Práctica 2: Limpieza y análisis de datos

Maite Gracia

25 de December, 2020

Contents

1 Descripción del dataset

Se ha decidido utilizar un dataset de la web Kaggle para la presente práctica. [**enlace**](https://www.kaggle.com/kemical/kickstarter-projects) 8 de la web Kickstarter. Kickstarter es una plataforma de micro mecenazgo, es decir, gente de todo el mundo ayuda a financiar las ideas y proyectos de pequeñas empresas o particulares.

En la web de Kickstarter se pueden encontrar miles de campañas que buscan financiación para desarrollar productos de todo tipo. Desde películas independientes, a juegos de mesa o ropa, peluches, libros etc. Cada una de estas campañas tendrá un periodo de tiempo en el que cualquiera podrá aportar dinero al proyecto y si se consigue llegar al límite de dinero requerido la campaña será fundada.

Yo personalmente utilicé Kickstarter hace unos años para lanzar una serie de productos lo cual es una de las razones por las que he elegido el presente dataset. El objetivo principal sería poder crear un modelo que predijera que probabilidad tiene cualquier tipo de producto de conseguir recaudar dinero mediante una campaña de Kickstarter antes de ser lanzado.

2 Integración y selección de los datos de interés a analizar

Las variables que componen el dataset son:

- ID: identificador interno de Kickstarter
- name: nombre del proyecto
- category: categoría específica en la que se encuentra el proyecto
- main_category: categoría principal de la campaña
- currency: divisa en la que se creó el proyecto
- deadline: fecha límite
- goal: cantidad de dinero que el creador necesita para completar el proyecto
- launched: fecha lanzamiento
- pledged: cantidad total aportada al proyecto
- state: condición en la que se encuentra el Proyecto (failed, successful, canceled, live, undefined)
- backers: total de mecenas.
- country: país en el que se encuentra el Proyecto.
- usd_pledged: conversión en dólares de la columna pledged hecha por Kickstarter
- usd_pledged_real: conversión en dólares de la columna pledged hecha a través de Fixer.io API
- usd_goal_real: conversión en dólares de la columna goal hecha a través de Fixer.io API

Antes de cargar el archivo en R se hace una inspección de los datos. Al tratarse de un archivo con extensión .csv, hay que cerciorarse del tipo de separador utilizado (en este caso la ",") y posteriormente se procede a su carga teniendo en cuenta el separador antes mencionado:

```
# Asignamos los datos del fichero cargado a una variable denominada dataSet
dataSet <- read.csv('../data/ks-projects-201801.csv')
nrow(dataSet)
```

```
## [1] 378661
```

```
names(dataSet)
```

```
##  [1] "ID"               "name"             "category"         "main_category"
##  [5] "currency"         "deadline"         "goal"             "launched"
##  [9] "pledged"          "state"            "backers"          "country"
## [13] "usd.pledged"      "usd_pledged_real" "usd_goal_real"
```

Vemos que el dataset original se compone de 378,661 muestras y 13 variables. Ya que se trata de una cantidad de muestras muy elevadas, se ha decidido aplicar una técnica para reducir la cantidad de estas, se empleará la técnica de muestreo aleatorio simple sin sustitución, es decir, se van a extraer 3000 muestras aleatorias del conjunto de datos, donde la probabilidad de escoger cada una de las muestras será la misma para todas, 1/378,661.

Para ello generaremos un fichero al que llamaremos sample_ks.csv que contendrá las 3000 muestras.

```
library(sampling)
indices <- sample( 1:nrow( dataSet ), 3000 )
dataSet <- dataSet[ indices, ]
```

A partir de ahora cuando se haga referencia al dataset, estaremos hablando del dataset que cotiene las 3000 muestras, no el dataset original.

3 Limpieza de los datos

```
# Muestra de las 5 primeras líneas del dataset completo
head(dataSet, 5)
```

```
##                ID                                                      name
## 105958  153850277                                        From Case to Coast
## 295519  574159336                    Little Readers Interactive eBook Publishing
## 22237  1112688723      The Beauty For Ashes Project Presents: You're Beautiful
## 142075 1721941336                               Rockwell: The Fluctuating Market
## 238877  284653041 Walkpad. Listen your music through the DualShock4 (Canceled)
##              category main_category currency   deadline goal
## 105958          Music         Music      USD 2015-01-18 1000
## 295519           Apps    Technology      USD 2014-06-16  299
## 22237    Performances         Dance      USD 2015-03-22 2500
## 142075 Tabletop Games         Games      EUR 2015-12-21 1500
## 238877        Gadgets    Technology      EUR 2015-11-27 5000
##                    launched pledged      state backers country usd.pledged
## 105958 2014-12-29 14:42:01     360     failed      11      US      360.00
## 295519 2014-05-22 04:00:08     314 successful       4      US      314.00
## 22237  2015-02-20 06:07:40    2700 successful      33      US     2700.00
## 142075 2015-12-01 15:38:16    6890 successful     234      BE     7294.41
## 238877 2015-10-28 18:55:51       0   canceled       0      IT        0.00
##        usd_pledged_real usd_goal_real
## 105958            360.0       1000.00
## 295519            314.0        299.00
## 22237            2700.0       2500.00
## 142075           7542.5       1642.05
## 238877              0.0       5290.00
```

```
# Análisis descriptivo del dataset
summary(dataSet)
```

```
##        ID                 name              category         main_category
##  Min.   :2.237e+05   Length:3000        Length:3000        Length:3000
##  1st Qu.:5.436e+08   Class :character   Class :character   Class :character
##  Median :1.061e+09   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.072e+09
##  3rd Qu.:1.601e+09
##  Max.   :2.147e+09
##
##    currency           deadline              goal           launched
##  Length:3000        Length:3000        Min.   :      1   Length:3000
##  Class :character   Class :character   1st Qu.:   2000   Class :character
##  Mode  :character   Mode  :character   Median :   5476   Mode  :character
##                                        Mean   :  28458
```

```
##                                      3rd Qu.:  15716
##                                      Max.   :9000000
##
##     pledged            state              backers        country
##  Min.   :      0.0   Length:3000        Min.   :    0   Length:3000
##  1st Qu.:     30.0   Class :character   1st Qu.:    2   Class :character
##  Median :    539.5   Mode  :character   Median :   11   Mode  :character
##  Mean   :   8623.1                      Mean   :  111
##  3rd Qu.:   4032.0                      3rd Qu.:   55
##  Max.   :1936825.0                      Max.   :35549
##
##   usd.pledged        usd_pledged_real    usd_goal_real
##  Min.   :      0.0   Min.   :      0.0   Min.   :      1
##  1st Qu.:     16.7   1st Qu.:     30.0   1st Qu.:   2000
##  Median :    332.6   Median :    538.4   Median :   5431
##  Mean   :   6544.3   Mean   :   8380.4   Mean   :  26521
##  3rd Qu.:   2939.0   3rd Qu.:   3891.8   3rd Qu.:  15289
##  Max.   :1936825.0   Max.   :1936825.0   Max.   :7112938
##  NA's   :29
```

```
# Comprobamos si hay NA en el dataset original
sapply(dataSet, function(x) sum(is.na(x)))
```

```
##               ID             name         category    main_category
##                0                0                0                0
##         currency         deadline             goal         launched
##                0                0                0                0
##          pledged            state          backers          country
##                0                0                0                0
##      usd.pledged usd_pledged_real    usd_goal_real
##               29                0                0
```

3.1 Normalización de los datos

Basándonos en la estadística descriptiva de la muestra y en la descripción de cada variable podemos ver que todas las variables menos ID son de tipo carácter. Para poder analizar de forma eficaz los datos haremos las siguientes conversiones:

- Variables category, main_category, currency y country van a convertirse a tipo factor para poder agrupar proyectos.

```
dataSet$category <- as.factor(dataSet$category)
dataSet$main_category <- as.factor(dataSet$main_category)
dataSet$currency <- as.factor(dataSet$currency)
dataSet$country <- as.factor(dataSet$country)

# Valores que toman las variables currency y country
unique(dataSet$currency)
```

```
##  [1] USD EUR GBP CAD AUD MXN SEK NZD SGD DKK NOK HKD
## Levels: AUD CAD DKK EUR GBP HKD MXN NOK NZD SEK SGD USD
```

```
unique(dataSet$country)
```

```
##  [1] US   BE   IT   GB   FR   CA   AU   DE   MX   NL   N,0" IE   AT   SE   NZ
## [16] ES   SG   DK   NO   HK   LU
## Levels: AT AU BE CA DE DK ES FR GB HK IE IT LU MX N,0" NL NO NZ SE SG US
```

Vemos que country tiene un carácter especial en algunos de los casos, vamos a sustituirlos por NA y más adelante imputaremos estos valores basándonos en la variable currency.

```
dataSet$country[dataSet$country == 'N,0"'] <- NA
```

- Las variables deadline y launched se convertirán a tipo Date.

```
dataSet$deadline <- as.Date(dataSet$deadline, '%Y-%m-%d')
dataSet$launched <- as.Date(dataSet$launched, '%Y-%m-%d')
```

- goal, pledged y usd.pledged van a pasar a ser tipo numérico.

```
dataSet$goal <- as.numeric(dataSet$goal)
dataSet$pledged <- as.numeric(dataSet$pledged)
dataSet$usd_pledged <- as.numeric(dataSet$usd_pledged)
dataSet$usd_pledged_real <- as.numeric(dataSet$usd_pledged_real)
dataSet$usd_goal_real <- as.numeric(dataSet$usd_goal_real)
```

- La variable state, como ya se ha explicado, detalla el estado en el que acabó o estaba en ese momento la campaña. Vemos que hay 5 estados failed, successful, canceled, suspended y undefined. Ya que undefined no está detallado que significa, se ha decidido añadir una nueva columna status, que contendrá dos valores, 0 si el proyecto no ha sido fundado y 1 si el proyecto ha recaudado los fondos suficientes.

```
dataSet['status'] <- as.factor(ifelse(dataSet$pledged > dataSet$goal, 1, 0))
```

- Se va a añadir una columna nueva euros_pledged que contendrá la conversión de usd_pledged_real a euros. Se utilizará la conversión 1€ = 1.23$ a 19 de diciembre.

```
dataSet['euros_pledged'] <- as.numeric(
  format(as.numeric(dataSet$usd_pledged_real)/1.23), nsmall = 1)
```

- Se va a añadir una columna nueva proyect_length de tipo numérico, que contendrá el total de días que el proyecto ha estado abierto a financiación. Esta nueva columna será resultado de la diferencia entre la columna deadline y launched.

```
dataSet['proyect_length'] <- as.numeric(dataSet$deadline - dataSet$launched)
```

```
# Muestra set de datos
head(dataSet, 5)
```

```
##                ID                                                    name
## 105958  153850277                                       From Case to Coast
## 295519  574159336                  Little Readers Interactive eBook Publishing
## 22237  1112688723      The Beauty For Ashes Project Presents: You're Beautiful
## 142075 1721941336                            Rockwell: The Fluctuating Market
## 238877  284653041 Walkpad. Listen your music through the DualShock4 (Canceled)
##               category main_category currency   deadline goal   launched pledged
## 105958           Music         Music      USD 2015-01-18 1000 2014-12-29     360
## 295519            Apps    Technology      USD 2014-06-16  299 2014-05-22     314
## 22237    Performances         Dance      USD 2015-03-22 2500 2015-02-20    2700
## 142075 Tabletop Games         Games      EUR 2015-12-21 1500 2015-12-01    6890
## 238877         Gadgets    Technology      EUR 2015-11-27 5000 2015-10-28       0
##             state backers country usd.pledged usd_pledged_real usd_goal_real
## 105958     failed      11      US      360.00            360.0       1000.00
## 295519 successful       4      US      314.00            314.0        299.00
## 22237  successful      33      US     2700.00           2700.0       2500.00
## 142075 successful     234      BE     7294.41           7542.5       1642.05
## 238877   canceled       0      IT        0.00              0.0       5290.00
##        usd_pledged status euros_pledged proyect_length
```

```
## 105958       360.0      0      292.6829             20
## 295519       314.0      1      255.2846             25
## 22237       2700.0      1     2195.1220             30
## 142075      7542.5      1     6132.1140             20
## 238877         0.0      0        0.0000             30
```

```
summary(dataSet)
```

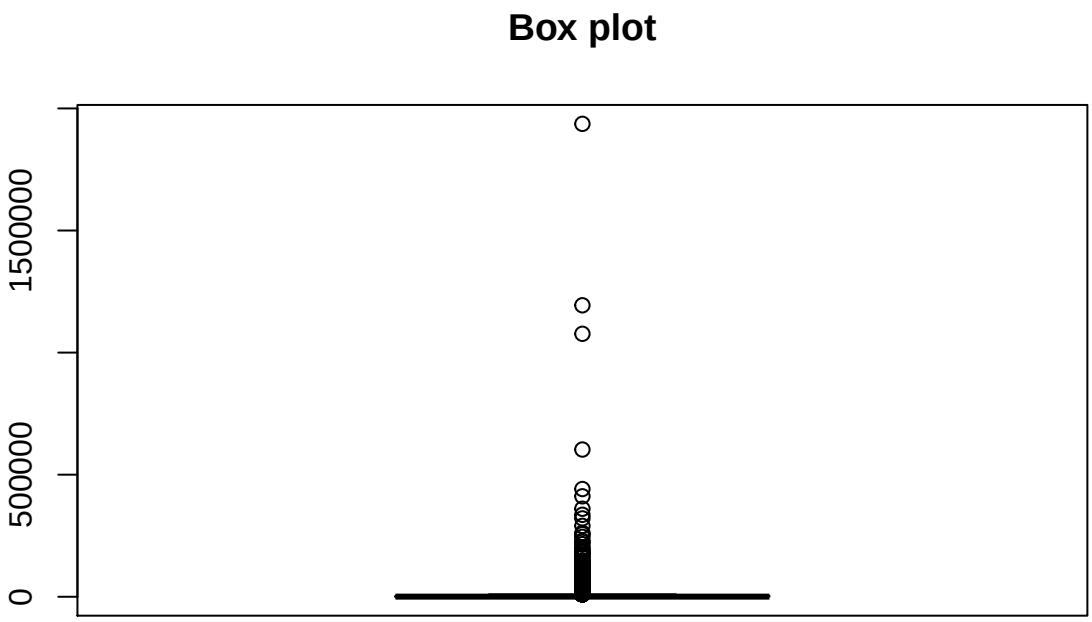
```
##        ID                 name                    category
##  Min.   :2.237e+05   Length:3000        Product Design: 191
##  1st Qu.:5.436e+08   Class :character   Documentary   : 132
##  Median :1.061e+09   Mode  :character   Music         : 122
##  Mean   :1.072e+09                      Tabletop Games: 117
##  3rd Qu.:1.601e+09                      Video Games   :  98
##  Max.   :2.147e+09                      Food          :  90
##                                         (Other)       :2250
##       main_category    currency       deadline                goal
##  Film & Video:488   USD    :2324   Min.   :2009-06-30   Min.   :      1
##  Music       :385   GBP    : 277   1st Qu.:2013-06-23   1st Qu.:   2000
##  Publishing  :336   EUR    : 146   Median :2015-02-11   Median :   5476
##  Games       :292   CAD    : 123   Mean   :2014-11-13   Mean   :  28458
##  Technology  :259   AUD    :  72   3rd Qu.:2016-05-17   3rd Qu.:  15716
##  Design      :251   MXN    :  12   Max.   :2018-02-15   Max.   :9000000
##  (Other)     :989   (Other):  46
##     launched             pledged             state              backers
##  Min.   :2009-04-29   Min.   :      0.0   Length:3000        Min.   :    0
##  1st Qu.:2013-05-23   1st Qu.:     30.0   Class :character   1st Qu.:    2
##  Median :2015-01-10   Median :    539.5   Mode  :character   Median :   11
##  Mean   :2014-10-10   Mean   :   8623.1                      Mean   :  111
##  3rd Qu.:2016-04-13   3rd Qu.:   4032.0                      3rd Qu.:   55
##  Max.   :2018-01-01   Max.   :1936825.0                      Max.   :35549
##
##     country      usd.pledged        usd_pledged_real    usd_goal_real
##  US     :2304   Min.   :      0.0   Min.   :      0.0   Min.   :      1
##  GB     : 272   1st Qu.:     16.7   1st Qu.:     30.0   1st Qu.:   2000
##  CA     : 122   Median :    332.6   Median :    538.4   Median :   5431
##  AU     :  72   Mean   :   6544.3   Mean   :   8380.4   Mean   :  26521
##  DE     :  32   3rd Qu.:   2939.0   3rd Qu.:   3891.8   3rd Qu.:  15289
##  (Other): 169   Max.   :1936825.0   Max.   :1936825.0   Max.   :7112938
##  NA's   :  29   NA's   :29
##   usd_pledged        status   euros_pledged       proyect_length
##  Min.   :      0.0   0:1937   Min.   :      0.0   Min.   : 1.00
##  1st Qu.:     30.0   1:1063   1st Qu.:     24.4   1st Qu.:30.00
##  Median :    538.4            Median :    437.7   Median :30.00
##  Mean   :   8380.4            Mean   :   6813.3   Mean   :34.33
##  3rd Qu.:   3891.8            3rd Qu.:   3164.0   3rd Qu.:39.00
##  Max.   :1936825.0            Max.   :1574654.0   Max.   :91.00
##
```

3.2 Valores atípicos

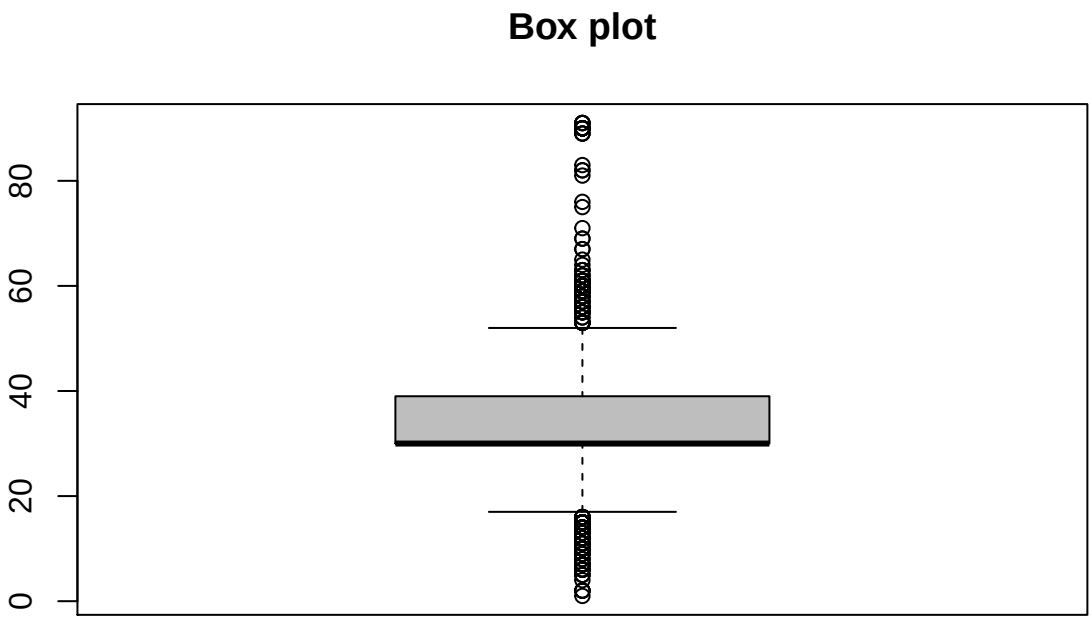
Volviendo a la estadística descriptiva vemos que la diferencia entre la media y el máximo y mínimo valor de muestras de la variable pledged y proyect_length es bastante significativa, lo que puede indicar la presencia de outliers. Vamos a comprobar si tenemos outliers mediante diagrama de cajas.

```r
# Importamos la librería ggplot2
library(ggplot2)

# Diagrama de cajas para la variable pledged y proyect_length
boxplot(dataSet$pledged, main="Box plot", col="gray")
```




```r
boxplot(dataSet$proyect_length, main="Box plot", col="gray")
```



Vemos que ambas variables tienen valores extremos por lo que vamos a analizar para determinar cómo proceder con ellos.

```r
tail(sort(dataSet$pledged), 10)
```

```
##  [1]  290429.0  320890.0  335650.0  360499.3  411363.4  441199.0  602960.5
##  [8] 1076940.1 1193776.5 1936825.0
```

```
tail(sort(dataSet$proyect_length), 10)
```

```
##  [1] 90 90 90 90 90 90 91 91 91 91
```

Vemos que hay bastante diferencia entre la media de la variable pledged y los valores más altos, pero haciendo un poco de investigación online se ha encontrado que algún proyecto ha llegado a recaudar más de 20,000,000$, [**enlace**](https://www.marketwatch.com/story/10-kickstarter-products-that-raised-the-most-money-2017-06-22-10883052). Por este motivo se ha decidido aceptar dichos outliers y tratarlos como datos válidos.

Por otro lado, para la variable proyect_length vemos que el valor máximo de esta son 16,739 días, lo que corresponde a más de 3 años, implicaría que un proyecto ha estado recaudando fondos durante todo ese tiempo. Haciendo un poco de investigación sobre la normativa de Kickstarter, se ha encontrado [**enlace**](https://help.kickstarter.com/hc/en-us/articles/115005128434-What-is-the-maximum-project-duration-#:~:text=Projects%20on%20Kickstarter%20can%20last,at%2030%20days%20or%20less.) que, hoy en día, la duración máxima por proyecto es de 60 días. También en este otro artículo de explica que hasta el año 2011 la duración máxima era de 90 días [**enlace**](https://www.kickstarter.com/blog/shortening-the-maximum-project-length).

```
outliersDays <- tail(sort(dataSet$proyect_length), 7)
outliersDays
```

```
## [1] 90 90 90 91 91 91 91
```

```
indexes <- which(dataSet$proyect_length %in% outliersDays)
indexes
```

```
##  [1]   58  221  565  634  692  930  945  973  976 1322 1355 1395 1424 1444 1542
## [16] 1572 2080 2394
```

```
dataSet$launched[indexes]
```

```
##  [1] "2009-09-15" "2010-10-27" "2011-01-11" "2009-11-01" "2010-05-13"
##  [6] "2011-01-08" "2011-05-09" "2011-05-25" "2010-02-23" "2011-05-31"
## [11] "2011-01-25" "2010-11-16" "2011-05-22" "2009-09-14" "2009-07-27"
## [16] "2011-06-07" "2010-01-07" "2010-12-05"
```

Al comprobar el valor launchdate para cada una de las muestras en las que se encontraban los outliers vemos que la fecha es 1970-01-01 01:00:00, se han recogido mal al guardar los datos, de ahí que salgan valores tan extremos para proyect_length.

Por todo ello se ha decidido que cualquier duración significativamente mayor de 90 días se va a tratar como outlier y se reemplazará por NA para posteriormente imputarlo con un valor de 90.

```
# Se reemplazan los valores en las posiciones indexes por NA
dataSet$proyect_length[indexes] <- NA
```

Muestra de todos las variables y sus valores NA's.

```
# Comprobamos si quedan NA's
sapply(dataSet, function(x) sum(is.na(x)))
```

```
##               ID             name         category    main_category
##                0                0                0                0
##         currency         deadline             goal         launched
##                0                0                0                0
##          pledged            state          backers          country
##                0                0                0               29
##      usd.pledged usd_pledged_real    usd_goal_real      usd_pledged
##               29                0                0                0
##           status    euros_pledged   proyect_length
```

```
##                0                0               18
```

3.3 Imputación de valores

- Cómo se ha mencionado anteriormente los valores NA de la variable proyect_length se van a reemplazar por 90 ya que es el máximo número de días que un proyecto puede estar recaudando dinero.

```
dataSet$proyect_length[indexes] <- 90
```

- Imputación de valores perdidos para la variable country.

```
idx <- which(is.na(dataSet$country))
# encontrar las combinaciones únicas de country y currency pero no cuando country
# es NA
uniques <- unique(dataSet[c('country', 'currency')])
uniques <- uniques[!is.na(uniques$country),]

# reemplazar los NA's de country con los valores únicos asociados con currency
na.country <- which(is.na(dataSet$country))
na.currency <- dataSet$currency[na.country]
dataSet$country[idx] <- uniques$country[match(na.currency, uniques$currency)]
```

Por último comprobamos si quedan NA's en los datos.

```
# Comprobamos si quedan NAs
sapply(dataSet, function(x) sum(is.na(x)))
```

```
##               ID             name         category    main_category
##                0                0                0                0
##         currency         deadline             goal         launched
##                0                0                0                0
##          pledged            state          backers          country
##                0                0                0                0
##      usd.pledged usd_pledged_real    usd_goal_real      usd_pledged
##               29                0                0                0
##           status    euros_pledged   proyect_length
##                0                0                0
```

3.4 Selección de datos

A continuación, vamos a detallar que atributos hemos descartado y cuales hemos decidido sean imprescindibles para el análisis:

- Se ha decidido borrar del dataset la columna de usd_pledged, esta representa la conversión a dolares por parte de Kickstarter del atributo pledged, pero se han descubierto bastantes inconsistencias. El creador del dataset, por este mismo motivo, decidió incluir un nuevo atributo con una conversión más precisa de pledged, que es la que vamos a usar.
- También se ha decidido descartar la variable usd_goal_pledged porque no resulta significativa para el estudio.
- Por otro lado, vamos a prescindir de la variable state. Como se ha mencionado anteriormente, un Kickstarter es satisfactorio si el proyecto consigue recaudar el dinero marcado como objetivo en el tiempo estimado, por lo que no es necesario para nuestro estudio si dicho proyecto se ha cancelado, o se ha suspendido o sigue activo. Se puede dar el caso por ejemplo que un proyecto llegue al objetivo económico marcado dentro de tiempo, pero el organizador, por cualquier motivo decida suspenderlo. En ese caso el proyecto aparecerá como cancelado, pero desde el punto de vista del objetivo del proyecto, la recaudación ha sido satisfactoria.

```r
# Quitar columnas usd_pledged y usd_goal_pledged del daataset
drops <- c('usd.pledged', 'usd_goal_pledged', 'state')
dataSet <- dataSet[ , !(names(dataSet) %in% drops)]
# Anális descriptivo del dataset limpio
summary(dataSet)
```

```
##        ID                 name                    category
##  Min.   :2.237e+05   Length:3000        Product Design: 191
##  1st Qu.:5.436e+08   Class :character   Documentary   : 132
##  Median :1.061e+09   Mode  :character   Music         : 122
##  Mean   :1.072e+09                      Tabletop Games: 117
##  3rd Qu.:1.601e+09                      Video Games   :  98
##  Max.   :2.147e+09                      Food          :  90
##                                         (Other)       :2250
##       main_category    currency       deadline                goal
##  Film & Video:488   USD    :2324   Min.   :2009-06-30   Min.   :      1
##  Music       :385   GBP    : 277   1st Qu.:2013-06-23   1st Qu.:   2000
##  Publishing  :336   EUR    : 146   Median :2015-02-11   Median :   5476
##  Games       :292   CAD    : 123   Mean   :2014-11-13   Mean   :  28458
##  Technology  :259   AUD    :  72   3rd Qu.:2016-05-17   3rd Qu.:  15716
##  Design      :251   MXN    :  12   Max.   :2018-02-15   Max.   :9000000
##  (Other)     :989   (Other):  46
##     launched             pledged             backers          country
##  Min.   :2009-04-29   Min.   :      0.0   Min.   :    0   US     :2324
##  1st Qu.:2013-05-23   1st Qu.:     30.0   1st Qu.:    2   GB     : 277
##  Median :2015-01-10   Median :    539.5   Median :   11   CA     : 123
##  Mean   :2014-10-10   Mean   :   8623.1   Mean   :  111   AU     :  72
##  3rd Qu.:2016-04-13   3rd Qu.:   4032.0   3rd Qu.:   55   DE     :  32
##  Max.   :2018-01-01   Max.   :1936825.0   Max.   :35549   FR     :  31
##                                                           (Other): 141
##  usd_pledged_real    usd_goal_real      usd_pledged        status
##  Min.   :      0.0   Min.   :      1   Min.   :      0.0   0:1937
##  1st Qu.:     30.0   1st Qu.:   2000   1st Qu.:     30.0   1:1063
##  Median :    538.4   Median :   5431   Median :    538.4
##  Mean   :   8380.4   Mean   :  26521   Mean   :   8380.4
##  3rd Qu.:   3891.8   3rd Qu.:  15289   3rd Qu.:   3891.8
##  Max.   :1936825.0   Max.   :7112938   Max.   :1936825.0
##
##  euros_pledged       proyect_length
##  Min.   :      0.0   Min.   : 1.00
##  1st Qu.:     24.4   1st Qu.:30.00
##  Median :    437.7   Median :30.00
##  Mean   :   6813.3   Mean   :34.33
##  3rd Qu.:   3164.0   3rd Qu.:39.00
##  Max.   :1574654.0   Max.   :90.00
##
```

3.5 Exportación de los datos limpios

Una vez el procesado de los datos ha finalizado, se genera un archivo csv con nombre "ks-projects-201801_clean.csv", que contendrá el dataset con 3000 muestras limpias.

```r
# Exportación de los datos limpios en .csv
write.csv(dataSet, '../data/ks-projects-201801_sample_clean.csv')
```

4 Análisis de los datos

4.1 Selección de los grupos de datos a analizar

De todo el conjunto de datos, se han seleccionado los siguientes atributos para poder ser analizados creyendo que son estos los que aportarán más valor al análisis posterior:

- main_category: recoge las 15 principales categorías presentes.

```
unique(dataSet$main_category)
```

```
##  [1] Music        Technology   Dance        Games        Art
##  [6] Photography  Design       Journalism   Food         Film & Video
## [11] Crafts       Publishing   Comics       Fashion      Theater
## 15 Levels: Art Comics Crafts Dance Design Fashion Film & Video Food ... Theater
```

- proyect_length: tiempo de duración de cada proyecto, expresado en días.
- euros_pledged: conversión a € de la variable usd_pledged_real.
- country: país en el que se publicó el proyecto Kickstarter. La variable currency representa la moneda de dicho país por lo que nos resulta redundante.
- backers: cantidad total de mecenas del proyecto.
- status: estado del proyecto, 0 no ha conseguido el objetivo, 1 si lo ha conseguido.

4.2 Normalidad y homocedasticidad

A la hora de identificar los métodos de análisis más adecuados se debe conocer antes las características de los datos, por ejemplo, si estos siguen una distribución normal o si presentan homocedasticidad. Por ello vamos a comprobar que las variables numéricas elegidas siguen una distribución normal o presentan homogeneidad de la varianza.

- Test de normalidad

Se va a utilizar el test Shapiro-Wilk, asumiendo un intervalo de confianza del 95%. Esto quiere decir que si el p-valor es menor o igual que el nivel de significancia con un valor de 0.05, entonces podemos rechazar la presunción de normalidad, es decir, la variable no sigue una distribución normal.

```
shapiro.test(dataSet$proyect_length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataSet$proyect_length
## W = 0.84502, p-value < 2.2e-16
```

```
shapiro.test(dataSet$euros_pledged)
```

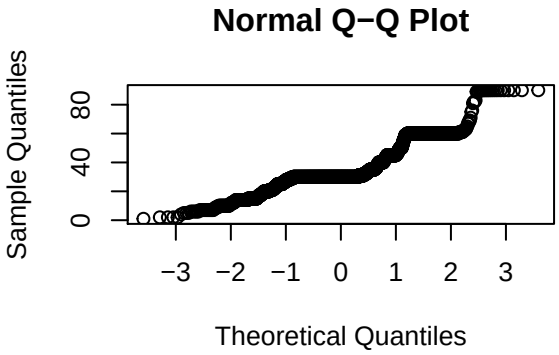
```
##
##  Shapiro-Wilk normality test
##
## data:  dataSet$euros_pledged
## W = 0.11157, p-value < 2.2e-16
```

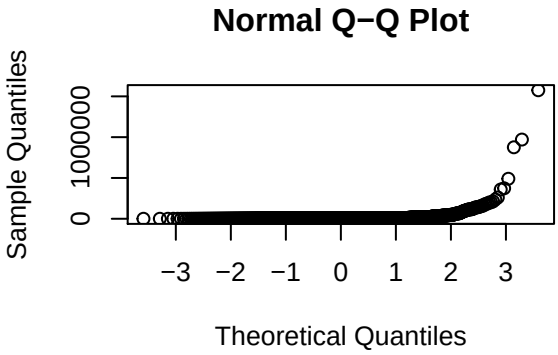
```
shapiro.test(dataSet$backers)
```

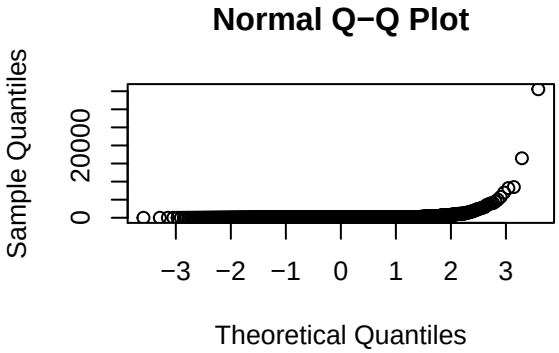
```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  dataSet$backers
## W = 0.084774, p-value < 2.2e-16
```

```
# Representación de la distribución
par(mfrow=c(2,2))
qqnorm(dataSet$proyect_length)
qqnorm(dataSet$euros_pledged)
qqnorm(dataSet$backers)
```







Se puede apreciar que los datos no siguen una distribución normal ya que en el total de las comprobaciones el p-valor del Test de Shapiro-Wilk el p-value < 2.2e-16 rechazando dicha distribución normal.

- Test de homocedasticidad

Ya que hemos comprobado que nuestros datos no siguen una distribución normal (p-value < 2.2e-16 en todos los casos), para el test de homocedasticidad tendremos utilizaremos el de Fligner-Killeen. La hipótesis nula asume la igualdad de varianzas, por lo que p-values inferiores al nivel de significancia (0.05), indicarán heterocedasticidad.

Para ello comprobaremos distintos grupos de datos entre sí:

```
fligner.test(dataSet$proyect_length ~ dataSet$euros_pledged, data = dataSet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  dataSet$proyect_length by dataSet$euros_pledged
## Fligner-Killeen:med chi-squared = 1450.7, df = 1935, p-value = 1
```

```
fligner.test(dataSet$euros_pledged ~ dataSet$backers, data = dataSet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
```

```
##
## data:  dataSet$euros_pledged by dataSet$backers
## Fligner-Killeen:med chi-squared = 1962.2, df = 404, p-value < 2.2e-16
```

```
fligner.test(dataSet$proyect_length ~ dataSet$backers, data = dataSet)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  dataSet$proyect_length by dataSet$backers
## Fligner-Killeen:med chi-squared = 287.48, df = 404, p-value = 1
```

De este análisis podemos observar dos casos, para las variables proyect_length-euros_pledged y proyect_length-backers el test de Fligner-Killen da un p-value mayor que 0.05 (1 y 0.942 respectivamente), por lo que se asume homocedasticidad.

Por otro lado, la prueba para euros_pledged y backers se resuelve con un p-value $<$ 2.2e-16, por lo que en este caso si se puede rechazar la hipótesis nula de homocedasticidad y se concluye que la variable euros_pledged presenta varianzas estadísticamente diferentes para los diferentes grupos de backers.

4.3 Pruebas estadísticas

Ya hemos comprobado anteriormente que las variables no siguen la distribución normal por lo para el contraste de hipótesis deberemos aplicar pruebas no paramétricas como Wilcoxon o Mann-Whitney.

TBC

4.3.1 Contraste de hipótesis de dos muestras

TBC

4.3.2 Modelo de regresión lineal muúltiple

TBC

4.3.3 Modelo de regresión logística

TBC

5 Representación de los resultados

TBC

6 Conclusiones

TBC

7 Agradecimientos

En primer lugar, agradecer y reconocer el trabajo de Mickaël Mouillé **[enalce]**(https://www.kaggle.com/kemical), creador del dataset por trabajo para recolectar datos durante tantos años y publicarlos para el uso público.

También agradecer a todas aquellas personas que han publicado sus dudas sobre el dataset para beneficio de todos.

8 Tabla de contribuciones

Contribuciones	Firma
Investigación previa	M.G.
Redacción de las respuestas	M.G.
Desarrollo código	M.G.