11.2 Assignment: Case Study

Predictors of Life Expectancy

Joi Chu-Ketterer

DSC - 550

Bellevue University

November 13, 2019

lchuketterer@my365.bellevue.edu

Instructor: Becky Deitenbeck

Case Study: Predictors of Life Expectancy

Introduction

There is no doubt that life expectancy has increased over the years. Due to advances in technology and medicines, society has been able to significantly increase the average life span. As society continues to advance in technology, now with the ability to create learning algorithms and predictive models, it would be especially valuable to understand the impacts on life expectancy. This study aims to create a successful predictive model that can be used as a baseline for future research to hopefully continue increasing the average life span.

Data Preparation

This study explores the predictors of life expectancy around the world using the *Life Expectancy (WHO)* dataset found on *Kaggle.* The dataset contains 2938 entries and 22 attributes. Many of the column data types were objects despite using numbers, so they were converted into floats. Converting them into floats will allow them to be used in predictive analytics later on. The attributes of the data are as follows:

- Country
- Year
- Status
- 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

Data Exploration

Bar plots, histograms, scatterplots, correlation matrices, parallel coordinates, and stacked bar plots were created to better understand relationships between the variables.

Bar Plots and Histograms

Bar plots were created to understand trends within the dataset, while histograms were created to understand the distribution of attributes. *Figure 1* demonstrates a relatively normal distribution for Life Expectancy, and heavily right-skewed distributions for Adult Mortality, Infant Death, and Under-Five Deaths. The skewed distributions for Adult Mortality, Infant Deaths, and Under-Five Deaths indicate that additional transformations will be needed to be applied to those attributes before they are used for modeling. This is to alleviate any extremes in the outcomes due to high variation in the data.

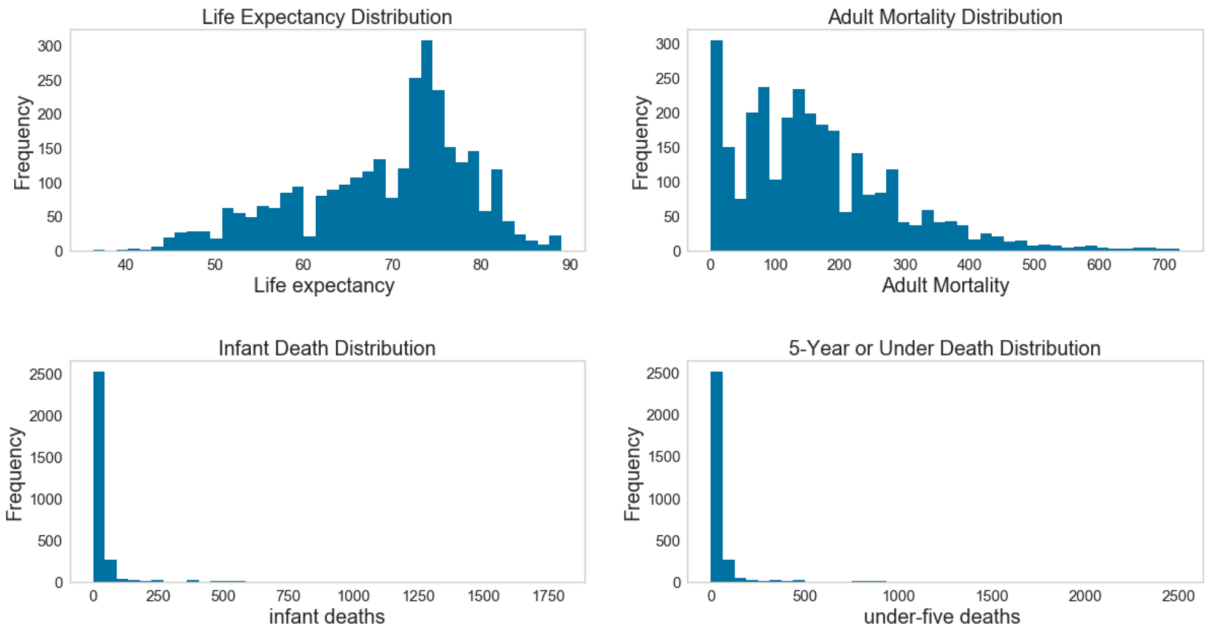


Figure 1. Histograms for Life Expectancy, Adult Mortality, Infant Deaths, and Under-Five Deaths

The bar plots in *Figure* 2 compare adult mortality rates with infant mortality rates. While they were not plotted on the same graph, it is evident far more infants die per year than adults.

Additionally, adult mortality seems to fluctuate from year to year, while infant deaths have steadily been declining.

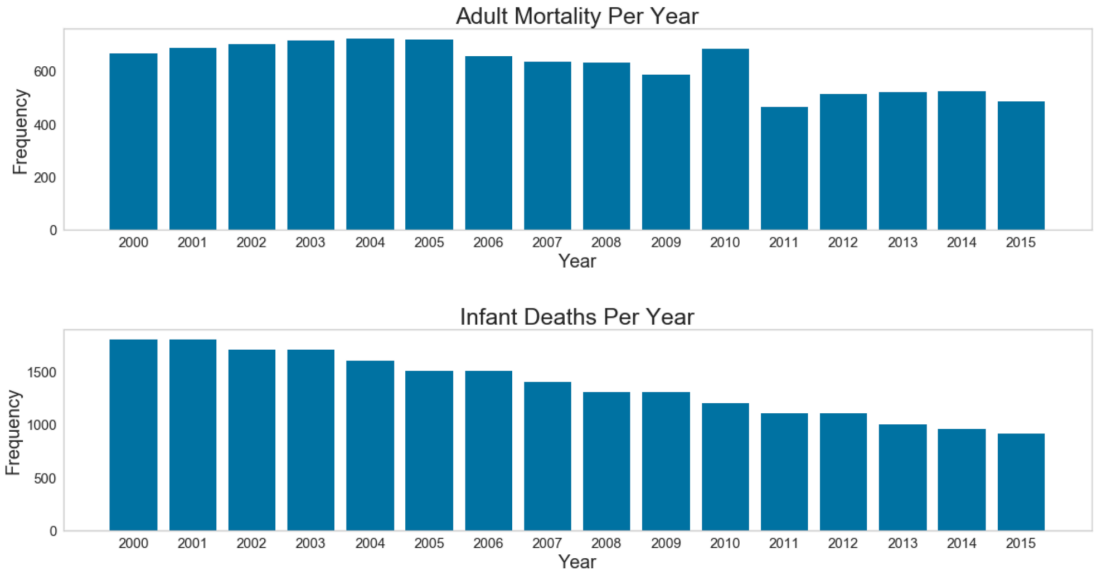


Figure 2. Bar plot comparison between Adult Mortality and Infant Deaths over the years

A second bar plot was used to compare the observation counts of developing countries and developed countries included in the dataset. *Figure 3* demonstrates the dataset has far more observations within developing countries than in developed. While no insight into the relationship of the categorical data can be made from the plot, it is useful to understand the imbalance in data collection within the dataset.

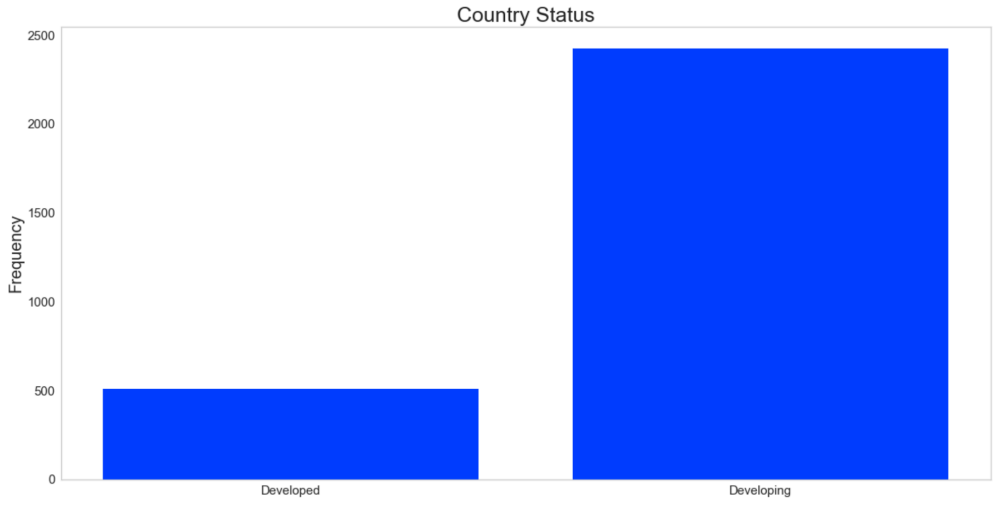


Figure 3. Comparison bar plot for life expectancy counts for developed and developing countries

Scatterplots

Scatterplots were used to gain insight into how the attributes correlate with one another. *Figure 4* shows the relationship between various attributes, showing there are no strong correlations between any of them. Although, it could be argued there is a positive correlation between Polio and Diphtheria. Despite expecting a strong correlation between Alcohol and Adult Morality, the data does not support this.

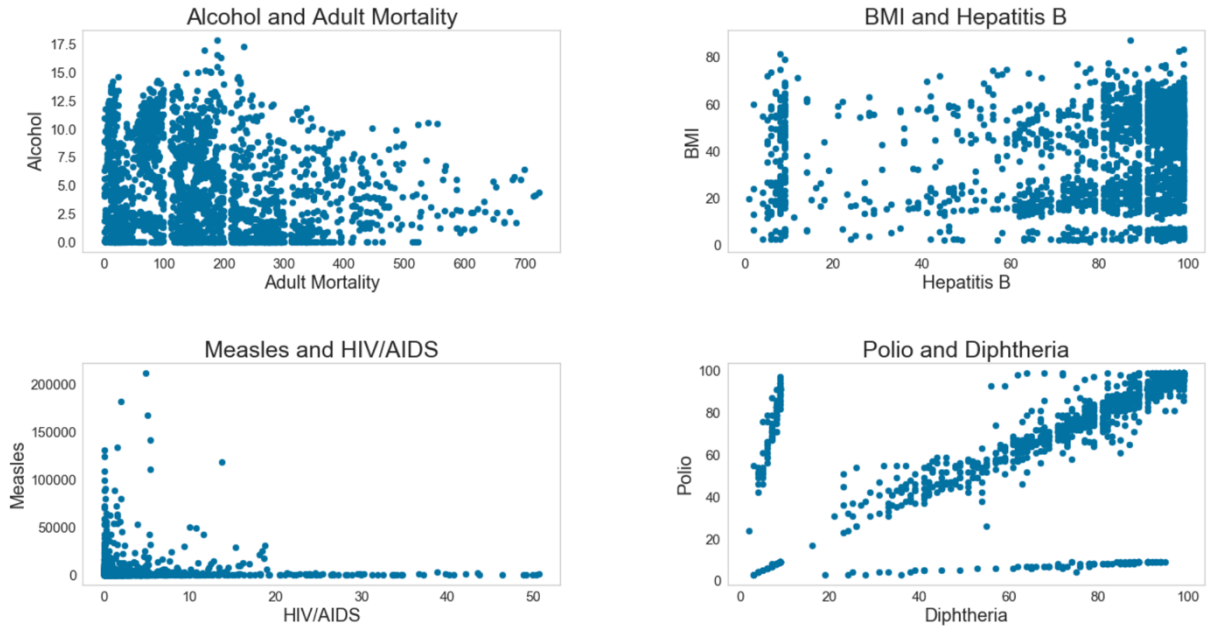


Figure 4. Scatterplots showing the relationship between several attributes

Similarly, *Figure 5* shows no correlation between Life Expectancy and various diseases. It could be argued there is a negative exponential correlation between HIV/AIDS and Life Expectancy, but it is as compelling as the positive correlation between Polio and Diphtheria.

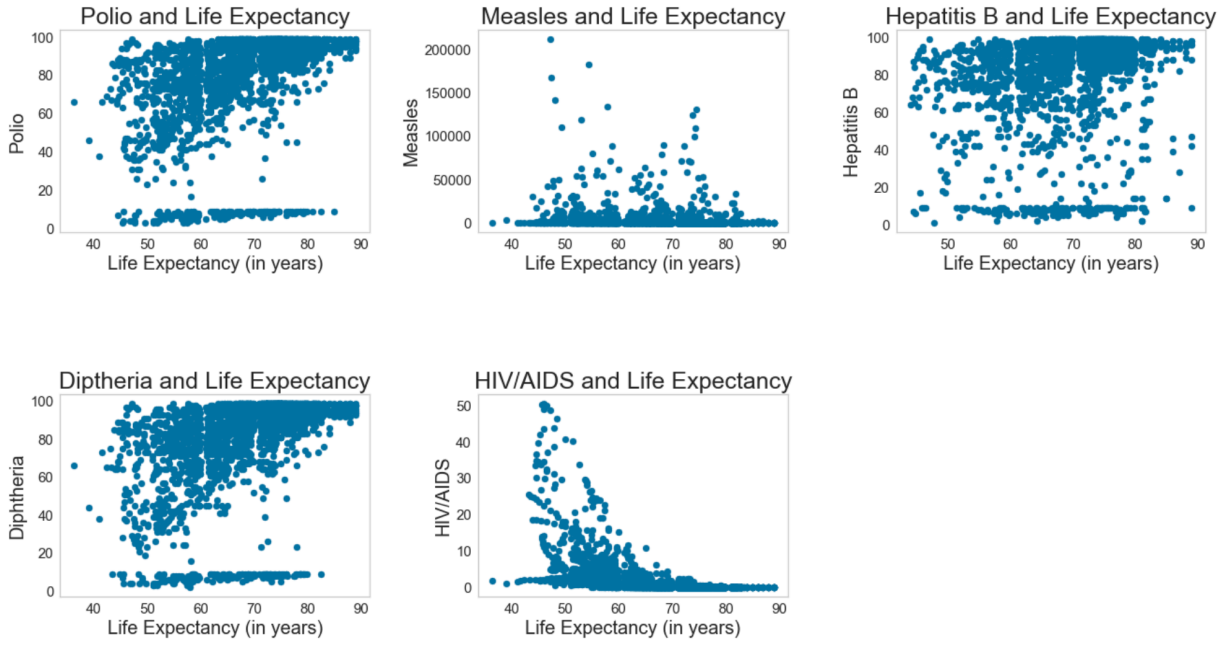


Figure 5. Scatterplots showing the relationship between Life Expectancy and diseases

Correlation Matrix

To confirm the weak correlations between the attributes, a correlation matrix featuring the diseases was created. Just as the previous scatter plots show, *Figure 6* confirms there is little to no correlation between the diseases and life expectancy. The only correlation that was detected is between HIV/AIDS and Measles and appears to have a slightly negative correlation.

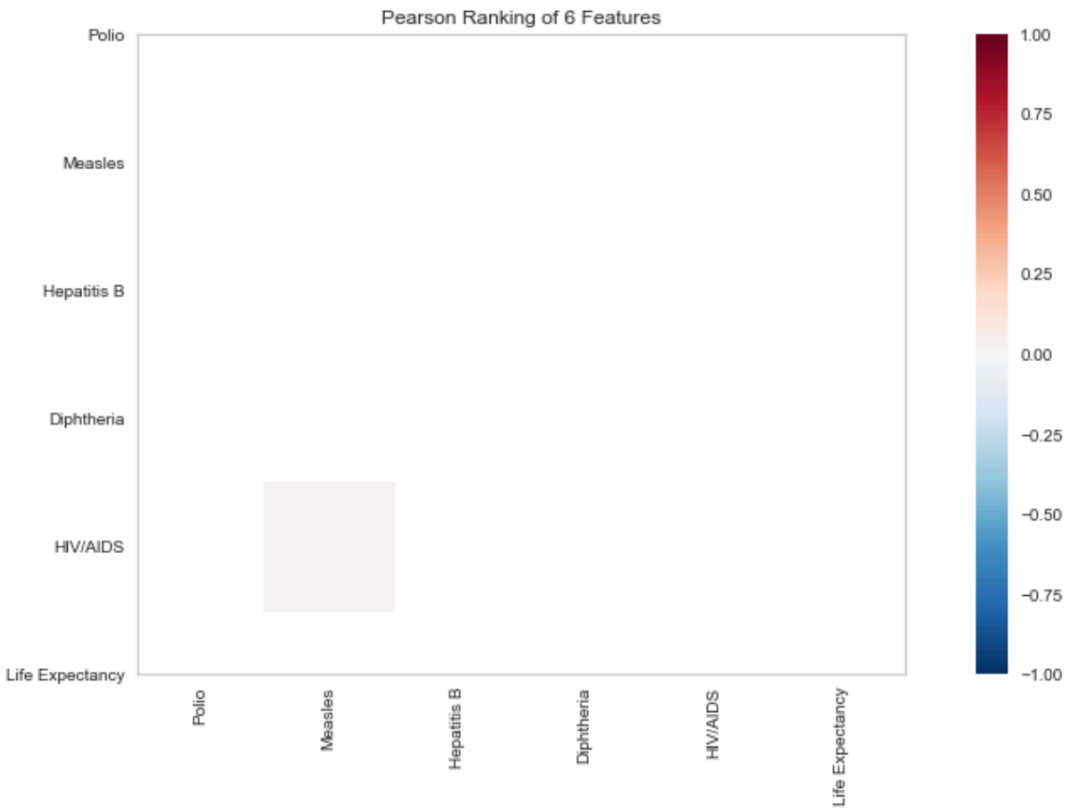


Figure 6. Correlation matrix of different diseases

Parallel Coordinates

Correlation was explored once more using a parallel coordinate graph. Using a parallel coordinate graph allows the introduction of an additional dimension. *Figure* 7 shows the correlation between diseases divided between developing countries and developed countries. Remembering *Figure 3,* there are more data entries for developing countries than developed, so the parallel coordinate graph is used to just understand the relationship rather than magnitude.

The graph shows the correlation between diseases among the two groups are generally the same. The only time they deviate is between Diphtheria and HIV/AIDS, suggesting those in developed countries contract Diphtheria more, while those in developing countries contract HIV/AIDS more.

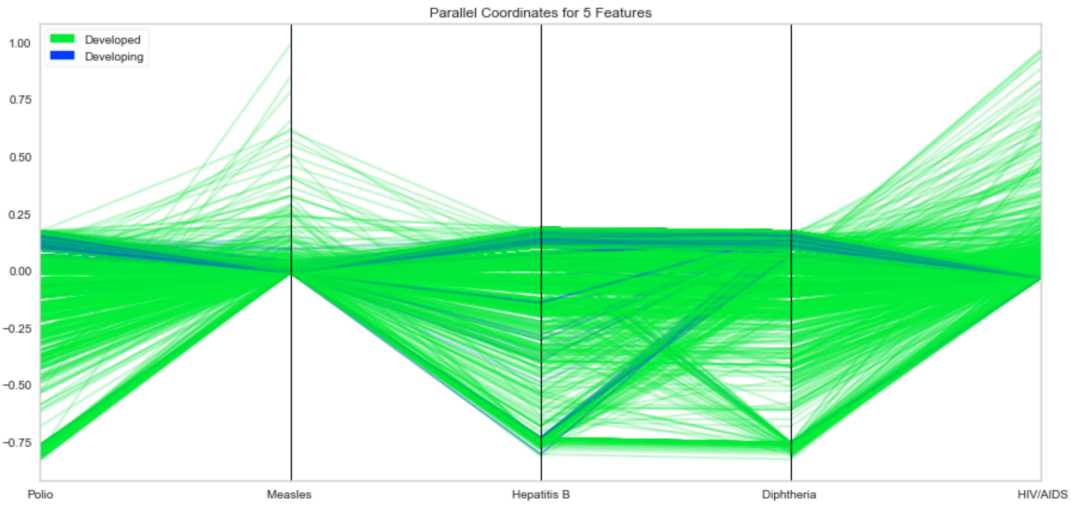


Figure 7. Parallel Coordinate graph for diseases divided into developing and developed countries

Stacked Bar Plots

Further exploration into the difference among developing and developed countries was explored using stacked bar plots. Three diseases were plotted where *Figure 8* shows far greater deaths per diseases occurred in developing countries.

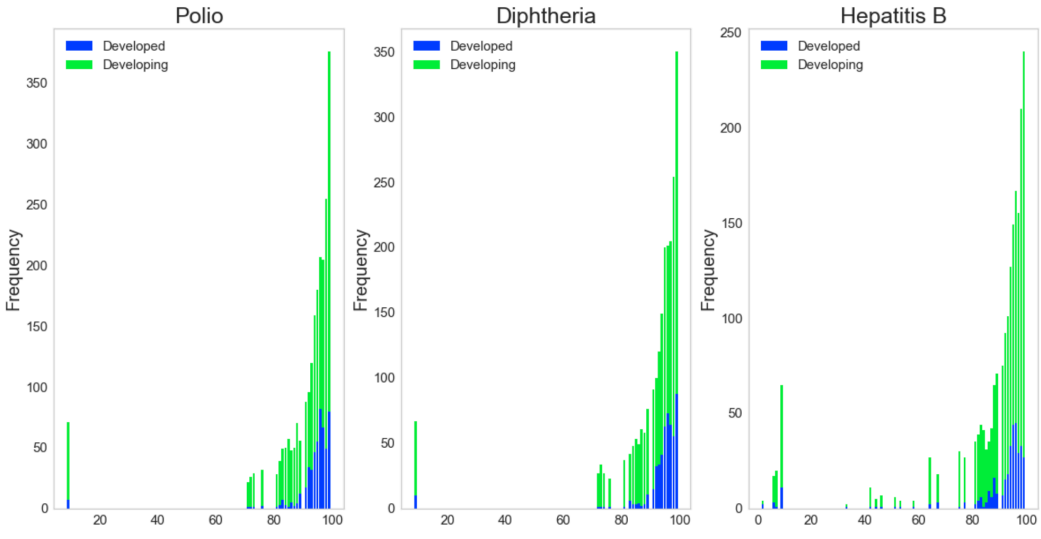


Figure 8. Stacked Bar Plots showing the distribution of Polio among developed and developing countries

Missing Data, Feature Transformation, Validation

Missing Data

Most commonly, analysts will find missing data entries in their dataset. The *Life Expectancy (WHO)* is no exception to this. During preliminary analysis, missing data entries were identified through the *isnull().sum()* method. All the categories of interest that were missing data were numerical, and so the mean of each attribute was calculated, and the resulting value replaced the null values. *Table 1* shows the difference in summary statistics from before and after the null values were replaced.

Table 1. Selective summary of attribute summary statistics displaying differences before and after missing data replacement

	Before Replacement			After Replacement			Variance (%)		
Attribute	Count	Mean	Std	Count	Mean	Std	Count	Mean	Std
Life Expectancy	2928	69.22	9.52	2938	69.23	9.05	0.34	0.01	(4.94)
Adult Mortality	2928	164.8	124.29	2938	164.73	124.09	0.34	(0.04)	(0.16)
Alcohol	2744	4.6	4.05	2938	4.55	3.92	7.07	(1.09)	(3.21)
Hepatitis B	2385	80.94	25.07	2938	83.02	23	23.19	2.57	(8.26)
BMI	2904	38.32	20.04	2938	38.38	42.04	1.17	0.16	109.78
Polio	2919	82.55	23.43	2938	82.62	23.38	0.65	0.08	(0.21)
Diphtheria	2919	82.32	23.72	2938	82.39	23.66	0.65	0.09	(0.25)

As *Table 1* suggests, the variability in summary statistics for the different attributes ranges greatly with the smallest variance having a magnitude of 0.01% and the largest having 109.78%. While taking the average value of an attribute to fill in missing data, it is certainly a risky approach. Conversely, removing all entries with any missing data could have similar repercussions.

Feature Transformation and Validation

As mentioned previously, the skewed distribution for Adult Mortality, Infant Deaths and Five-Under Deaths were heavily right-skewed calls for data transformation. As such, a log transformation was applied to each attribute. Taking the log minimizes the effects of the distribution skew. Additionally, the *Figure 1* shows a majority of the data points are small, and so to accurately account for small differences in value, the *numpy np.log1p()* method is used. *Table 2* shows the difference in summary statistics from before and after the log transformation.

Table 2. Selective summary of attribute summary statistics displaying differences before and after log transformation

	Before Log Transformation			After Log Transformation			Variance (%)		
Attribute	Mean	Std	Max	Mean	Std	Max	Mean	Std	Max
Adult Mortality	164.7	124.1	723	4.73	1.04	6.58	(97.13)	(99.16)	(99.09)
Infant Deaths	30.3	117.9	1800	1.77	1.66	7.5	(94.16)	(98.59)	(99.58)
Under-Five Deaths	42.06	160.5	2500	1.94	1.78	7.82	(95.39)	(98.89)	(99.69)

As *Table 2* suggests, the variability in summary statistics for the different attributes ranges greatly with the smallest variance having a magnitude of 94.16% and the largest having 99.69%. This means most of the values were reduced by almost ten times their value. As one can imagine, this greatly changes the histogram shapes. The difference is shown in *Figure 9*.

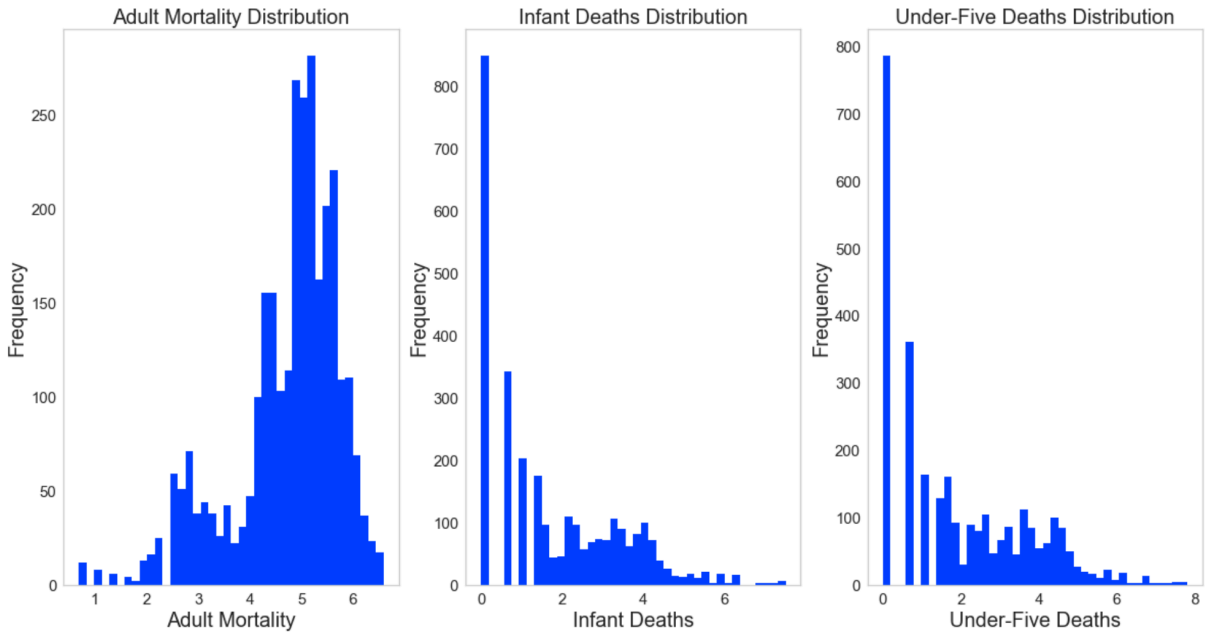


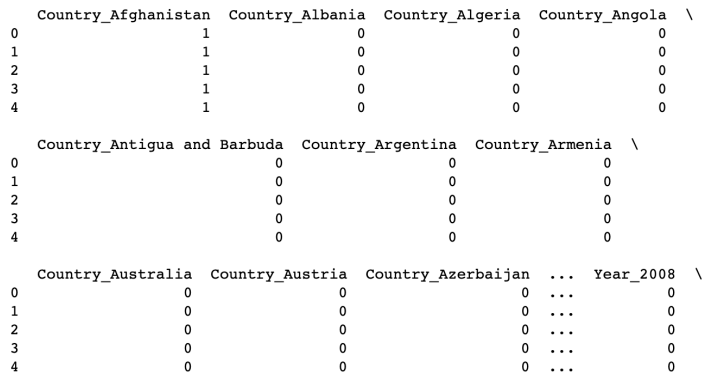
Figure 9. Histograms for Life Expectancy, Adult Mortality, Infant Deaths, and Under-Five Deaths after log transformation

A clear difference is seen between the original distributions in *Figure 1* to those in *Figure 9*. Additionally, with a smaller difference between the minimum and maximum values, the patterns of the distributions are clearer. This is especially evident for Infant Deaths and Under-Five Deaths.

One Hot Encoding

While most of the attributes within the dataset were numerical, there were some categorial data as well (developed vs. developing). Most algorithms cannot accurately handle categorical data, and so encoding methods are used to make it easier. In this study, One Hot Encoding is used to encode the categorical attributes Year, Country, and Status. Previous preliminary analytics revealed there are 193 unique countries, 16 unique years, and 2 unique statuses. While encoding the attributes, each unique entry is assigned a column. For this reason, the resulting encoded

matrix has the dimensions (2938, 211), indicating there are 2938 entries across 211 columns. For this reason, the full encoded matrix is not included in this report; however, *Figure 10* provides an example of some of the rows and columns.



	Country_Afghanistan	Country_Albania	Country_Algeria	Country_Angola	\
0	1	0	0	0	
1	1	0	0	0	
2	1	0	0	0	
3	1	0	0	0	
4	1	0	0	0	

	Country_Antigua and Barbuda	Country_Argentina	Country_Armenia	\
0	0	0	0	
1	0	0	0	
2	0	0	0	
3	0	0	0	
4	0	0	0	

	Country_Australia	Country_Austria	Country_Azerbaijan	...	Year_2008	\
0	0	0	0	...	0	
1	0	0	0	...	0	
2	0	0	0	...	0	
3	0	0	0	...	0	
4	0	0	0	...	0	

Figure 10. Screenshot of the resulting encoded matrix for the categorical attributes Status, Country, and Year

The encoded matrix is interpreted by identifying where there is a 1 present. For instance, the first five entries in *Figure 10* show 1 for Country_Afghanistan, which means in the original dataset the attribute Country would have Afghanistan as the entry.

Model Selection and Assessment

Data Understanding and Model Selection

Prior to choosing a model, the expected type of input and output data needs to be determined. For this study, the input data is both categorical and numerical. However, the categorical data was encoded into numerical data in previous analysis, so we can say all input data is numerical. Since the goal of the study is to predict life expectancy, the output data is also numerical. Thus, a linear model should be able to accurately describe and fit the dataset.

Previously, correlation graphs were created to determine the relationship amongst disease contraction. However, moving forward in the case study a new correlation matrix was created to determine the relationship amongst all attributes (*Figure 11*).

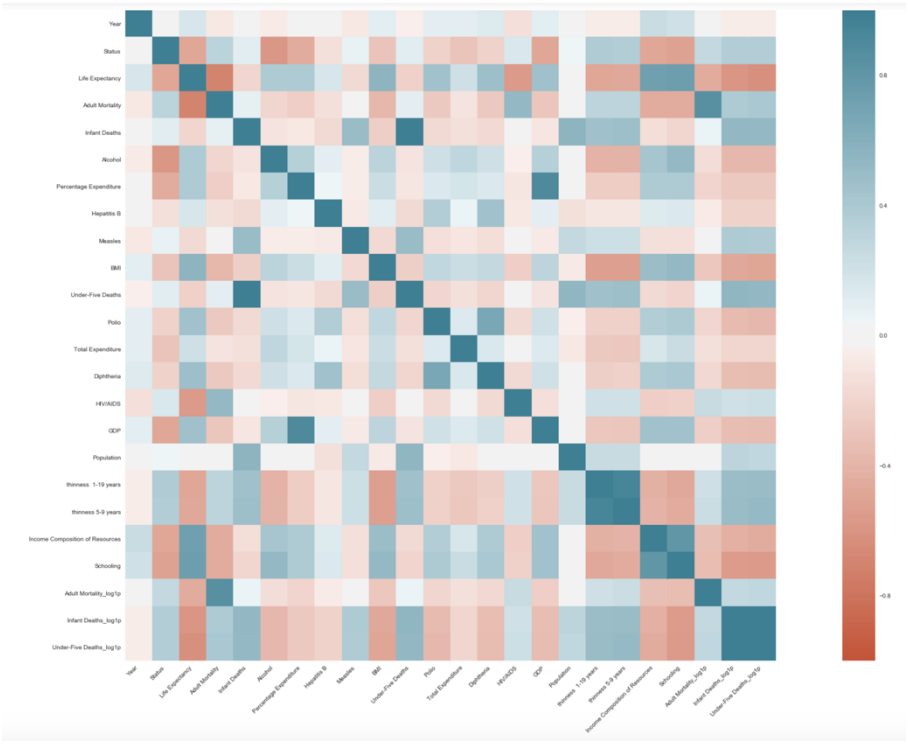


Figure 11. Correlation matrix for all attributes in the dataset including log transformed attributes

Figure 11 demonstrates that while there are no strong correlations between the attributes and life expectancy, there is a slight correlation with the following attributes:

- Status
- Adult Mortality
- BMI
- HIV/AIDS
- Income Composition of Resources
- Schooling
- log of Infant Deaths
- log of Under Five Deaths

As a result, these attributes will be used as predictors in the linear model. Since Schooling and Income Composition of Resources were not analyzed previously, some EDA was necessary to get these attributes ready for modelling. Schooling had 163 null values while Income Composition of Resources had 167. These null values were filled in using the same process described in *Table 1*.

With the data ready for the model, training and testing subsets were created where the training subset was 70% of the original dataset and the testing subset was 30% of the original subset. This resulted in 2056 data points in the training set and 882 data points in the testing set.

Model Assessment

Once the data was fitted to the linear model, predictions were created and evaluated against with the actual values. *Figure 12* demonstrates the relationship between the predicted and actual values of the linear model.

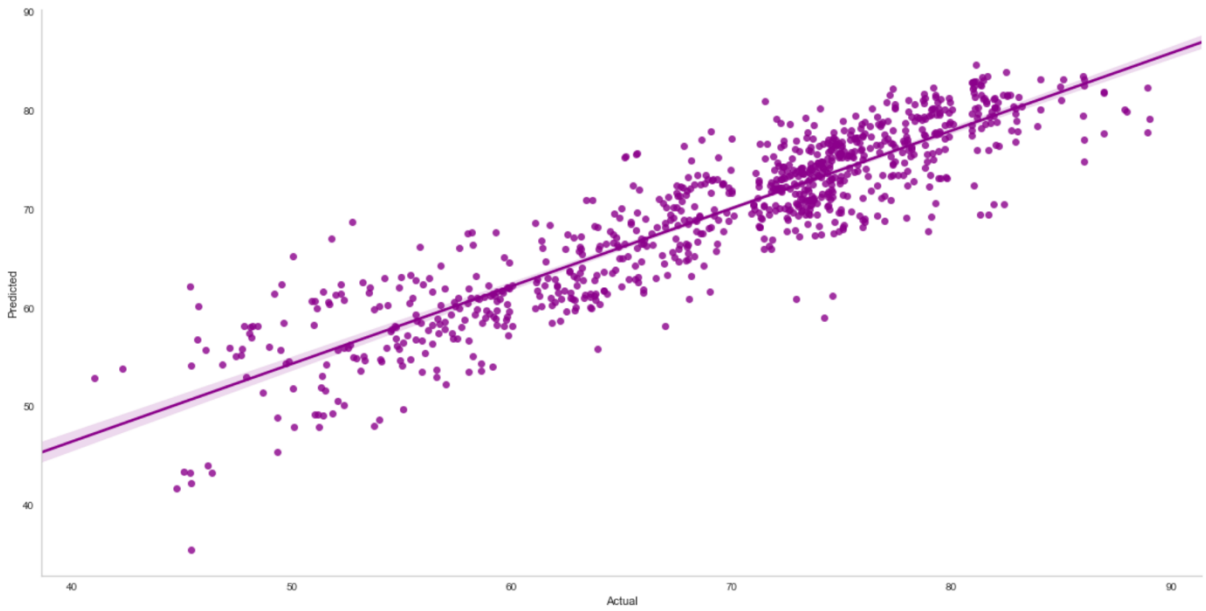


Figure 12. Regression plot showing the predicted values in relation to their actual values

The regression plot in *Figure 12* shows that for the most part the data seemed to follow the regression model fairly accurately. To get a concrete understanding of how well the model fit the data, the model score was taken (R^2 value), which was 0.816. This means the model was performing within 81.6% accuracy, which is pretty good for an initial model attempt.

Conclusion

Life expectancy is something all countries around the world aim to increase each year. With so many tools and so much data made available, it only seems fit to create a model that can help identify the predictors of longevity. This study aimed to do just that and was able to create a model that performs within 81.6% accuracy. While there are certainly many predictors that were not included in the dataset, this model can used as a baseline for future research.

References

Life Expectancy (WHO). (n.d.). Retrieved from https://www.kaggle.com/kumarajarshi/life-expectancy-who/download