# Predictive model using comparative analysis of correlated effects socio-economic factors play in relation to life expectancy

Task 2

by

Evan Boyd

Bachelors of Applied Science : Data Management & Analytics Capstone Task 2

Submitted to School of Technology Department

Western Governors University

July 2024

# A. Proposal Overview

## Research Question

How does corruption affect life expectancy?

## Context

Understanding the factors that contribute to a long and healthy life can yield invaluable insights and guide policy decisions effectively. For instance, if research reveals that communicable diseases cause a significant detriment to life expectancy compared to nutrition, then public health efforts can be reorganized to prioritize disease prevention and treatment. Similarly, if a strong correlation is found between high CO2 emissions and shorter life spans, this could prompt societal and governmental action to reduce carbon output, potentially leading to both improved public health and environmental benefits. Furthermore, uncovering a link between corruption levels and life expectancy might incentivize leaders to combat corruption more vigorously, recognizing that reducing corruption could directly enhance the standard of living and overall well-being of their populations. From formulating targeted interventions, to mobilizing resources and public support towards the most impacted areas, and even creating a life

expectancy calculator for leisure use, these insights will ultimately foster healthier and longer lives for communities worldwide.

## Summary of Compounding Works

The fallout of rampant corruption can ruin a country in more ways than one. Dreher and Herzfeld tell us in *The Economic Costs of Corruption: A Survey and New Evidence* many empirical approaches have already born evidence indicating corruption lowers foreign investments, reduces public expenditure, and the negative effects on GDP are substantial. Life expectancy drops by a substantial 2.5 years for every 1 index point increase in corruption. As the main determinants for corruption are also affected by corruption, Dreher and Herzfeld imply that without efforts to reduce corruption it will simply compound upon itself rippling negative economic effects throughout the country. Their conclusion aligns with my research, and I wholeheartedly agree with the sentiment of corruption being the single greatest obstacle to economic and social development.

While Dreher and Herzfeld reviewed World Bank data of many developing nations, the University of Nigeria took a more localized approach to studying life expectancy in Nigeria. *Climate change and Life Expectancy in a Developing Country: Evidence from Greenhouse Gas (CO2 ) Emission in Nigeria 2018* explores the possibility of CO2 emissions affecting the average Nigerian's life expectancy. Their analysis determined CO2 is not a threat to longevity in Nigeria, and even found evidence suggesting CO2 emissions increase life expectancy. Although the findings were not statistically

significant enough to draw conclusions, the results highlight the importance of my own research to assist in drawing nearer to an answer on CO2 and life expectancy.

Uchendu & Abolarin link corruption to food security and life expectancy using analytical techniques to determine relationships between CPI, FSI, LE, and population in LCC and MCC. They found that countries with low longevity were associated with low food security and more corruption. Suggesting policies discouraging corrupt practices and promoting good governance could be enacted to eradicate malnutrition in developing countries. If eradicating malnutrition can be expected to raise life expectancy, I expect my research to find similar conclusions for CO2 or corruption.

## Summary of Data Analytics Solution

Employing Python and its associated libraries (Matplotlib, Pandas, etc.), this study seeks to graphically represent the correlation between life expectancy and various socio-economic factors, with an emphasis on corruption, through the use of heatmaps, bar charts, and violin plots. The data was sourced through the World Bank and includes C02, health, and other socio-economic data that will prove useful in building a model for life expectancy. Visualizing my findings through heatmaps, bar charts, and violin plots will effectively communicate the insights derived from the data. Armed with these discoveries, I will clean and preprocess the data before constructing a model capable of accurately predicting life expectancy across various socio-economic conditions. The finalized model will be saved as a pickle file and shared with stakeholders for

implementation as deemed appropriate. This entire process will adhere to a waterfall methodology, ensuring each step is completed before proceeding to the next in a linear sequence.

### Benefits & Support of Decision-Making Process

The insights and economic model derived through this project will provide citizens and leaders alike a clear and accurate view of the effects corruption can have on populations & communities. The model could also be used to make a life expectancy calculator for people from diverse socio-economic backgrounds.

# B. Data Analytics Project Plan

### Goals, Objectives, & Deliverables

- Create a usable model that accurately predicts life expectancy based on various socio-economic factors.
    - Acquire data from source
        - Save acquired data in an excel file
    - Create visualization charts that
        - Save acquired data in an excel file
    - Clean and Preprocess data

        - Save preprocessed data in a dataframe
    - Build a model that accurately predicts life expectancy based on socio-economic factors
        - Save the model in a pickle file

## Scope of Project

This project will encapsulate data collection from World bank, data cleaning/preprocessing, exploratory data analysis, feature selection, Model development, model evaluation, and documentation. This project will focus on identifying correlations and building a predictive model, not establishing causal relationships. This project will not involve collecting any additional data. This project will not focus on deployment or integration into a production environment at this time.

## Standard Methodology

Utilizing the waterfall approach I planned to work through my project linearly, completing each step before the next. Requirements gathering encompasses collecting my data and defining my project scope. Working data for this project is available publicly from the World Bank, and the scope is defined above. Analysis & Design will

cover the selection of my models, libraries, charts, and processing steps. The specific details of my analysis & design will be included in a .ipynb file linked here. Implementation will begin with cleaning the data and creating visualizations, then end with the development and training of the model. Verification will be conducted using evaluation metrics to provide a z-score, p-value, and other metrics. Deployment and Maintenance will be foregone until the model is used in a production environment.

## Timeline & Milestones

- Explore relationship between features and life expectancy 14 days

  06/12/24 - 06/26/24

  - Create Visualizations 7 days

    06/12/24 - 06/19/24

  - Explain observed patterns 7 days

    06/19/24 - 06/26/24

- Build a model that accurately predicts life expectancy based on socio-economic factors 21 days

  06/26/24 - 07/17/24

  - Train machine learning model capable of predicting life expectancy 7 days

    06/26/24 - 07/03/24

- A comprehensive report summarizing the findings, including insights from the visualizations and model evaluations saved in a .ipynb file 21 days

  06/26/24 - 07/17/24

### Resources & Costs

1. Google Colab $19.98($9.99/month)
2. 60 Total hours(12 hrs/week for 5 weeks)
3. Desktop $1200

### Criteria for Success

A complete analysis with visualizations, a working model, and detailed documentation are the only criteria for success of this project.

# C. Design of Data Analytics Solution

### Hypothesis

The more corrupt the country, the shorter life expectancy they'll have.

## Analytical Method

This project will implement a predictive analytics solution. I'll implement random forest trees, linear regression, or stochastic gradient descent to build a model from the data and predict life expectancy. This method is ideal for using data to answer questions. The use of descriptive analytics will serve to extrapolate what is currently happening in the data and then provide some visualization highlighting those happenings. Both of these analytical methods are necessary to serve the end goal of answering how corruption affects life expectancy.

## Tools & Environment

I will utilize python and multiple libraries inside a notebook environment saved as a .ipynb file.

## Methods & Metrics to Evaluate Statistical Significance

I will use a Z-test to provide a Z-statistic and p-value, as well as use an R-squared value to help communicate the statistical significance of my findings. The z-test will compare the mean predicted value from the model and compare it to known test data. This will tell us the accuracy of the model, with a positive score being above the mean and negative below. The p-value will be used to assess the strength of evidence against a null hypothesis. A p-value above 0.05 suggests weak evidence, while below 0.05 suggests strong evidence against the null hypothesis. The R-Squared value will represent from 0 to 1 how well my model predicts my test data.

## Practical Significance

The practical significance of the project can be measured in the insights disclosed throughout the analysis of the data and evaluation of the model. Serving people and policy makers alike, the information included in this project will arm them in their decision making processes. To illustrate this picture a leader with limited resources trying to tackle an issue and maximize the longevity of their people. Great intentions can still bestow barren results, so allocating resources to the *correct* issue is important when working with limited resources. Guided by the information in this project, leaders could ascertain if spending more money on healthcare or to reduce corruption will have a greater impact on their countries life expectancy. From another perspective, a young person living in a country with a low life expectancy may use the information to influence their decision on where to relocate.

## Visual Communication

I will incorporate bar charts, heatmaps, and tabular data to visualize my data to assist with the analysis. All this will be performed using python and its associated libraries.

- Figure 1 will display null values in each column, highlighting areas where more accurate reporting is needed and take an honest look at the deficiencies in our data.
- Figure 2 will show a correlative heatmap measuring the pairwise correlation relationship between all features scaled between 1 to -1.

- Figure 3 will likely display the same information as Figure 2, but in a 2-way bar chart isolating the features relationship to life expectancy.
- Figure 4 will strive to visualize the average level of corruption by region.
- Figure 5 will feature a violin chart to show the diversity and range of life spans in each region

# D. Description of Dataset

### Source of Data

All data was sourced from the World Bank, then compiled and hosted on kaggle where I downloaded the data in csv format.

### Appropriateness of Dataset

This dataset is one of the most complete sets of annual socio-economic data that's publicly available. The World Bank is the premiere organization on gathering macro data worldwide to explore data driven solutions. With our goal of creating an accurate model to predict life expectancy, having accurate data from a reputable organization guarantees the integrity and reliability of our findings, ultimately enhancing the model's prediction.

## Data Collection Methods

The data is publicly available via the world bank and was compiled then posted to kaggle. I took steps to validate the authenticity of the data, then downloaded from kaggle.

## Observations on Quality & Completeness of Data

The data is very accurate, but missing large amounts of corruption data from Sub Saharan Africa, the Middle East, and North Africa

## Data Governance, Privacy, Security, Ethical, Legal, & Regulatory Compliances

As this data is publicly available I do not have to worry about any violations, legal or regulatory. Ethically speaking I need to be aware of how some may interpret the results of my model.

## Project Sources

*Amuka, J. I., Asogwa, F., Omeje, A., Onyechi, T., & Ugwuanyi, R. (2018). Climate change and Life Expectancy in a Developing Country: Evidence from Greenhouse Gas (CO2 ) Emission in Nigeria . International Journal of Economics and Financial Issues, 8(4).*

*Chavan, S. S. (2023a, September 5). Life expectancy & socio-economic (world bank). Kaggle. https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank*

*Dreher, Axel and Herzfeld, Thomas, (June 2005). The Economic Costs of Corruption: A Survey and New Evidence. Available at SSRN: https://ssrn.com/abstract=734184 or http://dx.doi.org/10.2139/ssrn.734184*

*Indicators. World Bank Open Data. (2020). https://data.worldbank.org/indicator*

*Uchendu, F. N., & Abolarin, T. O. (2015). Corrupt practices negatively influenced food security and live expectancy in developing countries. The Pan African medical journal, 20, 110. https://doi.org/10.11604/pamj.2015.20.110.5311*

## Glossary

FSI – Food Security Index

CPI - Corruption Perceptions Index

MCC – Most Corrupt Country

LCC – Least Corrupt Country

LE – Life Expectancy