MADA Course Project

Risk Factors Affecting Life Expectancy

Annabella Hines

4/25/23

## 0.1 Load Required Packages

# 1. Summary/Abstract

With new unique risks arising to human health, along with older ones becoming less prevalent, the factors that tie in to life expectancy are changing over time. Broad vaccine coverage has virtually eliminated the presence of diseases like polio and measles in most developed countries and populations, while air pollution and poor nutrition/weight issues have become an increasing problem for many people. Through analysis of data collect by the World Health Organization on the most prevalent risk factors for death and average life expectancy for each country, I aim to draw initial associations that may be areas for further study or targets of public health measures.

# 2. Introduction

## 2.1 General Background Information

Many factors can play a part in life expectancy in different countries from availability of medical care and vaccines to mental health and happiness.Factors may differ in the weight of their effect depending on developmental status of the country as well, and some previous research has shown the more obvious variables like GDP or health expenditure may not always be as influential as one may assume in these cases (Kabir 2008).

## 2.2 Description of data and data source

I decided to analyze two data sets, one is a life expectancy data set collected by the World Health organization which I obtained at this link https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who from Kaggle. This data specifically looks at immunization rates and the human development index for each country along with their overall life expectancy from. The other data set is Worldwide Deaths by country and risk factor which was also downloaded from Kaggle at https://www.kaggle.com/datasets/varpit94/worldwide-deaths-by-risk-factors. This data is as the title states, with the number of deaths for different risk factors in countries by year. Some examples include unsafe water sources and hygiene, child wasting and thinness, lifestyle and behavioral factors, and pollution.

#load in each dataset  
lifeexp <- readRDS("../../data/processed\_data/processeddata\_life.rds")  
riskfactor <- readRDS("../../data/processed\_data/processeddata\_risk.rds")  
#glimpse of each  
glimpse(lifeexp)

Rows: 2,938  
Columns: 22  
$ Country <chr> "Afghanistan", "Afghanistan", "Afghani…  
$ Year <int> 2015, 2014, 2013, 2012, 2011, 2010, 20…  
$ Status <chr> "Developing", "Developing", "Developin…  
$ Life.expectancy <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, 58…  
$ Adult.Mortality <int> 263, 271, 268, 272, 275, 279, 281, 287…  
$ Infant.deaths <int> 62, 64, 66, 69, 71, 74, 77, 80, 82, 84…  
$ Alcohol <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, 0.…  
$ Percentage.expenditure <dbl> 71.279624, 73.523582, 73.219243, 78.18…  
$ Hepatitis.B <int> 65, 62, 64, 67, 68, 66, 63, 64, 63, 64…  
$ Measles <int> 1154, 492, 430, 2787, 3013, 1989, 2861…  
$ BMI <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, 16…  
$ Under.five.deaths <int> 83, 86, 89, 93, 97, 102, 106, 110, 113…  
$ Polio <int> 6, 58, 62, 67, 68, 66, 63, 64, 63, 58,…  
$ Total.expenditure <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, 9.…  
$ Diphtheria <int> 65, 62, 64, 67, 68, 66, 63, 64, 63, 58…  
$ HIV.AIDS <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1…  
$ GDP <dbl> 584.25921, 612.69651, 631.74498, 669.9…  
$ Population <dbl> 33736494, 327582, 31731688, 3696958, 2…  
$ Thinness.Age1to19 <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, 18…  
$ Thinness.Age5to9 <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, 18…  
$ Income.composition.of.resources <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0.4…  
$ Years.of.education <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9, 8…

glimpse(riskfactor)

Rows: 3,696  
Columns: 31  
$ Country <chr> "Afghanistan", "Afghanistan",…  
$ Year <int> 2000, 2001, 2002, 2003, 2004,…  
$ Unsafe.water.source <dbl> 11121.653280, 10431.364440, 9…  
$ Unsafe.sanitation <dbl> 8588.517136, 8046.884118, 752…  
$ No.access.to.handwashing.facility <dbl> 7910.366272, 7501.548129, 705…  
$ Household.air.pollution.from.solid.fuels <dbl> 28329.3801, 27721.6413, 26679…  
$ Non.exclusive.breastfeeding <dbl> 4566.91865, 4219.42000, 3852.…  
$ Discontinued.breastfeeding <dbl> 235.1014722, 220.5293275, 207…  
$ Child.wasting <dbl> 34320.36460, 32430.50689, 305…  
$ Child.stunting <dbl> 16400.03695, 15848.94208, 150…  
$ Low.birth.weight.for.gestation <dbl> 19140.9054, 19273.2247, 18251…  
$ Secondhand.smoke <dbl> 5641.0470, 5600.9639, 5511.82…  
$ Alcohol.use <dbl> 167.5068, 169.7844, 182.0612,…  
$ Drug.use <dbl> 401.62980, 418.06897, 448.493…  
$ Diet.low.in.fruits <dbl> 10723.7285, 10859.4389, 10891…  
$ Diet.low.in.vegetables <dbl> 9452.4577, 9508.0230, 9475.42…  
$ Unsafe.sex <dbl> 530.28838, 543.74323, 567.856…  
$ Low.physical.activity <dbl> 4829.9912, 4878.3356, 4883.78…  
$ High.fasting.plasma.glucose <dbl> 25116.079, 25350.866, 25410.1…  
$ High.total.cholesterol <dbl> 17654.150, NA, NA, NA, NA, 20…  
$ High.body.mass.index <dbl> 8669.565, 8891.946, 9288.213,…  
$ High.systolic.blood.pressure <dbl> 32603.110, 32800.165, 32771.7…  
$ Smoking <dbl> 7409.526, 7516.432, 7566.578,…  
$ Iron.deficiency <dbl> 1203.0192300, 1201.4087450, 1…  
$ Vitamin.A.deficiency <dbl> 15582.315920, 15256.831640, 1…  
$ Low.bone.mineral.density <dbl> 405.80751, 401.99121, 401.023…  
$ Air.pollution <dbl> 33714.056, 33084.899, 32052.8…  
$ Outdoor.air.pollution <dbl> 5587.68, 5569.97, 5580.75, 61…  
$ Diet.high.in.sodium <dbl> 2928.851, 2924.408, 2929.457,…  
$ Diet.low.in.whole.grains <dbl> 13621.077, 13729.218, 13770.8…  
$ Diet.low.in.nuts.and.seeds <dbl> 8869.1214, 8920.5527, 8901.79…

Above shows a quick overview of some of the variables in each data set. There is a broad range of information that could be investigated for correlations to life expectancy.

## 2.3 Questions/Hypotheses to be addressed

Which variables are the most highly correlated to life expectancy? Are there differences in the most associated factors between the years 2000 and 2015?

# 3. Methods

The data for each set was separated by country and year, with the metrics represented in either percentages or number per 100,000 people. The full explanation of each variable can be found in the supplementary material. Cleaning the data involved matching the years of both data sets and combining them, along with renaming certain variables for more clarification. The analysis starts with some simple one variable linear models focused on the association to the variable of interest, life expectancy. Then a few machine learning approaches were used, including the decision tree model.

## 3.1 Data import and cleaning

The cleaning of the raw data can be found in datacleaning.qmd file in the code folder and processing\_code subfolder. I filtered down the risk factor data set to only include the years from 2000 to 2015 in order to match the life expectancy data. I also renamed variables in each set to make the information clearer and more ubiquitous.

## 3.2 Statistical analysis

Initially some generalized linear models were ran with all variables and with certain subsets of variables. After finding those results, some machine learning methods were used for more specialized conclusions and predictive modeling.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

First, here are summary tables for each data set to see an overview of the data.

#Risk factor dataset summary table  
readRDS(file = "../../results/summarytableriskfactor.rds")

Data summary

|  |  |
| --- | --- |
| Name | risk |
| Number of rows | 3696 |
| Number of columns | 31 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 30 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | 0 | 1 | 4 | 48 | 0 | 231 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 0 | 1.00 | 2007.50 | 4.61 | 2000.00 | 2003.75 | 2007.50 | 2011.25 | 2015.00 | ▇▆▆▆▆ |
| Unsafe.water.source | 0 | 1.00 | 28837.15 | 139635.91 | 0.01 | 9.98 | 235.59 | 4596.37 | 1727905.47 | ▇▁▁▁▁ |
| Unsafe.sanitation | 0 | 1.00 | 20980.85 | 102630.78 | 0.01 | 4.53 | 132.75 | 3364.85 | 1315192.21 | ▇▁▁▁▁ |
| No.access.to.handwashing.facility | 0 | 1.00 | 17326.48 | 82011.84 | 0.10 | 16.31 | 222.42 | 3593.10 | 1030042.30 | ▇▁▁▁▁ |
| Household.air.pollution.from.solid.fuels | 0 | 1.00 | 39437.15 | 170161.09 | 0.02 | 74.98 | 934.27 | 8770.48 | 2314854.16 | ▇▁▁▁▁ |
| Non.exclusive.breastfeeding | 0 | 1.00 | 5094.12 | 23006.68 | 0.00 | 4.14 | 77.50 | 1100.29 | 343019.17 | ▇▁▁▁▁ |
| Discontinued.breastfeeding | 0 | 1.00 | 331.50 | 1506.07 | 0.00 | 0.24 | 4.95 | 63.92 | 22083.23 | ▇▁▁▁▁ |
| Child.wasting | 0 | 1.00 | 35807.42 | 164393.28 | 0.11 | 36.06 | 586.94 | 8763.87 | 2528264.40 | ▇▁▁▁▁ |
| Child.stunting | 0 | 1.00 | 9231.53 | 44769.83 | 0.00 | 1.52 | 57.22 | 1530.04 | 713553.29 | ▇▁▁▁▁ |
| Low.birth.weight.for.gestation | 0 | 1.00 | 28616.51 | 124093.54 | 0.35 | 130.40 | 1109.70 | 8155.71 | 1713464.78 | ▇▁▁▁▁ |
| Secondhand.smoke | 0 | 1.00 | 23477.64 | 96860.55 | 4.29 | 277.07 | 1145.93 | 5715.34 | 1195523.54 | ▇▁▁▁▁ |
| Alcohol.use | 0 | 1.00 | 53525.40 | 207596.75 | -254.67 | 391.86 | 3147.60 | 13234.09 | 2773307.33 | ▇▁▁▁▁ |
| Drug.use | 0 | 1.00 | 9743.42 | 37658.71 | 1.56 | 125.24 | 472.66 | 2394.75 | 549156.84 | ▇▁▁▁▁ |
| Diet.low.in.fruits | 0 | 1.00 | 46713.33 | 189007.82 | 1.58 | 547.64 | 2508.41 | 10726.39 | 2388571.41 | ▇▁▁▁▁ |
| Diet.low.in.vegetables | 0 | 1.00 | 29005.50 | 113097.53 | 0.78 | 417.28 | 1841.96 | 7520.50 | 1445070.77 | ▇▁▁▁▁ |
| Unsafe.sex | 0 | 1.00 | 32606.78 | 143422.32 | 1.44 | 148.35 | 952.07 | 7083.52 | 1771140.67 | ▇▁▁▁▁ |
| Low.physical.activity | 0 | 1.00 | 22038.34 | 85338.61 | 3.50 | 268.34 | 1213.65 | 5866.06 | 1215129.40 | ▇▁▁▁▁ |
| High.fasting.plasma.glucose | 0 | 1.00 | 107906.76 | 411651.80 | 34.90 | 2224.10 | 8424.72 | 36930.53 | 6218325.33 | ▇▁▁▁▁ |
| High.total.cholesterol | 2804 | 0.24 | 52529.56 | 270933.14 | 11.30 | 894.66 | 4002.37 | 17658.72 | 3988170.13 | ▇▁▁▁▁ |
| High.body.mass.index | 0 | 1.00 | 75015.83 | 287371.62 | 32.43 | 1308.25 | 5213.28 | 23645.15 | 4462996.03 | ▇▁▁▁▁ |
| High.systolic.blood.pressure | 0 | 1.00 | 181980.62 | 709770.18 | 31.41 | 2817.56 | 11234.82 | 50903.66 | 10070692.40 | ▇▁▁▁▁ |
| Smoking | 0 | 1.00 | 136120.99 | 541514.83 | 15.14 | 1351.00 | 6024.21 | 31695.73 | 6936294.57 | ▇▁▁▁▁ |
| Iron.deficiency | 0 | 1.00 | 1702.77 | 8078.96 | 0.01 | 1.98 | 25.73 | 408.10 | 115979.06 | ▇▁▁▁▁ |
| Vitamin.A.deficiency | 0 | 1.00 | 9193.29 | 44430.61 | 0.00 | 1.53 | 52.59 | 1611.07 | 724028.15 | ▇▁▁▁▁ |
| Low.bone.mineral.density | 0 | 1.00 | 5041.70 | 20478.76 | 0.48 | 42.07 | 259.81 | 1183.30 | 311543.94 | ▇▁▁▁▁ |
| Air.pollution | 0 | 1.00 | 95253.09 | 388704.11 | 11.65 | 1083.96 | 6140.33 | 22530.25 | 4875230.89 | ▇▁▁▁▁ |
| Outdoor.air.pollution | 0 | 1.00 | 58760.02 | 242221.29 | 7.58 | 599.62 | 2282.81 | 14368.84 | 3318502.52 | ▇▁▁▁▁ |
| Diet.high.in.sodium | 0 | 1.00 | 55914.84 | 251939.68 | 6.02 | 393.27 | 1989.15 | 9811.41 | 3079919.69 | ▇▁▁▁▁ |
| Diet.low.in.whole.grains | 0 | 1.00 | 55389.59 | 217961.41 | 12.84 | 813.88 | 3536.78 | 14886.99 | 2983998.11 | ▇▁▁▁▁ |
| Diet.low.in.nuts.and.seeds | 0 | 1.00 | 36438.14 | 141580.42 | 7.60 | 562.83 | 2347.63 | 10407.50 | 2006782.21 | ▇▁▁▁▁ |

#Life expectancy dataset summary table.  
readRDS("../../results/summarytablelifeexp.rds")

Data summary

|  |  |
| --- | --- |
| Name | life |
| Number of rows | 2938 |
| Number of columns | 22 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 20 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

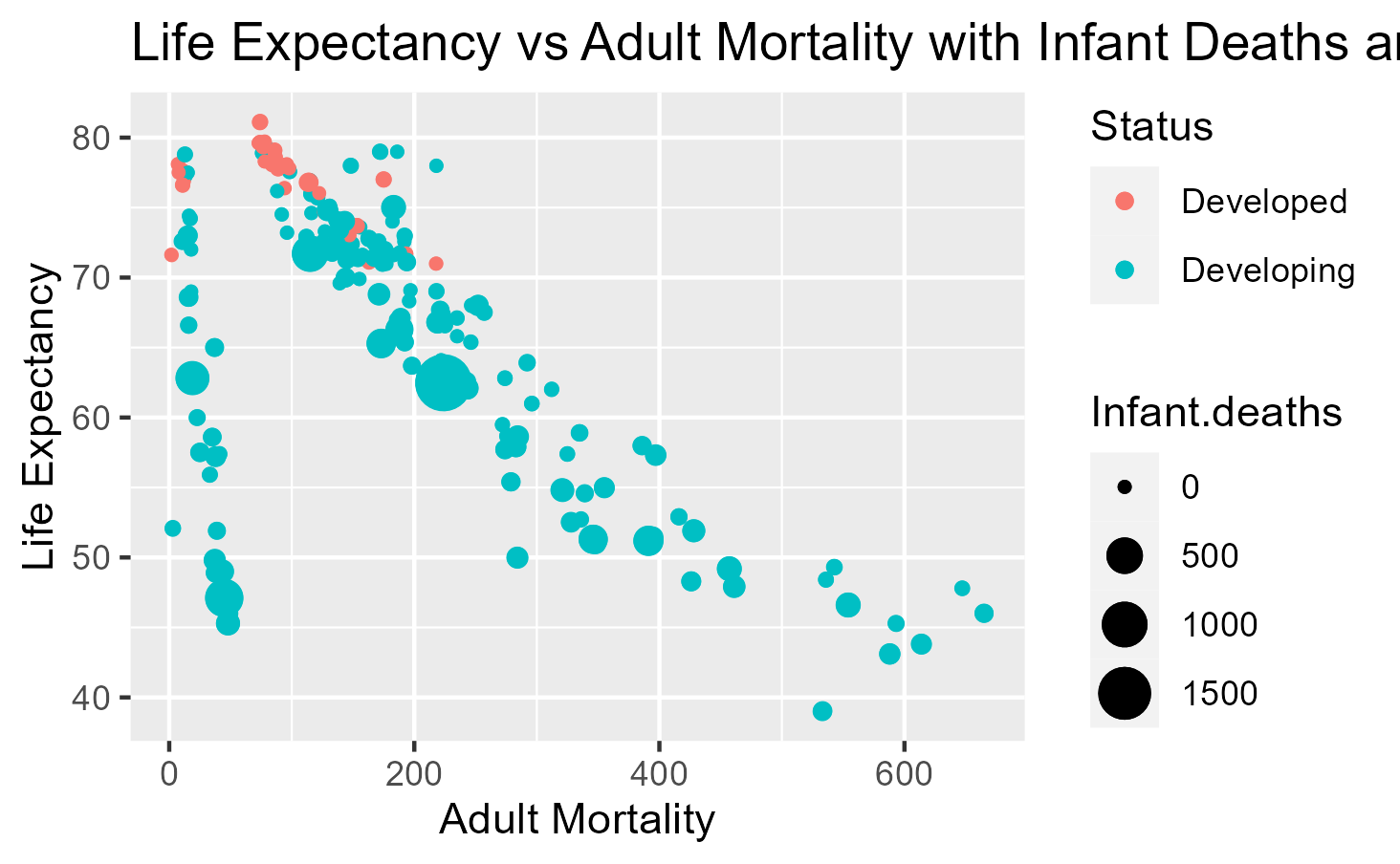
**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | 0 | 1 | 4 | 52 | 0 | 193 | 0 |
| Status | 0 | 1 | 9 | 10 | 0 | 2 | 0 |

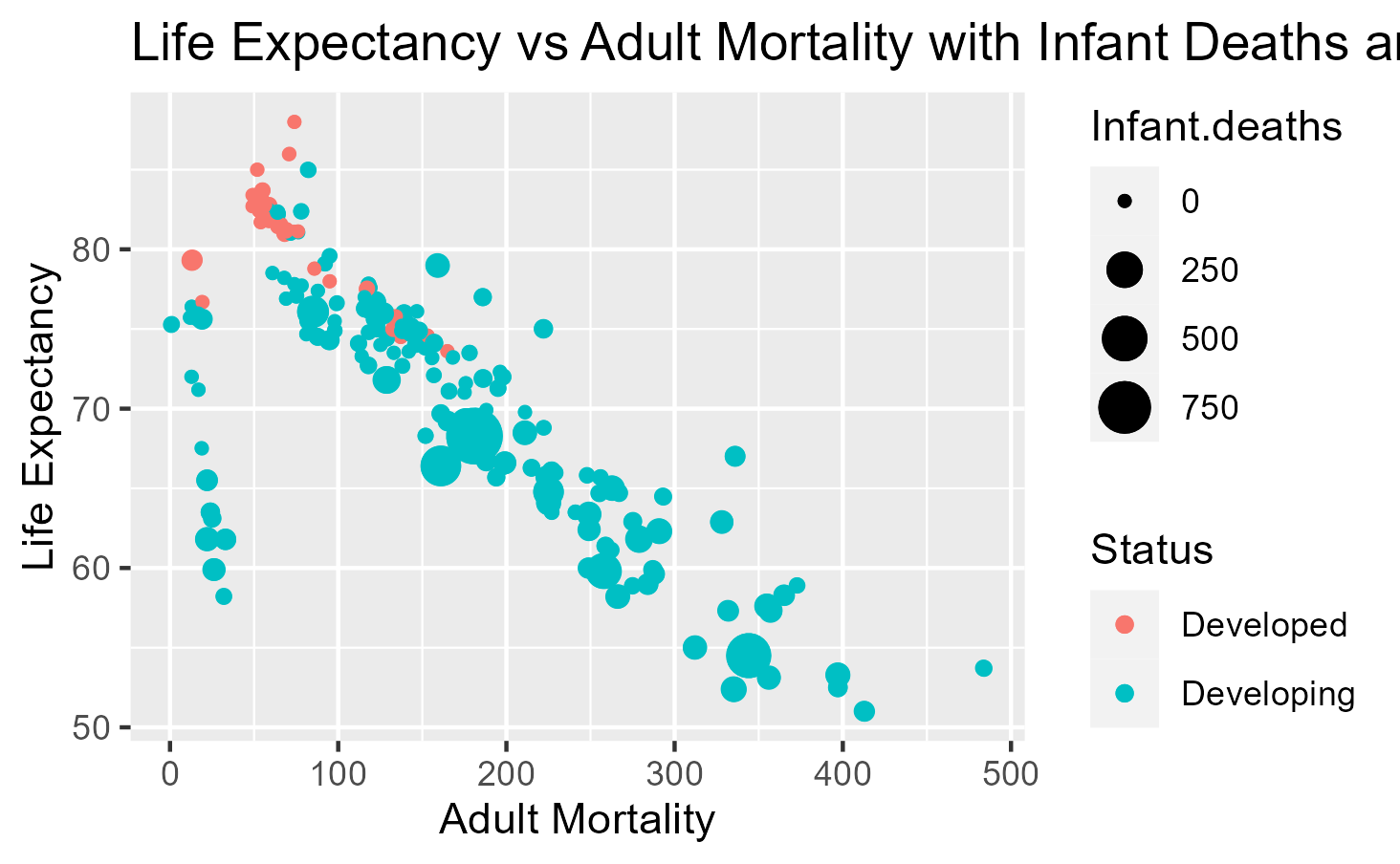
**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 0 | 1.00 | 2007.52 | 4.61 | 2000.00 | 2004.00 | 2008.00 | 2012.00 | 2.015000e+03 | ▇▆▆▆▆ |
| Life.expectancy | 10 | 1.00 | 69.22 | 9.52 | 36.30 | 63.10 | 72.10 | 75.70 | 8.900000e+01 | ▁▂▃▇▂ |
| Adult.Mortality | 10 | 1.00 | 164.80 | 124.29 | 1.00 | 74.00 | 144.00 | 228.00 | 7.230000e+02 | ▇▆▂▁▁ |
| Infant.deaths | 0 | 1.00 | 30.30 | 117.93 | 0.00 | 0.00 | 3.00 | 22.00 | 1.800000e+03 | ▇▁▁▁▁ |
| Alcohol | 194 | 0.93 | 4.60 | 4.05 | 0.01 | 0.88 | 3.76 | 7.70 | 1.787000e+01 | ▇▃▃▂▁ |
| Percentage.expenditure | 0 | 1.00 | 738.25 | 1987.91 | 0.00 | 4.69 | 64.91 | 441.53 | 1.947991e+04 | ▇▁▁▁▁ |
| Hepatitis.B | 553 | 0.81 | 80.94 | 25.07 | 1.00 | 77.00 | 92.00 | 97.00 | 9.900000e+01 | ▁▁▁▂▇ |
| Measles | 0 | 1.00 | 2419.59 | 11467.27 | 0.00 | 0.00 | 17.00 | 360.25 | 2.121830e+05 | ▇▁▁▁▁ |
| BMI | 34 | 0.99 | 38.32 | 20.04 | 1.00 | 19.30 | 43.50 | 56.20 | 8.730000e+01 | ▅▅▅▇▁ |
| Under.five.deaths | 0 | 1.00 | 42.04 | 160.45 | 0.00 | 0.00 | 4.00 | 28.00 | 2.500000e+03 | ▇▁▁▁▁ |
| Polio | 19 | 0.99 | 82.55 | 23.43 | 3.00 | 78.00 | 93.00 | 97.00 | 9.900000e+01 | ▁▁▁▂▇ |
| Total.expenditure | 226 | 0.92 | 5.94 | 2.50 | 0.37 | 4.26 | 5.76 | 7.49 | 1.760000e+01 | ▃▇▃▁▁ |
| Diphtheria | 19 | 0.99 | 82.32 | 23.72 | 2.00 | 78.00 | 93.00 | 97.00 | 9.900000e+01 | ▁▁▁▂▇ |
| HIV.AIDS | 0 | 1.00 | 1.74 | 5.08 | 0.10 | 0.10 | 0.10 | 0.80 | 5.060000e+01 | ▇▁▁▁▁ |
| GDP | 448 | 0.85 | 7483.16 | 14270.17 | 1.68 | 463.94 | 1766.95 | 5910.81 | 1.191727e+05 | ▇▁▁▁▁ |
| Population | 652 | 0.78 | 12753375.12 | 61012096.51 | 34.00 | 195793.25 | 1386542.00 | 7420359.00 | 1.293859e+09 | ▇▁▁▁▁ |
| Thinness.Age1to19 | 34 | 0.99 | 4.84 | 4.42 | 0.10 | 1.60 | 3.30 | 7.20 | 2.770000e+01 | ▇▃▁▁▁ |
| Thinness.Age5to9 | 34 | 0.99 | 4.87 | 4.51 | 0.10 | 1.50 | 3.30 | 7.20 | 2.860000e+01 | ▇▃▁▁▁ |
| Income.composition.of.resources | 167 | 0.94 | 0.63 | 0.21 | 0.00 | 0.49 | 0.68 | 0.78 | 9.500000e-01 | ▁▁▅▇▆ |
| Years.of.education | 163 | 0.94 | 11.99 | 3.36 | 0.00 | 10.10 | 12.30 | 14.30 | 2.070000e+01 | ▁▂▇▇▁ |

#Plot of Life Expectancy and Adult mortality in 2000  
#| label: fig-exp1  
#| fig-cap: "Life Expectancy and Adult Mortality 2000"  
#| echo: FALSE  
knitr::include\_graphics(here("results","p1\_2000.png"))



#Plot of Life Expectancy and Adult mortality in 2015  
#| label: fig-exp2  
#| fig-cap: "Life Expectancy and Adult Mortality 2015"  
#| echo: FALSE  
knitr::include\_graphics(here("results","p2\_2015.png"))



|  |
| --- |
| Figure 1: Life Expectancy Status Boxplot in 2000 and 2015. |

|  |
| --- |
| Figure 2: Deaths Due to Unsafe Water Source, Unsafe Sanitation, or No Access to Handwashing Facilities in the United States from 2000 to 2015 |

## 4.2 Basic statistical analysis

Here I explored some factors that would seem to be reasonably significant with simple single predictor models to life expectancy, the outcome of interest.

First, a logistic regression with country status and life expectancy.

lifeexp$Status <- as.factor(lifeexp$Status)  
glm\_mod <- logistic\_reg() %>% set\_engine("glm")  
glm1 <- glm\_mod %>% fit(Status ~ Life.expectancy, data=lifeexp)  
#checking results  
tidy(glm1)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 30.2 1.39 21.7 6.03e-104  
2 Life.expectancy -0.383 0.0181 -21.1 9.39e- 99

Life expectancy is negatively correlated with the country status of developing.

Then a linear regression comparing year to life expectancy.

lm\_mod <- linear\_reg() %>% set\_engine("lm")  
fitlm1 <- lm\_mod %>%  
 fit(Life.expectancy ~ Year, data=lifeexp)  
#checking results  
fitlm1

parsnip model object  
  
  
Call:  
stats::lm(formula = Life.expectancy ~ Year, data = data)  
  
Coefficients:  
(Intercept) Year   
 -635.8715 0.3512

tidy(fitlm1)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) -636. 75.5 -8.42 5.96e-17  
2 Year 0.351 0.0376 9.33 1.96e-20

Year is positively correlated with life expectancy.

Another linear regression with BMI.

fitlm2 <- lm\_mod %>%  
 fit(Life.expectancy ~ BMI, data=lifeexp)  
#checking results  
fitlm2

parsnip model object  
  
  
Call:  
stats::lm(formula = Life.expectancy ~ BMI, data = data)  
  
Coefficients:  
(Intercept) BMI   
 59.0325 0.2699

tidy(fitlm2)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 59.0 0.314 188. 0   
2 BMI 0.270 0.00727 37.1 8.92e-247

In this instance, BMI is positively correlated with life expectancy.

fitlm3 <- lm\_mod %>% fit(Life.expectancy~Polio, data=lifeexp)  
fitlm3

parsnip model object  
  
  
Call:  
stats::lm(formula = Life.expectancy ~ Polio, data = data)  
  
Coefficients:  
(Intercept) Polio   
 53.7042 0.1888

tidy(fitlm3)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 53.7 0.571 94.0 0   
2 Polio 0.189 0.00666 28.4 1.96e-156

Polio doesn’t seem significantly correlated.

fitlm4 <- lm\_mod %>%  
 fit(Life.expectancy ~ Measles, data=lifeexp)  
#checking results  
fitlm4

parsnip model object  
  
  
Call:  
stats::lm(formula = Life.expectancy ~ Measles, data = data)  
  
Coefficients:  
(Intercept) Measles   
 69.5421700 -0.0001307

tidy(fitlm4)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 69.5 0.178 391. 0   
2 Measles -0.000131 0.0000151 -8.63 9.73e-18

Measles does not seem significantly correlated with life expectancy.

combo\_both <- right\_join(lifeexp, riskfactor, by=c("Country", "Year")) %>% filter(Year==c(2000, 2015))  
fitlm5 <- lm\_mod %>%  
 fit(Life.expectancy ~ Air.pollution, data=combo\_both)  
#checking results  
fitlm5

parsnip model object  
  
  
Call:  
stats::lm(formula = Life.expectancy ~ Air.pollution, data = data)  
  
Coefficients:  
 (Intercept) Air.pollution   
 6.942e+01 -3.175e-06

tidy(fitlm5)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 69.4 0.944 73.5 1.75e-91  
2 Air.pollution -0.00000317 0.00000599 -0.530 5.97e- 1

Air pollution is negatively correlated with life expectancy.

fitlm6 <- lm\_mod %>%  
 fit(Life.expectancy ~ Low.physical.activity, data=combo\_both)  
#checking results  
fitlm6

parsnip model object  
  
  
Call:  
stats::lm(formula = Life.expectancy ~ Low.physical.activity,   
 data = data)  
  
Coefficients:  
 (Intercept) Low.physical.activity   
 6.910e+01 3.679e-05

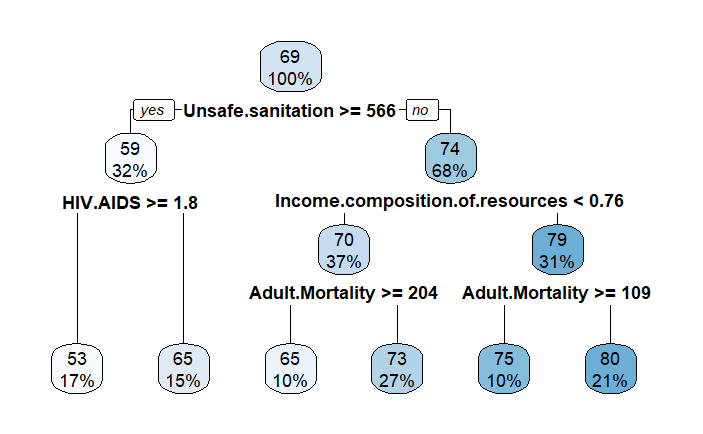
tidy(fitlm6)

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 69.1 0.967 71.4 3.26e-90  
2 Low.physical.activity 0.0000368 0.0000482 0.763 4.47e- 1

## 4.3 Full analysis

Here is a decision tree model with the predictions for different stratum of life expectancy.

#Decision Tree Graphic  
#| label: fig-result5  
#| fig-cap: "Decision tree"  
#| echo: FALSE  
knitr::include\_graphics(here("results","tree.png"))



The initial split was on unsafe sanitation, with the split reported as yes further split with HIV and AIDS into the lower life expectancy strata. If there is no unsafe sanitation, it further is split into income composition of resources and then adult mortality. This a simpler and less nuanced model but has an overall look at the predictive value of the data.

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

Some strengths include how large and expansive the data set is, as it has many variables for nearly every country and with a large range of years. A weakness that goes along with that is that it was difficult to work with such a large data set and pick out what the most important things to explore were. Also life expectancy is a nuanced estimate that has to do with a combination of multiple factors and also things that are not included in the available data, so a truly accurate analysis may be out of my reach.

## 5.3 Conclusions

*What are the main take-home messages?*

# 6. References

Kabir, Mahfuz. 2008. “Determinants of Life Expectancy in Developing Countries.” *The Journal of Developing Areas* 41 (2): 185–204. <https://doi.org/10.1353/jda.2008.0013>.