**VDSS\_Final\_Data.R - README**

**Author: Dylan Craig**

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

**File Name: VDSS\_Final\_Data.R**

**Purpose**

This script processes and integrates multiple datasets to create a final dataset with ZIP classification, ZCTA crosswalk, treatment, and quality data. It generates a comprehensive output, including summary statistics and unique ZIP-related counts.

**Overview**

The script, **VDSS\_Final\_Data.R**, processes and merges datasets related to ZIP codes, counties, LDSS offices, and associated attributes. It combines data sources such as ZIP-County data, ZIP-LDSS office associations (zip\_treat), ZCTA crosswalks, and bad ZIP-county rate data. The output is a cleaned and consolidated dataset saved as an Excel file.

**Datasets and Tools**

**Input Files:**

1. **ZIP\_COUNTY\_032010 to ZIP\_COUNTY\_122024**: Excel files containing ZIP to county relationships.
2. **Bailey\_Helmuth\_Zip\_Changes.xlsx**: Lists ZIP codes and their effective removal dates.
3. **VDSS\_Zip\_Type.xlsx**: Contains ZIP code classifications.
4. **HRSA\_ZCTA.xlsx**: Crosswalk linking ZIP codes to ZCTAs.
5. **bad\_zipcounty\_rate.dta**: Stata file with bad ZIP-county rate data.
6. **zip\_treat.dta**: Stata file with ZIP-LDSS office associations.

**Output File:**

* **Final\_Data.xlsx**: Includes the processed dataset, summary statistics, and unique ZIP-related counts.

**Dependencies:**

* **R Libraries:**
  + dplyr: Data manipulation
  + readxl: Reading Excel files
  + haven: Reading Stata files
  + writexl: Writing Excel files
  + stringr: String operations
  + lubridate: Date operations

**Script Functionality**

**1. Loading and Preprocessing Data:**

* **ZIP-County Data:** Processes multiple files, converts column names to uppercase, and associates time periods.
* **Bailey Helmuth ZIP Changes:** Filters out deleted ZIP codes based on effective dates.
* **ZIP Type Data:** Adds ZIP classifications (Standard, Unique, PO Box).
* **ZCTA Crosswalk:** Associates each ZIP with its corresponding ZCTA.
* **Additional Data:** Processes bad ZIP-county rate and ZIP-LDSS office associations.

**2. Data Integration:**

* Combines datasets using multiple joins to align ZIP-County data with other attributes (e.g., LDSS office).
* Ensures chronological clarity and comprehensive coverage.

**3. Classifications:**

* **Residential Status:** Classifies ZIPs as "Residential" or "Non-Residential" based on RES\_RATIO.
* **ZIP Type:** Differentiates ZIPs into categories (Standard, Unique, PO Box).

**4. Filtering for Unique Combinations:**

* Retains distinct combinations of key variables such as ZIP, COUNTY, and MONTH\_YEAR.

**5. Generating Summary Statistics:**

* Calculates:
  + Total unique ZIPs.
  + ZIP-county combinations by residential classification.
  + Monthly counts of unique ZIPs and ZIP-county combinations.

**6. Saving Output:**

* Outputs processed data and summary statistics as **Final\_Data.xlsx**.

**File Output**

**Final\_Data.xlsx**:

* **Sheets:**
  1. Final\_Data\_With\_ZCTA\_and\_Treat: Processed data.
  2. Summary\_Statistics: Summary statistics.
  3. Unique\_ZIP\_Counts: Monthly counts of unique ZIPs.
  4. Unique\_ZIP\_County\_Counts: Monthly counts of ZIP-county combinations.

**Usage**

1. Place all input files in the appropriate directories.
2. Install required R libraries.
3. Run the script within the designated R project environment.
4. The output will be saved to the specified folder.

**Notes**

* Ensure input files are named and located correctly.
* Check for any missing data before running the script.
* Results depend on the accuracy and completeness of the source files.