* Show All Code
* Hide All Code

# Geffo

#### Olushina A. Junior Ale

#### 2024-09-23

* [1 **1. Part 1: Data Cleaning**](#part-1-data-cleaning)
  + [1.1 **Load the dataset and check the structure:**](#load-the-dataset-and-check-the-structur)
  + [1.2 **b. Identify the columns with missing values. How many missing values are in the severe wasting column?**](#b.-identify-the-columns-with-missing-va)
  + [1.3 **Impute missing values in the selected columns with their respective means:**](#impute-missing-values-in-the-selected-c)
  + [1.4 **Convert variables into factors:**](#convert-variables-into-factors)
  + [1.5 **Remove non-essential variables:**](#remove-non-essential-variables)
  + [1.6 **Save the cleaned dataset:**](#save-the-cleaned-dataset)
* [2 **2. Part 2: Data Analysis**](#part-2-data-analysis)
  + [2.1 **Create a correlation matrix and interpret relationships:**](#create-a-correlation-matrix-and-interpr)
  + [2.2 **Linear regression analysis:**](#linear-regression-analysis)
  + [2.3 **Identify top 10 countries with the highest “overweight” values:**](#identify-top-10-countries-with-the-high)
* [3 **Part 3: Data Visualization in R**](#part-3-data-visualization-in-r)
  + [3.1 **Scatter plot of “Wasting” vs “Stunting” with color by “LLDC or SIDS”:**](#scatter-plot-of-wasting-vs-stunting-wit)
  + [3.2 **Box plot of “Overweight” by “Income Classification”:**](#box-plot-of-overweight-by-income-classi)
  + [3.3 **Histogram of “Underweight” and interpret skewness:**](#histogram-of-underweight-and-interpret-)

# 1 ****1. Part 1: Data Cleaning****

## 1.1 ****Load the dataset and check the structure:****

tibble [924 × 19] (S3: tbl\_df/tbl/data.frame)

$ ISO code : chr [1:924] "AFG" "AFG" "AFG" "AFG" ...

$ Country : chr [1:924] "AFGHANISTAN" "AFGHANISTAN" "AFGHANISTAN" "AFGHANISTAN" ...

$ Survey Year : chr [1:924] "1997" "2004" "2013" "2018" ...

$ Year : num [1:924] 1997 2004 2013 2018 1997 ...

$ Income Classification: num [1:924] 0 0 0 0 2 2 2 2 2 2 ...

$ LDC : num [1:924] 1 1 1 1 0 0 0 0 0 0 ...

$ LIFD : num [1:924] 1 1 1 1 0 0 0 0 0 0 ...

$ LLDC or SIDS : num [1:924] 1 1 1 1 0 0 0 0 0 0 ...

$ Survey Sample (N) : chr [1:924] "4846" "946" "44,26,469" NA ...

$ Severe Wasting : num [1:924] NA 3.5 4 1.6 NA 6.2 3.7 5.9 0.5 NA ...

$ Wasting : num [1:924] 18.2 8.6 9.5 5.1 8.1 12.2 7.3 9.6 1.6 4 ...

$ Overweight : num [1:924] 6.5 4.6 5.3 4.1 9.5 30.1 24.8 23.2 16.4 NA ...

$ Stunting : num [1:924] 53.2 59.3 40.4 38.2 20.4 39.2 26.7 23.2 11.3 16.9 ...

$ Underweight : num [1:924] 44.9 32.9 24.6 19.1 7.1 17 6.6 6.3 1.5 8 ...

$ Notes : chr [1:924] "Converted estimates" NA NA NA ...

$ Report Author : chr [1:924] "CIET International" "Ministry of Public Health (Afghanistan), UNICEF, CDC, National Institute for Research on Food and Nutrition (It"| \_\_truncated\_\_ "Ministry of Public Health, UNICEF and the Aga Khan University (AKU)." "KIT Royal Tropical Institute" ...

$ Source : chr [1:924] "Afghanistan 1997 multiple indicator baseline (MICS). Report to UNICEF. Communitiy Information and Epidemiologic"| \_\_truncated\_\_ "Summary report of the national nutrition survey, 2004. Kabul, Islamic Republic of Afghanistan: Ministry of Publ"| \_\_truncated\_\_ "Afghanistan National Nutrition Survey 2013." "Afghanistan Health Survey 2018" ...

$ Short Source : chr [1:924] "MICS" "NNS" "SMART" "Other" ...

$ U5 Population ('000s): num [1:924] 3839 4789 5445 5601 309 ...

## 1.2 ****b. Identify the columns with missing values. How many missing values are in the severe wasting column?****

ISO code Country Survey Year

0 0 0

Year Income Classification LDC

0 0 0

LIFD LLDC or SIDS Survey Sample (N)

0 0 63

Severe Wasting Wasting Overweight

228 47 136

Stunting Underweight Notes

37 22 597

Report Author Source Short Source

0 0 0

U5 Population ('000s)

0

[1] 228

## 1.3 ****Impute missing values in the selected columns with their respective means:****

## 1.4 ****Convert variables into factors:****

## 1.5 ****Remove non-essential variables:****

## 1.6 ****Save the cleaned dataset:****

# 2 ****2. Part 2: Data Analysis****

## 2.1 ****Create a correlation matrix and interpret relationships:****

Year Severe Wasting Wasting Overweight

Year 1.000000000 -0.09575465 -0.1385169 -0.002869286

Severe Wasting -0.095754651 1.00000000 0.7564331 0.040241304

Wasting -0.138516894 0.75643312 1.0000000 -0.251987140

Overweight -0.002869286 0.04024130 -0.2519871 1.000000000

Stunting -0.231953156 0.38290519 0.5620605 -0.269852382

Underweight -0.179736996 0.51948907 0.7907981 -0.439728562

U5 Population ('000s) -0.035117199 0.11123042 0.1805507 -0.081590457

Stunting Underweight U5 Population ('000s)

Year -0.2319532 -0.1797370 -0.03511720

Severe Wasting 0.3829052 0.5194891 0.11123042

Wasting 0.5620605 0.7907981 0.18055067

Overweight -0.2698524 -0.4397286 -0.08159046

Stunting 1.0000000 0.8251564 0.13515673

Underweight 0.8251564 1.0000000 0.22238195

U5 Population ('000s) 0.1351567 0.2223819 1.00000000

**Interpretation:**

## 2.2 ****Linear regression analysis:****

Call:

lm(formula = Underweight ~ `Severe Wasting` + Wasting + Overweight +

Stunting + Year + `U5 Population ('000s)`, data = data)

Residuals:

Min 1Q Median 3Q Max

-20.3904 -2.2876 -0.4459 2.0610 22.9589

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.293e+01 3.438e+01 -0.376 0.7069

`Severe Wasting` -2.617e-01 1.445e-01 -1.811 0.0705 .

Wasting 1.187e+00 5.455e-02 21.764 < 2e-16 \*\*\*

Overweight -5.037e-01 3.831e-02 -13.148 < 2e-16 \*\*\*

Stunting 4.141e-01 1.176e-02 35.212 < 2e-16 \*\*\*

Year 6.008e-03 1.711e-02 0.351 0.7256

`U5 Population ('000s)` 4.180e-05 8.848e-06 4.724 2.67e-06 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.433 on 917 degrees of freedom

Multiple R-squared: 0.8722, Adjusted R-squared: 0.8713

F-statistic: 1043 on 6 and 917 DF, p-value: < 2.2e-16

**Is the model a good fit for predicting underweight prevalence? Why or why not?**

**What is the estimated effect of “overweight” on “underweight”, holding other variables constant?**

## 2.3 ****Identify top 10 countries with the highest “overweight” values:****

# A tibble: 10 × 3

Country Year Overweight

<chr> <dbl> <dbl>

1 ALBANIA 2000 30.1

2 LIBYA 2014 29.6

3 UKRAINE 2000 26.5

4 BOSNIA AND HERZEGOVINA 2006 25.7

5 ALBANIA 2005 24.8

6 ALBANIA 2009 23.2

7 LIBYA 2007 22.4

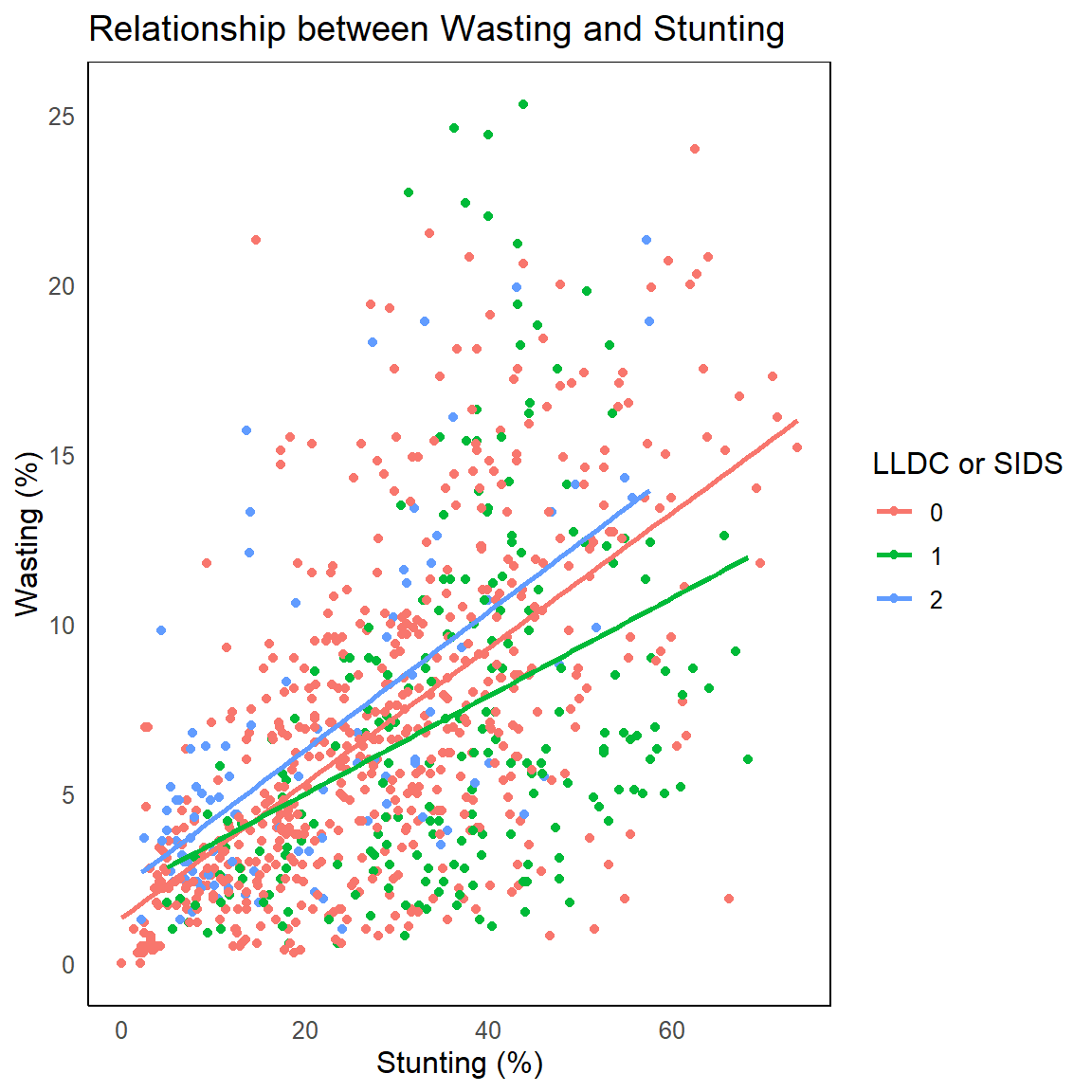
8 MONTENEGRO 2013 22.3

9 AUSTRALIA 2017 22

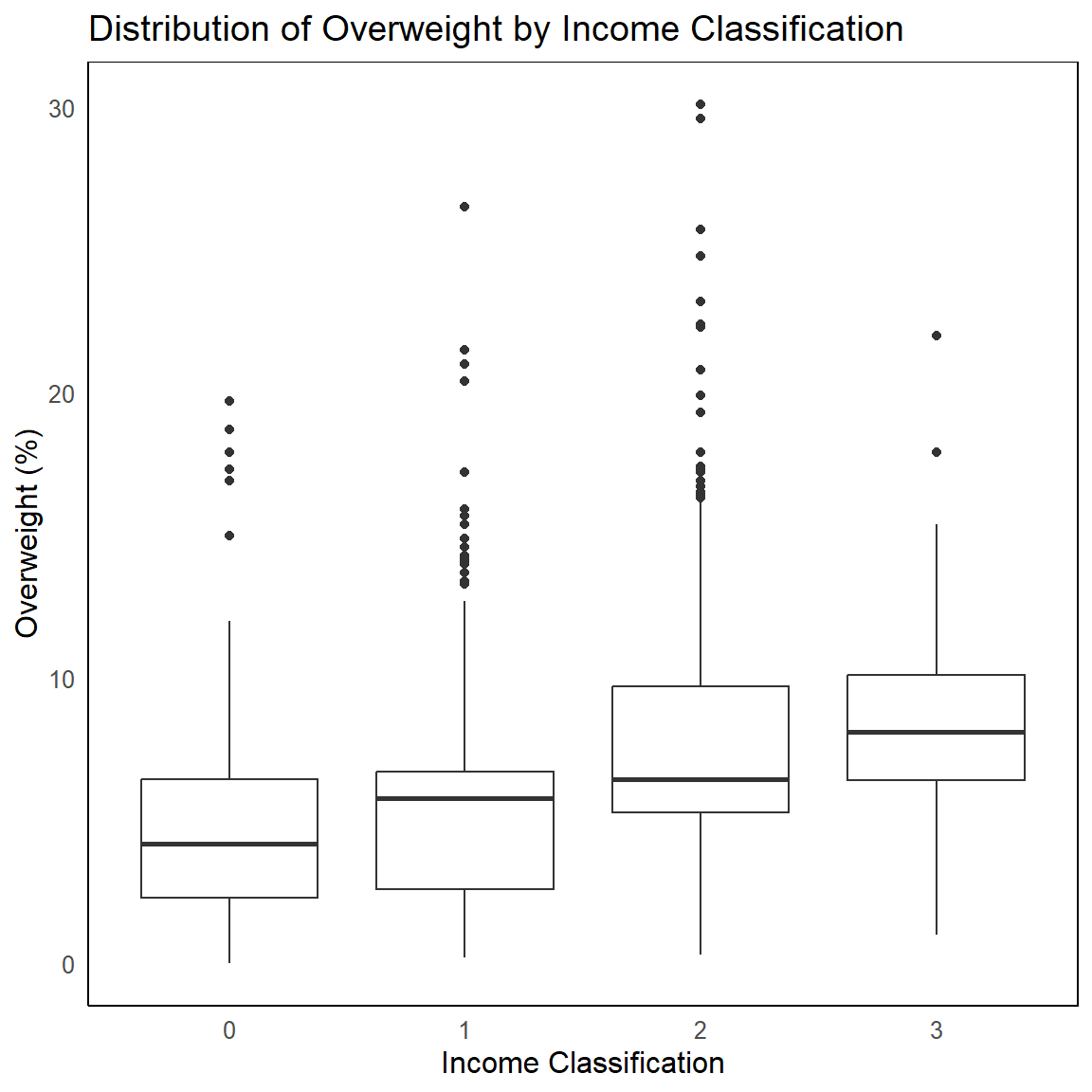
10 COMOROS (THE) 2000 21.5

# 3 ****Part 3: Data Visualization in R****

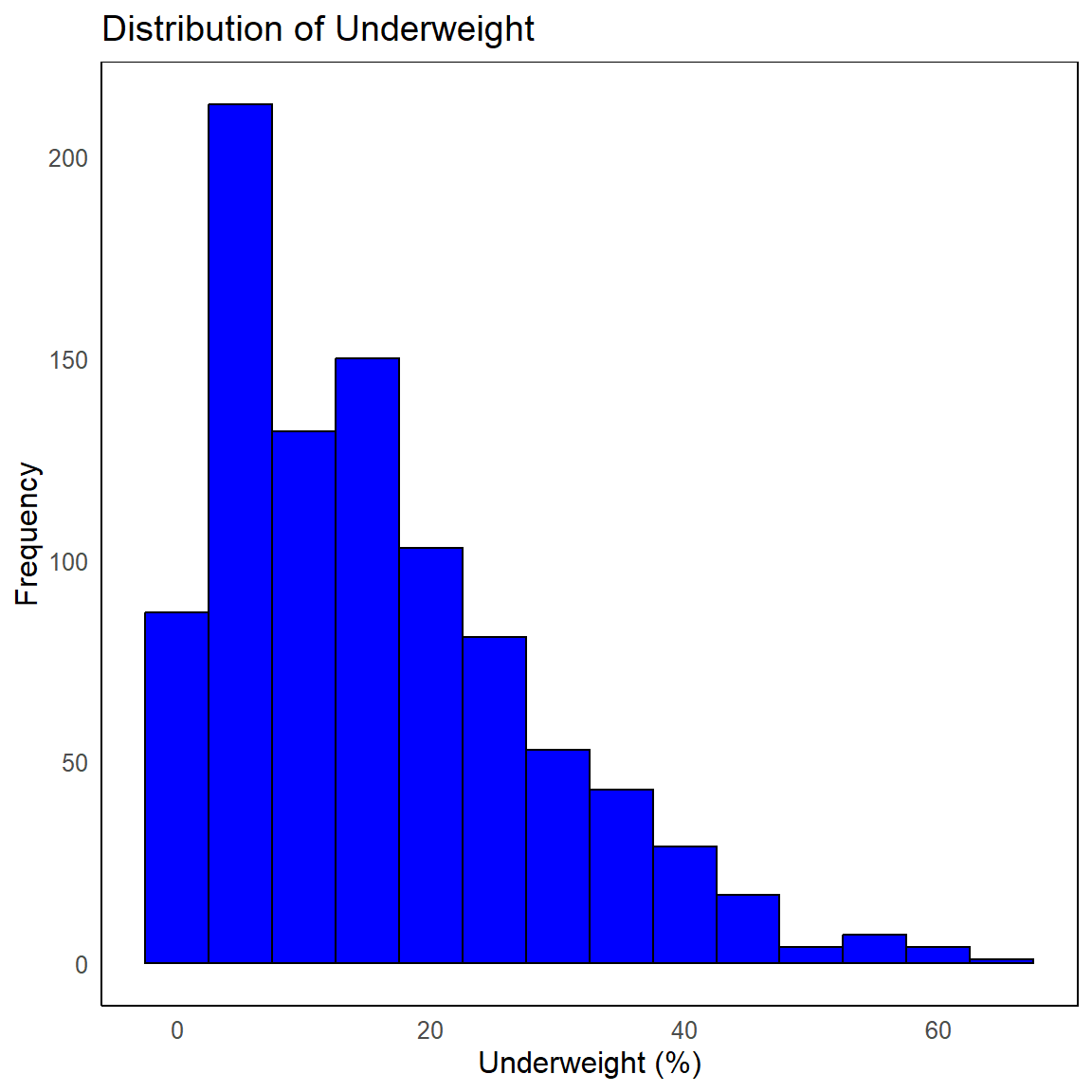
## 3.1 ****Scatter plot of “Wasting” vs “Stunting” with color by “LLDC or SIDS”:****



## 3.2 ****Box plot of “Overweight” by “Income Classification”:****



## 3.3 ****Histogram of “Underweight” and interpret skewness:****



’