Introduction à la science des données

Instructeur: Patrick Boily

uOttawa

Institut de développement professionnel
Professional Development Institute

TRAITEMENT DES DONNÉES

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

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.

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

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

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 !

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

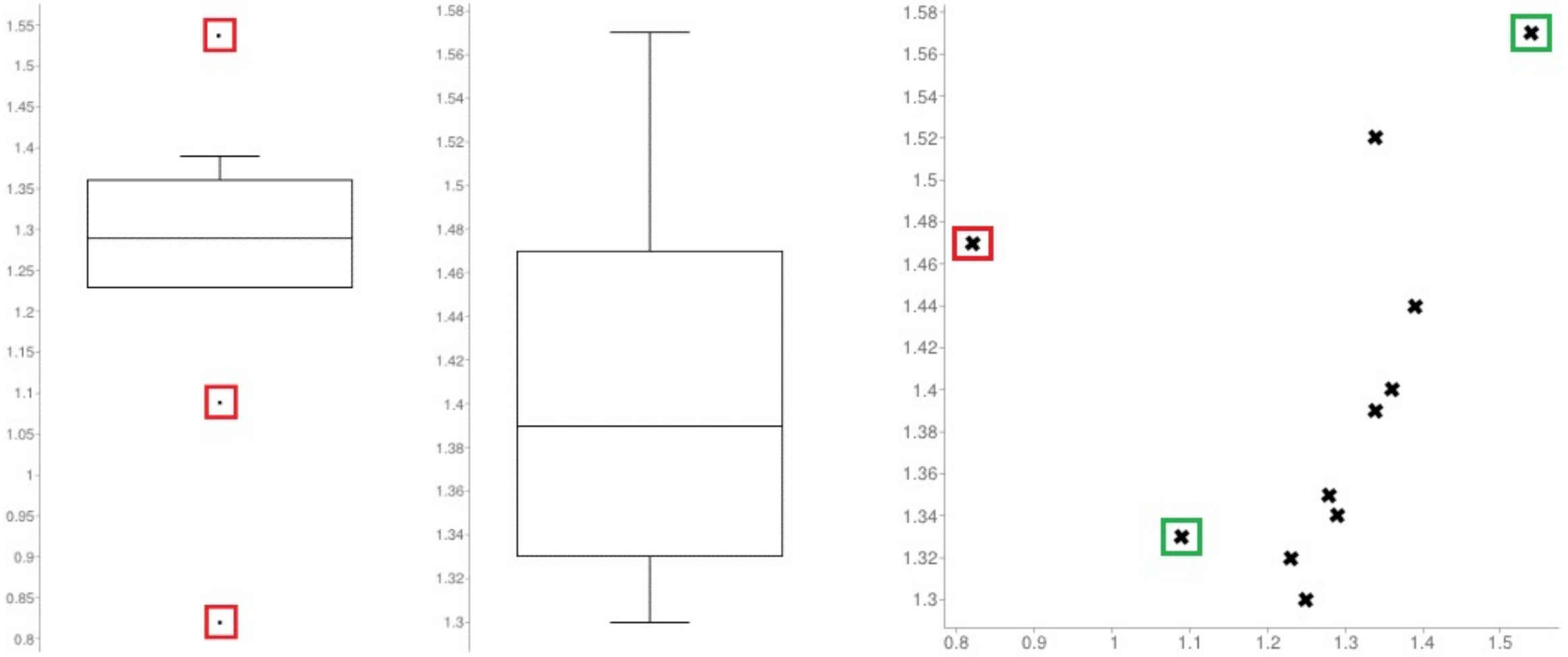
Il existe de nombreuses méthodes pour identifier les observations anormales ; **aucune d'entre elles n'est infaillible** et il faut faire preuve de discernement.

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)

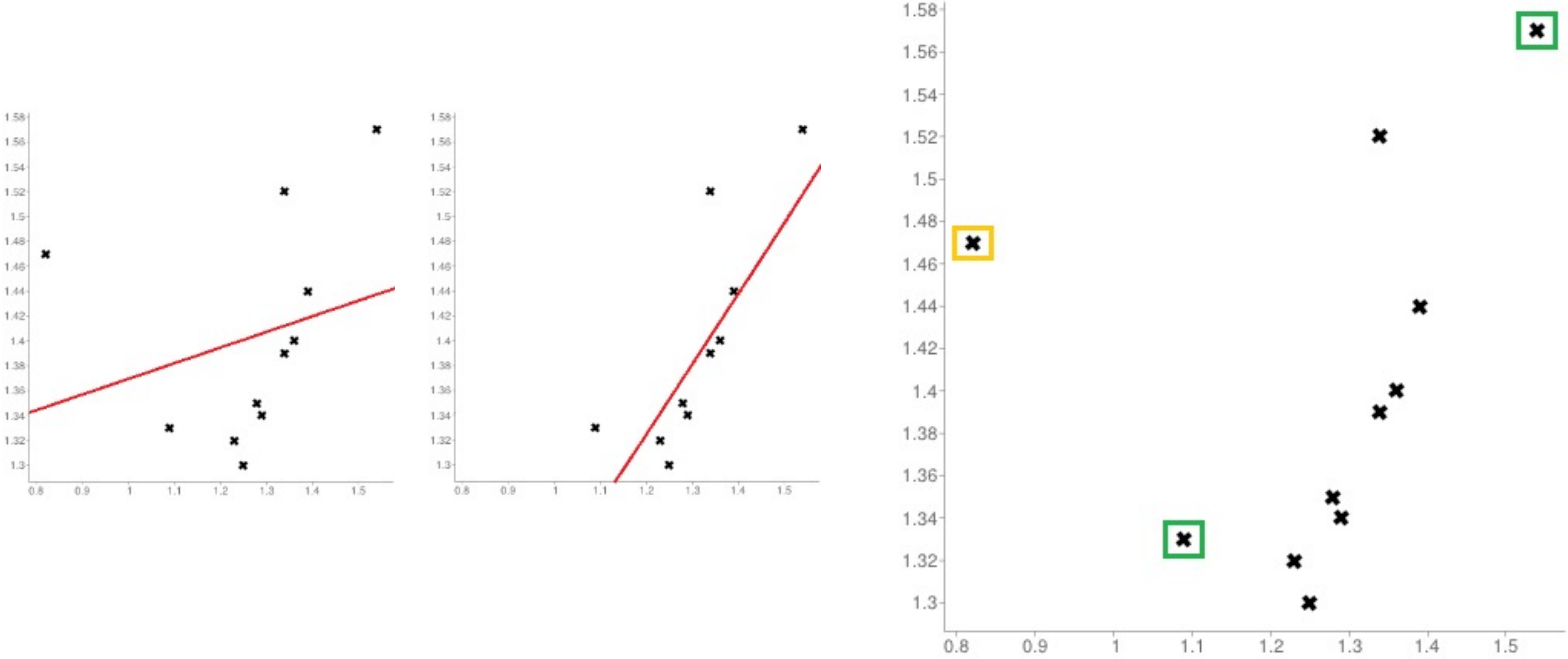
Une fois que les observations anormales ont été supprimées de l'ensemble de données, des unités auparavant « régulières » peuvent devenir anormales.

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.

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.

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).

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

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

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

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.