CEMA INTERNSHIP TASK

Kelvin Njenga

2023-07-19

# 1.Business understanding

CEMA (Center for Epidemiological Modelling and Analysis) has provided us with a comprehensive dataset containing monthly data for children under 5 years in Kenya, detailed at a county level. The dataset covers the period from January 2021 to June 2023 and encompasses vital health indicators for children.The dataset includes information on the total number of children dewormed, the prevalence of acute malnutrition, the number of stunted children in different age groups, the cases of diarrhea among children, and the prevalence of underweight children in various age categories.

## Objectives

The primary objective of this analysis is to conduct EDA to identify trends, patterns, and potential areas for targeted interventions to improve child health and well-being. State an appropriate research question to answer from the data.

## Research Question

What are the spatial and temporal patterns of “Total Dewormed” and “Diarrhea Cases” across the 47 counties in Kenya, and is there any correlation or clustering between these health indicators for children below 5 years?

# 2. Data Understanding

The dataset contains the following Features/Columns:

| Column | Description |
| --- | --- |
| period | The period (months from January 2021 to June 2023) |
| county | The name of the county in Kenya |
| Total Dewormed | Total number of children dewormed |
| Acute Malnutrition | Number of children <5 years with acute malnutrition |
| stunted 6-23 months | Number of children stunted (6-23 months) |
| stunted 0-<6 months | Number of children stunted (0-6 months) |
| stunted 24-59 months | Number of children stunted (24-59 months) |
| diarrhea cases | Number of children <5 years with diarrhea |
| Underweight 0-<6 months | Number of children underweight (0-6 months) |
| Underweight 6-23 months | Number of children underweight (6-23 months) |
| Underweight 24-59 Months | Number of children underweight (24-59 months) |

# 3.Data Preprocessing

## Importing the required libraries

# Importing required libraries  
  
library(tidyverse) # for data manipulation and visualization  
library(lubridate) # for dates  
library(dplyr) # for data manipulation and summarization  
library(ggplot2) # for creating data visualizations  
library(rgdal) # for geospatial data formats  
library(sf) # for spatial data  
library(foreign) # for reading and writing data files in various formats(.dbf etc)  
library(leaflet) # for interactive mapping visualizations  
library(plotly)  
library(corrplot) # for correlation plots  
library(knitr) # for dynamic reports  
library(RColorBrewer) # for color palettes

## Loading the Excel file

data <- read\_csv("data/cema\_internship\_task\_2023.csv")  
# Viewing the top 5 dataset values  
head(data)

## # A tibble: 6 × 11  
## period county `Total Dewormed` `Acute Malnutrition` `stunted 6-23 months`  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Jan-23 Baringo Co… 3659 8 471  
## 2 Jan-23 Bomet Coun… 1580 NA 1  
## 3 Jan-23 Bungoma Co… 6590 24 98  
## 4 Jan-23 Busia Coun… 7564 NA 396  
## 5 Jan-23 Elgeyo Mar… 1407 NA 92  
## 6 Jan-23 Embu County 3241 72 326  
## # ℹ 6 more variables: `stunted 0-<6 months` <dbl>,  
## # `stunted 24-59 months` <dbl>, `diarrhoea cases` <dbl>,  
## # `Underweight 0-<6 months` <dbl>, `Underweight 6-23 months` <dbl>,  
## # `Underweight 24-59 Months` <dbl>

## Dealing with Missing values and Duplicates

### Checking for Missing values

# Checking for missing values and other descriptive statistics  
summary(data)

## period county Total Dewormed Acute Malnutrition  
## Length:1410 Length:1410 Min. : 97 Min. : 1.0   
## Class :character Class :character 1st Qu.: 2454 1st Qu.: 15.0   
## Mode :character Mode :character Median : 4564 Median : 39.0   
## Mean : 11458 Mean : 125.4   
## 3rd Qu.: 8222 3rd Qu.: 143.5   
## Max. :392800 Max. :4123.0   
## NA's :355   
## stunted 6-23 months stunted 0-<6 months stunted 24-59 months diarrhoea cases  
## Min. : 1.0 Min. : 1.0 Min. : 1.0 Min. : 198   
## 1st Qu.: 69.5 1st Qu.: 36.5 1st Qu.: 22.0 1st Qu.: 1464   
## Median : 159.0 Median : 84.0 Median : 50.0 Median : 2158   
## Mean : 280.2 Mean : 139.8 Mean : 110.8 Mean : 2813   
## 3rd Qu.: 328.5 3rd Qu.: 157.0 3rd Qu.: 114.2 3rd Qu.: 3335   
## Max. :4398.0 Max. :7900.0 Max. :3169.0 Max. :15795   
## NA's :11 NA's :19 NA's :14   
## Underweight 0-<6 months Underweight 6-23 months Underweight 24-59 Months  
## Min. : 6.0 Min. : 16.0 Min. : 1.00   
## 1st Qu.: 87.0 1st Qu.: 249.0 1st Qu.: 51.25   
## Median : 162.5 Median : 456.0 Median : 120.50   
## Mean : 223.5 Mean : 652.3 Mean : 305.74   
## 3rd Qu.: 272.8 3rd Qu.: 791.8 3rd Qu.: 311.00   
## Max. :1937.0 Max. :5348.0 Max. :4680.00   
##

# Checking the percentage of missing values(NA)  
calculate\_missing\_percentage <- function(data) {  
 missing\_percentage <- colMeans(is.na(data)) \* 100  
 return(missing\_percentage)  
}  
missing\_percentages <- calculate\_missing\_percentage(data)  
kable(missing\_percentages, caption = "Missing Data Percentages")

Missing Data Percentages

|  | x |
| --- | --- |
| period | 0.0000000 |
| county | 0.0000000 |
| Total Dewormed | 0.0000000 |
| Acute Malnutrition | 25.1773050 |
| stunted 6-23 months | 0.7801418 |
| stunted 0-<6 months | 1.3475177 |
| stunted 24-59 months | 0.9929078 |
| diarrhoea cases | 0.0000000 |
| Underweight 0-<6 months | 0.0000000 |
| Underweight 6-23 months | 0.0000000 |
| Underweight 24-59 Months | 0.0000000 |

### Dealing With missing values

# Replacing all NA's with 0  
data <- data %>%  
 mutate(across(everything(), ~ifelse(is.na(.), 0, .)))  
  
# checking if the NA's have been replaced with 0  
missing\_percentages <- calculate\_missing\_percentage(data)  
kable(missing\_percentages, caption = "Missing Data Percentages")

Missing Data Percentages

|  | x |
| --- | --- |
| period | 0 |
| county | 0 |
| Total Dewormed | 0 |
| Acute Malnutrition | 0 |
| stunted 6-23 months | 0 |
| stunted 0-<6 months | 0 |
| stunted 24-59 months | 0 |
| diarrhoea cases | 0 |
| Underweight 0-<6 months | 0 |
| Underweight 6-23 months | 0 |
| Underweight 24-59 Months | 0 |

### Checking for duplicates

duplicates <- any(duplicated(data))  
print(duplicates)

## [1] FALSE

shape <- dim(data)  
print(paste("The shape of our dataset contains", shape[1], "rows and", shape[2], "columns."))

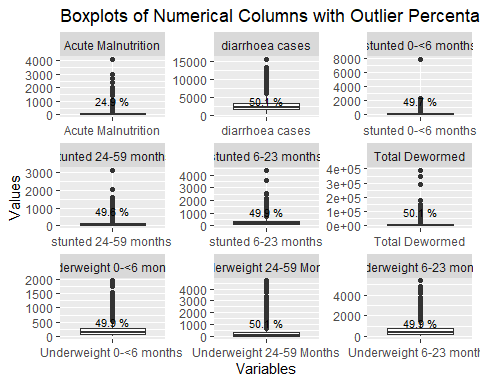
## [1] "The shape of our dataset contains 1410 rows and 11 columns."

### Renaming and converting our dataset columns to a 28 day calender(To include fulldates and February)

# removing county from county name entries  
data$county <- gsub(" County", "", data$county)  
  
# Convert the period date to a 28 day full date format  
data$full\_date <- dmy(paste0("28-", data$period))  
  
# Display the updated data with the new "full\_date" column  
view(data)

## Checking for Outliers

plot\_boxplots <- function(data) {  
 # Filter only the numerical columns  
 numerical\_data <- data %>% select\_if(is.numeric)  
  
 # Reshape the data for plotting  
 numerical\_data\_long <- numerical\_data %>%  
 pivot\_longer(everything(), names\_to = "variable", values\_to = "value")  
  
 # Calculate percentage of outliers for each variable  
 outlier\_percentage <- numerical\_data\_long %>%  
 group\_by(variable) %>%  
 summarize(outlier\_percentage = mean(value < quantile(value, 0.25) | value > quantile(value, 0.75)) \* 100)  
  
 # Plot boxplots with outlier percentage labels  
 ggplot(numerical\_data\_long, aes(x = variable, y = value)) +  
 geom\_boxplot() +  
 stat\_summary(fun = function(x) mean(x < quantile(x, 0.25) | x > quantile(x, 0.75)) \* 100,   
 geom = "text", aes(label = paste(round(after\_stat(y), 1), "%")),   
 vjust = -1, hjust = 0.5, size = 3) +  
 facet\_wrap(~ variable, scales = "free") +  
 labs(title = "Boxplots of Numerical Columns with Outlier Percentage",  
 x = "Variables", y = "Values")  
}  
plot\_boxplots(data)



In this analysis, I have chosen to ignore the outliers as it allows me to direct my attention towards the central and more common characteristics of the data. This approach enables me to gain deeper insights into the typical behavior and patterns exhibited by the majority of data points.

Furthermore, outliers can distort visualizations, making it challenging to perceive the overall trends and patterns effectively. Since my primary interest in this analysis is to understand the general trends and characteristics shared by the majority of the data, ignoring outliers allows me to focus on the factors that are more relevant to typical scenarios.

By excluding outliers, I aim to create visualizations and summary statistics that are more representative of the central tendency of the data. This ensures that my analysis is more accurate, reliable, and aligned with the research questions I am seeking to address.

# 4. External dataset validation

Diarrhoea is the second leading cause of death among children < 5 years of age. It accounts for one out of twenty seven child fatalities globally, with 80% of these occurring in low-middle-income countries. In Kenya, diarrhoea is responsible for 17% of all childhood diseases, with children < 5 years experiencing, on average, three incidences of diarrhoea annually.From the graph plotted below, it is noted that during the months of January to March 2021 to 2023, there is an increase in diarrhoea cases in the aggregated 47 counties. This could be attributed to various factors such as the dry season in Kenya during this period.

In the report [Extreme Weather Events in Kenya Between 2011 and 2020](https://meteo.go.ke/sites/default/files/downloads/Extreme%20Climate%20Events_Kenya%202011_to_2020.pdf) by the Kenya Meteorological Department, it is noted that the drought season exacerbated by La Nina conditions is experienced in the Country. Due to the decrease in rainfall, a water shortage is experienced in the country. This leads to prioritization of water usage for more crucial activities such as storing the little remaining water for drinking to sustain life. Additionally, this prompts households to also make use of any water that they may come across, whether clean or not.

In [Onyango, I. (2022). Determinants of Diarrheal Cases among Children under five Years in Households using Domestic Water in Kangemi, Nairobi County, Kenya](http://ir.jooust.ac.ke:8080/xmlui/handle/123456789/12138), the author attributes some of the determinants of the increase in diarrhea cases to inadequate sanitation and hygiene as well as tainted drinking water.

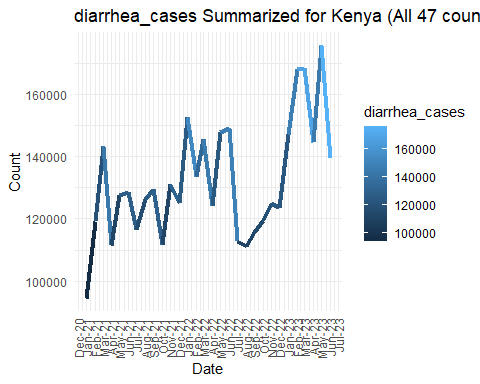
### Summarizing our dataset: Whole Kenya Overview

line\_data <- data %>%  
 group\_by(period) %>%  
 summarise(  
 total\_dewormed = sum(`Total Dewormed`, na.rm = TRUE),  
 diarrhea\_cases = sum(`diarrhoea cases`, na.rm = TRUE),  
 acute\_malnutrition = sum(`Acute Malnutrition`, na.rm = TRUE),  
 stunted\_6\_23\_months = sum(`stunted 6-23 months`, na.rm = TRUE),  
 stunted\_0\_6\_months = sum(`stunted 0-<6 months`, na.rm = TRUE),  
 stunted\_24\_59\_months = sum(`stunted 24-59 months`, na.rm = TRUE),  
 underweight\_0\_6\_months = sum(`Underweight 0-<6 months`, na.rm = TRUE),  
 underweight\_6\_23\_months = sum(`Underweight 6-23 months`, na.rm = TRUE),  
 underweight\_24\_59\_months = sum(`Underweight 24-59 Months`, na.rm = TRUE)  
 ) %>%  
 mutate(period = str\_c(period, "-01")) %>%  
 separate(period, into = c("month", "year", "day"), sep = "-", remove = FALSE) %>%   
 mutate(month\_number = as.character(match(month, month.abb)),  
 year = str\_replace(year, "^", "20")) %>%   
 mutate(date = make\_date(year, month\_number, day)) %>%   
 select(date, total\_dewormed, diarrhea\_cases, acute\_malnutrition,  
 stunted\_6\_23\_months, stunted\_0\_6\_months, stunted\_24\_59\_months,  
 underweight\_0\_6\_months, underweight\_6\_23\_months, underweight\_24\_59\_months) %>%   
 arrange(date)  
  
line\_data %>% head()

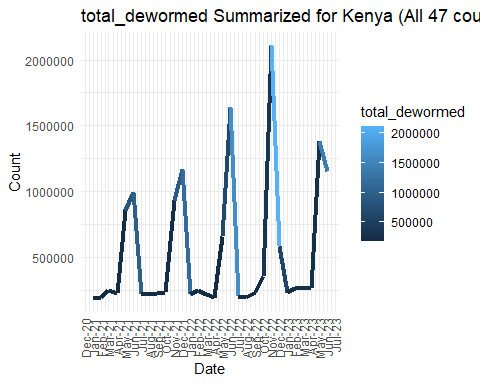
## # A tibble: 6 × 10  
## date total\_dewormed diarrhea\_cases acute\_malnutrition  
## <date> <dbl> <dbl> <dbl>  
## 1 2021-01-01 186487 94327 1500  
## 2 2021-02-01 196730 119174 2262  
## 3 2021-03-01 252827 143195 2485  
## 4 2021-04-01 227051 111542 3083  
## 5 2021-05-01 855315 127516 1502  
## 6 2021-06-01 996427 128450 1973  
## # ℹ 6 more variables: stunted\_6\_23\_months <dbl>, stunted\_0\_6\_months <dbl>,  
## # stunted\_24\_59\_months <dbl>, underweight\_0\_6\_months <dbl>,  
## # underweight\_6\_23\_months <dbl>, underweight\_24\_59\_months <dbl>

summarized\_data <- line\_data %>%  
 group\_by(date) %>%  
 summarize(  
 total\_dewormed = sum(total\_dewormed),  
 diarrhea\_cases = sum(diarrhea\_cases),  
 acute\_malnutrition = sum(acute\_malnutrition),  
 stunted\_6\_23\_months = sum(stunted\_6\_23\_months),  
 stunted\_0\_6\_months = sum(stunted\_0\_6\_months),  
 stunted\_24\_59\_months = sum(stunted\_24\_59\_months),  
 underweight\_0\_6\_months = sum(underweight\_0\_6\_months),  
 underweight\_6\_23\_months = sum(underweight\_6\_23\_months),  
 underweight\_24\_59\_months = sum(underweight\_24\_59\_months)  
 )  
view(summarized\_data)

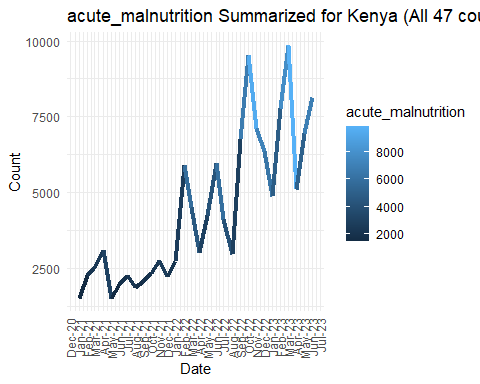
# Function to display the line trend of the health indicators from 2021 to 2023  
plot\_variable\_over\_time <- function(data, y\_variable) {  
 ggplot(data, aes(x = date, y = !!sym(y\_variable), color = !!sym(y\_variable))) +  
 geom\_line(linewidth = 1.5) +  
 labs(title = paste(y\_variable, "Summarized for Kenya (All 47 counties) Over Time"),  
 x = "Date",  
 y = "Count") +  
 scale\_x\_date(date\_labels = "%b-%y", date\_breaks = "1 month") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))  
}  
plot\_variable\_over\_time(summarized\_data, "diarrhea\_cases")



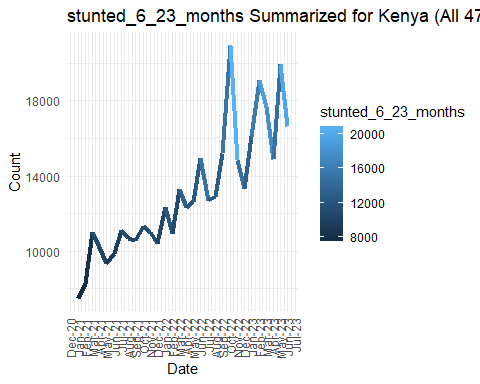
plot\_variable\_over\_time(summarized\_data, "total\_dewormed")



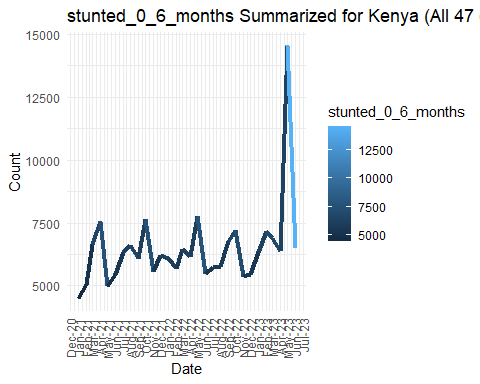
plot\_variable\_over\_time(summarized\_data, "acute\_malnutrition")



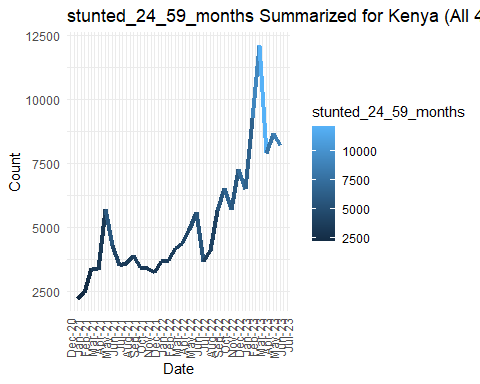
plot\_variable\_over\_time(summarized\_data, "stunted\_6\_23\_months")



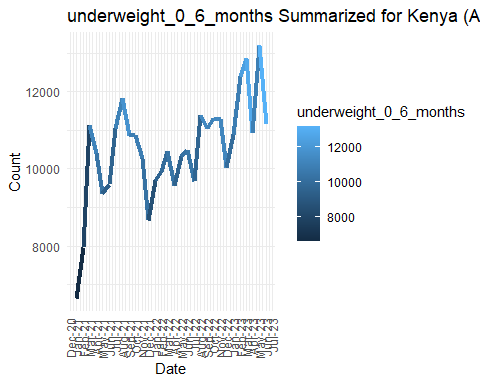
plot\_variable\_over\_time(summarized\_data, "stunted\_0\_6\_months")



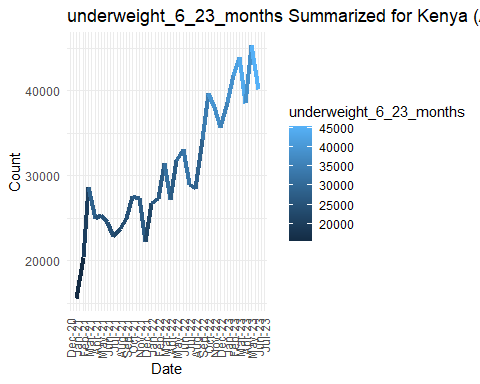
plot\_variable\_over\_time(summarized\_data, "stunted\_24\_59\_months")



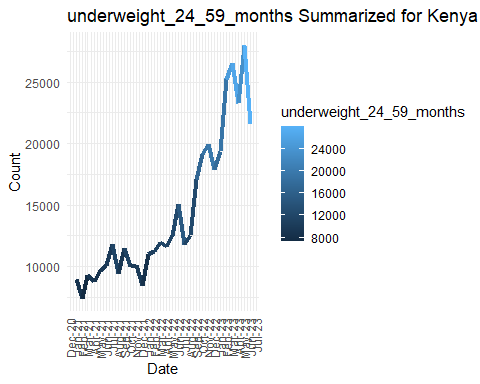
plot\_variable\_over\_time(summarized\_data, "underweight\_0\_6\_months")



plot\_variable\_over\_time(summarized\_data, "underweight\_6\_23\_months")



plot\_variable\_over\_time(summarized\_data, "underweight\_24\_59\_months")



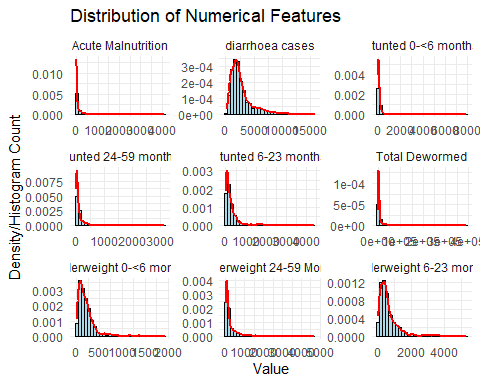
# 5. Exploratory Data Analysis(EDA)

## A.Univariate Analysis

### How are the numerical features distributions represented in the dataset?

plot\_histogram <- function(data) {  
 # Filter only the numerical columns  
 numerical\_data <- data %>% select\_if(is.numeric)  
  
 # Reshape the data for plotting  
 numerical\_data\_long <- numerical\_data %>%  
 pivot\_longer(everything(), names\_to = "variable", values\_to = "value")  
  
 # Plot Histogram and KDE  
 ggplot(numerical\_data\_long, aes(x = value)) +  
 geom\_histogram(aes(y = ..density..), fill = "lightblue", color = "black", bins = 30) +  
 geom\_density(color = "red", linewidth = 1) +  
 facet\_wrap(~ variable, scales = "free", ncol = 3) +  
 labs(title = "Distribution of Numerical Features",  
 x = "Value",  
 y = "Density/Histogram Count") +  
 theme\_minimal()  
}  
plot\_histogram(data)

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(density)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

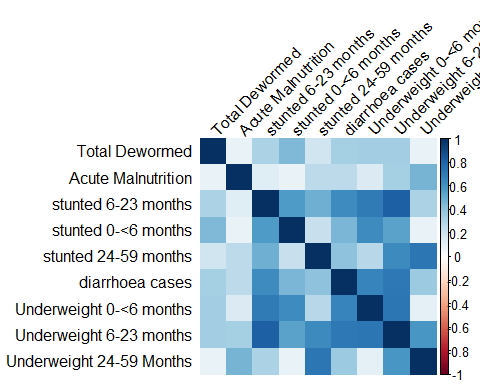


* Most of the numerical features exhibit a right-skewed distribution, indicating that the majority of data points have lower values, while a few extreme high values pull the distribution towards the right.
* There are some exceptions. The indicators “underweight 6-23 months,” “diarrhea cases,” and “underweight 0<6 months” exhibit a near-normal distribution. In these cases, the data is more balanced, with values spread symmetrically around the mean, indicating a more even distribution of numerical features.

## B.Bivariate Analysis

### How are health indicators correlated with each other, and which variables show significant associations?

# Correlation matrix  
correlation\_matrix <- cor(data[, c("Total Dewormed", "Acute Malnutrition", "stunted 6-23 months", "stunted 0-<6 months", "stunted 24-59 months", "diarrhoea cases", "Underweight 0-<6 months", "Underweight 6-23 months", "Underweight 24-59 Months")], use = "pairwise.complete.obs")  
  
# Create the correlation plot with upper and left labels switched  
corrplot(correlation\_matrix, method = "color", tl.col = "black", tl.srt = 45, tl.pos = "lt")



# Report summary for the Correlation plot  
correlation\_table <- round(correlation\_matrix, 2)  
# Print the table  
kable(correlation\_table, format = "html", caption = "Correlation Matrix")

Correlation Matrix

Total Dewormed

Acute Malnutrition

stunted 6-23 months

stunted 0-<6 months

stunted 24-59 months

diarrhoea cases

Underweight 0-<6 months

Underweight 6-23 months

Underweight 24-59 Months

Total Dewormed

1.00

0.09

0.32

0.44

0.19

0.33

0.35

0.35

0.10

Acute Malnutrition

0.09

1.00

0.13

0.09

0.27

0.26

0.15

0.34

0.47

stunted 6-23 months

0.32

0.13

1.00

0.56

0.48

0.62

0.70

0.81

0.32

stunted 0-<6 months

0.44

0.09

0.56

1.00

0.22

0.46

0.63

0.54

0.10

stunted 24-59 months

0.19

0.27

0.48

0.22

1.00

0.40

0.27

0.63

0.72

diarrhoea cases

0.33

0.26

0.62

0.46

0.40

1.00

0.66

0.71

0.37

Underweight 0-<6 months

0.35

0.15

0.70

0.63

0.27

0.66

1.00

0.73

0.12

Underweight 6-23 months

0.35

0.34

0.81

0.54

0.63

0.71

0.73

1.00

0.58

Underweight 24-59 Months

0.10

0.47

0.32

0.10

0.72

0.37

0.12

0.58

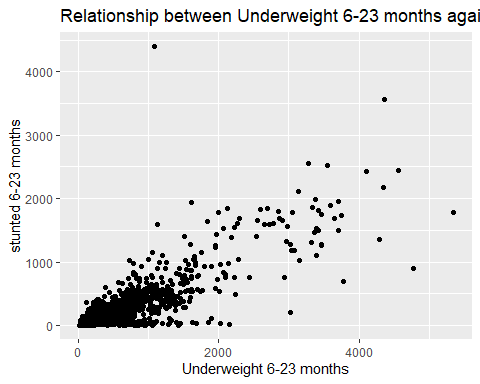
1.00

The analysis revealed strong positive correlations between certain health indicators in the dataset. Specifically:

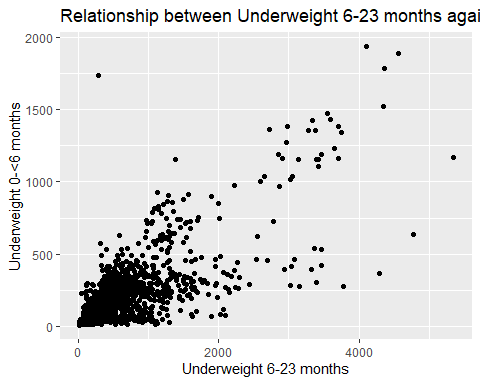
* The health indicators “Underweight 6-23 months” and “Stunted 6-23 months” exhibit the highest correlation of 0.81. This indicates that areas with higher rates of underweight children aged 6 to 23 months also tend to have higher rates of stunted children in the same age group.
* Following closely, “Underweight 6-23 months” and “Underweight 0-<6 months” show the second highest correlation of 0.73. This implies that regions with elevated levels of underweight children aged 6 to 23 months are also likely to have higher rates of underweight children aged less than 6 months.
* Additionally, the health indicators “Underweight 24-59 months” and “Stunted 24-59 months” demonstrate the third highest correlation of 0.72. This suggests that areas experiencing a higher prevalence of underweight children aged 24 to 59 months are also more likely to have increased rates of stunted children in the same age bracket.

### What is the relationship between the top 3 highest correlated(0.81,0.73 and 0.72) health indicators?

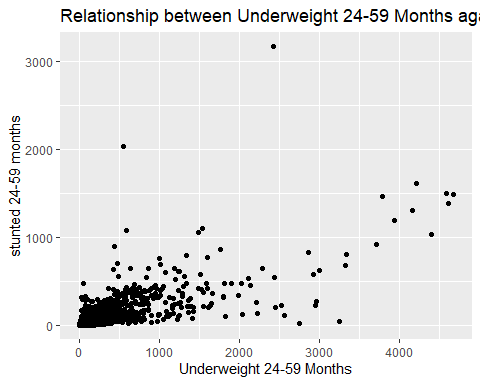
create\_scatter\_plot <- function(data, x\_column, y\_column) {  
 # scatterplot  
 ggplot(data, aes(x = !!sym(x\_column), y = !!sym(y\_column))) +  
 geom\_point() +  
 labs(title = paste("Relationship between", x\_column, "against", y\_column),  
 x = x\_column,  
 y = y\_column)  
}  
create\_scatter\_plot(data, "Underweight 6-23 months", "stunted 6-23 months")



create\_scatter\_plot(data, "Underweight 6-23 months", "Underweight 0-<6 months")



create\_scatter\_plot(data, "Underweight 24-59 Months", "stunted 24-59 months")



* For each plot, we observe a noticeable trend that indicates an increase or underlying pattern. In the first scatter plot, “Underweight 6-23 months” and “stunted 6-23 months” exhibit a positive relationship. As the values of “Underweight 6-23 months” increase, we also observe an increase in the values of “stunted 6-23 months.” This pattern suggests that there might be a connection between underweight and stunted growth for children in the 6-23 months age group.
* Similarly, in the second scatter plot, “Underweight 6-23 months” and “Underweight 0-<6 months” also show a positive relationship. As the values of “Underweight 6-23 months” increase, there is a corresponding increase in “Underweight 0-<6 months.” This trend indicates a link between underweight in the two age groups, highlighting a potential health concern among children under 6 months.
* Lastly, in the third scatter plot, “Underweight 24-59 Months” and “stunted 24-59 months” exhibit a positive correlation. As the values of “Underweight 24-59 Months” increase, there is a simultaneous increase in “stunted 24-59 months.” This observation suggests that underweight in children aged 24-59 months might be associated with stunted growth.

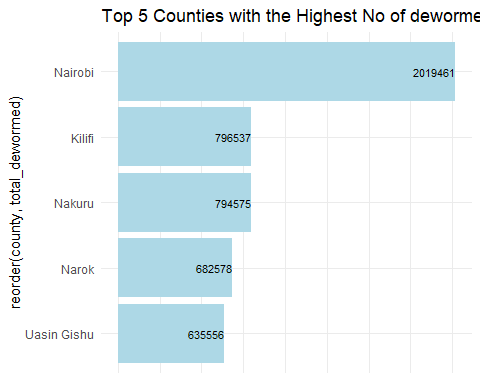
county\_data <- data %>%  
 group\_by(county,full\_date) %>%  
 summarise(  
 total\_dewormed = sum(`Total Dewormed`, na.rm = TRUE),  
 diarrhea\_cases = sum(`diarrhoea cases`, na.rm = TRUE),  
 acute\_malnutrition = sum(`Acute Malnutrition`, na.rm = TRUE),  
 stunted\_6\_23\_months = sum(`stunted 6-23 months`, na.rm = TRUE),  
 stunted\_0\_6\_months = sum(`stunted 0-<6 months`, na.rm = TRUE),  
 stunted\_24\_59\_months = sum(`stunted 24-59 months`, na.rm = TRUE),  
 underweight\_0\_6\_months = sum(`Underweight 0-<6 months`, na.rm = TRUE),  
 underweight\_6\_23\_months = sum(`Underweight 6-23 months`, na.rm = TRUE),  
 underweight\_24\_59\_months = sum(`Underweight 24-59 Months`, na.rm = TRUE),  
 .groups = "drop"  
  
 ) %>%  
 select(county,full\_date,total\_dewormed, diarrhea\_cases, acute\_malnutrition,  
 stunted\_6\_23\_months, stunted\_0\_6\_months, stunted\_24\_59\_months,  
 underweight\_0\_6\_months, underweight\_6\_23\_months, underweight\_24\_59\_months) %>%   
 arrange(county)  
  
county\_data %>% head()

## # A tibble: 6 × 11  
## county full\_date total\_dewormed diarrhea\_cases acute\_malnutrition  
## <chr> <date> <dbl> <dbl> <dbl>  
## 1 Baringo 2021-01-28 1917 895 4  
## 2 Baringo 2021-02-28 4376 1599 1  
## 3 Baringo 2021-03-28 4291 2331 8  
## 4 Baringo 2021-04-28 3306 1847 4  
## 5 Baringo 2021-05-28 13782 2131 2  
## 6 Baringo 2021-06-28 24348 2705 0  
## # ℹ 6 more variables: stunted\_6\_23\_months <dbl>, stunted\_0\_6\_months <dbl>,  
## # stunted\_24\_59\_months <dbl>, underweight\_0\_6\_months <dbl>,  
## # underweight\_6\_23\_months <dbl>, underweight\_24\_59\_months <dbl>

# Filter data for years 2021, 2022, and 2023  
county\_data$year <- year(county\_data$full\_date)  
  
# Group the filtered data by 'full\_date' and 'county', and then summarize  
summarized\_county\_data <- county\_data %>%  
 group\_by(year, county) %>%  
 summarize(  
 total\_dewormed = sum(`total\_dewormed`),  
 diarrhea\_cases = sum(`diarrhea\_cases`),  
 acute\_malnutrition = sum(`acute\_malnutrition`),  
 stunted\_6\_23\_months = sum(`stunted\_6\_23\_months`),  
 stunted\_0\_6\_months = sum(`stunted\_0\_6\_months`),  
 stunted\_24\_59\_months = sum(`stunted\_24\_59\_months`),  
 underweight\_0\_6\_months = sum(`underweight\_0\_6\_months`),  
 underweight\_6\_23\_months = sum(`underweight\_6\_23\_months`),  
 underweight\_24\_59\_months = sum(`underweight\_24\_59\_months`),  
 .groups = "drop"  
 )  
view(summarized\_county\_data)

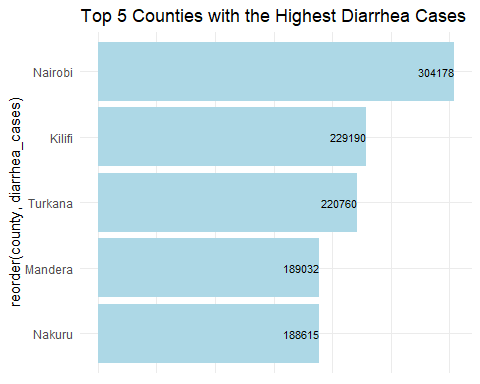
### What are the top 5 counties with the highest ‘Total Dewormed’ cases

top\_5\_dewormed\_counties <- data %>% group\_by(county) %>% summarise(  
 total\_dewormed = sum(`Total Dewormed`)) %>% arrange(desc(total\_dewormed)) %>%  
 slice\_max(total\_dewormed,n=5)  
# barplot  
ggplot(top\_5\_dewormed\_counties, aes(x = reorder(county, total\_dewormed), y = total\_dewormed)) +  
 geom\_bar(stat = "identity", fill = "lightblue") +  
 geom\_text(aes(label = total\_dewormed), hjust = 1, color = "black", size = 3) +   
 labs(title = "Top 5 Counties with the Highest No of dewormed childern",  
 y = "Total Dewormed Children") +  
 theme\_minimal() +  
 coord\_flip() +  
 theme(axis.text.x = element\_blank(),   
 axis.title.x = element\_blank())



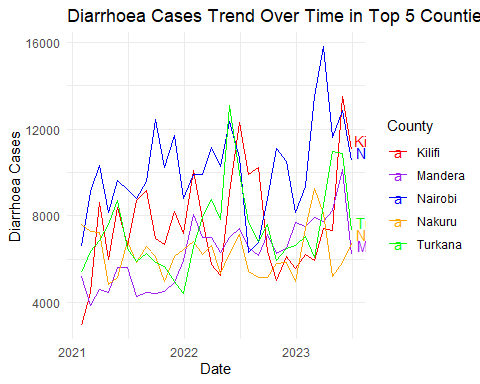
### What are the top 5 counties with the most diarrhoea cases?

top\_5\_counties <- data %>% group\_by(county) %>% summarise(  
 diarrhea\_cases = sum(`diarrhoea cases`)) %>% arrange(desc(diarrhea\_cases)) %>%  
 slice\_max(diarrhea\_cases,n=5)  
# barplot  
ggplot(top\_5\_counties, aes(x = reorder(county, diarrhea\_cases), y = diarrhea\_cases)) +  
 geom\_bar(stat = "identity", fill = "lightblue") +  
 geom\_text(aes(label = diarrhea\_cases), hjust = 1, color = "black", size = 3) +   
 labs(title = "Top 5 Counties with the Highest Diarrhea Cases",  
 y = "Diarrhea Cases") +  
 theme\_minimal() +  
 coord\_flip() +  
 theme(axis.text.x = element\_blank(),   
 axis.title.x = element\_blank())



### How does the number of Diarrhoea Cases vary over time in the top 5 counties?

selected\_counties <- c("Nairobi", "Kilifi", "Turkana", "Mandera", "Nakuru")  
selected\_data <- data %>%  
 filter(county %in% selected\_counties)  
  
# Create a line plot for each county  
ggplot(selected\_data, aes(x = full\_date, y = `diarrhoea cases`, color = county)) +  
 geom\_line() +  
 labs(title = "Diarrhoea Cases Trend Over Time in Top 5 Counties",  
 x = "Date",  
 y = "Diarrhoea Cases",  
 color = "County") +  
 theme\_minimal() +  
 scale\_color\_manual(values = c("Nairobi" = "blue", "Kilifi" = "red", "Turkana" = "green", "Mandera" = "purple", "Nakuru" = "orange")) +  
 geom\_text(data = selected\_data %>% filter(full\_date == max(full\_date)),  
 aes(label = county, color = county),  
 hjust = -0.1,  
 vjust = -0.2,  
 size = 4)



## C. Mutivariate Analysis

### Working with Geospatial Data

# Reading the .dbf file  
dbf\_data <- read.dbf("shapefiles/County.dbf")  
  
# View the data  
head(dbf\_data)

## fid OBJECTID ID Name Code Shape\_Leng Shape\_Area Area  
## 1 1 1 1 Mombasa MBA 0.8855862 0.02332511 286423166  
## 2 2 2 2 Kwale KLE 4.2841818 0.75826601 9309279431  
## 3 3 3 3 Kilifi KLF 5.3330801 1.02533838 12601873866  
## 4 4 4 4 Tana River TAN 10.2804494 3.18421264 39177464255  
## 5 5 5 5 Lamu LAU 3.7446892 0.74374311 9148878656  
## 6 6 6 6 Taita Taveta TVT 5.5844513 1.39385877 17126899316

# Reading the .shp file  
shp <- st\_read("shapefiles/County.shp")

## Reading layer `County' from data source   
## `D:\R studio work\internship\_task-main\shapefiles\County.shp'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 47 features and 8 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: 33.91028 ymin: -4.798828 xmax: 41.90613 ymax: 5.414124  
## Geodetic CRS: WGS 84

# View the data  
head(shp,2)

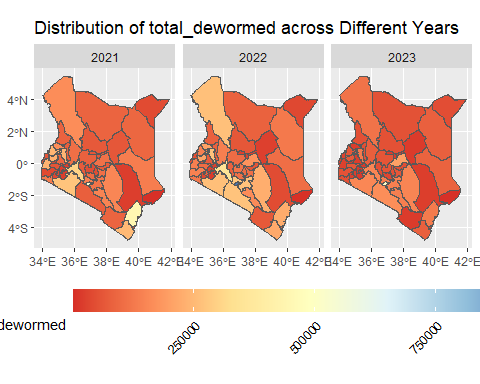
## Simple feature collection with 2 features and 8 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: 38.44611 ymin: -4.798828 xmax: 39.76147 ymax: -3.564514  
## Geodetic CRS: WGS 84  
## fid OBJECTID ID Name Code Shape\_Leng Shape\_Area Area  
## 1 1 1 1 Mombasa MBA 0.8855862 0.02332511 286423166  
## 2 2 2 2 Kwale KLE 4.2841818 0.75826601 9309279431  
## geometry  
## 1 MULTIPOLYGON (((39.6825 -4....  
## 2 MULTIPOLYGON (((39.32031 -3...

### Summarizing our dataset to each county

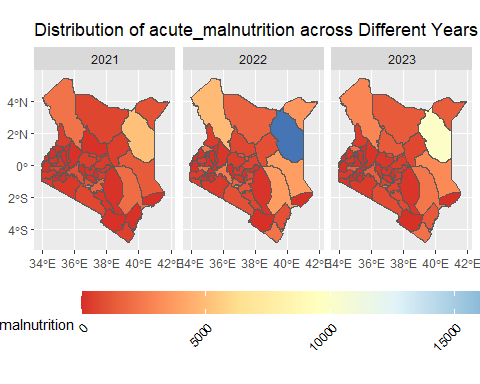
merged\_data <- left\_join(shp, summarized\_county\_data, by = c("Name" = "county"))  
view(merged\_data)

### How does the distribution of Total dewormed, Acute Malnutrition and diarrhoea cases vary across different years within the map?

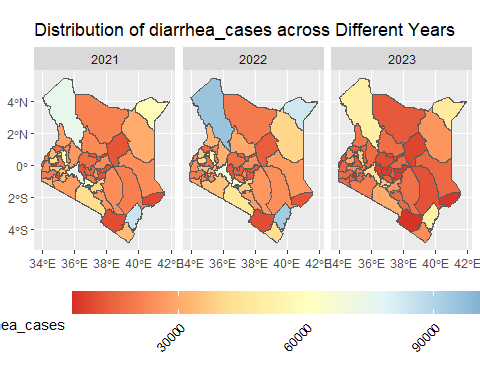
# Function to create a choropleth map for any of the health indicators  
create\_map <- function(data, variable\_name) {  
 ggplot() +  
 geom\_sf(data = data, aes(fill = !!sym(variable\_name), geometry = geometry)) +  
 scale\_fill\_distiller(palette = "RdYlBu", direction = 1) +   
 facet\_wrap(~year, ncol = 3) + # Using the 'year' column directly  
 labs(title = paste("Distribution of", variable\_name, "across Different Years"),  
 fill = variable\_name) +  
 theme(legend.position = "bottom",  
 legend.key.width = unit(2.5, "cm"),   
 legend.text = element\_text(angle = 45, hjust = 1))   
}  
  
# Map Distributions for each Health Indicator"  
create\_map(merged\_data, "total\_dewormed")



create\_map(merged\_data, "acute\_malnutrition")

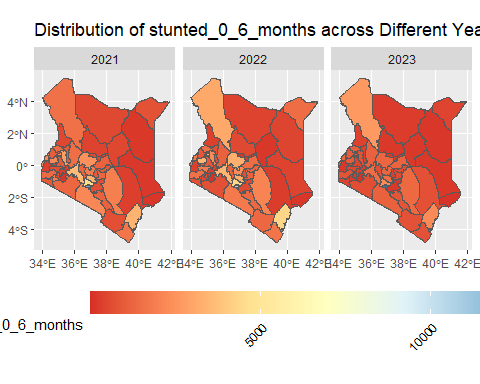


create\_map(merged\_data, "diarrhea\_cases")

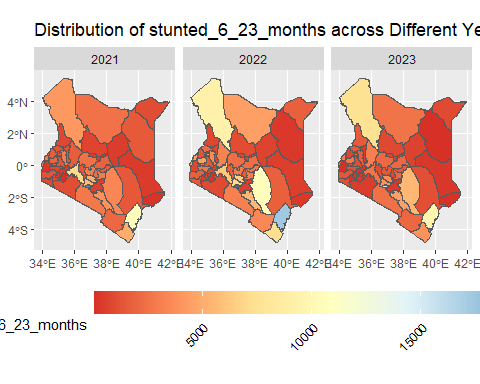


### How does the distribution of stunted groups of Children vary across different years within the map?

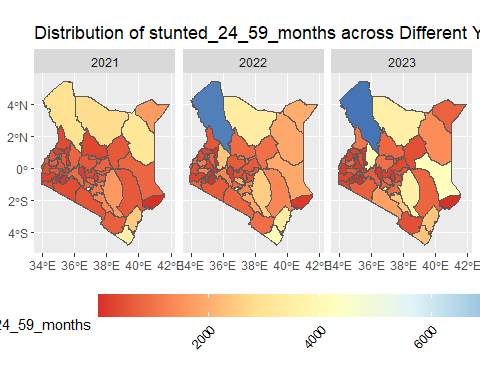
create\_map(merged\_data, "stunted\_0\_6\_months")



create\_map(merged\_data, "stunted\_6\_23\_months")

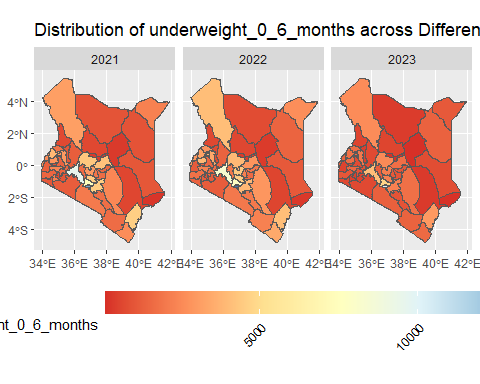


create\_map(merged\_data, "stunted\_24\_59\_months")

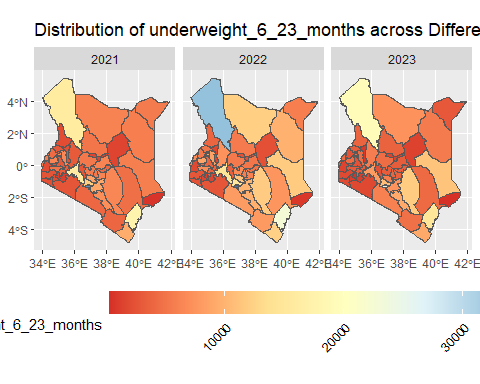


### How does the distribution of underweight groups of children vary across different years within the map?

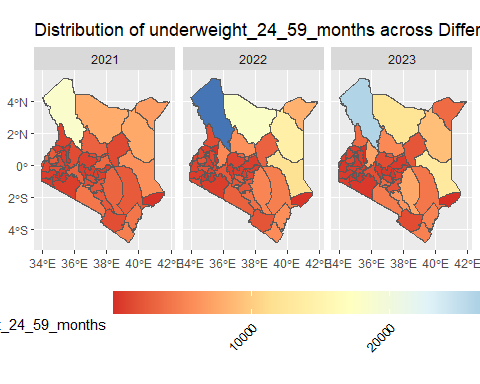
create\_map(merged\_data, "underweight\_0\_6\_months")



create\_map(merged\_data, "underweight\_6\_23\_months")



create\_map(merged\_data, "underweight\_24\_59\_months")



## 6. Conclusion

## Recommendations and Future Improvement ideas

I. Data Source Expansion:

* Consider incorporating additional data sources, such as demographic information, economic data, or environmental factors, to gain a more comprehensive understanding of health indicators’ determinants.

1. Collaboration:

* Collaborate with domain experts, health professionals, and other stakeholders to gain valuable insights into the data and the context of health-related issues.

1. Interactive Visualizations:

* Explore the use of interactive visualizations to allow users to interact with the data and gain more insights dynamically.

1. Automated Reports:

* Develop a mechanism to generate automated reports that summarize key findings and visualizations for different health indicators and counties.

V. Advanced Exploratory Analysis(EDA)

* Conduct an in-depth analysis on the remaining features in our dataset.