**Missing Data Analysis and Imputation Report**

**1. Introduction**

This report provides a comprehensive review of the missing data problem within the HDPSA datasets and details the strategies implemented to resolve it. Missing values are a common issue in survey-based and demographic datasets, and if left unaddressed, they can compromise the accuracy, reliability, and validity of subsequent analysis. The work presented here forms part of the Data Preparation phase of the CRISP-DM methodology and aims to ensure that the data is robust, consistent, and ready for advanced modeling.

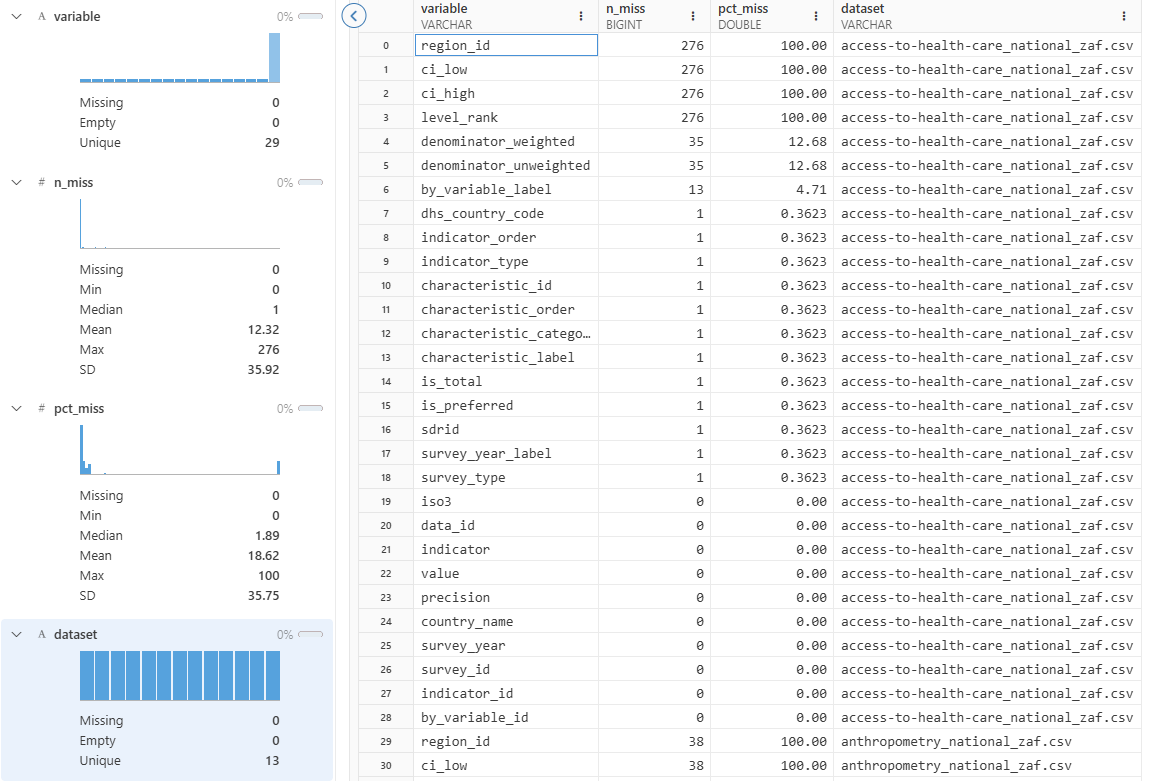
**2. Patterns of Missing Data**

The missing data investigation began with profiling all 13 datasets stored in the raw data directory. The analysis employed both numerical summaries and graphical outputs to uncover the extent and structure of missingness.

**2.1 Variable-Level Analysis**

Each dataset was examined at a variable level to measure the proportion of missing values per attribute. This analysis identified variables with negligible missingness as well as variables with significant gaps. For example:

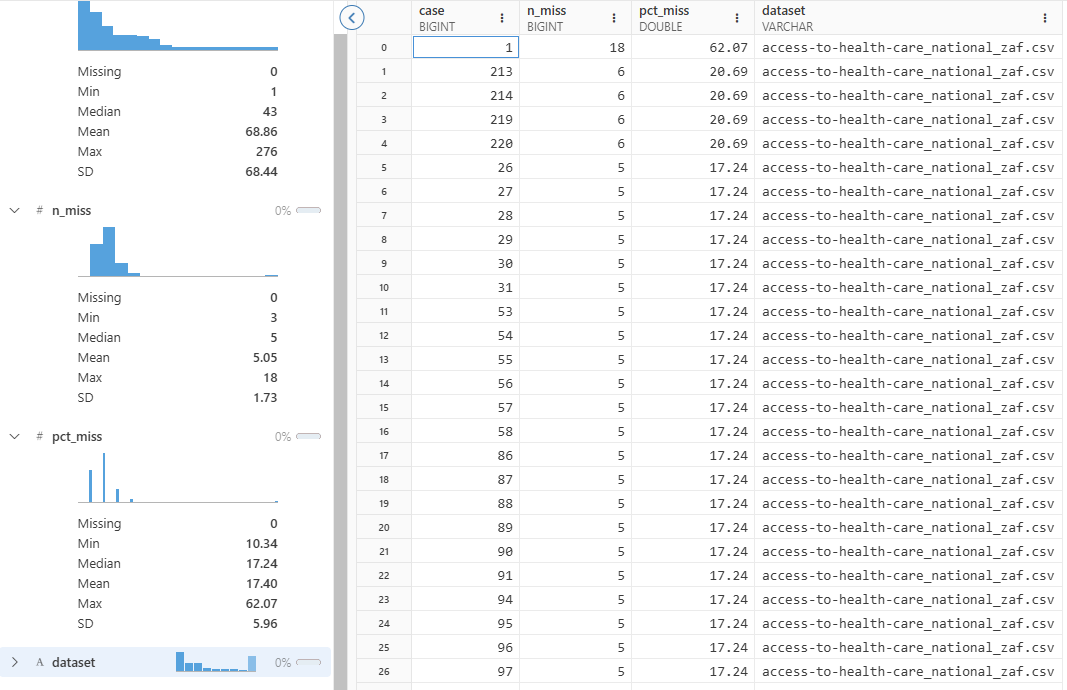
* Some categorical variables contained placeholders such as "99", "999", "Don’t know", or "Refused", which were standardized into NA.
* Certain percentage-based variables exhibited higher missingness due to incomplete survey years.



*Table 1: Missingness by Variable (generated output from missingness\_by\_variable\_all.csv).*

**2.2 Case-Level Analysis**

At the case level, the analysis revealed that certain respondents had multiple fields missing simultaneously. These cases represent incomplete survey records. Instead of discarding them, which would reduce the sample size and distort representativeness, they were retained for imputation.

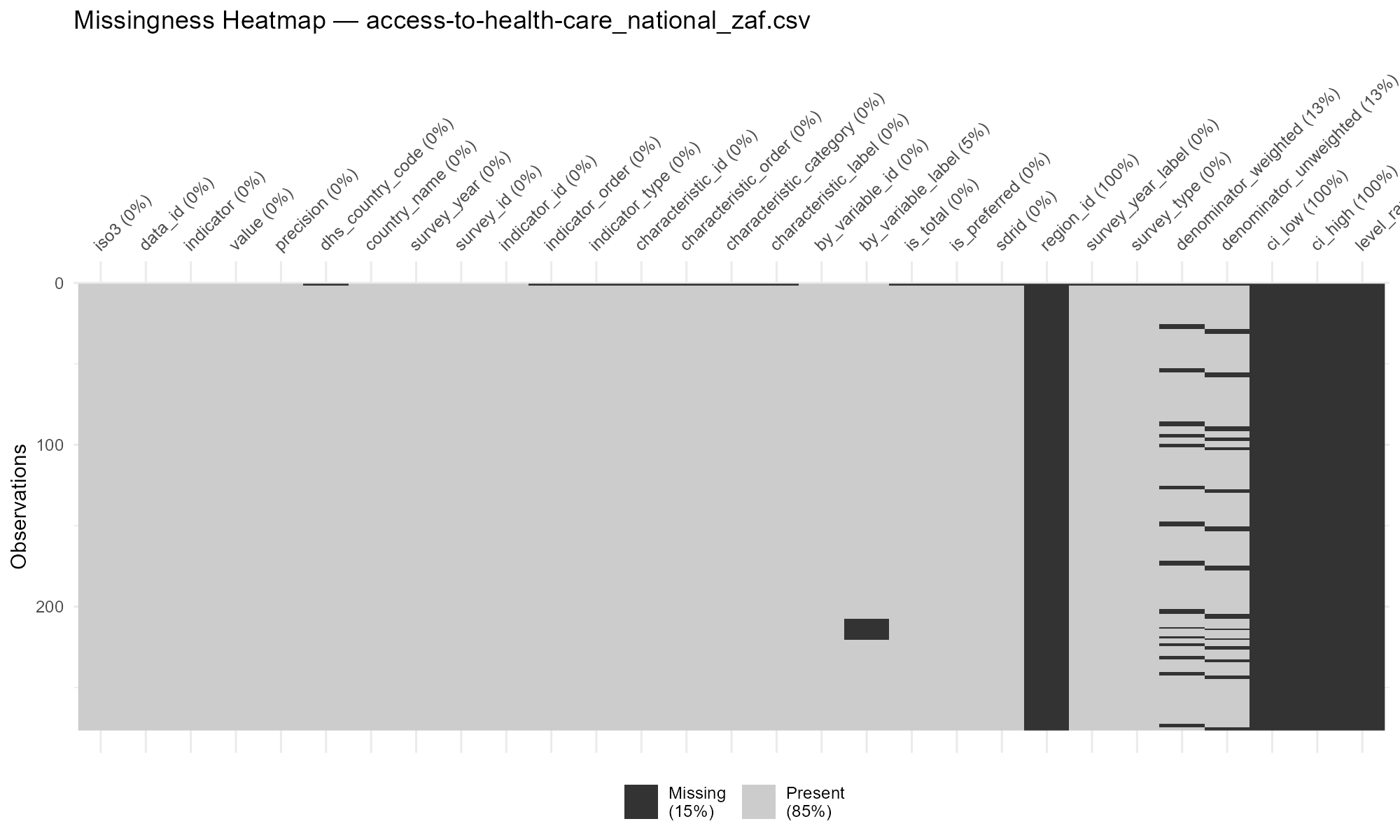


*Table 2: Missingness by Case (generated output from missingness\_by\_case\_all.csv).*

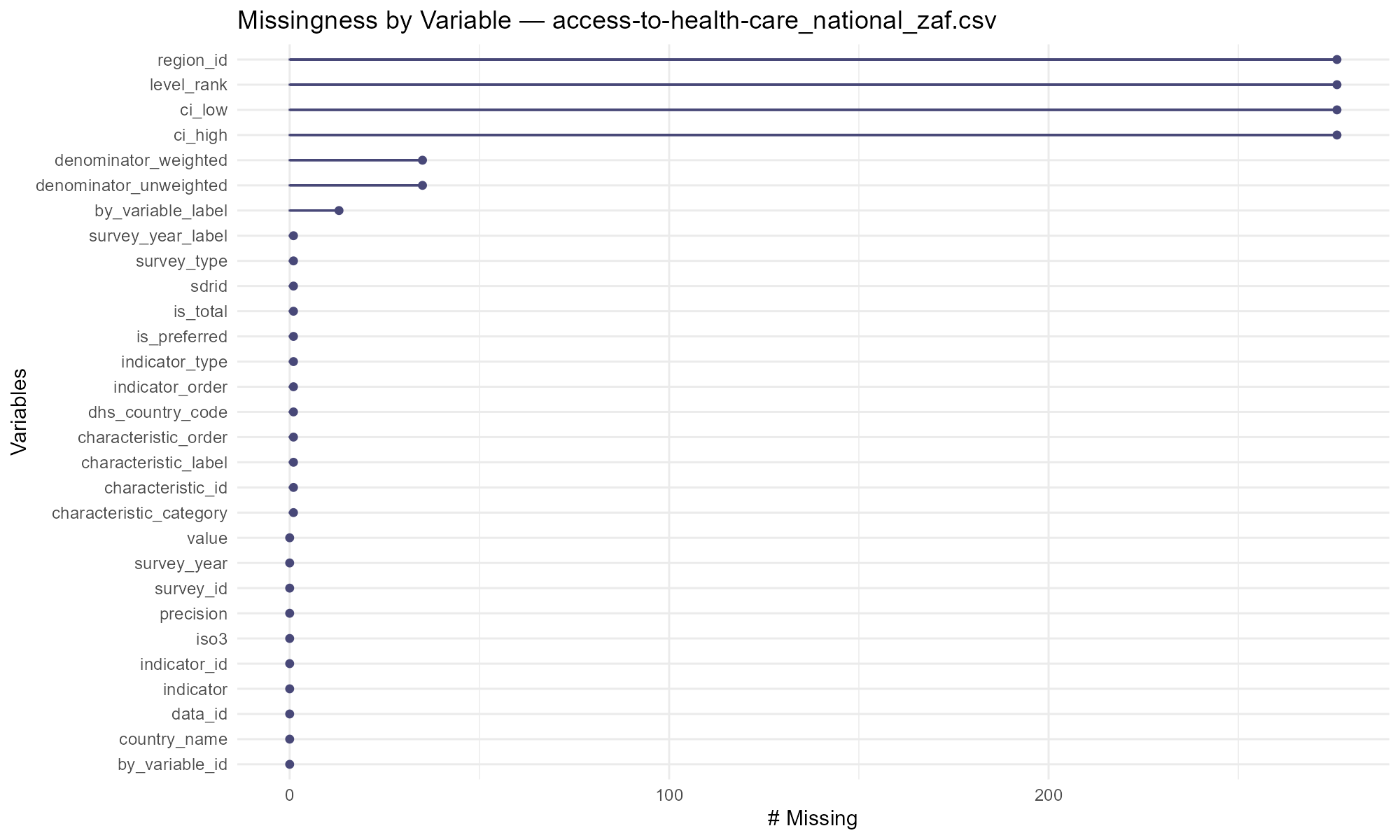
**2.3 Visual Exploration of Missingness**

Several diagnostic visuals were produced to highlight the structure of missingness:

* **Heatmaps** showed block patterns where entire groups of variables were absent for specific respondents.
* **Bar charts** ranked variables by their percentage of missingness.
* **UpSet plots** identified intersections of variables that were frequently missing together, highlighting structured dependencies.

**

*Figure 1: Missingness Heatmap (R script output).*

**  
*Figure 2: UpSet Plot of Joint Missingness (R script output).*

These findings demonstrated that missingness was not random but often systematic, influenced by survey design and respondent behavior.

**3. Imputation Strategies**

Imputation strategies were carefully chosen to address the identified patterns while preserving data integrity and usability. A hybrid framework was adopted, applying different methods depending on the data type.

**3.1 Categorical Variables**

For categorical attributes:

* Missing values were replaced with the **mode** (most frequent response), preserving distributional balance.
* Where no clear mode was available, an **“Unknown”** category was introduced to ensure that all records remained complete.

**3.2 Numeric Percentage Variables**

For survey-derived percentage variables:

* **Linear interpolation** was used where a time component was present. This leveraged year-to-year continuity to estimate plausible values for missing points.
* Values were clamped between 0 and 100 to prevent unrealistic results.

**3.3 Other Numeric Variables**

For general numeric variables:

* **Multiple Imputation by Chained Equations (MICE)** was used with Predictive Mean Matching, ensuring that values were imputed based on observed relationships in the data.
* In instances where MICE could not be applied effectively, a fallback **median imputation** was used to maintain central tendency while minimizing bias.

*A screenshot of a computer

AI-generated content may be incorrect.*

*Table 3: Imputation Plan by Variable (generated output from imputation\_plan\_all.csv).*

This approach ensured that every variable had a documented treatment plan, increasing transparency and reproducibility.

**4. Impact of Imputation**

The impact of the imputation strategies was measured both quantitatively and visually.

**4.1 Reduction in Missingness**

After imputation, the datasets were free of missing values. Attributes that previously contained high proportions of missing data were completed, ensuring no further information was lost during modeling.

*A screenshot of a computer

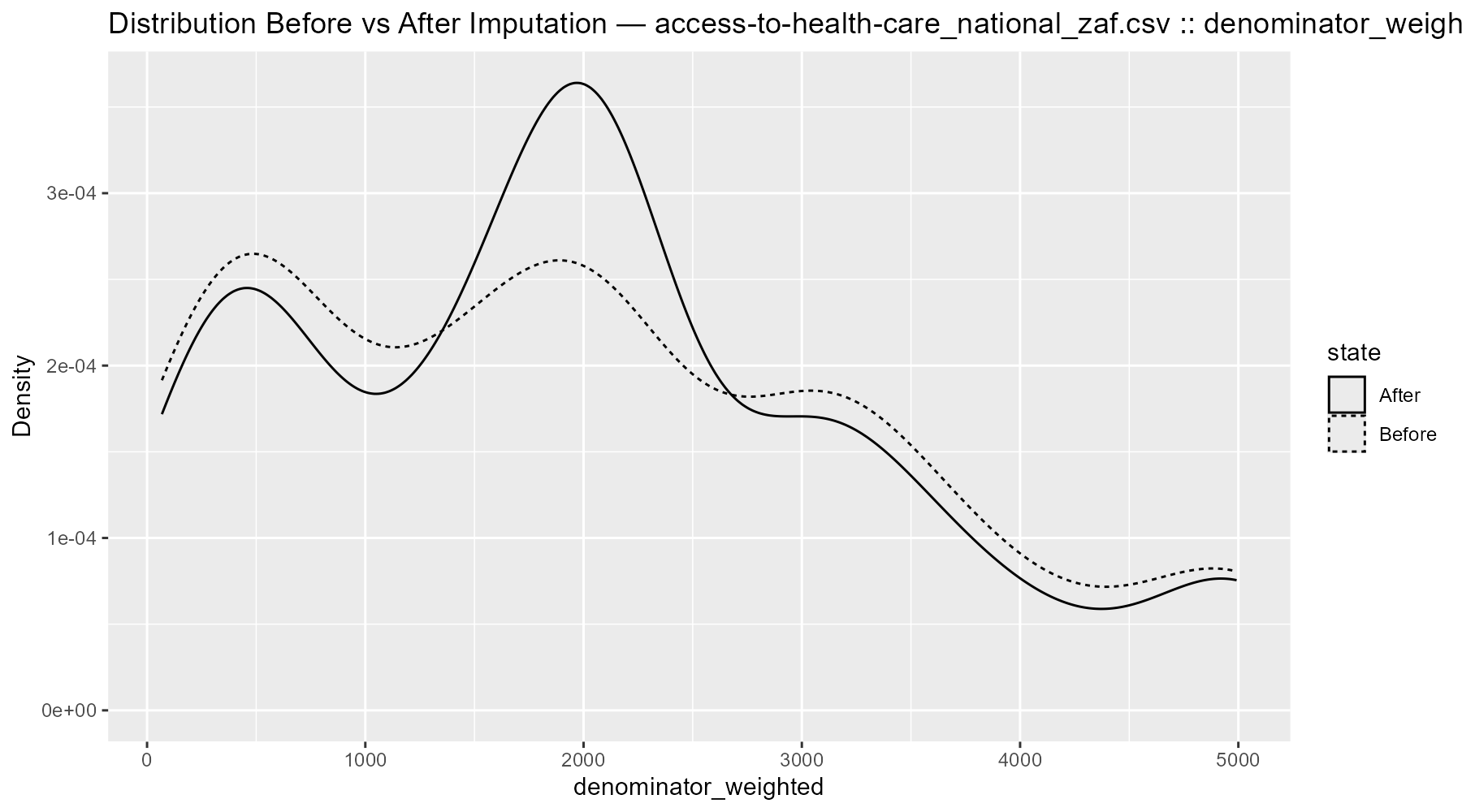
AI-generated content may be incorrect.*

*Table 4: Imputation Log Before vs After Missing Counts (generated output from imputation\_log\_all.csv).*

**4.2 Distributional Effects**

Density plots and comparative checks confirmed that the imputed data aligned closely with the original distributions:

* Median imputation preserved the central point but slightly reduced variability.
* MICE imputations produced values consistent with the underlying data structure, maintaining realistic variance.
* Interpolation preserved temporal trends and continuity in time-series data but naturally smoothed anomalies.

**

*Figure 3: Distribution Before vs After Imputation (example variable, R script output).*

**4.3 Considerations and Limitations**

* Mode imputation can increase the frequency of dominant categories, reducing diversity.
* Interpolation smooths outliers, which may mask genuine anomalies.
* MICE imputations depend on the accuracy of the predictive models, which may introduce synthetic correlations.

Despite these considerations, the chosen strategies provided a practical balance between completeness and representativeness.

**5. Conclusion**

The analysis confirmed that missingness across the HDPSA datasets was structured and non-random, often resulting from survey design choices and non-response behavior. The hybrid imputation strategy implemented combining mode, interpolation, MICE, and median imputation effectively resolved these gaps.

The cleaned and imputed datasets are now:

* Complete, with no missing values.
* Consistent, preserving the underlying distributions and relationships.
* Well-documented, with logs, imputation plans, and visual evidence supporting transparency.

This ensures that the data is ready for the next stages of the project, including attribute selection, modeling, and evaluation, with a solid foundation of data quality and integrity.