**Practice Task: R Plotting with Telomeric Repeat Variant (TRV) Synthetic Data**

**Objective:**  
Learn to create exploratory visualizations including boxplots, heatmaps, and scatter plots in R using synthetic TRV data.

**Instructions:**

**Step 1: Set Up Your R Environment**

* Open RStudio or your preferred R environment.
* Create a new R Markdown file.
* Make sure you have internet access to install packages.
* Install the required R packages: ggplot2, dplyr, etc.

**Step 2: Generate Synthetic TRV Data**

**2.1 Define Sample Metadata**

* Create a dataset with ~30 samples.
* Assign metadata columns for each sample:
  + state: disease state of the patient (e.g. "crc" (colorectal cancer), "brca" (breast cancer), "hh" (healthy control))
  + patient\_id: patient identifier (e.g. p1, p2, …)

**2.2 Define TRVs**  
Use the following list of TRVs:

* Canonical: TTAGGG
* Variants: TAGGG, TAAGGG, TTCGGG, TTAAGGG, TGAGGG, TTAGGGG GTAGGG, TTGGGG, TTGGG, CTAGGG, TTTAGGG, TCAGGG, TTAAGG, TAGGGG, TTTGGG, TTAGCG, TTTTAGGG, TGGGGG, TTATGG, TTACGG, TTAGGA, TTCCGA, ATCTGG, TCTGGG, TTGCGG, GGAGGG, ATAGGG, GTGGGG

**2.3 Simulate TRV Percentages**

* For each sample:
  + Assign ~90% to TTAGGG (canonical variant)
  + Distribute remaining ~10% randomly across the TRVs
  + Ensure that percentages sum to 100%

**2.4 Final Output Structure**  
Your final dataframe should have the following structure:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| patient\_id | state | TTAGGG | TAGGG | TAAGGG | … |
| p1 | crc | 91.2 | 1.5 | 0.5 | … |
| p2 | hh | 87 | 3.2 | 0.4 | … |
| … | … | … | … | … | … |

**Step 3: Prepare Data for Plotting**

* Reshape the dataframe as needed for the plots described in the next steps.

**Step 4: Create Boxplots**

* Create boxplots comparing TRV percentage distribution in each cancer group with healthy group across all TRVs. For each TRV you will need to have 2 boxplots colored by state.
* You can have a side-by-side boxplot layout:
  + Left panel: Non-canonical TRVs.
  + Right panel: Canonical TRV only.
* Overlay data points on each box.
* Annotate the plot with statistical comparison p-values (e.g., Mann–Whitney U with multiple testing correction) between cancer and healthy states for each TRV.
* Add clear axis labels, legend, and descriptive titles

**Step 5: Create a Heatmap**

* Generate a clustered heatmap.
* Cluster both rows (TRV variants) and columns (samples) hierarchically.
* Add a sample annotation bar showing the state of each sample using distinct colors.
* Use a diverging color gradient (e.g., blue–white–red) to reflect TRV percentage values.
* Include legends for:
  + TRV percentage color scale
  + Sample group (state) colors
* Label axes with TRV names and sample identifiers.
* Rotate x-axis labels if necessary for better readability.

**Step 6: Create a Scatter Plot**

* Generate a log-log scatter plot comparing TRV percentages between cancer and healthy groups (e.g., brca vs hh).
* Apply a log10 transformation to TRV percentages. Add a pseudocount before transformation if needed.
* For each TRV:
  + Calculate the mean percentage and standard error across each group.
* Plot each TRV as a point:
  + X-axis: Log10(%) in healthy
  + Y-axis: Log10(%) in cancer
* Add error bars representing standard error.
* Add a diagonal dashed line (y = x) as a reference.
* Annotate TRVs by sequence on the plot.
* Title the plot and label the axes accordingly.

**Step 7: Interpret the Plots**

* Observe whether different groups show distinct TRV patterns.

**Optional Challenge:**

* Add other types of visualisations of your choice for better understanding of TRV patterns and distributions across states.

**Step 8: Save Your Work**

* Save your .Rmd file with plots.

Good luck and have fun exploring your first TRV data!