Èlia Rovira Figueras
Anàlisi de Dades Òmiques
6 novembre 2024

Informe de la pràctica: Creació d'un Contenidor SummarizedExperiment i Anàlisi Inicial de Dades

1. Introducció

Aquest informe descriu el procés per descarregar, preparar i analitzar un conjunt de dades de l'estudi "Metabotips de resposta a la cirurgia bariàtrica independentment de la magnitud de la pèrdua de pes" en R mitjançant l'objecte SummarizedExperiment.

L'objectiu és integrar les dades principals amb les metadades i realitzar una análisi inicial, incloent visualitzacions i estadístiques descriptives.

2. Descàrrega i preparació de les Dades

Els fitxers CSV utilitzats es mostren a continuació:

- DataInfo_S013.csv: conté les metadades de les variables amb 695 files i 4 columnes (VarName, varTpe, Description) – es defineix com a data_info
- DataValues_S013.csv: conté les dades principals amb 39 files i 696 columnes – es defineix com a data_values

3. Comparació de noms i Preparació del Dataset

Es fa una comparació inicial dels noms per assegurar que les variables entre data_info i data_values coincideixen.

4. Creació del Contenidor SummarizedExperiment

Un cop els noms de les variables van coincidir, es va transposar data_values per convertir les variables en files, tal com requereix SummarizedExperiment. Seguidament s'executa la comanda de SummarizedExperiment.

L'objecte SummarizedExperiment creat té les característiques següents:

- Dimensions: Consta de 695 files (variables) i 39 columnes (mostres).
- Class: La classe de l'objecte és SummarizedExperiment, que és ideal per treballar amb conjunts de dades complexos que inclouen dades principals i metadades.
- Assays: Conté un únic assay anomenat counts, que representa la matriu de dades principals. Aquest assay emmagatzema les mesures quantitatives de les 695 variables per a cada una de les 39 mostres.
- Row Names: Les files tenen noms únics per a cada variable (ex., SUBJECTS, SURGERY, SM.C24.0_T5, etc.), cosa que facilita la identificació de cada mesura.

- Row Data: A més de les dades principals, inclou informació addicional a rowData, amb 4 columnes (...1, VarName, varTpe, Description) que proporcionen descripcions i tipus per a cada variable, millorant la comprensió de cada mesura.

- Col Names: Les columnes no tenen noms específics, ja que les mostres no inclouen noms individuals (colnames: NULL). Això indica que les mostres són anònimes o que no es necessita una identificació detallada per a l'anàlisi.

Aquest objecte proporciona una estructura organitzada que facilita l'accés i manipulació tant de les dades com de les metadades, cosa que serà útil per a anàlisis posteriors i per a integrar diverses fonts d'informació dins d'un únic contenidor.

5. Exploració del Dataset

5.1 Dimensions i Estructura

Es verifiquen les dimensions i la integritat de les dades.

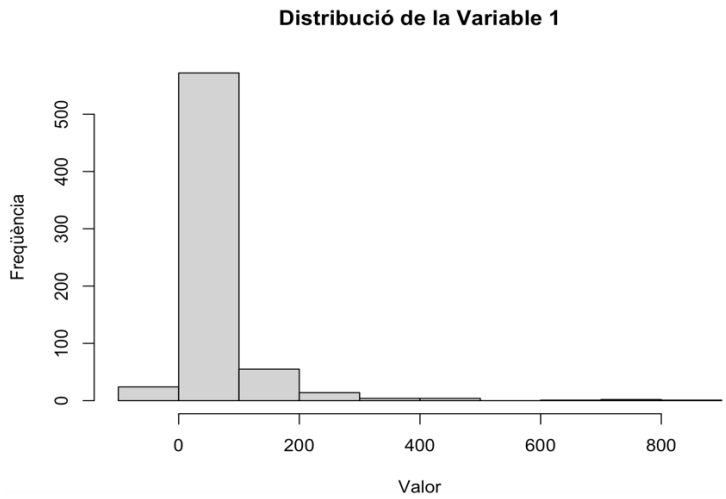
5.2 Estadístiques Descriptives i Conversió Numèrica

Com que algunes columnes contenen valors no numèrics, s'identifiquen i es converteixen a valors numèrics, substituint els valors que no es poden convertir amb NA.

6. Visualització de les Dades

Per comprendre millor la distribució de les dades, es creen els següents gràfics.

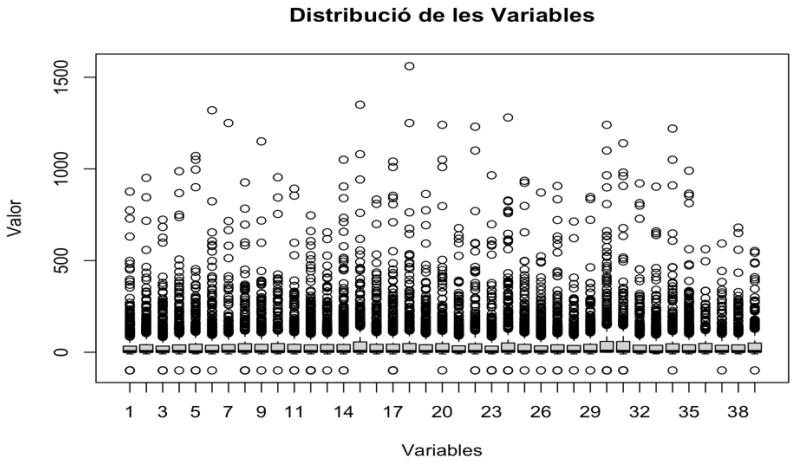
6.1 Histograma de valors



Imatge 1. Gràfic en forma d'histograma de la distribució de la variable 1.

L'histograma mostra la distribució dels valors de la Variable 1. Observem una alta freqüència de valors propers a zero, amb una distribució que decreix de manera brusca a mesura que augmenten els valors. La majoria dels registres es concentren en el rang de 0 a 200, amb alguns valors extrems que arriben fins a 800. Això suggereix que els valors de la Variable 1 no estan uniformement distribuïts, sinó que hi ha una acumulació significativa en valors baixos. Pot ser indicatiu d'una variable que mesura un fenomen esporàdic o amb molta variabilitat entre subjectes.

6.2 Boxplot per la distribució de totes les variables



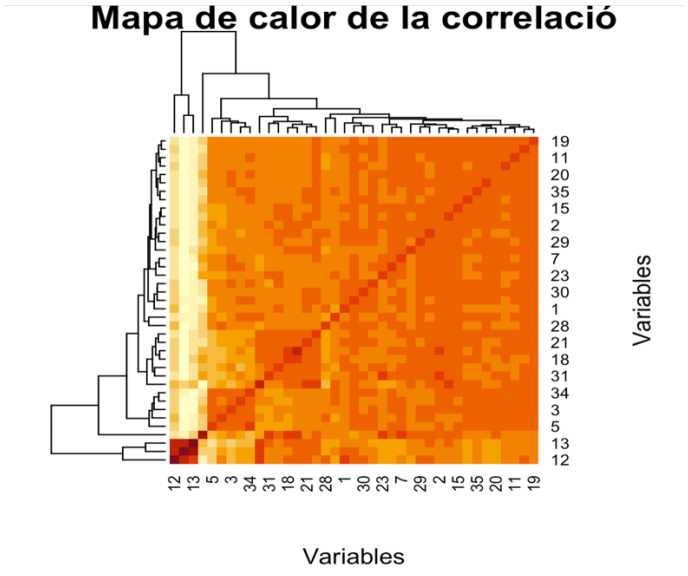
Imatge 2. Boxplot de la distribució de les diverses variables de l'estudi.

Cada columna del gràfic representa una variable i mostra la distribució dels seus valors. El boxplot visualitza el rang interquartílic (IQ) per cada variable, així com els valors mínims, màxims i els punts atípics.

La majoria de les dades es concentren prop de zero, però hi ha molts valors atípics (outliers) dispersos per sobre, indicant una variabilitat significativa. La presència de molts outliers suggereix que algunes variables tenen distribucions asimètriques o valors extrems. Aquest gràfic permet identificar les variables amb més variabilitat i aquelles amb valors fora de rang.

La distribució general sembla similar entre les variables, amb un patró de dispersió i outliers constant. Això podria suggerir que aquestes variables mesuren aspectes similars o tenen una estructura subjacent comuna.

6.3 Mapa de calor de la correlació



Imatge 2. Mapa de calor per observar la distribució de les diverses variables de l'estudi.

Aquest mapa de calor mostra la matriu de correlació entre les variables del conjunt de dades, on els colors més foscos (vermell intens) indiquen una correlació més alta, i els colors més clars (groc o blanc) una correlació baixa o nul·la.

- Relacions entre Variables: Es poden veure grups de variables amb correlacions significatives (zones fosques), suggerint relacions o mesures d'aspectes similars.

- Clústers de Variables: El dendrograma mostra agrupaments de variables segons la seva similitud, útils per identificar conjunts que poden representar fenòmens relacionats.

Aquest gràfic permet identificar ràpidament les relacions entre variables i facilita la selecció de grups per a l'anàlisi posterior o la reducció de dimensionalitat.

7. Conclusió

S'ha creat correctament un objecte SummarizedExperiment que conté 695 variables (files) i 39 mostres (columnes). Aquest objecte integra tant les dades principals com les metadades associades, permetent un anàlisi organitzat i eficient del conjunt de dades. Les visualitzacions inicials (histograma, boxplot i mapa de calor de la correlació) han proporcionat informació valuosa sobre la distribució, la variabilitat i les relacions entre les variables.

L'objecte SummarizedExperiment és una eina potent per treballar amb dades complexes, ja que facilita l'accés a les dades i metadades en un únic contenidor. Aquesta estructura modular serà útil per a anàlisis futures, permetent identificar patrons, outliers i possibles agrupaments de variables amb relacions significatives.

8. Publicació del Projecte a GitHub

Després d'analitzar les dades, el projecte complet és repostat a GitHub. El repositori inclou:

1. L'informe complet en format .Rmd i .html/.pdf.
2. L'objecte SummarizedExperiment en format .Rda:
3. El codi complet per a l'exploració de dades.
4. Els fitxers de dades originals en format CSV.
5. Un arxiu readme.me amb una descripción del projecte i les metadades.

El directori URL per al projecte es mostra a continuació:

https://github.com/erovirafigueras/Rovira-Figueras-Elia-PEC1.git