MADA Course Project

Risk Factors Affecting Life Expectancy

Annabella Hines

4/10/23

## 0.1 Load Required Packages

# 1. Summary/Abstract

*Write a summary of your project.*

# 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<- read\_csv("../../data/raw\_data/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.

riskfactor<- read\_csv("../../data/raw\_data/number-of-deaths-by-risk-factor.csv")

Rows: 6468 Columns: 31  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): Entity  
dbl (30): Year, Unsafe water source, Unsafe sanitation, No access to handwas...  
  
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.

#glimpse of each  
glimpse(lifeexp)

Rows: 2,938  
Columns: 22  
$ Country <chr> "Afghanistan", "Afghanistan", "Afgha…  
$ Year <dbl> 2015, 2014, 2013, 2012, 2011, 2010, …  
$ Status <chr> "Developing", "Developing", "Develop…  
$ `Life expectancy` <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, …  
$ `Adult Mortality` <dbl> 263, 271, 268, 272, 275, 279, 281, 2…  
$ `infant deaths` <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, …  
$ Alcohol <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, …  
$ `percentage expenditure` <dbl> 71.279624, 73.523582, 73.219243, 78.…  
$ `Hepatitis B` <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, …  
$ Measles <dbl> 1154, 492, 430, 2787, 3013, 1989, 28…  
$ BMI <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, …  
$ `under-five deaths` <dbl> 83, 86, 89, 93, 97, 102, 106, 110, 1…  
$ Polio <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 5…  
$ `Total expenditure` <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, …  
$ Diphtheria <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, …  
$ `HIV/AIDS` <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0…  
$ GDP <dbl> 584.25921, 612.69651, 631.74498, 669…  
$ Population <dbl> 33736494, 327582, 31731688, 3696958,…  
$ `thinness 1-19 years` <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, …  
$ `thinness 5-9 years` <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, …  
$ `Income composition of resources` <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0…  
$ Schooling <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9,…

glimpse(riskfactor)

Rows: 6,468  
Columns: 31  
$ Entity <chr> "Afghanistan", "Afghanistan…  
$ Year <dbl> 1990, 1991, 1992, 1993, 199…  
$ `Unsafe water source` <dbl> 7554.050, 7359.677, 7650.43…  
$ `Unsafe sanitation` <dbl> 5887.748, 5732.770, 5954.80…  
$ `No access to handwashing facility` <dbl> 5412.315, 5287.891, 5506.65…  
$ `Household air pollution from solid fuels` <dbl> 22388.50, 22128.76, 22873.7…  
$ `Non-exclusive breastfeeding` <dbl> 3221.139, 3150.560, 3331.34…  
$ `Discontinued breastfeeding` <dbl> 156.09755, 151.53985, 156.6…  
$ `Child wasting` <dbl> 22778.85, 22292.69, 23102.2…  
$ `Child stunting` <dbl> 10408.439, 10271.976, 10618…  
$ `Low birth weight for gestation` <dbl> 12168.56, 12360.64, 13459.5…  
$ `Secondhand smoke` <dbl> 4234.808, 4219.597, 4371.90…  
$ `Alcohol use` <dbl> 356.5293, 320.5985, 293.257…  
$ `Drug use` <dbl> 208.3254, 217.7697, 247.833…  
$ `Diet low in fruits` <dbl> 8538.964, 8642.847, 8961.52…  
$ `Diet low in vegetables` <dbl> 7678.718, 7789.773, 8083.23…  
$ `Unsafe sex` <dbl> 387.1676, 394.4483, 422.453…  
$ `Low physical activity` <dbl> 4221.303, 4252.630, 4347.33…  
$ `High fasting plasma glucose` <dbl> 21610.07, 21824.94, 22418.7…  
$ `High total cholesterol` <dbl> 9505.532, NA, NA, NA, NA, 1…  
$ `High body-mass index` <dbl> 7701.581, 7747.775, 7991.01…  
$ `High systolic blood pressure` <dbl> 28183.98, 28435.40, 29173.6…  
$ Smoking <dbl> 6393.667, 6429.253, 6561.05…  
$ `Iron deficiency` <dbl> 726.4313, 739.2458, 873.485…  
$ `Vitamin A deficiency` <dbl> 9344.132, 9330.182, 9769.84…  
$ `Low bone mineral density` <dbl> 374.8441, 379.8542, 388.130…  
$ `Air pollution` <dbl> 26598.01, 26379.53, 27263.1…  
$ `Outdoor air pollution` <dbl> 4383.83, 4426.36, 4568.91, …  
$ `Diet high in sodium` <dbl> 2737.198, 2741.185, 2798.56…  
$ `Diet low in whole grains` <dbl> 11381.38, 11487.83, 11866.2…  
$ `Diet low in nuts and seeds` <dbl> 7299.867, 7386.764, 7640.62…

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

## 2.3 Questions/Hypotheses to be addressed

Is there a correlation between certain risk factors and lower life expectancies by country? Have there been any specific trends over the years in which risk factors have become more prevalent?

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 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 or to better match each other.

## 3.2 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

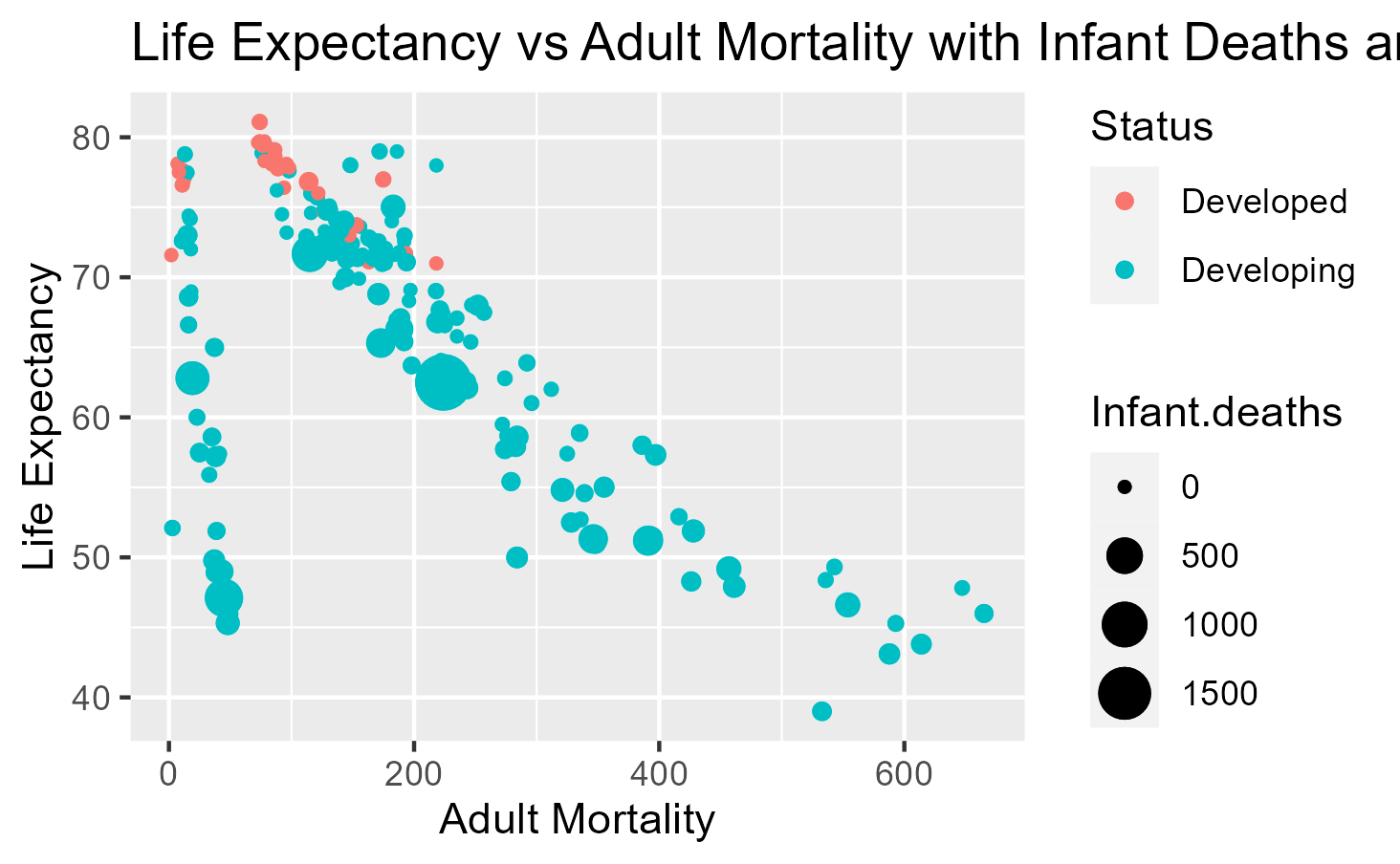
## 4.1 Exploratory/Descriptive analysis

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

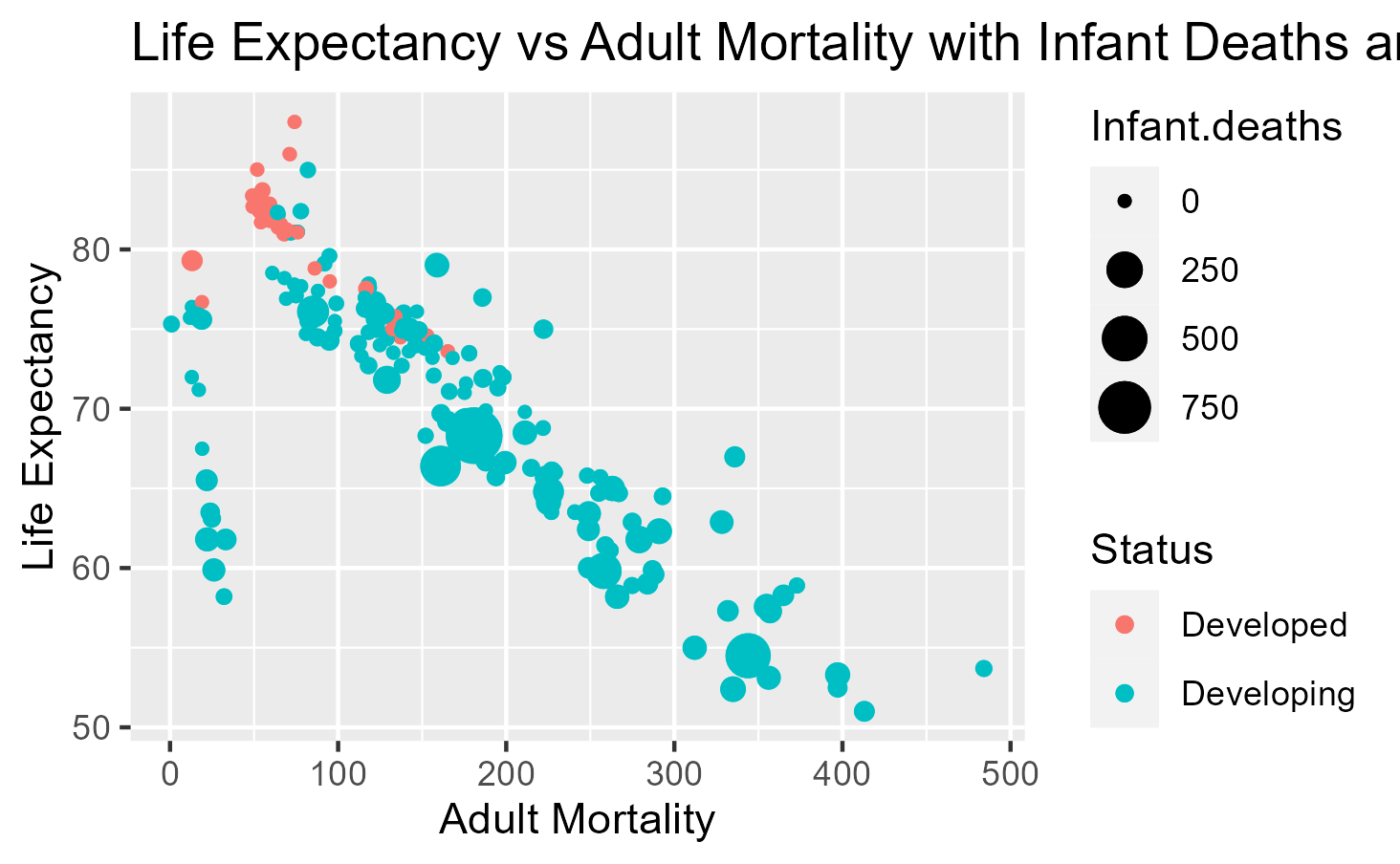
#Risk factor dataset summary table

#Life expectancy dataset summary table.

#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. |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

shows a scatterplot figure produced by one of the R scripts.

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example shows a summary of a linear model fit.

# 5. Discussion

## 5.1 Summary and Interpretation

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

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 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>.