Case Study: Life Expectancy Visual Analysis

Nidhi Joshi

Abstract— This study aims at assessing and predicting country-wise health indicator based on various factors like the country's economic status, mortality rates, disease counts, healthcare expenditure etc. In pursuit of determining relationship of Life Expectancy with various independent features, the data is prepared and transformed as needed to then plot histograms which help in identifying unique distributions. Then, the research goes on to explain the Analysis approach and process where it aims to determine probable association with the above mentioned factors through correlation matrix, visual analysis techniques and implementing models like K-Means Cluster and Linear Regression to arrive at possible answers to the posed problem statement.

1 Problem Statement

Health and demographics have constantly been measured and researched by public health authorities, researchers, governments, and healthcare organizations et al for their various respective application areas.

Numerous factors are monitored and evaluated over time for the sake of policy formulations, public health planning, epidemiological research, planning medical infrastructure and supply chains apart from being prepared for emergencies with strategical preparedness.

With a focus on life expectancy in different geographies, this study intends to answer the following important aspects:

- Change in health indicators based on specific economic factors over the years
- Health outcomes based on lifestyle factors and level of correlation between them
- Predicting the impact of region's economy and healthcare expenditure on the health indicators

This research aims to measure the factors involved in determining the life expectancy in different regions to help shape important health policies along with strategizing for disease prevention. It can not only help with economic and social development factors, but also improve the medical workforce planning and distribution giving a boost to the country's infrastructure.

The data is sufficiently suitable for analysing life expectancy in part of regions across the world. The considered features included Alcohol Intake, BMI, Schooling along with GDP, Percentage Expenditure on Healthcare, and Income Distribution on Resources. These aspects gave a comprehensive take on lifestyle choices, impact of education and region's economic factors to determine life expectancy. Most of the values are quantitative in nature which was a major help to predict health indicators through regression models.

2 State of the Art

Over the last century, one of humanity's most remarkable achievements has been the significant improvement in life expectancy. In the span of a hundred years, our world underwent a profound transformation, transitioning from a time when hardly any countries had a life expectancy surpassing 50 years to an era where many nations now enjoy an average life expectancy of 80 years. This dramatic increase was particularly evident in regions known for their long-lived populations. Initially, the extension of life was driven by the reduction of infectious diseases, especially among the younger population. As we successfully tackled and minimized deaths from infectious conditions, attention shifted to addressing cardiovascular diseases and cancer, which emerged as the predominant causes of mortality. Science and medicine then focused their efforts on understanding and combating these health challenges during the latter half of the previous century. [1]

The first paper from Lancet [2] which was referred, sourced its data from 8259 credible sources like UN Population Division, WHO, US Census etc. The data involved mortality rates from 918 overall locations including countries and other territories for both genders among 23 age groups starting from 1950 till 2017. This paper aimed to predict mortality rate based on age (which is basically Life Expectancy) by utilizing a spatiotemporal regression framework while blending the results of used methods across each intercensal period. This study crafted a continuous time series based on each individual's location. How this research experimented with death rates to determine Life Expectancy at birth gave me an insight for my analysis in such a way that I experimented with various available mortality rates in my data with Life Expectancy to discover a robust relationship.

The second research paper was about mental healthcare trends in US adults from 1999 to 2018 [3]. The paper sourced its data from National Health and Nutrition Examination Survey (NHANES) and utilised Logistic Regression model to determine trends and dependent features for mental health. There were 55,052 records to study. While this research was not exactly similar to my case study, it provided me insights on how mental health is directly linked to decreasing life expectancy and increased burden on medical expense for a country. Thus, it gave a prospective idea of enhancement to my research as future work.

The third study from National Library of Medicine [1] dealt with a National Center for Biotechnology Information (NCBI) database while visually analysing various disease related variables with the life expectancy as a final target. This study explored how negative changes in disease counts contributed to positive changes in Life Expectancy hence, prompting to analyse this side of aspect as well.

3 Properties of the Data

Human beings have been living longer, year on year. "In 1900, the average life expectancy of a newborn was 32 years. By 2021 this had more than doubled to 71 years." [4]

This research uses data from Kaggle [5] on Health and Demographics to evaluate health indicators leading to life expectancy across geographies. The data comprising of 22 columns and 1649 records involved country-wise health-related details spanning from year 2000 to 2015. The fields included Status ('Developed' or 'Developing'), Age-wise Mortality counts, Alcohol intake index, Average BMI, Total and Percentage expenditure on health care, Income composition of resources, GDP (per capita), Population Index, Schooling, Age-wise thinness index and disease counts for Hepatitis B, Measles, Polio, Diphtheria and HIV/AIDS with Life Expectancy as the variable to be predicted as an important health indicator later on through regression analysis in order to answer the third problem. While most of the fields are self-explanatory, Income composition of resources represents "a composite index that reflects the income distribution and access to resources in the country" [5] and Life expectancy is a calculated field equating to an average expected life (in years) for babies born in the documented year. Country, Year and Status consist of qualitative values while the rest of them are all numeric values. The complete analysis was code written in Python 3 [6] in Jupyter notebook 6.5.4 [7]

While no null or duplicate values were detected in the dataset, but some fields like Infant deaths, Percentage Expenditure, Measles, Under-five deaths, and Income composition of resources contained '0' values for some of the records. It could be assumed that either the data in some cases literally can amount to '0' or it has not been documented for that particular year for the respective country (i.e. Percentage Expenditure). In either case, the records couldn't be dropped as the rest of the variable values were of much importance to be discarded for analysis. Replacing these values with the column's mean would also have altered the analysis in a non-desirable way. Hence, the values were retained in their original form. A new field was added to the dataset combining all the available mortality counts into one as 'Total Deaths' which was calculated as the sum of 'Adult Mortality' and 'under-five deaths'. The field 'infant deaths' wasn't considered for the calculation as the count was assumed to be covered under 'under-five deaths' variable.

While functions like isnull(), duplicated() were used in order to determine the missing or duplicate values; distributions, skewness and outliers were examined by plotting histograms *(Figure 1)* for all the variables. Some fields like percentage expenditure, infant and under-five deaths, Adult mortality, Alcohol, diseases like Measles, HIV/AIDS came out to be right skewed understandably. Because only 15% countries had 'Developed' status, the right-skewedness of the GDP can also be justified. However, Hepatitis B, Polio and Diphtheria were alarmingly left-skewed indicating higher such cases in significant number of countries. These visualizations helped influence the foundation of further analysis through Clustering.

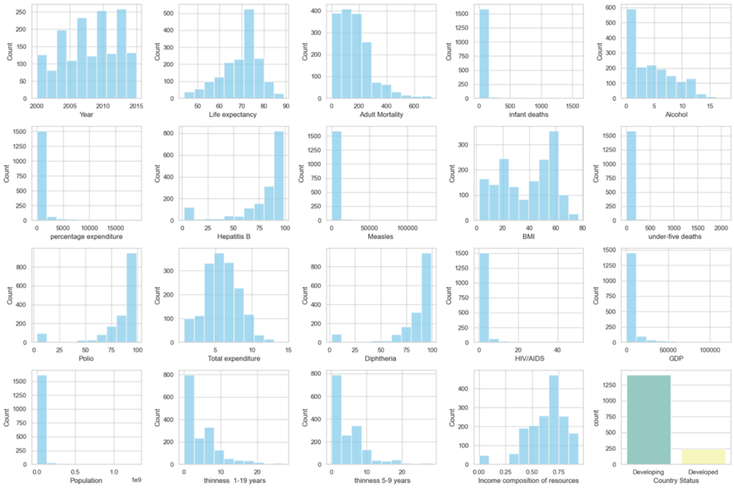


Figure 1: Histogram plots for all variables

4 Analysis

4.1 Approach

The Analysis Approach section deals with the steps taken to arrive at conclusions for the important research aspects.

The first and foremost task was to define a problem statement according to one's area of interest and then figure out the kind of data needed to cater to the problem statement. The related dataset was gathered from the available Kaggle data sources [5]. The features were examined carefully to check if they suit the research questions to be solved. Once the dataset was found to be convincing enough to be able to answer the questions, the approach could proceed to the next task i.e. Data Pre-processing.

The pre-processing of data involved Data wrangling and transformations as needed to arrive at comprehensive and sorted results. Data was checked for missing, erroneous, duplicate values or outliers, if any. There were none but some records were found to be carrying '0' value. Since, the fields involved were not much of a computational value and didn't make much difference to the analysis so by human judgement, it was decided to not drop those rows. Data transformation was done by combining multiple mortality related variables into a single measure i.e. 'Total deaths'. Data preparation was an important task since it lays the foundation for analysis and provides with a clear way to proceed. It could be revisited at times to distil our results better.

Visualizations combined with computational methods further aid the purpose of human intervention to effectively reason out insights out of the data. Histograms were plotted for each variable with an aim to assess value distributions and skewness of the data through human intervention. Furthermore, correlation matrix was employed to gauge the relations between different variables. One could humanly reason out from the matrix about the strongly related variables and plot them again on the correlation matrix, this time only with the selected features to analyse and compare further. Different related features could also be plotted on boxplots to measure the outliers along with comparison against median values.

Now, with the correlated values and selected features, a cluster model could be employed to see how different clusters were placed across on a world map. It would prove to be an interesting visualization at this point, as one could realise some missing geographies in the data by human judgement. To measure the efficiency of the cluster model, performance metrics could be calculated and visualized through an appropriate plot (in this case, a scatter plot). The parameters can be optimized at this point if the performance is not found to be satisfactory. If more questions could be answered by utilising another model, it could be done and measured again by plotting performance metrics on a graph. In this particular study, human judgement helped determine that the Linear Regression model could be used to derive conclusions.

The final results could then be visualised for assessing the predicted values with those of actual and conclusions can be drawn for the questions from the plot.

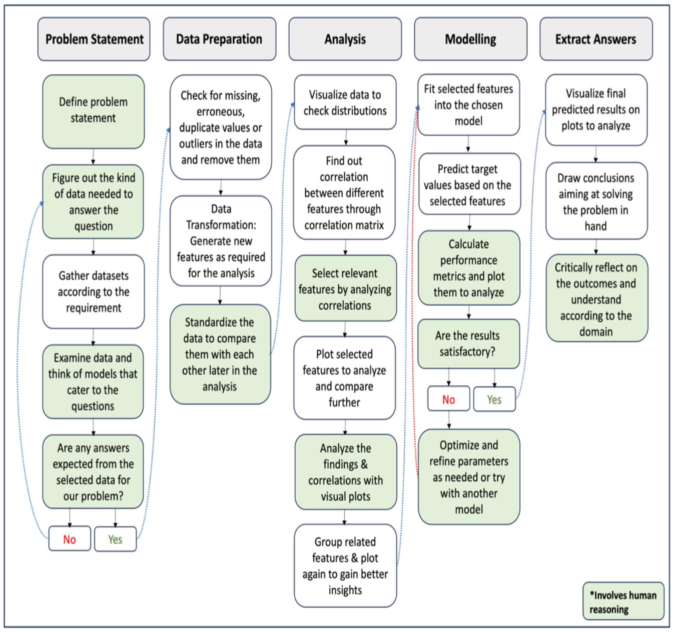

Problem Statement
Data Preparation
Analysis
Modelling
Extract Answers
Define problem statement
Figure out the kind of data needed to answer the question
Gather datasets according to the requirement
Examine data and think of models that cater to the questions
Are any answers expected from the selected data for our problem?
No
Yes
Check for missing, erroneous, duplicate values or outliers in the data and remove them
Data Transformation: Generate new features as required for the analysis
Standardize the data to compare them with each other later in the analysis
Visualize data to check distributions
Find out correlation between different features through correlation matrix
Select relevant features by analyzing correlations
Plot selected features to analyze and compare further
Analyze the findings & correlations with visual plots
Group related features & plot again to gain better insights
Fit selected features into the chosen model
Predict target values based on the selected features
Calculate performance metrics and plot them to analyze
Are the results satisfactory?
No
Yes
Optimize and refine parameters as needed or try with another model
Visualize final predicted results on plots to analyze
Draw conclusions aiming at solving the problem in hand
Critically reflect on the outcomes and understand according to the domain
*Involves human reasoning

Figure 2: Analysis Approach Diagram

4.2 Process

The process flow caters to this study's three main questions as mentioned below:

1. What is the change in health indicators based on specific economic factors over the years?
2. Are health indicators based on lifestyle factors in any way?
3. Can we predict the life expectancy based on region's economy and healthcare expenditure?

Question 1: What is the change in health indicators based on specific economic factors over the years?

With an aim to answer the first question, different countries were first filtered out and assigned to separate data frames based on their economic status – Developing or Developed. The two data frames were melted to be made compatible with Seaborn library and then were plotted on a seaborn line plot with 'Year' on the x-axis and 'Life Expectancy' on the y-axis with the Country 'Status' as the hue *(Figure 3).*

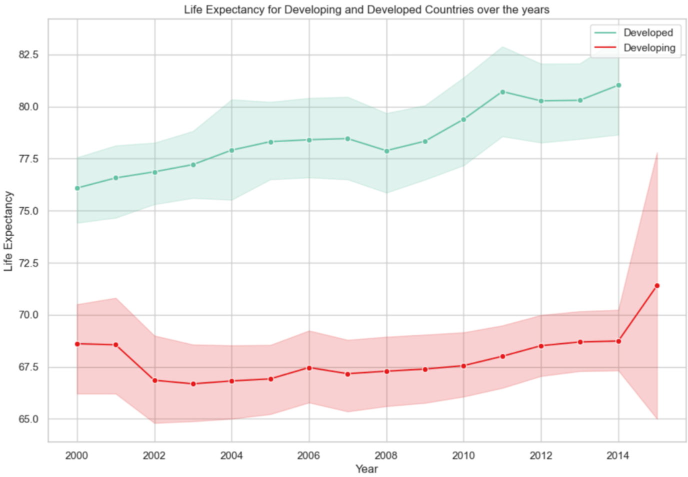
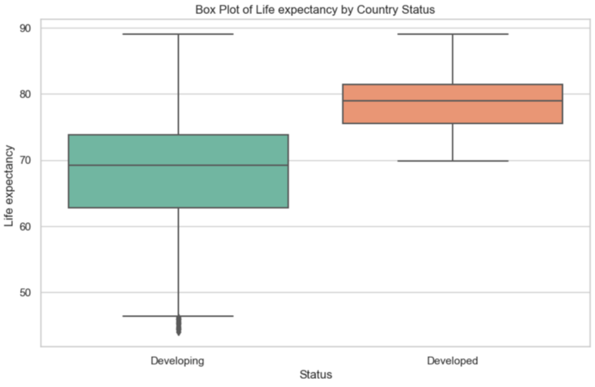


Figure 3: Year-wise Life Expectancy for different economies

It could be humanly judged from the plot that while 'Developed' economies continually displayed better Life Expectancy year-on-year, there was a surprising dip in Life Expectancy from 2001 to 2002 in case of 'Developing' economies. Even though, the numbers picked up and steadily rose in later years, but the reason for the dip between 2001 and 2002 can be a prospective research question for another subsequent study to this paper. The constant rise in Life Expectancy with each year can be attributed to improvements in medical infrastructure, better research in terms of pandemics and diseases and antibiotics development and availability along with better economic conditions and affordability for masses across the world. Based on the mentioned factors, it understandably makes sense why there is so much of distinction between Developed and Developing economies when it comes to Life Expectancies. It also gives rise to the ever-raging research and debate on whether rich countries really do have a higher life expectancy as compared to those nations in poverty. While this research question can be a matter of concern for another study, this particular plot clearly distinguishes between the results based on economic status and helps with clear comparison of the target variable, hence an appropriate graph to generate findings and answer our first question. However, a box plot was plotted as well to explore the outliers and summary stats were analysed for the target variable according to different economic regions *(Figure 4)*.



```
          count       mean       std   min    25%    50%   75%   max
Status
Developed   242.0  78.691736  4.273485  69.9  75.55  78.95  81.4  89.0
Developing 1407.0  67.687349  8.353983  44.0  62.80  69.20  73.8  89.0
```

Figure 4: Life Expectancy Boxplot and summary stats

Question 2: Are health indicators based on lifestyle factors in any way?

Now, to answer the second question, it was necessary to determine the relation between various features pertaining to lifestyle. To decide the same, a correlation matrix was plotted with all the quantitative variables in hand through a Seaborn heatmap along with the numeric correlated values for better judgement (*Figure 5a*). By human intervention, the relevant related features were identified along with a probable disease which was related to the lifestyle (HIV/AIDS) to measure the correlation again through Seaborn heatmap along with numeric measures values (*Figure 5b*). It was concluded through the plot that even though other lifestyle factors like Alcohol intake, BMI, or GDP per capita were strongly related to the health indicator in question – Life expectancy, Schooling was the most important factor. It could be successfully judged in human capacity that education played an important role in living longer by taking better health related decisions in one's lifetime. This was an interesting conclusion apart from the sought answer to the research question.

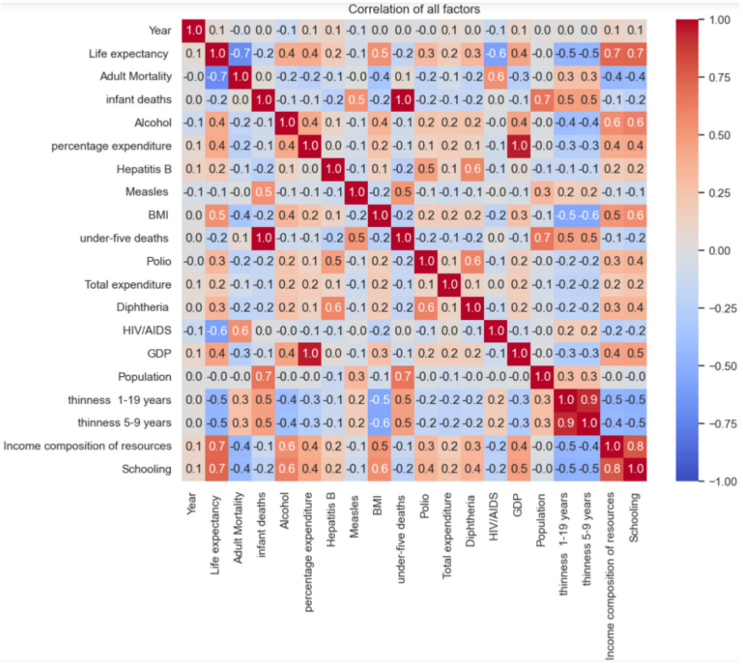


Figure 5a: Heatmap Correlation matrix of all variables

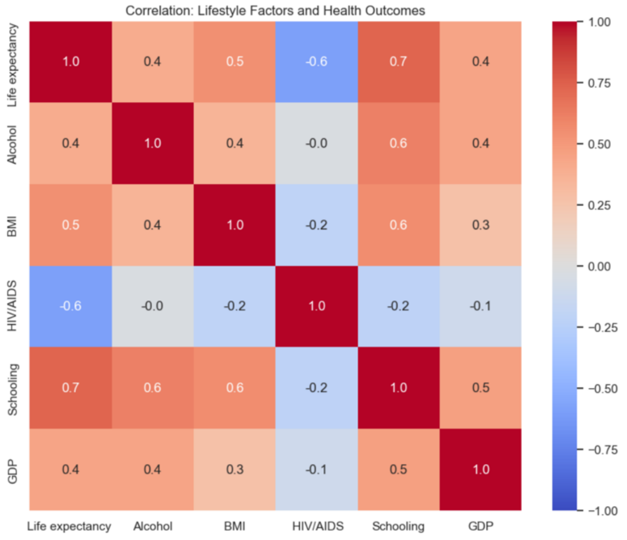


Figure 5a: Heatmap Correlation matrix of chosen variables

Even though our question was already answered but to further visualize other related factors to Life Expectancy, Scattergeo trace graph object was utilised from plotly.express() function (*Figure 6*). Initially, chloropleth maps were used to display the results but not so successfully as it was realised that the dataset missed many of the important regions like USA and some parts of Russia and Africa, so those regions were not displayed on the map at all resulting in a distorted view of the map and hence, not visually appealing. Scattergeo trace was considered as a better option consequently. Apart from the Life Expectancy itself, Healthcare expenditure and a newly formed feature ('Total Deaths') were traced on the graph. 'Total Deaths' was derived as a sum of Adult mortality and

Under-five mortality numbers. It was assumed that the infant mortality counts are already covered in the under-five death counts. Clearly, the regions with the maximum expenditure on healthcare amounted for better life expectancy and the ones with the minimum expenditure presented with much lower life expectancy figures. Even the death counts were higher in those regions. One interesting outcome which was observed through human judgement was that even though some regions in Africa didn't spend much on healthcare and didn't even score high on the Life Expectancy, but they still accounted for lower death counts as compared to the Indian subcontinent. This led to the use of K-Means Clustering Analysis to arrive at related clusters (*Figure 7*).

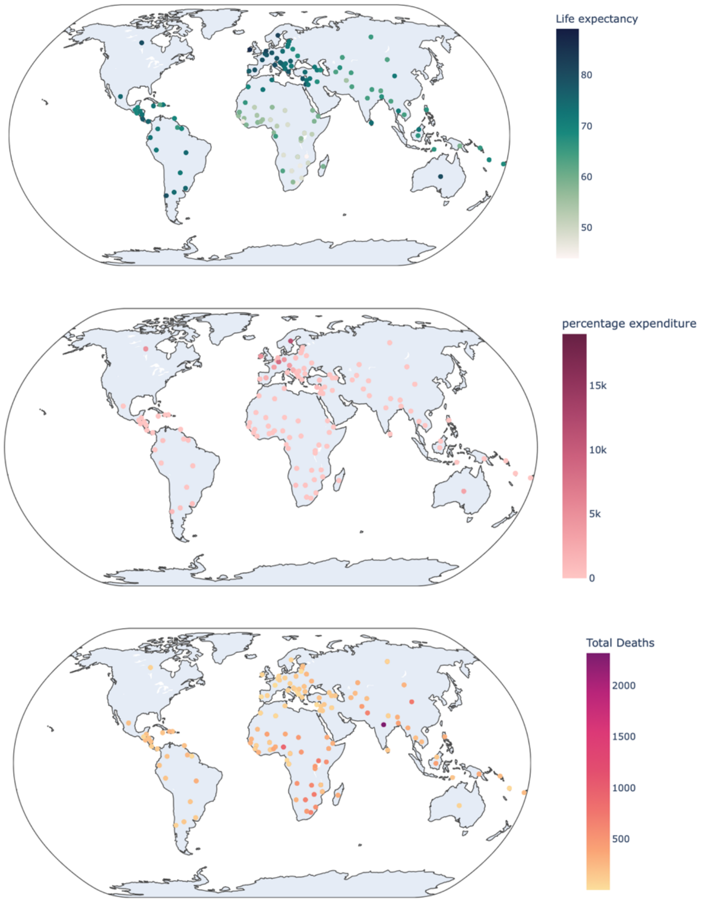


Figure 6: Scattergeo maps for Life Expectancy, Percentage Expenditure and Total Deaths

The considered values were standardized before use to ensure uniformity. Starting with 2 number of clusters, scattergeo was plotted but since there wasn't much detail with only 2 clusters, K-Mean Clustering was tried with 3, 4 and 5 clusters one by one. A decent average silhouette score (0.58) was achieved with 3 clusters along with better visual representation on the map as compared to 4 or 5 clusters on the map, hence the final clusters were decided to be 3 as displayed in (*Figure 7*). The silhouette scores were plotted on a scatter plot as well for human reasoning on the overall score as compared with the average (*Figure 8*).

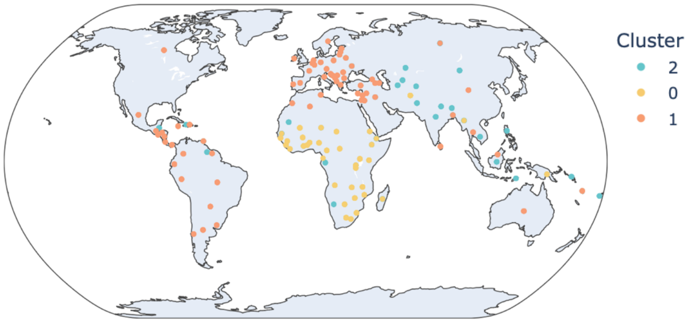


Figure 7: Scattergeo map displaying K-Means Clusters - 3

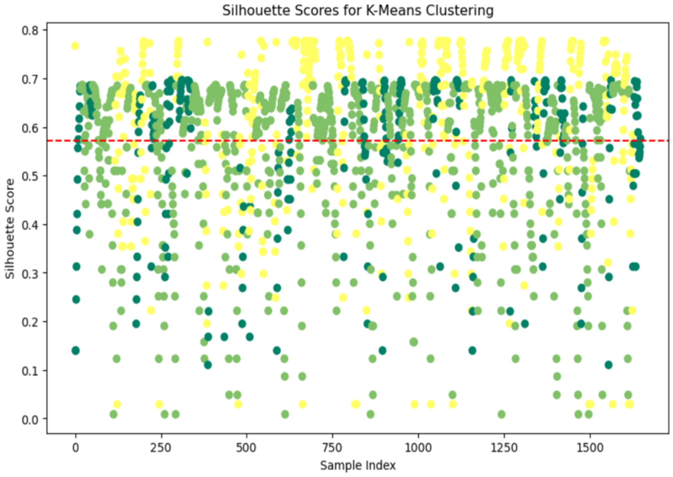


Figure 8: Scatter plot displaying Silhouette score data points along with average silhouette score

Question 3: Can we predict the life expectancy based on region's economy and healthcare expenditure?

Before one caters to this question, it is pertinent to look at disease factors and mortality related factors grouped separately to get an idea of what diseases and which age group deaths might probably contribute the most to the healthcare expenditure. To achieve the same, similar variables are grouped together and standardized before plotting box plots for both the categories (Figure 9). It can be determined by human understanding that Hepatitis B amounted for the maximum cases while HIV/AIDS for the minimum with maximum of outliers in case of Measles. It should be noted however, that HIV/AIDS counts even though small may ask for higher health expenditures as compared to those of Hepatitis B. In the Mortality boxplot, it could be observed that Adult mortality counts were much higher as compared to infant and under-five deaths. Under-five death counts, however, had the most outliers as compared to any other. Even though there was an effort to link these factors in some way to the healthcare expenditure, it was realised that they cannot be linked to the expenditure with certainty in any manner. However, since these give rise to another tangent in the research, hence still included in the study.

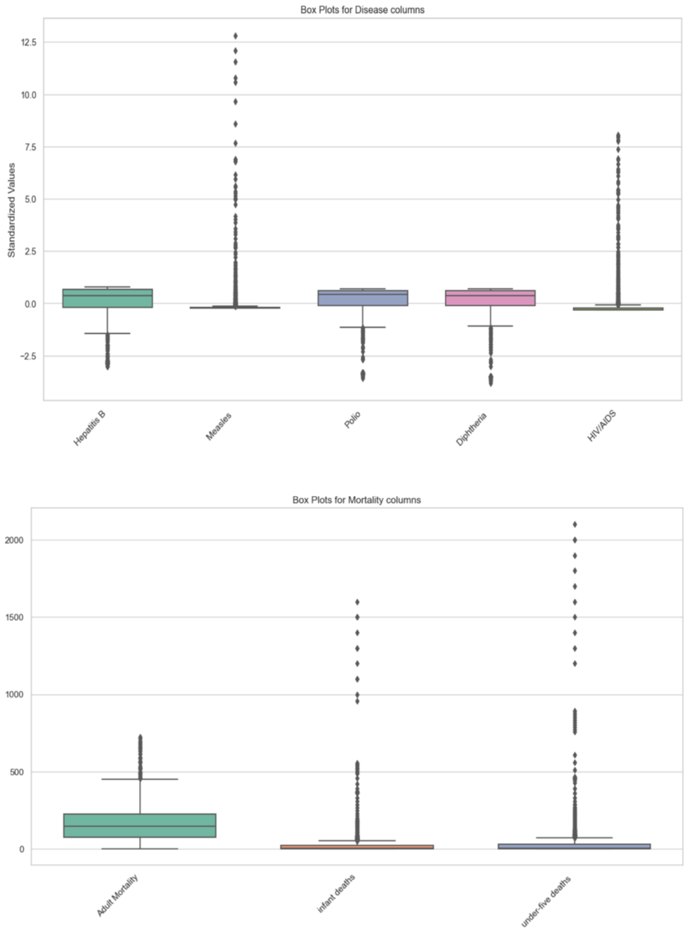


Figure 9: Boxplots for Diseases and Mortality rates respectively

Furthermore, to assess the health indicator based on a country's Developed or Developing status and its respective expenditure on healthcare, it was decided by human judgement to train the selected features to predict the Life Expectancy variable through a Linear Regression model. The model was trained with features like GDP per capita, Income composition of resources, Schooling and percentage expenditure to predict the life expectancy. For validation purpose, the dataset was split into an 80:20 ratio for training:testing and the final regression line was plotted along with the actual results for human reasoning (Figure 10a). Although, the plot displays the actual and predicted result lines to be with residuals but considering the field domain is social science, it can be assumed that these are fairly decent results to be accounted. From the Q-Q plot (Figure 10b) as well, one can determine that since the points are almost falling in a straight line, hence the residuals are almost in a uniform distribution.

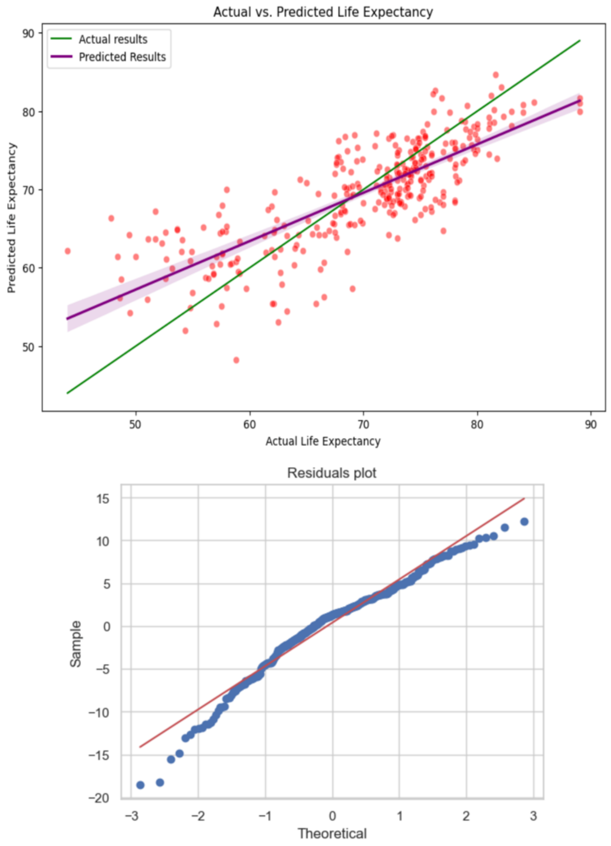


Figure 10: Linear Regression model with Q-Q plot

4.3 Results

The year-wise line plot of Life Expectancy was accurately able to depict a considerable difference in Life Expectancy levels based on a country's economic status. While the developed economies exhibited longer lives for its citizens, the developing economies on the other hand could not ensure a very high life expectancy for its residents in comparison. Hence, it can be said with certainty that there is a direct impact of a country's economy status to its Life Expectancy. Lifestyle factors also seemed to be affecting the health indicator as is evident from the K-Means clusters formed with a decent average Silhouette score of 0.58. Surprisingly, Schooling affected Life Expectancy more than factors like Alcohol intake or economic factors like GDP per capita. Thus, this stresses on the fact that educated masses are better at living longer hinting at their apparent awareness towards medical issues and available aids applied well in time when needed. Finally, Life expectancy could be fairly predicted based on a country's economic status and expenditure on healthcare through the Linear Regression model. This trained model can further be used by government authorities to take appropriate measure in balancing the health expenditure and formulate health policies accordingly.

5 Critical reflection

While the line plot was successfully able to distinguish between the Developed and Developing economies and provided for a good human reasoning but it couldn't explain the reason for the dip in life expectancy from 2001 to 2002, the reason for which can be further analysed in favour of a better research pursuit. Also, apart from the Life Expectancy, mortality counts could also have been compared simultaneously since those counts most probably can be expected to directly impact the health indicator in question.

Correlation matrices provided with an insightful depiction on which features really impacted Life expectancy, so it can be considered as an apt choice to base the K-Means clusters on later. Even though the average silhouette score came out to be 0.58 which could be considered as a suitably decent measure but there were still many datapoints ill-placed in clusters. Since, there are various factors considered for clustering, only Principal Component Analysis (PCA) couldn't suffice to refine the analysis. One could explore another approach involving multiple features to determine appropriate clusters resulting in their correct groups.

In the last approach a country's economic status and healthcare expenditure were intended to be related with Life Expectancy as health indicator. The initial analysis of grouping together disease-related counts and mortality counts to measure any relationship with the target variable was almost futile for the question to be answered. It was intended to establish a link between the deaths, diseases, and expenditure but apparently, this analysis could have answered a separate problem statement. Although, it was an interesting insight for an overall analysis. The employed linear regression model was an appropriate choice to predict Life Expectancy based on a region's economy and health expenditure as also represented by the Q-Q plot. However, other regression models could have been exercised to compare with the applied linear model to compare results and arrive at the best model. While examining the scattergeo maps, it was realised the data is missing for some of the major regions like USA, most parts of Russia and many parts of Africa. Hence, there can be a more comprehensive dataset which can be used for the same analysis with more information. This analysis can be expanded to include other aspects related to health like mental health, medical awareness index or happiness index to undertake an even more comprehensive study. This can be a lesson for a similar future analysis by analysts. The concerned study can prove to be good research covering all aspects around health.

Table of word counts

Abstract	105
Problem statement	246/250
State of the art	437/500
Properties of the data	495/500
Analysis: Approach	496/500
Analysis: Process	1370/1500
Analysis: Results	198/200
Analysis: Total	2064/2200
Critical reflection	416/500

References

[1] Crimmins, E.M. (2015) *Lifespan and Healthspan: Past, present, and promise*, *The Gerontologist*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4861644/ (Accessed: 06 January 2024).

[2] Multiple (2018) *Global, regional, and national age-sex-specific mortality ... - the lancet*, *The Lancet*. Available at: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31891-9/fulltext (Accessed: 06 January 2024).

[3] Wang, J., Qiu, Y. and Zhu, X. (2023) *Trends of mental health care utilization among US adults from 1999 to 2018 - BMC psychiatry*, *BioMed Central*. Available at: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-023-05156-2 (Accessed: 06 January 2024).

[4] Dattani, S. *et al.* (2023) *Life expectancy*, *Our World in Data*. Available at: https://ourworldindata.org/life-expectancy (Accessed: 06 January 2024).

[5] Tharmalingam, L. (2023) *Health and Demographics Dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/uom190346a/health-and-demographics-dataset/data (Accessed: 06 January 2024).

[6] *Download python* (no date) *Python.org*. Available at: https://www.python.org/downloads/ (Accessed: 06 January 2024).

[7] *Project jupyter* (no date) *Project Jupyter*. Available at: https://jupyter.org/ (Accessed: 06 January 2024).