

La base de données *visaprem* contiennent les profils de 1294 clients d'une institution bancaire française avant la zone euro. Les données ont été collectées lors d'une enquête mensuelle au mois M . Les variables incluses dans le fichier initial *visaprem* sont les suivantes :

Identifiant	Libellé
<i>matric</i>	Matricule (identifiant client)
<i>sexe</i>	Sexe, 0 pour homme, 1 pour femme
<i>age</i>	Âge en années
<i>famiq</i>	Situation familiale : un parmi mariée (<i>mar</i>), célibataire (<i>cel</i>), divorcée (<i>div</i>), union libre (<i>uli</i>), séparée (<i>sep</i>) ou veuve (<i>veu</i>)
<i>relat</i>	Ancienneté de relation en mois
<i>pcspq</i>	Catégorie socio-professionnelle (code numérique de l'INSEE)
<i>impnbs</i>	Nombre d'impayés en cours
<i>rejets</i>	Montant total des rejets en francs
<i>opgnb</i>	Nombre d'opérations par guichet dans le mois
<i>moyrv</i>	Moyenne des mouvements nets créditeurs des 3 mois en milliers de francs
<i>tavep</i>	Total des avoirs épargne monétaire en francs
<i>endet</i>	Taux d'endettement
<i>gaget</i>	Total des engagements en francs
<i>gagec</i>	Total des engagements court terme en francs
<i>gagem</i>	Total des engagements moyen terme en francs
<i>kvunb</i>	Nombre de comptes à vue
<i>qsmoy</i>	Moyenne des soldes moyens sur 3 mois
<i>qcred</i>	Moyenne des mouvements créditeurs en milliers de francs
<i>boppn</i>	Nombre d'opérations à $M - 1$
<i>facan</i>	Montant facturé dans l'année en francs
<i>lgagt</i>	Engagement long terme
<i>vienb</i>	Nombre de produits contrats vie
<i>viemt</i>	Montant des produits contrats vie en francs
<i>uemnb</i>	Nombre de produits épargne monétaire
<i>uemmts</i>	Montant des produits d'épargne monétaire en francs
<i>xlgnb</i>	Nombre de produits d'épargne logement
<i>xlgmt</i>	Montant des produits d'épargne logement en francs
<i>ylvnb</i>	Nombre de comptes sur livret
<i>ylvmt</i>	Montant des comptes sur livret en francs
<i>nbelts</i>	Nombre de produits d'épargne long terme
<i>mtelts</i>	Montant des produits d'épargne long terme en francs
<i>nbcats</i>	Nombre de produits épargne à terme
<i>mtcats</i>	Montant des produits épargne à terme
<i>nbbeccs</i>	Nombre de produits bons et certificats
<i>mtbeccs</i>	Montant des produits bons et certificats en francs

rocnb	Nombre de paiements par carte bancaire à $M - 1$
ntcas	Nombre total de cartes
nptag	Nombre de cartes point argent
segv2s	Segmentation version 2
itavc	Total des avoirs sur tous les comptes
zocnb	Nombre d'opérations par cartes
havef	Total des avoirs épargne financière en francs
nbgd1s	Nombre de jours à débit à M
nbgd2s	Nombre de jours à débit à $M - 1$
nbgd3s	Nombre de jours à débit à $M - 2$
carvp	Possession de la carte VISA Premier, soit oui (1), soit non (0)

1. Transformez la variable catégorielle `sexe` en variable binaire **numérique** avec `homme=0`, `femme=1`, et . pour les valeurs manquantes.
2. Fusionnez les situations familiales (`famiq`) selon que la personne est seule (`seu`) ou en couple (`cou`), **et . pour les valeurs manquantes**.
3. Éliminez les observations correspondant
 - aux comptes professionnels (**toute observation pour laquelle** à la fois `nbcats` et `nbbecs` **valent simultanément 1**),
 - aux **observations** pour lesquels la variable `age` est manquante
 - aux observations de clients âgés de moins de 18 ans et de plus de 65 ans.
4. Calculez le nombre total de jours à débit des trois derniers mois au sein de la variable `nbgd` et éliminez les variables utilisées lors de la création.
5. **Identifiez les observations pour lesquelles plus de 10% des valeurs des variables continues sont manquantes et supprimez-les.** ces observations sont déjà éliminées à ce stade.
6. Considérez le nombre total de cartes `ntcas`. Y a-t-il des incohérences en lien avec les autres variables?
7. Que représentent les variables manquantes résiduelles de `zocnb`? *Indice : voir la question précédente.* Expliquez pourquoi il serait logique de remplacer ces valeurs manquantes par des valeurs numériques (laquelle). Effectuez la modification.
8. Produisez un histogramme de la variable `ancienneté` du compte (`relat`). Que remarquez-vous?
9. Produisez un nuage de point de `relat` et `age` et commentez. Supprimez les valeurs aberrantes (*indice : quel est le lien entre `relat` et `age`?*)
10. Y a-t-il des variables exactement collinéaires? Si oui, identifiez lesquelles et éliminez une du lot pour chaque ensemble.
11. Créez un tableau de fréquence des variables `famiq` et `pcspq`. Expliquez en une phrase les conséquences de conserver des modalités dont la fréquence est basse.

Vous devez remettre avec votre rapport et votre code la base de données créée à la suite des manipulations. Nommez cette dernière selon la convention **d1_matricule.sas7bdat** (NB : les noms de fichier SAS doivent n'inclure que des chiffres, des lettres et une barre de soulignement (`_`), mais doivent obligatoirement commencer par une lettre.)