Assignment

Q1:

• Read both files into R

• Display the number of rows and columns for each dataset.

• Print the first 6 column names of the gene expression matrix

Q2:

• Convert all RACE and ETHNICITY values to lowercase and trim extra spaces.

• Create a numeric AGE\_num column.

• Add a new variable age\_group <18, 18–39, 40–64, 65+, or "unknown".

• Replace any missing DISEASE\_ONTOLOGY values with "unknown".

Q3:

• Create a subset of all lung-related cases

• Count how many male vs female patients are in this subset.

• Find the average age (ignoring NA) by DISEASE\_ONTOLOGY.

Q4:

* From the clinical file, select two columns that identify the sample (PATIENT\_ID, DISEASE\_ONTOLOGY).
* Perform a left\_join() to add disease info to each expression row by matching the sample name to PATIENT\_ID.
* Print how many rows successfully matched (non-NA DISEASE\_ONTOLOGY after join).

Q5:

* Using the clinical dataset, make a **bar plot** showing the number of patients in each DISEASE\_ONTOLOGY group.