

# Journey to Data Scientist : le cas Ulule

## Introduction - Business understanding

Dans la présente étude nous nous considérons comme une équipe de data scientists travaillant pour Ulule. L'objectif sera d'élaborer un modèle de machine learning permettant de prédire ou non le succès d'une campagne de crowdfunding à partir de données de la campagne; et de conseiller l'utilisateur derrière la campagne sur ce qu'il peut améliorer.

Dans la mesure où Ulule se rémunère en touchant une commission sur les projets ayant fonctionné, le site a tout intérêt à ce qu'un maximum de projets réussissent.

### Note :

Cette étude est basée principalement sur un set de données obtenu via l'API publique d'Ulule, avec l'autorisation du site par e-mail. Une vérification du set sera effectuée afin de ne pas traiter de données personnelles.

## Vérification du set de données - Data understanding

Dans la mesure où certains projets peuvent ou non avoir une vidéo de présentation, il est exclu de retirer toute ligne contenant un "NaN" (représentant un vide). On se contente donc de retirer les doublons et les colonnes constantes, dans un premier temps.

### Première analyse du set

Le set contient près de 50000 lignes correspondant à des projets, réussis ou non, et 96 colonnes contenant différents éléments comme le montant levé ou la description du projet dans différentes langues.

Les colonnes peuvent être groupées en quatre catégories :

- les données construites par Ulule (comme des listes d'urls)
- les données obsolètes ou constantes et qui seront retirées
- les données liées au projet (avant le lancement)
- les données liées à la campagne (après le lancement)

### Données construites par Ulule

Ces données sont indépendantes du possesseur du projet (l'utilisateur que nous cherchons à conseiller) et **ne seront donc pas utilisées dans cette étude**.

- `absolute_url`
- `discussion_thread_id`
- `id`
- `resource_uri`
- `slug`
- `urls`
- `user_role`

L'id du projet sera conservé pour disposer d'une variable indépendante du projet et simple à représenter, en abscisse notamment.

### Données obsolètes ou inutiles

Ces données proviennent d'anciennes versions de l'API ou sont constantes quelque soit le projet (dans ce data set) et sont donc à retirer.

- ~~address\_required~~
- ~~permissions~~
- ~~phone\_number\_required~~
- ~~required\_personal\_id\_number~~
- ~~image~~
- ~~status~~
- ~~is\_in\_extra\_time~~

### Données de la campagne

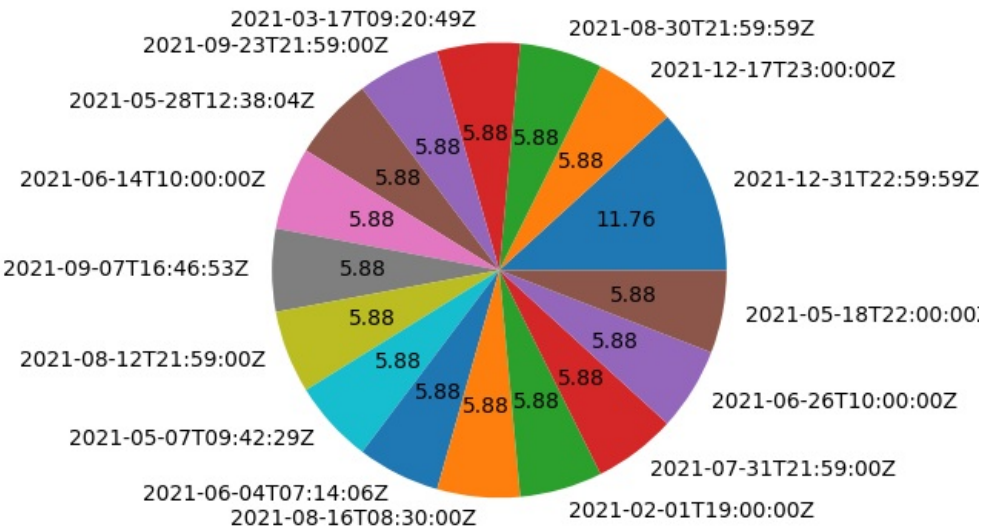
Ces données concernent le projet après son lancement.

- ~~amount\_raised~~
- ~~comments\_count~~
- ~~committed~~
- ~~date\_end~~
- ~~date\_end\_extra\_time~~
- ~~date\_goal\_raised~~
- ~~date\_start~~
- ~~fans\_count~~
- ~~finished~~
- ~~is\_cancelled~~
- ~~is\_in\_extra\_time~~
- ~~lowest\_contribution\_amount~~
- ~~nb\_days~~
- ~~nb\_products\_sold~~
- ~~news\_count~~
- ~~orders\_count~~
- ~~percent~~
- ~~sponsorships\_count~~
- ~~supporters\_count~~
- ~~time\_left~~
- ~~time\_left\_short~~

Afin de ne pas biaiser notre modèle, nous ne nous intéresserons pas aux projets encore en cours. Les variables **time\_left**, **time\_left\_short**, **is\_in\_extra\_time** ainsi que **finished** (après le retrait des projets inachevés) ne sont donc pas pertinentes. De même, les projets annulés doivent être retirés, ainsi que la colonne **is\_cancelled**.

date\_end\_extra\_time

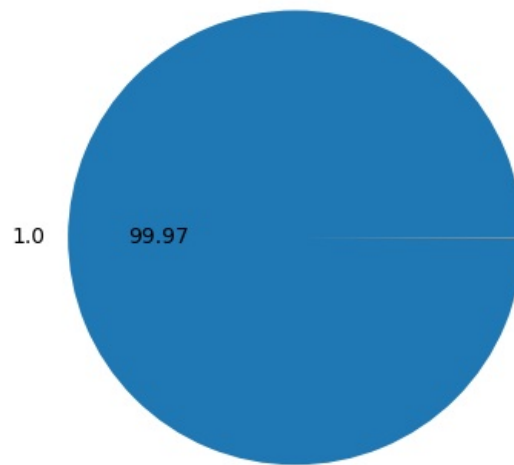
Extension de la durée de la campagne



La colonne date\_end\_extra\_time sera retirée car aucun projet ayant échoué n'y a fait appel et c'est un phénomène très minoritaire.

lowest\_contribution\_amount

## Répartition de la contribution minimale au sein des projets



Etant quasiment constante, la colonne **lowest\_contribution\_amount** peut également être retirée car non pertinente.

### **committed**

La colonne committed concerne les promesses faites par les supporters. Il y a deux cas de figure :

- Le projet est une campagne classique et les supporters promettent de l'argent (**amount\_raised**) pour atteindre un objectif (**goal**). Dans ce cas, **committed** est strictement égal à **amount\_raised**.
- Le projet est une prévente, les supporters promettent d'acheter un nombre de produits (**nb\_products\_sold**) pour atteindre un objectif de vente (**goal**). Dans ce cas, **committed** est strictement égal à **nb\_products\_sold**.

En conclusion, **committed** peut être retirée car inutile.

### **Données du projet**

Ces données concernent le projet avant son lancement.

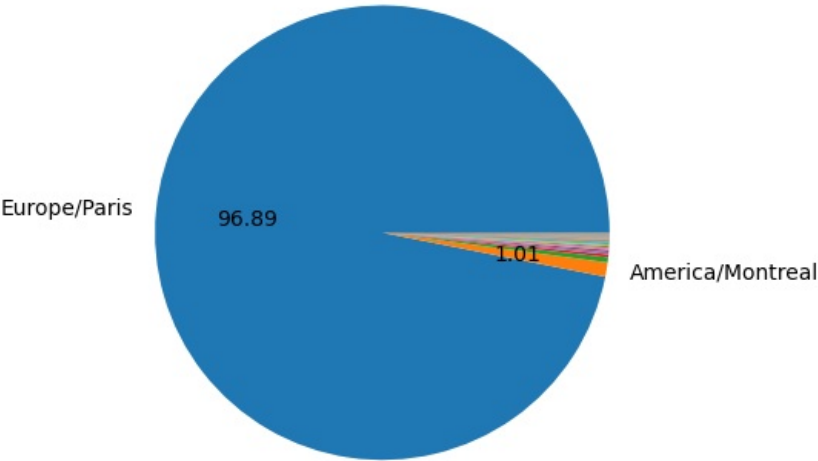
- analytics\_count
- background
- comments\_enabled
- country
- currency
- currency\_display
- delivery
- description\_ca
- description\_de
- description\_en
- description\_es
- description\_fr
- description\_it
- description\_nl
- description\_pt
- description\_funding\_ca
- description\_funding\_de
- description\_funding\_en
- description\_funding\_es
- description\_funding\_fr
- description\_funding\_it
- description\_funding\_nl
- description\_funding\_pt
- goal
- goal\_raised
- image
- lang
- location
- main\_image
- main\_tag
- name\_ca

- name\_de
- name\_en
- name\_es
- name\_fr
- name\_it
- name\_nl
- name\_pt
- owner
- payment\_methods
- rewards
- sponsorships\_count
- subtitle\_ca
- subtitle\_de
- subtitle\_en
- subtitle\_es
- subtitle\_fr
- subtitle\_it
- subtitle\_nl
- subtitle\_pt
- visible
- video
- type
- ~~timezone~~

Il ne nous a pas semblé pertinent de garder la colonne **delivery** car elle peut ne pas avoir de sens si le projet n'offre pas de récompense physique (comme un jeu vidéo ou un film).

**timezone**

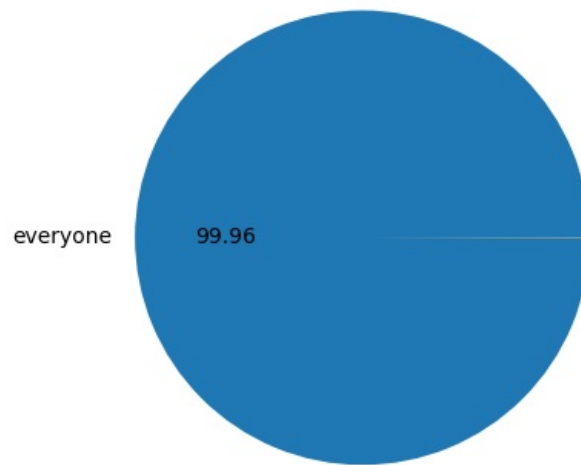
Répartition de la zone temporelle au sein des projets



L'immense majorité des projets a lieu dans la même zone, la colonne **timezone** est quasiment constante, elle peut être retirée.

**comments\_enabled**

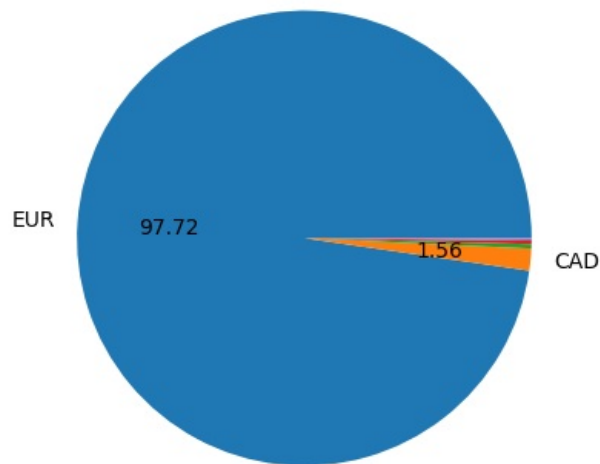
## Répartition des permissions de commentaires



Une écrasante majorité des projets autorise les commentaires pour tous les utilisateurs, la colonne **comments\_enabled** n'est donc pas pertinente.

## currency

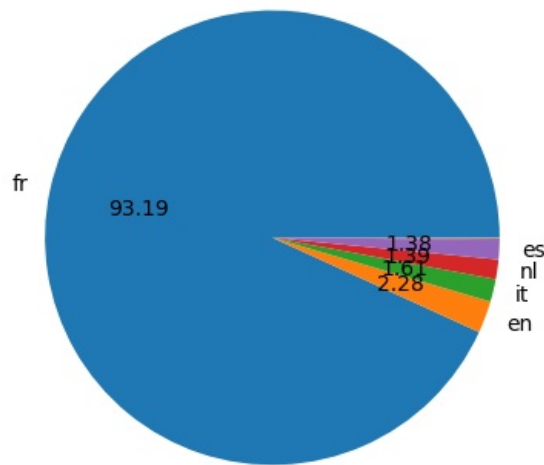
### Répartition de la monnaie utilisée au sein des projets



L'écrasante majorité des projets est en euro, il est donc possible de retirer la colonne **currency** ainsi que la colonne **currency\_display**, sans oublier les projets concernés.

## lang

## Répartition des langues au sein des projets



Les autres langues que le français étant très minoritaires, on peut retirer tous les projets concernés ainsi que les colonnes suivantes :

- **description\_[Langue!=fr]**
- **description\_funding\_[Langue!=fr]**
- **lang**
- **name\_[Langue!=fr]**
- **subtitle\_[Langue!=fr]**

### Bilan : colonnes restantes

Les colonnes suivantes sont conservées dans le dataset, mais peuvent nécessiter un travail supplémentaire, comme la colonne **video**. Nous n'allons en effet pas étudier la vidéo du projet en elle même mais plutôt le fait qu'elle existe ou non par exemple.

Le set contient une trentaine de colonnes pour environs 40.000 projets.

```
Index(['amount_raised', 'analytics_count', 'background', 'comments_count', 'date_end', 'date_goal_raised', 'date_start', 'description_fr', 'description_funding_fr', 'description_yourself_fr', 'fans_count', 'goal', 'goal_raised', 'id', 'location', 'main_tag', 'name_fr', 'nb_days', 'nb_products_sold', 'news_count', 'owner', 'payment_methods', 'percent', 'rewards', 'sponsorships_count', 'subtitle_fr', 'supporters_count', 'type', 'video', 'visible'], dtype='object')
```

Certaines colonnes doivent être binarisée pour représenter ou non la présence d'un objet (comme une vidéo).

Binarisation de **video** et **background**

**La colonne location contient un dictionnaire avec plusieurs attributs. On choisit de ne garder que la ville.**

**owner**

La colonne **owner** est inutilisable en tant que telle car seules les stats **anonymisées et concernant l'activité publique de lancement de projet** de l'owner nous intéressent.

**rewards**

Pour chaque projet, l'attribut reward propose un certain nombre de rewards dans une liste. Pour chaque reward, plusieurs informations sont disponibles, comme une date de livraison, un nombre de stock etc. Il est possible pour une reward d'avoir plusieurs variantes, par exemple une couleur pour un T-shirt, localisé dans l'attribut 'variants'.

Les stocks seront toujours nuls (les projets sont finis) mais il est possible de savoir combien de chacune des rewards ont été prises, et à quel prix. Il est donc possible de voir, pour un projet, ce qui a été le plus rentable i.e. plein de petites rewards ou peu de grosses; et de croiser avec tous les autres projets.

La colonne doit donc être retravaillée pour extraire une liste de dictionnaires par projet.

**main\_tag**

Dans la mesure où les projets ne se comportent pas de la même façon selon leur type, il peut être intéressant d'étudier les tags utilisés pour les décrire. Seuls nous intéressent l'id et le nom en français du tag, il faut donc les extraire.

## Retrait de lignes incomplètes

- Retrait de 38 lignes n'ayant aucune valeur dans la colonne **date\_start**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **date\_end**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **amount\_raised**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **comments\_count**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **date\_start**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **date\_end**
- Retrait de 2 lignes n'ayant aucune valeur dans la colonne **description\_fr**
- Retrait de 714 lignes n'ayant aucune valeur dans la colonne **description\_funding\_fr**
- Retrait de 399 lignes n'ayant aucune valeur dans la colonne **description\_yourself\_fr**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **fans\_count**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **goal**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **goal\_raised**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **id**
- Retrait de 53 lignes n'ayant aucune valeur dans la colonne **main\_tag**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **name\_fr**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **news\_count**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **owner**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **percent**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **rewards**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **sponsorships\_count**
- Retrait de 2 lignes n'ayant aucune valeur dans la colonne **subtitle\_fr**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **supporters\_count**

## nb\_days

La colonne "nb\_days" contient un tiers de valeurs vides, il faut la compléter.

## news\_per\_days

Création de la colonne news\_per\_days

## nb\_rewards

Création de la colonne nb\_rewards

## post\_covid

Nous allons étudier l'influence du COVID-19 sur les campagnes Ulule donc il est intéressant de rajouter une colonne 'post\_covid' indiquant si le projet prend fin après le mois de mars 2020. Création de la colonne "post\_covid"

## type

Les projets fonctionnent différemment selon qu'ils concernent des préventes ou une financement. Il convient donc de séparer le set en deux sous-sets.

## nb\_products\_sold

Retrait de la colonne **nb\_product\_sold** pour les projets n'étant pas sous la forme d'une prévente, car cette colonne est équivalente à la colonne **supporters\_count**.



# Journey to Data Scientist : le cas Ulule

## Introduction - Business understanding

Dans la présente étude nous nous considérons comme une équipe de data scientists travaillant pour Ulule. L'objectif sera d'élaborer un modèle de machine learning permettant de prédire ou non le succès d'une campagne de crowdfunding à partir de données de la campagne; et de conseiller l'utilisateur derrière la campagne sur ce qu'il peut améliorer.

Dans la mesure où Ulule se rémunère en touchant une commission sur les projets ayant fonctionné, le site a tout intérêt à ce qu'un maximum de projets réussissent.

**Note :**

Cette étude est basée principalement sur un set de données obtenu via l'API publique d'Ulule, avec l'autorisation du site par e-mail. Une vérification du set sera effectuée afin de ne pas traiter de données personnelles.

## Vérification du set de données - Data understanding

Dans la mesure où certains projets peuvent ou non avoir une vidéo de présentation, il est exclu de retirer toute ligne contenant un "NaN" (représentant un vide). On se contente donc de retirer les doublons et les colonnes constantes, dans un premier temps.

### Première analyse du set

Le set contient près de 50000 lignes correspondant à des projets, réussis ou non, et 96 colonnes contenant différents éléments comme le montant levé ou la description du projet dans différentes langues.

Les colonnes peuvent être groupées en quatre catégories :

- les données construites par Ulule (comme des listes d'urls)
- les données obsolètes ou constantes et qui seront retirées
- les données liées au projet (avant le lancement)
- les données liées à la campagne (après le lancement)

#### Données construites par Ulule

Ces données sont indépendantes du possesseur du projet (l'utilisateur que nous cherchons à conseiller) et **ne seront donc pas utilisées dans cette étude**.

- absolute\_url
- discussion\_thread\_id
- id
- resource\_uri
- slug
- urls
- user\_role

L'id du projet sera conservé pour disposer d'une variable indépendante du projet et simple à représenter, en abscisse notamment.

#### Données obsolètes ou inutiles

Ces données proviennent d'anciennes versions de l'API ou sont constantes quelque soit le projet (dans ce data set) et sont donc à retirer.

- address\_required
- permissions
- phone\_number\_required
- required\_personal\_id\_number
- image
- status
- is\_in\_extra\_time

#### Données de la campagne

Ces données concernent le projet après son lancement.

- amount\_raised
- comments\_count
- committed
- date\_end
- date\_end\_extra\_time
- date\_goal\_raised
- date\_start
- fans\_count
- finished
- is\_cancelled

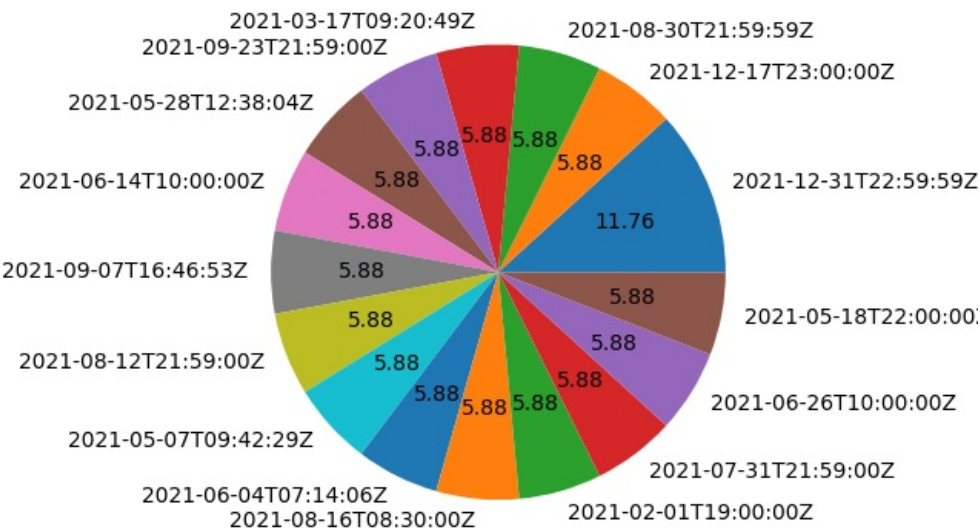


- is\_in\_extra\_time
- lowest\_contribution\_amount
- nb\_days
- nb\_products\_sold
- news\_count
- orders\_count
- percent
- sponsorships\_count
- supporters\_count
- time\_left
- time\_left\_short

Afin de ne pas biaiser notre modèle, nous ne nous intéresserons pas aux projets encore en cours. Les variables **time\_left**, **time\_left\_short**, **is\_in\_extra\_time** ainsi que **finished** (après le retrait des projets inachevés) ne sont donc pas pertinentes. De même, les projets annulés doivent être retirés, ainsi que la colonne **is\_cancelled**.

date\_end\_extra\_time

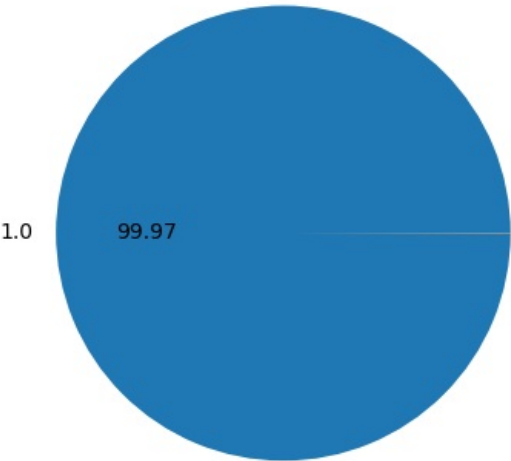
Extension de la durée de la campagne



La colonne date\_end\_extra\_time sera retirée car aucun projet ayant échoué n'y a fait appel et c'est un phénomène très minoritaire.

lowest\_contribution\_amount

Répartition de la contribution minimale au sein des projets



Etant quasiment constante, la colonne **lowest\_contribution\_amount** peut également être retirée car non pertinente.

**committed**

La colonne committed concerne les promesses faites par les supporters. Il y a deux cas de figure :

- Le projet est une campagne classique et les supporters promettent de l'argent (**amount\_raised**) pour atteindre un objectif (**goal**). Dans ce cas, **committed** est strictement égal à **amount\_raised**.
- Le projet est une prévente, les supporters promettent d'acheter un nombre de produits (**nb\_products\_sold**) pour atteindre un objectif de vente (**goal**). Dans ce cas, **committed** est strictement égal à **nb\_products\_sold**.

En conclusion, **committed** peut être retirée car inutile.

**Données du projet**

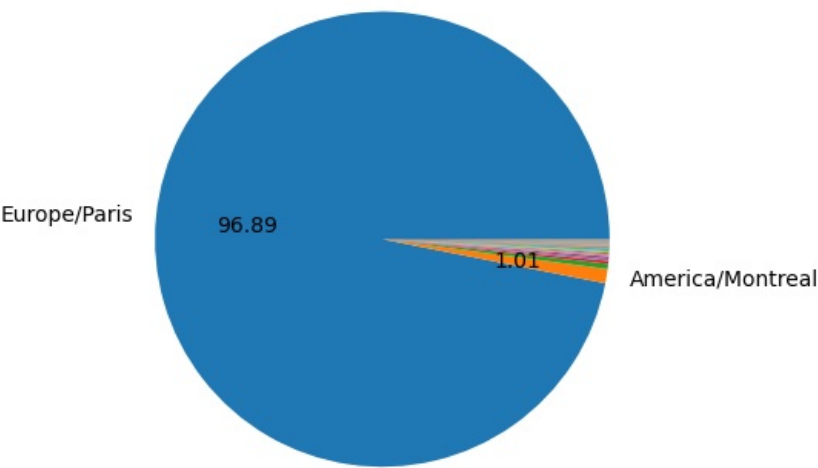
Ces données concernent le projet avant son lancement.

- analytics\_count
- background
- comments\_enabled
- country
- currency
- currency\_display
- delivery
- description\_ca
- description\_de
- description\_en
- description\_es
- description\_fr
- description\_it
- description\_nl
- description\_pt
- description\_funding\_ca
- description\_funding\_de
- description\_funding\_en
- description\_funding\_es
- description\_funding\_fr
- description\_funding\_it
- description\_funding\_nl
- description\_funding\_pt
- goal
- goal\_raised
- image
- lang
- location
- main\_image
- main\_tag
- name\_ca
- name\_de
- name\_en
- name\_es
- name\_fr
- name\_it
- name\_nl
- name\_pt
- owner
- payment\_methods
- rewards
- sponsorships\_count
- subtitle\_ca
- subtitle\_de
- subtitle\_en
- subtitle\_es
- subtitle\_fr
- subtitle\_it
- subtitle\_nl
- subtitle\_pt
- visible
- video
- type
- timezone

Il ne nous a pas semblé pertinent de garder la colonne **delivery** car elle peut ne pas avoir de sens si le projet n'offre pas de récompense physique (comme un jeu vidéo ou un film).

timezone

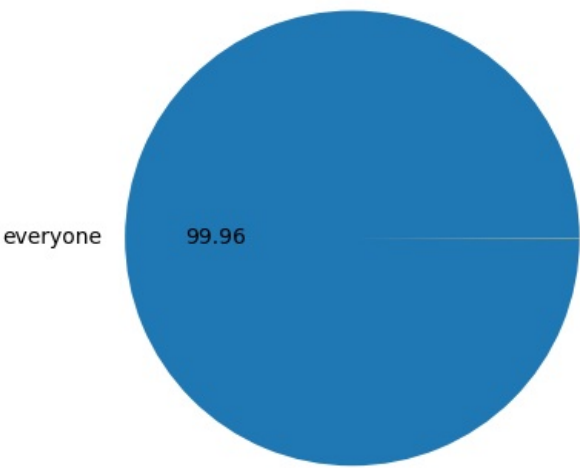
Répartition de la zone temporelle au sein des projets



L'immense majorité des projets a lieu dans la même zone, la colonne **timezone** est quasiment constante, elle peut être retirée.

comments\_enabled

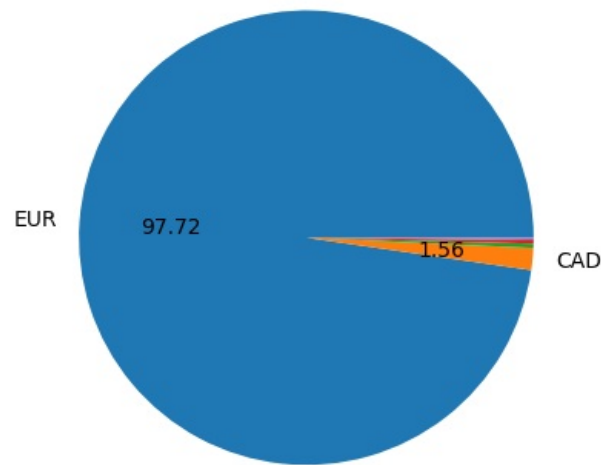
Répartition des permissions de commentaires



Une écrasante majorité des projets autorise les commentaires pour tous les utilisateurs, la colonne **comments\_enabled** n'est donc pas pertinente.

currency

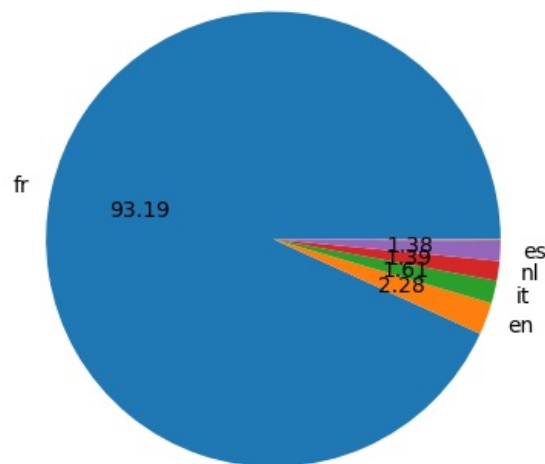
## Répartition de la monnaie utilisée au sein des projets



L'écrasante majorité des projets est en euro, il est donc possible de retirer la colonne **currency** ainsi que la colonne **currency\_display**, sans oublier les projets concernés.

**lang**

## Répartition des langues au sein des projets



Les autres langues que le français étant très minoritaires, on peut retirer tous les projets concernés ainsi que les colonnes suivantes :

- **description\_[Langue!=fr]**
- **description\_funding\_[Langue!=fr]**
- **lang**
- **name\_[Langue!=fr]**
- **subtitle\_[Langue!=fr]**

## Bilan : colonnes restantes

Les colonnes suivantes sont conservées dans le dataset, mais peuvent nécessiter un travail supplémentaire, comme la colonne **video**. Nous n'allons en effet pas étudier la vidéo du projet en elle même mais plutôt le fait qu'elle existe ou non par exemple.

Le set contient une trentaine de colonnes pour environ 40.000 projets.

Index(['amount\_raised', 'analytics\_count', 'background', 'comments\_count', 'date\_end', 'date\_goal\_raised', 'date\_start', 'description\_fr', 'description\_funding\_fr', 'description\_yourself\_fr', 'fans\_count', 'goal', 'goal\_raised', 'id', 'location', 'main\_tag', 'name\_fr', 'nb\_days', 'nb\_products\_sold', 'news\_count', 'owner', 'payment\_methods', 'percent', 'rewards', 'sponsorships\_count', 'subtitle\_fr', 'supporters\_count', 'type', 'video', 'visible'], dtype='object')

Certaines colonnes doivent être binarisée pour représenter ou non la présence d'un objet (comme une vidéo).  
Binarisation de **video** et **background**

La colonne **location** contient un dictionnaire avec plusieurs attributs. On choisit de ne garder que la ville.

**owner**

La colonne **owner** est inutilisable en tant que telle car seules les stats **anonymisées et concernant l'activité publique de lancement de projet** de l'owner nous intéressent.

**rewards**

Pour chaque projet, l'attribut reward propose un certain nombre de rewards dans une liste. Pour chaque reward, plusieurs informations sont disponibles, comme une date de livraison, un nombre de stock etc. Il est possible pour une reward d'avoir plusieurs variantes, par exemple une couleur pour un T-shirt, localisé dans l'attribut 'variants'.

Les stocks seront toujours nuls (les projets sont finis) mais il est possible de savoir combien de chacune des rewards ont été prises, et à quel prix. Il est donc possible de voir, pour un projet, ce qui a été le plus rentable i.e. plein de petites rewards ou peu de grosses; et de croiser avec tous les autres projets.

La colonne doit donc être retravaillée pour extraire une liste de dictionnaires par projet.

**main\_tag**

Dans la mesure où les projets ne se comportent pas de la même façon selon leur type, il peut être intéressant d'étudier les tags utilisés pour les décrire. Seuls nous intéressent l'id et le nom en français du tag, il faut donc les extraire.

Retrait de lignes incomplètes

- Retrait de 38 lignes n'ayant aucune valeur dans la colonne **date\_start**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **date\_end**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **amount\_raised**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **comments\_count**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **date\_start**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **date\_end**
- Retrait de 2 lignes n'ayant aucune valeur dans la colonne **description\_fr**
- Retrait de 714 lignes n'ayant aucune valeur dans la colonne **description\_funding\_fr**
- Retrait de 399 lignes n'ayant aucune valeur dans la colonne **description\_yourself\_fr**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **fans\_count**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **goal**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **goal\_raised**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **id**
- Retrait de 53 lignes n'ayant aucune valeur dans la colonne **main\_tag**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **name\_fr**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **news\_count**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **owner**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **percent**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **rewards**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **sponsorships\_count**
- Retrait de 2 lignes n'ayant aucune valeur dans la colonne **subtitle\_fr**
- Retrait de 0 lignes n'ayant aucune valeur dans la colonne **supporters\_count**

**nb\_days**

La colonne "nb\_days" contient un tiers de valeurs vides, il faut la compléter.

**news\_per\_days**

Création de la colonne news\_per\_days

**nb\_rewards**

Création de la colonne nb\_rewards

**post\_covid**

Nous allons étudier l'influence du COVID-19 sur les campagnes Ulule donc il est intéressant de rajouter une colonne

'post\_covid' indiquant si le projet prend fin après le mois de mars 2020. Création de la colonne "post\_covid"

## type

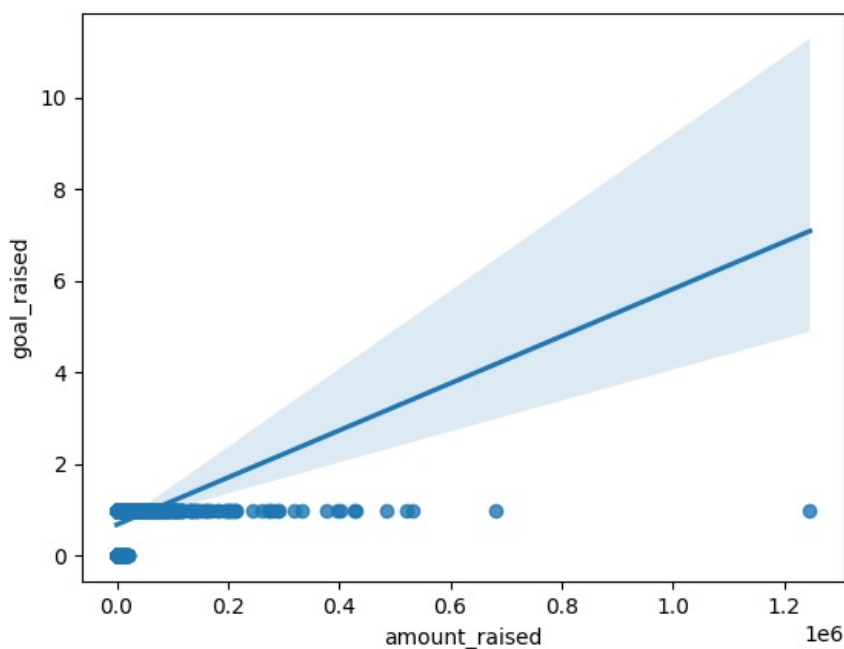
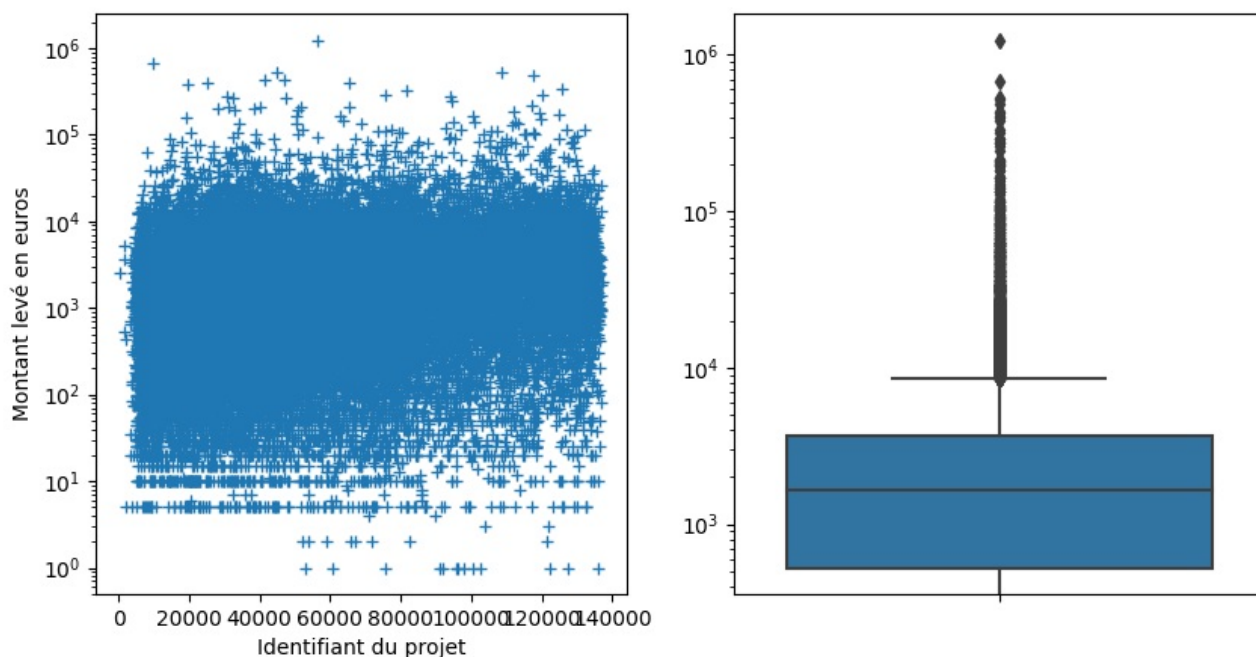
Les projets fonctionnent différemment selon qu'ils concernent des préventes ou une financement. Il convient donc de séparer le set en deux sous-sets.

## nb\_products\_sold

Retrait de la colonne **nb\_product\_sold** pour les projets n'étant pas sous la forme d'une prévente, car cette colonne est équivalente à la colonne **supporters\_count**.

# Statistiques descriptives

Montant levé par projet

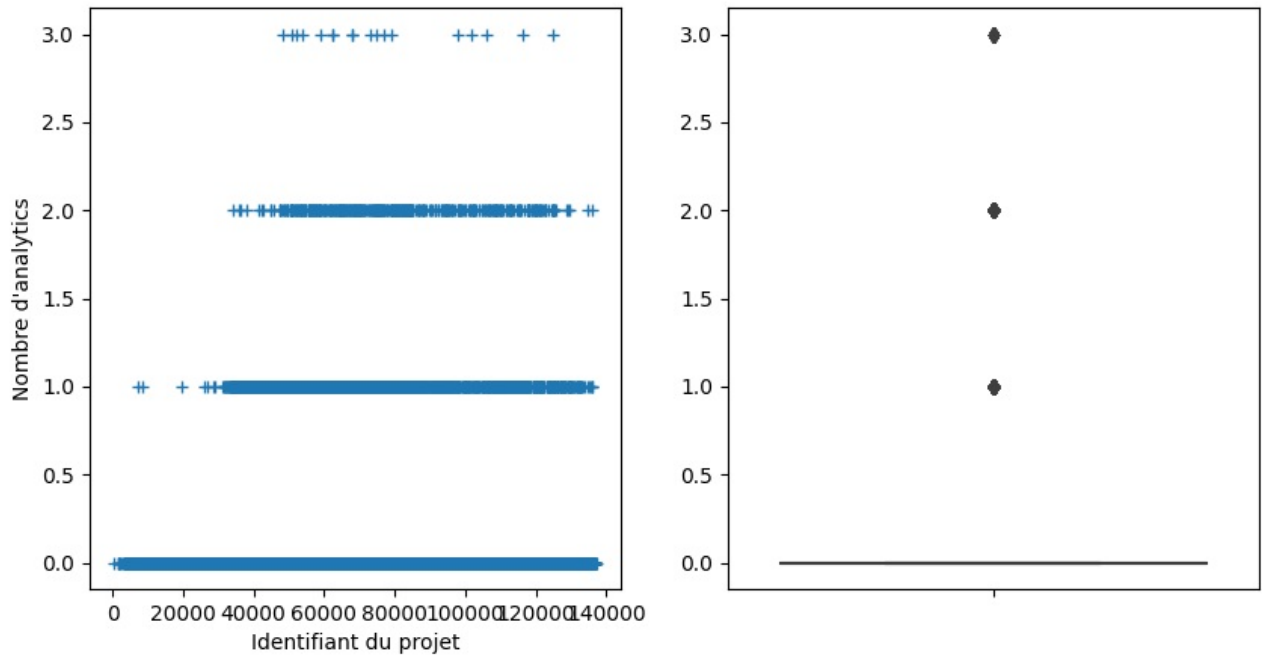


SpearmanrResult(correlation=0.6605276041694639, pvalue=0.0)

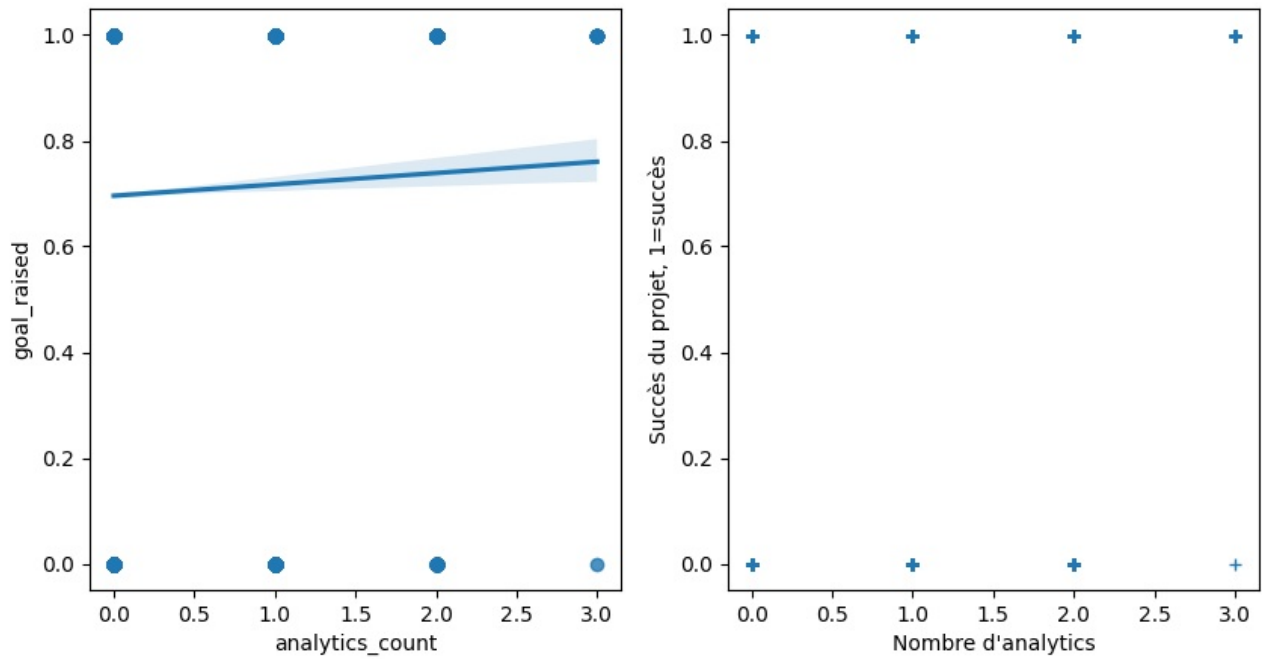
Les projets semblent assez homogènes dans les montants levés même si moins nombreux pour les plus hauts montants. Il semble y avoir une bonne corrélation entre le montant obtenu et le succès de la campagne (peu surprenant).

## analytics\_count

Nombre d'analytics par projet



Succès du projet en fonction du nombre d'analytics

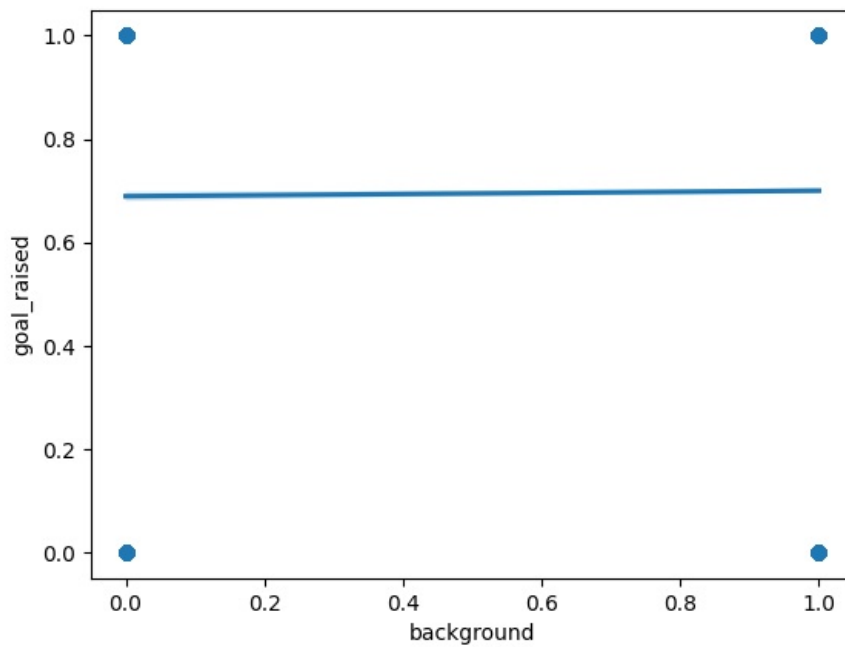
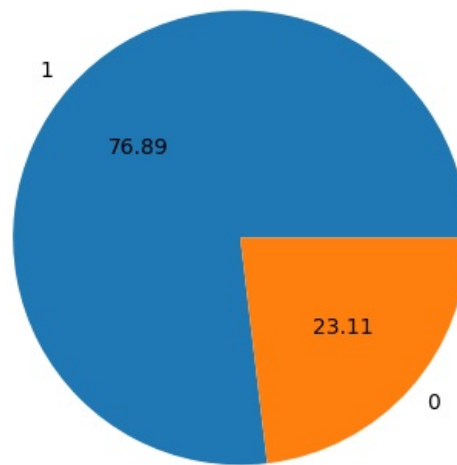


`SpearmanrResult(correlation=0.00653046420759979, pvalue=0.20824585283482497)`

On constate qu'il n'y a pas de corrélation entre le nombre d'analytics et la réussite d'un projet

**background**

## Présence d'un background dans le projet



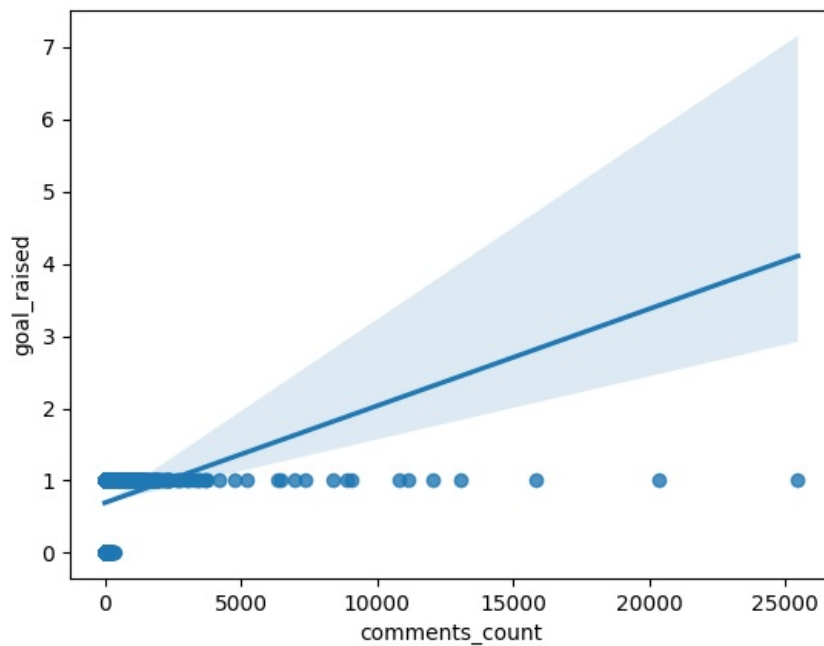
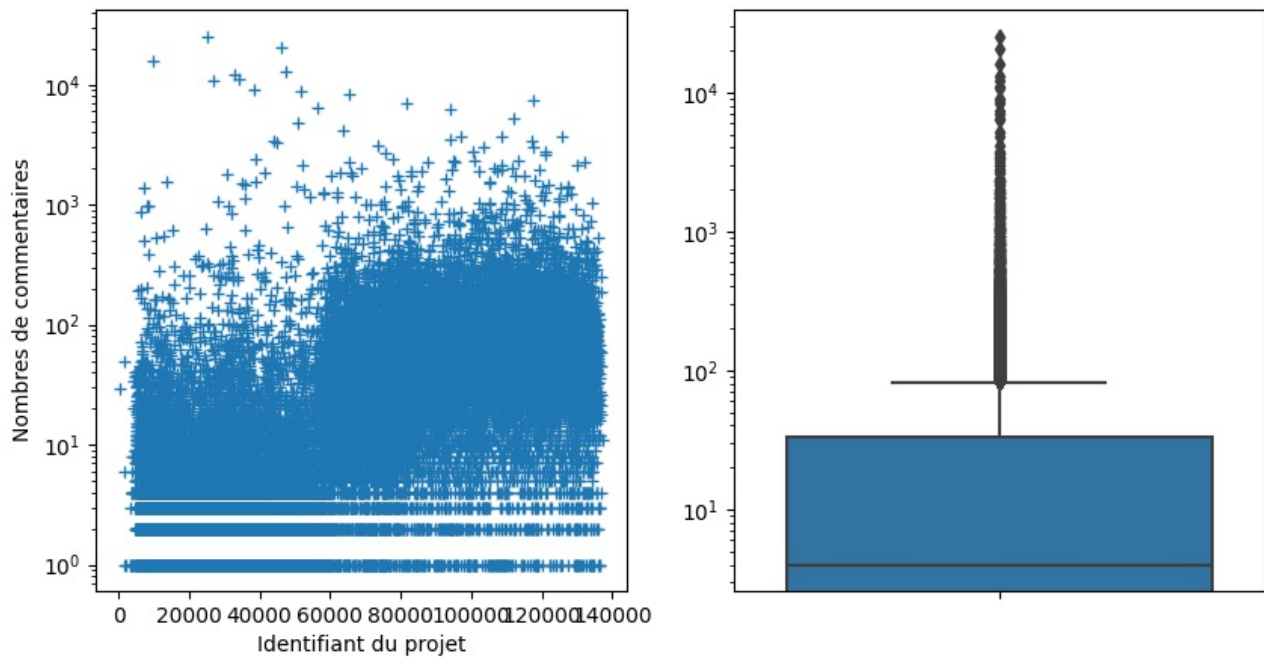
SpearmanrResult(correlation=0.009848313025372124, pvalue=0.057726360936618135)

Il n'y a pas de corrélation entre la présence d'un background et le succès de la campagne.

**comments\_count**



Nombre de commentaires par projet

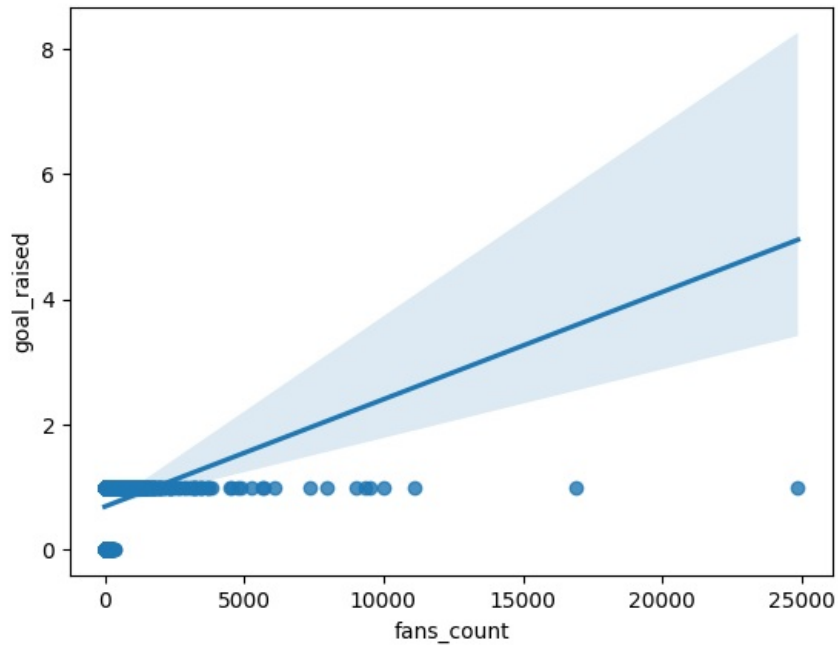
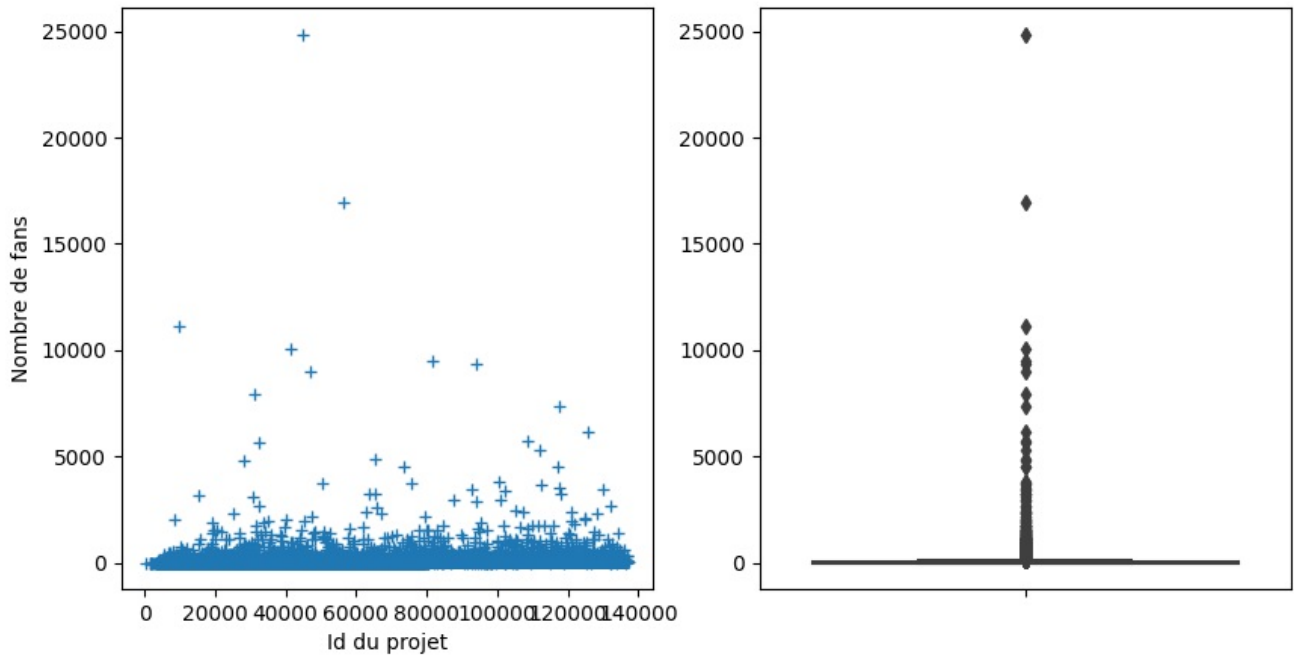


SpearmanrResult(correlation=0.3442342916010207, pvalue=0.0)

Les projets reçoivent globalement assez peu de commentaires. La plupart des projets qui ont beaucoup de commentaires, ont bien fonctionné. Il semble y avoir une petite corrélation entre le nombre de commentaires et le succès de la campagne.

**fans\_count**

## Représentation du nombre de fans

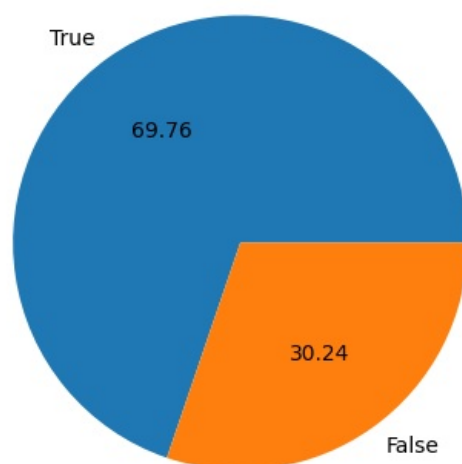


`SpearmanrResult(correlation=0.2378731377730896, pvalue=0.0)`

La présence de fans semble être un phénomène très minoritaire, il serait intéressant de vérifier si les projets suivis ont été plus réussis que les autres. Il est possible de conclure quand à l'existence d'une corrélation entre le succès de la campagne et le nombre de fans.

**goal\_raised**

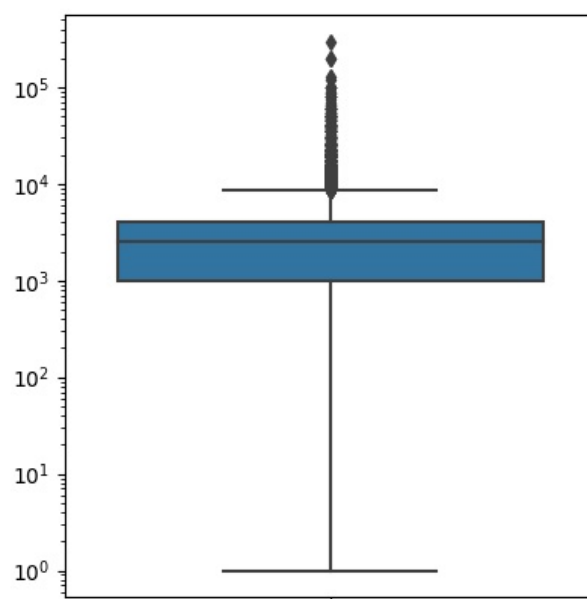
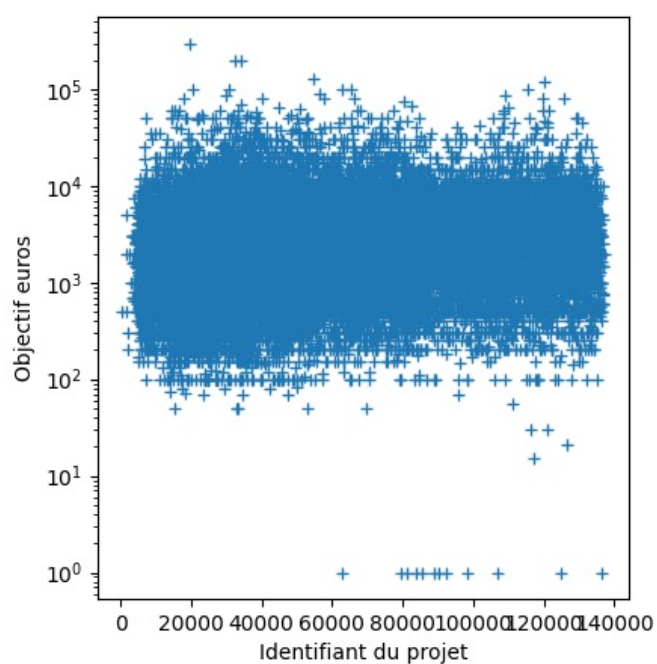
## Représentation du taux de succès des projets, succès=True

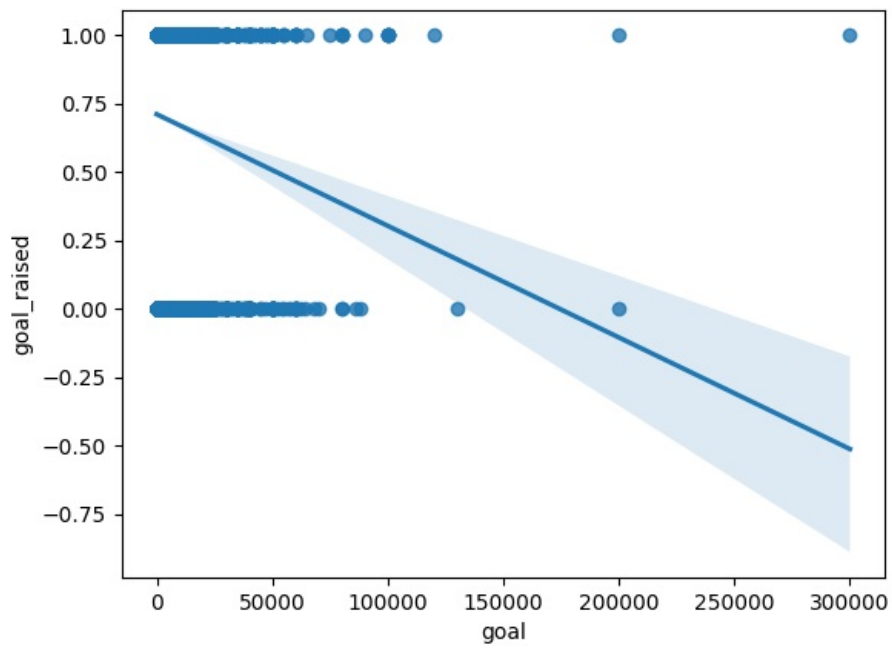


Pour rappel, le taux de succès des projets de la plateforme en 2021 est de 79%.

**goal**

## Objectif de validation du projet

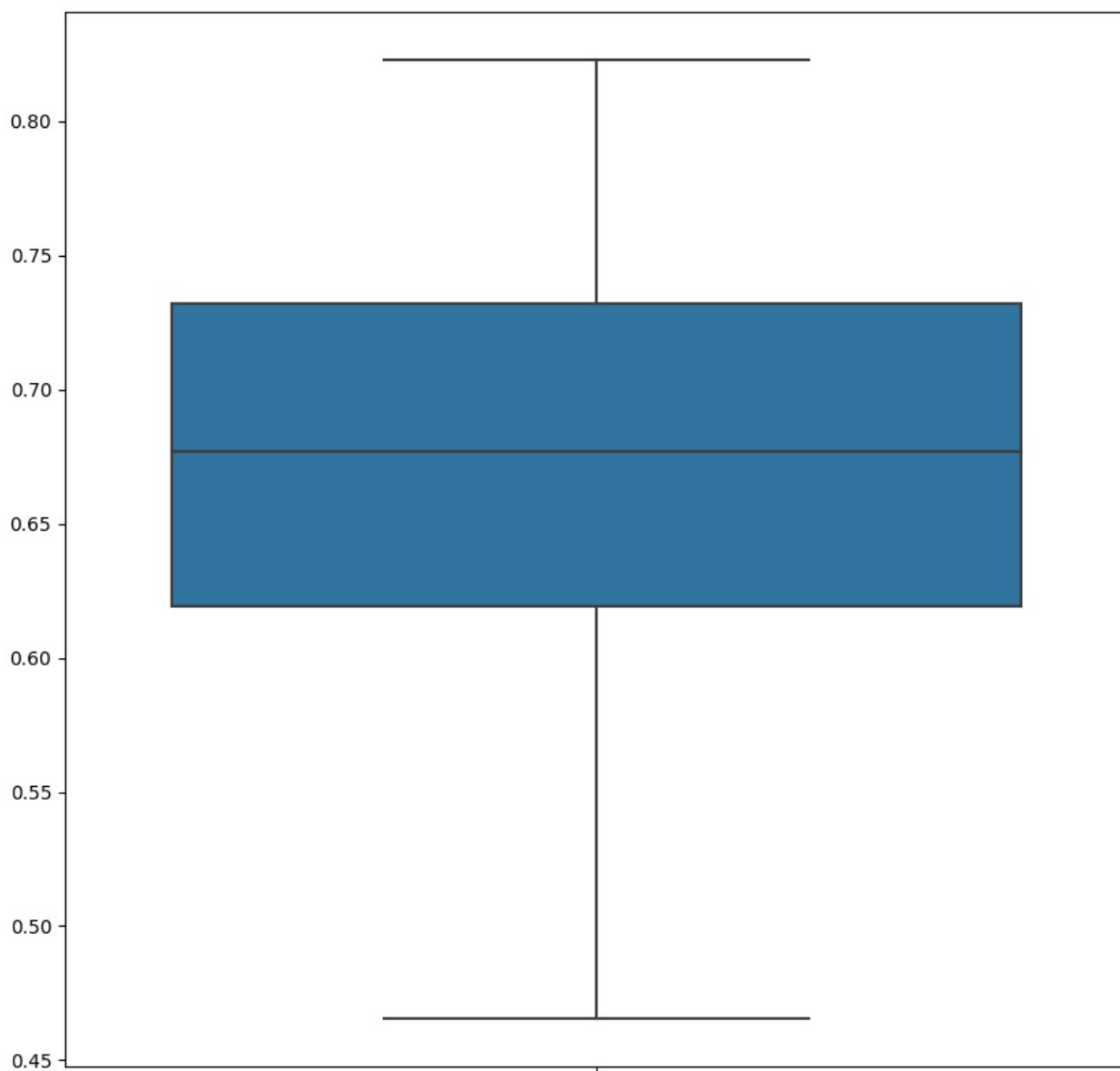




SpearmanrResult(correlation=-0.07454856591968433, pvalue=6.410987034414001e-47)

La majorité des projets semble se concentrer autour de la même fourchette de valeur, malgré quelques valeurs extrêmes. Il ne semble pas y avoir de corrélation entre le montant de succès de la campagne et son succès.

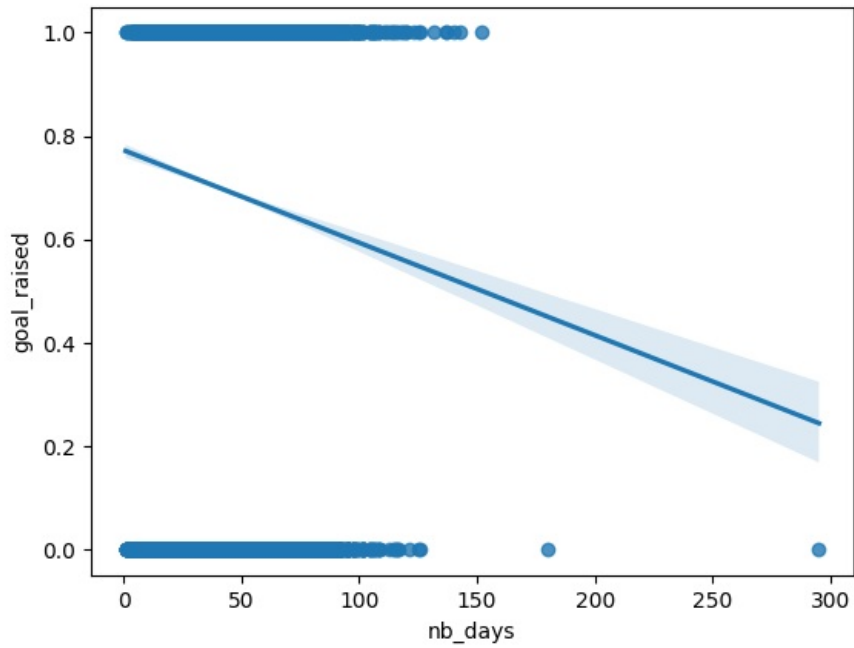
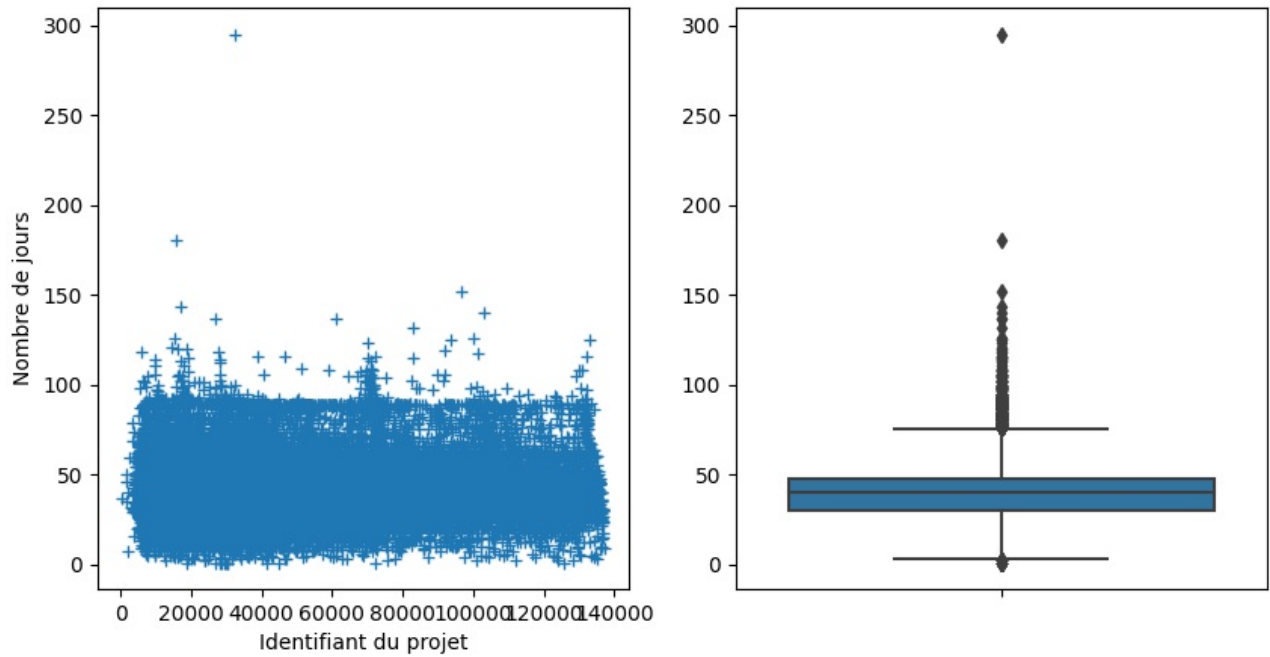
**success\_percent\_per\_main\_tag**



???

**nb\_days**

## Durée de la campagne de financement

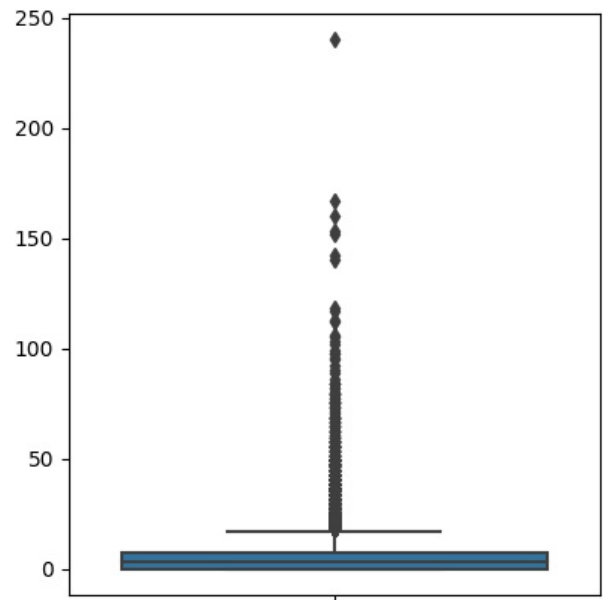
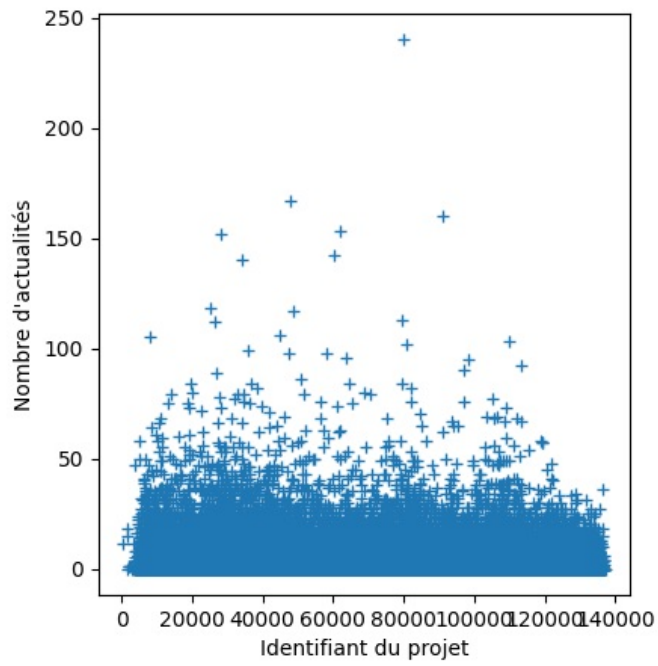


`SpearmanrResult(correlation=-0.06087830903719856, pvalue=7.775863519983873e-32)`

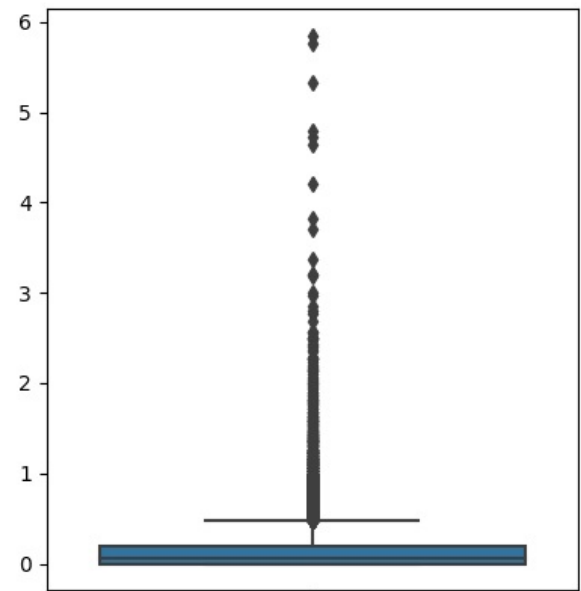
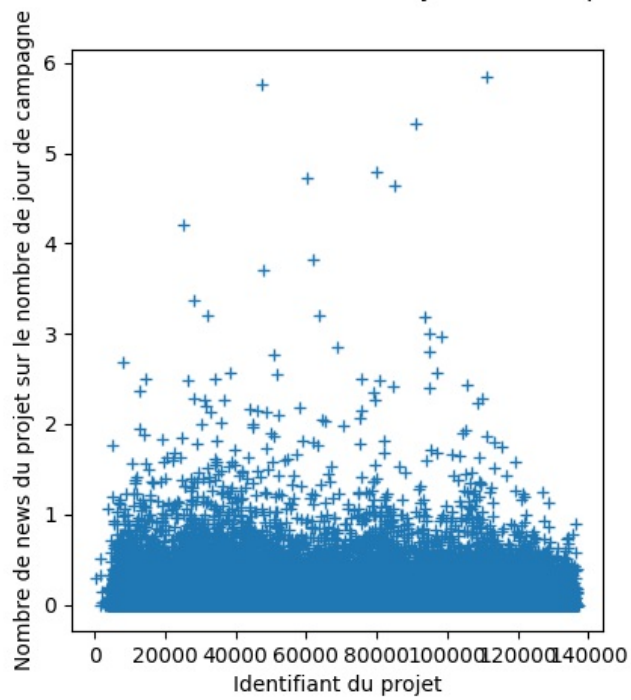
Les campagnes semblent durer plus ou moins un mois en grande majorité. Il ne semble n'y avoir aucune corrélation entre la durée de la campagne et son succès.

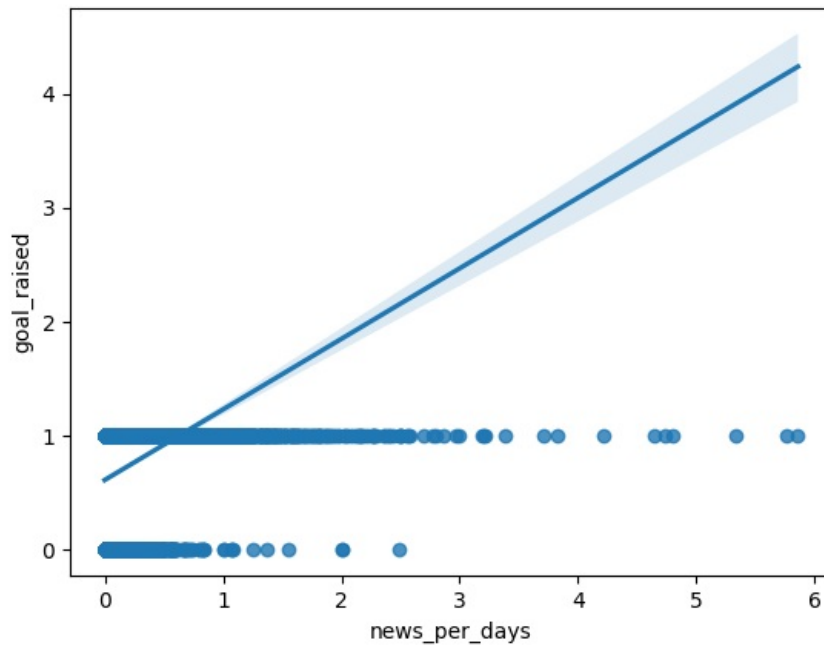
**news\_count**

Nombre d'actualités postées par projet



Nombre moyen de news par jour de campagne, par projet





SpearmanrResult(correlation=0.45036615456561185, pvalue=0.0)

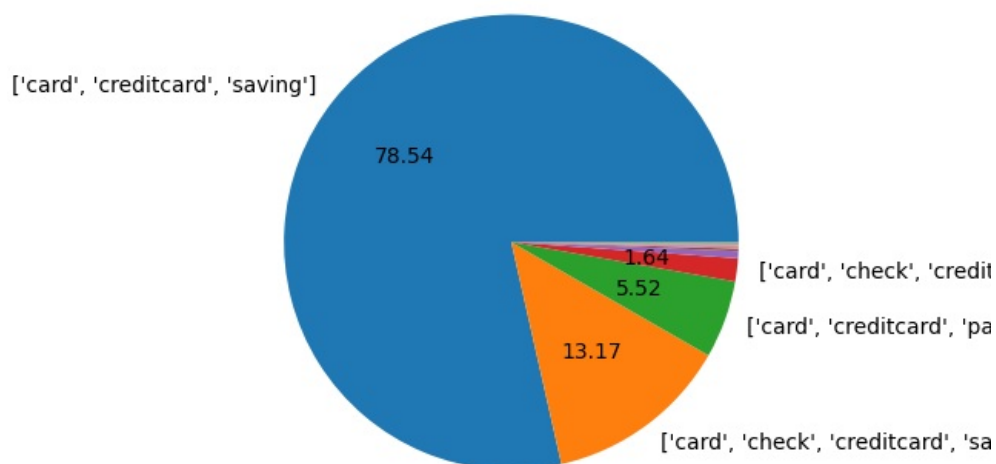
Un certain nombre de projets ne donne aucune nouvelle durant la campagne, la grande majorité n'en donne pas plus de cinq durant toute la campagne. La grande majorité des projets ne donne qu'une news tous les dix jours, au mieux. Il semble y avoir une légère corrélation entre le nombre de news par jour et le succès de la campagne.

#### owner

A reprendre -> s'assurer que si un owner lance 44 projets, il n'apparaisse pas 44 fois dans le plot.

#### payment\_methods

#### Représentation des moyens de paiement autorisés par les projets



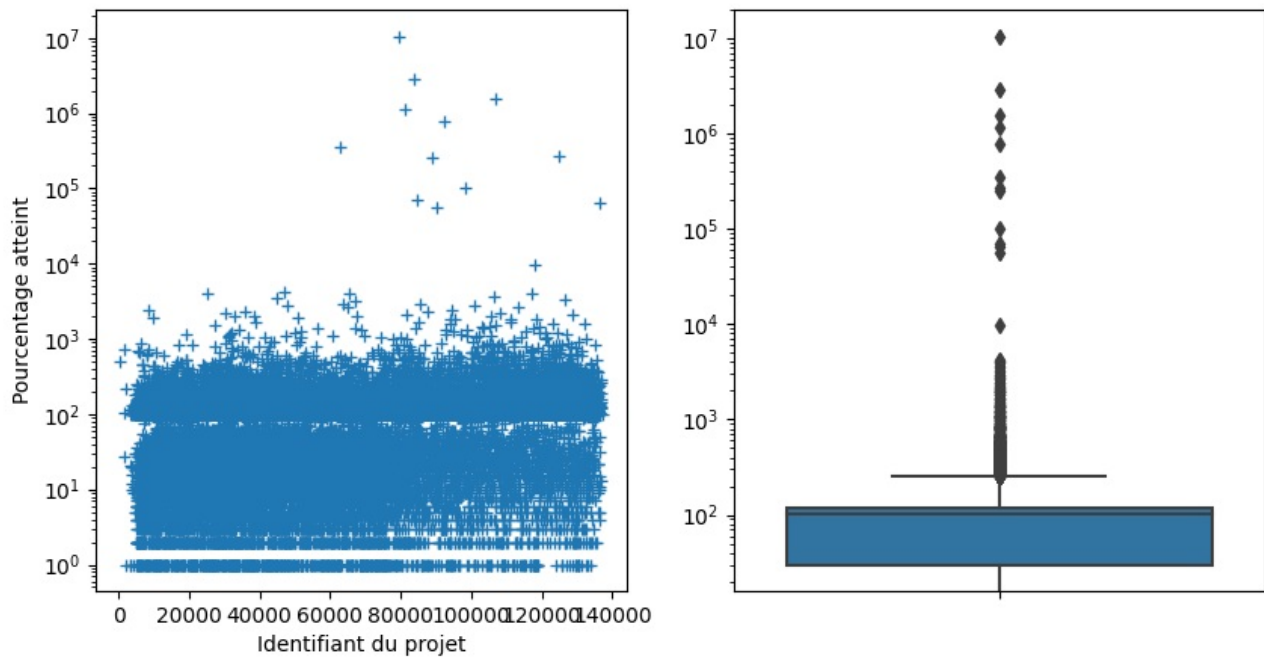
SpearmanrResult(correlation=nan, pvalue=nan)

La totalité autorise le trio "card, creditcard, saving" et 75% l'utilisent. Il ne semble y avoir aucune corrélation entre le moyen de paiement et le succès de la campagne.

#### percent



## Pourcentage du montant de succès demandé effectivement atteint par projet

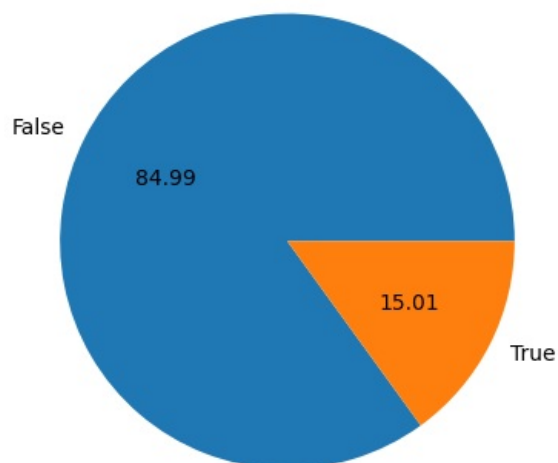


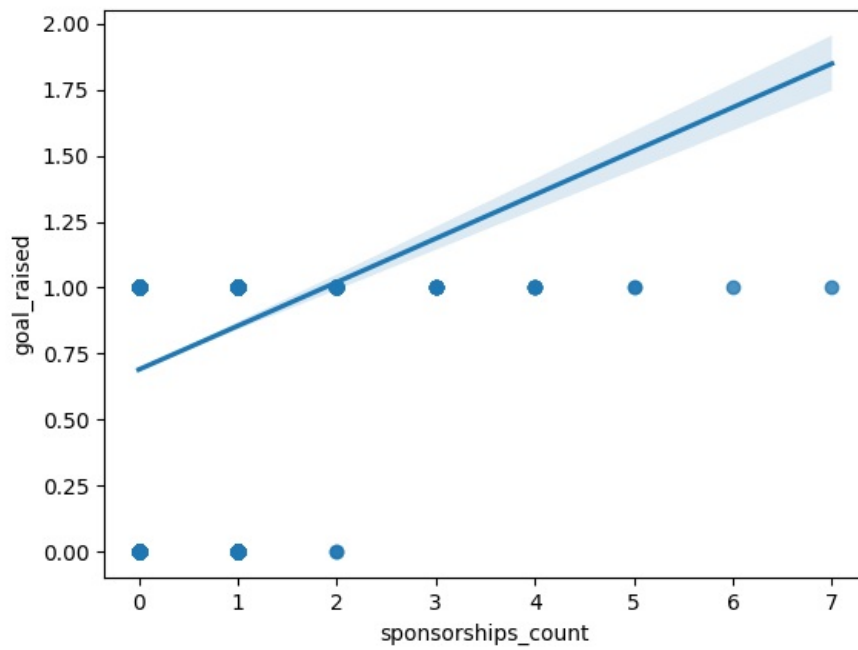
On note trois catégories de projets :

- ceux qui échouent complètement (moins de 50% \du montant demandé sont atteints)
- ceux qui réussissent "normalement" (entre 100% \et 175% \du montant demandé sont atteints)
- ceux qui réussissent "fortement" (au-delà de 200%)

**post\_covid**

Répartition des projets avant et après Mars 2020 (après=True)

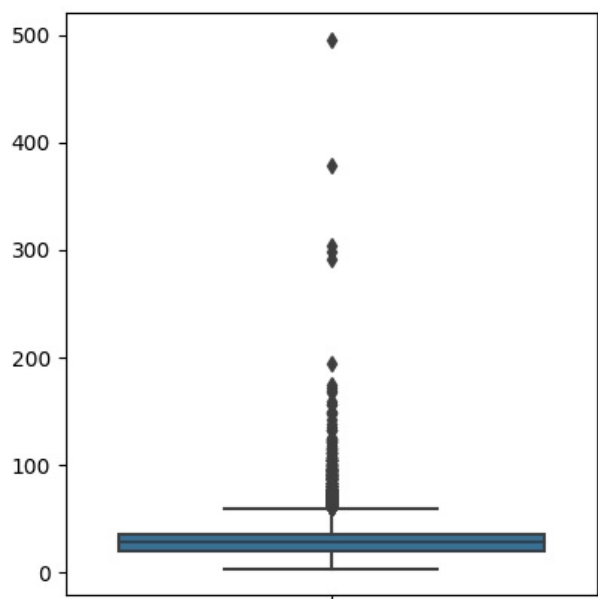
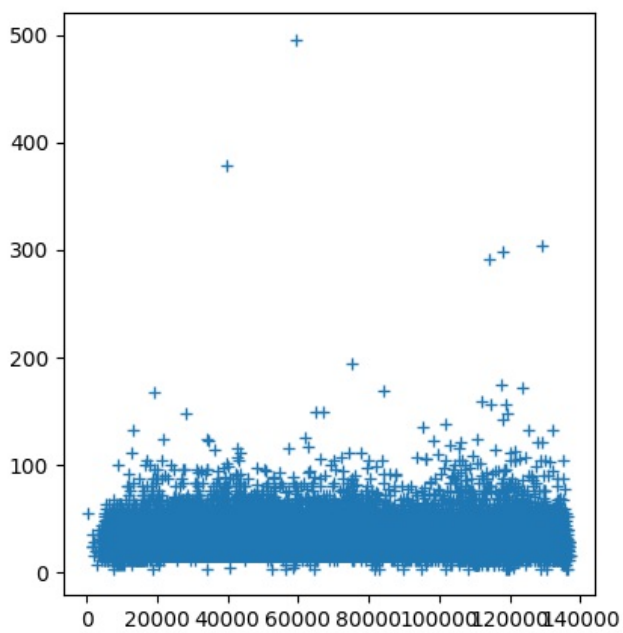


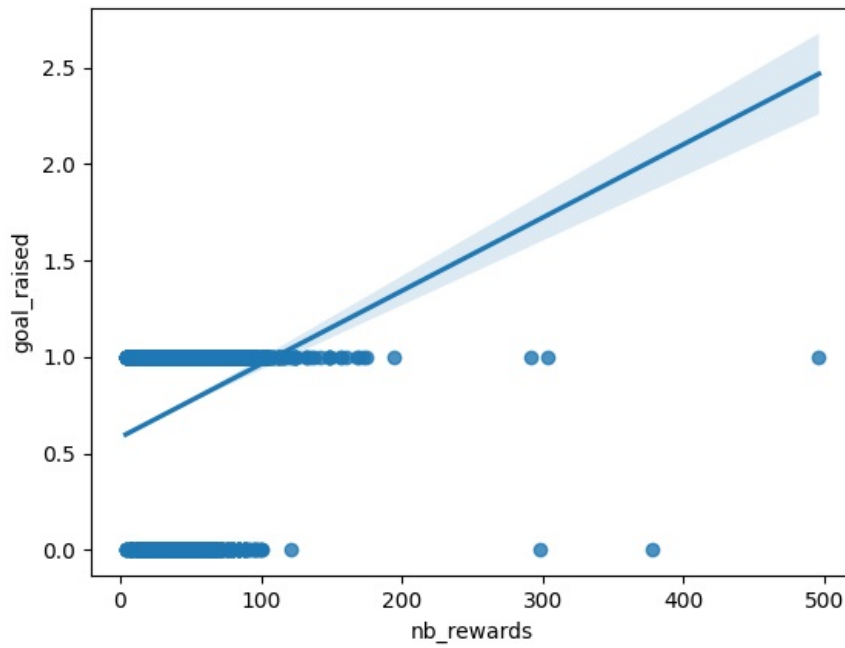


`SpearmanrResult(correlation=0.08698010948281582, pvalue=2.769548418451433e-63)`

L'obtention de sponsors semble être un phénomène très minoritaire. Il ne semble pas y avoir de corrélation entre le succès de la campagne et la présence de sponsors.

#### rewards



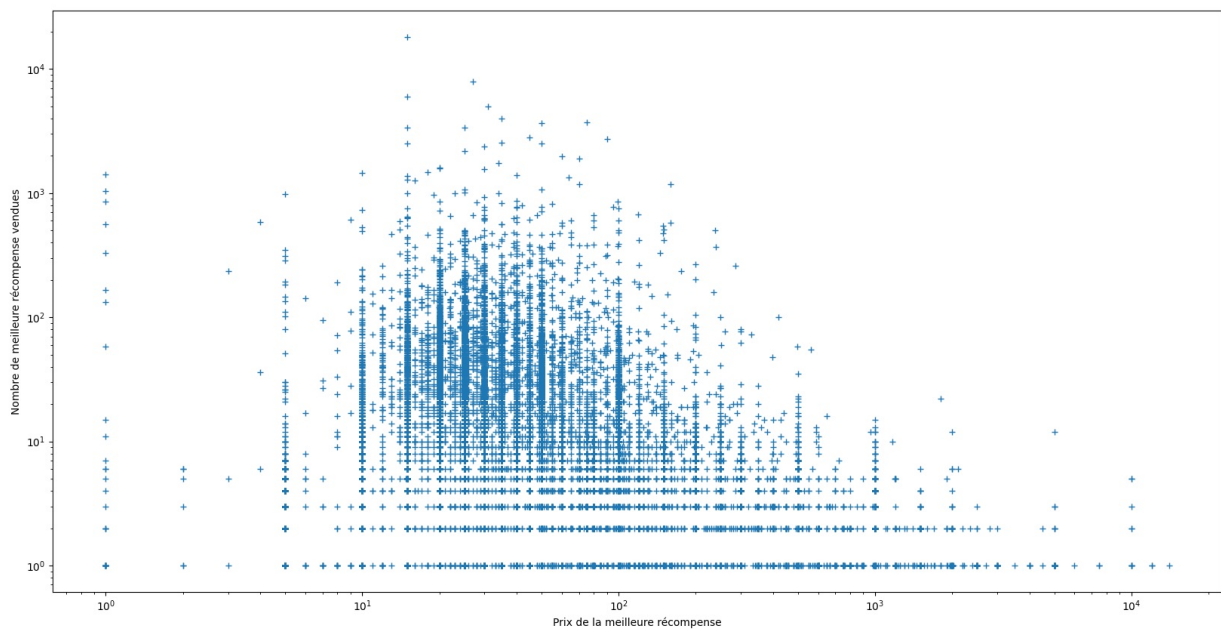


`SpearmanrResult(correlation=0.11324615380983429, pvalue=3.0649562544180267e-106)`

???

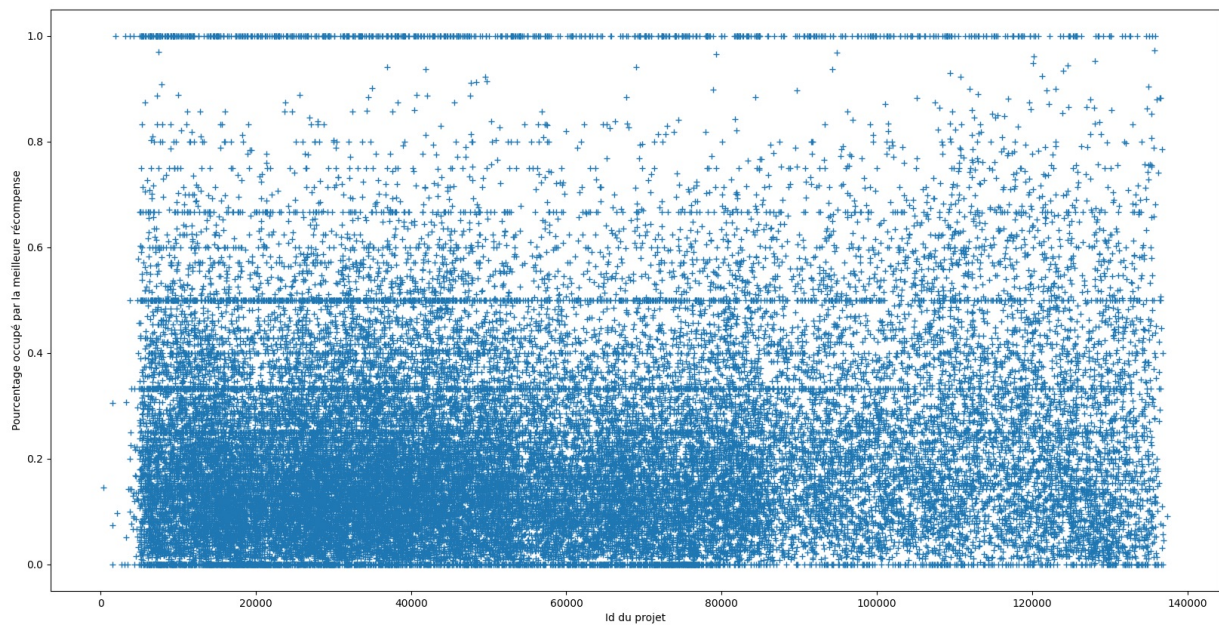
On appellera dans ce qui suit "meilleure récompense", la récompense maximisant le rapport de son prix sur le nombre d'achat. Ainsi une récompense vendue une seule fois à 500€ est meilleure qu'une récompense vendue 100 fois à 2€.

Représentation du nombre de meilleure récompense vendue par projet, en fonction de son prix



Les récompenses les plus rentables des projets semblent se situer entre 10 et 100€, pour une centaine de vente.

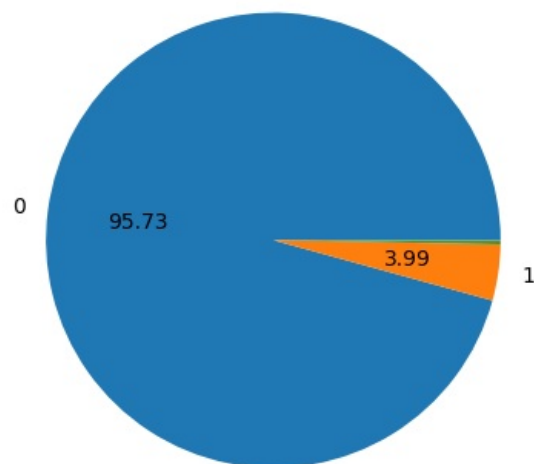
Représentation de la place occupée par la meilleure récompense au sein de toutes les récompenses, par projet, en fonction de l'id du projet

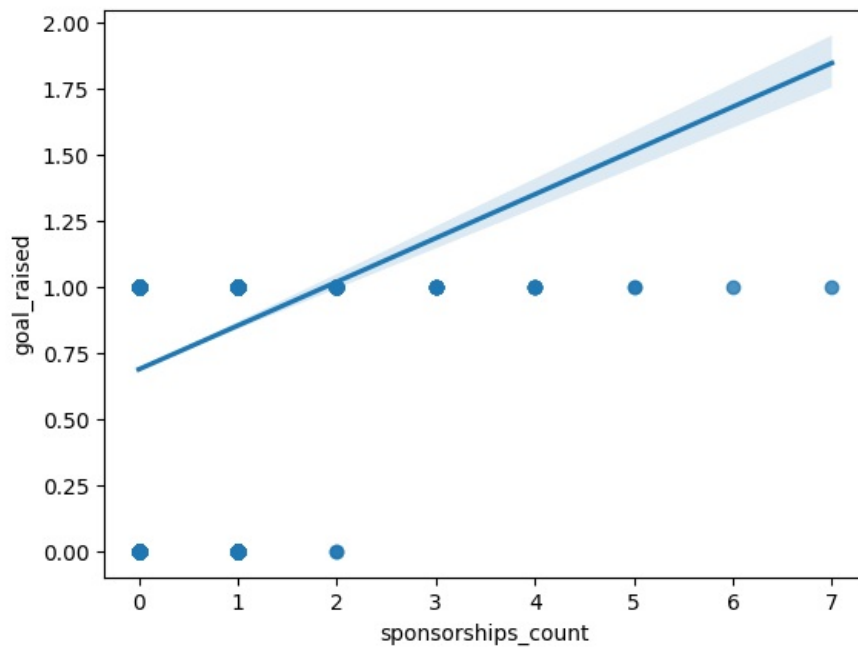


???

**sponsorships\_count**

Représentation du nombre de sponsors parmi les projets.



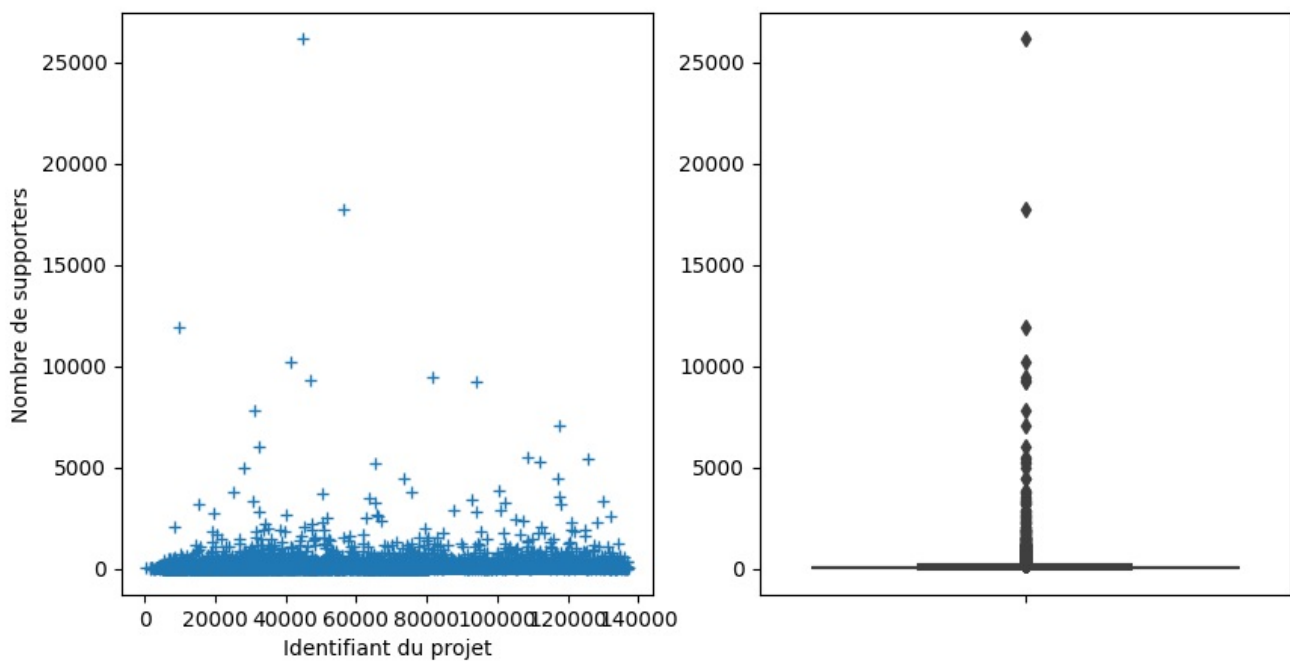


SpearmanrResult(correlation=0.08698010948281582, pvalue=2.769548418451433e-63)

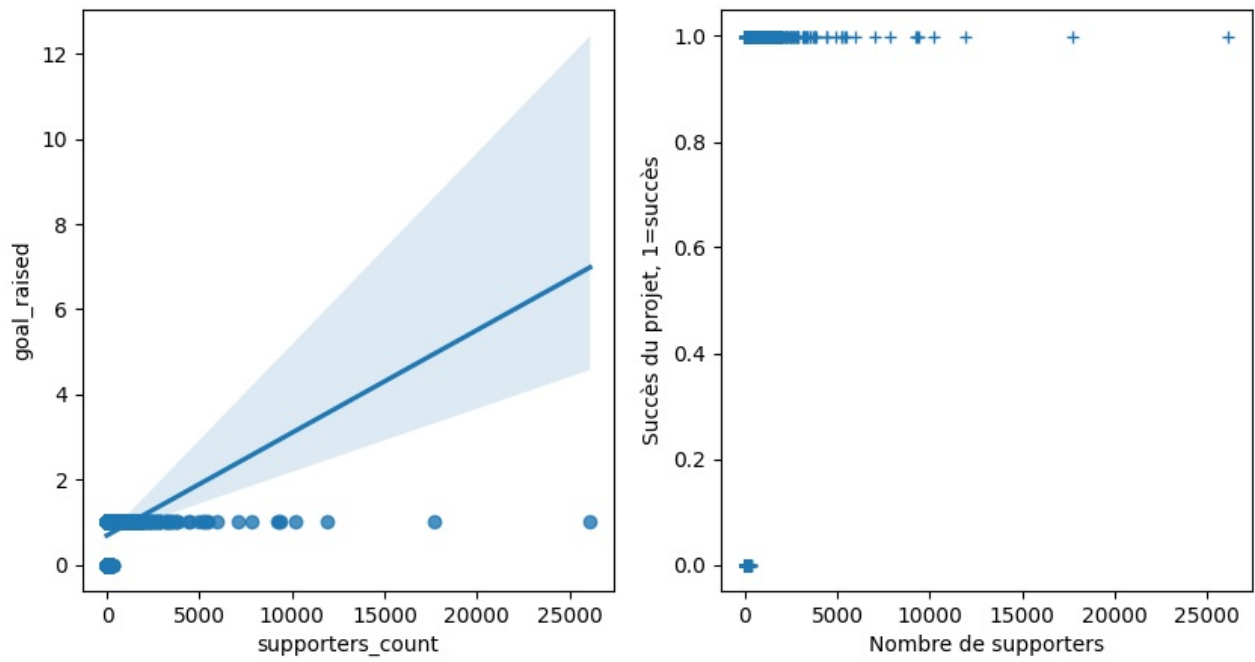
L'obtention de sponsors semble être un phénomène très minoritaire. Il ne semble pas y avoir de corrélation entre le succès de la campagne et la présence de sponsors.

**supporters\_count**

Nombre de supporters par projet



## Succès du projet en fonction du nombre de supporters

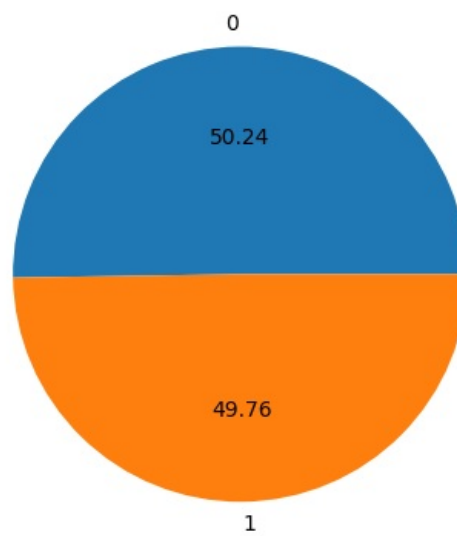


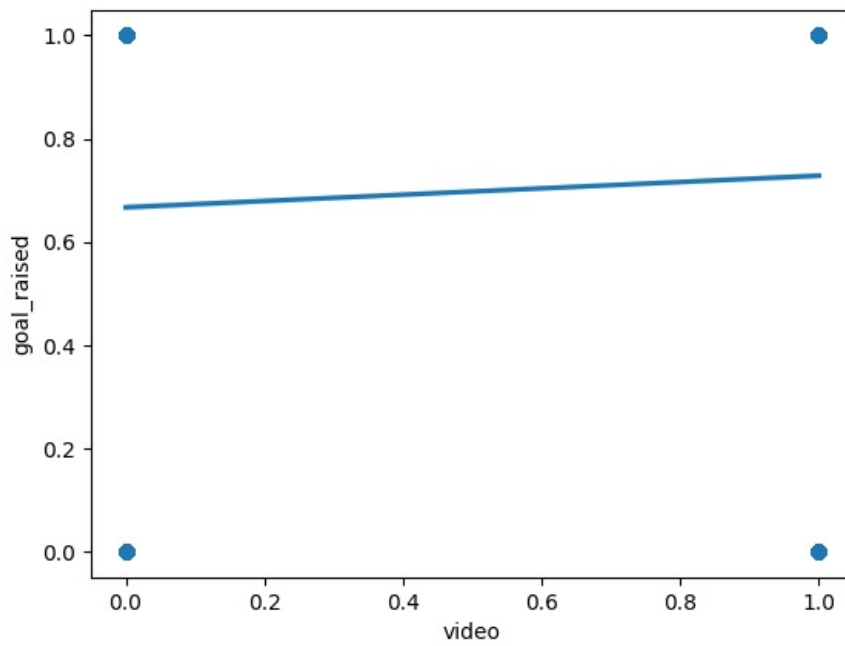
`SpearmanrResult(correlation=0.6697024008514539, pvalue=0.0)`

On constate une bonne corrélation entre le succès du projet et le nombre de supporters.

**video**

## Présence d'une vidéo dans le projet



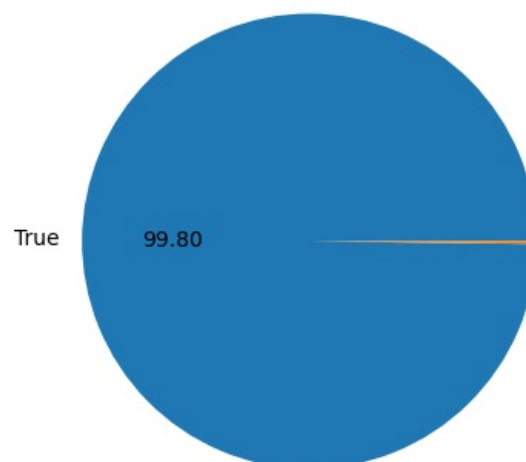


SpearmanrResult(correlation=0.06673662301049162, pvalue=6.289488654579654e-38)

Il ne semble pas y avoir de corrélation entre le fait de posséder une vidéo et le succès de la campagne

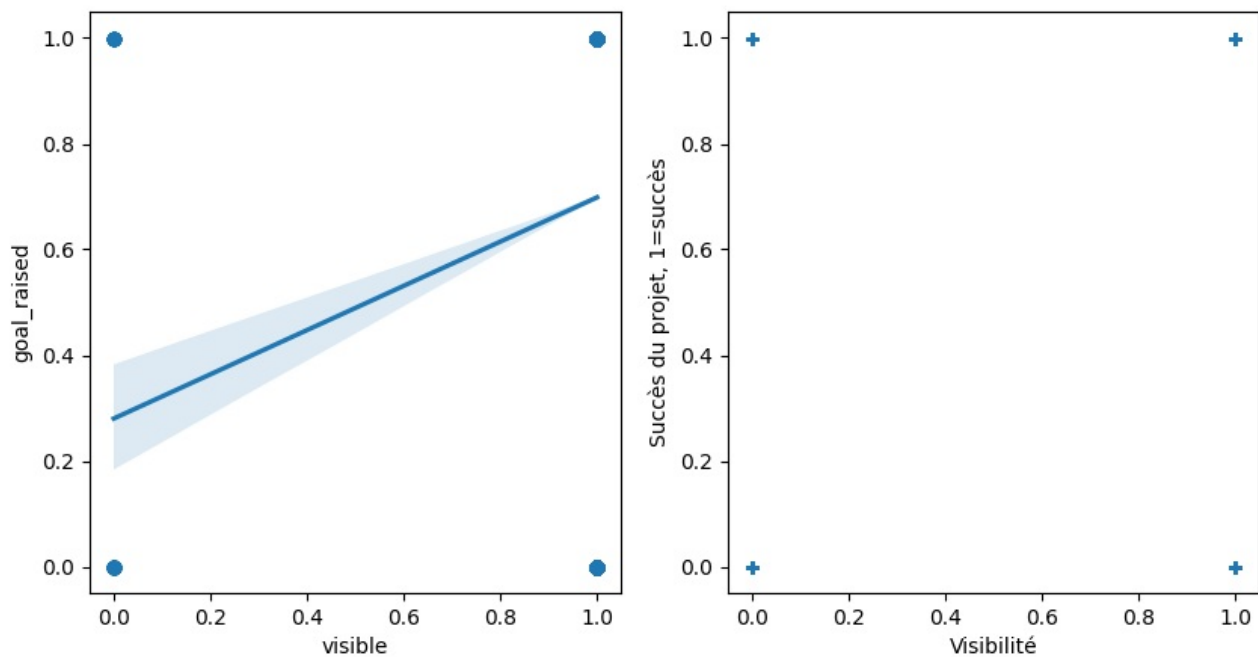
**visible**

Indexation sur les moteurs de recherches du projet





## Succès du projet en fonction de sa visibilité

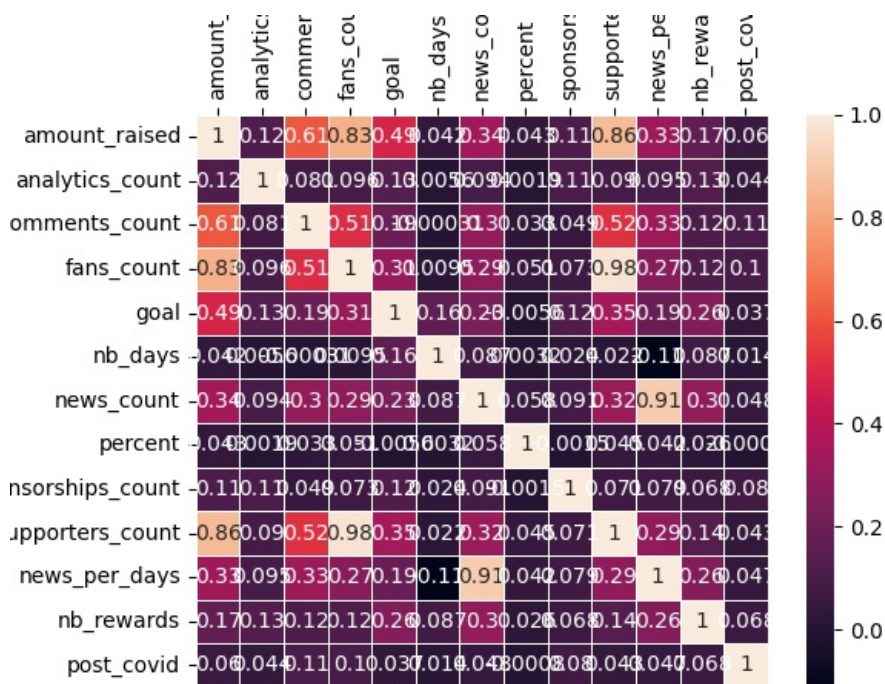


SpearmanrResult(correlation=0.04090737720076117, pvalue=3.1212076817719414e-15)

La plupart des projets sont indexés sur les moteurs de recherches. On ne constate cependant pas de corrélation entre la visibilité et le succès d'un projet

## Autres stats ? NOMMER CA AUTREMENT SVP

### Matrice de corrélation

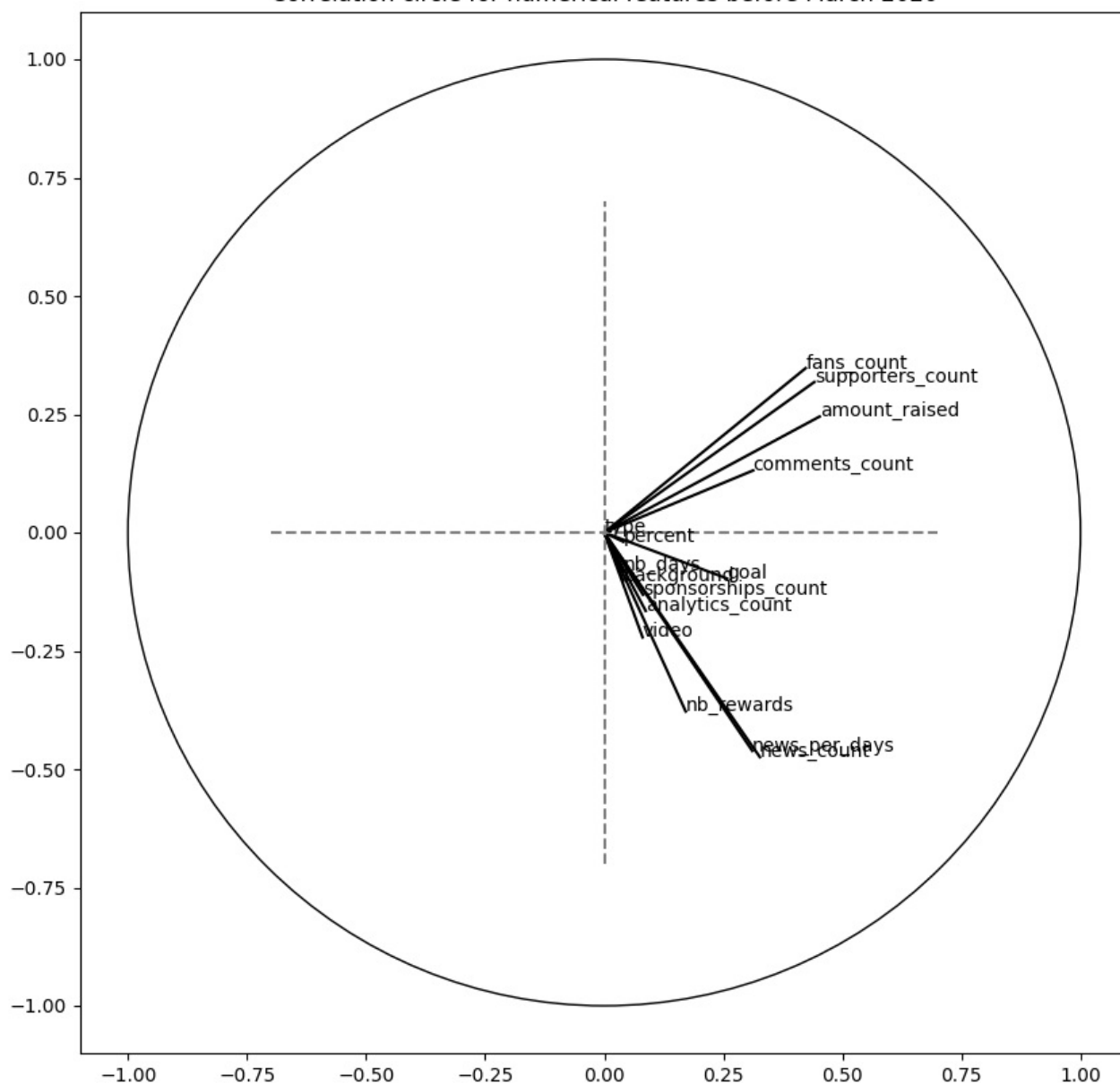


???

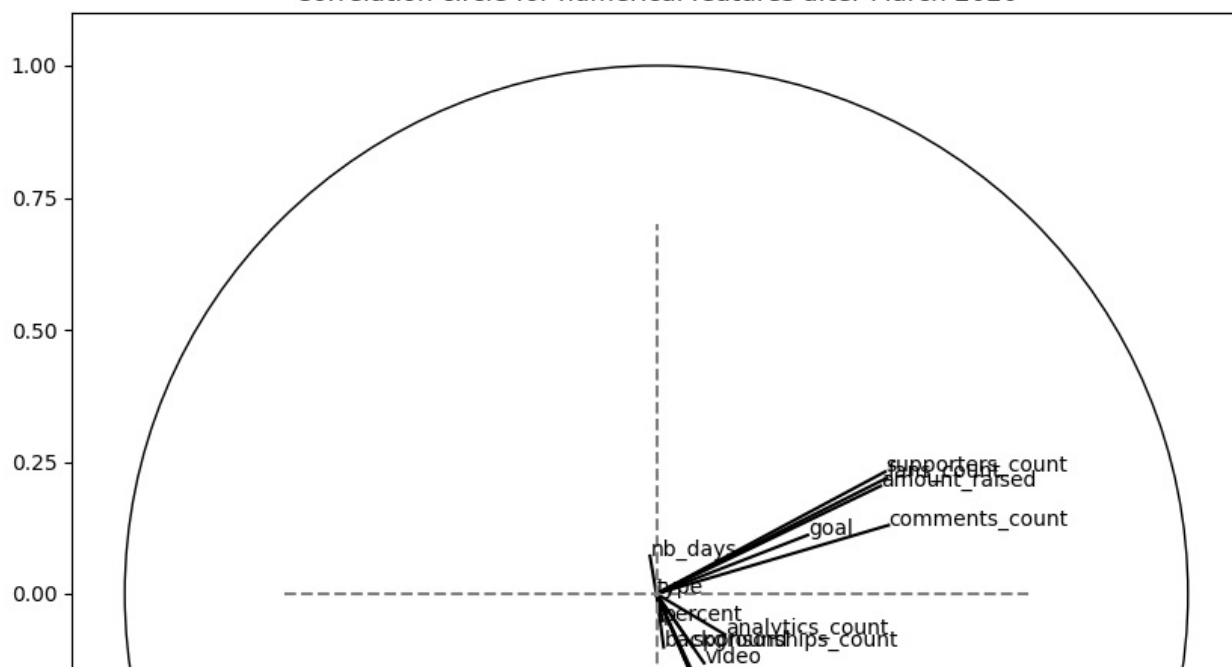
PCA

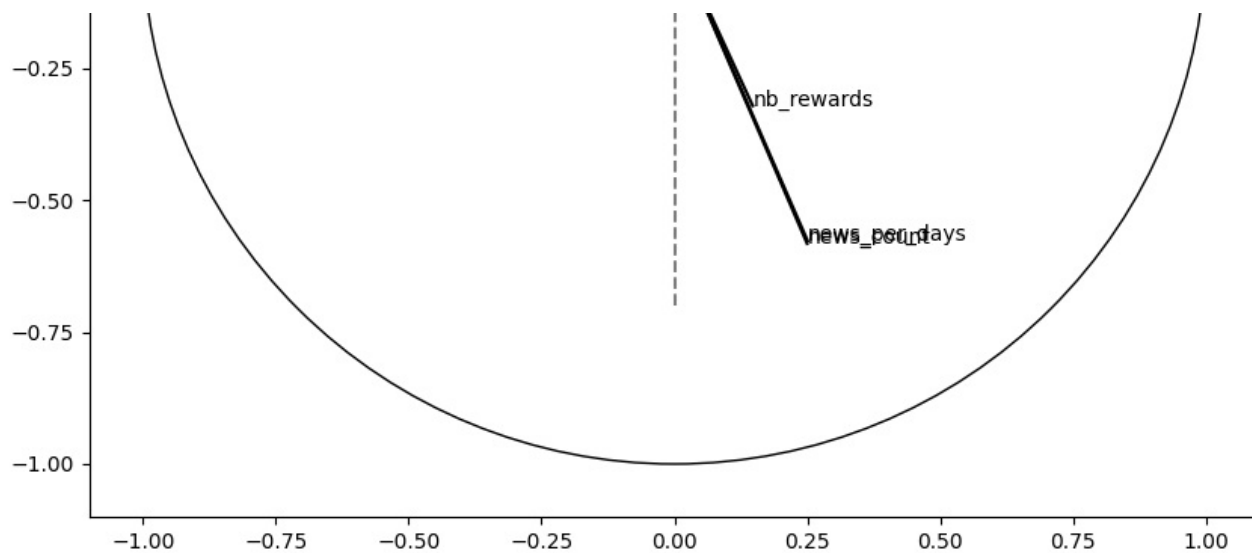


Correlation circle for numerical features before March 2020



Correlation circle for numerical features after March 2020





???