BDA800

Predicting & Analyzing Life Expectancy

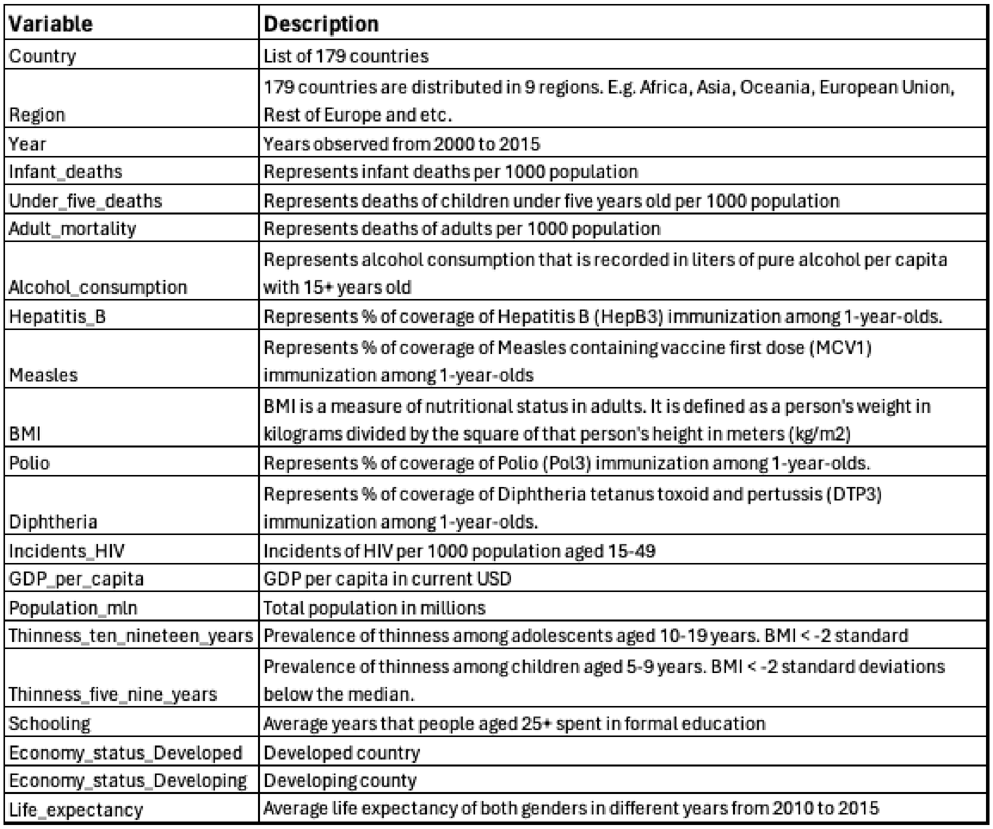
Introduction

Predicting and analyzing factors affecting life expectancy across globe is a critical endeavor and has implication for public health, policy formulation, and societal well-being. Primary audience for this study is researchers and academics in fields such as epidemiology, public health, demography, economics, and sociology who can contribute to the scientific understanding of life expectancy determinants and outcomes. Non-Governmental Organizations (WHO, UNICEF) can also be interested in leveraging the findings of the project to guide their programs and initiatives.

This project aims to analyze and forecast life expectancy using a comprehensive dataset sourced primarily from the World Health Organization (WHO). However, it's worth noting that the dataset utilized isn't directly sourced from WHO; rather, it's a "fixed" dataset that has undergone cleaning procedures, including the correction of null values and inaccuracies. This cleaning process was conducted by a Kaggle user using credible sources. The decision to utilize this "fixed" dataset was made to streamline the preprocessing phase and ensure a more robust foundation for subsequent analysis and forecasting endeavors related to life expectancy.

The dataset encompasses a wide array of variables spanning health indicators, immunization rates, economic factors, and demographic information for 179 countries from 2000 to 2015. The dataset has 21 variables and 2,864 rows.

Metadata available below:



Variable	Description
Country	List of 179 countries
Region	179 countries are distributed in 9 regions. E.g. Africa, Asia, Oceania, European Union, Rest of Europe and etc.
Year	Years observed from 2000 to 2015
Infant_deaths	Represents infant deaths per 1000 population
Under_five_deaths	Represents deaths of children under five years old per 1000 population
Adult_mortality	Represents deaths of adults per 1000 population
Alcohol_consumption	Represents alcohol consumption that is recorded in liters of pure alcohol per capita with 15+ years old
Hepatitis_B	Represents % of coverage of Hepatitis B (HepB3) immunization among 1-year-olds.
Measles	Represents % of coverage of Measles containing vaccine first dose (MCV1) immunization among 1-year-olds
BMI	BMI is a measure of nutritional status in adults. It is defined as a person's weight in kilograms divided by the square of that person's height in meters (kg/m2)
Polio	Represents % of coverage of Polio (Pol3) immunization among 1-year-olds.
Diphtheria	Represents % of coverage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization among 1-year-olds.
Incidents_HIV	Incidents of HIV per 1000 population aged 15-49
GDP_per_capita	GDP per capita in current USD
Population_mln	Total population in millions
Thinness_ten_nineteen_years	Prevalence of thinness among adolescents aged 10-19 years. BMI < -2 standard
Thinness_five_nine_years	Prevalence of thinness among children aged 5-9 years. BMI < -2 standard deviations below the median.
Schooling	Average years that people aged 25+ spent in formal education
Economy_status_Developed	Developed country
Economy_status_Developing	Developing county
Life_expectancy	Average life expectancy of both genders in different years from 2010 to 2015

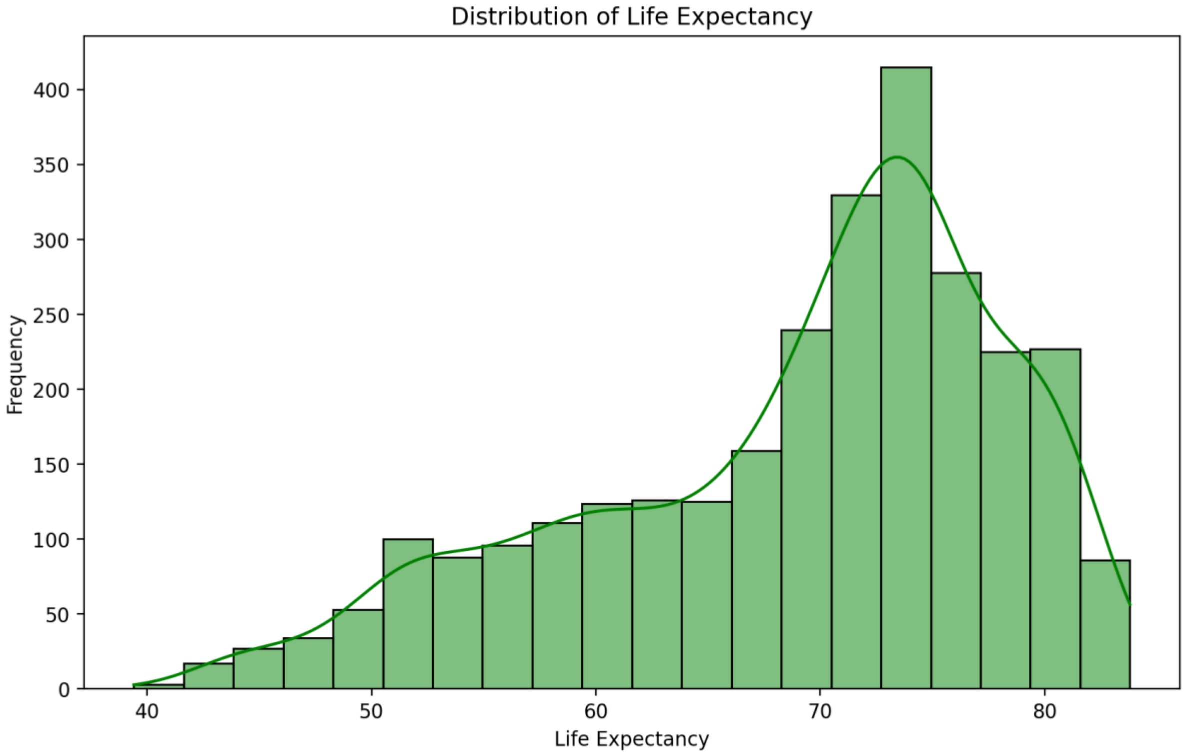
Project Objectives

As the dataset is widely available online and numerous predictive models have already been developed to forecast life expectancy, our focus for this project diverges from conventional analysis and predictive modeling. Instead, our objectives extend beyond model-building to encompass the following:

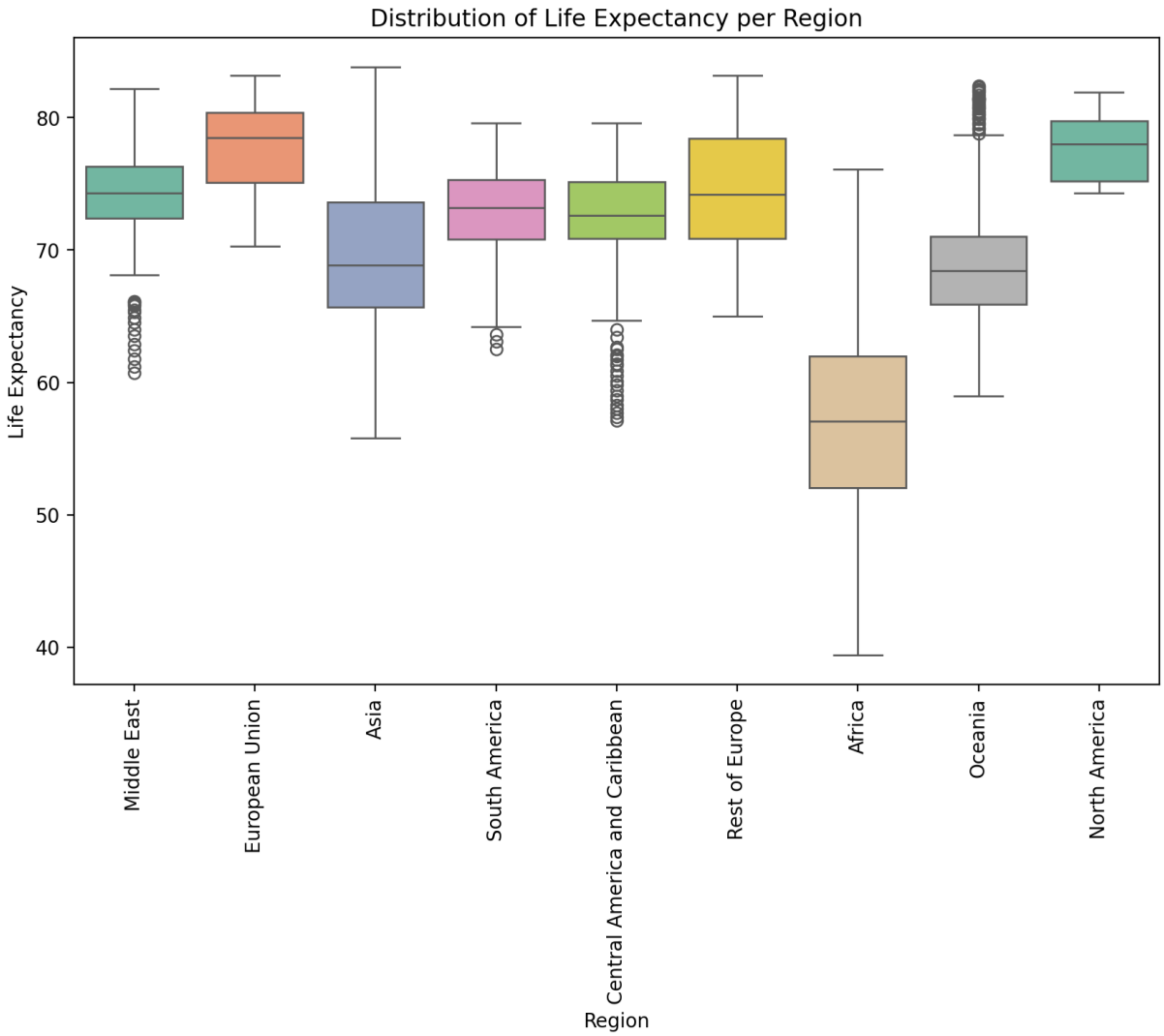
1. Conducting a thorough exploratory data analysis to uncover intriguing patterns or trends supported by existing research findings.
2. Introducing a new variable—population density—sourced from external data, to investigate its potential correlation with life expectancy. We aim to assess whether densely populated countries exhibit lower life expectancy, thereby accepting or rejecting this hypothesis.
3. Developing a user-friendly application to assess the best performing model. This application will enable users to predict life expectancy based on factors such as geographical region, adult mortality rates, educational attainment, among others.

Exploratory Data Analysis

As the dataset has been preprocessed beforehand, we will proceed directly to the exploratory data analysis phase.



Distribution of life expectancy over years for all countries. The graph shows that the most common life expectancy is around 74 years.

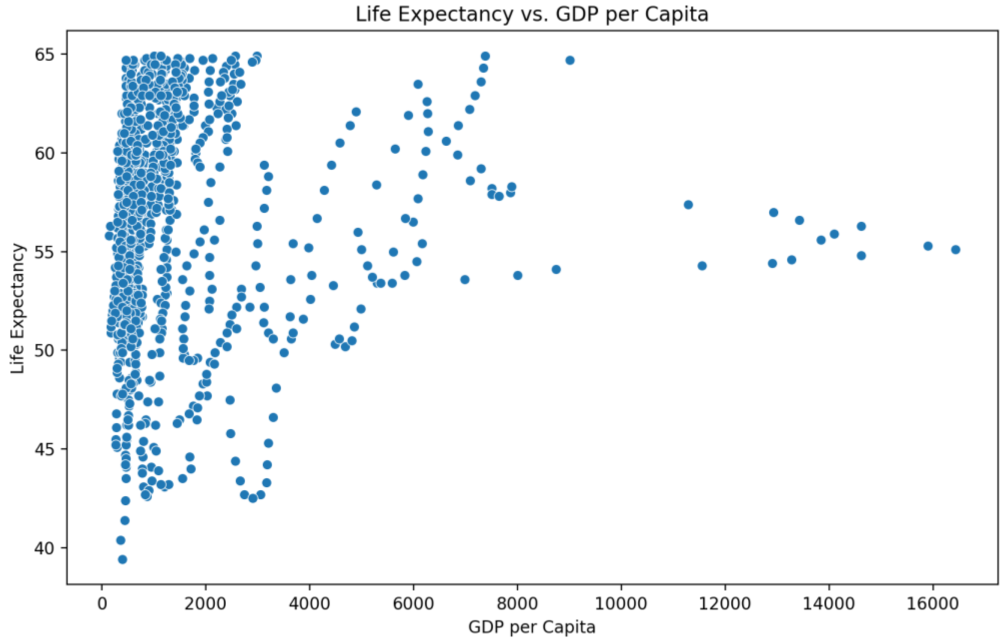


The box plots depicting the life expectancy distribution across regions.

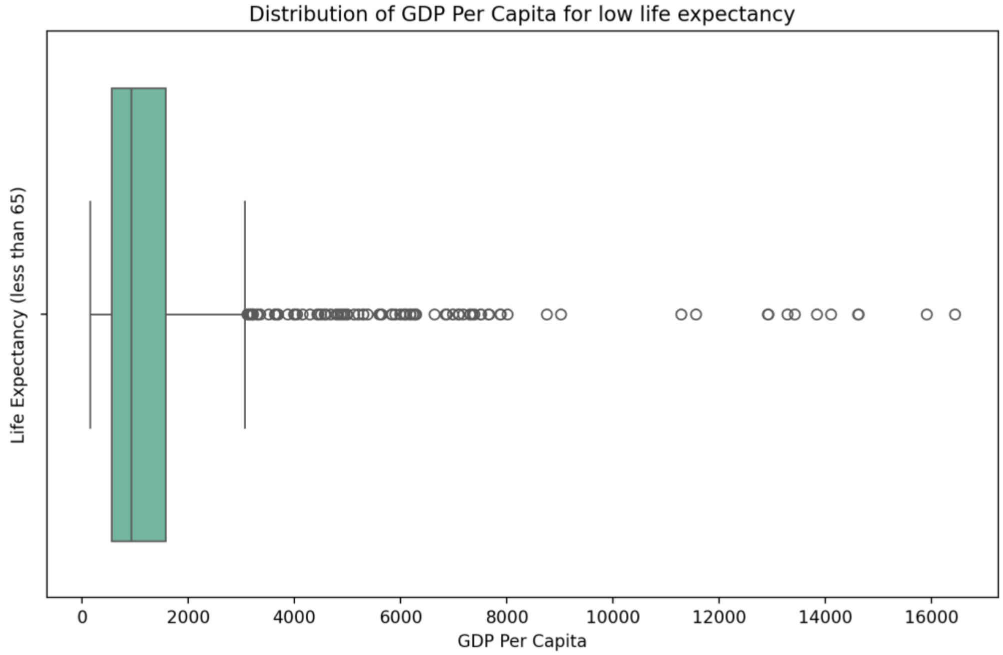
This graph illustrates that Africa possesses the lowest median life expectancy among all the regions shown. Additionally, the box plot for the African region exhibits a broad distribution of data. This indicates that the life expectancy in African countries varies widely, with the lowest reaching around 40 years and the highest extending up to approximately 75 years.

The box plots for the Middle East and Central America & Caribbean reveal a notable number of outliers, suggesting that there are several countries within these regions with life expectancies that are considerably lower than the regional median. These outliers point to significant variations within the regions, where some countries experience life expectancies that deviate from the common trends observed in these areas.

The box plots for the European Union and North America reveal that these regions possess the highest life expectancy when compared to all other regions analyzed. Moreover, the box plots corresponding to these two areas are notably narrow, indicating a smaller range of variation in life expectancy within each region.

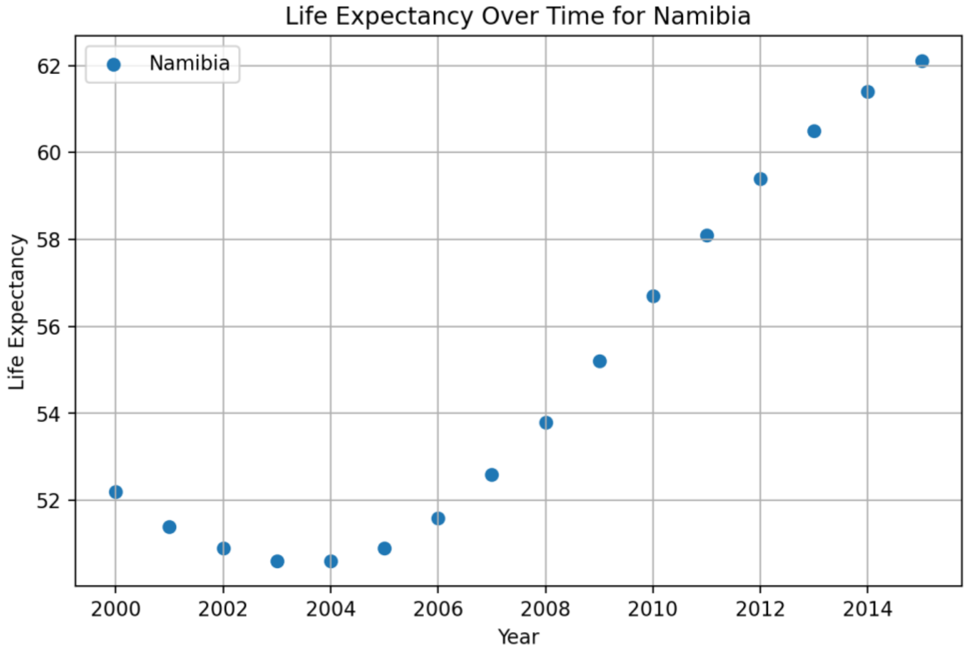


Relationship between GDPs per capita and life expectancy in nations where life expectancy is below 65 years.

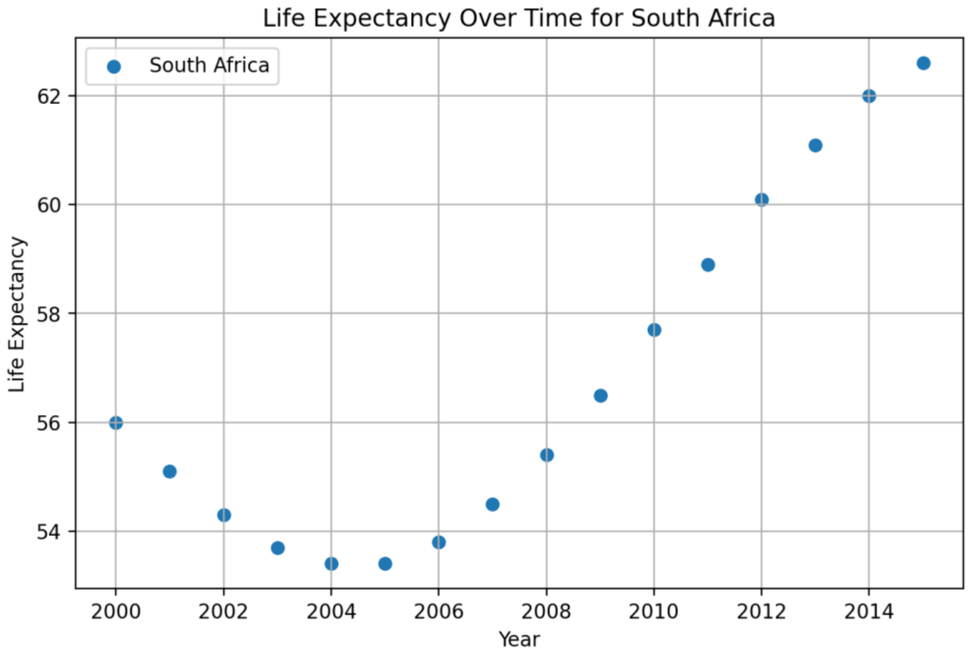


Box plot showing the GDP per capita distribution for countries with life expectancy below 65 years.

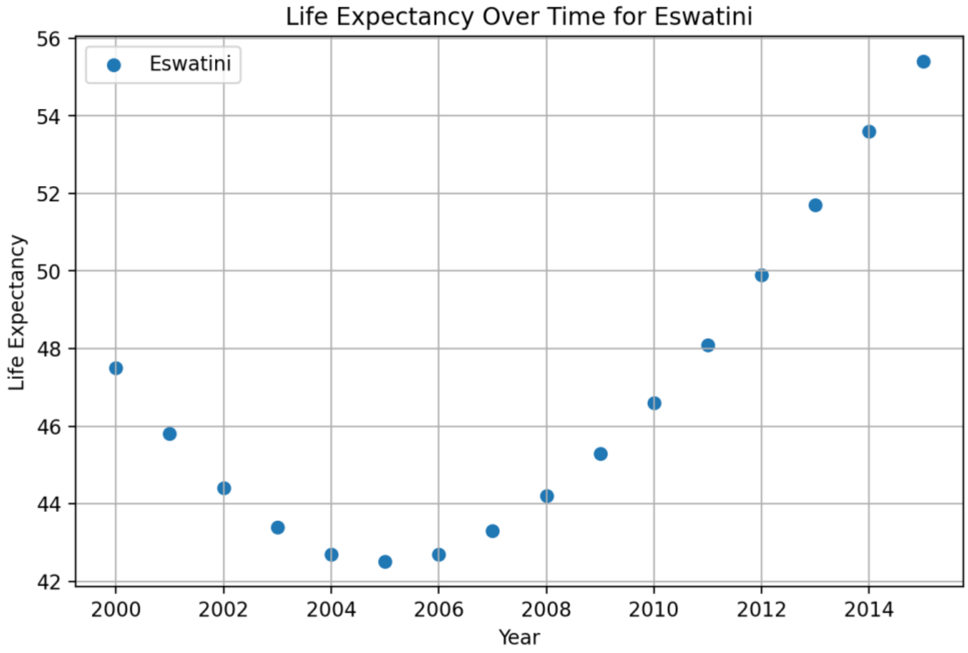
In this section of Exploratory Data Analysis, we aim to comprehensively examine the connection between GDPs per capita and life expectancy. Our focus is on analyzing the spread of per capita GDP among nations with a life expectancy under 65 years. The scatter plot reveals that the highest GDP per capita observed for these countries is around $16,000. Furthermore, the majority of GDP per capita values lie between $0 and $3,000, suggesting that nations with shorter life spans typically fall within this economic bracket. This finding is corroborated by the box plot of GDP per capita, which indicates that the upper quartile lies at or below $1,600. Additionally, the box plot features numerous outliers, signifying that a per capita GDP exceeding $3,000 is an anomaly for countries within this low life expectancy category.



Life Expectancy over time for Namibia (South Africa Region).

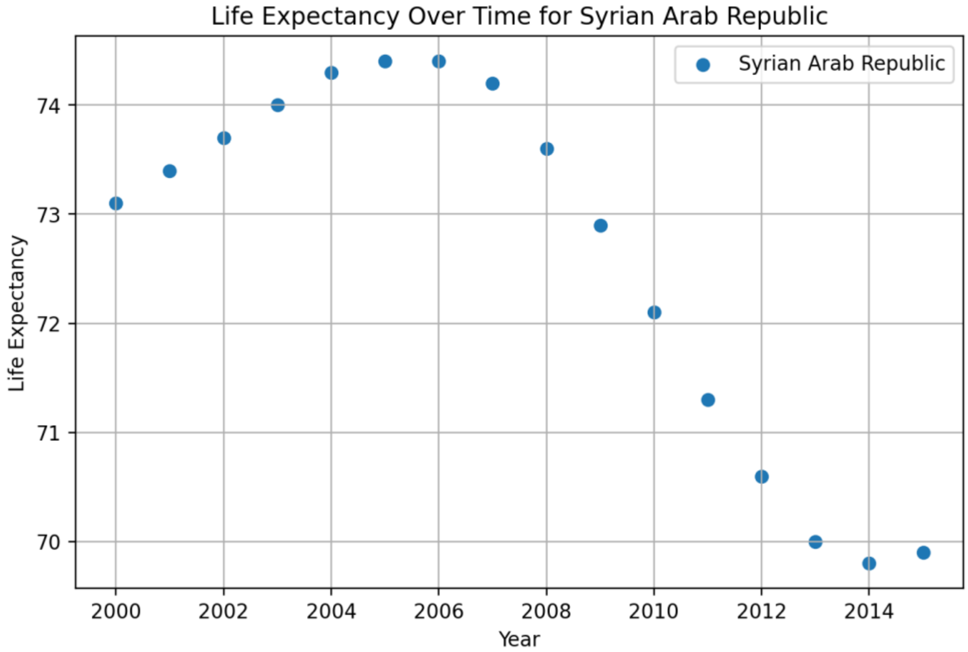


Life Expectancy over time for South Africa.



Life Expectancy over time for Eswatini (South Africa Region).

During our investigation into the trends of life expectancy over time in various countries, we encountered some notable patterns. The three graphs displayed depict the progression of life expectancy for Namibia, South Africa, and Eswatini. These neighboring countries, situated in the Southern Africa region, exhibit similar trends. We observed a decrease in life expectancy until the mid-2000s, followed by a subsequent rise. Our research revealed that the year 2006 marked the top of the HIV/AIDS crisis in Southern Africa, a period when mortality rates surpassed new HIV infections. The epidemic caused a drastic 20-year drop in life expectancy, with extremely high infant and maternal mortality. The initial response to the epidemic faced critique for being slow and ineffective. However, by the mid-2000s, South Africa began to intensify its efforts against HIV/AIDS, launching a comprehensive treatment program and initiating preventative strategies. This led to the strategic plan for HIV from 2007 to 2012, setting a promise for significant turnaround in the public health crisis.



Life Expectancy over time for Syria.

In our analysis, we uncovered a notable trend concerning Syria. The data reveals an increase in life expectancy up to the year 2006, followed by a sharp decrease afterwards. Further investigation into the events of 2006 unveiled some significant findings. Syria experienced its most severe drought in recorded history from 2006 to 2010, leading to the failure of 75 percent of the country's farms and the death of 85 percent of its livestock. This drought was the worst in the last 900 years. The emergence of conflict in 2011 intensified the crisis, significantly affecting the life expectancy in Syria.

Population Density

One of our objectives for the project was to see how density population correlates with life expectancy. This section is dedicated to explanation of the data preprocessing steps we completed to add population density to the main dataset.

The dataset, which includes information on the land area of each country measured in square miles, was obtained from the Kaggle platform and originally compiled by the U.S. Government's Central Intelligence Agency (CIA).

Example of the data is shown below.

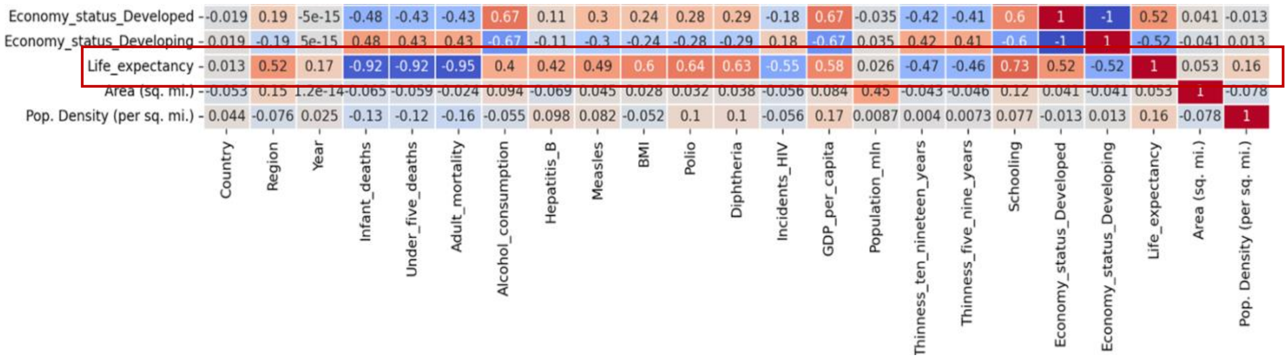
Country	Area (sq. mi.)
Afghanistan	647500
Albania	28748
Algeria	2381740
American Samoa	199
Andorra	468
Angola	1246700
Anguilla	102
Antigua & Barbuda	443
Argentina	2766890
Armenia	29800
Aruba	193
Australia	7686850

Since "Population" variable is already available in the main source of data, our objective is to introduce "Area" column and compute density by dividing population (in that specific year and) by area.

***Note**: an attempt to merge the two tables based on Country led to the occurrence of missing values since wordings of some countries are different in both datasets (e.g., Russia and Russian Republic). To overcome this issue, additional manual preprocessing steps to match names of the countries were required.*

Correlation Analysis

To determine which factors influence a country's life expectancy, we performed a correlation analysis to find the correlation coefficients linking life expectancy with other independent variables. We employed Pearson's method as our chosen technique for assessing these correlations.



	Country	Region	Year	Infant_deaths	Under_five_deaths	Adult_mortality	Alcohol_consumption	Hepatitis_B	Measles	BMI	Polio	Diphtheria	Incidents_HIV	GDP_per_capita	Population_mln	Thinness_ten_nineteen_years	Thinness_five_nine_years	Schooling	Economy_status_Developed	Economy_status_Developing	Life_expectancy	Area (sq. mi.)	Pop. Density (per sq. mi.)
Economy_status_Developed	-0.019	0.19	-5e-15	-0.48	-0.43	-0.43	0.67	0.11	0.3	0.24	0.28	0.29	-0.18	0.67	-0.035	-0.42	-0.41	0.6	1	-1	0.52	0.041	-0.013
Economy_status_Developing	0.019	-0.19	5e-15	0.48	0.43	0.43	-0.67	-0.11	-0.3	-0.24	-0.28	-0.29	0.18	-0.67	0.035	0.42	0.41	-0.6	-1	1	-0.52	-0.041	0.013
Life_expectancy	0.013	0.52	0.17	-0.92	-0.92	-0.95	0.4	0.42	0.49	0.6	0.64	0.63	-0.55	0.58	0.026	-0.47	-0.46	0.73	0.52	-0.52	1	0.053	0.16
Area (sq. mi.)	-0.053	0.15	1.2e-14	-0.065	-0.059	-0.024	0.094	-0.069	0.045	0.028	0.032	0.038	-0.056	0.084	0.45	-0.043	-0.046	0.12	0.041	-0.041	0.053	1	-0.078
Pop. Density (per sq. mi.)	0.044	-0.076	0.025	-0.13	-0.12	-0.16	-0.055	0.098	0.082	-0.052	0.1	0.1	-0.056	0.17	0.0087	0.004	0.0073	0.077	-0.013	0.013	0.16	-0.078	1

Correlation Matrix.

Variables having a coefficient more than the absolute value of 0.5 have strong correlation with life expectancy and will be chosen for future modelling.

List of significant attributes:

- Region
- Number of Infant Deaths per 1000
- Number of Under-Five Deaths per 1000
- Adult Mortality Rate (probability to die between 15 and 60 per 1000)
- BMI
- Polio Immunization Coverage among 1-years old
- Diphtheria Immunization Coverage among 1-years old
- Number of years in Schooling
- Incidents of HIVs among 15-49 ages
- GDP per capita

Predictive Modelling

Before we begin predictive modeling, it's essential to separate the dataset into training and testing subsets. For our project, we've allocated 30% of the data for testing purposes, while the remaining portion will be used for training.

To predict life expectancy (which is a continuous attribute) using significant variables listed above, we built 3 models: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.

In developing a predictive model, we encounter the challenge of hyperparameter tuning. To address this, we implemented GridSearchCV, a method designed to determine the optimal hyperparameter values. GridSearchCV systematically explores every potential combination of hyperparameters specified by the developer, constructs a model for each unique set, assesses their performance using Mean Squared Error (MSE), and then identifies the most effective parameter set according to the evaluation metric.

Best parameters for Decision Tree Regressor:
{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 5}

Best parameters for Random Forest Regressor:
{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}

Performance Evaluation

To evaluate the results of 3 models, R-squared and MSE metrics were chosen. Performance evaluation of each model is listed below.

Linear Regression:
R-squared: 97.80%
MSE: 1.895

Decision Tree Regressor:
R-squared: 98.92%
MSE: 0.931

Random Forest Regressor:
R-squared: 99.42%
MSE: 0.500

Observing the results, we see that they are quite impressive, with the Random Forest model achieving an R-squared value of 99%, indicating near-perfect accuracy. However, exceptionally high results can sometimes suggest that a model is overfitting, meaning it may not perform well on unseen data. To address this concern, we employed cross-validation technique, which confirmed the effectiveness of the models by giving similarly high scores, thereby rejecting the hypothesis of overfitting.

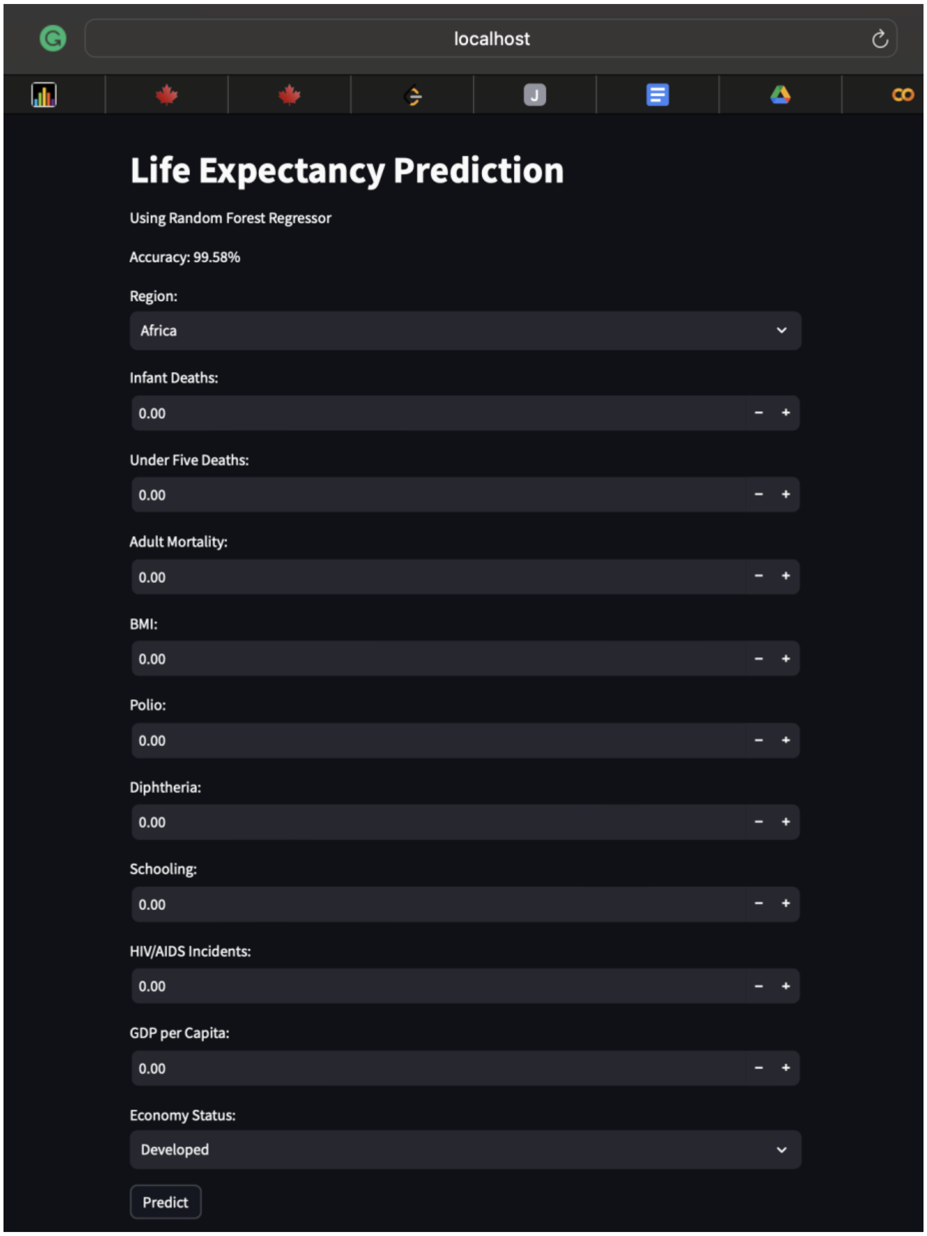
Results of cross-validation are listed below.

Linear Regression – MSE: 1.905
Decision Tree – MSE: 0.824
Random Forest – MSE: 0.357

Web Application

Finally, we are willing to introduce an accessible web application using streamlit library in Python designed to forecast life expectancy by utilizing the key factors we've previously pointed, including Region, BMI, Adult Mortality Rate, GDP per Capita, Economic Status, among others. This tool will be particularly useful for simulating the impact of variations in these attributes on life expectancy. For instance, the application could explore scenarios like the effect of a GDP per Capita increase from $40,000 to $60,000 on expected lifespan.

***Note**: code for application was sourced from Medium platform.*



Life Expectancy Prediction

Using Random Forest Regressor

Accuracy: 99.58%

Region: Africa

Infant Deaths: 0.00

Under Five Deaths: 0.00

Adult Mortality: 0.00

BMI: 0.00

Polio: 0.00

Diphtheria: 0.00

Schooling: 0.00

HIV/AIDS Incidents: 0.00

GDP per Capita: 0.00

Economy Status: Developed

Screenshot of a web application built. Code can be found in app.py.

Conclusion and Recommendations

This report has highlighted critical insights into factors influencing life expectancy across the globe, underscoring significant regional disparities and the impact of economic status. For our primary audience—researchers, academics, and NGOs like WHO and UNICEF—it's essential to consider these findings in their ongoing efforts to improve public health outcomes. We recommend that researchers further explore the correlation between life expectancy and the significant attributes identified, such as GDP per capita, Adult Mortality Rate, and health indicators. This deeper analysis can provide more targeted insights for policy formulation and health interventions.

For NGOs operating in regions with low life expectancy, it's vital to tailor health programs to address specific needs highlighted by our analysis, such as increasing immunization coverage and improving economic conditions. By utilizing our web application, stakeholders can simulate potential outcomes of various interventions, providing a practical tool for strategic planning and impact assessment. We encourage the use of this application to visualize the effects of changes in key attributes on life expectancy, aiding in more informed decision-making and policy development.

Reflection

Throughout this project on life expectancy, we learned a significant amount about the impact of various factors like health, economy, and education on life spans. One of the most interesting aspects was having to conduct external research to understand why life expectancy trends in certain countries would decline and then increase, or vice versa. This research allowed us to confirm the trends shown in our graphs with actual historical events and factors, providing a deeper understanding of the data. We particularly enjoyed using the web application to predict changes in life expectancy, which made our analysis feel practical and relevant. The project has increased our interest in public health, and we are excited to possibly continue exploring how specific policies can improve health outcomes beyond this course.

References

[1] Lasso, F. (n.d.). Countries of the World. Kaggle. https://www.kaggle.com/datasets/fernandol/countries-of-the-world/data?select=countries+of+the+world.csv

[2] Lasha (n.d.). Life Expectancy (WHO) Fixed. Kaggle. https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated

[3] Santiago, A. (2023, July 29). Life Expectancy Analysis (Regression Model). Medium. https://medium.com/@lexshie/life-expectancy-analysis-regression-model-4c696bb9a3f6

[4] Britannica, T. Editors of Encyclopaedia (2024, March 29). Syrian Civil War. Encyclopedia Britannica. https://www.britannica.com/event/Syrian-Civil-War

[5] Abdool Karim, S. S., Churchyard, G. J., Karim, Q. A., & Lawn, S. D. (2009). HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. Lancet (London, England), 374(9693), 921–933. https://doi.org/10.1016/S0140-6736(09)60916-8

[1] Source of “Area” attribute for all countries.

[2] Main source of the data used for the project. “Fixed” dataset.

[3] Medium article where we sourced code for an application.

[4] Article explaining the main events happening in Syria in 2000s.

[5] Article explaining the HIV/AIDS epidemic in South Africa.