DATA 583 Exploratory Data Analysis

Ujjwal Upadhyay, Shveta Sharma, Varshita Kyal

2023-03-12

Statistical Description

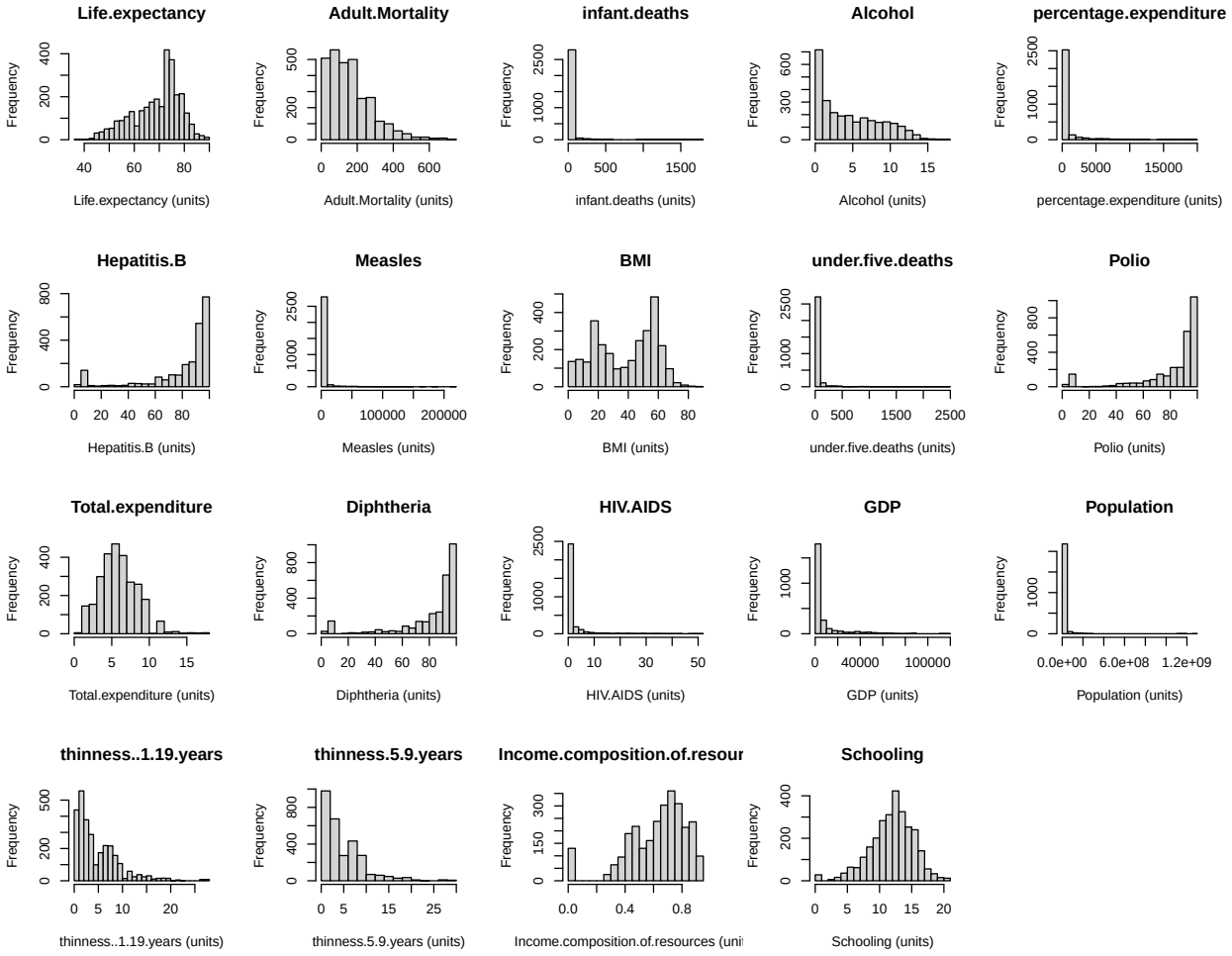
There are 3 categorical variables and 19 continuous numerical variables in the dataset."Life Expectancy" is our response variable. The dataset has few missing values which we will handle as part of data wrangling step. There are 2938 observations in the initial dataset.We will try to imputate missing values in columns by looping through columns and replace NA values with mean of each column for each country.1289 rows containing NAs were reduced to 810.Hence,we still have close to 27% of rows with some NA attribute. These observations will be deleted due to missingness while modelling.

Univariate Analysis

Table 1. Statistics for Numerical Predictors

	n	mean	sd	median	skew	kurtosis	se
Year	2938	2007.52	4.61	2008.00	-0.01	1.79	0.09
Life.expectancy	2928	69.22	9.52	72.10	-0.64	2.76	0.18
Adult.Mortality	2928	164.80	124.29	144.00	1.17	4.74	2.30
infant.deaths	2938	30.30	117.93	3.00	9.78	118.84	2.18
Alcohol	2744	4.60	4.05	3.76	0.59	2.20	0.08
percentage.expenditure	2938	738.25	1987.91	64.91	4.65	29.53	36.68
Hepatitis.B	2385	80.94	25.07	92.00	-1.93	5.76	0.51
Measles	2938	2419.59	11467.27	17.00	9.44	117.66	211.56
BMI	2904	38.32	20.04	43.50	-0.22	1.71	0.37
under.five.deaths	2938	42.04	160.45	4.00	9.49	112.56	2.96
Polio	2919	82.55	23.43	93.00	-2.10	6.77	0.43
Total.expenditure	2712	5.94	2.50	5.76	0.62	4.15	0.05
Diphtheria	2919	82.32	23.72	93.00	-2.07	6.55	0.44
HIV.AIDS	2938	1.74	5.08	0.10	5.39	37.83	0.09
GDP	2490	7483.16	14270.17	1766.95	3.20	15.31	285.98
Population	2286	12753375.12	61012096.51	1386542.00	15.91	300.36	1276079.80
thinness..1.19.years	2904	4.84	4.42	3.30	1.71	6.96	0.08
thinness.5.9.years	2904	4.87	4.51	3.30	1.78	7.35	0.08
Income.composition.of.resources	2771	0.63	0.21	0.68	-1.14	4.39	0.00
Schooling	2775	11.99	3.36	12.30	-0.60	3.88	0.06

Figure 1. Visualize the numerical predictors data distribution using histograms



After analyzing Table 1 and Figure 1, we have concluded as below:

1) The mean life expectancy across all countries is 69.22 years. The median life expectancy is slightly higher at 72.10 years, indicating that the distribution of life expectancies is slightly skewed towards lower values. The negative skewness value of -0.64 confirms this.

2) Adult mortality Rate(probability of dying between 15 and 60 years per 1000 population): The mean is 164.80, and the median is 144.00. The skewness is 1.17, indicating a moderately right-skewed distribution.This is an healthy indicator as most countries have lower value of adult mortality.

3) Infant deaths: The mean is 30.30, and the median is 3.00. The skewness is 9.78, indicating a highly right-skewed distribution.

4) The mean alcohol consumption across all countries is 4.60 liters per capita per year. The median consumption is lower at 3.76 liters per capita per year, indicating that the distribution of alcohol consumption is skewed towards higher values. The positive skewness value of 0.59 confirms this, indicating that the distribution is skewed towards the higher values which is correct as alcohol consumption is pretty common in most countries.

5) Percentage expenditure on health as a percentage of Gross Domestic Product per capita(%) : The mean is 738.25, and the median is 64.91. The skewness is 4.65, indicating a highly right-skewed distribution. It makes sense as only a small proportion of countries spend significant funds on healthcare.

6) Hepatitis B Immunization coverage of 1-year olds has mean 80.94, and the median is 92.00. The skewness is -1.93, indicating a moderately left-skewed distribution.This is again a decent indicator showing that most countries have partially complete to complete immunization coverage for this disease. Although, there is a significant number of countries where there is no immunization coverage.

7) Reported cases for Measles per 1000 population has mean 2419.59, and the median is 17.00. The skewness is 9.44, indicating a highly right-skewed distribution. In most developed countries, measles is now a rare disease due to widespread vaccination programs, so, the data makes sense.

8) Body Mass Index of entire population has mean 38.32, and the median is 43.50. The skewness is -0.22, indicating a roughly symmetrical distribution which is bi-modal in nature.

9) Under-five deaths: The mean is 42.04, and the median is 4.00. The skewness is 9.49, indicating a highly right-skewed distribution.The majority of these deaths (around 70%) occurred in sub-Saharan Africa and Southern Asia.

10) Polio immunization coverage among 1-year-olds (%) has mean 82.55, and the median is 93.00 which is quite high. The skewness is -2.10, indicating a moderately left-skewed distribution. Most countries have close to 100% Polio vaccine coverage, so, the data makes sense.

11) Total expenditure i.e.General government expenditure on health as a percentage of total government expenditure (%) has mean 5.94, and the median is 5.76. The skewness is 0.62, indicating a moderately right-skewed distribution.Again, this depends upon the goverment planning of countries and may vary between demographics, so it it expected to be skewed.

12) Diphtheria immunization coverage among 1-year-olds(%) has mean 82.32, and the median is 93.00. The skewness is -2.07, indicating a moderately left-skewed distribution which is expected as most countries have close to 100% coverage.

13) The mean HIV/AIDS prevalence across all countries is 1.74%. The median prevalence is significantly lower at 0.10%, indicating that the distribution of HIV/AIDS prevalence values is heavily skewed towards lower values. The large positive skewness value of 5.39 confirms this, indicating that the distribution is highly skewed towards the lower values which makes sense as HIV/AIDS is not prevalent is all countries.

14) Gross Domestic Product per capita (in USD) has mean 7483.16, and the median is 1766.95. The skewness is 3.20, indicating a highly right-skewed distribution whic is expected as developed nations tend to have higher GDP than developing nations.

15) Population has mean 12753375.12, and the median is 1386542.00. The skewness is 15.91, indicating a highly right-skewed distribution which is expected as a few countries are densely populated compared to a majority of nations that are sparsely populated.

16) Thinness(or malnutrition) 1-19 years as percentage has mean 4.84, and the median is 3.30. The skewness is 1.71, indicating a moderately right-skewed distribution.

17) Thinness 5-9 years: The mean is 4.87, and the median is 3.30. The skewness is 1.78, indicating a moderately right-skewed distribution.

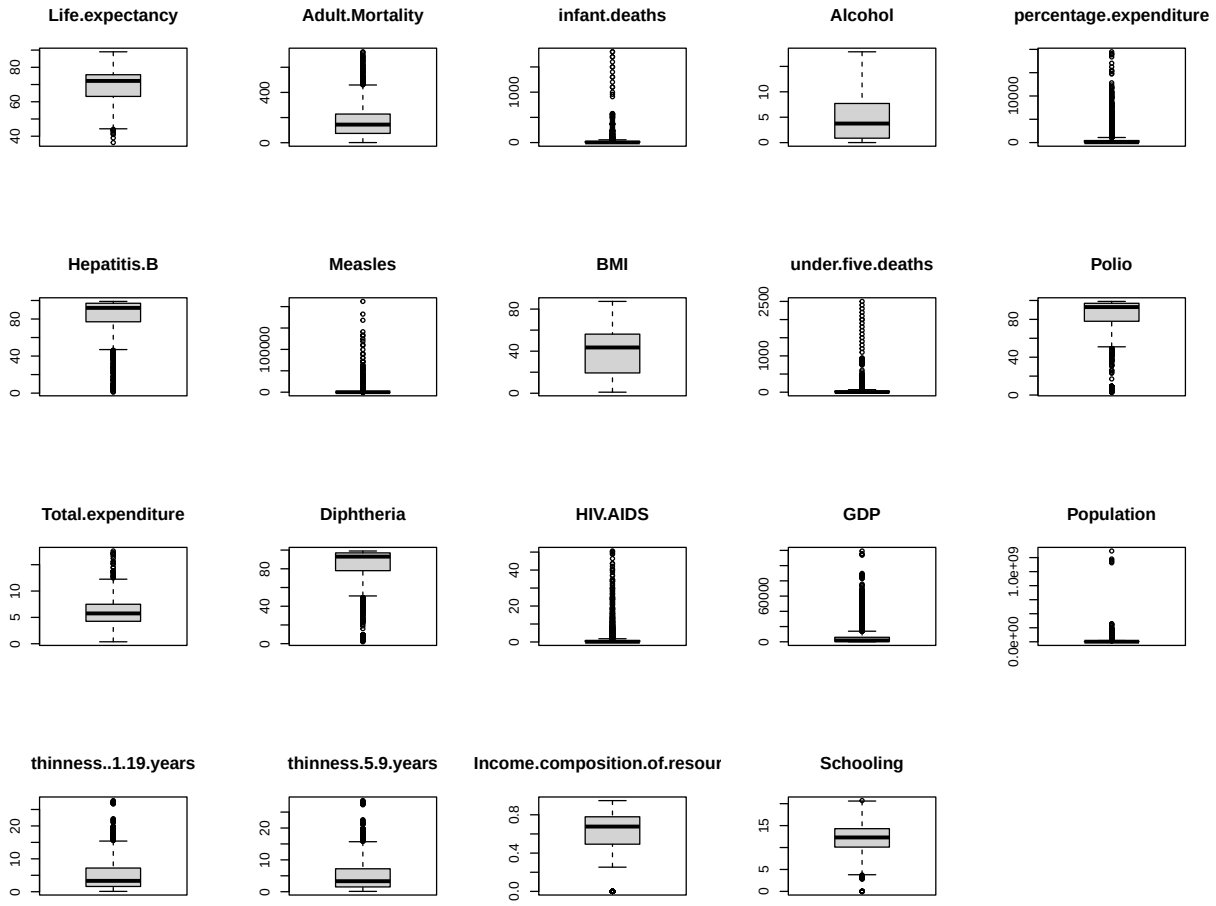
18) Income composition of resources, varies between 0 and 1 has mean 0.63, and the median is 0.68. The skewness is -1.14, indicating a moderately left-skewed distribution.

19) The mean schooling years across all countries is 11.99 years. The median schooling years are slightly higher at 12.30 years, indicating that the distribution of schooling years is slightly skewed towards higher values. The negative skewness value of -0.60 confirms this, indicating that the distribution is slightly skewed towards the higher values.

Modelling technique that may be affected by a right-skewed distribution is hypothesis testing. If the variable is not normally distributed, statistical tests that assume a normal distribution, such as t-tests and ANOVA, may give inaccurate results. In such cases, non-parametric tests may be more appropriate.

In some cases, a right-skewed distribution may also result in outliers, which can have a significant impact on the results of modelling. Outliers can affect the slope and intercept of regression models, and may also affect the accuracy of clustering and classification algorithms

Figure 2. Check for Outliers using Boxplots



As per Figure 2., we do observe outliers data in the diseases related to deaths predictors (HIV, Measles) and immunization predictors. We would retain these outliers because they represent natural variation in the data.Hence,Removing them may result in a loss of valuable information and insights. Also as our sample size is small, Outliers can have a disproportionate impact on the mean, median, and other statistics, and removing them can lead to biased results

Scientific questions about the you will try to answer?

1. What predictors are statistically significant in predicting the Life Expectancy?

2. How does Infant and Adult mortality rates impacts life expectancy?
3. Predicting Life Expectancy of a person based on statistically significant attributes of the underlying dataset.
4. Is there a positive or negative correlation between life expectancy and factors such as eating habits, lifestyle, exercise, and alcohol consumption?
5. How does schooling attribute affect the lifespan of humans?
6. Is there a positive or negative association between drinking alcohol and life expectancy?
7. How do different immunization rates (Hepatitis B, Polio, Diphtheria) impact life expectancy?
8. How does HIV/AIDS prevalence impact life expectancy, particularly in children under five years of age?
9. Does government expenditure on health impact life expectancy, and if so, to what extent?
10. Is there a difference in life expectancy between developed and developing countries, and if so, what factors explain these differences?

Statistical techniques to be used to answer research questions

The scientific inquiries mentioned are primarily concerned with analyzing the statistical significance of the variables rather than just making LifeExpectancy predictions. In order to examine the interrelationships between the highlighted variables, we will conduct regression analyses on the dataset.

1) To determine the significance of variables in predicting life expectancy, a statistical model was developed using a multiple regression analysis.
2) Correlation/Association study via Pearson's correlation coefficient to understand correlation/association between different predictors.
3) To determine which predictors are statistically significant in predicting life expectancy, you can use a statistical model such as multiple linear regression or a generalized linear model (GLM).
4) Built-in variable selection to get the best model
5) Interaction Linear regression model to compare the how the factors affecting life Expectancy vary between developed & developing nations.
6) Decision trees, random forests, or neural networks can also be used depending on the complexity and non-linearity of the data.

Justification for using Regression on Life Expectancy dataset

a) Relationship identification: Regression analysis can identify the relationship between a dependent variable and one or more independent variables. It can help determine the extent to which changes in the independent variables impact the dependent variable.

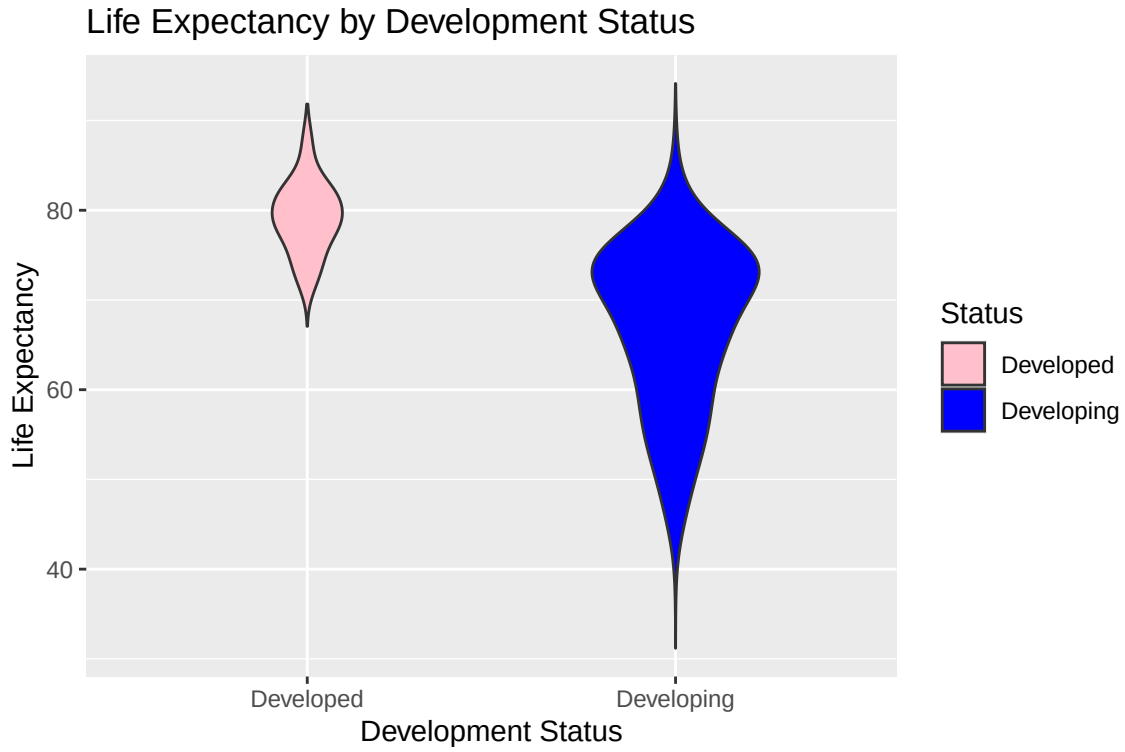
b) Prediction: Regression analysis can be used to predict future values of the dependent variable based on the values of the independent variables. This is particularly useful when there is a need to forecast future trends or patterns in data.

c) Model validation: Regression analysis provides a way to validate a statistical model by comparing the predicted values of the dependent variable with the actual values.

d) Model simplicity: Regression analysis allows for the creation of a simple model that can explain complex relationships between variables. This can be useful for understanding the essential factors that contribute to a particular outcome.

e) Statistical significance: Regression analysis can identify statistically significant relationships between variables, providing confidence that the observed relationships are not due to chance.

Figure 3. Distribution of Life Expectancy between Developed and Developing nations



From Figure 3., we can see the mean life.expectancy of Developed vs developing nations is closer, but the distribution of life.Expectancy in developing nations varies from mid 30s to mid 80s.In order to understand how different factors/predictors affecting lifeExpectancy vary between Developed vs Developing nation using an interaction model

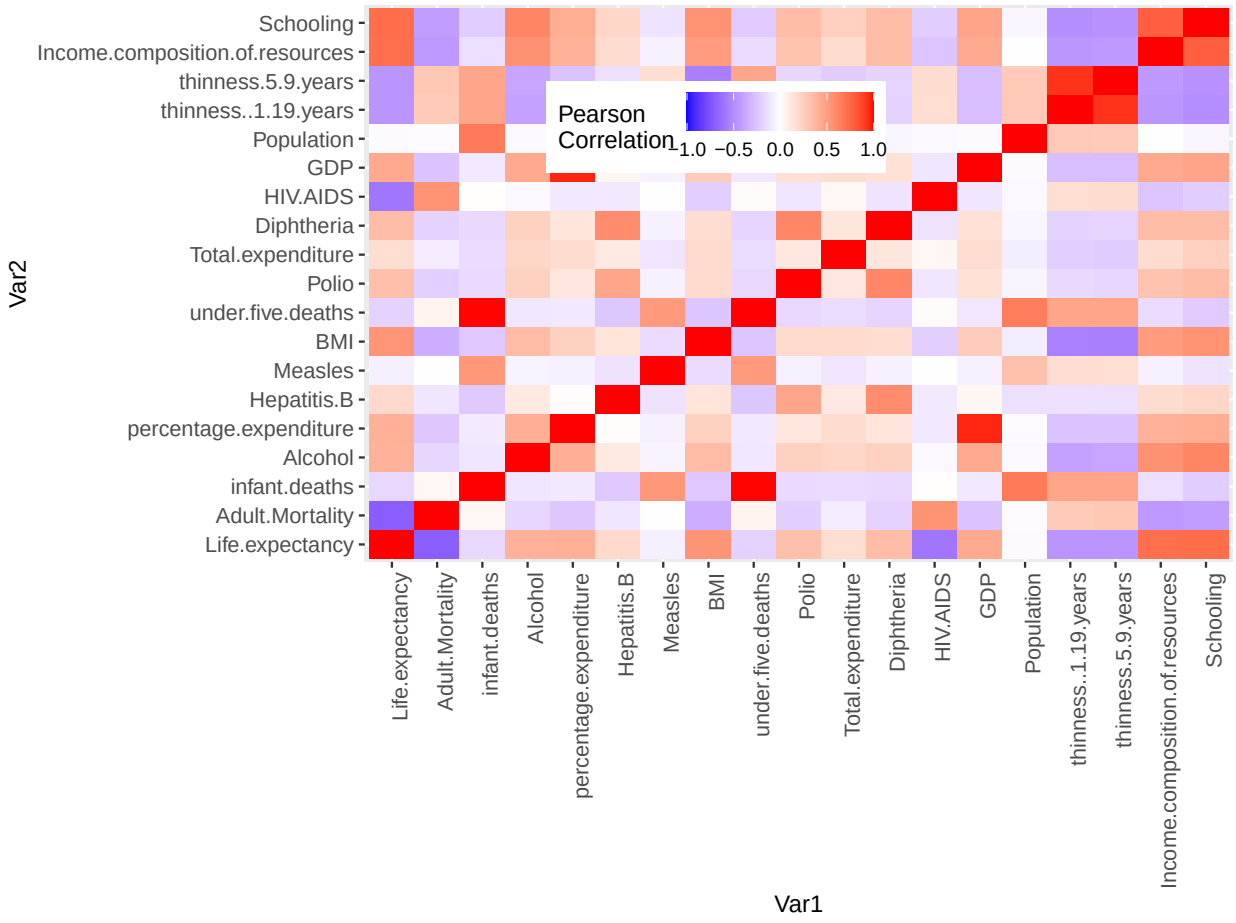
Using Interation LM model The interaction terms are given by the combination of each predictor variable with the "Status" variable. We have used StepAIC with backward variable selection to obtain an interaction LM model. As per the model, The Adult.Mortality, infant.deaths, Alcohol, under.five.deaths, and HIV.AIDS variables have a negative relationship with life expectancy, meaning that as these variables increase, life expectancy decreases. In contrast, variables such as Income.composition.of.resources, Schooling, and Polio have a positive relationship with life expectancy, meaning that as these variables increase, life expectancy also increases.

The model also includes interaction terms between some of the predictor variables and the "Developed" categorical variable. These interaction terms help to capture any differences in the relationship between the predictor variables and the response variable for developed countries compared to developing countries. For instance, the BMI:Developed interaction term indicates that the relationship between BMI and life expectancy is different for developed countries compared to developing countries.

The RSE for this model is 3.786, which means that the typical difference between the actual and predicted values of life expectancy is approximately 3.786 years.

Overall, this model provides a good fit for the data, with a high Adjusted R-squared value and several statistically significant predictor variables. However, it is important to note that the model assumes linearity, independence, normality, and equal variance of the errors, which should be assessed before making any inferences or predictions

Figure 4. Create a heatmap to study Pearson correlation between data



Pearson correlation measures only the linear relationship between two variables and cannot capture non-linear relationships.

Looking at the correlation matrix provided, we can interpret the correlations between different variables in the following ways:

- Life expectancy has a strong positive correlation with BMI (0.54), income composition of resources (0.72), and schooling (0.73), indicating that as these variables increase, life expectancy tends to increase as well.
- Life expectancy has a strong negative correlation with adult mortality (-0.70), HIV/AIDS (-0.59), and thinness in children under 5 years (-0.46), indicating that as these variables increase, life expectancy tends to decrease.
- Alcohol has a weak positive correlation with life expectancy (0.40), indicating that higher levels of alcohol consumption may be associated with higher life expectancy, although this correlation is not very strong.
- GDP has a moderate positive correlation with alcohol consumption (0.44) and percentage expenditure (0.96), indicating that as GDP increases, so does the amount of money spent on alcohol consumption and healthcare.

- Infant deaths and under-five deaths have strong negative correlations with vaccination rates (Polio and Diphtheria), indicating that higher vaccination rates may be associated with lower rates of infant and child mortality.
- Measles has a weak negative correlation with vaccination rates (Polio and Diphtheria), indicating that higher vaccination rates may be associated with lower rates of measles.
- Income composition of resources has a moderate positive correlation with BMI (0.51) and schooling (0.72), indicating that higher income may be associated with better education and nutrition.
- HIV/AIDS has a moderate positive correlation with adult mortality (0.55) and a weak negative correlation with percentage expenditure (-0.10), indicating that higher rates of HIV/AIDS may be associated with higher rates of adult mortality and lower levels of healthcare spending.
- Population has a weak negative correlation with percentage expenditure (-0.02), indicating that higher population size may be associated with lower levels of healthcare spending. However, this correlation is not very strong.

Applying GLM on dataset with family "Guassian"

We experimented with "Guassian" family as it assumes that the conditional distribution of the response variable (here life.expectancy) is normal, and the model is fit using maximum likelihood estimation. We have scaled data before applying GLM.

```
##
## Call:
## glm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     Alcohol + percentage.expenditure + BMI + under.five.deaths +
##     Total.expenditure + Diphtheria + HIV.AIDS + thinness.5.9.years +
##     Income.composition.of.resources + Schooling, family = gaussian,
##     data = na.omit(scaled_new_data))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.78996  -0.21590  -0.00264   0.23398   1.24845
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.003017   0.009392   0.321 0.748107
## Adult.Mortality                 -0.221505   0.012325 -17.972  < 2e-16 ***
## infant.deaths                    1.135786   0.123505   9.196  < 2e-16 ***
## Alcohol                         -0.021911   0.012953  -1.692 0.090924 .
## percentage.expenditure           0.097083   0.012032   8.069 1.36e-15 ***
## BMI                              0.070460   0.012546   5.616 2.29e-08 ***
## under.five.deaths               -1.170022   0.124827  -9.373  < 2e-16 ***
## Total.expenditure                0.020718   0.010653   1.945 0.051968 .
## Diphtheria                       0.037282   0.011285   3.304 0.000975 ***
## HIV.AIDS                        -0.233212   0.009510 -24.524  < 2e-16 ***
## thinness.5.9.years              -0.026791   0.012520  -2.140 0.032512 *
## Income.composition.of.resources  0.218859   0.018391  11.900  < 2e-16 ***
## Schooling                        0.312796   0.020681  15.125  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.1424026)
##
##     Null deviance: 1405.99  on 1648  degrees of freedom
## Residual deviance:  232.97  on 1636  degrees of freedom
## AIC: 1480.5
##
## Number of Fisher Scoring iterations: 2
```

The coefficients for the predictors provide information on how they influence the response variable. For example, the negative coefficient for Adult.Mortality (-0.221505) indicates that as Adult.Mortality increases, Life.expectancy decreases. The positive coefficient for infant.deaths (1.135786) suggests that as the number of infant deaths increases, Life.expectancy decreases. Similarly, the positive coefficient for Income.composition.of.resources (0.218859) suggests that as the income composition of resources increases, Life.expectancy also increases. Some predictors are found to be statistically significant based on the p-values, such as Adult.Mortality, infant.deaths, percentage.expenditure, under.five.deaths, HIV.AIDS, Income.composition.of.resources, and Schooling. These predictors are likely to have a significant impact on the response variable. Overall, this model suggests that factors such as mortality rates, infant and child mortality rates, income and education level, and disease prevalence play a significant role in predicting life expectancy in a country. However, this model may have limitations such as omitted variables, multicollinearity, and influential observations that may affect the accuracy of its predictions. We will analyse further and do model selection in main report.

Appendix

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

Background: There are 22 predictors in the Life Expectancy dataset(2938 records) categorized as below:

1) Character Type Predictors: Country and Status are character type variables which won't be particularly used for analysis.

2) Numeric Type Predictors: We have categorized these continuous variables into different groups to map them to relevant scientific research questions:

a) Immunization predictors: Hepatitis.B, Diphtheria, Polio

b) Disease related deaths predictors: HIV.AIDS, Measles

c) Lifestyle predictors: Alcohol, BMI

d) Social predictors: Adult.Mortality, infant.deaths, thinness..1.19.years, thinness.5.9.years, Schooling, Population

e) Economic predictors: percentage.expenditure, GDP, Income.composition.of.resources, Total.expenditure

3) DateTime Predictor: Year (This might just be used for analyzing trend in visualizations in final report)

Background (Column Descriptions)

1. **Country -** This column has character data for 193 countries for which we want to predict the Life Expectancy with no null values.

2. **Year -** This column has numeric data varying from 2000 to 2015 with no null values.
3. **Status -** This column has character data with only 2 values either 'Developed' or 'Developed' having no null values.
4. **Life.expectancy -** This column has numeric data varying from 36.3 to 89 with 10 null values. This column will be used as a response variable for our fitted model.
5. **Adult.Mortality -** This column has numeric data varying from 1 to 723 with 10 null values. It is define as Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
6. **Infant.deaths -** This column has numeric data varying from 0 to 1800 with no null values. It is define as number of infant deaths per 1000 population.
7. **Alcohol -** This column has Numeric data-type. It is the recorded per capita consumption (in litres of pure alcohol) for age group of more than 15 years with the range of 0.01 to 17.87 with mean 4.6029. It contains a total of 193 null-values.
8. **percentage.expenditure -** This column has Numeric data-type. It records the expenditure on health as a percentage of Gross Domestic Product per capita(%) with the range varing from 0 to 19479.912 with mean 738.251. It does not have any null values.
9. **Hepatitis.B -** This column has Numeric data-type . It records percentage Hepatitis B (HepB) immunization coverage among 1-year-olds with the range of 0 to 99 with mean at 80.94.It has total of 553 null values.
10. **Measles -** This column has Numeric data-type. It records the number of reported cases for Measles per 1000 population with the range of 0 to 212183.0 with mean at 2419.6. It does not have any null values.
11. **BMI -** This column has Numeric data-type. It records the average Body Mass Index of entire population with the range varying from 1 to 87.30 with mean at 38.32. It has total of 34 null values.
12. **under.five.deaths -** This column has Numeric data-type. It records the number of under-five deaths per 1000 population with the range from 0 to 2500 with mean at 42.04.It does not have any null values.
13. **Polio -** This column has numeric data varying from 3 to 99 with 19 null values. It is define as Polio (Pol3) immunization coverage among 1-year-olds (%).
14. **Total.expenditure -** This column has numeric data varying from 0.370 to 17.6 with 226 null values. It is define as General government expenditure on health as a percentage of total government expenditure (%).
15. **Diphtheria -** This attribute represents Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds(%). It is a float type attribute which has 19 null values.The value across world ranges from 2% to 99% with a mean of 82.32%
16. **HIV.AIDS -** This attribute represents Deaths per 1000 live births HIV/AIDS (0-4 years). It is a float type attribute that ranges between 0.1 to 50.6, with a mean value of 1.75. There are no nulls
17. **GDP -** This attribute is Gross Domestic Product per capita (in USD). It is a float column with around 448 null values that can be treated with data imputation methods. The mean value is 7483.16 across countries for those 15 years. We might be interested in year-wise mean here.
18. **Population -** This attribute describes Population of the country. it is a an integer type data with 652 null values that can be treated using mean from subsequent yearly population of that country.
19. **thinness..1.19.years -** This attribute represents Prevalence of malnutrition among children and adolescents for Age 10 to 19 in percentage. It is a float type attribute that varies between 0.1% to 27.7% with a mean of 4.84%. it has 34 null values.

20. **thinness.5.9.years -** This atttribute represents Prevalence of malnutrition among children for Age 5 to 9(%). It is a float % datatype.It varies in data from 0.1% to 28.6% with a mean value of 4.87%. We have around 34 nulls.

21. **Income Composition of resources -** This float type attribute represents Human Development Index in terms of income composition of resources (index ranging from 0 to 1). Our data lies between 0 and 0.95 with a mean value of 0.6. We have 167 null values.

22. **Schooling -** This attributes represents Number of years of Schooling(years). It is a float type attribute that varies between 0 and 20.7 with a mean of 11.99 years. We have around 163 null values that can be treated if required.