Project\_Objective 3

2025-03-25

# Load required packages  
library(survival)  
library(ggplot2)

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

library(dplyr)

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

##   
## Attaching package: 'dplyr'

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

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

library(tidyr)

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

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.4.3

## corrplot 0.95 loaded

library(survminer)

## Warning: package 'survminer' was built under R version 4.4.3

## Loading required package: ggpubr

## Warning: package 'ggpubr' was built under R version 4.4.3

##   
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':  
##   
## myeloma

library(MASS)

## Warning: package 'MASS' was built under R version 4.4.3

##   
## Attaching package: 'MASS'

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

# Load dataset  
data <- read.csv("global\_health.csv")  
summary(data)

## 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

data\_filled <- data %>%  
 dplyr::select(Country, CO2\_Exposure\_Percent, Fertility\_Rate, Sanitary\_Expense\_Per\_GDP, Unemployment\_Rate, Safe\_Water\_Access\_Percent, Immunization\_Rate, Infant\_Deaths, Hospital\_Beds\_Per\_1000) %>%  
 group\_by(Country) %>%  
 mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .))) %>%  
 ungroup()  
  
summary(data\_filled)

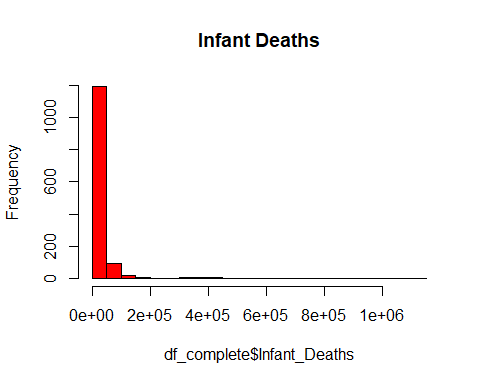
## Country CO2\_Exposure\_Percent Fertility\_Rate   
## Length:1880 Min. : 4.895 Min. :0.808   
## Class :character 1st Qu.: 13.877 1st Qu.:1.667   
## Mode :character Median : 21.183 Median :2.293   
## Mean : 26.323 Mean :2.754   
## 3rd Qu.: 32.449 3rd Qu.:3.623   
## Max. :107.145 Max. :7.400   
## NA's :10 NA's :20   
## Sanitary\_Expense\_Per\_GDP Unemployment\_Rate Safe\_Water\_Access\_Percent  
## Min. : 1.515 Min. : 0.100 Min. : 25.61   
## 1st Qu.: 4.382 1st Qu.: 3.716 1st Qu.: 65.16   
## Median : 6.236 Median : 5.811 Median : 91.06   
## Mean : 6.620 Mean : 7.684 Mean : 81.74   
## 3rd Qu.: 8.422 3rd Qu.:10.245 3rd Qu.: 99.67   
## Max. :24.283 Max. :35.707 Max. :100.00   
## NA's :20 NA's :140 NA's :210   
## Immunization\_Rate Infant\_Deaths Hospital\_Beds\_Per\_1000  
## Min. :19.00 Min. : 0 Min. : 0.170   
## 1st Qu.:84.00 1st Qu.: 250 1st Qu.: 1.150   
## Median :93.00 Median : 1996 Median : 2.440   
## Mean :87.94 Mean : 22859 Mean : 3.039   
## 3rd Qu.:97.00 3rd Qu.: 15004 3rd Qu.: 4.130   
## Max. :99.00 Max. :1104742 Max. :22.300   
## NA's :10 NA's :10 NA's :300

#View(data\_filled)

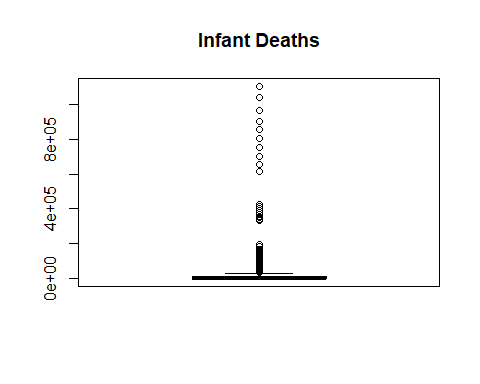
df\_complete <- data\_filled[complete.cases(data\_filled) &  
 !apply(data\_filled, 1, function(x) any(is.infinite(x))), ]  
#View(df\_complete)  
summary(df\_complete)

## Country CO2\_Exposure\_Percent Fertility\_Rate   
## Length:1330 Min. : 4.895 Min. :1.100   
## Class :character 1st Qu.: 14.537 1st Qu.:1.609   
## Mode :character Median : 20.917 Median :2.086   
## Mean : 26.651 Mean :2.522   
## 3rd Qu.: 32.152 3rd Qu.:2.959   
## Max. :107.145 Max. :7.400   
## Sanitary\_Expense\_Per\_GDP Unemployment\_Rate Safe\_Water\_Access\_Percent  
## Min. : 1.748 Min. : 0.100 Min. : 25.61   
## 1st Qu.: 4.450 1st Qu.: 3.692 1st Qu.: 73.28   
## Median : 6.346 Median : 5.822 Median : 91.94   
## Mean : 6.498 Mean : 7.474 Mean : 84.03   
## 3rd Qu.: 8.326 3rd Qu.: 9.842 3rd Qu.: 99.82   
## Max. :21.828 Max. :35.707 Max. :100.00   
## Immunization\_Rate Infant\_Deaths Hospital\_Beds\_Per\_1000  
## Min. :19.00 Min. : 8.0 Min. : 0.170   
## 1st Qu.:87.00 1st Qu.: 275.5 1st Qu.: 1.103   
## Median :94.00 Median : 1685.0 Median : 2.315   
## Mean :89.34 Mean : 21817.1 Mean : 2.896   
## 3rd Qu.:97.00 3rd Qu.: 12634.2 3rd Qu.: 4.200   
## Max. :99.00 Max. :1104742.0 Max. :13.330

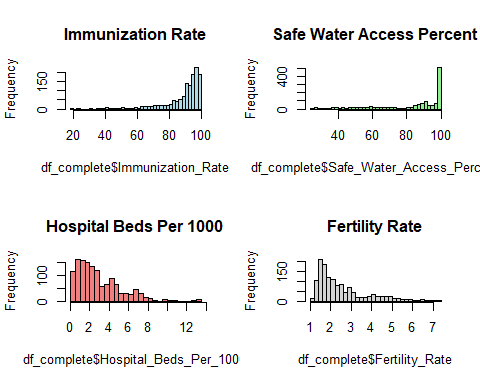
hist(df\_complete$Infant\_Deaths, main = "Infant Deaths", col = "red", breaks = 30)



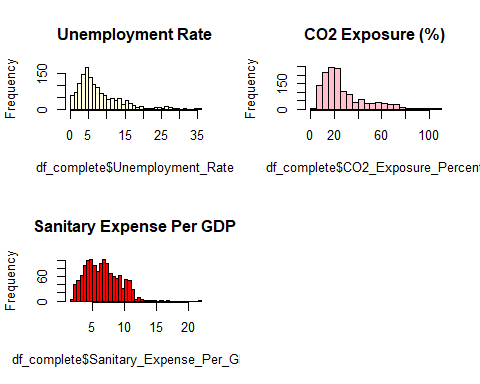
boxplot(df\_complete$Infant\_Deaths, main = "Infant Deaths", col = "red", breaks = 30)



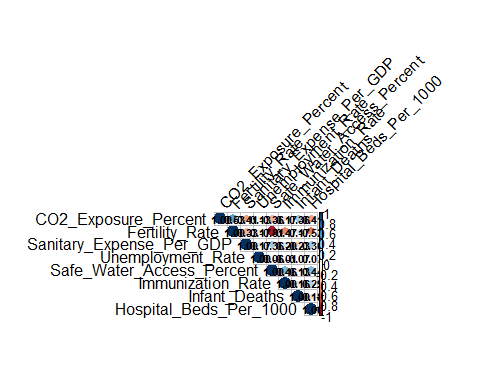
# Check distributions using histograms   
par(mfrow = c(2,2)) # Arrange plots in a 2x2 grid  
hist(df\_complete$Immunization\_Rate, main = "Immunization Rate", col = "lightblue", breaks = 30)  
hist(df\_complete$Safe\_Water\_Access\_Percent, main = "Safe Water Access Percent", col = "lightgreen", breaks = 30)  
hist(df\_complete$Hospital\_Beds\_Per\_1000, main = "Hospital Beds Per 1000", col = "lightcoral", breaks = 30)  
hist(df\_complete$Fertility\_Rate, main = "Fertility Rate", col = "lightgray", breaks = 30)



hist(df\_complete$Unemployment\_Rate, main = "Unemployment Rate", col = "lightyellow", breaks = 30)  
hist(df\_complete$CO2\_Exposure\_Percent, main = "CO2 Exposure (%)", col = "pink", breaks = 30)  
hist(df\_complete$Sanitary\_Expense\_Per\_GDP, main = "Sanitary Expense Per GDP", col = "red", breaks = 30)



# Reset the plotting parameters  
par(mfrow = c(1,1))  
  
# Create a correlation matrix to see relationships between variables  
numeric\_vars <- df\_complete %>%   
 dplyr::select(CO2\_Exposure\_Percent, Fertility\_Rate, Sanitary\_Expense\_Per\_GDP, Unemployment\_Rate, Safe\_Water\_Access\_Percent, Immunization\_Rate, Infant\_Deaths, Hospital\_Beds\_Per\_1000)  
  
cor\_matrix <- cor(numeric\_vars, use = "complete.obs")  
corrplot(cor\_matrix, method = "circle", type = "upper",   
 tl.col = "black", tl.srt = 45, addCoef.col = "black",   
 number.cex = 0.7)



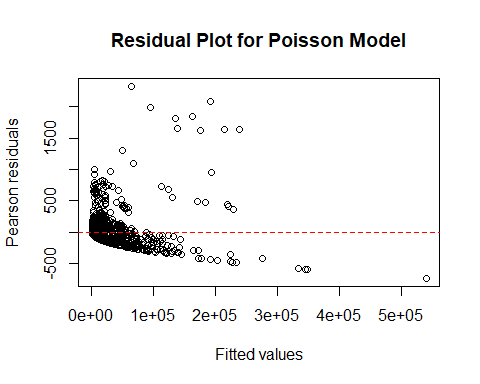
# 1. Fit Poisson Regression Model  
poisson\_model <- glm(Infant\_Deaths ~ CO2\_Exposure\_Percent+ Fertility\_Rate+ Sanitary\_Expense\_Per\_GDP+ Unemployment\_Rate+ Safe\_Water\_Access\_Percent+ Immunization\_Rate+ Hospital\_Beds\_Per\_1000,   
 family = poisson(link = "log"),   
 data = df\_complete)  
  
summary(poisson\_model)

##   
## Call:  
## glm(formula = Infant\_Deaths ~ CO2\_Exposure\_Percent + Fertility\_Rate +   
## Sanitary\_Expense\_Per\_GDP + Unemployment\_Rate + Safe\_Water\_Access\_Percent +   
## Immunization\_Rate + Hospital\_Beds\_Per\_1000, family = poisson(link = "log"),   
## data = df\_complete)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.408e+01 1.951e-03 7215.4 <2e-16 \*\*\*  
## CO2\_Exposure\_Percent 4.284e-02 1.034e-05 4144.2 <2e-16 \*\*\*  
## Fertility\_Rate -3.183e-01 2.328e-04 -1367.6 <2e-16 \*\*\*  
## Sanitary\_Expense\_Per\_GDP -2.137e-01 1.166e-04 -1832.2 <2e-16 \*\*\*  
## Unemployment\_Rate 1.683e-03 3.339e-05 50.4 <2e-16 \*\*\*  
## Safe\_Water\_Access\_Percent -1.027e-02 1.455e-05 -705.7 <2e-16 \*\*\*  
## Immunization\_Rate -2.956e-02 1.424e-05 -2076.4 <2e-16 \*\*\*  
## Hospital\_Beds\_Per\_1000 -1.778e-01 1.854e-04 -959.0 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 105588378 on 1329 degrees of freedom  
## Residual deviance: 52679065 on 1322 degrees of freedom  
## AIC: 52691544  
##   
## Number of Fisher Scoring iterations: 6

# 2. Check for overdispersion in Poisson model  
# Calculate dispersion parameter (should be close to 1 for Poisson)  
dispersion\_param <- sum(residuals(poisson\_model, type = "pearson")^2) / poisson\_model$df.residual  
cat("Dispersion parameter:", dispersion\_param, "\n")

## Dispersion parameter: 57771.43

# Visual check for overdispersion  
plot(fitted(poisson\_model), residuals(poisson\_model, type = "pearson"),  
 xlab = "Fitted values", ylab = "Pearson residuals",  
 main = "Residual Plot for Poisson Model")  
abline(h = 0, col = "red", lty = 2)



# 3. Fit Negative Binomial Regression Model to handle overdispersion  
  
nb\_model <- glm.nb(Infant\_Deaths ~ CO2\_Exposure\_Percent+ Fertility\_Rate+ Sanitary\_Expense\_Per\_GDP+ Unemployment\_Rate+ Safe\_Water\_Access\_Percent+ Immunization\_Rate+ Hospital\_Beds\_Per\_1000,  
 data = df\_complete)  
  
summary(nb\_model)

##   
## Call:  
## glm.nb(formula = Infant\_Deaths ~ CO2\_Exposure\_Percent + Fertility\_Rate +   
## Sanitary\_Expense\_Per\_GDP + Unemployment\_Rate + Safe\_Water\_Access\_Percent +   
## Immunization\_Rate + Hospital\_Beds\_Per\_1000, data = df\_complete,   
## init.theta = 0.4741703822, link = log)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 15.312692 0.542542 28.224 < 2e-16 \*\*\*  
## CO2\_Exposure\_Percent 0.048756 0.002870 16.990 < 2e-16 \*\*\*  
## Fertility\_Rate -0.221179 0.064693 -3.419 0.000629 \*\*\*  
## Sanitary\_Expense\_Per\_GDP -0.098706 0.017980 -5.490 4.02e-08 \*\*\*  
## Unemployment\_Rate -0.073367 0.007357 -9.973 < 2e-16 \*\*\*  
## Safe\_Water\_Access\_Percent -0.022645 0.003529 -6.417 1.39e-10 \*\*\*  
## Immunization\_Rate -0.041239 0.003768 -10.945 < 2e-16 \*\*\*  
## Hospital\_Beds\_Per\_1000 -0.136276 0.021274 -6.406 1.50e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(0.4742) family taken to be 1)  
##   
## Null deviance: 3098.8 on 1329 degrees of freedom  
## Residual deviance: 1701.9 on 1322 degrees of freedom  
## AIC: 25624  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 0.4742   
## Std. Err.: 0.0151   
##   
## 2 x log-likelihood: -25605.8080

dispersion\_param <- sum(residuals(nb\_model, type = "pearson")^2) / nb\_model$df.residual  
cat("Dispersion parameter:", dispersion\_param, "\n")

## Dispersion parameter: 1.972836

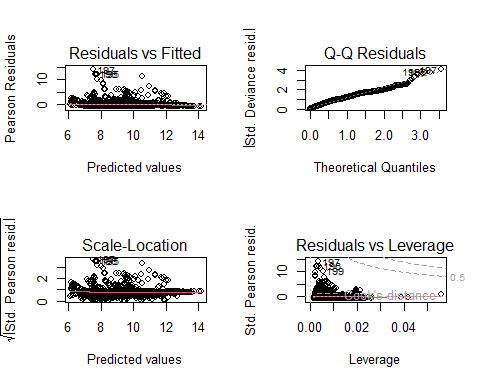
# 4. Compare models using AIC  
cat("Poisson model AIC:", AIC(poisson\_model), "\n")

## Poisson model AIC: 52691544

cat("Negative Binomial model AIC:", AIC(nb\_model), "\n")

## Negative Binomial model AIC: 25623.81

# 6. Diagnostic plots for the better model (assuming negative binomial is better)  
par(mfrow = c(2, 2))  
plot(nb\_model)



par(mfrow = c(1, 1))

# 7. Proportion of Deviance (Model Predictive Power)  
null\_deviance <- nb\_model$null.deviance  
residual\_deviance <- nb\_model$deviance  
  
proportion\_deviance <- (null\_deviance - residual\_deviance)/ null\_deviance  
  
proportion\_deviance

## [1] 0.4507717

#8. Transforming the coefficients (log of expected counts) fo interpretation  
coefficients <- summary(nb\_model)$coefficients  
  
percent\_change <- (exp(coefficients[, 1]) -1) \* 100  
  
percent\_change

## (Intercept) CO2\_Exposure\_Percent Fertility\_Rate   
## 4.469076e+08 4.996431e+00 -1.984265e+01   
## Sanitary\_Expense\_Per\_GDP Unemployment\_Rate Safe\_Water\_Access\_Percent   
## -9.399114e+00 -7.074041e+00 -2.239092e+00   
## Immunization\_Rate Hospital\_Beds\_Per\_1000   
## -4.040061e+00 -1.273985e+01