

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.

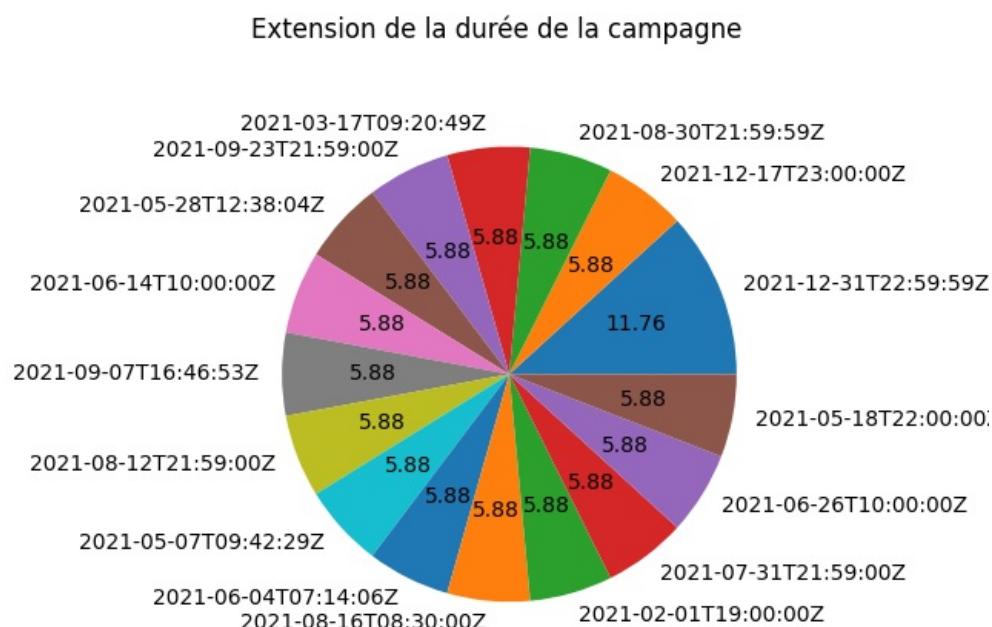
- amount_raised
- comments_count
- committed
- date_end
- date_end_extra_time
- date_goal_raised
- date_start
- fans_count
- finished
- is_cancelled
- is_in_extra_time
- lowest_contribution_amount
- nb_days
- nb_products_sold
- news_count
- orders_count
- percent
- sponsorships_count
- supporters_count
- 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_goal_raised

La colonne **date_goal_raised** est incompatible avec notre problématique : conseiller les lanceurs de projets pour qu'ils réussissent leur projet. Elle est donc retirée.

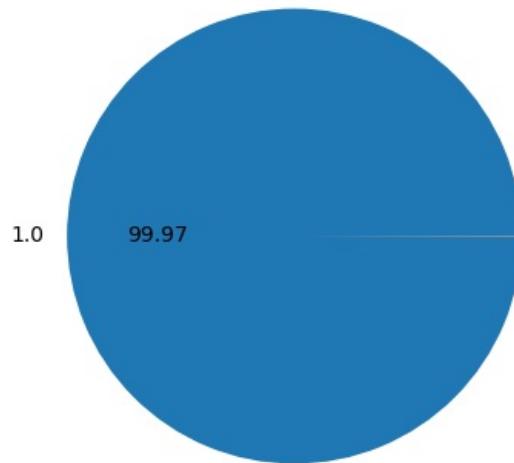
date_end_extra_time



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
- ecomments_enabled
- ecountry
- ecurrency
- 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).

location

La colonne location contient un dictionnaire avec plusieurs attributs. On choisit de ne rien garder de la localisation du owner car elle est indépendante du projet.

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.

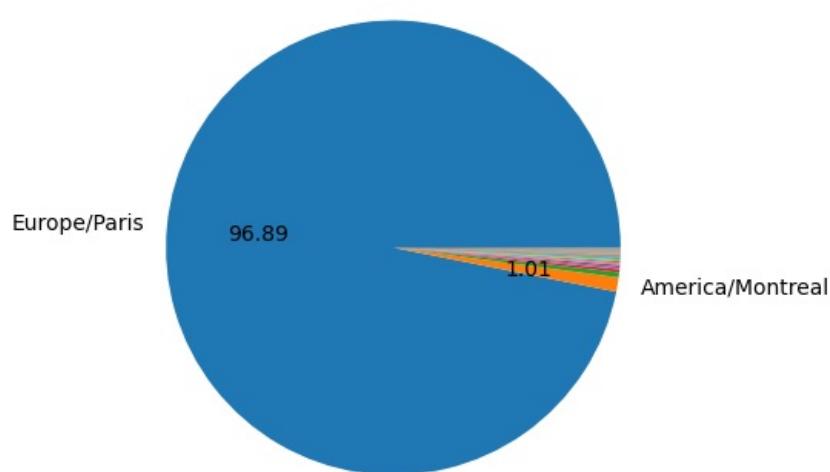
De plus, les stats des owners ne peuvent être utilisées : par exemple si un owner a lancé 44 projets, alors pour **chaque** projet, le nombre 44 apparaîtra, faisant grossir artificiellement les chiffres. La colonne est donc retirée.

payment_methods

La gestion des moyens de paiements ne nous a pas semblé pertinente. Colonne retirée.

timezone

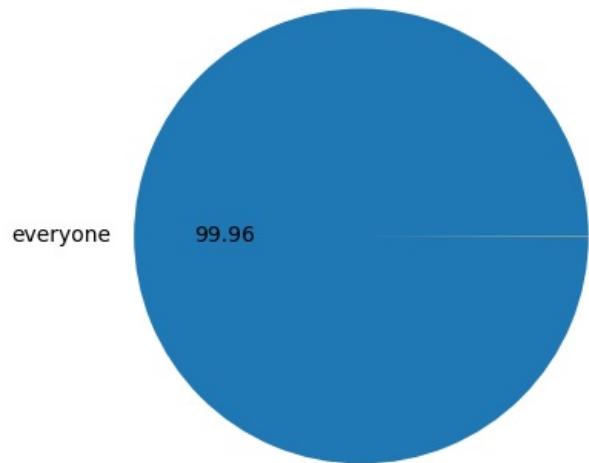
Répartition de la zone temporelle au sein des projets



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

comments_enabled

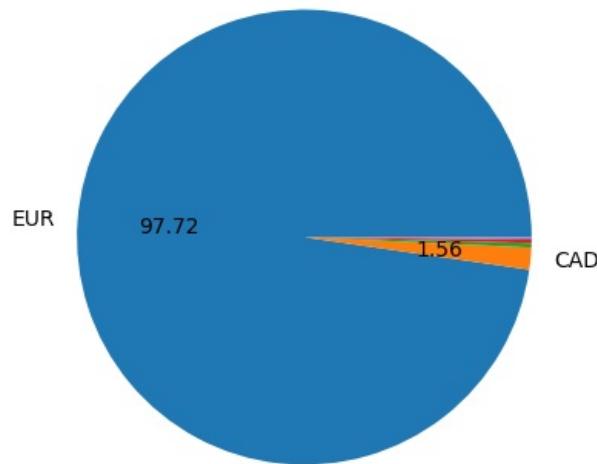
Répartition des permissions de commentaires



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

currency

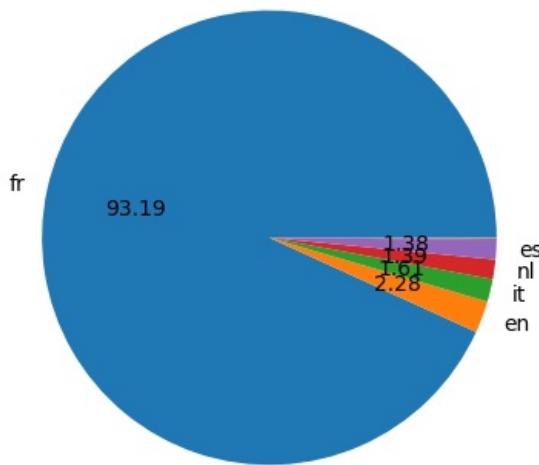
Répartition de la monnaie utilisée au sein des projets



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_start', 'description_fr', 'description_funding_fr', 'description_yourself_fr', 'fans_count', 'goal', 'goal_raised', 'id', 'main_tag', 'name_fr', 'nb_days', 'nb_products_sold', 'news_count', '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**

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 **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**

OneHotEncoding des main-tag

nb_rewards

Création de la colonne nb_rewards.

Afin de traiter les données textuelles comme la description du projet, la description de l'auteur du projet ainsi que le titre et le sous-titre du projet, nous gardons seulement la longueur, le nombre de points d'exclamation et le nombre de points d'interrogations dans ces données

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.

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"

date_*

Il n'est plus utile de conserver les dates de début et de fin si on dispose des colonnes nb_days et post_covid. Retirons les.

type

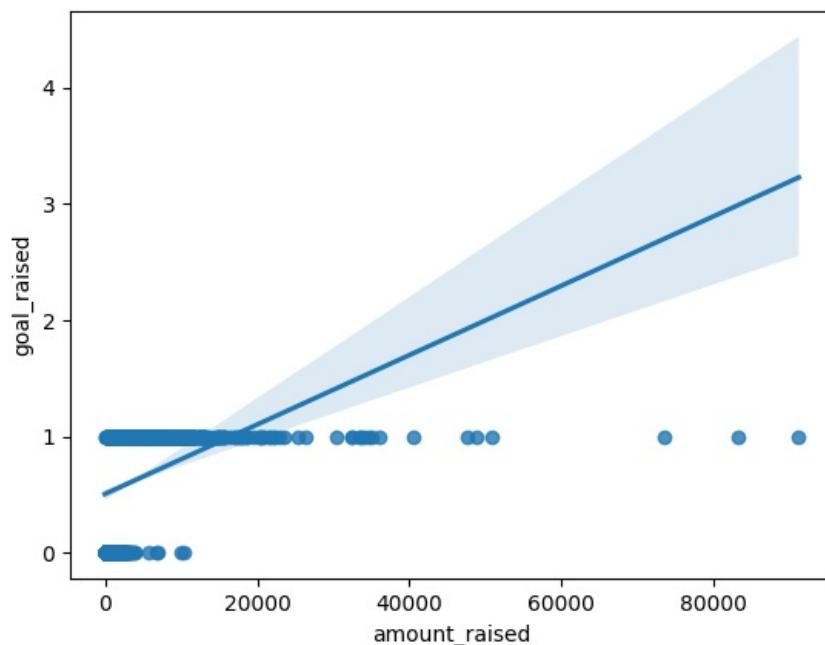
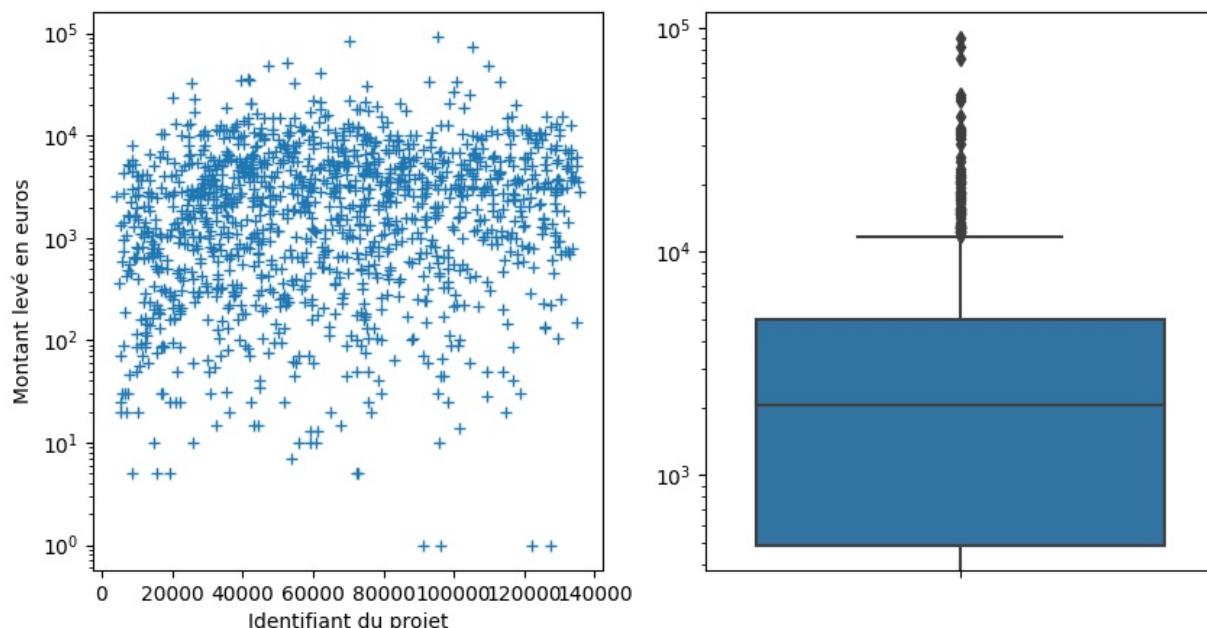
Les projets fonctionnent différemment selon qu'ils concernent des préventes ou une financement. Il convient donc de séparer le set en deux sous-sets.

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

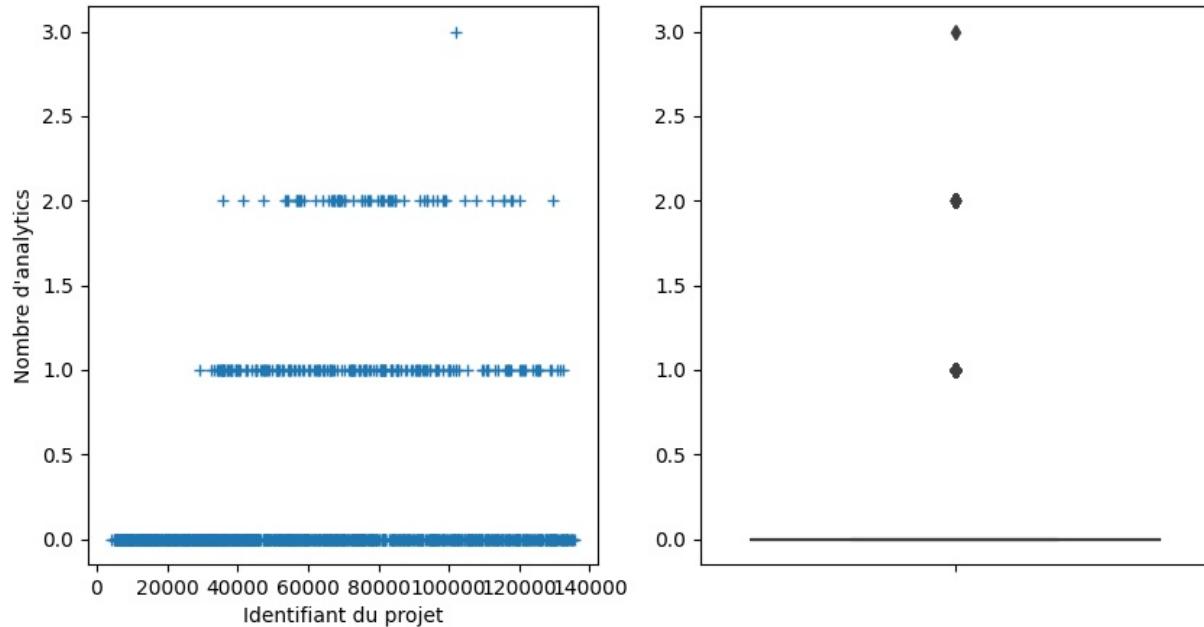


`SpearmanrResult(correlation=0.7398767789519061, pvalue=1.4817674190383845e-241)`

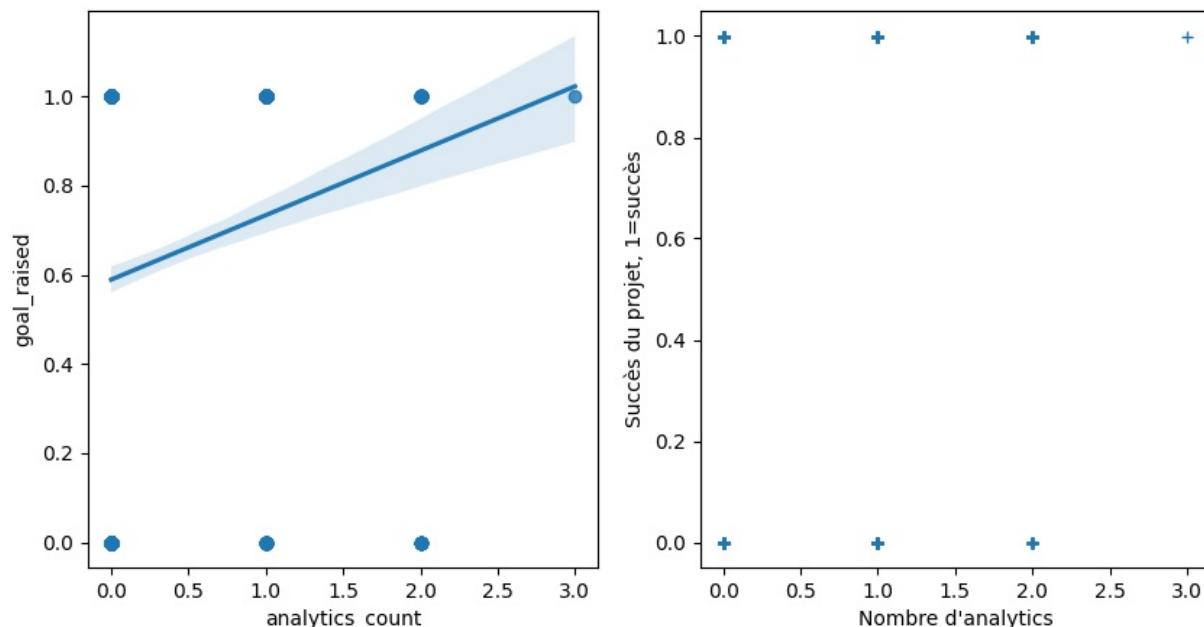
Les projets semblent assez homogènes dans les montants levés même si moins nombreux pour les plus hauts montants. Le coefficient de corrélation est assez élevé entre le montant obtenu et le succès de la campagne; plus une campagne lève de dons, plus elle est susceptible d'aboutir.

`analytics_count`

Nombre d'analytics par projet



Succès du projet en fonction du nombre d'analytics

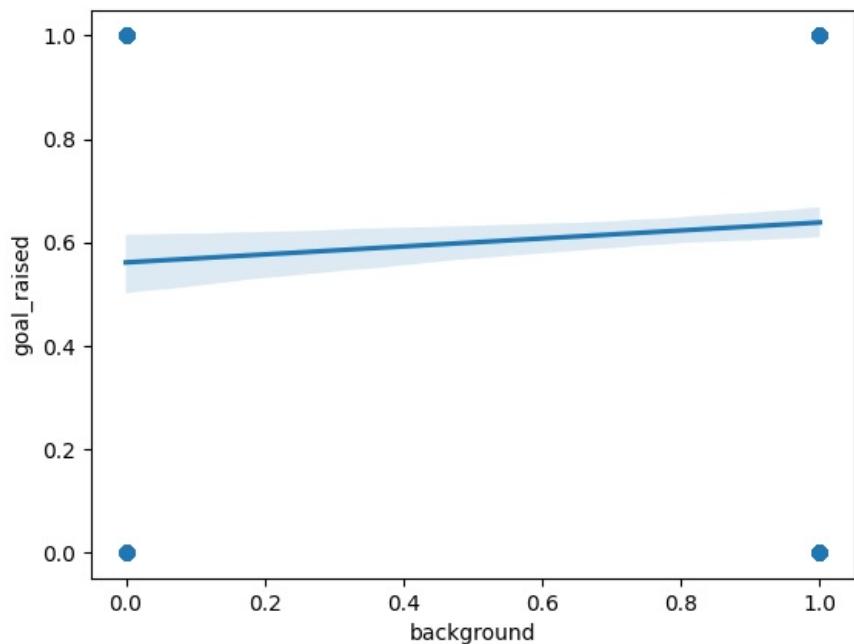
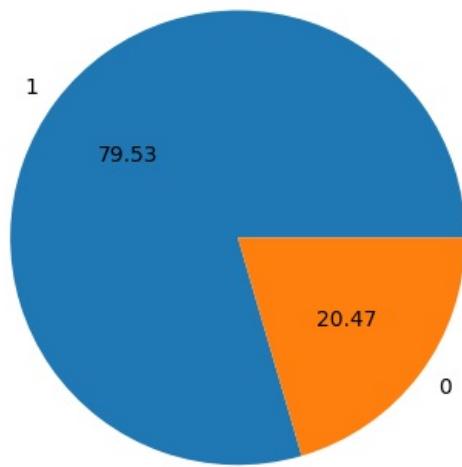


`SpearmanrResult(correlation=0.16252605727361744, pvalue=1.0679177670820515e-09)`

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

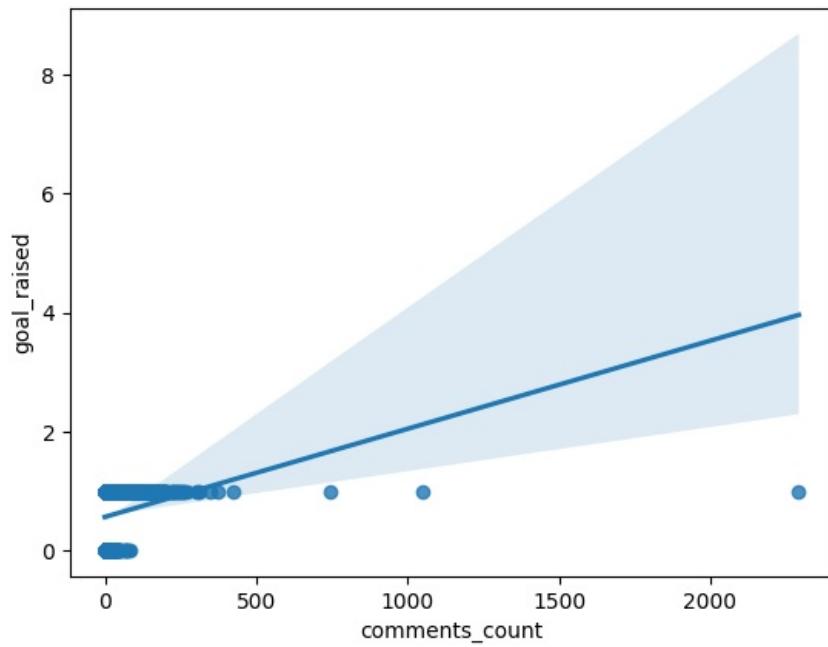
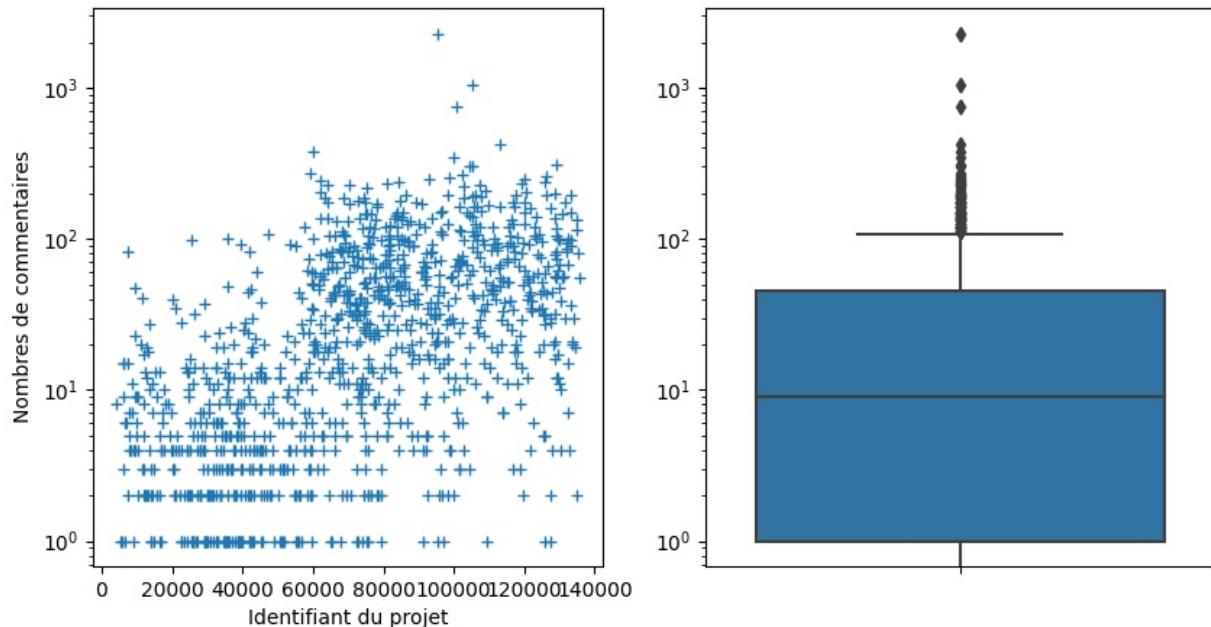


SpearmanrResult(correlation=0.06432194724046776, pvalue=0.016388331150595726)

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

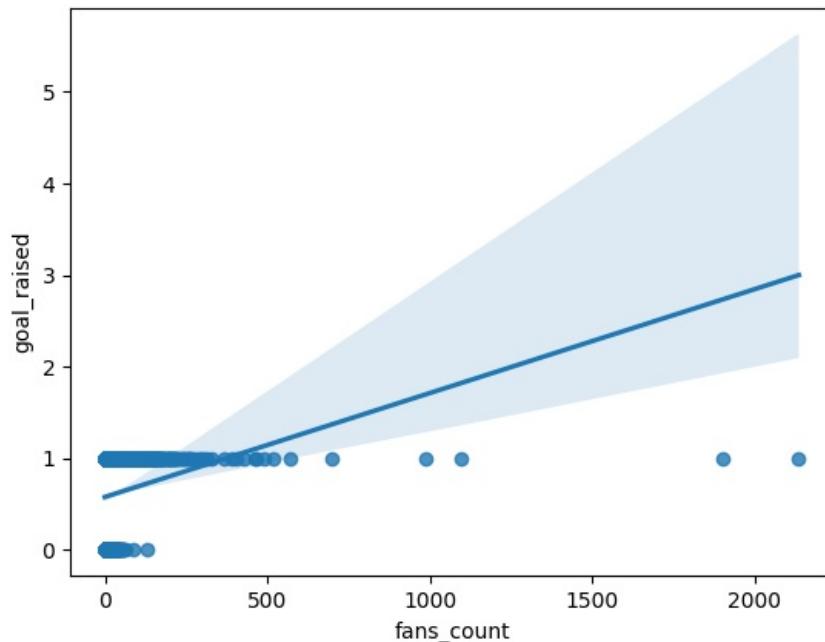
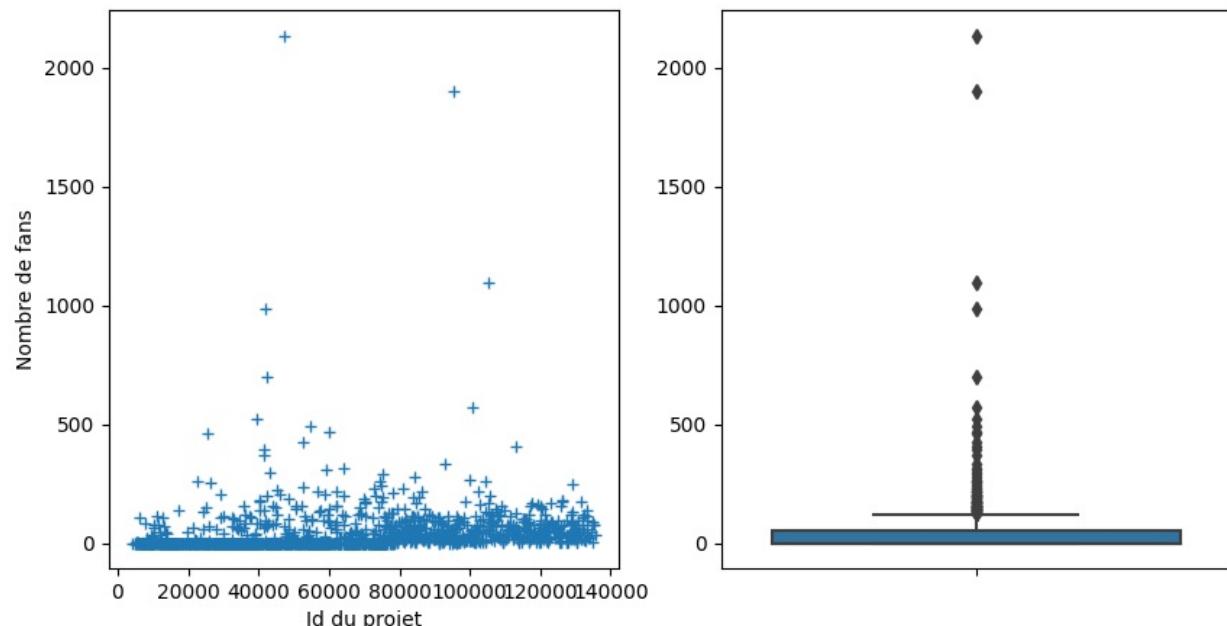


SpearmanResult(correlation=0.49529665504071824, pvalue=4.788486640207062e-87)

Les projets reçoivent globalement assez peu de commentaires. La plupart des projets qui ont beaucoup de commentaires, ont bien fonctionné. Il y a une corrélation entre le nombre de commentaires et le succès de la campagne; il semble que plus on en parle, plus elle soit susceptible de réussir, même si ce facteur ne semble pas être déterminant.

fans_count

Représentation du nombre de fans

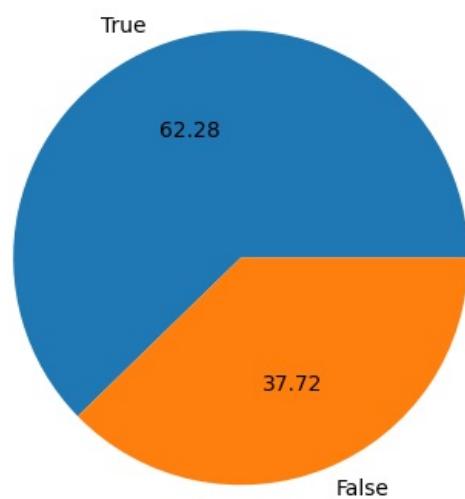


`SpearmanrResult(correlation=0.39243093603101564, pvalue=1.8053799236972974e-52)`

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.

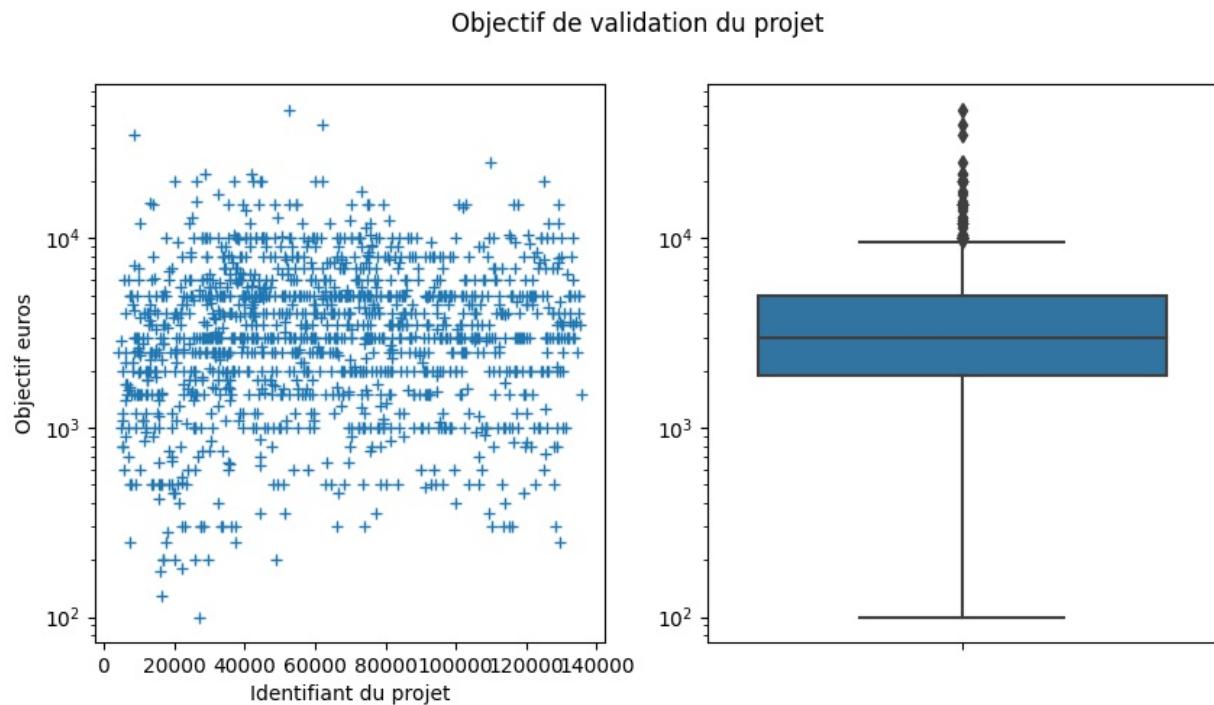
goal_raised

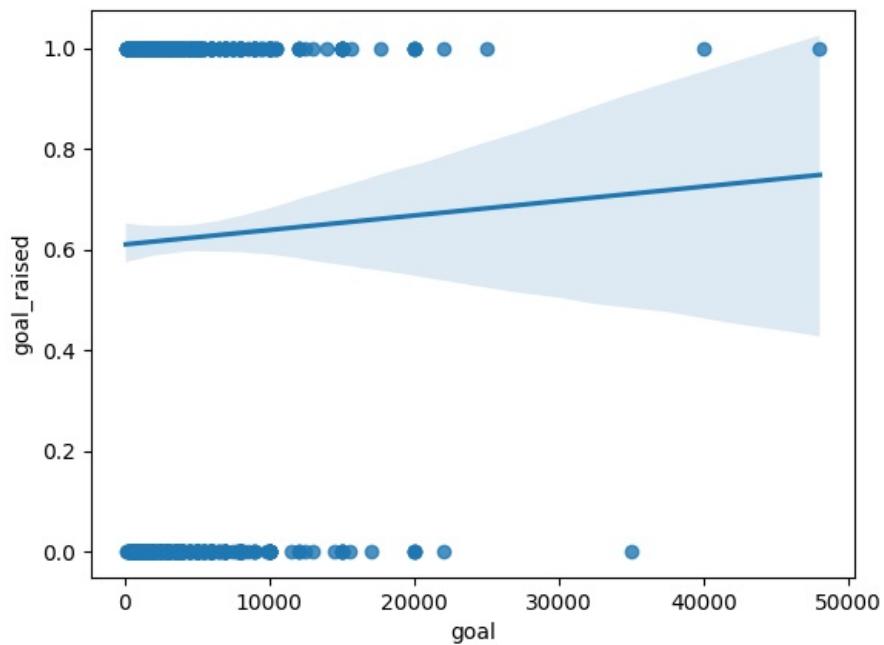
Représentation du taux de succès des projets, succès=True



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

goal



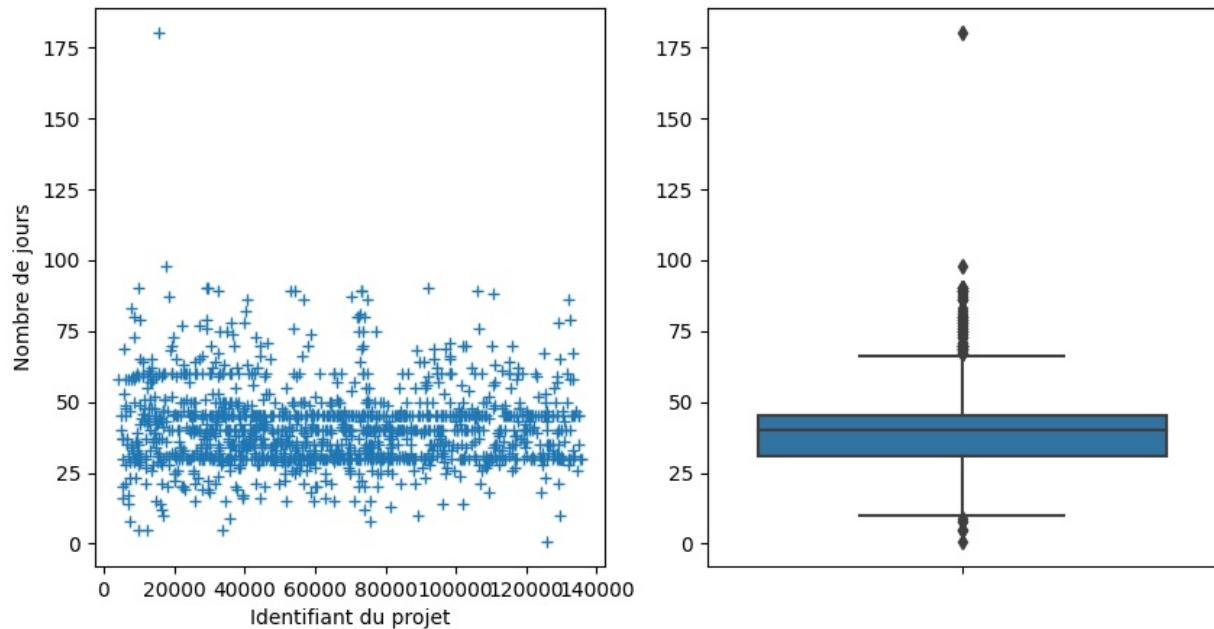


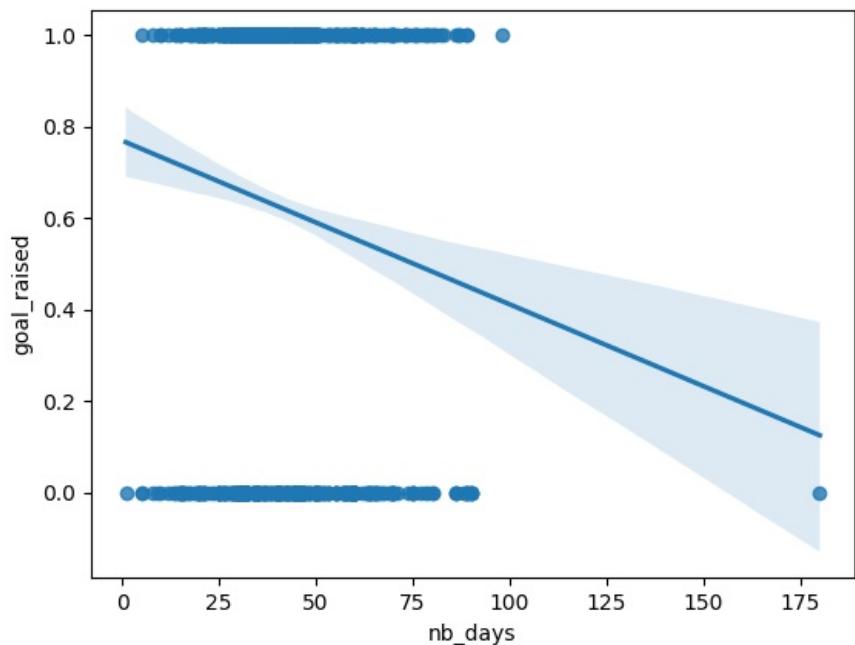
SpearmanrResult(correlation=0.009905089604229271, pvalue=0.7119554275711835)

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.

nb_days

Durée de la campagne de financement

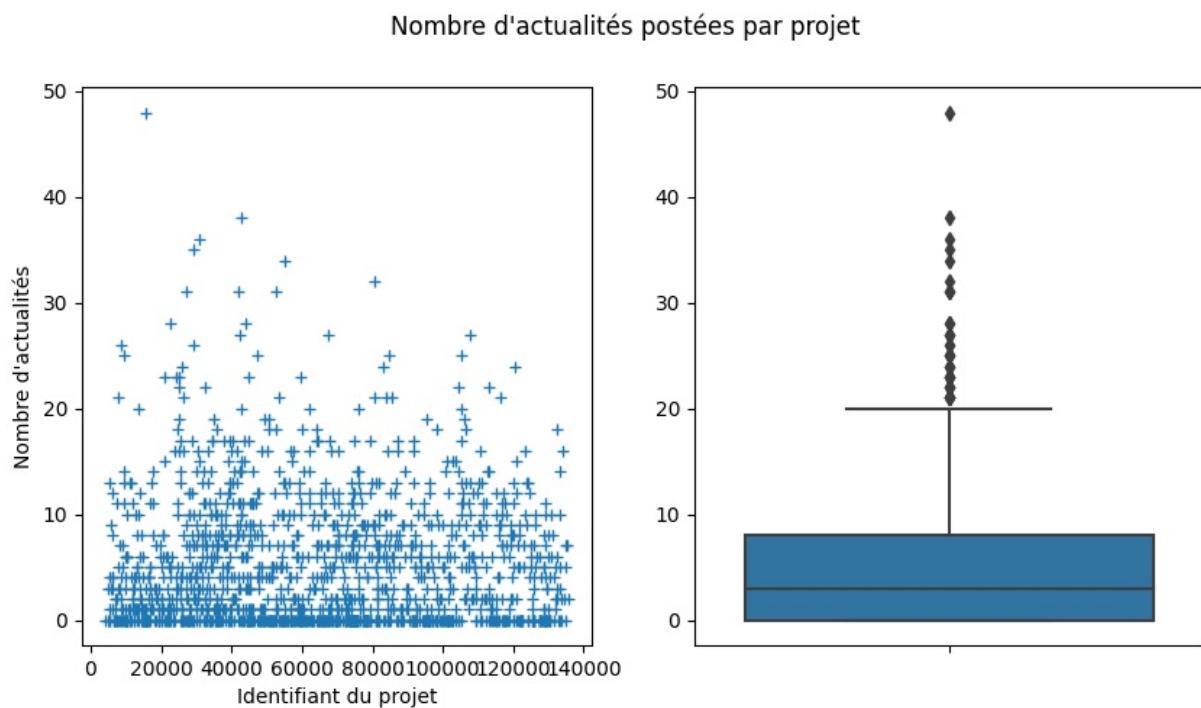




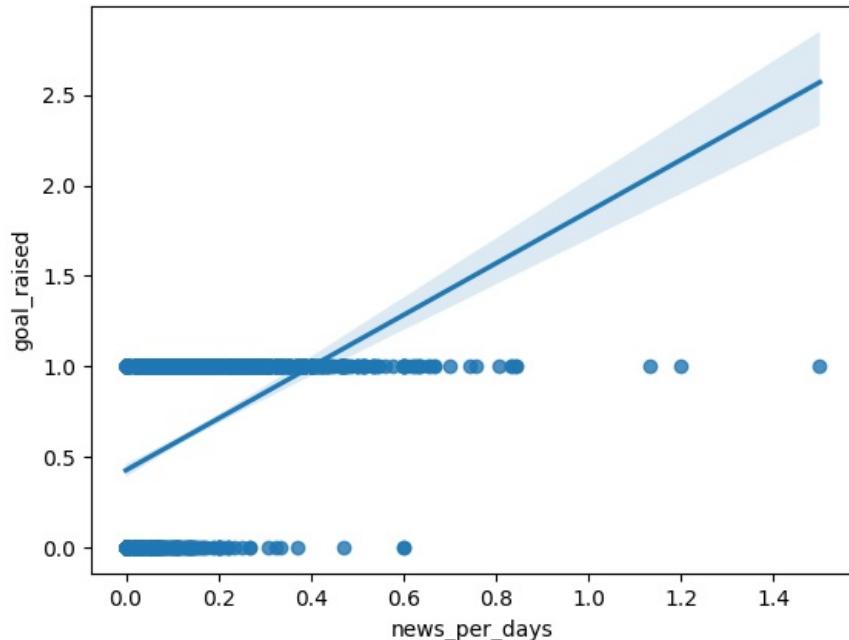
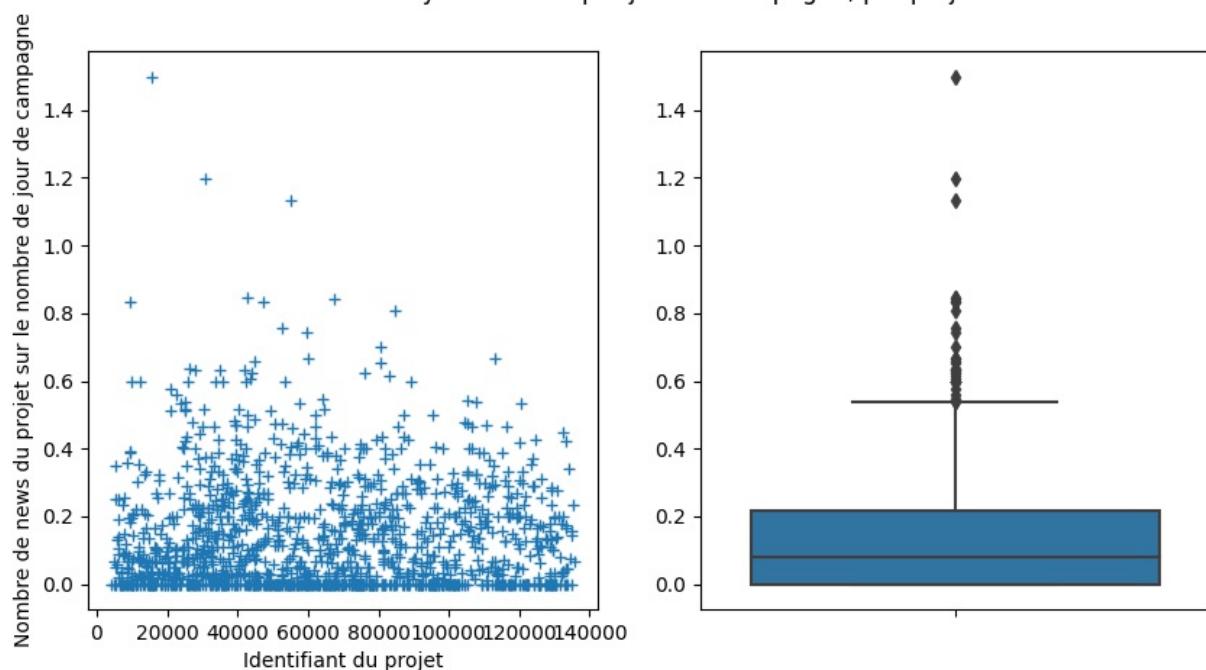
`SpearmanrResult(correlation=-0.10265741738840452, pvalue=0.0001246488187154437)`

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 moyen de news par jour de campagne, par projet

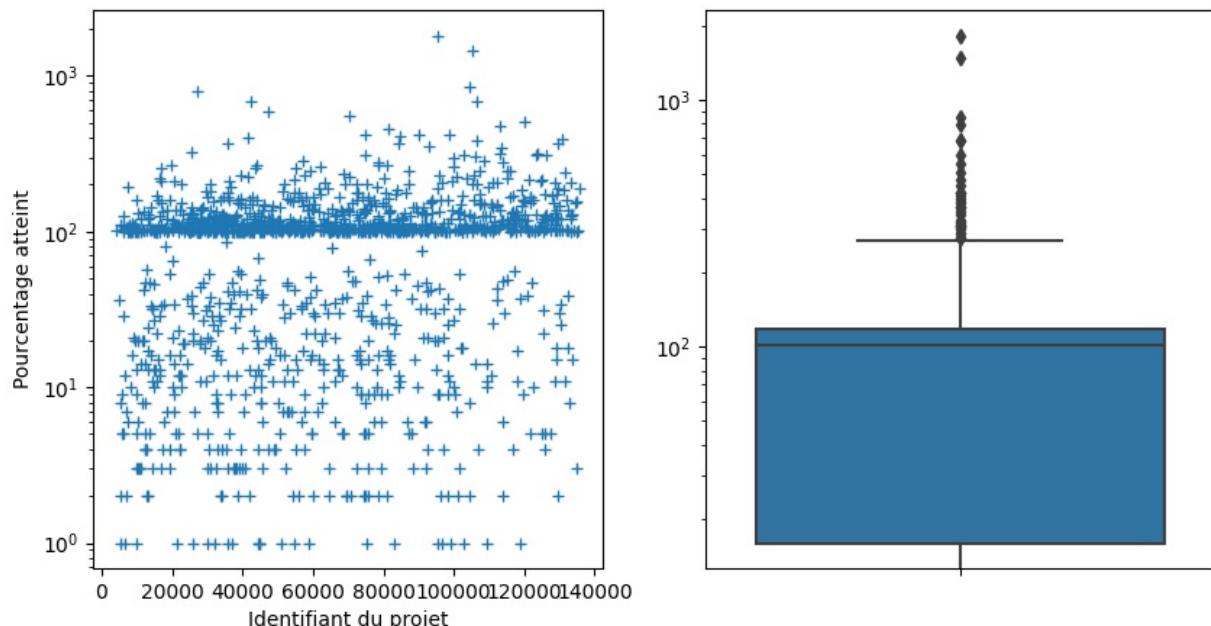


`SpearmanResult(correlation=0.588980340318175, pvalue=9.365828763801086e-131)`

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 que les campagnes qui ont le plus de succès fournisse plus de news par jour que les autres.

percent

Pourcentage du montant de succès demandé effectivement atteint par projet

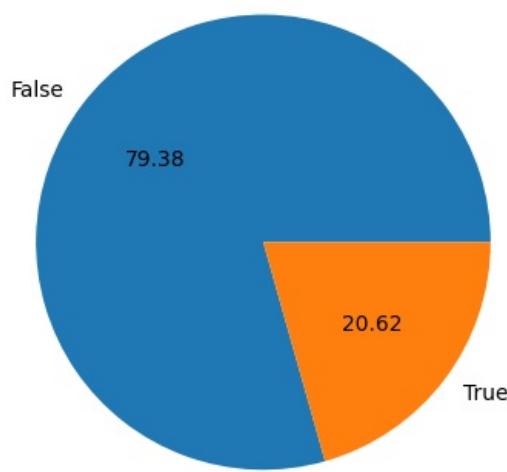


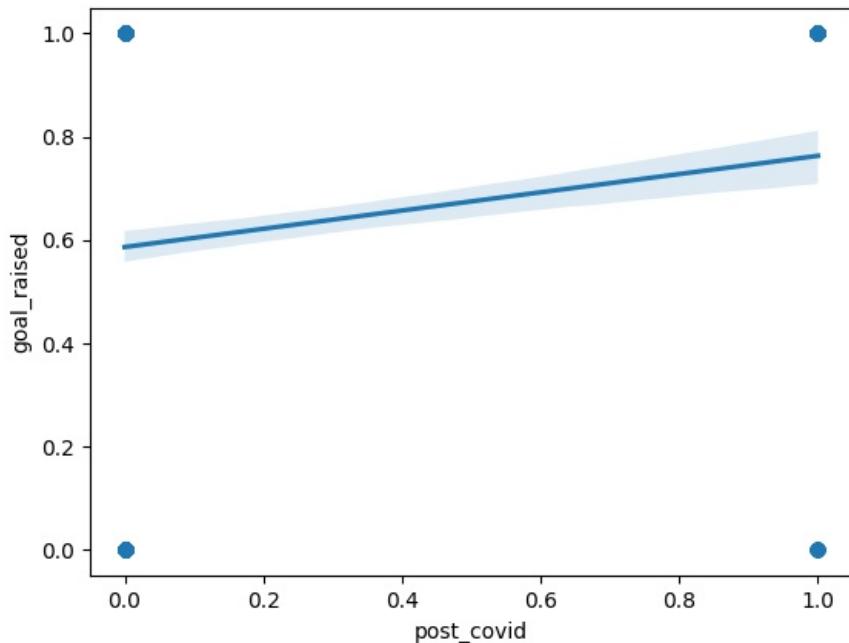
On note trois catégories de projets :

- 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)

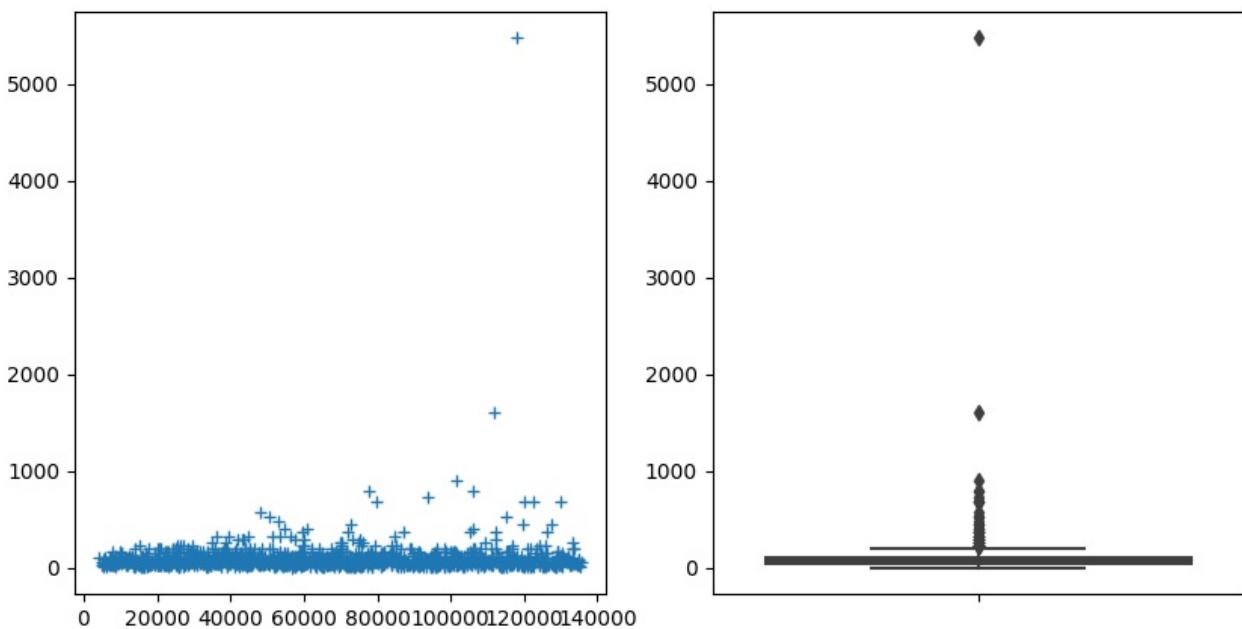


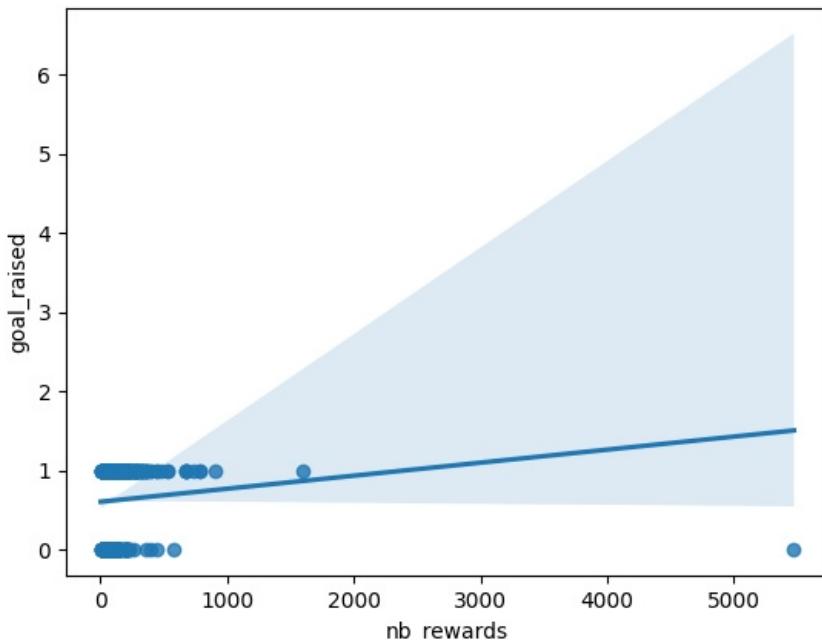


SpearmanResult(correlation=0.14744296784460775, pvalue=3.270302481736463e-08)

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.

nb_rewards





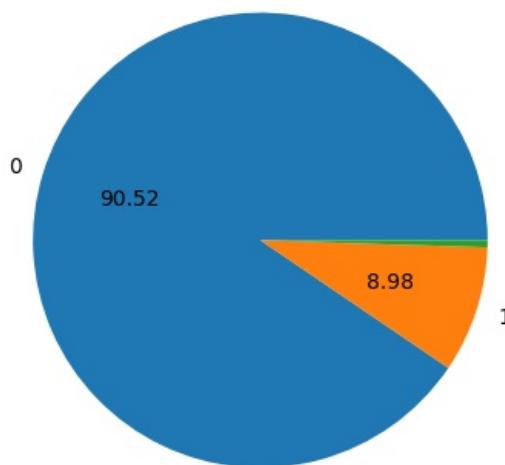
SpearmanResult(correlation=0.181219416894131, pvalue=9.651409030460442e-12)

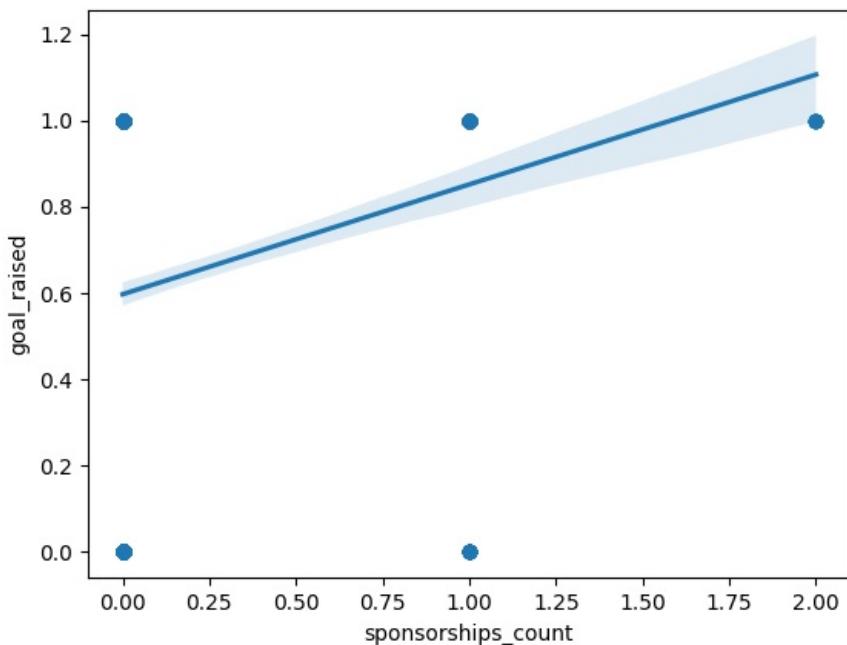
Une majorité de projets a entre 0 et 50 tiers de rewards. Les projets réussis ont une moyenne de nombre de tiers légèrement plus élevée que les projets ratés, mais cela ne semble pas réellement discriminer entre succès et échec d'un projet.

???

sponsorships_count

Représentation du nombre de sponsors parmi les projets.

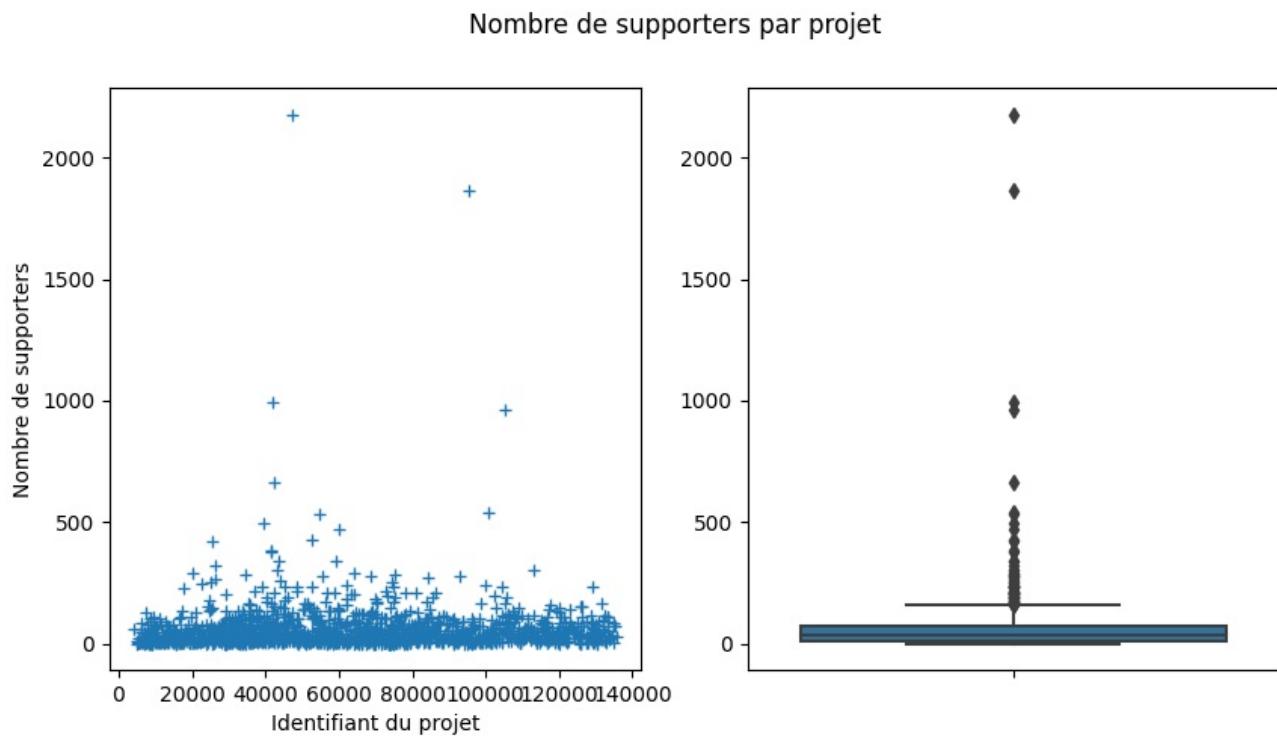




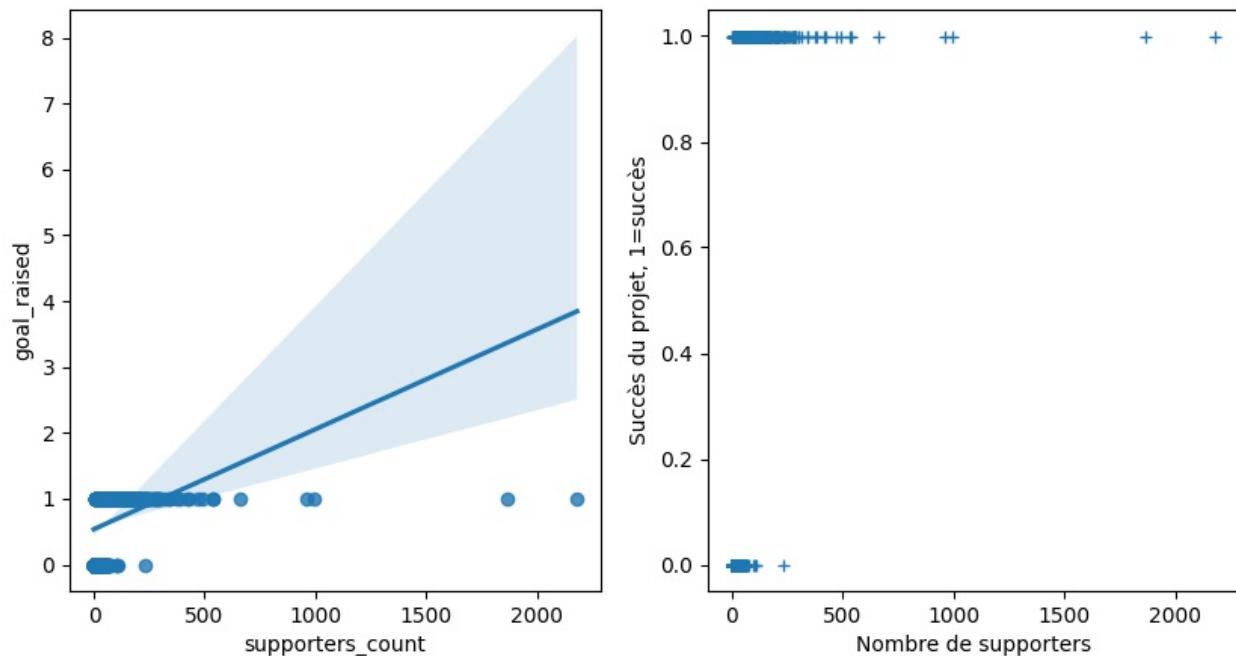
SpearmanResult(correlation=0.16625274211557997, pvalue=4.355729505598115e-10)

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



Succès du projet en fonction du nombre de supporters

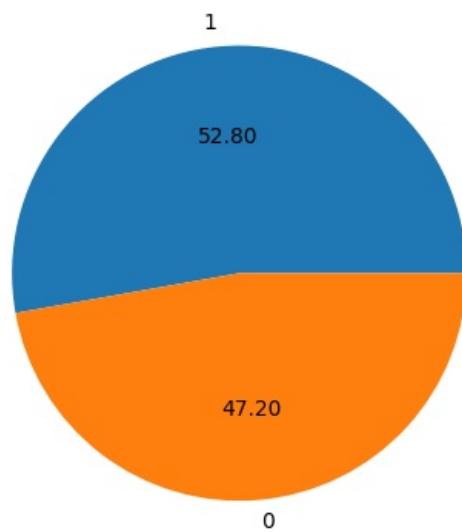


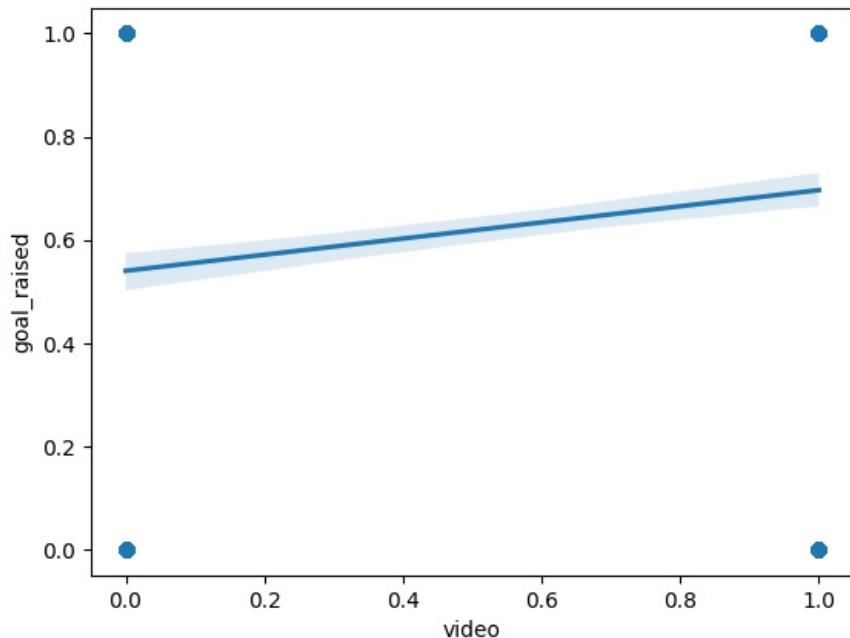
`SpearmanResult(correlation=0.7372606235631092, pvalue=5.4779730769548455e-239)`

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

video

Présence d'une vidéo dans le projet



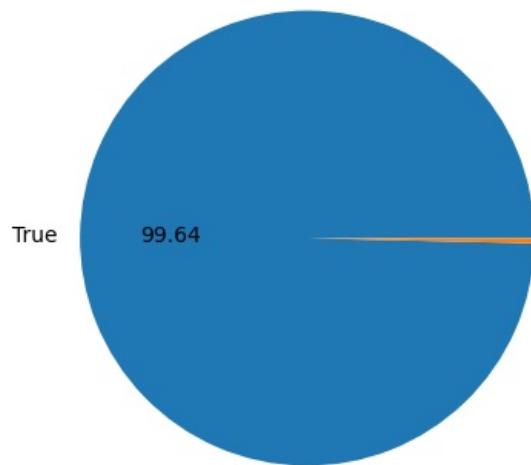


SpearmanResult(correlation=0.16095167132618254, pvalue=1.5502669766114702e-09)

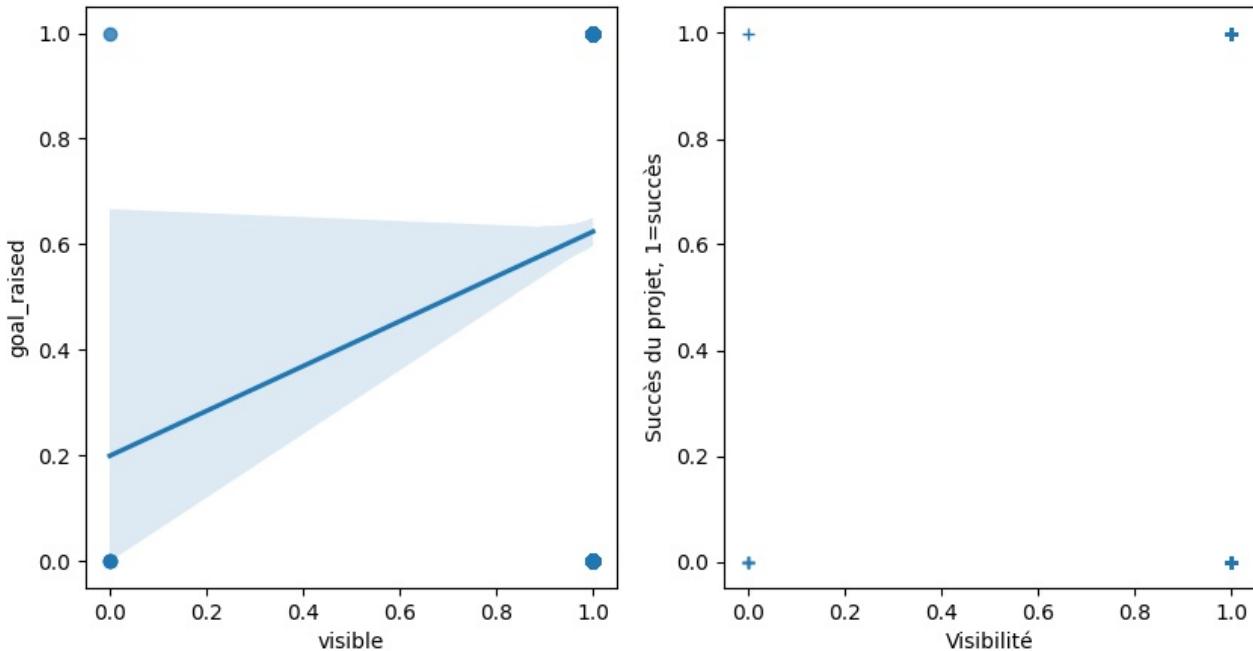
Il ne semble pas y avoir de corrélation entre le fait de posséder une vidéo et le succès de la campagne

visible

Indexation sur les moteurs de recherches du projet



Succès du projet en fonction de sa visibilité

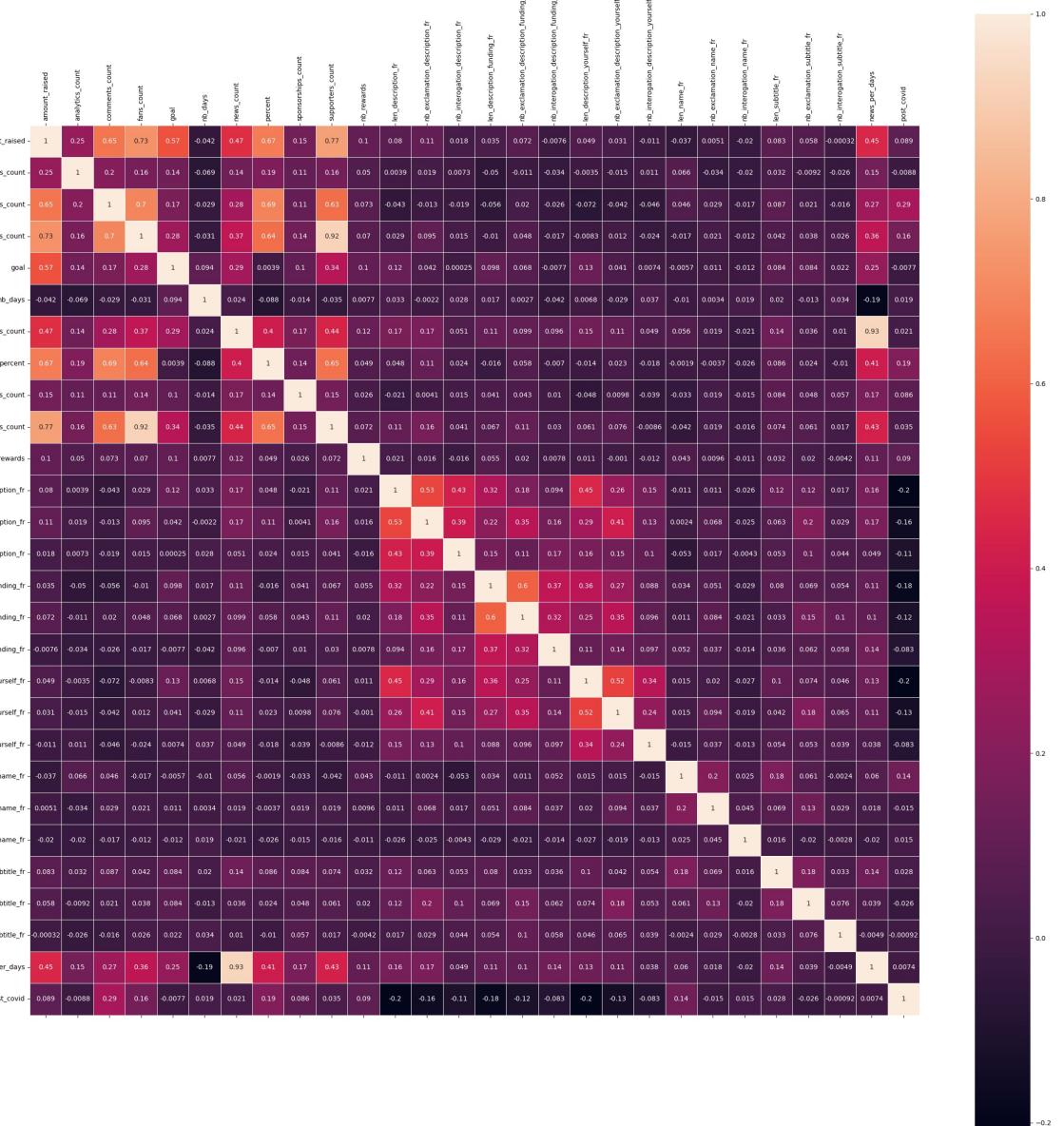


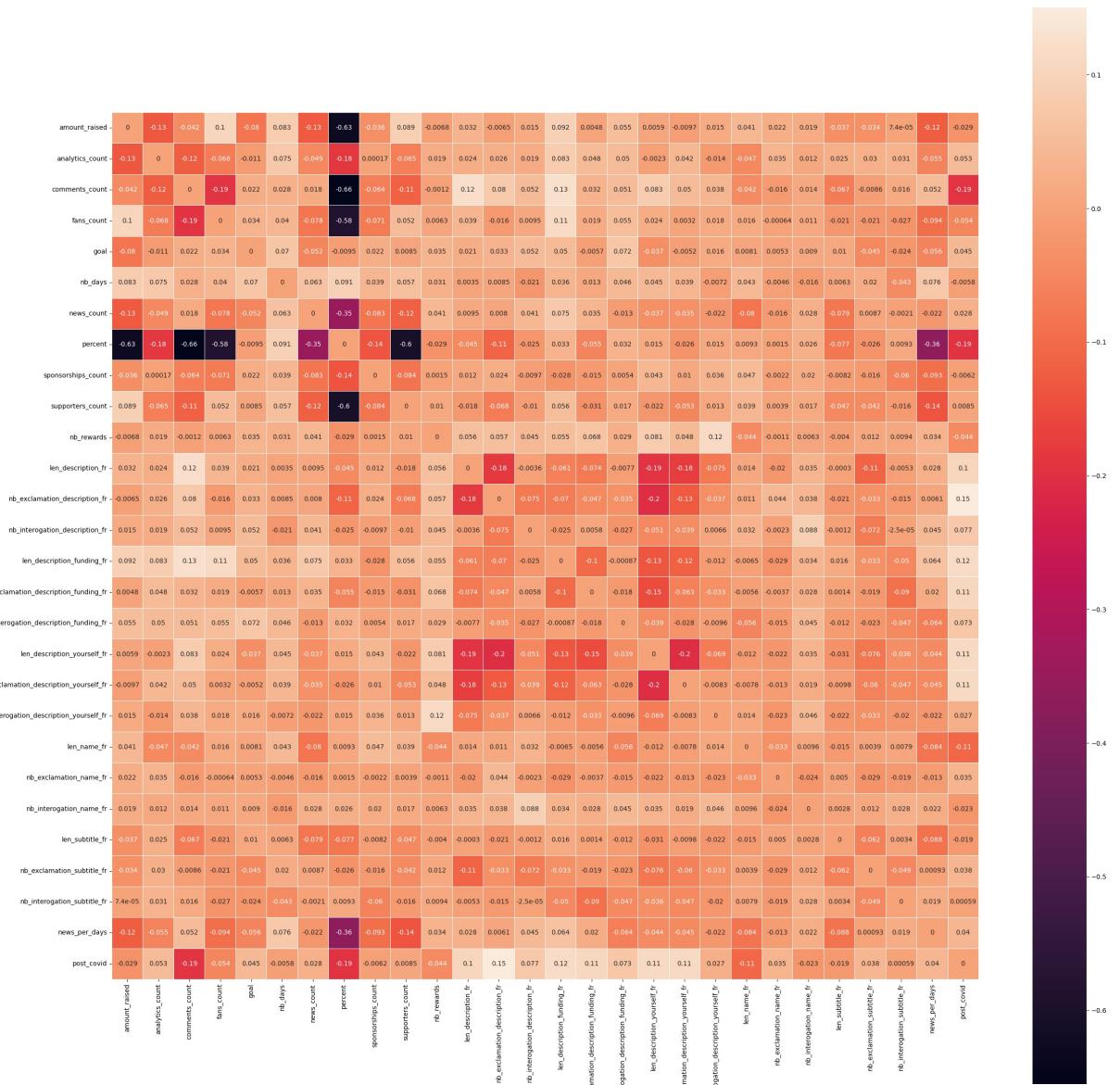
`SpearmanResult(correlation=0.052381487721841535, pvalue=0.05071142084030411)`

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

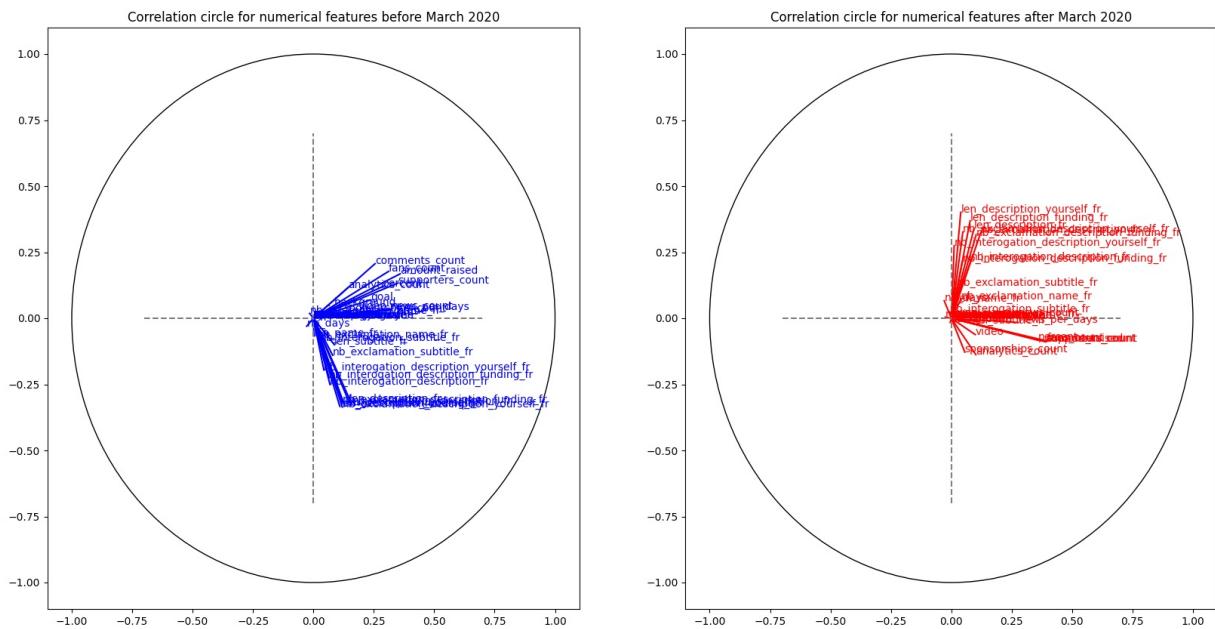
Matrice de corrélation





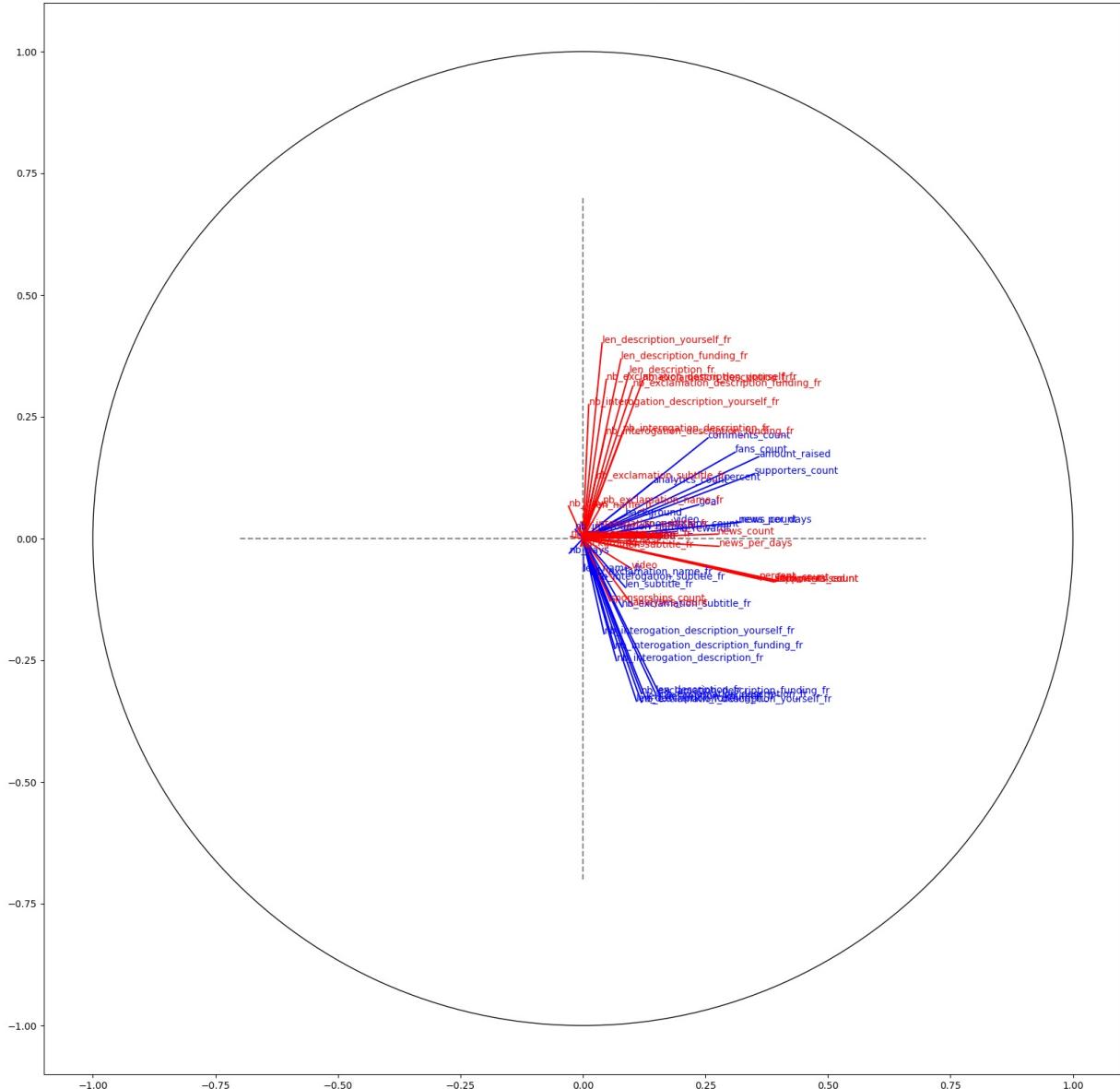
On explore les corrélations entre les différentes variables numériques du dataset. Le montant récolté est assez logiquement fortement corrélé aux nombres de fans, de supporter et de commentaires. Il semble également que mettre des news sur un projet soit une des façons de susciter des commentaires.

PCA



La PCA est une autre façon de représenter les corrélations entre nos différentes variables. Avant Mars 2020 (date que l'on considère dans notre cas comme date d'impact du covid sur Ulule), le nombre de fans, de supporters, de commentaires ainsi que le montant récolté étaient moins corrélés entre eux qu'à partir de Mars 2020, ce qui témoigne d'un aspect communautaire plus important avec le covid.

Correlation circle for numerical features before (blue) and after (red) March 2020



???