Pour que les classes d’une variable catégorielle soient fiables il faut qu’elles soient :

* **Significativement corrélées avec la variable cible :**

Dans notre cas, il faut que le taux de défaut soit significativement différent pour chaque classe, donc qu’il n’y ait pas indépendance entre les classes et la cible. Pour établir ce lien nous avons procédé à un test du khi deux. Deux variables ne sont, à priori pas indépendantes si on rejette H0, autrement si, la p-value tend vers 0. Ici, le seuil choisi était 5%.

METTRE GRAPHIQUE AVEC KHIDEUX SIGNIFICATIF VS NON SIGNIFICATIF

* **Stables dans le temps :**

Cela signifie que les modalités ne doivent pas se croiser. Par exemple, considérons une variable avec 2 modalités, et le taux de défaut pour chaque modalité. Si la modalité 1 à un taux de défaut de 10% en 2013 et la modalité 2 un taux de défaut de 15% en 2013, il faut que pour toutes les autres années, la modalités 2 est un taux de défaut supérieur à la modalité 1. Si ce n’est pas le cas, nous ne pouvons pas considérer la variable en question car l’information qu’elle nous donne sur la variable cible n’est pas robuste.

METTRE GRAPHIQUE AVEC STABILITE VS INSTABILIE

**Discrétisation des variables continues:**

Afin de discrétiser les variables nous avons suivi la méthodologie suivante :

1. Répartir les observations dans des catégories en nous basant sur les déciles, les quintiles, les quantiles ou la médiane ;
2. Pour chaque variables nous avons testé chaque technique (ie. déciles, quantiles, quartiles ou médiane), et nous avons choisis la répartition qui vérifiait :

* Une statistique du khi deux significative au seuil de 5% ;
* Des modalités stables dans le temps.

**Vérification et regroupement des modalités des variables qualitatives/catégorielles :**

Pour toutes les variables catégorielles initiales nous avons vérifié que leur répartition dans les classes dépendait de la cible (khi deux) et que les modalités étaient stables dans le temps.

Une variable au caractère particulier à retenu notre attention, la variable « emp\_title ». Cette variable correspondait au titre d’emploi que chaque client fourni lors de sa demande de prêt (exemple : manager, doctor). Cette variable possédait des fautes de frappes et n’était pas homogène. A l’aide des expressions régulières (regex) nous avons mis chaque individu dans différentes catégories d’emploi définies par la Standard Occupationnel Classification (SOC2018). Cette classification est utilisée par le « U.S bureau of labor statistics » pour classifier les emplois, et donc les individus, dans le cadre du traitement des données. Après avoir effectué ce traitement, la variable classé a été traité comme les autres variables catégorielles.