2Y(cum) Q2(cum) RMSEE pre ort pR2Y  pQ2
## Total    0.203    0.504    0.49 0.352   2   0 0.05 0.05
```



En esta sección se presentan los resultados del modelo PLS-DA ajustado con únicamente dos componentes, con el objetivo de facilitar la interpretación visual de la estructura de los datos. Aunque el modelo completo estima como óptimos más componentes, se ha decidido representar gráficamente solo los dos primeros para proyectar los individuos en un espacio bidimensional que permita observar de forma clara la separación entre los grupos Down y Control.

Como consecuencia del uso de únicamente dos componentes, los valores de R2Y y Q2 obtenidos en esta sección son menores que los del modelo completo, lo cual es esperable ya que no se está capturando toda la variabilidad posible. Aun así, los valores obtenidos (R2Y = 0.514 y Q2 = 0.502) siguen siendo aceptables y reflejan que incluso con solo dos componentes el modelo conserva una capacidad predictiva moderada. El test de

permutación confirma que estos resultados son estadísticamente significativos (pR2Y = 0.05, pQ2 = 0.05), lo que respalda la validez del modelo.

El gráfico de scores proyecta a los individuos en el plano definido por las componentes t1 y t2, y permite observar una separación clara entre los grupos Down y Control, lo que indica que las variables de expresión seleccionadas permiten distinguir entre ambos genotipos incluso en un espacio reducido. Por su parte, el gráfico de observaciones (observation diagnostics) ayuda a detectar posibles valores atípicos mediante la evaluación de la distancia ortogonal y la distancia de scores. En principio estos atípicos son aceptados ya que anteiormente en el estudio has sido tratados.

El análisis PLS-DA aplicado a los datos de expresión de los ratones ha demostrado ser una herramienta eficaz para discriminar entre los genotipos Down y Control. El modelo completo, ajustado con ocho componentes, presentó una elevada capacidad explicativa ($R2Y \approx 0.79$) y predictiva ($Q2 \approx 0.76$), con significación estadística validada mediante test de permutación. Además, aunque la representación gráfica se ha realizado utilizando únicamente los dos primeros componentes, esta ha permitido visualizar de forma clara la separación entre los grupos. En conjunto, los resultados indican que las variables de expresión consideradas contienen información relevante para diferenciar los genotipos y que el modelo PLS-DA es apropiado para abordar este tipo de problema de clasificación supervisada.

Métodos Opcionales

Análisis factorial de correspondencias (simple y múltiple):

El análisis factorial de correspondencias (AFC) es un método descriptivo no supervisado que analiza las relaciones entre **variables categóricas**. El AFC simple se aplica a una tabla de contingencia cruzada de dos variables categóricas mientras que el AFC múltiple se aplica a una base de datos compuesta por múltiples variables categóricas. Una ventaja de este método es que permite visualizar asociaciones entre distintos grupos de ratones según sus características categóricas, reduciendo la dimensionalidad del espacio. Algo inconveniente es la necesidad de convertir las variables numéricas a categóricas si quiere incluirlas en el análisis, discretizándolas. En este caso, la base de datos de los ratones tienen cuatro variables categóricas: *Genotipo*, *Tratamiento*, *Comportamiento* y *Clase*, y se podría aplicar un AFC múltiple a estas variables.

Reglas de asociación:

Las reglas de asociación son un método no supervisado para encontrar patrones frecuentes en la base de datos y permite conocer el comportamiento general de los individuos. Es necesario trabajar con **variables binarios**, entonces debería binarizar las variables numéricas de una base de datos, lo que puede reducir la información o crear demasiadas variables, introduciendo ruido. En esta base de datos, se podría binarizar las variables numéricas, convertiéndolas en transacciones, y aplicar el algoritmo apriori con soporte y confianza.

Análisis discriminante:

El análisis discriminate es un método supervisado para clasificar un individuo en su clase más cercana, buscando un conjunto de reglas discriminantes. Permite estudiar la contribución de cada variable, lo que ayuda en la explicación y significación. Algo inconveniente es que supone normalidad y homogeneidad de las variables para aplicar el método. Además puede ser sensible a valores atípicos. En este caso, se podría aplicar para predecir la *Clase* de un ratón en función de sus medidas de las proteínas.

Conclusiones del Estudio

Se ha investigado cómo determinadas proteinas pueden estar relacionadas con la capacidad de aprendizaje en ratones con síndrome de Down. Para ello, se ha trabajado con una base de datos detallada sobre la expresión de proteínas cerebrales en estos animales, considerando además factores como el tipo de tratamiento farmacológico recibido y la estimulación cerebral. Mediante análisis de componentes principales han destacado algunas proteínas cuya expresión parece especialmente relevante, como *pCAMKII_N*, *pKCG_N*, *NR2A_N* y *ERK_N*. Estas podrían estar implicadas en los mecanismos que afectan al aprendizaje, sobre todo en relación con la administración de memantina y los estímulos dirigidos a potenciar esta capacidad. Sin embargo, no es concluyente la efectividad de este fármaco. El análisis de clustering ha revelado que los ratones pueden dividirse en tres grupos bien diferenciados, lo cual respalda la idea de que existen perfiles distintos según las condiciones experimentales. Por otra parte, el modelo PLS-DA ha demostrado que es posible distinguir con bastante precisión entre ratones con y sin trisomía basándose únicamente en las expresiones proteicas, lo que refuerza el valor de esta información para fines de clasificación y diagnóstico. Es por ello que los resultados sugieren que hay diferencias relevantes a nivel molecular entre los distintos grupos de ratones, y que estas podrían estar influyendo en su comportamiento y capacidad de aprendizaje. Además en este estudio se han aplicado un conjunto de técnicas estadísticas variadas, desde el preprocesamiento de los datos hasta modelos supervisados y no supervisados, lo que muestra la potencia de análisis de la ciencia de datos en los campos como la biología.

Anexos

Anexo 1: Análisis exploratorio

```
##     DYRK1A_N         ITSN1_N           BDNF_N           NR1_N      
##  Min.   :0.1453   Min.   :0.2454   Min.   :0.1152   Min.   :1.331  
##  1st Qu.:0.2881   1st Qu.:0.4734   1st Qu.:0.2874   1st Qu.:2.057  
##  Median :0.3664   Median :0.5658   Median :0.3166   Median :2.297  
##  Mean   :0.4258   Mean   :0.6171   Mean   :0.3191   Mean   :2.297  
##  3rd Qu.:0.4877   3rd Qu.:0.6980   3rd Qu.:0.3482   3rd Qu.:2.528  
##  Max.   :2.5164   Max.   :2.6027   Max.   :0.4972   Max.   :3.758  
##  NA's   :3        NA's   :3        NA's   :3        NA's   :3      
##     NR2A_N          pAKT_N           pBRAF_N          pCAMKII_N      
##  Min.   :1.738   Min.   :0.06324   Min.   :0.06404   Min.   :1.344  
```

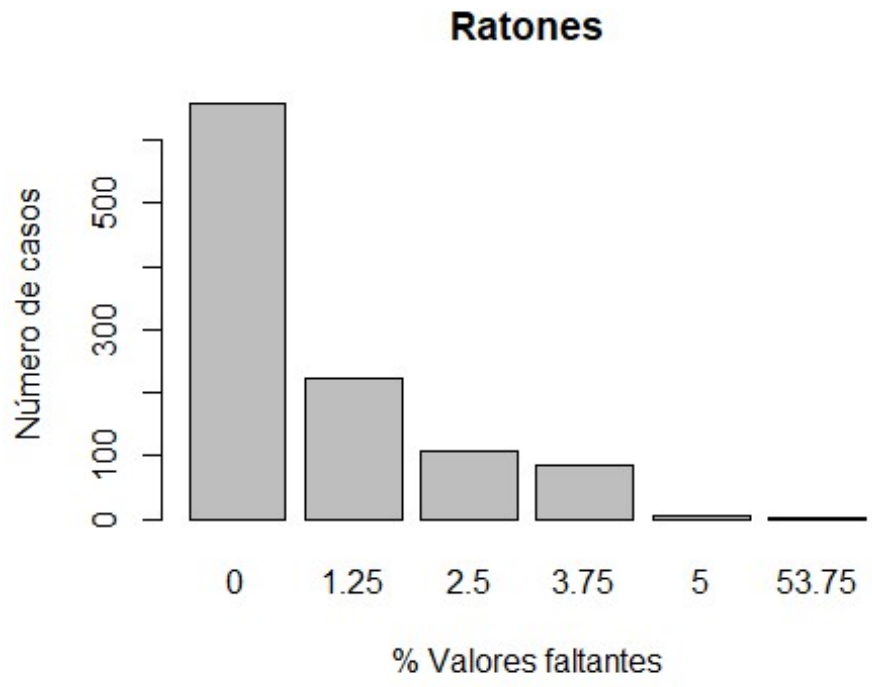
```
##  1st Qu.:3.156   1st Qu.:0.20575   1st Qu.:0.16459   1st Qu.:2.480
##  Median :3.761   Median :0.23118   Median :0.18230   Median :3.327
##  Mean   :3.844   Mean   :0.23317   Mean   :0.18185   Mean   :3.537
##  3rd Qu.:4.440   3rd Qu.:0.25726   3rd Qu.:0.19742   3rd Qu.:4.482
##  Max.   :8.483   Max.   :0.53905   Max.   :0.31707   Max.   :7.464
##  NA's   :3       NA's   :3         NA's   :3         NA's   :3
##     pCREB_N           pELK_N           pERK_N           pJNK_N
##  Min.   :0.1128   Min.   :0.429   Min.   :0.1492   Min.   :0.05211
##  1st Qu.:0.1908   1st Qu.:1.204   1st Qu.:0.3374   1st Qu.:0.28124
##  Median :0.2106   Median :1.356   Median :0.4436   Median :0.32133
##  Mean   :0.2126   Mean   :1.429   Mean   :0.5459   Mean   :0.31351
##  3rd Qu.:0.2346   3rd Qu.:1.561   3rd Qu.:0.6633   3rd Qu.:0.34871
##  Max.   :0.3062   Max.   :6.113   Max.   :3.5667   Max.   :0.49343
##  NA's   :3        NA's   :3       NA's   :3        NA's   :3
##      PKCA_N           pMEK_N            pNR1_N           pNR2A_N
##  Min.   :0.1914   Min.   :0.05682   Min.   :0.5002   Min.   :0.2813
##  1st Qu.:0.2818   1st Qu.:0.24429   1st Qu.:0.7435   1st Qu.:0.5903
##  Median :0.3130   Median :0.27739   Median :0.8211   Median :0.7196
##  Mean   :0.3179   Mean   :0.27503   Mean   :0.8258   Mean   :0.7269
##  3rd Qu.:0.3523   3rd Qu.:0.30345   3rd Qu.:0.8985   3rd Qu.:0.8486
##  Max.   :0.4740   Max.   :0.45800   Max.   :1.4082   Max.   :1.4128
##  NA's   :3        NA's   :3         NA's   :3        NA's   :3
##     pNR2B_N          pPKCAB_N          pRSK_N            AKT_N
##  Min.   :0.3016   Min.   :0.5678   Min.   :0.09594   Min.   :0.06442
##  1st Qu.:1.3813   1st Qu.:1.1683   1st Qu.:0.40414   1st Qu.:0.59682
##  Median :1.5637   Median :1.3657   Median :0.44060   Median :0.68247
##  Mean   :1.5620   Mean   :1.5253   Mean   :0.44285   Mean   :0.68224
##  3rd Qu.:1.7485   3rd Qu.:1.8859   3rd Qu.:0.48210   3rd Qu.:0.75969
##  Max.   :2.7240   Max.   :3.0614   Max.   :0.65096   Max.   :1.18217
##  NA's   :3        NA's   :3        NA's   :3         NA's   :3
##      BRAF_N          CAMKII_N          CREB_N           ELK_N
##  Min.   :0.1439   Min.   :0.2130   Min.   :0.1136   Min.   :0.4977
##  1st Qu.:0.2643   1st Qu.:0.3309   1st Qu.:0.1618   1st Qu.:0.9444
##  Median :0.3267   Median :0.3603   Median :0.1796   Median :1.0962
##  Mean   :0.3785   Mean   :0.3634   Mean   :0.1805   Mean   :1.1734
##  3rd Qu.:0.4136   3rd Qu.:0.3939   3rd Qu.:0.1957   3rd Qu.:1.3236
##  Max.   :2.1334   Max.   :0.5862   Max.   :0.3196   Max.   :2.8029
##  NA's   :3        NA's   :3        NA's   :3        NA's   :18
##      ERK_N          GSK3B_N           JNK_N            MEK_N
##  Min.   :1.132   Min.   :0.1511   Min.   :0.0463   Min.   :0.1472
##  1st Qu.:1.992   1st Qu.:1.0231   1st Qu.:0.2204   1st Qu.:0.2471
##  Median :2.401   Median :1.1598   Median :0.2449   Median :0.2734
##  Mean   :2.474   Mean   :1.1726   Mean   :0.2416   Mean   :0.2728
##  3rd Qu.:2.873   3rd Qu.:1.3097   3rd Qu.:0.2633   3rd Qu.:0.3008
##  Max.   :5.198   Max.   :2.4758   Max.   :0.3872   Max.   :0.4154
##  NA's   :3       NA's   :3        NA's   :3        NA's   :7
##      TRKA_N           RSK_N            APP_N          Bcatenin_N
##  Min.   :0.1987   Min.   :0.1074   Min.   :0.2356   Min.   :1.135
##  1st Qu.:0.6171   1st Qu.:0.1496   1st Qu.:0.3663   1st Qu.:1.827
##  Median :0.7050   Median :0.1667   Median :0.4020   Median :2.115
```

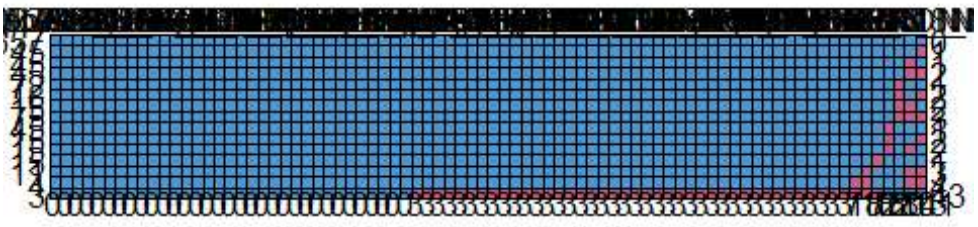
```
##  Mean   :0.6932   Mean   :0.1684   Mean   :0.4048   Mean   :2.147
##  3rd Qu.:0.7742   3rd Qu.:0.1845   3rd Qu.:0.4419   3rd Qu.:2.424
##  Max.   :1.0016   Max.   :0.3051   Max.   :0.6327   Max.   :3.681
##  NA's   :3        NA's   :3        NA's   :3        NA's   :18
##      SOD1_N           MTOR_N           P38_N          pMTOR_N
##  Min.   :0.2171   Min.   :0.2011   Min.   :0.2279   Min.   :0.1666
##  1st Qu.:0.3196   1st Qu.:0.4104   1st Qu.:0.3520   1st Qu.:0.6835
##  Median :0.4441   Median :0.4525   Median :0.4078   Median :0.7608
##  Mean   :0.5426   Mean   :0.4525   Mean   :0.4153   Mean   :0.7590
##  3rd Qu.:0.6958   3rd Qu.:0.4880   3rd Qu.:0.4663   3rd Qu.:0.8415
##  Max.   :1.8729   Max.   :0.6767   Max.   :0.9333   Max.   :1.1249
##  NA's   :3        NA's   :3        NA's   :3        NA's   :3
##     DSCR1_N          AMPKA_N          NR2B_N          pNUMB_N
##  Min.   :0.1553   Min.   :0.2264   Min.   :0.1848   Min.   :0.1856
##  1st Qu.:0.5309   1st Qu.:0.3266   1st Qu.:0.5149   1st Qu.:0.3128
##  Median :0.5767   Median :0.3585   Median :0.5635   Median :0.3474
##  Mean   :0.5852   Mean   :0.3684   Mean   :0.5653   Mean   :0.3571
##  3rd Qu.:0.6344   3rd Qu.:0.4008   3rd Qu.:0.6145   3rd Qu.:0.3927
##  Max.   :0.9164   Max.   :0.7008   Max.   :0.9720   Max.   :0.6311
##  NA's   :3        NA's   :3        NA's   :3        NA's   :3
##     RAPTOR_N         TIAM1_N          pP70S6_N         NUMB_N
##  Min.   :0.1948   Min.   :0.2378   Min.   :0.1311   Min.   :0.1180
##  1st Qu.:0.2761   1st Qu.:0.3720   1st Qu.:0.2811   1st Qu.:0.1593
##  Median :0.3049   Median :0.4072   Median :0.3777   Median :0.1782
##  Mean   :0.3158   Mean   :0.4186   Mean   :0.3945   Mean   :0.1811
##  3rd Qu.:0.3473   3rd Qu.:0.4560   3rd Qu.:0.4811   3rd Qu.:0.1972
##  Max.   :0.5267   Max.   :0.7221   Max.   :1.1292   Max.   :0.3166
##  NA's   :3        NA's   :3        NA's   :3
##     P70S6_N          pGSK3B_N         pPKCG_N          CDK5_N
##  Min.   :0.3441   Min.   :0.09998   Min.   :0.5988   Min.   :0.1812
##  1st Qu.:0.8267   1st Qu.:0.14925   1st Qu.:1.2968   1st Qu.:0.2726
##  Median :0.9313   Median :0.16021   Median :1.6646   Median :0.2938
##  Mean   :0.9431   Mean   :0.16121   Mean   :1.7066   Mean   :0.2924
##  3rd Qu.:1.0451   3rd Qu.:0.17174   3rd Qu.:2.1130   3rd Qu.:0.3125
##  Max.   :1.6800   Max.   :0.25321   Max.   :3.3820   Max.   :0.8174
##
##       S6_N          ADARB1_N       AcetylH3K9_N         RRP1_N
##  Min.   :0.1302   Min.   :0.5291   Min.   :0.05253   Min.   :-0.06201
##  1st Qu.:0.3167   1st Qu.:0.9305   1st Qu.:0.10357   1st Qu.: 0.14902
##  Median :0.4010   Median :1.1283   Median :0.15042   Median : 0.16210
##  Mean   :0.4292   Mean   :1.1974   Mean   :0.21648   Mean   : 0.16663
##  3rd Qu.:0.5349   3rd Qu.:1.3802   3rd Qu.:0.26965   3rd Qu.: 0.17741
##  Max.   :0.8226   Max.   :2.5399   Max.   :1.45939   Max.   : 0.61238
##
##      BAX_N            ARC_N           ERBB4_N          nNOS_N
##  Min.   :0.07233   Min.   :0.06725   Min.   :0.1002   Min.   :0.09973
##  1st Qu.:0.16817   1st Qu.:0.11084   1st Qu.:0.1470   1st Qu.:0.16645
##  Median :0.18074   Median :0.12163   Median :0.1564   Median :0.18267
##  Mean   :0.17931   Mean   :0.12152   Mean   :0.1565   Mean   :0.18130
##  3rd Qu.:0.19158   3rd Qu.:0.13196   3rd Qu.:0.1654   3rd Qu.:0.19857
```

```
##  Max.   :0.24114   Max.   :0.15875   Max.   :0.2087   Max.   :0.26074
##
##      Tau_N             GFAP_N           GluR3_N          GluR4_N
##  Min.   :0.09623   Min.   :0.08611   Min.   :0.1114   Min.   :0.07258
##  1st Qu.:0.16799   1st Qu.:0.11277   1st Qu.:0.1957   1st Qu.:0.10889
##  Median :0.18863   Median :0.12046   Median :0.2169   Median :0.12355
##  Mean   :0.21049   Mean   :0.12089   Mean   :0.2219   Mean   :0.12656
##  3rd Qu.:0.23394   3rd Qu.:0.12772   3rd Qu.:0.2460   3rd Qu.:0.14195
##  Max.   :0.60277   Max.   :0.21362   Max.   :0.3310   Max.   :0.53700
##
##      IL1B_N          P3525_N          pCASP9_N         PSD95_N
##  Min.   :0.2840   Min.   :0.2074   Min.   :0.8532   Min.   :1.206
##  1st Qu.:0.4756   1st Qu.:0.2701   1st Qu.:1.3756   1st Qu.:2.079
##  Median :0.5267   Median :0.2906   Median :1.5227   Median :2.242
##  Mean   :0.5273   Mean   :0.2913   Mean   :1.5483   Mean   :2.235
##  3rd Qu.:0.5770   3rd Qu.:0.3116   3rd Qu.:1.7131   3rd Qu.:2.420
##  Max.   :0.8897   Max.   :0.4437   Max.   :2.5862   Max.   :2.878
##
##      SNCA_N        Ubiquitin_N     pGSK3B_Tyr216_N      SHH_N
##  Min.   :0.1012   Min.   :0.7507   Min.   :0.5774   Min.   :0.1559
##  1st Qu.:0.1428   1st Qu.:1.1163   1st Qu.:0.7937   1st Qu.:0.2064
##  Median :0.1575   Median :1.2366   Median :0.8499   Median :0.2240
##  Mean   :0.1598   Mean   :1.2393   Mean   :0.8488   Mean   :0.2267
##  3rd Qu.:0.1733   3rd Qu.:1.3631   3rd Qu.:0.9162   3rd Qu.:0.2417
##  Max.   :0.2576   Max.   :1.8972   Max.   :1.2046   Max.   :0.3583
##
##      BAD_N            BCL2_N            pS6_N            pCFOS_N
##  Min.   :0.0883   Min.   :0.08066   Min.   :0.06725   Min.   :0.08542
##  1st Qu.:0.1364   1st Qu.:0.11555   1st Qu.:0.11084   1st Qu.:0.11351
##  Median :0.1523   Median :0.12947   Median :0.12163   Median :0.12652
##  Mean   :0.1579   Mean   :0.13476   Mean   :0.12152   Mean   :0.13105
##  3rd Qu.:0.1740   3rd Qu.:0.14823   3rd Qu.:0.13196   3rd Qu.:0.14365
##  Max.   :0.2820   Max.   :0.26151   Max.   :0.15875   Max.   :0.25653
##  NA's   :213      NA's   :285                         NA's   :75
##      SYP_N          H3AcK18_N          EGR1_N          H3MeK4_N
##  Min.   :0.2586   Min.   :0.07969   Min.   :0.1055   Min.   :0.1018
##  1st Qu.:0.3981   1st Qu.:0.12585   1st Qu.:0.1551   1st Qu.:0.1651
##  Median :0.4485   Median :0.15824   Median :0.1749   Median :0.1940
##  Mean   :0.4461   Mean   :0.16961   Mean   :0.1831   Mean   :0.2054
##  3rd Qu.:0.4908   3rd Qu.:0.19788   3rd Qu.:0.2045   3rd Qu.:0.2352
##  Max.   :0.7596   Max.   :0.47976   Max.   :0.3607   Max.   :0.4139
##                   NA's   :180       NA's   :210      NA's   :270
##      CaNA_N
##  Min.   :0.5865
##  1st Qu.:1.0814
##  Median :1.3174
##  Mean   :1.3378
##  3rd Qu.:1.5858
##  Max.   :2.1298
##
```

```
##         ERBB4_N         P3525_N           BAX_N          GFAP_N
pGSK3B_Tyr216_N
##      0.09631279      0.10304693      0.10499537      0.10946656
0.11111581
##         PSD95_N           ARC_N           pS6_N        pGSK3B_N
CDK5_N
##      0.11381116      0.11748174      0.11748174      0.11976813
0.12782326
##           SHH_N          nNOS_N     Ubiquitin_N           JNK_N
pNR1_N
##      0.12788720      0.13744589      0.14006649      0.14029269
0.14285225
##        CAMKII_N          MTOR_N          CREB_N         pBRAF_N
SYP_N
##      0.14407820      0.14473863      0.14610219      0.14870841
0.14892665
##          pRSK_N           MEK_N          SNCA_N           APP_N
NR1_N
##      0.15055325      0.15057607      0.15110706      0.15111557
0.15117642
##         pCREB_N          BDNF_N          IL1B_N          NR2B_N
GluR3_N
##      0.15329738      0.15476314      0.15560364      0.15602127
0.15724872
##        pCASP9_N         TIAM1_N         pMTOR_N          NUMB_N
PKCA_N
##      0.16025569      0.16074310      0.16131539      0.16178683
0.16429639
##          pJNK_N           RSK_N          pMEK_N         AMPKA_N
RAPTOR_N
##      0.16579654      0.16707183      0.16784725      0.16989296
0.17150265
##         DSCR1_N         pNR2B_N          TRKA_N         pNUMB_N
pAKT_N
##      0.17191332      0.17333115      0.17430596      0.17649461
0.17855997
##         pCFOS_N         P70S6_N           AKT_N           BAD_N
RRP1_N
##      0.18208401      0.18326233      0.18678854      0.18704142
0.19141420
##      Bcatenin_N          BCL2_N         GSK3B_N         GluR4_N
P38_N
##      0.20299655      0.20344813      0.20872531      0.21242840
0.21499498
##          EGR1_N          CaNA_N          NR2A_N         pNR2A_N
ERK_N
##      0.22063335      0.23705343      0.24274612      0.25863896
0.26405302
##        H3MeK4_N           ELK_N        ADARB1_N        pPKCAB_N
S6_N
```

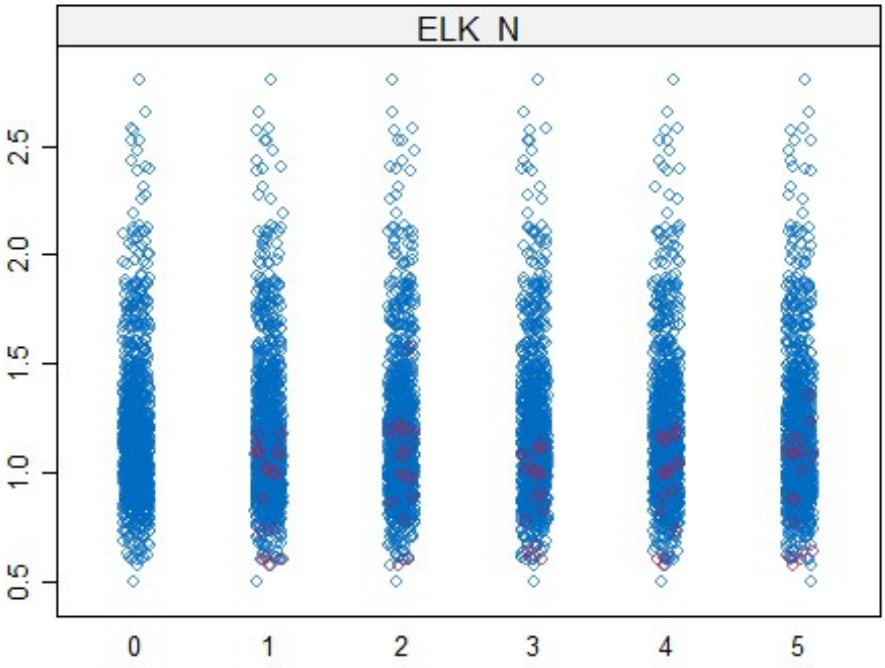
```
##      0.27022021      0.28616467      0.30210232      0.31585137
0.32020516
##          pELK_N           Tau_N         pPKCG_N       H3AcK18_N
pCAMKII_N
##      0.32680731      0.32785485      0.33894453      0.35023056
0.36616611
##        pP70S6_N         ITSN1_N          SOD1_N          BRAF_N
DYRK1A_N
##      0.39589338      0.40777706      0.51704784      0.57167051
0.58561769
##          pERK_N    AcetylH3K9_N
##      0.63254541      0.85599977

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.7787  1.0000 43.0000
```

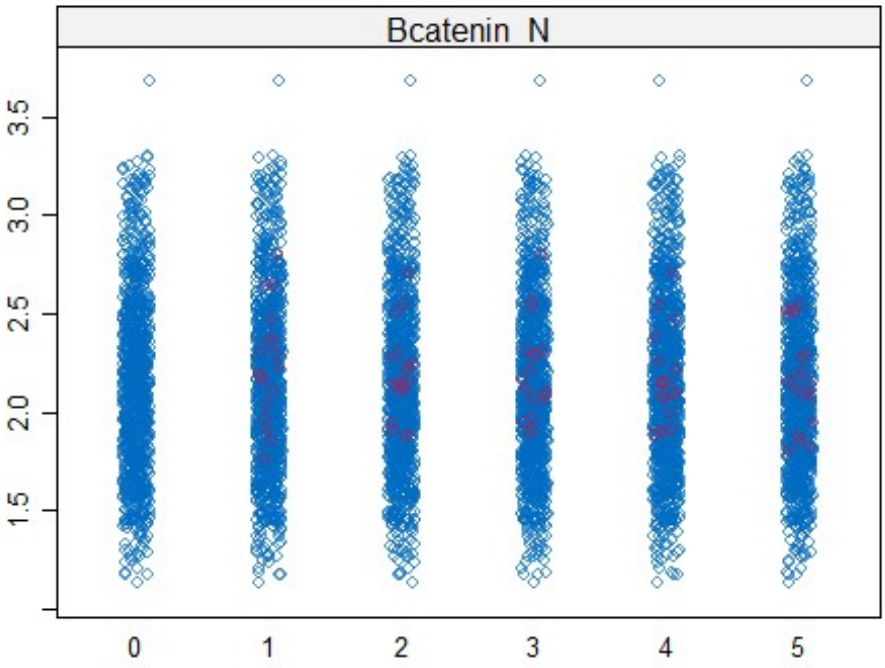


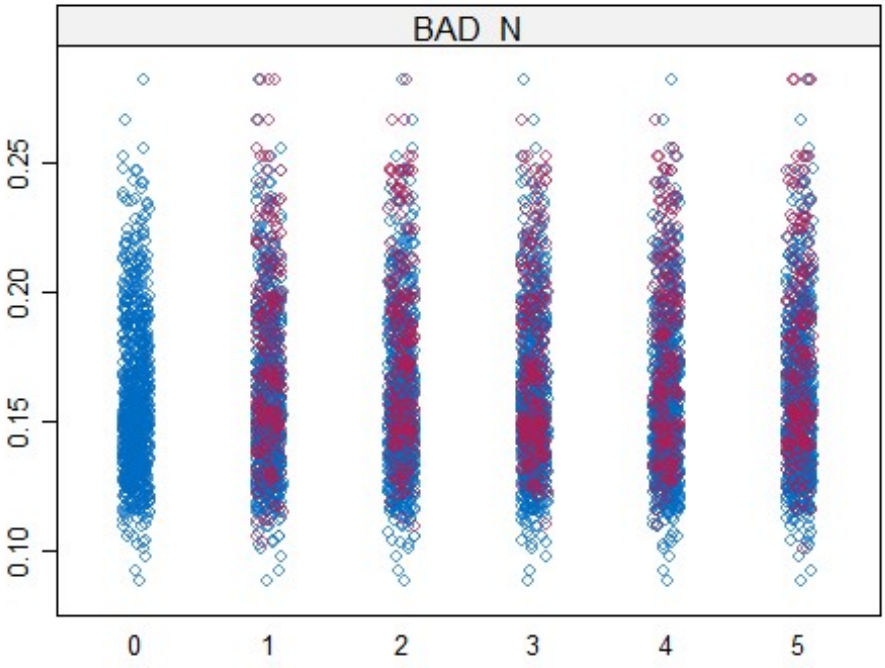


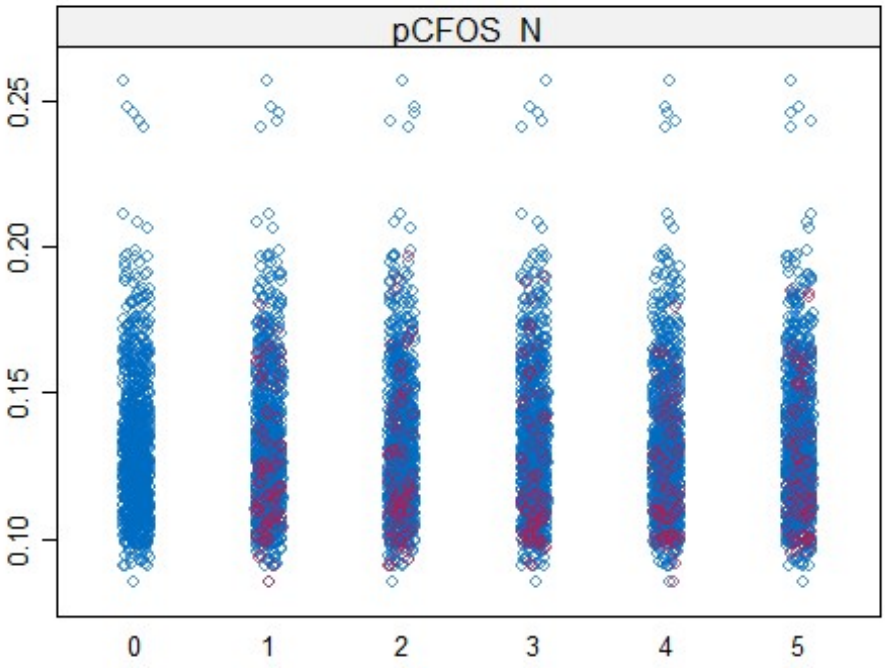
```
##               Variable numNA percNA
## ELK_N           ELK_N    18   1.67
## Bcatenin_N Bcatenin_N    18   1.67
## BAD_N           BAD_N   213  19.72
```

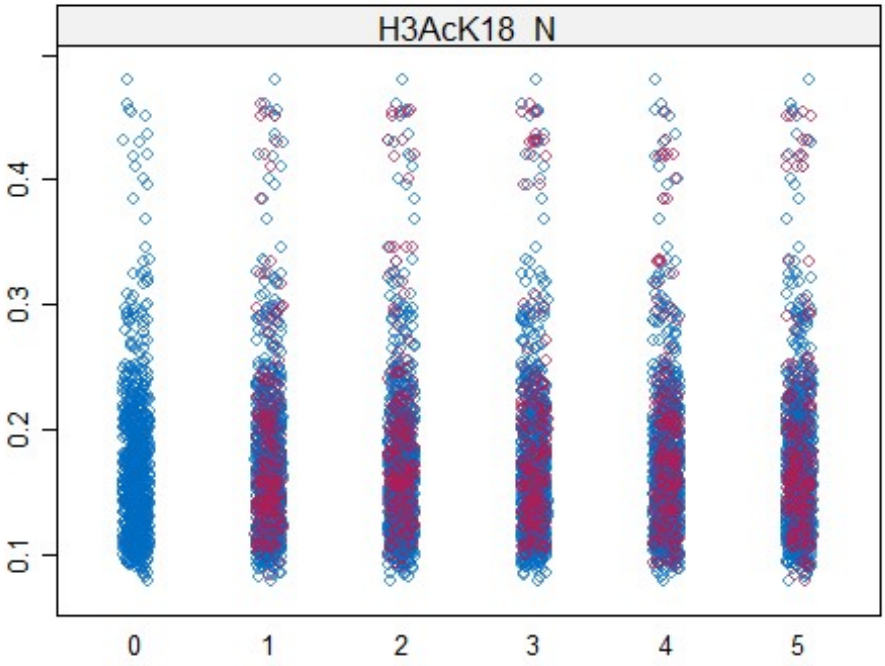
```
## pCFOS_N        pCFOS_N    75   6.94
## H3AcK18_N   H3AcK18_N   180  16.67
## EGR1_N         EGR1_N   210  19.44
```

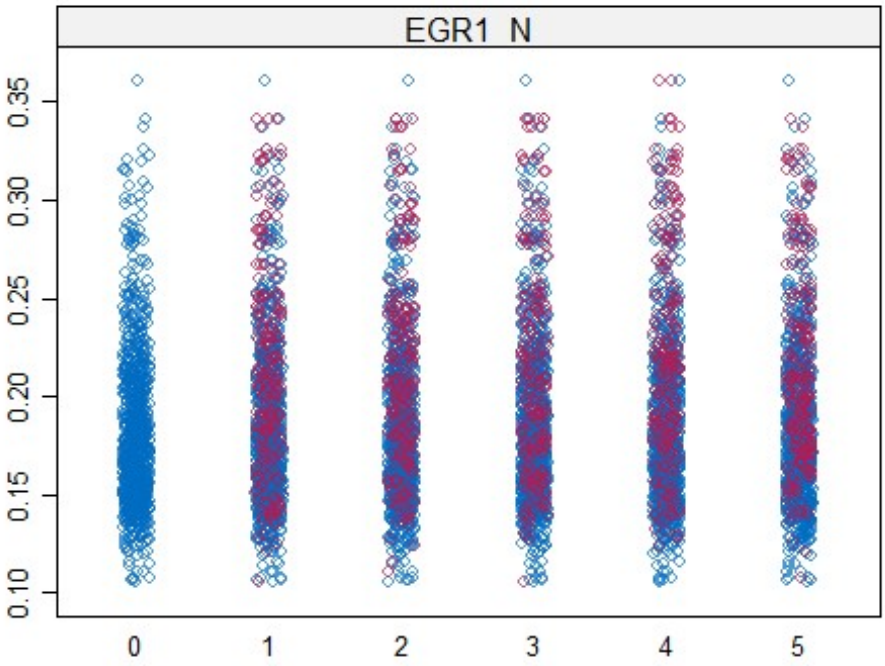


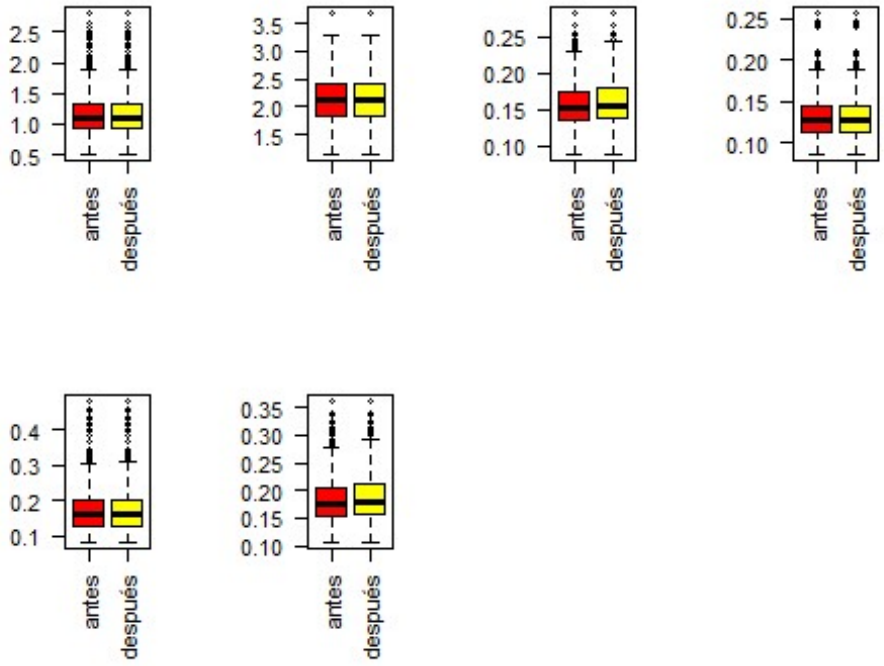



BAD N
0.10
0.15
0.20
0.25
0
1
2
3
4
5


pCFOS N
0.10
0.15
0.20
0.25
0
1
2
3
4
5

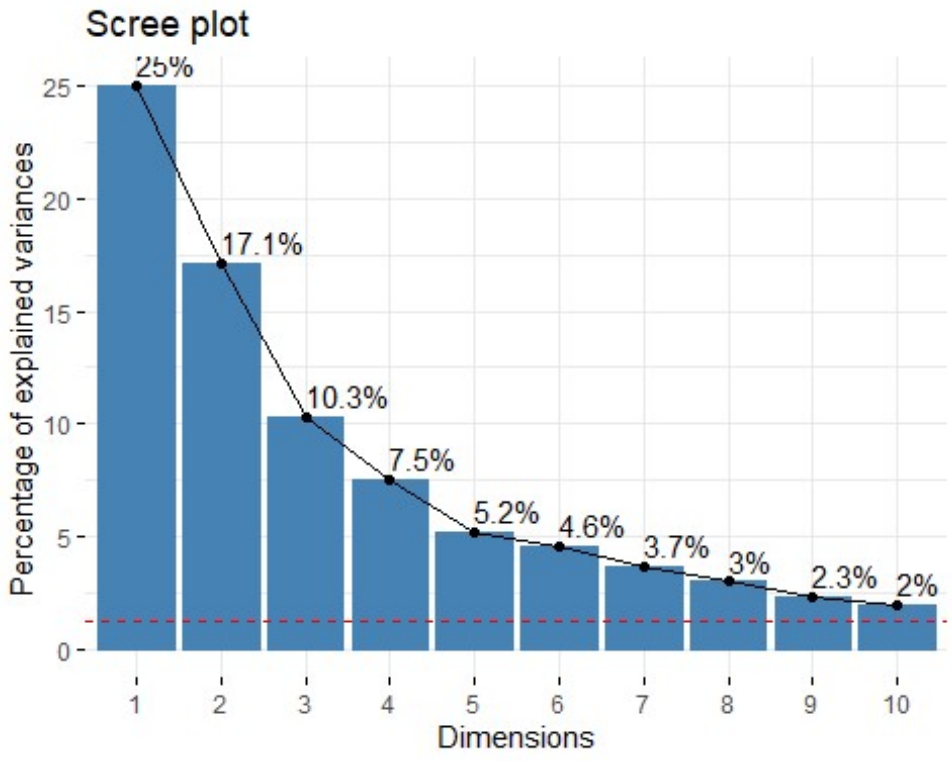

H3AcK18_N
0.1
0.2
0.3
0.4
0
1
2
3
4
5


EGR1_N
0.10
0.15
0.20
0.25
0.30
0.35
0
1
2
3
4
5



Anexo 2: PCA con escalado

En el PCA con escalado se aplica el mismo procedimiento que al PCA sin escalar. Se crea el PCA auxiliar para hacer el scree plot y al observarlo se escogen también 4 dimensiones para el PCA escalado. En este también se añaden como variables suplementarias todas las variables categóricas y binarias.

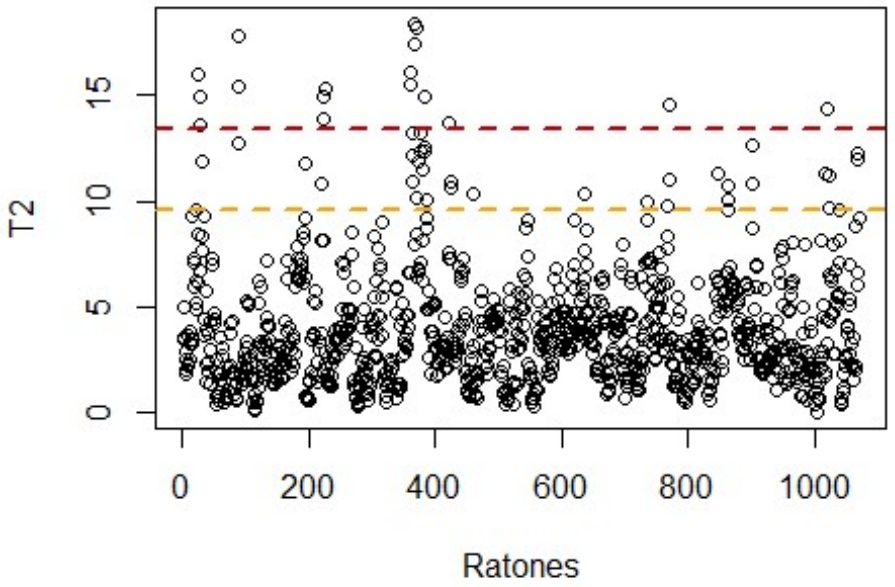


	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	19.484372	24.979964	24.97996
Dim.2	13.373964	17.146107	42.12607
Dim.3	8.059865	10.333160	52.45923
Dim.4	5.849918	7.499894	59.95913
Dim.5	4.051918	5.194766	65.15389
Dim.6	3.594851	4.608784	69.76267

Anomalos con escalar

En el caso del PCA escalado se grafica también el T2 de Hotelling. Se guarda la T2 del PCA con escalado en otra variable y se vuelven a calcular los límites F95 y F99. Se plotea otra

vez el gráfico incluyendo las dos líneas correspondientes a los límites.

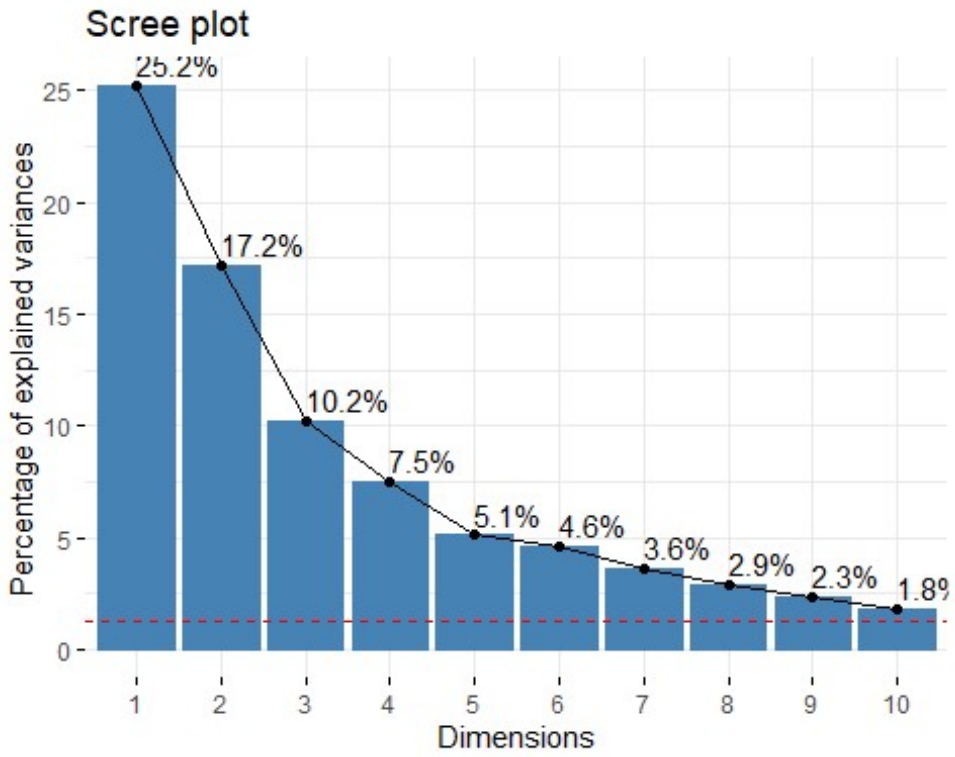



```
##   22   26   28   29   30   88   89   90  193  221  223  224  225  365  366
367
##   22   26   28   29   30   88   89   90  193  221  223  224  225  361  362
363
##  368  369  371  372  378  379  382  385  386  388  391  392  393  396  430
433
##  364  365  366  367  369  370  373  376  377  379  382  383  384  387  421
424
##  434  469  646  744  778  779  780  855  871  873  910  911 1027 1030 1033
1034
##  425  460  637  735  769  770  771  846  862  864  901  902 1018 1021 1024
1025
## 1050 1078 1079
## 1041 1069 1070
```

Usando el mismo criterio para mostrar los anómalos extremos (>F95*2) se puede ver que escalando hay muchos menos. Existen solo 3 observaciones anómalas que también estaban en el PCA sin escalado, correspondiendo las tres al mismo ratón.

PCA con escalar (sin anómalos)

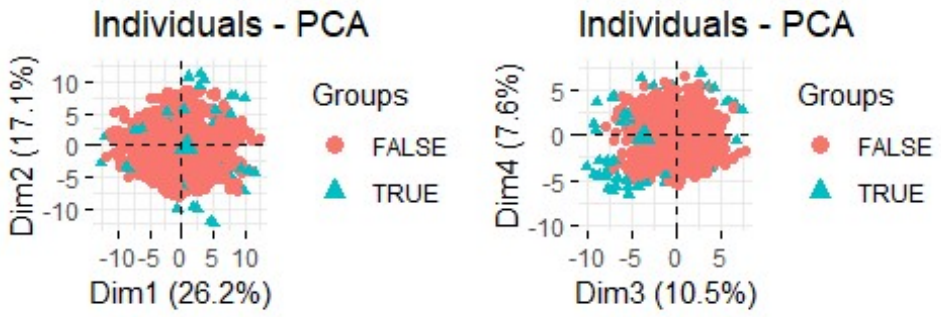
Se crea también el PCA escalado sin datos anómalos. Para ello se usa el mismo procedimiento. Se crea el scree plot con un PCA auxiliar y se puede ver que no cambia demasiado, por ello se crea el PCA escalado definitivo con 4 dimensiones también.



	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	19.915333	25.209282	25.20928
Dim.2	13.567007	17.173427	42.38271
Dim.3	8.077460	10.224633	52.60734
Dim.4	5.919307	7.492793	60.10014
Dim.5	4.047136	5.122957	65.22309
Dim.6	3.623349	4.586518	69.80961

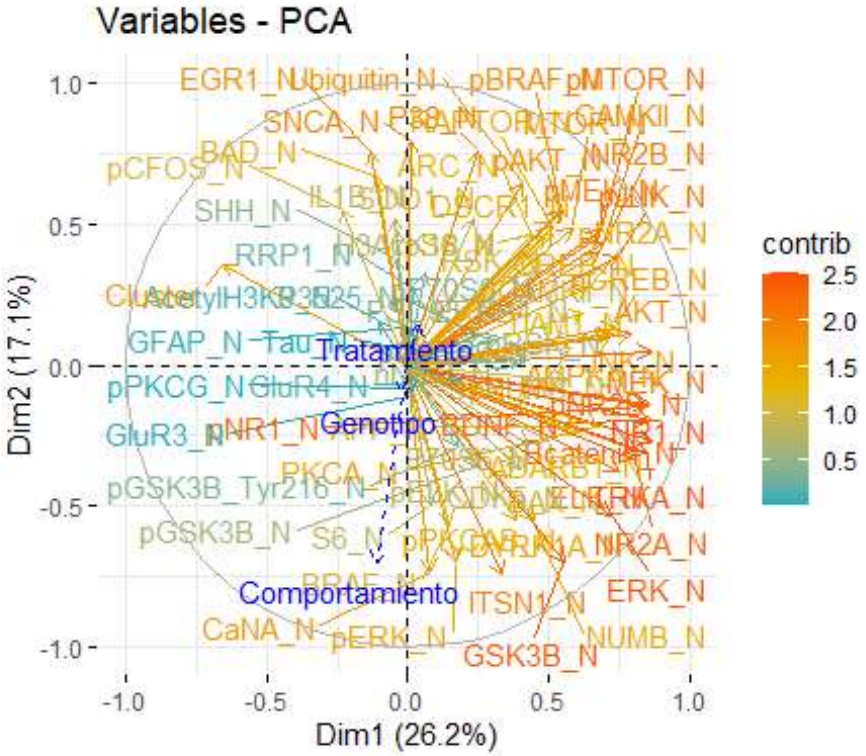
Graficos individuos con escalar

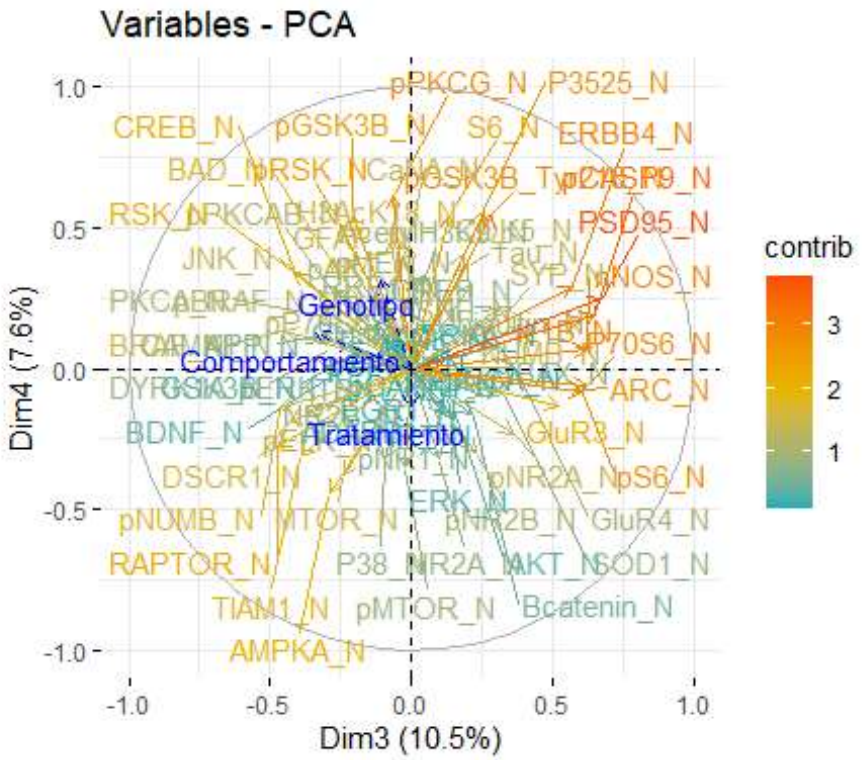
Se crea el mismo gráfico de individuos separando las observaciones anómalas incluyendo factor(T2 > F95) en habillage.



En el gráfico con escalado se ve que los que tienen valores atípicos no se separan tanto con los valores normales como en el gráfico sin escalar, pero se pueden diferenciar un poco aún.

Grafico variables con escalar





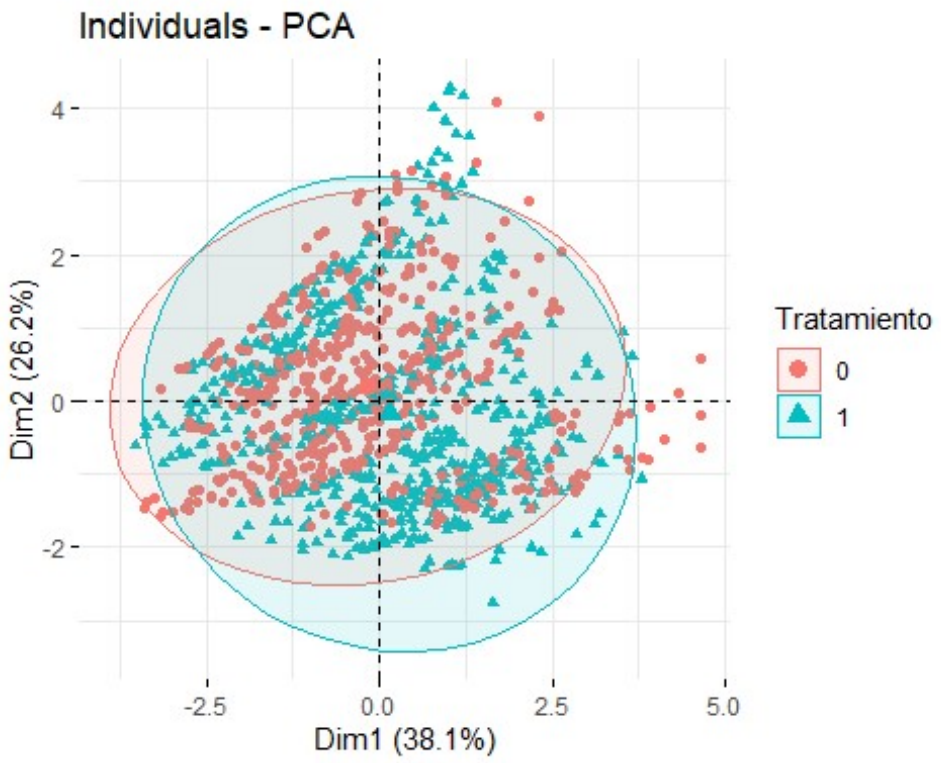
Anexo 3: Gráfico individuos por variable binaria

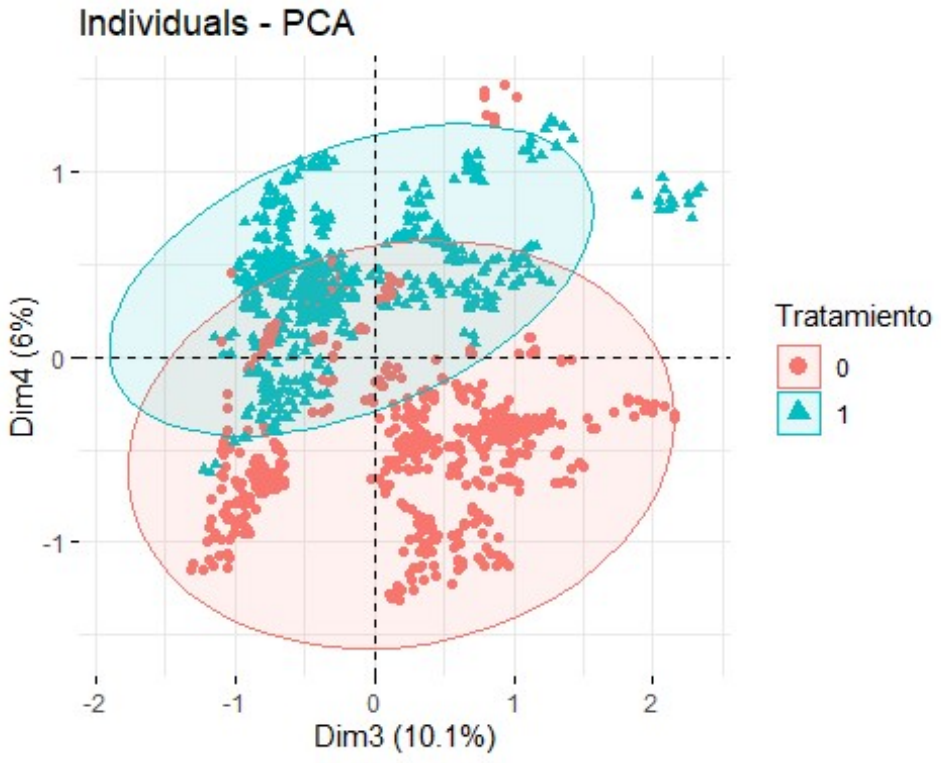
En este apartado se van a realizar gráficos de individuos agrupados por las diferentes variables binarias que hay en la base de datos, el propósito es ver si las dimensiones resultantes del PCA separan a los ratones por tipos (*Genotipo*, *Comportamiento*, *Tratamiento* o *Clase*).

Para ello se crearán gráficos de individuos con una variable binaria o categórica en el campo habillage, además se añadirán elipses en los grupos para ver mejor su distribución. Para cada variable que se desee estudiar se harán dos gráficos, uno para las dimensiones 1 y 2 y otro para las dimensiones 3 y 4.

Tratamiento

Primero, se ve si hay diferencia entre los ratones que se tratan con la droga memantina para recuperar la capacidad de aprender o no. Se añade en habillage la variable "Tratamiento".



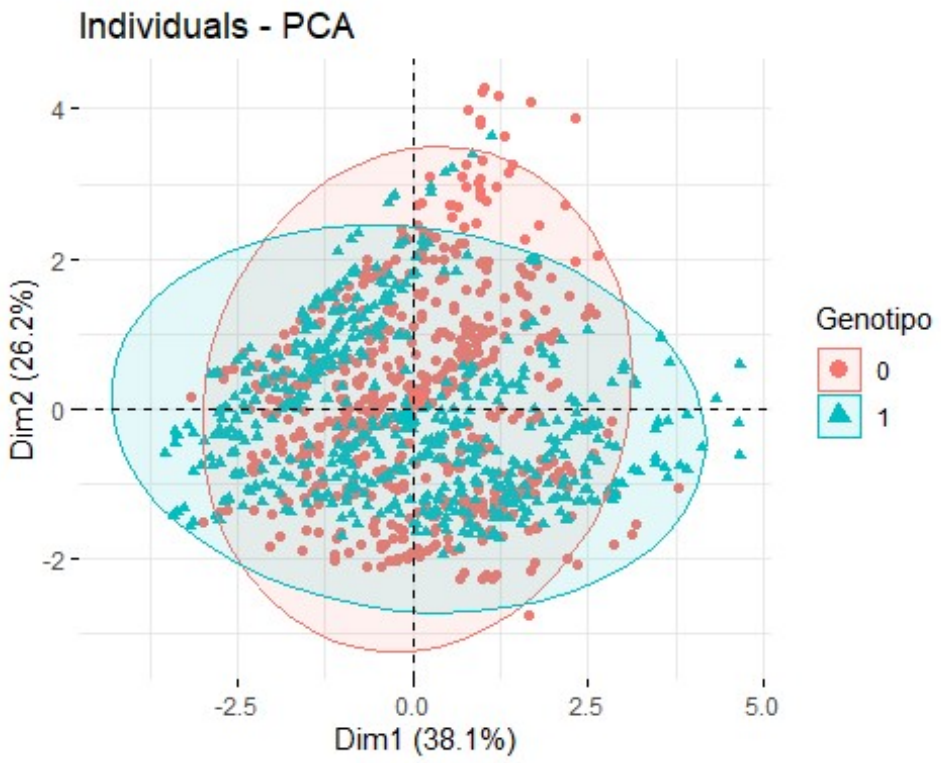


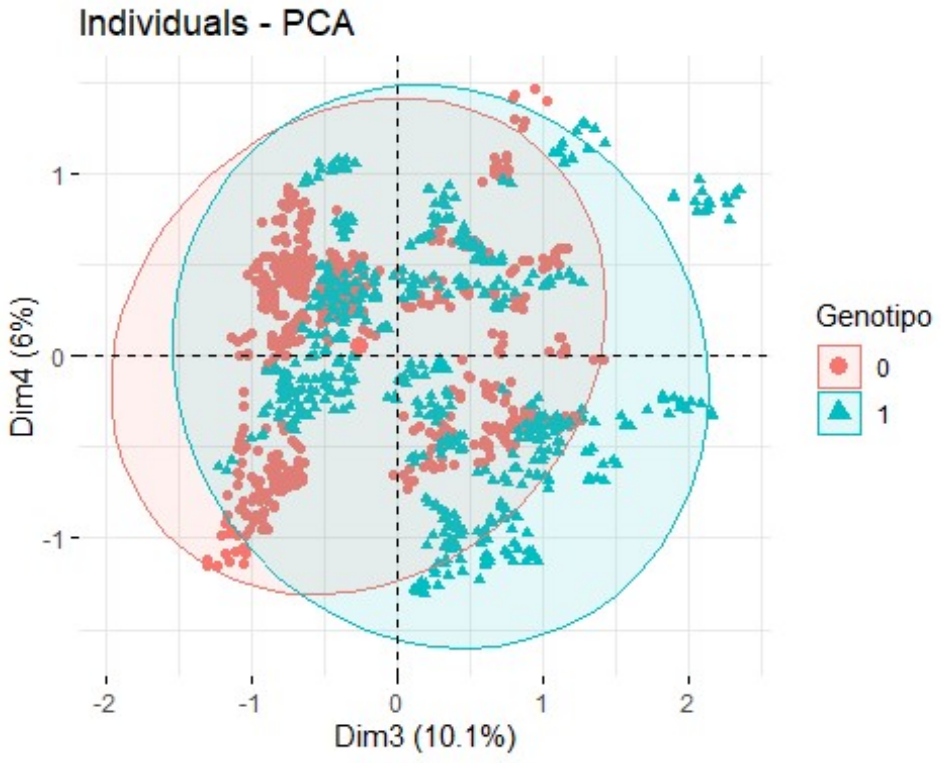
En el PCA ninguna de las dimensiones separa los ratones según su tratamiento. En todas las dimensiones los grupos de ratones que se tratan y que no se tratan se solapan en gran

parte de la elipse. Por tanto, con este análisis no se ve un efecto significativo del tratamiento con la droga memantina en general.

Genotipo

Después, se ve si habían diferencias entre las proteínas de ratones trisómicos (con síndrome de Down) y control. Para ello se agrega en el campo habillage la variable *Genotipo*.



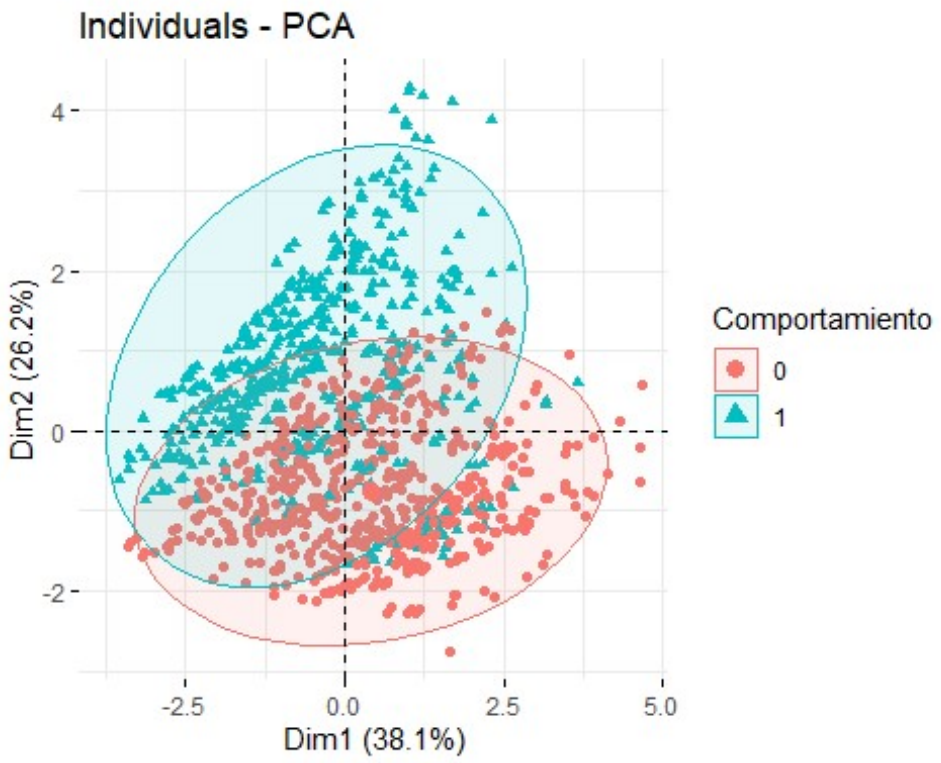


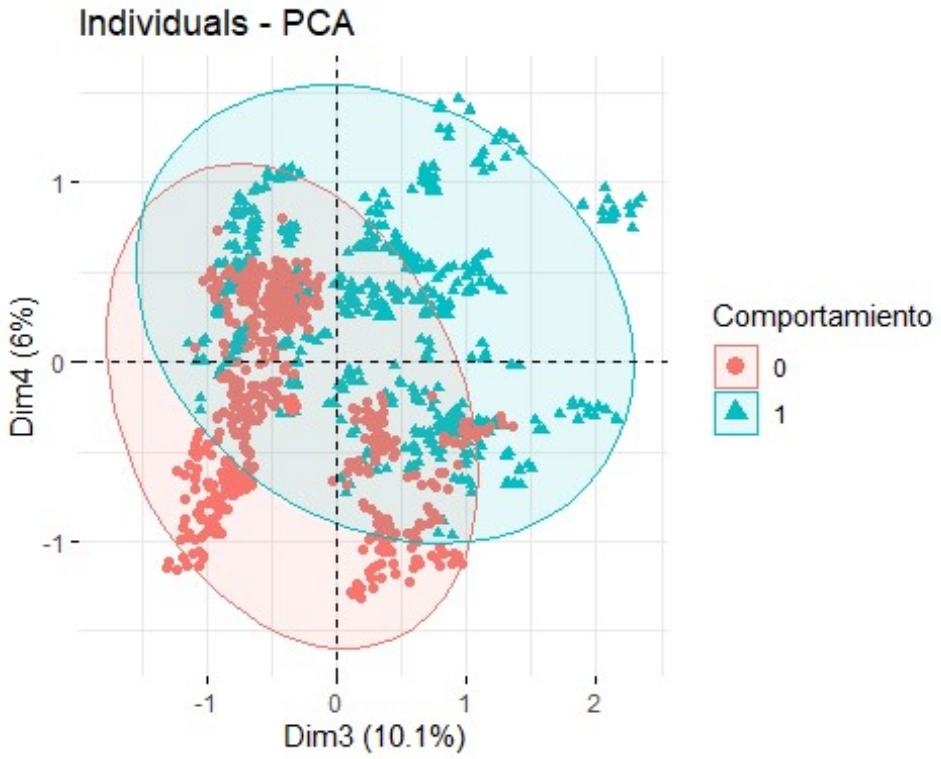
En los gráficos del PCA no se ve diferencias significativas entre los ratones trisómicos y control. En todas las dimensiones las distribuciones de ambos grupos se solapan. Al

menos con este análisis no se ven diferencias entre la expresión de proteínas de ratones trisómicos y control.

Comportamiento

A continuación, se visualiza los grupos de ratones según si estaban estimulados para aprender o no. Se agrega la variable *Comportamiento* en habillage.



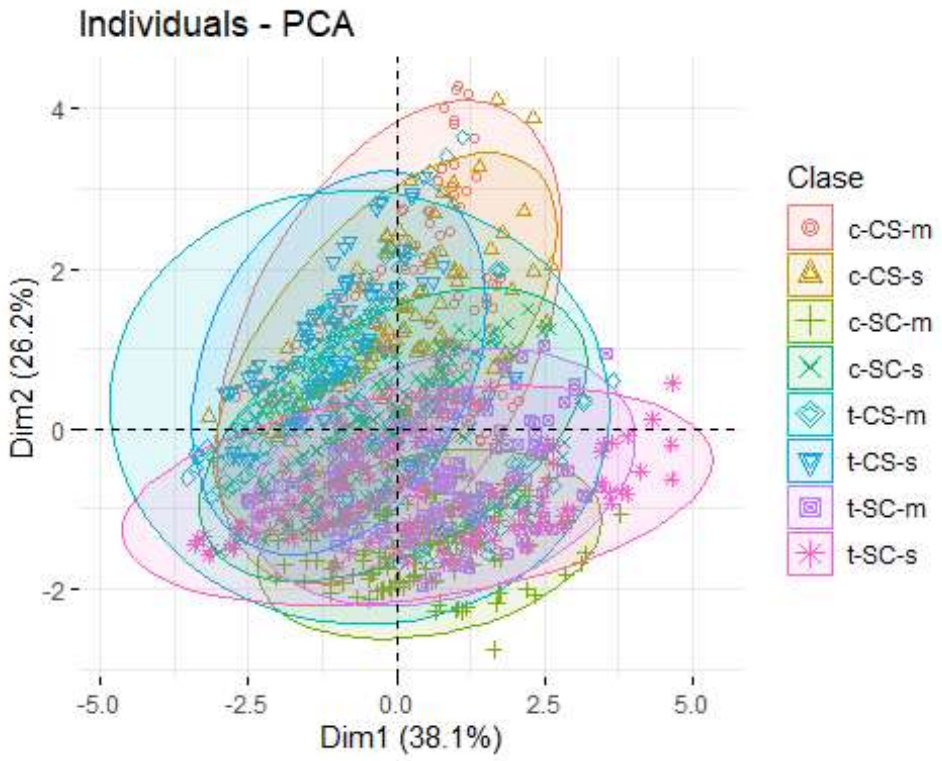


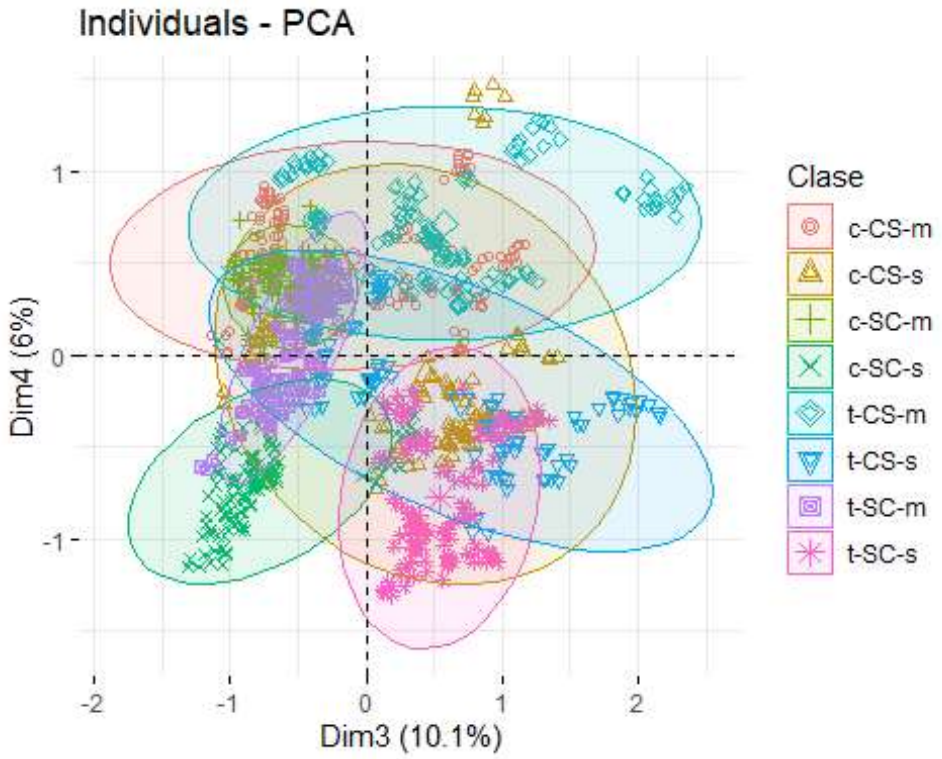
En los gráficos sin escalado se puede ver una diferencia entre la distribución de los individuos según su comportamiento. La segunda componente separa bastante bien los ratones según si están inducidos para aprender o no. Los ratones que están estimulados

para aprender tienen valores más altos en la dimensión 2 que los que no están estimulados para aprender. Con este análisis se ha visto que hay una diferencia significativa en la expresión de los ratones según si se les estimula para aprender o no.

Clase

Posteriormente, se quiere comparar todas las clases conjuntamente. Para ello se añade la variable *Clase* en habillage.

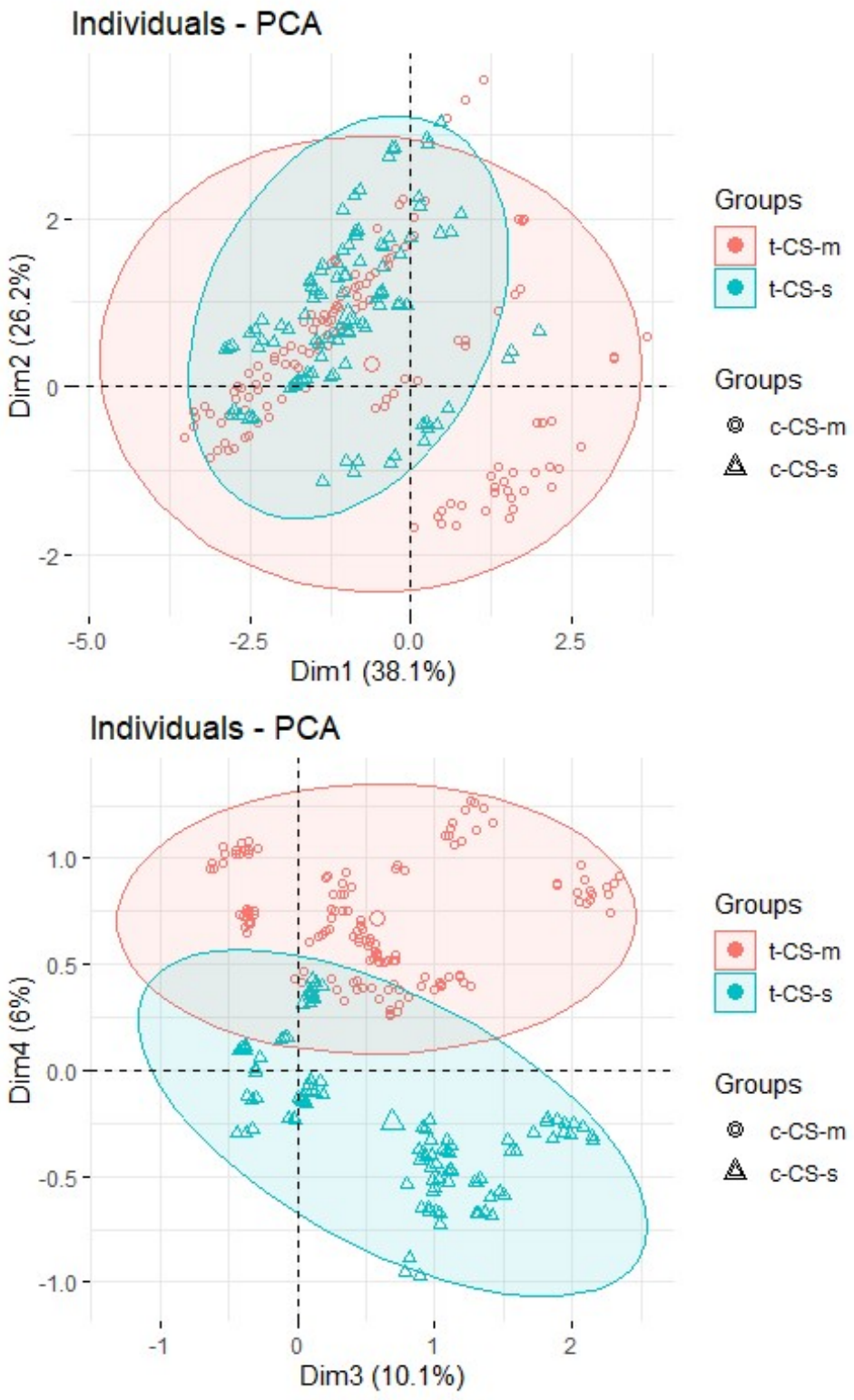




En el gráfico por clases de ratones del PCA se observa que no se separan las clases de ratones en ninguna de las dimensiones. Todas las elipses de las seis clases se solapan bastante sin ninguna distribución aparente.

Comparación entre clases t-CS-m y t-CS-s

Para ver si tiene efecto la droga memantina para devolverle a los ratones con down la capacidad de aprender en este apartado se comparan las clases de los ratones trisómicos, que están estimulados para aprender y que reciben memantina con ratones trisómicos, que están estimulados para aprender y no reciben memantina (*t-CS-m* con *t-CS-s*). Para ello se usa el campo **select.ind** poniendo solo las filas de observaciones que tienen una de esas dos clases.



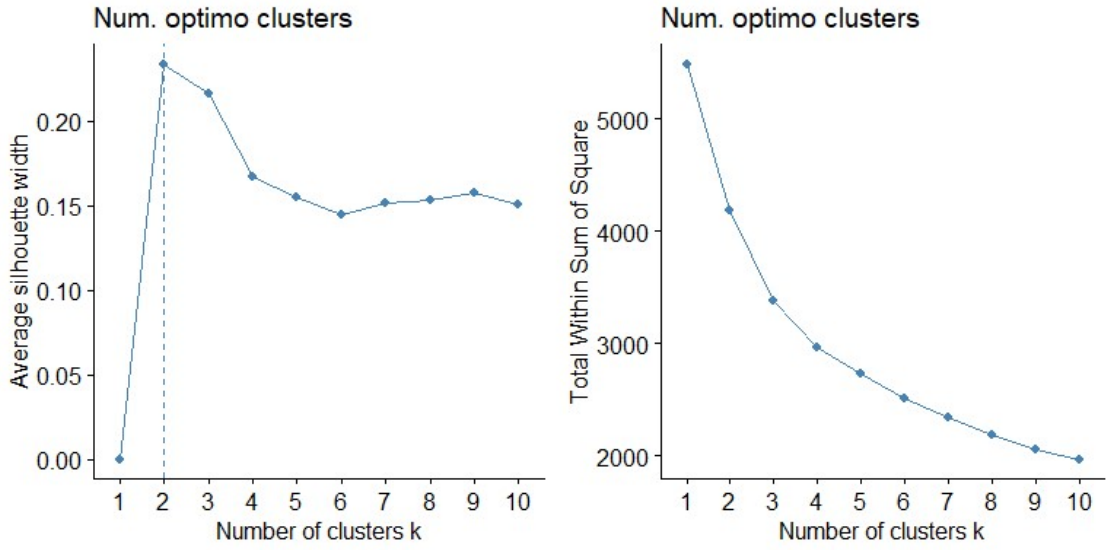
En el PCA no se ve que se separen estas dos clases, muchas de sus observaciones se solapan. Por tanto, no se logra ver un efecto significativo de la droga memantina que restaure la capacidad de aprender a los ratones con síndrome de Down.

Anexo 4: Clustering Métodos Jerárquicos

Se va a probar varios métodos jerárquicos para elegir el método más adecuado para aplicar el clustering. Se combina el análisis del coeficiente de Silhouette (maximizando) con la variabilidad intra-cluster (minimizando) para elegir el número de grupos con el que se realizará el clustering.

Método de Ward

Se obtendrá el número de clusters óptimo para el método de Ward.



Los resultados para el coeficiente Silhouette indican que el número óptimo de clusters es 2 clusters. Si observa la variabilidad intra-cluster, aún está bastante alta. Si elige el segundo óptimo, 3 clusters, la variabilidad intra-cluster ya baja bastante y, además, parece el punto en el que se crea el codo. Si coge el número de clusters como 4, baja la variabilidad intra-cluster pero también baja bastante el coeficiente Silhouette. Por tanto, se fijará el número de clusters en 3.

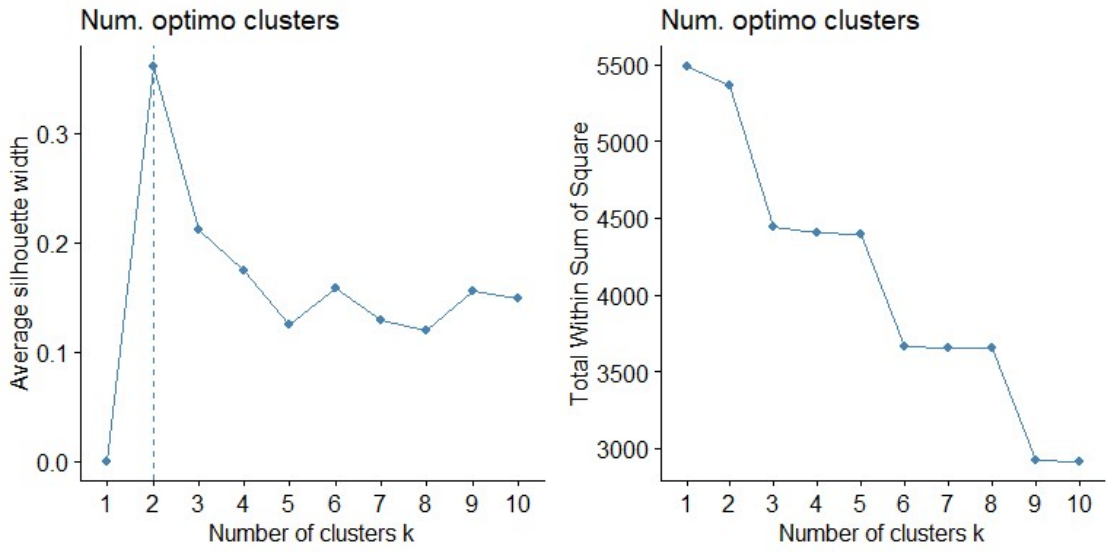
Se crea a continuación los 3 clusters con el método de Ward. No se generará el dendrograma debido a que el número de observaciones es elevado y no se podría observar claramente qué observaciones hay en cada clúster.

```
## grupos1
##   1   2   3
## 216 579 276
```

Se observan dos clústers que tienen más o menos la misma cantidad de observaciones y otro clúster que tiene más observaciones que los otros dos.

Método de la media

Ahora se estima el número óptimo de clusters para el método de la media.



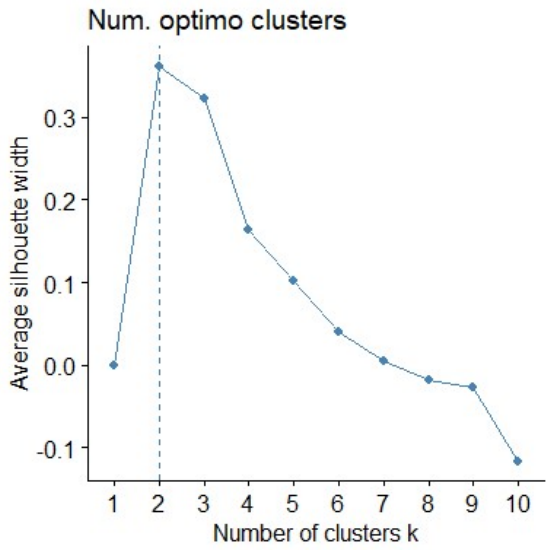
En la suma de cuadrados intra-cluster, se observan 3 codos. El primero codo es el 2 y coincide con el número óptimo del coeficiente de Silhouette. Sin embargo, la suma de cuadrados intra-cluster es demasiado elevada para 2 clusters. El segundo codo es el 4 bajando significativamente la suma de cuadrados intra-cluster y coincidiendo con el segundo número óptimo del coeficiente de Silhouette. Y, por último, el tercer codo es el 7 bajando aún más la suma de cuadrados intra-cluster y coincidiendo con el tercer número óptimo del coeficiente de Silhouette, el cual es prácticamente igual que el segundo óptimo. Sin embargo, teniendo en cuenta que hay 72 ratones, cada uno con 15 observaciones, 7 clústers es demasiado, por lo tanto se eligirá 4 clústers.

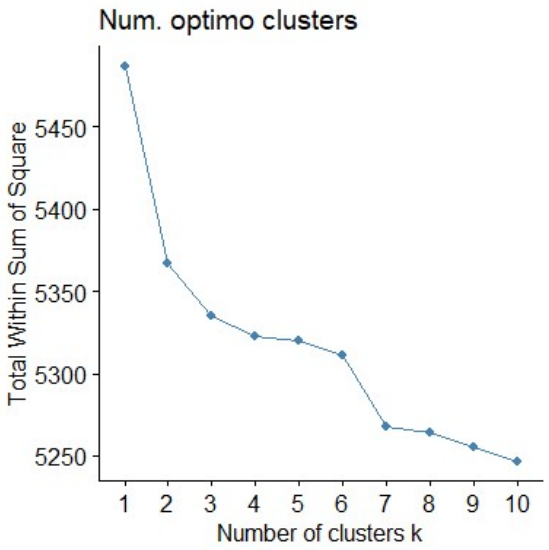
```
## grupos2
##   1   2   3   4
## 909 149   6   7
```

Sin embargo, cuando se observa el dendrograma y el número de observaciones por clúster, se observa que uno de los clústers que proporciona este método tiene una sola observación, siendo esta una cantidad demasiado reducida. Por lo tanto, se descartará este método.

Método Centroide

Ahora se estimará el número óptimo de clusters para el método centroide.





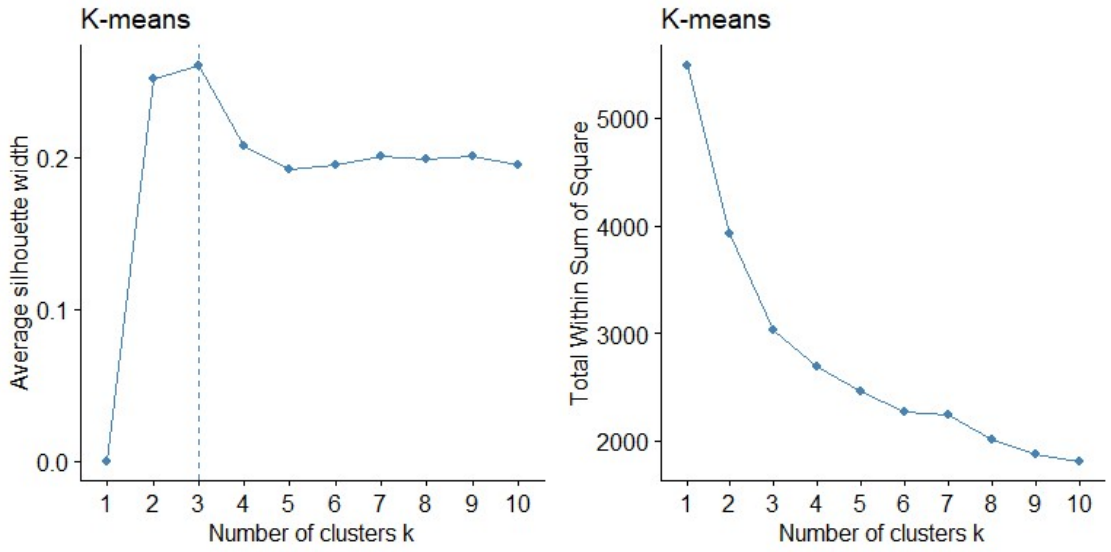
Observando el coeficiente de Silhouette, el número optimo de clústers es 2, mientras que con la suma de cuadrados intra-clúster se observa un codo con 5 clústers. Sin embargo, con 5 clústers el coeficiente de Silhouette es prácticamente 0. A partir del 2, el coeficiente de Silhouette baja significativamente y con 2 la suma de cuadrados intra-clúster baja un poco, por lo tanto se elegirán 2 clústers.

```
## grupos3
##    1    2
## 1064    7
```

El método de la media y el método centroide sucede el mismo problema, también se genera un clúster con una sola observación. En este caso, el número óptimo de clusters es 2 y los clusters son más desequilibrados que en el caso del método de la media. Por lo tanto, también se descarta este método.

Anexo 5: Clustering Métodos de Partición

K-means

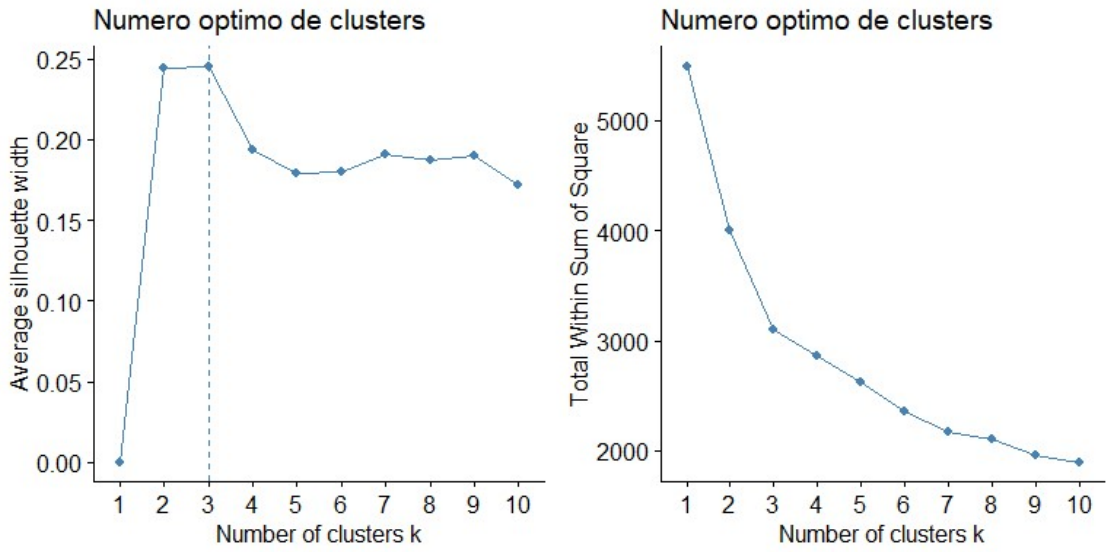


En el coeficiente de Silhouette, el número óptimo de clústers es 3. Esto coincide con el codo formado en la suma de cuadrados intra-clúster. Por lo que se eligen 3 clústers.

```
##
##   1   2   3
## 300 328 443
```

Con el método de k-means, se observa una distribución equilibrada entre los clústers formados.

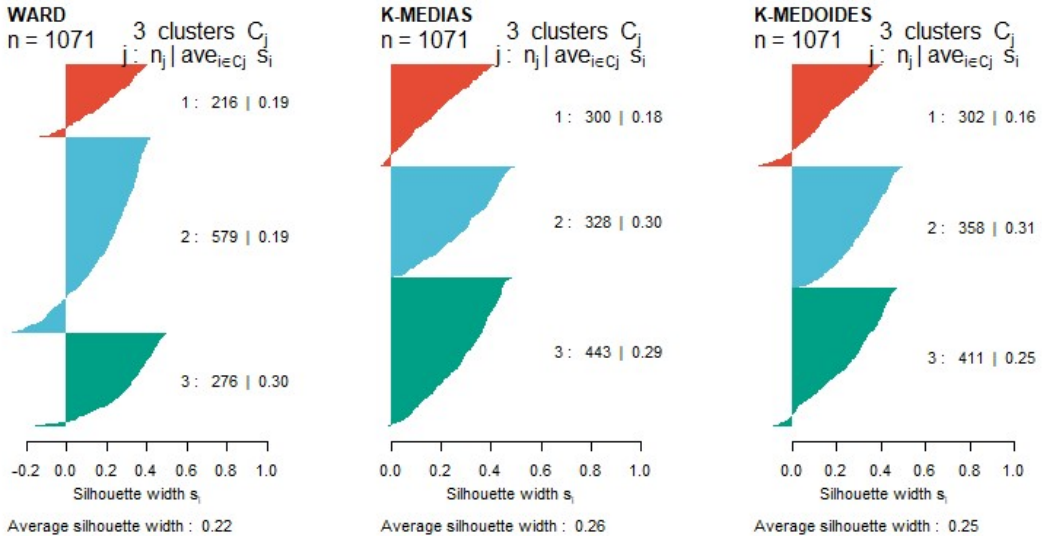
PAM ((Partitioning Around Medoids)



Se observa que con 3 clústers el coeficiente de Silhouette es muy similar que con 2 clústers, el número óptimo. Además, 3 clústers coincide con el codo más o menos visible de la suma de cuadrados intra-cluster. Por lo que se concluye con la elección de 3 clústers.

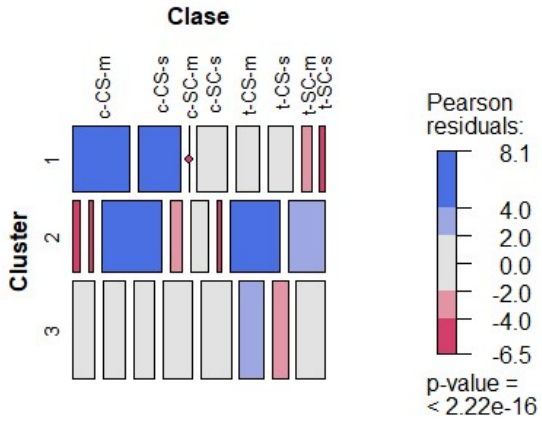
```
##
##   1   2   3
## 302 358 411
```

La aplicación del método PAM da lugar a una distribución homogénea entre los clústers generados.

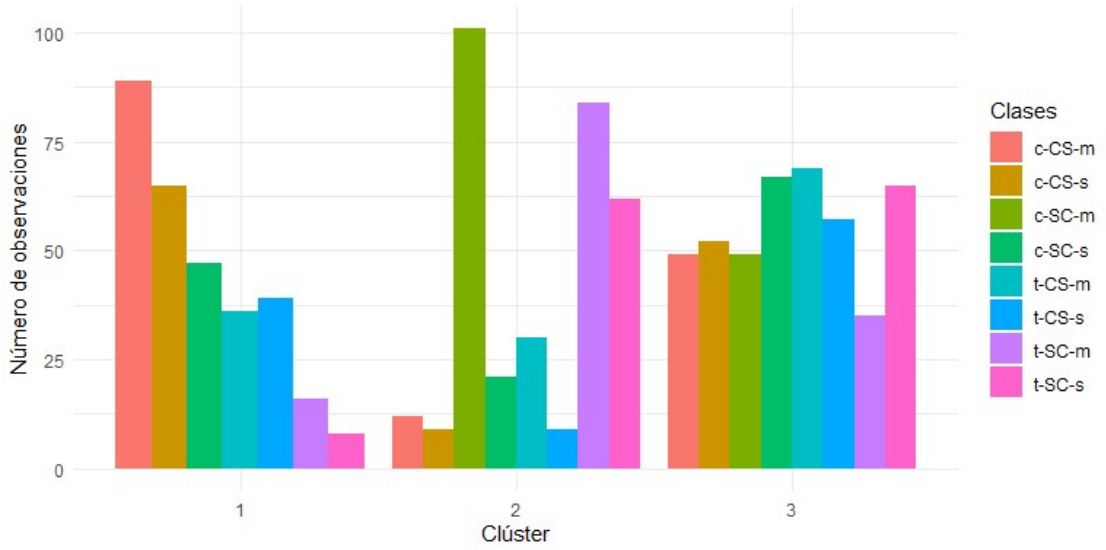


Observando los coeficientes de Silhouette de los resultados, se elige el método de k-means con 3 clústers ya es el que tiene mayor Silhouette media y menor cantidad de observaciones con coeficientes negativos, es decir, mal clasificados.

Anexo 6: Análisis con Clase Cluster



Dado que el valor p de la prueba de bondad de ajuste es menor a 0.05, se descarta la hipótesis de independencia entre las variables *Cluster* y *Clase*. En el clúster 1 no se observan grandes desviaciones, la mayoría se ajusta a lo esperado salvo los ratones trisómicos, con estimulación y sin medicina (*t-CS-s*) y los trisómicos, sin estimulación y con medicina (*t-SC-m*). Hay una ligera sobrerepresentación de los primeros y una ligera infrarrepresentación de los segundos. En el clúster 2 muestra una fuerte exceso representación de ratones neurotípicos, sin estimulación y con medicina (*c-SC-m*) y de trisómicos, sin estimulación y con medicina (*t-SC-m*). Además, hay una ligera sobrerrepresentación de ratones trisómicos, sin estimulación y sin medicina (*t-SC-s*). Por otra parte, hay infrarepresentación de 3 tipos de ratones, con estimulación (*CS*) y de los neurotípicos, sin estimulación y sin medicina (*c-SC-s*). En el clúster 3 destaca la sobrerepresentación de ratones neurotípicos, con estimulación y con medicina (*c-CS-m*) y de neurotípicos, con estimulación y sin medicina (*c-CS-s*). Por otra parte, hay infrarepresentación de 3 tipos de ratones sin estimulación, destacando entre estos tipos los neurotípicos, sin estimulación y con medicina (*c-SC-m*) ya que no hay ninguno.



En el gráfico de barras del clúster 1, se observan que los resultados coinciden más o menos con los del gráfico de mosaico. Aunque hay sobrerepresentación de ratones trisómicos con estimulación y sin medicina (*t-CS-s*) este no es el tipo del que hay más cantidad. Los ratones de los que hay mayor cantidad son los neurotípicos sin estimulación y sin la medicina y también los ratones trisómicos con estimulación y con la medicina. De estos dos tipos hay más o menos la misma cantidad en el clúster 1 y son justo lo contrario. Aunque se podría deducir que la estimulación y la medicina surgen efecto sobre ratones trisómicos ya que están en el mismo grupo que los neurotípicos a los que no se les ha aplicado nada, las otras clases también tienen una cantidad similar, por lo que no es concluyente. Además, en este clúster destaca como el tercer tipo del que hay más ratones: los trisómicos sin estimulación y sin medicina (*t-SC-s*). Tanto en el gráfico de barras del clúster 2 como en el del clúster 3, los resultados coinciden con lo anteriormente observado con los gráficos de mosaico.