Life_expectancy

2025-11-18

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(here)
```

```
## Warning: package ’here’ was built under R version 4.4.3
```

```
## here() starts at C:/Users/dhmha/OneDrive/Documents
```

```
library(janitor)
```

```
## Warning: package ’janitor’ was built under R version 4.4.3
```

```
##
## Attaching package: ’janitor’
##
## The following objects are masked from ’package:stats’:
##
##     chisq.test, fisher.test
```

```
library(zoo)
```

```
##
## Attaching package: ’zoo’
##
## The following objects are masked from ’package:base’:
##
##     as.Date, as.Date.numeric
```

```
library(broom)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: ’car’
##
## The following object is masked from ’package:dplyr’:
##
##     recode
##
## The following object is masked from ’package:purrr’:
##
##     some
```

```
library(glmnet)
```

```
## Warning: package ’glmnet’ was built under R version 4.4.3
```

```
## Loading required package: Matrix
##
## Attaching package: ’Matrix’
##
## The following objects are masked from ’package:tidyr’:
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

```
library(caret)
```

```
## Warning: package ’caret’ was built under R version 4.4.3
```

```
## Loading required package: lattice
##
## Attaching package: ’caret’
##
## The following object is masked from ’package:purrr’:
##
##     lift
```

```
library(scales)
```

```
##
## Attaching package: ’scales’
##
## The following object is masked from ’package:purrr’:
##
##     discard
##
## The following object is masked from ’package:readr’:
##
##     col_factor
```

```r
library(patchwork)
library(readr)
library(purrr)
library(Metrics)
```

```
## Warning: package ’Metrics’ was built under R version 4.4.3

##
## Attaching package: ’Metrics’
##
## The following objects are masked from ’package:caret’:
##
##     precision, recall
```

```r
library(treemapify)
```

```
## Warning: package ’treemapify’ was built under R version 4.4.3
```

```r
safe_read <- function(filepath) {
  if (!file.exists(filepath)) {
    stop(paste("Missing file:", filepath))
  }
  read_csv(filepath, show_col_types = FALSE)
}

files <- list(
  suicide = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\suicide_rates_data.csv",
  hale = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\healthy_life_expectancy_data.csv",
  child_u5 = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\child_deaths_less_than_5_yrs_data.cs
  child_5_9 = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\child_deaths_5_9_yrs_data.csv",
  adult_mort = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\adult_mortality_data.csv",
  homicide = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\homicide_data.csv",
  life_expectancy = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\life_expectancy_data.csv",
  ncd_deaths = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\ncd_deaths_data.csv",
  premature_ncd = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\premature_deaths_ncd_data.csv",
  poisoning = "C:\\Users\\dhmha\\OneDrive\\Documents\\IDMP_Project\\unintentional_poisoning_data.csv"
)
raw <- map(files, safe_read)


raw$hale_60 <- raw$hale %>% filter(Indicator == "Healthy life expectancy (HALE) at age 60 (years)")
raw$hale_birth <- raw$hale %>% filter(Indicator == "Healthy life expectancy (HALE) at birth (years)")
raw$life_60 <- raw$life_expectancy %>% filter(Indicator == "Life expectancy at age 60 (years)")
raw$life_birth <- raw$life_expectancy %>% filter(Indicator == "Life expectancy at birth (years)")

clean_data <- function(df, new_name) {
  df %>%
    select(ParentLocation, Location, Period, Dim1, FactValueNumeric) %>%
    rename(
      parentlocation = ParentLocation,
      country = Location,
      year = Period,
      gender = Dim1,
      value = FactValueNumeric
    ) %>%
    filter(year >= 2000, year <= 2019) %>%
    filter(gender != "Both sexes") %>%
    mutate(!!new_name := value) %>%
    select(-value)
}

cleaned <- list(
  suicide = clean_data(raw$suicide, "suicide_value"),
  hale_60 = clean_data(raw$hale_60, "hale_60_value"),
  hale_birth = clean_data(raw$hale_birth, "hale_value"),
  child_u5 = clean_data(raw$child_u5, "child_death_under_5_value"),
  child_5_9 = clean_data(raw$child_5_9, "child_death_over_5_value"),
  adult_mort = clean_data(raw$adult_mort, "mortality_value"),
  homicide = clean_data(raw$homicide, "homicide_value"),
  life_60 = clean_data(raw$life_60, "life_expectancy_60_value"),
  life_birth = clean_data(raw$life_birth, "life_expectancy_value"),
  ncd_deaths = clean_data(raw$ncd_deaths, "ncd_deaths_value"),
  premature_ncd = clean_data(raw$premature_ncd, "premature_ncd_value"),
  poisoning = clean_data(raw$poisoning, "poisoning_value")
)


cleaned$adult_mort <- cleaned$adult_mort %>% mutate(mortality_value = mortality_value * 1000)
cleaned$poisoning <- cleaned$poisoning %>% mutate(poisoning_value = poisoning_value * 100000)
by_cols <- c("parentlocation", "country", "year", "gender")
merged <- reduce(cleaned, full_join, by = by_cols)
if (all(c("ncd_deaths_value", "premature_ncd_value") %in% names(merged))) {
  merged <- merged %>% mutate(premature_ncd_deaths_count = round(ncd_deaths_value * (premature_ncd_valu
}

numeric_cols <- merged %>% select(where(is.numeric)) %>% names()
merged_imputed <- merged %>%
  arrange(country, gender, year) %>%
  group_by(country, gender) %>%
  mutate(across(all_of(numeric_cols), ~ na.locf(.x, na.rm = FALSE))) %>%
  mutate(across(all_of(numeric_cols), ~ na.locf(.x, fromLast = TRUE, na.rm = FALSE))) %>%
  ungroup()
reg_df <- merged_imputed %>% select(where(is.numeric)) %>% drop_na()
response <- "life_expectancy_value"
leakage_vars <- c("hale_value", "hale_60_value", "life_expectancy_60_value")
X <- reg_df %>%
  select(-all_of(response), -any_of(leakage_vars)) %>%
  mutate(across(everything(), scale))

y <- reg_df[[response]]
set.seed(123)
idx <- sample(seq_len(nrow(X)), size = 0.8 * nrow(X))
X_train <- X[idx, ]
X_test  <- X[-idx, ]
y_train <- y[idx]
```

```r
y_test  <- y[-idx]
formula_ols <- as.formula(paste(response, "~", paste(names(X_train), collapse = " + ")))
ols <- lm(formula_ols, data = bind_cols(X_train, life_expectancy_value = y_train))

# diagnostics
pred_test <- predict(ols, newdata = X_test)
metrics <- tibble(
  MAE = mae(y_test, pred_test),
  RMSE = rmse(y_test, pred_test),
  R2 = cor(y_test, pred_test)^2
)

# lasso model
lasso_cv <- cv.glmnet(as.matrix(X_train), y_train, alpha = 1, standardize = FALSE)
lasso_model <- glmnet(as.matrix(X_train), y_train, lambda = lasso_cv$lambda.min)
lasso_coefs <- coef(lasso_model)

# SAVing RESULTS
dir.create(here("output"), showWarnings = FALSE)
write_csv(merged_imputed, here("output", "merged_imputed.csv"))
write_csv(tidy(ols), here("output", "ols_coefficients.csv"))
write_csv(metrics, here("output", "regression_metrics.csv"))

#visualisations
#global trends
plot_df <- merged_imputed %>%
  mutate(YearGroup = cut(year, breaks = seq(2000, 2020, 4), right = FALSE,
                         labels = c("2000-2003", "2004-2007", "2008-2011", "2012-2015", "2016-2019")))
  group_by(YearGroup) %>%
  summarise(across(c(suicide_value, hale_60_value, hale_value), ~ mean(.x, na.rm = TRUE))) %>%
  pivot_longer(-YearGroup)

p1 <- ggplot(plot_df, aes(YearGroup, value, fill = YearGroup)) +
  geom_col() +
  facet_wrap(~ name, scales = "free_y") +
  scale_y_continuous(labels = comma) +
  labs(title = "Global Trends (Averages)", x = "Year Group", y = "Value") +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())

#TOP & BOTTOM COUNTRIES

country_avg <- merged_imputed %>%
  group_by(country) %>%
  summarise(avg_life_birth = mean(life_expectancy_value, na.rm = TRUE)) %>%
  filter(!is.nan(avg_life_birth) & avg_life_birth > 0) %>%
  arrange(desc(avg_life_birth))

top5_countries <- head(country_avg$country, 5)
bottom5_countries <- tail(country_avg$country, 5) # Now this won't be empty

top_df <- merged_imputed %>% filter(country %in% top5_countries) %>%
  group_by(country, year) %>% summarise(val = mean(life_expectancy_value, na.rm=TRUE), .groups="drop")
```

```
bottom_df <- merged_imputed %>% filter(country %in% bottom5_countries) %>%
  group_by(country, year) %>% summarise(val = mean(life_expectancy_value, na.rm=TRUE), .groups="drop")

common_theme <- theme_minimal() + theme(legend.position = "bottom")

p_top <- ggplot(top_df, aes(year, val, color = country)) + geom_line(size=1) +
  labs(title = "Top 5 Countries", y = "Life Expectancy") + common_theme
```

```
## 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.
```

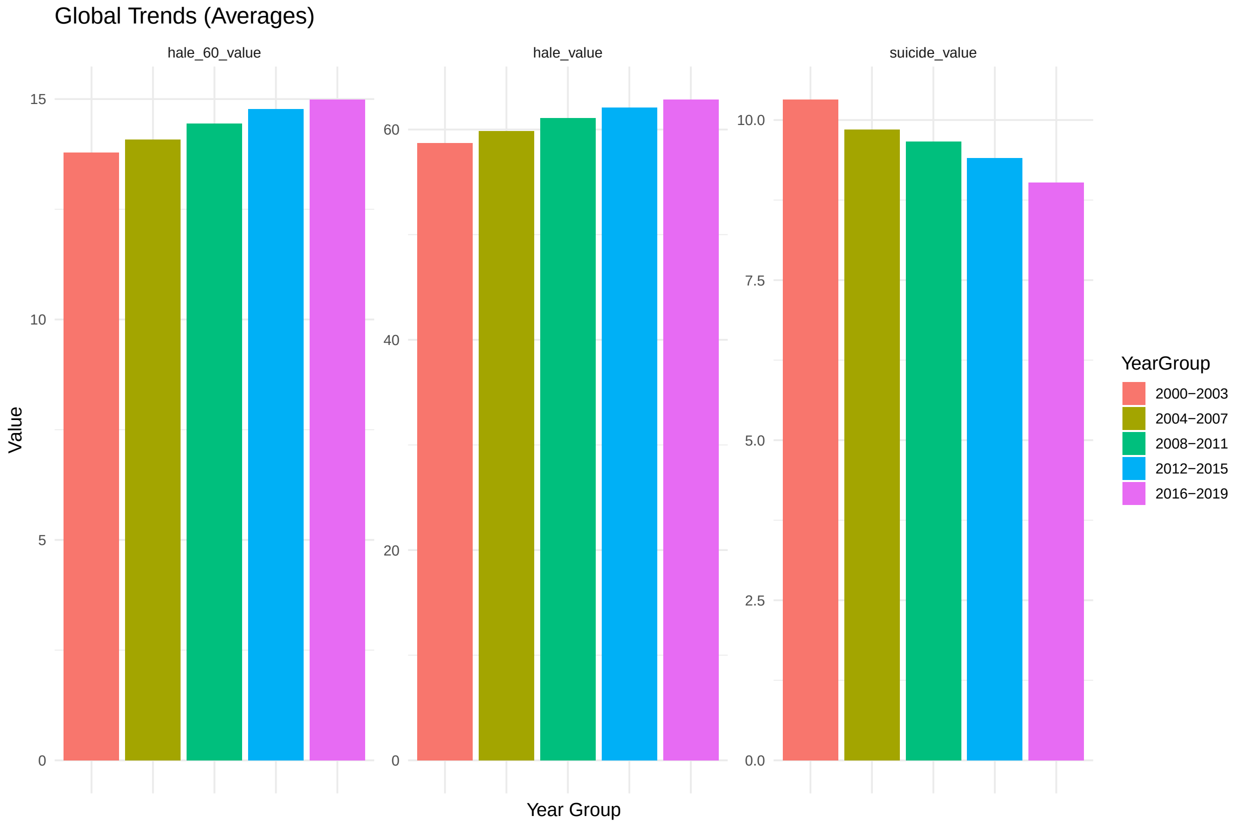
```
p_bottom <- ggplot(bottom_df, aes(year, val, color = country)) + geom_line(size=1) +
  labs(title = "Bottom 5 Countries", y = "Life Expectancy") + common_theme

combined_life <- p_top / p_bottom

#TREEMAP (COUNTS)
count_cols <- c("ncd_deaths_value", "premature_ncd_deaths_count")
df_counts <- merged_imputed %>% filter(year %in% c(2019)) %>%
  summarise(across(all_of(count_cols), ~sum(.x, na.rm = TRUE))) %>%
  pivot_longer(everything(), names_to = "Category", values_to = "Value") %>%
  mutate(Category = gsub("_value|_count", "", Category))

treemap_plot <- ggplot(df_counts, aes(area = Value, fill = Category, label = paste(Category, "\n", comma
  geom_treemap() + geom_treemap_text(color = "white", place = "centre") +
  labs(title = "Global Deaths (2019 Counts)")

# BAR CHART (RATES)
rate_cols <- c("mortality_value", "suicide_value", "homicide_value", "poisoning_value")
df_rates <- merged_imputed %>%
  filter(country %in% c(top5_countries, bottom5_countries)) %>%
  mutate(Group = ifelse(country %in% top5_countries, "Top 5", "Bottom 5")) %>%
  filter(year == 2019) %>%
  group_by(Group) %>%
  summarise(across(all_of(rate_cols), ~mean(.x, na.rm = TRUE))) %>%
  pivot_longer(-Group)

bar_plot <- ggplot(df_rates, aes(x = name, y = value, fill = Group)) +
  geom_col(position = "dodge") +
  facet_wrap(~name, scales="free") +
  theme_minimal() +
  labs(title = "Mortality Rates (2019)", x = "", y = "Rate") +
  theme(axis.text.x = element_blank())

# OUTPUTS!!!

# Stats
metrics
```

```
## # A tibble: 1 x 3
##     MAE  RMSE    R2
##   <dbl> <dbl> <dbl>
## 1  1.68  2.27 0.941
```

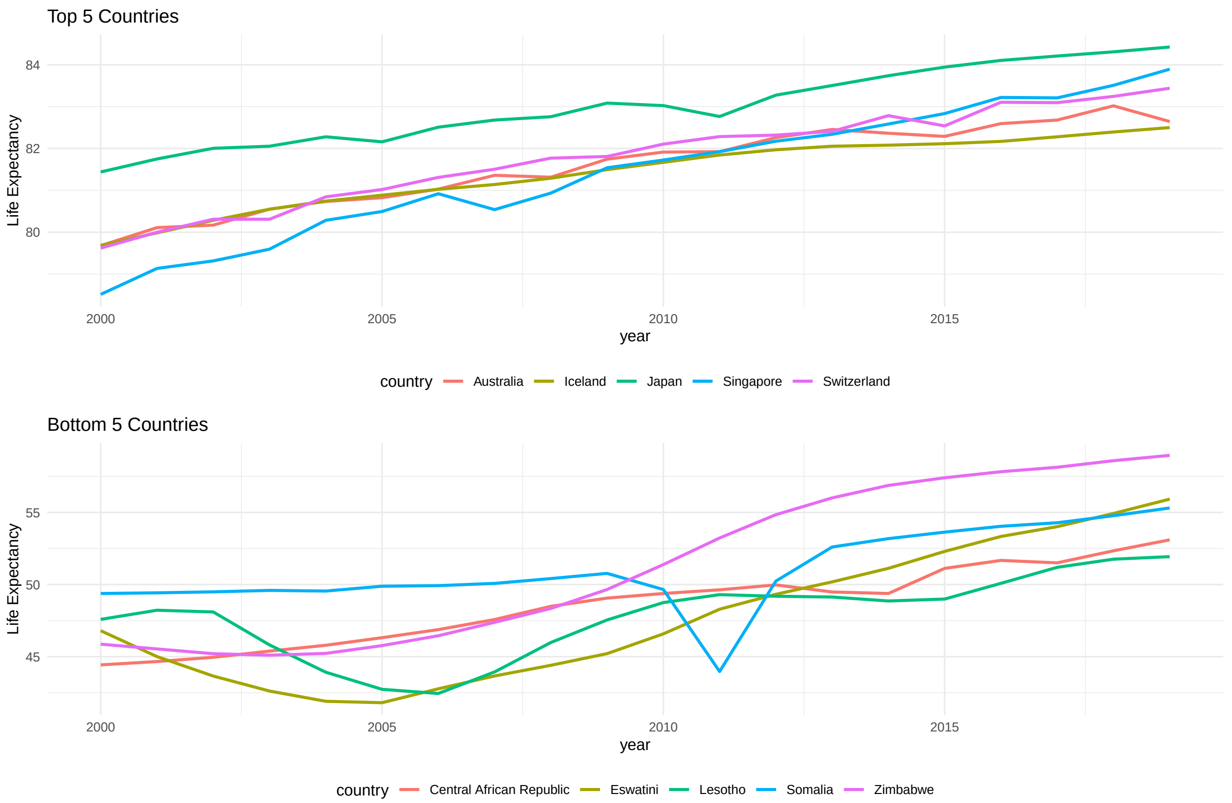
```
tidy(ols) %>% arrange(p.value) %>% head(5)
```

```
## # A tibble: 5 x 5
##   term                estimate std.error statistic   p.value
##   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)           70.0      0.0293    2384.  0
## 2 mortality_value       -7.05     0.0448    -158.  0
## 3 premature_ncd_value   -2.25     0.0426     -52.7 0
## 4 suicide_value          0.940    0.0370      25.4 1.90e-135
## 5 year                   0.465    0.0298      15.6 5.21e- 54
```

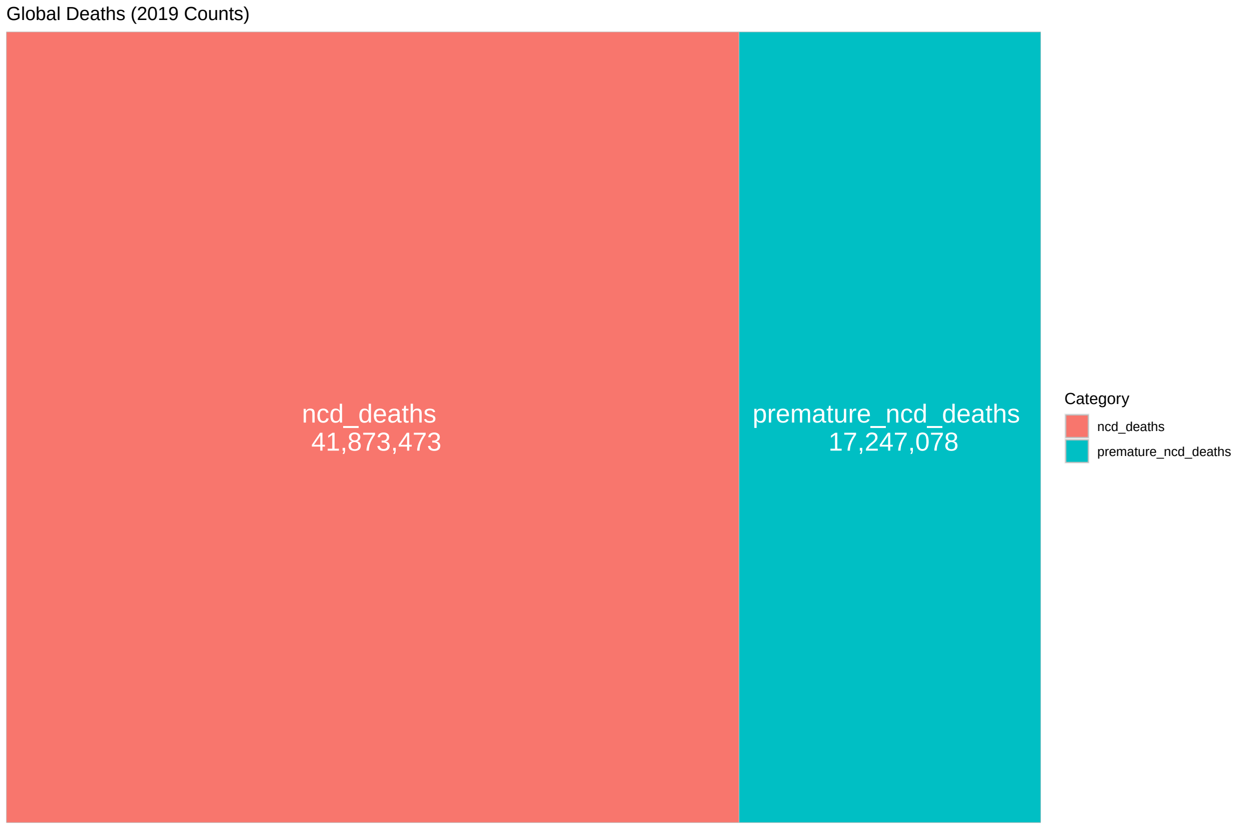
```
#visuals
p1
```



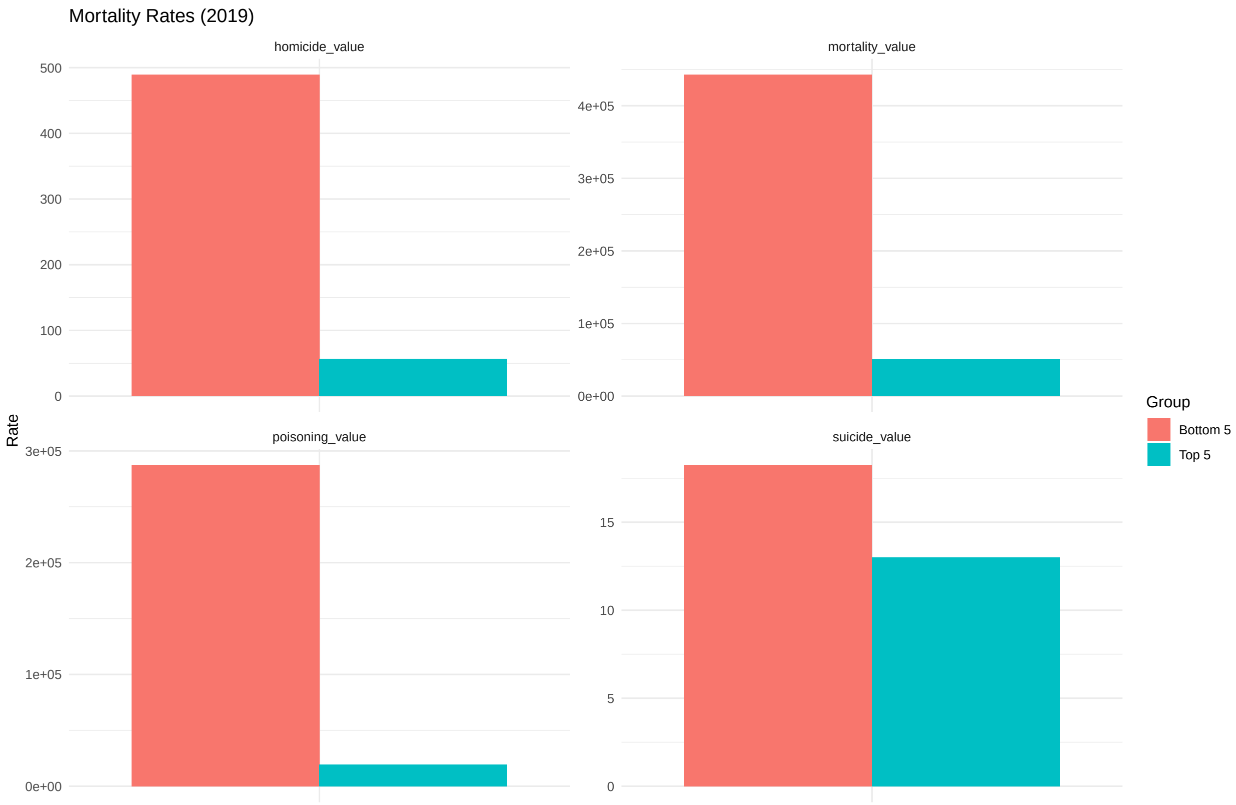
```
combined_life
```



```
treemap_plot
```


Global Deaths (2019 Counts)
ncd_deaths
41,873,473
premature_ncd_deaths
17,247,078
Category
🟥 ncd_deaths
🟦 premature_ncd_deaths

```
bar_plot
```


Mortality Rates (2019)
homicide_value
mortality_value
poisoning_value
suicide_value
Rate
500
400
300
200
100
0
4e+05
3e+05
2e+05
1e+05
0e+00
3e+05
2e+05
1e+05
0e+00
15
10
5
0
Group
Bottom 5
Top 5