**Rapport du projet**

**Transformation- manipulation de données**

# Étape 1 : On considère le fichier lending\_club\_loans.csv

Le jeu de données contient plusieurs types de données brutes, il a besoin d’être nettoyer et transformer, dans un premier temps on remarque qu’il y a des données de types catégorielles, numériques, décimales, alphanumériques, dates et aussi des variables alphanumériques qu'on peut extraire des valeurs numériques. On observe aussi qu’il y a des colonnes avec une seule valeur unique pour toutes les observations

* Le **dataset** contient **39786** lignes et **47** colonnes de données brutes non préparées.
* Il y a **2049** valeurs manquantes réparties sur **8** colonnes. les lignes contenant au moins une valeur manquante représente **4.74%** de l'ensemble des observations.
* Il faut prédire le risque de non-paiement du prêt. L'information de payement ou non payement peut se déduire de la colonne **loan\_status**.
* La variable **loan\_status** est de type catégorielle ces statuts sont : en cours de payement, un retard entre **16** et **30** jours, un retard entre **31** et **120** jours, en période de grâce ou bien manque de payement de plusieurs échéances.
* Plusieurs colonnes dans le **dataset** ne sont pas significatives pour la modélisation : colonnes contenant des clés d'identification, colonnes avec une seule classe ou valeur unique. Ces colonnes seront exclues de la modélisation :
  + **id','member\_id'** : colonnes de clés d'identification
  + 'zip\_code' : colonne semi-masquée
  + (**'pymnt\_plan', 'initial\_list\_status', 'application\_type', 'collections\_12\_mths\_ex\_med', 'chargeoff\_within\_12\_mths', 'delinq\_amnt', 'tax\_liens'**) : ne contiennent qu’une seule classe ou valeur unique
  + **'out\_prncp', 'out\_prncp\_inv'** : plus de **99.9%** des valeurs qui sont égales à 0.
  + **total\_pyment, funded\_amnt, funded\_amnt\_inv, total\_rec\_prncp, total\_rec\_int, total\_rec\_late\_fee, recoveries, collection\_recovery\_fee, last\_pymnt\_d, last\_pymnt\_amnt, total\_pymnt\_inv** sont des informations d'affaires qui arrivent après le prêt.
  + **pub\_rec\_bankruptcies** contient 1.75% de valeurs nulles et 95% de valeurs = 0.
  + Les variables **loan\_amnt** et **installment** sont importantes pour le modèle mais sont très fortement corrélées
  + On a 4 colonnes indiquant des dates sous le format **'Mois-année'**.

**last\_pymnt\_d, last\_credit\_pull\_d** sont des dates qui arrivent après le prêt et n'ont aucune importance pour le modèle.

* + Les variables numériques ne sont pas dans la même échelle et présentent des distributions différentes.
  + La variable home\_ownership cotient 3 classes rares **'OWN'**, **'OTHER'** et **'NONE'**.

Suite à la première étape les variables contenant des valeurs extrêmes doivent être mise à l'échelle avec la méthode Robuste et les autres variables peuvent être mises à l'échelle avec une méthode de normalisation rapprochant les distributions de la normale.

# Étape 2 : Étape de nettoyage

Suite à mon analyse descriptive des données on va appliquer les deux transformations:

1. ***Transformations I :***
   1. Eliminer les colonnes inutiles
   2. Eliminer la variable ‘**pub\_rec\_bankruptcies’** qui contient **1.75%** de valeurs nulles, et plus de **99%** de valeurs = **0**.
   3. Eliminer la variable **'installment'** fortement corrélée à la variable **‘loan\_amnt’**.
   4. Eliminer les lignes avec des valeurs manquantes.
   5. Garder que les classes **'Fully Paid'** et **'Charged Off'**
   6. Transformer les dates
   7. Calculer le nombre de mois entre la date du prêt et la date de première ligne de crédit.
   8. Transformer des données quantitatives en format décimales (‘**int\_rate’** et ‘**revol\_util’**)
   9. Transformer la colonne ‘**sub\_grade’** en variable ordinale et
   10. Supprimer la variable grade
   11. Calculer le nombre de mois entre la date du prêt et la date de première ligne de crédit.
   12. Transformer la colonne ‘**sub\_grade**’ en variable ordinale et supprimer la variable grade
   13. Supprimer les observations **NONE** de la variable ‘**home\_ownership’**.
   14. Extraire les observations extrêmes dans un jeu de données distinct.
2. ***Transformations II :***
   1. Encoder les variables catégorielles en methode one-hot-encoder.
   2. Regrouper les classes rares de la variables ‘**home\_ownership’** dans une seule classe **'OTHER'**
   3. Appliquer la standardisation et normalisation:
      * StandardScaler pour standardiser les variables,
      * la methode Yeo-Johnson pour la normalisation des variables continues (non ordinales),