METRO COLLEGE OF TECHNOLOGY
R-PROJECT REPORT

LIFE EXPECTANCY DATA SET
COLLECTED OVER 2000-2015 FROM 'W.H.O' DATA REPOSITORY
FOR 193 COUNTRIES.

EXPLORATORY DATA ANALYSIS

BY SHAHINA KHURAISHI

Shahina Khuraishi

[COMPANY NAME] [Company address]

HOW AND FROM WHERE THE DATA IS COLLECTED :

The project relies on accuracy of data. The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The data-sets are made available to public for the purpose of health data analysis. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. The data-set related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website. It has been observed that in the past 15 years , there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years.

Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The final merged file consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:Immunization related factors, Mortality factors, Economical factors and Social factors.

By going through the above meta data,

Cross-Industry Standard Process for Data Mining (CRISP-DM)

This model consisting of six phases. It is a cyclical process that provides a structured approach for the data mining process. The six phases can be implemented in any order but it would sometimes require backtracking to the previous steps and repetition of actions.

The six phases of CRISP-DM include:

1. BUISNESS UNDERSTANDING
2. DATA UNDERSTANDING
3. DATA PREPARATION
4. MODELLING
5. EVALUATION
6. DEPLOYMENT

1. BUISNESS UNDERSTANDING :

This project is based on factors affecting life expectancy considering demographic variables, income composition and mortality rates

In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well.

Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower the value of life expectancy.

This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

2. DATA UNDERSTANDING :

```r
#   Changing Directory
###############################################################
setwd("C:\\Users\\shahe\\Desktop\\R files\\FinalProject")
getwd()
###############################################################
```

```r
###############################################################
# Installing the package to read .xlsx file
###############################################################
install.packages("openxlsx")
library(openxlsx) # activating the library to use it
```

If the file is .csv we don’t have to install any package. As, it has inbuilt package available in R.

We can directly read the file by read.csv().

Reading the .xlsx file into ‘R’ environment

```r
df1<-read.xlsx("life_Expectency.xlsx",sheet = "Life Expectancy Data", colNames = TRUE, startRow = 1 )
```

Checking the Shape of the dataset :

```r
> dim(df1)                         # Checking the shape of the Data set
[1] 2938   22
```

Checking the first three (head) observations of the data set :

```r
> head(df1,3)                    # Checking  the head of the Data set
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles  BMI under-five.deaths Poli
1 Afghanistan 2015 Developing            65.0             263            62    0.01               71.27962          65    1154 19.1                83    6
2 Afghanistan 2014 Developing            59.9             271            64    0.01               73.52358          62     492 18.6                86   58
3 Afghanistan 2013 Developing            59.9             268            66    0.01               73.21924          64     430 18.1                89   62
  Total.expenditure Diphtheria HIV/AIDS      GDP Population thinness.1-19.years thinness.5-9.years Income.composition.of.resources Schooling
1              8.16         65      0.1 584.2592   33736494                17.2               17.3                           0.479      10.1
2              8.18         62      0.1 612.6965     327582                17.5               17.5                           0.476      10.0
3              8.13         64      0.1 631.7450   31731688                17.7               17.7                           0.470       9.9
```

Checking the last three (tail) observations of the data set :

```
> tail(df1,3)                       # Checking the tail of the Data set
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles  BMI under-five.deaths Polio
2936 Zimbabwe 2002 Developing            44.8              73            25    4.43                      0          73     304 26.3                40    73
2937 Zimbabwe 2001 Developing            45.3             686            25    1.72                      0          76     529 25.9                39    76
2938 Zimbabwe 2000 Developing            46.0             665            24    1.68                      0          79    1483 25.5                39    78
     Total.expenditure Diphtheria HIV/AIDS       GDP Population thinness.1-19.years thinness.5-9.years Income.composition.of.resources Schooling
2936              6.53         71     39.8  57.34834     125525                 1.2                1.3                           0.427      10.0
2937              6.16         75     42.1 548.58731   12366165                 1.6                1.7                           0.427       9.8
2938              7.10         78     43.5 547.35888   12222251                11.0               11.2                           0.434       9.8
```

Checking the Structure of the Data Set :

```
> str(df1)                         # To visualize the structure of DATA
'data.frame':	2938 obs. of  22 variables:
 $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ Year                           : num  2015 2014 2013 2012 2011 ...
 $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : num  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : num  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : num  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : num  1154 492 430 2787 3013 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ under-five.deaths              : num  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : num  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : num  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV/AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness.1-19.years            : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5-9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
>
```

Changing the column names which are not according to R standards :

```
colnames(df1)[c(12,16,19,20)]<-c("Under.five.deaths","HIV.AIDS","thinness.1_19.years","thinness.5_9.years")
```

Dropping Features : As we have no related information to extract meaningful data from it.

```
#dropping features as we don't have any knowledge to extract meaning full features from them

df1[,c("thinness.1_19.years","thinness.5_9.years")]<-NULL
```

Finding the Total number of duplicated data.

Duplicates are meaningful and useful for Data Analysis.

So, I am not deleting them.

```
#How many duplicated data are there?
sum(duplicated((df1))) # gives total number of duplicate data values, NO DUPLICATES FOUND
#r1<-which(duplicated(df1)) # gives row numbers of duplicate values
#df1<-df1[-r1,]   # removing the rows with duplicate values.
```

Handling Missing values :

```
df1[df1==' ']<-NA            # assigning missing values with 'NA'

sum(is.na(df1))              # To get total number of missing values
```

2495 missing values are found.

➔colsums(is.na(df1)) To get the number of missing values column wise.

```
> colSums(is.na(df1))            # To get the number of missing values column wise
                        Country                      Year                    Status
                              0                         0                         0
                Life.expectancy           Adult.Mortality             infant.deaths
                              0                         0                         0
                        Alcohol    percentage.expenditure               Hepatitis.B
                            193                         0                       553
                        Measles                       BMI         Under.five.deaths
                              0                        32                         0
                          Polio         Total.expenditure                Diphtheria
                             19                       226                        19
                       HIV.AIDS                       GDP                Population
                              0                       443                       644
Income.composition.of.resources                 Schooling            lifeExp.agegroup
                            160                       160                         0
>
```

➔To find the percentage of missing values each column has

```
> #Percentage of missing values:
> round(colMeans(is.na(df1))*100,2)# '2' represents the no. of digits after decimal
                        Country                      Year                    Status
                           0.00                      0.00                      0.00
                Life.expectancy           Adult.Mortality             infant.deaths
                           0.00                      0.00                      0.00
                        Alcohol    percentage.expenditure               Hepatitis.B
                           6.59                      0.00                     18.89
                        Measles                       BMI         Under.five.deaths
                           0.00                      1.09                      0.00
                          Polio         Total.expenditure                Diphtheria
                           0.65                      7.72                      0.65
                       HIV.AIDS                       GDP                Population
                           0.00                     15.13                     21.99
Income.composition.of.resources                 Schooling            lifeExp.agegroup
                           5.46                      5.46                      0.00
>
```

Interpretation from DATA understanding :

- We have 2930 observations and 22 Variables (columns)
- Life.Expectancy is the Target (Response) variable which is Continuous (Numeric) data type.

- Adult.Mortality is our second Target variable which is Continuous (Numeric) data type.
- And all other variables Status, Year , Measles, Hepatitis, Alcohol…are predicting variables.

Feature Engineering : Extracting variables from existing variables.

I am adding two categorical columns to the data set to make it meaningful.

```
df1$lifeExp.agegroup<-NA
df1

f1=function(x){
  if (is.na(x)) "N/A"
  else if (x<25)     "< 25"
  else if (x<= 35) "25-35"
  else if (x<= 45) "36-45"
  else if (x<= 55) "46-55"
  else if (x<= 65) "56-65"
  else if (x<= 75) "66-75"
  else if (x<= 85) "76-85"
  else if (x<= 95) "86-95"
  else              "95+"
}
# applying the function to 'life expectancy' column using 'sapply'
df1$lifeExp.agegroup<-sapply(df1$Life.expectancy ,f1)
df1
```

The column ‘lifeExp.agegroup’ is added to the Data set

```
$ Hepatitis.B           : num
$ Measles               : num
$ BMI                   : num
$ Under.five.deaths     : num
$ Polio                 : num
$ Total.expenditure     : num
$ Diphtheria            : num
$ HIV.AIDS              : num
$ GDP                   : num
$ Population            : num
$ lifeExp.agegroup      : chr
```

```
# Adding Another Categorical column to the Data S

df1$Year.groups<-NA
str(df1)

f2=function(x)  {
  if (x>=2000 && x<=2003) "2000-2003"
   else if (x>=2004 && x<=2007) "2004=2007"
    else if (x>=2008 && x<=2011) "2008-2011"
     else if (x>=2012 && x<=2015) "2012-2015"
}

df1$Year.groups<-sapply(df1$Year,f2)
str(df1)
df1
```

UNIVARIATE ANALYSIS for Categorical Variables : We have 3 categorical variables in this data set. 1. Country 2. Status 3. Life.Exp.agegroup

Summarization : table of frequency or percentage

Visualization : pie chart or bar chart

Making a copy of the data set before doing any changes.

```
df_org1<-df1                    # making a copy of the Data Set
```

1. 'Country'

Summarization : table of frequency or percentage

```
#1.SUMMARIZING Categorical Variables
sum(is.na(df1$Country)) # No missing values found.

levels(as.factor(df1$Country)) # give levels (different values) for column 'Country'
                               # shows Names of 193 countries
tb1<-table(df1$Country) #Viewing the frequency of each 'Country' which should be 16 for all countries
tb1                     #as we are considering data for 16 years.
                        # But, found frequency of '1' for few (10) countries.
```

- Levels gives the names of 193 countries.
- Since we are considering the data for 16 years (2000 – 2016).
- The frequency of each country should be '16'.
- But, 10 countries are found with '1' frequency.
- To, make the dataset balanced. I removed these countries

```
# Dropping the Rows with frequency '1'.

df1<-df1[-c(625,770,1651,1716,1813,1910,1959,2168,2217,2714), ]
```

```
levels(as.factor(df1$Country))
tb1<-table(df1$Country)
tb1
```

Now I have 183 countries with their frequency 16.

2. Status
   Summarization :

```
levels(as.factor(df1$Status))  # gives 2 levels for column Status.
] "Developed"  "Developing"
```

```
tb2<-table(df1$Status)    #Viewing the frequency of each level of 'Status'
tb2

Developed Developing
      512       2416
```

Since, it is a big dataset , I am dividing it into two parts ‘Developing’ and ‘Developed’ using subset().

```
sum(with(df1,Status == "Developing"))  #  Total number of observations with status ' Developing
developing<-subset(df1, Status == "Developing")  # Extracting the Data of "Developing" countrie
df2<-developing
```

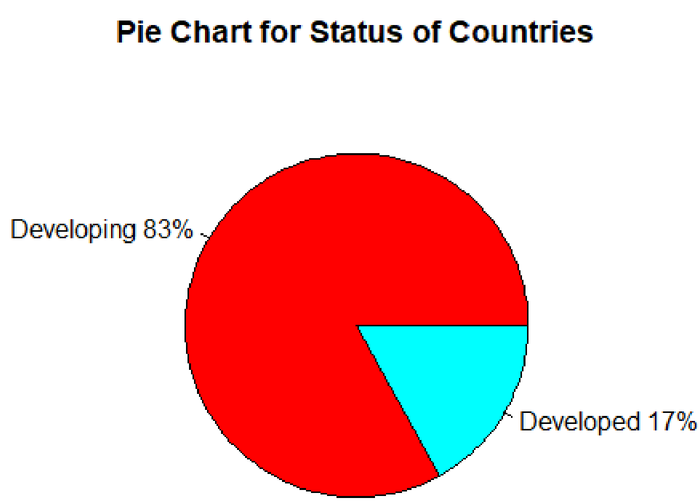
```
sum(with(df1,Status == "Developed"))
developed<-subset(df1, Status == "Developed")  # Extracting the Data of "Developed" countries
df3<-developed
```

```
# To view the levels and frequency
levels(as.factor(developing$Country))  # names of 161  developing countries found

tb3<-table(developing$Country)
tb3

levels(as.factor(developed$Country))  # names of 32 developed countries found

tb4<-table(developed$Country)
tb4
```

Visualization : pie chart or bar chart,...

Code for Pie chart: for viewing Status Categorical variable.

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

freq1 <- c(161,32)
pct <- round(freq1/sum(freq1)*100)
lbls <- c("Developing", "Developed")
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(pct,labels = lbls, col=rainbow(length(lbls)),#length(lbls) = 5
    main="Pie Chart for Status of Countries")
```



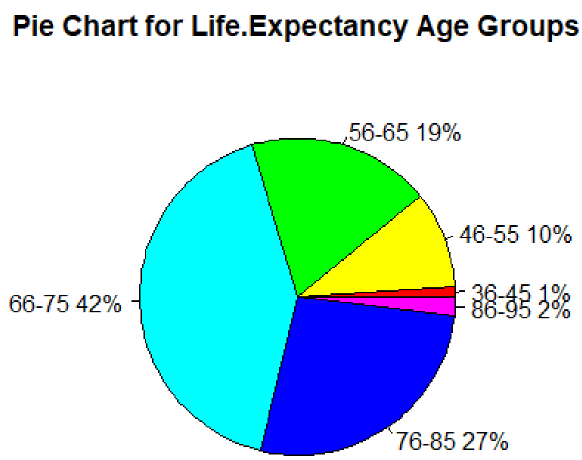
3. lifeExp.agegroup
   Summarization :

```r
levels(as.factor(df1$lifeExp.agegroup))  # 6 levels are retured
] "36-45" "46-55" "56-65" "66-75" "76-85" "86-95"
```

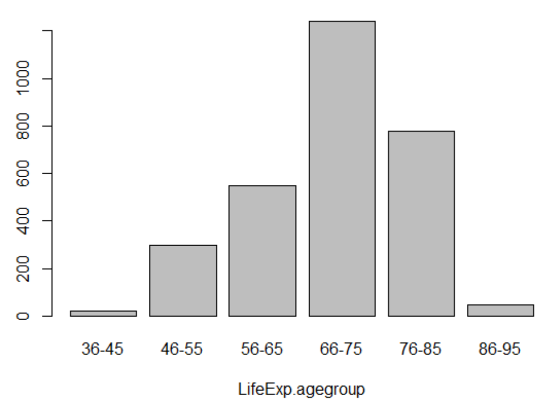
```r
> tb5<-table(df1$lifeExp.agegroup)        # frequency of life expectancy of each agegroup
> tb5

36-45 46-55 56-65 66-75 76-85 86-95
   19   296   549  1240   779    45
```

Visualization :

```r
# VISUALIZATION BY PIE CHART
par(mfrow = c(1, 1))

freq1 <- c(19,296,549,1240,779,45)
pct <- round(freq1/sum(freq1)*100)
lbls <- c("36-45","46-55","56-65","66-75","76-85","86-95")
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(pct,labels = lbls, col=rainbow(length(lbls)),#length(lbls) = 5
    main="Pie Chart for Life.Expectancy Age Groups")
```




```r
# 2.VISUALIZATION BY BARCHART
tblag<-table(df1$lifeExp.agegroup)
tblag
barplot(tblag,ylab="lifeExp.agegroup",horiz=T)
barplot(tblag,xlab="LifeExp.agegroup")
```



UNIVARIATE ANALYSIS FOR CONTINOUS (NUMERIC) VARIABLES :

Summarization : Central tendency(Mean, Median, Mode, Min, Max, 25th percentile, 75th percentile,Standard Deviation, Variance.

Visualization : Histogram, Densityplot, boxplot

Removing Categorical variables

```
str(df1)
dataset<-df1[,-c(1,3,21)]
str(dataset)
```

To check the Summary of all Numerical Variables.

```
summary(dataset)
```

```
      Year      Life.expectancy Adult.Mortality infant.deaths      Alcohol        percentage.expenditure
 Min.   :2000   Min.   :36.30   Min.   :  1.0   Min.   :   0.00   Min.   : 0.010   Min.   :    0.000
 1st Qu.:2004   1st Qu.:63.10   1st Qu.: 74.0   1st Qu.:   0.00   1st Qu.: 0.905   1st Qu.:    4.854
 Median :2008   Median :72.10   Median :144.0   Median :   3.00   Median : 3.770   Median :   65.611
 Mean   :2008   Mean   :69.22   Mean   :164.8   Mean   :  30.41   Mean   : 4.615   Mean   :  740.321
 3rd Qu.:2011   3rd Qu.:75.70   3rd Qu.:228.0   3rd Qu.:  22.00   3rd Qu.: 7.715   3rd Qu.:  442.614
 Max.   :2015   Max.   :89.00   Max.   :723.0   Max.   :1800.00   Max.   :17.870   Max.   :19479.912
                                                                  NA's   :193
```

```
 Hepatitis.B       Measles              BMI        Under.five.deaths     Polio        Total.expenditur
 Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00   Min.   : 3.00   Min.   : 0.37
 1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00   1st Qu.:78.00   1st Qu.: 4.26
 Median :92.00   Median :    17.0   Median :43.35   Median :   4.00   Median :93.00   Median : 5.75
 Mean   :80.96   Mean   :  2427.9   Mean   :38.24   Mean   :  42.18   Mean   :82.55   Mean   : 5.93
 3rd Qu.:97.00   3rd Qu.:   362.2   3rd Qu.:56.10   3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.: 7.49
 Max.   :99.00   Max.   :212183.0   Max.   :77.60   Max.   :2500.00   Max.   :99.00   Max.   :17.60
 NA's   :553                        NA's   :32                        NA's   :19      NA's   :226
```

```
   Diphtheria       HIV.AIDS           GDP              Population         Income.composition.of.resources
 Min.   : 2.00   Min.   : 0.100   Min.   :     1.68   Min.   :3.400e+01   Min.   :0.0000
 1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.85   1st Qu.:1.967e+05   1st Qu.:0.4930
 Median :93.00   Median : 0.100   Median :  1764.97   Median :1.392e+06   Median :0.6770
 Mean   :82.32   Mean   : 1.748   Mean   :  7494.21   Mean   :1.276e+07   Mean   :0.6274
 3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5932.90   3rd Qu.:7.427e+06   3rd Qu.:0.7792
 Max.   :99.00   Max.   :50.600   Max.   :119172.74   Max.   :1.294e+09   Max.   :0.9480
 NA's   :19                       NA's   :443         NA's   :644         NA's   :160
```

```
   Schooling
 Min.   : 0.0
 1st Qu.:10.1
 Median :12.3
 Mean   :12.0
 3rd Qu.:14.3
 Max.   :20.7
 NA's   :160
```

Interpretation :

In some variables like Adult.Mortality, infant deaths, percentage.expenditure, measels, under-five deaths there is a huge difference between the 3rd quartile and the maximum value.

Which can be considered as OUTLIERS, But, I am not considering them as outliers as the data belong to 183 different countries and each country have its own factors affecting these values.

Visualizing Target Variable :

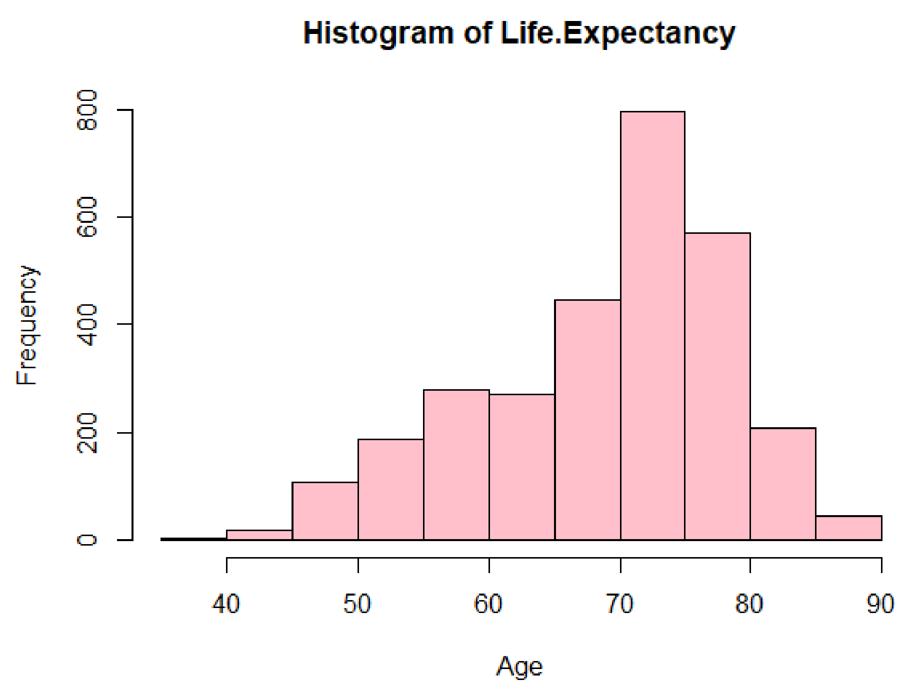
Q. 1. What is the distribution of Target variable. Which is Numeric.

By seeing the Histogram we can say that it is Normal with positive kurtosis.

Mean is 69.22 and standard Deviation is 9.5

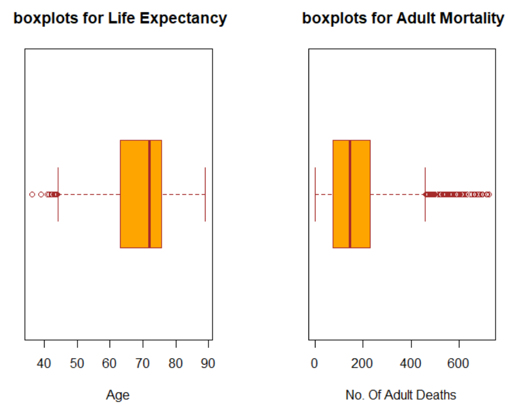
```
sd(df1$Life.expectancy, na.rm = TRUE)
[1] 9.523867
```

```
# 1. Problem : What is the distribution of Target (Numerical) variable
# Answer : By seeing the Histogram we can say that it is Normal with positive Kurtosis.
hist(df1$Life.expectancy,br=14,col="pink",xlab="Age",ylab="Frequency",
     freq=TRUE,main="Histogram of Life.Expectancy")
```



Visualization of Numeric Variables

Can see the Outliers Clearly in BOX



BIVARIATE ANALYSIS : CATEGORICAL VS. CATEGORICAL

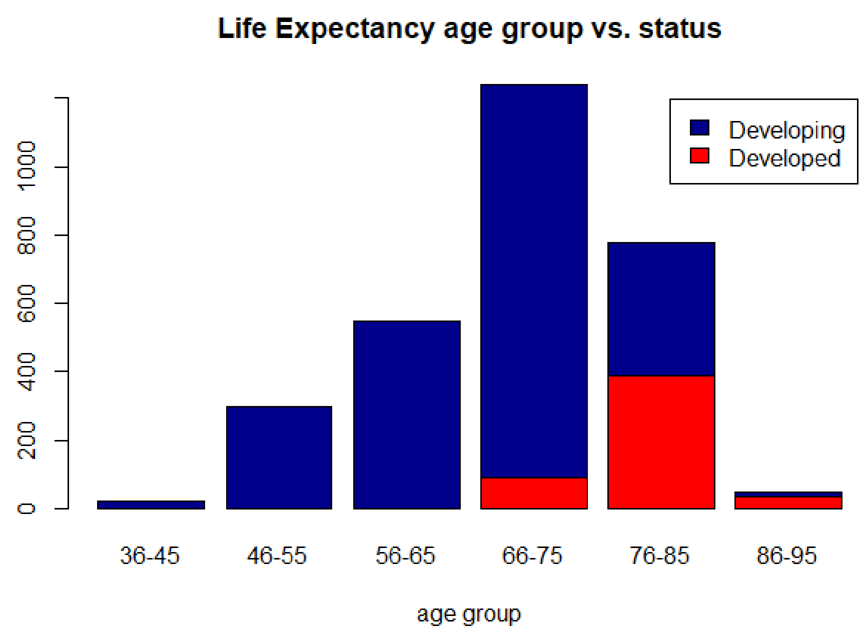
Summarizing : using Contingency Table

```
> tbl_ag_S<-xtabs(~ lifeExp.agegroup + Status, data=df1)
> tbl_ag_S
                Status
lifeExp.agegroup Developed Developing
           36-45         0         19
           46-55         0        296
           56-65         0        549
           66-75        90       1150
           76-85       387        392
           86-95        35         10
```

```
> tbl_ag_S.t<-t(tbl_ag_S)   # transpose tbl_ag_S
> tbl_ag_S.t
            lifeExp.agegroup
Status        36-45 46-55 56-65 66-75 76-85 86-95
  Developed       0     0     0    90   387    35
  Developing     19   296   549  1150   392    10
```

VISUALIZATION : Stacked bar plot

```
# stacked bar plot
par(mfrow = c(1, 1))
barplot(tbl_ag_S.t, main="Life Expectancy age group vs. status",
        xlab="age group", col=c("red","darkblue"),
        legend = rownames(tbl_ag_S.t))
```



2. Is there any relationship between the above 2 categorical variables. We need chisquare test for finding this.
   HO : No relation between 'lifeExp.agegroup' and 'Status'

```
> chisq.test(tbl_ag_S)

        Pearson's Chi-squared test

data:  tbl_ag_S
X-squared = 945.91, df = 5, p-value < 2.2e-16
```

Since p-value is less than 0.05 significance level. We reject the NULL hypothesis.
There is a relation between 'lifeExp.agegroup' and 'Status'

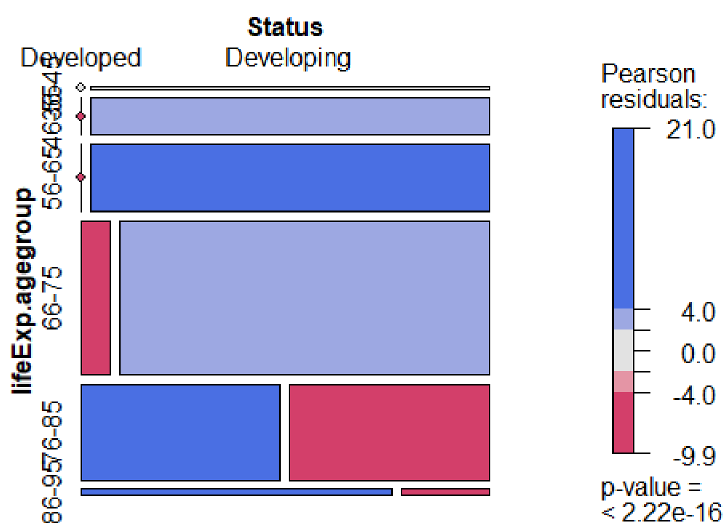
```
> prop.table(table(df1$lifeExp.agegroup, df1$Status))#calculates  probability of frequenc

         Developed   Developing
  36-45 0.000000000 0.006489071
  46-55 0.000000000 0.101092896
  56-65 0.000000000 0.187500000
  66-75 0.030737705 0.392759563
  76-85 0.132172131 0.133879781
  86-95 0.011953552 0.003415301
```

Since, the probability of contingency table is ‘0’. The 2 variables are DEPENDENT.

Mosaic plots provide a way to visualize contingency tables.

A mosaic plot is a visual representation of the association between two variables.

```
mosaic(tbl_ag_S, shade=TRUE, legend=TRUE)
```



BIVARIATE ANALYSIS : NUMERIC Vs. CATEGORICAL

Summarizing : Using aggregate function or t-apply

Mean of Target group by Status.

```
> tbba1<-aggregate(Life.expectancy~Status, data=df1, FUN=mean)
> tbba1
      Status Life.expectancy
1  Developed        79.19785
2 Developing        67.11147
>
```

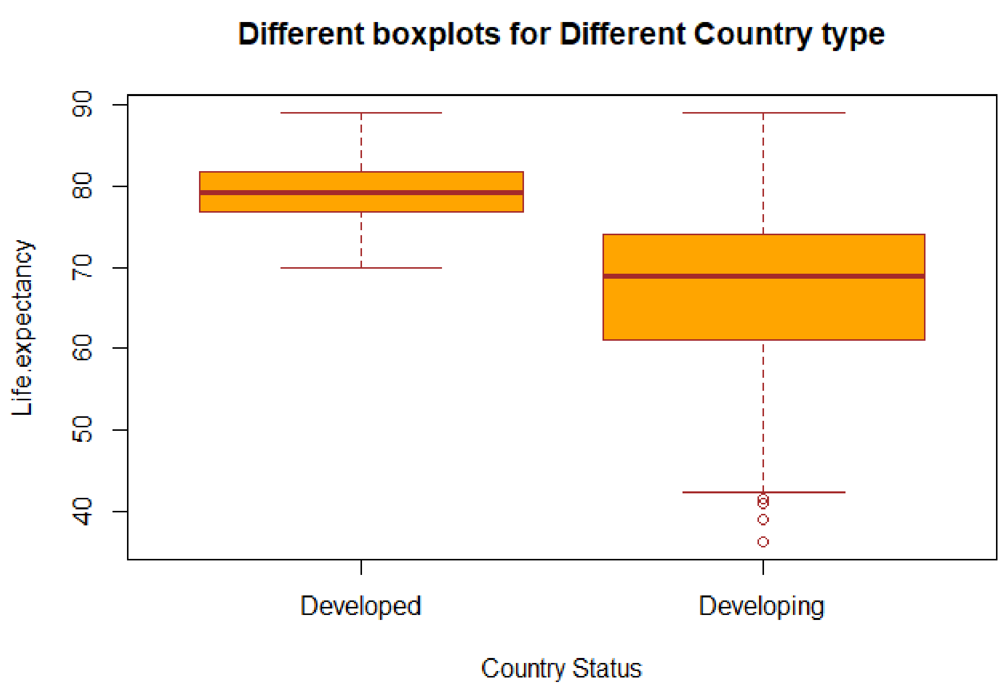
```
tbba2<-aggregate(Adult.Mortality~Status, data=df1, FUN=mean)
tbba2
    Status Adult.Mortality
 Developed        79.68555
Developing       182.83320
```

Here The life Expectancy in Developed countries is 79 years and Developing countries is 67 years.

Number of Adult deaths is less in Developed countries than Developing countries.

Visualization : Group Box Plot

```
# 2. VISUALIZING USING GROUP BOX PLOT

boxplot(Life.expectancy~Status,
        data=df1,
        main="Different boxplots for Different Country type",
        xlab="Country Status",
        ylab="Life.expectancy",
        col="orange",
        border="brown"
)
```



3. Problem: Is the mean of groups of ‘Status’ for 'life.Expectancy’ is Statistical Different
   Ho: The mean of 2 groups is equal

CHECKING whether mean of 2 groups is different USING T-TEST Since we have only 2 levels

```
# t-test
t.test(Life.expectancy~Status, data = df1, alternative = "greater")
```

The NULL Hypothesis is False. We Reject the Hypothesis.

Since p-value is < 0.05(5%) significance level . Mean of 2 groups is Statistically different from each Other.

```
> t.test(Life.expectancy~Status, data = df1, alternative = "greater")

	Welch Two Sample t-test

data:  Life.expectancy by Status
t = 47.868, df = 1807, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 11.67086      Inf
sample estimates:
 mean in group Developed mean in group Developing
                79.19785                 67.11147
```

BIVARIATE ANALYSIS for Life.Expectancy(Numerical) and

Year.groups (Categorical)

Q. 4 : Is there any relation between 'Year' and Target 'Life.expectancy'

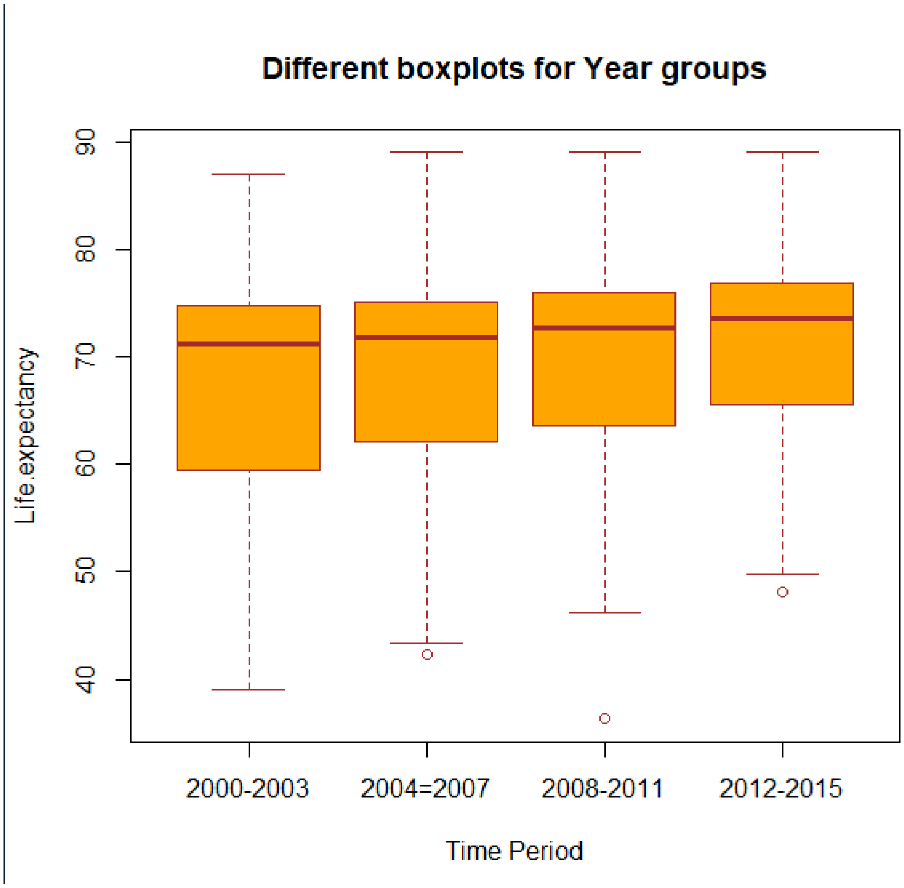
```
# Adding Another Categorical column to the Data S

df1$Year.groups<-NA
str(df1)

f2=function(x)  {
  if (x>=2000 && x<=2003) "2000-2003"
   else if (x>=2004 && x<=2007) "2004=2007"
    else if (x>=2008 && x<=2011) "2008-2011"
     else if (x>=2012 && x<=2015) "2012-2015"
}

df1$Year.groups<-sapply(df1$Year,f2)
str(df1)
df1
```

SUMMARIZING:

```
tbba4<-aggregate(Life.expectancy~Year.groups, data=df1, FUN=mean)
tbba4
Year.groups Life.expectancy
  2000-2003        67.16598
  2004=2007        68.38989
  2008-2011        70.01721
  2012-2015        71.32664
```

Visualizing :

```
boxplot(Life.expectancy~Year.groups,
        data=df1,
        main="Different boxplots for Year groups",
        xlab="Time Period",
        ylab="Life.expectancy",
        col="orange",
        border="brown"
)
```



To find the Relation between them we need to Run ANOVA test as the categorical variable Year.groups has more than 2 groups.

Ho : No difference between year.group means

```
> #ANOVA TEST
> one.way<-aov(Life.expectancy~Year.groups, data=df1)
> summary(one.way)
              Df Sum Sq Mean Sq F value Pr(>F)
Year.groups    3   7306  2435.5   27.58 <2e-16 ***
Residuals   2924 258184    88.3
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Pr(>F) in Anova test is the p-value.

Since p-value is < 0.05(5%) significance level . The NULL Hypothesis is False. We Reject the Hypothesis.

Mean of 4 groups is Statistically different from each Other.

BIVARIATE OR MULTIVARIATE ANALYSIS (CONTINOUS VS. CONTINOUS)

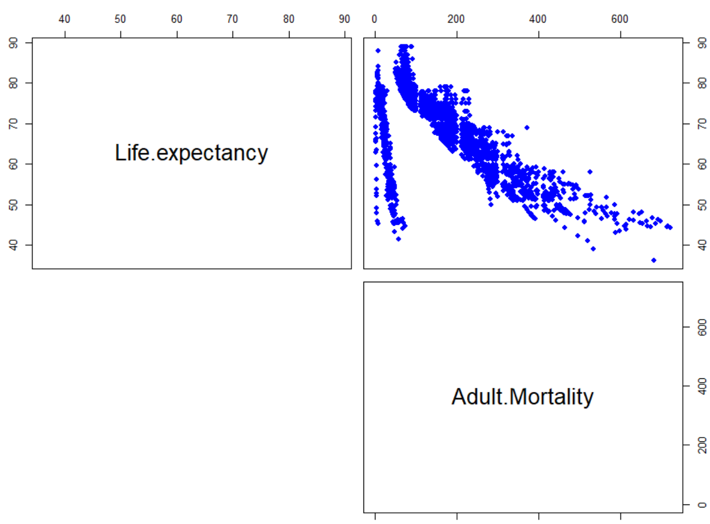
4.Does Life Expectancy have positive or negative relationship with Adult mortality?

```
# Problem :5 what kind of relation does Life Expectancy have with Adult Mortality
# SUMMARIZING

mydata1<-aggregate(Life.expectancy~Adult.Mortality, data=df1, FUN=mean)
mydata1
mydata1.cor<- cor(mydata1)
mydata1.cor

# The co-relation is -0.83, That means the above 2a variables are negatively corelated.
```

```
> mydata1.cor<- cor(mydata1)
> mydata1.cor
                Adult.Mortality Life.expectancy
Adult.Mortality       1.0000000      -0.8336785
Life.expectancy      -0.8336785       1.0000000
>
```

Since correlation is negative -0.83 , There is a strong negative linear relationship between Adult Mortality and Life.Expectancy.

```
#VISUALIZATION
pairs(df1[,c(4,5)], pch = 19,col="blue", lower.panel = NULL)
```

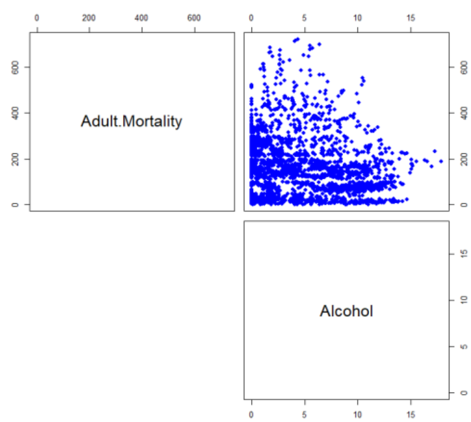


Q : 5. What kind of relation exist between Adult Mortality and Drinking Alcohol

```r
#Problem:5 What kind of relation exist between Adult Mortality and Drinking Alcohol
mydata2<-aggregate(Adult.Mortality~Alcohol, data=df1, FUN=mean)
mydata2
mydata2.cor<- cor(mydata2)
mydata2.cor
#VISUALIZATION
pairs(df1[,c(5,7)], pch = 19,col="blue", lower.panel = NULL)
```

The co-relation is -0.27. That means Adult Mortality and Drinking Alcohol are Weakly negatively co-related

```
> mydata2.cor<- cor(mydata2)
> mydata2.cor
                   Alcohol Adult.Mortality
Alcohol          1.0000000      -0.2737568
Adult.Mortality -0.2737568       1.0000000
> str(df1)
```

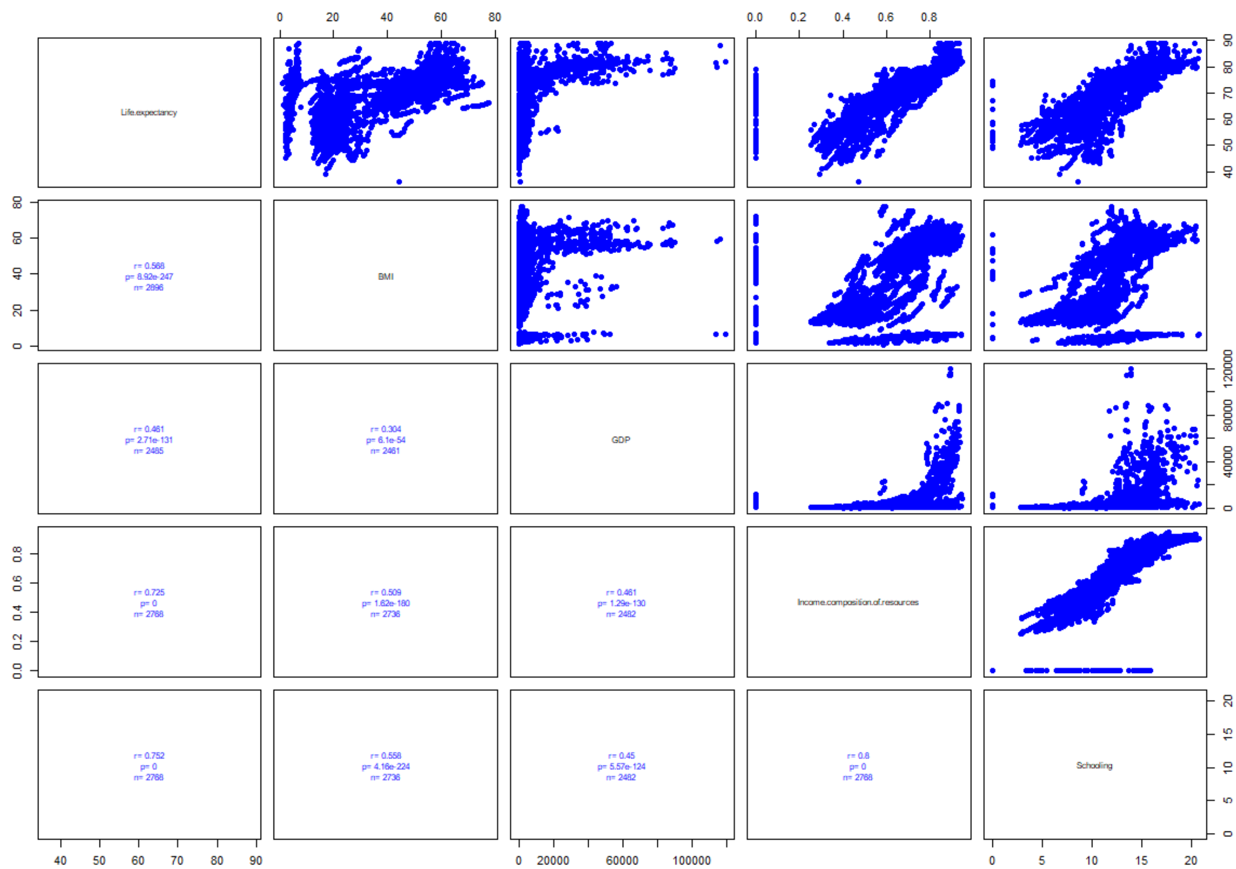


Q:6 Which predicting features are positively co-related.

After visualizing the scatter plots of Life Expectancy with all other Continuous

Variables using “pairs()” came to a conclusion that

```
pairs(df1[,c(4,11,17,19,20)], pch=19, col="blue",lower.panel = panel.cor1)
```



Life.Expectancy has positive co-relation with

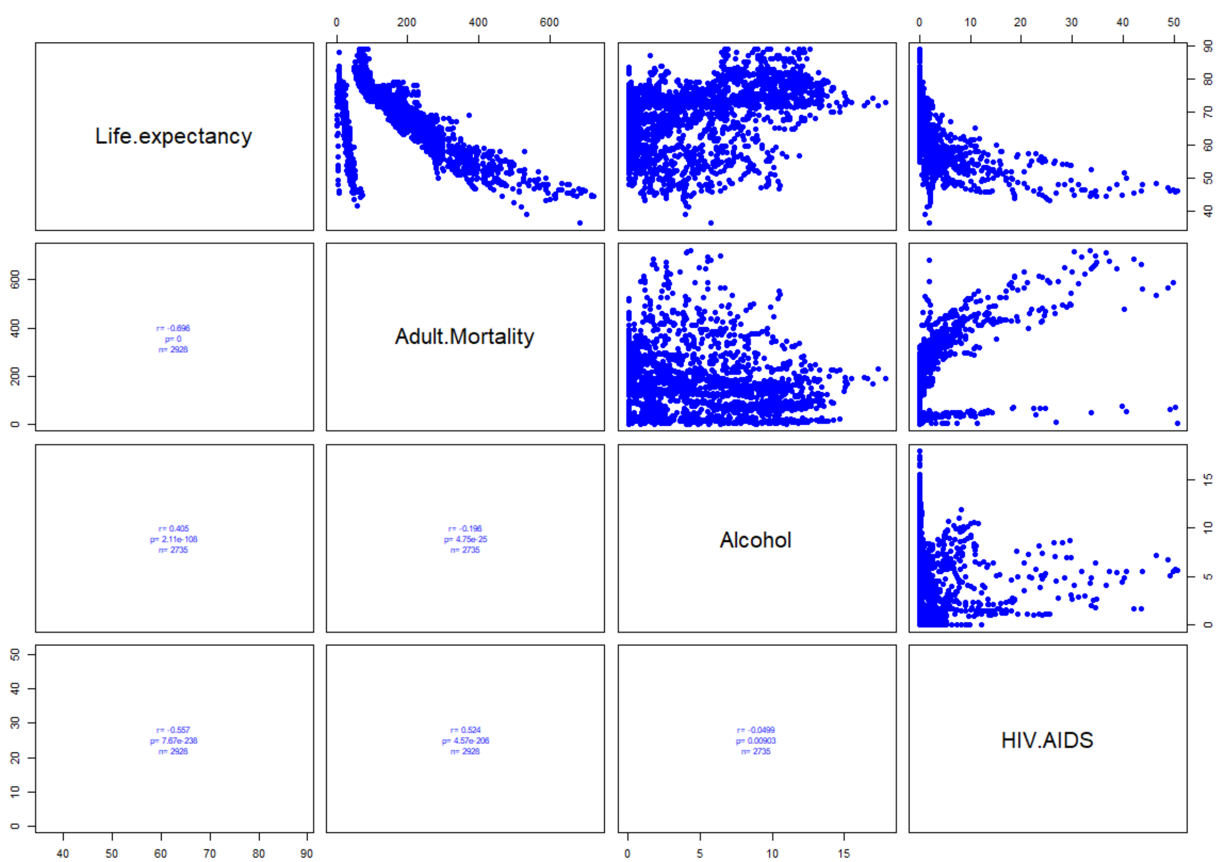
- GDP,
- Income.composition.of Resources
- Schooling
- BMI

```
panel.cor1 <- function(x, y, cex.cor = 0.8, method = "pearson", ...) {
  options(warn = -1)                   # Turn off warnings (e.g. tied
  usr <- par("usr"); on.exit(par(usr)) # Saves current "usr" and reset
  par(usr = c(0, 1, 0, 1))             # Set plot size to 1 x 1
  r <- cor(x, y, method = method, use = "pair")               # correl
  p <- cor.test(x, y, method = method)$p.val                  # p-valu
  n <- sum(complete.cases(x, y))                              # How ma
  txt <- format(r, digits = 3)                                # Format
  txt1 <- format(p, digits = 3)                                 # Form
  txt2 <- paste0("r= ", txt, '\n', "p= ", txt1, '\n', 'n= ', n) # Make
  text(0.5, 0.5, txt2, cex = cex.cor, ...)                      # Plac
  options(warn = 0)                                             # Rese
```

Q:7 Which features are negatively co-related.

Life.Expectancy shows strong negatively linear co-relation with Adult.Mortality and HIV.AIDS .

```
pairs(df1[,c(4,5,7,16)], pch=19, col="blue", lower.panel = panel.cor1)
```



8. What is the relation between Adult.Mortality and HIV.AIDS

   They have a moderate positive linear co-relation.

```r
# 5. What is the relation between Adult.Mortality and HIV.AIDS.
mydata3<-aggregate(Adult.Mortality~HIV.AIDS, data=df1, FUN=mean)
mydata3
mydata3.cor<- cor(mydata3)
mydata3.cor
#VISUALIZATION
pairs(df1[,c(5,16)], pch = 19,col="blue", lower.panel = NULL)
```

```
> mydata3.cor<- cor(mydata3)
> mydata3.cor
                 HIV.AIDS Adult.Mortality
HIV.AIDS        1.0000000       0.4421557
Adult.Mortality 0.4421557       1.0000000
>
```

BIVARIATE ANALYSIS :

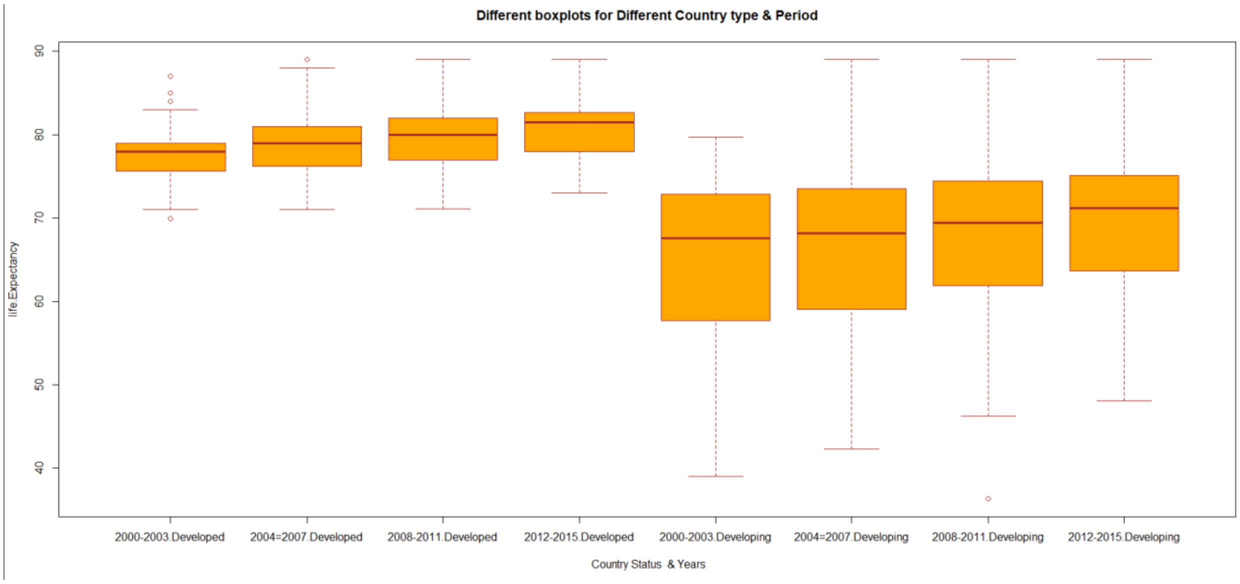
Q : 9 . What do you interpret regarding the Target Life Expectancy as the Years Pass by.

```r
# Bivariate Analysis of Target(numerical) and Year.groups & (categorical)
# 1.SUMMARIZING
aggregate(Life.expectancy~Status+Year.groups, data=df1, FUN=mean)  # grou
 str(df1)
 # 2. VISUALIZING
 boxplot(Life.expectancy~Year.groups+Status,
         data=df1,
         main="Different boxplots for Different Country type & Age group"
         xlab=" Country Status  & Years",
         ylab="life.Expectancy",
         col="orange",
         border="brown"
 )
```

Answer: The Range of Life.Expectancy for Developed countries is less than the Developing Countries. It is in increasing order from 2000 to 2015.

But, when we compare the Developed and Developing Countries, Life Expectancy of Developed countries is Higher than Developing Countries.

INTERPRETATION :



Q.10 : What do you conclude from the above EDA .

Conclusion:

By focusing on factors that are contributing for positive co-relation on Life-Expectancy.

Developing Countries can focus on them and try to improve those factors in order to improve their Life-Expectancy.

And by focusing on factors that are negatively co-related . Countries can try and take measures to mitigate their effect