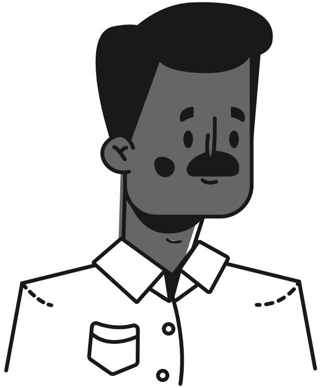
Life Expectancy

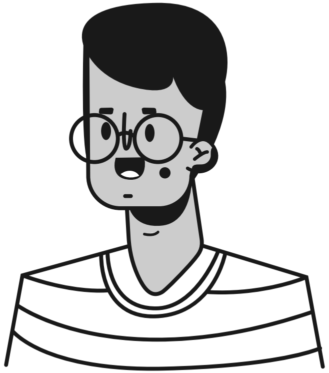
DSC 532 – Statistical Learning – Regression Problem

Andreas Papadopoulos



Foivos Lympouras

Constantinos Constantinou



Iosif Pintirishis

Contents

01

Introduction

02

Our Data Set

03

Data Pre - Processing

04

Exploratory Data Analysis

05

Feature Selection

06

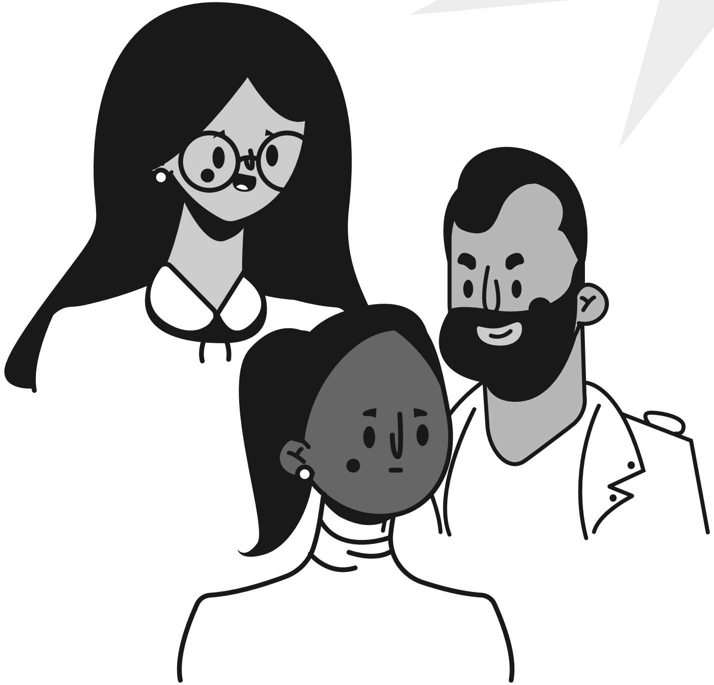
Models

07

Conclusion

01

Introduction



Investigating a wide array of factors impacting life expectancy, from health metrics to economic and social variables

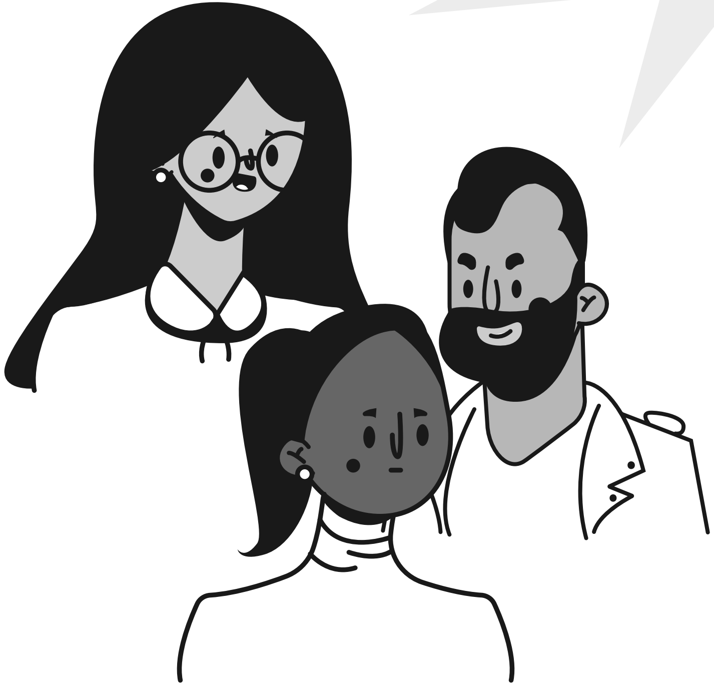
Utilizing a diverse dataset to uncover critical influences on life expectancy in various countries.

Aiming to offer actionable insights for nations to enhance citizens' longevity.

Predictive goal: Estimate life expectancy at birth using a robust set of predictors

02

Our Data Set



Our Data Set

Main file:

From a Kaggle competition

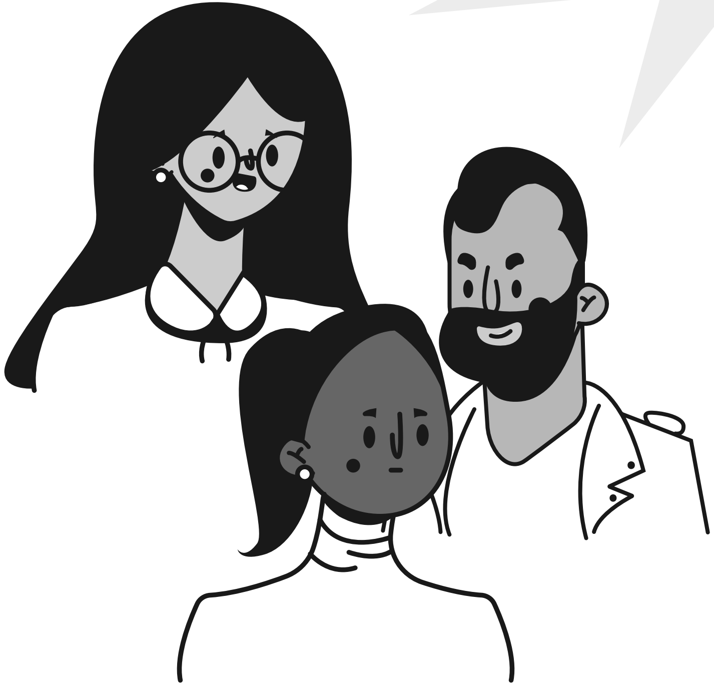
Other sources :

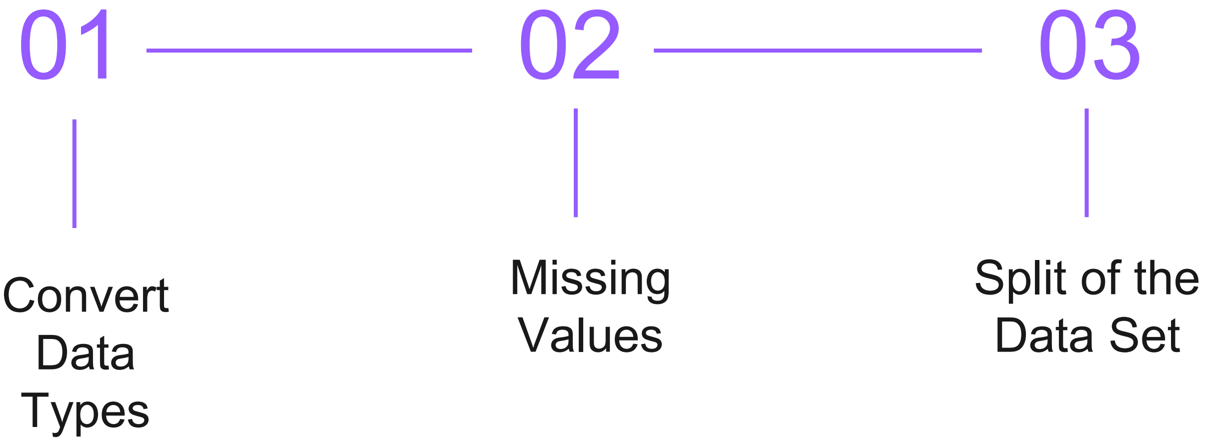
The Global Health Observatory (GHO) data repository under World Health Organization (WHO) and databank.worldbank

- Timeframe: Covers 179 countries from 2000 to 2015.
- Key variables influencing life expectancy include:
  - Geographic regions and country status (Developed/Developing).
  - Health indicators: Child mortality (Infant and Under-five deaths), Adult mortality, and Immunization coverage (Hepatitis B, Diphtheria, Polio).
  - Lifestyle factors: Alcohol consumption, Mean BMI, and HIV incidence.
  - Socioeconomic factors: Schooling years, GDP per capita, and Population.

03

Data Pre - Processing





01

Convert Data Types

02

Missing Values

03

Split of the Data Set

Convert Data types

- Created *'Economy_status'* column to classify countries by economic status: '**Developed**' or '**Developing**’ and then we transformed it to categorical factor for further analysis.

- Converted *'Population*' and *'GDP*' data types to numeric

- Ensured data types align with modeling requirements, improving data quality and analysis readiness.

Missing values

- Detected missing values in several features (9) including *Adult Mortality*, *Alcohol*, and *Hepatitis B*

- We assume missing values follow the Missing Completely at Random (MCAR) mechanism

- Hepatitis B:
  - 550 missing values, 19.2%.
  - Used Predictive Mean Matching for the missing values in Hepatitis B

- The rest of the variables, which have below 2% missing values, were imputed using the median of the training set

Split of the data set

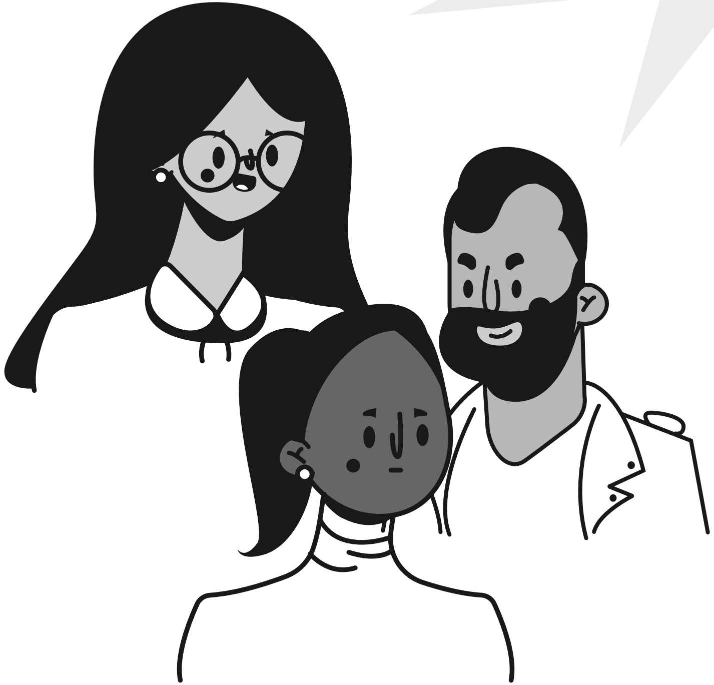
- Data split by year :

Training set (2000-2012)
Test set (2013-2015)

- 81.25% for Training set and 18.75% for Test set.

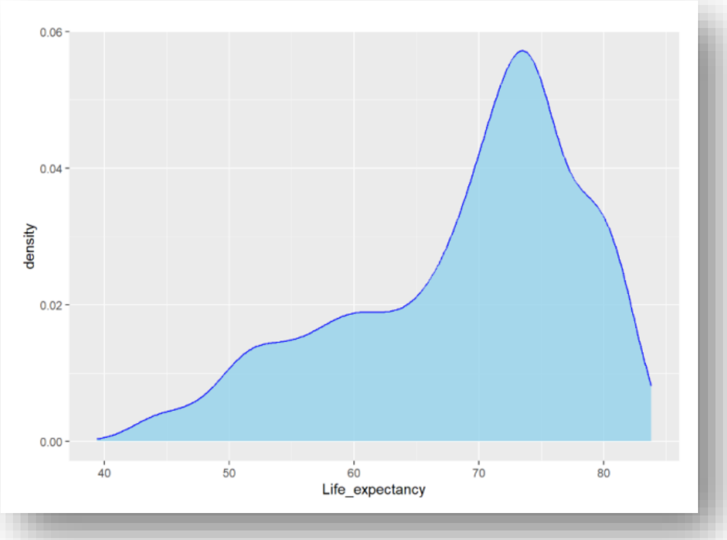
04

Exploratory Data Analysis

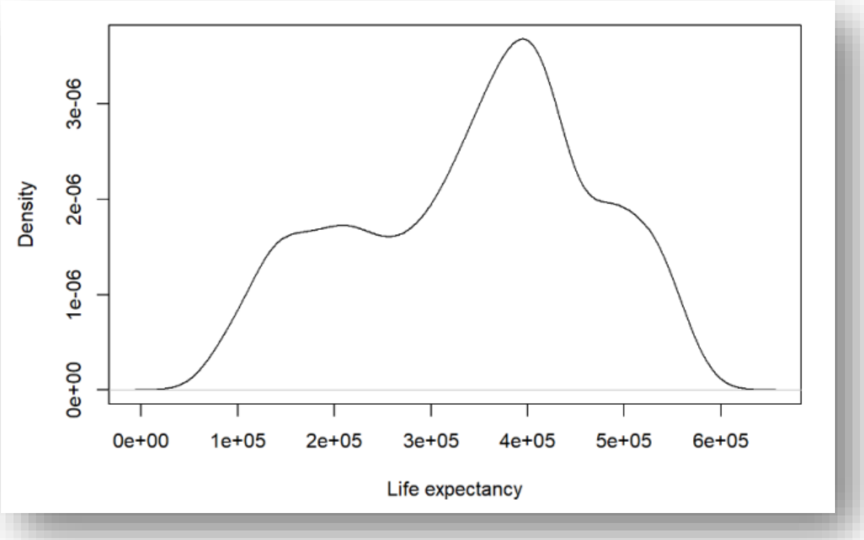


Life Expectancy

Life expectancy at birth, meaning the average number of years a newborn is expected to live



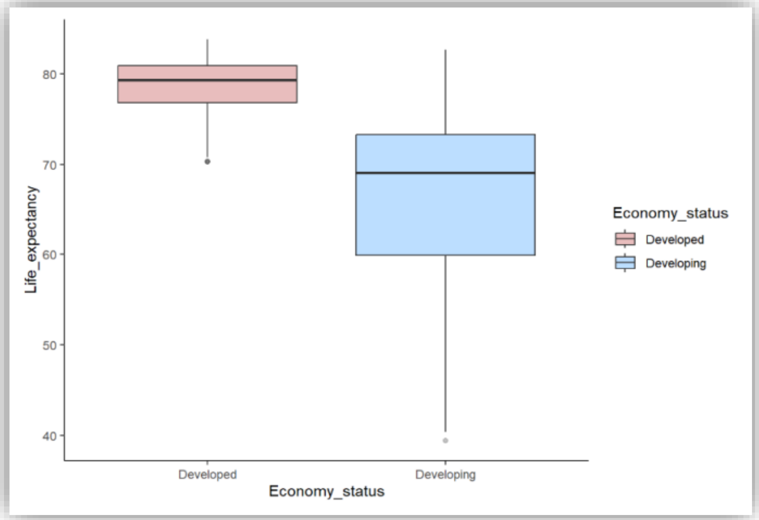
Density of Life Expectancy

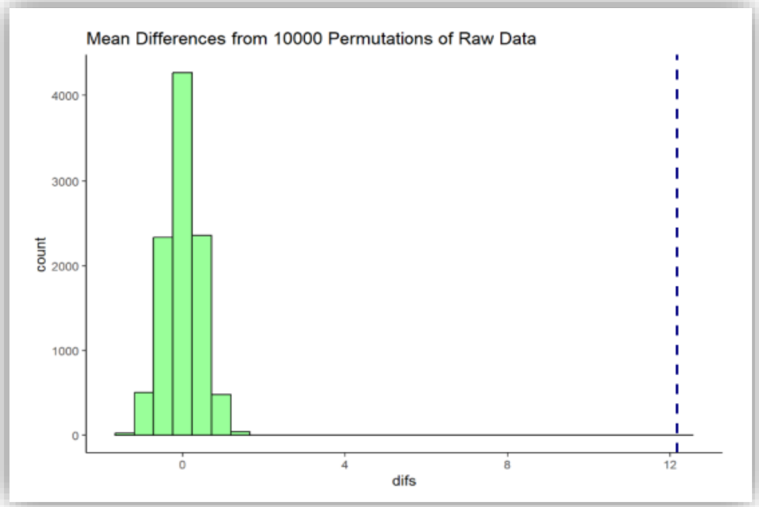


Density of $Life\ Expectancy^3$

Economy status

- Developed vs Developing countries
- Countries have been grouped according to their Gross National Income per capita
- People from Developed countries seem to have higher Life expectancy compared to people who come from Developing countries
- Permutation test to confirm
- We expect *Economy status* to be among the important features

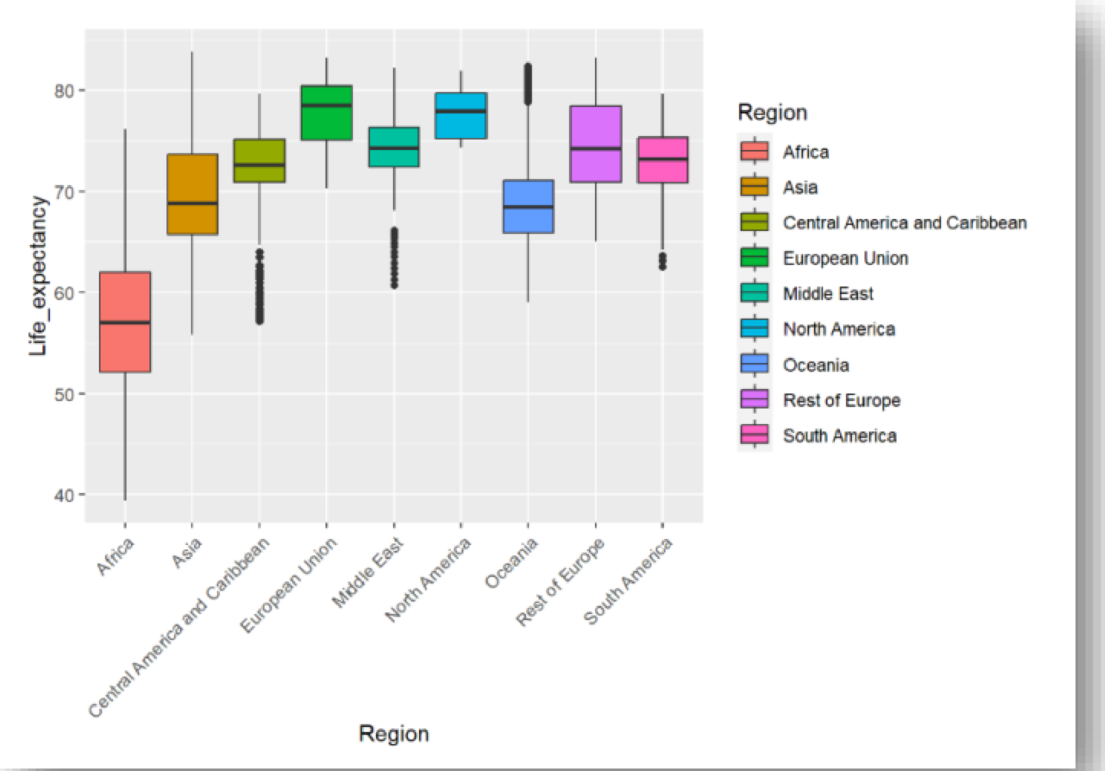




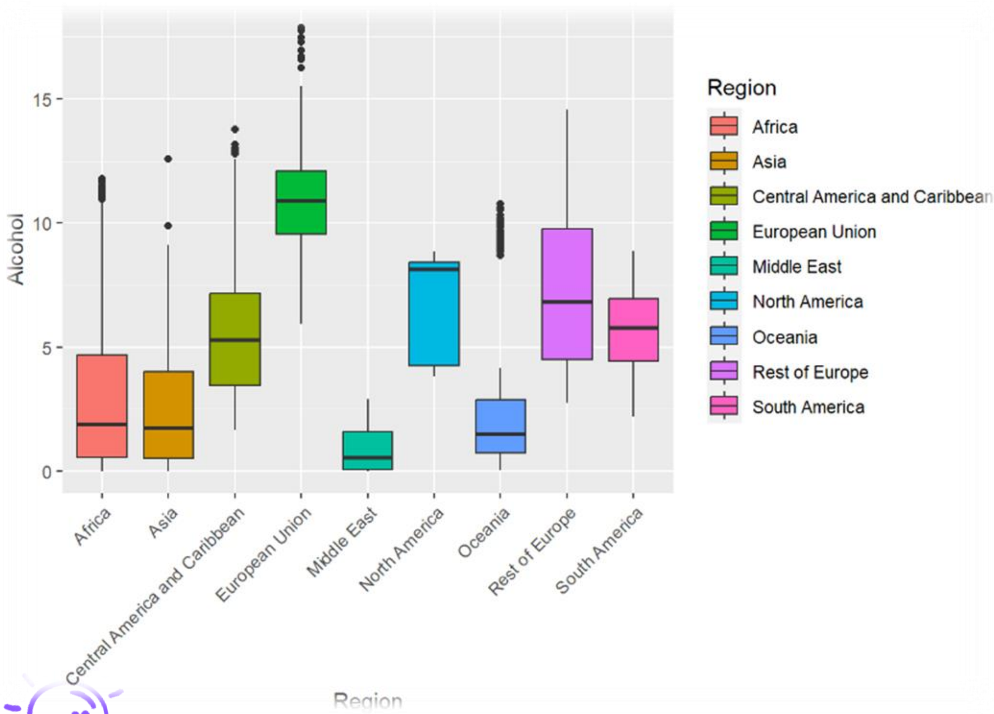
Region

- European Union and North America have the highest life expectancy

- Africa has the lowest with Asia being second lowest.

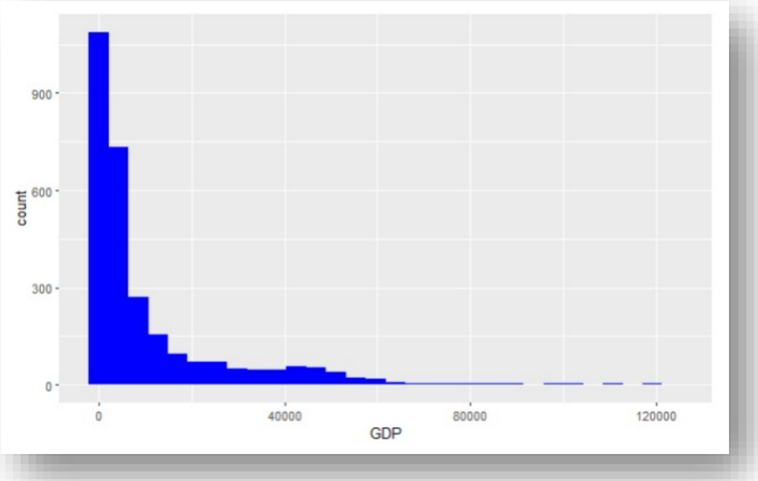


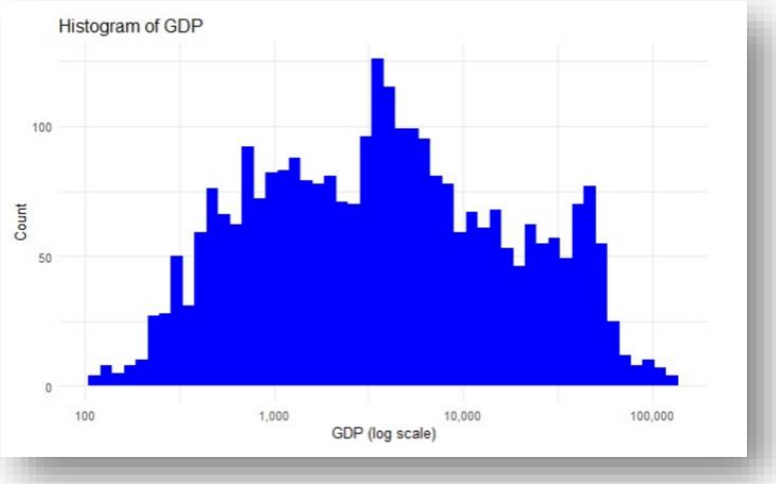
Alcohol Consumption



- The European Union has the Largest Alcohol Consumption
- The difference in consumption between regions is significant.

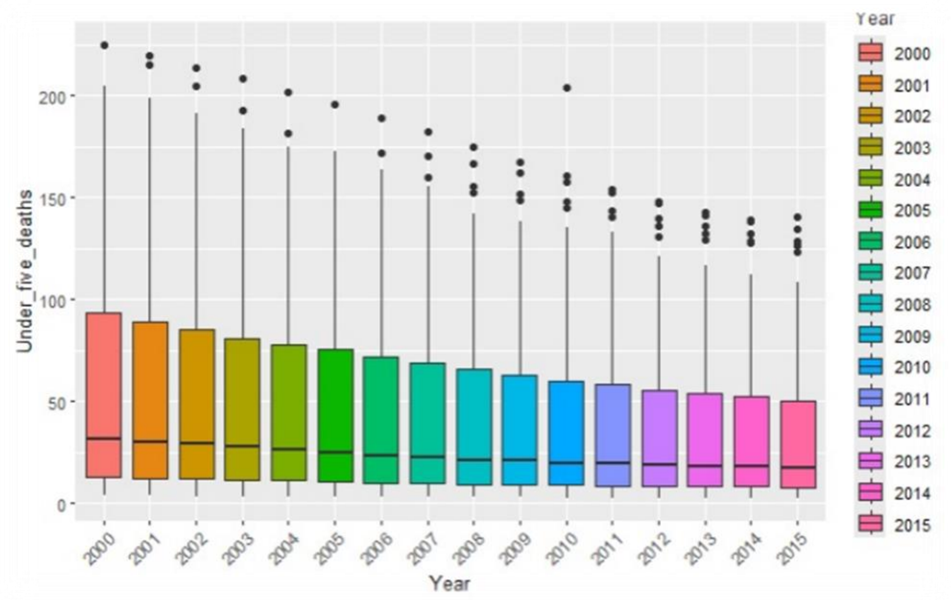
GDP per capita

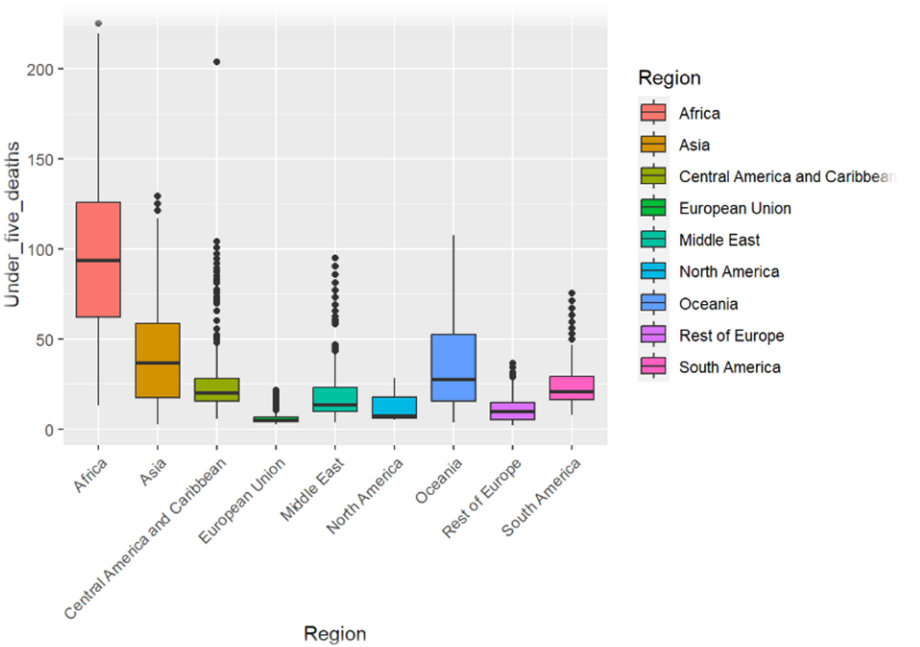




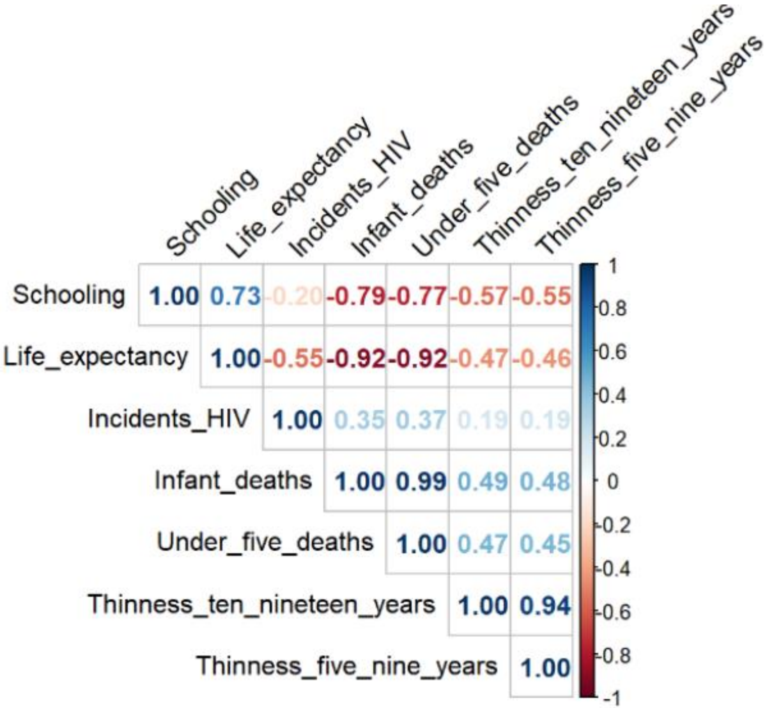
Gross Domestic Product (GDP) per capita shows a country's GDP divided by its total population

Under 5 deaths

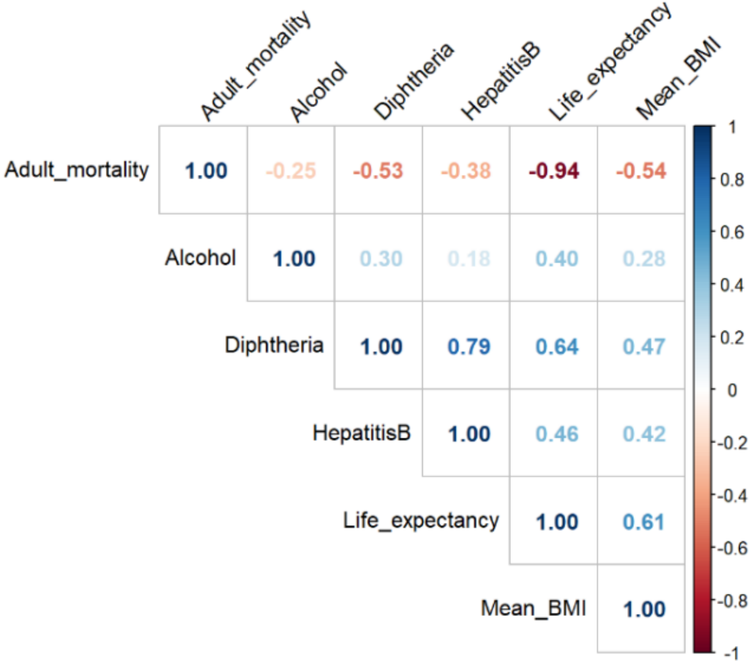




Correlations



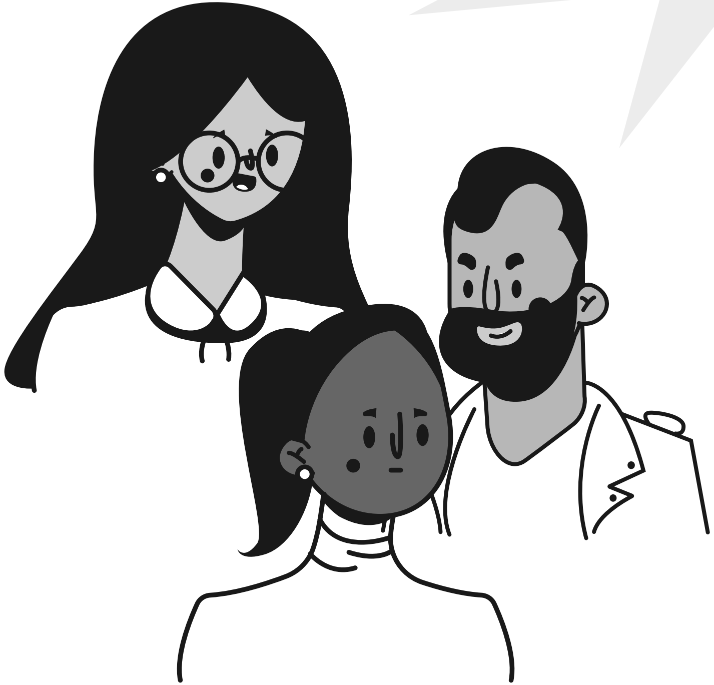
	Schooling	Life_expectancy	Incidents_HIV	Infant_deaths	Under_five_deaths	Thinness_ten_nineteen_years	Thinness_five_nine_years
Schooling	1.00	0.73	-0.20	-0.79	-0.77	-0.57	-0.55
Life_expectancy		1.00	-0.55	-0.92	-0.92	-0.47	-0.46
Incidents_HIV			1.00	0.35	0.37	0.19	0.19
Infant_deaths				1.00	0.99	0.49	0.48
Under_five_deaths					1.00	0.47	0.45
Thinness_ten_nineteen_years						1.00	0.94
Thinness_five_nine_years							1.00



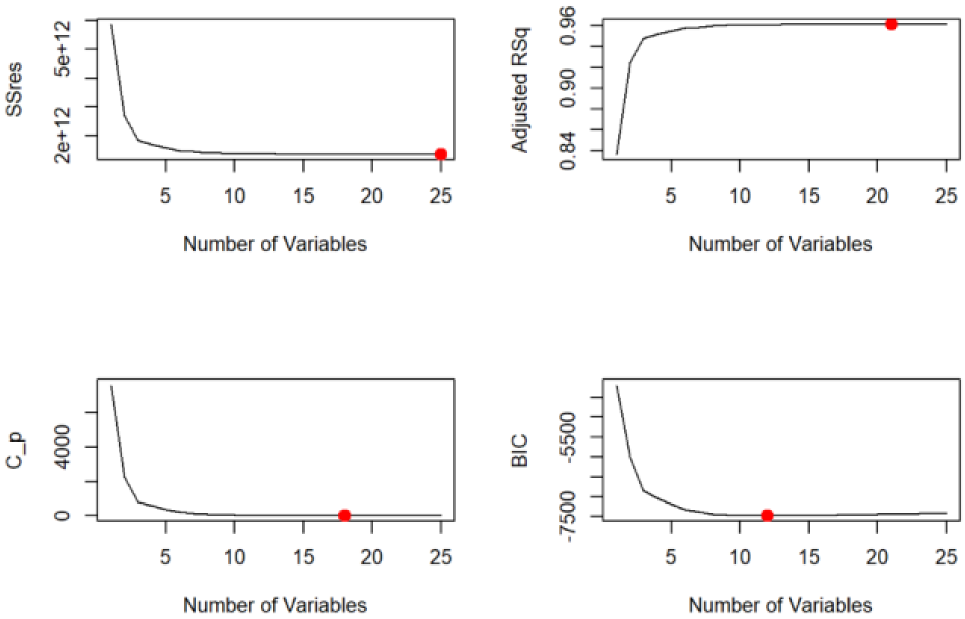
	Adult_mortality	Alcohol	Diphtheria	HepatitisB	Life_expectancy	Mean_BMI
Adult_mortality	1.00	-0.25	-0.53	-0.38	-0.94	-0.54
Alcohol		1.00	0.30	0.18	0.40	0.28
Diphtheria			1.00	0.79	0.64	0.47
HepatitisB				1.00	0.46	0.42
Life_expectancy					1.00	0.61
Mean_BMI						1.00

05

Feature Selection



Best subset selection



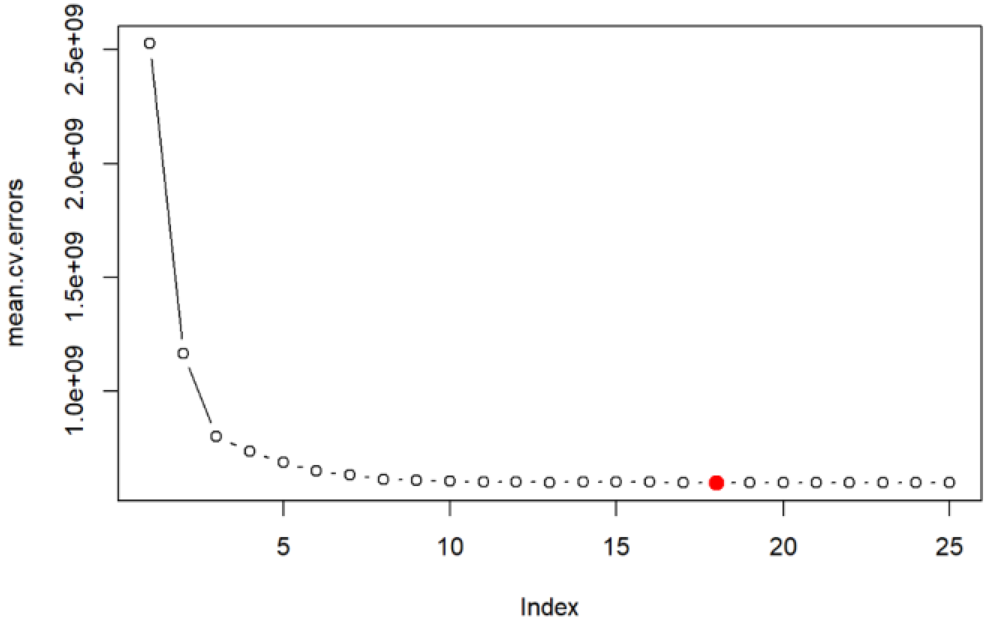
- p=25 which is less than 30 and n much grater than p → **BSS**

- From p=5 and after we have the same values approximately

- For p=5 and p=8 they share key variables such as Region, Infant_deaths, Economy-Status, Adult-Mortality and GDP

- With p=8 adds more dummy variables of Region

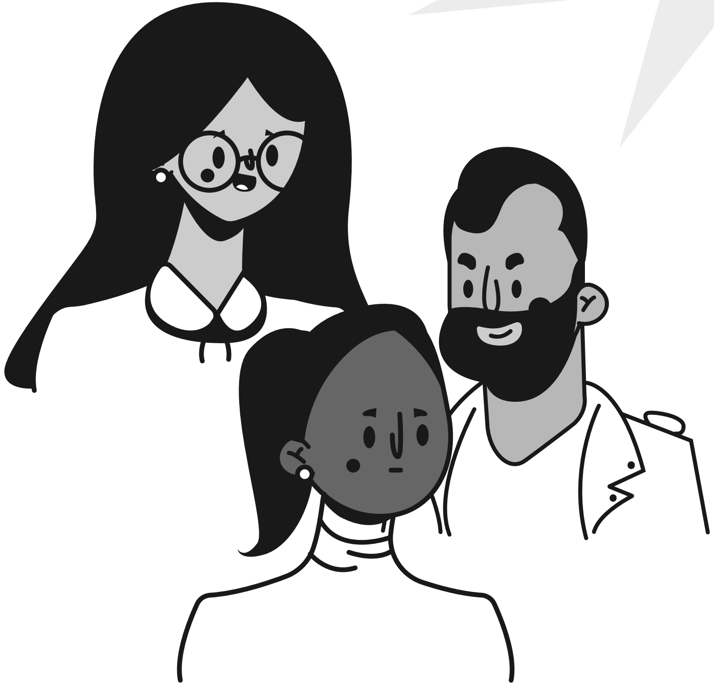
10 – Folds Cross Validation



- We use K=10
- Red dot indicates the lowest cross- validation error
- We apply One standard error rule, and we identify a model with 5 predictors as optimal
- The same 5 predictors as before

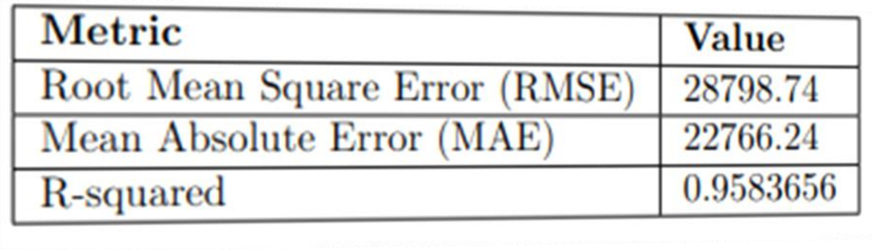
06

Models



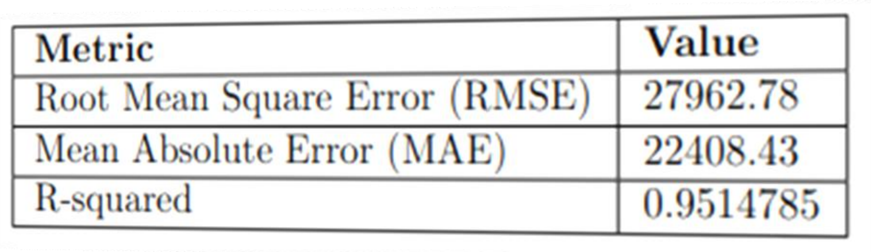
Random Forest Regression

Linear Regression



Metric	Value
Root Mean Square Error (RMSE)	28798.74
Mean Absolute Error (MAE)	22766.24
R-squared	0.9583656

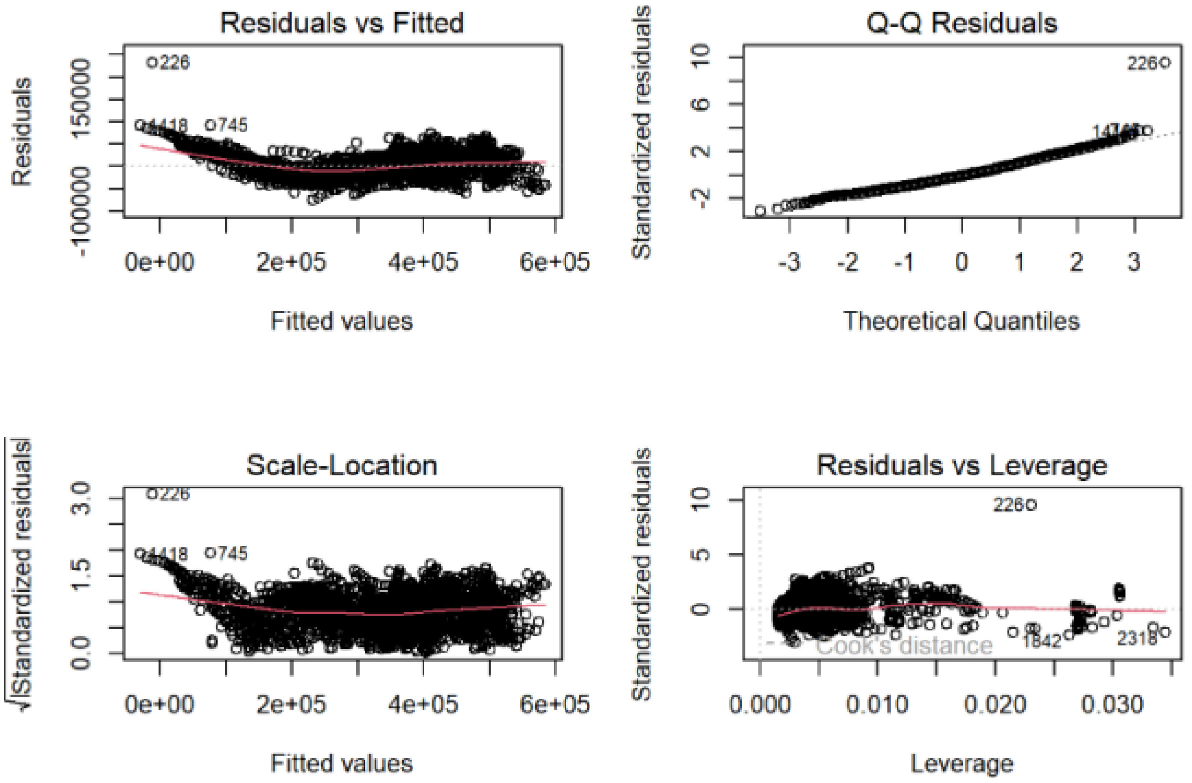
Random Forest Regression Metrics



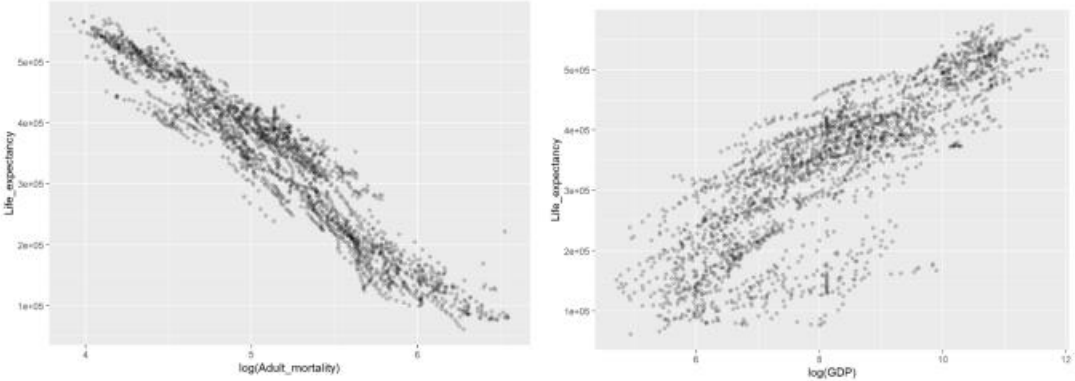
Metric	Value
Root Mean Square Error (RMSE)	27962.78
Mean Absolute Error (MAE)	22408.43
R-squared	0.9514785

Linear Model Regression Metrics

Linear Regression



Linear Regression

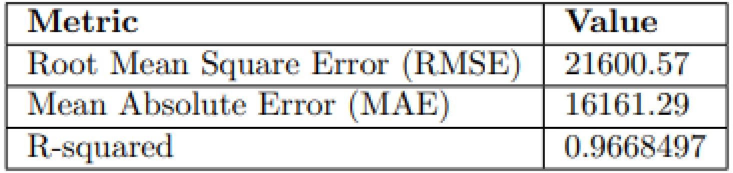


(a) $log(Adult\_Mortality)$

(b) $log(GDP)$

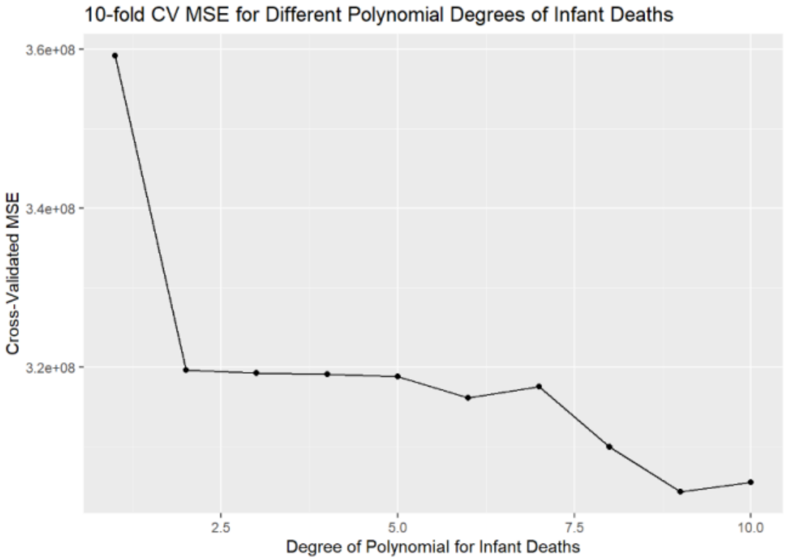
- When we log transform these two predictors, they have a linear relationship with our response

Linear Model Performance Metrics



Metric	**Value**
Root Mean Square Error (RMSE)	21600.57
Mean Absolute Error (MAE)	16161.29
R-squared	0.9668497

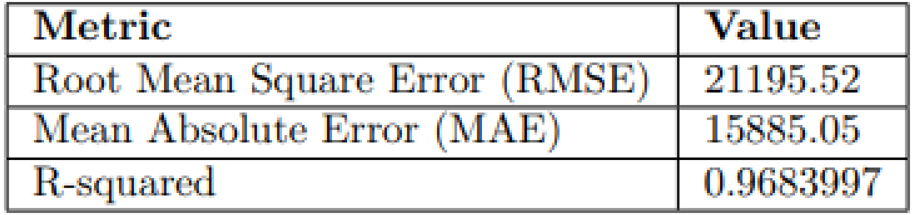
Linear Regression



- Checked the optimal complexity of *Infant deaths*

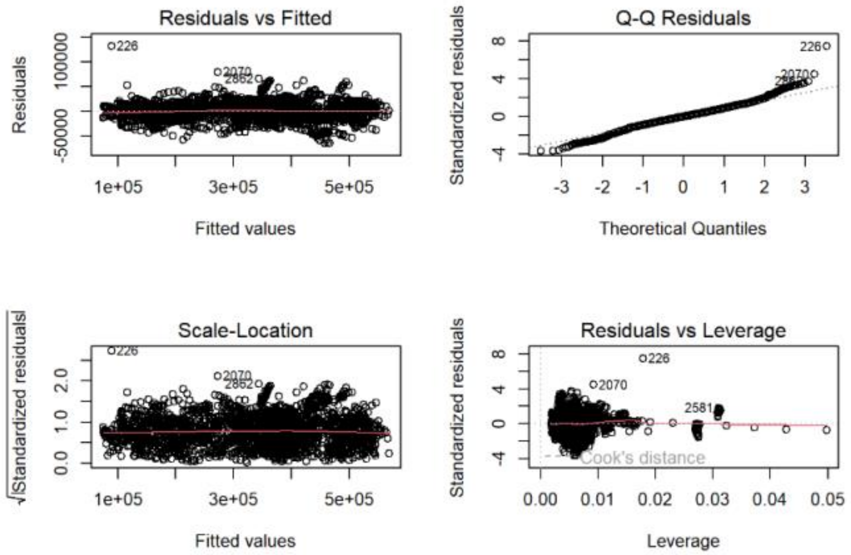
- *Infant deaths* polynomial degree 2

Performance Metrics of the Improved Linear Model



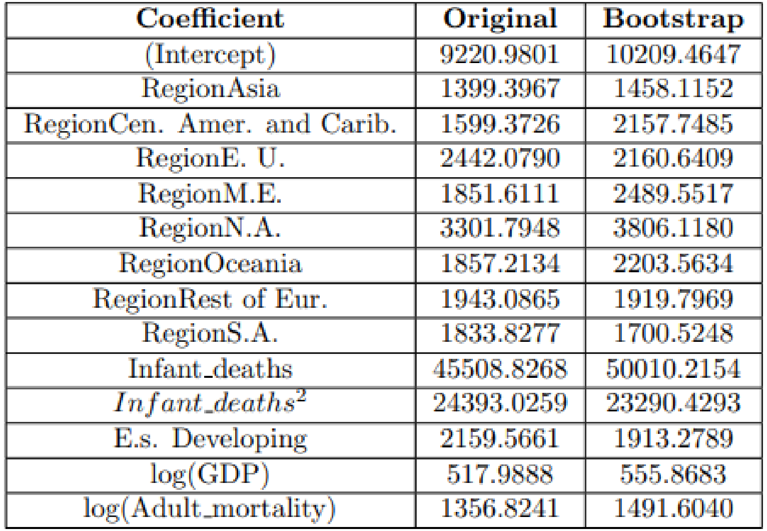
Metric	**Value**
Root Mean Square Error (RMSE)	21195.52
Mean Absolute Error (MAE)	15885.05
R-squared	0.9683997

Linear Regression



- Improvement on diagnostic plots
- Used the bootstrap approach to assess the variability of the coefficient estimates and the predictions

Standard errors



Coefficient	**Original**	**Bootstrap**
(Intercept)	9220.9801	10209.4647
RegionAsia	1399.3967	1458.1152
RegionCen. Amer. and Carib.	1599.3726	2157.7485
RegionE. U.	2442.0790	2160.6409
RegionM.E.	1851.6111	2489.5517
RegionN.A.	3301.7948	3806.1180
RegionOceania	1857.2134	2203.5634
RegionRest of Eur.	1943.0865	1919.7969
RegionS.A.	1833.8277	1700.5248
Infant_deaths	45508.8268	50010.2154
$Infant\_deaths^2$	24393.0259	23290.4293
E.s. Developing	2159.5661	1913.2789
log(GDP)	517.9888	555.8683
log(Adult_mortality)	1356.8241	1491.6040

Linear Regression

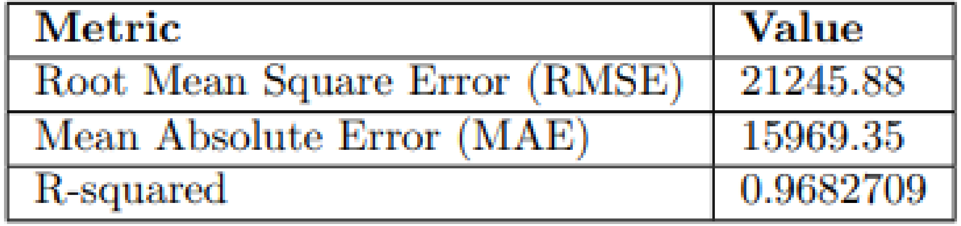
$$
\begin{aligned}
\widehat{Life\_expectancy}^{3} = & \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Asia} + \hat{\beta}_2 \cdot \text{Central America\&Caribbean} + \hat{\beta}_3 \cdot \text{European Union} \\
& + \hat{\beta}_4 \cdot \text{Middle East} + \hat{\beta}_5 \cdot \text{North America} + \hat{\beta}_6 \cdot \text{Oceania} \\
& + \hat{\beta}_7 \cdot \text{Rest of Europe} + \hat{\beta}_8 \cdot \text{South America} + \hat{\beta}_9 \cdot \text{Infant\_deaths} + \hat{\beta}_{10} \cdot \text{Infant\_deaths}^2 \\
& + \hat{\beta}_{11} \cdot \text{Economy\_status Developing} + \hat{\beta}_{12} \cdot log(\text{GDP}) + \hat{\beta}_{13} \cdot log(\text{Adult\_mortality}) \\
& + \hat{\beta}_{14} \cdot log(\text{GDP}) \cdot log(\text{Adult\_mortality})
\end{aligned}
$$

- Potential non-linearity in the relationship between a country’s GDP and its adult mortality rates

- Add an interaction term between them

- The scores were slightly worse

Performance Metrics



Metric	Value
Root Mean Square Error (RMSE)	21245.88
Mean Absolute Error (MAE)	15969.35
R-squared	0.9682709

Linear Regression – Best Model

$$
\begin{aligned}
\widehat{Life\_expectancy}^{3} = & \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Asia} + \hat{\beta}_2 \cdot \text{Central America\&Caribbean} + \hat{\beta}_3 \cdot \text{European Union} \\
& + \hat{\beta}_4 \cdot \text{Middle East} + \hat{\beta}_5 \cdot \text{North America} + \hat{\beta}_6 \cdot \text{Oceania} \\
& + \hat{\beta}_7 \cdot \text{Rest of Europe} + \hat{\beta}_8 \cdot \text{South America} + \hat{\beta}_9 \cdot \text{Infant\_deaths} + \hat{\beta}_{10} \cdot \text{Infant\_deaths}^2 \\
& + \hat{\beta}_{11} \cdot \text{Economy\_status Developing} + \hat{\beta}_{12} \cdot log(\text{GDP}) + \hat{\beta}_{13} \cdot log(\text{Adult\_mortality})
\end{aligned}
$$

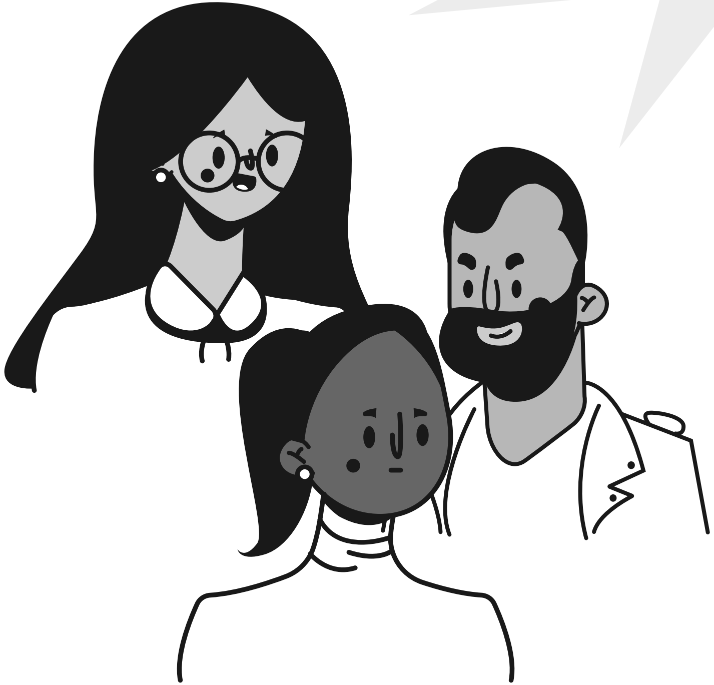
Metric	Value
RMSE	21195.52
MAE	15885.05
R^2	0.9683997

Prediction for Cyprus

- In 2016, adult mortality rate for Cyprus was 55 deaths per 1000 population (COVID 19)
- GDP per capita: 24600 USD
- Infant deaths: 2.3 deaths per 1000 population
- The 95% prediction interval is approximately [80.01 , 83.5]
- True value 81.06 of life expectancy.

07

Conclusion



Conclusion

- Our findings underscore the multifaceted nature of life expectancy, which is influenced by a complex interplay of economic, regional, and health-related factors

- Key predictors: region, infant deaths, economy status, GDP, and adult mortality

- A roadmap for policymakers and healthcare providers to target interventions effectively

- By focusing on the identified key predictors, countries can improve the health and well-being of their citizens

Thanks!

Do you have any questions?