04\_covid\_prevention

# Covid 19 Prevention

## 1. Load Libraries & data

# 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 COVID-19 prevention dataset  
# Load the COVID-19 prevention dataset, skipping first row if it contains metadata  
covid\_df <- read\_csv(here("data", "raw", "covid-19-prevention\_national\_zaf.csv"))

## Rows: 35 Columns: 29

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (17): ISO3, DataId, Indicator, Value, Precision, DHS\_CountryCode, Countr...  
## dbl (8): IndicatorOrder, CharacteristicId, CharacteristicOrder, IsTotal, Is...  
## lgl (4): RegionId, CILow, CIHigh, 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.

# Step 2: Remove first row (metadata)  
covid\_df <- covid\_df[-1, ]  
  
# Step 3: Reset row names  
rownames(covid\_df) <- NULL  
  
# Step 4: Optional: in  
cat("COVID-19 prevention dataset loaded successfully.\n")

## COVID-19 prevention dataset loaded successfully.

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

## Dimensions: 34 29

## 2. Initial Data Assessment and Column Renaming

* Column names are standardized to lowercase with underscores for readability. The dataset structure, summary statistics, and missingness are explored to identify potential quality issues.

# Quick glimpse of dataset  
glimpse(covid\_df)

## Rows: 34  
## Columns: 29  
## $ ISO3 <chr> "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF"…  
## $ DataId <chr> "795844", "795750", "795755", "795740", "795744…  
## $ Indicator <chr> "Population using an improved water source", "P…  
## $ Value <chr> "83.5", "36", "23.1", "19.3", "60.3", "80.2", "…  
## $ Precision <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1…  
## $ 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> "WS\_SRCE\_P\_IMP", "WS\_SRCE\_P\_PIP", "WS\_SRCE\_P\_PY…  
## $ IndicatorOrder <dbl> 250162010, 250162020, 250162025, 250162030, 250…  
## $ 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", "0", "0", "0…  
## $ ByVariableLabel <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ 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, 1, 1, 1, 1, 1, 1, 1, 1,…  
## $ SDRID <chr> "WSSRCEPIMP", "WSSRCEPPIP", "WSSRCEPPYD", "WSSR…  
## $ 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> 52007, 52007, 52007, 52007, 52007, 52007, 52007…  
## $ DenominatorUnweighted <dbl> 52465, 52465, 52465, 52465, 52465, 52465, 52465…  
## $ CILow <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ CIHigh <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ LevelRank <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…

# Summary of missingness  
skim(covid\_df)

Data summary

|  |  |
| --- | --- |
| Name | covid\_df |
| Number of rows | 34 |
| Number of columns | 29 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 17 |
| logical | 4 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ISO3 | 0 | 1 | 3 | 3 | 0 | 1 | 0 |
| DataId | 0 | 1 | 5 | 6 | 0 | 34 | 0 |
| Indicator | 0 | 1 | 32 | 75 | 0 | 20 | 0 |
| Value | 0 | 1 | 2 | 4 | 0 | 34 | 0 |
| Precision | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| DHS\_CountryCode | 0 | 1 | 2 | 2 | 0 | 1 | 0 |
| CountryName | 0 | 1 | 12 | 12 | 0 | 1 | 0 |
| SurveyYear | 0 | 1 | 4 | 4 | 0 | 2 | 0 |
| SurveyId | 0 | 1 | 9 | 9 | 0 | 2 | 0 |
| IndicatorId | 0 | 1 | 13 | 13 | 0 | 20 | 0 |
| IndicatorType | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| CharacteristicCategory | 0 | 1 | 5 | 5 | 0 | 1 | 0 |
| CharacteristicLabel | 0 | 1 | 5 | 5 | 0 | 1 | 0 |
| ByVariableId | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| ByVariableLabel | 34 | 0 | NA | NA | 0 | 0 | 0 |
| SDRID | 0 | 1 | 10 | 10 | 0 | 20 | 0 |
| SurveyType | 0 | 1 | 3 | 3 | 0 | 1 | 0 |

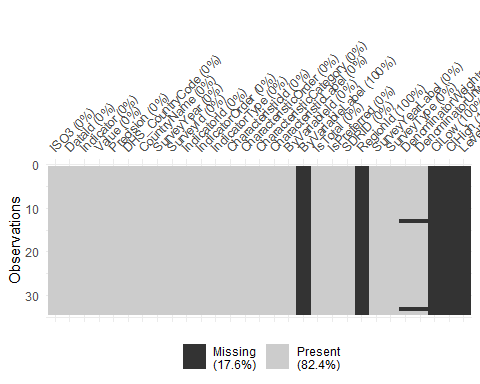
**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate | mean | count |
| --- | --- | --- | --- | --- |
| RegionId | 34 | 0 | NaN | : |
| CILow | 34 | 0 | NaN | : |
| CIHigh | 34 | 0 | NaN | : |
| LevelRank | 34 | 0 | NaN | : |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| IndicatorOrder | 0 | 1.00 | 252040162.06 | 4011155.25 | 250162010 | 250162190 | 250252010 | 250292085 | 260831120 | ▇▁▁▁▂ |
| CharacteristicId | 0 | 1.00 | 1000.00 | 0.00 | 1000 | 1000 | 1000 | 1000 | 1000 | ▁▁▇▁▁ |
| CharacteristicOrder | 0 | 1.00 | 0.00 | 0.00 | 0 | 0 | 0 | 0 | 0 | ▁▁▇▁▁ |
| IsTotal | 0 | 1.00 | 1.00 | 0.00 | 1 | 1 | 1 | 1 | 1 | ▁▁▇▁▁ |
| IsPreferred | 0 | 1.00 | 1.00 | 0.00 | 1 | 1 | 1 | 1 | 1 | ▁▁▇▁▁ |
| SurveyYearLabel | 0 | 1.00 | 2008.59 | 8.99 | 1998 | 1998 | 2016 | 2016 | 2016 | ▆▁▁▁▇ |
| DenominatorWeighted | 2 | 0.94 | 38815.16 | 12658.90 | 11066 | 37205 | 37205 | 52007 | 52007 | ▂▁▂▇▆ |
| DenominatorUnweighted | 2 | 0.94 | 39352.88 | 12765.80 | 11066 | 37925 | 37925 | 52465 | 52465 | ▂▁▂▇▆ |

# visualize missing values  
vis\_miss(covid\_df)



# Clean column names to lowercase with underscores  
covid\_df <- covid\_df %>% janitor::clean\_names()  
  
# Check new names  
colnames(covid\_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"

## Handle Duplicates

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

## Exact duplicate rows: 0

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

## Dimensions after duplicate removal: 34 29

## Convert Data Types

* Ensures all numeric, integer, and logical columns have correct types for downstream analysis. Prevents calculation errors and improves consistency.

# Convert numeric columns  
covid\_df <- covid\_df %>%  
 mutate(  
 across(c(value, precision, denominator\_weighted, denominator\_unweighted, ci\_low, ci\_high), as.numeric),  
 across(c(survey\_year, indicator\_order, characteristic\_id, characteristic\_order, survey\_year\_label, by\_variable\_id), as.integer),  
 across(c(is\_total, is\_preferred), ~as.logical(as.integer(.)))  
 )

## Drop Redundant 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")  
  
covid\_df <- covid\_df %>% select(-any\_of(redundant\_cols))  
  
# Remove columns that are entirely NA  
covid\_df <- covid\_df %>% select(where(~!all(is.na(.))))  
  
cat("Redundant and empty columns removed.\n")

## Redundant and empty columns removed.

cat("New dimensions:", dim(covid\_df), "\n")

## New dimensions: 34 14

## Handle Missing Values

* Numeric columns: filled with the median
* Character columns: filled with the most frequent value
* Logical columns: missing values set to FALSE
* Key metadata (survey\_year\_label, survey\_type) imputed explicitly for clarity

covid\_df <- covid\_df %>%  
 select(where(~!all(is.na(.))))  
  
  
impute\_mode <- function(x) {  
 ux <- na.omit(x)  
 if(length(ux) == 0) return(x)  
 rep(names(sort(table(ux), decreasing = TRUE))[1], length(x))  
}  
  
  
covid\_df <- covid\_df %>%  
 mutate(  
 # Numeric columns → median  
 across(where(is.numeric), ~ifelse(is.na(.), median(., na.rm = TRUE), .)),  
   
 # Character columns → mode  
 across(where(is.character), ~ifelse(is.na(.), impute\_mode(.), .)),  
   
 # Logical columns → set missing to FALSE (or TRUE if appropriate)  
 across(where(is.logical), ~ifelse(is.na(.), FALSE, .))  
 )  
  
  
missing\_summary <- data.frame(  
 Column = colnames(covid\_df),  
 n\_missing = colSums(is.na(covid\_df)),  
 total\_rows = nrow(covid\_df),  
 missing\_percent = round(colSums(is.na(covid\_df)) / nrow(covid\_df) \* 100, 2)  
)  
  
missing\_summary %>% arrange(desc(missing\_percent))

## Column n\_missing total\_rows  
## indicator indicator 0 34  
## value value 0 34  
## precision precision 0 34  
## survey\_year survey\_year 0 34  
## indicator\_order indicator\_order 0 34  
## indicator\_type indicator\_type 0 34  
## characteristic\_id characteristic\_id 0 34  
## characteristic\_order characteristic\_order 0 34  
## characteristic\_category characteristic\_category 0 34  
## characteristic\_label characteristic\_label 0 34  
## by\_variable\_id by\_variable\_id 0 34  
## is\_total is\_total 0 34  
## is\_preferred is\_preferred 0 34  
## survey\_year\_label survey\_year\_label 0 34  
## missing\_percent  
## indicator 0  
## value 0  
## precision 0  
## survey\_year 0  
## indicator\_order 0  
## indicator\_type 0  
## characteristic\_id 0  
## characteristic\_order 0  
## characteristic\_category 0  
## characteristic\_label 0  
## by\_variable\_id 0  
## is\_total 0  
## is\_preferred 0  
## survey\_year\_label 0

## Handle Outliers

* Extreme values in value are capped to the IQR boundaries (Winsorization), which reduces their influence while keeping most data intact.

# Quick check for extreme values in 'value' and denominators  
summary(covid\_df$value)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.50 6.45 31.20 36.74 56.85 96.00

summary(covid\_df$denominator\_weighted)

## Warning: Unknown or uninitialised column: `denominator\_weighted`.

## Length Class Mode   
## 0 NULL NULL

summary(covid\_df$denominator\_unweighted)

## Warning: Unknown or uninitialised column: `denominator\_unweighted`.

## Length Class Mode   
## 0 NULL NULL

# Identify potential outliers using IQR method  
outliers <- covid\_df %>%  
 filter(value > quantile(value, 0.75, na.rm = TRUE) + 1.5 \* IQR(value, na.rm = TRUE) |  
 value < quantile(value, 0.25, na.rm = TRUE) - 1.5 \* IQR(value, na.rm = TRUE))  
  
cat("Potential outlier rows in 'value':", nrow(outliers), "\n")

## Potential outlier rows in 'value': 0

## Final Validation

## Final Dataset Check Before Saving (existing columns only)  
  
# Check dataset dimensions and structure  
cat("Final dataset dimensions:", dim(covid\_df), "\n")

## Final dataset dimensions: 34 14

str(covid\_df)

## tibble [34 × 14] (S3: tbl\_df/tbl/data.frame)  
## $ indicator : chr [1:34] "Population using an improved water source" "Population using water piped into dwelling" "Population using water piped into yard/plot" "Population using a public tap/standpipe" ...  
## $ value : num [1:34] 83.5 36 23.1 19.3 60.3 80.2 3.3 8.4 46.4 40.8 ...  
## $ precision : num [1:34] 1 1 1 1 1 1 1 1 1 1 ...  
## $ survey\_year : int [1:34] 1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...  
## $ indicator\_order : int [1:34] 250162010 250162020 250162025 250162030 250162190 250162200 250162210 250202030 250262010 250262150 ...  
## $ indicator\_type : chr [1:34] "I" "I" "I" "I" ...  
## $ characteristic\_id : int [1:34] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 ...  
## $ characteristic\_order : int [1:34] 0 0 0 0 0 0 0 0 0 0 ...  
## $ characteristic\_category: chr [1:34] "Total" "Total" "Total" "Total" ...  
## $ characteristic\_label : chr [1:34] "Total" "Total" "Total" "Total" ...  
## $ by\_variable\_id : int [1:34] 0 0 0 0 0 0 0 0 0 0 ...  
## $ is\_total : logi [1:34] TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ is\_preferred : logi [1:34] TRUE TRUE TRUE TRUE TRUE TRUE ...  
## $ survey\_year\_label : int [1:34] 1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...

# Identify numeric columns that exist  
numeric\_cols <- covid\_df %>% select(where(is.numeric)) %>% colnames()  
  
# Summarize numeric columns for final inspection  
summary(select(covid\_df, all\_of(numeric\_cols)))

## value precision survey\_year indicator\_order   
## Min. : 1.50 Min. :0.0000 Min. :1998 Min. :250162010   
## 1st Qu.: 6.45 1st Qu.:1.0000 1st Qu.:1998 1st Qu.:250162190   
## Median :31.20 Median :1.0000 Median :2016 Median :250252010   
## Mean :36.74 Mean :0.9706 Mean :2009 Mean :252040162   
## 3rd Qu.:56.85 3rd Qu.:1.0000 3rd Qu.:2016 3rd Qu.:250292085   
## Max. :96.00 Max. :1.0000 Max. :2016 Max. :260831120   
## characteristic\_id characteristic\_order by\_variable\_id survey\_year\_label  
## Min. :1000 Min. :0 Min. :0 Min. :1998   
## 1st Qu.:1000 1st Qu.:0 1st Qu.:0 1st Qu.:1998   
## Median :1000 Median :0 Median :0 Median :2016   
## Mean :1000 Mean :0 Mean :0 Mean :2009   
## 3rd Qu.:1000 3rd Qu.:0 3rd Qu.:0 3rd Qu.:2016   
## Max. :1000 Max. :0 Max. :0 Max. :2016

# Confirm no remaining missing values in all columns  
missing\_summary <- covid\_df %>%  
 summarise(across(everything(), ~sum(is.na(.)))) %>%  
 pivot\_longer(cols = everything(), names\_to = "column", values\_to = "n\_missing") %>%  
 mutate(  
 total\_rows = nrow(covid\_df),  
 missing\_percent = round(n\_missing / total\_rows \* 100, 2)  
 )  
  
missing\_summary %>% arrange(desc(missing\_percent))

## # A tibble: 14 × 4  
## column n\_missing total\_rows missing\_percent  
## <chr> <int> <int> <dbl>  
## 1 indicator 0 34 0  
## 2 value 0 34 0  
## 3 precision 0 34 0  
## 4 survey\_year 0 34 0  
## 5 indicator\_order 0 34 0  
## 6 indicator\_type 0 34 0  
## 7 characteristic\_id 0 34 0  
## 8 characteristic\_order 0 34 0  
## 9 characteristic\_category 0 34 0  
## 10 characteristic\_label 0 34 0  
## 11 by\_variable\_id 0 34 0  
## 12 is\_total 0 34 0  
## 13 is\_preferred 0 34 0  
## 14 survey\_year\_label 0 34 0

## Save Dataset

# Define path to save cleaned CSV  
clean\_path <- here("data", "processed", "covid\_prevention\_cleaned\_zaf.csv")  
  
# Create folder if it doesn't exist  
if(!dir.exists(dirname(clean\_path))) dir.create(dirname(clean\_path), recursive = TRUE)  
  
# Write cleaned dataset  
write\_csv(covid\_df, clean\_path)  
  
cat("Cleaned COVID-19 prevention dataset saved successfully at:\n", clean\_path, "\n")

## Cleaned COVID-19 prevention dataset saved successfully at:  
## C:/Users/morul/School/3rd Year/BIN381/BIN381\_PROJECT/BIN381\_PROJECT/data/processed/covid\_prevention\_cleaned\_zaf.csv