Project 2: Life Expectancy Prediction

Kesav Adithya Venkidusamy

Bellevue University - Master of Science in Data Science

DSC680-T301 Applied Data Science (2231-1)

 Professor Catherine Williams

10/16/2022

Table of Contents

[**Business Problem** 3](#_Toc116717158)

[**Background/History** 3](#_Toc116717159)

[**Data Explanation** 3](#_Toc116717160)

[**Data Dictionary** 4](#_Toc116717161)

[**Data Preparation** 5](#_Toc116717162)

[**Data Visualization** 6](#_Toc116717163)

[**Methods** 12](#_Toc116717164)

[**Analysis** 12](#_Toc116717165)

[**Modeling Analysis** 12](#_Toc116717166)

[**Feature Analysis** 14](#_Toc116717167)

[**Conclusion** 16](#_Toc116717168)

[**Assumptions** 17](#_Toc116717169)

[**Limitations** 17](#_Toc116717170)

[**Challenges** 17](#_Toc116717171)

[**Future Uses/Additional Applications** 17](#_Toc116717172)

[**Recommendations** 17](#_Toc116717173)

[**Implementation Plan** 17](#_Toc116717174)

[**Ethical Assessment** 18](#_Toc116717175)

[**References** 18](#_Toc116717176)

## **Business Problem**

Everything has an expiration date; humans are no exception either. The term “life expectancy” refers to the number of years a person can expect to live. By definition, life expectancy is based on an estimate of the average age that members of a particular population group will be when they die. In this project, I aim to explore the parameters affecting the life span of individuals living in distinct countries and learn how the life span can be estimated with the help of machine learning models. I will also focus on exploring the parameters that greatly impact the lifespan of an individual

## **Background/History**

We’re in an unprecedented era where humans are living longer with increased access to modern science and healthcare. It’s no secret, though, that life expectancy varies widely across the globe. Life expectancy depends on several factors, the two most important being gender and birth year. Generally, females have a slightly higher life expectancy than males due to biological differences. Other factors that influence life expectancy include:

* Race and ethnicity
* Family medical history
* Risky lifestyles

## **Data Explanation**

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors of all countries. The datasets are made available to the public for health data analysis.

<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>

The data set related to life expectancy, and health factors for 193 countries have been collected from the same WHO data repository website, and its corresponding economic data was collected from the United Nation website.  Among all categories of health-related factors, only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in the health sector resulting in an improvement in human mortality rates, especially in developing nations in comparison to the past 30 years. Therefore, in this project, we have considered data from the years 2000-2015 for 193 countries for analysis. The data was collected from WHO and United Nations websites with the help of Deeksha Russell and Duan Wang.

### **Data Dictionary**

**Characteristics**

|  |  |
| --- | --- |
| **Data Set Characteristics** | Multivariate |
| **Attribute Characteristics** | Categorical, Integer |
| **Associated Tasks** | Classification |
| **Number of Instances** | 2938 |
| **Number of Attributes** | 22 |
| **Missing Values** | Yes |
| **Area** | Health |

**Attributes information**

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Feature Type** |
| Country | Country Observed | Discrete |
| Year | Year Observed | Continuous |
| Status | Status of the country; Developed or Developing Status | Discrete |
| Life expectancy | Life expectancy in age | Target |
| Adult Mortality | Adult Mortality Rates on both sexes (probability of dying between 15-60 years/1000 population). | Continuous |
| Infant deaths | Number of Infant Deaths per 1000 population | Continuous |
| Alcohol | Alcohol recorded per capita (15+) consumption (in liters of pure alcohol). | Continuous |
| Percentage expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%). | Continuous |
| Hepatitis B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) | Continuous |
| Measles | Number of reported Measles cases per 1000 population | Continuous |
| BMI | Average Body Mass Index of the entire population | Continuous |
| Under-five-deaths | Number of under-five deaths per 1000 population | Continuous |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) | Continuous |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) | Continuous |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) | Continuous |
| HIV/AIDS | Deaths per 1 000 live births of HIV/AIDS (0-4 years) | Continuous |
| GDP | Gross Domestic Product per capita (in USD) | Continuous |
| Population | The population of the country | Continuous |
| thinness 1-19 years | Prevalence of thinness among children and adolescents for Age 10 to 19 (%) | Continuous |
| thinness 5-9 years | Prevalence of thinness among children for Age 5 to 9(%) | Continuous |
| Income composition of resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) | Continuous |
| Schooling | Number of years of Schooling(years) | Continuous |

### **Data Preparation**

The problem statement of this project is to identify the dataset feature(s) which are mostly related to or affecting the life expectancy of a person. The data set contains approximately 3000 entries. Given the limited size of the data set, the model should only be expected to provide a modest improvement in the identification of attrition vs a random allocation of the probability of attrition. The dataset consists of 22 features of which 20 are numerical and two are categorical with “life expectancy” being the target.

The target variable “life expectancy” is a continuous variable with values ranging between 36.3 years and 89 years with a maximum number of values lying at 72.1 years. The mean for this variable lies at 69.22 years. The details are shown in figure 2.

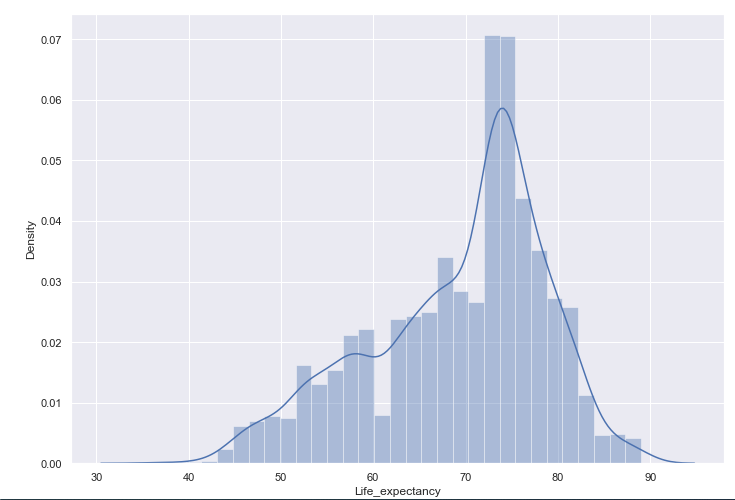


Figure . Target Variable “Life Expectancy” Analysis

As part of the EDA process, a null check has been performed initially on all the variables and found a null value present for the below features. The null value for all these variables is replaced with the forwarding fill method.

* Hepatitis B
* Alcohol
* Adult Mortality
* Polio
* Total Expenditure
* GDP
* Population
* Schooling
* Income Composition of resources
* Life Expectancy
* Diphtheria
* Thinness 5 to 9 years
* Thinness 1 to 19 years
* BMI

Then, a duplicate check has been performed, and found no duplicate in the dataset. None of the features has been removed from the dataset as I believe each feature correlates with the target variable “life expectancy”.

### **Data Visualization**

As mentioned before, the dataset contains 20 numerical features and 2 categorical features, and ‘life expectancy’ is the target variable. Below are some of the charts plotted to analyze these variables.

#### Histogram

The histogram chart is used to identify the distribution of numerical features present in the dataset. All the numerical features are considered for the histogram as shown in figure 2. The observation of each histogram is shown in figure 3.

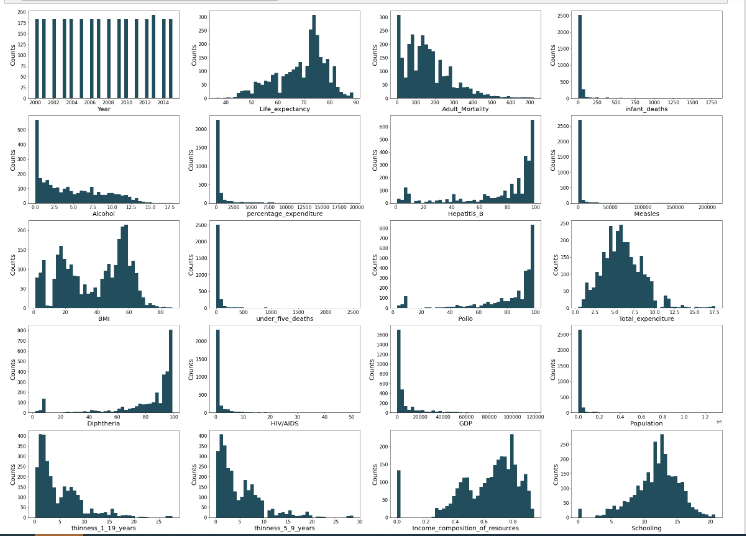


Figure Histogram for numerical variables

**Observation**

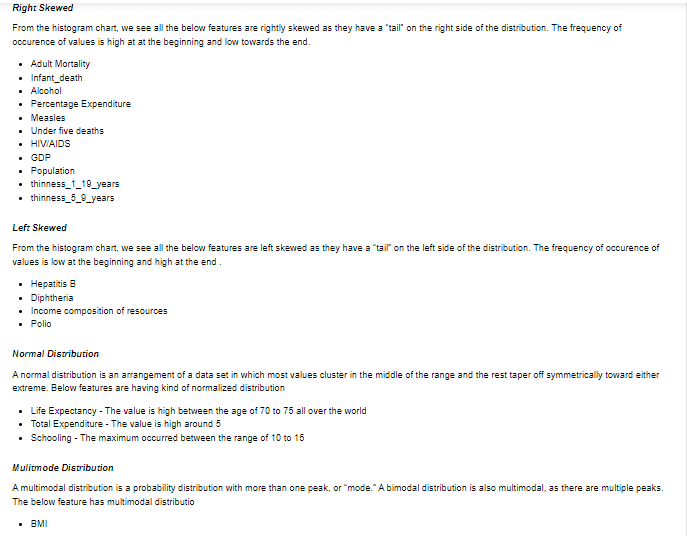


Figure : Observations from histogram charts

#### Violin Graph

A violin graph has been plotted for the categorical feature ‘Status’ and our target variable ‘Life Expectance’ to find the correlation between them. The details are shown in figure 4. From the graph, I could see developing countries have a low life expectancy and developed countries have high life expectancy all over the world.

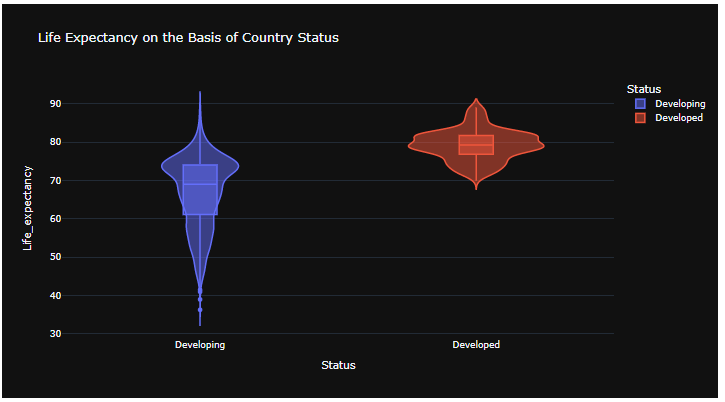


Figure Violin chart for Categorical Variable Status

#### Map Chart

A map chart as shown in figure 5 has been plotted between the categorical variable “country” and the target variable “life expectancy” to show the density of life expectancy across countries. Looking at the chart, we see that life expectancy is high for developed countries compared to developing countries. In addition, we could also notice that life expectancy increases over the years across the countries.

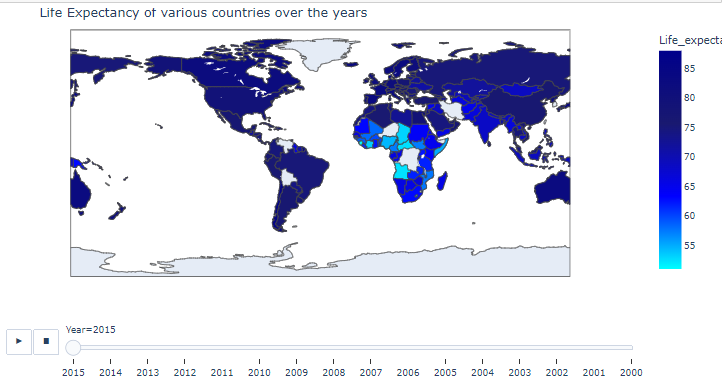


Figure : Map Chart showing the density of life expectancy

#### Heat Map

A heat map has been created to understand the correlation between various features present in the dataset. The details are shown in figure 6. The observations of the heat map are shown in figure 7.

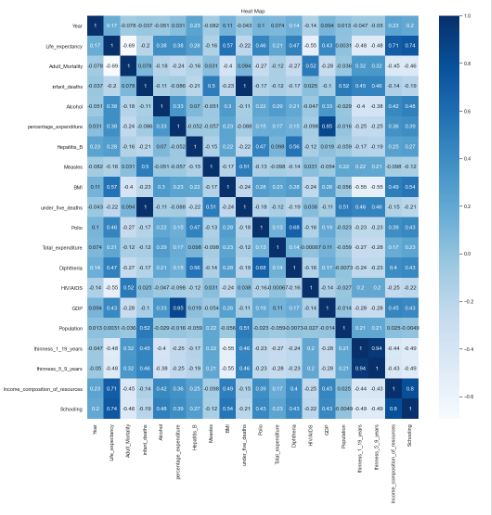
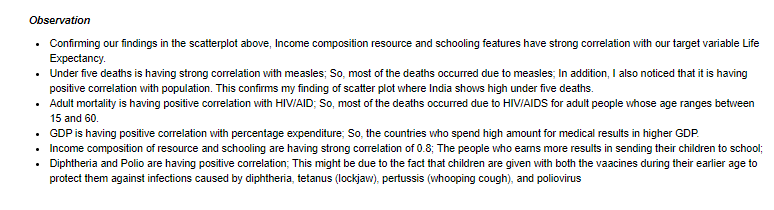


Figure : Heat Map to show the correlation

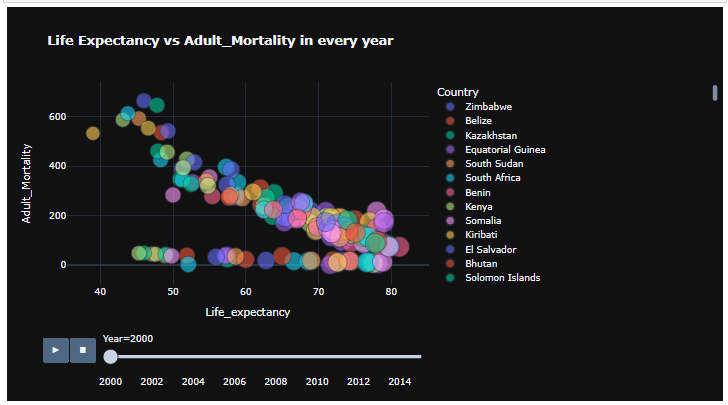
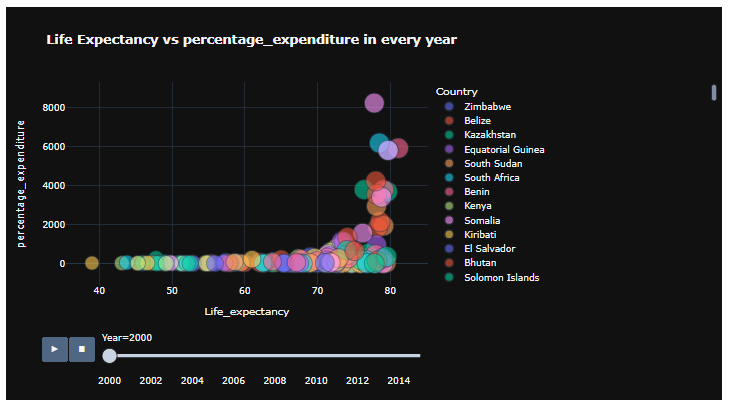


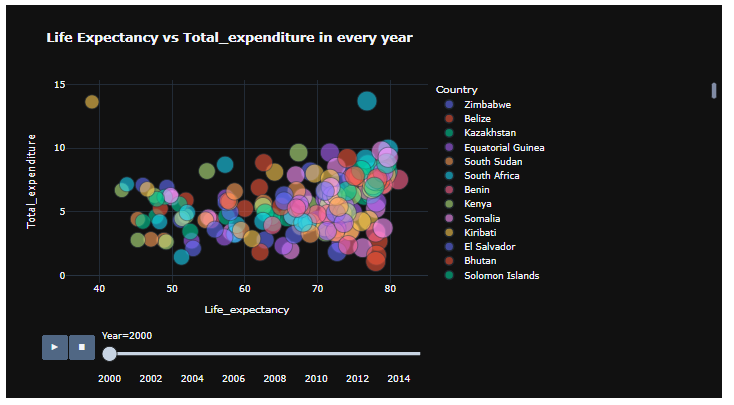
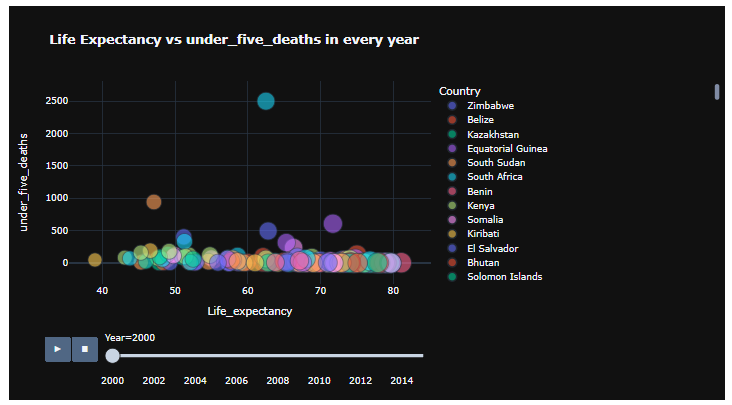
**Figure 7 Heat Map Correlations and observation**

#### Target Variable Analysis

The following scatter plots are created to analyze our target variable “life expectancy”. The details are shown in figure 8.

* **Life Expectancy vs Adult Mortality over years for each country:** We see the adult mortality rate is low for developed countries compared to developing countries. Due to the development of medicine in developing countries, the adult mortality rate has increased over the period. However, life expectancy has increased and adult mortality has decreased over years in the countries
* **Life Expectancy vs Percentage Expenditure over years for each country:** The medical spending contribution by developing countries is low resulting in low life expectancy compared to the spending by developed countries which results in high life expectancy. I could see the same pattern for all the years.
* **Life Expectancy vs Total Expenditure in every year and country:** Most of countries the total expenditure lies below 10 and only limited countries are having expenditures greater than 10. Most of the values for life expectancy lie in the range of 40 and 90. The United States is having high total expenditure and life expectancy. Life expectancy has increased over years for all the countries
* **Life Expectancy vs under\_five\_deaths in every year and country:** Most countries are having values of less than 500. I also noticed that India is having high under 5 deaths. This might be due to population as India is the 2nd largest population country in the world. However, over the year under-five deaths have been reducing in India.
* **Life Expectance vs BMI in every year and country:** Over the years, the BMI across the countries has increased which results in high life expectancy. This is because people are more cautious about their health nowadays compared to older years.
* **Life Expectance vs Schooling in every year and country:** Schooling also plays a major role in improving life expectancy. Schooling gradually increases over the years which results in high life expectancy.

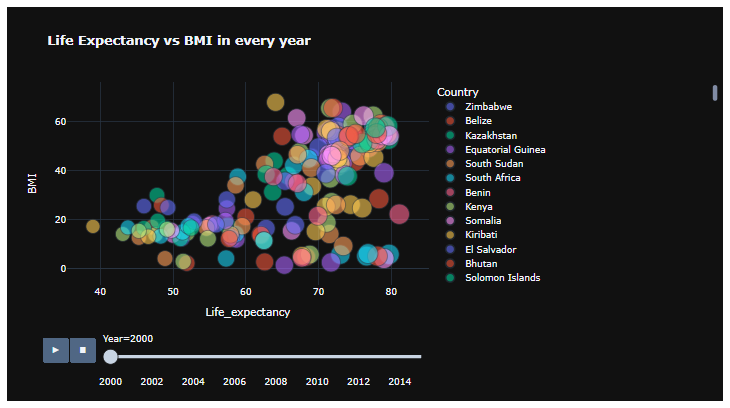
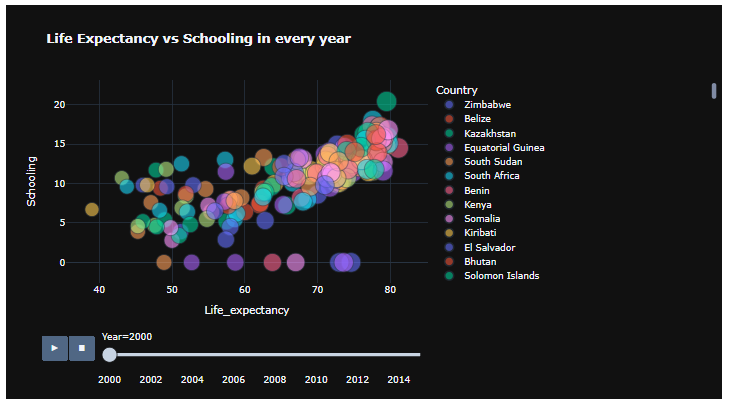
 

Figure : Scatter Charts

#### Regression Plot

The regression plot creates a regression line between 2 parameters and helps to visualize their linear relationships. The regression plots for the features (BMI, Income composition of resources, adult mortality, GDP, and infant deaths) have been created to understand the relationship with the target variable “life expectancy”. The details are shown in figure 9. I could see life expectancy increase for an increase in BMI, GDP, and income composition of resources and a decrease for an increase in adult mortality and infant deaths as expected.

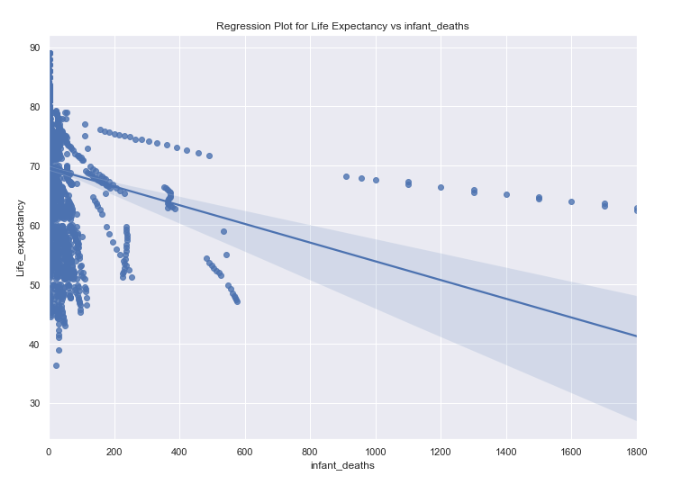
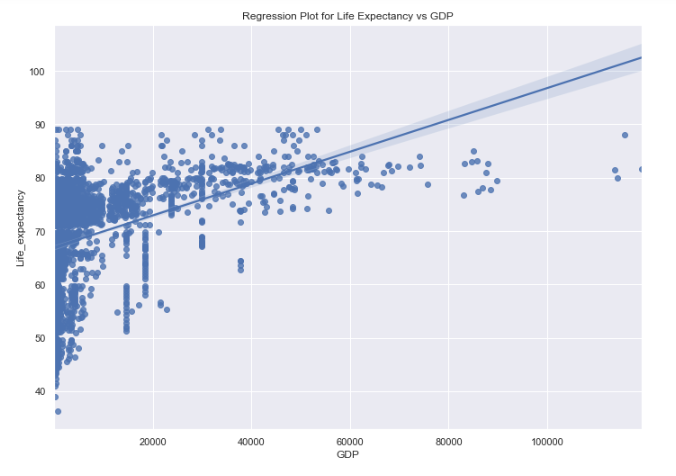
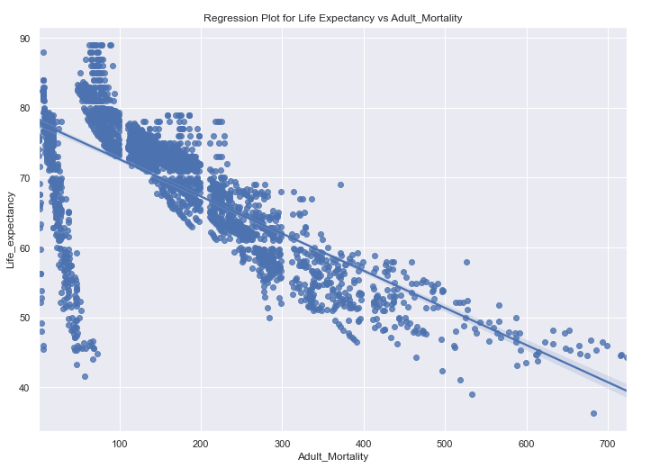
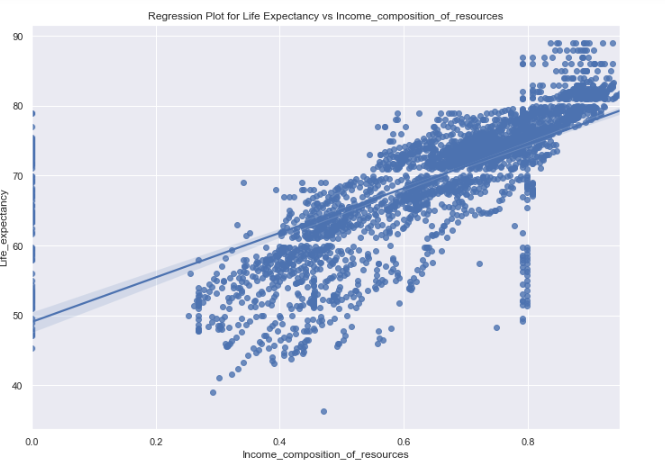


Figure : Regression Plot

## **Methods**

Since the target variable “Life Expectancy” is a continuous variable, the linear regression model has been executed on the dataset to determine the features that are mostly related or correlated to our target. Linear regression is commonly used for predictive analysis and modeling. Simple Linear Regression is a type of regression algorithm that models the relationship between a dependent variable and a single independent variable. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Feature selection is one of the most important parts of any ML model. We always want to select those features which have the maximum effect on our final output. I will perform the following operations for feature selection:

* Pearson Correlation Matrix
* Lasso Method
* OLS Regression

## **Analysis**

### **Modeling Analysis**

Initially, I ran the linear regression on the data without applying any normalization technique. Then, I applied the MinMaxScaler normalization technique which scales the minimum and maximum values to be 0 and 1 respectively and calculated the score. The score and mean square errors for these 2 approaches are given in table 1.

|  |  |  |
| --- | --- | --- |
| **Methods** | **R2 Score** | **RMS Error** |
| Without Normalization | 83.54% | 14.26 |
| With MinMaxScaler Normalization | 83.54% | 0.00513 |

Table 1: Linear Regression Scores for various techniques

In linear regression, accuracy cannot be measured due to continuous target variables. If we try to evaluate the accuracy, we will end up overfitting the model. So, the performance of the linear regression model is evaluated using the below metrics.

**R2 (R-Squared):**  R-squared (R2) is a measure of how close the data points are to the fitted line. It is also known as the coefficient of determination. Below is the acceptable range for R2 values.

* R-squared value < 0.3 this value is generally considered a None or Very weak effect size
* R-squared value 0.3 < r < 0.5 this value is generally considered a weak or low effect size
* R-squared value 0.5 < r < 0.7 this value is generally considered a Moderate effect size
* R-squared value r > 0.7 this value is generally considered a strong effect size

**Root Mean Square Error (RMSE):** Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are

Among the 2 linear regression models performed on the input data, the one with the MinMaxScaler normalization technique applied gives a better result with a low root mean square error value. The R2 value is the same for both techniques. Figure 10 represents a regression plot for the predicted values with and without the normalization technique.

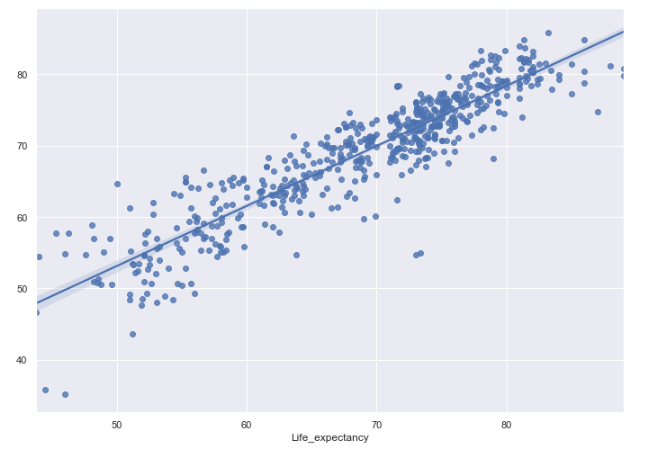
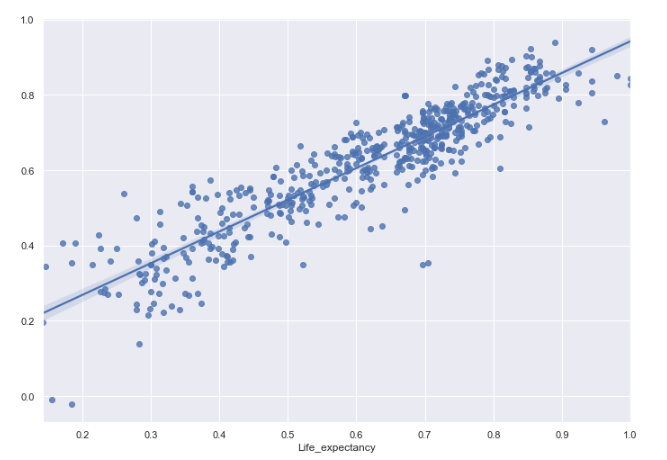
 

Figure : Regression Plot for predicted values without normalization and with normalization

### **Feature Analysis**

Using various methods as mentioned in the method section, I tried to find the best features from the dataset and the following are the best features in the dataset which show a higher impact on the target variable “life expectancy” compared to other features present in the dataset.

|  |  |
| --- | --- |
| **Test** | **Features** |
| Pearson's correlation matrix - Feature correlation to target variable "life expectancy" | **Positive Correlation:**   * Schooling * Income composition of resources * BMI   **Negative Correlation:**   * Adult Mortality * HIV/AIDS * Thinness\_1\_9\_years |
| Lasso Regression | BMI  Polio  Diphtheria  Income composition of resources  Schooling |
| Ordinary Least Square (OLS) | Year  Hepatitis\_B  GDP  Population  thinness\_1\_19\_years  thinness\_5\_9\_years |

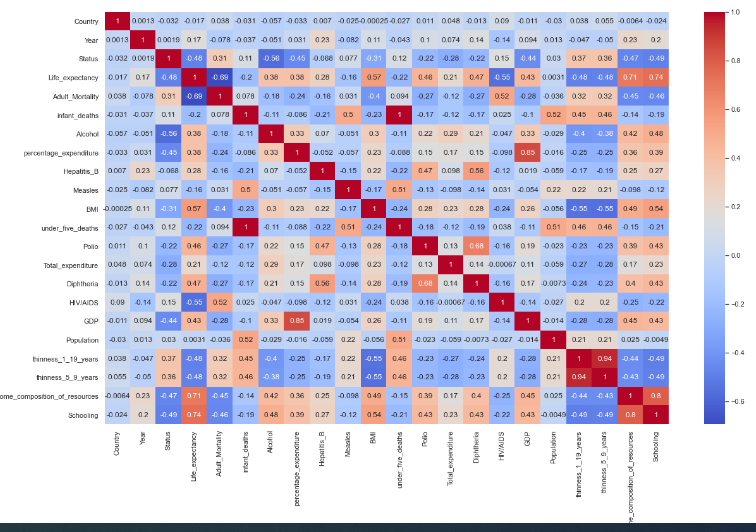


Figure : Pearson Correlation Matrix

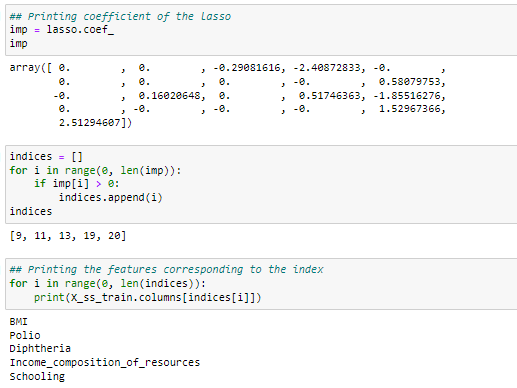


Figure : Best features using Lasso Regression

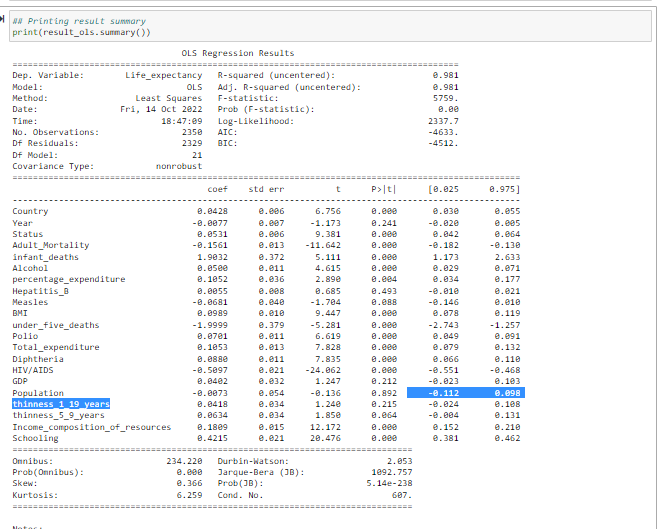


Figure : OLS Feature Extraction Summary

## **Conclusion**

The linear regression applied to the normalized data gave the best result and is used to predict life expectancy. Though the R2 value is the same for the linear regression applied to denormalized and normalized data, RMSE has been substantially reduced for the regression applied to normalized data.

Among various methods used to find the best features in the dataset, below are some of the top features having a high impact on the target variable “life expectancy”.

* BMI
* Income composition of resource
* Schooling
* Diphtheria
* Polio
* Population
* GDP
* Hepatitis-B

## **Assumptions**

The dataset contains null values across many features. I have substituted those null values with the ffill method (forward fill) available in the fillna method of pandas. I assume the values substituted would be correct as this method propagates the last valid observation forward.

Among various techniques available for feature extraction, I have considered only Pearson correlation, Lasso, and OLS methods as I assume these methods would give the best and expected results.

## **Limitations**

The dataset considered for this prediction analysis contains only limited rows (~3000 rows). So, the prediction and feature extraction performed during this analysis is based on this limited dataset. If this dataset doesn’t accurately reflect the original real-world data, modeling efforts cannot generate any useful insights.

## **Challenges**

There are a couple of challenges I faced during the data preparation step for model building. Identifying the correct features that contribute to the target, planning on how to handle the null values present across various features of the dataset, and deciding on which feature extraction model to use to name a few.

## **Future Uses/Additional Applications**

We see predicting factors affect life expectancy in one way or another. We saw leaving out some features does affect the average value of life expectancy. Similarly, the addition of a few features may also affect the result. With the addition of features to the dataset, this prediction model can be routinely run to identify the factors impacting life expectancy and take necessary action to increase the life span of human beings.

## **Recommendations**

Based on the available data, this model predicts the life expectancy and features impacting the life expectancy with better accuracy. However, this model should be regressed again when more data is available.

## **Implementation Plan**

With the current features available in the dataset, this model can be implemented to predict the life expectancy of a country. In addition, this model can be launched to evaluate various features impacting life expectancy. However, as additional features are added to the dataset, this model must be reevaluated to ensure there is no slippage due to added features.

## **Ethical Assessment**

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions made from this project’s study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project with a larger audience, the underlying assumptions and data considerations should be shared.

Another ethical consideration is to ensure no personal and sensitive information is present in the dataset. This dataset is an extract from Global Health Observatory (GHO) data repository governed by WHO and made available for public use. So, this dataset doesn’t have any sensitive information.

## **References**

KUMARRAJARSHI (2017). Life Expectancy (WHO).<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>

Arya Andhika (2019-08-12): Life Expectancy Prediction using Regression. <https://rstudio-pubs-static.s3.amazonaws.com/534874_2bdd7c6645804fd1b240e1ca3a9eb9d6.html>

World Health Organization: <https://www.who.int/data/gho>, <https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates>

Shashank Gupta. <https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression>

Caitlin McDonnell (2018-02-04). Machine learning to predict life expectancy. <https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1>