Memorandum

To: Derek Wu, Neil Cholli  
From: Dylan Craig  
Date Created: 12/27/2024  
Subject: Data Management Memo on ZIP-County Relationships and ZCTA Integration (VDSS)

**Purpose:** This memo explains the data cleaning and management process for analyzing ZIP-county relationships, integrating ZCTA crosswalks, classifying ZIPs, and merging additional treatment and quality metrics.

**Overview:** The final dataset consolidates ZIP-County relationships, ZCTAs, residential classifications, and treatment data. The unit of analysis is a unique combination of ZIP, county, month-year, and the LDSS ZIP code.

**Step 1: Processing ZIP-County Data (Monthly 2000–2024)**

Monthly datasets spanning 2000–2024 were processed to create a consistent format. Year-Quarter dates were extracted from file names and standardized as MM/01/YYYY. Only Virginia counties (FIPS codes starting with "51") were retained. Column names were standardized to uppercase, and extraneous columns were dropped. Any zip code that appeared at least once in the dataset was added to each Year-Quarter to ensure completeness of the dataset for each Year-Quarter, with corrections made in Step 2.

**Step 2: Removing Deleted ZIPs**

The Bailey Helmuth dataset identified ZIPs slated for deletion. These were matched with the ZIP-County data, and ZIPs with deletion dates earlier than their observed use in the dataset were excluded to ensure historical accuracy. This ended up only including 2 zip codes.

**Step 3: Classifying ZIPs as Residential or Non-Residential**

Using the VDSS ZIP type dataset, ZIPs were classified based on their residential ratio:

* “RESIDENTIAL”: Residential ratio > 0.
* “NON\_RESIDENTIAL”: Residential ratio = 0.

**Step 4: Classifying ZIPs as Standard, Unique, or PO Box**

Using the zipdatamaps.com dataset, zip codes were classified as the following:

* Standard: Typical address that can be delivered to by USPS
* Unique: High volume atypical address (e.g., a university, government building, business)
* PO Box: Areas served only by a PO Box (typically more rural; street delivery not possible)

**Step 5: Adding ZCTA Crosswalk Data**

The HRSA ZCTA crosswalk was used to map ZIPs to ZCTAs. Only ZIP and ZCTA columns were retained from the HRSA ZCTA dataset.

**Step 6: Integrating Treatment and Quality Data**

Two additional datasets were integrated:

1. **Treatment Flags (provided by Neil Cholli):** ZIP-to-LDSS office associations (e.g., above/below median distances).
2. **Bad ZIP-County Rates (provided by Neil Cholli):** Quality metrics for ZIP-County associations.

These datasets were cleaned, and data types were standardized to ensure seamless joins by ZIP codes.

**Step 7: Preparing the Final Dataset**

The dataset was filtered to retain unique combinations of key variables:

* ZIP, county, month-year, ZCTA, residential classification, LDSS ZIP code, and treatment metrics.

Summary tables were created to track:

* Total unique ZIPs and ZIP-County pairs.
* Residential vs. non-residential classifications.
* Monthly trends in ZIP-County data coverage.

**Final Unit of Analysis:**

The unit of analysis is a unique combination of ZIP, county, month-year, and the LDSS ZIP code. This allows for analyzing trends over time with precise spatial detail, including LDSS-specific relationships.

**Outputs:**

1. **Cleaned Dataset:** A comprehensive Excel file including ZIP-County relationships, ZCTAs, residential classifications, LDSS ZIP codes, and treatment metrics.
2. **Summary Statistics:** Tables summarizing ZIP trends and ZIP-County pair counts by month and classification.
3. **README:** Documentation detailing the cleaning steps, assumptions, and key data sources.