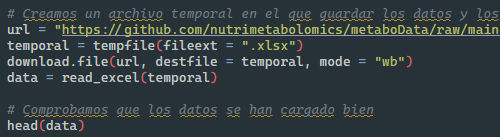
**JAVIER MARTÍNEZ DELGADO**

**PEC1-ÓMICAS**

**1.-Seleccionar un dataset de metabolómica**

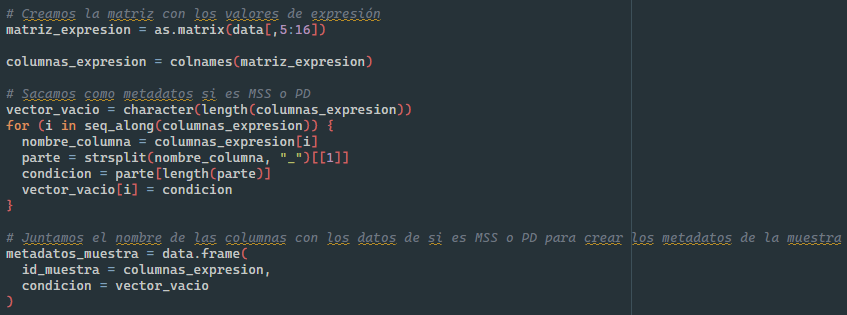
El dataset seleccionado es [metaboData/Datasets/2018-Phosphoproteomics at main · nutrimetabolomics/metaboData · GitHub](https://github.com/nutrimetabolomics/metaboData/tree/main/Datasets/2018-Phosphoproteomics).



Lo que he hecho, es descargar los datos de github directamente a un archivo temporal por lo que no estoy cargando un archivo que esté en mi ordenador y el código se puede utilizar desde cualquier dispositivo sin que de error. Una vez guardado en el archivo temporal, lo cargamos y lo visualizamos para comprobar que todo está bien

**2.-Crear un contenedor SummarizedExperiment que contenga los datos y metadatos**

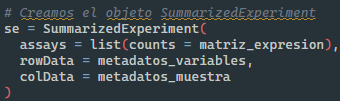
Lo primero que tenemos que hacer es crear la matriz con los valores de expresión por lo que extraeremos esas columnas determinadas del dataframe y guardaremos también el nombre de las columnas, las cuales procesaremos para obtener metadatos

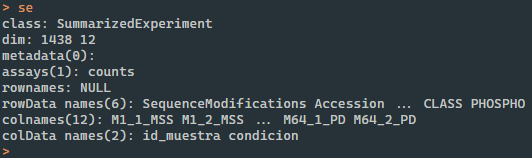


Con el resto de columnas, crearemos más metadatos, pero de las variables en este caso en lugar de metadatos de las muestras



Finalmente crearemos el contenedor SummarizedExperiment

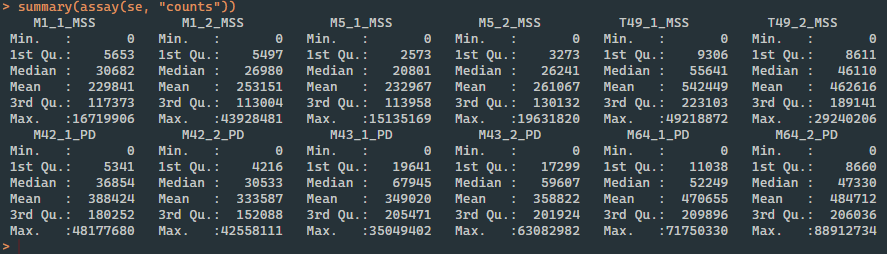




Las dimensiones del contenedor son 1438x12 al igual que el dataframe que habíamos cargado. En assays podemos comprobar que tenemos la matriz de expresión con los valores de expresión para cada muestra. Los valores de rowdata proporcionan información adicional sobre cada muestra y coldata nos dice que tenemos dos columnas de metadatos para las muestras: el id y si es MSS o PD.

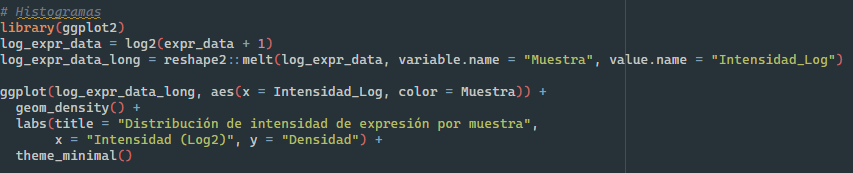
**3.-Exploración del dataset**

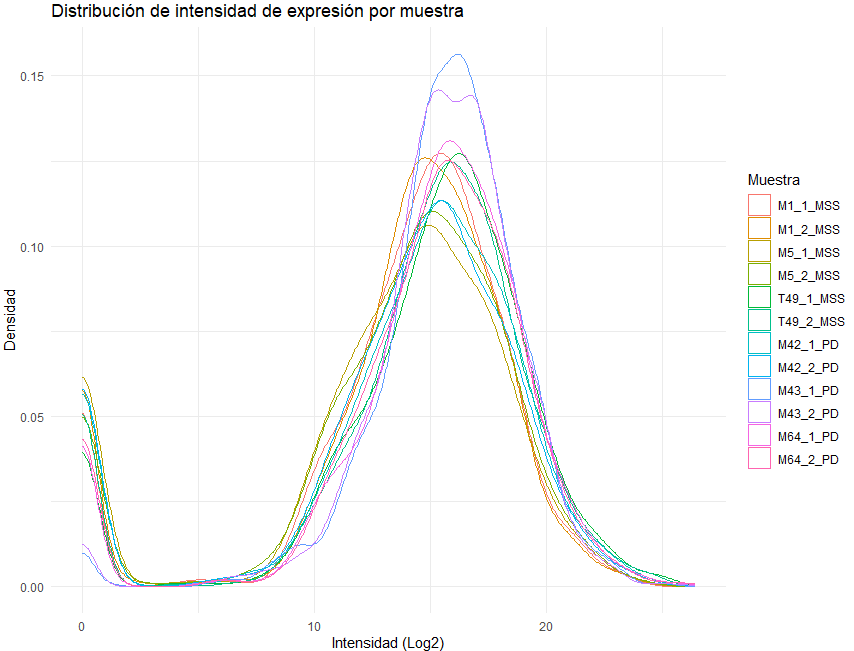
Lo primero que se suele hacer en un análisis exploratorio cuando se tienen variables cuantitativas es un resumen estadístico de las variables



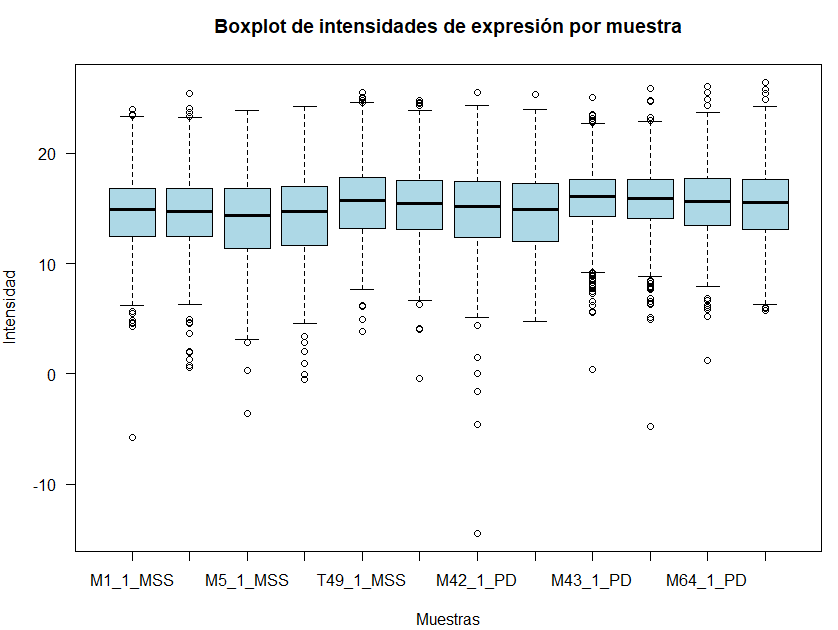
Llama mucho la atención el rango tan grande de las variables ya que el mínimo es 0 y el máximo es de decenas de millones.

A continuación, queremos ver la distribución de estos datos por lo que haremos una gráfica, pero antes, tendremos que aplicar el logaritmo a los datos ya que, al hacer la gráfica sin el logaritmo, como los datos están muy concentrados cerca del cero, no se puede apreciar nada



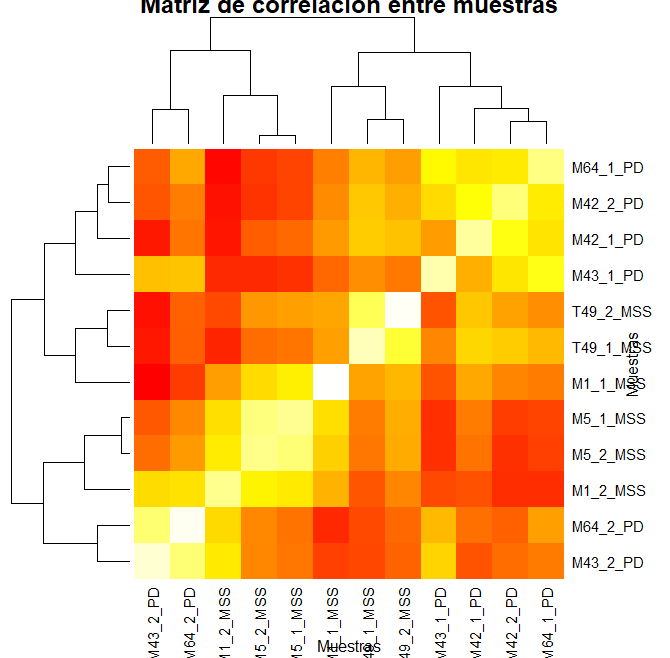


Podemos ver que las muestras se diferencian más entre sí a una intensidad muy baja cercana al cero o a una intensidad determinada formando un pico

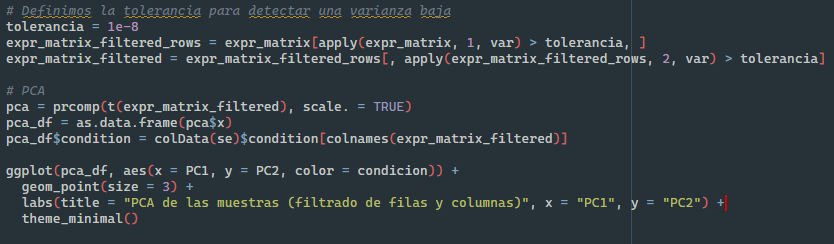


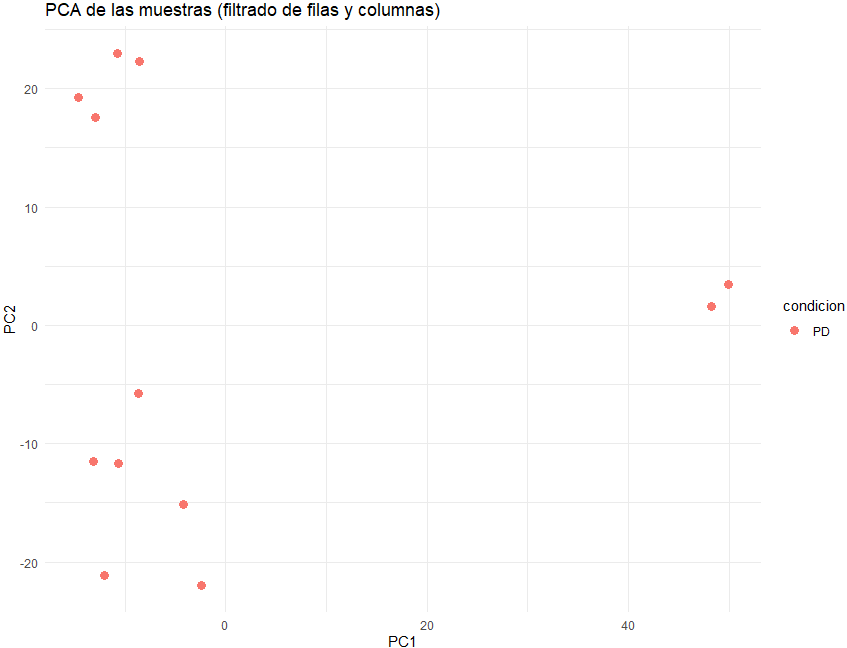
Al hacer un boxplot para ver cada variable y los outliers que tienen (también se ha tenido que usar los datos con logaritmo). Podemos ver que no hay ninguna que destaque sobre otras muestras pero sí que hay algunas con un rango mayor en los outliers

Si hacemos una matriz de correlación entre las variables podemos ver las que están fuertemente relacionadas



Para hacer un PCA, se ha tenido que hacer un threshold de intensidad para eliminar las de menor intensidad ya que el código devuelve un error al hacer el PCA

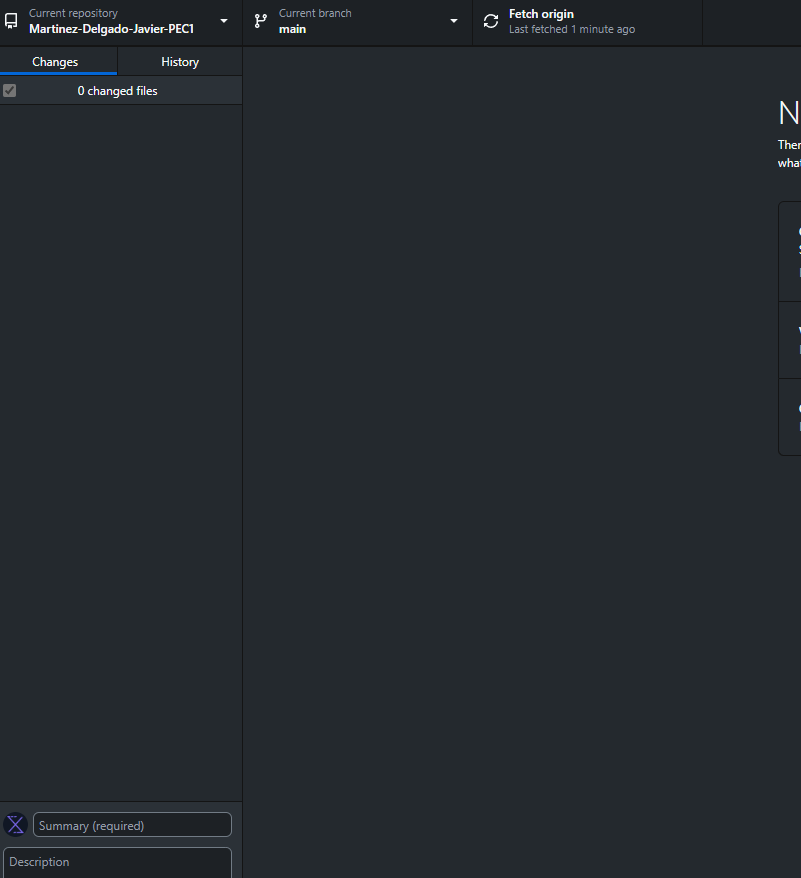


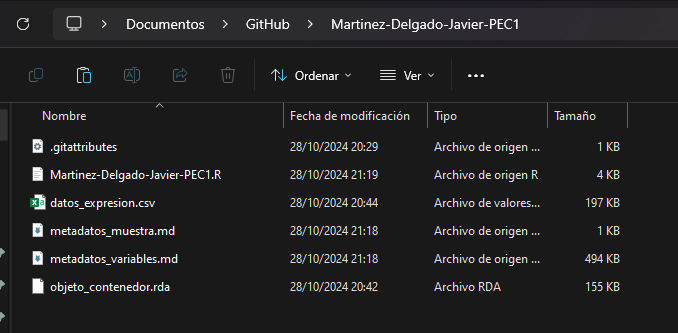


Como se puede ver a simple vista en el PCA, hay 3 grupos principales y Sólo quedan muestras del tipo PD, por lo que se han eliminado todas las muestras de tipo MSS

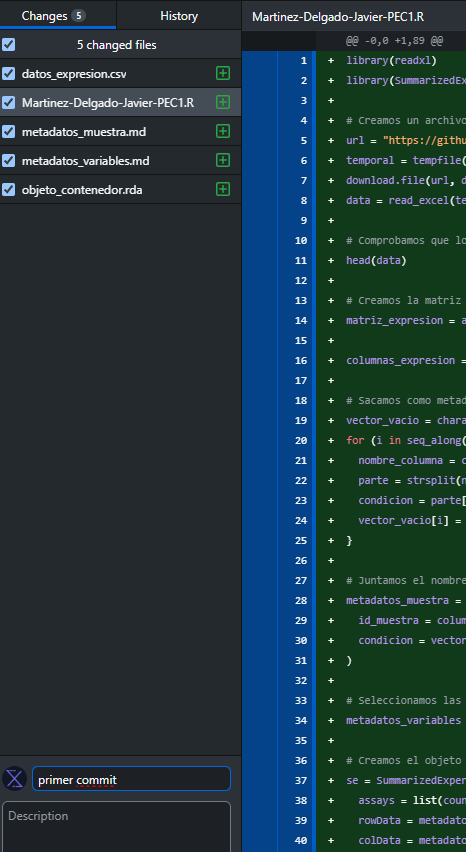
**4.-Reposición de los datos en GitHub**

En primer lugar, creamos el repositorio dentro de nuestra cuenta (en mi caso lo hago en github desktop ya que me es más fácil al tenerlo ya configurado y usarlo a menudo)





Metemos los ficheros en la carpeta del repositorio



Y hacemos un commit de los archivos

https://github.com/xshaikz/Martinez-Delgado-Javier-PEC1