Description of the Script

This document provides an overview and explanation of the 'z_scores_analysis.Rmd' script, detailing its main functions, statistical methods, and visualization purposes.

1. **Initial Setup**

    - The script begins by loading the necessary libraries and data files to ensure that all dependencies and datasets are available for analysis.

2. **Preprocessing**

    - Preparing the data for analysis by subsetting and filtering.

        i. **Subsetting by diagnostic Code**: The data is divided into subsets based on unique diagnostic codes (dcode).
        ii. **Column Filtering:** Metadata columns (e.g, sex, dcode, site) are removed, leaving only relevant columns for analysis.
        iii. **Supra- and Infra- Threshold Calculations**: For each subset, regions with z-scores greater than 1.96 (supra-values) and regions with z-scores less than -1.96 (infra-values) are calculated.

    - Depending on whether the data is single-column or multi-region, a similar code is applied for preprocessing. For both, the results are stored in summary data frames which include:
        i. Region name
        ii. Supra-normal count
        iii. Infra-normal count
        iv. Total subject count
        v. Percentages of supra- and infra-normal values

3. **Statistical Analysis**

    - The statistical analysis section conducts several tests and calculations aimed at comparing proportions between groups. Specifically, it compares the distribution of supra-normal and infra-normal values across different regions, evaluating whether the frequencies of these values differ significantly between groups.

        i. **Traditional Chi-square test:** The function **chisq.test()** is used to compare the proportions of supra-normal and infra-normal values between two groups for each region. We obtain p-value (p_val) and p-value corrected using the FDR (False Discovery Rate) method (p_val_corrected).

ii. **Permuted Chi-square test:** The permuted chi-square test is applied using the function **chisq_test(***-, -, distribution = approximate (nresample = 9999)***)**. The output includes: p-value obtained from the permuted chi-square test (perm_p_val) and p-value corrected using the FDR method for the permuted results (perm_pval_corrected).

iii. **Output:** The results are stored in two dataframes 'infra_results_df' and 'supra_results_df' in multi-regions dataframes. For the case of single-column dataframes, results are stored in 'results_df'.

4. t-SNE for Dimensionality Reduction and PAM Clustering:

- This section applies t-SNE (t-distributed Stochastic Neighbor Embedding) for dimensionality reduction and PAM (Partitioning Around Medoids) clustering. First, the optimal number of clusters (k) is calculated for each dataset. Then, the clustering is forced to k=3 to see how the data separates into 3 distinct groups. These results are visualized using density plots, scatter plots, and a distribution plot. The same plots are also depicted for the actual groups, with each group represented by different colors. Finally, all the plots are combined into a single figure with different sections to allow comparison of the results.