10_Maternal_mortality

#Loading Libraries

```
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(stringr)
library(readr)
library(here)

## here() starts at C:/Users/morul/School/3rd
Year/BIN381/BIN381_PROJECT/BIN381_PROJECT

library(purrr)
```

#Load Dataset

```
mam_df <- read_csv(here("data","raw", "maternal-mortality_national_zaf.csv"))

## Rows: 22 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.
```

#Disdplay Dataset content

```
head(mam_df)

## # A tibble: 6 × 29
##   ISO3   DataId Indicator Value Precision DHS_CountryCode CountryName
```

```
SurveyYear
##   <chr>  <chr>  <chr>     <chr> <chr>     <chr>           <chr>
<chr>
## 1 #coun… #meta… #indicat… #ind… #indicat… <NA>            #country+n…
#date+year
## 2 ZAF    91409  Female d… 5.5   1         ZA              South Afri… 1998
## 3 ZAF    91377  Number o… 19    0         ZA              South Afri… 1998
## 4 ZAF    768646 Years of… 1227… 0         ZA              South Afri… 1998
## 5 ZAF    768647 Years of… 1237… 0         ZA              South Afri… 1998
## 6 ZAF    535566 Pregnanc… 0.15  2         ZA              South Afri… 1998
## # ℹ 21 more variables: SurveyId <chr>, IndicatorId <chr>, IndicatorOrder
<dbl>,
## #   IndicatorType <chr>, CharacteristicId <dbl>, CharacteristicOrder
<dbl>,
## #   CharacteristicCategory <chr>, CharacteristicLabel <chr>,
## #   ByVariableId <chr>, ByVariableLabel <chr>, IsTotal <dbl>,
## #   IsPreferred <dbl>, SDRID <chr>, RegionId <lgl>, SurveyYearLabel <dbl>,
## #   SurveyType <chr>, DenominatorWeighted <dbl>, DenominatorUnweighted
<dbl>,
## #   CILow <dbl>, CIHigh <dbl>, LevelRank <lgl>
```

#Remove the first row(meta data)

```
mam_df <- mam_df[-1, ]
```

#dimensions

```
dim(mam_df)

## [1] 21 29
```

#Inspect Duplicated rows

```
dup_check <- mam_df %>%
  group_by(Indicator, SurveyYear, CharacteristicId, Value) %>%
  filter(n() > 1)

dup_check

## # A tibble: 0 × 29
## # Groups:   Indicator, SurveyYear, CharacteristicId, Value [0]
## # ℹ 29 variables: ISO3 <chr>, DataId <chr>, Indicator <chr>, Value <chr>,
## #   Precision <chr>, DHS_CountryCode <chr>, CountryName <chr>,
## #   SurveyYear <chr>, SurveyId <chr>, IndicatorId <chr>, IndicatorOrder
<dbl>,
## #   IndicatorType <chr>, CharacteristicId <dbl>, CharacteristicOrder
<dbl>,
## #   CharacteristicCategory <chr>, CharacteristicLabel <chr>,
## #   ByVariableId <chr>, ByVariableLabel <chr>, IsTotal <dbl>,
## #   IsPreferred <dbl>, SDRID <chr>, RegionId <lgl>, SurveyYearLabel <dbl>,
…
```

#perc mising values

```
data.frame(
  Column = names(mam_df),
  Missing_Percentage = paste0(round(colMeans(is.na(mam_df)) * 100, 2), "%")
  )
```

```
##                    Column Missing_Percentage
## 1                    ISO3                 0%
## 2                  DataId                 0%
## 3               Indicator                 0%
## 4                   Value                 0%
## 5               Precision                 0%
## 6         DHS_CountryCode                 0%
## 7             CountryName                 0%
## 8              SurveyYear                 0%
## 9                SurveyId                 0%
## 10            IndicatorId                 0%
## 11         IndicatorOrder                 0%
## 12          IndicatorType                 0%
## 13       CharacteristicId                 0%
## 14    CharacteristicOrder                 0%
## 15 CharacteristicCategory                 0%
## 16    CharacteristicLabel                 0%
## 17           ByVariableId                 0%
## 18        ByVariableLabel               100%
## 19                IsTotal                 0%
## 20            IsPreferred                 0%
## 21                  SDRID                 0%
## 22               RegionId               100%
## 23        SurveyYearLabel                 0%
## 24             SurveyType                 0%
## 25    DenominatorWeighted             90.48%
## 26  DenominatorUnweighted             71.43%
## 27                  CILow             85.71%
## 28                 CIHigh             85.71%
## 29              LevelRank               100%
```

```
mam_df <- mam_df %>%
  select(-RegionId, -LevelRank, -CILow, -CIHigh)  # 100% or 85% missing

# 2. Impute numeric columns with missing values
# Here, only DenominatorWeighted and DenominatorUnweighted
num_cols <- c("DenominatorWeighted", "DenominatorUnweighted")

mam_df <- mam_df %>%
  mutate(across(all_of(num_cols), ~ ifelse(is.na(.), median(., na.rm = TRUE),
.)))

# 3. Fill any remaining missing values using last observation carried
```

```
forward/backward
mam_df <- mam_df %>%
  fill(DenominatorWeighted, DenominatorUnweighted, .direction = "downup")

# 4. Check that missing values are gone
data.frame(
  Column = names(mam_df),
  Missing_Data = colSums(is.na(mam_df))
)
```

```
##                                         Column Missing_Data
## ISO3                                      ISO3            0
## DataId                                  DataId            0
## Indicator                            Indicator            0
## Value                                    Value            0
## Precision                            Precision            0
## DHS_CountryCode                DHS_CountryCode            0
## CountryName                        CountryName            0
## SurveyYear                          SurveyYear            0
## SurveyId                              SurveyId            0
## IndicatorId                        IndicatorId            0
## IndicatorOrder                  IndicatorOrder            0
## IndicatorType                    IndicatorType            0
## CharacteristicId              CharacteristicId            0
## CharacteristicOrder        CharacteristicOrder            0
## CharacteristicCategory CharacteristicCategory            0
## CharacteristicLabel        CharacteristicLabel            0
## ByVariableId                      ByVariableId            0
## ByVariableLabel                ByVariableLabel           21
## IsTotal                                IsTotal            0
## IsPreferred                        IsPreferred            0
## SDRID                                    SDRID            0
## SurveyYearLabel                SurveyYearLabel            0
## SurveyType                          SurveyType            0
## DenominatorWeighted        DenominatorWeighted            0
## DenominatorUnweighted    DenominatorUnweighted            0
```

#check data types

```
data.frame(
  Column = names(mam_df),
  paste0(sapply(mam_df, typeof))
)
```

```
##                     Column paste0.sapply.mam_df..typeof..
## 1                     ISO3                      character
## 2                   DataId                      character
## 3                Indicator                      character
## 4                    Value                      character
## 5                Precision                      character
## 6          DHS_CountryCode                      character
## 7             CountryName                      character
## 8              SurveyYear                      character
## 9                SurveyId                      character
## 10            IndicatorId                      character
## 11         IndicatorOrder                         double
## 12          IndicatorType                      character
## 13       CharacteristicId                         double
## 14    CharacteristicOrder                         double
## 15 CharacteristicCategory                      character
## 16    CharacteristicLabel                      character
## 17           ByVariableId                      character
## 18        ByVariableLabel                      character
## 19                IsTotal                         double
## 20            IsPreferred                         double
## 21                  SDRID                      character
## 22        SurveyYearLabel                         double
## 23             SurveyType                      character
## 24    DenominatorWeighted                         double
## 25  DenominatorUnweighted                         double
```

#Check The structure of the dataset

```
str(mam_df)
```

```
## tibble [21 × 25] (S3: tbl_df/tbl/data.frame)
##  $ ISO3                  : chr [1:21] "ZAF" "ZAF" "ZAF" "ZAF" ...
##  $ DataId                : chr [1:21] "91409" "91377" "768646" "768647"
...
##  $ Indicator             : chr [1:21] "Female deaths that are pregnancy-
related" "Number of pregnancy-related deaths" "Years of exposure to the risk
of mortality for women" "Years of exposure to the risk of mortality for women
(unweighted)" ...
##  $ Value                 : chr [1:21] "5.5" "19" "122701" "123738" ...
##  $ Precision             : chr [1:21] "1" "0" "0" "0" ...
##  $ DHS_CountryCode       : chr [1:21] "ZA" "ZA" "ZA" "ZA" ...
##  $ CountryName           : chr [1:21] "South Africa" "South Africa" "South
Africa" "South Africa" ...
##  $ SurveyYear            : chr [1:21] "1998" "1998" "1998" "1998" ...
##  $ SurveyId              : chr [1:21] "ZA1998DHS" "ZA1998DHS" "ZA1998DHS"
"ZA1998DHS" ...
##  $ IndicatorId           : chr [1:21] "MM_MMRT_W_FDP" "MM_MMRT_W_PDT"
"MM_MMRT_W_EXP" "MM_MMRT_W_EXU" ...
##  $ IndicatorOrder        : num [1:21] 7.7e+07 7.7e+07 7.7e+07 7.7e+07
7.7e+07 ...
##  $ IndicatorType         : chr [1:21] "I" "N" "D" "U" ...
##  $ CharacteristicId      : num [1:21] 10000 10000 10000 10000 10000 1000
1000 1000 1000 1000 ...
##  $ CharacteristicOrder   : num [1:21] 10000 10000 10000 10000 10000 0 0 0
0 0 ...
##  $ CharacteristicCategory: chr [1:21] "Total 15-49" "Total 15-49" "Total
```

```
15-49" "Total 15-49" ...
##  $ CharacteristicLabel   : chr [1:21] "Total 15-49" "Total 15-49" "Total 
15-49" "Total 15-49" ...
##  $ ByVariableId          : chr [1:21] "0" "0" "0" "0" ...
##  $ ByVariableLabel       : chr [1:21] NA NA NA NA ...
##  $ IsTotal               : num [1:21] 1 1 1 1 1 1 1 1 1 1 ...
##  $ IsPreferred           : num [1:21] 1 1 1 1 1 1 1 1 1 1 ...
##  $ SDRID                 : chr [1:21] "MMMMRTWFDP" "MMMMRTWPDT" 
"MMMMRTWEXP" "MMMMRTWEXU" ...
##  $ SurveyYearLabel       : num [1:21] 1998 1998 1998 1998 1998 ...
##  $ SurveyType            : chr [1:21] "DHS" "DHS" "DHS" "DHS" ...
##  $ DenominatorWeighted   : num [1:21] 92735 92735 92735 92735 122701 ...
##  $ DenominatorUnweighted : num [1:21] 93631 93631 123738 123738 123738 ...
```

#Convert Data Types

```
mam_df <- mam_df %>%
  mutate(
        Value = as.numeric(Value),
    Precision = as.numeric(Precision),
    SurveyYear = as.integer(SurveyYear),
    IndicatorOrder = as.integer(IndicatorOrder),
    CharacteristicId = as.integer(CharacteristicId),
    CharacteristicOrder = as.integer(CharacteristicOrder),
    IsTotal = as.logical(as.integer(IsTotal)),
    IsPreferred = as.logical(as.integer(IsPreferred)),
    SurveyYearLabel = as.integer(SurveyYearLabel),
    DenominatorWeighted = as.numeric(DenominatorWeighted),
    DenominatorUnweighted = as.numeric(DenominatorUnweighted),
  )
```

Summary table: column name, number of unique values, sample of unique values

```
n_sample <- 3

summary_tbl <- mam_df %>%
  map_df(~ tibble(
    n_unique = n_distinct(.),
    sample_values = paste(head(unique(.), n_sample), collapse = ", ")
  ), .id = "column")

summary_tbl
```

```
## # A tibble: 25 × 3
##    column          n_unique sample_values
##    <chr>              <int> <chr>
##  1 ISO3                   1 ZAF
##  2 DataId                21 91409, 91377, 768646
##  3 Indicator             11 Female deaths that are pregnancy-related,
Number of…
##  4 Value                 21 5.5, 19, 122701
##  5 Precision              4 1, 0, 2
##  6 DHS_CountryCode        1 ZA
##  7 CountryName            1 South Africa
##  8 SurveyYear             2 1998, 2016
##  9 SurveyId               2 ZA1998DHS, ZA2016DHS
## 10 IndicatorId           11 MM_MMRT_W_FDP, MM_MMRT_W_PDT, MM_MMRT_W_EXP
## # ℹ 15 more rows
```

Drop the countries only one unqiue value: reason, there is no useful information - county is also always za

```r
# See exact column names
colnames(mam_df)
```

```
##  [1] "ISO3"                   "DataId"                 "Indicator"
##  [4] "Value"                  "Precision"              "DHS_CountryCode"
##  [7] "CountryName"            "SurveyYear"             "SurveyId"
## [10] "IndicatorId"            "IndicatorOrder"         "IndicatorType"
## [13] "CharacteristicId"       "CharacteristicOrder"
"CharacteristicCategory"
## [16] "CharacteristicLabel"    "ByVariableId"           "ByVariableLabel"
## [19] "IsTotal"                "IsPreferred"            "SDRID"
## [22] "SurveyYearLabel"        "SurveyType"
"DenominatorWeighted"
## [25] "DenominatorUnweighted"
```

```r
# Then drop using safe selection
cols_to_drop <- c("iso3", "dhs_country_code", "country_name", "survey_id",
                  "by_variable_id", "by_variable_label", "is_total",
                  "region_id", "survey_year_label", "survey_type",
"characteristic_order")

# Only drop columns that exist
mam_df <- mam_df %>% select(-any_of(cols_to_drop))

# Confirm
colnames(mam_df)
```

```
##  [1] "ISO3"                   "DataId"                 "Indicator"
##  [4] "Value"                  "Precision"              "DHS_CountryCode"
##  [7] "CountryName"            "SurveyYear"             "SurveyId"
## [10] "IndicatorId"            "IndicatorOrder"         "IndicatorType"
## [13] "CharacteristicId"       "CharacteristicOrder"
"CharacteristicCategory"
## [16] "CharacteristicLabel"    "ByVariableId"           "ByVariableLabel"
## [19] "IsTotal"                "IsPreferred"            "SDRID"
## [22] "SurveyYearLabel"        "SurveyType"
"DenominatorWeighted"
## [25] "DenominatorUnweighted"
```

Outliers

```
# Statistical outlier detection
outlier_stats <- mam_df %>%
  summarise(
    mean_value = mean(Value, na.rm = TRUE),
    sd_value = sd(Value, na.rm = TRUE),
    outliers_upper = sum(Value > mean_value + 2*sd_value, na.rm = TRUE),
    outliers_lower = sum(Value < mean_value - 2*sd_value, na.rm = TRUE)
  )



print(outlier_stats)
```

```
## # A tibble: 1 × 4
##   mean_value sd_value outliers_upper outliers_lower
##        <dbl>    <dbl>          <int>          <int>
## 1     17882.   39767.              2              0
```

- The data set does not have any outliers so no need to handle

#save cleaned data

```
write_csv(mam_df, here("data","processed", "maternal-mortality_cleaned.csv"))
```