

**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 :

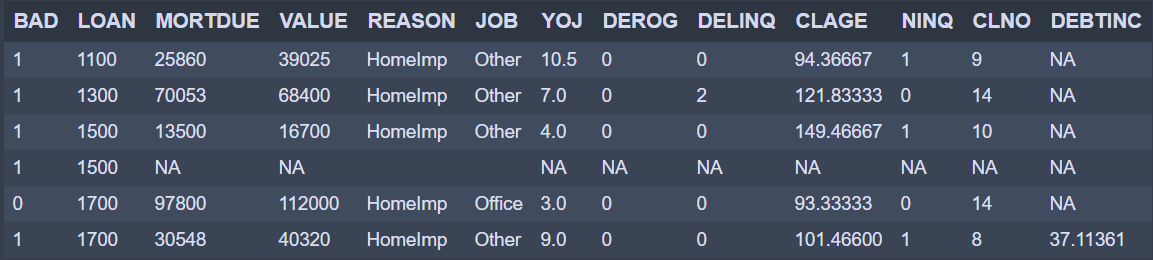
* des valeurs manquantes. Il nous faudra donc comprendre ces valeurs manquantes, et surtout les retraiter.
* des variables catégorielles (exemple pour la colonne JOB)
* des variables continues (exemple pour la colonne CLAGE)

Figure : 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.

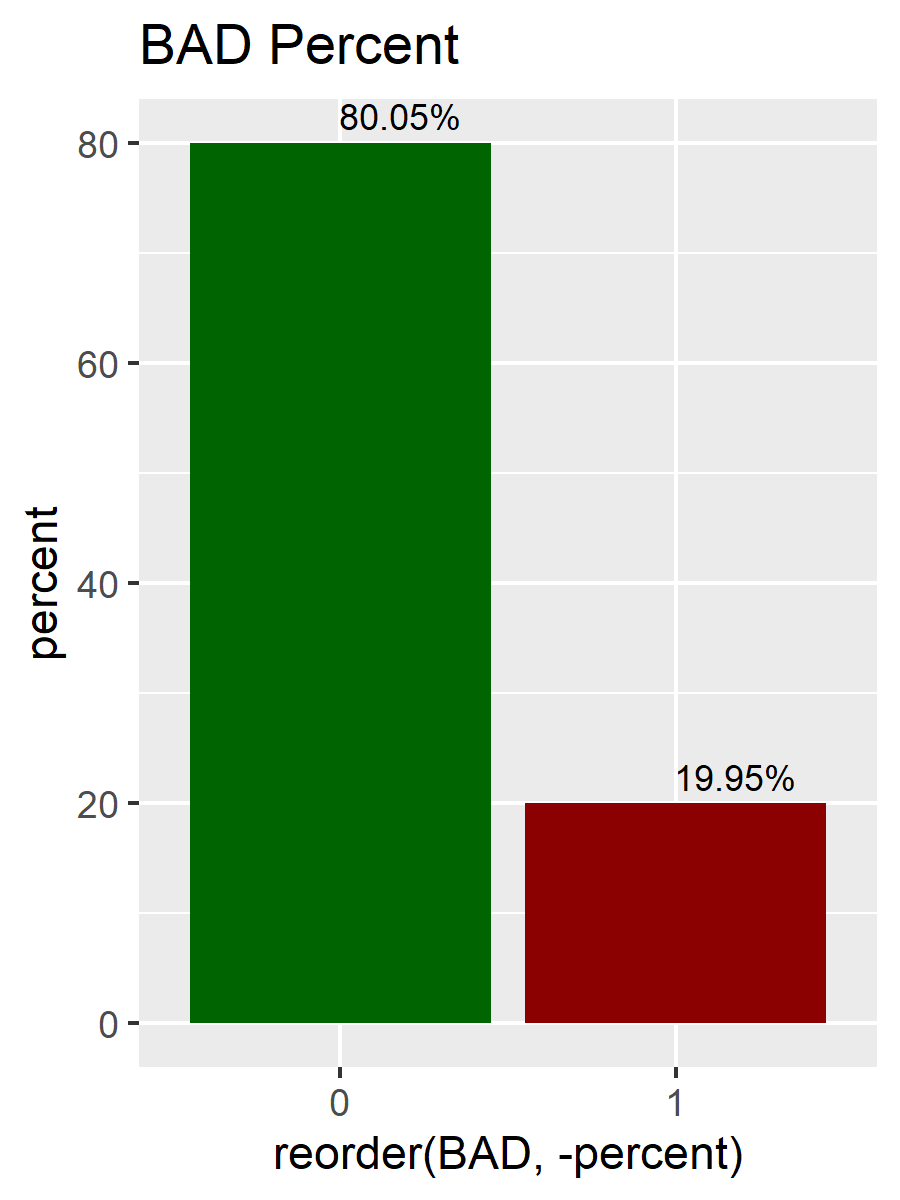
On peut s’estimer heureux car dans notre dataset, l’évènement faire défaut semble être commun avec environ 20% d’occurrence. Cela facilitera donc la modélisation.

Figure : Distribution des modalités pour la 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 tout 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

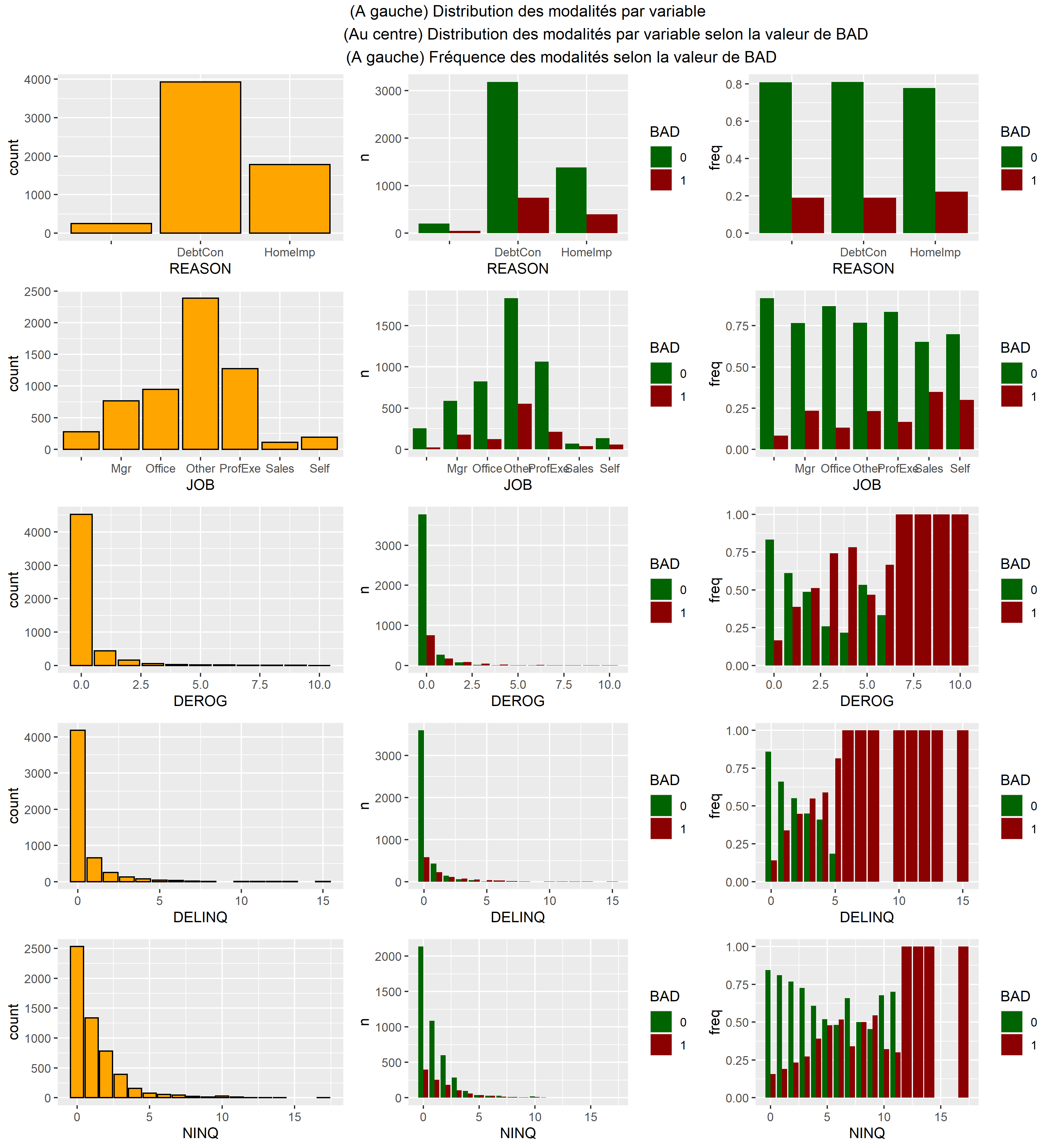


Figure : Distribution des variables catégorielles

Annexes :