03\_child\_mortality

# Data Cleaning: Child Mortality Rates

## 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)  
library(stringr)  
library(knitr)  
  
# Extras for cleaning and exploration  
library(janitor) # clean column names

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(visdat) # visualize missingness  
library(skimr) # summary stats  
library(ggplot2) # visualizations

## Load Dataset

# Load the child mortality dataset  
cmr\_df <- read\_csv(  
 here("data", "raw", "child-mortality-rates\_national\_zaf.csv"),  
 col\_types = cols() # suppress column guessing warnings  
)  
  
# Remove first metadata row if present  
cmr\_df <- cmr\_df[-1, ]  
rownames(cmr\_df) <- NULL  
  
cat("Dataset loaded successfully.\n")

## Dataset loaded successfully.

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

## Dimensions: 40 29

## Initial Data Assessment

# Clean column names  
cmr\_df <- janitor::clean\_names(cmr\_df)  
  
# Peek at structure  
glimpse(cmr\_df)

## Rows: 40  
## Columns: 29  
## $ iso3 <chr> "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF…  
## $ data\_id <chr> "85995", "794581", "785930", "56239", "101014"…  
## $ indicator <chr> "Neonatal mortality rate (5 year periods)", "P…  
## $ value <chr> "20", "26", "45", "15", "59", "20", "19", "26"…  
## $ precision <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "…  
## $ dhs\_country\_code <chr> "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "ZA"…  
## $ country\_name <chr> "South Africa", "South Africa", "South Africa"…  
## $ survey\_year <chr> "1998", "1998", "1998", "1998", "1998", "1998"…  
## $ survey\_id <chr> "ZA1998DHS", "ZA1998DHS", "ZA1998DHS", "ZA1998…  
## $ indicator\_id <chr> "CM\_ECMT\_C\_NNR", "CM\_ECMT\_C\_PNR", "CM\_ECMT\_C\_I…  
## $ indicator\_order <dbl> 63166010, 63166020, 63166030, 63166040, 631660…  
## $ indicator\_type <chr> "I", "I", "I", "I", "I", "I", "I", "I", "I", "…  
## $ characteristic\_id <dbl> 13000, 13000, 13000, 13000, 13000, 1000, 1000,…  
## $ characteristic\_order <dbl> 80000, 80000, 80000, 80000, 80000, 0, 0, 0, 0,…  
## $ characteristic\_category <chr> "Five year periods", "Five year periods", "Fiv…  
## $ characteristic\_label <chr> "0-4", "0-4", "0-4", "0-4", "0-4", "Total", "T…  
## $ by\_variable\_id <chr> "0", "0", "0", "0", "0", "14001", "14003", "14…  
## $ by\_variable\_label <chr> NA, NA, NA, NA, NA, "Five years preceding the …  
## $ is\_total <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
## $ is\_preferred <dbl> 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1…  
## $ sdrid <chr> "CMECMTCNNR", "CMECMTCPNR", "CMECMTCIMR", "CME…  
## $ region\_id <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ survey\_year\_label <dbl> 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998…  
## $ survey\_type <chr> "DHS", "DHS", "DHS", "DHS", "DHS", "DHS", "DHS…  
## $ denominator\_weighted <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ denominator\_unweighted <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…  
## $ ci\_low <dbl> 15, 20, 38, 9, 50, 15, 16, 20, 19, 38, 37, 9, …  
## $ ci\_high <dbl> 25, 31, 53, 20, 68, 25, 23, 31, 27, 53, 48, 20…  
## $ level\_rank <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA…

# Summary stats  
skim(cmr\_df)

Data summary

|  |  |
| --- | --- |
| Name | cmr\_df |
| Number of rows | 40 |
| 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 | 1.0 | 3 | 3 | 0 | 1 | 0 |
| data\_id | 0 | 1.0 | 5 | 6 | 0 | 40 | 0 |
| indicator | 0 | 1.0 | 11 | 56 | 0 | 15 | 0 |
| value | 0 | 1.0 | 1 | 4 | 0 | 27 | 0 |
| precision | 0 | 1.0 | 1 | 1 | 0 | 1 | 0 |
| dhs\_country\_code | 0 | 1.0 | 2 | 2 | 0 | 1 | 0 |
| country\_name | 0 | 1.0 | 12 | 12 | 0 | 1 | 0 |
| survey\_year | 0 | 1.0 | 4 | 4 | 0 | 2 | 0 |
| survey\_id | 0 | 1.0 | 9 | 9 | 0 | 2 | 0 |
| indicator\_id | 0 | 1.0 | 13 | 13 | 0 | 15 | 0 |
| indicator\_type | 0 | 1.0 | 1 | 1 | 0 | 3 | 0 |
| characteristic\_category | 0 | 1.0 | 5 | 17 | 0 | 3 | 0 |
| characteristic\_label | 0 | 1.0 | 3 | 11 | 0 | 3 | 0 |
| by\_variable\_id | 0 | 1.0 | 1 | 5 | 0 | 3 | 0 |
| by\_variable\_label | 20 | 0.5 | 30 | 31 | 0 | 2 | 0 |
| sdrid | 0 | 1.0 | 10 | 10 | 0 | 15 | 0 |
| survey\_type | 0 | 1.0 | 3 | 3 | 0 | 1 | 0 |

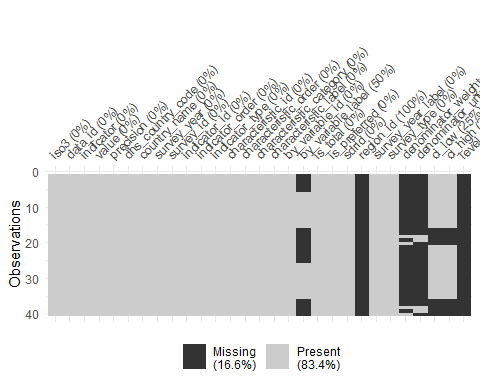
**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate | mean | count |
| --- | --- | --- | --- | --- |
| region\_id | 40 | 0 | NaN | : |
| level\_rank | 40 | 0 | NaN | : |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| indicator\_order | 0 | 1.00 | 63203530.00 | 25191.58 | 63166010 | 63196020.00 | 63206030.0 | 63213540.0 | 63236050 | ▃▁▇▁▃ |
| characteristic\_id | 0 | 1.00 | 6250.00 | 5424.30 | 1000 | 1000.00 | 5500.0 | 10750.0 | 13000 | ▇▁▁▃▃ |
| characteristic\_order | 0 | 1.00 | 22500.00 | 33873.82 | 0 | 0.00 | 5000.0 | 27500.0 | 80000 | ▇▁▁▁▂ |
| is\_total | 0 | 1.00 | 1.00 | 0.00 | 1 | 1.00 | 1.0 | 1.0 | 1 | ▁▁▇▁▁ |
| is\_preferred | 0 | 1.00 | 0.75 | 0.44 | 0 | 0.75 | 1.0 | 1.0 | 1 | ▂▁▁▁▇ |
| survey\_year\_label | 0 | 1.00 | 2007.00 | 9.11 | 1998 | 1998.00 | 2007.0 | 2016.0 | 2016 | ▇▁▁▁▇ |
| denominator\_weighted | 36 | 0.10 | 4348.00 | 890.27 | 3577 | 3577.00 | 4348.0 | 5119.0 | 5119 | ▇▁▁▁▇ |
| denominator\_unweighted | 36 | 0.10 | 4373.00 | 935.31 | 3563 | 3563.00 | 4373.0 | 5183.0 | 5183 | ▇▁▁▁▇ |
| ci\_low | 10 | 0.75 | 22.77 | 14.37 | 4 | 12.25 | 18.5 | 34.5 | 50 | ▇▆▂▅▃ |
| ci\_high | 10 | 0.75 | 35.20 | 17.60 | 10 | 20.75 | 30.0 | 51.0 | 68 | ▆▇▂▅▃ |

# Visualize missingness  
vis\_miss(cmr\_df)

 Purpose: Check structure, summary statistics, and missingness. Explanation: This gives an overview of column types, missing values, and potential issues before cleaning.

# Handle Duplicates

cat("Exact duplicates:", sum(duplicated(cmr\_df)), "\n")

## Exact duplicates: 0

cmr\_df <- cmr\_df %>% distinct()  
  
cat("Dimensions after deduplication:", dim(cmr\_df), "\n")

## Dimensions after deduplication: 40 29

## Drop Redundant / Empty Columns

redundant\_cols <- c(  
 "iso3", "data\_id", "dhs\_country\_code", "country\_name", "survey\_id",  
 "indicator\_id", "sdrid", "region\_id", "survey\_type", "level\_rank",  
 "denominator\_weighted", "denominator\_unweighted", "by\_variable\_label"  
)  
  
cmr\_df <- cmr\_df %>% select(-any\_of(redundant\_cols))  
  
cat("Dimensions after removing redundant/empty columns:", dim(cmr\_df), "\n")

## Dimensions after removing redundant/empty columns: 40 16

Columns that were unnecessary or fully empty were dropped:

* Redundant columns included identifiers such as iso3, data\_id, dhs\_country\_code, survey\_id, etc. ## Convert Column Types
* Ensure numeric, integer, and logical columns are typed correctly for analysis.

# Define expected columns by type (snake\_case!)  
numeric\_cols <- c("value", "precision", "ci\_low", "ci\_high")  
integer\_cols <- c("survey\_year", "indicator\_order", "characteristic\_id",  
 "characteristic\_order", "survey\_year\_label")  
logical\_cols <- c("is\_total", "is\_preferred")  
  
# Safe conversion function  
safe\_convert <- function(df, cols, fun) {  
 existing <- cols[cols %in% colnames(df)]  
 if(length(existing) > 0) {  
 df <- df %>% mutate(across(all\_of(existing), fun))  
 }  
 return(df)  
}  
  
cmr\_df <- cmr\_df %>%  
 safe\_convert(numeric\_cols, as.numeric) %>%  
 safe\_convert(integer\_cols, as.integer) %>%  
 safe\_convert(logical\_cols, ~as.logical(as.integer(.)))  
  
glimpse(cmr\_df)

## Rows: 40  
## Columns: 16  
## $ indicator <chr> "Neonatal mortality rate (5 year periods)", "P…  
## $ value <dbl> 20, 26, 45, 15, 59, 20, 19, 26, 23, 45, 42, 15…  
## $ precision <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
## $ survey\_year <int> 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998…  
## $ indicator\_order <int> 63166010, 63166020, 63166030, 63166040, 631660…  
## $ indicator\_type <chr> "I", "I", "I", "I", "I", "I", "I", "I", "I", "…  
## $ characteristic\_id <int> 13000, 13000, 13000, 13000, 13000, 1000, 1000,…  
## $ characteristic\_order <int> 80000, 80000, 80000, 80000, 80000, 0, 0, 0, 0,…  
## $ characteristic\_category <chr> "Five year periods", "Five year periods", "Fiv…  
## $ characteristic\_label <chr> "0-4", "0-4", "0-4", "0-4", "0-4", "Total", "T…  
## $ by\_variable\_id <chr> "0", "0", "0", "0", "0", "14001", "14003", "14…  
## $ is\_total <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…  
## $ is\_preferred <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRU…  
## $ survey\_year\_label <int> 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1998…  
## $ ci\_low <dbl> 15, 20, 38, 9, 50, 15, 16, 20, 19, 38, 37, 9, …  
## $ ci\_high <dbl> 25, 31, 53, 20, 68, 25, 23, 31, 27, 53, 48, 20…

## Handle Missing Values

# 1. Remove empty columns  
cmr\_df <- cmr\_df %>%  
 select(where(~!all(is.na(.))))  
  
# 2. Impute numeric with median  
num\_cols <- cmr\_df %>% select(where(is.numeric)) %>% names()  
cmr\_df <- cmr\_df %>%  
 mutate(across(all\_of(num\_cols), ~ifelse(is.na(.), median(., na.rm = TRUE), .)))  
  
# 3. Impute categorical with mode  
cat\_cols <- cmr\_df %>% select(where(is.character)) %>% names()  
impute\_mode <- function(x) {  
 ux <- na.omit(x)  
 if(length(ux) == 0) return(NA\_character\_)  
 names(sort(table(ux), decreasing = TRUE))[1]  
}  
cmr\_df <- cmr\_df %>%  
 mutate(across(all\_of(cat\_cols), ~ifelse(is.na(.), impute\_mode(.), .)))  
  
# 4. Summary after handling missing values  
missing\_summary <- cmr\_df %>%  
 summarise(across(everything(), ~sum(is.na(.)))) %>%  
 tidyr::pivot\_longer(cols = everything(), names\_to = "Variable", values\_to = "Missing\_Count")  
  
missing\_summary # this will be rendered in knit

## # A tibble: 16 × 2  
## Variable Missing\_Count  
## <chr> <int>  
## 1 indicator 0  
## 2 value 0  
## 3 precision 0  
## 4 survey\_year 0  
## 5 indicator\_order 0  
## 6 indicator\_type 0  
## 7 characteristic\_id 0  
## 8 characteristic\_order 0  
## 9 characteristic\_category 0  
## 10 characteristic\_label 0  
## 11 by\_variable\_id 0  
## 12 is\_total 0  
## 13 is\_preferred 0  
## 14 survey\_year\_label 0  
## 15 ci\_low 0  
## 16 ci\_high 0

* Completely empty columns were removed.
* Numeric columns: NAs imputed with median.
* Categorical columns: NAs imputed with mode.

## Handle Outliers

num\_cols <- cmr\_df %>% select(where(is.numeric))  
  
outlier\_bounds <- function(x) {  
 qnt <- quantile(x, probs=c(0.25, 0.75), na.rm=TRUE)  
 iqr <- diff(qnt)  
 c(lower=qnt[1]-1.5\*iqr, upper=qnt[2]+1.5\*iqr)  
}  
  
bounds <- map(num\_cols, outlier\_bounds)  
  
cmr\_df <- cmr\_df %>%  
 mutate(across(where(is.numeric),  
 ~pmin(pmax(., bounds[[cur\_column()]]["lower"]),  
 bounds[[cur\_column()]]["upper"])))

## Handle Noise / Special Values

cmr\_df <- cmr\_df %>%  
 mutate(across(where(is.numeric),  
 ~ifelse(. < 0, median(., na.rm = TRUE), .)))

* Negative numeric values were replaced with the column median.

## Save the Cleaned Dataset

# Save cleaned dataset to processed folder  
write\_csv(cmr\_df, here("data", "processed", "child-mortality-rates\_cleaned.csv"))  
  
cat("Cleaned dataset saved successfully.\n")

## Cleaned dataset saved successfully.

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

## Dimensions: 40 16