11_symptoms

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

```
ari_df <- read_csv(here("data", "raw","symptoms-of-acute-respiratory-
infection-ari_national_zaf.csv"))

## Rows: 27 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(ari_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    598577 Children… 21.9  1         ZA              South Afri… 1998
## 3 ZAF    397915 Children… 19.3  1         ZA              South Afri… 1998
## 4 ZAF    598578 Number o… 2912  0         ZA              South Afri… 1998
## 5 ZAF    384931 Number o… 4740  0         ZA              South Afri… 1998
## 6 ZAF    139860 Number o… 2958  0         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)

```
ari_df <- ari_df[-1, ]
```

#dimensions

```
dim(ari_df)

## [1] 26 29
```

#Inspect Duplicated rows

```
dup_check <- ari_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>,
…
```

#Percentage Missing Values

```
data.frame(
  Column = names(ari_df),
  Missing_Percentage = paste0(round(colMeans(is.na(ari_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                 0%
## 19                 IsTotal                 0%
## 20             IsPreferred                 0%
## 21                   SDRID                 0%
## 22                RegionId               100%
## 23         SurveyYearLabel                 0%
## 24              SurveyType                 0%
## 25     DenominatorWeighted             30.77%
## 26   DenominatorUnweighted             30.77%
## 27                   CILow               100%
## 28                  CIHigh               100%
## 29               LevelRank               100%
```

```
data.frame(
  Column = names(ari_df),
  Missing_Data = paste0(colSums(is.na(ari_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            0
## 19                  IsTotal            0
## 20              IsPreferred            0
## 21                    SDRID            0
## 22                 RegionId           26
## 23          SurveyYearLabel            0
## 24               SurveyType            0
## 25      DenominatorWeighted            8
## 26    DenominatorUnweighted            8
## 27                    CILow           26
## 28                   CIHigh           26
## 29                LevelRank           26
```

#check data types

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

```
##                      Column paste0.sapply.ari_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(ari_df)
```

```
## tibble [26 × 29] (S3: tbl_df/tbl/data.frame)
##  $ ISO3                   : chr [1:26] "ZAF" "ZAF" "ZAF" "ZAF" ...
##  $ DataId                 : chr [1:26] "598577" "397915" "598578" "384931"
...
##  $ Indicator              : chr [1:26] "Children with symptoms of ARI"
"Children with symptoms of ARI" "Number of children born in the last five (or
three) years" "Number of children born in the last five (or three) years" ...
##  $ Value                  : chr [1:26] "21.9" "19.3" "2912" "4740" ...
##  $ Precision              : chr [1:26] "1" "1" "0" "0" ...
##  $ DHS_CountryCode        : chr [1:26] "ZA" "ZA" "ZA" "ZA" ...
##  $ CountryName            : chr [1:26] "South Africa" "South Africa" "South
Africa" "South Africa" ...
##  $ SurveyYear             : chr [1:26] "1998" "1998" "1998" "1998" ...
##  $ SurveyId               : chr [1:26] "ZA1998DHS" "ZA1998DHS" "ZA1998DHS"
"ZA1998DHS" ...
##  $ IndicatorId            : chr [1:26] "CH_ARIS_C_ARI" "CH_ARIS_C_ARI"
"CH_ARIS_C_NUM" "CH_ARIS_C_NUM" ...
##  $ IndicatorOrder         : num [1:26] 9.4e+07 9.4e+07 9.4e+07 9.4e+07
9.4e+07 ...
##  $ IndicatorType          : chr [1:26] "I" "I" "D" "D" ...
##  $ CharacteristicId       : num [1:26] 1000 1000 1000 1000 1000 1000 1000
1000 1000 1000 ...
##  $ CharacteristicOrder    : num [1:26] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CharacteristicCategory: chr [1:26] "Total" "Total" "Total" "Total" ...
##  $ CharacteristicLabel    : chr [1:26] "Total" "Total" "Total" "Total" ...
##  $ ByVariableId           : chr [1:26] "14000" "14001" "14000" "14001" ...
##  $ ByVariableLabel        : chr [1:26] "Three years preceding the survey"
"Five years preceding the survey" "Three years preceding the survey" "Five
years preceding the survey" ...
##  $ IsTotal                : num [1:26] 1 1 1 1 1 1 1 1 1 1 ...
##  $ IsPreferred            : num [1:26] 0 1 0 1 0 1 0 1 0 1 ...
##  $ SDRID                  : chr [1:26] "CHARISCARI" "CHARISCARI"
```

```
"CHARISCNUM" "CHARISCNUM" ...
##  $ RegionId              : logi [1:26] NA NA NA NA NA NA ...
##  $ SurveyYearLabel       : num [1:26] 1998 1998 1998 1998 1998 ...
##  $ SurveyType            : chr [1:26] "DHS" "DHS" "DHS" "DHS" ...
##  $ DenominatorWeighted   : num [1:26] 2912 4740 NA NA 2912 ...
##  $ DenominatorUnweighted : num [1:26] 2958 4797 2958 4797 NA ...
##  $ CILow                 : logi [1:26] NA NA NA NA NA NA ...
##  $ CIHigh                : logi [1:26] NA NA NA NA NA NA ...
##  $ LevelRank             : logi [1:26] NA NA NA NA NA NA ...
```

#Convert Data Types

```
ari_df <- ari_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

```
library(dplyr)
library(purrr)

# Summary table: column name, number of unique values, sample of unique
values
n_sample <- 3

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



summary_tbl

## # A tibble: 29 × 3
##    column          n_unique sample_values
##    <chr>              <int> <chr>
##  1 ISO3                   1 ZAF
##  2 DataId                26 598577, 397915, 598578
##  3 Indicator              7 Children with symptoms of ARI, Number of
children b…
##  4 Value                 26 21.9, 19.3, 2912
##  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            7 CH_ARIS_C_ARI, CH_ARIS_C_NUM, CH_ARIS_C_UNW
## # ℹ 19 more rows
```

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

```r
ari_df <- ari_df %>%

 select(
     -ISO3,
    -DHS_CountryCode,
    -CountryName,
    -SurveyId,
    -ByVariableId,
    -ByVariableLabel,
    -IsTotal,
    -RegionId,
    -SurveyYearLabel,
    -SurveyType,
    -CharacteristicOrder
  )
```

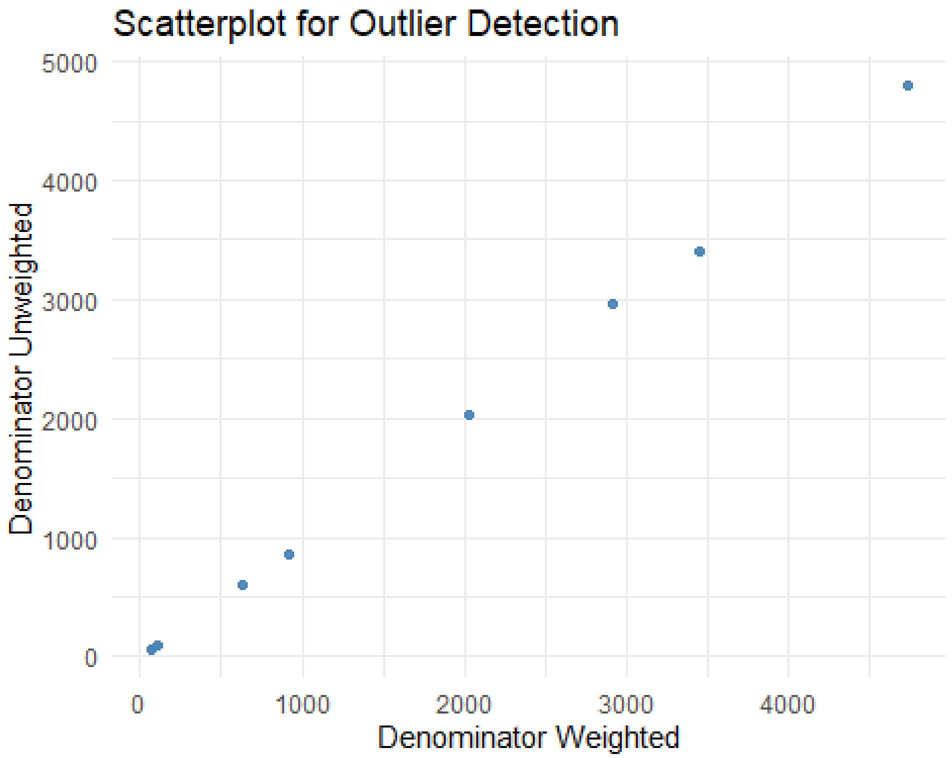
#Missing Values

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



ari_df <- ari_df %>%
   mutate(
    # 4740 <-> 4797
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 4740,
      4797,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 4797,
      4740,
      DenominatorWeighted
    ),
    # 2912 <-> 2958
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 2912,
      2958,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 2958,
      2912,
      DenominatorWeighted
    ),

    # 2025 <-> 2026
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 2025,
      2026,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 2026,
      2025,
      DenominatorWeighted
    ),

    # 3444 <-> 3413
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 3444,
      3413,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 3413,
      3444,
      DenominatorWeighted
    ),

    # 68 <-> 59
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 68,
      59,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 59,
      68,
      DenominatorWeighted
    ),

    # 107 <-> 94
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 107,
      94,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 94,
      107,
      DenominatorWeighted
    ),

    # 637 <-> 607
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 637,
      607,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 607,
      637,
      DenominatorWeighted
    ),

    # 913 <-> 862
    DenominatorUnweighted = if_else(
      is.na(DenominatorUnweighted) & DenominatorWeighted == 913,
      862,
      DenominatorUnweighted
    ),
    DenominatorWeighted = if_else(
      is.na(DenominatorWeighted) & DenominatorUnweighted == 862,
      913,
      DenominatorWeighted
    )
  )
```

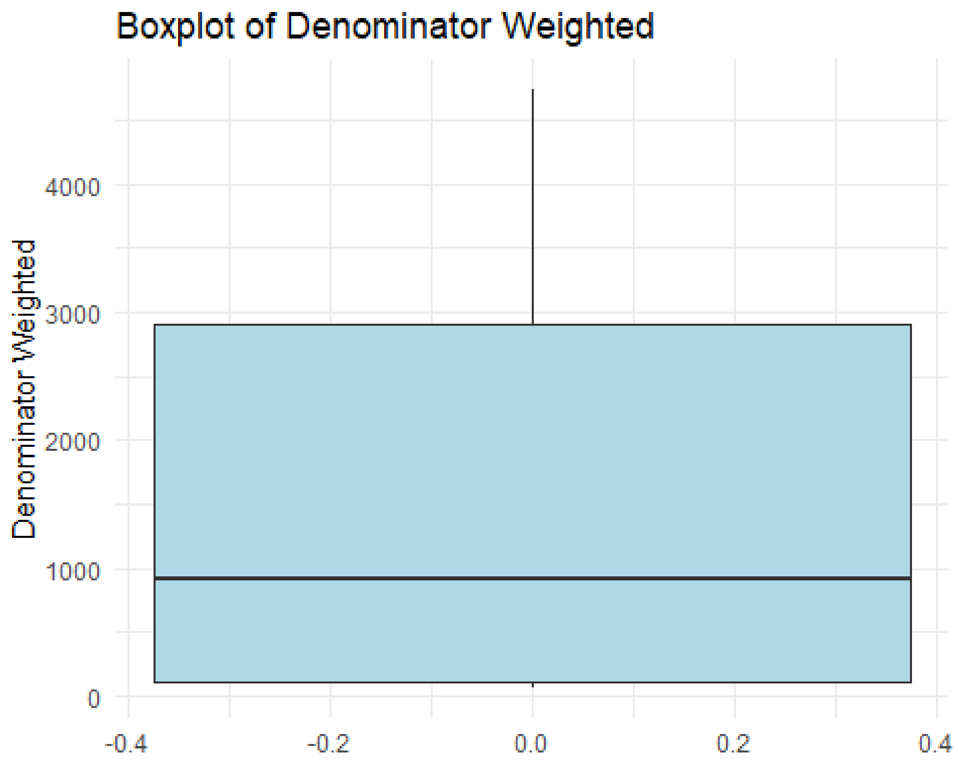
```
ari_df[
       c("DenominatorWeighted", "DenominatorUnweighted")]
```

```
## # A tibble: 26 × 2
##    DenominatorWeighted DenominatorUnweighted
##                  <dbl>                 <dbl>
##  1                2912                  2958
##  2                4740                  4797
##  3                2912                  2958
##  4                4740                  4797
##  5                2912                   2958
##  6                4740                   4797
##  7                 637                    607
##  8                 913                    862
##  9                 637                    607
## 10                 913                    862
## # ℹ 16 more rows
```

```
ggplot(ari_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(ari_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()
```



```
dim(ari_df)

## [1] 26 18
```

#save cleaned data

```
write_csv(ari_df, here("data","processed", "symptoms-of-acute-respiratory-infection-ari_cleaned.csv"))
```