Tipologia i cicle de vida de les dades

Pràctica 2. Kickstarter

Ivan Borrego Garcia / Cristina Sanchis Puerto

GENER 2020

Índex

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 tota classe, des de cinema independent fins productes tecnològics, passant per còmics, jocs o cuina.

Tot projecte té especificades una dada 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 També s'indica quina és la recaptació mínima de diners per considerar el mecenatge com a exitós, i per tant iniciar el projecte. Durant el període de mecenatge, qualsevol usuari pot participar en el mecenatge, fet que normalment es realitza escollint un dels diferents nivells d'aportació establerts pel creador del projecte, què 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 això que resulta de gran importància per a un nou creador revisar la trajectòria de projectes anteriors semblants. Específicament al nostre cas intentarem donar respostes a possibles dubtes a l'hora d'escollir les variables sobre les quals el creador té poder de decisió. Com podrien ser: la quantitat a demanar, la duració de la campanya o la ubicació, el moment del llançament, específicament per a un projecte de la categoria Jocs.

El creador d'un projecte de jocs, ens contracta per realitzar un estudi per a resoldre les qüestions que té, ja que vol començar una campanya a Kickstarter i vol saber si és millor publicar a agost o setembre o a la regió d'Europa o a Nord Amèrica. A més a més, el projecte ronda els 10.000 USD i no sap quant ha de durar la campanya. D'aquesta manera, s'estudiaran els projectes de la categoria "Games" més recents (2015 en endavant).

A continuació, detallem el conjunt de dades inicial, amb els seus atributs i la seua descripció:

- **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))
suppressPackageStartupMessages(library(nortest))
suppressPackageStartupMessages(library(scales))

## Warning: package 'scales' was built under R version 3.6.2

suppressPackageStartupMessages(library(corrplot))

# 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 1000007540 1000
011046 1000014025 1000023410 1000030581 1000034518 100004195 ...
##  $ name             : Factor w/ 375765 levels "","\177Not Twins - New EP! \
"The View from Down Here\"",..: 332541 135689 365010 344805 77349 206130 2934
62 69360 284139 290718 ...
##  $ category         : Factor w/ 159 levels "3D Printing",..: 109 94 94 91 5
6 124 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",..: 2
288 3042 1333 1017 2247 2463 1996 2448 1790 1863 ...
##  $ goal             : num  1000 30000 45000 5000 19500 50000 1000 25000 125
000 65000 ...
##  $ launched         : Factor w/ 378089 levels "1970-01-01 01:00:00",..: 243
292 361975 80409 46557 235943 278600 187500 274014 139367 153766 ...
##  $ 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 23 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 & Abullah
## 2 1000003930                Greeting From Earth: ZGAC Arts Capsule For ET
## 3 1000004038                                                Where is Hank?
## 4 1000007540          ToshiCapital Rekordz Needs Help to Complete Album
## 5 1000011046 Community Film Project: The Art of Neighborhood Filmmaking
## 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.pledged
## 1 2015-08-11 12:12:28       0     failed       0      GB           0
## 2 2017-09-02 04:43:57    2421     failed      15      US         100
## 3 2013-01-12 00:20:50     220     failed       3      US         220
## 4 2012-03-17 03:24:11       1     failed       1      US           1
## 5 2015-07-04 08:35:03    1283   canceled      14      US        1283
## 6 2016-02-26 13:38:27   52375 successful     224      US       52375
##   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 comprovació 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.

Una vegada, hem consultat si hi ha projectes duplicats, es pot prescindir de la columna ID, que com hem comentat abans, identifica el projecte. A banda, també s'elimina la variable name, ja que per a l'estudi que volem realitzar, l'atribut identificador i name no ens aporta valor.

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

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

```
##         category    main_category         currency         deadline
##                0                0                0                0
##             goal         launched          pledged            state
##                0                0                0                0
##          backers          country      usd.pledged usd_pledged_real
##                0                0             3797                0
##    usd_goal_real
##                0
```

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

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

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

```
# 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, distribuïdes en 15 categories principals. Com que treballarem només amb la categoria principal Games, s'eliminen per tant els atributs referents a categoria (category i main_category).

```
# Filtrem dades: escollim només la categoria Games
data               <- data[data$main_category=="Games",]
# Eliminem atribut category i main_category
data$category      <- NULL
data$main_category <- NULL
```

A continuació, formatem 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.

```
# 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")
```

Fem la consulta dels estats del conjunt de dades inicial.

```
# state
levels(data$state)

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

Posteriorment seleccionem els registres a partir de l'any 2015 i descartem els projectes amb estat "live"

```
# Filtrem dades
# Projectes del 2015 en endavant
data    <- data[data$launched >= "2015-01-01",]
# Projecte amb estat "failed" o "successful"
data    <- data[data$state != "live",]
# Eliminem els diferents nivells de la variable state
data <- droplevels(data)
# Verifiquem els diferents nivells
levels(data$state)
```

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

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).
Com a conseqüència eliminarem la variable deadline i farem una discretització de launched, on s'indicarà el mes de llançament. D'aquesta manera es podrà saber si el mes de l'any en què s'ha llançat el projecte afecta l'estat d'aquest.

```
# Creem la variable duration_days.
data$duration_days <- as.integer(data$deadline-data$launched)
# Discretitzem launched
data$launched      <- factor(month(data$launched), labels= c("January","Febru
ary","March","April","May","June","July","August","September","October","Nove
mber","December"))
# Reducció de la dimensionalitat, eliminem atribut deadline
data$deadline      <- NULL
```

Igualment serà interessant disposar del percentatge total assolit, així com dels valors mitjans de les aportacions per mecenes i de les aportacions per dia.

```
# Creem la variable mean_pledged
data$mean_pledged <- ifelse(data$backers==0, data$mean_pledged<-0, data$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 pledged_byday
data$pledged_byday <- data$usd_pledged_real/data$duration_days
```

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.

```
# Reducció de la dimensionalitat, eliminem atributs goal i pledged
data$goal    <- NULL
data$pledged <- NULL
```

Pel que fa a la quantitat requerida del projecte (usd_goal_real), considerem que és millor discretitzar per grups.

```
# Discretització de usd_goal_real
data$usd_goal_lvl  <-
            ifelse(data$usd_goal_real<2000, '<2000',
                   ifelse(data$usd_goal_real<10000, '>=2000 & <10000',
                          ifelse(data$usd_goal_real<50000, '>=10000 & <50000'
,
                                 ifelse(data$usd_goal_real<100000, '>=50000 &
<100000','>=100000'))))

data$usd_goal_lvl <- factor(data$usd_goal_lvl, levels= c("<2000",">=2000 & <1
0000",">=10000 & <50000",">=50000 & <100000",">=100000"))

#Categories de la variable usd_goal_real
levels(data$usd_goal_lvl)

## [1] "<2000"             ">=2000 & <10000"   ">=10000 & <50000"
## [4] ">=50000 & <100000" ">=100000"
```

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

```
# Creem la variable region.
data$region <- ifelse(data$country=='N,0"', data$region<-NA,
                      ifelse(data$country=="US"|data$country=="CA", data$regi
on<-"North America",
                             ifelse(data$country=="AU"|data$country=="NZ"|dat
a$country=="HK"|
                                      data$country=="SG"|data$country=="JP",
data$region<- "Asia & Pacific",
                                    ifelse(data$country=="MX", data$region<-
"Latin America",
                                           data$region<-"Europe")
                                    )
                             )
                      )

data$region <- as.factor(data$region)

#Categories de la variable region
levels(data$region)

## [1] "Asia & Pacific" "Europe"         "Latin America"  "North America"
```

Una vegada, realitzat el punt de neteja de les dades i decidir com gestionar els valors buits o valors extrems, s'explicarà com queda el conjunt de dades finals. Ja que, després

de l’estudi de l’apartat de neteja de dades, pot ser que també s’elimini algun atribut o descartem registres que puguen ser incoherents o erronis.

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))
```

```
##         currency         launched            state          backers
##                0                0                0                0
##          country      usd.pledged usd_pledged_real    usd_goal_real
##                0                5                0                0
##    duration_days     mean_pledged  percent_pledged    pledged_byday
##                0                0                0                0
##     usd_goal_lvl           region
##                0                5
```

En aquest cas, s'observa que l'única variable que conté valors buits és la variable usd_pledged i region. En canvi, s'observa que per a la variable usd_pledge_real no existeix cap valor buit. Ambdues variables (usd_pledged i usd_pledged_real), fan referència a la conversió a dòlars 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 càlcul 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. Però si sembla que la variable usd_pledged i region poden estar relacionades ja que contenen el mateix nombre d'elements buits. Per tant, també es consultarà el valor de country.

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

```
##        currency launched     state backers country usd.pledged
## 19391       USD    March undefined       0    N,0"          NA
## 47932       USD      May  canceled       0    N,0"          NA
## 119465      GBP      May  canceled       0    N,0"          NA
## 208022      EUR    March  canceled       0    N,0"          NA
## 324245      USD    April suspended       0    N,0"          NA
##        usd_pledged_real usd_goal_real duration_days mean_pledged
## 19391           7902.00       1000.00            40            0
## 47932          17425.00     108435.00            30            0
## 119465          9393.99      10824.23            30            0
## 208022          2999.04       1371.24            30            0
## 324245         16655.00      25000.00            45            0
##        percent_pledged pledged_byday     usd_goal_lvl region
## 19391        790.20000      197.5500            <2000   <NA>
## 47932         16.06953      580.8333         >=100000   <NA>
## 119465        86.78668      313.1330 >=10000 & <50000   <NA>
## 208022       218.71007       99.9680            <2000   <NA>
## 324245        66.62000      370.1111 >=10000 & <50000   <NA>
```

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

```
##        currency launched     state backers country usd.pledged
## 19391       USD    March undefined       0    N,0"          NA
## 47932       USD      May  canceled       0    N,0"          NA
## 119465      GBP      May  canceled       0    N,0"          NA
## 208022      EUR    March  canceled       0    N,0"          NA
## 324245      USD    April suspended       0    N,0"          NA
##        usd_pledged_real usd_goal_real duration_days mean_pledged
## 19391           7902.00       1000.00            40            0
## 47932          17425.00     108435.00            30            0
## 119465          9393.99      10824.23            30            0
## 208022          2999.04       1371.24            30            0
## 324245         16655.00      25000.00            45            0
##        percent_pledged pledged_byday     usd_goal_lvl region
## 19391        790.20000      197.5500            <2000   <NA>
## 47932         16.06953      580.8333         >=100000   <NA>
## 119465        86.78668      313.1330 >=10000 & <50000   <NA>
## 208022       218.71007       99.9680            <2000   <NA>
## 324245        66.62000      370.1111 >=10000 & <50000   <NA>
```

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

```
##      currency launched    state backers country usd.pledged
## 14        USD February   failed       0      US           0
## 215       USD  January canceled       0      US           0
```

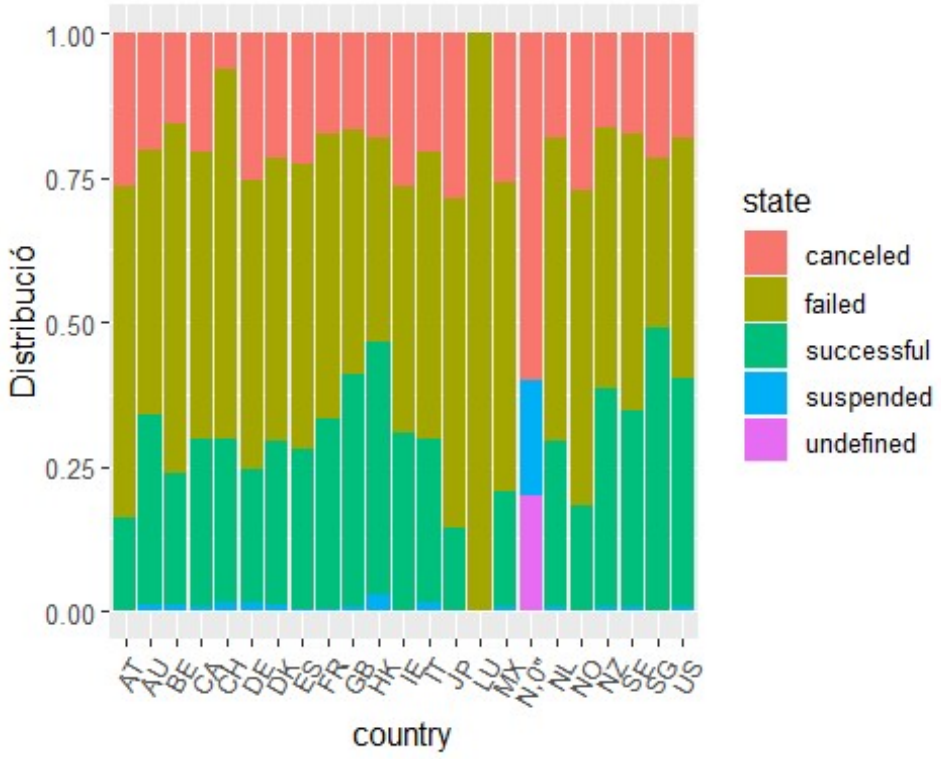
```
## 962       USD  October   failed       0      US           0
## 1044      AUD      May canceled       0      AU           0
## 1104      NOK February canceled       0      NO           0
## 1150      USD     June   failed       0      US           0
##      usd_pledged_real usd_goal_real duration_days mean_pledged
## 14                  0     200000.00            45            0
## 215                 0      15000.00            30            0
## 962                 0      10000.00            32            0
## 1044                0      15029.68            59            0
## 1104                0      87903.26            30            0
## 1150                0       1000.00            30            0
##      percent_pledged pledged_byday      usd_goal_lvl         region
## 14                 0             0          >=100000  North America
## 215                0             0  >=10000 & <50000  North America
## 962                0             0  >=10000 & <50000  North America
## 1044               0             0  >=10000 & <50000 Asia & Pacific
## 1104               0             0 >=50000 & <100000         Europe
## 1150               0             0             <2000  North America
```

S'observa que quan usd.pledge i region, tenen 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.

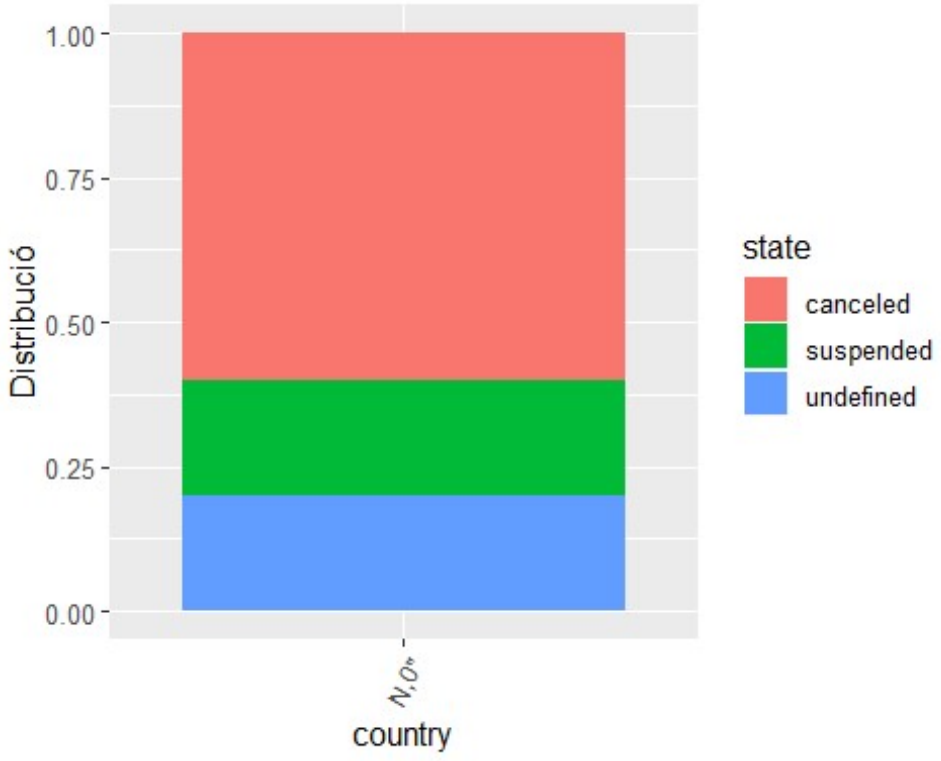
Les úniques variables amb valors que podríem considerar incoherents són state (undefined) i country (N,0"), i sembla que estan relacionades.

A través de la següents gràfiques comprovem, la relació entre state (undefined) i country(N,0"), i quan la variable usd.pledged i region quan contenen valors buits, està relacionada amb el valor country = N,0" i state = undefined.

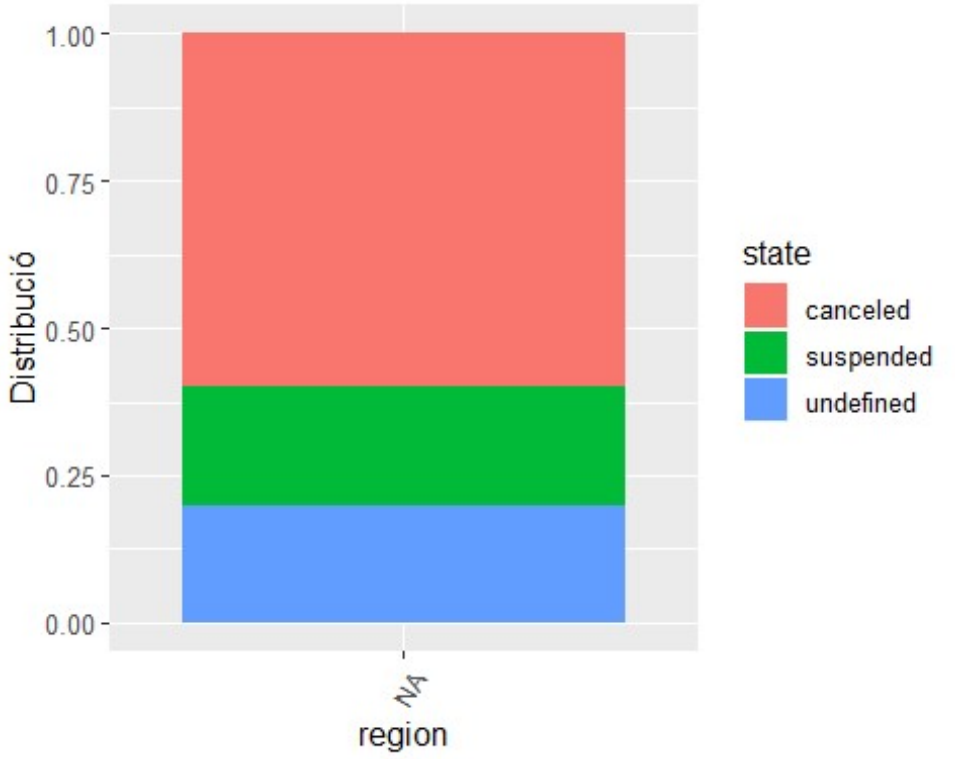
```
# Relació state vs country
ggplot(data,aes(x=country,fill=state))+geom_bar(position="fill")+ylab("Distri
bució")+
   theme(text = element_text(size=12),axis.text.x = element_text(angle=60, hj
ust=1))
```




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




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



A la primera gràfica, observem que l'estat "undefined" existeix tant sols per a quan country té el valor (N,0"). A més si ens fixem, tampoc existeixen els estats "failed" i "successful" quan country és (N,0").

A la segona gràfica, s'observa que quan usd.pledged té valor buit (NA), el country te tan sols el valor (N,0"), i el mateix passa a la tercera gràfica, que quan region té valor buit (NA) el country té tan sols el valor (N,0"), cosa que és normal, ja que la variable region s'ha construït a partir del country.

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
```

```
# Quants registres amb data$region == NA hi ha?
count(data[is.na(data$region),])
```

```
## # A tibble: 1 x 1
##       n
```

```
##   <int>
## 1     0
```

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

```
# Reducció de la dimensionalitat, eliminem atribut usd.pledged
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  1675
```

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

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

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

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  1675
```

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

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

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

```
## # A tibble: 1 x 1
##       n
```

```
##   <int>
## 1  1675

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

## # A tibble: 1 x 1
##       n
##   <int>
## 1  1675

# pledged_byday
count(data[which(data$pledged_byday == 0),])

## # A tibble: 1 x 1
##       n
##   <int>
## 1  1675
```

S'observa que per a les variables backers, usd_pledge_real, mean_pledged, percent_pledged, pledged_byday existeixen cap a uns 1675 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 que 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.

També té un sentit que les variables mean_pledged, percent_pledged pledged_byday tinguin el mateix nombre de registres a 0, ja que tal com s'han creat són dependents de la variable usd_pledge_real.

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"     "successful" "suspended"  "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   653
##   failed     951
##   successful   0
##   suspended   71
##   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   653
##   failed     951
##   successful   0
##   suspended   71
##   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.

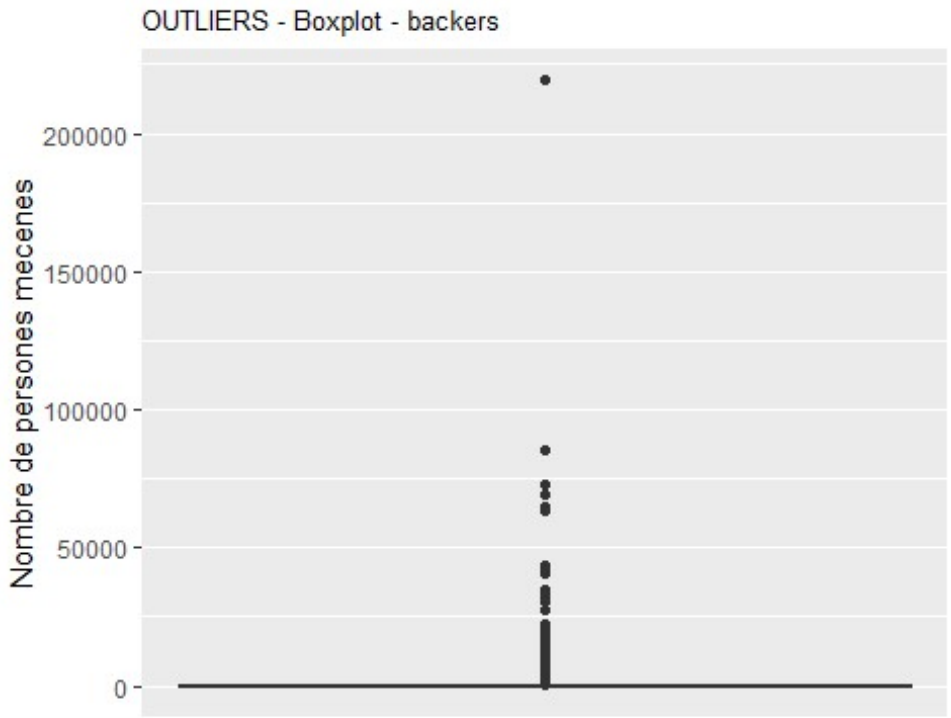
Identificació i tractament de valors extrems

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.

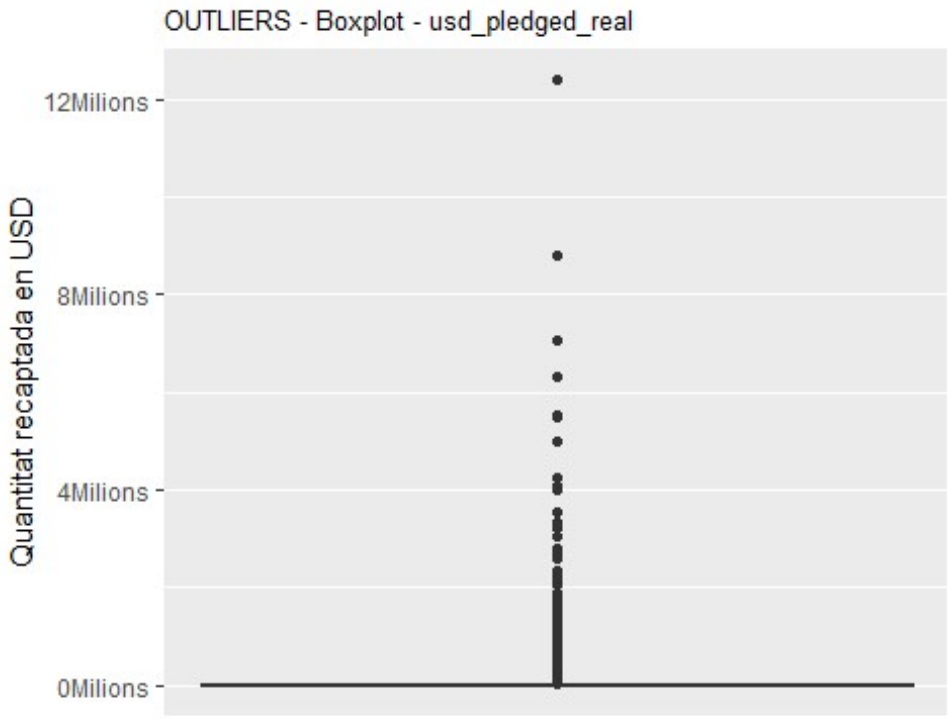
A continuació s'analitzen si hi ha valors extrems i si són errors de les dades o no.

```
# Funció per a formatejar les etiquetes dels imports
million <- function (x) { number_format(accuracy = 1,
```

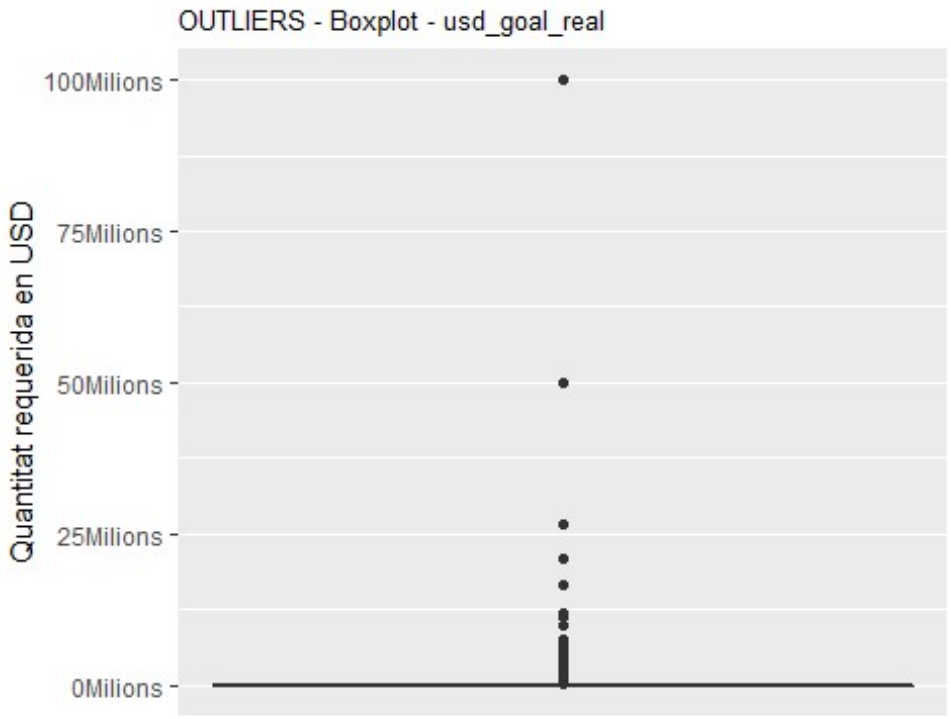
```
                                        scale = 1/1000000,
                                        suffix = "Milions",
                                        big.mark = ",")(x) }

# Funció per a formatejar les etiquetes dels imports
mil <- function (x) { number_format(accuracy = 1,
                                        scale = 1/1000,
                                        suffix = "Mil",
                                        big.mark = ",")(x) }
# Gràfic - Boxplot - backers
outlier1 <- ggplot(data, aes(y=backers)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Nombre de persones mecenes") +
            ggtitle("OUTLIERS - Boxplot - backers") +
            theme(plot.title = element_text(size=10),)
outlier1
```



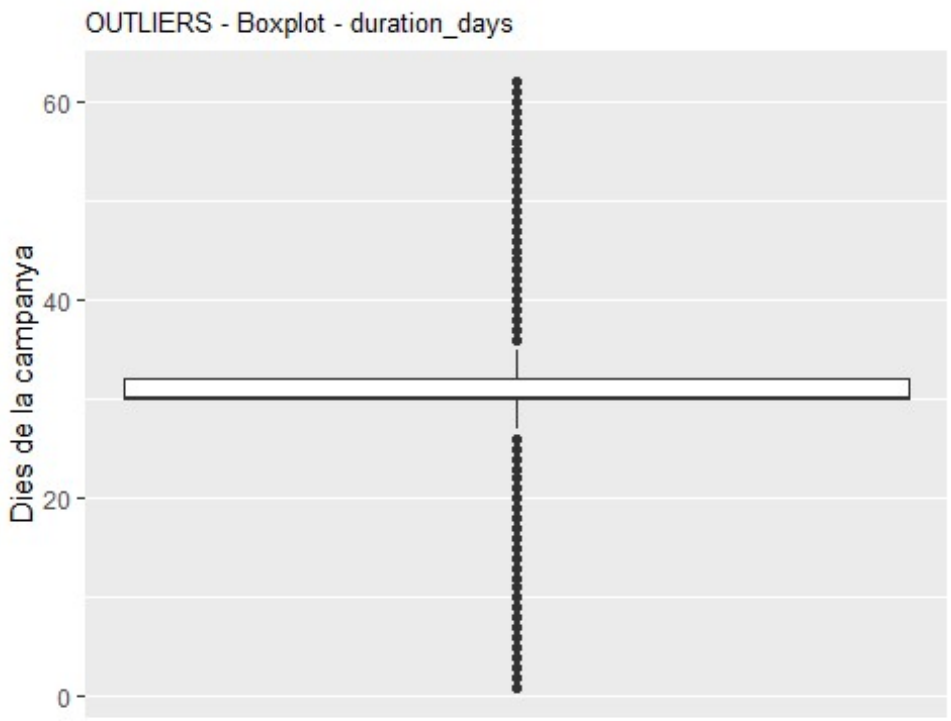

```
# Gràfic - Boxplot - usd_pledged_real
outlier2 <- ggplot(data, aes(y=usd_pledged_real)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Quantitat recaptada en USD", labels = mi
llion) +
            ggtitle("OUTLIERS - Boxplot - usd_pledged_real") +
            theme(plot.title = element_text(size=10),)
outlier2
```



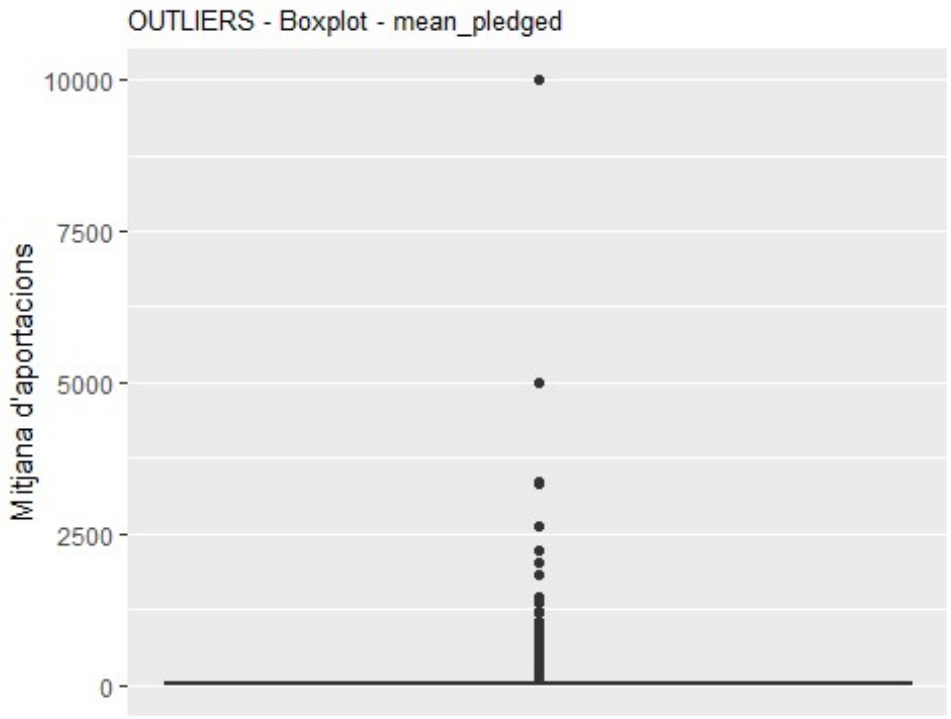

```
# Boxplot - usd_goal_real
outlier3 <- ggplot(data, aes(y=usd_goal_real)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Quantitat requerida en USD", labels = mi
llion) +
            ggtitle("OUTLIERS - Boxplot - usd_goal_real") +
            theme(plot.title = element_text(size=10),)
outlier3
```



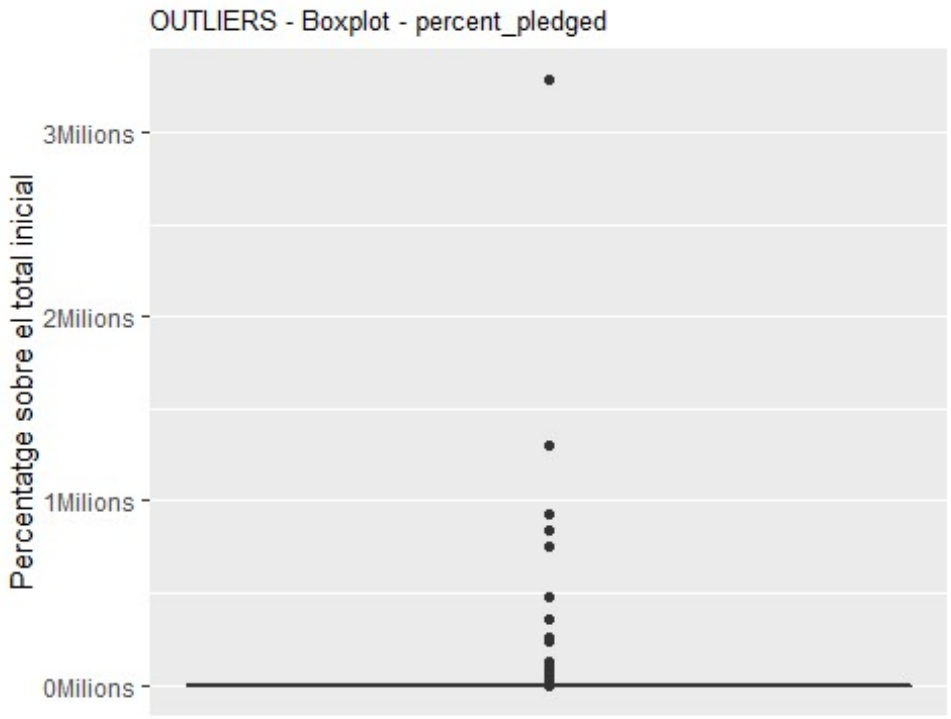

```r
# Boxplot - duration_days
outlier4 <- ggplot(data, aes(y=duration_days)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Dies de la campanya") +
            ggtitle("OUTLIERS - Boxplot - duration_days") +
            theme(plot.title = element_text(size=10),)
outlier4
```



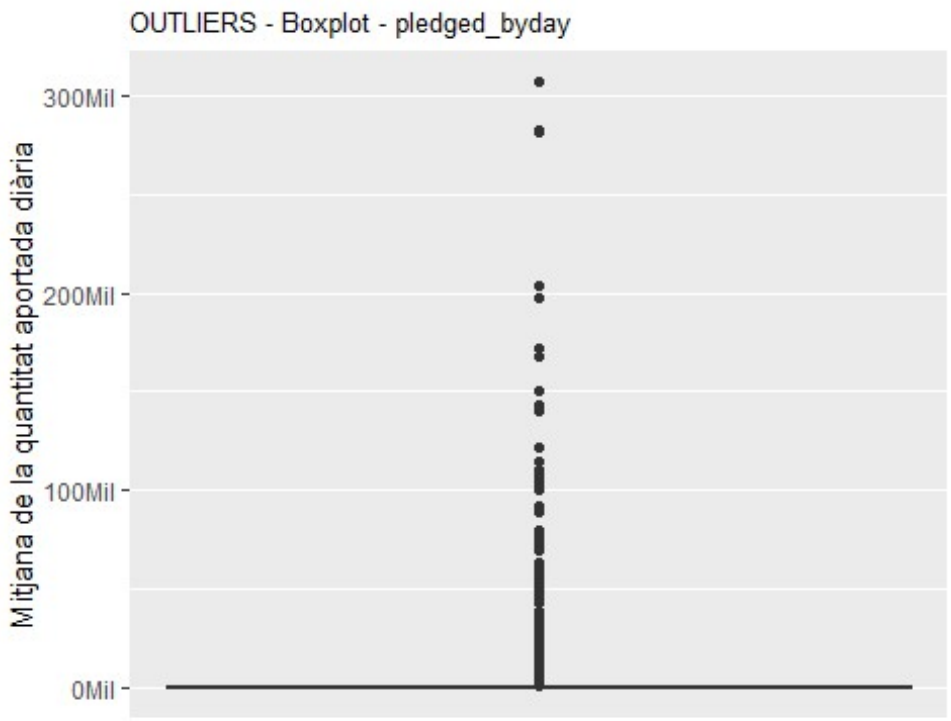

```r
# Boxplot - mean_pledged
outlier5 <- ggplot(data, aes(y=mean_pledged)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Mitjana d'aportacions") +
            ggtitle("OUTLIERS - Boxplot - mean_pledged") +
            theme(plot.title = element_text(size=10),)
outlier5
```




```
# Boxplot - percent_pledged
outlier6 <- ggplot(data, aes(y=percent_pledged)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Percentatge sobre el total inicial", lab
els = million) +
            ggtitle("OUTLIERS - Boxplot - percent_pledged") +
            theme(plot.title = element_text(size=10),)
outlier6
```




```r
# Boxplot - pledged_byday
outlier7 <- ggplot(data, aes(y=pledged_byday)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Mitjana de la quantitat aportada diària"
, labels = mil) +
            ggtitle("OUTLIERS - Boxplot - pledged_byday")+
            theme(plot.title = element_text(size=10),)
outlier7
```




```
# 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
outlier_pledged_byday <- boxplot(data$pledged_byday, plot = FALSE)$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_per
cent), length(outlier_pledged_byday))

outliersNumber

## [1] 2719 2813 2345 7083 1304 2518 2838
```

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…

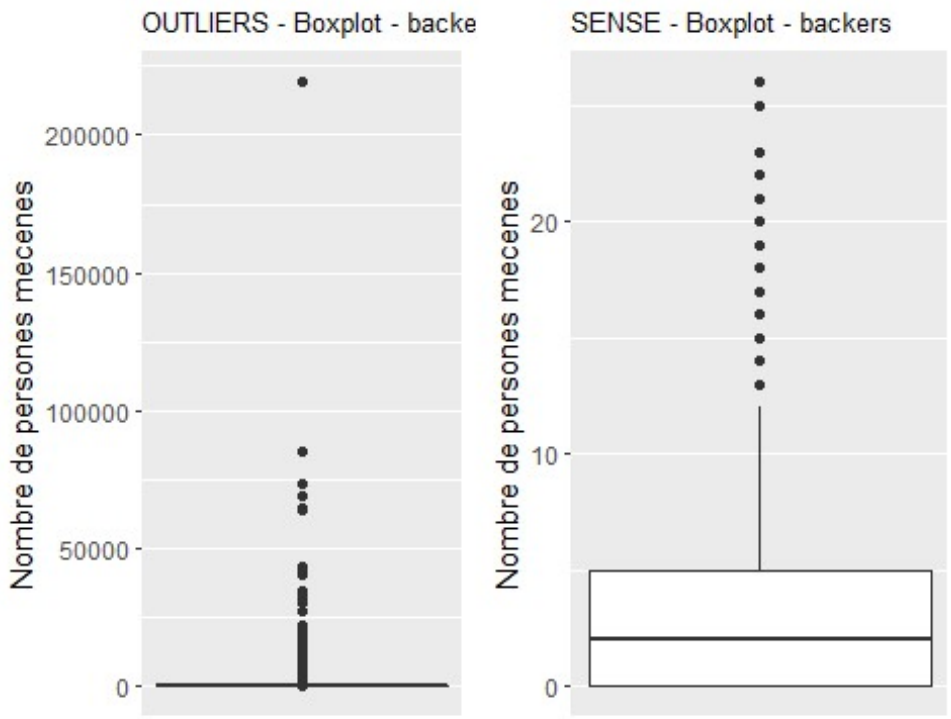
Optem per realitzar una comparativa de com quedarien els resultats si eliminem els valors extrems que estan més enllà d'un percentil determinat, en aquest cas del 0.5.

```
percentil <- 0.5
data_aux <- data[data$backers          < quantile(data$backers, percentil) &
                 data$usd_pledged_real < quantile(data$usd_pledged_real, perc
entil) &
                 data$usd_goal_real    < quantile(data$usd_goal_real, percent
il) &
                 data$duration_days    < quantile(data$duration_days, percent
il) &
                 data$mean_pledged     < quantile(data$mean_pledged, percenti
l) &
                 data$percent_pledged  < quantile(data$percent_pledged, perce
ntil) &
                 data$pledged_byday    < quantile(data$pledged_byday, percent
il),]

summary(data_aux$state)

##   canceled     failed successful  suspended  undefined
##        149        434          0         17          0

# Files
dim(data_aux)

## [1] 600  13
```

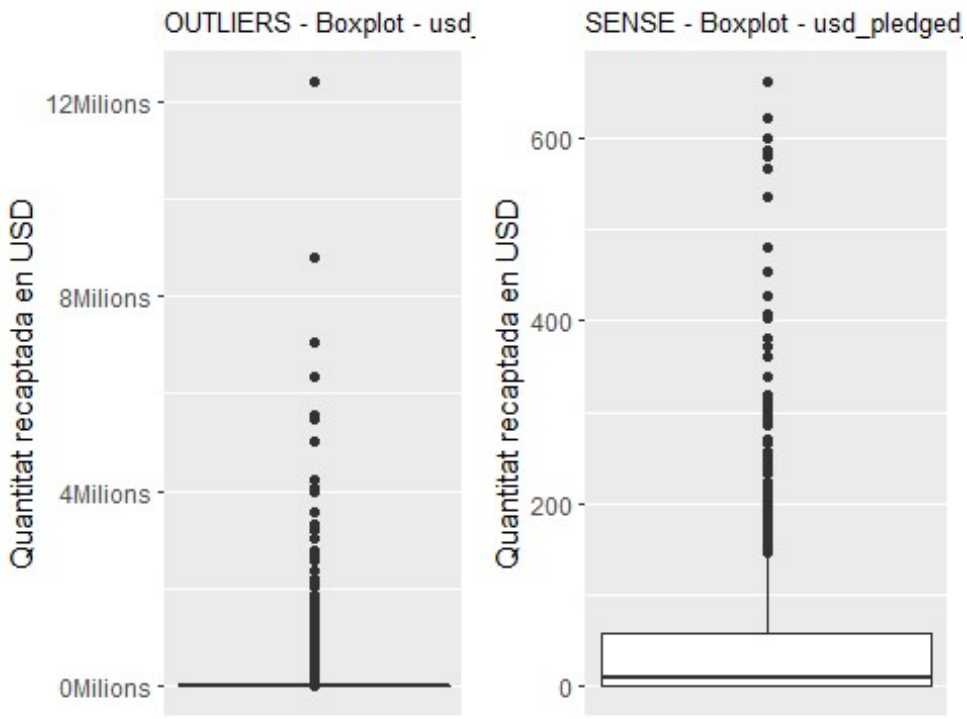
Eliminant els valors extrems que estan més enllà del percentil 0.5, el nombre de registres del conjunt de dades disminueix considerablement, per tant les mitjanes de les dades també es veuran afectades, com es pot comprovar a continuació a les gràfiques de comparatives del conjunt de dades amb valors extrems i sense valors extrems, d'altra banda, s'observa que tots els casos d'èxit són considerats outliers.

Per tant, considerant que el conjunt de dades disminueix molt seriosament i que sobre tot ens carregaríem un grup del conjunt de dades a estudiar que són els estats exitosos, decidim no eliminar els outliers i continuar amb les conclusions de l'estudi. Ja que aquests valors extrems no només no són incorrectes, sinó que, com ja hem indicat formen part de la població a estudiar.

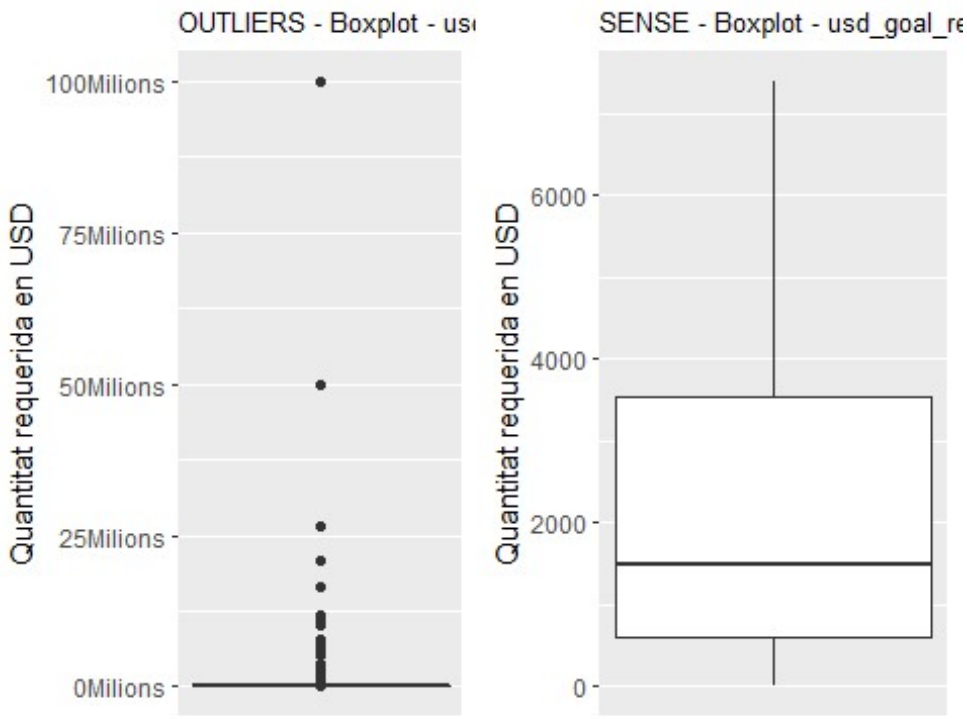
```
# Comparem amb els resultats anteriors
outlier8 <- ggplot(data_aux, aes(y=backers)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Nombre de persones mecenes") +
            ggtitle("SENSE - Boxplot - backers") +
            theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - backers
ggarrange(outlier1, outlier8, ncol = 2, nrow = 1)
```



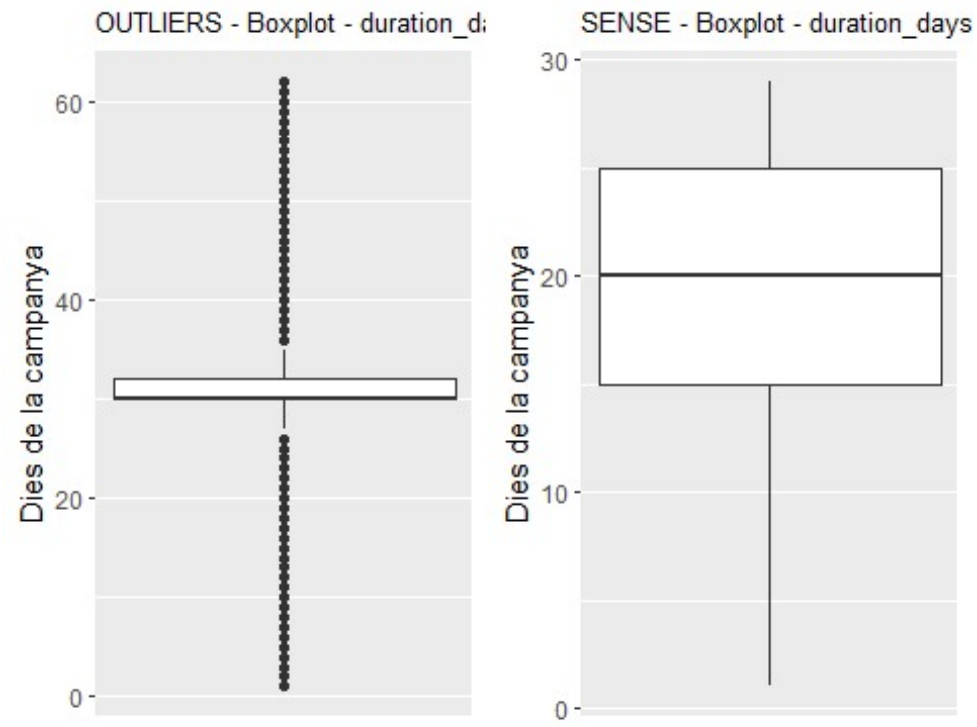

```r
# Gràfic - Boxplot - usd_pledged_real
outlier9 <- ggplot(data_aux, aes(y=usd_pledged_real)) +
            geom_boxplot() +
            scale_x_continuous(name="", breaks = NULL, labels = NULL) +
            scale_y_continuous(name="Quantitat recaptada en USD") +
            ggtitle("SENSE - Boxplot - usd_pledged_real") +
            theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - usd_pledged_real
ggarrange(outlier2, outlier9, ncol = 2, nrow = 1)
```



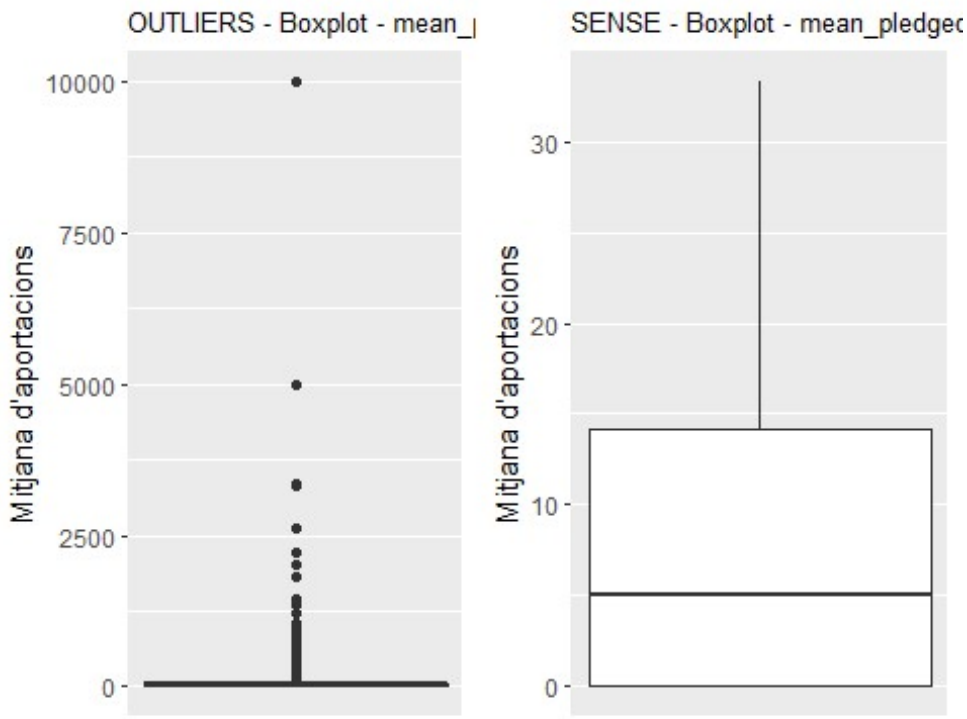

```r
# Boxplot - usd_goal_real
outlier10 <- ggplot(data_aux, aes(y=usd_goal_real)) +
             geom_boxplot() +
             scale_x_continuous(name="", breaks = NULL, labels = NULL) +
             scale_y_continuous(name="Quantitat requerida en USD") +
             ggtitle("SENSE - Boxplot - usd_goal_real") +
             theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - usd_goal_real
ggarrange(outlier3, outlier10, ncol = 2, nrow = 1)
```



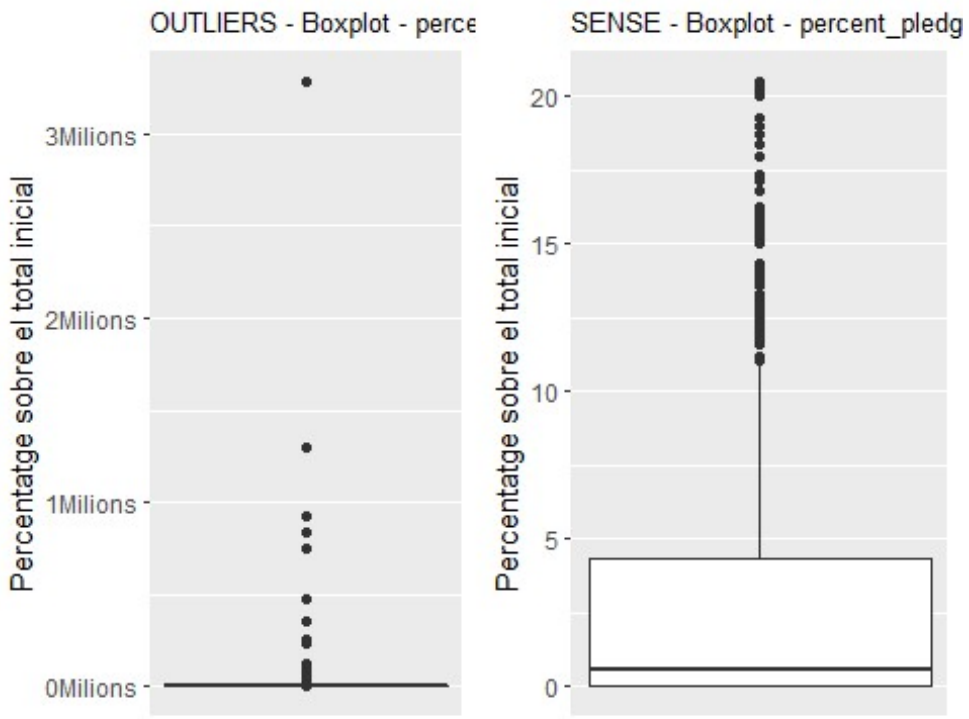

```r
# Boxplot - duration_days
outlier11 <- ggplot(data_aux, aes(y=duration_days)) +
             geom_boxplot() +
             scale_x_continuous(name="", breaks = NULL, labels = NULL) +
             scale_y_continuous(name="Dies de la campanya") +
             ggtitle("SENSE - Boxplot - duration_days") +
             theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - duration_days
ggarrange(outlier4, outlier11, ncol = 2, nrow = 1)
```



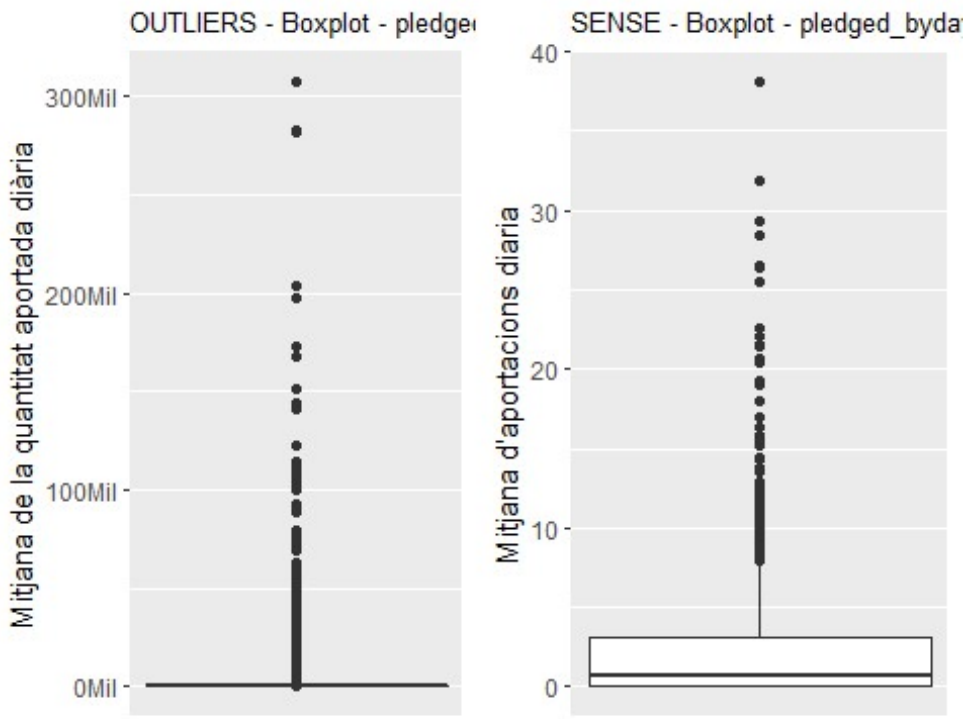

```r
# Boxplot - mean_pledged
outlier12 <- ggplot(data_aux, aes(y=mean_pledged)) +
             geom_boxplot() +
             scale_x_continuous(name="", breaks = NULL, labels = NULL) +
             scale_y_continuous(name="Mitjana d'aportacions") +
             ggtitle("SENSE - Boxplot - mean_pledged") +
             theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - mean_pledged
ggarrange(outlier5, outlier12, ncol = 2, nrow = 1)
```




```r
# Boxplot - mean_pledged
outlier13 <- ggplot(data_aux, aes(y=percent_pledged)) +
             geom_boxplot() +
             scale_x_continuous(name="", breaks = NULL, labels = NULL) +
             scale_y_continuous(name="Percentatge sobre el total inicial") +
             ggtitle("SENSE - Boxplot - percent_pledged") +
             theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - percent_pledged
ggarrange(outlier6, outlier13, ncol = 2, nrow = 1)
```




```r
# Boxplot - pledged_byday
outlier14 <- ggplot(data_aux, aes(y=pledged_byday)) +
             geom_boxplot() +
             scale_x_continuous(name="", breaks = NULL, labels = NULL) +
             scale_y_continuous(name="Mitjana d'aportacions diaria") +
             ggtitle("SENSE - Boxplot - pledged_byday") +
             theme(plot.title = element_text(size=10),)

# Grafica OUTLIERS vs SENSE - pledged_byday
ggarrange(outlier7, outlier14, ncol = 2, nrow = 1)
```



Una vegada, s'han analitzat les estadístiques de valors buits i els valors extrems, recapitulem com queda el conjunt de dades a analitzar al punt 4.

Eliminem les variables country, usd_pledged_real i usd_goal, no les hem eliminat abans perquè volíem estudiar el seu contingut dins del punt 3, per saber si els valors buits, 0… eren coherents o no.

Com que hem creat una variable region, considerem que tenir la variable country és informació redundant i la variable currency tampoc es farà servir, ja que els imports estan tots convertits al dòlar americà.

Eliminem la variable usd_pledged_real, ja que al crear la variable percent_pledged i tenir percentatges es més còmode treballar amb aquesta variable que amb la del conjunt de dades original.

El fet d'haver creat una variable usd_goal_lvl discretitzada, fa que tingam informació redundant i eliminem la variable usd_goal_real.

Creem una variable discretitzada per al percentatge sobre el total inicial assolit al final de la campanya, ja que com els valors extrems són tant dispersos, per a mostrar alguna que altra gràfica pot resultar complexa la seua comprensió. No s'elimina la variable quantitativa perquè és una variable objectiu per a prediure.

```
# Discretització de percent_pledged
data$percent_pledged_lvl  <-
            ifelse(data$percent_pledged <100, '<100',
                   ifelse(data$percent_pledged==100, '=100','>100'))

data$percent_pledged_lvl <- factor(data$percent_pledged_lvl, levels= c("<100"
,"=100",">100"))

#Categories de la variable usd_goal_real
levels(data$percent_pledged_lvl)

## [1] "<100" "=100" ">100"

# Reducció de la dimensionalitat, eliminem atributs country, usd_pledged_real
i usd_goal_real
data$country          <- NULL
data$currency         <- NULL
data$usd_pledged_real <- NULL
data$usd_goal_real    <- NULL
```

Finalment, es filtren els projectes amb estat “failed” i “successful” del conjunt de dades inicial, ja que l’objectiu de l’estudi és estudiar els estats d’aquests projectes.

```
data <- data[(data$state=="failed" | data$state=="successful"),]
data <- droplevels(data)

# Verifiquem els diferents nivells
levels(data$state)

## [1] "failed"     "successful"

# Resum final de l'estructura de les dades
str(data)

## 'data.frame':    16905 obs. of  10 variables:
##  $ launched          : Factor w/ 12 levels "January","February",..: 2 4 3
5 8 10 11 11 1 5 ...
##  $ state             : Factor w/ 2 levels "failed","successful": 1 2 1 2
1 2 2 1 1 1 ...
##  $ backers           : int  0 761 25 448 38 346 624 14 55 32 ...
##  $ duration_days     : int  45 28 25 30 60 30 38 5 60 30 ...
##  $ mean_pledged      : num  0 160.1 34.2 83 37.1 ...
##  $ percent_pledged   : num  0 1883.5 57.07 148.8 2.82 ...
##  $ pledged_byday     : num  0 4352 34.2 1240 23.5 ...
##  $ usd_goal_lvl      : Factor w/ 5 levels "<2000",">=2000 & <10000",..: 5
2 1 3 4 2 2 1 2 4 ...
##  $ region            : Factor w/ 4 levels "Asia & Pacific",..: 4 2 4 4 4
2 4 4 4 4 ...
```

```
##  $ percent_pledged_lvl: Factor w/ 3 levels "<100","=100",..: 1 3 1 3 1 3 3
1 1 1 ...
```

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

```
##       launched           state         backers         duration_days
##  October  :1644   failed    :9179   Min.   :     0.0   Min.   : 1.00
##  March    :1599   successful:7726   1st Qu.:     5.0   1st Qu.:30.00
##  May      :1542                     Median :    38.0   Median :30.00
##  April    :1527                     Mean   :   370.2   Mean   :31.26
##  September:1521                     3rd Qu.:   203.0   3rd Qu.:31.00
##  February :1457                     Max.   :219382.0   Max.   :62.00
##  (Other)  :7615
##   mean_pledged      percent_pledged   pledged_byday
##  Min.   :    0.00   Min.   :      0   Min.   :     0.00
##  1st Qu.:   17.75   1st Qu.:      2   1st Qu.:     3.71
##  Median :   34.34   Median :     45   Median :    57.52
##  Mean   :   49.97   Mean   :    823   Mean   :   925.29
##  3rd Qu.:   58.88   3rd Qu.:    181   3rd Qu.:   349.47
##  Max.   :10000.00   Max.   :3284300   Max.   :307511.17
##
##             usd_goal_lvl             region      percent_pledged_lvl
##  <2000            :4003   Asia & Pacific:  751   <100:9179
##  >=2000 & <10000  :5944   Europe        : 4371   =100:  47
##  >=10000 & <50000 :5354   Latin America :  103   >100:7679
##  >=50000 & <100000: 872   North America :11680
##  >=100000         : 732
##
##
```

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

```
##    launched      state backers duration_days mean_pledged percent_pledged
## 14 February     failed       0            45      0.00000         0.00000
## 15    April successful     761            28    160.12790      1883.49947
## 44    March     failed      25            25     34.24000        57.06667
## 60      May successful     448            30     83.03348       148.79600
## 80   August     failed      38            60     37.10526         2.82000
## 91  October successful     346            30     38.57639       224.76021
##    pledged_byday      usd_goal_lvl        region percent_pledged_lvl
## 14        0.0000          >=100000 North America               <100
## 15     4352.0475   >=2000 & <10000        Europe               >100
## 44       34.2400            <2000 North America               <100
## 60     1239.9667  >=10000 & <50000 North America               >100
## 80       23.5000 >=50000 & <100000 North America               <100
## 91      444.9143   >=2000 & <10000        Europe               >100
```

Per tant, el conjunt de dades final queda:

- **launched** → Mes de llançament de la campanya.
- **state** → Resultat de la campanya (successful/failed).
- **backers** → Nombre de persones mecenes.
- **duration_days** → Duració de la campanya en dies.
- **mean_pledged** → Mitjana d'aportacions al projecte per persona.
- **percent_pledged** → Percentatge sobre el total inicial assolit al final de la campanya.
- **pledged_byday** → Mitjana de la quantitat aportada diària.
- **region** → Regió d'origen del projecte.
- **usd_goal_lvl** → Quantitat demanada inicialment.
- **percent_pledged_lvl** → Rang de percentatge sobre el total inicial assolit al final de la campanya.

```
# 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)

Imaginem que som els creadors del projecte, i volem resoldre les següents qüestions en concret:

- Comprovar probabilitat d'èxit mitjançant percentatge assolit quan el projecte es troba entorn dels 10.000 USD.
- Tenim capacitat per publicar el projecte a Europa o a Amèrica del Nord, quina regió és millor?
- Com que som els creadors del projecte, sabem que el projecte estarà acabat a l'agost, volem saber si ens interessa començar la campanya immediatament o esperar a setembre.

Per tant seleccionarem els grups adients:

CAS ESPECÍFIC META

```
# Cas específic meta
data_goalinf <- data[(data$usd_goal_lvl==">=2000 & <10000"),]
data_goalsup <- data[(data$usd_goal_lvl==">=10000 & <50000"),]
```

Tenim capacitat per publicar el projecte a Europa o a Amèrica del Nord.

CAS ESPECÍFIC REGIÓ

```
# Cas específic regió
data_regeu <- data[data$region=="Europe",]
data_regna <- data[data$region=="North America",]
```

CAS ESPECÍFIC MES DE LLENçAMENT

```
# Cas especific mes de llençament
data_monau <- data[data$launched=="August",]
data_monse <- data[data$launched=="September",]
```

Comprovació de la normalitat i homogeneïtat de la variància.

Per a la comprovació de la normalitat, com que el conjunt de dades que fem servir te més de 5000 registres, no podem utilitzar shapiro.test ja que ens donaria error, per tant

s'utilitzarà ad.test on si p-value és menor que el valor 0.05 les variables no seguiran una distribució normal.

```
# Fem servir ad.test que si funciona per a conjunts de dades de mes de 5000 r
egistres
ad.test(data$backers)

##
##  Anderson-Darling normality test
##
## data:  data$backers
## A = 4825.4, p-value < 2.2e-16

ad.test(data$duration_days)

##
##  Anderson-Darling normality test
##
## data:  data$duration_days
## A = 1498.1, p-value < 2.2e-16

ad.test(data$mean_pledged)

##
##  Anderson-Darling normality test
##
## data:  data$mean_pledged
## A = 2812.4, p-value < 2.2e-16

ad.test(data$percent_pledged)

##
##  Anderson-Darling normality test
##
## data:  data$percent_pledged
## A = 6319.3, p-value < 2.2e-16

ad.test(data$pledged_byday)

##
##  Anderson-Darling normality test
##
## data:  data$pledged_byday
## A = 5147, p-value < 2.2e-16
```

Cap dada està normalitzada perquè el valor de p-value està per davall de 0.05

Mitjançant la funció qqnorm, també es pot observar que les dades no seguixen una normalització, aquest punt el mostrarem a l'apartat de les gràfiques.
Com que les dades no seguixen una distribució normal, pel que fa a la homogeneïtat de la variància fem servir el test de Fligner-Killeen. En aquest cas, la hipòtesi nul·la assumeix

igualtat de variàncies en els diferents grups de dades, de manera que p-valors inferiors al nivell de significació indicaran heteroscedasticitat.

```
# percent_pledged ~ usd_goal_lvl
fligner.test(percent_pledged ~ usd_goal_lvl, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  percent_pledged by usd_goal_lvl
## Fligner-Killeen:med chi-squared = 2429.9, df = 4, p-value <
## 2.2e-16

# percent_pledged ~ region
fligner.test(percent_pledged ~ region, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  percent_pledged by region
## Fligner-Killeen:med chi-squared = 355.31, df = 3, p-value <
## 2.2e-16

# percent_pledged ~ launched
fligner.test(percent_pledged ~ launched, data = data)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  percent_pledged by launched
## Fligner-Killeen:med chi-squared = 252.16, df = 11, p-value <
## 2.2e-16
```

Tenim un p-valor inferior a 0,05, per tant rebutgem la hipòtesi que les variàncies de les mostres són homogènies.

Aplicació de proves estadístiques

PROVES DE CONTRAST HIPÒTESIS

A continuació és realitzen les proves de contrast d'hipòtesis per respondre a les preguntes:

- L'èxit és independent de la regió?
- L'èxit és independent del moment del llançament?
- L'èxit és independent de la quantitat demanada?

L'èxit és independent de la quantitat demanada?

Si el projecte es troba entorn als 10.000 USD ens interessa comprovar l'èxit mitjançant percentatge assolit.

```
# Taula de contingencia pel total, obtenim chi-test
tcont <- as.matrix(xtabs(~usd_goal_lvl+state, data))
tcont
```

```
##                     state
## usd_goal_lvl         failed successful
##   <2000                1624       2379
##   >=2000 & <10000      2996       2948
##   >=10000 & <50000     3281       2073
##   >=50000 & <100000     657        215
##   >=100000              621        111
```

```
summary(tcont)
```

```
## Call: xtabs(formula = ~usd_goal_lvl + state, data = data)
## Number of cases in table: 16905
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 876.3, df = 4, p-value = 2.273e-188
```

```
# Test de wilcoxon pel cas específic
wilcox.test(data_goalinf$percent_pledged, data_goalsup$percent_pledged, corre
ct = FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data_goalinf$percent_pledged and data_goalsup$percent_pledged
## W = 18104427, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

L'èxit és independent de la regió?

Si podem escollir entre publicar el projecte en Europa o en Amèrica ens interessa comprovar-ho.

```
# Taula de contingencia pel total, obtenim chi-test
tcont <- as.matrix(xtabs(~region+state, data))
tcont
```

```
##                 state
## region           failed successful
```

```
##   Asia & Pacific    418        333
##   Europe           2555       1816
##   Latin America      75         28
##   North America    6131       5549
```

```
summary(tcont)
```

```
## Call: xtabs(formula = ~region + state, data = data)
## Number of cases in table: 16905
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 60.57, df = 3, p-value = 4.44e-13
```

```
# Test de wilcoxon pel cas especific
wilcox.test(data_regeu$percent_pledged, data_regna$percent_pledged, correct =
FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data_regeu$percent_pledged and data_regna$percent_pledged
## W = 23735366, p-value = 7.127e-12
## alternative hypothesis: true location shift is not equal to 0
```

L'èxit és independent del moment del llançament?

Si dubtem entre llençar el projecte a l'agost o al setembre podem comprovar-ho.

```
# Taula de contingencia pel total, obtenim chi-test
tcont <- as.matrix(xtabs(~launched+state, data))
tcont
```

```
##            state
## launched    failed successful
##   January      791        621
##   February     821        636
##   March        863        736
##   April        827        700
##   May          844        698
##   June         801        648
##   July         729        636
##   August       716        623
##   September    776        745
##   October      873        771
##   November     736        661
##   December     402        251
```

```
summary(tcont)
```

```
## Call: xtabs(formula = ~launched + state, data = data)
## Number of cases in table: 16905
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 28.61, df = 11, p-value = 0.002608
```

```
# test de wilcoxon pel cas especific
wilcox.test(data_monau$percent_pledged, data_monse$percent_pledged, correct =
FALSE)
```

```
##
##  Wilcoxon rank sum test
##
## data:  data_monau$percent_pledged and data_monse$percent_pledged
## W = 972358, p-value = 0.03703
## alternative hypothesis: true location shift is not equal to 0
```

En cap dels tres casos estudiats podem acceptar la hipòtesi nul·la, ja que els valors p tant de chi-quadrat com de Wilcoxon estan a prop del 0. Per tant els valors no són independents.

CORRELACIONS

La correlació medeix la relació lineal entre dues variables.

El coeficient va de -1 a 1.

- r = 1, la relació és positiva perfecta
- 0 < r < 1 la relació es positiva
- r = 0 no hi ha relació lineal
- -1 < r < 0 la relació es negativa
- r = -1 la relació és negativa perfecta

A continuació, es mostra la matriu de correlació per tal de conèixer la dependència/independència dels atributs, hem de tenir en compte que els atributs per a la matriu de correlació han de ser quantitatius.

Per tant, mitjançant les correlacions, intentem respondre a les següents preguntes:

- Com afecta el nombre de mecenes (backers) amb les quantitats aportades per persona, per dia i percentatge assolit?

Podem generar la taula de correlacions de les variables quantitatives, que són: backers, duration_days, mean_pledged, percent_pledged i pledged_byday.

```
cor(data[,4:7])
```

```
##                 duration_days mean_pledged percent_pledged pledged_byday
## duration_days     1.000000000 -0.004790837     -0.01902226   -0.04760937
## mean_pledged     -0.004790837  1.000000000      0.00303186    0.10032124
## percent_pledged  -0.019022258  0.003031860      1.00000000    0.01741098
## pledged_byday    -0.047609369  0.100321238      0.01741098    1.00000000
```

S'observa que hi ha una forta relació entre la variable backers i la variable pledged_byday, ja que com hem indicat anteriorment quan la relació més prop està del valor 1, més forta és aquesta. Hi ha una relació positiva entre percent_pledged i mean_pledged, encara que aquesta no és tan forta com la relació entre pledged_byday

A més també es pot observar que efectivament el nombre de mecenes afecta directament a les tres variables relacionades amb aportacions (quantitats aportades per persona, per dia, i percentatge assolit).

També podem veure que la duració de la campanya de mecenatge té un efecte lleugerament invers al resultat, però amb poca incidència.

PREDICCIONS

Les variables a predir seran percent_pledged i state. Aplicarem models de regressió utilitzant les variables sobre les quals podem decidir i que hem comprovat poden afectar el resultat.

```
# dataset per la regressió
data_r <- data[(data$usd_goal_lvl==">=2000 & <10000"|data$usd_goal_lvl==">=10
000 & <50000") &
               (data$region=="Europe"|data$region=="North America") &
               (data$launched=="August"|data$launched=="September"),]

# Per a la regressió discretitzem duració en <=30 o >30
data_r$duration_days <- as.factor(ifelse(data_r$duration_days<=30, "<=30",">3
0"))

# Model de Regressió Lineal Múltiple per predir percent_pledged (variable qua
ntitativa)
mlm <- lm(percent_pledged~usd_goal_lvl+region+launched+duration_days, data_r)
summary(mlm)

##
## Call:
## lm(formula = percent_pledged ~ usd_goal_lvl + region + launched +
##     duration_days, data = data_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
##  -248.6  -210.7  -137.5   -42.0 18415.4
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    172.068     48.502   3.548 0.000398 ***
## usd_goal_lvl>=10000 & <50000  -12.600     38.039  -0.331 0.740501
## regionNorth America            76.539     43.545   1.758 0.078963 .
## launchedSeptember              -9.188     38.333  -0.240 0.810597
## duration_days>30              -20.230     41.287  -0.490 0.624203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 814 on 1841 degrees of freedom
## Multiple R-squared:  0.001934,   Adjusted R-squared:  -0.0002343
## F-statistic: 0.892 on 4 and 1841 DF,  p-value: 0.4679
```

El coeficient de determinació R-squared és molt baix, indicant que l'explicació a la variància donada és inferior a l'1%, per tant no tenim un model predictiu acurat.

Els p-valors són molt alts, indicant que les variables són estadísticament irrellevants. El resultat és coherent a causa de la gran quantitat d'outliers i al càlcul fet amb variables qualitatives.

```
# Model de Regressió Logística per predir state (variable qualitativa)
mglm <- glm(state~usd_goal_lvl+region+launched+duration_days, family=binomial
, data_r)
summary(mglm)
```

```
##
## Call:
## glm(formula = state ~ usd_goal_lvl + region + launched + duration_days,
##     family = binomial, data = data_r)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3812  -1.0790  -0.8651   1.1444   1.6137
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.07445    0.12132  -0.614  0.53945
## usd_goal_lvl>=10000 & <50000 -0.70319    0.09565  -7.351 1.96e-13 ***
## regionNorth America           0.35940    0.11005   3.266  0.00109 **
## launchedSeptember             0.18225    0.09628   1.893  0.05837 .
## duration_days>30             -0.20685    0.10382  -1.992  0.04633 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2554.3  on 1845  degrees of freedom
## Residual deviance: 2484.7  on 1841  degrees of freedom
## AIC: 2494.7
##
## Number of Fisher Scoring iterations: 4
```

P-valors baixos indiquen rellevància estadística.

Les variables dicotòmiques indiquen quin valor utilitza la fórmula.

Les variables que més afecten al resultat són la quantitat objectiu >=10.000USD (negativament) i la regió Nord-Amèrica (positivament), seguits de duració>30 (negatiu) i llançament al setembre (positiu).

La diferència entre les desviacions nul·les i residuals (2554-2484) és petita, el qual ens indica que el model pot ser poc explicatiu.

```
# Model: ln(p/1-p) = -0.07445 + (-0.70319 * usd_goal_lvl>=10000) + (0.35940 *
region North America) + (0.18225 * launched September) + (-0.20685 * duration
_days>30)
exp(-0.07445+0.35940+0.18225)

## [1] 1.59552
```

Com que el que tenim és un logaritme de odds, el calculem fent l'exponent de e. El resultat de la predicció és un Odd Ratio de 1.59552.

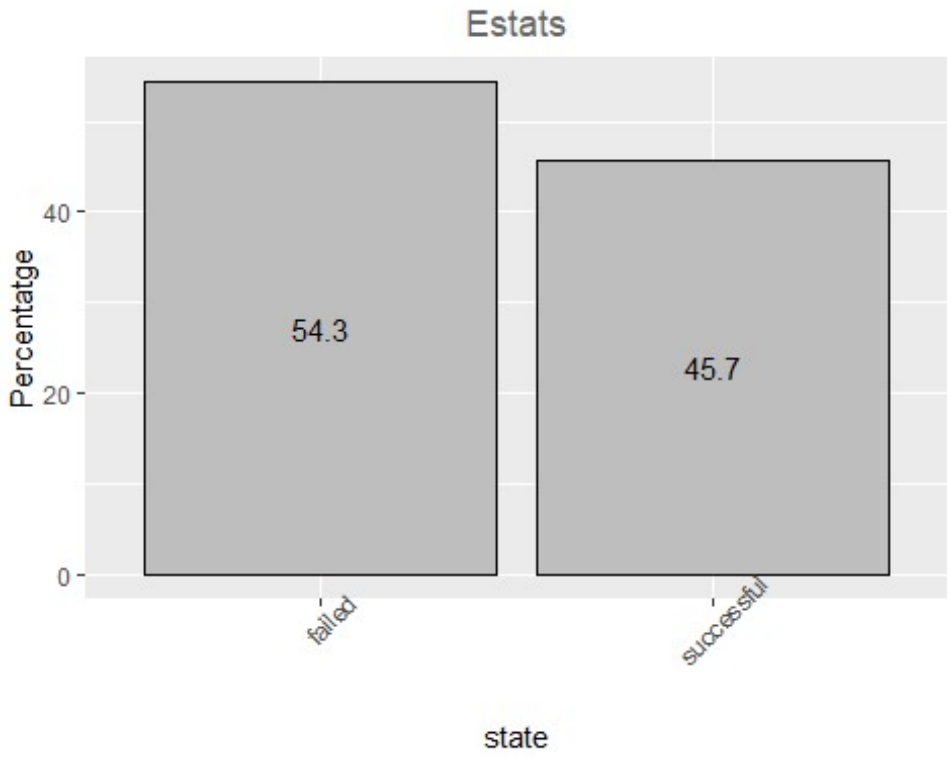
El resultat ens diu que, segons el model i responent als dubtes plantejats, per al nostre projecte de Joc seria recomanable fixar la meta en una quantitat inferior a 10.000USD, fer el llançament a nord-americà al setembre, i establir una duració de campanya no superior als 30 dies.

Les probabilitats d'èxit calculades en aquest cas seran de un 61.47%, per sobre del valor bàsic del model obtingut (48.13%)

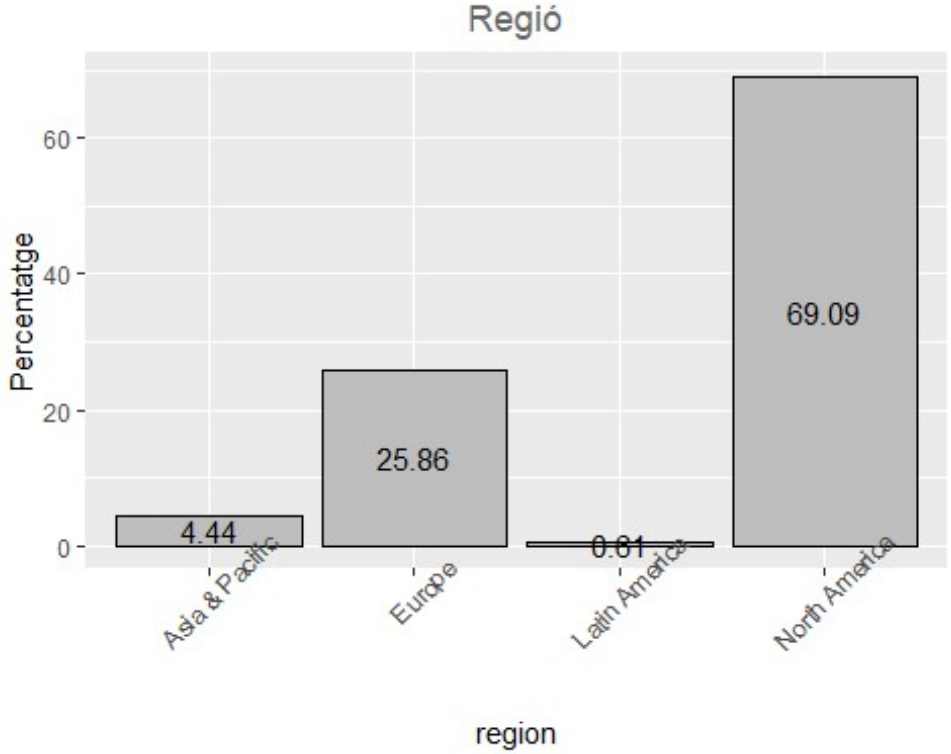
REPRESENTACIÓ DELS RESULTATS A PARTIR DE TAULES i GRÀFIQUES.

Distribucions variables qualitatives

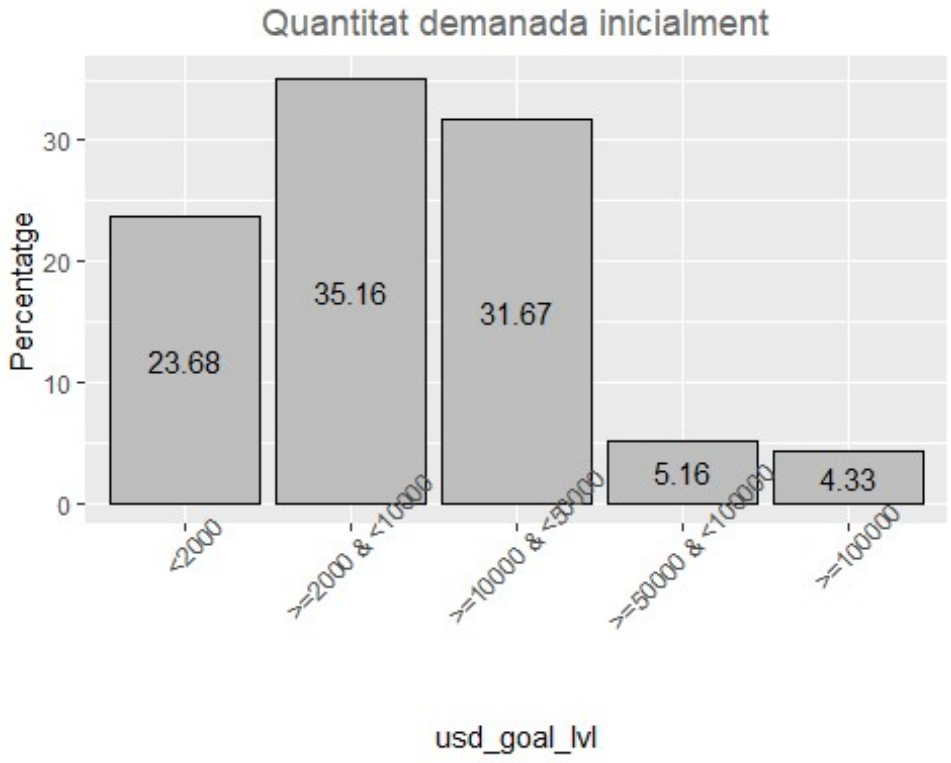
```
# Creem funcio per a construir els grafics amb percentatges
function_barres_percentatges <- function(var1,titolxlab,titol)
{
  # Càlcul freqüència relativa
  freq_rel <- as.data.frame(prop.table(table(var1)))
  # Calculem el percentatge a partir de la freqüència relativa
  freq_rel$Freq <- round((freq_rel$Freq * 100),2)
  # Preparant dades
  df_credit_profile <- data.frame(
                       Categories  = freq_rel$var1,
                       Percentatge = freq_rel$Freq)

  grafic <- ggplot(df_credit_profile, aes(x=Categories, y= Percentatge)) + ge
om_bar(color="black", fill="grey",stat="identity") +
  # Convert to pie (polar coordinates) and add labels
      geom_text(aes(label = Percentatge), position = position_stack(vjust = 0
.5)) +
      xlab(titolxlab)+theme(axis.text.x = element_text(angle = 45)) +
  # Add title
      ggtitle(titol) +
  # Tidy up the theme
      theme(plot.title = element_text(hjust = 0.5, color = "#666666"))

  return(grafic)
}

# state - Diagrama de barres
g4 <- function_barres_percentatges(data$state,"state","Estats")
g4
```



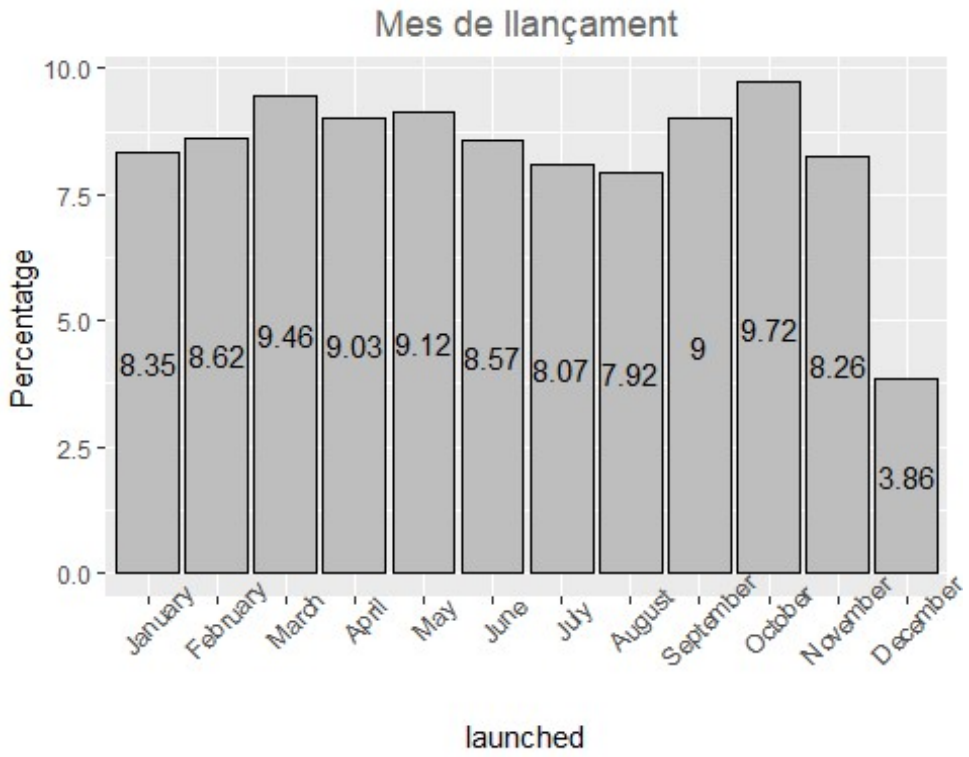

```
# Region - Diagrama de barres
g5 <- function_barres_percentatges(data$region,"region","Regió")
g5
```



```r
# Education - Diagrama de barres
g6 <- function_barres_percentatges(data$usd_goal_lvl,"usd_goal_lvl","Quantitat demanada inicialment")
g6
```



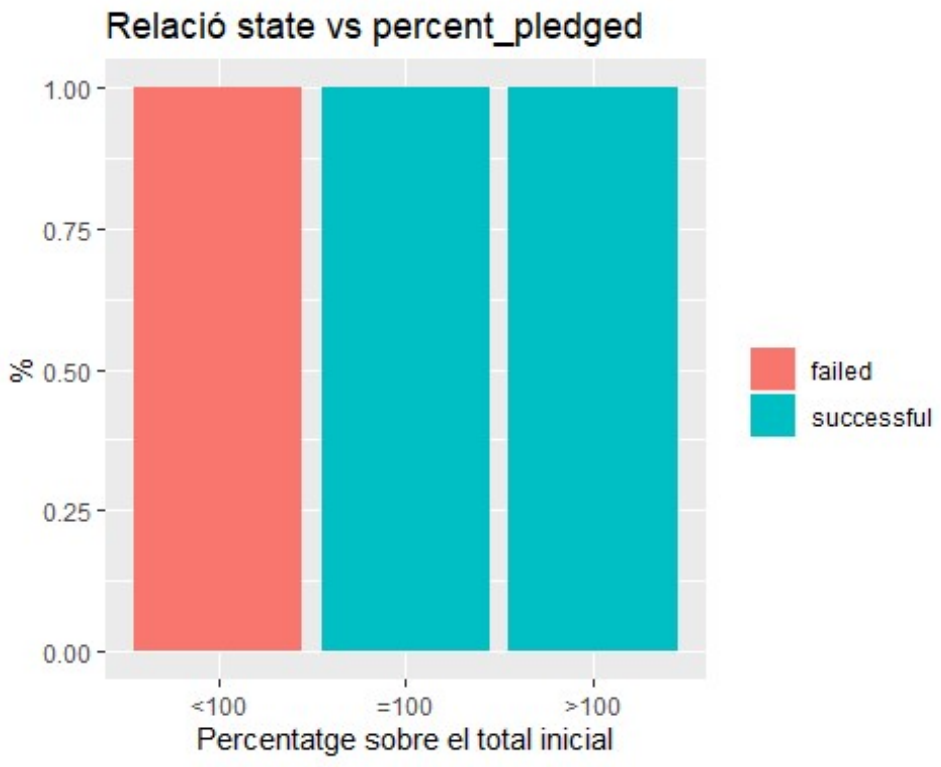

```r
# Mes de llançament - Diagrama de barres
g7 <- function_barres_percentatges(data$launched,"launched","Mes de llançament")
g7
```



Resposta a les preguntes plantejades amb gràfiques

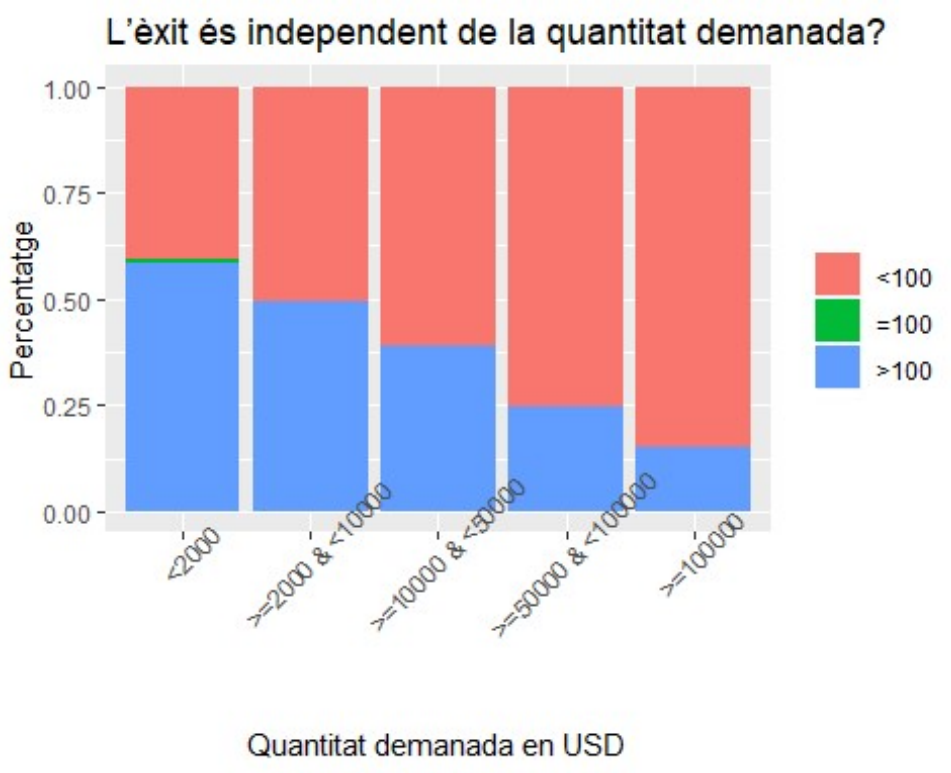
A la següent gràfica, es veu clarament que els projectes exitosos tenen el percentatge sobre el total inicial assolit al final de la campanya igual a 100 o major de 100, el que vol dir aquest percentatge és la quantitat aportada al final de la campanya és igual o superior a la quantitat demandada inicialment. Mentre que els projectes no exitosos, no arriben al 100, ja que la quantitat aportada al final no és igual o no supera la demandada inicialment.

```
# Gràfiques - Relació state vs percent_pledged
plot_by_state <- ggplot(data,aes(percent_pledged_lvl,fill=state)) +
                 geom_bar(position = "fill") +
                 labs(x="Percentatge sobre el total inicial", y="%") +
                 guides(fill=guide_legend(title="")) +
                 ggtitle("Relació state vs percent_pledged")

plot_by_state
```



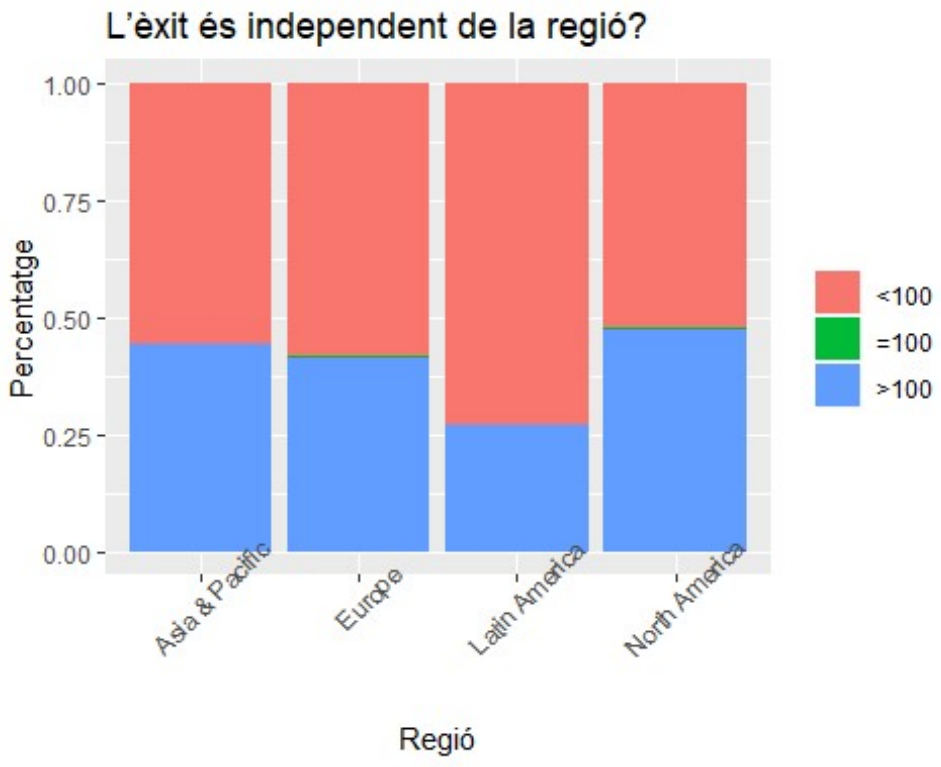
L'èxit és independent de la quantitat demanada?

```
# Gràfiques - Quantitat demanada inicialment (usd_goal_lvl)
ga <- ggplot(data, aes(usd_goal_lvl , fill=percent_pledged_lvl)) +
      geom_bar(position = "fill") +
      labs(x="Quantitat demanada en USD", y="Percentatge") +
      guides(fill=guide_legend(title="")) +
      ggtitle("L'èxit és independent de la quantitat demanada?") +
      theme(axis.text.x = element_text(angle = 45))
ga
```



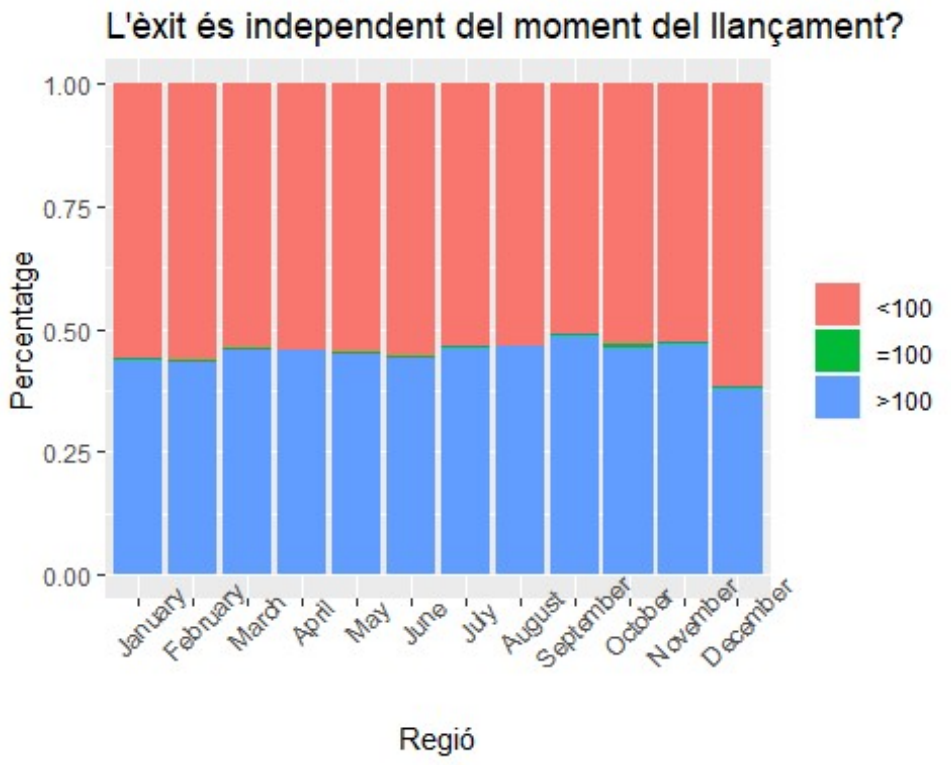
L'èxit és independent de la regió?

```
# Gràfiques - Regio
ga <- ggplot(data, aes(region , fill=percent_pledged_lvl)) +
      geom_bar(position = "fill") +
      labs(x="Regió", y="Percentatge") +
      guides(fill=guide_legend(title="")) +
      ggtitle("L'èxit és independent de la regió?") +
      theme(axis.text.x = element_text(angle = 45))
ga
```



L'èxit és independent del moment del llançament?

```
# Gràfiques - Month
ga <- ggplot(data, aes(launched , fill=percent_pledged_lvl)) +
      geom_bar(position = "fill") +
      labs(x="Regió", y="Percentatge") +
      guides(fill=guide_legend(title="")) +
      ggtitle("L'èxit és independent del moment del llançament?") +
      theme(axis.text.x = element_text(angle = 45))

ga
```



Per tant per a donar una resposta a les preguntes plantejades, podríem dir que la variable inicialment demandada influeix en l'èxit del projecte, on a partir de la gràfica, es pot veure que tal com augmenta la quantitat demanada hi ha menys percentatge de projectes exitosos.

A la segona gràfica, podem veure que la regió on menys percentatge de projectes exitosos hi ha és a Latina-Amèrica, la regió de Nord Amèrica està per damunt d'Àsia i Pacífic i d'Europa, però no és un percentatge destacable.

Consultant la tercera gràfica, es pot observar que el mes que destaca un poc més sobre la resta, és el mes de setembre.

Per tant al creador interessat del projecte, li podrem dir que seria recomanable fixar una meta inicial menor a 10000 dòlars americans, seria millor que publiqués el projecte a la regió de Nord-Amèrica i seria millor llançar el projecte al mes de setembre.

Als gràfics, pel que fa als projectes que tenen un percentatge sobre el total inicial = 100, es veu una fina línia, per tant si volem aprofundir més amb els resultats, es poden mostrar els percentatges, tal com s'indica a la següent taula:

```
# % èxit vs quantitat demanada
T_Freq_Abs_usd_goal_lvl <- table(data$usd_goal_lvl, data$percent_pledged_lvl)
T_Freq_Rel_usd_goal_lvl <- prop.table(T_Freq_Abs_usd_goal_lvl, margin = 1)
T_Freq_Rel_usd_goal_lvl
```

```
##
##                              <100         =100         >100
##   <2000              0.4056957282 0.0084936298 0.5858106420
##   >=2000 & <10000    0.5040376851 0.0020188425 0.4939434724
##   >=10000 & <50000   0.6128128502 0.0001867762 0.3870003736
##   >=50000 & <100000  0.7534403670 0.0000000000 0.2465596330
##   >=100000           0.8483606557 0.0000000000 0.1516393443
```

```
# % èxit vs regio
T_Freq_Abs_region <- table(data$region, data$percent_pledged_lvl)
T_Freq_Rel_region <- prop.table(T_Freq_Abs_region, margin = 1)
T_Freq_Rel_region
```

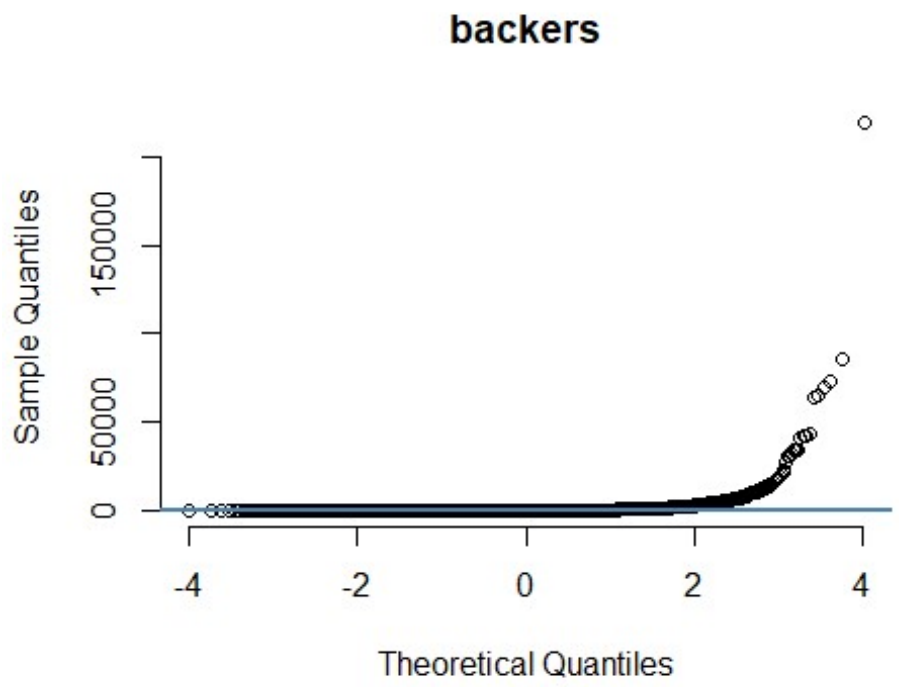
```
##
##                          <100        =100        >100
##   Asia & Pacific 0.556591212 0.002663116 0.440745672
##   Europe         0.584534431 0.003202928 0.412262640
##   Latin America  0.728155340 0.000000000 0.271844660
##   North America  0.524914384 0.002654110 0.472431507
```

```
# % èxit vs month
T_Freq_Abs_launched <- table(data$launched, data$percent_pledged_lvl)
T_Freq_Rel_launched <- prop.table(T_Freq_Abs_launched, margin = 1)
T_Freq_Rel_launched
```

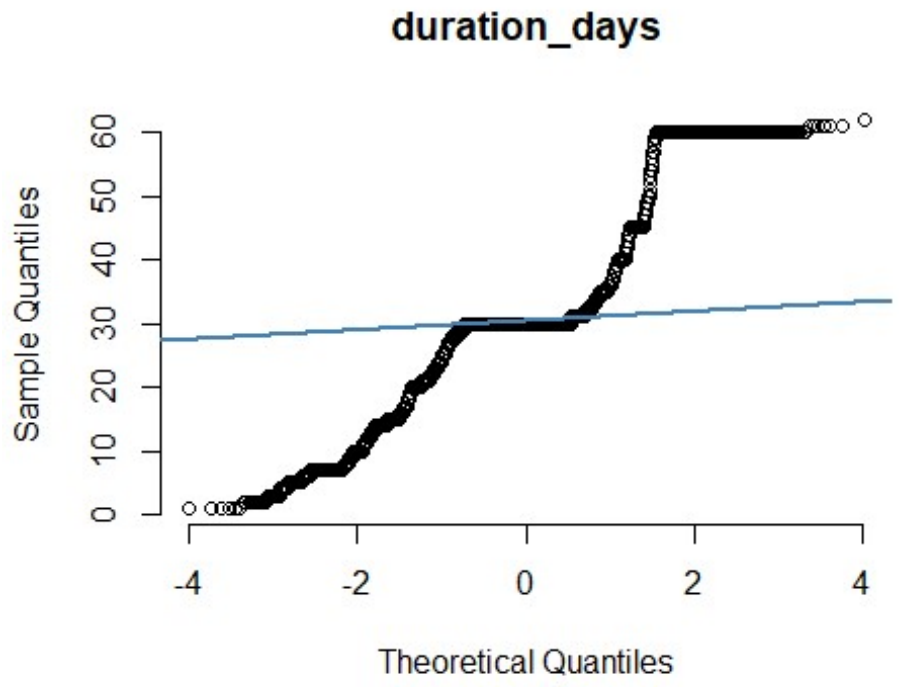
```
##
##                     <100        =100        >100
##   January   0.560198300 0.004249292 0.435552408
##   February  0.563486616 0.002059025 0.434454358
##   March     0.539712320 0.001876173 0.458411507
##   April     0.541584807 0.001309758 0.457105435
##   May       0.547341115 0.003891051 0.448767834
##   June      0.552795031 0.004140787 0.443064182
##   July      0.534065934 0.003663004 0.462271062
##   August    0.534727409 0.000000000 0.465272591
##   September 0.510190664 0.001314924 0.488494412
##   October   0.531021898 0.005474453 0.463503650
##   November  0.526843236 0.002147459 0.471009306
##   December  0.615620214 0.003062787 0.381316998
```

Gràfics normalització

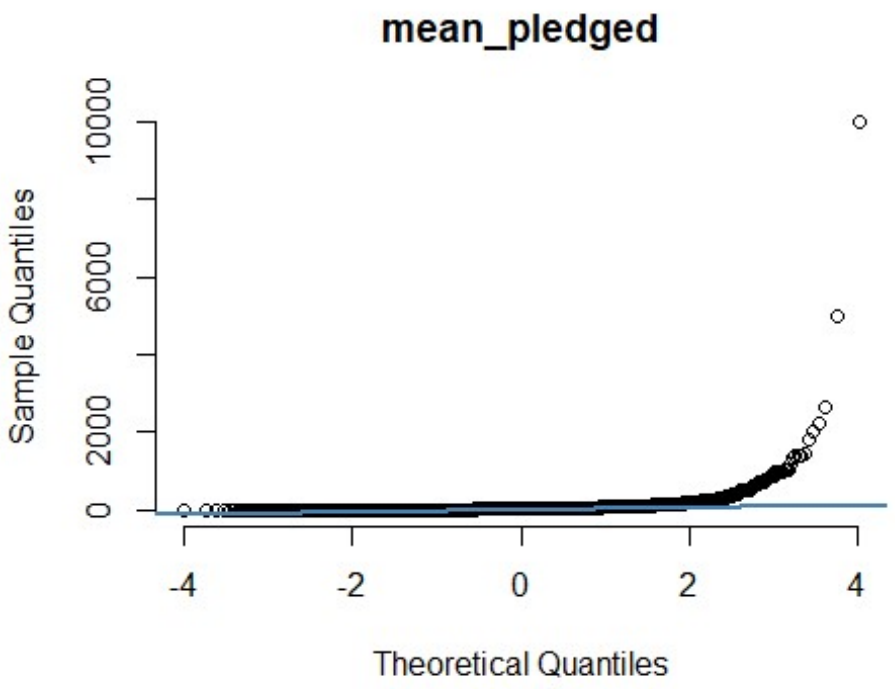
```
qqnorm(data$backers, pch = 1, frame = FALSE, main="backers")
qqline(data$backers, col = "steelblue", lwd = 2)
```



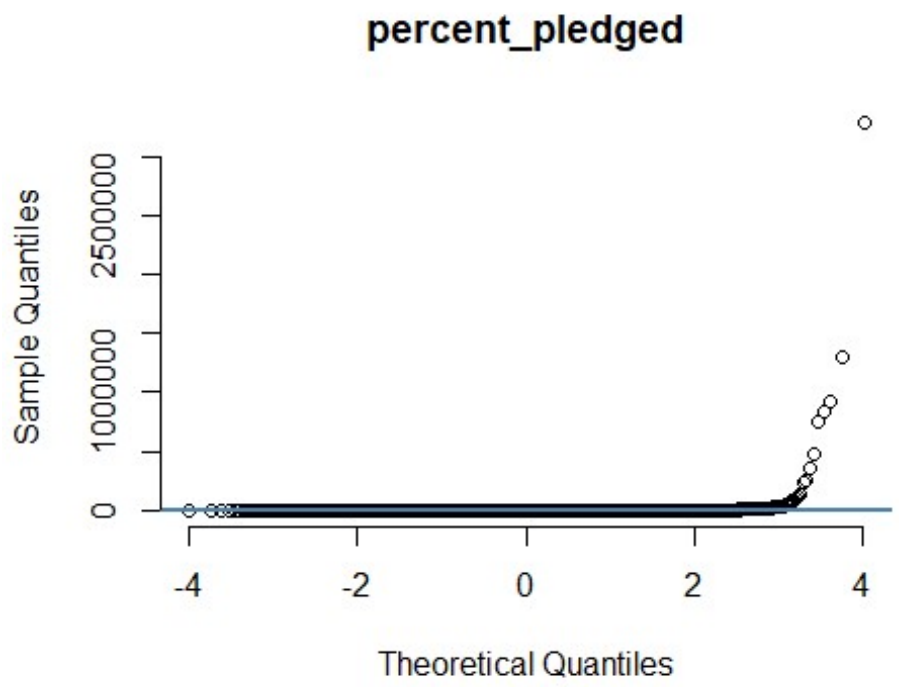

```r
qqnorm(data$duration_days, pch = 1, frame = FALSE, main="duration_days")
qqline(data$duration_days, col = "steelblue", lwd = 2)
```



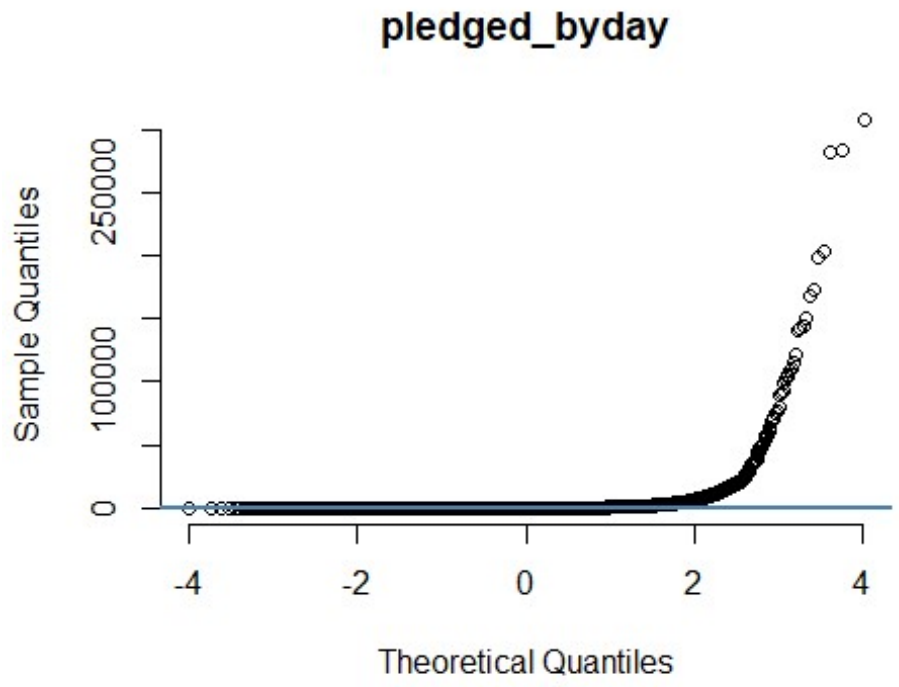
```r
qqnorm(data$mean_pledged, pch = 1, frame = FALSE, main="mean_pledged")
qqline(data$mean_pledged, col = "steelblue", lwd = 2, )
```




```r
qqnorm(data$percent_pledged, pch = 1, frame = FALSE, main="percent_pledged")
qqline(data$percent_pledged, col = "steelblue", lwd = 2)
```




```r
qqnorm(data$pledged_byday, pch = 1, frame = FALSE, main="pledged_byday")
qqline(data$pledged_byday, col = "steelblue", lwd = 2)
```



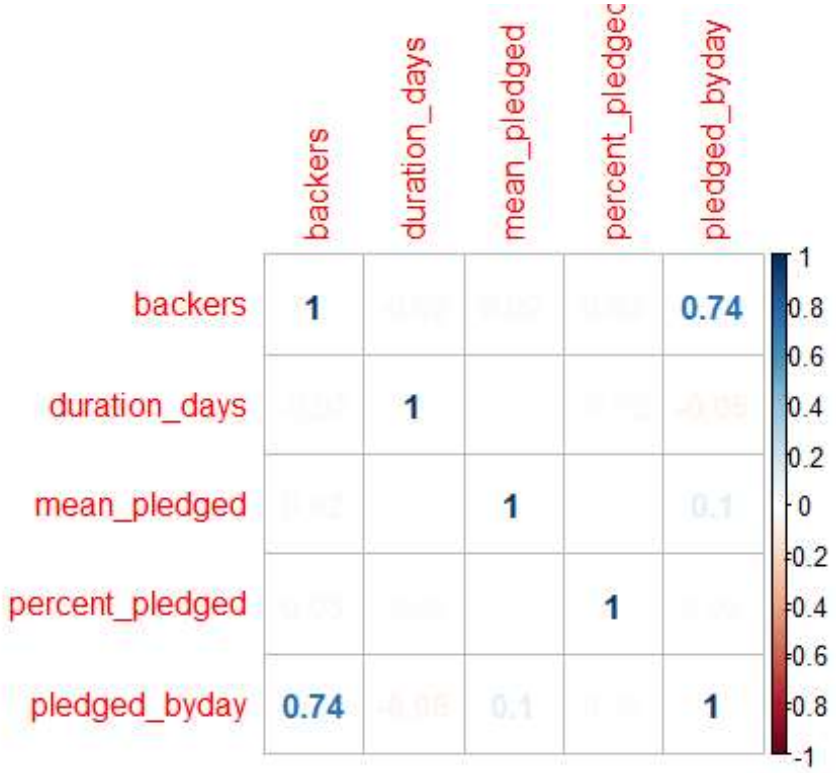
En les gràfiques que fan referència a la normalització, pel que fa a les variables quantitatives, s'observa que les dades tampoc segueixen una normalització a causa dels valors extrems, ja que es veuen punts que surten de la línia continuada de punts.

Gràfics correlació

```
cor(data[,3:7])
```

```
##                     backers duration_days mean_pledged percent_pledged
## backers          1.00000000  -0.023735822  0.024697306      0.02543862
## duration_days   -0.02373582   1.000000000 -0.004790837     -0.01902226
## mean_pledged     0.02469731  -0.004790837  1.000000000      0.00303186
## percent_pledged  0.02543862  -0.019022258  0.003031860      1.00000000
## pledged_byday    0.74451798  -0.047609369  0.100321238      0.01741098
##                  pledged_byday
## backers            0.74451798
## duration_days     -0.04760937
## mean_pledged       0.10032124
## percent_pledged    0.01741098
## pledged_byday      1.00000000
```

```
corrmatrix <- cor(data[,3:7])
corrplot(corrmatrix, method = "number")
```



A través de la gràfica de correlació, es veu clarament que hi ha una forta relació entre la variable backers i la variable pledged_byday, per tant a la pregunta com afecta el nombre de mecenes (backers) amb les quantitats aportades per persona, per dia i percentatge assolit podem dir que per a les quantitats aportades per dia té una relació considerablement forta, mentre que per a les quantitats aportades per persona i percentatge assolit, hi ha una relació positiva però és mínima.

RESOLUCIÓ DEL PROBLEMA

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

El fet de voler estudiar els casos exitosos dels projectes de la categoria de “Games” en concret, no ens permet eliminar els valors extrems, ja que si els llevarem, el conjunt de dades es veuria afectat amb un gran nombre de registres, i a banda també hem comprovat que aplicant un percentil del 0.5 arribaríem inclús a descartar els projectes amb un estat exitós, quedant-se amb un conjunt de dades del qual no podríem fer ús perquè no podríem estudiar les característiques dels projectes exitosos. En estar analitzant projectes fallits i projectes exitosos, és normal que el nombre de valors extrems siguin tan dispars, cosa que fa que les variables quantitaves no segueixin una distribució normal.

Com s’ha dit inicialment I s’ha pogut comprovar a les gràfiques I als mètodes d’anàlisis, un projecte Kickstarter es considera exitós quan l’import final de la campanya és igual o superior a l’import inicialment demandat, per tant tots els projectes que tenen un percentatge del total assolit menor a 100 seran projectes fallits. En el nostre conjunt final de dades, tots els projectes amb un percentatge del total assolit major o igual a 100 són projectes exitosos perquè hem descartat la resta d’estats, ja que els projectes en curs (state=live) o projectes que per exemple projectes que s’han considerat exitosos però finalment han sigut cancel·lats, no ens interessen perquè el que volem respondre és a les persones interessades a crear un projecte de jocs amb les característiques dels projectes considerats exitosos abans de començar.

Per característiques definim:

- Import inicial demanat
- Regió
- Mes llançament

Hem pogut respondre al creador del projecte, i li podem indicar que el millor moment per llançar un projecte entre agost I setembre, és setembre i si ho fa a la regió de Nord Amèrica, el projecte tindrà més probabilitat d’èxit. La variable més significativa a tenir en compte per a la creació del projecte, és l’import inicial demanat, on sempre serà millor crear un projecte amb un import menor a 10000 dòlars americans, ja que com major és l’import més probabilitat hi haurà que un projecte sigui fallit. Pel que fa a la normalització, ja hem dit que les dades no segueixen una distribució normal. El motiu es deu als valors extrems. El fet d’estar estudiant dos estats tan diferents on l’import final “defineix” l’estat del projecte, pot fer que hi hagi molts projectes amb valor 0 i molts altres amb imports vertaderament elevats I fer que les variables quantitatives no estiguin normalitzades ni tinguin una homogeneïtat de la variància per als grups estudiats en concret.

La variable duració no afecta gaire al resultat, però té una tendència negativa. Per tant en cas de dubte recomanaríem intentar no superar la primera meitat dels valors, centrada en 30 dies.

A banda també li podem dir al creador, que és interessant enfocar els esforços de publicitat a atreure la quantitat més gran de mecenes possible, més que a intentar que la mitjana d'aportacions per mecenes sigui alta.

Una opció haguera pogut ser seleccionar aquells projectes on l'import final fos major o igual a l'import demandat però d'aquesta manera tampoc tindrem els projectes amb un estat fallit I per tant no serviria per al nostre estudi.

CODI

Al llarg d'aquest document s'ha pogut veure el codi realitzat. A la carpeta Codi (https://github.com/csanchisp/Neteja-i-analisi-dades/tree/master/Codi) està el fitxer Rmd amb el codi de la pràctica.