# 06_HIV_Behavior

## HIV Behavior - National South Africa

### Load Libraries

```
# Data manipulation
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(readr)
library(here)

## here() starts at C:/Users/morul/School/3rd
Year/BIN381/BIN381_PROJECT/BIN381_PROJECT

library(purrr)

# Visualization and summaries
library(ggplot2)
library(skimr)
library(visdat)
```

### Load Data

```
# Load the HIV behavior dataset
hiv_df <- read_csv(
  here("data", "raw", "hiv-behavior_national_zaf.csv"),
  col_names = TRUE,   # use first row as column names
  col_types = cols()  # suppress guessing messages
)

# Step 2: Remove first row if it contains metadata
hiv_df <- hiv_df[-1, ]

# Step 3: Reset row names
rownames(hiv_df) <- NULL
```

```
cat("HIV behavior dataset loaded successfully.\n")
```

```
## HIV behavior dataset loaded successfully.
```

```
cat("Dimensions:", dim(hiv_df), "\n")
```

```
## Dimensions: 118 29
```

## Initial Assessment

```
# Quick glimpse
glimpse(hiv_df)
```

```
## Rows: 118
## Columns: 29
## $ ISO3                   <chr> "ZAF", "ZAF", "ZAF", "ZAF", "ZAF", "ZAF",
"ZAF"…
## $ DataId                 <chr> "795160", "795161", "796612", "795358",
"795240…
## $ Indicator              <chr> "Sex before the age of 15 [Women]", "Number
of …
## $ Value                  <chr> "8", "4324", "4459", "54.5", "2955",
"2993", "4…
## $ Precision              <chr> "1", "0", "0", "1", "0", "0", "1", "1",
"0", "0…
## $ DHS_CountryCode        <chr> "ZA", "ZA", "ZA", "ZA", "ZA", "ZA", "ZA",
"ZA",…
## $ CountryName            <chr> "South Africa", "South Africa", "South
Africa",…
## $ SurveyYear             <chr> "1998", "1998", "1998", "1998", "1998",
"1998",…
## $ SurveyId               <chr> "ZA1998DHS", "ZA1998DHS", "ZA1998DHS",
"ZA1998D…
## $ IndicatorId            <chr> "HA_AFSY_W_A15", "HA_AFSY_W_NM1",
"HA_AFSY_W_UN…
## $ IndicatorOrder         <dbl> 135763010, 135763020, 135763030, 135763040,
135…
## $ IndicatorType          <chr> "I", "D", "U", "I", "D", "U", "I", "I",
"D", "U…
## $ CharacteristicId       <dbl> 1000, 1000, 1000, 1000, 1000, 1000, 1000,
1000,…
## $ CharacteristicOrder    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ CharacteristicCategory <chr> "Total", "Total", "Total", "Total",
"Total", "T…
## $ CharacteristicLabel    <chr> "Total", "Total", "Total", "Total",
"Total", "T…
## $ ByVariableId           <chr> "0", "0", "0", "0", "0", "0", "0", "0",
"0", "0…
## $ ByVariableLabel        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ IsTotal                <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…
## $ IsPreferred            <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…
## $ SDRID                  <chr> "HAAFSYWA15", "HAAFSYWNM1", "HAAFSYWUN1",
"HAAF…
## $ RegionId               <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
NA,…
## $ SurveyYearLabel        <dbl> 1998, 1998, 1998, 1998, 1998, 1998, 1998,
1998,…
## $ SurveyType             <chr> "DHS", "DHS", "DHS", "DHS", "DHS", "DHS",
"DHS"…
## $ DenominatorWeighted    <dbl> 4324, NA, 55, 2955, NA, NA, 3721, 3721, NA,
372…
## $ DenominatorUnweighted  <dbl> 4459, 4459, NA, 2993, 2993, NA, 3857, 3857,
385…
## $ 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,…

# Summary of missingness
skim(hiv_df)
```

*Data summary*

| | |
|---|---|
| Name | hiv_df |
| Number of rows | 118 |
| 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 | 3 | 3 | 0 | 1 | 0 |
| DataId | 0 | 1 | 4 | 6 | 0 | 118 | 0 |

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---:|---:|---:|---:|---:|---:|---:|
| Indicator | 0 | 1 | 13 | 105 | 0 | 77 | 0 |
| Value | 0 | 1 | 1 | 4 | 0 | 99 | 0 |
| Precision | 0 | 1 | 1 | 1 | 0 | 2 | 0 |
| DHS_CountryCode | 0 | 1 | 2 | 2 | 0 | 1 | 0 |
| CountryName | 0 | 1 | 12 | 12 | 0 | 1 | 0 |
| SurveyYear | 0 | 1 | 4 | 4 | 0 | 2 | 0 |
| SurveyId | 0 | 1 | 9 | 9 | 0 | 2 | 0 |
| IndicatorId | 0 | 1 | 13 | 13 | 0 | 101 | 0 |
| IndicatorType | 0 | 1 | 1 | 1 | 0 | 3 | 0 |
| CharacteristicCategory | 0 | 1 | 5 | 11 | 0 | 2 | 0 |
| CharacteristicLabel | 0 | 1 | 5 | 11 | 0 | 2 | 0 |
| ByVariableId | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| ByVariableLabel | 118 | 0 | NA | NA | 0 | 0 | 0 |
| SDRID | 0 | 1 | 10 | 10 | 0 | 101 | 0 |
| SurveyType | 0 | 1 | 3 | 3 | 0 | 1 | 0 |

**Variable type: logical**

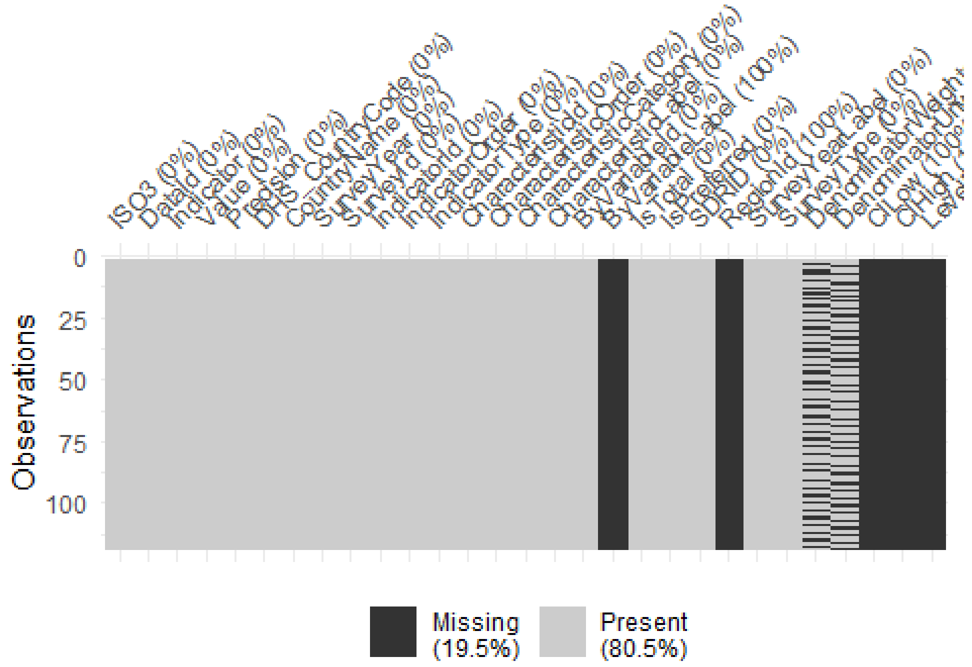
| skim_variable | n_missing | complete_rate | mean | count |
|---|---:|---:|---:|---|
| RegionId | 118 | 0 | NaN | : |
| CILow | 118 | 0 | NaN | : |
| CIHigh | 118 | 0 | NaN | : |
| LevelRank | 118 | 0 | NaN | : |

**Variable type: numeric**

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---:|---:|---:|---:|---:|---:|---:|---:|---:|---|
| IndicatorOrder | 0 | 1.00 | 135657340.85 | 177821.00 | 135403010 | 135451388 | 135763045 | 135804128 | 135864060 | ▆▃▁▃█ |
| CharacteristicId | 0 | 1.00 | 4889.83 | 4477.45 | 1000 | 1000 | 1000 | 10000 | 10000 | █▁ |

| skim_variable | n_missing | complete_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | ▁▁▇ |
| CharacteristicOrder | 0 | 1.00 | 4322.03 | 4974.95 | 0 | 0 | 0 | 10000 | 10000 | ▇▁▁▁▇ |
| IsTotal | 0 | 1.00 | 1.00 | 0.00 | 1 | 1 | 1 | 1 | 1 | ▁▁▇▁▁ |
| IsPreferred | 0 | 1.00 | 1.00 | 0.00 | 1 | 1 | 1 | 1 | 1 | ▁▁▇▁▁ |
| SurveyYearLabel | 0 | 1.00 | 2013.41 | 6.35 | 1998 | 2016 | 2016 | 2016 | 2016 | ▂▁▁▁▇ |
| DenominatorWeighted | 39 | 0.67 | 2380.66 | 2320.80 | 15 | 544 | 1787 | 3202 | 8514 | ▇▇▂▁▂ |
| DenominatorUnweighted | 38 | 0.68 | 2566.29 | 2232.04 | 86 | 871 | 1995 | 3179 | 8514 | ▇▇▂▁▂ |

```r
# Visualize missing values
vis_miss(hiv_df)
```



## Handle Duplicates

```
# Check for exact duplicates
cat("Exact duplicates:", sum(duplicated(hiv_df)), "\n")

## Exact duplicates: 0

# Remove exact duplicates
hiv_df <- hiv_df %>% distinct()
cat("Dimensions after duplicate removal:", dim(hiv_df), "\n")

## Dimensions after duplicate removal: 118 29
```

## Covert Data Types

```
# Convert numeric columns safely
num_cols <- c("value", "precision", "denominator_weighted",
"denominator_unweighted",
              "ci_low", "ci_high", "survey_year", "indicator_order",
              "characteristic_id", "characteristic_order",
"survey_year_label")
num_cols <- num_cols[num_cols %in% colnames(hiv_df)] # only existing columns

hiv_df <- hiv_df %>%
  mutate(across(all_of(num_cols), as.numeric))

# Logical columns
```

```
logical_cols <- c("is_total", "is_preferred")
logical_cols <- logical_cols[logical_cols %in% colnames(hiv_df)]
hiv_df <- hiv_df %>% mutate(across(all_of(logical_cols),
~as.logical(as.integer(.))))

# Check structure
str(hiv_df)

## tibble [118 × 29] (S3: tbl_df/tbl/data.frame)
##  $ ISO3                   : chr [1:118] "ZAF" "ZAF" "ZAF" "ZAF" ...
##  $ DataId                 : chr [1:118] "795160" "795161" "796612" "795358"
...
##  $ Indicator              : chr [1:118] "Sex before the age of 15 [Women]"
"Number of young women" "Number of young women (unweighted)" "Sex before the
age of 18 [Women]" ...
##  $ Value                  : chr [1:118] "8" "4324" "4459" "54.5" ...
##  $ Precision              : chr [1:118] "1" "0" "0" "1" ...
##  $ DHS_CountryCode        : chr [1:118] "ZA" "ZA" "ZA" "ZA" ...
##  $ CountryName            : chr [1:118] "South Africa" "South Africa"
"South Africa" "South Africa" ...
##  $ SurveyYear             : chr [1:118] "1998" "1998" "1998" "1998" ...
##  $ SurveyId               : chr [1:118] "ZA1998DHS" "ZA1998DHS" "ZA1998DHS"
"ZA1998DHS" ...
##  $ IndicatorId            : chr [1:118] "HA_AFSY_W_A15" "HA_AFSY_W_NM1"
"HA_AFSY_W_UN1" "HA_AFSY_W_A18" ...
##  $ IndicatorOrder         : num [1:118] 1.36e+08 1.36e+08 1.36e+08 1.36e+08
1.36e+08 ...
##  $ IndicatorType          : chr [1:118] "I" "D" "U" "I" ...
##  $ CharacteristicId       : num [1:118] 1000 1000 1000 1000 1000 1000 1000
1000 1000 1000 ...
##  $ CharacteristicOrder    : num [1:118] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CharacteristicCategory: chr [1:118] "Total" "Total" "Total" "Total" ...
##  $ CharacteristicLabel    : chr [1:118] "Total" "Total" "Total" "Total" ...
##  $ ByVariableId           : chr [1:118] "0" "0" "0" "0" ...
##  $ ByVariableLabel        : chr [1:118] NA NA NA NA ...
##  $ IsTotal                : num [1:118] 1 1 1 1 1 1 1 1 1 1 ...
##  $ IsPreferred            : num [1:118] 1 1 1 1 1 1 1 1 1 1 ...
##  $ SDRID                  : chr [1:118] "HAAFSYWA15" "HAAFSYWNM1"
"HAAFSYWUN1" "HAAFSYWA18" ...
##  $ RegionId               : logi [1:118] NA NA NA NA NA NA ...
##  $ SurveyYearLabel        : num [1:118] 1998 1998 1998 1998 1998 ...
##  $ SurveyType             : chr [1:118] "DHS" "DHS" "DHS" "DHS" ...
##  $ DenominatorWeighted    : num [1:118] 4324 NA 55 2955 NA ...
##  $ DenominatorUnweighted : num [1:118] 4459 4459 NA 2993 2993 ...
##  $ CILow                  : logi [1:118] NA NA NA NA NA NA ...
##  $ CIHigh                 : logi [1:118] NA NA NA NA NA NA ...
##  $ LevelRank              : logi [1:118] NA NA NA NA NA NA ...
```

- Numeric columns (value, precision, denominator_weighted, denominator_unweighted, ci_low, ci_high, survey_year, indicator_order, characteristic_id, characteristic_order, survey_year_label) were converted to numeric.

- Logical columns (is_total, is_preferred) were converted to boolean values.
- Conversion ensures accurate calculations and proper visualization. ## Handle Missing Values

```
# Impute survey_year_label with survey_year if missing
if ("survey_year_label" %in% colnames(hiv_df)) {
  hiv_df <- hiv_df %>%
    mutate(survey_year_label = ifelse(is.na(survey_year_label), survey_year,
survey_year_label))
}

# Impute survey_type with "Unknown" if missing
if ("survey_type" %in% colnames(hiv_df)) {
  hiv_df <- hiv_df %>%
    mutate(survey_type = ifelse(is.na(survey_type), "Unknown", survey_type))
}

# Recalculate missing summary
missing_summary <- data.frame(
  Column = colnames(hiv_df),
  n_missing = colSums(is.na(hiv_df)),
  total_rows = nrow(hiv_df),
  missing_percent = round(colSums(is.na(hiv_df))/nrow(hiv_df)*100, 2)
)

# Impute denominators with median of available values
hiv_df <- hiv_df %>%
  mutate(
    DenominatorWeighted = ifelse(is.na(DenominatorWeighted),
                                  median(DenominatorWeighted, na.rm = TRUE),
                                  DenominatorWeighted),
    DenominatorUnweighted = ifelse(is.na(DenominatorUnweighted),
                                    median(DenominatorUnweighted, na.rm =
TRUE),
                                    DenominatorUnweighted)
  )

# Function to calculate mode
get_mode <- function(x) {
  ux <- unique(x[!is.na(x)])
  ux[which.max(tabulate(match(x, ux)))]
}

# Impute missing values with most frequent value
hiv_df <- hiv_df %>%
  mutate(
    DHS_CountryCode = ifelse(is.na(DHS_CountryCode),
get_mode(DHS_CountryCode), DHS_CountryCode),
    IndicatorOrder = ifelse(is.na(IndicatorOrder), get_mode(IndicatorOrder),
IndicatorOrder),
    IndicatorType = ifelse(is.na(IndicatorType), get_mode(IndicatorType),
IndicatorType),
    CharacteristicId = ifelse(is.na(CharacteristicId),
get_mode(CharacteristicId), CharacteristicId),
    CharacteristicOrder = ifelse(is.na(CharacteristicOrder),
get_mode(CharacteristicOrder), CharacteristicOrder),
    CharacteristicCategory = ifelse(is.na(CharacteristicCategory),
get_mode(CharacteristicCategory), CharacteristicCategory),
    CharacteristicLabel = ifelse(is.na(CharacteristicLabel),
get_mode(CharacteristicLabel), CharacteristicLabel),
    IsTotal = ifelse(is.na(IsTotal), get_mode(IsTotal), IsTotal),
    IsPreferred = ifelse(is.na(IsPreferred), get_mode(IsPreferred),
IsPreferred),
    SDRID = ifelse(is.na(SDRID), get_mode(SDRID), SDRID),
    SurveyYearLabel = ifelse(is.na(SurveyYearLabel),
get_mode(SurveyYearLabel), SurveyYearLabel),
    SurveyType = ifelse(is.na(SurveyType), get_mode(SurveyType), SurveyType)
  )
# Drop columns that are 100% missing
cols_to_drop <- c("ByVariableLabel", "RegionId", "CILow", "CIHigh",
"LevelRank")
cols_to_drop <- intersect(cols_to_drop, colnames(hiv_df))  # only if they
exist

hiv_df <- hiv_df %>% select(-all_of(cols_to_drop))
cat("Dropped completely missing columns:\n")
```

```
## Dropped completely missing columns:
```

```
print(cols_to_drop)
```

```
## [1] "ByVariableLabel" "RegionId"        "CILow"           "CIHigh"
## [5] "LevelRank"
```

```
# Verify that missing values are handled
colSums(is.na(hiv_df))
```

```
##                   ISO3                 DataId              Indicator
##                      0                      0                      0
##                  Value              Precision        DHS_CountryCode
##                      0                      0                      0
##            CountryName             SurveyYear               SurveyId
##                      0                      0                      0
##            IndicatorId         IndicatorOrder          IndicatorType
##                      0                      0                      0
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                      0                      0                      0
##    CharacteristicLabel           ByVariableId                IsTotal
##                      0                      0                      0
##            IsPreferred                  SDRID        SurveyYearLabel
##                      0                      0                      0
##             SurveyType    DenominatorWeighted  DenominatorUnweighted
##                      0                      0                      0
```

Handling Missing Values

Strategies applied:

1. Survey Year Label: Filled missing survey_year_label with survey_year.
2. Survey Type: Filled missing survey_type with “Unknown”.
3. Denominator columns: Filled with median of available values.
4. Categorical columns: Filled missing values with the mode (most frequent value).
5. Dropped columns that were 100% missing (ByVariableLabel, RegionId, CILow, CIHigh, LevelRank).
6. Outcome: No missing values remain, ensuring the dataset is analysis-ready. ## Handle Outliers

```
# Winsorize HIV Behavior 'Value' at 1st and 99th percentiles

# First, check the structure and type of Value column
cat("Structure of Value column:\n")

## Structure of Value column:

str(hiv_df$Value)

##  chr [1:118] "8" "4324" "4459" "54.5" "2955" "2993" "40.3" "48.7" "3721"
...

cat("\nClass of Value column:", class(hiv_df$Value), "\n")

##
## Class of Value column: character

cat("First few values:", head(hiv_df$Value), "\n")

## First few values: 8 4324 4459 54.5 2955 2993

# Check for any non-numeric values
cat("\nNon-numeric values in Value column:\n")

##
## Non-numeric values in Value column:
```

```r
print(hiv_df$Value[!is.na(hiv_df$Value) & !is.numeric(hiv_df$Value)])
```

```
##   [1] "8"    "4324" "4459" "54.5" "2955" "2993" "40.3" "48.7" "3721"
"3857"
##  [11] "21.8" "1811" "1858" "4324" "4459" "2343" "2390" "57.6" "6586"
"6489"
##  [21] "60"   "3793" "3866" "68.7" "2603" "2532" "68.4" "1787" "1799" "4.5"
##  [31] "8514" "8514" "57.6" "387"  "394"  "3.9"  "7205" "7182" "17"
"3202"
##  [41] "3179" "65.3" "544"  "535"  "14.7" "2488" "2467" "1.5"  "3.1"
"8514"
##  [51] "8514" "68.1" "387"  "394"  "4.7"  "12.1" "3202" "3179" "71"   "544"
##  [61] "535"  "4.7"  "2.9"  "3202" "3179" "83.1" "92"   "86"   "6.1"
"2842"
##  [71] "2913" "50.3" "1984" "1995" "14.6" "1235" "1307" "66.2" "848"  "888"
##  [81] "37.4" "57.1" "2508" "2621" "62.7" "1431" "1471" "30.7" "62.5"
"1191"
##  [91] "1268" "75.9" "744"  "783"  "4.6"  "2842" "2913" "7.5"  "1757"
"1754"
## [101] "61.4" "132"  "153"  "20.7" "1235" "1307" "32.4" "788"  "820"
"72.9"
## [111] "256"  "244"  "5.9"  "575"  "1153" "0.1"  "287"  "308"
```

```r
# Convert to numeric if necessary (handling any character values)
hiv_df$Value <- as.numeric(as.character(hiv_df$Value))

# Check for NAs introduced by conversion
cat("\nNA values after conversion:", sum(is.na(hiv_df$Value)), "\n")
```

```
##
## NA values after conversion: 0
```

```r
# Now proceed with winsorization
lower_val <- quantile(hiv_df$Value, 0.01, na.rm = TRUE)
upper_val <- quantile(hiv_df$Value, 0.99, na.rm = TRUE)

cat("\n1st percentile (lower bound):", lower_val, "\n")
```

```
##
## 1st percentile (lower bound): 1.738
```

```r
cat("99th percentile (upper bound):", upper_val, "\n")
```

```
## 99th percentile (upper bound): 8514
```

```r
hiv_df <- hiv_df %>%
  mutate(
    Value = pmax(pmin(Value, upper_val), lower_val)
  )

# Create log transformation (using log1p to handle zeros)
```

```r
hiv_df <- hiv_df %>%
  mutate(Value_log = log1p(Value))

# Check summary
cat("\nSummary of Value after winsorization:\n")
```

```
##
## Summary of Value after winsorization:
```

```r
summary(hiv_df$Value)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.738   60.350  659.500 1605.543 2585.250 8514.000
```

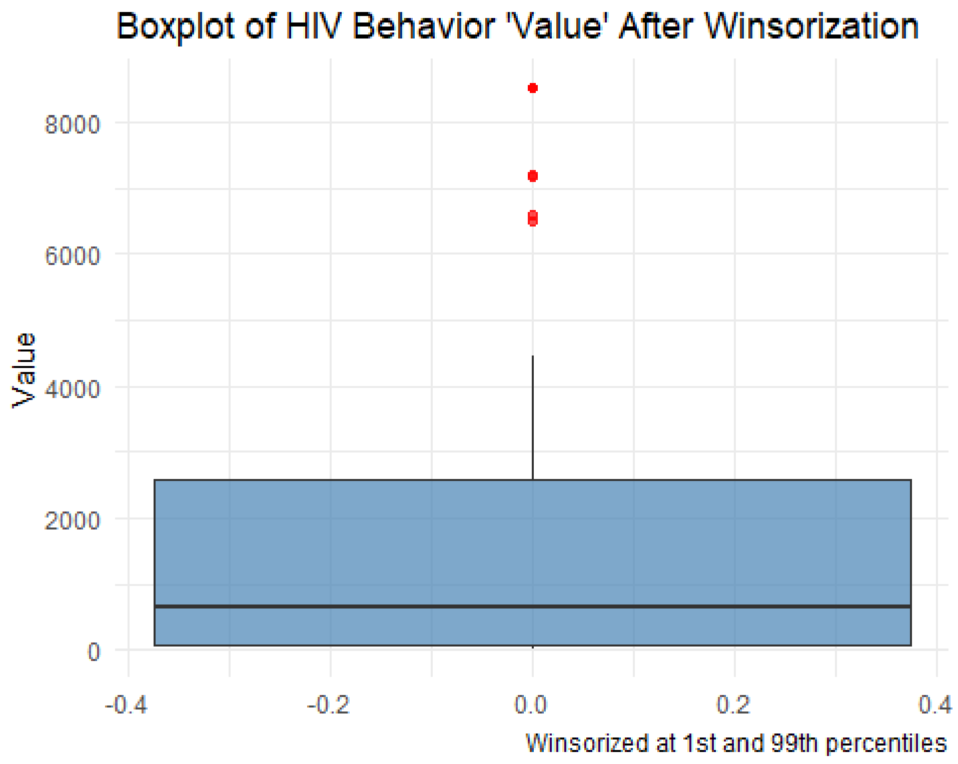
```r
cat("\nSummary of log-transformed Value:\n")
```

```
##
## Summary of log-transformed Value:
```

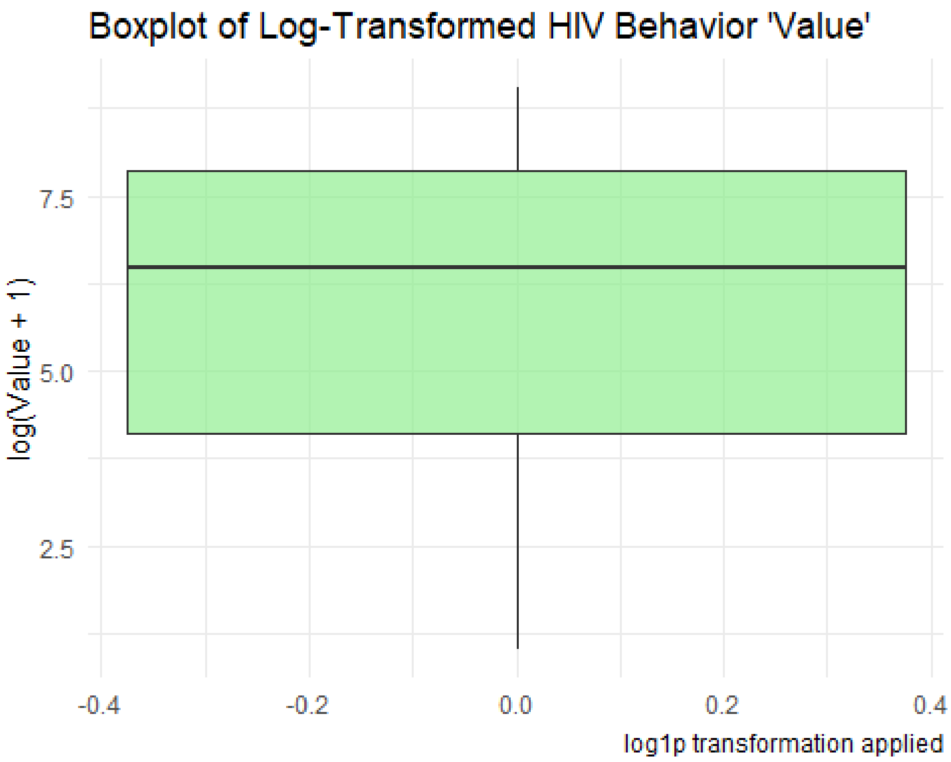
```r
summary(hiv_df$Value_log)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.007   4.117   6.485   5.852   7.858   9.050
```

```r
# Visualize with boxplot
ggplot(hiv_df, aes(y = Value)) +
  geom_boxplot(fill = "steelblue", outlier.color = "red", alpha = 0.7) +
  labs(
    title = "Boxplot of HIV Behavior 'Value' After Winsorization",
    y = "Value",
    caption = "Winsorized at 1st and 99th percentiles"
  ) +
  theme_minimal()
```



```r
# Additional visualization for log-transformed values
ggplot(hiv_df, aes(y = Value_log)) +
  geom_boxplot(fill = "lightgreen", outlier.color = "red", alpha = 0.7) +
  labs(
    title = "Boxplot of Log-Transformed HIV Behavior 'Value'",
    y = "log(Value + 1)",
    caption = "log1p transformation applied"
  ) +
  theme_minimal()
```



## Handling Outliers

- The numeric column Value was Winsorized at the 1st and 99th percentiles to reduce the influence of extreme values.
- Log transformation (log1p) was applied to Value to normalize distributions and handle zero values.
- Boxplots were created to visualize both the Winsorized and log-transformed values.

```
# Quick check of structure and summary
str(hiv_df)

## tibble [118 × 25] (S3: tbl_df/tbl/data.frame)
##  $ ISO3                      : chr [1:118] "ZAF" "ZAF" "ZAF" "ZAF" ...
##  $ DataId                    : chr [1:118] "795160" "795161" "796612" "795358"
...
##  $ Indicator                 : chr [1:118] "Sex before the age of 15 [Women]"
"Number of young women" "Number of young women (unweighted)" "Sex before the
age of 18 [Women]" ...
##  $ Value                     : num [1:118] 8 4324 4459 54.5 2955 ...
##  $ Precision                 : chr [1:118] "1" "0" "0" "1" ...
##  $ DHS_CountryCode           : chr [1:118] "ZA" "ZA" "ZA" "ZA" ...
##  $ CountryName               : chr [1:118] "South Africa" "South Africa"
"South Africa" "South Africa" ...
##  $ SurveyYear                : chr [1:118] "1998" "1998" "1998" "1998" ...
##  $ SurveyId                  : chr [1:118] "ZA1998DHS" "ZA1998DHS" "ZA1998DHS"
"ZA1998DHS" ...
```

```
##  $ IndicatorId           : chr [1:118] "HA_AFSY_W_A15" "HA_AFSY_W_NM1"
"HA_AFSY_W_UN1" "HA_AFSY_W_A18" ...
##  $ IndicatorOrder        : num [1:118] 1.36e+08 1.36e+08 1.36e+08 1.36e+08
1.36e+08 ...
##  $ IndicatorType         : chr [1:118] "I" "D" "U" "I" ...
##  $ CharacteristicId      : num [1:118] 1000 1000 1000 1000 1000 1000 1000
1000 1000 1000 ...
##  $ CharacteristicOrder   : num [1:118] 0 0 0 0 0 0 0 0 0 0 ...
##  $ CharacteristicCategory: chr [1:118] "Total" "Total" "Total" "Total" ...
##  $ CharacteristicLabel   : chr [1:118] "Total" "Total" "Total" "Total" ...
##  $ ByVariableId          : chr [1:118] "0" "0" "0" "0" ...
##  $ IsTotal               : num [1:118] 1 1 1 1 1 1 1 1 1 1 ...
##  $ IsPreferred           : num [1:118] 1 1 1 1 1 1 1 1 1 1 ...
##  $ SDRID                 : chr [1:118] "HAAFSYWA15" "HAAFSYWNM1"
"HAAFSYWUN1" "HAAFSYWA18" ...
##  $ SurveyYearLabel       : num [1:118] 1998 1998 1998 1998 1998 ...
##  $ SurveyType            : chr [1:118] "DHS" "DHS" "DHS" "DHS" ...
##  $ DenominatorWeighted   : num [1:118] 4324 1787 55 2955 1787 ...
##  $ DenominatorUnweighted : num [1:118] 4459 4459 1995 2993 2993 ...
##  $ Value_log             : num [1:118] 2.2 8.37 8.4 4.02 7.99 ...
```

```
summary(hiv_df)
```

```
##      ISO3              DataId           Indicator             Value         
##  Length:118         Length:118         Length:118         Min.   :   1.738  
##  Class :character   Class :character   Class :character   1st Qu.:  60.350  
##  Mode  :character   Mode  :character   Mode  :character   Median : 659.500  
##                                                           Mean   :1605.543  
##                                                           3rd Qu.:2585.250  
##                                                           Max.   :8514.000  
##   Precision         DHS_CountryCode    CountryName         SurveyYear       
##  Length:118         Length:118         Length:118         Length:118        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##    SurveyId         IndicatorId        IndicatorOrder      IndicatorType     
##  Length:118         Length:118         Min.   :135403010   Length:118        
##  Class :character   Class :character   1st Qu.:135451388   Class 
:character  
##  Mode  :character   Mode  :character   Median :135763045   Mode  
:character  
##                                        Mean   :135657341                     
##                                        3rd Qu.:135804128                     
##                                        Max.   :135864060                     
##  CharacteristicId CharacteristicOrder CharacteristicCategory
##  Min.   : 1000    Min.   :    0       Length:118            
##  1st Qu.: 1000    1st Qu.:    0       Class :character      
##  Median : 1000    Median :    0       Mode  :character      
##  Mean   : 4890    Mean   : 4322
##  3rd Qu.:10000    3rd Qu.:10000
##  Max.   :10000    Max.   :10000
##  CharacteristicLabel ByVariableId          IsTotal   IsPreferred
##  Length:118          Length:118         Min.   :1   Min.   :1
##  Class :character    Class :character   1st Qu.:1   1st Qu.:1
##  Mode  :character    Mode  :character   Median :1   Median :1
##                                         Mean   :1   Mean   :1
##                                         3rd Qu.:1   3rd Qu.:1
##                                         Max.   :1   Max.   :1
##     SDRID           SurveyYearLabel  SurveyType        DenominatorWeighted
##  Length:118         Min.   :1998    Length:118         Min.   :  15
##  Class :character   1st Qu.:2016    Class :character   1st Qu.:1191
##  Mode  :character   Median :2016    Mode  :character   Median :1787
##                     Mean   :2013                       Mean   :2184
##                     3rd Qu.:2016                       3rd Qu.:2579
##                     Max.   :2016                       Max.   :8514
##  DenominatorUnweighted   Value_log
##  Min.   :  86          Min.   :1.007
##  1st Qu.:1307          1st Qu.:4.117
##  Median :1995          Median :6.485
##  Mean   :2382          Mean   :5.852
##  3rd Qu.:2913          3rd Qu.:7.858
##  Max.   :8514          Max.   :9.050
```

```r
# Check for any remaining NAs
colSums(is.na(hiv_df))
```

```
##                   ISO3                 DataId              Indicator
##                      0                      0                      0
##                  Value              Precision        DHS_CountryCode
##                      0                      0                      0
##            CountryName             SurveyYear               SurveyId
##                      0                      0                      0
##            IndicatorId         IndicatorOrder          IndicatorType
##                      0                      0                      0
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                      0                      0                      0
##    CharacteristicLabel           ByVariableId                IsTotal
##                      0                      0                      0
##            IsPreferred                  SDRID        SurveyYearLabel
##                      0                      0                      0
##             SurveyType    DenominatorWeighted  DenominatorUnweighted
##                      0                      0                      0
##              Value_log
##                      0
```

```r
# Save cleaned dataset
write_csv(hiv_df, here("data", "processed", "hiv-
behavior_national_zaf_clean.csv"))
```

```r
cat("HIV behavior dataset cleaned and saved to data/processed/hiv-behavior_national_zaf_clean.csv\n")
```

```
## HIV behavior dataset cleaned and saved to data/processed/hiv-behavior_national_zaf_clean.csv
```