CL_HIV_Behavior_08

Group K

2025-09-19

Display Dataset content

```
## # A tibble: 5 x 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   795160 Sex befo~ 8     1         ZA              South Afri~ 1998
## 3 ZAF   795161 Number o~ 4324  0         ZA              South Afri~ 1998
## 4 ZAF   796612 Number o~ 4459  0         ZA              South Afri~ 1998
## 5 ZAF   795358 Sex befo~ 54.5  1         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 <lgl>, CIHigh <lgl>, LevelRank <lgl>
```

#Remove the first row(meta data)

```
hiv_df <- hiv_df[-1, ]
```

check data types

```
##                    Column paste0.sapply.hiv_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               RegionId                        logical
## 23        SurveyYearLabel                         double
## 24             SurveyType                      character
## 25    DenominatorWeighted                         double
## 26  DenominatorUnweighted                         double
## 27                  CILow                        logical
## 28                 CIHigh                        logical
## 29              LevelRank                        logical
```

#Convert Data Types

check for unique values

```
## # A tibble: 29 x 3
##    column          n_unique sample_values
##    <chr>              <int> <chr>
##  1 ISO3                   1 ZAF
##  2 DataId               106 795160, 795161, 796612
##  3 Indicator             77 Sex before the age of 15 [Women], Number of young w~
##  4 Value                 99 8, 4324, 4459
##  5 Precision              2 1, 0
##  6 DHS_CountryCode        1 ZA
##  7 CountryName            1 South Africa
##  8 SurveyYear             2 1998, 2016
##  9 SurveyId               2 ZA1998DHS, ZA2016DHS
## 10 IndicatorId           91 HA_AFSY_W_A15, HA_AFSY_W_NM1, HA_AFSY_W_UN1
## # i 19 more rows
```

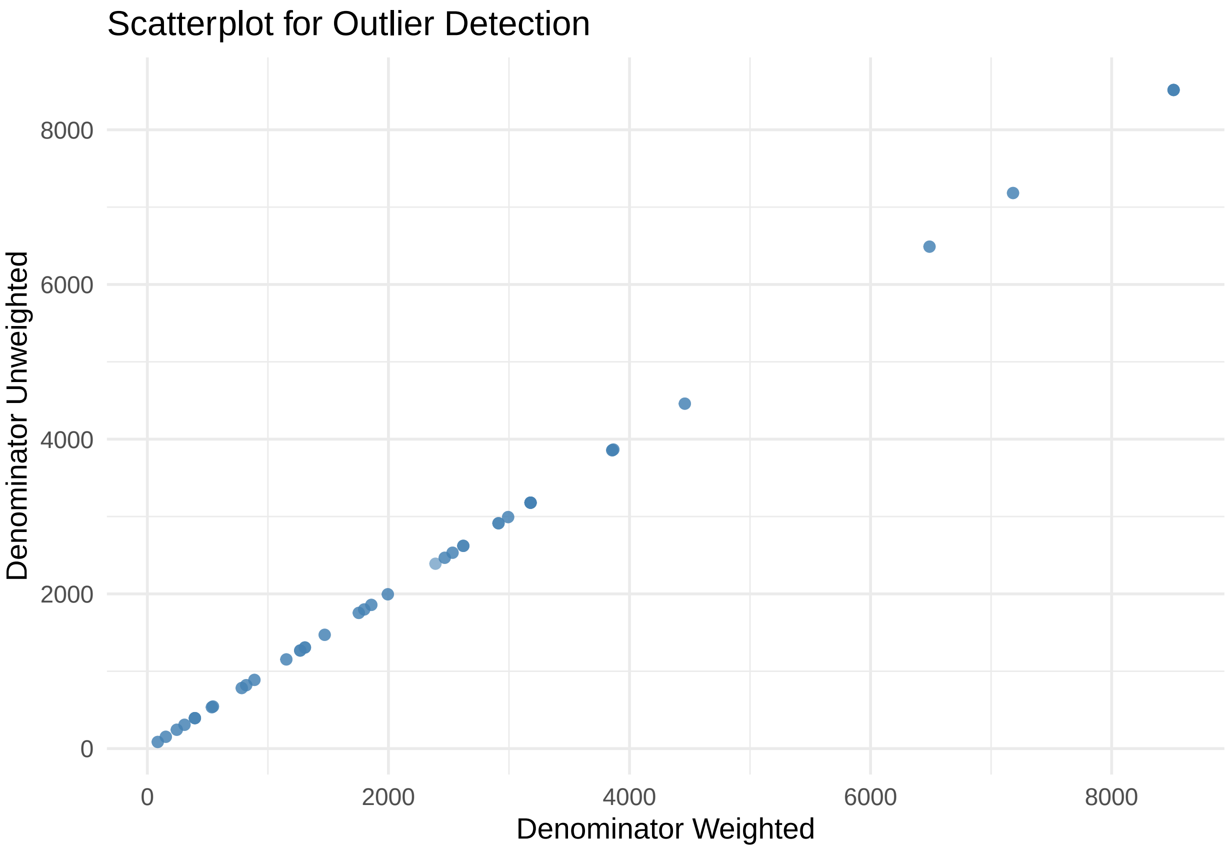
#Drop the countries only one unqiue value: reason, there is no useful information - county is also always za

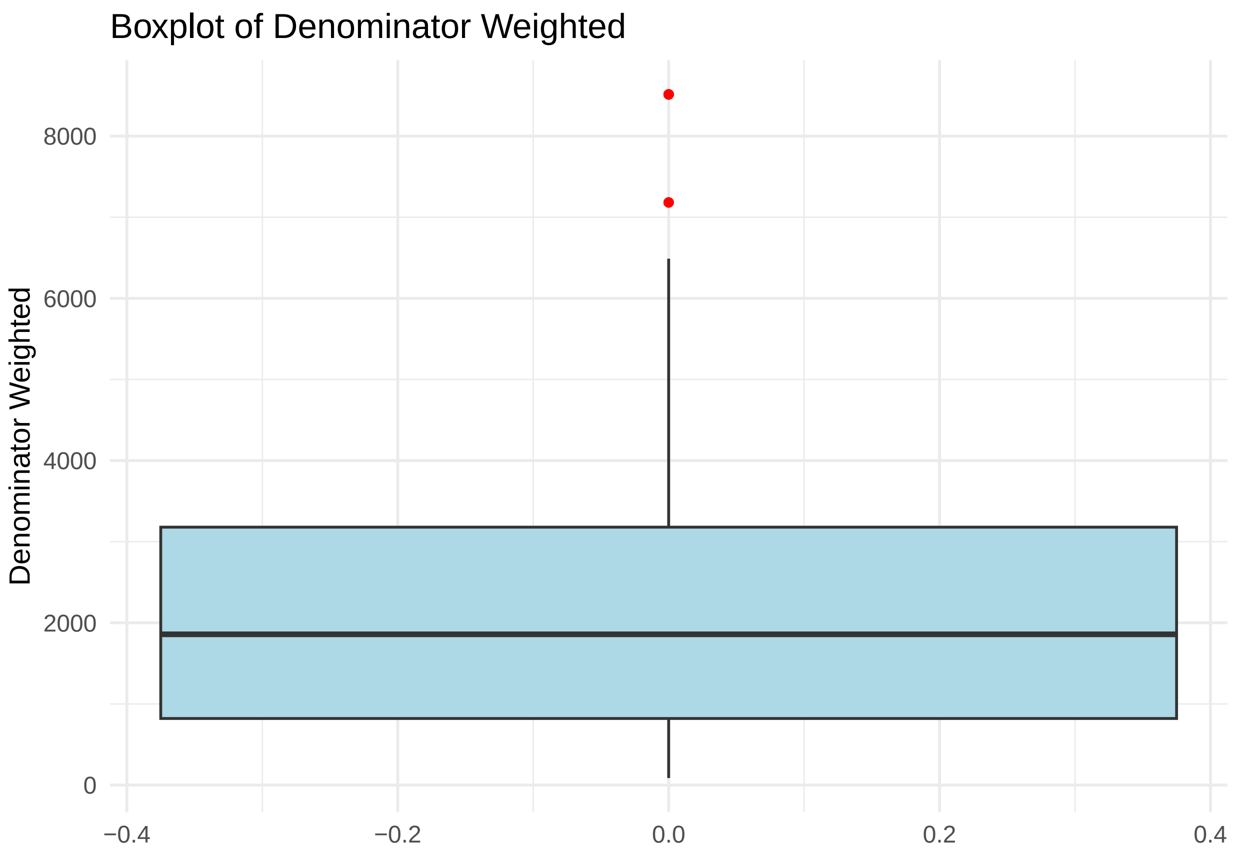
#Assumed pattern, the missing values can be filled with the previous non missing value in the opposite attribute

```r
library(dplyr)
library(tidyr)

hiv_df <- hiv_df %>%
  mutate(
    DenominatorWeighted = if_else(DataId == "796612", 2955, DenominatorUnweighted),
    DenominatorWeighted = if_else(DataId == "795149", 2343, DenominatorUnweighted),
    DenominatorWeighted = if_else(DataId == "393814", 2842, DenominatorUnweighted),
    DenominatorWeighted = if_else(DataId == "407073", 1984, DenominatorUnweighted),
    DenominatorWeighted = if_else(DataId == "253304", 1235, DenominatorUnweighted),
    DenominatorWeighted = if_else(DataId == "253300", 848, DenominatorUnweighted)
  )
```

```
hiv_df[
        c("DataId", "DenominatorWeighted", "DenominatorUnweighted")]
```





#Outlier Handling

```
# Calculate IQR boundaries
Q1_w <- quantile(hiv_df$DenominatorWeighted, 0.25, na.rm = TRUE)
Q3_w <- quantile(hiv_df$DenominatorWeighted, 0.75, na.rm = TRUE)
IQR_w <- Q3_w - Q1_w
lower_w <- Q1_w - 1.5 * IQR_w
upper_w <- Q3_w + 1.5 * IQR_w

Q1_uw <- quantile(hiv_df$DenominatorUnweighted, 0.25, na.rm = TRUE)
Q3_uw <- quantile(hiv_df$DenominatorUnweighted, 0.75, na.rm = TRUE)
IQR_uw <- Q3_uw - Q1_uw
lower_uw <- Q1_uw - 1.5 * IQR_uw
upper_uw <- Q3_uw + 1.5 * IQR_uw

# Cap values to the IQR limits
hiv_df <- hiv_df %>%
  mutate(
    DenominatorWeighted = pmin(pmax(DenominatorWeighted, lower_w), upper_w),
    DenominatorUnweighted = pmin(pmax(DenominatorUnweighted, lower_uw), upper_uw)
  )
```