DS 301 - Final Project, Spring 2023, Team B^3

World Health & Economics

By Abbie, Amanda, Brett, Yen

Background/Motivation

- **Goal:** See how different factors affect life expectancy across the world.
- **Why:**

Health and Wellbeing: Access to healthcare and available medical supplies varies greatly from country to country.

Research: Understanding importance of factors can lead research efforts and focus time, effort, and funds where it is needed the most.

Policy Making : Knowing how healthcare decisions can impact a country can help governments enact more beneficial policies.

Public Awareness : Communicating health recommendations can serve as preventive care and improve overall quality of life.

About the Dataset: Life Expectancy (WHO)

- Data contains life expectancy, health, immunization, and economic and demographic information about **179 countries** from **2000-2015 years**.
- The adjusted dataset has **21 variables** and **2864 rows**.
- Missing values are replaced by the average of the past three years.
- Data are gathered from reliable sources such as WHO, Oxford University, World Bank Data, and United Nations
- Variables:
    - **Country**
    - **Year**
    - Under_five_deaths
    - Alcohol_consumption
    - Measles
    - Polio
    - Incidents_HIV
    - Population_mln
    - Thinness_five_nine_years
    - Economy_status_Developed
    - **Life_expectancy**
    - Region
    - Infant_deaths
    - Adult_mortality
    - Hepatitis_B
    - BMI
    - Diphtheria
    - **GDP_per_capita**
    - Thinness_ten_nineteen_years
    - Schooling
    - Economy_status_Developing

About the Dataset: Life Expectancy (WHO)

Location & Time	Economy & Environment	
Country	**GDP per capita**	Measured in USD
Region	**Total Population**	Measured in Millions
9 unique regions	**Economy Status Developed (1 = developed)** **Economy Status Developing (0 = developed)**	Indicator variables with opposite meanings
Year	**Schooling**	Average years in education at 25+ years old

Death	Immunization	Diet	
Per 1000 People in Population	% Coverage among 1 year olds	**Alcohol Consumption**	Liters consumed per capita by those 15+ years old
Under Five **Infant** **Adult** **Life Expectancy** (Average across gender)	**Hepatitis B** **Measles** **Polio** **Diphtheria** **Incidents HIV**	**BMI**	Nutritional status in adults. (weight / height^2)
		Thinness (10 - 19 yrs) **Thinness (5 - 9 yrs)**	Proportion of thinness in the population of the corresponding ages

Main Questions

Prediction

What health predictors contribute to a countries average life expectancy?

Prediction Model: Predict life expectancy given these predictors.

Hypothesis : Significant predictors will be...

- Mortality rates
- BMI
- Vaccination rates

Classification

What health related factors should countries focus on to exceed the world average GDP per capita?

Classification Model: Classify a country above or below the world average given these predictors?

Hypothesis: Significant predictors will be...

- Education
- Mortality Rates
- Vaccination Rates

Methods – Test & Training Sets

Full Dataset: Included information from 179 countries over the years 2000 - 2015

- Total number of observations = 2,864
- Each observation = predictors from one country during one year.
- Removed one predictor variable, Economy_status_Developed.
- To prevent using time series data we performed the following changes.

Training Set: 2000-2014 Averaged Data

For each country

- Gather all observations from 2000-2014.
- Take the average of each predictor from those years.
- Place averages into a single observation and add it to the training set.

Test Set: 2015 data

For each country,

- Place observation for 2015 into the testing set.

Training & Testing Set Dimensions :

179 observations (countries) by 19 predictors (removed year)

Question #1

What health predictors contribute to a countries average life expectancy?

Prediction Model

Can we predict life expectancy given these predictors?

Exploring Data: Life Expectancy

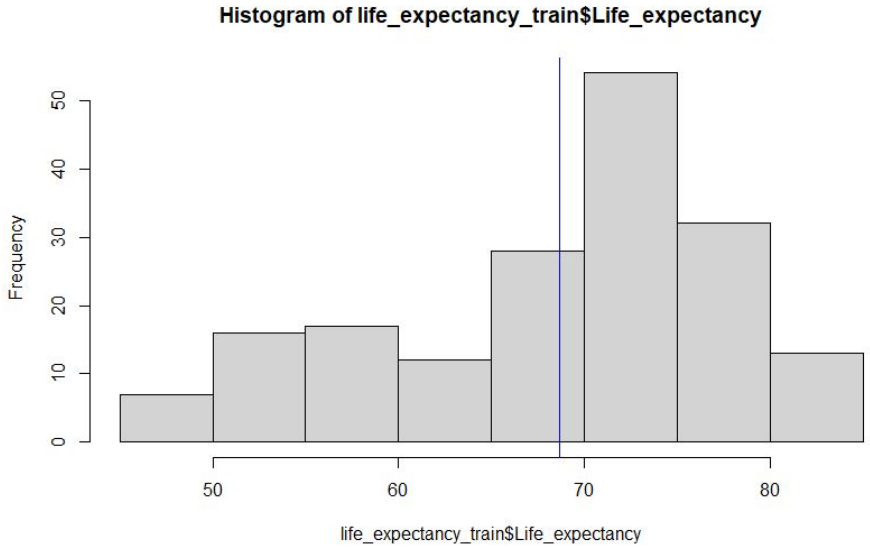
- **Life Expectancy**
  - Life expectancy refers to the average number of years a person is expected to live based on statistical data.
  - It is typically calculated at birth and can vary depending on various factors such as
    - Gender
    - Location
    - Lifestyle

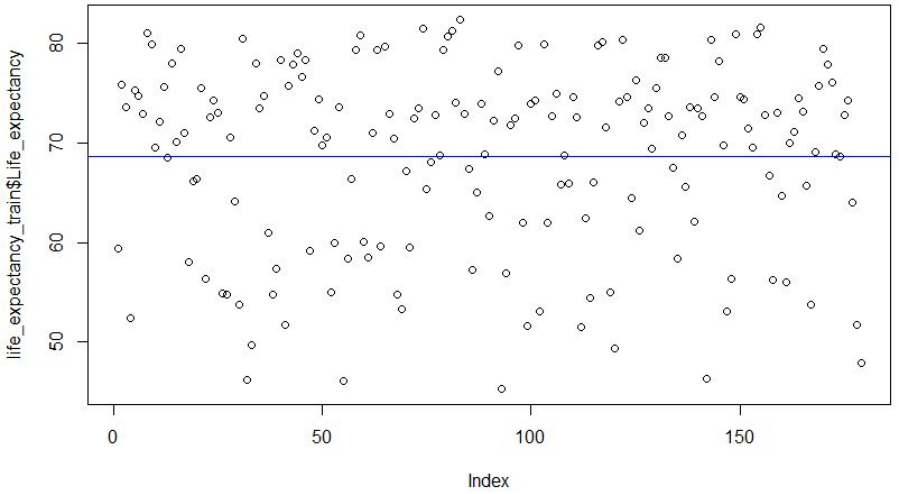
(Training Set)	Average	Minimum	25th Percentile	Median	75th Percentile	Maximum
Life_Expectancy	68.68223	45.24667	62.05000	71.46667	74.90000	82.36667

Goal : Identify health predictors that contribute to a countries average life expectancy.

Life Expectancy:

Observations:



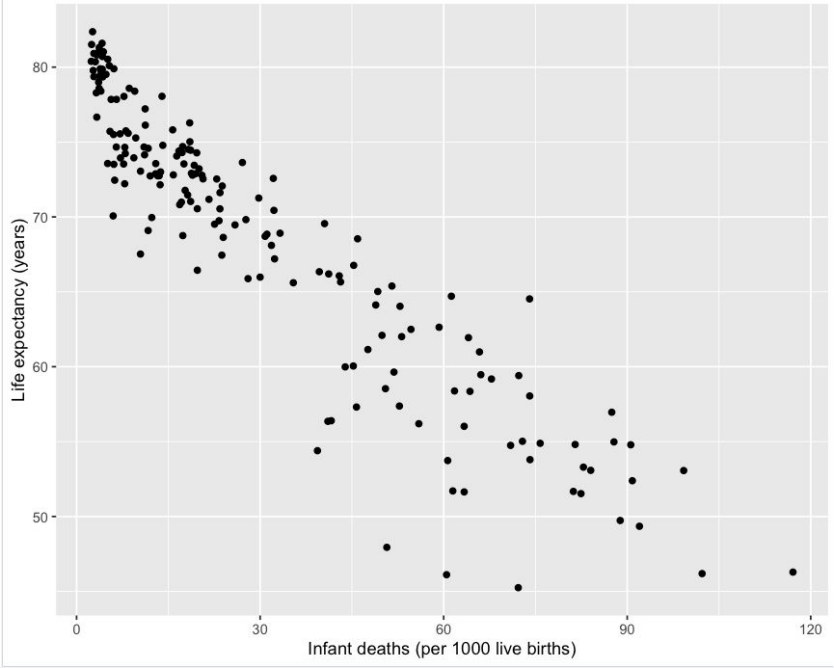


Exploring Our Hypothesis

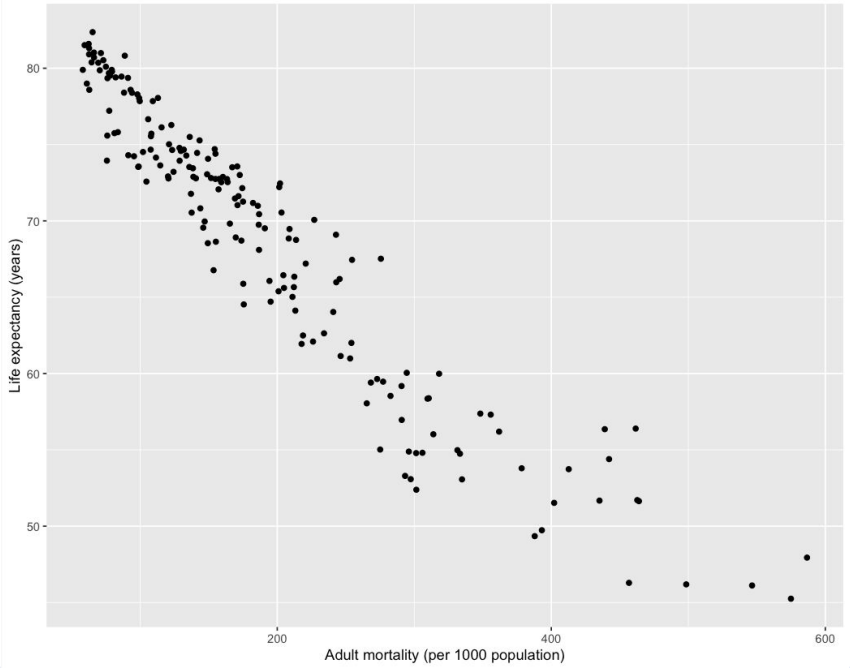
Hypothesis: Significant predictors will be...

- Mortality rates
- BMI
- Vaccination rates

Infant deaths vs. Life Expectancy



Adult mortality vs. Life Expectancy

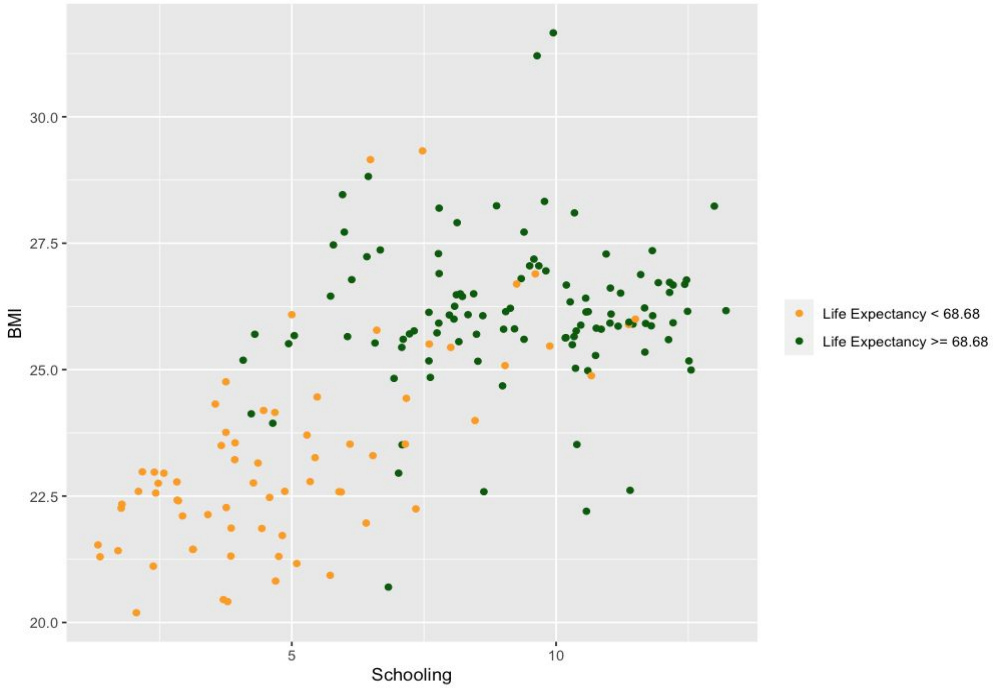


Exploring Our Hypothesis

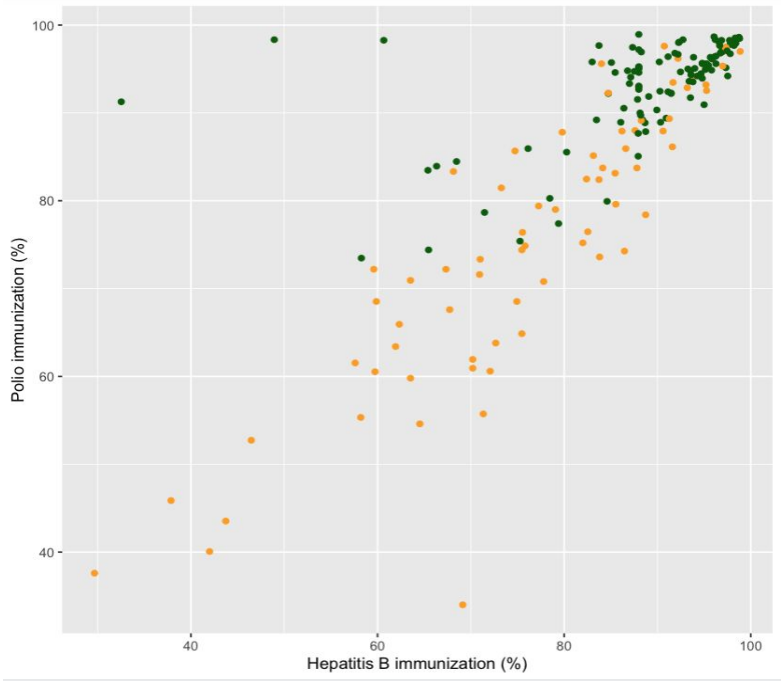
Hypothesis: Significant predictors will be...

- Mortality rates
- BMI
- Vaccination rates

Years of Schooling vs. BMI



Hep B Immunization vs. Polio Immunization



Prediction:

Results

Answer Questions

Question 1: What health predictors contribute to a countries average life expectancy?

Prediction: Given a set of health-related predictors, what is a country's average life expectancy?

- Infant_deaths
- Under_five_deaths
- Adult_mortality
- Alcohol_consumption
- Hepatitis_B
- BMI
- Diphtheria
- Incidents_HIV
- GDP_per_capita
- Population_mln
- Thinness_ten_nineteen_years
- Schooling

Predicted World Average:

71.11032

Test Set World Average:

71.46369

Methods/Processes

Methods we tried:

- Linear Regression
- Ridge
- Lasso
- Best Subset (AIC, BIC, CP, adjr2)
    - **BIC: M5** AIC: M9 cp: M8 adjsr2: M10
    - We chose BIC because it had the smallest test MSE with the chosen predictors
    - BIC is often preferred when the sample size is smaller
- CV with best subset selection

Methods/Processes: Results

Linear Regression

- Test MSE = 2.1146

Ridge Regression

- Test MSE = 2.4130

Lasso

- Test MSE = 2.1378

Best Subset (BIC → M5)

- Test MSE = 2.1225

Methods/Processes: Lasso

After performing model selection Lasso kept the following 10 predictors:

- Infant_deaths
- Under_five_deaths
- Adult_mortality
- Alcohol_consumption
- Hepatitis_B
- BMI
- GDP_per_capita
- Thinness_ten_nineteen_years
- Schooling

Kept all the same predictors as our optimal model except for Diphtheria and Population

Unexpected Results

- Although all of our test MSE's were low, least squares had a slightly lower one than both ridge and lasso
- Reasons for this could be:
  - Our dataset doesn't have many predictors so regularization might not help as much
  - There is a linear correlation between our predictors so ridge/lasso are overfitting the data

Methods/Processes: CV + Best Subset

Combining CV with 'best' subset selection

1. Use the "regsubsets" function on the training data to perform best subset selection.
2. Set up a loop to run cross-validation on each model subset, and record the validation error for each.
3. Identify the model subset with the smallest validation error, and fit a model using that subset of predictors.
4. Calculate the predicted values using the test data, and compute the test MSE to evaluate the model's performance on the test data.

```
> summary(model_train)

Call:
lm(formula = Life_expectancy ~ Infant_deaths + Under_five_deaths +
    Adult_mortality + Alcohol_consumption + Hepatitis_B + BMI +
    Diphtheria + Incidents_HIV + GDP_per_capita + Population_mln +
    Thinness_ten_nineteen_years + Schooling, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1701 -0.7887 -0.0317  0.7096  3.1950

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.700e+01  2.426e+00  35.863  < 2e-16 ***
Infant_deaths               -3.357e-02  2.373e-02  -1.415 0.159074
Under_five_deaths           -7.135e-02  1.511e-02  -4.723 4.91e-06 ***
Adult_mortality             -4.808e-02  2.341e-03 -20.534  < 2e-16 ***
Alcohol_consumption          1.130e-01  3.387e-02   3.338 0.001041 **
Hepatitis_B                 -2.693e-02  1.269e-02  -2.122 0.035323 *
BMI                         -1.979e-01  7.161e-02  -2.763 0.006375 **
Diphtheria                   1.700e-02  1.653e-02   1.029 0.305096
Incidents_HIV                8.751e-02  6.824e-02   1.282 0.201460
GDP_per_capita               2.967e-05  7.915e-06   3.749 0.000245 ***
Population_mln              -3.850e-04  7.357e-04  -0.523 0.601478
Thinness_ten_nineteen_years -4.827e-02  3.525e-02  -1.370 0.172689
Schooling                    2.571e-02  6.412e-02   0.401 0.688916
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 166 degrees of freedom
Multiple R-squared:  0.9837,    Adjusted R-squared:  0.9825
F-statistic:   834 on 12 and 166 DF,  p-value: < 2.2e-16
```

Test MSE: 2.103 (final decision)

Methods/Processes: Forward Selection

1) Adult Mortality
    a) A higher mortality rate brings down life expectancy
2) Infant Deaths
3) GDP
    a) A lower GDP per capita means less government revenue and less spending on programs such as healthcare and education
4) Under 5 deaths
5) Alcohol Consumption
    a) Alcohol is considered one of the most harmful substances to the user and to others

It is important to consider forward selection because it allows us to systematically explore the space of possible models and identify the most important predictors for a given response variable

Justification

Reason for choosing CV + best subset

- Gives the **lowest** Test MSE
- **Prevent Overfitting**: The method identifies the best subset of predictors that provides the best trade-off between bias and variance
- **Flexibility**: The method allows for the inclusion of a large number of predictors and the selection of the best subset of predictors based on their performance
- **Accuracy**: The method helps to identify the most important predictors that are relevant to the outcome variable, and can remove irrelevant predictors

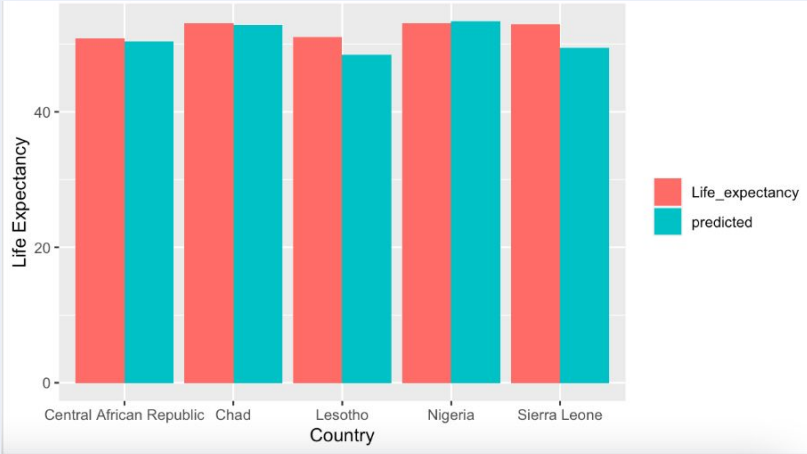
Prediction:

Summary

Final Summary & Answer Questions

Lowest Life Expectancy

Predicted	Test (2015)
Lesotho	Central Africa Republic
Sierra Leone	Lesotho
Central Africa Republic	Sierra Leone
Chad	Chad
Nigeria	Nigeria

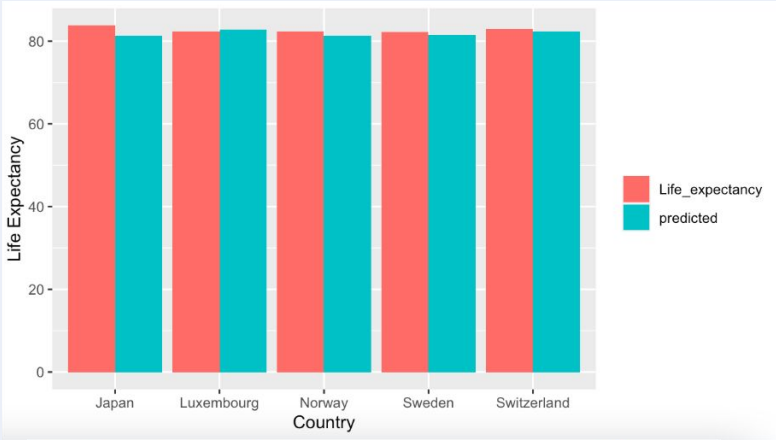


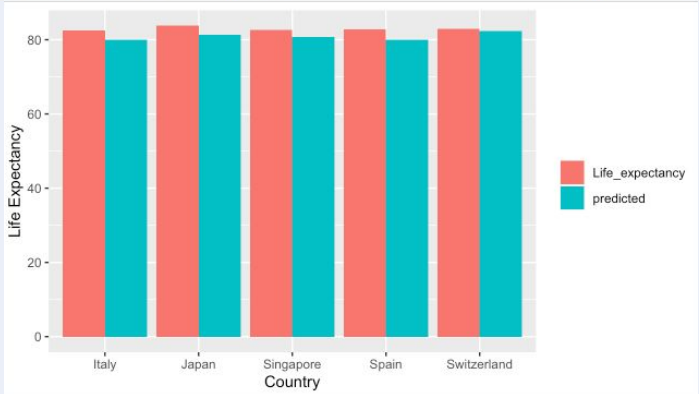
Prediction

Final Summary & Answer Questions

Highest Life Expectancy

Predicted	**Test (2015)**
Luxembourg (9th in test)	Japan
Switzerland	Switzerland
Sweden (11th in test)	Spain
Norway (8th in test)	Singapore
Japan	Italy





Prediction

Compare the Actual Result with Our Hypothesis

Actual Result (CV + best subset)

- Infant_deaths (2)
- Under_five_deaths (4)
- Adult_mortality (1)
- Alcohol_consumption (5)
- Hepatitis_B
- BMI
- Diphtheria
- Incidents_HIV
- GDP_per_capita (3)
- Population_mln
- Thinness_ten_nineteen_years
- Schooling

Forward Selection

1. Adult Mortality
2. Infant Deaths
3. GDP
4. Under five deaths
5. Alcohol Consumption

Our Hypothesis:

- Mortality rates
- BMI
- Vaccination rates

To Conclude:

Based on the information provided, it seems that the hypothesis was supported by the results, as the predicted life expectancy was fairly close to the test set, and the variables deemed important were selected using cross-validation + best subset selection.

Question #2

What health related factors should countries invest in to meet the world average GDP per capita?

Classification Model

Can we classify a country above or below the world average given these predictors?

Exploring Data: GDP per capita

- **GDP = Gross Domestic Product**
  - Measures monetary value of **all final goods and services.** Including...
    - how consumers spend on goods
    - how much consumers and companies invest in the stock market
    - how much the government spends
    - net imports and exports

- **GDP per capita =** Measure of how much the **average individual** within that country is spending in terms of consumption and investment and what the government is supplying to them in terms of good and services.

(Training Set)	Average	Minimum	25th Percentile	Median	75th Percentile	Maximum
GDP per Capita	12617.3	255.5	1391.20	4335.20	11909.03	102806.7

Goal : Identify health predictors that contribute to a countries GDP per capita being above or below the world average (12617.3).

Classification
Above 12617.3 = 1 Below 12617.3 = 0

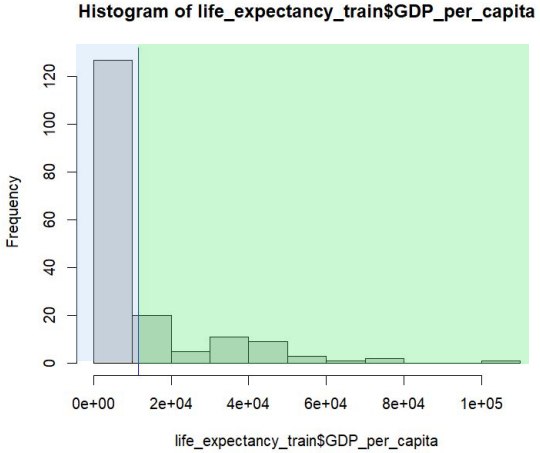
GDP Per Capita:

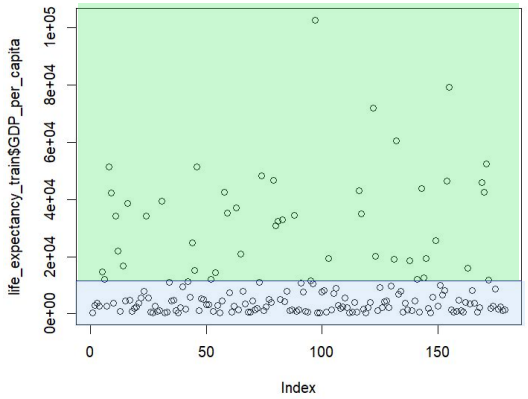
Large percentage of the world's countries are below the mean world GDP per capita. (133/179 = 0.74%)

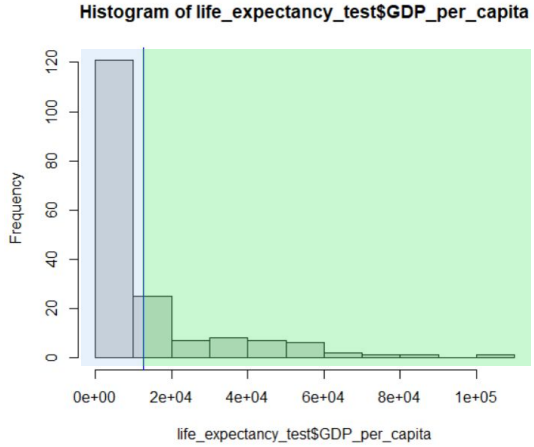
We see a heavy right skewed distribution for countries GDP per capita.

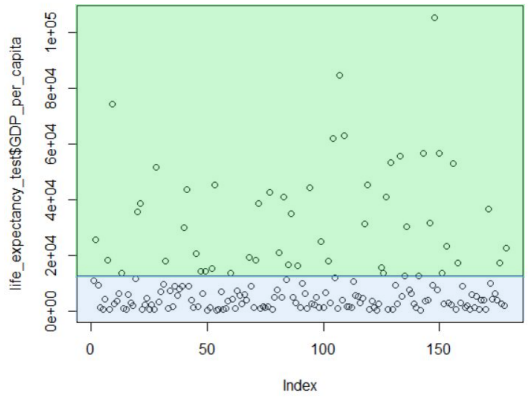
Blue
Below the world average

Green
Above world average







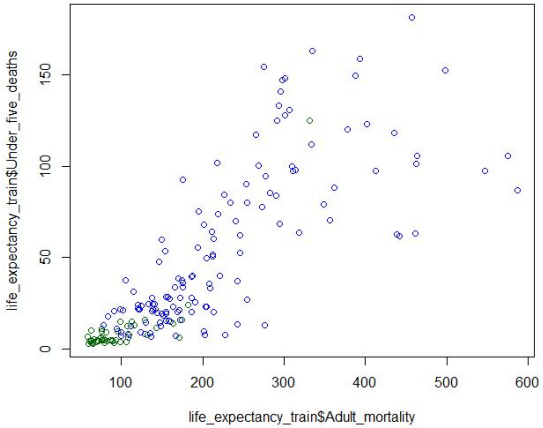


Exploring Our Hypothesis

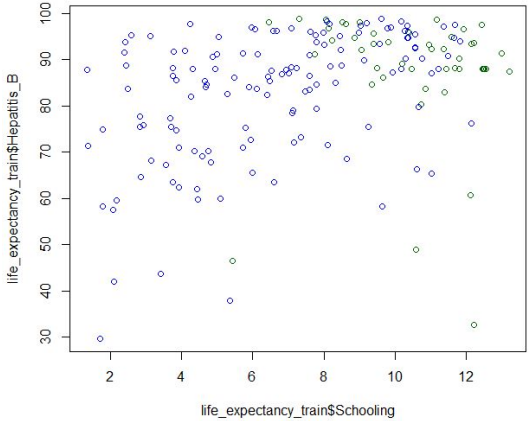
Hypothesis: Significant predictors will be...

- Education
- Mortality Rates
- Vaccination Rates

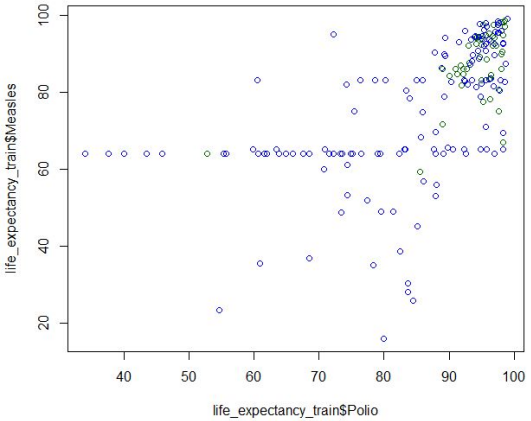
Adult mortality vs. Child mortality



Schooling vs. Hepatitis B



Polio vs. Measles



We see some grouping but many of the predictors do not have a clear decision boundary in regards to classification of the country having a GDP above or below the world average.

Big Picture

Classification: What health related factors should countries invest in to meet the world average GDP per capita?

- Enables countries to focus on health related factors that can increase their GDP per capita or maintain it above the average.
- Can optimize and invest in these areas.
- Higher GDP per capita
  - Attracts business
  - Enables spending and investment
  - Productive economy
  - Improved quality of life

Classification:

Results

Answer Questions

Question 2: What health related factors should countries invest in to meet the world average GDP per capita?

Classification: Given a set of health related predictors, is a countries GDP per capita above or below the world average?

- Under_five deaths
- Population_mln:
- Infant_deaths:
- Hepatitis_B
- BMI

Classification

Methods/Processes

We attempted to use:

- **Decision Tree** for classification and model selection.
  - Create a simple decision tree
  - Create an ensemble decision tree
  - Compare their RSS values, lowest RSS is the best at classification
  - Use the model corresponding to the optimal tree for KNN
- **KNN** for classification.
  - Using the predictors from the tree model...
    - Perform cross validation to find the optimal number of neighbors (k)
    - Predict on the test set.
- We choose to use a combination of the two because we wanted to be able to track feature selection.

Methods – Test & Training Sets

- Training and Test Sets have the following modifications.
  - **Created** a new variable = above_or_below_GDP
  - **Removed** GDP_per_capita
    - Do not want to use this as a predictor as it technically is what we are calculating for our response.
  - **Did not use** country or region when training the models

- Preformed Classification on Training Set
  - Response : above_or_below_GDP
  - Predictors: all besides GDP_per_capita, Country, Region

Training & Testing Set Dimensions :

179 observations (countries) by 19 predictors

Classification : Decision Trees

Goal: Create a classification model with a decision tree.

Simple - One Tree

- Simple decision trees are easily interpreted and we can understand how one predictor effects the response.
- Issue :
  - Weak learner
  - Highly flexible - high variance

Ensemble - Helps us reduce variance

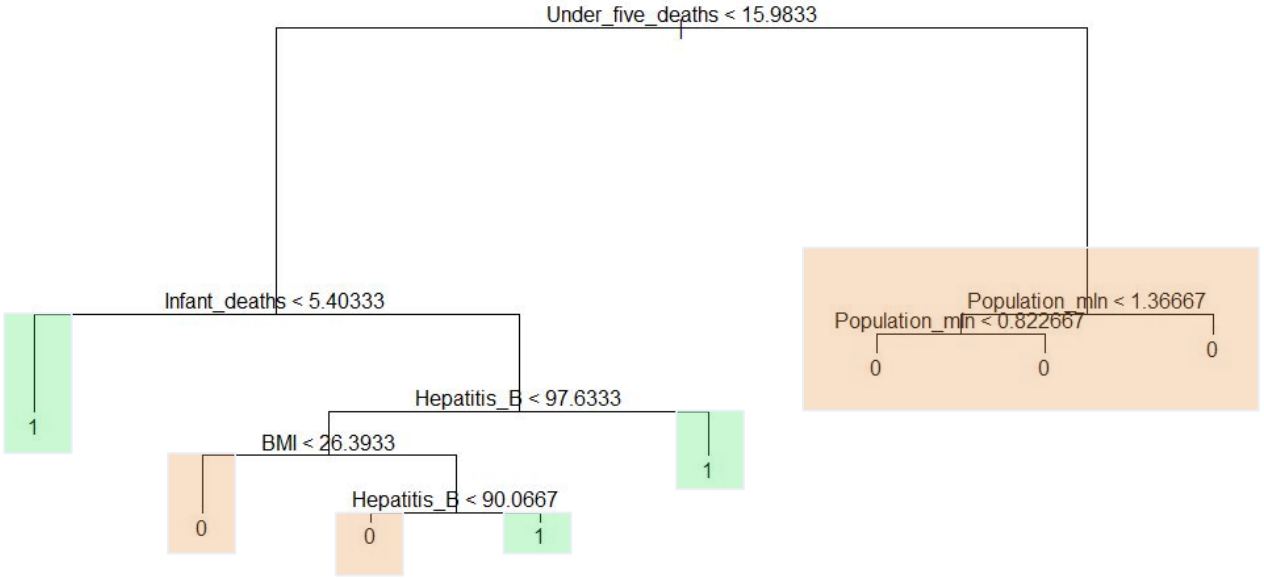
- Improves prediction accuracy, combines multiple modes into one.
- Issue:
  - Loss of interpretability
  - For our problem this poses challenges because we want to identify predictors

Classification: Decision Tree Attempt

Residual Mean Deviance: 0.1288
Number of terminal nodes : 8

Above

Below

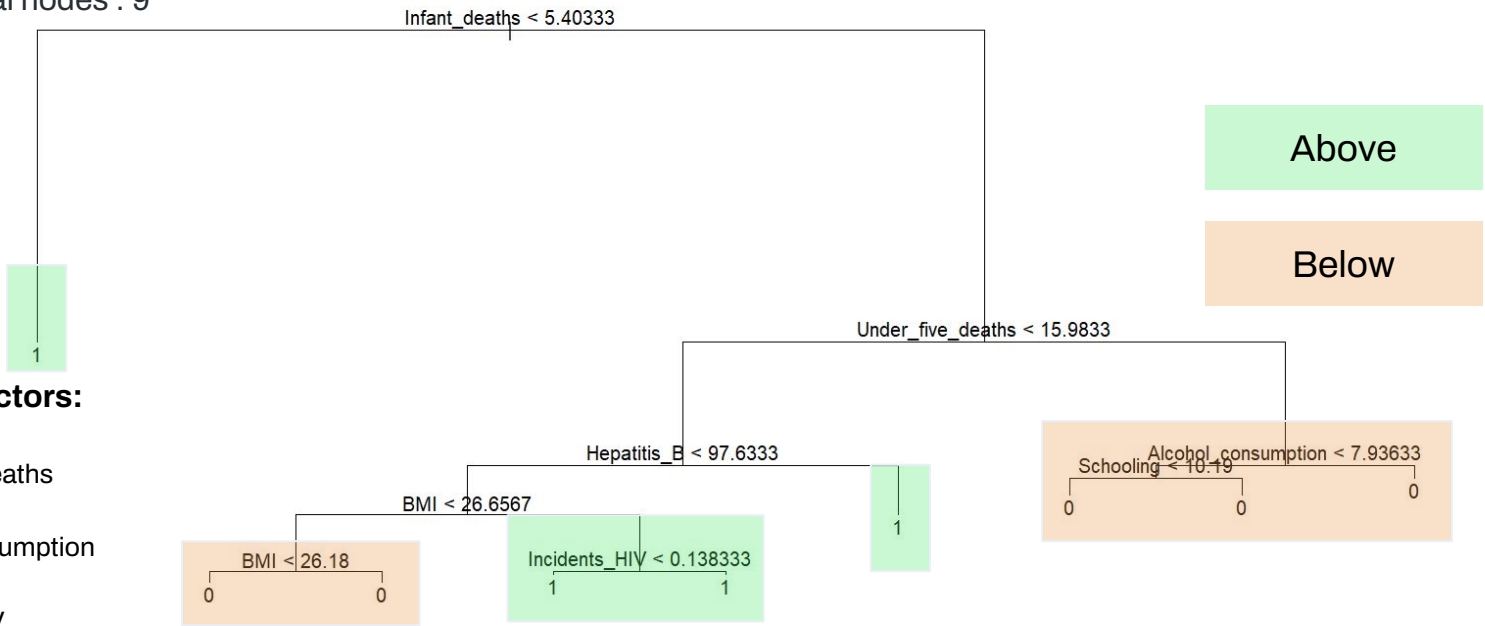


Significant Predictors:

- Under_five deaths
- Population_mln
- Infant_deaths
- Hepatitis_B
- BMI

Classification : Decision Tree - Ensemble

Residual Mean Deviance: 0.1672
Number of terminal nodes : 9



Significant Predictors:

- Infant_deaths
- Under_five deaths
- Hepatitis_B
- Alcohol Consumption
- Schooling
- Incidents_HIV
- BMI

Classification : *K*-Nearest Neighbor

Goal: Carry out KNN classification using the predictors found using a simple decision tree and ensemble tree and compare results

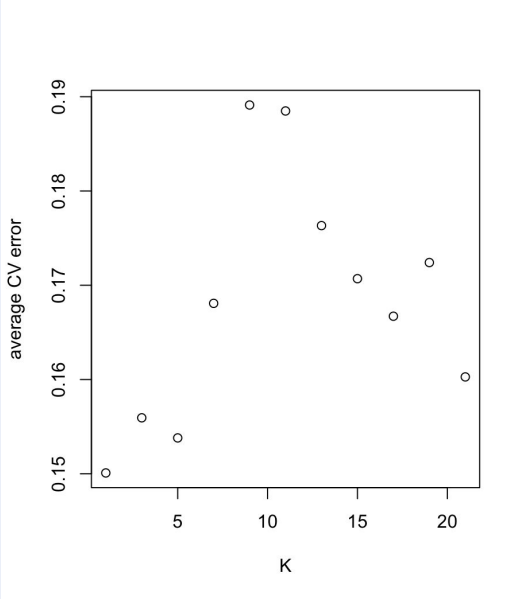
- KNN is able to carry out classification.
  - Pros
    - Non-parametric
    - Highly flexible - low bias
    - Has very irregular and flexible decision boundaries
    - Will give us misclassification error
  - Cons
    - Does not create a model, therefore no feature selection
    - Highly flexible - high variance

Classification : *K*-Nearest Neighbor

Identify optimal *K*

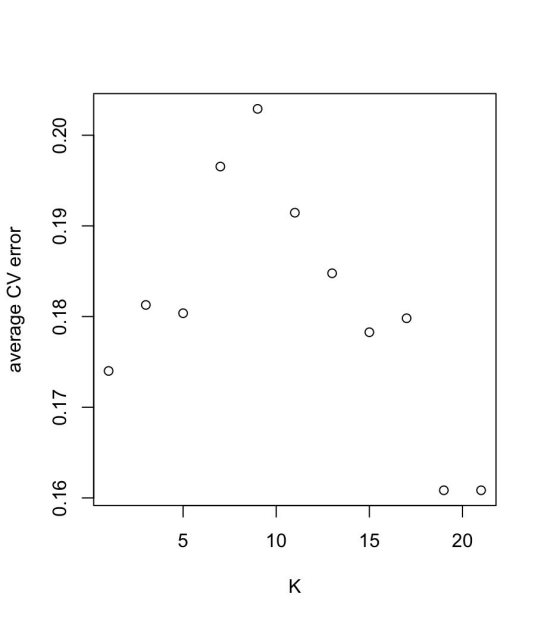
- Use 10-fold cross validation to identify optimal *K*
- Selected from *K* = {1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21}

Using Predictors from Simple Tree



K = 1

Using Predictors from Ensemble Tree



K = 19

Note: higher K = higher bias/ lower variance

Classification: *K*-Nearest Neighbor

From Simple Tree

	Below Avg. GDP	Above Avg. GDP
Below Avg. GDP	116	16
Above Avg. GDP	12	35

Misclassification rate = 0.1564246

From Ensemble Tree

	Below Avg. GDP	Above Avg. GDP
Below Avg. GDP	118	18
Above Avg. GDP	10	33

Misclassification rate = 0.1564246

Classification : False +/-

False Positive : Given a countries predictors, we classify them as being over the world average GDP per capita when they actually are not.

False Negative : Given a countries predictors, we classify them as being under the world average GDP per capita when they actually are.

False + is worse: Tell a country to invest in one area and then they do not see a return on investment in the form of GDP per capita over the world average.

False positive rate for classification using predictors from simple tree:

- 0.1212121

False positive rate for classification using predictors from ensemble tree:

- 0.1323529

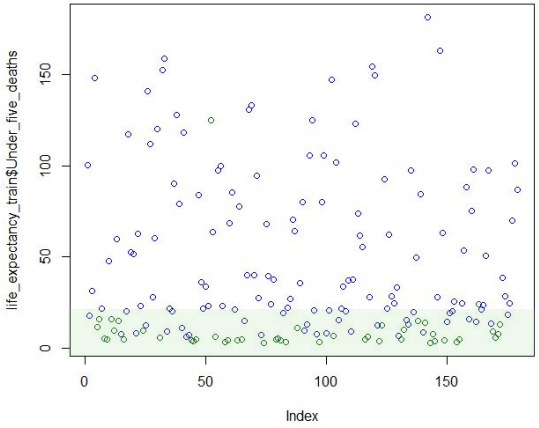
Classification Results

We saw that the simple tree had a lower RSS and false positive rate at the cost of having more variance.

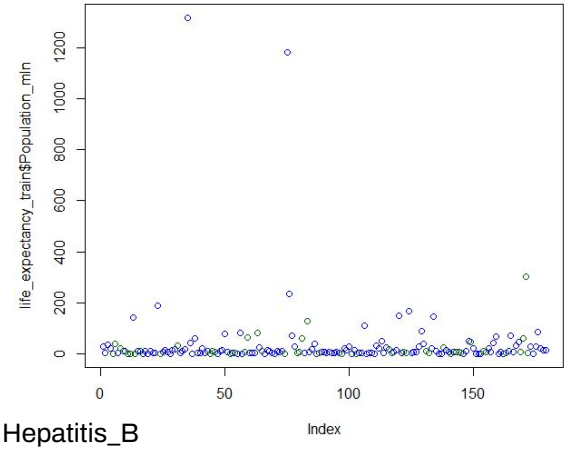
Tree Classifier	Simple Tree	Ensemble
Residual Mean Deviance	**0.1288**	0.1672
KNN		
# Neighbors	1 (High variance)	19 (High bias)
Misclassification Error	**0.1564246**	**0.1564246**
False Positive Rate	**0.1212121**	0.1323529
Choose the simple decision tree as our final classification model		

Viewing Predictor Tendencies

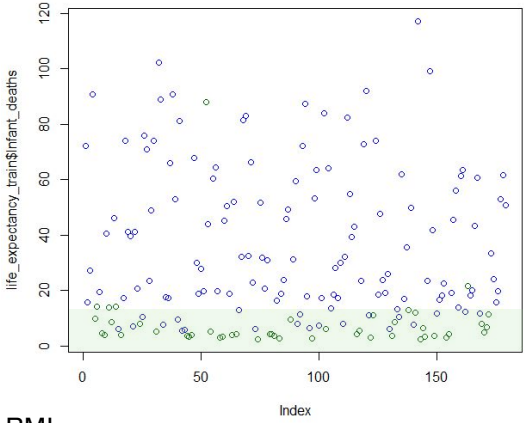
Under_five_deaths



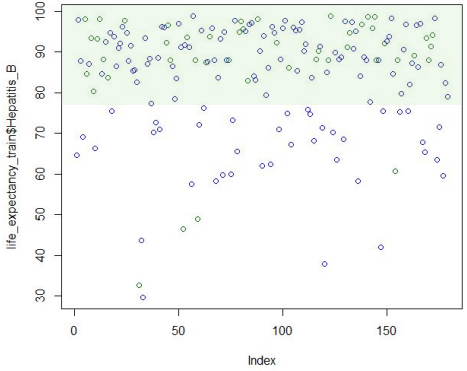
Population_mln



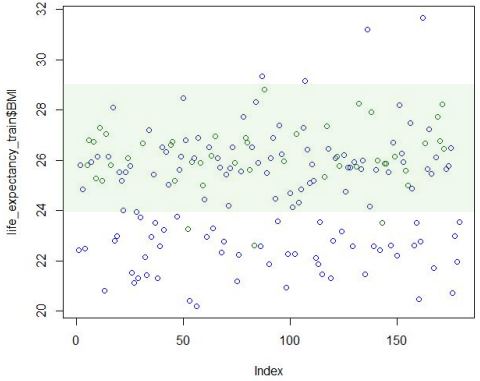
Infant_deaths



Hepatitis_B



BMI



Significant Predictors:

- Under_five deaths
- Population_mln
- Infant_deaths
- Hepatitis_B
- BMI

Colors indicate classification, data is from training set.

Green = Above
Blue = Below

Unexpected Results

- We expected that the misclassification rate of the classification using predictors from the simple tree would be higher than that using predictors of the ensemble tree
  - Because ensemble classification trees have more power and higher prediction accuracy than one simple tree
  - Instead they were equal

- Selected Population_mln in our model but virtually all population sizes go to 0 (under the average)
  - This is odd because we would expect some level of population to correlate with above the average as well.
  - Also saw that there was not a clear grouping for this in the previous plots

Classification:

Summary

Final Summary & Answer Questions

Question 2: What health related factors should countries invest in to meet the world average GDP per capita?

		Decision Boundary (to be classified as over world average GDP)
Mortality	Under_five deaths	Under 15.98 (per 1000 People in Population)
	Infant_deaths	Under 5.4 (per 1000 People in Population)
Vaccination	Hepatitis_B	Over 97.63% (Coverage among 1 year olds)
Lifestyle & Environment	BMI	Under 26.39
	Population_mln	Saw virtually all population sizes go to 0

Compare the Actual Result with Our Hypothesis

Our hypothesis was generally correct.

- Two mortality predictors
- One vaccination predictor

Although we did not see any significance to the schooling predictor and did not think that population or BMI would affect these results

Hypothesis: Significant predictors will be...

- Education
- Mortality Rates
- Vaccination Rates

From Our Results

- Under_five deaths
- Population_mln:
- Infant_deaths:
- Hepatitis_B
- BMI

Potential Error

Strength

- KNN is a powerful model for nonparametric data that may need irregular decision boundaries
- Because KNN cannot do model selection, we can use

Weakness

- We used the average GDP_per_capita, using the median may have been better for comparison
    - Is a better indicator of comparison because the average can be inflated to one side to to extremely wealthy countries.
- GDP per capita can only tell us a small amount about a country's economy so our recommendations should be taken into consideration with other factors.
- Both models have high variance.

Prediction & Classification:

Summary

Final Results

Prediction

What health predictors contribute to a countries average life expectancy?

- Adult_mortality (1)
- Infant_deaths (2)
- GDP_per_capita (3)
- Under_five_deaths (4)
- Alcohol_consumption (5)
- Hepatitis_B
- BMI
- Diphtheria
- Incidents_HIV
- Population_mln
- Thinness_ten_nineteen_years
- Schooling

Classification

What health related factors should countries focus on to exceed the world average GDP per capita?

- Under_five deaths
- Infant_deaths
- Hepatitis_B
- BMI
- Population_mln

Practical Insights & Further Questions

Both life expectancy and GDP per capita can be improved by similar efforts to improving healthcare, lifestyle, and the economy.

Improve mortality rates

- Countries should prioritize improving adult, child, and infant mortality rates.

Increase Vaccination

- A country may benefit from research about vaccination development to prioritize spending and effort.

Study diet and lifestyle

- A country may benefit from improving their population diet and lifestyle habits.

- What can be done to improve healthcare and lower mortality rates.
- How can vaccination rates be improved?
  - What diseases are most important to protect against?
- Should more studies be done on weight and diet and its impact on individual longevity and prosperity of a country?
- How or what the Countries can do to improve?
- Do they know the problems?
- What can affect those top predictors?

Thanks!

Questions?

References

Slide Theme :
https://slidesgo.com/theme/vital-signs-assessment-case-study#search-Health&position-63&results-1434

Dataset:

https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated