Données manquantes

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Données manquantes

Plusieurs champs d'une base de donnée peuvent être manquants

- non-réponse
- valeurs erronées (erreur d'encodage)
- perte de suivi et censure
- plusieurs versions de formulaires (question optionnelles)

Pourquoi s'en préoccuper

La plupart des procédures ne gèrent que les cas complets (toute observation avec des valeurs manquantes est éliminée).

Les données manquantes réduisent l'information disponible.

Classification des données manquantes (MCAR)

Cas 1: Données manquantes de façon complètement aléatoire (*missing completely at random*)

La probabilité que la valeur soit manquante ne dépend ni de la valeur, ni de celles des autres variables.

Exemple: questionnaire trop long, la personne ne répond pas à tout (sans lien avec les questions posées).

Hypothèse souvent irréaliste en pratique.

Classification des données manquantes (MAR)

Cas 2: données manquantes de façon aléatoire (*missing at random*): la probabilité que la valeur soit manquante ne dépend pas de la valeur *une fois qu'on a contrôlé pour les autres variables*.

Exemple: les hommes sont plus susceptibles dans l'ensemble de divulguer leur âge que les femmes.

Classification des données manquantes (MNAR)

Cas 3: données manquantes de façon non-aléatoire (*missing not at random*): la probabilité que la mesure soit manquante dépend de la valeur elle-même, pas déterminable avec d'autres variables

Exemple: une personne transgenre ne répond pas à la question genre (si seulement deux choix, homme/femme) et aucune autre question ne se rattache au genre ou à l'identité sexuelle.

Comment déterminer le type de données manquantes

Une personne ne divulgue pas son salaire? Données manquante de manière aléatoire ou non aléatoire?

Hypothèse pas testable, dépend du contexte et des variables auxiliaires disponibles.

Valeur logique

Souvent, les données manquantes ont une valeur logique:

- un client qui n'a pas de carte de crédit a un solde de 0!

D'où l'importance des validations d'usage et du nettoyage préliminaire de la base de données.

Méthodes d'imputation: cas complet

Retirer les observations avec données manquantes pour conserver les cas complets.

- Valide uniquement pour MCAR, sinon estimations biaisées.
- On perd de la précision en utilisant moins d'observations.

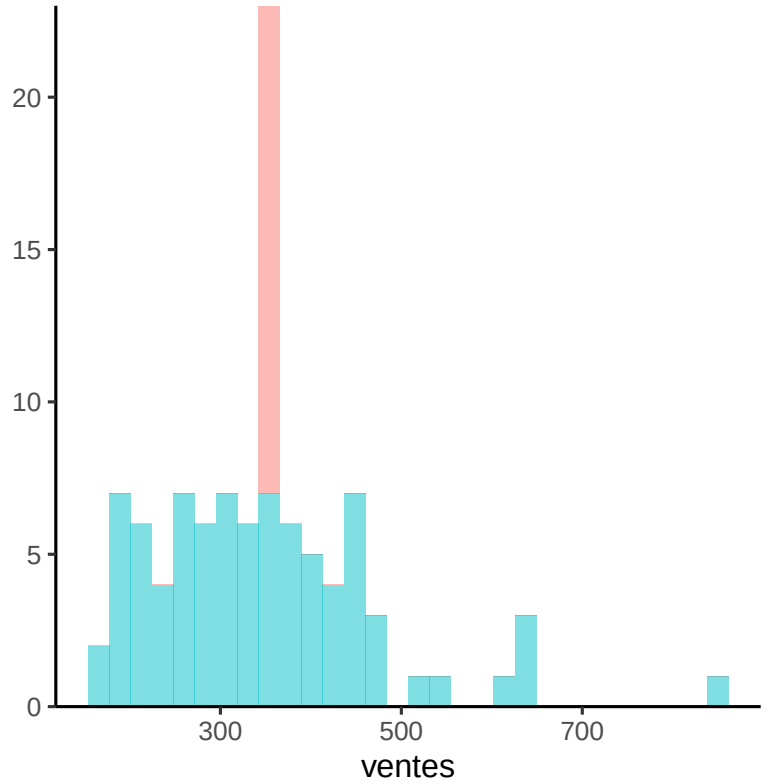
Méthodes d'imputation: imputation simple

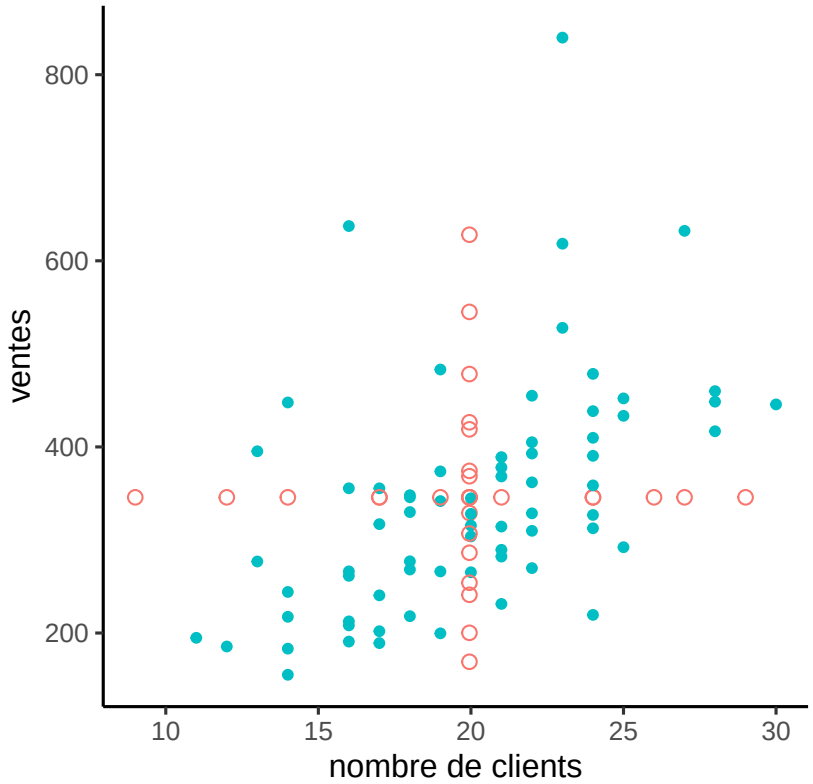
Remplacer les valeurs manquantes par un seul nombre. Par exemple, la moyenne (variables continues) ou le mode (variables catégorielles).

Approche pas recommandée (pourquoi?)

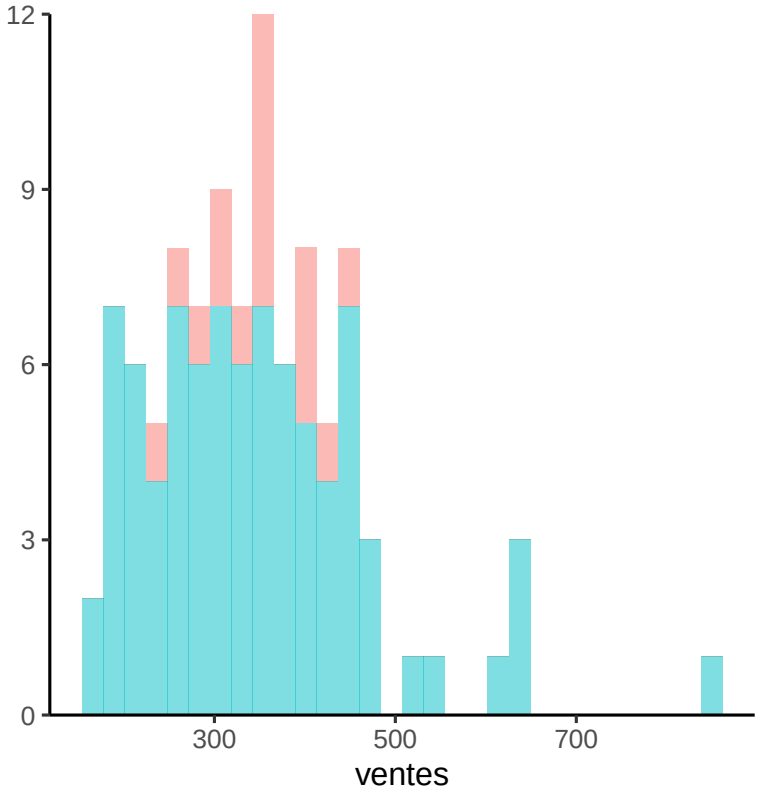
Imputation par la moyenne

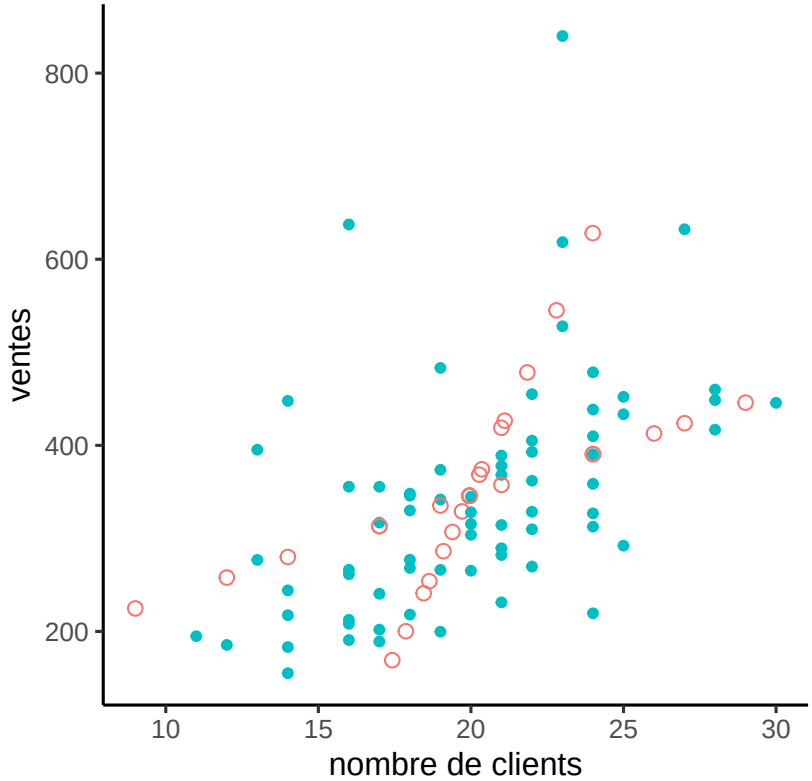
Dilution de la relation (corrélation) entre variables explicatives.
Réduction de la variabilité





Sous-estimation de l'incertitude.





Modèles prédictifs

Si le modèle est mal spécifié, les imputations seront erronées.

Problèmes de l'imputation simple

On ne tient pas compte du fait que des valeurs ont été remplacées (on fait comme si c'était de vraies observations).

On sous-évalue la variabilité dans les données

- les écarts-type des estimations sont trop petits.

Imputation multiple

Valide pour données MAR et MCAR.

1. procéder à une imputation aléatoire pour obtenir un échantillon complet
2. ajuster le modèle d'intérêt avec cet échantillon
3. répéter ce processus plusieurs fois
4. combiner les résultats obtenus.

Algorithme MICE

Algorithme d'imputation multiple par équation de chaine (MICE).

Faut-il toujours imputer?

Il faut utiliser son jugement.

Une observation imputée ne remplacera jamais une vraie observation.

- Si la proportion d'observations manquantes est petite (moins de 5%), on pourrait faire une analyse avec les cas complets (et valider au besoin en utilisant l'imputation multiple).
- Si la proportion de valeurs manquantes est 30% et que cette proportion baisse à 3% lorsque vous éliminez quelques variables peu importantes pour votre étude, alors procédez à leur élimination.

Inspection des valeurs manquantes

Il est donc nécessaire d'examiner la configuration des valeurs manquantes avant de faire quoi que ce soit.

```
data(manquantes, package = 'hecmulti')
summary(manquantes)
# Pourcentage de valeurs manquantes
apply(manquantes, 2, function(x){mean(is.na(x))})
# Voir les configurations de valeurs manquantes
md.pattern(missing1) # graphique diapo suivante
```

Table 1: Nombre et pourcentage de valeurs manquantes par variable.

	x1	x2	x3	x4	x5	x6	y
nombre	192	49	0	184	0	0	0
pourcentage	38.4	9.8	0	36.8	0	0	0

Configuration des valeurs manquantes

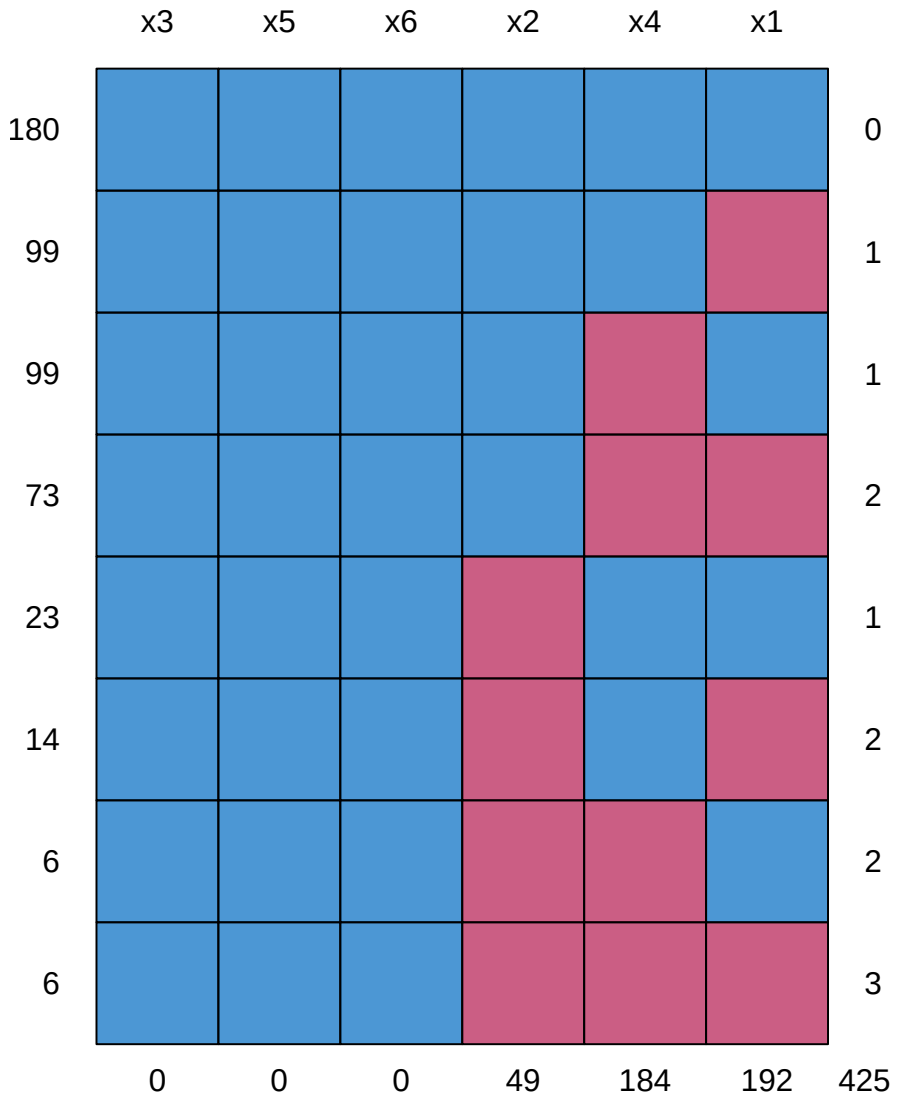


Figure 1: Configurations des valeurs manquantes pour la base de données `manquantes`.