Infographic Description

Ferdinand David Anggono (2702299661) - Yosia (2702300240) - Rainer Alexander Irawan (2702261196) - Alvin Febrian (2702370814)

2025-06-04

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(readxl)
library(readr)
library(ggrepel)
library(ggalt)
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```

```
library(countrycode)
library(rnaturalearth)
library(sf)
```

```
## Linking to GEOS 3.13.0, GDAL 3.10.1, PROJ 9.5.1; sf_use_s2() is TRUE
```

```
library(RColorBrewer)
library(plotly)
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

#Introduction

This report presents an analysis of data related to drinking water quality and sanitation. The data used is taken from two CSV files: - **drinking_water_dataset.csv** - **sanitation_dataset.csv**

2. Objectives

The objectives of this analysis are: - Understand the distribution of drinking water and sanitation quality - Identify important variables - Conduct reliability tests and interpret results

3. Data

3.1 Reading Data

```
drinking_data <- read.csv("drinking_water_dataset.csv")
sanitation_data <- read.csv("sanitation_dataset.csv")

# Tampilkan ringkasan data
summary(drinking_data)
```

```
##      ISO3             Country           Residence.Type     Service.Type
##  Length:16540       Length:16540       Length:16540       Length:16540
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       Year         Coverage           Population        Service.level
##  Min.   :2007   Min.   :  0.00000   Min.   :0.000e+00   Length:16540
##  1st Qu.:2010   1st Qu.:  0.07363   1st Qu.:3.270e+02   Class :character
##  Median :2014   Median :  3.45741   Median :1.330e+05   Mode  :character
##  Mean   :2014   Mean   : 21.54776   Mean   :7.088e+06
##  3rd Qu.:2018   3rd Qu.: 28.20768   3rd Qu.:2.035e+06
##  Max.   :2022   Max.   :100.00000   Max.   :1.416e+09
```

```
summary(sanitation_data)
```

```
##      ISO3             Country           Residence.Type     Service.Type
##  Length:16467       Length:16467       Length:16467       Length:16467
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       Year         Coverage          Population        Service.level
##  Min.   :2007   Min.   :  0.0000   Min.   :        0   Length:16467
##  1st Qu.:2010   1st Qu.:  0.5969   1st Qu.:     4177   Class :character
##  Median :2014   Median :  8.0440   Median :   267900   Mode  :character
##  Mean   :2014   Mean   : 21.7040   Mean   :  7129919
##  3rd Qu.:2018   3rd Qu.: 30.8592   3rd Qu.:  2700525
##  Max.   :2022   Max.   :100.0000   Max.   :974545978
```

Explanation: This section reads the available datasets. The dataset `drinking_water_dataset.csv` contains drinking water quality data, while `sanitation_dataset.csv` contains sanitation data. The `summary()` function is used to display the initial statistical summary.

4. Data Analysis

4.1 Question 1: Available Variables

Question: What are the available variables in both datasets?

```
names(drinking_data)
```

```
## [1] "ISO3"           "Country"        "Residence.Type" "Service.Type"
## [5] "Year"           "Coverage"       "Population"     "Service.level"
```

```
names(sanitation_data)
```

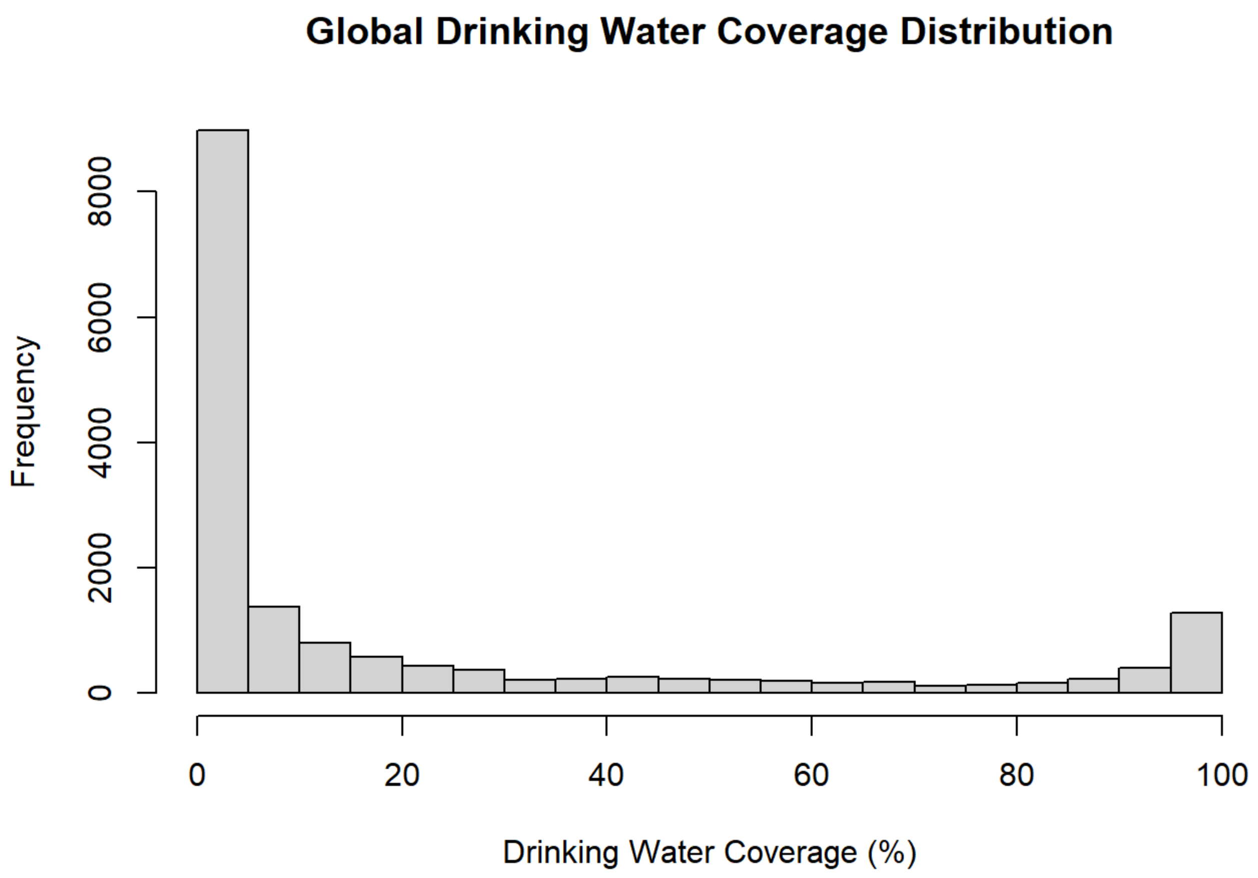
```
## [1] "ISO3"          "Country"       "Residence.Type" "Service.Type"
## [5] "Year"          "Coverage"      "Population"     "Service.level"
```

Explanation: This section displays the names of the columns or variables available in the drinking water and sanitation dataset. This is important as a basis for further analysis.

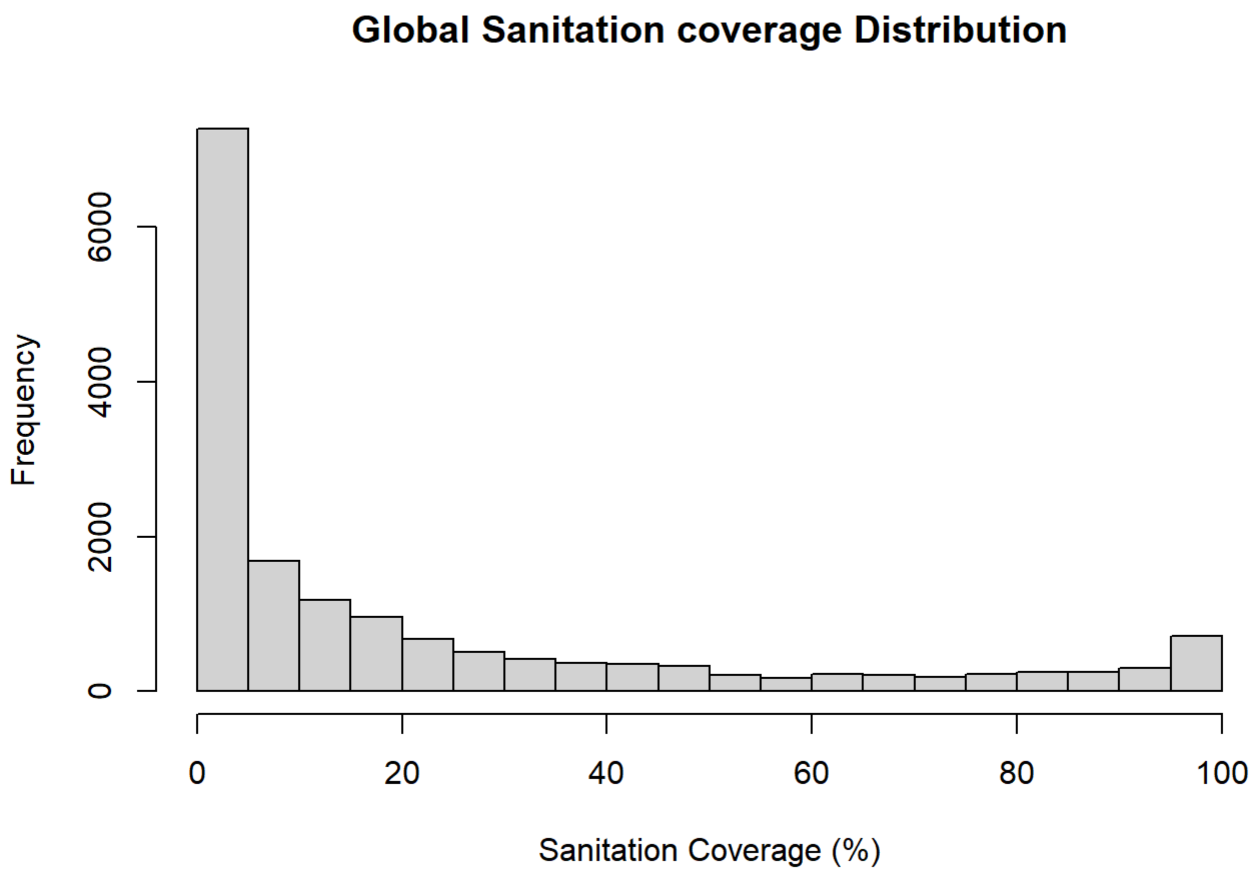
4.2 Question 2: Data Distribution

Question: How is the data distributed on each of the important variables?

```
hist(drinking_data$Coverage, main = "Global Drinking Water Coverage Distribution", xlab = "Dr
inking Water Coverage (%)")
```




```
hist(sanitation_data$Coverage, main = "Global Sanitation coverage Distribution", xlab = "Sani
tation Coverage (%)")
```



Explanation: The histogram visualization is used to see the distribution of values in the variables `Coverage` . The distribution helps us understand whether the data is skewed to the left, right, or normal. Both distribution are right-skewed, indicating a large number of regions still have very low access to sanitation and drinking water facility. Meanwhile, only a few region achieved close to 100% coverage.

5. Data Visualization

5.1 Plot 1

Question: What has been the trend in access to safely managed drinking water services in the world over the past fifteen years?

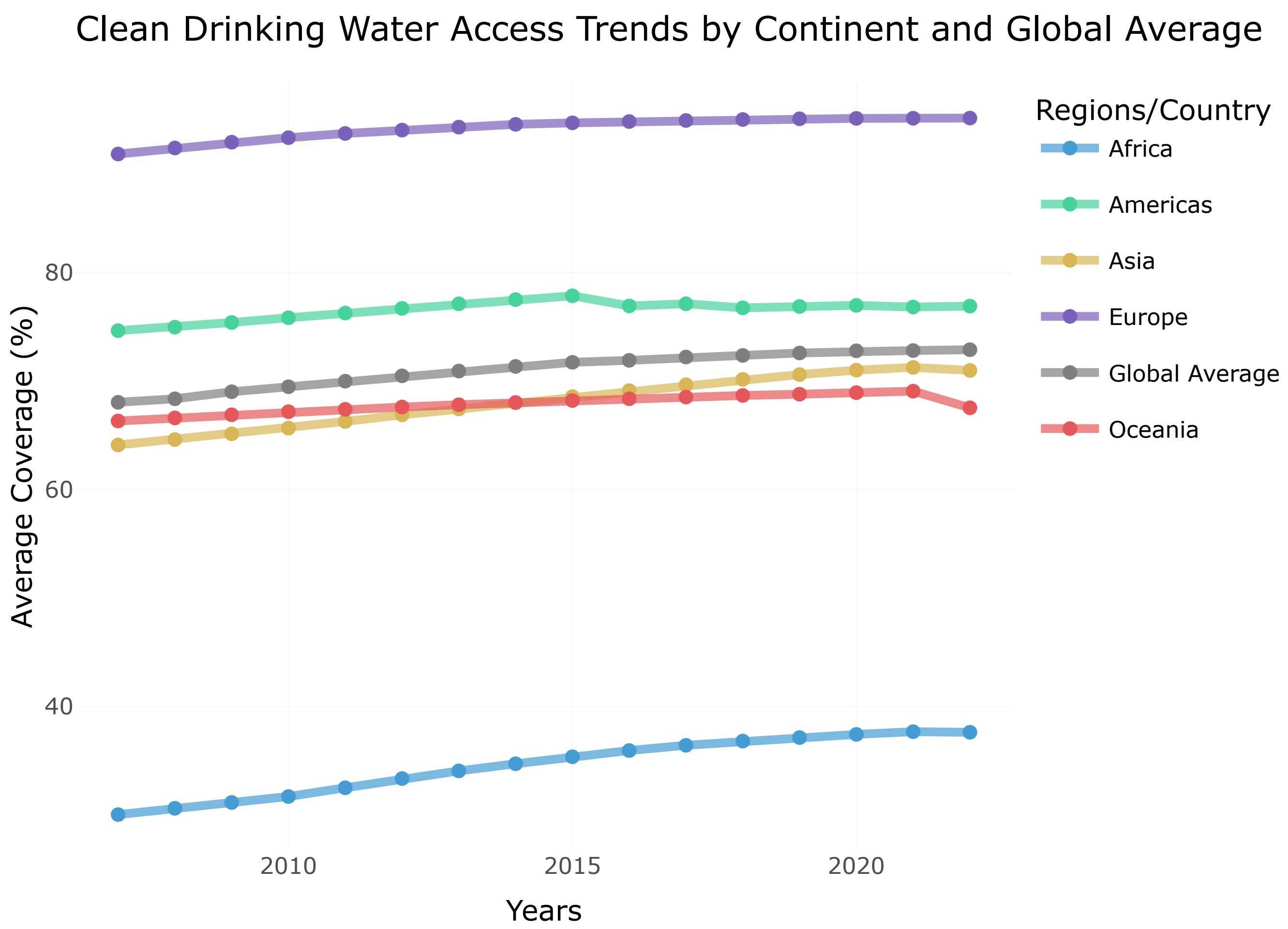
```
# Merge country based on continent
drinking_data <- drinking_data %>%
  mutate(Continent = countrycode(ISO3, origin = "iso3c", destination = "continent"))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `Continent = countrycode(ISO3, origin = "iso3c", destination =
##   "continent")`.
## Caused by warning:
## ! Some values were not matched unambiguously: CHI
```

```
# Filter only safe drinking water service
drinking_data_filtered <- drinking_data %>%
  filter(`Service.level` == "Safely managed service")

# Remove N/A value after
drinking_data_filtered <- drinking_data_filtered %>%
  filter(!is.na(Continent))

# Trend per continent
trend_by_continent <- drinking_data_filtered %>%
  group_by(Continent, Year) %>%
  summarise(Average_Coverage = mean(Coverage, na.rm = TRUE), .groups = "drop")

# Global Trend (average per country)
trend_global <- drinking_data_filtered  %>%
  group_by(Year) %>%
  summarise(Average_Coverage = mean(Coverage, na.rm = TRUE), .groups = "drop") %>%
  mutate(Continent = "Global Average")  # pakai label baru

# Merge those two trends
trend_combined <- bind_rows(trend_by_continent, trend_global)

# Plot
p <- ggplot(trend_combined, aes(
  x = Year,
  y = Average_Coverage,
  color = Continent,
  group = Continent,
  text = paste("Tahun:", Year,
               "<br>Wilayah:", Continent,
               "<br>Cakupan:", round(Average_Coverage, 1), "%"))
) +
  geom_line(size = 1.2, alpha = 0.7) +
  geom_point() +
  labs(
    title = "Clean Drinking Water Access Trends by Continent and Global Average",
    x = "Years",
    y = "Average Coverage (%)",
    color = "Regions/Country"
  ) +
  scale_color_manual(
    values = c(
      "Africa" = "#449cd3",
      "Asia" = "#d8b655",
      "Europe" = "#7a61ba",
      "Americas" = "#46d39a",
      "Oceania" = "#e55759",
      "Global" = "#121212"
    )
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
ggplotly(p, tooltip = "text")
```



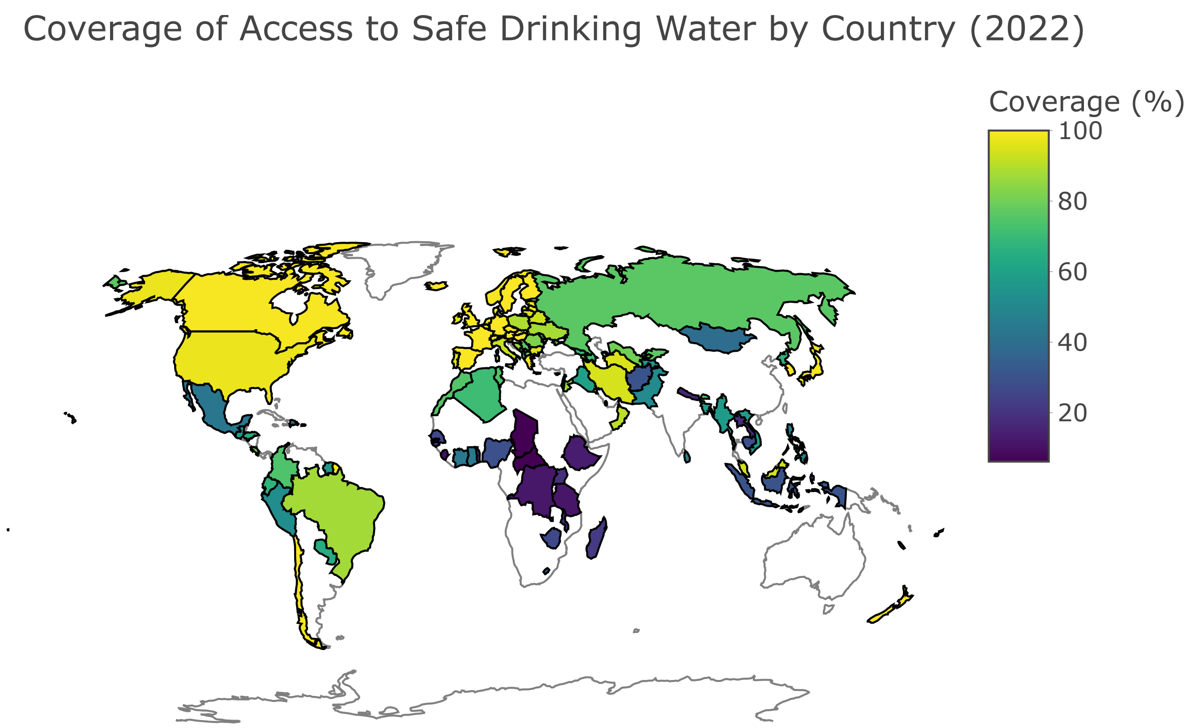

```
# Shows the average clean water coverage for each country (Table)
table_output <- trend_combined %>%
  pivot_wider(names_from = Continent, values_from = Average_Coverage) %>%
  arrange(Year)

print(table_output)
```

```
## # A tibble: 16 × 7
##     Year Africa Americas  Asia Europe Oceania `Global Average`
##    <int>  <dbl>    <dbl> <dbl>  <dbl>   <dbl>            <dbl>
##  1  2007   30.0     74.6  64.1   90.9    66.3             68.0
##  2  2008   30.6     75.0  64.6   91.5    66.6             68.4
##  3  2009   31.2     75.4  65.1   92.0    66.9             69.0
##  4  2010   31.7     75.8  65.7   92.4    67.2             69.5
##  5  2011   32.5     76.3  66.3   92.8    67.4             70.0
##  6  2012   33.4     76.7  66.9   93.1    67.6             70.5
##  7  2013   34.1     77.1  67.4   93.4    67.8             70.9
##  8  2014   34.7     77.5  68.0   93.7    68.0             71.3
##  9  2015   35.4     77.9  68.6   93.8    68.2             71.7
## 10  2016   35.9     76.9  69.1   93.9    68.4             71.9
## 11  2017   36.4     77.1  69.7   94.0    68.5             72.2
## 12  2018   36.8     76.7  70.2   94.1    68.7             72.4
## 13  2019   37.1     76.9  70.6   94.2    68.8             72.6
## 14  2020   37.4     77.0  71.0   94.2    68.9             72.8
## 15  2021   37.7     76.8  71.3   94.2    69.1             72.8
## 16  2022   37.6     76.9  71.0   94.2    67.5             72.9
```

5.2 Plot 2

Question: What is the global geographic distribution of access to safely managed drinking water services in 2022?

```
# Filter for Years:"2022" and Service level: "Safely managed service"
drinking_data_2022 <- drinking_data %>%
  filter(`Service.level` == "Safely managed service", Year == 2022) %>%
  filter(!is.na(Coverage)) %>%
  select(Country, ISO3, Coverage)

# Interactive map using plotly::plot_geo
fig <- plot_geo(drinking_data_2022)

fig <- fig %>%
  add_trace(
    z = ~Coverage,
    color = ~Coverage,
    colors = "viridis",
    text = ~paste("Negara:", Country,
                  "<br>Cakupan:", round(Coverage, 1), "%"),
    locations = ~ISO3,
    locationmode = "ISO-3",
    type = "choropleth"
  ) %>%
  colorbar(title = "Coverage (%)") %>%
  layout(
    title = "Coverage of Access to Safe Drinking Water by Country (2022)",
    geo = list(
      showframe = FALSE,
      showcoastlines = TRUE,
      coastlinecolor = "gray",
      projection = list(type = "natural earth")
    )
  )

fig
```



```
# Table form of the plot
print(drinking_data_2022)
```

```
##                                                  Country ISO3  Coverage
## 1                                            Afghanistan  AFG  30.03410
## 2                                                Albania  ALB  70.73607
## 3                                                Andorra  AND  90.64000
## 4                                                Armenia  ARM  82.41172
## 5                                                Austria  AUT  98.89632
## 6                                             Azerbaijan  AZE  71.61170
## 7                                                Belgium  BEL  99.73945
## 8                                             Bangladesh  BGD  59.06945
## 9                                               Bulgaria  BGR  95.65433
## 10                                               Bahrain  BHR  98.90398
## 11                                Bosnia and Herzegovina  BIH  86.97068
## 12                                      Saint Barthélemy  BLM 100.00000
## 13                                               Belarus  BLR  93.09976
## 14                                                Brazil  BRA  87.25872
## 15                                                Bhutan  BTN  73.34199
## 16                              Central African Republic  CAF   6.12644
## 17                                                Canada  CAN  99.03973
## 18                                           Switzerland  CHE  96.70000
## 19                                                 Chile  CHL  98.77136
## 20                                         Côte d'Ivoire  CIV  43.89216
## 21                      Democratic Republic of the Congo  COD  11.58434
## 22                                              Colombia  COL  73.85604
## 23                                            Costa Rica  CRI  80.50758
## 24                                                Cyprus  CYP  99.76506
## 25                                               Czechia  CZE  97.88341
## 26                                               Germany  DEU  99.91641
## 27                                               Denmark  DNK  99.91883
## 28                                    Dominican Republic  DOM  44.94118
## 29                                               Algeria  DZA  70.59793
## 30                                               Ecuador  ECU  67.08951
## 31                                                 Spain  ESP  99.56714
## 32                                               Estonia  EST  97.01938
## 33                                              Ethiopia  ETH  13.23790
## 34                                               Finland  FIN  99.64203
## 35                                                  Fiji  FJI  41.86358
## 36                                                France  FRA  99.70415
## 37  United Kingdom of Great Britain and Northern Ireland  GBR  99.80415
## 38                                               Georgia  GEO  69.14185
## 39                                                 Ghana  GHA  44.46808
## 40                                             Gibraltar  GIB 100.00000
## 41                                            Guadeloupe  GLP  95.70962
## 42                                                Gambia  GMB  47.67465
## 43                                         Guinea-Bissau  GNB  23.87160
## 44                                                Greece  GRC  98.87964
## 45                                             Guatemala  GTM  56.29331
## 46                                         French Guiana  GUF  91.48601
## 47                                                  Guam  GUM  99.05979
## 48                                    China, Hong Kong SAR  HKG 100.00000
## 49                                              Honduras  HND  65.20680
## 50                                               Hungary  HUN 100.00000
## 51                                             Indonesia  IDN  30.26617
## 52                                           Isle of Man  IMN  99.70983
## 53                                               Ireland  IRL  95.99141
## 54                              Iran (Islamic Republic of)  IRN  94.22125
```

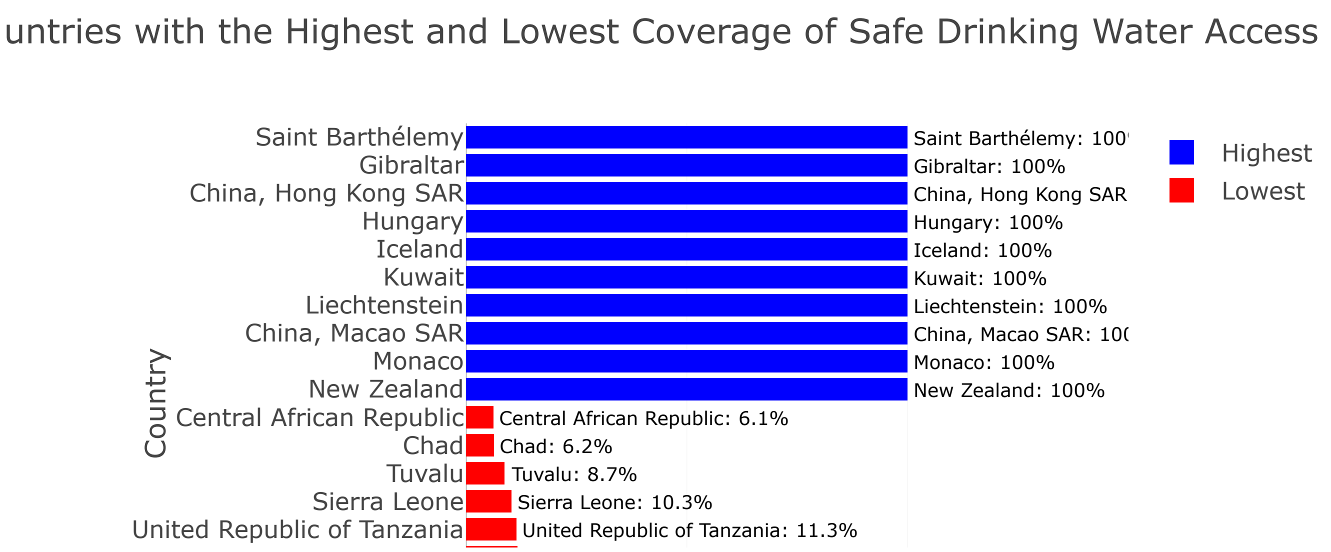
```
## 55                                     Iraq  IRQ  59.74263
## 56                                  Iceland  ISL 100.00000
## 57                                   Israel  ISR  99.47116
## 58                                    Italy  ITA  92.71056
## 59                                   Jordan  JOR  85.70913
## 60                                    Japan  JPN  98.65855
## 61                               Kyrgyzstan  KGZ  76.48715
## 62                                 Cambodia  KHM  29.13128
## 63                                 Kiribati  KIR  14.41407
## 64                        Republic of Korea  KOR  99.28012
## 65                                   Kuwait  KWT 100.00000
## 66         Lao People's Democratic Republic  LAO  17.87208
## 67                                  Lebanon  LBN  47.70000
## 68                            Liechtenstein  LIE 100.00000
## 69                                Sri Lanka  LKA  47.12753
## 70                                  Lesotho  LSO  28.21749
## 71                                Lithuania  LTU  94.98150
## 72                               Luxembourg  LUX  99.53408
## 73                                   Latvia  LVA  97.11127
## 74                         China, Macao SAR  MAC 100.00000
## 75               Saint Martin (French Part)  MAF  96.62876
## 76                                  Morocco  MAR  74.82420
## 77                                   Monaco  MCO 100.00000
## 78                      Republic of Moldova  MDA  75.22434
## 79                               Madagascar  MDG  22.23922
## 80                                   Mexico  MEX  43.03788
## 81                          North Macedonia  MKD  80.44703
## 82                                    Malta  MLT  99.77242
## 83                                  Myanmar  MMR  57.39650
## 84                               Montenegro  MNE  85.11892
## 85                                 Mongolia  MNG  39.28016
## 86                 Northern Mariana Islands  MNP  90.63918
## 87                               Martinique  MTQ  98.76963
## 88                                   Malawi  MWI  17.75708
## 89                                 Malaysia  MYS  93.94192
## 90                                  Mayotte  MYT  92.46111
## 91                            New Caledonia  NCL  96.86529
## 92                                  Nigeria  NGA  28.98488
## 93                                     Niue  NIU  93.54150
## 94              Netherlands (Kingdom of the)  NLD  99.96789
## 95                                   Norway  NOR  98.82311
## 96                                    Nepal  NPL  16.11664
## 97                              New Zealand  NZL 100.00000
## 98                                     Oman  OMN  90.85175
## 99                                 Pakistan  PAK  50.60178
## 100                                    Peru  PER  51.98521
## 101                             Philippines  PHL  47.90190
## 102                                   Palau  PLW  90.44085
## 103                                  Poland  POL  88.91450
## 104                             Puerto Rico  PRI  99.87311
## 105  Democratic People's Republic of Korea  PRK  66.53140
## 106                                Portugal  PRT  95.15673
## 107                                Paraguay  PRY  64.22060
## 108                      State of Palestine  PSE  80.33018
## 109                        French Polynesia  PYF  81.81326
## 110                                   Qatar  QAT  96.65482
```

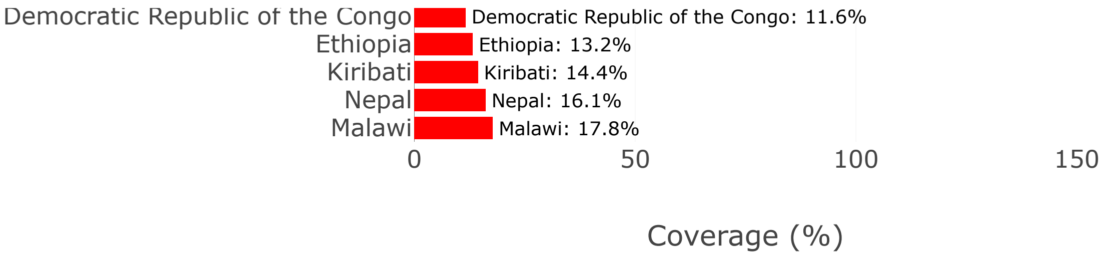
```
## 111                         Réunion  REU  95.75359
## 112                         Romania  ROU  82.07232
## 113              Russian Federation  RUS  76.23311
## 114                         Senegal  SEN  26.69259
## 115                       Singapore  SGP 100.00000
## 116                    Saint Helena  SHN  89.22964
## 117                    Sierra Leone  SLE  10.26242
## 118                      San Marino  SMR 100.00000
## 119                          Serbia  SRB  75.07525
## 120           Sao Tome and Principe  STP  36.30069
## 121                        Suriname  SUR  55.79615
## 122                        Slovakia  SVK  99.18473
## 123                        Slovenia  SVN  98.27432
## 124                          Sweden  SWE  99.73888
## 125        Turks and Caicos Islands  TCA  47.08508
## 126                            Chad  TCD   6.24705
## 127                            Togo  TGO  19.41779
## 128                      Tajikistan  TJK  55.29210
## 129                    Turkmenistan  TKM  94.88054
## 130                           Tonga  TON  29.52870
## 131                         Tunisia  TUN  74.30346
## 132                          Tuvalu  TUV   8.70710
## 133     United Republic of Tanzania  TZA  11.33565
## 134                          Uganda  UGA  18.68142
## 135                         Ukraine  UKR  87.62073
## 136        United States of America  USA  97.46839
## 137                      Uzbekistan  UZB  79.84530
## 138                        Viet Nam  VNM  57.78141
## 139       Wallis and Futuna Islands  WLF  68.88058
## 140                           Samoa  WSM  62.19171
## 141                        Zimbabwe  ZWE  26.51643
```

5.3 Plot 3

Question: Which countries have the highest and lowest clean water coverage? (Shows 10 countries each)

```
# Find the top 10 highest
top10 <- drinking_data_2022 %>%
  arrange(desc(Coverage)) %>%
  slice_head(n = 10)

# Find the top 10 lowest
bottom10 <- drinking_data_2022 %>%
  arrange(Coverage) %>%
  slice_head(n = 10)

# Merge and give them label
top10 <- top10 %>% mutate(Group = "Highest")
bottom10 <- bottom10 %>% mutate(Group = "Lowest")

plot_data <- bind_rows(top10, bottom10) %>%
  # Urutkan untuk tampilan barplot (descending untuk tertinggi, ascending untuk terendah)
  mutate(Country = factor(Country, levels = unique(Country)))

# Plotting
fig <- plot_ly(plot_data,
               x = ~Coverage,
               y = ~Country,
               color = ~Group,
               colors = c("Lowest" = "red", "Highest" = "blue"),
               type = "bar",
               text = ~paste0(Country, ": ", round(Coverage, 1), "%"),  # Teks yang akan dita
mpilkan
               textposition = "outside",
               textfont = list(color = "black"),
               orientation = "h",
               hovertemplate = "Country: %{y}<br>Coverage: %{x:.1f}%")  %>%
  layout(title = "10 Countries with the Highest and Lowest Coverage of Safe Drinking Water Ac
cess (2022)",
         xaxis = list(title = "Coverage (%)", range = c(0, 150) ),
         yaxis = list(title = "Country", autorange = "reversed"),
         barmode = "group",
         margin = list(l = 200, r = 50, t = 80, b = 120))

fig
```

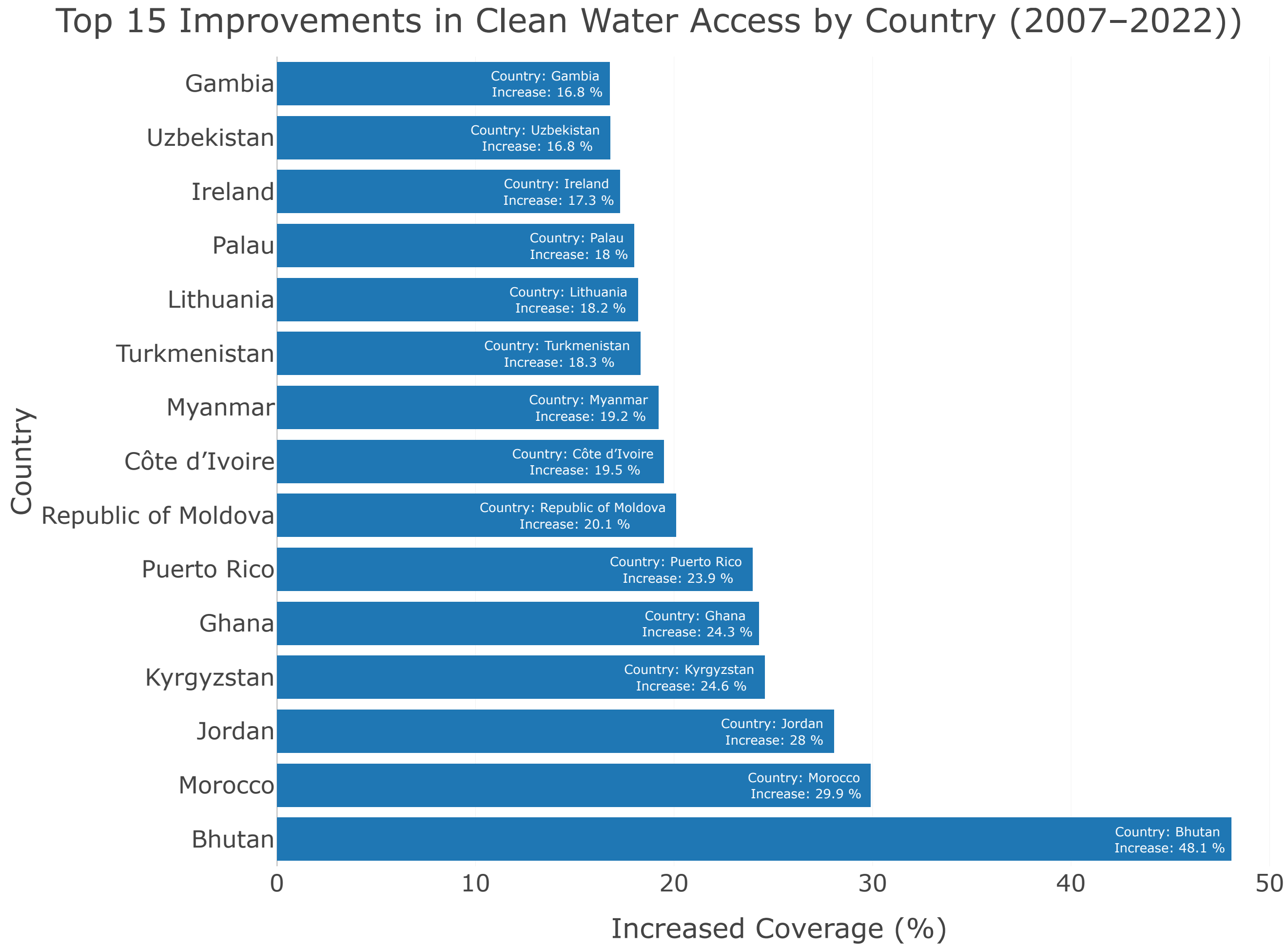




5.4 Plot 4

Question: Which country experienced the highest increase in access to clean water during the period 2007–2022?

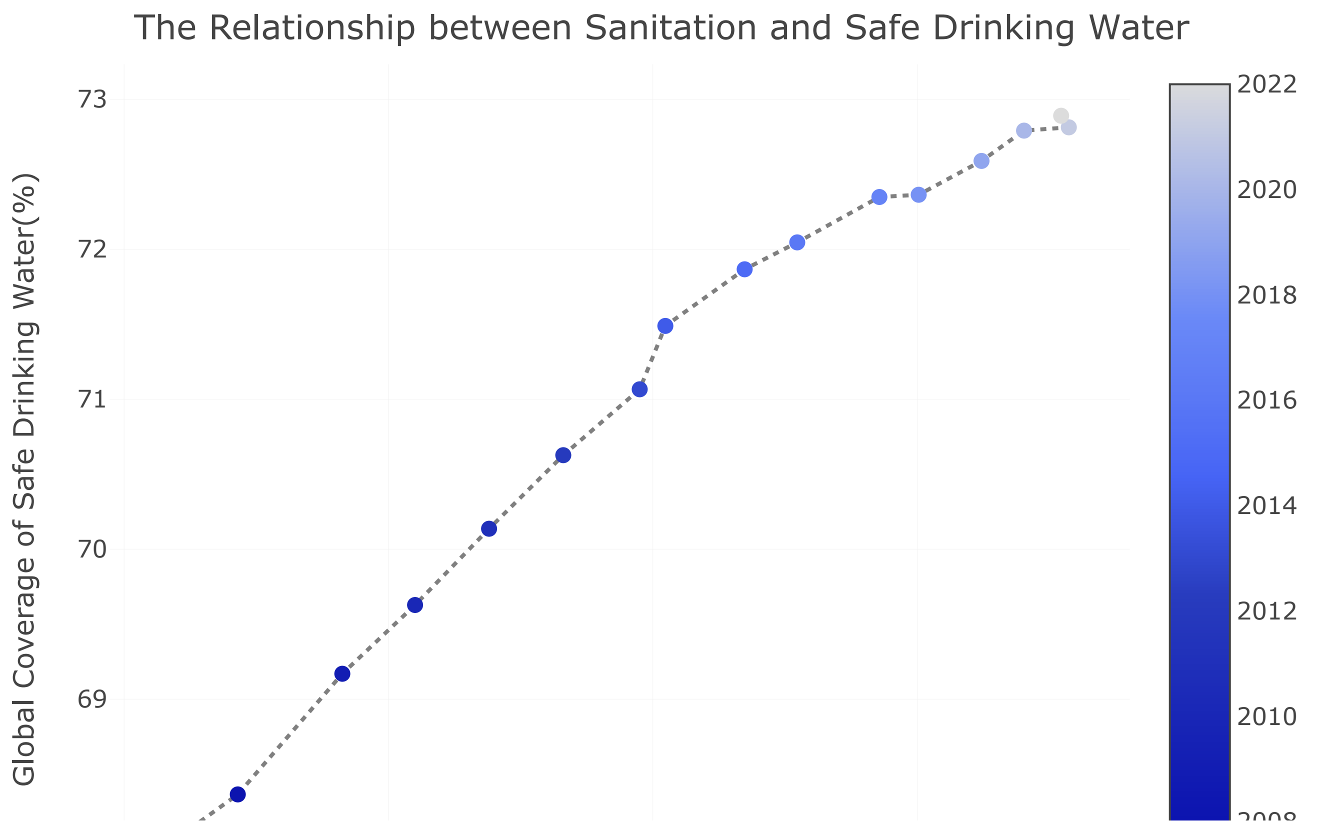
```
# Filter data service level: safely managed
drinking_increase <- drinking_data %>%
  filter(`Service.level` == "Safely managed service", Year %in% c(2007, 2022)) %>%
  select(Country, ISO3, Year, Coverage)

# Separate 2007 and 2022 data to calculate their difference
water_start <- drinking_increase %>%
  filter(Year == 2007) %>%
  select(Country, ISO3, Start = Coverage)

water_end <- drinking_increase %>%
  filter(Year == 2022) %>%
  select(Country, ISO3, End = Coverage)

water_change <- left_join(water_start, water_end, by = c("Country", "ISO3")) %>%
  mutate(Change = End - Start) %>%
  filter(!is.na(Change))

# Find top 15 country
top_increase <- water_change %>%
  arrange(desc(Change)) %>%
  slice_head(n = 15) %>%
  mutate(Country = factor(Country, levels = rev(Country)))

# Plotting
fig <- plot_ly(top_increase,
               x = ~Change,
               y = ~Country,
               type = 'bar',
               orientation = 'h',
               marker = list(color = 'viridis'),
               text = ~paste("Country:", Country,
                             "<br>Increase:", round(Change, 1), "%"),
               hoverinfo = "text") %>%
  layout(title = "Top 15 Improvements in Clean Water Access by Country (2007-2022)",
         xaxis = list(title = "Increased Coverage (%)"),
         yaxis = list(title = "Country", autorange = "reversed"))

fig
```

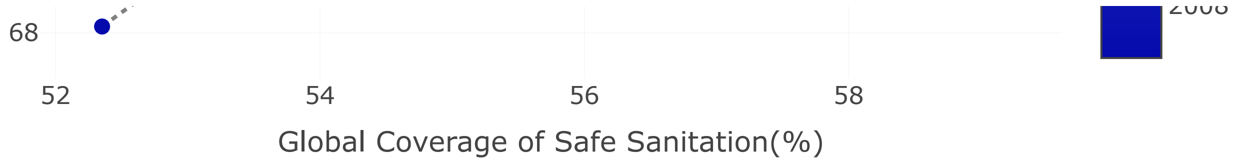


5.4 Plot 5

Question: Is there a positive correlation between increased access to clean drinking water and better sanitation practices?

```
#Filter data service level: safely managed service
clean_water <- drinking_data %>%
   filter(`Service.level` == "Safely managed service") %>%
  group_by(Year) %>%
  summarise(Global_Water = mean(Coverage, na.rm = TRUE), .groups = "drop")

clean_sanitation <- sanitation_data %>%
  filter(`Service.level` == "Safely managed service") %>%
  group_by(Year) %>%
  summarise(Global_Sanitation = mean(Coverage, na.rm = TRUE), .groups = "drop")

#Merge them by years
global_trend <- inner_join(clean_water, clean_sanitation, by = "Year")

#Plotting
plot_ly(global_trend,
        x = ~Global_Sanitation,
        y = ~Global_Water,
        type = 'scatter',
        mode = 'markers+lines',
        text = ~paste("Tahun:", Year,
                      "<br>Sanitasi Aman:", round(Global_Sanitation, 1), "%",
                      "<br>Air Minum Aman:", round(Global_Water, 1), "%"),
        hoverinfo = "text",
        textposition = "top center",
        marker = list(size = 8, color = ~Year, colorscale = 'Blues', showscale = TRUE),
        line = list(color = 'gray', dash = 'dot')) %>%
  layout(
    title = "The Relationship between Sanitation and Safe Drinking Water",
    xaxis = list(title = "Global Coverage of Safe Sanitation(%)"),
    yaxis = list(title = "Global Coverage of Safe Drinking Water(%)"),
    hovermode = "closest"
  )
```





Correlation Test

```
cor_value <- cor(global_trend$Global_Sanitation, global_trend$Global_Water, use = "complete.o
bs", method = "pearson")
print(cor_value)
```

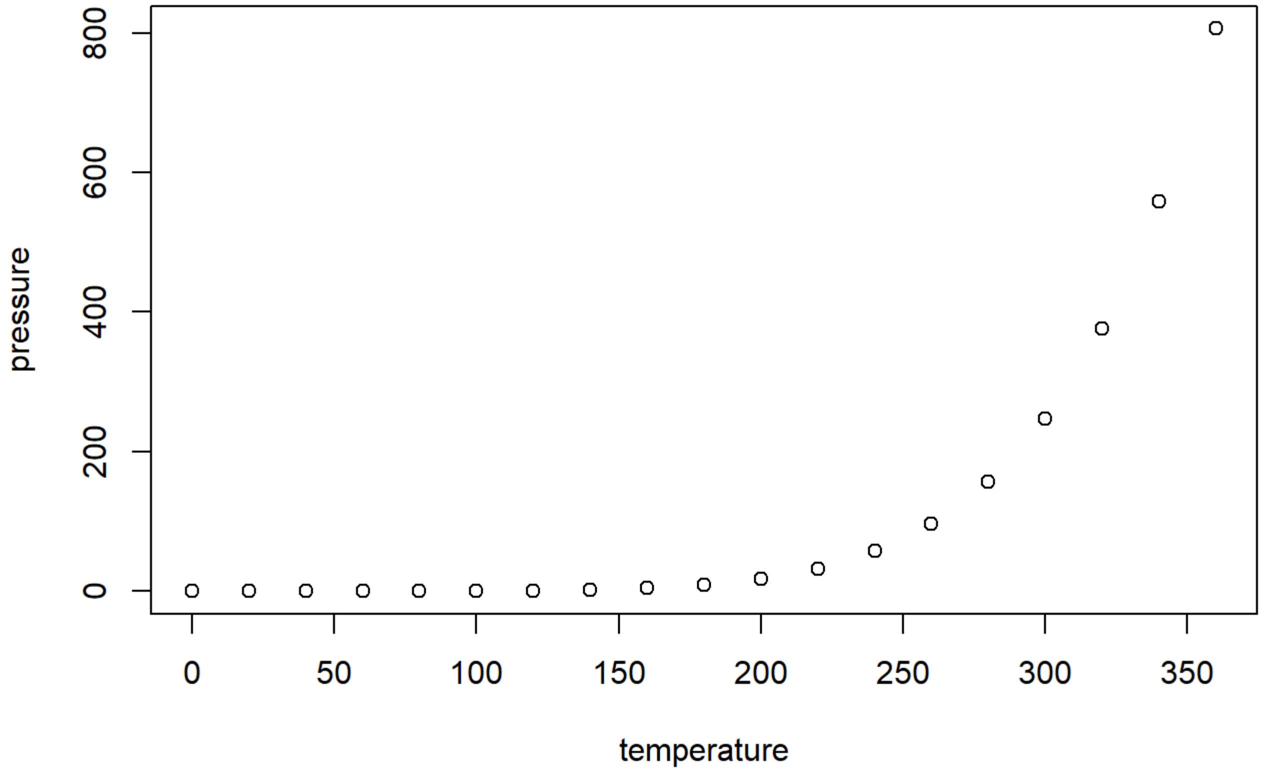
```
## [1] 0.9863773
```

```
cor.test(global_trend$Global_Sanitation, global_trend$Global_Water)
```

```
##
##  Pearson's product-moment correlation
##
## data:  global_trend$Global_Sanitation and global_trend$Global_Water
## t = 22.436, df = 14, p-value = 2.252e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9601297 0.9953861
## sample estimates:
##       cor
## 0.9863773
```

#Interpretation of Correlation Values and Significance Tests - Pearson Correlation Value (r) = 0.986, indicating a very strong positive relationship between safe sanitation coverage and access to safe drinking water - p-value = 2.252e-12, a value less than 0.05, indicating a very significant relationship Based on the analysis of WASH Household data from 2007 to 2022, a very strong correlation was found between safe sanitation coverage and access to safe drinking water globally, with a Pearson correlation value of r = 0.986 (p <0.001).

This shows that countries that improve sanitation infrastructure also tend to experience significant increases in access to safe drinking water.



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.