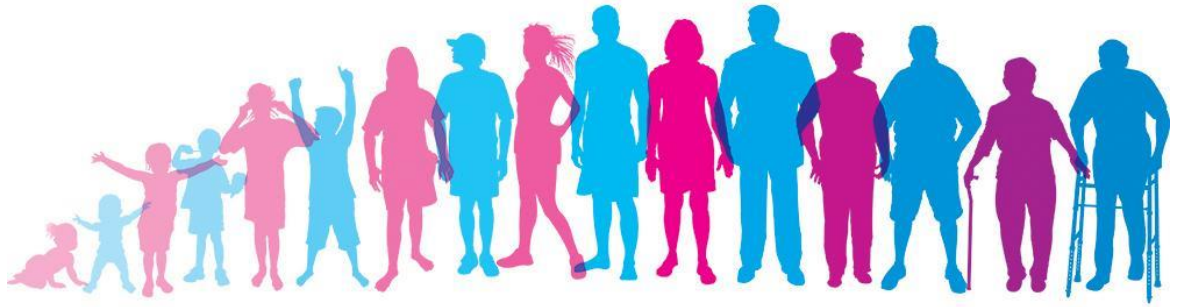
Sarah Carrier

Springboard Capstone 2 Project

February 2023

# Predicting Life Expectancy



# Motivation and Framing of Project

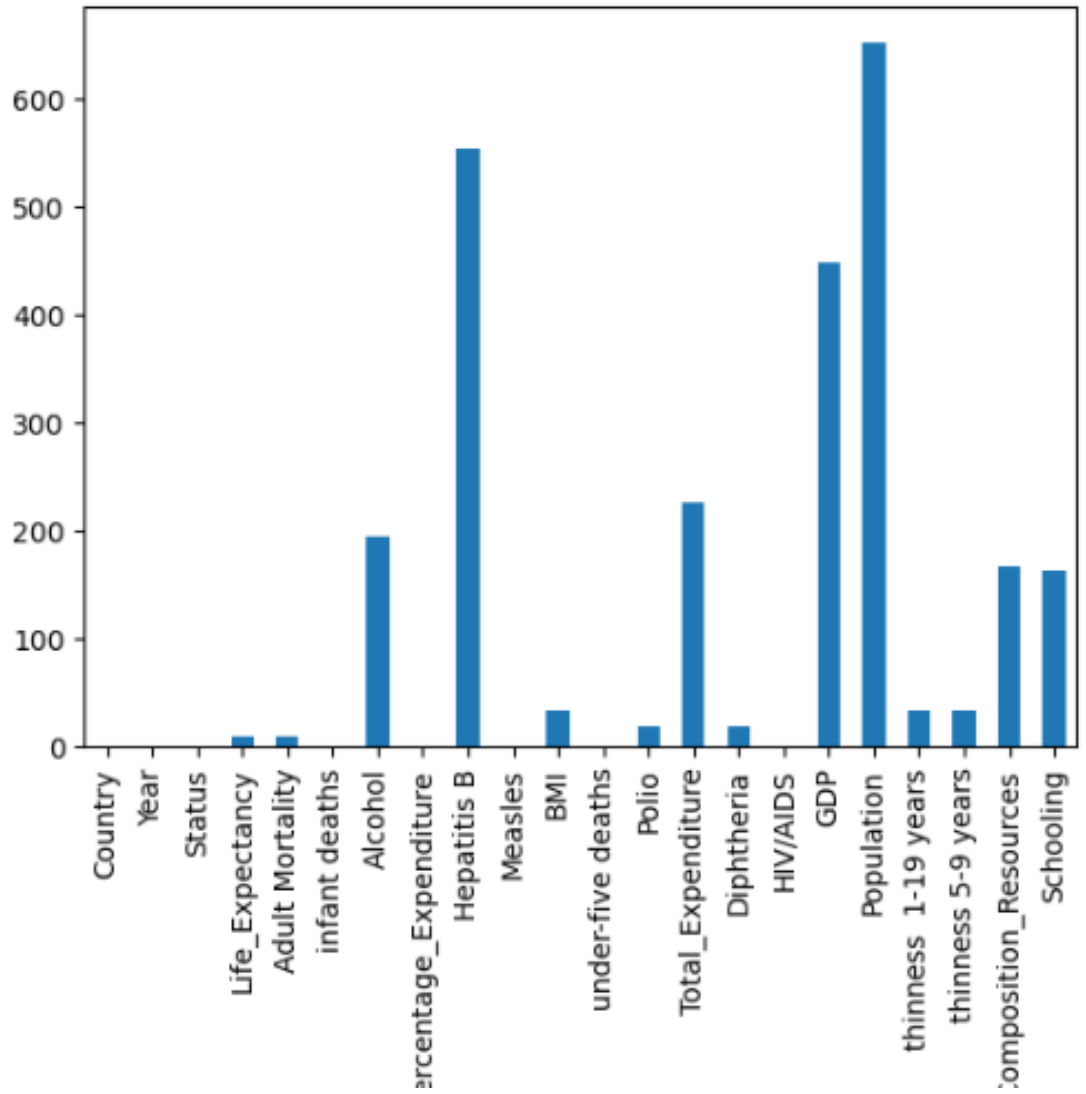
- Data proves that life expectancy has been increasing across the globe in both developing and developed countries. However, there is still room for prolonging human life spans, and maintaining healthy life expectancy should remain at the forefront of concern, particularly for countries deemed “developing” where life expectancy trails that of developed nations.
- The World Health Organization tracks life expectancy by country as well as potentially correlating factors such as alcohol consumption, GDP, and immunization rates. The dataset tracks variables such as life expectancy over the course of 15 years, 2000-2015.
- Factors like immunization rates probably influence life expectancy. Can the potentially correlating features be used to predict life expectancy for a country? If so, international leader can then focus on those factors in order to boost life expectancy.

# Audience and Use Cases

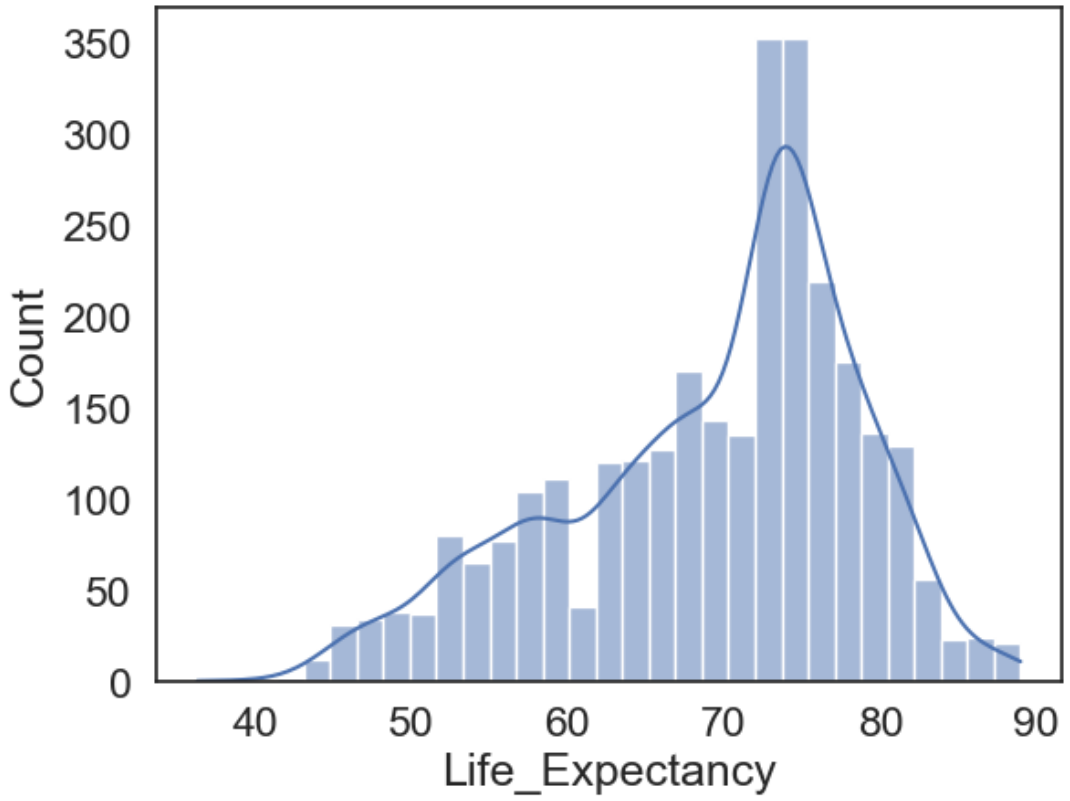
- politicians and economic leaders in nations across the world
- public health and medical experts and policy makers
- health providers, hospitals, and insurance companies looking to avoid overspending while also improving services
- community groups and advocates concerned about public health and welfare
- employers, recruiters, industry representatives interested in labor and employee welfare

# Data Wrangling and Cleaning

- In total, 2938 rows and 22 features
- There were a lot of NaNs in the dataset, *as seen at right*, and in fields where it was imperative that the values be generated to work with the data. To correct for this, I employed backwards and forwards interpolation to create values where needed.
- Concerning 0s: many of the values needed to be 0, so I did not commit to a blanket replacement of all zeros.



# Life Expectancy: The target variable

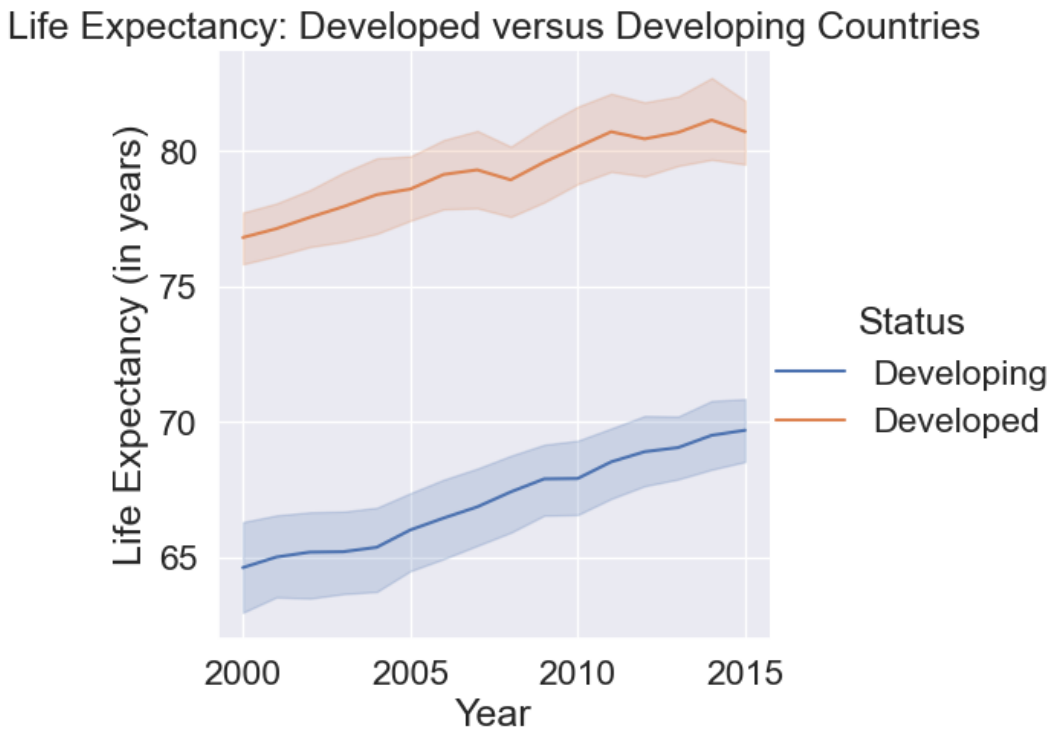



This visualization shows both the range and the mean of life expectancy:

Range 45-90 years

Average lifespan 69 years

# Life Expectancy by Country Status



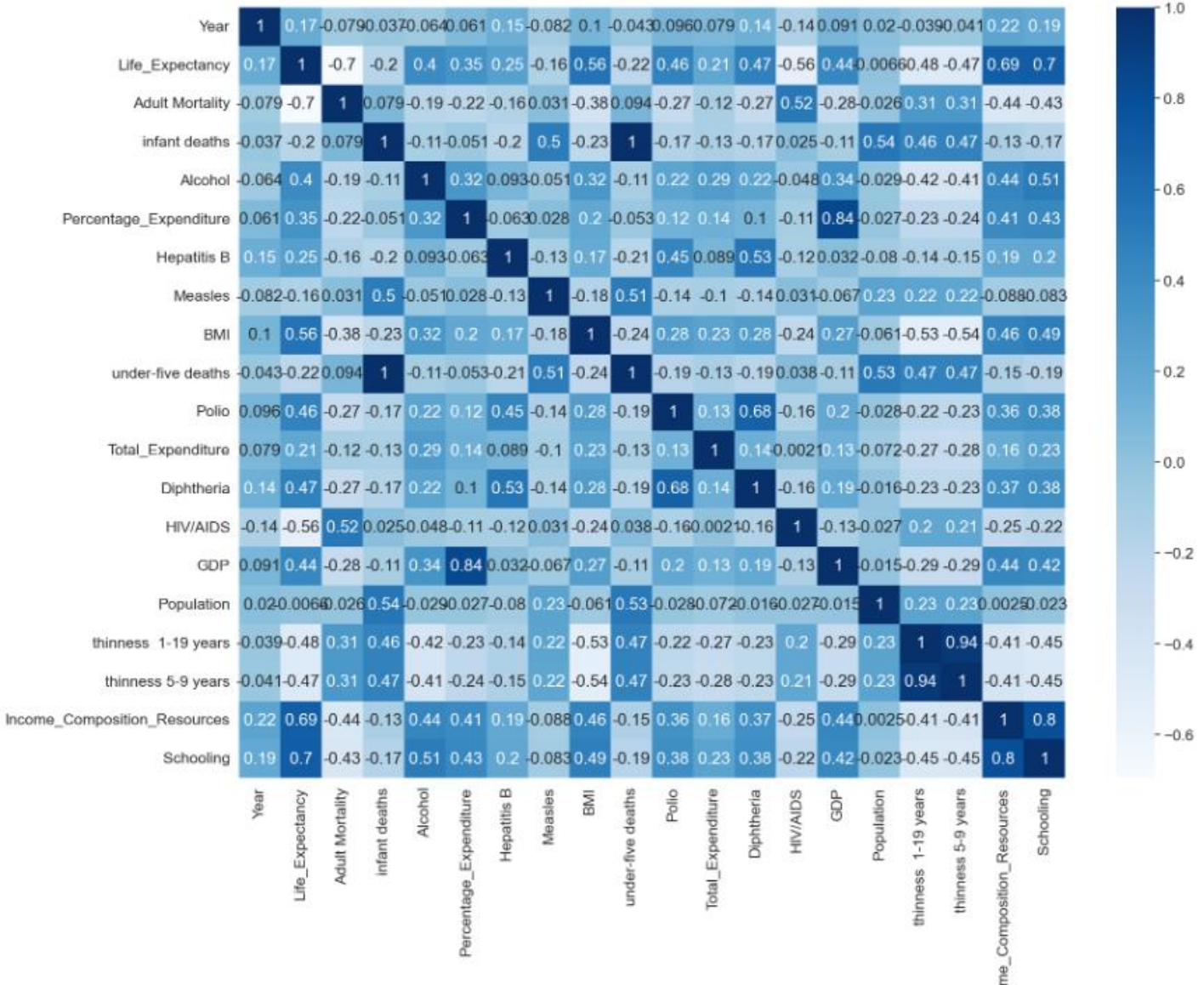

STATUS indicates whether a county is deemed “developing” or “developed”

This is helpful to visualize because it shows 1. developing countries have lower life expectancy, generally, but 2. life expectancy has also been steadily increasing for both categories.

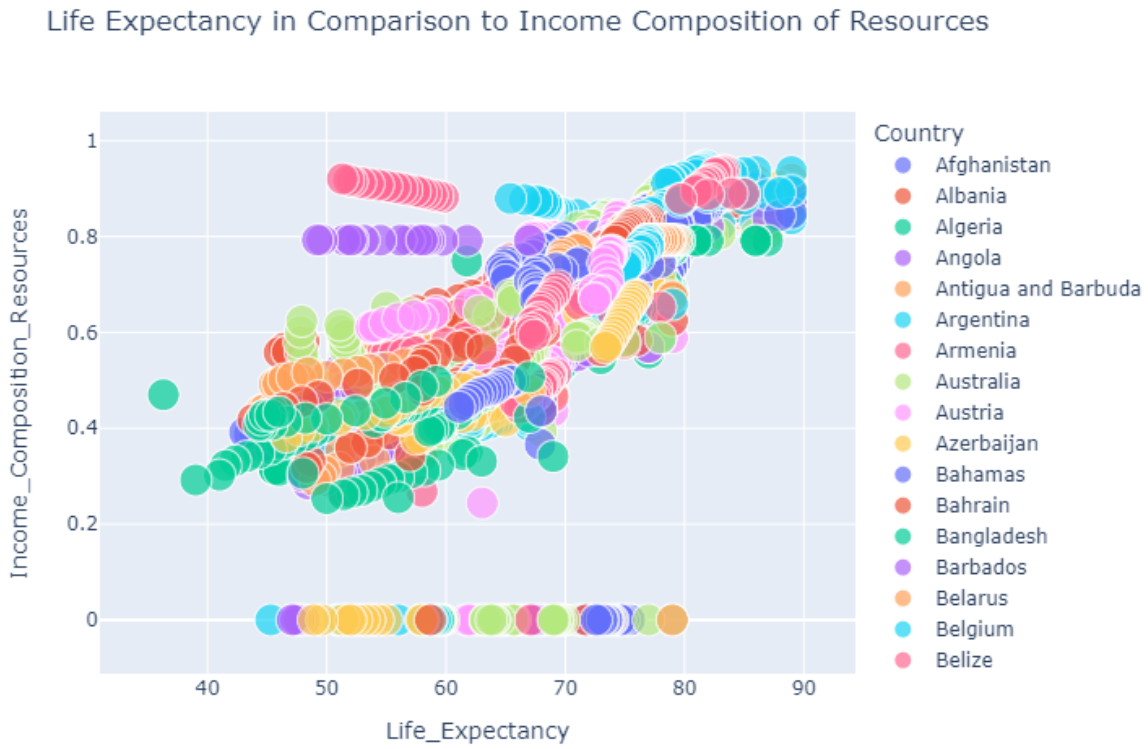
# EDA and Summary Statistics

Correlation heat map

This map shows the highest correlations between life expectancy and schooling, then income composition of resources.



# Visualizations



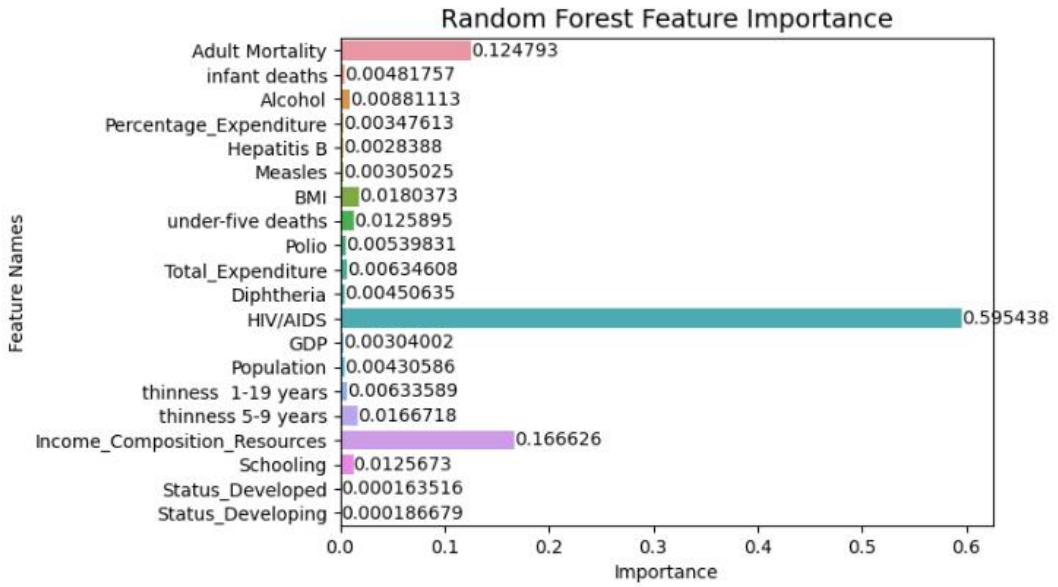

Scatter plots revealed relationships between features and the dependent variable life expectancy. For example, plotting income composition of resources reveals a linear relationship with life expectancy.

Outliers turned out to be due to a problem with the dataset itself.

# Developing Regression Models

Linear Regression, K Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting

# Testing the Models




- Comparison based on R2 scores, Mean Squared Error
- Result: Random Forest performed the best, although all the models performed remarkably well
- Feature importance with Random Forest

# Determination and Conclusions

- Random Forest feature importance confirmed one of the feature correlations that became apparent during the EDA step: income composition of resources (ICR)
- What is INCOME COMPOSITION OF RESOURCES?
  - the extent to which the income composition in capital and labor income is distributed across the income distribution. High levels of income composition inequality are associated with class-fragmented societies, whereas low levels are typical of multiple-sources-of-income societies.
- Income composition of resources is scored 0-1, with more equal societies have a score close to one.
- Returning to some of the insights extracted during the EDA process:
  - The country with the highest ICR in the dataset is NORWAY. Its average life expectancy across 15 years represented in the dataset is 82 – this is much higher than the average across the dataset of 69 years.

# Recommendations

- It is recommended that the first immediate step for increasing a country's life expectancy is to DECREASE inequality in income composition of resources.
- The country of NORWAY should be looked at as it has a very high life expectancy and the highest ICR in the entire dataset of almost 1 (0.984). What is Norway doing to ensure an equitable income composition of resources? This approach should be modeled.

# Future Work

- Two features that indicated potential correlations but were not consistent across analysis or modeling should be looked into further to determine if more scaling or cleaning could yield further insights:
  - SCHOOLING – this was the feature of highest correlation according to the heat map
    - Schooling in this dataset is the average number of years a country's citizens spend in school.
  - HIV/AIDS – this feature was by far scored the most important by the Random Forest model. In previous analysis, this feature did not seem to be important.
    - In this dataset, HIV/AIDS is "Deaths per 1 000 live births HIV/AIDS (0-4 years)"