Health Care Analytics

Vamsi Tallapu Reddy, Sujan Barama, Atharva Dastane, Anuradha Mysore Ravishankar, Anita T

9/23/2021

Final Project DTSC 5301-001: Data Science as a Field

Key Question - How have global health trends changed over the last 2 decades?

What are the basic health metric trends?

What are the major disease trends?

How have vaccination rates changed over time?

What is the financial impact caused by these changes?

Loading all the data

Data Source:https://apps.who.int//nha//database//Home//IndicatorsDownload//en

Other data sources:

1. https://www.who.int/teams/immunization-vaccines-and-biologicals/vaccine-access/planning-and-financing/immunization-financing-indicators

2. https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete

github link - https://github.com/adastane100/DTSC5301__PROJECT.git

Short data description

WHO data - data of health expenditure of 190 countries at a country x year level. Sample columns are; GDP, population, current health expenditure as a % of GDP, out of pocket expenses, voluntary health insurance, etc.

Kaggle data - No of doctors, nurses, life expectancy, mortality rate, tuberculosis, malaria cases, etc at a country year level

WHO immunization data - Amount spent on immunization at a country year level

Code starts here:

Loading libraries

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Load the main WHO data & clean data

```
## # A tibble: 3,648 x 12
##    country year  income_group region phc_che phc_usd_pc      gdp    pop che_gdp
##    <chr>   <chr> <chr>        <chr>    <dbl>      <dbl>    <dbl>  <dbl>   <dbl>
##  1 Algeria 2000  Up-Mid       AFR         NA         NA  4123500 31042.    3.49
##  2 Algeria 2001  Up-Mid       AFR         NA         NA  4227100 31452.    3.84
##  3 Algeria 2002  Up-Mid       AFR         NA         NA  4522800 31855.    3.73
##  4 Algeria 2003  Up-Mid       AFR         NA         NA  5252300 32264.    3.60
##  5 Algeria 2004  Up-Mid       AFR         NA         NA  6149100 32692.    3.54
##  6 Algeria 2005  Up-Mid       AFR         NA         NA  7562000 33150.    3.24
##  7 Algeria 2006  Up-Mid       AFR         NA         NA  8501636 33641.    3.36
##  8 Algeria 2007  Up-Mid       AFR         NA         NA  9352886 34167.    3.82
##  9 Algeria 2008  Up-Mid       AFR         NA         NA 11043704 34731.    4.20
## 10 Algeria 2009  Up-Mid       AFR         NA         NA  9968025 35334.    5.36
## # ... with 3,638 more rows, and 3 more variables: che_pc_usd <dbl>,
## #   vhi_che <dbl>, oops_che <dbl>
```

Create functions for everything:

```
#To load data
load_data <-function(x){
  data<- read.csv(x)
}
#To select relevant data
filter_data <- function(df){
  if ("Dim1" %in% colnames(df) == TRUE){
    data <- filter(df, Dim1 == "Both sexes")
```

```
    data <- select(data, country = Location, year = Period, value = "First Tooltip")
  }
  else{
    data <- select(df, country = Location, year = Period, value = "First Tooltip")
  }
}
```

Extracting the data from the URL

```
url_in <-"https://raw.githubusercontent.com/adastane100/DTSC5301_PROJECT/main/Data/"
file_names <- c("crudeSuicideRates.csv","infantMortalityRate.csv","lifeExpectancyAtBirth.csv",
                "medicalDoctors.csv","neonatalMortalityRate.csv","nursingAndMidwife.csv",
                "pharmacists.csv","under5MortalityRate.csv","30-70cancerChdEtc.csv","newHivInfections.c
                "3-total-expenditure-routine-immunization.xlsx")
urls <- str_c(url_in,file_names)
```

Loading Kaggle data from the github

```
suicides <- read_csv(urls[1], show_col_types = FALSE)
mortality <- read_csv(urls[2], show_col_types = FALSE)
life_expectancy <- read_csv(urls[3], show_col_types = FALSE)
doctors <- read_csv(urls[4], show_col_types = FALSE)
neo_mortality <- read_csv(urls[5], show_col_types = FALSE)
nursing <- read_csv(urls[6], show_col_types = FALSE)
pharmacists <- read_csv(urls[7], show_col_types = FALSE)
kids_mortality <- read_csv(urls[8], show_col_types = FALSE)
cancer_cases <- read_csv(urls[9], show_col_types = FALSE)
hiv_cases <- read_csv(urls[10], show_col_types = FALSE)
tuberculosis_cases <- read_csv(urls[11], show_col_types = FALSE)
malaria_cases <- read_csv(urls[12], show_col_types = FALSE)
hepatitis_cases <- read_csv(urls[13], show_col_types = FALSE)
```

Transform the data that is extracted from Github and rename the columns in the main data set

```
#Transform Data
mortality_clean <- filter_data(mortality)
life_expectancy_clean <- filter_data(life_expectancy)
doctors_clean <- filter_data(doctors)
neo_mortality_clean <- filter_data(neo_mortality)
nursing_clean <- filter_data(nursing)
```

```
pharmacists_clean <- filter_data(pharmacists)
kids_mortality_clean <- filter_data(kids_mortality)
cancer_cases_clean <- filter_data(cancer_cases)
hiv_cases_clean <- filter_data(hiv_cases)
tuberculosis_cases_clean <- filter_data(tuberculosis_cases)
malaria_cases_clean <- filter_data(malaria_cases)
hepatitis_cases_clean <- filter_data(hepatitis_cases)

# Rename value column

mortality_clean <- rename(mortality_clean, mortality_rate = value)
life_expectancy_clean <- rename(life_expectancy_clean, life_expectancy = value)
doctors_clean <- rename(doctors_clean, doctors = value)
neo_mortality_clean <- rename(neo_mortality_clean, neo_mortality_rate = value)
nursing_clean <- rename(nursing_clean, nurses_midwives = value)
pharmacists_clean <- rename(pharmacists_clean, pharmacists = value)
kids_mortality_clean <- rename(kids_mortality_clean, suicide_rate = value)
nursing_clean <- filter(nursing_clean, year >= 2000)
cancer_cases_clean <- rename(cancer_cases_clean, cancer_cases = value)
hiv_cases_clean <- rename(hiv_cases_clean, hiv_cases = value)
tuberculosis_cases_clean <- rename(tuberculosis_cases_clean, tuberculosis_cases = value)
malaria_cases_clean <- rename(malaria_cases_clean, malaria_cases = value)
hepatitis_cases_clean <- rename(hepatitis_cases_clean, hepatitis_cases = value)
```

Merge all dataframes to create a master dataframe

```
merge_cols = c("country, year")
merge1 <- merge(newdata, kids_mortality_clean, on = merge_cols, all.x = TRUE)
nrow(newdata)
```

```
## [1] 3648
```

```
merge2 <- merge(merge1, pharmacists_clean, on = merge_cols, all.x = TRUE)
merge23 <- merge(merge2, nursing_clean, on = merge_cols, all.x = TRUE)
merge3 <- merge(merge23, neo_mortality_clean, on = merge_cols, all.x = TRUE)
new_merge <- merge(merge3, doctors_clean, on = merge_cols, all.x = TRUE)
mer1 <- merge(new_merge, cancer_cases_clean, on = merge_cols, all.x = TRUE)
mer2 <- merge(mer1, tuberculosis_cases_clean, on = merge_cols, all.x = TRUE)
mer3 <- merge(mer2, malaria_cases_clean, on = merge_cols, all.x = TRUE)
mer4 <- merge(mer3, hiv_cases_clean, on = merge_cols, all.x = TRUE)
mer5 <- merge(mer4, hepatitis_cases_clean, on = merge_cols, all.x = TRUE)
new_merge2 <- merge(mer5, life_expectancy_clean, on = merge_cols, all.x = TRUE)
master_df <- merge(new_merge2, mortality_clean, on = merge_cols, all.x = TRUE)
#master_df <- merge(new_merge3, suicides_clean, on = merge_cols, all.x = TRUE)
```

Transform data further to begin analysis

```
master_df["health_expenditure"] = (master_df$che_gdp/100)*master_df$gdp
master_df <- filter(master_df, income_group == "Low" | income_group == "Low-Mid")
master_df$che <- substr(master_df$suicide_rate,1,5)
master_df$mortality_rate <- substr(master_df$mortality_rate,1,5)
master_df$suicide_rate <- as.numeric(master_df$suicide_rate)
```

```
## Warning: NAs introduced by coercion
```

```
master_df$mortality_rate <- as.numeric(master_df$mortality_rate)
```

```
## Warning: NAs introduced by coercion
```

How have doctor numbers changed from 2000 to 2018?

```
# Filtering data for 2000 & 2018 to compare the two
doc2000 <- filter(master_df, year == 2000 | year == 2001, doctors >= 0)
doc2000 <- select(doc2000, country, year, doctors)

doc2018 <- filter(master_df, year == 2017 | year == 2018, doctors >= 0)
doc2018 <- select(doc2018, country, year, doctors)

#Selecting countries that are common for 2000,2001 and 2017,2018 as comparison can only be made when th

countries_2018 <- unique(doc2018$country)
countries_2000 <- unique(doc2000$country)

unique_countries = countries_2000[countries_2000 %in% countries_2018]

#Checking mean of doctors for 2000 for common countries
doc2000 <- filter(doc2000, country == "Bangladesh" | country == "Chad" | country == "Honduras" |
         country == "India" | country == "Lao People's Democratic Republic" | country == "Pakistan" |
         country == "Papua New Guinea" |
         country == "Tunisia" | country == "Zimbabwe")

mean(doc2000$doctors)
```

```
## [1] 4.351667
```

```
#Checking mean of doctors for 2018 for common countries
doc_2018 <- filter(doc2018, country == "Bangladesh" | country == "Chad" | country == "Honduras" |
                     country == "India" | country == "Lao People's Democratic Republic" | country == "Pa
                     country == "Papua New Guinea"  |
                     country == "Tunisia" | country == "Zimbabwe")

doc_2018
```

```
##                              country year doctors
## 1                         Bangladesh 2017    5.43
## 2                         Bangladesh 2018    5.81
## 3                               Chad 2017    0.43
## 4                           Honduras 2017    3.09
## 5                              India 2017    7.78
## 6                              India 2018    8.57
## 7  Lao People's Democratic Republic 2017    3.73
## 8                           Pakistan 2017   10.01
## 9                           Pakistan 2018    9.80
## 10                  Papua New Guinea 2018    0.70
## 11                           Tunisia 2017   13.03
## 12                          Zimbabwe 2017    1.86
## 13                          Zimbabwe 2018    2.10
```

```
mean(doc_2018$doctors)
```

```
## [1] 5.564615
```

```
 #5.5

# Biggest Winners - Bangladesh, India
```

How have pharmacists changed from 2000 to 2018?

```
# Filtering data for 2000 & 2018 to compare the two
pharm2000 <- filter(master_df, year == 2000| year == 2001 | year == 2002,  nurses_midwives >= 0)
pharm2000 <- select(pharm2000, country, year, pharmacists_2000 = pharmacists)

pharm2018 <- filter(master_df, year == 2016 | year == 2017 | year == 2018, nurses_midwives >= 0)
pharm2018 <- select(pharm2018, country, year, pharmacists_2018 = pharmacists)

#checking for common countries among the two sets
countries_2018 <- unique(pharm2018$country)
countries_2000 <- unique(pharm2000$country)

head(countries_2018)
```

```
## [1] "Afghanistan"  "Angola"       "Bangladesh"   "Benin"        "Bhutan"
## [6] "Burkina Faso"
```

```
unique_countries = countries_2000[countries_2000 %in% countries_2018]
head(unique_countries)
```

```
## [1] "Cambodia" "Chad"     "Eswatini" "Guinea"   "Honduras" "India"
```

```r
pharmacy <- merge(pharm2000, pharm2018, all = TRUE)
head(pharmacy)
```

```
##       country year pharmacists_2000 pharmacists_2018
## 1 Afghanistan 2016               NA             0.47
## 2 Afghanistan 2017               NA               NA
## 3      Angola 2018               NA               NA
## 4  Bangladesh 2016               NA               NA
## 5  Bangladesh 2017               NA             1.61
## 6  Bangladesh 2018               NA             1.81
```

```r
#Biggest Winners - Nepal, East Timor
```

How have nurses/midwives changed from 2000 and 2018

```r
#Filtering out data for 2000 and 2018
nurses2000 <- filter(master_df, year == 2002| year == 2003 | year == 2004,  pharmacists >= 0)
nurses2000 <- select(nurses2000, country, year, nurses_2000 = nurses_midwives)

nurses2018 <- filter(master_df, year == 2016 |year == 2017 | year == 2018, pharmacists >= 0)
nurses2018 <- select(nurses2018, country, year, nurses_2018 = nurses_midwives)

# Performing an inner join to get common comparable data
nurses <- merge(nurses2000, nurses2018, all = TRUE)
head(nurses)
```

```
##       country year nurses_2000 nurses_2018
## 1 Afghanistan 2016          NA        1.48
## 2      Angola 2004        9.85          NA
## 3  Bangladesh 2002          NA          NA
## 4  Bangladesh 2017          NA        3.16
## 5  Bangladesh 2018          NA        4.12
## 6       Benin 2004        7.47          NA
```

```r
#From this table
#Biggest Winners - Nepal, Indonesia
#Change from 9.625 -  13.7
```

How has health expenditure per capita changed from 2000 to 2018

```r
#Filtering out data for 2000 and 2018
expenditure2000 <- filter(master_df, year == 2002| year == 2003 | year == 2004,  che_pc_usd >= 0)
expenditure2000 <- select(expenditure2000, country, year, health_expenditure_2000 = che_pc_usd)
```

```r
expenditure2018 <- filter(master_df, year == 2016 |year == 2017 | year == 2018, che_pc_usd >= 0)
expenditure2018 <- select(expenditure2018, country, year, health_expenditure_2018 = che_pc_usd)

#Inner join to get common countries
expenditure <- merge(expenditure2000, expenditure2018, all = TRUE)
head(expenditure)
```

```
##       country year health_expenditure_2000 health_expenditure_2018
## 1 Afghanistan 2002                15.80316                      NA
## 2 Afghanistan 2003                17.03574                      NA
## 3 Afghanistan 2004                20.41276                      NA
## 4 Afghanistan 2016                      NA                60.18867
## 5 Afghanistan 2017                      NA                65.70602
## 6 Afghanistan 2018                      NA                49.84261
```

```r
#From the table
#Biggest Winners - Sudan, Myanmar ~300 dollars/person, No point in this

#Change from 32.78 - 100
```

How has life expectancy changed from 2000 to 2018

```r
#Filter out data for 2000 and 2018
life_expectancy2000 <- filter(master_df, year == 2000| year == 2005 ,  life_expectancy >= 0)
life_expectancy2000 <- select(life_expectancy2000, country, year, life_expectancy_2000 = life_expectancy

life_expectancy2018 <- filter(master_df, year == 2015 |year == 2019 , life_expectancy >= 0)
life_expectancy2018 <- select(life_expectancy2018, country, year, life_expectancy2018 = life_expectancy)

#Inner join to get common countries
expectancy <- merge(life_expectancy2000, life_expectancy2018, all = TRUE)
head(expectancy)
```

```
##       country year life_expectancy_2000 life_expectancy2018
## 1 Afghanistan 2000                54.99                  NA
## 2 Afghanistan 2015                   NA               61.65
## 3      Angola 2000                49.30                  NA
## 4      Angola 2015                   NA               61.72
## 5  Bangladesh 2000                65.59                  NA
## 6  Bangladesh 2015                   NA               73.58
```

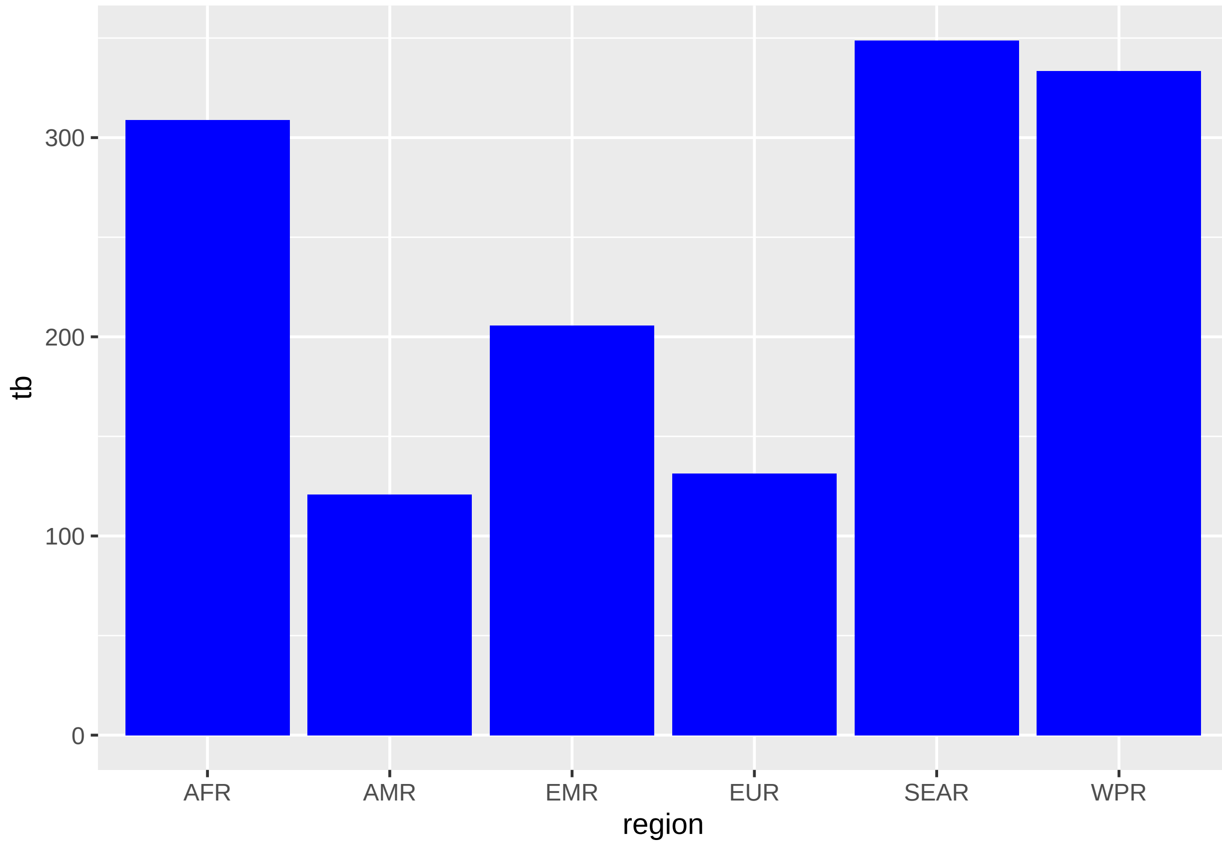
```r
#latest data available was 2015
#From the table
#Biggest Winners - Tunisia & Bangladesh, Rwanda & Burundi  - Tunisia ~75, Malaysia, South Africa
#Change from 58 - 65
```

How has the trends in health expenditure impacted Tuberculosis?

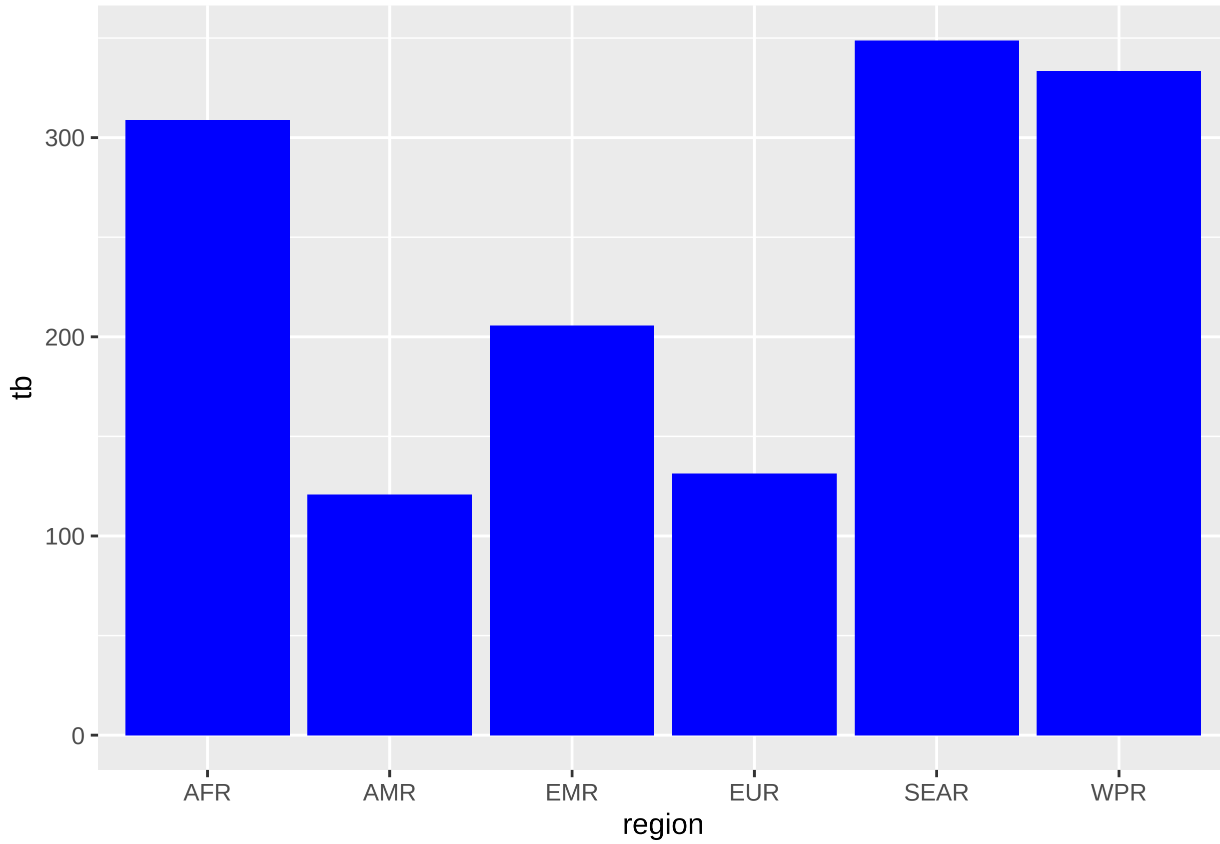
```
# Calculation of Current Health Expenditure from GDP
master_df<-master_df%>% mutate(CurrentHealthExp= (che_gdp*gdp)/100)

#Preparing the Required Data
gfg <- master_df$tuberculosis_cases
res = as.numeric(gsub(".*?([0-9]+).*", "\\1", gfg))
tuberculosis<-master_df%>%
  mutate(tuberculosis_cases_per_1000_pop = res)%>%
  group_by(year)%>%
  summarize(m=mean(tuberculosis_cases_per_1000_pop,na.rm=TRUE))
che<- master_df %>% group_by(year)%>%
  summarise(che1=log(mean(CurrentHealthExp,na.rm=TRUE)))

f<-merge(tuberculosis,che,on=year)
```

```
#Plotting health expenditure vs Tuberculosis per 1000 thousand population
g2<-ggplot(data = f, aes(x = year,group=1)) +
    geom_line(aes(y= m,group=1),color= "blue")+
  geom_line(aes(y=che1,group=1), color="green")
```

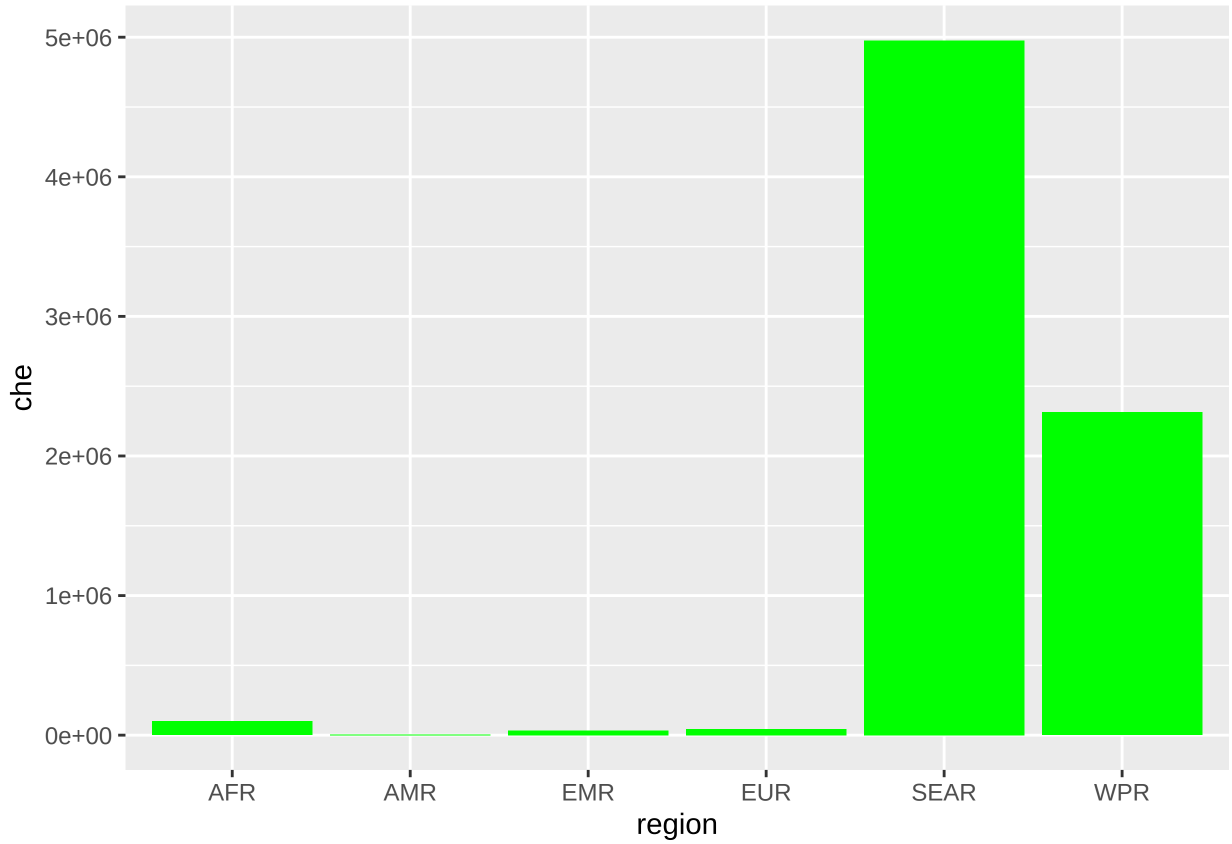
```
#Analyzing and Plotting Region Wise from 2000 to 2018 for TB per thousand vs Health Expenditure
master_df<-master_df%>%
  mutate(tuberculosis_cases_per_1000_pop = res)
region<- master_df%>%filter(year==2000)%>% group_by(region)%>% summarize (tb= mean(tuberculosis_cases_pe
region1<- master_df%>%filter(year==2000)%>% group_by(region)%>% summarize (che= mean(CurrentHealthExp, n
ggp <- ggplot(region, aes(x = region, y = tb,group=1)) +  # Create stacked bar chart
  geom_bar(stat = "identity",  fill= 'Blue')
ggp
```



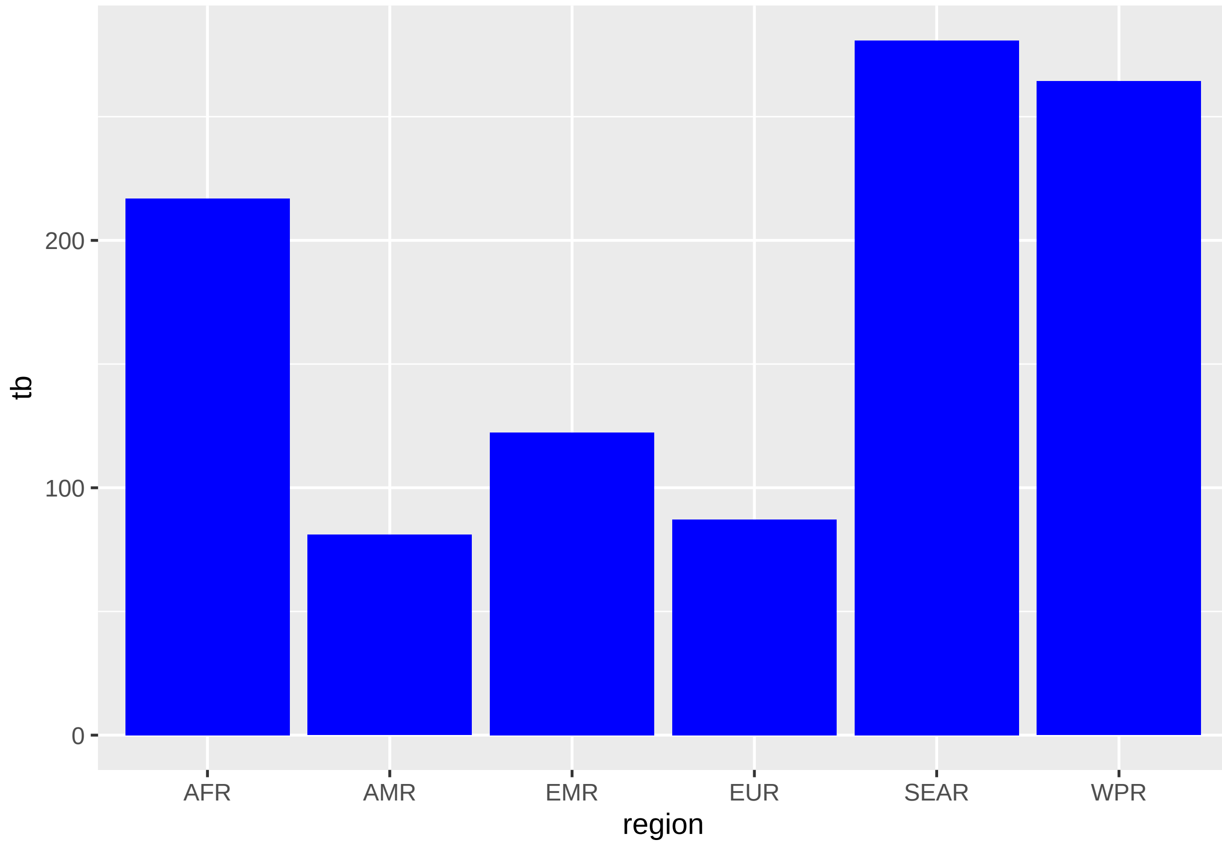

```
ggp1<- ggplot(region1,aes(x = region, y=che,group=1))+
  geom_bar(stat="identity", fill='Green')
ggp
```



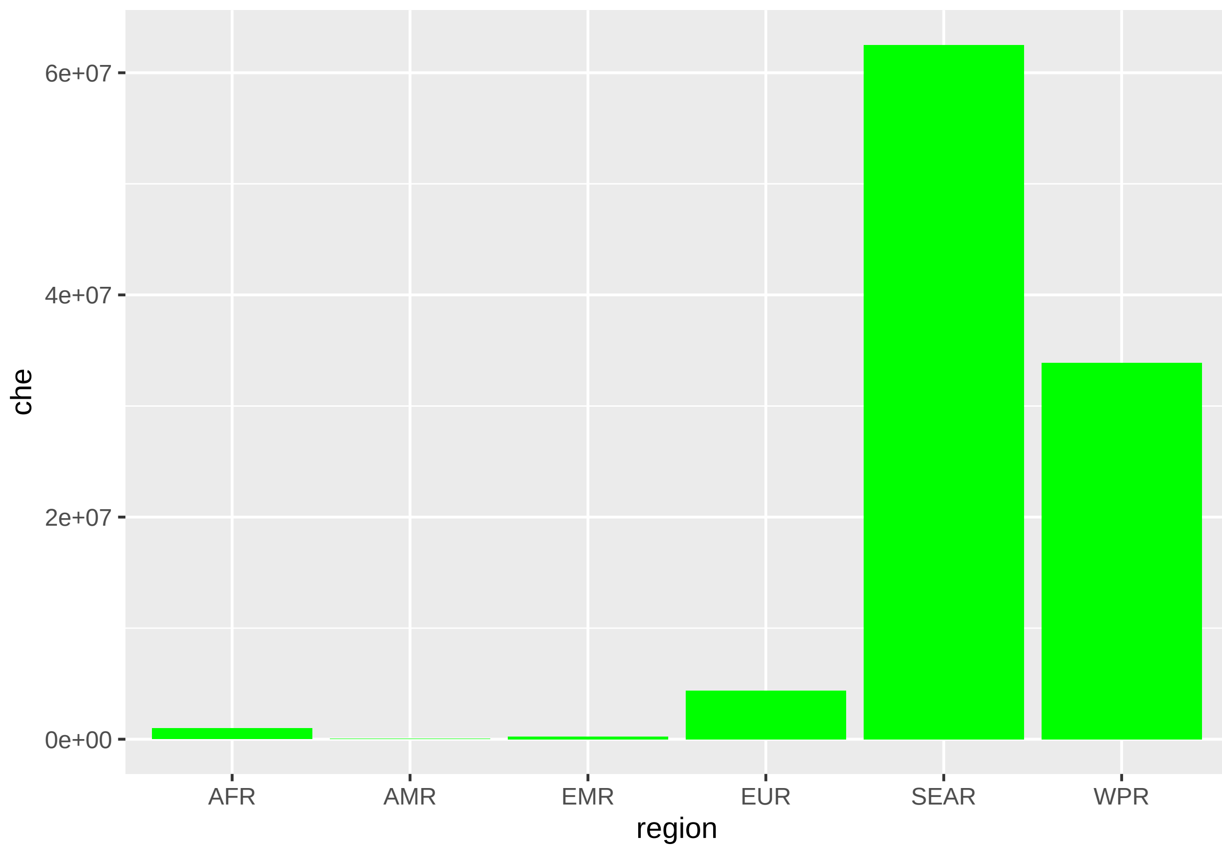

```
ggp1
```




```
region2018<- master_df%>%filter(year==2018)%>% group_by(region)%>% summarize (tb= mean(mean(tuberculosis
region2<- master_df%>%filter(year==2018)%>% group_by(region)%>% summarize (che= mean(CurrentHealthExp, n
ggp2018 <- ggplot(region2018, aes(x = region, y = tb,group=1) ) +  # Create stacked bar chart
  geom_bar(stat = "identity", fill='Blue')
ggp2018
```




```
ggpche<- ggplot(region2,aes(x = region, y=che,group=1))+
  geom_bar(stat="identity", fill= 'Green')
ggpche
```




```
library('ggpubr')
figure= ggarrange(ggp1,ggpche,ggp,ggp2018,labels=c('a','b','c','d'), ncol=2, nrow=2)
```

Conclusion: As the current health expenditure increased from 2000 to 2018, the tuberculosis cases decreased from 2000 to 2018 but we can see that the African and SEAR regions didn't show much improvement because in Africa, the patients who died due to TB has HIV and are considered as TB patients instead OF HIV. In SEAR, there is a country named Combodia in which 64% of people suffer from TB which increases the overall cases in TB. The bias in this is 40% of people are tested negative though they have the TB which makes the TB to spread.

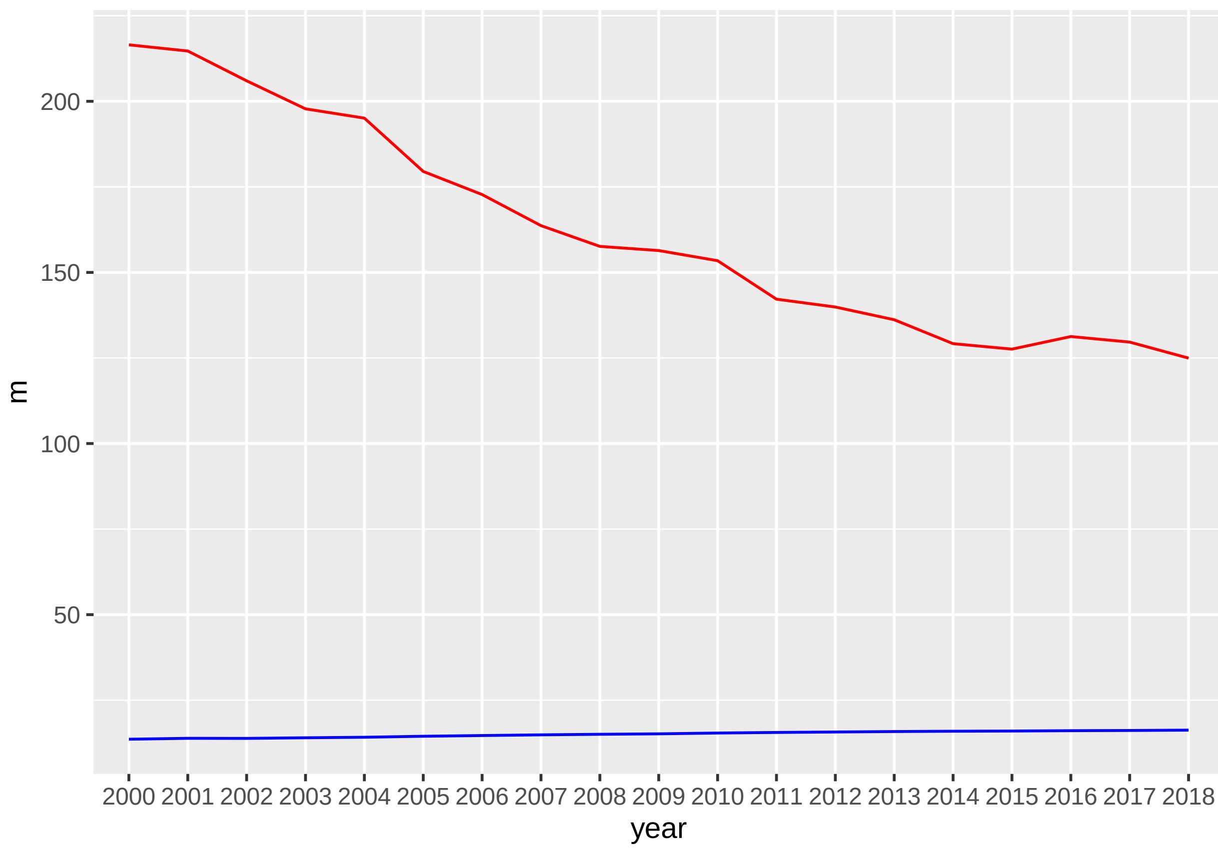
How have trends in Malaria changed from 2000 to 2018 with the growing Heath Expenditure?

```
#Calculating the Actual Current Health Expenditure From GDP
master_df<-master_df%>%mutate(CurrentHealthExp = (che_gdp/100)*gdp)

#Plotting for Current Health Expenditure vs Malaria cases per thousand Population
malaria <- master_df %>% group_by(year)%>%
  summarise(m=mean(malaria_cases,na.rm=TRUE))
gdp1<- master_df %>% group_by(year)%>%
  summarise(gdp=log(mean(CurrentHealthExp,na.rm=TRUE)))
```

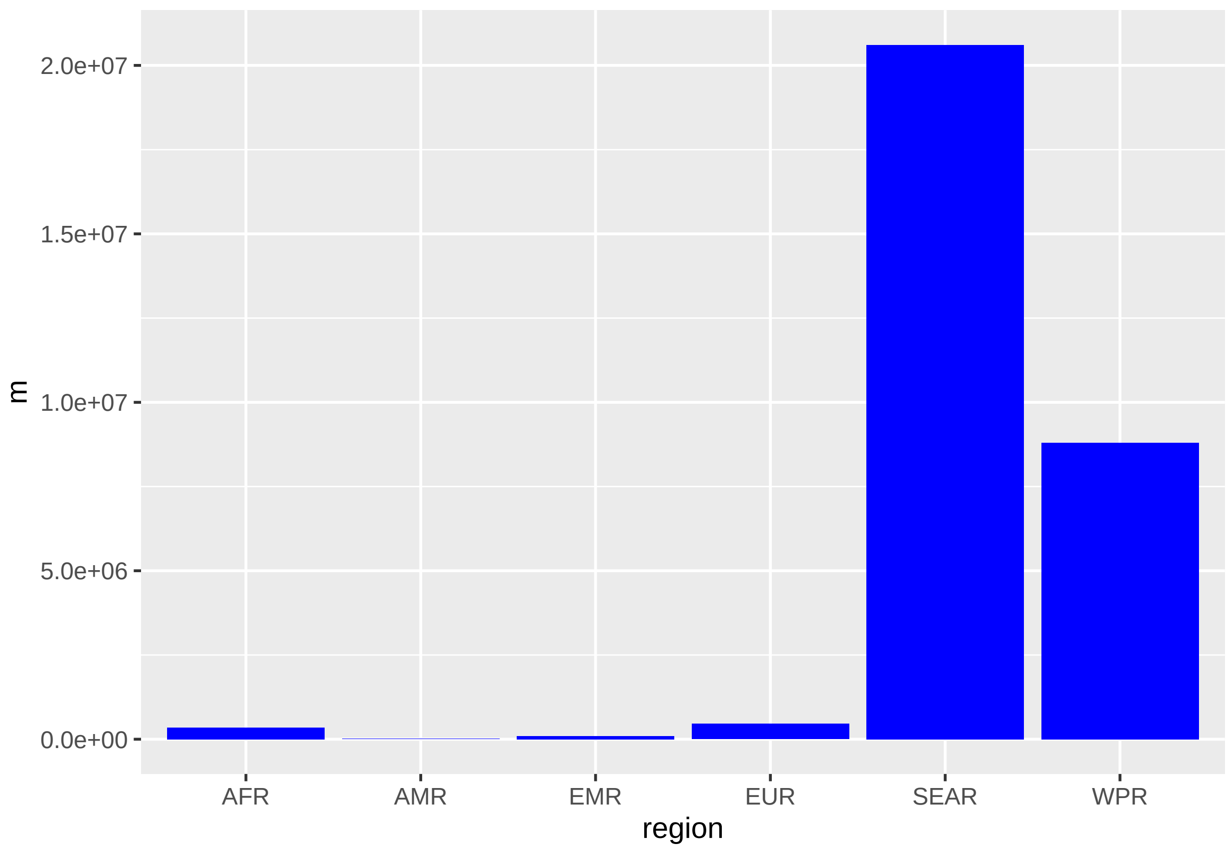
```r
f<-merge(malaria,gdp1,on=year)
f
```

```
##    year        m      gdp
## 1  2000 216.5042 13.59971
## 2  2001 214.6957 13.86423
## 3  2002 205.9895 13.84147
## 4  2003 197.8005 14.02347
## 5  2004 195.0900 14.17576
## 6  2005 179.5038 14.46476
## 7  2006 172.7509 14.67351
## 8  2007 163.6735 14.88031
## 9  2008 157.6053 15.04009
## 10 2009 156.3901 15.17645
## 11 2010 153.4213 15.40692
## 12 2011 142.1757 15.57595
## 13 2012 139.8865 15.71454
## 14 2013 136.1782 15.84648
## 15 2014 129.1847 15.93907
## 16 2015 127.5859 15.98915
## 17 2016 131.2589 16.09352
## 18 2017 129.6441 16.15806
## 19 2018 124.9629 16.25628
```

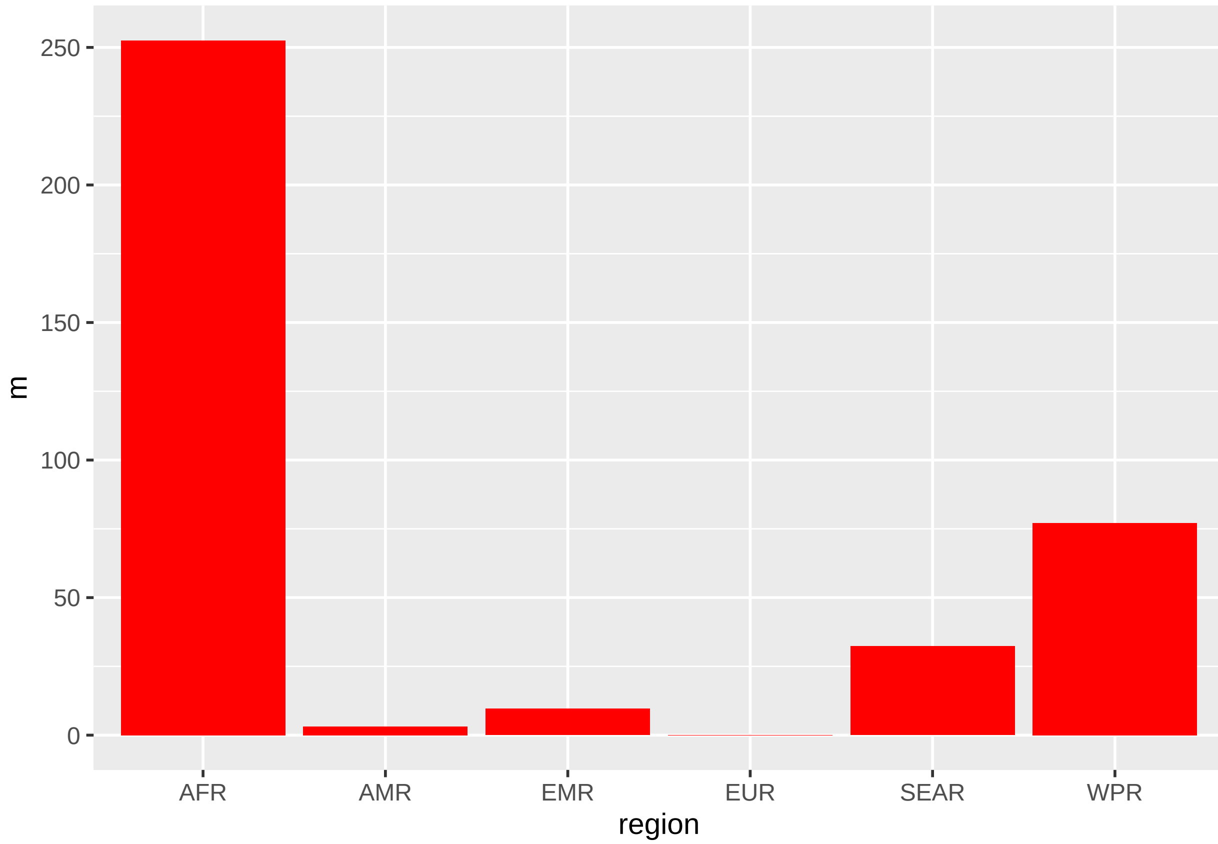
```r
write.csv(f,'analysis.csv')
g=ggplot(data=f,aes(x=year,group=1))+geom_line(aes(y=m,group=1),color='red')+geom_line(aes(y=gdp,group=
g
```



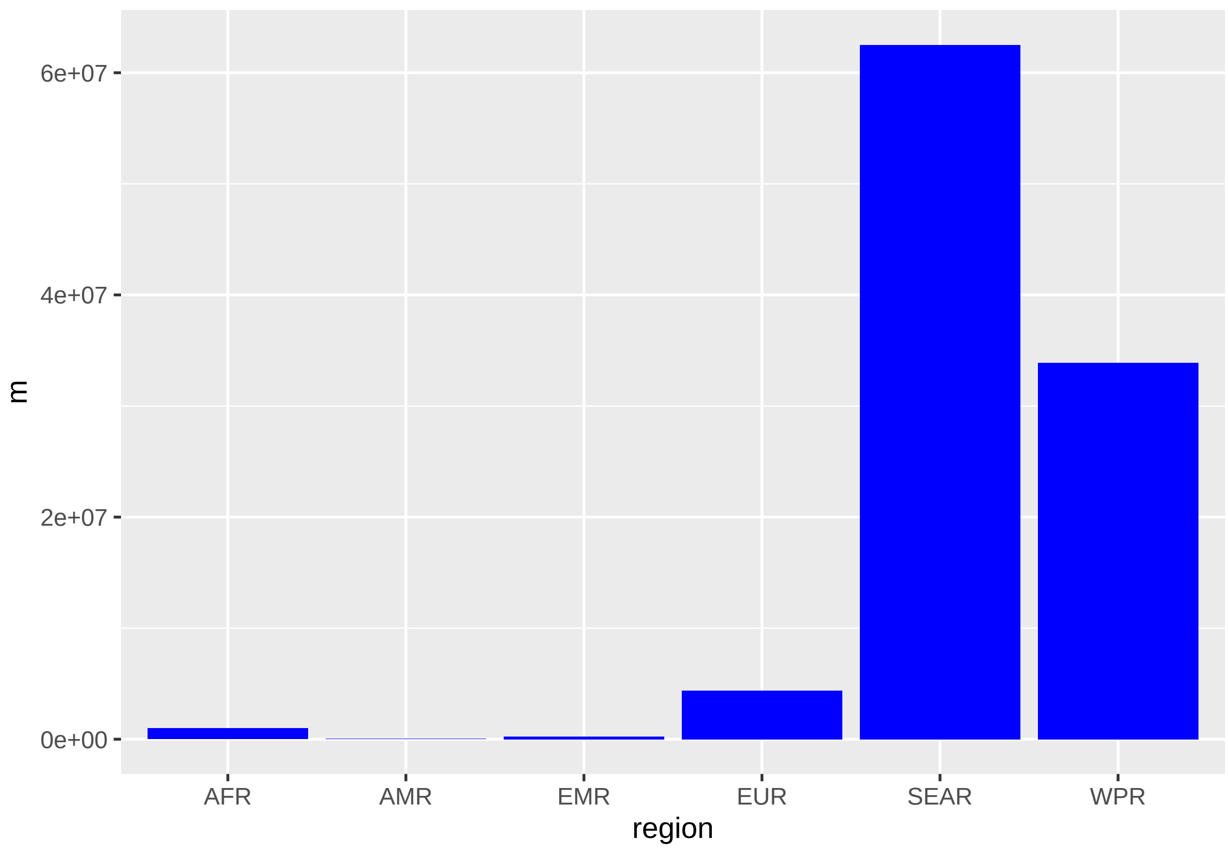

```r
#Analyzing and Plotting Region Wise from 2008 to 2018 for Malaria per thousand vs Health Expenditure

region_CHE_2008 <- master_df %>%
  filter(year == 2008)%>%
  group_by(region)%>%
  summarize(m=mean(CurrentHealthExp,na.rm=TRUE))
write.csv(region_CHE_2008,'analysis1.csv')
g_che_2008=ggplot(data=region_CHE_2008,aes(x=region,y=m,group=1))+geom_bar(stat='identity',fill='blue')
g_che_2008
```



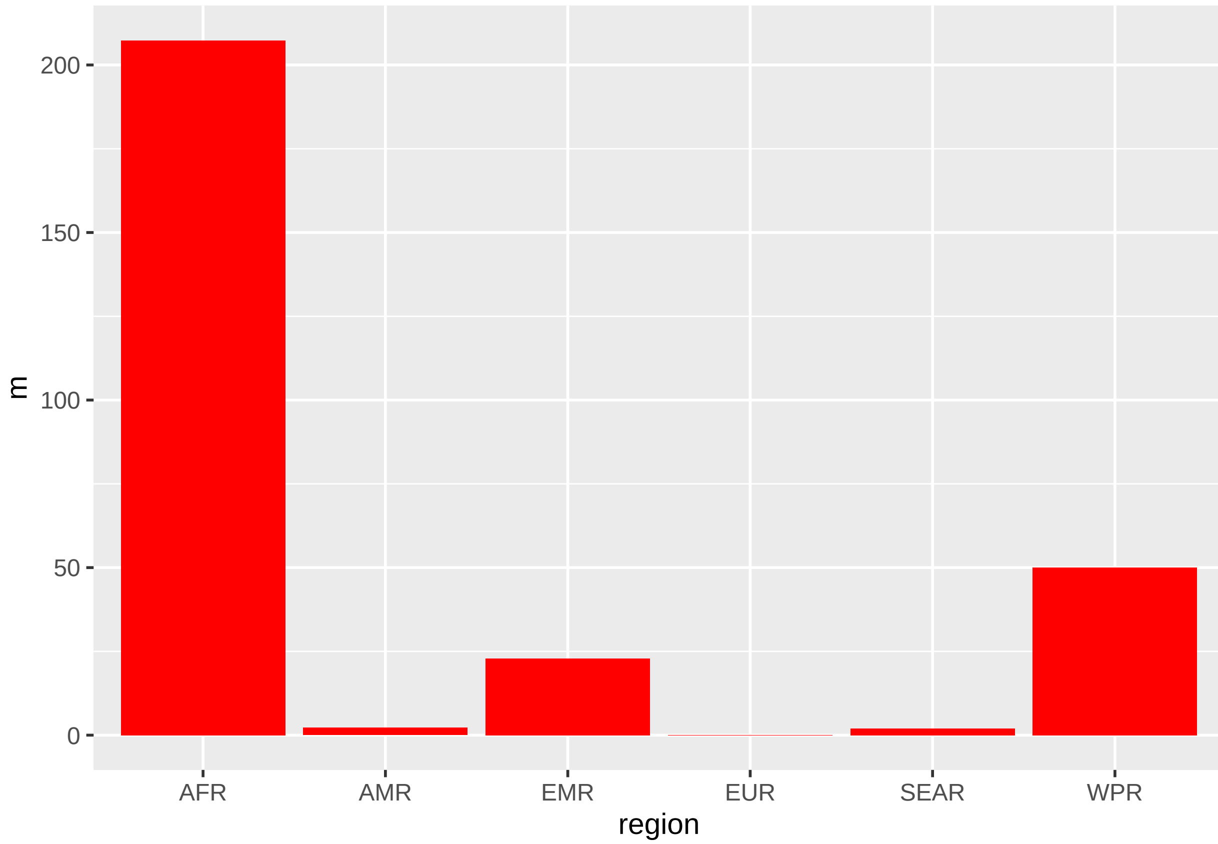

```
region_malaria_2008 <- master_df %>%
  filter(year == 2008)%>%
  group_by(region)%>%
  summarize(m=mean(malaria_cases,na.rm=TRUE))
write.csv(region_malaria_2008,'analysis2.csv')
g_mal_2008=ggplot(data=region_malaria_2008,aes(x=region,y=m,group=1))+geom_bar(stat='identity',fill='re
g_mal_2008
```



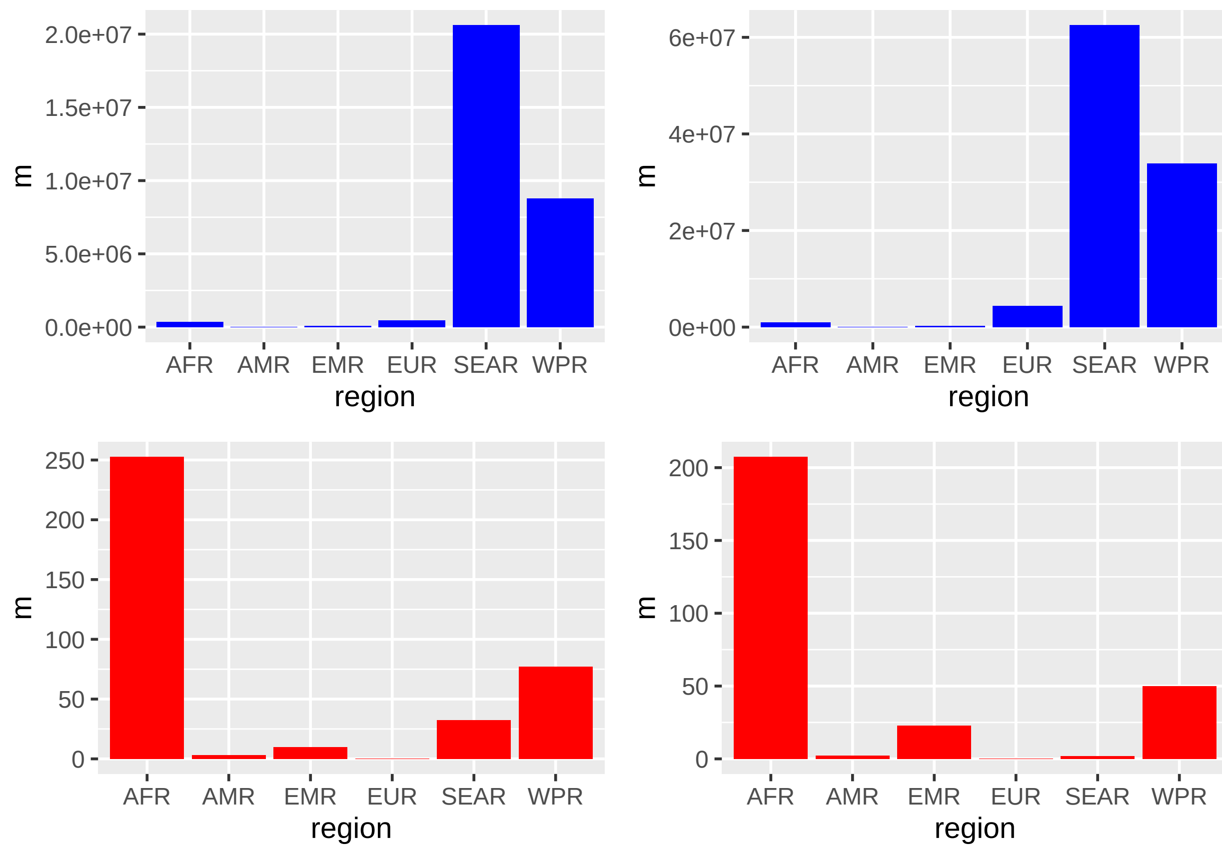

```
region_CHE_2018 <- master_df %>%
  filter(year == 2018)%>%
  group_by(region)%>%
  summarize(m=mean(CurrentHealthExp,na.rm=TRUE))
#write.csv(region_CHE_2018,'analysis3.csv')
g_che_2018=ggplot(data=region_CHE_2018,aes(x=region,y=m,group=1))+geom_bar(stat='identity',fill='blue')
g_che_2018
```




```
region_malaria_2018 <- master_df %>%
  filter(year == 2018)%>%
  group_by(region)%>%
  summarize(m=mean(malaria_cases,na.rm=TRUE))
write.csv(region_malaria_2018,'analysis4.csv')
g_mal_2018=ggplot(data=region_malaria_2018,aes(x=region,y=m,group=1))+geom_bar(stat='identity',fill='re
g_mal_2018
```




```
figure<-ggarrange(g_che_2008,g_che_2018,g_mal_2008,g_mal_2018,
                  ncol = 2, nrow = 2)
figure
```

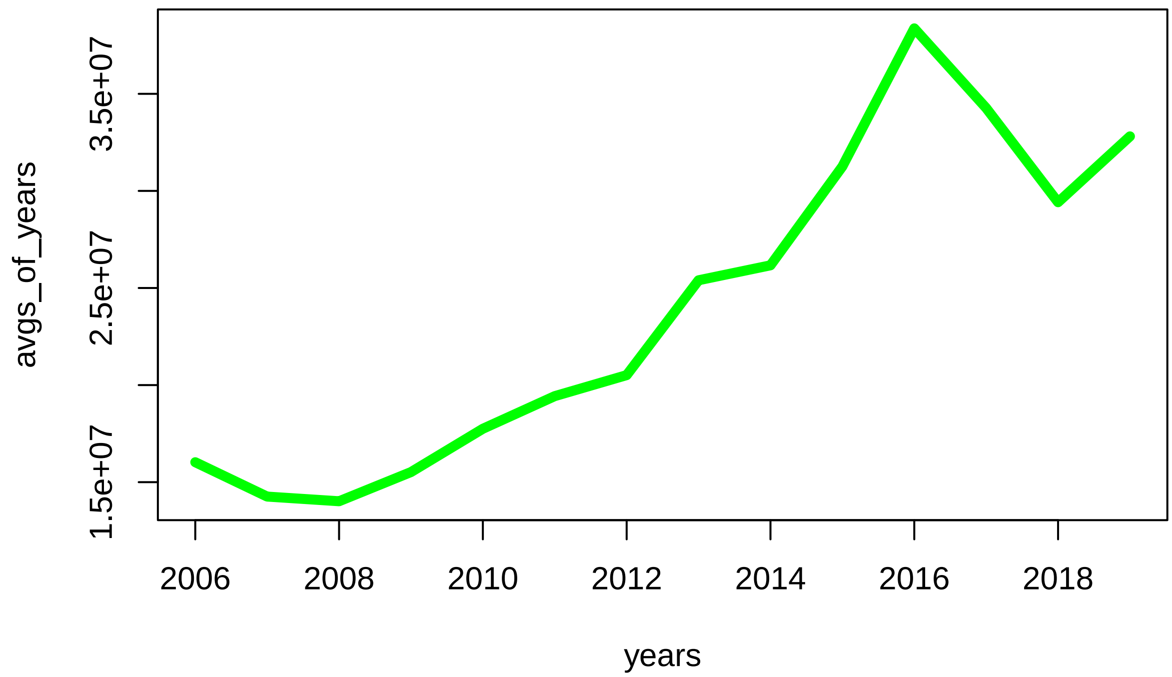


4 important conclusions:-

1- Significant increase in health expenditure led to significant decrease in malaria cases.This is evident from the SEA and WPR regions. 2-Poor investment in health expenditure led to significant increase in malaria cases.African countries have highest number of cases than any other region. 3-Significant increase in health expenditure did not affect the malaria cases.This is evident from American region.Maybe malaria was not a disease of concern for them. 4-No data for Europe meant European regions were Malaria free.On Googling it ,we found that Europe was malaria free in 2015.

How have vaccination rates changed over time?

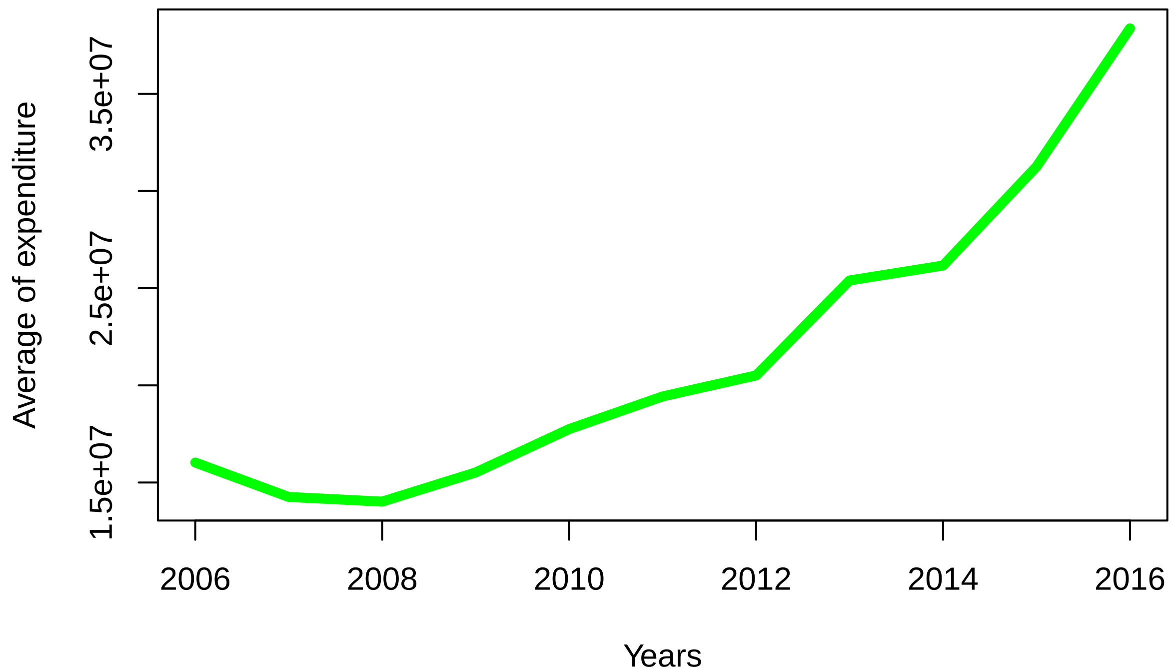
```
#Preparing data and filtering out Low and Low-Mid Countries.


im <- immu_data%>%filter(income == 'Low' |income == 'Low-Mid')
avgs_of_years = colMeans(im[5:18],na.rm = TRUE)
avgs_of_years_2016 = colMeans(im[5:15],na.rm = TRUE)
avgs_of_years = c(avgs_of_years)
years = c(2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019)
plot_of_avg <- plot(years,avgs_of_years, type = 'l', col="green", lwd=5)
```




```
typeof(avgs_of_years)
```

```
## [1] "double"
```

```
years_2016 = c(2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016)
plot_of_avg <- plot(years_2016,avgs_of_years_2016, type = 'l',xlab = "Years",ylab = "Average of expendi
```

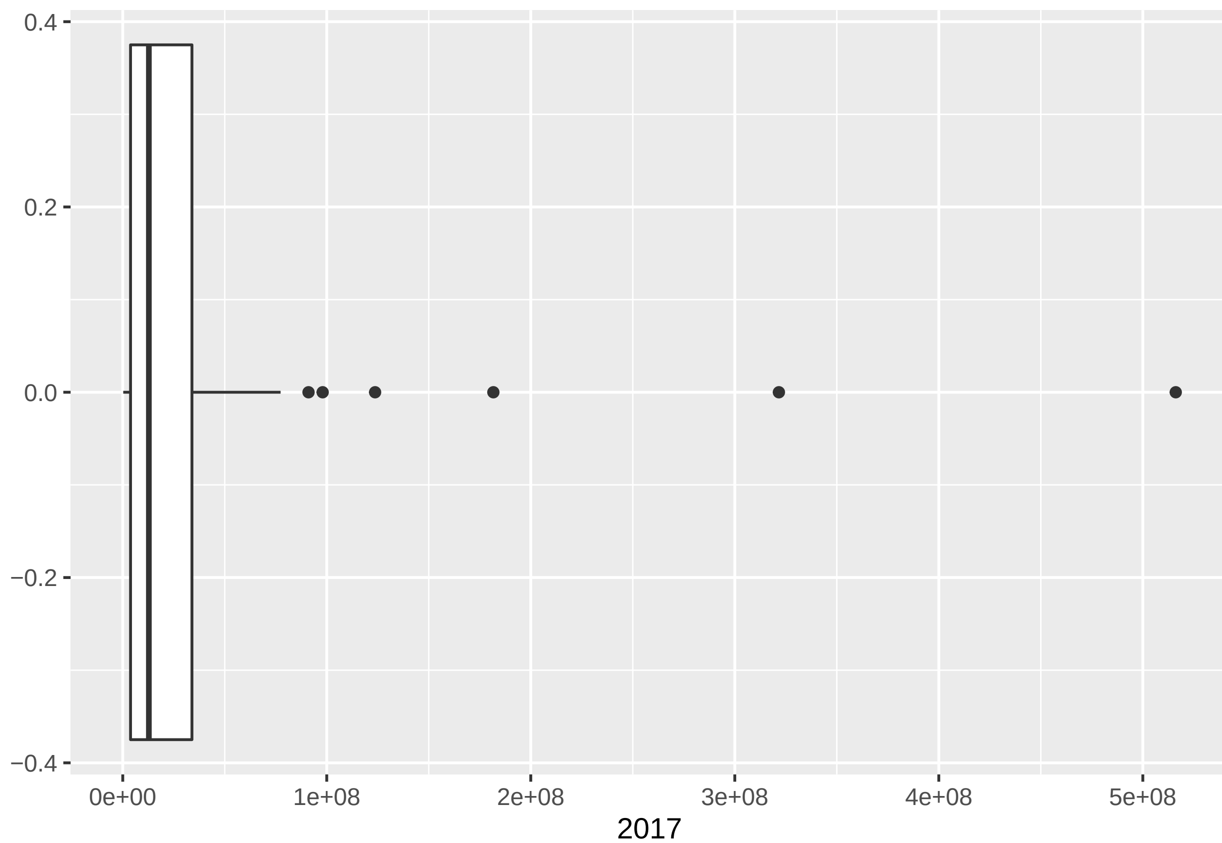



```r
avgs_of_years_2016 = c(avgs_of_years_2016)
typeof(avgs_of_years_2016)
```

```
## [1] "double"
```

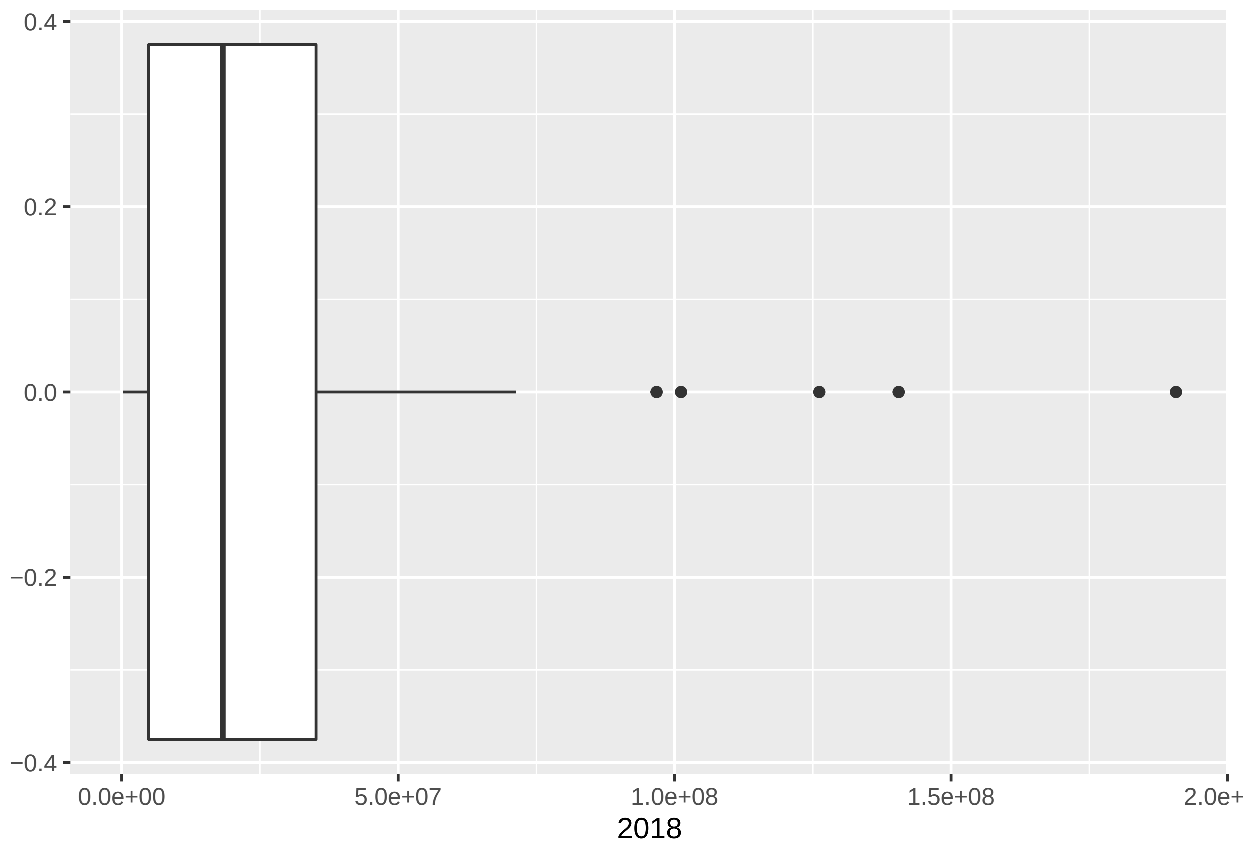
```r
#BOXPLOT FOR 2017
plot_2017 <- ggplot(data = im)+geom_boxplot(aes(x=`2017`))
plot_2017
```

```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```




```r
#BOXPLOT FOR 2018
plot_2018 <- ggplot(data = im)+geom_boxplot(aes(x=`2018`))
plot_2018
```

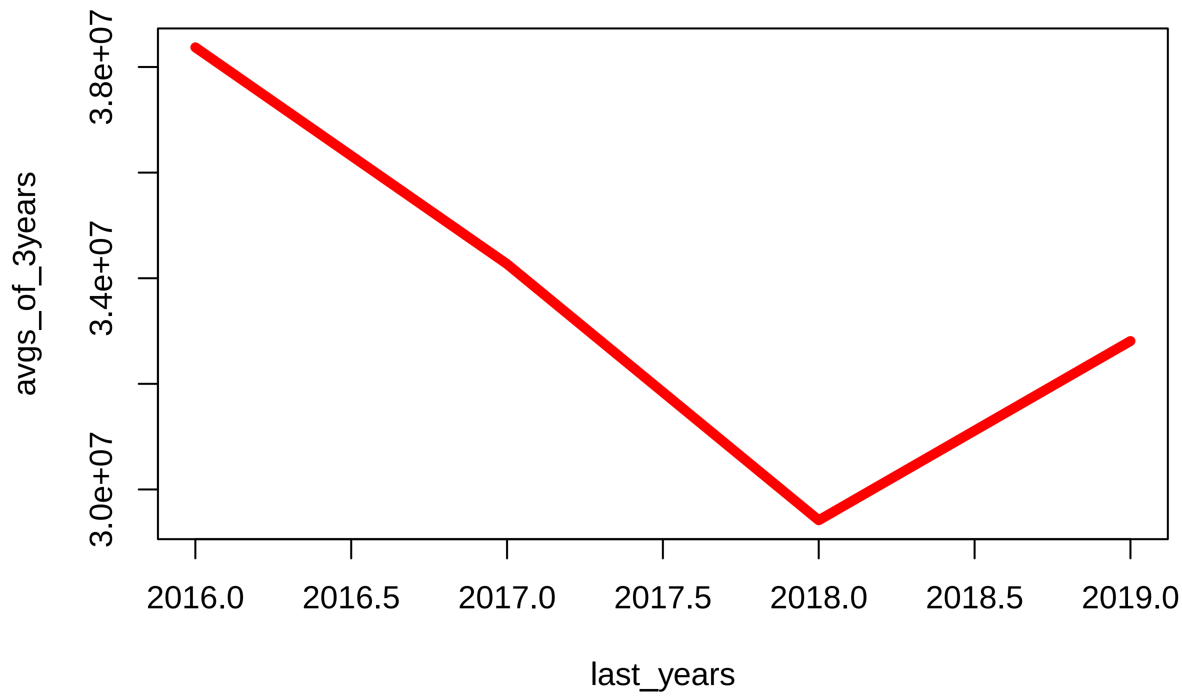
```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).
```




```r
# LINE PLOT FOR CLOSER LOOK
avgs_of_3years = colMeans(im[15:18],na.rm = TRUE)
avgs_of_3years = c(avgs_of_3years)
typeof(avgs_of_3years)
```

```
## [1] "double"
```

```r
last_years = c(2016,2017,2018,2019)
plot_of_avg <- plot(last_years,avgs_of_3years, type = 'l', col="red", lwd=5)
```




```
plot_of_avg
```

```
## NULL
```

```
#WHAT HAPPEND from 2016 to 2018 , region wise divison
region_2016<-im%>%
  group_by(Region)%>%
  summarize(mean(`2016`,na.rm = TRUE))
region_2016
```

```
## # A tibble: 6 x 2
##   Region `mean(\`2016\`, na.rm = TRUE)`
##   <chr>                            <dbl>
## 1 AFRO                         42080055.
## 2 AMRO                         14450317.
## 3 EMRO                         41413629.
## 4 EURO                         10372581.
## 5 SEARO                        70283948.
## 6 WPRO                         17119609.
```

```
region_2017<-im%>%
  group_by(Region)%>%
  summarize(mean(`2017`,na.rm = TRUE))
region_2017
```

```
## # A tibble: 6 x 2
##   Region `mean(\`2017\`, na.rm = TRUE)`
##   <chr>                           <dbl>
## 1 AFRO                        29242542.
## 2 AMRO                        15303204.
## 3 EMRO                        41968748.
## 4 EURO                        10162321.
## 5 SEARO                      106264622.
## 6 WPRO                        13487254.
```

```
region_2018<-im%>%
  group_by(Region)%>%
  summarize(mean(`2018`,na.rm = TRUE))
region_2018
```

```
## # A tibble: 6 x 2
##   Region `mean(\`2018\`, na.rm = TRUE)`
##   <chr>                           <dbl>
## 1 AFRO                        25786874.
## 2 AMRO                        15522892.
## 3 EMRO                        57965483.
## 4 EURO                        11902644.
## 5 SEARO                       30216798.
## 6 WPRO                        35552810.
```

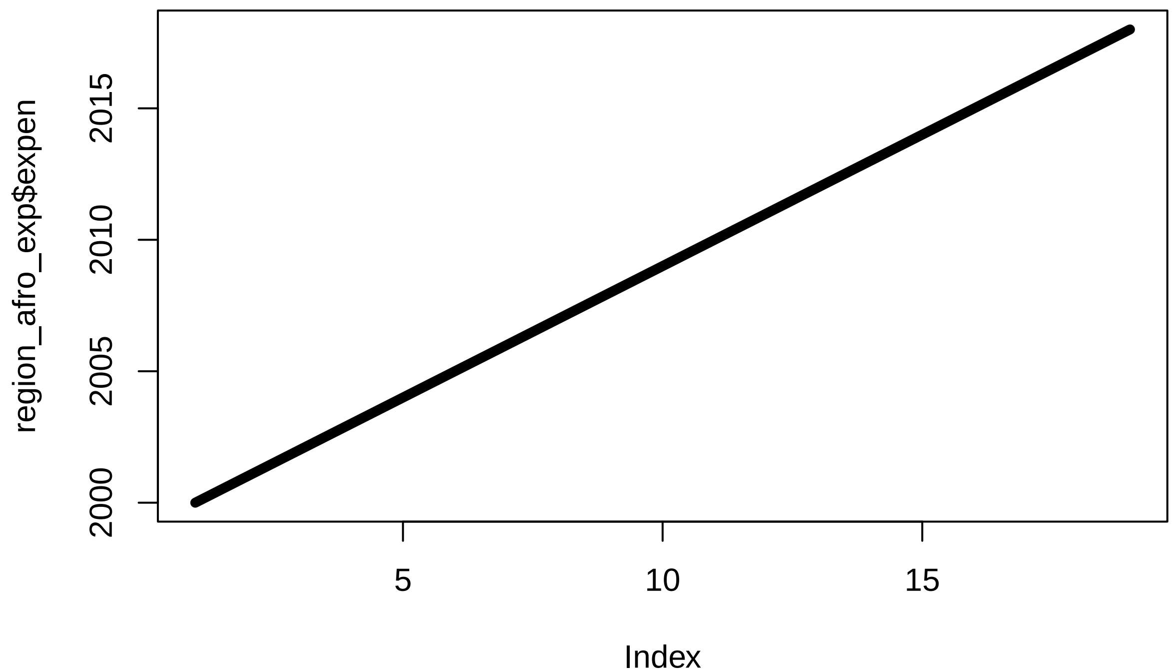
```
#From the values we can clearly see that there's a drastic drop in AFRO region and  SEARO region has in
#Health exp trend in just AFR region.
master_df$expen = master_df$che_gdp*master_df$gdp
region_afro_exp <- master_df%>%
  filter(region == 'AFR')%>%
  group_by(year)%>%
  summarize(mean(expen,na.rm = TRUE))
region_afro_exp
```

```
## # A tibble: 19 x 2
##    year  `mean(expen, na.rm = TRUE)`
##    <chr>                        <dbl>
##  1 2000                     10035395.
##  2 2001                     11436379.
##  3 2002                     13009918.
##  4 2003                     15815483.
##  5 2004                     19460709.
##  6 2005                     23324744.
##  7 2006                     28249734.
##  8 2007                     31191054.
##  9 2008                     34892450.
## 10 2009                     39230690.
## 11 2010                     44741387.
## 12 2011                     50389180.
## 13 2012                     54333358.
## 14 2013                     59639711.
## 15 2014                     71114295.
## 16 2015                     77214055.
```

```
## 17 2016                    83968673.
## 18 2017                    88912618.
## 19 2018                    97372941.
```

```
plot(region_afro_exp$year,region_afro_exp$expen, type = 'l', lwd = 5)
```

```
## Warning: Unknown or uninitialised column: ‘expen‘.
```




```
#Top five countries in expenditure
im_678 <- im%>%
  select(`2006`,`2007`,`2008`)
im$min <- apply(im_678,1,FUN=min)
im_1516 <- im%>%
  select(`2015`,`2016`)
im$max <- apply(im_678,1,FUN=max)
im$per_change <- (im$max - im$min)/im$min
im$per_change
```

```
##  [1] 1.27911492         NA         NA         NA 0.60764015 0.48105939
##  [7] 0.81760553 0.52026399         NA 0.29386926 3.34219111         NA
## [13] 0.73424123 3.46616541         NA 0.28211137         NA         NA
## [19]         NA         NA         NA 3.35551478         NA         NA
## [25] 1.71328804         NA         NA 0.72918134         NA         NA
## [31] 0.70342890         NA 0.70447775         NA         NA 2.42342134
```

```
## [37]         NA 1.11103943         NA         NA         NA         NA
## [43]         NA         NA         NA         NA 0.55071240 0.84840854
## [49] 2.66524705         NA         NA         NA 0.22010577 0.46372666
## [55]         NA         NA         NA 0.22808666         NA         NA
## [61]         NA 1.19951303         NA 0.72635503         NA         NA
## [67]         NA 0.56961135 1.11036295 0.44197931 0.43453993 1.08906863
## [73] 0.01546045 1.45016372 1.16698044         NA         NA
```

```
 im <- im[order(im$per_change, decreasing = TRUE), ]
im
```

```
## # A tibble: 77 x 21
##    `ISO Code` Country  Region income `2006` `2007` `2008` `2009`  `2010`  `2011`
##    <chr>      <chr>    <chr>  <chr>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
##  1 COM        Comoros  AFRO   Low-M~ 3.52e5 1.17e6 1.57e6 1.21e6  1.25e6  9.98e5
##  2 GIN        Guinea   AFRO   Low    4.24e6 1.10e6 9.74e5 1.14e6  9.08e6  1.70e7
##  3 CMR        Cameroon AFRO   Low-M~ 6.92e6 2.95e7 3.00e7 2.10e7  3.08e7  2.62e7
##  4 PAK        Pakistan EMRO   Low-M~ 3.36e7 1.23e8 1.03e8 8.41e7  4.21e7  7.48e7
##  5 MAR        Morocco  EMRO   Low-M~ 5.15e6 9.80e6 1.76e7 1.41e7 NA      NA
##  6 HND        Honduras AMRO   Low-M~ 8.21e6 1.46e7 2.23e7 1.50e7  1.54e7  1.62e7
##  7 VUT        Vanuatu  WPRO   Low-M~ 4.52e4 9.76e4 1.11e5 1.11e5  2   e5 NA
##  8 AFG        Afghani~ EMRO   Low    4   e7 1.93e7 1.76e7 2.23e7  2.52e7  2.49e7
##  9 SWZ        Eswatini AFRO   Low-M~ 1.10e6 5.02e5 8.46e5 1.53e6  2.21e6  2.81e6
## 10 YEM        Yemen    EMRO   Low    6.08e6 1.23e7 1.32e7 1.54e7  1.67e7  2.52e7
## # ... with 67 more rows, and 11 more variables: 2012 <dbl>, 2013 <dbl>,
## #   2014 <dbl>, 2015 <dbl>, 2016 <dbl>, 2017 <dbl>, 2018 <dbl>, 2019 <dbl>,
## #   min <dbl>, max <dbl>, per_change <dbl>
```

```
# so the top five countries are per change
im%>%
  select("Country",'Region', "per_change")
```

```
## # A tibble: 77 x 3
##    Country     Region per_change
##    <chr>       <chr>       <dbl>
##  1 Comoros     AFRO         3.47
##  2 Guinea      AFRO         3.36
##  3 Cameroon    AFRO         3.34
##  4 Pakistan    EMRO         2.67
##  5 Morocco     EMRO         2.42
##  6 Honduras    AMRO         1.71
##  7 Vanuatu     WPRO         1.45
##  8 Afghanistan EMRO         1.28
##  9 Eswatini    AFRO         1.20
## 10 Yemen       EMRO         1.17
## # ... with 67 more rows
```

```
#so the top five countries are in 2016
im <- im[order(im$`2016`, decreasing = TRUE), ]
im
```

```
## # A tibble: 77 x 21
```

```
##    `ISO Code` Country     Region income    `2006` `2007` `2008` `2009` `2010`
##    <chr>      <chr>       <chr>  <chr>      <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 NGA        Nigeria     AFRO   Low-Mid   1.07e8 NA     NA     NA     2.73e7
##  2 IND        India       SEARO  Low-Mid   9.86e7 6.42e7 5.70e7 7.23e7 9.84e7
##  3 PAK        Pakistan    EMRO   Low-Mid   3.36e7 1.23e8 1.03e8 8.41e7 4.21e7
##  4 ETH        Ethiopia    AFRO   Low      NA     NA     NA     NA     8.21e7
##  5 BGD        Bangladesh  SEARO  Low-Mid   2.30e7 2.74e7 3.41e7 6.05e7 6.80e7
##  6 COD        Democratic~ AFRO   Low      NA     NA     NA     NA     3.80e6
##  7 UGA        Uganda      AFRO   Low       2.69e7 2.02e7 1.87e7 1.72e7 1.39e7
##  8 IDN        Indonesia   SEARO  Low-Mid  NA      3.92e7 3.11e7 5.48e7 6.09e7
##  9 PHL        Philippines WPRO   Low-Mid  NA     NA     NA     NA     5.32e7
## 10 TZA        Tanzania    AFRO   Low       7   e6 1.09e7 1.48e7 2.84e7 4.37e7
## # ... with 67 more rows, and 12 more variables: 2011 <dbl>, 2012 <dbl>,
## #   2013 <dbl>, 2014 <dbl>, 2015 <dbl>, 2016 <dbl>, 2017 <dbl>, 2018 <dbl>,
## #   2019 <dbl>, min <dbl>, max <dbl>, per_change <dbl>
```

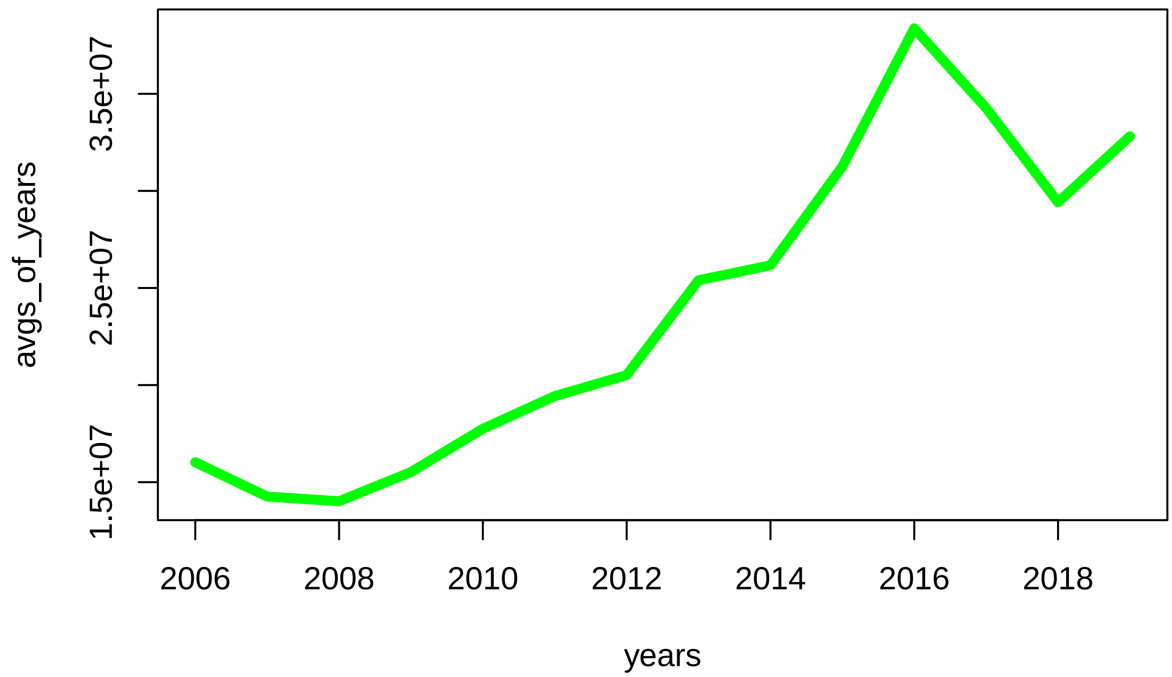
```
im%>%
  select("Country",'Region',`2016`)
```

```
## # A tibble: 77 x 3
##    Country                          Region     `2016`
##    <chr>                            <chr>       <dbl>
##  1 Nigeria                          AFRO   639150767
##  2 India                            SEARO  254043035.
##  3 Pakistan                         EMRO   129319797.
##  4 Ethiopia                         AFRO   121270813
##  5 Bangladesh                       SEARO  115183628
##  6 Democratic Republic of the Congo AFRO   102927251
##  7 Uganda                           AFRO    93658753
##  8 Indonesia                        SEARO   89146268.
##  9 Philippines                      WPRO    82118272.
## 10 Tanzania                         AFRO    72740226
## # ... with 67 more rows
```

```
im <- immu_data%>%
  filter(income == 'Low' | income == 'Low-Mid')
avgs_of_years = colMeans(im[5:18],na.rm = TRUE)
avgs_of_years_2016 = colMeans(im[5:15],na.rm = TRUE)
avgs_of_years = c(avgs_of_years)
write.csv(avgs_of_years,'analysis_till_2019.csv')
typeof(avgs_of_years)
```

```
## [1] "double"
```

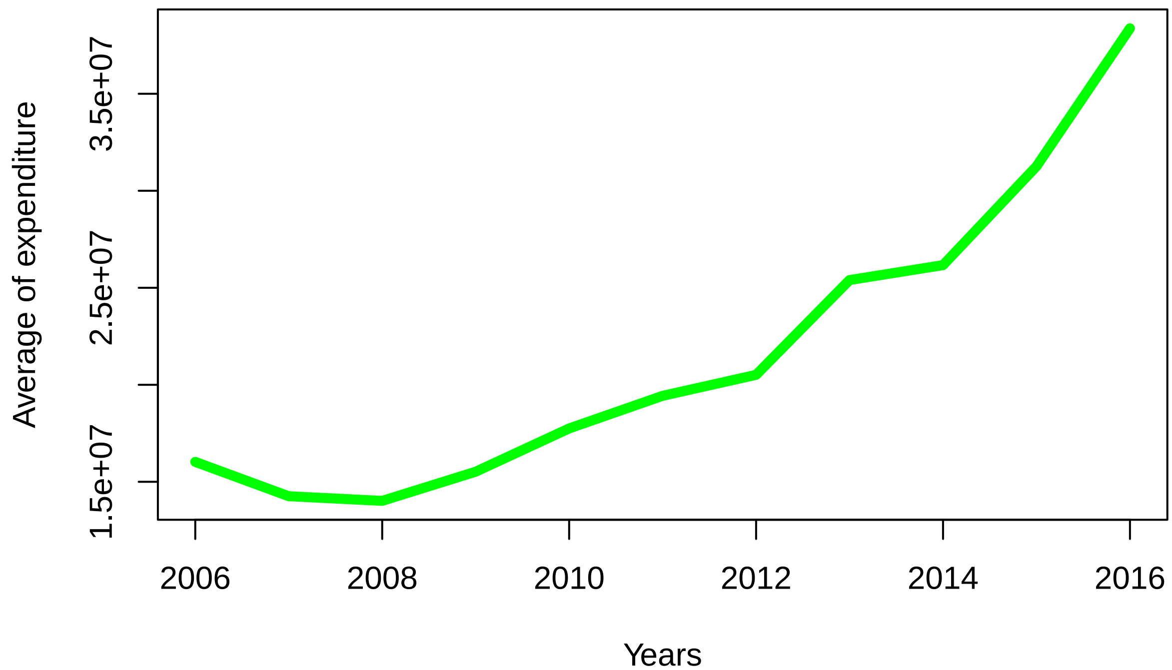
```
years = c(2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017,2018,2019)
plot_of_avg <- plot(years,avgs_of_years, type = 'l', col="green", lwd=5)
```




```
typeof(avgs_of_years)
```

```
## [1] "double"
```

```
years_2016 = c(2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016)
plot_of_avg <- plot(years_2016,avgs_of_years_2016, type = 'l',xlab = "Years",ylab = "Average of expendi
```



Health Expenditure on Vaccination rates increased by 141.6% from 2006 to 2016.

What is the financial impact?

Let us consider the financial impact of the WHO expenditure on the society of low mid and low income groups The first plot clearly shows the upward trend in Voluntary health expenses and downward trend in the Out of Pocket Expenditure

```
library(tidyverse)
library(readxl)
library("ggplot2")
theme_set(theme_bw())
library("sf")
```

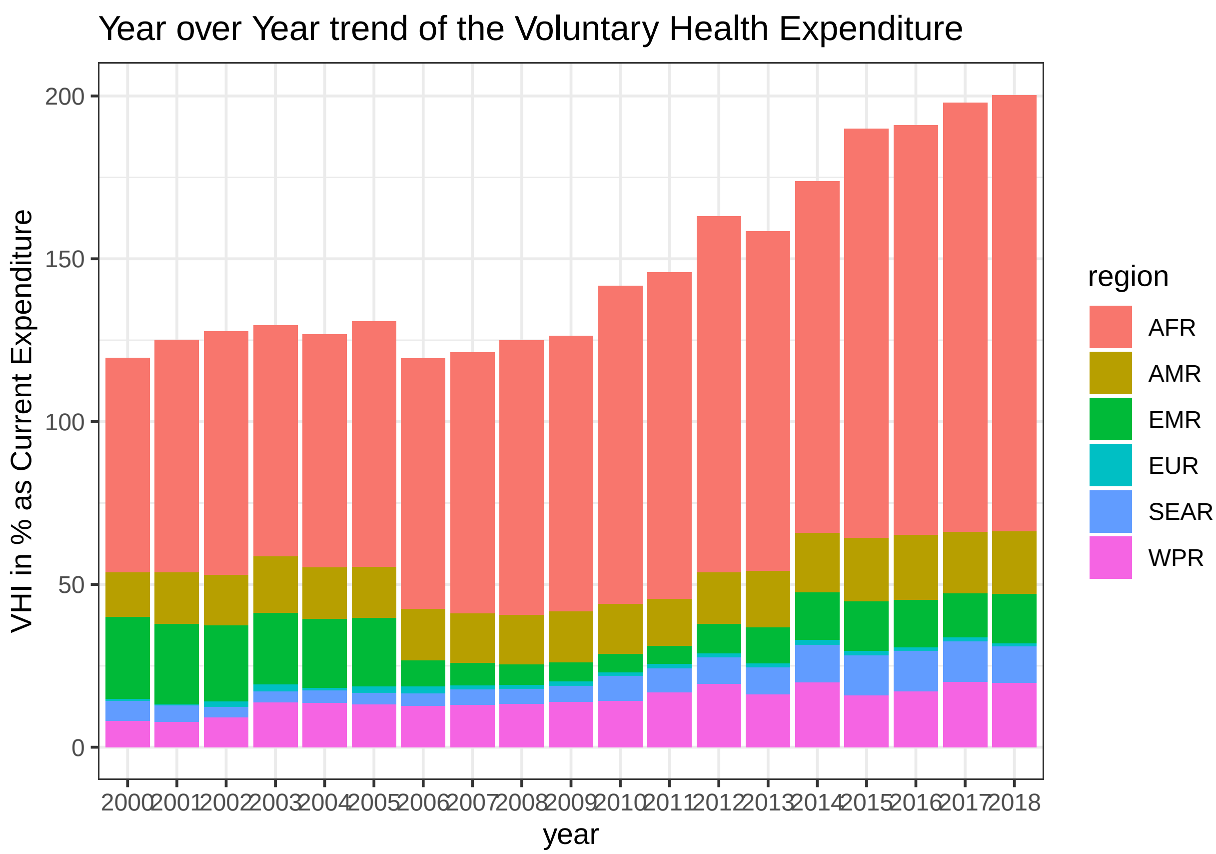
```
## Linking to GEOS 3.9.0, GDAL 3.2.1, PROJ 7.2.1
```

```
#Load the main WHO data
who_url = "https://apps.who.int//nha//database//Home//IndicatorsDownload//en"
who_download<- download.file(url = who_url, destfile = "data_Who.xlsx", mode="wb")
master_data <- read_excel("data_Who.xlsx")
#Keep the relevant data

head(newdata)
```

```
## # A tibble: 6 x 12
##   country year  income_group region phc_che phc_usd_pc     gdp    pop che_gdp
##   <chr>   <chr> <chr>        <chr>    <dbl>      <dbl>   <dbl>  <dbl>   <dbl>
## 1 Algeria 2000  Up-Mid       AFR         NA         NA 4123500 31042.    3.49
## 2 Algeria 2001  Up-Mid       AFR         NA         NA 4227100 31452.    3.84
## 3 Algeria 2002  Up-Mid       AFR         NA         NA 4522800 31855.    3.73
## 4 Algeria 2003  Up-Mid       AFR         NA         NA 5252300 32264.    3.60
## 5 Algeria 2004  Up-Mid       AFR         NA         NA 6149100 32692.    3.54
## 6 Algeria 2005  Up-Mid       AFR         NA         NA 7562000 33150.    3.24
## # ... with 3 more variables: che_pc_usd <dbl>, vhi_che <dbl>, oops_che <dbl>
```

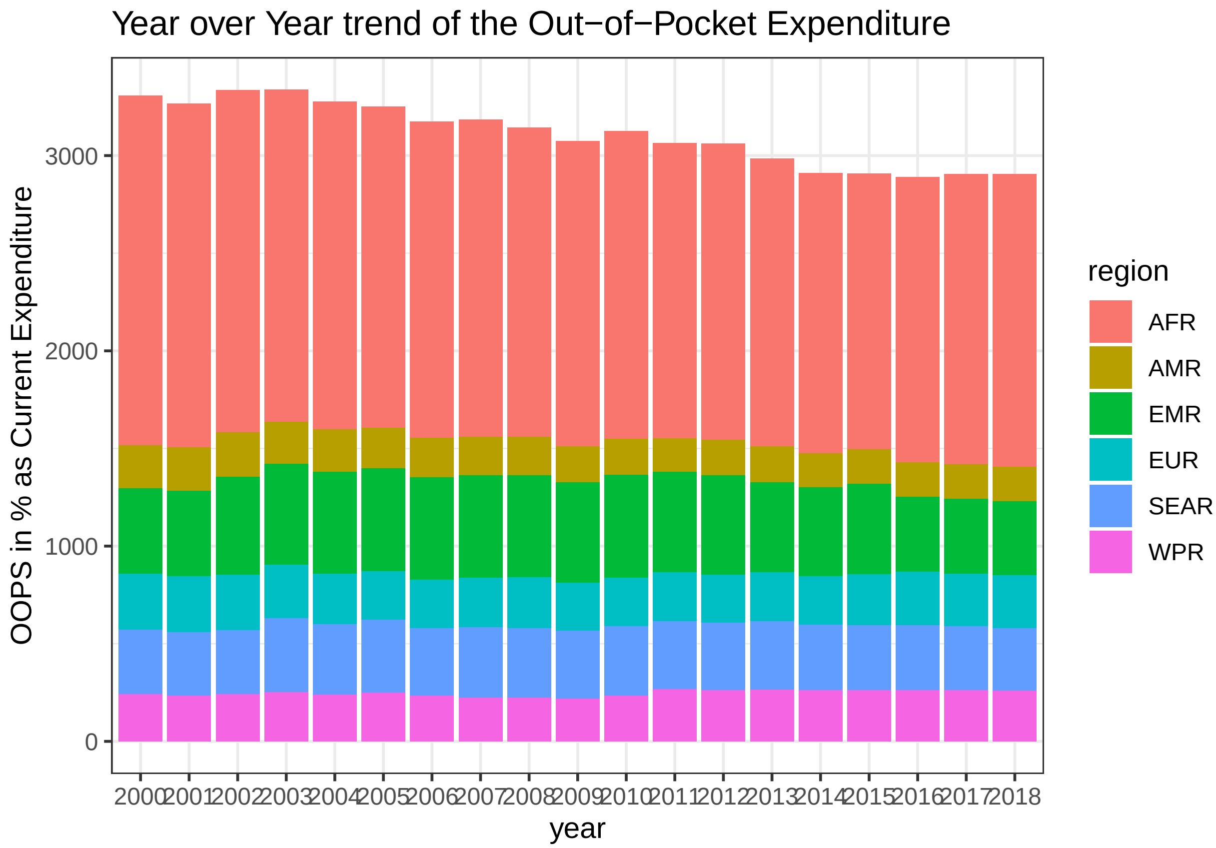
```
# Financial impact of such schemes on individuals
master_df <- filter(newdata, income_group == "Low" | income_group == "Low-Mid")

# trends and observations in the out of pocket expenses by an individual


master_df_region <- master_df %>%
  group_by(region) %>%
  summarise(mean_vhi = mean(vhi_che, na.omit = TRUE),
            mean_oops = mean(oops_che, na.omit = TRUE))

#View(master_df1)
# Low income and low mid - Europe
# Clearly individuals in European countries have more out of pocket expenses, and so they are not makin
# they are not only offered to the individuals under govt jobs but also are additional benefits other t

master_df_country <- master_df %>%
                  group_by(country) %>%
                  summarise(mean_vhi = mean(vhi_che, na.omit = TRUE),
                  mean_oops = mean(oops_che, na.omit = TRUE))

# it is visible also at the country level

# Visuals to cross verify the numbers

voluntary_expense_trend <- ggplot(data = master_df, aes(x=year, y=vhi_che))+
  geom_bar(stat="identity",(aes(fill=region))) +
  ggtitle("Year over Year trend of the Voluntary Health Expenditure") +
  xlab("year") +
  ylab("VHI in % as Current Expenditure")
voluntary_expense_trend
```

```
## Warning: Removed 41 rows containing missing values (position_stack).
```




```
out_of_expense_trend <- ggplot(data = master_df, aes(x=year, y=oops_che))+
  geom_bar(stat="identity",(aes(fill=region))) +
  ggtitle("Year over Year trend of the Out-of-Pocket Expenditure") +
  xlab("year") +
  ylab("OOPS in % as Current Expenditure")
out_of_expense_trend
```

```
## Warning: Removed 41 rows containing missing values (position_stack).
```



This is an immense improvement in the healthcare investment by the government and less out of pocket expenses for an individual —

Biases in our analysis

We can see 2 different kinds of biases: biases from data sources & collection and human biases Biases from data sources & collection 1.Data unavailability - This data consists of a lot of null values, missing values and NA values. Data for some of the fiscal years are missing, so we had to break the dataset in order to showcase the continuous trends.

2. Inaccuracies in data collection – Any and every inaccurate data point could result in erroneous results.
3. Restricted time frame – We have chosen 2000 to 2018 and this could possibly be a source of bias

Human biases –

1. Confirmation bias – bias induced by existing inherent beliefs

2.Dunning Kruger effect – This bias leads people to view an idea or event as simplistic because they don't have a lot of information on the subject. While under the influence of Dunning-Kruger, people overestimate their knowledge of something and it prevents them from being curious and seeking out information

3.Cultural bias – Bias introduced by cultural differences. For example, hygiene isn't important to country X. Initially, there were assumptions regarding the less focus on AFR countries. From the inference, we see that they have been doing well in healthcare sector

Conclusion

We would like to conclude our analysis with a relevant quote -

"*People often call me an optimist, because I show them the enormous progress they didn't know about. That makes me angry. I'm not an optimist. That makes me sound naive. I'm a very serious "possibility". That's something I made up. It means someone who neither hopes without reason, nor fears without reason, someone who constantly resists the over dramatic worldview. As a possibility, I see all this progress, and it fills me with conviction and hope that further progress is possible. This is not optimistic. It is having a clear and reasonable idea about how things are. It is having a worldview that is constructive and useful*"

Our analysis shows that things are genuinely getting better in global health. There are a lot of challenges facing low and low-mid countries today, but this analysis is an acknowledgement of their efforts in improving their health in general. Some of these are -

1. There has been a 200% increment in health expenditure
2. Life expectancy has increased to 65 from 58
3. Healthcare has become more affordable as compared to 2000
4. Money spent on immunizations has increased by 141.6%
5. Europe has eliminated Malaria
6. Significant reduction in Tuberculosis & Malaria cases