Health Data Analysis

Bartley Dawud Okiya

2023-07-21

Loading the libraries

```
library(tidyverse)
library(ggplot2)
library(visdat)
library(PerformanceAnalytics)
```

Load the data and print first few rows

```
data <- read_csv("MainData.csv")

head(data)
```

```
## # A tibble: 6 x 11
##   period    county   ‘Total Dewormed‘ ‘Acute Malnutrition‘ ‘stunted 6-23 months‘
##   <chr>     <chr>               <dbl>                <dbl>                 <dbl>
## 1 1/23/2023 Baringo~             3659                    8                   471
## 2 1/23/2023 Bomet C~             1580                   NA                     1
## 3 1/23/2023 Bungoma~             6590                   24                    98
## 4 1/23/2023 Busia C~             7564                   NA                   396
## 5 1/23/2023 Elgeyo ~             1407                   NA                    92
## 6 1/23/2023 Embu Co~             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>
```

Summary of the data

```
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         
## 
```

Print the structure of the data

```
dim(data)
```

```
## [1] 1410   11
```

Summary statistics for numerical variables

```
summary(data[, 3:ncol(data)])
```

```
##  Total Dewormed   Acute Malnutrition stunted 6-23 months stunted 0-<6 months
##  Min.   :    97   Min.   :   1.0     Min.   :   1.0      Min.   :   1.0     
##  1st Qu.:  2454   1st Qu.:  15.0     1st Qu.:  69.5      1st Qu.:  36.5     
##  Median :  4564   Median :  39.0     Median : 159.0      Median :  84.0     
##  Mean   : 11458   Mean   : 125.4     Mean   : 280.2      Mean   : 139.8     
##  3rd Qu.:  8222   3rd Qu.: 143.5     3rd Qu.: 328.5      3rd Qu.: 157.0     
##  Max.   :392800   Max.   :4123.0     Max.   :4398.0      Max.   :7900.0     
##                   NA’s   :355        NA’s   :11          NA’s   :19         
##  stunted 24-59 months diarrhoea cases Underweight 0-<6 months
##  Min.   :   1.0       Min.   :  198   Min.   :   6.0         
##  1st Qu.:  22.0       1st Qu.: 1464   1st Qu.:  87.0         
##  Median :  50.0       Median : 2158   Median : 162.5         
##  Mean   : 110.8       Mean   : 2813   Mean   : 223.5         
##  3rd Qu.: 114.2       3rd Qu.: 3335   3rd Qu.: 272.8         
##  Max.   :3169.0       Max.   :15795   Max.   :1937.0         
##  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
##
```

Convert the “Period” column to date format

```
data$period <- as.Date(data$period, format = "%m/%d/%y")
str(data)
```

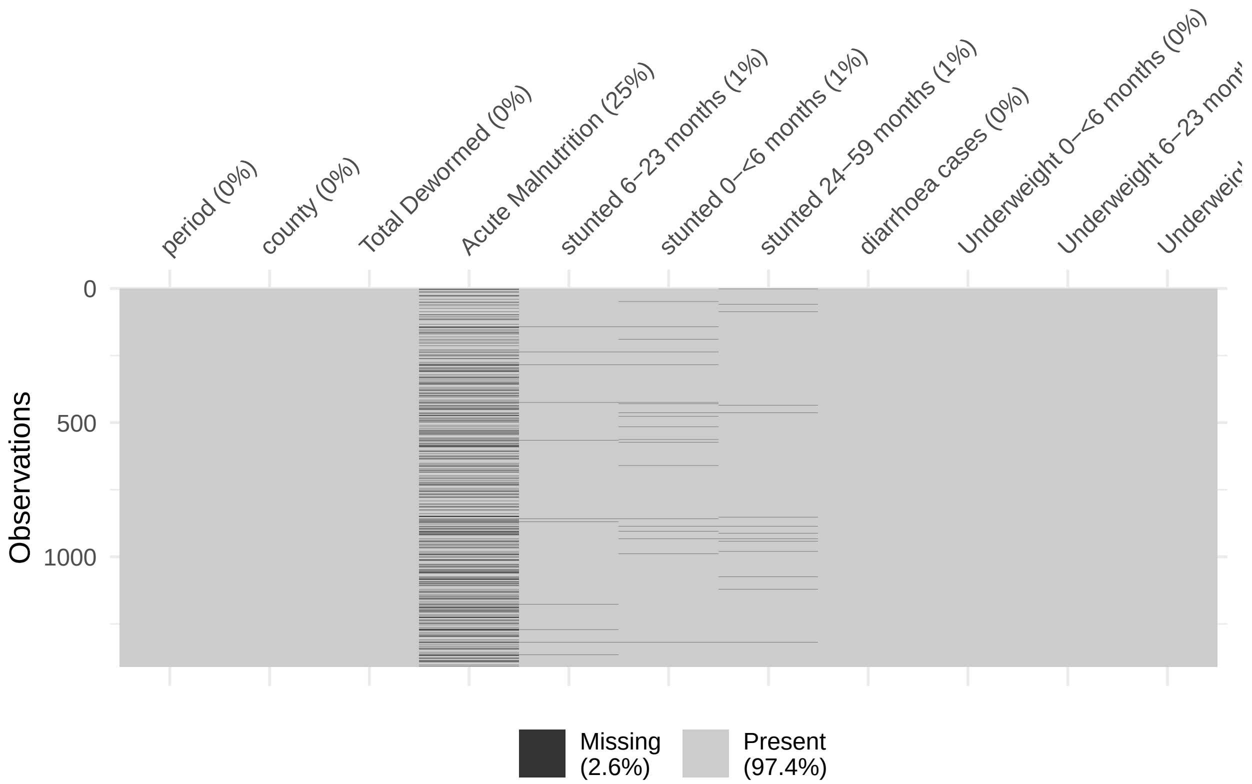
```
## spc_tbl_ [1,410 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ period                  : Date[1:1410], format: "2020-01-23" "2020-01-23" ...
##  $ county                  : chr [1:1410] "Baringo County" "Bomet County" "Bungoma County" "Busia Cou
##  $ Total Dewormed          : num [1:1410] 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>
```

Dealing with missing values

```
colSums(is.na(data))
```

```
##                   period                   county           Total Dewormed
##                        0                        0                        0
##       Acute Malnutrition      stunted 6-23 months      stunted 0-<6 months
##                      355                       11                       19
##     stunted 24-59 months          diarrhoea cases  Underweight 0-<6 months
##                        14                        0                        0
##  Underweight 6-23 months Underweight 24-59 Months
##                         0                        0
```

```
vis_miss(data)
```



```
data %>% drop_na()
```

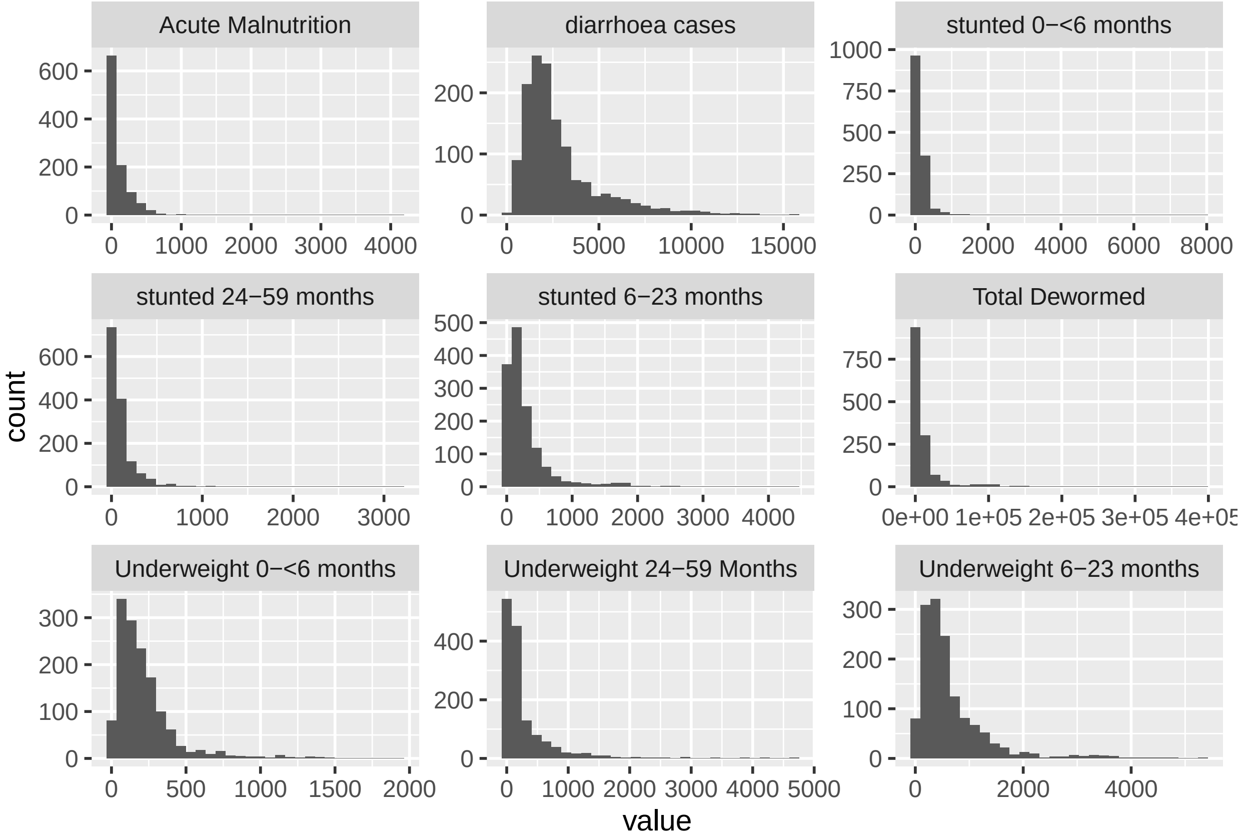
```
## # A tibble: 1,048 x 11
##    period     county `Total Dewormed` `Acute Malnutrition` `stunted 6-23 months`
##    <date>     <chr>              <dbl>                <dbl>                 <dbl>
##  1 2020-01-23 Barin~              3659                    8                   471
##  2 2020-01-23 Bungo~              6590                   24                    98
##  3 2020-01-23 Embu ~              3241                   72                   326
##  4 2020-01-23 Garis~              6751                  250                    40
##  5 2020-01-23 Homa ~              4691                    9                   209
##  6 2020-01-23 Isiol~               790                   26                    51
##  7 2020-01-23 Kajia~              7532                  104                   319
##  8 2020-01-23 Kakam~              8044                   36                   252
##  9 2020-01-23 Kiamb~              7891                  183                   530
## 10 2020-01-23 Kilif~              9991                   81                  1690
## # i 1,038 more rows
## # 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>
```

Exploring Distributions

Histogram for each numerical variable

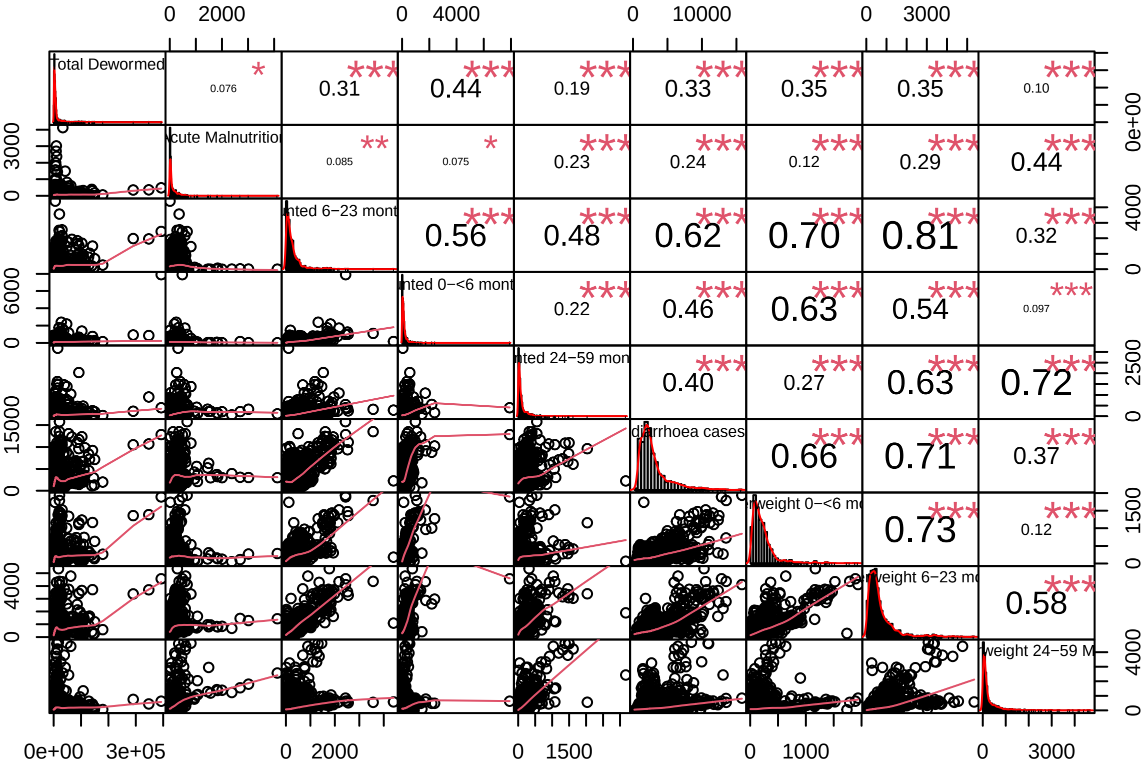
```
data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



Analyzing Relationships Between Variables

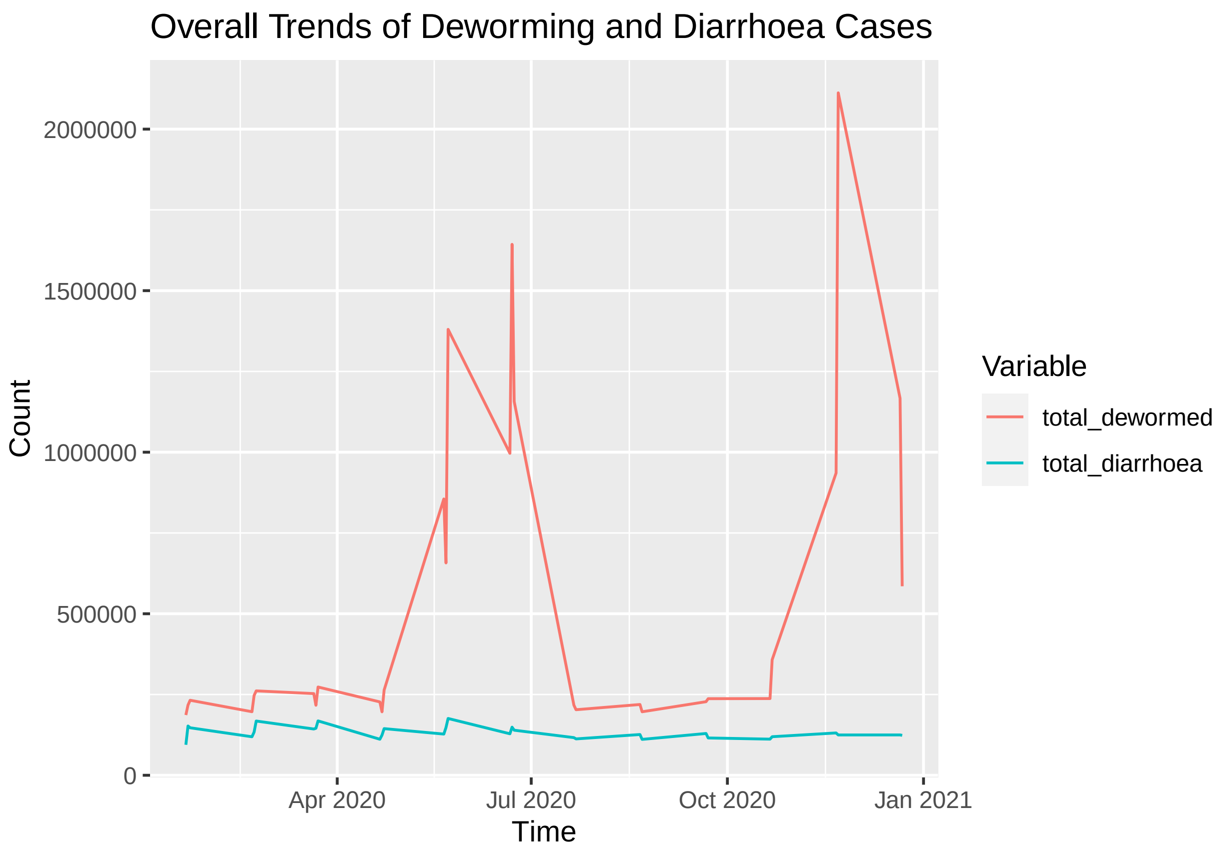
```
numerics <- data[,3:ncol(data)]


chart.Correlation(numerics, histogram=TRUE, method = "pearson")
```



## Research Question: Does the total deworming effort significantly reduce the diarrhoea cases in children under 5 years over time?

Overall Trends Visualization

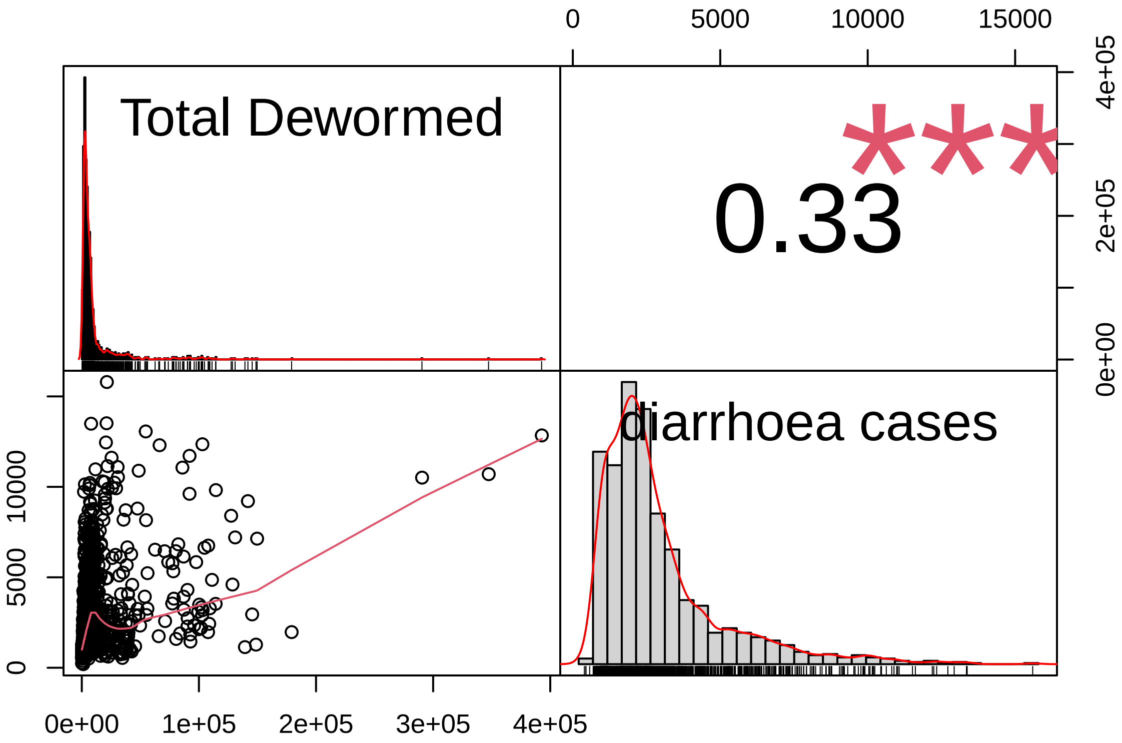
```
data %>%
  group_by(period) %>%
  summarise(total_dewormed = sum(`Total Dewormed`, na.rm = TRUE),
            total_diarrhoea = sum(`diarrhoea cases`, na.rm = TRUE)) %>%
  gather(key = "variable", value = "count", -period) %>%
  ggplot(aes(x = period, y = count, color = variable)) +
  geom_line() +
  labs(x = "Time", y = "Count", title = "Overall Trends of Deworming and Diarrhoea Cases", color = "Var
```



It is apparent that the deworming rate was driven by diarrhea cases. When there was a growth in people being dewormed, the cases stabilized.

Correlation Analysis for select variables

```
correlation_data <- data[, c("Total Dewormed", "diarrhoea cases")]


chart.Correlation(correlation_data, histogram=TRUE, method = "pearson")
```



A correlation of 0.33 between the number of children dewormed and the number of diarrhea cases indicates a moderate positive linear relationship. As the total number of dewormed children increased, the number of diarrhea cases also tends to increase. However, the strength of this relationship is moderate, so there's a lot of variability that is not explained by this relationship.

Regression Analysis

```
model <- lm(`diarrhoea cases` ~ `Total Dewormed`, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = ‘diarrhoea cases‘ ~ ‘Total Dewormed‘, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5565.4 -1246.2  -564.6   532.9 12701.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.491e+03  5.965e+01   41.76   <2e-16 ***
## ‘Total Dewormed‘ 2.816e-02  2.143e-03   13.14   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2041 on 1408 degrees of freedom
## Multiple R-squared:  0.1092, Adjusted R-squared:  0.1086
## F-statistic: 172.6 on 1 and 1408 DF,  p-value: < 2.2e-16
```

The p-value for Total Dewormed is less than 0.05, indicating that the Total Dewormed is a statistically significant predictor of the diarrhoea cases.

the R-squared value is 0.1092, which means that only about 10.92% of the variability in diarrhoea cases is explained by Total Dewormed. The rest of the variability is unexplained by this model, suggesting there may be other variables not included in this model that could explain the number of diarrhoea cases.

We will chose to add more variables:

1. Acute Malnutrition: It's plausible that malnourished children have weaker immune systems, and therefore may be more susceptible to diarrheal diseases. This could help explain additional variability in the data.
2. Underweight 0-<6 months, Underweight 6-23 months, Underweight 24-59 Months: These variables represent underweight children at different age categories. Being underweight might also make children more susceptible to diseases including diarrhea.
3. stunted 0-<6 months, stunted 6-23 months, stunted 24-59 months: These variables could also be relevant, as stunting is a sign of chronic malnutrition, which can be linked to susceptibility to disease.

Adding more variables (Multiple Lienar Regression)

```r
model <- lm(`diarrhoea cases` ~ `Total Dewormed` + `Acute Malnutrition` +
            `Underweight 0-<6 months` + `Underweight 6-23 months` + `Underweight 24-59 Months` +
            `stunted 0-<6 months` + `stunted 6-23 months` + `stunted 24-59 months`, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = ‘diarrhoea cases‘ ~ ‘Total Dewormed‘ + ‘Acute Malnutrition‘ +
##     ‘Underweight 0-<6 months‘ + ‘Underweight 6-23 months‘ + ‘Underweight 24-59 Months‘ +
##     ‘stunted 0-<6 months‘ + ‘stunted 6-23 months‘ + ‘stunted 24-59 months‘,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5341.5  -948.1  -152.3   689.5  7911.5
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.076e+03  7.149e+01  15.055  < 2e-16 ***
## ‘Total Dewormed‘            5.025e-03  1.883e-03   2.668  0.00775 **
## ‘Acute Malnutrition‘        3.648e-01  2.064e-01   1.768  0.07740 .
## ‘Underweight 0-<6 months‘   3.129e+00  3.389e-01   9.234  < 2e-16 ***
## ‘Underweight 6-23 months‘   1.040e+00  1.682e-01   6.187 8.79e-10 ***
## ‘Underweight 24-59 Months‘  4.858e-01  1.385e-01   3.509  0.00047 ***
## ‘stunted 0-<6 months‘      -1.011e-01  2.032e-01  -0.498  0.61892
## ‘stunted 6-23 months‘       5.806e-01  2.040e-01   2.847  0.00450 **
## ‘stunted 24-59 months‘     -4.688e-01  3.349e-01  -1.400  0.16189
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1489 on 1039 degrees of freedom
##   (362 observations deleted due to missingness)
## Multiple R-squared:  0.586,  Adjusted R-squared:  0.5828
## F-statistic: 183.8 on 8 and 1039 DF,  p-value: < 2.2e-16
```

The model’s R-squared is 0.586, which means that about 58.6% of the variation in diarrhoea cases can be explained by these predictors

1. Among all predictors, five variables (Total Dewormed, Acute Malnutrition, Underweight 0-<6 months, Underweight 6-23 months, and Underweight 24-59 Months) show a significant positive correlation with diarrhoea cases. The variable stunted 6-23 months is also positively correlated with the response variable.
2. The coefficient of Total Dewormed (0.005) indicates that for every additional unit of Total Dewormed, the number of diarrhoea cases is predicted to increase by about 0.005 units, holding all other predictors constant. However, the practical significance of this effect might not be substantial, given the scale of these variables.
3. Acute Malnutrition (0.3648), Underweight 0-<6 months (3.129), Underweight 6-23 months (1.040) and Underweight 24-59 Months (0.486) also show positive coefficients, meaning that increases in these predictors are associated with increases in the number of diarrhoea cases, all else being equal.
4. The stunted 6-23 months variable (0.5806) shows a positive relationship with diarrhoea cases too, suggesting that the higher the number of children stunted at this age, the higher the number of diarrhea cases, given the other predictors in the model remain constant.
5. The variables stunted 0-<6 months and stunted 24-59 months were not found to be significantly related to diarrhoea cases at the standard 0.05 level.

Conclusion

The analysis has been conducted aiming to investigate the effect of total deworming efforts on the reduction of diarrhoea cases in children under 5 years over time. From the correlation analysis and the regression models created, several conclusions can be drawn.

1. Firstly, a correlation of 0.33 was found between total deworming efforts and diarrhoea cases, which indicates a moderate positive relationship between the two variables. This suggests that as deworming efforts increase, diarrhoea cases also tend to increase, contrary to the expectation that deworming would lead to a decrease in diarrhoea cases. However, it’s crucial to note that this correlation does not imply causation, and the positive correlation could be driven by other confounding factors.
2. The simple linear regression model showed that the total dewormed variable is statistically significant, with an increase in total deworming associated with an increase in diarrhoea cases. However, the model’s R-squared value was 0.1092, meaning that only about 10.92% of the variation in diarrhoea cases can be explained by deworming efforts alone.

3. The multiple linear regression model included additional variables to account for more factors that may influence diarrhoea cases. Variables such as Acute Malnutrition, Underweight 0-<6 months, Underweight 6-23 months, and Underweight 24-59 Months were all positively associated with diarrhoea cases, and significantly so. This indicates that these factors play a crucial role in the prevalence of diarrhoea cases, and they should be addressed in conjunction with deworming efforts. The multiple regression model had a considerably higher R-squared value of 0.586, implying that the model, including all the predictors, explains approximately 58.6% of the variation in diarrhoea cases.

In conclusion, while deworming efforts are essential in addressing children's health issues, our analysis suggests that they are not directly linked to a reduction in diarrhoea cases. Other health and socio-economic factors significantly contribute to the diarrhoea case rate and should not be overlooked in efforts to improve children's health. Further research is required to gain a deeper understanding of these relationships and how they can be utilized to formulate more effective public health strategies.