La préparation des données

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

7. La qualité et le traitement des données

Le bordel total

"Les données sont désordonnées, vous savez."
"Même après avoir été nettoyées ?"
"*Surtout* après avoir été nettoyées."

Le nettoyage, le **traitement** et la **manipulation** des données sont des aspects essentiels des projets de science des données.

Les analystes peuvent consacrer **jusqu'à 80 % de** leur temps à la **préparation des données**.

La manipulation et le “tidyverse”

Les données **“tidy”** ont une structure spécifique :

- chaque variable se retrouve dans une seule colonne
- chaque observation se retrouve dans une seule rangée
- chaque type d'unité d'observation dans un seul tableau

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

VS.

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

Fonctionnalité de traitement

Les fonctions de traitement des données doivent permettre à l'analyste de :

- **extraire** un sous-ensemble de **variables** de la trame de données
- **extraire** un sous-ensemble d'**observations** de la trame de données
- **trier** les données selon toute combinaison de variables dans un ordre croissant/décroissant
- **créer de nouvelles variables** à partir de variables existantes
- **créer des tableaux croisés dynamiques**, par groupes d'observation
- **jouer** avec les **banques de données** (jointures, etc.)
- etc.

Le nettoyage des données

Il y a deux approches **philosophiques** de nettoyage/validation des données :

- méthodique
- narrative

L'approche **méthodique** consiste à passer en revue une **liste de contrôle** des problèmes potentiels et à signaler ceux qui s'appliquent aux données.

L'approche **narrative** consiste à **explorer** l'ensemble de données et à essayer de repérer les schémas improbables et irréguliers.

Le nettoyage des données

Méthodique (syntaxe)

- Pour : la liste de contrôle est **indépendante du contexte** ; les pipelines sont **faciles à implémenter** ; les erreurs courantes/observations invalides sont **facilement identifiées**
- Contre : peut s'avérer **chronophage** ; impossible d'identifier de nouveaux types d'erreurs

Narration (sémantique)

- Pour : le processus peut simultanément permettre de **comprendre les données** ; les faux départs sont (au maximum) aussi coûteux que le passage à l'approche méthodique
- Contre : peut manquer d'importantes sources d'erreurs et d'observations invalides pour les données comportant un **nombre élevé de caractéristiques** ; la connaissance du domaine peut biaiser le processus en négligeant les zones inintéressantes de l'ensemble de données

La solidité des données

L'ensemble de données idéal aura le moins de problèmes possible par rapport à ...

- **validité :** type de données, plage, réponse obligatoire, unicité, valeur, expressions régulières
- **exhaustivité :** observations manquantes
- **exactitude et précision :** liées aux erreurs de mesure et de saisie des données ; diagrammes de cibles (exactitude = biais, précision = erreur standard)
- **cohérence :** observations contradictoires
- **uniformité :** les unités sont-elles utilisées de manière uniforme ?

La vérification des problèmes liés à la qualité des données dès le départ peut vous éviter des maux de tête plus tard dans l'analyse.

La solidité des données

exact et
précis

précis, mais
pas exact

exact, mais
pas précis

ni exact,
ni précis

Les sources d'erreurs communes

Lorsque vous traitez des ensembles de données **hérités** ou **combinés** (c'est-à-dire des ensembles de données sur lesquels vous n'avez pas contrôle de la collecte et du traitement initial) :

- données manquantes avec un code
- 'NA'/'blank' avec un code
- erreur de saisie de données
- erreur de codage
- erreur de mesure
- entrées dupliquées
- accumulation ("heaping")

[Bill Watterson, *Calvin and Hobbes*, 15 sept. 1990].

La détection d'entrées non valides

Les entrées potentiellement invalides peuvent être détectées à l'aide de :

- **statistiques descriptives univariées**
  compte, étendue, score-z, moyenne, médiane, écart-type, contrôle logique
- **statistiques descriptives multivariées**
  tableaux croisés, contrôle logique
- **visualisation des données**
  nuage de points, histogramme, etc.

La détection d’entrées non valides

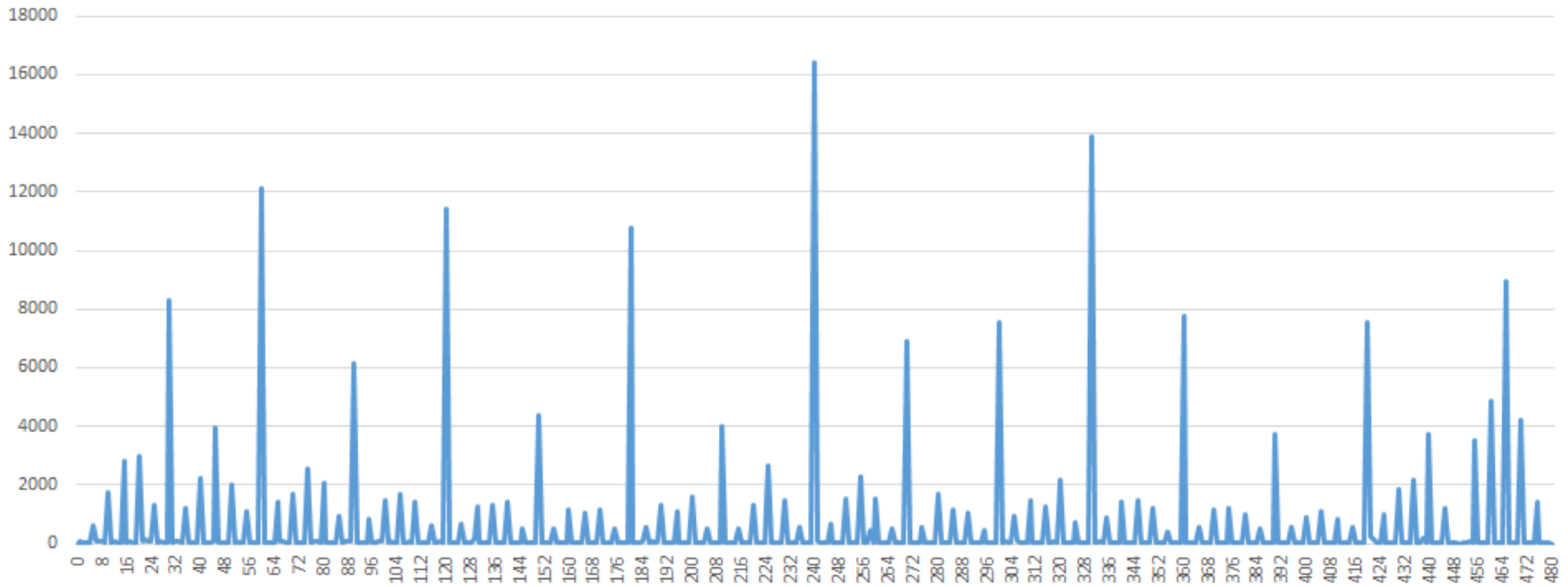
Les tests univariés ne montrent pas toujours **tout ce qui se passe**.

Cette étape pourrait permettre d'identifier les valeurs aberrantes potentielles.

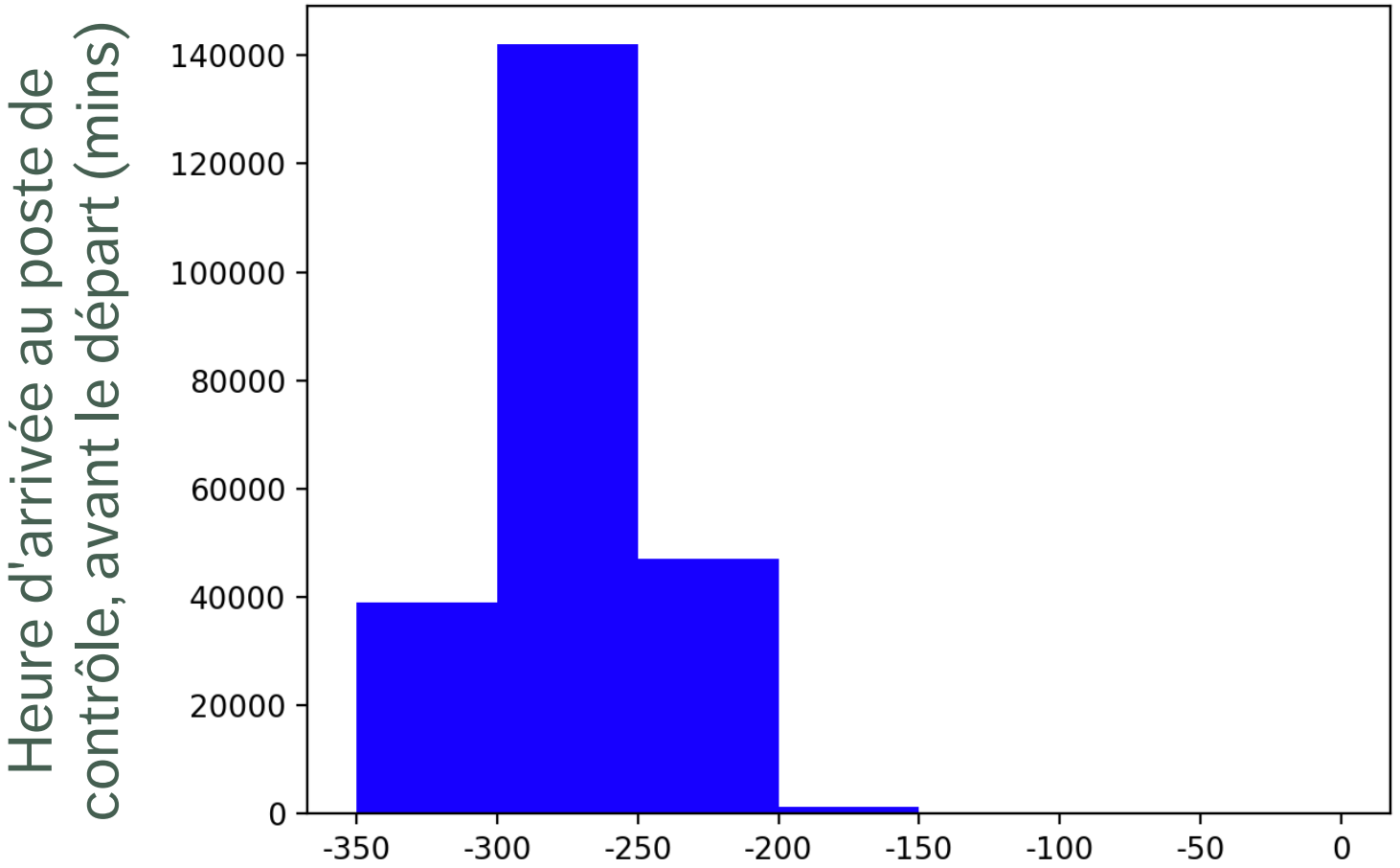
Défaut de détection des entrées non valides ≠ toutes les entrées sont valides.

Un petit nombre d'entrées non valides devrait être recodées comme étant "manquantes".

La détection d'entrées non valides



La détection d'entrées non valides



La détection d'entrées non valides

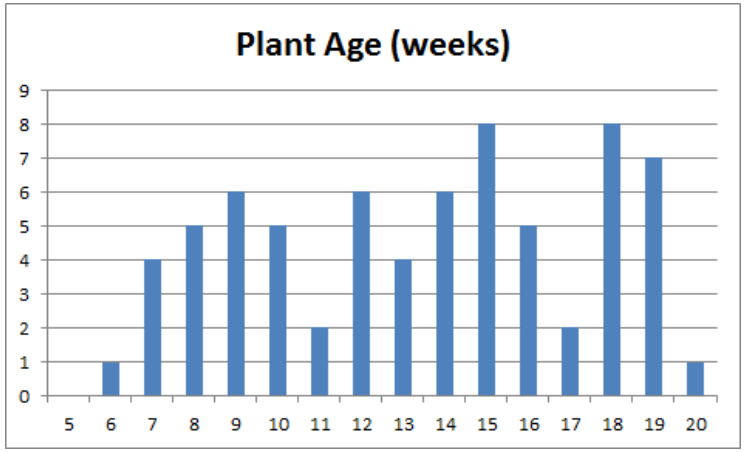
Sex	
Male	19
Female	17
(blank)	2
Total	**38**

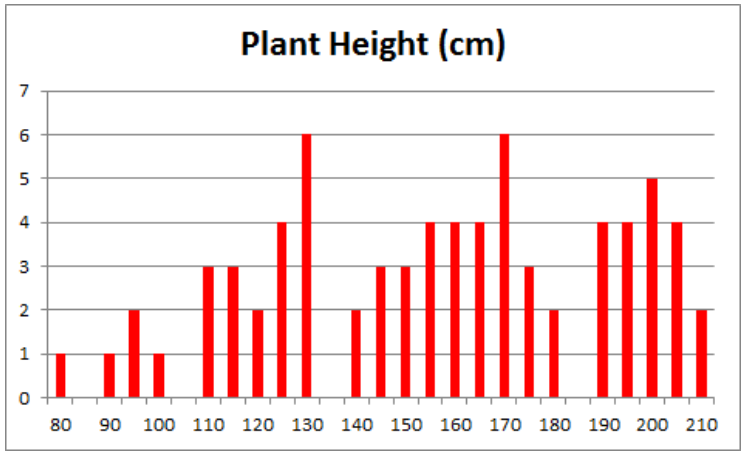
Pregnant	
Yes	7
No	27
99	1
(blank)	3
Total	**38**

vs.

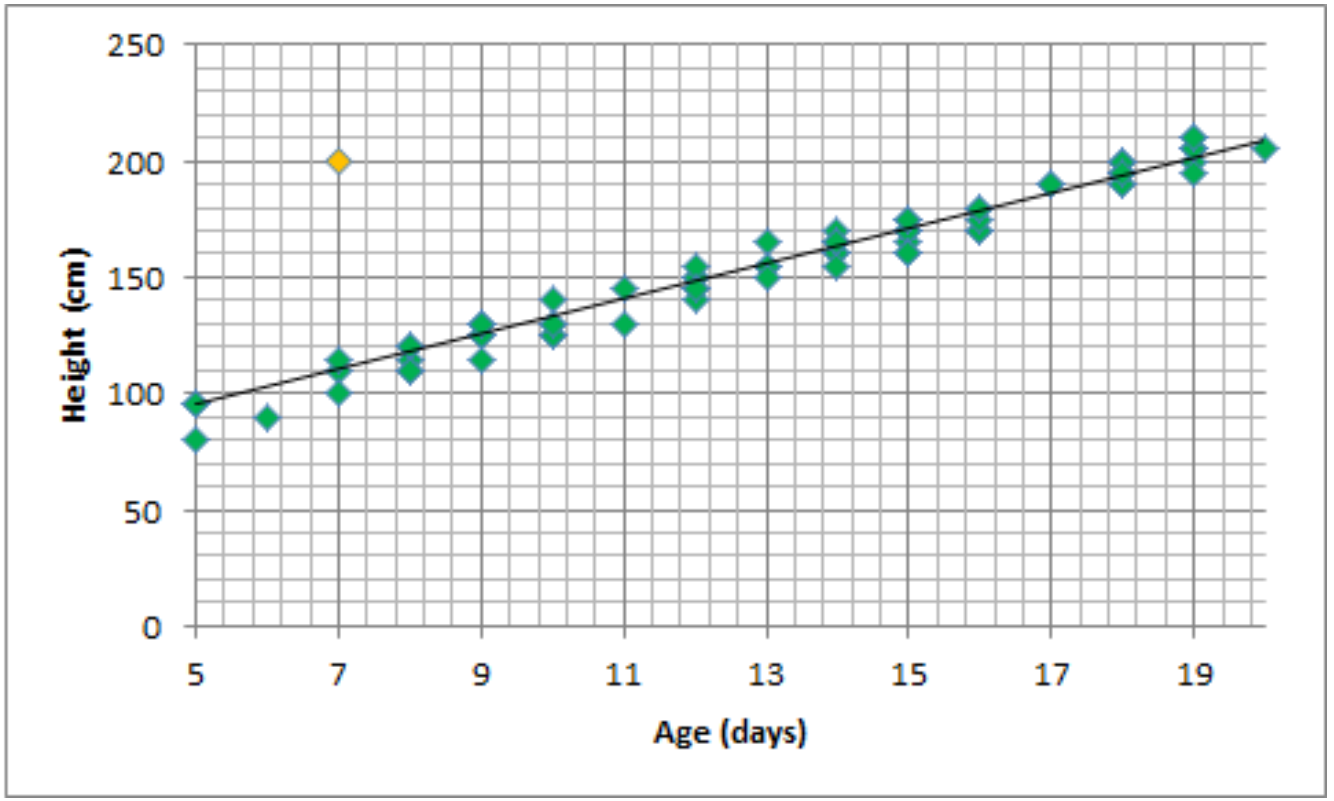
	Pregnant				Total
	Yes	No	99	(blank)	
Male	1	17	1	0	**19**
Female	6	9	0	2	**17**
(blank)	0	1	0	1	**2**
Total	**7**	**27**	**1**	**3**	**38**

La détection d'entrées non valides



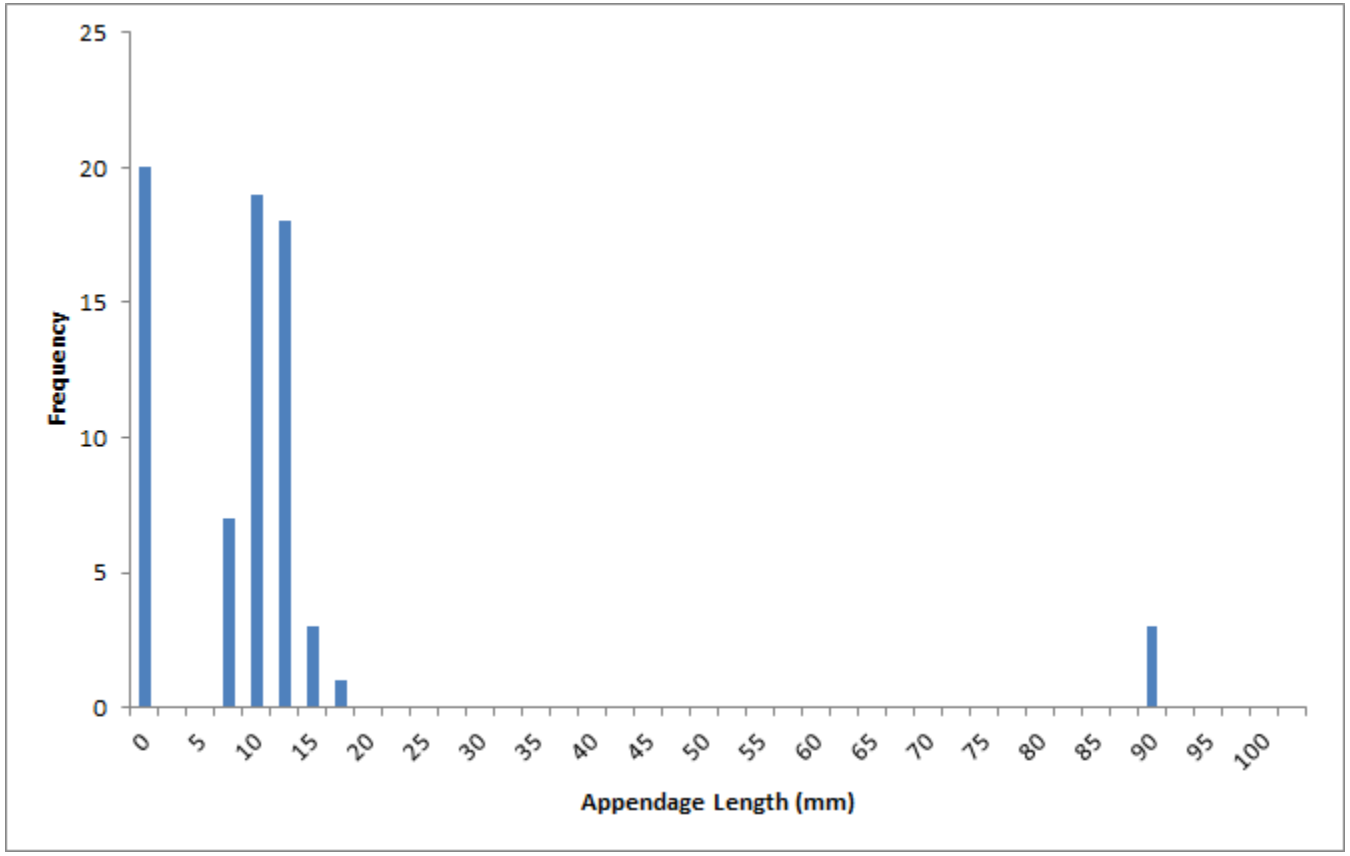


vs.



La détection d'entrées non valides

Appendage length (mm)	
Mean	10.35
Standard Deviation	16.98
Kurtosis	16.78
Skewness	4.07
Minimum	0
First Quartile	0
Median	8.77
Third Quartile	10.58
Maximum	88
Range	88
Interquartile Range	10.58
Mode	0
Count	71



Lectures conseillées

La qualité et le traitement des données

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

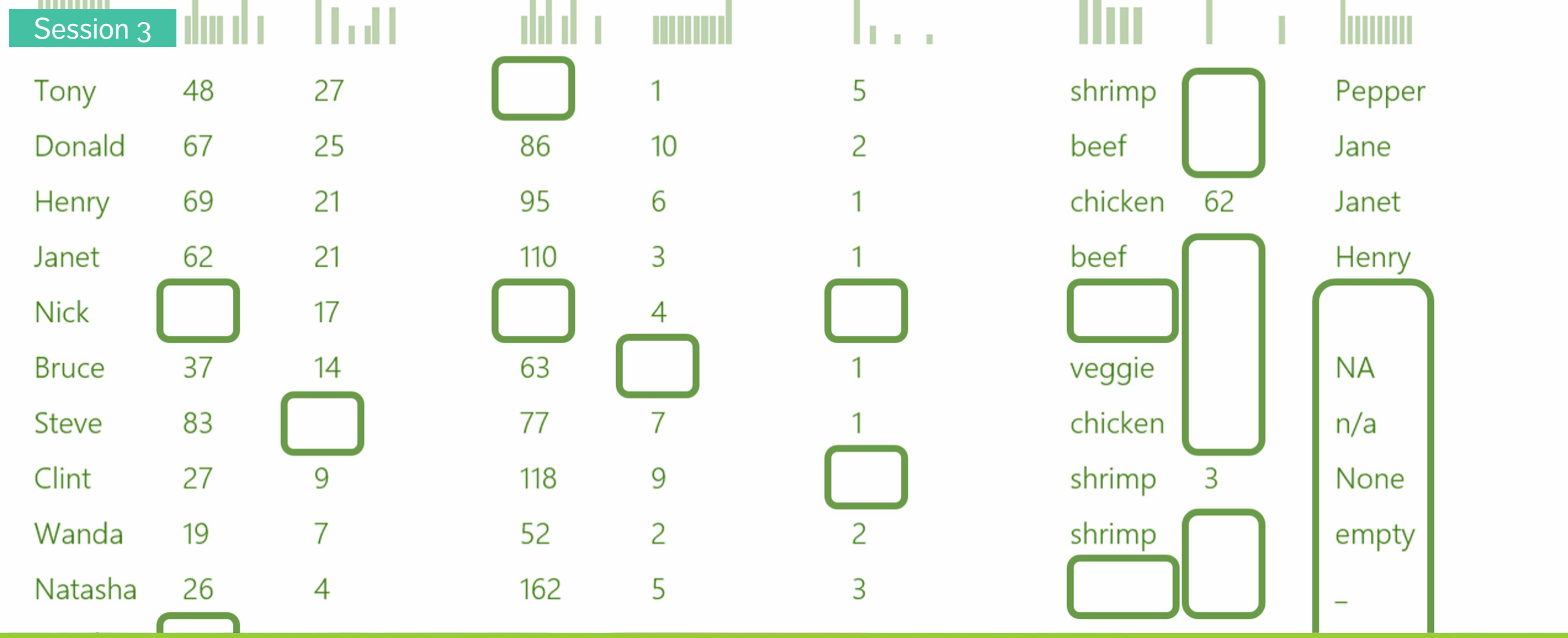
15. Data Preparation
    15.1 Introduction
    15.2 General Principles
    - Approaches to Data Cleaning
    - Pros and Cons
    - Tools and Methods

    15.3 Data Quality
    - Common Error Sources
    - Detecting Invalid Entries

Exercices

La qualité et le traitement des données

1. Recréez les exemples du Tidyverse.
2. Transformez le fichier cities.txt en ensemble de données “tidy”.
3. L'ensemble de données trouvé dans le fichier cities.txt semble-t-il être de bonne qualité (est-il “sain” ? comporte-t-il des entrées invalides ?)
4. Créez une liste d'éléments qui pourraient être utilisés dans une liste de contrôle de nettoyage méthodique des données. Utilisez des données que vous avez rencontrées dans le passé comme source d'inspiration (données numériques, catégorielles, textuelles).



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

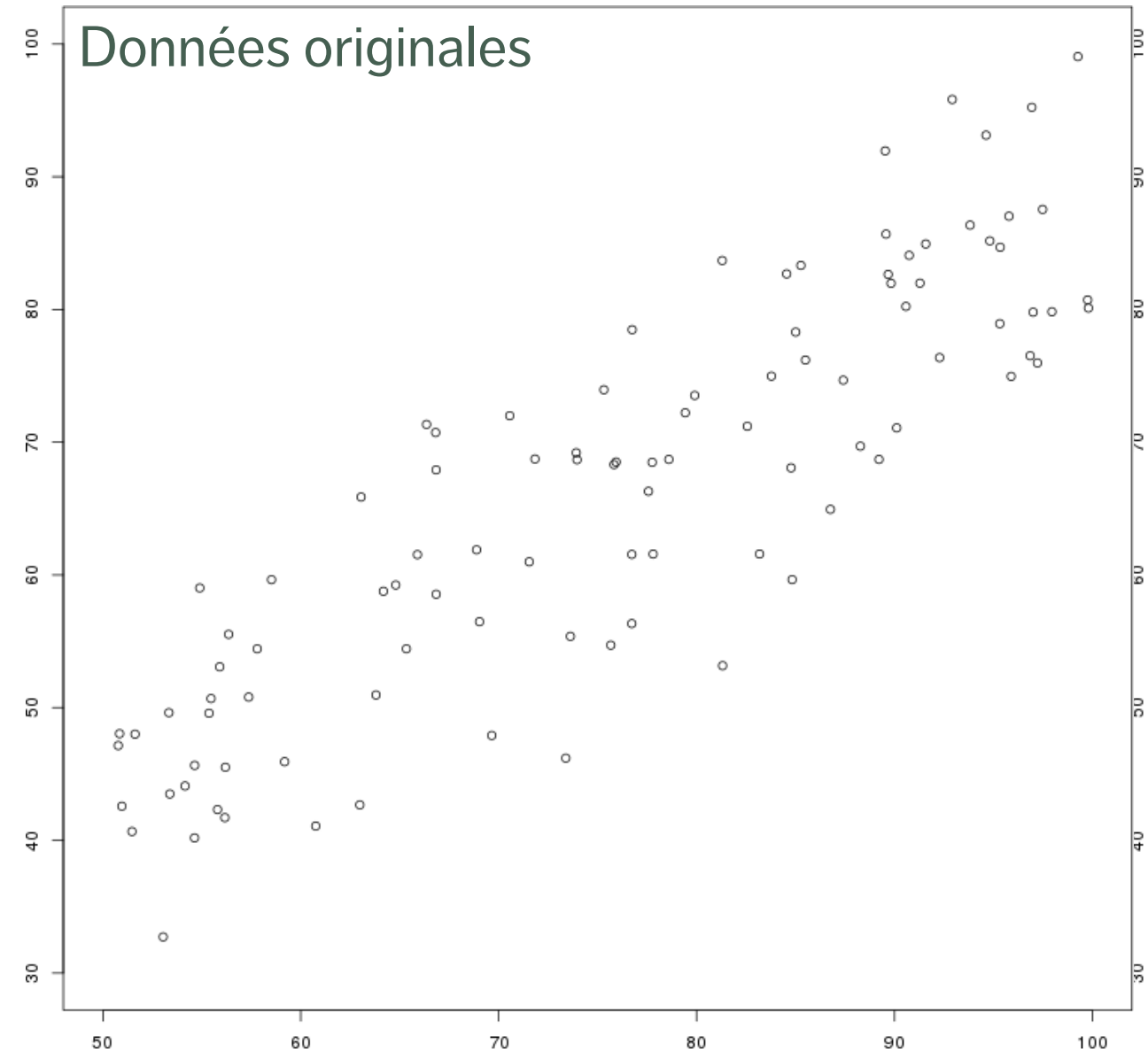
Dernière observation reportée : remplacer les valeurs manquantes par les dernières valeurs précédentes (dans une étude longitudinale)

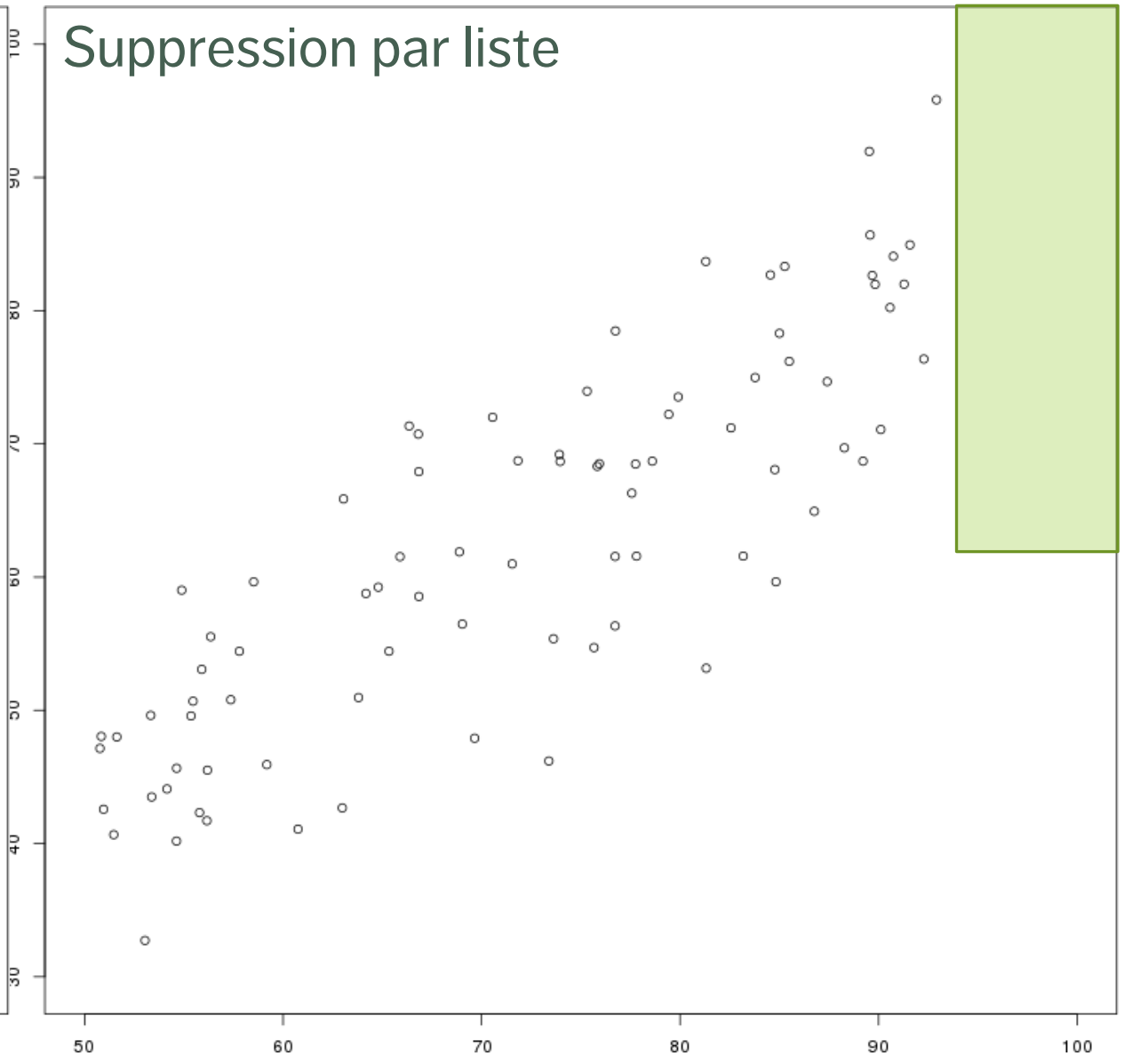
- **hypothèse :** MCAH, les valeurs ne varient pas beaucoup au fil du temps
- **contre :** peut être trop "généreux", selon la nature de l'étude

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

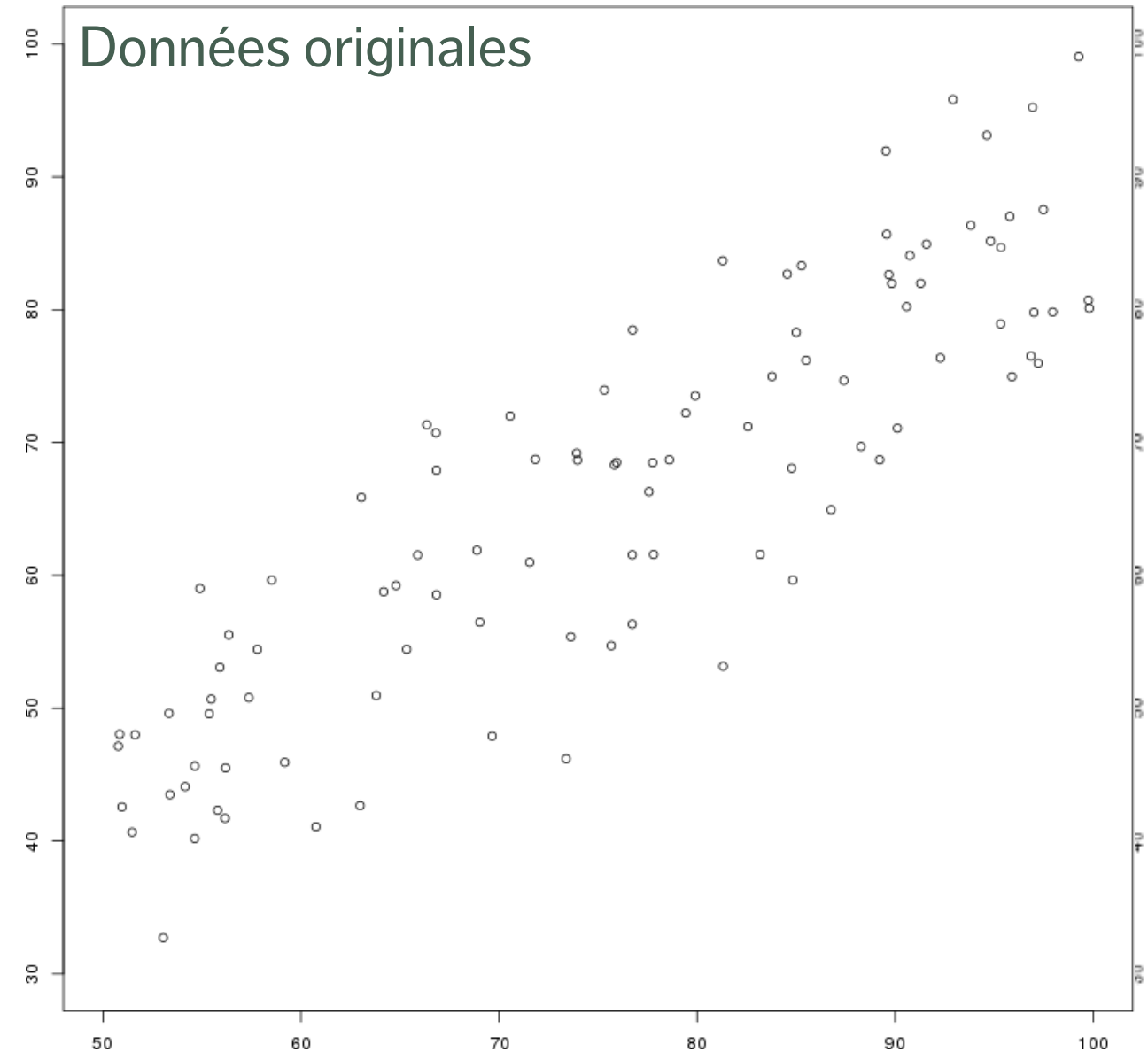
- **hypothèse :** MAH
- **contre :** difficile de choisir une valeur appropriée de k; distorsion possible dans la structure des données

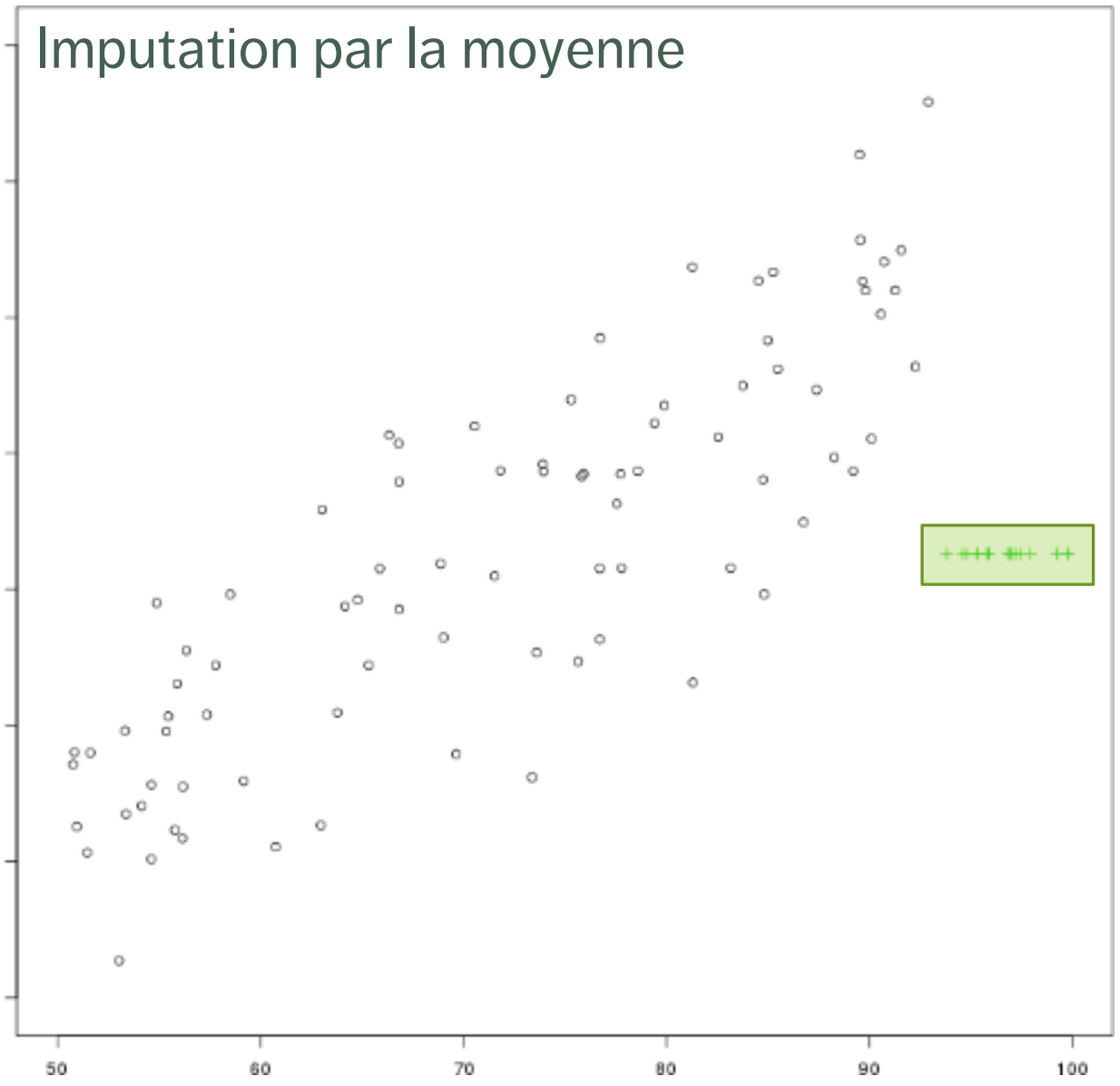
Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur



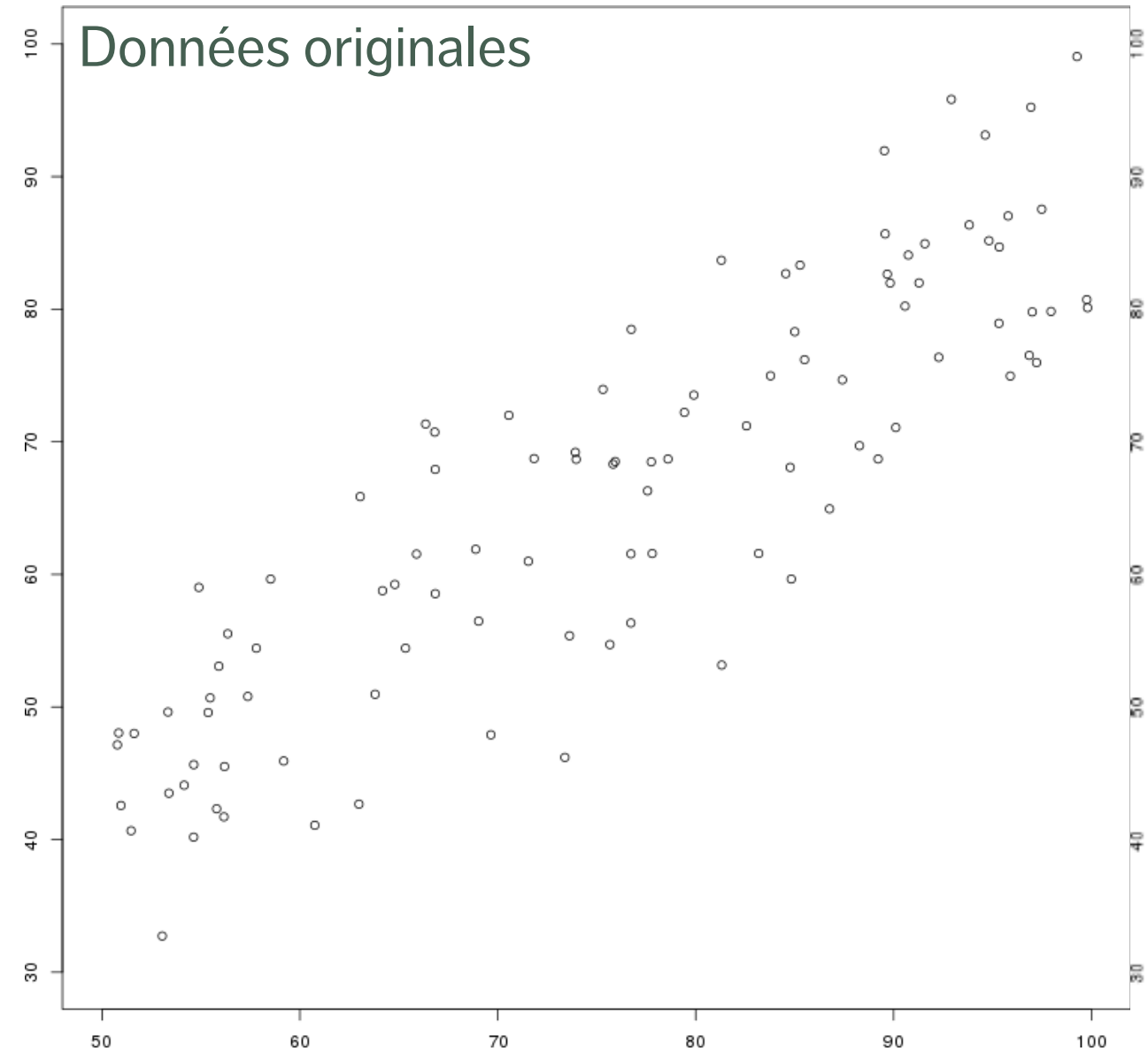


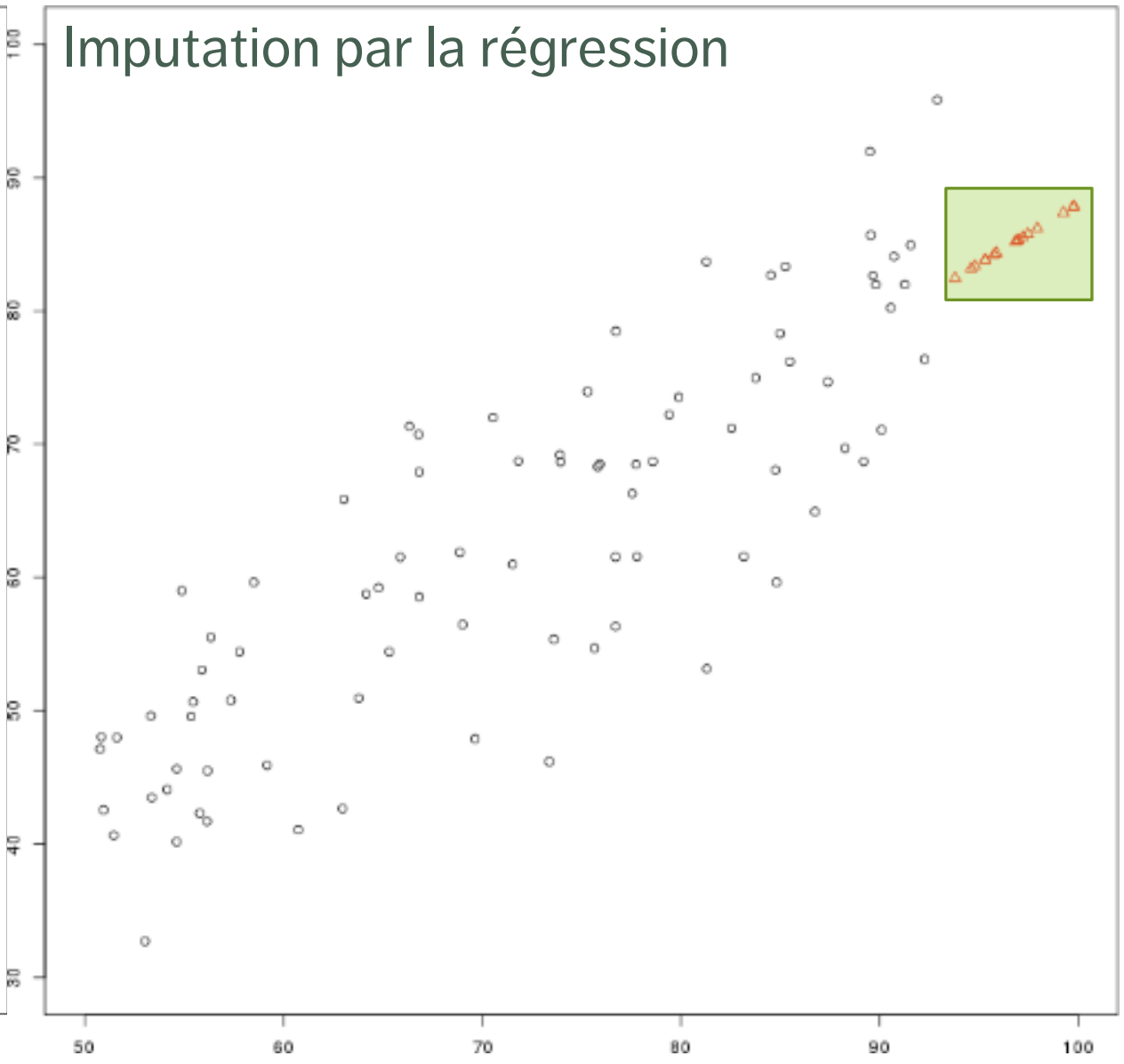
Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur



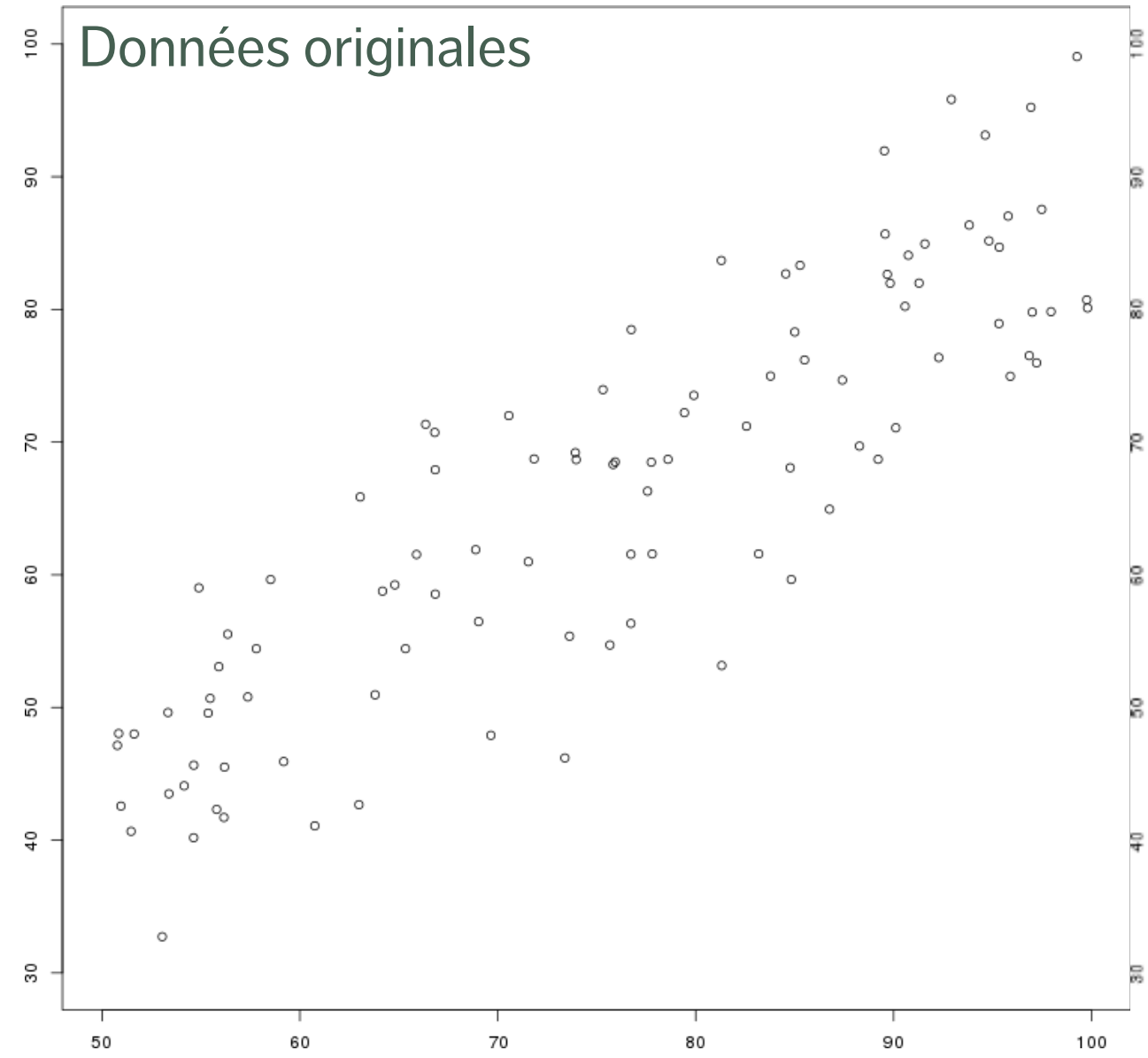


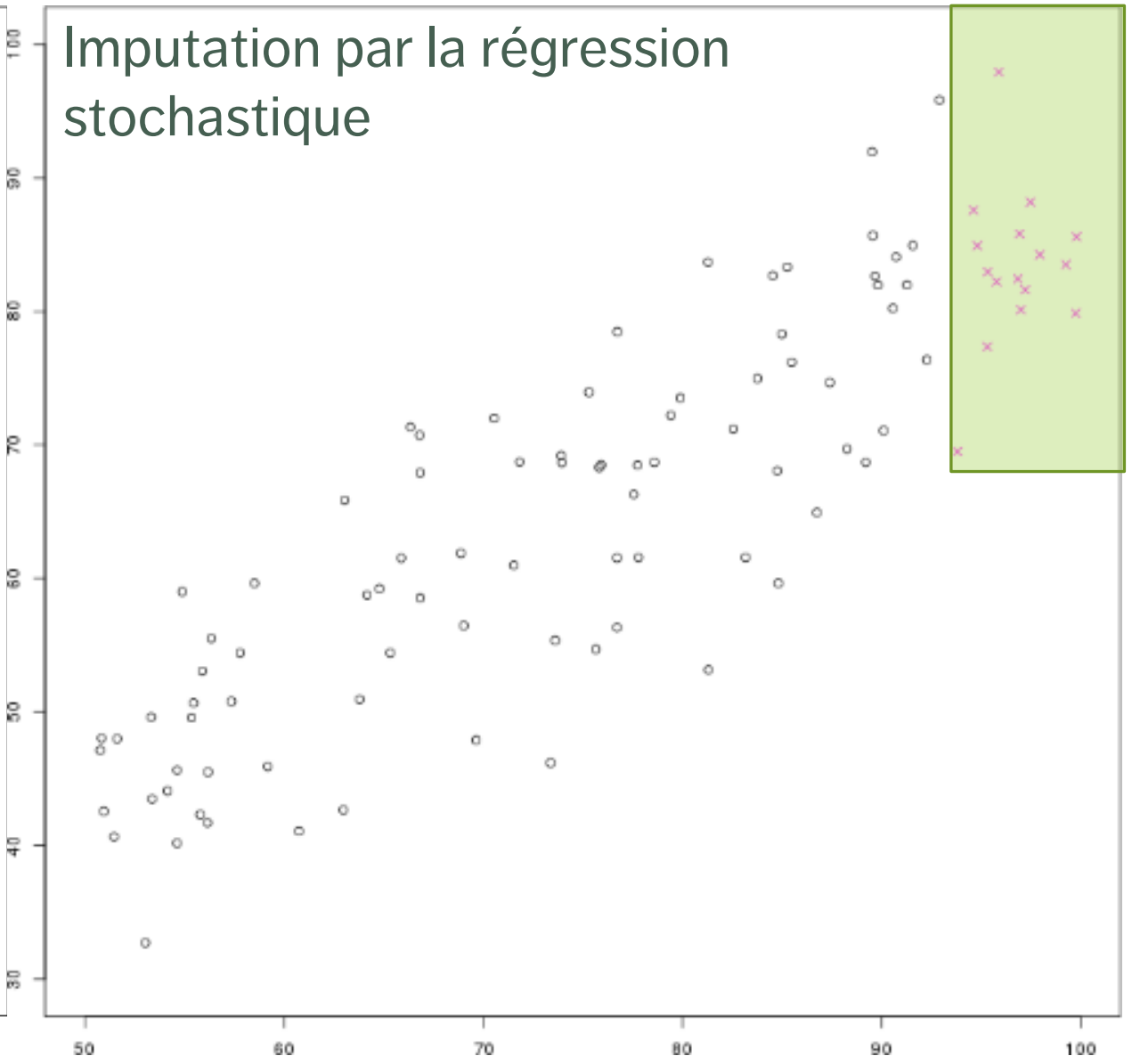
Données artificielles : les valeurs de tous les points pour lesquels $x > 92$ ont été effacé par erreur





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 conseillées

Les valeurs manquantes

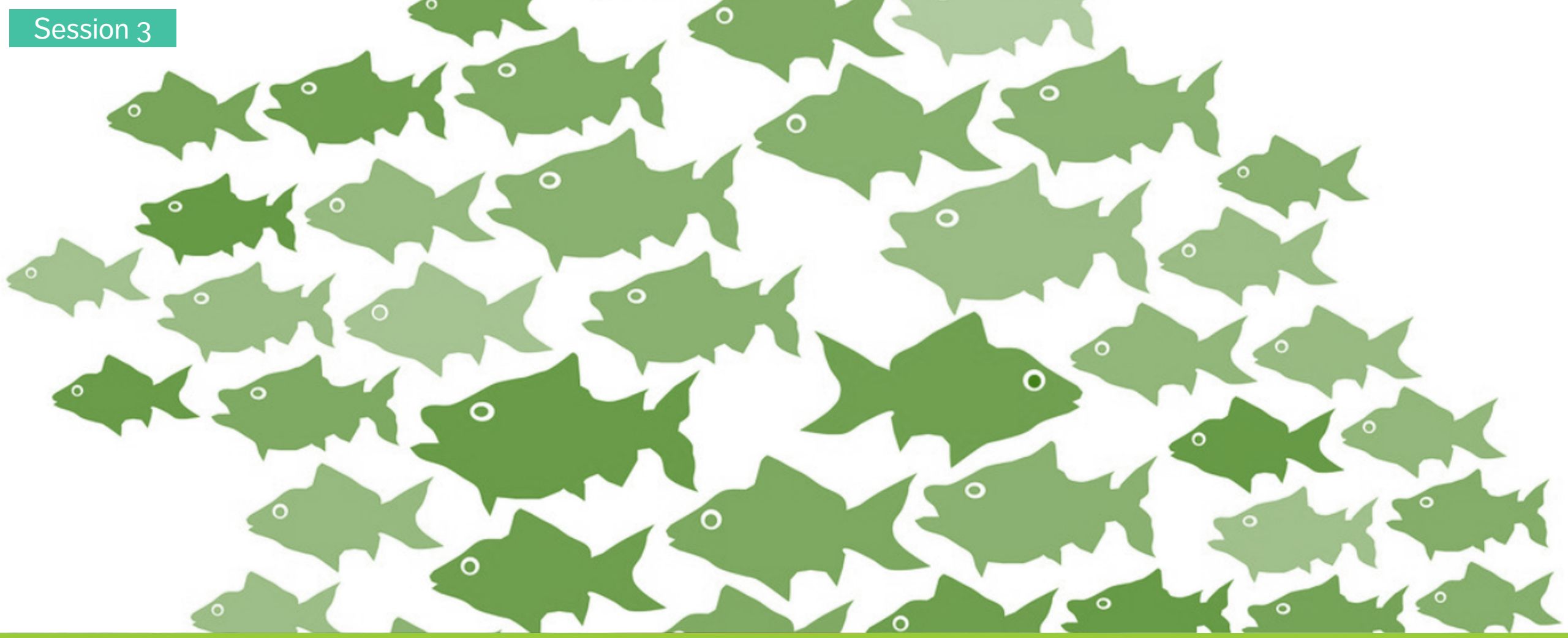
Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    15.4 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.



9. Les observations anormales

Les observations anormales

En pratique, une **observation anormale** peut se présenter comme

- un **"mauvais" objet/mesure :** artefacts de données, fautes, valeurs mal imputées, etc. ;
- une **observation mal classée :** selon les modèles de données existants, l'observation aurait dû être étiquetée différemment ;
- une observation dont les mesures se trouvent dans les **queues de distribution** d'un nombre suffisamment grand d'éléments ;
- un **inconnu inconnu** : un type d'observation totalement nouveau dont l'existence était jusqu'alors insoupçonnée.

Les observations anormales

Une observation peut être anormale dans un contexte, mais pas dans un autre

- un homme adulte de 1.80 m se situe dans le 86^{e} percentile pour les hommes canadiens (grand, mais pas inhabituel).
- en Bolivie, le même homme serait dans le 99.9^{e} percentile (très grand, inhabituel)

La détection des anomalies soulève des **questions intéressantes** pour les analystes et les experts en la matière : dans ce cas, pourquoi existe-t-il un écart aussi important entre les deux populations ?

Les valeurs aberrantes (“outliers”)

Les **observations aberrantes** sont des observations qui sont **atypiques** par rapport aux :

- autres caractéristiques à même l'unité (“*within units*”), et
- valeurs des caractéristiques des autres unités (“*between-units*”)

Les valeurs aberrantes sont des observations qui **ne ressemblent pas aux autres cas** ou qui **contredisent des dépendances** ou des règles **connues**.

Une étude minutieuse est nécessaire pour déterminer si ces valeurs aberrantes doivent être conservées ou supprimées de l'ensemble de données.

La détection des anomalies

Les valeurs aberrantes peuvent être anormales par rapport à une variables de l'unité, ou en combinaison.

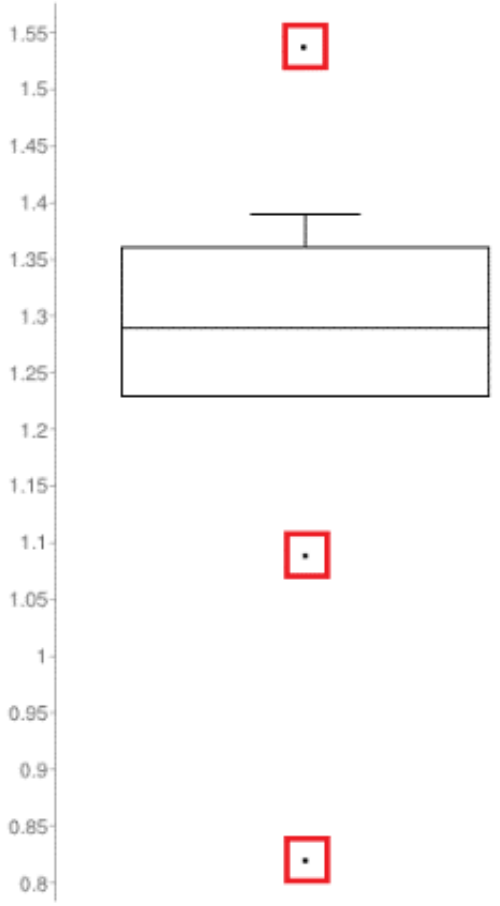
Les anomalies sont par définition **peu fréquentes**, et généralement entourées d'**incertitude en** raison de la petite taille des échantillons.

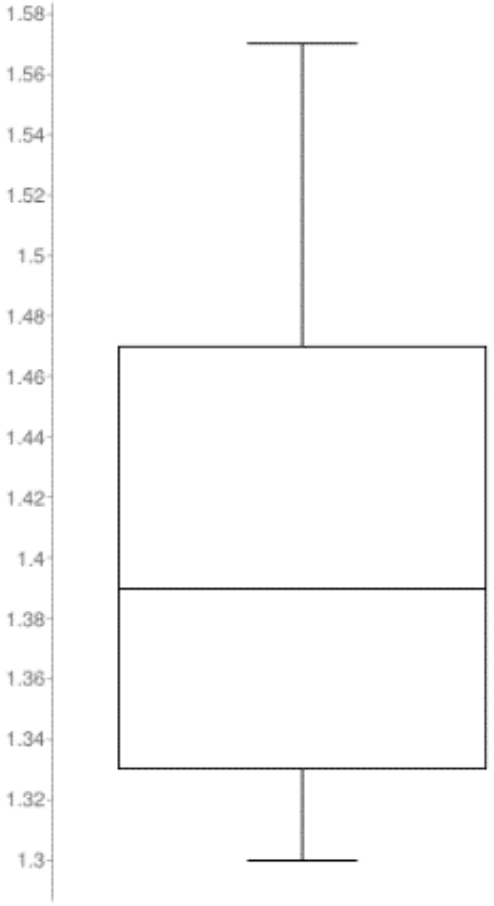
Il est difficile de différencier les anomalies du bruit ou des erreurs de saisie.

Les limites entre les unités normales et déviantes peuvent être **floues**.

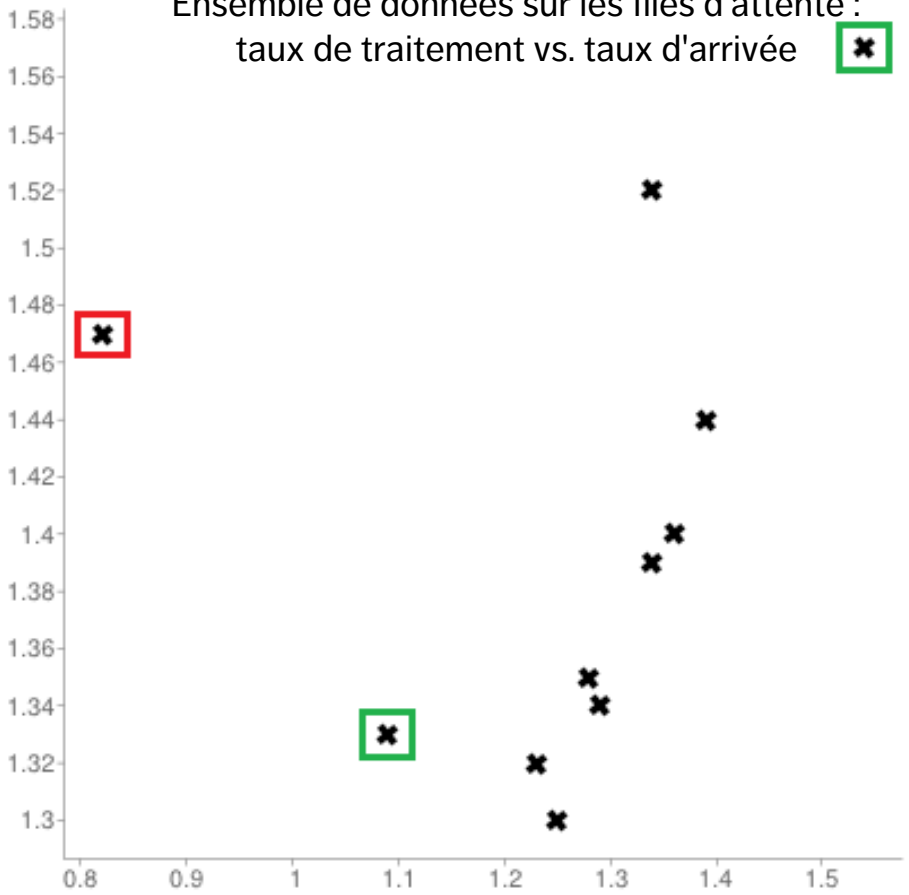
Les anomalies liées à des activités malveillantes sont généralement **déguisées**.

La détection des anomalies





Ensemble de données sur les files d'attente : taux de traitement vs. taux d'arrivée



La détection des anomalies

Il y a de nombreuses méthodes pour identifier les observations anormales ; **aucune d'entre elles n'est infaillible** et il faut faire preuve de discernement

Les méthodes graphiques sont faciles à mettre en œuvre et à interpréter.

- **observations périphériques**

  box-plots, nuages de points, matrices de nuages de points, tour 2D, distance de Cooke, tracés qq normaux

- **données influentes**

  un certain niveau d'analyse doit être effectué (effet de levier)

Attention : si les observations anormales ont été retirées de l'ensemble de données, des unités auparavant "régulières" peuvent devenir anormales !

Algorithmes de détection d'anomalie

Les **méthodes supervisées** utilisent un historique d'observations anormales étiquetées :

- l'expertise du domaine est requise pour étiqueter
- tâche de classification ou de régression
- problème d'occurrence rare

		Prédictions	
		Normales	Anormales
Réalité	Normales	*VN*	*FP*
	Anormales	*FN*	*VP*

Les **méthodes non supervisées** n'utilisent pas d'informations externes :

- méthodes et tests traditionnels
- problème de regroupement ou de règles d'association

Les algorithmes de détection

Le coût des erreurs de classification est souvent supposé être symétrique, ce qui peut conduire à des résultats **techniquement corrects, mais inutiles**.

Par exemple, la grande majorité des passagers aériens (99.99%+) n'emportent pas d'armes ; un modèle qui prédit qu'aucun passager ne fait passer une arme en fraude serait précis à 99.99%+, mais il passerait à côté du problème.

Pour l'**agence de sécurité**, le coût de penser à tort qu'un passager :

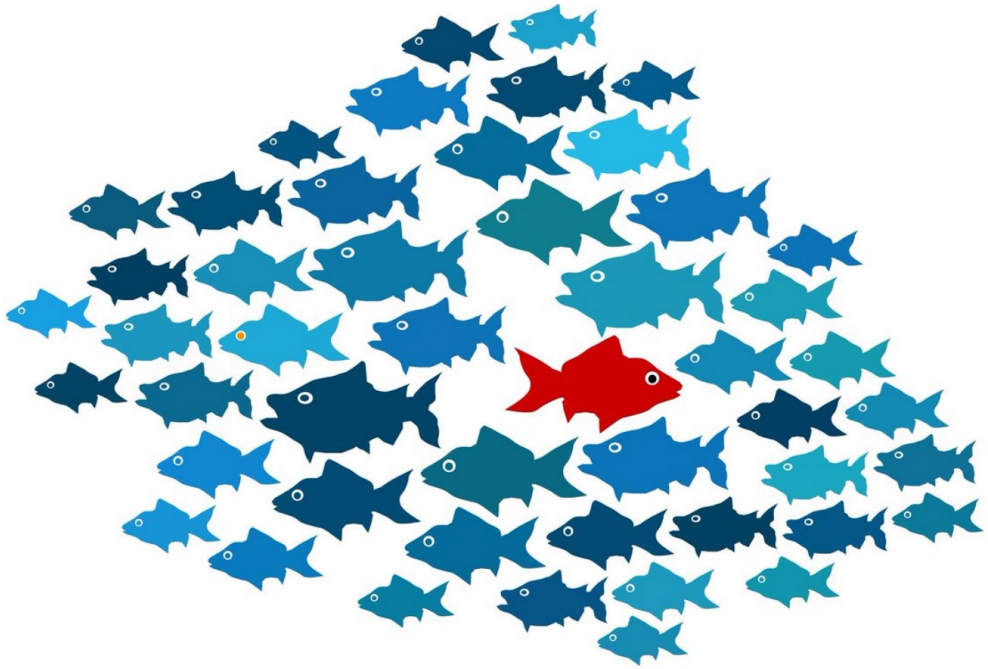
- introduit clandestinement une arme ⟹ coût d'une seule fouille
- ne fait pas passer une arme en fraude ⟹ catastrophe (potentiellement)

Les personnes injustement visées auront un point de vue différent à ce sujet !

Les algorithmes de détection

Si tous les participants à un atelier à l'exception d'un seul, peuvent visionner les conférences par vidéo, cette personne, cette connexion Internet et cet ordinateur sont **anormaux**, car ils ne se comportment pas comme les autres.

Mais cela **NE SIGNIFIE PAS** nécessairement que le comportement différent est celui qui nous intéresse...



Tests simples de valeurs aberrantes

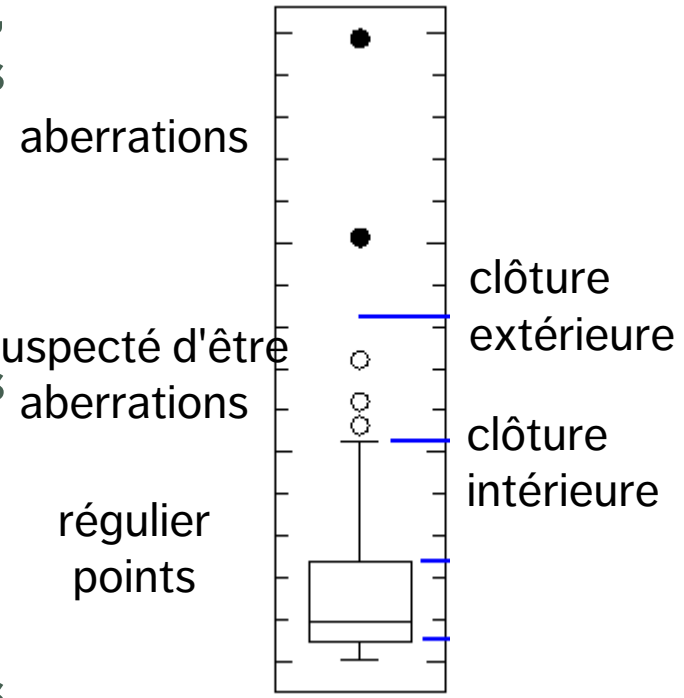
Test Boxplot de Tukey : pour les données normalement distribuées, les observations régulières se situent généralement entre les **clôtures intérieures :**

$$Q_1 - 1.5\times(Q_3 - Q_1) \text{ et } Q_3 + 1.5\times(Q_3 - Q_1).$$

Les **valeurs aberrantes suspectes** se situent entre les **clôtures intérieures** et les **clôtures extérieures :**

$$Q_1 - 3\times(Q_3 - Q_1) \text{ et } Q_3 + 3\times(Q_3 - Q_1).$$

Les valeurs aberrantes se trouvent au-delà des **clôtures extérieures**.



Tests simples de valeurs aberrantes

Le **test Q de Dixon** est utilisé dans les sciences expérimentales pour trouver des valeurs aberrantes dans des ensembles de données (extrêmement) petits (validité douteuse).

La **distance de Mahalanobis** (liée à l'effet de levier) peut être utilisée pour trouver des valeurs aberrantes multidimensionnelles (lorsque les relations sont linéaires).

Autres tests simples :

- **Grubbs** (univarié)
- **Tietjen-Moore** (pour un nombre spécifique de valeurs aberrantes)
- **écart généralisé extrême studentisé** (pour un nombre inconnu de valeurs aberrantes)
- **chi-deux (**les valeurs aberrantes affectant la qualité de l'ajustement)

Test sophistiqués des valeurs aberrantes

- **DBSCAN**, **OR$_h$** , et **LOF** (détection non supervisée des valeurs aberrantes)
- méthode **rang-puissance** (détection supervisée des valeurs aberrantes)
- méthodes **basées sur la distance** ou la **densité** (avec des mesures de distance exotiques)
- **autoencodeurs et erreur de reconstruction** (méthode d'apprentissage profond)
- méthodes d'**occurrences rares** (suréchantillonnage, sous-échantillonnage, CREDOS, PN, SHRINK, SMOTE, DRAMOTE, SMOTEBoost, RareBoost, MetaCost, AdaCost, CSB, SSTBoost, etc.)
- **AVF**, algorithmes **Greedy** (données catégoriques)
- **PCA**, **DOBIN** et autres méthodes de **projection** (pour les données à haute dimension)
- méthodes **subspatiales** et méthodes d'**ensemble**

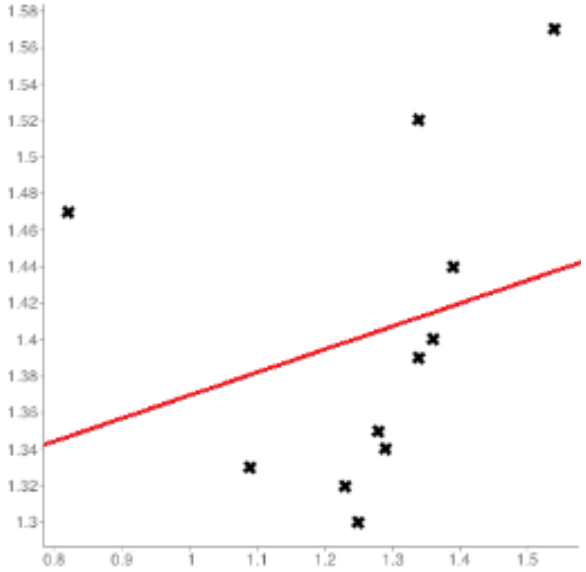
Observations influentes

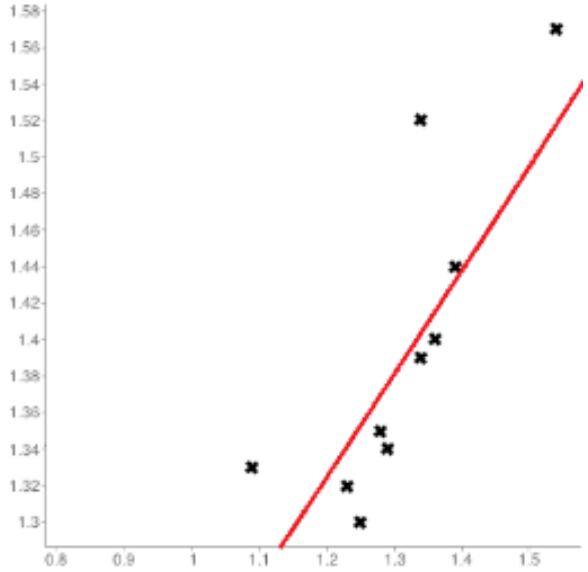
Les **observations influentes** sont des observations dont l'absence entraîne des résultats d'analyse **nettement différents**.

Lorsque des observations influentes sont identifiées, des **mesures correctives** (telles que des transformations de données) peuvent être nécessaires pour minimiser leurs effets indus.

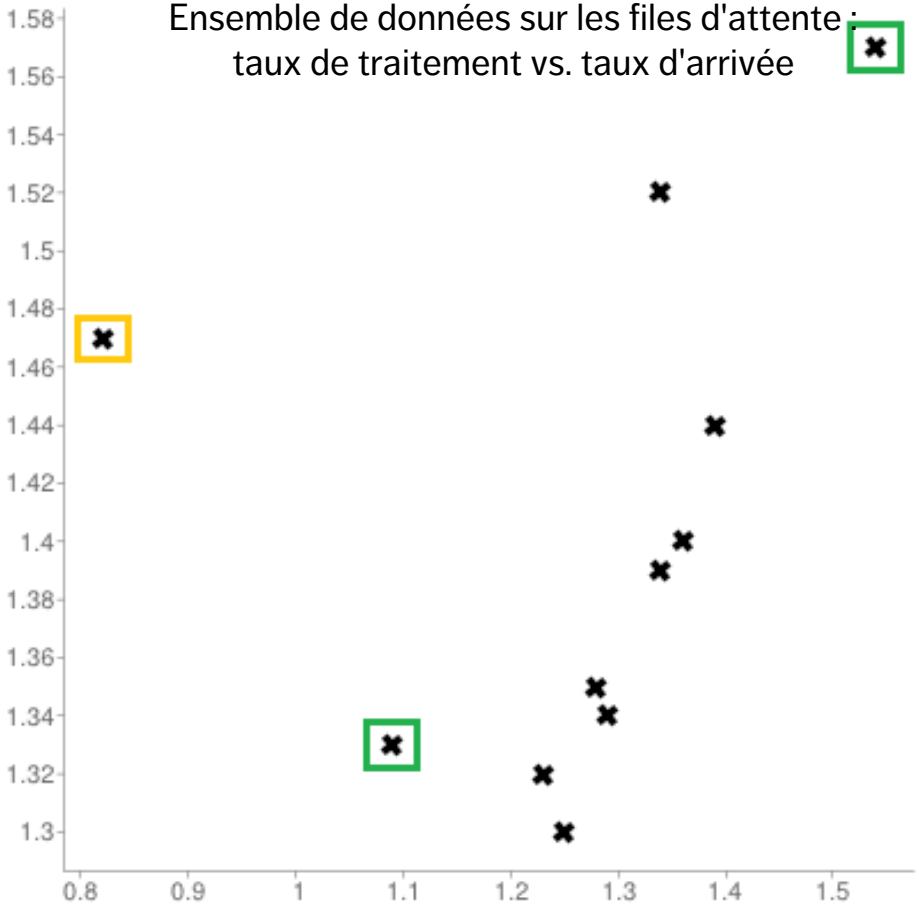
Les valeurs aberrantes peuvent être des observations influentes ; les observations influentes ne sont pas nécessairement des valeurs aberrantes (et *vice-versa)*.

Observations influentes





Ensemble de données sur les files d'attente : taux de traitement vs. taux d'arrivée



Remarques

L'identification des observations influentes est un **processus itératif** car les différentes analyses doivent être exécutées à plusieurs reprises.

L'identification et la suppression entièrement automatisées des observations anormales **ne sont PAS recommandées**.

Utilisez des transformations de données si les données **ne** sont **PAS normalement distribuées**.

Le fait qu'une observation soit une valeur aberrante ou non dépend de **divers facteurs** ; les observations qui finissent par être influentes dépendent de l'**analyse spécifique à effectuer**.

Lectures conseillées

Les observations anormales

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    15.5 Anomalous Observations
    - Anomaly Detection
    - Outlier Tests
    - Visual Outlier Detection

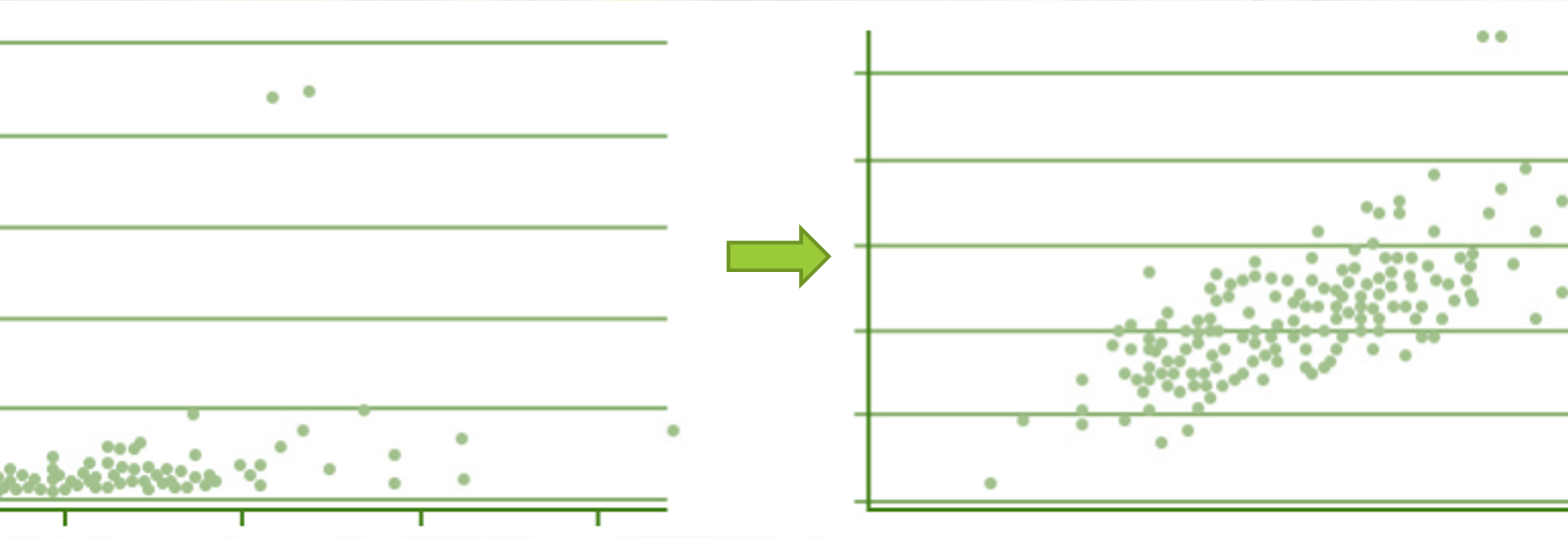
Volume 4: Techniques of Data Analysis

26. Anomaly Detection and Outlier Analysis

Exercices

Les observations anormales

1. Recréez le processus de détection des anomalies utilisé dans Example: Algae Bloom.
2. Trouvez les observations anormales dans les ensembles de données cities.txt et grades (le cas échéant).
3. Trouvez les observations anormales dans un ensemble de données de votre choix.



10. La dimensionnalité et les transformations de données

La dimensionnalité des données

En analyse des données, la **dimension** est le nombre d'attributs qui sont rassemblés dans un ensemble de données (le **nombre de colonnes**).

Nous pouvons considérer le nombre de variables utilisées pour décrire chaque objet (ligne) comme un vecteur décrivant cet objet : la dimension est simplement la **taille** de ce vecteur.

(**Remarque :** le terme "dimension" est utilisé différemment dans les contextes de “business intelligence”)

Dimensionnalité élevée et “Big Data”

Les ensembles de données peuvent être “massifs" de différentes manières :

- trop grand pour la **gestion** (ne peut être stockée, accédée, manipulée correctement en raison du nombre d'observations, du nombre de caractéristiques, de la taille globale)
- les dimensions peuvent aller à l'encontre des **hypothèses de modélisation** (# de caractéristiques > # d'observations)

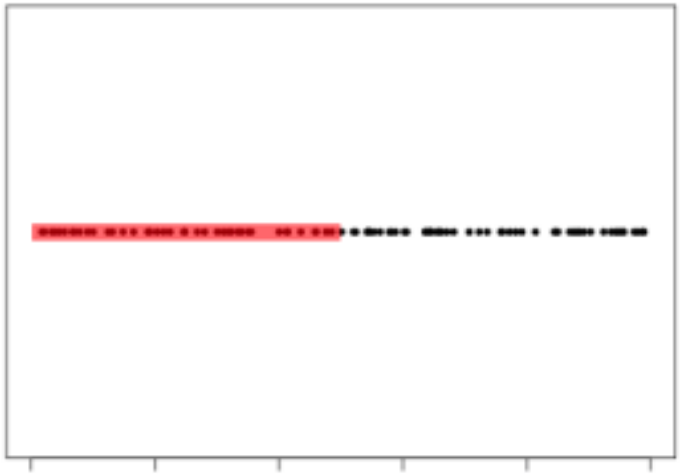
Exemples :

- plusieurs capteurs enregistrant plus de 100 observations par seconde dans une vaste zone géographique sur une longue période = **données massives**
- dans la *matrice terme-document d'*un corpus (colonnes = termes, rangées = documents), le nombre de termes est généralement beaucoup plus élevé que le nombre de documents, ce qui conduit à des **données éparses**

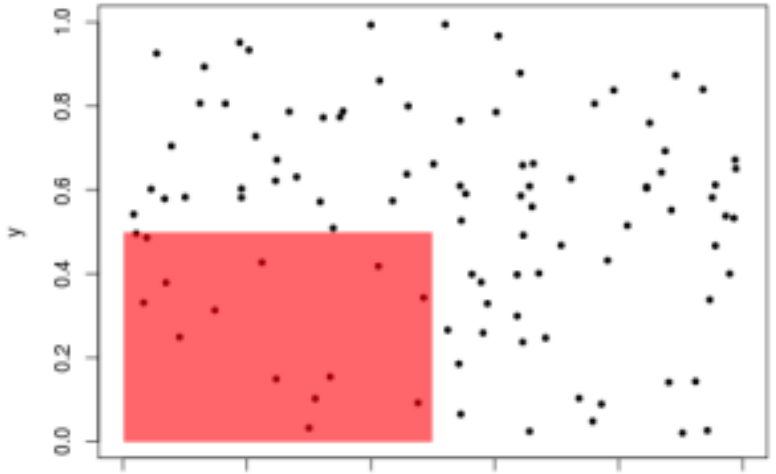
[http://simplystatistics.org]

Le fléau de la dimensionnalité

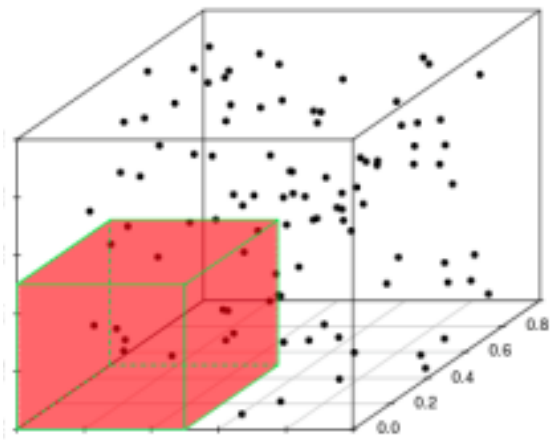
42% des données sont capturées



14% des données sont capturées



7% des données sont capturées



$N = 100$ observations, uniformément distribuées sur $[0,1]^d, d = 1, 2, 3.$
% des observations capturées par $[0,0.5]^d, d = 1,2,3.$

L’échantillonnage d’observations

Question : est-ce que toutes les données doivent être utilisée ?

Si les rangées sont sélectionnées au hasard (avec/sans remise), l'échantillon résultant peut être **représentatif** de l'ensemble des données.

Inconvénients :

- si le signal d'intérêt est rare, l'échantillonnage peut le noyer complètement
- si l'agrégation se produit en fin de parcours, l'échantillonnage affectera nécessairement les chiffres (passagers vs. vols)
- sur un fichier massif, même les opérations simples (e.g., trouver le # d’instances) peuvent être coûteuses – utilisez des **informations préalables sur la structure de l'ensemble** !

La sélection de caractéristiques

La suppression des variables **non pertinentes/redondantes** est une tâche courante du traitement des données.

Motivations :

- les outils de modélisation ne les gèrent pas bien (inflation de la variance, etc.)
- réduction de la dimension (# variables ≫ # observations)

Approches :

- filtre vs. enveloppe (“filter” vs. “wrapper”)
- non supervisé vs. supervisé

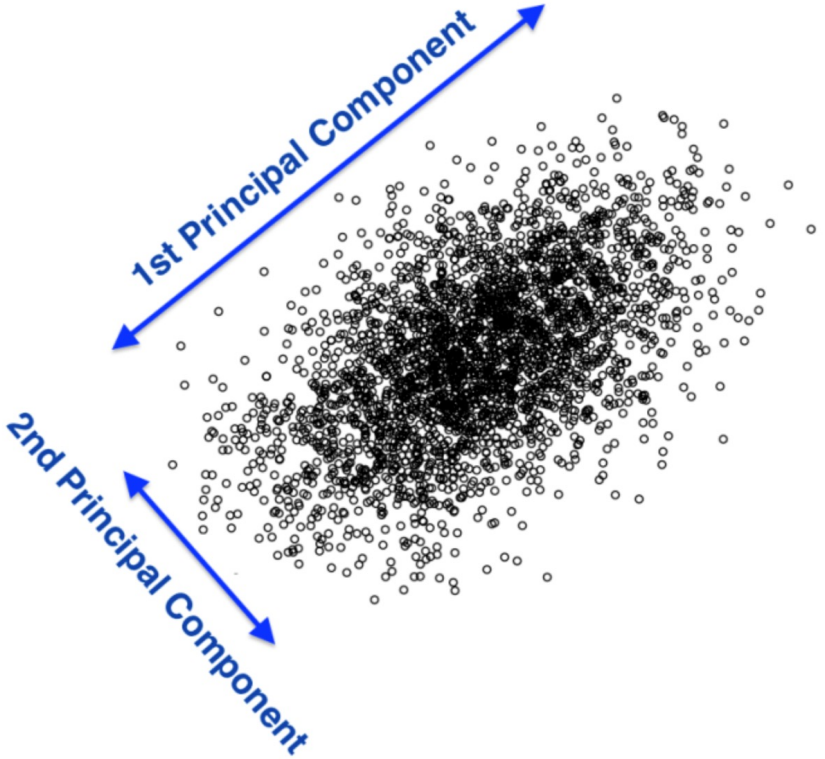
[A. Ng, K. Soo, *Numsense !*]

La réduction de dimension : ACP

Motivation : contenu nutritionnel des aliments

Quelle est la meilleure façon de différencier les produits alimentaires ? La teneur en vitamines, en matières grasses, ou en protéines ? Un peu de tout ?

L'analyse en composantes principales (ACP) peut être utilisée pour trouver les combinaisons de variables le long desquelles les observations sont **les plus répartis** (réduction de la dimension)**.**



[A. Ng, K. Soo, *Numsense !*]

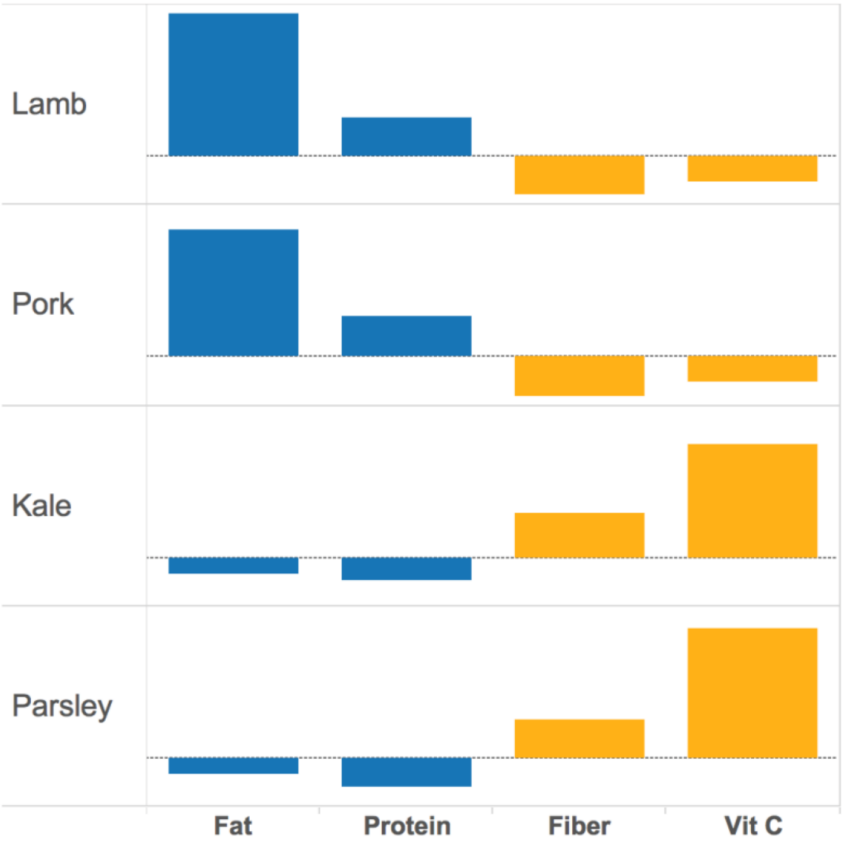
La réduction de dimension : ACP

La présence de nutriments semble être **corrélée** entre les différents aliments.

Dans un (petit) échantillon, les niveaux de *graisses* et de *protéines* semblent en phase, tout comme ceux des *fibres* et de la *vitamine C*.

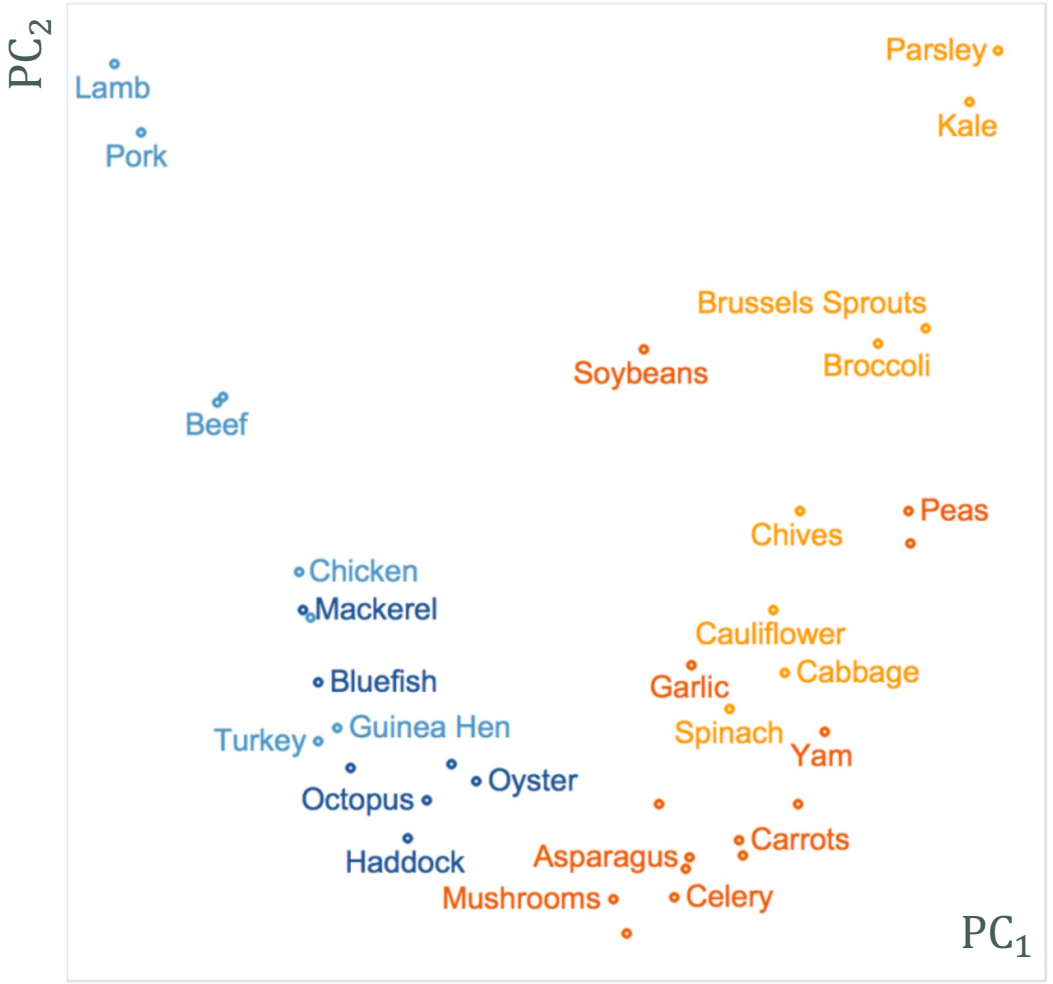
Dans un ensemble de données plus vaste, les corrélations sont $r = 0.56$ et $r = 0.57$

2 variables **dérivées** peuvent-elles expliquer cela ?



$$\mathrm{PC}_1 = -0.45\times\mathrm{Fat} - 0.55\times\mathrm{Protein} + 0.55\times\mathrm{Fiber} + 0.44\times\mathrm{Vitamin\ C}$$
$$\mathrm{PC}_2 = 0.66\times\mathrm{Fat} + 0.21\times\mathrm{Protein} + 0.19\times\mathrm{Fiber} + 0.70\times\mathrm{Vitamin\ C}$$

La différenciation ACP



différencie les légumes des viandes ; différencie 2 **sous-catégories** au sein de celles-ci :

- les **viandes** sont concentrées sur la gauche (PC_1 faibles)
- **les légumes** sont concentrés sur la droite (PC_1 élevé)
- **les fruits de mer** ont une plus faible teneur en *matières grasses* (PC_2 faible) et sont concentrés en bas
- les **légumes non feuillus** ont une teneur plus faible en *vitamine C* (PC_2 faible) et sont également regroupés en bas

Les transformations communes

Les modèles exigent parfois que certaines hypothèses relatives aux données soient respectées (normalité des résidus, linéarité, etc.).

Si les données brutes ne répondent pas aux exigences, nous pouvons soit :

- abandonner le modèle
- tenter de **transformer** les données

La deuxième approche nécessite une **transformation inverse** pour pouvoir tirer des conclusions sur les **données d'origine**.

Les transformations communes

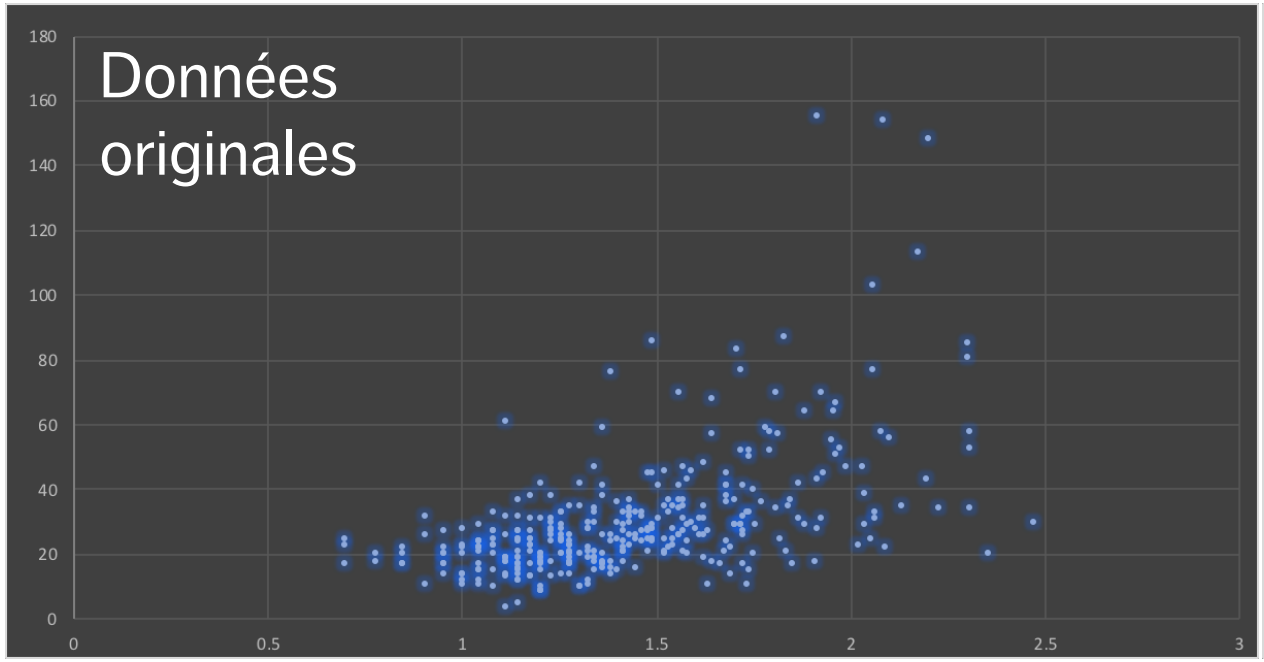
Dans le contexte de l'analyse des données, les transformations sont **monotones :**

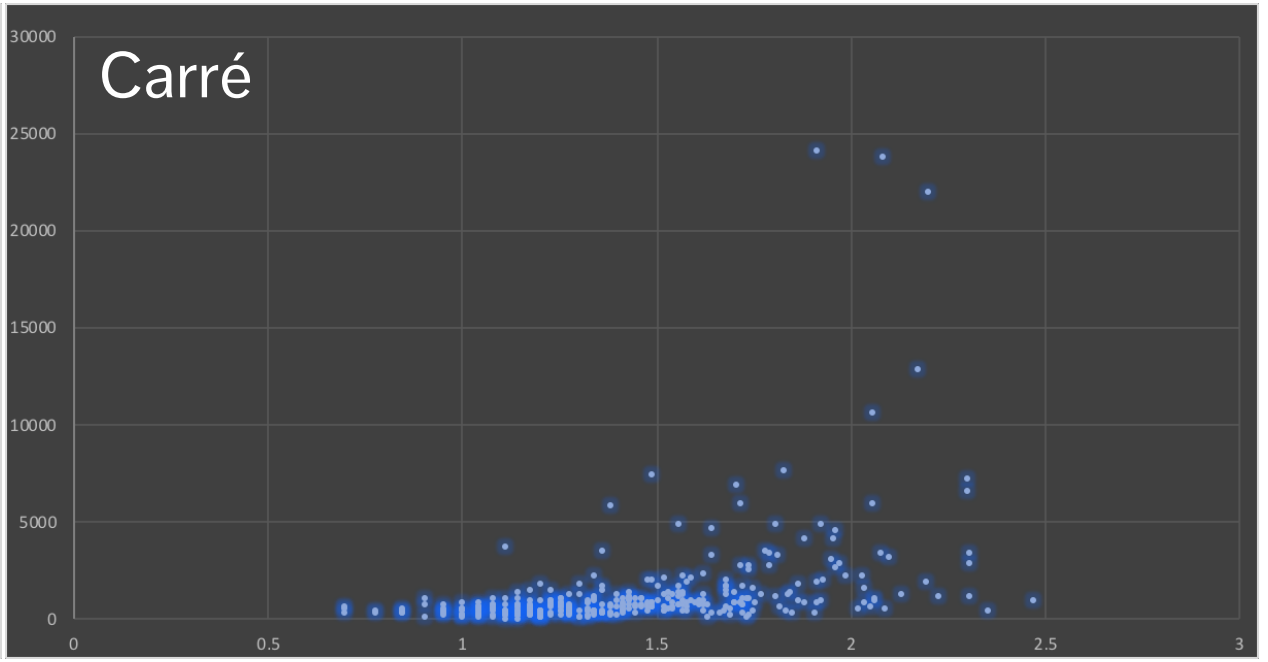
- logarithmique
- racine carrée, inverse, puissance :
- exponentielle
- Box-Cox, etc.

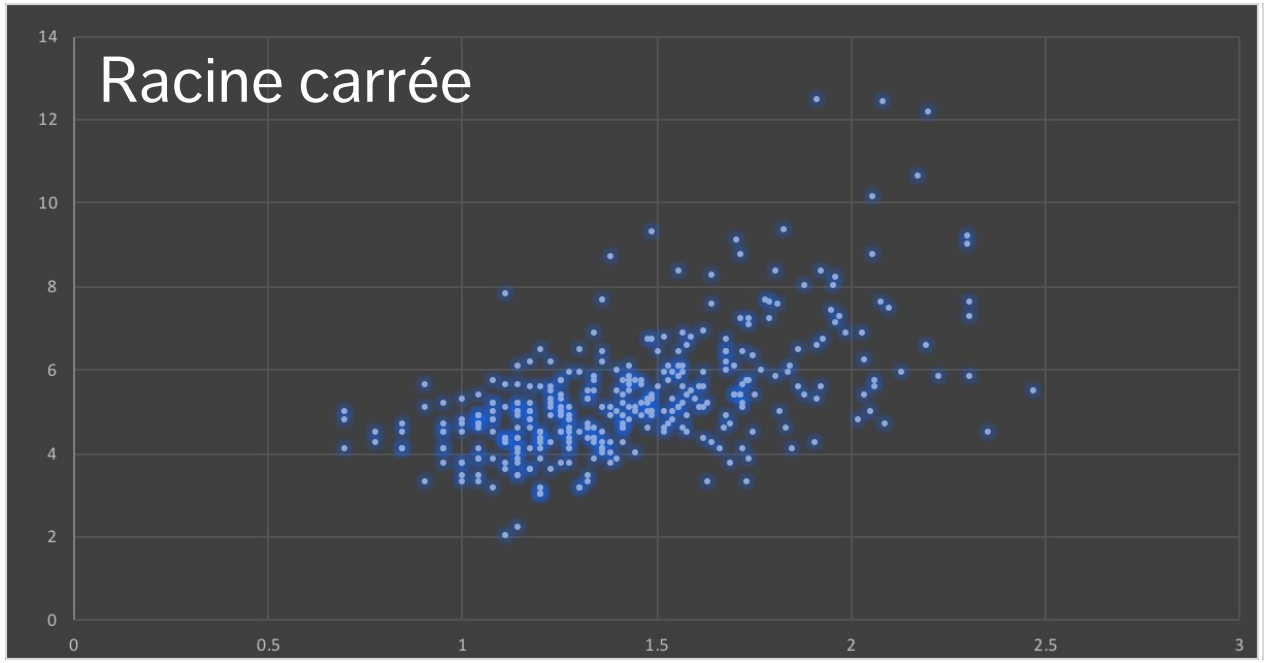
Les transformations sur les prédicteurs X peuvent atteindre la linéarité, mais à un prix (les corrélations ne sont pas préservées, par exemple).

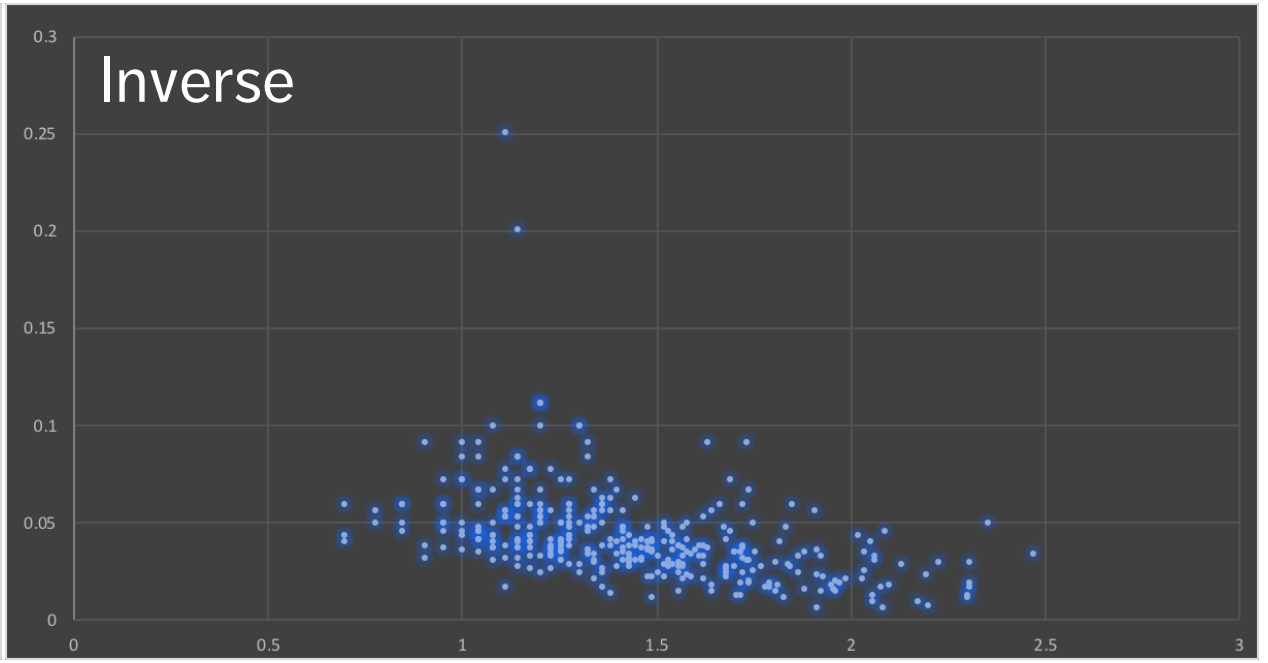
Les transformations sur la réponse Y peuvent aider avec la non-normalité et la variance inégale des termes d'erreur.

Session 4


Données originales
180
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3


Carré
30000
25000
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3


Racine carrée
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Inverse
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3

= moyenne géométrique

La transformation de Box-Cox

Supposons le modèle habituel $Y_j = \sum_i \beta_i X_{j,i} + \varepsilon_j$ avec soit

- des résidus asymétriques ;
- une variance non constante, et/ou
- une tendance non linéaire.

La **transformation de Box-Cox** $Y_j \mapsto Y_j'(\lambda)$ suggère un choix : sélectionnez λ qui maximise la log-vraisemblance correspondante

$$Y_j'(\lambda) = \begin{cases} \mathrm{gm}(\boldsymbol{Y}) \times \ln(Y_j), & \lambda = 0 \\ \lambda^{-1} \mathrm{gm}(\boldsymbol{Y})^{1-\lambda} \times (Y_j{}^{\lambda} - 1), & \lambda \neq 0 \end{cases}$$

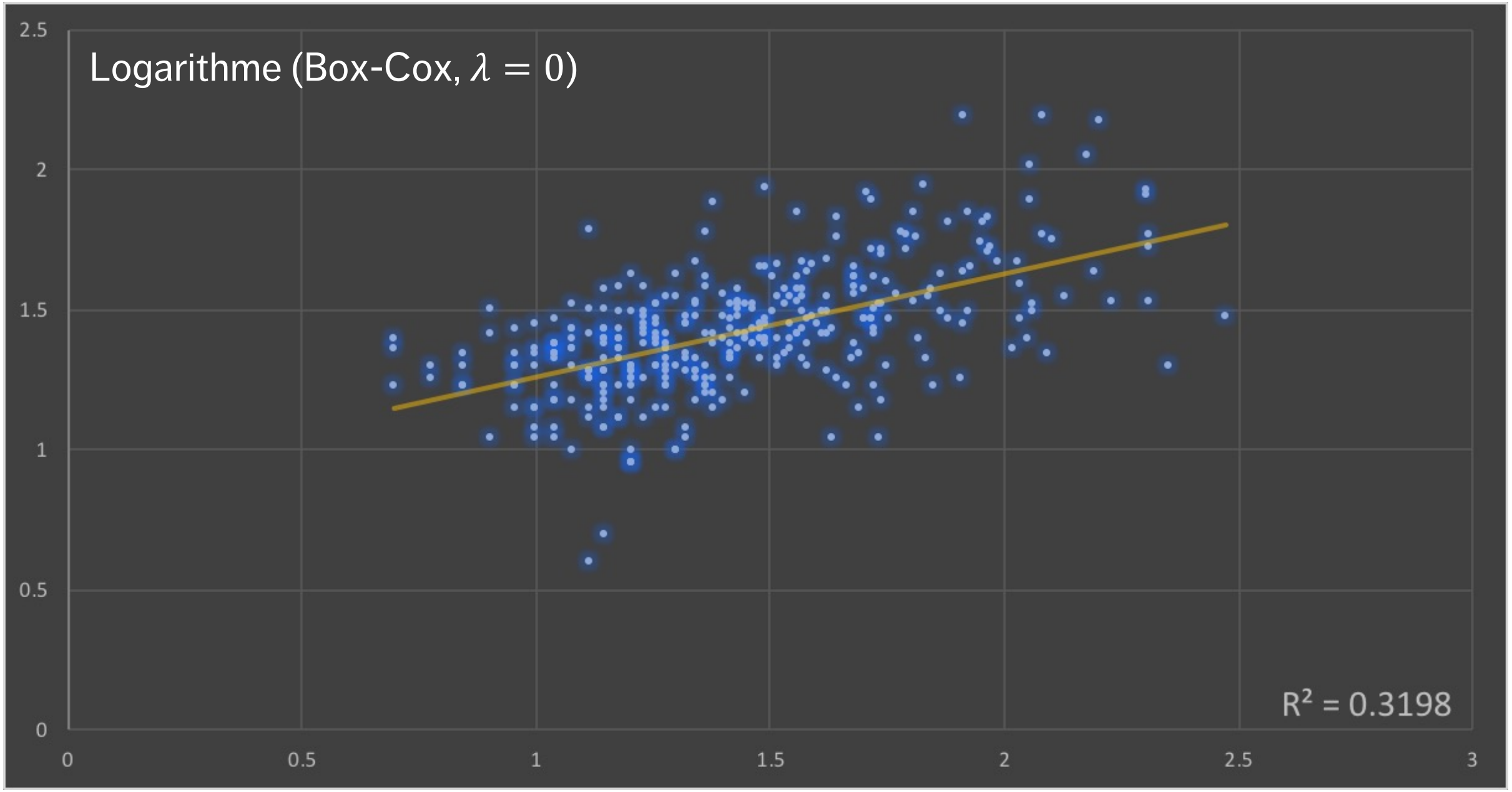
La transformation de Box-Cox

La procédure fournit un **guide** pour sélectionner une transformation.

Des justifications théoriques/pratiques peuvent exister pour un choix de λ.

Une analyse résiduelle est encore nécessaire pour s'assurer que le choix était approprié.

Mieux vaut travailler avec (ou interpréter) les données **transformées**.


Logarithme (Box-Cox, $\lambda = 0$)
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
$R^2 = 0.3198$

La mise à l'échelle

Les variables numériques peuvent avoir différentes **échelles** (e.g., des poids et des hauteurs).

La variance d'une variable à grande échelle est généralement supérieure à celle d'une variable à petite échelle, ce qui peut introduire un biais.

La **standardisation** crée une variable avec une moyenne 0 et un écart-type 1 :

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

La **normalisation** crée une variable dans l'intervalle $[0,1]$: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

La discrétisation

Pour réduire la complexité des calculs, il peut être nécessaire de remplacer une variable numérique par une variable **ordinale** (e.g., passer de la *taille* à "*petit*", "*moyen*", "*grand*").

L'**expertise de domaine** peut être utilisée pour déterminer les limites des bacs (bien que cela puisse introduire un biais inconscient dans les analyses).

En absence d'une telle expertise, on peut fixer les limites de sorte que soit :

- les bacs contiennent chacun le même nombre d'observations
- les bacs ont tous la même largeur
- la performance d'un certain outil de modélisation est maximisée

La création de variables

Il peut être nécessaire d'introduire de nouvelles variables :

- des **relations fonctionnelles** d'un certain sous-ensemble de caractéristiques disponibles
- pour imposer l'**indépendance des observations**
- pour imposer l'**indépendance des caractéristiques**
- pour simplifier l'analyse en examinant des **résumés agrégés** (en analyse de texte)

Dépendances temporelles ⟶ analyse des séries chronologiques (décalages ?)

Dépendances spatiales ⟶ analyse spatiale (voisins ?)

Lectures conseillées

La dimensionnalité et les transformations de données

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    15.6 Data Transformations
    - Common Transformations
    - Box-Cox Transformation
    - Scaling
    - Discretizing
    - Creating Variables

Volume 3: Spotlight on Machine Learning

23. Feature Selection and Dimension Reduction

Exercices

La dimensionnalité et les transformations de données

1. En utilisant Example: Algae Bloom comme base, mettez à l'échelle, discrétisez et créez de nouvelles variables à partir de l'ensemble de données algae blooms.
2. Mettez à l'échelle, discrétisez et créez de nouvelles variables à partir des ensembles de données grades et cities.txt.
3. Mettez à l'échelle, discrétisez et créez de nouvelles variables à partir d'un ensemble de données de votre choix.