EDA_Anthropometry_National_02

Group K

2025-09-09

```
## [1] "C:/Users/Caitlin/Documents/GitHub/BIN381-Project/scripts"
```

```
## character(0)
```

Load data and show the structure

```
anz_df <- read_csv(here("raw_data", "anthropometry_national_zaf.csv"))
```

```
## Rows: 38 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.
```

#Dislay the first and last 10 rows

```
head(anz_df, 10)
```

```
## # A tibble: 10 x 29
##    ISO3  DataId Indicator Value Precision DHS_CountryCode CountryName SurveyYear
##    <chr> <chr>  <chr>     <chr> <chr>     <chr>           <chr>       <chr>
##  1 #cou~ #meta~ #indicat~ #ind~ #indicat~ <NA>            #country+n~ #date+year
##  2 ZAF   198690 Children~ 9.8   1         ZA              South Afri~ 2016
##  3 ZAF   198687 Children~ 27.4  1         ZA              South Afri~ 2016
##  4 ZAF   198688 Mean hei~ -1.1  1         ZA              South Afri~ 2016
##  5 ZAF   597227 Children~ 0.6   1         ZA              South Afri~ 2016
##  6 ZAF   597228 Children~ 2.5   1         ZA              South Afri~ 2016
##  7 ZAF   597229 Children~ 13.3  1         ZA              South Afri~ 2016
##  8 ZAF   597226 Mean wei~ 0.6   1         ZA              South Afri~ 2016
##  9 ZAF   26824  Children~ 1.1   1         ZA              South Afri~ 2016
## 10 ZAF   26823  Children~ 5.9   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>
```

```
#tail(anz_df, 10)
```

```
anz_df[-1,]
```

```
## # A tibble: 37 x 29
##    ISO3  DataId Indicator Value Precision DHS_CountryCode CountryName SurveyYear
##    <chr> <chr>  <chr>     <chr> <chr>     <chr>           <chr>       <chr>
##  1 ZAF   198690 Children~ 9.8   1         ZA              South Afri~ 2016
##  2 ZAF   198687 Children~ 27.4  1         ZA              South Afri~ 2016
##  3 ZAF   198688 Mean hei~ -1.1  1         ZA              South Afri~ 2016
##  4 ZAF   597227 Children~ 0.6   1         ZA              South Afri~ 2016
##  5 ZAF   597228 Children~ 2.5   1         ZA              South Afri~ 2016
##  6 ZAF   597229 Children~ 13.3  1         ZA              South Afri~ 2016
##  7 ZAF   597226 Mean wei~ 0.6   1         ZA              South Afri~ 2016
##  8 ZAF   26824  Children~ 1.1   1         ZA              South Afri~ 2016
##  9 ZAF   26823  Children~ 5.9   1         ZA              South Afri~ 2016
## 10 ZAF   26825  Children~ 4.5   1         ZA              South Afri~ 2016
## # i 27 more rows
## # 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>, ...
```

#Dimenison of the data

```
dim(anz_df)
```

```
## [1] 38 29
```

#Provide a summary of the datset

```
summary(anz_df)
```

```
##      ISO3              DataId           Indicator            Value
##  Length:38          Length:38          Length:38          Length:38
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Precision         DHS_CountryCode    CountryName         SurveyYear
##  Length:38          Length:38          Length:38          Length:38
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    SurveyId         IndicatorId        IndicatorOrder      IndicatorType
##  Length:38          Length:38          Min.   :104236010   Length:38
##  Class :character   Class :character   1st Qu.:104236100   Class :character
##  Mode  :character   Mode  :character   Median :114563080   Mode  :character
##                                        Mean   :111493275
##                                        3rd Qu.:114564170
##                                        Max.   :114564260
##                                        NA's   :1
##  CharacteristicId CharacteristicOrder CharacteristicCategory
##  Min.   : 1000    Min.   :    0       Length:38
##  1st Qu.: 1000    1st Qu.:    0       Class :character
##  Median : 1000    Median :    0       Mode  :character
##  Mean   : 4162    Mean   : 3514
##  3rd Qu.:10000    3rd Qu.:10000
##  Max.   :10000    Max.   :10000
##  NA's   :1        NA's   :1
##  CharacteristicLabel ByVariableId       ByVariableLabel       IsTotal
##  Length:38           Length:38          Length:38          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:38          Mode:logical   Min.   :2016
##  1st Qu.:1    Class :character   NA's:38        1st Qu.:2016
##  Median :1    Mode  :character                  Median :2016
##  Mean   :1                                      Mean   :2016
##  3rd Qu.:1                                      3rd Qu.:2016
##  Max.   :1                                      Max.   :2016
##  NA's   :1                                      NA's   :1
##   SurveyType        DenominatorWeighted DenominatorUnweighted  CILow
##  Length:38          Min.   :1384        Min.   :1449          Mode:logical
##  Class :character   1st Qu.:1416        1st Qu.:1479          NA's:38
##  Mode  :character   Median :2336        Median :2457
##                     Mean   :2337        Mean   :2433
##                     3rd Qu.:3081        3rd Qu.:3210
##                     Max.   :3272        Max.   :3405
##                     NA's   :5           NA's   :5
##   CIHigh        LevelRank
##  Mode:logical   Mode:logical
##  NA's:38        NA's:38
##
##
##
##
##
```

#Convert Value to a numeric

```
anz_df$Value <- as.numeric(anz_df$Value)
```

```
## Warning: NAs introduced by coercion
```

#Find duplicated values

```
sum(duplicated(anz_df))
```

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

#Inspect columns for missing/empty data as values and pecentages respectively

```
colSums(is.na(anz_df))
```

```
##                   ISO3                 DataId              Indicator
##                      0                      0                      0
##                  Value              Precision        DHS_CountryCode
##                      1                      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                     37
##                IsTotal            IsPreferred                  SDRID
##                      1                      1                      1
##               RegionId        SurveyYearLabel             SurveyType
##                     38                      1                      1
##    DenominatorWeighted  DenominatorUnweighted                  CILow
##                      5                      5                     38
##                 CIHigh              LevelRank
##                     38                     38
```

```
round(colMeans(is.na(anz_df)) * 100, 2)
```

```
##                   ISO3                 DataId              Indicator
##                   0.00                   0.00                   0.00
##                  Value              Precision        DHS_CountryCode
##                   2.63                   0.00                   2.63
##            CountryName             SurveyYear               SurveyId
##                   0.00                   0.00                   0.00
##            IndicatorId         IndicatorOrder          IndicatorType
##                   0.00                   2.63                   2.63
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                   2.63                   2.63                   2.63
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                   2.63                   0.00                  97.37
```

```
##                 IsTotal            IsPreferred                  SDRID
##                    2.63                   2.63                   2.63
##               RegionId       SurveyYearLabel             SurveyType
##                  100.00                   2.63                   2.63
##    DenominatorWeighted DenominatorUnweighted                  CILow
##                   13.16                  13.16                 100.00
##                 CIHigh             LevelRank
##                 100.00                 100.00
```

Check for unique values

```
sapply(anz_df, function(x) length(unique(x)))
```

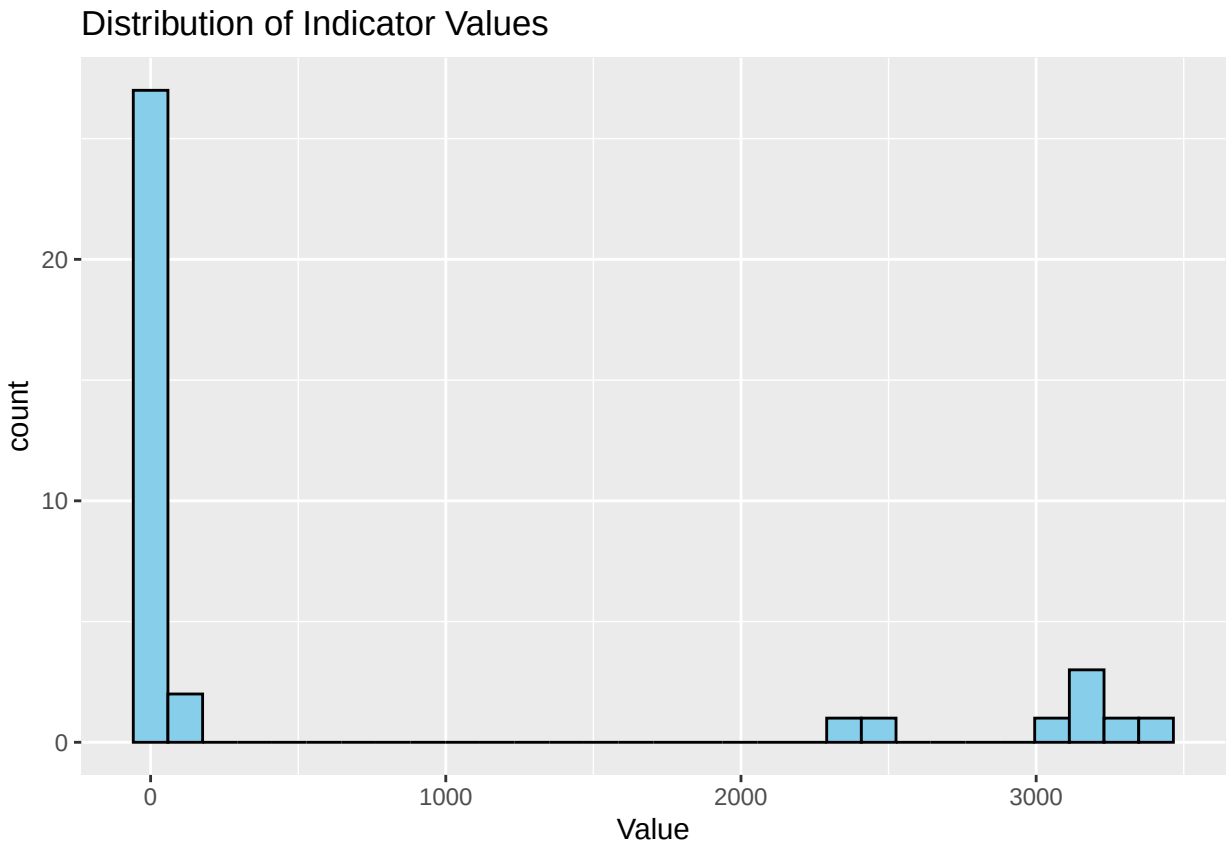
```
##                    ISO3                  DataId               Indicator
##                       2                      38                      34
##                   Value               Precision        DHS_CountryCode
##                      37                       3                       2
##             CountryName             SurveyYear                SurveyId
##                       2                       2                       2
##             IndicatorId         IndicatorOrder           IndicatorType
##                      38                      38                       4
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                       3                       3                       3
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                       3                       2                       2
##                IsTotal            IsPreferred                  SDRID
##                       2                       2                      38
##               RegionId       SurveyYearLabel             SurveyType
##                       1                       2                       2
##    DenominatorWeighted DenominatorUnweighted                  CILow
##                       8                       8                       1
##                  CIHigh             LevelRank
##                       1                       1
```

#Plots

```
#Histogram of Values

ggplot(anz_df, aes(x = Value)) +
geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Indicator Values")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```




```
ggsave("../outputs/visuals/IndicatorValuesDistribution_histo.png", width = 6, height = 4)
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```