Third Assignment

Web scraping and data cleanig

As we said in the previous assignment, we chose the country of Benin to workon our project. In particular we would like to adrees a specific question: **What impact does decentralization have on the performance of the health provision system in Benin?**

To do so,we need both data for the country and for the single municipalities to better analyze the situation and to find an appropriate answer to our main question. Thus, our first step is to web scrape data from the WB and the Who on Benin as a country, related to public health expenditure, health expenditure per capita, improved sanitation facilities, density of health infrastructures per 100000 population. In fact, as we alredy mentioned, decentralization in Benin only begun in the 2000s with establishment of 77 municipalities and transfer of competencies from central to local governments.Consequently, let's see the trend of the last years concerning this country on behalf of health related variables.

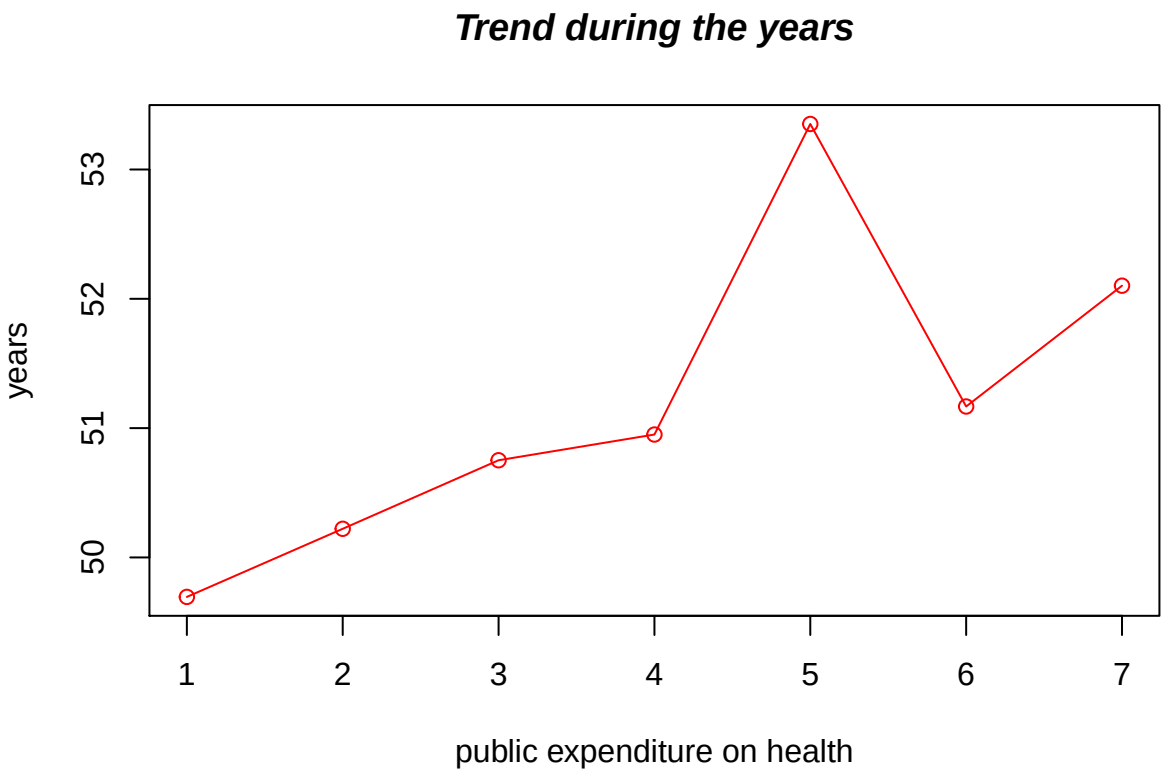
Firstly, we need to load the dataset and eliminate the column we do not need.

```
library(WDI)
```

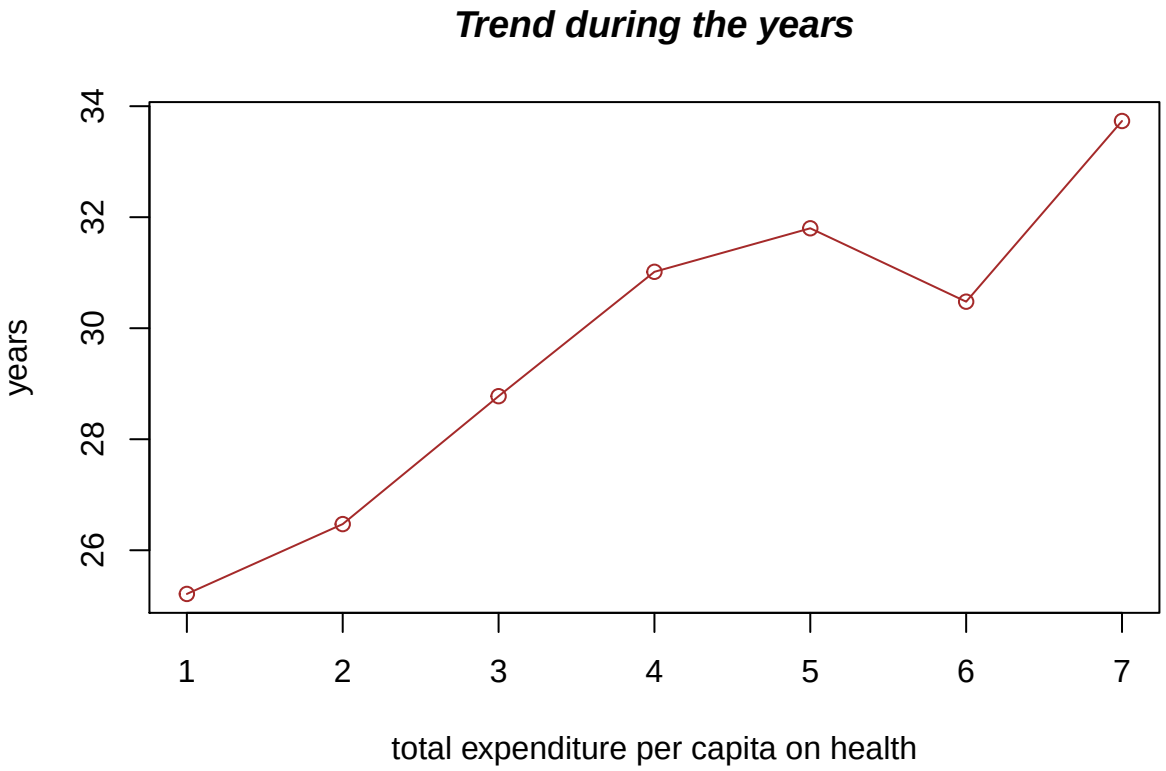
```
## Loading required package: RJSONIO
```

```
healthpublicexpend<-WDI(indicator= c('SH.XPD.PUBL', 'SH.XPD.PCAP','SH.STA.ACSN'))
Benin<-subset(healthpublicexpend,country=='Benin')
Benin$iso2c<-NULL


plot(Benin$SH.XPD.PUBL, type="o", col="red", Benin$Year, xlab="public expenditure on health", ylab="yea
title(main="Trend during the years", col.main="Black", font.main=4)
```



```r
plot(Benin$SH.XPD.PCAP, type="o", col="brown", Benin$Year, xlab="total expenditure per capita on health
title(main="Trend during the years", col.main="Black", font.main=4)
```



This dataset includes these variables over a seven year timespan (2005:2011):

(i) **health expenditure per capita (SH.XPD.PCAP)**, meaninig the sum of public and private health expenditures as a ratio of total population. It covers the provision of health services (preventive and curative), family planning activities, nutrition activities, and emergency aid designated for health but does not include provision of water and sanitation. Data are in current U.S. dollars.

(ii) **public health expenditure (SH.XPD.PUBL)**,which consists of recurrent and capital spending from government (central and local) budgets, external borrowings and grants (including donations from international agencies and nongovernmental organizations), and social (or compulsory) health insurance funds.

(iii) **improved sanitation facilities(SH.STA.ACSN)**, which actually refers to the access to improved sanitation facilities as the percentage of the population using improved sanitation facilities. The indicator includes flush/pour flush (to piped sewer system, septic tank, pit latrine), ventilated improved pit (VIP) latrine, pit latrine with slab, and composting toilet.

The second dataset is taken from the WHO indicators, in particular, from the Global Health Obervatory Data Repository. What we want is the dataset that refers to the density of health infrastructures per 100000 population, including:

(i) Health posts, that are either community centres or health environments with a very limited number of beds with limited curative and preventive care resources normally assisted by health workers or nurses,

(ii) Health centers,which includes the number of health centres from the public and private sectors, per 100,000 population

(iii) Number of district/rural hospitals from the public and private sectors, per 100,000 population,

(iv) Number of provincial hospitals from the public and private sectors, per 100,000 population,

(v) Number of specialized hospitals delivering mainly tertiary care from the public and private sectors, per 100,000 population. These specialized hospitals could be regional, specialized, research hospitals or Federal/National Institutes.

(vi) Number of specialized hospitals delivering mainly tertiary care from the public and private sectors, per 100,000 population. These specialized hospitals could be: regional, specialized, research hospitals or Federal/National Institutes.

However, these variables are only know for the years 2010 and 2013. We import this Dataset with an URL from the WHO website, that we call infrastructures.From this dataset we definetly have to extract only the country we are interested in (BEN). From this point, the dataset has to be cleaned of all those columns that we really don't need, since the original one shows also elements such as if it was published, comments,NA elements etc...

```
infrastructures <- read.csv("/var/folders/ln/gccsc1ts04d6l8_0qk_rlhp00000gn/T//RtmphC4lgc/data1b454d0b2
Benininfra<-subset(infrastructures,Country=='Benin')
Benininfra$PUBLISH.STATES<-NULL
Benininfra$WHO.region<-NULL
Benininfra$Low<-NULL
Benininfra$High<-NULL
Benininfra$Comments<-NULL
```

Merging datasets

Once we have our datasets, it is useful to merge them together so as to work easily and quickly with one dataset that would include all the needed variables.

```
names(Benin)[names(Benin)=="country"] <- "Country"
names(Benin)[names(Benin)=="year"] <- "Year"
```