PEC1

Ariel Ernesto Cariaga Martínez

Contenido

Resumen ejecutivo

Este estudio examina datos de metabolómica, obtenidos en pacientes sometidos a cirugía bariátrica, mediante análisis exploratorio y reducción de dimensionalidad. Utilizando un contenedor `SummarizedExperiment`, se estructuraron los datos metabolómicos, antropométricos y clínicos, aplicándose imputación kNN para gestionar valores faltantes. Las diversas visualizaciones sugieren ciertos patrones entre los datos provenientes de las muestras analizadas, incluyendo los resultados de un Análisis de Componentes Principales (PCA), un clúster jerárquico y un mapa de distancias, sugiriendo una posible relación con la intervención quirúrgica. Sin embargo, los primeros componentes principales explican solo una pequeña fracción de la varianza total, indicando alta complejidad en los datos. El estudio presenta limitaciones pero este enfoque inicial sienta las bases para estudios más complejos que integren técnicas avanzadas multidimensionales y de aprendizaje automático aplicados a esta *ómica*.

Objetivos del estudio.

Esta PEC completa la introducción a las ómicas mediante un ejercicio de repaso y ampliación que nos permitirá trabajar con algunas de las herramientas de este curso, en concreto, Bioconductor y la exploración multivariante de datos.Para llevar a cabo óptimamente esta primera parte, habrá que estar familiarizado con:

- las tecnologías ómicas,
- las herramientas para trabajar con ellas,
  - Bioconductor y
  - *github*,
- los contenedores de datos ómicos, como los `expressionSets`,
- y con las herramientas de exploración de datos, introducidas en la tercera actividad.

En concreto, los objetivos buscados son:

1. Seleccionar un dataset de metabolómica a obtener de este repositorio de github: https://github.com/nutrimetabolomics/metaboData/
   - O usar algún dataset del repositorio de metabolomicsWorkbench
2. Una vez descargados los datos, crear un contenedor del tipo `SummarizedExperiment` que contenga los datos y los metadatos (información acerca del dataset, las filas y las columnas).
3. Llevar a cabo una exploración del dataset que proporcione una visión general del mismo.
4. Elaborar un informe que describa el proceso realizado, incluyendo la descarga de los datos, la creación del contenedor, la exploración de los datos y la reposición de los datos en *github*. El nombre del repositorio tiene que ser el siguiente: Apellido1-Apellido2-Nombre-PEC1.
5. Crear un repositorio de *github* que contenga o
   - El informe.
   - El objeto contenedor con los datos y los metadatos en formato binario (.Rda).
   - El código R para la exploración de los datos
   - Los datos en formato texto.
   - Los metadatos acerca del dataset en un archivo Rmarkdown.

   La dirección (url) del repositorio deberá estar incluida en la última sección del informe de forma clara.

Materiales y métodos.

Origen y naturaleza de los datos.

Los diversos datasets para la generación del objeto `SummarizedExperiment` fueron tomados del trabajo publicado por Palau-Rodriguez et al. Estos datos contienen tanto datos en crudo, como metadatos. Más concretamente los datos fueron obtenidos de 39 pacientes que fueron sometidos a cirugías bariátricas (Hospital Virgen de la Victoria, Málaga) y en cuyos sueros se analizaron diversos metabolitos, con adquisiciones de datos a tiempos diferentes (1, 3 y 6 meses) tras la cirugía. Los datos recogidos incluyeron variables cualitativas y sociodemográficas (tales como sexo y edad y el tipo de cirugía bariátrica realizada). Por otra parte, también se incluyen valores antropométricos cuantitativos tales como peso, índice de masa corporal (BMI), perímetro de cintura, ratio cintura-cadera y una estratificación (grupo) según el grado del síndrome metabólico desarrollado por cada paciente. Otros datos clínicos cuantitativos también incluidos son los niveles de hemoglobina glicosilada, glucosa sérica e insulina, tensión arterial y datos de los perfiles lipídicos (colesterol total, colesterol VLDL, LDL y HDL). La toma de los datos cuantitativos para el análisis metabolómico se realizó por cromatografía líquida unida a espectrometría de masas.

Herramientas utilizadas.

En el presente estudio, los datos fueron procesados utilizando el software R (versión 4.4.1) y Bioconductor (version 3.20) (en particular con el paquete `summaryzedExperiments`). Siguiendo la decisión inicial del trabajo publicado, los valores faltantes fueron imputados con una aproximación de vecinos más cercanos (kNN) utilizando K=10. Para el caso del análisis exploratorio de los datos se usaron paquetes base de R, `tidyverse` (para la parte de manipulación y gráficos), `factoExtra` (para ciertas visualizaciones) y `DMwR2` (para las imputaciones de datos faltantes), entre otros.

Resultados

Generación del objeto `summaryzedExperiment`.

El archivo global de datos (DataValues_S013.csv) contenía 5 filas con metadatos de los pacientes. En concreto 5 filas que consistían en el ID del paciente, el sexo, la edad, el tipo de cirugía y la estratificación por grupo. Por lo tanto, estas 5 filas pasaron a formar parte del parámetro `colData` .

Por otra parte, el archivo (DataInfo_S013.csv) contenía información sobre los distintos metabolitos medidos. Tras eliminar las filas que contenían metadatos (y que forman parte del parámetro `colData`) el número de dimensiones y la disposición de los elementos fue la correcta

para la generación del objeto. En consecuencia, se generó el objeto `summaryzedExperiment` con:

- DataValues (sin las 5 filas comentadas y transpuesto) fue parte del parámetro `assays`.
- Las 5 filas indicadas fueron el parámetro `colData` .
- DataInfo_S013 fue parte del parámetro `rowData` .

Por lo tanto, la instrucción para la creación del objeto de interés para este estudio fue la siguiente:

```
se <- SummarizedExperiment(assays = list(counts = as.matrix(assay_1)), rowData
= rowdata, colData = coldata)
```

Análisis exploratorio.

Análisis descriptivo.

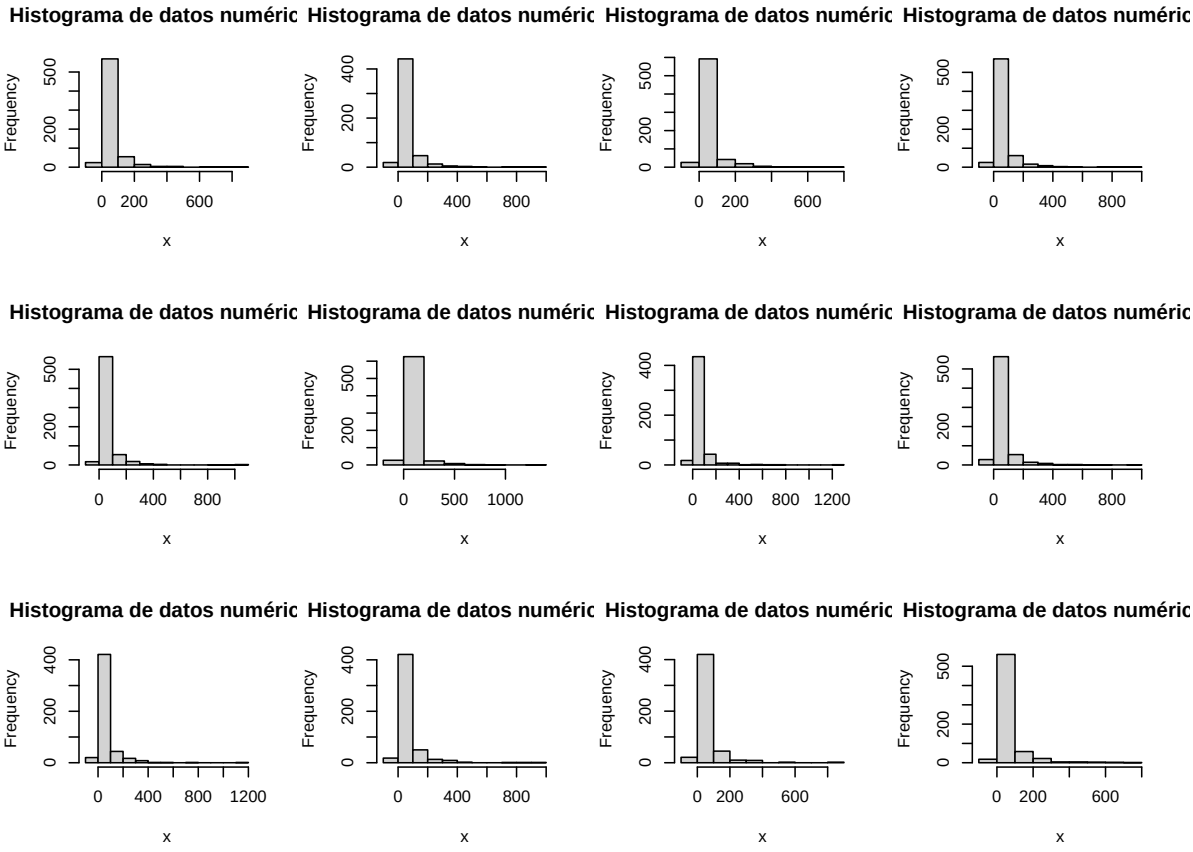
Se realizó un análisis descriptivo cuyos principales resultados resumidos fueron los siguientes:

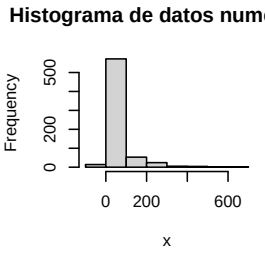
vars	n	mean	sd	min	max
1	674	37.43852	88.22440	-99	876
2	532	42.23858	94.61677	-9	951
3	689	34.42935	78.17078	-99	723
4	689	41.04681	93.74628	-99	987
5	668	39.80874	100.82470	-99	1070
6	688	42.68742	102.35981	-99	1320
7	516	41.35016	98.67353	-9	1250
8	675	40.03973	89.98608	-99	926
9	514	43.99287	96.98236	-99	1150
10	516	43.52808	95.71700	-99	954
11	509	40.57426	90.99756	-99	892
12	673	41.45322	89.30448	-99	746
13	675	39.31407	79.69230	-99	654
14	677	46.64704	108.75944	-99	1050
15	680	50.60993	114.17067	-9	1350
16	682	42.81330	91.40906	-9	833
17	683	46.11216	110.48559	-99	1040
18	685	49.26624	118.92715	-9	1560
19	680	42.58932	89.90755	-9	863
20	682	45.52494	106.62775	-99	1240
21	686	35.32131	77.81654	-9	677
22	507	48.87436	117.37097	-99	1230
23	668	36.02079	85.61359	-99	965

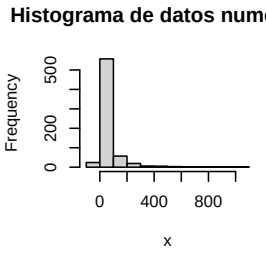
24	682	51.30868	119.08357	-99	1280
25	686	41.31902	93.76710	-99	935
26	659	35.08162	75.49721	-99	871
27	666	40.47683	94.36902	-99	907
28	356	38.73385	81.57957	-99	713
29	534	40.47637	88.32663	-99	846
30	680	57.43178	129.87929	-99	1240
31	509	52.66577	121.97636	-99	1140
32	663	39.46436	89.80033	-9	921
33	668	37.18463	83.24780	-9	903
34	541	46.91373	112.00372	-99	1220
35	668	42.23227	97.37140	-9	990
36	362	38.36775	71.79838	-9	562
37	333	37.62628	75.70017	-99	593
38	503	36.79207	77.30304	-9	680
39	362	44.71657	86.48241	-99	552

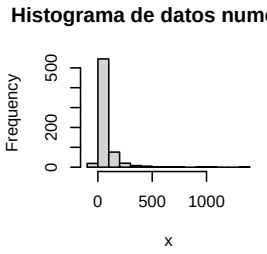
Análisis gráfico básico.

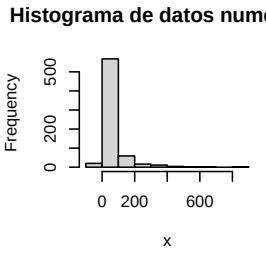
Se llevó a cabo un primer análisis gráfico de las distintas variables para ver sus características principales.

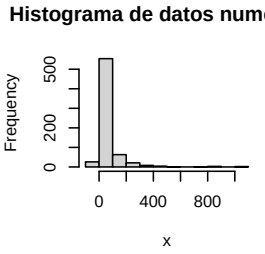

Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x
Histograma de datos numéric
Frequency
x

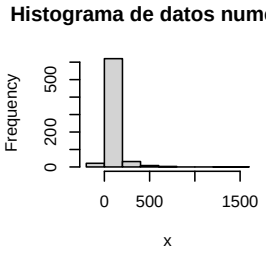

Histograma de datos numéric
Frequency
0
200
500
0
200
600
x

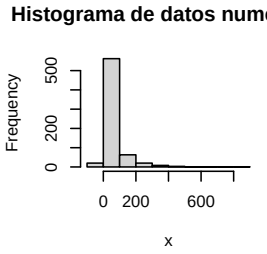

Histograma de datos numéric
Frequency
0
200
500
0
400
800
x

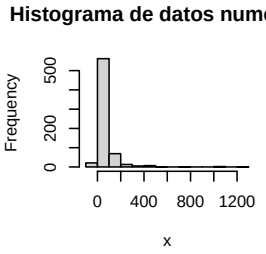

Histograma de datos numéric
Frequency
0
200
500
0
500
1000
x

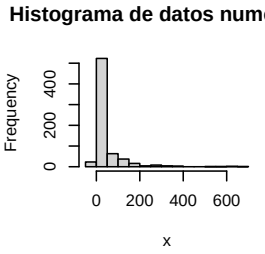

Histograma de datos numéric
Frequency
0
200
500
0
200
600
x

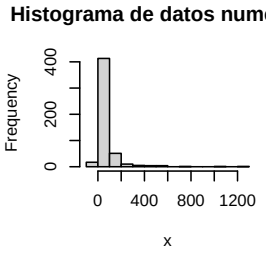

Histograma de datos numéric
Frequency
0
200
500
0
400
800
x

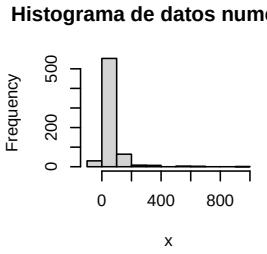

Histograma de datos numéric
Frequency
0
200
500
0
500
1500
x

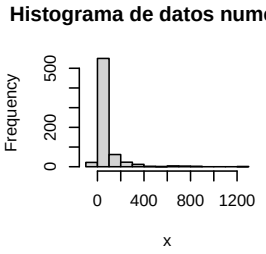

Histograma de datos numéric
Frequency
0
200
500
0
200
600
x

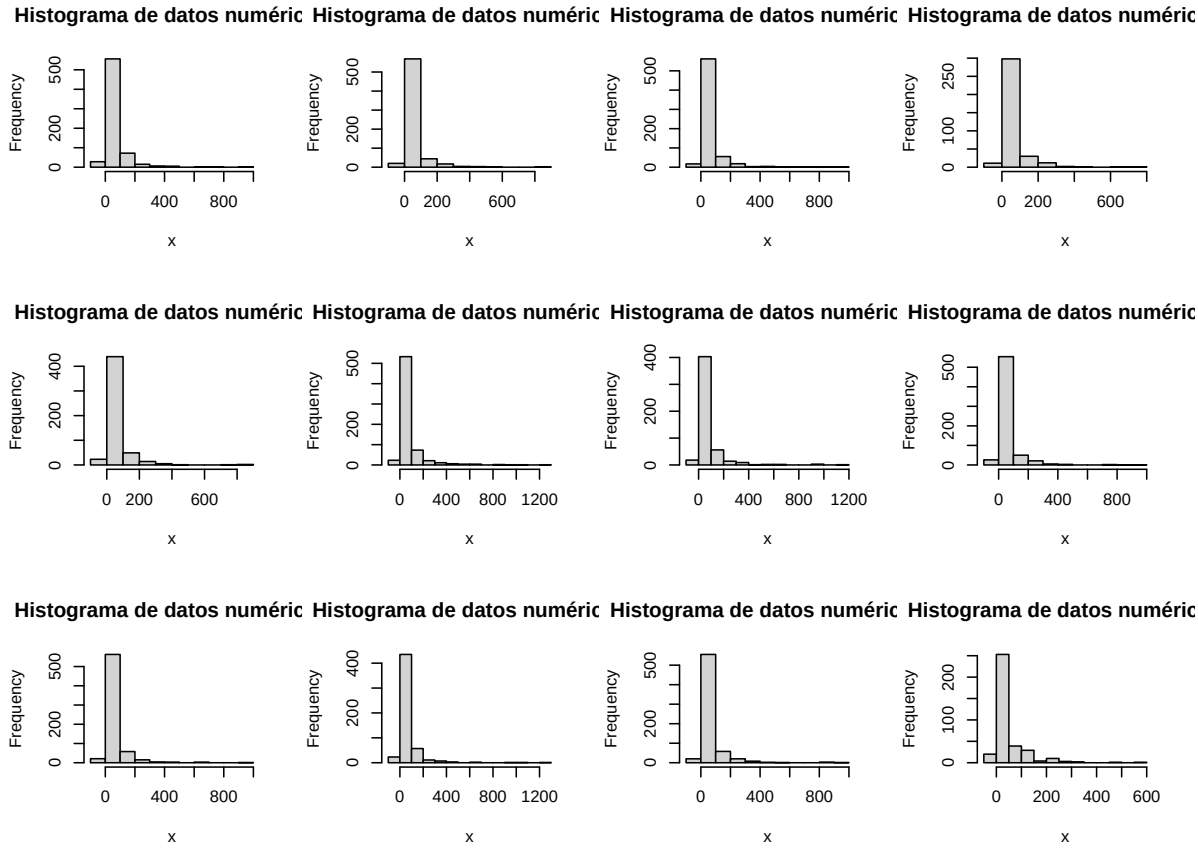

Histograma de datos numéric
Frequency
0
200
500
0
400
800
1200
x


Histograma de datos numéric
Frequency
0
200
400
0
200
400
600
x

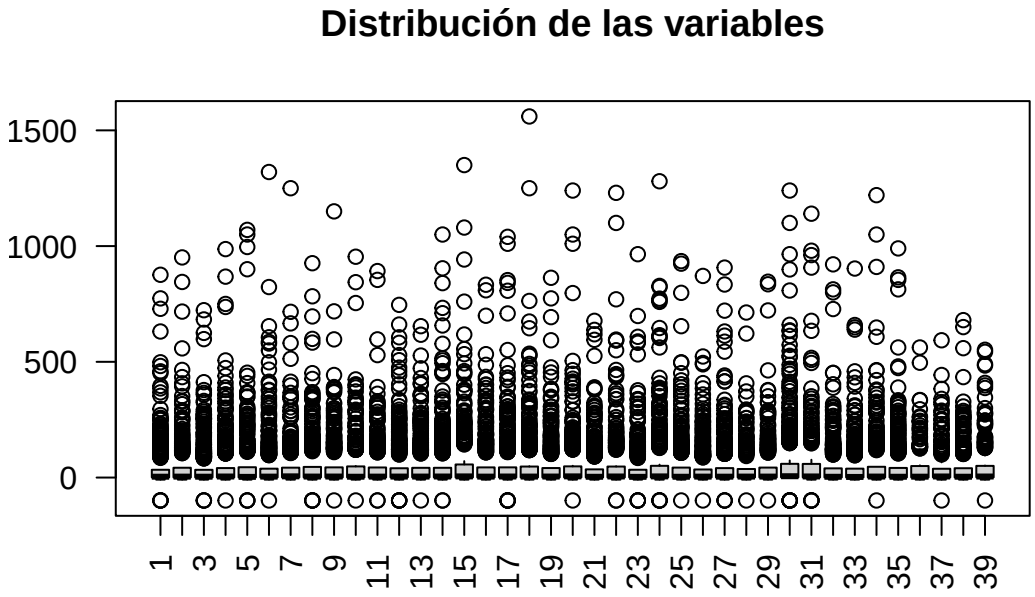

Histograma de datos numéric
Frequency
0
200
400
0
400
800
1200
x


Histograma de datos numéric
Frequency
0
200
500
0
400
800
x

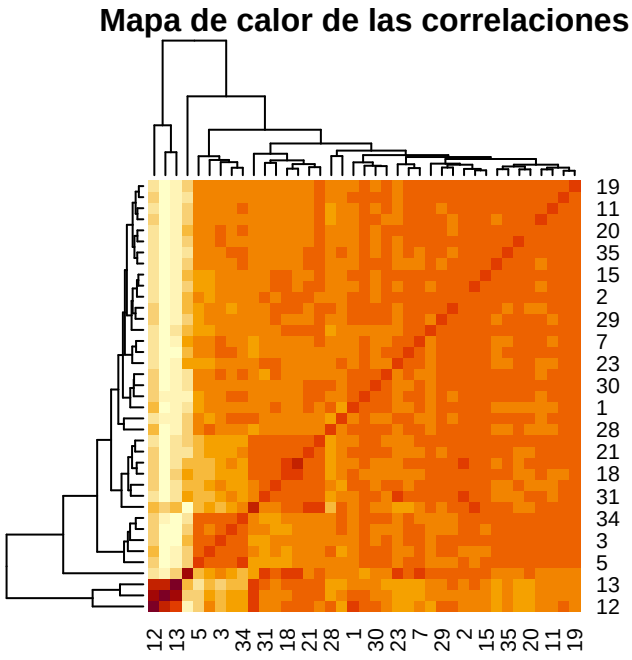

Histograma de datos numéric
Frequency
0
200
500
0
400
800
1200
x



En la búsqueda de una mejor comprensión de las características de los datos, se realizaron visualizaciones de tipo *boxplot* que, además permitieron una mejor observación de posibles valores *outliers*.

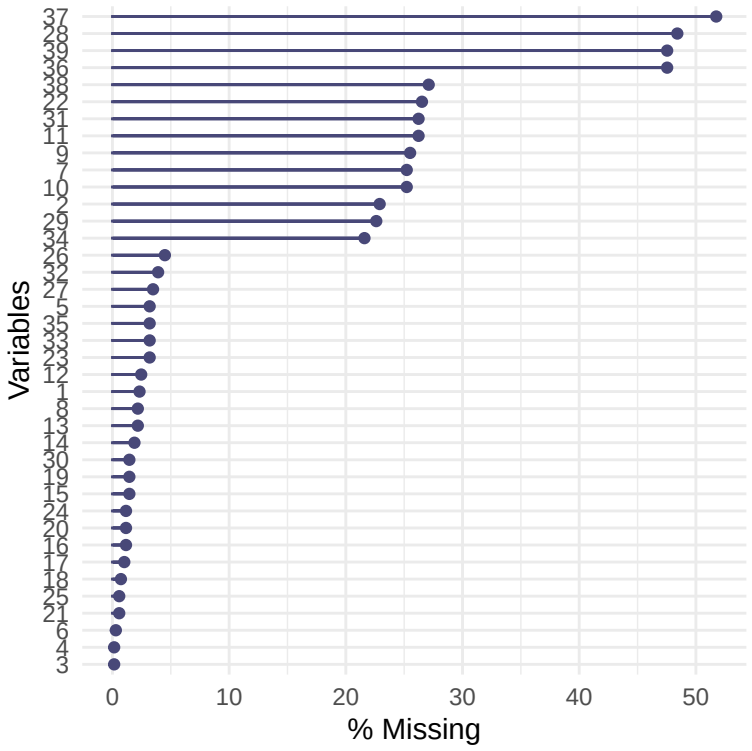


También se realizó un análisis exploratorio de las correlaciones.

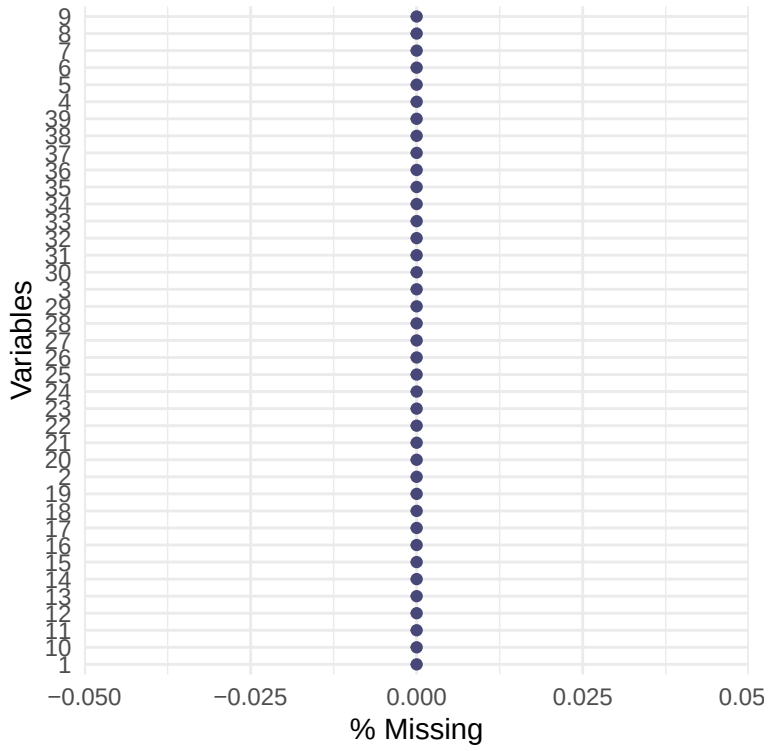


Preprocesado de los datos.

Un posible problema de cara a los siguientes pasos fue que la matriz de datos presentó datos faltantes, tal como se puede observar en el siguiente gráfico:



Por lo tanto, se realizó la imputación indicada previamente y, en consecuencia, el patrón de valores faltantes se eliminó.

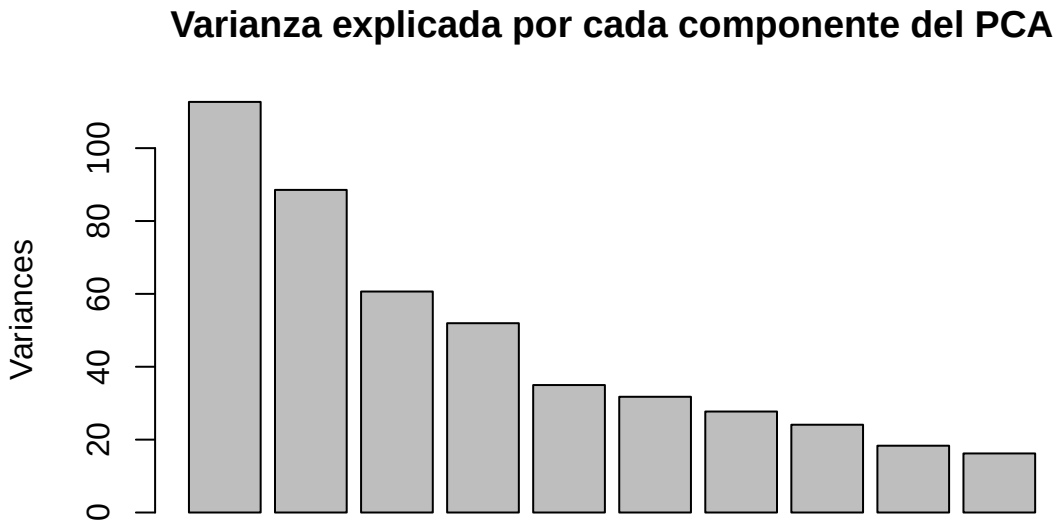


Reducción de la dimensionalidad (PCA).

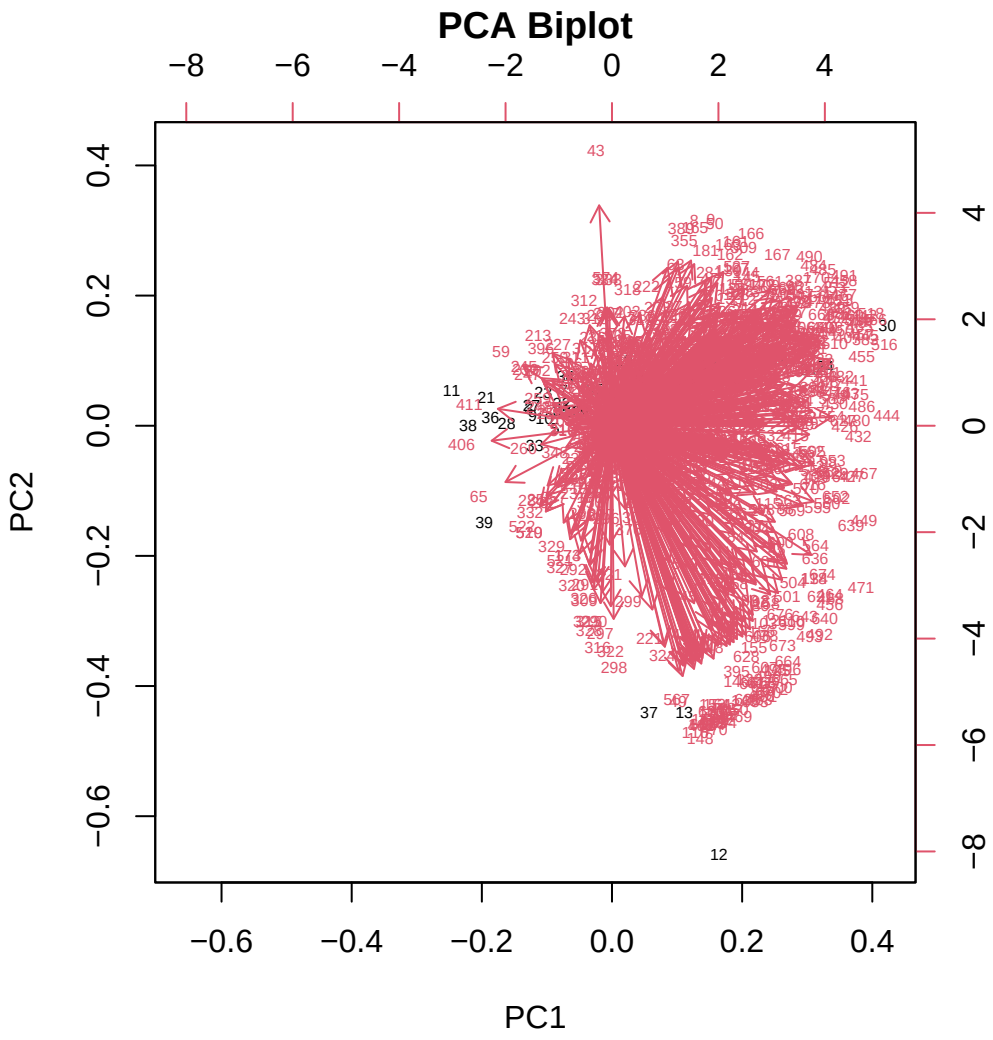
Para continuar con la comprensión del conjunto de datos se optó por una reducción de la dimensionalidad como un primer paso exploratorio.

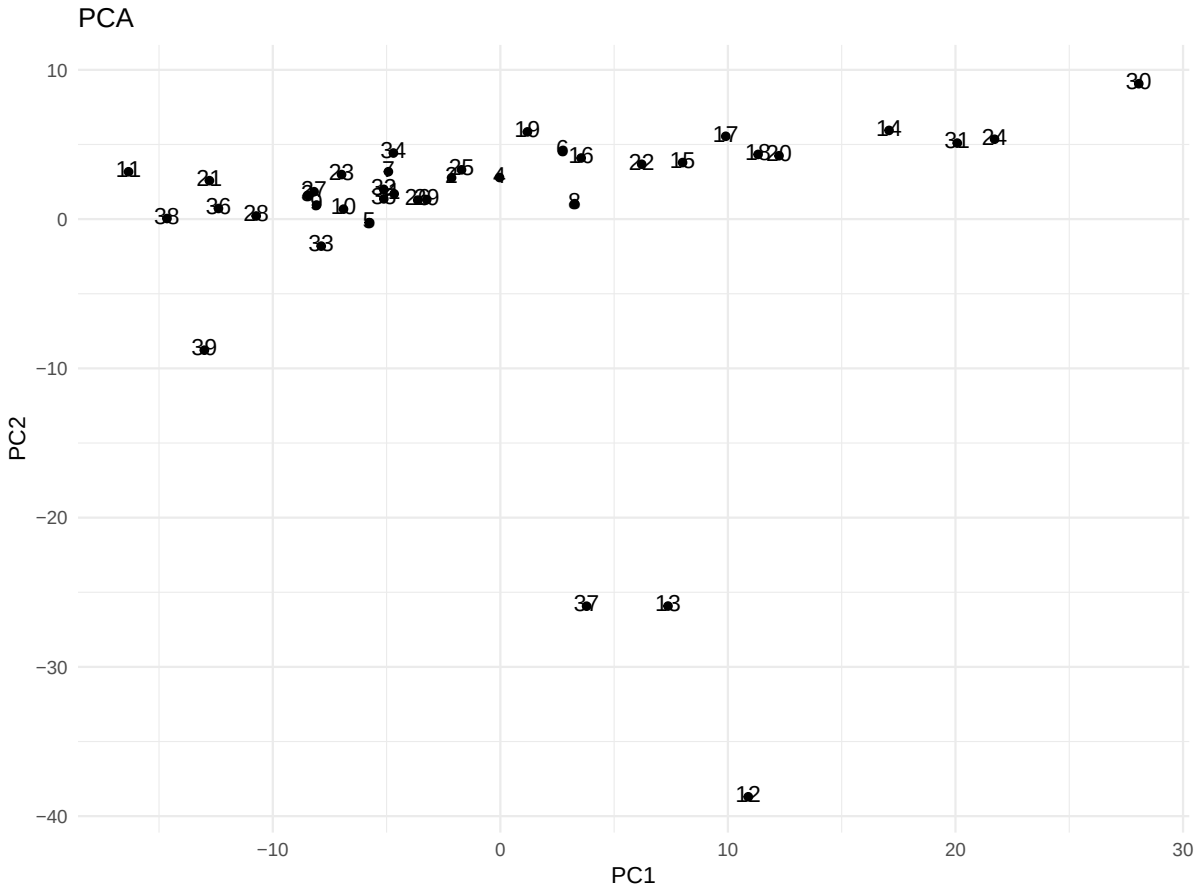
```
Importance of components:
                          PC1    PC2     PC3    PC4    PC5     PC6     PC7
Standard deviation     10.6155 9.4096 7.78755 7.2081 5.9144 5.63574 5.26443
Proportion of Variance  0.1633 0.1283 0.08789 0.0753 0.0507 0.04603 0.04017
Cumulative Proportion   0.1633 0.2916 0.37953 0.4548 0.5055 0.55155 0.59172
                          PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     4.90830 4.28152 4.02625 3.89794 3.85646 3.66975 3.49916
Proportion of Variance 0.03492 0.02657 0.02349 0.02202 0.02155 0.01952 0.01775
Cumulative Proportion  0.62664 0.65320 0.67670 0.69872 0.72027 0.73979 0.75753
                         PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     3.4348 3.34016 3.29652 3.22361 3.20965 3.01082 2.99240
Proportion of Variance 0.0171 0.01617 0.01575 0.01506 0.01493 0.01314 0.01298
Cumulative Proportion  0.7746 0.79080 0.80655 0.82161 0.83654 0.84968 0.86266
                          PC22   PC23    PC24    PC25    PC26    PC27   PC28
Standard deviation     2.89779 2.8165 2.73548 2.70148 2.64128 2.60374 2.4783
```

```
Proportion of Variance 0.01217 0.0115 0.01084 0.01058 0.01011 0.00983 0.0089
Cumulative Proportion  0.87483 0.8863 0.89717 0.90774 0.91785 0.92768 0.9366
                          PC29    PC30    PC31    PC32    PC33    PC34    PC35
Standard deviation     2.42070 2.37190 2.32610 2.20568 2.11135 2.07449 2.00571
Proportion of Variance 0.00849 0.00815 0.00784 0.00705 0.00646 0.00624 0.00583
Cumulative Proportion  0.94507 0.95323 0.96107 0.96812 0.97458 0.98082 0.98665
                          PC36    PC37    PC38      PC39
Standard deviation     1.88797 1.69350 1.66767 1.278e-14
Proportion of Variance 0.00517 0.00416 0.00403 0.000e+00
Cumulative Proportion  0.99181 0.99597 1.00000 1.000e+00
```



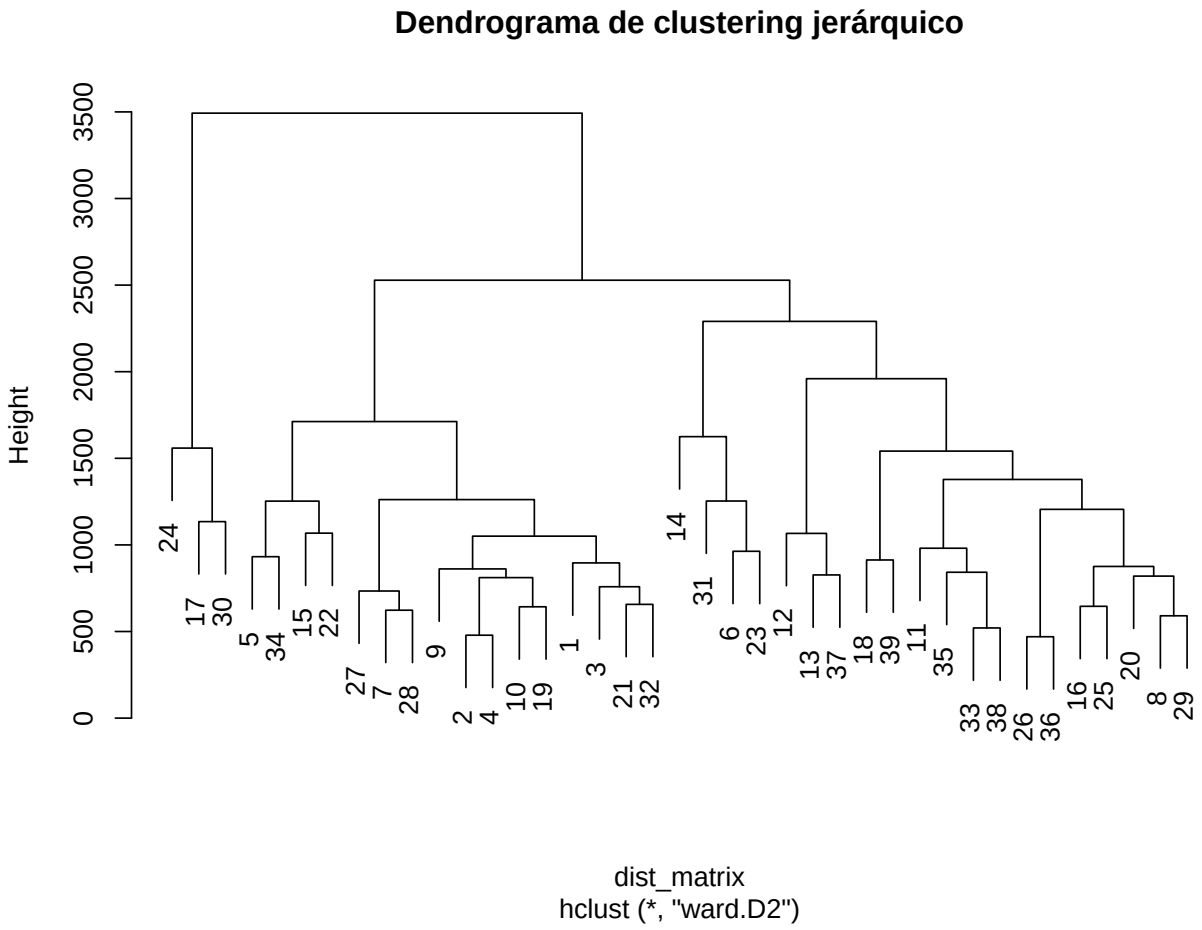
De donde vemos que los 2 primeros componentes explican alrededor del 30% de la varianza total. Viendo las *cargas* del primer componente vemos que corresponden valores adquiridos durante tiempos tardíos en el desarrollo del experimento y parecen incluir a ciertos glicerofosfolípidos, esfingolípidos y a los niveles del aminoácido Alanina. En principio, la vista gráfica podría ayudar a una mejor comprensión pero en este caso la gran cantidad de variables, afectó esta visualización.


PCA Biplot
PC1
PC2
−8
−6
−4
−2
0
2
4
−0.6
−0.4
−0.2
0.0
0.2
0.4
43
30
11
21
38
36
28
33
406
65
39
37
13
12

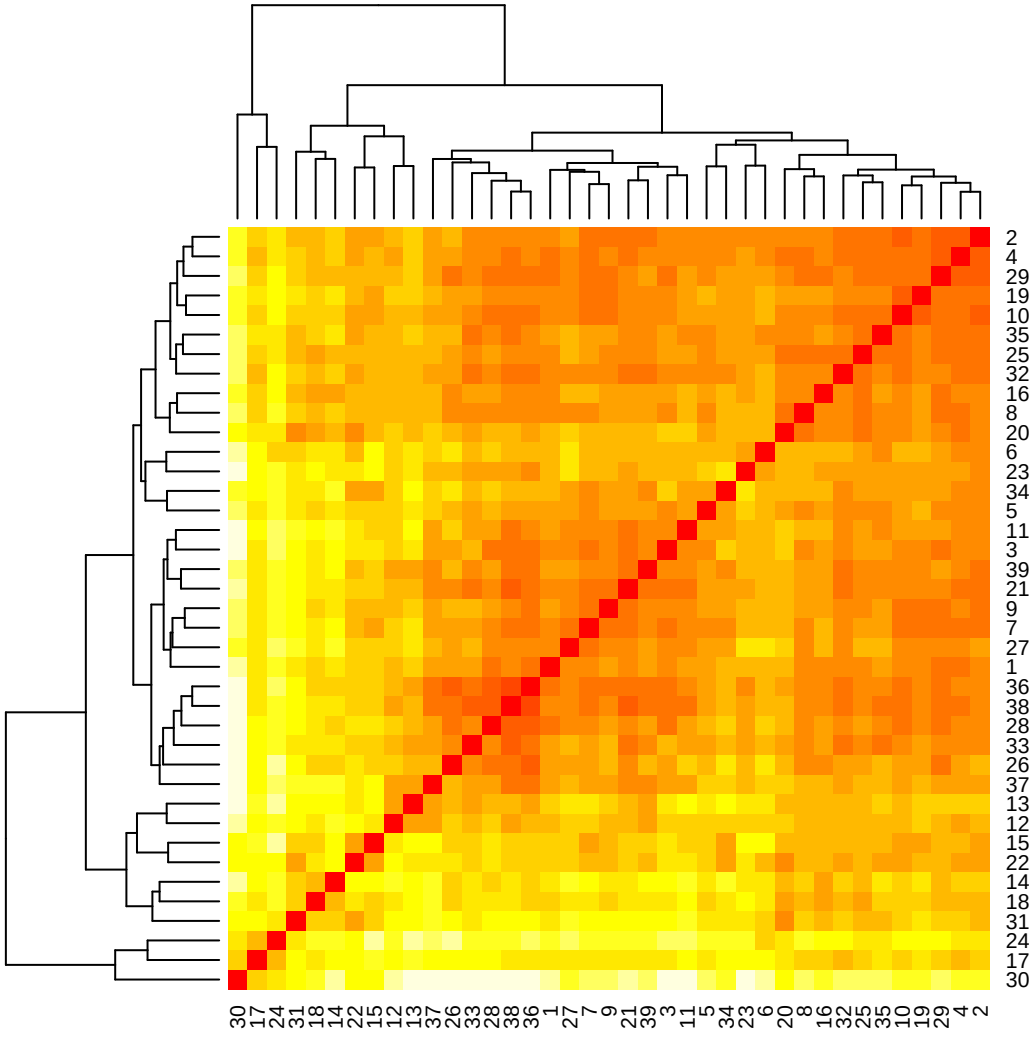


Otras opciones de visualización.

Se optó por otras formas de visualización para facilitar el acercamiento al conjunto de datos tal como un *clúster jerárquico* en donde podemos ver ciertas agrupaciones entre las muestras.



También se realizó un *heatmap de distancias* donde los colores que más tienden al rojo indican una cercanía mayor entre las muestras (siendo el color rojo la máxima cercanía y sería la de una muestra consigo misma y que se representa en la diagonal).



Discusión, limitaciones y conclusiones del estudio.

En este estudio, se ha trabajado con datos de metabolómica obtenidos de pacientes sometidos a cirugías bariátricas, lo que nos permite explorar cómo cambian ciertos metabolitos en función del tiempo y de las características individuales de los pacientes. La estructura de los datos fue organizada en un objeto `SummarizedExperiment`, lo cual facilitó la gestión tanto de los valores de los metabolitos como de los metadatos asociados, permitiendo realizar un análisis de los datos más eficiente. El análisis descriptivo y las visualizaciones iniciales proporcionaron información sobre la distribución de los valores de los metabolitos, destacando su variabilidad y sugiriendo la presencia de valores extremos en algunos. La imputación de valores faltantes mediante kNN contribuyó a la integridad de los datos, aunque esta técnica puede introducir cierto sesgo, especialmente en las variables con alta cantidad de datos faltantes. Posteriormen-

te, la reducción de dimensionalidad mediante PCA y el uso de otras visualizaciones brinda una vía inicial de análisis e interpretación. A pesar de lo anterio, el estudio presenta algunas limitaciones importantes: Baja capacidad explicativa del PCA: los primeros componentes principales explican una fracción limitada de la varianza total, lo que indica que los datos tienen una alta dimensionalidad y complejidad, y que es probable que existan relaciones no capturadas por el análisis lineal de PCA. Por otra parte, se requeriría un análisis más detallado para poder comprender y explicar con mayor detalle los hallazgos ya que en este estudio solo nos hemos detenido en un análisis exploratorio menor. En conclusión, la creación del contenedor SummarizedExperiment facilitó la estructuración de los datos y permitió llevar a cabo un análisis exploratorio con visualizaciones descriptivas y la búsqueda de patrones con ciertas visualizaciones. Sin embargo, debido a las limitaciones mencionadas, los resultados deben interpretarse con cautela. Probablmente, estudios más profundos que integren modelos no lineales y técnicas de aprendizaje automático, podrían mejorar la interpretación y predicción en el ámbito de la metabolómica y su relación con intervenciones médicas como la cirugía bariátrica.

Repositorio Github.

https://github.com/cariagamartinez/Cariaga-Martinez-Ariel-PEC1