DATA MINING AND VISUALIZATION

MILESTONE 2

ANGELITE ARENDSE

KAGISO SEBATI

SIMOEN LE ROUX

IAN FAASEN

Table of Contents

SECTION 1: DATA SELECTION AND DESCRIPTION

1. DATE DESCRIPTION

We started with 13 health and demographic datasets for this project. After carefully applying inclusion and exclusion criteria looking at relevance, data quality, and whether the data could be used effectively with our tools we narrowed it down to nine datasets that align closely with our goals. Together, they offer a rich mix of health outcomes, service and nutrition indicators, and demographic and environmental variables. In short, they give us a strong foundation to explore the questions we're asking.

Here's a quick look at the datasets we selected:

1.1 Access to Health Care

This dataset comes from the Demographic and Health Surveys (DHS) and includes details like provider type (doctor, nurse/midwife, other), survey year, service coverage, denominators, and confidence intervals. It's especially useful for understanding access to healthcare services — with a spotlight on maternal and reproductive health. It helps us see how healthcare is delivered across South Africa and where gaps might exist.

1.2 Child Mortality

Also sourced from DHS, this dataset focuses on child mortality rates and related healthcare access. It includes similar attributes to the Access to Health Care dataset and gives us a clearer picture of how child health outcomes are influenced by service availability and provider type.

1.3 Maternal Mortality

Pulled from both DHS and the World Health Organization (WHO), this dataset tracks maternal death ratios across different survey years. It's essential for calculating mortality rates during pregnancy and childbirth, and for identifying areas where maternal health services may be lacking.

1.4 HIV Quickstats

This one’s from UNAIDS and DHS. It tracks HIV prevalence and antiretroviral treatment (ART) coverage, broken down by age and gender across different survey years. This dataset is a vital tool for understanding the trajectory of South Africa’s HIV epidemic not just where it stands, but how treatment programs are performing and who they’re reaching. It helps us spot trends, identify gaps, and measure progress over time.

1.5 Immunization

This dataset tracks how many children under five are getting vaccinated for BCG, DPT, Polio, Measles, the big ones. It’s sourced from DHS and WHO, and it’s one of our clearest indicators of how well preventive care is reaching families. If coverage is high, it’s a good sign. If not, we know where to look closer.

1.6 IYCF

This one dives into how babies and toddlers are being fed: breastfeeding habits, dietary diversity, and other nutrition markers. Pulled from DHS and UNICEF, it gives us a window into early childhood health. These are the building blocks of survival and development, and this dataset helps us see where support is strong and where it’s slipping.

1.7 Literacy

This dataset combines data from national education surveys and DHS, offering literacy rates broken down by age, gender, region, and year. Literacy isn’t just an education metric — it’s a powerful predictor of health. People with higher literacy levels tend to have better health knowledge, make informed decisions, and access services more effectively.

1.8 Water

Sourced from DHS and Statistics South Africa, this dataset looks at household access to safe drinking water, including the type of water source and survey year. Clean water is a basic human need, and this dataset helps us assess how well communities are being served — and where environmental health risks may be lurking.

1.9 Toilet Facilities

Also, from DHS and Statistics South Africa, this dataset details household sanitation types from flush toilets to pit latrines, or no facilities at all. Sanitation plays a huge role in preventing infectious diseases and maintaining overall health. This dataset gives us a window into environmental health conditions and infrastructure gaps.

2. INCLUSION AND EXCLUSION CRITERIA

The datasets were evaluated using four key criteria: relevance, data quality, technical feasibility, and coverage. Relevance assessed whether the dataset directly aligns with project objectives related to health outcomes, service delivery, nutrition, and demographic or environmental determinants. Data quality evaluated completeness, consistency, and minimal missing values in critical fields. Technical feasibility considered the ability to analyse the data efficiently in R and Power BI, excluding datasets that were overly large or complex without sufficient analytical value. Coverage and redundancy ensured that each dataset contributed unique insights; datasets that overlapped substantially with others or did not provide additional information were excluded.

Based on these criteria, 9 datasets were included in the analysis:

- **Access to Health Care** was selected for its direct relevance to service access and maternal/child health services.
- **Child Mortality** and **Maternal Mortality** were included as core health outcome indicators.
- **HIV Quickstats** was selected due to its importance in monitoring a major national health burden.
- **Immunization** and **IYCF** datasets were included for their focus on preventive health and child nutrition.

- **Literacy** was included as education strongly influences health knowledge and behaviour.
- **Water** and **Toilet Facilities** were selected as key environmental determinants.

Conversely, datasets such as Anthropometry Rates, COVID-19 Prevention, and ARI Symptoms were excluded due to redundancy, narrow focus, and limited relevance. The DHS Behaviour dataset was considered optional and may be used only if specific behavioural analysis is required.

1 PRELIMINARY DATA EXPLORATION

1 Access to Health Care

This dataset has 276 rows and 29 columns, tracking health service access such as antenatal care providers and type of healthcare professionals. Most values are numeric percentages or survey-weighted estimates, with survey years ranging back to 1998. Missing values are low and mostly in confidence interval columns. It's a wide dataset with a mix of indicators and demographic breakdowns.

2 Child Mortality Rates

Containing 72 rows, this dataset summarizes infant, child, and under-5 mortality rates. Indicators are survey-based with denominators and confidence intervals included. Coverage spans multiple survey years, allowing trend analysis of child survival rates over time.

3 Maternal Mortality

With 17 rows, this dataset is one of the smallest, but highly significant. It tracks maternal mortality ratio estimates from surveys and modelled data. It's concise and ready for quick analysis, with percentages and ratios clearly provided.

Data Quality Verification

Significance and Correlation

In this phase of the project we need to ensure that the correct fields are included. We do this by testing the correlation and the significance of each field.

Access to Health Care

This table and the following tables in regards of correlation and significance tests, present the result of a correlation analysis between different fields in the dataset “Access to Health Care”. Each row shows the relationship between two specific variables, providing three key metrics: Correlation coefficient, p-value and the 95% confidence interval (CI).

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
DataId	Value	-15.7	0.009126	(-0.2702, -0.0394)
DataId	Precision	15.58	0.009664	(0.0382, 0.2691)
DataId	SurveyYear	16.25	0.006918	(0.0451, 0.2755)
DataId	IndicatorOrder	25.15	0.000025	(0.1373, 0.3591)
DataId	CharacteristicId	NA	NA	(NA, NA)
DataId	CharacteristicOrder	NA	NA	(NA, NA)
DataId	ByVariableId	20.7	0.000551	(0.091, 0.3175)

DataId	IsTotal	NA	NA	(NA, NA)
DataId	IsPreferred	-14.09	0.019401	(-0.2549, -0.023)
DataId	SurveyYearLabel	16.25	0.006918	(0.0451, 0.2755)
DataId	DenominatorWeighted	-8.17	0.206394	(-0.2059, 0.0452)
DataId	DenominatorUnweighted	-7.88	0.222889	(-0.2031, 0.048)
Value	Precision	-77.74	0	(-0.8203, -0.7259)
Value	SurveyYear	-9.12	0.131359	(-0.2073, 0.0274)
Value	IndicatorOrder	-9.08	0.132966	(-0.2069, 0.0278)
Value	CharacteristicId	NA	NA	(NA, NA)
Value	CharacteristicOrder	NA	NA	(NA, NA)
Value	ByVariableId	1.72	0.776624	(-0.1013,

				0.1352)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	-7.47	0.216974	(-0.1913, 0.044)
Value	SurveyYearLabel	-9.12	0.131359	(-0.2073, 0.0274)
Value	DenominatorWeighted	22.81	0.000358	(0.1047, 0.3445)
Value	DenominatorUnweighted	22.6	0.000405	(0.1026, 0.3426)
Precision	SurveyYear	-2.79	0.645088	(-0.1457, 0.0907)
Precision	IndicatorOrder	-8.27	0.171419	(-0.199, 0.0359)
Precision	CharacteristicId	NA	NA	(NA, NA)
Precision	CharacteristicOrder	NA	NA	(NA, NA)
Precision	ByVariableId	3.12	0.606409	(-0.0874, 0.1489)

Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	-2.25	0.710697	(-0.1404, 0.0961)
Precision	SurveyYearLabel	-2.79	0.645088	(-0.1457, 0.0907)
Precision	DenominatorWeighted	-0.53	0.934872	(-0.1316, 0.1212)
Precision	DenominatorUnweighted	-0.39	0.952225	(-0.1302, 0.1225)
SurveyYear	IndicatorOrder	9.04	0.134832	(-0.0282, 0.2065)
SurveyYear	CharacteristicId	NA	NA	(NA, NA)
SurveyYear	CharacteristicOrder	NA	NA	(NA, NA)
SurveyYear	ByVariableId	-20.79	0.000522	(-0.3183, -0.0918)
SurveyYear	IsTotal	NA	NA	(NA, NA)
SurveyYear	IsPreferred	7.37	0.22311	(-0.045,

				0.1903)
SurveyYear	SurveyYearLabel	100	0	(1.0, 1.0)
SurveyYear	DenominatorWeighted	-37.41	0.000000002006077	(-0.4779, -0.2601)
SurveyYear	DenominatorUnweighted	-38.34	0.0000000073600018	(-0.4862, -0.2701)
IndicatorOrder	CharacteristicId	NA	NA	(NA, NA)
IndicatorOrder	CharacteristicOrder	NA	NA	(NA, NA)
IndicatorOrder	ByVariableId	15.29	0.011105	(0.0353, 0.2664)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	11.62	0.054242	(-0.0021, 0.2313)
IndicatorOrder	SurveyYearLabel	9.04	0.134832	(-0.0282, 0.2065)
IndicatorOrder	DenominatorWeighted	-66.25	0	(-0.7279, -

				0.5851)
IndicatorOrder	DenominatorUnweighted	-66.12	0	(-0.7269, -0.5836)
CharacteristicId	CharacteristicOrder	NA	NA	(NA, NA)
CharacteristicId	ByVariableId	NA	NA	(NA, NA)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	NA	NA	(NA, NA)
CharacteristicId	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicId	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicId	DenominatorUnweighted	NA	NA	(NA)
CharacteristicOrder	ByVariableId	NA	NA	(NA, NA)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	NA	NA	(NA, NA)
CharacteristicOrder	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicOrder	DenominatorWeighted	NA	NA	(NA, NA)

CharacteristicOrder	DenominatorUnweighted	NA	NA	(NA, NA)
ByVariableId	IsTotal	NA	NA	(NA, NA)
ByVariableId	IsPreferred	-26.06	0.000012	(-0.3676, -0.1468)
ByVariableId	SurveyYearLabel	-20.79	0.000522	(-0.3183, -0.0918)
ByVariableId	DenominatorWeighted	10.16	0.115522	(-0.025, 0.2251)
ByVariableId	DenominatorUnweighted	10.41	0.107047	(-0.0226, 0.2274)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	7.37	0.22311	(-0.045, 0.1903)

IsPreferred	DenominatorWeighted	-41.1	0	(-0.5109, -0.3003)
IsPreferred	DenominatorUnweighted	-41.25	0	(-0.5122, -0.3019)
SurveyYearLabel	DenominatorWeighted	-37.41	0.000000002006077	(-0.4779, -0.2601)
SurveyYearLabel	DenominatorUnweighted	-38.34	0.0000000007360018	(-0.4862, -0.2701)
DenominatorWeighted	DenominatorUnweighted	99.99	0	(0.9998, 0.9999)

Include or Exclude

Inclusion Criteria:

Fields were considered for inclusion if they demonstrated a statistically significant linear relationship with the target variable, 'Value'. This significance was formally assessed using the p-value from Pearson's correlation tests, with a threshold set at $p<0.05$. This criterion ensures that the selected predictors have a non-random association with the outcome. Furthermore, a thorough data quality assessment was performed on each field. This involved evaluating the extent of missing data and the presence of extreme outliers. Variables with an excessive number of missing values or those containing outliers that could disproportionately influence model training were flagged for potential exclusion or require specific imputation strategies. Finally, the domain relevance of each field was evaluated against the primary modeling objective:

predicting health care access. Variables were retained only if they provided meaningful insight into the target variable from a public health or socio-economic perspective.

Exlusion Criteria:

Several criteria were applied to systematically exclude variables that could compromise the model's integrity or efficiency. Fields containing constant or non-numeric values were automatically excluded from the correlation analysis, as their lack of variance or quantifiable measure prevents them from contributing to a linear predictive model. Examples include categorical identifiers that do not possess an inherent order or scale. Weak or non-significant relationships were another key reason for exclusion. Variables with a correlation coefficient close to zero or a p-value greater than the significance threshold ($p \geq 0.05$) were deemed to have insufficient predictive power and were thus removed. A crucial step was the identification and removal of redundant variables. This was done to address multicollinearity, a condition where two or more predictor variables are highly correlated. The analysis revealed a near-perfect correlation between 'DenominatorUnweighted' and 'DenominatorWeighted' ($r = 0.9999$). To maintain model parsimony and avoid redundant information, 'DenominatorUnweighted' was excluded, as its information content was almost entirely captured by 'DenominatorWeighted'. Similarly, 'SurveyYear' and 'SurveyYearLabel' were found to be perfectly correlated ($r=1$) and thus 'SurveyYearLabel' was removed.

Fields to Include:

Fields with a significant correlation, typically with a p-value less than 0.05, with other variables should be included, as they are likely to have meaningful relationship. The following field relationships is included:

- Value and Precision: They have a strong negative correlation (-77.74%), suggesting that as one increase, the other significantly decreases.
- Indicator Order and Denomitor Weighted/Unweighted: They have a strong negative correlation (-66.25% and -66.12%) indicating that they are closely related.
- Survey Year and Denomitor Weighted/Unweighted: Moderate negative correlation (-37.41% and -38.34%)
- Is preferred and Denomitor Weighted/Unweighted: Moderate negative correlation (-41.1% and -41.25%)

- Denomitor Weighted and Denomitor Unweighted: These fields have a near-perfect positive correlation (99.99%) indicating these fields are essentially the same and one could be dropped.
- SurveyYear and SurveyYearLabel: A perfect positive correlation (100%) means these fields are redundant and one should be excluded.
- DataId and several fields: DataId shows significant correlations with Value (-15.7%), Precision (15.58%), SurveyYear (16.25%), IndicatorOrder (25.15%), ByVariableId (20.7%) and IsPreferred (-14.09%). This suggests that DataId is a meaningful identifier that captures information about these other variables.

Fields to Exclude

Fields with a non-statistically significant correlation with most other varialbes should be considerd for exlusion, as they may not add predictive power or meaningful information to the dataset.

- Value: This field does not have a significant correlation with SurveyYear (p = 0.13), IndicatorOrder (p = 0.13), ByVariableId (p = 0.78), or IsPreferred (p = 0.22), despite having strong correlations with Precision and Denominator fields. Its inclusion should be based on its fundamental importance as a primary measure.
- Precision: This field shows non-significant correlations with all tested variables except DataId and Value.
- SurveyYear and SurveyYearLabel: These have a perfect correlation. Keep one and exclude the other to avoid redundancy.
- CharacteristicId and CharacteristicOrder: The correlation tests were "NA" for these fields, which suggests they might be categorical or have missing data. Their relationship to other fields cannot be determined from these results. Further analysis is needed to decide on their inclusion.
- IsTotal: All tests involving this field resulted in "NA," suggesting it might also be a categorical or non-numeric field, or have missing values. Its relationship to other fields cannot be determined from these results.
- ByVariableId: This field has significant correlations with DataId, SurveyYear and IndicatorOrder, but not with Value, Precision, or Denominator fields. Its inclusion depends on the importance of its relationship with the other variables.

- IsPreferred: This field has significant correlations with DataId and Denominator fields, but not with Value, Precision, SurveyYear, IndicatorOrder, or ByVariableId. Its inclusion depends on its purpose in the dataset.

To summarise what should be kept:

DataId, Value, Precision, IndicatorOrder, SurveyYear (and drop SurveyYearLabel), DenominatorWeighted (and drop DenominatorUnweighted), IsPreferred and ByVariableId.

Attribute Weighting Strategy

To quantify the relative importance of each selected predictor, an attribute weighting strategy was implemented based on the absolute value of the correlation coefficient with the target variable, 'Value'. This approach is grounded in the principle that a variable's predictive strength is directly proportional to the magnitude of its linear relationship with the outcome.

The weights for each attribute were calculated as follows:

1. Identify the absolute correlation coefficient for each field that passed the inclusion criteria. For the target variable 'Value', the most significant correlations were found with 'Precision' ($|r| = 0.7774$) and the denominator fields, 'DenominatorWeighted' ($|r| = 0.2281$) and 'DenominatorUnweighted' ($|r| = 0.2260$).

2. Sum these absolute values to create a normalization factor. The sum of the absolute correlation coefficients for these three fields is $0.7774+0.2281+0.2260=1.2315$.

3. Normalize each absolute value by dividing it by the sum. This transforms the correlation coefficients into a set of weights that sum to 1, providing a clear representation of each variable's contribution relative to the others.

The resulting weights were:

- Precision**:** $0.7774/1.2315\approx0.6313$
- DenominatorWeighted**:** $0.2281/1.2315\approx0.1852$
- DenominatorUnweighted**:** $0.2260/1.2315\approx0.1835$

This weighting scheme explicitly reflects the greater predictive influence of 'Precision' on 'Value' compared to the other attributes. By assigning higher weights to more strongly correlated variables, the model can prioritize the most impactful features during the learning process, thereby optimizing its predictive performance. This method provides a transparent and justifiable basis for feature prioritization, a key component of robust academic research.

Child Mortality Rates

Variable 1	Variable 2	Correlation Coefficient	p-value	95% CI
Value	Precision	NA	NA	(NA, NA)
Value	IndicatorOrder	0.4332	0.00524	(0.1406, 0.6561)
Value	CharacteristicId	0.2317	0.150308	(-0.086, 0.5066)
Value	CharacteristicOrder	-0.124	0.445688	(-0.4194, 0.195)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	0.1907	0.238394	(-0.1284, 0.4741)
Value	SurveyYearLabel	-0.0636	0.696539	(-0.3678, 0.2529)
Value	CILow	0.9948	0	(0.9889, 0.9975)
Value	CIHigh	0.9969	0	(0.9934, 0.9985)
Precision	IndicatorOrder	NA	NA	(NA, NA)
Precision	CharacteristicId	NA	NA	(NA, NA)
Precision	CharacteristicOrder	NA	NA	(NA, NA)

Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	NA	NA	(NA, NA)
Precision	SurveyYearLabel	NA	NA	(NA, NA)
Precision	CILow	NA	NA	(NA, NA)
Precision	CIHigh	NA	NA	(NA, NA)
IndicatorOrder	CharacteristicId	-0.2955	0.064099	(-0.5559, 0.0176)
IndicatorOrder	CharacteristicOrder	-0.8038	0	(-0.892, -0.6567)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	-0.058	0.722102	(-0.363, 0.2581)
IndicatorOrder	SurveyYearLabel	0	1	(-0.3115, 0.3115)
IndicatorOrder	CILow	0.0538	0.777866	(-0.3126, 0.4062)
IndicatorOrder	CIHigh	0.0044	0.981483	(-0.3564, 0.3641)
CharacteristicId	CharacteristicOrder	0.8059	0	(0.6601, 0.8932)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	0.5659	0.000141	(0.3089, 0.7459)

CharacteristicId	SurveyYearLabel	0	1	(-0.3115, 0.3115)
CharacteristicId	CILow	-0.0534	0.779323	(-0.4059, 0.3129)
CharacteristicId	CIHigh	-0.0041	0.982906	(-0.3638, 0.3567)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	0.3884	0.013267	(0.0874, 0.6244)
CharacteristicOrder	SurveyYearLabel	0	1	(-0.3115, 0.3115)
CharacteristicOrder	CILow	-0.0534	0.779323	(-0.4059, 0.3129)
CharacteristicOrder	CIHigh	-0.0041	0.982906	(-0.3638, 0.3567)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	CILow	NA	NA	(NA, NA)
IsTotal	CIHigh	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	0	1	(-0.3115, 0.3115)
IsPreferred	CILow	-0.1068	0.574381	(-0.4497, 0.2636)
IsPreferred	CIHigh	-0.0082	0.965819	(-0.3674, 0.3531)

SurveyYearLabel	CILow	-0.2666	0.154377	(-0.572, 0.1036)
SurveyYearLabel	CIHigh	-0.1772	0.348923	(-0.5052, 0.1956)
CILow	CIHigh	0.9846	0	(0.9675, 0.9927)

Include and Exclude

Inclusion Criteria

Fields were selected for inclusion if they demonstrated a statistically significant linear relationship with another key variable and did not introduce redundancy.

- Value and IndicatorOrder: IndicatorOrder has a statistically significant correlation with Value (0.4332, p-value = 0.00524), making it a strong candidate for inclusion as a predictor.
- CharacteristicOrder, CharacteristicId and IsPreferred: These fields are all highly correlated with each other. CharacteristicOrder is strongly negatively correlated with IndicatorOrder (-0.8038) and positively correlated with CharacteristicId (0.8059) and IsPreferred (0.3884). This relationships suggests that these fields contain valuable, interconnected information and should be included.

Exclusion Criteria

Fields were excluded for having weak, non-significant, or redundant relationships with other variables.

- Precision, IsTotal and SurveyYearLabel: These fields show "NA" for all correlations or a p-value of 1 with other variables, indicating no computable or meaningful linear relationship. They should be excluded from the model.
- CILow and CIHigh: These fields are almost perfectly correlated with Value (0.9948 and 0.9969). This is a strong sign of data leakage and redundancy, meaning they are likely calculated directly from the 'Value' field. Including them would make the model artificially accurate on the training data but useless for new, unseen data. Therefore, they must be excluded.

- Value and CharacteristicId/CharacteristicOrder/IsPreferred/SurveyYearLabel: The p-values for these correlations with Value are all greater than 0.05, indicating a non-significant linear relationship. While some of these fields (e.g., CharacteristicId) are correlated with other key variables, they don't appear to be good direct predictors of 'Value' and may be better explored in a more complex model.

Attribute Weighting Strategy

An attribute weighting strategy is not needed at this stage of the analysis. The correlation tests revealed that only IndicatorOrder has a statistically significant linear relationship with the target variable, Value. The other significant correlations are between the predictor variables themselves (e.g., IndicatorOrder ~ CharacteristicOrder). A weighting strategy would be more appropriate for a later stage when multiple features are selected as direct predictors and their relative importance needs to be prioritized.

Maternal Mortality

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
Value	Precision	-26.24	0.250539	(-0.6235, 0.1909)
Value	IndicatorOrder	-47.64	0.029007	(-0.7532, -0.0563)
Value	CharacteristicId	47.66	0.028947	(0.0565, 0.7533)
Value	CharacteristicOrder	47.66	0.028947	(0.0565, 0.7533)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	NA	NA	(NA, NA)
Value	SurveyYearLabel	-16.75	0.467882	(-0.5588, 0.2847)

Precision	IndicatorOrder	-2.67	0.908556	(-0.4532, 0.4097)
Precision	CharacteristicId	2.71	0.907004	(-0.4093, 0.4535)
Precision	CharacteristicOrder	2.71	0.907004	(-0.4093, 0.4535)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	NA	NA	(NA, NA)
Precision	SurveyYearLabel	-2.71	0.907004	(-0.4535, 0.4093)
IndicatorOrder	CharacteristicId	-100	0	(-1.0, -1.0)
IndicatorOrder	CharacteristicOrder	-100	0	(-1.0, -1.0)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	NA	NA	(NA, NA)
IndicatorOrder	SurveyYearLabel	4.55	0.844596	(-0.3939, 0.468)
CharacteristicId	CharacteristicOrder	100	0	(1.0, 1.0)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	NA	NA	(NA, NA)
CharacteristicId	SurveyYearLabel	-4.55	0.844888	(-0.468, 0.394)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	NA	NA	(NA, NA)

CharacteristicOrder	SurveyYearLabel	-4.55	0.844888	(-0.468, 0.394)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	NA	NA	(NA, NA)

Include And Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships with the target variable, Value and other key fields.

Value, IndicatorOrder, CharacteristicId and CharacteristicOrder: These fields are all highly interconnected with statistically significant correlations. IndicatorOrder is significantly correlated with Value (-47.64%, p = 0.029) and it also has a perfect negative correlation with CharacteristicId and CharacteristicOrder (-100%, p = 0). CharacteristicId and CharacteristicOrder are perfectly correlated with each other (100%, p = 0) and significantly correlated with Value (47.66%, p = 0.029). This network of strong relationships suggests that all three fields are crucial for the model. However, since IndicatorOrder, CharacteristicId and CharacteristicOrder are perfectly correlated, only one of them should be chosen to represent this group in the model to avoid redundancy.

Exclusion Criteria

Fields were excluded if they lacked a statistically significant relationship with **Value** or were not computable.

Precision, IsTotal, IsPreferred and SurveyYearLabel: These fields show non-significant p-values (p > 0.05) when correlated with Value. Their relationships are too weak to be useful for linear prediction. Furthermore, IsTotal and IsPreferred have no computable correlations ("NA"), indicating they are likely non-numeric or have data quality issues preventing a standard correlation analysis. These fields should be excluded.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to rank the importance of the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The only fields with a statistically significant linear correlation to Value are IndicatorOrder, CharacteristicId and CharacteristicOrder.
2. Use Absolute Correlation Values:
   - |Correlation of Value with IndicatorOrder| = |−0.4764| = 0.4764
   - |Correlation of Value with CharacteristicId| = |0.4766| = 0.4766
   - |Correlation of Value with CharacteristicOrder| = |0.4766| = 0.4766
3. Choose One Redundant Field: Since IndicatorOrder, CharacteristicId and CharacteristicOrder are all perfectly correlated, we only need to include one of them in the final model. For instance, if we select IndicatorOrder, its weight would be 0.4764. The other two fields would not be included in the weighting strategy.

Final weights for the model would be determined after a single representative from the highly correlated group (e.g., IndicatorOrder) is chosen and potentially combined with any other significant predictors (none were found in this analysis).

DHS Quickstat

Variable 1	Variable 2	Correlation Coefficient	p-value	95% CI
Value	Precision	-0.3888	0.004392	(-0.5982, -0.1297)
Value	IndicatorOrder	0.0258	0.85613	(-0.2489, 0.2966)
Value	CharacteristicId	-0.1042	0.462154	(-0.3667, 0.1736)

Value	CharacteristicOrder	-0.1042	0.462154	(-0.3667, 0.1736)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	0.0194	0.891238	(-0.2548, 0.2908)
Value	SurveyYearLabel	0.1031	0.466871	(-0.1747, 0.3657)
Value	DenominatorWeighted	0.1874	0.288532	(-0.161, 0.4943)
Value	DenominatorUnweighted	0.2323	0.186083	(-0.1149, 0.5289)
Value	CILow	0.9976	0	(0.9921, 0.9993)
Value	CIHigh	0.9998	0	(0.9992, 0.9999)
Precision	IndicatorOrder	0.2241	0.110196	(-0.052, 0.4684)
Precision	CharacteristicId	0.237	0.09075	(-0.0384, 0.4789)
Precision	CharacteristicOrder	0.237	0.09075	(-0.0384, 0.4789)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	0.1929	0.170641	(-0.0844, 0.4425)

Precision	SurveyYearLabel	0.0535	0.7063	(-0.2226, 0.3217)
Precision	DenominatorWeighted	NA	NA	(NA, NA)
Precision	DenominatorUnweighted	NA	NA	(NA, NA)
Precision	CILow	-0.3334	0.244029	(-0.7341, 0.2395)
Precision	CIHigh	-0.2872	0.319487	(-0.7096, 0.2872)
IndicatorOrder	CharacteristicId	0.6141	0.000001	(0.4098, 0.7597)
IndicatorOrder	CharacteristicOrder	0.6141	0.000001	(0.4098, 0.7597)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	0.0406	0.774881	(-0.2349, 0.3101)
IndicatorOrder	SurveyYearLabel	0.1696	0.229409	(-0.1083, 0.4229)
IndicatorOrder	DenominatorWeighted	0.6519	0.00003	(0.4025, 0.8112)
IndicatorOrder	DenominatorUnweighted	0.6989	0.000004	(0.4724, 0.8388)
IndicatorOrder	CILow	-0.1773	0.544342	(-0.647, 0.39)
IndicatorOrder	CIHigh	-0.1244	0.671779	(-0.6144, 0.4349)

CharacteristicId	CharacteristicOrder	1	0	(1.0, 1.0)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	0.1125	0.427162	(-0.1655, 0.3739)
CharacteristicId	SurveyYearLabel	0.2729	0.050261	(0, 0.508)
CharacteristicId	DenominatorWeighted	0.3551	0.039332	(0.0192, 0.6189)
CharacteristicId	DenominatorUnweighted	0.4149	0.014693	(0.0893, 0.6604)
CharacteristicId	CILow	-0.3334	0.244029	(-0.7341, 0.2395)
CharacteristicId	CIHigh	-0.2872	0.319487	(-0.7096, 0.2872)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	0.1125	0.427162	(-0.1655, 0.3739)
CharacteristicOrder	SurveyYearLabel	0.2729	0.050261	(0, 0.508)
CharacteristicOrder	DenominatorWeighted	0.3551	0.039332	(0.0192, 0.6189)
CharacteristicOrder	DenominatorUnweighted	0.4149	0.014693	(0.0893, 0.6604)
CharacteristicOrder	CILow	-0.3334	0.244029	(-0.7341, 0.2395)
CharacteristicOrder	CIHigh	-0.2872	0.319487	(-0.7096, 0.2872)

IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsTotal	CILow	NA	NA	(NA, NA)
IsTotal	CIHigh	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	0.0535	0.7063	(-0.2226, 0.3217)
IsPreferred	DenominatorWeighted	0.227	0.196641	(-0.1204, 0.5249)
IsPreferred	DenominatorUnweighted	0.221	0.209127	(-0.1266, 0.5203)
IsPreferred	CILow	0.2189	0.452167	(-0.3527, 0.6715)
IsPreferred	CIHigh	0.2672	0.355727	(-0.3069, 0.6987)
SurveyYearLabel	DenominatorWeighted	-0.1722	0.330205	(-0.4823, 0.1763)
SurveyYearLabel	DenominatorUnweighted	-0.1163	0.512592	(-0.4372, 0.231)
SurveyYearLabel	CILow	0.167	0.568154	(-0.3989, 0.6408)
SurveyYearLabel	CIHigh	0.208	0.475537	(-0.3626, 0.6652)

DenominatorWeighted	DenominatorUnweighted	0.9789	0	(0.9578, 0.9895)
CILow	CIHigh	0.9958	0	(0.9864, 0.9987)

Include And Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships with the target variable, Value, and other key fields.

- Value and Precision: The negative correlation of -0.3888 with a p-value of 0.004392 indicates a statistically significant relationship. Precision should be included as a predictor.
- IndicatorOrder, CharacteristicId, and CharacteristicOrder: This group of fields has strong correlations with each other (e.g., IndicatorOrder and CharacteristicId have a correlation of 0.6141 with p = 1e-06). They also have significant correlations with the denominator fields. To avoid redundancy, only one of these fields should be selected to represent the group.
- DenominatorWeighted and DenominatorUnweighted: These fields have a very high correlation of 0.9789 with p = 0, indicating strong redundancy. They also have significant correlations with IndicatorOrder, CharacteristicId, and CharacteristicOrder. To avoid multicollinearity, only one of these should be included.

Exclusion Criteria

Fields were excluded if they lacked a statistically significant relationship with Value or were not computable.

- IndicatorOrder, CharacteristicId, CharacteristicOrder, IsTotal, IsPreferred, and SurveyYearLabel: While some of these fields (e.g., IndicatorOrder) have significant correlations with other variables, their correlations with the target variable, Value, are not statistically significant ($p > 0.05$), except for Precision, and they are either redundant with other fields or not computable. IsTotal and IsPreferred have no computable correlations ("NA"), and should be excluded. SurveyYearLabel has weak and non-significant correlations with all other fields and should be excluded.

- CILow and CIHigh: These fields are extremely highly correlated with Value (0.9976 and 0.9998). This is a strong sign of data leakage, as they are likely derived from 'Value'. Including them would lead to an overfit model that performs poorly on new data. They must be excluded.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to prioritize the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The only field with a statistically significant linear correlation to Value is Precision.
2. HandleRedundancy:The other highly correlated groups (IndicatorOrder/ CharacteristicId/CharacteristicOrder)and(DenominatorWeighted/DenominatorUnweighted) do not have a significant correlation with Value. Therefore, they should not be included in the weighting strategy as direct predictors.
3. Calculate Weights: The only predictor with a significant relationship is Precision, with an absolute correlation of |−0.3888| = 0.3888.
4. Normalize Weights: Since there is only one predictor, its normalized weight is 1.

The weighting strategy is not very useful in this case, as there is only one direct predictor. A more complex model might benefit from including the other correlated fields, but based on this correlation analysis, Precision is the only statistically significant linear predictor of Value.

Immunization

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
DataId	Value	9.45	0.312815	(-0.0893, 0.2722)

DataId	Precision	5.65	0.546795	(-0.1271, 0.2364)
DataId	SurveyYear	-27.24	0.00309	(-0.4332, -0.0948)
DataId	IndicatorOrder	22.97	0.013128	(0.0494, 0.3954)
DataId	CharacteristicId	-41.32	0.000004	(-0.5538, -0.2497)
DataId	CharacteristicOrder	-41.32	0.000004	(-0.5538, -0.2497)
DataId	ByVariableId	-41.32	0.000004	(-0.5538, -0.2497)
DataId	IsTotal	NA	NA	(NA, NA)
DataId	IsPreferred	44.75	0	(0.2888, 0.5823)
DataId	SurveyYearLabel	-27.24	0.00309	(-0.4332, -0.0948)
DataId	DenominatorWeighted	43.63	0.000002	(0.2695, 0.5776)
DataId	DenominatorUnweighted	41.65	0.000007	(0.247, 0.5613)
Value	Precision	-74.76	0	(-0.8184, -0.6545)
Value	SurveyYear	-13.11	0.160698	(-0.3061, 0.0525)

Value	IndicatorOrder	19.51	0.0358	(0.0133, 0.3645)
Value	CharacteristicId	-11.9	0.203092	(-0.295, 0.0647)
Value	CharacteristicOrder	-11.9	0.203092	(-0.295, 0.0647)
Value	ByVariableId	-11.9	0.203092	(-0.295, 0.0647)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	7	0.454974	(-0.1137, 0.2492)
Value	SurveyYearLabel	-13.11	0.160698	(-0.3061, 0.0525)
Value	DenominatorWeighted	18.17	0.05984	(-0.0075, 0.3584)
Value	DenominatorUnweighted	18.14	0.060292	(-0.0079, 0.3581)
Precision	SurveyYear	3.46	0.712238	(-0.1486, 0.2156)
Precision	IndicatorOrder	-2.55	0.785678	(-0.2069, 0.1575)
Precision	CharacteristicId	1.38	0.883099	(-0.1689, 0.1956)
Precision	CharacteristicOrder	1.38	0.883099	(-0.1689, 0.1956)

Precision	ByVariableId	1.38	0.883099	(-0.1689, 0.1956)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	-0.79	0.933137	(-0.1899, 0.1747)
Precision	SurveyYearLabel	3.46	0.712238	(-0.1486, 0.2156)
Precision	DenominatorWeighted	-1.23	0.899638	(-0.2008, 0.1771)
Precision	DenominatorUnweighted	-1.23	0.899328	(-0.2008, 0.1771)
SurveyYear	IndicatorOrder	-7.97	0.395158	(-0.2583, 0.1041)
SurveyYear	CharacteristicId	4.9	0.601707	(-0.1346, 0.2292)
SurveyYear	CharacteristicOrder	4.9	0.601707	(-0.1346, 0.2292)
SurveyYear	ByVariableId	4.9	0.601707	(-0.1346, 0.2292)
SurveyYear	IsTotal	NA	NA	(NA, NA)
SurveyYear	IsPreferred	-2.79	0.765931	(-0.2092, 0.1552)
SurveyYear	SurveyYearLabel	100	0	(1.0, 1.0)
SurveyYear	DenominatorWeighted	-46.77	0	(-0.6033, -0.3057)

SurveyYear	DenominatorUnweighted	-46.81	0	(-0.6037, -0.3062)
IndicatorOrder	CharacteristicId	-85.25	0	(-0.8956, -0.7935)
IndicatorOrder	CharacteristicOrder	-85.25	0	(-0.8956, -0.7935)
IndicatorOrder	ByVariableId	-85.25	0	(-0.8956, -0.7935)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	48.64	0	(0.3337, 0.6143)
IndicatorOrder	SurveyYearLabel	-7.97	0.395158	(-0.2583, 0.1041)
IndicatorOrder	DenominatorWeighted	72.73	0	(0.6241, 0.8055)
IndicatorOrder	DenominatorUnweighted	72.56	0	(0.6219, 0.8043)
CharacteristicId	CharacteristicOrder	100	0	(1.0, 1.0)
CharacteristicId	ByVariableId	100	0	(1.0, 1.0)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	-57.06	0	(-0.682, -0.4334)
CharacteristicId	SurveyYearLabel	4.9	0.601707	(-0.1346, 0.2292)

CharacteristicId	DenominatorWeighted	-44.77	0.000001	(-0.587, -0.2827)
CharacteristicId	DenominatorUnweighted	-44.4	0.000001	(-0.584, -0.2784)
CharacteristicOrder	ByVariableId	100	0	(1.0, 1.0)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	-57.06	0	(-0.682, -0.4334)
CharacteristicOrder	SurveyYearLabel	4.9	0.601707	(-0.1346, 0.2292)
CharacteristicOrder	DenominatorWeighted	-44.77	0.000001	(-0.587, -0.2827)
CharacteristicOrder	DenominatorUnweighted	-44.4	0.000001	(-0.584, -0.2784)
ByVariableId	IsTotal	NA	NA	(NA, NA)
ByVariableId	IsPreferred	-57.06	0	(-0.682, -0.4334)
ByVariableId	SurveyYearLabel	4.9	0.601707	(-0.1346, 0.2292)
ByVariableId	DenominatorWeighted	-44.77	0.000001	(-0.587, -0.2827)
ByVariableId	DenominatorUnweighted	-44.4	0.000001	(-0.584, -0.2784)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)

IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	-2.79	0.765931	(-0.2092, 0.1552)
IsPreferred	DenominatorWeighted	26.35	0.005856	(0.0785, 0.431)
IsPreferred	DenominatorUnweighted	25.47	0.007813	(0.069, 0.4233)
SurveyYearLabel	DenominatorWeighted	-46.77	0	(-0.6033, -0.3057)
SurveyYearLabel	DenominatorUnweighted	-46.81	0	(-0.6037, -0.3062)
DenominatorWeighted	DenominatorUnweighted	100	0	(0.9999, 1.0)

Include and Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships with the target variable, Value, and other key fields.

- Value and Precision: The negative correlation of -74.76% with a p-value of 0 indicates a statistically significant and strong linear relationship. Precision should be included as a predictor.
- Value and IndicatorOrder: The positive correlation of 19.51% with a p-value of 0.0358 is statistically significant ($p < 0.05$). This field should also be included as a predictor.
- DataId, SurveyYear, SurveyYearLabel, and DenominatorWeighted/Unweighted: DataId and SurveyYear/SurveyYearLabel are significantly correlated with each other and with the Denominator fields. These fields could be useful for a more complex

model (e.g., a time-series model), but as direct predictors of Value, their relationships are weak or non-significant.

- IndicatorOrder, CharacteristicId, CharacteristicOrder, and ByVariableId: This group of fields is highly interconnected with strong correlations (e.g., IndicatorOrder and CharacteristicId have a correlation of -85.25%, p = 0). To avoid redundancy, only one of these fields should be selected to represent the group.

- DenominatorWeighted and DenominatorUnweighted: These fields have a perfect correlation of 100% (p = 0), indicating strong redundancy. Only one should be included, preferably DenominatorWeighted, as it has a slightly stronger correlation with Value (p = 0.05984) than DenominatorUnweighted (p = 0.060292), although neither is statistically significant.

Exclusion Criteria

Fields were excluded if they lacked a statistically significant relationship with Value or were not computable.

- CharacteristicId, CharacteristicOrder, ByVariableId, IsPreferred, SurveyYear, SurveyYearLabel, DenominatorWeighted, and DenominatorUnweighted: While some of these fields (e.g., CharacteristicId) have significant correlations with other variables, their linear correlations with the target variable, Value, are not statistically significant ($p > 0.05$). Therefore, they should be excluded as direct predictors.

- IsTotal: This field has no computable correlations ("NA"), indicating it is likely non-numeric or has data quality issues. It should be excluded.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to prioritize the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The fields with a statistically significant linear correlation to Value are Precision and IndicatorOrder.

2. Calculate Weights: The absolute correlations with Value are:

   - |Correlation of Value with Precision| = |−0.7476| = 0.7476

- |Correlation of Value with IndicatorOrder| = |0.1951| = 0.1951

3. Normalize Weights: Sum the absolute values to get the normalization factor: 0.7476 + 0.1951 = 0.9427.
    - Precision: 0.7476/0.9427 = 0.7930
    - IndicatorOrder: 0.1951/0.9427 = 0.2070

IYCF

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
DataId	Value	38.12	0.080043	(-0.0481, 0.6917)
DataId	Precision	NA	NA	(NA, NA)
DataId	SurveyYear	-58.4	0.004324	(-0.8069, -0.2154)
DataId	IndicatorOrder	-51.81	0.013502	(-0.7713, -0.1235)
DataId	CharacteristicId	-0.29	0.98966	(-0.424, 0.4192)
DataId	CharacteristicOrder	-0.29	0.98966	(-0.424, 0.4192)
DataId	ByVariableId	NA	NA	(NA, NA)
DataId	IsTotal	NA	NA	(NA, NA)
DataId	IsPreferred	NA	NA	(NA, NA)
DataId	SurveyYearLabel	-58.4	0.004324	(-0.8069, -0.2154)

DataId	DenominatorWeighted	44.28	0.050577	(0.0003, 0.7402)
DataId	DenominatorUnweighted	43.52	0.055127	(-0.009, 0.736)
Value	Precision	NA	NA	(NA, NA)
Value	SurveyYear	-3.43	0.87939	(-0.4494, 0.393)
Value	IndicatorOrder	-25.25	0.256906	(-0.6093, 0.1892)
Value	CharacteristicId	-11.28	0.617253	(-0.5101, 0.3242)
Value	CharacteristicOrder	-11.28	0.617253	(-0.5101, 0.3242)
Value	ByVariableId	NA	NA	(NA, NA)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	NA	NA	(NA, NA)
Value	SurveyYearLabel	-3.43	0.87939	(-0.4494, 0.393)
Value	DenominatorWeighted	43.04	0.058157	(-0.0149, 0.7333)
Value	DenominatorUnweighted	42.35	0.062797	(-0.0234, 0.7293)
Precision	SurveyYear	NA	NA	(NA, NA)
Precision	IndicatorOrder	NA	NA	(NA, NA)

Precision	CharacteristicId	NA	NA	(NA, NA)
Precision	CharacteristicOrder	NA	NA	(NA, NA)
Precision	ByVariableId	NA	NA	(NA, NA)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	NA	NA	(NA, NA)
Precision	SurveyYearLabel	NA	NA	(NA, NA)
Precision	DenominatorWeighted	NA	NA	(NA, NA)
Precision	DenominatorUnweighted	NA	NA	(NA, NA)
SurveyYear	IndicatorOrder	44.27	0.039099	(0.0259, 0.7284)
SurveyYear	CharacteristicId	-25.88	0.24489	(-0.6135, 0.1828)
SurveyYear	CharacteristicOrder	-25.88	0.24489	(-0.6135, 0.1828)
SurveyYear	ByVariableId	NA	NA	(NA, NA)
SurveyYear	IsTotal	NA	NA	(NA, NA)
SurveyYear	IsPreferred	NA	NA	(NA, NA)
SurveyYear	SurveyYearLabel	100	0	(1.0, 1.0)
SurveyYear	DenominatorWeighted	-25.55	0.276896	(-0.6271, 0.2108)
SurveyYear	DenominatorUnweighted	-26.44	0.260016	(-0.6329, 0.2018)

IndicatorOrder	CharacteristicId	-44.45	0.038227	(-0.7294, -0.0281)
IndicatorOrder	CharacteristicOrder	-44.45	0.038227	(-0.7294, -0.0281)
IndicatorOrder	ByVariableId	NA	NA	(NA, NA)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	NA	NA	(NA, NA)
IndicatorOrder	SurveyYearLabel	44.27	0.039099	(0.0259, 0.7284)
IndicatorOrder	DenominatorWeighted	-31.86	0.171016	(-0.6671, 0.1443)
IndicatorOrder	DenominatorUnweighted	-31.73	0.172835	(-0.6663, 0.1457)
CharacteristicId	CharacteristicOrder	100	0	(1.0, 1.0)
CharacteristicId	ByVariableId	NA	NA	(NA, NA)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	NA	NA	(NA, NA)
CharacteristicId	SurveyYearLabel	-25.88	0.24489	(-0.6135, 0.1828)
CharacteristicId	DenominatorWeighted	-62.66	0.003115	(-0.837, -0.2547)
CharacteristicId	DenominatorUnweighted	-62.3	0.003342	(-0.8353, -0.2492)
CharacteristicOrder	ByVariableId	NA	NA	(NA, NA)

CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	NA	NA	(NA, NA)
CharacteristicOrder	SurveyYearLabel	-25.88	0.24489	(-0.6135, 0.1828)
CharacteristicOrder	DenominatorWeighted	-62.66	0.003115	(-0.837, -0.2547)
CharacteristicOrder	DenominatorUnweighted	-62.3	0.003342	(-0.8353, -0.2492)
ByVariableId	IsTotal	NA	NA	(NA, NA)
ByVariableId	IsPreferred	NA	NA	(NA, NA)
ByVariableId	SurveyYearLabel	NA	NA	(NA, NA)
ByVariableId	DenominatorWeighted	NA	NA	(NA, NA)
ByVariableId	DenominatorUnweighted	NA	NA	(NA, NA)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	NA	NA	(NA, NA)
IsPreferred	DenominatorWeighted	NA	NA	(NA, NA)
IsPreferred	DenominatorUnweighted	NA	NA	(NA, NA)
SurveyYearLabel	DenominatorWeighted	-25.55	0.276896	(-0.6271, 0.2108)

SurveyYearLabel	DenominatorUnweighted	-26.44	0.260016	(-0.6329, 0.2018)
DenominatorWeighted	DenominatorUnweighted	99.98	0	(0.9994, 0.9999)

Include and Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships.

- SurveyYear and SurveyYearLabel: These fields have a perfect positive correlation of 100% (p = 0). They also have significant correlations with other variables like DataId and IndicatorOrder. To avoid redundancy, only one of them should be included, for instance, SurveyYear.
- IndicatorOrder, CharacteristicId, and CharacteristicOrder: This group of fields is significantly correlated with each other and with the SurveyYear and Denominator fields. To avoid redundancy, only one field from this highly correlated group should be included in the model, as their correlation with Value is not statistically significant.
- DenominatorWeighted and DenominatorUnweighted: These fields have a very high correlation of 99.98% (p = 0), indicating strong redundancy. They also have significant correlations with the Characteristic fields. To avoid multicollinearity, only one of them should be included, preferably DenominatorWeighted, as it has a slightly stronger correlation with Value (p = 0.058) than DenominatorUnweighted (p = 0.062), although neither is statistically significant.
- DataId: This field has significant correlations with SurveyYear, IndicatorOrder, CharacteristicId, and the Denominator fields. It also has a correlation with Value of 38.12%, but the p-value is 0.08, which is above the standard significance threshold of 0.05. It should be considered for inclusion, but with caution.

Exclusion Criteria

Fields were excluded because they lacked a statistically significant relationship with Value or were not computable.

- Value, IndicatorOrder, CharacteristicId, CharacteristicOrder, IsTotal, IsPreferred, ByVariableId, and Precision:
  - Value is the target variable, so it is not a predictor.
  - IsTotal, IsPreferred, ByVariableId, and Precision have many "NA" correlations, suggesting they contain non-numeric data or have other issues preventing correlation analysis. These should be excluded.
  - IndicatorOrder, CharacteristicId, and CharacteristicOrder have non-significant p-values ($p > 0.05$) when correlated with Value. Therefore, they should be excluded as direct predictors of Value.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to prioritize the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The only predictor with a p-value close to the significance threshold is DataId ($p = 0.08$). The DenominatorWeighted and DenominatorUnweighted fields also have p-values close to 0.05. No fields have a p-value less than 0.05.
2. Handle Redundancy: If we decide to include the fields with p-values close to the significance threshold, we must address the redundancy:
   - Choose one from the highly correlated group: CharacteristicId and CharacteristicOrder.
   - Choose one from the highly correlated group: DenominatorWeighted and DenominatorUnweighted.
3. Calculate Weights: Given that no fields have a statistically significant linear correlation with Value, a weighting strategy based purely on correlation coefficients is not appropriate. A more complex model or a different feature selection method (e.g., recursive feature elimination, which does not rely on linear correlation) would be needed.

This analysis shows that in this particular dataset, linear correlation alone is not sufficient to identify strong predictors for the Value field.

Literacy

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
DataId	Value	-47.74	0.033269	(-0.7595, -0.0443)
DataId	Precision	55.87	0.010447	(0.1544, 0.8028)
DataId	SurveyYear	NA	NA	(NA, NA)
DataId	IndicatorOrder	-52.01	0.018736	(-0.7825, -0.1008)
DataId	CharacteristicId	NA	NA	(NA, NA)
DataId	CharacteristicOrder	NA	NA	(NA, NA)
DataId	ByVariableId	NA	NA	(NA, NA)
DataId	IsTotal	NA	NA	(NA, NA)
DataId	IsPreferred	NA	NA	(NA, NA)
DataId	SurveyYearLabel	NA	NA	(NA, NA)
DataId	DenominatorWeighted	47.59	0.045893	(0.0116, 0.7714)
DataId	DenominatorUnweighted	99.9	0	(0.9972, 0.9996)
Value	Precision	-85.03	0.000002	(-0.9394, -0.6538)

Value	SurveyYear	NA	NA	(NA, NA)
Value	IndicatorOrder	-18.69	0.430127	(-0.5813, 0.2787)
Value	CharacteristicId	NA	NA	(NA, NA)
Value	CharacteristicOrder	NA	NA	(NA, NA)
Value	ByVariableId	NA	NA	(NA, NA)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	NA	NA	(NA, NA)
Value	SurveyYearLabel	NA	NA	(NA, NA)
Value	DenominatorWeighted	17.43	0.488994	(-0.3184, 0.593)
Value	DenominatorUnweighted	14.52	0.565341	(-0.345, 0.5732)
Precision	SurveyYear	NA	NA	(NA, NA)
Precision	IndicatorOrder	-4.14	0.862521	(-0.4752, 0.4086)
Precision	CharacteristicId	NA	NA	(NA, NA)
Precision	CharacteristicOrder	NA	NA	(NA, NA)
Precision	ByVariableId	NA	NA	(NA, NA)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	NA	NA	(NA, NA)
Precision	SurveyYearLabel	NA	NA	(NA, NA)

Precision	DenominatorWeighted	0	1	(-0.4669, 0.4669)
Precision	DenominatorUnweighted	0	1	(-0.4669, 0.4669)
SurveyYear	IndicatorOrder	NA	NA	(NA, NA)
SurveyYear	CharacteristicId	NA	NA	(NA, NA)
SurveyYear	CharacteristicOrder	NA	NA	(NA, NA)
SurveyYear	ByVariableId	NA	NA	(NA, NA)
SurveyYear	IsTotal	NA	NA	(NA, NA)
SurveyYear	IsPreferred	NA	NA	(NA, NA)
SurveyYear	SurveyYearLabel	NA	NA	(NA, NA)
SurveyYear	DenominatorWeighted	NA	NA	(NA, NA)
SurveyYear	DenominatorUnweighted	NA	NA	(NA, NA)
IndicatorOrder	CharacteristicId	NA	NA	(NA, NA)
IndicatorOrder	CharacteristicOrder	NA	NA	(NA, NA)
IndicatorOrder	ByVariableId	NA	NA	(NA, NA)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	NA	NA	(NA, NA)
IndicatorOrder	SurveyYearLabel	NA	NA	(NA, NA)
IndicatorOrder	DenominatorWeighted	-99.84	0	(-0.9994, -0.9955)

IndicatorOrder	DenominatorUnweighted	-99.86	0	(-0.9995, -0.996)
CharacteristicId	CharacteristicOrder	NA	NA	(NA, NA)
CharacteristicId	ByVariableId	NA	NA	(NA, NA)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	NA	NA	(NA, NA)
CharacteristicId	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicId	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicId	DenominatorUnweighted	NA	NA	(NA, NA)
CharacteristicOrder	ByVariableId	NA	NA	(NA, NA)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	NA	NA	(NA, NA)
CharacteristicOrder	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicOrder	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicOrder	DenominatorUnweighted	NA	NA	(NA, NA)
ByVariableId	IsTotal	NA	NA	(NA, NA)
ByVariableId	IsPreferred	NA	NA	(NA, NA)
ByVariableId	SurveyYearLabel	NA	NA	(NA, NA)
ByVariableId	DenominatorWeighted	NA	NA	(NA, NA)
ByVariableId	DenominatorUnweighted	NA	NA	(NA, NA)
IsTotal	IsPreferred	NA	NA	(NA, NA)

IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	NA	NA	(NA, NA)
IsPreferred	DenominatorWeighted	NA	NA	(NA, NA)
IsPreferred	DenominatorUnweighted	NA	NA	(NA, NA)
SurveyYearLabel	DenominatorWeighted	NA	NA	(NA, NA)
SurveyYearLabel	DenominatorUnweighted	NA	NA	(NA, NA)
DenominatorWeighted	DenominatorUnweighted	100	0	(1.0, 1.0)

Include and Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships.

- Value and Precision: The negative correlation of -85.03% with a p-value of 0.000002 indicates a statistically significant and very strong linear relationship. Precision should be a primary predictor.
- DataId and Value: The negative correlation of -47.74% with a p-value of 0.033269 is statistically significant ($p < 0.05$). This field should also be included as a predictor.
- DataId and Precision: The positive correlation of 55.87% with a p-value of 0.010447 is statistically significant. While not a direct predictor of Value, its strong correlation with another key predictor (Precision) indicates its potential importance.
- DataId and IndicatorOrder: The negative correlation of -52.01% with a p-value of 0.018736 is statistically significant. This field is also correlated with DenominatorWeighted and DenominatorUnweighted, suggesting its role in the data structure.
- DataId and DenominatorWeighted: The positive correlation of 47.59% with a p-value of 0.045893 is statistically significant. This field should be included.

Exclusion Criteria

Fields were excluded if they lacked a statistically significant relationship with the target variable Value or were not computable.

- IndicatorOrder, DenominatorWeighted, and DenominatorUnweighted: While these fields have strong correlations with other variables, their linear correlations with the target variable, Value, are not statistically significant ($p > 0.05$). DenominatorWeighted and DenominatorUnweighted have a perfect correlation of 100%, making them redundant.
- SurveyYear, CharacteristicId, CharacteristicOrder, ByVariableId, IsTotal, IsPreferred, and SurveyYearLabel: These fields show "NA" for all or most correlations, indicating they are likely non-numeric or have data quality issues that prevent correlation analysis. These fields should be excluded.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to prioritize the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The fields with a statistically significant linear correlation to Value are Precision and DataId.
2. Calculate Weights: The absolute correlations with Value are:
    - $|\text{Correlation of Value with Precision}| = |-0.8503| = 0.8503$
    - $|\text{Correlation of Value with DataId}| = |-0.4774| = 0.4774$
3. Normalize Weights: Sum the absolute values to get the normalization factor: $0.8503 + 0.4774 = 1.3277$.
    - Precision: $0.8503/1.3277 = 0.6404$
    - DataId: $0.4774/1.3277 = 0.3596$

Water

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
DataId	Value	6.81	0.500727	(-0.13, 0.261)
DataId	Precision	-2.25	0.8239	(-0.218, 0.1747)
DataId	SurveyYear	-97.34	0	(-0.982, -0.9606)
DataId	IndicatorOrder	-22.53	0.024184	(-0.4039, -0.0303)
DataId	CharacteristicId	NA	NA	(NA, NA)
DataId	CharacteristicOrder	NA	NA	(NA, NA)
DataId	ByVariableId	NA	NA	(NA, NA)
DataId	IsTotal	NA	NA	(NA, NA)
DataId	IsPreferred	NA	NA	(NA, NA)
DataId	SurveyYearLabel	-97.34	0	(-0.982, -0.9606)
DataId	DenominatorWeighted	36.55	0.00025	(0.1781, 0.5273)
DataId	DenominatorUnweighted	35.22	0.000434	(0.1632, 0.5162)
Value	Precision	-84.13	0	(-0.8906, -0.7726)

Value	SurveyYear	-9.75	0.334394	(-0.2884, 0.1008)
Value	IndicatorOrder	36.35	0.000201	(0.1799, 0.5226)
Value	CharacteristicId	NA	NA	(NA, NA)
Value	CharacteristicOrder	NA	NA	(NA, NA)
Value	ByVariableId	NA	NA	(NA, NA)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	NA	NA	(NA, NA)
Value	SurveyYearLabel	-9.75	0.334394	(-0.2884, 0.1008)
Value	DenominatorWeighted	12.55	0.223254	(-0.077, 0.3179)
Value	DenominatorUnweighted	12.46	0.226497	(-0.0778, 0.3172)
Precision	SurveyYear	7.29	0.471018	(-0.1253, 0.2655)
Precision	IndicatorOrder	-42.92	0.000008	(-0.577, -0.2542)
Precision	CharacteristicId	NA	NA	(NA, NA)
Precision	CharacteristicOrder	NA	NA	(NA, NA)
Precision	ByVariableId	NA	NA	(NA, NA)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	NA	NA	(NA, NA)

Precision	SurveyYearLabel	7.29	0.471018	(-0.1253, 0.2655)
Precision	DenominatorWeighted	-1.27	0.9025	(-0.2126, 0.1883)
Precision	DenominatorUnweighted	-1.23	0.90563	(-0.2122, 0.1887)
SurveyYear	IndicatorOrder	21.57	0.0311	(0.0202, 0.3954)
SurveyYear	CharacteristicId	NA	NA	(NA, NA)
SurveyYear	CharacteristicOrder	NA	NA	(NA, NA)
SurveyYear	ByVariableId	NA	NA	(NA, NA)
SurveyYear	IsTotal	NA	NA	(NA, NA)
SurveyYear	IsPreferred	NA	NA	(NA, NA)
SurveyYear	SurveyYearLabel	100	0	(1.0, 1.0)
SurveyYear	DenominatorWeighted	-23.53	0.021009	(-0.4162, -0.0365)
SurveyYear	DenominatorUnweighted	-22.77	0.02566	(-0.4095, -0.0285)
IndicatorOrder	CharacteristicId	NA	NA	(NA, NA)
IndicatorOrder	CharacteristicOrder	NA	NA	(NA, NA)
IndicatorOrder	ByVariableId	NA	NA	(NA, NA)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	NA	NA	(NA, NA)

IndicatorOrder	SurveyYearLabel	21.57	0.0311	(0.0202, 0.3954)
IndicatorOrder	DenominatorWeighted	-4.2	0.684611	(-0.2404, 0.1598)
IndicatorOrder	DenominatorUnweighted	-4.01	0.698401	(-0.2386, 0.1617)
CharacteristicId	CharacteristicOrder	NA	NA	(NA, NA)
CharacteristicId	ByVariableId	NA	NA	(NA, NA)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	NA	NA	(NA, NA)
CharacteristicId	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicId	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicId	DenominatorUnweighted	NA	NA	(NA, NA)
CharacteristicOrder	ByVariableId	NA	NA	(NA, NA)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	NA	NA	(NA, NA)
CharacteristicOrder	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicOrder	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicOrder	DenominatorUnweighted	NA	NA	(NA, NA)
ByVariableId	IsTotal	NA	NA	(NA, NA)
ByVariableId	IsPreferred	NA	NA	(NA, NA)
ByVariableId	SurveyYearLabel	NA	NA	(NA, NA)

ByVariableId	DenominatorWeighted	NA	NA	(NA, NA)
ByVariableId	DenominatorUnweighted	NA	NA	(NA, NA)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	NA	NA	(NA, NA)
IsPreferred	DenominatorWeighted	NA	NA	(NA, NA)
IsPreferred	DenominatorUnweighted	NA	NA	(NA, NA)
SurveyYearLabel	DenominatorWeighted	-23.53	0.021009	(-0.4162, -0.0365)
SurveyYearLabel	DenominatorUnweighted	-22.77	0.02566	(-0.4095, -0.0285)
DenominatorWeighted	DenominatorUnweighted	99.99	0	(0.9999, 1.0)

Include and Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships.

- Value and Precision: The negative correlation of -84.13% with a p-value of 0 indicates a statistically significant and very strong linear relationship. Precision should be a primary predictor.
- Value and IndicatorOrder: The positive correlation of 36.35% with a p-value of 0.000201 is statistically significant ($p < 0.05$). This field should also be included as a predictor.

- DataId, SurveyYear, and SurveyYearLabel: DataId has a near-perfect negative correlation with SurveyYear and SurveyYearLabel (-97.34%, p = 0). To avoid redundancy, only one of these should be used. DataId also has a significant correlation with IndicatorOrder and the Denominator fields, making it a good proxy for the others.
- SurveyYear and IndicatorOrder: The positive correlation of 21.57% (p = 0.0311) is statistically significant.
- SurveyYearLabel and DenominatorWeighted/Unweighted: The negative correlation of -23.53% (p = 0.021) and -22.77% (p = 0.026) are statistically significant, suggesting the survey year is a relevant factor.

Exclusion Criteria

Fields were excluded if they lacked a statistically significant relationship with the target variable Value or were not computable.

- DataId, DenominatorWeighted, and DenominatorUnweighted: While these fields have significant correlations with other variables, their linear correlations with the target variable, Value, are not statistically significant ($p > 0.05$). DenominatorWeighted and DenominatorUnweighted have a perfect correlation of 99.99%, making them redundant.
- CharacteristicId, CharacteristicOrder, ByVariableId, IsTotal, and IsPreferred: These fields show "NA" for all or most correlations, indicating they are likely non-numeric or have data quality issues that prevent correlation analysis. These fields should be excluded.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to prioritize the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The fields with a statistically significant linear correlation to Value are Precision and IndicatorOrder.
2. Calculate Weights: The absolute correlations with Value are:
   - $|\text{Correlation of Value with Precision}| = |-0.8413| = 0.8413$

- |Correlation of Value with IndicatorOrder| = |0.3635| = 0.3635

3. Normalize Weights: Sum the absolute values to get the normalization factor: 0.8413 + 0.3635 = 1.2048.
    - Precision: 0.8413/1.2048 = 0.6983
    - IndicatorOrder: 0.3635/1.2048 = 0.3017

Toilet Facilities

Variable 1	Variable 2	Correlation Coefficient (%)	p-value	95% CI
DataId	Value	14.27	0.344037	(-0.154, 0.4158)
DataId	Precision	-7.78	0.607082	(-0.36, 0.2174)
DataId	SurveyYear	-85.15	0	(-0.9155, -0.7455)
DataId	IndicatorOrder	43.9	0.00227	(0.1705, 0.6469)
DataId	CharacteristicId	NA	NA	(NA, NA)
DataId	CharacteristicOrder	NA	NA	(NA, NA)
DataId	ByVariableId	NA	NA	(NA, NA)
DataId	IsTotal	NA	NA	(NA, NA)
DataId	IsPreferred	NA	NA	(NA, NA)
DataId	SurveyYearLabel	-85.15	0	(-0.9155, -0.7455)

DataId	DenominatorWeighted	51.59	0.000471	(0.2514, 0.7087)
DataId	DenominatorUnweighted	53.44	0.000267	(0.2752, 0.7212)
Value	Precision	-82.76	0	(-0.9014, -0.7073)
Value	SurveyYear	-13.94	0.3556	(-0.413, 0.1573)
Value	IndicatorOrder	29.28	0.048293	(0.0027, 0.5374)
Value	CharacteristicId	NA	NA	(NA, NA)
Value	CharacteristicOrder	NA	NA	(NA, NA)
Value	ByVariableId	NA	NA	(NA, NA)
Value	IsTotal	NA	NA	(NA, NA)
Value	IsPreferred	NA	NA	(NA, NA)
Value	SurveyYearLabel	-13.94	0.3556	(-0.413, 0.1573)
Value	DenominatorWeighted	19.22	0.222662	(-0.1187, 0.4688)
Value	DenominatorUnweighted	19.09	0.225819	(-0.12, 0.4677)
Precision	SurveyYear	10.22	0.499178	(-0.1939, 0.3812)
Precision	IndicatorOrder	-8.79	0.561478	(-0.3688, 0.2077)

Precision	CharacteristicId	NA	NA	(NA, NA)
Precision	CharacteristicOrder	NA	NA	(NA, NA)
Precision	ByVariableId	NA	NA	(NA, NA)
Precision	IsTotal	NA	NA	(NA, NA)
Precision	IsPreferred	NA	NA	(NA, NA)
Precision	SurveyYearLabel	10.22	0.499178	(-0.1939, 0.3812)
Precision	DenominatorWeighted	-1.87	0.906313	(-0.3208, 0.2868)
Precision	DenominatorUnweighted	-1.81	0.909313	(-0.3203, 0.2874)
SurveyYear	IndicatorOrder	-6.14	0.685364	(-0.3455, 0.2331)
SurveyYear	CharacteristicId	NA	NA	(NA, NA)
SurveyYear	CharacteristicOrder	NA	NA	(NA, NA)
SurveyYear	ByVariableId	NA	NA	(NA, NA)
SurveyYear	IsTotal	NA	NA	(NA, NA)
SurveyYear	IsPreferred	NA	NA	(NA, NA)
SurveyYear	SurveyYearLabel	100	0	(1.0, 1.0)
SurveyYear	DenominatorWeighted	-23.54	0.133413	(-0.5033, 0.0738)
SurveyYear	DenominatorUnweighted	-22.78	0.146733	(-0.4973, 0.0818)

IndicatorOrder	CharacteristicId	NA	NA	(NA, NA)
IndicatorOrder	CharacteristicOrder	NA	NA	(NA, NA)
IndicatorOrder	ByVariableId	NA	NA	(NA, NA)
IndicatorOrder	IsTotal	NA	NA	(NA, NA)
IndicatorOrder	IsPreferred	NA	NA	(NA, NA)
IndicatorOrder	SurveyYearLabel	-6.14	0.685364	(-0.3455, 0.2331)
IndicatorOrder	DenominatorWeighted	95.21	0	(0.9122, 0.9742)
IndicatorOrder	DenominatorUnweighted	95.52	0	(0.9176, 0.9758)
CharacteristicId	CharacteristicOrder	NA	NA	(NA, NA)
CharacteristicId	ByVariableId	NA	NA	(NA, NA)
CharacteristicId	IsTotal	NA	NA	(NA, NA)
CharacteristicId	IsPreferred	NA	NA	(NA, NA)
CharacteristicId	SurveyYearLabel	NA	NA	(NA, NA)
CharacteristicId	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicId	DenominatorUnweighted	NA	NA	(NA, NA)
CharacteristicOrder	ByVariableId	NA	NA	(NA, NA)
CharacteristicOrder	IsTotal	NA	NA	(NA, NA)
CharacteristicOrder	IsPreferred	NA	NA	(NA, NA)
CharacteristicOrder	SurveyYearLabel	NA	NA	(NA, NA)

CharacteristicOrder	DenominatorWeighted	NA	NA	(NA, NA)
CharacteristicOrder	DenominatorUnweighted	NA	NA	(NA, NA)
ByVariableId	IsTotal	NA	NA	(NA, NA)
ByVariableId	IsPreferred	NA	NA	(NA, NA)
ByVariableId	SurveyYearLabel	NA	NA	(NA, NA)
ByVariableId	DenominatorWeighted	NA	NA	(NA, NA)
ByVariableId	DenominatorUnweighted	NA	NA	(NA, NA)
IsTotal	IsPreferred	NA	NA	(NA, NA)
IsTotal	SurveyYearLabel	NA	NA	(NA, NA)
IsTotal	DenominatorWeighted	NA	NA	(NA, NA)
IsTotal	DenominatorUnweighted	NA	NA	(NA, NA)
IsPreferred	SurveyYearLabel	NA	NA	(NA, NA)
IsPreferred	DenominatorWeighted	NA	NA	(NA, NA)
IsPreferred	DenominatorUnweighted	NA	NA	(NA, NA)
SurveyYearLabel	DenominatorWeighted	-23.54	0.133413	(-0.5033, 0.0738)
SurveyYearLabel	DenominatorUnweighted	-22.78	0.146733	(-0.4973, 0.0818)
DenominatorWeighted	DenominatorUnweighted	99.99	0	(0.9999, 1.0)

Include and Exclude

Inclusion Criteria

Fields were selected for their statistically significant relationships.

- Value and Precision: The negative correlation of -82.76% with a p-value of 0 indicates a statistically significant and very strong linear relationship. Precision should be a primary predictor.
- Value and IndicatorOrder: The positive correlation of 29.28% with a p-value of 0.048293 is statistically significant (p < 0.05). This field should also be included as a predictor.
- DataId, SurveyYear, and SurveyYearLabel: DataId has a near-perfect negative correlation with SurveyYear and SurveyYearLabel (-85.15%, p = 0). To avoid redundancy, only one of these should be used. DataId also has significant correlations with IndicatorOrder and the Denominator fields, making it a good proxy for the others.
- DataId and IndicatorOrder: The positive correlation of 43.9% (p = 0.00227) is statistically significant.
- DataId and DenominatorWeighted/Unweighted: The positive correlation of 51.59% (p = 0.000471) and 53.44% (p = 0.000267) are statistically significant, suggesting that DataId is a strong proxy for these fields.

Exclusion Criteria

Fields were excluded if they lacked a statistically significant relationship with the target variable Value or were not computable.

- DataId, SurveyYear, SurveyYearLabel, and DenominatorWeighted/Unweighted: While these fields have significant correlations with other variables, their linear correlations with the target variable, Value, are not statistically significant (p > 0.05). DenominatorWeighted and DenominatorUnweighted have a near-perfect correlation of 99.99%, making them redundant.
- CharacteristicId, CharacteristicOrder, ByVariableId, IsTotal, and IsPreferred: These fields show "NA" for all or most correlations, indicating they are likely non-numeric or have data quality issues that prevent correlation analysis. These fields should be excluded.

Attribute Weighting Strategy

An attribute weighting strategy is necessary to prioritize the selected features for model training. The weights will be based on the absolute value of the correlation coefficient with the target variable, Value.

1. Identify Correlated Predictors: The fields with a statistically significant linear correlation to Value are Precision and IndicatorOrder.
2. Calculate Weights: The absolute correlations with Value are:
   - |Correlation of Value with Precision| = |−0.8276| = 0.8276
   - |Correlation of Value with IndicatorOrder| = |0.2928| = 0.2928
3. Normalize Weights: Sum the absolute values to get the normalization factor: 0.8276 + 0.2928 = 1.1204.
   - Precision: 0.8276/1.1204 = 0.7386
   - IndicatorOrder: 0.2928/1.1204 = 0.2614

Conclusion of Data Quality Verification

The data quality verification phase of this project successfully ensured that the correct fields were selected for analysis by using significance and correlation tests. This methodology involved a thorough assessment of each dataset, considering data quality, modeling requirements and the rationale for including or excluding specific variables.

Based on this process, the following fields were identified for inclusion in their respective datasets:

- Access to Health Care:
  - DataId: Shows significant correlations with several other fields, suggesting it's a meaningful identifier.
  - Value: This field is considered a primary measure and is fundamentally important.
  - Precision: Has significant correlations with DataId and Value.
  - IndicatorOrder: Has a significant correlation with DataId and other fields.
  - SurveyYear: Chosen over SurveyYearLabel due to their perfect correlation, with one being dropped to avoid redundancy.

    - DenominatorWeighted: Chosen over DenominatorUnweighted to avoid redundancy, as they have a near-perfect correlation.
    - IsPreferred: Shows significant correlations with DataId and the Denominator fields.
    - ByVariableId: Has significant correlations with DataId, SurveyYear and IndicatorOrder.
- Child Mortality Rates:
    - Value: As the target variable, it is a key field for the analysis.
    - IndicatorOrder: Has a statistically significant correlation with Value.
    - CharacteristicOrder, CharacteristicId and IsPreferred: These fields are highly correlated with each other and are considered to contain valuable, interconnected information that should be included.
- Maternal Mortality:
    - Value: As the target variable, it is crucial to the analysis.
    - IndicatorOrder, CharacteristicId and CharacteristicOrder: These fields are all highly interconnected with statistically significant correlations to the target variable 'Value'. However, because they are perfectly correlated, only one of them would be chosen to represent this group in the final model to avoid redundancy, the chosen field will be IndicatorOrder'
- DHS Quickstat:
    - Value: As the target variable, it is a key field for analysis.
    - Precision: Shows a statistically significant negative correlation with Value (-0.3888) and should be included as a predictor.
- Immunization:
    - Value: As the target variable, it is a key field for analysis.
    - Precision: The negative correlation of -74.76% with a p-value of 0 indicates a statistically significant and strong linear relationship. Precision should be included as a predictor.
    - IndicatorOrder: The positive correlation of 19.51% with a p-value of 0.0358 is statistically significant and should also be included as a predictor.
- IYCF:

    - DataId: Has a correlation with Value of 38.12%, but the p-value is 0.08, which is above the standard significance threshold. It should be considered for inclusion, but with caution.
    - SurveyYear: Chosen over SurveyYearLabel due to their perfect positive correlation.
    - DenominatorWeighted: Chosen over DenominatorUnweighted to avoid redundancy, as they have a very high correlation.
- Literacy:
    - Value: As the target variable, it is a key field for analysis.
    - Precision: The negative correlation of -85.03% with a p-value of 0.000002 indicates a statistically significant and very strong linear relationship. Precision should be a primary predictor.
    - DataId: The negative correlation of -47.74% with a p-value of 0.033269 is statistically significant and should also be included as a predictor.
- Water:
    - Value: As the target variable, it is a key field for analysis.
    - Precision: The negative correlation of -84.13% with a p-value of 0 indicates a statistically significant and very strong linear relationship. Precision should be a primary predictor.
    - IndicatorOrder: The positive correlation of 36.35% with a p-value of 0.000201 is statistically significant and should also be included as a predictor.
    - DataId: Has a near-perfect negative correlation with SurveyYear and SurveyYearLabel (-97.34%), and is a good proxy to use for them.
- Toilet Facilities:
    - Value: As the target variable, it is a key field for analysis.
    - Precision: The negative correlation of -82.76% with a p-value of 0 indicates a statistically significant and very strong linear relationship. Precision should be a primary predictor.
    - IndicatorOrder: The positive correlation of 29.28% with a p-value of 0.048293 is statistically significant and should also be included as a predictor.
    - DataId: Has a significant correlation with IndicatorOrder (43.9%, p = 0.00227) and the Denominator fields (DenominatorWeighted (51.59%, p = 0.000471) and

DenominatorUnweighted (53.44%, p = 0.000267)), making it a strong proxy for these fields.