Course: STQD 6444

Topic: Bad lifestyle in Life Expectancy using Statistical Analysis

Name: Syed Norman Daniel Bin Syed Mahadir(P125754)

1.0 Data background and source

This dataset is collected by the Global Health Observatory (GHO) data repository, which is under the World Health Organization (WHO) that keeps track of health status for all countries. This dataset consists of 193 countries, 22 Columns and 2938 rows which means 20 predicting variables. However, the data is extracted from Kaggle.com

The variables are namely as below table:

Variable name	**Explanation**
Country	Name of a country
Year	Year
Status	Developing or developed country
Life expectancy	Life expectancy in age
Adult morality	Adult morality rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant deaths	Number of Infant Deaths per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)

Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Measles - number of reported cases per 1000 population
BMI	Average Body Mass Index of entire population
Under Five deaths	Number of under-five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/ AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
Thinness 1-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)

Thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9(%)
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of Schooling(years)

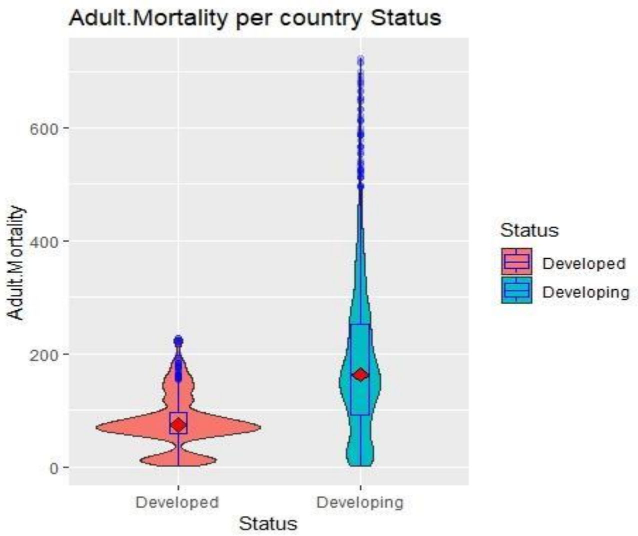
In this project, there are one objectives that will be discussed:

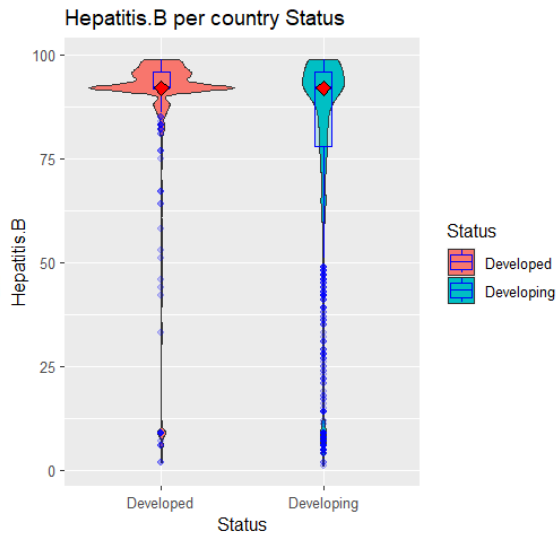
1. To investigate the relationship between bad Health on life expectancy for different status countries.

2.0 Description analysis of data

2.1 Objective : To investigate the relationship between bad Health on life expectancy for different status countries.

Adult Mortality & Hepatitis

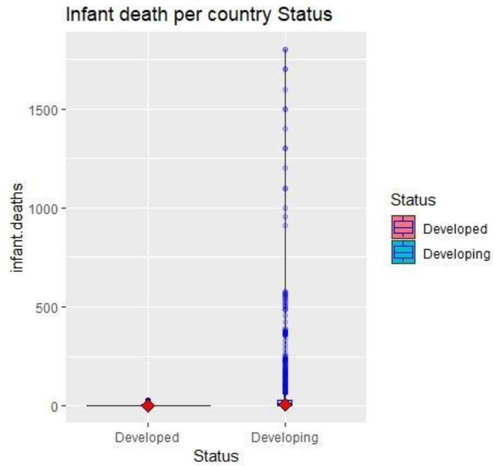


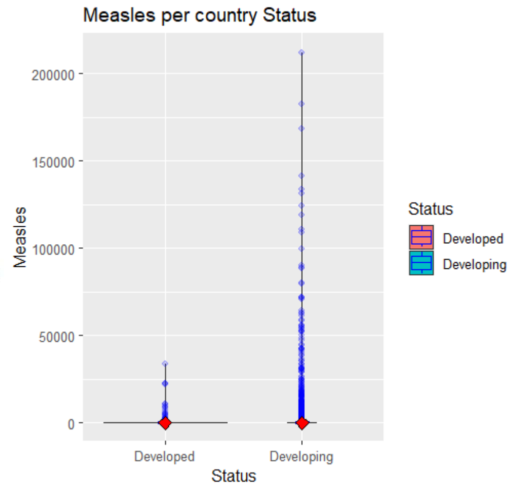


For the violin plot from Adult Mortality, we can see that for developed countries wider spots on 50 to 100 numbers of death adult mortality and for developing to be seen not quite large wider spots compare developed countries in the range 100 to 200 death. The wider spots indicate the density of data points at a given value. The frequency or concentration of data points at that value is indicated by the breadth of the spot. The larger the number of data points, the broader the spot. The violin plot is a compact depiction of a variable's distribution that aids in the visualization of skewness and kurtosis in data.

For the violin plot from Hepatitis B per country status, we can see that both of them contain many outliers.For the wider spots, we can see that countries developed in range 87 to 100 have better wide spots than countries developing which have smaller wide spots on 87 to 100.We can conclude that, for country developed the wider spots state that the immunization are commonly take as to prevent the hepatitis B disease and as we can see it can affect to reduce Adult Mortality and extend Life expectancy.

Infant death & Measles



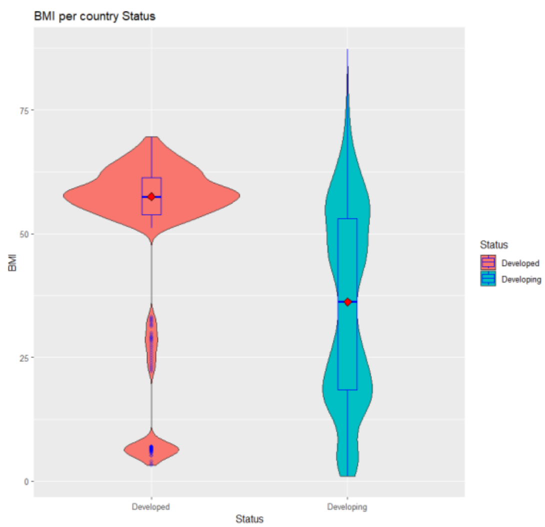


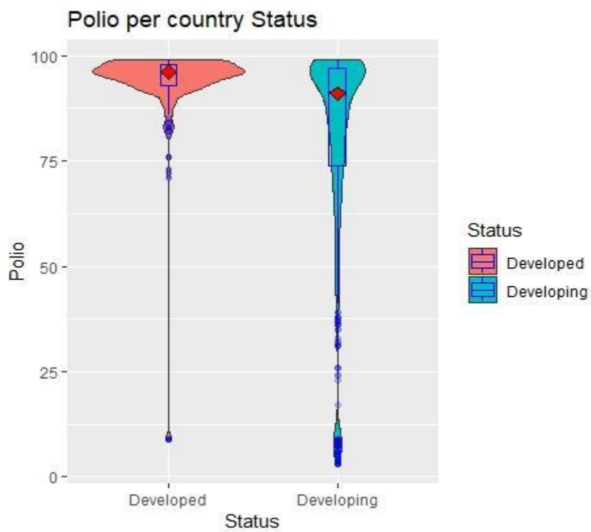
For the violin plot from Infant Death & Measles,we cannot see the boxplot and violin plot because maybe this data is very flat or outliers have too big of a value and cannot see the plot. That's why we cannot see the distribution of this graph even though Infant death for Developed and Developing has inter quartiles and median. Below are shows the infant death and measles for both status,developed and developing countries:

	infant.deaths	infant.deaths	Measles	Measles
Min.	: 0.000	: 0.00	: 0.0	: 0.0
1st Qu.	: 0.000	: 1.00	: 0.0	: 0.0
Median	: 0.000	: 6.00	: 12.0	: 18.0
Mean	: 1.494	: 36.38	: 499.0	: 2824.9
3rd Qu.	: 1.000	: 28.00	: 96.5	: 514.5
Max.	:28.000	:1800.00	:33812.0	:212183.0

As we cannot see the the both plot because both of them have small median and quartiles,plus the higher outlier in are the main factor we cannot plot for this variable dataset.The reason why outliers that country Developing has so many outliers is because developing countries children has lack of access in healthcare (For example:immunization Hepatitis B,Diphtheria,etc) and most them are extremely vulnerable to dangerous diseases due to antibody kids system.

BMI & Polio

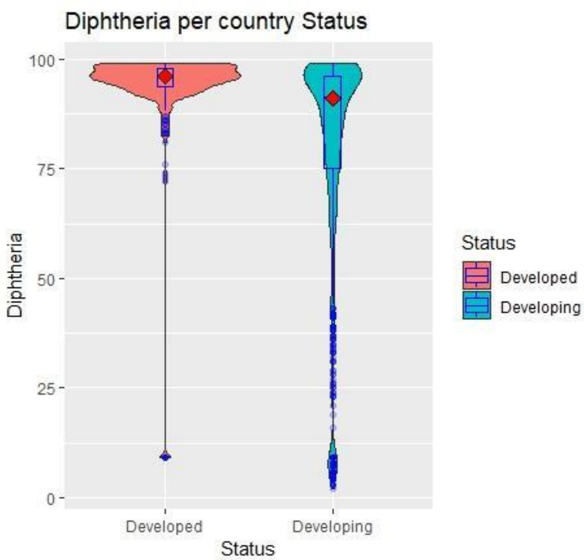


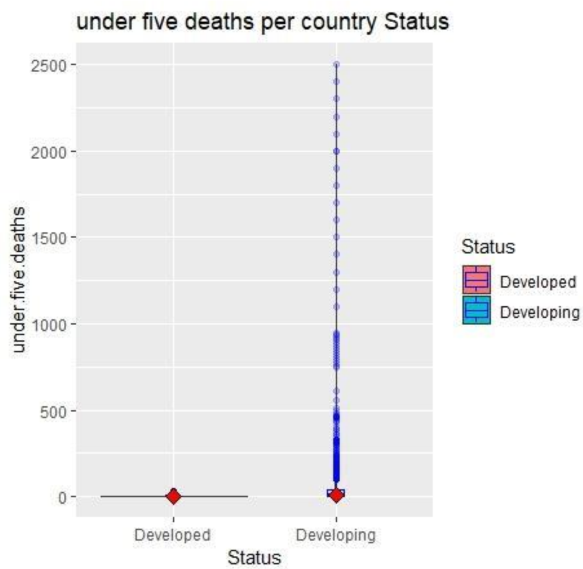


For the violin plot from BMI per country status, we can see that both of them contain many outliers.For the wider spots, we can see that countries developed in range 50 to 62.5, 15 to 35 and 3 to 12 have better wide spots than countries developing which have smaller wide spots on 12.5 to 25 and 37.5 to 62.5.This indicates that the country developed has three wider spots means that for the first wider spots is maybe underweight and ideal bmi, another spot on 15 to 35 is having an overweight bmi among people in country developed and last wider spot shows the obese weight which has a larger density in this violin graph which having higher bmi is not good for your health and can lead to heart disease and death .Next, for country developing we can see that the denser spots in 12.5 to 25 show thats those people developing countries having a good underweight and ideal bmi better widers spot than developed country means that in terms of bmi this developing countries has healthy lifestyle although has lack of healthcare and safety.

For the violin plot from Polio per country status, we can see that both of them contain many outliers.For the wider spots, we can see that countries developed in range 87.5 to 100 have better wide spots than countries developing which have smaller wide spots on 87.5 to 100.The wider spots indicates developed country has more take the polio vacination than developing countries due to good access healthcare and financial.

Diphtheria & under five deaths





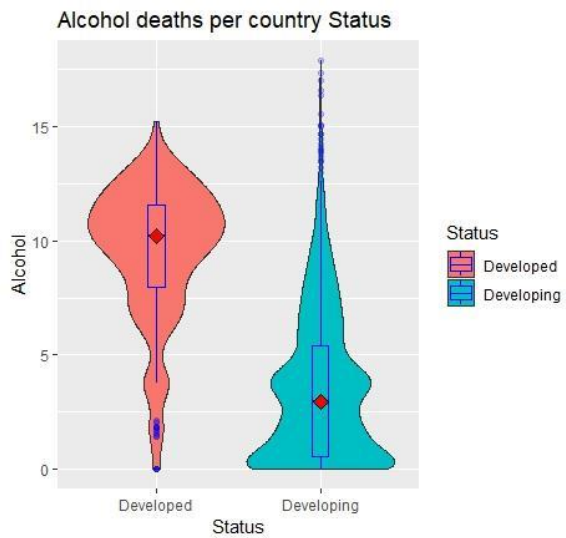
For the violin plot from Diphtheria per country status, we can see that both of them contain many outliers.For the wider spots, we can see that countries developed in range 87.5 to 100 have better wide spots than countries developing which have smaller wide spots on 87.5 to 100.The wider spots indicates developed country has more take the polio vacination than developing countries due to good access healthcare and financial.The outliers from this plot we can says that due to no awareness of the benefits of taking this vaccine and lack of good financial are the main reason why there is has outliers.

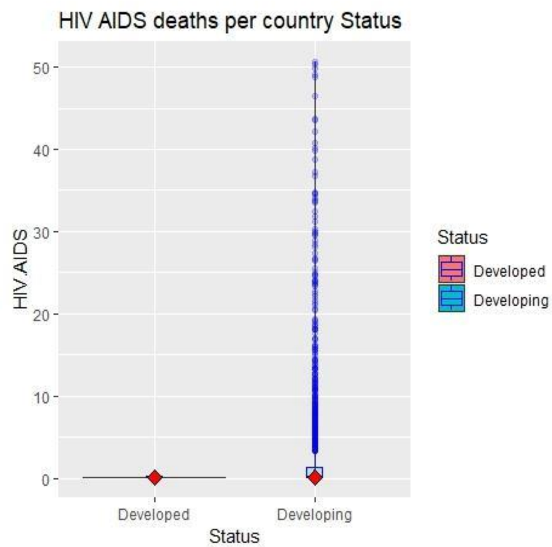
For the violin plot from under five deaths per country status,we cannot see boxplot or violin plot because the data may be very flat or the outliers have too high of values. Because of this, even if the distribution of this graph shows the median and interquartile ranges for infant mortality for developed and developing countries, we are unable to see it due to higher scalar on y-axis and extremely outlier. Below are shows the under five deaths for both status,developed and developing countries:

```
under.five.deaths  under.five.deaths
Min.   : 0.000     Min.   :   0.00
1st Qu.: 0.000     1st Qu.:   1.00
Median : 0.000     Median :   7.00
Mean   : 1.811     Mean   :  50.53
3rd Qu.: 2.000     3rd Qu.:  39.00
Max.   :33.000     Max.   :2500.00
```

We can see that clearly, there is not flat data and due to high values of outliers can effect the violin and boxplot. For country developed to be seen that having a good access in terms healthcare(taking vaccine) and financial are able to reduce the outlier in developed country.

Alcohol & HIV AIDS





For the violin plot from Alcohol per country status, we can see that both of them contain many outliers.For the wider spots, we can see that countries developed in range 5 to 10 have better wide spots than countries developing which have smaller wide spots on 0 to 5 per litre.Consumption of alcohol for countries developed has higher consumption alcohol than developing countries.

From the HIV/AIDS death per country status,We are unable to see the boxplot and violin plot for the HIV/AIDS data because the data may be relatively flat or because outliers have values that are too large to be visible on the plot. Because of this, even if the distribution of this graph shows the median and interquartile ranges for infant mortality for developed and developing countries, we are unable to see it.For country Developed, we can see from the dataset it only contains 0.1 and that's why there is no outlier in country Developed. Below are shows the under five deaths for both status,developed and developing countries:

```
   HIV.AIDS
 Min.   :0.1
 1st Qu.:0.1
 Median :0.1
 Mean   :0.1
 3rd Qu.:0.1
 Max.   :0.1
```

```
    HIV.AIDS
 Min.   : 0.100
 1st Qu.: 0.100
 Median : 0.100
 Mean   : 2.089
 3rd Qu.: 1.400
 Max.   :50.600
```

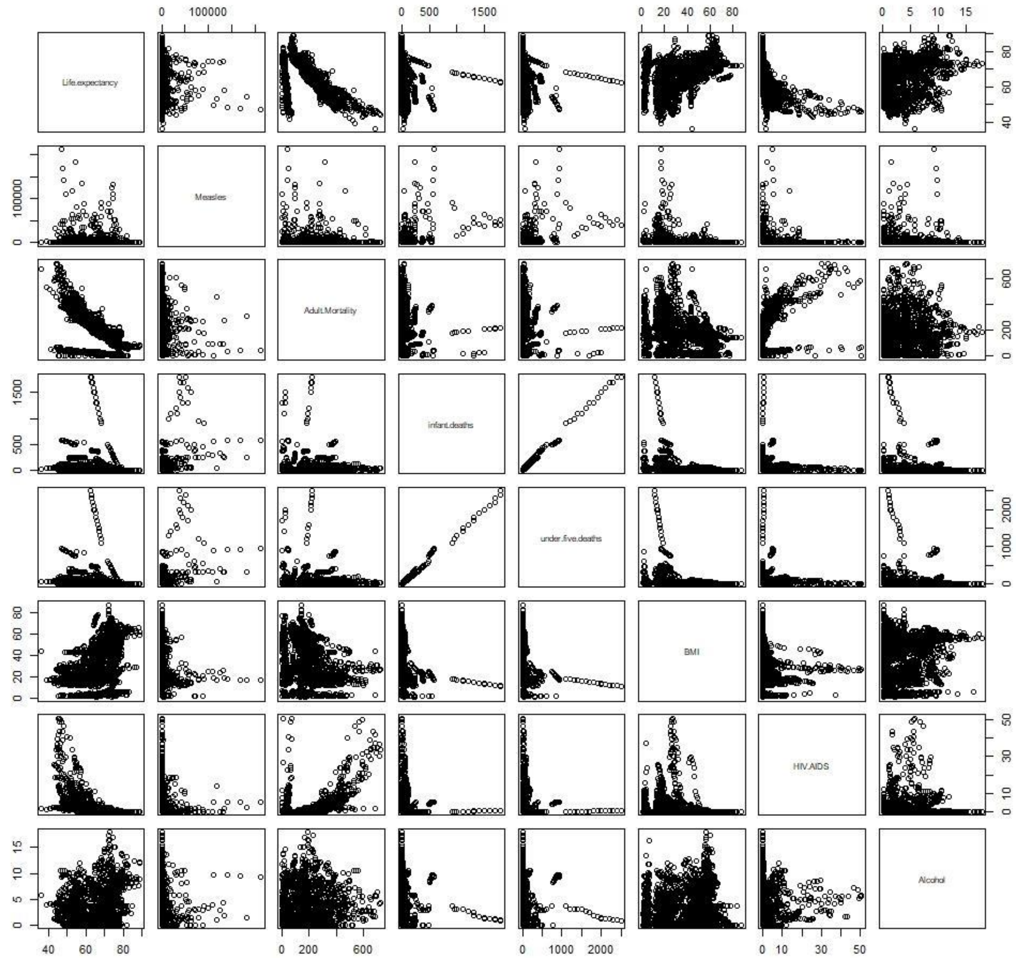
We can see that clearly, there is flat data on developed countries and due to high values of outliers in developing HIV.AIDS can affect the violin and boxplot. For country developed to be seen that having a good access in terms healthcare(taking vaccine) and financial are able to reduce the outlier in developed country.

3.0 Regression analysis

Objective : To investigate the relationship between bad Health on life expectancy for different status countries.

3.1 Graph plots

3.1.1 Developing countries



From the plot for life expectancy, we have a plot for a bad health lifestyle which can shorten someone's life expectancy. The bad health lifestyle for this data set is adult mortality, measles, infant death, under five years death ,BMI, HIV.AIDS and alcohol. Those dataset variables should have negative correlation to this graph.

For the first plot, we can see that life expectancy with measles shows that there is a slightly negative correlation that means that the life expectancy can decrease if the measles is increased.

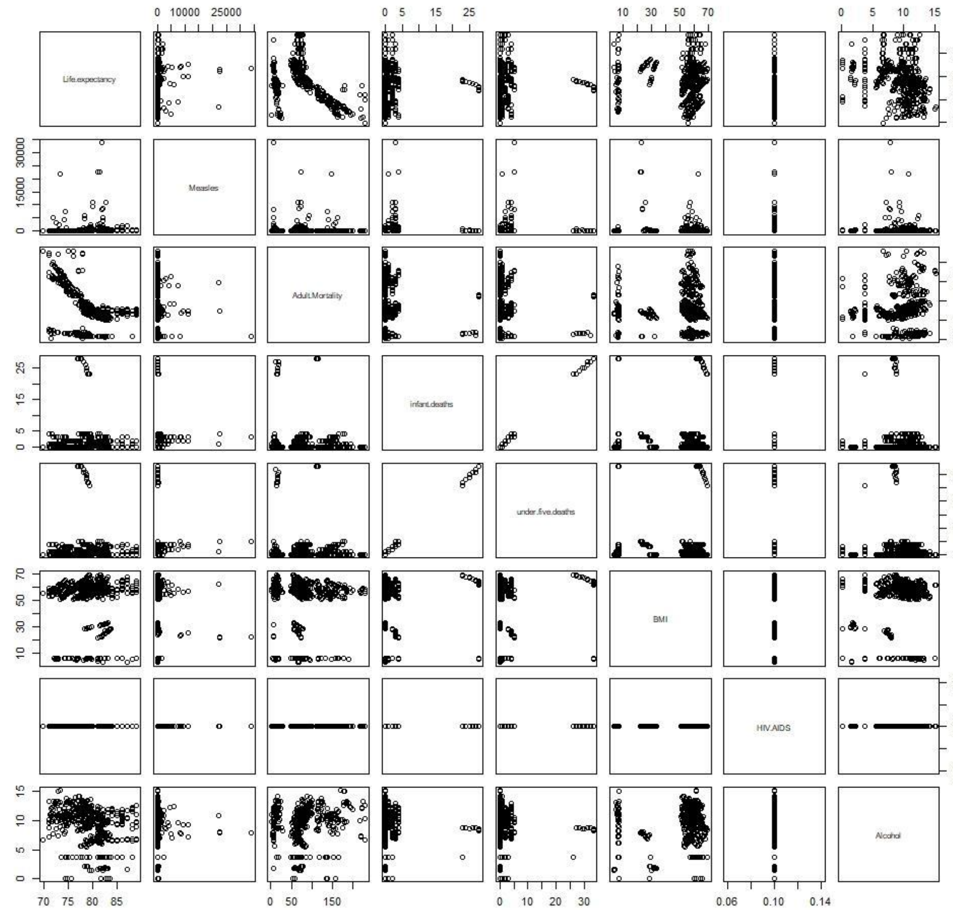
For the plot of life expectancy with adult mortality, we can see that the correlation between life expectancy and adult mortality is quite strong compared to another plot.This shows that adult mortality is the reason why life expectancy becomes shorter when there is more adult mortality and we can conclude that this relationship are the main factor life and high rate of adult mortality can decrease life expectancy.

The plot of life expectancy with infant death and with death under five shows that the correlation of this plot is quite similar with one to another. Both of them we can see that have similar negative correlation plots.

However, for the plot of life expectancy with BMI and with alcohol, we can see that both of them have not strong positive correlation and we can see clearly from the plot that the life.For the country developing, access to healthcare, good food, and clean water can be limited in poor nations, affecting overall health and life expectancy. Certain additional factors may alter the association between life expectancy and alcohol intake in these nations, making it difficult to establish a significant positive correlation.

Lastly, an inverse relationship between life expectancy and HIV/AIDS is frequently found. Life expectancy is reduced in locations where HIV/AIDS prevalence is high. This is owing to the disease's severe health consequences, which might result in death. The correlation coefficient between the two variables would almost certainly be close to -1, suggesting a significant inverse link. Other variables, such as access to healthcare and treatment, can also have an impact on the life expectancy of those living with HIV/AIDS.

3.1.2 Developed countries



From the plot Life expectancy with measles, we can see that there is no correlation value means that there is no relationship between both of them. The reason why there is no correlation is because the developed country may have developed a cure or good access healthcare that can prevent this disease from happening.

From the plot Life expectancy with adult mortality, we can see that this plot has quite a strong negative correlation, the same as developing countries but both have different life spans.For example, developed countries have better lifespan developing countries due to lifestyle, food,nutrients, environment and good healthcare.

From the plot Life expectancy with infant death, death under five and bmi, we can see that there is no correlation value means life expectancy has no relationship with infant death, death under five and bmi. We can conclude that this is due to good hospitality, the government cares about the children and the public health.

From the plot Life expectancy with HIV/AIDS shows that NA correlation means that there is a similar disease HIV/AIDS for every country and year in this developed country even has a different life expectancy.

Lastly, there is an inverse relationship between life expectancy with alcohol and has moderate correlation between it. Consuming alcohol can lead to death and It has been demonstrated that excessive alcohol use can cause a variety of health issues and raise the chance of premature mortality. Moderate alcohol use, on the other hand, has been linked to a modestly higher life expectancy. The reason why those people consume alcohol is because they want to enjoy their life and release their stress but consuming high alcohol can lead to their own death and also other people too.(Example:Car crash,Killing someone,etc).

3.4 Hypothesis Testing

Bad health
lifestyle=Measles+HIV.AIDS+Adult.Mortality+infant.deaths+under.five.deaths+BMI+Alcohol

H0= **there is no correlation between Life expectancy with Bad health lifestyle**

H1= **there is correlation between life expectancy with Bad health lifestyle**

Developing Country

Data	Correlation value	P-value	Hypothesis Test
Life expectancy With Adult Mortality	-0.6609854	p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With Infant deaths	-0.1669865	p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With under five deaths	-0.1958457	p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With Measles	-0.1421993	p-value = 1.977e-12(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With HIV.AIDS	-0.570906	p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With BMI	0.5429241	p-value = p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With Alcohol	0.1954332	p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis

Bad health Life on Expectancy and Adult Mortality
Life.Expectancy Adult.Mortality has value -0.6609854 which indicates that it has strong negative correlation between Life Expectancy and adult mortality in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and Infant deaths
Life.Expectancy and Infant deaths has value -0.1669865 which indicates that it has weak negative correlation between Life Expectancy and Infant deaths in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and under five deaths

Life.Expectancy and under five deaths has value -0.1958457 which indicates that it has weak negative correlation between Life Expectancy and under five deaths in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and HIV-AIDS

Life.Expectancy HIV-AIDS has value -0.570906 which indicates that it has weak negative correlation between Life Expectancy and HIV-AIDS in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis..

Bad health Life on Expectancy and Measles

Life.Expectancy and Measles has value -0.1421993 which indicates that it has weak negative correlation between Life Expectancy and Measles in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and BMI

Life.Expectancy and BMI has value 0.5429241 which indicates that it has positive negative correlation between Life Expectancy and BMI in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and Alcohol

Life.Expectancy and Alcohol has value 0.1954332 which indicates that it has positive correlation between Life Expectancy and Alcohol in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Developed Country

Data	Correlation Value	P-value	Hypothesis Test
Life expectancy With Adult Mortality	-0.4854888	p-value < 2.2e-16(<0.05)	have enough evidence to reject null hypothesis
Life expectancy With Infant deaths	-0.05476379	p-value = 0.2161(<0.05)	fail to reject null hypothesis
Life expectancy With under five deaths	-0.04795308	p-value = 0.2788(<0.05)	fail to reject null hypothesis
Life expectancy With Measles	0.03780051	p-value = 0.3934	fail to reject null hypothesis
Life expectancy With HIV.AIDS	NA	NA	-
Life expectancy With BMI	-0.04396246	p-value = 0.3208	fail to reject null hypothesis
Life expectancy With Alcohol	-0.2865784	p-value =3.901e-11	have enough evidence to reject null hypothesis

Bad health Life on Expectancy and Adult Mortality

Life.Expectancy Adult.Mortality has value -0.4854888 which indicates that it has strong negative correlation between Life Expectancy and adult mortality in developing countries.Therefore, since p-value < 2.2e-16 and p value < 0.05, we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and Infant deaths

Life.Expectancy and Infant deaths has value -0.05476379 which indicates that it has no correlation between Life Expectancy and Infant deaths in developing countries.Therefore, since p-value =0.2161 and p-value>(0.05) ,this shown that it fail to reject null hypothesis.

Bad health Life on Expectancy and under five deaths

Life.Expectancy and under five deaths has value -0.04795308 which indicates that it has no correlation even tho has negative value correlation between Life Expectancy and under five deaths in developing countries.Therefore, since p-value =0.2788 and p-value>(0.05) ,this shown that it fail to reject null hypothesis.

Bad health Life on Expectancy and Measles

Life.Expectancy and measles has value 0.03780051 which indicates that it has no correlation even tho has correlation positive value correlation Life Expectancy and Measles in developing countries.Therefore, since p-value =0.3934 and p-value>(0.05) ,this shown that it fail to reject null hypothesis.

Bad health Life on expectancy With Alcohol

Life.Expectancy and measles has value -0.2865784 which indicates that it has no correlation even tho has correlation positive value correlation Life Expectancy and Measles in developing countries.Therefore, since p-value =3.901e-11 and p-value<(0.05) ,this shown that it we have enough evidence to reject null hypothesis.

Bad health Life on Expectancy and BMI

Life.Expectancy and BMI has value -0.04396246 which indicates that it has no correlation even tho has correlation positive value correlation between Life Expectancy and BMI in developing countries.Therefore, since p-value =0.3208 and p-value>(0.05) ,this shown that it fail to reject null hypothesis.

Bad health Life on Expectancy and HIV AIDS

Life.Expectancy and Alcohol shown NA as a result the data means that for result Alcohol is similar to all countries that is to say 0.1 for these developed countries.

4.0 R code

```r
# ////////////////////////// DATA CLEANING ////////////////////
dat_dat<- read.csv(file.choose())
life_dat <- dat_dat
life<- dat_dat
head(life_dat)
str(life_dat)
library(tidyverse)
library(mice)

#to visualise the NA values using MICE
is.na(life_dat)
mice::md.pattern(life_dat)   # too messy, need to check each variable one by one

sum(is.na(life_dat$Life.expectancy))
sum(is.na(life_dat$Adult.Mortality))
sum(is.na(life_dat$infant.deaths))   # no NA
sum(is.na(life_dat$Alcohol))
sum(is.na(life_dat$percentage.expenditure))  # no NA
sum(is.na(life_dat$Hepatitis.B))
sum(is.na(life_dat$Measles))    # no NA
sum(is.na(life_dat$BMI))
sum(is.na(life_dat$under.five.deaths))    # no NA
sum(is.na(life_dat$Polio))
sum(is.na(life_dat$Total.expenditure))
sum(is.na(life_dat$Diphtheria))
sum(is.na(life_dat$HIV.AIDS))     # no NA
sum(is.na(life_dat$GDP))
sum(is.na(life_dat$Population))
sum(is.na(life_dat$thinness..1.19.years))
sum(is.na(life_dat$thinness.5.9.years))
sum(is.na(life_dat$Income.composition.of.resources))
sum(is.na(life_dat$Schooling))


# fill missing values of with median

imp_med2 <- life_dat %>% mutate(across(where(is.numeric), ~replace_na(., median(., na.rm=TRUE))))
str(imp_med2)
head(imp_med2,10)
sum(is.na(imp_med2))
sum(is.na(life_dat))  # to compare with original data for sum of NA values
str(life_dat)
mice::md.pattern(imp_med2)
```

```r
# ///////////// EDA ///////////////////
life_data<- imp_med2

# objective 1: to analyse the relationship between economic factors against life expectancy in developed and developing countries

# defining economic factors in the dataset
str(life_data)

## ---------- descriptive analysis using graph and histogram --------------------


# looking at overall countries

ggplot(life_data, aes(x=Life.expectancy)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Life.expectancy in Country developed")

summary(life_data$Life.expectancy)
# we can observe that the life expectancy for the whole population is skewed to the left
# with mean of the population are 69.23 years which are slightly lesser than the median of 72 years

hist(life_data$Life.expectancy)       #relatively normal, slightly skewed to left


hist(life_data$percentage.expenditure)      #skewed to right
hist(life_data$GDP)                         #skewed to right
hist(life_data$Total.expenditure)     # relatively normal, slightly skewed to right


# looking at developing countries

devping_life_data<- life_data%>%
  filter(life_data$Status=="Developing")%>%
  select(Country,Status,Life.expectancy, percentage.expenditure,GDP,Total.expenditure)

head(devping_life_data)

ggplot(devping_life_data, aes(x=Life.expectancy)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Life.expectancy in Country developed")


hist(devping_life_data$Life.expectancy)
summary(devping_life_data$Life.expectancy)
# we can observe that the life expectancy for the developing countries is skewed to the left
# with mean of the developing countries are 67.23 years which are slightly lesser than the median of 69.05 years
```

```r
ggplot(devping_life_data, aes(x=percentage.expenditure)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of percentage.expenditure in Country developed")

hist(devping_life_data$percentage.expenditure)   #skewed to right
summary(devping_life_data$percentage.expenditure)
# we can observe that the percentage expenditure for the developing countries is skewed to the right



ggplot(devping_life_data, aes(x=GDP)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of GDP in Country developed")

hist(devping_life_data$GDP)                       #skewed to right
summary(devping_life_data$GDP)
# we can observe that the GDP per capita for the developing counties is skewed to the right


ggplot(devping_life_data, aes(x=Total.expenditure)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Total.expenditure in Country developed")

hist(devping_life_data$Total.expenditure)
summary(devping_life_data$Total.expenditure)
# we can observe that the total government expenditure (%) on healthcare for the developing counties is skewed to the right


# looking at developed countries

devped_life_data<- life_data%>%
  filter(life_data$Status=="Developed")%>%
  select(Country,Status,Life.expectancy, percentage.expenditure,GDP,Total.expenditure)

head(devped_life_data)

ggplot(devped_life_data, aes(x=Life.expectancy)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Life.expectancy in Country developed")

hist(devped_life_data$Life.expectancy)
summary(devped_life_data$Life.expectancy)

ggplot(devped_life_data, aes(x=percentage.expenditure)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Life.expectancy in Country developed")

hist(devped_life_data$percentage.expenditure)   #skewed to right
summary(devped_life_data$percentage.expenditure)
```

```r
# //////////////////////////// MULTIPLE LINEAR REGRESSION ANALYSIS ////////////////////////////

# we believe that multiple economical variables will have an effect on life expectancy for both developing and developed countries
# therefore, we are performing multiple linear regression of the economical variables for both developing and developed countires


# multi linear regression plot for developing countries
par(mfrow=c(2,2))
fit <- lm(Life.expectancy~percentage.expenditure+GDP+Total.expenditure , data=devping_life_data)
#plot(fit)
summary (fit)


# multi linear regression plot for developed countries
fit2 <- lm(Life.expectancy~percentage.expenditure+GDP+Total.expenditure , data=devped_life_data)
#plot(fit2)
summary (fit2)
fit3<- lm(Life.expectancy~GDP , data=devped_life_data)
summary (fit3)

# -----------------------------------------------------------------------------------------------
# -----------------------------------------------------------------------------------------------
# /////////////////////////// Objective 2: bad health on life expectancy ///////////////////////////

str(Life)
Life_Developing<-filter(Life,Status=="Developing")
Life_Developed<-filter(Life,Status=="Developed")
sum(is.na(Life$Status))
sum(is.na(Life$Life.expectancy))
sum(is.na(Life$Alcohol))
sum(is.na(Life$Hepatitis.B))
sum(is.na(Life$Measles))    # no NA
sum(is.na(Life$BMI))
sum(is.na(Life$Polio))
sum(is.na(Life$Diphtheria))
sum(is.na(Life$HIV.AIDS))     # no NA
#replace NA with 0
Life_Developing<-Life_Developing%>%replace(is.na(.),0)
Life_Developed<-Life_Developed%>%replace(is.na(.),0)
#select column for first our main objective
Data_Developed<-subset(Life_Developed,select=c(Country,Year,Status,Life.expectancy,Hepatitis.B,Measles,Polio,Diphtheria,HIV.AIDS,
                                               Adult.Mortality,infant.deaths,under.five.deaths,BMI,Alcohol))
Data_Developing<-subset(Life_Developing,select=c(Country,Year,Status,Life.expectancy,Hepatitis.B,Measles,Polio,Diphtheria,HIV.AIDS,
                                                 Adult.Mortality,infant.deaths,under.five.deaths,BMI,Alcohol))
summary(Data_Developed)
summary(Data_Developing)
#
```

```r
#-------------------------------------------
#Graph distribution
##Developed
ggplot(Data_Developed, aes(x=Life.expectancy)) +
  geom_density(alpha=.3, fill="blue", color="blue", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Life.expectancy in Country developed")
skewness(Data_Developed$Life.expectancy)
###Developing
ggplot(Data_Developing, aes(x=Life.expectancy)) +
  geom_density(alpha=.3, fill="yellow", color="yellow", linewidth=1.5)+
  geom_vline(aes(xintercept=mean(Life.expectancy)), linewidth=1)+
  ggtitle("Distribution density of Life.expectancy in Country developing")
skewness(Data_Developing$Life.expectancy)
#violinplot with boxplot
set_plot_dimensions(20,10)
par(mfrow=c(2,7))
ggplot(Life ,aes(x= Status,y=Life.expectancy, fill= Status)) +
  geom_violin() +  geom_boxplot(width=0.1, color="blue", alpha=0.2) + stat_summary(fun.y = median, geom = "point",
                                                                                    shape = 23, size = 3, fill = "red")+
  ggtitle("Life expectancy per country Status")
ggplot(Life ,aes(x= Status,y=Adult.Mortality, fill= Status)) +
  geom_violin() +  geom_boxplot(width=0.1, color="blue", alpha=0.2) + stat_summary(fun.y = median, geom = "point",
                                                                                    shape = 23, size = 3, fill = "red")
gplot(Life ,aes(x= Status,y=Adult.Mortality, fill= Status)) +
  geom_violin() +  geom_boxplot(width=0.1, color="blue", alpha=0.2) + stat_summary(fun.y = median, geom = "point",
                                                                                    shape = 23, size = 3, fill = "red")
ggtitle("Adult.Mortality per country Status")
ggplot(Life ,aes(x= Status,y=Hepatitis.B, fill= Status)) +
  geom_violin() +   geom_boxplot(width=0.1, color="blue", alpha=0.2) +
  ggtitle("Hepatitis.B per country Status") + stat_summary(fun.y = median, geom = "point", shape = 23, size = 3, fill = "red")
ggplot(Life ,aes(x= Status,y=Measles, fill= Status)) +
  geom_violin() +  geom_boxplot(width=0.1, color="blue", alpha=0.2) +
  ggtitle("Measles per country Status") +stat_summary(fun.y = median, geom = "point", shape = 23, size = 3, fill = "red")
ggplot(Life ,aes(x= Status,y=BMI, fill= Status)) +
  geom_violin() +  geom_boxplot(width=0.1, color="blue", alpha=0.2) +stat_summary(fun.y = median, geom = "point",
                                                                                   shape = 23, size = 3, fill = "red")+
  ggtitle("BMI per country Status")
ggplot(Life ,aes(x= Status,y=Polio, fill= Status)) +
  geom_violin() +  geom_boxplot(width=0.1, color="blue", alpha=0.2) + stat_summary(fun.y = median, geom = "point",
                                                                                    shape = 23, size = 3, fill = "red")+
  ggtitle("Polio per country Status")
ggplot(Life ,aes(x= Status,y=BMI, fill= Status)) +
  geom_violin() + geom_boxplot(width=0.1, color="blue", alpha=0.2) + +stat_summary(fun.y = median, geom = "point",
                                                                                    shape = 23, size = 3, fill = "red")+
  ggtitle("BMI per country Status")
ggplot(Life ,aes(x= Status,y=Diphtheria, fill= Status)) +
  geom_violin() +geom_boxplot(width=0.1, color="blue", alpha=0.2) + stat_summary(fun.y = median, geom = "point",
                                                                                  shape = 23, size = 3, fill = "red")+
  ggtitle("Diphtheria per country Status")
```

```r
-----------------------------------------------------------------
  #cor test
  ##DevelopedCountry
  cor.test(Data_Developed$Life.expectancy,Data_Developed$Adult.Mortality)
##DevelopingCountry
cor.test(Data_Developing$Life.expectancy,Data_Developing$Adult.Mortality)

#correlation and collinearity
##Developing
cor(Data_Developing$Measles)
#--------------------------------------------------
#Multivariables
##Developing
par(mfrow=c(2,2))
A<-lm(Life.expectancy~Measles+HIV.AIDS+Adult.Mortality+infant.deaths+under.five.deaths+BMI+Alcohol,data=Data_Developing )
plot(A)
summary(A)
##Developed
B<-lm(Life.expectancy~Measles+HIV.AIDS+Adult.Mortality+infant.deaths+under.five.deaths+BMI+Alcohol,data=Data_Developed)
plot(B)
summary(B)
```

5.0 Results and findings on multi linear regression analysis

It is infer that multiple economic variables will have an effect on life expectancy for both developing and developed countries based on the hypothesis testing, multiple regression analysis is used to further assess the strength of the relationship between the outcome (life expectancy) and several predictor variables, and also to assess the importance (association) of each of the predictors variables to the relationship

Objective : To investigate the relationship between bad Health on life expectancy for different status countries.

Developing Countries

```
> A<-lm(Life.expectancy~Measles+HIV.AIDS+Adult.Mortality+infant.deaths
+under.five.deaths+BMI+Alcohol,data=Data_Developing)
> summary(A)

Call:
lm(formula = Life.expectancy ~ Measles + HIV.AIDS + Adult.Mortality +
    infant.deaths + under.five.deaths + BMI + Alcohol, data = Data_Dev
eloping)

Residuals:
     Min       1Q   Median       3Q      Max
-26.5012  -2.8614   0.3332   3.0751  15.8521

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.717e+01  3.328e-01 201.870  < 2e-16 ***
Measles           -2.514e-05  9.601e-06  -2.618  0.00889 **
HIV.AIDS          -4.780e-01  2.194e-02 -21.785  < 2e-16 ***
Adult.Mortality   -2.606e-02  9.856e-04 -26.437  < 2e-16 ***
infant.deaths      1.749e-01  9.969e-03  17.543  < 2e-16 ***
under.five.deaths -1.321e-01  7.374e-03 -17.912  < 2e-16 ***
BMI                1.261e-01  5.980e-03  21.087  < 2e-16 ***
Alcohol            4.608e-01  3.278e-02  14.055  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.075 on 2418 degrees of freedom
Multiple R-squared:  0.6824,    Adjusted R-squared:  0.6815
F-statistic: 742.2 on 7 and 2418 DF,  p-value: < 2.2e-16
```

In developing countries ,the first step to analyzing multiple regression analysis is to look at the F-statistic and related p-value, which can be found at the bottom of the model summary.The p-value of the F-statistic in our example is 2.2e-16, which is extremely significant. This signifies that at least one of the predictor factors is connected to the outcome variable in a substantial way . As we can see in p-value, almost all have great significant value and only measles have P-value less than 0.05 implies that the connection between the predictor and the dependent variable(life expectancy) is statistically significant and is not likely to have happened by coincidence.Besides that, adjusted R-squared for developing countries has better value fit than developed country.

Life.Expectancy=(6.643e+01)+(-2.524e-05)Measles+(-5.231e-01)HIV.AIDS+(-2.144e-02)Adult.Mortality+ (1.810e-01)Infant.deaths+(-1.364e01)under.five.deaths+(1.188e-01)BMI+(5.001e-01)Alcohol

Developed Countries

```
> B<-lm(Life.expectancy~Measles+Adult.Mortality+infant.deaths+under.five.deaths+BMI
+HIV.AIDS+Alcohol,data=Data_Developed)
> summary(B)

Call:
lm(formula = Life.expectancy ~ Measles + Adult.Mortality + infant.deaths +
    under.five.deaths + BMI + HIV.AIDS + Alcohol, data = Data_Developed)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4677  -1.8664  -0.0215   1.4969  10.4935

Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.533e+01  6.617e-01 128.952  < 2e-16 ***
Measles           -5.511e-06  5.920e-05  -0.093   0.9259
Adult.Mortality   -3.775e-02  3.074e-03 -12.280  < 2e-16 ***
infant.deaths     -1.344e+00  4.908e-01  -2.738   0.0064 **
under.five.deaths  1.081e+00  4.183e-01   2.584   0.0100 *
BMI               -3.694e-03  8.585e-03  -0.430   0.6672
HIV.AIDS                  NA         NA      NA       NA
Alcohol           -3.030e-01  4.897e-02  -6.187 1.27e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.295 on 505 degrees of freedom
Multiple R-squared:  0.3054,	Adjusted R-squared:  0.2972
F-statistic: 37.01 on 6 and 505 DF,  p-value: < 2.2e-16
```

In developed countries, the F-statistic and associated p-value, which are located at the bottom of the model summary, are the first things to be examined when evaluating multiple regression analysis. In our example, the F-p-value statistics is 2.2e-16, which is quite significant. This shows that at least one of the predictor factors is significantly linked to the outcome variable.

We found Measles,BMI, and HIV/AIDS is not significant in the multiple regression model. This means we need to remove Measles,BMI, and HIV/AIDS and it is straightforward from the model because it is not statistically significant:

```
> B1<-lm(Life.expectancy~Adult.Mortality+infant.deaths+under.five.deaths+Alcohol,da
ta=Data_Developed)
> summary(B1)

Call:
lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
    under.five.deaths + Alcohol, data = Data_Developed)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4833  -1.8595  -0.0434   1.4927  10.4630

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       85.151719   0.521669 163.229  < 2e-16 ***
Adult.Mortality   -0.037716   0.003067 -12.299  < 2e-16 ***
infant.deaths     -1.346912   0.480276  -2.804  0.00523 **
under.five.deaths  1.082933   0.409070   2.647  0.00837 **
Alcohol           -0.305084   0.048598  -6.278 7.39e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.29 on 507 degrees of freedom
Multiple R-squared:  0.3052,	Adjusted R-squared:  0.2997
F-statistic: 55.67 on 4 and 507 DF,  p-value: < 2.2e-16
```

Finally, we can see that Adult Mortality and Alcohol is quite significant than infant death and under five deaths.This is our model equation :

Life.Expectancy=(85.151719)+(-0.037716)Adult.Mortality+(-1.346912)Infant.deaths+(-2.144e-02)Under.five.deaths+(-0.305084)Alcohol

6.0 conclusion

As a conclusion from the bad health lifestyle, we can say that the main factors can that affect life expectancy are:

For country Developing:

1. HIV-AIDS
2. Adult Mortality
3. Infant Death
4. Under Five Death
5. Alcohol
6. BMI

For country Developed:

1. Adult Mortality
2. Alcohol