Data description and selection

Group K

2025-09-13

datasets <- list(  
 "Access\_To\_Healthcare\_01" = here("raw\_data", "access-to-health-care\_national\_zaf.csv"),  
 "Child\_Mortality\_02" = here("raw\_data", "child-mortality-rates\_national\_zaf.csv"),  
 "HIV\_Behaviour\_03" = here("raw\_data", "hiv-behavior\_national\_zaf.csv"),  
 "Immunization\_04" = here("raw\_data", "immunization\_national\_zaf.csv"),  
 "Toilet\_05" = here("raw\_data", "toilet-facilities\_national\_zaf.csv"),  
 "Water\_06" = here("raw\_data", "water\_national\_zaf.csv"),  
 "Covid\_07" = here("raw\_data", "covid-19-prevention\_national\_zaf.csv"),  
 "Maternal\_mortality\_08" = here("raw\_data", "maternal-mortality\_national\_zaf.csv"),  
 "Literacy\_09" = here("raw\_data", "literacy\_national\_zaf.csv"),  
 "ARI\_10" = here("raw\_data", "symptoms-of-acute-respiratory-infection-ari\_national\_zaf.csv"),  
 "Anthro\_11" = here("raw\_data", "anthropometry\_national\_zaf.csv"),  
 "IYCF\_12" = here("raw\_data", "iycf\_national\_zaf.csv")  
)

summarize\_dataset <- function(file\_path, dataset\_name){  
 df <- read\_csv(file\_path, show\_col\_types = FALSE)  
 n\_rows <- nrow(df)  
 n\_cols <- ncol(df)  
   
 n\_numeric <- sum(sapply(df, is.numeric))  
 n\_character <- sum(sapply(df, is.character))  
 n\_factor <- sum(sapply(df, is.factor))  
 n\_logical <- sum(sapply(df, is.logical))  
 n\_date <- sum(sapply(df, lubridate::is.Date)) # use lubridate to detect dates  
   
   
 data.frame(  
 Dataset = dataset\_name,  
 Rows = n\_rows,  
 Columns = n\_cols,  
 Numeric\_Columns = n\_numeric,  
 Character\_Columns = n\_character,  
 Factor\_Columns = n\_factor,  
 Logical\_Columns = n\_logical,  
 Date\_Columns = n\_date  
 )  
}  
  
# Apply function to all datasets  
summary\_table <- do.call(rbind, lapply(names(datasets), function(name) {  
 summarize\_dataset(datasets[[name]], name)  
}))  
  
# Display table in Word  
knitr::kable(summary\_table, caption = "Overview of all datasets with detailed column types")

Overview of all datasets with detailed column types

| Dataset | Rows | Columns | Numeric\_Columns | Character\_Columns | Factor\_Columns | Logical\_Columns | Date\_Columns |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Access\_To\_Healthcare\_01 | 276 | 29 | 8 | 17 | 0 | 4 | 0 |
| Child\_Mortality\_02 | 41 | 29 | 10 | 17 | 0 | 2 | 0 |
| HIV\_Behaviour\_03 | 119 | 29 | 8 | 17 | 0 | 4 | 0 |
| Immunization\_04 | 117 | 29 | 8 | 17 | 0 | 4 | 0 |
| Toilet\_05 | 47 | 29 | 8 | 17 | 0 | 4 | 0 |
| Water\_06 | 101 | 29 | 8 | 17 | 0 | 4 | 0 |
| Covid\_07 | 35 | 29 | 8 | 17 | 0 | 4 | 0 |
| Maternal\_mortality\_08 | 22 | 29 | 10 | 17 | 0 | 2 | 0 |
| Literacy\_09 | 21 | 29 | 8 | 17 | 0 | 4 | 0 |
| ARI\_10 | 27 | 29 | 8 | 17 | 0 | 4 | 0 |
| Anthro\_11 | 38 | 29 | 8 | 17 | 0 | 4 | 0 |
| IYCF\_12 | 23 | 29 | 8 | 17 | 0 | 4 | 0 |

# Function to calculate data quality measures  
data\_quality\_summary <- function(file\_path, dataset\_name) {  
 df <- read\_csv(file\_path, show\_col\_types = FALSE)  
   
 n\_duplicates <- sum(duplicated(df))  
 missing\_pct <- round(sum(is.na(df)) / (nrow(df) \* ncol(df)) \* 100, 2)  
   
 # Simple checks for outliers/invalids  
 numeric\_cols <- df %>% select(where(is.numeric))  
 outliers <- sum(sapply(numeric\_cols, function(x) sum(x < 0 | x > 100, na.rm = TRUE))) # example for percentages  
   
 # Inconsistent coding (non-numeric, non-date factors)  
 char\_cols <- df %>% select(where(is.character))  
 inconsistent <- sum(sapply(char\_cols, function(x) sum(grepl("unknown|Unknown|N/A|NA|n/a", x))))  
   
 # Unexpected zeros  
 zeros <- sum(sapply(numeric\_cols, function(x) sum(x == 0, na.rm = TRUE)))  
   
 data.frame(  
 Dataset = dataset\_name,  
 Duplicate\_Rows = n\_duplicates,  
 Missing\_Values\_Percent = missing\_pct,  
 Outliers\_Invalid = outliers,  
 Inconsistent\_Coding = inconsistent,  
 Unexpected\_Zeros = zeros  
 )  
}  
  
# Apply function to all datasets  
quality\_table <- bind\_rows(  
 lapply(seq\_along(datasets), function(i) {  
 data\_quality\_summary(datasets[[i]], names(datasets)[i])  
 })  
)  
knitr::kable(quality\_table, caption = "Data Quality Summary for All Datasets")

Data Quality Summary for All Datasets

| Dataset | Duplicate\_Rows | Missing\_Values\_Percent | Outliers\_Invalid | Inconsistent\_Coding | Unexpected\_Zeros |
| --- | --- | --- | --- | --- | --- |
| Access\_To\_Healthcare\_01 | 0 | 14.98 | 1298 | 0 | 434 |
| Child\_Mortality\_02 | 0 | 17.66 | 148 | 0 | 30 |
| HIV\_Behaviour\_03 | 0 | 19.85 | 555 | 0 | 67 |
| Immunization\_04 | 0 | 16.33 | 624 | 0 | 86 |
| Toilet\_05 | 0 | 18.78 | 222 | 0 | 46 |
| Water\_06 | 0 | 17.96 | 492 | 4 | 100 |
| Covid\_07 | 0 | 18.92 | 166 | 2 | 34 |
| Maternal\_mortality\_08 | 0 | 23.67 | 86 | 0 | 11 |
| Literacy\_09 | 0 | 20.03 | 116 | 0 | 0 |
| ARI\_10 | 0 | 17.62 | 105 | 0 | 39 |
| Anthro\_11 | 0 | 19.15 | 190 | 0 | 24 |
| IYCF\_12 | 0 | 19.79 | 116 | 2 | 12 |