Irene Castellanos González

ANÁLISIS DE DATOS ÓMICOS PEC 1

1.- Elección del dataset y descarga

Del repositorio de GitHub, nutrimetabolomics/metaboData**,** he descargado primero 'Data_Catalog.xlsx' para ver los datasets que hay. Finalmente he optado por el dataset '2023-CIMCBTutorial' y lo he descargado desde el repositorio del profesor.

2.- Creación del contenedor

No he tenido que instalar Bioconductor porque ya lo tenía de antes por haber practicado los ejercicios propuestos de entrenamiento.

Primero, he cargado en R las dos hojas del archivo GastricCancer_NMR, Data y Peak, aunque realmente la que he utilizado es Data ya que es la que contiene todos los datos tanto la información de las muestras (ID de las muestras, tipo de muestra, y clase) como las medidas de los metabolitos (columnas desde M1 hasta M149), la hoja Peak la he usado para conocer los nombres de los metabolitos.

Dado que algunos de los valores numéricos en las columnas de los metabolitos están formateados con comas (","), los he reemplazado por puntos (".") para que R los interprete correctamente como números decimales. Esto lo he hecho utilizando gsub() para hacer el reemplazo y as.numeric() para convertir las columnas a formato numérico.

A continuación, he extraído las columnas correspondientes a los metabolitos (M1 hasta M149) de sample_data, creando una matriz numérica. Asigno los nombres de las muestras como nombres de las filas de esta matriz, utilizando la columna SampleID de sample_data.

He tenido que transponer la matriz porque las columnas y las filas estaban al revés y arrojaba el error 'Error in rownames<-(*tmp*, value = .get_colnames_from_first_assay(assays)) invalid rownames length'. Entonces, para evitar este problema relacionado con la orientación de los datos en análisis posteriores, se transpone la matriz de metabolitos. Esto intercambia las filas por las columnas, asegurando que cada fila represente un metabolito y cada columna una muestra.

El siguiente paso ha sido crear un objeto 'colData' que contenga metadatos sobre las muestras. Este objeto incluye información como los identificadores de las muestras (SampleID), el tipo de muestra (SampleType), y la clase de la muestra (Class).

Finalmente, he creado el objeto SummarizedExperiment, que es un contenedor especializado para almacenar datos biológicos. Este contenedor incluye:

- Una matriz de datos (assays), que en este caso es la matriz transpuesta de los metabolitos (metabolite_data_t).
- Un conjunto de metadatos sobre las muestras (colData).

Por último, imprimo el objeto SummarizedExperiment para asegurarme de que se ha creado correctamente. Esto incluye información sobre el número de metabolitos, muestras y los metadatos asociados.

3.- Exploración de los datos

Para una exploración del dataset, he utilizado varias herramientas de R para tener una visión general de los datos:

- **Resumen de los datos de metabolitos**: utilicé la función summary() para generar un resumen estadístico básico de la matriz de metabolitos. Este resumen incluye información como los valores mínimos, máximos, medias, y la distribución general de los metabolitos. Así se puede identificar la naturaleza de los datos y posibles anomalías, como valores extremos o atípicos.

- **Estructura y contenido de los metadatos**: a través del comando head() exploré los metadatos asociados con las muestras para ver las primeras filas, y str() para observar la estructura del objeto col_data. Esto permite confirmar que las columnas clave (SampleID, SampleType, y Class) están correctamente formateadas y contienen los tipos de datos esperados.

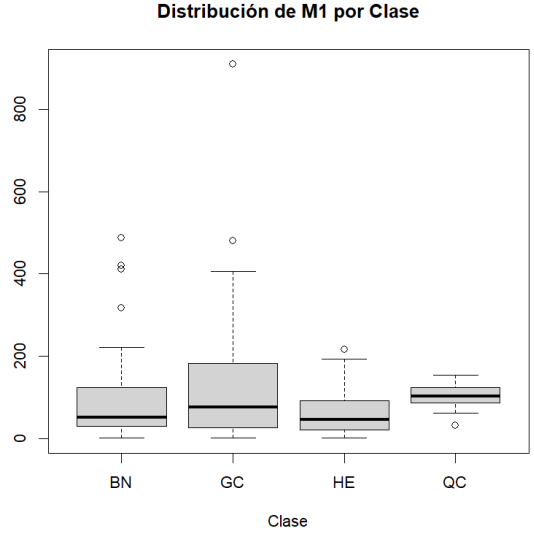
  Las primeras filas del dataframe muestran que hay 140 muestras en total, de las cuales algunas pertenecen al tipo QC (control de calidad) y otras son Sample (muestras de pacientes o grupos de estudio) y que las clases están distribuidas en cuatro categorías: BN, GC, HE, y QC.

  En cuanto a la estructura del dataframe muestra que cada columna contiene datos del tipo character (cadena de caracteres), lo que es adecuado para las categorías de muestras y que no hay metadatos adicionales asociados a este objeto, lo que sugiere que los datos principales están bien estructurados.

- **Análisis de la unicidad en las columnas de metadatos**. Mediante la función sapply(), se cuenta el número de valores únicos presentes en cada columna de los metadatos. Este análisis revela que había 140 muestras únicas en la columna SampleID, 2 tipos distintos de muestras en SampleType (QC y Sample) y 4 clases diferentes en Class (BN, GC, HE, y QC).

- **Frecuencia de los tipos y clases de muestras**. Creo una tabla de frecuencias para las variables categóricas SampleType y Class, lo que permite observar la distribución de los diferentes tipos de muestras y clases presentes en el dataset de forma que puedo evaluar si las clases están balanceadas o si hay predominancia de un grupo sobre otros. Para SampleType hay 17 muestras de tipo QC y 123 muestras de tipo Sample y para Class la distribución de las clases fue de 40 para BN, 43 para GC, 40 para HE, y 17 para QC.

- **Visualizar la distribución de un metabolito en concreto.** He realizado un gráfico tipo boxplot para visualizar la distribución del metabolito M1 en función de las clases (Class). Este gráfico me permite identificar posibles diferencias entre las clases en relación con los niveles del metabolito M1 y proporciona información sobre la dispersión de los datos dentro de cada clase.



El gráfico de boxplot muestra que el metabolito M1 tiene una mayor concentración en la clase GC en comparación con las otras clases. En BN, HE, y QC, la mediana es más baja y la dispersión es menor, lo que indica que las concentraciones de M1 son más homogéneas en esas clases. Se observan algunos outliers en varias clases, particularmente en GC, lo que podría ser indicativo de diferencias biológicas o técnicas en esas muestras.

- **Análisis de valores faltantes en el conjunto de datos.** Para ello, se verificó la calidad de los datos, se calculó el porcentaje de datos faltantes por cada metabolito. Este análisis nos muestra si había una cantidad significativa de datos faltantes.

4.- Reposición de los datos en GitHub

Antes de nada, me he creado una cuenta en GitHub ya que no tenía. Una vez creada desde la página oficial he creado el repositorio y he incluido todos los entregables que menciona la PEC1. La URL del repositorio es la siguiente:
https://github.com/irenecastellanos/Castellanos-Gonzalez-Irene-PEC1.git