Data Analytics Project

Dataset: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download

Table of Contents

[Introduction 1](#_Toc203940186)

[Master the data 2](#_Toc203940187)

[Database model ERD (foreign and primary keys) 2](#_Toc203940188)

[DB browser for SQLite 4](#_Toc203940189)

[Cleaning data 4](#_Toc203940190)

[Perform test plan 4](#_Toc203940191)

[Summary statistics (descriptive analytics) 4](#_Toc203940192)

[Regression analysis (predictive analytics) 5](#_Toc203940193)

[Python code 6](#_Toc203940194)

[Documenting the outcome 7](#_Toc203940195)

[Address, refine results & Communicate insights 8](#_Toc203940196)

[Three talking points & visualisations 8](#_Toc203940197)

[Track outcomes (summary) 11](#_Toc203940198)

[Conclusion 11](#_Toc203940199)

[References 12](#_Toc203940200)

## **Introduction**

This academic report aims to analyse the chosen dataset ‘Life Expectancy (WHO)’ and will identify and address the following primary research questions:

1. How does country GP affect life expectancy?
2. How could schooling and country status influence the average income consumption of resources?
3. What is the relationship between countries, population and the polio immunization?

The chosen dataset depicts data regarding sample populations globally and includes demographical and socioeconomical information about each country. Moreover, the significant role of this report is to recognize and present possible relationships derived from the dataset’s information about each country and how they might affect the life expectancy of each country’s overall population.

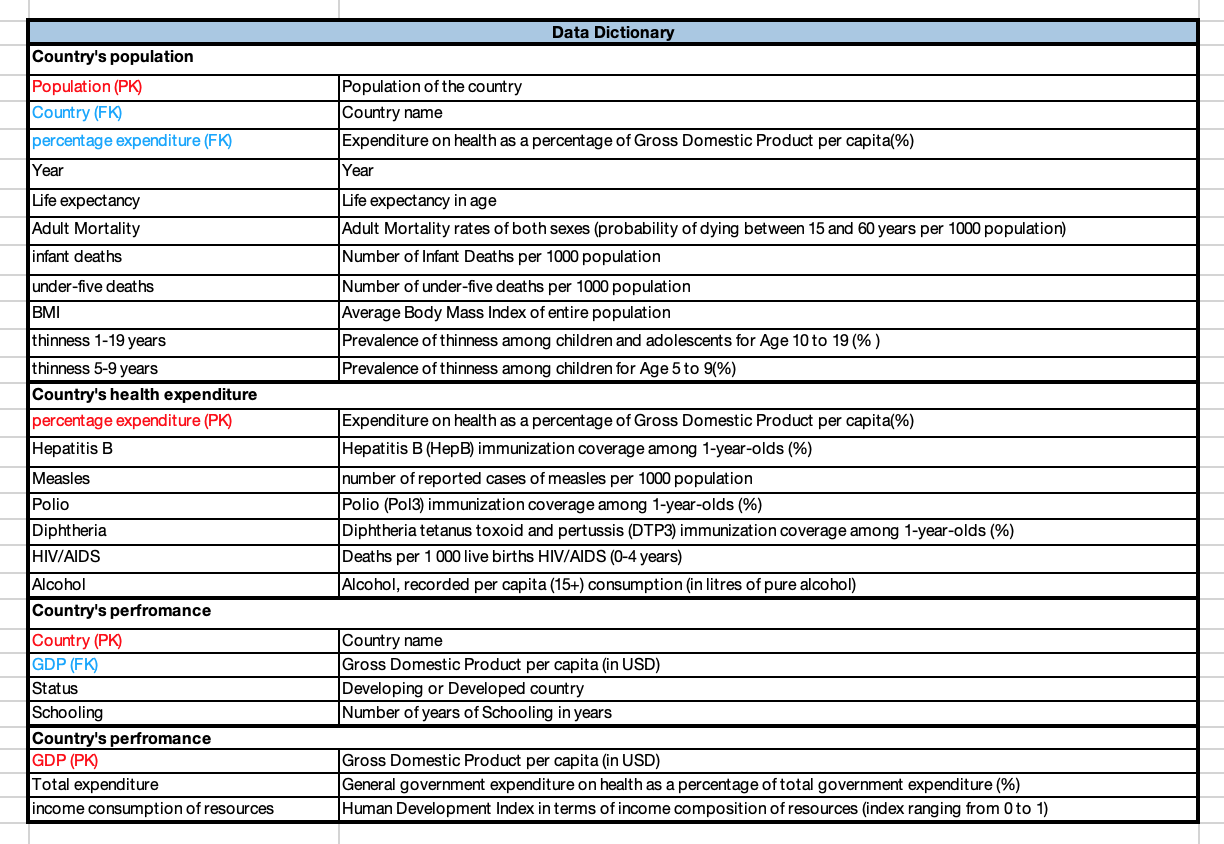
Furthermore, this report aims to also evaluate and examine the health, income, government expenditure, GDP, country status, and mortality factors included in the dataset and conclude about how these factors can impact the life expectancy of each country’s population, therefore providing an insight into the possible relationships present between the previously mentioned factors and how this information can be utilized to predict their impact on life expectancies for each population.

Lastly, this report provides descriptive statistical information about the dataset, which will illuminate readers about the data’s sample size, mean and median values for our primary variables.

## **Master the data**

### Database model ERD (foreign and primary keys)

Figure 1.1



Diagram

Description automatically generatedFigure 1.2

Figure 1.1 shows the relationships between the normalized data. It shows that the percentage expenditure of healthcare can be used to describe many different populations, this indicates a ‘one-to-many relationship’. Many-to-many relationships are also represented through the country’s population table, showing many countries and many different percentage expenditures.

### DB browser for SQLite

Figure 1.2

Here we took a normalised dataset and deformalised it in the SQL DB browser. We took the main many-to-many relationship table ‘Countrypopulation’ and joined it with the ‘Countryperfromace’ table using the Country foreign key. This step was repeated for all tables and then ran, importing the file to excel for analysing and cleaning.

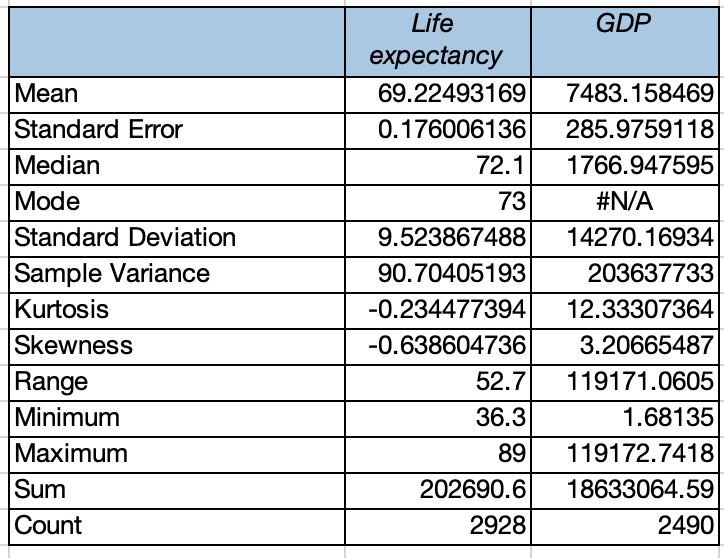
### Cleaning data

During our data analysis and interpretation, we utilised a normalised dataset and then denormalised it to produce data tables which we could draw conclusions from. We also derived interconnected relationships between the data. Our process included selecting a specific year for each data row, to avoid country name repetition. At first, we chose the most recent year, ‘2015’, however this year had many blank values and insufficient data, therefore we chose ‘2014’. Furthermore, whilst cleaning up our data, we also had to remove rows with missing values, and ignore modes of 0 when selecting our Primary and Foreign Keys. By doing this, we had our desired data points ready for data analysis.

## **Perform test plan**

### Summary statistics (descriptive analytics)

Figure 2.1



Descriptive statistics are crucial tools that can help summarize several key characteristics of a dataset and provide information about the dataset’s measures of central tendency such as mean, median, and mode. Furthermore, they can also provide measures of variability such as minimum and maximum variables and standard deviation (Hayes, 2023).

In Figure 2.1, it can be seen that the ‘Life Expectancy’ variable has a mean of 69.22, meaning the total sum of values divided by the number of observations produces a mean of 69.22, whereas GDP has a mean of 7,483.15. Moreover, we can derive from the table above that the median for the ‘Life Expectancy’ variable is 72.1 whereas the media for the ‘GDP’ variable is 1766.94, showing that the middle values for each variable in the dataset are 72.1 and 1766.94, respectively. The mode for the ‘Life Expectancy’ variable is 73, showing that the most frequently seen data point for ‘Life Expectancy’ is 73, whereas the mode for the ‘GDP’ variable is not available, showing that values are not repeated for the variable’s values, and therefore make the variable a suitable choice for primary key selection.

Additionally, the standard deviation shows how disseminated the data is for each variable. The standard deviation for ’Life Expectancy’ is 9.52 whereas its mean is 69.22, showing that they are not close and there proving that the data point is an outlier. Similarly, the standard deviation for ‘GDP’ is 14,270 whereas its mean is 7,483, which is not close to the mean and therefore also showing that the data point is an outlier. In addition, it is observed that the skewness for the ‘Life Expectancy’ variable is negative showing that it is to the left, whereas the skewness for the ‘GDP’ variable is positive showing that it is to the right. Additionally, we can see the minimum values of the ‘Life expectancy’ variable which consist of the age of 36.3, whereas the maximum value is the age of 89, showing a significant difference between life expectancies across different countries.

### Regression analysis (predictive analytics)

Table

Description automatically generatedFigure 2.2

Model Information:

* Dependent variable - Life Expectancy
* X Variables - Percentage Expenditure, Income Composition of Resources, Schooling
* Intercept  - α
* Slope - β

Regression refers to a model or statistical concept which can be utilised to explain a relationship between a dependent targeted variable and a singular or plural independent variables. For example, the regression model can show how the dependent variables are affected by changes affecting the independent variables.

From this equation, the following information can be derived:

These findings also correlate with the Human Development Index (HDI), a system made by the United Nations to measure separate human development across countries and uses metrics such as schooling and economics per country to calculate life expectancies (The Investopedia Team, 2022). This significant fact strengthens the regression model’s findings even further. Whereby, if a country has a percentage expenditure on health of 72%, an income composition index of 0.68 and schooling years of 16 the corresponding life expectancy will return:

By using the above calculation, we conclude that the life expectancy would be 74, using the appropriate numerical values accordingly.

The R square shows the ‘amount of variance explained’, if returned as 1, these variables would explain 100% of the life expectancy and thus a value of 59% proves not entirely useful but shows some of its relationships, this may prove there are too many x-variables making the data less useful than intended.

The p-value shows how statistically significant the data is and that we can be more than 0.01% sure of our assumptions.

### Python code

Graphical user interface

Description automatically generatedFigure 2.3

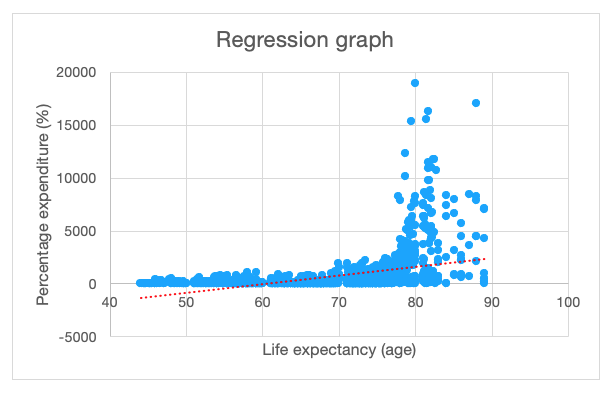
A picture containing table

Description automatically generatedFigure 2.4

We used python in google colab to establish variables, pick samples and covert between metric and categorical variables by implementing a ‘Life risk’ of high, low and mid ages and their corresponding colours and therefore ensure that the data would be colour-coded for efficient interpretation. This was imported to excel and further used in our pivot table analysis for our talking points.

### Documenting the outcome

Chart, scatter chart

Description automatically generated Figure 2.3 Figure 2.4

Chart, scatter chart

Description automatically generated Figure 2.5

These regression graphs show the relationships between life expectancy and each of the dependent variables and can also be used to predict expected outcomes (by accountants). The relationship however doesn’t imply causality.

Figure 2.3 depicts an almost static line until the life expectancy reaches 70 to where the line turns vertical which suggests poorer countries with a lower life expectancy can’t afford to allocate a lot of their expenditure to healthcare.

Figure 2.4 presents a positive slope and indicates that the number of schooling years would rise proportionality with life expectancy when showing their relationship.

Figure 2.5 presents a steeper positive slope, suggesting that an increase in life expectancy links to in a more than proportionate increase in the income consumption of resources and thus a positive correlation.

## **Address, refine results & Communicate insights**

### Three talking points & visualisations

1. How does country GP affect life expectancy?

Figure 3.1

Text

Description automatically generated with medium confidence

This represents an ‘elif’ statement to link various amounts GDP to a corresponding colour, this replaced the need for multiple variables and thus this function was stored under ‘gdpamount’. Similar python coding was used for all talking points.

Figure 3.2

Graphical user interface, text, application

Description automatically generated

We then created a pivot table to link the two variables using one categorical variable of ‘GDP risk level’ and another numerical life expectancy which we performed the average of to get an average of all countries.

Figure 3.3

Chart, pie chart

Description automatically generated

Figure 3.3 represents our produced a colour coded pie chart for this data. Our findings suggest that there is a present relationship between GDP status and life expectancy. By visualising our results, we show that low GDP, colour coded red, is correlated with the lowest average life expectancy which is 65.43. Moreover, medium GDP, colour coded amber, is correlated with the medium average of life expectancy which is 72.53. Lastly, the highest level of GDP colour coded green, is correlated with the highest average of life expectancy which is 81.52.

Most importantly, our analysis has distinguished that the higher levels of GDP per country, the higher the average life expectancy per country there is.

1. How could schooling and country status influence the average income consumption of resources?

Figure 3.4

Graphical user interface

Description automatically generated with low confidence

Figure 3.5

Graphical user interface, application

Description automatically generated

We then converted this data into pivot tables whereby it presents the relationship between the average schooling years and average of income composition of resources by showing the links to having a good, sufficient or poor education level in developed or developing countries. A second pivot table was made to further condense this data

Chart, funnel chart

Description automatically generatedFigure 3.6

The funnel chart in figure 3.6 represents these results. We can determine that the education level links closely with income consumption of resources whereby they both follow a red, amber and green trend in the same order. It highlights the relationship between how well developed a country is in relation to its education system and its income consumption of resources i.e. its indicators of HDI.

1. What is the relationship between countries, population and the polio immunization?

Figure 3.7

Table

Description automatically generated with low confidence

Figure 3.8

Graphical user interface, application

Description automatically generated

We transferred this data into a pivot table to show the relationships between the % of polio immunizations and its effect on countries and its population. Here we filtered the data to show three differing risk countries in 2014 showing how much of their average population is immunized. Argentina proves covered with nearly 100% in comparison to Lesotho with only a 9% immunization rate which may pose a risk to their population.

Figure 3.9

Map

Description automatically generated

We created a map chart to illustrate our findings for all countries. It shows that areas in Asia seem to be more covered with the immunization in comparison to a few countries in Europe being lower down on the scale. Overall Africa can be seen to have the most countries in the bottom end of the scale i.e. a small percentage of their country immunized from polio which may link to poor health expenditure.

## **Track outcomes (summary)**

### Conclusion

In conclusion, this report has evaluated ‘Title’ and examined several elements and factors that influence life expectancy across countries. Through several forms of data analysis, such as ERD models, SQLite, cleaning data, descriptive and regression analysis, Python code, and Excel Tables and Charts this report has asserted the presence of multiple relationships between variables present in the ‘Life Expectancy (WHO)’ database following the ‘IMPACT’ method and has provided answers to our primary research questions.

Through data analytics, our conclusive findings determine the presence of a positive relationship between GDP and Life Expectancy. Moreover, there is also a present trend which suggests another positive relationship between schooling and country status with average income consumption of resources. Lastly, we have also proved through our data that there is also a present positive relationship in regard to population numbers and polio immunisation %. Moreover, our study has encountered several limitations such as we came across involved a low r square in our regression which may indicate our choice of x-variables wasn’t the best suited to our y variable ‘life expectancy’ and so may of returned less useful results. Others included, various blanks In our dataset which meant a lot of countries had to be removed from our overall analysis which may not give an accurate depiction globally. Finally, variables were measured in varying values which proved difficult when choosing variables to compare throughout the project as language and the overall context had to be adjusted. In summary, through this report we have developed a deeper understanding regarding country data elements and life expectancies across global populations.

## **References**

Hayes, A. (2023) Descriptive Statistics: Definition, Overview, Types, Example. Available at: https://www.investopedia.com/terms/d/descriptive\_statistics.asp (Accessed: 30th March 2023)

Richardson, V. et al. (2020) *Data Analytics for Accounting.* 2nd Edition. ProQuest Ebook Central. ISBN: 9781260571097

The Investopedia Team (2022) *What is the Human Development Index (HDI)?.* Available at: https://www.investopedia.com/terms/h/human-development-index-hdi.asp (Accessed: 30th March 2023)