# 

# 

# 

# **Student Name**

# 

# 

# 

# 

# **Life Expectancy**

Contents

* Introduction
* Data Description
  + Reading Data
  + Exploratory data analysis
  + Data cleaning
  + Statistics on given data
  + Data wrangling & feature selection/Extraction
  + Linear Regression Diagnostics
* Checking the statistical significance
* Checking the Assumption
* Conclusion

.

**Introduction:**

We have a data set we wanted to determine the life expectancy from 2000 to 2015. This data is on the economics, the healthcare, the immunization levels and some other information collected from each country and in each year, we are supposed to give the recommendations to the government on how they can improve on the life expectancy.

**Data Description:**

The Data set are based on the different countries, the project is relying on the accuracy of the data World Health Organization (WHO) keeps track of the health status of all countries and many other related factors

The World Health Organization (WHO) has started producing annual life tables for all members in the states in 1999, these reports are basically WHO estimates the global, regional, country-level pattern and trends in the all-cause.

The life expectancy, estimate the average of additional years that a person of a given person, the most common measure of life expectancy is life expectancy of the birth, the data has give to us has noise or not clean we do the below steps to understand the data better and give clear picture about the data to the WHO

We are analyzing the data set in the below steps:

1. Read data

2. Exploratory data analysis

3. Feature Engineering and feature selection

4. Model selection

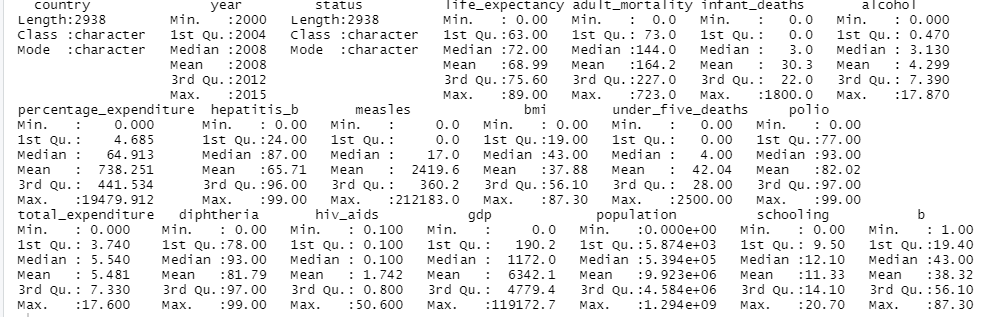
5. Model optimization

**Reading Data:**

The data is saved in the csv format Life\_expectency.csv The data contains 2938 rows and 19 columns, with the country names status (either developed or developing) all the data in the numeric format.

**Exploratory data analysis:**

In the Exploratory data analysis (EDA) part we are cleaning the data as in the data set will need to check the null values and impute the null values with NAN, mean, and median.



Using above summary we can able to find out is database having correct column type or not

like above life\_expectancy has type character that is why we have to convert that into numerical.

As in the data there are many null values which are Alcohol, Hepatitis, Audlt\_mortality, BMI, Polio, Total\_exp, GDP, Population, and schooling have imputed the nan , mean and median.

I have selected Alcohol and schooling values for imputing the Nan by observing the trends in between selected intervals. These values are mean values of the selected interval of another feature.

**Data Cleaning:**

As data cleaning,

* we have removed whitespaces
* we are able to find out if the database has the correct column type or not, like above life\_expectancy has type character that is why we have to convert that into numerical. Converting all character type columns as a numerical type.

Here our plan is to fill out all missing values with mean of respective column, that is why why we have to follow below 4 steps,

* Checking if dataset has any NA value
* Replacing all na values with 0
* Converting values into numeric
* Filling out all missing values with mean of respective columns, Note that we are performing this only for columns that have numerical data, it will help us to get proper stats.

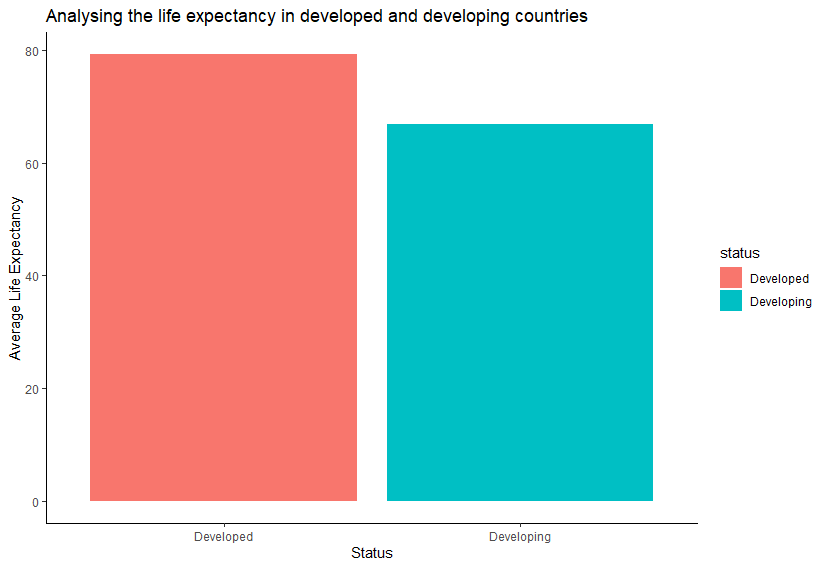
**Statistics on given data:**

Removing the year column, we don't need it for now.

Average life expectancy of all over the worlds: **68.98931**

Check the avg life expectancy by countries

We can see that the life expectancy in developed countries is 20% greater than developing countries.

****

We can see that the life expectancy is higher in developed nations like the UK, Canada and there is a 20% drop in life expectancy in developing nations like Afghanistan, Yemen, Etc.

Why life expectancy is greater in developed nations,

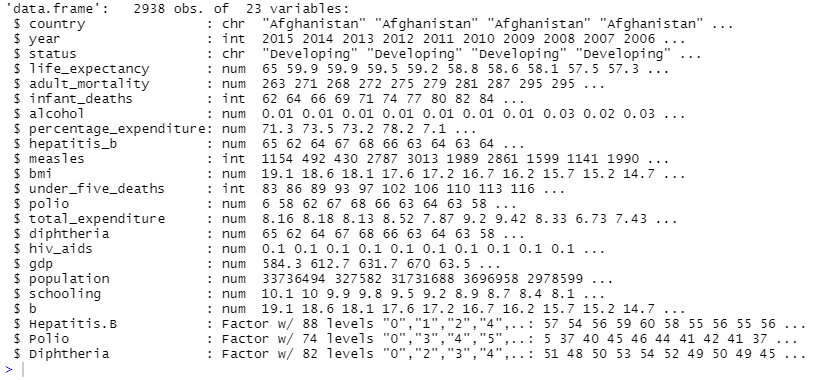
Now on the previous step we are able to know that life expectancy in developed countries is 20% greater than developing countries.

We have to check now WHY?

We will find it in a while….

**Data wrangling & feature selection/Extraction:**

All countries need to reach ≥90% national coverage for all vaccines in the country’s routine immunization schedule by 2020. Based on that statement, we are going to mutate the Hepatitis.B, Polio, and Diphtheria into a categorical variable, with 2 values: “Under 90% Covered” and “Covered by 90% or More”. By doing this, hopefully we can get a better view on the immunization impact to Life.expectancy.



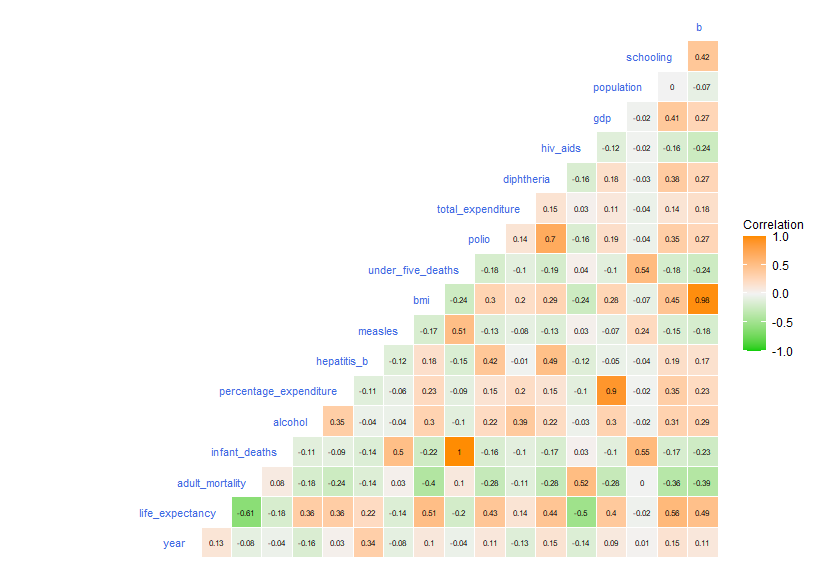
Mutate hepatitis\_b: Range between min value & the Quartile is very wide and has to be manipulated.

Mutate diphtheria: Range between min value & the Quartile is very wide and has to be manipulated.

Mutate polio: Range between min value & the Quartile is very wide and has to be manipulated.

Correlations and Variances,

Looks like life\_expectancy dependent variable strong positive with alcohol (0.36), polio (0.43), diphtheria (0.44), schooling (0.56)



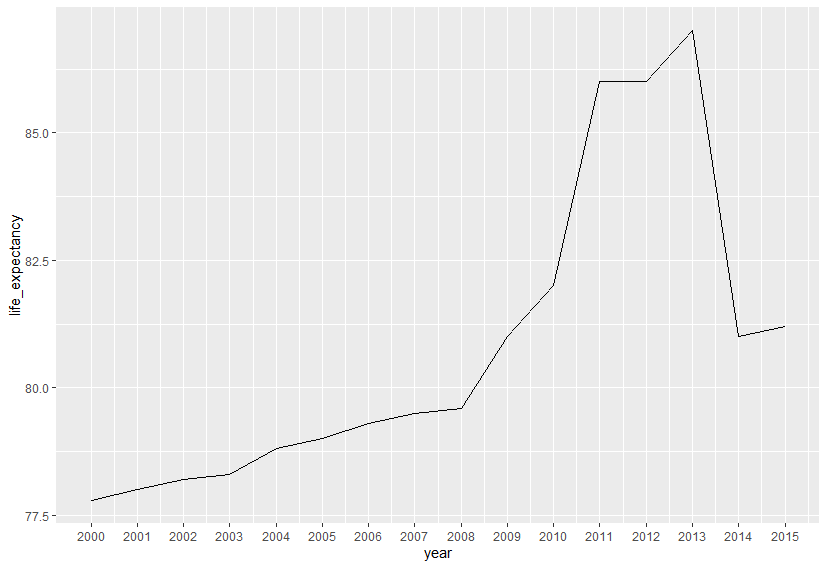
Taking one developed country data

We are going to take United Kingdom of Great Britain and Northern Ireland data

Here we are able to analyse that life expectancy is increasing till 2013 and suddenly it gets decreased.

WHY suddenly it gets decreased after 2013

Why its increasing till 2013



Correlation graph,



Looks like the dependent variable life expectancy is strongly correlated with measles (0.89), polio (0.82), diphtheria (0.82)

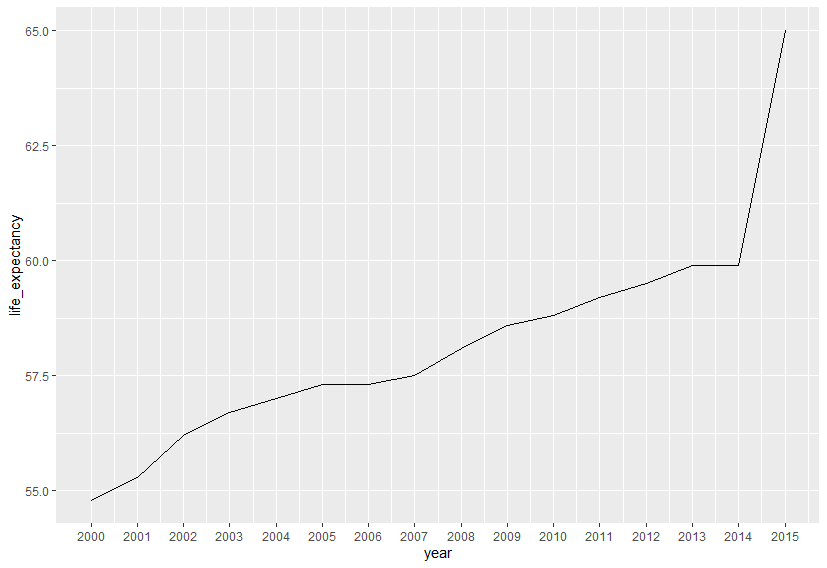
We are able to see that dependent variable life expectancy is negatively correlated with adult\_mortality, which is valid because if the mortality rate of adults is high, then the life expectancy of people will be low.

Now,

Taking one developing country data

We are going to take Afghanistan data

Here we are able to analyse that life expectancy is increasing till 2013 and suddenly it gets decreased. WHY suddenly it gets increased after 2013



Looks like the dependent variable life expectancy is strongly correlated with percentage\_expenditure (0.89), gdp (0.68), diphtheria (0.81), schooling(0.84)

We are able to see that dependent variable life expectancy is negatively correlated with adult\_mortality, which is valid because if the mortality rate of adults is high, then the life expectancy of people will be low.



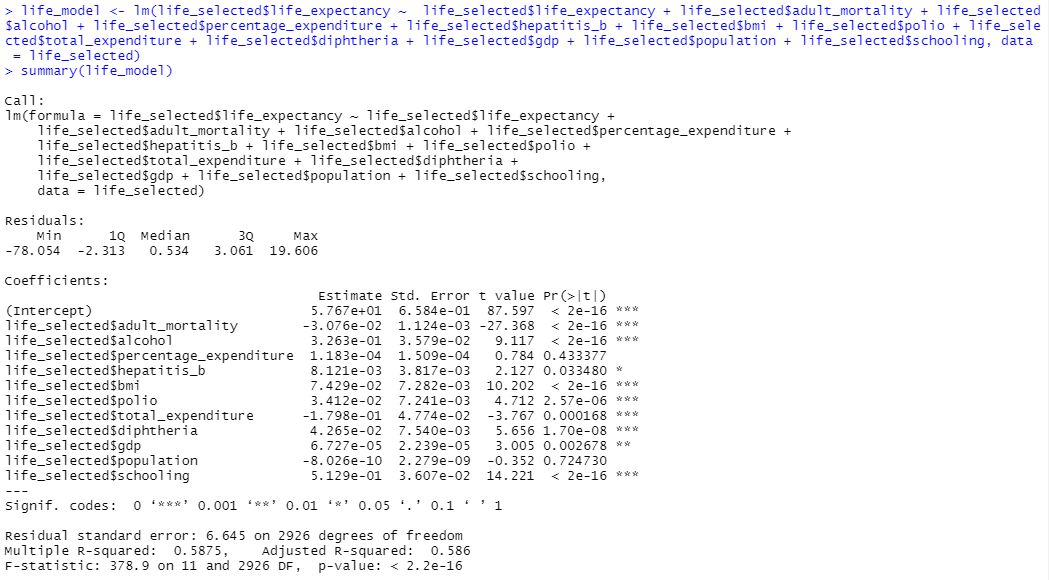
**Linear Regression Diagnostics:**

We are going to predict the Life.expectancy by using Selected Variables. And this is the full linear prediction model.

Coefficients interpretations: Regarding Coefficients the interesting analysis is adult\_mortality, total\_expenditure, population may give negative effects, Indicating additional of these variables may lead to decrease the Life expectancy. gdp and hepatitis\_b has a big positive effect on the Life expectancy.

Adj. R-squared interpretation: Approximately 58.6% of the observed variation can be explained by the model’s inputs, this is not better but we can say good results But still indicating that we are on the right path to create a good linear model.

Significance of Predictors: adult\_mortality, alcohol, polio, total\_expenditure, diphtheria and schooling As seen on the p-value and this are the most significant Predictors.



**Checking the statistical significance:**

Backward direction

The Adjusted R-squared and R-squared both are simpler approximately **82.46%**

Forward Direction

The Adjusted R-squared and R-squared both are simpler approximately **82.45%**

**Checking the Assumption:**

Plotting the residual on histogram like normal distribution

Checking the outliers, there are some outliers in the data, so we are taking the life expectancy below 50.

After removing the outlier, we are building the model. Unfortunately Adj R-squares fall from 0.8245 to 0.8047 so will try with original data with the outliers.

Trying to transform the data using log since we know that model backward is the best

fit.

The transformed data giving the Adjusted R-squared is smaller than the model backward we will not take the box-cox transformed log transformed is my new model

After testing all the independent variables are not correlated with each other, in the linearity test all the variables selected have linearity with the dependent variable.

**Conclusion:**

Linear model that fits to predict life expectancy based on the,

* Adj. R-Squared value
* Error Value
* pass 2 of 4 Assumption Check

Which is the Multicollinearity and Linearity Test.

Normality and Homoscedasticity do not give expected results.

Even when we look at the visualization the residuals plot seems following Normal Distribution & Homoscedasticity principle, but the statistical test gives different results.

Linear Model can be used to explain the linear correlation between life expectancy & the selected independent variables.

Since this model is highly sensitive to outliers (which quite massive occured in this data and taking it out is not a good option), it is very highly recommended to see the outliers pattern if you still wish to use this model on the new set of Life.expectancy data.