

**Projet de Scoring**

Session 2019 - 2020



Plan

I ).

EDA

La première étape dans tout projet de data science est l’exploration de nos données. Cette étape est primordiale car elle 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 :

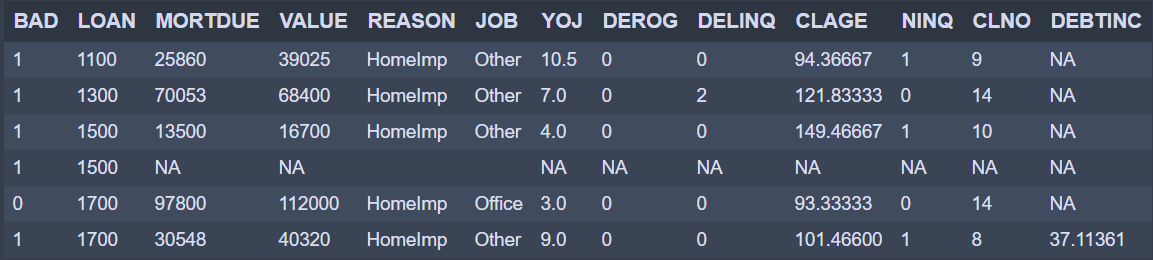
* des valeurs manquantes. Il nous faudra donc comprendre ces valeurs manquantes, et surtout les retraiter.
* des variables catégorielles (exemple pour la colonne JOB)
* des variables numériques, continues ou non (exemple pour la colonne CLAGE)

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 client 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.

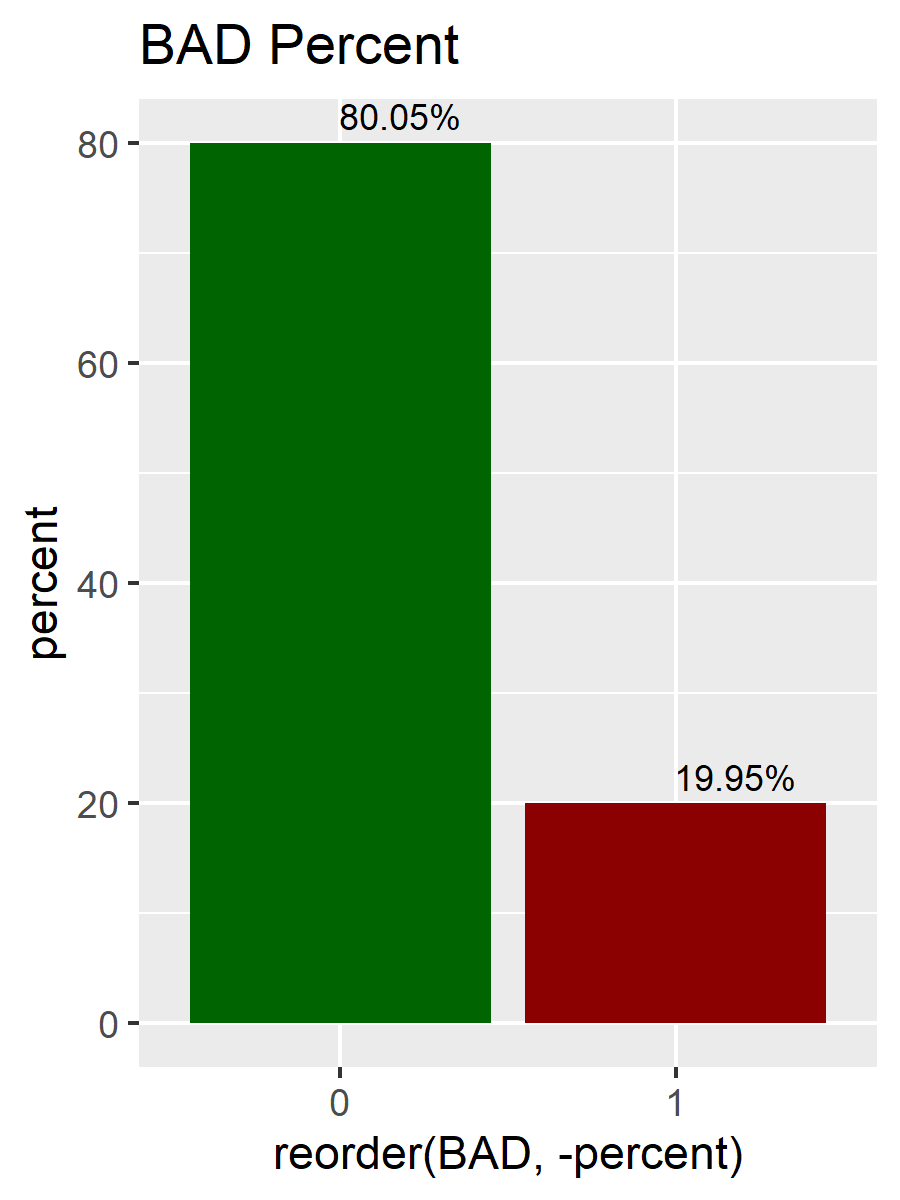
On peut s’estimer heureux car dans notre dataset, l’évènement faire défaut semble être commun avec environ 20% d’occurrence. Cela facilitera donc la modélisation.

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

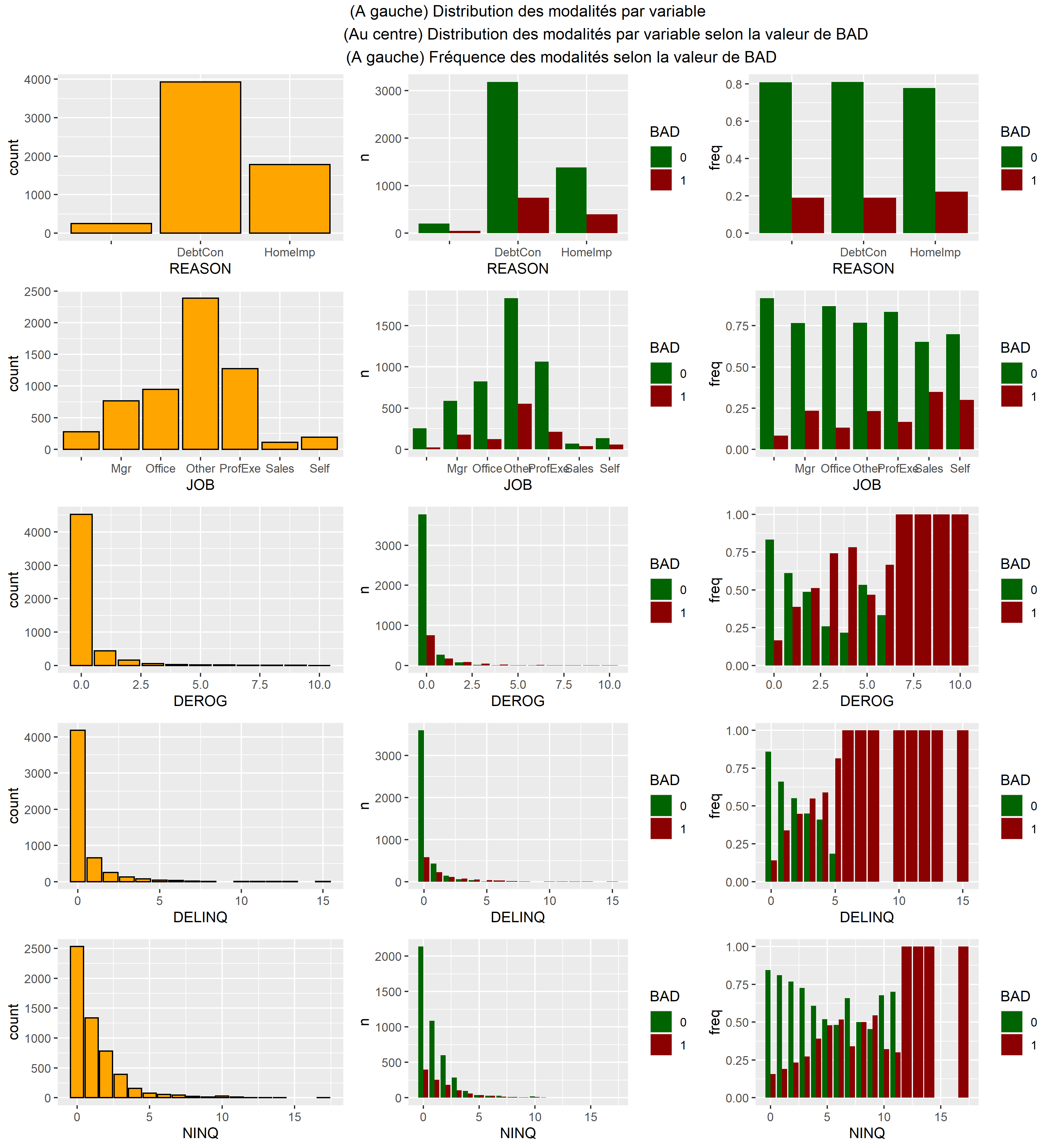
mais par une absence de valeur. Par conséquent, on a environ 3% de valeurs manquantes pour la variables REASON et à peu près 5% pour la variable JOB. Par rapport à la variable DEROG, on se rend compte que la plus part des observations n’ont pas eu de problème de paiement (DEROG = 0), et que le fait d’en avoir beaucoup semble inhabituel.

Figure : Distribution des variables catégorielles

Au centre, on peut observer la distribution en fonction de la valeur de la variable cible. Il est intéressant de considérer des distributions condistionnelles car elles peuvent nous permettre de rendre compte visuellement des différences 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 aucunes des caractéristiques ne semblent sortir du lot. C’est pourquoi nous avons la colonne de droite, qui représente la distribution par 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.

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

Passons aux variables numériques, en adoptant la même méthode d’analyse. On a décidé dans un premier temps de nous concentrer sur la distribution de ces variables.

C’est ce que nous avons représenté en Annexe 1. Par exemple, on peut se rendre compte que la plus part des crédits se situent entre 2 000 et 25 000 d’unité monétaire (aucune devise n’est spécifié), que le plus gros de la distribution de la valeur des maisons se situe entre 30 000 et 250 000. Chose intéressante, certaines variables ont des distributions asymétrique (avec une queue vers la droite), c’est-à-dire que la moyenne est supérieure à la médiane. C’est ce qu’on observe pour la variable VALUE, LOAN, MORTDUE et YOJ. Probablement qu’effectuer une log transformation de ces variables pourrait-être intéressant pour l’analyse, car elle nous permettrait de transformer la distribution en gaussienne. Finalement, on observe un certain nombre de valeurs aberrantes en apparance, en observant les queues des distributions. On verra prochainement à l’aide de boxplots s’il s’agit réelement d’outliers.

Dans un deuxième temps, toujours sur l’Annexe 1, on a regardé la distribution selon la valeur de la variable BAD. Initialement, nous voulions savoir si il était possible de dissocier les distributions selon la valeur de la variable cible. Dans notre cas, aucune distribtuion n’est différenciable selon la valeur de BAD.

On a énoncé précédemment la possibilité que nos variables contiennent des valeurs aberrantes, et donc dans la Figure 4, nous avons tracé les boxplots de nos variables avec et sans la considération de notre variable cible.

Et en effet, toutes les variables ont bien des outliers. Mais, pour nous, aucune ne représente des valeurs aberrantes. En effet, nous n’avons pas le contexte de ces prêts, donc on ne peut pas dire si un LOAN a 80 000 est une aberration. Si cela avait été un prêt à la consommation pour de l’électroménager par exemple, alors probablement que 80 000 serait aberrant. Mais pour une voiture ce n’est pas le cas. Pareil pour la VALUE, car une maison peut très facilement valoir 750 000. Pour YOJ, travailler 40 ans pour la même entreprise n’est pas commun, mais cela ne peut être considéré comme absurde. Enfin pour finir, la variable DEBTINC est exprimée en pourcentage, donc un ratio de 200 (i.e. une dette 3 fois supérieure au revenu) n’est pas aberrant non plus.

Ce qui est maintenant pertinent, c’est de voir à quelle modalité de BAD ces valeurs particulièrement fortes appartiennent. C’est pourquoi nous pouvons observer un boxpot par variable selon la valeur de la cible. Au-delà de pouvoir attribuer un outlier à une modalité de la cible, elle nous permet d’observer plus facilement la répartition des valeurs qu’avec les histogrammes de l’Annexe 1.

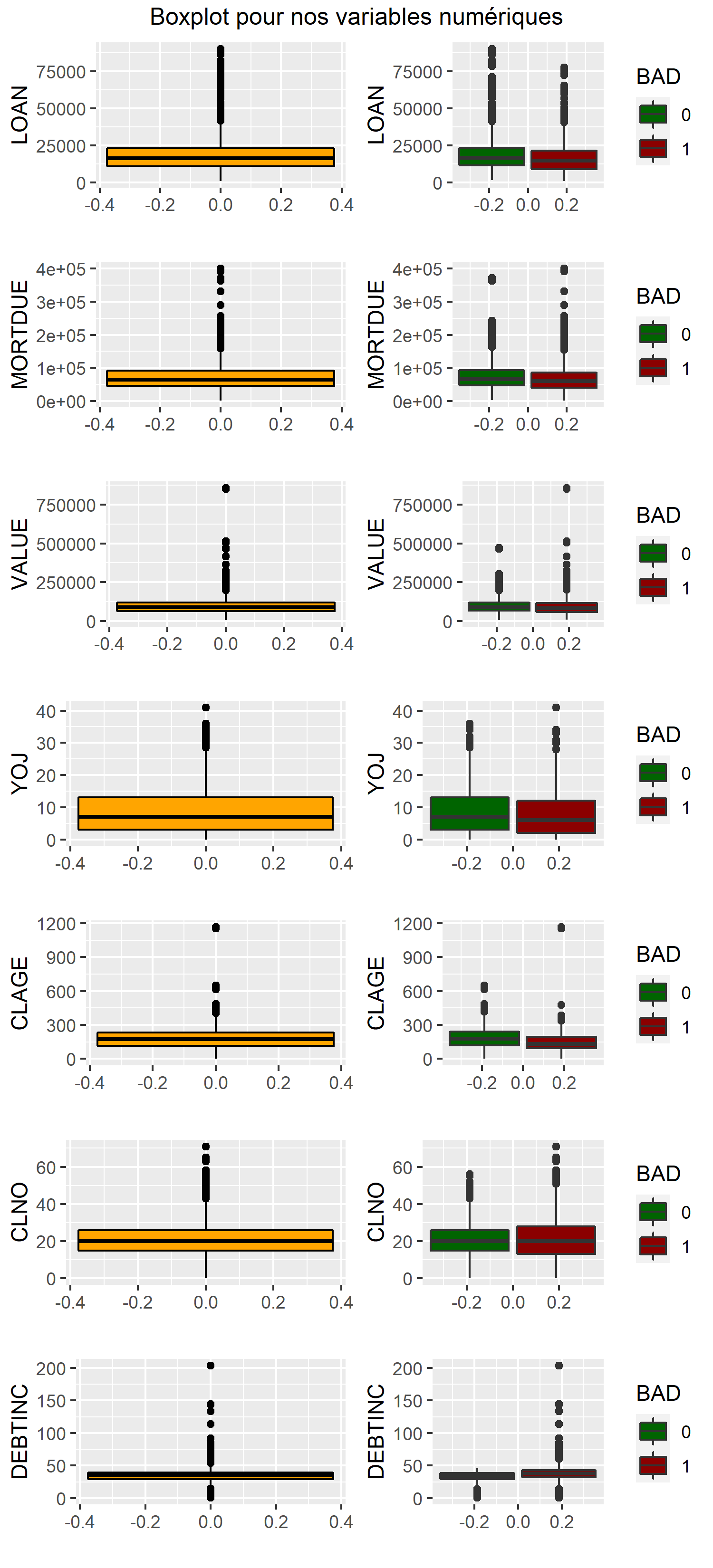


Figure : Boxplot

De cette façon, on a l’occasion d’expliquer certaines observations qui peuvent sembler aberrantes comme étant une caractéristique d’un client qui fait défaut. Mais comme expliqué lors de l’étude des histogrammes (Annexe 1), il est difficile de différencier la distribution selon la modalité de BAD. Néanmoins, nous avons un contre-exemple avec la variable DEBTINC. Une personne qui fait défaut a tendance à être surendettée, et c’est ce qu’on peut remarquer sur le boxplot : toutes les valeurs jugées aberrantes pour cette variable sont associées au fait de faire défaut. Cela pourrait être une des caractéristiques importantes pour prédire le défaut d’un client.

Ainsi, de cette analyse, on peut en ressortir 2 grands points :

1. Graphiquement, c’est compliqué de déterminer une caractéristique associée au défaut ou au non-défaut. Ce qui pourrait nous être défavorable lors de l’utilisation de modèles linéaires simples.
2. Lors de notre analyse, nous nous sommes d’avantage reposée sur une connaissance métier que sur une interprétation graphique de nos valeurs.

II) PREPROCESSING & ANALYSE

PREPROCESSING

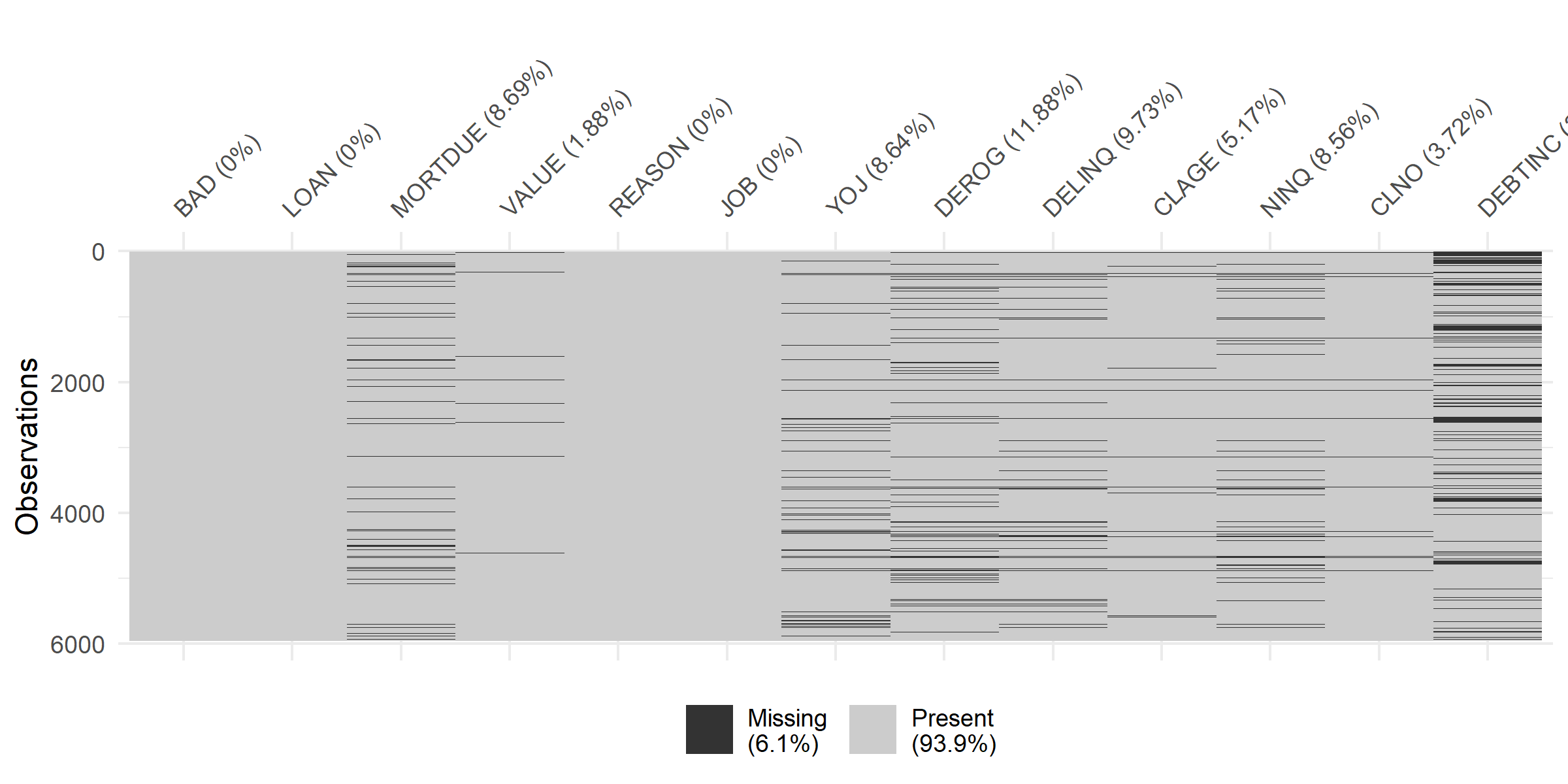
Pour être honnête, l’analyse que nous venons de vous présenter reste incomplète étant donné que nous n’avons pas pris en considération, pour les variables numériques du moins, les valeurs manquantes. A l’aide de la Figure 5, nous pouvons observer la disposition des valeurs manquantes, numériques uniquement :

Figure : Position des valeurs manquantes par variable numérique

Nous pouvons alors remarquer que la variable avec le plus de valeur manquante est DEBTINC, qui malheureusement était la seule variable avec laquelle nous avions pu associer certaines de ses valeurs à la variable cible. Or, le raisonnement est peut-être biaisé par le manque de valeur. Deuxièmement, on peut observer des « patterns » de valeurs manquantes. C’est-à-dire que visuellement, on a des plusieurs traits qui traverse toutes les variables numériques. Cela semble nous indiquer que nous sommes en présence pour la plupart des valeurs manquantes à des MAR (missing at random) et des MNAR (missing not at random). Traiter ces valeurs manquantes sera donc plus compliqué, surtout si on souhaite éviter de générer du biais pour notre future analyse.

Dans un premier temps, on a cherché savoir si les valeurs manquantes par colonnes étaient plus associées aux observations qui ont fait défaut ou non. Par conséquent, nous avons tracé la disposition des valeurs manquantes par modalité de BAD (Annexe 2). Premier fait marquant, nous avons 65% de valeurs manquantes pour la variable DEBTINC quand BAD = 1, contre 10% quand BAD = 0, alors que sans considérer la valeur de BAD, nous avions 23% de valeurs manquantes. Ainsi, la plupart des NA de cette colonne sont liées au fait d’avoir fait défaut. Probablement que les personnes qui ont fait défaut n’ont pas fourni leurs informations à propos de leur endettement ; mais, quand on supprime les lignes qui possèdes plus de 37%[[1]](#footnote-1) de valeurs manquantes, nous n’avons plus ce déséquilibre de NA pour la colonne DEBTINC (Annexe 3), sans pour autant détraquer les proportions[[2]](#footnote-2) pour les modalités de la variable cible.

Il nous faut donc traiter les valeurs manquantes. Après avoir supprimé les lignes qui possédaient plus de 37% de NA, nous avons :

* Remplacé par la moyenne pour les variables MORTDUE, VALUE, YOJ, CLAGE et DEBTINC (variables numériques continues)
* Remplacé par la médiane pour les variables DEROG, DELINQ, NINQ et CLNO (variables numériques non continues)
* Remplacé par la valeur « missing\_value » pour les variables JOB et REASON (variables catégorielles)

Pour finir on a cherché à comparer l’effet d’une log transformation pour certaines variables sur la forme de leur distribution (Annexe 4) : Le fait d’appliquer un log nous permet de fit une gaussienne ? On a donc tracé un QQPlot avant et après la log transformation, qui permet de comparer la position de notre distribution par rapport à différent quantile de la distribution gaussienne, il permet donc de comparer la similarité entre 2 distributions (ici une loi normale et la distribution de notre variable). Avant la transformation, on observe un QQPlot avec une courbe « convexe », ce qui représente bien la distribution asymétrique avec une queue vers la droite. Il y a énormément à dire sur ces graphiques, mais on va se concentrer sur le graphique après la log transformation. Nous remarquons alors que pour la variable YOJ[[3]](#footnote-3), cette log transformation nous éloigne d’une distribution gaussienne, avec beaucoup de valeurs nulles. On n’appliquera donc pas cette transformation sur cette variable (en plus de la distribution de sa transformation qui s’éloigne d’une gaussienne). Pour toutes les autres variables, les queues s’éloignent un peu de la loi normale, mais l’ensemble fit bien avec la gaussienne.

C’est pourquoi cette log transformation a été appliqué aux variables LOAN, VALUE et MORTDUE, et pas YOJ pour les raisons qui ont été cités précédemment.

Maintenant que nous avons traiter les valeurs manquantes, la suite du projet est de scaler nos données pour éviter les problèmes d’échelles, et d’encoder nos variables catégorielles.

Les deux variables facteurs de notre jeu de données, à savoir REASON et JOB, sont des variables catégorielles disposant de peu de valeurs uniques. En effet, la variable REASON contient les modalités suivantes :

* HomeImpt
* DebtCon
* missing\_value

Et la variable JOB :

* Office
* Sales
* Mgr
* ProfExe
* Other
* missing\_value

Ainsi, ces deux variables sont éligibles à deux techniques simples de préprocessing permettant de les rendre compréhensible par un modèle linéaire : la labellisation et le OneHotEncoding (aussi connu sous le nom de dummying). Mais nous avons de la chance car le merveilleux langage R permet de typer les variables catégorielles comme factor, ce qui nous offre la possibilité de passé outre l’encodage.

Après l’encodage, nous avons appliqué du standard scaling (retrait de la moyenne, puis division du tout pas l’écart-type) à nos variables numériques dans le but d’obtenir une meilleure interprétabilité de nos variables, et surtout des paramètres des régressions linéaires que nous allons appliquer par la suite. En effet, si les variables disponibles sont à la même échelle, alors les coefficients de la régression linéaire le sont aussi, et peuvent être comparés l’un à l’autre.

En Annexe 5, nous pouvons constater que le standard scaling conserve la distribution de nos variables, et que seule l’échelle est impactée par cette transformation.

ANALYSE

Maintenant on a des données propres, qui vont nous servir à modéliser la probabilité de défaut. Mais avant, on va chercher à observer les liens entre nos variables, pour avoir à la fois une direction pour la modélisation et surtout des pistes d’analyse. L’idée de ce point et de pouvoir analyser les variables entres elles, la plupart du temps face à la variable cible mais pas uniquement. Cette analyse n’a pas pour but de faire de la sélection variable, mais de nous donner les outils pour comprendre pourquoi telle ou telle variable n’a pas été sélectionné par nos méthodes de feature selection.

Le premier lien que nous avons décidé d’étudier était celui entre toutes nos variables catégorielles et la variable cible. C’est pourquoi nous nous sommes orientés vers un test du Chi² d’indépendance, qui nous permet de tester l’absence de lien entre 2 variables, dans notre exemple JOB-BAD et REASON-BAD. On va donc tester l’hypothèse suivante H0 : « X et Y sont indépendants » (i.e. les valeurs de Y dépendent de X).

Dans notre cas, nous avons obtenu une p-value de 0.0125 pour le test de l’hypothèse « REASON et BAD sont indépendants », et une p-value de 0 pour le test « JOB et BAD sont indépendants ». Par conséquent, pour un seuil de confiance de 5%, nous sommes obligés de refuser l’hypothèse H0, car la p-value est inférieure à notre seuil. On peut alors en conclure que nos 2 variables explicatives REASON et JOB ne sont pas inintéressantes à utiliser car les valeurs que peut prendre la variable BAD dépend des valeurs que prennent REASON et JOB (avec un risque de se tromper de 5%).

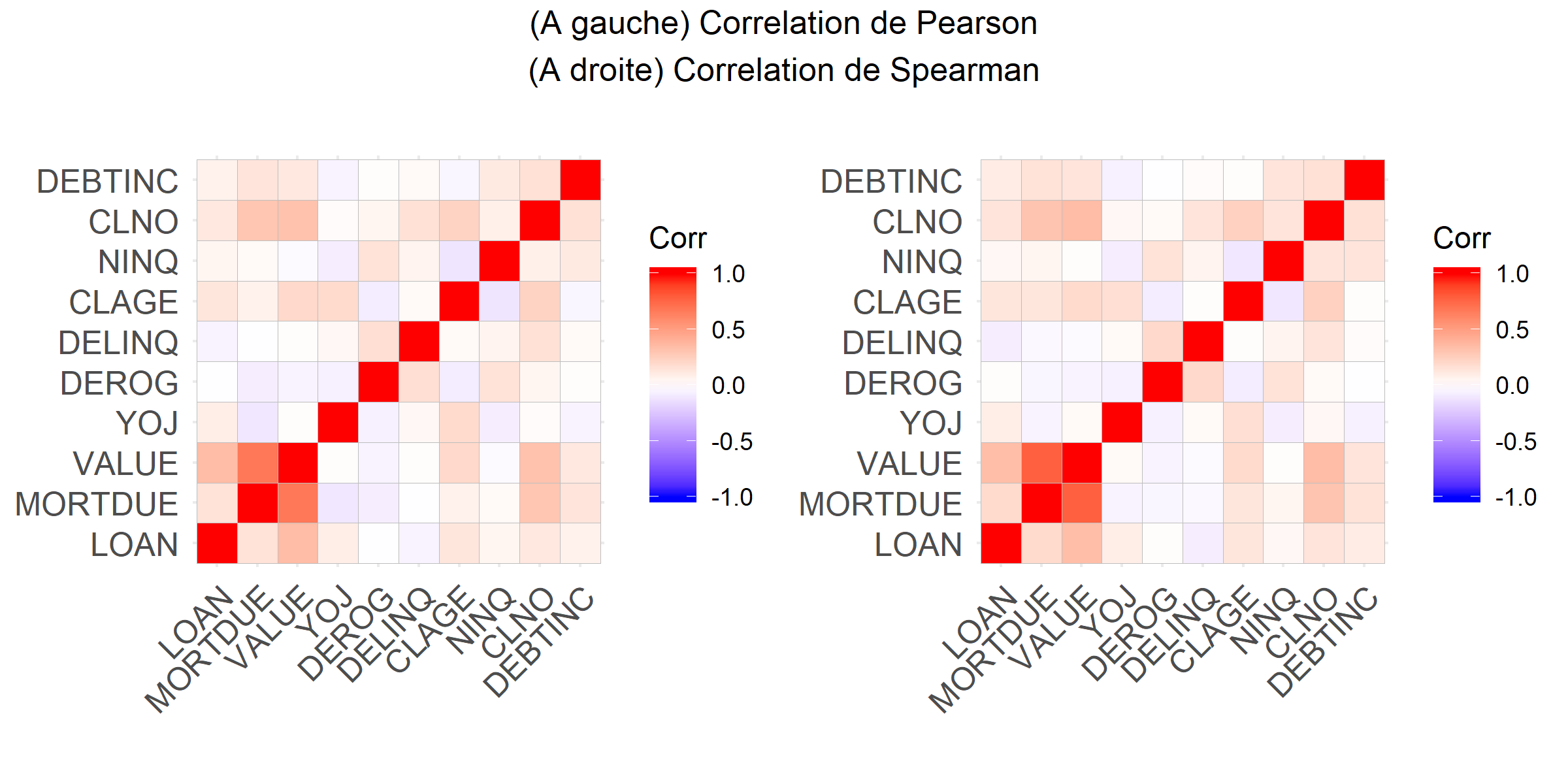
Le deuxième lien que nous avons décidé d’étudier est simplement la corrélation entre toutes les variables numériques (on exclut donc la variable cible). C’est ce que nous pouvons observer à travers la Figure 6 ci-dessous. Nous avons choisi de calculer aussi la corrélation de Spearman car comme l’analyse exploratoire nous l’avait montrée, nous avions des outliers (même si d’un point de vu « Business » ce n’était pas des observations extrêmes) et par conséquent, pour éviter d’avoir un biais dans notre analyse nous avons décidé de compléter avec une corrélation du rang ; étant donné que nous avions décidé de garder ces observations dans notre jeu de données. A première vue, nous n’avons pas de différence de couleur entres les cellules de nos 2 graphique de corrélation, ce qui est une bonne chose étant donné qu’on pas pouvoir se concentrer sur un des graphiques. On se rend compte que la variable VALUE fortement corrélé à la variable MORTDUE, ce qui est logique, étant donné VALUE exprime la valeur de la maison, et que MORTDUE représente la valeur du prêt hypothécaire. Il faudra faire attention plus tard à vérifier qu’il n’y a pas de colinéarité entre ces variables. On a aussi une corrélation élevée entre LOAN et VALUE, mais mise à part ces 2 cas, rien d’anormale. Pour compléter cette analyse, voici un Variance Inflation Factor analysis (VIF), qui va nous permettre d’observer les problèmes de colinéarité et de multi-colinéarité. On cherche donc à savoir si une variable peut s’exprimer en fonction des autres, et est peut alors biaiser notre modèle et notre analyse. Voici Figure 7 ci-dessous les résultats de notre VIF :

Figure : Corrélation plot

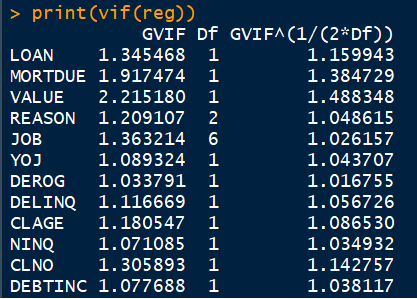


Figure : Sortie R du VIF

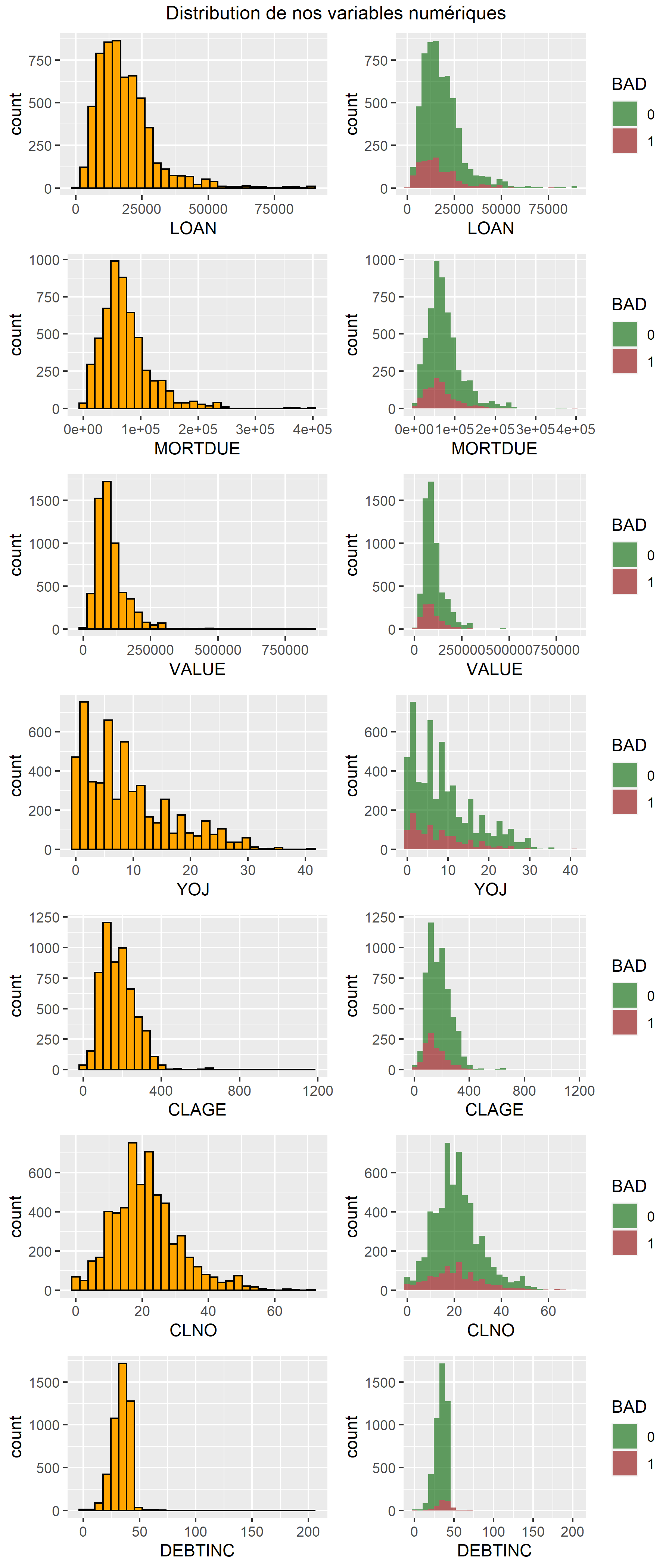
Le VIF permet « d’estimer de combien la variance d’un coefficient est augmentée en raison d’une relation linéaire avec d’autres prédicteurs »[[4]](#footnote-4).

On va regarder la sortie GVIF. Si la valeur est supérieure à 2.5 ou 5 (cela dépend des statisticiens), alors on a de la multi-colinéarité sévère dans nos variables. Dans notre cas aucune valeurs de dépasse les 2.2 pour la variable VALUE. Cela veut dire que la variance de coefficient est de près 2 fois plus forte que la variance que l’on aurait pu observer si ce facteur était absolument décorrélé de toutes les autres variables. Cela explique alors la forte corrélation que nous avons pu observer entre VALUE et MORTDUE. Mais son VIF est inférieur à 2.5, donc nous allons tout de même la conserver et nous verront si cette variable sera sélectionné ou pas par le processus de feature selection.

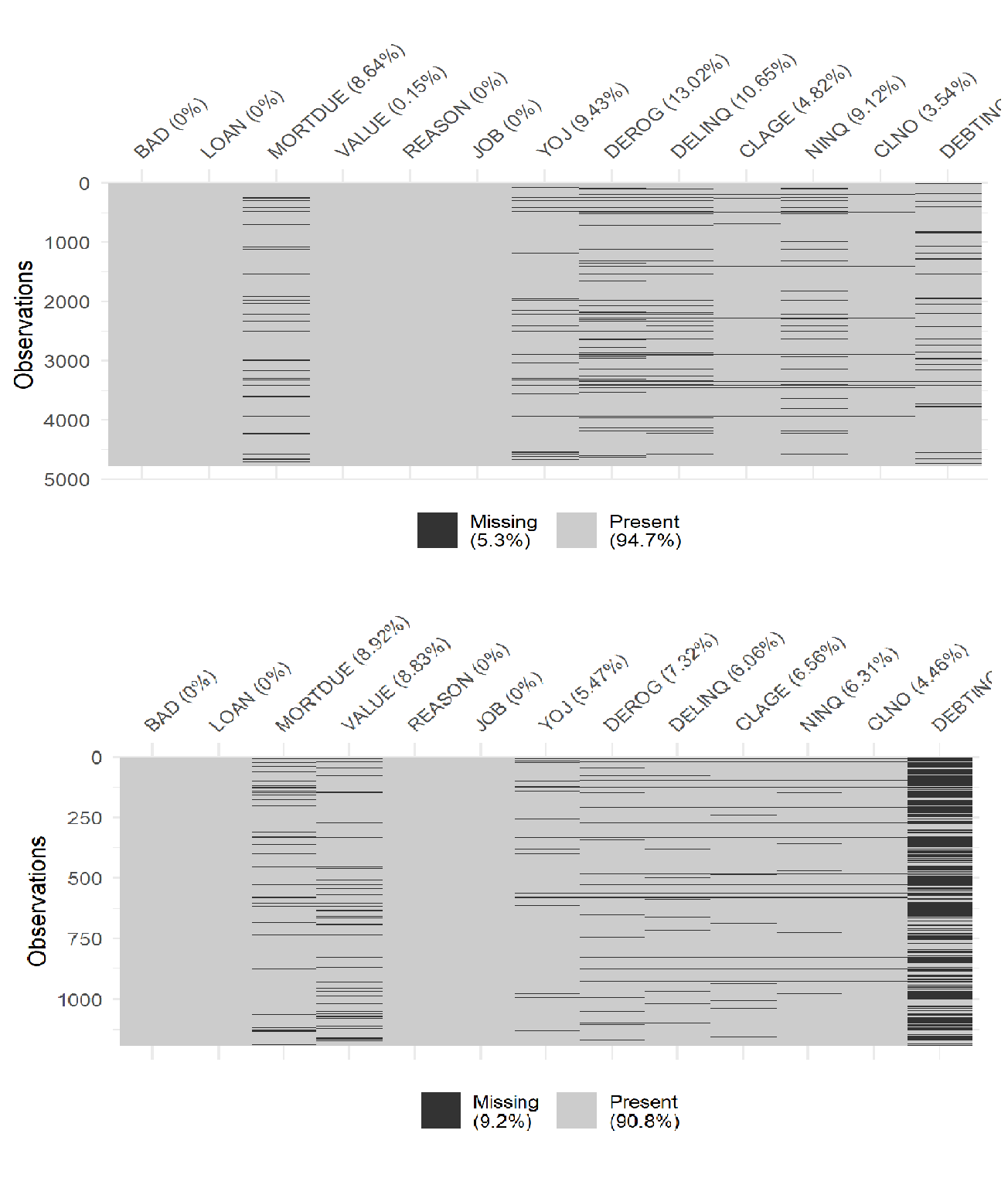
Nous pouvons aussi remarquer que toutes les autres variables ont un VIF qui avoisine 1, ce qui caractérise des variables, peu corrélé entres-elles, ce que nous avions pu observer Figure 6.

Hyp normalite « verified » avec preprocessing et log transfo ^^

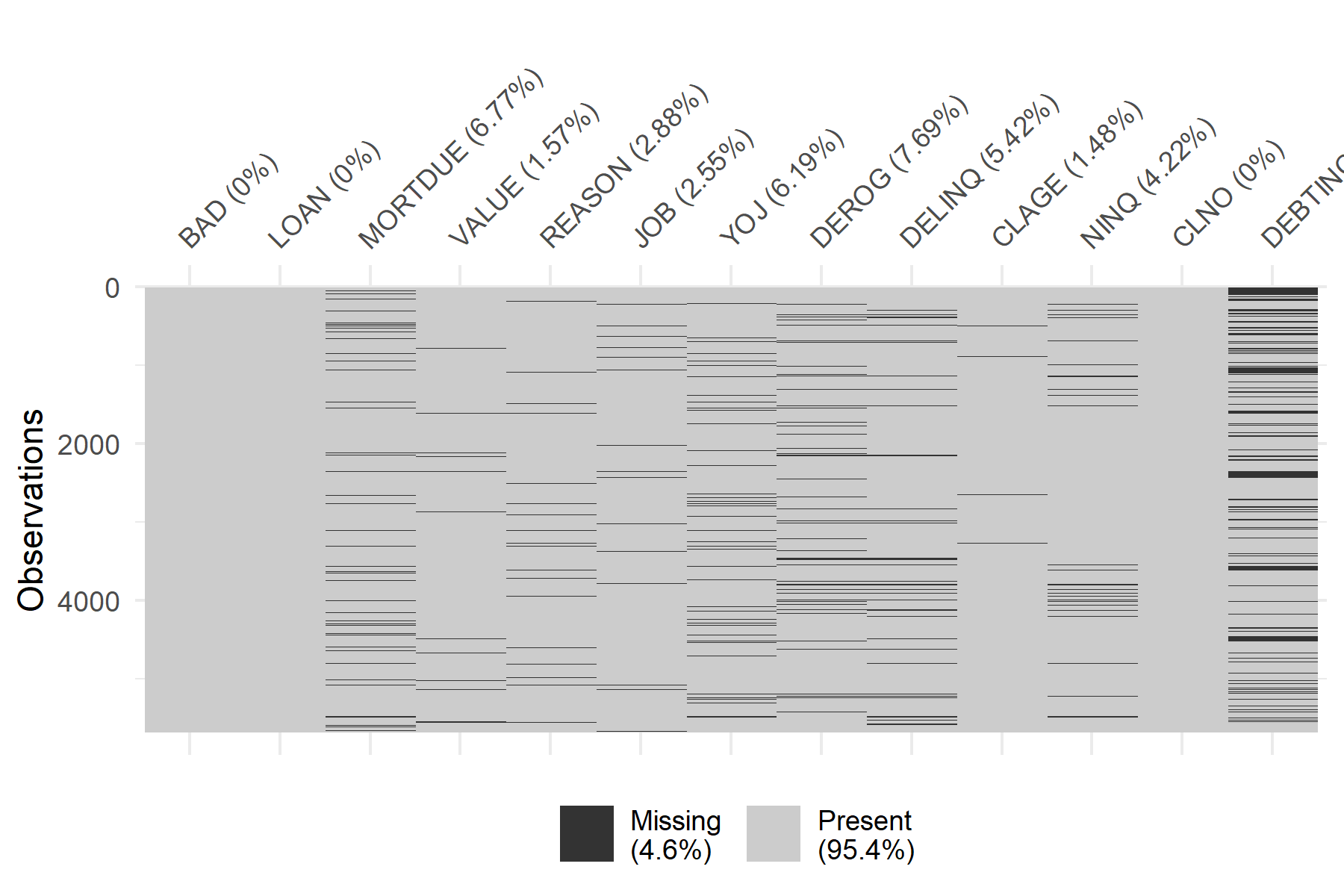
Annexes :



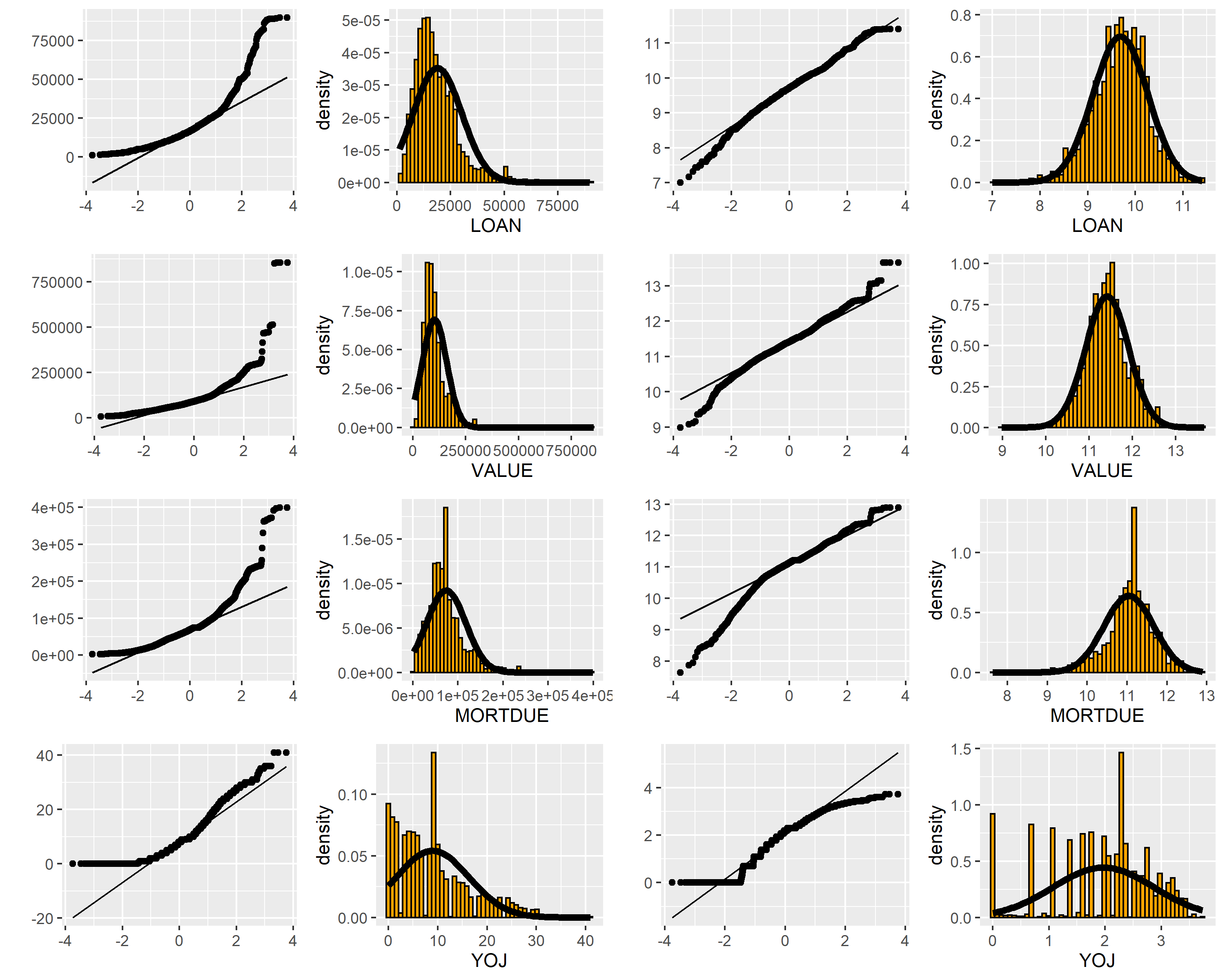
Annexe : Distribution des variables continues



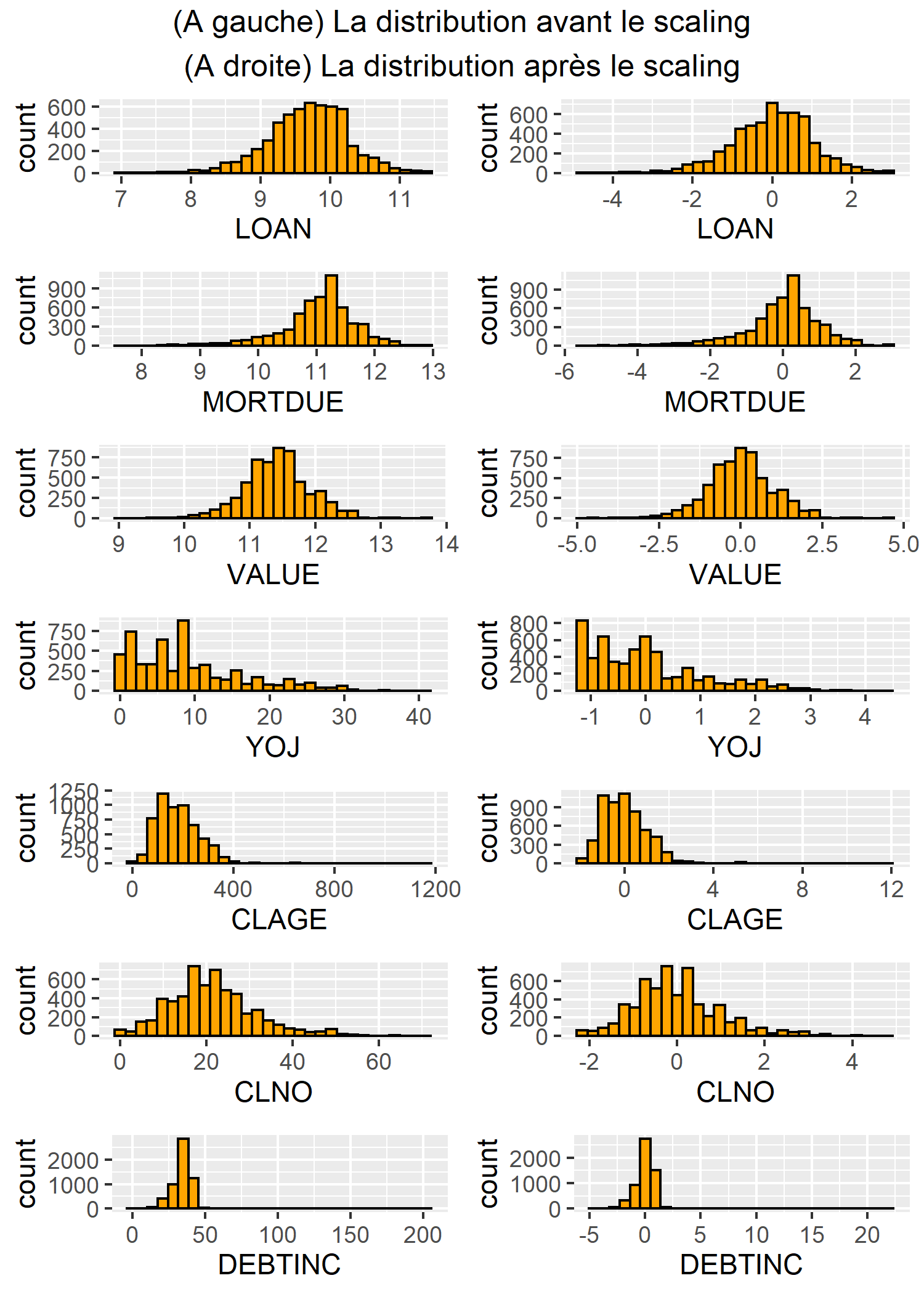
Annexe : (En haut) Valeurs manquantes pour BAD = 0, (En bas) Valeurs manquantes pour BAD = 1



Annexe : Variables manquantes après suppresion des lignes avec 37% ou plus de NA



Annexe : QQPLot avant / après log transformation



Annexe 5 : Distribution avant / après le scale des variables numériques continues

1. Ce seuil correspond à la valeur inférieure du ratio 5/13, choisi pour éviter d’avoir des valeurs manquantes simultanément dans les colonnes DEROG, DELINQ, CLAGE, NINQ et CLNO. [↑](#footnote-ref-1)
2. ~80% pour BAD = 0 [↑](#footnote-ref-2)
3. La variable YOJ possède des valeurs nulles, donc nous avons préféré faire une log + 1 transformation. [↑](#footnote-ref-3)
4. <http://larmarange.github.io/analyse-R/multicolinearite.html> [↑](#footnote-ref-4)