

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.

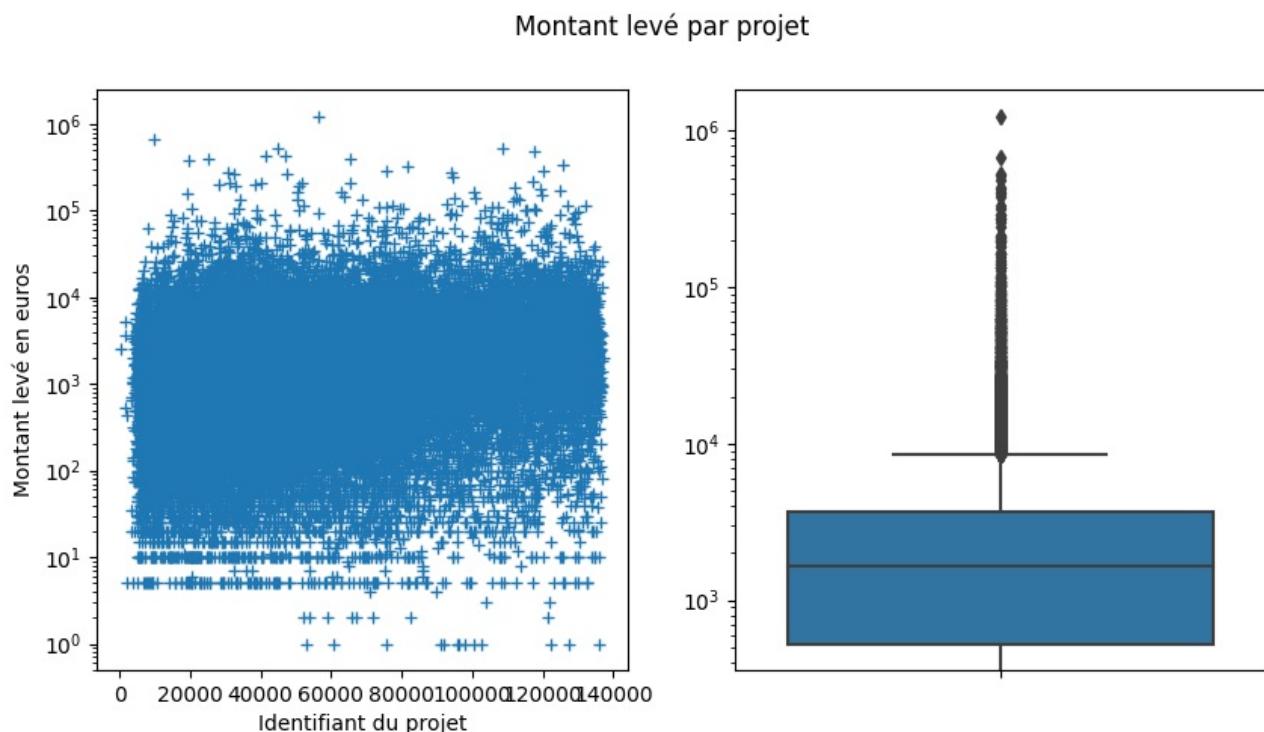
Dans la mesure où Ulule se rémunère en touchant une commission sur les projets ayant fonctionné, le site a tout intérêt à ce qu'un maximum de projets réussissent.

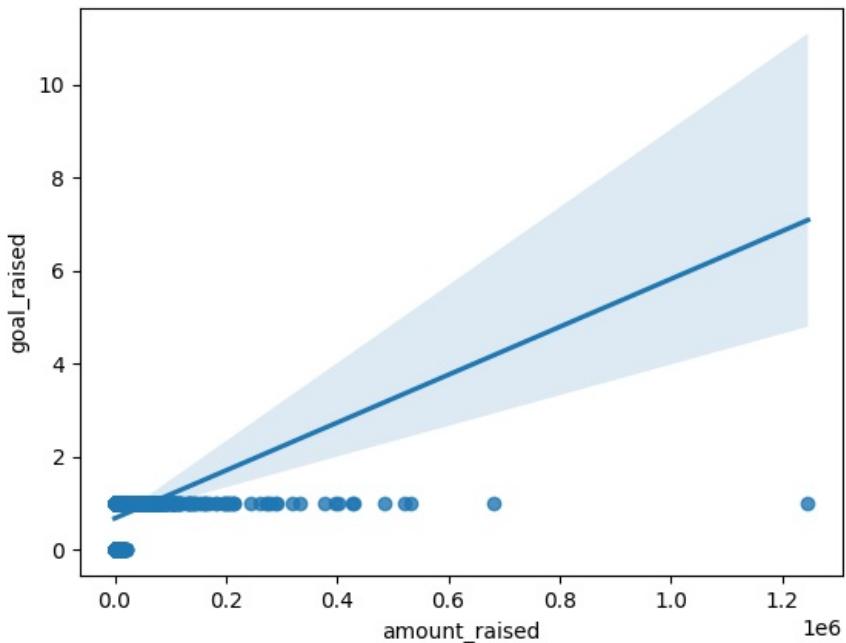
Note :

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.

Chargement des données du CSV pré-nettoyé

Statistiques descriptives

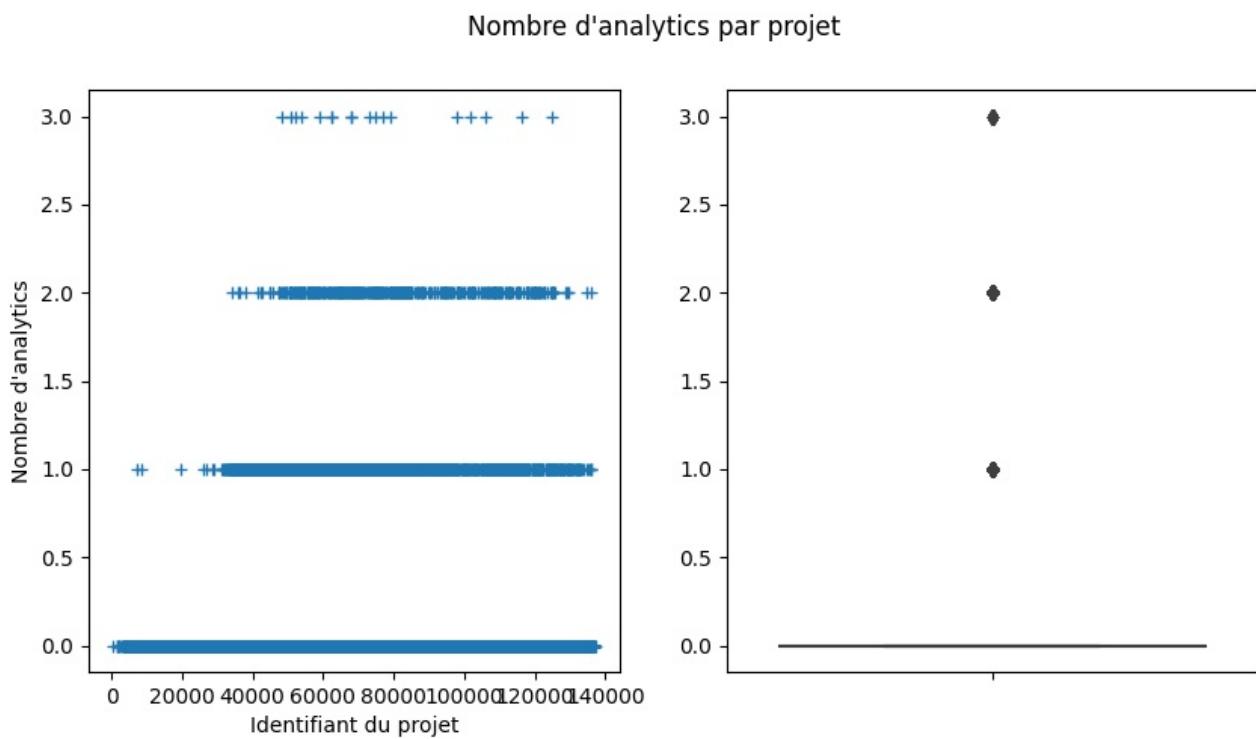




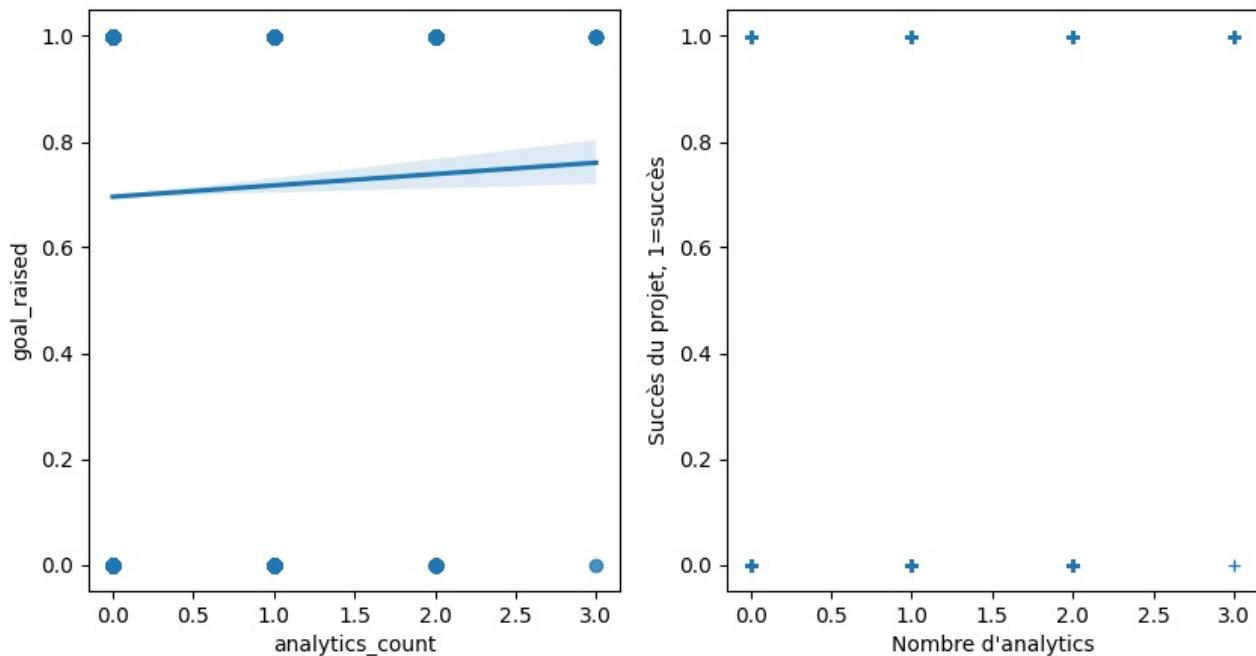
```
SpearmannResult(correlation=0.6605276041694639, pvalue=0.0)
```

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



Succès du projet en fonction du nombre d'analytics

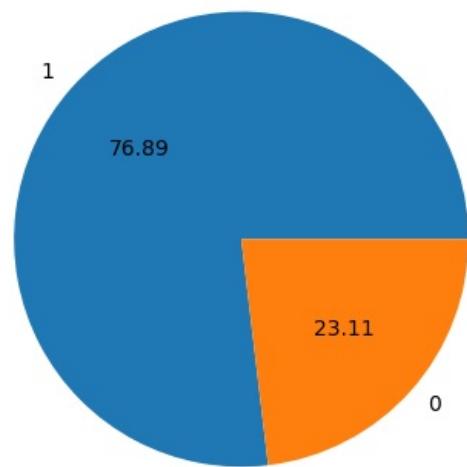


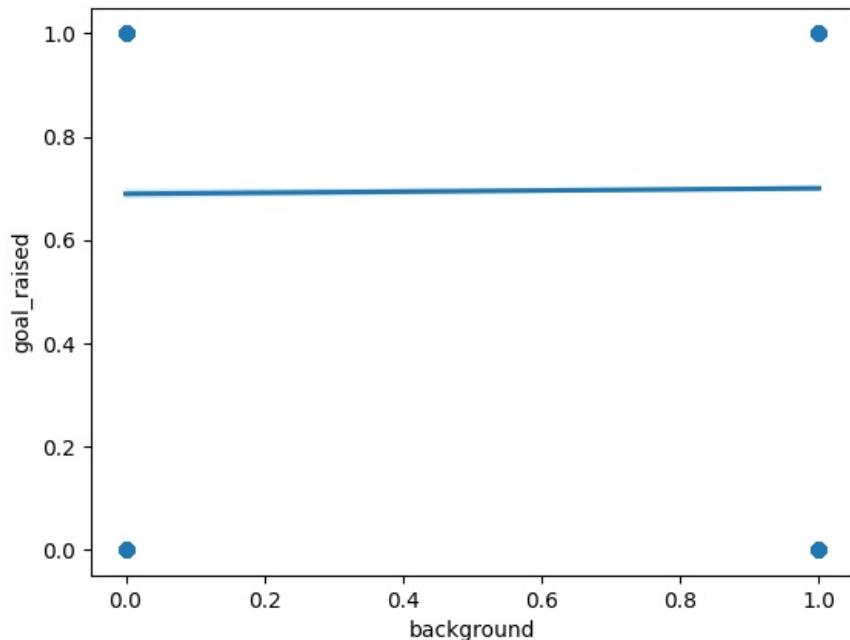
`SpearmanResult(correlation=0.00653046420759979, pvalue=0.20824585283482497)`

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



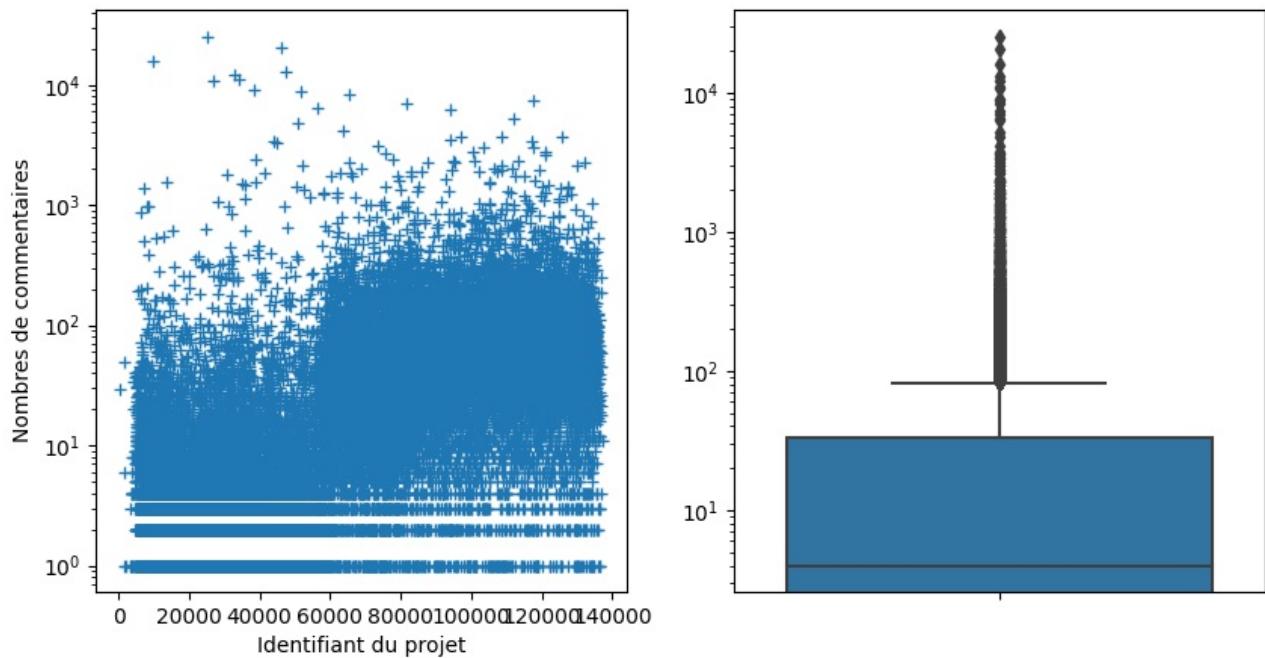


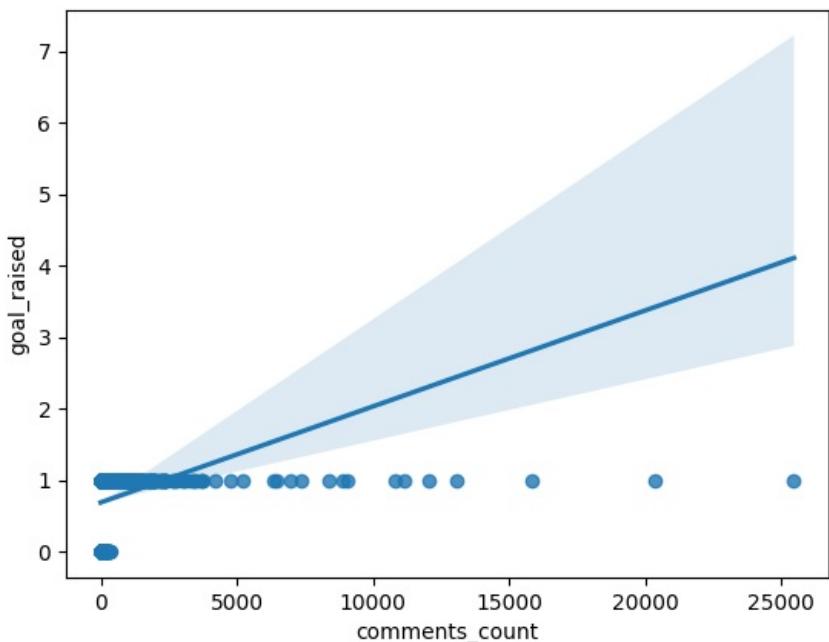
SpearmanResult(correlation=0.009848313025372124, pvalue=0.057726360936618135)

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



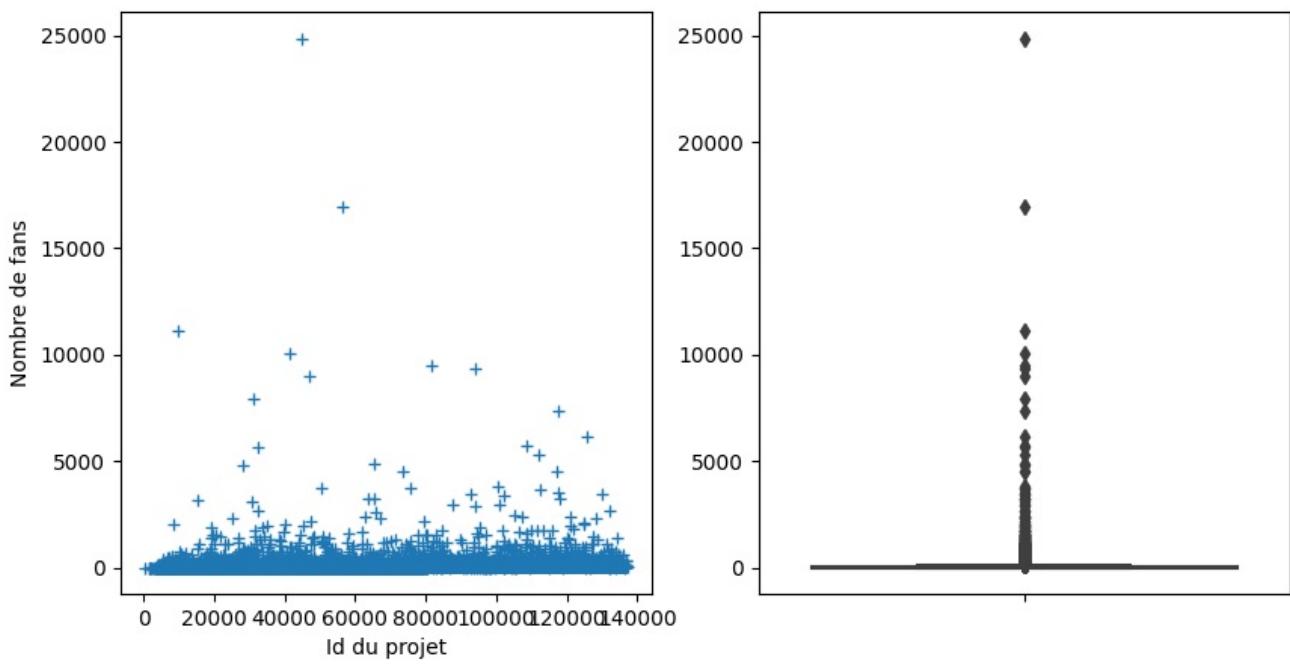


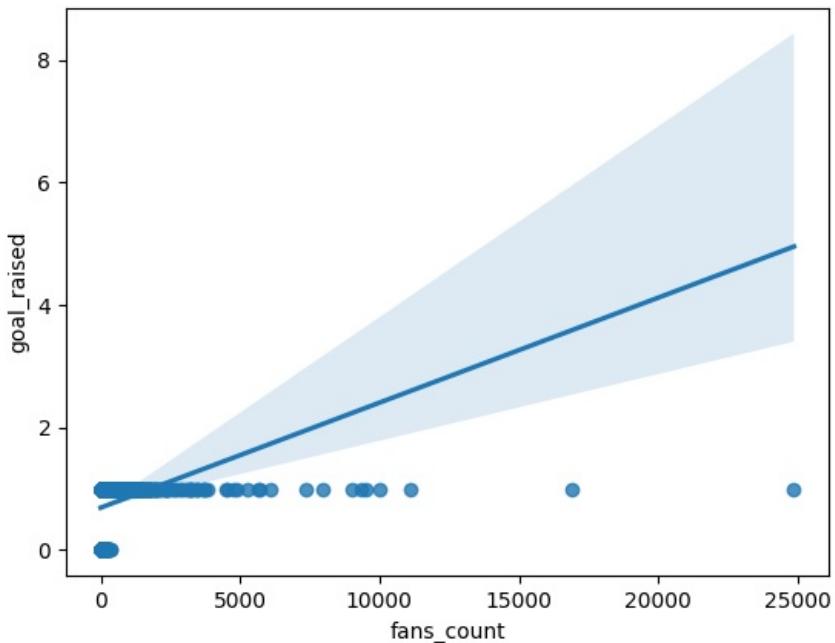
`SpearmannResult(correlation=0.3442342916010207, pvalue=0.0)`

Les projets reçoivent globalement assez peu de commentaires. La plupart des projets qui ont beaucoup de commentaires, ont bien fonctionné. Il 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



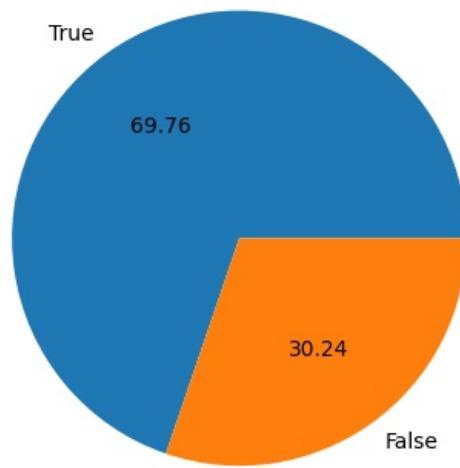


`SpearmannResult(correlation=0.2378731377730896, pvalue=0.0)`

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

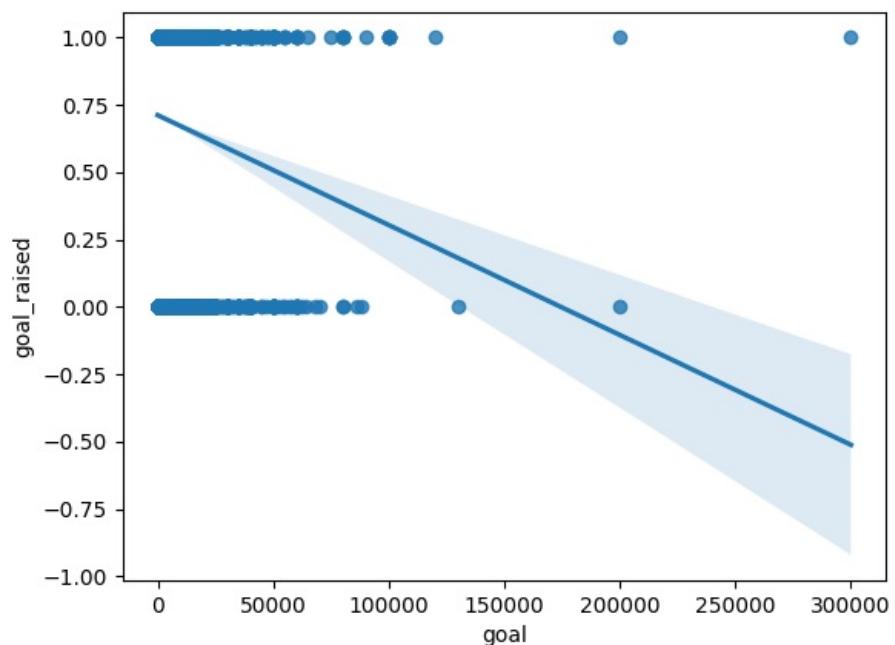
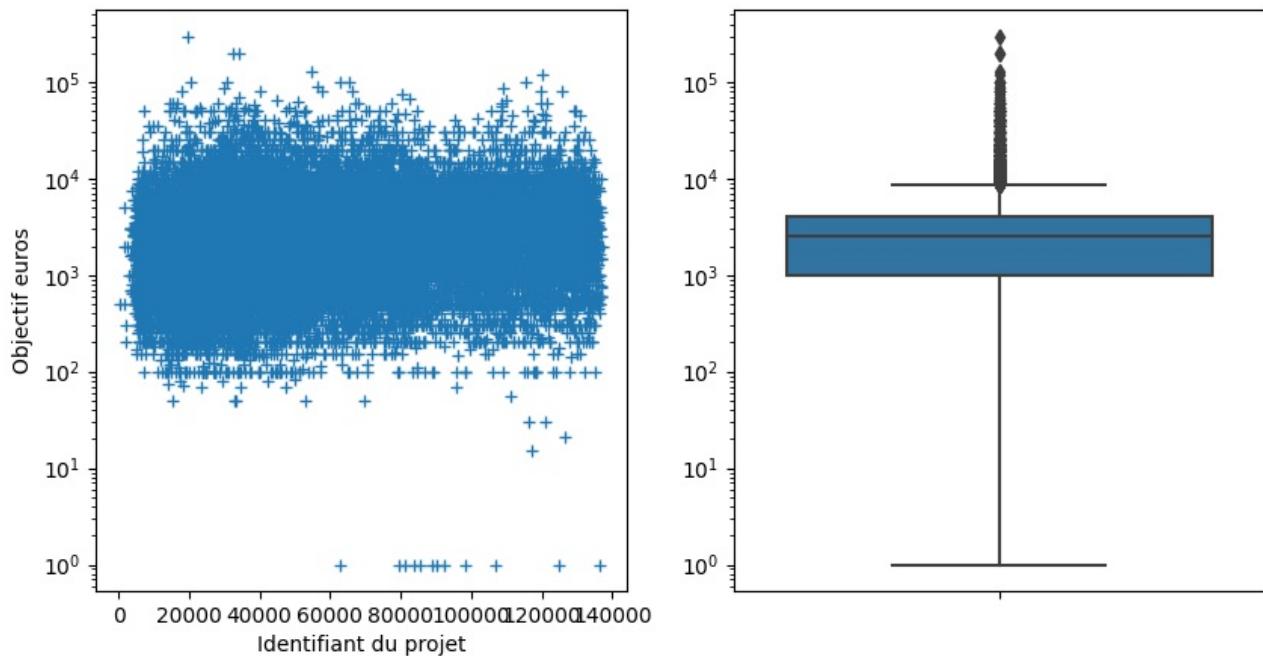
Représentation du taux de succès des projets, succès=True



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

goal

Objectif de validation du projet

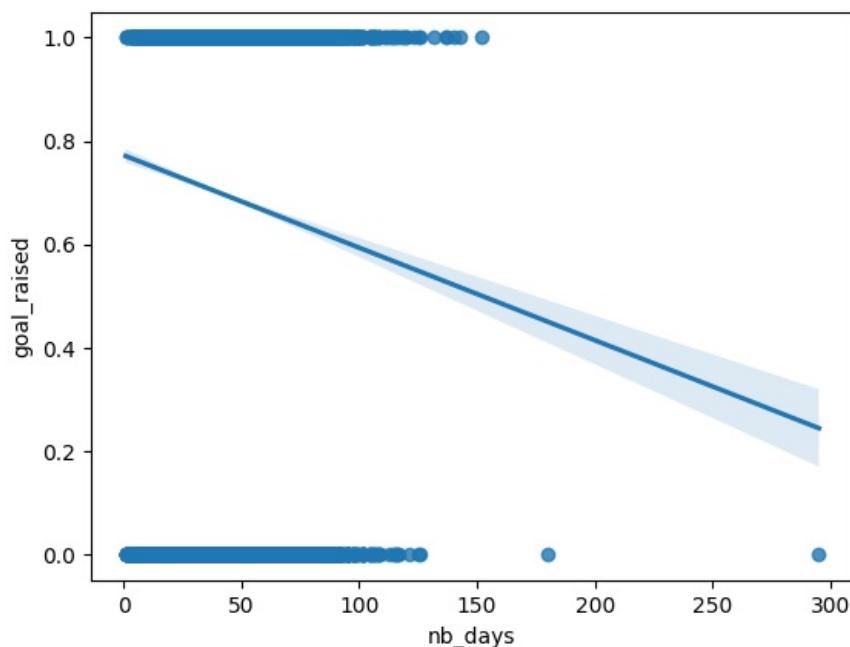
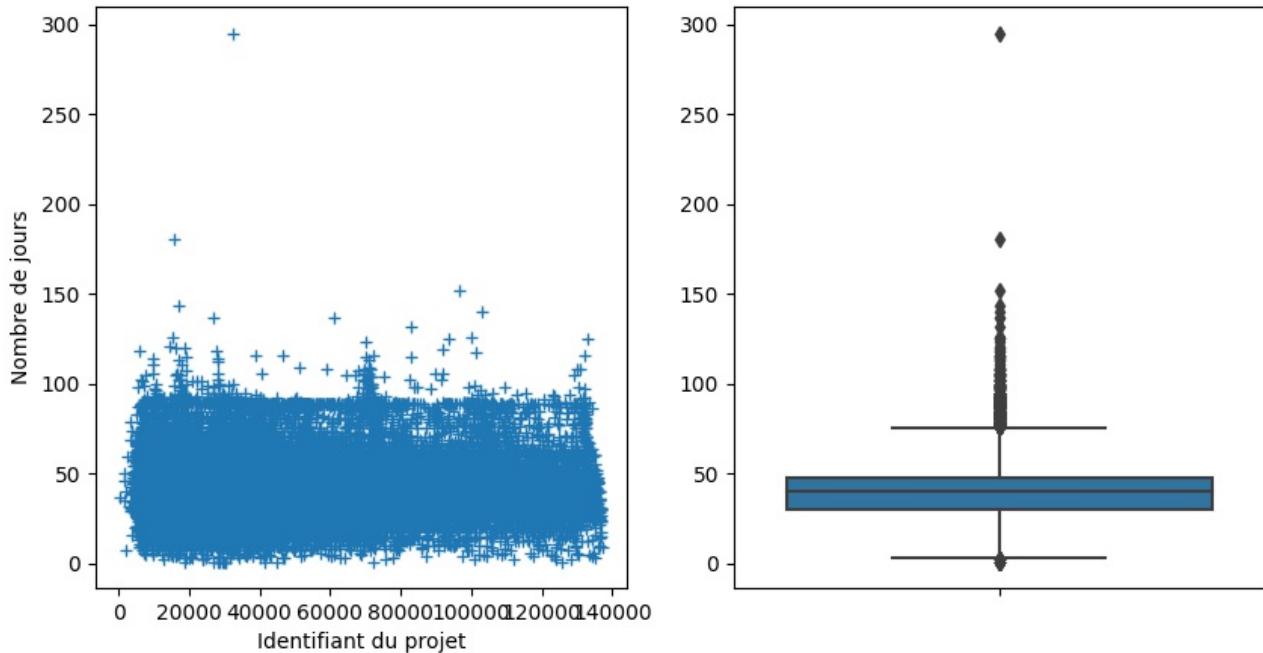


```
SpearmanResult(correlation=-0.07454856591968433, pvalue=6.410987034414001e-47)
```

La majorité des projets semblent 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

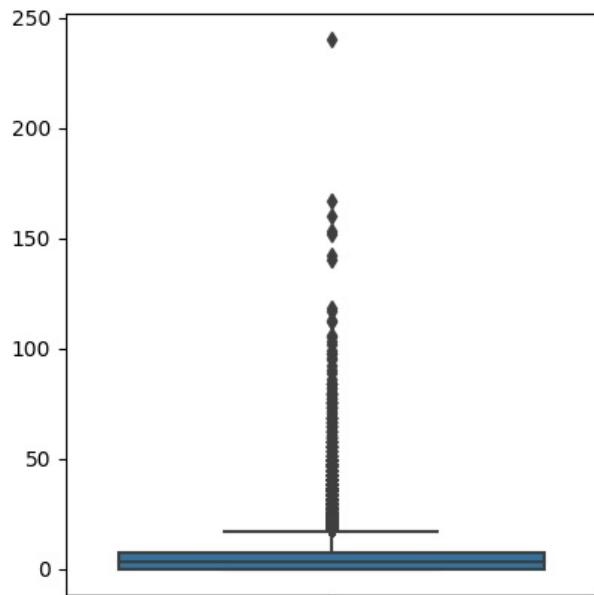
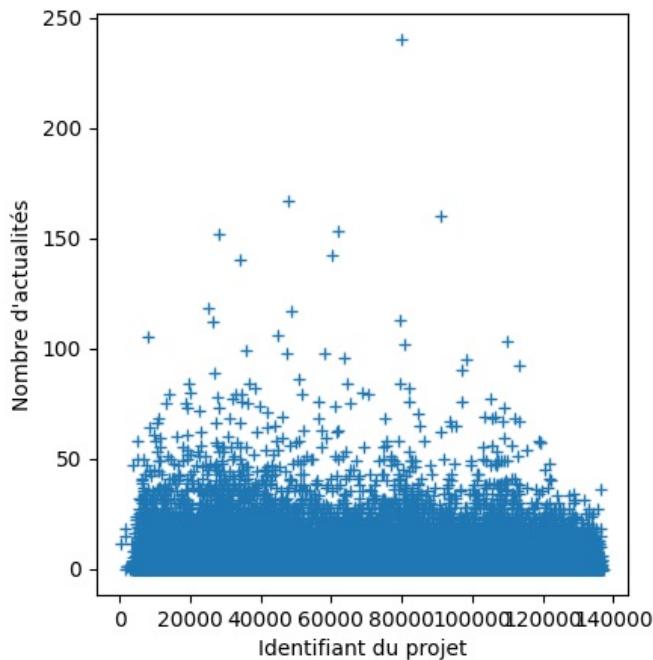


SpearmanResult(correlation=-0.06087830903719856, pvalue=7.775863519983873e-32)

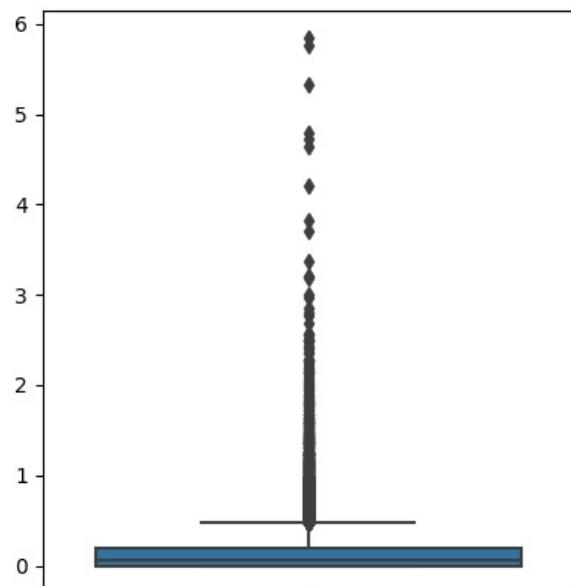
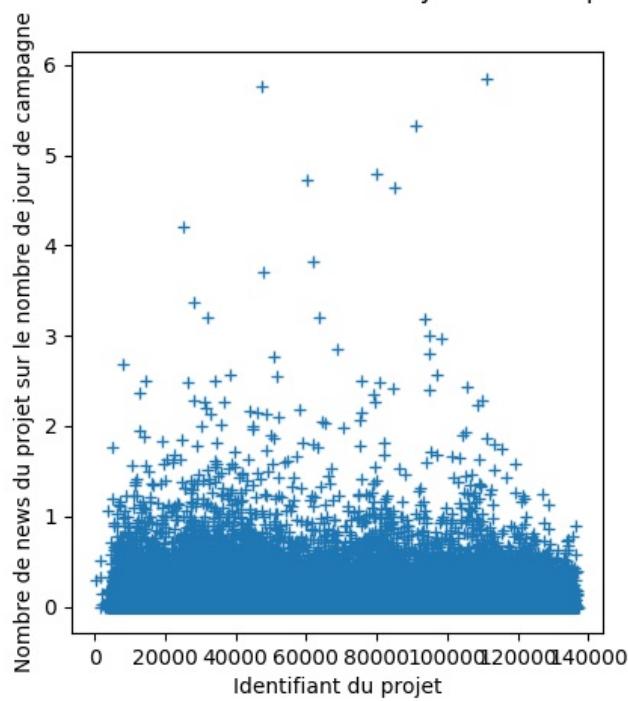
Les campagnes semblent durer plus ou moins un mois en grande majorité. Il ne semble n'y avoir aucun corrélation entre la durée de la campagne et son succès.

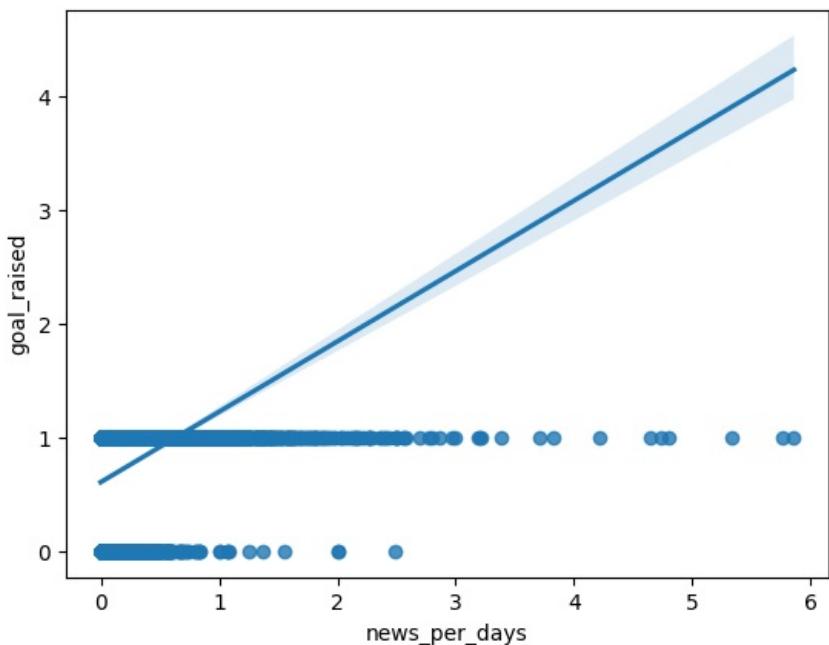
news_count

Nombre d'actualités postées par projet



Nombre moyen de news par jour de campagne, par projet



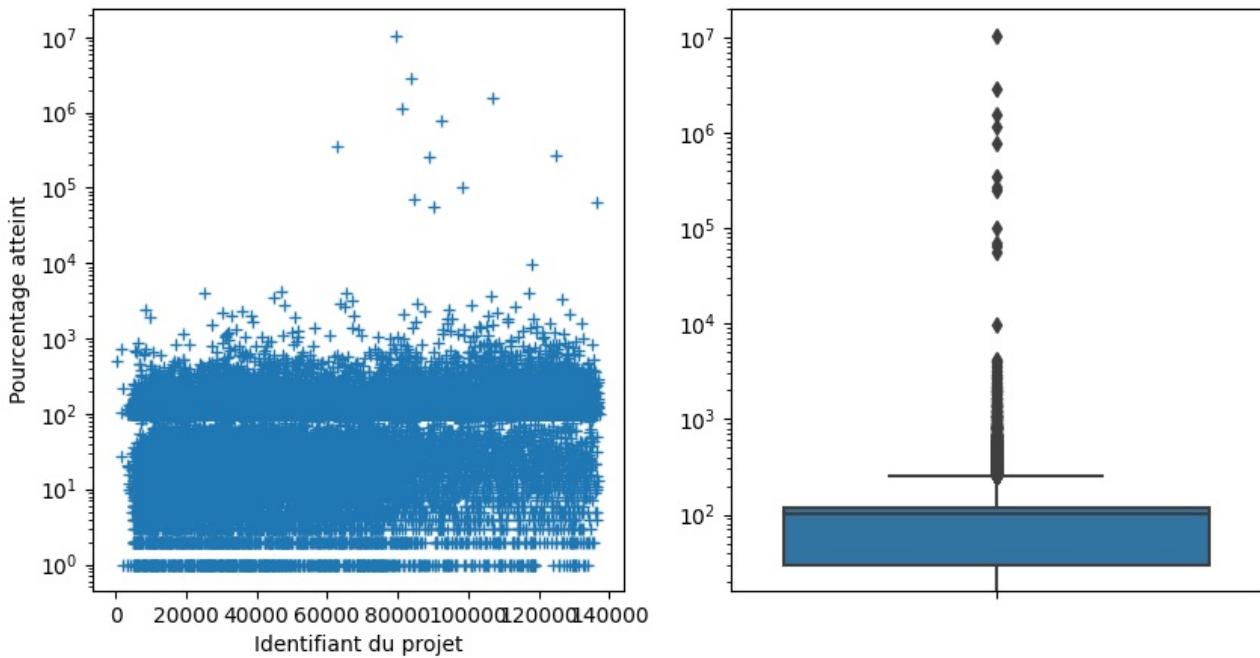


`SpearmannResult(correlation=0.45036615456561185, pvalue=0.0)`

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 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

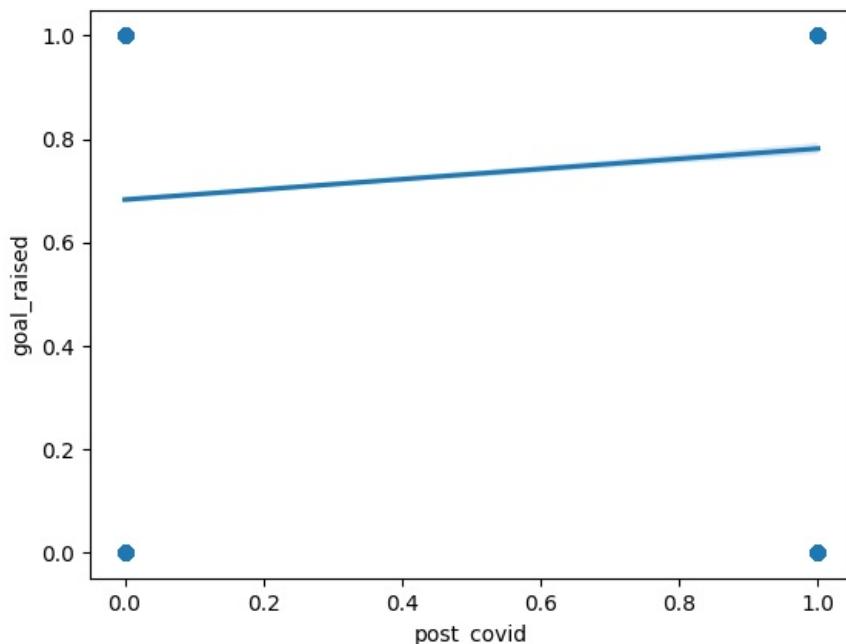
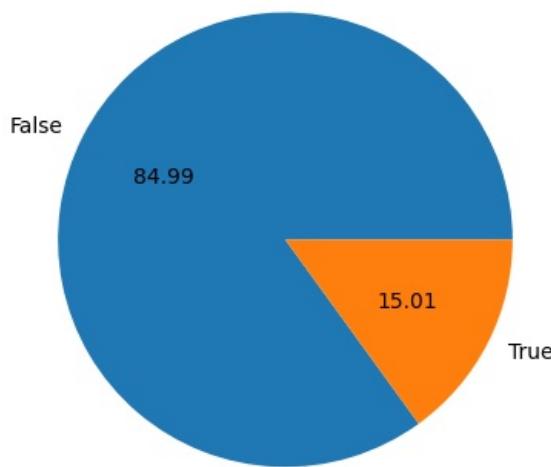


On note trois catégories de projets :

- 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

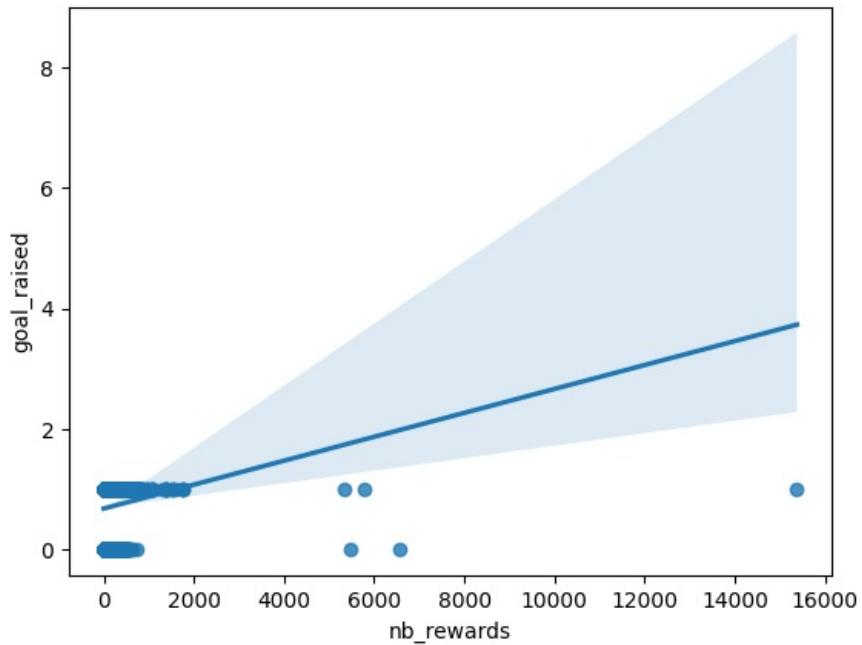
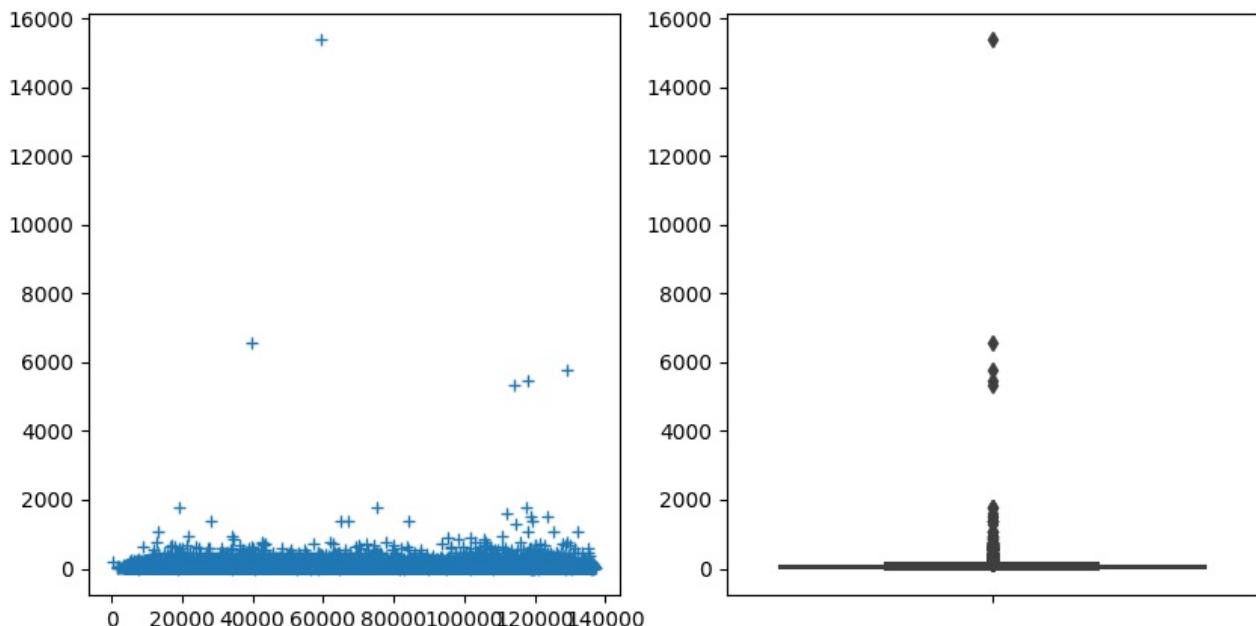
Répartition des projets avant et après Mars 2020 (après=True)



SpearmanResult(correlation=0.07676912106516919, pvalue=1.1733117696819341e-49)

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



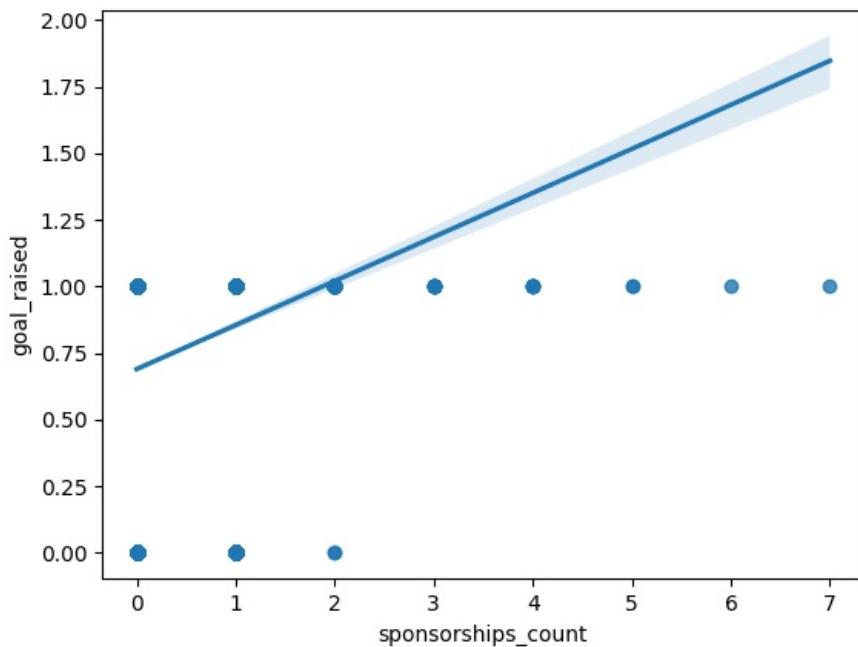
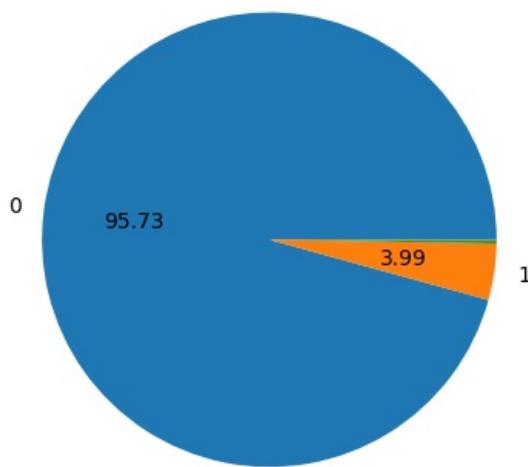
SpearmanResult(correlation=0.11169033379810844, pvalue=2.240426896268747e-103)

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.

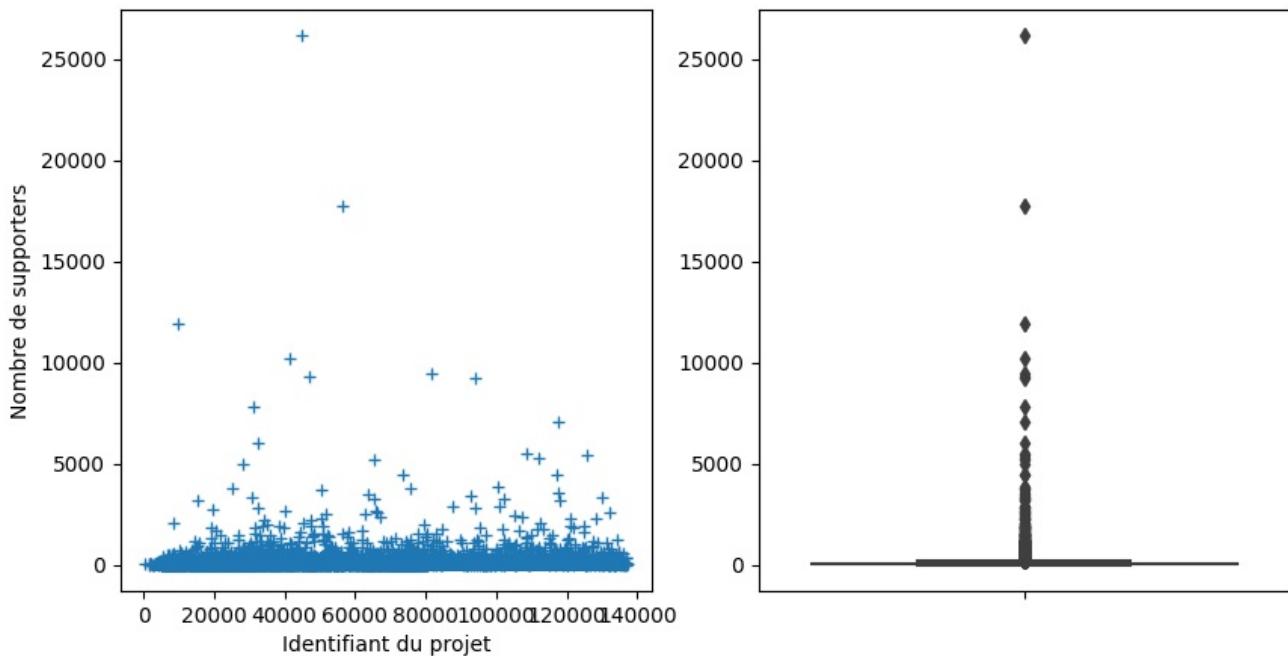


SpearmanResult(correlation=0.08698010948281582, pvalue=2.769548418451433e-63)

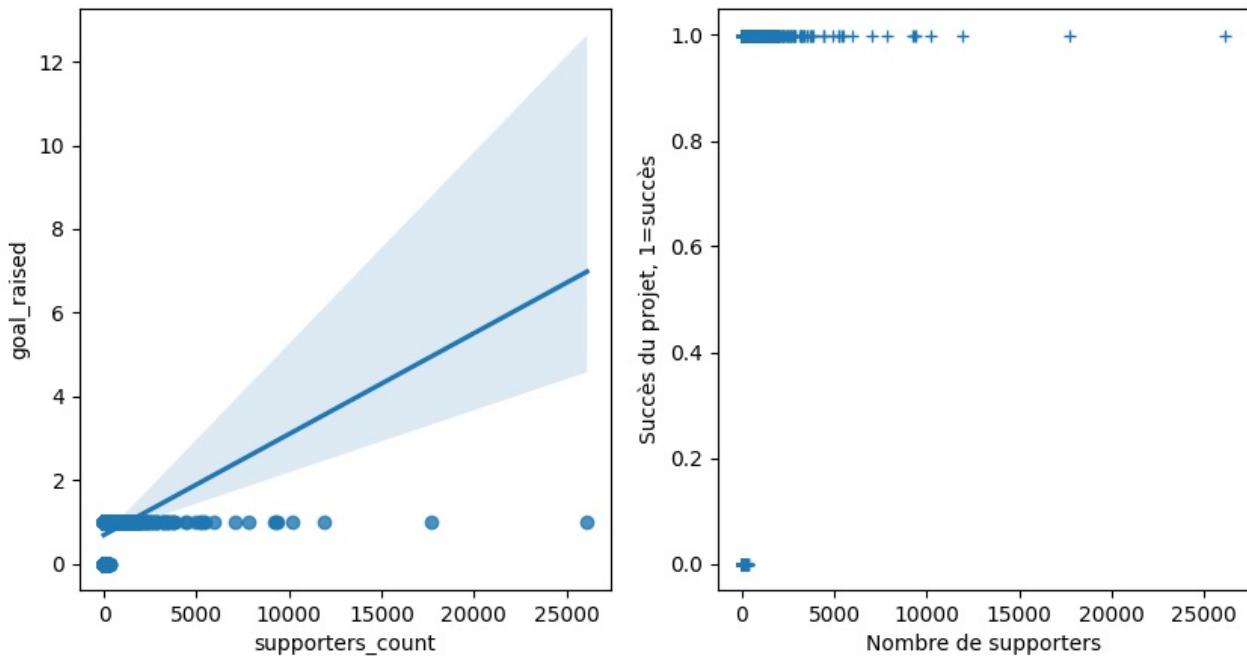
L'obtention de sponsors semble être un phénomène très minoritaire. Il ne semble pas y avoir de corrélation entre le succès de la campagne et la présence de sponsors.

supporters_count

Nombre de supporters par projet



Succès du projet en fonction du nombre de supporters

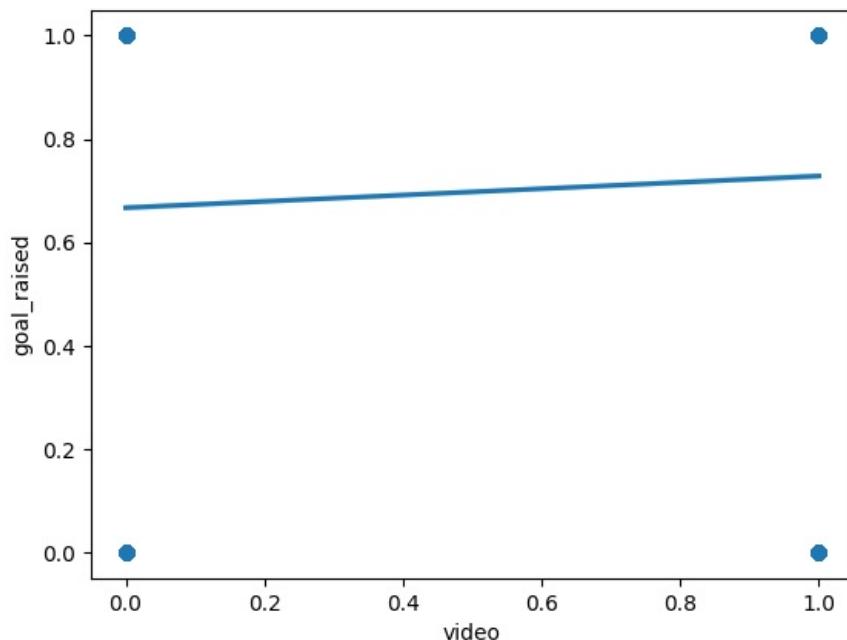
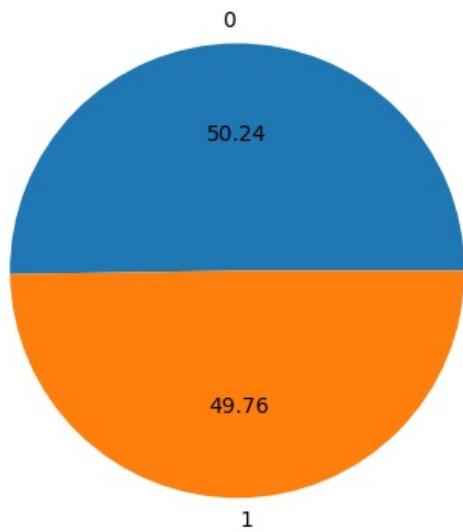


`SpearmanrResult(correlation=0.6697024008514539, pvalue=0.0)`

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

video

Présence d'une vidéo dans le projet

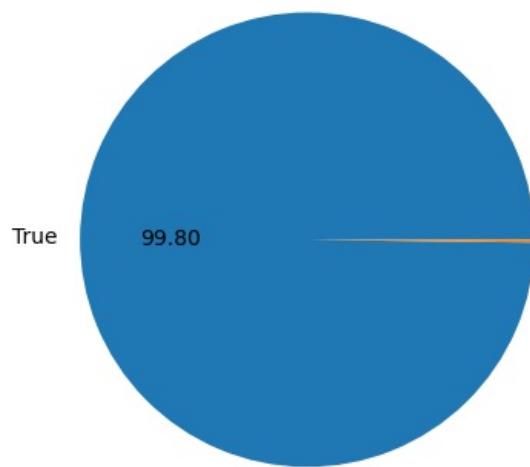


SpearmanrResult(correlation=0.06673662301049162, pvalue=6.289488654579654e-38)

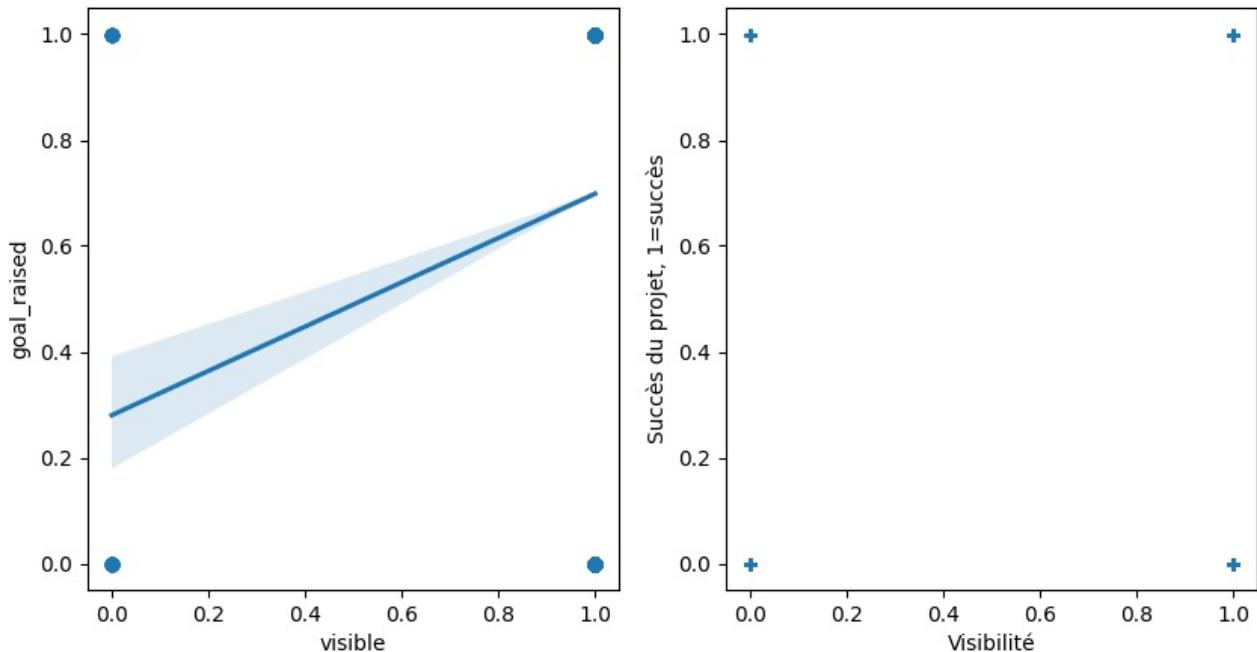
Il ne semble pas y avoir de corrélation entre le fait de posséder une vidéo et le succès de la campagne

visible

Indexation sur les moteurs de recherches du projet



Succès du projet en fonction de sa visibilité

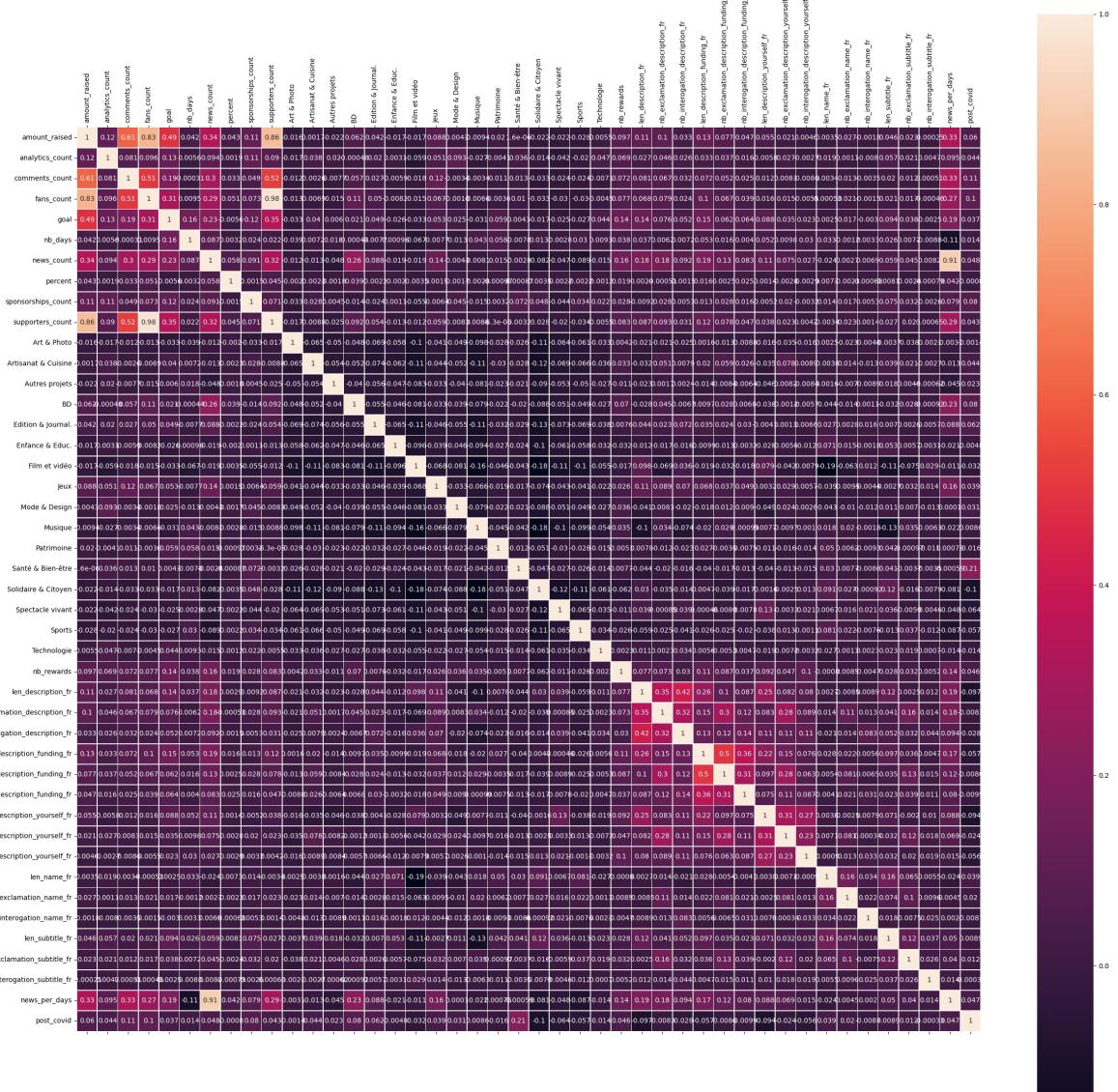


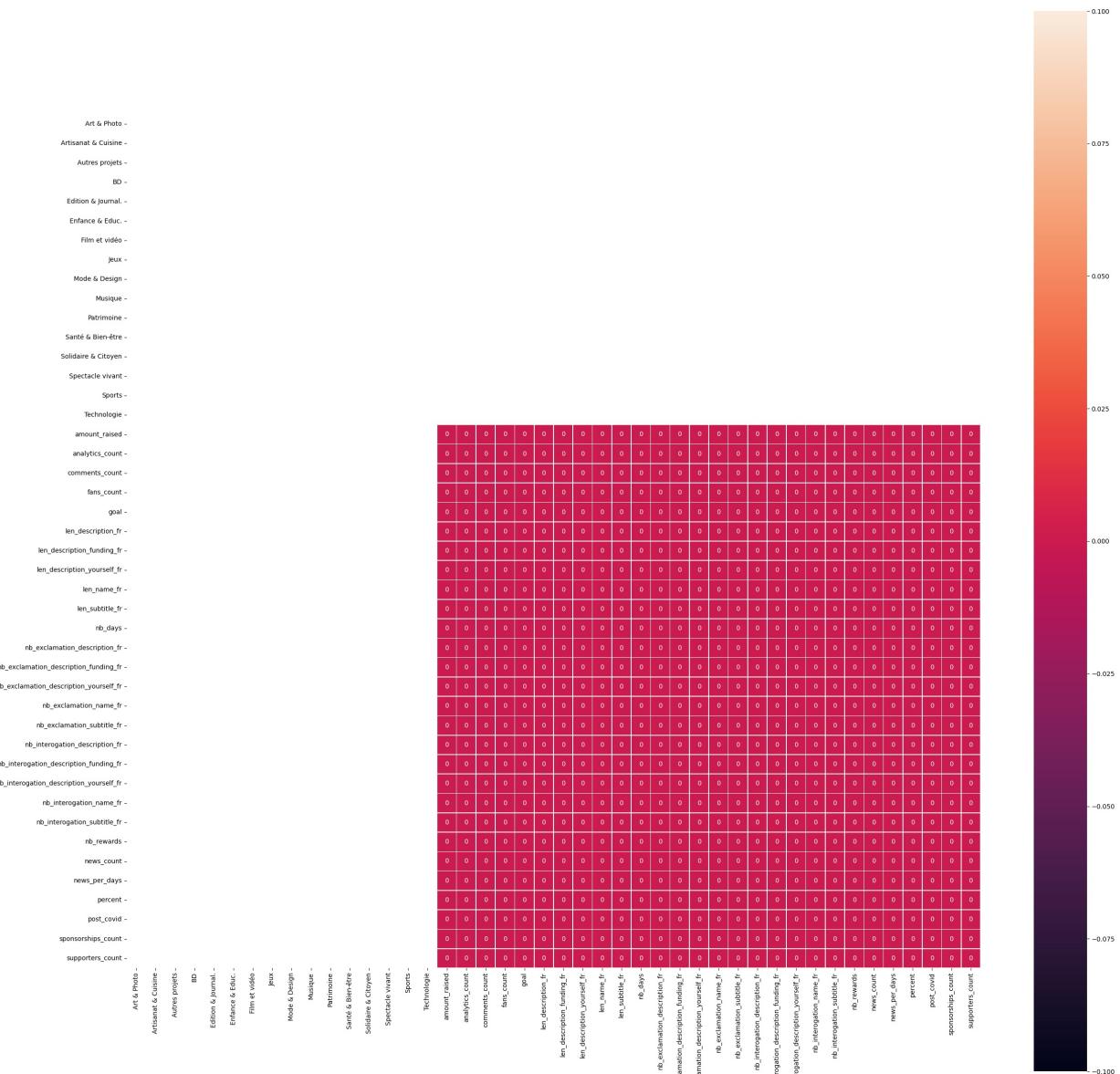
SpearmanResult(correlation=0.04090737720076117, pvalue=3.1212076817719414e-15)

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

Autres stats ? NOMMER CA AUTREMENT SVP

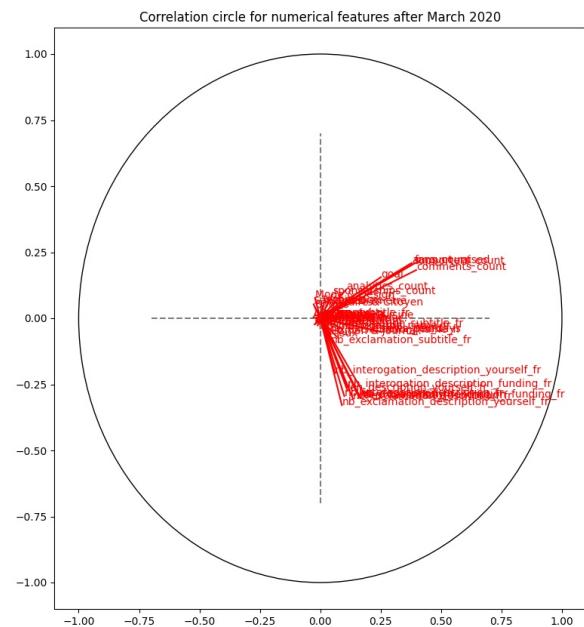
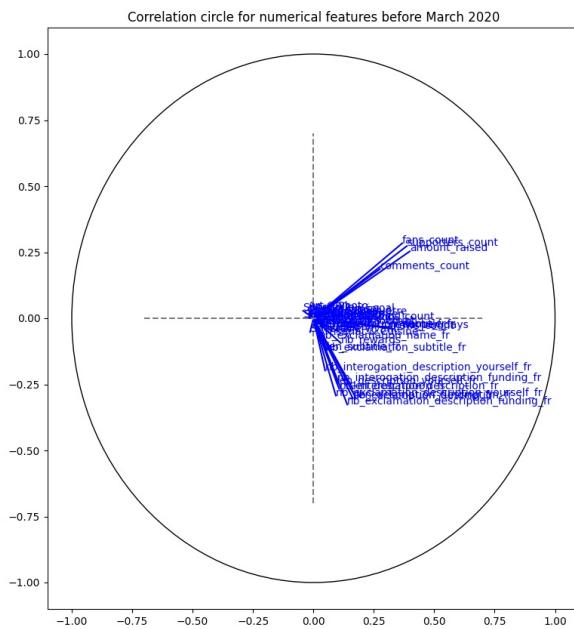
Matrice de corrélation



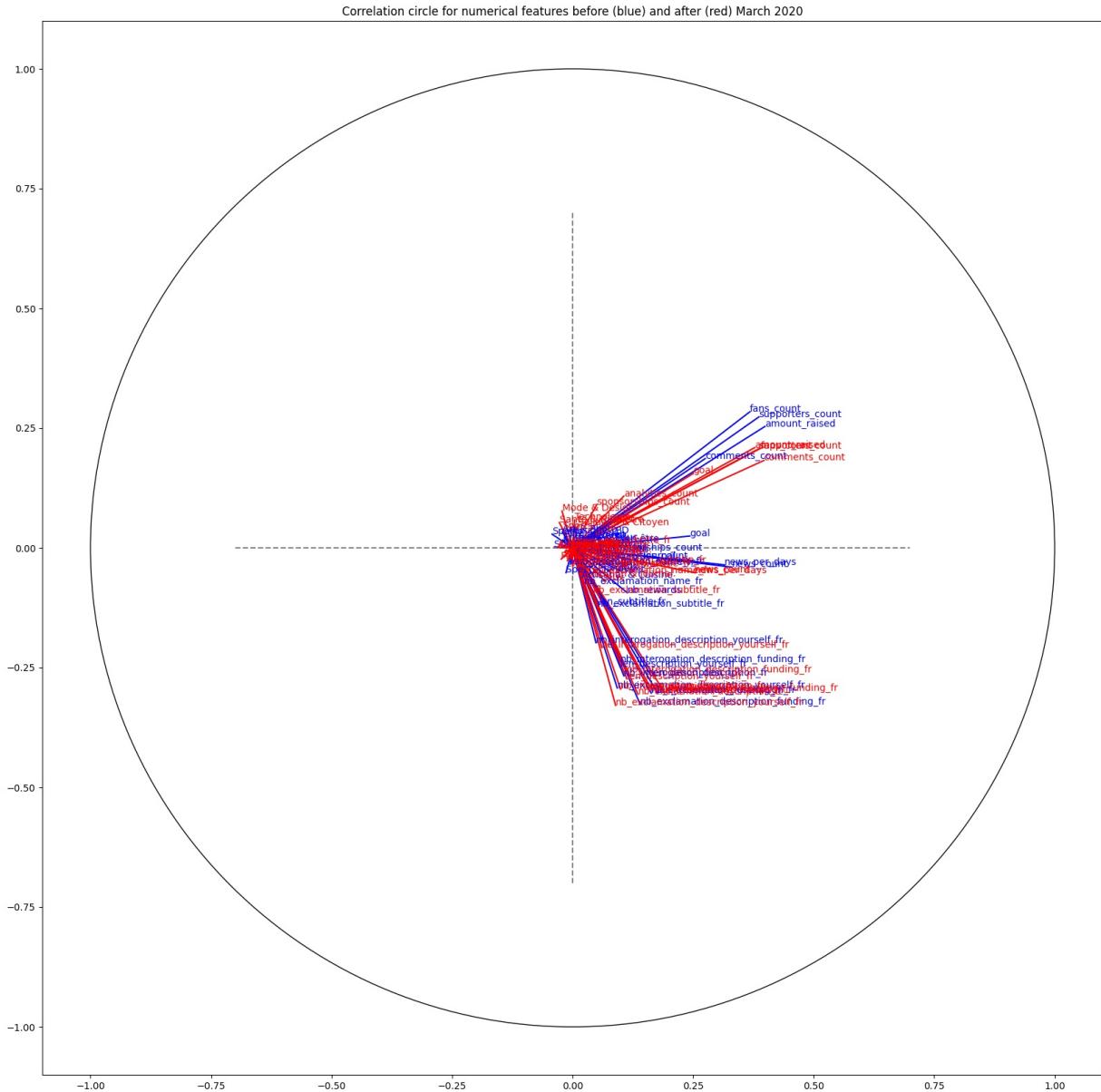


On explore les corrélations entre les différentes variables numériques du dataset. Le montant récolté est assez logiquement fortement corrélé aux nombres de fans, de supporter et de commentaires. Il semble également que mettre des news sur un projet soit une des façons de susciter des commentaires.

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.



???

Tuning et performance des différents modèles

Régression logistique

Le score de la régression logistique après tuning des paramètres est 0.7783283067643257. Avec les paramètres: solver = lbfgs, penalty = l2 et C = 100

Predicted label

