MADA Course Project

Risk Factors Affecting Life Expectancy

Annabella Hines

5/5/23

# 1. Abstract

Life expectancy is an important general metric in comparing countries’ health of their populations on the same scale, but the factors that affect it are numerous and nuanced.The goal is to create a model to better understand which risks are the most influential on life expectancy by combining two data sets containing information on life expectancy, demographics, and prevalent risk factors. Through analysis of data collected by the World Health Organization on the average life expectancy and most prevalent risk factors for death for each country over the years, I aim to draw initial associations that may be areas for further study or targets of public health measures. Through a surface level exploration of the data, the trends in life expectancy compared to other factors can be observed and then further examined with simple statistical analyses. After that point, machine learning models are employed to more accurately predict life expectancy from provided variables.

# 2. Introduction

## 2.1 General Background Information

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(Goodson et al. 2017) (Yin et al. 2020). The risk exposures that have increased the most globally from 2010 to 2019 are ambient particulate matter pollution, drug use, high fasting plasma glucose, and high body-mass index while the ones that have decreased are household air pollution, unsafe water/sanitation/handwashing, child growth failure, tobacco smoking, and lead exposure (Murray et al. 2020). Although the leading risk factors resulting in death vary by age and location, the overall highest were high systolic blood pressure and and tobacco use(Murray et al. 2020).

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). There have been attempts to create formulas to mathematically determine a population’s life expectancy based on risk factors within that population but each are so specific that its difficult to create an overall generalizable model.

## 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. The summary table of the risk factor data set (**table-1?**) and life expectancy data set (**table-2?**) can be seen in the exploratory and descriptive analysis section of the manuscript.

## 2.3 Questions/Hypotheses to be addressed

Which variables are the most highly correlated to life expectancy? What risk factors or demographics are the best predictors of 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, such as ‘Entity’ to ‘Country’ in the risk factor data.

## 3.2 Statistical analysis

Initially some generalized linear models were ran with all variables and with certain subsets of variables. Life expectancy is negatively correlated with the status of developing and positively correlated with the status of developed. After finding those results, some machine learning methods were used for more specialized conclusions and predictive modeling. A decision tree **?@fig-result5** was developed initially that illustrated simple predictive variables leading to the strata of life expectancy.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

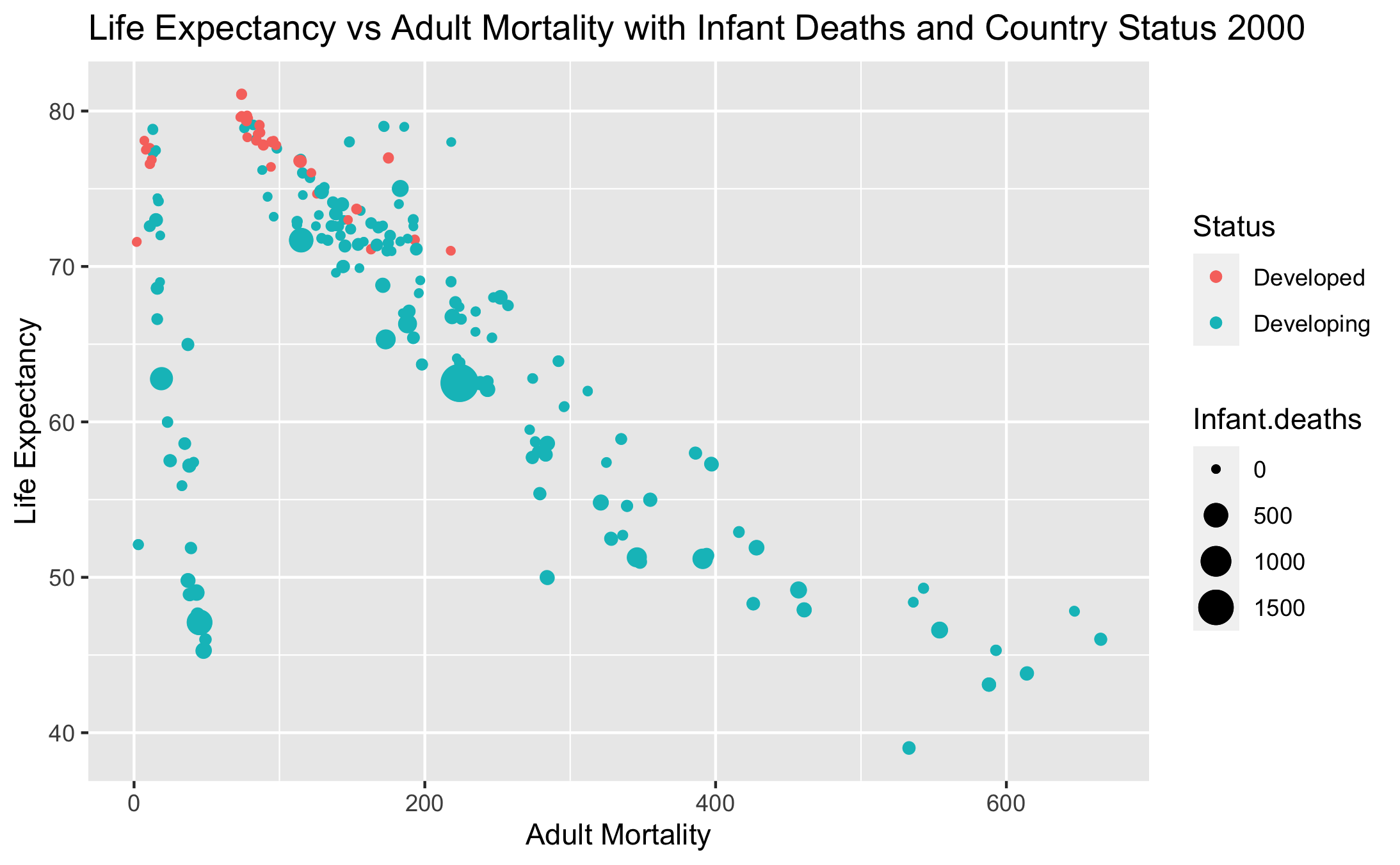
First, here are summary tables for each data set to see an overview of the data.

#Risk factor dataset summary table  
#| label: table-1  
#| fig-cap: "Risk Factor Data Summary Tabel"  
#| echo: FALSE  
readRDS(file = "../../results/summarytableriskfactor.rds")

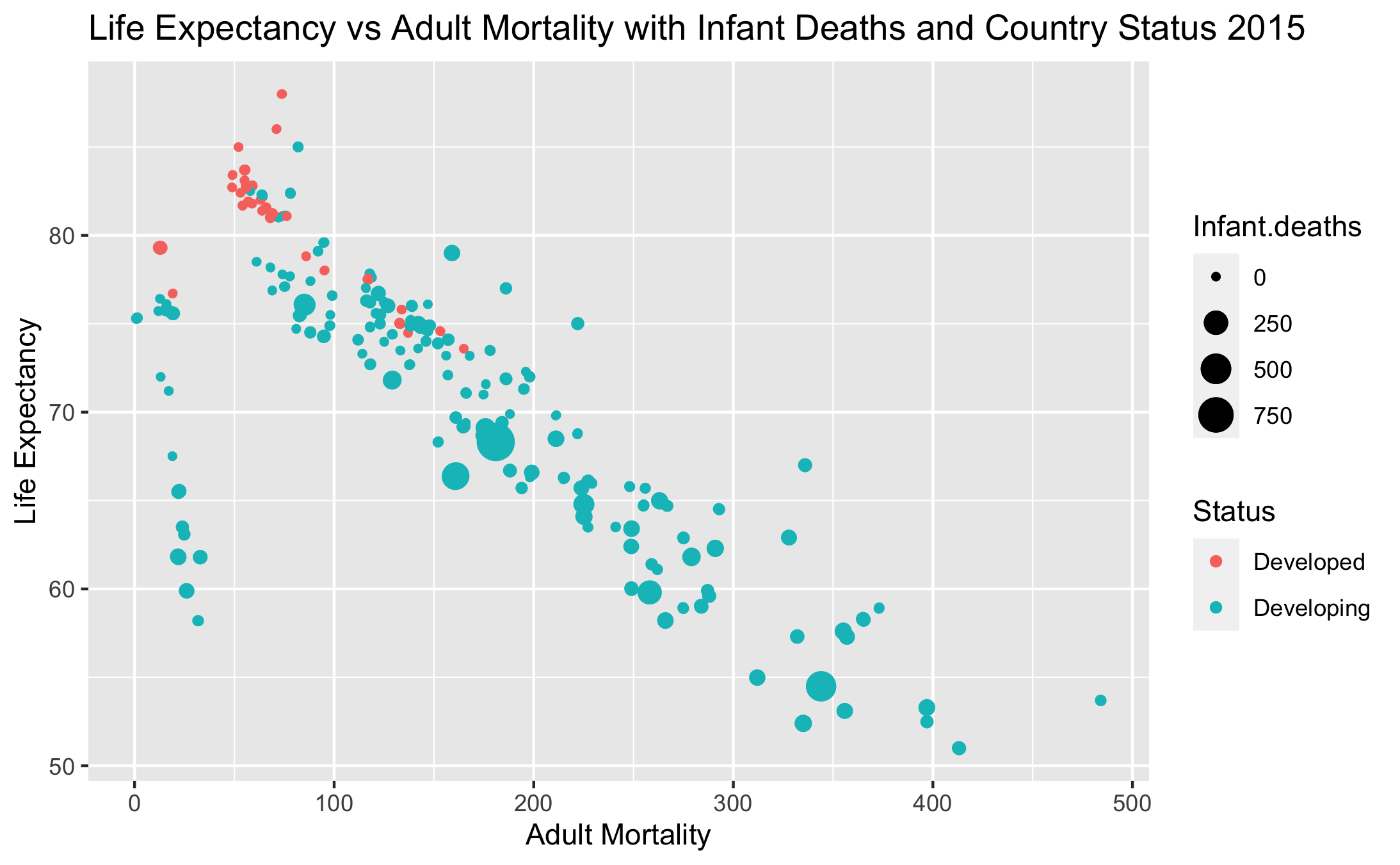
# A tibble: 31 × 17  
 skim\_type skim\_variable n\_missing complete\_rate character.min character.max  
 \* <chr> <chr> <int> <dbl> <int> <int>  
 1 character Country 0 1 4 48  
 2 numeric Year 0 1 NA NA  
 3 numeric Unsafe.water.s… 0 1 NA NA  
 4 numeric Unsafe.sanitat… 0 1 NA NA  
 5 numeric No.access.to.h… 0 1 NA NA  
 6 numeric Household.air.… 0 1 NA NA  
 7 numeric Non.exclusive.… 0 1 NA NA  
 8 numeric Discontinued.b… 0 1 NA NA  
 9 numeric Child.wasting 0 1 NA NA  
10 numeric Child.stunting 0 1 NA NA  
# ℹ 21 more rows  
# ℹ 11 more variables: character.empty <int>, character.n\_unique <int>,  
# character.whitespace <int>, numeric.mean <dbl>, numeric.sd <dbl>,  
# numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>, numeric.p75 <dbl>,  
# numeric.p100 <dbl>, numeric.hist <chr>

# A tibble: 22 × 20  
 skim\_type skim\_variable n\_missing complete\_rate character.min character.max  
 \* <chr> <chr> <int> <dbl> <int> <int>  
 1 character Country 0 1 4 52  
 2 factor Status 0 1 NA NA  
 3 numeric Year 0 1 NA NA  
 4 numeric Life.expectancy 10 0.997 NA NA  
 5 numeric Adult.Mortality 10 0.997 NA NA  
 6 numeric Infant.deaths 0 1 NA NA  
 7 numeric Alcohol 194 0.934 NA NA  
 8 numeric Percentage.exp… 0 1 NA NA  
 9 numeric Hepatitis.B 553 0.812 NA NA  
10 numeric Measles 0 1 NA NA  
# ℹ 12 more rows  
# ℹ 14 more variables: character.empty <int>, character.n\_unique <int>,  
# character.whitespace <int>, factor.ordered <lgl>, factor.n\_unique <int>,  
# factor.top\_counts <chr>, numeric.mean <dbl>, numeric.sd <dbl>,  
# numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>, numeric.p75 <dbl>,  
# numeric.p100 <dbl>, numeric.hist <chr>

#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"))



The trend of life expectancy decreasing as adult mortality increases is similar in 2000 and 2015, with developed countries shown at the end with high life expectancy and low adult mortality.

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

The life expectancy distribution for developed and developing countries were similar in 2000 and 2015, with developed countries having an overall higher distribution.

|  |
| --- |
| 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.

# 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 appears to be negatively correlated with the country status of developing.

Then a linear regression comparing year to life expectancy.

# 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.

# 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.

# 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.

# 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.

# 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

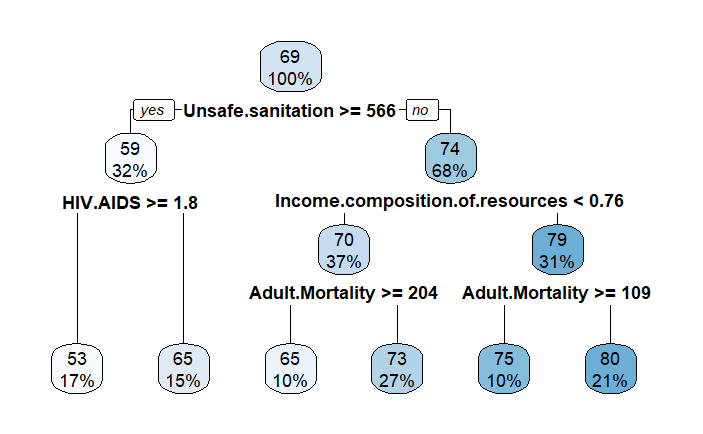
Air pollution is negatively correlated with life expectancy.

# 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. It also comes from a very reliable source, the World Health Organization. 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

Shaw, J. W., Horrace, W. C., & Vogel, R. J. (2005, April 1). The determinants of life expectancy: An analysis of the OECD health … Retrieved April 27, 2023, from https://onlinelibrary.wiley.com/doi/10.1002/j.2325-8012.2005.tb00675.x

GBD 2019 Risk Factors Collaborators (2020). Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet (London, England), 396(10258), 1223–1249. https://doi.org/10.1016/S0140-6736(20)30752-2

Yin, P., Brauer, M., Cohen, A. J., Wang, H., Li, J., Burnett, R. T., Stanaway, J. D., Causey, K., Larson, S., Godwin, W., Frostad, J., Marks, A., Wang, L., Zhou, M., & Murray, C. J. L. (2020). The effect of air pollution on deaths, disease burden, and life expectancy across China and its provinces, 1990-2017: an analysis for the Global Burden of Disease Study 2017. The Lancet. Planetary health, 4(9), e386–e398. https://doi.org/10.1016/S2542-5196(20)30161-3

Goodson, James L., James P. Alexander, Robert W. Linkins, and Walter A. Orenstein. 2017. “Measles and Rubella Elimination: Learning from Polio Eradication and Moving Forward with a Diagonal Approach.” *Expert Review of Vaccines* 16 (12): 1203–16. <https://doi.org/10.1080/14760584.2017.1393337>.

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>.

Murray, Christopher J L, Aleksandr Y Aravkin, Peng Zheng, Cristiana Abbafati, Kaja M Abbas, Mohsen Abbasi-Kangevari, Foad Abd-Allah, et al. 2020. “Global Burden of 87 Risk Factors in 204 Countries and Territories, 19902019: A Systematic Analysis for the Global Burden of Disease Study 2019.” *The Lancet* 396 (10258): 1223–49. <https://doi.org/10.1016/s0140-6736(20)30752-2>.

Yin, Peng, Michael Brauer, Aaron J Cohen, Haidong Wang, Jie Li, Richard T Burnett, Jeffrey D Stanaway, et al. 2020. “The Effect of Air Pollution on Deaths, Disease Burden, and Life Expectancy Across China and Its Provinces, 19902017: An Analysis for the Global Burden of Disease Study 2017.” *The Lancet Planetary Health* 4 (9): e386–98. <https://doi.org/10.1016/s2542-5196(20)30161-3>.