**VDSS\_ZIP\_BORDER\_STATUS\_TYPE\_BAD\_ZIPCOUNTY\_RATE.R - README**

**Author: Dylan Craig**

**Date Last Updated: 12/29/2024**

**File Name: VDSS\_ZIP\_BORDER\_STATUS\_TYPE\_BAD\_ZIPCOUNTY\_RATE.R**

**Purpose**

This script merges datasets related to ZIP code types, border statuses, and bad ZIP-County rates. It calculates summary statistics and performs pairwise t-tests on the data, grouped by BORDER\_STATUS and TYPE. The results provide insights into service discrepancies and variations in ZIP code characteristics.

**Overview**

The script performs the following tasks:

1. Loads and merges data on ZIP types, border statuses, and bad ZIP-County rates.
2. Calculates weighted metrics such as the weighted bad ZIP-County rate.
3. Summarizes statistics grouped by BORDER\_STATUS and TYPE.
4. Performs t-tests to compare group means within BORDER\_STATUS and TYPE categories.
5. Saves the results to an Excel file for further analysis.

**Datasets and Tools**

**Input Files:**

1. **VDSS\_Zip\_Type.xlsx:** Contains ZIP code type classifications (e.g., PO Box, Standard, Unique).
2. **VDSS\_Zip\_FIPS\_Border\_Status\_Final\_Designation.xlsx:** Includes ZIP code border status classifications (e.g., Bordering, Interior, Overlapping).
3. **bad\_zipcounty\_rate.dta:** Provides the rate of incorrect VDSS services (bad\_zipcounty\_rate) and the number of cases (N) for each ZIP code.

**Output File:**

1. **VDSS\_Summary\_Statistics\_and\_T\_Tests.xlsx:** Summary statistics and t-test results for BORDER\_STATUS and TYPE categories.

**Dependencies:**

* **R Libraries:**
  + dplyr: For data manipulation.
  + readxl: For reading Excel files.
  + writexl: For writing data to Excel files.
  + haven: For reading .dta files.

**Script Functionality**

**1. Loading and Merging Data:**

* Reads ZIP type, border status, and bad ZIP-County rate datasets.
* Merges these datasets by ZIP code to create a unified dataset for analysis.

**2. Calculating Weighted Metrics:**

* Computes a weighted bad ZIP-County rate (weighted\_bad\_zipcounty) by multiplying bad\_zipcounty by N and dividing by the total number of cases.

**3. Summary Statistics:**

* Groups data by BORDER\_STATUS and calculates:
  + Average number of cases (avg\_N).
  + Average bad ZIP-County rate (avg\_bad\_zipcounty).
  + Weighted bad ZIP-County rate.
  + Counts of non-missing values for N and bad\_zipcounty.
* Groups data by TYPE to calculate the same metrics.

**4. Performing T-Tests:**

* Conducts pairwise t-tests for:
  + Differences in means of N, bad\_zipcounty, and weighted\_bad\_zipcounty between BORDER\_STATUS groups.
  + Differences in means of the same variables between TYPE groups (e.g., PO Box vs Standard).

**5. Saving the Results:**

* Saves summary statistics and t-test results in an Excel file for further review.

**File Output**

**VDSS\_Summary\_Statistics\_and\_T\_Tests.xlsx:**

* Contains:
  + Summary statistics by BORDER\_STATUS and TYPE.
  + Pairwise t-test results for each variable grouped by BORDER\_STATUS and TYPE.

**Usage**

1. Place input datasets in the specified directories.
2. Install required R libraries.
3. Execute the script in an R project environment.
4. Access the output file in the Data Outputs/VDSS\_Zip\_Border\_Status\_Type\_Bad\_ZipCounty\_Rate/ directory.

**Notes**

* Ensure the input datasets are complete and correctly formatted.
* T-tests compare the means of groups to determine statistical significance. Interpret results carefully, especially for small sample sizes.
* The accuracy of the analysis depends on the correctness of the merged datasets.