BIN371_m1

2025-08-08

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
df1 <- read.csv("anthropometry_national_zaf.csv")
df2 <- read.csv("literacy_national_zaf.csv")
df3 <- read.csv("maternal-mortality_national_zaf.csv")
df4 <- read.csv("symptoms-of-acute-respiratory-infection-ari_national_zaf.csv")

df_list <- list(df1, df2, df3, df4)
```

```
check_duplicates <- function(data, show = TRUE) {
  duplicate_rows <- data[duplicated(data), ]
  if (show) {
    if (nrow(duplicate_rows) > 0) {
      cat("Duplicates found:\n")
      print(duplicate_rows)
    } else {
      cat("No duplicate rows found.\n")
    }
  }
  invisible(duplicate_rows)
}

check_empty_values <- function(data) {
  na_counts <- sapply(data, function(x) sum(is.na(x)))
  empty_counts <- sapply(data, function(x) sum(x == "", na.rm = TRUE))
  string_na_counts <- sapply(data, function(x) sum(x == "NA", na.rm = TRUE))
  result <- data.frame(
    Column = names(data),
    NA_Count = na_counts,
    Empty_String_Count = empty_counts,
    "Text_'NA'_Count" = string_na_counts
  )
  result <- result[rowSums(result[,-1]) > 0, ]
  if (nrow(result) == 0) {
    cat("No missing or empty values found.\n")
  } else {
    cat("Columns with missing/empty values:\n")
    print(result)
```

```
  }
  invisible(result)
}
```

```
cat("Dataframe Column Names\n")
```

```
## Dataframe Column Names
```

```
for(i in seq_along(df_list)){
  cat("Dataframe", i, "\n")
  print(names(df_list[[i]]))
  cat("\n")
}
```

```
## Dataframe 1
##  [1] "ISO3"                  "DataId"                "Indicator"
##  [4] "Value"                 "Precision"             "DHS_CountryCode"
##  [7] "CountryName"           "SurveyYear"            "SurveyId"
## [10] "IndicatorId"           "IndicatorOrder"        "IndicatorType"
## [13] "CharacteristicId"      "CharacteristicOrder"   "CharacteristicCategory"
## [16] "CharacteristicLabel"   "ByVariableId"          "ByVariableLabel"
## [19] "IsTotal"               "IsPreferred"           "SDRID"
## [22] "RegionId"              "SurveyYearLabel"       "SurveyType"
## [25] "DenominatorWeighted"   "DenominatorUnweighted" "CILow"
## [28] "CIHigh"                "LevelRank"
##
## Dataframe 2
##  [1] "ISO3"                  "DataId"                "Indicator"
##  [4] "Value"                 "Precision"             "DHS_CountryCode"
##  [7] "CountryName"           "SurveyYear"            "SurveyId"
## [10] "IndicatorId"           "IndicatorOrder"        "IndicatorType"
## [13] "CharacteristicId"      "CharacteristicOrder"   "CharacteristicCategory"
## [16] "CharacteristicLabel"   "ByVariableId"          "ByVariableLabel"
## [19] "IsTotal"               "IsPreferred"           "SDRID"
## [22] "RegionId"              "SurveyYearLabel"       "SurveyType"
## [25] "DenominatorWeighted"   "DenominatorUnweighted" "CILow"
## [28] "CIHigh"                "LevelRank"
##
## Dataframe 3
##  [1] "ISO3"                  "DataId"                "Indicator"
##  [4] "Value"                 "Precision"             "DHS_CountryCode"
##  [7] "CountryName"           "SurveyYear"            "SurveyId"
## [10] "IndicatorId"           "IndicatorOrder"        "IndicatorType"
## [13] "CharacteristicId"      "CharacteristicOrder"   "CharacteristicCategory"
## [16] "CharacteristicLabel"   "ByVariableId"          "ByVariableLabel"
## [19] "IsTotal"               "IsPreferred"           "SDRID"
## [22] "RegionId"              "SurveyYearLabel"       "SurveyType"
## [25] "DenominatorWeighted"   "DenominatorUnweighted" "CILow"
## [28] "CIHigh"                "LevelRank"
##
## Dataframe 4
##  [1] "ISO3"                  "DataId"                "Indicator"
```

```
##  [4] "Value"                  "Precision"              "DHS_CountryCode"
##  [7] "CountryName"            "SurveyYear"             "SurveyId"
## [10] "IndicatorId"            "IndicatorOrder"         "IndicatorType"
## [13] "CharacteristicId"       "CharacteristicOrder"    "CharacteristicCategory"
## [16] "CharacteristicLabel"    "ByVariableId"           "ByVariableLabel"
## [19] "IsTotal"                "IsPreferred"            "SDRID"
## [22] "RegionId"               "SurveyYearLabel"        "SurveyType"
## [25] "DenominatorWeighted"    "DenominatorUnweighted"  "CILow"
## [28] "CIHigh"                 "LevelRank"
```

```
cat("Dataframe Duplication Tests\n")
```

```
## Dataframe Duplication Tests
```

```
for(i in seq_along(df_list)){
  cat("Dataframe", i, "\n")
  check_duplicates(df_list[[i]])
  cat("\n")
}
```

```
## Dataframe 1
## No duplicate rows found.
##
## Dataframe 2
## No duplicate rows found.
##
## Dataframe 3
## No duplicate rows found.
##
## Dataframe 4
## No duplicate rows found.
```

```
cat("Dataframe Empty Values Check\n")
```

```
## Dataframe Empty Values Check
```

```
for(i in seq_along(df_list)){
  cat("Dataframe", i, "\n")
  check_empty_values(df_list[[i]])
  cat("\n")
}
```

```
## Dataframe 1
## Columns with missing/empty values:
##                                        Column NA_Count Empty_String_Count
## DHS_CountryCode               DHS_CountryCode        0                  1
## IndicatorOrder                 IndicatorOrder        1                  0
## IndicatorType                   IndicatorType        0                  1
## CharacteristicId             CharacteristicId        1                  0
## CharacteristicOrder       CharacteristicOrder        1                  0
## CharacteristicCategory CharacteristicCategory        0                  1
```

```
## CharacteristicLabel       CharacteristicLabel        0                  1
## ByVariableLabel               ByVariableLabel        0                 37
## IsTotal                               IsTotal        1                  0
## IsPreferred                       IsPreferred        1                  0
## SDRID                                   SDRID        0                  1
## RegionId                             RegionId       38                  0
## SurveyYearLabel               SurveyYearLabel        1                  0
## SurveyType                         SurveyType        0                  1
## DenominatorWeighted       DenominatorWeighted        5                  0
## DenominatorUnweighted   DenominatorUnweighted        5                  0
## CILow                                 CILow       38                  0
## CIHigh                               CIHigh       38                  0
## LevelRank                         LevelRank       38                  0
##                       Text_.NA._Count
## DHS_CountryCode                     0
## IndicatorOrder                      0
## IndicatorType                       0
## CharacteristicId                    0
## CharacteristicOrder                 0
## CharacteristicCategory              0
## CharacteristicLabel                 0
## ByVariableLabel                     0
## IsTotal                             0
## IsPreferred                         0
## SDRID                               0
## RegionId                            0
## SurveyYearLabel                     0
## SurveyType                          0
## DenominatorWeighted                 0
## DenominatorUnweighted               0
## CILow                               0
## CIHigh                              0
## LevelRank                           0
##
## Dataframe 2
## Columns with missing/empty values:
##                                     Column NA_Count Empty_String_Count
## DHS_CountryCode               DHS_CountryCode        0                  1
## IndicatorOrder                 IndicatorOrder        1                  0
## IndicatorType                   IndicatorType        0                  1
## CharacteristicId             CharacteristicId        1                  0
## CharacteristicOrder       CharacteristicOrder        1                  0
## CharacteristicCategory CharacteristicCategory        0                  1
## CharacteristicLabel       CharacteristicLabel        0                  1
## ByVariableLabel               ByVariableLabel        0                 20
## IsTotal                               IsTotal        1                  0
## IsPreferred                       IsPreferred        1                  0
## SDRID                                   SDRID        0                  1
## RegionId                             RegionId       21                  0
## SurveyYearLabel               SurveyYearLabel        1                  0
## SurveyType                         SurveyType        0                  1
## DenominatorWeighted       DenominatorWeighted        3                  0
## DenominatorUnweighted   DenominatorUnweighted        3                  0
## CILow                                 CILow       21                  0
```

```
## CIHigh                                CIHigh       21                  0
## LevelRank                          LevelRank       21                  0
##                         Text_.NA._Count
## DHS_CountryCode                       0
## IndicatorOrder                        0
## IndicatorType                         0
## CharacteristicId                      0
## CharacteristicOrder                   0
## CharacteristicCategory                0
## CharacteristicLabel                   0
## ByVariableLabel                       0
## IsTotal                               0
## IsPreferred                           0
## SDRID                                 0
## RegionId                              0
## SurveyYearLabel                       0
## SurveyType                            0
## DenominatorWeighted                   0
## DenominatorUnweighted                 0
## CILow                                 0
## CIHigh                                0
## LevelRank                             0
##
## Dataframe 3
## Columns with missing/empty values:
##                                      Column NA_Count Empty_String_Count
## DHS_CountryCode               DHS_CountryCode        0                  1
## IndicatorOrder                 IndicatorOrder        1                  0
## IndicatorType                   IndicatorType        0                  1
## CharacteristicId             CharacteristicId        1                  0
## CharacteristicOrder       CharacteristicOrder        1                  0
## CharacteristicCategory CharacteristicCategory        0                  1
## CharacteristicLabel       CharacteristicLabel        0                  1
## ByVariableLabel               ByVariableLabel        0                 21
## IsTotal                               IsTotal        1                  0
## IsPreferred                       IsPreferred        1                  0
## SDRID                                   SDRID        0                  1
## RegionId                             RegionId       22                  0
## SurveyYearLabel               SurveyYearLabel        1                  0
## SurveyType                         SurveyType        0                  1
## DenominatorWeighted       DenominatorWeighted       20                  0
## DenominatorUnweighted   DenominatorUnweighted       16                  0
## CILow                                 CILow       19                  0
## CIHigh                               CIHigh       19                  0
## LevelRank                         LevelRank       22                  0
##                         Text_.NA._Count
## DHS_CountryCode                       0
## IndicatorOrder                        0
## IndicatorType                         0
## CharacteristicId                      0
## CharacteristicOrder                   0
## CharacteristicCategory                0
## CharacteristicLabel                   0
## ByVariableLabel                       0
```

```
## IsTotal                                0
## IsPreferred                            0
## SDRID                                  0
## RegionId                               0
## SurveyYearLabel                        0
## SurveyType                             0
## DenominatorWeighted                    0
## DenominatorUnweighted                  0
## CILow                                  0
## CIHigh                                 0
## LevelRank                              0
##
## Dataframe 4
## Columns with missing/empty values:
##                                       Column NA_Count Empty_String_Count
## DHS_CountryCode               DHS_CountryCode        0                   1
## IndicatorOrder                 IndicatorOrder        1                   0
## IndicatorType                   IndicatorType        0                   1
## CharacteristicId             CharacteristicId        1                   0
## CharacteristicOrder       CharacteristicOrder        1                   0
## CharacteristicCategory CharacteristicCategory        0                   1
## CharacteristicLabel       CharacteristicLabel        0                   1
## IsTotal                               IsTotal        1                   0
## IsPreferred                       IsPreferred        1                   0
## SDRID                                 SDRID        0                   1
## RegionId                           RegionId       27                   0
## SurveyYearLabel                SurveyYearLabel        1                   0
## SurveyType                         SurveyType        0                   1
## DenominatorWeighted       DenominatorWeighted        9                   0
## DenominatorUnweighted   DenominatorUnweighted        9                   0
## CILow                                 CILow       27                   0
## CIHigh                               CIHigh       27                   0
## LevelRank                         LevelRank       27                   0
##                         Text_.NA._Count
## DHS_CountryCode                       0
## IndicatorOrder                        0
## IndicatorType                         0
## CharacteristicId                      0
## CharacteristicOrder                   0
## CharacteristicCategory                0
## CharacteristicLabel                   0
## IsTotal                               0
## IsPreferred                           0
## SDRID                                 0
## RegionId                              0
## SurveyYearLabel                       0
## SurveyType                            0
## DenominatorWeighted                   0
## DenominatorUnweighted                 0
## CILow                                 0
## CIHigh                                0
## LevelRank                             0
```

```r
clean_df <- function(df){
  names(df) <- gsub("\\s+", "_", names(df))
  names(df) <- gsub("[^A-Za-z0-9_]", "", names(df))
  if ("Value" %in% names(df)) {
    df$Value_num <- readr::parse_number(as.character(df$Value))
  }
  df
}

df_list <- lapply(df_list, clean_df)

df_names <- c("Anthropometry","Literacy","MaternalMortality","ARI_Symptoms")
names(df_list) <- df_names

for (nm in names(df_list)) {
  df <- df_list[[nm]]
  if ("Value" %in% names(df)) {
    bad <- suppressWarnings(sum(is.na(as.numeric(df$Value)) & !is.na(df$Value)))
    cat(nm, "- non-numeric Value entries initially:", bad, "\n")
  }
}
```

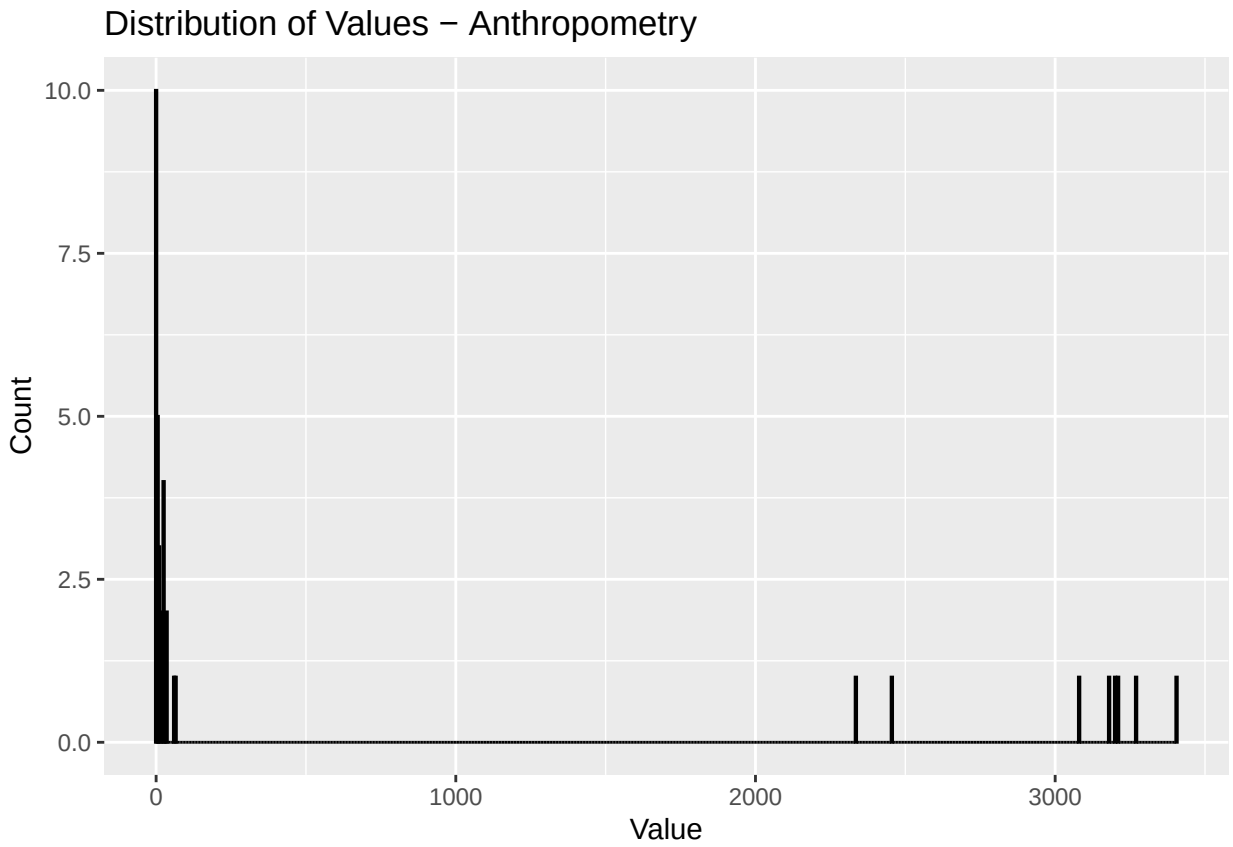
```
## Anthropometry - non-numeric Value entries initially: 1
## Literacy - non-numeric Value entries initially: 1
## MaternalMortality - non-numeric Value entries initially: 1
## ARI_Symptoms - non-numeric Value entries initially: 1
```

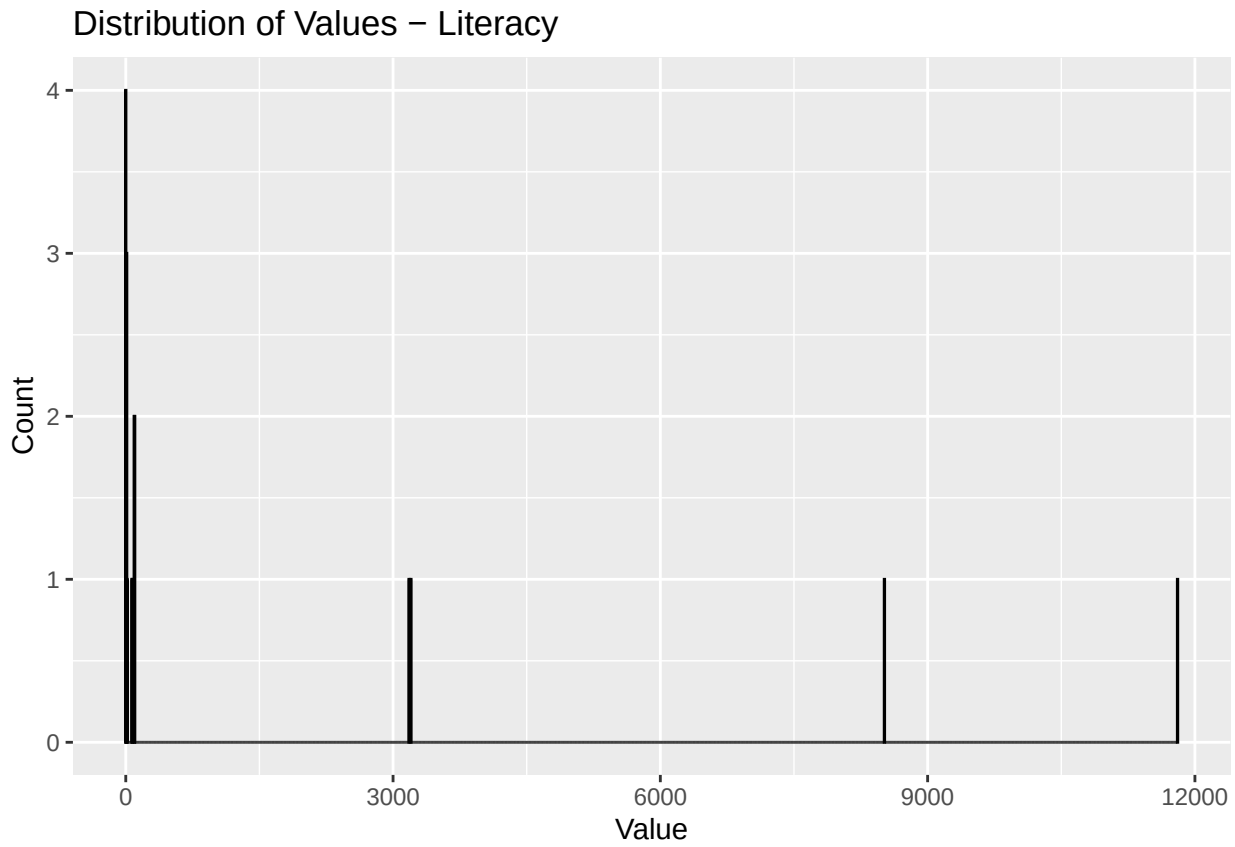
```r
library(dplyr)

stats_list <- lapply(df_list, function(df){
  if ("Value_num" %in% names(df)) {
    tibble(
      Mean = mean(df$Value_num, na.rm = TRUE),
      Median = median(df$Value_num, na.rm = TRUE),
      Variance = var(df$Value_num, na.rm = TRUE),
      Missing = sum(is.na(df$Value_num))
    )
  } else {
    tibble(Mean=NA, Median=NA, Variance=NA, Missing=NA)
  }
})

names(stats_list) <- names(df_list)
stats_list
```

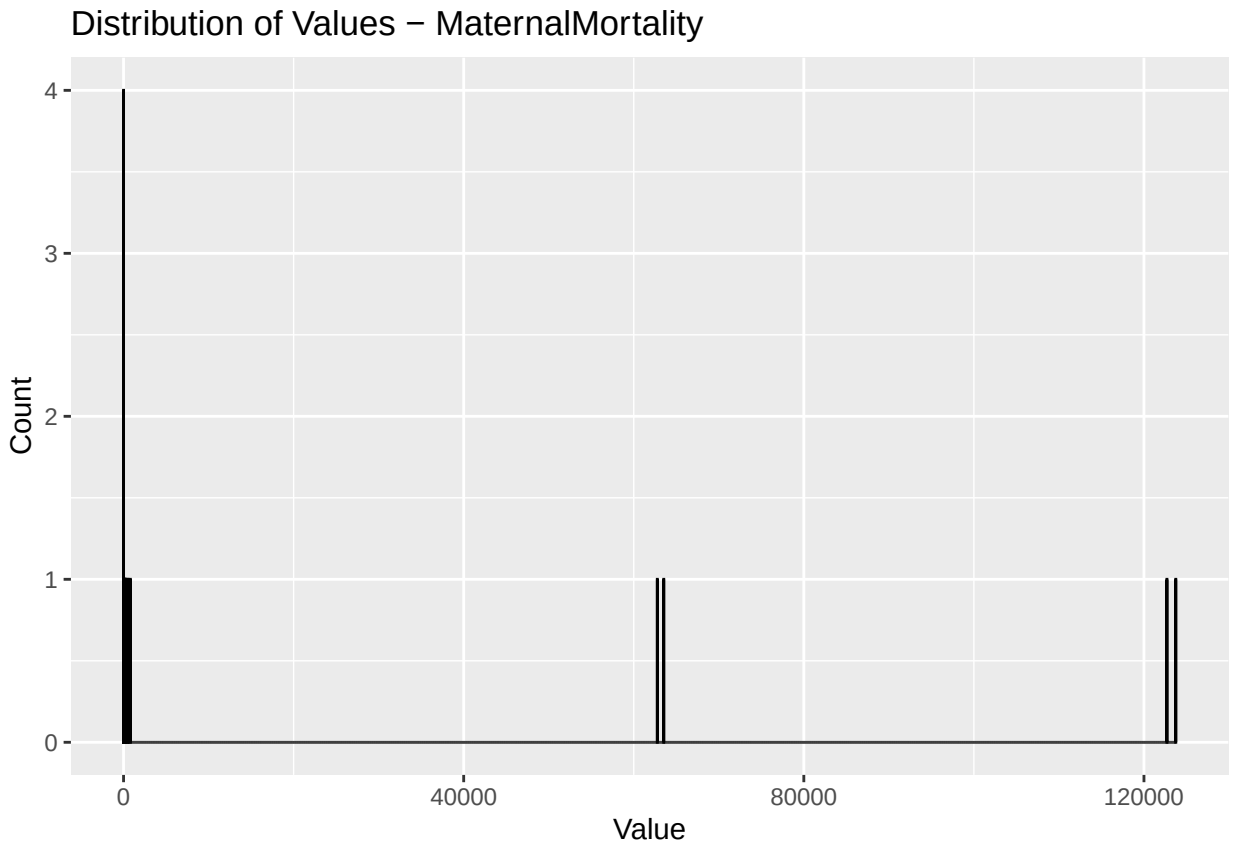
```
## $Anthropometry
## # A tibble: 1 x 4
##    Mean Median Variance Missing
##   <dbl>  <dbl>    <dbl>   <int>
## 1  664.   13.3 1601734.       1
##
## $Literacy
## # A tibble: 1 x 4
```

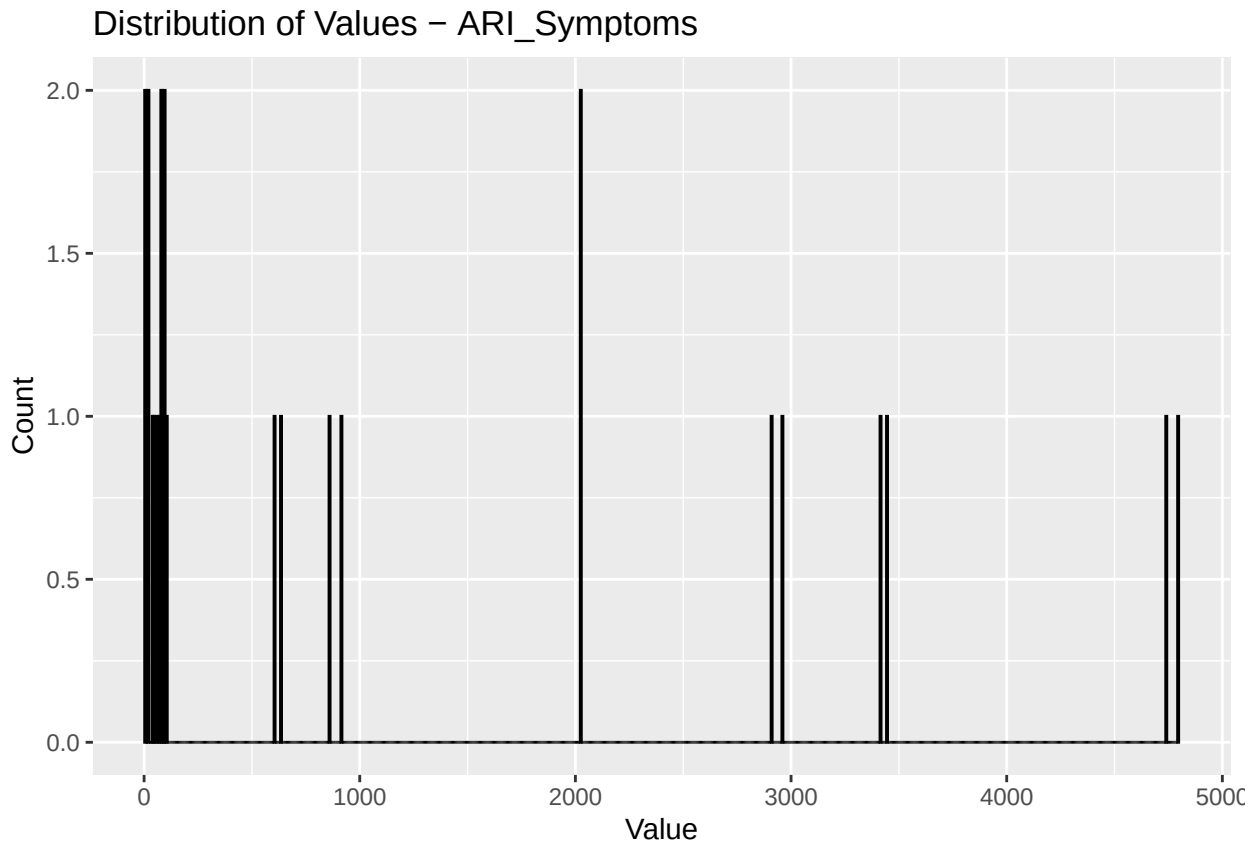
```
##    Mean Median  Variance Missing
##   <dbl>  <dbl>     <dbl>   <int>
## 1 1365.   42.1 10263925.       1
##
## $MaternalMortality
## # A tibble: 1 x 4
##     Mean Median    Variance Missing
##    <dbl>  <dbl>       <dbl>   <int>
## 1 17882.     97 1581411126.       1
##
## $ARI_Symptoms
## # A tibble: 1 x 4
##    Mean Median Variance Missing
##   <dbl>  <dbl>    <dbl>   <int>
## 1 1159.   102. 2475643.       1
```

```r
library(ggplot2)

for (nm in names(df_list)) {
  df <- df_list[[nm]]
  if ("Value_num" %in% names(df)) {
    print(
      ggplot(df, aes(x = Value_num)) +
        geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
        labs(title = paste("Distribution of Values -", nm),
             x = "Value", y = "Count")
    )
  }
}
```

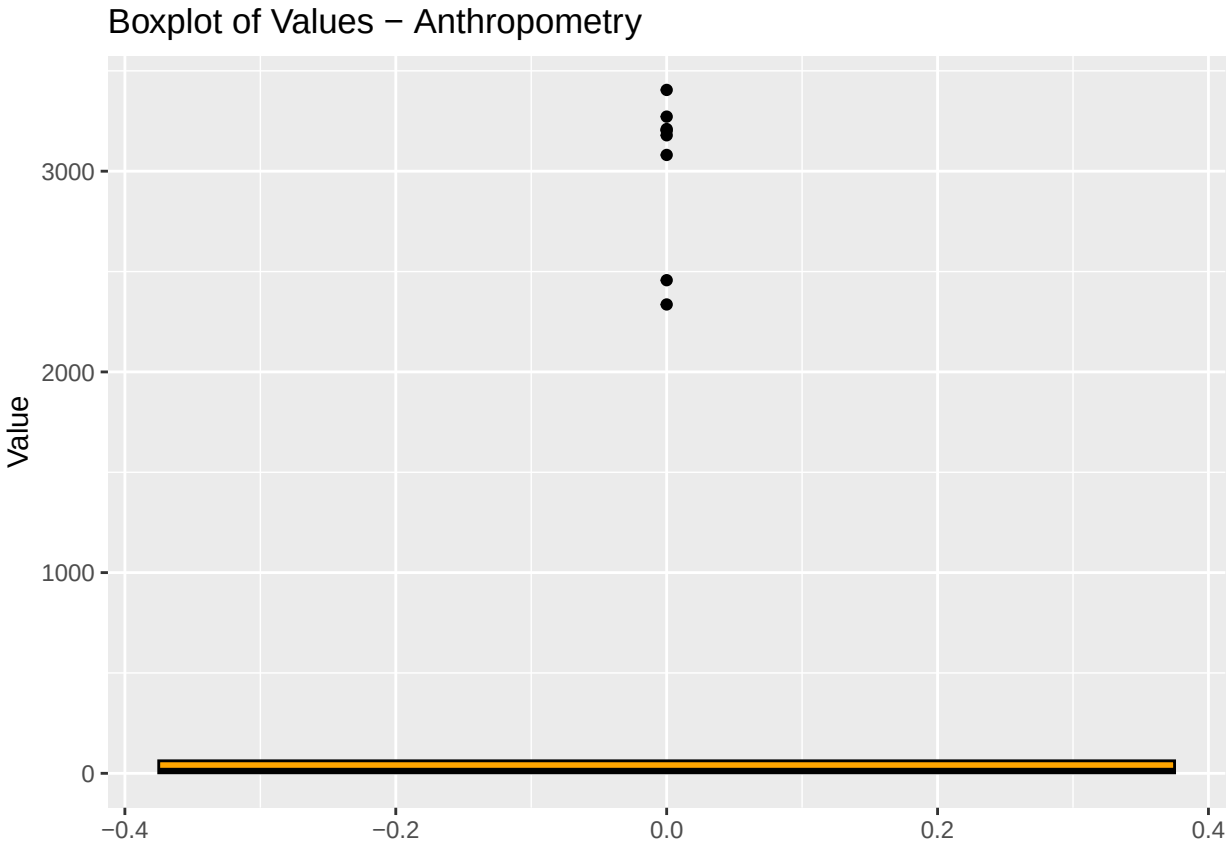

Distribution of Values − Anthropometry
Count
10.0
7.5
5.0
2.5
0.0
0
1000
2000
3000
Value

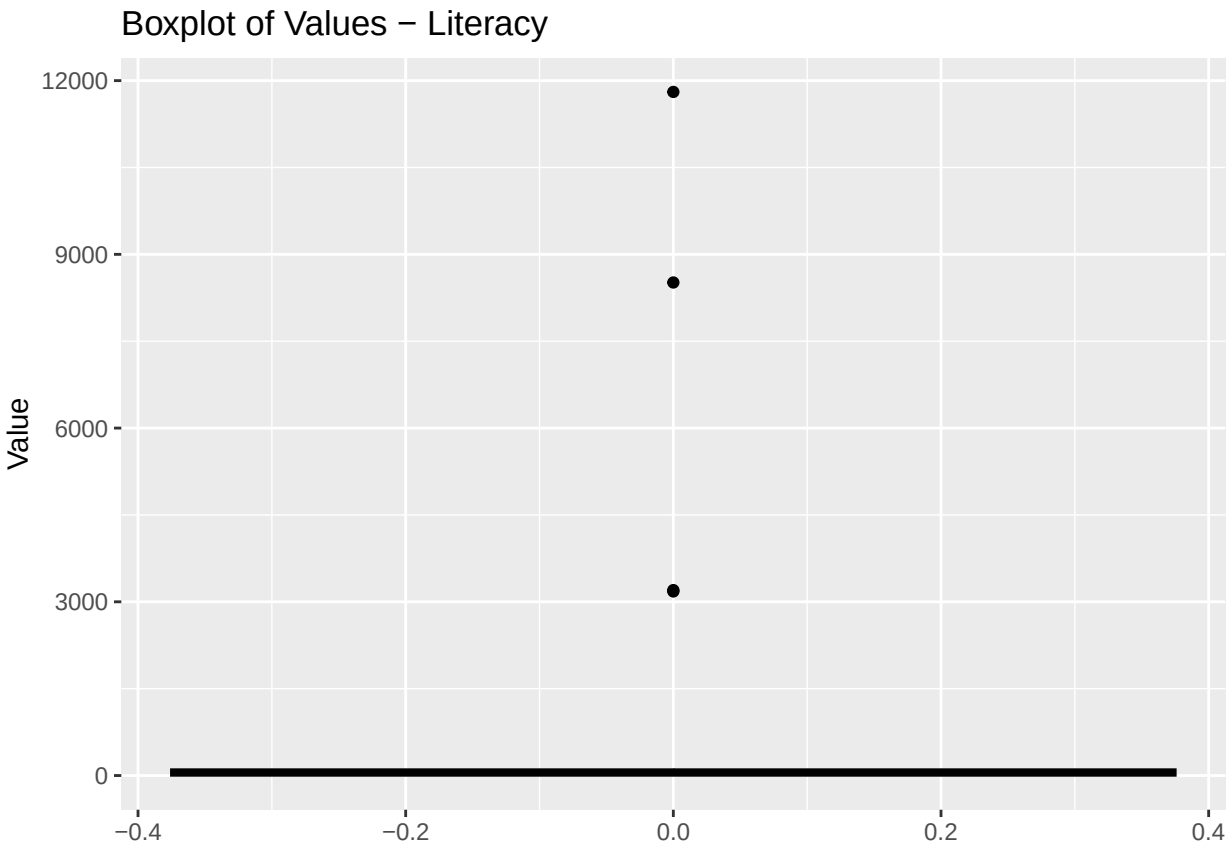

Distribution of Values − Literacy
Count
4
3
2
1
0
0
3000
6000
9000
12000
Value

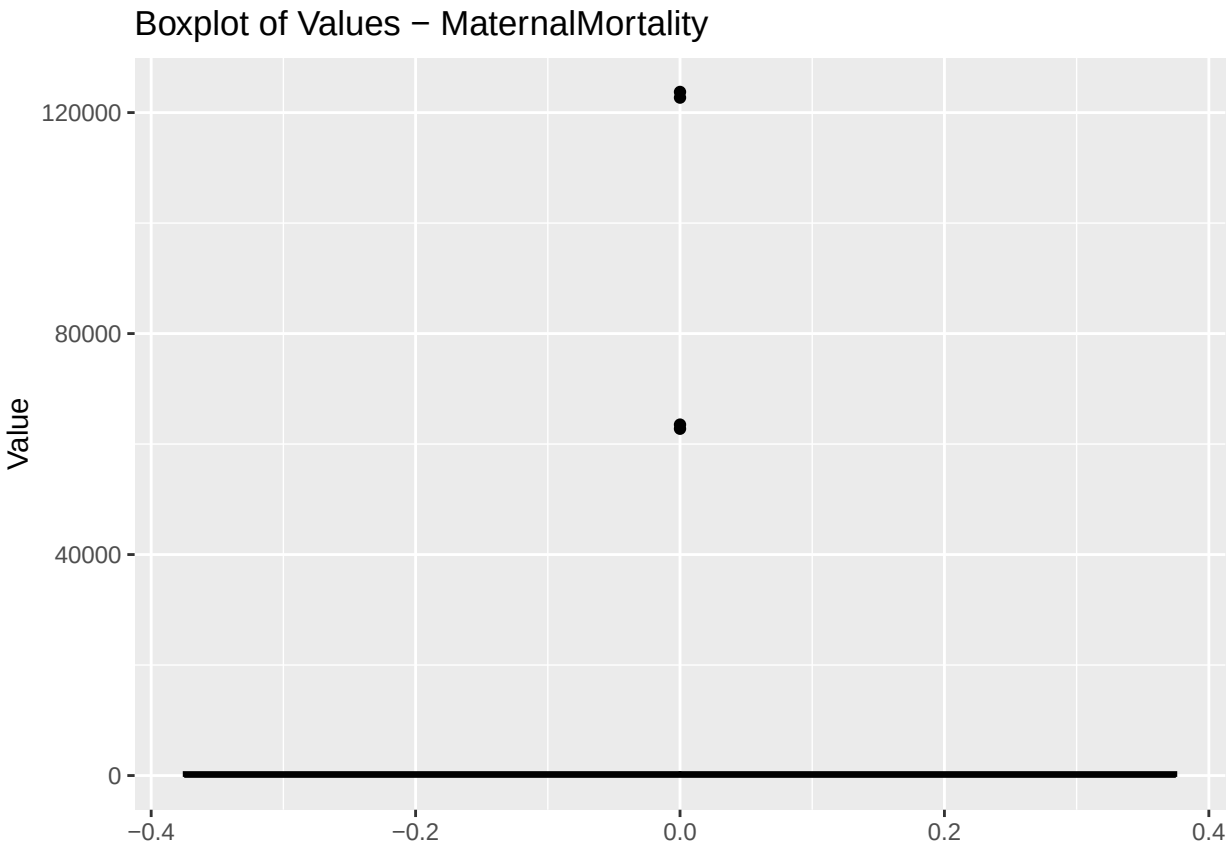

Distribution of Values − MaternalMortality
Count
4
3
2
1
0
0
40000
80000
120000
Value

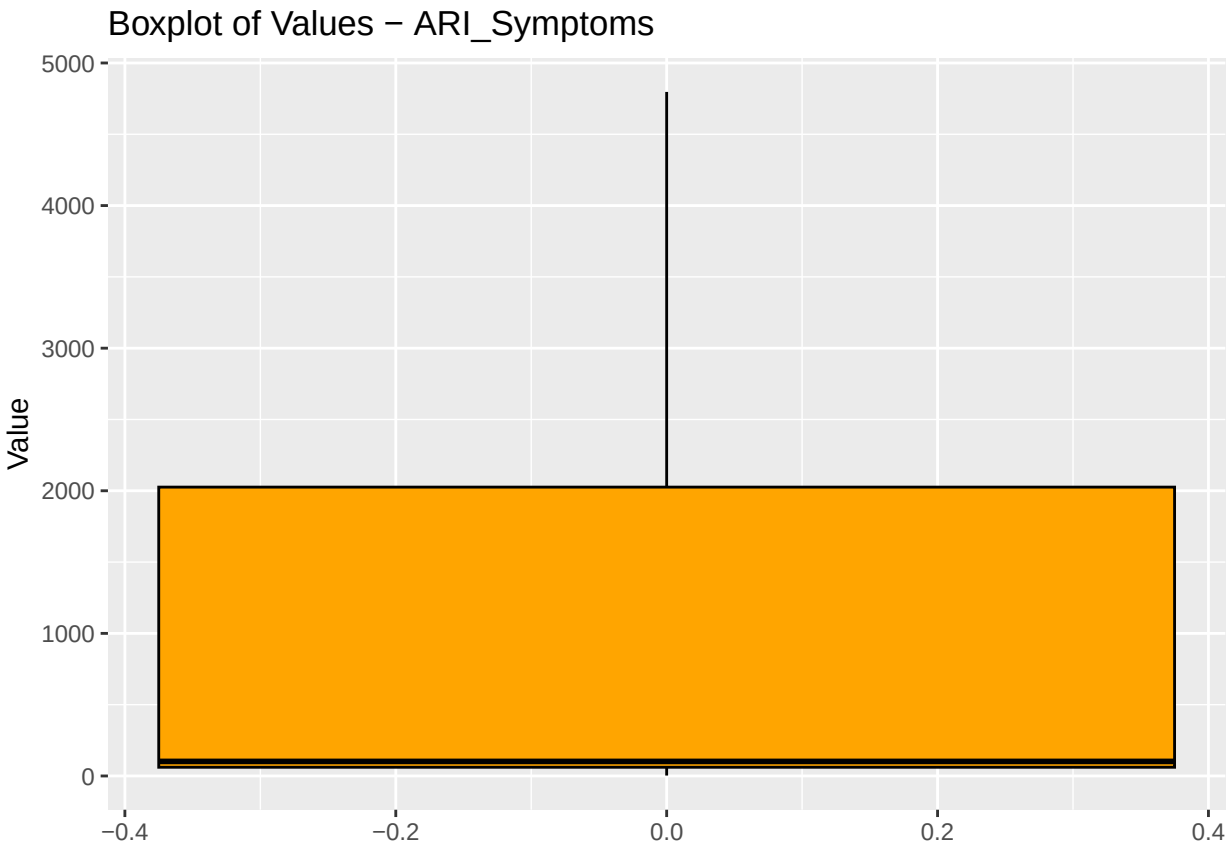



```
for (nm in names(df_list)) {
  df <- df_list[[nm]]
  if ("Value_num" %in% names(df)) {
    print(
      ggplot(df, aes(y = Value_num)) +
        geom_boxplot(fill = "orange", color = "black") +
        labs(title = paste("Boxplot of Values -", nm),
             y = "Value")
    )
  }
}
```


Boxplot of Values − Anthropometry
Value
3000
2000
1000
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Values − Literacy
Value
12000
9000
6000
3000
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Values − MaternalMortality
Value
120000
80000
40000
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Values − ARI_Symptoms
Value
5000
4000
3000
2000
1000
0
−0.4
−0.2
0.0
0.2
0.4