Edgar-Jandi-CEMA-Project-2023.R

r1138441

2023-07-21

#I USED R-STUDIO CLOUD IN THIS PROJECT  
#I UPLOADED THE DOWNLOADED CSV FILE DIRECTLY TO R-STUDIO WHICH I IMPORTED DIRECTLY TO THE ENVIRONMENT FOR VIEWING, CLEANING AND MANIPULATION  
#THEN I STARTED WRITING THE SCRIPT  
#INSTALLING NECESSARY PACKAGES  
install.packages("tidyverse")

## Error in install.packages : Updating loaded packages

library(tidyverse)  
install.packages("lubridate")

## Error in install.packages : Updating loaded packages

library(lubridate)  
  
# Load the dataset  
data <- cema\_internship\_task\_2023  
  
# Explore the structure of the dataset  
str(data)

## spc\_tbl\_ [1,410 × 11] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ period : chr [1:1410] "Jan-23" "Jan-23" "Jan-23" "Jan-23" ...  
## $ county : chr [1:1410] "Baringo County" "Bomet County" "Bungoma County" "Busia County" ...  
## $ Total Dewormed : num [1:1410] 3659 1580 6590 7564 1407 ...  
## $ Acute Malnutrition : num [1:1410] 8 NA 24 NA NA 72 250 9 26 104 ...  
## $ stunted 6-23 months : num [1:1410] 471 1 98 396 92 326 40 209 51 319 ...  
## $ stunted 0-<6 months : num [1:1410] 34 3 154 143 71 86 13 87 6 102 ...  
## $ stunted 24-59 months : num [1:1410] 380 NA 23 111 5 24 99 58 50 155 ...  
## $ diarrhoea cases : num [1:1410] 2620 1984 4576 2239 2739 ...  
## $ Underweight 0-<6 months : num [1:1410] 85 41 231 251 57 141 223 140 13 139 ...  
## $ Underweight 6-23 months : num [1:1410] 739 86 315 608 104 ...  
## $ Underweight 24-59 Months: num [1:1410] 731 16 120 125 21 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. period = col\_character(),  
## .. county = col\_character(),  
## .. `Total Dewormed` = col\_double(),  
## .. `Acute Malnutrition` = col\_double(),  
## .. `stunted 6-23 months` = col\_double(),  
## .. `stunted 0-<6 months` = col\_double(),  
## .. `stunted 24-59 months` = col\_double(),  
## .. `diarrhoea cases` = col\_double(),  
## .. `Underweight 0-<6 months` = col\_double(),  
## .. `Underweight 6-23 months` = col\_double(),  
## .. `Underweight 24-59 Months` = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

# Summary statistics  
summary(data)

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

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

## [1] 399

# Remove rows with missing values  
data <- na.omit(data)  
  
# Load the 'zoo' package  
install.packages("zoo")

## Error in install.packages : Updating loaded packages

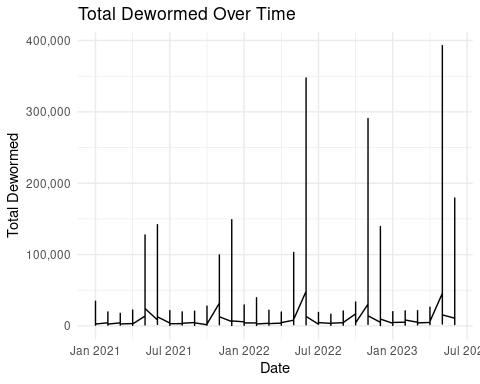
library(zoo)  
  
# Convert the 'period' column to a suitable format for plotting  
data$period\_plot <- as.yearmon(data$period, format = "%b-%y")  
  
# View the first few rows of the cleaned and transformed dataset  
head(data)

## # A tibble: 6 × 12  
## period county `Total Dewormed` `Acute Malnutrition` `stunted 6-23 months` `stunted 0-<6 months`  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Jan-23 Baringo C… 3659 8 471 34  
## 2 Jan-23 Bungoma C… 6590 24 98 154  
## 3 Jan-23 Embu Coun… 3241 72 326 86  
## 4 Jan-23 Garissa C… 6751 250 40 13  
## 5 Jan-23 Homa Bay … 4691 9 209 87  
## 6 Jan-23 Isiolo Co… 790 26 51 6  
## # ℹ 6 more variables: `stunted 24-59 months` <dbl>, `diarrhoea cases` <dbl>,  
## # `Underweight 0-<6 months` <dbl>, `Underweight 6-23 months` <dbl>,  
## # `Underweight 24-59 Months` <dbl>, period\_plot <yearmon>

# view the entire dataset in a separate data viewer window  
View(data)  
  
# Verify column names  
colnames(data)

## [1] "period" "county" "Total Dewormed"   
## [4] "Acute Malnutrition" "stunted 6-23 months" "stunted 0-<6 months"   
## [7] "stunted 24-59 months" "diarrhoea cases" "Underweight 0-<6 months"   
## [10] "Underweight 6-23 months" "Underweight 24-59 Months" "period\_plot"

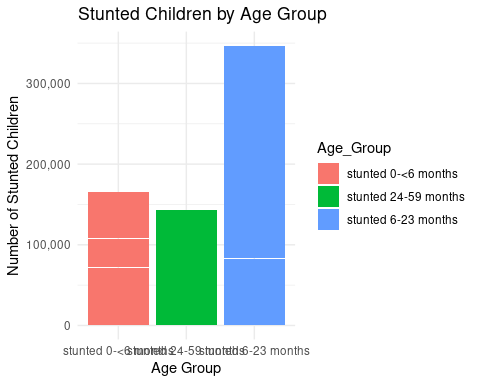
#RESEARCH QUESTIONS  
  
#1. How does the total number of children dewormed vary over time?  
# Load the scales package  
library(scales)  
  
# Convert the 'period' column to a suitable format for plotting  
data$period\_plot <- as.yearmon(data$period, format = "%b-%y")  
  
# Remove rows with missing values  
data <- na.omit(data)  
  
# Data Visualization - Time series plot for Total Dewormed  
ggplot(data, aes(x = period\_plot, y = `Total Dewormed`)) +  
 geom\_line() +  
 labs(x = "Date", y = "Total Dewormed", title = "Total Dewormed Over Time") +  
 theme\_minimal() +  
 scale\_y\_continuous(labels = comma) # Format y-axis labels with commas for thousands separator



#2. Which county has the highest number of children with acute malnutrition?  
# Find the highest number of children with acute malnutrition  
max\_acute\_malnutrition <- max(data$`Acute Malnutrition`, na.rm = TRUE)  
  
# Find the county with the highest number of children with acute malnutrition  
highest\_acute\_malnutrition\_county <- data %>%  
 filter(`Acute Malnutrition` == max\_acute\_malnutrition)  
  
# Print the result  
highest\_acute\_malnutrition\_county

## # A tibble: 1 × 12  
## period county `Total Dewormed` `Acute Malnutrition` `stunted 6-23 months` `stunted 0-<6 months`  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Oct-22 Wajir Cou… 33469 4123 8 3  
## # ℹ 6 more variables: `stunted 24-59 months` <dbl>, `diarrhoea cases` <dbl>,  
## # `Underweight 0-<6 months` <dbl>, `Underweight 6-23 months` <dbl>,  
## # `Underweight 24-59 Months` <dbl>, period\_plot <yearmon>

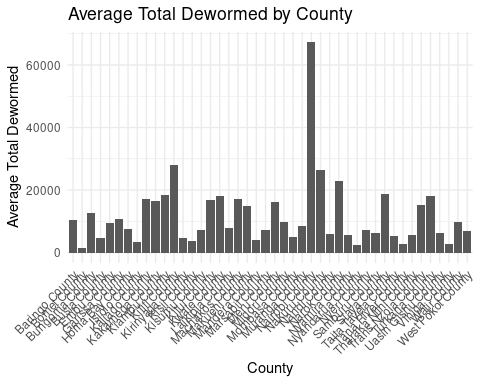
#3. How does the number of stunted children change across different age groups?  
data\_long <- data %>%  
 pivot\_longer(cols = starts\_with("stunted"), names\_to = "Age\_Group", values\_to = "Stunted\_Count")  
ggplot(data\_long, aes(x = Age\_Group, y = Stunted\_Count, fill = Age\_Group)) +  
 geom\_bar(stat = "identity") +  
 labs(x = "Age Group", y = "Number of Stunted Children", title = "Stunted Children by Age Group") +  
 theme\_minimal() +   
 scale\_y\_continuous(labels = comma) # Format y-axis labels with commas for thousands separator



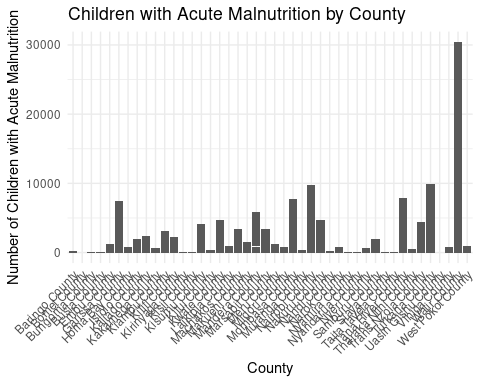
# 4. Is there any correlation between the number of children with diarrhea and the number of underweight children?  
# Calculate the correlation between 'diarrhoea cases' and 'Underweight 24-59 Months'  
correlation <- cor(data$`diarrhoea cases`, data$`Underweight 24-59 Months`, use = "complete.obs")  
  
# Print the correlation coefficient  
correlation

## [1] 0.3518845

#5. What is the overall trend of total dewormed children for each county?  
# Calculate the average total dewormed children for each county  
county\_trends <- data %>%  
 group\_by(county) %>%  
 summarize(Avg\_Total\_Dewormed = mean(`Total Dewormed`, na.rm = TRUE)) %>%  
 ggplot(aes(x = county, y = Avg\_Total\_Dewormed)) +  
 geom\_bar(stat = "identity") +  
 labs(x = "County", y = "Average Total Dewormed", title = "Average Total Dewormed by County") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
# Display the plot  
print(county\_trends)



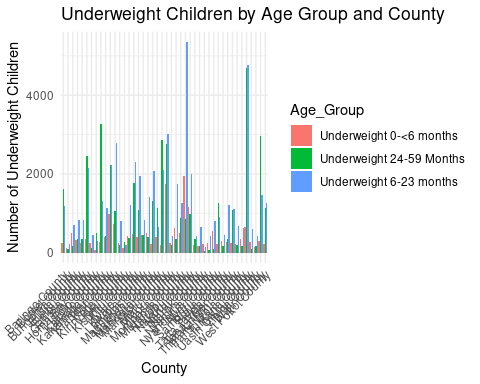
#6. How does the number of children with acute malnutrition vary across counties?  
county\_acute\_malnutrition <- data %>%  
 ggplot(aes(x = county, y = `Acute Malnutrition`)) +  
 geom\_bar(stat = "identity") +  
 labs(x = "County", y = "Number of Children with Acute Malnutrition", title = "Children with Acute Malnutrition by County") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
# Display the plot  
print(county\_acute\_malnutrition)



# Load necessary libraries  
library(dplyr)  
library(ggplot2)  
  
# 7. How do the different types of underweight children (0-6 months, 6-23 months, 24-59 months) compare in different counties?  
  
# Extract column names that start with "Underweight"  
startsWithUnderweight <- colnames(data)[startsWith(colnames(data), "Underweight")]  
  
# Print the column names that start with "Underweight" (for verification)  
print(startsWithUnderweight)

## [1] "Underweight 0-<6 months" "Underweight 6-23 months" "Underweight 24-59 Months"

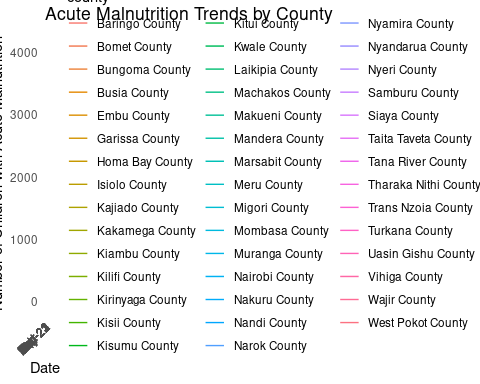
# Pivot the data\_long\_underweight to long format for visualization  
data\_long\_underweight <- data %>%  
 pivot\_longer(cols = startsWithUnderweight, names\_to = "Age\_Group", values\_to = "Underweight\_Count") %>%  
 filter(!is.na(Underweight\_Count))  
# Plot the bar chart to compare underweight children by age group and county  
ggplot(data\_long\_underweight, aes(x = county, y = Underweight\_Count, fill = Age\_Group)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(x = "County", y = "Number of Underweight Children", title = "Underweight Children by Age Group and County") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



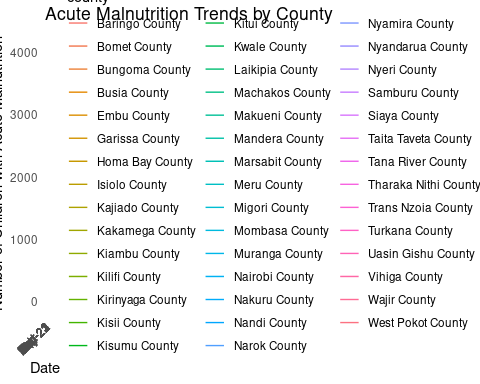
#8. What is the trend of acute malnutrition in each county over time?  
sum(is.na(data$`Acute Malnutrition`))

## [1] 0

county\_trends\_acute\_malnutrition <- data %>%  
 group\_by(county) %>%  
 ggplot(aes(x = period, y = `Acute Malnutrition`, group = county, color = county)) +  
 geom\_line() +  
 labs(x = "Date", y = "Number of Children with Acute Malnutrition", title = "Acute Malnutrition Trends by County") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
print(county\_trends\_acute\_malnutrition)



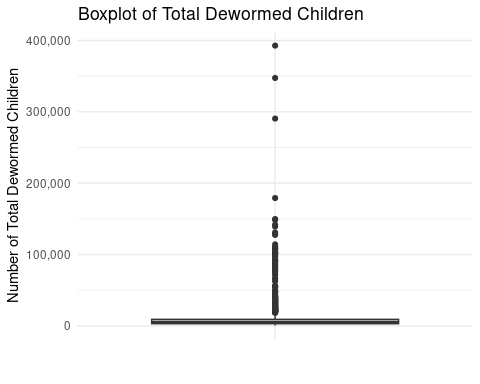
# Print the plot  
print(county\_trends\_acute\_malnutrition)



#9. Are there any outliers in the number of total dewormed children?  
print(colnames(data))

## [1] "period" "county" "Total Dewormed"   
## [4] "Acute Malnutrition" "stunted 6-23 months" "stunted 0-<6 months"   
## [7] "stunted 24-59 months" "diarrhoea cases" "Underweight 0-<6 months"   
## [10] "Underweight 6-23 months" "Underweight 24-59 Months" "period\_plot"

# Load necessary library  
library(ggplot2)  
  
# Check for missing values in "Total Dewormed" column  
data <- data[complete.cases(data), ]  
  
# Plot the boxplot of "Total Dewormed" column  
ggplot(data, aes(x = "", y = `Total Dewormed`)) +  
 geom\_boxplot() +  
 labs(x = "", y = "Number of Total Dewormed Children", title = "Boxplot of Total Dewormed Children") +  
 theme\_minimal() +  
 scale\_y\_continuous(labels = comma) # Format y-axis labels with commas for thousands separator



#THE END