HIS Project - Analysis of world's WASH condition

Giorgio Bonardi - 727671

Introduction

In this report I will analyze some dataset concerning the exposure and burden of diseases related to **water, sanitation, and hygiene**. The main objective is to examine the historical progression and current status of WASH condition in different countries worldwide. Additionally, we aim to identify any potential correlation between WASH conditions and mortality rates from specific diseases.

Technical details

The datasets were retrieved from the website of the World Health Organization: https://www.who.int. More specifically, from https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/water-sanitation-and-hygiene-exposure and https://www.who.int/data/gho/data/themes/topics/water-sanitation-and-hygiene-burden-of-disease.

Preparation

Library

```
suppressPackageStartupMessages({
  library("tidyverse")
  library("maps")
  library("janitor")
  library("tibble")
  library("gridExtra")

  library("forecast")
})
```

Note: the package 'tidyverse' include a set of packages among which: dplyr, stringr, ggplot2, forcats.

Import data set

We use the *read.csv* function, which gives us a data frame. Let's create four primary data frames, from which sub-data frames will be created based on the requirements.

```
drinking_water <- read.csv("drinking-water_services.csv")
sanitation_services <- read.csv("managed_sanitation_services.csv")

wash_deaths2019 <- read.csv("wash_deaths.csv")
diarrhea_deaths_2019 <- read.csv("diarrhea_deaths_w_s_h.csv")
```

Data Cleaning

Now that we have imported the data, we can proceed to clean it.

Names

The janitor package provides the clean_names() function which handles problematic variable names.

Cleaning Steps:

- Returns names with only lowercase letters, using _ as a separator.
- Handles special characters and spaces.
- Appends numbers to duplicated names.
- Converts "%" to "percent" to retain meaning.

Example of column names before cleaning the names:

```
colnames(drinking_water)
```

```
##  [1] "IndicatorCode"              "Indicator"
##  [3] "ValueType"                  "ParentLocationCode"
##  [5] "ParentLocation"             "Location.type"
##  [7] "SpatialDimValueCode"        "Location"
##  [9] "Period.type"                "Period"
## [11] "IsLatestYear"               "Dim1.type"
## [13] "Dim1"                       "Dim1ValueCode"
## [15] "Dim2.type"                  "Dim2"
## [17] "Dim2ValueCode"              "Dim3.type"
## [19] "Dim3"                       "Dim3ValueCode"
## [21] "DataSourceDimValueCode"     "DataSource"
## [23] "FactValueNumericPrefix"     "FactValueNumeric"
## [25] "FactValueUoM"               "FactValueNumericLowPrefix"
## [27] "FactValueNumericLow"        "FactValueNumericHighPrefix"
## [29] "FactValueNumericHigh"       "Value"
## [31] "FactValueTranslationID"     "FactComments"
## [33] "Language"                   "DateModified"
```

```
drinking_water <- drinking_water %>% clean_names()
sanitation_services <- sanitation_services %>% clean_names()

wash_deaths2019 <- wash_deaths2019 %>% clean_names()
diarrhea_deaths_2019 <- diarrhea_deaths_2019 %>% clean_names()
```

Example of column names after cleaning the names:

```
colnames(drinking_water)
```

```
##  [1] "indicator_code"              "indicator"
##  [3] "value_type"                  "parent_location_code"
##  [5] "parent_location"             "location_type"
##  [7] "spatial_dim_value_code"      "location"
##  [9] "period_type"                 "period"
## [11] "is_latest_year"              "dim1_type"
## [13] "dim1"                        "dim1value_code"
## [15] "dim2_type"                   "dim2"
## [17] "dim2value_code"              "dim3_type"
```

```
## [19] "dim3"                          "dim3value_code"
## [21] "data_source_dim_value_code"    "data_source"
## [23] "fact_value_numeric_prefix"     "fact_value_numeric"
## [25] "fact_value_uo_m"               "fact_value_numeric_low_prefix"
## [27] "fact_value_numeric_low"        "fact_value_numeric_high_prefix"
## [29] "fact_value_numeric_high"       "value"
## [31] "fact_value_translation_id"     "fact_comments"
## [33] "language"                      "date_modified"
```

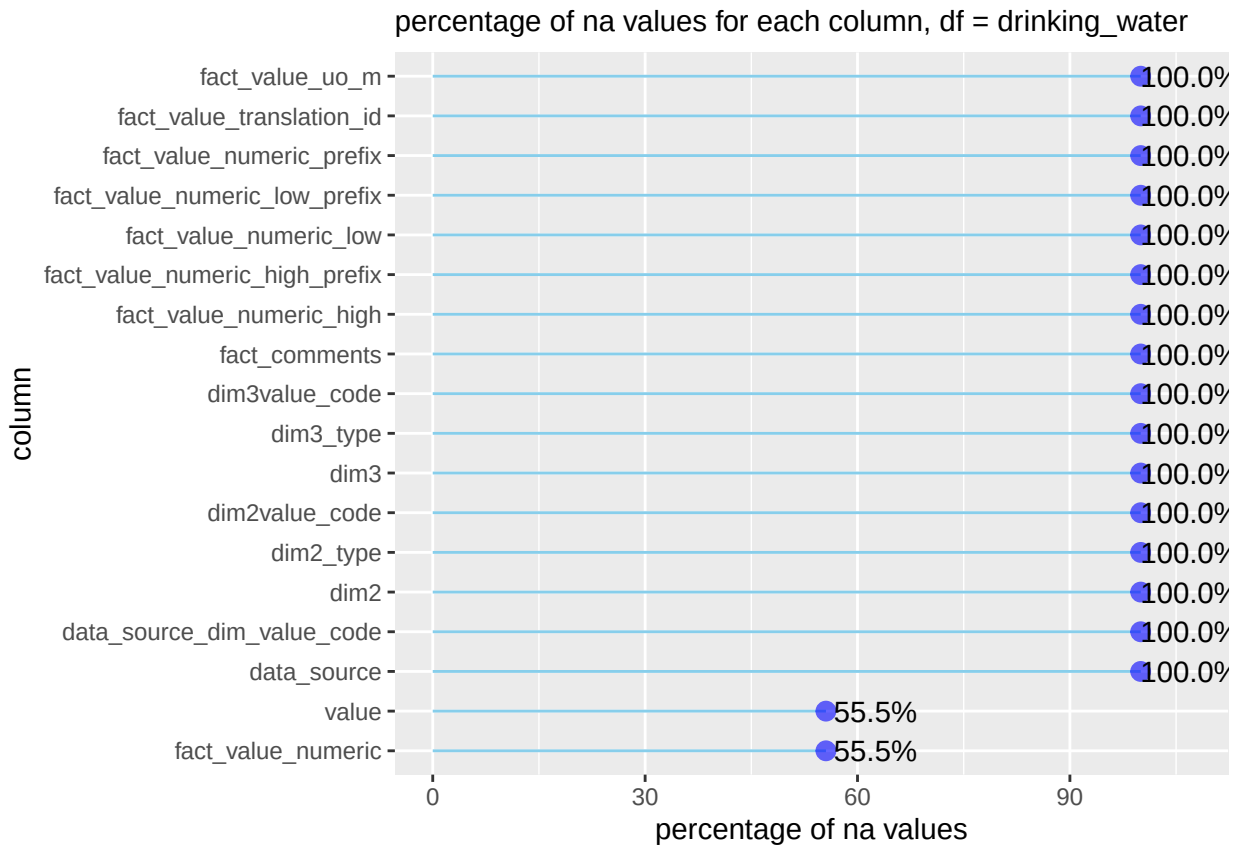
NA Value

We can analyze the percentage of NA values in the columns to evaluate the reliability of the data.

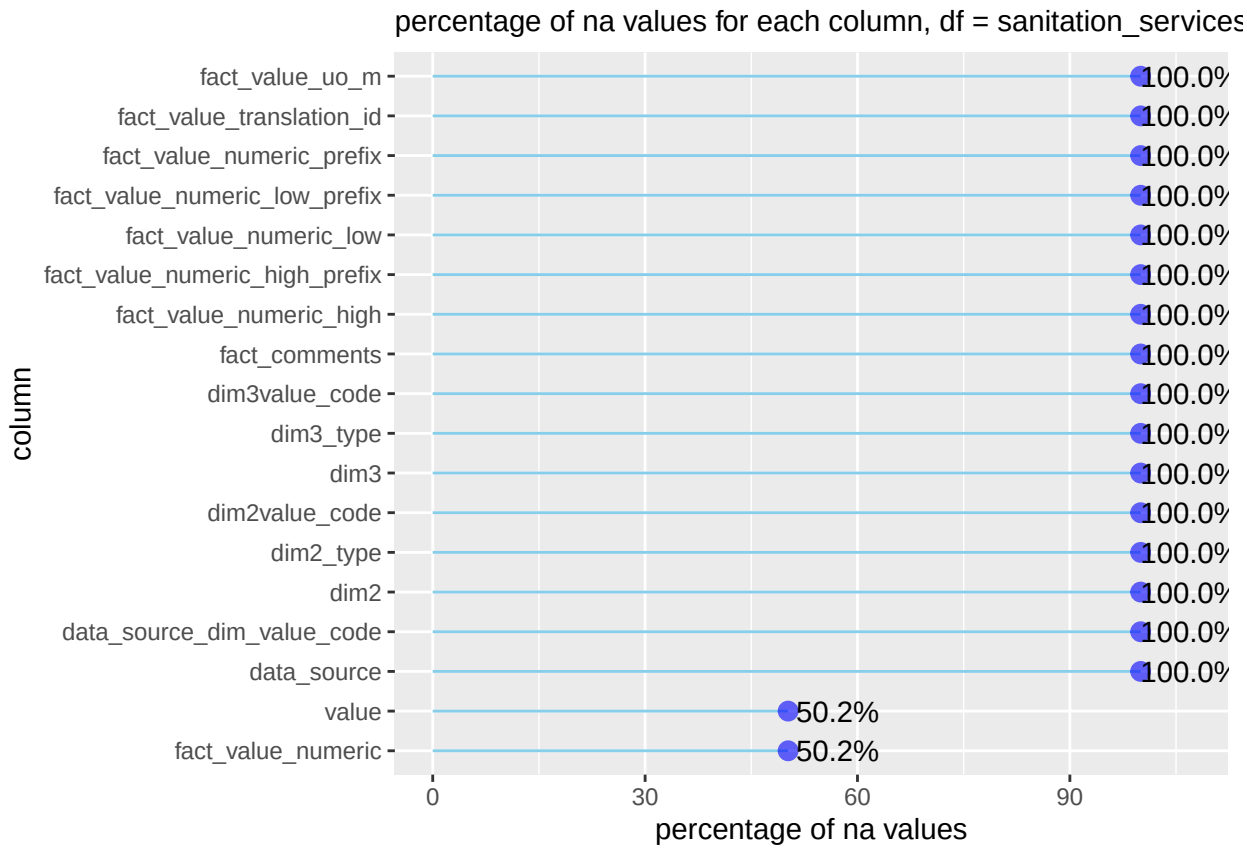
Note that the plots will only show the columns that have a % greater than 0.

```
missing <- colMeans(is.na(drinking_water)) * 100
missing <- missing[missing >= 1]
missing <- data.frame(missing)
missing <- rownames_to_column(missing, var = "row_name")
missing_drinking_water<- ggplot(data = missing, aes(
  x = fct_reorder(row_name, missing),
  y = missing,
  label = sprintf("%0.1f%%", missing)
)) +
  geom_segment(aes(xend = row_name, y = 0, yend = missing), color = "skyblue") +
  geom_point(color = "blue",
             size = 3,
             alpha = 0.6) +
  geom_text(nudge_y = 7) +
  coord_flip() +
  labs(title = "percentage of na values for each column, df = drinking_water",
       x = "column", y = "percentage of na values") +
   theme(plot.title = element_text(size = 11))
```

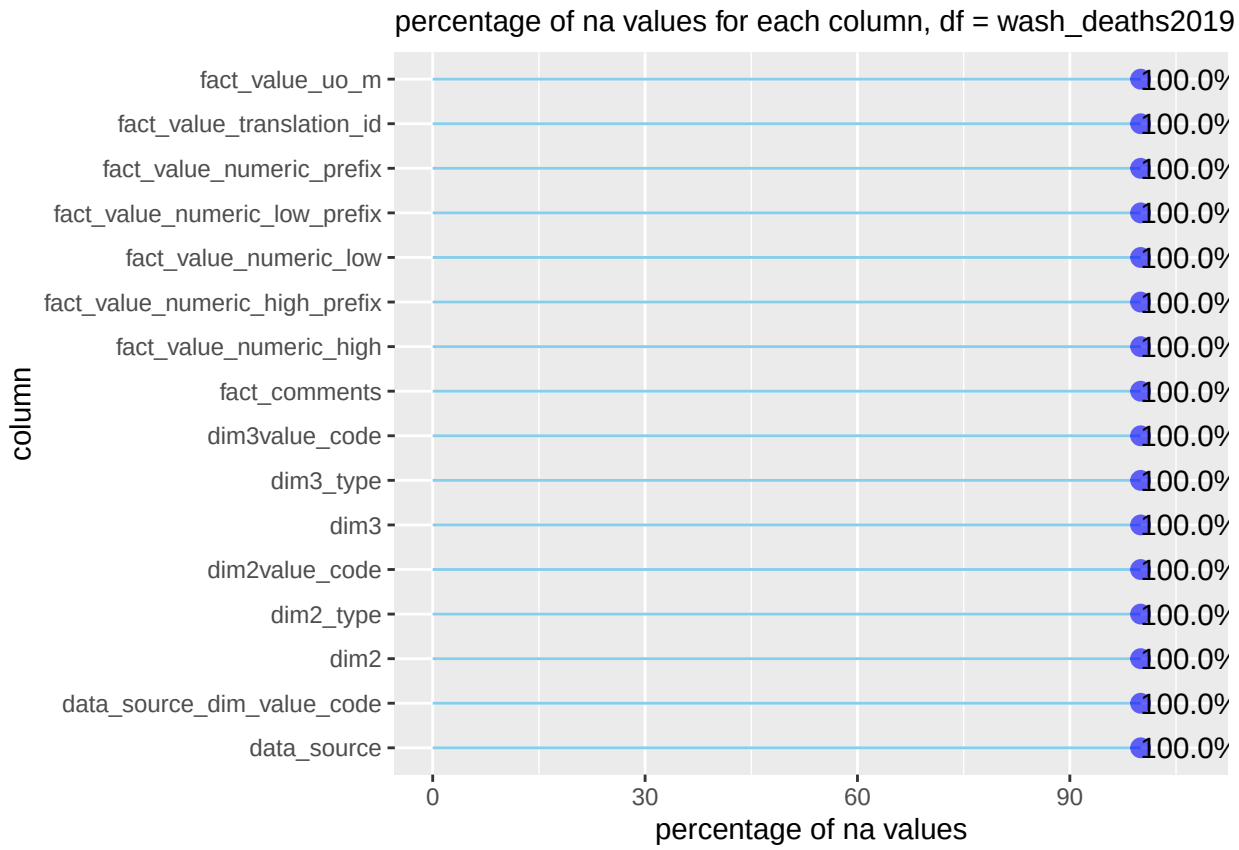
```
print(missing_drinking_water)
```



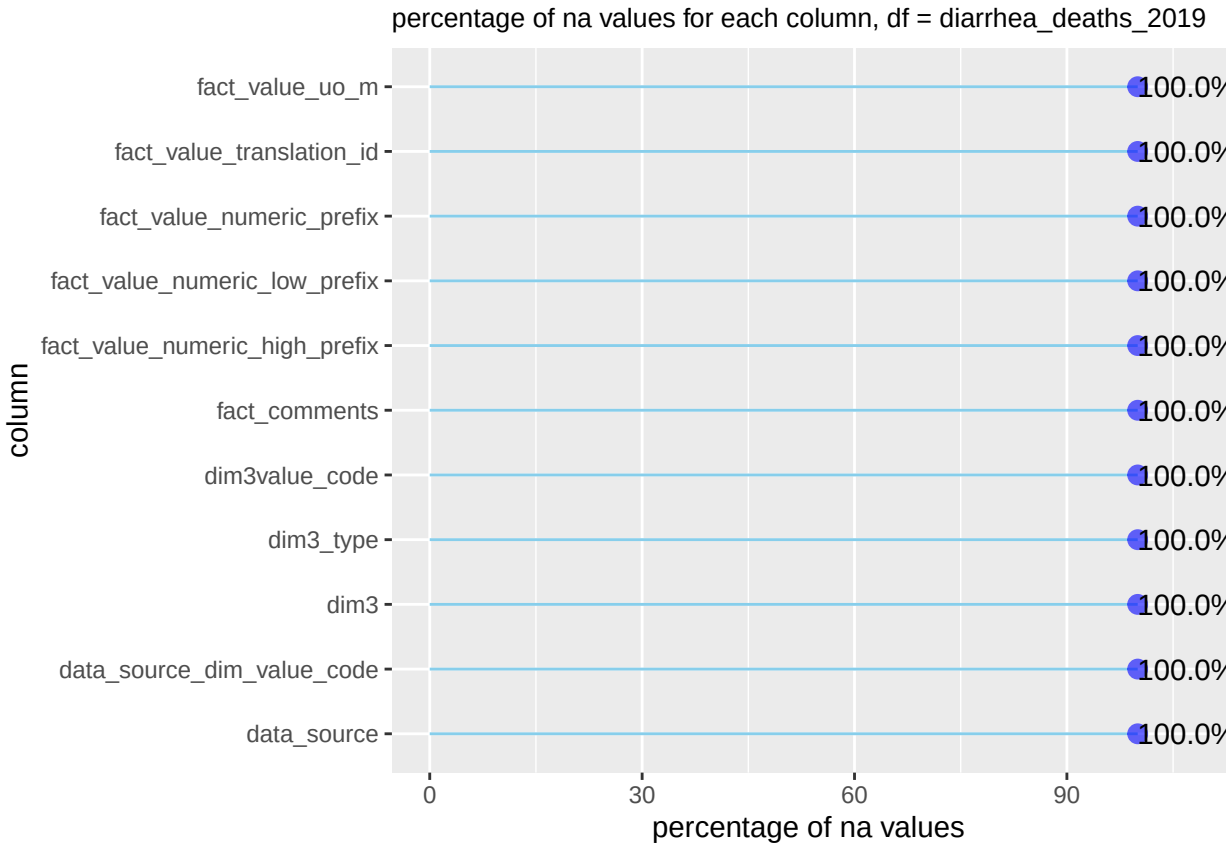

```
print(missing_sanitation_services)
```




```
print(missing_wash_deaths2019)
```




```
print(missing_diarrhea_deaths_2019)
```



Columns that contain only NA values can be removed. Additionally, columns such as 'indicator_code' or 'indicator' that have the same value for all rows can also be removed.

```
drinking_water <- subset(drinking_water, select =
                           c("parent_location", "location", "period", "dim1", "value"))
sanitation_services <- subset(sanitation_services, select =
                                c("parent_location", "location", "period", "dim1", "value"))

wash_deaths2019 <- subset(wash_deaths2019, select =
                           c("parent_location", "location", "dim1", "fact_value_numeric"))
diarrhea_deaths_2019 <- subset(diarrhea_deaths_2019, select =
                           c("parent_location", "location", "dim1", "dim2", "fact_value_numeric"))
```

Rename columns

Now we will rename some columns to make them more meaningful

```
colnames(drinking_water)[colnames(drinking_water) == "location"] = "region"
colnames(sanitation_services)[colnames(sanitation_services) == "location"] = "region"

colnames(wash_deaths2019)[colnames(wash_deaths2019) == "location"] = "region"
colnames(diarrhea_deaths_2019)[colnames(diarrhea_deaths_2019) == "location"] = "region"
```

```
colnames(drinking_water)[colnames(drinking_water) == "value"] <- "water_value"
colnames(sanitation_services)[colnames(sanitation_services) == "value"] <- "sanitation_value"
```

```
colnames(wash_deaths2019)[colnames(wash_deaths2019) == "fact_value_numeric"] = "wash_deaths"
colnames(diarrhea_deaths_2019)[colnames(diarrhea_deaths_2019)
                               == "fact_value_numeric"] ="diarrhea_deaths"
```

```r
colnames(drinking_water)[colnames(drinking_water) == "dim1"] = "ambient"
colnames(sanitation_services)[colnames(sanitation_services) == "dim1"] <- "ambient"

colnames(wash_deaths2019)[colnames(wash_deaths2019) == "dim1"] = "sex"
colnames(diarrhea_deaths_2019)[colnames(diarrhea_deaths_2019) == "dim1"] = "sex"
```

```r
nrow(drinking_water)
```

```
## [1] 12348
```

```r
nrow(sanitation_services)
```

```
## [1] 12348
```

```r
nrow(wash_deaths2019)
```

```
## [1] 549
```

```r
nrow(diarrhea_deaths_2019)
```

```
## [1] 732
```

The **drinking water** and **sanitation service** datasets have 12348 rows.

The **wash deaths dataset** has 549 rows and the **diarrhea deaths dataset** has 732 rows.

Merge dataset

Now let's proceed to create two sub-dataframes from the main data frames. One will contain the data related to drinking water and sanitation services and the other will contain the data related to deaths.

```r
completeWASHDf <- merge(drinking_water, sanitation_services,  by =
                          c("parent_location", "region", "period", "ambient"))
complete_death <- merge(wash_deaths2019, diarrhea_deaths_2019, by =
                          c("parent_location", "region", "sex"))
```

```r
nrow(completeWASHDf)
```

```
## [1] 12348
```

```r
nrow(complete_death)
```

```
## [1] 732
```

The **complete WASH** dataset after merging the drinking water and sanitation service datasets still has 12348 rows.

The **complete death** dataset after merging the wash deaths and diarrhea deaths datasets has now 732 rows.

Subsetting

I want only the Total field of their values to remain, so we need to discard all Rural/Urban values

```r
completeWASHDf <- completeWASHDf %>% filter(ambient == "Total")
```

Let's create a relevant subset for the 'deaths' data frame that we can use later

```r
both_sex_death <- subset(complete_death, sex == "Both sexes" & dim2 == "All age groups (total)")
```

```r
nrow(completeWASHDf)
```

```
## [1] 4116
```

```r
nrow(both_sex_death)
```

```
## [1] 182
```

The **complete WASHDf** dataset after the subsetting has 4116 rows.

The **both sex death** dataset after subsetting the complete death dataset has now 182 rows.

Population Dataset

Now we will introduce a dataset containing the population for each country so that we can "normalize" the data. However, before doing that, we need to adjust the dataset.

```r
worldPopulation <- read.csv("world_population.csv")
worldPopulation <- worldPopulation %>% clean_names()
worldPopulation <- subset(worldPopulation, select = c("country", "population"))
colnames(worldPopulation)[colnames(worldPopulation) == "country"] = "region"
```

Let's standardize the region names to ensure compatibility across the various datasets.

First of all, let's identify the regions with different names and correct them:

```r
setdiff(both_sex_death$region, worldPopulation$region)
```

```
##  [1] "Côte d'Ivoire"
##  [2] "Democratic Republic of the Congo"
##  [3] "Sao Tome and Principe"
##  [4] "United Republic of Tanzania"
##  [5] "Bolivia (Plurinational State of)"
##  [6] "United States of America"
##  [7] "Venezuela (Bolivarian Republic of)"
##  [8] "Iran (Islamic Republic of)"
##  [9] "Syrian Arab Republic"
## [10] "Czechia"
## [11] "Republic of Moldova"
## [12] "Russian Federation"
## [13] "The former Yugoslav Republic of Macedonia"
## [14] "Türkiye"
## [15] "United Kingdom of Great Britain and Northern Ireland"
## [16] "Democratic People's Republic of Korea"
## [17] "Brunei Darussalam"
## [18] "Lao People's Democratic Republic"
## [19] "Micronesia (Federated States of)"
## [20] "Republic of Korea"
## [21] "Viet Nam"
```

```r
worldPopulation <- worldPopulation %>%
  mutate(region = recode(str_trim(region), "United States" = "USA",
                            "United Kingdom" = "UK",
                            "Congo" = "Republic of Congo",
                         "DR Congo"="Democratic Republic of the Congo",
                         "Bolivia (Plurinational State of)" = "Bolivia",
                         "Iran (Islamic Republic of)" = "Iran",
                         "Türkiye" = "Turkey",
                         "Côte d'Ivoire" = "Ivory Coast",
                         "Sao Tome & Principe" = "Sao Tome and Principe",
                         "Czech Republic (Czechia)" = "Czech Republic"))
```

```
both_sex_death <- both_sex_death %>%
  mutate(region = recode(str_trim(region), "United States of America" = "USA",
                            "United Kingdom of Great Britain and Northern Ireland" = "UK",
                            "Democratic People's Republic of Korea" = "North Korea",
                            "Republic of Korea" = "South Korea",
                          "Congo"="Republic of Congo",
                          "Russian Federation" = "Russia",
                          "Iran (Islamic Repulic of)" = "Iran",
                          "United Republic of Tanzania" = "Tanzania",
                          "Bolivia (Plurinational State of)" = "Bolivia",
                          "Iran (Islamic Republic of)" = "Iran",
                          "Türkiye" = "Turkey",
                          "Côte d'Ivoire" = "Ivory Coast",
                          "Viet Nam"= "Vietnam",
                          "Lao People's Democratic Republic" = "Laos",
                          "Brunei Darussalam" = "Brunei",
                          "Republic of Moldova" = "Moldova",
                          "The former Yugoslav Republic of Macedonia" = "North Macedonia",
                          "Czechia" = "Czech Republic",
                          "Syrian Arab Republic" = "Syria",
                          "Venezuela (Bolivarian Republic of)" = "Venezuela",
                          "Micronesia (Federated States of)"= "Micronesia"))
```

Now, let's merge the 'worldPopulation' data frame with the 'wash_deaths2019' and create new columns that represent the ratios between the values and the population.

```
both_sex_death <- left_join(both_sex_death, worldPopulation, by = "region")
```

```
both_sex_death <- both_sex_death %>%
  mutate(ratio_wash_to_region = (wash_deaths / population)*100)
```

```
both_sex_death <- both_sex_death %>%
  mutate(ratio_diarrhea_to_region = (diarrhea_deaths / population)*100)
```

We will follow the same process with the 'completeWASHD' data frame without showing the code.

Life Expectancy Dataset

Now let's import a final dataset that we will later use to identify correlations among the different datasets. We will select only the year 2019 since it is the only year present in the other datasets. We will use the same process as we did with the other datasets before.

```
life_exp <- read.csv("life_expectancy.csv")
life_exp <- life_exp %>% clean_names()
life_exp_2019 <- subset(life_exp, year == 2019)
```

Analyses and Plot

Population using safely managed drinking-water and sanitation services

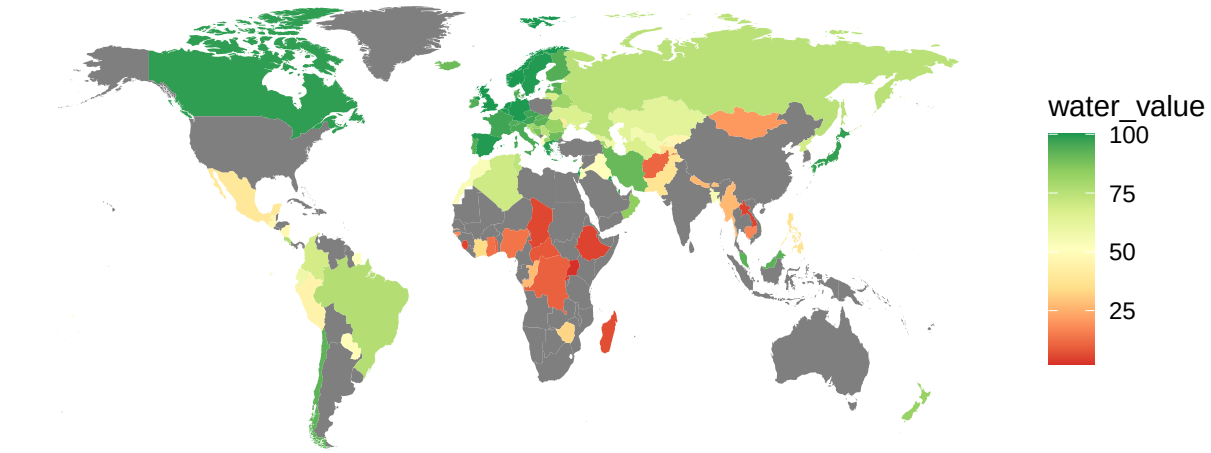
World Map

Now we will represent the values of the dataframes using a world map.

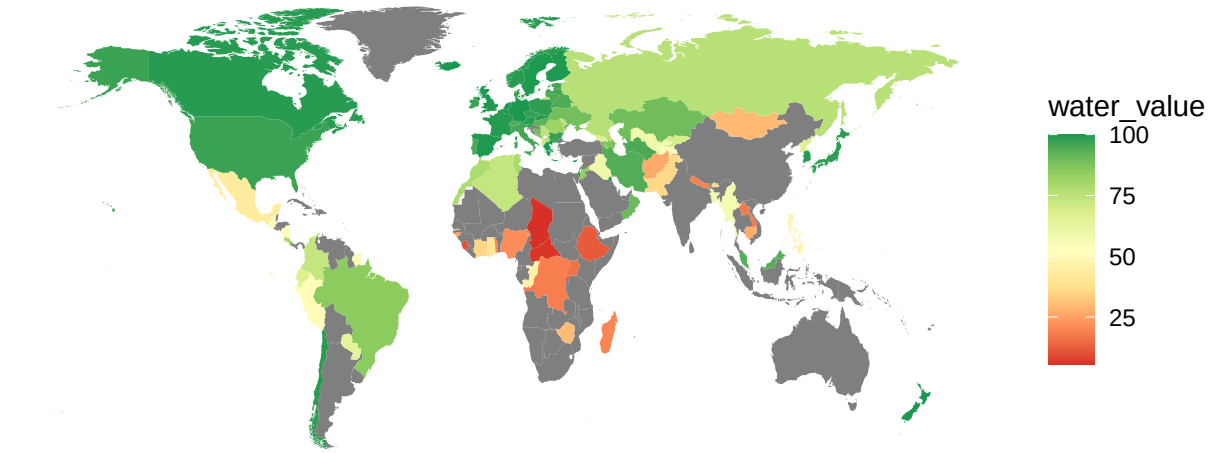
```
world <- map_data("world")
```

Note: The *gray* color represents the regions that did not have a value within the dataset.

```
plot(worldValueWater2000)
```



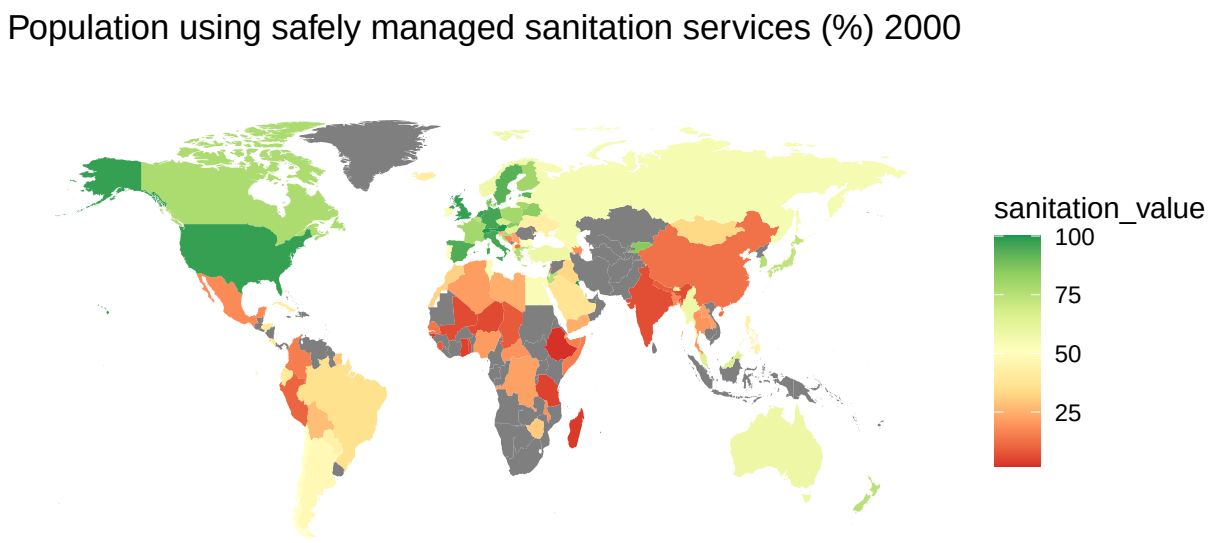

```
plot(worldValueWater2019)
```



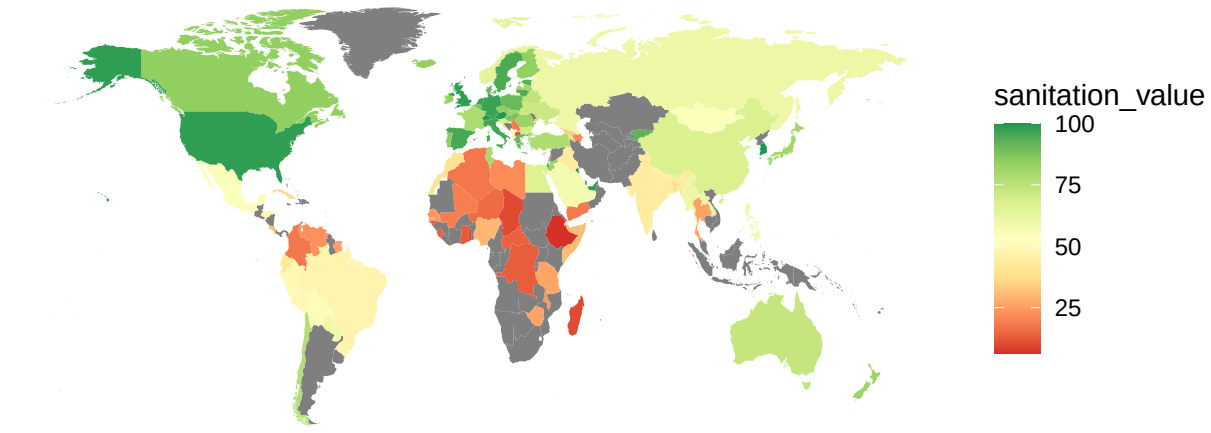
Population using safely managed drinking-water services (%): This plot represents the population using safely managed drinking-water services in 2019 across different regions of the world. The color shading on the map indicates the percentage of the population that has access to safe drinking-water services, with greener shades representing higher percentages. This plot helps visualize the disparities in access to safe drinking water across regions. It is evident that some regions, particularly in Europe and parts of North

America, have higher percentages of the population with access to safe drinking water, while certain regions in Africa and parts of Asia have lower percentages. The plot provides an overview of the global distribution of access to safe drinking water. We can conclude that there was an improvement over the years, but a different type of plot later on will assist us in better visualizing the variations between the years.

```
plot(worldValueSanitation2000)
```




```
plot(worldValueSanitation2019)
```



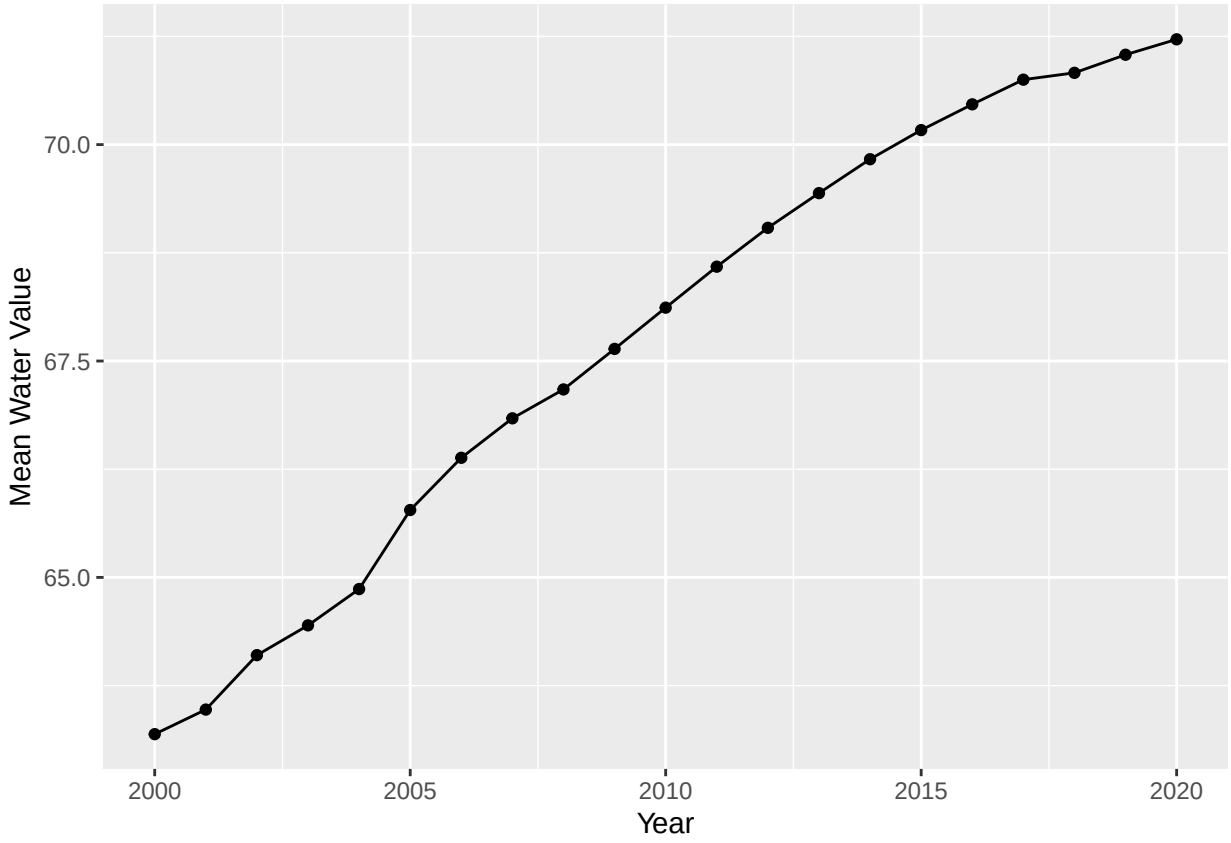
Population using safely managed sanitation services (%): This displays the population using safely managed sanitation services in 2019. Similar to the previous plot, the color shading represents the percentage of the population with access to safe sanitation services, with greener shades indicating higher percentages. The plot shows the variation in access to safe sanitation services across different regions of the world. It reveals that regions like Europe and North America have higher percentages of the population with access to safe sanitation, while regions in Africa and parts of Asia have lower percentages. This plot provides insights into the global disparities in access to safe sanitation facilities. We can clearly observe the improvement between the years in several countries, such as **China** and **India**, which increased from **<25%** to **>50%**.

Overall, these plots effectively visualize the regional differences in access to safe drinking water and sanitation services, highlighting the areas that require more attention and resources to improve access to these essential services.

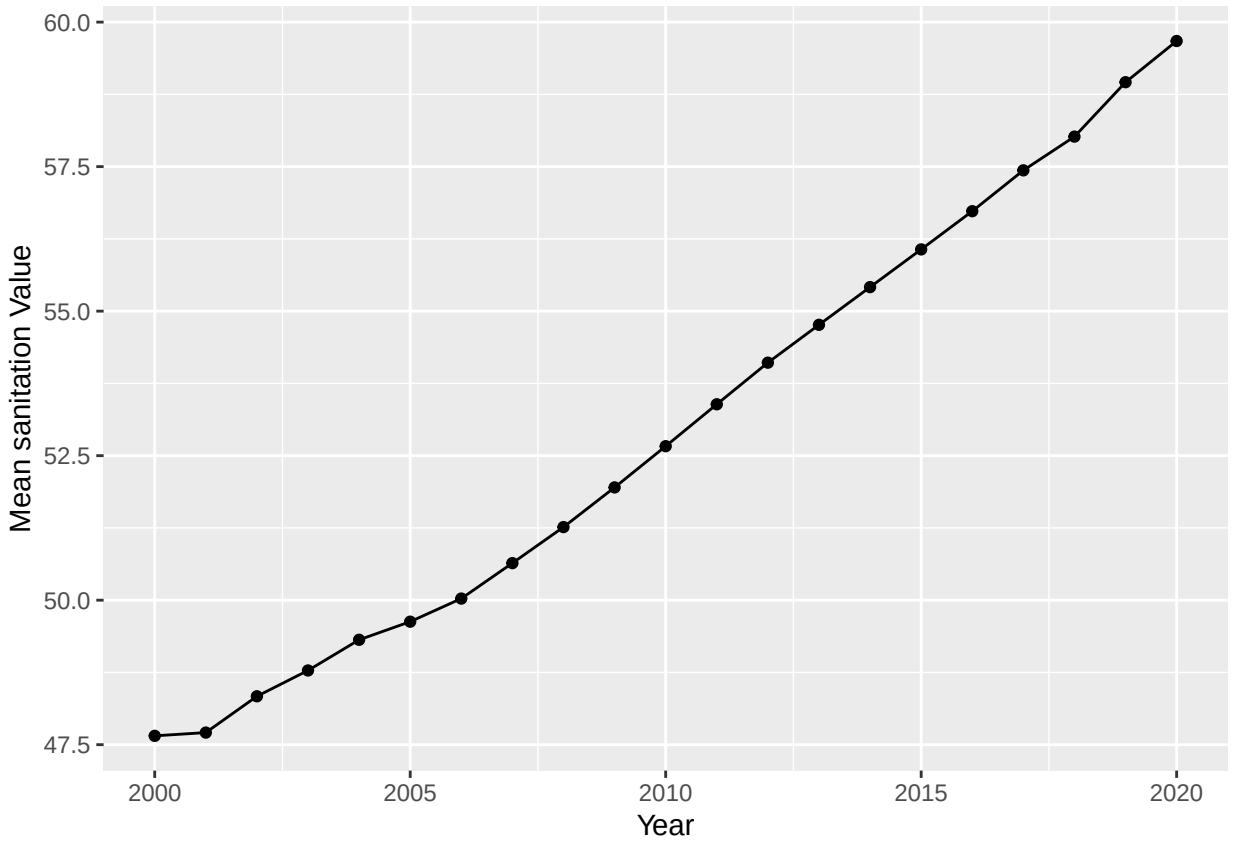
Analysis over the years

Now we want to analyze the data of all continents over the years using other plots. As shown in the world plot earlier, we anticipate that the values will increase.

```
print(over_year_water)
```



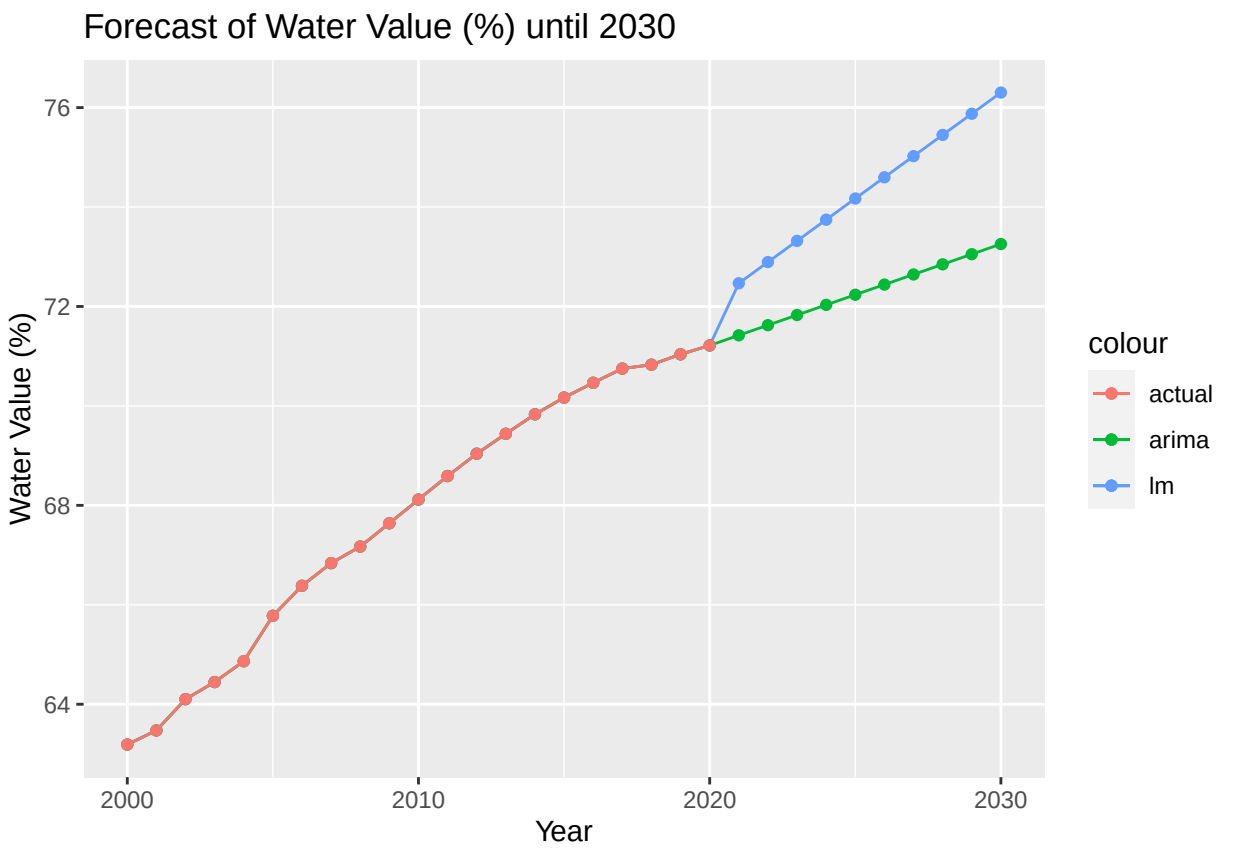

```
print(over_year_sanitation)
```



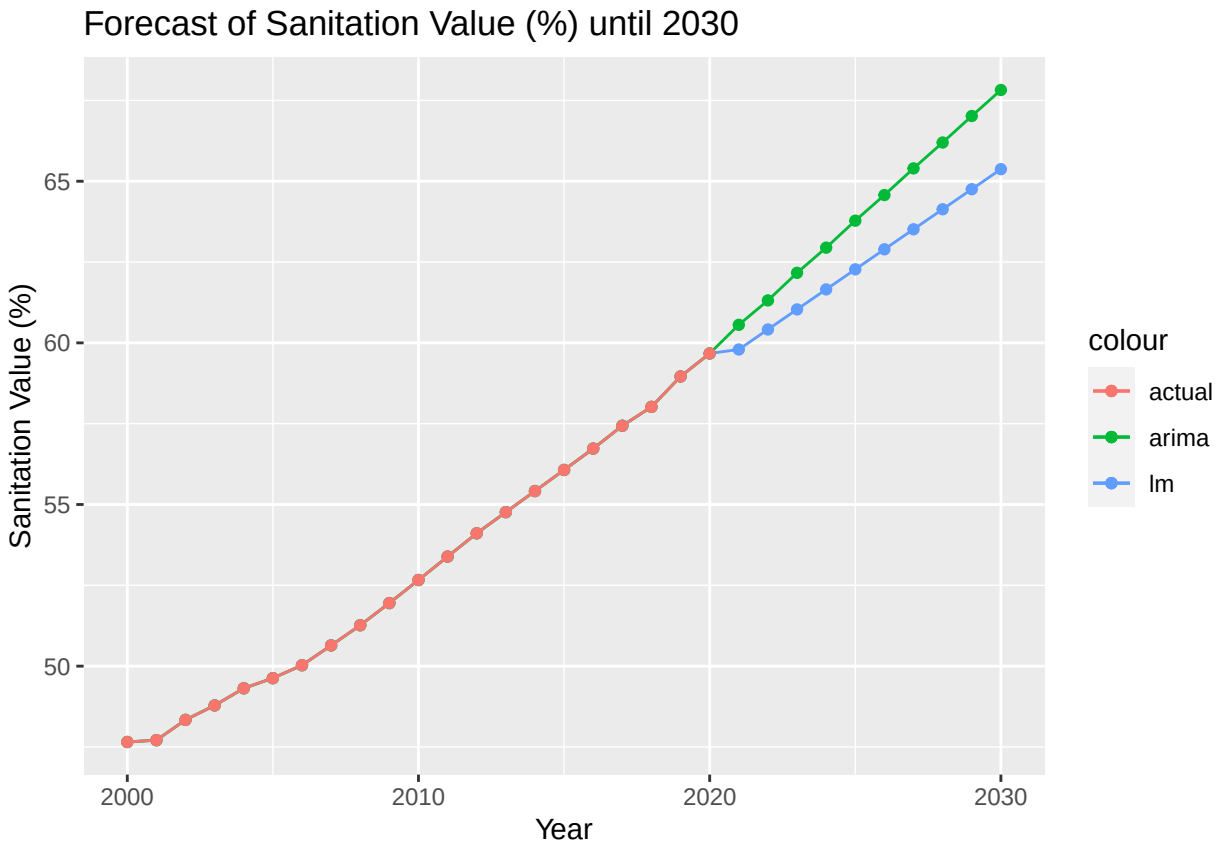
The graphs show the mean values of water and sanitation services over time. Both graphs show an increasing trend, indicating an improvement in access to these essential services. These positive trends reflect the global efforts to enhance water and sanitation infrastructure and promote hygiene practices.

Predictions

```
plot(water_over_years)
```




```
plot(sanitation_over_years)
```

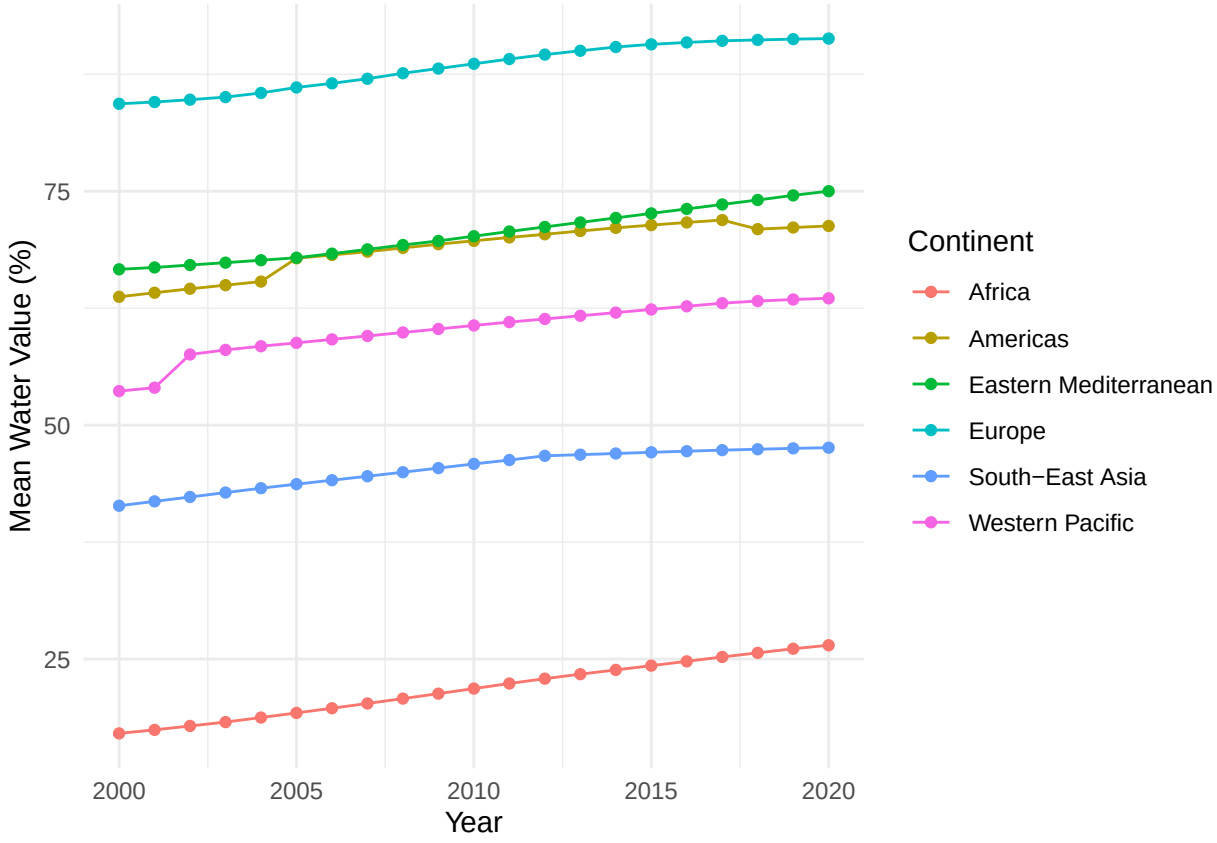


lm, linear regression, is a model that identifies a relationship between the years and the water-related value. It tries to fit a straight line to the existing data to make future predictions along this line. On the other hand, **arima**, AutoRegressive Integrated Moving Average, uses past data to identify patterns of trend, seasonality, and noise in the data, and then use these patterns to make future predictions.

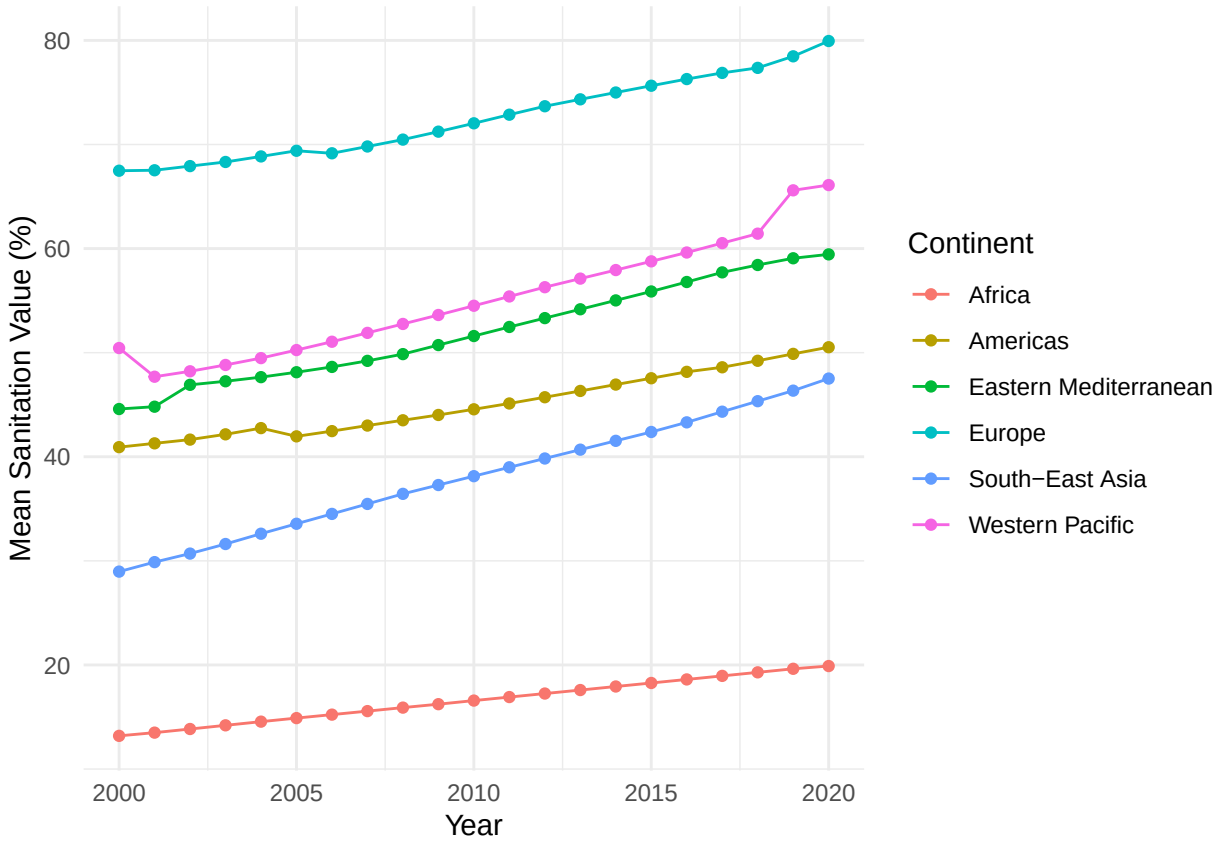
We can observe that, according to predictions, if there will be no significant changes from now until 2030, the water and sanitation problems will persist.

Analysis over the year for each continent

```
print(mean_over_year_water)
```




```
print(mean_over_year_sanitation)
```

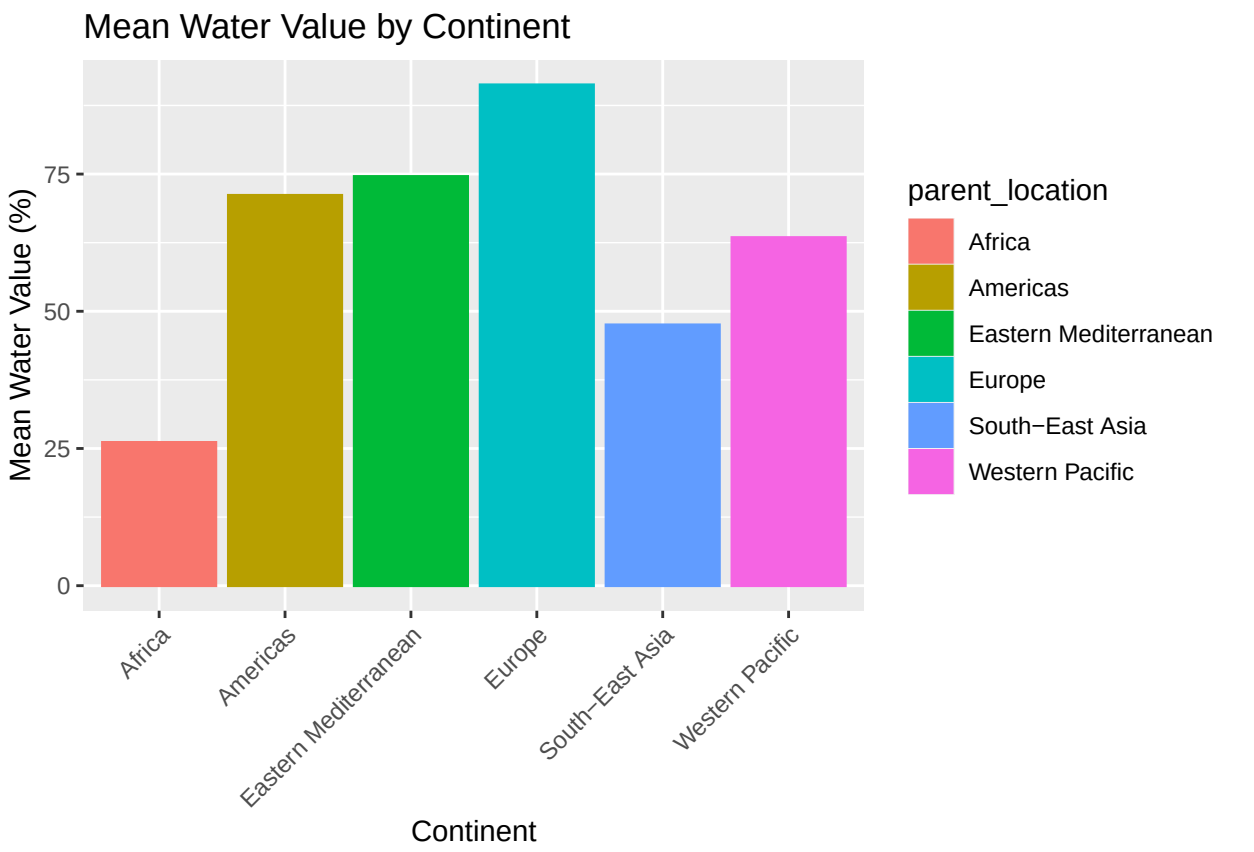


The graphs display the mean values of water and sanitation services over time for each continent. By examining the trends across continents, we can observe variations in the progress of access to these services. The plot illustrates differences and progress in different continents, emphasizing the need for specific actions and investments to address the unique challenges each continent faces in providing sufficient water and sanitation services.

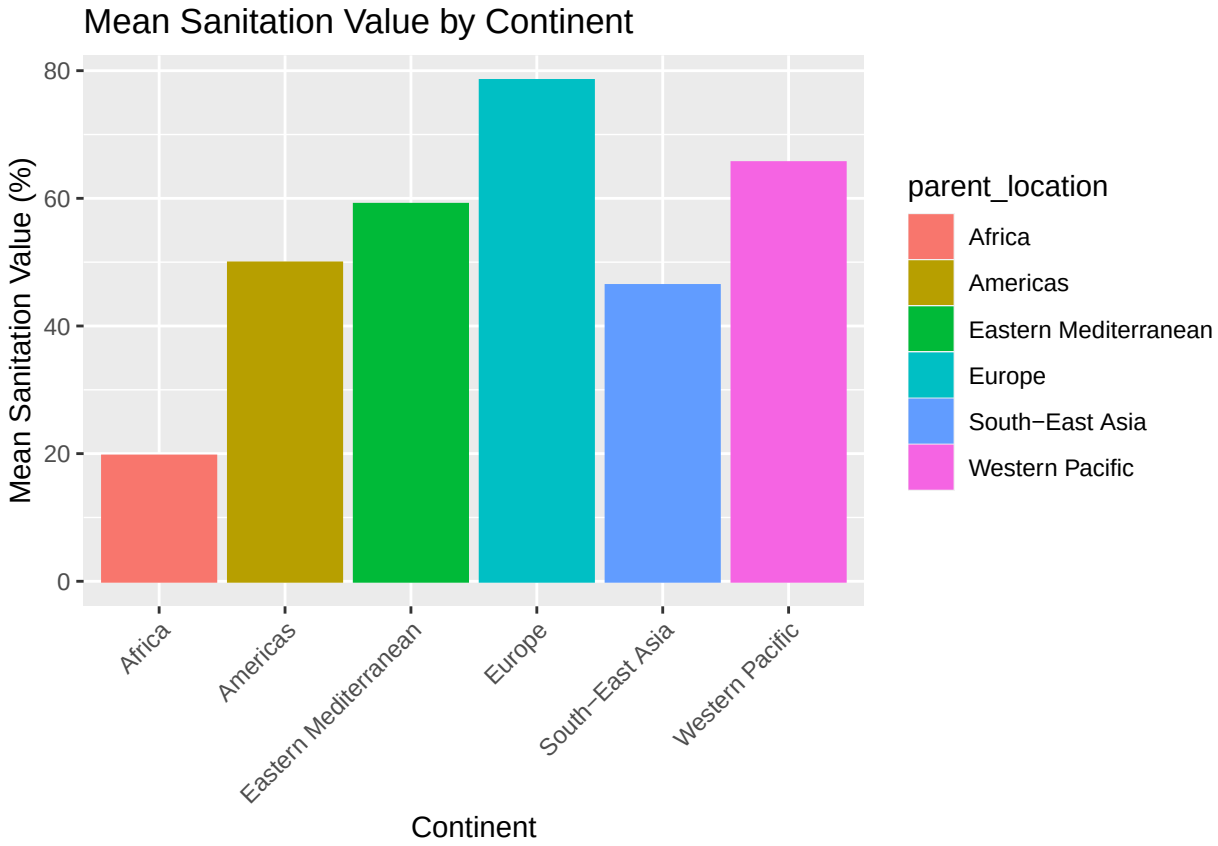
As expected, Europe, America, and the Eastern Mediterranean are above average, while Africa shows a decreasing trend in the mean values. This indicates that we need to focus more on the region of Africa, which is likely being addressed by relevant efforts.

Mean value by continent

```
print(mean_bar_water)
```




```
print(mean_bar_sanitation)
```



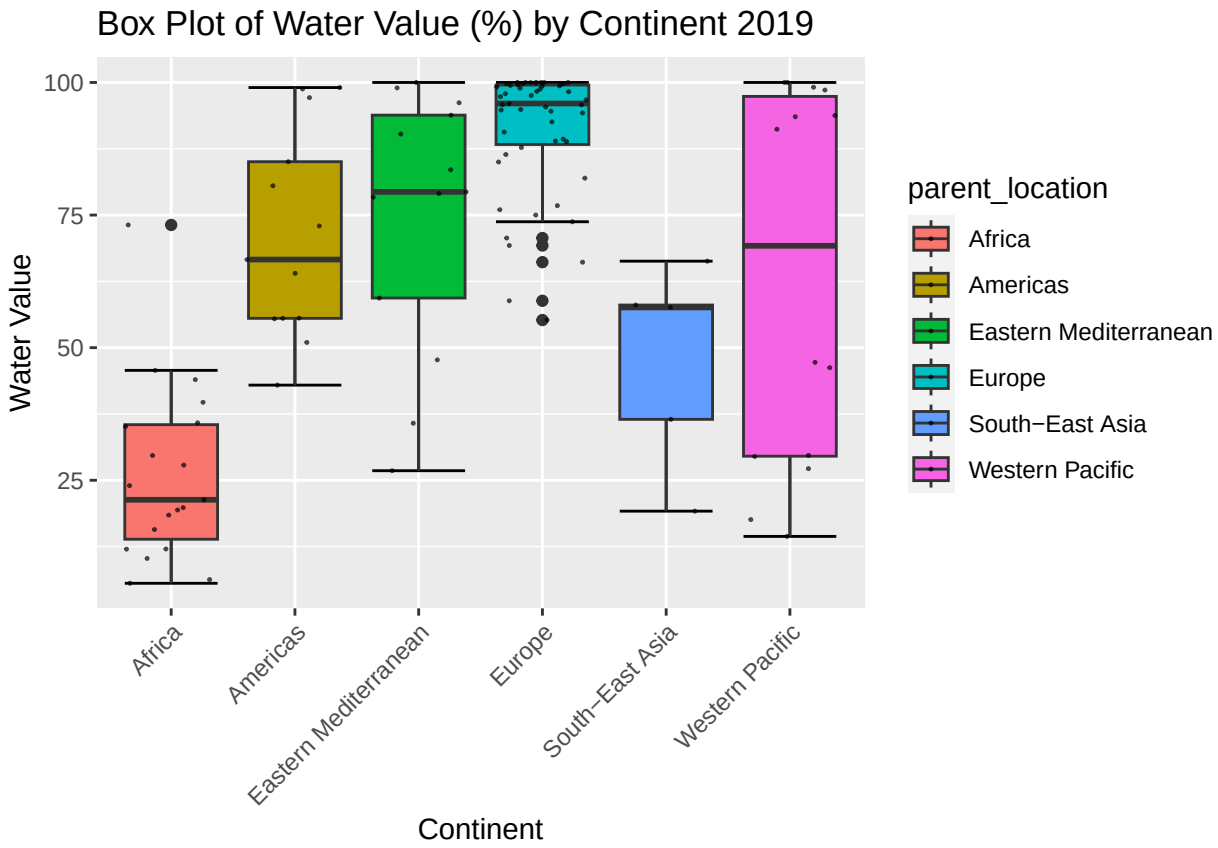
The bar charts represent the mean values of water and sanitation services by continent for the year 2019. The height of each bar indicates the average value for the respective continent. These plots better highlight the differences between each continent. For instance we can observe that, for both sanitation and water value, Africa is less than one-fourth of Europe.

Box Plot

The box plots represent the distribution of water and sanitation values across different continents. The boxes indicate the interquartile range (IQR) of the data, with the median value displayed as a horizontal line within the box. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, while any points outside this range are considered outliers.

Comparing the box plots for water and sanitation values allows us to observe differences in the central tendencies and ranges across continents. The height of the boxes and the position of the medians provide information about the typical values for each continent. The whiskers and outliers indicate the presence of extreme values or potential data anomalies.

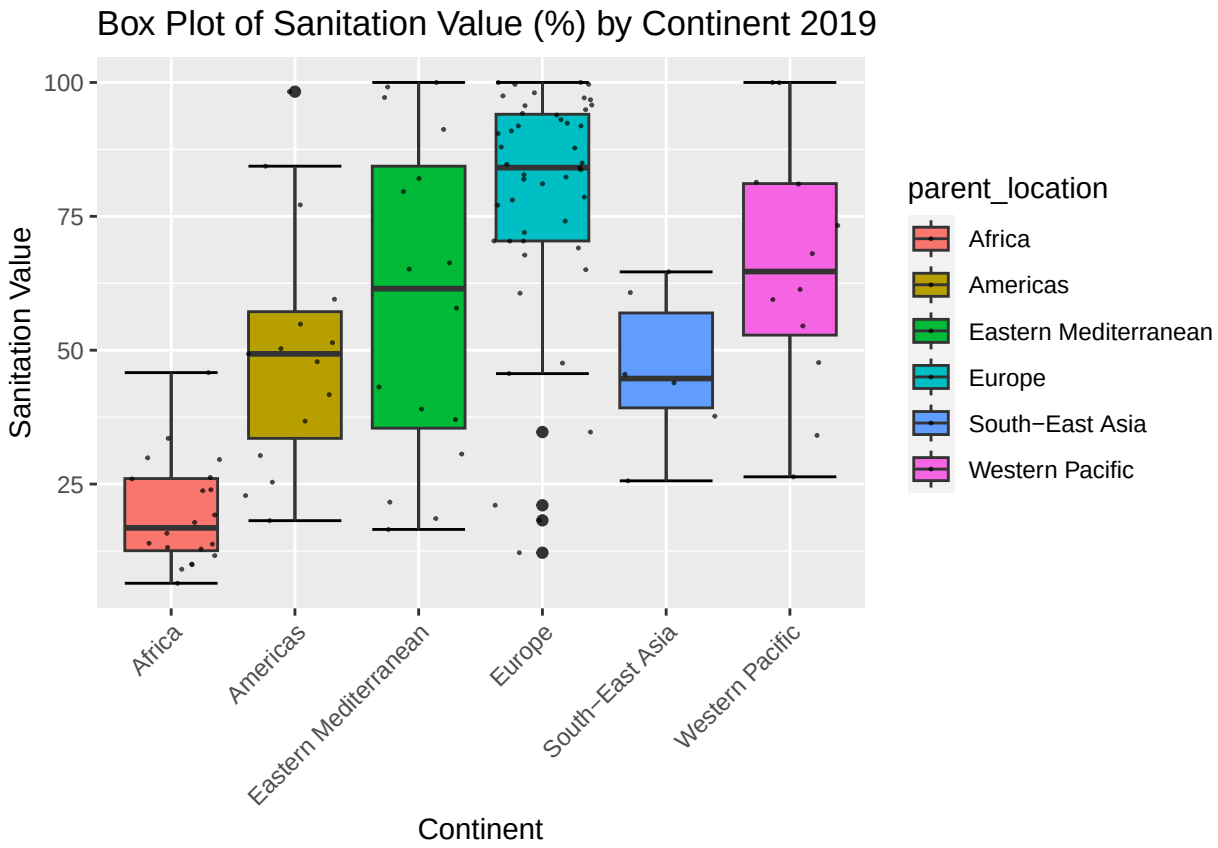
```
plot(boxplot_water)
```



Note: Each point represent the water value of a country in 2019

We can notice how the **box plot** also reflects the situation discovered in the other graphs. However, we can observe that there are many outliers in the box plot for water values in Europe. Additionally, we can notice the presence of outliers in Africa, indicating that there are some regions well above the average. These outliers are also useful to highlight that despite a high average in Europe, there are still countries that need improvements. On the other hand, in Africa, despite the very low average, there are some countries that are doing well.

```
print(boxplot_sanitation)
```



Note: Each point represent the sanitation value of a country in 2019

For the box plot related to the sanitation value, we have, in general, more 'standard' results, with some outliers for **Africa**, indicating that there are countries with a much better situation than the average. However, as seen in the world plot, we observe relatively low values on average for **America**, with lower values in South America and higher values in Central/North America.

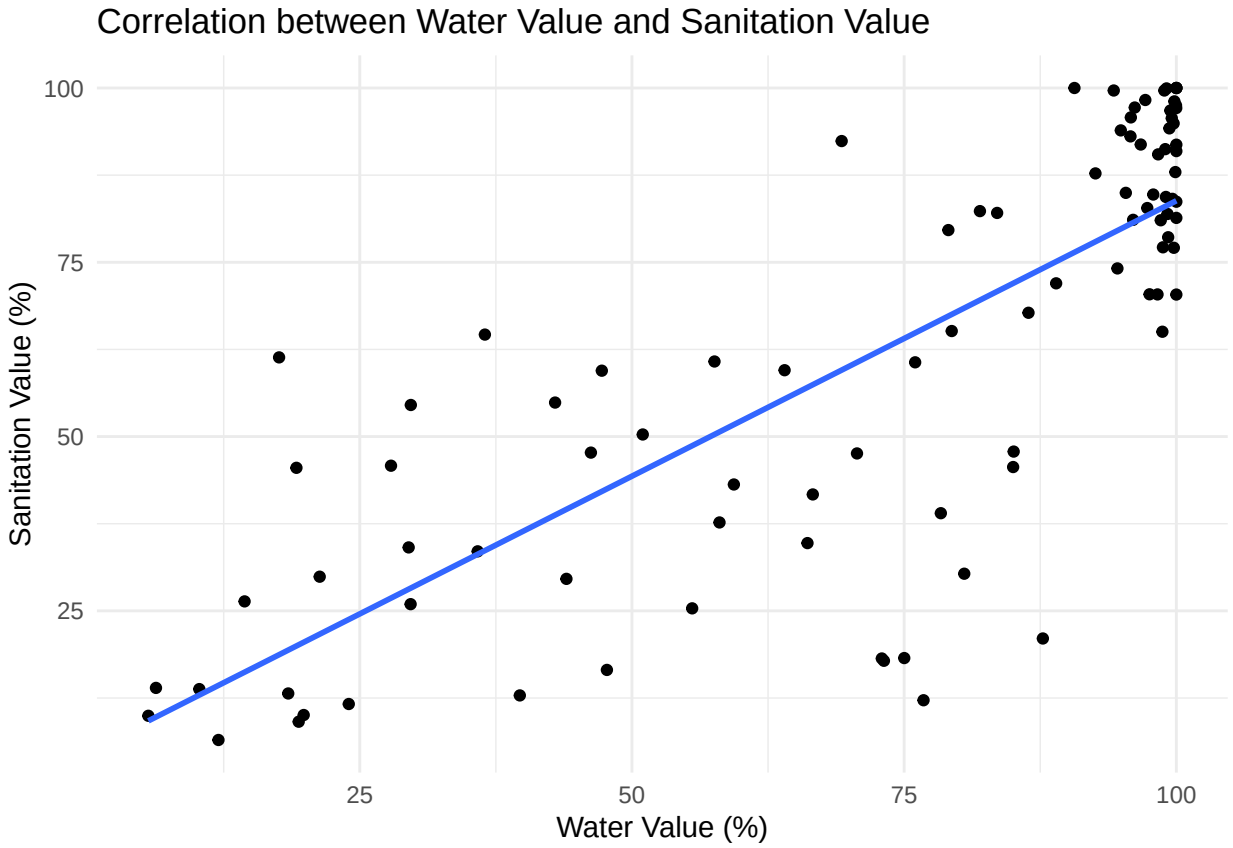
Scatter plot

```
water_sanitation_cor
```

```
## [1] 0.7925022
```

```
print(correlation_plot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The *water_value* and *sanitation_value* have a correlation of **0.792**. This positive relationship indicates that when there is an improvement in water quality in a specific area, it is likely that there will also be an improvement in sanitation quality. Furthermore, the positive correlation is consistent with expectations, suggesting that good water quality is often associated with better sanitary conditions.

Relation with burden disease

Let's now move on to the second part of the analysis, which involves studying and comparing a second significant dataset on the *burden of disease*.

Analysis on the wash deaths dataset

How many people died due to unsafe WASH condition in 2019?

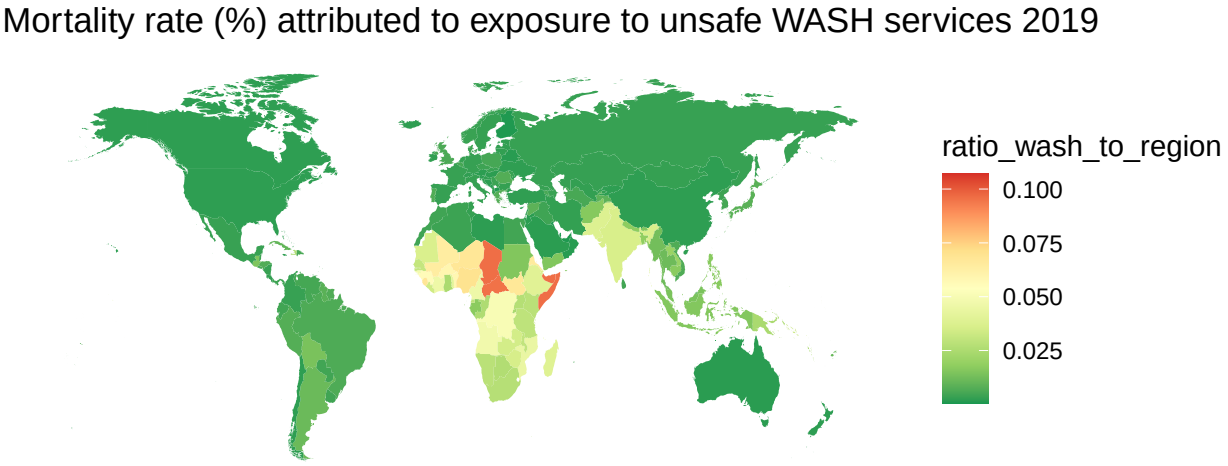
```
sum(both_sex_death$wash_deaths)
```

```
## [1] 1400694
```

1.4 Milion of people died for unsafe wash condition.

World Map

```
print(worldValueDeathWash)
```

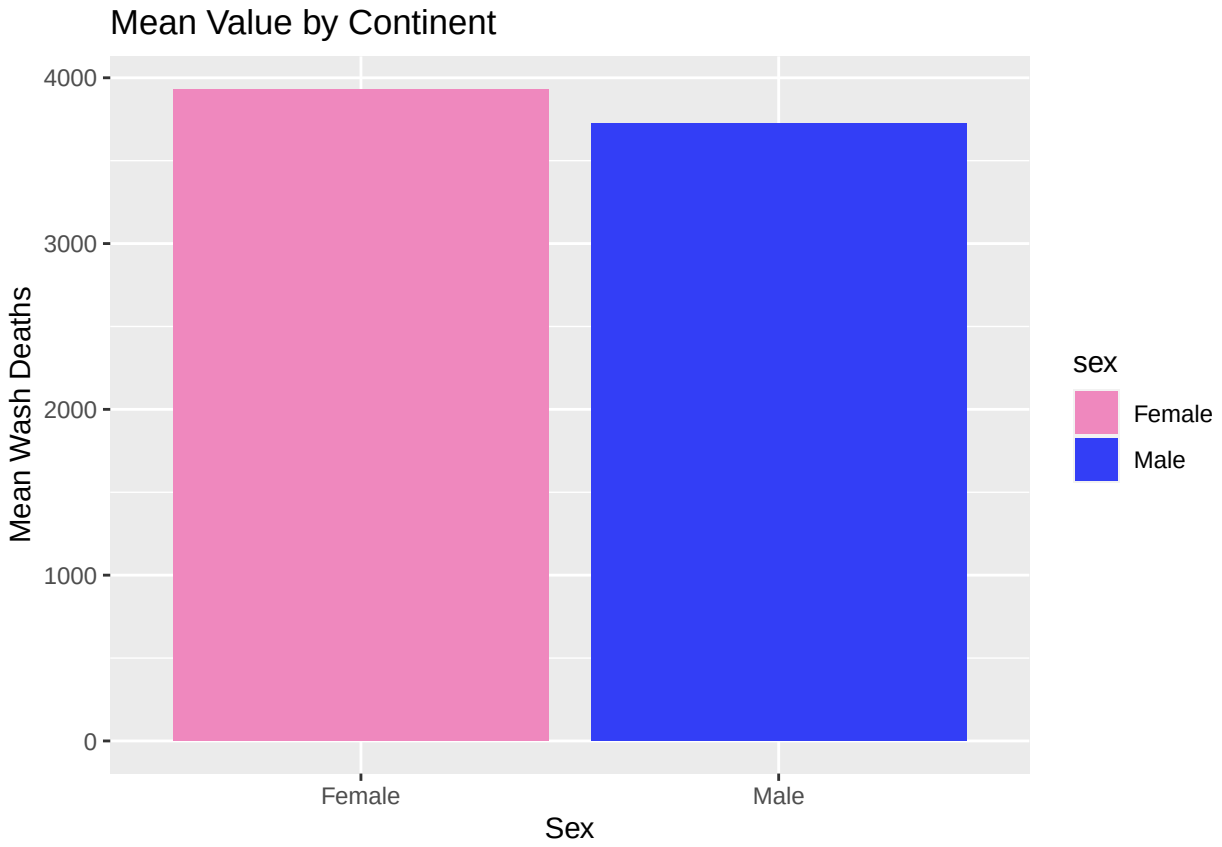


This plot represents, for each state of every continent, the percentage of deaths due to WASH conditions in the total population. this plot help us visualize how the deaths are concentrated around the world, and one continent stands out among the others, which is **Africa**. This observation is consistent with the previous analysis we conducted. We can see that also in **India** we have a relatively high ratio.

Graphs on Men and Women

Let's see the average of the man and women who died across countries

```
print(man_women_death)
```



From the bar plot, we can observe that, on average, more women than men died. Additionally, we can calculate the ratio between male and female deaths.

```r
sum(subset(complete_death, sex == "Female")$wash_deaths) /
  sum(subset(complete_death, sex == "Male")$wash_deaths)
```

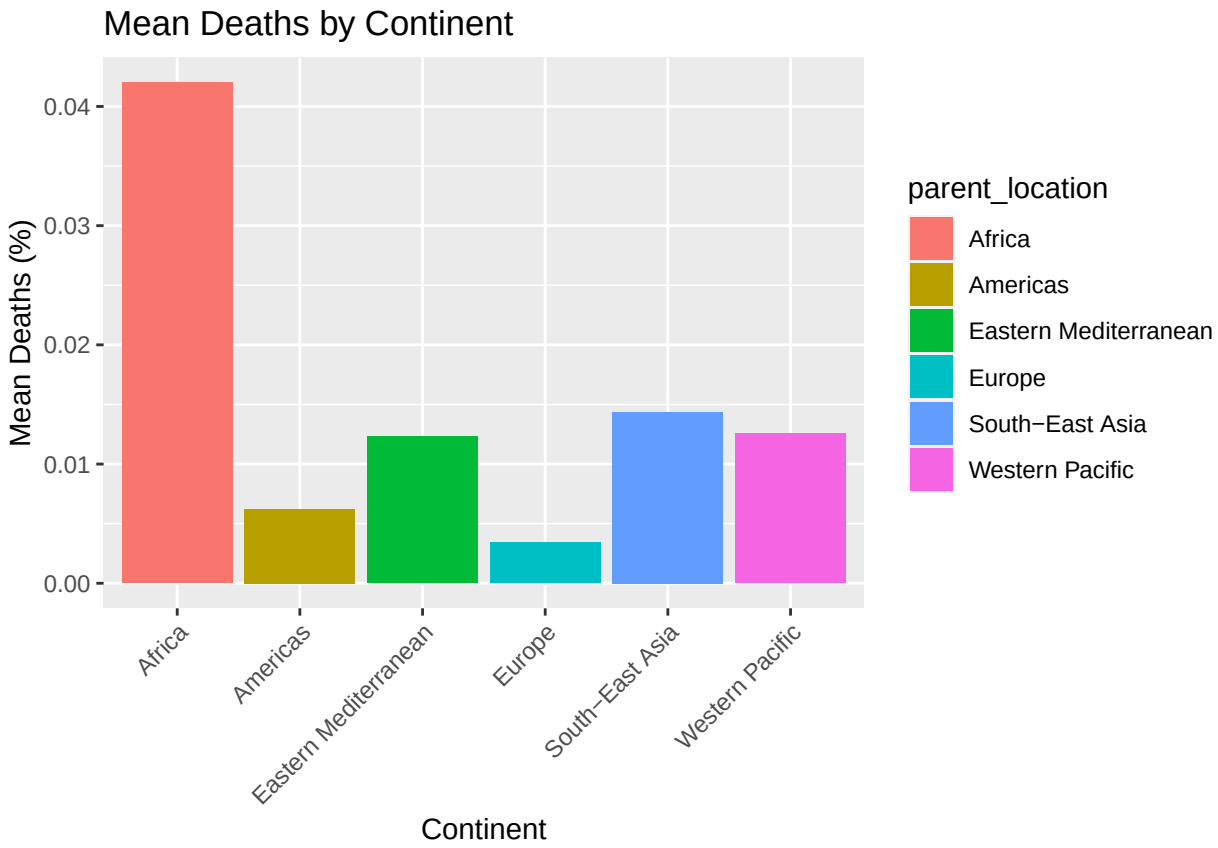
```
## [1] 1.056662
```

We can see that there are slightly more female deaths than male deaths.

Using the normalized data

Now, we will use a bar plot to compare the deaths across continents, with a focus on observing the difference in deaths between Africa and other continents. Note that the total deaths for a continent are divided by the continent's population to obtain normalized data

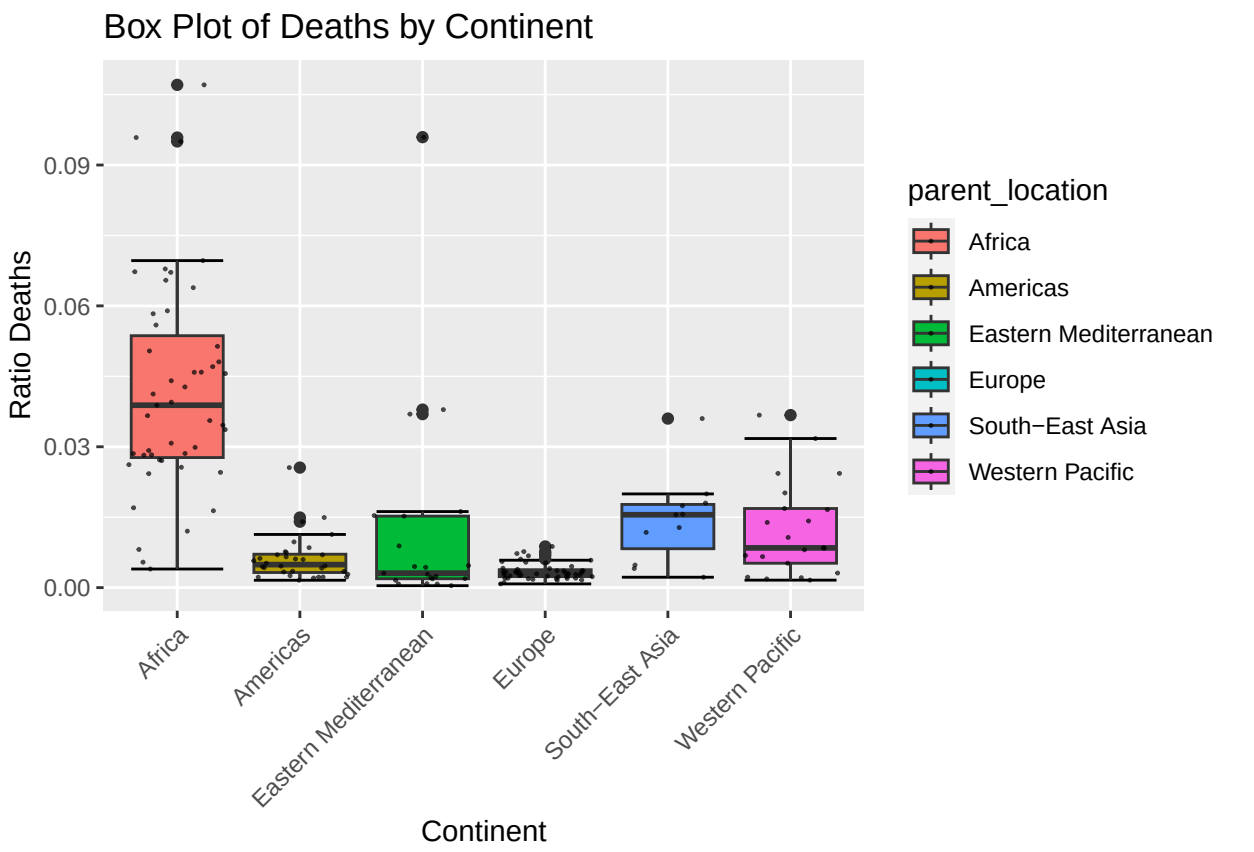
```r
print(mean_deaths_continent_bar)
```



As expected, Africa stands out from the other continents, but Asia, Western Pacific, and the Eastern Mediterranean also have higher values compared to America and Europe.

Box plot

```
print(ratio_wash_death_box)
```



This box plot helps us see how in continents like **America** and **Europe**, the number of deaths is low across all regions, resulting in squeezed-down plots. On the other hand, in other continents, there is more differentiation between regions.

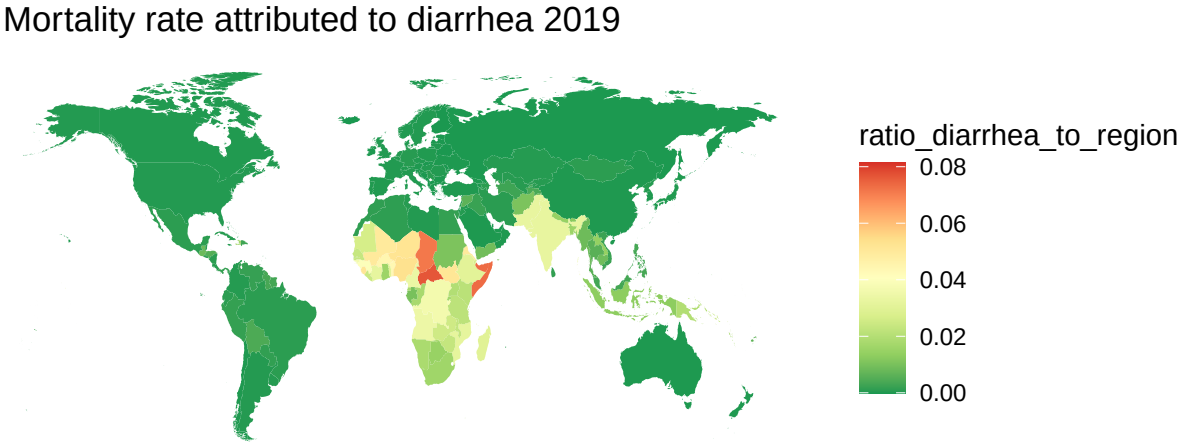
As before, **Africa** has higher values compared to other continents, and it also exhibits some outliers that are significantly higher than the other values. Later on, we will study the outliers in Africa.

Analysis on the diarrhea deaths dataset

Now we will analyze the dataset related to deaths caused by diarrhea.

World Map

```
print(worldValuediarrheaDeath)
```

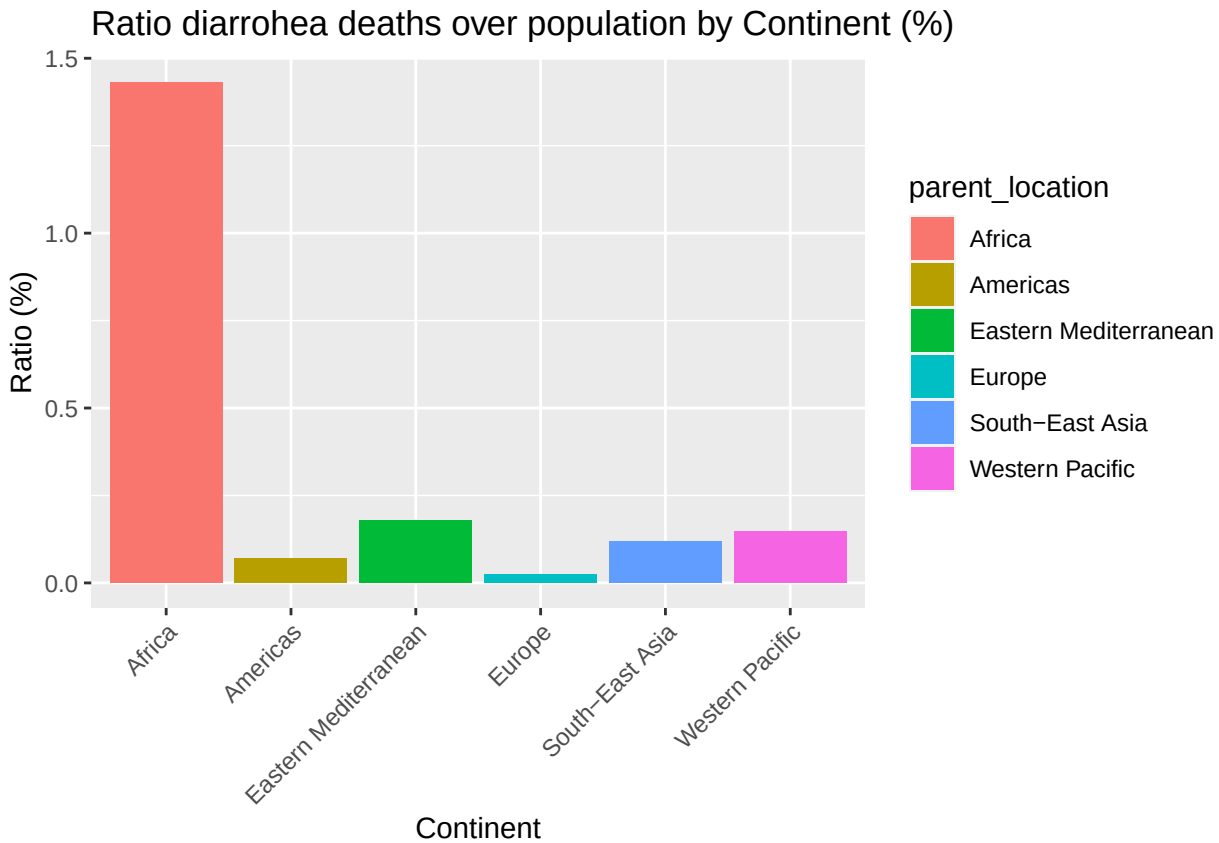


The first thing we notice is that this plot is very similar to the one about WASH deaths. Later on, we will investigate this further.

Similarly to the other plot, we can observe that Africa is the continent with the most deaths caused by diarrhea, followed by Asia, where India has a high number of deaths.

Bar plot

```
print(sum_death_bar)
```



Note: The sum has been calculated based on the ratio between the diarrhea deaths and the population of each continent. Otherwise, America and Asia would have higher numbers due to their larger population sizes. This plot confirms what was seen previously and makes it easier to observe the gap between Africa and the other continents.

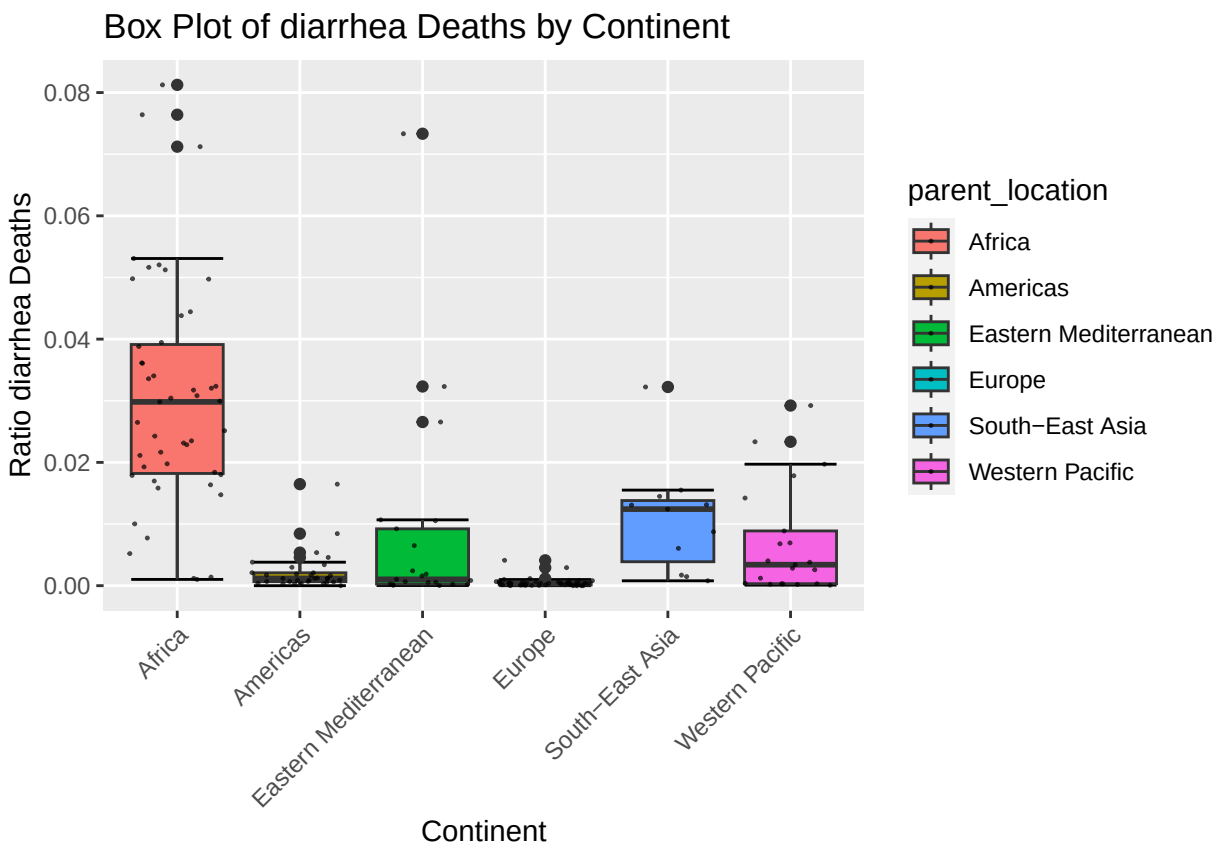
```r
sum(both_sex_death$diarrhea_deaths)
```

```
## [1] 1035171
```

1 milion people died from diarrhea due to unsafe WASH in 2019.

Box plot

```
print(ratio_diarrhea_death_box)
```



Similar to the previous plot, we can observe a lot of similarity in this box plot. We notice some outliers for **Africa** and the **Eastern Mediterranean**, which are very far from the other points. However, in general, the situation remains consistent, with Africa having a higher mean and America and Europe having very low values.

Relation between wash and diarrhea

Since the dataset on diarrhea deaths contains the number of diarrhea deaths from inadequate water, sanitation, and hygiene, let's see how many of the WASH-related deaths are specifically caused by diarrhea.

```
sum(both_sex_death$diarrhea_deaths) / sum(both_sex_death$wash_deaths) * 100
```

```
## [1] 73.90415
```

Approximately **74%** of WASH-related deaths are caused by diarrhea.

Study the country with the highest deaths ratio

Let's now study the country with the highest ratio of deaths because, as seen before in the box plot, there were some of them with higher values compared to the average.

```
subset(both_sex_death, ratio_wash_to_region == max(both_sex_death$ratio_wash_to_region))$region
```

```
## [1] "Lesotho"
```

```
subset(both_sex_death, region == "Lesotho")$ratio_wash_to_region
```

```
## [1] 0.1070766
```

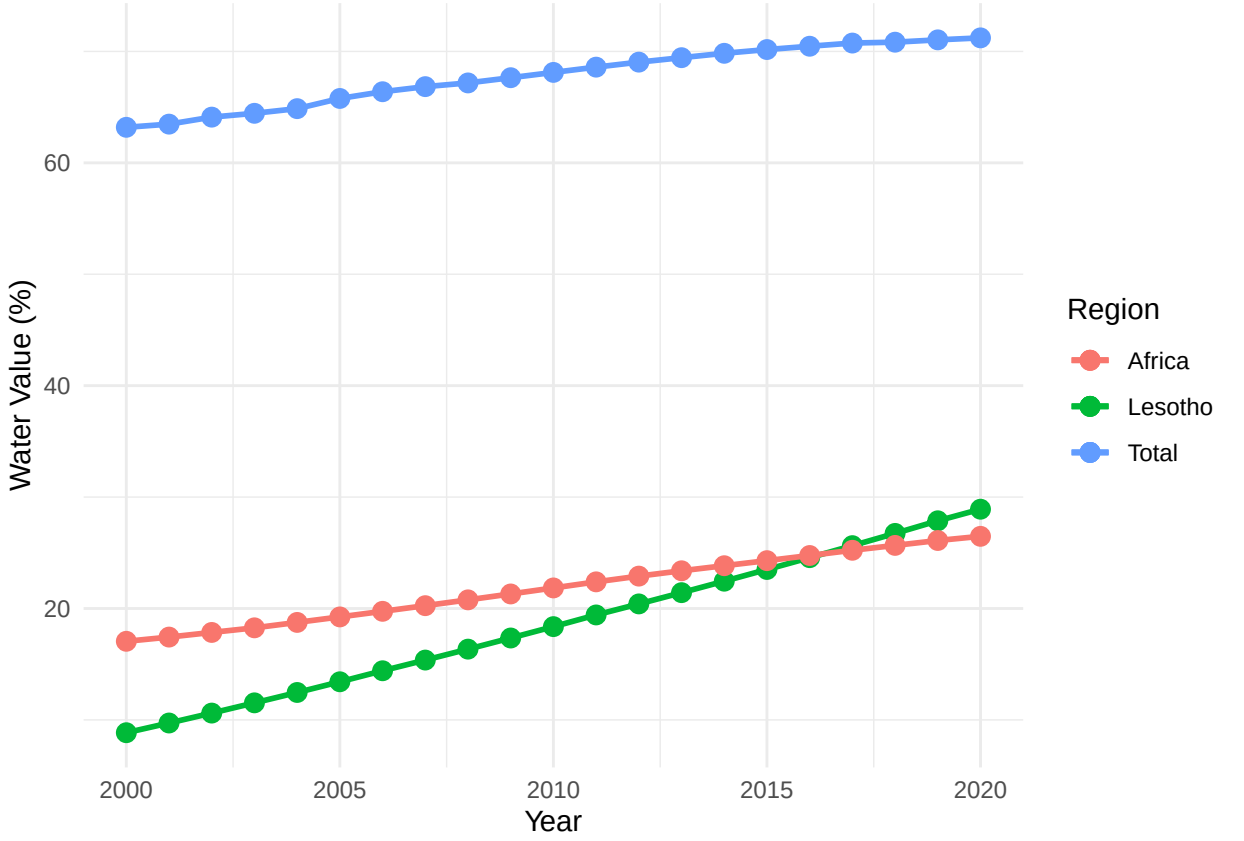
The region with the highest mortality rate due to WASH-related causes is **Lesotho**, with **0.1%** of deaths per population.

Let's create a data frame that contains the data related to Lesotho so we can perform some further analysis.

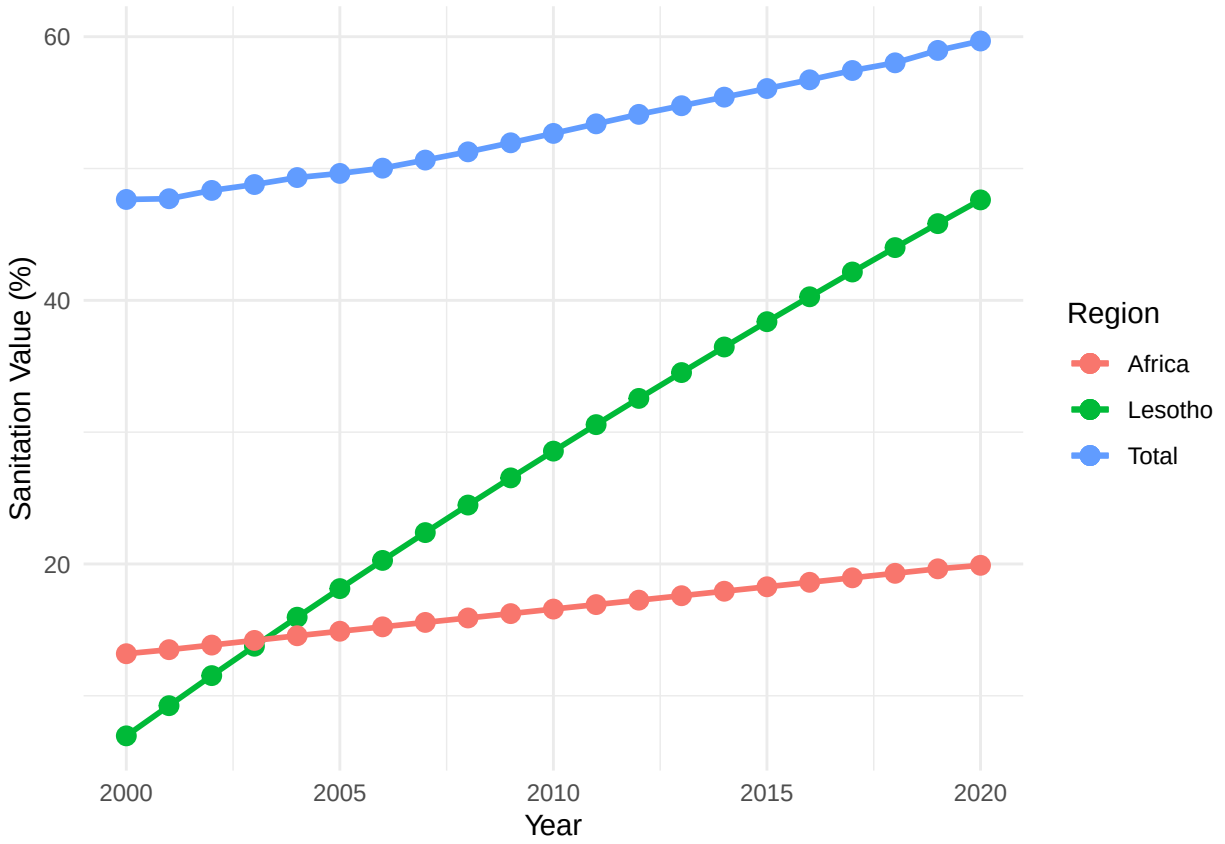
```
lesotho_df <- merge(subset(both_sex_death, region == "Lesotho"),subset(completeWASHDf, region == "Lesotl
                    by = c("parent_location", "region"))
```

Now we can try comparing the trend over time of Lesotho with the average overall trend of countries.

```
print(plot_water_lesotho)
```



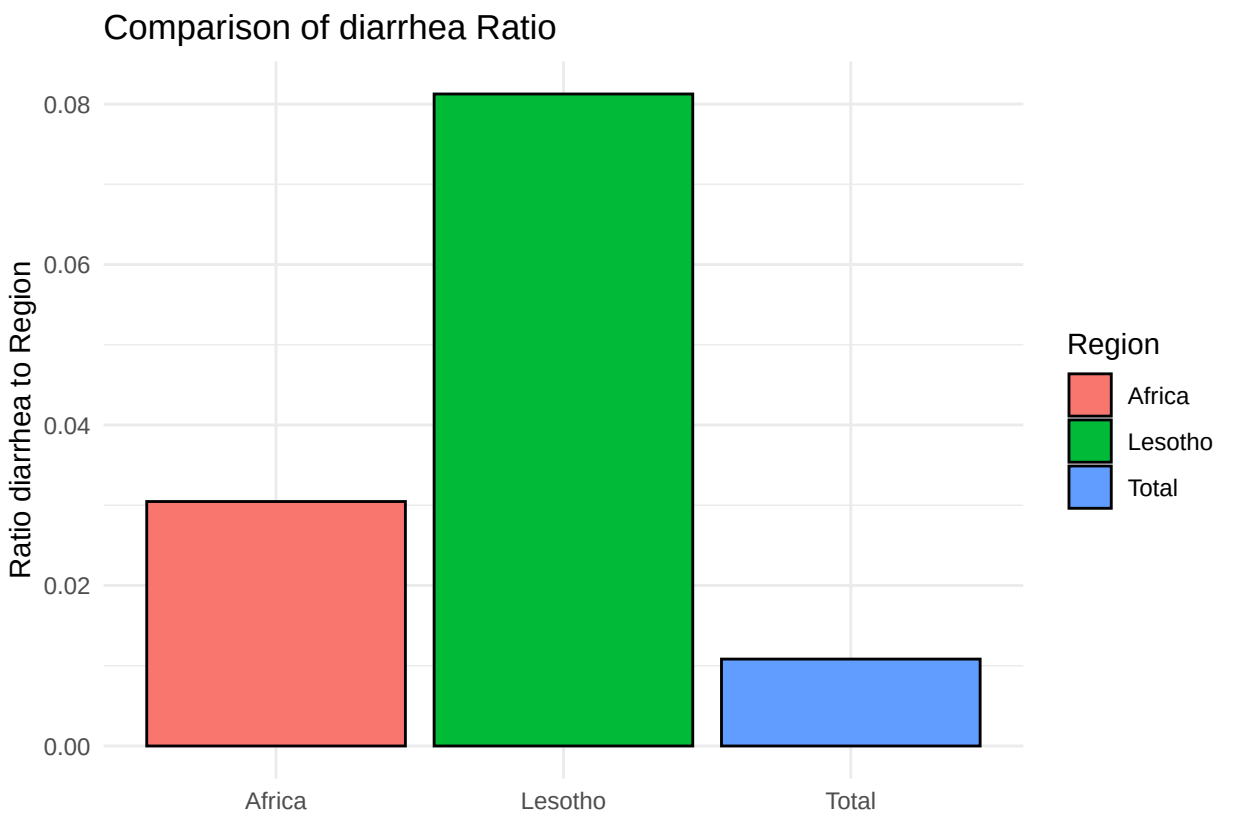

```
print(plot_sanitation_lesotho)
```



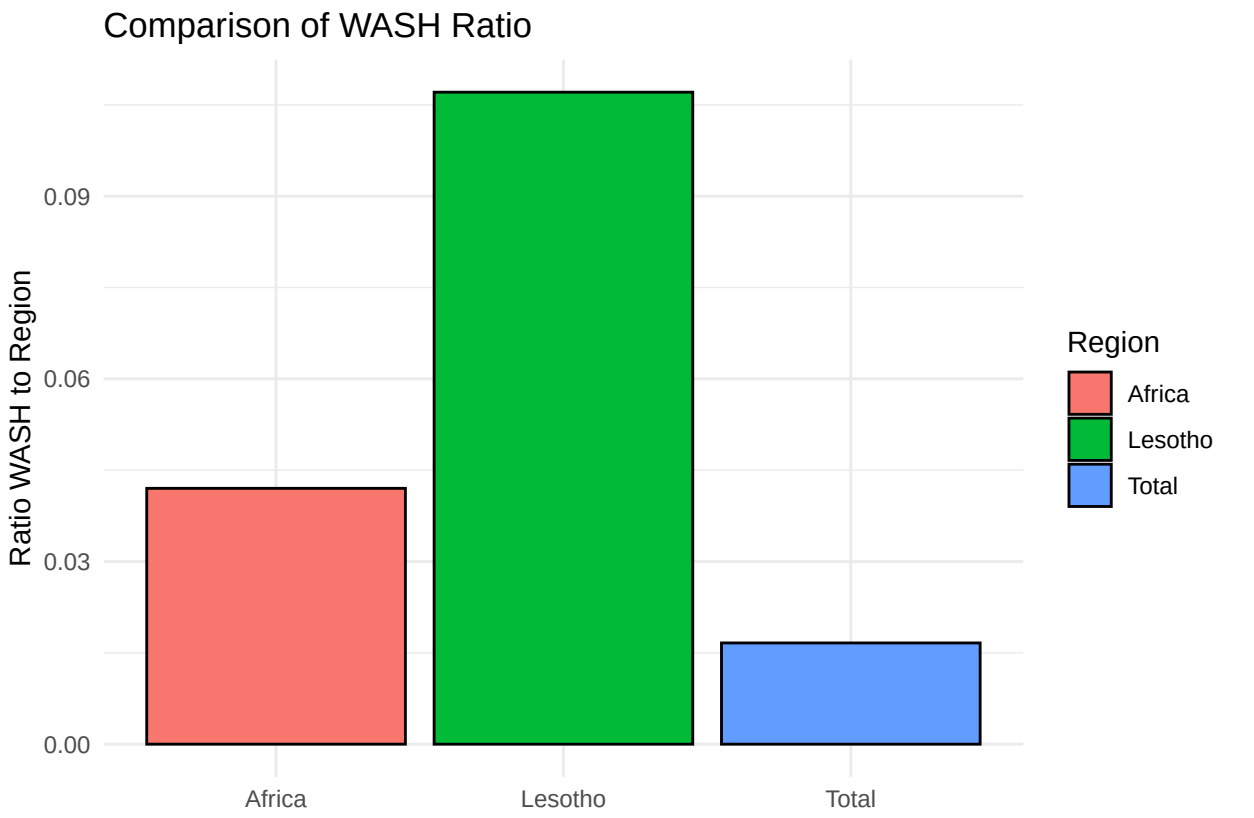
As expected, Lesotho has values well below the average. In terms of water quality, we can see that the "curve" is still far from the overall average. On the other hand, in terms of sanitation, we can observe a significant improvement in recent years that tends to approach the overall average.

Now let's calculate the average number of deaths in Lesotho compared to the world.

```
print(barplot_diarr_lesotho)
```



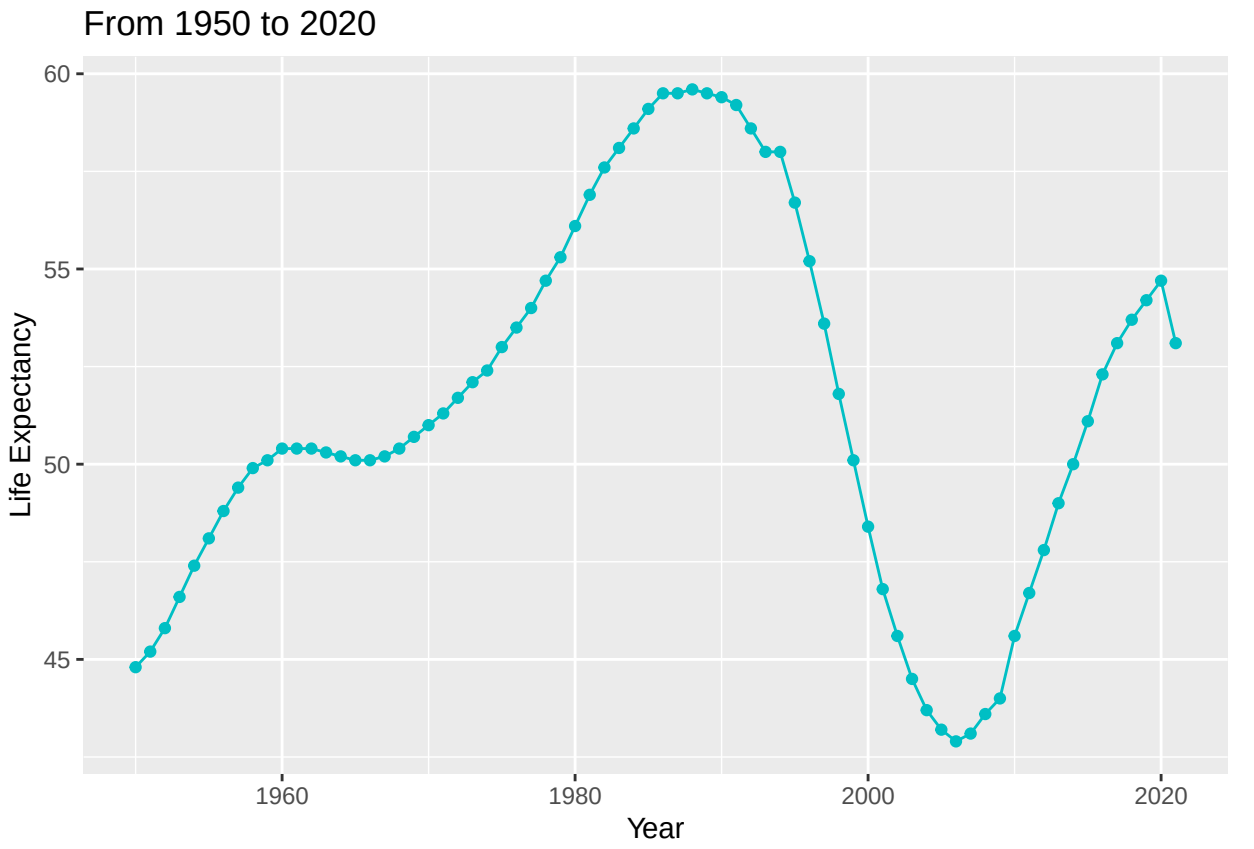

```
print(barplot_wash_lesotho)
```



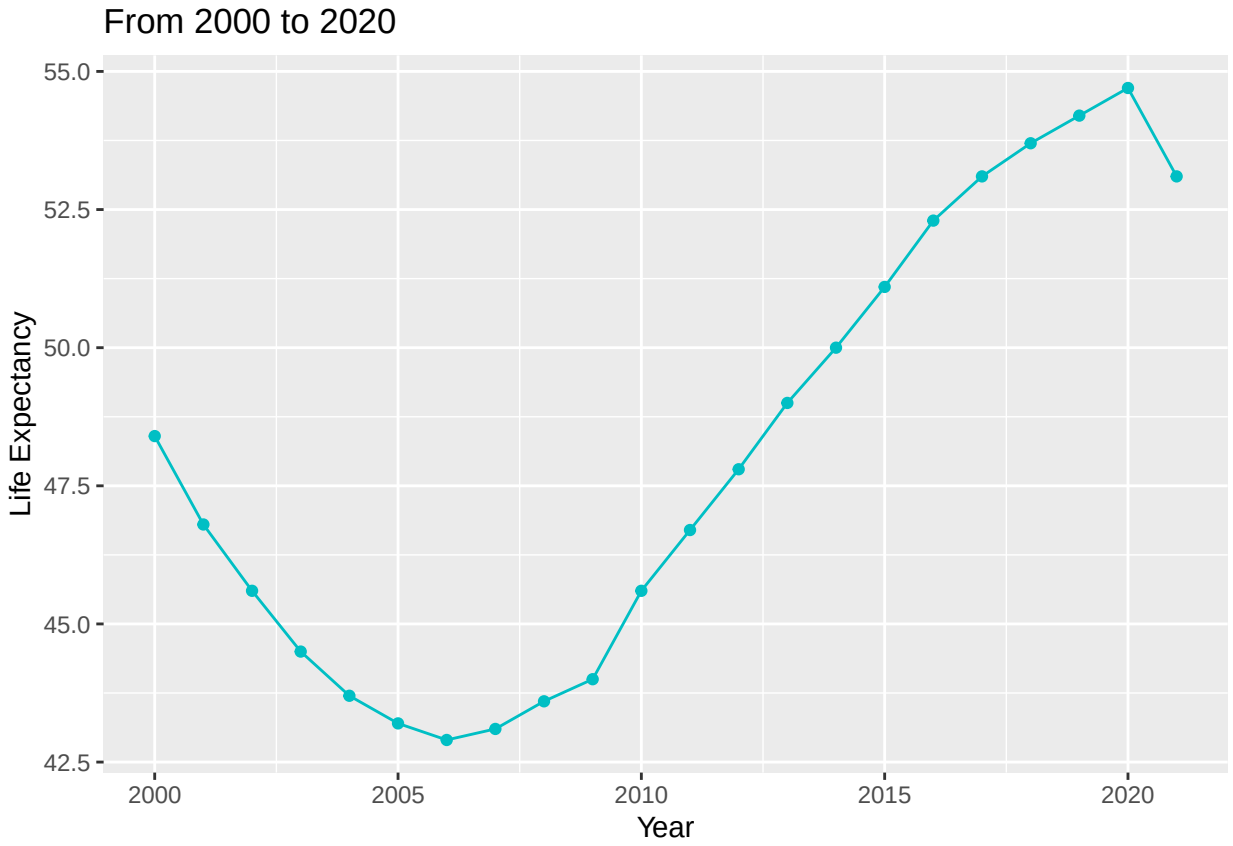
We can see, as expected, that Lesotho is above the average.

Now let's look at the life expectancy in Lesotho.

```
print(complete_life_exp_lesotho)
```




```
print(partial_life_exp_lesotho)
```



From the **1950-2020** plot, we can observe a drastic reduction in life expectancy after the 1990s. Since we don't have data on water quality in that years, we cannot establish a correlation between the two factors. However, we can notice that in the **2000-2020** graph, both water quality and healthcare (after the 2000s) have improved, as well as life expectancy. Therefore, between the 1990s and 2000s, we can imagine that these two variables decreased (or remained constant). However, we can definitely say that something happened around those years that caused a significant decline in life expectancy.

Study the country with the highest increase over the years

The region that experienced the highest improvement in *water quality* between the period 2000 and 2020 is: **Moldova**.

The water value in **2000s** was: **40.4%** , and in **2020**, it increased to : **74.07%**

The region that experienced the highest improvement in *sanitation quality* between the period 2000 and 2020 is: **Andorra**.

The sanitation value in **2000s** was: **14.6%** , and in **2020**, it increased to : **100%**

Correlation between datasets

Subset

```
merged_df <- merge(life_exp_2019, both_sex_death, by = c("region"))
merged_df <- merge(merged_df, totalDf2019, by = c("region", "period", "parent_location"))
```

Analysis on correlation

Let's calculate the correlation with the **ratio** for each continent:

```
## [1] "Eastern Mediterranean"
## [1] -0.7968705
## [1] "Europe"
## [1] -0.1296597
## [1] "Africa"
## [1] -0.8150216
## [1] "Americas"
## [1] -0.6089404
## [1] "Western Pacific"
## [1] -0.6405645
## [1] "South-East Asia"
## [1] -0.4986985
```

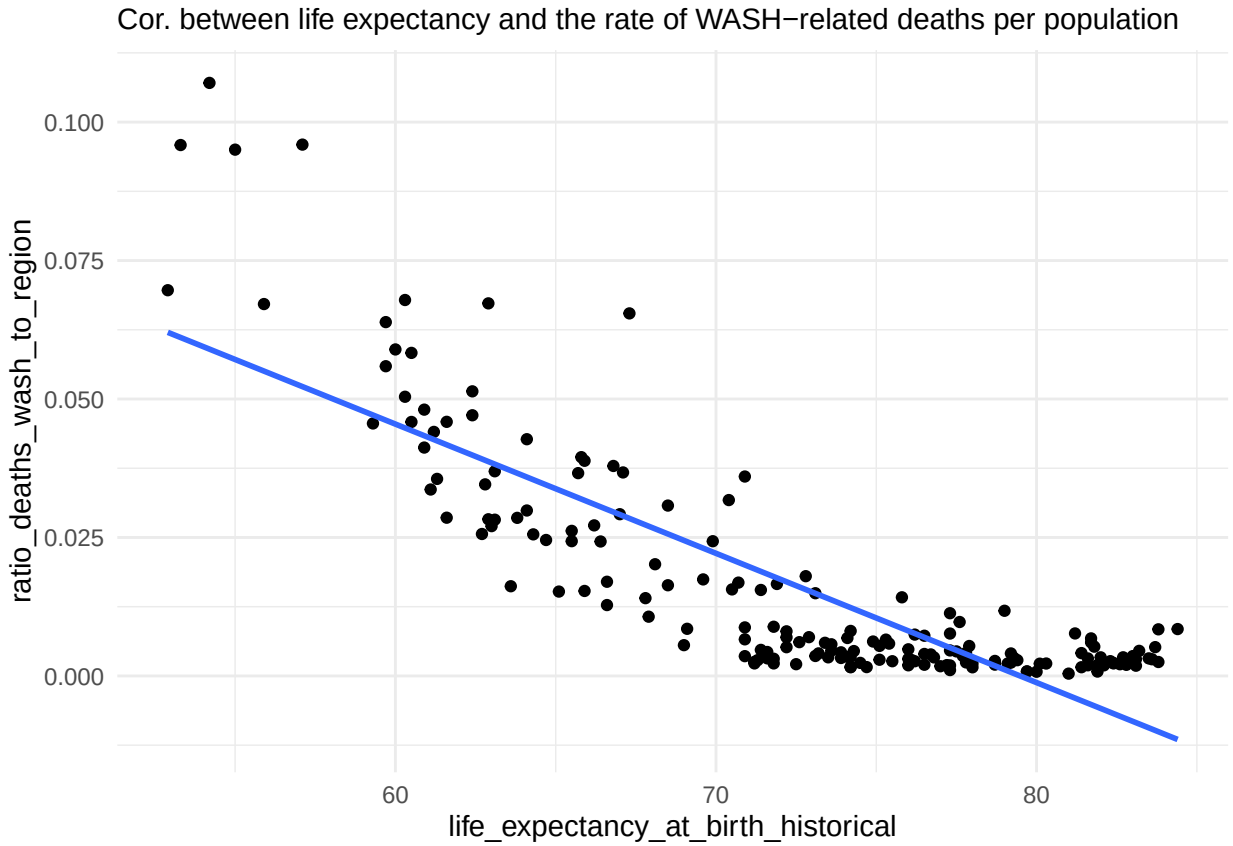
Based on the correlation values obtained between life expectancy and the mortality rate from WASH-related causes relative to the population for each continent, we can observe *significantly high values* for regions like the **Eastern Mediterranean**, **Africa**, and **Western Pacific**. These high values indicate a negative correlation between life expectancy and the mortality rate from WASH-related causes relative to the population. This suggests that as life expectancy increases, the mortality rate tends to decrease, and vice versa.

However, for other continents, we observe more moderate values, indicating a lower correlation. This implies that life expectancy is influenced by additional factors that are not currently being considered.

Now we can create a plot based on the correlations.

```
print(scatter_plot_wash)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
print(correlation_wash)
```

```
## [1] -0.8335848
```

The obtained correlation value is **-0.83**, demonstrating a negative correlation between life expectancy and the mortality rate from WASH-related causes relative to the population. This means that as life expectancy increases, the mortality rate from WASH-related causes relative to the population tends to decrease, and vice versa.

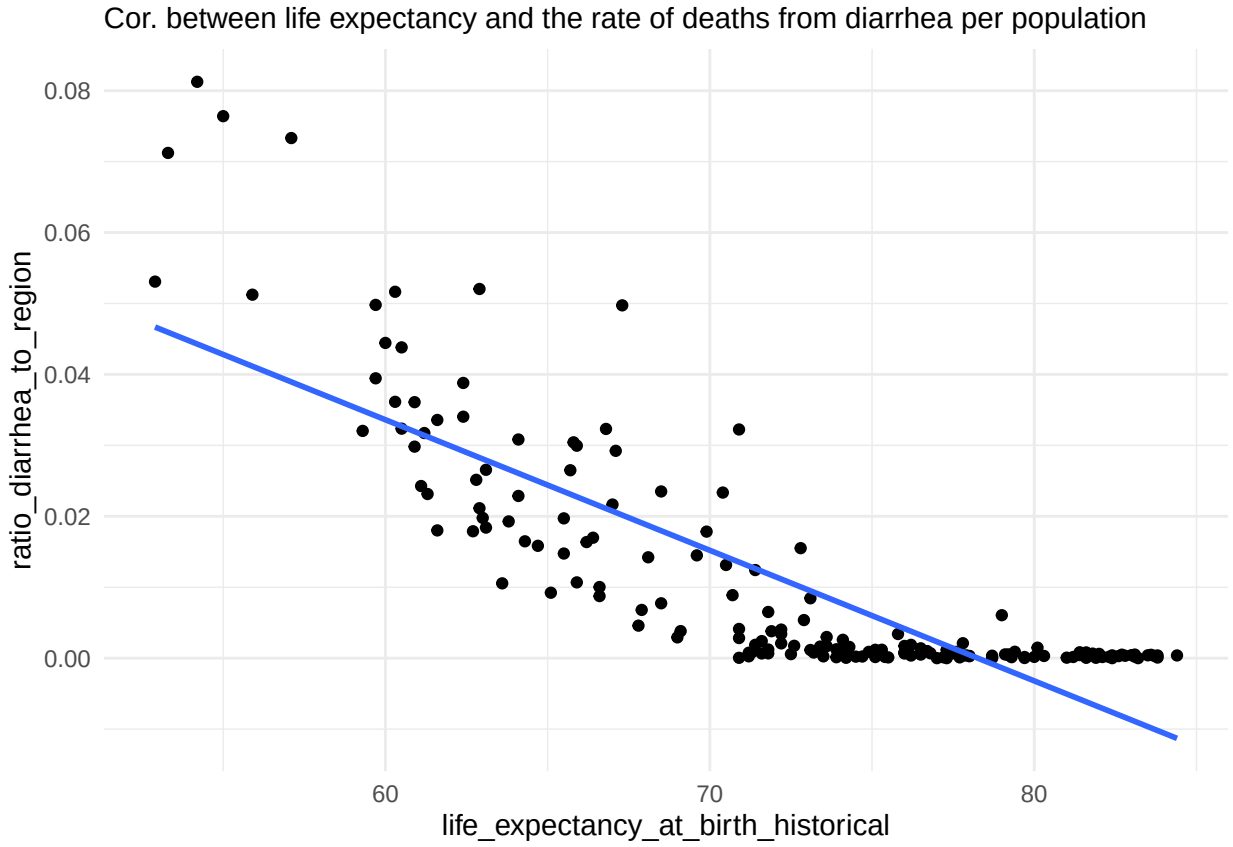
From the graph, we can observe the data points tending to follow the linear regression line, confirming the negative correlation between the two variables.

Furthermore, it is notable that the life expectancy range between **70 and 90 years** include the majority of the data points, which are characterized by an approximately **0% mortality rate** from WASH-related causes. This graph highlights the significance of ensuring access to clean drinking water, sanitation services, and proper hygiene practices to enhance health conditions and increase life expectancy among individuals.

Correlation with the **rario_diarrhea**

```
print(scatter_plot_diarr)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
print(correlation_diarr)
```

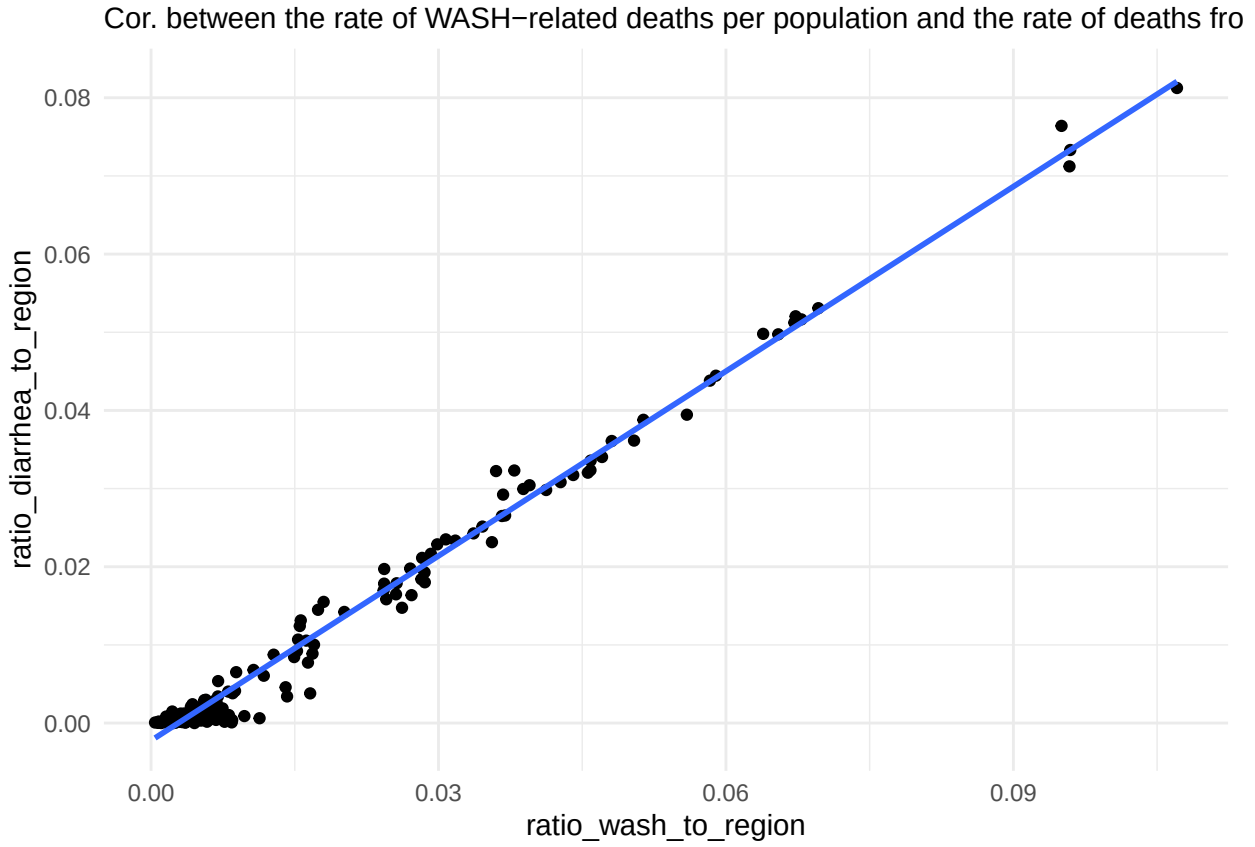
```
## [1] -0.8302084
```

We can see an equal value for the correlation **-0.83**, and we can also observe that this plot exhibits a strong similarity to the previous one, illustrating the same pattern where the range of 70-90 years encompasses the majority of the data points.

Now we make a plot between the ratio of wash and diarrhea

```
print(scatter_plot_diarr_wash)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
print(correlation_diarr_wash)
```

```
## [1] 0.994381
```

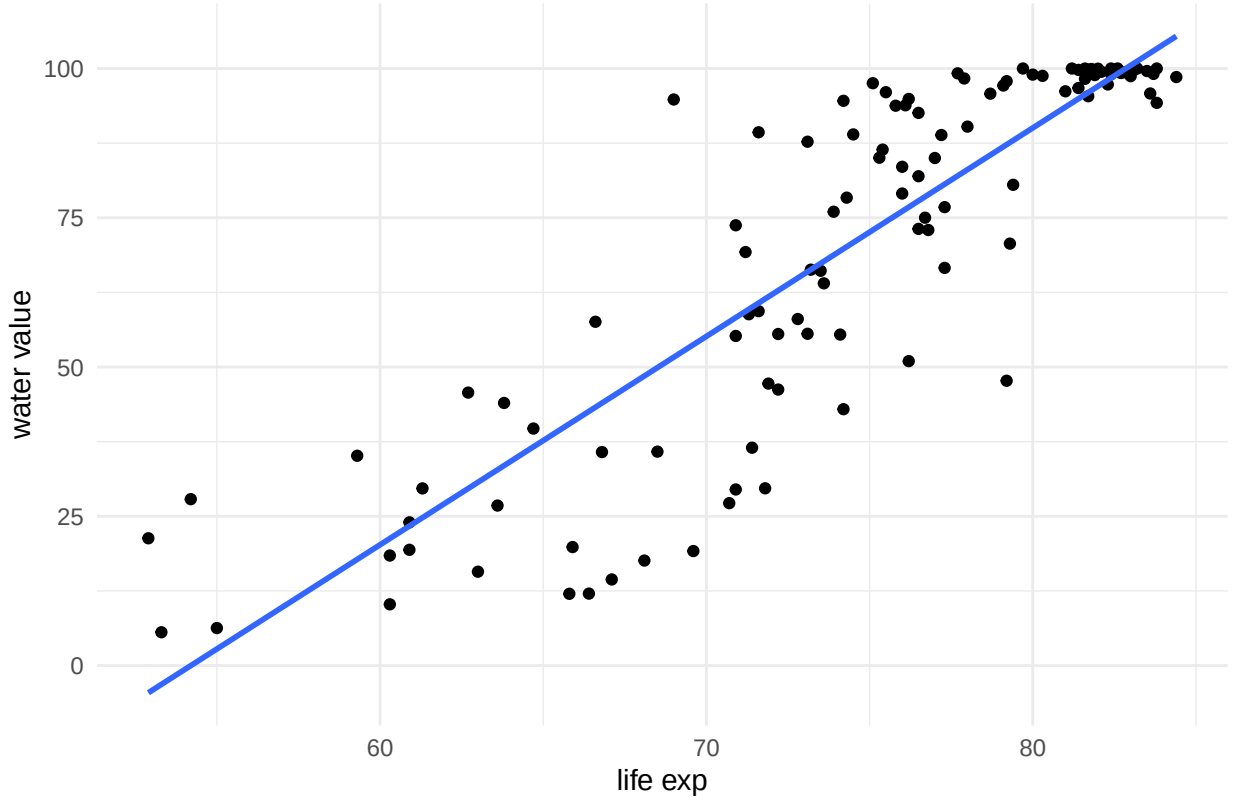
As expected, they are highly correlated (**cor = 0.9943196**), as seen before, with diarrhea accounting for 74% of the WASH-related deaths.

Lets see now the correlation with the water and sanitation data frames.

```
print(scatter_plot_water)
```

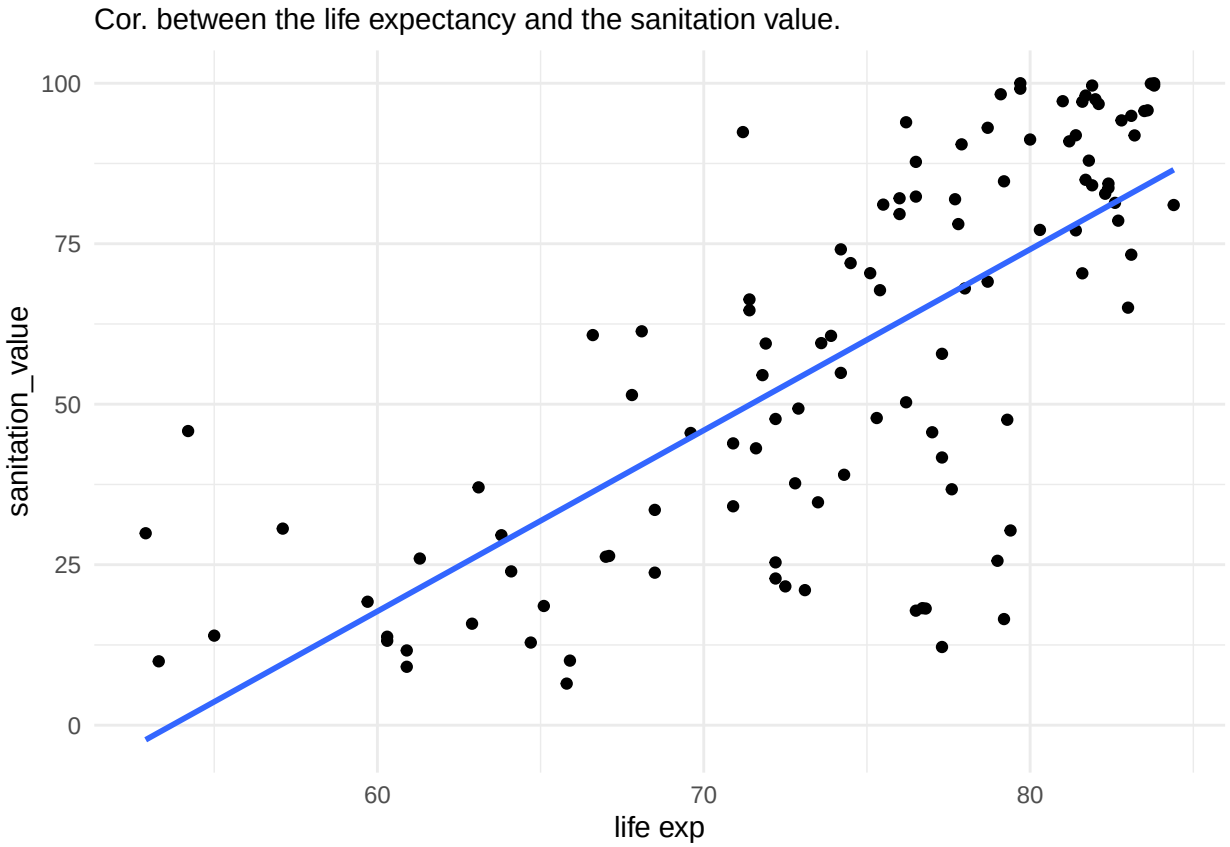
```
## `geom_smooth()` using formula = 'y ~ x'
```

Cor. between the life expectancy and the water value.




```
print(scatter_plot_sanitation)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Water correlation value: **0.868**

Sanitation correlation value: **0.731**

In general, we can observe that the quality of water and sanitation are two factors that significantly impact life expectancy. It is evident that as both the quality of water and sanitation improve, life expectancy tends to increase, and conversely, as these factors decrease, life expectancy decreases.

Conclusion

From the analysis of the datasets related to Water, Sanitation, and Hygiene (WASH), we observed significant disparities among continents, with Africa emerging as a critical area in terms of access to adequate water and sanitation services. The very low values for water and sanitation in Africa are concerning and demand more attention to improve the hygiene and sanitary conditions of the population.

We noticed that the improvement in water and sanitation quality correlates positively with life expectancy. This connection suggests that improving WASH services could have a significant impact on people's health and life expectancy. Specifically, countries with better WASH services exhibit higher life expectancy and a reduction in WASH-related deaths.

However, we observed some 'positive' outliers, such as the Eastern Mediterranean and certain African regions, which have values well above the average. These results may indicate that some areas are making significant progress in improving their WASH conditions, despite the overall situation of the continent.

Diarrhea is a major cause of death in places with poor WASH services. This shows that we really need to focus on fixing these problems, especially in Africa.

In conclusion, our study highlights the need for global action to improve access to clean water, sanitation, and hygiene, with particular attention to continents like Africa. Investing in adequate WASH services can

have a significant impact on people's health and quality of life, leading to a reduction in diseases and deaths related to to the previously mentioned problems.