Description of the Script

This document provides an overview and explanation of the 'final_z_analysis_figures.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 Functions**

   ○ Several functions are defined to prepare the data for analysis. These include:

      i. **subsetAndLabel:** Divides the data into different data frames based on diagnostic codes 'dcode'. Metadata columns (e.g, sex, dcode, site) are removed, leaving only relevant columns for analysis.
      ii. **subsetByGender:** Filters the data to include only male subjects.
      iii. **setAboveThresholdToOne:** Converts regions with z-scores greater than 1.96 (supra-normal regions) into a binary matrix where values above the threshold are set to 1, and others to 0.
      iv. **setBelowThresholdToOne:** Similar to setAboveThresholdToOne, but for regions with z-scores less than -1.96 (infra-normal regions).
      v. **calculate_total_ones:** Computes the total number of supra-normal and infra-normal regions for each subject.
      vi. **calculate_percentage_subjects:** Calculates the percentage of subjects with a specific number of supra-normal or infra-normal regions.

3. **Statistical tests**

   ○ The script incorporates statistical tests to compare distributions and proportions:
      i. **Welch's t-test (t.test()):** Used to compare z-distributions across regions.
      ii. **Wilcoxon rank-sum test (wilcox_test())**: Applied to compare skewed proportions.
      iii. **T-tests and Cohen's D (t.test(), cohen.d()):** These are applied after one of the plots below to compare the mean z-scores between groups, evaluating statistical significance and effect size.

4. **Plots**

   ○ **Bar Plots for Supra and Infra-normal Regions:** The x-axis represents the number of affected regions while the y-axis shows the proportion of subjects with that number of affected regions. Different colors distinguish the groups.

○ **Density plot for Z-scores:** A density plot is generated to show the distribution of z-scores for each group. The ±1.96 values are marked with dashed lines to indicate significant regions.

○ **Z Distribution plot by region:** This plot visualizes the distribution of z-scores by brain region for each group. The plots are arranged in a grid to facilitate comparison across regions.