EDA_IYCF_12

Group K

2025-09-09

1. Loading the dataset

```
## # A tibble: 6 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    795971 Children~ 87.4  1         ZA              South Afri~ 1998
## 3 ZAF    795973 Children~ 38.9  1         ZA              South Afri~ 1998
## 4 ZAF    621666 Children~ 6.9   1         ZA              South Afri~ 1998
## 5 ZAF    621667 Children~ 6.3   1         ZA              South Afri~ 1998
## 6 ZAF    621670 Children~ 40.9  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>
```

2. Data Overview

Summary statistics

```
##      ISO3              DataId           Indicator            Value
##  Length:23          Length:23          Length:23          Length:23
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Precision         DHS_CountryCode    CountryName         SurveyYear
##  Length:23          Length:23          Length:23          Length:23
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

```
##    SurveyId         IndicatorId        IndicatorOrder      IndicatorType
##  Length:23          Length:23          Min.   :104266010   Length:23
##  Class :character   Class :character   1st Qu.:104306022   Class :character
##  Mode  :character   Mode  :character   Median :104306080   Mode  :character
##                                        Mean   :104336967
##                                        3rd Qu.:104411025
##                                        Max.   :104436150
##                                        NA's   :1
##  CharacteristicId CharacteristicOrder CharacteristicCategory
##  Min.   :  1000   Min.   :    0       Length:23
##  1st Qu.:  1000   1st Qu.:    0       Class :character
##  Median :  1000   Median :    0       Mode  :character
##  Mean   :134637   Mean   : 9546
##  3rd Qu.:295001   3rd Qu.:21001
##  Max.   :295001   Max.   :21001
##  NA's   :1        NA's   :1
##  CharacteristicLabel ByVariableId       ByVariableLabel       IsTotal
##  Length:23           Length:23          Length:23          Min.   :1
##  Class :character    Class :character   Class :character   1st Qu.:1
##  Mode  :character    Mode  :character   Mode  :character   Median :1
##                                                            Mean   :1
##                                                            3rd Qu.:1
##                                                            Max.   :1
##                                                            NA's   :1
##   IsPreferred    SDRID           RegionId       SurveyYearLabel
##  Min.   :1    Length:23          Mode:logical   Min.   :1998
##  1st Qu.:1    Class :character   NA's:23        1st Qu.:1998
##  Median :1    Mode  :character                  Median :2016
##  Mean   :1                                      Mean   :2009
##  3rd Qu.:1                                      3rd Qu.:2016
##  Max.   :1                                      Max.   :2016
##  NA's   :1                                      NA's   :1
##   SurveyType        DenominatorWeighted DenominatorUnweighted  CILow
##  Length:23          Min.   : 345.0      Min.   : 346.0        Mode:logical
##  Class :character   1st Qu.: 363.0      1st Qu.: 380.2        NA's:23
##  Mode  :character   Median : 501.5      Median : 505.0
##                     Mean   : 750.4      Mean   : 755.7
##                     3rd Qu.: 867.0      3rd Qu.: 877.0
##                     Max.   :2010.0      Max.   :2041.0
##                     NA's   :3           NA's   :3
##   CIHigh        LevelRank
##  Mode:logical   Mode:logical
##  NA's:23        NA's:23
##
##
##
##
##
```

3. Data Quality

Missing values per column

```
##                   ISO3                 DataId              Indicator
##                      0                      0                      0
##                  Value              Precision        DHS_CountryCode
##                      0                      0                      1
##            CountryName             SurveyYear               SurveyId
##                      0                      0                      0
##            IndicatorId         IndicatorOrder          IndicatorType
##                      0                      1                      1
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                      1                      1                      1
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                      1                      0                     22
##                IsTotal            IsPreferred                  SDRID
##                      1                      1                      1
##               RegionId        SurveyYearLabel             SurveyType
##                     23                      1                      1
##    DenominatorWeighted  DenominatorUnweighted                  CILow
##                      3                      3                     23
##                 CIHigh              LevelRank
##                     23                     23
```

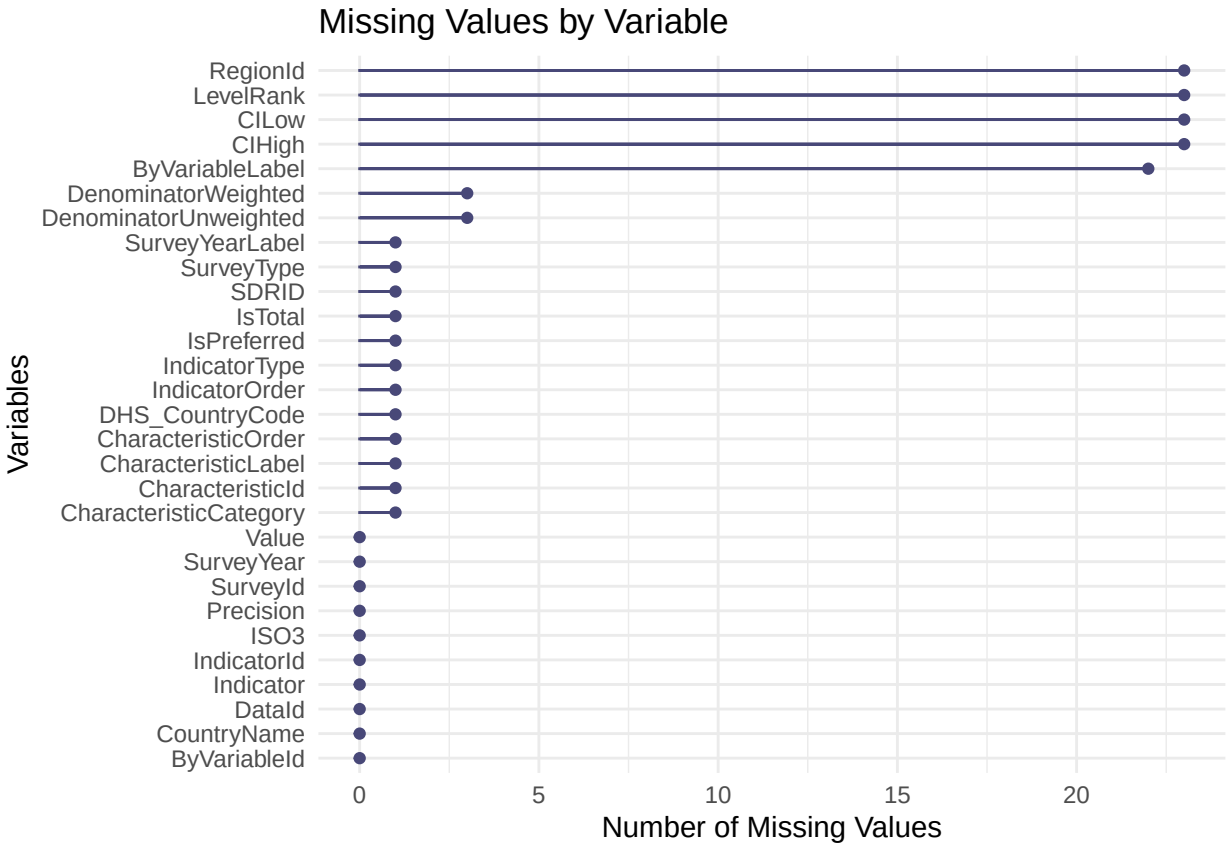
Dupilcates

```
## [1] 0
```

4. Visualizations

Plot showing which columns have missing values across the 23 rows

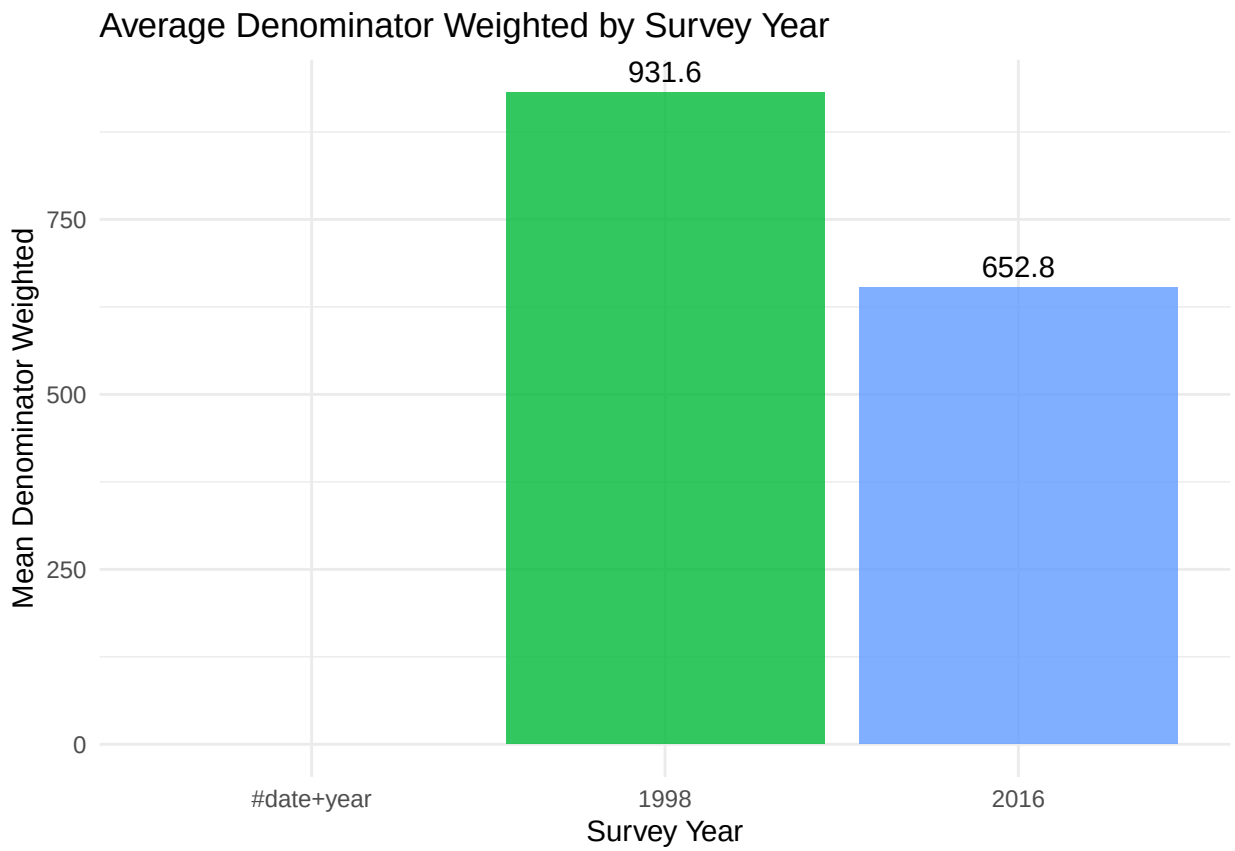
```
##
## Attaching package: 'naniar'

## The following object is masked from 'package:skimr':
##
##     n_complete
```



Bar Plot to compare overall survey size between 1998 and 2016.

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_col()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_text()`).
```




```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_col()‘).
## Removed 1 row containing missing values or values outside the scale range
## (‘geom_text()‘).
```

5. Summary Tables

```
## Warning: There were 4 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘across(...)‘.
## Caused by warning in ‘min()‘:
## ! no non-missing arguments to min; returning Inf
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 3 remaining warnings.
```

```
## # A tibble: 20 x 3
##    Variable             Statistic  Value
##    <chr>                <chr>      <dbl>
##  1 CIHigh               mean       NaN
##  2 CIHigh               sd          NA
##  3 CIHigh               min        Inf
##  4 CIHigh               max       -Inf
##  5 CILow                mean       NaN
##  6 CILow                sd          NA
##  7 CILow                min        Inf
```

```
##  8 CILow                  max       -Inf
##  9 DenominatorUnweighted mean       756.
## 10 DenominatorUnweighted sd         539.
## 11 DenominatorUnweighted min        346
## 12 DenominatorUnweighted max       2041
## 13 DenominatorWeighted   mean       750.
## 14 DenominatorWeighted   sd         534.
## 15 DenominatorWeighted   min        345
## 16 DenominatorWeighted   max       2010
## 17 Value                 mean        41.2
## 18 Value                 sd          27.1
## 19 Value                 min          6.3
## 20 Value                 max         87.4
```

6. Data dictionary

Table 1: Data Dictionary for the DHS Maternal Mortality Dataset

Column	DataType	Description
ISO3	character	ISO3 country code
DataId	character	Unique ID for the dataset row
Indicator	character	Name of the health indicator
Value	character	Numeric value of the indicator (currently character)
Precision	character	Number of decimal places or precision
DHS_CountryCode	character	Country code from DHS dataset
CountryName	character	Full country name
SurveyYear	character	Year survey was conducted
SurveyId	character	Unique ID for the survey
IndicatorId	character	DHS indicator code
IndicatorOrder	numeric	Order of indicator in dataset
IndicatorType	character	Type of indicator (I,N,D,U,C)
CharacteristicId	numeric	ID for the population characteristic
CharacteristicOrder	numeric	Order of the characteristic
CharacteristicCategory	character	Category of the population (e.g., Total 15-49)
CharacteristicLabel	character	Label describing the population characteristic
ByVariableId	character	ID for any breakdown variable
ByVariableLabel	character	Label for breakdown variable
IsTotal	numeric	1 if row represents a total, 0 otherwise
IsPreferred	numeric	1 if this is the preferred row for this indicator, 0 otherwise
SDRID	character	DHS short code for the indicator
RegionId	logical	Region identifier (mostly NA)
SurveyYearLabel	numeric	Year label for survey (redundant with SurveyYear)
SurveyType	character	Type of survey (e.g., DHS)
DenominatorWeighted	numeric	Weighted denominator used to calculate indicator
DenominatorUnweighted	numeric	Unweighted denominator used to calculate indicator
CILow	numeric	Lower bound of the confidence interval
CIHigh	numeric	Upper bound of the confidence interval
LevelRank	logical	Level rank (mostly NA; optional hierarchy indicator)