Life Expectancy Trends from 2000-2015

Final Project IST 652

Daylin Hernandez, Jessica Krumm, & Zimra Panitz

Introduction

Lifespan is a common trait among all people yet varies vastly, reflecting the average number of years a newborn is expected to live. Disparities in life expectancy have been observed across different parts of the world. While some countries demonstrated substantial improvements in life expectancy, other countries continue to struggle with low lifespan. However, the life expectancy of an individual cannot be determined solely by their location, gender, or genetic makeup. Several factors can influence a person's lifespan causing it to fluctuate. Identifying these factors is the primary step in understanding how to achieve a longer lifespan.

Therefore, the purpose of this project is to identify key factors and patterns that impact a person's lifespan, and understand which countries have the highest life expectancy. This study will involve the analysis of three data sources addressing the life expectancy and the factors that could influence a person's lifespan.

The focus of the data analysis will be to answer the following questions:

1. What factors have a bigger impact on life expectancy from 2000 to 2015?
2. Which countries present the longest and shortest lifespan during this period?

About the data

To accomplish this task, two datasets from Kaggle were used. The datasets can be accessed using the links below:

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

https://www.kaggle.com/datasets/vrec99/life-expectancy-2000-2015?select=Life_Expectancy_00_15.csv

The first dataset, Life Expectancy (WHO), offers an opportunity to look at health data, economic information, and demographic indicators for different countries so these influencing factors can be understood. This dataset includes 2938 entries with 22 columns for the following attributes: Country, Year, Life expectancy, Adult Mortality, Infant Deaths, Alcohol consumption, Percentage Expenditure, Hepatitis B immunization, Measles cases, BMI, Under-five deaths, Polio immunization, Total expenditure, Diphtheria immunization, HIV/AIDS, GDP, Population, Thinness (10-19 years and 5-9 years), Income composition of resources, and Schooling.

The second dataset, Life Expectancy 2000-2015, is more focused on life expectancy in the context of environmental and economic factors containing 1904 entries with 17 column for the following attributes: Country, year, population, CO2 emission, electric power consumptions, forest area, GDP per capita, Military expenditure, Obesity among adults, drinking water services, People practicing open defecation, Individuals using the Internet, Beer consumption per capita, Health expenditure, and Life Expectancy.

Both datasets contain data ranging from 2000 to 2015. By combining the two datasets for the analysis, there is a better chance of covering the main factors impacting the lifespan. Data integrity is a vital factor when running an analysis but not all datasets will contain information that correctly represents the subject.

The final source for data is a semistructured text data collected from an article found online. This article is titled ‘Exploring the Factors that Affect Human Longevity’ while the first two data sets are structured, tabular data. Using unstructured text data offers an opportunity to collect information about the article that could not be represented in a structured, tabular manner. This article lists various factors that can predict the life expectancy of an individual. Since the data structure of the third data source is different than the other two datasets, the analysis used on it will also differ.

Preprocessing Steps

The first step in the analysis is to clean the data and ensure it is working to avoid errors or faulty results later. The first step is loading the datasets by calling the read_csv function in Python. To understand the necessary steps to clean the data, the head() function is called on the raw data. This gives a brief preview of the first 5 rows of the data. While Kaggle datasets usually contain quality data, the preprocessing differs for everyone depending on the goal the analysis aims to achieve. Next, all trailing spaces are removed using strip(). After the white space is removed, the data is evaluated for null values. Because the dataset is large and not all variables can be replaced, the null value treatment selected is to drop the null values using dropna().

As previously mentioned, there is a wide variety of data included in these datasets. For this reason, there is a variety of data types in the set as well. To run an effective analysis, categorical variables are converted to numeric variables (dummy variables) where a number represents a category. In addition to converting the categorical, the numeric variables are scaled. This puts all the data on the same scale so when the results are being interpreted, there is no misleading information. There are two different datasets being evaluated in this analysis so scaling the variables gives a more cohesive result when interpreting the output. For this specific task, creating a dataframe separate from the raw data allowed the scaling and conversion to be done without losing the information in the raw data.

The text data also needs to be preprocessed and prepared for analysis. The first step is defining the URL to the article and then pulling and parsing the text. Parsing the data gives the normally unstructured data some sort of structure for easier handling and interpretation. After the HTML has been parsed, web scraping is performed to extract the data. This function creates a string of all the words in the document. Next, stopwords are defined to eliminate words that do not add value to the document. This includes things like conjunctions and articles. Since the ‘nltk’ package was creating issues when attempting to use predetermined stopwords, the list of words was hard coded and defined. Then, the punctuation is removed from the data. Removing the punctuation will make aggregation easier so there are no repeats of words just because one was tokenized with a punctuation mark.

The article used contained a URL for source citing at the bottom. For this analysis, the link is not necessary and is removed using re.sub(). In addition to removing the URL, re.sub() was also used to remove non-alphanumeric characters and the tag that was present in the raw document. For better results, the data is all formatted to be lower case using lower(). Finally, the leading and trailing white spaces are eliminated. These steps give a cleaner version of the raw data, but there are some steps taken to make it compatible with different analyses.

Analysis of Structured Data

Now that the data is preprocessed, the next step is performing exploratory data analysis (EDA) to get to know the data domain better. To avoid hard coding for both datasets, functions were defined and executed one after the other. Creating reusable functions saves time in the future and reduces errors one may encounter when hard coding. To get an understanding of the spread of the data now that it has been scaled, a histogram is used (Figure 1). The first histogram of the data shows how often each number representing life expectancy appears. The bars in the histogram show the data before it has been smoothed. The line shows distribution after smoothing the data. The data does not quite have normal distribution. It is left-skewed which is likely due to the nature of the dataset. Normalizing the data may make it more symmetrical but data symmetry does not impact this particular analysis. The second dataset is also left-skewed despite the difference in sources.

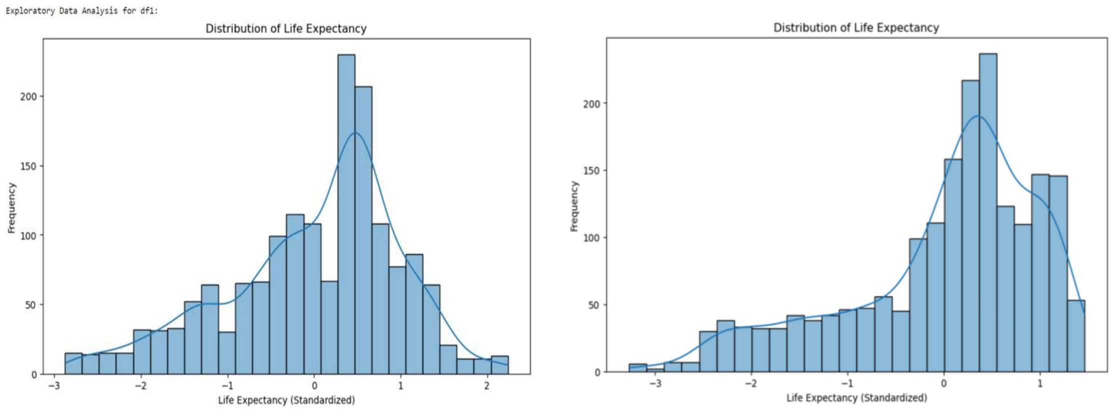
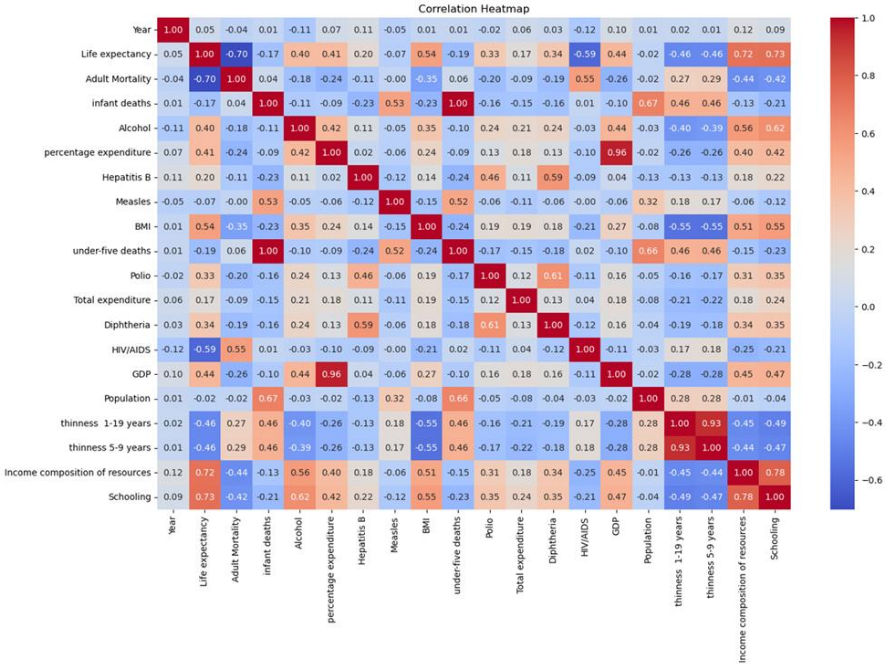


Figure 1: Distribution of Life Expectancy for structured datasets 1 and 2.

The next step taken was creating a correlation matrix. For better interpretation, the matrix was used to create a correlation heatmap (Figure 2). The correlation matrix helps identify when variables are influencing one another. Some analysis processes assume variable independence, so identifying possible multicollinearity is vital for accurate results. The correlation matrix ranks variables on a scale of -1 to 1. The farther from 0, the stronger the correlation. If the value is negative, it indicates a negative correlation, and positive values indicate positive correlation.



	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
Year	1.00	0.05	-0.04	0.01	-0.11	0.07	0.11	-0.05	0.01	0.01	-0.02	0.06	0.03	-0.12	0.10	0.01	0.02	0.01	0.12	0.09
Life expectancy	0.05	1.00	-0.70	-0.17	0.40	0.41	0.20	-0.07	0.54	-0.19	0.33	0.17	0.34	-0.59	0.44	-0.02	-0.46	-0.46	0.72	0.73
Adult Mortality	-0.04	-0.70	1.00	0.04	-0.18	-0.24	-0.11	-0.00	-0.35	0.06	-0.20	-0.09	-0.19	0.55	-0.26	-0.02	0.27	0.29	-0.44	-0.42
infant deaths	0.01	-0.17	0.04	1.00	-0.11	-0.09	-0.23	0.53	-0.23	1.00	-0.16	-0.15	-0.16	0.01	-0.10	0.67	0.46	0.46	-0.13	-0.21
Alcohol	-0.11	0.40	-0.18	-0.11	1.00	0.42	0.11	-0.05	0.35	-0.10	0.24	0.21	0.24	-0.03	0.44	-0.03	-0.40	-0.39	0.56	0.62
percentage expenditure	0.07	0.41	-0.24	-0.09	0.42	1.00	0.02	-0.06	0.24	-0.09	0.13	0.18	0.13	-0.10	0.96	-0.02	-0.26	-0.26	0.40	0.42
Hepatitis B	0.11	0.20	-0.11	-0.23	0.11	0.02	1.00	-0.12	0.14	-0.24	0.46	0.11	0.59	-0.09	0.04	-0.13	-0.13	-0.13	0.18	0.22
Measles	-0.05	-0.07	-0.00	0.53	-0.05	-0.06	-0.12	1.00	-0.15	0.52	-0.06	-0.11	-0.06	-0.00	-0.06	0.32	0.18	0.17	-0.06	-0.12
BMI	0.01	0.54	-0.35	-0.23	0.35	0.24	0.14	-0.15	1.00	-0.24	0.19	0.19	0.18	-0.21	0.27	-0.08	-0.55	-0.55	0.51	0.55
under-five deaths	0.01	-0.19	0.06	1.00	-0.10	-0.09	-0.24	0.52	-0.24	1.00	-0.17	-0.15	-0.18	0.02	-0.10	0.66	0.46	0.46	-0.15	-0.23
Polio	-0.02	0.33	-0.20	-0.16	0.24	0.13	0.46	-0.06	0.19	-0.17	1.00	0.12	0.61	-0.11	0.16	-0.05	-0.16	-0.17	0.31	0.35
Total expenditure	0.06	0.17	-0.09	-0.15	0.21	0.18	0.11	-0.11	0.19	-0.15	0.12	1.00	0.13	0.04	0.18	-0.08	-0.21	-0.22	0.18	0.24
Diphtheria	0.03	0.34	-0.19	-0.16	0.24	0.13	0.59	-0.06	0.18	-0.18	0.61	0.13	1.00	-0.12	0.16	-0.04	-0.19	-0.18	0.34	0.35
HIV/AIDS	-0.12	-0.59	0.55	0.01	-0.03	-0.10	-0.09	-0.00	-0.21	0.02	-0.11	0.04	-0.12	1.00	-0.11	-0.03	0.17	0.18	-0.25	-0.21
GDP	0.10	0.44	-0.26	-0.10	0.44	0.96	0.04	-0.06	0.27	-0.10	0.16	0.18	0.16	-0.11	1.00	-0.02	-0.28	-0.28	0.45	0.47
Population	0.01	-0.02	-0.02	0.67	-0.03	-0.02	-0.13	0.32	-0.08	0.66	-0.05	-0.08	-0.04	-0.03	-0.02	1.00	0.28	0.28	-0.01	-0.04
thinness 1-19 years	0.02	-0.46	0.27	0.46	-0.40	-0.26	-0.13	0.18	-0.55	0.46	-0.16	-0.21	-0.19	0.17	-0.28	0.28	1.00	0.93	-0.45	-0.49
thinness 5-9 years	0.01	-0.46	0.29	0.46	-0.39	-0.26	-0.13	0.17	-0.55	0.46	-0.17	-0.22	-0.18	0.18	-0.28	0.28	0.93	1.00	-0.44	-0.47
Income composition of resources	0.12	0.72	-0.44	-0.13	0.56	0.40	0.18	-0.06	0.51	-0.15	0.31	0.18	0.34	-0.25	0.45	-0.01	-0.45	-0.44	1.00	0.78
Schooling	0.09	0.73	-0.42	-0.21	0.62	0.42	0.22	-0.12	0.55	-0.23	0.35	0.24	0.35	-0.21	0.47	-0.04	-0.49	-0.47	0.78	1.00

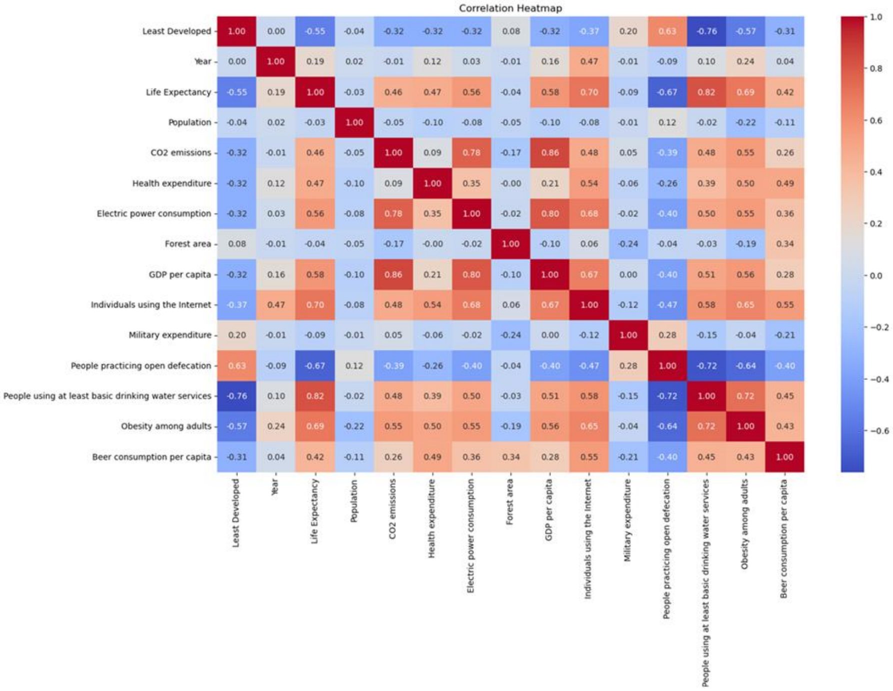


Figure 2: Correlation heatmap for structured datasets 1 and 2.

Summary statistics are another way to show the distribution of the data. The summary statistics show mean, median, standard deviation, minimum values, and maximum values. This summary is broken down by column. Since the data has been scaled, the numbers do not show the actual number of years for the life expectancy. The top 20 mean values for life expectancy and the correlating columns are then plotted on a histogram (Figure 3). This identifies which countries have the longest expected lifespan. Since both datasets include country data, a comparison between the two can be made based on which countries were in the top 20 and the associated value for life expectancy.

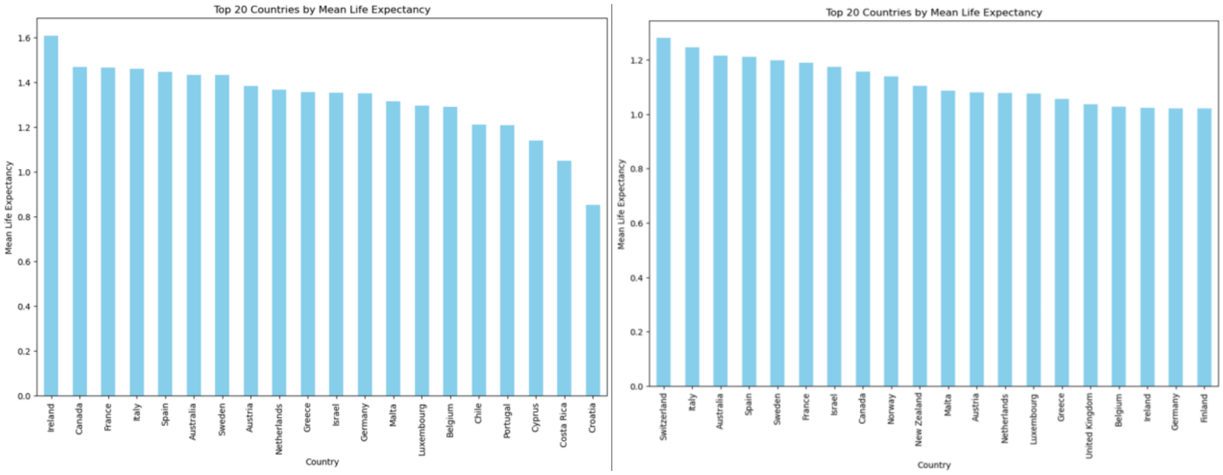
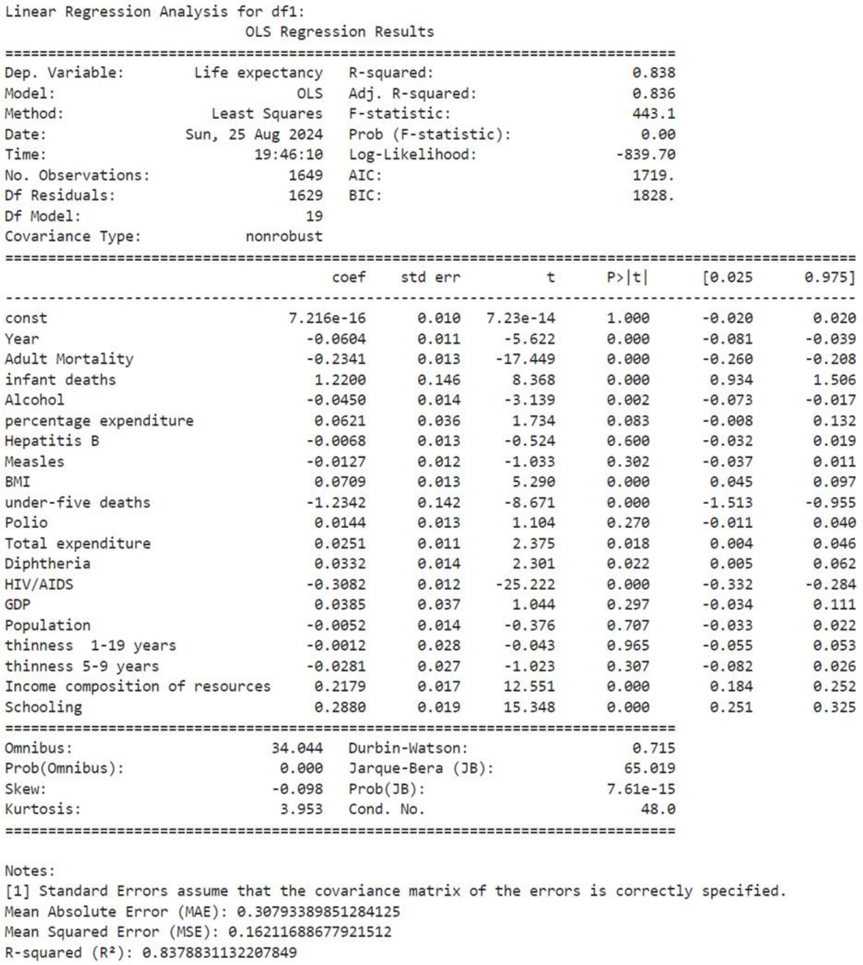


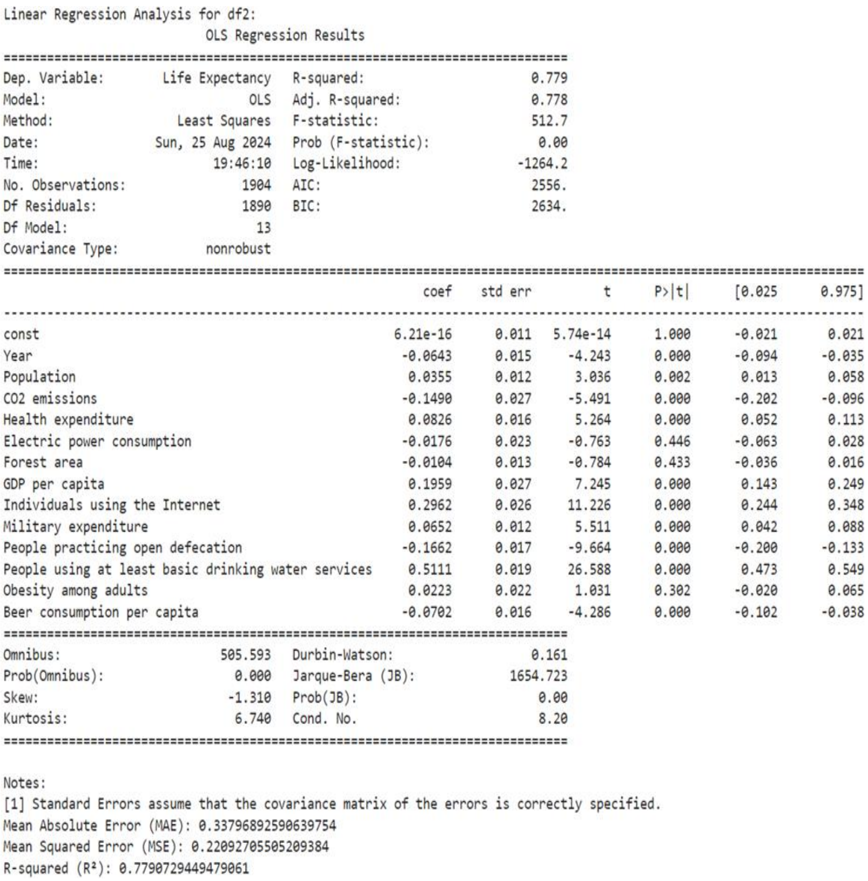
Figure 3: Histogram of top 20 countries Life Expectancy for structured datasets 1 and 2.

Then, a linear regression was run on the variables with the dependent variable being life expectancy. Linear regression is a great tool for this project because it shows if the correlation between the dependent and independent variables are statistically significant. It also produces and R-squared value that represents the percentage of change in the dependent variable that can be explained by the independent variables (Figure 4).



```
Linear Regression Analysis for df1:
                            OLS Regression Results
==============================================================================
Dep. Variable:       Life expectancy   R-squared:                       0.838
Model:                           OLS   Adj. R-squared:                  0.836
Method:                Least Squares   F-statistic:                     443.1
Date:               Sun, 25 Aug 2024   Prob (F-statistic):               0.00
Time:                       19:46:10   Log-Likelihood:                -839.70
No. Observations:               1649   AIC:                             1719.
Df Residuals:                   1629   BIC:                             1828.
Df Model:                         19
Covariance Type:           nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                            7.216e-16      0.010   7.23e-14      1.000      -0.020       0.020
Year                               -0.0604      0.011     -5.622      0.000      -0.081      -0.039
Adult Mortality                    -0.2341      0.013    -17.449      0.000      -0.260      -0.208
infant deaths                       1.2200      0.146      8.368      0.000       0.934       1.506
Alcohol                            -0.0450      0.014     -3.139      0.002      -0.073      -0.017
percentage expenditure              0.0621      0.036      1.734      0.083      -0.008       0.132
Hepatitis B                        -0.0068      0.013     -0.524      0.600      -0.032       0.019
Measles                            -0.0127      0.012     -1.033      0.302      -0.037       0.011
BMI                                 0.0709      0.013      5.290      0.000       0.045       0.097
under-five deaths                  -1.2342      0.142     -8.671      0.000      -1.513      -0.955
Polio                               0.0144      0.013      1.104      0.270      -0.011       0.040
Total expenditure                   0.0251      0.011      2.375      0.018       0.004       0.046
Diphtheria                          0.0332      0.014      2.301      0.022       0.005       0.062
HIV/AIDS                           -0.3082      0.012    -25.222      0.000      -0.332      -0.284
GDP                                 0.0385      0.037      1.044      0.297      -0.034       0.111
Population                         -0.0052      0.014     -0.376      0.707      -0.033       0.022
thinness  1-19 years               -0.0012      0.028     -0.043      0.965      -0.055       0.053
thinness 5-9 years                 -0.0281      0.027     -1.023      0.307      -0.082       0.026
Income composition of resources     0.2179      0.017     12.551      0.000       0.184       0.252
Schooling                           0.2880      0.019     15.348      0.000       0.251       0.325
==============================================================================
Omnibus:                      34.044   Durbin-Watson:                   0.715
Prob(Omnibus):                 0.000   Jarque-Bera (JB):               65.019
Skew:                         -0.098   Prob(JB):                     7.61e-15
Kurtosis:                      3.953   Cond. No.                         48.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Mean Absolute Error (MAE): 0.30793389851284125
Mean Squared Error (MSE): 0.16211688677921512
R-squared (R²): 0.8378831132207849
```

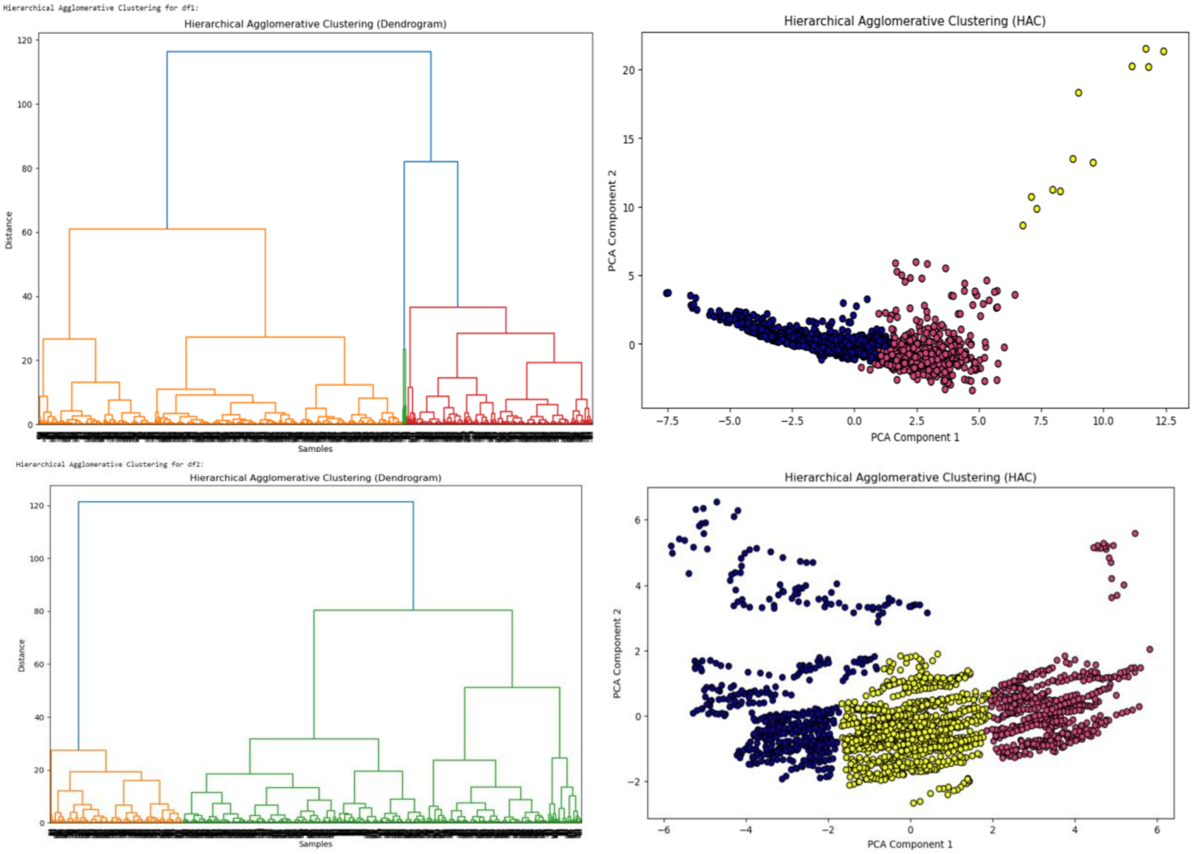


```
Linear Regression Analysis for df2:
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life Expectancy   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.778
Method:                 Least Squares   F-statistic:                     512.7
Date:                Sun, 25 Aug 2024   Prob (F-statistic):               0.00
Time:                        19:46:10   Log-Likelihood:                -1264.2
No. Observations:                1904   AIC:                             2556.
Df Residuals:                    1890   BIC:                             2634.
Df Model:                          13
Covariance Type:            nonrobust
=======================================================================================================================
                                                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------------
const                                                 6.21e-16      0.011   5.74e-14      1.000      -0.021       0.021
Year                                                   -0.0643      0.015     -4.243      0.000      -0.094      -0.035
Population                                              0.0355      0.012      3.036      0.002       0.013       0.058
CO2 emissions                                          -0.1490      0.027     -5.491      0.000      -0.202      -0.096
Health expenditure                                      0.0826      0.016      5.264      0.000       0.052       0.113
Electric power consumption                             -0.0176      0.023     -0.763      0.446      -0.063       0.028
Forest area                                            -0.0104      0.013     -0.784      0.433      -0.036       0.016
GDP per capita                                          0.1959      0.027      7.245      0.000       0.143       0.249
Individuals using the Internet                          0.2962      0.026     11.226      0.000       0.244       0.348
Military expenditure                                    0.0652      0.012      5.511      0.000       0.042       0.088
People practicing open defecation                      -0.1662      0.017     -9.664      0.000      -0.200      -0.133
People using at least basic drinking water services     0.5111      0.019     26.588      0.000       0.473       0.549
Obesity among adults                                    0.0223      0.022      1.031      0.302      -0.020       0.065
Beer consumption per capita                            -0.0702      0.016     -4.286      0.000      -0.102      -0.038
==============================================================================
Omnibus:                      505.593   Durbin-Watson:                   0.161
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1654.723
Skew:                          -1.310   Prob(JB):                         0.00
Kurtosis:                       6.740   Cond. No.                         8.20
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Mean Absolute Error (MAE): 0.33796892590639754
Mean Squared Error (MSE): 0.22092705505209384
R-squared (R²): 0.7790729449479061
```

Figure 4: Linear regression outcome for structured datasets 1 and 2.

A k-means cluster analysis is then run on the data. This process creates clusters based on data categorization. Variables are then divided into groups most closely related to the other variables in their group, and less like variables in other groups. After the clustering is completed, there is a visual representation of the output. The next process ran is a hierarchical agglomerative clustering (HAC) that puts the data in a hierarchy. It creates several 'branches' and the amount of 'branches' is up to the analyzer, unlike k-means. To represent the HAC results, they are put into a dendrogram. The dendrogram is difficult to interpret, so a scatter plot was also produced. (Figure 5)

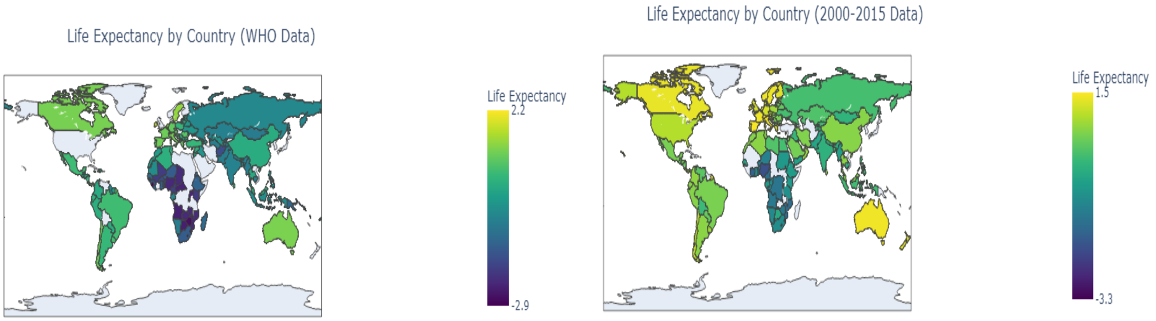


```
Hierarchical Agglomerative Clustering for df1:
```

```
Hierarchical Agglomerative Clustering for df2:
```

Figure 5: HAC outcome for structured datasets 1 and 2.

Since countries are an area of interest, a choropleth map was created with shading to represent life expectancy to further assess the behavior of life expectancy across the countries from 2000 to 2015. This gives a global overview of lifespan trends that allows for patterns to be identified (Figure 6). To focus closely on countries, the top 5 highest expected lifespan and the bottom 5 lowest expected lifespan are identified and plotted with a line graph to show the variation over the years (Figure 7).



Life Expectancy by Country (WHO Data)
Life Expectancy
2.2
-2.9
Life Expectancy by Country (2000-2015 Data)
Life Expectancy
1.5
-3.3

Figure 6: Choropleth map of Life Expectancy for structured datasets 1 and 2.

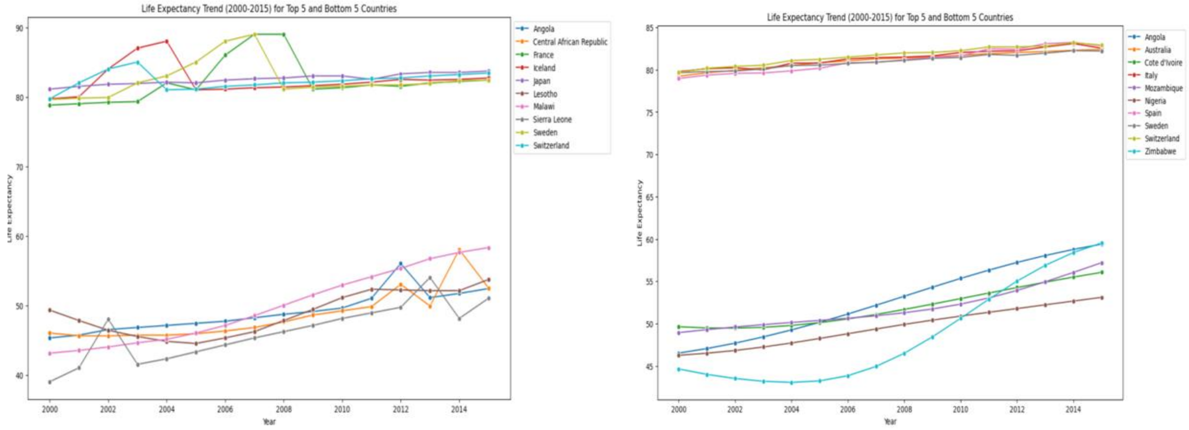


Figure 7: Life Expectancy trend lines for top and bottom five countries for structured datasets 1 and 2.

Analysis of Semistructured Data

Since the text data is an entirely different structure, the possible analyses are different than the ones used for structured data. To start, a document term matrix (DTM) is created. From this DTM a dataframe is created for better readability. Since there is now structure in the data, the repeated words can be combined and the frequency values in the rows will be summed. This is completed by using groupby() and sum(). Now there is a structured cohesive dataframe representing the data. The first step in the analysis is looking at the highest frequency words in the document. The top 20 were plotted on a bar graph and range from 3-12. The bottom 20 were also plotted, but this did not provide useful information because several values have a total of 1 (Figure 8). For a better understanding of the spread of data, iloc is used to separate out the variables that only occur once and sum them to see how many there are.

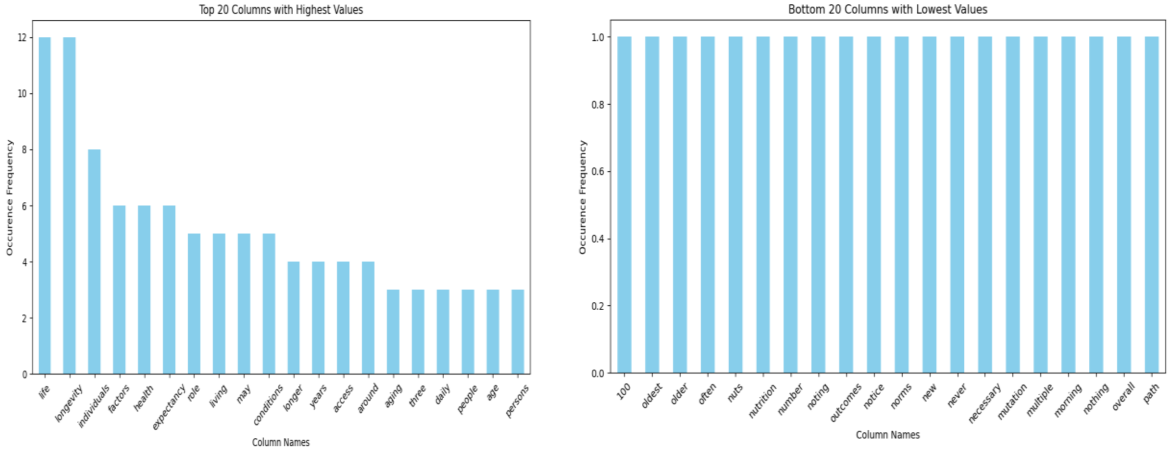


Figure 8: Frequency word distribution for semiestructured data.

Given the large number of variables that only occur once, another dataframe is created only containing variables that occur more than once. At this point, the stop words have been removed and the single occurrence values have been removed leaving 84 words. Using those words, a word cloud is created. Since a word cloud can be misleading, a sentiment analysis was run. The sentiment analysis shows polarity and subjectivity. Polarity tells if the document is negative, neutral, or positive and is ranked on a –1 to 1 scale. Subjectivity tells if the article is opinion or fact based. This is on a 0 to 1 scale, 0 being all facts, 1 being all opinions. The next step was to create a wordcloud from all the words, including the ones that only occur once to see the difference between the two (Figure 9). Since a larger dataset was used, there are more words included in the second word cloud. The final step was converting the dataframe to a term-frequency inverse document-frequency (TF-IDF). This ranks the importance of the words in the document. Since there are many variables, the TF-IDF was used to show the distribution of frequencies for the words in the document. This was visually represented with a scatter plot.

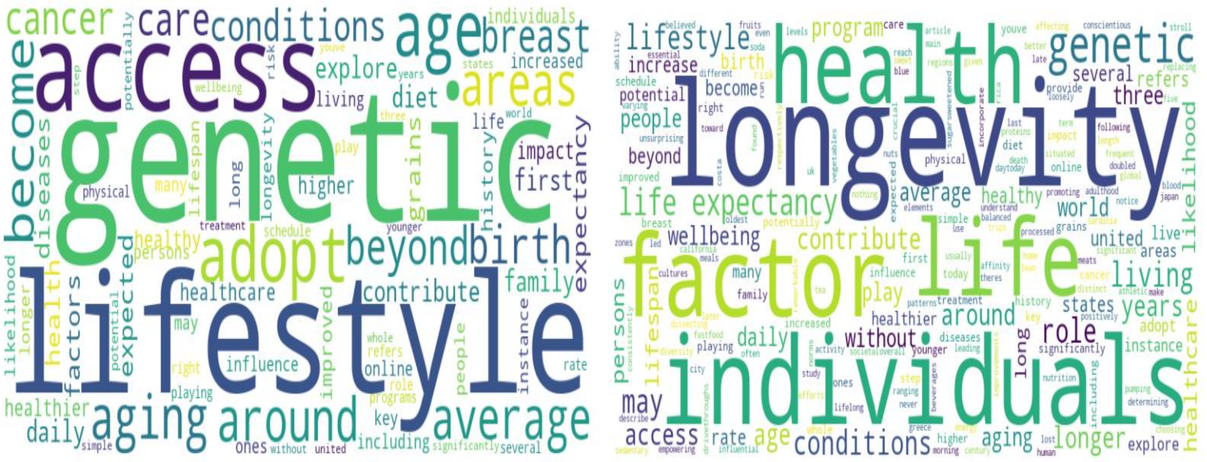


Figure 9: Wordclouds for semistructured data.

Results

Since there are three sources for data, each dataset will be individually interpreted and then put into the context of all three sources and what they mean. Looking at the WHO dataset results first, the top 5 countries with the highest mean life expectancy are Ireland, Canada, France, Italy, and Spain. However, that has not always been the case. Looking at the Life expectancy trends from 2000-2015, France, Iceland, Japan, Sweden, and Switzerland are the top 5. The bottom 5 countries with the lowest life expectancy are Sierra Leone, Central African Republic, Lesotho, Angola, and Malawi. To understand the implications of these differing results, the important variables should be considered.

Based on the linear regression results, alcohol consumption, BMI, the total expenditure on health as a percentage of GDP, increased prevalence of HIV/Aids, Diphtheria vaccinations, Income Composition of Resources (an index combining income, literacy rate, gross enrollment ratio, and life expectancy at birth), and schooling are all significant factors that impact life expectancy. The r-squared value indicates that 84% of the change in one's life expectancy can be explained by those variables. All the factors, excluding schooling, the total expenditure on health as a percentage of GDP, and Diphtheria vaccinations, have a negative correlation with life expectancy. This indicates that as those variables increase, life expectancy decreases. The two variables with positive correlation indicate that as one receives more schooling and as more people are vaccinated for Diphtheria, life expectancy increases.

Looking at the correlation matrix, a high positive correlation was observed between life expectancy and schooling, income, and BMI, while other factors such as adult mortality and HIV/AIDS were negatively correlated. Variables like the population and year did not seem to have a big impact on life expectancy.

The second dataset names Switzerland, Italy, Australia, Spain, and Sweden as the top 5 countries with highest mean life expectancy. The graph showing trends over time names the same 5 countries as the mean highest life expectancy graph. The bottom 5 with the lowest life expectancy names Zimbabwe, Nigeria, Mozambique, Cote d'Ivoire, and Angola. The linear regression names C02 emissions, health expenditure, GDP per capita, individuals using internet, people practicing open defecation, basic drinking water services, and beer consumption as significant variables. C02 emissions, people practicing open defecation, and beer consumption are all negatively correlated with life expectancy. The remaining variables are all positively correlated. The r-squared value is at 78%.

The correlation matrix shows additional details on the variables impact on life expectancy. There is a high positive correlation between life expectancy and basic drinking water services at .82. C02 Emissions and power consumption have a high positive correlation at .78. GDP per capita and C02 emissions are also positively correlated at .86. Open defecation and basic drinking water services have a high negative correlation at -.72. All these results seem intuitive. Some of them point to the country being less developed. Based on the correlation matrix, country development and life expectancy are negatively correlated at -.55. Access to clean drinking water is essential to all life so the high positive correlation is expected.

The text data results are a bit less telling than the structured data. However, the word cloud produced interesting results. Some of the important words are genetic, lifestyle, access, conditions, diseases, cancer, diet, and physical. The second word cloud produced similar results with some variation. Additional words from the second word cloud are world, wellbeing, and healthcare. The sentiment analysis determined that the article is neutral but leans towards positive. The subjectivity score indicates that there are opinions present in the article, but it leans towards factual. The top 20 most used words are not a surprise as they include life, factors, health, conditions, etc.

Conclusion

The data sources produced a myriad of meaningful results when analyzing the data individually, but when analyzed together they create a more complex picture. Sweden and Switzerland are the most commonly occurring countries in the top life expectancy, except for the mean life expectancy results from the WHO. While the specific countries differ for the bottom 5 shortest life expectancy, they are all in Africa. The factors contributing to a shorter life expectancy differ between datasets, but all follow a pattern. Almost all the variables are commonplace in underdeveloped countries. Lack of access to clean water, open defecation, literacy rate, less opportunity to attend school, and BMI are all issues underdeveloped countries struggle with. Additionally, the bottom 5 countries in Africa continue the trend as Africa has the highest number of underdeveloped countries. The word cloud showed words such as diseases, physical, lifestyle, access, and conditions.

The disparities in the results can be explained due to the accuracy of the dataset, the collection methods and measurement for life expectancy and the different factors analyzed in each case.

Given the results of the two structured datasets and the text data, all three have produced the same conclusion: location is directedly correlated to life expectancy. Countries with higher life expectancy are well developed and generally benefit from robust healthcare systems, higher income levels, and better access to education and essential services. In contrast, countries with lower life expectancy often struggle with challenges such as high disease burden, poor healthcare infrastructure, and economic instability. These results can be used to guide the country governments on taking action to extend their respective population lifespan.

Tasks by Group Members

Project tasks	Responsible member
Cleaned, prepped and develop program for dataset 1 (Structured data)	Daylin Hernandez
Cleaned, prepped and develop program for dataset 2 (Structured data)	Daylin Hernandez
Cleaned, prepped and develop program for the third data source (Semistructured data)	Jessica Krumm
Combined Jupyter notebooks with all the code.	Jessica Krumm
Power Point presentation and mockup style for the report.	Zimra Panitz
Final Report draft	Jessica Krumm
Review, comment and editing of final report and power presentation.	Daylin Hernandez, Jessica Krumm, Zimra Panitz