Affaires mondiales Canada | Global Affairs Canada

Canada

CANADIAN FOREIGN SERVICE INSTITUTE | L'INSTITUT CANADIEN DU SERVICE EXTÉRIEUR

Introduction à l'analyse des données

TRAITEMENT DES DONNÉES

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

NETTOYAGE DES DONNÉES

TRAITEMENT DES DONNÉES

« Évidemment, la meilleure façon de traiter les données manquantes est de ne pas en avoir. »

T. Orchard, M. Woodbury

« La phrase la plus excitante à entendre, celle qui annonce le plus de découvertes, n'est pas « Eurêka ! » mais "C'est drôle... »

I. Asimov

4 REMARQUES TRÈS IMPORTANTES

Ne travaillez **JAMAIS** sur l'ensemble de données original. Faites des copies en cours de route.

Documentez **TOUTES** vos étapes et procédures de nettoyage.

Si vous vous surprenez à nettoyer une trop grande partie de vos données, **ARRÊTEZ**. Il y a peut-être un problème avec la procédure de collecte des données.

Réfléchissez **à deux fois** avant de rejeter une observation entière.

APPROCHES DU NETTOYAGE DES DONNÉES

Il existe deux approches philosophiques du nettoyage et de la validation des données :

- méthodique
- narratif

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 des données et à essayer de repérer les schémas improbables et irréguliers.

AVANTAGES ET INCONVÉNIENTS

Méthodique (syntaxe)

- Avantages : liste de contrôle est **indépendante du contexte** ; pipelines **faciles à mettre en œuvre** ; erreurs courantes/observations non valides sont **facilement identifiées**
- Inconvénients : peut **s'avérer chronophage** ; impossible d'identifier de nouveaux types d'erreurs

Narratif (sémantique)

- Avantages : peut simultanément permettre de **comprendre les données** ; les faux départs sont (au maximum) aussi coûteux que le passage à une approche mécanique
- Inconvénients : peut passer à côté d'importantes sources d'erreurs et d'observations non valables pour les ensembles de données comportant un **grand nombre de caractéristiques** ; la connaissance du domaine peut fausser le processus en négligeant des zones inintéressantes de l'ensemble de données

OUTILS ET MÉTHODES

Méthodique

- liste des problèmes potentiels (Bingo du nettoyage des données)
- code pouvant être réutilisé dans différents contextes

Narratif

- visualisation
- résumé des données
- tableaux de distribution
- petits multiples
- analyse des données

Bingo du nettoyage des données

random missing values	outliers	values outside of expected range - numeric	factors incorrectly/iconsistently coded	date/time values in multiple formats
impossible numeric values	leading or trailing white space	badly formatted date/time values	non-random missing values	logical inconsistencies across fields
characters in numeric field	values outside of expected range - date/time	DCB!	inconsistent or no distinction between null, 0,not available, not applicable,missing	possible factors missing
multiple symbols used for missing values	???	fields incorrectly separated in row	blank fields	logical iconsistencies within field
entire blank rows	character encoding issues	duplicate value in unique field	non-factor values in factor	numeric values in character field

APPROCHES DU NETTOYAGE DES DONNÉES

L'approche narrative s'apparente à l'élaboration d'une grille de mots croisés avec un stylo et à l'inscription **occasionnelle** de réponses potentiellement fausses, pour voir où cela vous mène.

L'approche mécanique s'apparente à travailler avec un crayon, un dictionnaire et à ne jamais noter une réponse à moins d'être certain qu'elle est correcte.

Vous résoudrez plus de grilles (et ce sera plus « flashy ») de la première manière, mais vous vous tromperez rarement à l'aide de la seconde.

C'est la même chose avec les données : les analystes doivent être à l'aise avec les deux approches.

LES TYPES D'OBSERVATIONS MANQUANTES

Les champs vides se déclinent en 4 saveurs :

- **Non-réponse**
  une observation était attendue mais aucune n'avait été saisie
- **Problème de saisie des données**
  une observation a été enregistrée mais n'a pas été saisie dans l'ensemble de données
- **Entrée non valide**
  une observation enregistrée a été considérée comme non valide et a été supprimée
- **Champ vide attendu**
  un champ a été laissé vide, mais comme prévu

Un trop grand nombre de valeurs manquantes (des 3 premiers types) peut indiquer des problèmes liés au processus de **collecte des données** (nous y reviendrons plus tard) ; un trop grand nombre de valeurs manquantes (du 4e type) peut indiquer une mauvaise **conception du questionnaire**.

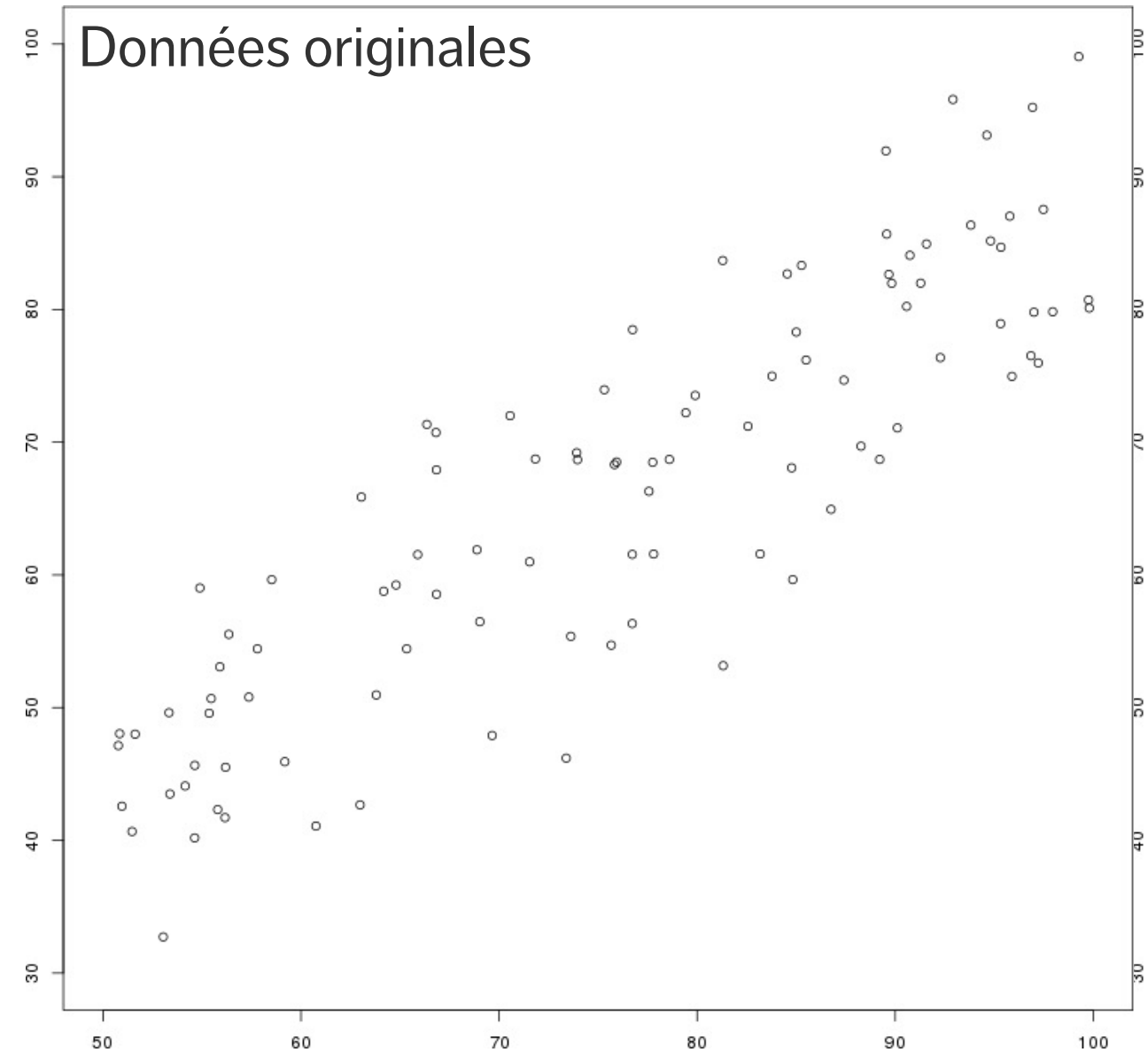
L'ARGUMENT EN FAVEUR DE L'IMPUTATION

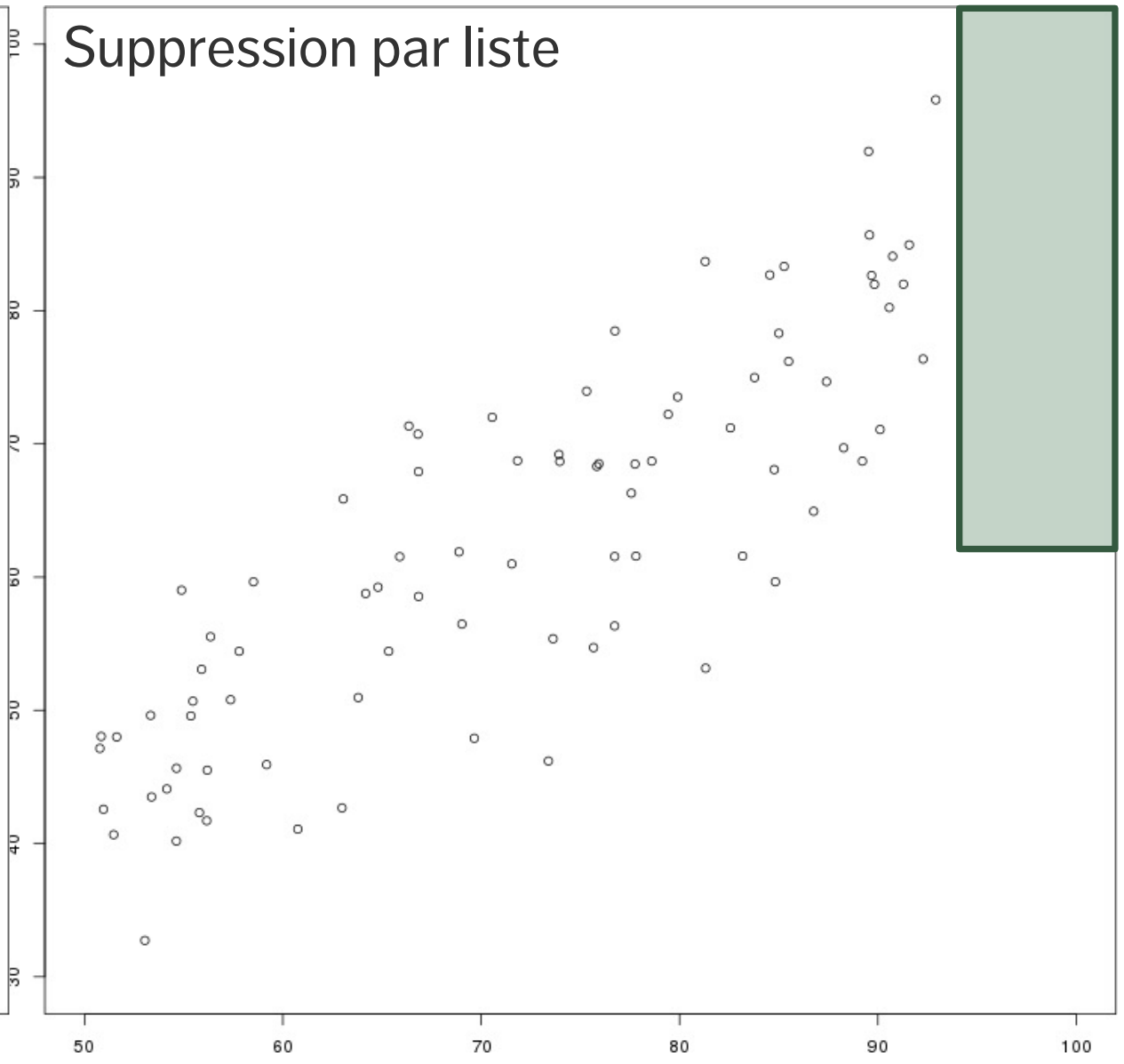
Toutes les méthodes d'analyse ne peuvent pas facilement accommoder de telles observations :

- **Supprimer** l'observation manquante
  - non recommandé, sauf si les données sont manquantes de manière complètement aléatoire dans l'ensemble de l'ensemble de données
  - acceptable dans certaines situations (comme un petit nombre de valeurs manquantes dans un grand ensemble de données)

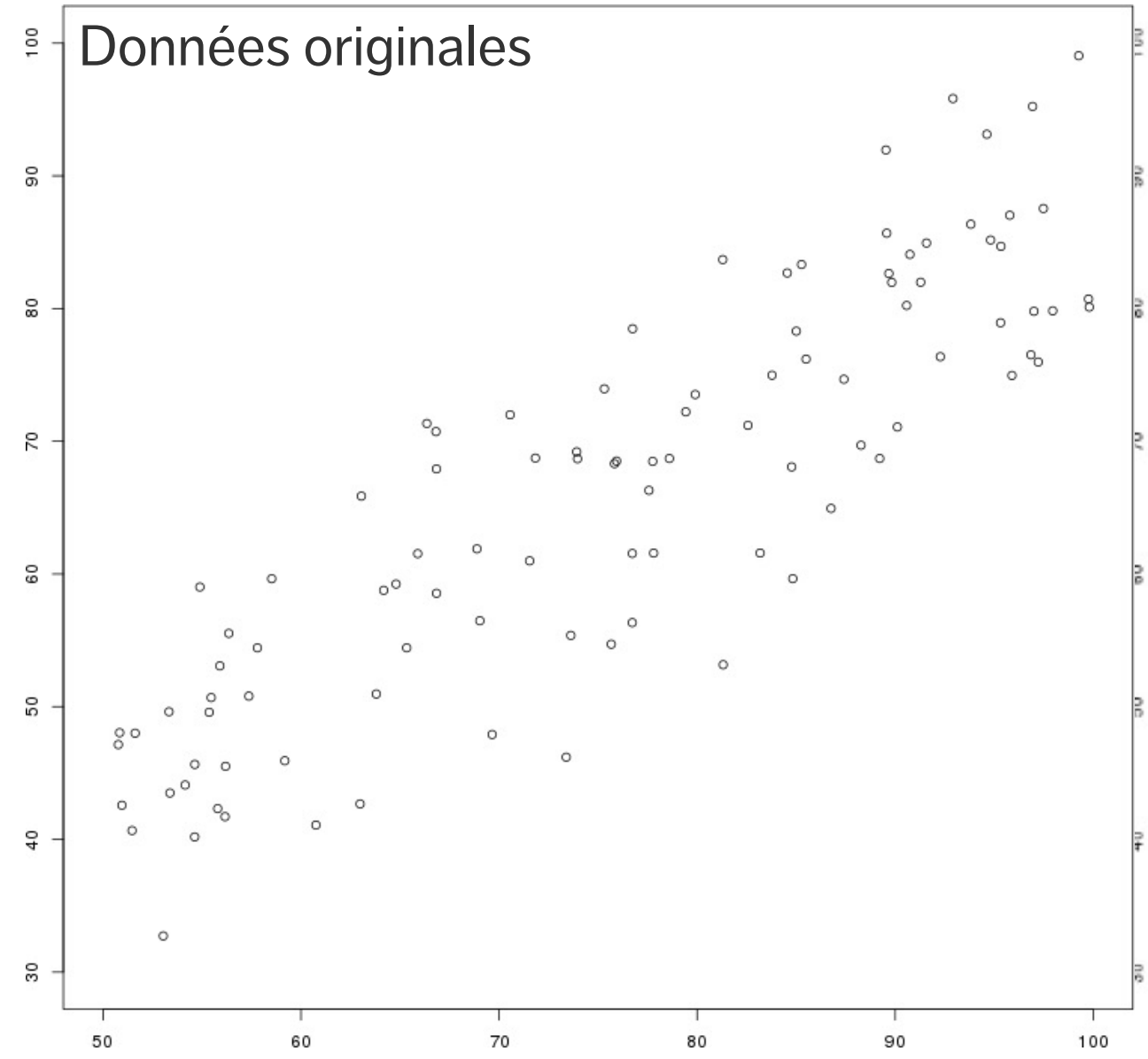
- Trouver une **valeur de remplacement (imputation)**
  - principal inconvénient : nous ne savons jamais quelle aurait été la valeur réelle
  - souvent la meilleure option disponible

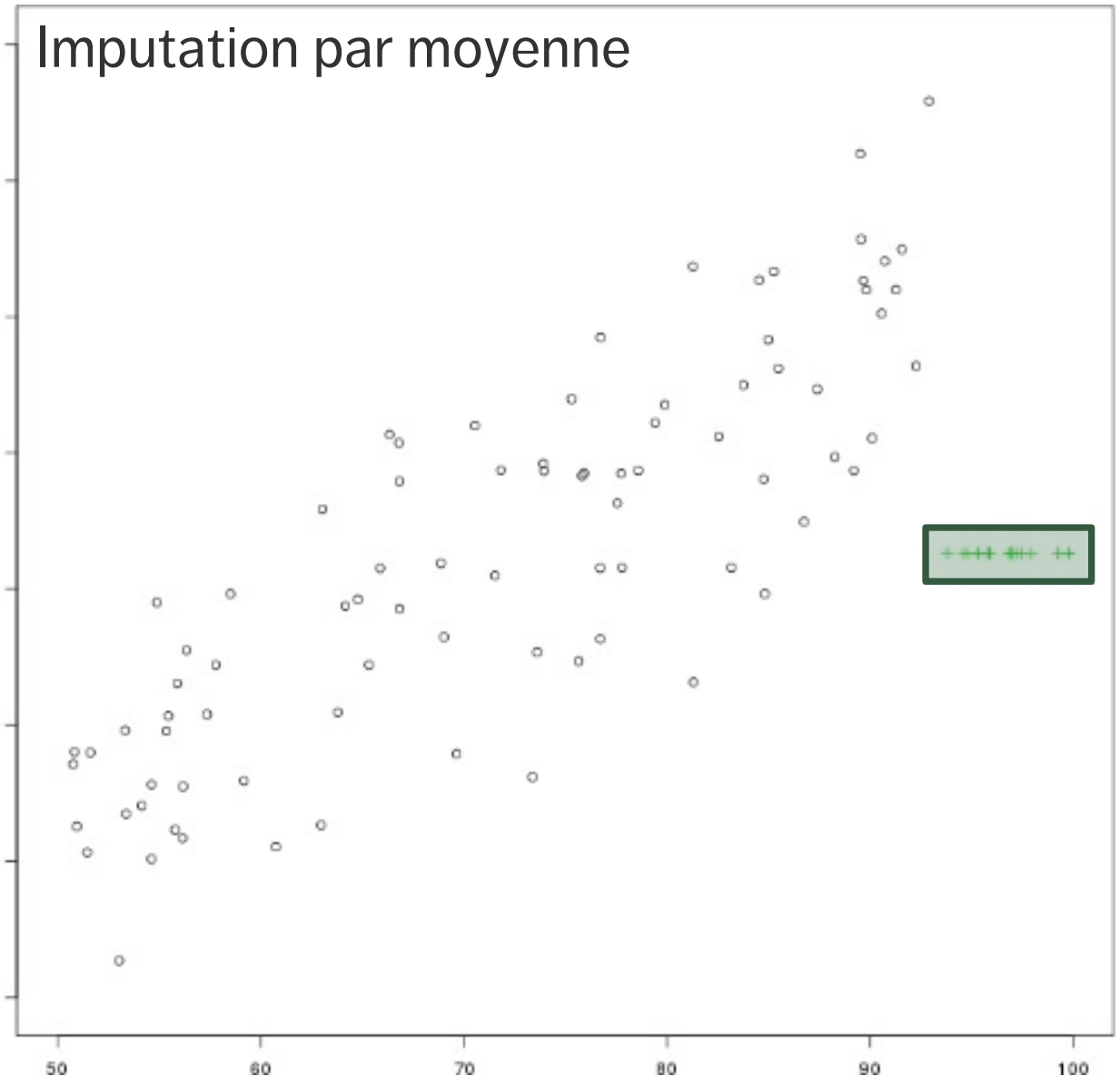
Données artificielles : les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.



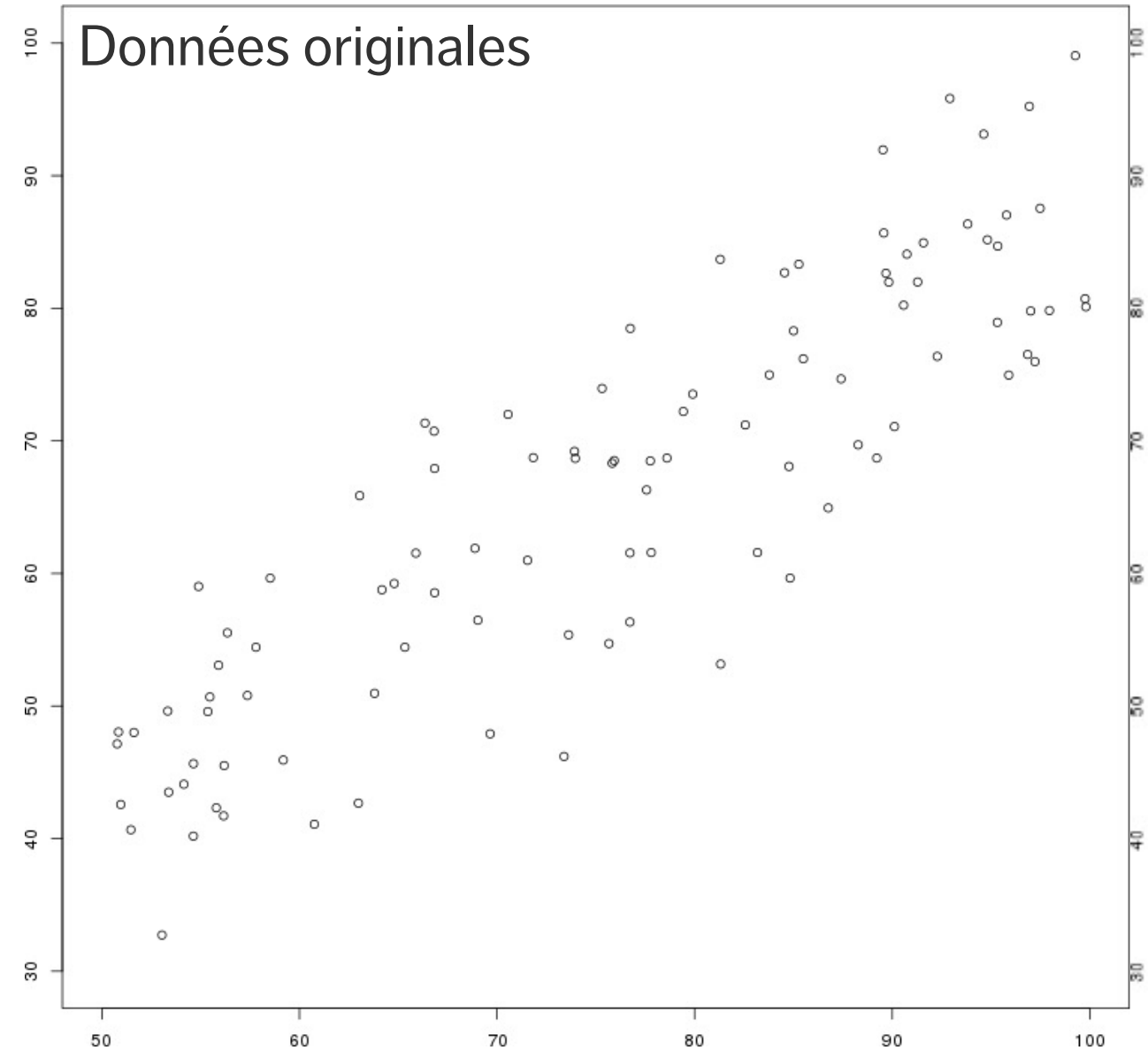


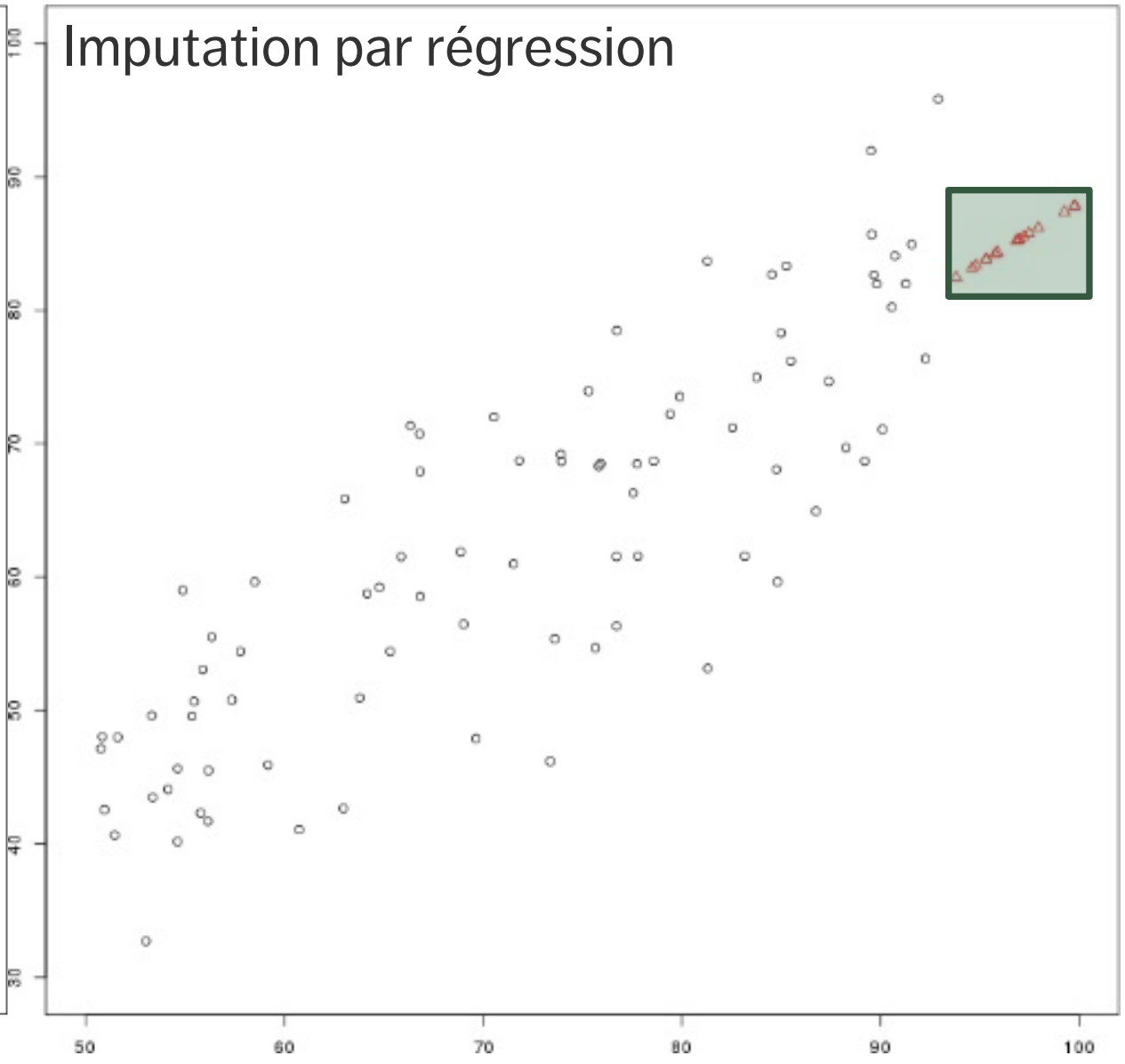
Données artificielles : les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.



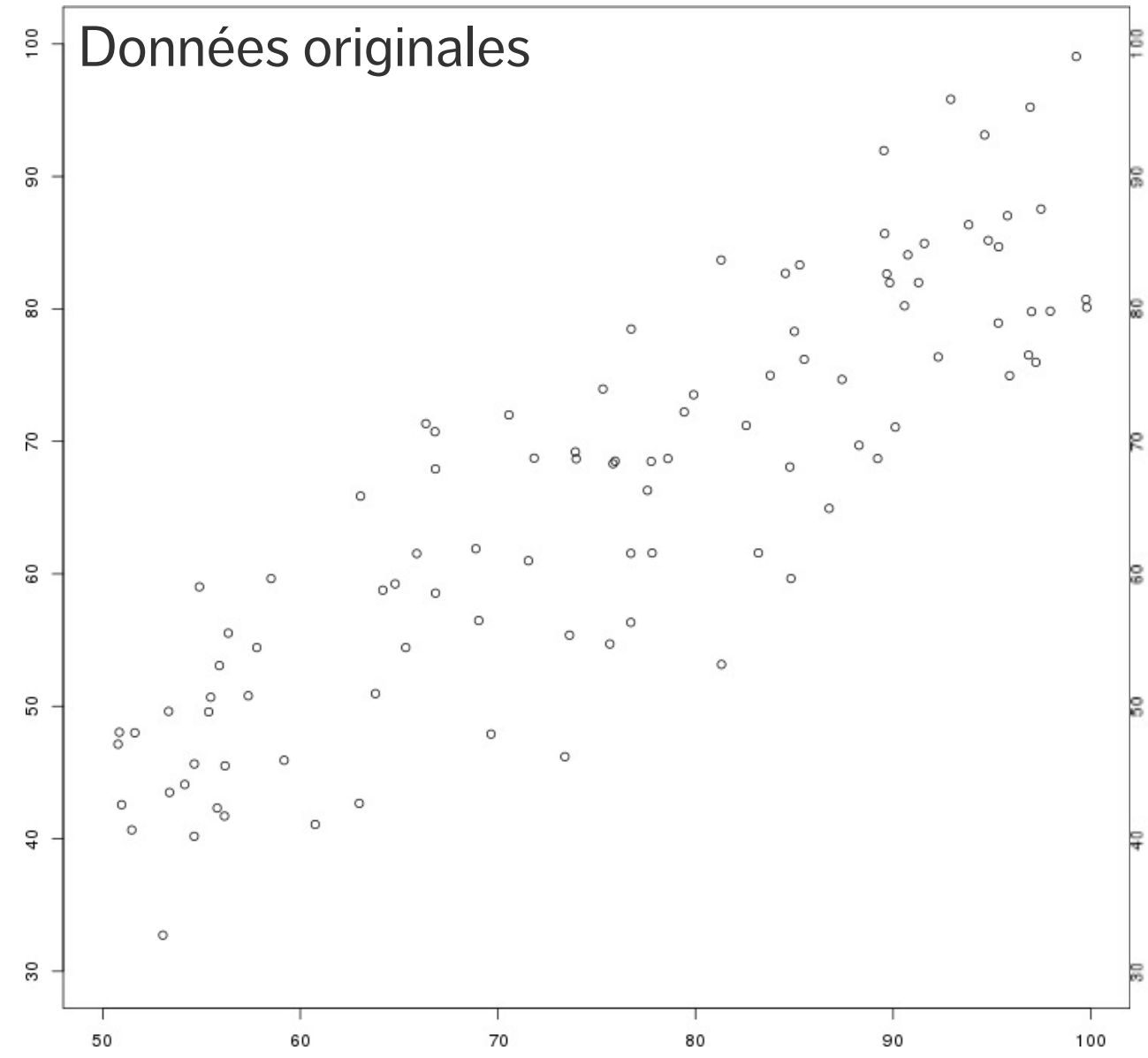


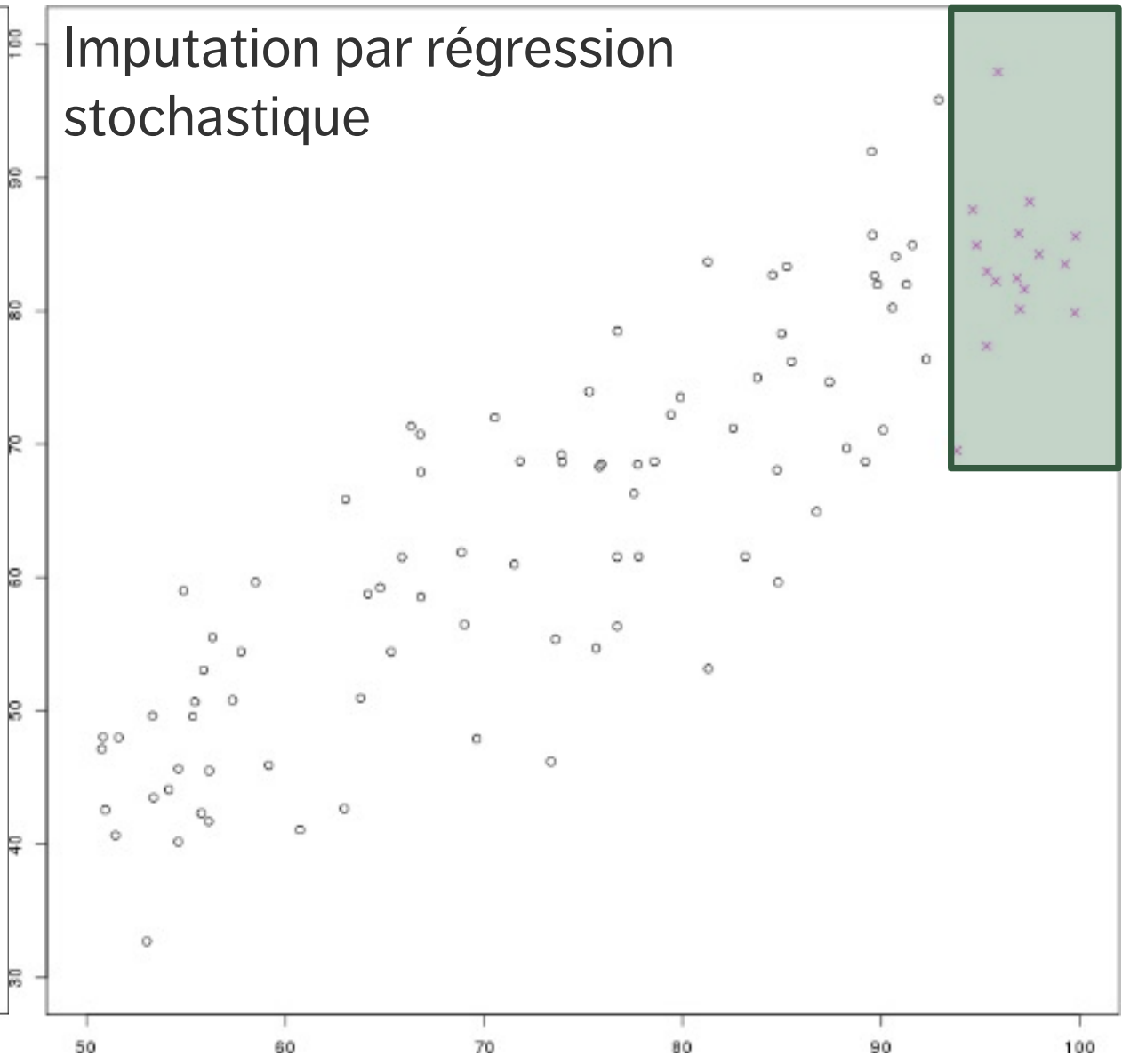
Données artificielles : les valeurs y de tous les points pour lesquels $x > 92$ ont été effacées par erreur.





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





POINTS 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 le mieux lorsque les valeurs manquent complètement au hasard, mais les méthodes d'imputation ont également tendance à produire des estimations biaisées.

Dans le cas d'une 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 n'oubliez pas que il n'y a rien de gratuit !

POINTS DE DONNÉES SPÉCIAUX

Les **observations périphériques** sont des points de données qui sont **atypiques** par rapport aux :

- autres caractéristiques de l'unité (**à l'intérieur de l'unité**)
- mesures sur le terrain pour d'autres unités (**entre les unités**)

ou dans le cadre d'un sous-ensemble collectif d'observations.

Les observations aberrantes sont des observations qui **ne ressemblent pas à d'autres cas** ou qui contredisent des **dépendances** ou des règles connues.

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

POINTS DE DONNÉES SPÉCIAUX

Les points de données **influents** sont des observations dont l'absence entraîne des **résultats d'analyse** sensiblement **différents**.

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

Les valeurs aberrantes **peuvent être** des points de données influents, les points influents **ne sont pas nécessairement** des valeurs aberrantes.

LA DÉTECTION D'ANOMALIES

Les valeurs aberrantes peuvent être anormales pour n'importe quelle variable de l'unité, ou pour une combinaison de variables.

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 de données.

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

Lorsque les anomalies sont associées à des activités malveillantes, elles sont généralement **déguisées**.

LA DÉTECTION D'ANOMALIES

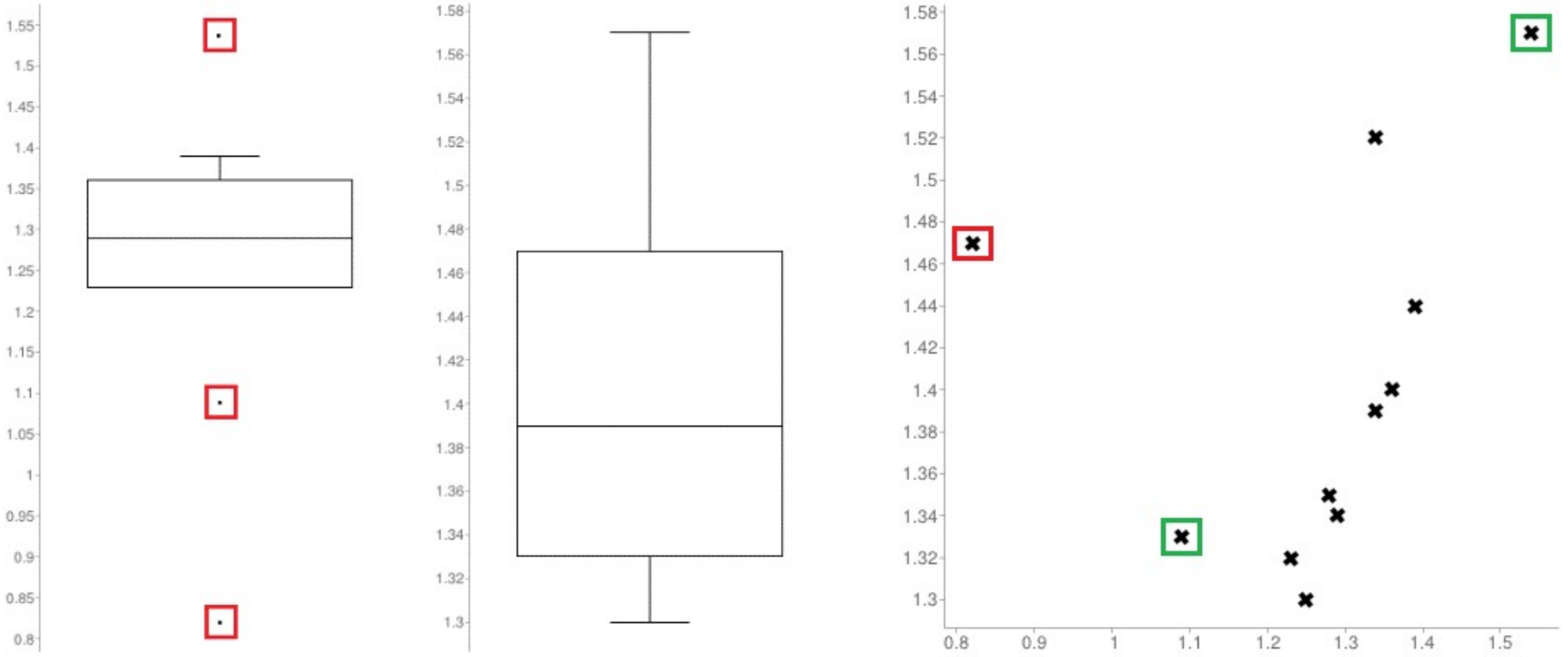
Il existe de nombreuses méthodes pour identifier les observations anormales ; **aucune d'entre elles n'est infaillible** et il faut faire preuve de discernement.

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

- **Observations périphériques**
  diagrammes en boîte, diagrammes de dispersion, matrices de diagrammes de dispersion, distance de Cooke, diagrammes qq normaux
- **Données influentes**
  un certain niveau d'analyse doit être effectué (effet de levier)

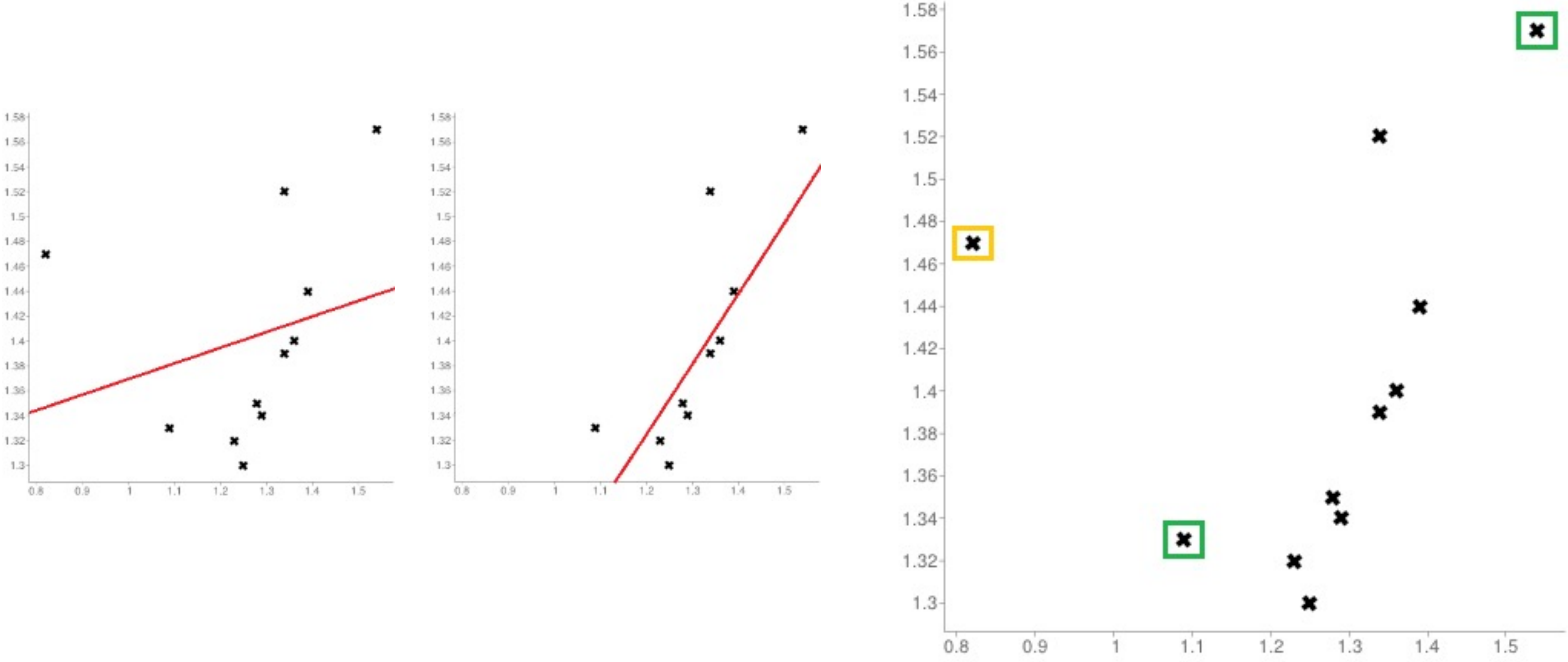
Une fois que les observations anormales ont été supprimées de l'ensemble de données, des unités auparavant « régulières » peuvent devenir anormales.

VALEURS ABERRANTES



Ensemble de données sur les files d'attente : taux de traitement par rapport au taux d'arrivée

OBSERVATIONS INFLUENTES



Ensemble de données sur les files d'attente : taux de traitement par rapport au taux d'arrivée

POINTS À RETENIR

L'identification des points d'influence est un processus itératif car les différentes analyses doivent être exécutées de nombreuses fois.

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

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

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

RÉDUCTION ET TRANSFORMATION DES DONNÉES

TRAITEMENT DES DONNÉES

LA DIMENSIONNALITÉ DES DONNÉES

Dans l'analyse des données, la **dimension** des données est le nombre de variables (ou d'attributs) qui sont rassemblées dans un ensemble de données, représenté par le nombre de colonnes.

Le terme dimension est une extension de l'utilisation du terme pour désigner la taille d'un vecteur.

Nous pouvons considérer les variables utilisées pour décrire chaque objet (ligne) comme un **vecteur** décrivant cet objet.

Note : le terme dimension est utilisé différemment dans les contextes de business intelligence.

Mais plus de données, c'est toujours mieux, non ?

Ça dépend.

DIMENSIONNALITÉ ÉLEVÉE ET BIG DATA

Les ensembles de données peuvent être « grands » de différentes manières :

- trop grandes pour que le **matériel** puisse les gérer (ne peuvent être stockées ou consultées correctement en raison du nombre d'observations, du nombre de caractéristiques ou de la taille globale).
- la taille peut aller à l'encontre des **hypothèses de modélisation** (nombre de caractéristiques > nombre d'observations).

Exemples :

- De multiples capteurs enregistrant 100+ observations par seconde dans une vaste zone géographique sur une longue période = **très gros ensemble de données**.
- Dans la matrice terme-document d'un corpus (col = termes, lignes = documents), le nombre de termes est généralement beaucoup plus élevé que le nombre de documents, ce qui conduit à des données **excessivement éparses**.

LA MALÉDICTION DE LA DIMENSIONNALITÉ

À moins que la taille de l'ensemble de données ne croisse de façon exponentielle avec sa dimension, les performances de tout modèle que nous construisons risquent de souffrir de la **malédiction de la dimensionnalité**.

Solutions possibles :

- **observations d'échantillonnage**
- **sélection des caractéristiques** (facile) et/ou réduction des dimensions (difficile).

Nous cherchons des moyens de préserver le signal tout en réduisant la dimension : il est plus facile de trouver des aiguilles dans de petites bottes de foin !

(Il s'agit en fait d'un problème difficile... mais nous éviterons les détails techniques dans ce cours).

ÉCHANTILLONNAGE D'OBSERVATIONS

Question : chaque observation (ligne de l'ensemble de données) doit-elle être utilisée ?

Si les lignes sont sélectionnées au hasard, 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 a lieu plus tard, l'échantillonnage affectera nécessairement les chiffres (passagers vs. vols)
- même des opérations simples sur un grand fichier (trouver le nombre de lignes, par exemple) peuvent être lourdes en termes de mémoire et de temps de calcul – **des informations préalables sur la structure de l'ensemble de données peuvent aider**.

SÉLECTION DES CARACTÉRISTIQUES

La suppression des variables non pertinentes ou redondantes est une tâche commune du traitement des données.

Motivations :

- les outils de modélisation ne les gèrent pas bien ces tâches (inflation de la variance due à la multi-colinéarité, etc.)
- réduction de la dimension (nombre de variables > nombre d'observations)

Approches :

- filtre vs. emballage (« wrapper »)
- non supervisé vs. supervisé

MÉTHODES DE SÉLECTION DES CARACTÉRISTIQUES

Les **méthodes de filtrage** inspectent chaque variable individuellement et les évaluent en fonction d'une certaine **métrique d'importance**.

Les caractéristiques les moins pertinentes (c'est-à-dire dont le score d'importance est inférieur à un certain seuil) sont ensuite supprimées.

Les **méthodes enveloppantes** recherchent des sous-ensembles de caractéristiques pour lesquels le critère d'évaluation utilisé par la méthode analytique éventuelle est "optimisé".

Le processus est **itératif**, et généralement intensif en termes de calcul : les sous-ensembles candidats sont utilisés dans l'analyse jusqu'à ce que l'on obtienne une métrique d'évaluation acceptable pour l'analyse.

MÉTHODES DE SÉLECTION DES CARACTÉRISTIQUES

Les méthodes non supervisées déterminent l'importance d'une caractéristique en se basant uniquement sur ses valeurs.

Les méthodes supervisées évaluent l'importance de chaque caractéristique en étudiant sa relation avec une caractéristique cible (corrélation, etc.).

Les méthodes enveloppantes sont généralement supervisées.

Méthodes de filtrage non supervisées : suppression des variables constantes, des variables de type ID (différentes sur toutes les observations), des caractéristiques à faible variabilité, etc.

DISCRÉTISATION

Pour réduire la complexité du calcul, il peut être nécessaire de remplacer une variable numérique par une variable **ordinale** (de la valeur de la taille à *petit*, *moyen*, *grand*, par exemple).

L'expertise du domaine peut être utilisée pour déterminer la taille des groupes (*bins*), bien que cela puisse introduire un biais inconscient dans les analyses.

En l'absence d'une telle expertise, les limites peuvent être fixées de sorte que soit

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

QUALITÉ DES DONNÉES ET VALIDATION DES DONNÉES

TRAITEMENT DES DONNÉES

Martin : Les données, c'est le bordel...
Allison : Même quand elles ont été nettoyées ?
Martin : Surtout quand elles ont été nettoyées.

P. Boily, *Introduction au conseil quantitatif*

DONNÉES FIABLES

L'ensemble de données idéal aura le moins de problèmes possible avec :

- **Validité** : type de données, plage de données, réponse obligatoire, unicité, valeur, expressions régulières.
- **Intégralité :** observations manquantes
- **Exactitude et précision :** liées aux erreurs de mesure et/ou de saisie des données ; diagrammes cibles (exactitude en tant que biais, précision en tant qu'erreur standard).
- **Cohérence :** observations contradictoires
- **Uniformité :** les unités sont-elles utilisées de manière uniforme dans les ensembles de données ?

Vérifier les problèmes de qualité des données à un stade précoce peut éviter des difficultés plus tard dans l'analyse.

DONNÉES FIABLES

exact et précis

précis, mais inexact

exact, mais imprécis

ni exact, ni précis

SOURCES COMMUNES D'ERREUR

Lorsque vous traitez des ensembles de données **hérités** ou **combinés** (c'est-à-dire des ensembles de données sur lesquels vous avez peu de contrôle) :

- un code est attribué aux données manquantes
- un code est attribué à « N/A / vide »
- erreur de saisie des données
- erreur de codage
- erreur de mesure
- entrées en double
- entassement

DÉTECTER LES ENTRÉES NON VALIDES

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

- **Statistiques descriptives univariées**
  compte, étendue, z-score, moyenne, médiane, écart type, contrôle logique
- **Statistiques descriptives multivariées**
  table n-way, contrôle logique
- **Visualisation des données**
  nuage de points, matrice de nuage de points, histogramme, histogramme conjoint, etc.

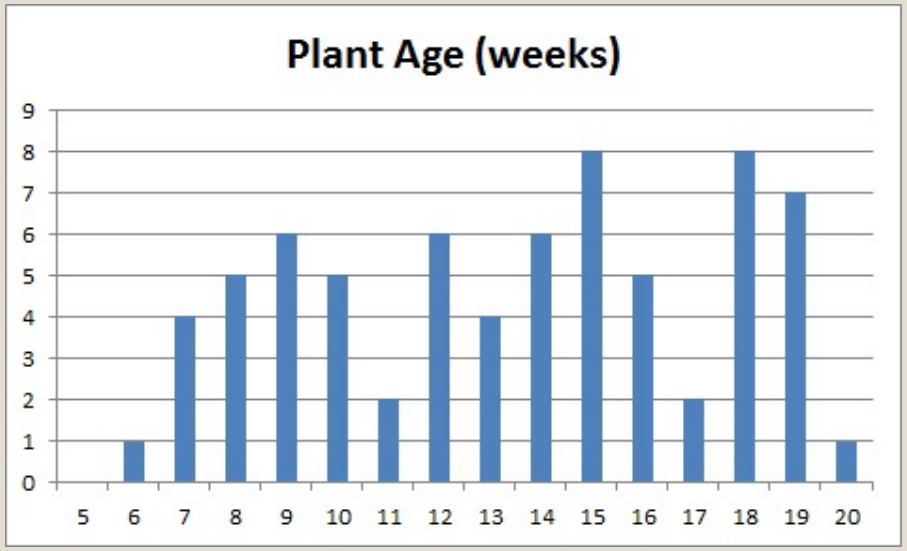
Cette étape pourrait permettre d'identifier les valeurs aberrantes potentielles.

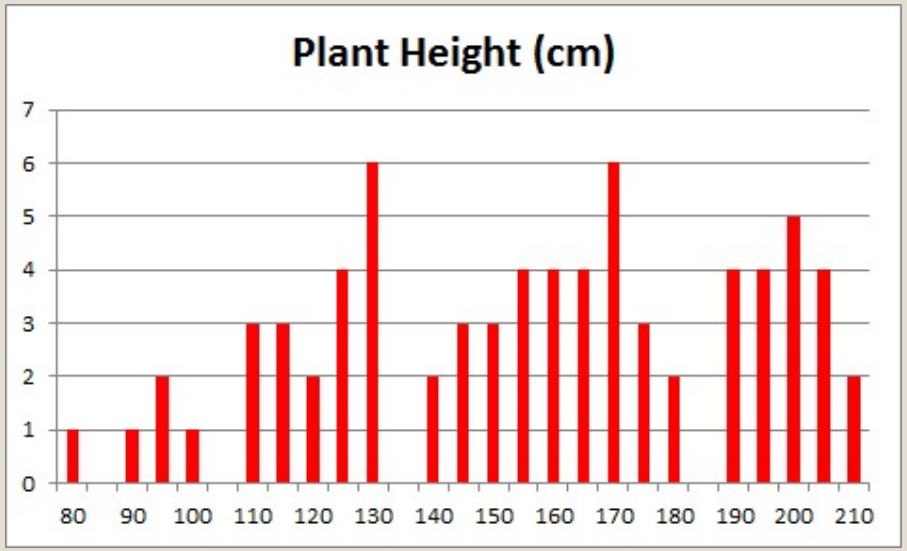
Le fait de ne pas détecter d’entrées non valides **ne signifie pas** que toutes les entrées sont valides.

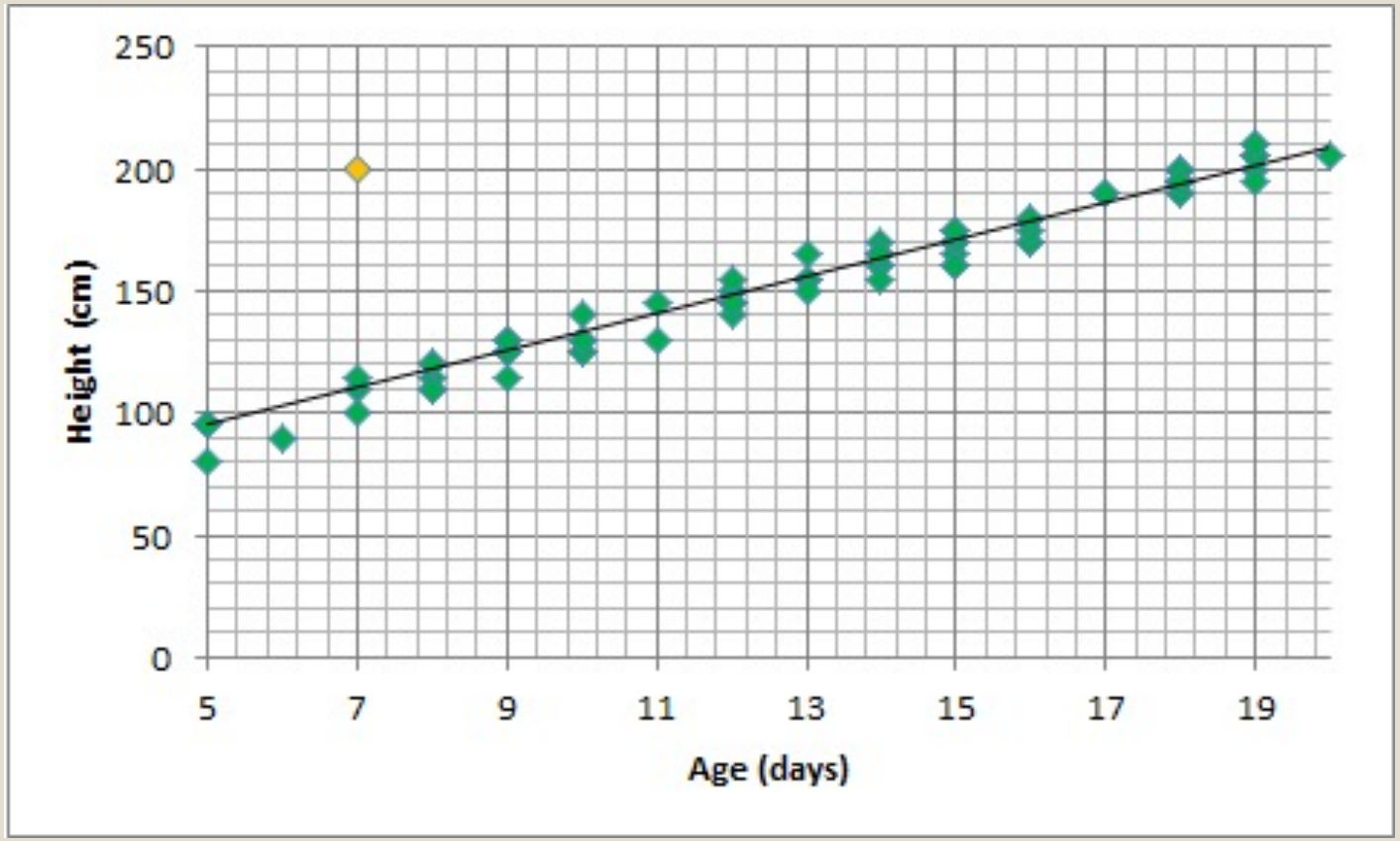
Un petit nombre d'entrées non valides sont recodées comme manquantes.

[Dossier personnel]

ILLUSTRATION

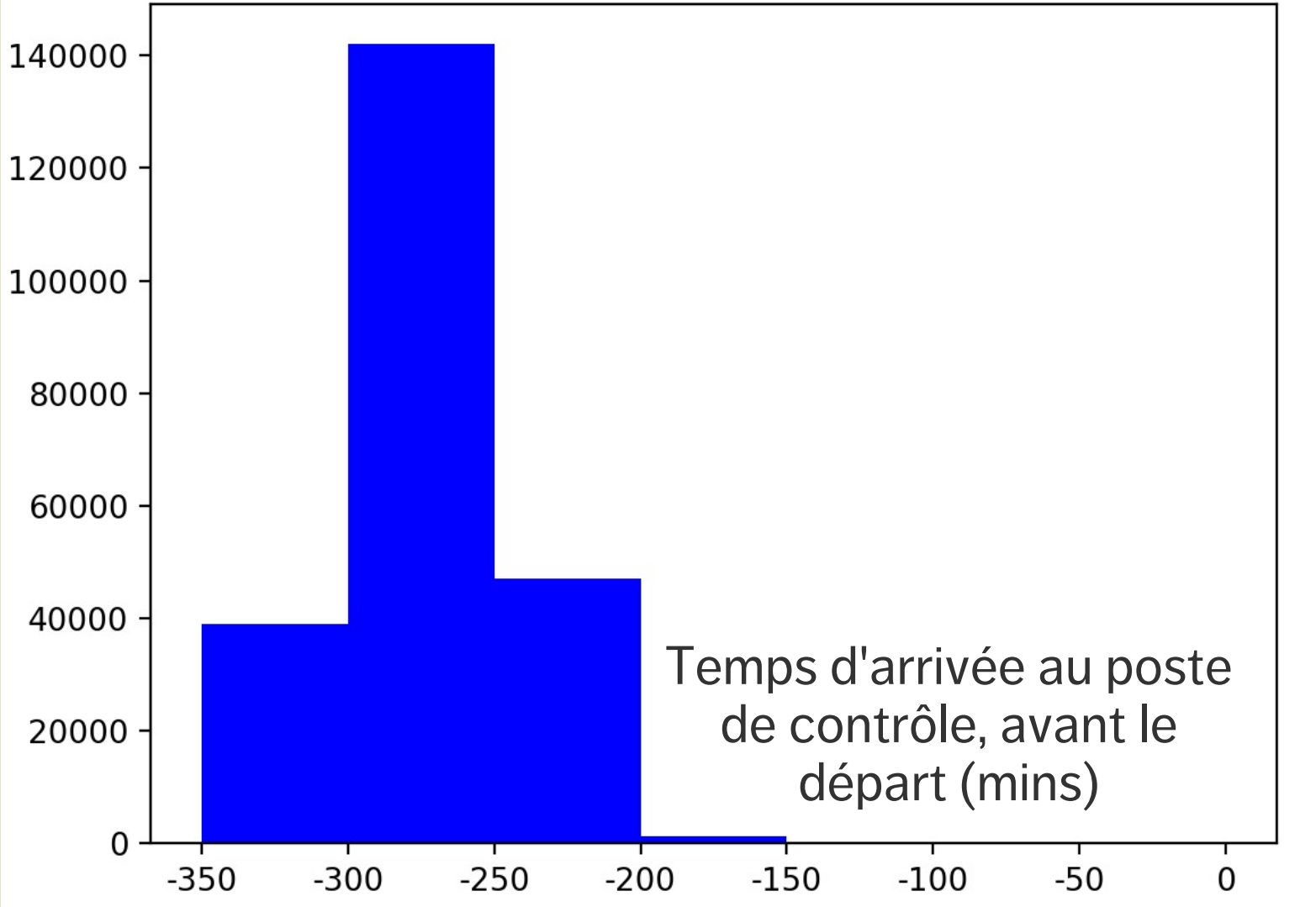






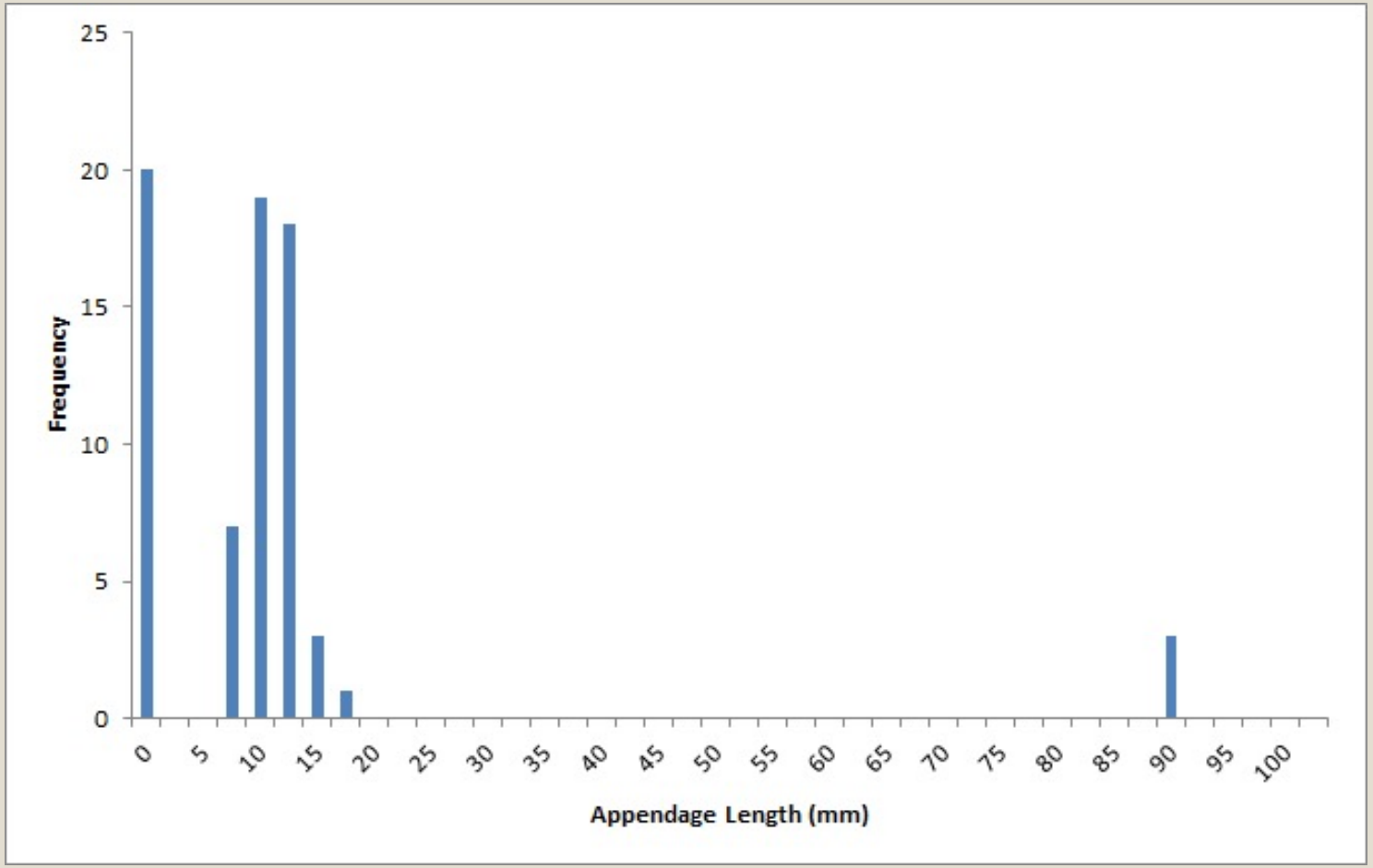
[Dossier personnel]

ILLUSTRATION



ILLUSTRATION

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



POINTS À RETENIR

N'attendez pas que l'analyse soit terminée pour découvrir qu'il y avait un problème de qualité des données.

Les tests univariés ne révèlent pas toujours toute l'histoire.

Les visualisations peuvent aider.

Le contexte est crucial – vous pouvez avoir besoin de plus de contexte sur les données pour les comprendre... mais quelle que soit la situation, vous devez comprendre la qualité de l'ensemble de données.