M2.851 - Tipología y ciclo de vida de los datos Práctica 2

Ana Caudevilla

2018-01-08

1. Descripción del dataset

Este dataset busca comprender mejor el estado de salud dental de la población. En este caso se plantea si a los países con mayor riqueza les supone un empeoramiento de la salud dental, porque consumen más dulces, o si al contrario les representa una ventaja, por disponer de mejor asistencia sanitaria o incluso de un mejor nivel educativo.

2. Carga y limpieza de los datos

```
> adultliteracy <-
+   read.table("C:/Dropbox/Formacion/UOC-BD/UOCAsig/TipoCiclo/Bloque3/Bad_Teeth/adultliteracy.csv",
+    header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
```

```
> badteeth <-
+   read.table("C:/Dropbox/Formacion/UOC-BD/UOCAsig/TipoCiclo/Bloque3/Bad_Teeth/badteeth.csv",
+    header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
```

```
> gdp <-
+   read.table("C:/Dropbox/Formacion/UOC-BD/UOCAsig/TipoCiclo/Bloque3/Bad_Teeth/gdp.csv",
+    header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
```

```
> healthexpend <-
+   read.table("C:/Dropbox/Formacion/UOC-BD/UOCAsig/TipoCiclo/Bloque3/Bad_Teeth/healthexpend.csv",
+    header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
```

```
> sugar_comsumption <-
+   read.table("C:/Dropbox/Formacion/UOC-BD/UOCAsig/TipoCiclo/Bloque3/Bad_Teeth/sugar_consumption.csv
+    header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
```

adultliteracy - % adultos nivel educativo

Referencia: *Literacy rate, adult total (% of people ages 15 and above, UNESCO)*

```
> dim(adultliteracy)
```

```
[1] 262  38
```

```
> names(adultliteracy)
```

```
 [1] "Adult..15...literacy.rate......Total"
 [2] "X1975"
 [3] "X1976"
 [4] "X1977"
 [5] "X1978"
 [6] "X1979"
 [7] "X1980"
 [8] "X1981"
 [9] "X1982"
[10] "X1983"
[11] "X1984"
[12] "X1985"
[13] "X1986"
[14] "X1987"
[15] "X1988"
[16] "X1989"
[17] "X1990"
[18] "X1991"
[19] "X1992"
[20] "X1993"
[21] "X1994"
[22] "X1995"
[23] "X1996"
[24] "X1997"
[25] "X1998"
[26] "X1999"
[27] "X2000"
[28] "X2001"
[29] "X2002"
[30] "X2003"
[31] "X2004"
[32] "X2005"
[33] "X2006"
[34] "X2007"
[35] "X2008"
[36] "X2009"
[37] "X2010"
[38] "X2011"
```

```
> head(adultliteracy)
```

```
  Adult..15...literacy.rate......Total X1975 X1976 X1977 X1978    X1979
1                          Afghanistan    NA    NA    NA    NA 18.15768
2                              Albania    NA    NA    NA    NA       NA
3                              Algeria    NA    NA    NA    NA       NA
4                              Andorra    NA    NA    NA    NA       NA
5                               Angola    NA    NA    NA    NA       NA
6                             Anguilla    NA    NA    NA    NA       NA
  X1980 X1981 X1982 X1983   X1984 X1985 X1986    X1987 X1988 X1989 X1990
1    NA    NA    NA    NA      NA    NA    NA       NA    NA    NA    NA
2    NA    NA    NA    NA      NA    NA    NA       NA    NA    NA    NA
3    NA    NA    NA    NA      NA    NA    NA 49.63088    NA    NA    NA
4    NA    NA    NA    NA      NA    NA    NA       NA    NA    NA    NA
5    NA    NA    NA    NA      NA    NA    NA       NA    NA    NA    NA
6    NA    NA    NA    NA 95.4071    NA    NA       NA    NA    NA    NA
  X1991 X1992 X1993 X1994 X1995 X1996 X1997 X1998 X1999 X2000    X2001
1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA       NA
2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA 98.71298
3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA       NA
4    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA       NA
5    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA 67.40542
6    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA       NA
    X2002 X2003 X2004 X2005    X2006 X2007    X2008 X2009 X2010    X2011
1      NA    NA    NA    NA       NA    NA       NA    NA    NA 39.00000
2      NA    NA    NA    NA       NA    NA 95.93864    NA    NA 96.84530
3 69.8735    NA    NA    NA 72.64868    NA       NA    NA    NA       NA
4      NA    NA    NA    NA       NA    NA       NA    NA    NA       NA
5      NA    NA    NA    NA       NA    NA       NA    NA    NA 70.36242
6      NA    NA    NA    NA       NA    NA       NA    NA    NA       NA
```

badteeth - Dientes estropeados por niño

Referencia: *Bad teeth per child (12 yr, WHO)*

```
> dim(badteeth)
```

```
[1] 191   5
```

```
> names(badteeth)
```

```
[1] "NA."   "X2004" "NA..1" "NA..2" "NA..3"
```

```
> head(badteeth)
```

```
                 NA. X2004 NA..1 NA..2 NA..3
1         Afghanistan  2.90    NA    NA    NA
2             Albania  3.02    NA    NA    NA
3             Algeria  2.30    NA    NA    NA
4              Angola  1.70    NA    NA    NA
5            Anguilla  2.50    NA    NA    NA
6 Antigua and Barbuda  0.70    NA    NA    NA
```

gdp - PIB per capita

Referencia: *GDP/capita (US$, inflation-adjusted, World Bank)*

```
> dim(gdp)
```

```
[1] 275  53
```

```
> names(gdp)
```

```
 [1] "Income.per.person..fixed.2000.US.."
 [2] "X1960"
 [3] "X1961"
 [4] "X1962"
 [5] "X1963"
 [6] "X1964"
 [7] "X1965"
 [8] "X1966"
 [9] "X1967"
[10] "X1968"
[11] "X1969"
[12] "X1970"
[13] "X1971"
[14] "X1972"
[15] "X1973"
[16] "X1974"
[17] "X1975"
[18] "X1976"
[19] "X1977"
[20] "X1978"
[21] "X1979"
[22] "X1980"
[23] "X1981"
[24] "X1982"
[25] "X1983"
[26] "X1984"
[27] "X1985"
[28] "X1986"
[29] "X1987"
[30] "X1988"
[31] "X1989"
[32] "X1990"
[33] "X1991"
[34] "X1992"
[35] "X1993"
[36] "X1994"
[37] "X1995"
[38] "X1996"
[39] "X1997"
[40] "X1998"
[41] "X1999"
[42] "X2000"
[43] "X2001"
[44] "X2002"
[45] "X2003"
[46] "X2004"
[47] "X2005"
[48] "X2006"
[49] "X2007"
[50] "X2008"
[51] "X2009"
[52] "X2010"
[53] "X2011"
```

```
> head(gdp)
```

```
  Income.per.person..fixed.2000.US..    X1960    X1961   X1962    X1963
1                           Abkhazia       NA       NA      NA       NA
2                        Afghanistan       NA       NA      NA       NA
3               Akrotiri and Dhekelia       NA       NA      NA       NA
4                            Albania       NA       NA      NA       NA
5                            Algeria 1280.385 1085.415 855.948 1128.416
6                     American Samoa       NA       NA      NA       NA
     X1964    X1965    X1966    X1967    X1968    X1969   X1970    X1971
1       NA       NA       NA       NA       NA       NA      NA       NA
2       NA       NA       NA       NA       NA       NA      NA       NA
3       NA       NA       NA       NA       NA       NA      NA       NA
4       NA       NA       NA       NA       NA       NA      NA       NA
5 1170.324 1215.016 1127.614 1200.558 1291.864 1359.491 1436.13 1235.664
6       NA       NA       NA       NA       NA       NA      NA       NA
     X1972    X1973   X1974    X1975   X1976    X1977    X1978    X1979
1       NA       NA      NA       NA      NA       NA       NA       NA
2       NA       NA      NA       NA      NA       NA       NA       NA
3       NA       NA      NA       NA      NA       NA       NA       NA
4       NA       NA      NA       NA      NA       NA       NA       NA
5 1527.646 1538.306 1603.35 1632.296 1714.07 1747.665 1848.438 1923.291
6       NA       NA      NA       NA      NA       NA       NA       NA
     X1980    X1981    X1982    X1983    X1984    X1985    X1986    X1987
1       NA       NA       NA       NA       NA       NA       NA       NA
2       NA       NA       NA       NA       NA       NA       NA       NA
3       NA       NA       NA       NA       NA       NA       NA       NA
4 1060.685 1099.513 1110.512 1101.336 1065.235 1059.866 1091.561 1054.248
5 1876.076 1869.621 1924.614 1963.365 2008.472 2020.087 1969.764 1902.061
6       NA       NA       NA       NA       NA       NA       NA       NA
     X1988    X1989     X1990     X1991     X1992     X1993     X1994
1       NA       NA        NA        NA        NA        NA        NA
2       NA       NA        NA        NA        NA        NA        NA
3       NA       NA        NA        NA        NA        NA        NA
4 1013.629 1092.475  977.7655  687.9919  643.2858  714.2414  784.5831
5 1833.153 1864.713 1832.7434 1766.6608 1755.9737 1680.3799 1630.3815
6       NA       NA        NA        NA        NA        NA        NA
      X1995     X1996    X1997    X1998    X1999    X2000    X2001
1        NA        NA       NA       NA       NA       NA       NA
2        NA        NA       NA       NA       NA       NA       NA
3        NA        NA       NA       NA       NA       NA       NA
4  899.7829  990.6532  895.561 1013.514 1118.172 1200.137 1281.843
5 1660.0042 1698.3338 1690.238 1750.651 1781.142 1794.405 1814.415
6        NA        NA       NA       NA       NA       NA       NA
     X2002    X2003    X2004    X2005    X2006    X2007    X2008    X2009
1       NA       NA       NA       NA       NA       NA       NA       NA
2       NA       NA       NA       NA       NA       NA       NA       NA
3       NA       NA       NA       NA       NA       NA       NA       NA
4 1313.723 1381.041 1454.023 1525.724 1594.495 1681.614 1804.419 1857.353
5 1871.922 1971.513 2043.136 2115.186 2124.958 2155.485 2173.788 2192.704
6       NA       NA       NA       NA       NA       NA       NA       NA
     X2010    X2011
1       NA       NA
2       NA       NA
3       NA       NA
4 1915.424 1965.707
5 2231.980 2255.225
6       NA       NA
```

healthexpend - gastos en salud por persona

Referencia: *Government health spending per person (US$, WHO)*

```
> dim(healthexpend)
```

```
[1] 265  17
```

```
> names(healthexpend)
```

```
 [1] "Per.capita.government.expenditure.on.health.at.average.exchange.rate..US.."
 [2] "X1995"
 [3] "X1996"
 [4] "X1997"
 [5] "X1998"
 [6] "X1999"
 [7] "X2000"
 [8] "X2001"
 [9] "X2002"
[10] "X2003"
[11] "X2004"
[12] "X2005"
[13] "X2006"
[14] "X2007"
[15] "X2008"
[16] "X2009"
[17] "X2010"
```

```
> head(healthexpend)
```

```
  Per.capita.government.expenditure.on.health.at.average.exchange.rate..US..
1                                                                   Abkhazia
2                                                                Afghanistan
3                                                    Akrotiri and Dhekelia
4                                                                    Albania
5                                                                    Algeria
6                                                             American Samoa
     X1995    X1996    X1997    X1998    X1999    X2000    X2001
1       NA       NA       NA       NA       NA       NA       NA
2       NA       NA       NA       NA       NA       NA       NA
3       NA       NA       NA       NA       NA       NA       NA
4 13.94059 17.06207 14.16477 18.62585 28.13971 27.16051 30.50962
5 46.77146 47.96005 49.73840 48.67055 45.54382 45.91111 52.50942
6       NA       NA       NA       NA       NA       NA       NA
       X2002     X2003    X2004     X2005     X2006      X2007      X2008
1         NA        NA       NA        NA        NA         NA         NA
2  0.8326431  1.250118  1.61416  2.525066  2.813779   3.503426   3.744613
3         NA        NA       NA        NA        NA         NA         NA
4 32.5499020 40.609457 63.93560 71.356600 75.552514  88.762634 109.074284
5 54.0783807 62.637209 63.22940 69.295636 81.679706 108.904747 147.820706
6         NA        NA       NA        NA        NA         NA         NA
       X2009      X2010
1         NA         NA
2   3.908887   4.390408
3         NA         NA
4 106.893745  94.023613
5 143.160577 138.840923
6         NA         NA
```

sugar_comsumption - consumo de azúcar por día y persona

Referencia: *Sugar comsumption per person (g per day, FAO)*

```
> dim(sugar_comsumption)
```

```
[1] 278  46
```

```
> names(sugar_comsumption)
```

```
 [1] "NA."   "X1961" "X1962" "X1963" "X1964" "X1965" "X1966" "X1967"
 [9] "X1968" "X1969" "X1970" "X1971" "X1972" "X1973" "X1974" "X1975"
[17] "X1976" "X1977" "X1978" "X1979" "X1980" "X1981" "X1982" "X1983"
[25] "X1984" "X1985" "X1986" "X1987" "X1988" "X1989" "X1990" "X1991"
[33] "X1992" "X1993" "X1994" "X1995" "X1996" "X1997" "X1998" "X1999"
[41] "X2000" "X2001" "X2002" "X2003" "X2004" "NA..1"
```

```
> head(sugar_comsumption)
```

```
                  NA. X1961 X1962 X1963 X1964 X1965 X1966 X1967 X1968
1             Abkhazia    NA    NA    NA    NA    NA    NA    NA    NA
2           Afghanistan    NA    NA    NA    NA    NA    NA    NA    NA
3 Akrotiri and Dhekelia    NA    NA    NA    NA    NA    NA    NA    NA
4               Albania 30.14 30.14 32.88 35.62 35.62 35.62 38.36 38.36
5               Algeria 46.58 49.32 46.58 49.32 46.58 46.58 49.32 49.32
6        American Samoa    NA    NA    NA    NA    NA    NA    NA    NA
  X1969 X1970 X1971 X1972 X1973 X1974 X1975 X1976 X1977 X1978 X1979 X1980
1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
4 38.36 38.36 41.10 41.10 41.10 43.84 43.84 43.84 41.10 43.84 46.58 46.58
5 46.58 43.84 52.06 49.32 46.58 49.32 63.01 65.75 71.23 71.23 82.19 82.19
6    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
  X1981 X1982 X1983 X1984 X1985 X1986 X1987 X1988 X1989 X1990 X1991 X1992
1    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
2    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
3    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
4 46.58 46.58 46.58 46.58 46.58 49.32 49.32 49.32 49.32 52.06 38.36 52.06
5 82.19 73.97 82.19 82.19 79.45 84.93 93.15 79.45 84.93 82.19 79.45 73.97
6    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
  X1993  X1994 X1995 X1996 X1997 X1998 X1999 X2000 X2001 X2002 X2003 X2004
1    NA     NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
2    NA     NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
3    NA     NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
4 79.45 101.37 54.80 68.49 60.27 60.27 57.53 65.75 68.49 71.23 65.75 65.75
5 76.71  73.97 73.97 73.97 79.45 54.80 60.27 82.19 79.45 82.19 84.93 84.93
6    NA     NA    NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
  NA..1
1    NA
2    NA
3    NA
4    NA
5    NA
6    NA
```

Variables y observaciones

```
> colnames(adultliteracy)[1]
```

```
[1] "Adult..15...literacy.rate......Total"
```

```
> colnames(adultliteracy)[1] <- "Country"
```

```
> colnames(adultliteracy)[1]
```

```
[1] "Country"
```

```
> colnames(badteeth)[1]
```

```
[1] "NA."
```

```
> colnames(badteeth)[1] <- "Country"
```

```
> colnames(gdp)[1]
```

```
[1] "Income.per.person..fixed.2000.US.."
```

```
> colnames(gdp)[1] <- "Country"
```

```
> colnames(healthexpend)[1]
```

```
[1] "Per.capita.government.expenditure.on.health.at.average.exchange.rate..US.."
```

```
> colnames(healthexpend)[1] <- "Country"
```

```
> colnames(healthexpend)[1]
```

```
[1] "Country"
```

```
> colnames(sugar_comsumption)[1]
```

```
[1] "NA."
```

```
> colnames(sugar_comsumption)[1] <- "Country"
```

```
> colnames(sugar_comsumption)[1]
```

```
[1] "Country"
```

```
> names(adultliteracy)
```

```
 [1] "Country" "X1975"   "X1976"   "X1977"   "X1978"   "X1979"   "X1980"
 [8] "X1981"   "X1982"   "X1983"   "X1984"   "X1985"   "X1986"   "X1987"
[15] "X1988"   "X1989"   "X1990"   "X1991"   "X1992"   "X1993"   "X1994"
[22] "X1995"   "X1996"   "X1997"   "X1998"   "X1999"   "X2000"   "X2001"
[29] "X2002"   "X2003"   "X2004"   "X2005"   "X2006"   "X2007"   "X2008"
[36] "X2009"   "X2010"   "X2011"
```

```
> names(badteeth)
```

```
[1] "Country" "X2004"   "NA..1"   "NA..2"   "NA..3"
```

```
> names(gdp)
```

```
 [1] "Country" "X1960"   "X1961"   "X1962"   "X1963"   "X1964"   "X1965"
 [8] "X1966"   "X1967"   "X1968"   "X1969"   "X1970"   "X1971"   "X1972"
[15] "X1973"   "X1974"   "X1975"   "X1976"   "X1977"   "X1978"   "X1979"
[22] "X1980"   "X1981"   "X1982"   "X1983"   "X1984"   "X1985"   "X1986"
[29] "X1987"   "X1988"   "X1989"   "X1990"   "X1991"   "X1992"   "X1993"
[36] "X1994"   "X1995"   "X1996"   "X1997"   "X1998"   "X1999"   "X2000"
[43] "X2001"   "X2002"   "X2003"   "X2004"   "X2005"   "X2006"   "X2007"
[50] "X2008"   "X2009"   "X2010"   "X2011"
```

```
> names(healthexpend)
```

```
 [1] "Country" "X1995"   "X1996"   "X1997"   "X1998"   "X1999"   "X2000"
 [8] "X2001"   "X2002"   "X2003"   "X2004"   "X2005"   "X2006"   "X2007"
[15] "X2008"   "X2009"   "X2010"
```

```
> names(sugar_comsumption)
```

```
 [1] "Country" "X1961"   "X1962"   "X1963"   "X1964"   "X1965"   "X1966"
 [8] "X1967"   "X1968"   "X1969"   "X1970"   "X1971"   "X1972"   "X1973"
[15] "X1974"   "X1975"   "X1976"   "X1977"   "X1978"   "X1979"   "X1980"
[22] "X1981"   "X1982"   "X1983"   "X1984"   "X1985"   "X1986"   "X1987"
[29] "X1988"   "X1989"   "X1990"   "X1991"   "X1992"   "X1993"   "X1994"
[36] "X1995"   "X1996"   "X1997"   "X1998"   "X1999"   "X2000"   "X2001"
[43] "X2002"   "X2003"   "X2004"   "NA..1"
```

```
> colnames(adultliteracy) <- gsub("X","", colnames(adultliteracy))
```

```
> colnames(badteeth) <- gsub("X","", colnames(badteeth))
```

```
> colnames(gdp) <- gsub("X","", colnames(gdp))
```

```
> colnames(healthexpend) <- gsub("X","", colnames(healthexpend))
```

```
> colnames(sugar_comsumption) <- gsub("X","", colnames(sugar_comsumption))
```

El número de campos es variable, los años de las observaciones no coinciden para los distintos datasets:

- adultliteracy: 1975 - 2011
- badteeth: 2004 y dos NAs
- gdp: 1960 - 2011
- healthexpend: 1995 - 2010
- sugar_comsumption: 1961 - 2004 y uno adicional NA

Valores nulos

En la carga de datos, se indicó que los valores de text “NA” se consideraran como la no existencia del dato (en read.csv: na.strings=“NA”)

```
> sapply(adultliteracy, function(x)(sum(is.na(x))))
```

```
Country    1975    1976    1977    1978    1979    1980    1981    1982
      2     256     255     261     260     255     244     244     255
   1983    1984    1985    1986    1987    1988    1989    1990    1991
    260     257     256     258     259     258     252     246     240
   1992    1993    1994    1995    1996    1997    1998    1999    2000
    252     258     250     257     254     259     257     254     219
   2001    2002    2003    2004    2005    2006    2007    2008    2009
    232     243     252     232     243     237     230     235     234
   2010    2011
    227     178
```

```
> sapply(adultliteracy, function(x)(sprintf("%.2f%%",
+   sum(is.na(x))*100/nrow(adultliteracy))))
```

```
 Country     1975     1976     1977     1978     1979     1980     1981
 "0.76%" "97.71%" "97.33%" "99.62%" "99.24%" "97.33%" "93.13%" "93.13%"
    1982     1983     1984     1985     1986     1987     1988     1989
"97.33%" "99.24%" "98.09%" "97.71%" "98.47%" "98.85%" "98.47%" "96.18%"
    1990     1991     1992     1993     1994     1995     1996     1997
"93.89%" "91.60%" "96.18%" "98.47%" "95.42%" "98.09%" "96.95%" "98.85%"
    1998     1999     2000     2001     2002     2003     2004     2005
"98.09%" "96.95%" "83.59%" "88.55%" "92.75%" "96.18%" "88.55%" "92.75%"
    2006     2007     2008     2009     2010     2011
"90.46%" "87.79%" "89.69%" "89.31%" "86.64%" "67.94%"
```

```
> apply(adultliteracy, 1, function(x)(all(is.na(x))))
```

```
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[166] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[188] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[199] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[210] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[221] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[232] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[243] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[254] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
```

```
> adultliteracy[261,]
```

```
    Country 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986
261    <NA>   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
    1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
261   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
    2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011
261   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
```

```
> adultliteracy[262,]
```

```
    Country 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986
262    <NA>   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
    1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
262   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
    2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011
262   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
```

```
> apply(adultliteracy, 2, function(x)(all(is.na(x))))
```

```
Country    1975    1976    1977    1978    1979    1980    1981    1982
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
   1983    1984    1985    1986    1987    1988    1989    1990    1991
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
   1992    1993    1994    1995    1996    1997    1998    1999    2000
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
   2001    2002    2003    2004    2005    2006    2007    2008    2009
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE
   2010    2011
  FALSE   FALSE
```

Hay un porcentaje muy elevado de datos vacios, incluso tenemos dos observaciones con todos los campos nulos. Así, en primer lugar eliminamos esas dos filas completamente nulas:

```
> adultliteracy<-adultliteracy[-c(261, 262),]
```

```
> sapply(badteeth, function(x)(sum(is.na(x))))
```

```
Country    2004   NA..1   NA..2   NA..3
      1       1     191     191     191
```

```
> nrow(badteeth)
```

```
[1] 191
```

```
> apply(badteeth, 2, function(x)(all(is.na(x))))
```

```
Country    2004   NA..1   NA..2   NA..3
  FALSE   FALSE    TRUE    TRUE    TRUE
```

```
> which(is.na(badteeth[,1]))
```

```
[1] 191
```

```
> which(is.na(badteeth[,2]))
```

```
[1] 191
```

Eliminamos la fila y las 3 columnas con todos los datos vacios:

```
> badteeth<-badteeth[-c(191),]
```

```
> badteeth<-badteeth[,-c(3,4,5)]
```

```
> head(badteeth)
```

```
              Country 2004
1         Afghanistan 2.90
2             Albania 3.02
3             Algeria 2.30
4              Angola 1.70
5            Anguilla 2.50
6 Antigua and Barbuda 0.70
```

```
> sapply(gdp, function(x)(sum(is.na(x))))
```

```
Country    1960    1961    1962    1963    1964    1965    1966    1967
      0     179     178     178     178     178     172     170     169
   1968    1969    1970    1971    1972    1973    1974    1975    1976
    168     168     156     156     156     156     155     151     150
   1977    1978    1979    1980    1981    1982    1983    1984    1985
    145     145     144     132     128     125     125     121     118
   1986    1987    1988    1989    1990    1991    1992    1993    1994
    115     111     109     108      94      92      91      90      88
   1995    1996    1997    1998    1999    2000    2001    2002    2003
     84      84      82      81      79      75      79      80      80
   2004    2005    2006    2007    2008    2009    2010    2011
     81      81      82      82      84      86      93     100
```

```
> sapply(gdp, function(x)(sprintf("%.2f%%",sum(is.na(x))*100/nrow(gdp))))
```

```
 Country     1960     1961     1962     1963     1964     1965     1966
 "0.00%" "65.09%" "64.73%" "64.73%" "64.73%" "64.73%" "62.55%" "61.82%"
    1967     1968     1969     1970     1971     1972     1973     1974
"61.45%" "61.09%" "61.09%" "56.73%" "56.73%" "56.73%" "56.73%" "56.36%"
    1975     1976     1977     1978     1979     1980     1981     1982
"54.91%" "54.55%" "52.73%" "52.73%" "52.36%" "48.00%" "46.55%" "45.45%"
    1983     1984     1985     1986     1987     1988     1989     1990
"45.45%" "44.00%" "42.91%" "41.82%" "40.36%" "39.64%" "39.27%" "34.18%"
    1991     1992     1993     1994     1995     1996     1997     1998
"33.45%" "33.09%" "32.73%" "32.00%" "30.55%" "30.55%" "29.82%" "29.45%"
    1999     2000     2001     2002     2003     2004     2005     2006
"28.73%" "27.27%" "28.73%" "29.09%" "29.09%" "29.45%" "29.45%" "29.82%"
    2007     2008     2009     2010     2011
"29.82%" "30.55%" "31.27%" "33.82%" "36.36%"
```

Es un número muy considerable de campos vacios.

```
> sapply(healthexpend, function(x)(sum(is.na(x))))
```

```
Country    1995    1996    1997    1998    1999    2000    2001    2002
      0      76      75      75      74      74      75      74      75
   2003    2004    2005    2006    2007    2008    2009    2010
     75      75      75      75      75      75      75      78
```

```
> sapply(healthexpend, function(x)(sprintf("%.2f%%",
+   sum(is.na(x))*100/nrow(healthexpend))))
```

```
 Country     1995     1996     1997     1998     1999     2000     2001
 "0.00%" "28.68%" "28.30%" "28.30%" "27.92%" "27.92%" "28.30%" "27.92%"
    2002     2003     2004     2005     2006     2007     2008     2009
"28.30%" "28.30%" "28.30%" "28.30%" "28.30%" "28.30%" "28.30%" "28.30%"
    2010
"29.43%"
```

Igualmente un número considerable de valores vacíos.

```
> sapply(sugar_comsumption, function(x)(sum(is.na(x))))
```

```
Country    1961    1962    1963    1964    1965    1966    1967    1968
     19     123     123     123     123     123     123     123     123
   1969    1970    1971    1972    1973    1974    1975    1976    1977
    123     123     123     123     123     123     123     123     123
   1978    1979    1980    1981    1982    1983    1984    1985    1986
    123     123     123     123     123     123     123     123     123
   1987    1988    1989    1990    1991    1992    1993    1994    1995
    123     123     123     123     123     105     104     104     104
   1996    1997    1998    1999    2000    2001    2002    2003    2004
    103     103     103     103     103     103     103     103     103
  NA..1
    278
```

```
> sapply(sugar_comsumption, function(x)(sprintf("%.2f%%",
+   sum(is.na(x))*100/nrow(sugar_comsumption))))
```

```
  Country      1961      1962      1963      1964      1965      1966
  "6.83%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"
     1967      1968      1969      1970      1971      1972      1973
 "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"
     1974      1975      1976      1977      1978      1979      1980
 "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"
     1981      1982      1983      1984      1985      1986      1987
 "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"  "44.24%"
     1988      1989      1990      1991      1992      1993      1994
 "44.24%"  "44.24%"  "44.24%"  "44.24%"  "37.77%"  "37.41%"  "37.41%"
     1995      1996      1997      1998      1999      2000      2001
 "37.41%"  "37.05%"  "37.05%"  "37.05%"  "37.05%"  "37.05%"  "37.05%"
     2002      2003      2004     NA..1
 "37.05%"  "37.05%"  "37.05%" "100.00%"
```

```
> apply(sugar_comsumption, 1, function(x)(all(is.na(x))))
```

```
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[166] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[188] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[199] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[210] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[221] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[232] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[243] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[254] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[265]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[276]  TRUE  TRUE  TRUE
```

Existe un porcentaje elevado de datos que faltan, incluso una columna completamente vacia y varias filas. En primer lugar se borran las entradas completamente nulas:

```
> sugar_comsumption <- sugar_comsumption[,-c(ncol(sugar_comsumption))]
```

```
> sugar_comsumption <- sugar_comsumption[-c(260:278),]
```

Datos de Paises

Los espacios en blanco que pudieran existir al inicio y final de los nombres ya se han eliminado con la carga de los datos (strip.white=TRUE)

En *badteeh* no están repetidos:

```
> nrow(badteeth)
```

```
[1] 190
```

```
> nrow(unique(badteeth))
```

```
[1] 190
```

```
> length(intersect(sugar_comsumption[,1],intersect(healthexpend[,1],
+   intersect(gdp[,1], intersect(badteeth[,1], adultliteracy[,1])))))
```

```
[1] 176
```

Revisamos las diferencias entre conjuntos de datos:

```
> length(setdiff(gdp[,1],badteeth[,1]))
```

```
[1] 99
```

```
> length(setdiff(badteeth[,1],gdp[,1]))
```

```
[1] 14
```

```
> setdiff(badteeth[,1],gdp[,1])
```

```
 [1] "Central African Rep."            "Cook Islands"
 [3] "Cote D'Ivoire"                   "Czech Rep."
 [5] "Dominican Rep."                  "Korea, Dem. Rep."
 [7] "Korea, Rep."                     "Kyrgyzstan"
 [9] "Laos"                            "Saint Kitts and Nevis"
[11] "Saint Lucia"                     "Saint Vincent and the Grenadines"
[13] "Slovak republic"                 "Yemen, Rep."
```

```
> setdiff(gdp[,1], badteeth[,1])
```

```
 [1] "Abkhazia"                      "Akrotiri and Dhekelia"
 [3] "American Samoa"                "Andorra"
 [5] "Aruba"                         "Azerbaijan"
 [7] "British Virgin Islands"        "Central African Republic"
 [9] "Chad"                          "Channel Islands"
[11] "Christmas Island"              "Cocos Island"
[13] "Comoros"                       "Congo, Rep."
[15] "Cook Is"                       "Cote d'Ivoire"
[17] "Czech Republic"                "Czechoslovakia"
[19] "Dominican Republic"            "East Germany"
[21] "Equatorial Guinea"             "Eritrea"
[23] "Eritrea and Ethiopia"          "Faeroe Islands"
[25] "Falkland Is (Malvinas)"        "French Guiana"
[27] "Greenland"                     "Guadeloupe"
[29] "Guam"                          "Guernsey"
[31] "Guinea"                        "Holy See"
[33] "Isle of Man"                   "Jersey"
[35] "North Korea"                   "South Korea"
[37] "United Korea (former)\n"       "Kosovo"
[39] "Kyrgyz Republic"               "Lao"
[41] "Marshall Islands"              "Mayotte"
[43] "Monaco"                        "Montenegro"
[45] "Montserrat"                    "Nauru"
[47] "Netherlands Antilles"          "Ngorno-Karabakh"
[49] "Norfolk Island"                "Northern Cyprus"
[51] "Northern Mariana Islands"      "Palau"
[53] "Pitcairn"                      "Qatar"
[55] "St. BarthÃ©lemy"               "St. Helena"
[57] "St. Kitts and Nevis"           "St. Lucia"
[59] "St. Martin"                    "St. Vincent and the Grenadines"
[61] "St.-Pierre-et-Miquelon"        "Sao Tome and Principe"
[63] "Serbia"                        "Serbia excluding Kosovo"
[65] "Slovak Republic"               "Somaliland"
[67] "South Ossetia"                 "Svalbard"
[69] "Taiwan"                        "Timor-Leste"
[71] "Transnistria"                  "Turks and Caicos Islands"
[73] "USSR"                          "Wallis et Futuna"
[75] "West Bank and Gaza"            "West Germany"
[77] "Western Sahara"                "Virgin Islands (U.S.)"
[79] "North Yemen (former)"          "South Yemen (former)"
[81] "Yemen"                         "Yugoslavia"
[83] "Ã<U+0085>land"                        "South Sudan"
[85] "Christian"                     "Coastline"
[87] "CuraÃ§ao"                      "Sint Maarten (Dutch part)"
[89] "St. Martin (French part)"      "Antarctica"
[91] "Virgin Islands, British"       "Hawaiian Trade Zone"
[93] "U.S. Pacific Islands"          "Wake Island"
[95] "Bonaire"                       "Sark"
[97] "Chinese Taipei"                "Saint Eustatius"
[99] "Saba"
```

Algunas diferencias se observa que corresponden al mismo país pero escrito de diferentes forma, se unifican. En los casos poco claros de nombres, nos apoyamos en la lista de países de la wikipedia: https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

```
> badteeth[grep("Central",badteeth$Country),1]
```

```
[1] Central African Rep.
190 Levels: Afghanistan Albania Algeria Angola ... Zimbabwe
```

```
> gdp[grep("Central",gdp$Country),1]
```

```
[1] Central African Republic
275 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> badteeth$Country <- gsub("Central.+","Central African Republic",
+   badteeth$Country)
```

```
> badteeth[grep("Central",badteeth$Country),1]
```

```
[1] "Central African Republic"
```

```
> adultliteracy[grep("Central",adultliteracy$Country),1]
```

```
[1] Central African Rep.
260 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> healthexpend[grep("Central",healthexpend$Country),1]
```

```
[1] Central African Republic
265 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> sugar_comsumption[grep("Central",sugar_comsumption$Country),1]
```

```
[1] Central African Rep.
259 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> adultliteracy$Country <- gsub("Central.+","Central African Republic", adultliteracy$Country)
```

```
> sugar_comsumption$Country <- gsub("Central.+","Central African Republic", sugar_comsumption$Country)
```

```
> badteeth$Country <- gsub("Cote.+","Cote d'Ivoire", badteeth$Country)
```

```
> adultliteracy[grep("Cote",adultliteracy$Country),1]
```

```
[1] "Cote d'Ivoire"
```

```
> sugar_comsumption[grep("Cote",sugar_comsumption$Country),1]
```

```
[1] "Cote d'Ivoire"
```

```
> healthexpend[grep("Cote",healthexpend$Country),1]
```

```
[1] Cote d'Ivoire
265 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> badteeth$Country <- gsub("Dominican.+","Dominican Republic",
+   badteeth$Country)
```

```
> adultliteracy[grep("Dominican",adultliteracy$Country),1]
```

```
[1] "Dominican Rep."
```

```
> sugar_comsumption[grep("Dominican",sugar_comsumption$Country),1]
```

```
[1] "Dominican Rep."
```

```
> healthexpend[grep("Dominican",healthexpend$Country),1]
```

```
[1] Dominican Republic
265 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> adultliteracy$Country <- gsub("Dominican.+","Dominican Republic", adultliteracy$Country)
```

```
> sugar_comsumption$Country <- gsub("Dominican.+","Dominican Republic", sugar_comsumption$Country)
```

```
> adultliteracy[grep("Lao",adultliteracy$Country),1]
```

```
[1] "Laos"
```

```
> healthexpend[grep("Lao",healthexpend$Country),1]
```

```
[1] Lao
265 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> sugar_comsumption[grep("Lao",sugar_comsumption$Country),1]
```

```
[1] "Laos"
```

```
> gdp[grep("Lao",gdp$Country),1]
```

```
[1] Lao
275 Levels: Ã<U+0085>land Abkhazia Afghanistan Akrotiri and Dhekelia ... Zimbabwe
```

```
> badteeth[grep("Lao",badteeth$Country),1]
```

```
[1] "Laos"
```

```
> healthexpend$Country <- gsub("Lao","Laos", healthexpend$Country)
```

```
> gdp$Country <- gsub("Lao","Laos", gdp$Country)
```

```
> adultliteracy[grep("Lucia",adultliteracy$Country),1]
```

```
[1] "Saint Lucia"
```

```
> healthexpend[grep("Lucia",healthexpend$Country),1]
```

```
[1] "St. Lucia"
```

```
> sugar_comsumption[grep("Lucia",sugar_comsumption$Country),1]
```

```
[1] "Saint Lucia"
```

```
> gdp[grep("Lucia",gdp$Country),1]
```

```
[1] "St. Lucia"
```

```
> badteeth[grep("Lucia",badteeth$Country),1]
```

```
[1] "Saint Lucia"
```

```
> healthexpend$Country <- gsub(".+Lucia","Saint Lucia", healthexpend$Country)
```

```
> gdp$Country <- gsub(".+Lucia","Saint Lucia", gdp$Country)
```

```
> adultliteracy[grep("Slova",adultliteracy$Country),1]
```

```
[1] "Slovak Republic"
```

```
> healthexpend[grep("Slova",healthexpend$Country),1]
```

```
[1] "Slovak Republic"
```

```
> sugar_comsumption[grep("Slova",sugar_comsumption$Country),1]
```

```
[1] "Slovak Republic"
```

```
> gdp[grep("Slova",gdp$Country),1]
```

```
[1] "Slovak Republic"
```

```
> badteeth[grep("Slova",badteeth$Country),1]
```

```
[1] "Slovak republic"
```

```
> badteeth$Country <- gsub("Slova.+","Slovak Republic", badteeth$Country)
```

```
> adultliteracy[grep("Cook",adultliteracy$Country),1]
```

```
[1] "Cook Islands"
```

```
> healthexpend[grep("Cook",healthexpend$Country),1]
```

```
[1] "Cook Is"
```

```
> sugar_comsumption[grep("Cook",sugar_comsumption$Country),1]
```

```
[1] "Cook Islands"
```

```
> gdp[grep("Cook",gdp$Country),1]
```

```
[1] "Cook Is"
```

```
> badteeth[grep("Cook",badteeth$Country),1]
```

```
[1] "Cook Islands"
```

```
> healthexpend$Country <- gsub("Cook.+","Cook Islands", healthexpend$Country)
```

```
> gdp$Country <- gsub("Cook.+","Cook Islands", gdp$Country)
```

```
> adultliteracy[grep("Czech",adultliteracy$Country),1]
```

```
[1] "Czech Rep."     "Czechoslovakia"
```

```
> healthexpend[grep("Czech",healthexpend$Country),1]
```

```
[1] "Czech Republic" "Czechoslovakia"
```

```
> sugar_comsumption[grep("Czech",sugar_comsumption$Country),1]
```

```
[1] "Czech Rep."     "Czechoslovakia"
```

```
> gdp[grep("Czech",gdp$Country),1]
```

```
[1] "Czech Republic" "Czechoslovakia"
```

```
> badteeth[grep("Czech",badteeth$Country),1]
```

```
[1] "Czech Rep."
```

```
> healthexpend$Country <- gsub("Czech R.+","Czech Rep.", healthexpend$Country)
```

```
> gdp$Country <- gsub("Czech R.+","Czech Rep.", gdp$Country)
```

```
> adultliteracy[grep("Kyrgy",adultliteracy$Country),1]
```

```
[1] "Kyrgyzstan"
```

```
> healthexpend[grep("Kyrgy",healthexpend$Country),1]
```

```
[1] "Kyrgyz Republic"
```

```
> sugar_comsumption[grep("Kyrgy",sugar_comsumption$Country),1]
```

```
[1] "Kyrgyzstan"
```

```
> gdp[grep("Kyrgy",gdp$Country),1]
```

```
[1] "Kyrgyz Republic"
```

```
> badteeth[grep("Kyrgy",badteeth$Country),1]
```

```
[1] "Kyrgyzstan"
```

```
> healthexpend$Country <- gsub("Kyrgy.+","Kyrgyzstan", healthexpend$Country)
```

```
> gdp$Country <- gsub("Kyrgy.+","Kyrgyzstan", gdp$Country)
```

```
> adultliteracy[grep("Vincent",adultliteracy$Country),1]
```

```
[1] "Saint Vincent and the Grenadines"
```

```
> healthexpend[grep("Vincent",healthexpend$Country),1]
```

```
[1] "St. Vincent and the Grenadines"
```

```
> sugar_comsumption[grep("Vincent",sugar_comsumption$Country),1]
```

```
[1] "Saint Vincent and the Grenadines"
```

```
> gdp[grep("Vincent",gdp$Country),1]
```

```
[1] "St. Vincent and the Grenadines"
```

```
> badteeth[grep("Vincent",badteeth$Country),1]
```

```
[1] "Saint Vincent and the Grenadines"
```

```
> healthexpend$Country <- gsub(".+Vincent.+","Saint Vincent and the Grenadines", healthexpend$Country)
```

```
> gdp$Country <- gsub(".+Vincent.+","Saint Vincent and the Grenadines", gdp$Country)
```

```
> adultliteracy[grep("Yemen",adultliteracy$Country),1]
```

```
[1] "Yemen, Rep."                  "Yemen Arab Republic (Former)"
[3] "Yemen Democratic (Former)"
```

```
> healthexpend[grep("Yemen",healthexpend$Country),1]
```

```
[1] "North Yemen (former)" "South Yemen (former)" "Yemen"
```

```
> sugar_comsumption[grep("Yemen",sugar_comsumption$Country),1]
```

```
[1] "Yemen Arab Republic (Former)" "Yemen Democratic (Former)"
[3] "Yemen, Rep."
```

```
> gdp[grep("Yemen",gdp$Country),1]
```

```
[1] "North Yemen (former)" "South Yemen (former)" "Yemen"
```

```
> badteeth[grep("Yemen",badteeth$Country),1]
```

```
[1] "Yemen, Rep."
```

```
> adultliteracy[grep("Korea",adultliteracy$Country),1]
```

```
[1] "Korea, Dem. Rep." "Korea, Rep."      "Korea, United"
```

```
> healthexpend[grep("Korea",healthexpend$Country),1]
```

```
[1] "North Korea"             "South Korea"
[3] "United Korea (former)\n"
```

```
> sugar_comsumption[grep("Korea",sugar_comsumption$Country),1]
```

```
[1] "Korea, Dem. Rep." "Korea, Rep."      "Korea, United"
```

```
> gdp[grep("Korea",gdp$Country),1]
```

```
[1] "North Korea"             "South Korea"
[3] "United Korea (former)\n"
```

```
> badteeth[grep("Korea",badteeth$Country),1]
```

```
[1] "Korea, Dem. Rep." "Korea, Rep."
```

```
> healthexpend$Country <- gsub("North Korea","Korea, Dem. Rep.", healthexpend$Country)
```

```
> gdp$Country <- gsub("North Korea","Korea, Dem. Rep.", gdp$Country)
```

```
> healthexpend$Country <- gsub("South Korea","Korea, Rep.", healthexpend$Country)
```

```
> gdp$Country <- gsub("South Korea","Korea, Rep.", gdp$Country)
```

```
> gdp$Country <- gsub("United Korea.+","Korea, United", gdp$Country)
```

```
> healthexpend$Country <- gsub("United Korea.+","Korea, United", healthexpend$Country)
```

```
> healthexpend$Country <- gsub("North Yemen.+","Yemen Arab Republic (Former)", healthexpend$Country)
```

```
> gdp$Country <- gsub("North Yemen.+","Yemen Arab Republic (Former)", gdp$Country)
```

```
> healthexpend$Country <- gsub("South Yemen.+","Yemen Democratic (Former)", healthexpend$Country)
```

```
> gdp$Country <- gsub("South Yemen.+","Yemen Democratic (Former)", gdp$Country)
```

```
> healthexpend$Country <- gsub("Yemen","Yemen, Rep.", healthexpend$Country)
```

```
> gdp$Country <- gsub("Yemen","Yemen, Rep.", gdp$Country)
```

```
> adultliteracy[grep("Kitts",adultliteracy$Country),1]
```

```
[1] "Saint Kitts and Nevis"
```

```
> healthexpend[grep("Kitts",healthexpend$Country),1]
```

```
[1] "St. Kitts and Nevis"
```

```
> sugar_comsumption[grep("Kitts",sugar_comsumption$Country),1]
```

```
[1] "Saint Kitts and Nevis"
```

```
> gdp[grep("Kitts",gdp$Country),1]
```

```
[1] "St. Kitts and Nevis"
```

```
> badteeth[grep("Kitts",badteeth$Country),1]
```

```
[1] "Saint Kitts and Nevis"
```

```
> healthexpend$Country <- gsub(".+Kitts.+","Saint Kitts and Nevis", healthexpend$Country)
```

```
> gdp$Country <- gsub(".+Kitts.+","Saint Kitts and Nevis", gdp$Country)
```

```
> nrow(badteeth)
```

```
[1] 190
```

```
> length(intersect(sugar_comsumption[,1],intersect(healthexpend[,1],intersect(gdp[,1], intersect(badt
+   1], adultliteracy[,1])))))
```

```
[1] 190
```

Tipos de datos

```
> str(adultliteracy)
```

```
'data.frame':   260 obs. of  38 variables:
 $ Country: chr  "Afghanistan" "Albania" "Algeria" "Andorra" ...
 $ 1975   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1976   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1977   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1978   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1979   : num  18.2 NA NA NA NA ...
 $ 1980   : num  NA NA NA NA NA ...
 $ 1981   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1982   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1983   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1984   : num  NA NA NA NA NA ...
 $ 1985   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1986   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1987   : num  NA NA 49.6 NA NA ...
 $ 1988   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1989   : num  NA NA NA NA NA ...
 $ 1990   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1991   : num  NA NA NA NA NA ...
 $ 1992   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1993   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1994   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1995   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1996   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1997   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1998   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 1999   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 2000   : num  NA NA NA NA NA ...
 $ 2001   : num  NA 98.7 NA NA 67.4 ...
 $ 2002   : num  NA NA 69.9 NA NA ...
 $ 2003   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 2004   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 2005   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 2006   : num  NA NA 72.6 NA NA ...
 $ 2007   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 2008   : num  NA 95.9 NA NA NA ...
 $ 2009   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ 2010   : num  NA NA NA NA NA ...
 $ 2011   : num  39 96.8 NA NA 70.4 ...
```

```
> str(badteeth)
```

```
'data.frame':   190 obs. of  2 variables:
 $ Country: chr  "Afghanistan" "Albania" "Algeria" "Angola" ...
 $ 2004   : num  2.9 3.02 2.3 1.7 2.5 0.7 3.4 2.4 0.8 1 ...
```

```
> str(gdp)
```

```
'data.frame':   275 obs. of  53 variables:
 $ Country: chr  "Abkhazia" "Afghanistan" "Akrotiri and Dhekelia" "Albania" ...
 $ 1960   : num  NA NA NA NA 1280 ...
 $ 1961   : num  NA NA NA NA 1085 ...
 $ 1962   : num  NA NA NA NA 856 ...
 $ 1963   : num  NA NA NA NA 1128 ...
 $ 1964   : num  NA NA NA NA 1170 ...
 $ 1965   : num  NA NA NA NA 1215 ...
 $ 1966   : num  NA NA NA NA 1128 ...
 $ 1967   : num  NA NA NA NA 1201 ...
 $ 1968   : num  NA NA NA NA 1292 ...
 $ 1969   : num  NA NA NA NA 1359 ...
 $ 1970   : num  NA NA NA NA 1436 ...
 $ 1971   : num  NA NA NA NA 1236 ...
 $ 1972   : num  NA NA NA NA 1528 ...
 $ 1973   : num  NA NA NA NA 1538 ...
 $ 1974   : num  NA NA NA NA 1603 ...
 $ 1975   : num  NA NA NA NA 1632 ...
 $ 1976   : num  NA NA NA NA 1714 ...
 $ 1977   : num  NA NA NA NA 1748 ...
 $ 1978   : num  NA NA NA NA 1848 ...
 $ 1979   : num  NA NA NA NA 1923 ...
 $ 1980   : num  NA NA NA 1061 1876 ...
 $ 1981   : num  NA NA NA 1100 1870 ...
 $ 1982   : num  NA NA NA 1111 1925 ...
 $ 1983   : num  NA NA NA 1101 1963 ...
 $ 1984   : num  NA NA NA 1065 2008 ...
 $ 1985   : num  NA NA NA 1060 2020 ...
 $ 1986   : num  NA NA NA 1092 1970 ...
 $ 1987   : num  NA NA NA 1054 1902 ...
 $ 1988   : num  NA NA NA 1014 1833 ...
 $ 1989   : num  NA NA NA 1092 1865 ...
 $ 1990   : num  NA NA NA 978 1833 ...
 $ 1991   : num  NA NA NA 688 1767 ...
 $ 1992   : num  NA NA NA 643 1756 ...
 $ 1993   : num  NA NA NA 714 1680 ...
 $ 1994   : num  NA NA NA 785 1630 ...
 $ 1995   : num  NA NA NA 900 1660 ...
 $ 1996   : num  NA NA NA 991 1698 ...
 $ 1997   : num  NA NA NA 896 1690 ...
 $ 1998   : num  NA NA NA 1014 1751 ...
 $ 1999   : num  NA NA NA 1118 1781 ...
 $ 2000   : num  NA NA NA 1200 1794 ...
 $ 2001   : num  NA NA NA 1282 1814 ...
 $ 2002   : num  NA NA NA 1314 1872 ...
 $ 2003   : num  NA NA NA 1381 1972 ...
 $ 2004   : num  NA NA NA 1454 2043 ...
 $ 2005   : num  NA NA NA 1526 2115 ...
 $ 2006   : num  NA NA NA 1594 2125 ...
 $ 2007   : num  NA NA NA 1682 2155 ...
 $ 2008   : num  NA NA NA 1804 2174 ...
 $ 2009   : num  NA NA NA 1857 2193 ...
 $ 2010   : num  NA NA NA 1915 2232 ...
 $ 2011   : num  NA NA NA 1966 2255 ...
```

```
> str(healthexpend)
```

```
'data.frame':   265 obs. of  17 variables:
 $ Country: chr  "Abkhazia" "Afghanistan" "Akrotiri and Dhekelia" "Albania" ...
 $ 1995   : num  NA NA NA 13.9 46.8 ...
 $ 1996   : num  NA NA NA 17.1 48 ...
 $ 1997   : num  NA NA NA 14.2 49.7 ...
 $ 1998   : num  NA NA NA 18.6 48.7 ...
 $ 1999   : num  NA NA NA 28.1 45.5 ...
 $ 2000   : num  NA NA NA 27.2 45.9 ...
 $ 2001   : num  NA NA NA 30.5 52.5 ...
 $ 2002   : num  NA 0.833 NA 32.55 54.078 ...
 $ 2003   : num  NA 1.25 NA 40.61 62.64 ...
 $ 2004   : num  NA 1.61 NA 63.94 63.23 ...
 $ 2005   : num  NA 2.53 NA 71.36 69.3 ...
 $ 2006   : num  NA 2.81 NA 75.55 81.68 ...
 $ 2007   : num  NA 3.5 NA 88.8 108.9 ...
 $ 2008   : num  NA 3.74 NA 109.07 147.82 ...
 $ 2009   : num  NA 3.91 NA 106.89 143.16 ...
 $ 2010   : num  NA 4.39 NA 94.02 138.84 ...
```

```
> str(sugar_comsumption)
```

```
'data.frame':   259 obs. of  45 variables:
 $ Country: chr  "Abkhazia" "Afghanistan" "Akrotiri and Dhekelia" "Albania" ...
 $ 1961   : num  NA NA NA 30.1 46.6 ...
 $ 1962   : num  NA NA NA 30.1 49.3 ...
 $ 1963   : num  NA NA NA 32.9 46.6 ...
 $ 1964   : num  NA NA NA 35.6 49.3 ...
 $ 1965   : num  NA NA NA 35.6 46.6 ...
 $ 1966   : num  NA NA NA 35.6 46.6 ...
 $ 1967   : num  NA NA NA 38.4 49.3 ...
 $ 1968   : num  NA NA NA 38.4 49.3 ...
 $ 1969   : num  NA NA NA 38.4 46.6 ...
 $ 1970   : num  NA NA NA 38.4 43.8 ...
 $ 1971   : num  NA NA NA 41.1 52.1 ...
 $ 1972   : num  NA NA NA 41.1 49.3 ...
 $ 1973   : num  NA NA NA 41.1 46.6 ...
 $ 1974   : num  NA NA NA 43.8 49.3 ...
 $ 1975   : num  NA NA NA 43.8 63 ...
 $ 1976   : num  NA NA NA 43.8 65.8 ...
 $ 1977   : num  NA NA NA 41.1 71.2 ...
 $ 1978   : num  NA NA NA 43.8 71.2 ...
 $ 1979   : num  NA NA NA 46.6 82.2 ...
 $ 1980   : num  NA NA NA 46.6 82.2 ...
 $ 1981   : num  NA NA NA 46.6 82.2 ...
 $ 1982   : num  NA NA NA 46.6 74 ...
 $ 1983   : num  NA NA NA 46.6 82.2 ...
 $ 1984   : num  NA NA NA 46.6 82.2 ...
 $ 1985   : num  NA NA NA 46.6 79.5 ...
 $ 1986   : num  NA NA NA 49.3 84.9 ...
 $ 1987   : num  NA NA NA 49.3 93.2 ...
 $ 1988   : num  NA NA NA 49.3 79.5 ...
 $ 1989   : num  NA NA NA 49.3 84.9 ...
 $ 1990   : num  NA NA NA 52.1 82.2 ...
 $ 1991   : num  NA NA NA 38.4 79.5 ...
 $ 1992   : num  NA NA NA 52.1 74 ...
 $ 1993   : num  NA NA NA 79.5 76.7 ...
 $ 1994   : num  NA NA NA 101 74 ...
 $ 1995   : num  NA NA NA 54.8 74 ...
 $ 1996   : num  NA NA NA 68.5 74 ...
 $ 1997   : num  NA NA NA 60.3 79.5 ...
 $ 1998   : num  NA NA NA 60.3 54.8 ...
 $ 1999   : num  NA NA NA 57.5 60.3 ...
 $ 2000   : num  NA NA NA 65.8 82.2 ...
 $ 2001   : num  NA NA NA 68.5 79.5 ...
 $ 2002   : num  NA NA NA 71.2 82.2 ...
 $ 2003   : num  NA NA NA 65.8 84.9 ...
 $ 2004   : num  NA NA NA 65.8 84.9 ...
```

Modificamos el tipo para *Country*:

```
> adultliteracy$Country <- as.factor(adultliteracy$Country)
```

```
> str(adultliteracy$Country)
```

```
 Factor w/ 260 levels "Ã<U+0085>land","Abkhazia",..: 3 5 6 8 9 10 11 12 13 14 ...
```

```
> badteeth$Country <- as.factor(badteeth$Country)
```

```
> gdp$Country <- as.factor(gdp$Country)
```

```
> healthexpend$Country <- as.factor(healthexpend$Country)
```

```
> sugar_comsumption$Country <- as.factor(sugar_comsumption$Country)
```

Reducción de la dimensionalidad

Dado que el interés del estudio es la salud dental, podríamos reducir los datos a los que sean relacionables con *badteeth* (países presentes en *badteeth* y datos anteriores o iguales a 2004), pero por ahora los conservamos para apoyar también la relación entre riqueza y nivel educativo o consumo de azúcar. Ya lo realizaremos en la integración de los datos.

Normalización de datos

Transformaciones para que los datos sean comparables o faciliten su comprensión. Por ejemplo el valor de *badteeth* corresponde al número de dientes estropeados en niños de 12 años, y podría interesar utilizar un porcentaje en su lugar, para ello intersaría saber el total de dientes, a esa edad parecería que son 28, pero no lo aplicamos al no tener certeza.

Integración de los datos

Conversión de cada matriz de datos en un conjunto de [pais, año, valor] para poder integrar.

```
> badteethExt <- data.frame(Country=character(), Year=character(), BadTeeths=integer(), stringsAsFacto
```

```
> for (j in 2:ncol(badteeth))
+ {
+ year<- colnames(badteeth)[j]
+ for (i in 1:nrow(badteeth))
+ {
+ badteethExt[(j-2)*nrow(badteeth)+i,1] <- as.character(badteeth$Country[i])
+ badteethExt[(j-2)*nrow(badteeth)+i,2] <- year
+ badteethExt[(j-2)*nrow(badteeth)+i,3] <- badteeth[i,j]
+ }
+ }
```

```
> nrow(badteethExt)
```

```
[1] 190
```

```
> head(badteethExt)
```

```
              Country Year BadTeeths
1         Afghanistan 2004      2.90
2             Albania 2004      3.02
3             Algeria 2004      2.30
4              Angola 2004      1.70
5            Anguilla 2004      2.50
6 Antigua and Barbuda 2004      0.70
```

```
> adultliteracyExt <- data.frame(Country=character(), Year=character(), LiteracyRate=integer(), string
```

```
> for (j in 2:ncol(adultliteracy))
+ {
+ year<- colnames(adultliteracy)[j]
+ for (i in 1:nrow(adultliteracy))
+ {
+ adultliteracyExt[(j-2)*nrow(adultliteracy)+i,1] <- as.character(adultliteracy$Country[i])
+ adultliteracyExt[(j-2)*nrow(adultliteracy)+i,2] <- year
+ adultliteracyExt[(j-2)*nrow(adultliteracy)+i,3] <- adultliteracy[i,j]
+ }
+ }
```

```
> nrow(adultliteracyExt) == nrow(adultliteracy)*(ncol(adultliteracy)-1)
```

```
[1] TRUE
```

```
> adultliteracyExt[1:3,]
```

```
      Country Year LiteracyRate
1 Afghanistan 1975           NA
2     Albania 1975           NA
3     Algeria 1975           NA
```

```
> healthexpendExt <- data.frame(Country=character(), Year=character(), HealthExpend=integer(), string
```

```
> for (j in 2:ncol(healthexpend))
+ {
+ year<- colnames(healthexpend)[j]
+ for (i in 1:nrow(healthexpend))
+ {
+ healthexpendExt[(j-2)*nrow(healthexpend)+i,1] <- as.character(healthexpend$Country[i])
+ healthexpendExt[(j-2)*nrow(healthexpend)+i,2] <- year
+ healthexpendExt[(j-2)*nrow(healthexpend)+i,3] <- healthexpend[i,j]
+ }
+ }
```

```
> nrow(healthexpendExt) == nrow(healthexpend)*(ncol(healthexpend)-1)
```

```
[1] TRUE
```

```
> healthexpendExt[1:3,]
```

```
                Country Year HealthExpend
1              Abkhazia 1995           NA
2           Afghanistan 1995           NA
3 Akrotiri and Dhekelia 1995           NA
```

```
> gdpExt <- data.frame(Country=character(), Year=character(), GDP=integer(), stringsAsFactors=FALSE)
```

```
> for (j in 2:ncol(gdp))
+ {
+ year<- colnames(gdp)[j]
+ for (i in 1:nrow(gdp))
+ {
+ gdpExt[(j-2)*nrow(gdp)+i,1] <- as.character(gdp$Country[i])
+ gdpExt[(j-2)*nrow(gdp)+i,2] <- year
+ gdpExt[(j-2)*nrow(gdp)+i,3] <- gdp[i,j]
+ }
+ }
```

```
> nrow(gdpExt) == nrow(gdp)*(ncol(gdp)-1)
```

```
[1] TRUE
```

```
> gdpExt[1:3,]
```

```
                Country Year GDP
1              Abkhazia 1960  NA
2           Afghanistan 1960  NA
3 Akrotiri and Dhekelia 1960  NA
```

```
> sugar_comsumptionExt <- data.frame(Country=character(), Year=character(), SugarComsumption=integer()
+   stringsAsFactors=FALSE)
```

```
> for (j in 2:ncol(sugar_comsumption))
+ {
+ year<- colnames(sugar_comsumption)[j]
+ for (i in 1:nrow(sugar_comsumption))
+ {
+ sugar_comsumptionExt[(j-2)*nrow(sugar_comsumption)+i,1] <- as.character(sugar_comsumption$Country[i]
+ sugar_comsumptionExt[(j-2)*nrow(sugar_comsumption)+i,2] <- year
+ sugar_comsumptionExt[(j-2)*nrow(sugar_comsumption)+i,3] <- sugar_comsumption[i,j]
+ }
+ }
```

```
> nrow(sugar_comsumptionExt) == nrow(sugar_comsumption)*(ncol(sugar_comsumption)-1)
```

```
[1] TRUE
```

```
> sugar_comsumptionExt[1:3,]
```

```
                Country Year SugarComsumption
1              Abkhazia 1961               NA
2           Afghanistan 1961               NA
3 Akrotiri and Dhekelia 1961               NA
```

Integración de los datos en un único dataset:

```
> bt <- merge(badteethExt, gdpExt, by=c("Country", "Year"))
```

```
> bt <- merge(bt, healthexpendExt, by=c("Country", "Year"))
```

```
> bt <- merge(bt, sugar_comsumptionExt, by=c("Country", "Year"))
```

```
> bt <- merge(bt, adultliteracyExt, by=c("Country", "Year"))
```

Revisamos los datos integrados:

```
> nrow(bt)
```

```
[1] 190
```

```
> bt[1:4,]
```

```
      Country Year BadTeeths       GDP HealthExpend SugarComsumption
1 Afghanistan 2004      2.90        NA      1.61416               NA
2     Albania 2004      3.02 1454.0229     63.93560            65.75
3     Algeria 2004      2.30 2043.1357     63.22940            84.93
4      Angola 2004      1.70  353.2315     19.66478            35.62
  LiteracyRate
1           NA
2           NA
3           NA
4           NA
```

Al realizar el *merge* hemos reducido la dimensionalidad que se citaba anteriormente con los países sin datos de *badteeth* y con los años distintos de 2.004

```
> str(bt)
```

```
'data.frame':   190 obs. of  7 variables:
 $ Country         : chr  "Afghanistan" "Albania" "Algeria" "Angola" ...
 $ Year            : chr  "2004" "2004" "2004" "2004" ...
 $ BadTeeths       : num  2.9 3.02 2.3 1.7 2.5 0.7 3.4 2.4 0.8 1 ...
 $ GDP             : num  NA 1454 2043 353 NA ...
 $ HealthExpend    : num  1.61 63.94 63.23 19.66 NA ...
 $ SugarComsumption: num  NA 65.8 84.9 35.6 NA ...
 $ LiteracyRate    : num  NA NA NA NA NA NA NA NA NA NA ...
```

```
> bt$Country <- as.factor(bt$Country)
```

```
> str(bt)
```

```
'data.frame':   190 obs. of  7 variables:
 $ Country         : Factor w/ 190 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Year            : chr  "2004" "2004" "2004" "2004" ...
 $ BadTeeths       : num  2.9 3.02 2.3 1.7 2.5 0.7 3.4 2.4 0.8 1 ...
 $ GDP             : num  NA 1454 2043 353 NA ...
 $ HealthExpend    : num  1.61 63.94 63.23 19.66 NA ...
 $ SugarComsumption: num  NA 65.8 84.9 35.6 NA ...
 $ LiteracyRate    : num  NA NA NA NA NA NA NA NA NA NA ...
```

3. Análisis

```
> sapply(bt, function(x)(sum(is.na(x))))
```

```
        Country             Year        BadTeeths              GDP
              0                0                0               16
   HealthExpend SugarComsumption     LiteracyRate
             17               25              163
```

```
> sapply(bt, function(x)(sprintf("%.2f%%",sum(is.na(x))*100/nrow(bt))))
```

```
        Country             Year        BadTeeths              GDP
        "0.00%"          "0.00%"          "0.00%"          "8.42%"
   HealthExpend SugarComsumption     LiteracyRate
        "8.95%"         "13.16%"         "85.79%"
```

Falta el 86% de los datos respecto al nivel educativo (*LiteracyRate*), poco relevante para el estudio.

Los que no tienen datos, se borran:

```
> bta <- subset(bt, !is.na(GDP)| !is.na(HealthExpend)| !is.na(SugarComsumption) | !is.na(LiteracyRate)
```

```
> nrow(bta)
```

```
[1] 183
```

```
> btb <- subset(bt, !is.na(GDP)& !is.na(HealthExpend)& !is.na(SugarComsumption))
```

```
> nrow(btb)
```

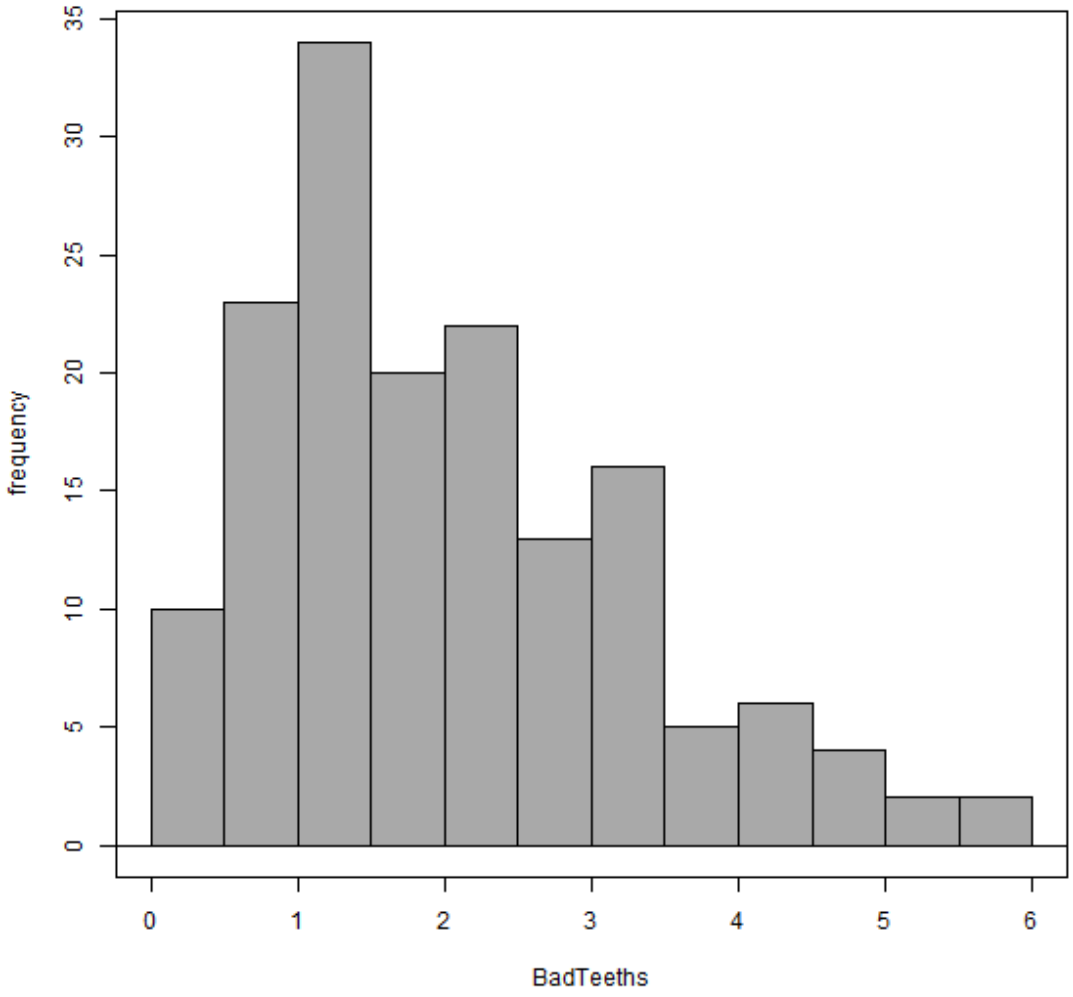
```
[1] 157
```

badteeths

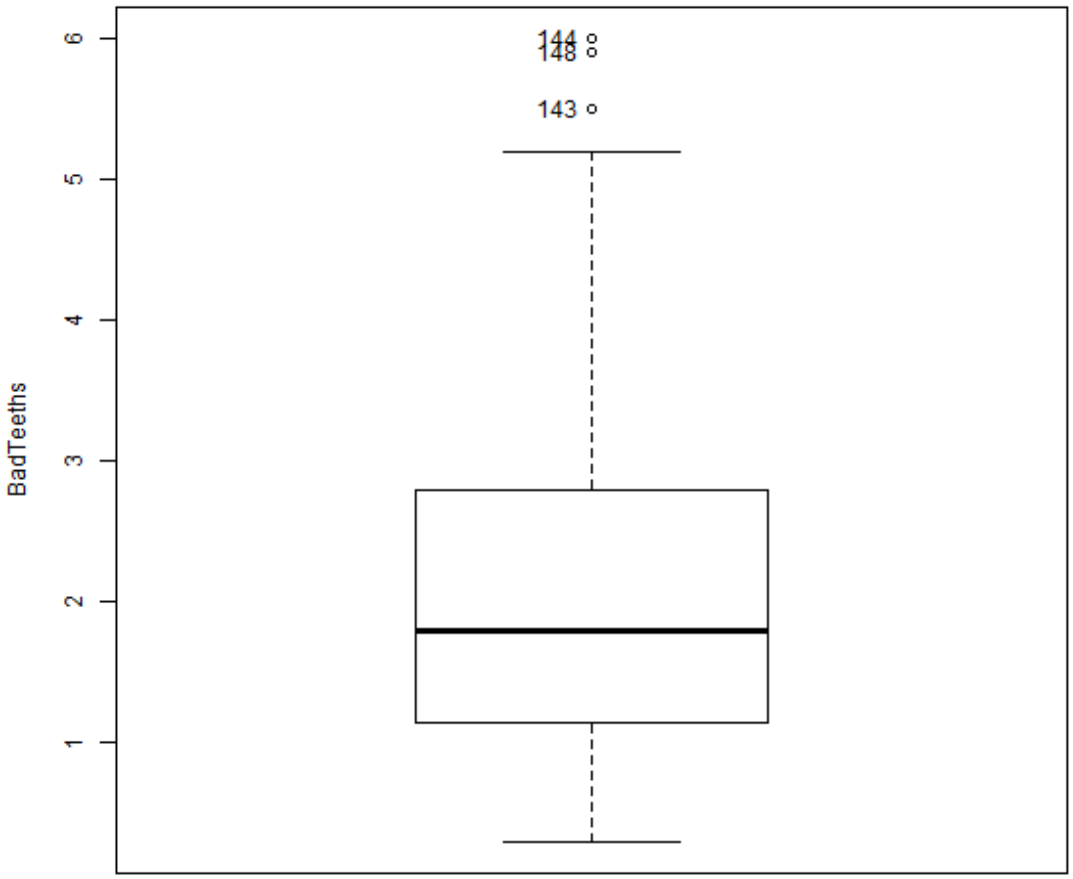
```
> summary(btb$BadTeeths)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.300   1.140   1.800   2.076   2.800   6.000
```

```
> with(btb, Hist(BadTeeths, scale="frequency", breaks="Sturges", col="darkgray"))
```




```
> Boxplot( ~ BadTeeths, data=btb, id.method="y")
```




```
[1] "143" "144" "148"
```

Boxplot o diagrama de caja representa el rango intercuartílico, entre el valor máximo y valor mínimo, formando una caja entre Q1 y Q3 con Q2 o mediana como línea cruzada. Los valores atípicos o outliers se presentan en los extremos alejados de la caja: “valores inferiores a Q1 - 1,5IQR o valores superiores a Q3 + IQR1,5”.

Pueden observarse valors atípicos en los valores superiores de la variable.

Dispersión usando las medidas: rango intercuartílico, varianza y desviación típica:

```
> IQR(btb$BadTeeths)
```

```
[1] 1.66
```

```
> var(btb$BadTeeths)
```

```
[1] 1.570963
```

```
> sd(btb$BadTeeths)
```

```
[1] 1.253381
```

```
> btb[which(btb$BadTeeths > 1.5*1.66+2.8),]
```

```
                  Country Year BadTeeths      GDP HealthExpend
143 Saint Kitts and Nevis 2004       5.5 9465.436     193.1087
144           Saint Lucia 2004       6.0 5078.464     159.4894
148          Saudi Arabia 2004       5.9 9261.922     285.5432
    SugarComsumption LiteracyRate
143           156.16           NA
144            98.63           NA
148            73.97     82.85774
```

Como vemos en el diagrama de caja, en este caso tenemos 3 valores atípicos (superiores a 1.5*IQR+Q3). Los conservamos, son outliers pero al mismo tiempo están alejados del máximo nivel posible (28 piezas dentales?)

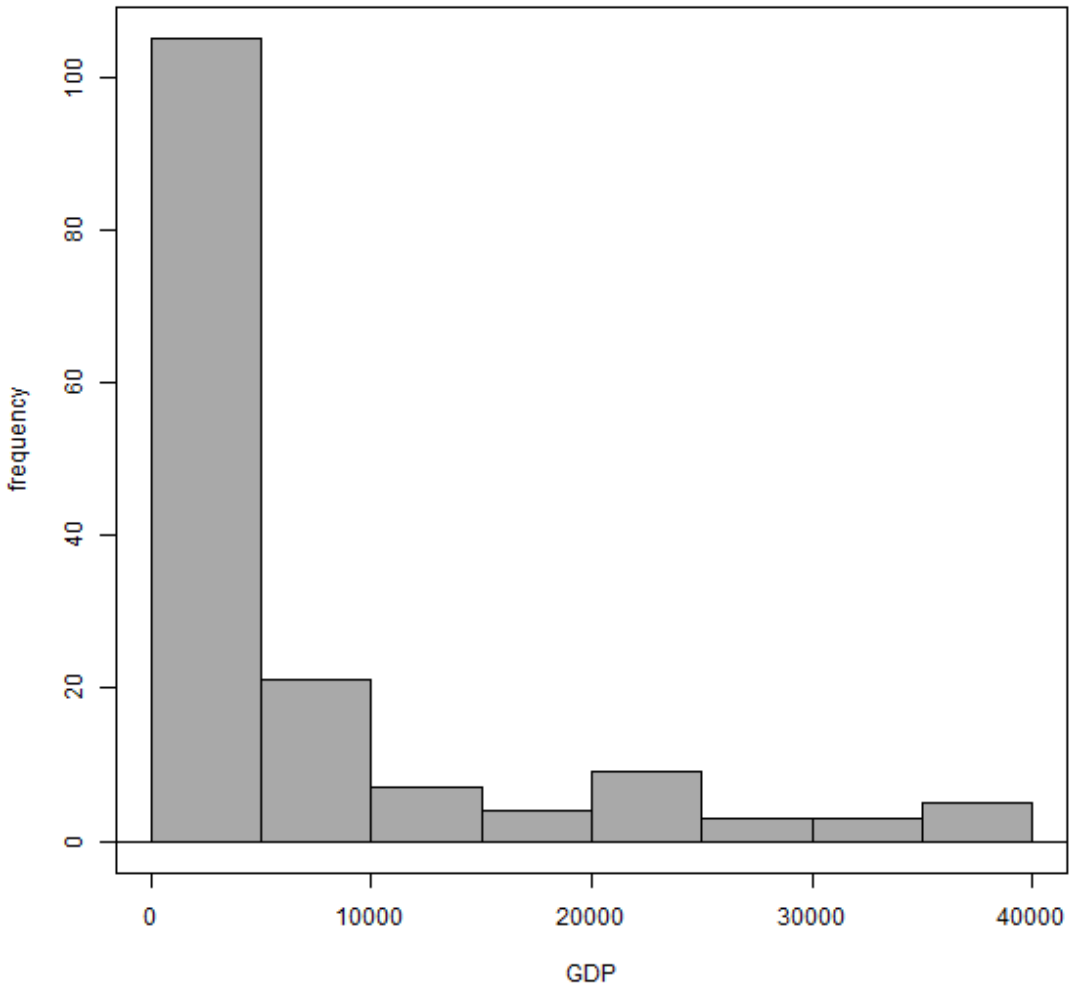
GDP

```
> smm<-summary(btb$GDP)
```

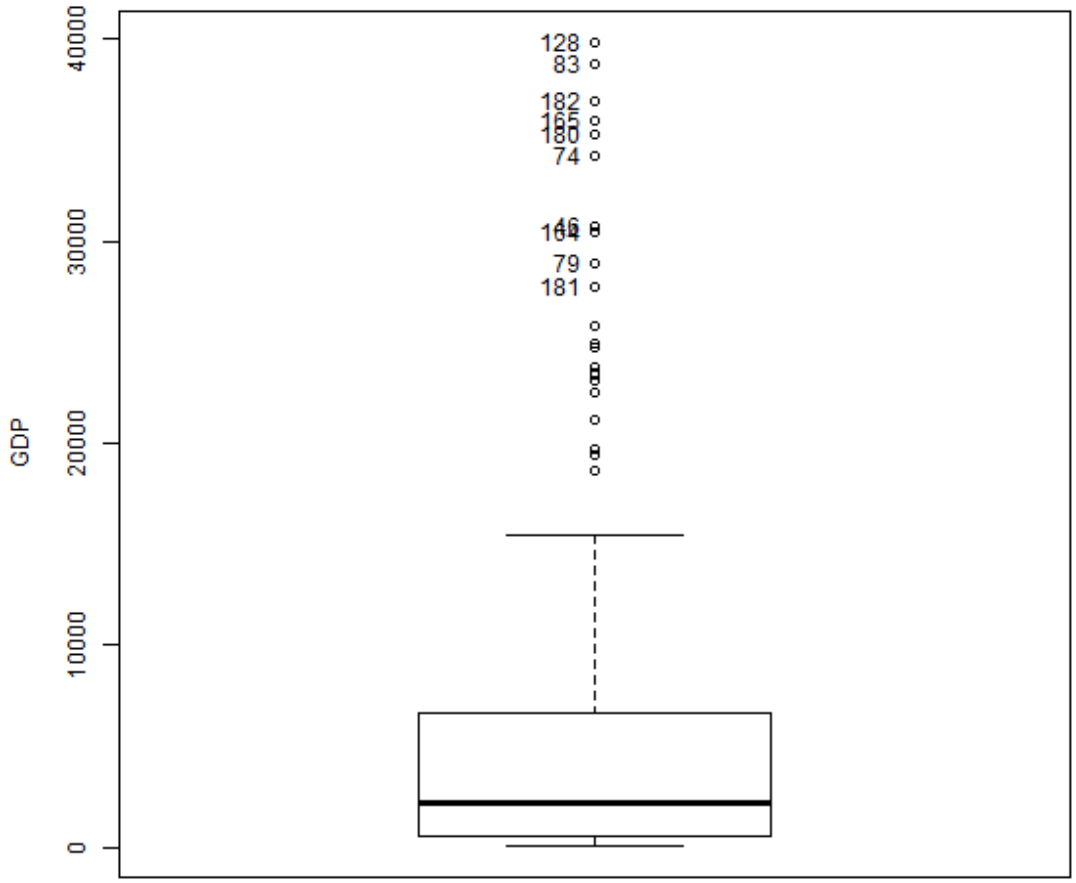
```
> smm
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   88.24   577.10  2164.63  6581.89  6588.90 39830.41
```

```
> with(btb, Hist(GDP, scale="frequency", breaks="Sturges", col="darkgray"))
```




```
> Boxplot( ~ GDP, data=btb, id.method="y")
```




```
 [1] "128" "83"  "182" "165" "180" "74"  "46"  "164" "79"  "181"
```

```
> iqr<-IQR(btb$GDP)
```

```
> iqr
```

```
[1] 6011.791
```

```
> var(btb$GDP)
```

```
[1] 92499236
```

```
> sd(btb$GDP)
```

```
[1] 9617.652
```

La varianza no es homogénea, mucha dispersión de los datos, los valores de los datos son elevados para su manejo, interesa estandarizar esta variable. En el diagrama de caja se ve que la distribución es asimétrica, el 50% de los casos por encima de la mediana tienen más dispersión que el 50% de los casos que son inferiores a la mediana

```
> btb[which(btb$GDP > 1.5*iqr+smm[5]),]
```

```
                Country Year BadTeeths      GDP HealthExpend
9             Australia 2004      0.80 23498.26    1915.1531
10              Austria 2004      1.00 24945.05    2787.6847
11              Bahamas 2004      1.60 21106.72     633.1497
16              Belgium 2004      1.10 23750.46    2677.6519
25               Brunei 2004      4.80 18609.15     557.9192
31               Canada 2004      2.10 24936.83    2132.4034
46              Denmark 2004      0.80 30773.71    3692.8056
56              Finland 2004      1.20 25774.06    2225.7377
57               France 2004      1.20 22495.24    2861.5022
62              Germany 2004      0.70 23390.86    2706.4983
74              Iceland 2004      1.40 34230.18    3714.6547
79              Ireland 2004      1.10 28937.33    2668.2599
80               Israel 2004      1.66 19366.34     880.5043
81                Italy 2004      1.10 19744.89    1952.3460
83                Japan 2004      1.70 38793.62    2352.0917
90               Kuwait 2004      2.60 23107.47     595.9849
121         Netherlands 2004      0.80 24746.84    2234.8590
128              Norway 2004      1.70 39830.41    4542.5184
164              Sweden 2004      1.00 30434.45    2977.2338
165         Switzerland 2004      0.86 36003.23    3213.2000
180 United Arab Emirates 2004      1.60 35316.34     588.3173
181       United Kingdom 2004      0.70 27752.91    2392.0480
182        United States 2004      1.19 36931.39    2785.6029
    SugarComsumption LiteracyRate
9             128.77           NA
10            123.29           NA
11            126.03           NA
16            150.69           NA
25            106.85           NA
31            172.60           NA
46            158.90           NA
56             93.15           NA
57            109.59           NA
62            123.29           NA
74            153.43           NA
79            115.07           NA
80            104.11           NA
81             84.93           NA
83             76.71           NA
90            101.37           NA
121           142.47           NA
128           120.55           NA
164           128.77           NA
165           164.38           NA
180           104.11           NA
181           112.33           NA
182           191.78           NA
```

Muchos *outliers*. Revisandolos someramente parecen datos correctos, corresponden a países con el PIB (GDP) más elevado del mundo.

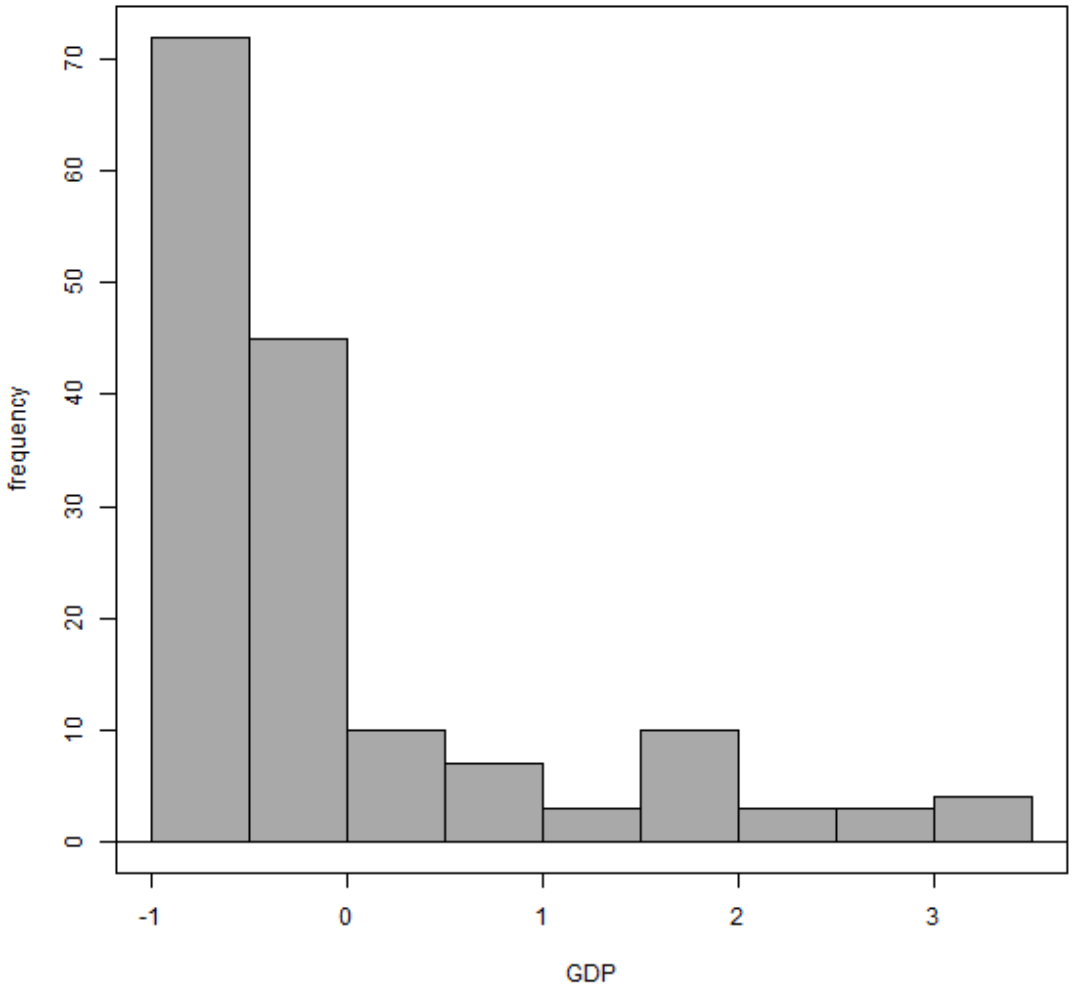
```
> btc<-btb
```

```
> btc$GDP<- (btb$GDP - mean(btb$GDP)) / sd(btb$GDP)
```

```
> summary(btc$GDP)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.675180 -0.624350 -0.459286  0.000000  0.000729  3.457031
```

```
> with(btc, Hist(GDP, scale="frequency", breaks="Sturges", col="darkgray"))
```



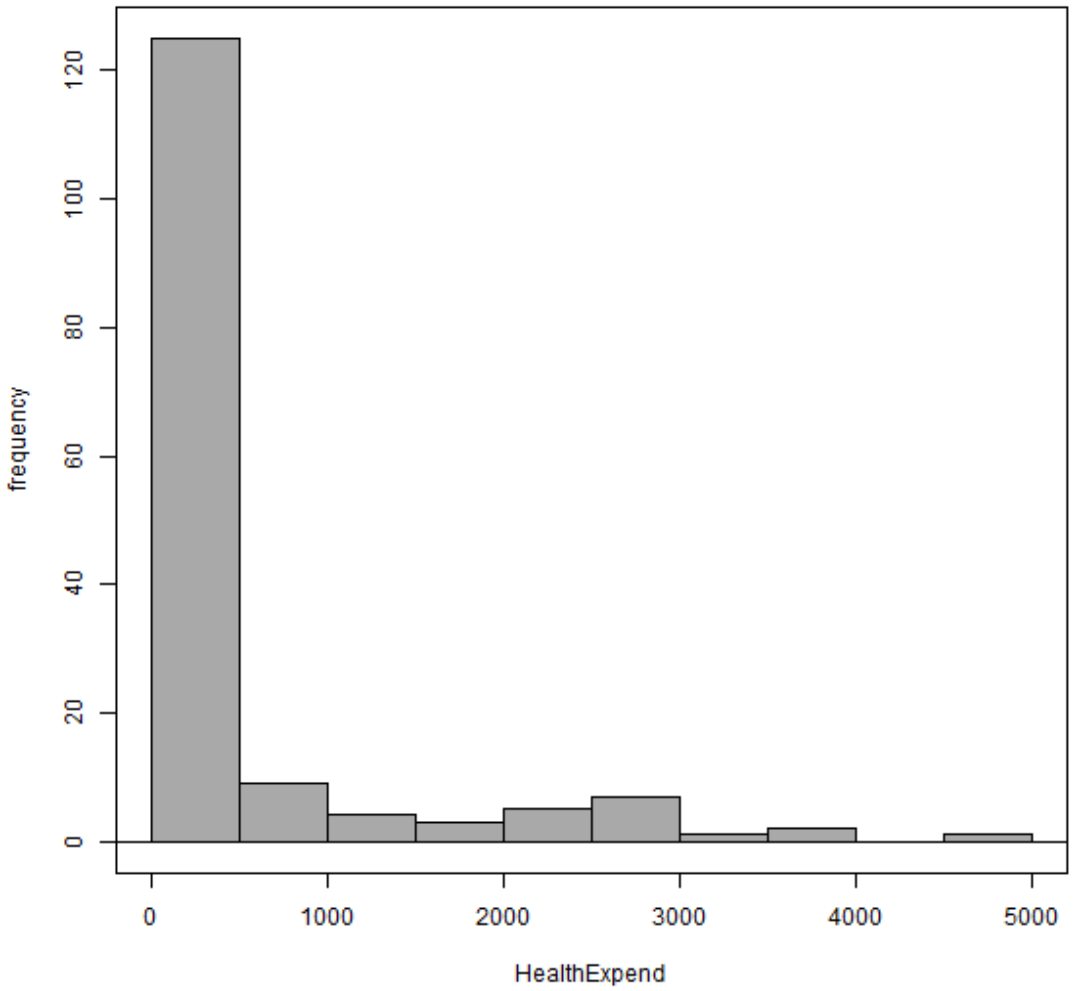
HealthExpend

```
> smm<-summary(btc$HealthExpend)
```

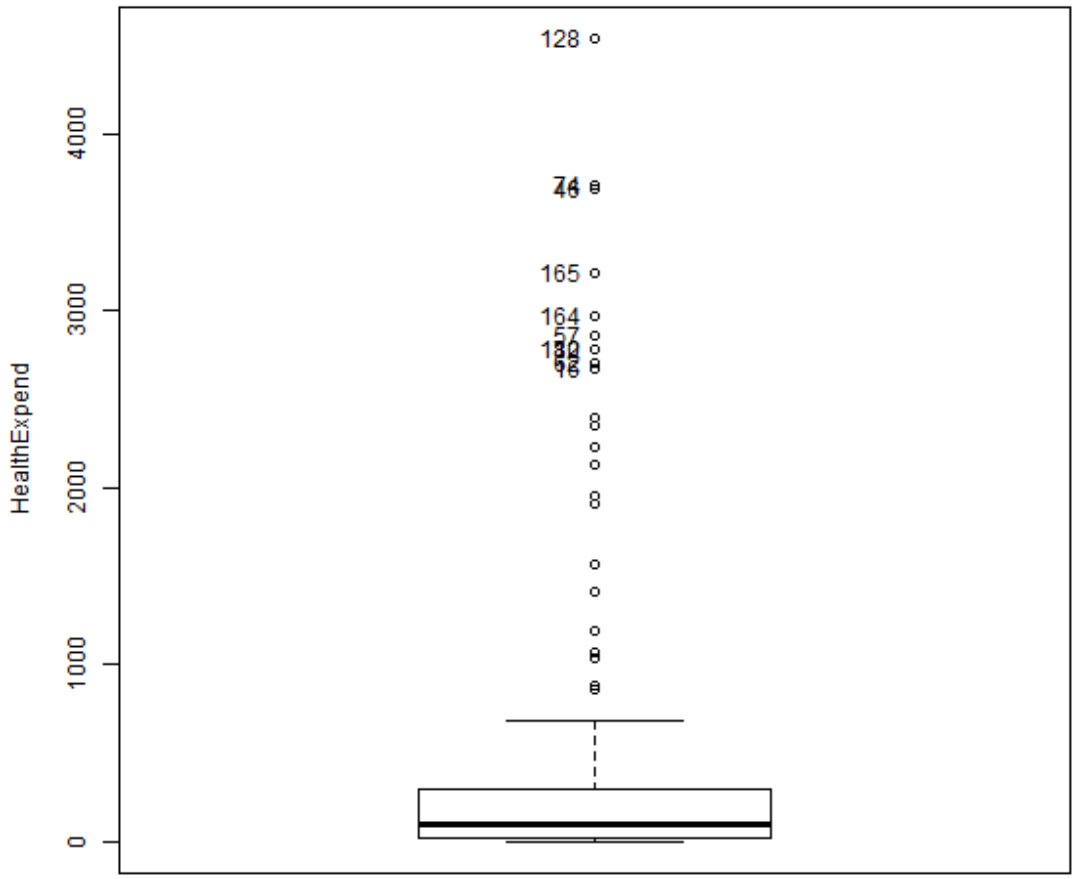
```
> smm
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   1.416   16.221   96.154  469.230  290.616 4542.518
```

```
> with(btc, Hist(HealthExpend, scale="frequency", breaks="Sturges", col="darkgray"))
```




```
> Boxplot( ~ HealthExpend, data=btc, id.method="y")
```




```
 [1] "128" "74"  "46"  "165" "164" "57"  "10"  "182" "62"  "16"
```

```
> iqr<-IQR(btc$HealthExpend)
```

```
> iqr
```

```
[1] 274.3949
```

```
> var(btc$HealthExpend)
```

```
[1] 807493.5
```

```
> sd(btc$HealthExpend)
```

```
[1] 898.6064
```

```
> btc[which(btc$HealthExpend > 1.5*iqr+smm[5]),]
```

```
           Country Year BadTeeths       GDP HealthExpend SugarComsumption
9        Australia 2004      0.80 1.7588882    1915.1531           128.77
10         Austria 2004      1.00 1.9093184    2787.6847           123.29
16         Belgium 2004      1.10 1.7851105    2677.6519           150.69
31          Canada 2004      2.10 1.9084635    2132.4034           172.60
46         Denmark 2004      0.80 2.5153559    3692.8056           158.90
56         Finland 2004      1.20 1.9955158    2225.7377            93.15
57          France 2004      1.20 1.6545990    2861.5022           109.59
62         Germany 2004      0.70 1.7477213    2706.4983           123.29
65          Greece 2004      2.20 0.7093563    1067.3828            95.89
74         Iceland 2004      1.40 2.8747448    3714.6547           153.43
79         Ireland 2004      1.10 2.3244180    2668.2599           115.07
80          Israel 2004      1.66 1.3292692     880.5043           104.11
81           Italy 2004      1.10 1.3686293    1952.3460            84.93
83           Japan 2004      1.70 3.3492305    2352.0917            76.71
108          Malta 2004      1.60 0.3497126     854.7466           131.51
121    Netherlands 2004      0.80 1.8887101    2234.8590           142.47
123    New Zealand 2004      1.60 0.8600206    1571.7167           164.38
128         Norway 2004      1.70 3.4570313    4542.5184           120.55
137       Portugal 2004      1.50 0.5165410    1190.6540            93.15
155       Slovenia 2004      1.80 0.5103941    1032.8624            41.10
159          Spain 2004      1.12 0.9177162    1413.5989            93.15
164         Sweden 2004      1.00 2.4800819    2977.2338           128.77
165    Switzerland 2004      0.86 3.0590985    3213.2000           164.38
181 United Kingdom 2004      0.70 2.2012674    2392.0480           112.33
182  United States 2004      1.19 3.1556045    2785.6029           191.78
    LiteracyRate
9             NA
10            NA
16            NA
31            NA
46            NA
56            NA
57            NA
62            NA
65            NA
74            NA
79            NA
80            NA
81            NA
83            NA
108           NA
121           NA
123           NA
128           NA
137           NA
155     99.65247
159           NA
164           NA
165           NA
181           NA
182           NA
```

```
> btc[which(btc$HealthExpend < smm[2]-1.5*iqr),]
```

```
[1] Country          Year             BadTeeths        GDP
[5] HealthExpend     SugarComsumption LiteracyRate
<0 rows> (or 0-length row.names)
```

Equivalente al caso GDP pero mucho más extremado. Se normaliza.

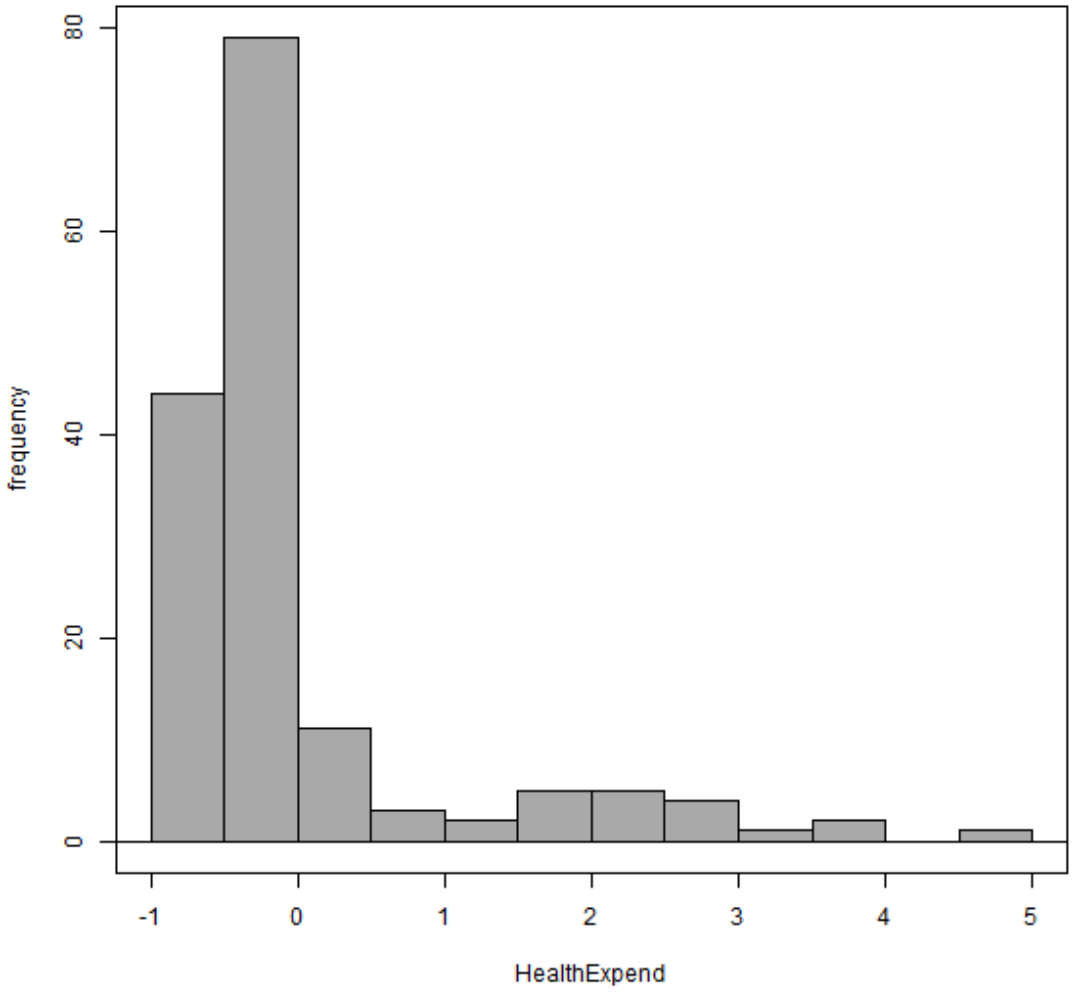
```
> btd<-btc
```

```
> btd$HealthExpend<- (btc$HealthExpend - mean(btc$HealthExpend)) / sd(btc$HealthExpend)
```

```
> summary(btd$HealthExpend)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-0.5206 -0.5041 -0.4152  0.0000 -0.1988  4.5329 
```

```
> with(btd, Hist(HealthExpend, scale="frequency", breaks="Sturges", col="darkgray"))
```



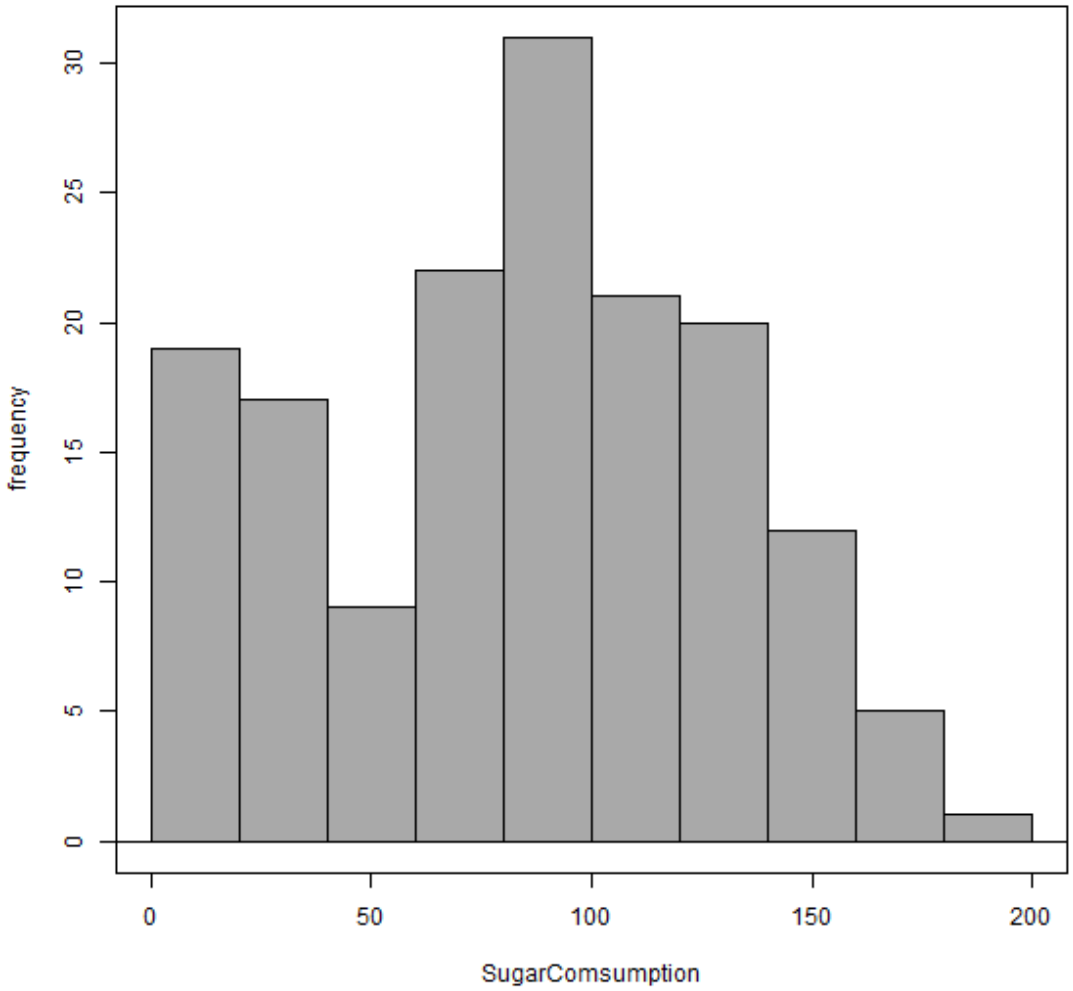
SugarComsumption

```
> smm<-summary(btd$SugarComsumption)
```

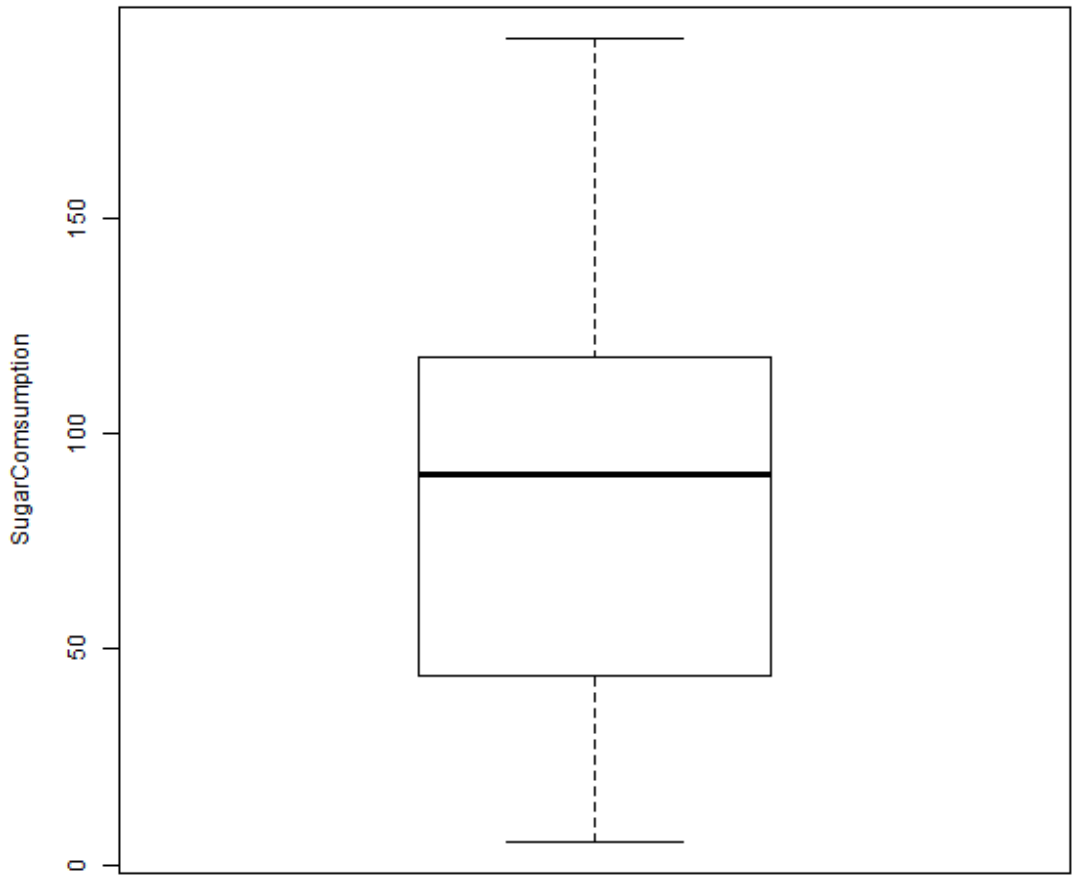
```
> smm
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.48   43.84   90.41   84.72  117.81  191.78
```

```
> with(btd, Hist(SugarComsumption, scale="frequency", breaks="Sturges", col="darkgray"))
```




```
> Boxplot( ~ SugarComsumption, data=btd, id.method="y")
```




```
> iqr<-IQR(btd$SugarComsumption)
```

```
> iqr
```

```
[1] 73.97
```

```
> var(btd$SugarComsumption)
```

```
[1] 2031.187
```

```
> sd(btd$SugarComsumption)
```

```
[1] 45.06869
```

```
> btc[which(btd$SugarComsumption > 1.5*iqr+smm[5]),]
```

```
[1] Country          Year             BadTeeths        GDP
[5] HealthExpend     SugarComsumption LiteracyRate
<0 rows> (or 0-length row.names)
```

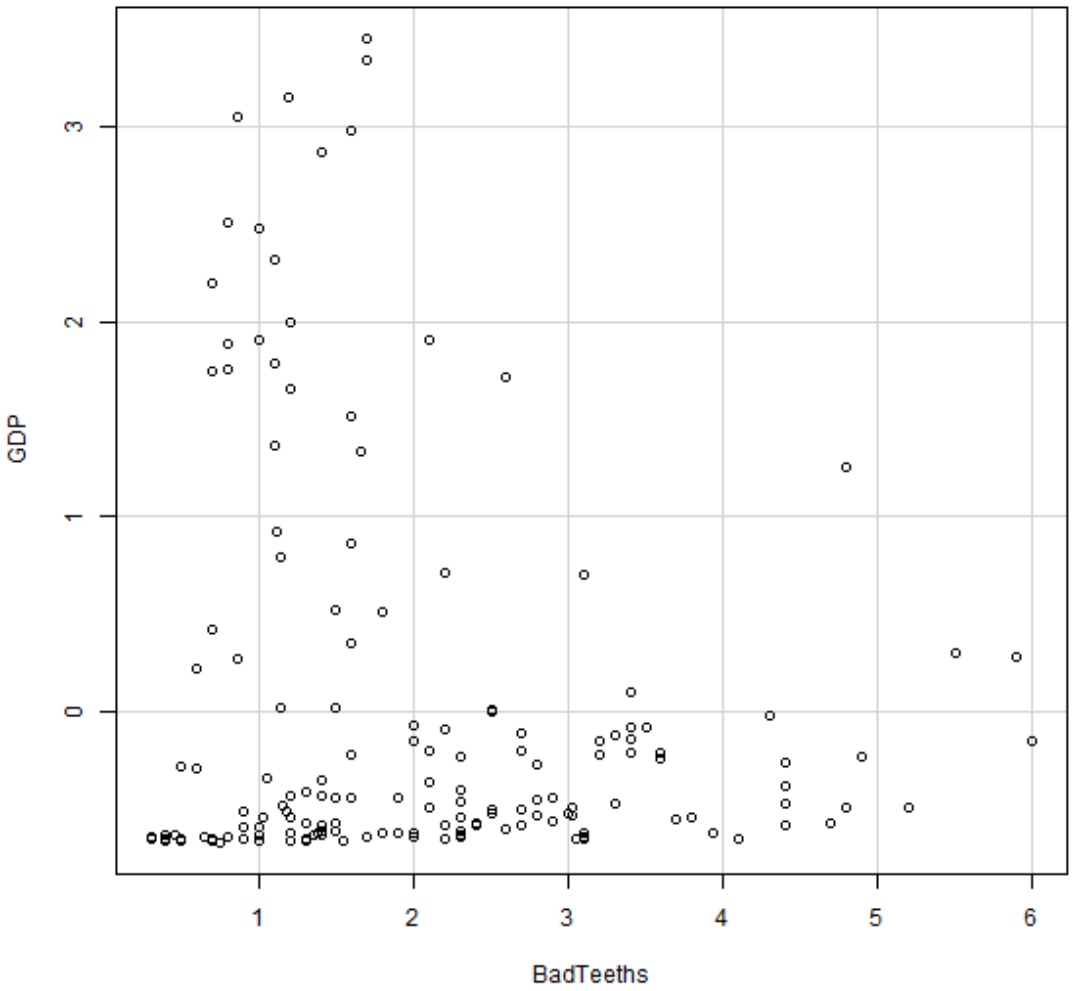
```
> btc[which(btd$SugarComsumption < smm[2]-1.5*iqr),]
```

```
[1] Country          Year             BadTeeths        GDP
[5] HealthExpend     SugarComsumption LiteracyRate
<0 rows> (or 0-length row.names)
```

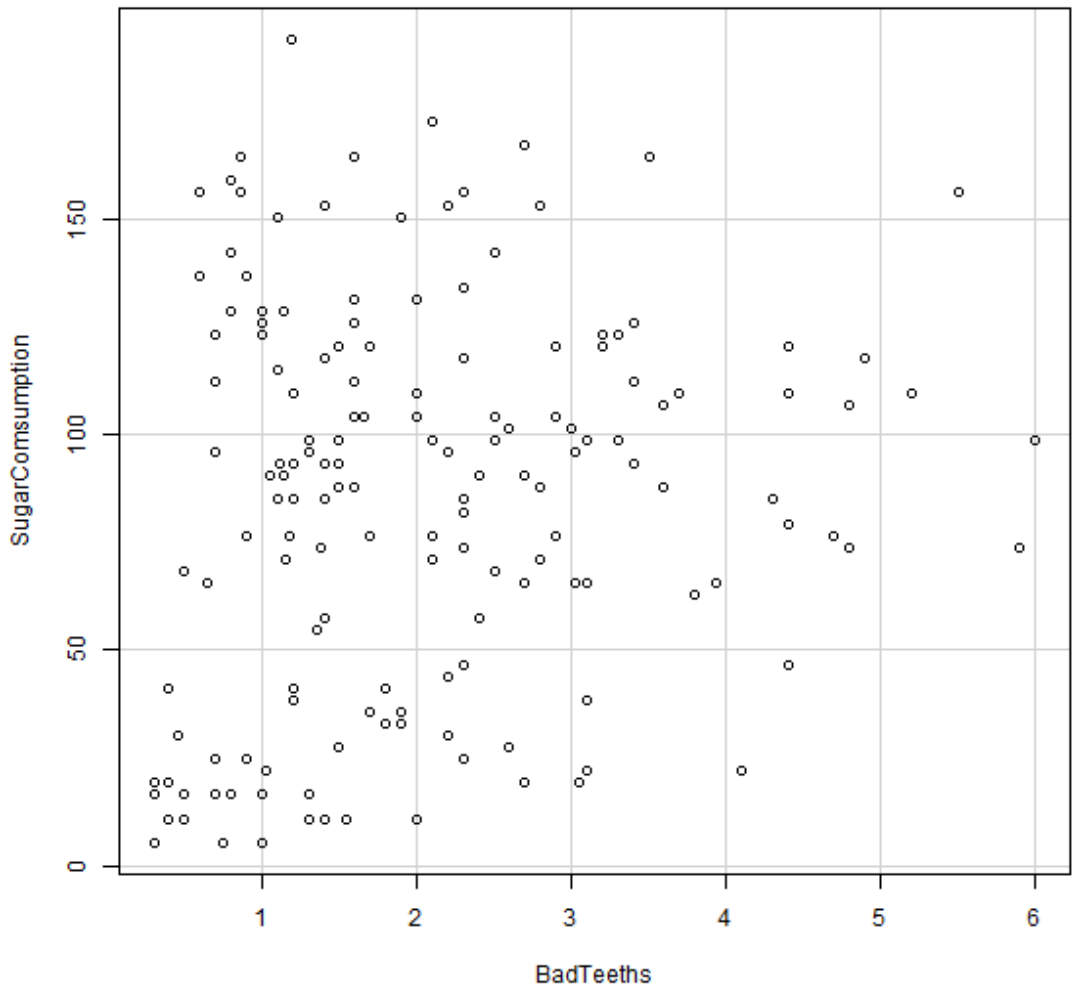
Sin outliers. Dispersión homogenea.

Plots con BadTeeths

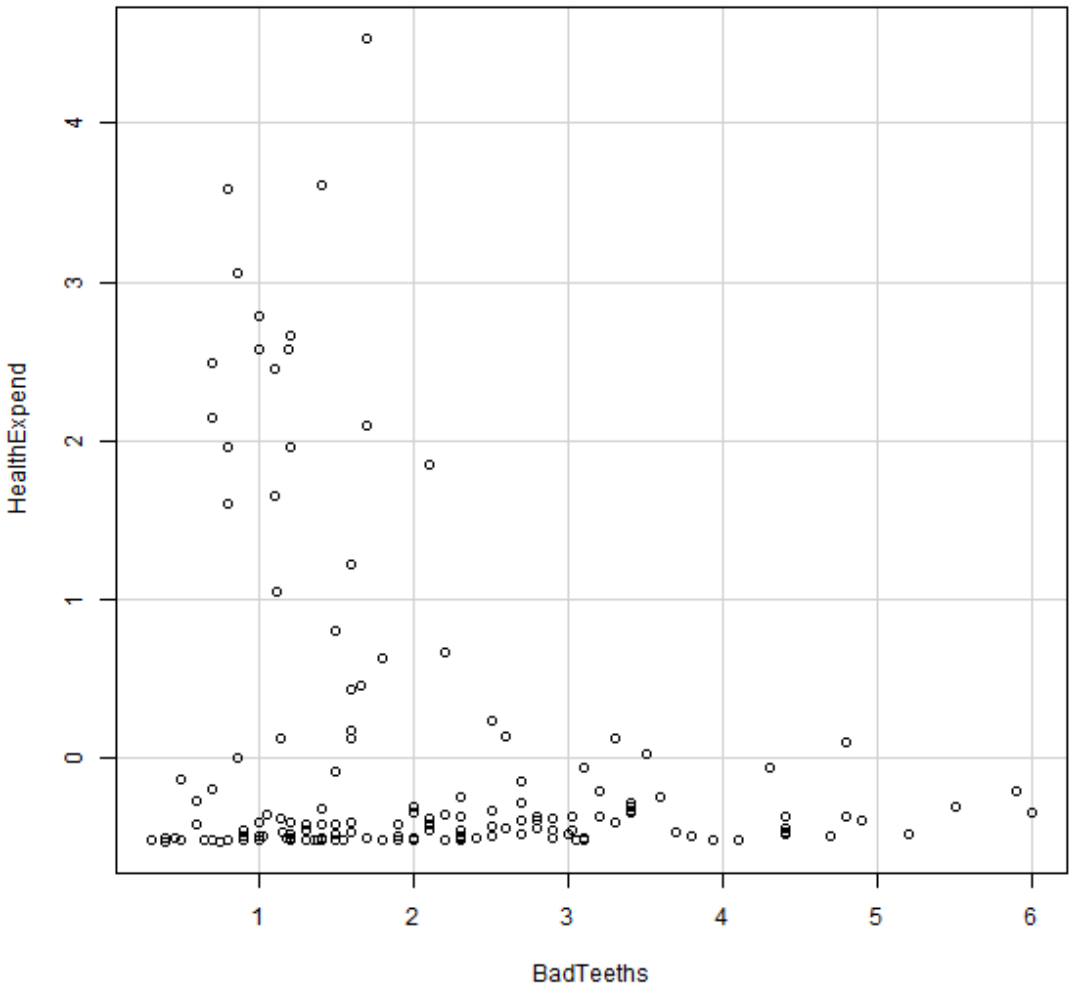
```
> scatterplot(GDP~BadTeeths, reg.line=FALSE, smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, el
+   levels=c(.5, .9), data=btd)
```



```
> scatterplot(SugarComsumption~BadTeeths, reg.line=FALSE, smooth=FALSE, spread=FALSE, boxplots=FALSE,
+   ellipse=FALSE, levels=c(.5, .9), data=btd)
```




```
> scatterplot(HealthExpend~BadTeeths, reg.line=FALSE, smooth=FALSE, spread=FALSE, boxplots=FALSE, spa
+    levels=c(.5, .9), data=btd)
```



Prueba de regresión lineal múltiple

```
> rlin <- lm(BadTeeths ~ GDP + HealthExpend + SugarComsumption, data=btd)
```

```
> summary(rlin)
```

```
Call:
lm(formula = BadTeeths ~ GDP + HealthExpend + SugarComsumption,
    data = btd)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3319 -0.8536 -0.1183  0.6573  3.7623

Coefficients:
                  Estimate Std. Error t value     Pr(>|t|)
(Intercept)       1.285081   0.226305   5.679 0.0000000664 ***
GDP               0.131646   0.239888   0.549     0.583957
HealthExpend     -0.615488   0.230505  -2.670     0.008401 **
SugarComsumption  0.009330   0.002436   3.831     0.000186 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.164 on 153 degrees of freedom
Multiple R-squared:  0.1539,    Adjusted R-squared:  0.1373
F-statistic: 9.273 on 3 and 153 DF,  p-value: 0.00001133
```

El coeficiente de bondad del ajuste es 0.1539, bajo, cuanto más cerca de 1 mejor es el modelo.

```
> rlin2 <- lm(BadTeeths ~ HealthExpend + SugarComsumption, data=btd)
```

```
> summary(rlin2)
```

```
Call:
lm(formula = BadTeeths ~ HealthExpend + SugarComsumption, data = btd)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3009 -0.8477 -0.1507  0.6474  3.8260

Coefficients:
                  Estimate Std. Error t value     Pr(>|t|)
(Intercept)       1.253347   0.218295   5.742 0.0000000485 ***
HealthExpend     -0.502982   0.105132  -4.784 0.0000039860 ***
SugarComsumption  0.009705   0.002333   4.160 0.0000526784 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.162 on 154 degrees of freedom
Multiple R-squared:  0.1522,    Adjusted R-squared:  0.1412
F-statistic: 13.82 on 2 and 154 DF,  p-value: 0.000003014
```

```
> rlin3 <- lm(BadTeeths ~ SugarComsumption, data=btd)
```

```
> summary(rlin3)
```

```
Call:
lm(formula = BadTeeths ~ SugarComsumption, data = btd)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7970 -0.9611 -0.2861  0.7824  3.8728

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.694338   0.211391   8.015 2.48e-13 ***
SugarComsumption 0.004499   0.002204   2.041   0.0429 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.241 on 155 degrees of freedom
Multiple R-squared:  0.02618,   Adjusted R-squared:  0.01989
F-statistic: 4.166 on 1 and 155 DF,  p-value: 0.04293
```

```
> rlin4 <- lm(BadTeeths ~ GDP, data=btd)
```

```
> summary(rlin4)
```

```
Call:
lm(formula = BadTeeths ~ GDP, data = btd)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9134 -0.8969 -0.1685  0.7061  3.8917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.07554    0.09894  20.978   <2e-16 ***
GDP         -0.20965    0.09926  -2.112   0.0363 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.24 on 155 degrees of freedom
Multiple R-squared:  0.02798,   Adjusted R-squared:  0.02171
F-statistic: 4.461 on 1 and 155 DF,  p-value: 0.03627
```

Evaluación del modelo y conclusiones

Estamos obteniendo unos coeficientes de bondad de los resultados muy bajos, se debería seguir investigando y analizando el caso para intentar mejorar.