# Predicting Life Expectancy

via
health factors and government spending

By Sailaja Karra and Brayton Hall

# Overview

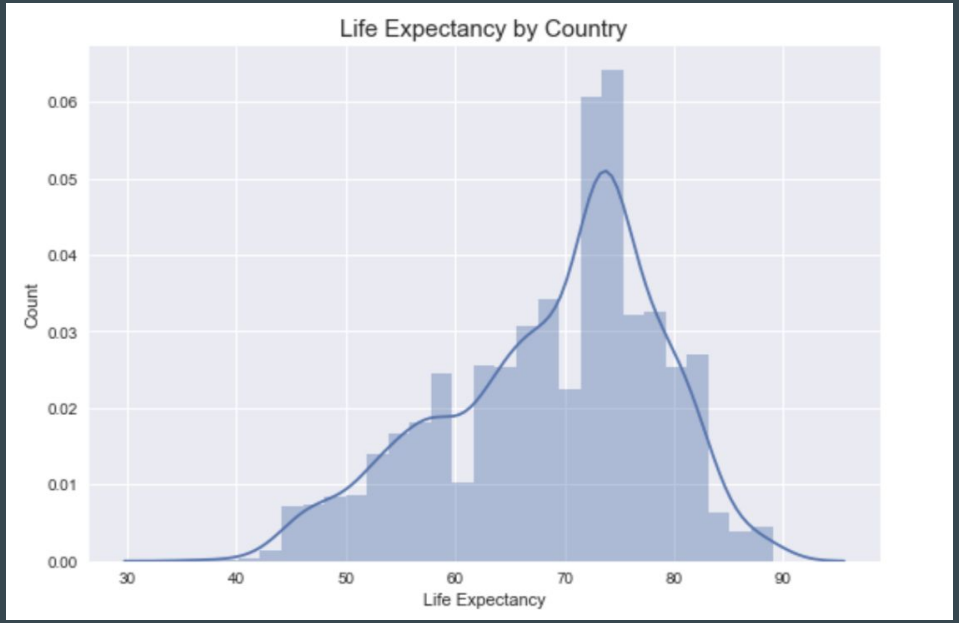
## What is life expectancy?

The *average expected lifespan* of people across different countries, which vary in their allocation of resources to health care.

Data Collected from:
World Health Organization
(via Kaggle)

Average Global Life Expectancy:
68.8 years

Standard Deviation:
9.8 years:



# CLEAN UP AND EDA

Initial DataFrame:
2938 SAMPLES
22 COLUMNS

14 COLUMNS MISSING DATA
Hepatitis B missing 553
Population 652

## Cleaning Process

1. Dropped Hepatitis B, Country, Year
2. Imputed median for 'schooling,', "alcohol', and 'GDP' and all economic features missing values
3. Turned our only binary categorical variable 'Status', into 0 or 1 for 'Developing' or 'Developed'
4. Dropped Remaining Missing Values

Cleaned DataFrame:

2244 SAMPLES
18 Columns

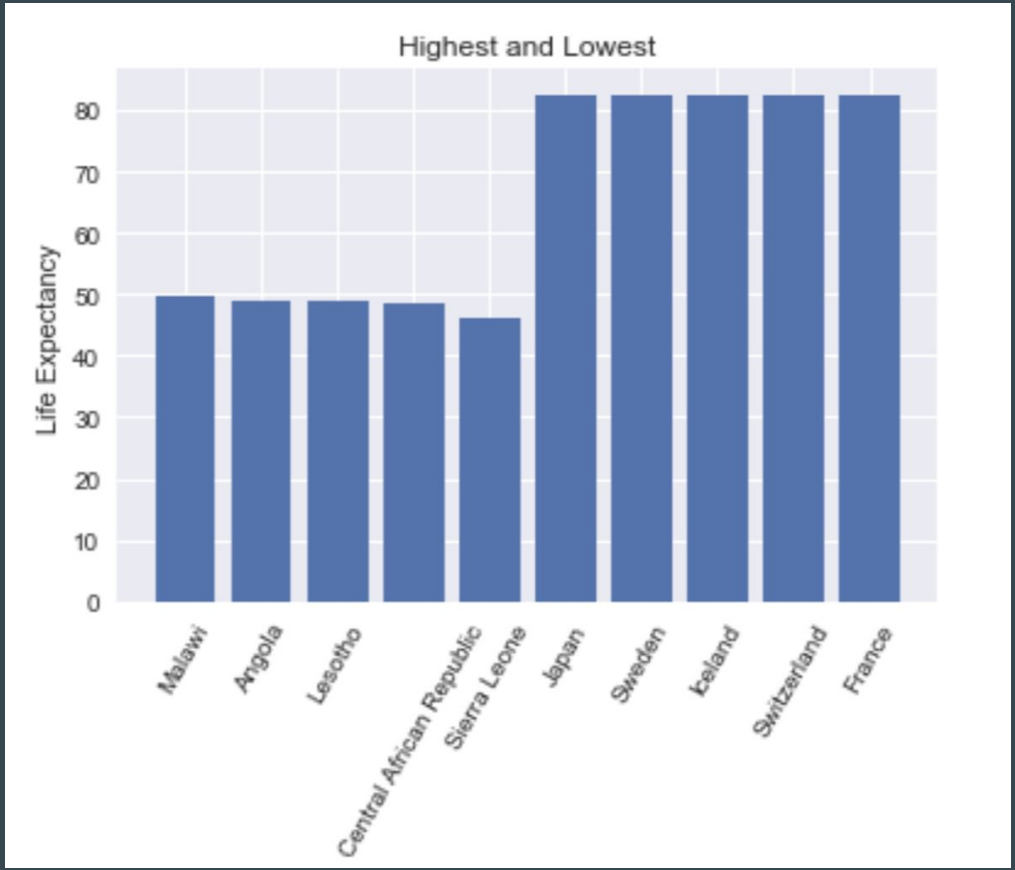
# COUNTRIES WITH HIGHEST AND LOWEST LIFE EXPECTANCIES

## LOWEST

Sierra Leone 46.1
Central African Republic 48.5
Lesotho 48.8
Angola 49.0
Malawi 49.9

## HIGHEST

Japan 82.5
Sweden 82.5
Iceland 82.4
Switzerland 82.3
France 82.2



# OUR GOAL

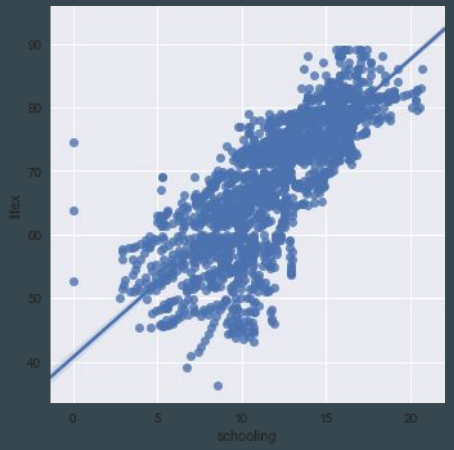
Determine what most affects life expectancy in order to effectively predict it, and produce actionable advice based on those features.

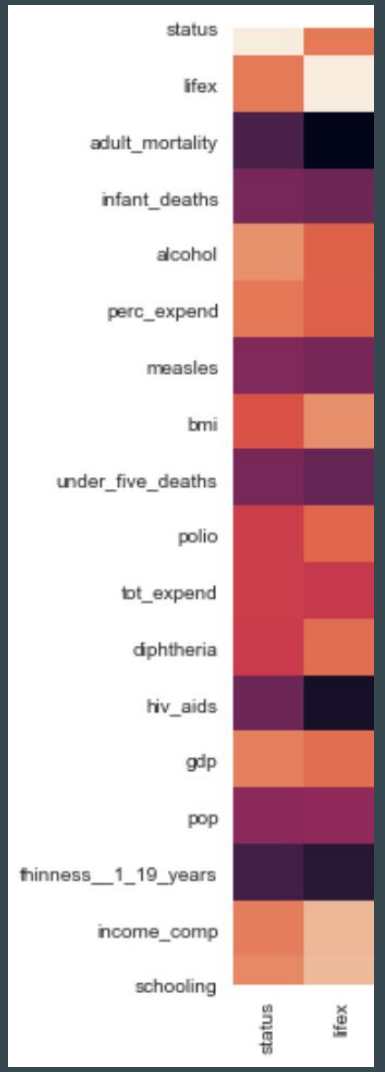
## Main relevant features in our data:

- education
- population
- alcohol
- bmi
- income index
- infant mortality
- adult mortality
- development status
- HIV/AIDS

Initial strong correlations:

1. schooling .78
2. adult_mort -.67
3. bmi .59
4. status .51



# MODEL BUILDING

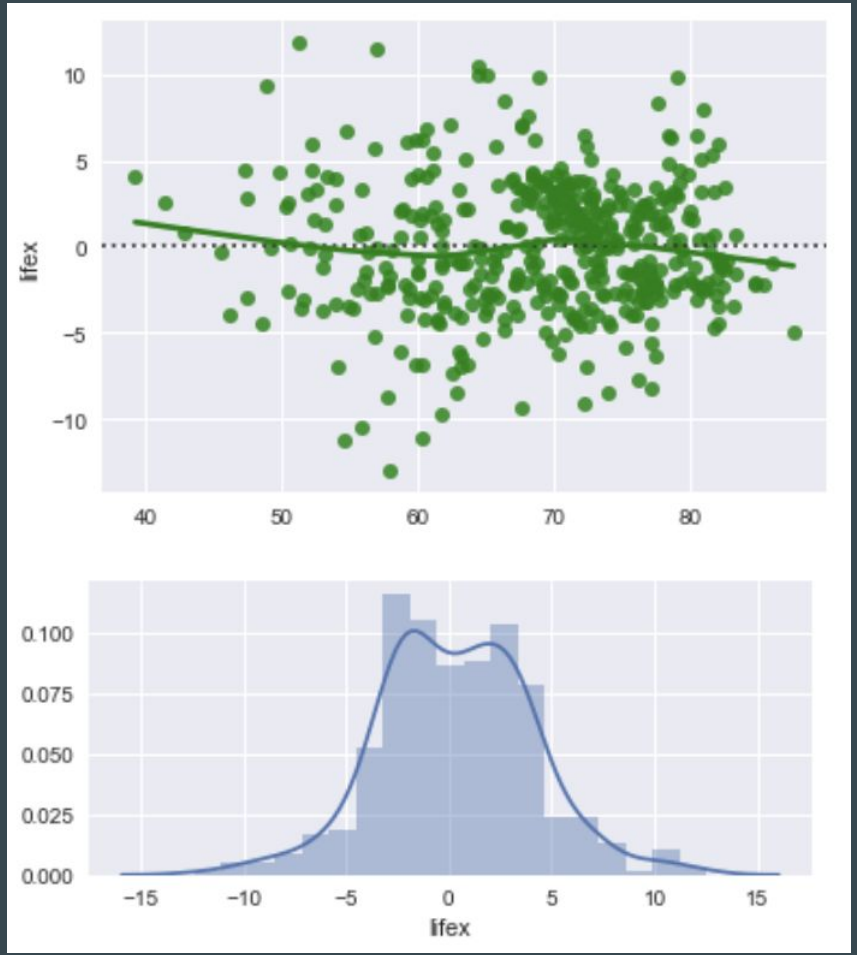
| | |
|---|---|
| Recursive Feature Elimination | Income, Schooling, HIV/AIDS |
| Linear model | TEST $R^2$ 0.83 RMSE 3.70 |
| Lasso L1 | TEST $R^2$ 0.83 RMSE 3.70 |
| Ridge L2 | TEST $R^2$ 0.82 RMSE 3.69 |

Ridge L2 (alpha: .01) was our best model, with a root mean squared error of 3.69, meaning it is, on average, 3.69 days off when predicting the true values.

# Coefficients of predictors

- ➔ Recursive feature elimination revealed the most important features in our model
- ➔ As expected **Income** is the main driving factor for life expectancy. Followed by **Schooling** and negatively affected by **HIV_AIDS**.
- ➔ Residuals are normally distributed

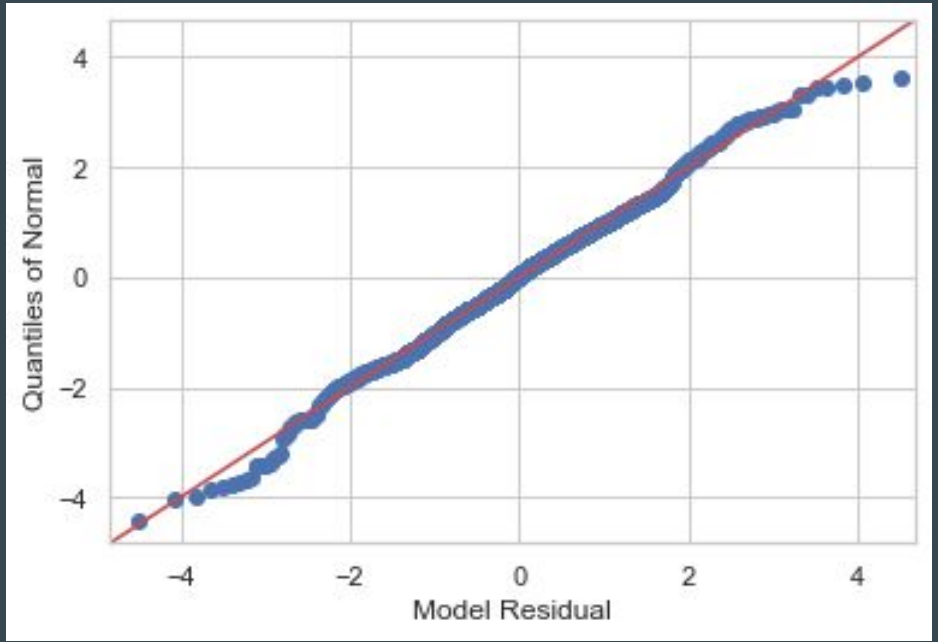
Residual Plots for Normality

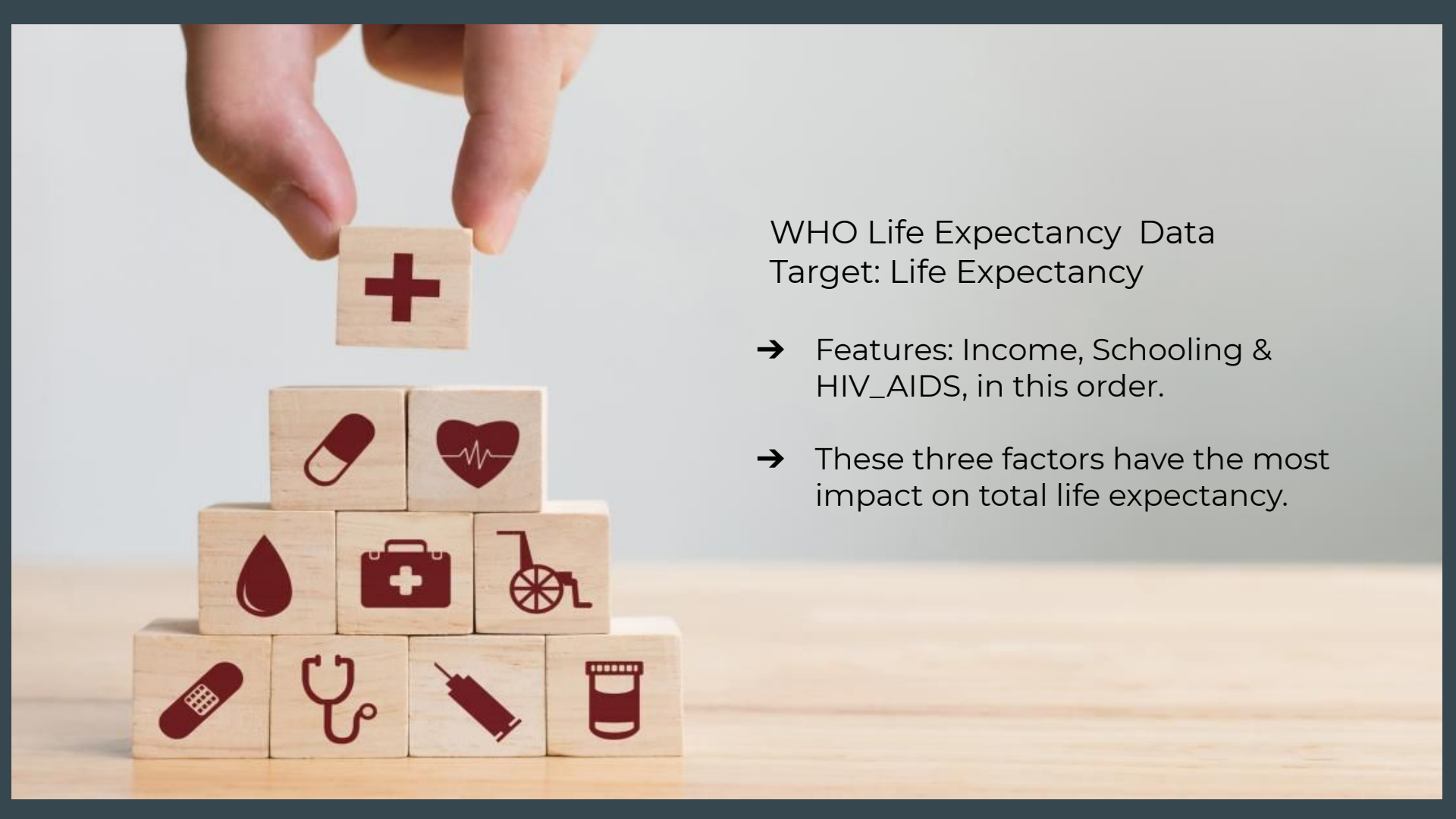


# Conclusion based on linear regression coefficients:

WHO Life Expectancy  Data
Target: Life Expectancy

- ➔ Feature: Income, Schooling & HIV_AIDS. In this order.
- ➔ These three factors have the most impact on total life expectancy.
- ➔ Total R2 is 76.3% from these three predictors.
- ➔ Used Stats Model to Build the Linear Regression Formula





WHO Life Expectancy  Data
Target: Life Expectancy

- ➔ Features: Income, Schooling & HIV_AIDS, in this order.
- ➔ These three factors have the most impact on total life expectancy.