Informe Suicidios

Ana Blanes Martinez - Xavier Castilla Carbonell

26/5/2019

Contents

1. Detalles de la actividad

1.1 Descipción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, dataset), orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Para resolver la actividad se integrará un informe donde esten unificados los conceptos teoricos, objetivos y resolución con código R.

El dataset usado se puede encontrar siguiendo el enlace.

1.2 Objetivos de la actividad

Los objetivos que queremos cumplir mediante el desarrollo de esta actividad son los siguientes:

- Aplicar conocimientos adquiridos en un ambito multidisciplinar en un entorno poco conocido
- Identificar datos relevantes y tratamientos necesarios
- Analizar los datos adecuadamente
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado
- Desarrollar las habilidades aprendidas
- Desarrollar la capacidad de búsqueda, gestión y uso de la información en el ámbito de la ciencia de datos

Para ello desarrollaremos con un caso práctico las seis etapas de un proyecto analítico en ciencia de datos, *identificar el problema*, *recopilar y almacenar datos* relacionados, *limpiar los datos*, *analizar*, *representar* y *responderemos a la pregunta planteada.*

1.3 Competencias

Así, las competencias del Máster en Data Science que se desarrollan son:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2. Resolución

2.1 Identificar el problema y objetivos del análisis

Los suicidios son uno de los principales problemas de la sanidad pública presente en todos los países del mundo independientemente de su nivel económico, población o ubicación geográfica. Con este estudio pretendemos desevelar que sectores de la población están más en riesgo considerando su *edad*, *sexo*, *renta per capita* y *continente* usando la información de distintos países a lo largo de los años.

La finalidad del análisis es generar un informe sobre el risgo de cada grupo de población con el fin de ayudar a enfocar futuros proyectos a identificar las causas del suicidio en los grupos más desfavorecidos

Para ello mostraremos y compararemos los distintos grupos de población en base a:

- Su sexo
- Su gdb_per_capita
- Su año
- Su HDI
- Su rango de edad

Además crearemos un modelo regresivo con los objetivos de:

- Completar los datos de HDI for year

2.2 Descripción del dataset

El siguiente dataset se ha construido a partir de distintos conjuntos de datos (referencias en la bibliografia) para su análisis. Se compone de 27820 registros con 12 campos. Los registros están almacenados en un fichero CSV llamado *suicides* extaído de la web *Kaggle*.

Estos son los campos que contiene:

- **country**: Nombre del país
- **year**: Año del registro
- **sex**: Genero de la población
- **age**: Grupo de edad de la población del país ese año
- **suicides_no**: Número de suicidios en el país ese año
- **population**: Habitantes del país ese año
- **suicides/100k pop** Numero de suicidios de la población por cada cien mil habitantes
- **country-year**: Identificador compuesto por los campos país y año
- **HDI for year**: (*Human Development Index*), es un indicador del desarrollo humano en base a la salud, la educación y la riqueza. El valor va de 0 a 1 siendo cuan más alto mejor.
- **gdp_for_year ($)**: (*Gross domestic product*) (PIB), expresa el valor monetario de la producción de bienes y servicios de un país
- **gdp_per_capita ($)**: relación entre el GDP y la cantidad de población de un país
- **generation**: Nombre de la generación a la que pertenece un grupo de población

2.3 Preparación de los datos

Instalamos las librerias que necesitaremos durante el desarrollo:

```
#Preparación del análisis
#install.packages("countrycode") ##Es necesario instalar el package para poder ejecutar el código
#install.packages("ggplot2")
library(readr)
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 3.4.4
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
library(plyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

El primer paso de nuestro analisis es realizar la lectura del fichero en formato CSV **suicide.csv** en el que se encuentran nuestros datos. El resultado devuelto devuelto por la función *read.csv* será un objeto *data.frame*:

```
#Lectura de datos
dfSuicidios <- read.csv("suicide.csv", encoding = "UTF-8", header = TRUE);
colnames(dfSuicidios)[colnames(dfSuicidios)=="X.U.FEFF.country"] <- "country"
```

Analicemos el dataset para conocer que tipo de información contiene y como está distribuída:

```
#Informe de los datos
summary(dfSuicidios)
```

```
##         country           year          sex                age
##  Austria    :  382   Min.   :1985   female:13910   15-24 years:4642
##  Iceland    :  382   1st Qu.:1995   male  :13910   25-34 years:4642
##  Mauritius  :  382   Median :2002                  35-54 years:4642
##  Netherlands:  382   Mean   :2001                  5-14 years :4610
##  Argentina  :  372   3rd Qu.:2008                  55-74 years:4642
##  Belgium    :  372   Max.   :2016                  75+ years  :4642
##  (Other)    :25548
##   suicides_no        population       suicides.100k.pop
##  Min.   :    0.0   Min.   :     278   Min.   :  0.00
##  1st Qu.:    3.0   1st Qu.:   97498   1st Qu.:  0.92
```

```
##  Median :   25.0   Median :  430150   Median :  5.99
##  Mean   :  242.6   Mean   : 1844794   Mean   : 12.82
##  3rd Qu.:  131.0   3rd Qu.: 1486143   3rd Qu.: 16.62
##  Max.   :22338.0   Max.   :43805214   Max.   :224.97
##
##       country.year    HDI.for.year           gdp_for_year....
##  Albania1987:   12   Min.   :0.483   1,002,219,052,968:   12
##  Albania1988:   12   1st Qu.:0.713   1,011,797,457,139:   12
##  Albania1989:   12   Median :0.779   1,016,418,229    :   12
##  Albania1992:   12   Mean   :0.777   1,018,847,043,277:   12
##  Albania1993:   12   3rd Qu.:0.855   1,022,191,296    :   12
##  Albania1994:   12   Max.   :0.944   1,023,196,003,075:   12
##  (Other)    :27748   NA's   :19456   (Other)          :27748
##  gdp_per_capita....           generation
##  Min.   :   251     Boomers        :4990
##  1st Qu.:  3447     G.I. Generation:2744
##  Median :  9372     Generation X   :6408
##  Mean   : 16866     Generation Z   :1470
##  3rd Qu.: 24874     Millenials     :5844
##  Max.   :126352     Silent         :6364
##
```

Una rápida inspección no permite ver que algunas de las columnas del dataset no son necesarias pues son el resultado de una operación entre otras columnas, tenemos **Country-year**, **suicides/100k pop** y **gdp_per_capita**.

Los casos de *suicides/100k pop* y *gdp_per_capita* son interesantes pues presentan una mediacion de los suicidios y el nivel de via mejor para realizar comparaciones ya que hay países con valores de población muy dispares (max: 43805214, min: 278) y lo mismo con el valor del PIB (max: 1.812e+13, min: 4.692e+07)

Por eso eliminaremos las columnas: Country-year, suicides_no y gdp_for_year ($).

Y trabajaremos con las columnas: *country*, *year*, *sex*, *age*, *suicides_100k_pop*, *HDI_for_year*, *gdp_per_capita_USDollar* y *generation* y agregaremos la columna *continent* para ubicar los distintos países en zonas geográficas.

Analicemos cada una de estas variables:

- **Country**

```r
#Informe de country
summary(dfSuicidios$country)
```

```
##                     Austria                     Iceland
##                         382                         382
##                   Mauritius                 Netherlands
##                         382                         382
##                   Argentina                     Belgium
##                         372                         372
##                      Brazil                       Chile
##                         372                         372
##                    Colombia                     Ecuador
##                         372                         372
##                      Greece                      Israel
##                         372                         372
##                       Italy                       Japan
##                         372                         372
##                  Luxembourg                       Malta
```

```
##                         372                         372
##                      Mexico                 Puerto Rico
##                         372                         372
##           Republic of Korea                   Singapore
##                         372                         372
##                       Spain                United Kingdom
##                         372                         372
##               United States                   Australia
##                         372                         360
##                    Bulgaria                  Costa Rica
##                         360                         360
##                      France                   Guatemala
##                         360                         360
##                     Ireland                      Norway
##                         360                         360
##                      Sweden                      Canada
##                         358                         348
##                     Finland                 New Zealand
##                         348                         348
##                Turkmenistan                      Belize
##                         348                         336
##                 Saint Lucia                    Suriname
##                         336                         336
##                     Ukraine                     Uruguay
##                         336                         336
##                     Romania                    Thailand
##                         334                         334
##         Antigua and Barbuda                    Paraguay
##                         324                         324
##                    Portugal          Russian Federation
##                         324                         324
##         Trinidad and Tobago               Czech Republic
##                         324                         322
##                     Germany                  Kazakhstan
##                         312                         312
##                   Kyrgyzstan                     Grenada
##                         312                         310
##                     Hungary                    Barbados
##                         310                         300
##                      Guyana                      Kuwait
##                         300                         300
##                      Panama Saint Vincent and Grenadines
##                         300                         300
##                     Armenia                        Cuba
##                         298                         288
##                 El Salvador                      Poland
##                         288                         288
##                     Bahamas                     Albania
##                         276                         264
##                     Denmark                     Georgia
##                         264                         264
##                    Slovakia                  Uzbekistan
##                         264                         264
##                     Croatia                   Lithuania
```

```
##                         262                         262
##                     Bahrain                     Belarus
##                         252                         252
##                     Estonia                      Latvia
##                         252                         252
##                    Slovenia                 Switzerland
##                         252                         252
##                South Africa                      Serbia
##                         240                         216
##                  Seychelles                     Jamaica
##                         216                         204
##                  Azerbaijan                 Philippines
##                         192                         180
##                      Cyprus                       Qatar
##                         178                         178
##                       Aruba                        Fiji
##                         168                         132
##                    Kiribati                   Sri Lanka
##                         132                         132
##                    Maldives                  Montenegro
##                         120                         120
##                      Turkey                   Nicaragua
##                          84                          72
##       United Arab Emirates                        Oman
##                          72                          36
##      Saint Kitts and Nevis                   San Marino
##                          36                          36
##      Bosnia and Herzegovina                  Cabo Verde
##                          24                          12
##                    Dominica                     (Other)
##                          12                          22
```

- **Year**

```
#Informe de country
summary(dfSuicidios$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1985    1995    2002    2001    2008    2016
```

- **Sex**

```
#Informe de country
summary(dfSuicidios$sex)
```

```
## female   male
##  13910  13910
```

- **Age**

```
#Informe de country
summary(dfSuicidios$age)
```

```
## 15-24 years 25-34 years 35-54 years  5-14 years 55-74 years   75+ years
##        4642        4642        4642        4610        4642        4642
```

- **suicides_100k_pop**

```
#Informe de country
summary(dfSuicidios$suicides.100k.pop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.92    5.99   12.82   16.62  224.97
```

- **HDI_for_year**

```
#Informe de country
summary(dfSuicidios$HDI.for.year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.483   0.713   0.779   0.777   0.855   0.944   19456
```

- **gdp_per_capita_USDollar**

```
#Informe de country
summary(dfSuicidios$gdp_per_capita....)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     251    3447    9372   16866   24874  126352
```

El campo HDI_for_year requiere un tratamiento especial pues contiene Na's, lo que haremos será completar este dato estimando la media entre los dos valores más cercanos (uno de años anteriores y otro de años posteriores).

2.4 Limpieza de los datos

El primer paso será crear la nueva columna **continent** mediante el campo *country* y usando la libreria *countrycode*:

```
dfSuicidios$continent <- countrycode(sourcevar = dfSuicidios[, "country"], origin = "country.name", des
```

Renombrarmos las columnas con espacios y simbolos para que sean más fáciles de manejar:

```
colnames(dfSuicidios)[colnames(dfSuicidios)=="gdp_per_capita...."] <- "gdp_per_capita_USDollar"
colnames(dfSuicidios)[colnames(dfSuicidios)=="gdp_for_year...."] <- "gdp_for_year_USDollar"
colnames(dfSuicidios)[colnames(dfSuicidios)=="HDI.for.year"] <- "HDI_for_year"
colnames(dfSuicidios)[colnames(dfSuicidios)=="country.year"] <- "country_year"
colnames(dfSuicidios)[colnames(dfSuicidios)=="suicides.100k.pop"] <- "suicides_100k_pop"
```

Y nos quedamos con las columnas deseadas:

```
keeps <- c("country","continent", "year","sex","age","suicides_100k_pop","HDI_for_year","gdp_per_capita
dfSuicidiosFiltered <- dfSuicidios[keeps]
```

Si hay algo que destaca en el dataset es que esta columna HDI, a pesar de su importancia, presenta muchos valores NA y esto se convierte en un problema para el análisis que queremos realizar. Si observamos el dataset también vemos que los datos se presentan cada 5 años, y es en esos huecos en los que aparece NA. Como técnica de tratamiento de los datos hemos optado para esos huecos rellenarlos con los valores medios a partir de los datos existentes anterior y posterior para cada uno de los países. Aunque existen en R diferentes librerías para la realización de este tipo de acciones, hemos optado por un desarrollo ad-hoc que se ajuste perfectamente a nuestras necesidades. La función es la siguiente:

```
completeHDI_For_YearDatoPosteriorMedia <- function(df){
        j <- 0;
        country <- "";
        vectorNA = c();
        lastValue <- NA
```

```
        for(i in 1:nrow(df)) {

            if(i> 1 && (df[i-1,]$country !=df[i,]$country)) {
                j <- 0;
                vectorNA = c();
                lastValue <- NA
            }


            if(is.na(df[i,]$HDI_for_year)) {
                vectorNA <- c(vectorNA,i);
                j <- j+1;
            } else {

                if(is.na(lastValue)) {
                    lastValue <-df[i,]$HDI_for_year
                }

                if(length(vectorNA)>0) {
                    for(value in vectorNA) {
                        hdi <- df[i,]$HDI_for_year
                        if(!is.na(lastValue)) {
                            hdi = (lastValue + df[i,]$HDI_for_year)/2

                        }

                        df[value,]$HDI_for_year <- hdi;
                        vectorNA = c();

                        j<-0;
                    }
                    lastValue <- df[i,]$HDI_for_year
                }
            }
        }
        return (df);

    }
```

Completamos los datos de *HDI_For_Year*

```
dfSuicidiosFiltered <- completeHDI_For_YearDatoPosteriorMedia(dfSuicidiosFiltered)
```

Agrupamos los países en periodos de 5 años, esto lo hacemos porque nos faltan algunos registros y así tenemos toda la información mediante las medias.

```
dfSuicidiosFiltered$yearRange5Years<-cut(dfSuicidiosFiltered$year, c(1980,1985,1990,1995,2000,2005,2010

# Muestra medias de suicidios, gdp, hdi agrupoado por country, continent, yearRange5Years. Si se quiere
dfSuicidiosFiltered2 <- ddply(dfSuicidiosFiltered, .(country, continent,yearRange5Years), summarize,  m
```

Para finalizar comprobamos que tenemos el subset adecuado

```
str(dfSuicidiosFiltered2)
```

```
## 'data.frame':    580 obs. of  6 variables:
```

```
##  $ country                    : Factor w/ 101 levels "Albania","Antigua and Barbuda",..: 1 1 1 1 1 1
##  $ continent                  : chr  "Europe" "Europe" "Europe" "Europe" ...
##  $ yearRange5Years            : Factor w/ 8 levels "(1980,1985]",..: 2 3 4 5 6 1 2 3 4 5 ...
##  $ mean_suicides_100k_pop     : num  2.71 2.57 4.8 3.95 2.98 ...
##  $ mean_HDI_for_year          : num  0.619 0.619 0.641 0.679 0.711 ...
##  $ mean_gdp_per_capita_USDollar: num  799 555 1049 2104 4103 ...
```

2.5 Análisis de los datos

Analisis de suicidios por población

La columna suicides_100k_pop nos informa de la tasa de suicidios por cada cien mil personas (S100K), por lo que es una variable que nos permite comparar los disitntos registros en una misma escala.

Calculemos la media, la mediana y la desviación estandar de S100K:

```
mean(dfSuicidiosFiltered$suicides_100k_pop)
```

```
## [1] 12.8161
```

```
median(dfSuicidiosFiltered$suicides_100k_pop)
```

```
## [1] 5.99
```

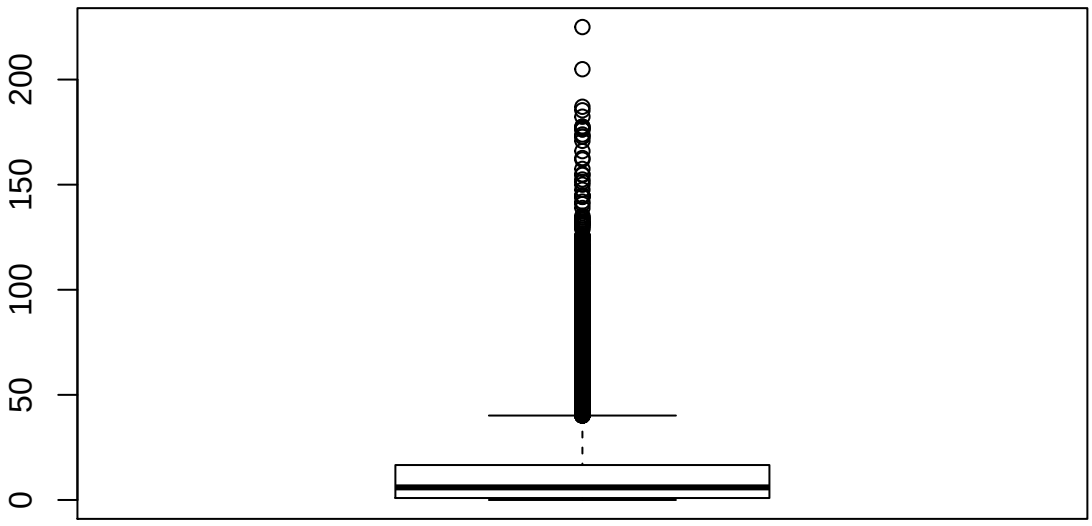
```
sd(dfSuicidiosFiltered$suicides_100k_pop)
```

```
## [1] 18.96151
```

La media es bastante más grande que la mediana lo que nos indica que los valores elevados de la columna son muy elevados respecto a los valores más bajos. Esto se ve confirmado por la desviación estándar que es muy elevada, lo que indica que el rango de valores es muy disperso.

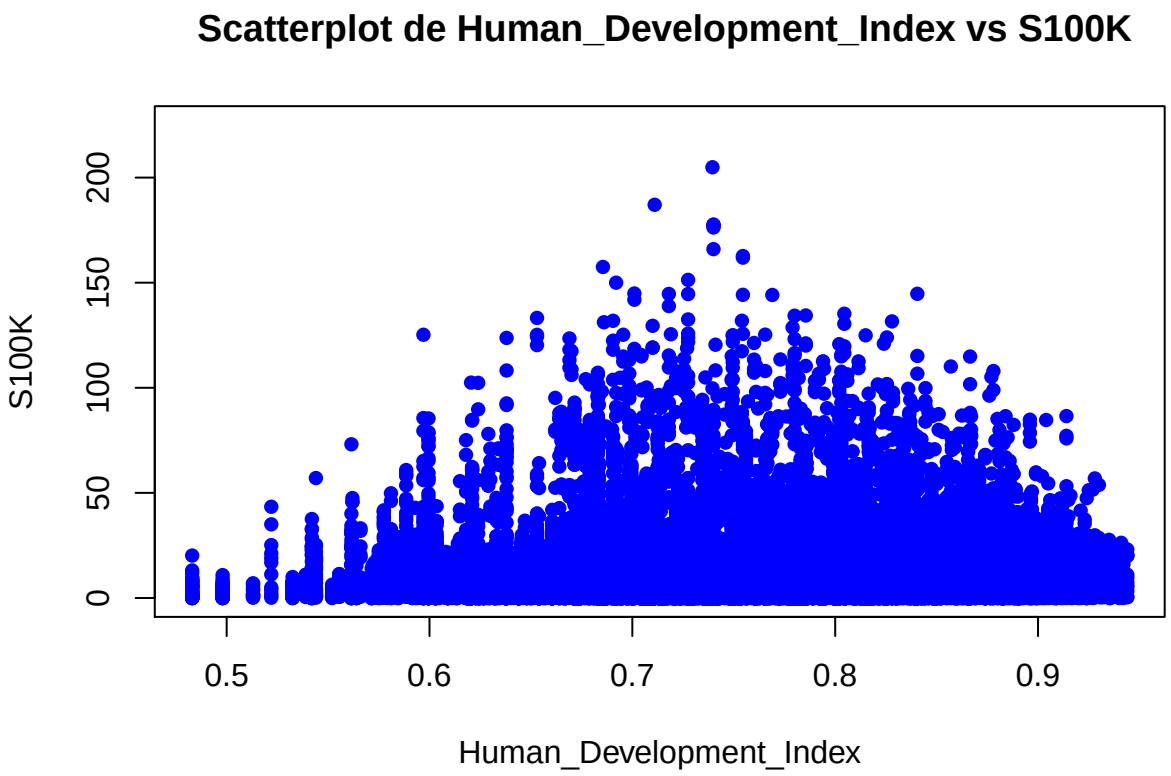
Por eso si representamos los datos en un boxplot nos queda la caja aplastada en el fondo, la mayoria de los datos son bajos pero los valores extremos son muy elevados. Aún así no creemos que sea necesario eliminar estos valores extremos ya que no dejan de ser valores reales (provinientes de fuentes como la *Organización Mundial de la Salud*) y el objetivo de nuestro estudio es identificar las causas que llevan al suicidio y no podemos obviar conjuntos de datos donde se muestra la mayor tasa de mortalidad.

```
boxplot(dfSuicidiosFiltered$suicides_100k_pop)
```



Visualicemos la relación de estos datos con el resto de variables (exceptuando Country y Sex)

```r
# Scatterplot relacionando Human_Development_Index y S100K
plot(dfSuicidiosFiltered$HDI_for_year, dfSuicidiosFiltered$suicides_100k_pop, xlab = "Human_Development_
     main = " Scatterplot de Human_Development_Index vs S100K")
```



Con este gráfico relacionando el HDI con los suicidios vemos que forma una campana de gauss desviada hacia la derecha. Los países menos desarrollados (HDI < 0.55) están en menor riesgo de suicidarse y ese riesgo va aumentando hasta un punto de inflexión (HDI > 0.75) donde la tónica es de descenso. Esto nos indica que las causas de suicidio pueden estar condicionadas por los factores que alteran el desarrollo de un país hasta que alcanza cierto punto en que la mejora estatal hace descender la tasa de suicidios. En este caso podemos identificar que los países en desarrollo son los que están más en riesgo.

Comprobaremos las observaciones con la hipotesis nula de que no existe relacióne entre las dos variables observadas con una significancia del 0.05.

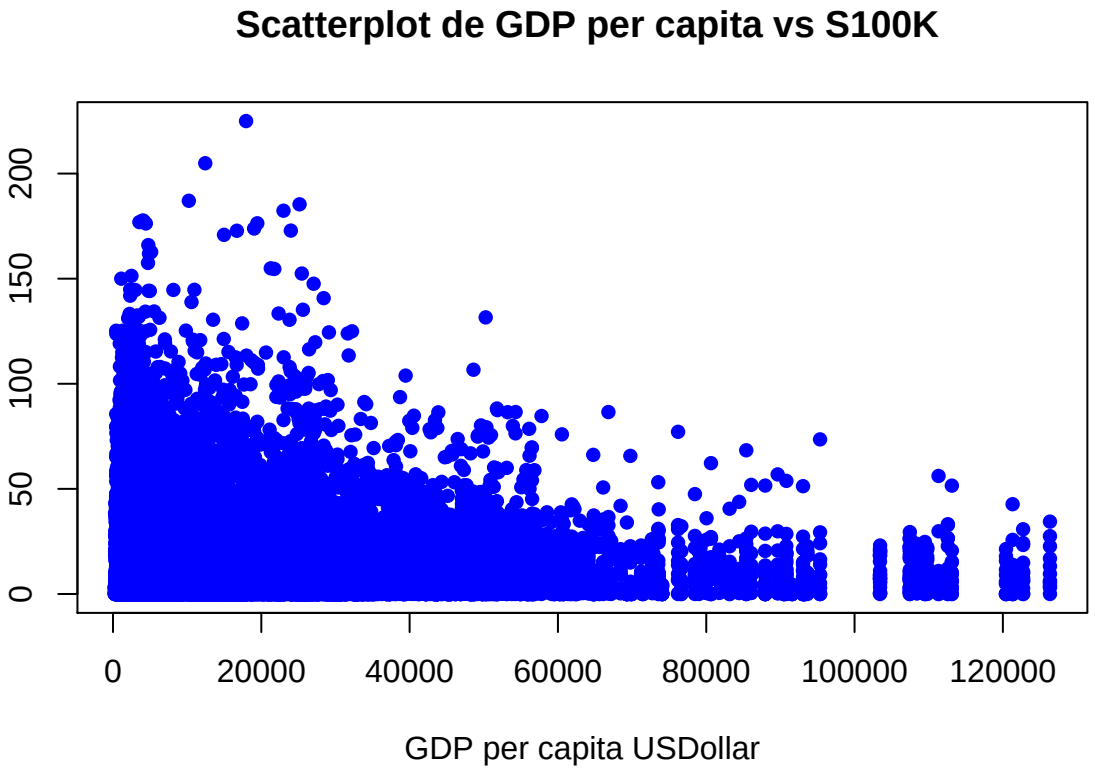
```r
HDI.lm = lm(dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$HDI_for_year, data=faithful)
summary(HDI.lm)
```

```
##
## Call:
## lm(formula = dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$HDI_for_year,
##     data = faithful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.580 -10.898  -6.561   3.656 192.708
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -0.0876     0.9429  -0.093    0.926
## dfSuicidiosFiltered$HDI_for_year  16.6324     1.2345  13.473   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.27 on 25318 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.007119,   Adjusted R-squared:  0.00708
## F-statistic: 181.5 on 1 and 25318 DF,  p-value: < 2.2e-16
```

Como el p-value es menor a 0.05 rechazamos la hipotesis nula y validamos que existe una fuerte relación entre las dos variables.

```
# Scatterplot relacionando gdp_per_capita y S100K
plot(dfSuicidiosFiltered$gdp_per_capita_USDollar, dfSuicidiosFiltered$suicides_100k_pop, xlab = "GDP pe
     main = " Scatterplot de GDP per capita vs S100K")
```



En el gráfico que relaciona el PIB per capita con la tasa de suicidios apreciamos que hay un rápido descenso de los suicidios a medida que aumenta la renta de los individuos pero sobre los 60K de renta la caída se convierte en una linea recta.

Esto nos indica que el nivel de vida es una de las causas principales de suicidio entre las rentas más bajas pero se vuelve más indiferente cuan más alto es, indicando que existen otros factores a tener en cuenta.

Comprobaremos las observaciones con la hipotesis nula de que no existe relacióne entre las dos variables observadas con una significancia del 0.05.

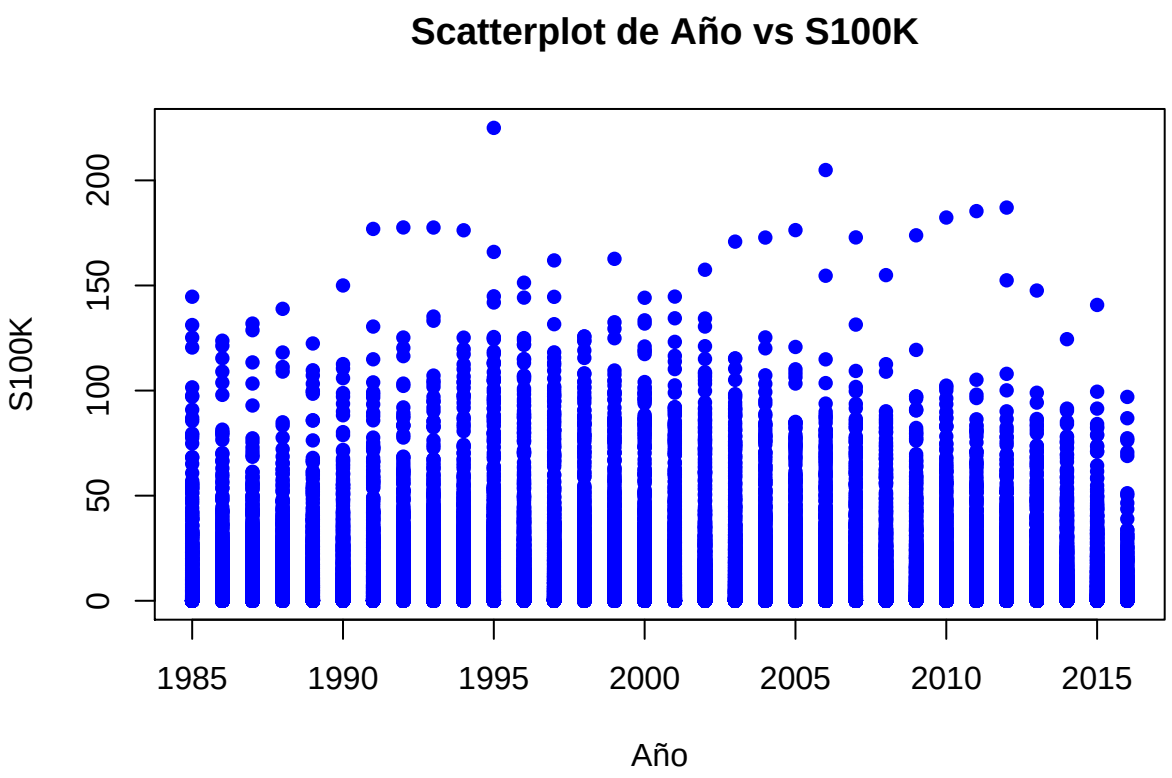
```
GDP.lm = lm(dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$gdp_per_capita_USDollar, data=fa
summary(GDP.lm)
```

```
##
## Call:
## lm(formula = dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$gdp_per_capita_USDollar,
##     data = faithful)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -13.012 -11.894  -6.827   3.802 212.152
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 1.279e+01  1.524e-01  83.888
## dfSuicidiosFiltered$gdp_per_capita_USDollar 1.792e-06  6.019e-06   0.298
##                                             Pr(>|t|)
## (Intercept)                                   <2e-16 ***
## dfSuicidiosFiltered$gdp_per_capita_USDollar    0.766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.96 on 27818 degrees of freedom
## Multiple R-squared:  3.187e-06,  Adjusted R-squared:  -3.276e-05
## F-statistic: 0.08865 on 1 and 27818 DF,  p-value: 0.7659
```

Como el p-value es mayor que 0.05 validamos la hipotesis nula por lo que no existe una fuerte relación entre las dos variables y no es un medidor adecuado a diferencia de nuestra observación original.

```
# Scatterplot relacionando año y S100K
plot(dfSuicidiosFiltered$year, dfSuicidiosFiltered$suicides_100k_pop, xlab = "Año", ylab = "S100K", pch
     main = " Scatterplot de Año vs S100K")
```



En el gráfico que relaciona los suicidios con los años apreciamos que no hay relación alguna entre las dos variables y los valores que destacan se deberán a factores alternativos.

Comprobaremos las observaciones con la hipotesis nula de que no existe relacióne entre las dos variables observadas con una significancia del 0.05.

```
YEAR.lm = lm(dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$year, data=faithful)
summary(YEAR.lm)
```

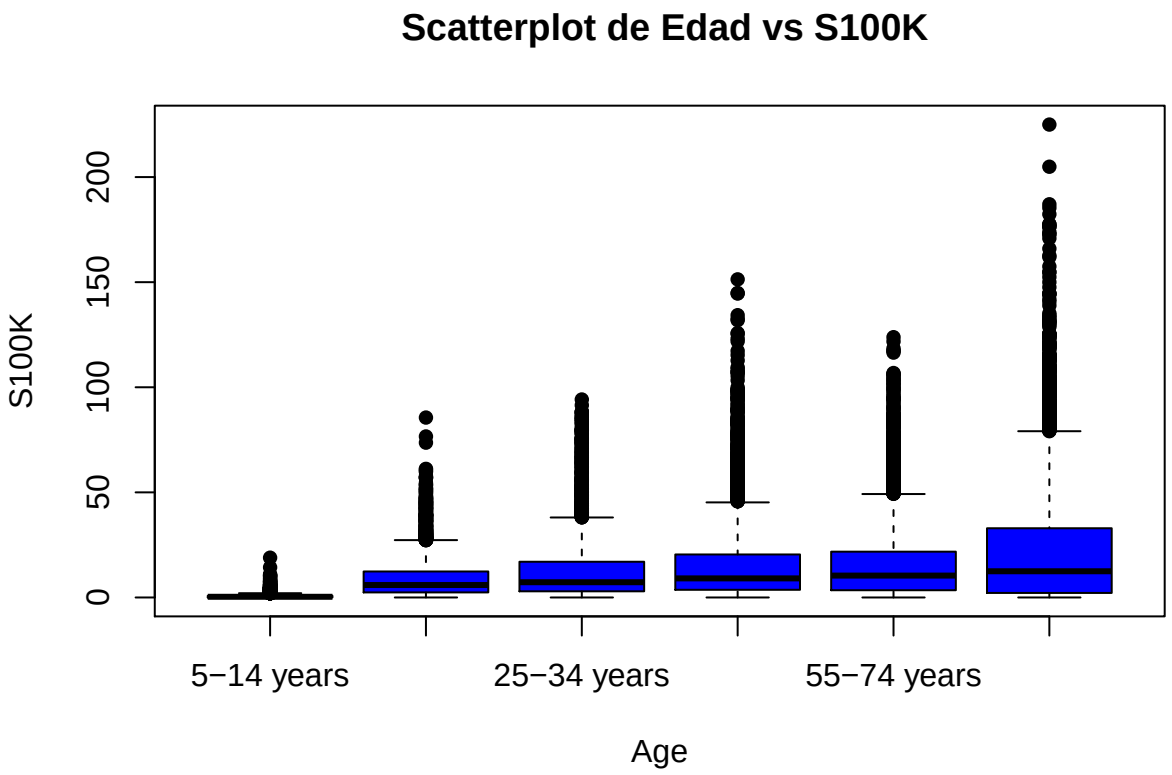
```
##
## Call:
## lm(formula = dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$year,
##     data = faithful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.237 -11.682  -6.867   3.802 211.607
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              187.72638   26.84423   6.993 2.75e-12 ***
## dfSuicidiosFiltered$year  -0.08740    0.01341  -6.516 7.35e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.95 on 27818 degrees of freedom
## Multiple R-squared:  0.001524,   Adjusted R-squared:  0.001488
## F-statistic: 42.46 on 1 and 27818 DF,  p-value: 7.353e-11
```

En este caso, se rechaza la hipotesis nula y se encuentra una fuerte relación entre el año y la tasa de suicidios

```
levels((dfSuicidios$age))
```

```
## [1] "15-24 years" "25-34 years" "35-54 years" "5-14 years"  "55-74 years"
## [6] "75+ years"
```

```
# Scatterplot relacionando Edad y S100K
plot(relevel(dfSuicidiosFiltered$age, "5-14 years"), dfSuicidiosFiltered$suicides_100k_pop, xlab = "Age
     main = " Scatterplot de Edad vs S100K")
```



Observamos en el conjunto de boxplots para cada rango de edad ascendente que existe una tendencia al

alza de suicidios a medida que la edad de la población aumenta así como los casos extremos que también aumentan con la edad.

Comprobaremos las observaciones con la hipotesis nula de que no existe relacióne entre las dos variables observadas con una significancia del 0.05.

```
AGE.lm = lm(dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$age, data=faithful)
summary(AGE.lm)
```

```
##
## Call:
## lm(formula = dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$age,
##     data = faithful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.955  -9.327  -2.087   2.493 201.015
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         8.9472     0.2580  34.675   <2e-16
## dfSuicidiosFiltered$age25-34 years  3.2397     0.3649   8.878   <2e-16
## dfSuicidiosFiltered$age35-54 years  6.0003     0.3649  16.443   <2e-16
## dfSuicidiosFiltered$age5-14 years  -8.3271     0.3655 -22.780   <2e-16
## dfSuicidiosFiltered$age55-74 years  7.2084     0.3649  19.754   <2e-16
## dfSuicidiosFiltered$age75+ years   15.0083     0.3649  41.129   <2e-16
##
## (Intercept)                        ***
## dfSuicidiosFiltered$age25-34 years ***
## dfSuicidiosFiltered$age35-54 years ***
## dfSuicidiosFiltered$age5-14 years  ***
## dfSuicidiosFiltered$age55-74 years ***
## dfSuicidiosFiltered$age75+ years   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.58 on 27814 degrees of freedom
## Multiple R-squared:  0.1406, Adjusted R-squared:  0.1404
## F-statistic: 909.8 on 5 and 27814 DF,  p-value: < 2.2e-16
```

En este caso, se rechaza la hipotesis nula y se encuentra una fuerte relación entre la edad y la tasa de suicidios

Ahora comprobemos la relación entre los S100k y todas las variables del dataset que hemos visualizado.

```
TOTAL.lm = lm(dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$year + dfSuicidiosFiltered$gdp
summary(TOTAL.lm)
```

```
##
## Call:
## lm(formula = dfSuicidiosFiltered$suicides_100k_pop ~ dfSuicidiosFiltered$year +
##     dfSuicidiosFiltered$gdp_per_capita_USDollar + dfSuicidiosFiltered$HDI_for_year +
##     dfSuicidiosFiltered$age, data = faithful)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.393  -9.309  -2.784   3.366 182.788
##
```

```
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             3.199e+02  2.725e+01   11.74
## dfSuicidiosFiltered$year                -1.698e-01  1.371e-02  -12.39
## dfSuicidiosFiltered$gdp_per_capita_USDollar -1.182e-04  8.856e-06  -13.34
## dfSuicidiosFiltered$HDI_for_year         4.047e+01  1.814e+00   22.31
## dfSuicidiosFiltered$age25-34 years       3.088e+00  3.654e-01    8.45
## dfSuicidiosFiltered$age35-54 years       5.858e+00  3.654e-01   16.03
## dfSuicidiosFiltered$age5-14 years       -8.252e+00  3.654e-01  -22.58
## dfSuicidiosFiltered$age55-74 years       6.927e+00  3.654e-01   18.96
## dfSuicidiosFiltered$age75+ years         1.440e+01  3.654e-01   39.41
##                                         Pr(>|t|)
## (Intercept)                               <2e-16 ***
## dfSuicidiosFiltered$year                  <2e-16 ***
## dfSuicidiosFiltered$gdp_per_capita_USDollar   <2e-16 ***
## dfSuicidiosFiltered$HDI_for_year          <2e-16 ***
## dfSuicidiosFiltered$age25-34 years        <2e-16 ***
## dfSuicidiosFiltered$age35-54 years        <2e-16 ***
## dfSuicidiosFiltered$age5-14 years         <2e-16 ***
## dfSuicidiosFiltered$age55-74 years        <2e-16 ***
## dfSuicidiosFiltered$age75+ years          <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.79 on 25311 degrees of freedom
##   (2500 observations deleted due to missingness)
## Multiple R-squared:  0.1616, Adjusted R-squared:  0.1614
## F-statistic:   610 on 8 and 25311 DF,  p-value: < 2.2e-16
```

Como podemos observar todas las variables presentan una alta relación con la tasa de suicidios, con un p-value mucho menor de 0.05, es interesante el caso de *gdp_per_capita_USDollar* que por si mismo no presentaba una relación con la tasa de suicidios que si es significativa dentro del conjunto de variables.

Esto es debido a que la causa de suicidios no se puede describir correctamente como causa de un solo factor si no de la relación de distintos factores entrelazados.

Análisis por genero

Analizamos el dataset y de las columnas disponibles, detectamos las que necesitamos para esta acción específica. Seleccionamos las columnas: "country", "year","sex","suicides_no", es decir, el país, el año, el sexo y el número de suicidios registrados.

De manera previa al análisis observamos los diferentes valores de cara a identificar valores nulos que puedan desvirtuar nuestros resultados. Valores NA no hay pero sí que vemos la existencia de valores "0", por tanto vamos a hacer un estudio previo para clarificar que los valores 0 son valores 0 reales (no hubo suicidios para ese país-año-franja de edad concreto) o estamos hablando de pérdida de datos.

Para ello primero seleccionamos todos los países que tienen valores 0 en la columna suicidios_no. Aquí observamos que hay muchos países, pero obviamente no es improbable que haya países que con franjas de edad 5-14 años o +75 en los que no se haya producido suicidios en algún año, así que eliminamos estas dos franjas de nuestro estudio preliminar.

Una vez realizados estos pasos, obtenemos esta lista:

Albania, Antigua and Barbuda, Armenia, Aruba, Azerbaijan, Bahamas,Bahrain, Barbados, Belize, Bosnia and Herzegovina, Cabo Verde, Costa Rica, Cyprus, Dominica, Fiji, Georgia, Greece, Grenada, Guatemala, Guyana, Iceland, Jamaica, Kiribati, Kuwait, Luxembourg, Macau, Maldives,Malta, Mauritius, Montenegro,

Oman, Panama, Paraguay,Qatar,Saint Kitts and Nevis, Saint Lucia, Saint Vincent and Grenadines, San Marino, Seychelles, Slovakia, Suriname,Trinidad and Tobago,Turkmenistan,United Arab Emirates

Es posible que los ceros se deban a datos perdidos, pero lo cierto es que todos estos países que presentan valores 0 son países de poca extensión y con poca población, por tanto puede encajar que haya alguna franja de edad de las todavía disponibles para algún año que no haya habido suicidios.

De manera que concluimos que los valores 0 que figuran en el dataset se deben a que el número de suicidios para ese país, franja de edad, año ha sido 0.

Para facilitar el análisis de los datos posterior, creamos la columna yearRange5Years que discretiza la columna year en valores de 5 en 5 años.

```
keeps <- c("country", "year","sex","suicides_no")
dfSuicidiosFilteredGenre <- dfSuicidios[keeps]
```

```
dfSuicidiosFilteredGenre$yearRange5Years<-cut(dfSuicidiosFilteredGenre$year, c(1980,1985,1990,1995,2000
```

Como queremos hacer el estudio con todos los países del mundo, realizamos una suma agrupada por únicamente por sexo y rango de años, de manera que obtenemos una lista con sexo, período, y suma total

```
dfSuicidiosFilteredSum <- ddply(dfSuicidiosFilteredGenre, .(sex,yearRange5Years), summarize,  sum_suici
```

Si analizamos el dataset obtenido, y teniendo en cuenta que el último año registrado y solo en algunos países es el 2016, estimamos oportuno eliminar este rango, al no estar completo prácticamente y no estar disponible para todos los países, por lo que puede desvirtuar los resultados obtenidos. Por tanto eliminamos la fila correspondiente al período 2015-2020.

```
dfSuicidiosFilteredYears <- dfSuicidiosFilteredSum[dfSuicidiosFilteredSum[,"yearRange5Years"]!="(2015,2
```

A continuación obtenemos a modo de información estadística, los valores para sexo hombre o mujer. En este momento ya podemos observar datos como media, mediana, cuartiles, etc. lo que nos permite ya decir que hay una gran diferencia en el número de suicidios entre hombres y mujeres.

Para continuar con la demostración de esta asunción, obtenemos un diagrama de cajas en los que podemos ver representados estos elementos.

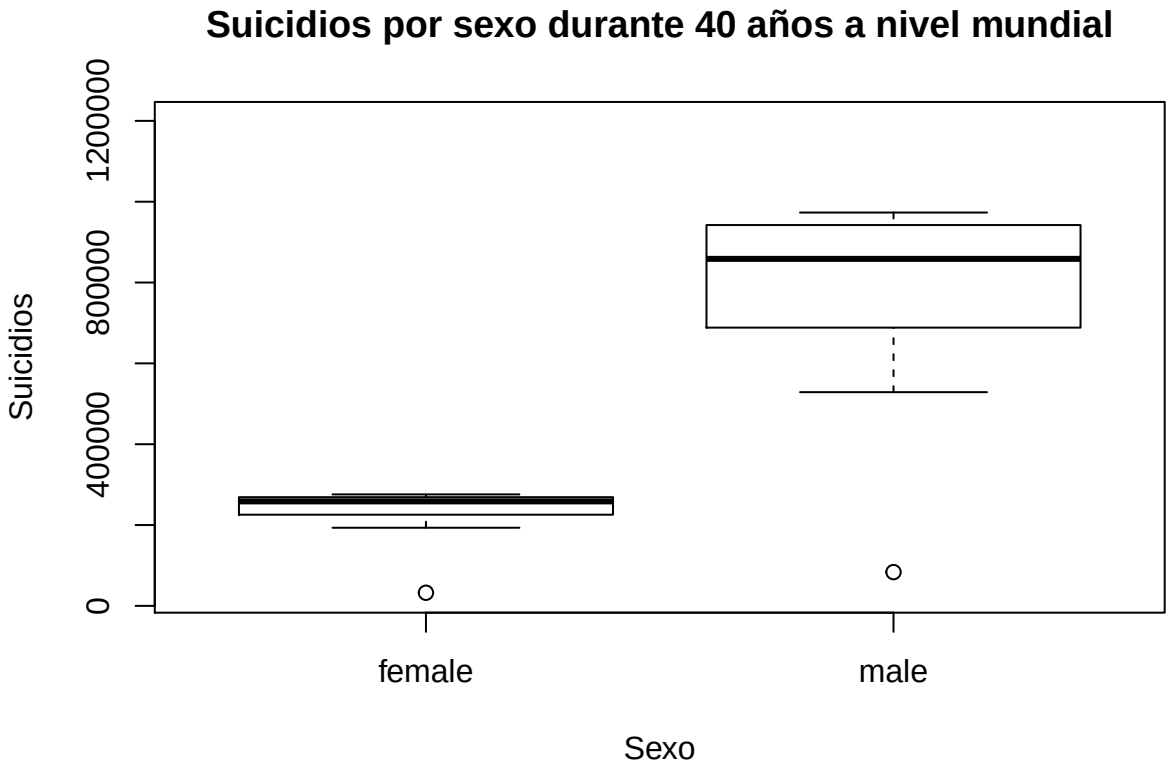
```
summary(dfSuicidiosFilteredYears[dfSuicidiosFilteredYears[,"sex"]=="female",])
```

```
##      sex       yearRange5Years sum_suicides_no
##  female:7   (1980,1985]:1      Min.   : 32479
##  male  :0   (1985,1990]:1      1st Qu.:225621
##             (1990,1995]:1      Median :258556
##             (1995,2000]:1      Mean   :222287
##             (2000,2005]:1      3rd Qu.:268960
##             (2005,2010]:1      Max.   :275809
##             (Other)    :1
```

```
summary(dfSuicidiosFilteredYears[dfSuicidiosFilteredYears[,"sex"]=="male",])
```

```
##      sex       yearRange5Years sum_suicides_no
##  female:0   (1980,1985]:1      Min.   : 83584
##  male  :7   (1985,1990]:1      1st Qu.:688450
##             (1990,1995]:1      Median :858577
##             (1995,2000]:1      Mean   :739544
##             (2000,2005]:1      3rd Qu.:942274
##             (2005,2010]:1      Max.   :973203
##             (Other)    :1
```

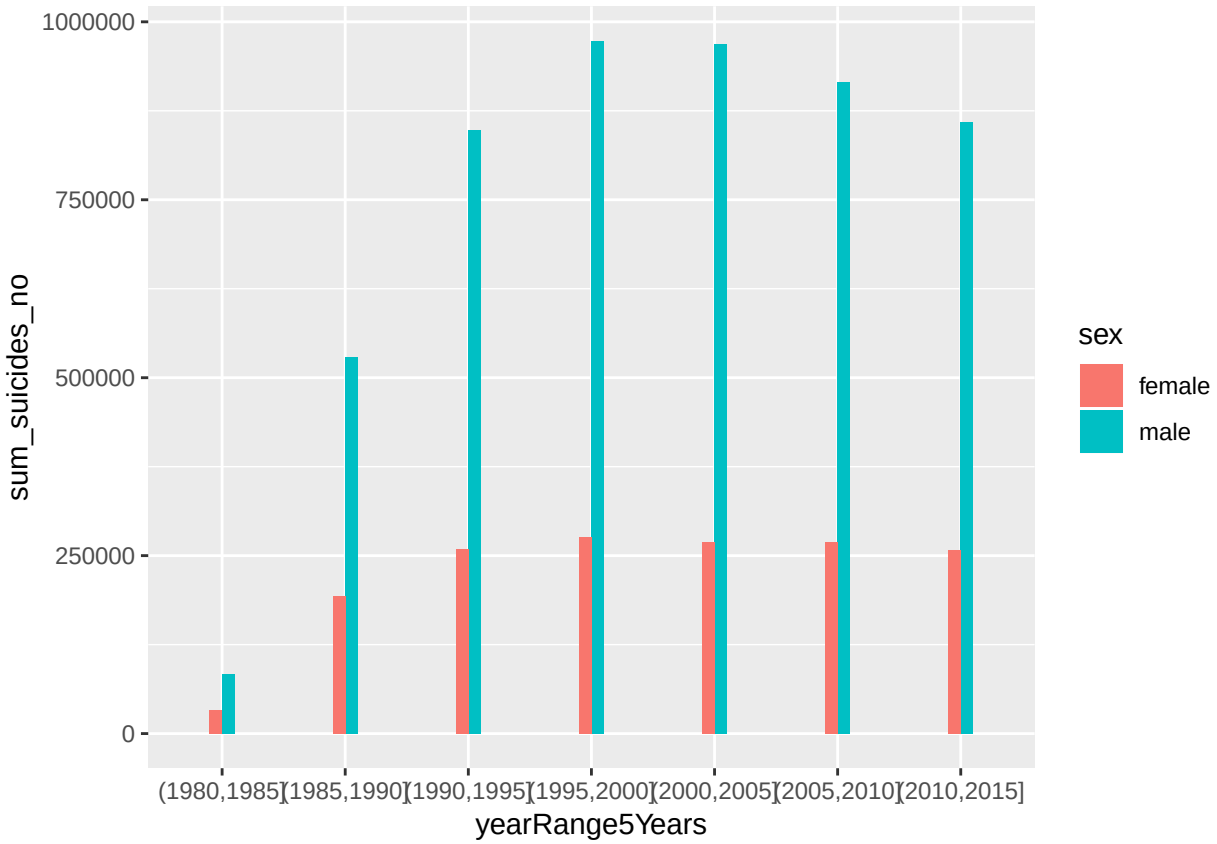
```
boxplot(sum_suicides_no~sex,data=dfSuicidiosFilteredYears, ylim=c(30000, 1200000), main="Suicidios por
```



Destaca la presencia de outliers para ambos sexos, que correspondencia con los valores del primer período 1980-1985. Esto nos lleva a plantearnos que posiblemente la cifra para este período no esté disponible para todos los países además que posiblemente en esos años no se llevase un recuento muy preciso de estos datos, de manera que en base a esto podríamos plantearnos la eliminación de este registro. Si eliminamos este rango, podemos volver a obtener los datos estadísticos y el diagrama de cajas y ver las diferencias.

Finalmente, hacemos uso de un diagrama de barras para observar de manera clara las diferencias del número de sucidios a lo largo de los años.

```
ggplot(dfSuicidiosFilteredYears, aes(yearRange5Years, sum_suicides_no, fill = sex)) + geom_bar(stat = "
```



3. Predicción del HDI en Europa en los próximos años

El siguiente análisis cambia el enfoque completamente, pero nos permitirá adquirir una información que será de utilidad en otros posibles análisis posteriores. Es por ello que estimamos que merece la pena profundidar un poco en el contenido de este dato y sus implicaciones. Para ello vamos a realizar una predicción de cómo va a evolucionar el HDI en Europa en los próximos años. Para ello hacemos un filtrado de de las columnas que necesitamos para este análisis:

```
# Predicción del HDI en Europa en los próximos años
keeps <- c("country", "continent","year","HDI_for_year")
dfHDI <- dfSuicidios[keeps]

dfHDIFiltered <- dfHDI[dfHDI[,"continent"]=="Europe",]
```

Llamamos la función *completeHDI_For_YearDatoPosteriorMedia*, definida en apartados anteriores

```
dfHDIFilteredComplete <- completeHDI_For_YearDatoPosteriorMedia(dfHDIFiltered)
```

Llegado a este punto obtenemos un listado de los países de Europa, con su año, su HDI calculado en base a valores medios del año anterior y posterior de cada país Como aún así obtenemos algunos datos NA en lo referente a los últimos años, los filtramos. Al fin y al cabo nuestro análisis predictivo nos ayudará a "completar" estos datos:

```
dfHDIFilteredCompleteWithoutNA <- dfHDIFilteredComplete[complete.cases(dfHDIFilteredComplete), ]
```

Finalmente realizamos una media de todos los valores disponibles para cada uno de los años:

```
dfHDIFilteredCompleteWithoutNAByYear <- ddply(dfHDIFilteredCompleteWithoutNA, .(year), summarize,  HDI_
```

Una vez obtenidos estos valores disponemos de un dataset compuesto por **year** y **HDI_for_year** y a continuación creamos un modelo de regresión lineal basado en esos datos:

```
model <- lm(HDI_for_year ~ year, data=dfHDIFilteredCompleteWithoutNAByYear)
```

Como carecemos de valores a partir del 2015, ejecutamos una predicción en base al model anterior para predecir los valores de HDI para los años: 2015, 2016, 2017, 2018, 2019, 2020

```
new.df <- data.frame(year=c(2015,2016,2017,2018,2019,2020))
predict(model, new.df)
```

```
##         1         2         3         4         5         6
## 0.8696999 0.8740208 0.8783417 0.8826626 0.8869835 0.8913043
```

Observamos que continúa la tendencia detectada con los años disponibles de un pequeño incremento en los años siguientes

4. Análisis de varianza

A continuación vamos a mostrar unos análisis ANOVA que nos permita comparar los resultados de K 'factores' con respecto a la variable dependiente o de interés, que en este caso va a ser el HDI y el número de suicidios.

Para ello vamos a volver a procesar los datos disponibles

```
# ANOVAS

keeps <- c("country","continent","year","suicides_no","gdp_per_capita_USDollar","HDI_for_year")
dfHDI <- dfSuicidios[keeps]

dfHDIFilteredComplete <- completeHDI_For_YearDatoPosteriorMedia(dfHDI)

dfHDIFilteredCompleteWithoutNA <- dfHDIFilteredComplete[complete.cases(dfHDIFilteredComplete), ]
dfHDIFilteredCompleteWithoutNASum <- ddply(dfHDIFilteredCompleteWithoutNA, .(country, continent, year,g
fit <- anova(lm(HDI_for_year ~ country+gdp_per_capita_USDollar, data=dfHDIFilteredCompleteWithoutNASum)
fit
```

```
## Analysis of Variance Table
##
## Response: HDI_for_year
##                           Df  Sum Sq Mean Sq F value    Pr(>F)
## country                   89 15.3419 0.17238  206.76 < 2.2e-16 ***
## gdp_per_capita_USDollar    1  1.2193 1.21926 1462.40 < 2.2e-16 ***
## Residuals               2019  1.6833 0.00083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que el p-valor obtenido es menor al nivel de significancia 0.05, se puede concluir que HDI_for_year muestra diferencias significativas para los diferentes países y producto interior bruto.

```
fit2 <- anova(lm(sum_suicides_no ~ country+gdp_per_capita_USDollar, data=dfHDIFilteredCompleteWithoutNA
fit2
```

```
## Analysis of Variance Table
##
## Response: sum_suicides_no
##                           Df     Sum Sq   Mean Sq  F value    Pr(>F)
```

```
## country                   89 5.8812e+10 660813947 943.3121 < 2.2e-16 ***
## gdp_per_capita_USDollar    1 4.7279e+06   4727948   6.7491  0.009447 **
## Residuals               2019 1.4144e+09    700525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que el p-valor obtenido es menor al nivel de significancia 0.05, se puede concluir que sum_suicides_no muestra diferencias significativas para los diferentes países y producto interior bruto.

5. Recursos

Dataset
https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

References
United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

[Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

World Health Organization. (2018). Suicide prevention. Retrieved from
http://www.who.int/mental_health/suicide-prevention/en/

Referencia de informe
Gutierrez_teguayco_Practica_2___Limpieza_y_validacion_de___27_12_2017_07_31_05

Human Development Index (HDI)
https://en.wikipedia.org/wiki/Human_Development_Index

Gross domestic product (GDP)
https://en.wikipedia.org/wiki/Gross_domestic_product

Package 'countrycode' https://cran.r-project.org/web/packages/countrycode/countrycode.pdf