Factors Affecting Life Expectancy: A Statistical Analysis

Pratik Prakash Brahmapurkar

Contents

1. Abstract

This report contains a Statistical Analysis of the Factors Influencing the Expected Life Expectancy. The information was gathered from the World Health Organization (WHO) and the United Nations website with the assistance of Deeksha Russell and Duan Wang. From 2000 to 2015, data was collected from 193 nations, covering a time span of 15 years. Out of all the categories of health-related factors, only those key elements that are more representative of the population were selected. Here the target variable for regression model is 'Life Expectancy' in the report. The report further explains the background of the research problem, step by step design, Results section which claim the actual output and finally the summarized conclusion.

2. Background

The word "life expectancy" refers to the number of years that an individual might reasonably expect living. By definition, life expectancy is a projection of the average age at which members of a certain population group will die. (Ortiz-Ospina, 2017)

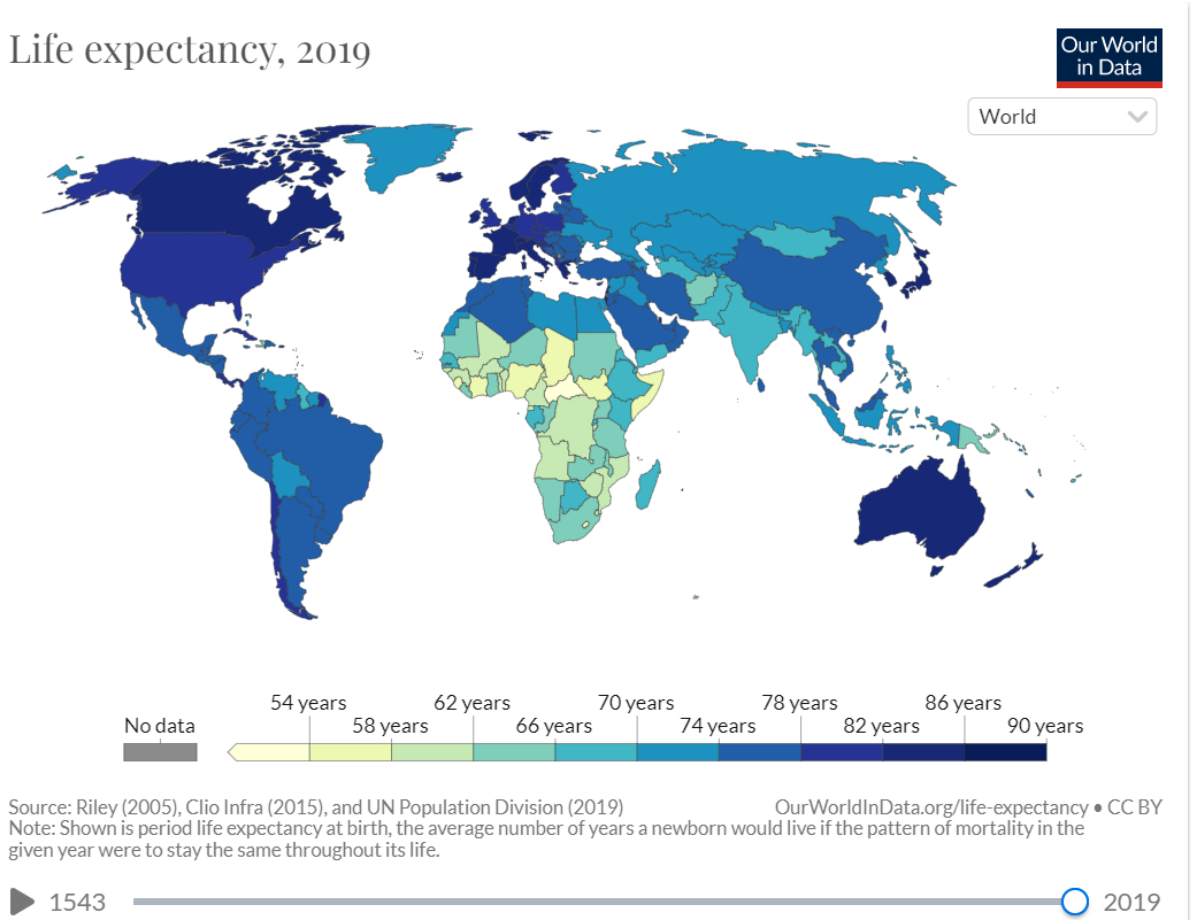
Since the Age of Enlightenment, life expectancy has climbed considerably. Life expectancy began to rise in the early industrialised countries in the early nineteenth century, but remained low in the rest of the world. This resulted in extremely significant inequality in the distribution of health throughout the planet. Excellent health in developed countries and continuously poor health in developing countries. Global inequality has declined during the previous few decades. (Roser, et al., 2013)

As a measure of a country's progress, life expectancy at birth has risen in most countries over the last ten years. Over the years, there have been big changes in things like poverty, nutrition, adult literacy, and access to safe drinking water, the burden of diseases, and sanitation. These changes would have had a positive effect on life expectancy. It's not just Sub-Saharan Africa where life expectancy has been going down, though. In many countries in the developing world, this has been happening. In some of the countries although income and health expenditure is increasing, but their life expectancy is falling down. (Kabir, 2008)

According to (Martikainen, et al., 2014) , In all high-income countries, there are big differences in mortality between rich and poor people. At 35, men at the top of the social ladder have a life expectancy 5 to 10 years longer than men at the bottom of the social ladder. This is like the difference between being a life-long smoker and never smoking. These differences are seen for all-cause mortality and for most of the specific causes of death. Many behavioural risk factors, such as drinking too much alcohol or smoking, are well-known causes of death. Usually, people who are less well-off smoke more than people who are better-off. Harmful alcohol use among. Men are more common in lower social classes, and women are less common. There is less consistency in the way that women do things.

As per (Nixon & Ulmann, 2006), Evidence shows that there are long-term trends in industrialised countries for better health outcomes and more money spent on health care. The average infant mortality rate in the countries of the European Union (EU) went from 3.3 deaths per 1,000 live births in 1960 to 0.6 deaths per 1,000 live births in 1995. The average life expectancy at birth for females went up from 72.5 to 80 years; the average male life expectancy at birth went from 67.6 to 73.6 years; and the total health expenditure as a share of GDP went from 3.4 to 7.7 percent.

Japan has the longest life expectancy at birth in the world, with men and women both averaging 82.0 years of life when they were born even after having world war 2. High-quality health care, a national health insurance system and a strong economy have all been linked to Japan's long life expectancy. Nutrition and environmental factors have been linked to the country's long life expectancy too. Analysis of time trends found that the rise in life expectancy preceded the rise in per capita GDP by 10 years. In education, almost everyone went to elementary school and more people went to high school for both sexes. This was linked to better health. (Sugiura, et al., 2007)



Visualisation 1. Shows world map of life expectancy in 2019 (Ortiz-Ospina, 2017)

According to (Roser, et al., 2013) In 2017, there were 56.5 million deaths worldwide; slightly more than half of these were people over the age of 70; 26% were people between the ages of 50 and 69; 13% were people between the ages of 15 and 49; only 1% were people older than 5 and younger than 14; and nearly 9% were children under the age of 5. Since 1990, there has been a dramatic shift in the average age at which individuals die. People are living longer lives and dying at a younger age. In 1990, children under the age of five accounted for roughly one-quarter of all fatalities. In 2019, this figure has dropped to little less than 9 percent. During this time span, however, the proportion of deaths occurring in the over-70s age category has climbed from a third to half of all fatalities.

3. Design and Methodology

A. Data Collection and loading the data into R-studio

The Life Expectancy (WHO) data is downloaded from Kaggle in .csv format. The work directory in R-studio has been set to the location where the World Health Organization's Life Expectancy data is saved. After that, the csv file is read into R-studio naming the data frame as 'life'. 2938 observations are included inside the data frame, which further comprises 22 attributes. In addition, the necessary libraries have been pre-loaded.

B. Analysing and Summarising the Data

The summary function is responsible for doing an analysis of the whole data set, 'life.' It should be noted that there are a significant number of NA values contained in the data. Using the sapply() method, it has also been possible to examine the mean summary. The describe() function has also revealed the range, skew, and kurtosis of the data.

C. Data Cleaning

The total number of NA values in the data set is around 2563. However, owing to inconsistency and incorrect data in a few columns, we will not consider them in this situation. Some attributes may be removed, and the NA values in those attributes can be replaced by the mean or median value in the attribute.

Life.expectancy	Adult.Mortality	Alcohol	Hepatitis.B	BMI	Polio	Total.expenditure	Diphtheria	GDP	Population	thinness	Income.composition	Schooling
10	10	194	553	34	19	226	19	448	652	34	167	163

Table 1. Shows columns consisting of NA values

1. infant.deaths: Infant deaths as 0 per 1000 seems like incorrect data for many countries. It won't be possible for to have infant rate zero for 848 rows. This column is highly unrealistic. So, it won't be considered for analyzing purpose.
2. BMI: The minimum BMI in the data set is 1 and maximum is value of BMI is 87.3. This column is highly fluctuate, in real life this won't be possible because BMI below 18.5 is underweight and BMI above 30 is obese. (Prevention, 2021) So, BMI column would be of no use.
3. under.five.deaths: 0 deaths under five seems unrealistic, same like Infant deaths. So, this column won't be considered for analysis.
4. Hepatitis.B: For Hepatitis.B there are 553 NA rows. So, due to insufficient data this column can be ignored.
5. Population: There are many outliers in population column. Here 34 is the lowest population and 1.29 Billion is the highest. So, anything below 15000 is been replaced by the median value.
6. Life.expectancy: Life Expectancy column is a target variable. So, the NA values are removed from the column.
7. For Alcohol, Polio, Total.expenditure, Schooling, Diphtheria, Thinness..1.19.years and 5.1 years the NA values are replaced by the mean or median values.
8. Column Country and Status are nominal variable and they converted into factors.
9. The final dataset consists of 1789 observations and 19 variables. It is named as 'lifex'.

D. Visualizing the Data

While data visualisation is essential for communicating insights, it is also a strong approach for exploratory data analysis. For new and current data, data visualisation is the simplest and most efficient tool for understanding the data. In this case data has been visualised using a library called 'ggplot2'. There are different types of plots available in ggplot library. Plots used for analysing purpose are Points, Boxplot and Histogram.

Points: Scatterplots are created with the help of the point geom. In order to visualise the connection between two continuous variables, the scatterplot is the most effective.

Boxplot: Boxplots show how evenly distributed data is in a collection. It quartiles the data. Data set maximum, median, minimum, first and third quartiles are shown in boxplot.

Histogram: A histogram is a graphical depiction of the estimation of numerical data. In a histogram, every bar represents a grouping of numbers into a range.

While visualising the data mainly it is been visualisied with respect to country status whether the country is developed or developing and the attribute life expectancy.

E. Correlation

A correlation coefficient is used to determine the degree of relationship between two variables. It is referred to as Pearson's correlation coefficient in certain areas. The relationship between Life Expectancy and other factors is explored in this study. The graph of the correlation matrix is shown with the help of the R corrplot function. When working with a correlation matrix, Corrplot offers a visual exploration tool that allows automated variable reordering to assist in the identification of hidden patterns within variables.

F. Splitting the dataset into training and test set

The data is divided into two parts in the ratio of 80:20 into two sets: a training set and a test set. To begin, the Index must be trained using the createDataPartition() method, with the target value Life.expectancy serving as the training variable. We'll keep p=0.8 since we want 80 percent of the training set and the remaining 20 percent for the test set, respectively.

G. Multiple Linear Regression

When trying to interpret the multiple linear regression model, the very first step is to determine which attributes in the Life Expectancy (WHO) data set should be taken into consideration while considering 'Life.Expectancy' as dependent variable. The independent variables were chosen by examining whether numerical variables had a strong correlation with the 'Life.Expectancy' in order to determine their significance. The variables which were chosen for all the models are:
"Year","Adult.Mortality","Alcohol","percentage.expenditure","Polio","Diphtheria","Schooling","HIV.AIDS", "Income.composition.of.resources", "Total.expenditure"
Mean Average Error (MAE), Root Mean Squared Error (RMSE) and R-Square is obtained by running this model. Variance inflation factor (or VIF), which evaluates how much the variance of a regression coefficient is exaggerated as a result of multicollinearity in the model, is computed for the Multiple linear regression model.

H. K-fold Cross Validation

Cross-validation is a resampling process that is used to assess machine learning models on a small sample of data. It is also known as cross validation. The process contains a single parameter, denoted by the letter k, which refers to the number of groups into which a given data sample should be divided. As a result, the process is also referred to as k-fold cross-validation. When a particular number for k is specified, it may be substituted for k in the reference to the model, with k=10 becoming 10-fold cross-validation. With the support of the boot library, the cv. glm() method is used. When applied to generalised linear models, this function computes the estimated K-fold cross-validation prediction error. Here, FOR loop is been applied 10 times, to get a proper cross validation error.

I. Best Subset Selection

When attempting to predict an outcome, the best subset selection approach seeks to identify the subset of independent variables that best predicts the result. It does this by examining all potential combinations of independent variables. In R-code regsubsets() function is used with dependent variable Life.expectancy and other independent variables. It is necessary to provide the variable nvmax, which specifies the maximum number of

predictors that may be included in the model. In this case nvmax = 5 indicates that the function will yield the best model with up to five variables, and so on.

J. Lasso and Ridge Regression

Lasso and Ridge regression are a technique that may be used to fit a regression model when there is a high degree of multicollinearity in the data. The ridge regression model is fitted with the help of the function glmnet(), with alpha=0 being specified. Furthermore, adjusting alpha = 1 is similar to using the Lasso Regression technique. X and Y variables are already predefined with the model matrix and target variable (Life Expectancy) respectively.
The optimal (best) lambda value for Ridge and Lasso tests that minimises the mean square error (MSE) is obtained. Finally, the R-squared of the model on the training data is calculated.

K. **Principal Components Regression**

In contrast to normal linear regression, Principal Component Regression (PCR) is a regression approach that has the same purpose as standard linear regression: to model the connection between a target variable and the predictor variables. In R-programming pcr() function is used with the help of pls package. The scale value is denoted as TRUE, this instructs R to scale each predictor variable's mean to 0 and standard deviation to 1. This ensures that if a predictor variable is measured in different units, it does not dominate the model. And the value of validation is CV, this instructs R to do k-fold cross-validation on the model. By default, this utilises k=10 folds.

4. Results and discussion

A. Summary after cleaning the Data

```
> summary(lifex)
       Country          Year            Status     Life.expectancy Adult.Mortality    Alcohol       percentage.expenditure
 Afghanistan:  16   Min.   :2000   Developed : 261   Min.   :44.00   Min.   :  1.0   Min.   : 0.010   Min.   :    0.00
 Albania    :  16   1st Qu.:2005   Developing:1528   1st Qu.:64.30   1st Qu.: 77.0   1st Qu.: 1.010   1st Qu.:   23.97
 Armenia    :  16   Median :2009                     Median :71.70   Median :148.0   Median : 3.755   Median :  104.31
 Austria    :  16   Mean   :2008                     Mean   :69.36   Mean   :167.9   Mean   : 4.459   Mean   :  645.06
 Belarus    :  16   3rd Qu.:2012                     3rd Qu.:75.00   3rd Qu.:227.0   3rd Qu.: 7.010   3rd Qu.:  466.74
 Belgium    :  16   Max.   :2015                     Max.   :89.00   Max.   :723.0   Max.   :17.870   Max.   :18961.35
 (Other)    :1693
  Hepatitis.B       Measles           Polio       Total.expenditure   Diphtheria       HIV.AIDS           GDP
 Min.   : 2.00   Min.   :     0   Min.   : 3.00   Min.   : 0.740    Min.   : 2.00   Min.   : 0.100   Min.   :     1.68
 1st Qu.:75.00   1st Qu.:     0   1st Qu.:81.00   1st Qu.: 4.510    1st Qu.:81.00   1st Qu.: 0.100   1st Qu.:   482.25
 Median :91.00   Median :    15   Median :93.00   Median : 5.755    Median :92.00   Median : 0.100   Median :  1631.42
 Mean   :79.25   Mean   :  2193   Mean   :83.38   Mean   : 5.947    Mean   :84.01   Mean   : 1.889   Mean   :  5573.63
 3rd Qu.:96.00   3rd Qu.:   372   3rd Qu.:97.00   3rd Qu.: 7.340    3rd Qu.:97.00   3rd Qu.: 0.700   3rd Qu.:  4850.00
 Max.   :99.00   Max.   :131441   Max.   :99.00   Max.   :14.390    Max.   :99.00   Max.   :50.600   Max.   :119172.74

   Population         thinness..1.19.years thinness.5.9.years Income.composition.of.resources   Schooling
 Min.   :     15328   Min.   : 0.100       Min.   : 0.100     Min.   :0.0000                  Min.   : 4.20
 1st Qu.:    423724   1st Qu.: 1.600       1st Qu.: 1.700     1st Qu.:0.5090                  1st Qu.:10.30
 Median :   1453684   Median : 3.100       Median : 3.200     Median :0.6730                  Median :12.30
 Mean   :  14561002   Mean   : 4.823       Mean   : 4.883     Mean   :0.6335                  Mean   :12.13
 3rd Qu.:   8121423   3rd Qu.: 7.000       3rd Qu.: 7.000     3rd Qu.:0.7540                  3rd Qu.:14.00
 Max.   :1293859294   Max.   :27.200       Max.   :28.200     Max.   :0.9370                  Max.   :20.70
```

Fig 1. Shows Summary after cleaning the data.

B. Visualization

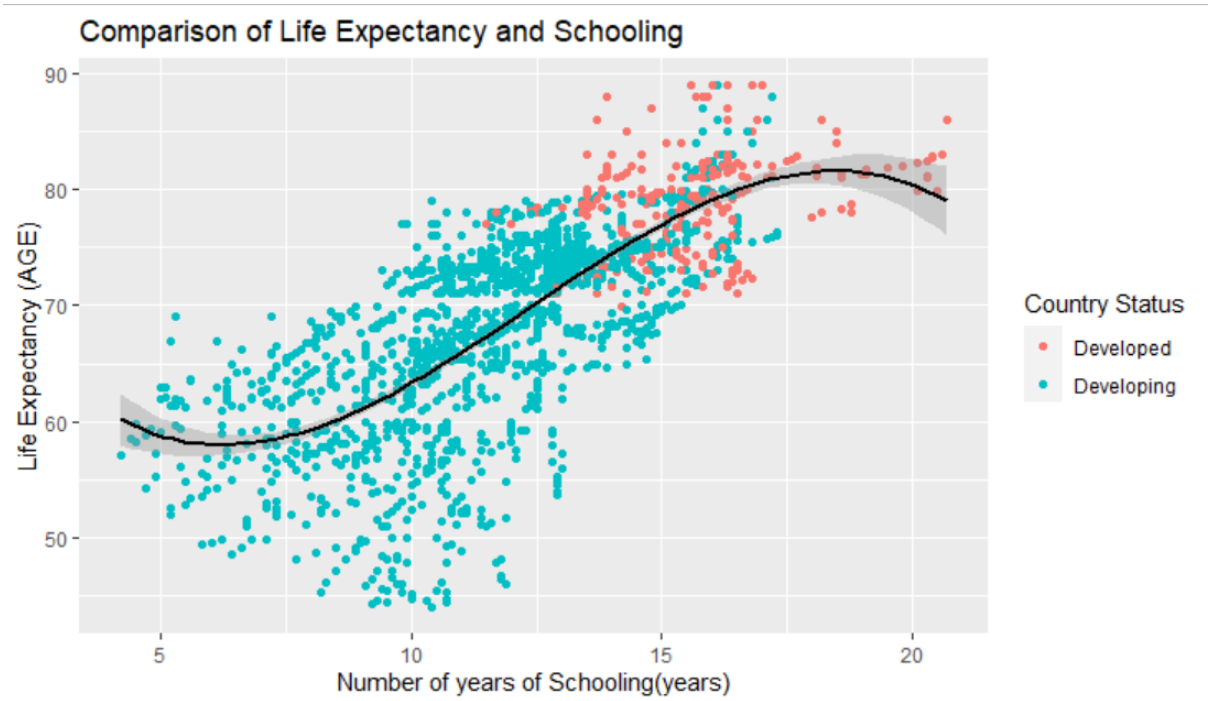


Fig 2. Shows life expectancy rate with respect to Number of years of Schooling (years)

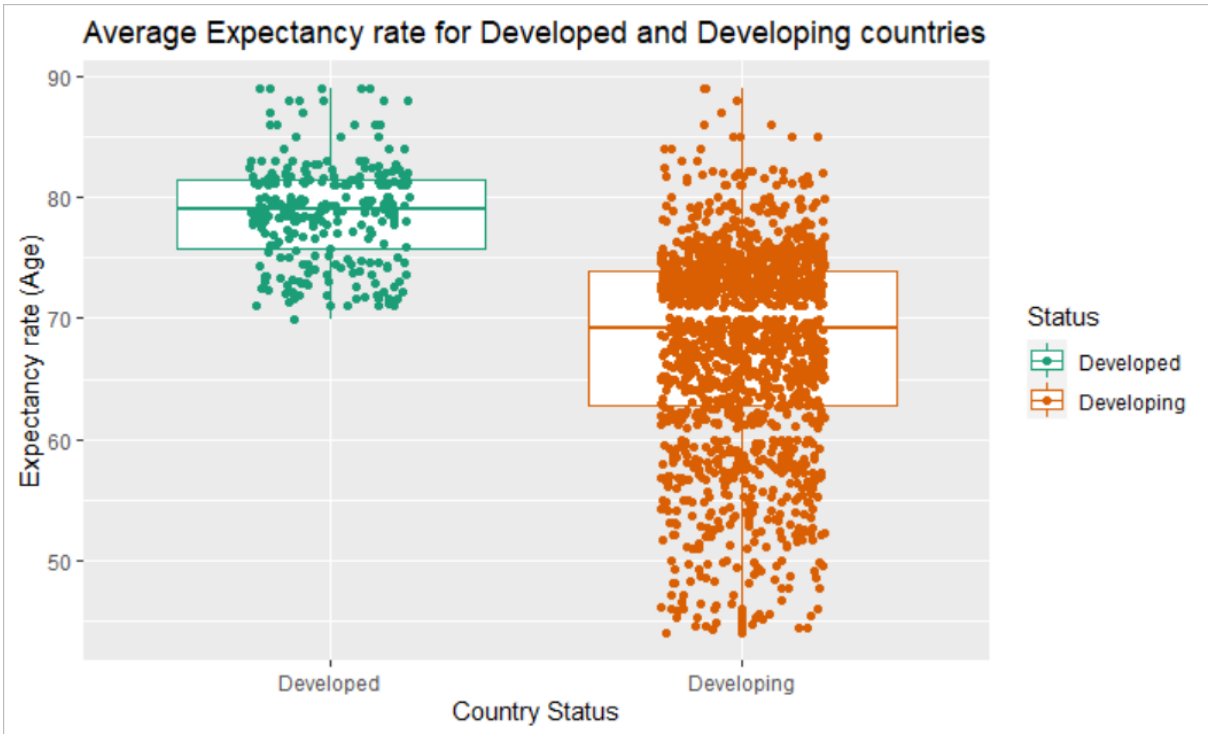


Fig 3. Shows average Expectancy rate for Developed and Developing countries

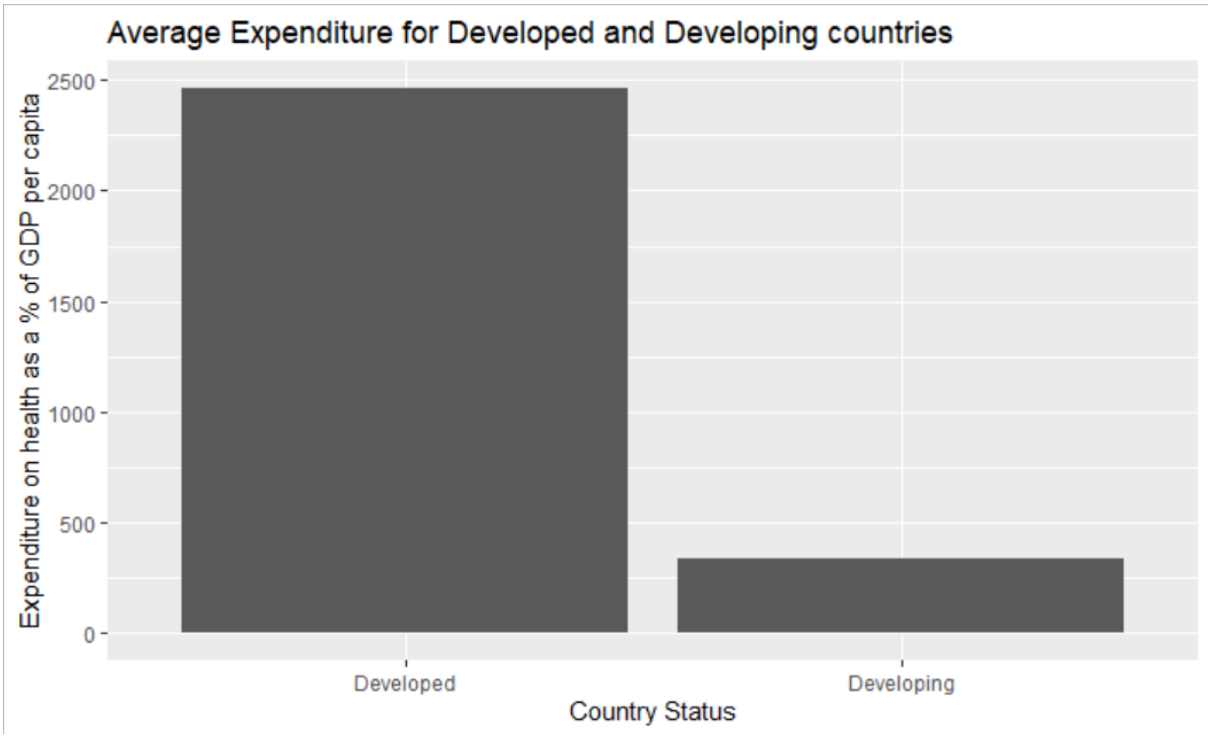


Fig 4. Shows. Expenditure for Developed and Developing countries

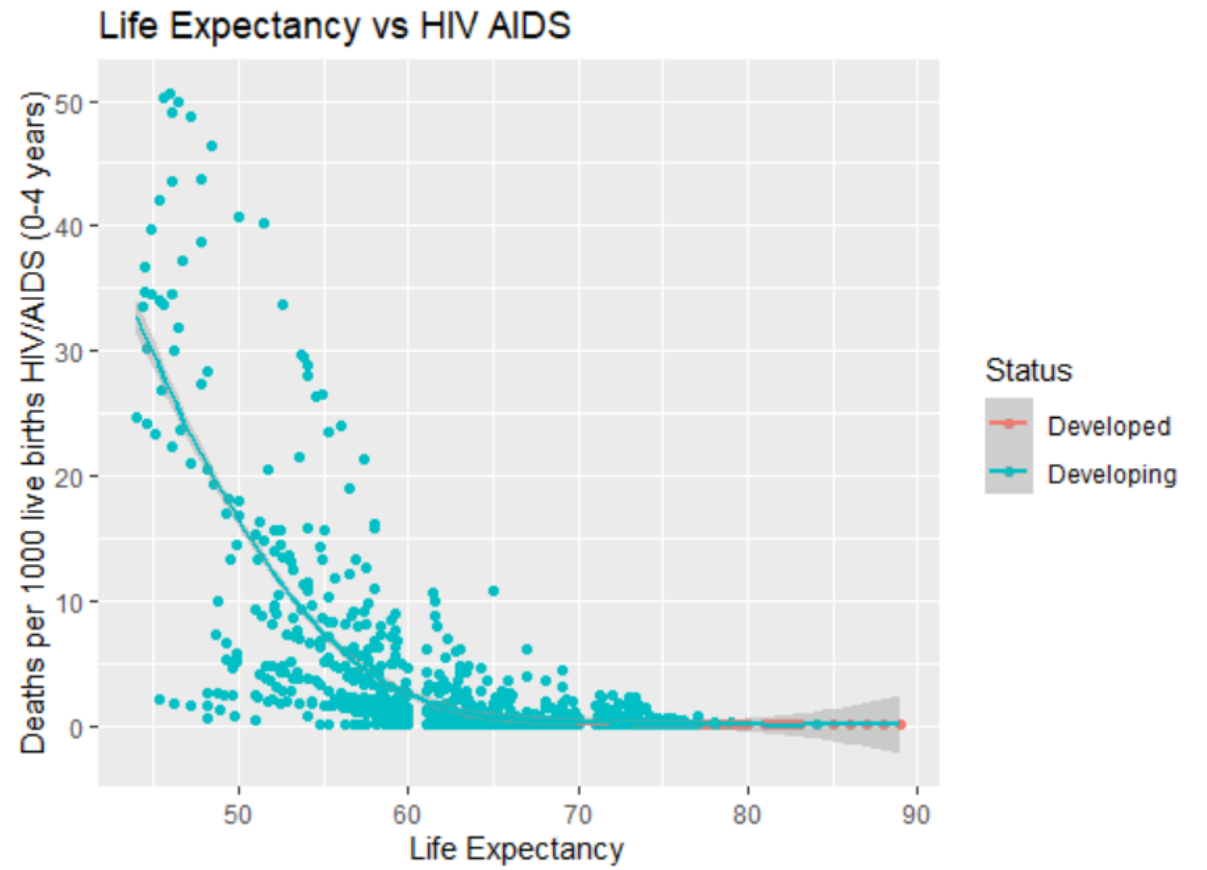


Fig 5. Shows Deaths per 1000 live births HIV/AIDS with respect to Life Expectancy

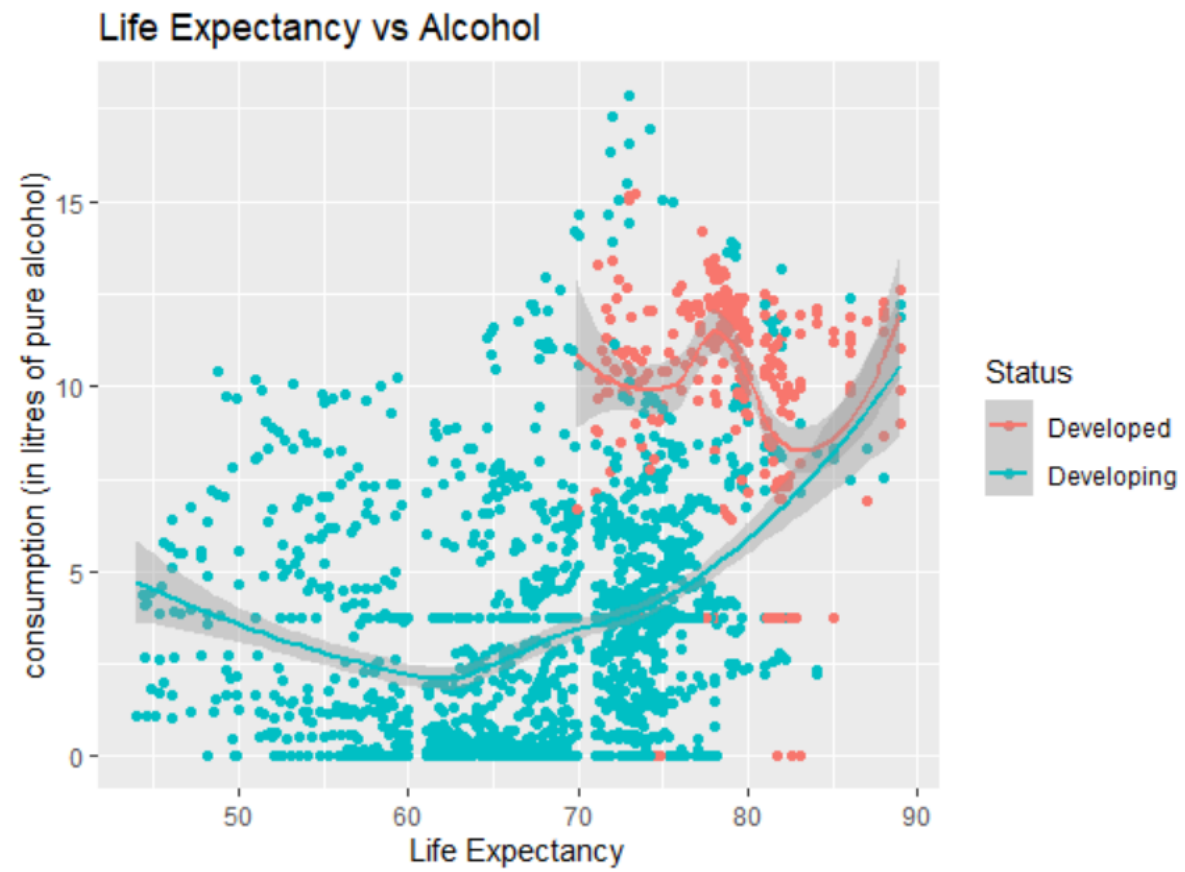


Fig 6. Shows Life Expectancy with respect to Alcohol consumption

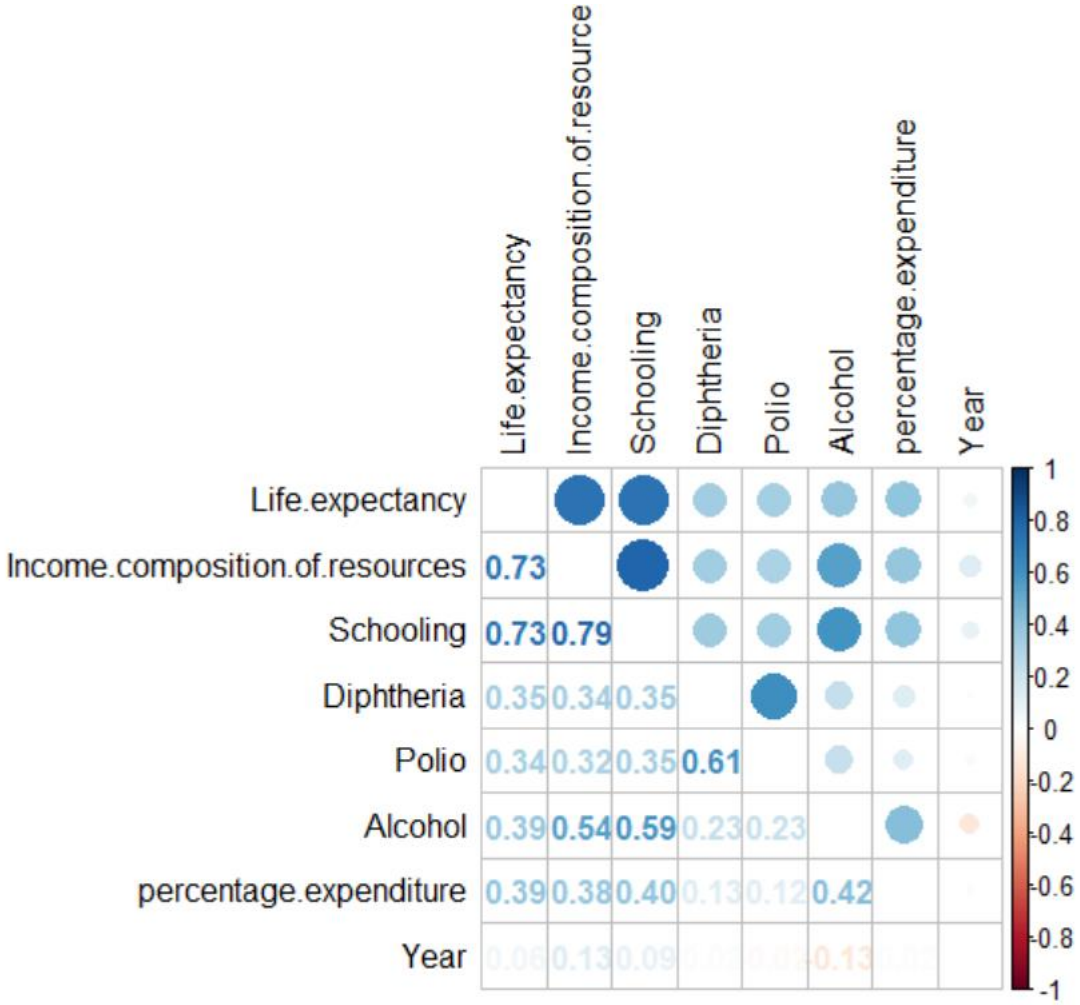


Fig 7. Shows Corrplot for different Variables

C. Multiple Linear Regression

```
Call:
lm(formula = Life.expectancy ~ (Schooling + Diphtheria + Polio +
    Alcohol + percentage.expenditure + Year + Adult.Mortality +
    HIV.AIDS + Income.composition.of.resources + Total.expenditure),
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6747  -2.2036   0.0354   2.3077  11.6023

Coefficients:
                                   Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                     322.89015996  47.66644624   6.774      0.0000000000183 ***
Schooling                         1.04653462   0.06104821  17.143 < 0.0000000000000002 ***
Diphtheria                        0.01865586   0.00574303   3.248              0.00119 **
Polio                             0.00842864   0.00552504   1.526              0.12735
Alcohol                          -0.14602687   0.03385248  -4.314      0.0000171681069 ***
percentage.expenditure            0.00041413   0.00006179   6.703      0.0000000000294 ***
Year                             -0.13564508   0.02379237  -5.701      0.0000000144454 ***
Adult.Mortality                  -0.01806747   0.00105090 -17.192 < 0.0000000000000002 ***
HIV.AIDS                         -0.47893971   0.02058372 -23.268 < 0.0000000000000002 ***
Income.composition.of.resources  12.12161654   0.91818587  13.202 < 0.0000000000000002 ***
Total.expenditure                 0.09087790   0.04571562   1.988              0.04701 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.698 on 1422 degrees of freedom
Multiple R-squared:  0.8231,    Adjusted R-squared:  0.8218
F-statistic: 661.6 on 10 and 1422 DF,  p-value: < 0.00000000000000022
```

Fig 8. Shows Summary for Multiple Linear Regression Model (MLRM)

RMSE	Rsquared	MAE
3.517973	0.834953	2.689395

Table 2. Shows RMSE, R-Square and Mean Absolute Error for MLRM

```
> vif(MLR.model)
                  Schooling                 Diphtheria                      Polio                    Alcohol
                   3.177660                   1.644285                   1.621020                   1.845629
     percentage.expenditure                       Year             Adult.Mortality                   HIV.AIDS
                   1.311786                   1.116404                   1.713115                   1.420623
Income.composition.of.resources        Total.expenditure
                   3.043653                   1.091531
> mean(vif(MLR.model))
[1] 1.79857
```

Fig 9. Shows Variance inflation factor (VIF)

D. Validation Set Approach

```
> cv.error.10 <- rep(0, 10)
> for (i in 1:10) {
+   glm.fit <- glm(Life.expectancy ~ poly(Schooling + Diphtheria + Polio + Alcohol + percentage.expenditure
 + Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources + Total.expenditure, i), data = life
x)
+   cv.error.10[i] <- cv.glm(lifex, glm.fit, K = 10)$delta[1]
+ }
> cv.error.10
 [1]  67.05813  64.56176  64.25924  64.67834  64.31866  70.11950  66.53351  67.16152  97.86232 233.24233
```

Fig 10. Show K-fold Cross Validation

E. Subset Selection Method

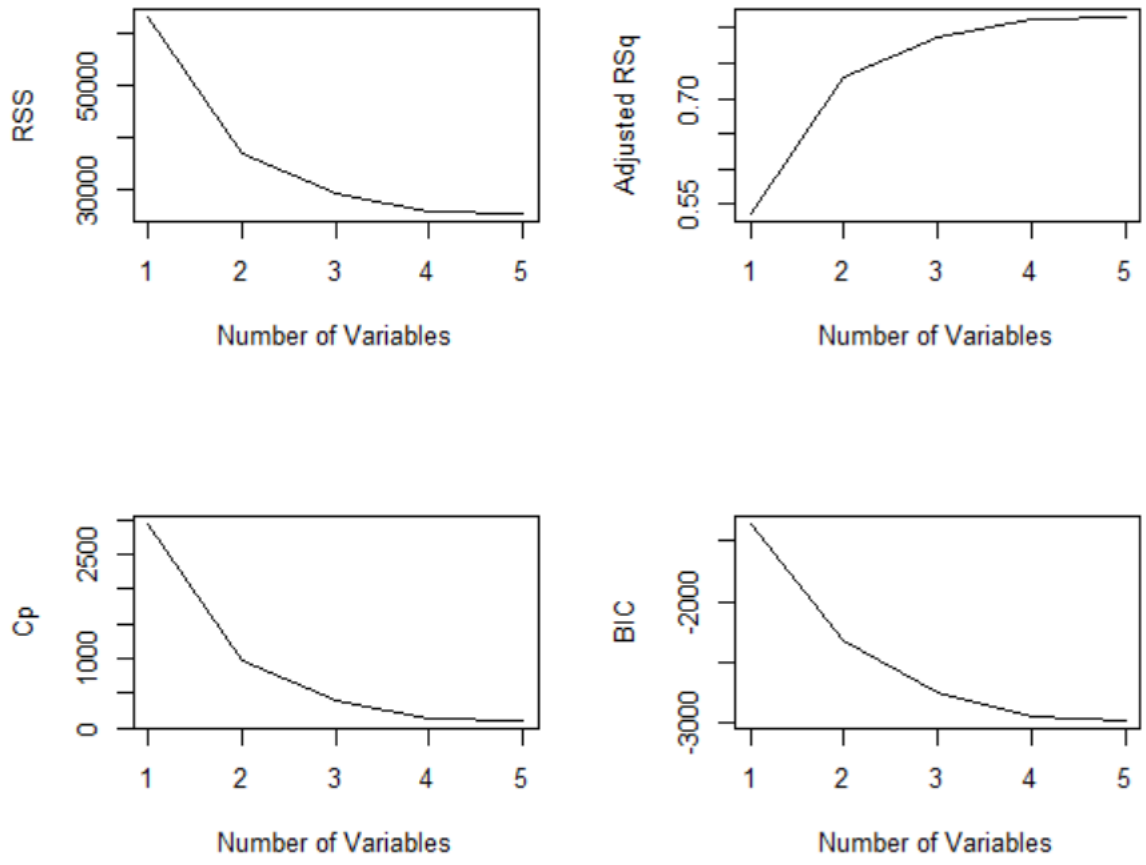


Fig 11. Shows graph of RSS, Adjusted R-Square, Cp and BIC (Bayesian information criteria)

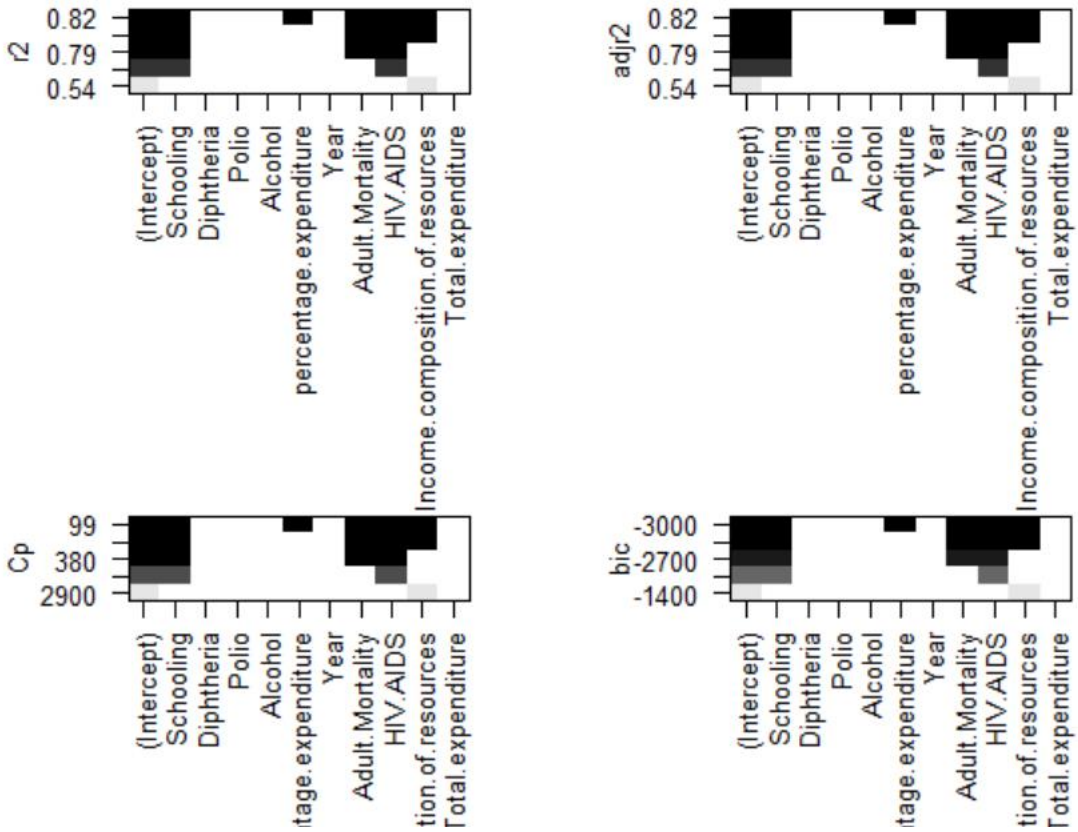


Fig 12. Shows plot of RSS, Adjusted R-Square, Cp and BIC with respect to Variables

```
> coef(regfit.full,5)
                  (Intercept)                      Schooling
                53.3811928207                   1.0250269588
       percentage.expenditure                Adult.Mortality
                 0.0003980492                  -0.0194686186
                    HIV.AIDS Income.composition.of.resources
                -0.4448067429                  11.6797788269
```

Fig 13. Shows Coefficient of Best Subset selection

F. Ridge Regression

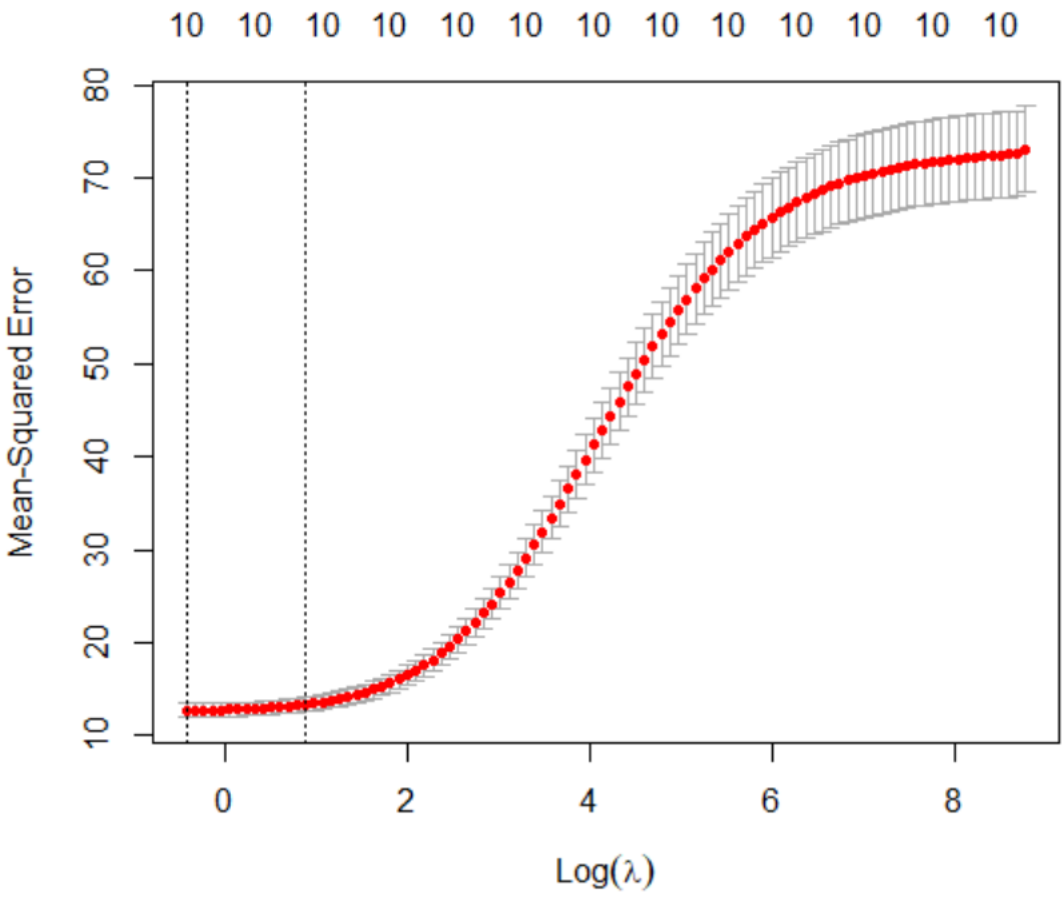


Fig 14. Shows MSE and Lambda Graph

```
> bestlam <- cv.out$lambda.min
> bestlam
[1] 0.6508003
```

Fig 15. Shows the lambda value that minimizes the test MSE

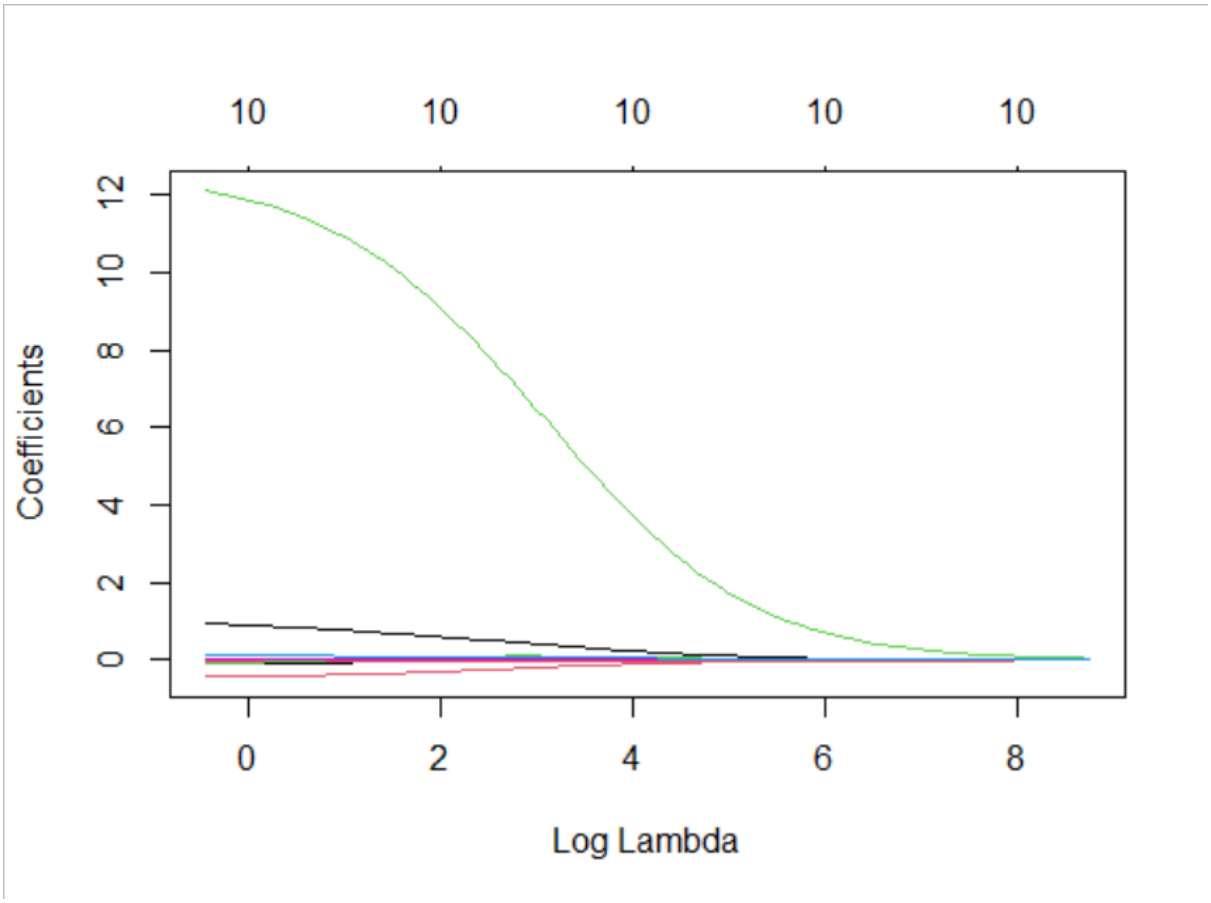


Fig 16. Shows coefficient estimates changed as a result of increasing lambda

```
> y_predicted <- predict(ridge.mod, s = bestlam, newx = x)
> #find SST and SSE
> sst <- sum((y - mean(y))^2)
> sse <- sum((y_predicted - y)^2)
> #find R-Squared
> ridge.rsq <- 1 - sse/sst
> ridge.rsq
[1] 0.8225971
```

Fig 17. Shows R-square for Ridge Regression

G. Lasso Regression

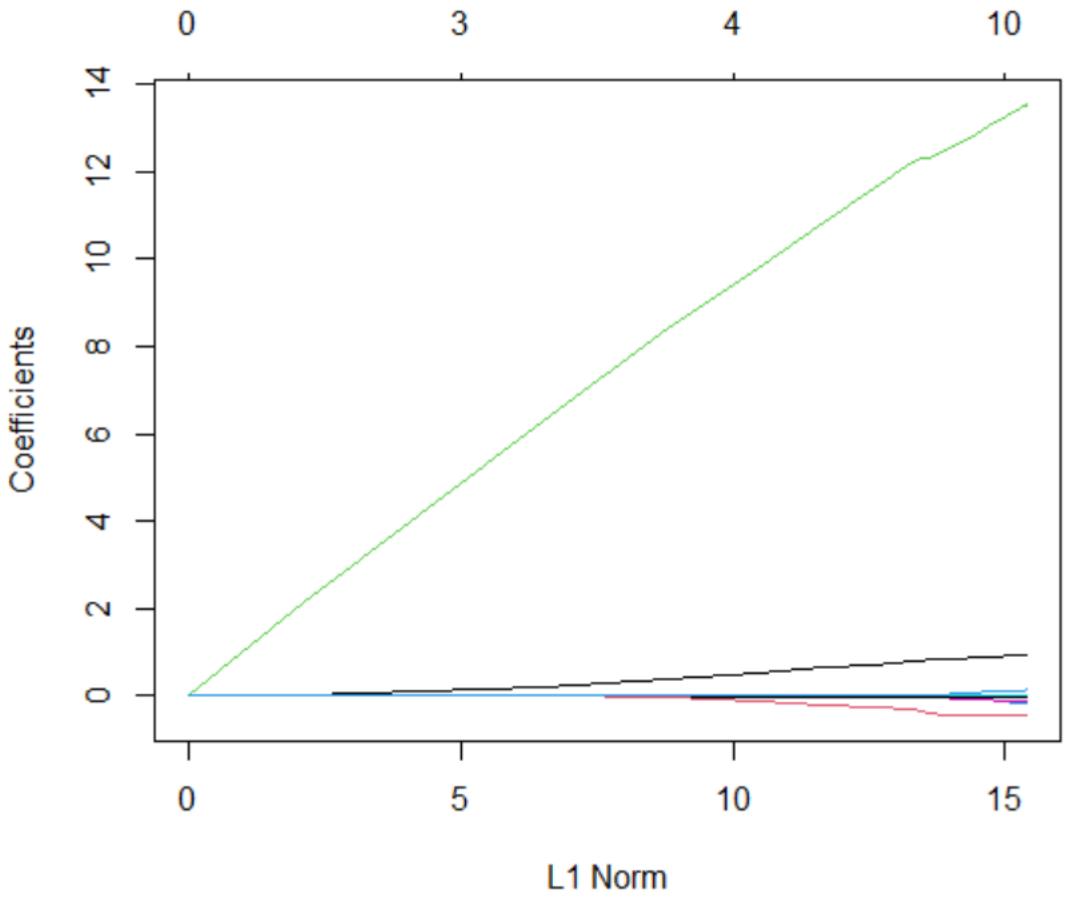
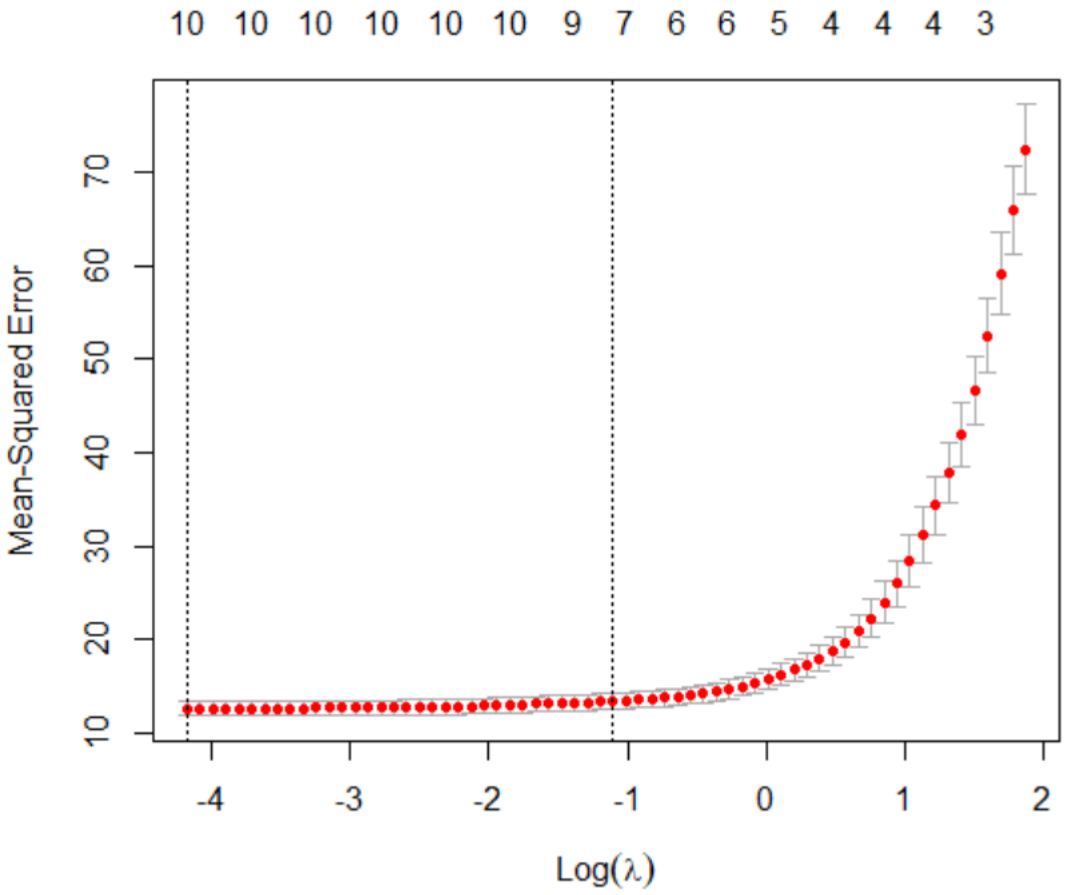


Fig 18. Shows Plot of Lasso model




```
> bestlam <- cv.out$lambda.min
> bestlam
[1] 0.01520234
```

Fig 19. Shows MSE – Lambda graph and best Lambda value for Lasso model

```
> #use fitted best model to make predictions
> y_predicted <- predict(lasso.mod, s = bestlam, newx = x)
> #find SST and SSE
> sst <- sum((y - mean(y))^2)
> sse <- sum((y_predicted - y)^2)
> #find R-Squared
> lasso.rsq <- 1 - sse/sst
> lasso.rsq
[1] 0.8242941
```

Fig 20. Shows R-Square for Lasso

H. **Principal Components Regression**

```
> summary(pcr.fit)
Data:   X dimension: 1789 10
        Y dimension: 1789 1
Fit method: svdpc
Number of components considered: 10

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           8.742    5.026    4.135    4.075    3.947    3.904    3.873
adjCV        8.742    5.026    4.133    4.074    3.946    3.903    3.872
       7 comps  8 comps  9 comps  10 comps
CV        3.85    3.853    3.692     3.689
adjCV     3.85    3.853    3.690     3.687

TRAINING: % variance explained
                1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X                 34.17    48.21    60.67    70.93    79.42    86.12
Life.expectancy   67.00    77.87    78.48    79.82    80.27    80.58
                7 comps  8 comps  9 comps  10 comps
X                 90.71    94.59    97.98    100.00
Life.expectancy   80.96    81.03    82.44     82.57
```

Fig 21. Shows Summary of Model Fitting

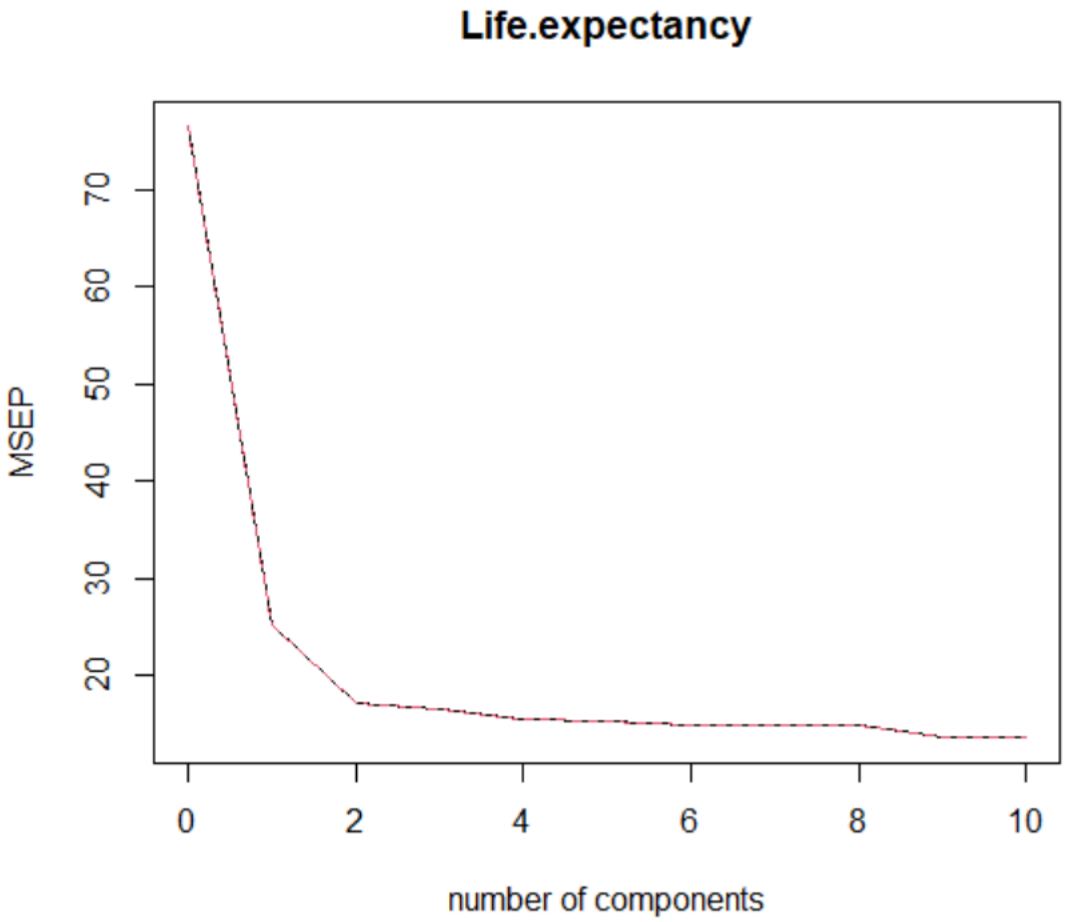


Fig 22. Shows MSEP with respect to number of components

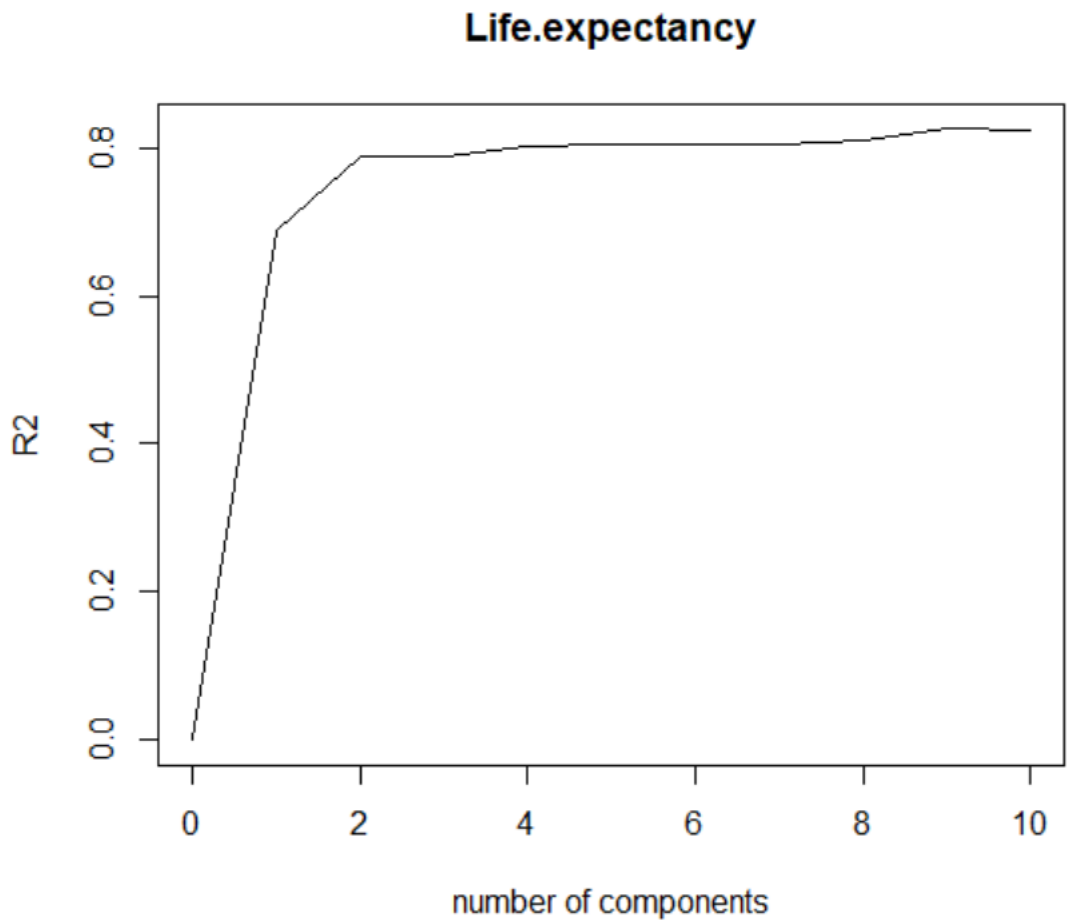


Fig 23. Shows R-Square with respect to number of components

```
> mean((pcr.pred - y.test)^2)
[1] 13.98661
> #calculate RMSE
> sqrt(mean((pcr.pred - y.test)^2))
[1] 3.739868
```

Fig 24. Shows Mean and RMSE value

I. Discussion

By cleaning and analysing the World Health Organization's (WHO) data collection, we can observe that Country Status and Life Expectancy are the most relevant attributes in terms of selecting the most insightful information. The ggplots figure illustrates the majority of the justifications. Figure 2. Shows that developed countries have a better education facilities than the developing ones. Schooling has a significant impact on improving the life expectancy among developing nations since it allows individuals to become considerably more educated and contributes to the improvement of the country's welfare and healthcare, as well as its economy. Figure 4. Shows expenditure on healthcare system by a developed country is much larger as compared to developing countries. So, this might also be one of the reason from Figure 3. that the minimum life expectancy of the developed country is 70 years. Fig 5. Depicts that there are maximum deaths due to HIV/AIDS in developing countries. The trend in figure 6. Depicts that richer nations can afford alcohol, and

alcohol use is more frequent among wealthier inhabitants. As a result, the relationship between developing countries with alcohol is positive, but the relationship between developed countries with alcohol is negative. Fig 7. Shows corrplot where the correlation of variables is being depicted.

Figure 8 shows that the multiple linear regression MLR.model has a significant p-value and an R-Squared value of 0.8231 (82.31 percent). Table 2 shows that the root mean square error is 3.518, which is a significant value. The assumption of no multicollinearity is shown in Fig 9, which has a mean VIF of 1.79. If the VIF had been more than 10, there would have been reason for concern.

Fig. 10 Shows K-fold Cross Validation where the value at 3rd fold tends to be the lowest which is 64.25 and the highest value is at 10th fold which is 233.24. It can be seen that the model with 3 variables is the best model. It has the lower prediction error. It could be used to assess a statistical learning method's test error to evaluate performance or choose the proper amount of flexibility.

Figure 11 and 12 shows plots for Best Subset selection which shows graph of RSS, Adjusted R-Square, Cp and BIC. Figure 13 shows the 5 coefficient of Best Subset selection as nvmax was given as 5.

Figure 14 shows the MSE and Lambda Graph where the best lambda value turns out to be 0.65 for ridge regression, it is the lambda value that minimizes the test MSE. Figure 16 illustrates a trace plot of ridge regression to demonstrate how raising lambda impacted the coefficient estimations. From figure 17. It can also be noted that R-Square value of ridge is 0.82259 (82.259%). Figure 19 shows the MSE and Lambda Graph where the best lambda value turns out to be 0.15 for lasso regression, it is the lambda value that minimizes the test MSE. Figure 18 illustrates a trace plot of lasso regression to demonstrate how raising lambda impacted the coefficient estimations. From figure 20. It can also be noted that R-Square value of ridge is 0.82429 (82.429%).

In Figure 21, you can see a summary of the Principal Components Regression (PCR). The following is the summary for Validation: The RMSEP shows that if we simply utilise the intercept component in the model, the test RMSE is 8.742, which is consistent with the data. When we include the first main component in the test, the RMSE of the test lowers to 5.026. When the second main component is included, the test RMSE decreases to 4.135. The training percent variance table informs us of the proportion of the variation in the response variable that can be explained by the primary components of the model. Just the first principal component may explain 34.17 percent of the variance in the response variable, which is a significant amount of variation. Incorporating the second main component into the model allows us to explain 48.21 percent of the variance in the outcome variable. Figure 23. Shows R-Square with respect to number of components.

Regression	R-Square	RMSE
Multiple Linear	0.835	3.518
Ridge	0.822	3.855
Lasso	0.824	3.834
Principal Components	-	3.739

Table 3. Shows Regression models with RMSE and/or R-Square

5. Conclusion

The dataset collected from WHO had many missing values, and we discovered that the majority of these missing values came from nations with a relatively small population and where data collecting is a time-consuming operation.

From table 3. It can also be observed that the lowest RMSE and highest R-Square value is for multiple linear regression model. The finalized model to analyze the Life expectancy data is Multiple Linear Regression model.

Alcoholism is a major problem in industrialised countries where people have a lot of disposable income, which demonstrates how irresponsible individuals are when it comes to their own wellbeing when it comes to drinking use. It is negatively correlated as per the data provided in the developed countries. The government can increase the subsidy on alcoholic beverages, as well as the number of healthcare and welfare camps, in order to raise awareness among the public about the dangers of binge drinking and the consequences of overindulging.

When we look at illnesses such as HIV/AIDS, polio, Hepatitis B, and Diphtheria, we can find that underdeveloped nations have much lower life expectancies than developed ones because they have good medical treatments available. In order to assist developing countries in eliminating illnesses that are threatening the lives of their citizens, developed countries should provide immunizations. For example covid-19 vaccines were being distributed by developed country or by country who had mass vaccine production. The government should place more emphasis on the education of children, who will one day be the face of the nation, and should ensure that they get nutritious meals and receive a decent education.

Japan, although having been severely damaged by World War II, has recovered very well and is today the nation with the longest life expectancy due to their countries emerging economy. And government looking after their people.

6. References

Kabir, M., 2008. Determinants of Life Expectancy in Developing Countries. *The Journal of Developing Areas,* 41(2), pp. 185-204.

Martikainen, P., Mäkelä, P., Peltonen, R. & Myrskylä, M., 2014. Income Differences in Life Expectancy: The Changing Contribution of Harmful Consumption of Alcohol and Smoking. *JSTOR,* 25(2), pp. 182-190.

Nixon, J. & Ulmann, P., 2006. The relationship between healthcare expenditure and health outcomes. *The European Journal of Health Economics,* pp. 7-18.

Ortiz-Ospina, E., 2017. *"Life Expectancy" – What does this actually mean?.* [Online]
Available at: https://ourworldindata.org/life-expectancy-how-is-it-calculated-and-how-should-it-be-interpreted
[Accessed 8 April 2022].

Prevention, C. f. D. C. a., 2021. *Healthy Weight, Nutrition, and Physical Activity.* [Online]
Available at:
https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#InterpretedAdults
[Accessed 7 April 2022].

Roser, M., Ortiz-Ospina, E. & Ritchie, H., 2013. *Life Expectancy.* [Online]
Available at: https://ourworldindata.org/life-expectancy
[Accessed 8 April 2022].

Sugiura, Y., Ju, Y.-S., Yasuoka, J. & Jimba, M., 2007. *Rapid increase in Japanese life expectancy after World War II.* [Online]
Available at: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/114/2012/10/RP245.pdf
[Accessed 10 April 2022].

7. Appendix / R-code used

```
> sqrt(mean((ridge.pred - y.test)^2))
[1] 3.855435
> ###RMSE
> sqrt(mean((lasso.pred - y.test)^2))
[1] 3.834575
>

> #calculate RMSE
> sqrt(mean((pcr.pred - y.test)^2))
[1] 3.739868
>
```

```
#Install the required packages
#Read the Packages
library(readxl)
library(psych)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(caret) #to split the data
library(Hmisc) #For rcorr() function
library(corrplot)
library(corrr)
library(data.table)
library(boot)
```

```
library(ISLR2)
library(leaps)
library(glmnet)
library(pls)

#Set Workind Directory
setwd('D:/Business Analytics/Advanced Analytics and Machine Learning/Assignment 1')

#Read the excel sheet into variable life
life <- read.csv('Life Expectancy Data.csv')

options(scipen = 100)

#Summarizing the Data
summary(life)

# get means for variables in data frame
# excluding missing values
sapply((life), mean, na.rm=TRUE)

describe(life)

#::::: Data Quality Issues and Action ::::::

#Caluclate Total NA values
sum(is.na(summary(life)))

#Remove infant.death, BMI, Under five deaths
life <- select(life, -6,-11:-12)

#Resummarize Population below 15000 to median value
life$Population[life$Population < 15000] <- median(life$Population, na.rm = TRUE)

#Re-summarize the Alcohol to Median for NAs
life$Alcohol[is.na(life$Alcohol)] <- median(life$Alcohol, na.rm = TRUE)

#Re-summarize the Polio to Median for NAs
life$Polio[is.na(life$Polio)] <- mean(life$Polio, na.rm = TRUE)

#Re-summarize the Diphtheria  to Mean for NAs
life1$Diphtheria[is.na(life1$Diphtheria)] <- mean(life1$Diphtheria, na.rm = TRUE)

#Re-summarize the thinness..1.19.years  to Median for NAs
life$thinness..1.19.years[is.na(life$thinness..1.19.years)] <- median(life$thinness..1.19.years, na.rm =
TRUE)

#Re-summarize the thinness.5.9.years  to Median for NAs
life$thinness.5.9.years[is.na(life$thinness.5.9.years)] <- median(life$thinness.5.9.years, na.rm =
TRUE)

#Re-summarize the Schooling to Mean for NAs
```

```
life$Schooling[is.na(life$Schooling)] <- median(life$Schooling, na.rm = TRUE)

#Re-summarize the Total Expenditure to Median for NAs
life$Total.expenditure[is.na(life$Total.expenditure)] <- median(life$Total.expenditure, na.rm =
TRUE)


#Convert Country into as.factor
life$Country <- as.factor(life$Country)

#Convert Status into as.factor
life$Status <- as.factor(life$Status)

lifex <- life %>% drop_na()
summary(lifex)

#:::::: VISUALIZATION USing GGPLOT::::::::::

#Life Expectancy Vs Schooling - Geom Points with respect
lifex %>% ggplot(aes(x=lifex$Schooling, y=(lifex$Life.expectancy), colour = lifex$Status))+
 geom_point()+
 geom_smooth(method = lm, formula = y ~ splines::bs(x, 1), colour = 'black')+
 labs(title = 'Comparison of Life Expectancy and Schooling',
     x="Number of years of Schooling(years)", y= "Life Expectancy (AGE)", colour = "Country Status")+
 scale_y_continuous(labels = function(x) format(x, scientific = FALSE))


#Life Expectancy for Developed and Developing countries
ggplot(data = lifex, aes(x=Status,y=Life.expectancy, color=Status)) +
 geom_boxplot()+
 scale_color_brewer(palette="Dark2") +
 geom_jitter(shape=16, position=position_jitter(0.2))+
 labs(title = 'Average Expectancy rate for Developed and Developing countries',
     y='Expectancy rate (Age)',x='Country Status')


#Average Expenditure for Developed and Developing countries
ggplot(lifex)+
 geom_histogram(mapping = aes(x = lifex$Status, y=(lifex$percentage.expenditure)),
            stat = "Summary", fun.y = "mean")+
 labs(title = "Average Expenditure for Developed and Developing countries", x="Country Status", y=
      "Expenditure on health as a % of GDP per capita")+
 scale_y_continuous(labels = function(x) format(x, scientific = FALSE))

#HIV AIDS vs Life Expectancy with respect to Status
ggplot (data = lifex, aes (x=Life.expectancy,y=HIV.AIDS, colour = Status)) + geom_point() +
geom_smooth(method="loess") +
 ggtitle('Life Expectancy vs HIV AIDS') +
 xlab('Life Expectancy') +
 ylab('Deaths per 1000 live births HIV/AIDS (0-4 years)')
```

```
#Life Expectancy vs Alcohol with respect to Status
ggplot (data = lifex, aes (x=Life.expectancy,y=Alcohol, colour = Status)) + geom_point() +
geom_smooth(method="loess") +
  ggtitle('Life Expectancy vs Alcohol') +
  xlab('Life Expectancy') +
  ylab('consumption (in litres of pure alcohol)')


#:::: RELATIONSHIPS BETWEEN DIFFERENT VARIABLES ::::::

#CORRPLOT
subdata <-
lifex[c("Life.expectancy","Income.composition.of.resources","Schooling","Diphtheria","Polio","Alcoh
ol","percentage.expenditure","Year")]
cor <- cor(subdata)
cor_sort <- as.matrix(sort(cor[,'Life.expectancy'], decreasing = TRUE))
corrplot.mixed(cor, tl.col="black", tl.pos="lt")

#Relationship between Life.expectancy and Income.composition.of.resources with p-value and
confidence interval
cor.test(x=lifex$Life.expectancy, y=lifex$Income.composition.of.resources)

#Relationship between Life.expectancy and Total.expenditure with p-value and confidence interval
cor.test(x=lifex$Life.expectancy, y=lifex$Total.expenditure)

#Relationship between Life.expectancy and Polio with p-value and confidence interval
cor.test(x=lifex$Life.expectancy, y=lifex$Polio)

#::::: SPLIT THE LIFEX DATA INTO TRAINING AND TEST:::

#to create a partition with 80%
set.seed(123) #generate a sequence of random numbers
index <- createDataPartition(lifex$Life.expectancy, p = 0.8, list = FALSE,)
train <- lifex[index, ] #first 80% for training
test <- lifex[-index, ] #bottom 20% for testing



#::::::::::::MULTIPLE LINEAR REGRESSION::::::::::::::

#Create a Multiple Linear regression model
MLR.model <- lm(Life.expectancy ~ (Schooling + Diphtheria + Polio + Alcohol +
percentage.expenditure + Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources +
Total.expenditure) , data = train)

#review the model
summary(MLR.model)

#prediction using the model
prediction_1 <- predict(MLR.model, newdata = test)
```

```
#(i.e. difference between the actual sale value and the predicted sale value)
postResample(pred = prediction_1, obs = test$Life.expectancy)

#No MultiColinearity
vif(MLR.model)
mean(vif(MLR.model))


#::::::VALIDATION SET APPROACH::::::::::

train <- sample(index, )

## $k$-Fold Cross-Validation
###
set.seed(17)
cv.error.10 <- rep(0, 10)
for (i in 1:10) {
  glm.fit <- glm(Life.expectancy ~ poly(Schooling + Diphtheria + Polio + Alcohol +
percentage.expenditure + Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources +
Total.expenditure, i), data = lifex)
  cv.error.10[i] <- cv.glm(lifex, glm.fit, K = 10)$delta[1]
}
cv.error.10

#:::::::::::::::: Subset Selection Method::::::::::::::::

### Best Subset Selection

###
#Check if no row has NA
sum(is.na(lifex$Life.expectancy))
###

regfit.full <- regsubsets(Life.expectancy ~ (Schooling + Diphtheria + Polio + Alcohol +
percentage.expenditure + Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources +
Total.expenditure ), lifex)
summary(regfit.full)
###
regfit.full <- regsubsets(Life.expectancy ~ (Schooling + Diphtheria + Polio + Alcohol +
percentage.expenditure + Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources +
Total.expenditure),data = lifex,
                    nvmax = 5)
reg.summary <- summary(regfit.full)
###
names(reg.summary)
###
reg.summary$rsq
###
par(mfrow = c(2, 2))
plot(reg.summary$rss, xlab = "Number of Variables",
    ylab = "RSS", type = "l")
```

```
plot(reg.summary$adjr2, xlab = "Number of Variables",
    ylab = "Adjusted RSq", type = "l")
###
which.max(reg.summary$adjr2)
points(11, reg.summary$adjr2[11], col = "red", cex = 2,
      pch = 20)
###
plot(reg.summary$cp, xlab = "Number of Variables",
    ylab = "Cp", type = "l")
which.min(reg.summary$cp)
points(10, reg.summary$cp[10], col = "red", cex = 2,
      pch = 20)
which.min(reg.summary$bic)
plot(reg.summary$bic, xlab = "Number of Variables",
    ylab = "BIC", type = "l")
points(6, reg.summary$bic[6], col = "red", cex = 2,
      pch = 20)
###
plot(regfit.full, scale = "r2")
plot(regfit.full, scale = "adjr2")
plot(regfit.full, scale = "Cp")
plot(regfit.full, scale = "bic")
###
coef(regfit.full,5)



###:::::::: Ridge Regression::::::::::::::::::::
### DEFINING x and y
x <- model.matrix(Life.expectancy ~ Schooling + Diphtheria + Polio + Alcohol +
percentage.expenditure + Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources +
Total.expenditure, lifex)
y <- lifex$Life.expectancy

###

grid <- 10^seq(10, -2, length = 100)
ridge.mod <- glmnet(x, y, alpha = 0, lambda = grid)
###
dim(coef(ridge.mod))
###
ridge.mod$lambda[50]
coef(ridge.mod)[, 50]
sqrt(sum(coef(ridge.mod)[-1, 50]^2))
###
ridge.mod$lambda[60]
coef(ridge.mod)[, 60]
sqrt(sum(coef(ridge.mod)[-1, 60]^2))
###
predict(ridge.mod, s = 50, type = "coefficients")[1:20, ]
###
```

```
set.seed(1)
train <- sample(1:nrow(x), nrow(x) / 2)
test <- (-train)
y.test <- y[test]
###
ridge.mod <- glmnet(x[train, ], y[train], alpha = 0,
                    lambda = grid, thresh = 1e-12)
ridge.pred <- predict(ridge.mod, s = 4, newx = x[test, ])
mean((ridge.pred - y.test)^2)
###
mean((mean(y[train]) - y.test)^2)
###
ridge.pred <- predict(ridge.mod, s = 1e10, newx = x[test, ])
mean((ridge.pred - y.test)^2)
###
ridge.pred <- predict(ridge.mod, s = 0, newx = x[test, ],
                      exact = T, x = x[train, ], y = y[train])
mean((ridge.pred - y.test)^2)
lm(y ~ x, subset = train)
predict(ridge.mod, s = 0, exact = T, type = "coefficients",
      x = x[train, ], y = y[train])[1:20, ]
###
set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 0)
plot(cv.out)
bestlam <- cv.out$lambda.min
bestlam
###
ridge.pred <- predict(ridge.mod, s = bestlam,
                      newx = x[test, ])

##produce Ridge trace plot
plot(glmnet(x = data.matrix(lifex[,
c("Year","Adult.Mortality","Alcohol","percentage.expenditure","Polio","Diphtheria","Schooling","HI
V.AIDS", "Income.composition.of.resources", "Total.expenditure")]) , y = lifex$Life.expectancy , alpha
= 0), xvar = "lambda")

mean((ridge.pred - y.test)^2)
###
out <- glmnet(x, y, alpha = 0)
predict(out, type = "coefficients", s = bestlam)[1:20, ]

#use fitted best model to make predictions
y_predicted <- predict(ridge.mod, s = bestlam, newx = x)

#find SST and SSE
sst <- sum((y - mean(y))^2)
sse <- sum((y_predicted - y)^2)

#find R-Squared
ridge.rsq <- 1 - sse/sst
```

```
ridge.rsq

##::::::::::::LASSO REGRESSION::::::::::::

###
lasso.mod <- glmnet(x[train, ], y[train], alpha = 1,
                    lambda = grid)
plot(lasso.mod)
###
set.seed(1)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 1)
plot(cv.out)
bestlam <- cv.out$lambda.min
lasso.pred <- predict(lasso.mod, s = bestlam,
                      newx = x[test, ])
mean((lasso.pred - y.test)^2)
###
out <- glmnet(x, y, alpha = 1, lambda = grid)
lasso.coef <- predict(out, type = "coefficients",
                      s = bestlam)
lasso.coef
lasso.coef[lasso.coef != 0]

#use fitted best model to make predictions
y_predicted <- predict(lasso.mod, s = bestlam, newx = x)

#find SST and SSE
sst <- sum((y - mean(y))^2)
sse <- sum((y_predicted - y)^2)

#find R-Squared
lasso.rsq <- 1 - sse/sst
lasso.rsq

###::::: Principal Components Regression:::::::

###
set.seed(2)
pcr.fit <- pcr(Life.expectancy ~ Schooling + Diphtheria + Polio + Alcohol + percentage.expenditure +
Year + Adult.Mortality + HIV.AIDS + Income.composition.of.resources + Total.expenditure, data =
lifex, scale = TRUE,
               validation = "CV")
###
summary(pcr.fit)
###
validationplot(pcr.fit, val.type = "MSEP")
###
set.seed(1)
pcr.fit <- pcr(Life.expectancy ~ Schooling + Diphtheria + Polio + Alcohol + percentage.expenditure +
Year + Adult.Mortality + + HIV.AIDS + Income.composition.of.resources + Total.expenditure, data =
lifex, subset = train,
```

```
            scale = TRUE, validation = "CV")
summary(pcr.fit)

#visualize cross-validation plots
validationplot(pcr.fit, val.type = "MSEP")
validationplot(pcr.fit, val.type="R2")
###

train <- lifex[index,]
test <- lifex[-index, ] #bottom 20% for testing
y.test <- lifex[-index,c("Life.expectancy")]
pcr.pred <- predict(pcr.fit, test, ncomp = 5)
mean((pcr.pred - y.test)^2)
###

#calculate RMSE
sqrt(mean((pcr.pred - y.test)^2))
```