# ***LINEAR REGRESSION (R-PROJECT)***

Approach, steps and a brief summary about the project

The World Health Organization (WHO) released a health data set concerning life expectancy with many factors in the 4 significant categories being health, social, economic, mortality, and immunization. Moreover, the data contains information from 193 countries from the period 2000–2015 & was made public to perform health analysis. The final database was quite large, consisting of 23 Columns and 2938 rows.

## Following are the steps which were followed to create the model:

Explored the data and did analysis to find the independent variables which were co-related and has impact on the Target Variable (dependent variable) which was the next year's **Life expectancy**.

**Step 1.**
uploaded the **Raw data**.

**Step 2.**
Explored the dataset and categories all the variables into three categories. Continuous, Categorical and Qualitative Columns.

Finding total no. of unique values in each variable at once

```
lengths(lapply(InputData,unique))
```

## We have distributed all the variables in three categories as ContinousCols(Unique values>20), CategoricalCols(Unique values<20) and Qualitative Columns.

```
ContinousCols=('Life_Expectancy','Adult_Mortality','Infant_Deaths','Alcohol','Percentage_Expenditure','Measles','BMI','Under.five_Deaths','Polio','Total_Expenditure','Diphtheria','HIV.AIDS','GDP','Per_Capita_GDP','Population','Thinness_1.19_Years','Thinness_5.9_Years','Income_Composition_of_Resources','Schooling')
```

```
CategoricalCols= ('Year','Status')
```

**Two columns, (Hepatitis_B and Country)** were considered as qualitative because **'Country'** having factor of 192 levels. We do not consider factor level more than 30.

**Hepatitis_B** – It is having 553 missing values as it isn't good to treat so many missing values as it will create dummy variable.

**Step 3. Identify the problem.**
Predicting next year's life expectancy value by making regression model.

**Step 4.**
Target Variable is Life_Expentancy.

**Step 5.**
The Target variable is a continuous variable hence we did Linear Regression.

** Linear Regression - Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

**Step 6.**

**Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

**Explored each and every potential predictor for quality and distribution.**

Explored the data set using **Histograms (Continuous Columns) and Bar plot (Categorical Columns).**

**Explore each potential predictor for distribution and quality**

**Library to generate professional colors**

```
library(RColorBrewer)
```

**Histogram for multiple Column at once**

**For splitting windows**

```
par(mfrow=c(5,4))


ColsForHist=c('Life_Expectancy','Adult_Mortality','Infant_Deaths','Alcohol','Percentage_Exp
enditure','Measles','BMI','Under.five_Deaths','Polio','Total_Expenditure','Diphtheria','HIV.AI
DS','GDP','Per_Capita_GDP','Population','Thinness_1.19_Years','Thinness_5.9_Years',
'Income_Composition_of_Resources','Schooling')
```

**Looping to create the histograms for each column**

```
for (contCol in ColsForHist) {
        hist(InputData[,c(contCol)], main=paste('Histogram of:', contCol),
        col=brewer.pal(8,"Paired"))

}
```

**Treating Outliers**

Treating outliers using 12500 as the substitution

```
#Percentage_Expenditure

hist(InputData[InputData$Percentage_Expenditure<12000,'Percentage_Expenditure'])

InputData[InputData$Percentage_Expenditure>12000,'Percentage_Expenditure']=12500

hist(InputData$Percentage_Expenditure)
```

**These columns have poor quality distribution**

Infant_Deaths, Measles, Under.five_Deaths, HIV.AIDS, GDP, Population

**Exploring MULTIPLE CATEGORICAL features**

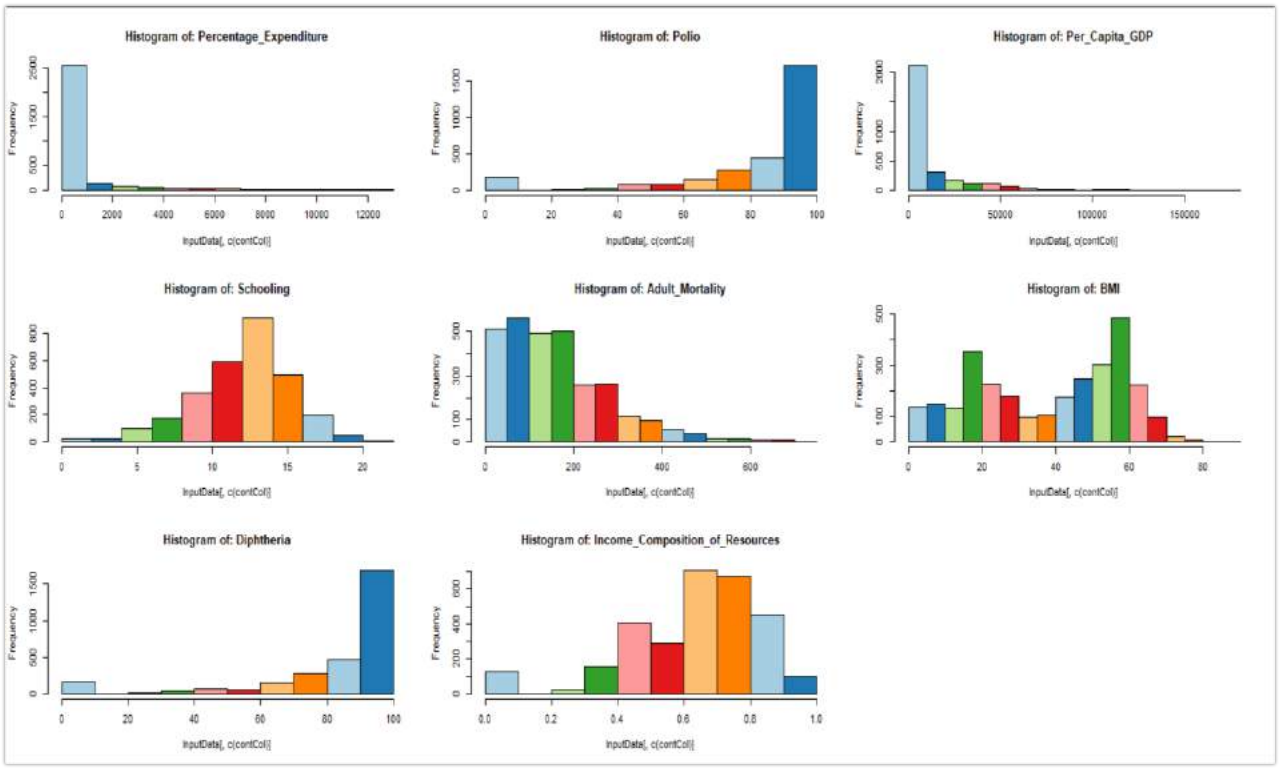
**Splitting the plot window into two parts**

```
par(mfrow=c(2,1))
```

**looping to create the Bar-Plots for each column**

```
CategoricalCols=c('Year','Status')


for (catCols in CategoricalCols){
 barplot(table(InputData[,c(catCols)]),  main=paste('BarPlot of:', catCols),
 col=brewer.pal(8,'Paired'))
}
```




**Data Distribution**

**Quality of distribution of Year is not good as it showing no variance**

**We will leave "Year"**

**Step 7.**

**Visual Relationship between Predictors and Target variable.**

Continuous Vs Continuous -- Scatter Plot.

Continuous Vs Categorical --- Box Plot.

**For multiple continuous columns at once-- Scatter Plot**

plot(x=InputData$Schooling, y=InputData$Life_Expectancy,col='blue')

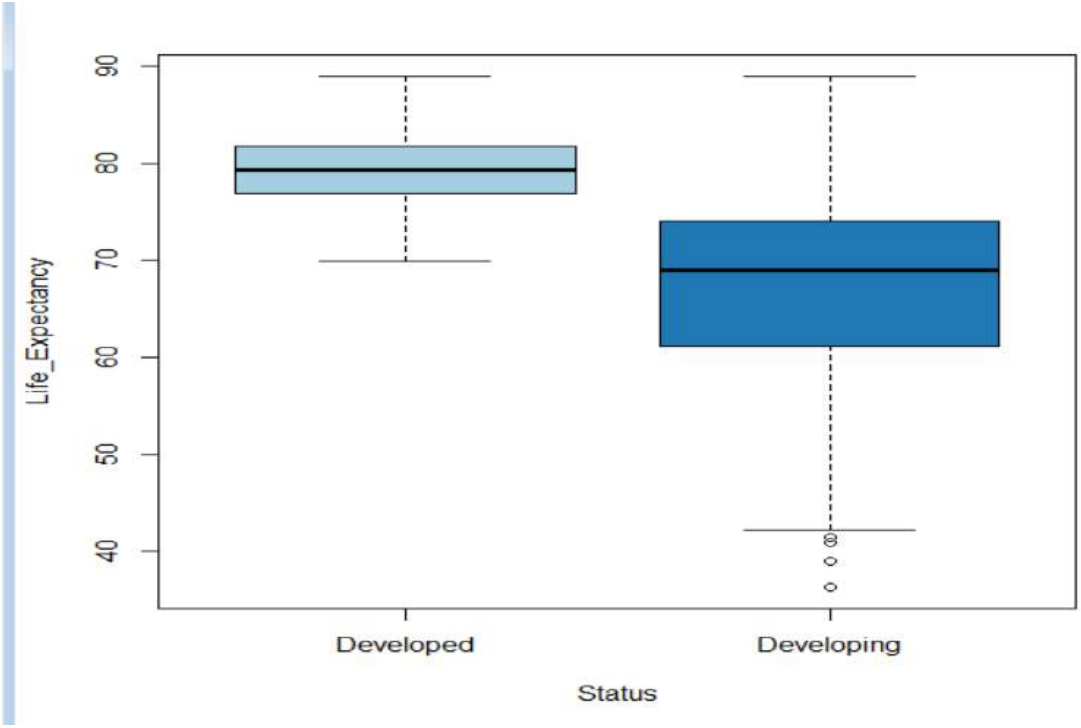
ContinousCols=c('Life_Expectancy','Adult_Mortality','Infant_Deaths','Alcohol','Percentage_Expenditure','Measles','BMI','Under.five_Deaths','Polio','Total_Expenditure','Diphtheria','HIV.AIDS','GDP','Per_Capita_GDP','Population','Thinness_1.19_Years','Thinness_5.9_Years','Income_Composition_of_Resources','Schooling')

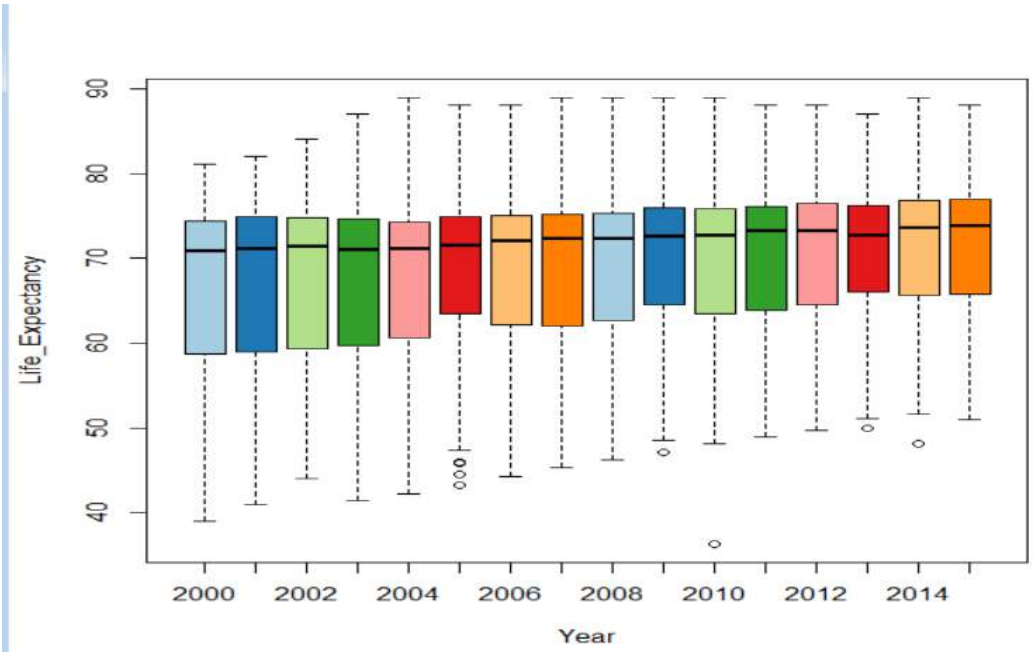
plot(InputData[,ContinousCols],col='blue')

**Continuous Vs Categorical Visual analysis: Boxplot**

boxplot(Life_Expectancy~Status,InputData,col=brewer.pal(8,"Paired"))

boxplot(Life_Expectancy~Year,InputData,col=brewer.pal(8,"Paired"))






**Year-No Variance**

**Step 8.**

**Statistical Relationship between predictors and target variable**
Continuous Vs Continuous ---- Correlation
Categorical Vs Continuous --- ANOVA

**Multiple Continuous Column at once-Correlation**

```
ColsForCor=c('Life_Expectancy','Adult_Mortality','Infant_Deaths','Alcohol','Percentage_Expe
nditure','Measles','BMI','Under.five_Deaths','Polio','Total_Expenditure','Diphtheria','HIV.AID
S','GDP','Per_Capita_GDP','Population','Thinness_1.19_Years','Thinness_5.9_Years','Income’,
Composition_of_Resources','Schooling')

CorData= cor(InputData[,c(ColsForCor)], use='complete.obs')

CorData
```

**Final columns which have high correlation with the target variable**

```
names(CorrData[,'Life_Expectancy'][abs(CorrData[,'Life_Expectancy'])>0.5])
```

Life_Expectancy
Adult_Mortality
BMI
HIV.AIDS
Per_Capita_GDP
Income_Composition_of_Resources
Schooling

**These variables get selected as it has 50% correlation with target variable**

**Continuous Vs Categorical correlation strength: ANOVA**

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples

F-Statistic is Mean Sq error/residual Mean Square error
H0: Variables are NOT correlated
Small P-Value--> Variables are correlated (H0 is rejected)
Large P-Value--> Variables are NOT correlated (H0 is accepted)

**ANOVA test for multiple columns at once**

```
ColsForAnova= c('Year','Status')


for (catCol in ColsForAnova){
        anovaData= InputData[, c('Life_Expectancy', catCol)]
        print(str(anovaData))
        print(summary(aov(Life_Expectancy~., anovaData)))

}
```

We also did our ANOVA test to check if our Ho assumption was true or false.

```
'data.frame':   2938 obs. of  2 variables:
 $ Life_Expectancy: num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Year           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
NULL
             Df Sum Sq Mean Sq F value              Pr(>F)
Year          1   7676    7676   87.11 <0.0000000000000002 ***
Residuals  2926 257815      88
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
10 observations deleted due to missingness
'data.frame':   2938 obs. of  2 variables:
 $ Life_Expectancy: num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Status         : chr  "Developing" "Developing" "Developing" "Developing" ...
NULL
             Df Sum Sq Mean Sq F value              Pr(>F)
Status        1  61715   61715   886.2 <0.0000000000000002 ***
Residuals  2926 203776      70
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
10 observations deleted due to missingness
```

The P-value after ANOVA test was 0.0000000000000002 so we rejected our Ho because the variables were correlated.

**Step 9.**
**Checked and treated the missing values.**

colSums(is.na(InputData))

**Instead of deleting, we treated individual columns using data imputation strategies as well.**

**We treated all the missing values with the median of that variable.**

**Step 10.**

**Generated the Data frame for Machine learning.**

TargetVariableName=c('Life_Expectancy')

TargetVariable=InputData[,c(TargetVariableName)]

**Choosing multiple Predictors which may have relation with Target Variable, based on the exploratory data analysis.**

BestPredictorName=c('Status','Adult_Mortality','Percentage_Expenditure','BMI','Polio','Diphtheria','Per_Capita_GDP','Income_Composition_of_Resources','Schooling')

```
PredictorVariables= InputData[,c(BestPredictorName)]

DataForML=data.frame(TargetVariable,PredictorVariables)

head(DataForML)

str(DataForML)
```

Converted character to factor

```
DataForML$Status=as.factor(DataForML$Status)
```

**Data Splitting:**

****Data splitting** is the act of partitioning available **data** into. two portions, usually for cross-validatory purposes. One. portion of the **data** is used to develop a predictive model. and the other to evaluate the model's performance.

We split out data into two portions: For training it's 70% and 30% for testing.

```
TrainingSamplingIndex= sample(1:nrow(DataForML),size=0.7*nrow(DataForML))

DataForMLTrain=DataForML[TrainingSamplingIndex,]

DataForMLTest=DataForML[-TrainingSamplingIndex,]

dim(DataForMLTrain)

dim(DataForMLTest)
```

**LINEAR REGRESSION:**

Regression analysis is described as "*Using the relationship between variables to find the best fit line or the regression equation that can be used to make predictions.*"

The regression analysis of the income composition of resources & life expectancy is a form of predictive modelling whose purpose is to investigate the relationship between a dependent and independent variable. In this case, the income composition of resources is the independent variable, and life expectancy is the dependent variable.

**Creating Predictive models on training data to check the accuracy of each algorithm**

```
startTime=Sys.time()

Model_Reg=lm(TargetVariable~.,data=DataForMLTrain)
summary(Model_Reg)
endTime=Sys.time()

endTime-startTime
```

```
Call:
lm(formula = TargetVariable ~ ., data = DataForMLTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9052  -2.1377   0.1114   2.4665  17.4651

Coefficients:
                              Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                 53.64966815  0.76231511  70.377 < 0.0000000000000002 ***
StatusDeveloping            -1.15540091  0.34468910  -3.352             0.000817 ***
Adult_Mortality             -0.03026954  0.00096434 -31.389 < 0.0000000000000002 ***
Percentage_Expenditure       0.00010820  0.00007138   1.516             0.129741
BMI                          0.06059282  0.00597123  10.147 < 0.0000000000000002 ***
Polio                        0.03579882  0.00594260   6.024  0.00000000201107547 ***
Diphtheria                   0.04594143  0.00588107   7.812  0.00000000000000894 ***
Per_Capita_GDP               0.00003197  0.00000870   3.675             0.000244 ***
Income_Composition_of_Resources  7.86700468  0.88520534   8.887 < 0.0000000000000002 ***
Schooling                    0.59503153  0.05523282  10.773 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.538 on 2046 degrees of freedom
Multiple R-squared:  0.7669,    Adjusted R-squared:  0.7659
F-statistic:   748 on 9 and 2046 DF,  p-value: < 0.00000000000000022

> endTime=Sys.time()
>
> endTime-startTime
Time difference of 0.10028 secs
>
```

**Multiple R-squared: 0.7669, Adjusted R-squared: 0.7659**
**F-statistic: 748 on 9 and 2046 DF, p-value: < 0.00000000000000022**

**Durbin Watson test**

durbinWatsonTest(Model_Reg)

lag Autocorrelation D-W Statistic p-value
1 -0.002591235 2.005028 0.944
Alternative hypothesis: rho != 0

**We checked the accuracy of each Algorithm.**

**Rejecting the Variables with High probability and accepting the ones which have probability close to zero.**

**Checked accuracy of the model on Testing Data.**

```
DataForMLTest$Predict_LM=predict(Model_Reg,DataForMLTest)

head(DataForMLTest)
```

**Calculating the Absolute Percentage Error for each prediction**

```
LM_APE=100*(abs(DataForMLTest$Predict_LM -
DataForMLTest$TargetVariable)/DataForMLTest$TargetVariable)

print(paste('## Mean Accuracy of Linear Model is:', 100 - mean(LM_APE)))

print(paste('## Median Accuracy of Linear Model is:', 100 - median(LM_APE)))
```

**## Mean Accuracy of Linear Model is: 94.9067884952327**
**## Median Accuracy of Linear Model is: 96.6434916893895**

## Insights:

Countless factors play into the country's average life expectancy and how long somebody may live. Making predictions using machine learning regressions gives us more insight into immunization and social factors we don't often think about. By understanding correlations between multiple variables, we learn about the factors that affect lifespan the most.

Using this model, I created, one thing we can do is predict a country's life expectancy as infrastructure, society, and resources change over the years. We can also experiment with certain factors and their correlation to a high life expectancy.

Most importantly though, we can utilize the model and dataset to find areas of improvement in specific countries. By knowing which factors have the most significant role in a lower life expectancy, a country can decide to spend more money & resource on those certain things. Without a doubt, it's crucial for nations to know where they are lacking & can do better, from factors in health, social, economic, mortality, and immunization categories. The more

knowledge we have about the importance of these factors, the more success we will have when it comes to extending a country's average life expectancy and creating a better quality of life for those who live there.

---

**Thanks & Regards,**

**S M Amanullah**