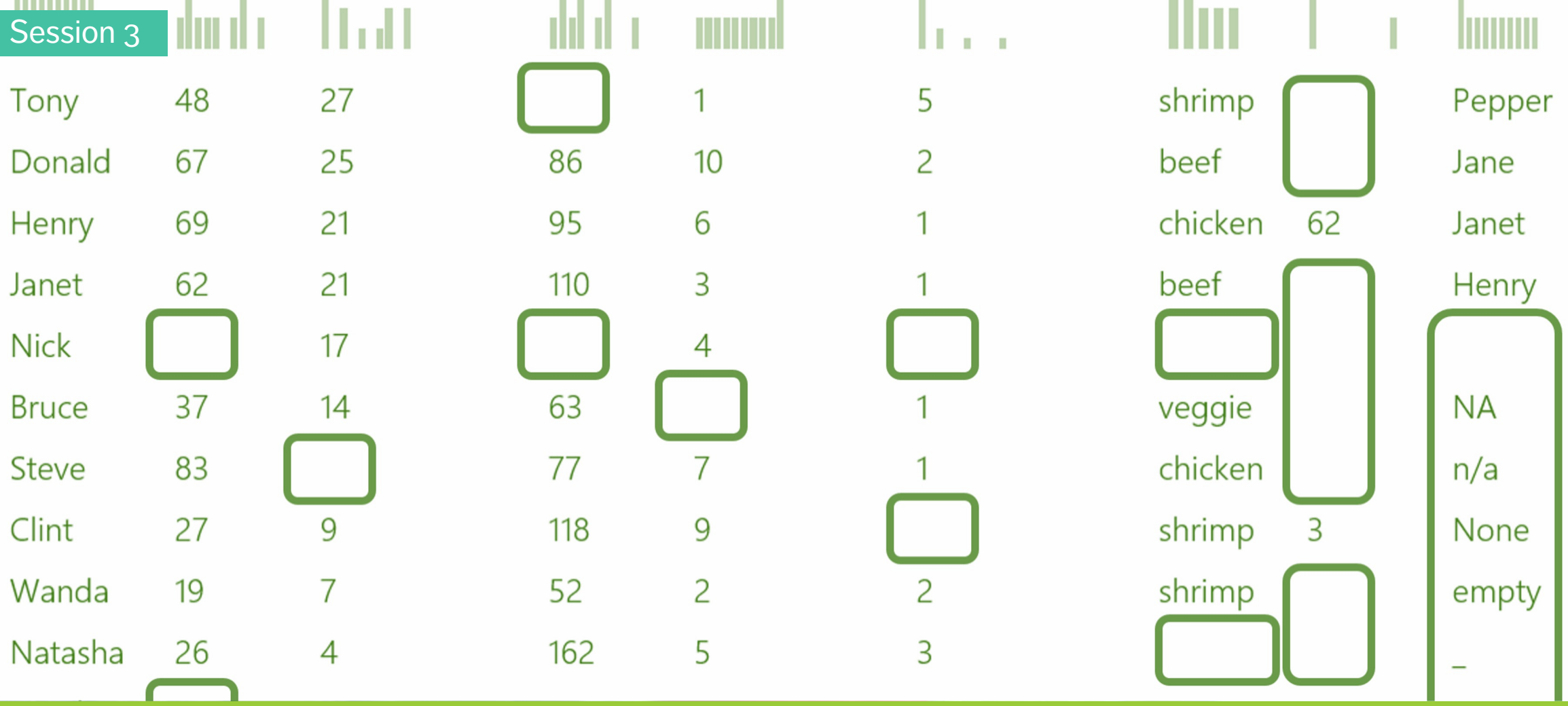


Tony	48	27		1	5	shrimp	Pepper
Donald	67	25	86	10	2	beef	Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef	Henry
Nick		17		4			
Bruce	37	14	63		1	veggie	NA
Steve	83		77	7	1	chicken	n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp	empty
Natasha	26	4	162	5	3			_

8. Les valeurs manquantes

Les types d'observations manquantes

Les champs vierges existent en 4 versions :

- **non-réponse**
  une observation était attendue mais aucune n'a été saisie
- **problème de saisie des données**
  une observation a été enregistrée mais n'a pas été saisie dans l'ensemble de données
- **entrée invalide**
  une observation a été enregistrée mais a été considérée comme non valide et a été supprimée
- **blanc attendu**
  un champ a été laissé vide, mais c'est normal

Les types d'observations manquantes

Trop de valeurs manquantes des trois premiers types peut indiquer des **problèmes dans le processus de collecte des données**.

Trop de valeurs manquantes du quatrième type peut indiquer une **mauvaise conception du questionnaire**.

Trouver les valeurs manquantes peut vous aider à traiter d'autres problèmes de science des données.

L’imputation

Les méthodes d'analyse ne s'accommodent pas facilement des observations manquantes :

- **écarter** l'observation manquante
  - non recommandé, à moins que les données manquantes soient MCAH
  - acceptable dans certaines situations (e.g., un petit nombre de valeurs manquantes dans un ensemble de données massives)
- trouver une **valeur de remplacement (imputation)**
  - principal inconvénient : nous ne savons jamais quelle aurait été la vraie valeur
  - mais cela demeure souvent la meilleure option disponible

Les mécanisme de valeurs manquantes

Manquant complètement au hasard (MCAH)

- l'absence de l'élément est indépendante de sa valeur ou des variables auxiliaires
- **exemple :** une surtension électrique supprime aléatoirement une observation dans l'ensemble de données

Manquant au hasard (MAH)

- l'absence d'un article n'est pas complètement aléatoire ; elle peut être expliquée par des variables auxiliaires avec des informations complètes.
- **exemple :** si les femmes sont moins susceptibles de vous dire leur âge que les hommes pour des raisons sociétales, mais pas à cause des valeurs d'âge elles-mêmes

Les mécanisme de valeurs manquantes

Ne manquant pas au hasard (NMAH)

- la raison de la non-réponse est liée à la valeur de l'item (également appelée **non-réponse non-ignorable**)
- **exemple :** si les consommateurs de drogues illicites sont moins susceptibles d'admettre leur consommation de drogues que les abstinents...

En général, le mécanisme manquant **ne peut pas être déterminé** avec certitude ; on devra émettre des hypothèses (l'expertise du domaine aide).

Les méthodes d'imputation

- suppression par liste
- imputation par la moyenne ou par la valeur la plus fréquente
- imputation par la régression ou la corrélation
- imputation par la régression stochastique
- report de la dernière observation
- report en arrière de l'observation suivante
- imputation par les k voisins les plus proches
- imputation multiple
- etc.

Les méthodes d'imputation

Suppression par liste : supprimer les unités avec au 1+ valeurs manquantes

- **hypothèse :** MCAH
- **Contre :** peut introduire un biais (si non MCAH), réduction de la taille de l'échantillon, augmentation de l'erreur standard

Imputation moyenne/la plus fréquente : remplacer les valeurs manquantes par la valeur moyenne/la plus fréquente.

- **hypothèse :** MCAH
- **contre :** distorsions de la distribution (pic à la moyenne) et des relations entre les variables

Les méthodes d'imputation

Imputation par régression/corrélation : remplacer les valeurs manquantes par des valeurs ajustées en se basant sur des variables avec des informations complètes.

- **hypothèse :** MAH
- **contre :** réduction artificielle de la variabilité, surestimation de la corrélation

Imputation par régression stochastique : imputation par la régression/la corrélation avec ajout d'un terme d'erreur aléatoire

- **hypothèse :** MAH
- **contre :** risque accru d'erreur de type I (faux positifs) en raison de la faible erreur-type

Les méthodes d'imputation

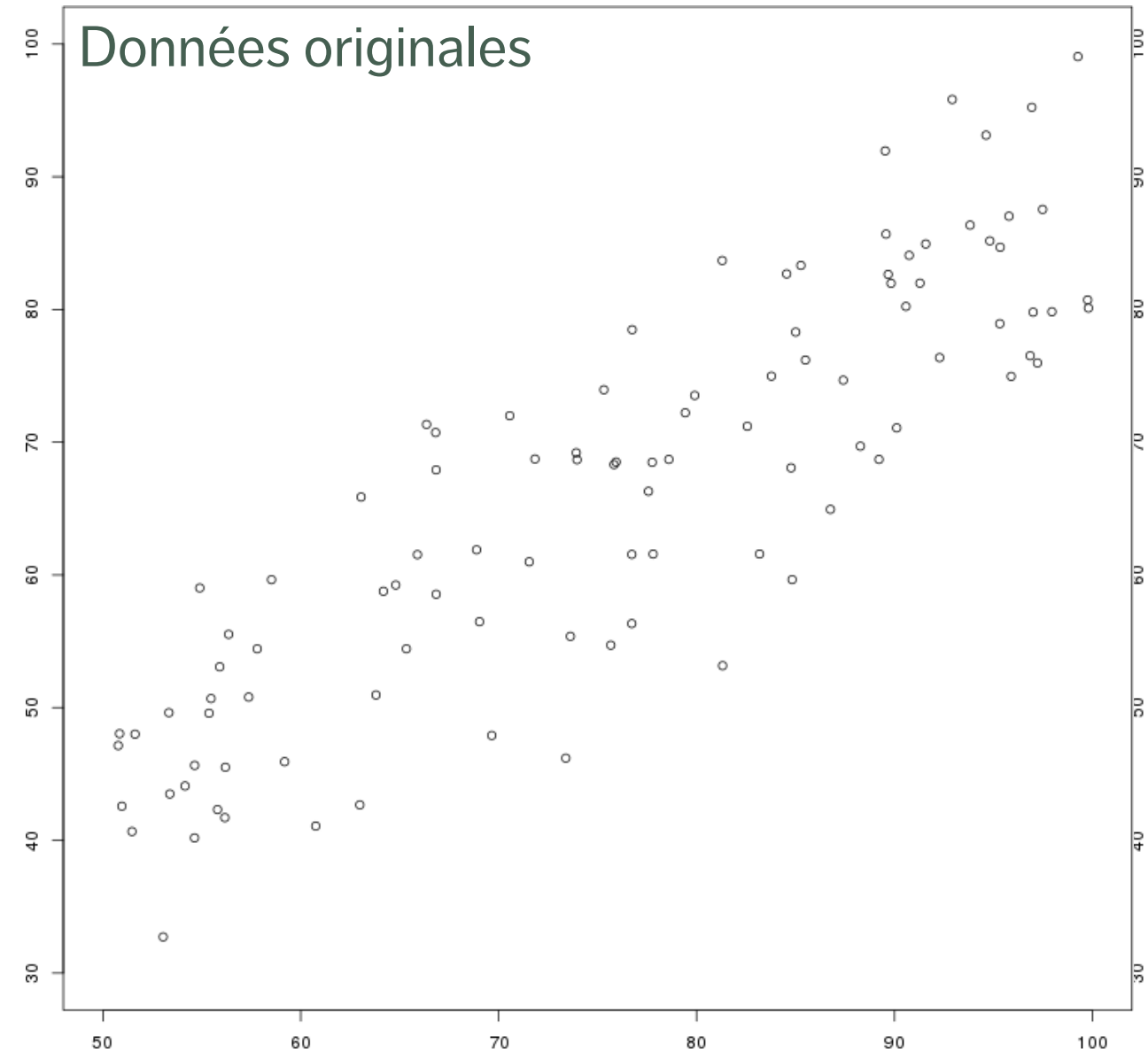
Dernière observation reportée : remplacer les valeurs manquantes par les dernières valeurs précédentes (dans une étude longitudinale)

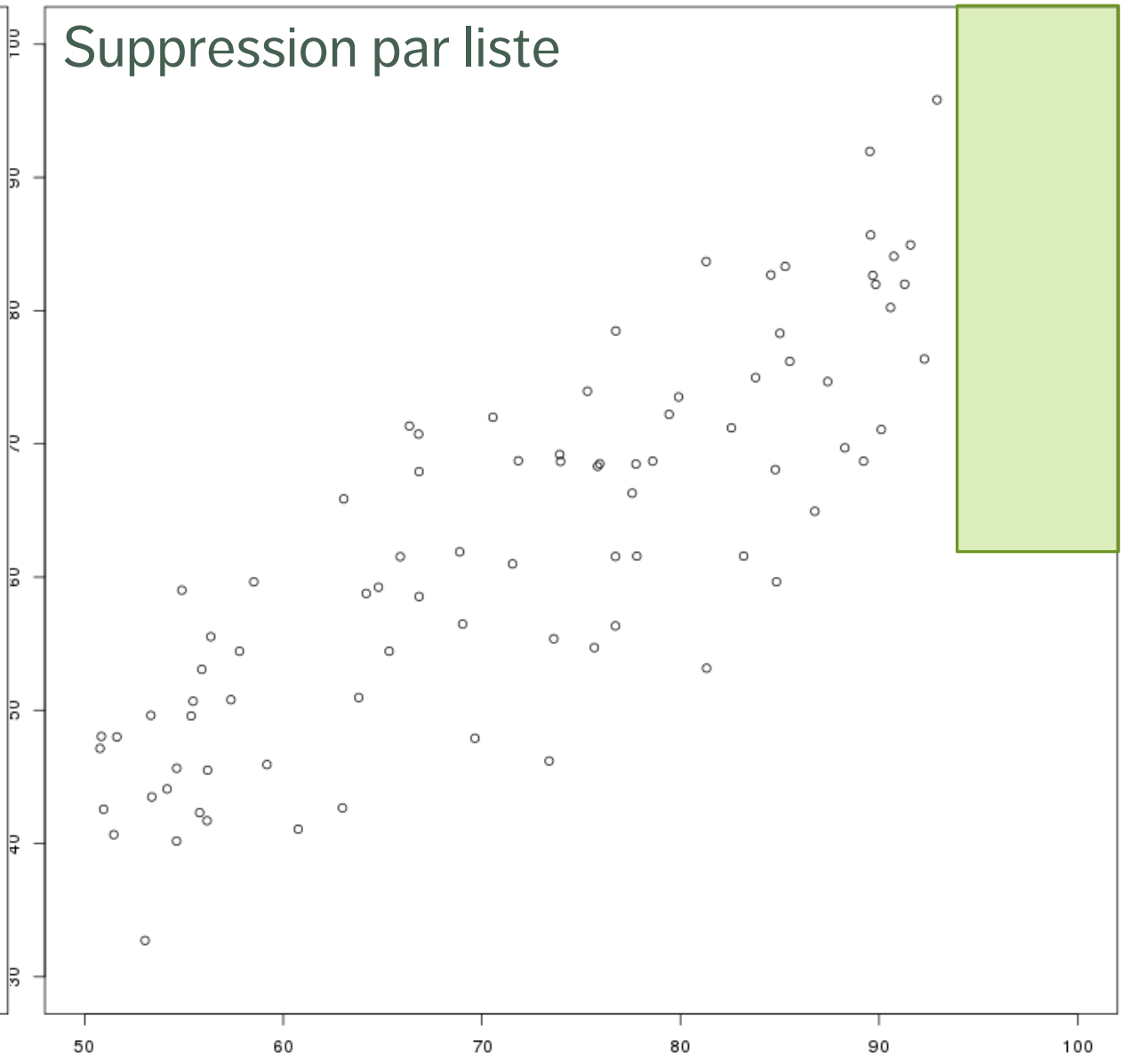
- **hypothèse :** MCAH, les valeurs ne varient pas beaucoup au fil du temps
- **contre :** peut être trop "généreux", selon la nature de l'étude

imputation par le plus proche voisin (kNN) : remplacer l'entrée manquante par la moyenne du groupe des k cas complets les plus similaires

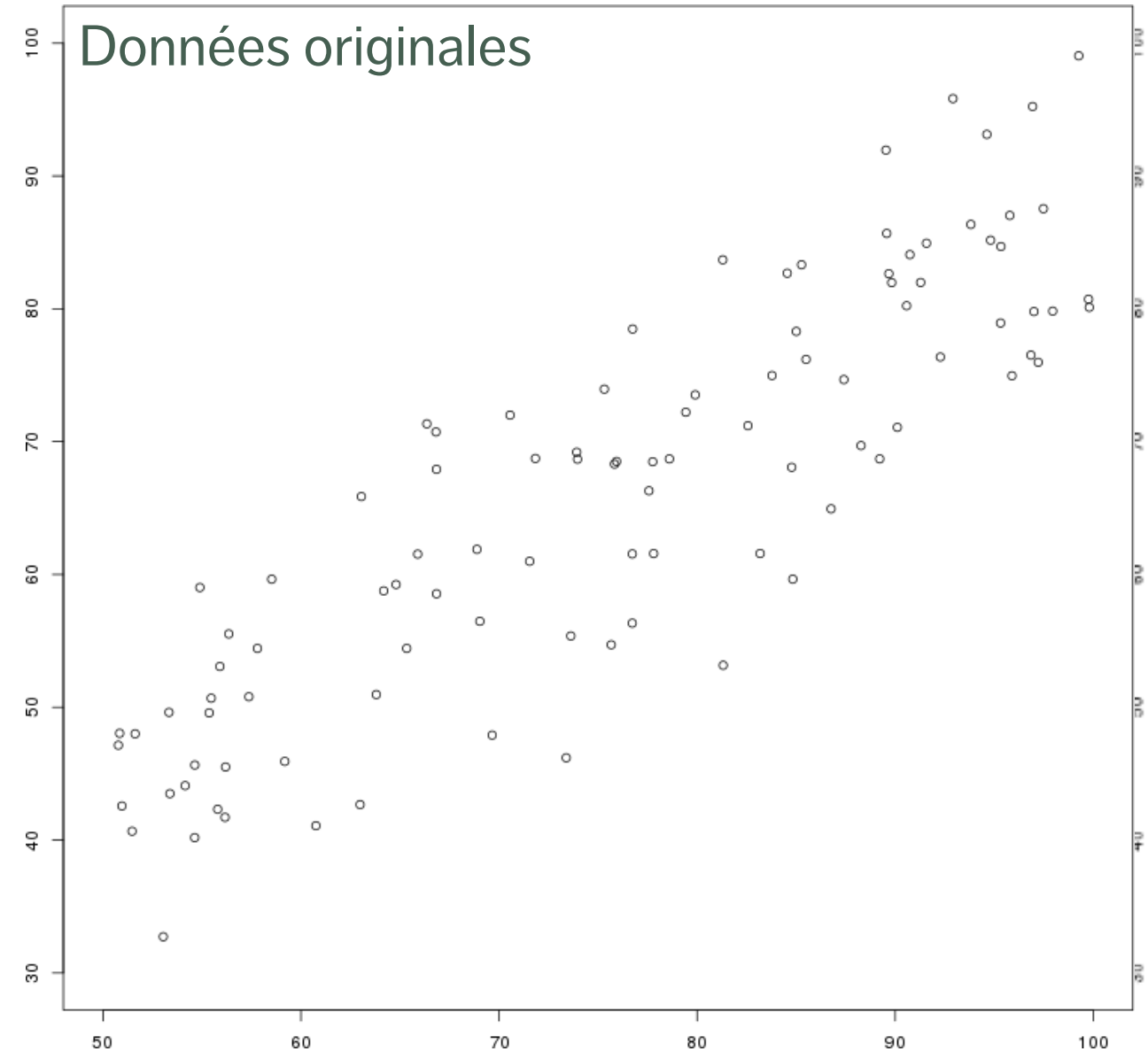
- **hypothèse :** MAH
- **contre :** difficile de choisir une valeur appropriée de k; distorsion possible dans la structure des données

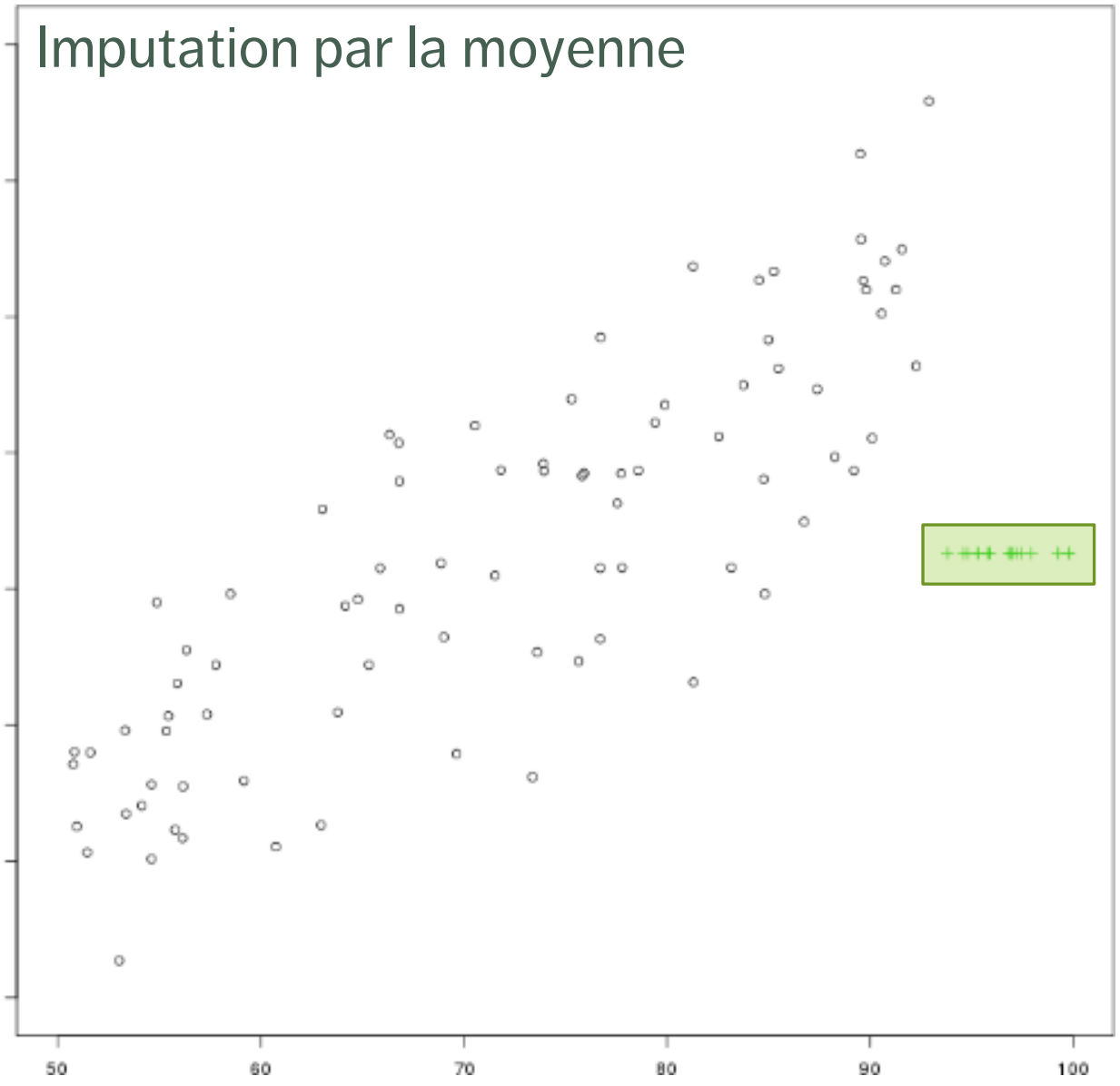
Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur



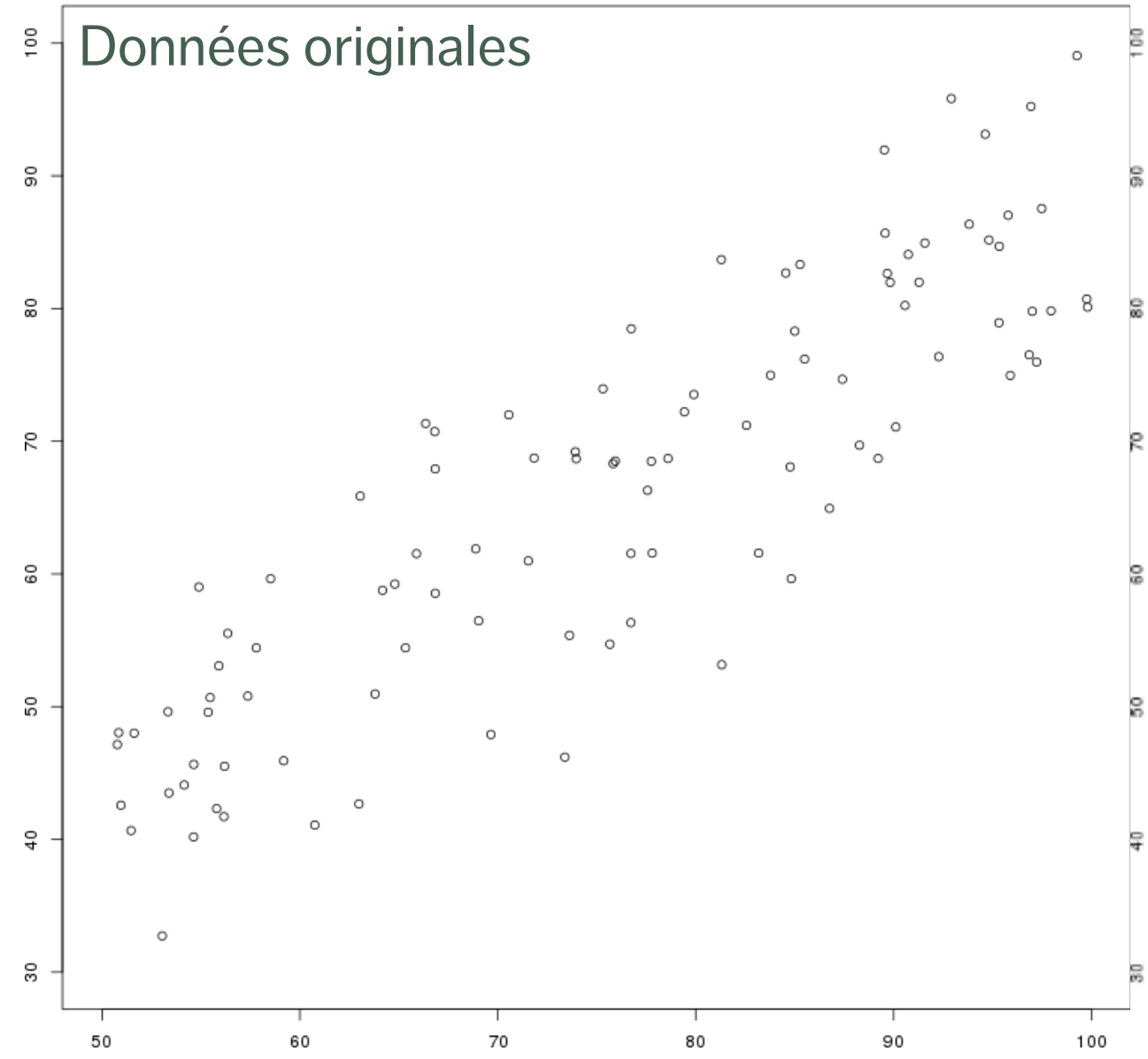


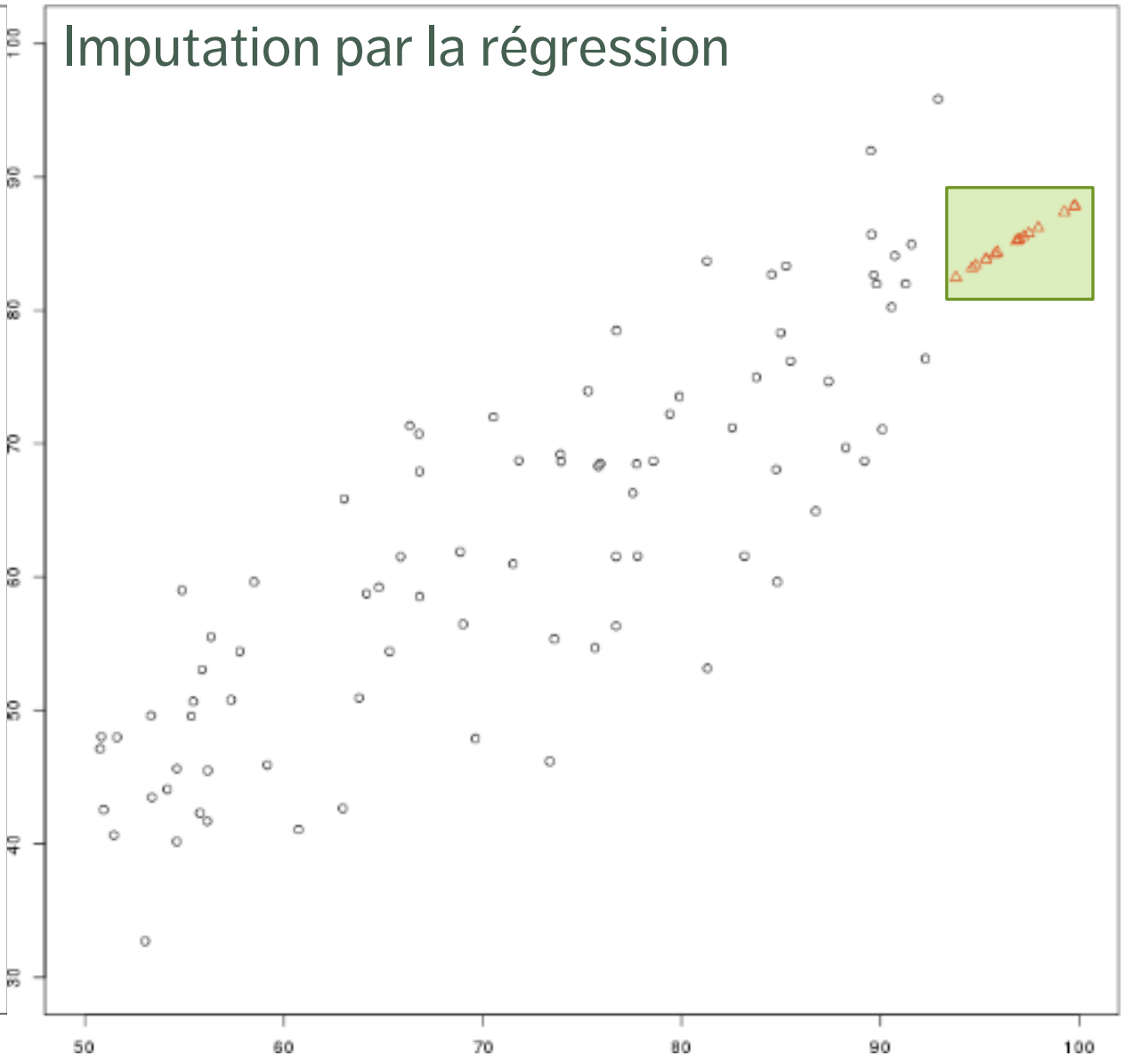
Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur



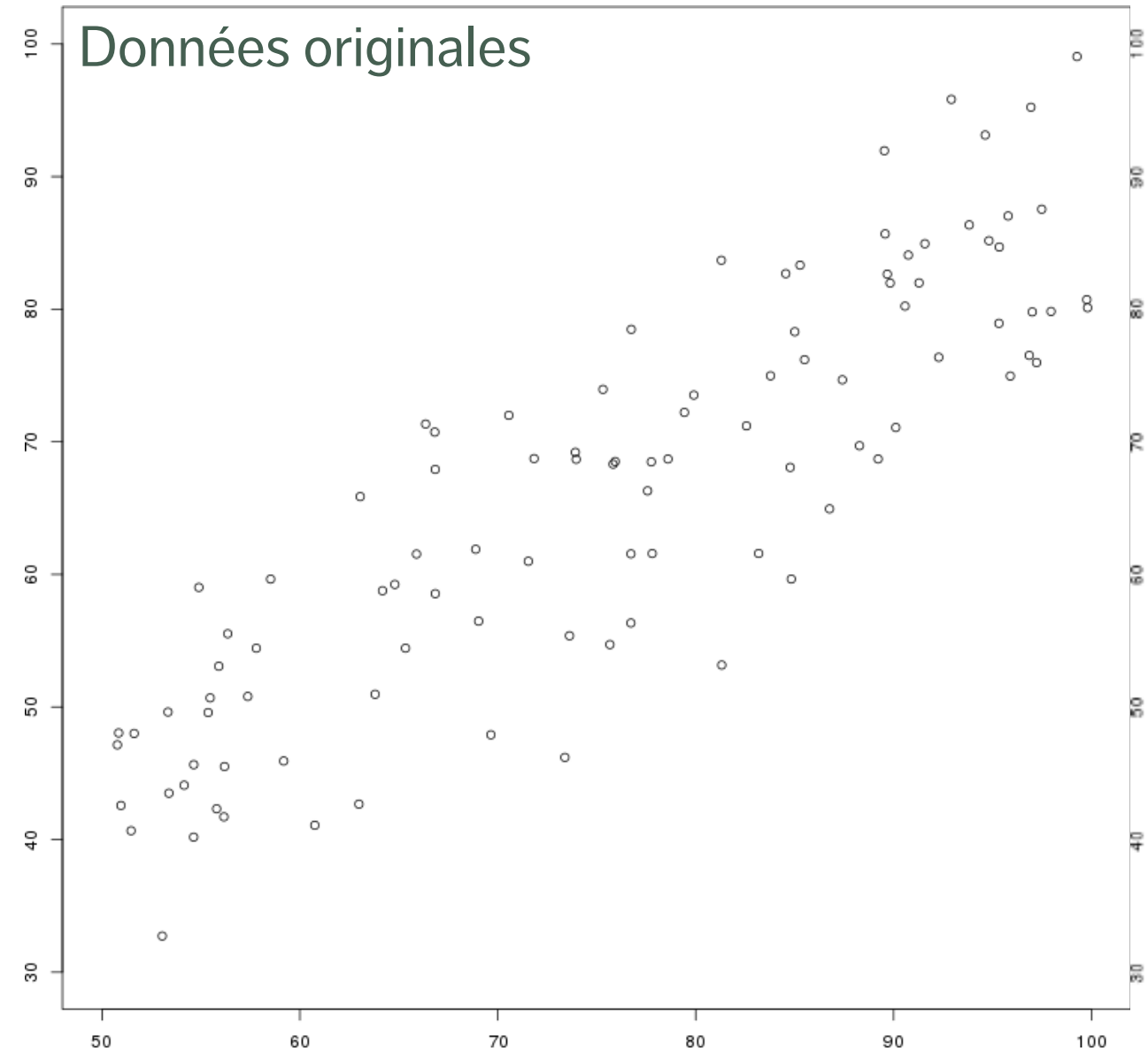


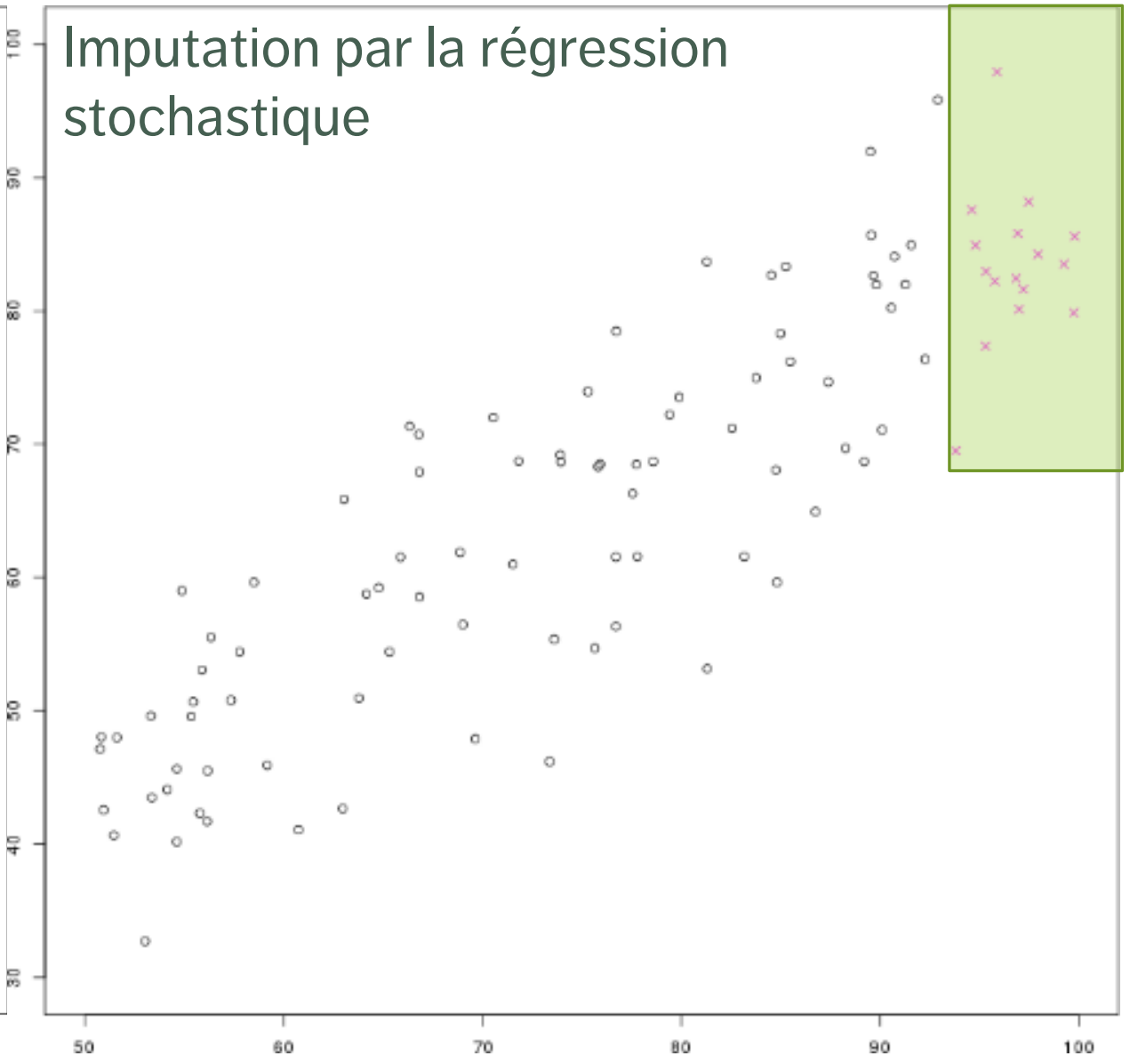
Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur





Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur





L’imputation multiple

Les imputations augmentent le “bruit” (l’incertitude) dans les données.

Dans le cas de l'**imputation multiple**, l'effet de ce bruit peut être mesuré en consolidant les résultats de l'analyse à partir de plusieurs répétitions de la procédure d’imputation dur l’ensembles de données manquantes

Étapes :

1. l'imputation répétée crée m versions de l'ensemble de données
2. chacun de ces m ensembles de données est analysé, ce qui donne m résultats
3. les m résultats sont regroupés en un seul résultat pour lequel la moyenne, la variance et les intervalles de confiance sont connus

L’imputation multiple

Avantages

- **flexible** ; peut être utilisé dans diverses situations (MCAH, MAH, voire NMAH dans certains cas)
- tient compte de l'**incertitude** des valeurs imputées
- assez facile à mettre en œuvre

Inconvénients

- m peut devoir être assez **grande** lorsqu'il y a plusieurs valeurs manquantes dans de nombreuses caractéristiques, ce qui ralentit les analyses
- si le résultat de l'analyse n'est pas une valeur unique mais un objet mathématique compliqué, cette approche a peu de chances d'être utile

À retenir

Les valeurs manquantes **ne peuvent pas être simplement ignorées**.

Le mécanisme manquant **ne peut généralement pas être déterminé** avec certitude.

Les méthodes d'imputation fonctionnent mieux lorsque les valeurs sont **MCAH** ou **MAH**; les méthodes d'imputation ont tendance à produire des estimations biaisées.

Dans l'imputation simple, les données imputées sont traitées comme les données réelles ; l'**imputation multiple** peut contribuer à réduire le bruit.

L'imputation stochastique est-elle la meilleure solution ? Dans notre exemple, oui - mais ... faîtes attention au **théorème du "No-Free Lunch"** !

Lectures suggérées

Les valeurs manquantes

Data Understanding, Data Analysis, Data Science
Data Preparation

Missing Values

- Missing Value Mechanisms
- Imputation Methods
- Multiple Imputation

Exercices

Les valeurs manquantes

1. Recréez les exemples de Imputation Methods.
2. Recréez le processus d'imputation des valeurs manquantes (nettoyage des données) utilisé dans Example: Algae Bloom.
3. Effectuez l'imputation kNN sur l'ensemble de données des grades avec différentes valeurs de k.
4. Effectuez une imputation multiple sur l'ensemble de données grades en utilisant la régression stochastique afin d'estimer la pente et l'ordonnée de la ligne de meilleur ajustement.