# Projet n°8 – Kickstarter : peut-on prévoir si un projet va réussir ou échouer ?

Jérémy Vangansberg - OpenClassrooms

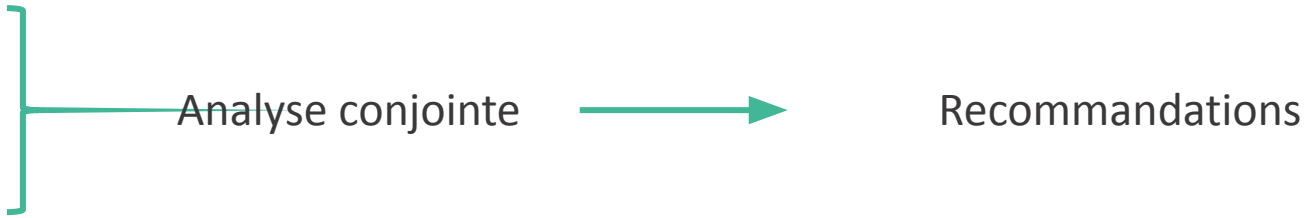
# Plan

1. Contexte et enjeux du projet
2. Présentation du rapport

# Contexte et enjeux

# Contexte et objectif du projet

- Sujet étudié - La plateforme de *crowdfunding* ou financement participatif : Kickstarter
- Objectif :
  - Prédire avec un maximum de précision si un projet va réussir ou échouer **avant qu'il ne soit lancé**
- Deux phases :
  - Analyse exploratoire
  - Modélisation

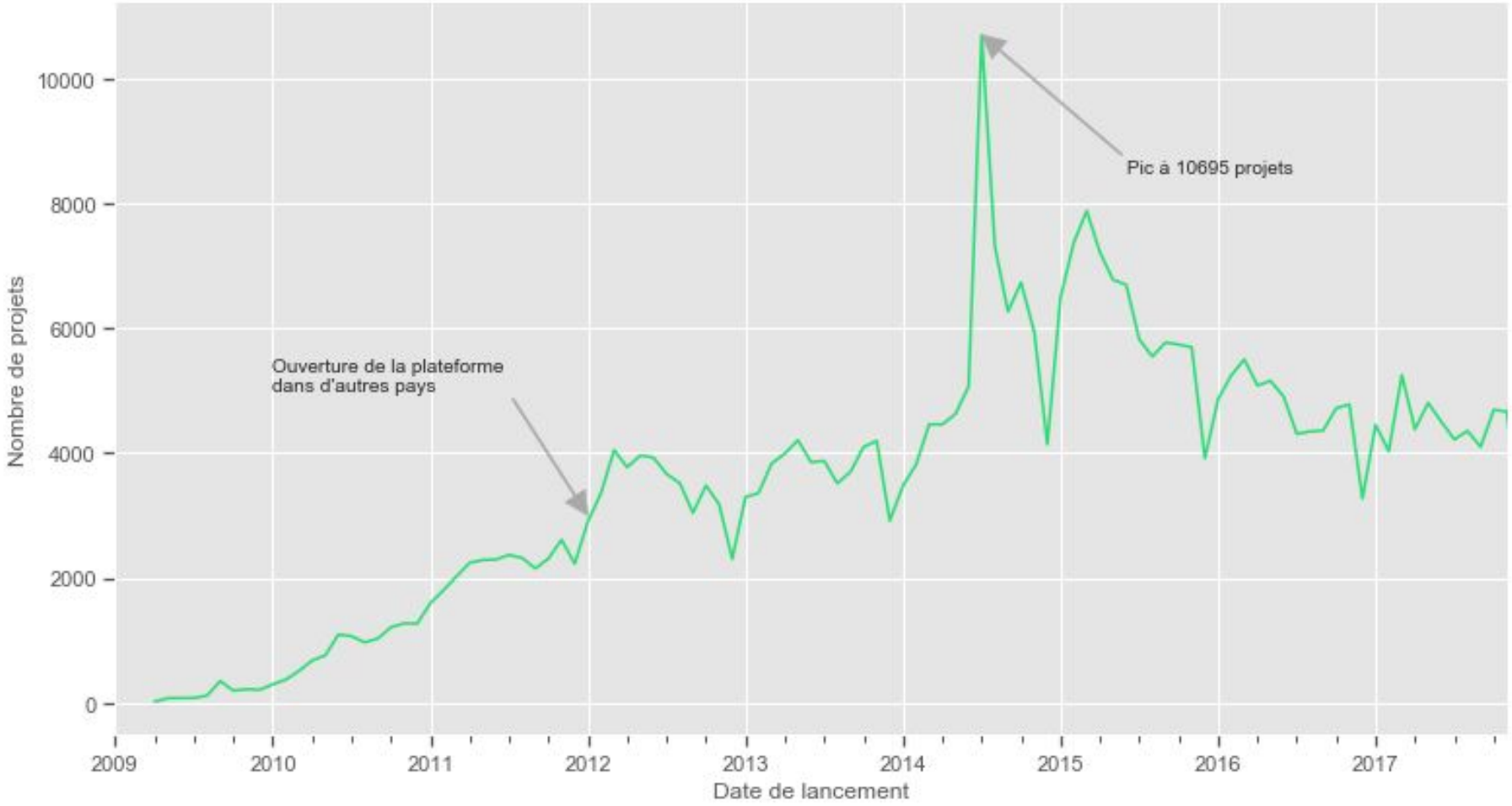


# A propos de Kickstarter

- Lancement en **2009** aux Etats-Unis
- En 2020 :
  - Accessible depuis 22 pays
  - 19 millions de personnes se sont engagées à hauteur de 5,4 milliards $
  - Plus de 500 000 projets proposés dont environ 192 000 financés
- Plage temporelle de mon analyse :
  - **2009-2017**

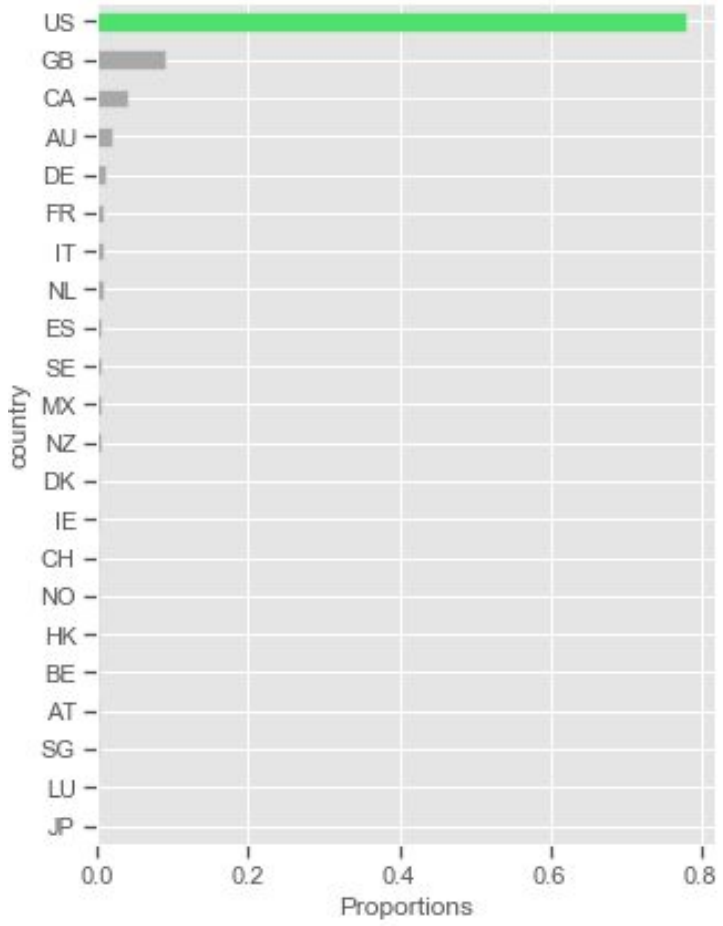
# A propos de Kickstarter

Nombre de projets lancés sur Kickstarter entre le 28 avril 2009 et le 30 novembre 2017



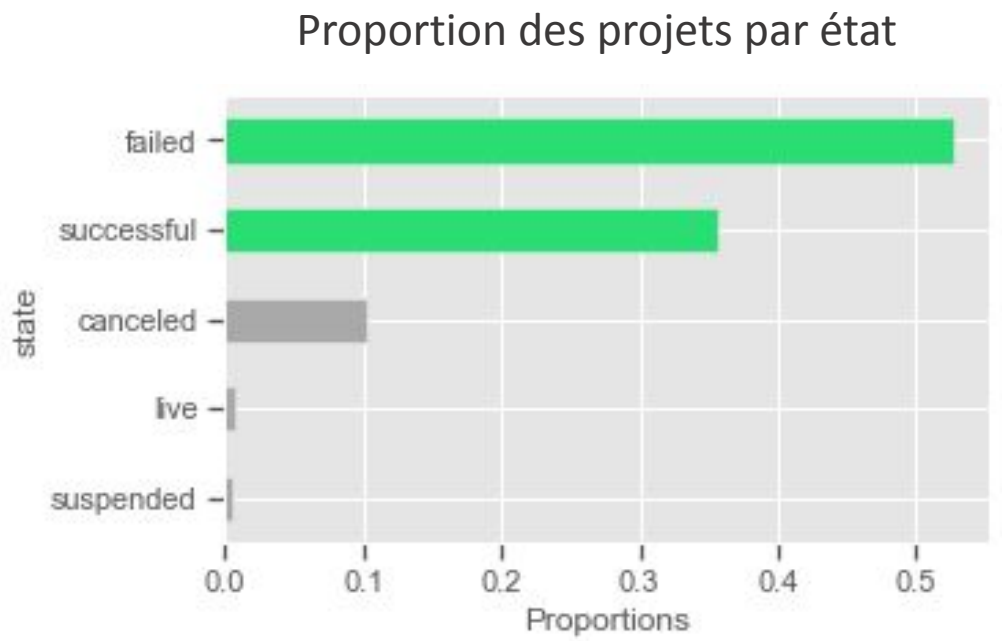
# A propos de Kickstarter

Proportion des pays d'origine



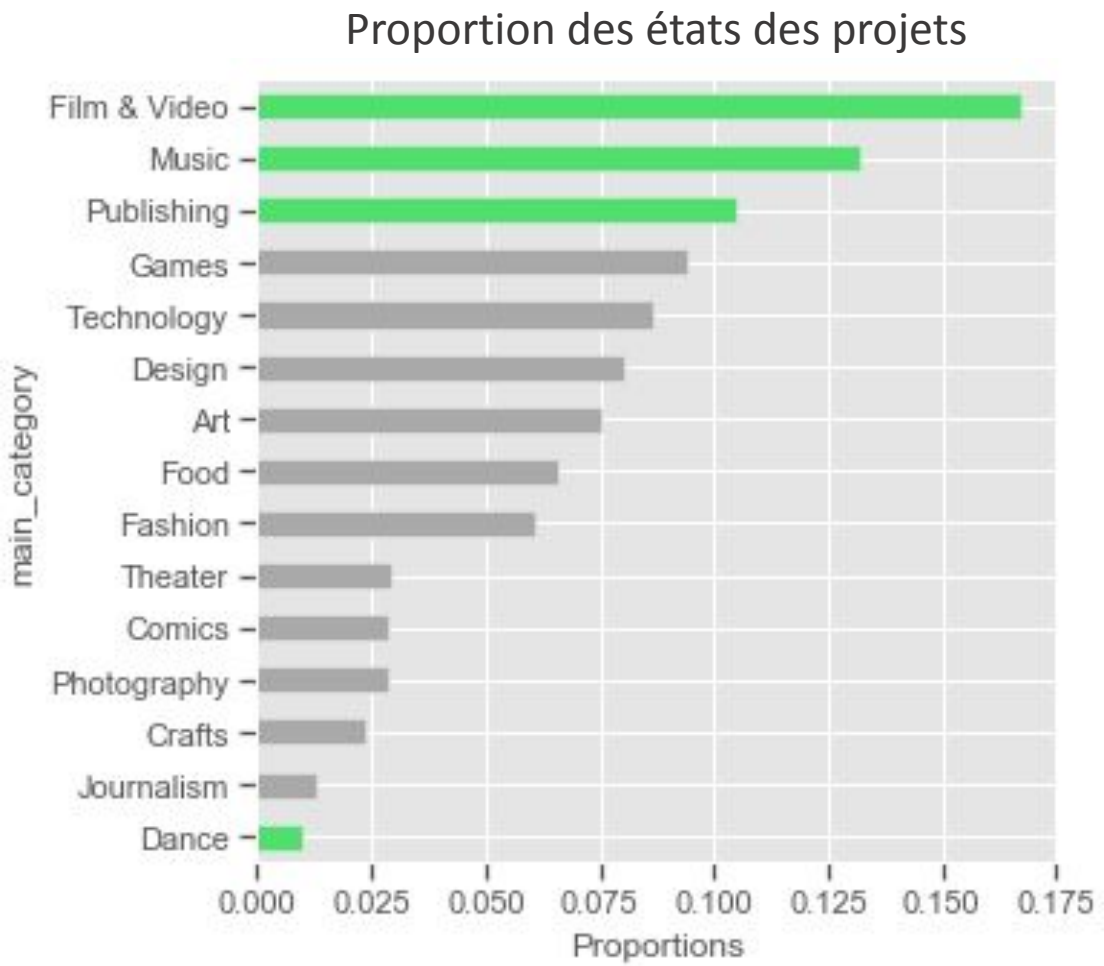

- Presque 80% des projets proviennent des Etats-Unis

# Combien de projets aboutissent ?



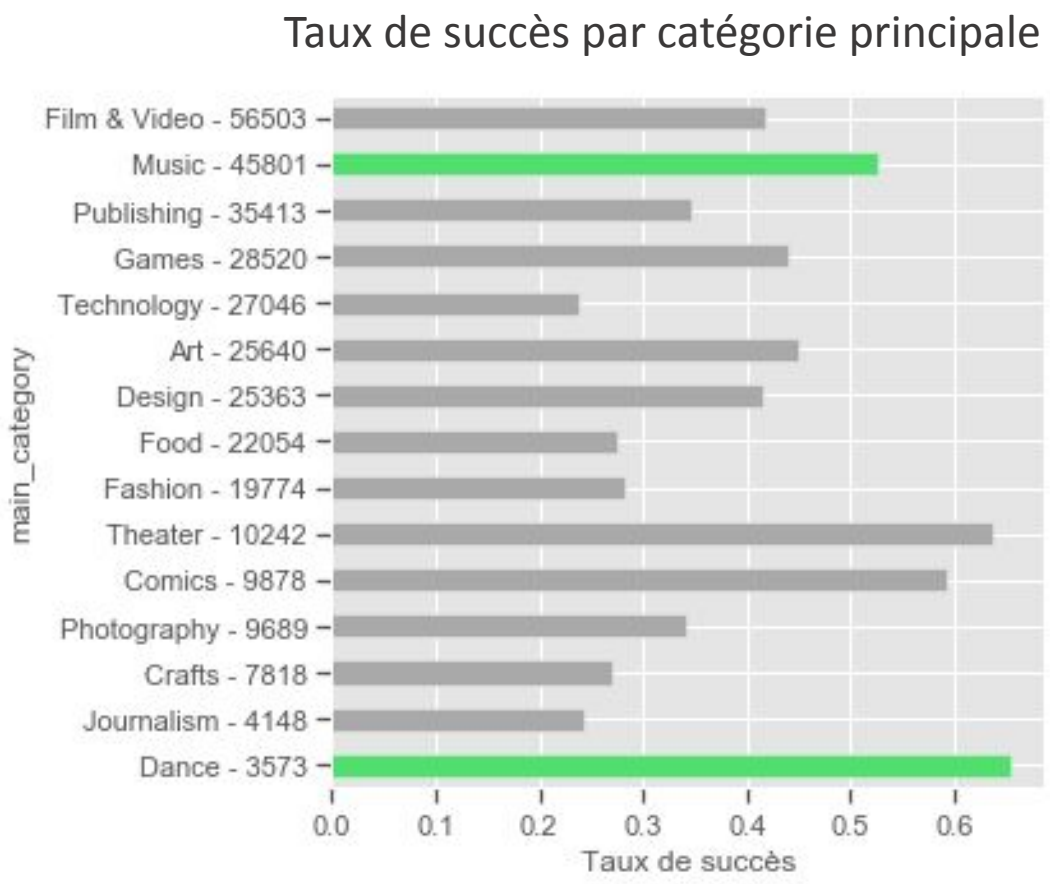

- + 50 % d’échecs
- Environ 35 % de succès et 10% d’annulations

# Nombre de projets par catégorie




- Les catégories les plus représentées sont :
  - *Film & Video*
  - *Music*
  - *Publishing*

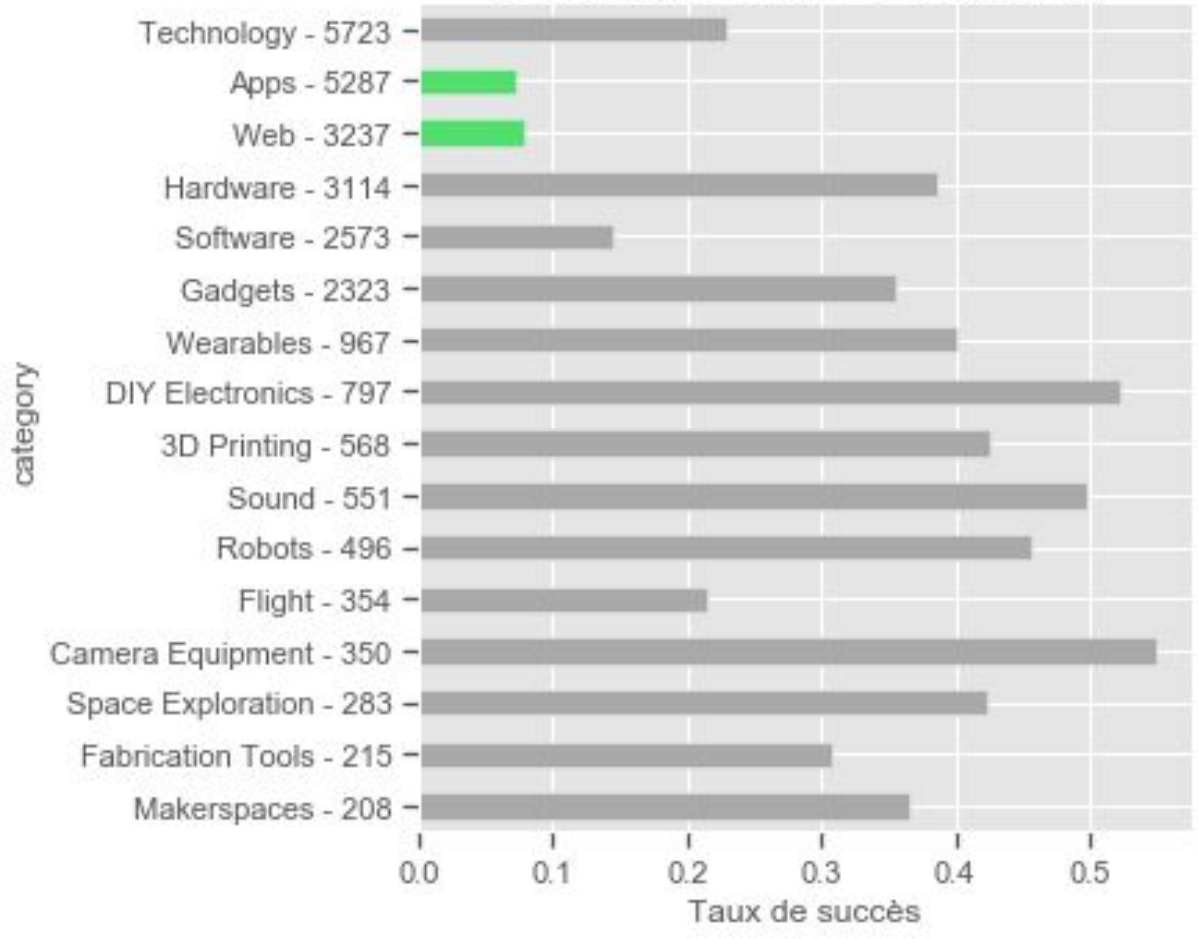
# Taux de succès par catégorie principale




- *Music* a également un taux de réussite élevé
- *Dance,* la catégorie la moins populaire à plus de 65% de taux de réussite

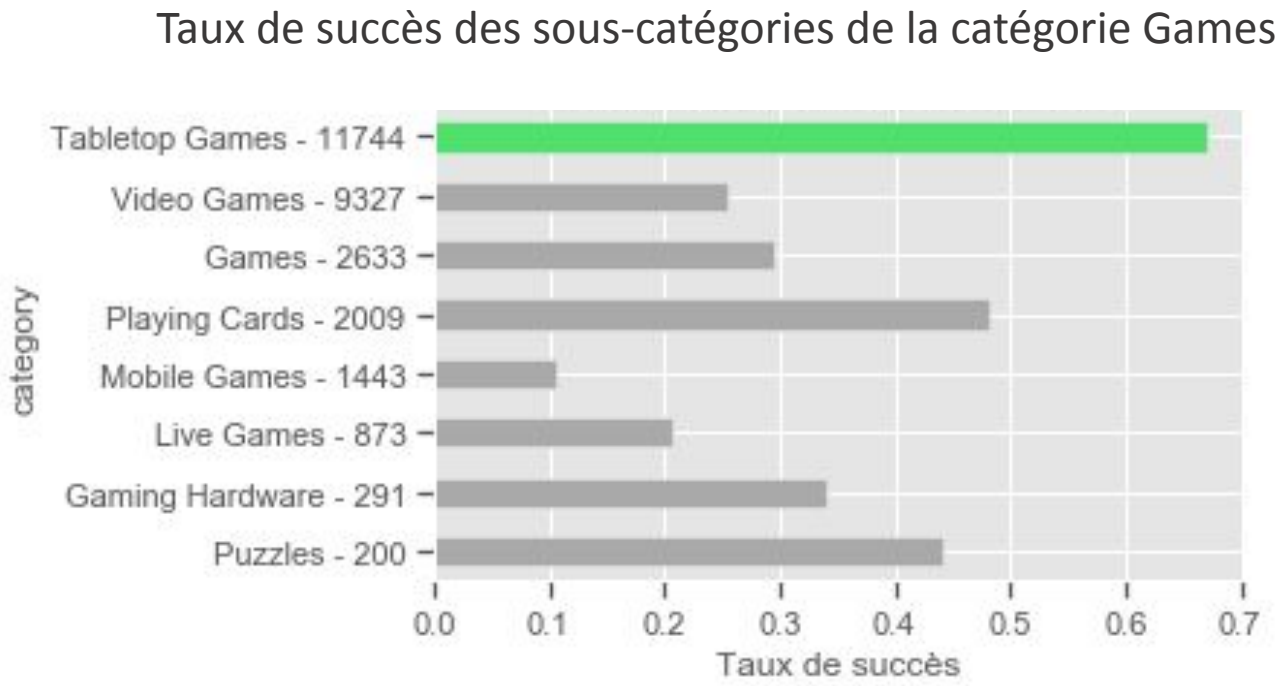
# Taux de succès des sous-catégories de *Technology*

Taux de succès des sous-catégories de la catégorie *Technology*




- Apps et web à sont à – de 10 % de taux de réussite

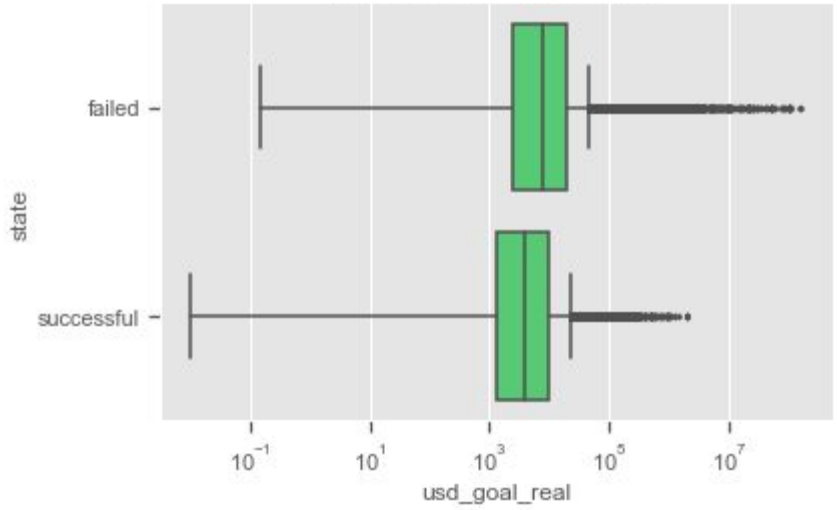
# Taux de succès des sous-catégories de *Games*




- Les jeux de société ont un taux de réussite de plus de 65%

# Montant de l'objectif en fonction de l'état

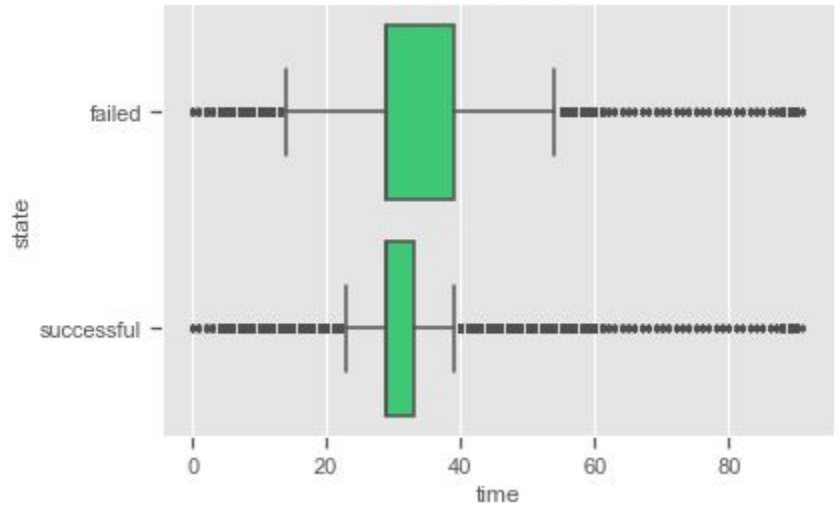
Distribution de `usd_goal_real` en fonction de l'état du projet




- Les projets qui réussissent ont un objectif de financement plus bas
- Cette information est à pondérer. En effet, un objectif plus bas est plus facile à atteindre

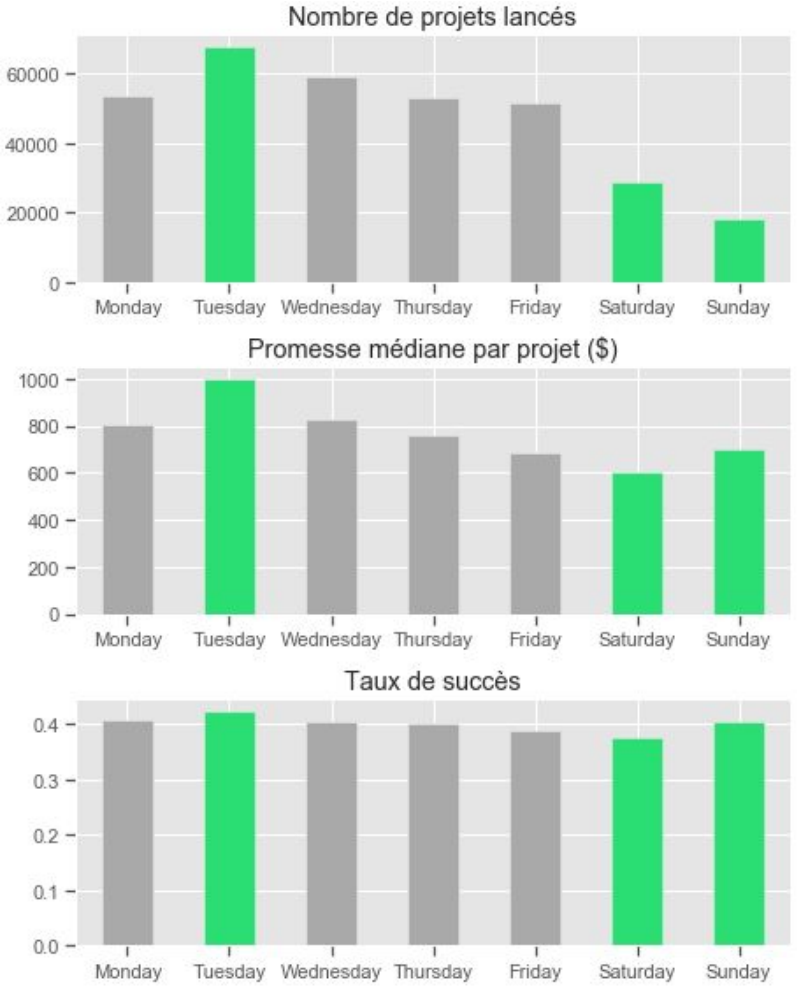
# Durée de mise en ligne, en fonction de l'état

Distribution de `time` par état



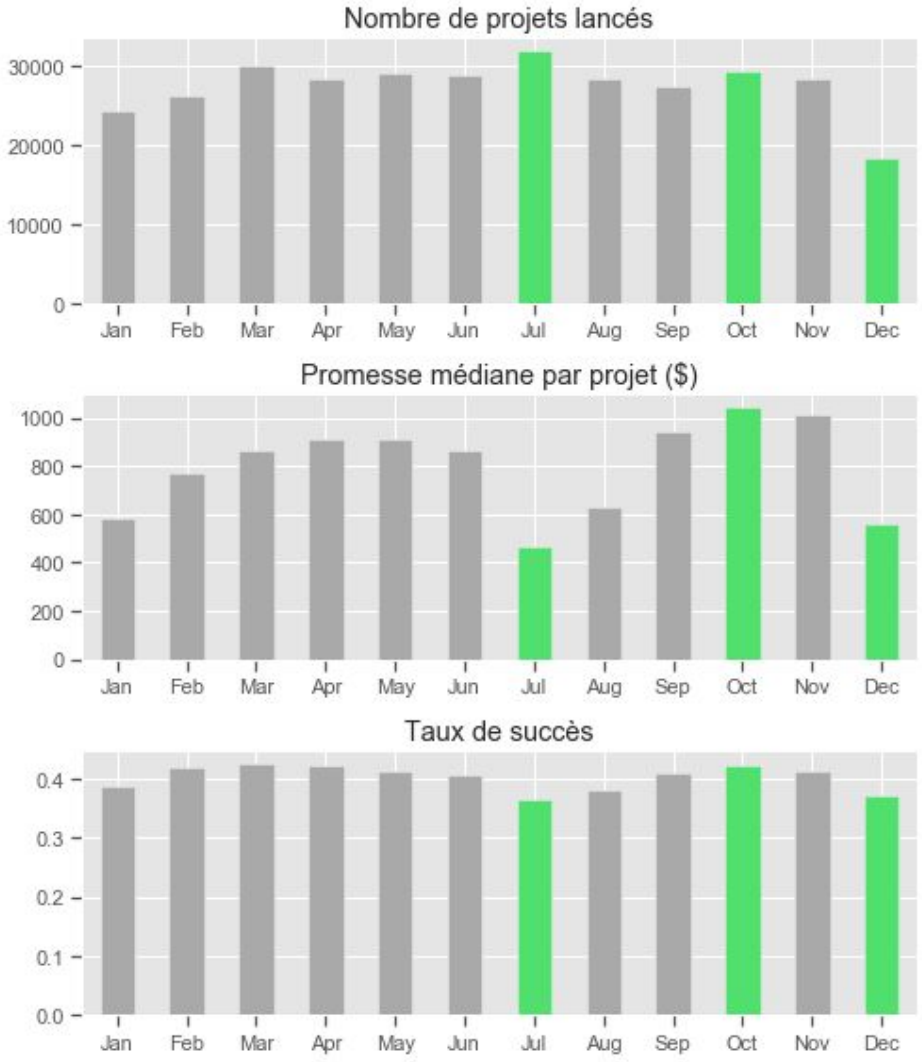

- Les projets qui échouent et ceux qui réussissent ont une médiane et un Q1 **égalent à 29**
- En général, les projets qui réussissent sont **plus courts**

# Les variables temporelles : les jours de la semaine



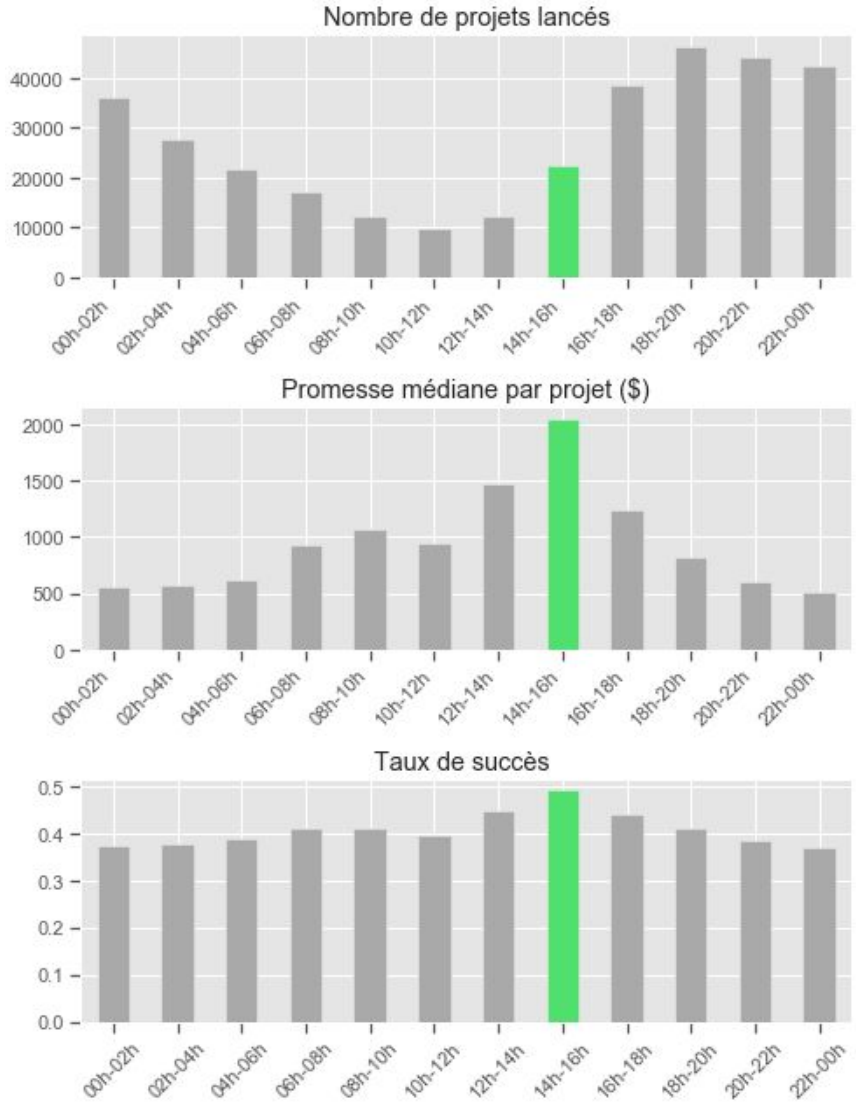

- Le mardi ⊕
- Le week-end ⊖

# Les variables temporelles : les mois




- Décembre ⛔
- Juillet ⛔
- Octobre ➕

# Les variables temporelles : les heures




- 14-16h ⊕
- La nuit et le matin ⊖

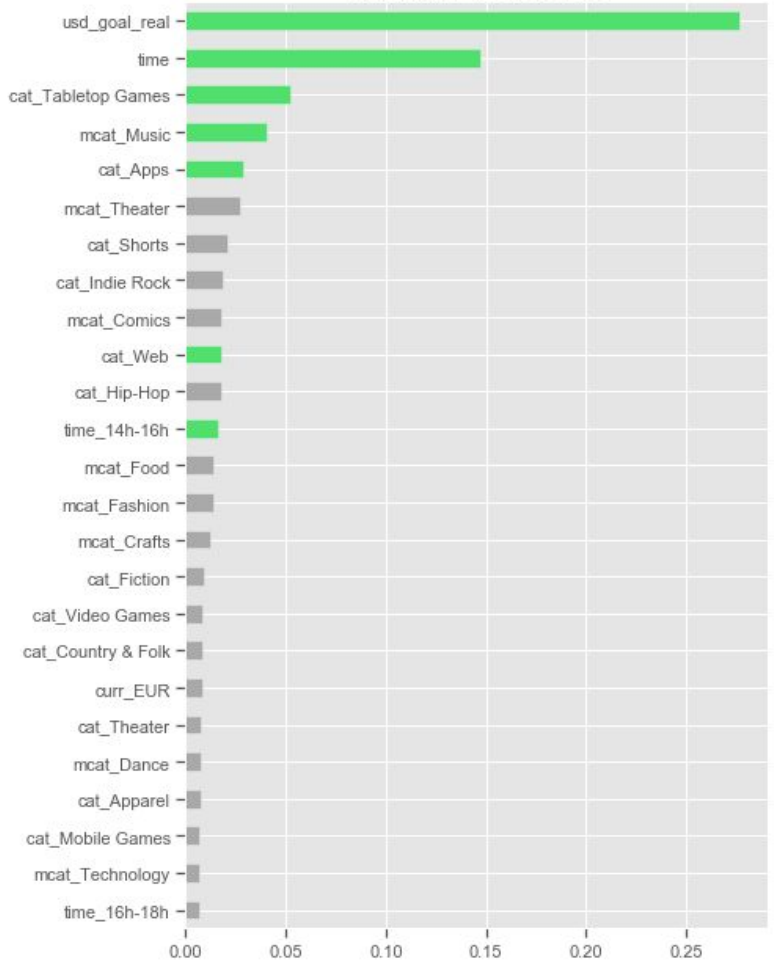
# Modélisation

- 3 modèles ont été utilisé :
  - Algorithmes de classification supervisés
  - Implémentation à l'aide de la librairie *sklearn*

| **Nom des algorithmes** | ***Accuracy*** |
|---|---|
| Régression logistique | 65,1% |
| *Support Vector Classifier* | 64,38% |
| *Stochastic Gradient Boosting* | 68,65% |

# L’importance des variables

L’importance des variables utilisées par le modèle *Stochastic Gradient Boosting*




- 2 **variables numériques** en tête :
  - L’objectif de financement et la durée
- … elles sont suivies par des catégories et des sous-catégories
- 12$^{e}$ place : le créneau 14-16h

# Recommandations business

A l'issu de mon analyse, je peux émettre plusieurs recommandations :

- Le montant de l'objectif de financement doit rester **modéré**
- La durée entre le lancement et la date butoir doit rester **courte**
- Les catégories à favoriser : **Jeux de société**, **bandes dessinées**
- Les catégories à éviter : **Apps**, **Web**
- Les moments les plus propices : entre **14** et **16** heures, le **mardi** et en **octobre**
- Les moments les moins propices : la **nuit** et le **matin**, le **week-end** et en **juillet** ou en **décembre**

# Hypothèses sur les catégories

Attrait pour les projets **tangibles** :

- Les jeux de société
- les bandes dessinées

L’importance de la **communauté** :

- La catégorie musicale

# Présentation du rapport

# Structure

Le rapport fait 22 pages et contient :

- 18 graphiques et 2 tableaux
- Page de garde avec le logo et la problématique

# Résumé

- Un résumé à la deuxième page :
  - Contexte
  - Résultats
  - Recommandations
- Il permet de **donner envie** au lecteur d'aller plus loin
- Commencer par le plus important


**Résumé**

Le financement participatif ou *crowdfunding* connaît un essor important depuis ces 10 dernières années[1]. C'est devenu une alternative aux moyens de financement classique comme les banques. Cette analyse porte sur l'une des plateformes les plus connues : Kickstarter[2]. Cette structure existe depuis 2009 et plus de 500 000 projets y ont été proposés. 192 000 ont atteint leur objectif[3]. L'entreprise s'est ouverte à l'international en 2014 mais elle reste principalement utilisée par des états-uniens.

Dans cette analyse, j'utilise le *Stochastic Gradient Boosting* sur un jeu de données extrait de *Kaggle* pour définir si un projet va réussir ou échouer avant qu'il soit lancé. Le modèle a un taux d'*accuracy* de 68.65%. Les variables les plus importantes pour la réussite d'un projet sont : un objectif de financement raisonnable et une période de financement courte. De plus, il existe des écarts importants entre les taux de réussite des différentes catégories. En effet, moins de 10% des *Apps* réussissent à atteindre leur objectif tandis que ce chiffre atteint plus de 65% pour les jeux de société. Cette plateforme est plutôt adaptée aux petits projets qualitatifs qu'aux projets qui nécessitent un apport en capital important.


[1] *Crowdfunding: Mapping EU markets and events study -European Commission*
[2] https://www.kickstarter.com/
[3] Chiffres en 2020

2

# Plan

- Plan :
  - Chronologique
  - 3 niveaux

**Table des matières**

# Rédaction : Syntaxe et temps

- Les temps
  - Utilisation du **présent** (notamment présent narratif) pour donner plus d'impact
  - Utilisation occasionnelle du passé et du futur (pour respecter une chronologie)
- Syntaxe
  - Une phrase = une idée

# Rédaction : Polices et attributs du texte

- Corps du texte : Times New Roman
- Référence aux éléments du code : `Courrier New`
- Attributs pré-attentifs : couleur, taille, *italique*, **gras**, etc.
- Utilisation de *l'italique* pour les mots anglais et les valeurs des variables catégorielles
- Les caractères gras sont réservés aux titres

Un projet est défini comme *successful* si `usd_goal_real` est inférieur à `usd_pledged_real`. Par conséquent les variables : `pledged`, `usd_pledged`, `usd_pledged_real` mais également `backers` sont des *data leakage*[5], c'est-à-dire que ces informations ne sont pas censées être disponibles au lancement du projet et donc au moment où l'on souhaite faire la prédiction. Ces informations ne seront donc pas utilisées dans les algorithmes.

# Charte graphique

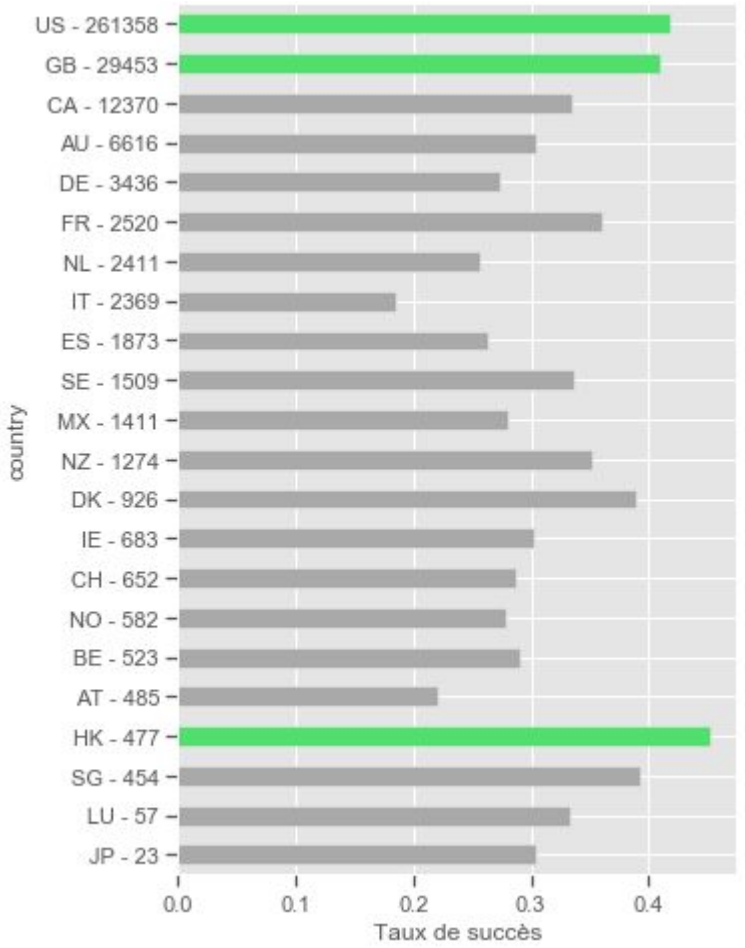
Rapport d'Analyse

KICKSTARTER

Kickstarter : Peut-on prédire si un projet va réussir ou échouer avant qu'il soit lancé ?

Jérémy Vangansberg - OpenClassrooms

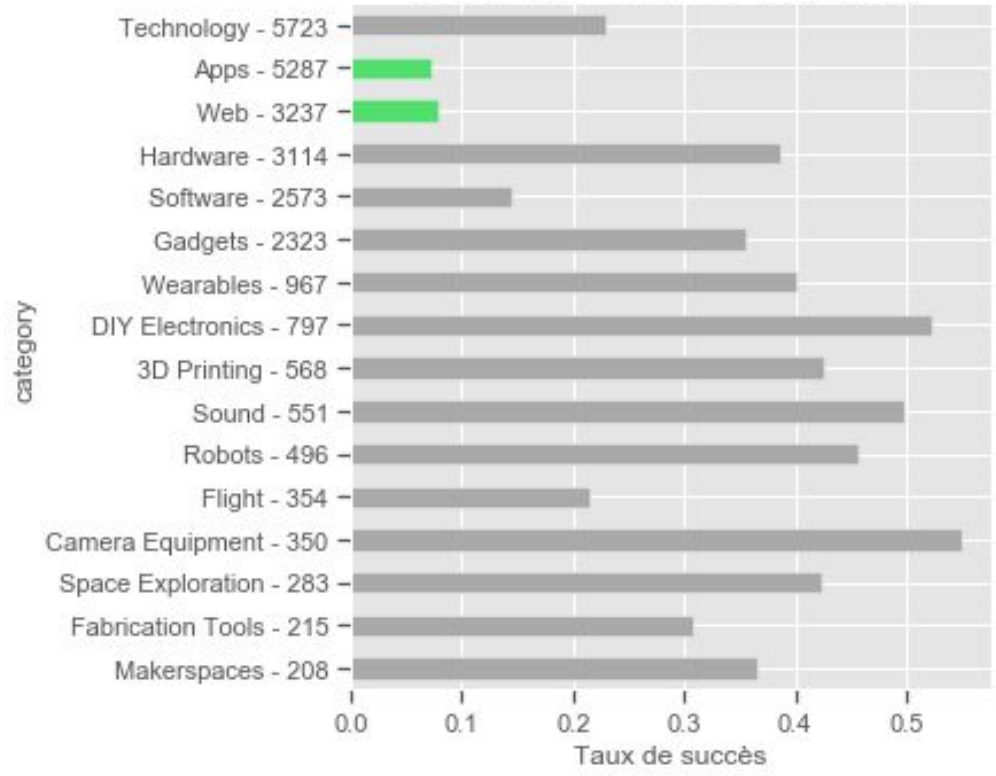
Proportion des pays d'origine




- Nuances de gris
- Reprise de la couleur de l'entreprise : '#2bde73'
- Utilisation du style 'ggplot' :
  - Grille
  - Fond gris
  - Sans bordure (distraction visuelle)

# Phase exploratoire et la phase explicative

Taux de succès des sous-catégories de la catégorie *Technology*




- **Mise en valeur** des éléments importants de chaque graphique

# Types de graphique

3 types de graphiques ont été utilisé :

Boîte à moustaches

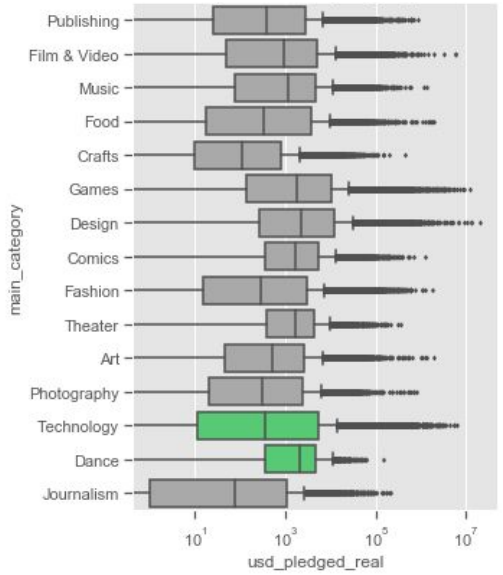



Diagramme à barres horizontales

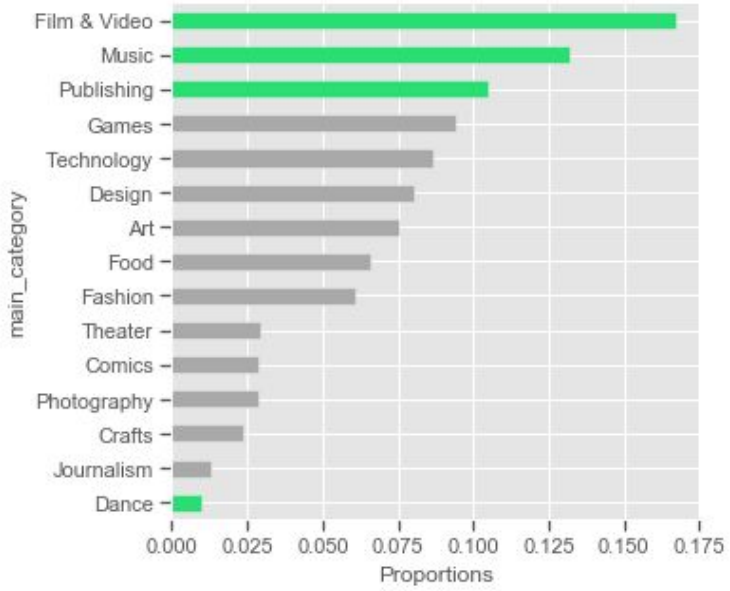
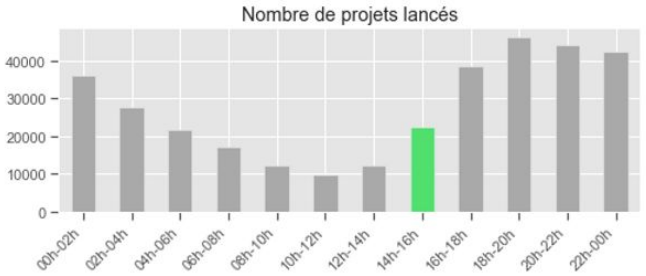



Diagramme à barres verticales




Les graphiques non-utilisés : diagramme circulaire, scatter plot, etc.

# Diagramme à barres vs pie-chart

Diagramme circulaire

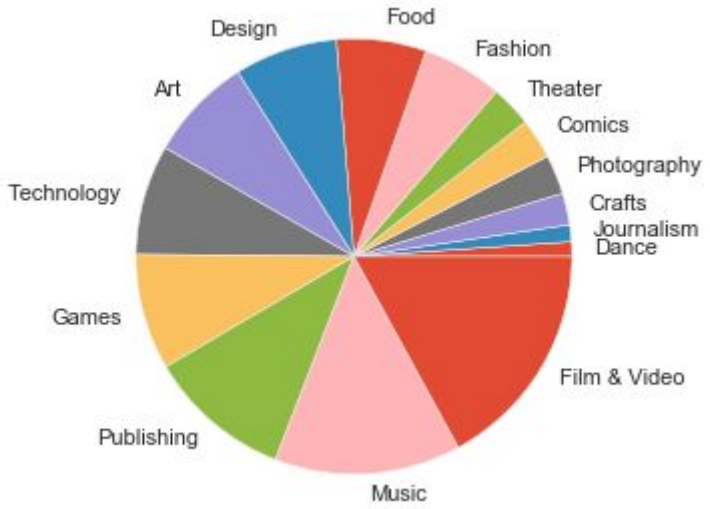
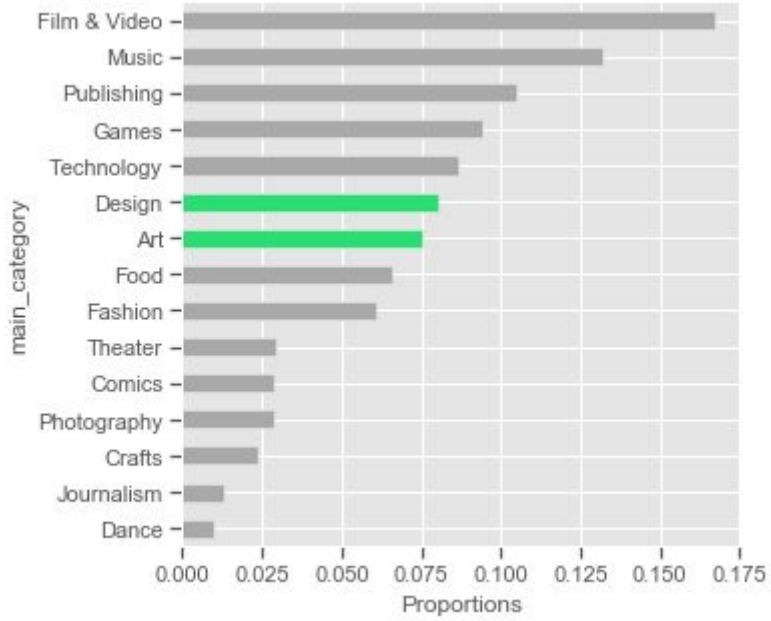



Diagramme à barres horizontales



# Tableaux

| Nom de la variable | Description de la variable | Type | Données non-nulles |
|---|---|---|---|
| ID | Numéro d'identifiant | Int64 | 378661 |
| name | Nom | Object | **378657** |
| main_category | La catégorie principale | Object | 378661 |
| category | La sous-catégorie | Object | 378661 |
| currency | La devise locale | Object | 378661 |
| country | Pays d'origine | Object | 378661 |
| goal | Objectif de financement fixé par le créateur dans sa devise | Float64 | 378661 |
| usd_goal_real | Objectif de financement en dollars fixé par le créateur | Float64 | 378661 |
| launched | Date de lancement | Object | 378661 |
| deadline | Date butoir | Object | 378661 |
| pledged | Promesse de financement dans la devise locale | Float64 | 378661 |
| usd_pledged | Promesse de financement en dollars (conversion faite par Kickstarter) | Float64 | **372041** |
| usd_pledged_real | Promesse de financement en dollars (conversion faite par Fixer.io API) | Float64 | 378661 |
| backers | Nombre de personnes qui soutiennent financièrement le projet | Float64 | 378661 |
| state | L'état du projet (*successful, failed, canceled, suspending ou still live* ; soit réussi, échoué, annulé, suspendu ou en cours | Object | 378661 |

TABLEAU 1 – Description du jeu de données

| Variable | Cardinalité |
|---|---|
| main_category | 15 |
| category | 159 |
| currency | 14 |
| country | 22 |
| state | 5 |

TABLEAU 2 – Cardinalité par variable catégorielle

- **Synthétiser** un grand nombre d'informations sur un espace réduit
- Permet de **résumer** une information autrement que par une phrase

# Références

- Utilisation des notes de bas de page
  - Effectuer une traduction
  - Citer une source
  - Donner plus de détails
- Alternative en fin de rapport

## 1. Introduction

Ma démarche consiste à tenter de prévoir si un projet proposé sur Kickstarter va réussir ou échouer avant qu'il soit lancé. Le jeu de données vient de la plateforme *Kaggle* et il contient environ 370 000 projets lancés entre 2009 et 2017. Afin d'effectuer des prédictions sur l'échec ou le succès d'un projet, je vais utiliser différents algorithmes de classification dont je comparerai la précision. Dans ce but, j'utiliserai un mélange entre des données catégorielles et des données numériques. Ces données seront préparées afin que les algorithmes puissent les utiliser dans un cadre optimal. Lorsque les différents modèles seront entraînés et testés, je sélectionnerai le meilleur d'entre eux pour approfondir son analyse.

## 2. Nettoyage des données

Le dataset contient 378661 entrées et 14 variables. Chaque entrée concerne un projet :

| Nom de la variable | Description de la variable | Type | Données non-nulles |
|---|---|---|---|
| `ID` | Numéro d'identifiant | Int64 | 378661 |
| `name` | Nom | Object | **378657** |
| `main_category` | La catégorie principale | Object | 378661 |
| `category` | La sous-catégorie | Object | 378661 |
| `currency` | La devise locale | Object | 378661 |
| `country` | Pays d'origine | Object | 378661 |
| `goal` | Objectif de financement fixé par le créateur dans sa devise | Float64 | 378661 |
| `usd_goal_real` | Objectif de financement en dollars fixé par le créateur | Float64 | 378661 |
| `launched` | Date de lancement | Object | 378661 |
| `deadline` | Date butoir | Object | 378661 |
| `pledged` | Promesse de financement dans la devise locale | Float64 | 378661 |
| `usd_pledged` | Promesse de financement en dollars (conversion faite par Kickstarter) | Float64 | **372041** |
| `usd_pledged_real` | Promesse de financement en dollars (conversion faite par Fixer.io API) | Float64 | 378661 |
| `backers` | Nombre de personnes qui soutiennent financièrement le projet | Float64 | 378661 |
| `state` | L'état du projet (*successful, failed, canceled, suspending ou still live*[4]) | Object | 378661 |

TABLEAU 1 – Description du jeu de données

Un projet est défini comme *successful* si `usd_goal_real` est inférieur à `usd_pledged_real`. Par conséquent les variables : `pledged`, `usd_pledged`, `usd_pledged_real` mais également `backers` sont des *data leakage*[5], c'est-à-dire que ces informations ne sont pas censées être disponibles au lancement du projet et donc au moment où l'on souhaite faire la prédiction. Ces informations ne seront donc pas utilisées dans les algorithmes.

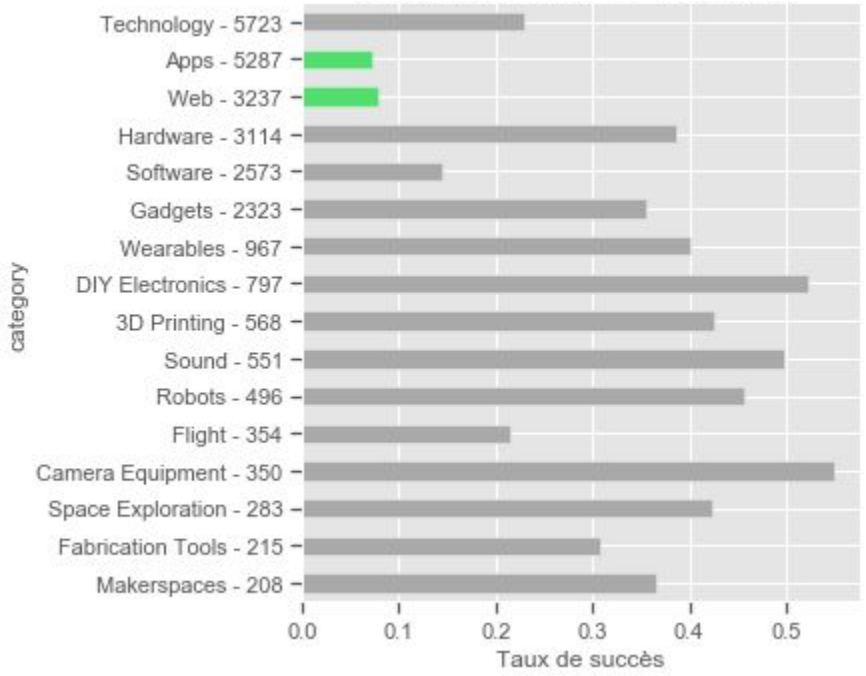
### 2.1. Données manquantes

Le tableau n°1 nous informe que deux variables contiennent des *NaN*[6]. Cependant certains *datasets* peuvent enregistrer les valeurs manquantes sous d'autres formes. C'est le cas ici pour

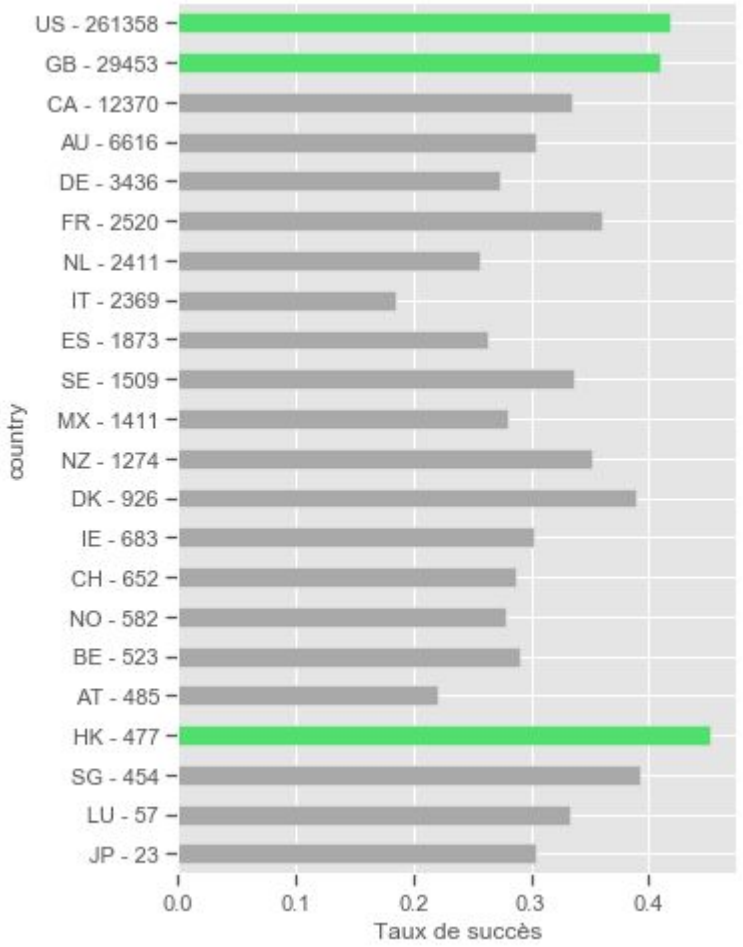
[4] Réussi, échoué, annulé, suspendu ou en cours
[5] Littéralement « Fuite de données »
[6] *Not a Number*

# Choix d’analyse - Taux de succès

Taux de succès des sous-catégories de la catégorie *Technology*

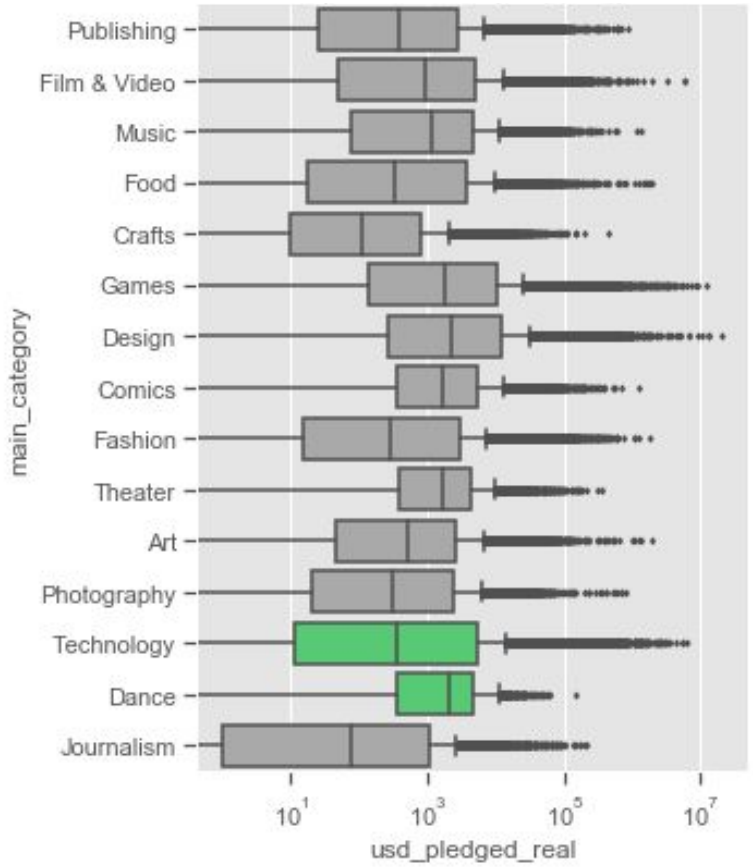



Taux de succès par pays

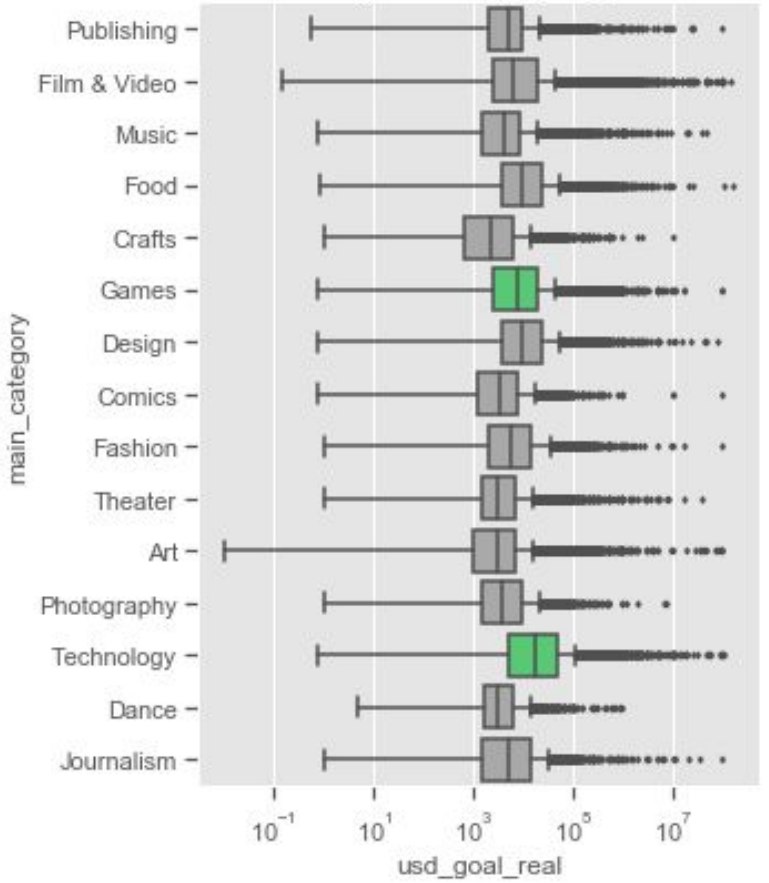


# Choix d’analyse – La distribution

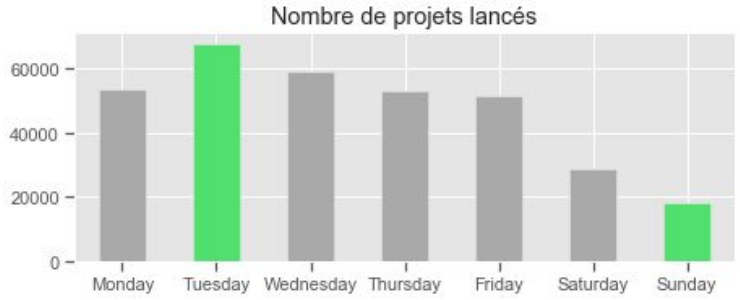
Distribution de `usd_pledged_real` par `main_category`

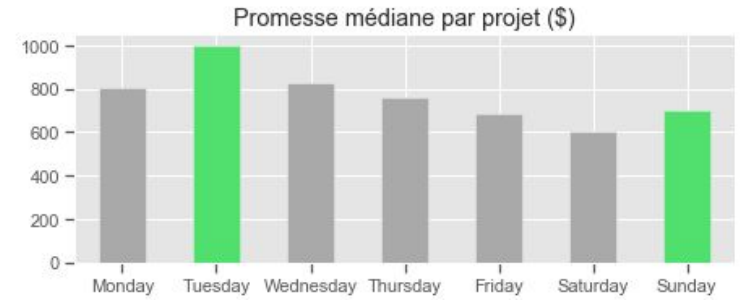


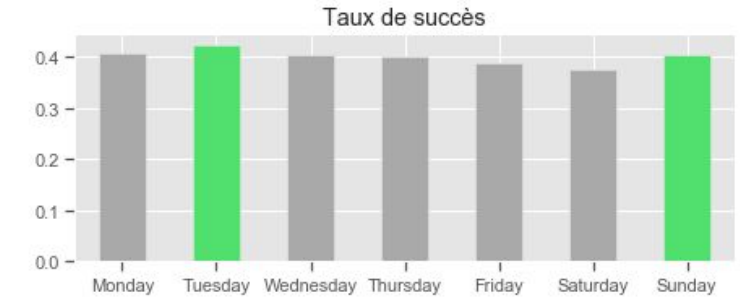

Distribution de `usd_goal_real` par `main_category`

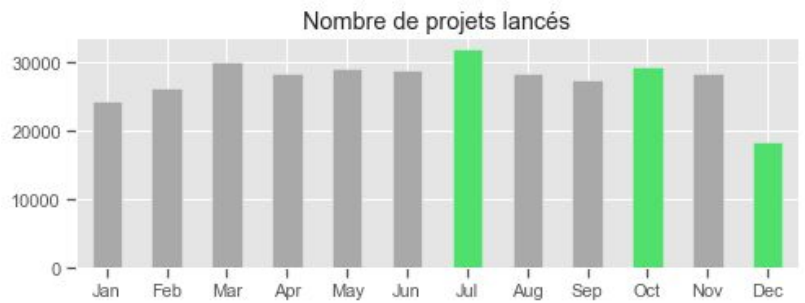


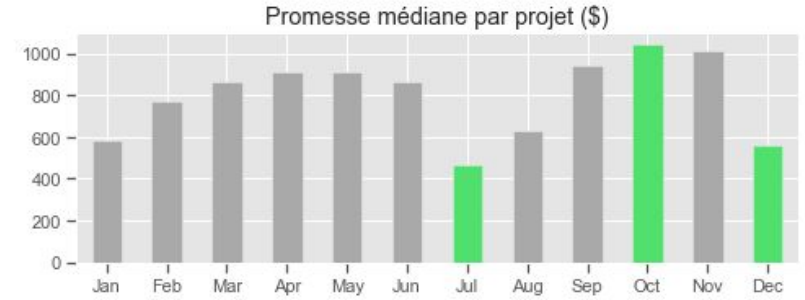
# Choix d’analyse – Nombre de projets / promesse médiane / taux de succès



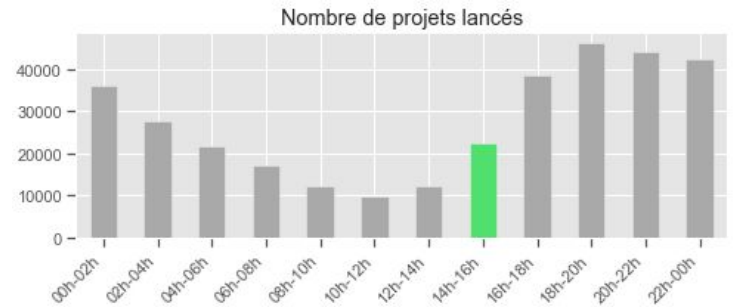



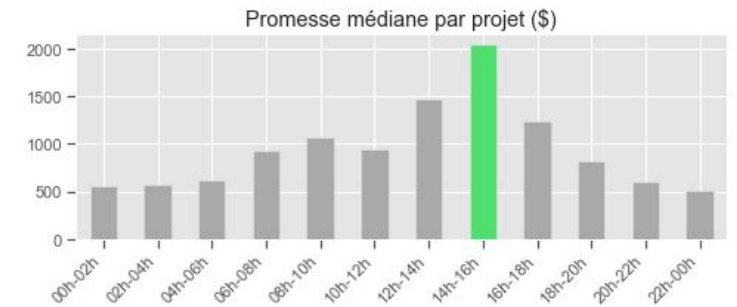








# Modélisation

- Choix des modèles
  - Deux modèles simples : **Régression logistique** et **SVC**
  - Ces méthodes permettent d'avoir un point de comparaison
  - Un modèle avancée : SGB
- Pourquoi le *Stochastic Gradient Boosting* ?
  - Méthode **ensembliste**
  - Permet de sélectionner un **sous-échantillon** du *dataset*

# Modélisation

- L'accent est mis
  - Sur **l'implémentation**
  - Le choix des paramètres
- Pas de description des modèles (lecteur technique)

### 5.2.2. Régression logistique

L'implémentation d'une régression logistique peut se faire en utilisant la fonction `SGDClassifier()` de `sklearn`. Il suffit d'attribuer la valeur `log` au paramètre `loss`.

Paramètres fixes :

- `penalty: 'elasticnet'`
- `class_weight: 'balanced'`

Paramètre à optimiser grâce à la fonction `GridSearchCV()` :

- `alpha: np.logspace(-6, 1, 6)`

Meilleur paramètre :

- `alpha: 2.51 * 10-5`

`accuracy_score` de la régression logistique : **65.1%**

# Les difficultés rencontrées

- Effectuer des modélisations sur un **volume de données important**
- Paramétrer les graphiques pour mettre en exergue les **informations souhaitées**
- Ne pas pouvoir traduire certains mots anglais (*accuracy*)
- Détecter les *Data leakage,* on n'a **pas accès** a ces **informations** au moment de la prédiction :
  - Montant des promesses de financement
  - Nombres de soutiens

# Conclusion

- Attributs pré-attentif : **Faciliter** la compréhension
- Charte graphique stricte : rendu professionnel
- Différence entre phase exploratoire et explicative

# Pistes pour approfondir le sujet

- Utiliser des algorithmes de *Natural Language Processing* pour analyser les titres des projets
- Analyser les caractéristiques des projets au fil des années
- Utiliser Google Collab