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

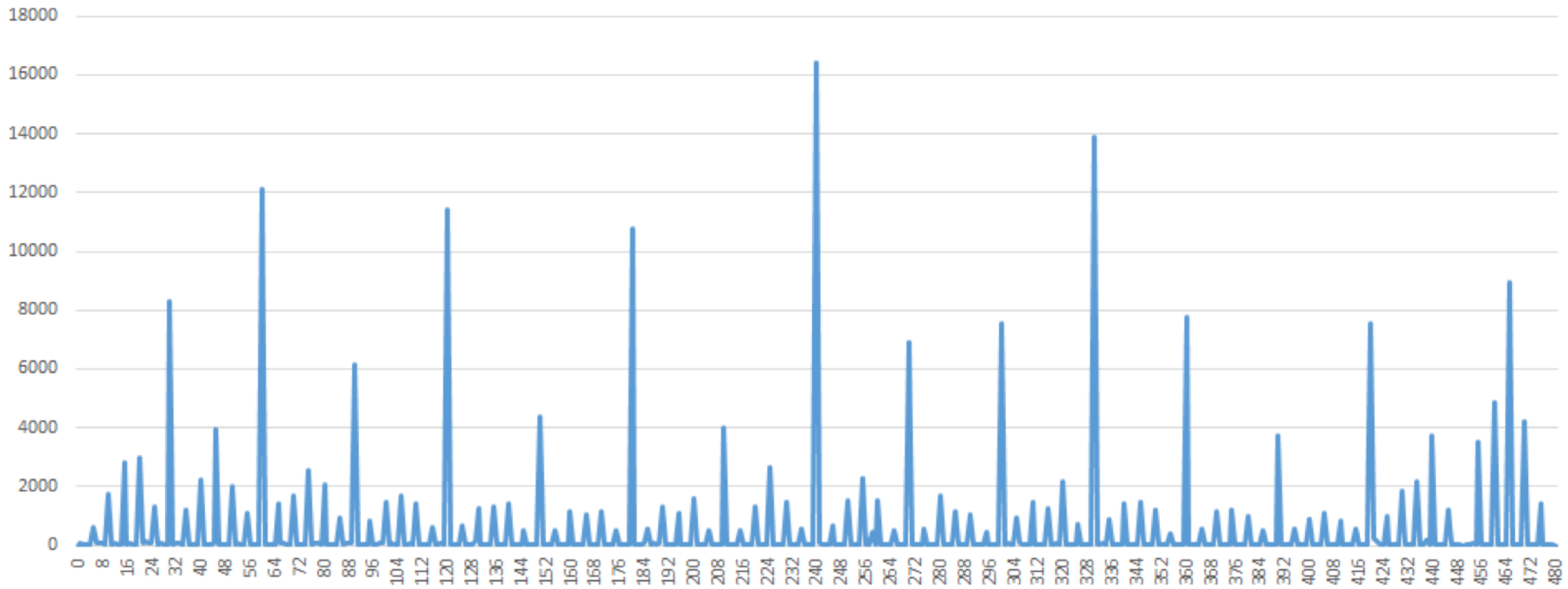
Les tests univariés ne montrent pas toujours **tout ce qui se passe**.

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

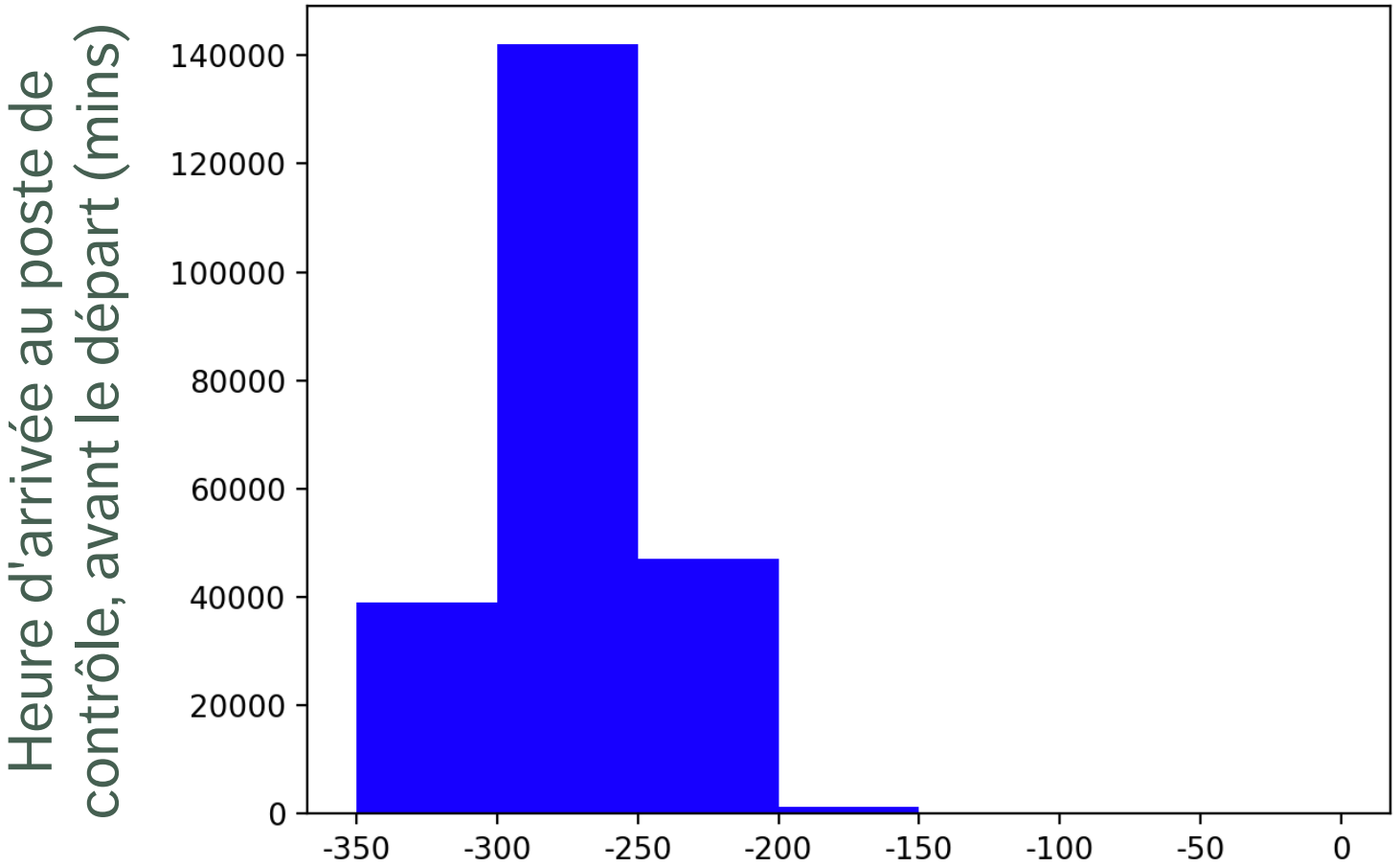
Défaut de détection des entrées non valides ≠ toutes les entrées sont valides.

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

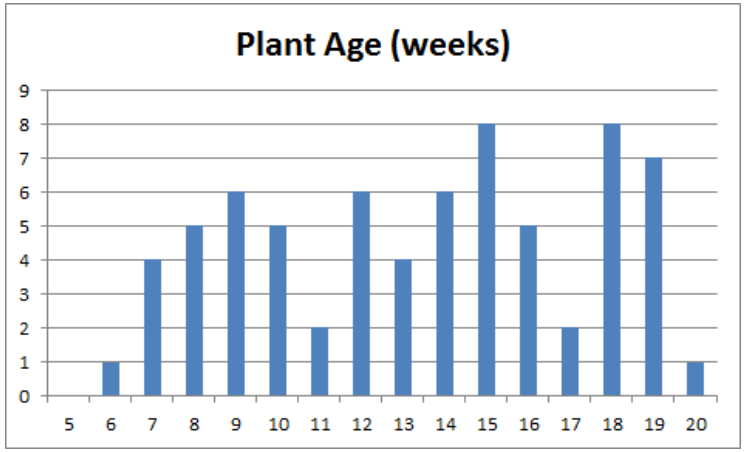
Sex	
Male	19
Female	17
(blank)	2
Total	**38**

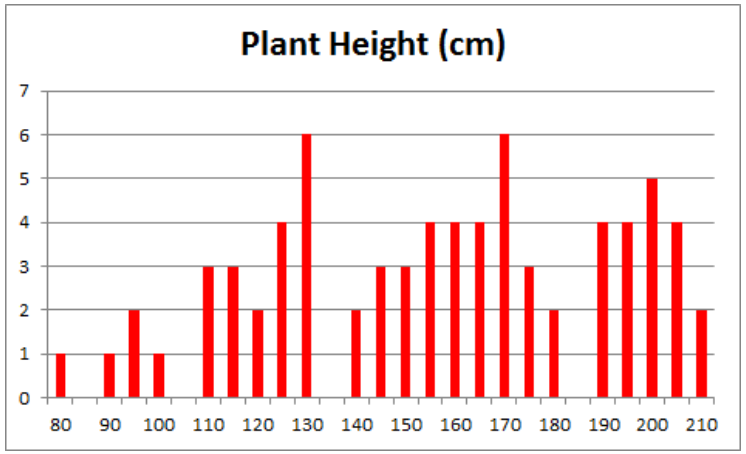
Pregnant	
Yes	7
No	27
99	1
(blank)	3
Total	**38**

vs.

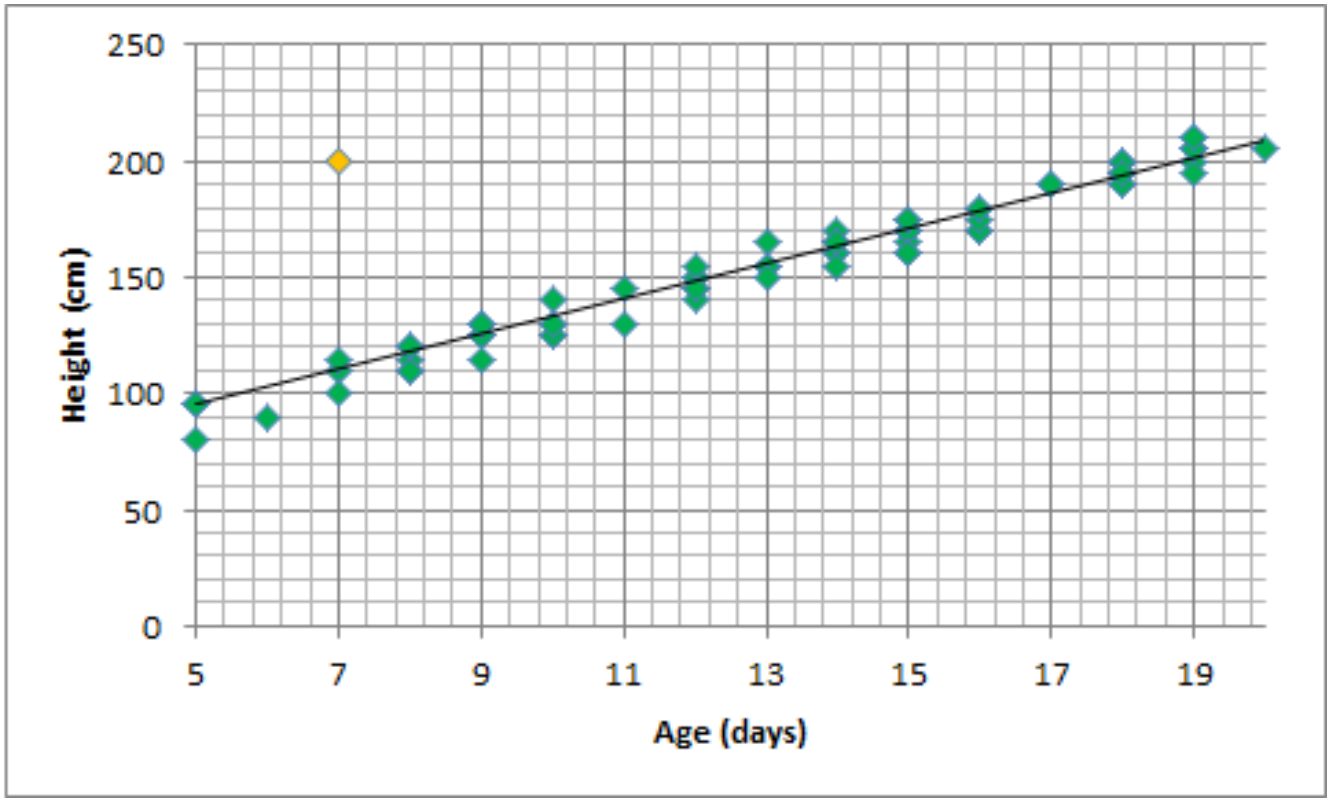
	Pregnant				Total
	Yes	No	99	(blank)	
Male	1	17	1	0	**19**
Female	6	9	0	2	**17**
(blank)	0	1	0	1	**2**
Total	**7**	**27**	**1**	**3**	**38**

La détection d'entrées non valides



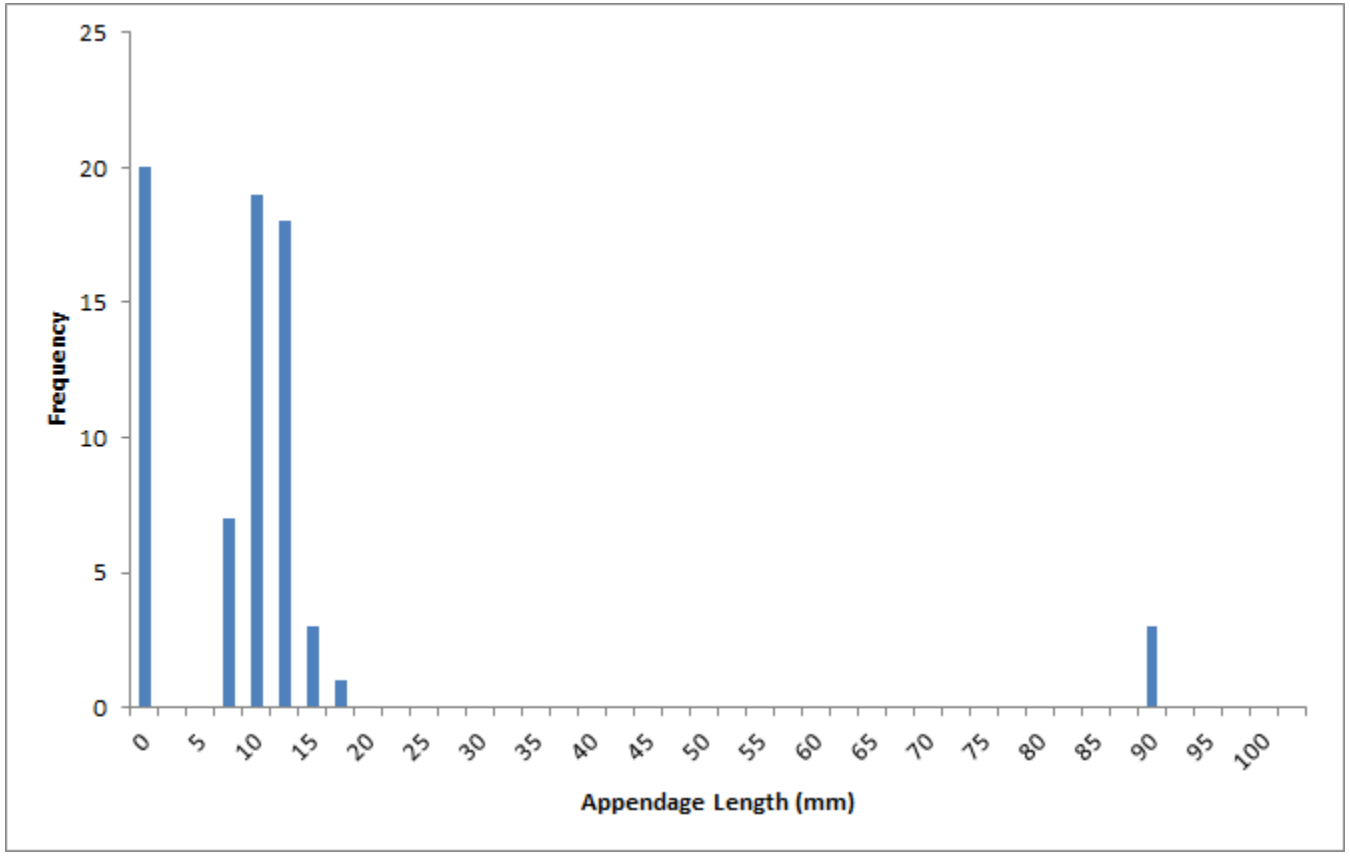


vs.



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 suggérées

La qualité et le traitement des données

Data Understanding, Data Analysis, Data Science
Data Preparation

Introduction

General Principles

- Approaches to Data Cleaning
- Pros and Cons
- Tools and Methods

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).