

**Projet de Scoring**

Session 2019 - 2020



Plan

I ).

La première étape dans tout projet de data science est l’exploration de nos données. Cette étape est primordiale nous permet à la fois, de comprendre nos données, mais aussi d’identifier des patterns d’analyses, des problèmes etc.

A partir de la Figure 1, on peut déjà remarquer :

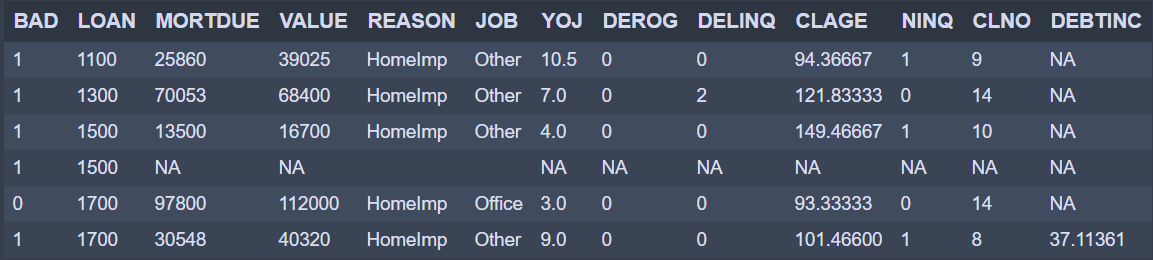
* des valeurs manquantes. Il nous faudra donc comprendre ces valeurs manquantes, et surtout les retraiter.
* des variables catégorielles (exemple pour la colonne JOB)
* des variables continues (exemple pour la colonne CLAGE)

Figure 1 : premières lignes du dataset

Tout d’abord, commençons par analyser nos variables catégorielles. En premier lieu, voici Figure 2 la distribution de notre variable cible, c’est-à-dire la variable qui nous indique si le détenteur de ce en question a fait défaut ou non. On sait que le fait d’avoir un défaut (i.e. BAD = 1) est censé être un évènement rare, et donc en général plus dur à modéliser. Avoir donc une idée sur la distribution des modalités de notre variable cible nous donnera des pistes sur la méthode à considérer pour modéliser cet évènement.

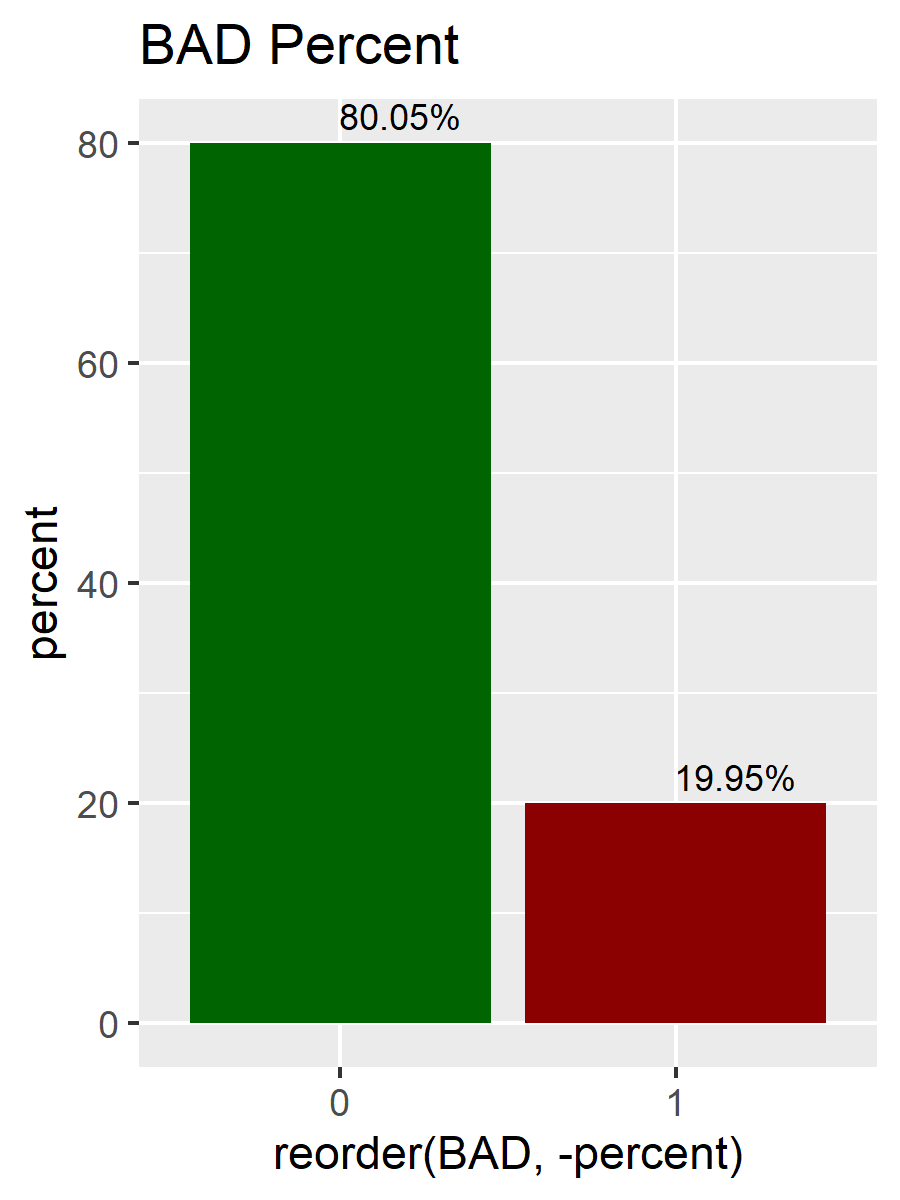
On peut s’estimer heureux car dans notre dataset, l’évènement faire défaut semble être commun avec environ 20% d’occurrence. Cela facilitera donc la modélisation.

Figure : Distribution de notre variable cible

Ensuite, il nous faut analyser toutes les autres variables catégorielles. Dans le dataset, on a considéré 5 variables catégorielles, comme on peut le voir dans sur Figure 3. Pour les variables REASON et JOB, aucun doute, ce sont bien des variables catégorielles. Mais pour les 3 autres, on a des variables numériques non continues. Ainsi, nous avons décidé de les considérer comme des variables catégorielles, en tout cas pour la partie exploration des données, car elles imposent une notion de rang (des variables ordinales). Par exemple, la variable DELINQ désigne le nombre de d’impayés pour l’observation. On peut alors classer les individus selon le nombre d’impayés.

Passons maintenant à l’analyse de la Figure 3. Tout à gauche, on peut observer la distribution des modalités de chaque variable. On remarque dans un premier temps que les valeurs manquantes pour les variables REASON et JOB ne sont pas représentées par des NA

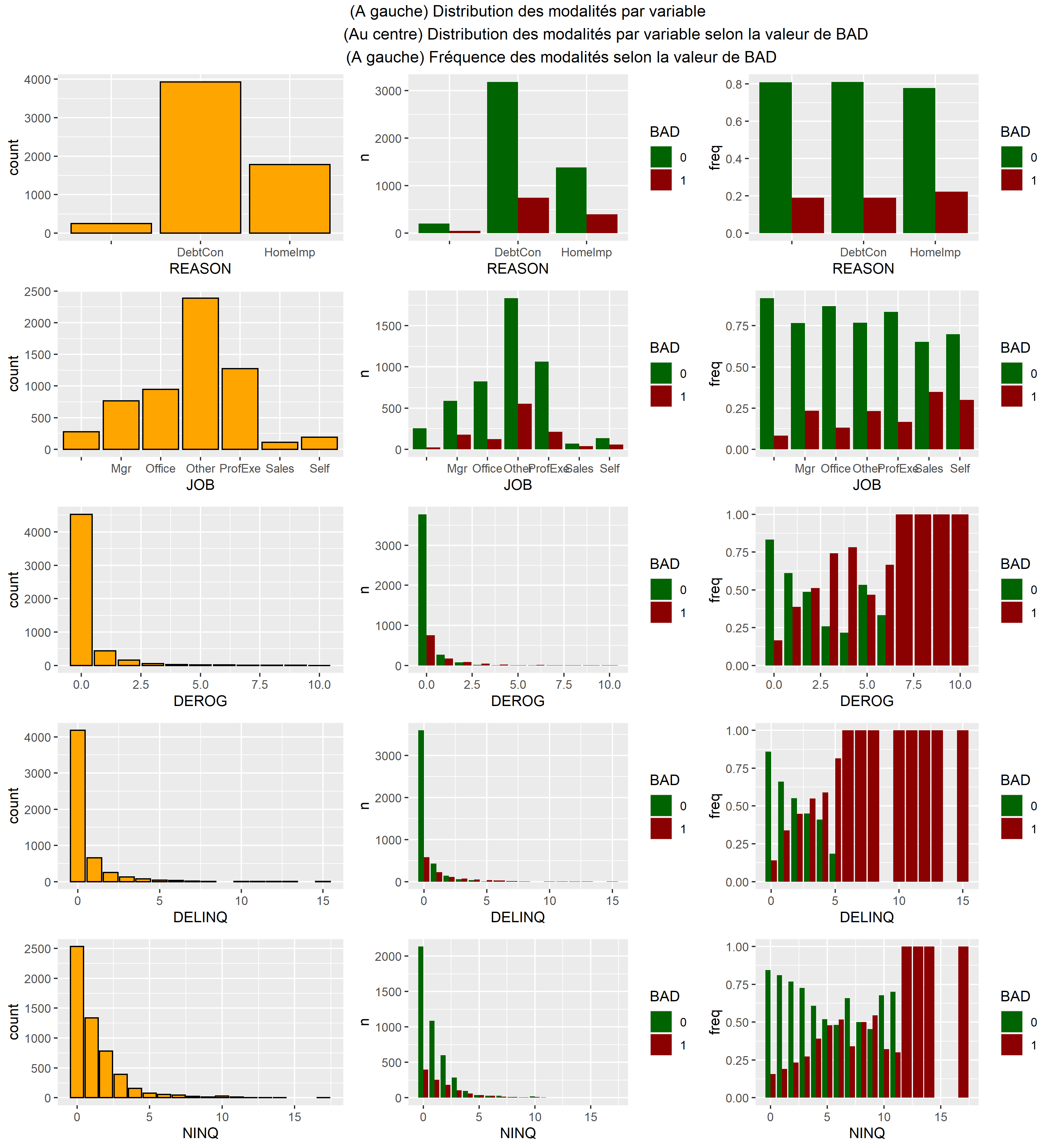
mais par une absence de valeur. Par conséquent, on a environ 3% de valeurs manquantes pour la variables REASON et à peu près 5% pour la variable JOB. Par rapport à la variable DEROG, on se rend compte que la plus part des observations n’ont pas eu de problème de paiment, et que le fait d’en avoir beaucoup semble inhabituel.

Figure 3 : Distribution des variables catégorielles

Au centre, on peut observer la distribution en fonction de la valeur de la variable cible. C’est intéressant de considéré des distributions condistionnelles car elles peuvent nous permettre de rendre compte de différence de comportement selon la valeur de la cible. Dans notre cas, il n’est pas pertinent de regarder ces distributions car la fréquence des modalités de la variable BAD ne sont pas égales, et aucune caractéristique ne semble sortir du lot. C’est pourquoi, nous avons la colonne de droite, qui représente la distribution de chaque modalité en fonction de la valeur de la cible. On peut alors remarquer pour la variable DEROG :

* les valeurs extrêmes sont toutes associées à la modalité BAD = 1 (i.e. a fait défaut). Peut-être que les valeurs ne sont pas extrêmes et font juste partie des caractéristiques d’un individu qui fait défaut.
* plus le nombre de dérogation augmente, plus la fréquence de personne ayant fait défaut augmente. C’est intéressant, et pas anodin, car cela semble explicable. En effet, peut-être qu’avant de faire défaut, le client a plusieurs retards de paiement (90 jours selon Bâle III).

On remarque les même éléments pour les variables DELINQ et NINQ.

Annexes :