05\_DHS\_Quickstats

# DHS QuickStats (National, South Africa)

## Load Libraries

# Data manipulation  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(readr)  
library(here)

## here() starts at C:/Users/morul/School/3rd Year/BIN381/BIN381\_PROJECT/BIN381\_PROJECT

library(purrr)  
  
# Visualization and summaries  
library(ggplot2)  
library(skimr)  
library(visdat)

## Load the DHS QuickStats dataset

dhs\_df <- read\_csv(here("data", "raw", "dhs-quickstats\_national\_zaf.csv"))

## Rows: 53 Columns: 29  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (17): ISO3, DataId, Indicator, Value, Precision, DHS\_CountryCode, Countr...  
## dbl (10): IndicatorOrder, CharacteristicId, CharacteristicOrder, IsTotal, Is...  
## lgl (2): RegionId, LevelRank  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Remove first row if it contains metadata  
dhs\_df <- dhs\_df[-1, ]  
  
# Reset row names  
rownames(dhs\_df) <- NULL  
  
cat("DHS QuickStats dataset loaded successfully.\n")

## DHS QuickStats dataset loaded successfully.

cat("Dimensions:", dim(dhs\_df), "\n")

## Dimensions: 52 29

## Initial Data Assessment

# Quick glimpse  
glimpse(dhs\_df)

## Rows: 52  
## Columns: 29  
## $ ISO3 <chr> "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF"…  
## $ DataId <chr> "796527", "795692", "795693", "795515", "795357…  
## $ Indicator <chr> "Total fertility rate 15-49", "Married women cu…  
## $ Value <chr> "2.9", "56.3", "55.1", "16.5", "75.7", "24.2", …  
## $ Precision <chr> "1", "1", "1", "1", "1", "1", "1", "0", "0", "0…  
## $ DHS\_CountryCode <chr> "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "ZA",…  
## $ CountryName <chr> "South Africa", "South Africa", "South Africa",…  
## $ SurveyYear <chr> "1998", "1998", "1998", "1998", "1998", "1998",…  
## $ SurveyId <chr> "ZA1998DHS", "ZA1998DHS", "ZA1998DHS", "ZA1998D…  
## $ IndicatorId <chr> "FE\_FRTR\_W\_TFR", "FP\_CUSM\_W\_ANY", "FP\_CUSM\_W\_MO…  
## $ IndicatorOrder <dbl> 11763080, 32633010, 32633020, 32933030, 3293315…  
## $ IndicatorType <chr> "I", "I", "I", "I", "I", "I", "I", "I", "I", "I…  
## $ CharacteristicId <dbl> 1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000,…  
## $ CharacteristicOrder <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…  
## $ CharacteristicCategory <chr> "Total", "Total", "Total", "Total", "Total", "T…  
## $ CharacteristicLabel <chr> "Total", "Total", "Total", "Total", "Total", "T…  
## $ ByVariableId <chr> "0", "0", "0", "0", "0", "0", "0", "14001", "14…  
## $ ByVariableLabel <chr> NA, NA, NA, NA, NA, NA, NA, "Five years precedi…  
## $ IsTotal <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…  
## $ IsPreferred <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1,…  
## $ SDRID <chr> "FEFRTRWTFR", "FPCUSMWANY", "FPCUSMWMOD", "FPNA…  
## $ RegionId <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ SurveyYearLabel <dbl> 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998,…  
## $ SurveyType <chr> "DHS", "DHS", "DHS", "DHS", "DHS", "DHS", "DHS"…  
## $ DenominatorWeighted <dbl> NA, 5077, 5077, 5077, 3695, NA, NA, NA, NA, NA,…  
## $ DenominatorUnweighted <dbl> NA, 4948, 4948, 4948, 3590, NA, NA, NA, NA, NA,…  
## $ CILow <dbl> NA, NA, NA, NA, NA, NA, NA, 38, 37, 50, 50, 77,…  
## $ CIHigh <dbl> NA, NA, NA, NA, NA, NA, NA, 53, 48, 68, 63, 223…  
## $ LevelRank <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…

# Summary of missingness  
skim(dhs\_df)

Data summary

|  |  |
| --- | --- |
| Name | dhs\_df |
| Number of rows | 52 |
| Number of columns | 29 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 17 |
| logical | 2 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ISO3 | 0 | 1.00 | 3 | 3 | 0 | 1 | 0 |
| DataId | 0 | 1.00 | 3 | 6 | 0 | 52 | 0 |
| Indicator | 0 | 1.00 | 15 | 76 | 0 | 27 | 0 |
| Value | 0 | 1.00 | 2 | 4 | 0 | 51 | 0 |
| Precision | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| DHS\_CountryCode | 0 | 1.00 | 2 | 2 | 0 | 1 | 0 |
| CountryName | 0 | 1.00 | 12 | 12 | 0 | 1 | 0 |
| SurveyYear | 0 | 1.00 | 4 | 4 | 0 | 2 | 0 |
| SurveyId | 0 | 1.00 | 9 | 9 | 0 | 2 | 0 |
| IndicatorId | 0 | 1.00 | 13 | 13 | 0 | 27 | 0 |
| IndicatorType | 0 | 1.00 | 1 | 1 | 0 | 1 | 0 |
| CharacteristicCategory | 0 | 1.00 | 5 | 11 | 0 | 2 | 0 |
| CharacteristicLabel | 0 | 1.00 | 5 | 11 | 0 | 2 | 0 |
| ByVariableId | 0 | 1.00 | 1 | 6 | 0 | 6 | 0 |
| ByVariableLabel | 33 | 0.37 | 12 | 32 | 0 | 5 | 0 |
| SDRID | 0 | 1.00 | 10 | 10 | 0 | 27 | 0 |
| SurveyType | 0 | 1.00 | 3 | 3 | 0 | 1 | 0 |

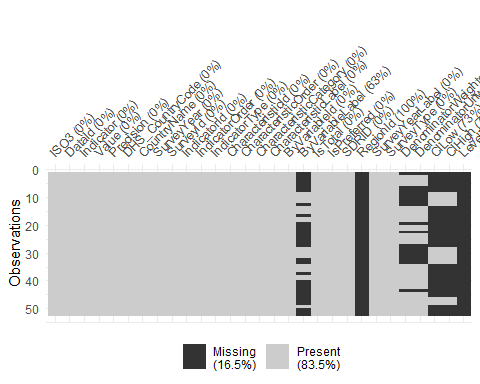
**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate | mean | count |
| --- | --- | --- | --- | --- |
| RegionId | 52 | 0 | NaN | : |
| LevelRank | 52 | 0 | NaN | : |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IndicatorOrder | 0 | 1.00 | 93178551.54 | 61143758.08 | 11763080.0 | 60330295.00 | 83566070.0 | 104261072.50 | 260321010 | ▃▇▂▁▂ |
| CharacteristicId | 0 | 1.00 | 2557.69 | 3438.04 | 1000.0 | 1000.00 | 1000.0 | 1000.00 | 10000 | ▇▁▁▁▂ |
| CharacteristicOrder | 0 | 1.00 | 1730.77 | 3820.05 | 0.0 | 0.00 | 0.0 | 0.00 | 10000 | ▇▁▁▁▂ |
| IsTotal | 0 | 1.00 | 1.00 | 0.00 | 1.0 | 1.00 | 1.0 | 1.00 | 1 | ▁▁▇▁▁ |
| IsPreferred | 0 | 1.00 | 0.79 | 0.41 | 0.0 | 1.00 | 1.0 | 1.00 | 1 | ▂▁▁▁▇ |
| SurveyYearLabel | 0 | 1.00 | 2008.73 | 8.92 | 1998.0 | 1998.00 | 2016.0 | 2016.00 | 2016 | ▆▁▁▁▇ |
| DenominatorWeighted | 18 | 0.65 | 3832.21 | 3328.81 | 246.0 | 1414.50 | 3050.0 | 5055.75 | 12247 | ▇▅▂▂▂ |
| DenominatorUnweighted | 18 | 0.65 | 3999.03 | 3682.57 | 256.0 | 1470.75 | 2841.0 | 4948.00 | 12247 | ▇▆▁▁▂ |
| CILow | 38 | 0.27 | 67.67 | 79.98 | 11.5 | 27.75 | 37.5 | 50.00 | 270 | ▇▁▁▁▂ |
| CIHigh | 38 | 0.27 | 160.54 | 258.61 | 17.3 | 45.00 | 52.0 | 66.75 | 802 | ▇▁▁▁▂ |

# Visualize missing values  
vis\_miss(dhs\_df)



# Standardize column names  
dhs\_df <- dhs\_df %>% janitor::clean\_names()  
colnames(dhs\_df)

## [1] "iso3" "data\_id"   
## [3] "indicator" "value"   
## [5] "precision" "dhs\_country\_code"   
## [7] "country\_name" "survey\_year"   
## [9] "survey\_id" "indicator\_id"   
## [11] "indicator\_order" "indicator\_type"   
## [13] "characteristic\_id" "characteristic\_order"   
## [15] "characteristic\_category" "characteristic\_label"   
## [17] "by\_variable\_id" "by\_variable\_label"   
## [19] "is\_total" "is\_preferred"   
## [21] "sdrid" "region\_id"   
## [23] "survey\_year\_label" "survey\_type"   
## [25] "denominator\_weighted" "denominator\_unweighted"   
## [27] "ci\_low" "ci\_high"   
## [29] "level\_rank"

### Glimpse of the Dataset

A quick look at the first few rows and column types revealed:

* 17 character columns (e.g., ISO3, Indicator, CountryName)
* 10 numeric columns (e.g., Value, Precision)
* 2 logical columns (RegionId, LevelRank)

### Summary of Missing Values

The skimr package summarized missingness:

* Columns such as ByVariableLabel, DenominatorWeighted, CILow, and CIHigh contained missing values.
* These columns would require imputation or handling in later steps.

### Visualization

vis\_miss() was used to create a visual map of missing data, which helped identify columns with high missingness at a glance.

Why this step matters: Understanding missing data is crucial for selecting appropriate imputation methods or deciding if columns should be dropped.

## Rename Columns Meaningfully

# Replace generic col\_1, col\_2, ... with actual names  
colnames(dhs\_df) <- c(  
 "iso3", "data\_id", "indicator", "value", "precision",  
 "dhs\_country\_code", "country\_name", "survey\_year", "survey\_id",  
 "indicator\_id", "indicator\_order", "indicator\_type", "characteristic\_id",  
 "characteristic\_order", "characteristic\_category", "characteristic\_label",  
 "by\_variable\_id", "by\_variable\_label", "is\_total", "is\_preferred",  
 "sdrid", "region\_id", "survey\_year\_label", "survey\_type",  
 "denominator\_weighted", "denominator\_unweighted", "ci\_low", "ci\_high",  
 "level\_rank"  
)  
  
cat("Columns renamed to meaningful names.\n")

## Columns renamed to meaningful names.

colnames(dhs\_df)

## [1] "iso3" "data\_id"   
## [3] "indicator" "value"   
## [5] "precision" "dhs\_country\_code"   
## [7] "country\_name" "survey\_year"   
## [9] "survey\_id" "indicator\_id"   
## [11] "indicator\_order" "indicator\_type"   
## [13] "characteristic\_id" "characteristic\_order"   
## [15] "characteristic\_category" "characteristic\_label"   
## [17] "by\_variable\_id" "by\_variable\_label"   
## [19] "is\_total" "is\_preferred"   
## [21] "sdrid" "region\_id"   
## [23] "survey\_year\_label" "survey\_type"   
## [25] "denominator\_weighted" "denominator\_unweighted"   
## [27] "ci\_low" "ci\_high"   
## [29] "level\_rank"

Column names were standardized to snake\_case using janitor::clean\_names().

Additionally, descriptive names were assigned to generic column names (e.g., col\_1, col\_2) to improve readability.

Example:

* ISO3 → iso3
* DataId → data\_id
* Value → value
* Precision → precision
* Benefit: This ensures consistency across analysis scripts and improves interpretability for readers.

## Remove Duplicates

# Check for exact duplicates  
exact\_dups <- sum(duplicated(dhs\_df))  
cat("Exact duplicate rows:", exact\_dups, "\n")

## Exact duplicate rows: 0

# Remove duplicates, keeping first occurrence  
dhs\_df <- dhs\_df %>%  
 distinct(indicator, survey\_year, characteristic\_id, value, .keep\_all = TRUE)  
  
cat("Dimensions after duplicate removal:", dim(dhs\_df), "\n")

## Dimensions after duplicate removal: 52 29

## Remove Redundant & Empty Columns

all\_na\_cols <- dhs\_df %>%  
 summarise(across(everything(), ~all(is.na(.)))) %>%  
 pivot\_longer(everything(), names\_to = "column", values\_to = "all\_na") %>%  
 filter(all\_na) %>%  
 pull(column)  
  
if(length(all\_na\_cols) > 0) {  
 dhs\_df <- dhs\_df %>% select(-all\_of(all\_na\_cols))  
 cat("Removed 100% missing columns:\n")  
 print(all\_na\_cols)  
} else {  
 cat("No columns were 100% missing.\n")  
}

## Removed 100% missing columns:  
## [1] "region\_id" "level\_rank"

### Removing Duplicates and Empty Columns

Exact duplicate rows were checked; none were found.

The dataset was then filtered to retain unique combinations of Indicator, Survey Year, Characteristic ID, and Value.

Fully empty columns (region\_id and level\_rank) were removed.

## Convert Data Types Safely

# Numeric columns  
numeric\_cols <- intersect(c("value", "precision", "ci\_low", "ci\_high"), colnames(dhs\_df))  
  
# Integer columns  
integer\_cols <- intersect(c("survey\_year", "indicator\_order", "characteristic\_id",  
 "characteristic\_order", "survey\_year\_label", "by\_variable\_id"), colnames(dhs\_df))  
  
# Logical columns  
logical\_cols <- intersect(c("is\_total", "is\_preferred"), colnames(dhs\_df))  
  
# Apply type conversion  
dhs\_df <- dhs\_df %>%  
 mutate(  
 across(all\_of(numeric\_cols), as.numeric),  
 across(all\_of(integer\_cols), as.integer),  
 across(all\_of(logical\_cols), ~as.logical(as.integer(.)))  
 )  
  
str(dhs\_df)

## tibble [52 × 27] (S3: tbl\_df/tbl/data.frame)  
## $ iso3 : chr [1:52] "ZAF" "ZAF" "ZAF" "ZAF" ...  
## $ data\_id : chr [1:52] "796527" "795692" "795693" "795515" ...  
## $ indicator : chr [1:52] "Total fertility rate 15-49" "Married women currently using any method of contraception" "Married women currently using any modern method of contraception" "Unmet need for family planning" ...  
## $ value : num [1:52] 2.9 56.3 55.1 16.5 75.7 24.2 18.4 45 42 59 ...  
## $ precision : num [1:52] 1 1 1 1 1 1 1 0 0 0 ...  
## $ dhs\_country\_code : chr [1:52] "ZA" "ZA" "ZA" "ZA" ...  
## $ country\_name : chr [1:52] "South Africa" "South Africa" "South Africa" "South Africa" ...  
## $ survey\_year : int [1:52] 1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...  
## $ survey\_id : chr [1:52] "ZA1998DHS" "ZA1998DHS" "ZA1998DHS" "ZA1998DHS" ...  
## $ indicator\_id : chr [1:52] "FE\_FRTR\_W\_TFR" "FP\_CUSM\_W\_ANY" "FP\_CUSM\_W\_MOD" "FP\_NADM\_W\_UNT" ...  
## $ indicator\_order : int [1:52] 11763080 32633010 32633020 32933030 32933150 41633090 51703090 63206030 63206030 63206050 ...  
## $ indicator\_type : chr [1:52] "I" "I" "I" "I" ...  
## $ characteristic\_id : int [1:52] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 ...  
## $ characteristic\_order : int [1:52] 0 0 0 0 0 0 0 0 0 0 ...  
## $ characteristic\_category: chr [1:52] "Total" "Total" "Total" "Total" ...  
## $ characteristic\_label : chr [1:52] "Total" "Total" "Total" "Total" ...  
## $ by\_variable\_id : int [1:52] 0 0 0 0 0 0 0 14001 14003 14001 ...  
## $ by\_variable\_label : chr [1:52] NA NA NA NA ...  
## $ is\_total : logi [1:52] TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ is\_preferred : logi [1:52] TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ sdrid : chr [1:52] "FEFRTRWTFR" "FPCUSMWANY" "FPCUSMWMOD" "FPNADMWUNT" ...  
## $ survey\_year\_label : int [1:52] 1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...  
## $ survey\_type : chr [1:52] "DHS" "DHS" "DHS" "DHS" ...  
## $ denominator\_weighted : num [1:52] NA 5077 5077 5077 3695 ...  
## $ denominator\_unweighted : num [1:52] NA 4948 4948 4948 3590 ...  
## $ ci\_low : num [1:52] NA NA NA NA NA NA NA 38 37 50 ...  
## $ ci\_high : num [1:52] NA NA NA NA NA NA NA 53 48 68 ...

Columns were converted to appropriate types:

* Numeric columns: value, precision, ci\_low, ci\_high
* Integer columns: survey\_year, indicator\_order, characteristic\_id, etc.
* Logical columns: is\_total, is\_preferred
* Purpose: Correct data types ensure proper calculations, comparisons, and visualizations

## Handle Missing Values

# Define mode function for categorical imputation  
impute\_mode <- function(x) {  
 ux <- na.omit(x)  
 if(length(ux) == 0) return(x)  
 rep(names(sort(table(ux), decreasing = TRUE))[1], length(x))  
}  
  
# Impute missing values  
dhs\_df <- dhs\_df %>%  
 mutate(  
 # Numeric → median  
 across(where(is.numeric), ~ifelse(is.na(.), median(., na.rm = TRUE), .)),  
   
 # Character → mode  
 across(where(is.character), ~ifelse(is.na(.), impute\_mode(.), .)),  
   
 # Logical → FALSE  
 across(where(is.logical), ~ifelse(is.na(.), FALSE, .))  
 )  
  
# Ensure survey\_year\_label filled  
dhs\_df <- dhs\_df %>%  
 mutate(survey\_year\_label = ifelse(is.na(survey\_year\_label), survey\_year, survey\_year\_label))  
  
# Recalculate missing values  
missing\_summary <- data.frame(  
 Column = colnames(dhs\_df),  
 n\_missing = colSums(is.na(dhs\_df)),  
 total\_rows = nrow(dhs\_df),  
 missing\_percent = round(colSums(is.na(dhs\_df))/nrow(dhs\_df)\*100, 2)  
)  
  
missing\_summary %>% arrange(desc(missing\_percent))

## Column n\_missing total\_rows  
## iso3 iso3 0 52  
## data\_id data\_id 0 52  
## indicator indicator 0 52  
## value value 0 52  
## precision precision 0 52  
## dhs\_country\_code dhs\_country\_code 0 52  
## country\_name country\_name 0 52  
## survey\_year survey\_year 0 52  
## survey\_id survey\_id 0 52  
## indicator\_id indicator\_id 0 52  
## indicator\_order indicator\_order 0 52  
## indicator\_type indicator\_type 0 52  
## characteristic\_id characteristic\_id 0 52  
## characteristic\_order characteristic\_order 0 52  
## characteristic\_category characteristic\_category 0 52  
## characteristic\_label characteristic\_label 0 52  
## by\_variable\_id by\_variable\_id 0 52  
## by\_variable\_label by\_variable\_label 0 52  
## is\_total is\_total 0 52  
## is\_preferred is\_preferred 0 52  
## sdrid sdrid 0 52  
## survey\_year\_label survey\_year\_label 0 52  
## survey\_type survey\_type 0 52  
## denominator\_weighted denominator\_weighted 0 52  
## denominator\_unweighted denominator\_unweighted 0 52  
## ci\_low ci\_low 0 52  
## ci\_high ci\_high 0 52  
## missing\_percent  
## iso3 0  
## data\_id 0  
## indicator 0  
## value 0  
## precision 0  
## dhs\_country\_code 0  
## country\_name 0  
## survey\_year 0  
## survey\_id 0  
## indicator\_id 0  
## indicator\_order 0  
## indicator\_type 0  
## characteristic\_id 0  
## characteristic\_order 0  
## characteristic\_category 0  
## characteristic\_label 0  
## by\_variable\_id 0  
## by\_variable\_label 0  
## is\_total 0  
## is\_preferred 0  
## sdrid 0  
## survey\_year\_label 0  
## survey\_type 0  
## denominator\_weighted 0  
## denominator\_unweighted 0  
## ci\_low 0  
## ci\_high 0

### Handling Missing Values

Strategy:

1. Numeric columns: Imputed using the median
2. Character columns: Imputed using the mode (most frequent value)
3. Logical columns: Missing values set to FALSE

Special handling: survey\_year\_label was filled with survey\_year where missing.

## Outlier Detection

# Identify potential outliers using IQR  
numeric\_cols <- intersect(c("value", "precision"), colnames(dhs\_df))  
  
for(col in numeric\_cols) {  
 Q1 <- quantile(dhs\_df[[col]], 0.25, na.rm = TRUE)  
 Q3 <- quantile(dhs\_df[[col]], 0.75, na.rm = TRUE)  
 IQR\_val <- Q3 - Q1  
 lower <- Q1 - 1.5\*IQR\_val  
 upper <- Q3 + 1.5\*IQR\_val  
 dhs\_df[[paste0(col, "\_outlier\_flag")]] <- dhs\_df[[col]] < lower | dhs\_df[[col]] > upper  
}  
  
# Winsorize 'value' at 1st and 99th percentile  
lower\_cap <- quantile(dhs\_df$value, 0.01, na.rm = TRUE)  
upper\_cap <- quantile(dhs\_df$value, 0.99, na.rm = TRUE)  
  
dhs\_df <- dhs\_df %>%  
 mutate(value = pmax(pmin(value, upper\_cap), lower\_cap))  
  
summary(dhs\_df$value)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.604 23.450 55.700 69.002 81.525 504.890

* Method: Interquartile Range (IQR) to identify extreme values
* Treatment: Values outside 1.5×IQR were flagged, then Winsorized at the 1st and 99th percentiles.

## Save Cleaned Data

write\_csv(dhs\_df, here("data", "processed", "dhs\_quickstats\_cleaned.csv"))  
cat("Cleaned DHS QuickStats dataset saved.\n")

## Cleaned DHS QuickStats dataset saved.