WIE2003 Team 15

Group Leader : AIN FARHANA BINTI AHMAD FAWAZZI

|  |  |
| --- | --- |
| **NAME** | **MATRIC NO.** |
| AIN FARHANA BINTI AHMAD FAWAZZI | 17082224/1 |
| HAFSAH HANNAH BINTI SASNI | 17094349/1 |
| LEE HONG YAN | 17207044/1 |
| SOO MUN CHONG | 17204674/1 |
| NG WEI HONG | 17203419/1 |

library(tidyverse)  
library(dplyr)  
library(ggplot2)  
library(reshape2)

# Part 1

# Introduction

**To study the average life expectancy across countries over time from 2000 to 2015. 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.**

**It is interesting that the trend of life expectancy varies across different countries around the globe. In this study, our objective is to find out what the underlying reasons are behind the trend.**

# Part 2

# Question

**What is the relationship of alcohol consumption, health expenditure, GDP, population and other factors with life expectancy across various countries from 2000 to 2015 ?**

# Part 3

# Objectives

1. To determine the correlation between total health expenditure, GDP, alcohol consumption, composition of income, population size with life expectancy.
2. To analyze the trend found in the factors affecting life expectancy
3. To determine the effect that each factor has on the life expectancy of citizens of each country
4. To examine which country is having the highest life expectancy from 2000-2015.

**At first, we import the library needed to process our data.**

library(dplyr)  
library(tidyverse)  
library(ggplot2)

**Then, we set our working directory and read the dataset from the excel file in the directory setwd(“C:\Users\Windows 10\Documents\RStudio\LifeExpectancyAnalysis\LifeExpectancy”)**

file <- read.csv("LifeExpectancyData.csv", na.strings = "NA")

# Part 4

# Data Cleaning

**We then select the columns that we need from the dataset.**

data <- file %>% select(Country, Year, Status, Life.expectancy, Alcohol, Total.expenditure, GDP, Population,Income.composition.of.resources)

**The dataset contains some empty values of Life expectancy. But most of the empty values are from a certain countries like Marshall Islands, Monaco, Nauru and Saint Kitts and Nevis. Those countries only have one record from one year, it means that we are not able to calculate median or mean of the country’s life expectancy to fill the empty value. We deleted the rows which contains empty life expectancy value.**

sum(is.na(data$Life.expectancy))

## [1] 10

data <- data %>% drop\_na(Life.expectancy)

**Total expenditure column contains some NA values. We then grouped the data by country, find the mean of total expenditure(General government expenditure on health as a percentage of total government expenditure (%)) and fill the rows which contain NA with the mean calculated for that particular country(if any). The reason we are doing this instead of replacing all NA values with mean of all the observations is that replacing each NA value with mean of that particular country is more of an accurate indicator or prediction.But some countries don’t have record at all for the total expenditure column, thus the mean we found was also NA for that country. We decided to delete those rows altogether.**

sum(is.na(data$Total.expenditure))

## [1] 226

data <- data %>% group\_by(Country) %>% mutate(Total.expenditure= replace(Total.expenditure, is.na(Total.expenditure), mean(Total.expenditure, na.rm=TRUE)))  
sum(is.na(data$Total.expenditure))

## [1] 32

index <- !is.na(data$Total.expenditure)  
data <- data[index,]  
head(data)

## # A tibble: 6 x 9  
## # Groups: Country [1]  
## Country Year Status Life.expectancy Alcohol Total.expenditu~ GDP Population  
## <fct> <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Afghan~ 2015 Devel~ 65 0.01 8.16 584. 33736494  
## 2 Afghan~ 2014 Devel~ 59.9 0.01 8.18 613. 327582  
## 3 Afghan~ 2013 Devel~ 59.9 0.01 8.13 632. 31731688  
## 4 Afghan~ 2012 Devel~ 59.5 0.01 8.52 670. 3696958  
## 5 Afghan~ 2011 Devel~ 59.2 0.01 7.87 63.5 2978599  
## 6 Afghan~ 2010 Devel~ 58.8 0.01 9.2 553. 2883167  
## # ... with 1 more variable: Income.composition.of.resources <dbl>

**For Alcohol Consumption per capita consumption in litre, most of the countries have only one or two missing values. So we calculated mean value of alcohol consumption of each countries from 2000-2015, and fill the missing value of alcohol for that country(if any). After that, we found that South Sudan and Palau don’t have records for alcohol consumption at all, so we deleted the rows of those countries.**

sum(is.na(data$Alcohol))

## [1] 191

data <- data %>% group\_by(Country) %>% mutate(Alcohol = replace(Alcohol, is.na(Alcohol), median(Alcohol, na.rm = TRUE)))  
sum(is.na(data$Alcohol))

## [1] 16

index1 <- which(is.na(data$Alcohol))  
data[index1,]

## # A tibble: 16 x 9  
## # Groups: Country [1]  
## Country Year Status Life.expectancy Alcohol Total.expenditu~ GDP  
## <fct> <int> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 South ~ 2015 Devel~ 57.3 NA 2.71 759.  
## 2 South ~ 2014 Devel~ 56.6 NA 2.74 1152.  
## 3 South ~ 2013 Devel~ 56.4 NA 2.62 1186.  
## 4 South ~ 2012 Devel~ 56 NA 2.77 958.  
## 5 South ~ 2011 Devel~ 55.4 NA 2.71 177.  
## 6 South ~ 2010 Devel~ 55 NA 2.71 1562.  
## 7 South ~ 2009 Devel~ 54.3 NA 2.71 1265.  
## 8 South ~ 2008 Devel~ 53.6 NA 2.71 1679.  
## 9 South ~ 2007 Devel~ 53.1 NA 2.71 NA   
## 10 South ~ 2006 Devel~ 52.5 NA 2.71 NA   
## 11 South ~ 2005 Devel~ 51.9 NA 2.71 NA   
## 12 South ~ 2004 Devel~ 51.4 NA 2.71 NA   
## 13 South ~ 2003 Devel~ 58 NA 2.71 NA   
## 14 South ~ 2002 Devel~ 52 NA 2.71 NA   
## 15 South ~ 2001 Devel~ 49.6 NA 2.71 NA   
## 16 South ~ 2000 Devel~ 48.9 NA 2.71 NA   
## # ... with 2 more variables: Population <dbl>,  
## # Income.composition.of.resources <dbl>

data <- data %>% drop\_na(Alcohol)

**For GDP, since most of the NA values for GDP come from the same countries(multiple NA GDP for one country), we deleted the rows which contain the NA values.**

data[which(is.na(data$GDP)),]

## # A tibble: 406 x 9  
## # Groups: Country [193]  
## Country Year Status Life.expectancy Alcohol Total.expenditu~ GDP  
## <fct> <int> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 Bahamas 2015 Devel~ 76.1 10.2 6.71 NA  
## 2 Bahamas 2014 Devel~ 75.4 9.45 7.74 NA  
## 3 Bahamas 2013 Devel~ 74.8 9.42 7.5 NA  
## 4 Bahamas 2012 Devel~ 74.9 9.5 7.43 NA  
## 5 Bahamas 2011 Devel~ 75 9.34 7.63 NA  
## 6 Bahamas 2010 Devel~ 75 9.19 7.44 NA  
## 7 Bahamas 2009 Devel~ 74.6 9.29 7.43 NA  
## 8 Bahamas 2008 Devel~ 74.5 10.2 7.3 NA  
## 9 Bahamas 2007 Devel~ 74.4 10.8 7.8 NA  
## 10 Bahamas 2006 Devel~ 74.2 11.1 6.93 NA  
## # ... with 396 more rows, and 2 more variables: Population <dbl>,  
## # Income.composition.of.resources <dbl>

data <- data %>% drop\_na(GDP)

**For Population, again it’s just that we have countries which contain the population record for all the year from 2000-2015, OR countries which don’t have records at all for population because it’s difficult to get population data from certain countries. Therefore, we deleted those rows.**

sum(is.na(data$Population))

## [1] 220

data[which(is.na(data$Population)),]

## # A tibble: 220 x 9  
## # Groups: Country [193]  
## Country Year Status Life.expectancy Alcohol Total.expenditu~ GDP  
## <fct> <int> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 Antigu~ 2015 Devel~ 76.4 7.84 4.79 13567.  
## 2 Antigu~ 2014 Devel~ 76.2 8.56 5.54 12888.  
## 3 Antigu~ 2013 Devel~ 76.1 8.58 5.33 12225.  
## 4 Antigu~ 2012 Devel~ 75.9 8.18 5.39 12565.  
## 5 Antigu~ 2011 Devel~ 75.7 7.84 5.65 11929.  
## 6 Antigu~ 2010 Devel~ 75.6 7.84 5.63 12127.  
## 7 Antigu~ 2009 Devel~ 75.4 7.82 4.86 1312.  
## 8 Antigu~ 2008 Devel~ 75.2 8.27 4.69 1473.  
## 9 Antigu~ 2007 Devel~ 75 8.64 4.27 14252.  
## 10 Antigu~ 2006 Devel~ 74.8 8.93 4.34 12724.  
## # ... with 210 more rows, and 2 more variables: Population <dbl>,  
## # Income.composition.of.resources <dbl>

data <- data %>% drop\_na(Population)

**We have only one NA value for income.composition of resources and that country only has one record of year 2013in our dataset, so we are not able to calculate median or mean of income for that country(From 2000-2015). We deleted that row.**

sum(is.na(data$Income.composition.of.resources))

## [1] 0

data[which(is.na(data$Income.composition.of.resources)),]

## # A tibble: 0 x 9  
## # Groups: Country [193]  
## # ... with 9 variables: Country <fct>, Year <int>, Status <fct>,  
## # Life.expectancy <dbl>, Alcohol <dbl>, Total.expenditure <dbl>, GDP <dbl>,  
## # Population <dbl>, Income.composition.of.resources <dbl>

data <- data %>% drop\_na(Income.composition.of.resources)

# Part 5

# Exploratory Graph

**We need to understand our data properties by looking at the structure of our data.**

head(data)

## # A tibble: 6 x 9  
## # Groups: Country [193]  
## Country Year Status Life.expectancy Alcohol Total.expenditu~ GDP Population  
## <fct> <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Afghan~ 2015 Devel~ 65 0.01 8.16 584. 33736494  
## 2 Afghan~ 2014 Devel~ 59.9 0.01 8.18 613. 327582  
## 3 Afghan~ 2013 Devel~ 59.9 0.01 8.13 632. 31731688  
## 4 Afghan~ 2012 Devel~ 59.5 0.01 8.52 670. 3696958  
## 5 Afghan~ 2011 Devel~ 59.2 0.01 7.87 63.5 2978599  
## 6 Afghan~ 2010 Devel~ 58.8 0.01 9.2 553. 2883167  
## # ... with 1 more variable: Income.composition.of.resources <dbl>

glimpse(data)

## Rows: 2,254  
## Columns: 9  
## Groups: Country [193]  
## $ Country <fct> Afghanistan, Afghanistan, Afghanist...  
## $ Year <int> 2015, 2014, 2013, 2012, 2011, 2010,...  
## $ Status <fct> Developing, Developing, Developing,...  
## $ Life.expectancy <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8,...  
## $ Alcohol <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01,...  
## $ Total.expenditure <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20,...  
## $ GDP <dbl> 584.25921, 612.69651, 631.74498, 66...  
## $ Population <dbl> 33736494, 327582, 31731688, 3696958...  
## $ Income.composition.of.resources <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, ...

**Summary of Life Expectancy**

summary(data$Life.expectancy)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 36.30 62.40 71.40 68.79 75.47 89.00

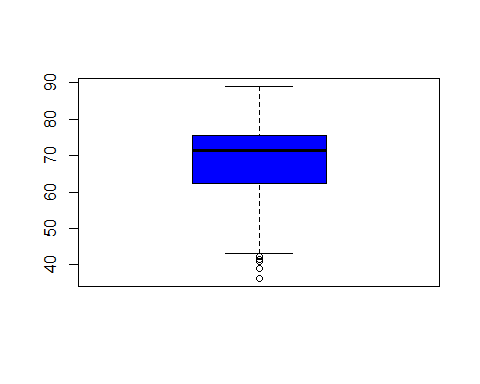
**Check the number of rows and columns the data frame has.**

dim(data)

## [1] 2254 9

**Boxplot of Life expectancy**

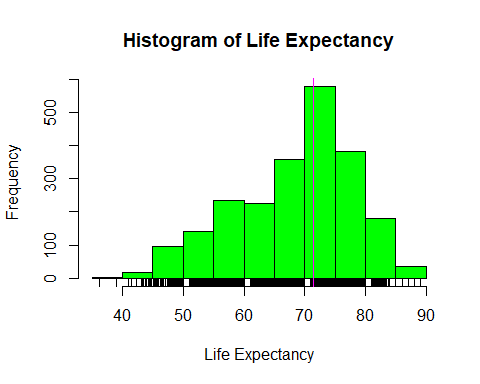
boxplot(data$Life.expectancy, col = "blue")



**The boxplot above shows that the median of life expectancy is about 70 years old. And the 25th percentile is about 60 years old, the 75th percentile is about 75 years old.**

**Histogram of Life Expectancy**

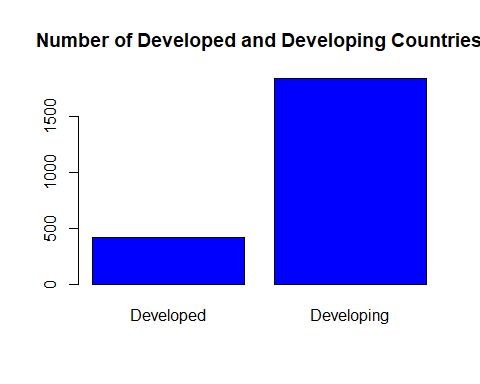
hist(data$Life.expectancy, main = "Histogram of Life Expectancy", xlab = "Life Expectancy", col = "green")  
rug(data$Life.expectancy)  
abline(v = median(data$Life.expectancy), col = "magenta")



**The histogram tells that the life expectancy is normally distributed with the median of around 72 years old. There are roughly 600 countries with life expectancy of 70-75 years old.**

**Barplot to see how development status varies**

barplot(table(data$Status), col = "blue", main = "Number of Developed and Developing Countries")



**The barplot above shows that the dataset we obtained contains unbalanced Country Status. Most the countries are developing countries. Developing countries are at least 3x more than developed countries.**

# Part 6

# Data Analysis

**Initially, we constructed a correlation matrix to gain understanding of how each of the factor affects the life expectancy. We only select the factors to build the correlation matrix.**

correlationFrame <- data  
correlationFrame <- ungroup(correlationFrame)  
correlationFrame <- correlationFrame %>% select(Life.expectancy, Alcohol, Total.expenditure, GDP, Population, Income.composition.of.resources)  
#cor(correlationFrame)  
cormat <- round(cor(correlationFrame),2)  
head(cormat)

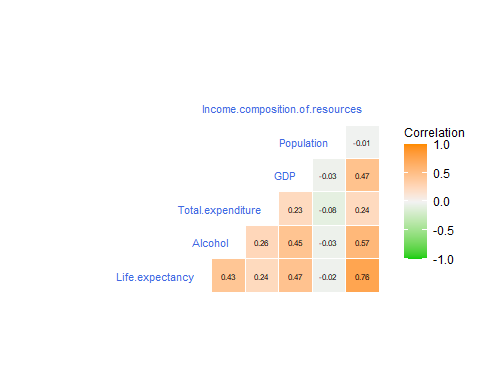
## Life.expectancy Alcohol Total.expenditure GDP  
## Life.expectancy 1.00 0.43 0.24 0.47  
## Alcohol 0.43 1.00 0.26 0.45  
## Total.expenditure 0.24 0.26 1.00 0.23  
## GDP 0.47 0.45 0.23 1.00  
## Population -0.02 -0.03 -0.08 -0.03  
## Income.composition.of.resources 0.76 0.57 0.24 0.47  
## Population Income.composition.of.resources  
## Life.expectancy -0.02 0.76  
## Alcohol -0.03 0.57  
## Total.expenditure -0.08 0.24  
## GDP -0.03 0.47  
## Population 1.00 -0.01  
## Income.composition.of.resources -0.01 1.00

library(reshape2)  
melted\_cormat <- melt(cormat)  
head(melted\_cormat)

## Var1 Var2 value  
## 1 Life.expectancy Life.expectancy 1.00  
## 2 Alcohol Life.expectancy 0.43  
## 3 Total.expenditure Life.expectancy 0.24  
## 4 GDP Life.expectancy 0.47  
## 5 Population Life.expectancy -0.02  
## 6 Income.composition.of.resources Life.expectancy 0.76

**We also constructed heatmap for correlation matrix to better visualize the effect of each factor on life expectancy.**

source("https://raw.githubusercontent.com/briatte/ggcorr/master/ggcorr.R")  
ggcorr(correlationFrame,   
 label = T,   
 label\_size = 2,  
 label\_round = 2,  
 hjust = 1,  
 size = 3,   
 color = "royalblue",  
 layout.exp = 5,  
 low = "green3",   
 mid = "gray95",   
 high = "darkorange",  
 name = "Correlation")



# Life expectancy over Alcohol consumption

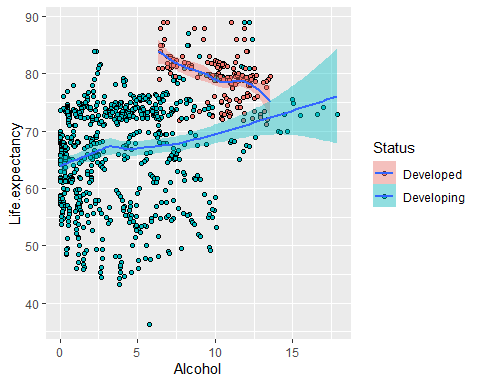
life\_year <- data %>%   
 filter (Year %in% c(2010, 2009, 2008, 2007, 2006, 2005))%>%  
 select (Country, Status, Year, Alcohol, Life.expectancy)   
  
head(life\_year)

## # A tibble: 6 x 5  
## # Groups: Country [1]  
## Country Status Year Alcohol Life.expectancy  
## <fct> <fct> <int> <dbl> <dbl>  
## 1 Afghanistan Developing 2010 0.01 58.8  
## 2 Afghanistan Developing 2009 0.01 58.6  
## 3 Afghanistan Developing 2008 0.03 58.1  
## 4 Afghanistan Developing 2007 0.02 57.5  
## 5 Afghanistan Developing 2006 0.03 57.3  
## 6 Afghanistan Developing 2005 0.02 57.3

**We use scatterplot to show the relation between Life Expectancy and Alcohol in year 2005-2010**

ggplot(life\_year, aes(Alcohol, Life.expectancy, fill = Status)) +geom\_point(shape = 21) + geom\_smooth(method = "loess") + scale\_color\_gradient(low = "yellow", high = "blue")

## `geom\_smooth()` using formula 'y ~ x'



**From this plot, we can see that there is a relationship between the two variables in the year 2005 to 2010. There’s a noticeable diferrence between developed and developing countries. For developed countries, the higher the alcohol consumption, the higher the life expectancy. The opposite holds true for developing countries.**

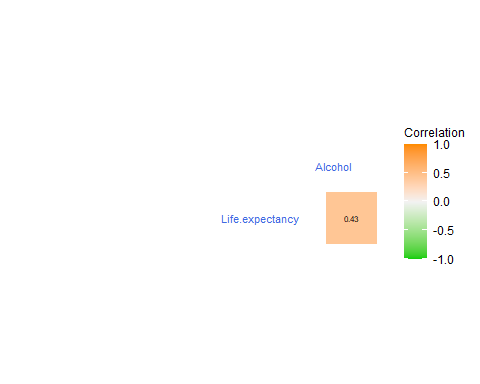
**Finding the correlation between Alcohol consumption with Life expectancy**

source("https://raw.githubusercontent.com/briatte/ggcorr/master/ggcorr.R")  
data\_num <- life\_year %>% select(Life.expectancy,Alcohol)

## Adding missing grouping variables: `Country`

ggcorr(data\_num,   
 label = T,   
 label\_size = 2,  
 label\_round = 2,  
 hjust = 1,  
 size = 3,   
 color = "royalblue",  
 layout.exp = 5,  
 low = "green3",   
 mid = "gray95",   
 high = "darkorange",  
 name = "Correlation")

## Warning in ggcorr(data\_num, label = T, label\_size = 2, label\_round = 2, : data  
## in column(s) 'Country' are not numeric and were ignored

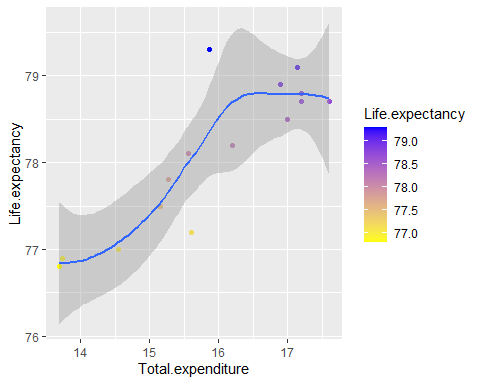


**Using ggcorr function, there is +0.4 correlation between alcohol consumption with life expectancy which indicates positive relationship.**

# Life expectancy over Total expenditure in United States of America, a developed country

dat2 <- file %>% filter (Country == "United States of America") %>% mutate(Total.expenditure = replace(Total.expenditure,is.na(Total.expenditure), mean(Total.expenditure, na.rm = TRUE)))  
ggplot(dat2, aes(x=Total.expenditure, y=Life.expectancy, color=Life.expectancy))+geom\_point() + scale\_color\_gradient(low = "yellow", high = "blue")+ geom\_smooth()

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

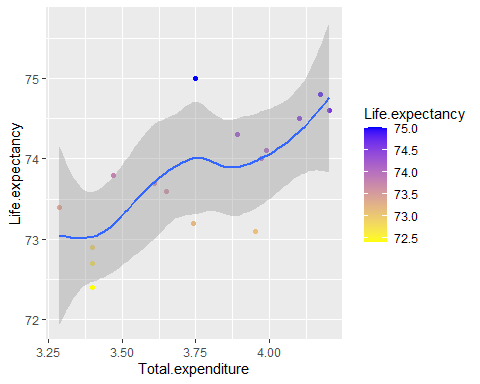


#df2 <- data %>% group\_by(Country) %>% mutate(meanLifeExp = mean(Life.expectancy), meanExpenditure = mean(percentage.expenditure)) %>% #distinct(Country, .keep\_all = TRUE) %>% ungroup() %>% select(Status, meanLifeExp, meanExpenditure)   
#graph2 <- ggplot(df2, aes(x = meanExpenditure,y = meanLifeExp, color = Status)) + geom\_point() + facet\_wrap(~Status) + ggtitle("Relationship #between Life Expectancy and Expenditure on health as a percentage of Gross Domestic Product per capita(%)") + xlab("Percentage of #Expenditure on health compared to GDP per capita") + ylab("Life Expectancy") + scale\_x\_log10()  
#graph2

**Life expenctancy over Total expenditure in Malaysia, a developing country.**

dat3 <- data %>%  
 filter (Country == "Malaysia", Year )  
ggplot(dat3, aes(x=Total.expenditure, y=Life.expectancy, color=Life.expectancy))+geom\_point() + scale\_color\_gradient(low = "yellow", high = "blue")+  
 geom\_smooth()

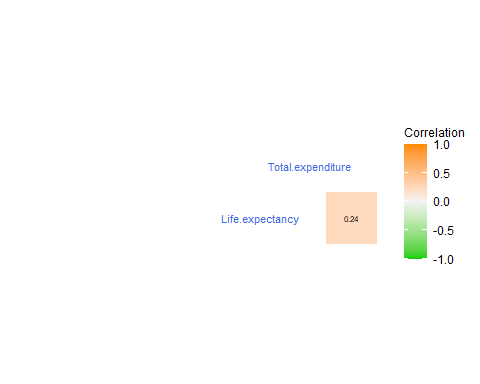
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



**Here I used scatter plot to see the relationship between life expectancy and total expenditure in 2 different countries(developed and developing), What i found is that there proves to be a relationship between these 2 variables. The higher the total expenditure, the higher their life expectancy will be.**

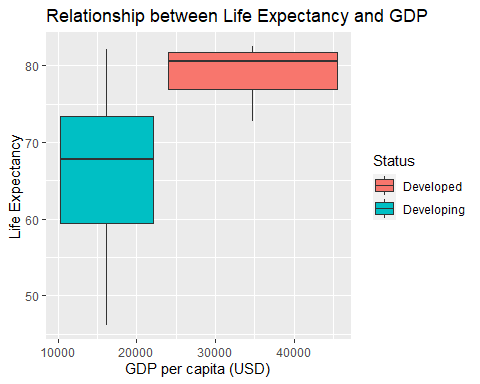
**Finding the correlation between life expectancy with total expenditure**

source("https://raw.githubusercontent.com/briatte/ggcorr/master/ggcorr.R")  
sub\_data <- data %>% select(Country,Life.expectancy, Total.expenditure)  
data\_num <- sub\_data %>% ungroup() %>% select\_if(is.numeric)  
ggcorr(data\_num,   
 label = T,   
 label\_size = 2,  
 label\_round = 2,  
 hjust = 1,  
 size = 3,   
 color = "royalblue",  
 layout.exp = 5,  
 low = "green3",   
 mid = "gray95",   
 high = "darkorange",  
 name = "Correlation")

 **Moving on, here I used the ggcorr function to find the correlation between total expenditure and life expentancy. The correlation between these 2 is at +0.22. This implies that there is a positive relationship between these two variables, although weak and likely insignicant**

# Graph of Life Expectancy and Gross Domestic Product per capita (GDP)

df3 <- data %>% group\_by(Country) %>% mutate(meanLifeExp = mean(Life.expectancy), meanGDP = mean(GDP)) %>% distinct(Country, .keep\_all = TRUE) %>% ungroup() %>% select(Status, meanLifeExp, meanGDP)  
graph3 <- ggplot(df3, aes(x = meanGDP, y = meanLifeExp, fill = Status)) + geom\_boxplot() + ggtitle("Relationship between Life Expectancy and GDP") + xlab("GDP per capita (USD)") + ylab("Life Expectancy")  
graph3

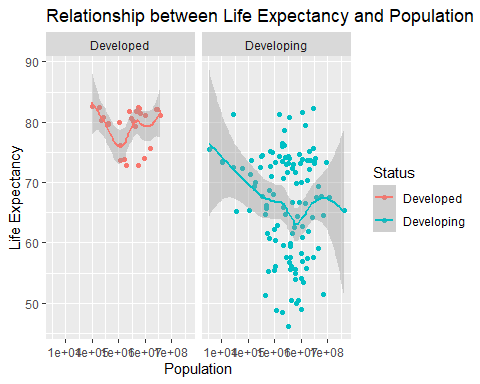


**Boxplot is plotted to see the relationship between GDP and Life Expectancy of Developed Countries and Developing Countries. What I have found is that developing countries tend to have 10k to 25k GDP per capita (USD) and the median of Life Expectancy is around 68 years old. This is lower as compared to developed countries which tend to have higher GDP per capita(USD) and higher life expectancy of around 80 years old.**

# Graph of Life Expectancy and Population

df4 <- data %>% group\_by(Country) %>% mutate(meanLifeExp = mean(Life.expectancy), meanPop = mean(Population)) %>% distinct(Country, .keep\_all = TRUE) %>% ungroup() %>% select(Status, meanLifeExp, meanPop)  
graph4 <- ggplot(df4, aes(x = meanPop, y = meanLifeExp, color = Status)) + geom\_point() + ggtitle("Relationship between Life Expectancy and Population") + xlab("Population") + ylab("Life Expectancy") + scale\_x\_log10() + facet\_wrap(~Status) + geom\_smooth()  
graph4

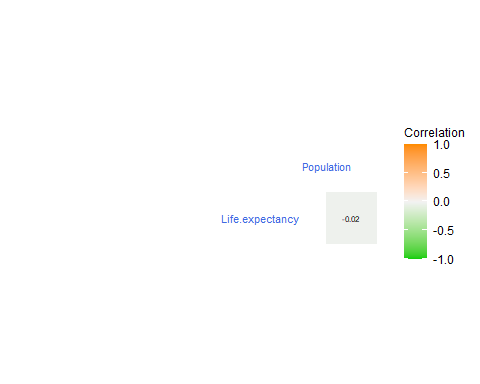
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



**From the graph of Life expectancy and population above, it is clear that life expectancy is negatively correlated to population size in developing countries and the correlation is not clear in developed countries.**

# Finding the correlation between life expectancy with population

source("https://raw.githubusercontent.com/briatte/ggcorr/master/ggcorr.R")  
sub\_data <- data %>% select(Country,Life.expectancy, Population)  
data\_num <- sub\_data %>% ungroup() %>% select\_if(is.numeric)  
ggcorr(data\_num,   
 label = T,   
 label\_size = 2,  
 label\_round = 2,  
 hjust = 1,  
 size = 3,   
 color = "royalblue",  
 layout.exp = 5,  
 low = "green3",   
 mid = "gray95",   
 high = "darkorange",  
 name = "Correlation")



**It is also shown above that life expectancy is slightly negatively correlated with population.**

CountryLifeExp <- data %>% group\_by(Country) %>% select(Country, Life.expectancy, Alcohol, Total.expenditure, GDP, Population, Income.composition.of.resources) %>% summarize(meanLifeExp = mean(Life.expectancy), meanAlcohol = mean(Alcohol), meanExpenditure = mean(Total.expenditure), meanGDP = mean(GDP), meanPopulation = mean(Population))  
  
CountryLifeExp[which.max(CountryLifeExp$meanLifeExp),]

## # A tibble: 1 x 6  
## Country meanLifeExp meanAlcohol meanExpenditure meanGDP meanPopulation  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Japan 82.5 7.07 6.47 24893. 97384.

**The Country with the highest life expectancy is Japan with the mean life expectancy of 82 years old from 2000-2015. Japan has the alcohol consumption of about 7 litre of pure alcohol per capita. Japan spent about 6% expenditure on health as of total government expenditure, and have Gross Domestic Product Per capita(USD) of 24892.54 dollars. Japan have around 100,000 citizens from 2000-2015.**

# Part 6: Correlation

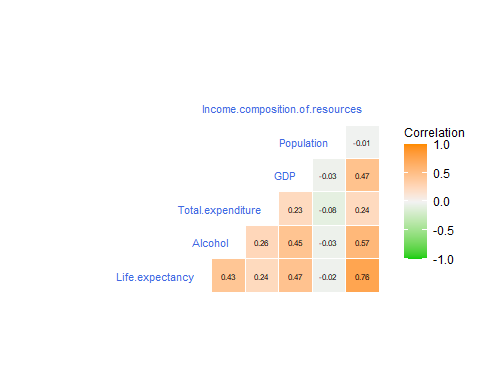
library(DT)

## Warning: package 'DT' was built under R version 3.6.3

subset <- subset(data, select = c(Life.expectancy,Alcohol,Total.expenditure,GDP,Population))  
  
cor(subset) %>% datatable() %>% formatRound(columns = c("Life.expectancy","Alcohol","Total.expenditure","GDP","Population"), digits = 3)



source("https://raw.githubusercontent.com/briatte/ggcorr/master/ggcorr.R")  
ggcorr(correlationFrame,   
 label = T,   
 label\_size = 2,  
 label\_round = 2,  
 hjust = 1,  
 size = 3,   
 color = "royalblue",  
 layout.exp = 5,  
 low = "green3",   
 mid = "gray95",   
 high = "darkorange",  
 name = "Correlation")

 # **Part 7 : Conclusion of Descriptive Analysis**

**From the analysis above, we can reasonably conclude that higher alcohol consumption is linked to higher life expectancy with a correlation of about 0.4. Government expenditure on health as a percentage of total government expenditure(%) is slightly linked to higher life expectancy, though the effect is not obvious with a correlation of about 0.2. For GDP per capita(USD), it has a relatively higher effect on the life expectancy. It makes sense as the income of a country or economic performance of a country is directly reflected by the GDP per capita. Hence, the life expectancy increases with the increment in GDP per capita of a country. Countries with high population are associated with lower life expectancy of its citizens. This is because the resources of the country have to be distributed across its citizens, resulting each citizen getting lesser health resources.**

# Part 8: Multivariable Linear Regression

# Clean data where we treat developing as 0 and developed as 1 to ease linear modelling

clean<-file %>% drop\_na()   
clean <- clean %>%  
 mutate(clean\_status = ifelse(Status == "Developing",0,1))  
linear\_model<-lm(Life.expectancy~.-Country-Year-Status,data=clean)  
linear\_model

##   
## Call:  
## lm(formula = Life.expectancy ~ . - Country - Year - Status, data = clean)  
##   
## Coefficients:  
## (Intercept) Adult.Mortality   
## 5.348e+01 -1.663e-02   
## infant.deaths Alcohol   
## 9.350e-02 -9.140e-02   
## percentage.expenditure Hepatitis.B   
## 3.673e-04 -6.525e-03   
## Measles BMI   
## -7.865e-06 3.376e-02   
## under.five.deaths Polio   
## -7.035e-02 7.935e-03   
## Total.expenditure Diphtheria   
## 7.586e-02 1.490e-02   
## HIV.AIDS GDP   
## -4.370e-01 8.738e-06   
## Population thinness..1.19.years   
## -6.425e-10 -1.238e-02   
## thinness.5.9.years Income.composition.of.resources   
## -4.798e-02 9.817e+00   
## Schooling clean\_status   
## 8.665e-01 9.684e-01

# y = c + m1x1 + m2x2 + ....

**p value<2.2e-16 means there is at least one variable that is highly significant to life expectancy.** **To see which predictor variables are significant (Significant test)**

coefficient<-summary(linear\_model)$coefficient  
coefficient

## Estimate Std. Error t value  
## (Intercept) 5.348274e+01 7.375397e-01 72.5150681  
## Adult.Mortality -1.663174e-02 9.494415e-04 -17.5173908  
## infant.deaths 9.349971e-02 1.065327e-02 8.7766178  
## Alcohol -9.139501e-02 3.316341e-02 -2.7558989  
## percentage.expenditure 3.673363e-04 1.801074e-04 2.0395401  
## Hepatitis.B -6.524647e-03 4.448550e-03 -1.4666908  
## Measles -7.865434e-06 1.078595e-05 -0.7292296  
## BMI 3.375565e-02 5.998098e-03 5.6277248  
## under.five.deaths -7.034836e-02 7.711285e-03 -9.1227806  
## Polio 7.935254e-03 5.151979e-03 1.5402342  
## Total.expenditure 7.585822e-02 4.067445e-02 1.8650089  
## Diphtheria 1.489933e-02 5.927886e-03 2.5134300  
## HIV.AIDS -4.369640e-01 1.784256e-02 -24.4899830  
## GDP 8.737938e-06 2.837049e-05 0.3079939  
## Population -6.424645e-10 1.748706e-09 -0.3673942  
## thinness..1.19.years -1.238499e-02 5.300123e-02 -0.2336736  
## thinness.5.9.years -4.798356e-02 5.231461e-02 -0.9172113  
## Income.composition.of.resources 9.816570e+00 8.321251e-01 11.7969891  
## Schooling 8.665032e-01 5.940312e-02 14.5868300  
## clean\_status 9.683668e-01 3.379401e-01 2.8654984  
## Pr(>|t|)  
## (Intercept) 0.000000e+00  
## Adult.Mortality 4.421388e-63  
## infant.deaths 4.167811e-18  
## Alcohol 5.918326e-03  
## percentage.expenditure 4.155722e-02  
## Hepatitis.B 1.426532e-01  
## Measles 4.659661e-01  
## BMI 2.146026e-08  
## under.five.deaths 2.098456e-19  
## Polio 1.236975e-01  
## Total.expenditure 6.235983e-02  
## Diphtheria 1.205217e-02  
## HIV.AIDS 4.926216e-113  
## GDP 7.581264e-01  
## Population 7.133727e-01  
## thinness..1.19.years 8.152678e-01  
## thinness.5.9.years 3.591677e-01  
## Income.composition.of.resources 7.022005e-31  
## Schooling 2.172561e-45  
## clean\_status 4.216969e-03

**Ignore those variables in our model where p value>0.05(not significant).**

linear\_model<-lm(Life.expectancy~Adult.Mortality+infant.deaths+Alcohol+percentage.expenditure+BMI+under.five.deaths+Total.expenditure+Diphtheria+HIV.AIDS+Income.composition.of.resources+Schooling+clean\_status,data=clean)  
linear\_model

##   
## Call:  
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +   
## Alcohol + percentage.expenditure + BMI + under.five.deaths +   
## Total.expenditure + Diphtheria + HIV.AIDS + Income.composition.of.resources +   
## Schooling + clean\_status, data = clean)  
##   
## Coefficients:  
## (Intercept) Adult.Mortality   
## 52.9409973 -0.0168480   
## infant.deaths Alcohol   
## 0.0900533 -0.0774547   
## percentage.expenditure BMI   
## 0.0004285 0.0381165   
## under.five.deaths Total.expenditure   
## -0.0687366 0.0807168   
## Diphtheria HIV.AIDS   
## 0.0146054 -0.4385139   
## Income.composition.of.resources Schooling   
## 9.9998733 0.8752576   
## clean\_status   
## 0.9426198

# Residual Standard Error (RSE)

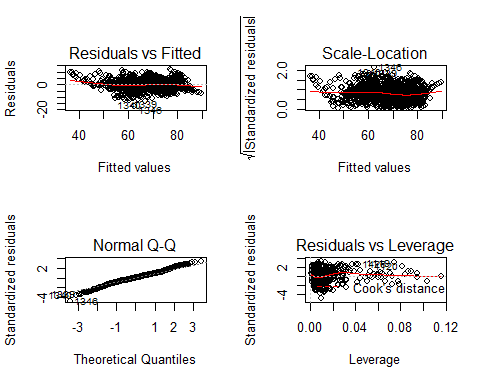
error<-sigma(linear\_model)/mean(clean$Life.expectancy)   
error

## [1] 0.05180732

**In our multiple regression example, the RSE is 0.0518 corresponding to 5% error rate.**

**fitted(linear\_model) predicted values** **regression diagnostics, give the error from actual life expectancy**

error\_from\_actual<-influence(linear\_model)   
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page  
plot\_regression\_lineplot<-plot(linear\_model)



# Link to our Shiny App

<https://hongyanlee.shinyapps.io/WIE2003_Intro_to_Data_Science/>