Tipologia i cicle de vida de les dades - Pràctica 2. Kickstarter

Ivan Borrego Garcia / Cristina Sanchis Puerto

GENER 2020

Índex

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç Github (*https://github.com*) on es trobin les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen a la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Malgrat que no es tracta del mateix enunciat, els següents exemples d'edicions anteriors us poden servir com a guia:

- Exemple: *https://github.com/Bengis/nba-gap-cleaning*
- Exemple complex (fitxer adjunt).

Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

● Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.

● Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

Objectius

Els objectius concrets d'aquesta pràctica són:

● Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.

● Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.

● Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.

● Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.

● Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.

● Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.

● Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (*https://www.kaggle.com*). Alguns exemples de dataset amb els que podeu treballar són:

● Red Wine Quality (*https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009*).

● Titanic: Machine Learning from Disaster (*https://www.kaggle.com/c/titanic*).

L'últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició. Seguint

les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

Pràctica:

DESCRIPCIÓ DEL DATASET

Perquè és important i quina pregunta/problema pretén respondre?

Kickstarter és una plataforma de micromecenatge (crowdfunding) nord-americana. En ella és possible finançar projectes creatius de tot tipus, des de cinema independent fins productes tecnològics, passant per còmics, jocs o cuina.

Tot projecte especificarà des de el començament les dates d'inici i final de l'activitat a la plataforma, que seran les que el creador consideri oportunes, tot i que el rang normalment es troba entre unes poques setmanes i diversos mesos. El projecte també haurà d'indicar quina és la recaptació mínima de diners per considerar el mecenatge com a exitós i per tant iniciar el projecte.

Durant aquest temps qualsevol usuari pot participar en el mecenatge, fet que normalment es realitza escollint un dels diferents nivells d'aportació establerts pel creador del projecte, que solen incloure una còpia o participació del producte final que es vol crear. Cal tenir en compte que les aportacions es fan efectives un cop hagi acabat el temps d'activitat fixat i només si s'ha assolit l'objectiu monetari inicial. Si no és el cas, el projecte es considera fracassat i no es realitza cap pagament.

És per tant, important per a un nou creador revisar la trajectòria de projectes anteriors semblants, per estudiar les característiques tant dels projectes exitosos com dels que no.

* **ID** : Identificador únic del projecte.
* **name** : Nom del projecte.
* **category** : Categoria específica del projecte.
* **main_category** : Categoria general del projecte.
* **currency** : Moneda en la que gestiona la recaptació.
* **deadline** : Data del final de la campanya de mecenatge.
* **goal** : Quantitat mínima de diners aportats per considerar el projecte exitós.
* **launched** : Data de llançament de la campanya de mecenatge.
* **pledged** : Quantitat aportada al final de la campanya.
* **state** : Estat en que es troba el projecte.
* **backers** : Nombre de persones mecenes.

* **country** : País d'origen del projecte.
* **usd.pledged** : Conversió a dòlars americans de la quantitat recaptada (feta per Kickstarter)
* **usd_pledged_real** : Conversió a dòlars americans de la quantitat recaptada (feta per plataforma independent fixer.io)
* **usd_goal_real** : Conversió a dòlars americans de la quantitat requerida (feta per plataforma independent fixer.io)

INTEGRACIÓ I SELECCIÓ DE LES DADES D'INTERÈS A ANALITZAR

```
# Paquets
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(lubridate))
suppressPackageStartupMessages(library(ggpubr))
# Llegim les dades
data<- read.csv("ks-projects-201801.csv", header=T, sep=",")
# Verifiquem l'estructura del joc de dades
str(data)
```

```
## 'data.frame':    378661 obs. of  15 variables:
##  $ ID               : int  1000002330 1000003930 1000004038 10000075
40 1000011046 1000014025 1000023410 1000030581 1000034518 100004195 ..
.
##  $ name             : Factor w/ 375765 levels "","\177Not Twins - Ne
w EP! \"The View from Down Here\"",..: 332541 135689 365010 344805 773
49 206130 293462 69360 284139 290718 ...
##  $ category         : Factor w/ 159 levels "3D Printing",..: 109 94
94 91 56 124 59 42 114 40 ...
##  $ main_category    : Factor w/ 15 levels "Art","Comics",..: 13 7 7
11 7 8 8 8 5 7 ...
##  $ currency         : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14
14 14 14 14 14 14 14 14 ...
##  $ deadline         : Factor w/ 3164 levels "2009-05-03","2009-05-16
",..: 2288 3042 1333 1017 2247 2463 1996 2448 1790 1863 ...
##  $ goal             : num  1000 30000 45000 5000 19500 50000 1000 25
000 125000 65000 ...
##  $ launched         : Factor w/ 378089 levels "1970-01-01 01:00:00",
..: 243292 361975 80409 46557 235943 278600 187500 274014 139367 15376
6 ...
##  $ pledged          : num  0 2421 220 1 1283 ...
##  $ state            : Factor w/ 6 levels "canceled","failed",..: 2 2
2 2 1 4 4 2 1 1 ...
##  $ backers          : int  0 15 3 1 14 224 16 40 58 43 ...
```

```
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 2
3 23 23 23 23 23 23 23 ...
##  $ usd.pledged     : num  0 100 220 1 1283 ...
##  $ usd_pledged_real: num  0 2421 220 1 1283 ...
##  $ usd_goal_real   : num  1534 30000 45000 5000 19500 ...

# Consultem les primeres files del conjunt de dades
head(data)

##           ID
name
## 1 1000002330                            The Songs of Adelaide & Abu
llah
## 2 1000003930              Greeting From Earth: ZGAC Arts Capsule Fo
r ET
## 3 1000004038                                          Where is H
ank?
## 4 1000007540          ToshiCapital Rekordz Needs Help to Complete A
lbum
## 5 1000011046 Community Film Project: The Art of Neighborhood Filmma
king
## 6 1000014025                                     Monarch Espresso
Bar
##         category main_category currency   deadline  goal
## 1         Poetry    Publishing      GBP 2015-10-09  1000
## 2 Narrative Film  Film & Video      USD 2017-11-01 30000
## 3 Narrative Film  Film & Video      USD 2013-02-26 45000
## 4          Music         Music      USD 2012-04-16  5000
## 5   Film & Video  Film & Video      USD 2015-08-29 19500
## 6    Restaurants          Food      USD 2016-04-01 50000
##              launched pledged      state backers country usd.pledge
d
## 1 2015-08-11 12:12:28       0     failed       0      GB
0
## 2 2017-09-02 04:43:57    2421     failed      15      US         10
0
## 3 2013-01-12 00:20:50     220     failed       3      US         22
0
## 4 2012-03-17 03:24:11       1     failed       1      US
1
## 5 2015-07-04 08:35:03    1283   canceled      14      US        128
3
## 6 2016-02-26 13:38:27   52375 successful     224      US       5237
5
##   usd_pledged_real usd_goal_real
## 1                0       1533.95
```

```
## 2              2421      30000.00
## 3               220      45000.00
## 4                 1       5000.00
## 5              1283      19500.00
## 6             52375      50000.00
```

Observem que tenim, 378661 observacions i 15 atributs.

Verifiquem que no hi hagi projectes duplicats. La comprobació la fem a partir de la variable ID que és l'identificador únic del projecte.

```
# Projectes duplicats?
length(unique(data$ID))
```

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

No hi ha registres duplicats, ja que hi ha 378661 valors diferents de la variable ID, que és el nombre total d'observacions que conté el conjunt de dades.

```
# Estadístiques de valors buits
colSums(is.na(data))
```

```
##               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
##             3797                0                0
```

```
colSums(data=="")
```

```
##               ID             name         category    main_category
##                0                4                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
##               NA                0                0
```

Una vegada, hem consultat si hi ha projectes duplicats, es pot prescindir de la columna ID, que com hem comentat abans, és l'identificador únic del projecte. Per a l'anàlisi de les dades que volem realitzar, l'atribut identificador no ens aporta valor.

```
# Eliminem atribut ID
data$ID <- NULL
```

A continuació formategem els atributs que fan referència a la data: launched and deadline. Pel que fa a la variable launched, les hores, minuts i segons, no ens interessa. El format per

ambdues variables serà YYYY-MM-DD. Posteriorment seleccionarem els registres a partir de l'any 2015.

```
# Formategem launched i deadline
data$launched <- as.Date(substr(as.character(data$launched), 1, 10), "
%Y-%m-%d")
data$deadline <- as.Date(as.character(data$deadline), "%Y-%m-%d")
# Filtrem dades: Projecte del 2015 en endavant
data <- data[data$launched >= "2015-01-01",]
# Registres després del filtratge
dim(data)
```

```
## [1] 186808     14
```

El nombre de registres ha passat de 378661 observacions, a 186808 registres. Seleccionant els registres a partir de l'any 2015, hem realitzat una reducció de la quantitat del conjunt de dades.

Considerem que és interessant, conèixer la durada dels projectes, per tant, creem una nova variable "duration_days", que serà la duració en dies del projecte, des de la data de llançament de la campanya (launched) fins a la data final de la campanya (deadline).

Igualment serà interessant disposar dels valors mitjans de les aportacions i del percentatge total assolit.

També s'inclou un nou atribut, perquè volem distribuir els països per regió.

```
# Creem la variable duration_days
data$duration_days <- as.integer(data$deadline-data$launched)
# Creem la variable mean_pledged
data$mean_pledged <- ifelse(data$backers==0, data$mean_pledged<-0, dat
a$mean_pledged<-data$usd_pledged_real/data$backers)
# Creem la variable percent_pledged
data$percent_pledged <- (data$usd_pledged_real/data$usd_goal_real)*100
# Creem la variable region
data$region <- ifelse(data$country=='N,0"', data$region<-NA,
                      ifelse(data$country=="US"|data$country=="CA", da
ta$region<-"North America",
                             ifelse(data$country=="AU"|data$country=="
NZ"|data$country=="HK"|
                                      data$country=="SG"|data$country=
="JP", data$region<- "Asia & Pacific",
                                    ifelse(data$country=="MX", data$re
gion<- "Latin America",
                                           data$region<-"Europe")
                                    )
                             )
                      )
data$region <- as.factor(data$region)
# Consultem les primeres files del conjunt de dades
head(data)
```

```
##                                                                name
## 1                                The Songs of Adelaide & Abullah
## 2                  Greeting From Earth: ZGAC Arts Capsule For ET
## 5   Community Film Project: The Art of Neighborhood Filmmaking
## 6                                             Monarch Espresso Bar
## 8              Chaser Strips. Our Strips make Shots their B*tch!
## 14 G-Spot Place for Gamers to connect with eachother & go pro!
##          category main_category currency   deadline   goal   launch
ed
## 1          Poetry    Publishing      GBP 2015-10-09   1000 2015-08-
11
## 2  Narrative Film  Film & Video      USD 2017-11-01  30000 2017-09-
02
## 5    Film & Video  Film & Video      USD 2015-08-29  19500 2015-07-
04
## 6     Restaurants          Food      USD 2016-04-01  50000 2016-02-
26
## 8          Drinks          Food      USD 2016-03-17  25000 2016-02-
01
## 14          Games         Games      USD 2016-03-25 200000 2016-02-
09
##    pledged      state backers country usd.pledged usd_pledged_real
## 1        0     failed       0      GB           0                0
## 2     2421     failed      15      US         100             2421
## 5     1283   canceled      14      US        1283             1283
## 6    52375 successful     224      US       52375            52375
## 8      453     failed      40      US         453              453
## 14       0     failed       0      US           0                0
##    usd_goal_real duration_days mean_pledged percent_pledged
region
## 1        1533.95            59      0.00000        0.000000
Europe
## 2       30000.00            60    161.40000        8.070000 North A
merica
## 5       19500.00            56     91.64286        6.579487 North A
merica
## 6       50000.00            35    233.81696      104.750000 North A
merica
## 8       25000.00            45     11.32500        1.812000 North A
merica
## 14     200000.00            45      0.00000        0.000000 North A
merica
```

Podem prescindir de les variables, goal, pledge. La variable goal correspon a la quantitat mínima de diners aportats per considerar el projecte exitós, i la variable pledge és la quantitat aportada al final de la campanya. Existeixen tres variables més que fan referència a l'import, on

s'ha realitzat una conversió a dòlars americans (USD), per tant per a realitzar l'anàlisi dels projectes, s'utilitzaran les conversions en dòlars, que correspon a les variables usd_goal_real, usd_plegded_real i usd.pledged. No considerem eliminar l'atribut currency, perquè pensem que és interessant, conèixer en quantes divises diferents s'està treballant, i saber si la divisa del projecte és un factor per a garantir el seu èxit o no.

```
# Reducció de la dimensionalitat, eliminem variables...
data$goal        <- NULL
data$pledged     <- NULL
```

Si s'observa el resum de l'estructura de les dades, la variable name està definida com una variable de tipus factor. Canviem el seu tipus, ja que el nom d'un projecte, no es deu considerar de tipus factor, sinó més bé de tipus string.

També es consulta la quantitat de categories pel que fa a la variable category i main_category

```
# La variable name és un string
data$name <- as.character(data$name)
# category
length(levels(data$category))
```

```
## [1] 159
```

```
# main_category
length(levels(data$main_category))
```

```
## [1] 15
```

```
levels(data$main_category)
```

```
##  [1] "Art"          "Comics"       "Crafts"       "Dance"
##  [5] "Design"       "Fashion"      "Film & Video" "Food"
##  [9] "Games"        "Journalism"   "Music"        "Photography"
## [13] "Publishing"   "Technology"   "Theater"
```

Existeixen 159 categories per als projectes, distribuides en 15 categories principals. Per tant, es treballarà amb la categoria principal del projecte. S'elimina l'atribut category.

```
# Reducció de la dimensionalitat, eliminem variable category
data$category <- NULL
# Resum final de l'estructura de les dades
str(data)
```

```
## 'data.frame':    186808 obs. of  15 variables:
##  $ name            : chr  "The Songs of Adelaide & Abullah" "Greeti
ng From Earth: ZGAC Arts Capsule For ET" "Community Film Project: The
Art of Neighborhood Filmmaking" "Monarch Espresso Bar" ...
##  $ main_category   : Factor w/ 15 levels "Art","Comics",..: 13 7 7
8 8 9 9 5 13 6 ...
##  $ currency        : Factor w/ 14 levels "AUD","CAD","CHF",..: 6 14
14 14 14 14 6 14 14 1 ...
```

```
##  $ deadline        : Date, format: "2015-10-09" "2017-11-01" ...
##  $ launched        : Date, format: "2015-08-11" "2017-09-02" ...
##  $ state           : Factor w/ 6 levels "canceled","failed",..: 2 2
1 4 2 2 4 2 2 2 ...
##  $ backers         : int  0 15 14 224 40 0 761 11 20 1 ...
##  $ country         : Factor w/ 23 levels "AT","AU","BE",..: 10 23 2
3 23 23 23 10 23 23 2 ...
##  $ usd.pledged     : num  0 100 1283 52375 453 ...
##  $ usd_pledged_real: num  0 2421 1283 52375 453 ...
##  $ usd_goal_real   : num  1534 30000 19500 50000 25000 ...
##  $ duration_days   : int  59 60 56 35 45 45 28 30 30 30 ...
##  $ mean_pledged    : num  0 161.4 91.6 233.8 11.3 ...
##  $ percent_pledged : num  0 8.07 6.58 104.75 1.81 ...
##  $ region          : Factor w/ 4 levels "Asia & Pacific",..: 2 4 4
4 4 4 2 4 4 1 ...
```

```
# Resum de les dades
summary(data)
```

```
##      name                main_category      currency
##  Length:186808      Technology  :22732   USD    :126695
##  Class :character   Film & Video:22677   GBP    : 20251
##  Mode  :character   Games       :21270   EUR    : 16535
##                     Music       :19229   CAD    :  9834
##                     Design      :18488   AUD    :  5474
##                     Publishing  :18417   MXN    :  1752
##                     (Other)     :63995   (Other):  6267
##     deadline             launched                  state
##  Min.   :2015-01-05   Min.   :2015-01-01   canceled  :21814
##  1st Qu.:2015-08-19   1st Qu.:2015-07-15   failed    :99784
##  Median :2016-05-04   Median :2016-04-01   live      : 2799
##  Mean   :2016-06-04   Mean   :2016-05-01   successful:58199
##  3rd Qu.:2017-03-13   3rd Qu.:2017-02-08   suspended : 1405
##  Max.   :2018-03-03   Max.   :2018-01-02   undefined : 2807
##
##     backers            country        usd.pledged       usd_pledged
_real
##  Min.   :     0.0   US     :124533   Min.   :       0   Min.   :
0
##  1st Qu.:     1.0   GB     : 19889   1st Qu.:       0   1st Qu.:
13
##  Median :     8.0   CA     :  9679   Median :     143   Median :
400
##  Mean   :   114.1   AU     :  5385   Mean   :    6423   Mean   :
10538
##  3rd Qu.:    51.0   DE     :  4171   3rd Qu.:    1626   3rd Qu.:
```

```
3691
##  Max.   :219382.0   N,0"   :  3028   Max.   :20338986   Max.   :203
38986
##                     (Other): 20123   NA's   :3028
##  usd_goal_real       duration_days    mean_pledged
##  Min.   :        0   Min.   : 1.00   Min.   :    0.000
##  1st Qu.:     2000   1st Qu.:30.00   1st Qu.:    5.667
##  Median :     6238   Median :30.00   Median :   35.167
##  Mean   :    60524   Mean   :33.28   Mean   :   64.388
##  3rd Qu.:    20000   3rd Qu.:35.00   3rd Qu.:   73.949
##  Max.   :151395870   Max.   :90.00   Max.   :10000.000
##
##  percent_pledged              region
##  Min.   :       0   Asia & Pacific:  7592
##  1st Qu.:       0   Europe        : 40224
##  Median :       8   Latin America :  1752
##  Mean   :     447   North America :134212
##  3rd Qu.:     105   NA's          :  3028
##  Max.   :10427789
##
```

NETEJA DE LES DADES

Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Els valors 0 no sempre fa referència a un valor perdut, pot ser un valor buit legítim. A continuació, s'analitzen les variables del conjunt per tal de saber si contenen zeros o elements buits, per saber com gestionar-los i considerar si es tracta d'errors o no. Per exemple, sense fer una anàlisi exhaustiva, podríem dir, que en les variables relacionades amb els diners si existís un zero, no hauria de ser un error, ja que poden haver-hi projectes en els quals no hi hagi cap mena d'aportació econòmica, en canvi veure un element buit ens podria generar dubte, ja que podria ser un zero que no hi ha hagut aportació econòmica o bé que no se sap que ha passat.

Consultem les estadístiques de valors buits.

```
# Estadístiques de valors buits per atributs
colSums(is.na(data))
```

```
##             name    main_category         currency         deadline
##                0                0                0                0
##         launched            state          backers          country
##                0                0                0                0
##      usd.pledged usd_pledged_real    usd_goal_real    duration_days
##             3028                0                0                0
```

```
##     mean_pledged  percent_pledged           region
##                0                0             3028
```

En aquest cas, s'observa que l'única variable que conté valors buits és la variable usd_pledge. En canvi, s'observa que per a la variable usd_pledge_real no existeix cap valor buit. Ambdues variables, fan referència a la conversió a dolars americans de la quantitat recaptada, amb la diferència que la conversió de usd_pledged està feta per Kickstarter i usd_pledged_real està feta per una plataforma independent fixer.io. En aquest cas, eliminar l'atribut usd_pledged, no suposa una pèrdua d'informació, ja que també tenim la informació a la variable uds_pledge_real i és més consistent. Ara bé, abans d'eliminar usd_pledged, es consultarà quins valors conté usd_pledged_real quan usd_pledged no està informada. És clar que aquestes dues variables no són dependents, ja que el calcul de la conversió es realitza a partir de la variable goal, on a l'inici de la pràctica s'ha vist que aquesta variable goal tampoc contenia valors buits com la variable usd_pledge_real.

```
# Consultem dades quan usd.pledged conté valors buits
head(data[is.na(data$usd.pledged),])
```

```
##                                                      name
## 329                              Duncan Woods - Chameleon EP
## 633                  The Making of Ashley Kelley's Debut Album
## 648                               Butter Side Down Debut Album
## 750                                    Chase Goehring debut EP
## 845 LUKAS LIGETI'S 50TH BIRTHDAY FESTIVAL: ORIGINAL NEW MUSIC!
## 865              The Battle For Breukelen: A Neighborhood Epic
##     main_category currency   deadline   launched     state backers
country
## 329         Music      AUD 2015-08-25 2015-08-04 undefined       0
N,0"
## 633         Music      USD 2015-04-09 2015-03-10 undefined       0
N,0"
## 648         Music      USD 2015-11-26 2015-11-02 undefined       0
N,0"
## 750         Music      USD 2016-03-21 2016-02-23 undefined       0
N,0"
## 845         Music      USD 2015-06-11 2015-05-15 undefined       0
N,0"
## 865  Film & Video      USD 2015-11-07 2015-10-10 undefined       0
N,0"
##     usd.pledged usd_pledged_real usd_goal_real duration_days mean_p
ledged
## 329          NA          3402.08       3211.53            21
0
## 633          NA          3576.00       3500.00            30
0
## 648          NA          7007.80       6000.00            24
0
```

```
## 750          NA          3660.38       3000.00             27
0
## 845          NA          6370.00       5000.00             27
0
## 865          NA          6695.00       6000.00             28
0
##     percent_pledged region
## 329        105.9333   <NA>
## 633        102.1714   <NA>
## 648        116.7967   <NA>
## 750        122.0127   <NA>
## 845        127.4000   <NA>
## 865        111.5833   <NA>
```

```
# Consultem dades quan usd_pledge_real és 0 per saber si usd_pledge ta
mbé té valors buits
head(data[which(data$usd_pledged_real == 0),])
```

```
##                                                              name
## 1                              The Songs of Adelaide & Abullah
## 14 G-Spot Place for Gamers to connect with eachother & go pro!
## 27                                        Superhero Teddy Bear
## 57                             Following the Call - A Novel
## 66                                                 Safer Home
## 68                             Shreddit - Privacy on Reddit
##    main_category currency   deadline   launched  state backers coun
try
## 1     Publishing      GBP 2015-10-09 2015-08-11 failed       0
GB
## 14         Games      USD 2016-03-25 2016-02-09 failed       0
US
## 27        Crafts      GBP 2016-01-05 2015-12-06 failed       0
GB
## 57    Publishing      USD 2016-02-01 2016-01-02 failed       0
US
## 66    Technology      CAD 2015-07-03 2015-06-03 failed       0
CA
## 68    Technology      GBP 2017-07-02 2017-06-02 failed       0
GB
##    usd.pledged usd_pledged_real usd_goal_real duration_days mean_pl
edged
## 1            0                0       1533.95            59
0
## 14           0                0     200000.00            45
0
## 27           0                0      17489.65            30
```

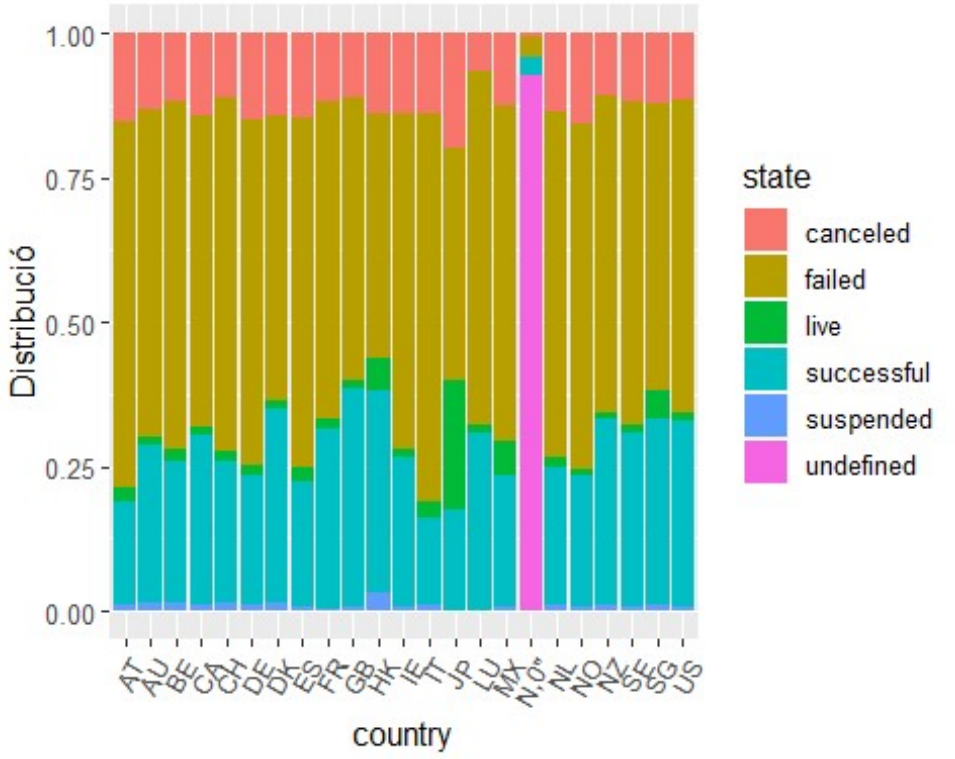
```
0
## 57           0               0      13500.00            30
0
## 66           0               0      39739.31            30
0
## 68           0               0       2579.35            30
0
##    percent_pledged        region
## 1                0        Europe
## 14               0 North America
## 27               0        Europe
## 57               0 North America
## 66               0 North America
## 68               0        Europe
```

S'observa que quan usd.pledge té valors buits, la variable country és N,0 i la variable state és undefined, per tant hi ha una relació entre aquestes variables. I quan la variable usd_pledged_real és 0, usd_pledge no té valors buits.

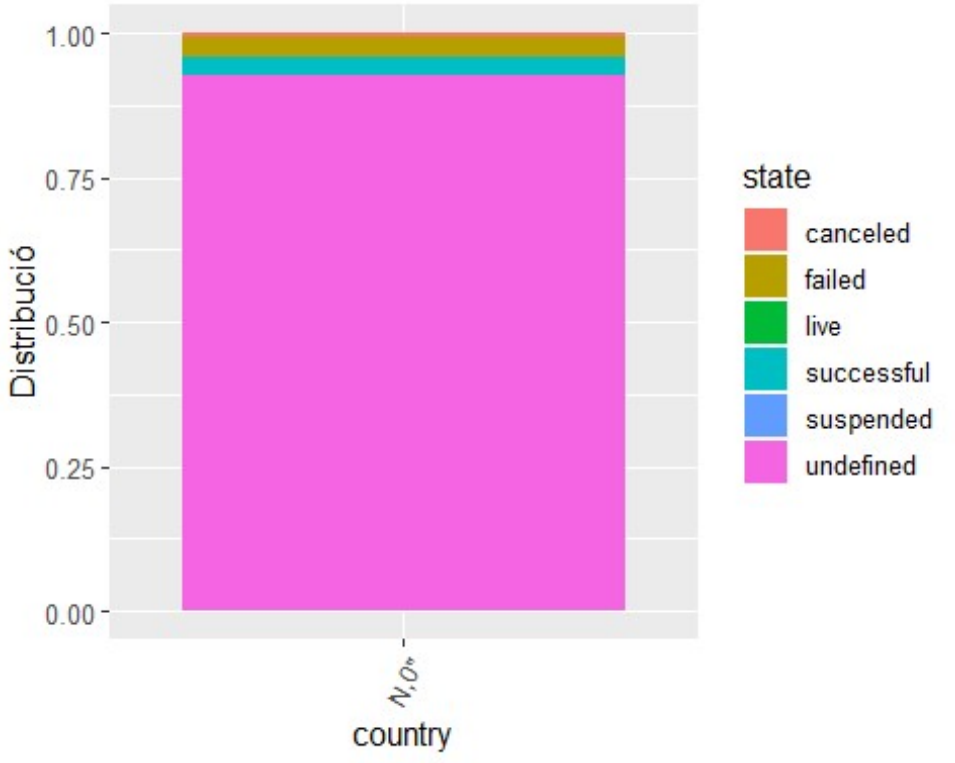
Les úniques variables amb valors que podríem considerar incoherents són state (undefined) i country (N,0"), i sembla que estan relacionades.

A través de la següents gràfiques comprovem que la variable usd.pledged quan conté valors buits, està relacionada amb el valor country = N,0" i state = undefined.

```
# Relació state vs country
ggplot(data,aes(x=country,fill=state))+geom_bar(position="fill")+ylab(
"Distribució")+
   theme(text = element_text(size=12),axis.text.x = element_text(angle
=60, hjust=1))
```




```
# usd.pledged == NA vs state vs country
ggplot(data[is.na(data$usd.pledged),],aes(x=country,fill=state))+geom_
bar(position="fill")+ylab("Distribució")+
   theme(text = element_text(size=12),axis.text.x = element_text(angle
=60, hjust=1))
```



Per tant considerem els valors country = N,0" com a incorrectes i els eliminem de l'estudi.

```
# Eliminem country = N,0"
data <- data[data$country!='N,0"',]
# Quants registres amb data$usd.pledged == NA hi ha?
count(data[is.na(data$usd.pledged),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

Una vegeda, eliminats els registres amb country = N,0", comprovem que no hi ha cap observació al conjunt de dades amb usd.pledged amb valors buits. Finalment, decidim eliminar la variable usd.pledged, perquè usd_pledged_real és més consistent.

```
data$usd.pledged <- NULL
```

A continuació, es comprova per a les variables quantitatives, si el fet de contenir 0 o no es tracta d'un error, o en canvi, és un valor buit legítim.

```
# backers - nombre de persones mecenes
count(data[which(data$backers == 0),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 29901
```

```r
# duration_days
count(data[which(data$duration_days == 0),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
# usd_pledged_real
count(data[which(data$usd_pledged_real == 0),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 29900
```

```r
# usd_goal_real
count(data[which(data$usd_goal_real == 0),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```r
# mean_pledged
count(data[which(data$mean_pledged == 0),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 29901
```

```r
# percent_pledged
count(data[which(data$percent_pledged == 0),])
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 29900
```

S'observa que per a les variables backers, usd_pledge_real, mean_pledged i percent_pledged existeixen cap a uns 30 mil registres informats amb el valor 0. En aquest cas, pot tenir un sentit, és a dir, poden haver-hi projectes en els quals no s'hagi recaptat diners i projectes en els no hi hagi cap nombre de persones mecenes. En aquest cas, el valor té sentit, ara bé per exemple, si el nombre de persones mecenes és zero i el projecte és considerat com a exitós no tindria sentit. I el mateix amb la variable usd_pledged_real, el fet que no es recapten diners per a un projecte i el seu estat és exitós tampoc tindria sentit.

Pel que fa a la variable usd_goal_real i duration_days, no contenen cap valor zero. Per aquestes variables, trobar-se un valor zero, seria considerat un valor perdut o un error. Ja que el més normal és que per a un projecte s'informi d'una quantitat requerida, i almenys hauria d'haver-hi un dia entre la data de llançament del mecenatge i la data final de la campanya de mecenatge del projecte.

Per tant, s'analitza si tenen sentit els valors a zero per a les variables backers i usd_pleged_real.

```
# Te sentit el 0 de backers i usd_pleged_real?
# Valors de la variable state
levels(data$state)
```

```
## [1] "canceled"   "failed"     "live"       "successful" "suspended"
## [6] "undefined"
```

```
# Relació backers vs state
data_aux  <- data[which(data$backers == 0),]
tabla_aux <- table(data_aux$state, data_aux$backers)
tabla_aux
```

```
##
##                 0
##   canceled    7058
##   failed     21831
##   live         548
##   successful     0
##   suspended    464
##   undefined      0
```

```
# Relació usd_pledged_real vs state
data_aux  <- data[which(data$usd_pledged_real == 0),]
tabla_aux <- table(data_aux$state,data_aux$usd_pledged_real)
tabla_aux
```

```
##
##                 0
##   canceled    7058
##   failed     21830
##   live         548
##   successful     0
##   suspended    464
##   undefined      0
```

Després de consultar la taula de contingència, pel que fa a la relació entre la variable backers i state i la variable usd_pledge_real i state. S'observa que efectivament quan usd_pledge_real és 0, cap projecte ha sigut exitós, el mateix passa amb la variable backers.

Identificació i tractament de valors extrems.

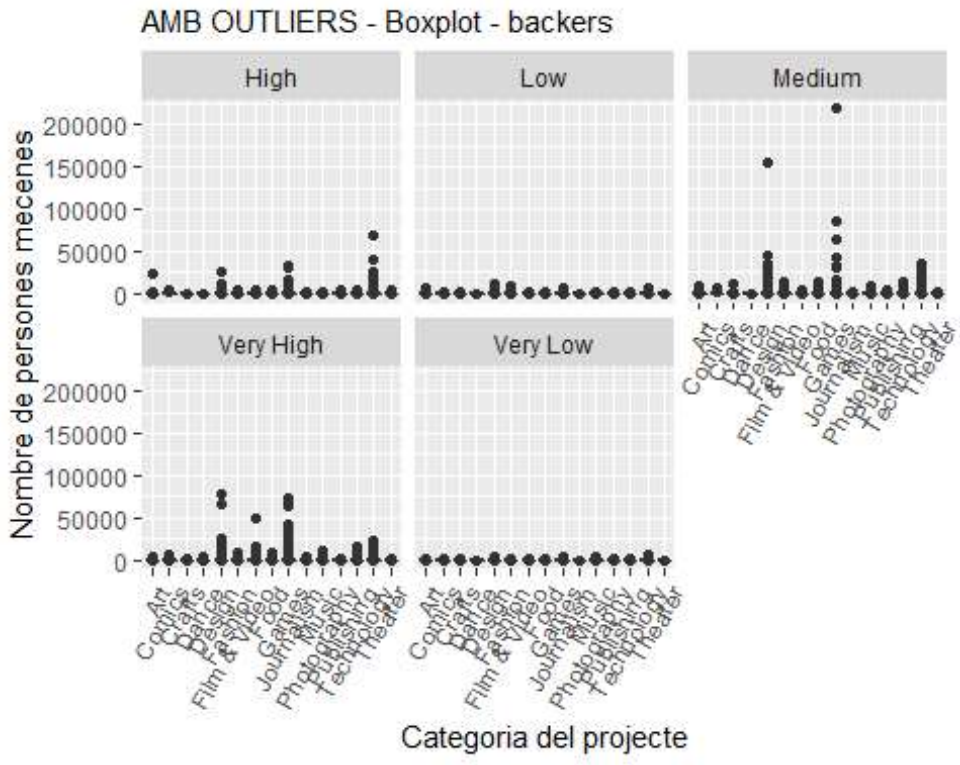
Com ja sabem, els valors extrems són les dades que difereixen significativament dels valors de les distribucions normals d'una variable. Els valors estan molt lluny respecte als altres, sobre 3 desviacions estàndard sobre la mitjana del conjunt. Es generen sospites si les dades han sigut generades amb el mateix mecanisme o no. Per tant, són una amenaça important per a la validesa i generalització dels resultats, poden causar problemes en l'anàlisi estadística de les dades, com augmentar la variància de l'error; si es distribueixen de forma no aleatòria, s'alteren les probabilitats de cometre errors de tipus I i II amb els contrasts d'hipòtesis; també poden influir o esbiaixar greument en les estimacions que poden ser d'interès important, ja que poden no ser generades per la població que ens interessa. En resum, augmenten de manera dràstica els errors d'inferència i redueix dràsticament la força i el poder de les proves estadístiques.

Pot ser interessant discretitzar goal per visualitzar amb més detall les variables resultat

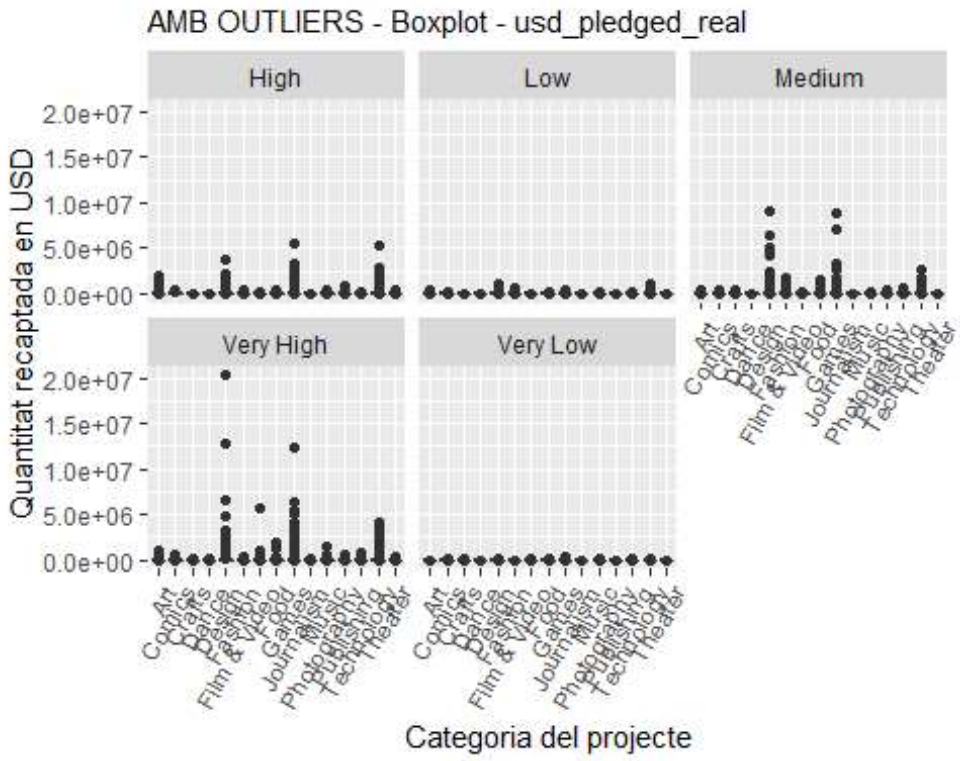
```
#centers <- sort(kmeans(data$usd_goal_real,5)$centers)
#cl <- kmeans(data$usd_goal_real,centers)
#
#vlow   <- subset(data$usd_goal_real, cl$cluster==1)
#low    <- subset(data$usd_goal_real, cl$cluster==2)
#med    <- subset(data$usd_goal_real, cl$cluster==3)
#high   <- subset(data$usd_goal_real, cl$cluster==4)
#vhigh  <- subset(data$usd_goal_real, cl$cluster==5)
#list(sort(vlow4),sort(low4),sort(med4),sort(high4),sort(vhigh4))
data$usd_goal_disc  <-
            ifelse(data$usd_goal_real<2000, 'Very Low',
                   ifelse(data$usd_goal_real<10000, 'Low',
                          ifelse(data$usd_goal_real<50000, 'Medium',
                                 ifelse(data$usd_goal_real<100000, 'Hi
gh','Very High'))))
data$usd_goal_disc <- as.factor(data$usd_goal_disc)
```

A continuació s'analitzen si hi ha valors extrems i si són errors de les dades o no.

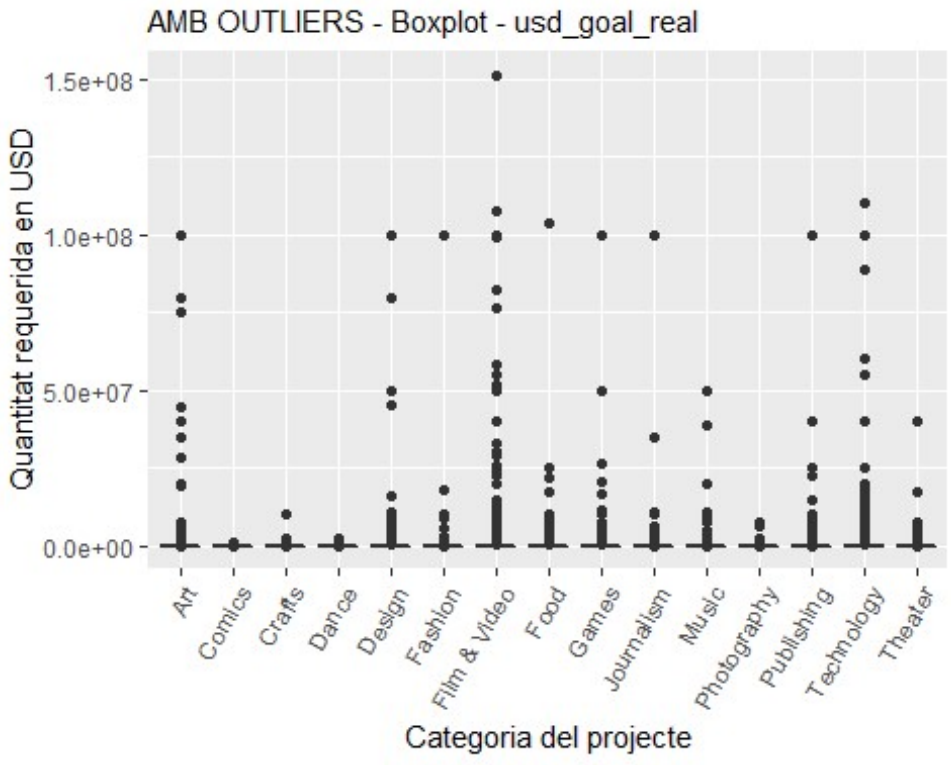
```
# Gràfic - Boxplot - backers
outlier1 <- ggplot(data, aes(x=main_category, y=backers)) +
            facet_wrap(~usd_goal_disc) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Nombre de persones mecenes") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("AMB OUTLIERS - Boxplot - backers")
outlier1
```



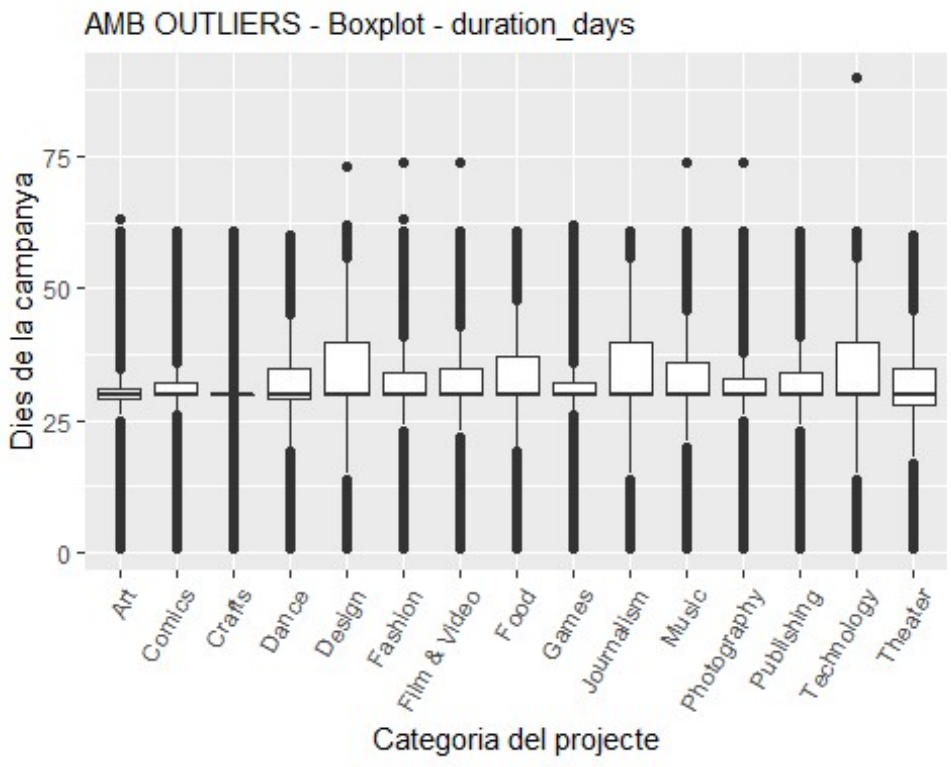

```r
# Gràfic - Boxplot - usd_pledged_real
outlier2 <- ggplot(data, aes(x=main_category, y=usd_pledged_real)) +
            facet_wrap(~usd_goal_disc) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Quantitat recaptada en USD") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("AMB OUTLIERS - Boxplot - usd_pledged_real")
outlier2
```



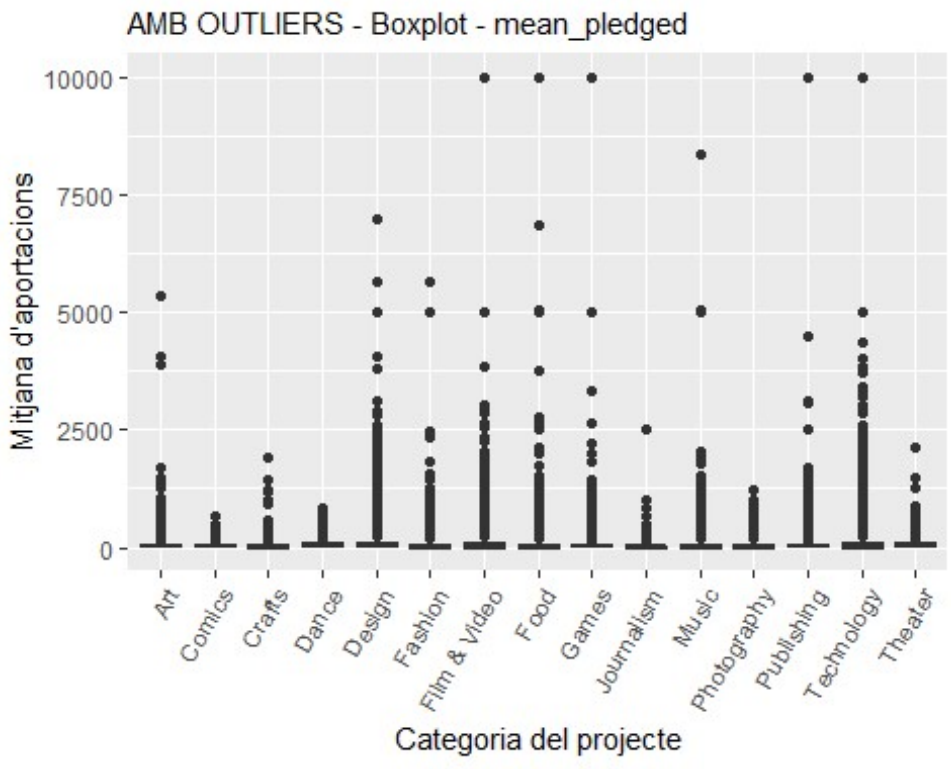

```
# Boxplot - usd_goal_real
outlier3 <- ggplot(data, aes(x=main_category, y=usd_goal_real)) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Quantitat requerida en USD") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("AMB OUTLIERS - Boxplot - usd_goal_real")
outlier3
```



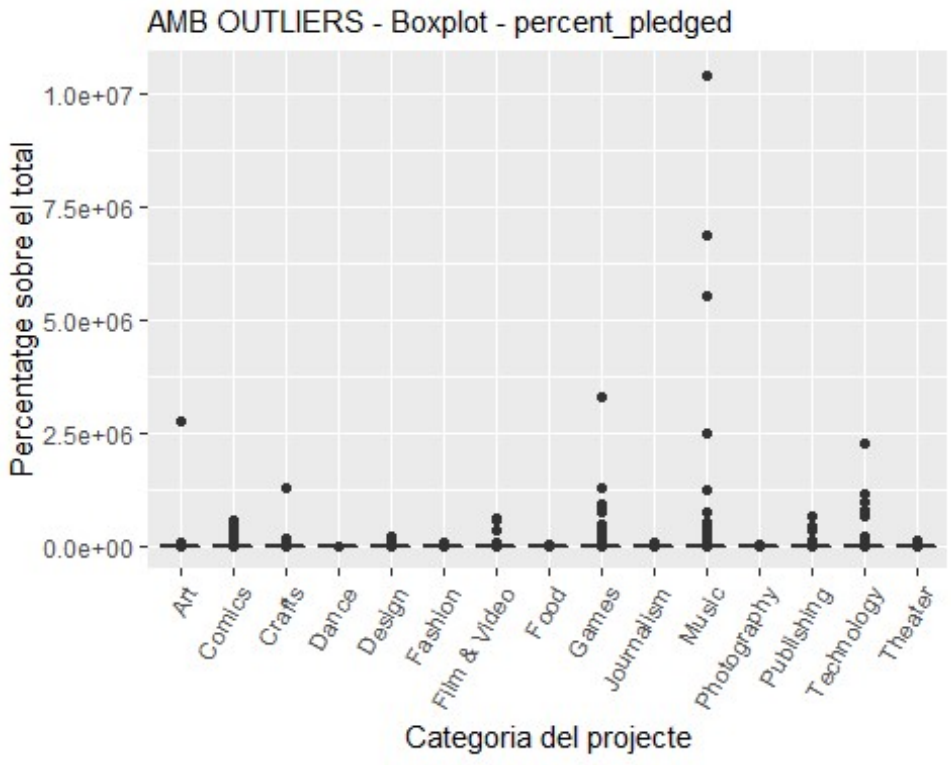

```
# Boxplot - duration_days
outlier4 <- ggplot(data, aes(x=main_category, y=duration_days)) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Dies de la campanya") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("AMB OUTLIERS - Boxplot - duration_days")
outlier4
```




```
# Boxplot - mean_pledged
outlier5 <- ggplot(data, aes(x=main_category, y=mean_pledged)) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Mitjana d'aportacions") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("AMB OUTLIERS - Boxplot - mean_pledged")
outlier5
```




```r
# Boxplot - percent_pledged
outlier6 <- ggplot(data, aes(x=main_category, y=percent_pledged)) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Percentatge sobre el total") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("AMB OUTLIERS - Boxplot - percent_pledged")
outlier6
```




```
# Agafem els valors dels outliers
outlier_backers <- boxplot(data$backers, plot = FALSE)$out
outlier_pledged <- boxplot(data$usd_pledged_real, plot = FALSE)$out
outlier_goal    <- boxplot(data$usd_goal_real, plot = FALSE)$out
outlier_days    <- boxplot(data$duration_days, plot = FALSE)$out
outlier_mean    <- boxplot(data$mean_pledged, plot = FALSE)$out
outlier_percent <- boxplot(data$percent_pledged, plot = FALSE)$out
# Recompte d'outliers, entesos com els que queden "fora" del boxplot
outliersNumber <- c(length(outlier_backers), length(outlier_pledged),
length(outlier_goal),
                    length(outlier_days), length(outlier_mean), length
(outlier_percent))
outliersNumber

## [1] 24444 27114 21736 53364 12890 11285
```

Si observem el recompte de valors extrems, és un nombre molt alt, on el fet d'eliminar tots els outliers pot tenir un efecte considerable sobre el conjunt de dades en el que estem treballant. Però el fet de deixar aquests valors, també afecta les mitjanes, variàncies…

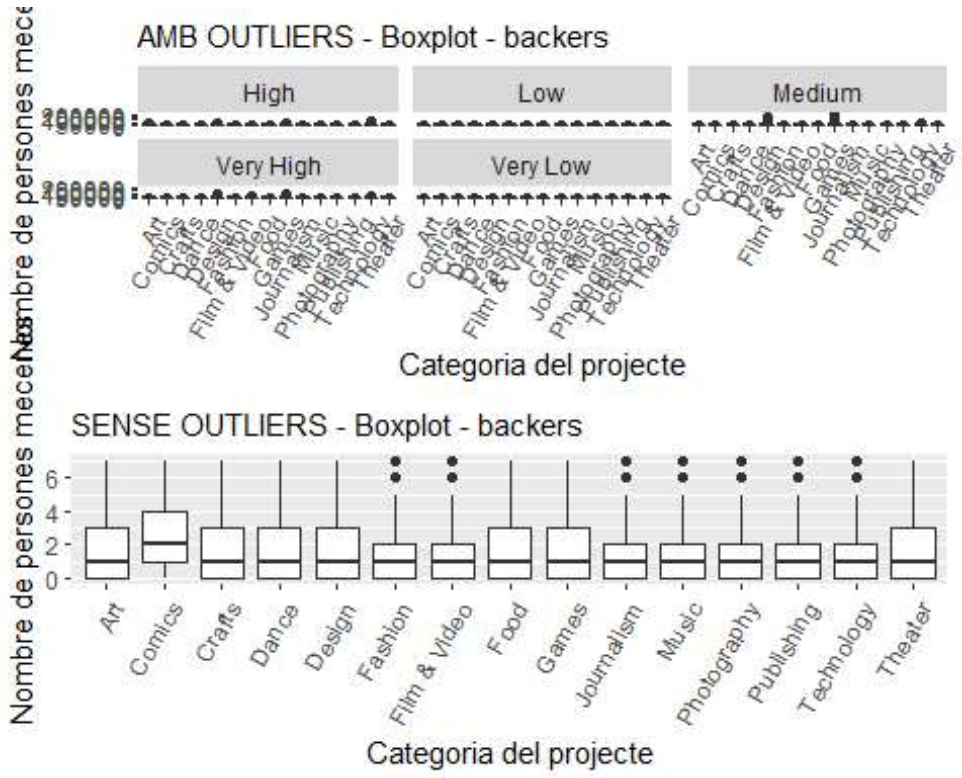
No podem però eliminar aquests outliers, no només formen part del conjunt sino que són casos d'interés per l'estudi.

Optem per realitzar una comparativa de com quedarien els resultats si eliminem els valors extrems que estàn més enllà d'un percentil determinat. Ja que pot ser interessant, deixar els valors extrems i saber quin és el perfil de projectes amb aquest tipus de dades.

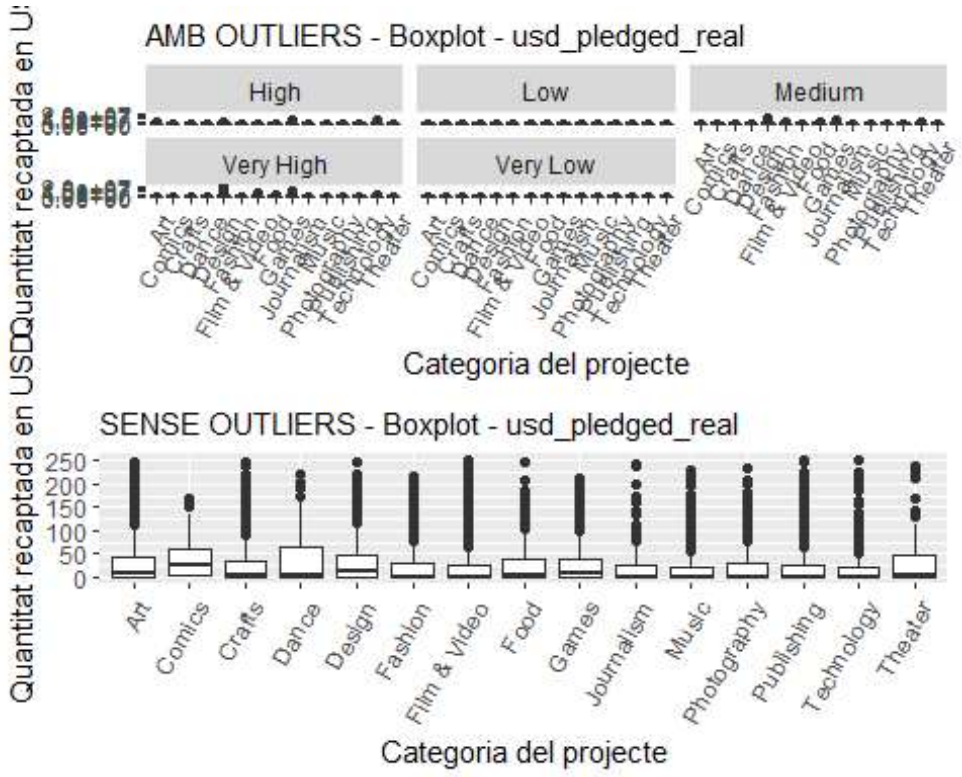
```r
percentil <- 0.5
data_aux <- data[data$backers          < quantile(data$backers, percen
til) &
                 data$usd_pledged_real < quantile(data$usd_pledged_rea
l, percentil) &
                 data$usd_goal_real    < quantile(data$usd_goal_real,
percentil) &
                 data$duration_days    < quantile(data$duration_days,
percentil) &
                 data$mean_pledged     < quantile(data$mean_pledged, p
ercentil),]
# Files
dim(data_aux)
```

```
## [1] 7607   15
```

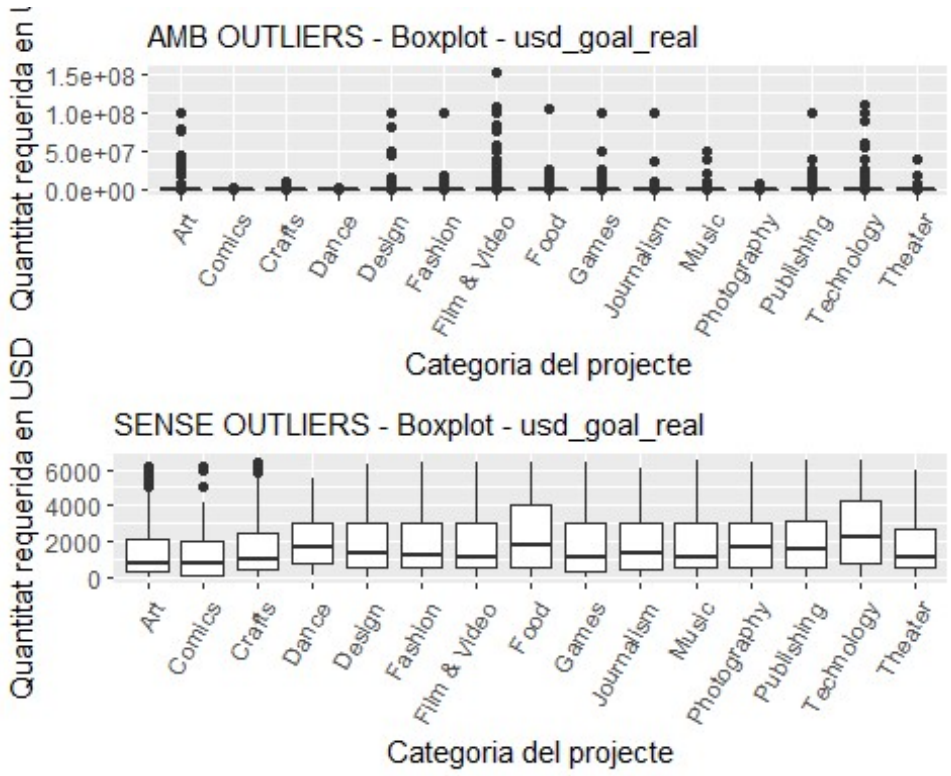
```r
# Comparem amb els resultats anteriors
outlier7 <- ggplot(data_aux, aes(x=main_category, y=backers)) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Nombre de persones mecenes") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("SENSE OUTLIERS - Boxplot - backers")
# Grafica AMB OUTLIERS vs SENSE OUTLIERS - backers
ggarrange(outlier1, outlier7, ncol = 1, nrow = 2)
```



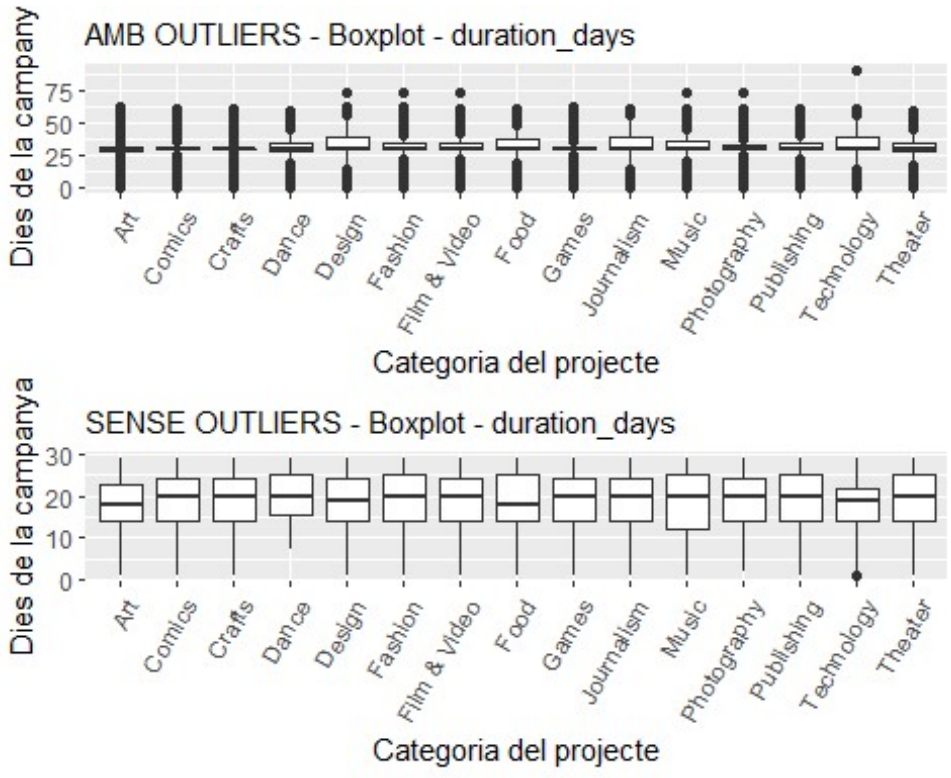
```
# Gràfic - Boxplot - usd_pledged_real
outlier8 <- ggplot(data_aux, aes(x=main_category, y=usd_pledged_real))
+
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Quantitat recaptada en USD") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("SENSE OUTLIERS - Boxplot - usd_pledged_real")
# Grafica AMB OUTLIERS vs SENSE OUTLIERS - usd_pledged_real
ggarrange(outlier2, outlier8, ncol = 1, nrow = 2)
```



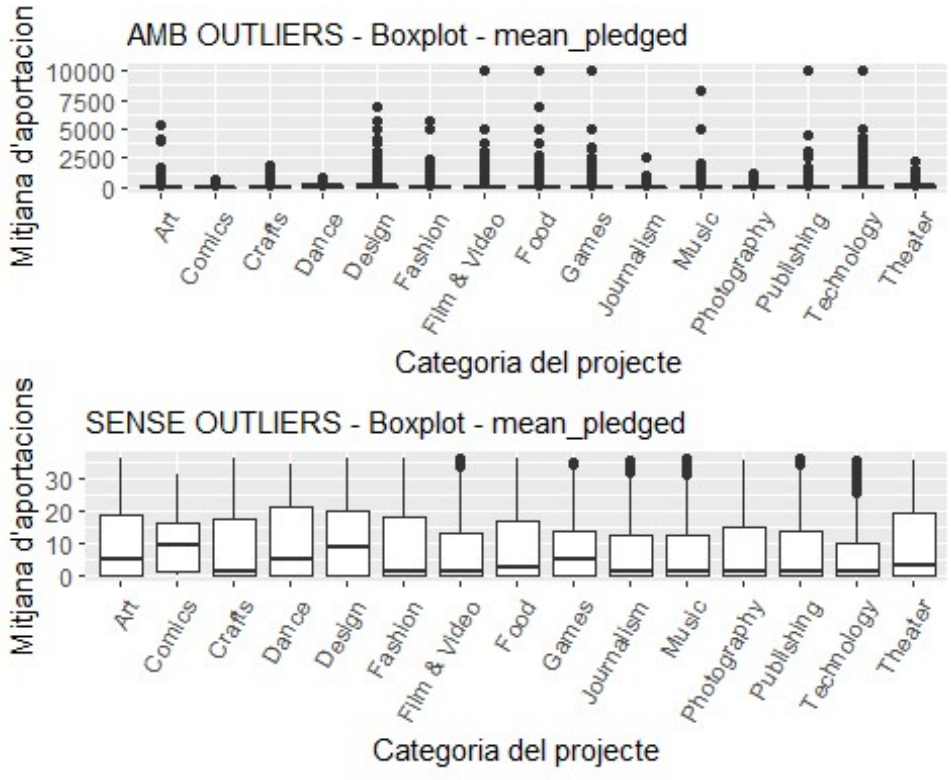

```
# Boxplot - usd_goal_real
outlier9 <- ggplot(data_aux, aes(x=main_category, y=usd_goal_real)) +
            geom_boxplot() + labs(x="Categoria del projecte") +
            scale_y_continuous(name="Quantitat requerida en USD") +
            theme(plot.title = element_text(size=11), axis.text.x= ele
ment_text(angle=60, hjust=1)) +
            ggtitle("SENSE OUTLIERS - Boxplot - usd_goal_real")
# Grafica AMB OUTLIERS vs SENSE OUTLIERS - usd_goal_real
ggarrange(outlier3, outlier9, ncol = 1, nrow = 2)
```



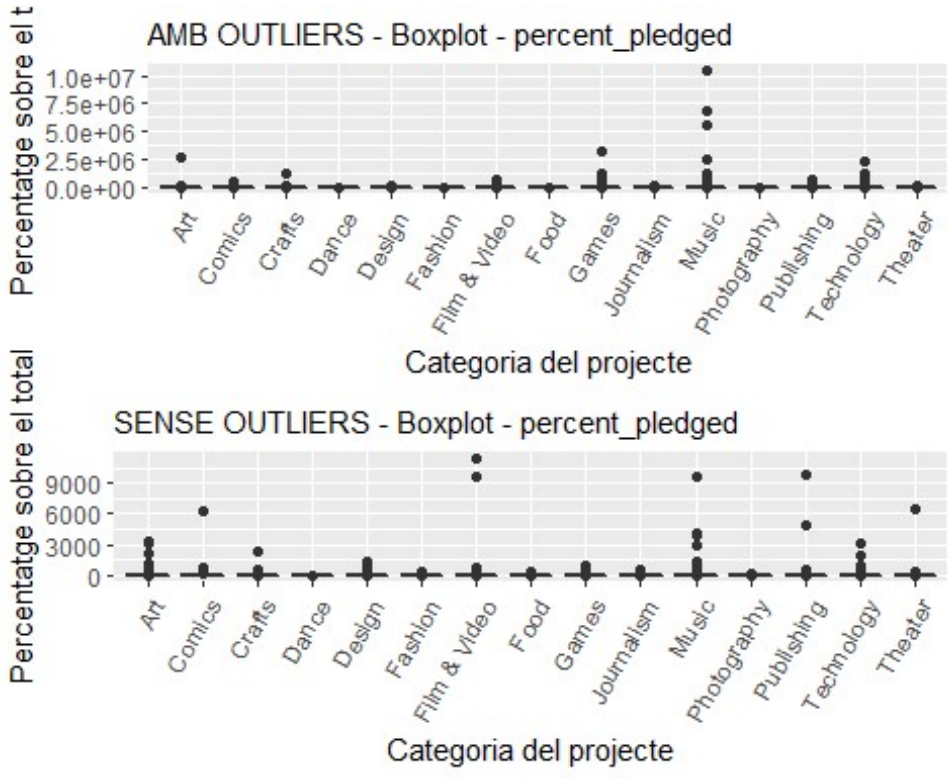

```
# Boxplot - duration_days
outlier10 <- ggplot(data_aux, aes(x=main_category, y=duration_days)) +
             geom_boxplot() + labs(x="Categoria del projecte") +
             scale_y_continuous(name="Dies de la campanya") +
             theme(plot.title = element_text(size=11), axis.text.x= el
ement_text(angle=60, hjust=1)) +
             ggtitle("SENSE OUTLIERS - Boxplot - duration_days")
# Grafica AMB OUTLIERS vs SENSE OUTLIERS - duration_days
ggarrange(outlier4, outlier10, ncol = 1, nrow = 2)
```




```
# Boxplot - mean_pledged
outlier11 <- ggplot(data_aux, aes(x=main_category, y=mean_pledged)) +
             geom_boxplot() + labs(x="Categoria del projecte") +
             scale_y_continuous(name="Mitjana d'aportacions") +
             theme(plot.title = element_text(size=11), axis.text.x= el
ement_text(angle=60, hjust=1)) +
             ggtitle("SENSE OUTLIERS - Boxplot - mean_pledged")
# Grafica AMB OUTLIERS vs SENSE OUTLIERS - mean_pledged
ggarrange(outlier5, outlier11, ncol = 1, nrow = 2)
```




```
# Boxplot - mean_pledged
outlier12 <- ggplot(data_aux, aes(x=main_category, y=percent_pledged))
+
             geom_boxplot() + labs(x="Categoria del projecte") +
             scale_y_continuous(name="Percentatge sobre el total") +
             theme(plot.title = element_text(size=11), axis.text.x= el
ement_text(angle=60, hjust=1)) +
             ggtitle("SENSE OUTLIERS - Boxplot - percent_pledged")
# Grafica AMB OUTLIERS vs SENSE OUTLIERS - percent_pledged
ggarrange(outlier6, outlier12, ncol = 1, nrow = 2)
```



Fitxer de sortida

```
# Escrivim les dades finals en un fitxer csv
write.csv(data, "ks-projects-201801_final.csv")
```

ANÀLISI DE LES DADES

Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar)

L'objectiu principal és analitzar els projectes finalitzats, concretament la relació entre resultat i altres variables relatives als mecenes, categories, aportacions, la localització geogràfica o la campanya.

```
# subset de state amb failed, canceled o succesful (estats de projecte
s amb la campanya finalitzada)
# És l'atribut més rellevant, podem comprovar per paisos, categories,
periode de l'any, o mitjana d'aportacions (backers/pledged)
datares <- data[(data$state=="failed" | data$state=="canceled" | data$
state=="successful"),]
summary(datares)
```

```
##      name                main_category      currency
##  Length:179579      Technology  :22007   USD    :121915
##  Class :character   Film & Video:21632   GBP    : 19471
```

```
##  Mode  :character   Games       :20803   EUR    : 15887
##                     Design      :18001   CAD    :  9460
##                     Publishing  :17576   AUD    :  5246
##                     Music       :16967   MXN    :  1632
##                     (Other)     :62593   (Other):  5968
##     deadline            launched                 state
##  Min.   :2015-01-05   Min.   :2015-01-01   canceled  :21792
##  1st Qu.:2015-08-19   1st Qu.:2015-07-15   failed    :99679
##  Median :2016-05-04   Median :2016-04-01   live      :    0
##  Mean   :2016-05-30   Mean   :2016-04-27   successful:58108
##  3rd Qu.:2017-03-04   3rd Qu.:2017-01-31   suspended :    0
##  Max.   :2018-02-24   Max.   :2018-01-02   undefined :    0
##
##     backers            country       usd_pledged_real
##  Min.   :     0.0   US     :121915   Min.   :       0
##  1st Qu.:     1.0   GB     : 19471   1st Qu.:      15
##  Median :     8.0   CA     :  9460   Median :     410
##  Mean   :   116.8   AU     :  5246   Mean   :   10767
##  3rd Qu.:    53.0   DE     :  4057   3rd Qu.:    3766
##  Max.   :219382.0   FR     :  2873   Max.   :20338986
##                     (Other): 16557
##  usd_goal_real       duration_days    mean_pledged      percent_ple
dged
##  Min.   :        0   Min.   : 1.00   Min.   :    0.00   Min.   :
0
##  1st Qu.:     2000   1st Qu.:30.00   1st Qu.:    8.12   1st Qu.:
0
##  Median :     6500   Median :30.00   Median :   36.36   Median :
8
##  Mean   :    60172   Mean   :33.17   Mean   :   65.47   Mean   :
459
##  3rd Qu.:    20000   3rd Qu.:35.00   3rd Qu.:   75.00   3rd Qu.:
105
##  Max.   :151395870   Max.   :90.00   Max.   :10000.00   Max.   :104
27789
##
##             region         usd_goal_disc
##  Asia & Pacific:  7335   High     :11371
##  Europe        : 39237   Low      :62095
##  Latin America :  1632   Medium   :55499
##  North America :131375   Very High: 9389
##                          Very Low :41225
##
##
```

Un segon set de dades inclou els projectes encara en actiu en el moment en què es van recollir les dades. En aquest cas farem una predicció del resultat basant-nos en les dades del set anterior

```
dataliv <- data[data$state=="live",]
summary(dataliv)
```

```
##      name                main_category    currency       deadline
##  Length:2798        Technology  :377   USD    :1740   Min.   :2016-
08-08
##  Class :character   Film & Video:332   EUR    : 329   1st Qu.:2018-
01-10
##  Mode  :character   Design      :305   GBP    : 279   Median :2018-
01-17
##                     Publishing  :299   CAD    : 132   Mean   :2018-
01-19
##                     Games       :287   MXN    : 107   3rd Qu.:2018-
01-28
##                     Music       :281   AUD    :  70   Max.   :2018-
03-03
##                     (Other)     :917   (Other): 141
##     launched                 state         backers             count
ry
##  Min.   :2016-07-25   canceled  :   0   Min.   :    0.00   US     :
1740
##  1st Qu.:2017-12-04   failed    :   0   1st Qu.:    1.00   GB     :
279
##  Median :2017-12-12   live      :2798   Median :    5.00   CA     :
132
##  Mean   :2017-12-10   successful:   0   Mean   :   68.12   MX     :
107
##  3rd Qu.:2017-12-20   suspended :   0   3rd Qu.:   26.00   IT     :
76
##  Max.   :2018-01-02   undefined :   0   Max.   :10748.00   DE     :
75
##                                                            (Other):
389
##  usd_pledged_real   usd_goal_real      duration_days    mean_pledge
d
##  Min.   :     0.0   Min.   :       1   Min.   : 4.00   Min.   :   0
.000
##  1st Qu.:    10.0   1st Qu.:    2049   1st Qu.:30.00   1st Qu.:   7
.335
##  Median :   259.6   Median :    6524   Median :33.00   Median :  32
.917
##  Mean   :  5871.3   Mean   :   62549   Mean   :39.81   Mean   :  64
.274
```

```
##  3rd Qu.:  1798.5   3rd Qu.:   18469   3rd Qu.:51.00   3rd Qu.:  72
.024
##  Max.   :724423.8   Max.   :99000000   Max.   :61.00   Max.   :2500
.000
##
##  percent_pledged              region       usd_goal_disc
##  Min.   :     0.0   Asia & Pacific: 153   High     : 152
##  1st Qu.:     0.1   Europe        : 666   Low      :1000
##  Median :     5.2   Latin America : 107   Medium   : 874
##  Mean   :   289.4   North America :1872   Very High: 139
##  3rd Qu.:    44.6                         Very Low : 633
##  Max.   :429800.0
##
```

Comprovació de la normalitat i homogeneïtat de la variància.

```
shapiro.test(sample(datares$duration_days,5000))

##
##  Shapiro-Wilk normality test
##
## data:  sample(datares$duration_days, 5000)
## W = 0.82515, p-value < 2.2e-16
```

Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Anotacions per a tenir en compte per a l'analisi:

- Per quins grups de dades separem? Estari bé separar per estat de projecte o projectes amb estat acabat (failed,canceled,successful) i projectes en curs (state=live)
- Es pot calcular la probabilitat que hi ha d'acabar amb èxit un projecte (state=live)
- Per al contrast d'hipotesis, podem veure si hi ha les mateixes possibilitats de tenir èxit als USA, a Europa, o a Àsia (de ser així podem crear una altra variable continent?)
- A major nombre de persones mecenes hi ha més probabilitat de tenir èxit al projecte?
- Les mitjanes d'aportació són les mateixes en totes les categories?
- Per a la normalitat aplicar qqplot i shapiro.test
- La variable a predir pot ser usd_pledged_real

REPRESENTACIÓ DELS RESULTATS A PARTIR DE TAULES i GRÀFIQUES.

RESOLUCIÓ DEL PROBLEMA

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema? ***** ## CODI *****

Recursos

Els següents recursos són d'utilitat per la realització de la pràctica:

● Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

● Megan Squire (2015). Clean Data. Packt Publishing Ltd.

● Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.

● Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.

● Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.

● Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc.

● Tutorial de Github *https://guides.github.com/activities/hello-world*.

Criteris de valoració

Tots els apartat són obligatoris. La ponderació dels exercicis és la següent:

● Els apartats 1, 2 i 6 valen 0,5 punts.

● Els apartats 3,5 i 7 valen 2 punts.

● L'apartat 4 val 2,5 punts.

Es valorarà la idoneïtat de les respostes, que han de ser clares i completes. Les diferents etapes han d'estar ben justificades i acompanyades del codi corresponent. També es valorarà la síntesi i claredat, a través de l'ús de comentaris, del codi resultant, així com la qualitat de les dades finals analitzades.

Format i data de lliurament

Durant la setmana del 23 de desembre el grup podrà lliurar al professor una entrega parcial opcional. Aquesta entrega parcial és molt recomanable per tal de rebre assessorament sobre la pràctica i verificar que la direcció presa és la correcta. Es lliuraran comentaris als estudiants que

hagin efectuat l'entrega parcial però no comptarà per la nota de la pràctica. En l'entrega parcial els estudiants hauran de lliurar per correu electrònic (*mcalvogonza@uoc.edu*) l'enllaç al repositori Github amb el que hagin avançat.

Pel que fa a l'entrega final, cal lliurar un únic fitxer que contingui l'enllaç a Github on hi hagi:

1. Una Wiki on hi hagi els noms dels components del grup i una descripció dels fitxers.
2. Un document Word, Open Office o PDF amb les respostes a les preguntes i els noms dels components del grup. A més, al final de document, haurà d'aparèixer la següent taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació de que l'integrant ha participat en aquell apartat. Tots els integrants han de participar en cadascun dels apartats, de manera que, idealment, els apartats hauran d'estar signats per tots els integrants.
3. Una carpeta amb el codi generat per analitzar les dades.
4. El fitxer CSV amb les dades originals.
5. El fitxer CSV amb les dades finals analitzades.

Aquest document de l'entrega final de la Pràctica 2 s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les 23:59 del dia 7 de gener. No s'acceptaran lliuraments fora de termini.