13\_Water

#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(ggplot2)

#Load Dataset

wtr\_df <- read\_csv(here("data","raw", "water\_national\_zaf.csv"))

## Rows: 101 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.

#Display Dataset content

head(wtr\_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 795195 Househol… 86.3 1 ZA South Afri… 1998   
## 3 ZAF 795196 Househol… 38.9 1 ZA South Afri… 1998   
## 4 ZAF 795198 Househol… 19.5 1 ZA South Afri… 1998   
## 5 ZAF 795199 Househol… 3 1 ZA South Afri… 1998   
## 6 ZAF 795212 Househol… 0.7 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)

wtr\_df <- wtr\_df[-1, ]

#dimensions

dim(wtr\_df)

## [1] 100 29

#Inspect Duplicated rows

dup\_check <- wtr\_df %>%  
 group\_by(Indicator, SurveyYear, CharacteristicId, Value, DenominatorWeighted) %>%  
 filter(n() > 1)  
  
dup\_check

## # A tibble: 8 × 29  
## # Groups: Indicator, SurveyYear, CharacteristicId, Value, DenominatorWeighted  
## # [4]  
## ISO3 DataId Indicator Value Precision DHS\_CountryCode CountryName SurveyYear  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 ZAF 795213 Household… 100 1 ZA South Afri… 1998   
## 2 ZAF 795752 Populatio… 100 1 ZA South Afri… 1998   
## 3 ZAF 795205 Household… 100 1 ZA South Afri… 1998   
## 4 ZAF 795759 Populatio… 100 1 ZA South Afri… 1998   
## 5 ZAF 295330 Household… 100 1 ZA South Afri… 2016   
## 6 ZAF 414154 Populatio… 100 1 ZA South Afri… 2016   
## 7 ZAF 295325 Household… 100 1 ZA South Afri… 2016   
## 8 ZAF 414167 Populatio… 100 1 ZA South Afri… 2016   
## # 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>

wtr\_df <- wtr\_df %>%  
 distinct(Indicator, SurveyYear, CharacteristicId, Value, DenominatorWeighted, .keep\_all = TRUE)

#Percentage Missing Values

data.frame(  
 Column = names(wtr\_df),  
 Missing\_Percentage = paste0(round(colMeans(is.na(wtr\_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 4.17%  
## 26 DenominatorUnweighted 4.17%  
## 27 CILow 100%  
## 28 CIHigh 100%  
## 29 LevelRank 100%

data.frame(  
 Column = names(wtr\_df),  
 Missing\_Data = paste0(colSums(is.na(wtr\_df)))  
 )

## Column Missing\_Data  
## 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 96  
## 19 IsTotal 0  
## 20 IsPreferred 0  
## 21 SDRID 0  
## 22 RegionId 96  
## 23 SurveyYearLabel 0  
## 24 SurveyType 0  
## 25 DenominatorWeighted 4  
## 26 DenominatorUnweighted 4  
## 27 CILow 96  
## 28 CIHigh 96  
## 29 LevelRank 96

#check data types

data.frame(  
 Column = names(wtr\_df),  
 paste0(sapply(wtr\_df, typeof))  
)

## Column paste0.sapply.wtr\_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

#Check The structure of the dataset

str(wtr\_df)

## tibble [96 × 29] (S3: tbl\_df/tbl/data.frame)  
## $ ISO3 : chr [1:96] "ZAF" "ZAF" "ZAF" "ZAF" ...  
## $ DataId : chr [1:96] "795195" "795196" "795198" "795199" ...  
## $ Indicator : chr [1:96] "Households using an improved water source" "Households using water piped into dwelling" "Households using a public tap/standpipe" "Households using a tubewell/borehole" ...  
## $ Value : chr [1:96] "86.3" "38.9" "19.5" "3" ...  
## $ Precision : chr [1:96] "1" "1" "1" "1" ...  
## $ DHS\_CountryCode : chr [1:96] "ZA" "ZA" "ZA" "ZA" ...  
## $ CountryName : chr [1:96] "South Africa" "South Africa" "South Africa" "South Africa" ...  
## $ SurveyYear : chr [1:96] "1998" "1998" "1998" "1998" ...  
## $ SurveyId : chr [1:96] "ZA1998DHS" "ZA1998DHS" "ZA1998DHS" "ZA1998DHS" ...  
## $ IndicatorId : chr [1:96] "WS\_SRCE\_H\_IMP" "WS\_SRCE\_H\_PIP" "WS\_SRCE\_H\_TAP" "WS\_SRCE\_H\_TUB" ...  
## $ IndicatorOrder : num [1:96] 2.5e+08 2.5e+08 2.5e+08 2.5e+08 2.5e+08 ...  
## $ IndicatorType : chr [1:96] "I" "I" "I" "I" ...  
## $ CharacteristicId : num [1:96] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 ...  
## $ CharacteristicOrder : num [1:96] 0 0 0 0 0 0 0 0 0 0 ...  
## $ CharacteristicCategory: chr [1:96] "Total" "Total" "Total" "Total" ...  
## $ CharacteristicLabel : chr [1:96] "Total" "Total" "Total" "Total" ...  
## $ ByVariableId : chr [1:96] "0" "0" "0" "0" ...  
## $ ByVariableLabel : chr [1:96] NA NA NA NA ...  
## $ IsTotal : num [1:96] 1 1 1 1 1 1 1 1 1 1 ...  
## $ IsPreferred : num [1:96] 1 1 1 1 1 1 1 1 1 1 ...  
## $ SDRID : chr [1:96] "WSSRCEHIMP" "WSSRCEHPIP" "WSSRCEHTAP" "WSSRCEHTUB" ...  
## $ RegionId : logi [1:96] NA NA NA NA NA NA ...  
## $ SurveyYearLabel : num [1:96] 1998 1998 1998 1998 1998 ...  
## $ SurveyType : chr [1:96] "DHS" "DHS" "DHS" "DHS" ...  
## $ DenominatorWeighted : num [1:96] 12247 12247 12247 12247 12247 ...  
## $ DenominatorUnweighted : num [1:96] 12247 12247 12247 12247 12247 ...  
## $ CILow : logi [1:96] NA NA NA NA NA NA ...  
## $ CIHigh : logi [1:96] NA NA NA NA NA NA ...  
## $ LevelRank : logi [1:96] NA NA NA NA NA NA ...

#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),  
 )

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

wtr\_df <- wtr\_df %>%  
 select(  
 -ISO3,   
 -DHS\_CountryCode,   
 -CountryName,   
 -SurveyId,  
 -ByVariableId,   
 -ByVariableLabel,   
 -IsTotal,  
 -RegionId,   
 -SurveyYearLabel,   
 -SurveyType,  
 -CharacteristicOrder  
 )

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

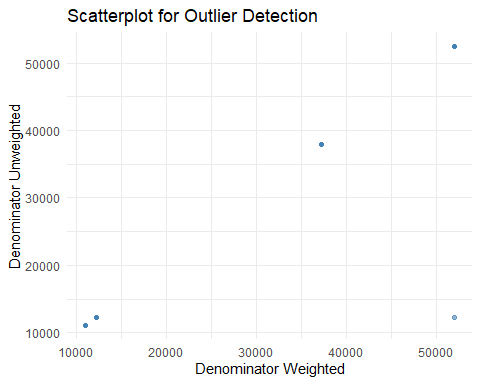
wtr\_df <- wtr\_df %>%  
 fill(DenominatorWeighted, DenominatorUnweighted, .direction = "down")  
  
wtr\_df[  
 c("DataId","DenominatorWeighted", "DenominatorUnweighted")]

## # A tibble: 96 × 3  
## DataId DenominatorWeighted DenominatorUnweighted  
## <chr> <dbl> <dbl>  
## 1 795195 12247 12247  
## 2 795196 12247 12247  
## 3 795198 12247 12247  
## 4 795199 12247 12247  
## 5 795212 12247 12247  
## 6 795201 12247 12247  
## 7 795207 12247 12247  
## 8 795211 12247 12247  
## 9 795200 12247 12247  
## 10 795202 12247 12247  
## # ℹ 86 more rows

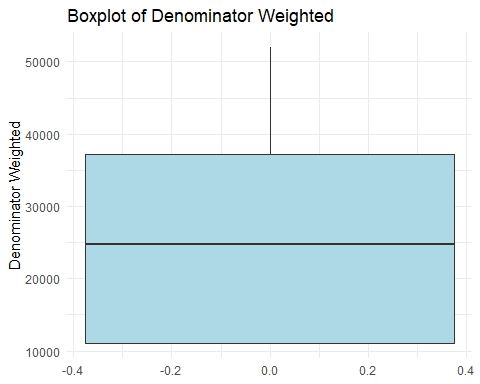
# Replace DenominatorUnweighted for a specific dataid

wtr\_df$DenominatorUnweighted[wtr\_df$DataId == "795270"] <- 12247

ggplot(wtr\_df, aes(x = DenominatorWeighted, y = DenominatorUnweighted)) +  
 geom\_point(alpha = 0.6, color = "steelblue") +  
 labs(title = "Scatterplot for Outlier Detection",  
 x = "Denominator Weighted",  
 y = "Denominator Unweighted") +  
 theme\_minimal()



ggplot(wtr\_df, aes(y = DenominatorWeighted)) +  
 geom\_boxplot(fill = "lightblue", outlier.color = "red", outlier.shape = 16) +  
 labs(title = "Boxplot of Denominator Weighted",  
 y = "Denominator Weighted") +  
 theme\_minimal()



unique(wtr\_df$Indicator)

## [1] "Households using an improved water source"   
## [2] "Households using water piped into dwelling"   
## [3] "Households using a public tap/standpipe"   
## [4] "Households using a tubewell/borehole"   
## [5] "Households using rainwater"   
## [6] "Households using tanker truck"   
## [7] "Households using bottled water"   
## [8] "Households using an unimproved water source"   
## [9] "Households using surface water"   
## [10] "Households using other water source"   
## [11] "Households with don't know or missing information on water source"   
## [12] "Households: Total"   
## [13] "Population using an improved water source"   
## [14] "Population using water piped into dwelling"   
## [15] "Population using a public tap/standpipe"   
## [16] "Population using a tubewell/borehole"   
## [17] "Population using rainwater"   
## [18] "Population using tanker truck"   
## [19] "Population using bottled water/demi john"   
## [20] "Population using an unimproved water source"   
## [21] "Population using surface water"   
## [22] "Population using other water source"   
## [23] "Population with don't know or missing information on water source"   
## [24] "Population: Total"   
## [25] "Households with water on the premises"   
## [26] "Households with water 30 minutes or less away round trip"   
## [27] "Households with water more than 30 minutes away round trip"   
## [28] "Household with unknown or missing information on round trip time to water"   
## [29] "Population with water on the premises"   
## [30] "Population with water 30 minutes or less away round trip"   
## [31] "Population with water more than 30 minutes away round trip"   
## [32] "Population with unknown or missing information on round trip time to water"  
## [33] "Number of households"   
## [34] "Number of households (unweighted)"   
## [35] "Number of persons"   
## [36] "Number of persons (unweighted)"   
## [37] "Households using a protected well"   
## [38] "Households using a protected spring"   
## [39] "Households using an unprotected well water"   
## [40] "Households using an unprotected spring"   
## [41] "Population using a protected well"   
## [42] "Population using a protected spring"   
## [43] "Population using an unprotected well water"   
## [44] "Population using an unprotected spring"   
## [45] "Households treating water by boiling"   
## [46] "Households treating water by adding bleach/chlorine"   
## [47] "Households treating water by straining through a cloth"   
## [48] "Households treating water using a ceramic, sand or other filter"   
## [49] "Households treating water using solar disinfection"   
## [50] "Households treating water using other methods"   
## [51] "Households not treating water"   
## [52] "Households with missing information on treatment of water"   
## [53] "Households using an appropriate treatment method"   
## [54] "Population treating water by boiling"   
## [55] "Population treating water by adding bleach/chlorine"   
## [56] "Population treating water by straining through a cloth"   
## [57] "Population treating water using a ceramic, sand or other filter"   
## [58] "Population treating water using solar disinfection"   
## [59] "Population treating water using other methods"   
## [60] "Population not treating water"   
## [61] "Population with missing information on treatment of water"   
## [62] "Population using an appropriate treatment method"

#save cleaned data

write\_csv(wtr\_df, here("data","processed", "water\_cleaned.csv"))