belgium campus itversity

Project Milestone 2

**Module:** Business Intelligence 381  
**Project:** Health & Demographic Patterns in South Africa (HDPSA)  
**Milestone 2:** Data Preparation

Petrus Human 577842

Frederik Knoetze 600965

Teleki Shai 601377

Moloko Rakumako 601352

Table of Contents

[Introduction 2](#_Toc209205914)

[Overall Project Goals 2](#_Toc209205915)

[Milestone 2 Data Preparation Goals 2](#_Toc209205916)

[R Script Descriptions 3](#_Toc209205917)

[Milestone\_2\_combine\_datasets.R 3](#_Toc209205918)

[Milestone\_2\_data\_cleaning.R 3](#_Toc209205919)

[Milestone\_2\_feature\_selection.R 4](#_Toc209205920)

[Data Transformation: Column Comparison 5](#_Toc209205921)

[Raw Data Columns Example from access-to-health-care\_national\_zaf.csv 5](#_Toc209205922)

[Final Processed Data Columns Example from feature\_selected\_cleaned\_combined\_dataset.csv 6](#_Toc209205923)

[Columns Removed During Processing 6](#_Toc209205924)

[Rationale for Final Column Selection 7](#_Toc209205925)

[Data Preparation Considerations for Modelling 8](#_Toc209205926)

[Train-Test Split Strategy with Limited Time Points 8](#_Toc209205927)

[Data Scaling and Normalization 9](#_Toc209205928)

[Conclusion 10](#_Toc209205929)

Introduction

This document provides an overview of the R scripts developed for the Health & Demographic Patterns in South Africa (HDPSA) project, specifically addressing the data preparation phase for Milestone 2. These scripts are designed to transform raw data into a clean, pre-processed, and feature-selected format suitable for future modelling and data analysis, aligning with the project's goals for business understanding and data quality.

# Overall Project Goals

The primary data mining goals for the HDPSA project are to identify variables correlating with child mortality, understand how changes in variables over time impact child mortality, and create a model to predict child mortality. Key success criteria for data preparation (Phase 2) include:

* Data is cleaned and pre-processed (missing values handled, duplicates removed, outliers addressed).\*\* This directly aligns with the functionality of the R scripts.
* Model-ready datasets with <5% unprofiled missingness on critical KPIs. (This is an outcome goal that the scripts contribute to).
* The Milestone 1 analysis highlighted data quality issues such as missing data (14.98% to 23.67% range), systematic missing patterns, and the need for careful handling of duplicates and outliers.

Milestone 2 Data Preparation Goals

Based on the Milestone 2 instructions, the R scripts contribute significantly to the following key objectives:

* \*\*Select Data:\*\* This involves focusing on indicators with sufficient data coverage and utilizing `IsPreferred` flags for data selection.
* Verify Data Quality:\*\* This encompasses implementing appropriate missing data treatment, standardizing formats, validating ranges, and ensuring temporal consistency.

R Script Descriptions

Milestone\_2\_combine\_datasets.R

Purpose: To consolidate all individual raw CSV datasets into a single, unified dataset, ensuring a consistent and clean starting point for further processing.

Functionality:

* Reads all CSV files located in the `Raw Dataset’s folder.
* Identifies and extracts the true header (column names) from the first CSV file.
* Reads all subsequent CSV files, explicitly skipping the first two rows (the actual header and a metadata row present in each source file) to prevent duplicate headers from becoming data rows.
* Combines all data into a single data frame.
* Assigns the extracted true header to the combined data.
* Removes any existing `combined\_dataset.csv` to ensure a fresh output.
* Saves the resulting unified dataset as `combined\_dataset.csv` in the `Cleaned Dataset’s folder.

Relation to Data Quality (Milestone 2): This script is foundational for data quality by ensuring all raw data is brought together consistently. It directly addresses the "Standardize formats" requirement by handling disparate raw files and resolving the initial structural quality issue of multiple header rows in source files, thus preparing the data for further quality verification. It also contributes to "Temporal consistency" by combining data across different survey years into a single structure.

## Milestone\_2\_data\_cleaning.R

Purpose: To clean and preprocess the combined dataset, addressing critical data quality issues as outlined in Milestone 2 instructions.

Functionality:

* Reads the `combined\_dataset.csv` from the `Cleaned Dataset’s folder.
* Removes Duplicate Rows: Directly addresses the "duplicates removed" success criteria by identifying and eliminating exact duplicate rows across the entire dataset.
* Imputes Missing Values: Implements a "Missing data strategy" and "Handle missing values" by performing imputation for missing data points. For numerical columns, missing values are replaced with the mean of that columnwithin groups defined by 'Indicator' and 'SurveyYear. For categorical columns, missing values are replaced with the mode (most frequent value) within groups defined by 'Indicator' and 'SurveyYear'\*This group-wise approach ensures context-aware imputation.
* Handles Outliers: Directly addresses the "outliers addressed" success criteria. Outliers in numerical columns are capped using the Interquartile Range (IQR) method (values falling outside 1.5 times the IQR from the first (Q1) or third (Q3) quartiles are capped at the respective bounds). This capping is performed within groups defined by 'Indicator' and 'SurveyYear', contributing to "Validate ranges" by ensuring data falls within reasonable, group-specific limits.
* Saves the thoroughly cleaned and preprocessed dataset as `cleaned\_combined\_dataset.csv` in the `Cleaned Datasets` folder.

Relation to Data Quality (Milestone 2): This script is central to the "Verify Data Quality" objective. It systematically handles missing values, duplicates, and outliers, ensuring the dataset is robust, reliable, and adheres to the quality standards required for analysis. The group-wise cleaning methods enhance the integrity and contextual accuracy of the data.

## Milestone\_2\_feature\_selection.R

Purpose: To select a subset of features (columns) and rows that are most relevant for future modelling and data analysis, based on predefined criteria from Milestone 2.

Functionality:

* Reads the `cleaned\_combined\_dataset.csv` from the `Cleaned Dataset’s folder.
* Filters Rows: Directly implements the "Quality flags: Utilize IsPreferred flags for data selection" and "Indicator selection: Focus on indicators with sufficient data coverage" requirements by retaining only those rows where the `IsPreferred` column has a value of `1`. This ensures that only preferred and relevant data points are carried forward.
* Selects Columns:\*\* Keeps only the `Indicator`, `Value`, and `SurveyYear` columns. This aligns with focusing on key variables for analysis and reducing dimensionality, contributing to efficient data selection.
* Saves the filtered and column-selected dataset as `feature\_selected\_cleaned\_combined\_dataset.csv` in the `Cleaned Datasets` folder.

Relation to Data Selection (Milestone 2): This script directly addresses the "Select Data" requirement by focusing on preferred indicators and reducing the dataset to essential variables. By using the `IsPreferred` flag, it ensures that the selected data is of higher quality and relevance for subsequent modelling phases.

# Data Transformation: Column Comparison

To illustrate the data transformation process, a comparison between the columns of a raw dataset and the final `feature\_selected\_cleaned\_combined\_dataset.csv` is provided below. This highlights the columns that have been retained and those that were removed during the data preparation and feature selection phases.

## Raw Data Columns Example from access-to-health-care\_national\_zaf.csv

* `ISO3`
* `DataId`
* `Indicator`
* `Value`
* `Precision`
* `DHS\_CountryCode`
* `CountryName`
* `SurveyYear`
* `SurveyId`
* `IndicatorId`
* `IndicatorOrder`
* `IndicatorType`
* `CharacteristicId`
* `CharacteristicOrder`
* `CharacteristicCategory`
* `CharacteristicLabel`
* `ByVariableId`
* `ByVariableLabel`
* `IsTotal`
* `IsPreferred`
* `SDRID`
* `RegionId`
* `SurveyYearLabel`
* `SurveyType`
* `DenominatorWeighted`
* `DenominatorUnweighted`
* `CILow`
* `CIHigh`
* `LevelRank`

## Final Processed Data Columns Example from feature\_selected\_cleaned\_combined\_dataset.csv

* `Indicator`
* `Value`
* `SurveyYear`

## Columns Removed During Processing

The following columns were removed during the data cleaning and feature selection process, primarily by the `Milestone\_2\_feature\_selection.R` script, which specifically selected for `Indicator`, `Value`, and `SurveyYear` after filtering by `IsPreferred = 1`.

* `ISO3`
* `DataId`
* `Precision`
* `DHS\_CountryCode`
* `CountryName`
* `SurveyId`
* `IndicatorId`
* `IndicatorOrder`
* `IndicatorType`
* `CharacteristicId`
* `CharacteristicOrder`
* `CharacteristicCategory`
* `CharacteristicLabel`
* `ByVariableId`
* `ByVariableLabel`
* `IsTotal`
* `IsPreferred`
* `SDRID`
* `RegionId`
* `SurveyYearLabel`
* `SurveyType`
* `DenominatorWeighted`
* `DenominatorUnweighted`
* `CILow`
* `CIHigh`
* `LevelRank`

## Rationale for Final Column Selection

The decision to retain only the `Indicator`, `Value`, and `SurveyYear` columns in the final processed dataset (`feature\_selected\_cleaned\_combined\_dataset.csv`) was driven by a combination of visual analysis, the project's core objectives, and the identification of redundant or less critical information for the primary modeling task.

Key reasons for this focused selection include:

* Redundancy of Country-Specific Identifiers:\*\* Columns such as `ISO3`, `DHS\_CountryCode`, and `CountryName` all specify the country as South Africa. Given that the project's scope is exclusively focused on South Africa, these columns provide redundant information and do not contribute unique variance to the analysis. Their removal simplifies the dataset without losing relevant geographical context.
* Focus on Core Measurement and Temporal Trends: The `Indicator` column clearly defines *\*what\** health or demographic aspect is being measured. The `Value` column represents *\*the actual measurement\** for that indicator. The `SurveyYear` column provides the *\*temporal context\** of the measurement. These three attributes are fundamental for understanding trends, patterns, and relationships over time for specific indicators, which aligns directly with the project's goal of analyzing health and demographic patterns.
* Minimizing Noise and Irrelevant Metadata: Many other columns, including various IDs (`DataId`, `SurveyId`, `IndicatorId`, `CharacteristicId`, `ByVariableId`, `SDRID`, `RegionId`), orderings (`IndicatorOrder`, `CharacteristicOrder`), and descriptive metadata (`IndicatorType`, `CharacteristicCategory`, `CharacteristicLabel`, `ByVariableLabel`, `SurveyYearLabel`, `SurveyType`), while useful for data management or detailed exploration, were deemed less critical for the direct modeling of `Value` against `SurveyYear`. Including them could introduce unnecessary complexity or noise without significant analytical gain for the primary objective.
* Efficiency and Model Simplicity: By focusing on the most pertinent variables, the dataset becomes more streamlined, which can lead to more efficient model training and easier interpretation of results. This aligns with the principle of parsimony in model building.
* Visual Analysis Insights: Initial visual analysis of the raw and combined datasets revealed that many of the removed columns exhibited little variance or direct correlation with the `Value` when considered in the context of `SurveyYear` and `Indicator`. This visual evidence supported the decision to narrow down the feature set to the most impactful and non-redundant attributes.

This selective approach ensures that the final dataset is lean, focused, and optimized for the subsequent modelling and data analysis phases, directly supporting the project's objectives of identifying key patterns and predicting child mortality.

Data Preparation Considerations for Modelling

## Train-Test Split Strategy with Limited Time Points

Given that the dataset contains only two distinct `SurveyYear` values, a traditional random train-test split for predictive modelling is not recommended, as it risks severe data leakage and unreliable model evaluation. Instead, for any predictive modelling tasks, a \*\*time-series validation approach\*\* is most appropriate.

My professional recommendation would be to:

1.  Utilize the earlier `SurveyYear` data for training your model.

2.  Utilize the later `SurveyYear` data exclusively for testing and evaluating your model's predictive performance.

This approach respects the temporal nature of the data, allowing you to assess how well a model trained on historical patterns can predict outcomes in a subsequent period.

Important Considerations and Limitations:

* Limited Generalizability: With only two time points, the model's ability to generalize to *\*future\** years beyond the test year will be highly constrained. The evaluation results will be specific to the transition from the earlier to the later year.
* Data Scarcity for Training: Training a model on data from only a single year might limit its ability to learn complex patterns or robust relationships.
* Focus on Time-Series Analysis: If the primary objective is to understand historical trends, patterns, and relationships within the data (descriptive time-series analysis) rather than making robust predictions, then using the entire dataset for analysis without a formal train-test split is appropriate. However, if prediction is the goal, some form of temporal validation is essential, even with limited data.

## Data Scaling and Normalization

It is important to note that the `Value` column, which represents the core measurement for each indicator, is already expressed in percentages or thousands (as indicated by the nature of the health and demographic data). Therefore, explicit data scaling or normalization (e.g., min-max scaling, standardization) of this column is generally \*\*not necessary\*\* for most modelling algorithms. The values are already within a comparable range, and further scaling might not provide significant benefits or could even obscure the direct interpretability of the `Value` itself.

This ensures that the data remains interpretable and directly reflects the real-world magnitudes of the indicators.

# Conclusion

These three R scripts collectively form a robust data preparation pipeline for Milestone 2 of the HDPSA project. They ensure that the raw, disparate datasets are combined, thoroughly cleaned, and appropriately feature-selected, providing a high-quality foundation for in-depth data analysis and predictive modelling, in direct alignment with the project's success criteria and data preparation goals.