Práctica 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS

Alberto Lucas Navío

Índice

1. Introducción:

Esta actividad consiste en llevar a cabo un caso práctico de tratamiento de datos. Como parte de este tratamiento se incluye, la limpieza del conjunto de datos, seleccionar los datos relevantes para llevar a cabo un análisis, así como utilizar de forma eficiente las herramientas de integración, limpieza, validación y análisis de los datos.

2. Resolución:

2.1. Descripción del conjunto de datos:

El dataset elegido para llevar a cabo el proyecto se denomina *"Countries of the World"* y se puede encontrar en Kaggle a través de este enlace: https://www.kaggle.com/fernandol/countries-of-the-world. Este conjunto de datos contiene indicadores económicos como tasa de migración, producto interior bruto per cápita, etc., sobre prácticamente todos los países del mundo. A continuación, describiremos el conjunto de variables que forman este dataset.

Este conjunto de datos se ha enriquecido con datos adicionales extraídos de otro dataset, denominado *"Country_profile_variables".* Este conjunto de datos se puede encontrar también en Kaggle, concretamente: https://www.kaggle.com/sudalairajkumar/undata-country-profiles. Concretamente, hemos escogido una variable de este dataset, que es la tasa de desempleo de cada país, y lo hemos agregado al conjunto datos anterior.

Así pues, nuestro dataset inicial está formado por los siguientes atributos.

- Country: nombre del país.
- Region: region geográfica en la que se encuentra el país en cuestión
- Population: número de habitantes del país.
- Area (sq. mi.): superficie del país en millas cuadradas.
- Pop. Density (per sq.mi.): densidad de población del país.
- Coastline (coast/area ratio): el porcentaje de línea de costa del país con relación al área total del mismo.
- Net migration: la tasa de migración del país. Será un valor positivo si el país en cuestión tiene más inmigrantes que emigrantes. Aquellas observaciones con un valor negativo para esta variable, experimentan más emigración que inmigración.
- Infant mortality (per 1000 births): número de muertes por cada 1000 niños nacidos en un año.
- GDP_MILLION: product interior bruto del país.
- GDP($ per capita): producto interior bruto del país dividido por el número de habitantes.
- Literacy (%): porcentaje de personas del país que saben leer y escribir.
- Phones (per 1000): número total de teléfonos móviles en uso por cada 1000 habitantes.
- Arable (%): porcentaje del territorio que puede ser utilizado para el cultivo.
- Crops (%): índice que mide la producción agrícola del país.

- Other (%)
- Climate: tipo de clima del país en cuestión
- Birthrate: número de nacimientos por 1000 habitantes en el país.
- Deathrate: el ratio de muertes por 1000 habitantes en el país.
- Agriculture: porcentaje del producto interior bruto generado por el sector primario.
- Industry: porcentaje del producto interior bruto generado por el sector secundario.
- Service: procentaje del product interior bruto generado por el sector terciario.
- Unemployment: porcentaje de población activa que no tiene trabajo.

Partimos pues de un conjunto de datos con 227 observaciones y 22 atributos para cada observación.

2.2. Importancia y objetivos de los análisis:

Este conjunto de datos está formado por una lista de prácticamente todos los países del mundo y una serie de atributos que son un conjunto de indicadores económicos de cada país. En el análisis que vamos a realizar va a girar en torno a la tasa neta de inmigración de los países.

Concretamente, vamos a intentar extraer conclusiones acerca de qué factores económicos de un país tienen más influencia sobre fenómenos como es la migración. Así mismo, vamos a comprobar si existe una relación estadística entre el producto interior bruto de un país o su producto interior bruto per cápita y su tasa de migración.

Este análisis pretende entender un poco mejor un fenómeno tan complejo y que centra tantos debates en la actualidad como la migración. Entender qué factores llevan a las personas a desplazarse de un país a otro es muy relevante en la sociedad actual.

Este tipo de análisis tienen gran relevancia en una gran diversidad de ámbitos. A nivel profesional se podría asumir que historiadores, geógrafos y cualquier persona que trabaje en el ámbito de las Ciencias Sociales serían los principales receptores de un análisis de este tipo. Sin embargo, este análisis también puede ser de interés para cualquier persona interesada en entender mejor el mundo en el que vive, independientemente de su profesión.

2.3. Limpieza de datos

El primer paso del proyecto será importar el conjunto de datos con el que vamos a trabajar. Éste se encuentra en un fichero CS, que importamos a la interfaz de R en formato data.frame, mediante el comando read.csv.

```
> #Importamos los datos
> country_data = read.csv("initial_dataset.csv", header=TRUE, dec=",")
> head(country_data)
        Country                               Region Population Area..sq..mi.. Pop..Density..per.sq..mi.. Coastline..coast.area.ratio. Net.migration Infant.mortality..per.1000.births.
1   Afghanistan        ASIA (EX. NEAR EAST)            31056997        647500                    48.0                        0.00        23.06                            163.07
2       Albania   EASTERN EUROPE                     3581655         28748                   124.6                        1.26        -4.93                             21.52
3       Algeria   NORTHERN AFRICA                   32930091       2381740                    13.8                        0.04        -0.39                             31.00
4 American Samoa  OCEANIA                              57794           199                   290.4                       58.29       -20.71                              9.27
5       Andorra   WESTERN EUROPE                       71201           468                   152.1                        0.00         6.60                              4.05
6        Angola   SUB-SAHARAN AFRICA                12127071       1246700                     9.7                        0.13         0.00                            191.19
  GDP_MILLION_. GDP....per.capita. Literacy.... Phones..per.1000. Arable.... Crops.... Other.... Climate Birthrate Deathrate Agriculture Industry Service Unemployment
1             0               700         36.0              3.2       12.13      0.22      87.65       1     46.60     20.34       0.380    0.240   0.380           35
2          4100              4500         86.5             71.2       21.09      4.42      74.49       3     15.11      5.22       0.232    0.188   0.579           14
3        108700              6000         70.0             78.1        3.22      0.25      96.53       1     17.14      4.61       0.101    0.600   0.298         11,7
4           #N/A              8000         97.0            259.5       10.00     15.00      75.00       2     22.46      3.27          NA       NA      NA         29,8
5          1000             19000        100.0            497.2        2.22      0.00      97.78       3      8.71      6.25          NA       NA      NA          3,7
6           #N/A              1900         42.0              7.8        2.41      0.24      97.35      NA     45.11     24.20       0.096    0.658   0.246          #N/A
>
```

Observamos qué tipo asigna R a cada variable. Como se puede ver disponemos de varias variables numéricas, así como algunas variables a las que R les ha asignado el tipo factor.

```
> ## Tipo de dato asignado a las variables del dataframe
> sapply(country_data, function(x) class(x))
                        Country                          Region
                       "factor"                        "factor"
                     Population                    Area..sq..mi..
                      "integer"                       "integer"
     Pop..Density..per.sq..mi..       Coastline..coast.area.ratio.
                      "numeric"                       "numeric"
                  Net.migration Infant.mortality..per.1000.births.
                      "numeric"                       "numeric"
                  GDP_MILLION_.                GDP....per.capita.
                      "integer"                       "integer"
                   Literacy....                 Phones..per.1000.
                      "numeric"                       "numeric"
                     Arable....                         Crops....
                      "numeric"                       "numeric"
                     Other....                           Climate
                      "numeric"                       "numeric"
                     Birthrate                         Deathrate
                      "numeric"                       "numeric"
                   Agriculture                          Industry
                      "numeric"                       "numeric"
                       Service                      Unemployment
                      "numeric"                       "numeric"
>
```

2.3.1. Selección de datos de interés:

Vamos a iniciar el tratamiento de los datos, seleccionando los atributos relevantes para nuestro análisis. Para el análisis que vamos a realizar podemos prescindir de varios atributos que están más relacionados con factores geográficos de un país que con indicadores estrictamente económicos. Así pues, prescindiremos de los siguientes atributos*: pop. Density, coastline, phones (per 1000), arable, crops* and *climate*.

```
> country_data <- country_data[, -(5:6)]

> country_data <- country_data[, -(10:14)]
```

El resultado por tanto, es un conjunto de datos con 227 observaciones y 15 atributos. Una vez seleccionados los atributos, es posible tratar los valores ausentes del dataset.

2.3.2. Tratamiento de nulos y ceros:

El siguiente paso del proyecto será el tratamiento de nulos y ceros de nuestro conjunto de datos. Antes de nada, es necesario comprobar el número de valores nulos de cada variable. Así mismo, hay que considerar si es posible que una variable contenga el valor 0, o si este valor no forma parte del dominio del atributo. En este caso habría que tratar este 0 como un valor nulo.

Comprobamos pues, el número de nulos que hay por cada variable, como se puede ver a continuación:

```
> sapply (country_data, function(x) sum(is.na(x)))
                         Country                           Region
                               0                                0
                      Population                   Area..sq..mi..
                               0                                0
                   Net.migration Infant.mortality..per.1000.births.
                               0                                3
                   GDP_MILLION_.                GDP....per.capita.
                             142                                0
                    Literacy....                        Birthrate
                              18                                3
                      Deathrate                      Agriculture
                               4                               15
                        Industry                          Service
                              16                               15
                    Unemployment
                              37
```

Una vez comprobado el número de valores nulos, hay que decidir qué estrategia es la más adecuada para tratar con estos. Se podrían las observaciones con valores nulos. Sin embargo, este dataset sólo contiene 227 observaciones. Por tanto, eliminar las observaciones con valores nulos supondría una pérdida de información importante.

Por otra parte, hay variables con gran cantidad de valores nulos y otras con muy pocos. Por lo tanto, utilizaremos una estrategia diferente para tratar los valores nulos de cada variable. Cabe destacar que en aquellos casos en los que es necesario introducir valores de forma manual, la información se ha extraído del siguiente sitio: https://www.indexmundi.com/

Una vez aclarado esto, vamos analizando atributo por atributo:

- GDP_MILLION_$:

  Para este atributo, no disponemos del valor de 142 observaciones. Además, en 35 observaciones el valor es 0. Sin embargo, 0 no puede formar parte del dominio de este atributo. Por tanto, podemos decir que tenemos 172 valores nulos para esta variable. Es decir, más del 75% de observaciones son valores nulos.

  Como hay tantos valores ausentes, no sería muy efectivo aplicar técnicas de imputación de valores como la imputación basada en K vecinos. Sin embargo, podemos calcular este atributo fácilmente a partir de otras variables de nuestro conjunto de datos. Concretamente, vamos a multiplicar el producto interior bruto per cápita de un país por su número de habitantes para obtener el producto interior bruto del país.

  Así pues, añadimos una nueva columna a nuestro *dataframe*, que será igual al producto de las columnas GDP ($ per capita) y population. Una vez hecho esto, eliminamos el atributo GDP_MILLION_$:

```
> country_data$GDP <- as.integer((country_data$GDP....per.capita. * country_data$Population)/ 1000000)
>
```

```
>
> country_data <- country_data[,-(7)]
>
```

- GDP ($ per capita):

  Sólo 1 valor ausente. Como es una única observación, no supone un esfuerzo excesivo hallar el valor del este atributo. Por tanto, podemos buscar el valor de la misma e incluirlo en nuestro dataset. Esta forma de tratar los valores nulos nos permite completar nuestro conjunto de datos de la forma más precisa, lo cual nos permitirá realizar análisis más precisos.

  Además, es sencillo encontrar la información que necesitamos y añadir el valor a la observación correspondiente.

```
> country_data$Country[224]
[1] Western Sahara
227 Levels: Afghanistan Albania Algeria American Samoa Andorra Angola ... Zimbabwe
>
> country_data$GDP....per.capita.[224]
[1] NA
>
> country_data$GDP....per.capita.[224]<- 2500
> country_data$GDP....per.capita.[224]
[1] 2500
```

- Literacy: 18 observaciones sin este valor.

  En este caso vamos a utilizar el método de kNN, del paquete *VIM,* un método de imputación de valores basado en la similitud entre las observaciones. Es decir, utilizamos una técnica para predecir el valor del atributo de una observación, a partir del valor de las observaciones más parecidas.

```
> suppressWarnings (suppressMessages(library(VIM)))
```

  Con esto asignamos valores a las observaciones que no tienen valor en nuestro conjunto de datos.

```
> country_data$Literacy_rate <- kNN(country_data)$Literacy_rate
```

- Birthrate: 3 observaciones sin este valor.

  De nuevo, como sólo hay 3 observaciones sin valor, es razonable buscar estos valores y añadirlos al conjunto de datos.

```
> country_data$Birthrate[182]
[1] NA
>
> country_data$Birthrate[182]<- 9.2
> country_data$Birthrate[182]
[1] 9.2
>
```

```
>
> country_data$Birthrate[222]<- 5.5
> country_data$Birthrate[224]<- 28.9
>
```

- Deathrate: 4 observaciones sin este valor.

  En este caso hay 4 valores nulos. Aplicamos de nuevo la técnica anterior de encontrar la información que necesitamos y aplicarla a nuestro conjunto de datos.

```
> country_data$Deathrate[48]
[1] NA
>
> country_data$Deathrate[48]<- 8.6
> country_data$Deathrate[48]
[1] 8.6
>
```

```
> country_data$Deathrate[48]
[1] NA
>
> country_data$Deathrate[48]<- 8.6
> country_data$Deathrate[48]
[1] 8.6
>
> country_data$Deathrate[222]<- 5.5
> country_data$Deathrate[224]<- 11.49
>
```

```
> country_data$Deathrate[182]
[1] NA
>
> country_data$Deathrate[182]<- 13.2
> country_data$Deathrate[182]
[1] 13.2
>
```

- Agriculture, industry, service: 15 observaciones sin valor.

  Estos tres atributos van a ser reemplazados por un único atributo que describa simplemente cuál es el sector principal en la economía del país. Para nuestro análisis no es necesario disponer del porcentaje de cada sector en la economía de cada país. Así pues, primero se reemplazarán los atributos y luego se insertarán los valores de forma manual.

- Unemployment: 37 observaciones sin valor

  Como se ha hecho anteriormente, reemplazamos los valores nulos utilizando Knn.

```
> country_data$Unemployment_rate <- kNN(country_data)$Unemployment_rate
```

- Net_migration: 3 observaciones sin valor.

  De nuevo para este atributo, disponemos sólo de 3 observaciones sin valor. Cabe destacar que hay valores 0, pero en este caso, el valor 0 sí que forma parte del dominio del atributo. Como hemos hecho anteriormente, vamos a encontrar la información necesaria e introducir los valores para la observación correspondiente.

```
> country_data$Net.migration[48]
[1] NA
>
> country_data$Net.migration[48] <- -2.2
> country_data$Net.migration[48]
[1] -2.2
>
```

```
> country_data$Net.migration[222]
[1] NA
>
> country_data$Net.migration[222]<- -4.6
> country_data$Net.migration[222]
[1] -4.6
>
```

```
> country_data$Net.migration[224]
[1] NA
>
> country_data$Net.migration[224]<- 5.4
> country_data$Net.migration[224]
[1] 5.4
>
```

- Infant_mortality: 3 observaciones sin valor.

De nuevo, sólo 3 observaciones con valores nulos. Como en los casos anteriores buscamos la información necesaria y la introducimos en el dataframe.

```
> country_data$Infant.mortality[48]<- 12.6
> country_data$Infant.mortality[222]<- 4.3
> country_data$Infant.mortality[224]<- 50.5
>
```

2.3.3. Reducción de la dimensionalidad

Una vez hemos tratado los valores ausentes, vamos a llevar a cabo un ejercicio de reducción de dimensionalidad. Disponemos de tres atributos que indican el porcentaje de importancia que tiene cada sector económico.

Vamos a crear un nuevo atributo que indique el cuál es el sector más importante en la economía de un país, que va a reemplazar las tres variables que indican el peso de cada sector económico.

```
> country_data$Main_sector <- ifelse ((country_data$Agriculture > country_data$Industry) & (country_data$Agriculture > country_data$Service), "Primario",
+ ifelse((country_data$Industry > country_data$Agriculture) & (country_data$Industry > country_data$Service), "Secundario",
+ ifelse ((country_data$Service > country_data$Agriculture) & (country_data$Service > country_data$Industry), "Terciario", "Desconocido")))
>
```

Una vez hecho esto, hay varias observaciones para las que no tenemos valor. Sin embargo, es posible encontrar cuál es el sector más importante en la economía de cada país y completar nuestro conjunto de datos.

```
> country_data$Main_sector[4] <- "Primario"
> country_data$Main_sector[5] <- "Terciario"
> country_data$Main_sector[79] <- "Terciario"
> country_data$Main_sector[81] <- "Secundario"
> country_data$Main_sector[84] <- "Terciario"
> country_data$Main_sector[135] <- "Primario"
> country_data$Main_sector[139] <- "Terciario"
> country_data$Main_sector[141] <- "Terciario"
> country_data$Main_sector[145] <- "Terciario"
> country_data$Main_sector[154] <- "Secundario"
> country_data$Main_sector[172] <- "Primario"
> country_data$Main_sector[175] <- "Terciario"
> country_data$Main_sector[178] <- "Terciario"
> country_data$Main_sector[209] <- "Terciario"
> country_data$Main_sector[222] <- "Primario"
> country_data$Main_sector[224] <- "Primario"
>
```

Una vez hecho esto, podemos eliminar los atributos “agriculture, industry, service” ya que una vez creado el atributo “Main_sector”, esta información resulta redundante para nuestro análisis.

```
>
> country_data <- country_data[,-(10:12)]
```

Eliminamos también el atributo “Infant.Mortality.per1000births”. Como disponemos de un atributo denominado “Infant.Mortality” que mide el mismo factor, es redundante conservar las dos variables. Eliminamos por lo tanto una de ellas, como se ve a continuación.

```
>
> country_data <- country_data[,-(6)]
```

- **Reemplazar los nombres de las columnas**

  Una última operación que vamos a realizar antes de finalizar la fase de limpieza de los datos es la de cambiar el nombre de las columnas, para facilitar el uso de las variables durante la fase de análisis.

```
>
> colnames(country_data) <- c("Country", "Region", "Population", "Area", "Net_migration", "GDP_per_capita", "Literacy_rate", "Birthrate",
+ "Deathrate", "Unemployment_rate", "Infant_mortality", "GDP", "Main_sector")
>
```

Una vez hemos tratado los valores nulos de cada atributo, comprobamos que efectivamente nuestro conjunto de datos ya no tiene valores nulos.

```
> sapply(country_data, function(x) sum(is.na(x)))
          Country            Region        Population              Area      Net_migration
                0                 0                 0                 0                 0
   GDP_per_capita     Literacy_rate         Birthrate         Deathrate Unemployment_rate
                0                 0                 0                 0                 0
 Infant_mortality               GDP       Main_sector
                0                 0                 0
> 
```

Una vez completada esta operación, procedemos a tratar los valores extremos.

2.3.4. Tratamiento de valores extremos

Los valores extremos son aquellos valores que se encuentran muy alejados de la distribución normal de una variable. Estos valores pueden aparecer por distintos motivos y pueden ser indicadores de que hay algún error en los datos que puede afectar al resultado de los análisis.

Vamos a aplicar la función *boxplot.stats()* de R, en aquellas variables en que los valores extremos puedan ser indicadores. Es decir, obteniendo los valores extremos para la variable población se pueden extraer muchas conclusiones, ya que estamos comparando países con poblaciones muy diferentes. Sin embargo, con atributos como la tasa de migración, los valores extremos sí que pueden ser relevantes.

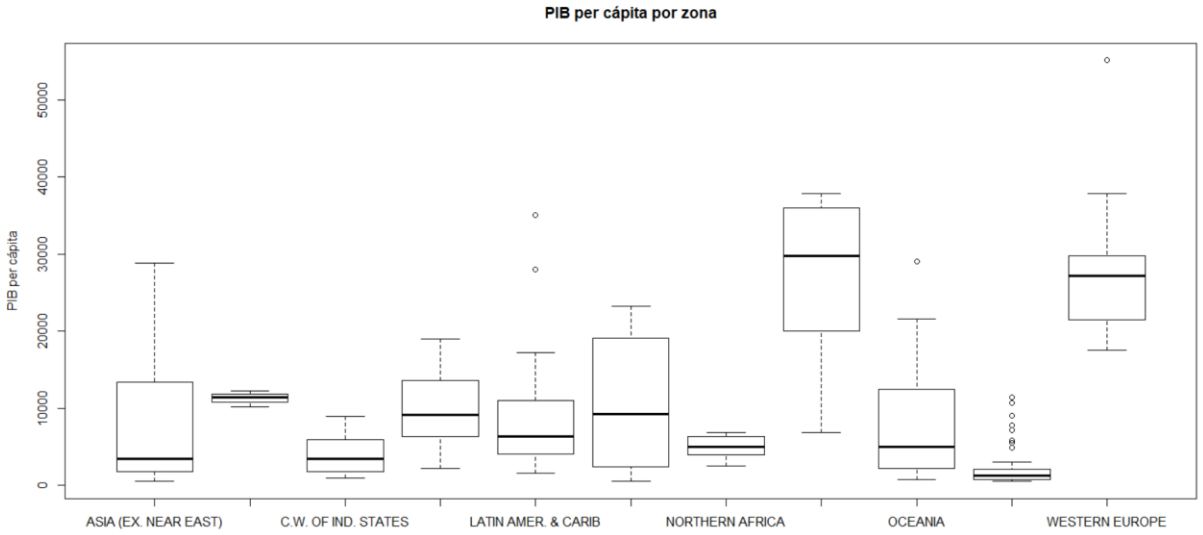
Vemos a continuación el cálculo de los valores extremos para nuestro conjunto de datos:

```
> boxplot.stats(country_data$Net_migration)$out
 [1]  23.06  -4.93 -20.71   6.60  10.76  -6.15  -6.47  -4.90  10.01  -4.58   5.96 -12.07  18.75 -13.87
[15]  -8.58   6.27  -4.70  -8.37 -13.92   5.24   4.99   5.36  -4.92   6.59  14.18   4.85   8.97   4.86
[29]  -6.04   6.78  -4.87 -20.99   7.75   4.05   9.61  16.29  -7.11  -4.86  -7.64 -11.70  10.98  -5.69
[43]  11.53   5.37  -8.81   4.05 -10.83  11.68  -8.94  -4.60   5.40
>
> boxplot.stats(country_data$GDP_per_capita)$out
[1] 55100 37800 37800
>
> boxplot.stats(country_data$Literacy_rate)$out
[1] 36.0 26.6 35.9 17.6 31.4 37.8
>
> boxplot.stats(country_data$Birthrate)$out
numeric(0)
>
> boxplot.stats(country_data$Deathrate)$out
 [1] 20.34 24.20 29.50 18.65 19.31 28.71 23.10 19.33 21.35 18.86 20.91 23.03 22.00 29.74 19.93 21.84
>
> boxplot.stats(country_data$Unemployment_rate)$out
 [1] 35.0 29.8 77.0 60.0 28.0 26.7 33.5 40.6 40.0 30.6 28.1 30.0 36.0 29.8 28.1 48.0 27.6 28.0 50.0 26.4
[21] 26.7 27.0 95.0
>
> boxplot.stats(country_data$Infant_mortality)$out
[1] 163.07 191.19 128.87 130.79 143.64
>
```

Observando los resultados podemos ver que nuestro conjunto de datos presenta valores extremos en cada una de las variables cuantitativas. Para tomar una decisión acerca de como tratar estos valores, es útil visualizar de forma gráfica estos valores. Por tanto, vamos a generar un *boxplot* para cada variable con valores extremos. Vamos a distribuir los valores en función de la región ya que esto nos dará una mejor idea sobre si un valor extremo puede darse en realidad o no.

En primer lugar, miramos al tributo “GDP_per_capita”:

```
> boxplot(country_data$GDP_per_capita ~ country_data$Region, data = country_data, main = "PIB per cápita por zona", xlab ="Región", ylab= "PIB per cápita")
```



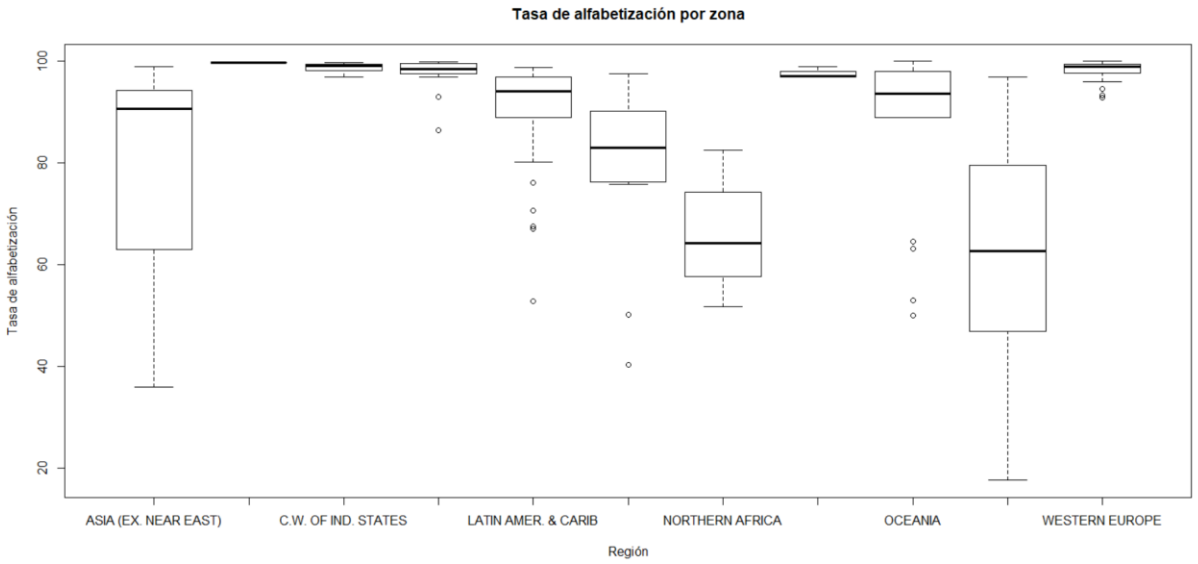
Los dos primeros *outliers* que vemos están en el grupo de países de Sudamérica y el Caribe. Estos dos valores extremos se corresponden con Aruba (que ha pertenecido a Holanda muchos años) y las Islas Caimán (que pertenecen a Gran Bretaña). Son dos territorios con circunstancias especiales y con el PIB per cápita más alto del continente.

En lo que respecta a Oceanía, el valor extremo corresponde con Australia. Esto no es de extrañar, teniendo en cuenta que es con diferencia el país más grande de Oceanía con muchísima diferencia. En cuanto a la región África subsahariana vemos que presenta un número más alto de valores extremos. Sin embargo, vemos que estos valores no son muy elevados. De hecho, son inferiores al valor mínimo de otro grupo como es Europa Occidental. Sin embargo, aparecen como valores extremos debido al bajo PIB per cápita de muchos países de África.

Por último, hay un valor extremo de la región Europa Occidental. Este valor pertenece a Luxemburgo, que es uno de los países con mayor calidad de vida del mundo. Por tanto, vamos a dejar los valores extremos de esta variable como están.

Vamos a considerar a continuación los valores extremos en la tasa de alfabetización.

```
> boxplot(country_data$Literacy_rate ~ country_data$Region, data = country_data, main = "Tasa de alfabetización por zona", xlab ="Región", ylab= "Tasa de alfabetización")
```



Los primeros valores extremos que observamos están en la región Europa del este y pertenecen a Albania y Macedonia. Estos valores en cualquier caso son superiores al 85% de alfabetización, por lo que podemos considerar que sean plausibles, aunque estén por debajo de la media europea.

Miramos a continuación a la región de Sudamérica. Los valores extremos pertenecen a países de Centroamérica con unas circunstancias muy similares (Guatemala, Haiti, Honduras, NIcaragua). Como todos los valores extremos pertenecen a una región muy concreta de Sudamérica y estos valores indican que en este grupo de países la tasa de alfabetización es más baja que en el resto del continente.

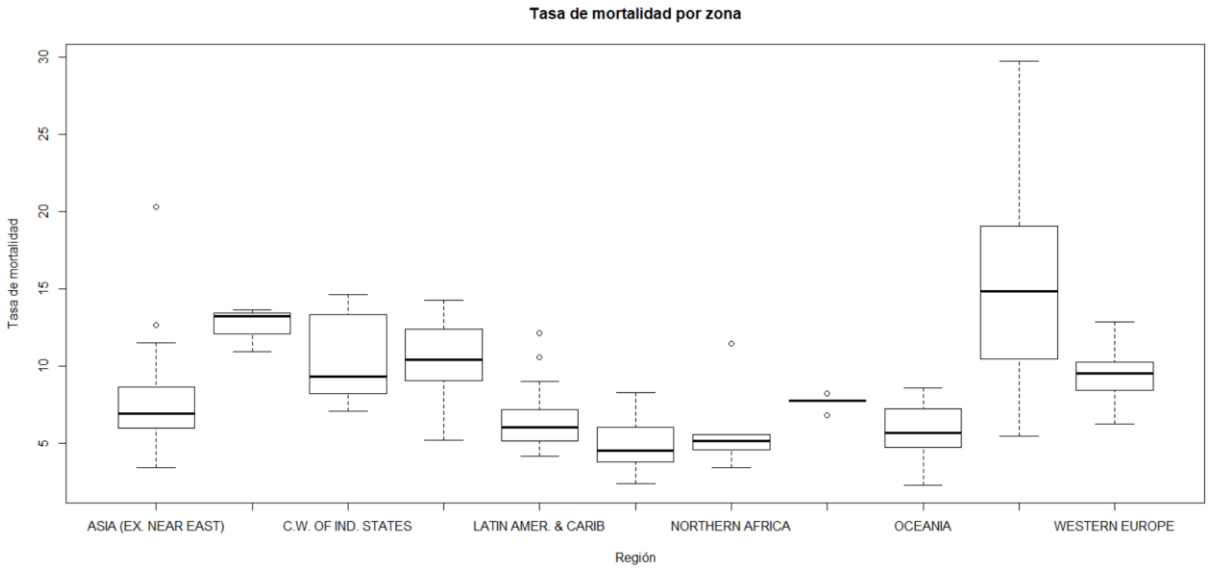
Pasamos a la siguiente región, Medio Oriente. Hay valores extremos que son mucho más bajos que en el resto de la región. Sin embargo, vemos que en Asia hay países con una tasa de alfabetización incluso más baja. Teniendo en cuenta que Medio Oriente pertenece a Asia, estos valores extremos son también posibles.

Por su parte en Oceanía, los valores extremos pertenecen a países como Papua Nueva Guinea, Vanuatu o Wallis y Futuna. Países muy pequeños, con unas circunstancias muy particulares que nos lleva a pensar que es posible que tengan una tasa de alfabetización muy alejada de la media del continente.

Por último, aunque en Europa Occidental haya valores extremos, todos los valores son superiores al 90%. Por tanto, no se puede considerar que estos valores sean imposibles. De esta forma, en esta variable dejamos los valores extremos como están.

Miramos a continuación al siguiente atributo, la tasa de mortalidad:

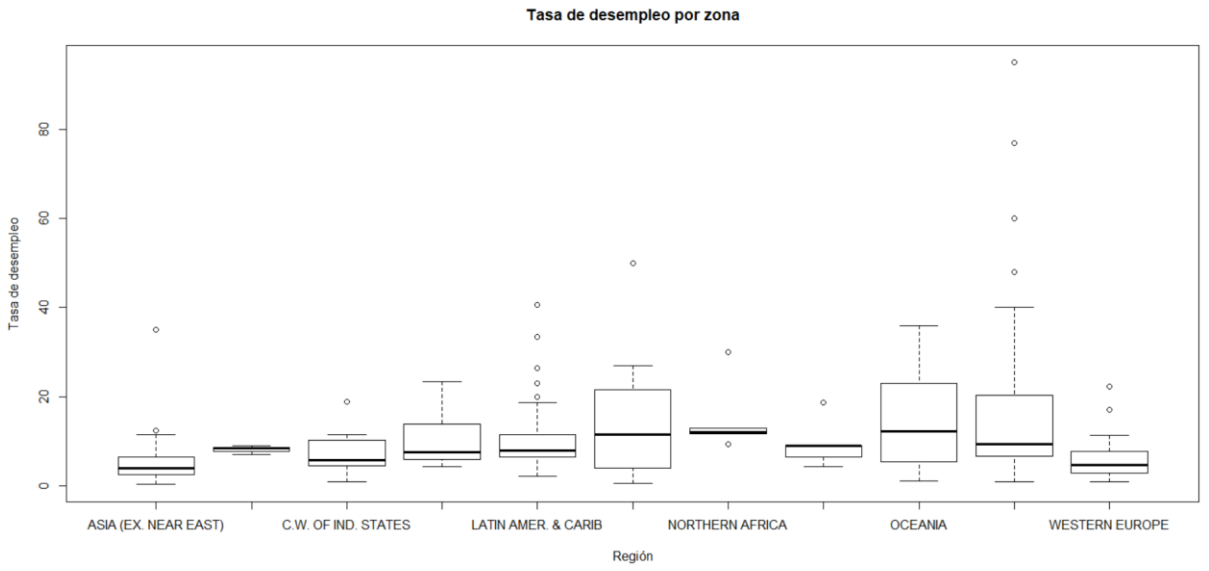
```
> boxplot(country_data$Deathrate ~ country_data$Region, data = country_data, main = "Tasa de mortalidad por zona", xlab ="Región", ylab= "Tasa de mortalidad")
```



En el caso de la tasa de mortalidad, vemos que, aunque algunos valores aparezcan como valores extremos en su región, estos no son valores que su puedan considerar imposibles. Sí que merecen mención uno de los valores extremos de la zona de Asia Oriental. Este valor pertenece a Afganistán y se debe a las circunstancias que ha vivido ese país en los últimos tiempos. De la misma manera, el valor extremo en la región del norte de África se corresponde con Sahara del Oeste, una región que también ha vivido unas circunstancias extremas, que explicarían este valor, tan superior a la media de la región.

Por tanto, estos valores de nuevo se dejan como están.

```
> boxplot(country_data$Unemployment_rate ~ country_data$Region, data = country_data, main = "Tasa de desempleo por zona", xlab ="Región", ylab= "Tasa de desempleo")
```



Mirando a la tasa de desempleo, de nuevo el valor extremo de Asia pertenece a Afganistán, y el de Medio Oriente Siria. De nuevo estos valores pueden ser considerados normales teniendo en cuenta las circunstancias. Por su parte, en la región de estados indoeuropeos, el valor extremo es de Armenia, y es de aproximadamente un 20%. No es un valor descabellado, teniendo en cuenta que países como España han experimentado una tasa de desempleo superior.

En cuanto a Sudamérica, vemos que países que también han sufrido circunstancias convulsas como Haití o Venezuela, aparecen como valores extremos, con una tasa de desempleo cercana al 40%. De nuevo, podemos aceptar estos valores, así como el valor extremo del norte de África, que pertenece a Libia.

Sin embargo, en África subsahariana, los valores extremos están muy alejados del resto de valores, incluso alcanzando el 95%. Concretamente, hay 4 valores que parecen demasiado altos para ser del dominio de este atributo. Vamos a comprobar los valores reales de estos países: Zimbabwe, Burkina Faso, Djobouti y Senegal.

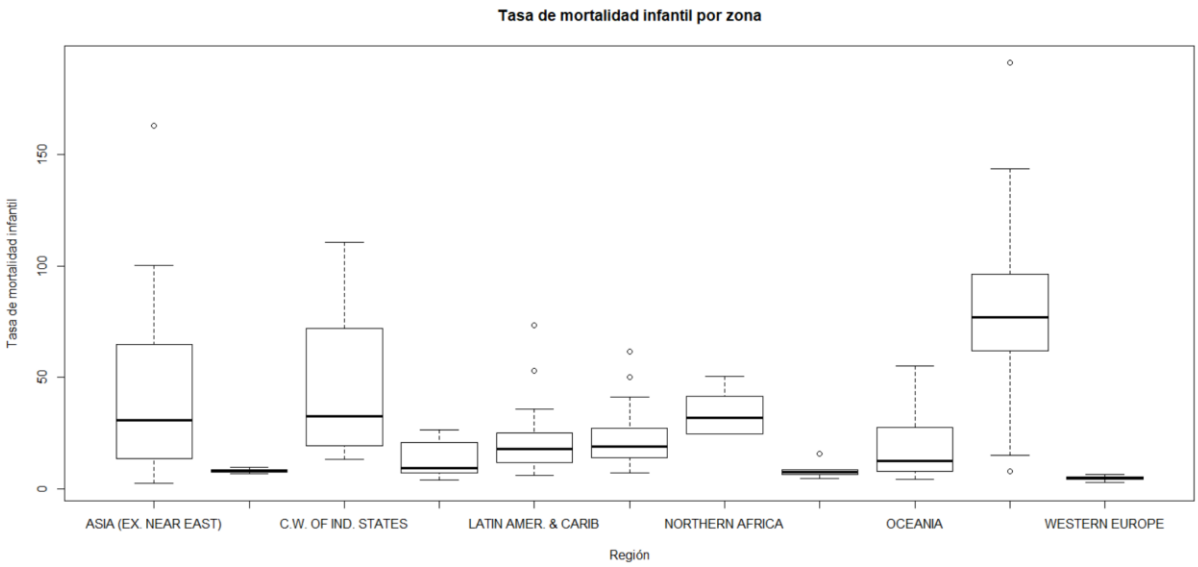
- Zimbabwe: 4.17
- Burkina Faso: 3.4
- Djibouti: 13%
- Senegal: 8%

Al comprobar el valor real para estas observaciones, vemos que los datos que teníamos eran erróneos y por tanto, vamos a reemplazar estos valores.

```
> country_data$Unemployment_rate[227]
[1] 95
>
> country_data$Unemployment_rate[227]<- 4.2
> country_data$Unemployment_rate[32]
[1] 77
> country_data$Unemployment_rate[32]<- 3.4
>
> country_data$Unemployment_rate[56]
[1] 60
> country_data$Unemployment_rate[56]<- 13.0
>
> country_data$Unemployment_rate[181]
[1] 48
>
> country_data$Unemployment_rate[181]<- 8.0
>
```

El resto de valores se dejan como estaban.

```
> boxplot(country_data$Infant_mortality ~ country_data$Region, data = country_data, main = "Tasa de mortalidad infantil por zona", xlab ="Región", ylab= "Tasa de mortalidad infantil")
```



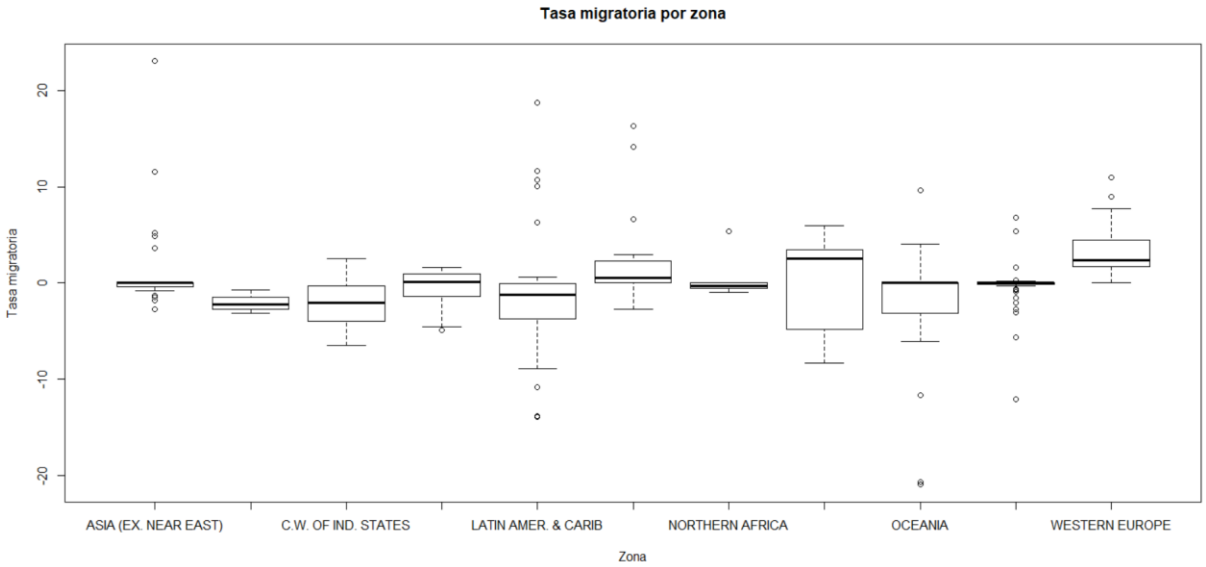
A continuación, comprobamos los valores extremos de la variable mortalidad infantil.

De nuevo Afganistán presenta una mortalidad infantil mucho más elevada que el resto de Asia, y se trata de un valor veraz. En Sudamérica Haití y Bolivia son los valores extremos, pero que son plausibles. Por su parte, en medio Oriente Iraq presenta también un valor elevado, aunque no tanto como para considerarlo extraño.

Por su parte, en África hay un valor que pertenece a Angola y que es tan elevado que de nuevo parece erróneo. Sin embargo, en este caso, el valor es correcto y por tanto, de nuevo para esta variable dejamos los valores como están.

Por último, vamos a representar los valores de la tasa migratoria de cada país dividida por regiones, ya que esto nos permitirá ver más fácilmente si hay algún valor erróneo.

```
>
> boxplot (country_data$Net_migration ~ country_data$Region, data = country_data, main = "Tasa migratoria por zona", xlab = "Zona", ylab="Tasa migratoria")
>
```



El primer valor que resulta llamativo es el país de Asia que tiene una tasa de migración positiva superior a 20. Inicialmente, se ha considerado que este dato es erróneo. Sin embargo, este dato se corresponde con Afganistán y es resultado de que muchos refugiados que emigraron por la guerra han regresado al país. Por tanto, este valor es correcto.

El resto de valores extremos se debe principalmente a países con unas circunstancias especiales como las Islas Vírgenes Británicas en Sudamérica, Singapur en Asia y otros territorios como colonias y países con tratados migratorios especiales.

No obstante, hay un valor que es erróneo es el de Western Sahara. Vemos como hay un valor extremo en la región de África del norte. Éste se corresponde con Western Sahara y se debe a que el valor introducido anteriormente para este país pertenece a un año distinto al del resto de países. Por tanto, nos aseguramos de averiguar el valor correcto y sustituimos el valor anterior por el dato correcto. No vamos a realizar ningún cambio sobre el resto de los datos.

Con esto concluye la fase de preparación de los datos. En este momento nuestro conjunto de datos está preparado para realizar los análisis correspondientes.

2.3.5. Exportación de los datos preprocesados

Una vez finalizada la fase de integración, validación y limpieza del conjunto de datos, procedemos a exportar un fichero con el conjunto de datos generado. El fichero se denominará "country_data.csv", y se generará mediante la función *write.csv().*

```
> write.csv(country_data, "country_data.csv")
>
```

2.4. Análisis de los datos

Después de haber manipulado y transformado los datos, hemos obtenido un dataset sobre el que vamos a realizar los análisis indicados anteriormente.

Antes de nada, vamos a utilizar la función *summary* para tener una idea de la distribución de valores de cada variable. A continuación se puede ver un resumen de los variables que tiene cada atributo en nuestro dataset.

```
> summary(country_data)
        Country                    Region     Population            Area          Net_migration       GDP_per_capita  Literacy_rate      Birthrate       Deathrate
 Afghanistan   :  1   SUB-SAHARAN AFRICA  :51   Min.   :7.026e+03   Min.   :       2   Min.   :-20.990000   Min.   :  500   Min.   : 17.60   Min.   : 5.50   Min.   : 2.290
 Albania       :  1   LATIN AMER. & CARIB :45   1st Qu.:4.376e+05   1st Qu.:    4648   1st Qu.: -1.135000   1st Qu.: 1900   1st Qu.: 74.50   1st Qu.:12.56   1st Qu.: 5.910
 Algeria       :  1   ASIA (EX. NEAR EAST):28   Median :4.787e+06   Median :   86600   Median :  0.000000   Median : 5500   Median : 92.60   Median :18.79   Median : 8.100
 American Samoa:  1   WESTERN EUROPE      :28   Mean   :2.874e+07   Mean   :  598227   Mean   : -0.007753   Mean   : 9658   Mean   : 83.28   Mean   :22.01   Mean   : 9.249
 Andorra       :  1   OCEANIA             :21   3rd Qu.:1.750e+07   3rd Qu.:  441811   3rd Qu.:  0.980000   3rd Qu.:15700   3rd Qu.: 98.00   3rd Qu.:29.77   3rd Qu.:10.800
 Angola        :  1   NEAR EAST           :16   Max.   :1.314e+09   Max.   :17075200   Max.   : 23.060000   Max.   :55100   Max.   :100.00   Max.   :50.73   Max.   :29.740
 (Other)       :221   (Other)             :38
Unemployment_rate Infant_mortality      GDP          Main_sector
Min.   : 0.30     Min.   :  2.29   Min.   :     12   Length:227
1st Qu.: 4.40     1st Qu.:  8.11   1st Qu.:   2559   Class :character
Median : 8.00     Median : 20.97   Median :  16289   Mode  :character
Mean   :11.24     Mean   : 35.33   Mean   : 232205
3rd Qu.:12.70     3rd Qu.: 55.34   3rd Qu.:  86666
Max.   :95.00     Max.   :191.19   Max.   :11281191
```

2.4.1. Selección de grupos de datos a analizar

A continuación, tenemos la posibilidad de agrupar observaciones de nuestro conjunto de datos. Esto puede ser útil para comparar distintos grupos. En nuestro caso, puede ser interesante comparar la situación entre distintas regiones. Podemos crear un grupo para cada región mediante los siguientes comandos:

```
> country_data.asia <- country_data[country_data$Region == "ASIA (EX. NEAR EAST)",]
> country_data.europa_este <- country_data[country_data$Region == "EASTERN EUROPE                     ",
> country_data.africa_norte <- country_data[country_data$Region == "NORTHERN AFRICA",]
> country_data.oceania <- country_data[country_data$Region == "OCEANIA",]
> country_data.europa <- country_data[country_data$Region == "WESTERN EUROPE",]
> country_data.africa <- country_data[country_data$Region == "SUB-SAHARAN AFRICA",]
> country_data.sudamerica <- country_data[country_data$Region == "LATIN AMER. & CARIB",]
> country_data.medio_oriente <- country_data[country_data$Region == "NEAR EAST",]
> country_data.norteamerica <- country_data[country_data$Region == "NORTHERN AMERICA",]
>
```

De la misma manera, una comparación interesante es la de los países en función de su producto interior bruto per cápita. Está claro que las condiciones de vida de un país serán diferentes en función de esta variable. Así pues, comparando atributos de países con una renta per cápita alta, con los que tienen una renta per cápita baja.

Establecemos dos grupos, dividiendo los países en función de si su renta per cápita está por encima de la media de este atributo o por debajo.

```
> country_data.PIB_pc_alto <- country_data[country_data$GDP_per_capita > 9660,]
> country_data.PIB_pc_bajo <- country_data[country_data$GDP_per_capita < 9660,]
>
> country_data.PIB_alto <- country_data[country_data$GDP > 20000,]
> country_data.PIB_bajo <- country_data[country_data$GDP < 20000,]
>
```

2.4.2. Comprobación de la normalidad y homocesdasticidad

El siguiente de la fase analítica será comprobar si las variables cuantitativas de nuestro conjunto de datos siguen una distribución normal. Para ello, utilizamos el paquete *nortest.* Además, establecemos como nivel de significación 0.05.

Para esta comprobación usamos la prueba de normalidad *Anderson-Darling*. Por lo tanto, establecemos un bucle que recorra todas las variables numéricas de nuestro dataset y compruebe su p-valor. Si el p-valor es superior al nivel de significación establecido, se considera que la variable sigue una distribución normal.

```
> library (nortest)
>
> alpha = 0.05
> col.names = colnames(country_data)
>
> for (i in 1:ncol(country_data)){
+     if (i == 1) cat("Variables que no siguen una distribución normal:\n")
+     if (is.integer (country_data[,i]) | is.numeric (country_data[,i])) {
+         p_val = ad.test(country_data[,i])$p.value
+         if(p_val < alpha) {
+             cat(col.names[i])
+
+
+             if (i < ncol(country_data) - 1) cat(",")
+             if (i %% 3 == 0) cat("\n")
+           }
+         }
+ }
Variables que no siguen una distribución normal:
Population,
Area,Net_migration,GDP_per_capita,
Literacy_rate,Birthrate,Deathrate,
Unemployment_rate,Infant_mortality,GDP
>
```

Observando el resultando de la prueba de normalidad, podemos ver que ninguna de nuestras variables cuantitativas sigue una distribución normal.

La siguiente prueba que vamos a realizar es la de la homogeneidad de varianzas mediante la aplicación de la prueba de Fligner-Killen. El test de Fligner-Killen es el más adecuado, ya que como hemos visto anteriormente, las variables no cumplen la condición de normalidad. En este caso, vamos a comparar la homogeneidad de las varianzas de las variables cualitativas de nuestro conjunto de datos para cada cada una de las regiones que aparecen en el *dataset.*

```
> fligner.test (Net_migration ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  Net_migration by Region
Fligner-Killeen:med chi-squared = 38.309, df = 10, p-value = 3.353e-05
```

```
> fligner.test (GDP_per_capita ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  GDP_per_capita by Region
Fligner-Killeen:med chi-squared = 53.266, df = 10, p-value = 6.645e-08
```

```
> fligner.test (Literacy_rate ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  Literacy_rate by Region
Fligner-Killeen:med chi-squared = 83.824, df = 10, p-value = 8.9e-14
```

```
> fligner.test (Birthrate ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  Birthrate by Region
Fligner-Killeen:med chi-squared = 58.819, df = 10, p-value = 6.059e-09
```

```
> fligner.test (Deathrate ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  Deathrate by Region
Fligner-Killeen:med chi-squared = 68.811, df = 10, p-value = 7.517e-11
```

```
> fligner.test (Unemployment_rate ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  Unemployment_rate by Region
Fligner-Killeen:med chi-squared = 39.681, df = 10, p-value = 1.929e-05
```

```
> fligner.test (Infant_mortality ~ Region, data = country_data)

        Fligner-Killeen test of homogeneity of variances

data:  Infant_mortality by Region
Fligner-Killeen:med chi-squared = 94.943, df = 10, p-value = 5.575e-16
```

Para cada una de las variables, el resultado del test arroja un p-valor menor que el nivel de significación. Al ser menor que el nivel de significación, debemos aceptar que la varianza entre las distintas regiones del conjunto de datos no es homogénea para ninguna variable.

Este resultado indica que existen diferencias estadísticamente significativas para cada una de las regiones que componen el conjunto de datos. A tenor de este resultado, las pruebas estadísticas se realizarán sobre los grupos indicados en apartados anteriores y no sobre la totalidad de los datos, ya que obtendremos resultados más precisos.

2.5. Pruebas estadísticas

2.5.1. Test de Wilcoxon

Antes de analizar las regiones que componen nuestro conjunto de datos en profundidad, vamos a responder una de las cuestiones las que queríamos dar respuesta con este análisis. ¿Se puede demostrar estadísticamente que la tasa de inmigración es diferente según el PIB y el PIB per cápita de cada país?

Para ello, vamos a categorizar cada país en función de su PIB y su PIB per cápita, como se puede ver a continuación.

Creamos una columna para categorizar países por PIB

```
> country_data$GDP_category <- ifelse((country_data$GDP > 20000), "Alto",
+ ifelse ((country_data$GDP < 20000), "Bajo", "Desconocido"))
```

Creamos una columna para categorizar países por PIB per cápita

```
> country_data$GDP_pc_category <- ifelse((country_data$GDP_per_capita > 9660), "Alto",
+ ifelse ((country_data$GDP_per_capita < 9660), "Bajo", "Desconocido"))
```

Diferencias estadísticas en tasa de migración para países con PIB alto

```
> wilcox.test(Net_migration ~ GDP_category, data = country_data)

        Wilcoxon rank sum test with continuity correction

data:  Net_migration by GDP_category
W = 7010.5, p-value = 0.2325
alternative hypothesis: true location shift is not equal to 0
```

En el resultado del test, vemos que el p-valor obtenido es mayor que el nivel de significación. Esto indica que no se puede afirmar que haya diferencias estadísticas en la variable tasa migratoria, en función del PIB de un país.

A continuación, repetimos la prueba para la variable “GDP_per_capita”, es decir, el PIB per cápita de un país.

```
> wilcox.test(Net_migration ~ GDP_pc_category, data = country_data)

        Wilcoxon rank sum test with continuity correction

data:  Net_migration by GDP_pc_category
W = 9039.5, p-value = 1.927e-12
alternative hypothesis: true location shift is not equal to 0
```

En este caso, el p-valor es menor que el nivel de significación. Por tanto, sí que se puede afirmar que existe una diferencia estadística en la tasa de migración en función de si los países tienen un PIB per cápita bajo o elevado.

Parece lógico pensar que los países un PIB per cápita elevado tendrán una tasa migratoria elevada, es decir, reciben más inmigrantes que aquellos con un PIB per cápita más bajo. En estos casos la tasa de migración será inferior, un mayor número de habitantes emigran de estos países.

2.5.2. Variables con mayor influencia sobre la tasa de migración

Con este análisis, el objetivo es averiguar qué variables continuas tienen más influencia sobre la tasa de inmigración. Como hemos visto anteriormente, existen diferencias estadísticas entre todas las regiones que componen nuestro conjunto de datos. Esto sugiere que en cada región, las variables que tengan más influencia sobre la tasa migratoria, pueden ser diferentes.

Vamos a centrarnos en tres regiones en la que la tasa de migración en la mayoría de países es negativa, es decir, experimentan más emigración que inmigración: Europa del este, Sudamérica y África subsahariana.

Para ello, vamos a realizar un análisis de correlación para estas regiones y comparar el coeficiente de correlación de cada variable. De nuevo, como tenemos atributos que no siguen una distribución normal, el método más adecuado será el método de Spearman, y en concreto la función *cor.test().*

De esta forma, lo que hacemos es un bucle que recorra el dataframe y ejecute la función *cor.test()* entre “Net_migration” y cada una de las variables numéricas del conjunto de datos. Rellenamos la matriz con el valor del coeficiente de correlación y el p-valor para cada variable.

Una vez hallamos averiguado cuál es la variable con una mayor correlación con la variable explicada, vamos a representar gráficamente el modelo de regresión entre la variable explicativa y la tasa de migración.

EUROPA DEL ESTE

Cabe destacar que en este grupo vamos a incluir todos los países categorizados como “EASTERN EUROPE”, “C.W. INDO. STATES” y “BALTICS”. Es posible agrupar estos países, ya que siguen una distribución homogénea.

Se puede asumir que estos tres grupos pueden agruparse para el mismo análisis. Sin embargo, se ha comprobado que estas tres regiones tienen una varianza homogénea antes de agruparlas. En este caso, el p-value es mayor que el nivel de significación y por tanto, aceptamos que la varianza es homogénea entre las tres regiones.

```
> country_data.europa_este <- country_data[country_data$Region == "EASTERN EUROPE" | country_data$Region == "C.W. OF IND. STATES" | country_data$Region == "BALTICS",]
>
> fligner.test (Net_migration ~ Region, data = country_data.europa_este)

        Fligner-Killeen test of homogeneity of variances

data:  Net_migration by Region
Fligner-Killeen:med chi-squared = 1.9013, df = 2, p-value = 0.3865

> |
```

Procedemos pues, con el análisis de Europa del este. Para ello, creamos una matriz, en la que mostraremos el coeficiente de correlación y el p-valor. Este valor puede proporcionar información acerca del peso estadístico de la correlación obtenida.

```
> matrix_eureste <- matrix(nc =2 , nr = 0)
> colnames(matrix_eureste) <- c("estimate", "p-value")
```

Creamos el bucle que recorra el conjunto de datos ejecutando la función *cor.test()* sobre las variables cuantitativas.

```
> for(i in 1:(ncol(country_data.europa_este)-1)) {
+     if (is.integer(country_data.europa_este[,i]) | is.numeric(country_data.europa_este[,i])){
+         spearman_test = cor.test(country_data.europa_este[,i], country_data.europa_este[,5], method = "spearman")
+         corr_coef = spearman_test$estimate
+         p_val = spearman_test$p.value
+
```

Rellenamos la matriz con los valores obtenidos.

```
+ pair = matrix(ncol = 2, nrow = 1)
+ pair[1][1] = corr_coef
+ pair[2][1] = p_val
+ matrix_eureste <- rbind(matrix_eureste, pair)
+ rownames (matrix_eureste)[nrow(matrix_eureste)] <- colnames(country_data.europa_este)[i]
+ }
+ }
```

A continuación, se puede ver la matriz obtenida. Observamos que el coeficiente de correlación siempre tiene un valor de entre 1 y -1. Las variables más correlaciones con el atributo

"Net_migration" son aquellas cuyo coeficiente de correlación es más próximo a 1 ó a -1. Por el contrario, si el valor del coeficiente de correlación es cercano a 0, la correlación será más baja.

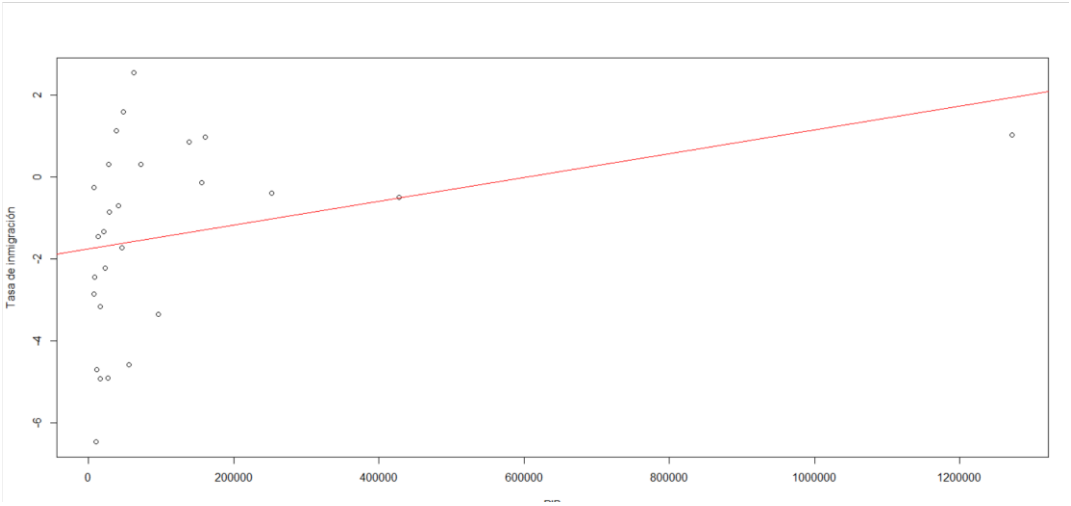
```
> print (matrix_eureste)
                     estimate     p-value
Population         0.25213675 0.203743933
Area               0.09218559 0.646305207
Net_migration      1.00000000 0.000000000
GDP_per_capita     0.50312930 0.007471365
Literacy_rate      0.37402591 0.054611985
Birthrate         -0.45726496 0.017399420
Deathrate          0.43101343 0.025765645
Unemployment_rate -0.26717558 0.177904419
Infant_mortality  -0.50854701 0.007441407
GDP                0.55738706 0.002947282
> 
```

Las variables más relevantes son por tanto, "GDP", "Infant_mortality" y "GDP_per_capita", con un coeficiente de correlación superior a 0.5.

Una vez hemos comprobado cuál es la variable con una mayor correlación, representamos gráficamente el modelo de regresión. Para ello simplemente usamos la función *plot* y utilizamos *abline* y *lm* para representar la recta de regresión.

```
> plot(country_data.europa_este$GDP, country_data.europa_este$Net_migration, xlab = "PIB", ylab= "Tasa de inmigración")
> abline(lm(country_data.europa_este$Net_migration ~ country_data.europa_este$GDP), col = "red")
```

El modelo de regresión obtenido es el siguiente para la variable GDP por tasa de migración en Europa del Este es el siguiente:



Se puede ver en la gráfica que a mayor PIB, mayor tiende a ser la tasa migratoria, es decir son variables directamente proporcionales. Esto quiere decir que a medida que el PIB de un país es más elevado, éste tiende a recibir más inmigrantes.

SUDAMÉRICA

Repetimos la misma operación pero en este caso para los países de Sudamérica y el Caribe.

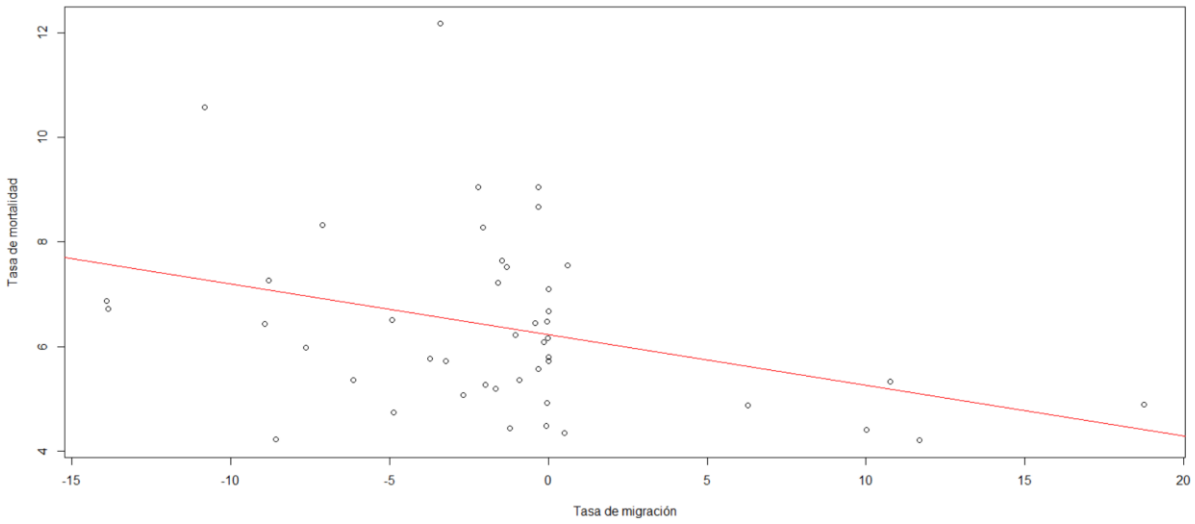
```
> matrix_sud <- matrix(nc =2 , nr = 0)
> colnames(matrix_sud) <- c("estimate", "p-value")
>
> for(i in 1:(ncol(country_data.sudamerica)-1)) {
+     if (is.integer(country_data.sudamerica[,i]) | is.numeric(country_data.sudamerica[,i])){
+         spearman_test = cor.test(country_data.sudamerica[,i], country_data.sudamerica[,5], method = "spearman")
+         corr_coef = spearman_test$estimate
+         p_val = spearman_test$p.value
+
+ pair = matrix(ncol = 2, nrow = 1)
+ pair[1][1] = corr_coef
+ pair[2][1] = p_val
+ matrix_sud <- rbind(matrix_sud, pair)
+ rownames (matrix_sud)[nrow(matrix_sud)] <- colnames(country_data.sudamerica)[i]
+ }
+ }
```

```
> print(matrix_sud)
                   estimate    p-value
Population      -0.06319813 0.68002518
Area            -0.01064321 0.94467952
Net_migration    1.00000000 0.00000000
GDP_per_capita   0.33512392 0.02442798
Literacy_rate    0.23550561 0.11938534
Birthrate       -0.18733319 0.21785555
Deathrate       -0.34196654 0.02149194
Unemployment_rate -0.32266149 0.03063684
Infant_mortality  -0.22695972 0.13379800
GDP              0.02853471 0.85239146
```

Curiosamente, en este caso los atributos con mayor relación sobre la tasa migratoria son diferentes con respecto a Europa del este. En Sudamérica, las variables más influyentes son “Deathrate” (relación negativa), “GDP_per_capita” y “Unemployment_rate”.

Vemos a continuación, el modelo de regresión para la tasa de mortalidad y tasa de migración.

```
> plot(country_data.sudamerica$Net_migration, country_data.sudamerica$Deathrate, xlab = "Tasa de migración", ylab= "Tasa de mortalidad")
> abline(lm(country_data.sudamerica$Deathrate ~ country_data.sudamerica$Net_migration), col = "red")
```



Es fácil ver que estas variables son inversamente proporcionales. Es decir, a medida que la tasa de mortalidad disminuye, la tasa de migración tiende a ser más elevada. Es decir, los países con tasa de mortalidad baja atraen más inmigrantes.

MISMA OPERACIÓN ÁFRICA

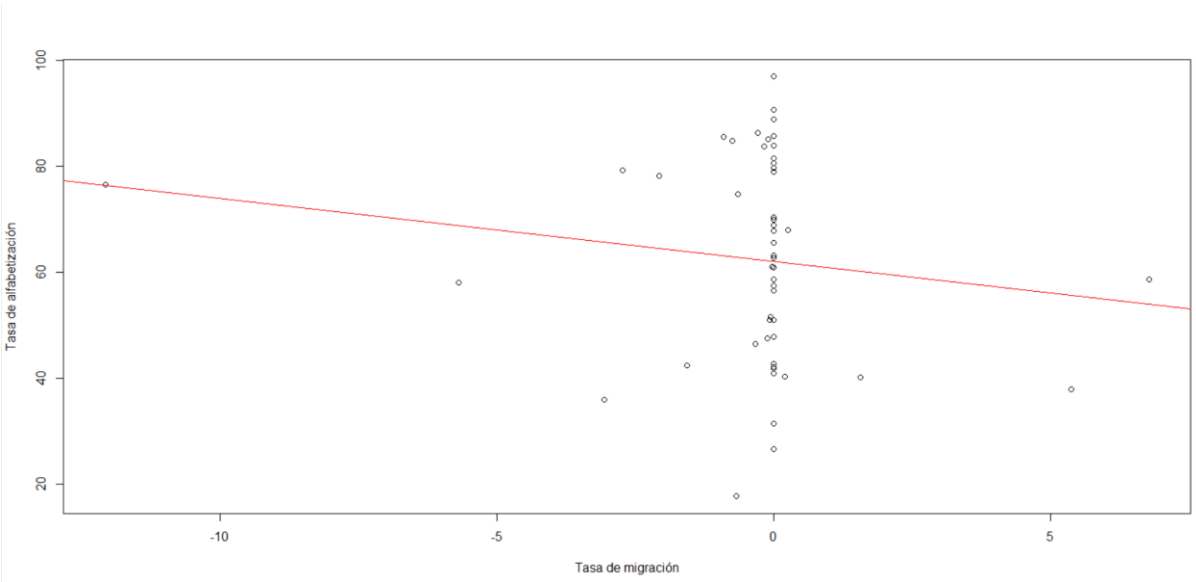
El tercer grupo que vamos a analizar es la región de África subsahariana.

```
> matrix_africa <- matrix(nc =2 , nr = 0)
> colnames(matrix_africa) <- c("estimate", "p-value")
>
> for(i in 1:(ncol(country_data.africa)-1)) {
+     if (is.integer(country_data.africa[,i]) | is.numeric(country_data.africa[,i])){
+         spearman_test = cor.test(country_data.africa[,i], country_data.africa[,5], method = "spearman")
+         corr_coef = spearman_test$estimate
+         p_val = spearman_test$p.value
+
+ pair = matrix(ncol = 2, nrow = 1)
+ pair[1][1] = corr_coef
+ pair[2][1] = p_val
+ matrix_africa <- rbind(matrix_africa, pair)
+ rownames (matrix_africa)[nrow(matrix_africa)] <- colnames(country_data.africa)[i]
+ }
+ }
```

```
> print(matrix_africa)
                      estimate   p-value
Population          0.029271181 0.8384321
Area                0.001337262 0.9925692
Net_migration       1.000000000 0.0000000
GDP_per_capita     -0.075327165 0.5993366
Literacy_rate      -0.149702446 0.2944046
Birthrate           0.128723856 0.3679949
Deathrate           0.090564408 0.5273698
Unemployment_rate   0.110892184 0.4385231
Infant_mortality    0.134072904 0.3482550
GDP                -0.018771197 0.8959800
>
```

Se puede ver que en este caso los valores de correlación estimados son más bajos que en las otras regiones. La variable con más influencia sería “Literacy_rate”, que es negativa. Si bien hay que recalcar que es complicado establecer valores para la correlación, ya que la tasa de migración en muchas de las observaciones es de 0.

```
> plot(country_data.africa$Net_migration, country_data.africa$Literacy_rate, xlab = "Tasa de migración", ylab= "Tasa de alfabetización")
> abline(lm(country_data.africa$Literacy_rate ~ country_data.africa$Net_migration), col = "red")
```



Representamos el modelo de regresión para la tasa de migración por la tasa de alfabetización. Sugiere una relación inversamente proporcional. A mayor tasa de alfabetización, más emigrantes dejan el país.

2.5.3. Modelo de regresión lineal

- **¿Se puede predecir la tasa de migración de un país a partir de sus indicadores económicos?**

Los modelos de regresión lineal son muy útiles para realizar predicciones sobre el valor de una variable dadas unas condiciones determinadas. En el apartado anterior hemos representado un modelo de regresión entre la tasa de migración y la variable explicativa. En este apartado, vamos a crear un modelo de regresión utilizando tanto las variables cuantitativas como cualitativas de nuestro conjunto de datos. La finalidad es poder realizar predicciones sobre la tasa de inmigración de un país si se producen cambios en las circunstancias del mismo. Por ejemplo, sería interesante predecir la tasa de migración de un país si su PIB per cápita aumenta en un uno por ciento.

Para obtener un modelo de regresión lo más eficiente posible, vamos a obtener varios modelos de regresión y calcular el coeficiente de determinación de cada uno de ellos. Así, escogeremos el que tenga mayor coeficiente de determinación, de forma que podamos obtener unas predicciones tan precisas como sea posible. De nuevo, vamos a crear un modelo para cada una de estas regiones.

EUROPA DEL ESTE

Renombramos las variables, para que sea más sencillo trabajar con ellas:

```
> # Variable a predecir
>
> tasa_migracion = country_data.europa_este$Net_migration
>
> # Variables cuantitativas con mayor coeficiente de correlación con respecto a la tasa migratoria
>
> PIB_per_capita = country_data.europa_este$GDP_per_capita
> mortalidad_infantil = country_data.europa_este$Infant_mortality
> tasa_desempleo = country_data.europa_este$Unemployment_rate
> tasa_nacimientos = country_data.europa_este$Birthrate
> tasa_alfabetizacion = country_data.europa_este$Literacy_rate
> tasa_mortalidad = country_data.europa_este$Deathrate
>
> # Regresos cualitativos
>
> sector_economico = country_data.europa_este$Main_Sector
>
```

```
> sector_economico = country_data.europa_este$Main_sector

> PIB = country_data.europa_este$GDP
```

Y procedemos a crear los modelos. Se han creado tres modelos. El primero contiene todas las variables del conjunto de datos. En los dos siguientes, hemos eliminado las variables menos relevantes según la tabla de correlación creada anteriormente:

Una vez creados los modelos, vamos a visualizar el resultado de los mismos en una matriz. Mostramos el coeficiente de determinación y nos quedaremos con el que tenga mejor coeficiente.

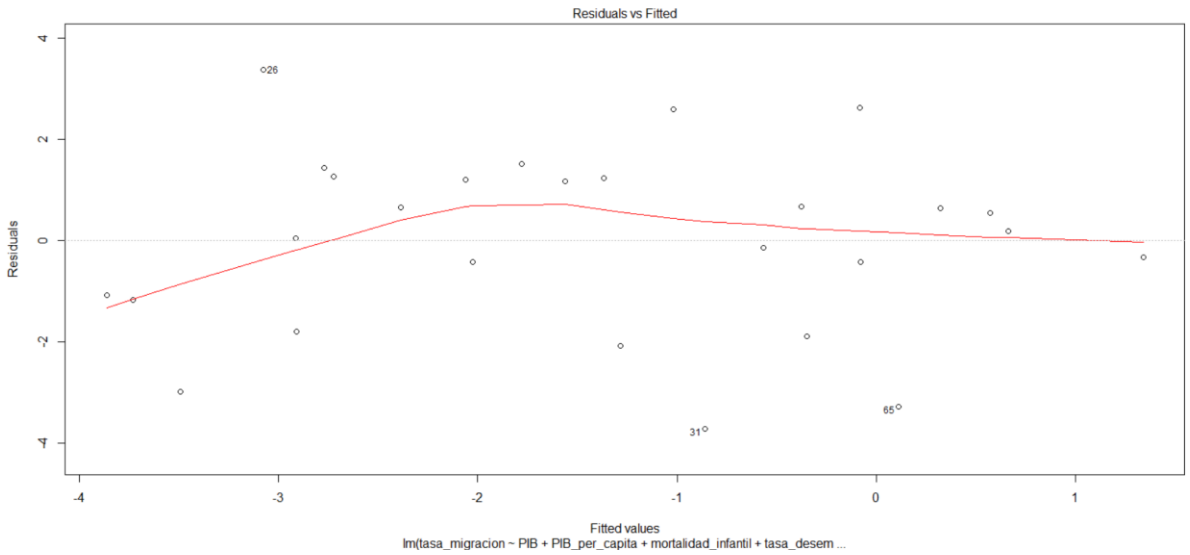
```
> modelo1_eureste <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil + tasa_desempleo + tasa_nacimientos +tasa_alfabetizacion + tasa_mortalidad + sector_economico, data =
> modelo2_eureste <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil + tasa_nacimientos + tasa_mortalidad, data = country_data.europa_este)
> modelo3_eureste <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil, data = country_data.europa_este)
>
>
> tabla.coeficientes <- matrix(c(1, summary(modelo1_eureste)$r.squared,
+                      2, summary(modelo2_eureste)$r.squared,
+                      3, summary(modelo3_eureste)$r.squared),
+                      ncol = 2, byrow = TRUE)
>
> colnames(tabla.coeficientes) <- c("Modelo", "coeficiente")
>
>  tabla.coeficientes
     Modelo coeficiente
[1,]      1   0.4007977
[2,]      2   0.3754966
[3,]      3   0.2967651
>
```

En este caso, se puede ver que el primer modelo (aunque tiene un resultado muy similar al segundo), es el modelo con mayor coeficiente de determinación y por tanto, será el que utilicemos a la hora de realizar predicciones. Cabe destacar, no obstante, que no es un coeficiente de determinación muy alto, por lo que la relación entre estas variables y la tasa de migración no es muy fuerte.

Habrá que tener esto en cuenta a la hora de realizar una predicción y tomar el resultado obtenido como una estimación y no como un resultado preciso.

```
> plot(modelo1_eureste)
```

Por último, representamos gráficamente el modelo obtenido. El gráfico representa la línea de predicción y los valores reales de la observación (residuos).



Ciertas observaciones están muy alejadas del valor estimado por el modelo. Por eso, el coeficiente de determinación del modelo no es muy elevado.

SUDAMÉRICA

Repetimos la operación para crear un modelo que nos permita predecir la tasa de migración de países de la región de Sudamérica.

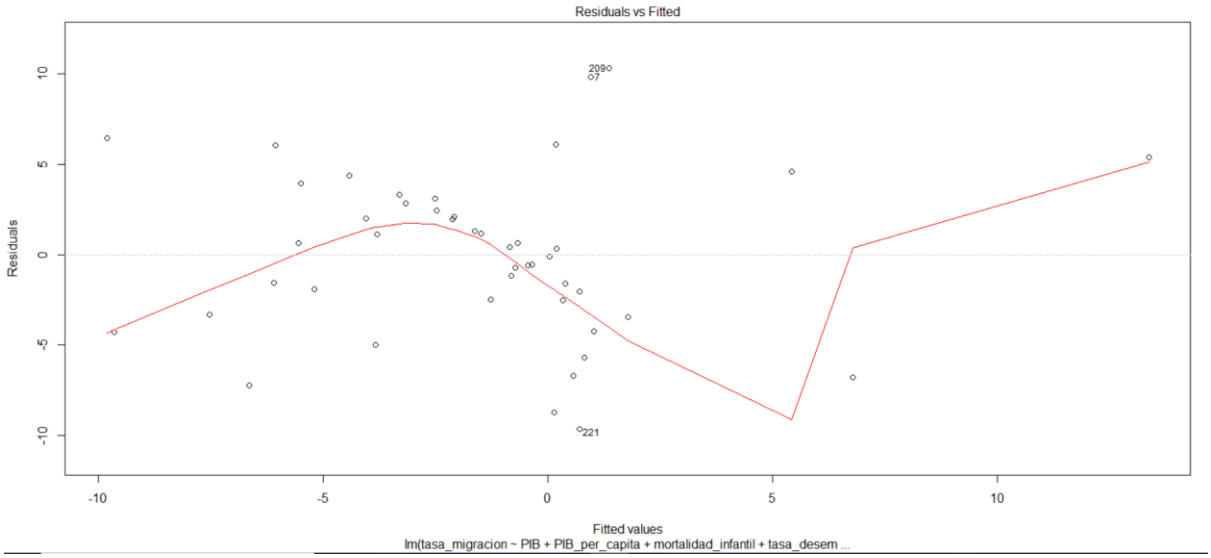
```
> tasa_migracion = country_data.sudamerica$Net_migration
>
> PIB_per_capita = country_data.sudamerica$GDP_per_capita
>
> PIB = country_data.sudamerica$GDP
>
> mortalidad_infantil = country_data.sudamerica$Infant_mortality
>
> tasa_desempleo = country_data.sudamerica$Unemployment_rate
>
> tasa_nacimientos = country_data.sudamerica$Birthrate
>
> tasa_alfabetizacion = country_data.sudamerica$Literacy_rate
>
> sector_economico = country_data.sudamerica$Main_sector
>
>
```

```
> modelo1_sud <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil + tasa_desempleo + tasa_nacimientos +tasa_alfabetizacion + tasa_mortalidad + sector_economico, data = count
>  modelo2_sud <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil + tasa_nacimientos + tasa_mortalidad + sector_economico, data = country_data.sudamerica)
> modelo3_sud <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil, data = country_data.sudamerica)
>
```

```
> tabla.coeficientes <- matrix(c(1, summary(modelo1_sud)$r.squared,
+                       2, summary(modelo2_sud)$r.squared,
+                       3, summary(modelo3_sud)$r.squared),
+                       ncol = 2, byrow = TRUE)
>
> colnames(tabla.coeficientes) <- c("Modelo", "coeficiente")
> tabla.coeficientes
     Modelo coeficiente
[1,]      1   0.4455591
[2,]      2   0.3970756
[3,]      3   0.2162465
>
```

De nuevo, el mejor modelo es el primero, que tiene un coeficiente de determinación mayor que el modelo creado para Europa del Este. Representamos gráficamente este modelo para obtener el siguiente gráfico:

```
> plot(modelo1_sud)
```



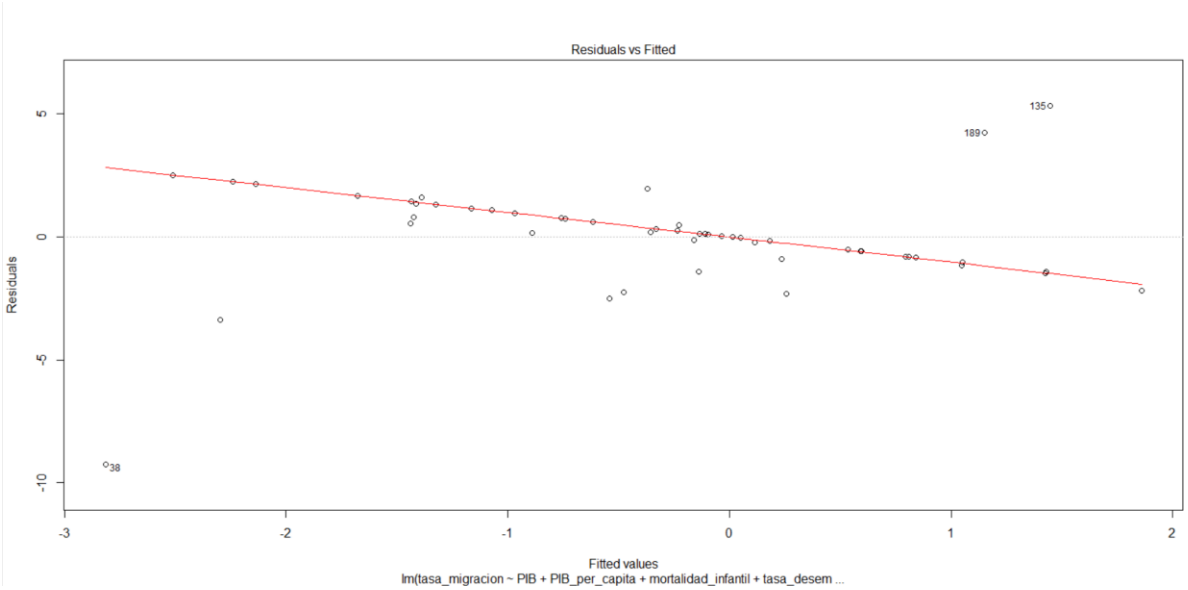
ÁFRICA

Vamos a mirar por último a la región de África subsahariana.

```
> tasa_migracion = country_data.PIB_pc_bajo$Net_migration
> PIB_per_capita = country_data.PIB_pc_bajo$GDP_per_capita
> PIB = country_data.PIB_pc_bajo$GDP
> mortalidad_infantil = country_data.PIB_pc_bajo$Infant_mortality
> tasa_desempleo = country_data.PIB_pc_bajo$Unemployment_rate
> tasa_nacimientos = country_data.PIB_pc_bajo$Birthrate
> tasa_alfabetizacion = country_data.PIB_pc_bajo$Literacy_rate
> sector_economico = country_data.PIB_pc_bajo$Main_sector
> tasa_mortalidad = country_data.PIB_pc_bajo$Deathrate
>
>
> modelo1 <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil + tasa_desempleo + tasa_nacimientos +tasa_alfabetizacion + tasa_mortalidad + sector_economico, data = country_d$
> modelo1 <- lm(tasa_migracion ~ PIB_per_capita + mortalidad_infantil + tasa_nacimientos +tasa_alfabetizacion + tasa_mortalidad, data = country_data.PIB_pc_bajo)
>
> modelo1 <- lm(tasa_migracion ~ PIB + PIB_per_capita + mortalidad_infantil + tasa_desempleo + tasa_nacimientos +tasa_alfabetizacion + tasa_mortalidad + sector_economico, data = country_d$
> modelo2 <- lm(tasa_migracion ~ PIB_per_capita + mortalidad_infantil + tasa_nacimientos +tasa_alfabetizacion + tasa_mortalidad, data = country_data.PIB_pc_bajo)
> modelo3 <- lm(tasa_migracion ~ mortalidad_infantil + tasa_nacimientos +tasa_alfabetizacion, data = country_data.PIB_pc_bajo)
>
> tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
+                                2, summary(modelo2)$r.squared,
+                                3, summary(modelo3)$r.squared),
+                                ncol = 2, byrow = TRUE)
>
> colnames(tabla.coeficientes) <- c("Modelo", "coeficiente")
>
> tabla.coeficientes
     Modelo coeficiente
[1,]      1  0.21557897
[2,]      2  0.10252061
[3,]      3  0.08677468
>
```

Se puede ver que los modelos obtenidos tienen los coeficientes de determinación más bajos. Es decir, el modelo va a ser menos fiable que los anteriores. Vamos de nuevo a representar el modelo gráficamente.

```
> plot(modelo1_afr)
```



Una vez hemos seleccionado el modelo más eficiente para cada región podemos proceder a realizar predicciones de la tasa de migración para un país, dado el valor de una serie de variables. Vamos a realizar predicciones para un país de Europa del este y otro de Sudamérica, ya que son modelos más fiables.

Los datos que hemos utilizado para "entrenar" nuestros modelos son valores de los países de años anteriores. Vamos a comprobar la precisión del modelo utilizando el valor de las variables para el año 2018.

```
> albania_new <- data.frame(PIB = 130400, PIB_per_capita = 4538, mortalidad_infantil= 7.8, tasa_desempleo = 13.8, tasa_nacimientos = 13.2, tasa_alfabetizacion = 97.6, tasa_mortalidad $
> predict (modelo1_eureste, albania_new)
       1
-2.316558
```

El modelo estima que la tasa de migración en Albania en 2018 fue de -2.3. El valor real fue de -3.3. Como se ha mencionado anteriormente, hay que ser cauto a la hora de utilizar los resultados. Sin embargo, sí que obtenemos una idea del valor de la tasa migratoria en ese país.

```
> ecuador_new <- data.frame(PIB = 189750, PIB_per_capita = 11500, mortalidad_infantil= 15.9, tasa_desempleo = 4.6, tasa_nacimientos = 17.6, tasa_alfabetizacion = 96.1, tasa_mortalidad$
> predict (modelo1_sud, ecuador_new)
         1
-0.1553499
```

Por su parte, para el modelo obtenido para Sudamérica, vamos a predecir la tasa migratoria en Ecuador. Usamos de nuevo los indicadores económicos de 2018 y el valor que devuelve el modelo es una tasa de migratoria de -0.15. El valor real es de 0. Vemos que este modelo es algo más preciso que el anterior, y de nuevo nos permite estimar cuál será la tasa de migración de un país a partir de unas condiciones dadas.

3. CONCLUSIONES

Así pues, partiendo de un conjunto de datos inicial que recoge una serie de indicadores económicos de prácticamente todos los países del mundo, hemos llevado a cabo un proceso de análisis para intentar responder las preguntas planteadas al inicio del proyecto. Se han realizado varias pruebas estadísticas sobre este conjunto de datos, con el objetivo fundamental de entender si se puede dar una explicación estadística a la tasa neta de migración que hay en cada país.

Con la prueba de correlación hemos intentado comprobar qué variables tienen más influencia sobre la tasa de migración. Además, mediante el contraste de hipótesis, se ha intentado demostrar que hay una relación estadística algo que puede parecer obvio de forma empírica: que el PIB per cápita de un país afecta directamente sobre la tasa migratoria de un país. Los resultados de estos resultados se han mostrado gráficamente tanto en forma de tablas, como representando el modelo de regresión de forma gráfica.

Por último, una vez averiguadas las relaciones que existen entre la tasa de migración y el resto de variables, se han creado modelos de predicción. Con estos, podemos estimar el valor de la tasa migratoria para unas condiciones concretas. De nuevo, el modelo de regresión ha sido representado gráficamente.

Antes de llevar a cabo estos análisis se ha realizado una fase de limpieza y preparación de los datos. Esto incluye el tratamiento de valores nulos y de valores extremos (outliers). El objetivo de esta fase has sido disponer de un conjunto de datos tan preciso como sea posible sin perder información. Por tanto, hemos ido viendo como tratar tanto los valores nulos como los extremos, estudiando cada variable y siguiendo una estrategia diferente para cada variable.

4. RECURSOS

Jiawei Han, Micheline Kamber and Jian Pei. *Data Mining: Concepts and Techniques* 3rd Ed. Burlington : Morgan Kaufmann, cop. 2012. 332 p.

Osborne, W. J. (2010). Best Practices in Dealing with Extreme Scores, *Newborn and Infant Nursing Reviews* (pp. 37-43) Retrieved from (http://www.sciencedirect.com/science/article/pii/S1527336909001779)

Squire, M. *Clean Data*, Packt Publishing, Limited, 2015. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/bibliouocsp-ebooks/detail.action?docID=2057549.

Subirats, L., Calvo, M., Trenard, P. (2019). *Introducción a La Limpieza y Análisis de Datos*. Editorial UOC.