# 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.

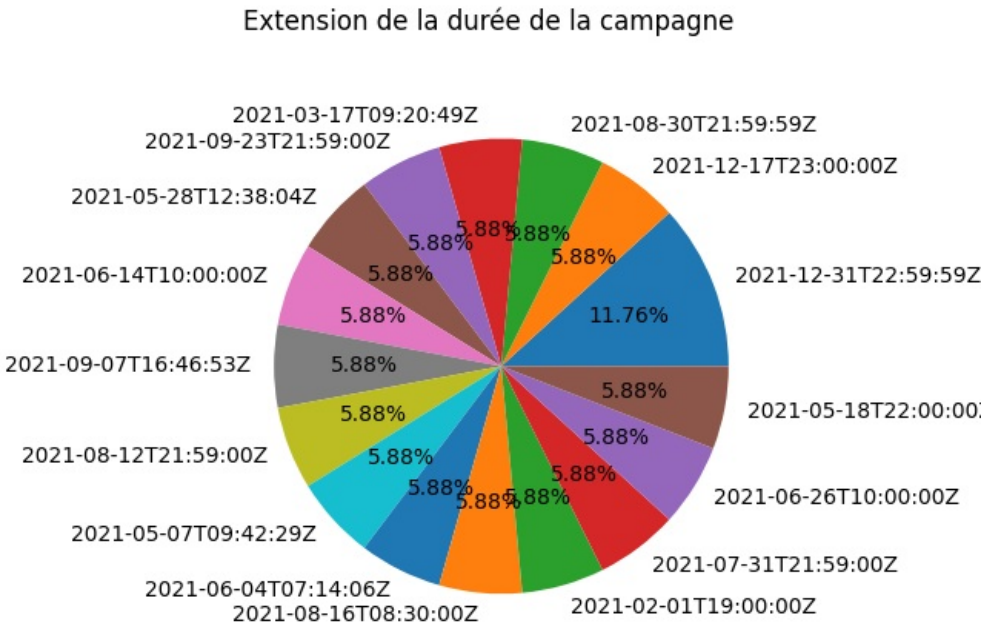
- amount_raised
- comments_count
- ~~committed~~
- date_end
- date_end_extra_time
- date_goal_raised
- date_start
- fans_count
- ~~finished~~
- ~~is_cancelled~~
- ~~is_in_extra_time~~
- ~~lowest_contribution_amount~~
- nb_days
- nb_products_sold
- news_count
- orders_count
- percent
- sponsorships_count
- 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**.

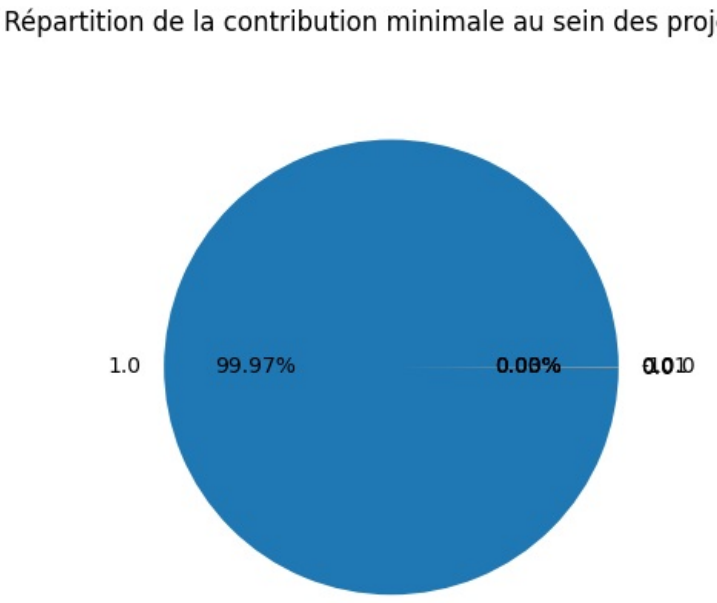
#### date_end_extra_time

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

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**

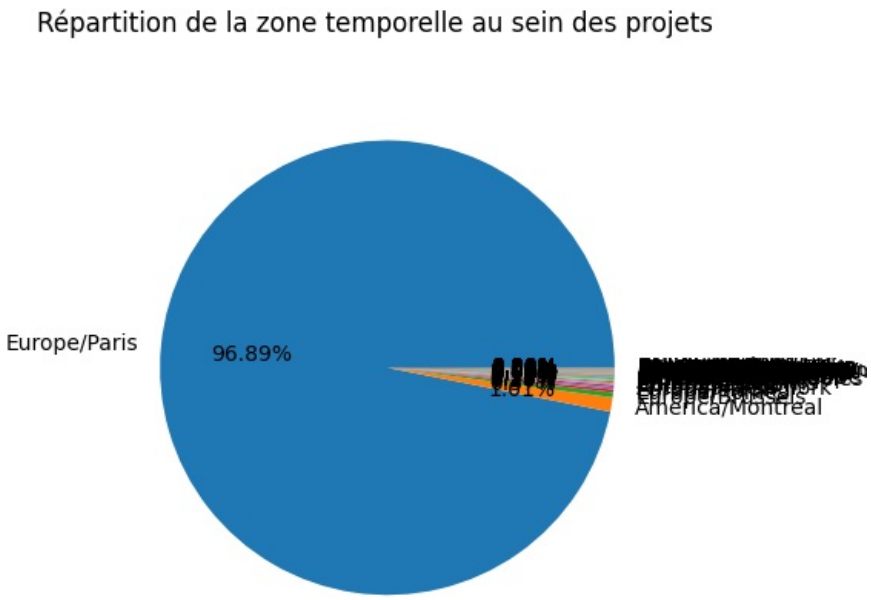
Ces données concernent le projet avant son lancement.

- 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).

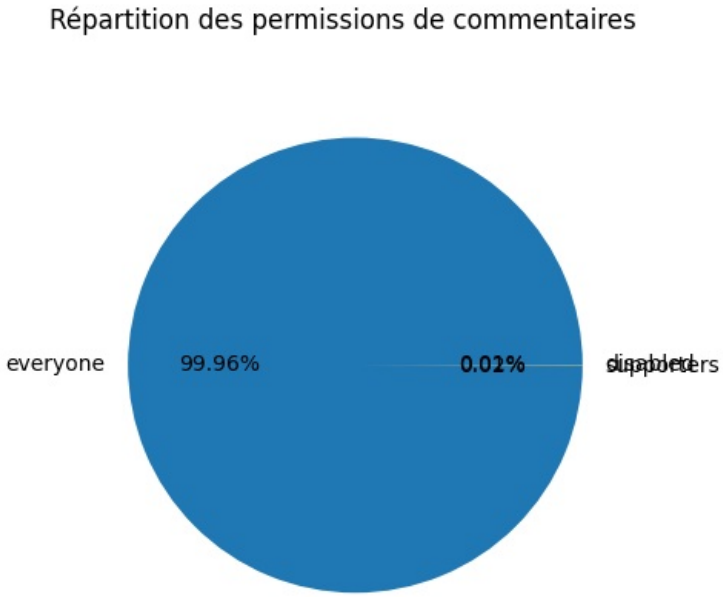
#### timezone

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



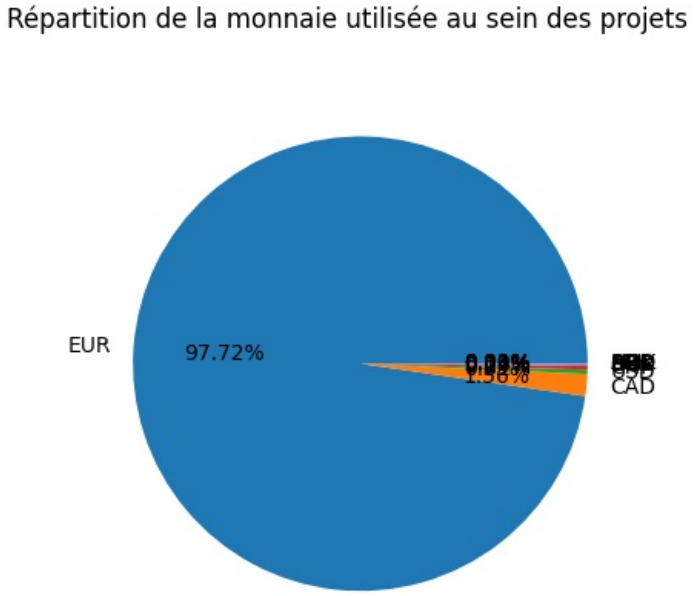

#### comments_enabled

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




#### currency

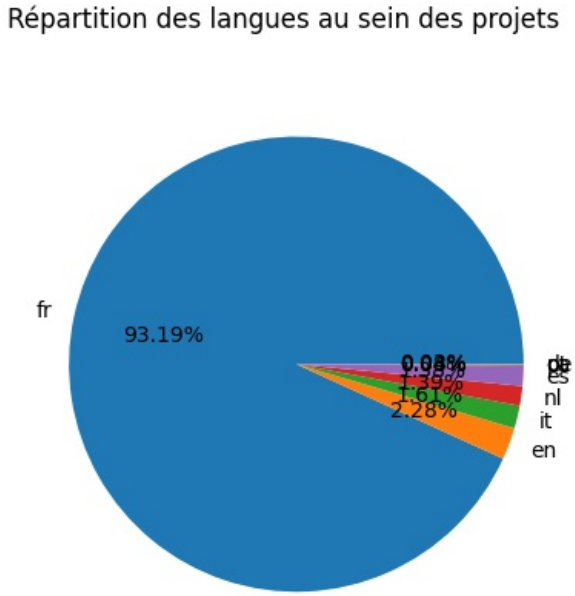
L'écrasante majorité des projets est en euro, il est donc possible de retirer la colonne **currency** ainsi que la colonne **currency_display**, sans oublier les projets concernés.




**lang**

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).

#### 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 aucun valeur dans la colonne date_start
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne date_end
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne amount_raised
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne comments_count
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne date_start
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne date_end
- Retrait de 2 lignes n'ayant aucun valeur dans la colonne description_fr
- Retrait de 714 lignes n'ayant aucun valeur dans la colonne description_funding_fr
- Retrait de 399 lignes n'ayant aucun valeur dans la colonne description_yourself_fr
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne fans_count
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne goal
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne goal_raised
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne id
- Retrait de 53 lignes n'ayant aucun valeur dans la colonne main_tag
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne name_fr
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne news_count
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne owner
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne percent
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne rewards
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne sponsorships_count
- Retrait de 2 lignes n'ayant aucun valeur dans la colonne subtitle_fr
- Retrait de 0 lignes n'ayant aucun valeur dans la colonne supporters_count

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**.