Exploring Health Indicators for Children Under 5 Years at a County Level

Paul Muriithi

2023-07-20

Abstract

This data analysis explores health indicators for children under 5 years at a county level in Kenya, focusing on the period from January 2021 to June 2023. The dataset contains monthly data on various variables, including the total number of children dewormed, cases of acute malnutrition, stunted children, children with diarrhea cases, and underweight children in different age groups. The primary goal of this analysis is to identify trends, patterns, and potential relationships between these health indicators to gain insights into child health at a regional level.

Introduction

Child health is a critical aspect of public health, and monitoring key health indicators can provide valuable insights into the well-being of young children. This analysis aims to explore the health indicators for children under 5 years in Kenya's counties to better understand the health status and identify potential areas for improvement.

The dataset used in this analysis consists of granular information at a county level, allowing us to investigate variations in health indicators across regions. We will begin by performing exploratory data analysis (EDA) to understand the data distribution, handle missing values, and visualize key health indicators over time. Subsequently, we will conduct regression analysis to assess the relationship between deworming efforts, stunted growth, underweight cases, and acute malnutrition.

Through this analysis, we hope to provide valuable insights into the health status of children under 5 years in different Kenyan counties, which can be utilized to inform targeted interventions and policies to improve child health outcomes.

1. Exploratory Data Analysis (EDA)

In this section, we perform exploratory data analysis on the provided dataset containing monthly data for children under 5 years, disaggregated at a county level for the period January 2021 to June 2023. The dataset includes information on various variables, such as the total number of children dewormed, number of children with acute malnutrition, stunted children, children with diarrhea cases, and underweight children in different age groups.

```r
# Load necessary libraries
library(tidyverse)
library(lubridate)
library(psych)
library(knitr)
library(gridExtra)
library(lares)
```

```r
library(ggridges)
library(forcats)
library(hrbrthemes)
library(viridis)
library(hrbrthemes)
```

1.1 Load the Data

Next, we load the dataset from the provided URL and display the first few rows to get an overview of the data structure.

```r
# Load the data from the provided URL
data_url <- "https://raw.githubusercontent.com/cema-uonbi/internship_task/main/data/cema_internship_tas
data <- read.csv(data_url)

# View the first few rows of the dataset
head(data[c(1:4)])
```

```
  period                  county Total.Dewormed Acute.Malnutrition
1 Jan-23         Baringo County           3659                  8
2 Jan-23           Bomet County           1580                 NA
3 Jan-23         Bungoma County           6590                 24
4 Jan-23           Busia County           7564                 NA
5 Jan-23 Elgeyo Marakwet County           1407                 NA
6 Jan-23            Embu County           3241                 72
```

1.2 Data Preprocessing

In this step, we rename the column names to make them more descriptive and check for any missing values in the dataset. If there are missing values, we replace them with the median of the corresponding column.

```r
# Rename column names
colnames(data) <- c("Period", "County", "Dewormed", "AcuteMalnutrition", "Stunted(6-23m)",
                    "Stunted(0-<6m)", "Stunted(24-59m)", "DiarrheaCases", "Underweight(0-<6m)",
                    "Underweight(6-23m)", "Underweight(24-59m)")



# Check for missing values
print(sum(is.na(data)))
```

```
[1] 399
```

```r
# replace missing with average
data <- data %>%
  mutate(across(c(3:11), ~replace_na(., median(., na.rm=TRUE))))

# Check for missing values
sum(is.na(data))
```

```
[1] 0
```

1.3 Data Transformation

We convert the 'Period' column to datetime format and arrange the data by 'Period' in ascending order. Additionally, we extract the year from the 'Period' column to facilitate time-series analysis.

```r
# Convert 'Period' column to datetime format
data$Period <- dmy(paste0("01-", data$Period)) # Adding "01-" for day to create valid date format

# Arrange data by 'Period' in ascending order
data <- data %>% arrange(Period)

# Extract year from 'Period' column
data$Year <- year(data$Period)




str(data)
```

```
'data.frame':   1410 obs. of  12 variables:
 $ Period              : Date, format: "2021-01-01" "2021-01-01" ...
 $ County              : chr  "Baringo County" "Bomet County" "Bungoma County" "Busia County" ...
 $ Dewormed            : int  1917 1306 4367 885 1767 817 4888 1377 1093 4866 ...
 $ AcuteMalnutrition   : int  4 39 39 39 39 10 65 9 28 63 ...
 $ Stunted(6-23m)      : int  66 40 46 149 56 15 37 56 24 147 ...
 $ Stunted(0-<6m)      : int  555 17 44 883 53 4 6 165 1 77 ...
 $ Stunted(24-59m)     : int  17 33 22 7 2 50 67 8 55 43 ...
 $ DiarrheaCases       : int  895 4255 2045 514 1881 384 1514 829 892 2690 ...
 $ Underweight(0-<6m)  : int  78 58 154 70 58 83 64 31 24 170 ...
 $ Underweight(6-23m)  : num  90 96 190 70 85 211 370 184 105 264 ...
 $ Underweight(24-59m): num  21 33 54 13 11 ...
 $ Year                : num  2021 2021 2021 2021 2021 ...
```

1.4 Data Description

We generate descriptive statistics for numerical variables, including mean, standard deviation, median, minimum, maximum, range, and standard error.

```r
# Describe the data
kable(describe(data[c(3:11)]) %>%
  select(n, mean, sd, median, min, max, range, se), signif = 3, caption = "Summary Statistics")
```

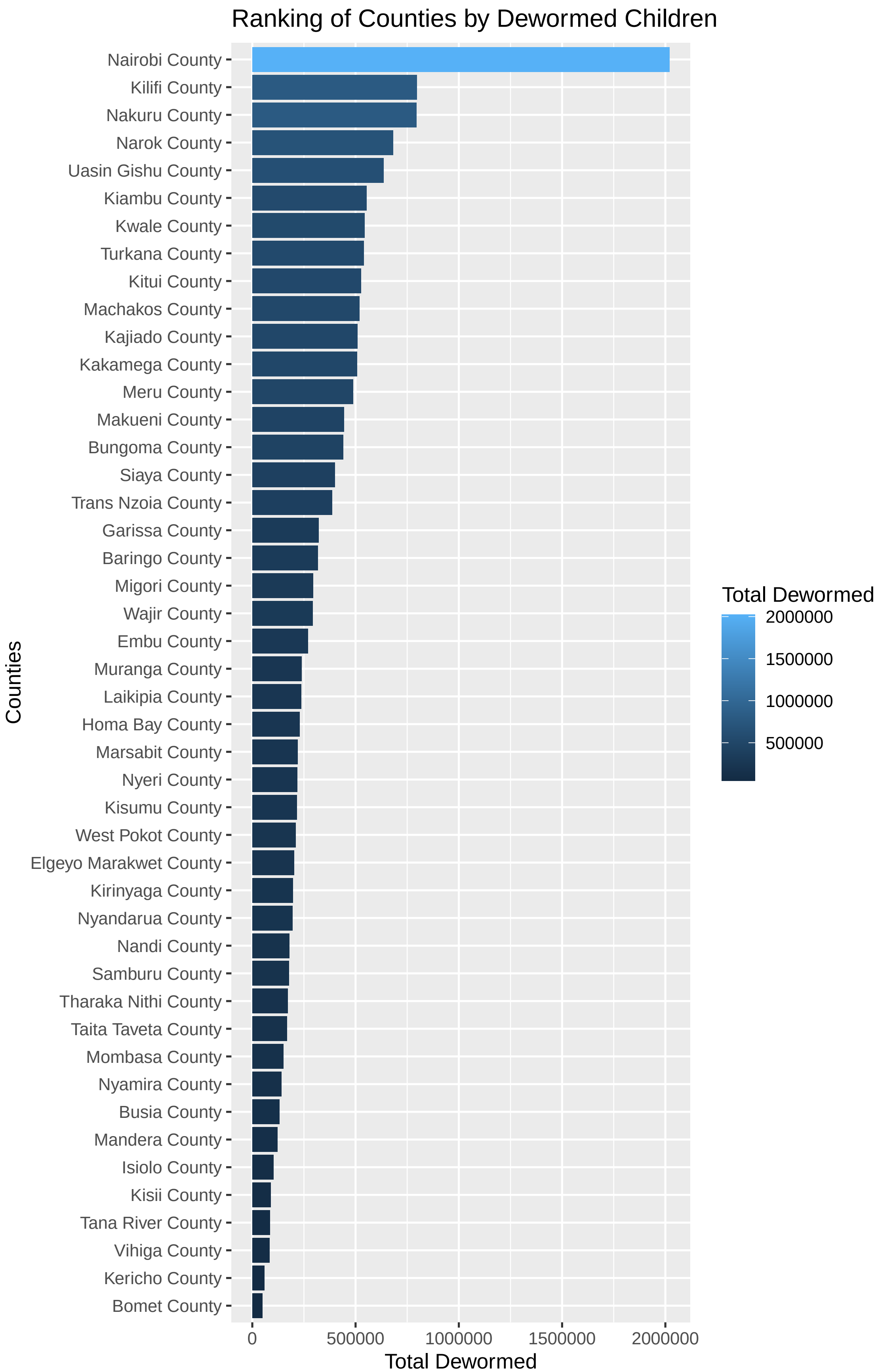
Table 1: Summary Statistics

	n	mean	sd	median	min	max	range	se
Dewormed	1410	11457.9184	25372.4261	4564.5	97	392800	392703	675.697698
AcuteMalnutrition	1410	103.6468	233.5185	39.0	1	4123	4122	6.218874
Stunted(6-23m)	1410	279.2121	379.2081	159.0	1	4398	4397	10.098761
Stunted(0-<6m)	1410	139.0397	278.4202	84.0	1	7900	7899	7.414658
Stunted(24-59m)	1410	110.1617	192.5275	50.0	1	3169	3168	5.127236
DiarrheaCases	1410	2813.3823	2161.8961	2158.0	198	15795	15597	57.573850
Underweight(0-<6m)	1410	223.4709	228.5319	162.5	6	1937	1931	6.086075
Underweight(6-23m)	1410	652.2595	669.5775	456.0	16	5348	5332	17.831641
Underweight(24-59m)	1410	305.7372	538.4616	120.5	1	4680	4679	14.339870

1.5 Ranking of Total Dewormed by Counties

We create a bar plot to rank the counties based on the total number of children dewormed. The height of each bar represents the total dewormed count for the respective county.

```
data1 <- data%>%
  group_by(County) %>%
  summarise(Dewormed = sum(Dewormed)) %>%
  filter(Dewormed> 25000)

  ggplot(data=data1,aes(x=reorder(County,Dewormed, top = 10),y=Dewormed)) +
  geom_bar(stat ='identity',aes(fill=Dewormed))+
  coord_flip() +
  theme_grey() +
  scale_fill_gradient(name="Total Dewormed")+
  labs(title = 'Ranking of Counties by Dewormed Children',
       y='Total Dewormed',x='Counties')
```

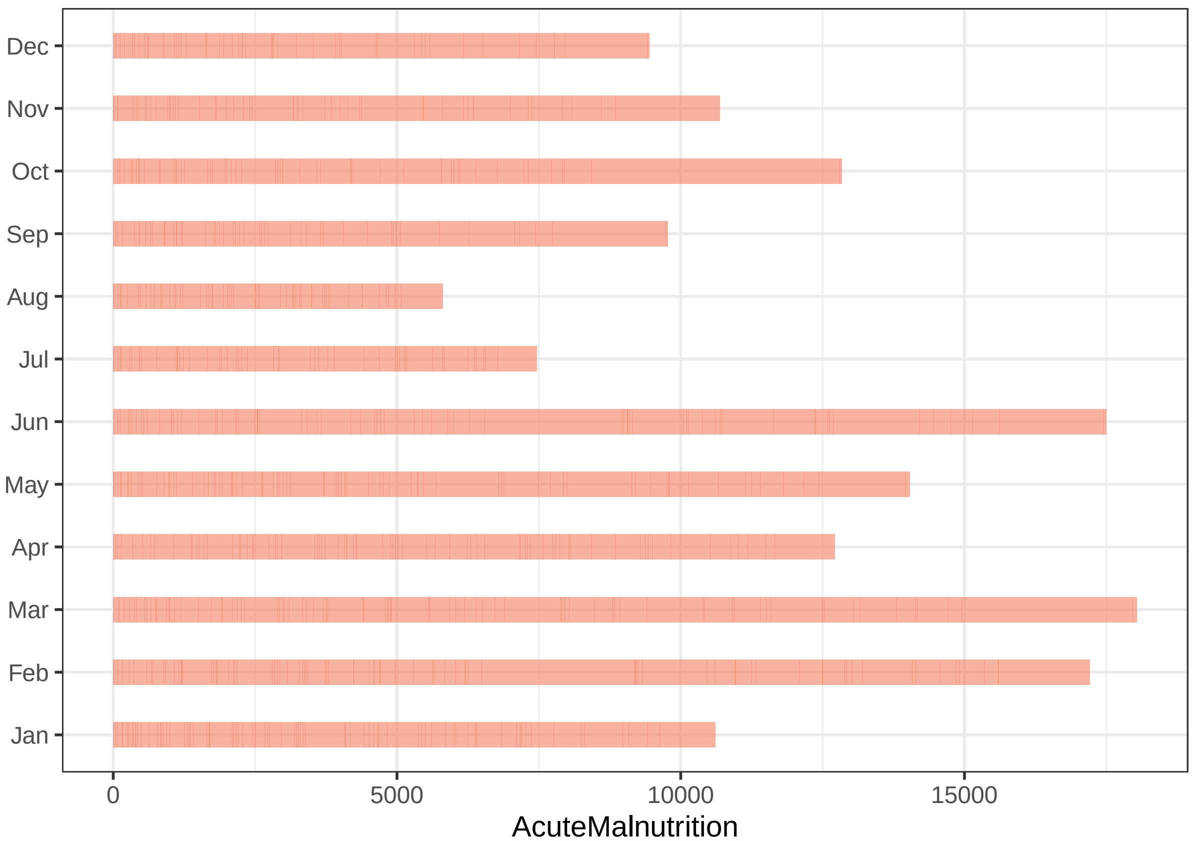

Ranking of Counties by Dewormed Children
Nairobi County
Kilifi County
Nakuru County
Narok County
Uasin Gishu County
Kiambu County
Kwale County
Turkana County
Kitui County
Machakos County
Kajiado County
Kakamega County
Meru County
Makueni County
Bungoma County
Siaya County
Trans Nzoia County
Garissa County
Baringo County
Migori County
Wajir County
Embu County
Muranga County
Laikipia County
Homa Bay County
Marsabit County
Nyeri County
Kisumu County
West Pokot County
Elgeyo Marakwet County
Kirinyaga County
Nyandarua County
Nandi County
Samburu County
Tharaka Nithi County
Taita Taveta County
Mombasa County
Nyamira County
Busia County
Mandera County
Isiolo County
Kisii County
Tana River County
Vihiga County
Kericho County
Bomet County
Counties
0
500000
1000000
1500000
2000000
Total Dewormed
Total Dewormed
2000000
1500000
1000000
500000

1.6 Monthly Distribution of Acute Malnutrition Cases

In this section, we focus on visualizing the monthly distribution of acute malnutrition cases. We start by extracting the month from the 'Period' column and create a new column 'Month' with abbreviated month names.

To display the distribution effectively, we reorder the months based on the number of acute malnutrition cases. This arrangement ensures that the months are displayed in descending order of acute malnutrition cases.

```
# Extract the month from 'Period'
data$Month <- month(data$Period, label = TRUE, abbr = TRUE)


# Reorder following the value of another column:
data %>%
  mutate(name = fct_reorder(Month, AcuteMalnutrition)) %>%
  ggplot( aes(Month, AcuteMalnutrition)) +
  geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
  coord_flip() +
  xlab("") +
  theme_bw()
```

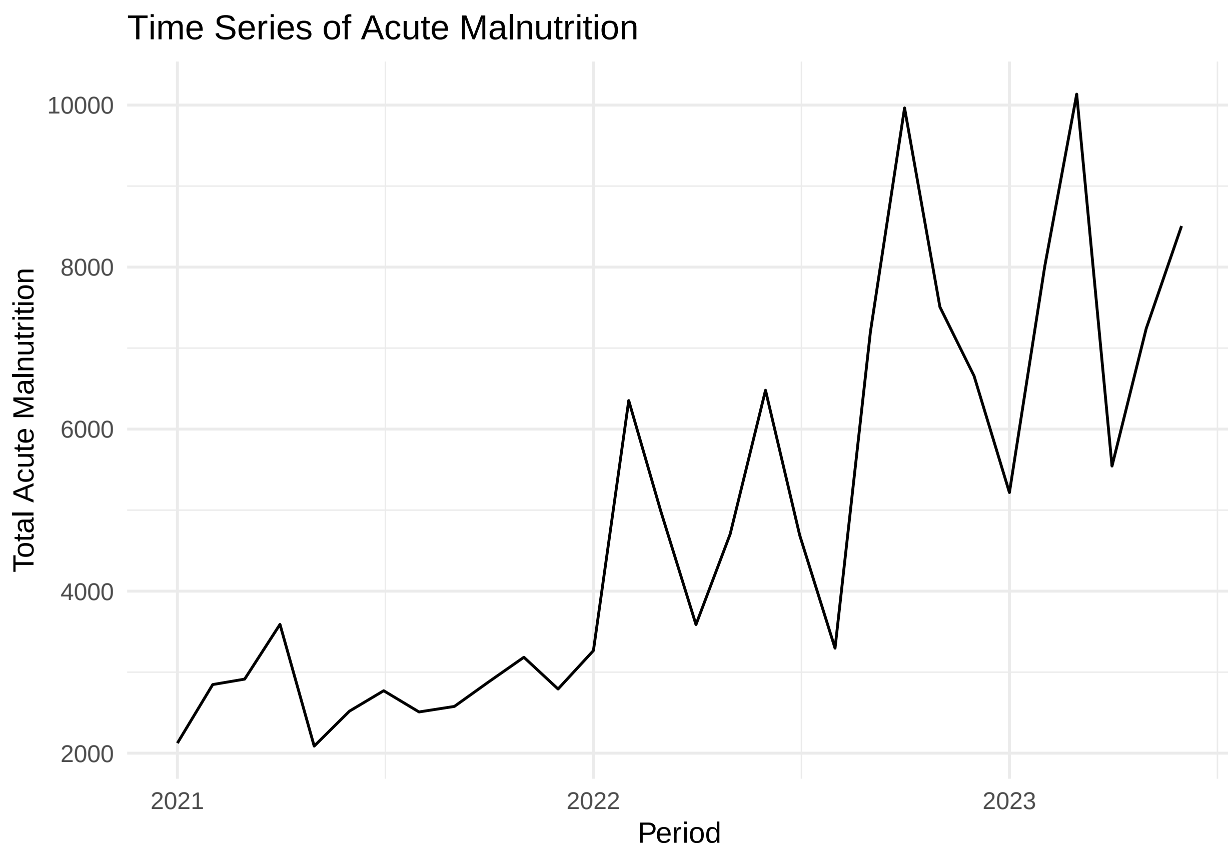


** 1.7 Time Series of Acute Malnutrition**

We visualize the time series of acute malnutrition cases over the study period. The plot shows the trend of acute malnutrition cases for each month.

```
# Group data by Period and calculate monthly sum of Acute Malnutrition cases
data_time_series <- data %>%
```

```r
  group_by(Period) %>%
  summarise(Total_Acute_Malnutrition = sum(AcuteMalnutrition, na.rm = TRUE)) %>%
  ungroup()

# Time series plot
ggplot(data_time_series, aes(x = Period, y = Total_Acute_Malnutrition)) +
  geom_line() +
  labs(title = "Time Series of Acute Malnutrition",
       x = "Period",
       y = "Total Acute Malnutrition") +
  theme_minimal()
```



2. Data Analysis

2.1 Research Question:

Before conducting the data analysis, let's define the research question based on the dataset:

- ***Research Question:** How does the total number of children with Acute Malnutrition vary across counties, and what is the relationship between deworming efforts, stunted growth, underweight and acute malnutrition cases in different counties?

2.2 Top 10 Counties with Highest Total Dewormed

We identify the top 10 counties with the highest total dewormed count and display their corresponding acute malnutrition values in a table.

```
top_10_counties <- data %>%
  group_by(County) %>%
  summarise(Total_Dewormed = median(Dewormed),
            Acute_Malnutrition = median(AcuteMalnutrition)) %>%
  top_n(10, Total_Dewormed) %>%
  arrange(desc(Total_Dewormed))

# Display the top 10 counties and their corresponding Acute Malnutrition values
kable(top_10_counties, caption = "Top 10 counties with highest Deworming rate")
```

Table 2: Top 10 counties with highest Deworming rate

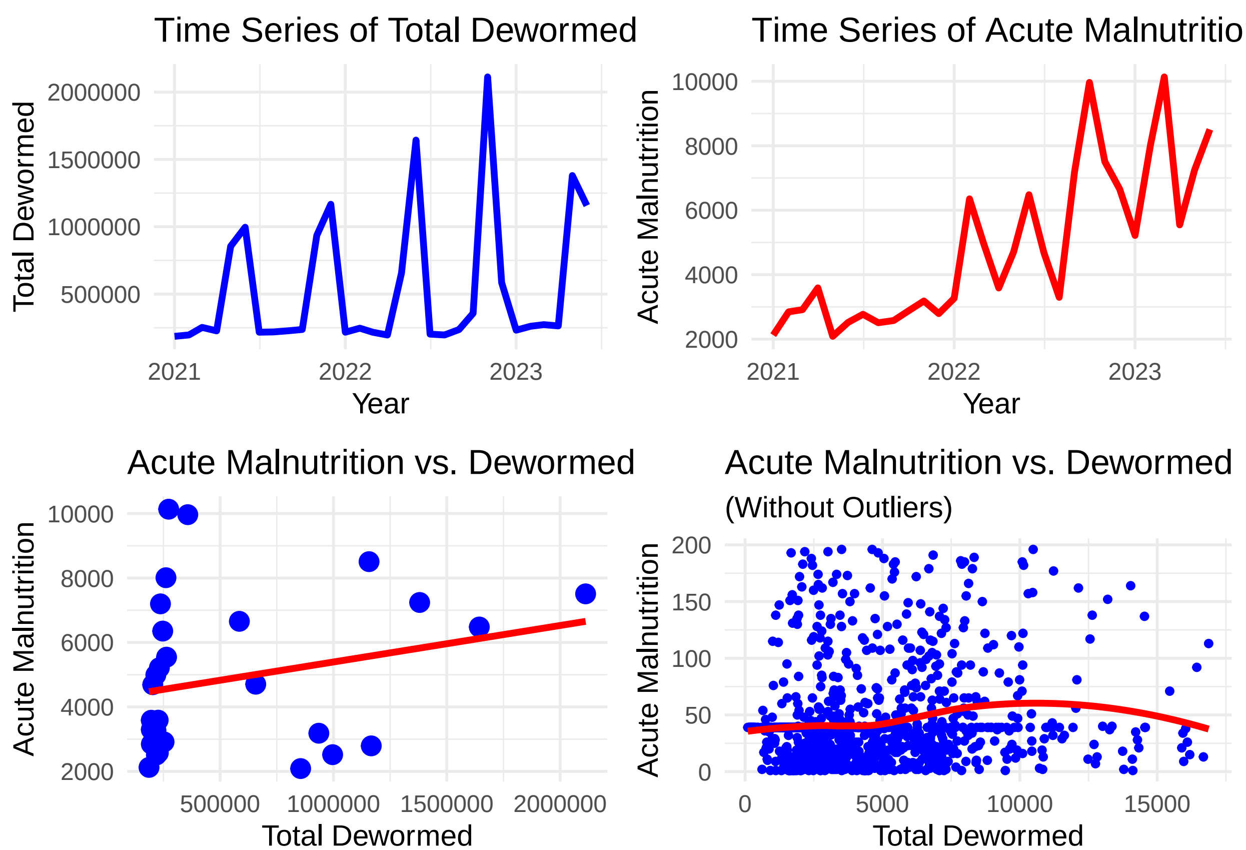
County	Total_Dewormed	Acute_Malnutrition
Nairobi County	22066.0	313.0
Turkana County	11144.5	291.5
Nakuru County	10386.5	161.0
Kakamega County	10088.5	26.0
Garissa County	8160.0	227.0
Kiambu County	7924.0	92.5
Bungoma County	7738.5	39.0
Kilifi County	7539.0	33.0
Kwale County	7508.5	83.0
Uasin Gishu County	7392.0	39.0

2.3 Analysing Total Dewormed vs. Acute Malnutrition

We create separate time series plots for the total number of children dewormed and acute malnutrition cases over the study period. These plots allow us to observe any trends or patterns in the two variables.

```
# Time series plot for Total Dewormed and Acute Malnutrition (separate lines)
data_time_series <- data %>%
  group_by(Period) %>%
  summarise(Total_Dewormed = sum(Dewormed, na.rm = TRUE),
            Acute_Malnutrition = sum(AcuteMalnutrition, na.rm = TRUE))

# Plot for Total Dewormed
plot_total_dewormed <- ggplot(data_time_series, aes(x = Period, y = Total_Dewormed)) +
  geom_line(color = "blue", size = 1.2) +
  labs(title = "Time Series of Total Dewormed",
       x = "Year",
       y = "Total Dewormed") +
  theme_minimal()

# Plot for Acute Malnutrition
plot_acute_malnutrition <- ggplot(data_time_series, aes(x = Period, y = Acute_Malnutrition)) +
  geom_line(color = "red", size = 1.2) +
  labs(title = "Time Series of Acute Malnutrition",
       x = "Year",
       y = "Acute Malnutrition") +
  theme_minimal()


# Scatter plot
```

```
scatter_plot <- ggplot(data_time_series, aes(x = Total_Dewormed, y = Acute_Malnutrition)) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", color = "red", se = FALSE, size = 1.2) +
  labs(title = "Acute Malnutrition vs. Dewormed",
       x = "Total Dewormed",
       y = "Acute Malnutrition") +
  theme_minimal()
# Calculate IQR and filter out outliers for 'Total Dewormed' and 'Acute Malnutrition'
outlier_removed_data <- data %>%
  filter(between(Dewormed, quantile(Dewormed, 0.25) - 1.5*IQR(Dewormed), quantile(Dewormed, 0.75)+
                 1.5*IQR(Dewormed)),
         between(AcuteMalnutrition, quantile(AcuteMalnutrition, 0.25) - 1.5*IQR(AcuteMalnutrition),
                 quantile(AcuteMalnutrition, 0.75) + 1.5*IQR(AcuteMalnutrition)))

# Scatter plot without outliers
without_outliers <- ggplot(outlier_removed_data, aes(x = Dewormed, y = AcuteMalnutrition)) +
  geom_point(color = "blue", size = 1) +
  geom_smooth(method = "loess", color = "red", se = FALSE, size = 1.2) +
  labs(title = "Acute Malnutrition vs. Dewormed",
       subtitle = "(Without Outliers)",
       x = "Total Dewormed",
       y = "Acute Malnutrition") +
  theme_minimal()

# Combine both plots using grid.arrange
grid.arrange(plot_total_dewormed, plot_acute_malnutrition, scatter_plot, without_outliers, ncol = 2)
```



2.4 Comparing Stunted and Underweight Cases by County

In this section, we compare the total number of stunted and underweight children in each county. We group the data by county and calculate the aggregate sum of stunting cases (combining cases for age groups 0-6 months, 6-23 months, and 24-59 months) and underweight cases (combining cases for age groups 0-6 months, 6-23 months, and 24-59 months) for each county.

The table below shows the top 10 counties with the highest number of stunted and underweight children.

```
# Calculate the stunting cases for each county
stunted_cases <- data %>%
  group_by(County) %>%
  summarise(Stunted = sum(`Stunted(0-<6m)`, `Stunted(6-23m)`, `Stunted(24-59m)`),
            Underweight  = sum(`Underweight(0-<6m)`, `Underweight(6-23m)`, `Underweight(24-59m)`)) %>%
  arrange(desc(Underweight)) %>%
  top_n(10, Underweight)

kable(stunted_cases, caption = "Number of Stunted and Underweight Children by county")
```

Table 3: Number of Stunted and Underweight Children by county

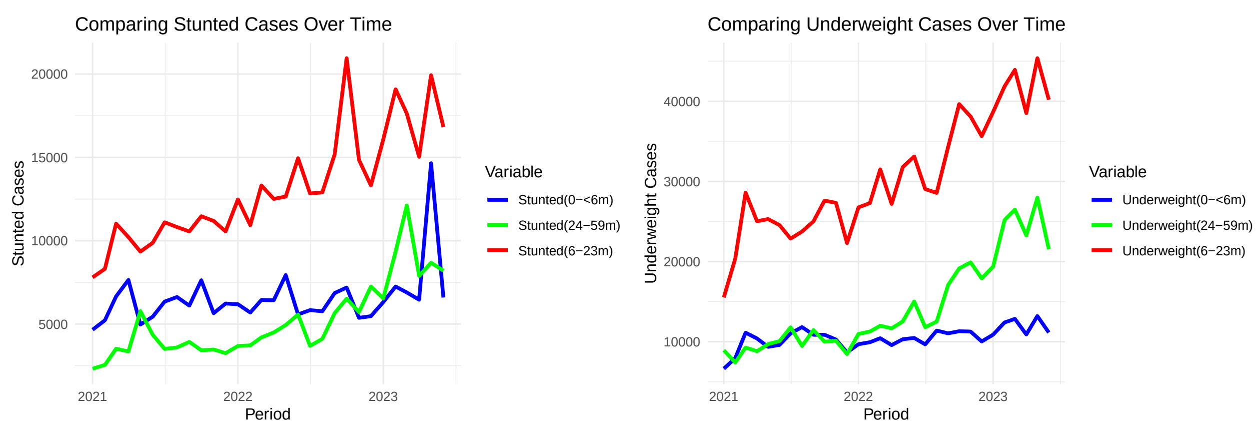
County	Stunted	Underweight
Turkana County	47667	156954
Nairobi County	106321	156930
Kilifi County	55695	82130
Nakuru County	26994	67272
Marsabit County	20334	65555
Kiambu County	30303	65325
Garissa County	8954	62778
Wajir County	9255	55566
Kwale County	30222	53607
Kitui County	32309	52957

2.5 Comparing Stunted and Underweight Cases Over Time

We compare the time series of stunted and underweight cases for different age groups (0-6 months, 6-23 months, and 24-59 months) over the study period. The plots display the trend of stunted and underweight cases for each age group.

```
# Time series plots
stunted_underweight_ts <- data %>%
  group_by(Period) %>%
  summarise(`Stunted(0-<6m)` = sum(`Stunted(0-<6m)`),
            `Stunted(6-23m)` = sum(`Stunted(6-23m)`),
            `Stunted(24-59m)` = sum(`Stunted(24-59m)`),
            `Underweight(0-<6m)` = sum(`Underweight(0-<6m)`),
            `Underweight(6-23m)` = sum(`Underweight(6-23m)`),
            `Underweight(24-59m)` = sum(`Underweight(24-59m)`))

stunted_plot <- ggplot(stunted_underweight_ts, aes(x = Period)) +
  geom_line(aes(y = `Stunted(0-<6m)`, color = "Stunted(0-<6m)"), size = 1.2) +
  geom_line(aes(y = `Stunted(6-23m)`, color = "Stunted(6-23m)"), size = 1.2) +
  geom_line(aes(y = `Stunted(24-59m)`, color = "Stunted(24-59m)"), size = 1.2) +
  labs(title = "Comparing Stunted Cases Over Time",
       x = "Period",
```

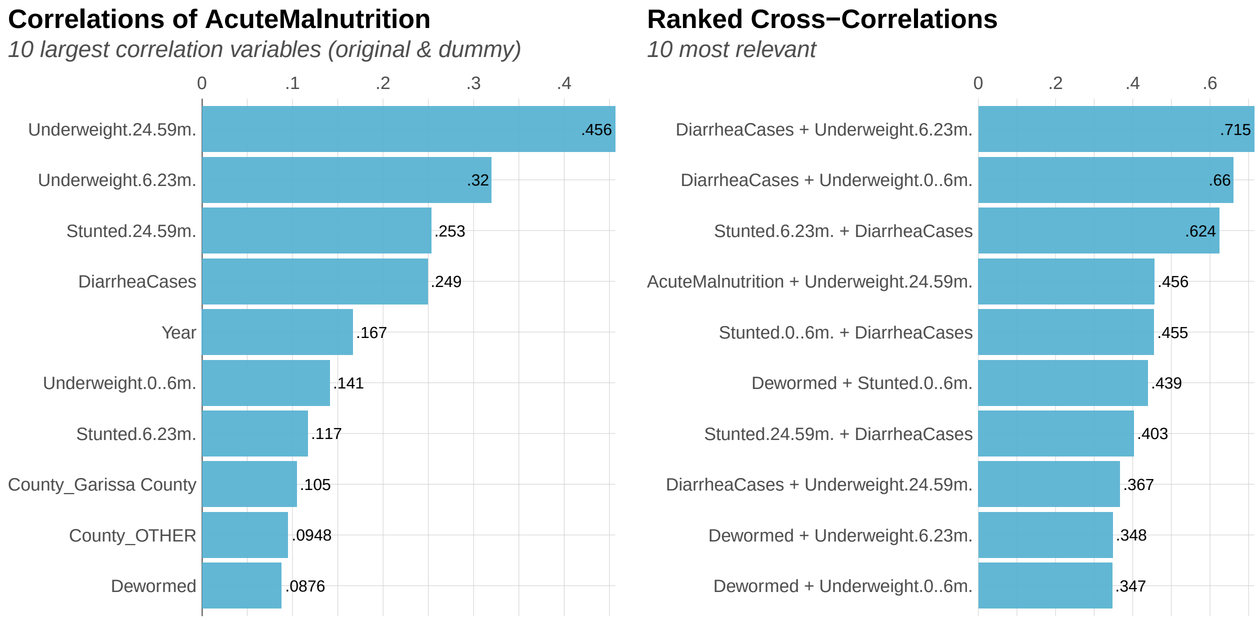
```r
       y = "Stunted Cases",
       color = "Variable") +
  scale_color_manual(name = "Variable",
                     values = c("Stunted(0-<6m)" = "blue",
                                "Stunted(6-23m)" = "red",
                                "Stunted(24-59m)" = "green")) +
  theme_minimal()


#
underweight_plot <- ggplot(stunted_underweight_ts, aes(x = Period)) +
  geom_line(aes(y = `Underweight(0-<6m)`, color = "Underweight(0-<6m)"), size = 1.2) +
  geom_line(aes(y = `Underweight(6-23m)`, color = "Underweight(6-23m)"), size = 1.2) +
  geom_line(aes(y = `Underweight(24-59m)`, color = "Underweight(24-59m)"), size = 1.2) +
  labs(title = "Comparing Underweight Cases Over Time",
       x = "Period",
       y = "Underweight Cases",
       color = "Variable") +
  scale_color_manual(name = "Variable",
                     values = c("Underweight(0-<6m)" = "blue",
                                "Underweight(6-23m)" = "red",
                                "Underweight(24-59m)" = "green")) +
  theme_minimal()

# Combine both plots using grid.arrange
grid.arrange(stunted_plot, underweight_plot, ncol = 2)
```



2.6 Correlation Analysis

Finally, we perform a correlation analysis to identify the top 10 correlations with acute malnutrition. We display the correlation matrix and cross-correlations for these top correlations.

```r
# Add 'TotalStunted' and 'TotalUnderweight' columns
#data$TotalStunted <- data$`Stunted(0-<6m)` + data$`Stunted(6-23m)` + data$`Stunted(24-59m)`
#data$TotalUnderweight <- data$`Underweight(0-<6m)` + data$`Underweight(6-23m)` + data$`Underweight(24-


# Show only top 5 corrrelations
acute_mal_corr <- data%>%corr_var(AcuteMalnutrition, top = 10)

grid.arrange(acute_mal_corr, corr_cross(data, top = 10), ncol=2)
```



2.7 Regression Analysis

```
# Perform linear regression
regression_model <- lm(log(AcuteMalnutrition) ~ Dewormed + DiarrheaCases +
                        `Stunted(6-23m)` + `Stunted(0-<6m)` + `Stunted(6-23m)` +
                        `Stunted(24-59m)` + `Underweight(0-<6m)` + `Underweight(6-23m)` +
                        `Underweight(24-59m)`, data = data)

# Extract only the coefficients table from the summary of the regression model
coefficients <- summary(regression_model)$coefficients

# Print the coefficients table
kable(coefficients, caption = "Regression coefficients")
```

Table 4: Regression coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.8901371	0.0562187	51.4087935	0.0000000
Dewormed	-0.0000001	0.0000015	-0.0497244	0.9603491
DiarrheaCases	0.0001283	0.0000234	5.4862666	0.0000000
`Stunted(6-23m)`	-0.0009606	0.0001655	-5.8051498	0.0000000
`Stunted(0-<6m)`	-0.0001249	0.0001655	-0.7547037	0.4505536
`Stunted(24-59m)`	-0.0005384	0.0002769	-1.9445297	0.0520315
`Underweight(0-<6m)`	0.0005245	0.0002806	1.8694384	0.0617703
`Underweight(6-23m)`	0.0006160	0.0001359	4.5317934	0.0000063
`Underweight(24-59m)`	0.0007892	0.0001086	7.2693011	0.0000000