In [ ]:
```r
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(tidyr)
library(readxl)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

In [ ]:
```r
# Load the dataset
data <- read_excel("/content/Sanitation (1) (1).xlsx")
```

In [ ]:
```r
str(data)
head(data)
```

```
tibble [242 × 8] (S3: tbl_df/tbl/data.frame)
 $ Country               : chr [1:242] "Afghanistan" "Albania" "Algeria" "Am
erican Samoa" ...
 $ Year                  : chr [1:242] "2022" "2022" "2022" "2021-2022" ...
 $ Safely managed service: num [1:242] NA 56.4 62.4 37 100 ...
 $ At least basic service: num [1:242] 56 NA NA NA NA ...
 $ Basic service         : num [1:242] NA 43 23.4 17.1 0 ...
 $ Limited               : num [1:242] 12.04 0.57 10.82 44.32 0 ...
 $ Unimproved            : num [1:242] 23.16 0.13 3.36 1.53 0 ...
 $ Open defecation       : num [1:242] 8.84 0 0 0.79 0 ...
```

A tibble: 6 × 8

Country	Year	Safely managed service	At least basic service	Basic service	Limited	Unimproved	Oper defecatior
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Afghanistan	2022	NA	55.95	NA	12.04	23.16	8.84
Albania	2022	56.35	NA	42.95	0.57	0.13	0.00
Algeria	2022	62.41	NA	23.42	10.82	3.36	0.00
American Samoa	2021-2022	36.99	NA	17.15	44.32	1.53	0.79
Andorra	2022	100.00	NA	0.00	0.00	0.00	0.00
Angola	2022	NA	52.18	NA	21.28	9.25	17.29

In [ ]:
```r
# Function to calculate mode
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

In [ ]:
```r
# Replace missing values
data_filled <- data %>%
  mutate(across(where(is.numeric), ~ifelse(is.na(.), mean(., na.rm = TRUE),
  mutate(across(where(is.character), ~ifelse(is.na(.), get_mode(.), .)))

# View the filled dataset
head(data_filled)
```

A tibble: 6 × 8

Country (or area), SDG region, world	Year	Safely managed service	At least basic service	Basic service	Limited	Unimproved	O defecat
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<d
Afghanistan	2022.0	59.57732	55.95000	22.43034	12.04	23.16	
Albania	2022.0	56.35000	76.64883	42.95000	0.57	0.13	
Algeria	2022.0	62.41000	76.64883	23.42000	10.82	3.36	
American Samoa	2021-2022	36.99000	76.64883	17.15000	44.32	1.53	
Andorra	2022.0	100.00000	76.64883	0.00000	0.00	0.00	
Angola	2022.0	59.57732	52.18000	22.43034	21.28	9.25	1

In [ ]:
```r
install.packages(c("ggmap", "maps", "reshapes2", "viridis", "sf"))
```

```
Installing packages into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Warning message:
“package ‘reshapes2’ is not available for this version of R

A version of this package for your version of R might be available elsewher
e,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-pac
kages”
also installing the dependencies ‘proxy’, ‘e1071’, ‘wk’, ‘png’, ‘plyr’, ‘jpe
g’, ‘bitops’, ‘gridExtra’, ‘classInt’, ‘s2’, ‘units’
```

In [ ]:
```r
# Load required libraries
library(dplyr)
library(ggplot2)
```

```
library(ggmap)
library(maps)
library(ggplot2)
library(viridis)
library(reshape2)
library(sf)
```

```
Linking to GEOS 3.11.1, GDAL 3.6.4, PROJ 9.1.1; sf_use_s2() is TRUE
```

In [ ]:
```
str(data)
summary(data)
```

```
tibble [242 × 8] (S3: tbl_df/tbl/data.frame)
 $ Country                : chr [1:242] "Afghanistan" "Albania" "Algeria" "Am
erican Samoa" ...
 $ Year                   : chr [1:242] "2022" "2022" "2022" "2021-2022" ...
 $ Safely managed service: num [1:242] NA 56.4 62.4 37 100 ...
 $ At least basic service: num [1:242] 56 NA NA NA NA ...
 $ Basic service          : num [1:242] NA 43 23.4 17.1 0 ...
 $ Limited                : num [1:242] 12.04 0.57 10.82 44.32 0 ...
 $ Unimproved             : num [1:242] 23.16 0.13 3.36 1.53 0 ...
 $ Open defecation        : num [1:242] 8.84 0 0 0.79 0 ...
   Country              Year           Safely managed service
 Length:242         Length:242         Min.   :  0.00
 Class :character   Class :character   1st Qu.: 32.05
 Mode  :character   Mode  :character   Median : 62.41
                                       Mean   : 59.58
                                       3rd Qu.: 88.74
                                       Max.   :100.00
                                       NA's   :93
 At least basic service Basic service      Limited         Unimproved
 Min.   :  5.64         Min.   : 0.00   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 59.89         1st Qu.: 5.03   1st Qu.: 0.090   1st Qu.: 0.050
 Median : 88.67         Median :15.18   Median : 2.420   Median : 1.530
 Mean   : 76.65         Mean   :22.43   Mean   : 6.778   Mean   : 7.582
 3rd Qu.: 98.03         3rd Qu.:30.36   3rd Qu.:10.270   3rd Qu.: 9.650
 Max.   :100.00         Max.   :86.80   Max.   :44.550   Max.   :64.540
 NA's   :148            NA's   :93      NA's   :1        NA's   :1
 Open defecation
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 0.110
 Mean   : 5.184
 3rd Qu.: 4.540
 Max.   :67.000
 NA's   :1
```

In [ ]:
```
# Overview of the dataset
str(data_filled)
summary(data_filled)
```

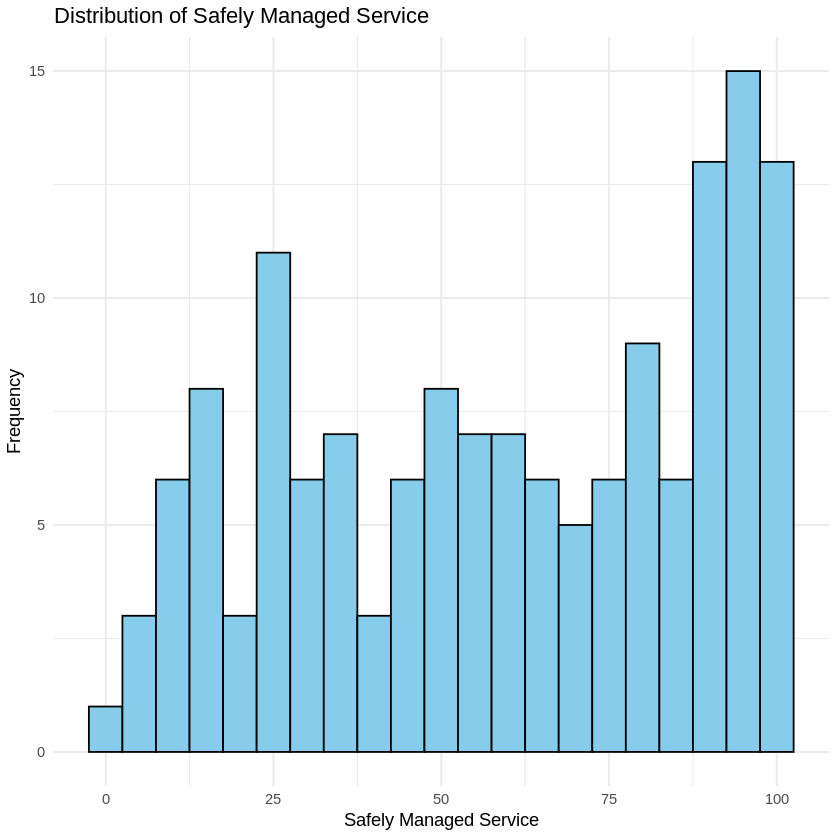
```
Error: object 'data_filled' not found
Traceback:

1. .handleSimpleError(function (cnd)
 . {
 .     watcher$capture_plot_and_output()
 .     cnd <- sanitize_call(cnd)
 .     watcher$push(cnd)
 .     switch(on_error, continue = invokeRestart("eval_continue"),
 .         stop = invokeRestart("eval_stop"), error = invokeRestart("eval_er
ror",
 .             cnd))
 . }, "object 'data_filled' not found", base::quote(eval(expr, envir)))
```

In [ ]:
```
colnames(data_filled)
```

'Country (or area), SDG region, world' · 'Year' · 'Safely managed service' · 'At least basic service' · 'Basic service' · 'Limited' · 'Unimproved' · 'Open defecation'

In [ ]:
```
library(ggplot2)
library(dplyr)

# Assuming your dataset is loaded into a variable called 'data'

# Create the histogram
ggplot(data, aes(x = `Safely managed service`)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Safely Managed Service",
       x = "Safely Managed Service",
       y = "Frequency")

# Find the top 5 countries with the highest frequencies of Safely Managed Se
top_countries <- data %>%
  filter(!is.na(`Safely managed service`)) %>%
  arrange(desc(`Safely managed service`)) %>%
  head(5) %>%
  select(Country, `Safely managed service`)

print(top_countries)
```

```
Warning message:
“Removed 93 rows containing non-finite outside the scale range (`stat_bin()
`).”
# A tibble: 5 × 2
  Country   `Safely managed service`
  <chr>                        <dbl>
1 Andorra                        100
2 Kuwait                         100
3 Monaco                         100
4 Singapore                      100
5 Qatar                           99.9
```



1. **Skewness:**

The distribution is right-skewed (positively skewed). This means that there is a tail of higher values on the right side of the plot.

2. **Central Tendency:**

The median is likely between 75 and 100, based on the peak of the distribution. The mean would be higher than the median due to the right skew.

3. **Spread:**

The data has a relatively wide spread with values ranging from 0 to 100. There are several modes (peaks) in the distribution, indicating multiple groups or clusters within the data.

4. **Outliers:**

A few outliers are present on the left side of the plot (values below 10), but they are not as extreme as in a heavily skewed distribution.

5. **Overall Distribution:**

A majority of the observations have a high level of Safely Managed Service. A smaller proportion has lower levels of Safely Managed Service.

In [ ]:
```
services <- c('Safely.managed.service', 'At.least.basic.service', 'Basic.ser
```

In [ ]:
```
install.packages("mice")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘minqa’, ‘nloptr’, ‘ucminf’, ‘numDeriv’, ‘i
terators’, ‘lme4’, ‘ordinal’, ‘foreach’, ‘shape’, ‘RcppEigen’, ‘pan’, ‘jom
o’, ‘glmnet’, ‘mitml’
```

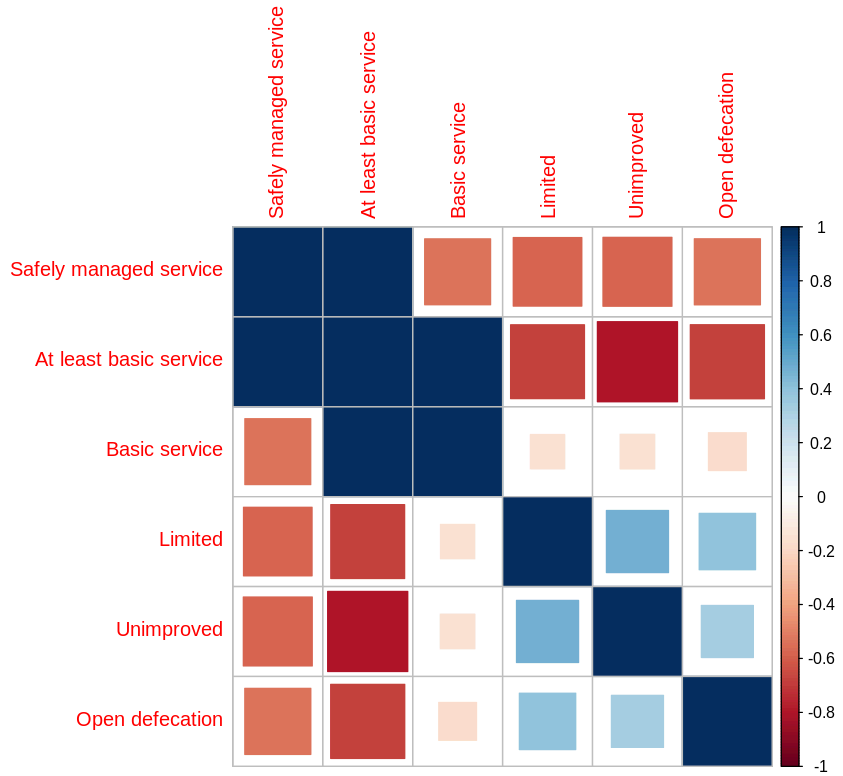
In [ ]:
```
correlation_matrix <- cor(data[, 3:8], use = "complete.obs")

# Display the correlation matrix
print(correlation_matrix)
```

```
                       Safely managed service At least basic service
Safely managed service                      1                      1
At least basic service                      1                      1
Basic service                               1                      1
Limited                                    -1                     -1
Unimproved                                 -1                     -1
Open defecation                             1                      1
                       Basic service Limited Unimproved Open defecation
Safely managed service             1      -1         -1                1
At least basic service             1      -1         -1                1
Basic service                      1      -1         -1                1
Limited                           -1       1          1               -1
Unimproved                        -1       1          1               -1
Open defecation                    1      -1         -1                1
```

In [ ]:
```
install.packages("corrplot")  # Run this line if you need to install the pac
library(corrplot)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

corrplot 0.94 loaded
```

In [ ]:
```
# Compute the correlation matrix
correlation_matrix <- cor(data[, 3:ncol(data)], use = "pairwise.complete.obs

# Plot the correlation matrix
corrplot(correlation_matrix, method = "square")
```



In [ ]:
```
install.packages("treemap")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘gridBase’, ‘igraph’
```

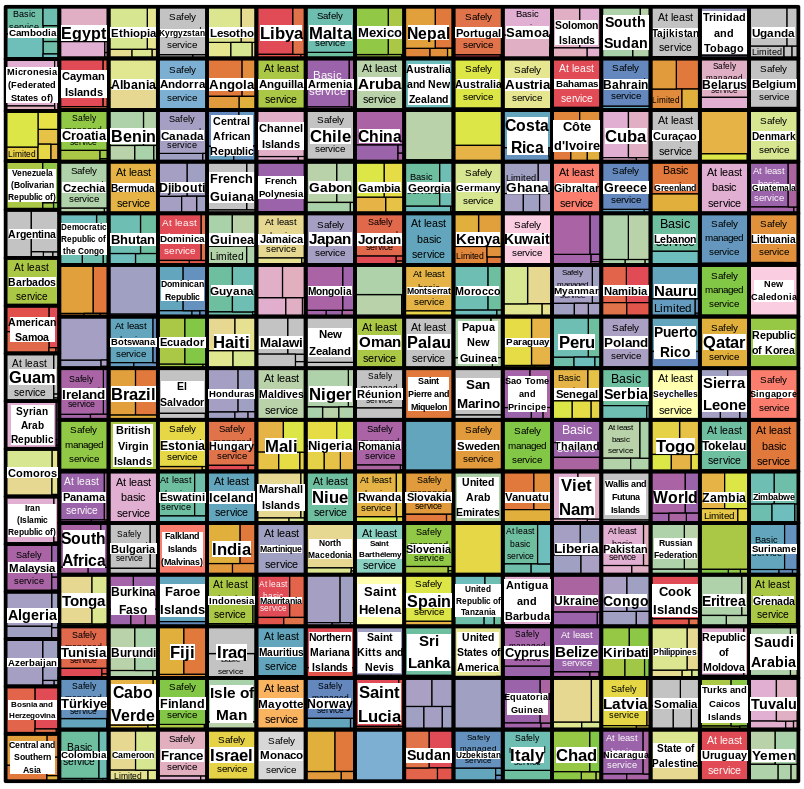
In [ ]:
```
library(treemap)
```

In [ ]:
```
data_long <- data %>%
  pivot_longer(cols = c(`Safely managed service`, `At least basic service`,
               names_to = "Service_Type",
               values_to = "Percentage")

# Remove rows with NA values for plotting
data_long <- data_long %>% filter(!is.na(Percentage))

# Create the treemap
treemap(data_long,
        index = c("Country", "Service_Type"),
        vSize = "Percentage",
        vColor = "Service_Type",
        draw = TRUE,
        title = "Sanitation Services by Country",
        fontfamily.title = "Arial",
        fontsize.title = 14,
        fontsize.labels = 10,
```

```
        bg.labels = "white",
        palette = "Set3")
```

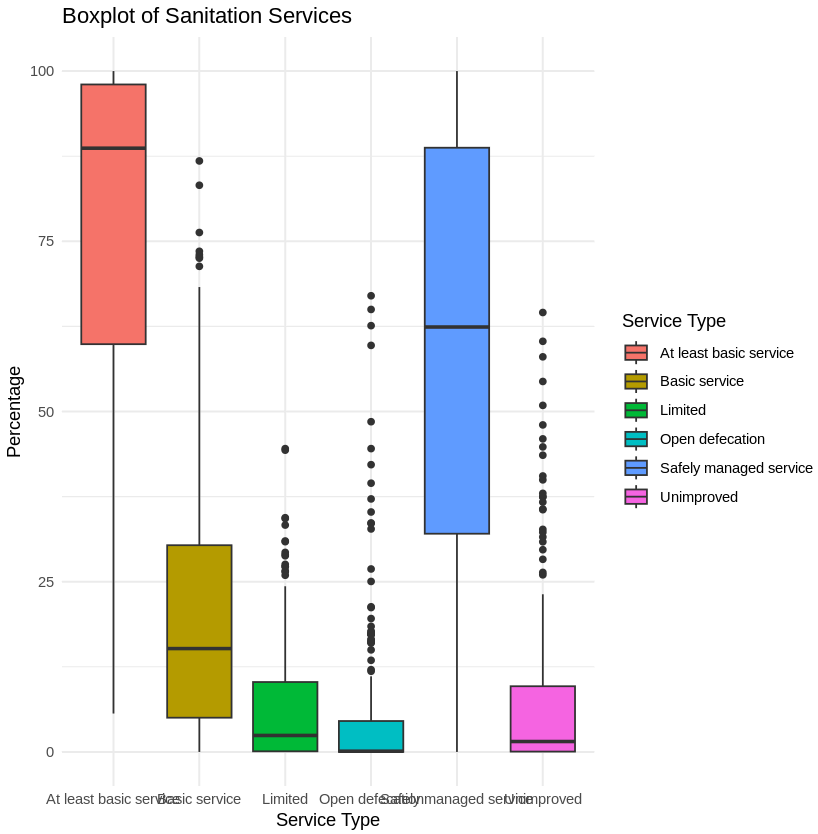
```
Warning message in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
$x, x$y, :
“font family 'Arial' not found in PostScript font database”
Warning message in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x
$x, x$y, :
“font family 'Arial' not found in PostScript font database”
```

Sanitation Services by Country



In [ ]:
```
ggplot(data_long, aes(x = `Service Type`, y = Percentage, fill = `Service Ty
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Sanitation Services",
       x = "Service Type",
       y = "Percentage")
```

```
Warning message:
“Removed 337 rows containing non-finite outside the scale range
(`stat_boxplot()`).”
```



1. Outliers:

All service types have outliers, as indicated by the individual data points outside the whiskers. Open defecation and Limited have the most outliers, suggesting more extreme values in these categories. 2. Overlap:

There is some overlap between the distributions of different service types, indicating that some regions or countries may have a mix of different sanitation services.

Summary:

- Safely managed service is the most prevalent and consistent type of sanitation service, with a relatively high median and low variability.
- Open defecation is the least prevalent and most variable type, with a low median and a wide range of percentages.
- At least basic service and Basic service have similar distributions, with moderate median values and variability.
- Limited and Unimproved categories also show significant variability and overlap with other categories, indicating a diverse range of sanitation conditions across different regions.

In [ ]:
```
install.packages("viridis")
```

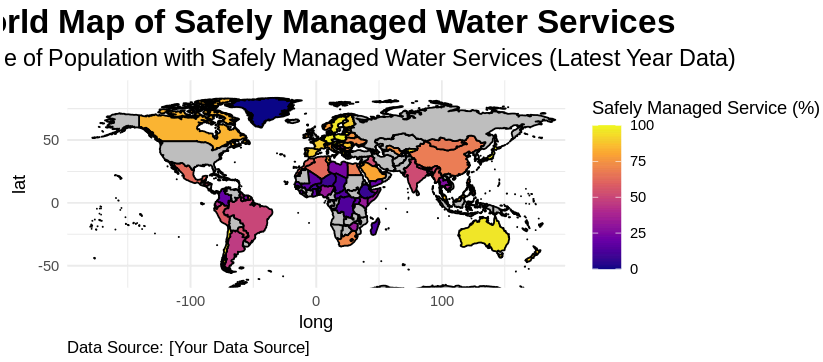
```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependency ‘gridExtra’
```

In [ ]:
```r
install.packages("maps")
```

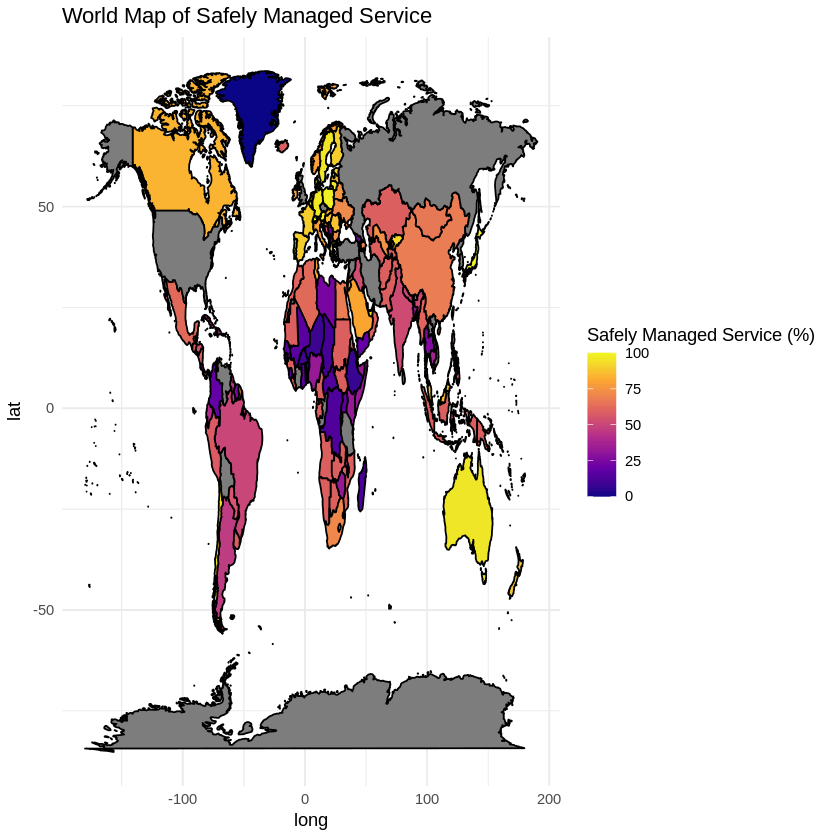
```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

In [ ]:
```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(viridis)
library(maps)


# Get world map data
world_map <- map_data("world")

# Merge with your dataset
data_map <- left_join(world_map, data, by = c("region" = "Country"))

# Plot the map
ggplot(data_map, aes(long, lat, group = group, fill = `Safely managed servic
  geom_polygon(color = "black") +
  scale_fill_viridis_c(option = "plasma", na.value = "grey", name = "Safely
  labs(title = "World Map of Safely Managed Water Services",
       subtitle = "Percentage of Population with Safely Managed Water Servic
       caption = "Data Source: [Your Data Source]") + # Add your data source
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 20, face = "bold"),
    plot.subtitle = element_text(hjust = 0.5, size = 14),
    plot.caption = element_text(hjust = 0, size = 10)
  ) +
  coord_fixed(xlim = c(-180, 180), ylim = c(-60, 90)) # Adjust the limits if
```

```
Attaching package: ‘maps’


The following object is masked from ‘package:viridis’:

    unemp
```



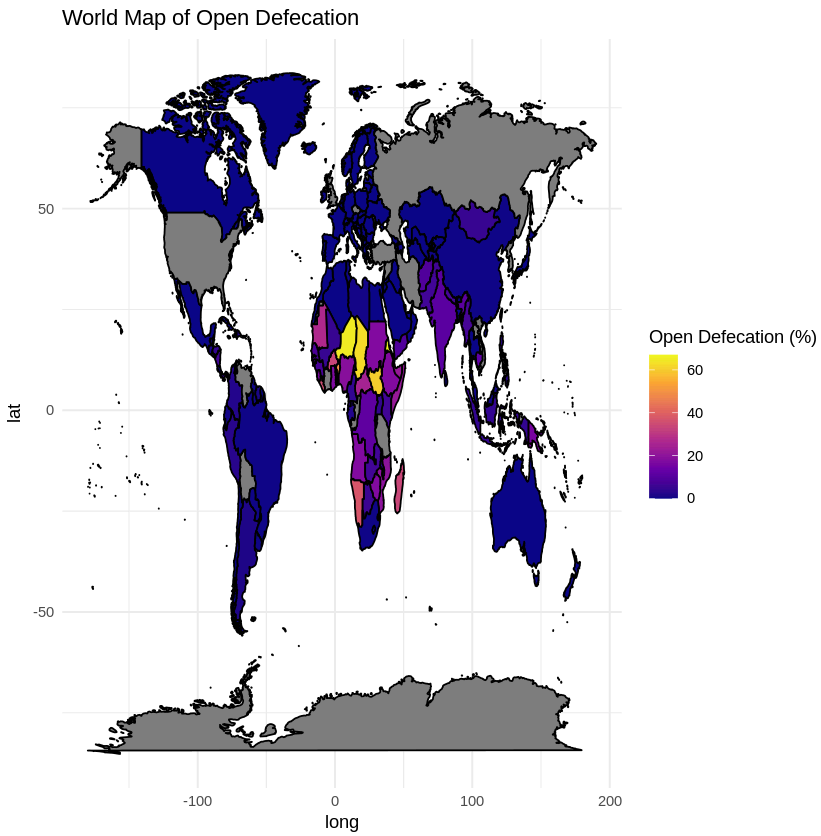
In [ ]:
```
# Get world map data
world_map <- map_data("world")

# Merge with your data
data_map <- left_join(world_map, data_filled, by = c("region" = "Country (or

# Plot the map
ggplot(data_map, aes(long, lat, group = group, fill = `Safely managed servic
  geom_polygon(color = "black") +
  scale_fill_viridis_c(option = "plasma", name = "Safely Managed Service (%)
  labs(title = "World Map of Safely Managed Service") +
  theme_minimal()
```



Summary:

- The world map highlights significant disparities in access to Safely Managed Service across different regions. Some regions, particularly in Europe, North America, and Australia, have high levels of Safely Managed Service, while others, especially in Africa and South Asia, have low levels.
- Developed countries generally have higher rates than developing countries, and urban areas often have better access than rural areas. Income inequality and economic development are likely factors influencing the distribution of Safely Managed Service.
- Regional clustering influence the distribution of Safely Managed Service. Environmental factors such as climate, topography, and water availability may also play a role.

In [ ]:
```
# Plot the map for Open Defecation
ggplot(data_map, aes(long, lat, group = group, fill = `Open defecation`)) +
  geom_polygon(color = "black") +
  scale_fill_viridis_c(option = "plasma", name = "Open Defecation (%)") +
  labs(title = "World Map of Open Defecation") +
  theme_minimal()
```



Summary:

- The world map highlights significant disparities in access to sanitation facilities and the practice of Open Defecation across different regions. Many countries in Sub-Saharan Africa have high rates, while those in Western Europe and North America have low rates.
- Developing countries generally have higher rates than developed countries, and rural areas often have higher prevalence than urban areas.
- Rural areas often have higher prevalence of Open Defecation than urban areas. Coastal regions may have lower rates of Open Defecation than landlocked regions.