Alvarez-de-la-Sierra-Daniel-PEC1

Daniel Álvarez de la Sierra

2025-03-23

Contents

URL al repositorio en GitHub **2**

Descripción del dataset **2**

Análisis de los datos **3**

Descripción de las muestras 3

Aánalisis descriptivo de los datos experimentales 4

Análisis exploratorio multivariante 6

URL al repositorio en GitHub

Los ficheros asociados a esta PEC se encuentran en el siguiente repositorio PEC1 Daniel Álvarez de la Sierra.

Descripción del dataset

Se selecciona uno de los datasets del repositorio proporcionado en las instrucciones de la PEC1. En concreto el conjunto de datos utilizado por el CIMCB para su tutorial de análisis de datos ómicos. Estos datos, accesibles también en el repositorio Metabolomics Workbench bajo el identificador PR000699, contienen la concentración de distintos metabolitos en la orina de individuos con cáncer gástrico, enfermedad gástrica benigna, y controles sanos.

En primer lugar visualizamos la estructura general de las dos hojas de excel que hemos cargado:

Table 1: Encabezado de la hoja "data" del excel.

Idx	Day of Expt	Sample_Type	QC	Batch	Order	Sample_id	M1	M2	M3
1	2014-12-08	QC	1	1	1	sample_1	90.1	491.6	202.9
2	2014-12-08	Sample	0	1	2	sample_2	43.0	525.7	130.2
3	2014-12-08	Sample	0	1	3	sample_3	214.3	10703.2	104.7
4	2014-12-08	Sample	0	1	4	sample_4	31.6	59.7	86.4
5	2014-12-08	Sample	0	1	5	sample_5	81.9	258.7	315.1
6	2014-12-08	Sample	0	1	6	sample_6	196.9	128.2	862.5

Table 2: Encabezado de la hoja "peak" del excel.

Idx	Name	Label
1	M1	1_3-Dimethylurate
2	M2	1_6-Anhydro- -D-glucose
3	M3	1_7-Dimethylxanthine
4	M4	1-Methylnicotinamide
5	M5	2-Aminoadipate
6	M6	2-Aminobutyrate

La primera contiene el conjunto de datos del estudio, la primera columna corresponde a los índices de las filas, desde la segunda (Day of Expt) a la séptima (Sample_id) se encuentran los metadatos asociados a cada muestra, y desde la octava (M1) hasta la última tenemos las concentraciones correspondientes a todos los metabolitos analizados para cada una de las muestras. La segunda tabla cargada contiene la equivalencia entre los codigos asignados a cada metabolito (M1:M129) y sus nombres reales.

A continuación, utilizamos ambos dataframes para crear nuestro objeto de tipo `SummarizedExperiment`:

```
## class: SummarizedExperiment
## dim: 129 140
## metadata(3): ID_proyecto Publicacion_asociada Autor_principal
## assays(1): metabolitos
## rownames(129): 1_3-Dimethylurate 1_6-Anhydro- -D-glucose ...
##   pi_Methylhistidine tau_Methylhistidine
## rowData names(0):
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(6): Day of Expt Sample_Type ... Order Sample_id
```

La principal diferencia entre un objeto de tipo `ExpressionSet` y uno de tipo `SummarizedExperiment` es que este último puede considerarse una extensión del primero, pero es más flexible en cuanto a la accesibilidad a la información de sus filas.

Análisis de los datos

Descripción de las muestras

```
##   Day.of.Expt         Sample_Type  QC      Batch      Order
##  Min.   :2014-12-08   QC    : 17   0:123   1:35   Min.   :  1.00
##  1st Qu.:2014-12-09   Sample:123   1: 17   2:35   1st Qu.: 35.75
##  Median :2014-12-11                        3:34   Median : 70.50
##  Mean   :2014-12-12                        4:36   Mean   : 70.50
##  3rd Qu.:2014-12-18                               3rd Qu.:105.25
##  Max.   :2014-12-18                               Max.   :140.00
##   Sample_id
##  Length:140
##  Class :character
##  Mode  :character
##
##
##
```

El dataset contiene información sobre 140 muestras distintas.

Los metadatos de las muestras contienen información sobre la fecha en la que se realizó el experimento, el tipo de muestra (control o problema), el batch de procesamiento, y la codificación de la muestra.

El tipo de muestra esta codificado en dos columnas distintas, "Sample_Type" y "QC", por lo que en primer lugar conviene comprobar que ambas son equivalentes y no existen incongruencias.

```
##
##            0   1
##   QC       0  17
##   Sample 123   0
```

La coincidencia es pefecta, por lo que ambas contienen la misma información.

De la misma manera, es lógico pensar que el Batch hace referencia a la fecha del procesamiento experimental de la muestra, por lo que podría ser que columna "Day.of.Expt" y "Batch" sean equivalentes:

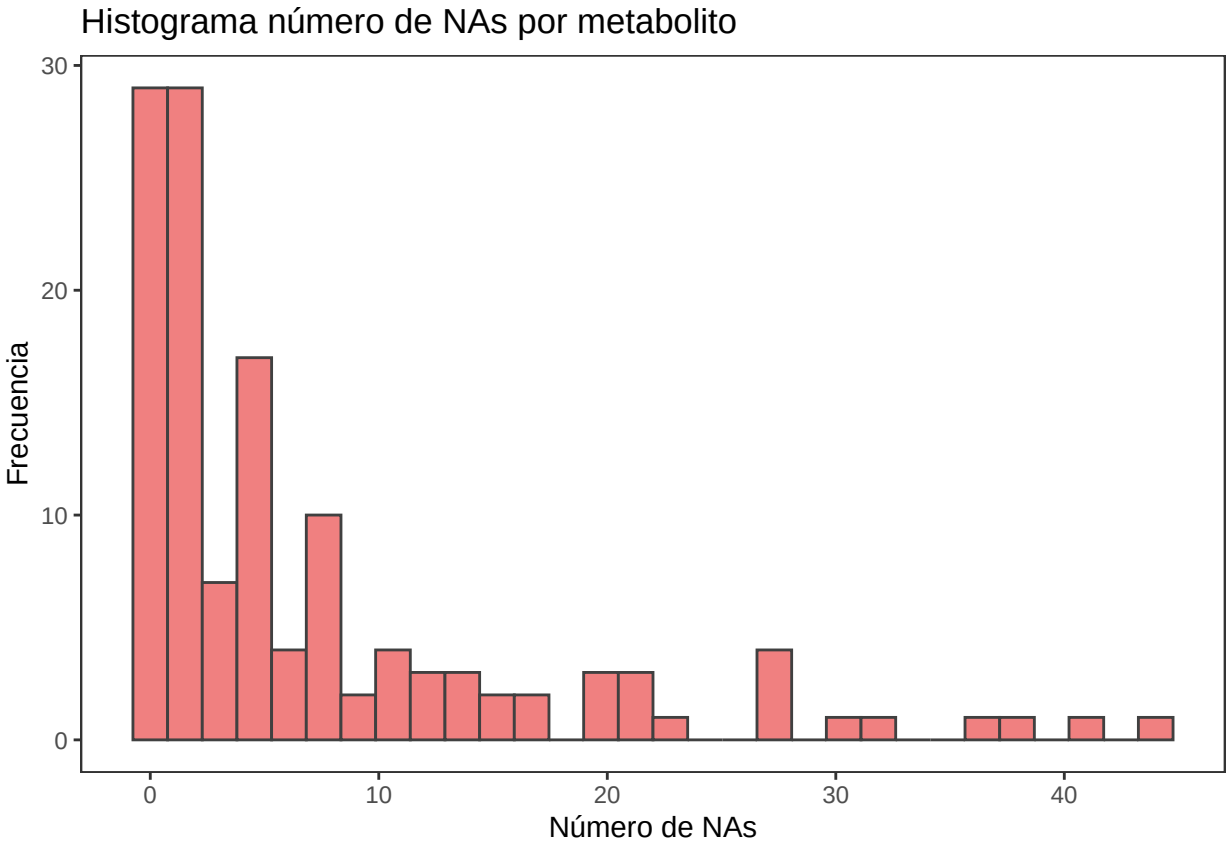
```
##
##               1  2  3  4
##   2014-12-08 35  0  0  0
##   2014-12-10  0 35  0  0
##   2014-12-12  0  0 34  0
##   2014-12-18  0  0  0 36
```

De nuevo, hay una correspondencia perfecta entre ambas, por lo que cada batch hace referencia a un único momento de realización de los experimentos, y estas dos columnas son equivalentes.

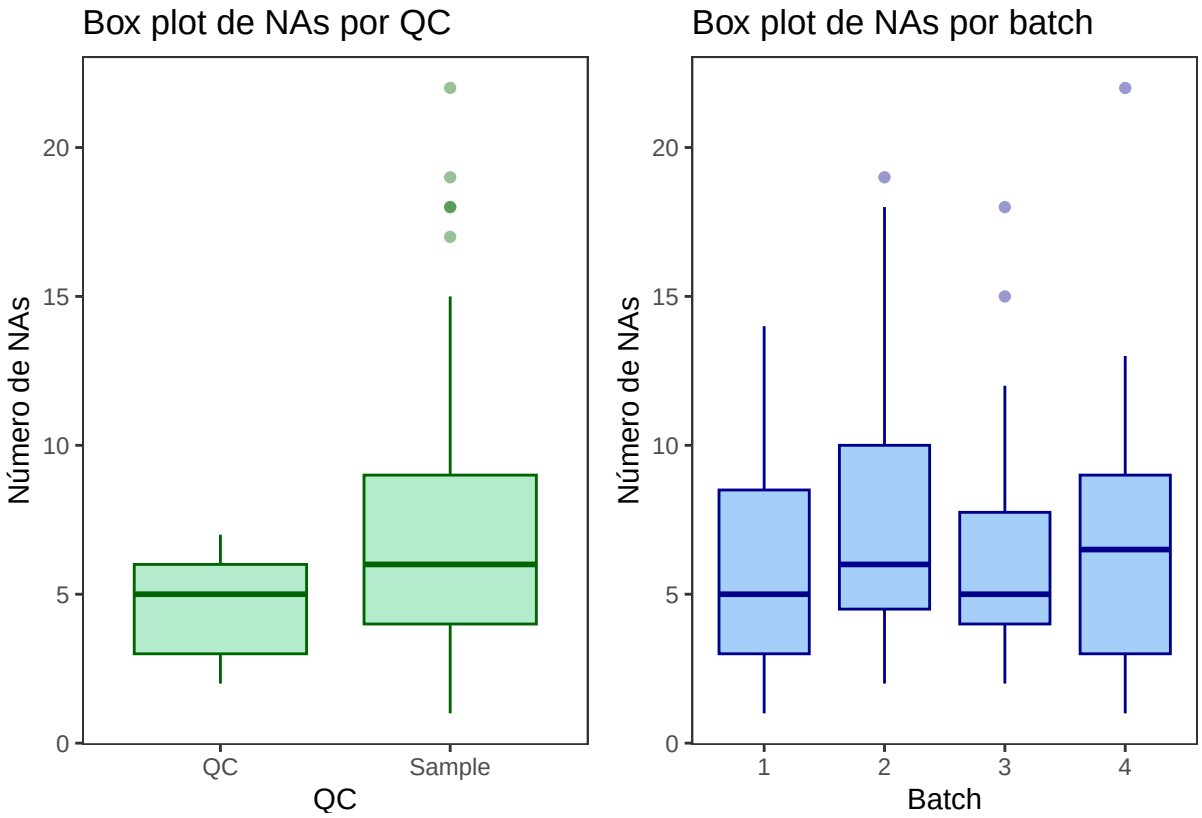
Aánalisis descriptivo de los datos experimentales

En este experimento se ha cuantificado en la orina de pacientes y controles la concentración de 129 analitos diferentes.

Al crear el objeto SummarizedExperiment, el ensayo "metabolitos" fue añadido en forma de matriz, por lo que sabemos que todas las variables incluidas son de tipo numérico. En primer lugar comprobaremos si existen datos ausentes en la matriz de datos:



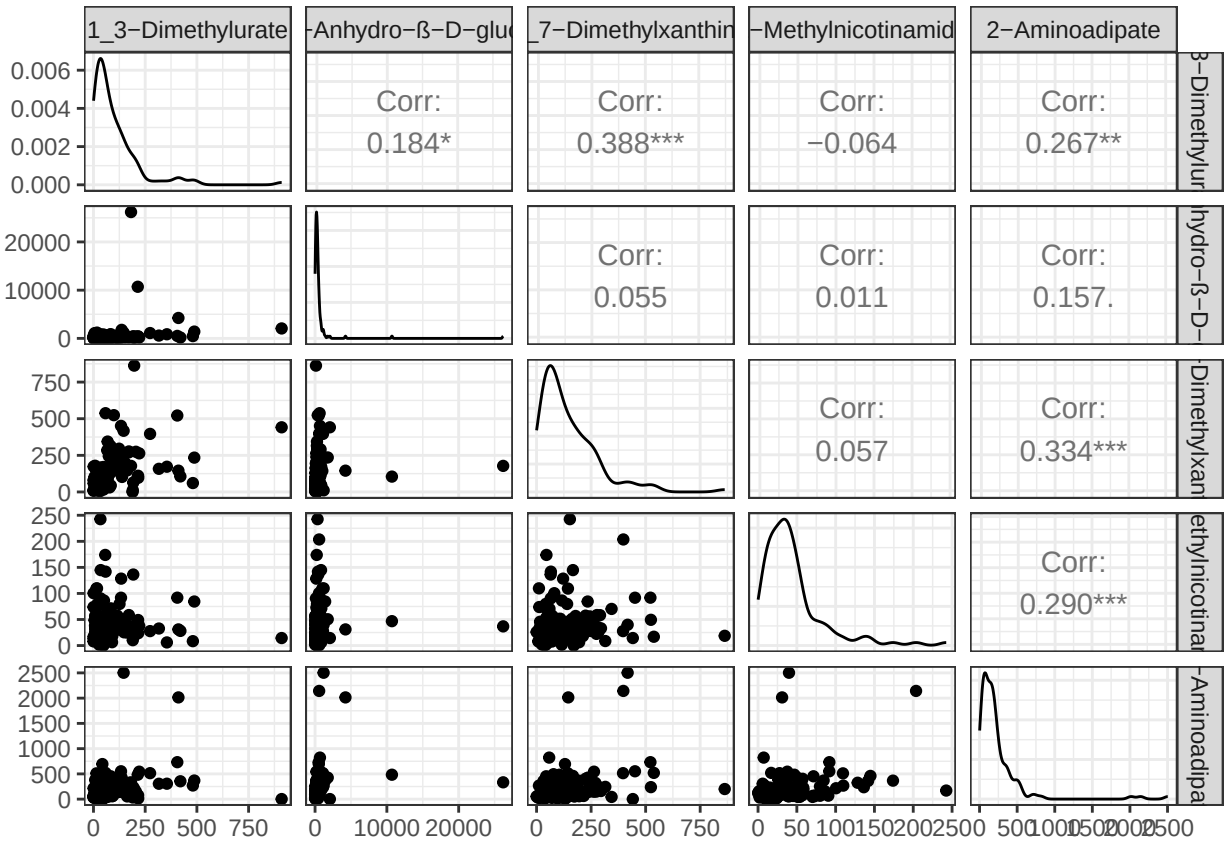
Como vemos, hay un gran número de valores ausentes. Lo primero sería realizar un análisis preliminar para ver si el número de NAs no depende del batch de las muestras o del tipo de muestra.



A primera vista no parece que haya una diferencia significativa en cuanto al número de NAs según el tipo de muestra o el batch de análisis, lo cual podría reflejar un problema a nivel experimental.

El siguiente paso sería realizar una descripción básica de los datos. Podríamos obtener los estadísticos básicos univariantes (media aritmética, mediana, SD…). Sin embargo, con un número tan elevado de variables esto no tiene tanta utilidad debido a lo laborioso que sería su interpretación.

Ocurre lo mismo en el caso del análisis bivariante. El siguiente ejemplo nos permitiría realizar un análisis visual de la distribución de las variables y de la correlación entre ellas en experimentos con un número más reducido de variables:



Análisis exploratorio multivariante

En casos con un número tan elevado de variables lo más recomendable es realizar un análisis multivariante.

Preprocesamiento

En primer lugar, realizamos la imputación de los datos ausentes. Para ello utilizamos la función PomaImpute del paquete POMA.

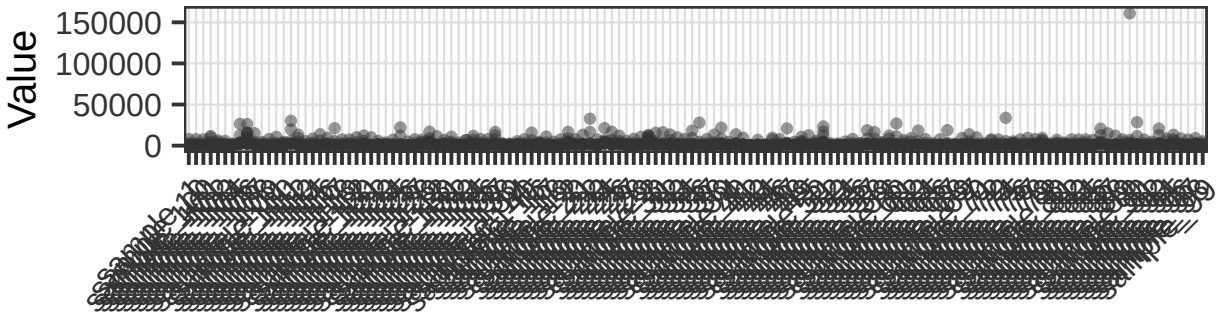
```
## 0 features removed.
```

```
## class: SummarizedExperiment
## dim: 129 140
## metadata(0):
## assays(1): ''
## rownames(129): 1_3-Dimethylurate 1_6-Anhydro- -D-glucose ...
##   pi_Methylhistidine tau_Methylhistidine
## rowData names(0):
## colnames(140): sample_1 sample_2 ... sample_139 sample_140
## colData names(6): Day of Expt Sample_Type ... Order Sample_id
```

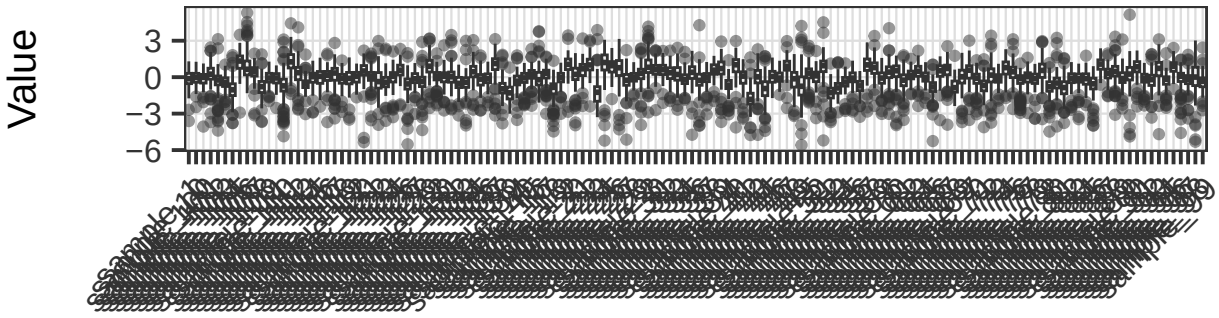
A continuación, realizamos la normalización de los datos para evitar que cuyos valores se encuentro en distintos rangos puedan afectar a los análisis posteriores.

Observamos el efecto de la normalización:

Datos no normalizados



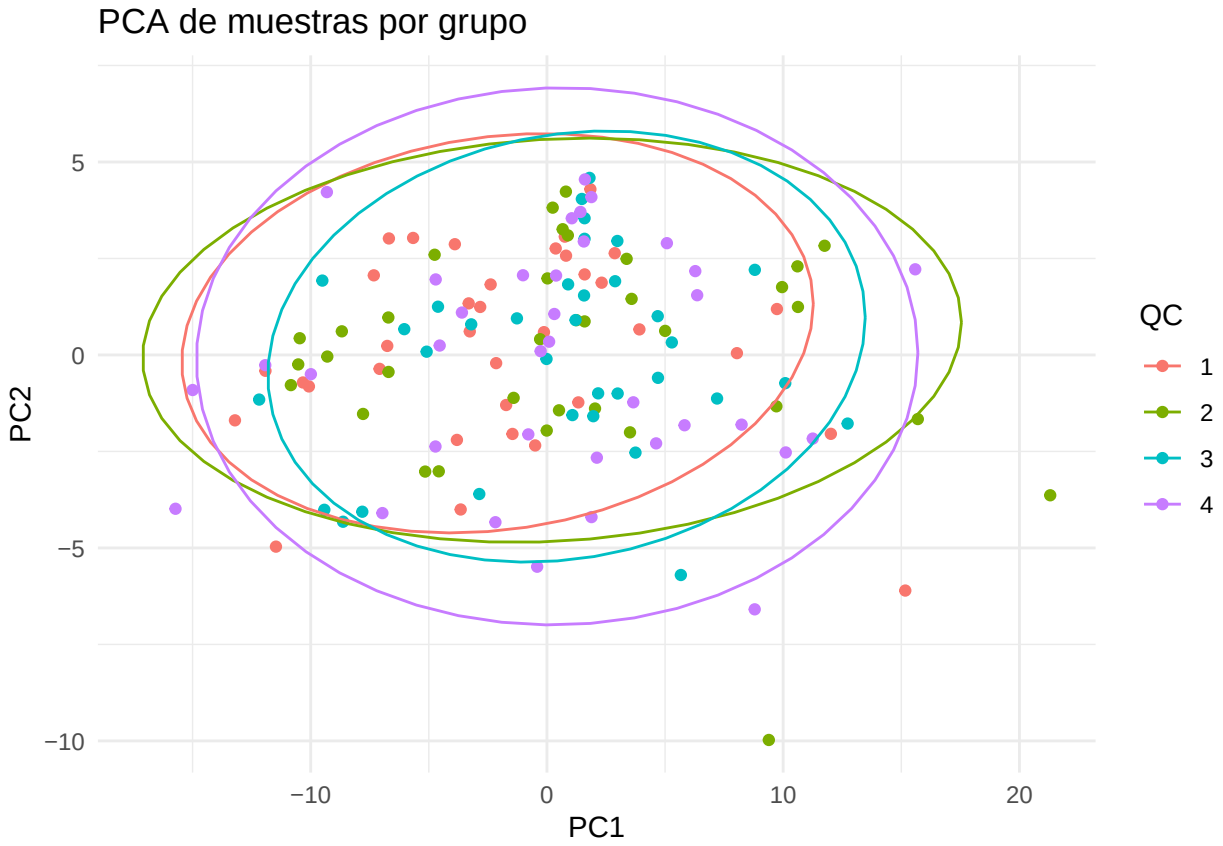
Datos normalizados



Análisis multivariante

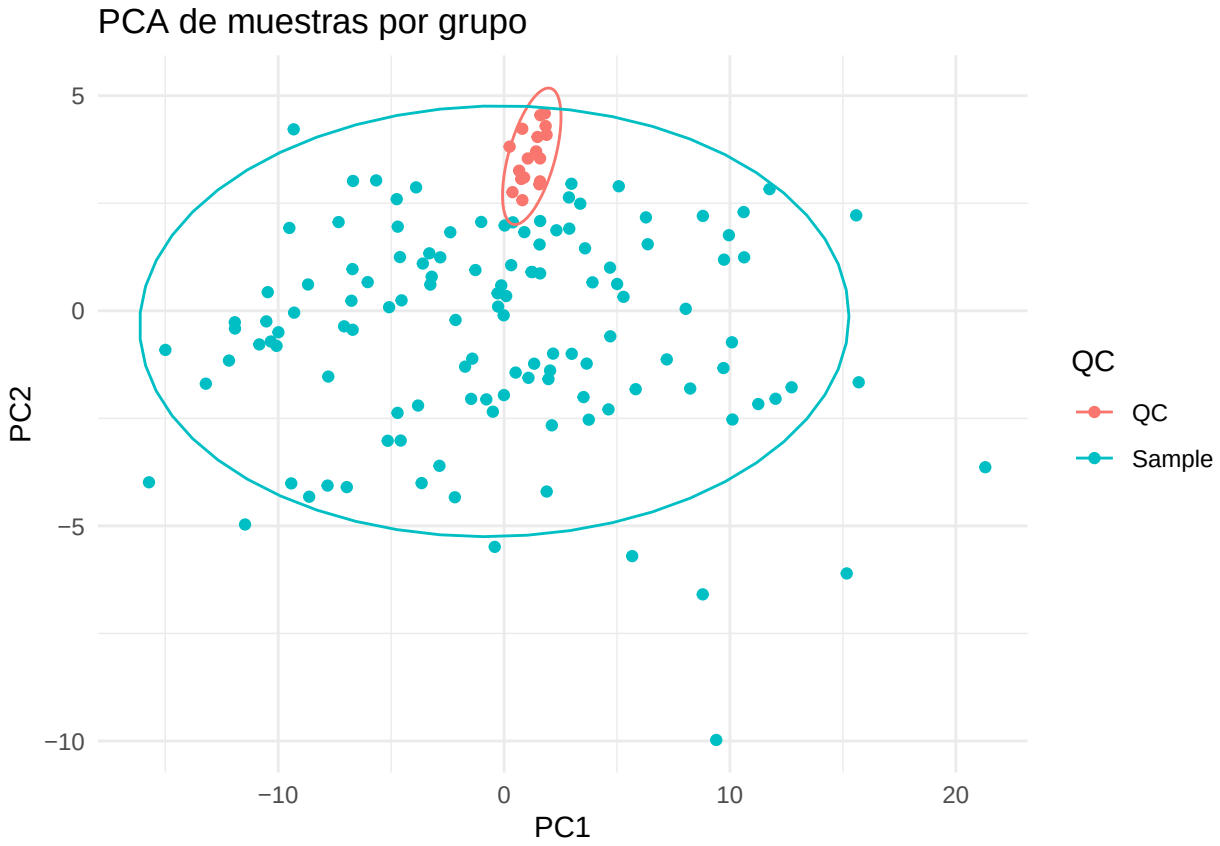
Para este tipo de datos uno de los mejores análisis exploratorios que podría realizar es un PCA. Con ello conseguimos reducir el número de dimensiones que vamos a utilizar creando componentes principales que sean combinaciones lineales de las variables existentes de tal forma que capturen la mayor parte de la variabilidad posible.

En primer lugar realizamos el análisis de PCA y visualizamos los resultados utilizando como outcome el batch:



Parece que las muestras no se agrupan en función del momento del procesamiento, por lo que podemos descartar que haya un efecto asociado al batch.

A continuación visualizamos el mismo gráfico agrupando las muestras según sean de controles o de pacientes problema:



Observamos que las muestras de los controles son muy similares entre ellas y se agrupan juntas. Aunque las muestras problema muestran mucha más variabilidad pueden separarse de forma clara de los controles.

El análisis preliminar sugiere que los pacientes con cáncer gástrico tienen un perfil de secreción de metabolitos en orina distinto al de los controles sanos. Sería necesario ampliar el estudio para encontrar que metabolitos contribuyen en mayor medida a esta diferencia, y averiguar si podría usarse una combinación de ellos como marcador de cáncer gástrico para mejorar el diagnóstico de esta patología.