Analysis of Factors affecting Life Expectancy

Beria, Anisha

Abstract— This project analyzes a data-set related to life expectancy for 193 countries from the years 2000-2015. The data focuses on economic, mortality, immunization, social and other health related factors in different countries. By analyzing the observations in this data-set, I aim to find the predictive factors that result in higher/lower values of life expectancy in both developing and developed countries. The analysis will also help determine what areas need improvement in different countries in order to enhance life expectancy of their respective populations.

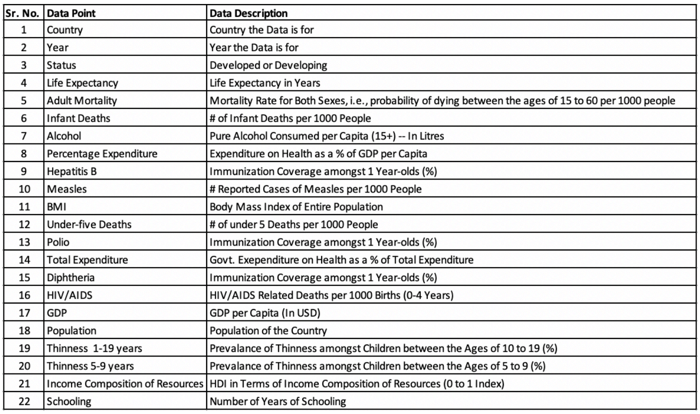
1 Introduction

Life expectancy is a key metric for assessing population health. Unlike infant and child mortality, it captures mortality along the entire life course. In the pre-modern world, life expectancy was around 30 years in most regions of the world. However, life expectancy has rapidly increased since the Age of Enlightenment. In the early 19th century, life expectancy began increasing in industrialized countries and remained low in other parts of the world leading to high inequality in health standards across the globe. Over the last few decades, though this inequality has decreased substantially with countries suffering from poor health standards catching up quickly, it still persists in less developed regions of the world like the Central African Republic (~53 years) whereas, developed regions like Japan have a life expectancy of 83 years. Global life expectancy is now above 70 years.

The aim of this project is to unpack and analyze WHO data of life expectancy for 193 countries from the year 2000- 2015. After a basic description of the data and overarching trends, I will be looking at how certain variables have changed over the course of 15 years, keeping in mind economic, mortality, immunization, social and other health related factors that may have affected these numeric and statistical changes.

2 Description of Data

Data for this project was sourced online from Kaggle and is titled "Life Expectancy Data.csv". The dataset consists of 22 columns and 2938 rows which include 20 predicting variables—1 ordinal variable, 2 nominal variables and 18 continuous numerical variables. Below is a detailed data dictionary that outlines the specificity of the multiple columns:



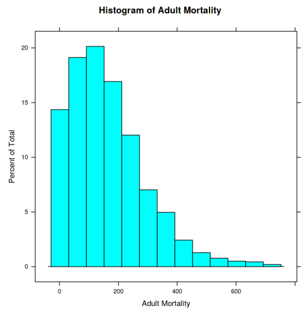
Sr. No.	Data Point	Data Description
1	Country	Country the Data is for
2	Year	Year the Data is for
3	Status	Developed or Developing
4	Life Expectancy	Life Expectancy in Years
5	Adult Mortality	Mortality Rate for Both Sexes, i.e., probability of dying between the ages of 15 to 60 per 1000 people
6	Infant Deaths	# of Infant Deaths per 1000 People
7	Alcohol	Pure Alcohol Consumed per Capita (15+) -- In Litres
8	Percentage Expenditure	Expenditure on Health as a % of GDP per Capita
9	Hepatitis B	Immunization Coverage amongst 1 Year-olds (%)
10	Measles	# Reported Cases of Measles per 1000 People
11	BMI	Body Mass Index of Entire Population
12	Under-five Deaths	# of under 5 Deaths per 1000 People
13	Polio	Immunization Coverage amongst 1 Year-olds (%)
14	Total Expenditure	Govt. Exependiture on Health as a % of Total Expenditure
15	Diphtheria	Immunization Coverage amongst 1 Year-olds (%)
16	HIV/AIDS	HIV/AIDS Related Deaths per 1000 Births (0-4 Years)
17	GDP	GDP per Capita (In USD)
18	Population	Population of the Country
19	Thinness 1-19 years	Prevalance of Thinness amongst Children between the Ages of 10 to 19 (%)
20	Thinness 5-9 years	Prevalance of Thinness amongst Children between the Ages of 5 to 9 (%)
21	Income Composition of Resources	HDI in Terms of Income Composition of Resources (0 to 1 Index)
22	Schooling	Number of Years of Schooling

The dataset had 2563 missing values. The 10 missing values for life expectancy and adult mortality are from 10 countries for which data points are only available for one year. Thus, I removed these 10 countries from my analysis. Next, I created a dummy variable to record the status of a country where, 1 indicates a developed country and 0 indicates a developing country.

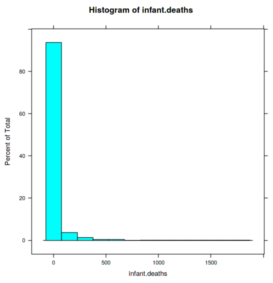
Univariate Analysis:

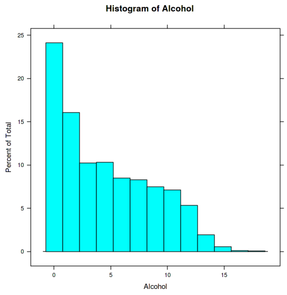
I created histograms for the multiple variables to analyze data distribution: -

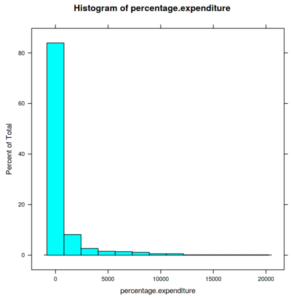
The histogram for *adult mortality* indicated a somewhat normal distribution.



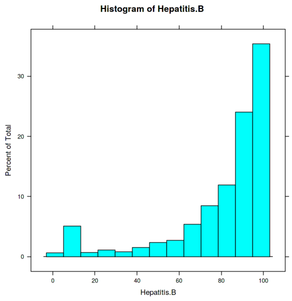
Infant Death, *Alcohol* and *Percentage Expenditure* were right-skewed with most of the data concentrated on the left side of the plots.



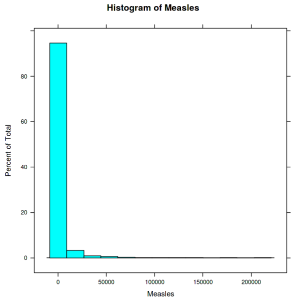




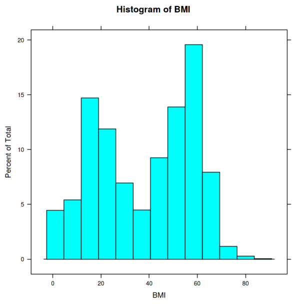
Hepatitis B was left-skewed, with most of the data concentrated on the right side of the plot.



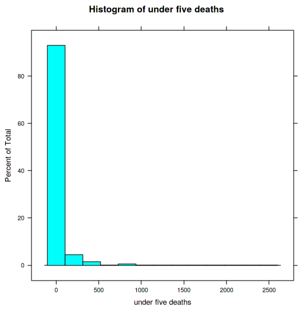
Measles was again, right-skewed



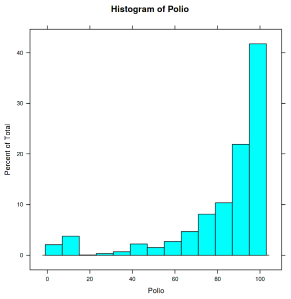
BMI had a bimodal distribution.

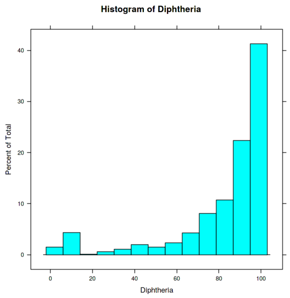


Under 5 Deaths was right-skewed.

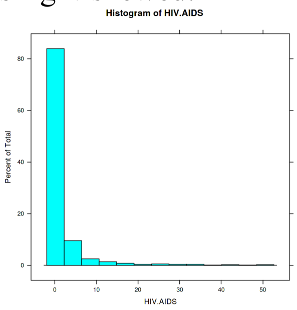


The histograms for *Polio* and *Diphtheria* were left-skewed:

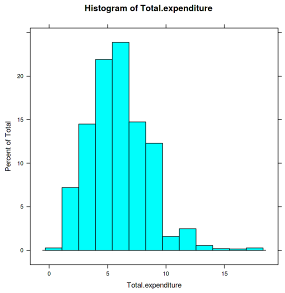


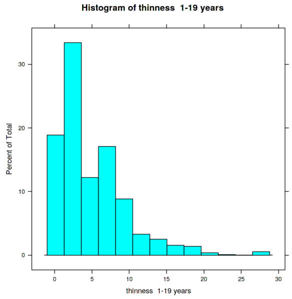


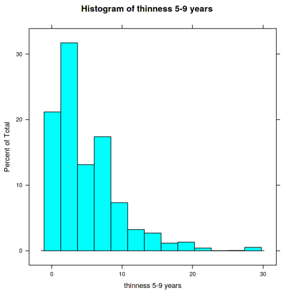
HIV/AIDs was right-skewed:

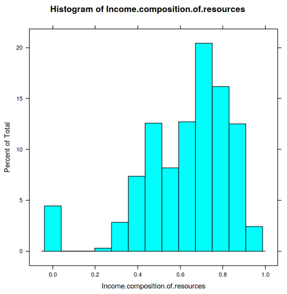


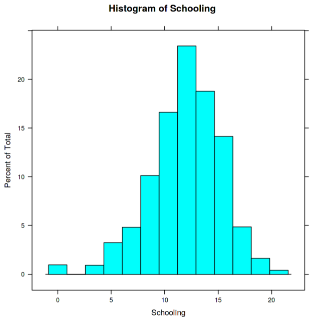
Finally, the histograms for *Total Expenditure, Thinness, Income Composition of Resources* and *Schooling* were all normally distributed:





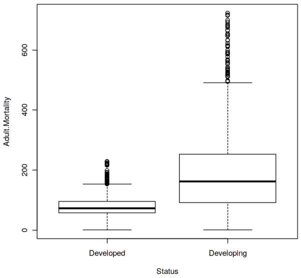


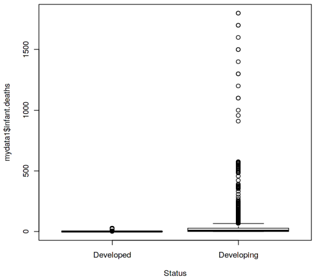


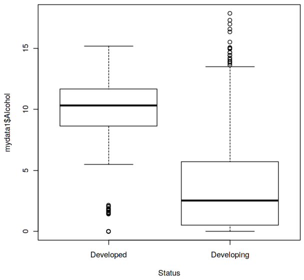


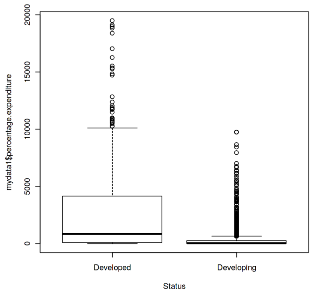
IDENTIFYING OUTLIERS:

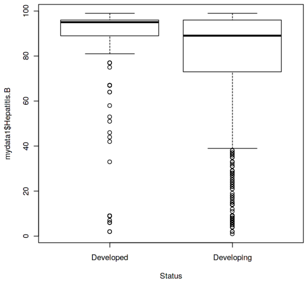
Next, I created boxplots for all the variables to identify outliers:

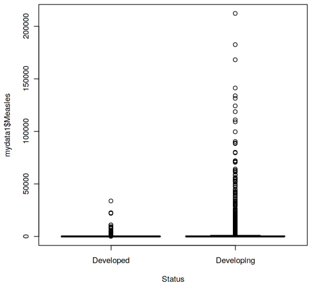


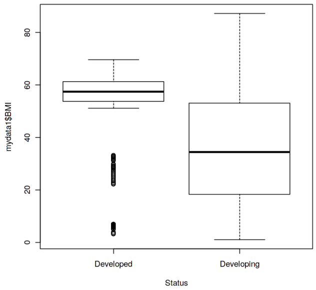


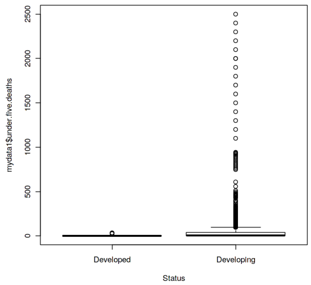


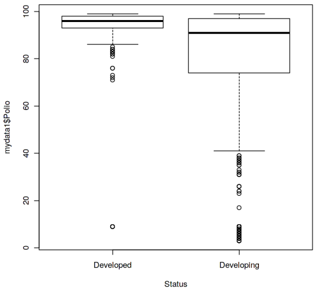


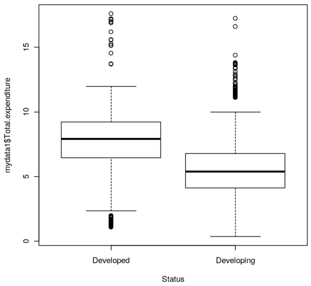


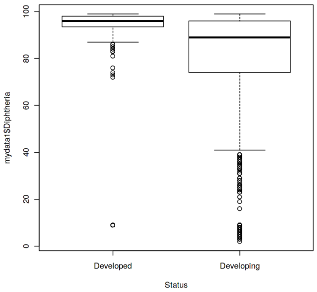


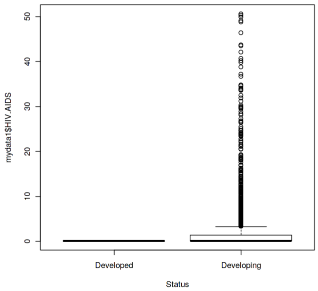


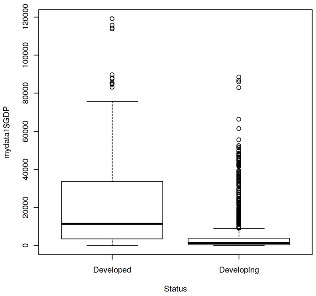


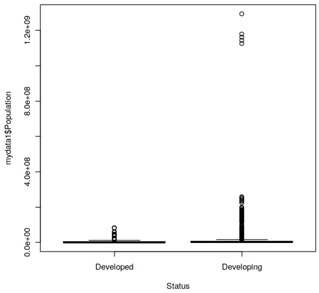


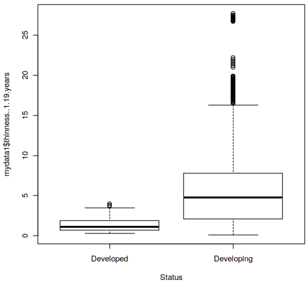


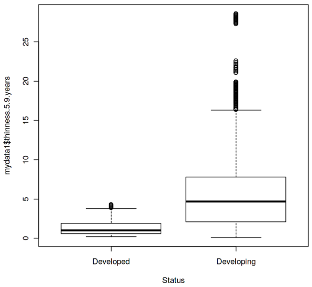


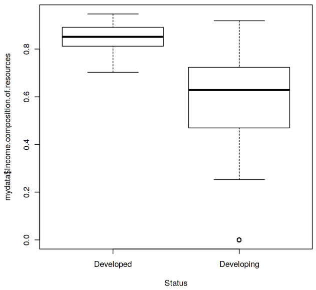


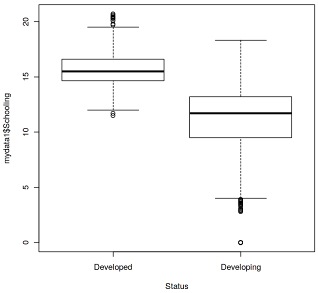








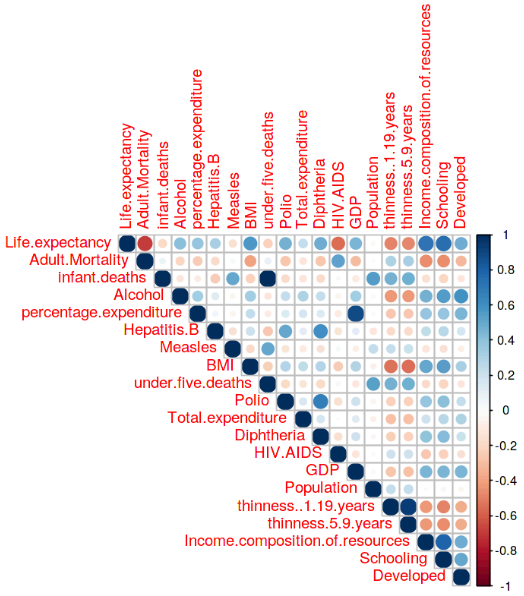




Analzying these boxplots also helped me identify the pattern of predictive factors between developing and developed countries. In general, developed countries have lower mean adult mortality, slightly lower mean infant deaths, higher mean alcohol consumption, higher mean % expenditure, larger mean immunization coverages, higher mean BMI's, lower mean thinness, higher mean GDP, higher mean schooling and lastly, higher mean total expenditure. The mean population between developed and developing countries does not have any significance difference.

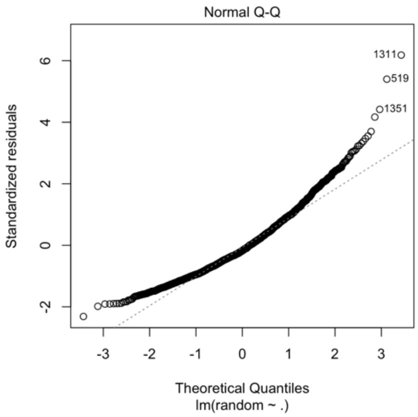
Paramteric Tests:

Additivity is used to verify the correlation between variables. The graph below is a dot representation where blue represents positive correlation and red represents negative correlation. The size of the dot represents the level of correlation, i.e., the larger the dot, the larger is the correlation.



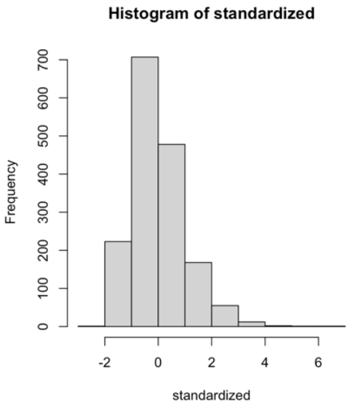
As we can see, life expectancy is negatively correlated to adult mortality, infant deaths, measles, under 5 deaths, thinness and HIV/AIDs. Whereas, it is positively correlated to schooling, income composition of resources, % expenditure, immunization coverages, developed status and BMI. As expected, infant death and under 5 deaths have a very strong positive correlation because they essentially measure the same thing. GDP is highly correlated to % expenditure because a higher GDP natutrally allows for higher healthcare expenditure. Additionally, it is important to make note of *multicollinearity* since correlation exists between independent variables.

Next, I ran the test for *linearity* across the numeric variables. I ran a fake regression and generated a QQ plot to see how linear the variables are, as shown below:



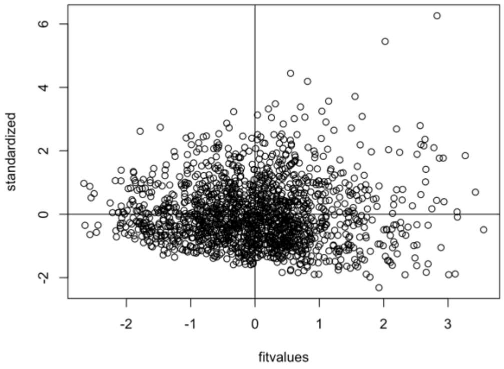
Since the data points are not following the ab line, the variables did not meet the criteria for linearity.

In order to test for *normality*, I ran a histogram for the standardized values of the fake regression references above. I also used skewness, kurtosis, and Mahalanobis analysis to verify the normality of univariate and multivariate scenarios.

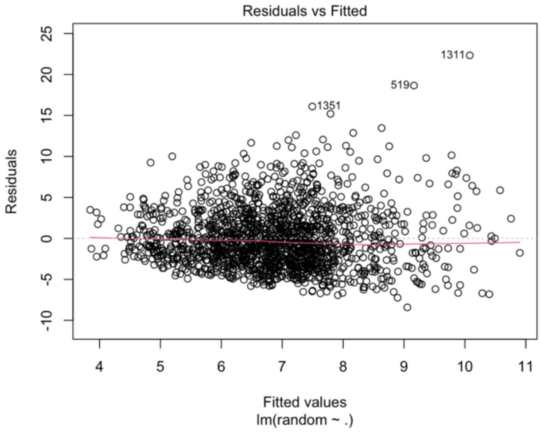


From the histogram, the data appears to be somewhat normally distributed with a slight skewness towards the right.

I also ran a model of the fitted and standardized values to see the distribution along the origin and test for *homogeneity*. From the graph below, we can see that the data is quite homogenous. Besides a few data points, the majority of the data values are quite evenly distributed along 0.



Lastly, I plotted residuals versus fitted values to test for *homoscedasticity*. Since the residuals are equally spread out around the y=0 line, the assumption for homoscedasticity is also met.



3 EXPLORATORY DATA ANALYSIS

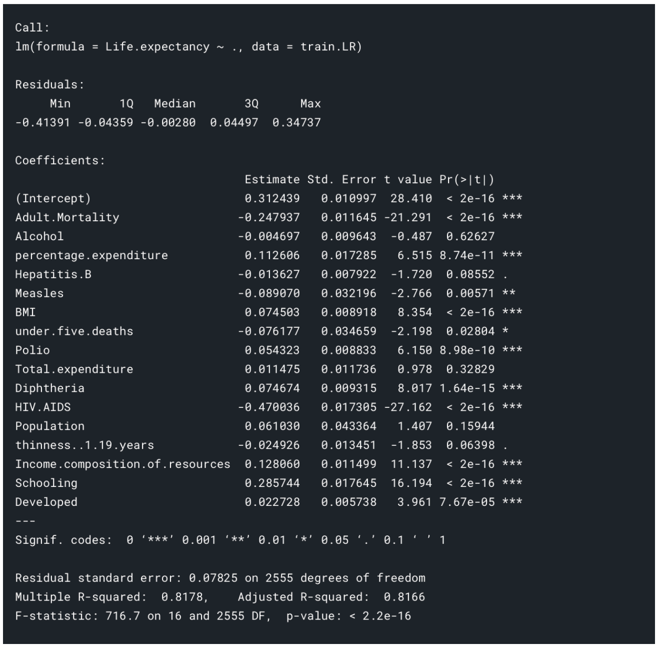
The first thing that I notices during EDA is that life expectancy has clearly increased over the years. In developed countries, vices such as adult mortality, the prevalance of thinness, infant deaths and HIV/AIDs related deaths are significantly lower than those in developing countries. I believe that this difference is what ultimately leads to higher life expectancy in developed nations. Furthermore, pros such as better schooling, larger immunization coverages, income composition of resources, total expenditure and % expenditure are also higher in developed nations. These further cements the gap of of life expectancy between developed and developing countries.

4 MODELLING

Before developing my model, I first performed a stepwise removal of variables using the variance inflation factor (vif) function. This allowed me to identify predicting variables that are more related to other predictors rather than to the response. These variables included infant deaths, GDP and the prevalence of thinness between the ages of 5 to 9, with a vif>5 and were thus, removed from the final model.

Next, I proceeded to partition my data into train and test sets. The train set comprised of data between the years 2000 to 2013 (2572 rows and 19 columns), whereas, the test set comprised of data between 2014 to 2015 (366 rows and 19 columns). I also performed data scaling on all the numeric variables.

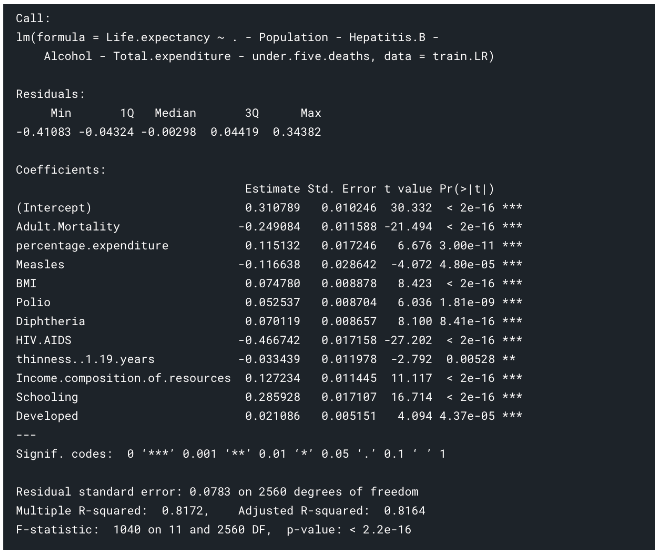
I ran a *linear regression model* on the data sets and first, derived a base model with all the variables:



```
Call:
lm(formula = Life.expectancy ~ ., data = train.LR)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41391 -0.04359 -0.00280  0.04497  0.34737 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     0.312439   0.010997  28.410  < 2e-16 ***
Adult.Mortality                -0.247937   0.011645 -21.291  < 2e-16 ***
Alcohol                        -0.004697   0.009643  -0.487  0.62627    
percentage.expenditure          0.112606   0.017285   6.515 8.74e-11 ***
Hepatitis.B                    -0.013627   0.007922  -1.720  0.08552 .  
Measles                        -0.089070   0.032196  -2.766  0.00571 ** 
BMI                             0.074503   0.008918   8.354  < 2e-16 ***
under.five.deaths              -0.076177   0.034659  -2.198  0.02804 *  
Polio                           0.054323   0.008833   6.150 8.98e-10 ***
Total.expenditure               0.011475   0.011736   0.978  0.32829    
Diphtheria                      0.074674   0.009315   8.017 1.64e-15 ***
HIV.AIDS                       -0.470036   0.017305 -27.162  < 2e-16 ***
Population                      0.061030   0.043364   1.407  0.15944    
thinness..1.19.years           -0.024926   0.013451  -1.853  0.06398 .  
Income.composition.of.resources  0.128060   0.011499  11.137  < 2e-16 ***
Schooling                       0.285744   0.017645  16.194  < 2e-16 ***
Developed                       0.022728   0.005738   3.961 7.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07825 on 2555 degrees of freedom
Multiple R-squared:  0.8178,    Adjusted R-squared:  0.8166 
F-statistic: 716.7 on 16 and 2555 DF,  p-value: < 2.2e-16
```

As you can see, total expenditure, hepatitis B, population and thinness between 1 to 19 are not significant. Therefore, I chose to move forward with only the top 5 variables affecting life expectancy—(i) Adult Mortality; (ii) BMI; (iii)

HIV/AIDs; (iv) Schooling and; (v) Income Composition of Resources.



```
Call:
lm(formula = Life.expectancy ~ . - Population - Hepatitis.B -
    Alcohol - Total.expenditure - under.five.deaths, data = train.LR)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41083 -0.04324 -0.00298  0.04419  0.34382

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.310789   0.010246  30.332  < 2e-16 ***
Adult.Mortality                -0.249084   0.011588 -21.494  < 2e-16 ***
percentage.expenditure          0.115132   0.017246   6.676 3.00e-11 ***
Measles                        -0.116638   0.028642  -4.072 4.80e-05 ***
BMI                             0.074780   0.008878   8.423  < 2e-16 ***
Polio                           0.052537   0.008704   6.036 1.81e-09 ***
Diphtheria                      0.070119   0.008657   8.100 8.41e-16 ***
HIV.AIDS                       -0.466742   0.017158 -27.202  < 2e-16 ***
thinness..1.19.years           -0.033439   0.011978  -2.792  0.00528 **
Income.composition.of.resources  0.127234   0.011445  11.117  < 2e-16 ***
Schooling                       0.285928   0.017107  16.714  < 2e-16 ***
Developed                       0.021086   0.005151   4.094 4.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0783 on 2560 degrees of freedom
Multiple R-squared:  0.8172,	Adjusted R-squared:  0.8164
F-statistic:  1040 on 11 and 2560 DF,  p-value: < 2.2e-16
```

PREDICTION:

I used the the predict function in R for my regression model. The mean absolute error (mae) and mean squared error (mse) for the train set were 0.058 and 0.078 respectively. As we know, MAE measures the average magnitude of errors in a set of predictions, without considering their direction. The lower the MSE, the higher the accuracy of prediction as there would be an excellent match between the actual and predicted data set. MSE on the other hand, tells us how close a regression line is to a set of points. Again, lower values are better. For my model, I believe that the low values of MAE and MSE coupled with an R2 of ~82% indicate that my model has good predictive power. The MAE and MSE for the test set were 0.064 and 0.087 respectively.

Though my model is relatively accurate, I do believe that a tree regression could potentially give us lower mean absolute errors and thus, will probably have a higher predictive power.

5 CONCLUSIONS

The aim of this report was to see which out of the 20 predicting variables in the dataset truly affect life expectancy and then analyze the correlation between life expectancy and factors such as schooling, immunizations and health care expenditure. As already mentioned earlier, I found that only 5 out of the 20 predicting variables truly affect life expectancy. These 5 variables include —(i) Adult Mortality; (ii) BMI; (iii) HIV/AIDs; (iv) Schooling and; (v) Income Composition of Resources. Lower adult mortality, higher BMI, less HIV/AIDs related deaths, more schooling and a higher income composition of resources leads to higher life expectancy. Naturally, all of these are more prevalent in developed nations as compared to developing nations, and thus, life expectancy is higher. By focusing their resources towards these predictive variables, developing nations can increase the life expectancy of their population.

REFERENCES

[1] Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2013, May 23).
Life Expectancy. Retrieved December 11, 2020, from https://ourworldindata.org/life-expectancy

[2] KumarRajarshi. (2018, February 10).
Life Expectancy (WHO). Retrieved December 11, 2020, from https://www.kaggle.com/kumarajarshi/life-expectancy-who