Visually Exploring a Dataset

Written Assignment on a Dataset Based on Official

World Health Organization Data

Exploratory Data Analysis and Visualization (DLBDSEDAV01)

22.07.2025

Developed by: **Alexander Stankov**

Matriculation number: 92127236

Tutor: **Prof. Dr. Visieu Lac**

Table of Contents

1. Introduction

To improve public health outcomes, policy makers, healthcare professionals, and researchers need to understand the factors that contribute to public health. One of the most widely studied, easily comprehensible and clearly signifying indicators of health status is life expectancy [1]. The current project deals with an analysis of a dataset derived from the Global Health Observatory (GHO) data repository of the World Health Organization (WHO) [2]. It aims to identify important patterns and correlations related to life expectancy and other health-important variables. Although it does not aim to dive any deeper into the dataset than to conduct a structured exploratory data analysis (EDA), this study seeks to draw insights that can form the basis for future research or even policy action.

The complete coding and raw analysis, as well as examination, not included in this report, can be found in the following GitHub repository: https://github.com/AlStankov/eda-and-visualization

2. Dataset Overview

The current project examines a dataset, which has been created using merging of several data files from the Global Health Observatory (GHO) data repository under World Health Organization (WHO). Moreover, some variables and observations have been filtered out in favor of others [2]. This dataset preparation has been conducted by researchers, which are not related to the current analysis, and will be referred to as “authors” in this section. The analysis of the current project uses the already prepared dataset.

The dataset has been obtained from the Global Health Observatory (GHO) data repository under World Health Organization (WHO). The dataset contains data related to different health factors and life expectancy. However, the dataset analyzed in this project consists only of a sample of 22 variables, which have been considered most representative by the authors of the dataset.

The same authors have observed a huge development in the health sector in the time frame 2000 - 2015, which has led to an improvement in the mortality rate, compared to the 30 years period prior that. Therefore, the dataset has been limited to the years from 2000 to 2015.

Since the dataset has been derived from the World Health Organization, it can be supposed that it does not contain any dramatically inaccurate data. During the dataset preparation, conducted by its authors, it has been identified that most of the missing data was for the variables population, Hepatitis B and GDP. Indeed, it has also been determined that the missingness was primarily typical of less-known countries such as Vanuatu, Tonga, Togo, Cabo Verde etc. Due to the high

complexity of finding this information and the inaccurateness of imputing it, those countries have been dropped from the final version of the dataset.

3. Basic EDA

Although the dataset is considered to be clean and error-free, it was examined once again. Soon, a naming error in one of the variables was identified. Among the others, there are two variables related to thinness, namely “thinness 1-19 years” and “thinness 5-9 years”. Apparently, the two variables seem overlapping and thus raise the question whether there has not been made a mistake.

After thorough examination of the variables’ descriptions in Kaggle, it was identified that the variable named “thinness 1-19 years” actually is supposed to describe the “Prevalence of thinness among children and adolescents for Age 10 to 19 (%)” (CITAT). Thus, it can be concluded that the name of the variable is a result of a typing error, and the correct name has to be “thinness 10-19 years”. Consequently, the variable name was changed.

Afterwards, all variable names were changed according to the Python convention. All letters were converted to lowercase and all spaces were replaced with underscores. However, the variables will be denoted with spaces instead of underscores in this report for clarity purposes.

The examination of datatypes did not show anything questionable or irrational. The percentage of missing values per column is shown in Fig. 1.

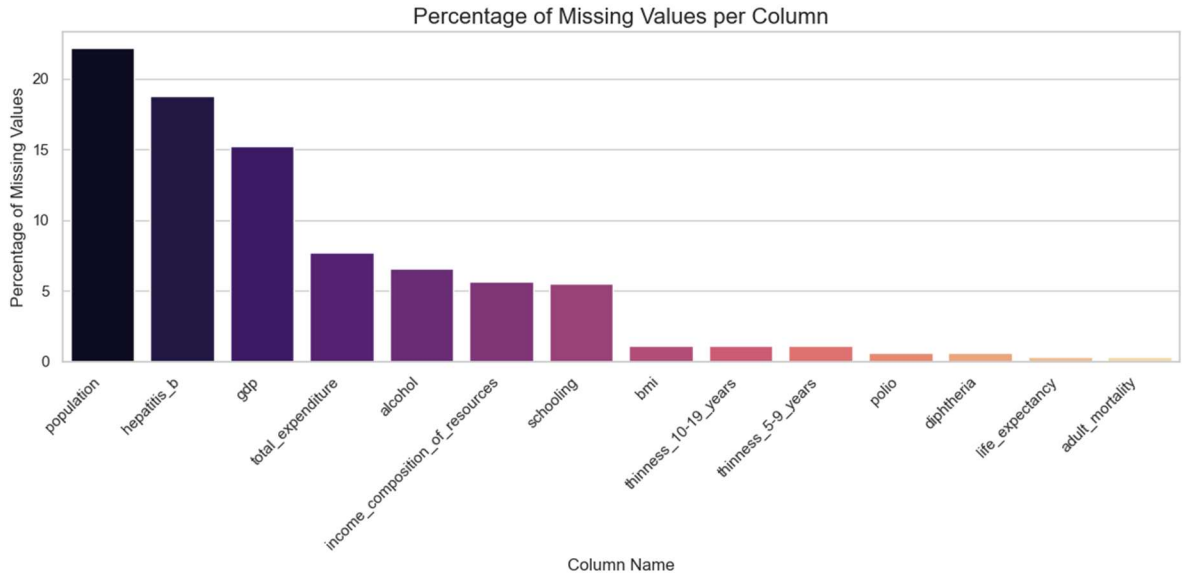

Fig. 1 Percentage of missing values per column

The only variables with more than 10% missing values are “population”, “hepatitis_b” and “gdp”. Overall, it can be concluded that the missing values are not broadly spread and are not too many.

No duplicate values were found, and no odd values were identified while examining the number of unique values per variable. At this point, the dataset was considered clean enough and the general examination of the variables as a whole was completed.

The first variable particularly examined was ‘year’. Fig. 2 shows the number of observations per year. As it can be seen, there are slightly more observations for the year 2013 than for the rest of the years, all of which have equal number of observations.

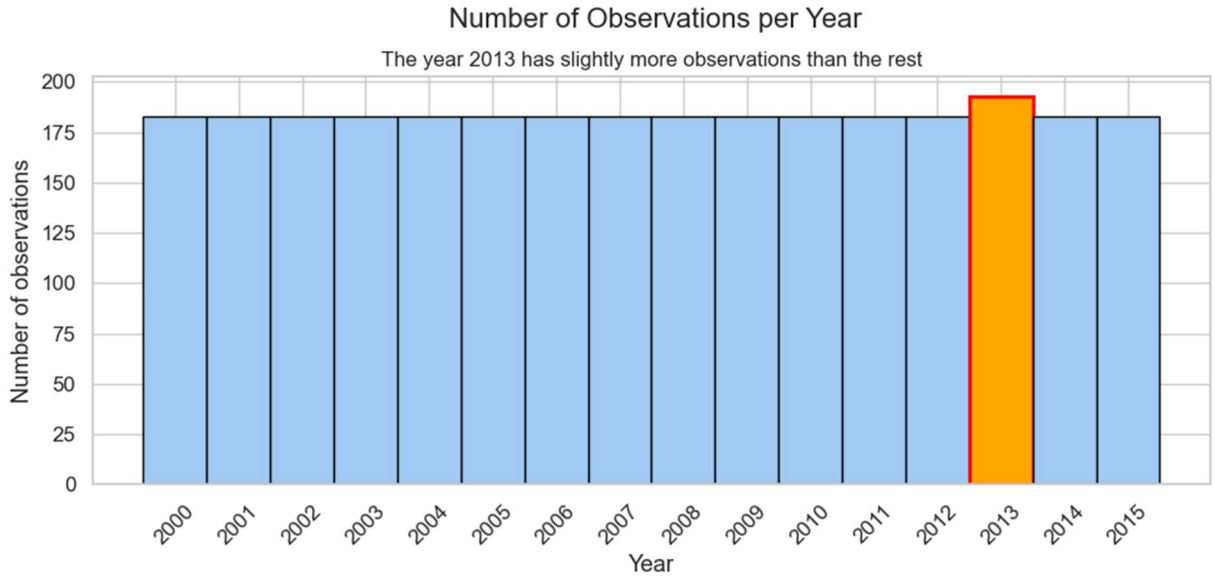

Fig. 2 Number of observations per year

The next examined variable is life expectancy. Its distribution, kernel density estimator (KDE) graph, mean value and standard deviation can be seen in Fig. 3.

The value that occurs the most is slightly above the arithmetic mean of 69 years. A standard deviation of 9.5 years shows moderately high difference in life expectancy for different countries. Additionally, the histogram shows the even values below 45 are existent, despite them being very low in count.

Although it is not displayed on Fig. 3, the trimmed mean for the life expectancy was also calculated, where the lowest and highest 10% of the values were trimmed. The result was very

close to the arithmetic mean with 70 years. The results for 15% and 20% of the values from both tails of the distribution trimmed were as well similar with 70.28 and 70.61 years respectively.

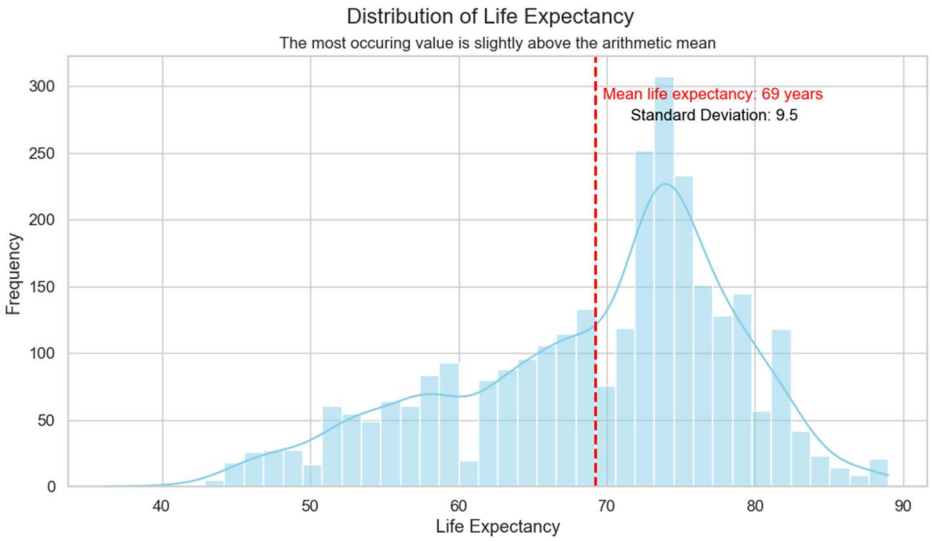

Fig. 3 Distribution of life expectancy

Similarly interesting is the variable adult mortality, which is shown with the violin plot in Fig. 4. It is worth noting that the variable shows the mortality rate per 1000 inhabitants, aged between 15 and 60. Apparently, this is a highly skewed distribution with a long right tail. Although the mean value is 144, there are extreme value of up to close to 800. Nevertheless, the 25th and 75th percentile are relatively close to the mean value, which shows that the data is distributed near the mean value.

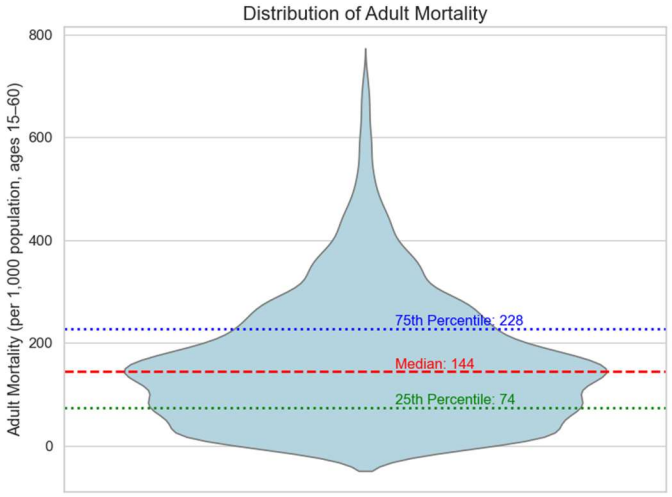

Fig. 4 Distribution of adult mortality

4. Deeper EDA

The variable life expectancy can be considered one of the most significant in the dataset in terms of its public importance. Therefore, it makes sense to check how it is correlated with the other dataset features. It should be noted though that the variables year, country and status have been filtered out of this examination due to their non-numeric nature. Any type of encoding of these variables would spread out beyond the scope of the project, so this strategy was decided against.

The correlation coefficient used was Spearman's correlation coefficient. Pearson's and Kendall's coefficients were also considered, but were not acknowledged as suitable.

Figure 5 shows the correlation between the variable life expectancy and all other examined variables. Spearman's correlation coefficient has a value between -1 and 1, where -1 denotes a perfect negative correlation and 1 is to be interpreted as perfect positive correlation [3].

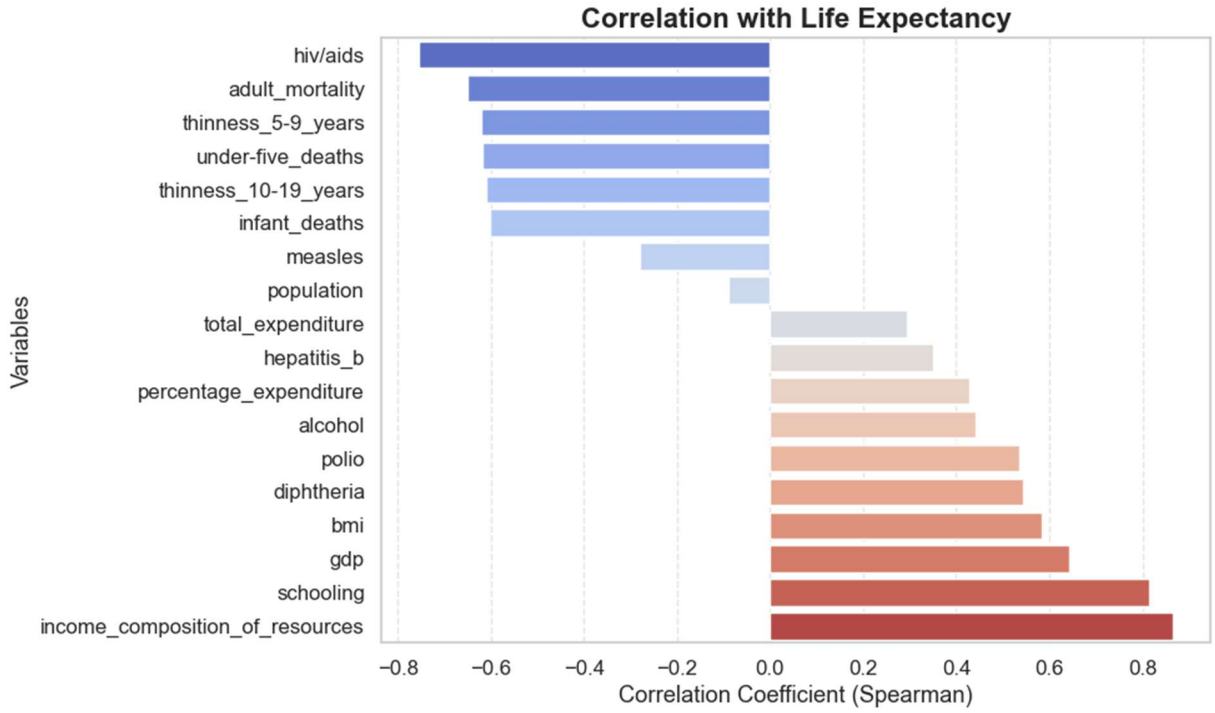

Fig. 5 Correlation of life expectancy with other variables

Overall, it can be concluded that most of the variables have a high correlation coefficient with life expectancy (above 0.5 or below -0.5), which means that they heavily influence the life

expectancy value [3]. Apparently, there are two variables with very high positive correlation, namely income composition of resources and schooling, as well as two variables with high negative correlation, namely hiv/aids and adult mortality.

Perhaps those highly correlated with life expectancy variables are correlated with each other as well. Therefore, the correlation between all variables (again, without year, country and status) was examined. Figure 6 shows the correlation coefficient between those variables, where its value is at least 0.4 or at most -0.4.

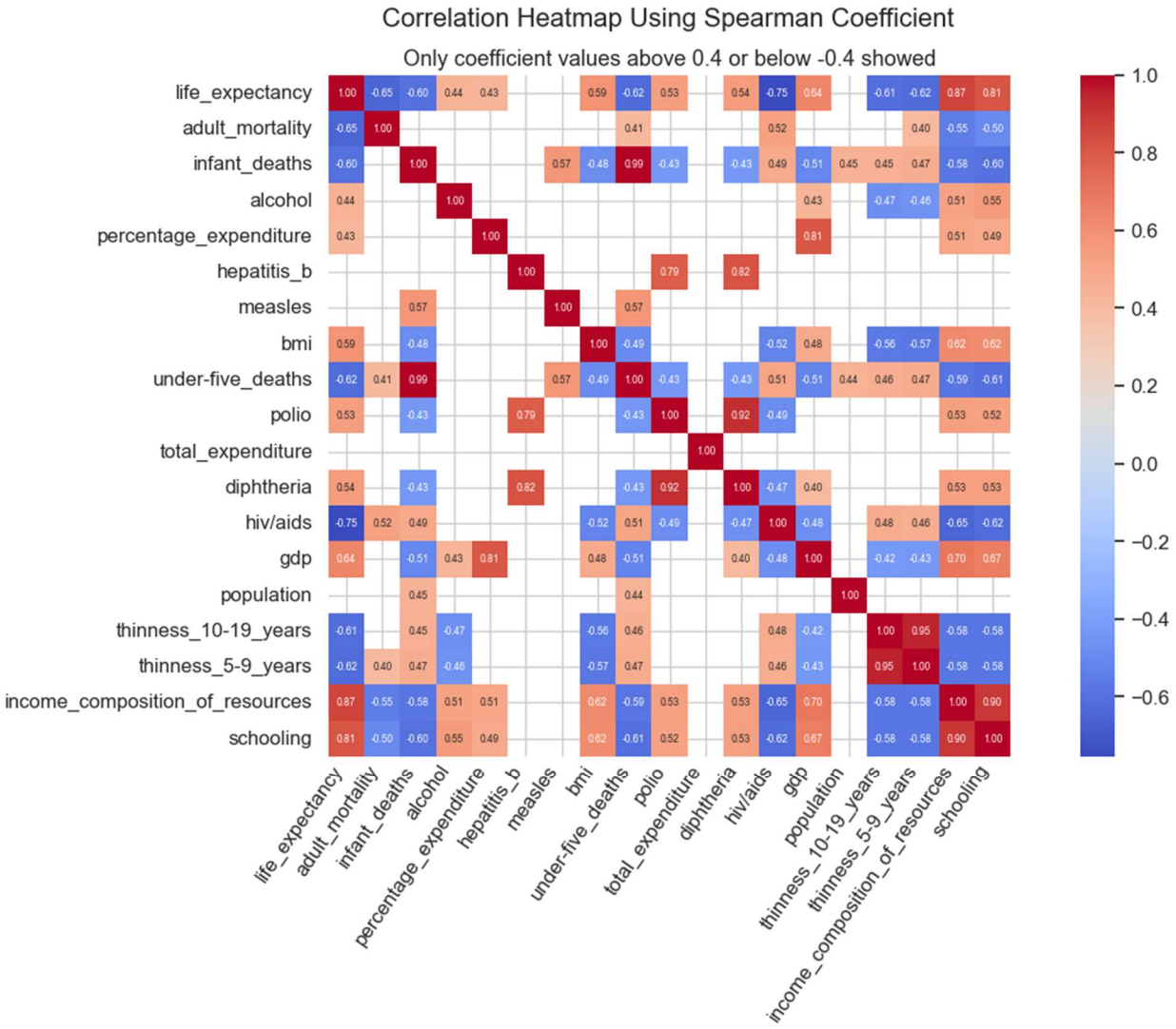

Fig. 6 Correlation between variables (only values greater than 0.4 or smaller than -0.4 showed)

Indeed, a very high correlation between schooling and income composition of resources is observed (0.90). This creates an opportunity for further exploration of the reasons behind this high correlation.

Apart from that, diphtheria and polio are also exceptionally high correlated (0.92). Although both diseases are biologically different and therefore unrelated, perhaps the factors contributing to their spread, such as the way they are transmitted and possible prevention through vaccination, are similar, if not even identical. Moreover, both diseases are highly correlated with hepatitis b as well (0.82 for diphtheria and 0.79 for polio, respectively).

Another interesting observation is the extremely high correlation between thinness 5-9 years and thinness 10-19 years (0.95). This correlation value raises the question whether the differentiation of the two age groups is accurate and meaningful. Moreover, the correlation values of both variables with the other variables from the figure above are highly similar.

Next, the high correlation between GDP (Gross Domestic Product) and percentage expenditure (0.81) shows that countries with higher GDP are more likely to spend more on health than poorer countries.

Finally, a relatively high negative correlation is observed between schooling and the variables infant deaths, under-five deaths and hiv/aids(-0.60, -0.61 and -0.62 respectively). However, we can hardly assume a causative relationship between those variables. Instead, they might be characteristics of developed and underdeveloped countries, for example. Nevertheless, the correlation values still leave room for further analysis.

5. Comparison with Other Studies on the Same Dataset

One of the intriguing studies on the same dataset has been performed by a user named Karina [4]. They have also begun their study with an examination of the missing values. Apparently, the main goal of their study was to be able to predict the variable life expectancy. Thus, it was compared with other variables, similarly to the study presented here; still, the visualizations provided were not as aesthetical, and not very informative since some of them raise confusion – both due to inappropriate choice of a visualization and missing guiding comments and/or subtitle. More importantly, they compare the variable life expectancy with other variables separately and do not use any correlation metrics, which creates an impression of lacking a greater perspective of the entire dataset. In the current study presented, a more general

approach was chosen, in which the correlation between the variable and all other meaningful variables has been simultaneously shown and compared.

Another study on the same dataset has been conducted by a user named Bianca Boykin [5]. The variable of highest interest for them was again the life expectancy. They begin their study with a removal of the outliers, whereby an outlier is supposed to be a value, determined by calculations based on the interquartile range (a more detailed discussion of the approach from a technical perspective would go beyond the scope of this section). However, this approach creates a risk of removing non-outlier values, which are actually to be considered anomalies, not outliers. For example, although it was not presented in the current final report, the study conducted during the current project found extreme values of 0 for the variable income composition of resources. However, after deeper examination, they were found to be meaningful and thus were not considered outliers. The analysis of this variable can be found in the code repository provided in the Introduction.

6. Conclusion and Future Perspectives

Although the current project does not aim to dive deeper into the dataset than was shown in this report, the intriguing findings presented here may serve as an inspiration for further analysis of the dataset. One area for further development is the distribution of the variable life expectancy. It is worth examining what factors contribute the most in the countries with the lowest life expectancy, so that they can be treated with priority by the relevant authorities. Another direction for further development are the diseases discussed in section 4. It would be important to examine which factors do the at first unrelated diseases have in common, so that their spread can be better controlled and preferably stopped. Finally, although a causative relationship between the variable schooling and the variables infant deaths, under-five deaths and hiv/aids was considered implausible, it is worth examining what is this negative correlation a sign of. Perhaps it shows a trend, which can only be determined by deeper analysis.

The patterns observed in this project, particularly those involving life expectancy, can form the basis for predictive modeling with machine learning techniques. Multivariate regression or clustering analysis are both possible further steps. Of course, they should be preceded by additional refinement of the dataset, such as treating the missing values in an appropriate manner. Overall, the dataset has the potential to provide future researchers with crucial new insights into the global health perspective.

7. References

1. Robine, J. M., & Ritchie, K. (1991). Healthy life expectancy: evaluation of global indicator of change in population health. *British medical journal*, *302*(6774), 457-460.
2. Kumar, A. (n.d.). *Life expectancy (WHO)* [Data set]. Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data
3. Ali Abd Al-Hameed, K. (2022). Spearman's correlation coefficient in statistical analysis. *International Journal of Nonlinear Analysis and Applications*, *13*(1), 3249-3255.
4. Ohmammamia. (n.d.). *EDA & analysis of life expectancy dataset* [Notebook]. Kaggle. https://www.kaggle.com/code/ohmammamia/eda-analysis-of-life-expectancy-dataset/notebook
5. Boykin, B. (n.d.). *WHO life expectancy* [Notebook]. Kaggle. https://www.kaggle.com/code/biancaboykin/who-life-expectancy