CL_Water_11

Group K

2025-09-19

Display Dataset content

```
## # A tibble: 3 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    795195 Househol~ 86.3  1         ZA              South Afri~ 1998
## 3 ZAF    795196 Househol~ 38.9  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)

Convert Data Types

```
wtr_df <- wtr_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),
  )
```

Check for unique values

```
## # A tibble: 29 x 3
```

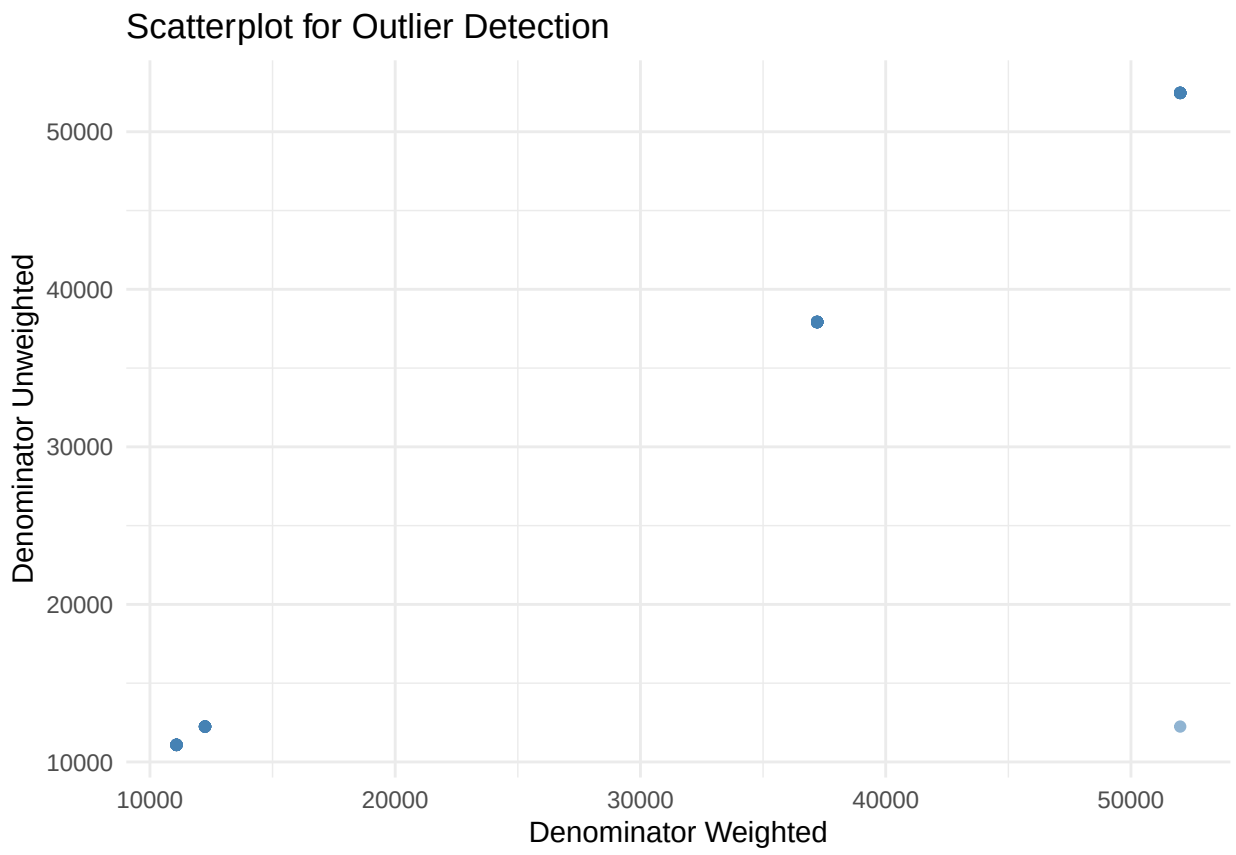
```
##    column          n_unique sample_values
##    <chr>              <int> <chr>
##  1 ISO3                   1 ZAF
##  2 DataId                96 795195, 795196, 795198
##  3 Indicator             62 Households using an improved water source, Househol~
##  4 Value                 67 86.3, 38.9, 19.5
##  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           62 WS_SRCE_H_IMP, WS_SRCE_H_PIP, WS_SRCE_H_TAP
## # i 19 more rows
```

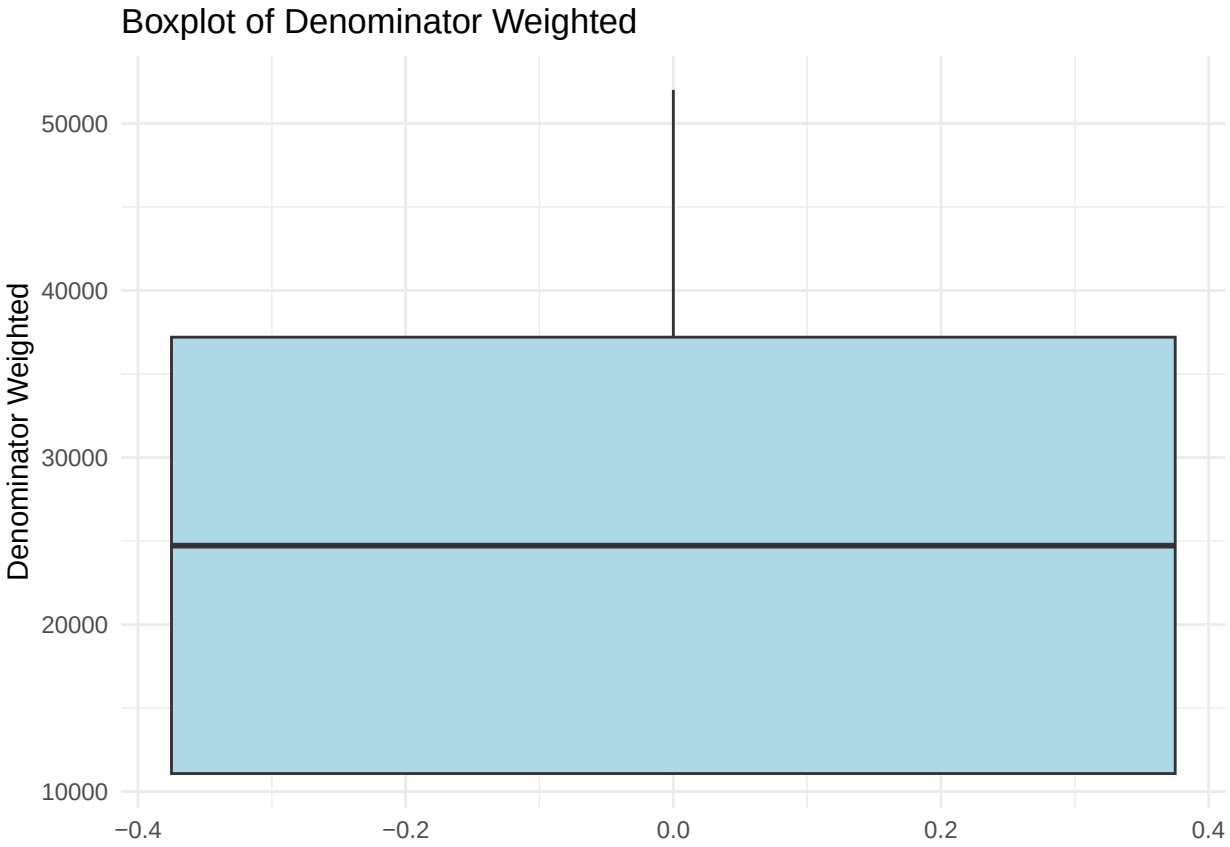
#Drop the countries only one unqiue value: reason, there is no useful information - county is also always za

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

Replace DenominatorUnweighted for a specific dataid

```
wtr_df$DenominatorUnweighted[wtr_df$DataId == "795270"] <- 12247
```




Boxplot of Denominator Weighted
Denominator Weighted
50000
40000
30000
20000
10000
−0.4
−0.2
0.0
0.2
0.4