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"),
```

```r
  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.
```

```r
cat("Dimensions:", dim(cmr_df), "\n")
```

```
## Dimensions: 40 29
```

Initial Data Assessment

```r
# 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…
```

```r
# 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

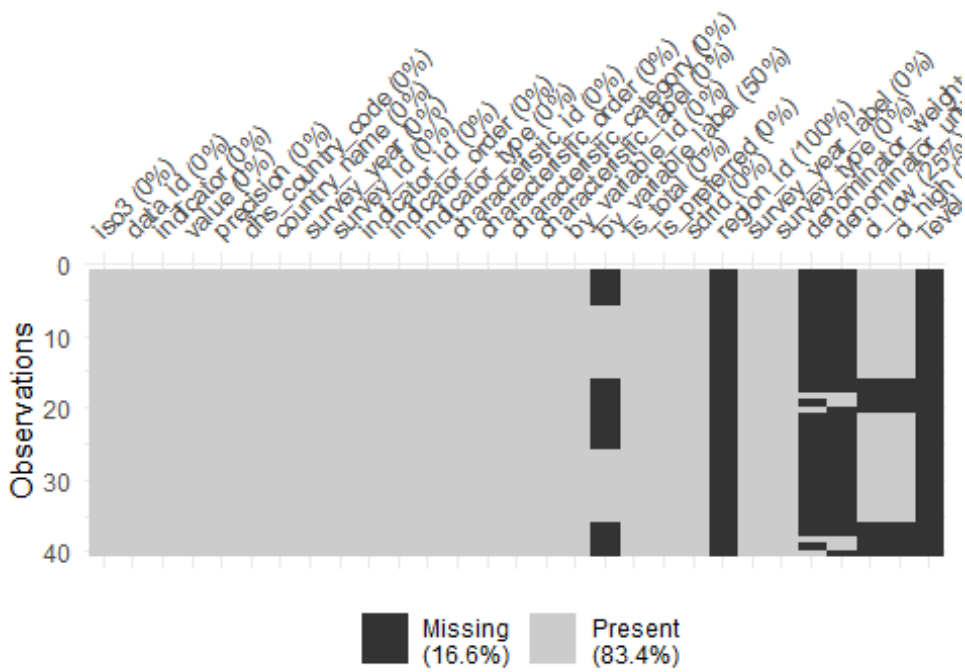
skim_variable	n_missing	complete_rate	mean	count
region_id	40	0	NaN	:
level_rank	40	0	NaN	:

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	▇▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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) %>%
```

```r
  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

```r
# 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),
.)))
```

```r
# 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

```r
num_cols <- cmr_df %>% select(where(is.numeric))

outlier_bounds <- function(x) {
```

```r
  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

```r
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

```r
# 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
```