**Projet Interpromo 2022**

**Crédit Agricole**

**Nettoyage des données**

Nous avons commencé le projet par nettoyer le jeu de données, qui est une étape très importante.

Cette première étape consiste à effectuer des premières analyses sur les variables pour mieux les connaître. Pour cela, nous avons réalisé plusieurs graphiques simples sur certaines des variables afin d’observer la répartition des données, et observer un lien éventuel avec la variable de décision « TOP\_FRAUDE ».

Nous avons également regardé quelques corrélations entre les variables, ce qui pouvait nous aider à remplir des valeurs manquantes dans certaines variables.

Beaucoup de variables avaient des données manquantes. Pour ces variables, nous avons regardé le pourcentage de données manquantes pour savoir quelle méthode appliquer ensuite. Une variable avec peu de données manquantes a été remplie par la valeur 0, ou par une autre méthode comme l’utilisation de la médiane ou la moyenne (pour une variable quantitative par exemple). Pour les variables qualitatives, le mode (valeur la plus fréquente) peut être utilisé, ou encore la valeur précédente ou suivante.

Les variables qualitatives ont été binarisées (à l’aide de la fonction get\_dummies). Cette méthode consiste à diviser une variable en d’autant de variables binaires qu’il y a de modalités.