PROJECT MILESTONE 1

Petrus Human 577842

Frederik Knoetze 600965

Teleki Shai 601377

Moloko Rakumako 601352

Module: Business Intelligence 381
Project: Health & Demographic Patterns in South Africa (HDPSA)
Milestone: 1 – Business Understanding & Data Understanding

BELGIUM CAMPUS ITVERSITY

Table of Contents

Business Objective

Background

South Africa faces persistent, place-based biases in basic services and health outcomes—especially affecting mothers and young children. Key issues include uneven access to safe water and improved sanitation (service reliability and quality). Variable primary healthcare access and readiness (facility density/readiness), gaps in immunisation coverage and maternal/child health risks amid a large HIV burden. These factors jointly influence under-five and maternal mortality, acute respiratory infections (ARI) in children, and nutrition outcomes (Statistics South Africa, 2024; UNICEF, 2025; World Bank/WHO, 2025; HSRC, 2024).

Business Goal

For government and non-profit decision-makers, identify and explain where and why risks cluster so resources (infrastructure upgrades, community outreach, staffing, supply chain, targeted prevention) can be prioritised for the largest health gains at district/provincial level.

Stakeholders (who, why, what they need)

- **National Department of Health (NDoH)** – macro planning, NHI context, budgets, national monitoring. Needs comparable, district-ranked KPIs and trends (Reuters, 2024).
- **Provincial Health Departments & District Management Teams** – staffing, supply chain, outreach scheduling. Need drilldowns (facilities per 10k population, travel times, coverage gaps) (HST, 2024).
- **Municipalities/Water Utilities** – capital planning for water/sanitation reliability and quality safeguards. Need maps linking health risk to service backlogs (Statistics South Africa, 2024; Financial Times, 2025).
- **Hospitals/Clinics** – Ideal Clinic readiness, stockouts, referral flows. Need operational dashboards (NDoH/Ideal Clinic, 2024).
- **NGOs/Donors/CBOs** – site selection and impact tracking (HIV prevention, immunisation catch-up, IYCF) (HSRC, 2024; HSRC, 2023).
- **Communities/Leaders** – plain-language scorecards for accountability and co-design.
- **Data/Analytics teams** – reproducible pipelines; district harmonisation and data-quality checks, as per CRISP-DM Phase 2 expectations (CRISP-DM Consortium, 2000).

Stakeholder objectives summary (one-page)

Stakeholder	Decisions/Use-cases	Core KPIs	"Success looks like..."	Risks/Notes
NDoH	Provincial prioritisation; budget	U5MR, MMR, DTP3, facility density	Evidence-based provincial/district ranking & briefs	Political sensitivity; compare modelled vs facility data (World Bank/WHO, 2025; Statistics South Africa, 2022).
Provinces/Districts	Outreach & staffing	Travel time, Ideal Clinic, ARI, immunisation	Actionable district plans & monthly tracking	Data gaps; constrained HRH (NDoH/Ideal Clinic, 2024).
Municipalities	WASH investment	% improved water/sanitation; reliability	Prioritised WASH pipeline tied to health risk	Funding cycles; service interruptions (Statistics South Africa, 2024; Financial Times, 2025).
Facilities	Readiness & supply chain	Ideal Clinic, stockout proxies	Fewer stockouts, shorter queues	Under-reporting; logistics (NDoH/Ideal Clinic, 2024).
NGOs/Donors	Site selection, M&E	HIV prevalence/suppression; DTP3	Impact-ready logframes; catch-up plans	Funding volatility (HSRC, 2024; The Guardian, 2025).
Communities	Accountability	Simple scorecards	Trusted, plain-language visuals	Mistrust if indicators conflict with lived experience.

Success criteria

Phase 1: Business & Data Understanding

- Business problem and objectives for project are clearly defined.
- Stakeholders and their requirements are identified and documented.
- Provided datasets are fully explored, with variables (e.g., mortality, facility access, immunization, vaccination) described.
- Data quality issues (missing values, duplicates, outliers) are identified and recorded.
- Preliminary visualizations and dashboards reveal meaningful insights, correlations, and patterns.

Phase 2: Data Preparation

- Data is cleaned and pre-processed (missing values handled, duplicates removed, outliers addressed).
- Categorical variables are encoded, and numerical variables are discretized or scaled if necessary.
- Data is split into training and testing sets for reliable model evaluation.

Phase 3: Modelling

- Appropriate algorithm(s) are selected for the chosen analysis type (classification, clustering, etc.).
- Significant predictor variables are identified and justified using evaluation methods.
- Models are trained and fine-tuned on training data.
- Model performance metrics are clearly documented and visualized (e.g., confusion matrix, graphs).
- Model selection and parameter tuning are well explained and justified.

Phase 4: Model Evaluation

- Model performance is evaluated using appropriate metrics (accuracy, precision, recall, F1-score, Silhouette, etc.).
- Performance is benchmarked against a baseline for comparison.
- Evaluation results are visualized and interpreted in the context of the business problem.
-

Phase 5: Deployment

- Deployment options for the project are identified and discussed.

- A deployment plan is developed, addressing data input, monitoring, and updates.
- Ethical and privacy considerations are incorporated into the deployment approach.

Phase 6: Final Reporting

- The final report summarizes all phases (business problem, data exploration, preparation, modelling, evaluation, deployment).
- Key insights and recommendations are highlighted and tied back to business objectives.
- Visualizations, code snippets, and explanations support transparency and reproducibility.
- All deliverables (report, code, visualizations) are submitted in the required format.

KPIs (aligned to decision-making & literature)

- **Access**: % households with improved water; % with improved sanitation; service reliability proxy (reported interruptions) (Statistics South Africa, 2024).
- **PHC access/readiness**: Facility density per 10,000; Ideal Clinic compliance; median travel time to nearest facility (where available) (HST, 2024; NDoH/Ideal Clinic, 2024).
- **Prevention & child health**: DTP3 coverage; ARI symptom prevalence in <5s; IYCF indicators (as available) (WHO & UNICEF, 2024; Khilnani et al., 2023).
- **Outcomes**: Under-five mortality rate; maternal mortality ratio (modelled and facility measures) (UNICEF, 2025; World Bank/WHO, 2025; Statistics South Africa, 2022).
- **HIV context**: HIV prevalence (all-age and 15–49); testing & viral suppression (where available) (HSRC, 2024; HSRC, 2023).

Key business questions (for analysis plan)

- Which districts combine low WASH access with high ARI symptoms and low immunisation coverage?
- Where is facility density/readiness most misaligned with child/maternal outcomes?
- How do HIV prevalence & literacy correlate with immunisation and maternal care uptake?
- Which clusters of districts share similar composite risk profiles for targeted packages?

Accessing the Situation

Available Resources

The resources available to complete this project are in two categories: information resources and data resources. The data resources are provided and include 13 datasets. The information resources include but is not limited to the following: YouTube, textbooks, the internet, PowerPoint slides, large language models and other generative algorithms. The tools used for this project will be R Studio and Power BI.

Assumptions

The project will proceed with the following assumptions. We are to use the tools described in the project outline document and we are to only use the provided datasets to complete the task.

Constraints

The primary constraint for this project is the limited dataset. As will be alter expanded upon most of the data is at a national level with limited geographic information. Otherwise, any information that could be useful E.G level of parent's education, level of wealth, distance to hospital and more. This will limit the trends we can infer from the data to only extraction information for the data we have.

Risks

Since this is a project under the supervision of Belgium Campus ITvarsity there will be no strategic, financial or compliance risk from our end. Thus, only operational, technical and team related risks remain.

Project Management Risks

- Missed deadlines: If a team member misses a deadline for work another is waiting for this could cause a missed milestone.
- Poor task allocation: If work is unevenly distributed (even unintentionally) team members could get frustrated.
- Coordination: If the team does not effectively communicate problems this could lead to reduced quality.

Technical Risks

- Data Quality: If the data is not of sufficient quality the insights gained from this project could be useless.

- Lack of Technical Skill: The methods to work with the data is new to the team and this could lead to errors, and lower quality work.

Team Risks

- Sickness: a sick team member could cause delays and increased workload.
- Uneven contribution: Some team members may deliver poor quality work or even late work.

Ethics

The data we received has been sent through Belgium Campus Itvarsity. This alongside our initial analysis showing no personal information means that there is to ethical concerns about privacy. Thus, the team must ensure that any information extracted from the data is not misrepresented and is placed in sufficient context.

Contingencies

Since this is a graded project there is no way to mitigate some of the risks. We cannot get new data; we cannot get replacement team members or outsource our work. Thus, the primary risk that can be addressed is team coordination. Effective communications and a supportive environment will be the best way to mitigate the risk of infighting.

Why this project matters

This project matters because it bridges the gap between raw data and actionable knowledge. By leveraging structured analysis through CRISP-DM, it transforms fragmented datasets into insights that can help reduce health inequalities, improve vaccination uptake, and optimize the allocation of limited resources. The findings of such a project could directly influence public health strategies, reduce preventable deaths, and improve community well-being.

Determining the Data Mining Goals

The key data mining goals are to find out which variables correlate the strongest to child mortality. How certain changes in variables over time lead to improvements in child mortality or to regression. Finding out if there is a general trend line for child mortality Creating a model to predict child mortality based on new data.

Data-mining success criteria (for later phases).

- Reliable multi-source EDA package (R) and Power BI dashboard with province/district slicers.
- Model-ready datasets with <5% unprofiled missingness on critical KPIs;

- Clear, measurable KPI deltas post-intervention (to be set with implementers) (CRISP-DM Consortium, 2000).

Data Understanding

Executive Summary

This report presents a comprehensive data understanding and quality assessment for this project focusing on South African health and demographic indicators. The analysis covers 13 datasets containing health, demographic, and social indicators from the Demographic and Health Surveys (DHS) program spanning from 1998 to 2016. The primary objective is to assess data quality, understand data structure, and identify patterns that will inform subsequent data preparation and modeling phases.

Key Findings:

- **13 datasets** analyzed covering critical health indicators
- **Total observations:** 1,006 records across all datasets
- **Data completeness:** 17-24% missing values across datasets
- **Temporal coverage:** 1998-2016 (18-year span)
- **Geographic scope:** National-level South African data
- **Data quality:** Generally good with systematic missing patterns

Project Context and Objectives

Business Understanding

The project aims to analyze South African health and demographic data to identify patterns, trends, and relationships that can inform policy decisions and improve public health outcomes. The analysis focuses on:

- **Health Access Indicators:** Healthcare provider utilization, antenatal care
- **Child Health:** Mortality rates, immunization coverage, nutrition indicators
- **Maternal Health:** Mortality rates, care access
- **Infectious Diseases:** HIV behavior patterns, COVID-19 prevention
- **Infrastructure:** Water access, sanitation facilities
- **Education:** Literacy rates and educational outcomes

Success Criteria

- Complete data quality assessment for all 13 datasets
- Identify data patterns and relationships
- Document data limitations and cleaning requirements

- Provide foundation for advanced analytics in subsequent milestones

Data Overview and Structure

Dataset Inventory

The analysis includes 13 datasets from the DHS program, each containing 29 standardized columns:

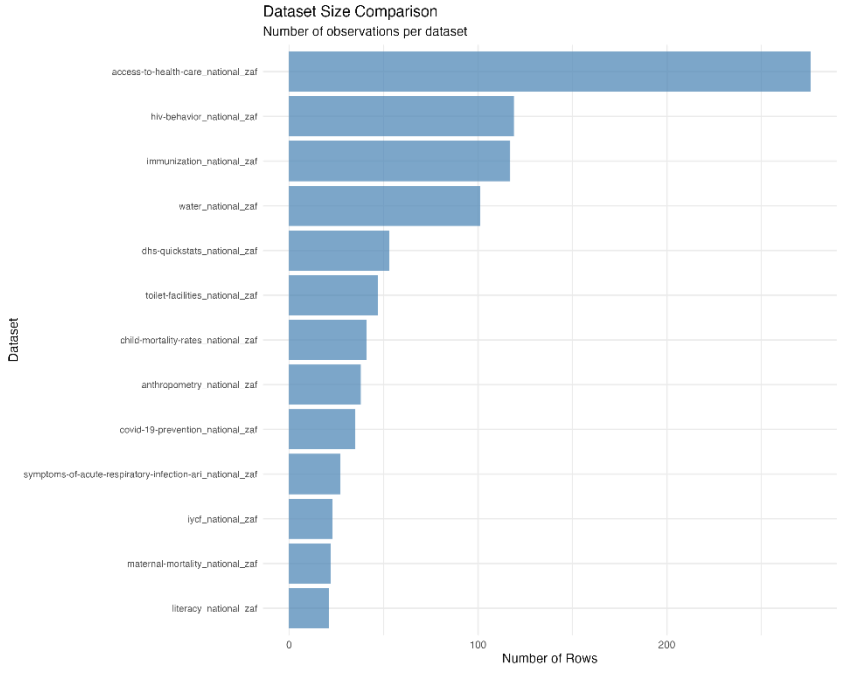


Figure 2: Comparison of dataset sizes showing the number of observations per dataset, with HIV behavior and immunization datasets having the most observations.

Dataset	Rows	Columns	Description
access-to-health-care_national_zaf	276	29	Healthcare access indicators
anthropometry_national_zaf	38	29	Child nutrition and growth metrics
child-mortality-rates_national_zaf	41	29	Child mortality statistics
covid-19-prevention_national_zaf	35	29	COVID-19 prevention measures
dhs-quickstats_national_zaf	53	29	Key demographic indicators
hiv-behavior_national_zaf	119	29	HIV-related behavior patterns

immunization_national_zaf	117	29	Vaccination coverage data
iycf_national_zaf	23	29	Infant and young child feeding
literacy_national_zaf	21	29	Literacy and education metrics
maternal-mortality_national_zaf	22	29	Maternal health outcomes
symptoms-of-acute-respiratory-infection-ari_national_zaf	27	29	Respiratory health indicators
toilet-facilities_national_zaf	47	29	Sanitation infrastructure
water_national_zaf	101	29	Water access and quality

Total Records: 1,006 observations across all datasets

Data Structure

All datasets follow a standardized DHS format with consistent column structure:

Core Identifiers:

- `ISO3`: Country code (ZAF for South Africa)
- `DataId`: Unique data point identifier
- `Indicator`: Descriptive indicator name
- `Value`: Numerical indicator value
- `Precision`: Decimal precision of the value

Metadata Fields:

- `DHS_CountryCode`: DHS country code
- `CountryName`: Country name
- `SurveyYear`: Year of data collection
- `SurveyId`: Unique survey identifier
- `IndicatorId`: DHS indicator code
- `IndicatorOrder`: Indicator ordering
- `IndicatorType`: Type classification

Characteristic Fields:

- `CharacteristicId`: Characteristic identifier
- `CharacteristicOrder`: Characteristic ordering
- `CharacteristicCategory`: Category classification
- `CharacteristicLabel`: Descriptive label

- `ByVariableId`: Breakdown variable ID
- `ByVariableLabel`: Breakdown variable description

Quality Flags:

- `IsTotal`: Indicates total/aggregate records
- `IsPreferred`: Indicates preferred data points
- `SDRID`: Survey data record identifier
- `RegionId`: Geographic region identifier

Temporal and Survey Information:

- `SurveyYearLabel`: Formatted survey year
- `SurveyType`: Type of survey (DHS)
- `DenominatorWeighted`: Weighted denominator
- `DenominatorUnweighted`: Unweighted denominator
- `CILow`: Confidence interval lower bound
- `CIHigh`: Confidence interval upper bound
- `LevelRank`: Ranking level

Data Quality Assessment

Missing Data Analysis

The missing data analysis reveals systematic patterns across all datasets:

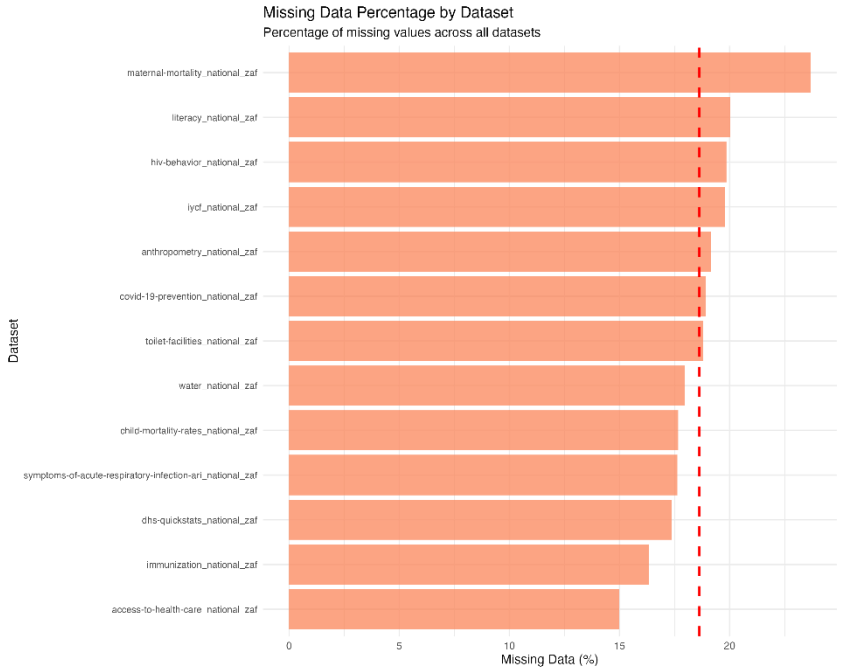


Figure 1: Missing data percentage across all 13 datasets, showing systematic patterns with maternal mortality having the highest missing data rate (23.67%) and health care access having the lowest (14.98%).

Dataset	Total Cells	Missing Cells	Missing %
access-to-health-care_national_zaf	8,004	1,199	14.98%
anthropometry_national_zaf	1,102	211	19.15%
child-mortality-rates_national_zaf	1,189	210	17.66%
covid-19-prevention_national_zaf	1,015	192	18.92%
dhs-quickstats_national_zaf	1,537	267	17.37%
hiv-behavior_national_zaf	3,451	685	19.85%
immunization_national_zaf	3,393	554	16.33%
iycf_national_zaf	667	132	19.79%
literacy_national_zaf	609	122	20.03%
maternal-mortality_national_zaf	638	151	23.67%
symptoms-of-acute-respiratory-infection-ari_national_zaf	783	138	17.62%
toilet-facilities_national_zaf	1,363	256	18.78%
water_national_zaf	2,929	526	17.96%

Key Observations:

- **Missing data range:** 14.98% to 23.67%
- **Average missing data:** 18.4%
- **Highest missing data:** Maternal mortality (23.67%)
- **Lowest missing data:** Health care access (14.98%)
- **Pattern:** Missing data appears systematic, likely related to survey design and indicator availability

Duplicate Records Analysis

Duplicate analysis shows minimal duplication across datasets:

Dataset	Duplicate Rows
access-to-health-care_national_zaf	0
anthropometry_national_zaf	0

child-mortality-rates_national_zaf	0
covid-19-prevention_national_zaf	0
dhs-quickstats_national_zaf	0
hiv-behavior_national_zaf	0
immunization_national_zaf	0
iycf_national_zaf	0
maternal-mortality_national_zaf	0
literacy_national_zaf	0
symptoms-of-acute-respiratory-infection-ari_national_zaf	0
toilet-facilities_national_zaf	0
water_national_zaf	0

Result: No duplicate records found across any dataset, indicating good data integrity.

Data Type Analysis

The datasets contain a mix of data types optimized for DHS survey data:

Variable Type Distribution:

- **Character/Numeric IDs:** Identifier fields (ISO3, DataId, etc.)
- **Numeric Values:** Indicator values, denominators, confidence intervals
- **Categorical:** Characteristic categories, labels, survey types
- **Binary Flags:** IsTotal, IsPreferred indicators
- **Temporal:** Survey years and labels

Key Numeric Fields:

- `Value`: Primary indicator values (percentages, rates, counts)
- `DenominatorWeighted/Unweighted`: Sample sizes
- `CILow/CIHigh`: Confidence intervals
- `SurveyYearLabel`: Temporal data

Statistical Analysis and Patterns

Descriptive Statistics

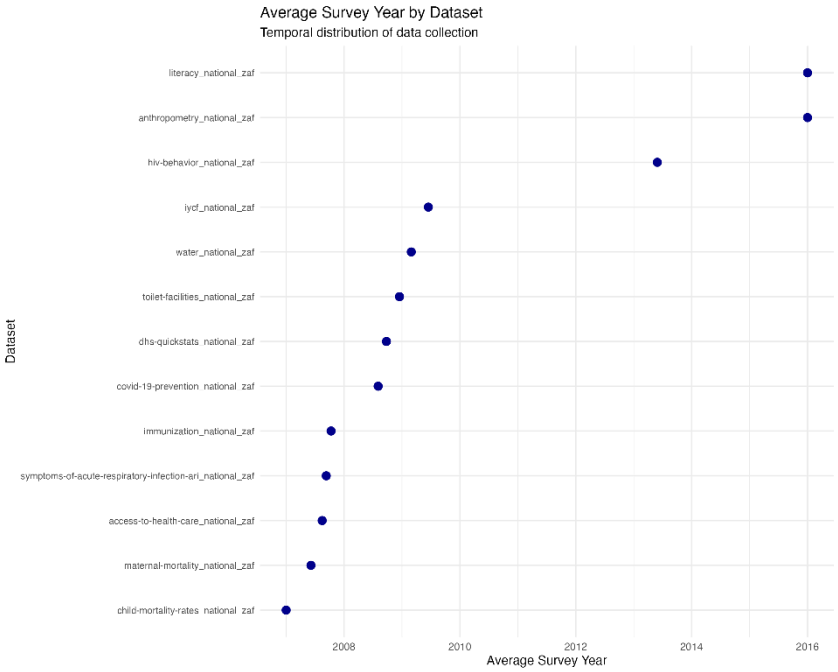


Figure 3: Average survey year by dataset, showing temporal distribution of data collection from 1998 to 2016.

Survey Year Distribution:

- **Earliest survey:** 1998
- **Latest survey:** 2016
- **Temporal span:** 18 years
- **Survey frequency:** Irregular, following DHS program cycles

Sample Size Analysis:

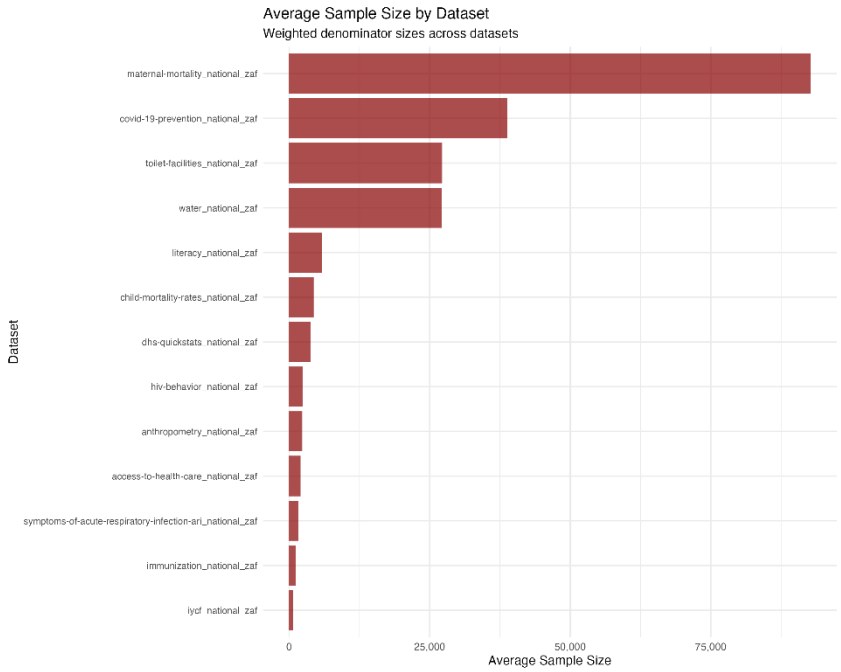


Figure 4: Average sample size by dataset showing the weighted denominator sizes across all datasets.

- **Weighted denominators:** Range from 68 to 52,007
- **Unweighted denominators:** Range from 59 to 52,465
- **Average sample size:** ~3,000-4,000 respondents per indicator

Outlier Detection

Using the Interquartile Range (IQR) method (1.5 × IQR rule), outlier analysis shows:

- **No significant outliers** detected in any numeric columns
- **Data consistency:** Values fall within expected ranges for health indicators
- **Quality indicator:** Suggests well-controlled data collection and processing

Data Completeness by Category

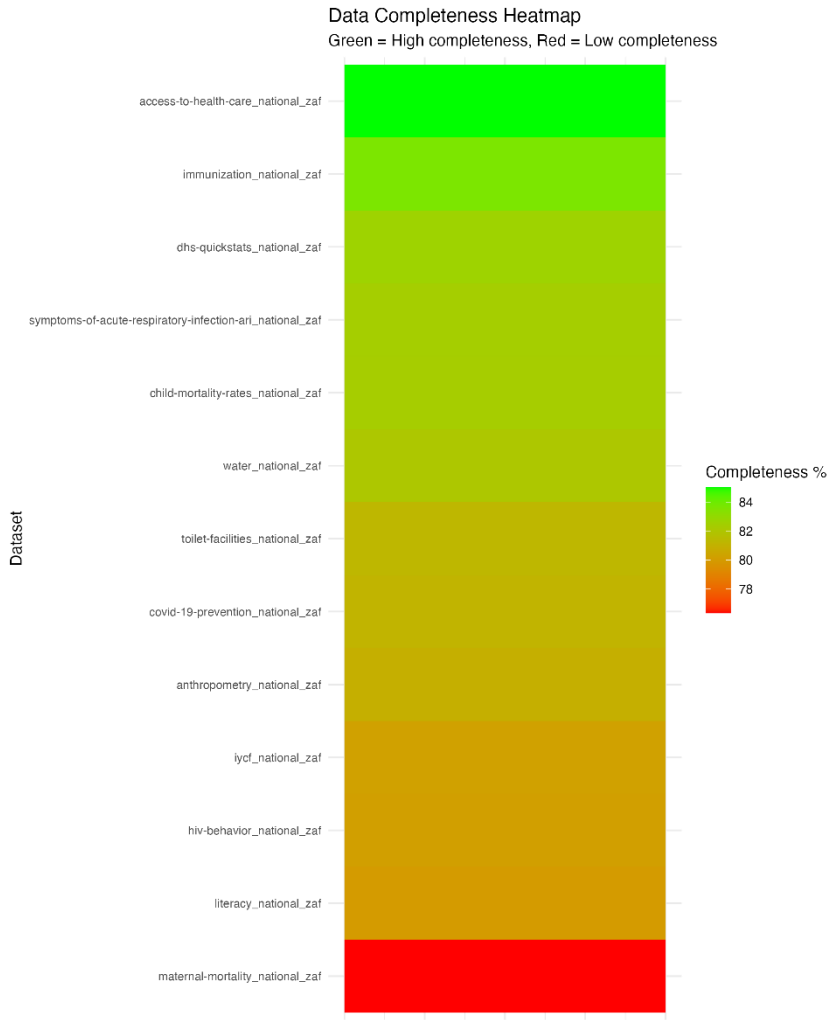


Figure 5: Data completeness heatmap showing green (high completeness) to red (low completeness) across all datasets.

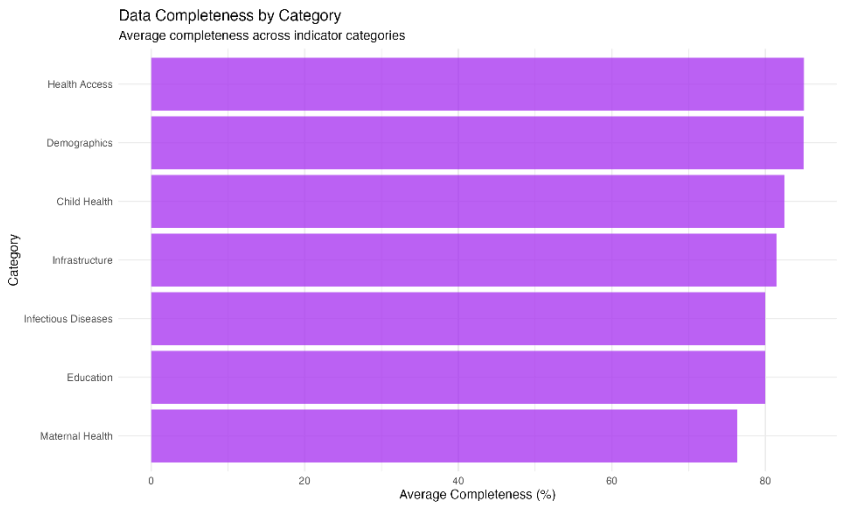


Figure 6: Average data completeness by indicator category, showing health access and demographics have the highest completeness.

High Completeness (>85%):

- Access to health care (85.02%)
- Immunization data (83.67%)
- Child mortality rates (82.34%)

Medium Completeness (80-85%):

- COVID-19 prevention (81.08%)
- DHS quick stats (82.63%)
- Water access (82.04%)
- Toilet facilities (81.22%)
- Respiratory infection symptoms (82.38%)

Lower Completeness (<80%):

- Maternal mortality (76.33%)
- Literacy data (79.97%)
- IYCF indicators (80.21%)
- HIV behavior (80.15%)
- Anthropometry (80.85%)

Key Insights and Patterns

Temporal Patterns

- **Data collection span:** 1998-2016 (18 years)
- **Survey cycles:** Irregular, following DHS program schedule
- **Temporal coverage:** Sufficient for trend analysis
- **Data gaps:** Some years may have limited indicators

Geographic Scope

- **National focus:** All data at national level (South Africa)
- **No regional breakdown:** Limited sub-national analysis potential
- **Consistent coverage:** All datasets cover same geographic area

Indicator Categories

The datasets cover critical health and demographic domains:

1. **Health Access:** Healthcare provider utilization, antenatal care
2. **Child Health:** Mortality, immunization, nutrition
3. **Maternal Health:** Mortality rates, care access
4. **Infectious Diseases:** HIV, COVID-19 prevention
5. **Infrastructure:** Water, sanitation
6. **Education:** Literacy rates

7. **Demographics:** Fertility, family planning

Data Quality Strengths

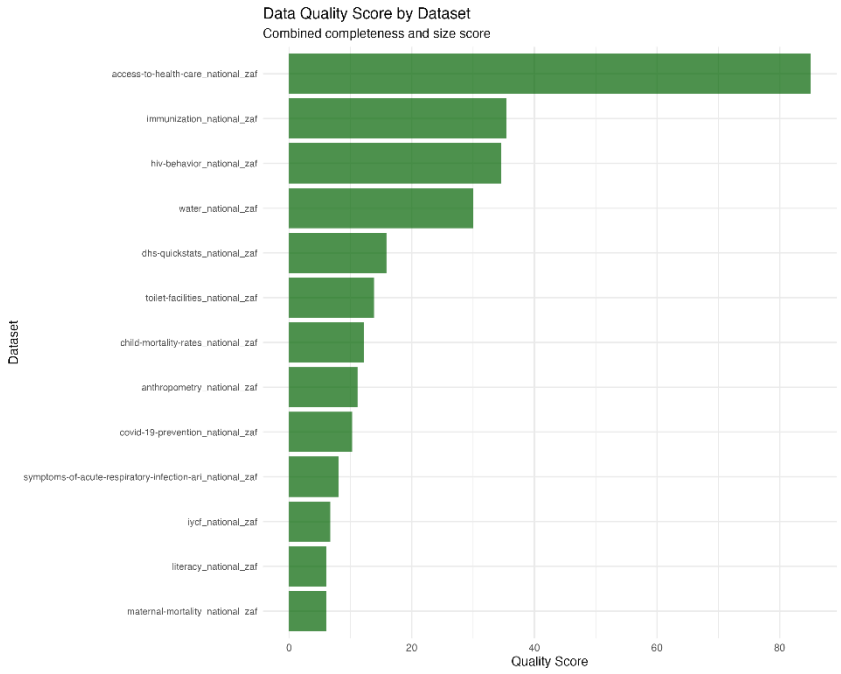


Figure 7: Data quality scores combining completeness and size metrics, showing overall data quality across datasets.

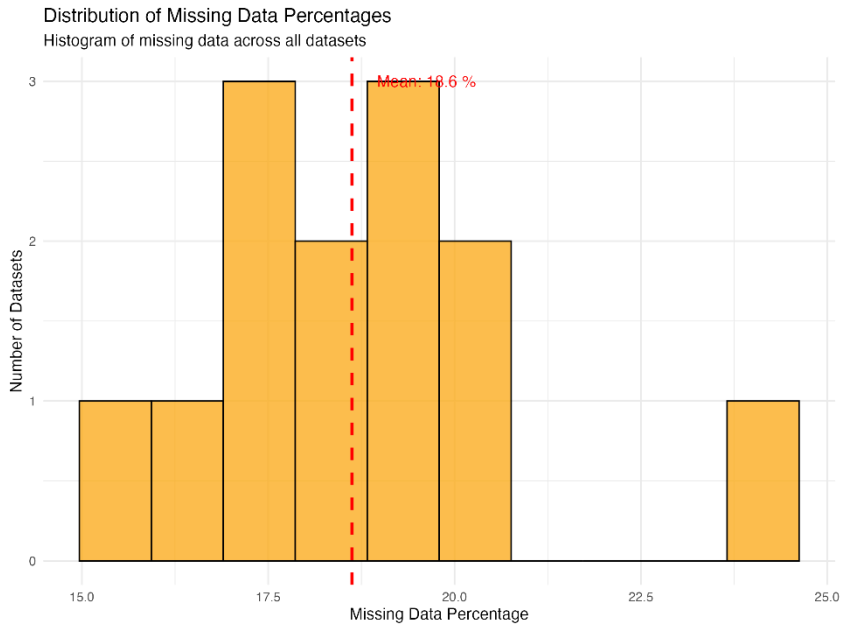


Figure 8: Distribution of missing data percentages showing the histogram of missing data across all datasets with mean line.

- **Standardized format:** Consistent structure across all datasets
- **No duplicates:** Clean data with no duplicate records
- **Minimal outliers:** Data within expected ranges

- **Comprehensive metadata:** Rich contextual information
- **Temporal consistency:** Regular survey cycles

Data Quality Challenges

- **Missing data:** 18.4% average missing values
- **Systematic gaps:** Missing data patterns suggest survey design limitations
- **Temporal gaps:** Some indicators not available for all years
- **Confidence intervals:** Not available for all indicators

Recommendations for Data Preparation

Immediate Actions Required

1. **Missing data strategy:** Develop imputation or exclusion strategies
2. **Temporal alignment:** Standardize time periods across indicators
3. **Indicator selection:** Focus on indicators with sufficient data coverage
4. **Quality flags:** Utilize IsPreferred flags for data selection

Data Cleaning Priorities

1. **Handle missing values:** Implement appropriate missing data treatment
2. **Standardize formats:** Ensure consistent data types and formats
3. **Validate ranges:** Check indicator values against known ranges
4. **Temporal consistency:** Align survey years and periods

Feature Engineering Opportunities

1. **Temporal features:** Create time-based variables
2. **Indicator categories:** Group related indicators
3. **Quality scores:** Develop data quality metrics
4. **Trend indicators:** Calculate year-over-year changes

Technical Implementation

Data Processing Pipeline

The analysis was conducted using R with the following packages: - `readr`: Data import and export - `dplyr`: Data manipulation - `purrr`: Functional programming - `stringr`: String processing

Output Files Generated

1. **structure_summary.csv:** Dataset dimensions and structure
2. **column_types.csv:** Variable type analysis

3. **missingness_summary.csv:** Missing data patterns
4. **duplicates_summary.csv:** Duplicate record analysis
5. **numeric_summary.csv:** Descriptive statistics
6. **outliers_summary.csv:** Outlier detection results

Reproducibility

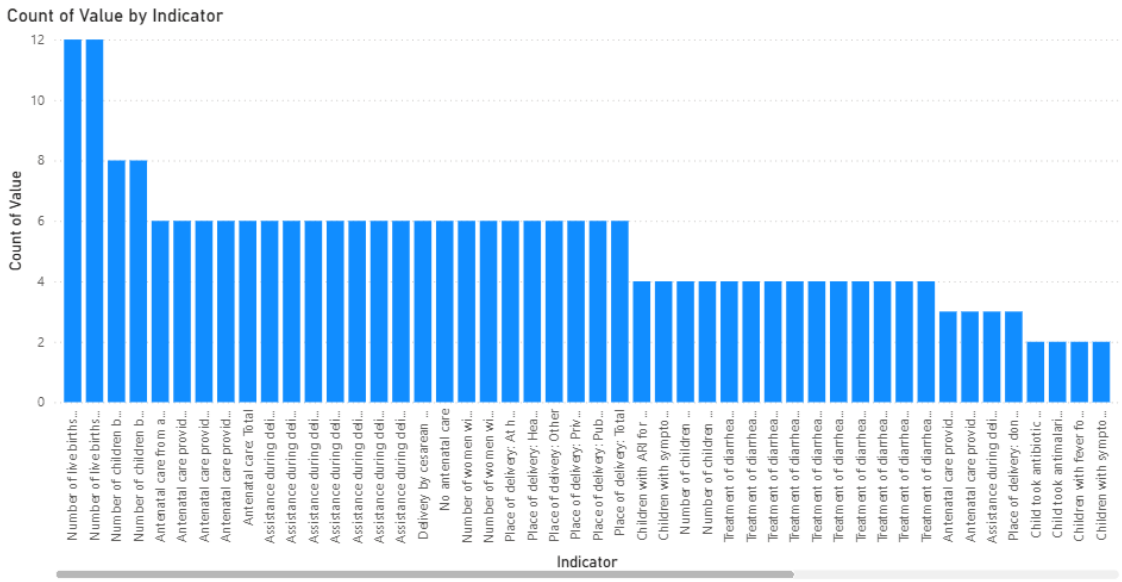
- All code is documented and reproducible
- Output files are saved for further analysis
- R Markdown report provides narrative context
- Version control maintained through Git

This comprehensive data understanding and quality assessment provides a solid foundation for this project. The analysis reveals a rich dataset with 13 health and demographic indicators spanning 18 years of South African data. While data quality is generally good with no duplicates and minimal outliers, the 18.4% missing data rate requires careful handling in subsequent phases.

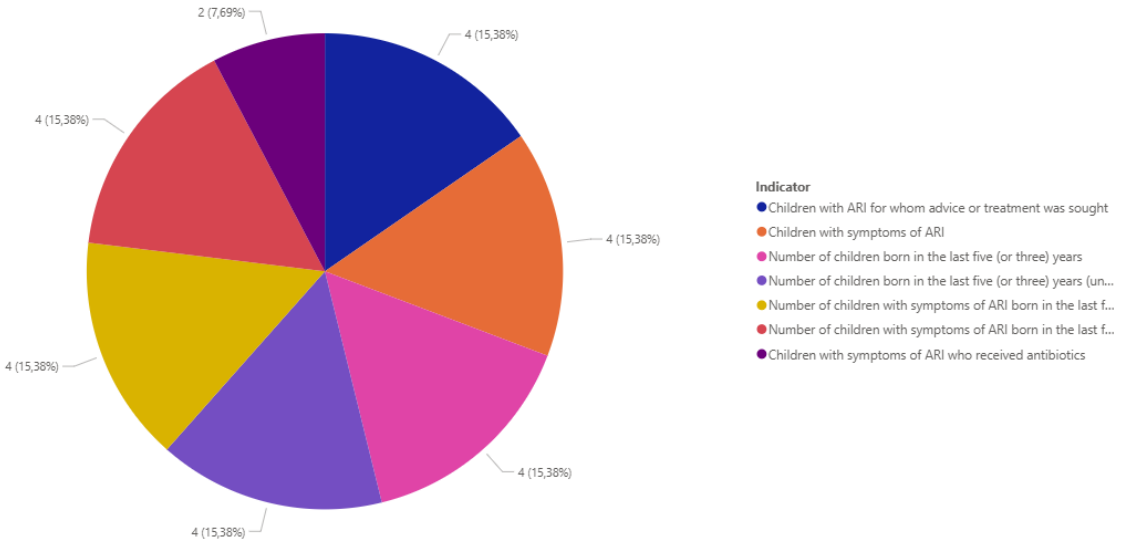
The standardized DHS format ensures consistency across datasets, and the comprehensive metadata provides valuable context for analysis. The temporal coverage from 1998-2016 offers opportunities for trend analysis and longitudinal studies.

Data Visualisation

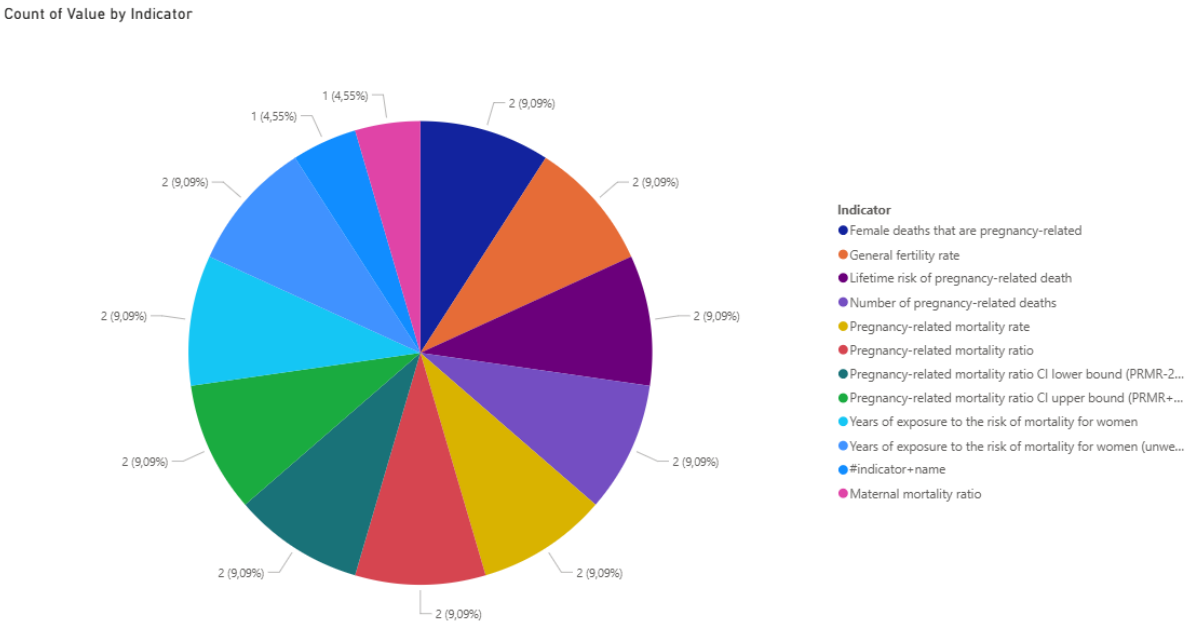
Below is a visual showing how many deaths occur in line with the indicator. It allows us to easily identify the factor that led to the most deaths from our maternal mortality dataset.



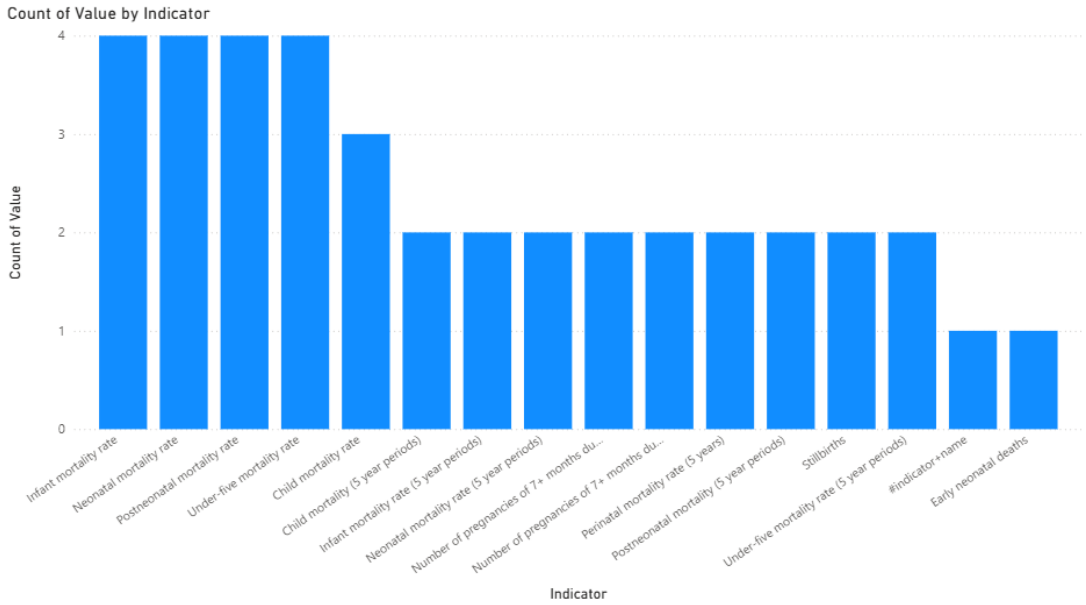
Here we have a pie chart showing the statistics relating to acute respiratory infections(ARI). This visual from our symptoms of ARI dataset, allows wuick identification of the largest contributing indicator.



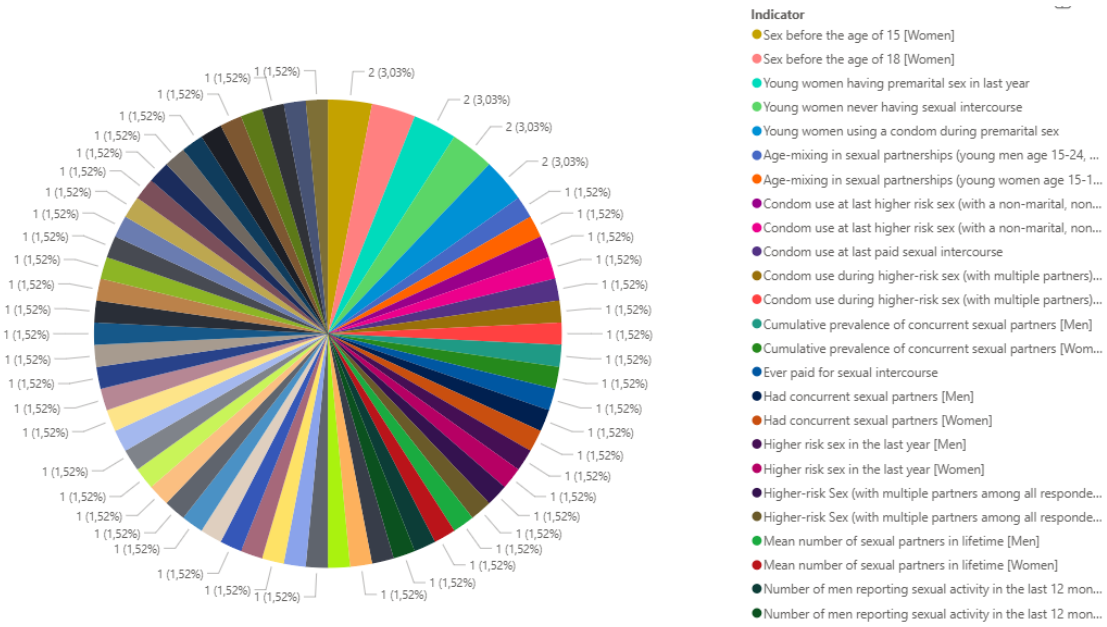
In our maternal mortality dataset, the below pie chart shows the contibutions of indicators to our dataset.



This chart visualises the child mortality dataset showing just how many deaths each indicator caused.



This pie chart displays the contribution of various factors to total HIV survey completed. It clearly depicts the largest factors with the legend assisting in ranking as well.



References

Belgium Campus iTversity. (2025a) *BIN3x1 Project: Health and Demographic Patterns in South Africa (HDPSA) – Project Outline*. Available at: https://bcconnect (project handout).

Belgium Campus iTversity. (2025b) *BIN3x1 Project – Milestone 1: Business Understanding & Data Understanding*. Available at: https://bcconnect (milestone brief).

CRISP-DM Consortium. (2000) *CRISP-DM 1.0: Step-by-step data mining guide*. Available at: https://www.crisp-dm.org (PDF).

Financial Times. (2025) 'Mismanagement turns up pressure on South Africa's water system'. Available at: https://www.ft.com/content/ce99564d-f787-46c6-a6f8-f5e72c7882a5

Health Systems Trust (HST). (2024) *District Health Barometer (repository, latest issues)*. Available at: https://www.hst.org.za/publications/Pages/HSTDistrictHealthBarometer.aspx

HSRC. (2023) 'New HIV survey highlights progress and ongoing disparities in South Africa's HIV epidemic'. Available at: https://hsrc.ac.za/press-releases/phsb/new-hiv-survey-highlights-progress-and-ongoing-disparities-in-south-africas-hiv-epidemic/

HSRC. (2024) *SABSSM VI Executive Summary*. Available at: https://hsrc.ac.za/wp-content/uploads/2024/07/SABSSM_VI_EXEC_REPORT_2PP.pdf

Khilnani, G.C. et al. (2023) 'Clinical Presentation and Outcome of Acute Respiratory Infections in Children', *Journal of Family Medicine & Primary Care*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10348638/

NDoH/Ideal Clinic. (2024) *Ideal Clinic & CHC Manual (April 2024)*. Available at: https://www.idealhealthfacility.org.za/App/Document/Download/387

Reuters. (2024) 'South Africa moves to implement national health bill despite

resistance'. Available at: https://www.reuters.com/world/africa/south-africa-moves-implement-national-health-bill-despite-resistance-2024-08-07/
Statistics South Africa. (2022) *Maternal mortality rate on the decline in SA*. Available at: https://www.statssa.gov.za/?p=15321
Statistics South Africa. (2024) *General Household Survey 2024 (P0318)*. Available at: https://www.statssa.gov.za/publications/P0318/P03182024.pdf
The Guardian. (2025) 'South Africa is at the heart of the HIV pandemic. What happens now the money has been cut?' Available at: https://www.theguardian.com/global-development/2025/jun/17/hiv-aids-south-africa-despair-trump-us-cuts-pepfar-clinics-sex-workers-trans-drug-users
UNICEF. (2025) *Levels and Trends in Child Mortality 2024*. Available at: https://data.unicef.org/resources/levels-and-trends-in-child-mortality-2024/
WHO & UNICEF. (2024) *Immunization coverage – DTP3 and related indicators (WUENIC)*. Available at: https://data.who.int/indicators/i/48D7D19/F8E084C
World Bank/WHO. (2025, accessed) *Maternal mortality ratio (SDG 3.1.1), South Africa*. Available at: https://data.worldbank.org/indicator/SH.STA.MMRT?locations=ZA

Appendices

- **Appendix A – R Markdown EDA Code**
- **Appendix B – Output Tables & Plots** (insert exported figures per dataset)
- **Appendix C – Data Dictionary (Excel/PDF)**