Exploratory Data Analysis- Food Security Indicators in the Sub-Saharan African Regions

Meta Data

The dataset contains information about various Food Security indicators from over the world. It has data ranging from 2017 to 2022, for 205 countries. The data set was obtained from Food and Agricultural Organization of the United Nations (FAOSTAT). The original data set contains 45 indicators/variables. However, only 21 have been chosen for the analysis due to similarities in the variables.

However, the data set contains missing values in the form of NAs for different countries or years in certain indicators. These missing values and 'problematic' values form for errors in the codes and make creating charts a complex task. The health related variables from over the world are being compared with countries in the Sub Saharan African Region-Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius,Mozambique,Nambia, Niger, Nigeria, Rwanda Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Sudan,Togo,Uganda,Zambia,Zimbabwe. Since food and health are unescapable parts of every human's life, it is interesting to see how despite the importance of it, different regions/countries have different levels of accessibility to it due to other variables.

Collecting the data was easy as it readily available on a United Nations website. However, a relatively more complex task was cleaning it, sorting it as there were initially 37000 rows of data.

The data was cleaned using Excel and was formatted according to the years. Further, the data was put into R to get the graphs used for visual representation throughout the report. The codes for the following are mentioned.

In addition to that, Notes and Flags columns were made for each of the indicators. For e.g. RailLinesDensitykm100sqkm the flag 'X' signifies countries who have recorded an entry in that particular year. Not all indicators are thoroughly analyzed but touched upon. A lot of scatter plots have been made due to the nature of the data-it is more important to define the relationship between variables due to the interconnectedness.

ETHICAL ISSUES ABOUT THE DATA AND TOPIC: Ethical issues involving the data set for this analysis could be the variables that have been used and analyzed. For example. GDPpc-as it doesn't consider the socioeconomic impact of reporting GDPpc and its potential impact on perceptions of different regions. Ensuring data bout political stability is presented responsibly due to again the impact on perceptions about the countries. Ensuring confidentiality of individuals with low birth rate and aggregating the data effectively.

Area: Country names represented in the dataset.

Year: Indicating the specific year in which the observations were recorded.

Variables

1. RailLinesDensityKmPer100sqkm: Density of rail lines per 100 square kilometers.
2. LowBirthweightPCT: Percentage of infants with low birth weight.
3. PopUsingSafeSanitationPCT: Percentage of the population using safe sanitation facilities.
4. ExclusiveBreastfeedInfants0_5monthsPCT: Percentage of infants aged 0 to 5 months exclusively breastfed.
5. WomenRepAge15_49WithAnaemiaPCT: Percentage of women aged 15 to 49 with anemia.
6. PoliticalStabilityNoViolTerrorINDEX: An index measuring political stability, absence of violence, and the absence of terrorism incidents.
7. PopUsingSafeDrinkingWaterPCT: Percentage of the population with access to safe drinking water.
8. PopUsingAtLeastBasicSanitationPCT: Percentage of the population using at least basic sanitation facilities.
9. PopAtLeastBasicDrinkingWaterPCT: Percentage of the population with access to at least basic drinking water services.
10. ChildrenUder5OverweightPCT: Percentage of children under 5 years old classified as overweight.
11. ChildrenUnder5OnerweightPCT: Percentage of children under 5 years old classified as underweight.
12. ChildrenUnder5WithWastingPCT: Percentage of children under 5 years old with wasting (low weight-for-height).
13. FoodProdPCVariabilityCONSTDOLLARS: Coefficient of variation in per capita food production in constant dollars.

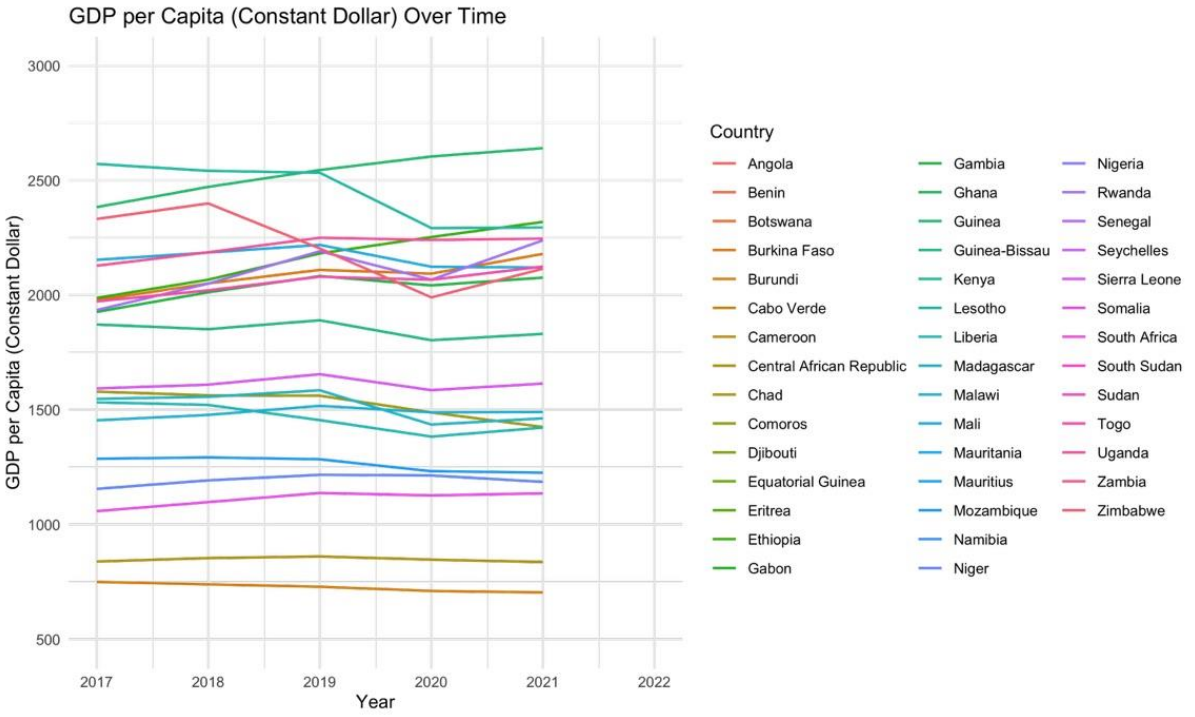
14. NumWomenRepAge15_49WithAnaemiaMILLIONS: Number of women aged 15 to 49 with anemia in millions.
15. NumNewbornLowBirthweightMILLIONS: Number of newborns with low birth weight in millions.
16. NumChildUnder5StuntedMILLIONS: Number of children under 5 years old classified as stunted in millions.
17. NumChildUnder5OverweightMILLIONS: Number of children under 5 years old classified as overweight in millions.
18. NumChildUnder5WithWastingMILLIONS: Number of children under 5 years old with wasting (low weight-for-height) in millions.
19. MinDietEnergyReqKCALperCAPperDAY: Minimum dietary energy requirements in kilocalories per capita per day.
20. CaloricLossRetailDistPCT: Percentage of caloric loss during retail distribution.
21. AvDietaryEnergyReqKCALperCAPperDAY: Average dietary energy requirements in kilocalories per capita per day.

First step of this report is analyzing the GDPpc (Gross Domestic Product Per Capita) or the income per head in the Sub-Saharan African regions as the income or GDP of a country can be directly linked to other variables which will be shown below. The line chart shows the same. This chart was made using R with the code:

```
# Load the necessary library
library(ggplot2)

# Your data frame name might be different, please replace "ssa_data" with the actual name
your data <- ssa_data

# Create a line chart for GDPpcConstDollar
line_chart <- plot(your_data, aes(x = Year, y = GDPpcConstDollar, color = Area, group = Area))
+
  geom_line(size = 0.7) +
  labs(title = "GDP per Capita (Constant Dollar) Over Time",
       x = "Year",
       y = "GDP per Capita (Constant Dollar)",
       color = "Country") +
  scale_y_continuous(limits = c(500, 3000), breaks = seq(500, 3000, by = 500)) +
  theme_minimal()
```

```
# Print the plot
print(line_chart)

# Save the plot as an image file
ggsave("line_chart_sub_saharan_africa.png", line_chart, width = 10, height = 6, dpi = 300)
```

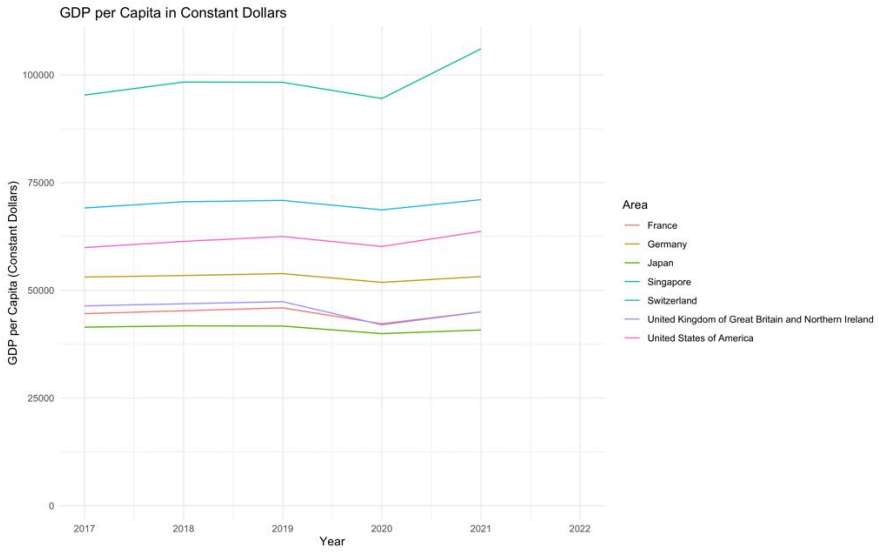
It is visible that countries with a lower GDPpc have a smoother and more constant GDPpc throughout the years whereas countries with a relatively higher one has more changes over the years where there more drops and pickups.



The GDPpc in the countries of the Sub-Saharan African Region, ranges from 750(Burkina Faso -2017) to 2250(Gabon-2021), these areas show relatively low income compared to other countries like Switzerland where it goes up to 70000's in 2018 and 2019. A list of 'developed' countries has been curated to compare effectively. These include-United States of America, United Kingdom of Great Britain and Northern Ireland, Singapore, Switzerland, Germany, Japan, France. A line chart created with the following code is given below:

```
# Filter data for the specified countries
selected_countries <- c("United States of America", "United Kingdom of Great Britain and Northern Ireland",
                        "Singapore", "Switzerland", "Germany", "Japan", "France")

filtered_data <- dt[dt$Area %in% selected_countries, ]

# Load necessary libraries
```

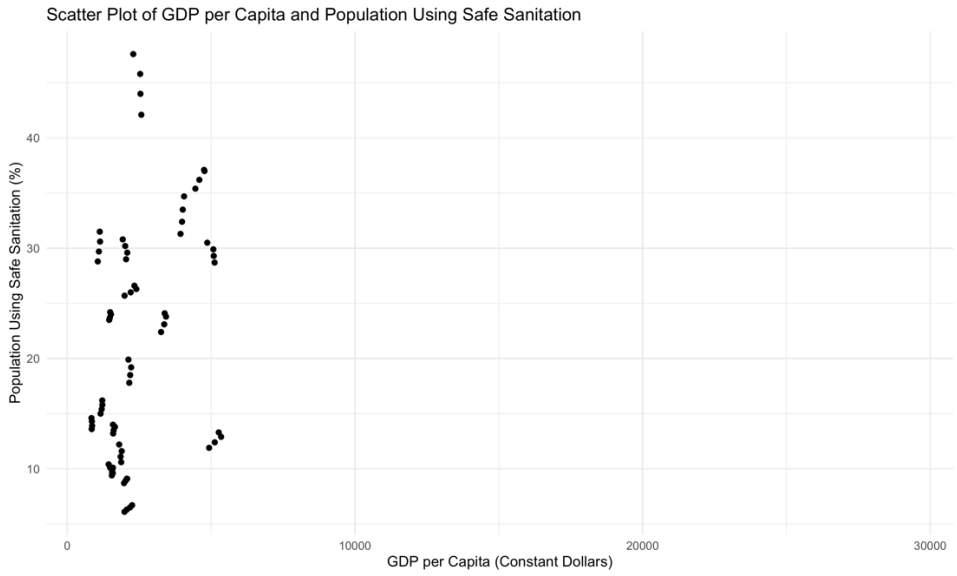
```
library(ggplot2)

# Find the maximum value of GDPpcConstDollar
max_value <- max(filtered_data$GDPpcConstDollar, na.rm = TRUE)

# Create a line chart
plot <- ggplot(filtered_data, aes(x = Year, y = GDPpcConstDollar, color = Area)) +
  geom_line() +
  labs(title = "GDP per Capita in Constant Dollars",
       x = "Year",
       y = "GDP per Capita (Constant Dollars)",
       color = "Area") +
  theme_minimal() +
  ylim(2000, max_value)  # Set the y-axis limits

# Save the plot as a PNG file
ggsave("GDP_per_Capita_Plot.png", plot, width = 10, height = 6)
```



GDP per capita can be directly linked with essential services like drinking water and sanitation services. This is shown in the following scatter plots made using the following codes:

```
# Load necessary libraries
library(ggplot2)

# Filter data for Sub-Saharan African countries
sub_saharan_african_countries <- c("Angola", "Benin", "Botswana", "Burkina Faso", "Burundi", "Cabo Verde", "Cameroon", "Central African Republic", "Chad", "Comoros", "Congo", "Cote d'Ivoire", "Djibouti", "Equatorial Guinea", "Eritrea", "Eswatini", "Ethiopia", "Gabon", "Gambia", "Ghana", "Guinea", "Guinea-Bissau", "Kenya", "Lesotho", "Liberia", "Madagascar", "Malawi", "Mali", "Mauritania", "Mauritius", "Mozambique", "Namibia", "Niger", "Nigeria", "Rwanda", "Sao Tome and Principe", "Senegal", "Seychelles", "Sierra Leone", "Somalia", "South Africa", "South Sudan", "Sudan", "Tanzania", "Togo", "Uganda", "Zambia", "Zimbabwe")
```

```
sub_saharan_data <- dt[dt$Area %in% sub_saharan_african_countries, ]

# Create a scatter plot
scatter_plot <- ggplot(sub_saharan_data, aes(x = GDPpcConstDollar, y = PopUsingSafeSanitationPCT)) +
  geom_point() +
  labs(title = "Scatter Plot of GDP per Capita and Population Using Safe Sanitation",
       x = "GDP per Capita (Constant Dollars)",
       y = "Population Using Safe Sanitation (%)") +
  theme_minimal()

# Save the scatter plot as a PNG file
ggsave("Scatter_Plot_SubSaharan.png", scatter_plot, width = 10, height = 6)
```

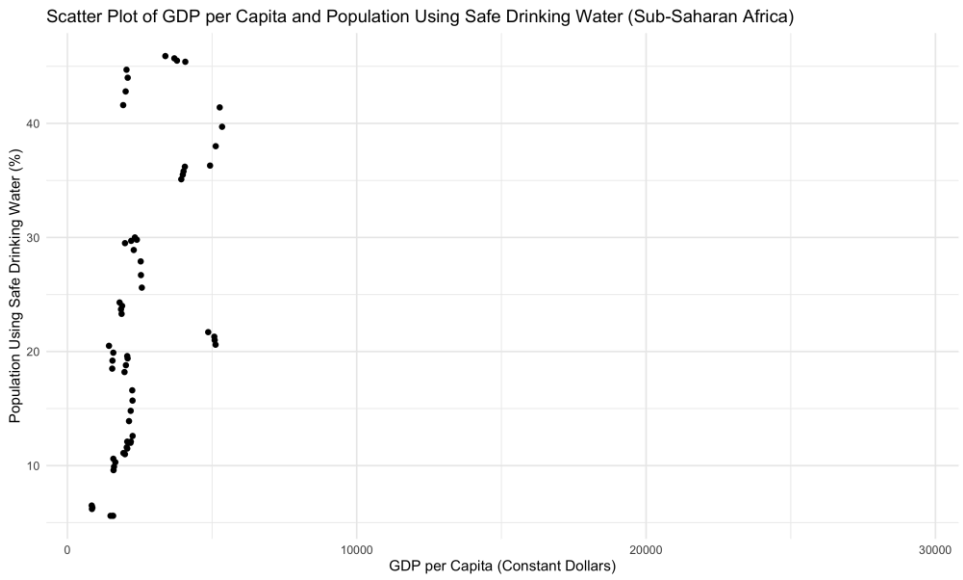


The hypothesis that GDP per capita is directly linked to population using safe sanitation is not entirely confirmed through the graph. When the GDPpc is near 2000(average for countries of this region) it is starting off 10% which is positive as only 10% population is using safe sanitation, but gradually it starts increasing and goes up to 40% for the same amount of GDPpc. However, when the GDPpc is higher at around 5000 the population using safe is still around 12%. Therefore, a more in-depth analysis required to confirm this.

Code 2:

```
# Load necessary libraries
library(ggplot2)
```

```
# Filter data for Sub-Saharan African countries
sub_saharan_african_countries <- c("Angola", "Benin", "Botswana", "Burkina Faso", "Burundi", "Cabo Verde", "Cameroon", "Central African Republic", "Chad", "Comoros", "Congo", "Cote d'Ivoire", "Djibouti", "Equatorial Guinea", "Eritrea", "Eswatini", "Ethiopia", "Gabon", "Gambia", "Ghana", "Guinea", "Guinea-Bissau", "Kenya", "Lesotho", "Liberia", "Madagascar", "Malawi", "Mali", "Mauritania", "Mauritius", "Mozambique", "Namibia", "Niger", "Nigeria", "Rwanda", "Sao Tome and Principe", "Senegal", "Seychelles", "Sierra Leone", "Somalia", "South Africa", "South Sudan", "Sudan", "Tanzania", "Togo", "Uganda", "Zambia", "Zimbabwe")

sub_saharan_data <- dt[dt$Area %in% sub_saharan_african_countries, ]

# Create a scatter plot
scatter_plot <- ggplot(sub_saharan_data, aes(x = GDPpcConstDollar, y = PopUsingSafeDrinkingWaterPCT)) +
  geom_point() +
  labs(title = "Scatter Plot of GDP per Capita and Population Using Safe Drinking Water (Sub-Saharan Africa)",
       x = "GDP per Capita (Constant Dollars)",
       y = "Population Using Safe Drinking Water (%)") +
  theme_minimal()

# Save the scatter plot as a PNG file
ggsave("Scatter_Plot_GDPvsSafeDrinkingWater_SubSaharan.png", scatter_plot, width = 10, height = 6)
```



For a hypothesis that tests whether GDPpc and %population using safe drinking is directly correlated, it is visible in the graph a lower GDppc($1000-$2000) has only 3%of its population with safe drinking water. Gradually at the same level of GDPpc it grows hence rejects the idea

that it is positively linked. However, the direct link can also be seen when at $5000 it is at high but also breaks it at the same. Hence a more detailed analysis is required for the same.

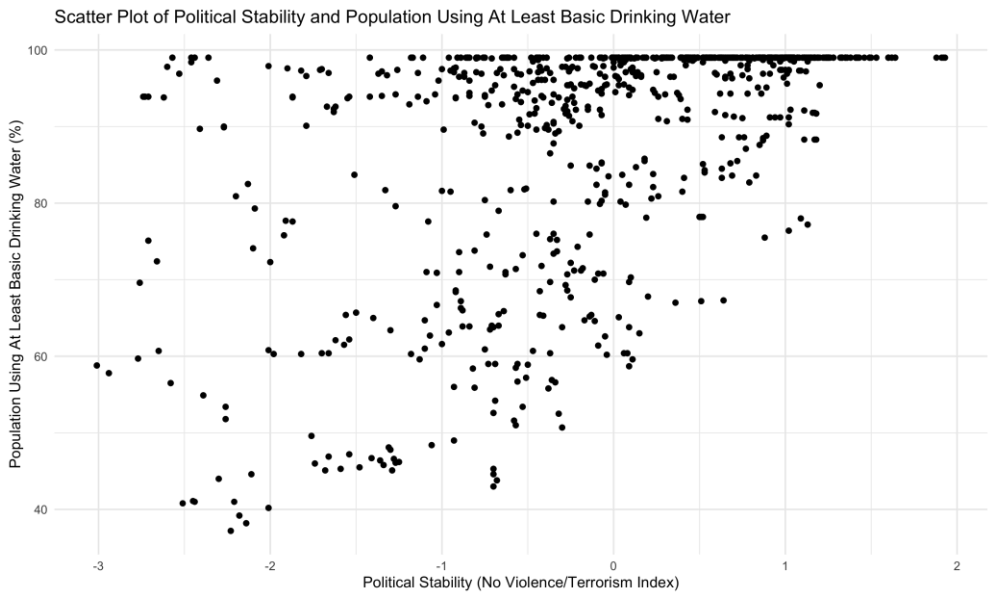
Another factor affecting these variables is Political Stability and absence of violence and terrorism index. This index is now compared with a slightly different index-%population using at least basic sanitation and population using at least basic drinking water. However, this time it is more general and not specific to a region.

Code for basic sanitation services Scatter:

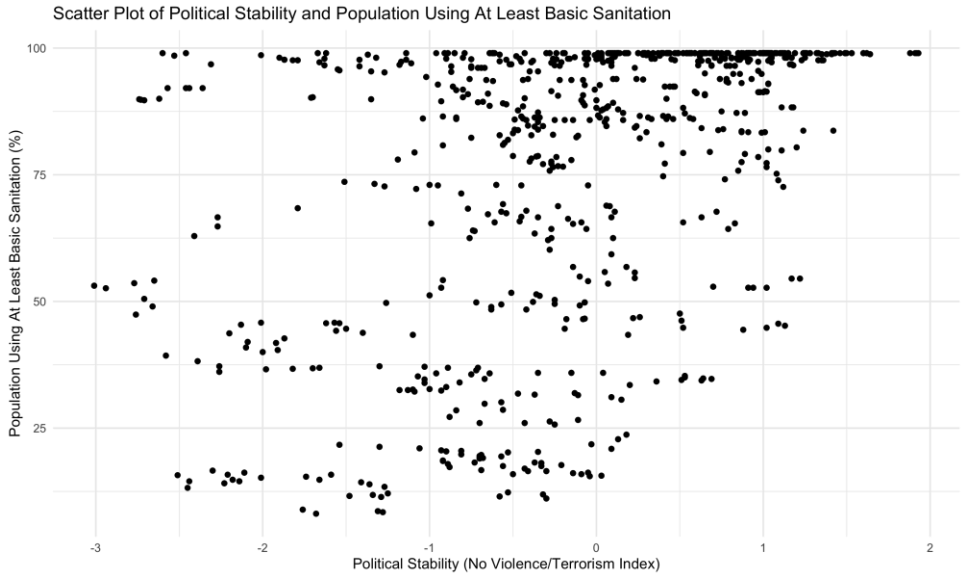
```
# Load necessary libraries
library(ggplot2)

# Create a scatter plot for PoliticalStabilityNoViolTerrorINDEX and PopUsingAtLeastBasicSanitationPCT
scatter_plot <- ggplot(dt, aes(x = PoliticalStabilityNoViolTerrorINDEX, y = PopUsingAtLeastBasicSanitationPCT)) +
  geom_point() +
  labs(title = "Scatter Plot of Political Stability and Population Using At Least Basic Sanitation",
       x = "Political Stability (No Violence/Terrorism Index)",
       y = "Population Using At Least Basic Sanitation (%)") +
  theme_minimal()

# Save the scatter plot as a PNG file
ggsave("Scatter_Plot_PoliticalStabilityVsPopSanitation.png", scatter_plot, width = 10, height = 6)
```

Code for Basic Drinking Water Facilities Scatter:

```
# Load necessary libraries
library(ggplot2)

# Create a scatter plot for PoliticalStabilityNoViolTerrorINDEX and PopAtLeastBasicDrinkingWaterPCT
scatter_plot <- ggplot(dt, aes(x = PoliticalStabilityNoViolTerrorINDEX, y = PopAtLeastBasicDrinkingWaterPCT)) +
  geom_point() +
  labs(title = "Scatter Plot of Political Stability and Population Using At Least Basic Drinking Water",
       x = "Political Stability (No Violence/Terrorism Index)",
       y = "Population Using At Least Basic Drinking Water (%)") +
  theme_minimal()

# Save the scatter plot as a PNG file
ggsave("Scatter_Plot_PoliticalStabilityVsPopWater.png", scatter_plot, width = 10, height = 6)
```



The political stability index has a clear direct correlation to the percentage of the population that has access to at least basic drinking water. When the index is negative a lower amount of population (40%) has access, this could be due to wars, even in low stability times, some of them 100% access to at least basic water services, these could be the more stable or non-violent areas or the more developed countries where a proper system for drinking eater is in place. As the political stability index is increasing, the level of access to at least basic drinking water is sticking to 100% or slighter lower.



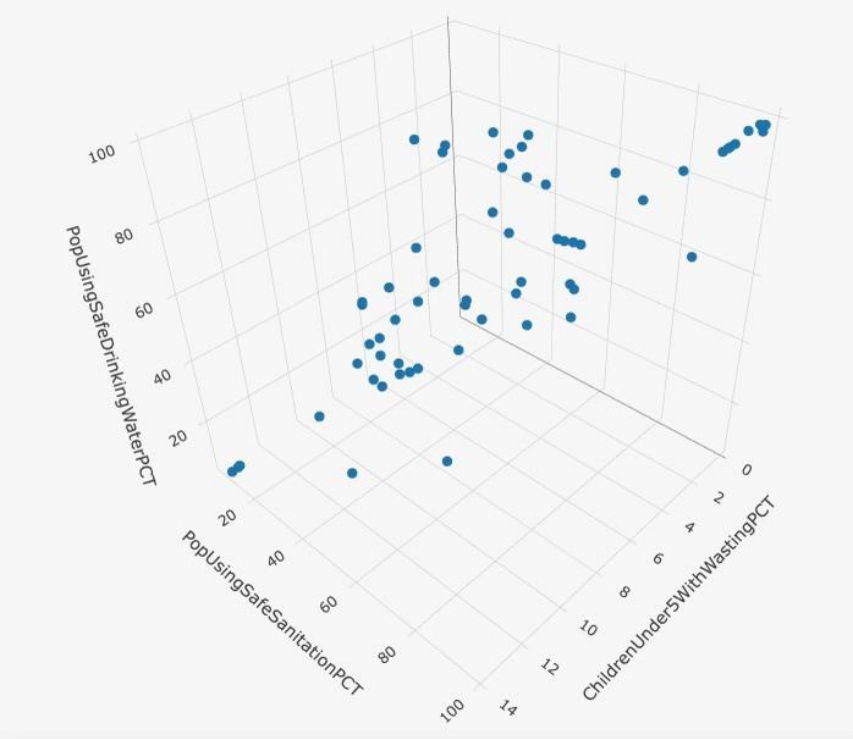
The similar analysis goes for political stability and basic sanitation facilities. However, from the chart it's a slightly a different pattern i.e. at 0 which is slightly more unstable than negative,

sanitation services are still at a low but also at 100% and since this is for the world and not a specific country, the difference in countries is clearly visible as there may be a system in place to stop the violence and terror and for it to not damage and rupture the lives for their citizens by taking away their basic rights to these services. A system that is well put in place will not be able to get damaged by political instability whereas this can happen in lesser developed areas where these services themselves may have an unstable and unreliable source.
As now it is clear that all variables are interconnected, the next step is to analyze, the impact of the level accessibility of the basic services on the health of the people-Children under 5 years of age. This relation is shown using the 3D scatter plot below.

Code:

```
# Select relevant columns for the 3D scatter plot
selected_columns_3d <- c("ChildrenUnder5WithWastingPCT", "PopUsingSafeSanitationPCT", "PopUsingSafeDrinkingWaterPCT")

# Create 3D scatter plot
scatter_3d <- plot_ly(
  dt,
  x = dt[, selected_columns_3d[1]],
  y = dt[, selected_columns_3d[2]],
  z = dt[, selected_columns_3d[3]],
  type = "scatter3d",
  mode = "markers",
  marker = list(size = 5)
) %>%
  layout(scene = list(xaxis = list(title = selected_columns_3d[1]),
                      yaxis = list(title = selected_columns_3d[2]),
                      zaxis = list(title = selected_columns_3d[3])))

# Show the plot
scatter_3d

# Save the plot as an image file (adjust the filename and format as needed)
save_image(scatter_3d, file = "scatter_3d_plot.png")
```

The variables are population using safe drinking water and safe sanitation and children under 5 with wasting i.e. low weight for height.

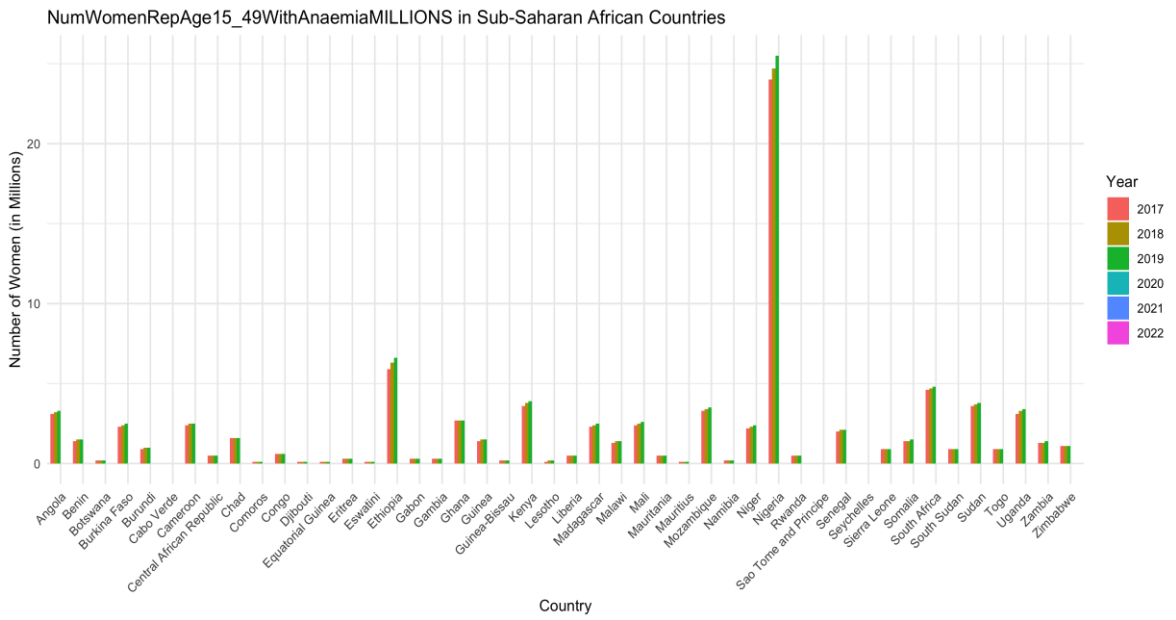


The way the analysis of the scatter plot can be carried out is straightforward. Population using safe drinking water and sanitation services is directly correlated. However, Population with safe sanitation does not much of a relation to wasting. However, population using drinking water is more related since both are directly affected to nutrition rather than the overall of a child. Between 4 to 8% of children with wasting also fall under the category where only less than 40% of the population has access to safe drinking water services. The highest level of % of wasting-14%, also shows water service accessibility of less than 20%.

Code for the bar chart shown below (Number of Women between 15 and 49 with Anemia):

```
# Load necessary libraries
library(ggplot2)

# Filter data for Sub-Saharan African countries
sub_saharan_african_countries <- c("Angola", "Benin", "Botswana", "Burkina Faso", "Burundi", "Cabo Verde", "Cameroon", "Central African Republic", "Chad", "Comoros", "Congo", "Cote d'Ivoire", "Djibouti", "Equatorial Guinea", "Eritrea", "Eswatini", "Ethiopia", "Gabon", "Gambia", "Ghana", "Guinea", "Guinea-Bissau", "Kenya", "Lesotho", "Liberia", "Madagascar", "Malawi", "Mali", "Mauritania", "Mauritius", "Mozambique", "Namibia", "Niger", "Nigeria", "Rwanda", "Sao Tome and Principe", "Senegal", "Seychelles", "Sierra Leone", "Somalia", "South Africa", "South Sudan", "Sudan", "Tanzania", "Togo", "Uganda", "Zambia", "Zimbabwe")

sub_saharan_data <- dt[dt$Area %in% sub_saharan_african_countries, ]

# Bar chart for NumWomenRepAge15_49WithAnaemiaMILLIONS
```

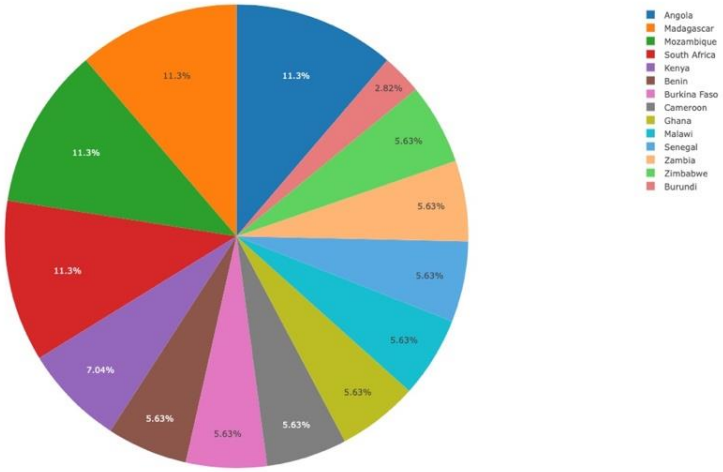
```
bar_chart <- ggplot(sub_saharan_data, aes(x = Area, y = NumWomenRepAge15_49WithAnaemiaMILLIONS, fill = factor(Year))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "NumWomenRepAge15_49WithAnaemiaMILLIONS in Sub-Saharan African Countries",
       x = "Country", y = "Number of Women (in Millions)",
       fill = "Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Save the bar chart as an image (PNG)
ggsave("Bar_Chart_NumWomenRepAge15_49WithAnaemiaMILLIONS_SubSaharanAfrica.png", bar_chart, width = 12, height = 6)
```



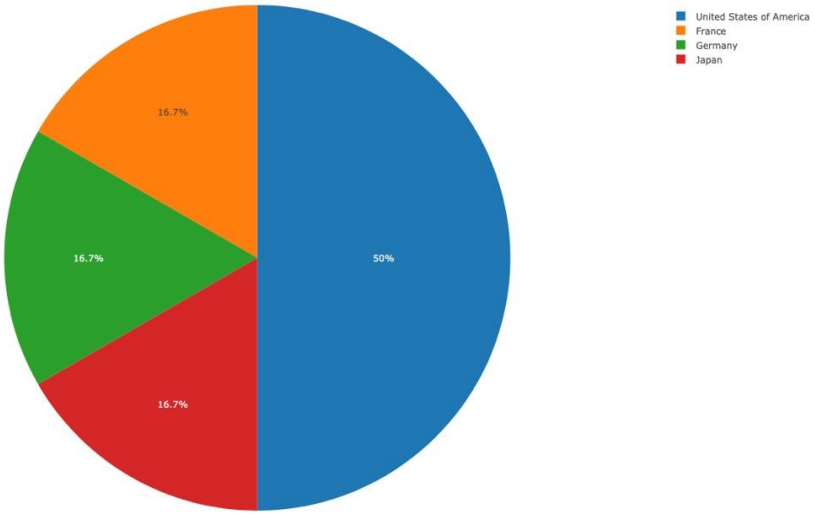
One of the main observations from the graph above is that no data is available for the years 2020,2021,2022 for any of the countries. Moreover, no data is available for Cabo Verde, Sao Tome and Principe and Seychelles.

The year with the highest number of women with anemia between the reproductive age of 15-49 is 2019(higher than 2017 and/or same as 2018). The country with the highest number is Nigeria with up to 26 million women with anemia in 2019 and around 24,23 million in 2018& 2017 respectively. Second highest being Ethiopia at around 6 million, followed by South Africa, Kenya, Sudan, Uganda, and Mozambique. Countries like Djibouti, Equatorial Guinea, Eswatini, Mauritius being the lowest. A common pattern observed is that every country in all three years has a similar number i.e. no big differences with a new year.

Since Nigeria has the highest number of women with anemia in the reproductive age, it is best to compare results for this with other indicators such as low birthweight, percentage of infants who get breastfed.



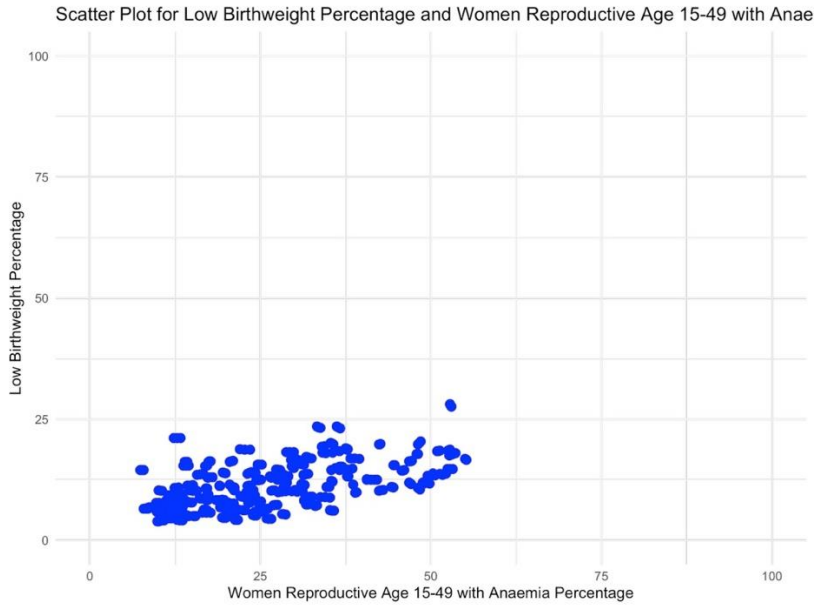
The pie chart shows the Number of Newborns with low birthweight in the Sub-Saharan African region. The above was made through R(Plotly). The countries with the highest number of newborns with low birthweight from the chart are Angola, Madagascar, Mozambique all of whom have a relatively high number of women with anaemia in their reproductive age hence a possible link could be drawn. However, problems the pie chart are that it is an aggregate of all years -not one specific and that some data is missing (Nigeria). The pie chart below shows the same variable however, for a group of 'developed' countries.



This pie chart has however produced interesting results when it gives the United States 50% of the pie while counting the number of children with low birthweight. Because as a developed country which has a good healthcare sector, it should not be as high, hence the either the data

is incorrect (missing values or problematic values) or not represented accurately. This situation can be regarded as a negative result in the analysis.

A more general idea of the same is represented below, however, the scatter plot shows % of women instead of number in millions.



In the graph above, a cluster/ group is formed between 12% and 25% and then near 50%, This shows the percentage of women who have anemia between the reproductive age of 15-49 and this cluster is formed where a low birthweight in children is at a LOW (less than 25%). So hence, it can be said there is not directly correlation.

Cluster Analysis- The following is a cluster analysis done with the variables that represent Number of women of reproductive age of 15-49, %Infants breastfeed between the age of 0-5months, Number of Newborn with a low birthweight. A cluster analysis makes clusters or groups segments of data together that will help find underlying relationships and helps find patterns or trends. Missing or problematic values are found in the data and the processed is stopped if any are found. Ensures the data is clean and ready to use. sets a random seed for reproducibility, specifies the number of clusters (k), and attempts to perform k-means clustering. If clustering is successful, this assigns cluster labels to the original dataset based on the k-means results. This attempts to aggregate the data by cluster, calculating the mean for each variable within each cluster. The number of clusters (k) is set to 3 in this example. The goal is to identify patterns or groups within the data based on the chosen variables.

```
> # Display the cluster centers
> print(km_model$centers)
  NumNewbornLowBirthweightMILLIONS ExclusiveBreastfeedInfants0_5monthsPCT NumWomenRepAge15_49WithAnaemiaMILLIONS
1                        0.3196018                              45.34474                                2.908994
2                        0.3274808                              23.97500                                4.728682
3                        6.6666667                              49.15853                              184.733333
>
> # Check the size of each cluster
> table(dt$Cluster)

   1    2    3
1872   56    6
```

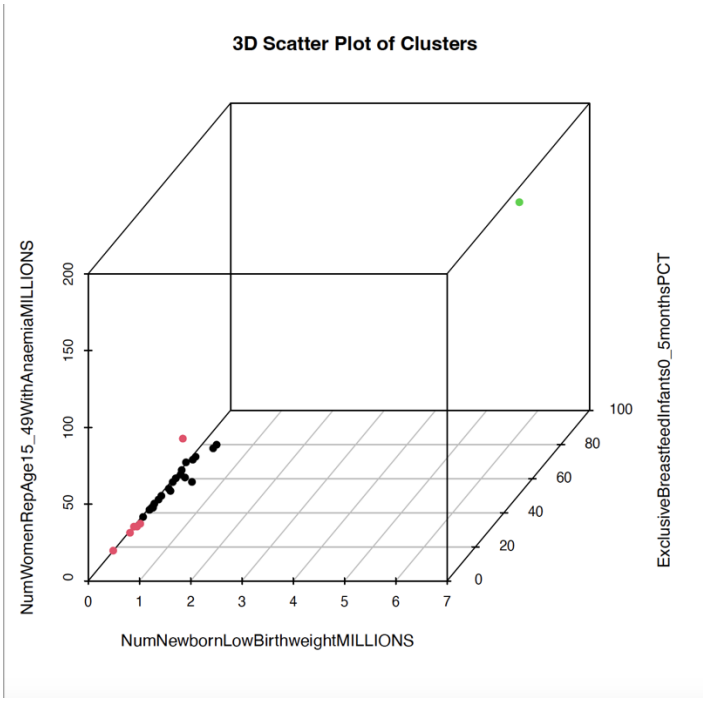
```
>
> # Explore mean values by cluster
> mean_values_by_cluster <- aggregate(. ~ Cluster, data = selected_variables_analysis, FUN = mean)
> print(mean_values_by_cluster)
  Cluster NumNewbornLowBirthweightMILLIONS ExclusiveBreastfeedInfants0_5monthsPCT
NumWomenRepAge15_49WithAnaemiaMILLIONS
1       1                       0.2255814                              51.15116
5.172093
2       2                       0.2833333                              25.90000
13.200000
3       3                       6.8000000                              58.00000
182.200000
```

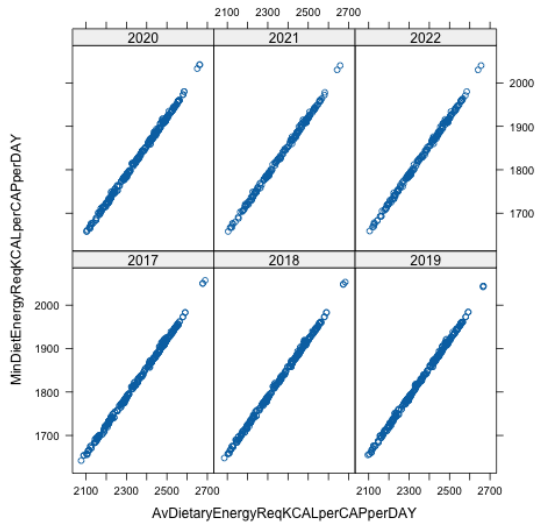
Cluster 1: this cluster appears to represent regions with a moderate number of newborns with low birth weight, a high rate of exclusive breastfeeding, and a low number of women aged 15-49 with anaemia.

Cluster 2: This cluster seems to represent regions with moderate values across all three indicators, suggesting a balanced profile.

Cluster 3 : This cluster may represent regions with a high number of newborns with low birth weight, a relatively high rate of exclusive breastfeeding, and a high number of women aged 15-49 with anaemia.

These results are shown using the following 3Dscatter plot:





The xy plot above is generated using a lattice package on R it shows the relationship between the dietary requirements and Minimum for an aggregate of the world not specific to the Sub-Saharan Region over the years. For interpretation, we can see that it is a direct positive relationship as there is not a lot of difference between the two indicators. A reason for this could the knowledge people use to keep their energy and calories in check. The values are

concentrated till 2000kcals after that there are exceptions as several factors are involved in the analysis of these. The rest can be considered “outliers “for these indicators.