Assignment 4 – Instruction

Part A: Breast Cancer Wisconsin Dataset

1. Histogram and QQplot of `texture_mean`

- **What to do:** Load the dataset, make a histogram and a QQplot of `texture_mean`.
- **Hint:** Use `hist()` for the histogram and `qqnorm()` + `qqline()` for the QQplot. The QQplot checks normality.

1a. Interpretation

- **What to do:** Decide whether the distribution looks normal.
- **Hint:** In a QQplot, data that is normal will fall close to the diagonal line. Deviations suggest skew or heavy tails.

1b. Log Transformation

- **What to do:** Apply a log transform, make new plots, compare.
- **Hint:** If a variable is right-skewed, a log transform often makes it closer to normal. Use `log()` in R.

2. Linear Model

- **What to do:** Fit a linear regression with `lm(texture_mean ~ area_mean).`
- **Hint:** After fitting, check the p-value of the slope coefficient. This tells you if `area_mean` significantly predicts `texture_mean`.

3. Visualization

- **What to do:** Plot `texture_mean` vs `area_mean` with a regression line. Add the equation in the title.
- **Hint:** `geom_smooth(method="lm")` will add the regression line in ggplot. You can find the slope and intercept from `summary(model).`

Part B: RNA-Seq Differential Expression

4. Filter Genes

- **What to do:** Remove genes with 2 or fewer samples with counts. Report how many genes and samples remain.
- **Hint:** Think of this as keeping rows (genes) with enough data. Use `rowSums()` on the count matrix to filter.

5. Run DESeq

- **What to do:** Fit the model using DESeq.
- **Hint:** The function is `DESeq()` on your DESeq2 object. It estimates size factors, dispersions, and fits the GLM.

6. Differential Expression – DPN Group

- **What to do:** Extract results comparing Control vs. DPN. Find top 5 genes, and count up- vs down-regulated.
- **Hint:** Use the `results()` function with a contrast. Then order by adjusted p-value. For counts of up/down, check the sign of `log2FoldChange`.

7. Highest Log-Fold Change Gene

- **What to do:** Find the gene with the maximum log-fold change and visualize its expression.
- **Hint:** `which.max()` will help you identify the index of the highest log-fold change. Use `plotCounts()` to see expression across treatments.

8. Volcano Plot

- **What to do:** Make a volcano plot for Control vs DPN, highlight genes with $p < 0.05$.
- **Hint:** X-axis should be `log2FoldChange`, Y-axis should be `-log10(pvalue)`. Genes that are both far left/right and high up are the most interesting.