STATISTICAL ANALYSIS ON FACTORS INFLUENCING LIFE EXPECTANCY

BY

1. Ahmad Mustapha Wali (407)
2. Lara Tomeh (407)
3. Alexandru-Ștefan Ghiță (407)

Being our project for Exploratory Data Analysis

ABOUT THE DATASET

The World Health Organization's (WHO) Global Health Observatory (GHO) data collection monitors a wide range of health-related factors across all nations. The datasets are made accessible to the public with the intention of studying health data. There is a WHO data repository for life expectancy and health characteristics for 193 countries, and the United Nations website provides economic information for the same number. Among the various categories of health-related characteristics, only those that are most representative of the population were chosen. Over the last 15 years, there has been a dramatic improvement in human mortality rates in the health sector, particularly in developing nations, when compared to the previous 30 years. This is why, in this project's research, data from 2000 to 2015 for 193 countries was used. By integrating all of the original data files, a single dataset has been formed. When the data was originally examined, several values were discovered to be missing. The most often missing figures were discovered to be population, Hepatitis B, and GDP. Tonga, Togo, and Cape Verde were among the countries where data was absent. The final merged file (final dataset) has 22 columns and 2938 rows, yielding 20 predictor variables. Following that, all variables that may be utilized to create predictions were divided into four primary categories: vaccination, mortality, economy, and society.

This project aims to answer the following key questions:

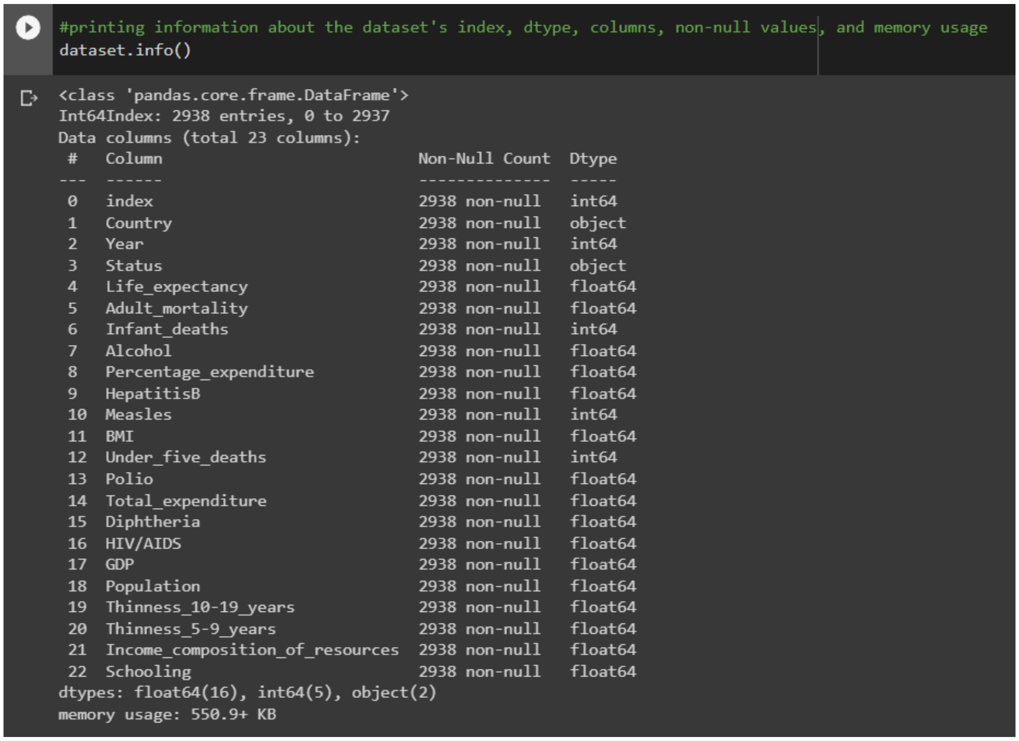
1. Do the numerous predictive factors that were previously considered have any effect on life expectancy?
2. What are the true determinants of life expectancy that can be predicted?

3. Should a country with a lower life expectancy (<65) spend more on healthcare in order to boost its average lifespan?
4. How do death rates for infants and adults impact life expectancy?
5. Do eating habits, lifestyle, exercise, drinking, have a favorable or negative impact on life expectancy, among individuals.
6. What effect does education have on human lifespan?
7. Does life expectancy in highly populated nations tend to be lower?
8. What effect does vaccination coverage have on life expectancy?

The dataset has the following 22 columns and 2938 rows:

1. Country - the country from which the indicators originate.
2. Year - the calendar year in which the indicators were created.
3. Status - whether a country is classified as 'Developing' or 'Developed' by the World Health Organization.
4. Life_expectancy - the life expectancy of a person in years for a specific country and year.
5. Adult_mortality - the adult mortality rate as a percentage of the total population (i.e., number of people dying between 15 and 60 years per 1000 population).
6. Infant_deaths - Infant mortality rate per 1000 population.
7. Alcohol- The rate of alcohol consumption in a nation is expressed in liters of pure alcohol consumed per capita.
8. Percentage_expenditure - health expenditures as a proportion of GDP.
9. Hepatitis_b - Number of one-year-olds immunized against Hepatitis B as a percentage of all one-year-olds in the population.
10. Measles - Measles cases recorded per 1000 population.
11. BMI - A country's whole population's average Body Mass Index (BMI).
12. Under-five_deaths - number of people under the age of five deaths per 1000 population.
13. Polio - The proportion of one-year-olds immunized against Polio compared to the total population of one-year-olds.
14. Total_expenditure - Health expenditures as a proportion of overall government spending.
15. Diphtheria - Immunization rate of one-year-olds against diphtheria, tetanus toxoid, and pertussis (DTP3).

16. HIV/AIDS - HIV/AIDS-related deaths per 1000 live births among children under the age of five; number of children under the age of five that die as a result of HIV/AIDS per 1000 live births.
17. GDP - Gross Domestic Product per capita.
18. Population - population of a country.
19. Thinness_1-19_years - rate of thinness among people aged 10-19.
20. Thinness_5-9_years - rate of thinness among people aged 5-9.
21. Income_composition_of_resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1).
22. Schooling - average number of years of schooling of a population.



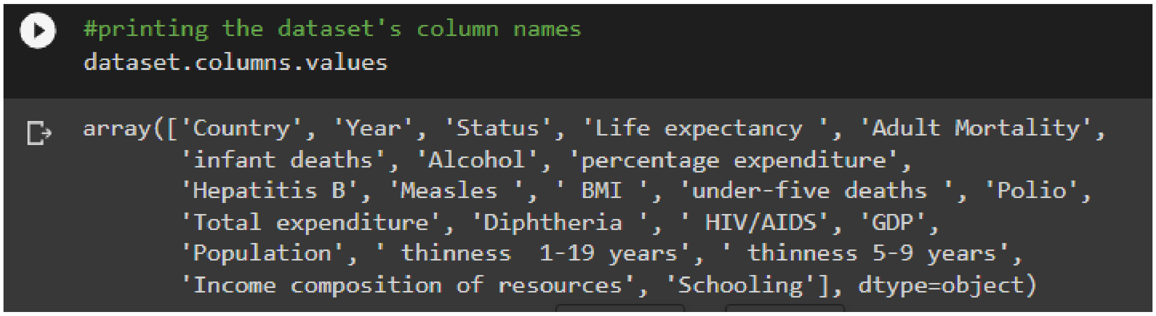
```
#printing information about the dataset's index, dtype, columns, non-null values, and memory usage
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2938 entries, 0 to 2937
Data columns (total 23 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   index                            2938 non-null   int64  
 1   Country                          2938 non-null   object 
 2   Year                             2938 non-null   int64  
 3   Status                           2938 non-null   object 
 4   Life_expectancy                  2938 non-null   float64
 5   Adult_mortality                  2938 non-null   float64
 6   Infant_deaths                    2938 non-null   int64  
 7   Alcohol                          2938 non-null   float64
 8   Percentage_expenditure           2938 non-null   float64
 9   HepatitisB                       2938 non-null   float64
 10  Measles                          2938 non-null   int64  
 11  BMI                              2938 non-null   float64
 12  Under_five_deaths                2938 non-null   int64  
 13  Polio                            2938 non-null   float64
 14  Total_expenditure                2938 non-null   float64
 15  Diphtheria                       2938 non-null   float64
 16  HIV/AIDS                         2938 non-null   float64
 17  GDP                              2938 non-null   float64
 18  Population                       2938 non-null   float64
 19  Thinness_10-19_years             2938 non-null   float64
 20  Thinness_5-9_years               2938 non-null   float64
 21  Income_composition_of_resources  2938 non-null   float64
 22  Schooling                        2938 non-null   float64
dtypes: float64(16), int64(5), object(2)
memory usage: 550.9+ KB
```

Dataset Information

There are 3 categorical variables and 19 discrete numeric columns in total. 21 of the columns are also features, and one to serve as the label (Life_expectancy).

DATA PREPROCESSING



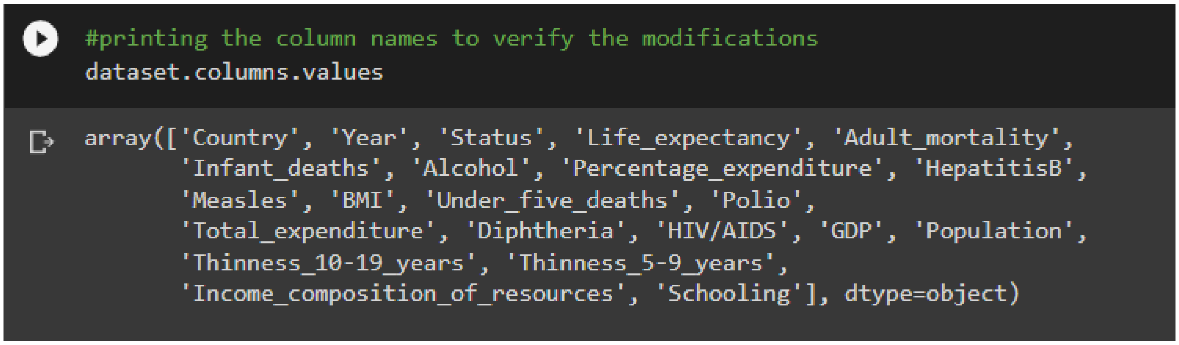
```
#printing the dataset's column names
dataset.columns.values
```

```
array(['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'], dtype=object)
```

Initial Dataset Columns

It could be observed that the column names in the dataset are inconsistently spaced and contain several white spaces. They were renamed to have no spaces but underscores.

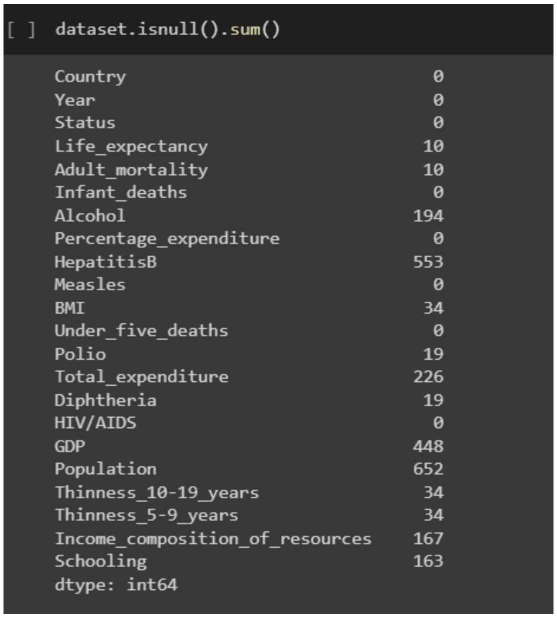
The column “ thinness 1-19 years” was also renamed to “Thinness_10-19_years” as specified in the dataset description.



```
#printing the column names to verify the modifications
dataset.columns.values
```

```
array(['Country', 'Year', 'Status', 'Life_expectancy', 'Adult_mortality',
       'Infant_deaths', 'Alcohol', 'Percentage_expenditure', 'HepatitisB',
       'Measles', 'BMI', 'Under_five_deaths', 'Polio',
       'Total_expenditure', 'Diphtheria', 'HIV/AIDS', 'GDP', 'Population',
       'Thinness_10-19_years', 'Thinness_5-9_years',
       'Income_composition_of_resources', 'Schooling'], dtype=object)
```

Columns After Renaming

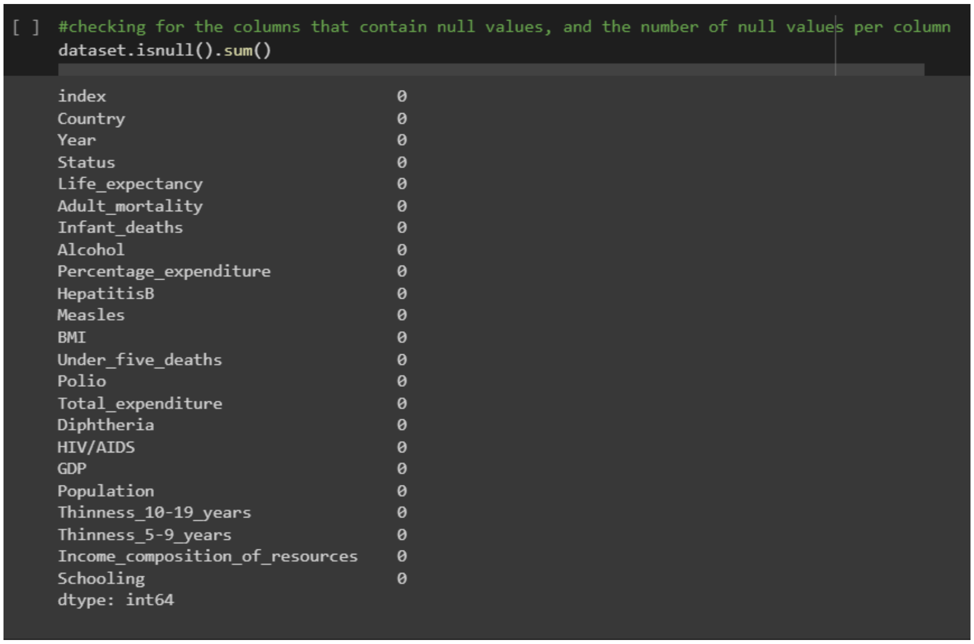


```
[ ]  dataset.isnull().sum()

     Country                            0
     Year                               0
     Status                             0
     Life_expectancy                   10
     Adult_mortality                   10
     Infant_deaths                      0
     Alcohol                          194
     Percentage_expenditure             0
     HepatitisB                       553
     Measles                            0
     BMI                               34
     Under_five_deaths                  0
     Polio                             19
     Total_expenditure                226
     Diphtheria                        19
     HIV/AIDS                           0
     GDP                              448
     Population                       652
     Thinness_10-19_years              34
     Thinness_5-9_years                34
     Income_composition_of_resources  167
     Schooling                        163
     dtype: int64
```

Null Values

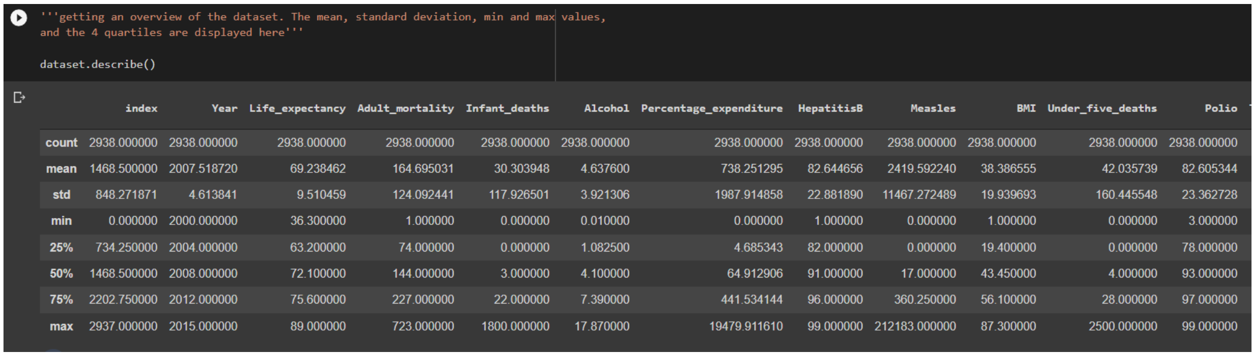
The data contained at least 652 null values. We initially attempted to fill in those numbers using linear interpolation, which did not work, and then filled them in using the median of each column.



```
[ ]  #checking for the columns that contain null values, and the number of null values per column
     dataset.isnull().sum()

     index                              0
     Country                            0
     Year                               0
     Status                             0
     Life_expectancy                    0
     Adult_mortality                    0
     Infant_deaths                      0
     Alcohol                            0
     Percentage_expenditure             0
     HepatitisB                         0
     Measles                            0
     BMI                                0
     Under_five_deaths                  0
     Polio                              0
     Total_expenditure                  0
     Diphtheria                         0
     HIV/AIDS                           0
     GDP                                0
     Population                         0
     Thinness_10-19_years               0
     Thinness_5-9_years                 0
     Income_composition_of_resources    0
     Schooling                          0
     dtype: int64
```

Sum of Null Values After Filling



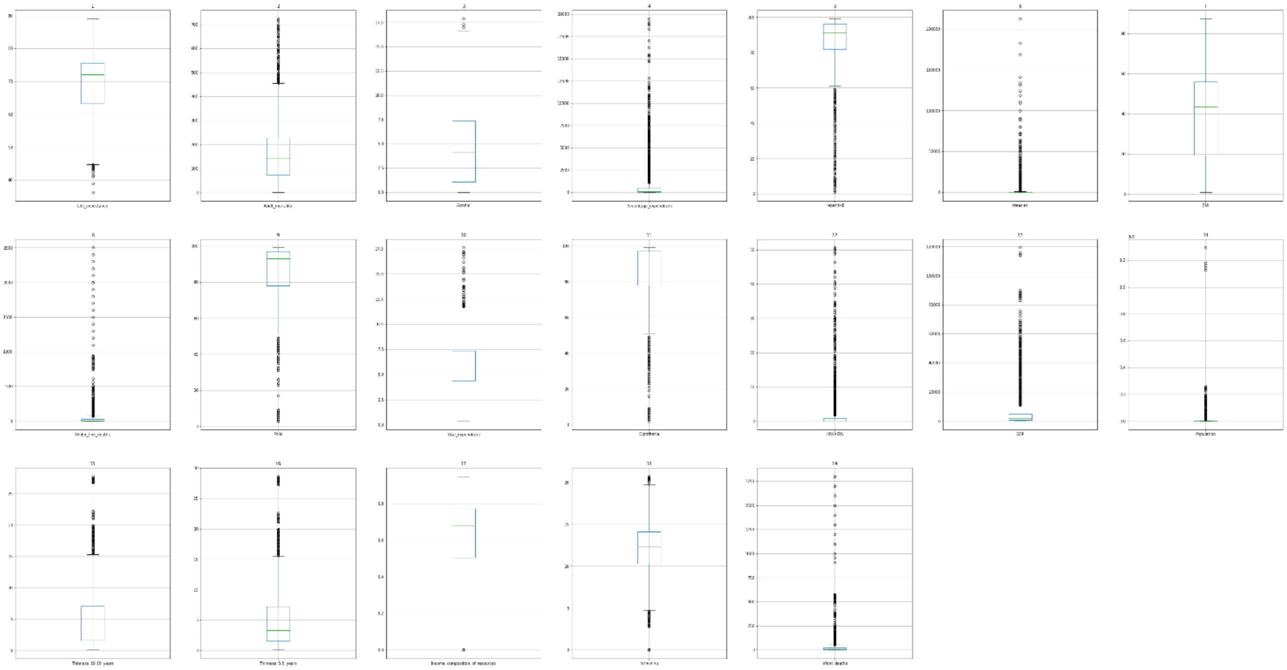
```
'''getting an overview of the dataset. The mean, standard deviation, min and max values,
and the 4 quartiles are displayed here'''

dataset.describe()
```

```
             index         Year  Life_expectancy  Adult_mortality  Infant_deaths      Alcohol  Percentage_expenditure   HepatitisB        Measles          BMI  Under_five_deaths        Polio
count  2938.000000  2938.000000      2938.000000      2938.000000    2938.000000  2938.000000             2938.000000  2938.000000    2938.000000  2938.000000        2938.000000  2938.000000
mean   1468.500000  2007.518720        69.238462       164.695031      30.303948     4.637600              738.251295    82.644656    2419.592240    38.386555          42.035739    82.605344
std     848.271871     4.613841         9.510459       124.092441     117.926501     3.921306             1987.914858    22.881890   11467.272489    19.939693         160.445548    23.362728
min       0.000000  2000.000000        36.300000         1.000000       0.000000     0.010000                0.000000     1.000000       0.000000     1.000000           0.000000     3.000000
25%     734.250000  2004.000000        63.200000        74.000000       0.000000     1.082500                4.685343    82.000000       0.000000    19.400000           0.000000    78.000000
50%    1468.500000  2008.000000        72.100000       144.000000       3.000000     4.100000               64.912906    91.000000      17.000000    43.450000           4.000000    93.000000
75%    2202.750000  2012.000000        75.600000       227.000000      22.000000     7.390000              441.534144    96.000000     360.250000    56.100000          28.000000    97.000000
max    2937.000000  2015.000000        89.000000       723.000000    1800.000000    17.870000            19479.911610    99.000000  212183.000000    87.300000        2500.000000    99.000000
```

Dataset Description

A brief inspection of the dataset description shows the presence of many outliers. Adult mortality of 0.1%, for example, seems unreasonable (min Adult_mortality = 1), and a country cannot spend 19479.911610% of its income on health, etc. Various methods of checking and correcting outliers exist. For this project, we used boxplots for easy visual detection of the outliers.



Checking for Outliers Using Boxplots

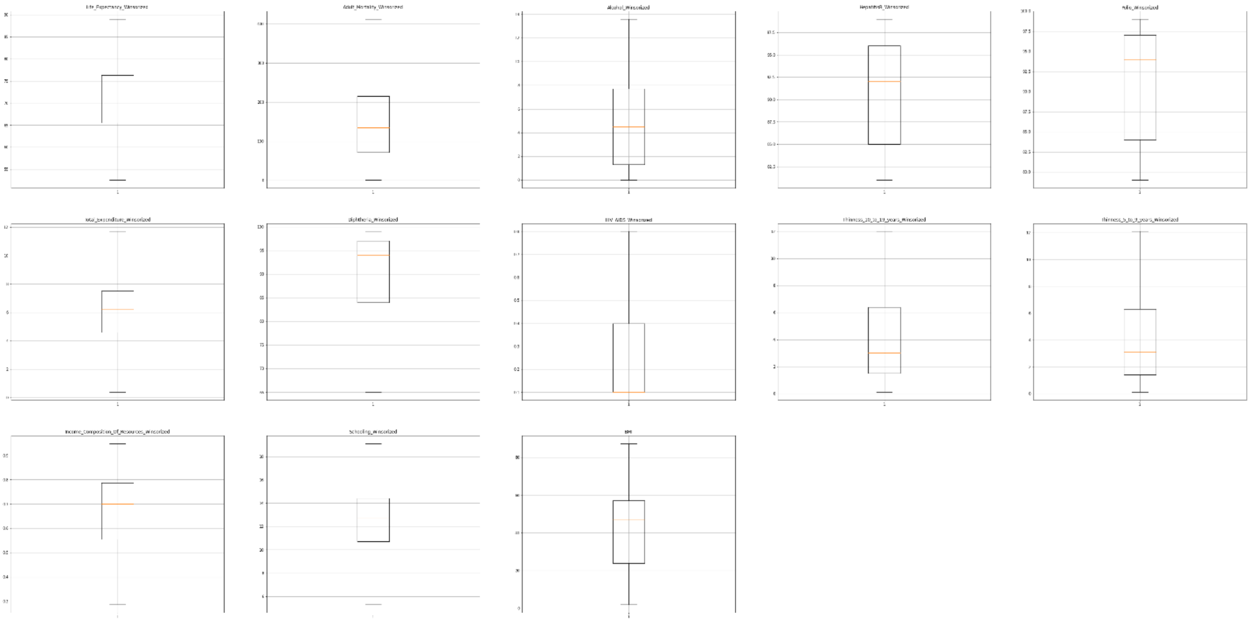
There seem to be many outliers and irregularities with the dataset.

i) *Infant_deaths*, *Measles*, and *Under_five_deaths*, which are per 1000 population (thus shouldn't be above 1000) all have values above 1000. To deal with this, we dropped all values not less than 1000 since they are probably a mistake, and there is no way we can know their actual values.

ii) *Percentage_expenditure* is unreasonably high. Also, *Population* and *GDP* have been affected by the few countries having a very high population of more than 1 billion (China and India) and GDP per capita (Qatar, Norway, etc.) respectively to create too many outliers. To deal with this category, we took the log of these columns.

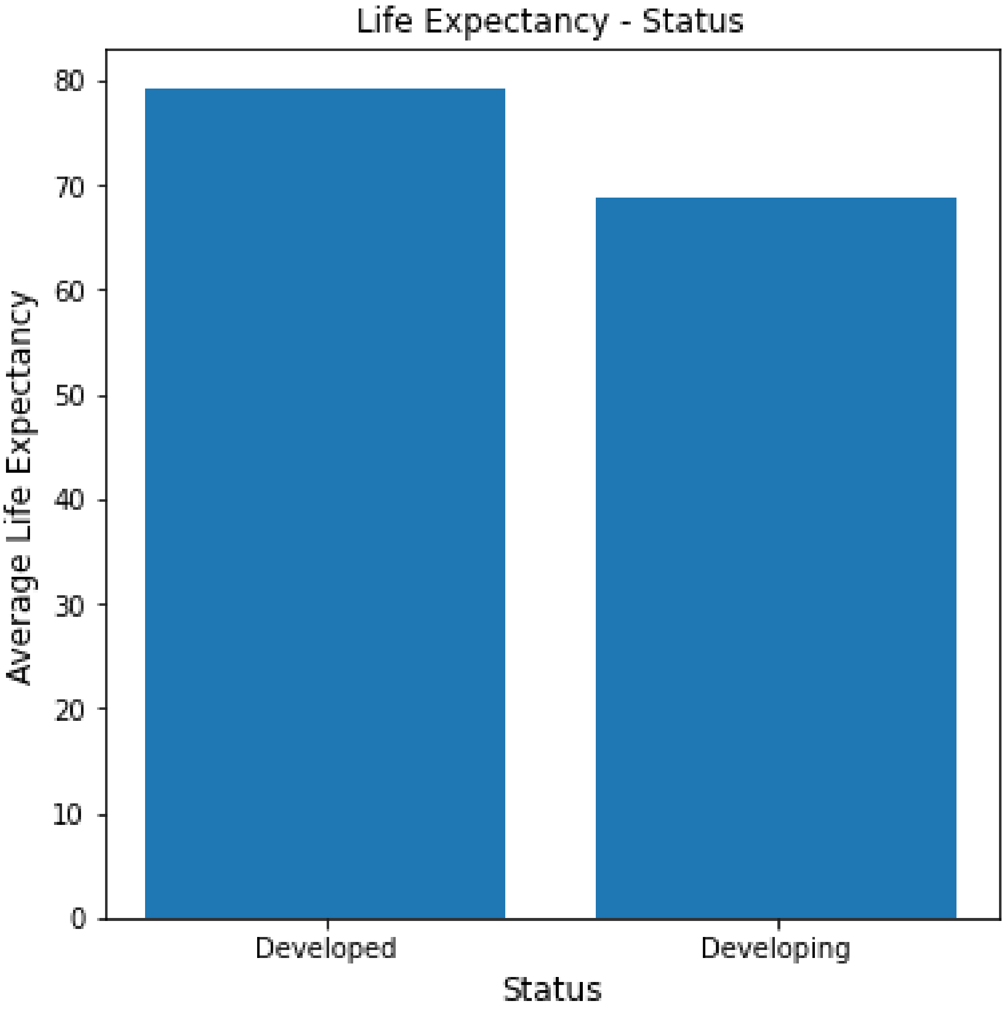
iii) For the rest of the columns with outliers, we winsorized them by changing their quartile limits.

A second boxplot of the winsorized data shows that the outliers have been removed.



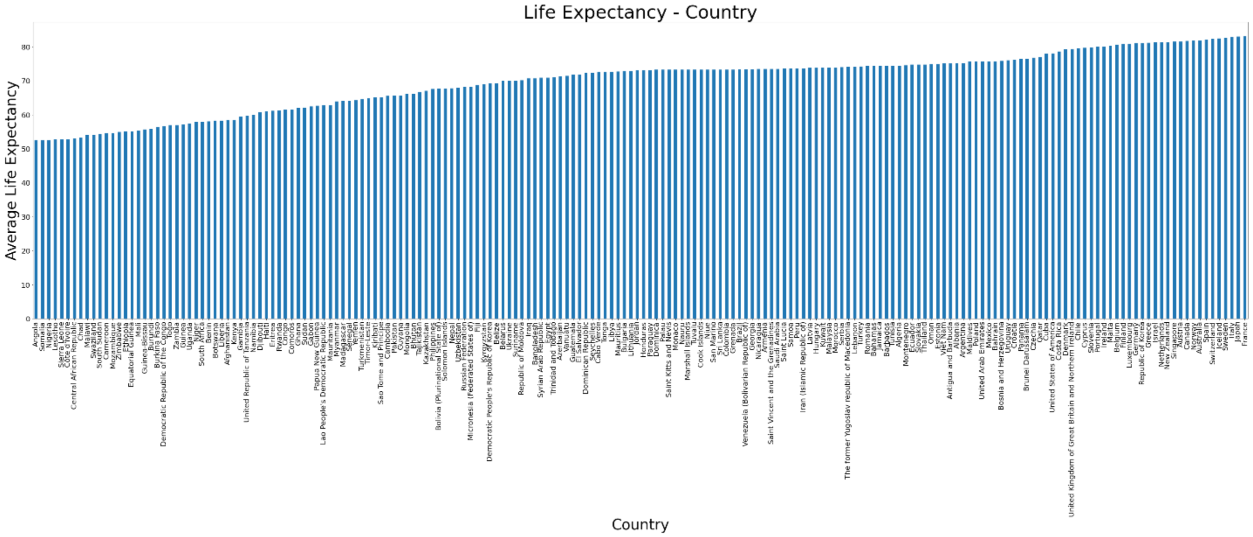
Boxplot After Winsorization

DATA ANALYSIS

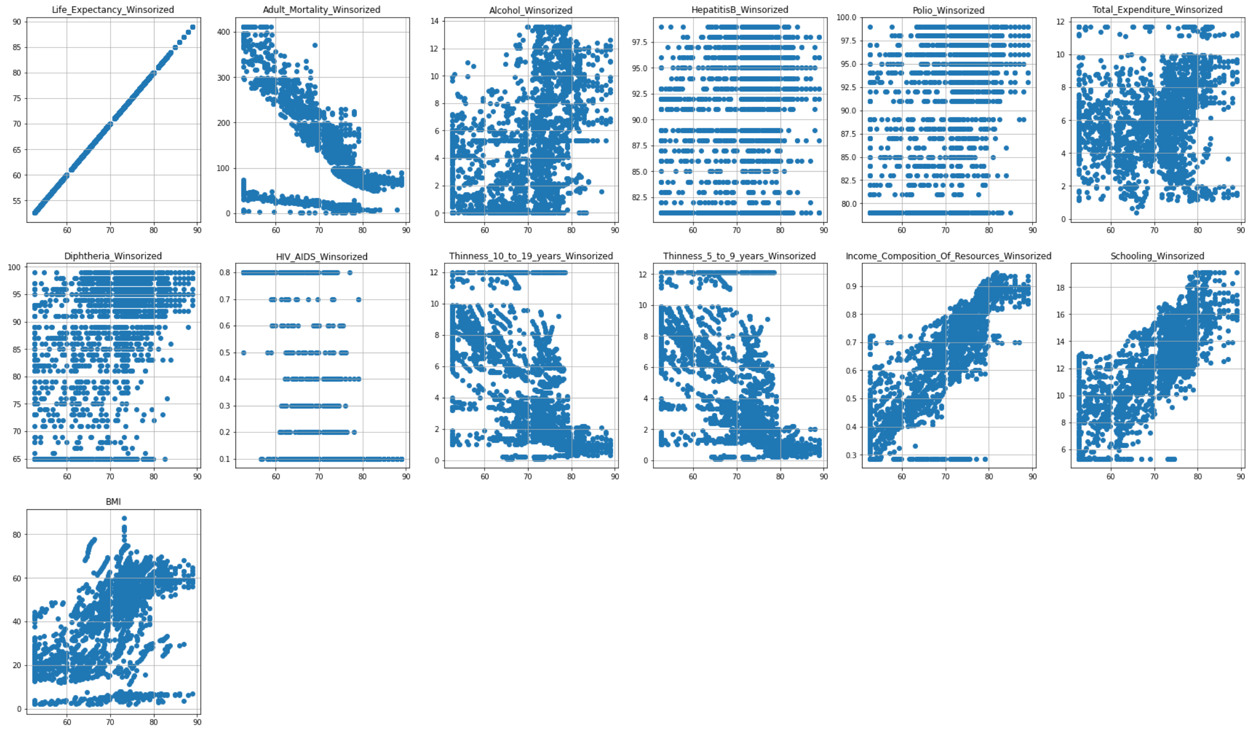


Mean Life Expectancy Grouped by Country Status

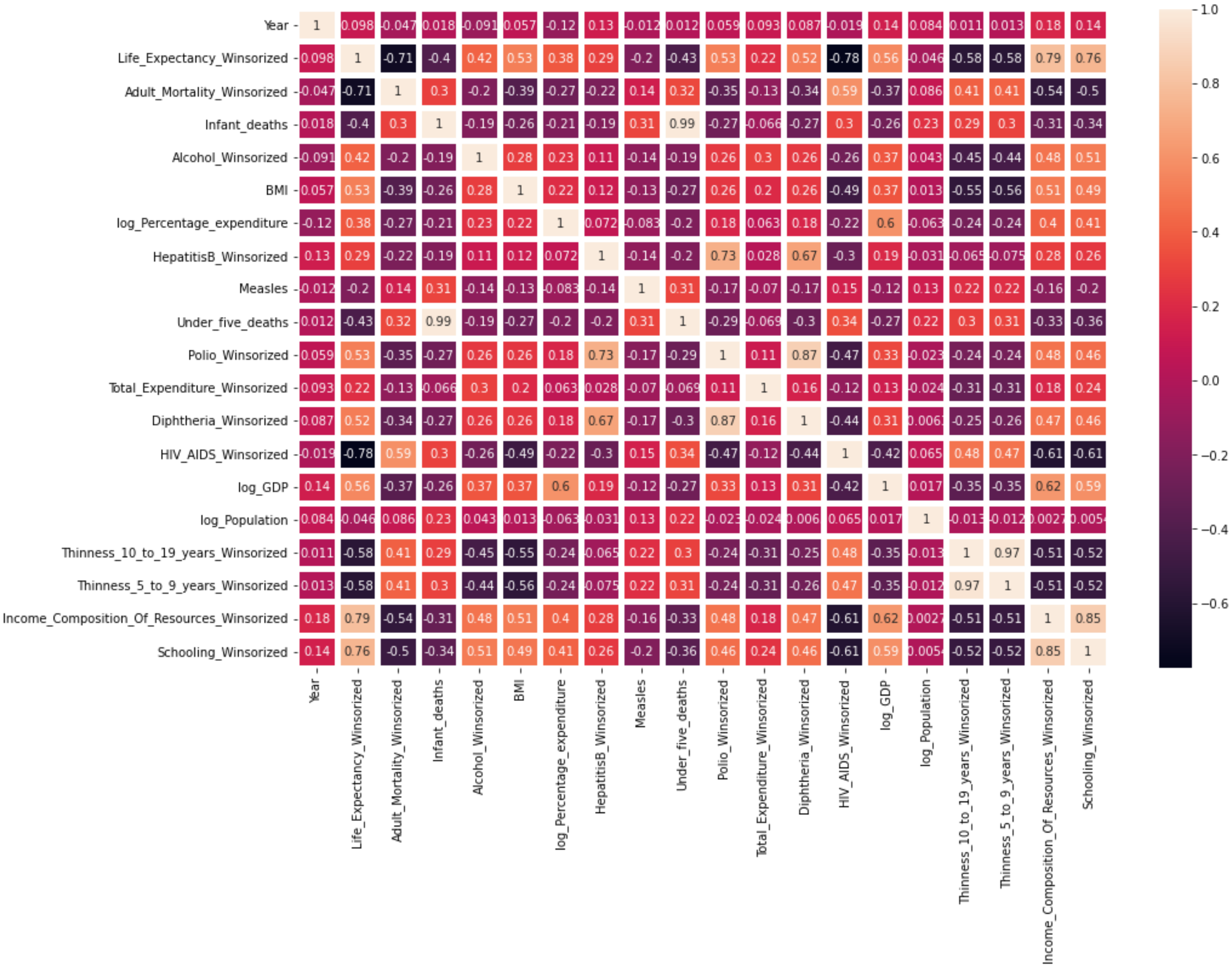
Residents in developed countries have, on average, 10 years higher life expectancy than those in developing countries on average (70 years). As can also be observed in the figure below, the country with the lowest life expectancy is *Angola*, a developing country; while the one with the one with the highest is *France*, a developed country. On closer observation, most of the countries below the 75th percentile are developing countries, while most of those above it are developed nations.



Countries' Life Expectancy Sorted in Ascending Order



Plot of Life Expectancy Against the other Variables

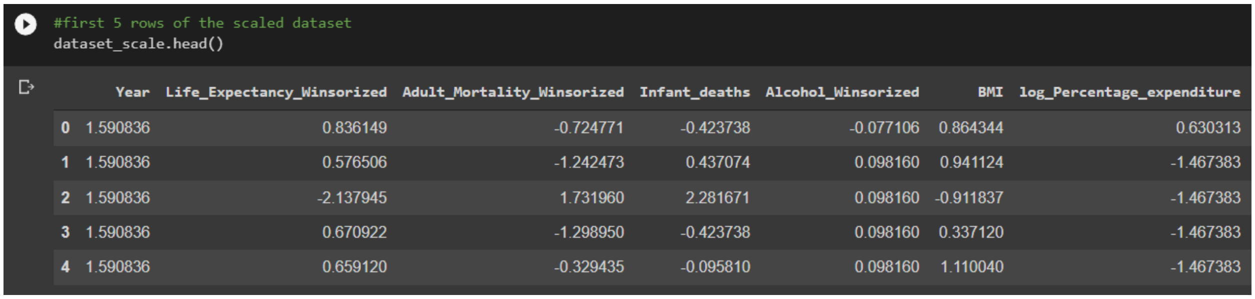


Correlation Heatmap

The 2 figures above indicate the correlations between a predictor and the other predictors. In the case of life expectancy, we can see that it has strong positive correlation with *Income_Composition_Of_resources* and *Schooling*. To a lesser extent, it is also affected positively by *Polio*, and *Diphtheria*. This is because polio and diphtheria immunization greatly reduce the chances of children dying from those ailments. It turns out that public healthcare expenditure is not the most important predictor determining life expectancy (weakly positively correlated by 0.22).

Life expectancy is also observed to be strongly negatively correlated with *Adult_mortality*, *HIV/AIDS*, *Thinness_10-19_Years*, and *Thinness_5-9_Years*.

APPLYING REGRESSION MODELS

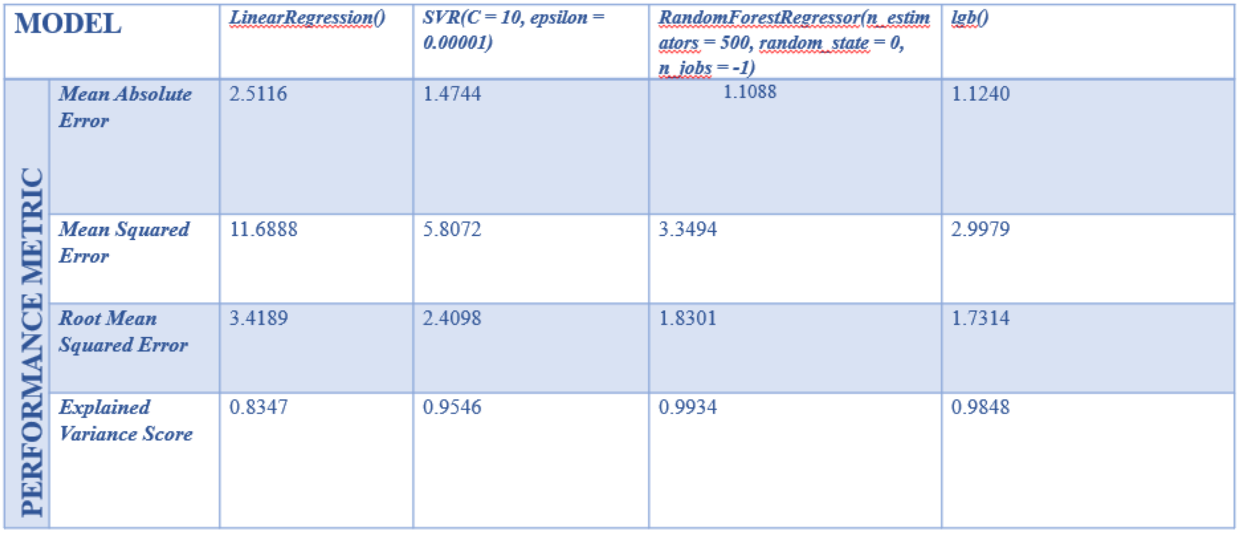


```
#first 5 rows of the scaled dataset
dataset_scale.head()
```

```
       Year  Life_Expectancy_Winsorized  Adult_Mortality_Winsorized  Infant_deaths  Alcohol_Winsorized        BMI  log_Percentage_expenditure
0  1.590836                    0.836149                   -0.724771      -0.423738           -0.077106   0.864344                    0.630313
1  1.590836                    0.576506                   -1.242473       0.437074            0.098160   0.941124                   -1.467383
2  1.590836                   -2.137945                    1.731960       2.281671            0.098160  -0.911837                   -1.467383
3  1.590836                    0.670922                   -1.298950      -0.423738            0.098160   0.337120                   -1.467383
4  1.590836                    0.659120                   -0.329435      -0.095810            0.098160   1.110040                   -1.467383
```

A Part of the Dataset After Scaling

Being a regression task, we scaled the dataset using *StandardScaler* then fit it into 4 regression models using 4 evaluation metrics of mean absolute error, mean squared error, root mean squared error, and explained variance score. The models used are the linear, support vector, random forest, and LightGBM regressors. The figure below shows a summary of the models' performance.



MODEL		***LinearRegression()***	***SVR(C = 10, epsilon = 0.00001)***	***RandomForestRegressor(n_estimators = 500, random_state = 0, n_jobs = -1)***	***lgb()***
PERFORMANCE METRIC	***Mean Absolute Error***	2.5116	1.4744	1.1088	1.1240
	Mean Squared Error	11.6888	5.8072	3.3494	2.9979
	Root Mean Squared Error	3.4189	2.4098	1.8301	1.7314
	Explained Variance Score	0.8347	0.9546	0.9934	0.9848

Model Performance Summary

From the above figure, the best models for this task are the LightGBM and the random forest regressor, and the support vector regressor performed fairly good as well. The linear regressor, however, performed relatively bad compared to the others.

CONCLUSION

As indicated by the correlation matrix, some of the predictive factors initially considered had little effect on life expectancy. Government expenditure on healthcare, population, and infant measles rates all have a negligible influence on a country's life expectancy. Life expectancy is, however, significantly influenced by a country's HDI and the average number of years spent in school by its citizens. It is also favorably influenced, albeit to a lesser extent, by a country's GDP, polio and diphtheria vaccination rates, and the average BMI of its residents.

Adult mortality has a large negative effect on life expectancy (-0.7), but infant mortality has a comparatively little negative effect (-0.4). Countries' life expectancy is also heavily influenced negatively by the number of children dying of HIV under the age of five. Other significant, although less substantial, negative correlations include the prevalence of thinness among children aged 5-9 and adolescents aged 10-19.

As previously stated, a country's public health expenditure is weakly correlated to its citizens' life expectancy (0.22). Life expectancy is positively correlated with lifestyle choices such as eating habits and exercise (which eventually impact the BMI) (0.53). Additionally, alcohol use is favorably connected with life expectancy (0.53). A plausible explanation for this might be because persons who consume more alcohol typically have more money.

As seen in the matrix, education is the second most significant predictor in the dataset (0.76), while population has a negligible effect on life expectancy (0.046). Countries that invest more in education will almost certainly see an increase in life expectancy.