Final Project Report:
Analyzing Life Expectancy Data

Group 3
Mehrdad Ghyabi, Sagar D. Goswami, Sai Jaswanth Kumar

George Mason University

STAT-515: Statistics: Visualization for Statistics
Dr. Isuru Dassanayake

October 18, 2021

Abstract:

The World Health Organization (WHO) provides access to essential Health Indicators Data through the Open Data APIs, which provides guidance and framework for analysis, and planning to Governments and Private entities. This project is about analyzing a dataset consisting of historical life expectancy data from different countries and 18 related variables. The variables cover a variety of socio-economic parameters from GDP to schooling rates and from various disease statistics to mortality rates. The study was done in three parts, addressing a specific research question.

The first part talks about the distribution of data in various variables, second part found the most critical variables to regress life expectancy. In the last part, a principal component analysis was performed to see if the information from this dataset could be meaningfully conveyed with a smaller number of variables. The results showed that the model could reasonably estimate life expectancy using the first four principal components.

About the Data:

The WHO Life Expectancy Data has the following features:

It has 21 variables and 2938 records with data for 183 Countries for years ranging from 2000-2015. For the sake of simplicity and recency, only data from 2010-2015 was considered for most of the analysis.

The following table shows the description of each variable for the WHO Life Expectancy Data:

S.No	Attribute	Information
1	Country	Country
2	Year	Year
3	Status	Developed or Developing status
4	Life Expectancy	Age(years)

5	Adult Mortality	Adult Mortality Rates of both sexes(probability of dying between 15&60 years per 1000 population)
6	Infant Deaths	Number of Infant Deaths per 1000 population
7	Alcohol	Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
8	Percent Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
9	Hep B	Hepatitis B (HepB) immunization coverage among 1-year olds(%)
10	Measles	Number of reported measles cases per 1000 population
11	BMI	Average Body Mass Index of the entire population
12	U-5 Deaths	Number of under-five deaths per 1000 population
13	Polio	Polio(Pol3) immunization coverage among 1-year olds(%)
14	Total Expenditure	General government expenditure on health as a percentage of total government expenditure(%)
15	Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year olds (%)
16	HIV/AIDS	Deaths per 1000 live births HIV/AIDS(0-4 years)
17	GDP	Gross Domestic Product per capita(in USD)
18	Population	Population Thinness 10-19- Prevalence of thinness among children and adolescents for Age 10 to 19(%)
19	Thinness 5	Prevalence of thinness among children for Age 5 to 9(%)
20	Income Composition	Human Development Index in terms of income composition of resources(0
21	Schooling	Number of years of Schooling

The WHO data in its raw form has a lot of missing values, outliers, and garbage data. Moreover, the population data was added separately using the merge function. Various Exploratory Analysis was done on the raw data, then cleaned and manipulated to appropriate form to prepare it for analysis. The research methodology employed for the project is as follows:

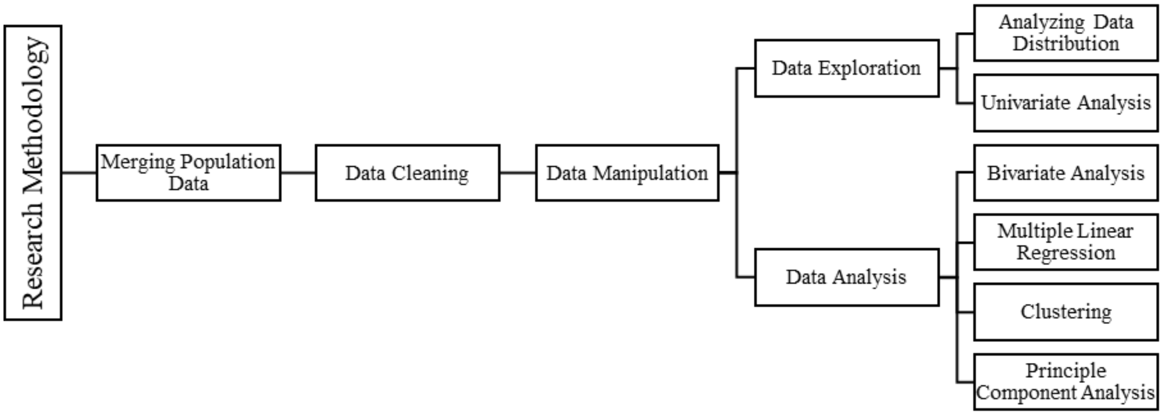


Figure 1 Research Project Methodology

Data Exploration

The Raw Data has country-wise and Year-wise data. It was visualized on maps, using the map_data left_join and ggplot functions. Following observations were made using the Map plots:

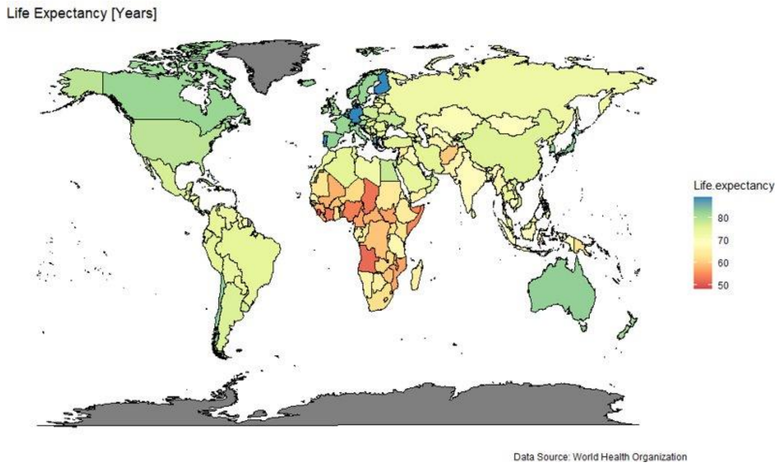
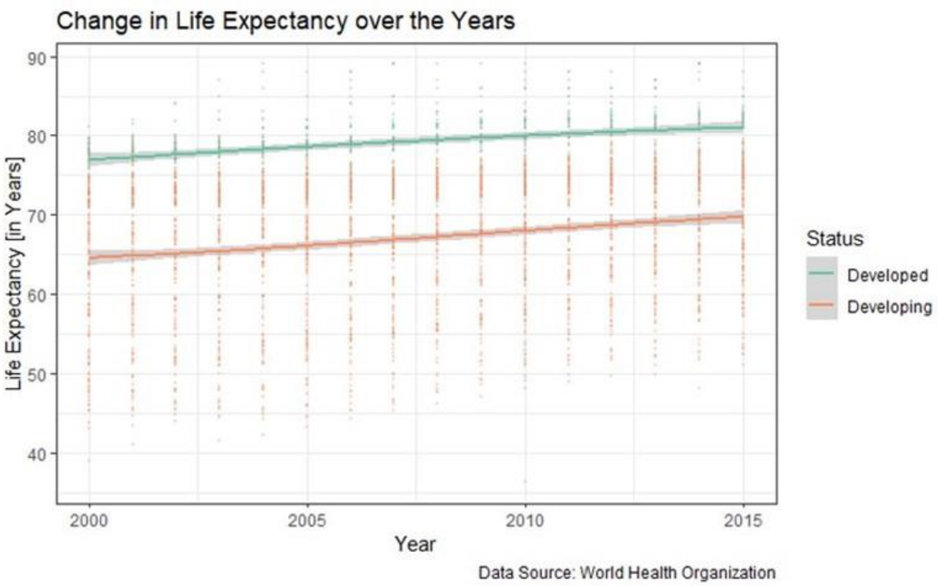


Figure 2 Life Expectancy [in Years] for all the countries in 2014.

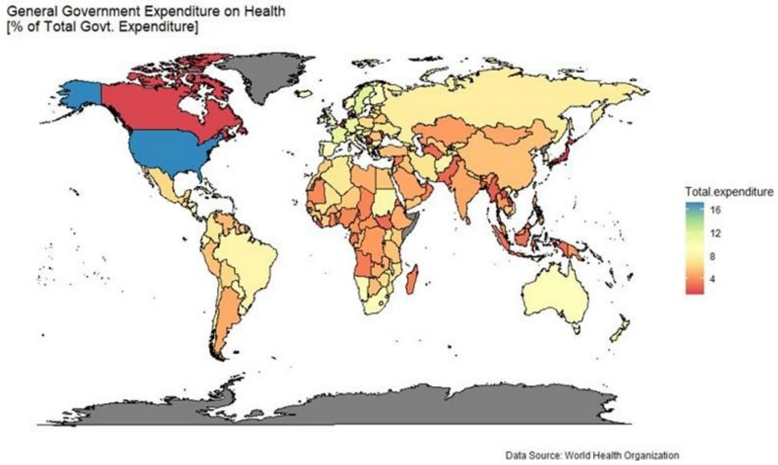
Life Expectancy appears to be lower for African Continent. In contrast, European Countries like Finland, Germany, Portugal, Albania, and Greece were observed to have the highest Life Expectancy of more than 75 Years.



Data Source: World Health Organization

Figure 3 Change in Life Expectancy over the Years

The Life Expectancy was observed to be increasing with years for both (Developed and Developing Countries.) The average difference between them was around 10-12 Years.

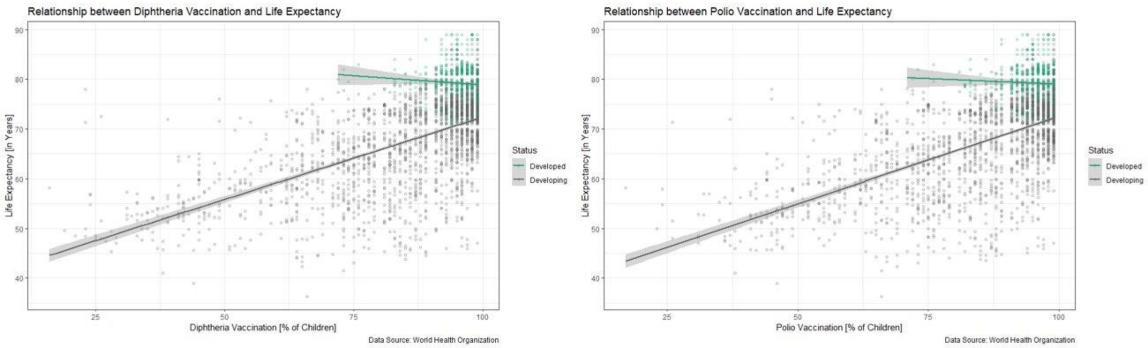

General Government Expenditure on Health
[% of Total Govt. Expenditure]
Total.expenditure
16
12
8
4

Data Source: World Health Organization

Figure 4 Government Expenditure as (%) of total expenditure

Canada appears to have the lowest proportion of government spending, yet it has one of the highest Life Expectancy. For most of the countries, Government Expenditure lies between 4-12% of total expenditure.

Bi-Variate Analysis of key Variables

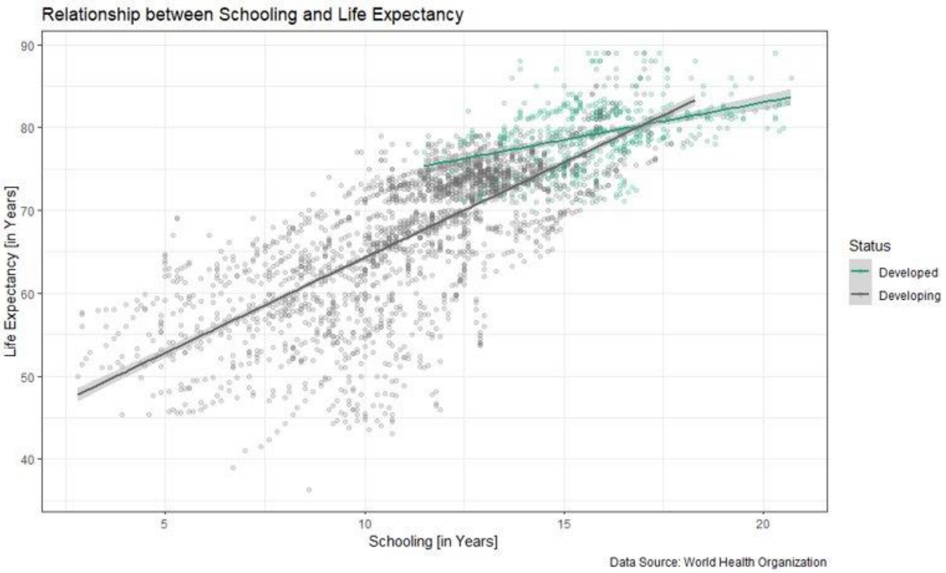


Data Source: World Health Organization

Data Source: World Health Organization

Figure 5 Correlation between Vaccination and Life Expectancy

It can be validated from the above plots that increasing the Vaccination % will increase the Life Expectancy of a country by a considerable margin, especially for Developing Countries.

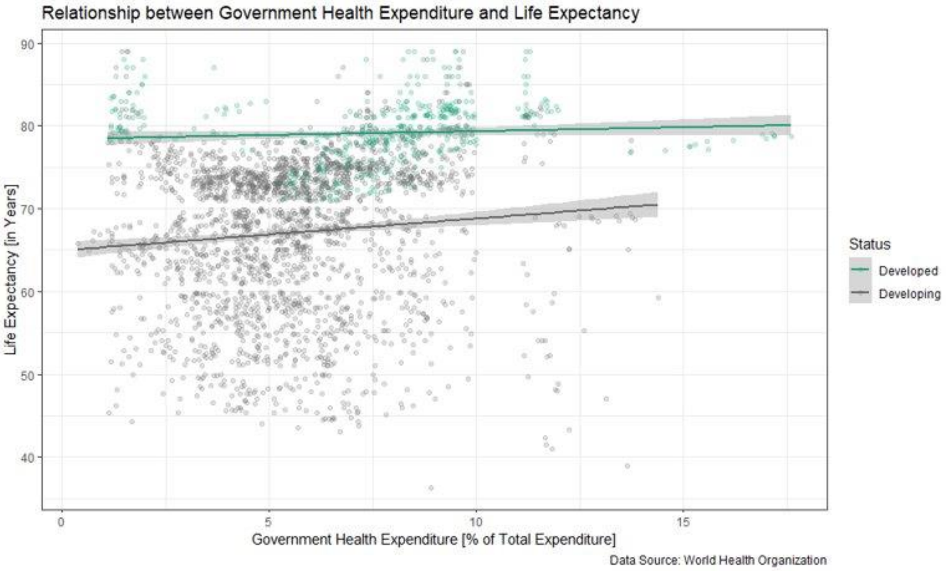


Data Source: World Health Organization

Figure 6 Correlation between Education and Life Expectancy

A similar pattern can be noticed for Education as well. The higher the years spent in Schooling, the more Life Expectancy can be observed.

Interestingly enough, the same cannot be said for the Government Spending on Healthcare. However, it could probably be due to some underlying collinearity with other variables like GDP, cost of healthcare services, and the facilities available for a given country.



Data Source: World Health Organization

Analysis of the factors determining the life expectancy:

Figure 7 shows that developing countries are 33% behind in average life expectancy rates. While the average life expectancy of around 80 years the developing countries are struck at 69%. While this bar graph gives an overall estimate of the data looking these values in time series helps us to determine whether If there is any improvement in life expectancy over the years.

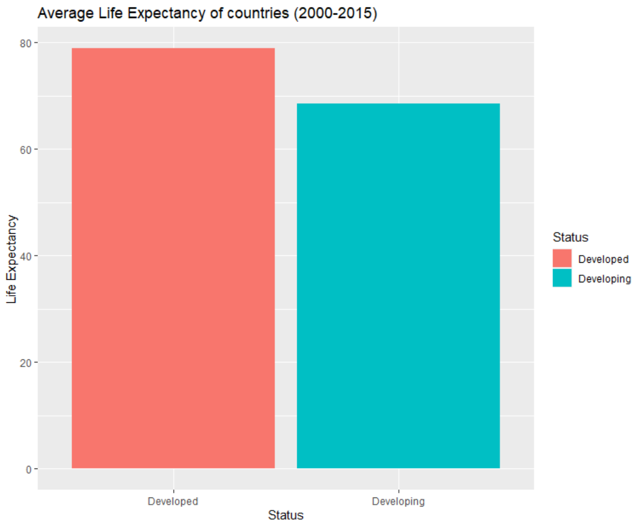


Figure 7- Life Expectancy

Line graph in Figure 8, it is evident that developing countries are doing their best to improve the life expectancy. The results show that the average life expectancy in developing countries has increased by almost 3% starting with 67.7 years in 2010 and 70.3 years in 2015 respectively. On the other hand, in case of developed countries we can see that there are highs and lows in average expectancy, and it has settled to 79 to 80 years.

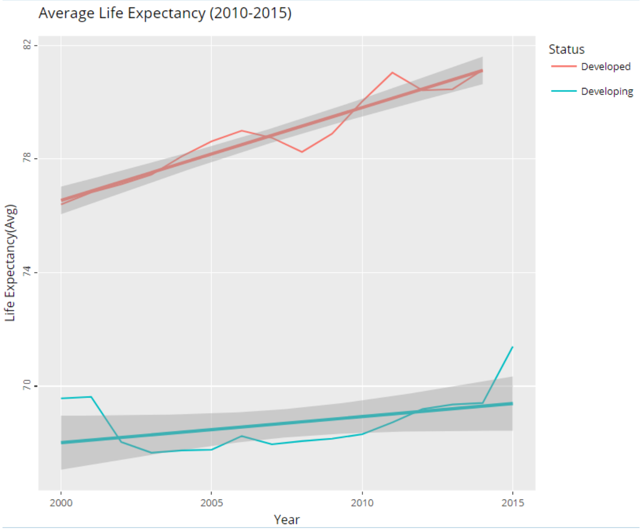


Figure 8-Life expectancy trends

Figure 9 shows that Life expectancy is positively correlated with factors such as Income composition of resources and Schooling whereas it is negatively correlated with Adult mortality and HIV(Aids).

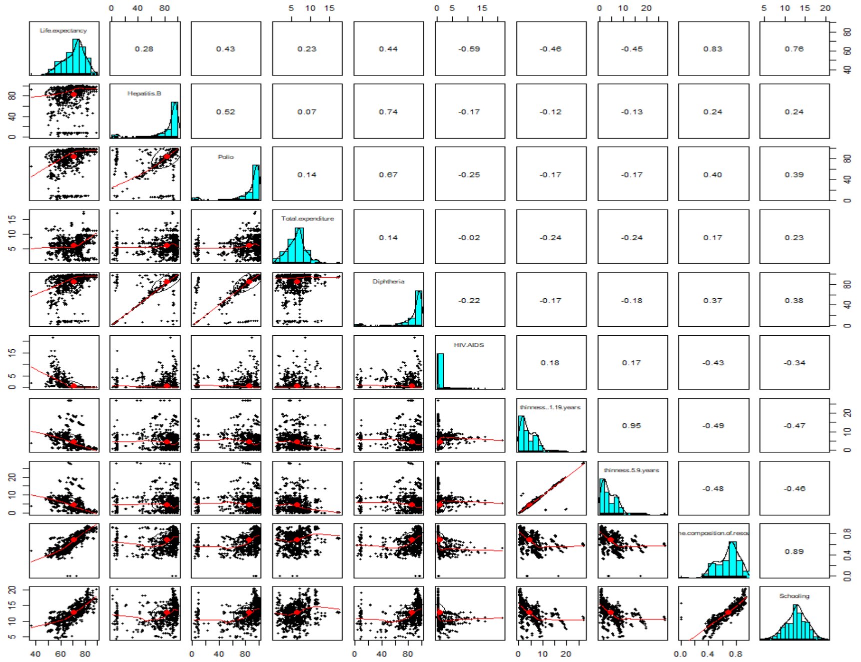
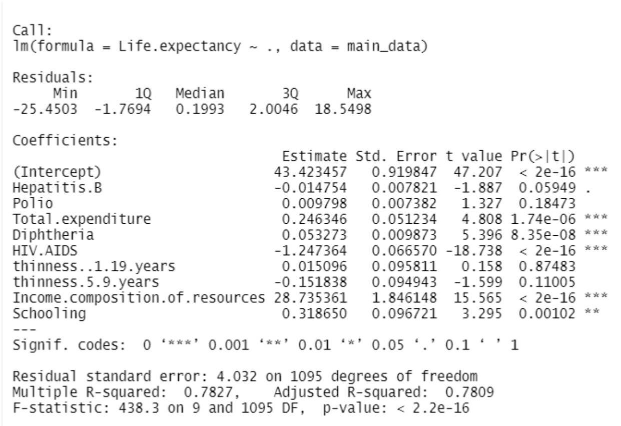


Figure 9- Distributin Data

Research Question 1: What Factors play important role in determining the Life expectancy?

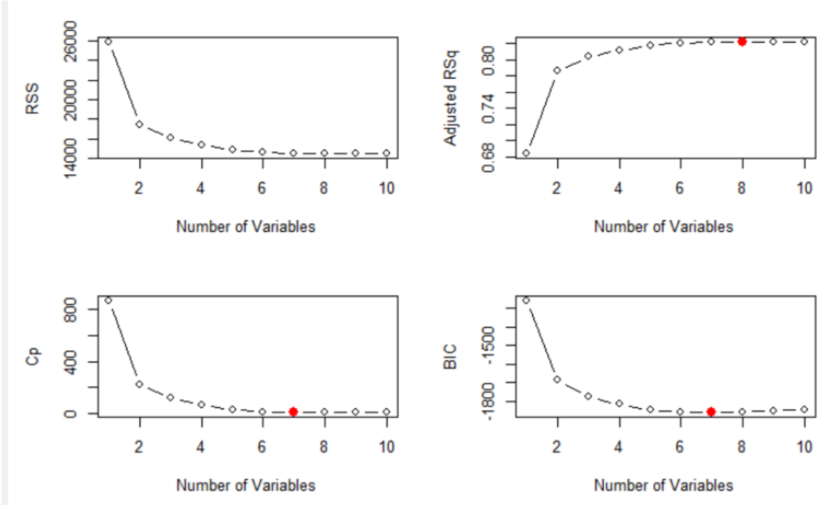
After fitting the linear regression model with Life expectancy as determining variable and all other as predicting variables, the results showed that health related variables such as Hepatitis, Polio, and thinness 10-19 years are not explaining the life expectancy very well. It also explains that variables such as Adult Mortality, Income composition of resources, HIV and Diphtheria and Schooling are major factors in determining the life expectancy as the p-value is less than 2e-16 rejecting the null hypothesis.



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

Residuals:
     Min       1Q   Median       3Q      Max 
-25.4503  -1.7694   0.1993   2.0046  18.5498 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     43.423457   0.919847  47.207  < 2e-16 ***
Hepatitis.B                     -0.014754   0.007821  -1.887  0.05949 .  
Polio                            0.009798   0.007382   1.327  0.18473    
Total.expenditure                0.246346   0.051234   4.808 1.74e-06 ***
Diphtheria                       0.053273   0.009873   5.396 8.35e-08 ***
HIV.AIDS                        -1.247364   0.066570 -18.738  < 2e-16 ***
thinness..1.19.years             0.015096   0.095811   0.158  0.87483    
thinness.5.9.years              -0.151838   0.094943  -1.599  0.11005    
Income.composition.of.resources 28.735361   1.846148  15.565  < 2e-16 ***
Schooling                        0.318650   0.096721   3.295  0.00102 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.032 on 1095 degrees of freedom
Multiple R-squared:  0.7827,	Adjusted R-squared:  0.7809 
F-statistic: 438.3 on 9 and 1095 DF,  p-value: < 2.2e-16
```

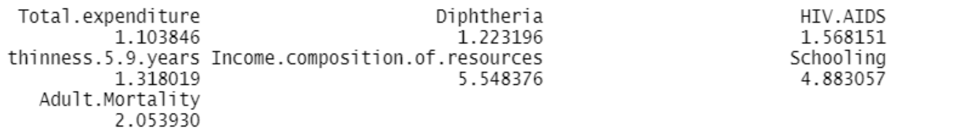
Model Improvement using Backward subset Selection

To improve the model best backward subset selection has been applied on the previous multiple linear regression model.



After comparing the r2,adjusted r2,cp and BIC values the best subset model tells 7 predictor variables can explain the Life expectancy best.

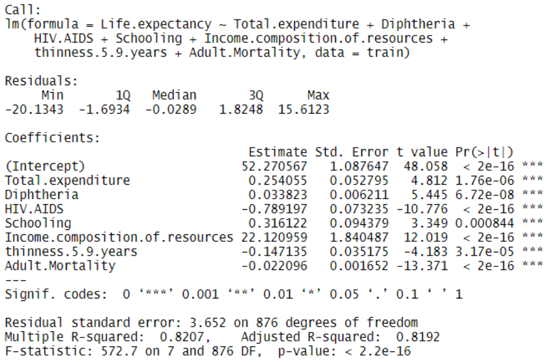
Multicollinearity Test



```
   Total.expenditure                        Diphtheria                          HIV.AIDS
            1.103846                          1.223196                          1.568151
  thinness.5.9.years Income.composition.of.resources                         Schooling
            1.318019                          5.548376                          4.883057
     Adult.Mortality
            2.053930
```

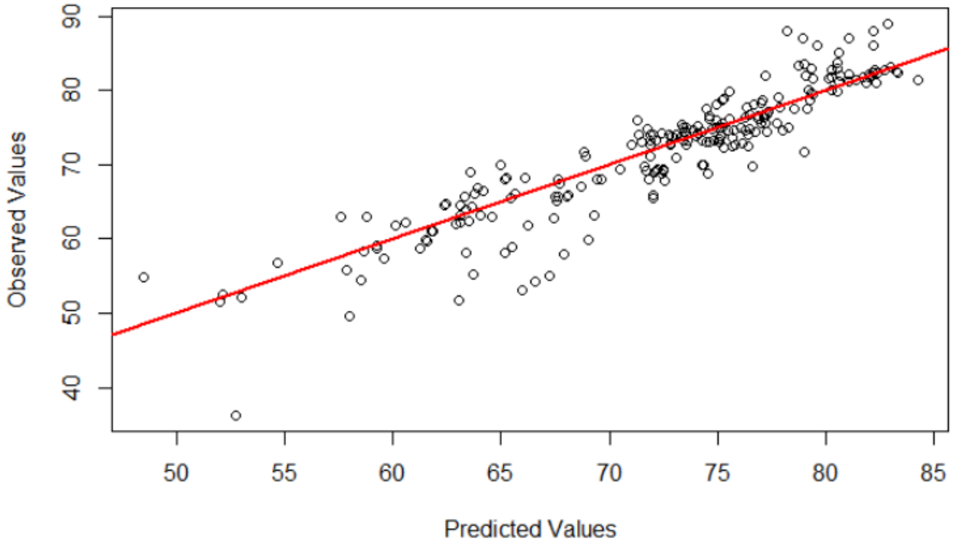
To determine variances between the predictor variables a multicollinearity test has been made using variance inflation factor method. As a rule of thumb if the variables show up variance greater than 10, it indicates that these variables that are showing difficulty in estimating the coefficients. After applying the test on the above model the results showed that there is no presence of multicollinearity which is a positive sign.

Model Fitting and Results:



```
Call:
lm(formula = Life.expectancy ~ Total.expenditure + Diphtheria + 
    HIV.AIDS + Schooling + Income.composition.of.resources + 
    thinness.5.9.years + Adult.Mortality, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.1343  -1.6934  -0.0289   1.8248  15.6123 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     52.270567   1.087647  48.058  < 2e-16 ***
Total.expenditure                0.254055   0.052795   4.812 1.76e-06 ***
Diphtheria                       0.033823   0.006211   5.445 6.72e-08 ***
HIV.AIDS                        -0.789197   0.073235 -10.776  < 2e-16 ***
Schooling                        0.316122   0.094379   3.349 0.000844 ***
Income.composition.of.resources 22.120959   1.840487  12.019  < 2e-16 ***
thinness.5.9.years              -0.147135   0.035175  -4.183 3.17e-05 ***
Adult.Mortality                 -0.022096   0.001652 -13.371  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.652 on 876 degrees of freedom
Multiple R-squared:  0.8207,	Adjusted R-squared:  0.8192 
F-statistic: 572.7 on 7 and 876 DF,  p-value: < 2.2e-16
```

The above linear model has been trained with 80% of the data and tested on the remaining 20%. Then the error and accuracy are calculated using the RMSE which gave values 3.57 and R2 of 0.83 when compared to 3.58 and 0.83 of model1.



Although there hasn't been much improvement between these models, it can be concluded that model 2 did a better job in predicting the life expectancy with less no of predictors. The results show that the selected variables are actually significant and using them for regression results in an accurate model.

Research Question 3: Is it possible to use PCA to make prediction about life expectancy?

A research question that can be asked about this dataset is that if it is possible to gain a prospective of life expectancy in different countries if the life expectancy column was not included in the dataset. Too address this in a simple way, countries were categorized in two groups. In this dataset the mean life expectancy is 71.21 years and the median is 73.30 years, and countries were categorized into countries with life expectancy over 70 years and the ones with life expectancy of below 70 years. The result is visualized in Figure 10 on a map.

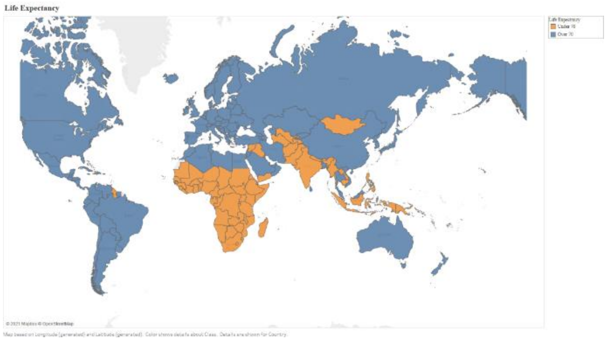

Life Expectancy
Life Expectancy
🟧 Under 70
🟦 Over 70
© 2021 Mapbox © OpenStreetMap
Map based on Longitude (generated) and Latitude (generated). Color shows details about Class. Details are shown for Country.

Figure 10- Countries categorized based on life expectancy

A general conception is that life expectancy is correlated with the level of development of each country. To test this theory countries were categorized into two groups, "developed" and "developing", based on the only categorical variable present in the dataset. The result is visualized in Figure 11 on a map. A visual comparison of Figure 10 and Figure 11, shows that the level of life expectancy in different countries can not be simply explained by looking at their development status, and a deeper analysis is required to answer this research question.

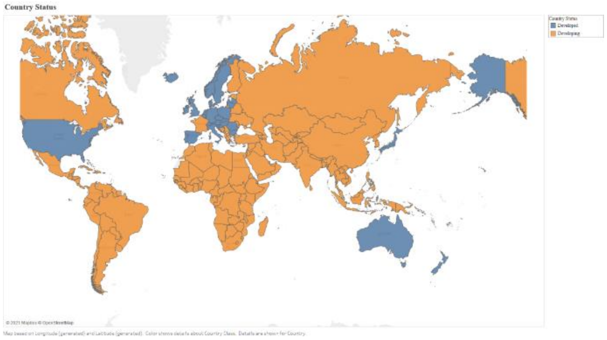

Country Status
Country Status
Developed
Developing
© 2021 Mapbox © OpenStreetMap
Map based on Longitude (generated) and Latitude (generated). Color shows details about Country Status. Details are shown for Country.

Figure 11- Countries categorized based on their status

Principal Component Analysis (PCA) is a useful technique for exploratory data analysis, allowing one to better visualize the variations present in a dataset with many variables. It is particularly helpful in the case of "wide" datasets, where there are many variables for each sample. PCA was selected as the method to answer this research question, and it was performed in R platform.

In this case, where many variables are present, the data cannot be simply visualized in its raw format, making it difficult to get a sense of the trends present within. PCA allow us to see the overall trend of the data, identifying which countries are similar to one another and which are very different. This can enable us to identify groups of samples that are similar in terms of life expectancy.

PCA is a type of linear transformation on a given data set that has values for a certain number of variables (coordinates) for a certain amount of spaces. This linear transformation fits this dataset to a new coordinate system in such a way that the most significant variance is found on the first coordinate, and each subsequent coordinate is orthogonal to the last and has a lesser variance. In this way, the algorithm transforms a set of x correlated variables over y samples to a set of p uncorrelated principal components over the same samples. The mathematical details of PCA is out of the scope of this report.

Variables which correlate with one another, will all contribute strongly to the same principal component. Each principal component sums up a certain percentage of the total variation in the dataset. If initial variables are strongly correlated with one another, it is possible to approximate most of the complexity in the dataset with just a few principal components. As more principal components are added, more and more of the original dataset is summarized. Adding additional components makes the estimate of the total dataset more accurate, but also more unwieldy.

Figure 12 shows proportion of variance explained by the principal components for different years. present the cumulative proportion of variance as the number of principal components grows for the same years. Based on these figures, a total number of 4 principal components were selected to analyze the data. It is noteworthy that by looking at 4 principal component the expected arm shape is visible in

Figure 12. And it captures more than 65% of the variance existing in the dataset.

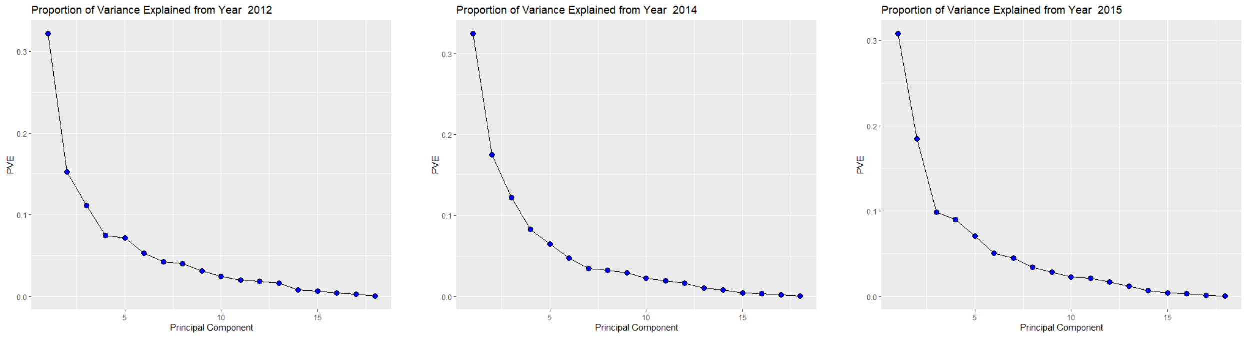


Figure 12- PVE for different years

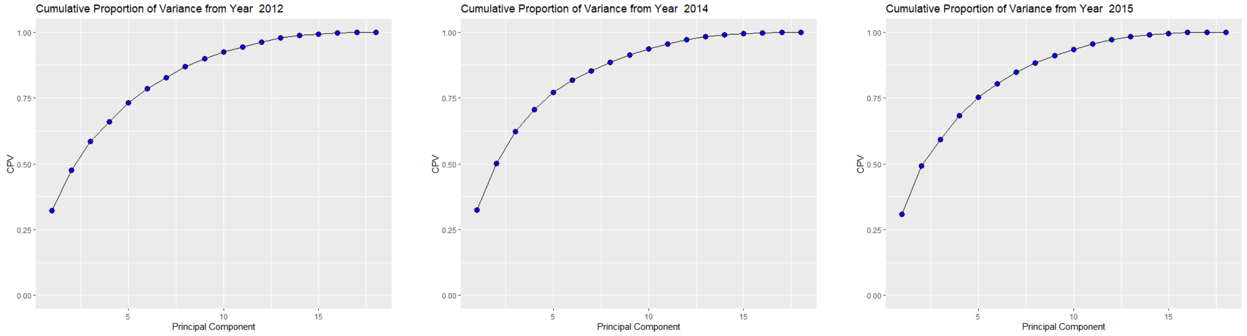


Figure 13- CPV for Different Years

To study the research question, a K Nearest Neighbors (KNN) clustering algorithm with K=2 was used to categorize countries into two groups using the principal components. To visually investigate the results, data points (countries) were plotted in 1D based on the first principal component, and in 2D based on the first two principal components separately. Then, categories were color coded on the plots. Figure 14 shows the countries color coded based on their development status in principal component coordinates in 1D and 2D. Figure 15 shows the countries color coded based on their levels of life expectancy in principal component coordinates in 1D and 2D. In both Figure 14 and Figure 15 two fairly separated groups are visible however there is some overlap between them.

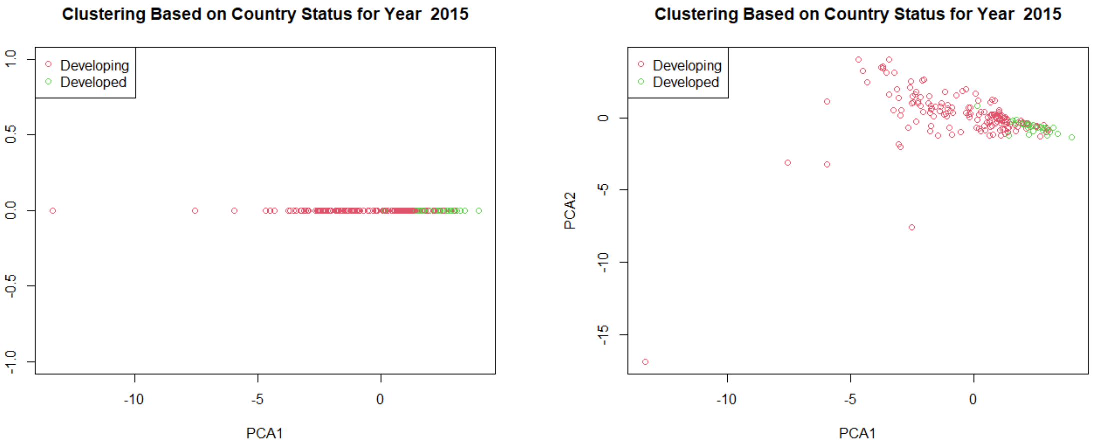


Figure 14- Clusters based on status

As it was mentioned before, the 4 first principal components were used for analysis. The countries were clustered based on their 4 first principal component with KNN (K=2), and the results are presented in the first principal component coordinate, and in the first 2 principal components coordinates in Figure 16. This figure shows a pattern similar to those of Figure 14 and Figure 15, however, there is a better way to visualize this similarity in trends.

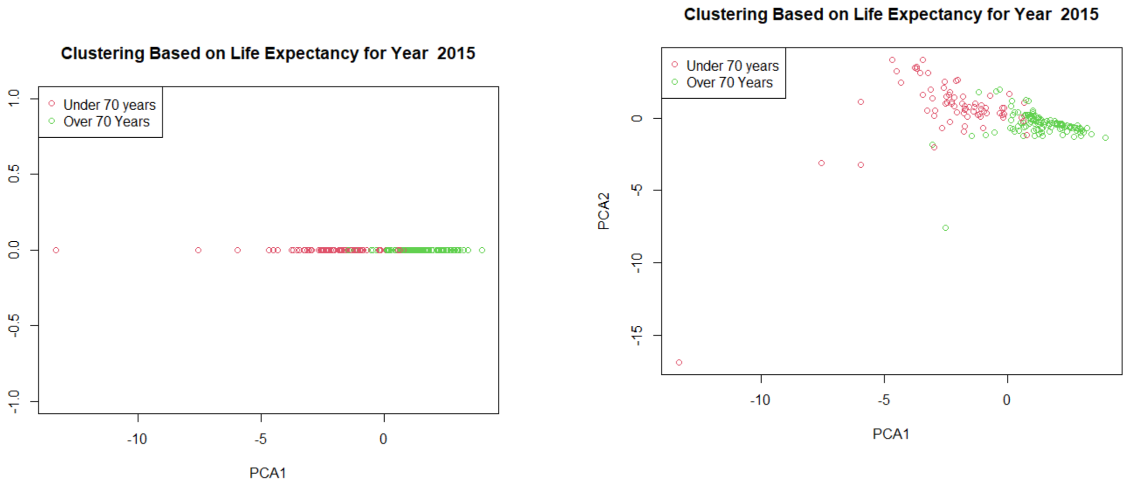


Figure 15- Clusters based on life expectancy

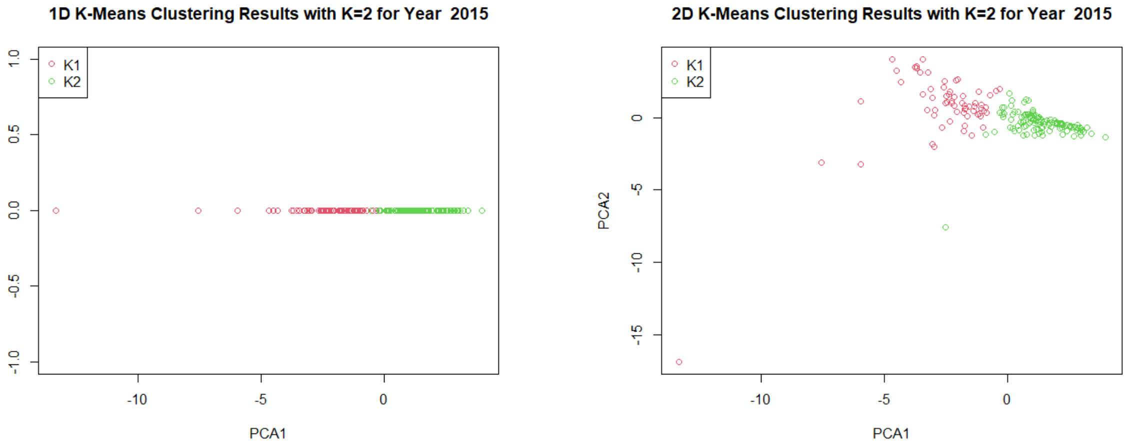


Figure 16- Clusters based on 2NN

Result of the 2NN clustering based on PCA are represented on a map along with the map of life expectancy levels from Figure 10, in **Error! Reference source not found.**. The maps look like each other with only a few exceptions. This shows that even if the life expectancy column was absent from the dataset, it would be possible to make assumption about the levels of life expectancy in different countries with a fairly good accuracy.

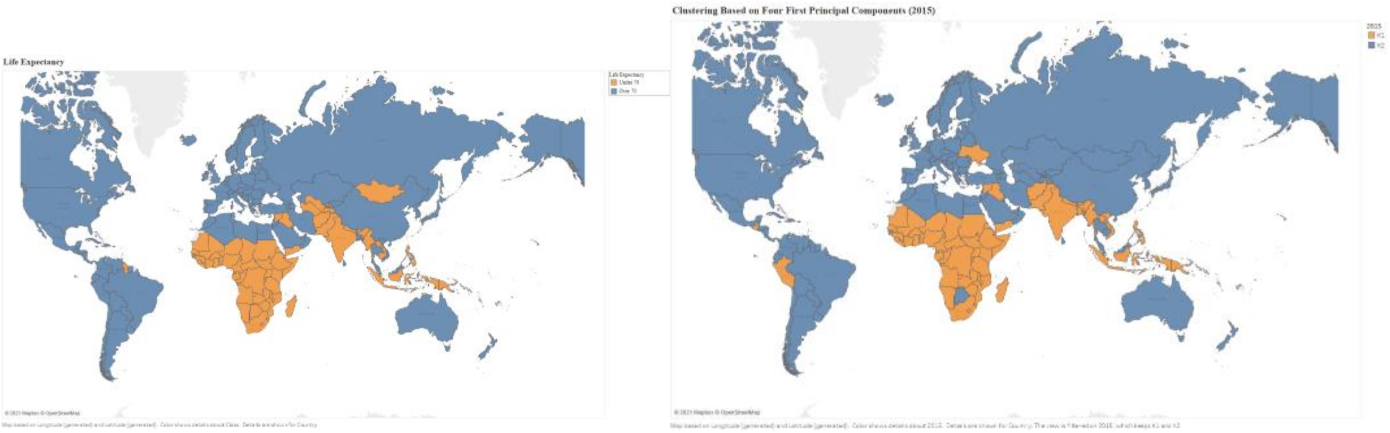

Life Expectancy
Life Expectancy
Under 70
Over 70
Clustering Based on Four First Principal Components (2015)
2015
K1
K2
© 2021 Mapbox © OpenStreetMap

Figure 17- Final visualization along with life expectancy levels