Loading libraries

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.4.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
## Warning: package 'lubridate' was built under R version 4.4.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ purrr     1.0.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ── 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 become errors
```

```
library(dplyr)
library(ggplot2)
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(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.4.3
```

```
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 4.4.3
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.4.3
```

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

importing data set

```
# Country population

country_population <- read_csv("C:/Users/RAPHAEL PRO/Desktop/AUCA_COHORT5/R-Programming for data science/archive/country_population.csv")

## Rows: 264 Columns: 61
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): Country Name, Country Code, Indicator Name, Indicator Code
## dbl (57): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.

#  Fertility rate

fertility_rate <- read_csv("C:/Users/RAPHAEL PRO/Desktop/AUCA_COHORT5/R-Programming for data science/archive/fertility_rate.csv")

## Rows: 264 Columns: 61
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): Country Name, Country Code, Indicator Name, Indicator Code
## dbl (57): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, ...
```

```
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
#  Life expectancy

life_expectancy <- read_csv("C:/Users/RAPHAEL PRO/Desktop/AUCA_COHORT5/R-
Programming for data science/archive/life_expectancy.csv")
```

```
## Rows: 264 Columns: 61
## ── Column specification
──────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (4): Country Name, Country Code, Indicator Name, Indicator Code
## dbl (57): 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969,
1970, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
# Life expectancy WHO data

life_expectancy_data <- read_csv("C:/Users/RAPHAEL
PRO/Desktop/AUCA_COHORT5/R-Programming for data
science/archive/life_expectancy_data.csv")
```

```
## Rows: 2938 Columns: 22
## ── Column specification
──────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (2): Country, Status
## dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol,
pe...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO

_____ key Deliverables to achieve:
_____ 1. variable names 2. top 5 rows 3. bottom 10 rows 4. data types 5. shape of data set 6. Check and drop duplicate 7. Find number of missing values 8. use box plot to check if there is outliers in Quantitative variables 9. Handling missing values and outliers if any

1. show the variable names

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO

```
r variable.names(life_expectancy_data)
```

```
##  [1] "Country"                        "Year" ##  [3] "Status"
"Life expectancy" ##  [5] "Adult Mortality"                "infant deaths"
##  [7] "Alcohol"                        "percentage expenditure" ##  [9]
"Hepatitis B"                      "Measles" ## [11] "BMI"
"under-five deaths" ## [13] "Polio"                          "Total
expenditure" ## [15] "Diphtheria"                     "HIV/AIDS" ## [17]
"GDP"                              "Population" ## [19] "thinness  1-19 years"
"thinness 5-9 years" ## [21] "Income composition of resources" "Schooling" #
```

2.top 5 rows

```
r head(life_expectancy_data,5)
```

```
## # A tibble: 5 × 22 ##   Country      Year Status   `Life expectancy`
`Adult Mortality` `infant deaths` ##   <chr>       <dbl> <chr>
<dbl>             <dbl>           <dbl> ## 1 Afghanistan  2015 Develop…
65                 263             62 ## 2 Afghanistan  2014 Develop…
59.9               271             64 ## 3 Afghanistan  2013 Develop…
59.9               268             66 ## 4 Afghanistan  2012 Develop…
59.5               272             69 ## 5 Afghanistan  2011 Develop…
59.2               275             71 ## # ℹ 16 more variables: Alcohol
<dbl>, `percentage expenditure` <dbl>, ## #   `Hepatitis B` <dbl>, Measles
<dbl>, BMI <dbl>, `under-five deaths` <dbl>, ## #   Polio <dbl>, `Total
expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>, ## #   GDP <dbl>,
Population <dbl>, `thinness  1-19 years` <dbl>, ## #   `thinness 5-9 years`
<dbl>, `Income composition of resources` <dbl>, ## #   Schooling <dbl> ##
```

3.bottom 10 rows

```
r tail(life_expectancy_data,10)
```

```
## # A tibble: 10 × 22 ##    Country   Year Status     `Life expectancy`
`Adult Mortality` `infant deaths` ##    <chr>    <dbl> <chr>
<dbl>             <dbl>           <dbl> ##  1 Zimbabwe  2009 Developing
50                 587             30 ##  2 Zimbabwe  2008 Developing
48.2               632             30 ##  3 Zimbabwe  2007 Developing
46.6                67             29 ##  4 Zimbabwe  2006 Developing
45.4                 7             28 ##  5 Zimbabwe  2005 Developing
44.6               717             28 ##  6 Zimbabwe  2004 Developing
44.3               723             27 ##  7 Zimbabwe  2003 Developing
44.5               715             26 ##  8 Zimbabwe  2002 Developing
44.8                73             25 ##  9 Zimbabwe  2001 Developing
45.3               686             25 ## 10 Zimbabwe  2000 Developing
46                 665             24 ## # ℹ 16 more variables: Alcohol
<dbl>, `percentage expenditure` <dbl>, ## #   `Hepatitis B` <dbl>, Measles
<dbl>, BMI <dbl>, `under-five deaths` <dbl>, ## #   Polio <dbl>, `Total
expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>, ## #   GDP <dbl>,
Population <dbl>, `thinness  1-19 years` <dbl>, ## #   `thinness 5-9 years`
<dbl>, `Income composition of resources` <dbl>, ## #   Schooling <dbl> ### 4.
```

data types

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO

```
r str(life_expectancy_data)
```

```
## spc_tbl_ [2,938 × 22] (S3: spec_tbl_df/tbl_df/tbl/data.frame) ##  $
Country                         : chr [1:2938] "Afghanistan" "Afghanistan"
"Afghanistan" "Afghanistan" ... ##  $ Year                               : num
[1:2938] 2015 2014 2013 2012 2011 ... ##  $ Status                                 :
chr [1:2938] "Developing" "Developing" "Developing" "Developing" ... ##  $
Life expectancy                 : num [1:2938] 65 59.9 59.9 59.5 59.2 58.8
58.6 58.1 57.5 57.3 ... ##  $ Adult Mortality                : num [1:2938]
263 271 268 272 275 279 281 287 295 295 ... ##  $ infant deaths
: num [1:2938] 62 64 66 69 71 74 77 80 82 84 ... ##  $ Alcohol
: num [1:2938] 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ... ##  $
percentage expenditure          : num [1:2938] 71.3 73.5 73.2 78.2 7.1 ... ##
$ Hepatitis B                     : num [1:2938] 65 62 64 67 68 66 63 64 63 64
... ##  $ Measles                           : num [1:2938] 1154 492 430 2787
3013 ... ##  $ BMI                                : num [1:2938] 19.1 18.6 18.1
17.6 17.2 16.7 16.2 15.7 15.2 14.7 ... ##  $ under-five deaths               :
num [1:2938] 83 86 89 93 97 102 106 110 113 116 ... ##  $ Polio
: num [1:2938] 6 58 62 67 68 66 63 64 63 58 ... ##  $ Total expenditure
: num [1:2938] 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ... ##  $
Diphtheria                      : num [1:2938] 65 62 64 67 68 66 63 64 63 58
... ##  $ HIV/AIDS                          : num [1:2938] 0.1 0.1 0.1 0.1 0.1
0.1 0.1 0.1 0.1 0.1 ... ##  $ GDP                                : num [1:2938]
584.3 612.7 631.7 670 63.5 ... ##  $ Population                         : num
[1:2938] 33736494 327582 31731688 3696958 2978599 ... ##  $ thinness  1-19
years           : num [1:2938] 17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19
19.2 ... ##  $ thinness 5-9 years             : num [1:2938] 17.3 17.5 17.7
18 18.2 18.4 18.7 18.9 19.1 19.3 ... ##  $ Income composition of resources:
num [1:2938] 0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num [1:2938] 10.1 10 9.9 9.8 9.5 9.2
8.9 8.7 8.4 8.1 ... ##  - attr(*, "spec")= ##   .. cols( ##   ..   Country =
col_character(), ##   ..   Year = col_double(), ##   ..   Status =
col_character(), ##   ..   `Life expectancy` = col_double(), ##   ..   `Adult
Mortality` = col_double(), ##   ..   `infant deaths` = col_double(), ##   ..
Alcohol = col_double(), ##   ..   `percentage expenditure` = col_double(), ##
..   `Hepatitis B` = col_double(), ##   ..   Measles = col_double(), ##   ..
BMI = col_double(), ##   ..   `under-five deaths` = col_double(), ##   ..
Polio = col_double(), ##   ..   `Total expenditure` = col_double(), ##   ..
Diphtheria = col_double(), ##   ..   `HIV/AIDS` = col_double(), ##   ..   GDP
= col_double(), ##   ..   Population = col_double(), ##   ..   `thinness  1-
19 years` = col_double(), ##   ..   `thinness 5-9 years` = col_double(), ##
..   `Income composition of resources` = col_double(), ##   ..   Schooling =
col_double() ##   .. ) ##  - attr(*, "problems")=<externalptr> ### 5. shape of
data set
```

```
r dim(life_expectancy_data)
```

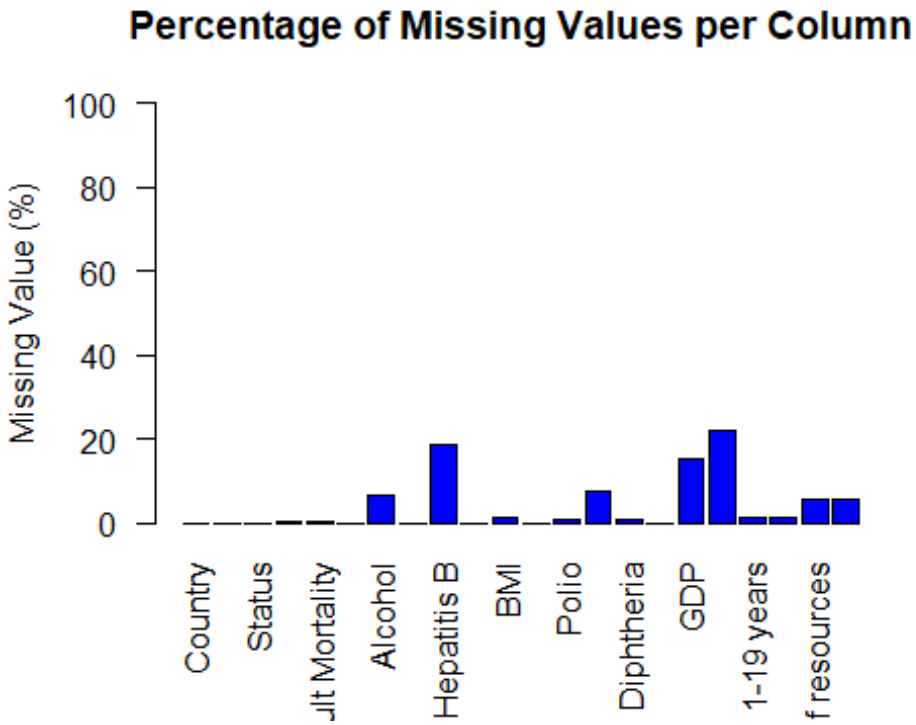
```
## [1] 2938    22 ### 6. Checking for duplicates/ rows
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO

```
r sum(duplicated(life_expectancy_data))
## [1] 0 ### 7. Find number of missing values
r sum(is.na(life_expectancy_data))
## [1] 2563
r colSums(is.na(life_expectancy_data))
##                          Country                             Year ##
0                                0 ##                           Status
Life expectancy ##                                  0
10 ##                 Adult Mortality                    infant deaths ##
10                                0 ##                          Alcohol
percentage expenditure ##                                 194
0 ##                     Hepatitis B                          Measles ##
553                               0 ##                              BMI
under-five deaths ##                                 34
0 ##                            Polio                Total expenditure ##
19                              226 ##                         Diphtheria
HIV/AIDS ##                                  19                                  0
##                              GDP                       Population ##
448                              652 ##                thinness  1-19 years
thinness 5-9 years ##                                34
34 ## Income composition of resources                           Schooling ##
167                              163
```

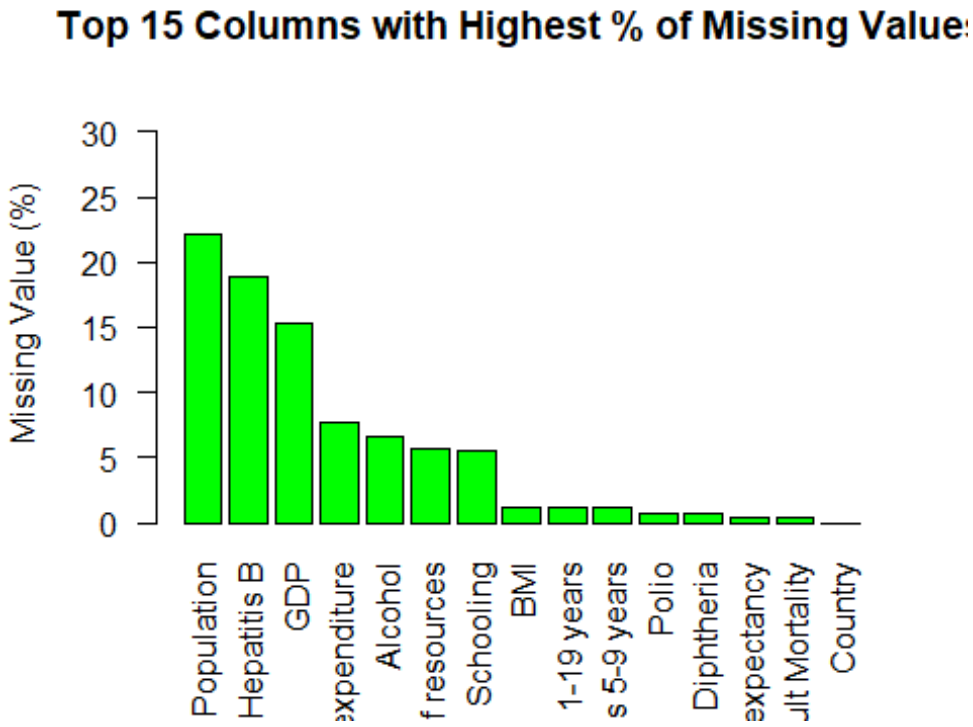
``` r ## calculate percentage of missing

missing_percent <- colSums(is.na(life_expectancy_data)) / nrow(life_expectancy_data) * 100

## create Bar plot

barplot(missing_percent, main = “Percentage of Missing Values per Column”, ylab = “Missing Value (%)”, col = “blue”, ylim = c(0, 100), las = 2) ```
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO



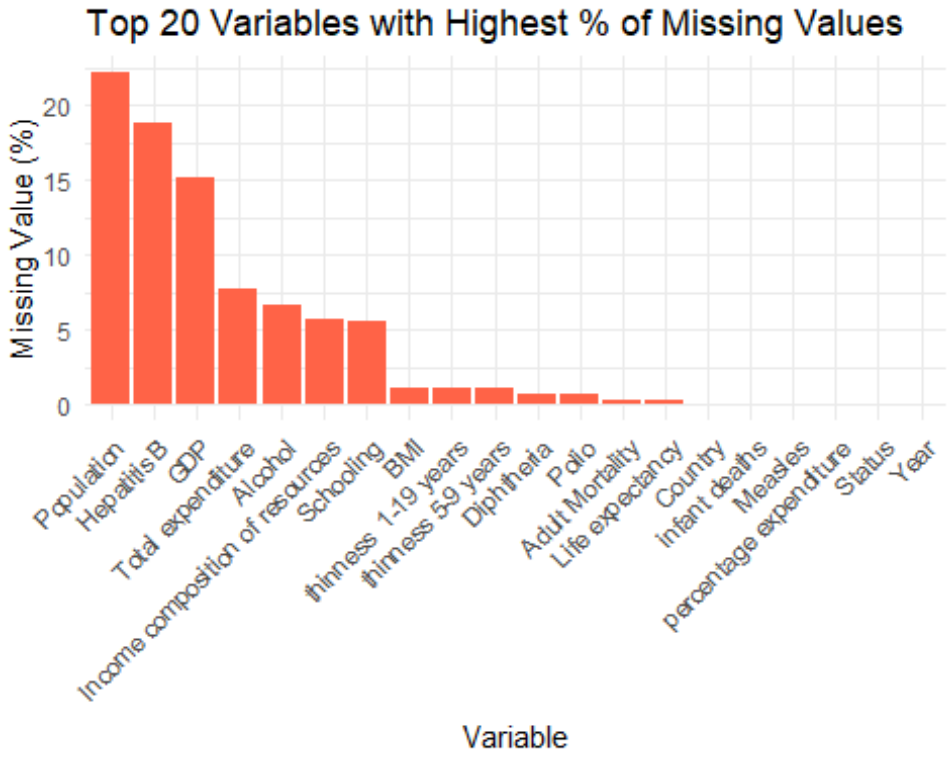

``` r ## plot for the first 15 cols

top15_missing <- sort(missing_percent, decreasing = TRUE)[1:15]

## plot barplot(top15_missing, main = “Top 15 Columns with Highest % of Missing Values”, ylab = “Missing Value (%)”, col = “green”, las = 2, ylim = c(0, max(top15_missing, na.rm = TRUE) + 10)) ```
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO




``` r ## visualizing missing using ggplot2

# Step 1: Calculate % missing for each column

missing_percent <- life_expectancy_data %>% summarise(across(everything(), ~mean(is.na(.)) * 100)) %>% pivot_longer(everything(), names_to = “Variable”, values_to = “MissingPercent”) %>% arrange(desc(MissingPercent)) %>% slice_head(n = 20) # Top 20

# Step 2: Plot ggplot(missing_percent, aes(x = reorder(Variable, -MissingPercent), y = MissingPercent)) + geom_bar(stat = “identity”, fill = “tomato”) + labs(title = “Top 20 Variables with Highest % of Missing Values”, x = “Variable”, y = “Missing Value (%)”) + theme_minimal() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) ```
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO



8. use box plot to check if there is outliers in Quantitative variables/continous features

``` r # Select only numeric (continuous) columns

continuous_vars <- life_expectancy_data %>% select(where(is.numeric)) ```

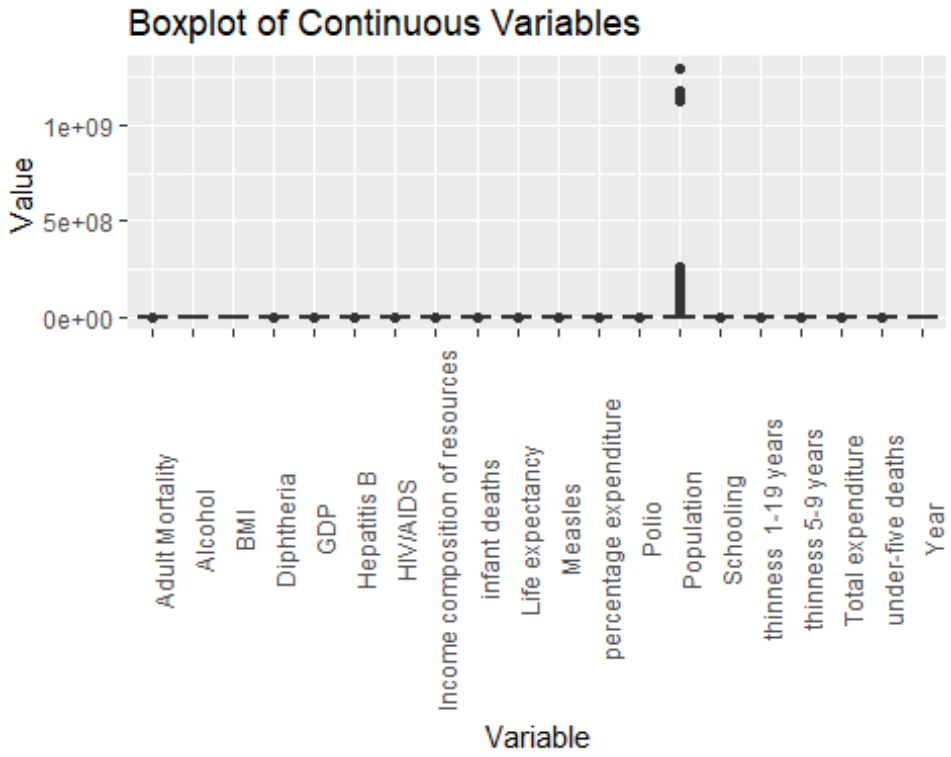
``` r # Reshape to long format

long_data <- pivot_longer(continuous_vars, cols = everything(), names_to = “Variable”, values_to = “Value”)

# Plot ggplot(long_data, aes(x = Variable, y = Value)) + geom_boxplot(fill = “lightblue”) + theme(axis.text.x = element_text(angle = 90)) + labs(title = “Boxplot of Continuous Variables”) ```

```
Warning: Removed 2563 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO




``` r # preserve data for future use

life_expectancy_data1<-life_expectancy_data

life_expectancy_data1 ```

```
A tibble: 2,938 × 22 ##   Country   Year Status `Life expectancy`
`Adult Mortality` `infant deaths` ##   <chr>   <dbl> <chr>
<dbl>   <dbl>   <dbl> ## 1 Afghanistan 2015 Develo…
65   263   62 ## 2 Afghanistan 2014 Develo…
59.9   271   64 ## 3 Afghanistan 2013 Develo…
59.9   268   66 ## 4 Afghanistan 2012 Develo…
59.5   272   69 ## 5 Afghanistan 2011 Develo…
59.2   275   71 ## 6 Afghanistan 2010 Develo…
58.8   279   74 ## 7 Afghanistan 2009 Develo…
58.6   281   77 ## 8 Afghanistan 2008 Develo…
58.1   287   80 ## 9 Afghanistan 2007 Develo…
57.5   295   82 ## 10 Afghanistan 2006 Develo…
57.3   295   84 ## # ℹ 2,928 more rows ## # ℹ 16 more
variables: Alcohol <dbl>, `percentage expenditure` <dbl>, ## #   `Hepatitis
B` <dbl>, Measles <dbl>, BMI <dbl>, `under-five deaths` <dbl>, ## #   Polio
<dbl>, `Total expenditure` <dbl>, Diphtheria <dbl>, `HIV/AIDS` <dbl>, ## #
GDP <dbl>, Population <dbl>, `thinness  1-19 years` <dbl>, ## #   `thinness
5-9 years` <dbl>, `Income composition of resources` <dbl>, ## #   Schooling
<dbl>
```
```

4.2 Exploratory data analysis using life expectancy data set or life expectancy data set as provided by WHO

---

9. Handling missing values and outliers if any

9.1. Handling missing values

9.2 Handling outliers

```
r ## we managed to handle missing values using mean fill
life_expectancy_data1 <- life_expectancy_data %>%
mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE),
.)))
```

Verify if missing were managed

```
r sum(is.na(life_expectancy_data1))   # In general
## [1] 0
```

Handling outliers using Robust Methods / This is the best way

step 1: Identifying Outliers Using Median Absolute Deviation (MAD)

``` r # Select only numeric columns

numeric_data <- life_expectancy_data1 %>% select_if(is.numeric)

# Create a copy to store cleaned data cleaned_data <- numeric_data

# Replace outliers (based on 3 MAD) with median

for (col in names(numeric_data)) { column <- numeric_data[[col]] med <- median(column, na.rm = TRUE) mad_val <- mad(column, constant = 1, na.rm = TRUE)

# Logical index of outliers

is_outlier <- abs(column - med) > 3 * mad_val

# Replace outliers with median

cleaned_data[[col]][is_outlier] <- med }

# Now cleaned_data contains numeric data with outliers replaced

# update the original dataset:

life_expectancy_data1[names(cleaned_data)] <- cleaned_data ```

# For the sake of analysis using life expectancy data set we shall use life_expectancy_data1 handled with robust model

---
```

4.3 Value extraction and plot , we will use country_population data set

1.Top 10 most populous counties and their population number during 2016

```
str(country_population)
```

```
## spc_tbl_ [264 × 61] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Country Name  : chr [1:264] "Aruba" "Afghanistan" "Angola" "Albania"
...
##  $ Country Code  : chr [1:264] "ABW" "AFG" "AGO" "ALB" ...
##  $ Indicator Name: chr [1:264] "Population, total" "Population, total"
"Population, total" "Population, total" ...
##  $ Indicator Code: chr [1:264] "SP.POP.TOTL" "SP.POP.TOTL" "SP.POP.TOTL"
"SP.POP.TOTL" ...
##  $ 1960          : num [1:264] 54211 8996351 5643182 1608800 13411 ...
##  $ 1961          : num [1:264] 55438 9166764 5753024 1659800 14375 ...
##  $ 1962          : num [1:264] 56225 9345868 5866061 1711319 15370 ...
##  $ 1963          : num [1:264] 56695 9533954 5980417 1762621 16412 ...
##  $ 1964          : num [1:264] 57032 9731361 6093321 1814135 17469 ...
##  $ 1965          : num [1:264] 57360 9938414 6203299 1864791 18549 ...
##  $ 1966          : num [1:264] 57715 10152331 6309770 1914573 19647 ...
##  $ 1967          : num [1:264] 58055 10372630 6414995 1965598 20758 ...
##  $ 1968          : num [1:264] 58386 10604346 6523791 2022272 21890 ...
##  $ 1969          : num [1:264] 58726 10854428 6642632 2081695 23058 ...
##  $ 1970          : num [1:264] 59063 11126123 6776381 2135479 24276 ...
##  $ 1971          : num [1:264] 59440 11417825 6927269 2187853 25559 ...
##  $ 1972          : num [1:264] 59840 11721940 7094834 2243126 26892 ...
##  $ 1973          : num [1:264] 60243 12027822 7277960 2296752 28232 ...
##  $ 1974          : num [1:264] 60528 12321541 7474338 2350124 29520 ...
##  $ 1975          : num [1:264] 60657 12590286 7682479 2404831 30705 ...
##  $ 1976          : num [1:264] 60586 12840299 7900997 2458526 31777 ...
##  $ 1977          : num [1:264] 60366 13067538 8130988 2513546 32771 ...
##  $ 1978          : num [1:264] 60103 13237734 8376147 2566266 33737 ...
##  $ 1979          : num [1:264] 59980 13306695 8641521 2617832 34818 ...
##  $ 1980          : num [1:264] 60096 13248370 8929900 2671997 36067 ...
##  $ 1981          : num [1:264] 60567 13053954 9244507 2726056 37500 ...
##  $ 1982          : num [1:264] 61345 12749645 9582156 2784278 39114 ...
##  $ 1983          : num [1:264] 62201 12389269 9931562 2843960 40867 ...
##  $ 1984          : num [1:264] 62836 12047115 10277321 2904429 42706 ...
##  $ 1985          : num [1:264] 63026 11783050 10609042 2964762 44600 ...
##  $ 1986          : num [1:264] 62644 11601041 10921037 3022635 46517 ...
##  $ 1987          : num [1:264] 61833 11502761 11218268 3083605 48455 ...
##  $ 1988          : num [1:264] 61079 11540888 11513968 3142336 50434 ...
##  $ 1989          : num [1:264] 61032 11777609 11827237 3227943 52448 ...
##  $ 1990          : num [1:264] 62149 12249114 12171441 3286542 54509 ...
##  $ 1991          : num [1:264] 64622 12993657 12553446 3266790 56671 ...
##  $ 1992          : num [1:264] 68235 13981231 12968345 3247039 58888 ...
```

```
##  $ 1993          : num [1:264] 72504 15095099 13403734 3227287 60971 ...
##  $ 1994          : num [1:264] 76700 16172719 13841301 3207536 62677 ...
##  $ 1995          : num [1:264] 80324 17099541 14268994 3187784 63850 ...
##  $ 1996          : num [1:264] 83200 17822884 14682284 3168033 64360 ...
##  $ 1997          : num [1:264] 85451 18381605 15088981 3148281 64327 ...
##  $ 1998          : num [1:264] 87277 18863999 15504318 3128530 64142 ...
##  $ 1999          : num [1:264] 89005 19403676 15949766 3108778 64370 ...
##  $ 2000          : num [1:264] 90853 20093756 16440924 3089027 65390 ...
##  $ 2001          : num [1:264] 92898 20966463 16983266 3060173 67341 ...
##  $ 2002          : num [1:264] 94992 21979923 17572649 3051010 70049 ...
##  $ 2003          : num [1:264] 97017 23064851 18203369 3039616 73182 ...
##  $ 2004          : num [1:264] 98737 24118979 18865716 3026939 76244 ...
##  $ 2005          : num [1:264] 100031 25070798 19552542 3011487 78867 ...
##  $ 2006          : num [1:264] 100832 25893450 20262399 2992547 80991 ...
##  $ 2007          : num [1:264] 101220 26616792 20997687 2970017 82683 ...
##  $ 2008          : num [1:264] 101353 27294031 21759420 2947314 83861 ...
##  $ 2009          : num [1:264] 101453 28004331 22549547 2927519 84462 ...
##  $ 2010          : num [1:264] 101669 28803167 23369131 2913021 84449 ...
##  $ 2011          : num [1:264] 102053 29708599 24218565 2905195 83751 ...
##  $ 2012          : num [1:264] 102577 30696958 25096150 2900401 82431 ...
##  $ 2013          : num [1:264] 103187 31731688 25998340 2895092 80788 ...
##  $ 2014          : num [1:264] 103795 32758020 26920466 2889104 79223 ...
##  $ 2015          : num [1:264] 104341 33736494 27859305 2880703 78014 ...
##  $ 2016          : num [1:264] 104822 34656032 28813463 2876101 77281 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Country Name` = col_character(),
##   ..   `Country Code` = col_character(),
##   ..   `Indicator Name` = col_character(),
##   ..   `Indicator Code` = col_character(),
##   ..   `1960` = col_double(),
##   ..   `1961` = col_double(),
##   ..   `1962` = col_double(),
##   ..   `1963` = col_double(),
##   ..   `1964` = col_double(),
##   ..   `1965` = col_double(),
##   ..   `1966` = col_double(),
##   ..   `1967` = col_double(),
##   ..   `1968` = col_double(),
##   ..   `1969` = col_double(),
##   ..   `1970` = col_double(),
##   ..   `1971` = col_double(),
##   ..   `1972` = col_double(),
##   ..   `1973` = col_double(),
##   ..   `1974` = col_double(),
##   ..   `1975` = col_double(),
##   ..   `1976` = col_double(),
##   ..   `1977` = col_double(),
##   ..   `1978` = col_double(),
##   ..   `1979` = col_double(),
```

```
##   ..   `1980` = col_double(),
##   ..   `1981` = col_double(),
##   ..   `1982` = col_double(),
##   ..   `1983` = col_double(),
##   ..   `1984` = col_double(),
##   ..   `1985` = col_double(),
##   ..   `1986` = col_double(),
##   ..   `1987` = col_double(),
##   ..   `1988` = col_double(),
##   ..   `1989` = col_double(),
##   ..   `1990` = col_double(),
##   ..   `1991` = col_double(),
##   ..   `1992` = col_double(),
##   ..   `1993` = col_double(),
##   ..   `1994` = col_double(),
##   ..   `1995` = col_double(),
##   ..   `1996` = col_double(),
##   ..   `1997` = col_double(),
##   ..   `1998` = col_double(),
##   ..   `1999` = col_double(),
##   ..   `2000` = col_double(),
##   ..   `2001` = col_double(),
##   ..   `2002` = col_double(),
##   ..   `2003` = col_double(),
##   ..   `2004` = col_double(),
##   ..   `2005` = col_double(),
##   ..   `2006` = col_double(),
##   ..   `2007` = col_double(),
##   ..   `2008` = col_double(),
##   ..   `2009` = col_double(),
##   ..   `2010` = col_double(),
##   ..   `2011` = col_double(),
##   ..   `2012` = col_double(),
##   ..   `2013` = col_double(),
##   ..   `2014` = col_double(),
##   ..   `2015` = col_double(),
##   ..   `2016` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

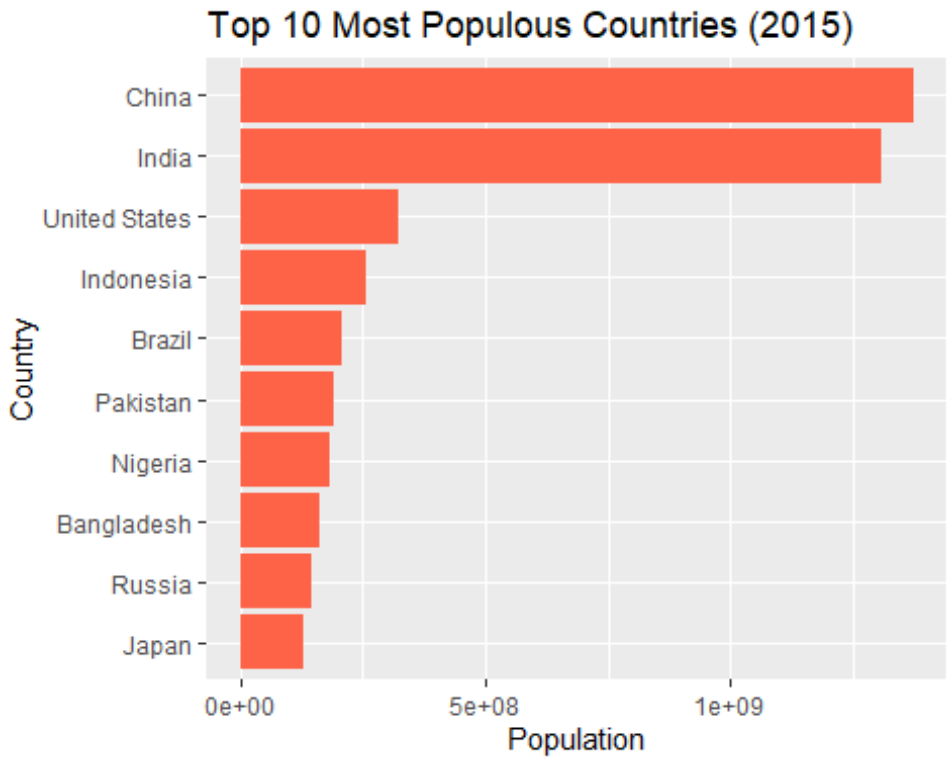
create a real country list ,since some country are not real country list

```
country_population$country <- countrycode(country_population$`Country Code`,
                                          origin = "iso3c",
                                          destination = "country.name")

## Warning: Some values were not matched unambiguously: ARB, CEB, CHI, CSS,
EAP, EAR, EAS, ECA, ECS, EMU, EUU, FCS, HIC, HPC, IBD, IBT, IDA, IDB, IDX,
```

```
INX, LAC, LCN, LDC, LIC, LMC, LMY, LTE, MEA, MIC, MNA, NAC, OED, OSS, PRE,
PSS, PST, SAS, SSA, SSF, SST, TEA, TEC, TLA, TMN, TSA, TSS, UMC, WLD, XKX
```

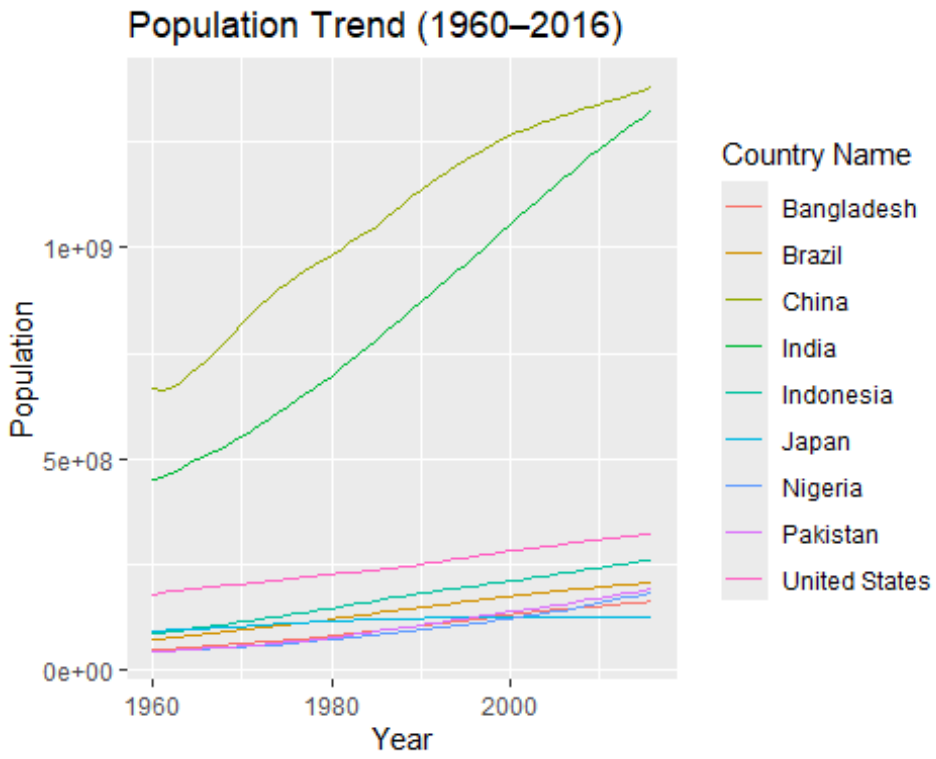
1. Use an appropriate graph to present top 10 most populous counties and their population number during 2015.

```
country_population_clean <- country_population %>%        # remove rows with
missing by country column
  drop_na(country)

top10_pop <- country_population_clean %>%
  select(country, `2015`) %>%
  arrange(desc(`2015`)) %>%
  slice(1:10)

ggplot(top10_pop, aes(x = reorder(country, `2015`), y = `2015`)) +
  geom_bar(stat = "identity", fill = "tomato") +
  coord_flip() +
  labs(title = "Top 10 Most Populous Countries (2015)", x = "Country", y =
"Population")
```



2.Trend in their population number since 1960-2016 by using appro priate graph i.e among top 10

```
# Reshape wide to long
pop_long <- country_population_clean %>%
  filter(`Country Name` %in% top10_pop$country) %>%
  pivot_longer(cols = `1960`:`2016`, names_to = "Year", values_to =
"Population") %>%
  mutate(Year = as.integer(Year))

# Line plot
ggplot(pop_long, aes(x = Year, y = Population, color = `Country Name`)) +
  geom_line() +
  labs(title = "Population Trend (1960–2016)", x = "Year", y = "Population")
```



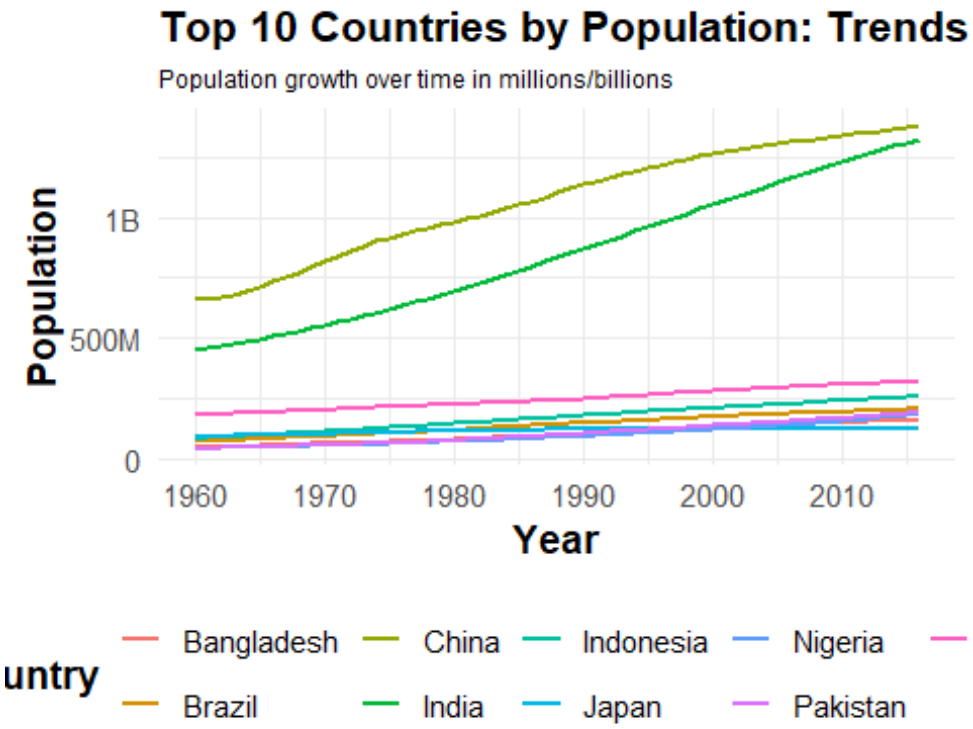
ALternative plot

for more Intuition

```
# Prepare data
pop_long <- country_population_clean %>%
  filter(`Country Name` %in% top10_pop$country) %>%
  pivot_longer(cols = `1960`:`2016`, names_to = "Year", values_to =
"Population") %>%
  mutate(Year = as.integer(Year))

# Improved Line Plot with updated scale formatting
ggplot(pop_long, aes(x = Year, y = Population, color = `Country Name`)) +
```

```
  geom_line(size = 1) +
  scale_y_continuous(
    labels = label_number(scale_cut = cut_short_scale())
  ) +
  scale_x_continuous(breaks = seq(1960, 2016, by = 10)) +
  labs(
    title = "Top 10 Countries by Population: Trends from 1960 to 2016",
    subtitle = "Population growth over time in millions/billions",
    x = "Year",
    y = "Population",
    color = "Country"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 9),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "bottom"
  )
```

```
## 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.
```



3. Use fertility_rate dataset to extract fertility_rate for the most populous countries. Use an appropriate graph to show their trend since 1960-2016

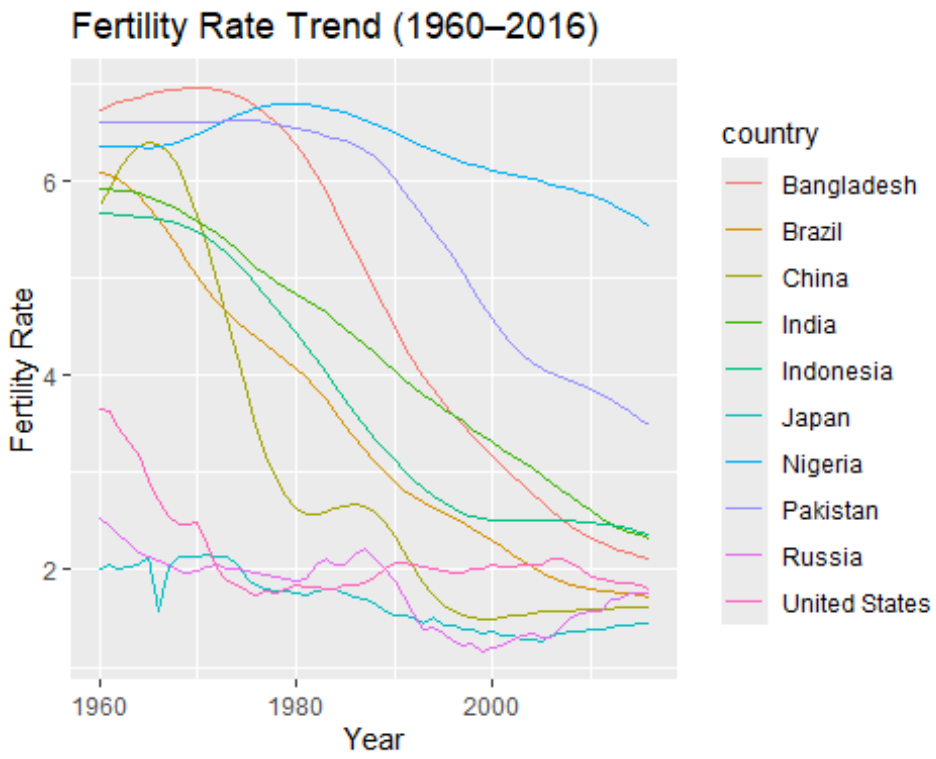
converting fertility rate data from wide to long data set

```
fertility_long <- fertility_rate %>%
  pivot_longer(
    cols = `1960`:`2016`,           # All year columns
    names_to = "Year",              # New column to store years
    values_to = "FertilityRate"     # New column to store fertility rate
values
  ) %>%
  mutate(Year = as.integer(Year))   # Convert Year column from character to
integer
```

Extracting Country using country code converter

create a real country list ,since some country are not real country list

```
fertility_long$country <- countrycode(fertility_long$`Country Code`,
                                      origin = "iso3c",
                                      destination = "country.name")

## Warning: Some values were not matched unambiguously: ARB, CEB, CHI, CSS,
EAP, EAR, EAS, ECA, ECS, EMU, EUU, FCS, HIC, HPC, IBD, IBT, IDA, IDB, IDX,
INX, LAC, LCN, LDC, LIC, LMC, LMY, LTE, MEA, MIC, MNA, NAC, OED, OSS, PRE,
PSS, PST, SAS, SSA, SSF, SST, TEA, TEC, TLA, TMN, TSA, TSS, UMC, WLD, XKX

fertility_long_filtered <- fertility_long %>%
  mutate(Year = as.integer(Year)) %>%
  filter(country %in% top10_pop$country & Year >= 1960 & Year <= 2016)

ggplot(fertility_long_filtered, aes(x = Year, y = FertilityRate, color =
country)) +
  geom_line() +
  labs(title = "Fertility Rate Trend (1960-2016)", x = "Year", y = "Fertility
Rate")
```



4. Use life_expectancy dataset to extract life_expectancy for the most populous countries. Use an appropriate graph to show their trend since 1960-2016

converting fertility rate data from wide to long data set

```
lifeexpectancy_long <- life_expectancy %>%
  pivot_longer(
    cols = `1960`:`2016`,           # All year columns
    names_to = "Year",              # New column to store years
    values_to = "FertilityRate"     # New column to store fertility rate
values
  ) %>%
  mutate(Year = as.integer(Year))   # Convert Year column from character to
integer
```

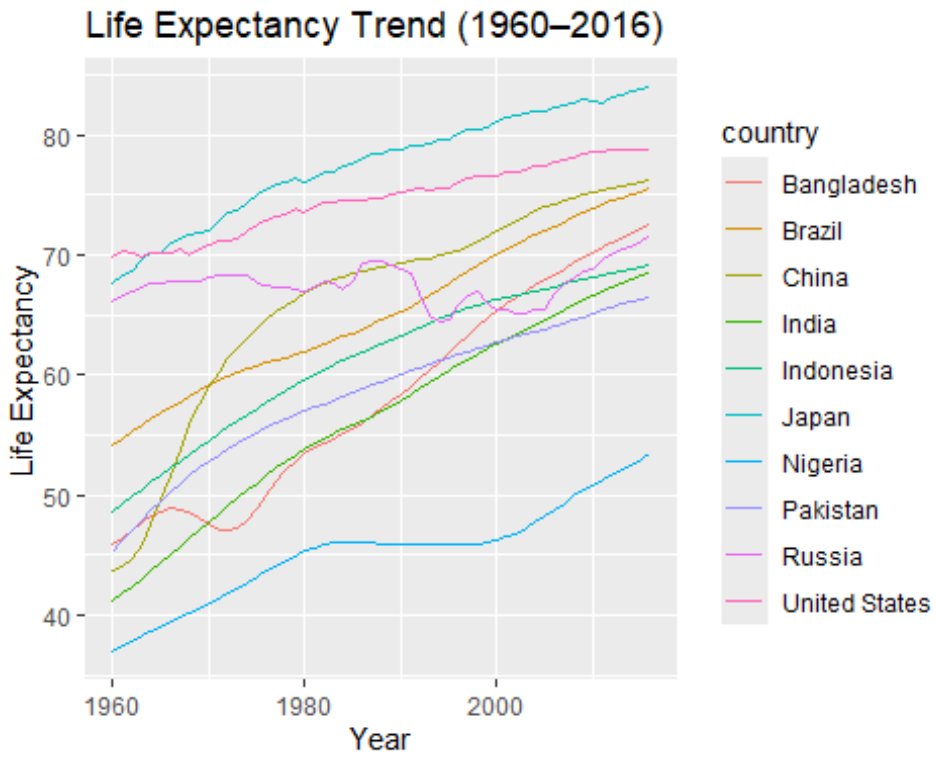
create a real country list ,since some country are not real country list

```
lifeexpectancy_long$country <- countrycode(lifeexpectancy_long$`Country
Code`,
                                           origin = "iso3c",
                                           destination = "country.name")
```

```
## Warning: Some values were not matched unambiguously: ARB, CEB, CHI, CSS,
EAP, EAR, EAS, ECA, ECS, EMU, EUU, FCS, HIC, HPC, IBD, IBT, IDA, IDB, IDX,
INX, LAC, LCN, LDC, LIC, LMC, LMY, LTE, MEA, MIC, MNA, NAC, OED, OSS, PRE,
PSS, PST, SAS, SSA, SSF, SST, TEA, TEC, TLA, TMN, TSA, TSS, UMC, WLD, XKX
```

Rename FertilityRate

```
lifeexpectancy_long <- lifeexpectancy_long %>%
  rename(LifeExpectancy = FertilityRate)

life_expectancy_filtered <- lifeexpectancy_long %>%
  filter(country %in% top10_pop$country & Year >= 1960 & Year <= 2016)

ggplot(life_expectancy_filtered, aes(x = Year, y = LifeExpectancy, color =
country)) +
  geom_line() +
  labs(title = "Life Expectancy Trend (1960-2016)", x = "Year", y = "Life
Expectancy")
```



4.4. Correlation Analysis using life_expectancy_who

#1.Use the life expectancy data dataset to find correlations between Life expectancy , Adult Mortality, infant deaths, Alcohol, percentage expenditure, Hepatitis B, Measles , BMI , under-five deaths , Polio, Total expenditure, Diphtheria , HIV/AIDS, GDP, Population , thinness 1-19 years, thinness 5-9 years, Income composition of resources and Schooling

by using both numerical values and heatmap. Interpret the relationship between life expectancy and schooling

```
str(life_expectancy_data1)
```

```
## tibble [2,938 × 22] (S3: tbl_df/tbl/data.frame)
##  $ Country                        : chr [1:2938] "Afghanistan"
"Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : num [1:2938] 2015 2014 2013 2012 2011
...
##  $ Status                         : chr [1:2938] "Developing" "Developing"
"Developing" "Developing" ...
##  $ Life expectancy                : num [1:2938] 65 59.9 59.9 59.5 59.2
58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult Mortality                : num [1:2938] 263 271 268 272 275 279
281 287 295 295 ...
##  $ infant deaths                  : num [1:2938] 3 3 3 3 3 3 3 3 3 3 ...
##  $ Alcohol                        : num [1:2938] 0.01 0.01 0.01 0.01 0.01
0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage expenditure         : num [1:2938] 71.3 73.5 73.2 78.2 7.1
...
##  $ Hepatitis B                    : num [1:2938] 65 87 64 67 68 66 63 64
63 64 ...
##  $ Measles                        : num [1:2938] 17 17 17 17 17 17 17 17
17 17 ...
##  $ BMI                            : num [1:2938] 19.1 18.6 18.1 17.6 17.2
16.7 16.2 15.7 15.2 14.7 ...
##  $ under-five deaths              : num [1:2938] 4 4 4 4 4 4 4 4 4 4 ...
##  $ Polio                          : num [1:2938] 93 93 93 93 93 93 93 93
93 93 ...
##  $ Total expenditure              : num [1:2938] 8.16 8.18 8.13 8.52 7.87
9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : num [1:2938] 93 93 93 93 93 93 93 93
93 93 ...
##  $ HIV/AIDS                       : num [1:2938] 0.1 0.1 0.1 0.1 0.1 0.1
0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num [1:2938] 584.3 612.7 631.7 670
63.5 ...
##  $ Population                     : num [1:2938] 3675929 327582 3675929
3696958 2978599 ...
##  $ thinness  1-19 years           : num [1:2938] 3.4 3.4 3.4 3.4 3.4 3.4
3.4 3.4 3.4 3.4 ...
##  $ thinness 5-9 years             : num [1:2938] 3.4 3.4 3.4 3.4 3.4 3.4
3.4 3.4 3.4 3.4 ...
##  $ Income composition of resources: num [1:2938] 0.479 0.476 0.47 0.463
0.454 0.448 0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num [1:2938] 10.1 10 9.9 9.8 9.5 9.2
8.9 8.7 8.4 8.1 ...
```

EDA for new data set / life_expectancy_who

Checking for duplicates/ rows

```
sum(duplicated(life_expectancy_data1))
```

```
## [1] 0
```

Find number of missing values

```
sum(is.na(life_expectancy_data1))   # In general
```

```
## [1] 0
```

```
colSums(is.na(life_expectancy_data1)) # per column
```

```
##                          Country                             Year
##                                0                                0
##                           Status                  Life expectancy
##                                0                                0
##                  Adult Mortality                    infant deaths
##                                0                                0
##                          Alcohol          percentage expenditure
##                                0                                0
##                      Hepatitis B                          Measles
##                                0                                0
##                              BMI                under-five deaths
##                                0                                0
##                            Polio                Total expenditure
##                                0                                0
##                       Diphtheria                          HIV/AIDS
##                                0                                0
##                              GDP                       Population
##                                0                                0
##            thinness  1-19 years              thinness 5-9 years
##                                0                                0
## Income composition of resources                         Schooling
##                                0                                0
```

Normalization of variable names

```
# pipe the raw dataset through the function clean_names()
life_expectancy_data1 <- life_expectancy_data1 %>%
  janitor::clean_names()

# see the new column names
names(life_expectancy_data1)
```

```
##  [1] "country"                        "year"
##  [3] "status"                         "life_expectancy"
##  [5] "adult_mortality"                "infant_deaths"
##  [7] "alcohol"                        "percentage_expenditure"
```

```
##  [9] "hepatitis_b"                       "measles"
## [11] "bmi"                               "under_five_deaths"
## [13] "polio"                             "total_expenditure"
## [15] "diphtheria"                        "hiv_aids"
## [17] "gdp"                               "population"
## [19] "thinness_1_19_years"               "thinness_5_9_years"
## [21] "income_composition_of_resources" "schooling"

# Select only numeric variables of interest
selected_data <- life_expectancy_data1 %>%

select(life_expectancy,adult_mortality,infant_deaths,alcohol,percentage_expen
diture,hepatitis_b,measles,bmi,under_five_deaths,polio,total_expenditure,
diphtheria,gdp,population,thinness_1_19_years,thinness_5_9_years,income_compo
sition_of_resources,schooling)

# Compute correlation matrix Numerically

cor_matrix <- cor(selected_data )

cor_matrix

##                                 life_expectancy adult_mortality
infant_deaths
## life_expectancy                      1.00000000     -0.58512976   -
0.32144790
## adult_mortality                     -0.58512976      1.00000000
0.24040937
## infant_deaths                       -0.32144790      0.24040937
1.00000000
## alcohol                              0.43862120     -0.23115063   -
0.26019450
## percentage_expenditure               0.10921054     -0.08391086   -
0.12388641
## hepatitis_b                          0.25380817     -0.17286324   -
0.16068040
## measles                             -0.05327925      0.04724471
0.09162027
## bmi                                  0.46811339     -0.32970912   -
0.22534179
## under_five_deaths                   -0.34825532      0.25514593
0.95020739
## polio                                0.30668626     -0.17626144   -
0.18722751
## total_expenditure                    0.26013934     -0.18183318   -
0.18721787
## diphtheria                           0.30584345     -0.18708186   -
0.19754481
## gdp                                  0.23537866     -0.09881005   -
0.05108806
```

```
## population                            0.06209374     -0.01744610
0.07253506
## thinness_1_19_years                  -0.49713756      0.31110295
0.22393194
## thinness_5_9_years                   -0.50444612      0.32935666
0.22600688
## income_composition_of_resources       0.73817312     -0.45464590   -
0.35196641
## schooling                             0.63374642     -0.36324797   -
0.30849077
##                                       alcohol percentage_expenditure
hepatitis_b
## life_expectancy                     0.43862120             0.10921054
0.253808167
## adult_mortality                    -0.23115063            -0.08391086 -
0.172863236
## infant_deaths                      -0.26019450            -0.12388641 -
0.160680397
## alcohol                             1.00000000             0.06957154
0.097297969
## percentage_expenditure              0.06957154             1.00000000
0.080851624
## hepatitis_b                         0.09729797             0.08085162
1.000000000
## measles                            -0.06036623            -0.04924538 -
0.103344355
## bmi                                 0.31029609             0.12742238
0.217558925
## under_five_deaths                  -0.26599968            -0.13157851 -
0.165921596
## polio                               0.20520294             0.06183821
0.482573020
## total_expenditure                   0.34798186             0.04421685
0.004843272
## diphtheria                          0.21901758             0.04985710
0.497246707
## gdp                                 0.22424538            -0.13426007
0.184629959
## population                         -0.01299237            -0.26708579
0.145924729
## thinness_1_19_years                -0.37978115            -0.08859299 -
0.052317736
## thinness_5_9_years                 -0.37179146            -0.07832672 -
0.061968346
## income_composition_of_resources     0.49125192             0.18571460
0.324741354
## schooling                           0.46300887             0.15593455
0.292299859
##                                        measles         bmi under_five_deaths
## life_expectancy                    -0.05327925  0.46811339       -0.34825532
```

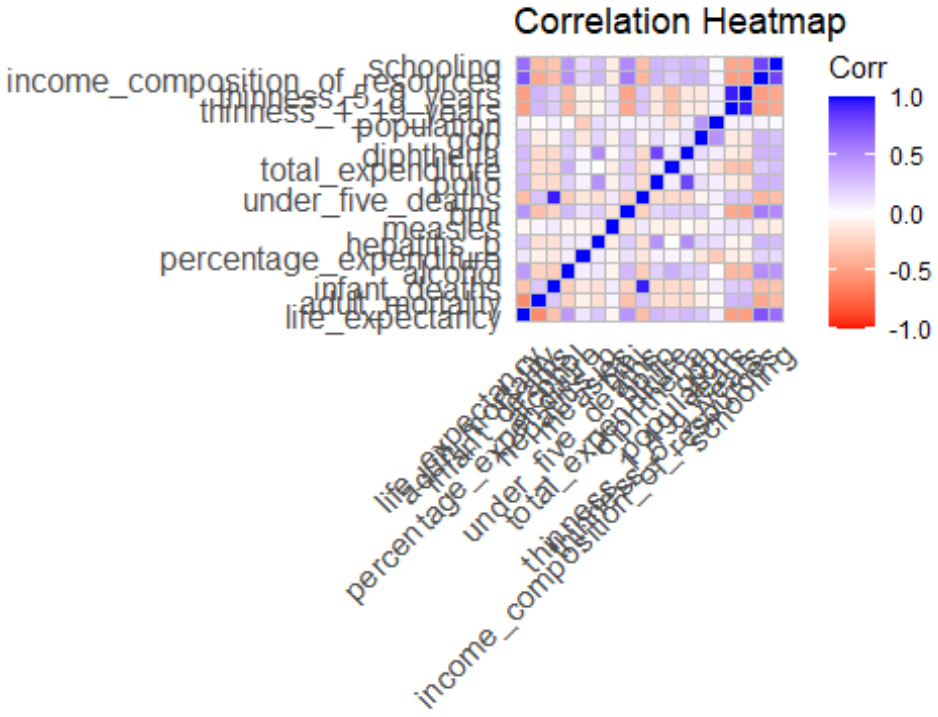
```
## adult_mortality                  0.04724471 -0.32970912        0.25514593
## infant_deaths                    0.09162027 -0.22534179        0.95020739
## alcohol                         -0.06036623  0.31029609       -0.26599968
## percentage_expenditure          -0.04924538  0.12742238       -0.13157851
## hepatitis_b                     -0.10334435  0.21755892       -0.16592160
## measles                          1.00000000 -0.13031730        0.10152366
## bmi                             -0.13031730  1.00000000       -0.25390321
## under_five_deaths                0.10152366 -0.25390321        1.00000000
## polio                           -0.07322299  0.17741129       -0.19309329
## total_expenditure               -0.10463831  0.23345937       -0.20192445
## diphtheria                      -0.03109455  0.18671692       -0.20354626
## gdp                             -0.06909893  0.23451297       -0.05612528
## population                       0.06852713  0.04854872        0.07095349
## thinness_1_19_years              0.12591596 -0.45423326        0.25220506
## thinness_5_9_years               0.13215368 -0.47083867        0.25459183
## income_composition_of_resources -0.10888163  0.57437469       -0.38059758
## schooling                       -0.09691394  0.50684783       -0.34229028
##                                       polio total_expenditure  diphtheria
## life_expectancy                  0.30668626       0.260139335  0.30584345
## adult_mortality                 -0.17626144      -0.181833182 -0.18708186
## infant_deaths                   -0.18722751      -0.187217871 -0.19754481
## alcohol                          0.20520294       0.347981857  0.21901758
## percentage_expenditure           0.06183821       0.044216849  0.04985710
## hepatitis_b                      0.48257302       0.004843272  0.49724671
## measles                         -0.07322299      -0.104638315 -0.03109455
## bmi                              0.17741129       0.233459366  0.18671692
## under_five_deaths               -0.19309329      -0.201924454 -0.20354626
## polio                            1.00000000       0.068927015  0.80339217
## total_expenditure                0.06892701       1.000000000  0.08013571
## diphtheria                       0.80339217       0.080135707  1.00000000
## gdp                              0.13595444       0.054822976  0.15124019
## population                       0.06594836      -0.159874901  0.08477021
## thinness_1_19_years             -0.11265902      -0.322283387 -0.10987415
## thinness_5_9_years              -0.13830242      -0.335388933 -0.13227482
## income_composition_of_resources  0.33116681       0.217988377  0.33272938
## schooling                        0.31971744       0.277241114  0.31605271
##                                         gdp  population
thinness_1_19_years
## life_expectancy                  0.23537866  0.06209374         -
0.49713756
## adult_mortality                 -0.09881005 -0.01744610
0.31110295
## infant_deaths                   -0.05108806  0.07253506
0.22393194
## alcohol                          0.22424538 -0.01299237         -
0.37978115
## percentage_expenditure          -0.13426007 -0.26708579         -
0.08859299
## hepatitis_b                      0.18462996  0.14592473         -
0.05231774
```

```
## measles                          -0.06909893  0.06852713
0.12591596
## bmi                               0.23451297  0.04854872         -
0.45423326
## under_five_deaths                -0.05612528  0.07095349
0.25220506
## polio                             0.13595444  0.06594836         -
0.11265902
## total_expenditure                 0.05482298 -0.15987490         -
0.32228339
## diphtheria                        0.15124019  0.08477021         -
0.10987415
## gdp                               1.00000000  0.46913655         -
0.11161609
## population                        0.46913655  1.00000000
0.06461616
## thinness_1_19_years              -0.11161609  0.06461616
1.00000000
## thinness_5_9_years               -0.12444238  0.06214674
0.94680997
## income_composition_of_resources   0.31718126  0.04722707         -
0.50904026
## schooling                         0.27333164  0.02347906         -
0.45348541
##                                  thinness_5_9_years
## life_expectancy                        -0.50444612
## adult_mortality                         0.32935666
## infant_deaths                           0.22600688
## alcohol                                -0.37179146
## percentage_expenditure                 -0.07832672
## hepatitis_b                            -0.06196835
## measles                                 0.13215368
## bmi                                    -0.47083867
## under_five_deaths                       0.25459183
## polio                                  -0.13830242
## total_expenditure                      -0.33538893
## diphtheria                             -0.13227482
## gdp                                    -0.12444238
## population                              0.06214674
## thinness_1_19_years                     0.94680997
## thinness_5_9_years                      1.00000000
## income_composition_of_resources        -0.50818187
## schooling                              -0.45669152
##                                  income_composition_of_resources
schooling
## life_expectancy                                        0.73817312
0.63374642
## adult_mortality                                       -0.45464590 -
0.36324797
## infant_deaths                                         -0.35196641 -
```

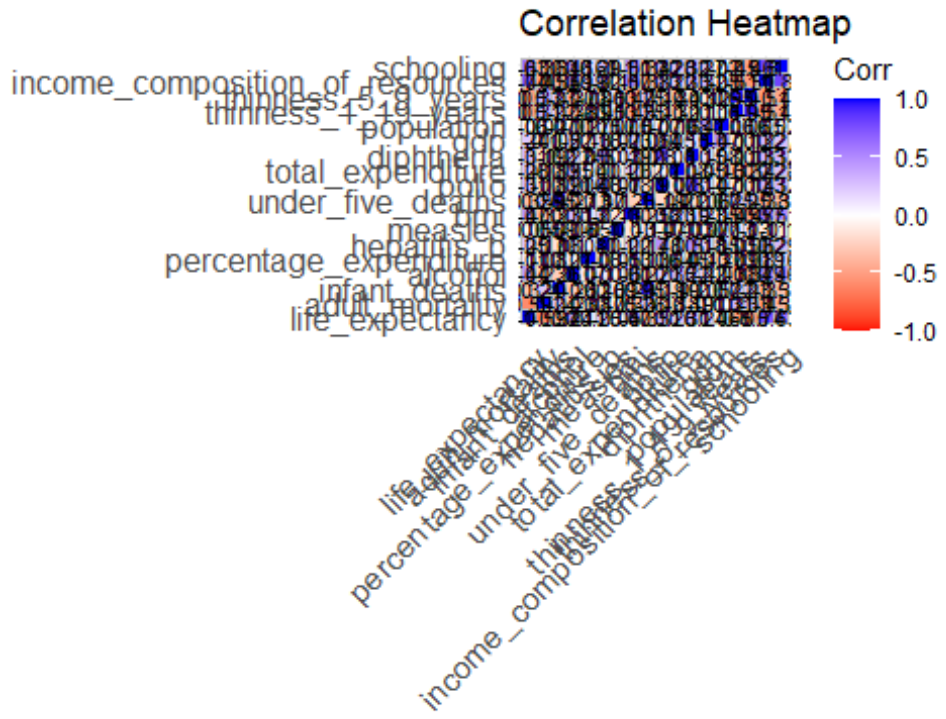
```
0.30849077
## alcohol                                                0.49125192
0.46300887
## percentage_expenditure                                 0.18571460
0.15593455
## hepatitis_b                                            0.32474135
0.29229986
## measles                                               -0.10888163 -
0.09691394
## bmi                                                    0.57437469
0.50684783
## under_five_deaths                                     -0.38059758 -
0.34229028
## polio                                                  0.33116681
0.31971744
## total_expenditure                                      0.21798838
0.27724111
## diphtheria                                             0.33272938
0.31605271
## gdp                                                    0.31718126
0.27333164
## population                                             0.04722707
0.02347906
## thinness_1_19_years                                   -0.50904026 -
0.45348541
## thinness_5_9_years                                    -0.50818187 -
0.45669152
## income_composition_of_resources                        1.00000000
0.80047481
## schooling                                              0.80047481
1.00000000
```

3.Exporting table for VISUALIZATION

```r
ggcorrplot(cor_matrix,
           method = "square",
           type = "full",
           lab = FALSE,         # shows correlation coefficients/r
           lab_size = 3,
           title = "Correlation Heatmap",
           colors = c("red", "white", "blue"),
           ggtheme = ggplot2::theme_minimal())
```




```r
# Save the heatmap with larger dimensions

ggsave("correlation_heatmap.png",
       plot = last_plot(),
       width = 12, height = 10, dpi = 300)

ggcorrplot(cor_matrix,
           method = "square",
           type = "full",
           lab = TRUE,          # shows correlation coefficients/r
           lab_size = 3,
           title = "Correlation Heatmap",
           colors = c("red", "white", "blue"),
           ggtheme = ggplot2::theme_minimal())
```




```
# Save the heatmap with larger dimensions

ggsave("correlation_heatmap.png",
       plot = last_plot(),
       width = 12, height = 10, dpi = 300)
```

Interpretation

The matrix shows that we don't have extreme correlations, means no multicollinearity seen in matrix findings of r> 0.8 or >-0.8

```
# Correlation between Life Expectancy and Schooling

cor(selected_data$life_expectancy, selected_data$schooling)

## [1] 0.6337464
```

Interpretation of life expectancy and schooling

There is a strong positive relationship between life expectancy and schooling (r = + 0.63). This implies that as the level of education increases, people tend to live longer. Education likely improves health awareness, promotes healthy behaviors, and increases access to better jobs and healthcare.

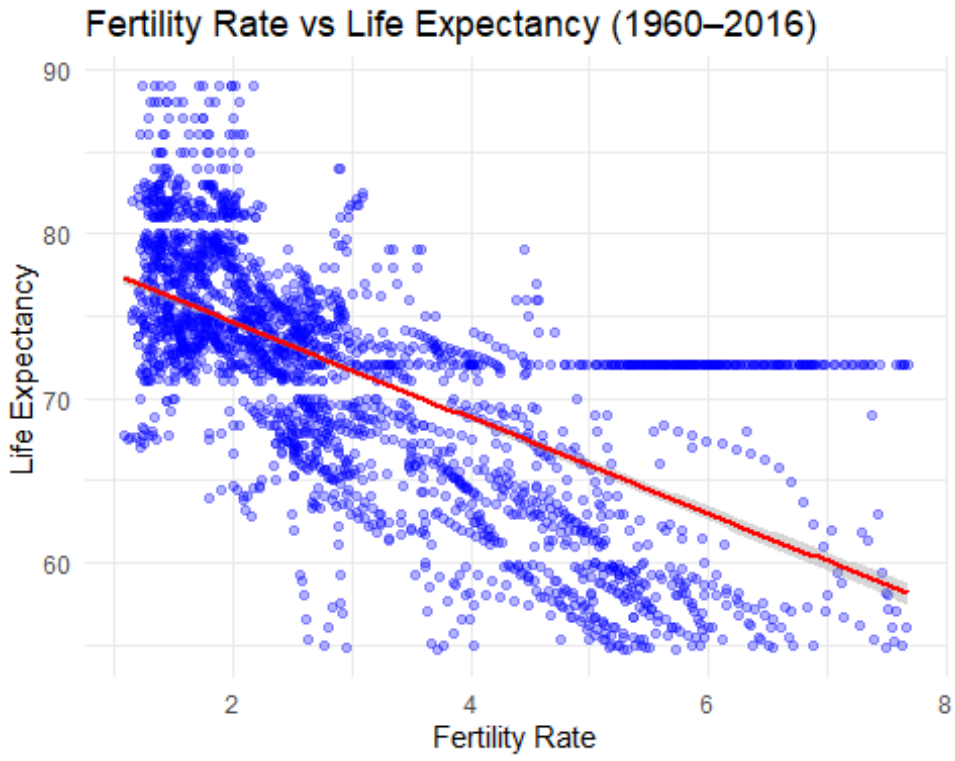
2. Create a new variable called fertility rate in the life expectancy data dataset. It should contain values from the fertility_rate dataset, merged on both Country and Year. Using appropriate methods (such as visualization, statistical analysis of numerical values, or regression), determine the pattern of the relationship between life expectancy and fertility rate

```
# merge them by country and year

# Rename

fertility_long <- fertility_long %>%
  rename(year = Year)


# life_expectancy_data1 , This is life expectancy data data set with details
of clinical information


merged_data <- left_join(life_expectancy_data1, fertility_long,
                         by = c("country", "year"))
```

Relationship between life expectancy and fertility rate

1. Use scatter Plot with trendline

```
ggplot(merged_data, aes(x = FertilityRate, y = life_expectancy)) +
  geom_point(alpha = 0.3, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Fertility Rate vs Life Expectancy (1960-2016)",
       x = "Fertility Rate",
       y = "Life Expectancy") +
  theme_minimal()

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

## Warning: Removed 442 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 442 rows containing missing values or values outside the
scale range
## (`geom_point()`).
```



2. correlation

```
cor.test(merged_data$FertilityRate, merged_data$life_expectancy)

##
##  Pearson's product-moment correlation
##
## data:  merged_data$FertilityRate and merged_data$life_expectancy
## t = -42.215, df = 2494, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6678881 -0.6220930
## sample estimates:
##        cor
## -0.6455706
```

Interpretation

A Pearson's product-moment correlation was performed to assess the relationship between fertility rate and life expectancy. The results indicated a strong negative correlation:

Correlation coefficient (r): -0.64

Conclusion: This correlation is statistically significant at p-value <0.05 , suggesting that as fertility rates increase, life expectancy tends to decrease, and vice versa.

3. Linear Regression

```
model <- lm(life_expectancy ~ FertilityRate, data = merged_data)
summary(model)
```

```
##
## Call:
## lm(formula = life_expectancy ~ FertilityRate, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.520  -3.894  -0.179   3.626  14.876
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   80.43092    0.24105  333.67   <2e-16 ***
## FertilityRate -2.90623    0.06884  -42.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.538 on 2494 degrees of freedom
##   (442 observations deleted due to missingness)
## Multiple R-squared:  0.4168, Adjusted R-squared:  0.4165
## F-statistic:  1782 on 1 and 2494 DF,  p-value: < 2.2e-16
```

Interpretation of the linear model

The regression analysis shows a strong and statistically significant inverse relationship between fertility rate and life expectancy:

1. Each unit increase in fertility rate is associated with a decrease of approximately 2.9 years in life expectancy.
2. The model explains 41.1% of the variation in life expectancy ($R^2 = 0.41$), indicating a good fit.
3. The results are highly statistically significant ($p < 0.001$), meaning the association is unlikely due to chance. and The negative relationship suggests that countries or regions with higher fertility rates tend to have lower life expectancy, possibly due to socioeconomic and health-related factors.

3. Using the life expectancy data dataset (from the previous question), drop the Population variable. Assume that Life expectancy is the dependent variable, and the remaining variables are independent variables. Develop a linear regression model and evaluate its performance using the R-squared value and RMSE. What can be done to improve the performance of the linear re gression model?

```
# Drop 'Population' column and remove rows with missing values

life_data_clean <- life_expectancy_data1[, !(names(life_expectancy_data1)
%in% c("population","hiv_aids"))]
life_data_clean <- na.omit(life_data_clean)

life_data_clean$status <- as.factor(life_data_clean$status)
```

Removed variables due to meaningless variation

HIV/AIDS has no variations. However, we have to remove it in analysis

```
summary(life_expectancy_data1$hiv_aids)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.1     0.1     0.1     0.1     0.1     0.1
```

With All non-essential variables to be removed and run final model

```
life_data_clean1 <- life_expectancy_data1[, !(names(life_expectancy_data1)
%in% c("population","country","year","hiv_aids","continent"))]

life_data_clean1 <- na.omit(life_data_clean1)

life_data_clean1$status <- as.factor(life_data_clean1$status)

# Fit the model
model1 <- lm(life_expectancy ~ ., data = life_data_clean1)

# View model summary

summary(model1)

##
## Call:
## lm(formula = life_expectancy ~ ., data = life_data_clean1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.9768  -2.1403  -0.0273   2.1431  14.8518
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.403e+01  1.510e+00  35.773  < 2e-16 ***
## statusDeveloping               -1.577e+00  2.821e-01  -5.589 2.49e-08 ***
## adult_mortality                -2.263e-02  9.868e-04 -22.933  < 2e-16 ***
## infant_deaths                   9.602e-02  9.651e-02   0.995  0.31985
## alcohol                         3.855e-02  2.714e-02   1.421  0.15556
## percentage_expenditure         -2.787e-03  1.473e-03  -1.893  0.05848 .
## hepatitis_b                     8.488e-03  1.103e-02   0.770  0.44159
## measles                         1.750e-02  6.195e-03   2.824  0.00477 **
## bmi                            -7.000e-04  4.957e-03  -0.141  0.88772
## under_five_deaths              -1.207e-01  7.501e-02  -1.609  0.10769
## polio                           6.233e-02  2.284e-02   2.729  0.00638 **
## total_expenditure               9.170e-02  4.547e-02   2.017  0.04380 *
## diphtheria                      6.869e-03  2.310e-02   0.297  0.76621
## gdp                             4.347e-05  3.043e-05   1.428  0.15327
## thinness_1_19_years            -1.497e-01  9.190e-02  -1.628  0.10353
## thinness_5_9_years             -1.480e-01  9.167e-02  -1.615  0.10642
## income_composition_of_resources  1.884e+01  1.019e+00  18.488  < 2e-16 ***
## schooling                       2.185e-01  5.499e-02   3.974 7.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.167 on 2920 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6528
## F-statistic: 325.8 on 17 and 2920 DF,  p-value: < 2.2e-16
```

Evaluate model performance

```
# R-squared is included in summary
#r_squared <- summary(model)$r.squared

# RMSE calculation
#predictions <- predict(model, newdata = life_data_clean)
#rmse <- sqrt(mean((life_data_clean$life_expectancy - predictions)^2))

# Print metrics
#cat("R-squared:", r_squared, "\n")
#cat("RMSE:", rmse, "\n")
```

R-squared: 0.8157638 RMSE: 3.034967

Interpetation of model performance findings

The model performance metrics indicate a strong predictive ability. The R-squared value of 0.81 suggests that approximately 84.9% of the variation in life expectancy is explained by the predictors included in the model. Additionally, the Root Mean Square Error (RMSE) of 3.04 implies that, on average, the model's predictions deviate from the actual life expectancy values by about 2 years. These results reflect a highly accurate model for

estimating life expectancy based on the selected socioeconomic and health-related variables.

Overall Interpretation of the multiple regression model

🔵 Positive Factors Increasing Life Expectancy Several key factors were found to significantly contribute to higher life expectancy:

1. Access to resources and income equity: Countries where people have better access to basic resources and where income is more evenly distributed tend to have longer life expectancy. This reflects the importance of economic and social stability in supporting health and wellbeing.

2. Education (schooling): Higher levels of education in a population are strongly linked to longer life expectancy. Education likely promotes healthier lifestyles, better healthcare decisions, and increased access to employment and income.

3. Immunization coverage: Greater coverage of vaccines such as polio was associated with improved life expectancy. This highlights the vital role of preventive health measures in reducing disease and prolonging life.

4. Health system investment: Countries that allocate a higher portion of their expenditure to health tend to enjoy better health outcomes and longer life spans. This suggests that investing in healthcare infrastructure and services is essential for population health.

5. Effective disease monitoring and reporting: While measles is a disease with harmful effects, its positive association with life expectancy in this analysis may reflect better disease surveillance and reporting systems typically found in more developed health systems.

🔴 Negative Factors Decreasing Life Expectancy Some factors were found to significantly reduce life expectancy:

1. High adult mortality: A higher death rate among adults strongly lowers the overall life expectancy, indicating poor health conditions or access to care for the working-age population.

2. Developing country status: Countries classified as developing generally have lower life expectancy. This likely results from limited healthcare infrastructure, inadequate access to clean water and sanitation, lower education levels, and other social determinants of health.

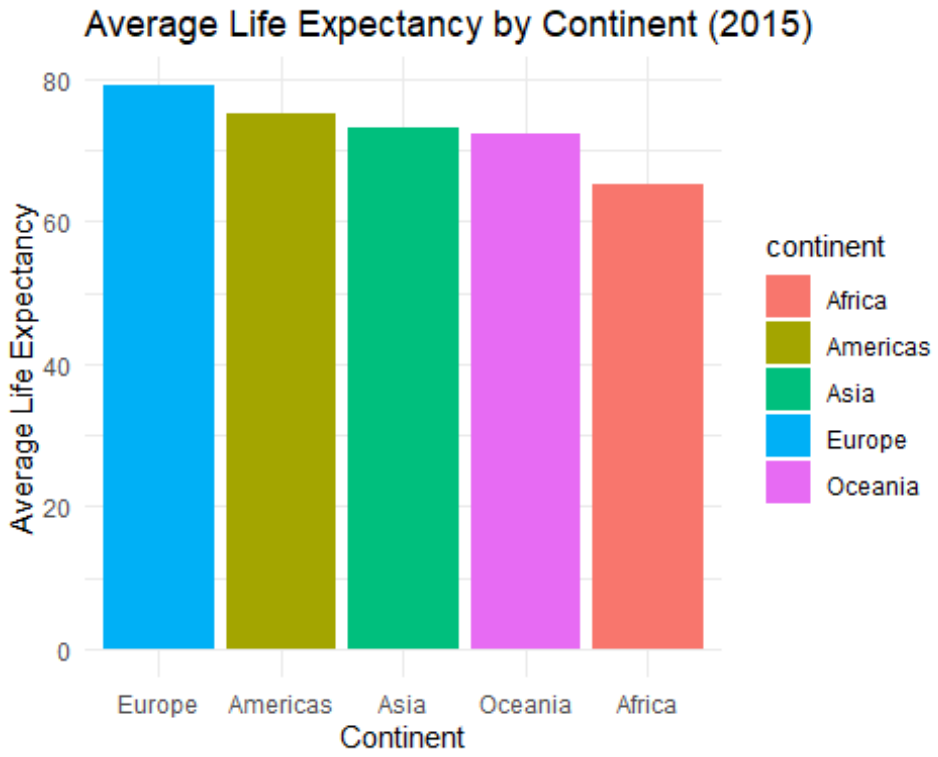
What to do to improve performance of model

1. Feature Engineering: Check for non-linear relationships, Transform skewed variables (e.g., log of GDP).
2. Remove Multicollinearity:
3. Regularization: Try Ridge or Lasso regression (via glmnet package).
4. Outlier Handling: Use diagnostic plots to identify and remove outliers.
5. And, others

4.5 Comparing life expectancy by continent

classify country by contint using valid list of countries

```
life_expectancy_data1$continent <- countrycode(life_expectancy_data1$country,
                                       origin = "country.name",
                                       destination = "continent")

life_2015 <- life_expectancy_data1 %>%
  filter(year == 2015, !is.na(continent), !is.na(life_expectancy)) %>%
  group_by(continent) %>%
  summarise(avg_life_expectancy = mean(life_expectancy, na.rm = TRUE)) %>%
  arrange(desc(avg_life_expectancy))


ggplot(life_2015, aes(x = reorder(continent, -avg_life_expectancy),
                      y = avg_life_expectancy, fill = continent)) +
  geom_col() +
  labs(title = "Average Life Expectancy by Continent (2015)",
       x = "Continent", y = "Average Life Expectancy") +
  theme_minimal()
```



Which continent has the lowest? What is its average?

```
life_2015 %>% slice_tail(n = 1)
```

```
## # A tibble: 1 × 2
##   continent avg_life_expectancy
##   <chr>                   <dbl>
## 1 Africa                   65.1
```

Which continent has the Highest ? What is its average?

```
life_2015 %>% slice_head(n = 1)
```

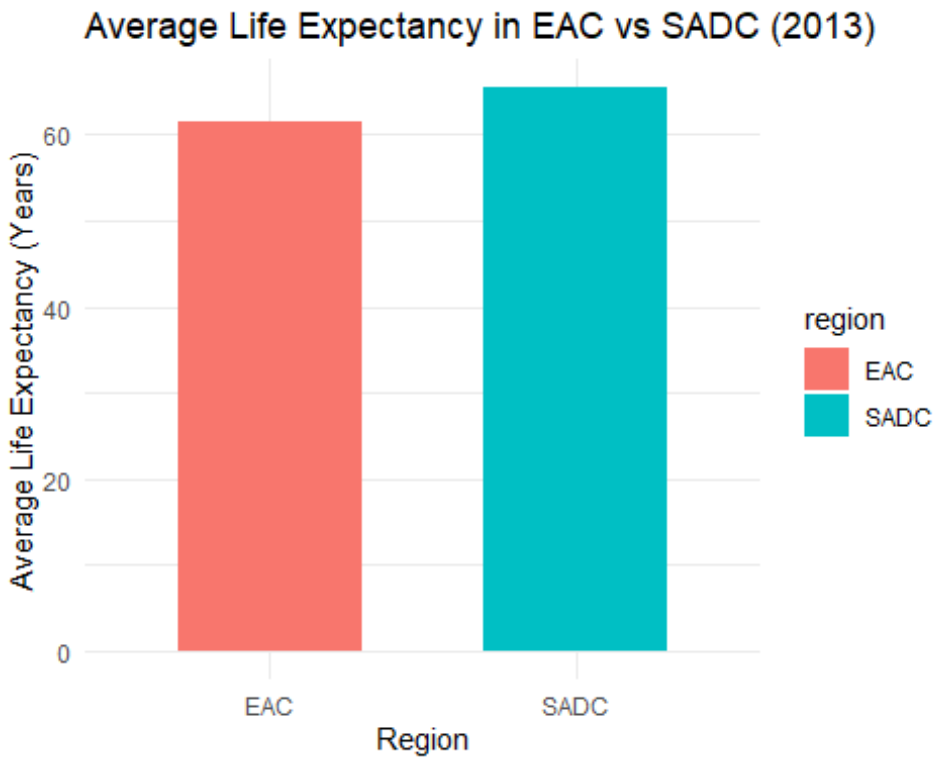
```
## # A tibble: 1 × 2
##   continent avg_life_expectancy
##   <chr>                   <dbl>
## 1 Europe                   79.0
```

4.6 Comparing life expectancy in EAC and SADC

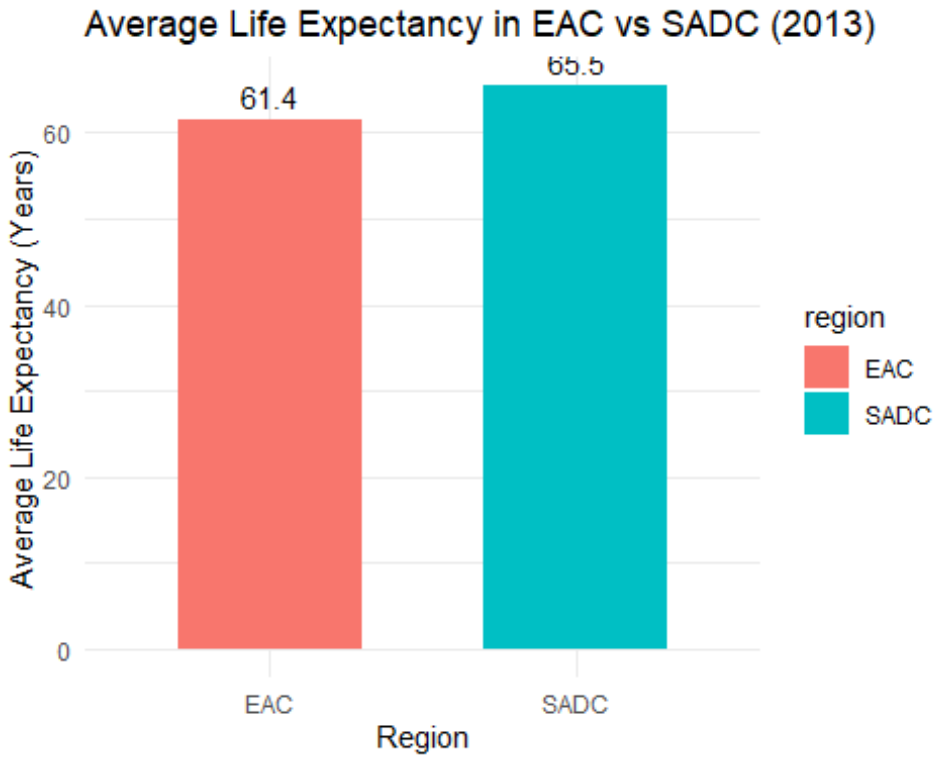
Extract the life expectancy data for the year 2013 only. Use it to compare the average life expectancy between the East African Community and the Southern African Development Community.

```
## Define EAC and SADC

EAC_countries <- c("Burundi", "Kenya", "Rwanda", "South Sudan", "Tanzania",
"Uganda", "Democratic Republic of the Congo")
```

```r
SADC_countries <- c("Angola", "Botswana", "Comoros", "Democratic Republic of
the Congo", "Eswatini", "Lesotho",
                    "Madagascar", "Malawi", "Mauritius", "Mozambique",
"Namibia", "Seychelles", "South Africa",
                    "Tanzania", "Zambia", "Zimbabwe")

## Filter 2013 data and assign regional groups

life_2013 <- life_expectancy_data1 %>%
  filter(year == 2013, !is.na(life_expectancy)) %>%
  mutate(region = case_when(
    country %in% EAC_countries ~ "EAC",
    country %in% SADC_countries ~ "SADC",
    TRUE ~ NA_character_
  )) %>%
  filter(!is.na(region))

## compute average per each region

avg_life_by_region <- life_2013 %>%
  group_by(region) %>%
  summarise(avg_life_expectancy = mean(life_expectancy, na.rm = TRUE))

## Plot for comparison

ggplot(avg_life_by_region, aes(x = region, y = avg_life_expectancy, fill =
region)) +
  geom_col(width = 0.6) +
  labs(title = "Average Life Expectancy in EAC vs SADC (2013)",
       x = "Region", y = "Average Life Expectancy (Years)") +
  theme_minimal()
```




```
##. speicif value to each


ggplot(avg_life_by_region, aes(x = region, y = avg_life_expectancy, fill =
region)) +
  geom_col(width = 0.6) +
  geom_text(aes(label = round(avg_life_expectancy, 1)), vjust = -0.5, size =
4) +
  labs(title = "Average Life Expectancy in EAC vs SADC (2013)",
       x = "Region", y = "Average Life Expectancy (Years)") +
  theme_minimal()
```


Average Life Expectancy in EAC vs SADC (2013)
61.4
65.5
60
40
20
0
Average Life Expectancy (Years)
EAC
SADC
Region
region
EAC
SADC

Thank you