El dataset proporciona una gran cantidad de indicadores a lo largo de un gran conjunto de años y de países, de manera que resulta complicado realizar una tarea de preparación y limpieza de datos que sea compatible a todos los análisis que se desean proporcionar.

Por ello, lo que planteamos es el tratamiento específico que realizamos para cada una de las preguntas y análisis propuestos, de manera que en función de lo que queramos demostrar o predecir, realizaremos un preparatorio de los datos ajustado y apropiado para tal fin.

Comparativa de suicidios por sexo a nivel mundial considerando los 40 años de datos disponibles.

Preparación y limpieza de los datos:
Analizamos el dataset y de las columnas disponibles, detectamos las que necesitamos para esta acción específica. Cargamos pues el csv y seleccionamos las columnas: "country", “year”,”sex”,”suicides_no”, es decir, el país, el año, el sexo y el número de suicidios registrados. Para facilitar el análisis de los datos posterior, creamos la columna yearRange5Years que discretiza la columna year en valores de 5 en 5 años. De manera previa cargamos las librerías necesarias para poder realizar éstas y las siguientes acciones:

```
# Suicidios Mundial por sexo y por rango de años de 5 en 5


#install.packages("countrycode")
library(countrycode)
library(dplyr)
library(plyr)



dfSuicidios <- read.csv("/Users/anablanesmartinez/MasterCienciaDatos/
TipologiaCicloVidaDatos/PRAC2/suicide.csv", encoding = "UTF-8", header =
TRUE);


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

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

```
dfSuicidiosFiltered$yearRange5Years<-cut(dfSuicidiosFiltered$year,
c(1980,1985,1990,1995,2000,2005,2010,2015,2020))
```

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(dfSuicidiosFiltered, .
(sex,yearRange5Years), summarize,  sum_suicides_no=sum(suicides_no))
```

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"]!
="(2015,2020]",]
```

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"]=="femal
e",])
summary(dfSuicidiosFilteredYears[dfSuicidiosFilteredYears[,"sex"]=="male"
,])

boxplot(sum_suicides_no~sex,data=dfSuicidiosFilteredYears, ylim=c(30000,
1200000), main="Suicidios por sexo durante 40 años a nivel
mundial",xlab="Sexo", ylab="Suicidios")
```

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 = "identity", width = 0.2, position =
"dodge") + labs(list(x = "x", y = "count",fill = "group"))
```

Podemos ver cómo el número de suicidios en los hombres a nivel mundial es alrededor de 4 veces mayor que el de las mujeres.

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 análisis posteriorres. Aunque el dataset es sobre suicidios, aparece una columna referida al Human Development Index (HDI) que puede ser de utilidad si queremos realizar un análisis sobre si existe una correlación entre este valor y el número de suicidios. 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 cargamos las librerías necesarias y el dataset, para a continuación hacer un filtrado de de las columnas que necesitamos para este análisis:

"country", “continent","year","HDI_for_year"

Como queremos hacerlo por continente y solo disponemos del campo country, hacemos uso de la librería de R countrycode que permite obtener el continente asociado a cada país. Una vez hecho esto, filtramos por continente = “Europa”.

```
# Predicción del HDI en Europa en los próximos años


#install.packages("countrycode")
library(countrycode)
library(dplyr)
library(plyr)

dfHDI <- read.csv("/Users/anablanesmartinez/MasterCienciaDatos/
TipologiaCicloVidaDatos/PRAC2/suicide.csv", encoding = "UTF-8", header =
TRUE);


dfHDI$continent <- countrycode(sourcevar = dfHDI[, "country"], origin =
"country.name", destination = "continent")
colnames(dfHDI)[colnames(dfHDI)=="gdp_per_capita...."] <-
"gdp_per_capita_USDollar"
colnames(dfHDI)[colnames(dfHDI)=="gdp_for_year...."] <-
"gdp_for_year_USDollar"
colnames(dfHDI)[colnames(dfHDI)=="HDI.for.year"] <- "HDI_for_year"
colnames(dfHDI)[colnames(dfHDI)=="country.year"] <- "country_year"
colnames(dfHDI)[colnames(dfHDI)=="suicides.100k.pop"] <-
"suicides_100k_pop"

keeps <- c("country", "continent","year","HDI_for_year")
```

```
dfHDI <- dfHDI[keeps]

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

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);

    }
```

```
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_for_year=mean(HDI_for_year, na.rm = TRUE))
```

Una vez obtenidos estos valores disponemos de un dataset compuesto por year y HDI 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)
```

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