CEMA Internship Task

Humphrey Kinoti

2023-07-21

A copy of this project is in: link to my github repo

Instructions

You have been provided with a dataset which contains monthly data for children <5 years, disaggregated at a county level for the period January 2021 to June 2023.

Dataset description

The dataset contains the following variables:

- `period`: months from January 2021 to June 2023
- `county`: the 47 counties in Kenya
- `Total Dewormed`: Total number of children dewormed
- `Acute Malnutrition`: Number of children <5 years with acute malnutrition
- `stunted 0-6 months`, `stunted 6-23 months`, `stunted 24-59 months`: Number of children stunted
- `diarrhea cases`: Number of children <5 years with diarrhea
- `underweight 0-6 months`, `underweight 6-23 months`, `underweight 24-59 months`: Number of children who are underweight

Objectives

Your task is to: -

- Conduct exploratory data analysis
- State an appropriate research question you would want to answer from the data
- Carry out appropriate data analysis to address the question you have stated above

Explanatory Data Analysis

Import Data

As the data is in CSV format, we will import data using `read_csv()` function in the `readr` package.

```
library(tidyverse)
cema_data <- read_csv("data/cema_internship_task_2023.csv")
```

After the importation, Let's have a look of how the data is structured to have a better understanding of the data. In this case, I will use the `skim` function from `skimr` package.

```r
library(skimr)
skim(cema_data) %>%
  dplyr::select(skim_type, skim_variable, n_missing,
                character.min, character.max, character.empty,
                character.n_unique,
                numeric.mean, numeric.sd, numeric.p0, numeric.p25, numeric.p50, numeric.p75, numeric.p1
```

Table 1: Data summary

Name	cema_data
Number of rows	1410
Number of columns	11
Column type frequency:	
character	2
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	min	max	empty	n_unique
period	0	6	6	0	30
county	0	11	22	0	47

Variable type: numeric

skim_variable	n_missing	mean	sd	p0	p25	p50	p75	p100
Total Dewormed	0	11457.92	25372.43	97	2454.50	4564.5	8222.50	392800
Acute Malnutrition	355	125.40	266.49	1	15.00	39.0	143.50	4123
stunted 6-23 months	11	280.16	380.55	1	69.50	159.0	328.50	4398
stunted 0-<6 months	19	139.79	280.24	1	36.50	84.0	157.00	7900
stunted 24-59 months	14	110.77	193.40	1	22.00	50.0	114.25	3169
diarrhoea cases	0	2813.38	2161.90	198	1464.25	2158.0	3335.25	15795
Underweight 0-<6 months	0	223.47	228.53	6	87.00	162.5	272.75	1937
Underweight 6-23 months	0	652.26	669.58	16	249.00	456.0	791.75	5348
Underweight 24-59 Months	0	305.74	538.46	1	51.25	120.5	311.00	4680

The above tables have some summary statistics of all the columns in the dataset and the nature missing values. However, the column names have some white spaces which will cause some tedious work during analysis. Lets replace the white spaces with an underscore(_). We will not remove the missing variables since we will lose some important data for other columns that are not missing.

Moreover, the `period` column has dates data but it is of character datatype. We need to convert this column to date format, so that we can use it in creating time series plots to visualize the trends and patterns over time.

Data Cleaning

Lets rename the column names

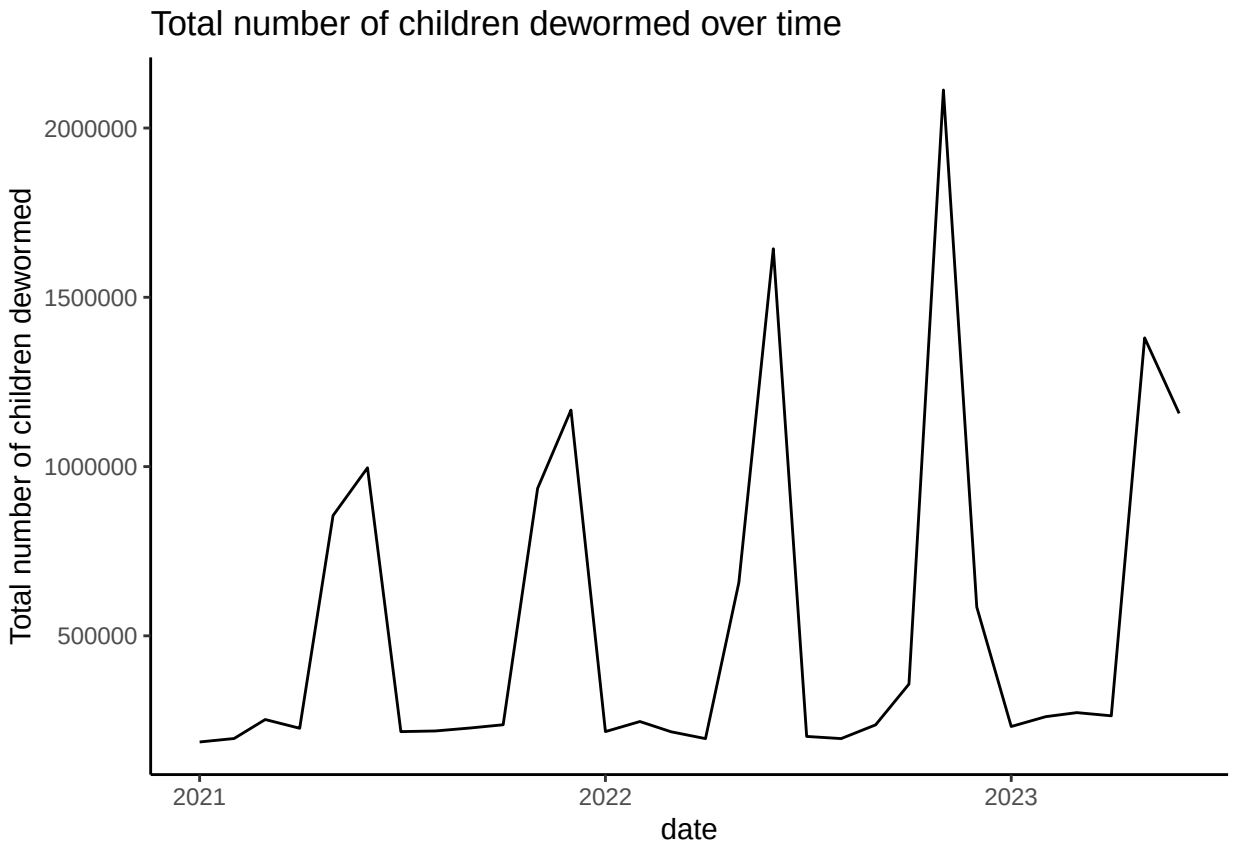
```
new_colnames <- c("period","county","total_dewormed","acute_malnutrition",
                  "stunted_6-23_months", "stunted_0-<6_months", "stunted_24-59_months",
                  "diarrhoea_cases", "underweight_0-<6_months", "underweight_6-23_months",
                  "underweight_24-59_months")
#create a copy of cema_data
child_data <- cema_data
names(child_data) <- new_colnames
#check if the colnames were updated
#colnames(child_data)
#Adding a date column
child_data$date <- my(child_data$period)
summary(child_data$date)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2021-01-01" "2021-08-01" "2022-03-16" "2022-03-17" "2022-11-01" "2023-06-01"
```

Data visualization

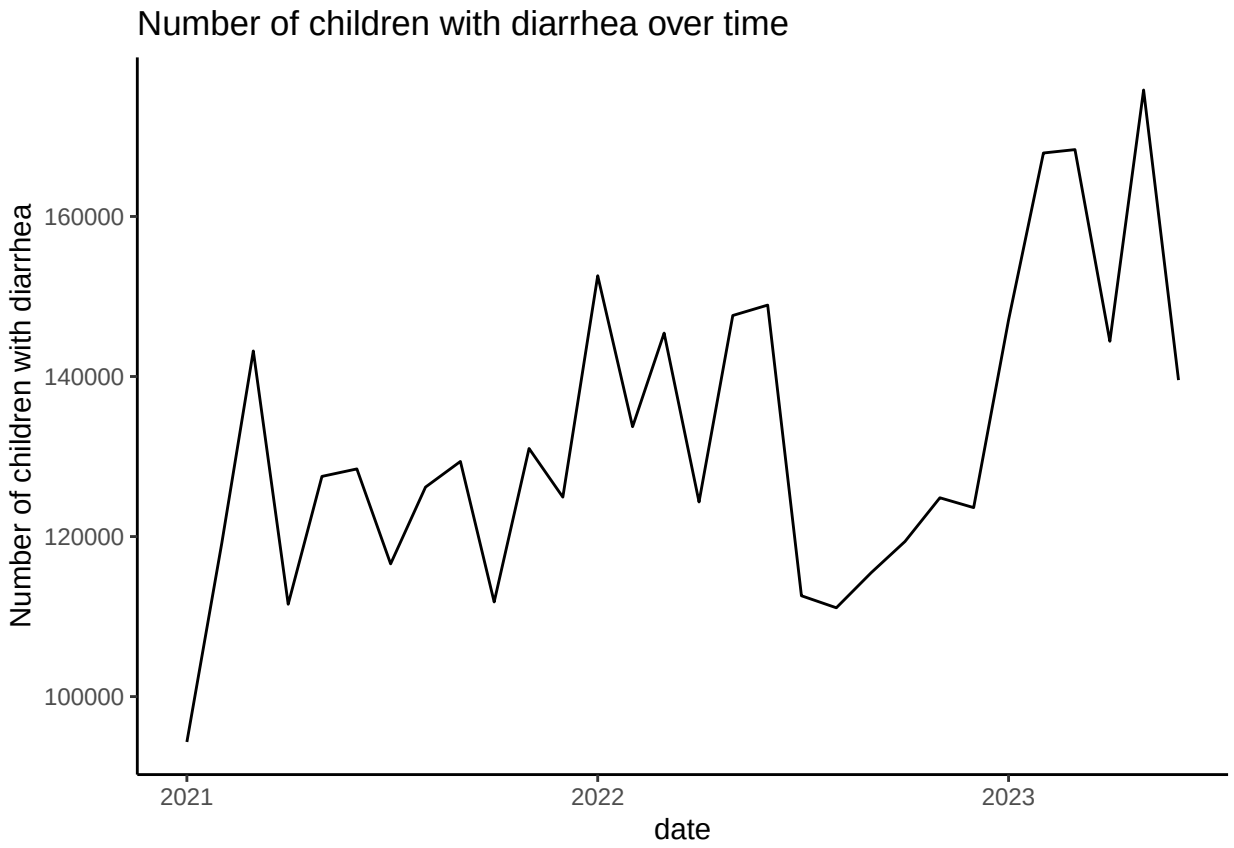
Since all the continuous variables are discrete, line plot is the best. Since we have the date column, we can use it in creating time series plots to visualize the trends and patterns over time. There are different trends that we can check: a. The trend of each variable over time. In this case, we To do this, I will aggregate the dataset, grouping with time.

```
#summing all colums by period
time_agg <- as.data.frame(child_data) %>%
  dplyr::select(-c(county,date)) %>%
  group_by(period) %>%
  summarise(across(.cols = is.numeric, .fns = sum, na.rm = TRUE))
#Adding the date column
time_agg$date <- my(time_agg$period)
#sorting by date column
time_agg <- time_agg %>%
  arrange(date)

deworm_data= time_agg %>%
  dplyr::select(period,date ,total_dewormed)

#visualize using a line plot
deworm_data %>%
  ggplot(aes(x= date, y= total_dewormed))+
  ggtitle("Total number of children dewormed over time")+
  xlab("date")+
  ylab("Total number of children dewormed")+
  geom_line()+
  theme_classic()
```



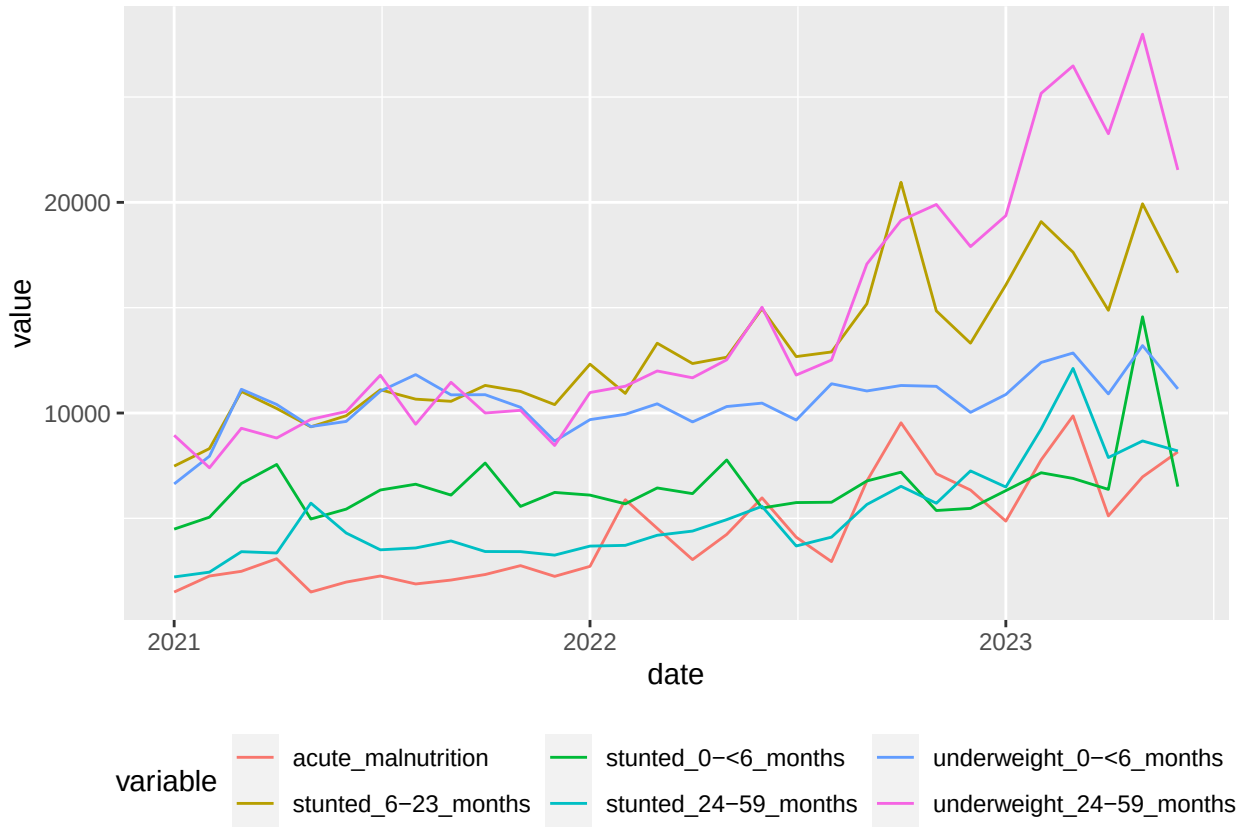
From the above plot, we can answer the following research question: - How has the deworming program's effectiveness evolved over time from January 2021 to June 2023? - Are there any notable trends or patterns in the total number of children dewormed over the months?

```
diarr_data= time_agg %>%
  dplyr::select(period,date ,diarrhoea_cases)

#visualize using a line plot
diarr_data %>%
  ggplot(aes(x= date, y= diarrhoea_cases))+
  ggtitle("Number of children with diarrhea over time")+
  xlab("date")+
  ylab("Number of children with diarrhea")+
  geom_line()+
  theme_classic()
```



This plot can help us answer the following research question: - Has there been a seasonal pattern in the prevalence of diarrhea cases over the study period?

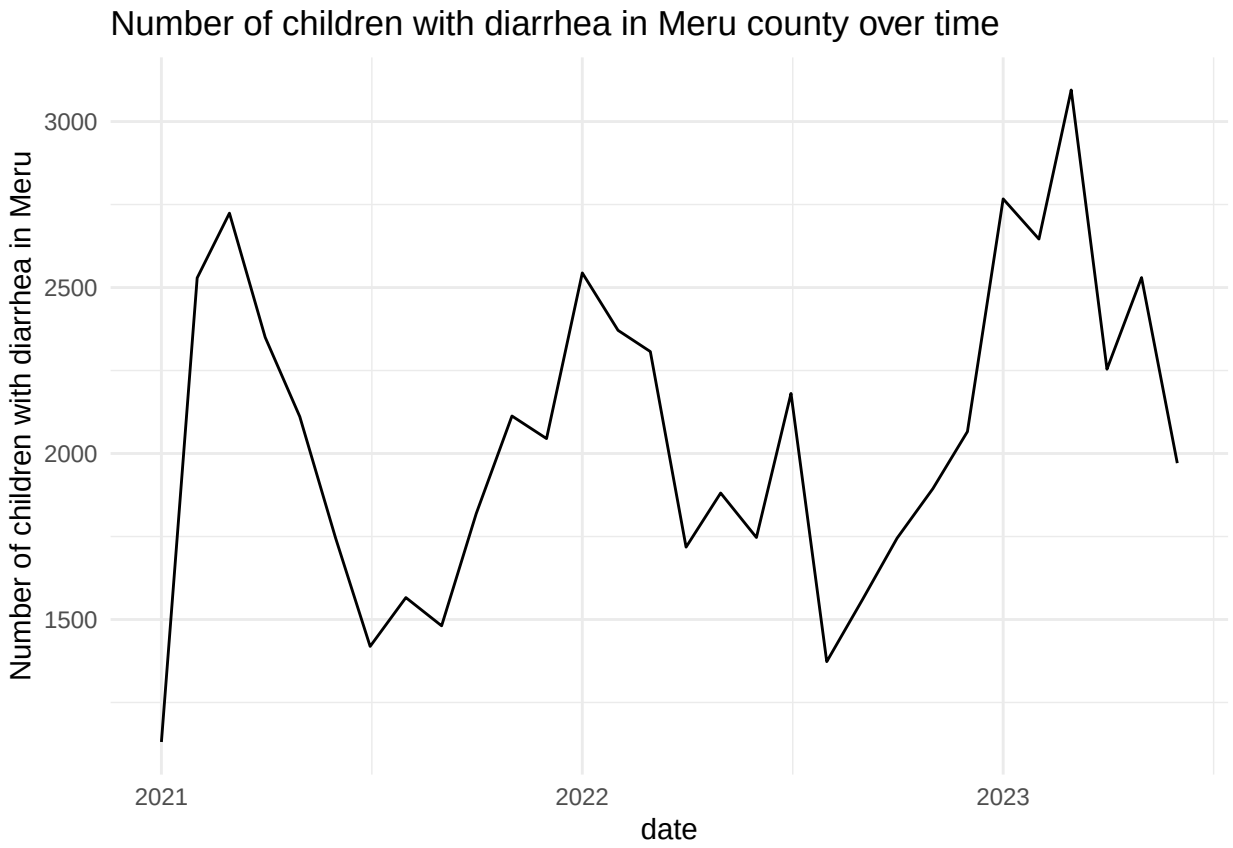
```
all_other <- time_agg %>%
  dplyr::select(-c(total_dewormed, diarrhoea_cases,`underweight_6-23_months`))
library(reshape2)
time_agg_long <- melt(all_other, id=c("period","date"))

time_agg_long %>%
  ggplot(aes(x= date, y= value, color= variable))+
  geom_line()+
  theme(legend.position = "bottom")
```



This plot can help us answer the following research question: - Has there been a seasonal pattern in the prevalence of acute malnutrition cases over the study period?

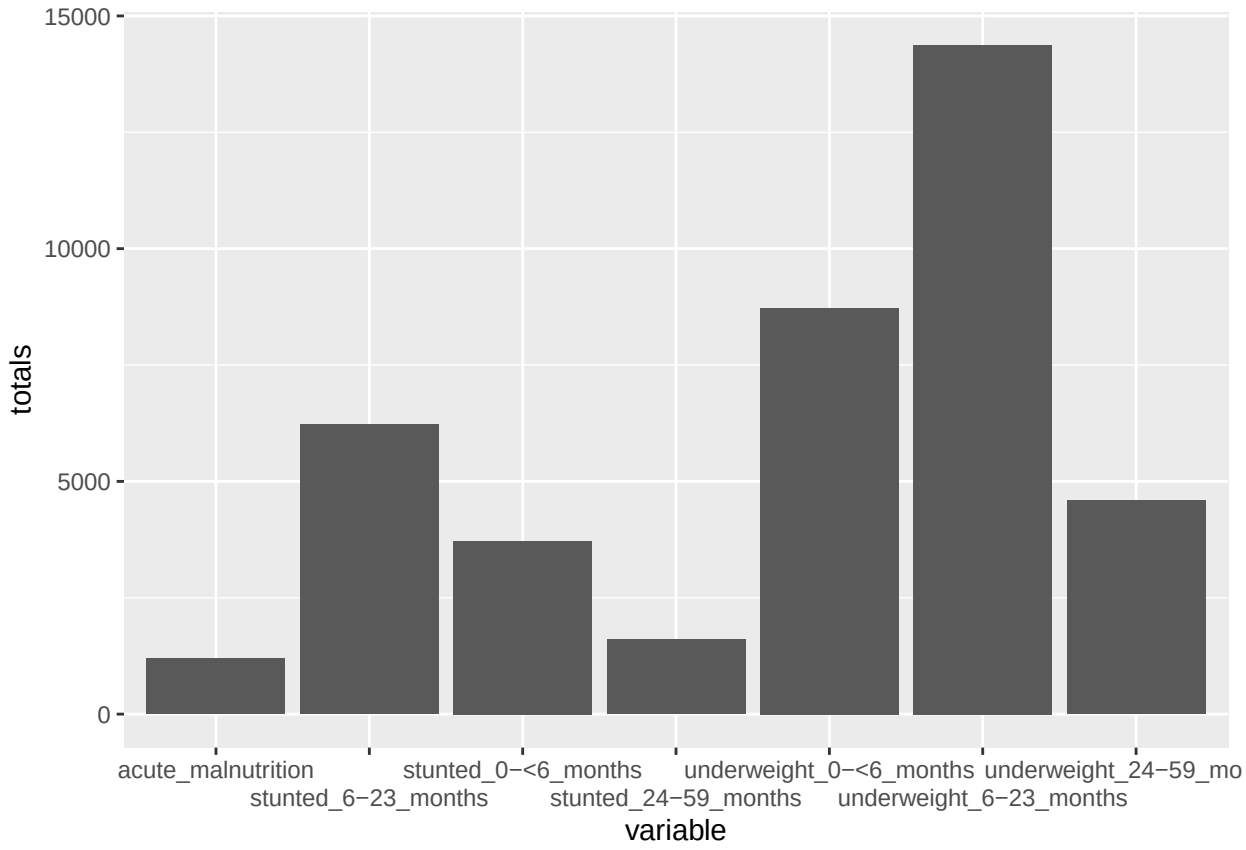
b. The trend of each variable over time in each county. For example, we can check the trend of number of children with diarrhea over time in Meru county

```
meru_diarr <- as.data.frame(child_data) %>%
  dplyr::select(period,county ,diarrhoea_cases) %>%
  dplyr::filter(county == "Meru County")
#Add time variable
meru_diarr$date <- my(meru_diarr$period)
#sort by date
meru_diarr <- meru_diarr %>% arrange(date)

#line plot
meru_diarr %>%
  ggplot(aes(x= date, y= diarrhoea_cases))+
  ggtitle("Number of children with diarrhea in Meru county over time")+
  xlab("date")+
  ylab("Number of children with diarrhea in Meru")+
  geom_line()+
  theme_minimal()
```



Geospatial Analysis First, we can have a simple bar plot to see the total distribution of all variables per county. For example in Meru county

```
county_agg <- as.data.frame(child_data) %>%
  dplyr::select(-c(date, period)) %>%
  group_by(county) %>%
  summarise(across(.cols = is.numeric, .fns = sum, na.rm = TRUE))

county_long = melt(county_agg, id = "county")
result <- county_long %>%
    filter(county == "Meru County") %>%
    group_by(variable) %>%
    summarise(totals= sum(value)) %>%
  filter(variable != c("total_dewormed","diarrhoea_cases"))

ggplot(result, aes(x = variable, y = totals)) +
  geom_bar(stat = "identity")+
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  NULL
```



Importing the shapefiles

```
library(sf)
sf_data <- read_sf("shapefiles/County.shp")
head(sf_data)
```

```
## Simple feature collection with 6 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 37.58447 ymin: -4.798828 xmax: 41.56909 ymax: -0.001525879
## Geodetic CRS:  WGS 84
## # A tibble: 6 x 9
##     fid OBJECTID    ID Name         Code  Shape_Leng Shape_Area         Area
##   <dbl>    <dbl> <dbl> <chr>        <chr>      <dbl>      <dbl>        <dbl>
## 1     1        1     1 Mombasa      MBA        0.886     0.0233   286423166.
## 2     2        2     2 Kwale        KLE        4.28      0.758   9309279431.
## 3     3        3     3 Kilifi       KLF        5.33      1.03   12601873866.
## 4     4        4     4 Tana River   TAN       10.3       3.18   39177464255.
## 5     5        5     5 Lamu         LAU        3.74      0.744   9148878656.
## 6     6        6     6 Taita Taveta TVT        5.58      1.39   17126899316.
## # i 1 more variable: geometry <MULTIPOLYGON [°]>
```

Make a simple plot

```
sf_data %>%
    ggplot() +
  geom_sf(aes(fill = Name),show.legend = F)+
   theme(legend.text.align = 1,
         legend.title.align = 0.5,
```

```
        plot.title = element_text(hjust = .1,vjust=0.2,size=18,family="Times"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.text.x=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks=element_blank(),
        axis.title.x=element_blank(),
        axis.title.y=element_blank())+
  geom_sf_text(aes(label= Code), size= 1.5)
```

Merging the dataframes

```
county_agg$county = gsub(" County", "", county_agg$county)
merged_sf <- merge(sf_data, county_agg, by.x= "Name", by.y= "county")
```

Now we can get insights of the dataframe since we have merged it.

Research Question

Is there a significant relationship between the number of children dewormed monthly and the Child Health Indicators (prevalence of acute malnutrition, stunting, underweight, or diarrhea cases) among children under 5 years old in the country?

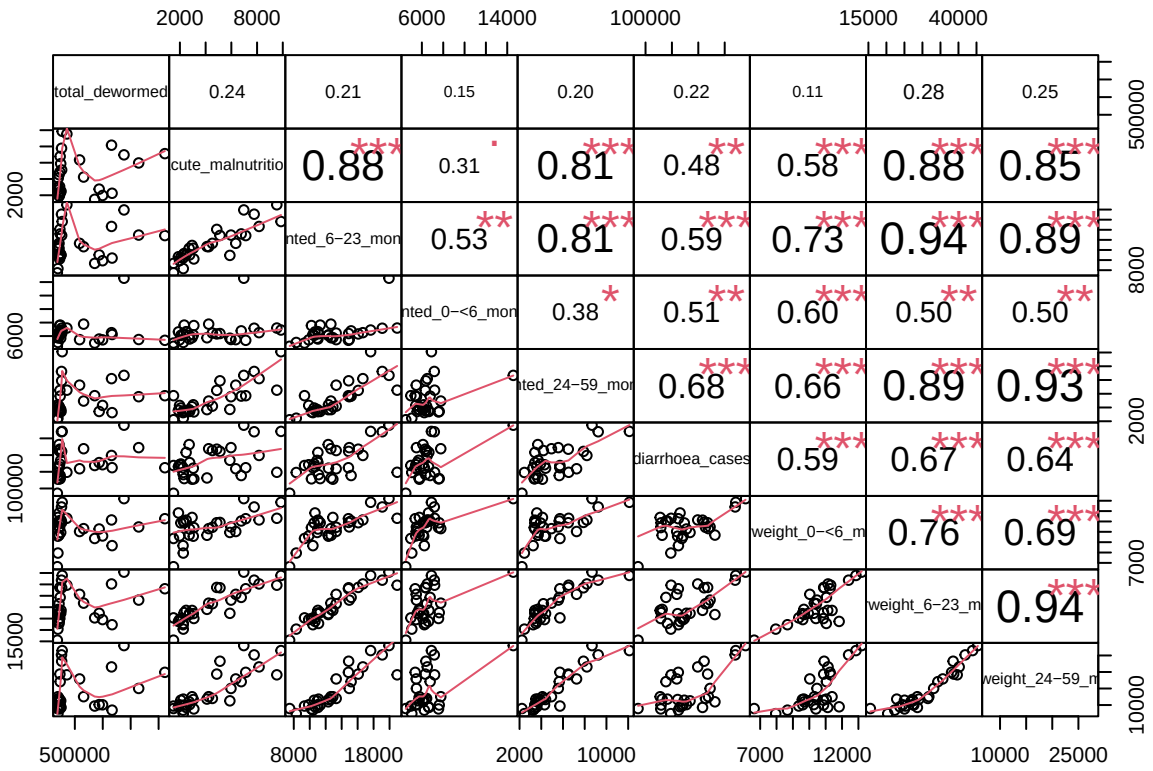
Solution To answer this question, we need to perform a Correlation Analysis. We will compute correlation coefficients (i.e., Pearson correlation) to assess the relationships between the number of children dewormed and child health indicators like acute malnutrition, stunting, underweight, and diarrhea cases in Kenya. However, another research should be conducted to control for each county level.

We need the monthly dataset for the aforementioned variables:

```
cor_data <- time_agg %>%
  dplyr::select(-c(period, date))
```

Let's do the correlation analysis:

```
library(PerformanceAnalytics)
chart.Correlation(cor_data, method = "pearson", histogram = F, pch = 16)
```



There is a very weak insignificant relationship between the number of children dewormed monthly and all the Child Health indicators. This means that deworming has no effect on the health indicators. Some improvements should be made.