Uncovering Health Inequities: A BI Exploration of Literacy, ARI Symptoms, and Maternal Mortality

BIN 371 Milestone 2

Suhil Jugroop Makhosandile Nkosi Nkosinathi Remoratile Sibusiso Mogorosi Michael Anyimadu

Contents

[1. Quick overview and goals 2](#_Toc206000007)

[2. Datasets and basic dimensions 2](#_Toc206000008)

[3. Selection rules and rationale 2](#_Toc206000009)

[4. Standardization and parsing 2](#_Toc206000010)

[5. Cleaning and imputation approach 2](#_Toc206000011)

[6. Missingness and quality snapshot after cleaning 3](#_Toc206000012)

[7. Feature selection testing and interpretation 3](#_Toc206000013)

[8. Transformations and data split 4](#_Toc206000014)

[9. Limitations and recommended next steps 4](#_Toc206000015)

[10. Short conclusion 4](#_Toc206000016)

# 1. Quick overview and goals

The objective was to take four DHS style datasets, get them into a clean and consistent shape, handle missing values sensibly, run simple feature selection tests, and prepare design matrices for modelling. The four datasets are Anthropometry, Literacy, Maternal Mortality, and ARI Symptoms. The outputs are cleaned CSVs and 70/30 training and test sets saved for each dataset that had an eligible response variable.

# 2. Datasets and basic dimensions

Each dataset was loaded from CSV. The row and column count before any selection were:

1. Anthropometry: 38 rows, 30 columns.
2. Literacy: 21 rows, 30 columns.
3. Maternal Mortality: 22 rows, 30 columns.
4. ARI Symptoms: 27 rows, 30 columns.

A short preview of the first eight column names shows a consistent structure across datasets with fields like ISO3, DataId, Indicator, Value, Precision, CountryName, and SurveyYear present. These shapes and previews are reported in the uploaded Milestone 2 document.

# 3. Selection rules and rationale

Columns kept if present were Indicator, Value, Value num, Sex, Age, Region, Period, and Subgroup. Any free text comments and near-constant columns were removed. Columns with more than 40 percent missing were dropped unless they were considered essential. I think this is a reasonable balance. It keeps the analysis focused on meaningful, mostly populated variables while preserving indicators and the numeric response that we want to model.

The 40 percent threshold is a pragmatic choice. It reduces noise from sparsely populated fields but leaves room for variables that may be important even if somewhat incomplete. An alternative would be to use a stricter threshold or to impute more aggressively. For this stage, the chosen rule gives a good compromise between data retention and quality.

# 4. Standardization and parsing

Column names were standardized by removing spaces and non-alphanumeric characters and replacing spaces with underscores. The Value field was parsed into a numeric column named Value\_num that handles formats like percentages and numbers with commas. The code reports one non-numeric Value entry per dataset before parsing. This step is important because downstream numeric operations rely on a clean numeric response column.

# 5. Cleaning and imputation approach

The cleaning pipeline did the following main steps.

1. Exact duplicates were removed.
2. Missing values for Value\_num were imputed using a hierarchical median approach.
   1. First attempt was median within the cell defined by Indicator and Sex.
   2. If that was not available, then median within Indicator.
   3. If still missing, then global median fallback.
3. Winsorization at the 1st and 99th percentiles were applied optionally for datasets with more than 10 non-missing observations, producing Value\_winz.
4. Categorical fields were cast to factors and a scaled response Value scaled was created for modelling.

The hierarchical median imputation is conservative and robust to skew. It uses domain structure first which helps keep imputations context specific. Using global median only as a last resort prevents extreme distortions when group information is missing. Winsorization is optional but useful to reduce the effect of extreme outliers before scaling.

# 6. Missingness and quality snapshot after cleaning

After cleaning the total missing counts were reduced to zero for the sets reported, meaning all required fields were filled or appropriately handled. The cleaned row counts remained the same as original because rows were not dropped unless duplicate. The reported outlier counts before winsorization on Value num were:

1. Anthropometry had 8 IQR outliers.
2. Literacy had 4 IQR outliers.
3. Maternal Mortality had 4 IQR outliers.
4. ARI Symptoms had 0 IQR outliers.

These counts justify the optional winsorization step for the datasets with notable outliers.

# 7. Feature selection testing and interpretation

Two simple statistical tests were run to identify candidate predictors.

1. Spearman correlation for numeric predictors against Value\_scaled. No eligible numeric predictors were present for Spearman in the datasets, so that analysis returned no significant results.
2. Kruskal Wallis tests for categorical predictors against Value\_scaled. Indicator stood out as the most significant categorical predictor in ARI Symptoms with a p-value around 0.0018. Maternal Mortality also showed Indicator with a p-value near 0.045. Anthropometry and Literacy did not show strong categorical associations in this test.

Interpretation  
 Indicator being significant in ARI Symptoms and Maternal Mortality suggests that the particular measure being reported is an important determinant of the scaled response. This is reasonable because many DHS indicators capture different health constructs and are expected to vary in level. The absence of numeric predictors limits the types of multivariate numeric analysis possible at this stage. Converting Period into a numeric proxy was considered but is noisy because it encodes factor levels rather than true numeric years.

# 8. Transformations and data split

For datasets with Value scaled a one-hot encoded design matrix was created using model matrix expansion. The pipeline generated training and testing sets with a 70/30 split using a fixed random seed for reproducibility. These objects were saved as RDS files and the cleaned CSVs were output to the outputs\_m2 directory. The Milestone note records the output path used on the developer machine.

# 9. Limitations and recommended next steps

Limitations to be mindful of:

1. Small sample sizes reduce statistical power and make many numeric tests inapplicable.
2. Period was encoded as factor to numeric levels which can misrepresent time. If true survey years are available, they should be parsed as integers.
3. Indicator is categorical and often multi-level. Treating it as a predictor via one-hot encoding is fine for simple models but can lead to high dimensionality for small N.

Recommendations:

1. If possible, augment datasets or combine related indicators to increase sample size.
2. Replace factor-encoded period with actual numeric survey year for meaningful temporal analysis.
3. Run simple regularized models such as Lasso or ridge on the one-hot expanded design matrices to identify stable predictors.
4. Report model stability via cross-validation and present effect sizes with confidence intervals rather than relying only on univariate tests.

# 10. Short conclusion

The Milestone 2 script does a solid, pragmatic job of preparing DHS-style datasets for modeling. The cleaning and imputation strategy is sensible for a project at this scale, and the early feature selection points to Indicator as a notable categorical driver in some datasets.