EDA_Water_04

Group K

2025-09-09

```
## # 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    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>
```

2) Data overview

```
## Rows: 101
## Columns: 29
## $ ISO3                  <chr> "#country+code", "ZAF", "ZAF", "ZAF", "ZAF", "Z~
## $ DataId                <chr> "#meta+id", "795195", "795196", "795198", "7951~
## $ Indicator             <chr> "#indicator+name", "Households using an improve~
## $ Value                 <chr> "#indicator+value+num", "86.3", "38.9", "19.5",~
## $ Precision             <chr> "#indicator+precision", "1", "1", "1", "1", "1"~
## $ DHS_CountryCode       <chr> NA, "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "~
## $ CountryName           <chr> "#country+name", "South Africa", "South Africa"~
## $ SurveyYear            <chr> "#date+year", "1998", "1998", "1998", "1998", "~
## $ SurveyId              <chr> "#survey+id", "ZA1998DHS", "ZA1998DHS", "ZA1998~
## $ IndicatorId           <chr> "#indicator+code", "WS_SRCE_H_IMP", "WS_SRCE_H_~
## $ IndicatorOrder        <dbl> NA, 250161010, 250161020, 250161030, 250161040,~
## $ IndicatorType         <chr> NA, "I", "I", "I", "I", "I", "I", "I", "I", "I"~
## $ CharacteristicId      <dbl> NA, 1000, 1000, 1000, 1000, 1000, 1000, 1000, 1~
## $ CharacteristicOrder   <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ CharacteristicCategory <chr> NA, "Total", "Total", "Total", "Total", "Total"~
## $ CharacteristicLabel   <chr> NA, "Total", "Total", "Total", "Total", "Total"~
## $ ByVariableId          <chr> "#indicator+label+code", "0", "0", "0", "0", "0~
## $ ByVariableLabel       <chr> "#indicator+label", NA, NA, NA, NA, NA, NA, NA,~
## $ IsTotal               <dbl> NA, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ IsPreferred           <dbl> NA, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ SDRID                 <chr> NA, "WSSRCEHIMP", "WSSRCEHPIP", "WSSRCEHTAP", "~
```

```
## $ RegionId               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ SurveyYearLabel        <dbl> NA, 1998, 1998, 1998, 1998, 1998, 1998, 1998, 1~
## $ SurveyType             <chr> NA, "DHS", "DHS", "DHS", "DHS", "DHS", "DHS", "~
## $ DenominatorWeighted    <dbl> NA, 12247, 12247, 12247, 12247, 12247, 12247, 1~
## $ DenominatorUnweighted  <dbl> NA, 12247, 12247, 12247, 12247, 12247, 12247, 1~
## $ CILow                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ CIHigh                 <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ LevelRank              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
##      ISO3               DataId           Indicator            Value
##  Length:101         Length:101         Length:101         Length:101
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Precision         DHS_CountryCode    CountryName         SurveyYear
##  Length:101         Length:101         Length:101         Length:101
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    SurveyId          IndicatorId        IndicatorOrder      IndicatorType
##  Length:101         Length:101         Min.   :250161010   Length:101
##  Class :character   Class :character   1st Qu.:250161158   Class :character
##  Mode  :character   Mode  :character   Median :250162140   Mode  :character
##                                        Mean   :250187767
##                                        3rd Qu.:250231012
##                                        Max.   :250232110
##                                        NA's   :1
##  CharacteristicId CharacteristicOrder CharacteristicCategory
##  Min.   :1000     Min.   :0           Length:101
##  1st Qu.:1000     1st Qu.:0           Class :character
##  Median :1000     Median :0           Mode  :character
##  Mean   :1000     Mean   :0
##  3rd Qu.:1000     3rd Qu.:0
##  Max.   :1000     Max.   :0
##  NA's   :1        NA's   :1
##  CharacteristicLabel ByVariableId       ByVariableLabel       IsTotal
##  Length:101          Length:101         Length:101         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:101         Mode:logical   Min.   :1998
##  1st Qu.:1    Class :character   NA's:101       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:101         Min.   :11083       Min.   :11083         Mode:logical
##  Class :character   1st Qu.:11083       1st Qu.:11083         NA's:101
##  Mode  :character   Median :24726       Median :25086
##                     Mean   :27138       Mean   :27449
##                     3rd Qu.:37205       3rd Qu.:37925
##                     Max.   :52007       Max.   :52465
##                     NA's   :5           NA's   :5
##   CIHigh        LevelRank
##  Mode:logical   Mode:logical
##  NA's:101       NA's:101
##
##
##
##
##
```

Table 1: Data summary

Name	waterDataset
Number of rows	101
Number of columns	29
Column type frequency:	
character	17
logical	4
numeric	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ISO3	0	1.00	3	13	0	2	0
DataId	0	1.00	6	8	0	101	0
Indicator	0	1.00	15	74	0	63	0
Value	0	1.00	1	20	0	68	0
Precision	0	1.00	1	20	0	3	0
DHS_CountryCode	1	0.99	2	2	0	1	0
CountryName	0	1.00	12	13	0	2	0
SurveyYear	0	1.00	4	10	0	3	0
SurveyId	0	1.00	9	10	0	3	0
IndicatorId	0	1.00	13	15	0	65	0
IndicatorType	1	0.99	1	1	0	5	0
CharacteristicCategory	1	0.99	5	5	0	1	0
CharacteristicLabel	1	0.99	5	5	0	1	0
ByVariableId	0	1.00	1	21	0	2	0
ByVariableLabel	100	0.01	16	16	0	1	0
SDRID	1	0.99	10	10	0	64	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
SurveyType	1	0.99	3	3	0	1	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
RegionId	101	0	NaN	:
CILow	101	0	NaN	:
CIHigh	101	0	NaN	:
LevelRank	101	0	NaN	:

Variable type: numeric

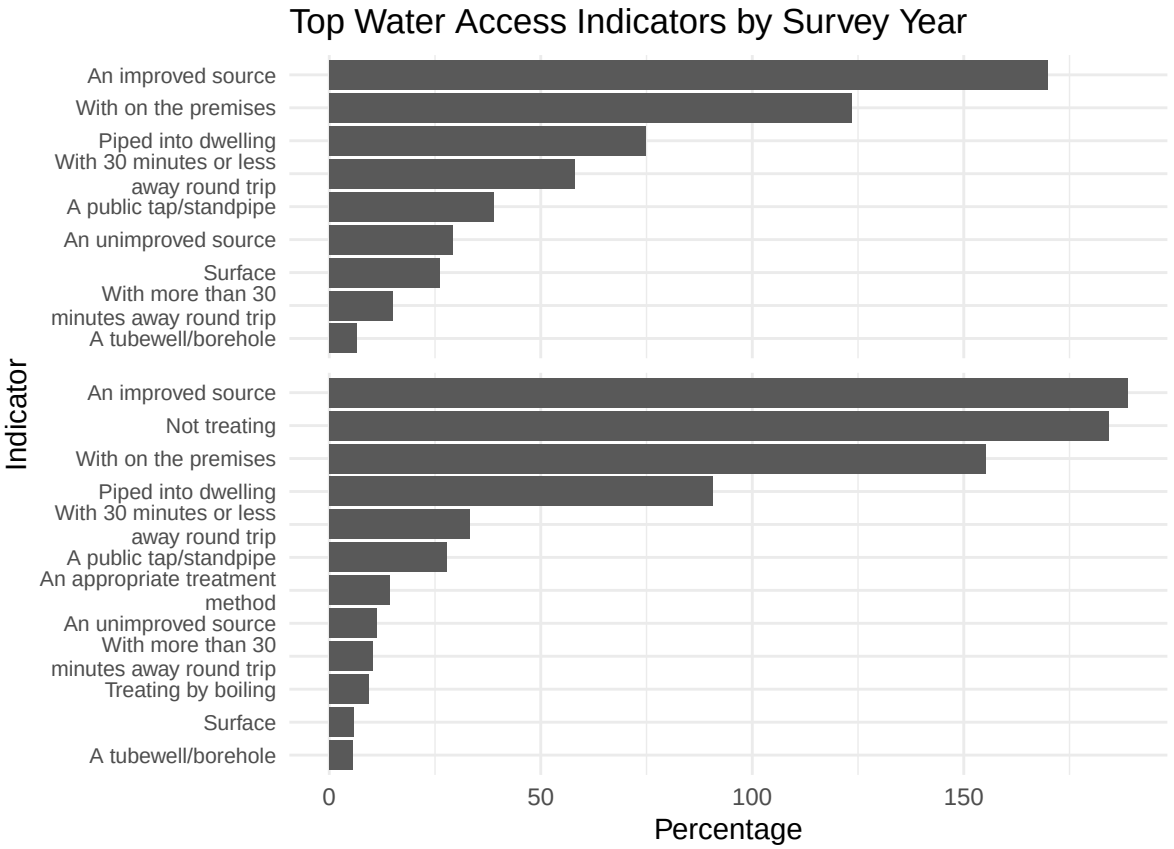
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
IndicatorOrder	1	0.99	250187766.56	30272.17	250161010	250161158	250162140	250231012	250232110	
CharacteristicId	1	0.99	1000.00	0.00	1000	1000	1000	1000	1000	
CharacteristicOrder	1	0.99	0.00	0.00	0	0	0	0	0	
IsTotal	1	0.99	1.00	0.00	1	1	1	1	1	
IsPreferred	1	0.99	1.00	0.00	1	1	1	1	1	
SurveyYearLabel	1	0.99	2009.16	8.78	1998	1998	2016	2016	2016	
DenominatorWeighted	5	0.95	27137.62	16510.60	11083	11083	24726	37205	52007	
DenominatorUnweighted	5	0.95	27448.50	16780.94	11083	11083	25086	37925	52465	

3) Data quality checks

```
##                  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                   100
##               IsTotal           IsPreferred                 SDRID
##                     1                     1                     1
##              RegionId       SurveyYearLabel            SurveyType
##                   101                     1                     1
##    DenominatorWeighted DenominatorUnweighted                 CILow
##                     5                     5                   101
##                CIHigh             LevelRank
##                   101                   101
```

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

```
## # A tibble: 1 x 2
##   out_of_bounds pct_rows_violating
##           <int>              <dbl>
## 1             8                  8
```



4) Visualizations —
5) Summary tables

Table 5: Summary Statistics for Key Numeric Variables in the Water Dataset

Variable	Statistic	Value
CIHigh	mean	NaN
CIHigh	sd	NA
CIHigh	min	Inf
CIHigh	max	-Inf
CILow	mean	NaN
CILow	sd	NA
CILow	min	Inf
CILow	max	-Inf
DenominatorUnweighted	mean	27448.500
DenominatorUnweighted	sd	16780.940
DenominatorUnweighted	min	11083.000
DenominatorUnweighted	max	52465.000
DenominatorWeighted	mean	27137.625
DenominatorWeighted	sd	16510.599
DenominatorWeighted	min	11083.000

Variable	Statistic	Value
DenominatorWeighted	max	52007.000
Value	mean	2283.723
Value	sd	9158.584
Value	min	0.000
Value	max	52465.000