Project-Objective 2

Group 1

2025-03-26

# What economic and social factors most influence lower life expectancy in gender?

## Dataset

# Load necessary libraries  
library(dplyr) # Data manipulation

## Warning: package 'dplyr' was built under R version 4.4.2

##   
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2) # Data visualization

## Warning: package 'ggplot2' was built under R version 4.4.2

library(caret) # Machine learning and feature selection

## Warning: package 'caret' was built under R version 4.4.2

## Cargando paquete requerido: lattice

library(corrplot) # Correlation plot

## Warning: package 'corrplot' was built under R version 4.4.2

## corrplot 0.95 loaded

library(car) # For regression diagnostics

## Warning: package 'car' was built under R version 4.4.2

## Cargando paquete requerido: carData

## Warning: package 'carData' was built under R version 4.4.2

##   
## Adjuntando el paquete: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(countrycode)

## Warning: package 'countrycode' was built under R version 4.4.2

# Load dataset  
global\_health <- read.csv("global\_health.csv")  
  
# Inspect dataset  
str(global\_health) # Check structure

## 'data.frame': 1880 obs. of 29 variables:  
## $ Country : chr "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
## $ Country\_Code : chr "AFG" "AFG" "AFG" "AFG" ...  
## $ Year : int 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 ...  
## $ Fertility\_Rate : num 5.83 5.7 5.56 5.41 5.26 ...  
## $ Urban\_Population\_Percent : num 24.2 24.4 24.6 24.8 25 ...  
## $ Total\_Population : num 30466479 31541209 32716210 33753499 34636207 ...  
## $ Water\_Access\_Percent : num 21.1 22 22.9 23.9 24.8 ...  
## $ Unemployment\_Rate : num 7.91 7.92 7.92 9.01 10.1 ...  
## $ Sanitary\_Expense\_Per\_GDP : num 7.9 8.81 9.53 10.11 11.82 ...  
## $ Life\_Expectancy : num 61.9 62.4 62.5 62.7 63.1 ...  
## $ Life\_Expectancy\_Female : num 63.5 64 64.3 64.6 65.1 ...  
## $ Life\_Expectancy\_Male : num 60.3 60.8 60.8 60.8 61.2 ...  
## $ Infant\_Deaths : num 70669 70906 70288 70068 69144 ...  
## $ GDP\_Per\_Capita : num 653 639 627 567 523 ...  
## $ Hospital\_Beds\_Per\_1000 : num 0.47 0.46 0.46 0.44 0.45 0.42 0.4 0.38 0.34 0.36 ...  
## $ Female\_Population : num 15067373 15594637 16172321 16682054 17115346 ...  
## $ Male\_Population : num 15399105 15946572 16543889 17071446 17520861 ...  
## $ Alcohol\_Consumption\_Per\_Capita: num 0.004 0.004 0.004 0.009 0.013 0.017 0.016 0.016 0.011 NA ...  
## $ Immunization\_Rate : num 67 64 62 64 66 66 72 72 70 66 ...  
## $ Sanitary\_Expense\_Per\_Capita : num 52.6 56.3 60.2 60.1 61.5 ...  
## $ CO2\_Exposure\_Percent : num 70.9 73.1 77.1 73.5 72.8 ...  
## $ Air\_Pollution : num 70.9 73.1 77.1 73.5 72.8 ...  
## $ Labour\_Force\_Total : num 7520865 7881567 8285362 8630724 8913938 ...  
## $ Tuberculosis\_Per\_100000 : num 189 189 189 189 189 189 189 189 183 185 ...  
## $ Suicide\_Rate\_Percent : num 3.68 3.66 3.6 3.57 3.61 3.44 3.52 3.61 NA NA ...  
## $ Obesity\_Rate\_Percent : num 10.7 11.6 10.4 11.2 12 ...  
## $ Underweight\_Rate\_Percent : num 10.15 10.79 10.17 10.52 7.88 ...  
## $ Overweight\_Rate\_Percent : num 31.6 32.7 34 35.2 36.5 ...  
## $ Safe\_Water\_Access\_Percent : num 46.7 49.5 52.2 55.1 58 ...

summary(global\_health) # Get basic statistics

## Country Country\_Code Year Fertility\_Rate   
## Length:1880 Length:1880 Min. :2012 Min. :0.808   
## Class :character Class :character 1st Qu.:2014 1st Qu.:1.670   
## Mode :character Mode :character Median :2016 Median :2.311   
## Mean :2016 Mean :2.763   
## 3rd Qu.:2019 3rd Qu.:3.646   
## Max. :2021 Max. :7.400   
## NA's :36   
## Urban\_Population\_Percent Total\_Population Water\_Access\_Percent  
## Min. : 11.19 Min. :1.044e+04 Min. : 5.863   
## 1st Qu.: 40.07 1st Qu.:2.062e+06 1st Qu.: 44.034   
## Median : 58.88 Median :8.693e+06 Median : 76.250   
## Mean : 58.62 Mean :3.960e+07 Mean : 68.157   
## 3rd Qu.: 77.63 3rd Qu.:2.918e+07 3rd Qu.: 97.108   
## Max. :100.00 Max. :1.412e+09 Max. :100.000   
## NA's :655   
## Unemployment\_Rate Sanitary\_Expense\_Per\_GDP Life\_Expectancy  
## Min. : 0.100 Min. : 1.515 Min. :47.84   
## 1st Qu.: 3.716 1st Qu.: 4.400 1st Qu.:65.37   
## Median : 5.811 Median : 6.280 Median :72.55   
## Mean : 7.684 Mean : 6.647 Mean :71.44   
## 3rd Qu.:10.245 3rd Qu.: 8.428 3rd Qu.:77.42   
## Max. :35.707 Max. :24.283 Max. :84.56   
## NA's :140 NA's :46 NA's :40   
## Life\_Expectancy\_Female Life\_Expectancy\_Male Infant\_Deaths   
## Min. :50.49 Min. :45.36 Min. : 0   
## 1st Qu.:68.06 1st Qu.:62.96 1st Qu.: 250   
## Median :76.09 Median :69.30 Median : 1996   
## Mean :74.10 Mean :68.89 Mean : 22859   
## 3rd Qu.:80.40 3rd Qu.:74.68 3rd Qu.: 15004   
## Max. :87.71 Max. :82.60 Max. :1104742   
## NA's :40 NA's :40 NA's :10   
## GDP\_Per\_Capita Hospital\_Beds\_Per\_1000 Female\_Population   
## Min. : 216.8 Min. : 0.170 Min. : 5111   
## 1st Qu.: 1984.5 1st Qu.: 1.300 1st Qu.: 1038946   
## Median : 5875.1 Median : 2.590 Median : 4273046   
## Mean : 15919.4 Mean : 3.153 Mean : 19670179   
## 3rd Qu.: 17585.2 3rd Qu.: 4.260 3rd Qu.: 14458779   
## Max. :235132.8 Max. :22.300 Max. :691219627   
## NA's :27 NA's :599   
## Male\_Population Alcohol\_Consumption\_Per\_Capita Immunization\_Rate  
## Min. : 5333 Min. : 0.000 Min. :19.00   
## 1st Qu.: 1026041 1st Qu.: 1.758 1st Qu.:84.00   
## Median : 4331481 Median : 4.920 Median :93.00   
## Mean : 19926179 Mean : 5.409 Mean :87.94   
## 3rd Qu.: 14404744 3rd Qu.: 8.592 3rd Qu.:97.00   
## Max. :726503429 Max. :17.180 Max. :99.00   
## NA's :256 NA's :10   
## Sanitary\_Expense\_Per\_Capita CO2\_Exposure\_Percent Air\_Pollution   
## Min. : 13.63 Min. : 4.895 Min. : 4.895   
## 1st Qu.: 86.75 1st Qu.: 13.894 1st Qu.: 13.894   
## Median : 353.00 Median : 21.181 Median : 21.181   
## Mean : 1158.32 Mean : 26.298 Mean : 26.298   
## 3rd Qu.: 1158.39 3rd Qu.: 32.333 3rd Qu.: 32.333   
## Max. :12012.24 Max. :107.145 Max. :107.145   
## NA's :47 NA's :197 NA's :197   
## Labour\_Force\_Total Tuberculosis\_Per\_100000 Suicide\_Rate\_Percent  
## Min. : 36260 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 1439398 1st Qu.: 13.0 1st Qu.: 2.21   
## Median : 4341022 Median : 48.0 Median : 10.42   
## Mean : 19161171 Mean : 118.5 Mean : 9.57   
## 3rd Qu.: 12177482 3rd Qu.: 167.8 3rd Qu.: 13.57   
## Max. :780709584 Max. :1180.0 Max. :147.80   
## NA's :140 NA's :210 NA's :608   
## Obesity\_Rate\_Percent Underweight\_Rate\_Percent Overweight\_Rate\_Percent  
## Min. : 0.47 Min. : 0.230 Min. : 5.46   
## 1st Qu.:11.35 1st Qu.: 1.350 1st Qu.:25.91   
## Median :16.49 Median : 3.060 Median :45.34   
## Mean :17.72 Mean : 5.509 Mean :42.33   
## 3rd Qu.:23.32 3rd Qu.: 9.572 3rd Qu.:55.01   
## Max. :69.08 Max. :31.090 Max. :88.81   
## NA's :230 NA's :230 NA's :230   
## Safe\_Water\_Access\_Percent  
## Min. : 25.61   
## 1st Qu.: 65.56   
## Median : 91.07   
## Mean : 81.96   
## 3rd Qu.: 99.68   
## Max. :100.00   
## NA's :232

head(global\_health) # Preview first few rows

## Country Country\_Code Year Fertility\_Rate Urban\_Population\_Percent  
## 1 Afghanistan AFG 2012 5.830 24.160  
## 2 Afghanistan AFG 2013 5.696 24.373  
## 3 Afghanistan AFG 2014 5.560 24.587  
## 4 Afghanistan AFG 2015 5.405 24.803  
## 5 Afghanistan AFG 2016 5.262 25.020  
## 6 Afghanistan AFG 2017 5.129 25.250  
## Total\_Population Water\_Access\_Percent Unemployment\_Rate  
## 1 30466479 21.12400 7.909  
## 2 31541209 22.03447 7.919  
## 3 32716210 22.94430 7.915  
## 4 33753499 23.85359 9.011  
## 5 34636207 24.76222 10.100  
## 6 35643418 25.67142 11.184  
## Sanitary\_Expense\_Per\_GDP Life\_Expectancy Life\_Expectancy\_Female  
## 1 7.897169 61.923 63.514  
## 2 8.805964 62.417 64.027  
## 3 9.528878 62.545 64.274  
## 4 10.105348 62.659 64.576  
## 5 11.818590 63.136 65.096  
## 6 12.620817 63.016 66.099  
## Life\_Expectancy\_Male Infant\_Deaths GDP\_Per\_Capita Hospital\_Beds\_Per\_1000  
## 1 60.317 70669 653.4175 0.47  
## 2 60.791 70906 638.7332 0.46  
## 3 60.812 70288 626.5129 0.46  
## 4 60.760 70068 566.8811 0.44  
## 5 61.193 69144 523.0530 0.45  
## 6 60.105 67644 526.1408 0.42  
## Female\_Population Male\_Population Alcohol\_Consumption\_Per\_Capita  
## 1 15067373 15399105 0.004  
## 2 15594637 15946572 0.004  
## 3 16172321 16543889 0.004  
## 4 16682054 17071446 0.009  
## 5 17115346 17520861 0.013  
## 6 17614722 18028696 0.017  
## Immunization\_Rate Sanitary\_Expense\_Per\_Capita CO2\_Exposure\_Percent  
## 1 67 52.61354 70.92232  
## 2 64 56.30554 73.13182  
## 3 62 60.18958 77.14373  
## 4 64 60.05854 73.49082  
## 5 66 61.48646 72.76591  
## 6 66 66.90922 65.86235  
## Air\_Pollution Labour\_Force\_Total Tuberculosis\_Per\_100000 Suicide\_Rate\_Percent  
## 1 70.92232 7520865 189 3.68  
## 2 73.13182 7881567 189 3.66  
## 3 77.14373 8285362 189 3.60  
## 4 73.49082 8630724 189 3.57  
## 5 72.76591 8913938 189 3.61  
## 6 65.86235 9254593 189 3.44  
## Obesity\_Rate\_Percent Underweight\_Rate\_Percent Overweight\_Rate\_Percent  
## 1 10.70 10.15 31.55  
## 2 11.55 10.79 32.73  
## 3 10.44 10.17 33.95  
## 4 11.19 10.52 35.19  
## 5 11.99 7.88 36.45  
## 6 10.04 7.39 37.73  
## Safe\_Water\_Access\_Percent  
## 1 46.68  
## 2 49.45  
## 3 52.25  
## 4 55.09  
## 5 57.97  
## 6 60.89

# Add continent column based on country names  
global\_health$Continent = countrycode(global\_health$Country, origin = "country.name",   
 destination = "continent")  
  
# View a few rows to confirm  
head(global\_health)

## Country Country\_Code Year Fertility\_Rate Urban\_Population\_Percent  
## 1 Afghanistan AFG 2012 5.830 24.160  
## 2 Afghanistan AFG 2013 5.696 24.373  
## 3 Afghanistan AFG 2014 5.560 24.587  
## 4 Afghanistan AFG 2015 5.405 24.803  
## 5 Afghanistan AFG 2016 5.262 25.020  
## 6 Afghanistan AFG 2017 5.129 25.250  
## Total\_Population Water\_Access\_Percent Unemployment\_Rate  
## 1 30466479 21.12400 7.909  
## 2 31541209 22.03447 7.919  
## 3 32716210 22.94430 7.915  
## 4 33753499 23.85359 9.011  
## 5 34636207 24.76222 10.100  
## 6 35643418 25.67142 11.184  
## Sanitary\_Expense\_Per\_GDP Life\_Expectancy Life\_Expectancy\_Female  
## 1 7.897169 61.923 63.514  
## 2 8.805964 62.417 64.027  
## 3 9.528878 62.545 64.274  
## 4 10.105348 62.659 64.576  
## 5 11.818590 63.136 65.096  
## 6 12.620817 63.016 66.099  
## Life\_Expectancy\_Male Infant\_Deaths GDP\_Per\_Capita Hospital\_Beds\_Per\_1000  
## 1 60.317 70669 653.4175 0.47  
## 2 60.791 70906 638.7332 0.46  
## 3 60.812 70288 626.5129 0.46  
## 4 60.760 70068 566.8811 0.44  
## 5 61.193 69144 523.0530 0.45  
## 6 60.105 67644 526.1408 0.42  
## Female\_Population Male\_Population Alcohol\_Consumption\_Per\_Capita  
## 1 15067373 15399105 0.004  
## 2 15594637 15946572 0.004  
## 3 16172321 16543889 0.004  
## 4 16682054 17071446 0.009  
## 5 17115346 17520861 0.013  
## 6 17614722 18028696 0.017  
## Immunization\_Rate Sanitary\_Expense\_Per\_Capita CO2\_Exposure\_Percent  
## 1 67 52.61354 70.92232  
## 2 64 56.30554 73.13182  
## 3 62 60.18958 77.14373  
## 4 64 60.05854 73.49082  
## 5 66 61.48646 72.76591  
## 6 66 66.90922 65.86235  
## Air\_Pollution Labour\_Force\_Total Tuberculosis\_Per\_100000 Suicide\_Rate\_Percent  
## 1 70.92232 7520865 189 3.68  
## 2 73.13182 7881567 189 3.66  
## 3 77.14373 8285362 189 3.60  
## 4 73.49082 8630724 189 3.57  
## 5 72.76591 8913938 189 3.61  
## 6 65.86235 9254593 189 3.44  
## Obesity\_Rate\_Percent Underweight\_Rate\_Percent Overweight\_Rate\_Percent  
## 1 10.70 10.15 31.55  
## 2 11.55 10.79 32.73  
## 3 10.44 10.17 33.95  
## 4 11.19 10.52 35.19  
## 5 11.99 7.88 36.45  
## 6 10.04 7.39 37.73  
## Safe\_Water\_Access\_Percent Continent  
## 1 46.68 Asia  
## 2 49.45 Asia  
## 3 52.25 Asia  
## 4 55.09 Asia  
## 5 57.97 Asia  
## 6 60.89 Asia

# View dataset   
View(global\_health)

# Convert Year to factor (categorical)  
global\_health$Year <- as.factor(global\_health$Year)

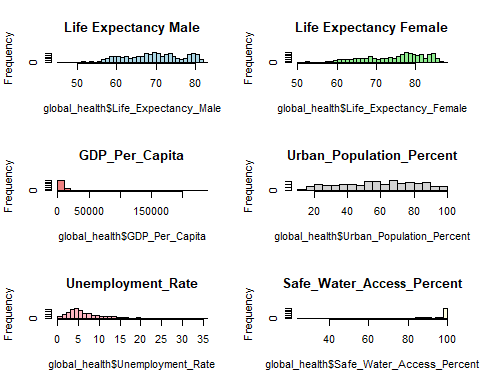
#Total missing values  
sum(is.na(global\_health))

## [1] 4220

colSums(is.na(global\_health))

## Country Country\_Code   
## 0 0   
## Year Fertility\_Rate   
## 0 36   
## Urban\_Population\_Percent Total\_Population   
## 0 0   
## Water\_Access\_Percent Unemployment\_Rate   
## 655 140   
## Sanitary\_Expense\_Per\_GDP Life\_Expectancy   
## 46 40   
## Life\_Expectancy\_Female Life\_Expectancy\_Male   
## 40 40   
## Infant\_Deaths GDP\_Per\_Capita   
## 10 27   
## Hospital\_Beds\_Per\_1000 Female\_Population   
## 599 0   
## Male\_Population Alcohol\_Consumption\_Per\_Capita   
## 0 256   
## Immunization\_Rate Sanitary\_Expense\_Per\_Capita   
## 10 47   
## CO2\_Exposure\_Percent Air\_Pollution   
## 197 197   
## Labour\_Force\_Total Tuberculosis\_Per\_100000   
## 140 210   
## Suicide\_Rate\_Percent Obesity\_Rate\_Percent   
## 608 230   
## Underweight\_Rate\_Percent Overweight\_Rate\_Percent   
## 230 230   
## Safe\_Water\_Access\_Percent Continent   
## 232 0

# Check distributions using histograms (raw data)  
par(mfrow = c(3,2)) # Arrange plots in a 2x2 grid  
hist(global\_health$Life\_Expectancy\_Male, main = "Life Expectancy Male", col = "lightblue", breaks = 30)  
hist(global\_health$Life\_Expectancy\_Female, main = "Life Expectancy Female", col = "lightgreen", breaks = 30)  
hist(global\_health$GDP\_Per\_Capita, main = "GDP\_Per\_Capita", col = "lightcoral", breaks = 30)  
hist(global\_health$Urban\_Population\_Percent, main = "Urban\_Population\_Percent", col = "lightgray", breaks = 30)  
hist(global\_health$Unemployment\_Rate, main = "Unemployment\_Rate", col = "lightpink", breaks = 30)  
hist(global\_health$Safe\_Water\_Access\_Percent, main = "Safe\_Water\_Access\_Percent", col = "lightyellow", breaks = 30)



library(dplyr)  
library(tidyr)

## Warning: package 'tidyr' was built under R version 4.4.2

data\_filled <- global\_health %>%  
 select(Country, Life\_Expectancy, Life\_Expectancy\_Female, Life\_Expectancy\_Male, GDP\_Per\_Capita, Unemployment\_Rate, Safe\_Water\_Access\_Percent, Urban\_Population\_Percent) %>%  
 group\_by(Country) %>%  
 mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .))) %>%  
 ungroup()

summary(data\_filled)

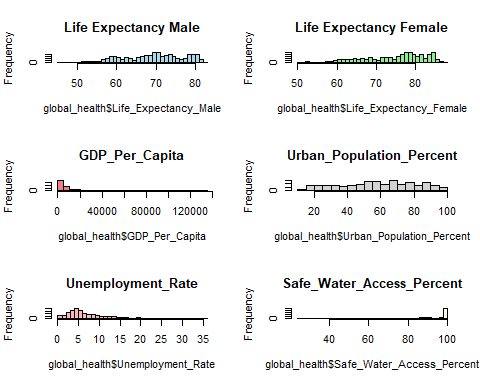
## Country Life\_Expectancy Life\_Expectancy\_Female Life\_Expectancy\_Male  
## Length:1880 Min. :47.84 Min. :50.49 Min. :45.36   
## Class :character 1st Qu.:65.37 1st Qu.:68.06 1st Qu.:62.96   
## Mode :character Median :72.55 Median :76.09 Median :69.30   
## Mean :71.44 Mean :74.10 Mean :68.89   
## 3rd Qu.:77.42 3rd Qu.:80.40 3rd Qu.:74.68   
## Max. :84.56 Max. :87.71 Max. :82.60   
## NA's :40 NA's :40 NA's :40   
## GDP\_Per\_Capita Unemployment\_Rate Safe\_Water\_Access\_Percent  
## Min. : 216.8 Min. : 0.100 Min. : 25.61   
## 1st Qu.: 1959.3 1st Qu.: 3.716 1st Qu.: 65.16   
## Median : 5872.4 Median : 5.811 Median : 91.06   
## Mean : 15833.4 Mean : 7.684 Mean : 81.74   
## 3rd Qu.: 17401.3 3rd Qu.:10.245 3rd Qu.: 99.67   
## Max. :235132.8 Max. :35.707 Max. :100.00   
## NA's :10 NA's :140 NA's :210   
## Urban\_Population\_Percent  
## Min. : 11.19   
## 1st Qu.: 40.07   
## Median : 58.88   
## Mean : 58.62   
## 3rd Qu.: 77.63   
## Max. :100.00   
##

global\_health <- data\_filled[complete.cases(data\_filled), ]

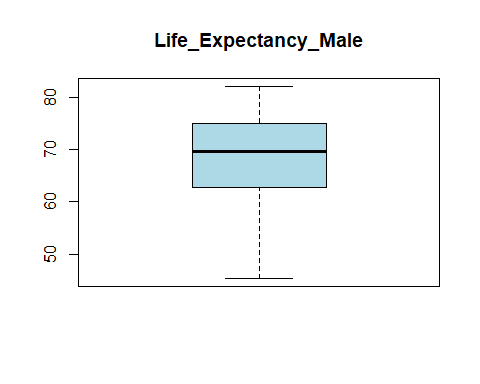
# Summary statistics after imputation  
summary(global\_health)

## Country Life\_Expectancy Life\_Expectancy\_Female Life\_Expectancy\_Male  
## Length:1540 Min. :47.84 Min. :50.49 Min. :45.36   
## Class :character 1st Qu.:65.31 1st Qu.:67.97 1st Qu.:62.82   
## Mode :character Median :72.81 Median :76.36 Median :69.66   
## Mean :71.50 Mean :74.13 Mean :68.98   
## 3rd Qu.:77.61 3rd Qu.:80.68 3rd Qu.:74.96   
## Max. :84.56 Max. :87.71 Max. :82.10   
## GDP\_Per\_Capita Unemployment\_Rate Safe\_Water\_Access\_Percent  
## Min. : 216.8 Min. : 0.100 Min. : 25.61   
## 1st Qu.: 1815.6 1st Qu.: 3.730 1st Qu.: 64.35   
## Median : 5650.9 Median : 5.811 Median : 90.33   
## Mean : 13930.1 Mean : 7.710 Mean : 81.16   
## 3rd Qu.: 16143.4 3rd Qu.:10.203 3rd Qu.: 99.68   
## Max. :133711.8 Max. :35.707 Max. :100.00   
## Urban\_Population\_Percent  
## Min. : 11.19   
## 1st Qu.: 40.15   
## Median : 58.12   
## Mean : 57.98   
## 3rd Qu.: 77.16   
## Max. :100.00

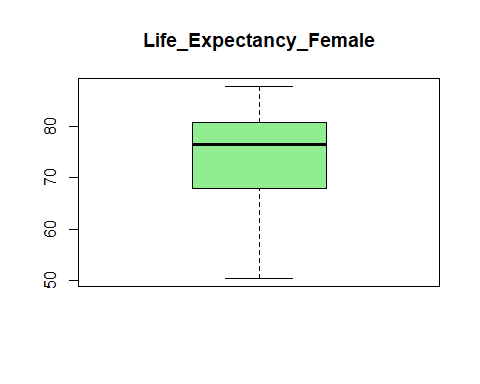
# Check distributions using histograms (raw data)  
par(mfrow = c(3,2)) # Arrange plots in a 2x2 grid  
hist(global\_health$Life\_Expectancy\_Male, main = "Life Expectancy Male", col = "lightblue", breaks = 30)  
hist(global\_health$Life\_Expectancy\_Female, main = "Life Expectancy Female", col = "lightgreen", breaks = 30)  
hist(global\_health$GDP\_Per\_Capita, main = "GDP\_Per\_Capita", col = "lightcoral", breaks = 30)  
hist(global\_health$Urban\_Population\_Percent, main = "Urban\_Population\_Percent", col = "lightgray", breaks = 30)  
hist(global\_health$Unemployment\_Rate, main = "Unemployment\_Rate", col = "lightpink", breaks = 30)  
hist(global\_health$Safe\_Water\_Access\_Percent, main = "Safe\_Water\_Access\_Percent", col = "lightyellow", breaks = 30)



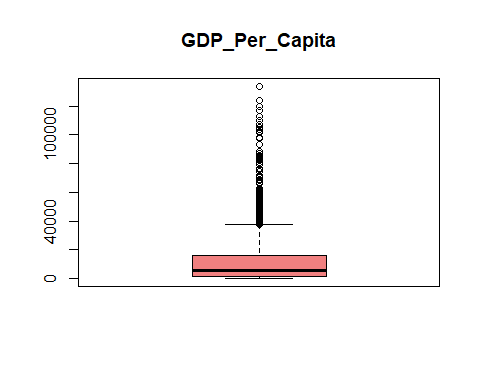
# Boxplots to detect outliers  
boxplot(global\_health$Life\_Expectancy\_Male, main = "Life\_Expectancy\_Male", col = "lightblue")



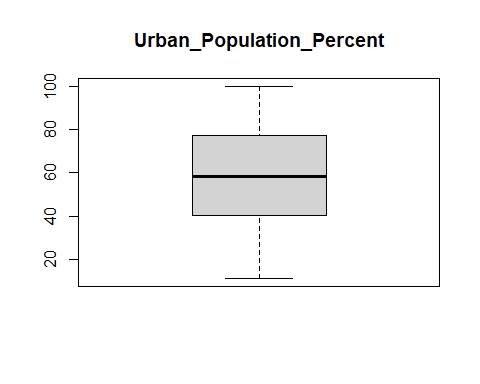
boxplot(global\_health$Life\_Expectancy\_Female, main = "Life\_Expectancy\_Female", col = "lightgreen")



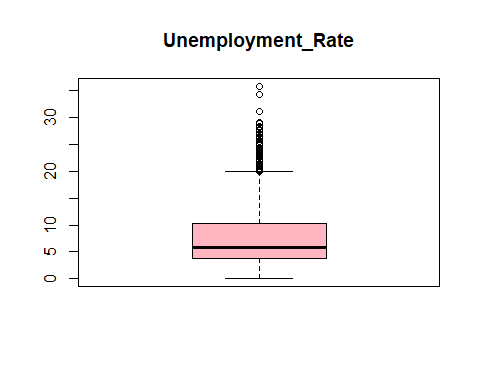
boxplot(global\_health$GDP\_Per\_Capita, main = "GDP\_Per\_Capita", col = "lightcoral")



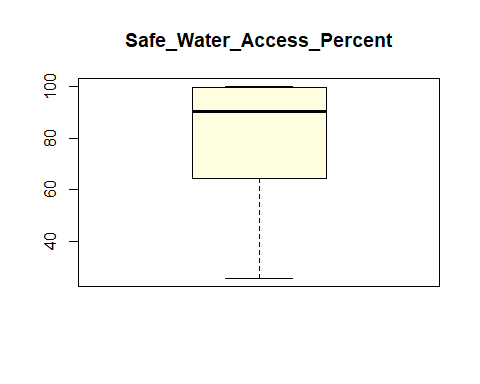
boxplot(global\_health$Urban\_Population\_Percent, main = "Urban\_Population\_Percent", col = "lightgray")



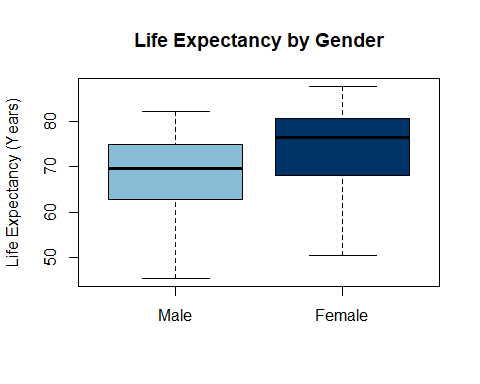
boxplot(global\_health$Unemployment\_Rate, main = "Unemployment\_Rate", col = "lightpink")



boxplot(global\_health$Safe\_Water\_Access\_Percent, main = "Safe\_Water\_Access\_Percent", col = "lightyellow")



# Combined boxplot for Male and Female Life Expectancy  
boxplot(global\_health$Life\_Expectancy\_Male,  
 global\_health$Life\_Expectancy\_Female,  
 names = c("Male", "Female"),  
 main = "Life Expectancy by Gender",  
 col = c("#88BBD6", "#003366"),  
 ylab = "Life Expectancy (Years)")



# Categorías: Low (≤75), High (>75)  
global\_health$Life\_Expectancy\_Category\_Male <- cut(global\_health$Life\_Expectancy\_Male,  
 breaks = c(-Inf, 75, Inf),  
 labels = c("Low", "High"))  
  
global\_health$Life\_Expectancy\_Category\_Female <- cut(global\_health$Life\_Expectancy\_Female,  
 breaks = c(-Inf, 75, Inf),  
 labels = c("Low", "High"))

# GDP per Capita (Low / High based on median)  
global\_health$GDP\_Category <- ifelse(global\_health$GDP\_Per\_Capita < median(global\_health$GDP\_Per\_Capita, na.rm = TRUE),  
 "Low GDP", "High GDP")  
  
# Urban Population  
global\_health$Urban\_Category <- ifelse(global\_health$Urban\_Population\_Percent < median(global\_health$Urban\_Population\_Percent, na.rm = TRUE),  
 "Low Urban", "High Urban")  
  
# Unemployment Rate  
global\_health$Unemployment\_Category <- ifelse(global\_health$Unemployment\_Rate < median(global\_health$Unemployment\_Rate, na.rm = TRUE),  
 "Low Unemployment Rate", "High Unemployment Rate")  
  
# Safe\_Water\_Access\_Percent  
global\_health$Safe\_Water\_Access\_Category <- ifelse(global\_health$Safe\_Water\_Access\_Percent < median(global\_health$Safe\_Water\_Access\_Percent, na.rm = TRUE),  
 "Low Safe Water Access", "High Safe Water Access")

# Female

## GDP Category vs. Female Life Expectancy

# Create contingency table  
table\_female\_gdp <- table(global\_health$Life\_Expectancy\_Category\_Female,  
 global\_health$GDP\_Category)  
table\_female\_gdp

##   
## High GDP Low GDP  
## Low 84 596  
## High 686 174

# Perform Chi-Square Test  
chisq.test(table\_female\_gdp, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_female\_gdp  
## X-squared = 690.32, df = 1, p-value < 2.2e-16

### Interpretation

-Since the p-value is far less than 0.05, we reject the null hypothesis. -This means there is a statistically significant association between GDP category and female life expectancy. -Countries with higher GDP are more likely to have higher female life expectancy.

## Urban Population Category vs. Female Life Expectancy

table\_female\_urban <- table(global\_health$Life\_Expectancy\_Category\_Female,  
 global\_health$Urban\_Category)  
table\_female\_urban

##   
## High Urban Low Urban  
## Low 108 572  
## High 662 198

chisq.test(table\_female\_urban,correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_female\_urban  
## X-squared = 566.96, df = 1, p-value < 2.2e-16

### Interpretation

-Since p < 0.05, you reject the null hypothesis. -There is a statistically significant relationship between urban population level and female life expectancy. -Countries with higher urbanization tend to have higher female life expectancy.

## Unemployment Category vs. Female Life Expectancy

table\_female\_unemp <- table(global\_health$Life\_Expectancy\_Category\_Female,  
 global\_health$Unemployment\_Category)  
table\_female\_unemp

##   
## High Unemployment Rate Low Unemployment Rate  
## Low 280 400  
## High 490 370

chisq.test(table\_female\_unemp, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_female\_unemp  
## X-squared = 37.921, df = 1, p-value = 7.368e-10

### Interpretion

-p < 0.05, so you reject the null hypothesis. -There is a significant association between unemployment rate and female life expectancy. -Even among high unemployment countries, many still have high life expectancy, but the association remains statistically significant.

## Safe Water Access Category vs. Female Life Expectancy

table\_female\_water <- table(global\_health$Life\_Expectancy\_Category\_Female,  
 global\_health$Safe\_Water\_Access\_Category)  
table\_female\_water

##   
## High Safe Water Access Low Safe Water Access  
## Low 96 584  
## High 674 186

chisq.test(table\_female\_water, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_female\_water  
## X-squared = 627.12, df = 1, p-value < 2.2e-16

### Interpretation

-There is a statistically significant association between safe water access and female life expectancy (p < 0.05). -This means that countries with high access to safe water tend to have higher female life expectancy. -In countries where people have better access to safe water, women are more likely to live longer.

# Male

## GDP Category vs. Male Life Expectancy

# Create contingency table  
table\_Male\_gdp <- table(global\_health$Life\_Expectancy\_Category\_Male,  
 global\_health$GDP\_Category)  
table\_Male\_gdp

##   
## High GDP Low GDP  
## Low 397 760  
## High 373 10

# Perform Chi-Square Test  
chisq.test(table\_Male\_gdp, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_Male\_gdp  
## X-squared = 457.93, df = 1, p-value < 2.2e-16

### Interpretation

-There is a statistically significant association between GDP per capita and male life expectancy. -Countries with higher GDP are far more likely to have higher male life expectancy. -In contrast, low GDP countries are overwhelmingly concentrated in the low life expectancy group.

## Urban Population Category vs. Male Life Expectancy

table\_Male\_urban <- table(global\_health$Life\_Expectancy\_Category\_Male,  
 global\_health$Urban\_Category)  
table\_Male\_urban

##   
## High Urban Low Urban  
## Low 424 733  
## High 346 37

chisq.test(table\_Male\_urban, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_Male\_urban  
## X-squared = 331.82, df = 1, p-value < 2.2e-16

### Interpretation

-There is a statistically significant association between urban population and male life expectancy. -Countries with high urbanization are much more likely to have higher male life expectancy. -On the other hand, low urbanization is strongly associated with low life expectancy in males.

## Unemployment Category vs. Male Life Expectancy

table\_Male\_unemp <- table(global\_health$Life\_Expectancy\_Category\_Male,  
 global\_health$Unemployment\_Category)  
table\_Male\_unemp

##   
## High Unemployment Rate Low Unemployment Rate  
## Low 581 576  
## High 189 194

chisq.test(table\_Male\_unemp, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_Male\_unemp  
## X-squared = 0.086882, df = 1, p-value = 0.7682

### Interpretation

-There is no statistically significant association between unemployment rate and male life expectancy. -The p-value (0.7682) is much greater than 0.05, indicating that male life expectancy appears to be independent of unemployment levels in this dataset.

## Safe Water Access Category vs. Male Life Expectancy

table\_Male\_water <- table(global\_health$Life\_Expectancy\_Category\_Male,  
 global\_health$Safe\_Water\_Access\_Category)  
table\_Male\_water

##   
## High Safe Water Access Low Safe Water Access  
## Low 411 746  
## High 359 24

chisq.test(table\_Male\_water, correct = FALSE)

##   
## Pearson's Chi-squared test  
##   
## data: table\_Male\_water  
## X-squared = 390.01, df = 1, p-value < 2.2e-16

### Interpretation

-There is a statistically significant association between safe water access and male life expectancy. -Countries with high access to safe water are much more likely to have higher male life expectancy. -Countries with poor water access tend to have lower male life expectancy.

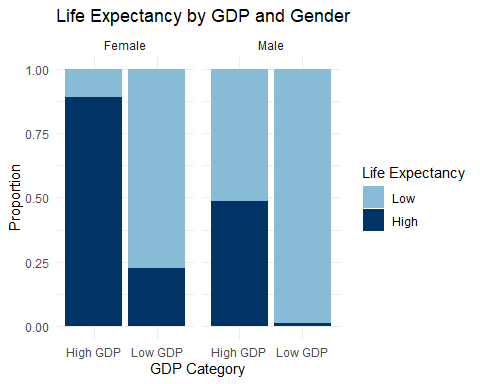
## Visualizations

library(ggplot2)

# Combine male and female data for faceted plotting  
library(dplyr)  
  
data\_female <- global\_health %>%  
 select(GDP\_Category, Urban\_Category, Unemployment\_Category, Safe\_Water\_Access\_Category,  
 Life\_Expectancy\_Category\_Female) %>%  
 mutate(Gender = "Female",  
 Life\_Expectancy\_Category = Life\_Expectancy\_Category\_Female)  
  
data\_male <- global\_health %>%  
 select(GDP\_Category, Urban\_Category, Unemployment\_Category, Safe\_Water\_Access\_Category,  
 Life\_Expectancy\_Category\_Male) %>%  
 mutate(Gender = "Male",  
 Life\_Expectancy\_Category = Life\_Expectancy\_Category\_Male)  
  
# Combine into one dataset  
gender\_data <- bind\_rows(data\_female, data\_male)

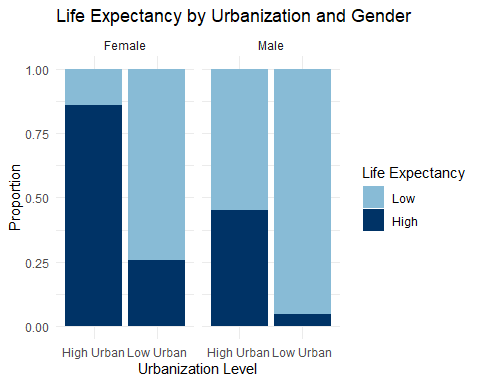
## Comparative Plot – Life Expectancy by GDP and Gender

ggplot(gender\_data, aes(x = GDP\_Category, fill = Life\_Expectancy\_Category)) +  
 geom\_bar(position = "fill") +  
 facet\_wrap(~Gender) +  
 labs(title = "Life Expectancy by GDP and Gender",  
 x = "GDP Category",  
 y = "Proportion",  
 fill = "Life Expectancy") +  
 scale\_fill\_manual(values = c("Low" = "#88BBD6", "High" = "#003366")) +  
 theme\_minimal()



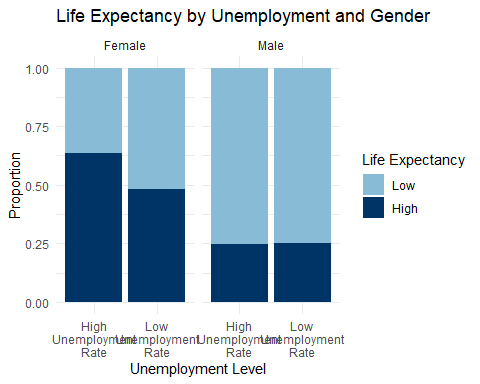
## Comparative Plot – Life Expectancy by Urbanization and Gender

ggplot(gender\_data, aes(x = Urban\_Category, fill = Life\_Expectancy\_Category)) +  
 geom\_bar(position = "fill") +  
 facet\_wrap(~Gender) +  
 labs(title = "Life Expectancy by Urbanization and Gender",  
 x = "Urbanization Level",  
 y = "Proportion",  
 fill = "Life Expectancy") +  
 scale\_fill\_manual(values = c("Low" = "#88BBD6",  
 "High" = "#003366")) +  
 theme\_minimal()



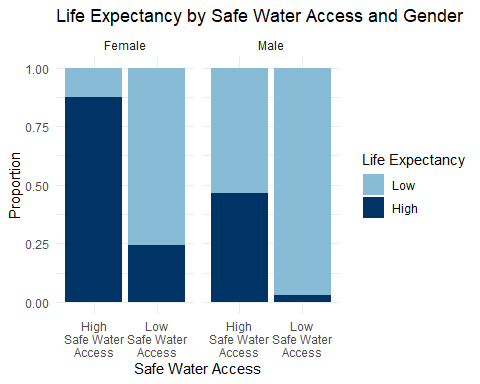
## Comparative Plot – Life Expectancy by Unemployment and Gender

gender\_data$Unemployment\_Category <- ifelse(  
 gender\_data$Unemployment\_Category == "High Unemployment Rate",   
 "High\nUnemployment\nRate",   
 "Low\nUnemployment\nRate"  
)  
  
ggplot(gender\_data, aes(x = Unemployment\_Category, fill = Life\_Expectancy\_Category)) +  
 geom\_bar(position = "fill") +  
 facet\_wrap(~Gender) +  
 labs(title = "Life Expectancy by Unemployment and Gender",  
 x = "Unemployment Level",  
 y = "Proportion",  
 fill = "Life Expectancy") +  
 scale\_fill\_manual(values = c("Low" = "#88BBD6", "High" = "#003366")) +  
 theme\_minimal()



## Comparative Plot – Life Expectancy by Safe Water Access and Gender

gender\_data$Safe\_Water\_Access\_Category <- ifelse(  
 gender\_data$Safe\_Water\_Access\_Category == "High Safe Water Access",  
 "High\nSafe Water\nAccess",  
 "Low\nSafe Water\nAccess"  
)  
  
ggplot(gender\_data, aes(x = Safe\_Water\_Access\_Category, fill = Life\_Expectancy\_Category)) +  
 geom\_bar(position = "fill") +  
 facet\_wrap(~Gender) +  
 labs(title = "Life Expectancy by Safe Water Access and Gender",  
 x = "Safe Water Access",  
 y = "Proportion",  
 fill = "Life Expectancy") +  
 scale\_fill\_manual(values = c("Low" = "#88BBD6", "High" = "#003366"))+  
 theme\_minimal()



# GLM

model\_female <- glm(Life\_Expectancy\_Category\_Female ~ GDP\_Category + Urban\_Category +  
 Safe\_Water\_Access\_Category + Unemployment\_Category,  
 family = binomial(link="logit"), data = global\_health)  
  
summary(model\_female)

##   
## Call:  
## glm(formula = Life\_Expectancy\_Category\_Female ~ GDP\_Category +   
## Urban\_Category + Safe\_Water\_Access\_Category + Unemployment\_Category,   
## family = binomial(link = "logit"), data = global\_health)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 3.5012 0.1811 19.329  
## GDP\_CategoryLow GDP -1.4576 0.1820 -8.009  
## Urban\_CategoryLow Urban -2.2627 0.1940 -11.660  
## Safe\_Water\_Access\_CategoryLow Safe Water Access -2.4548 0.1829 -13.425  
## Unemployment\_CategoryLow Unemployment Rate 0.2587 0.1748 1.480  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## GDP\_CategoryLow GDP 1.15e-15 \*\*\*  
## Urban\_CategoryLow Urban < 2e-16 \*\*\*  
## Safe\_Water\_Access\_CategoryLow Safe Water Access < 2e-16 \*\*\*  
## Unemployment\_CategoryLow Unemployment Rate 0.139   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2113.8 on 1539 degrees of freedom  
## Residual deviance: 1003.5 on 1535 degrees of freedom  
## AIC: 1013.5  
##   
## Number of Fisher Scoring iterations: 5

exp(coef(model\_female))

## (Intercept)   
## 33.15378651   
## GDP\_CategoryLow GDP   
## 0.23278457   
## Urban\_CategoryLow Urban   
## 0.10407244   
## Safe\_Water\_Access\_CategoryLow Safe Water Access   
## 0.08588118   
## Unemployment\_CategoryLow Unemployment Rate   
## 1.29518184

### Interpretation

-Holding all other variables constant, the odds of having high female life expectancy in countries with low GDP are approximately 77% lower than in countries with high GDP. Strong evidence of a GDP effect (p < 0.001).  
-Countries with low urbanization have about 90% lower odds of high female life expectancy compared to highly urbanized countries, after controlling for other variables. Strong urbanization effect (p < 0.001).  
-The odds of high female life expectancy in countries with poor access to safe water are about 91% lower than those with good access, controlling for GDP, urbanization, and unemployment. Very strong effect (p < 0.001).  
-Unemployment effect is not statistically significant (p = 0.139).

model\_male <- glm(Life\_Expectancy\_Category\_Male ~ GDP\_Category + Urban\_Category +  
 Safe\_Water\_Access\_Category + Unemployment\_Category,  
 family = binomial(link="logit"), data = global\_health)  
  
summary(model\_male)

##   
## Call:  
## glm(formula = Life\_Expectancy\_Category\_Male ~ GDP\_Category +   
## Urban\_Category + Safe\_Water\_Access\_Category + Unemployment\_Category,   
## family = binomial(link = "logit"), data = global\_health)  
##   
## Coefficients:  
## Estimate Std. Error z value  
## (Intercept) 0.3486 0.1126 3.096  
## GDP\_CategoryLow GDP -2.6683 0.3417 -7.809  
## Urban\_CategoryLow Urban -1.8758 0.2132 -8.799  
## Safe\_Water\_Access\_CategoryLow Safe Water Access -2.4044 0.2419 -9.938  
## Unemployment\_CategoryLow Unemployment Rate 0.7705 0.1650 4.671  
## Pr(>|z|)   
## (Intercept) 0.00196 \*\*   
## GDP\_CategoryLow GDP 5.79e-15 \*\*\*  
## Urban\_CategoryLow Urban < 2e-16 \*\*\*  
## Safe\_Water\_Access\_CategoryLow Safe Water Access < 2e-16 \*\*\*  
## Unemployment\_CategoryLow Unemployment Rate 3.00e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1727.6 on 1539 degrees of freedom  
## Residual deviance: 957.0 on 1535 degrees of freedom  
## AIC: 967  
##   
## Number of Fisher Scoring iterations: 7

exp(coef(model\_male))

## (Intercept)   
## 1.41713495   
## GDP\_CategoryLow GDP   
## 0.06937110   
## Urban\_CategoryLow Urban   
## 0.15323762   
## Safe\_Water\_Access\_CategoryLow Safe Water Access   
## 0.09031939   
## Unemployment\_CategoryLow Unemployment Rate   
## 2.16092463

### Interpretation

-Holding all other variables constant, countries with low GDP have 93% lower odds of high male life expectancy compared to those with high GDP. Strong GDP effect (p < 0.001).  
-Countries with low urbanization have 85% lower odds of high male life expectancy than highly urbanized countries, adjusting for other variables. Strong urbanization effect (p < 0.001).  
-Countries with poor access to safe water have 91% lower odds of high male life expectancy than countries with good access. Very strong effect (p < 0.001).  
-Countries with low unemployment have more than twice the odds (OR ≈ 2.16) of high male life expectancy compared to those with high unemployment. This time, the effect is statistically significant (p < 0.001).