1. Objectives/Vision

Everything has an expiration date; humans are no exception either. Our proposal today is to predict the life expectancy of a person given the essential parameters with the help of machine learning models. Life expectancy is affected by a variety of socioeconomic, genetic, and environmental factors. Therefore, it makes for a good area of study to investigate via statistical methods to better understand the measure of life expectancy.

We would like to explore the life expectancy (WHO) dataset on Kaggle. We would be (Moreover our analysis for the given score would be by utilizing the linear and more complex regression) analyzing the scores by applying linear and more complex regression (algorithms on our chosen dataset) and trying to come up with the most effective model which will give us a good prediction on the life expectancy of individuals. Furthermore, we will make an effort to propose the most effective model which will give us a good prediction of the life expectancy of individuals.

Life Expectancy measures the average time a human being is expected to live based on different parameters such as DOB, current age, gender, environment, Medical History, Lifestyle Choices, Ethnicity, Race, and many more. When Interpreting Life Expectancy, expect that very few people might die at the expected or precise age indicated by the life expectancy prediction. Good prediction helps determine the course of treatment, helps in the procurement of medical services and facilities, and probably improves the quality of the final phase of the life cycle.

………………………..

…………………………………

………………………………………….

…………………………………………………..

1. Stakeholders

Firstly, this study would be beneficial for the general public in a broader way that every individual may be (curious/interested) and would like to know how much they would be living based on the parameters discussed in our analysis. Secondly, government health bodies and private organizations would be (interested/concerned) in our findings which would help them give an understanding of where they are standing in terms of life expectancy of their individual and whether they should be increasing their health expenditure to address issues which might affect the life expectancy of their individuals.

Increased longevity of life can keep an individual mentally engaged with the work they enjoy, preventing or minimizing stress, and loneliness.

It also provides organizations to keep employees of higher age groups in the organization as they are amongst the most engaged age group. This also provides an opportunity for different generations to learn from each other and improve development amongst the employees.

…………………………….

………………………………………

………………………………………………

……………………………………………………….

1. Research Questions

Our analysis aims to answer the following research questions:

1} Which area should a country focus on to efficiently improve the life expectancy of its population?

2} What is the overall average life expectancy in most countries?

3} Do developed countries have more life expectancy than developing countries?

4} After what amount of alcohol per capita consumption does the life expectancy increase? Does Life Expectancy have a positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol, etc?

5} What is the ideal range of population for any country to have a higher life expectancy?

6} What is the impact of schooling on the lifespan of humans?

7} How do Infant and Adult mortality rates affect life expectancy?

8} By what age do the organisations need to provide incentives or guidance as the employees age so that they can have a good state of living?

1. Dataset range (Dataset Acquisitions & Merger) - Describe variables and granularity

Data Source:

KumarRajarshi. (2018, February 10). Life expectancy (WHO). Kaggle. Retrieved on August 26, 2022, from https://www.kaggle.com/kumarajarshi/life-expectancy-who.

Past studies were conducted on the above dataset which was for the period 2000 to 2015, but in our analysis (have) we have included data for the period 2000 to 2020 by researching various datasets from the public domain extracting and exploring other important parameters believed to affect the life expectancy prediction. These datasets that were acquired and merged to form a final dataset with variables as described below.

Past studies were conducted on the above dataset which was for the period 2000 to 2015. In our analysis to add more granularity, we have tried to combine multiple datasets to explore various other variables which we thought might have some or little effect in predicting the remaining life span of individuals. After combining all these data we have now information on variables starting from 1990 to 2020.

| Sr. no. | Variables | Description |
| --- | --- | --- |
| 1. | Country | Name of countries used for analysing life expectancy. Total of 193 countries are used for analysis. |
| 2. | Year | It has the year in which the particular data points are recorded. |
| 3. | InfantMortality | It shows the mortality rate on an Infant. |
| 4. | Population | It provides the population of a particular country under analysis. |
| 5. | GDP | Expenditure on health as a percentage of Gross Domestic Product per capita(%) |
| 6. | Number of executions(Amnesty International) |  |
| 7. | Deaths -XXX | Deaths caused because of different diseases/incidents. |

1. EDA - can bring up more questions
2. The areas where a country should focus on in order to improve the life expectancy of its population could be found out from the following visualisations which shows correlation between the predictor variables and our response variable “Life Expectancy”.

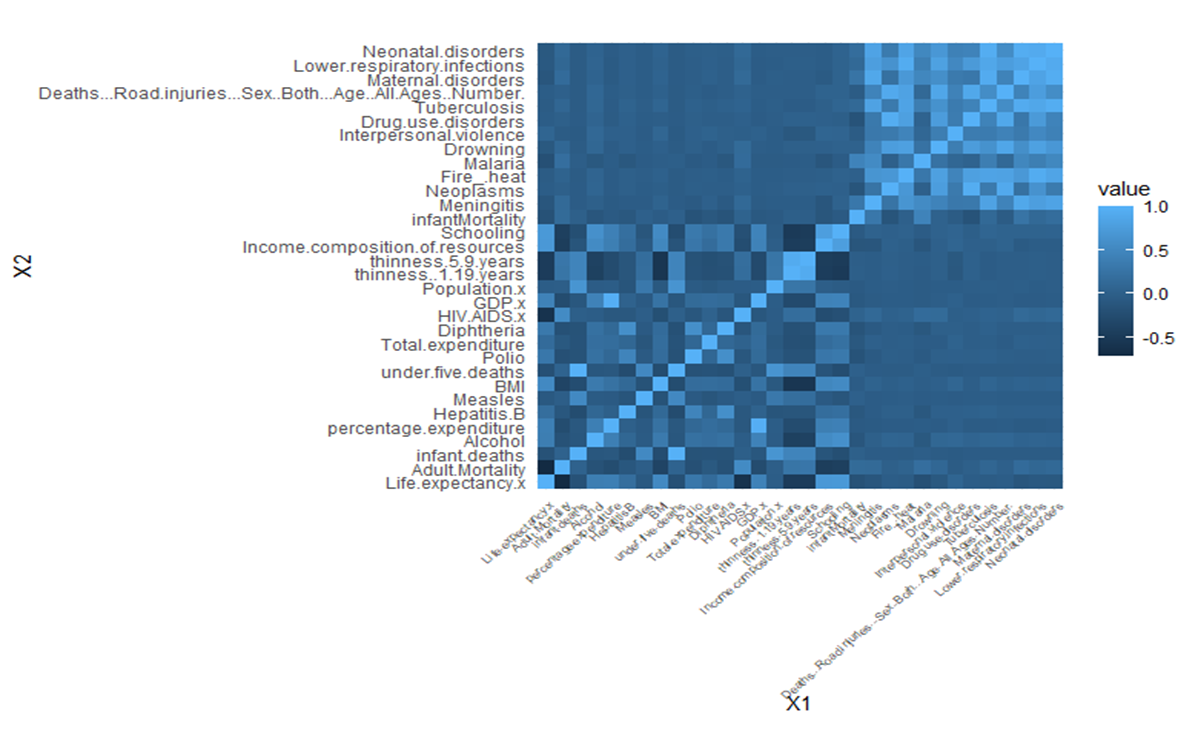
(To be removed once we get eda from our team, just keeping to fill the document).

Fig.1 Shows variables which has more correlation to Life Expectancy.

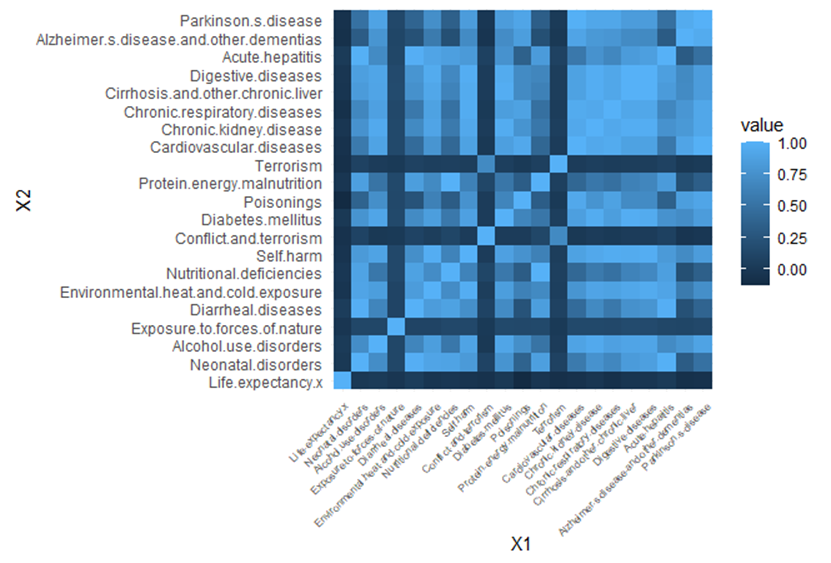


Fig 2. Factors having less or no impact on Life Expectancy

2} What is the overall average life expectancy in most of the countries.

3}Does developed countries have more life expectancy than developing countries.

The following visualisations shows that developed countries have more life expectancy compared to the developing or underdeveloped countries.

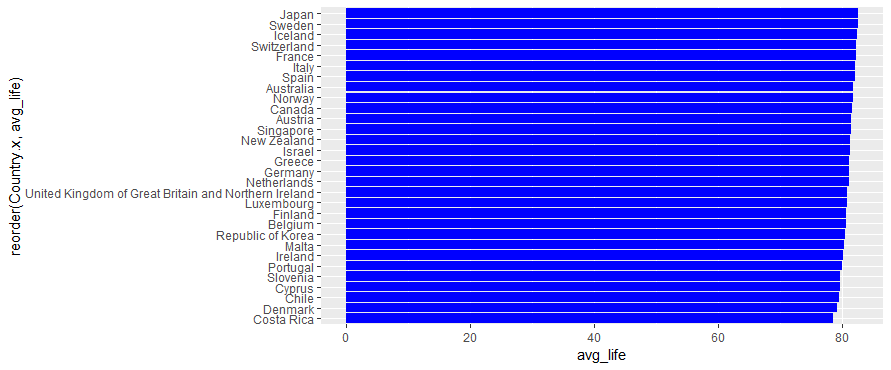


Figure3 Top 30 countries as per their average life expectancy

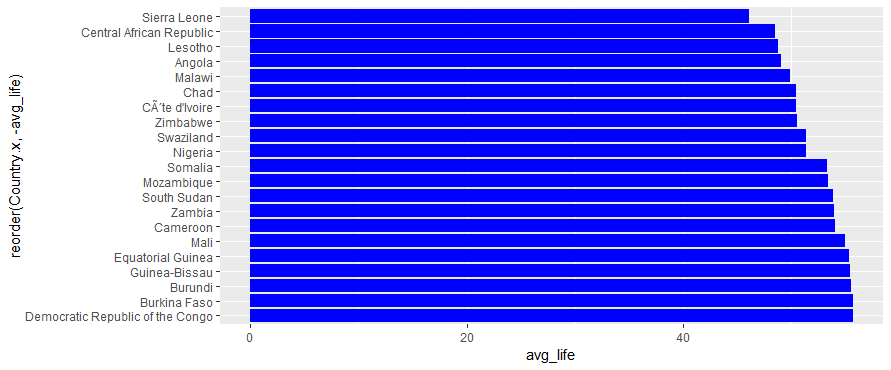


Figure 4 Bottom 30 countries as per their average life expectancy

1. Modeling techniques used

- Still need to discuss which ones to propose

- Reason why we chose the particular regression model. Starting from LR and proceeding with multiple Reg, Techniques.

1. Issues that can be anticipated

* Issues might arise while merging the datasets collected from different sources.
* Issues might arrive because of imbalanced datasets while training the regression model.
* We might not have enough data for training and testing the model, which might result in overfitting/underfitting models.

1. Appendix - Code for data Acquisitions and Merger
2. References ( Req. or not) - (Not sure?)

* Data set sources references
* Kaggle references.
* EDA references.