

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

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
nbjd1s	Nombre de jours à débit à M
nbjd2s	Nombre de jours à débit à $M - 1$
nbjd3s	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. Notez qu'une valeur manquante pour des variables de type `Alphanum` est une chaîne de caractère vide, pas un point (`.`)
3. Éliminez les observations correspondant aux observations
 - pour lesquelles la variable `age` est manquante
 - 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 `nbjd` et éliminez les variables utilisées lors de la création.
5. Considérez le nombre total de cartes `ntcas`. Y a-t-il des incohérences en lien avec les autres variables?
6. 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.
7. Produisez un histogramme de la variable ancienneté du compte (`relat`). Que remarquez-vous?
8. Produisez un nuage de point de `relat` et `age` et commentez. Supprimez les valeurs aberrantes (indice : quel est le lien entre `relat` et `age`)?
9. ~~Y a-t-il des variables exactement collinéaires? Si oui, identifiez-lesquelles et éliminez-une du lot pour chaque ensemble.~~ **question optionnelle**
10. 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.)