Module 7 : Activités post-collecte
Vidéo 2 : Synthèse, visualisation et édition

Sharan Sharma

THE WORLD BANK

MANNHEIM
BUSINESS SCHOOL

Les activités sont nombreuses...

1) Encodage des réponses ouvertes

2) Préparation des données

3) Synthèse et visualisation

4) Édition des données

5) Imputation et pondération

6) Contrôle de la divulgation

7) Traitement final, documentation et diffusion

1. Données manquantes

- Bien qu'il y ait une surveillance pendant la collecte des données (→ Module 6, Vidéo 4), une analyse est aussi nécessaire à cette étape.

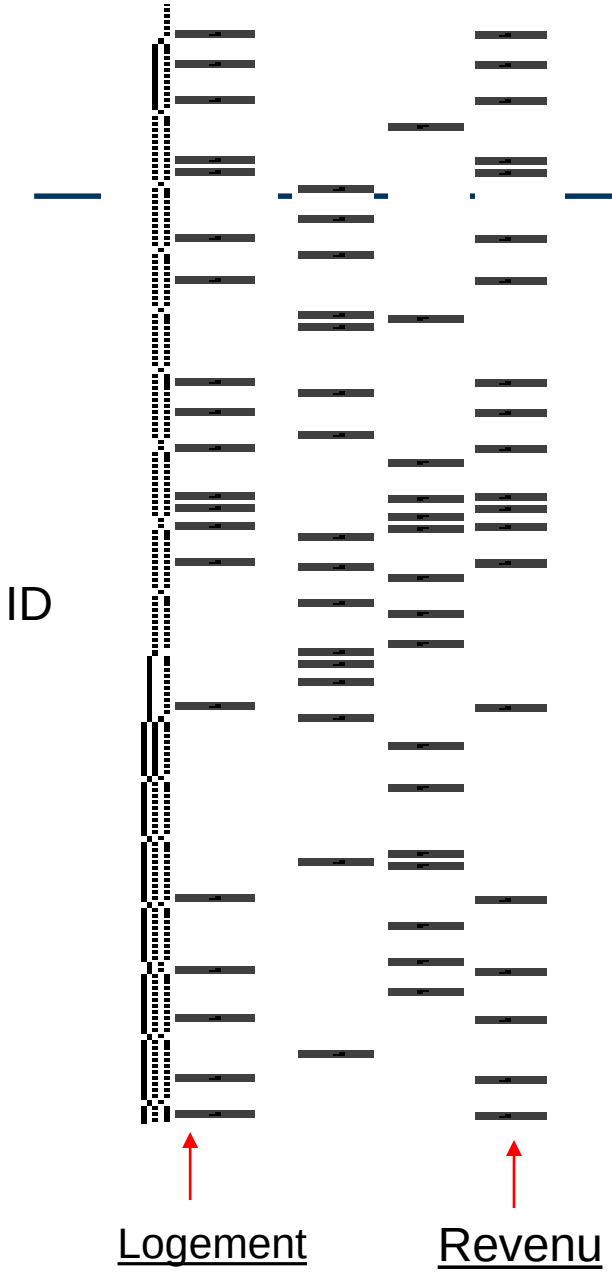
- Les données sont-elles manquantes en raison des schémas de saut (NA) ? Ou parce que le répondant ne sait pas ou refuse de répondre ? Ou en raison d'un autre problème ?
  - Il est intéressant d'analyser ces facteurs séparément

- Commencez par calculer le pourcentage de données manquantes pour chaque variable.
  - Concentrez-vous sur les variables dont le taux de données manquantes est > ___ %.
  - Comparez avec les vagues précédentes dans le cas d'une enquête en panel, ou avec des enquêtes transversales récurrentes.

Données manquantes...

- Examinez également les *schémas* de données manquantes. Certains problèmes sont-ils apparents ?
  - Par ex, les données manquantes font toutes parties de l'unité d'échantillonnage primaire PSU 10
  - Y a-t-il des données manquantes où il ne devrait pas en avoir ? Par ex. une question qui devrait être absolument posée au répondant (à ne pas sauter) ayant pour réponse NA.

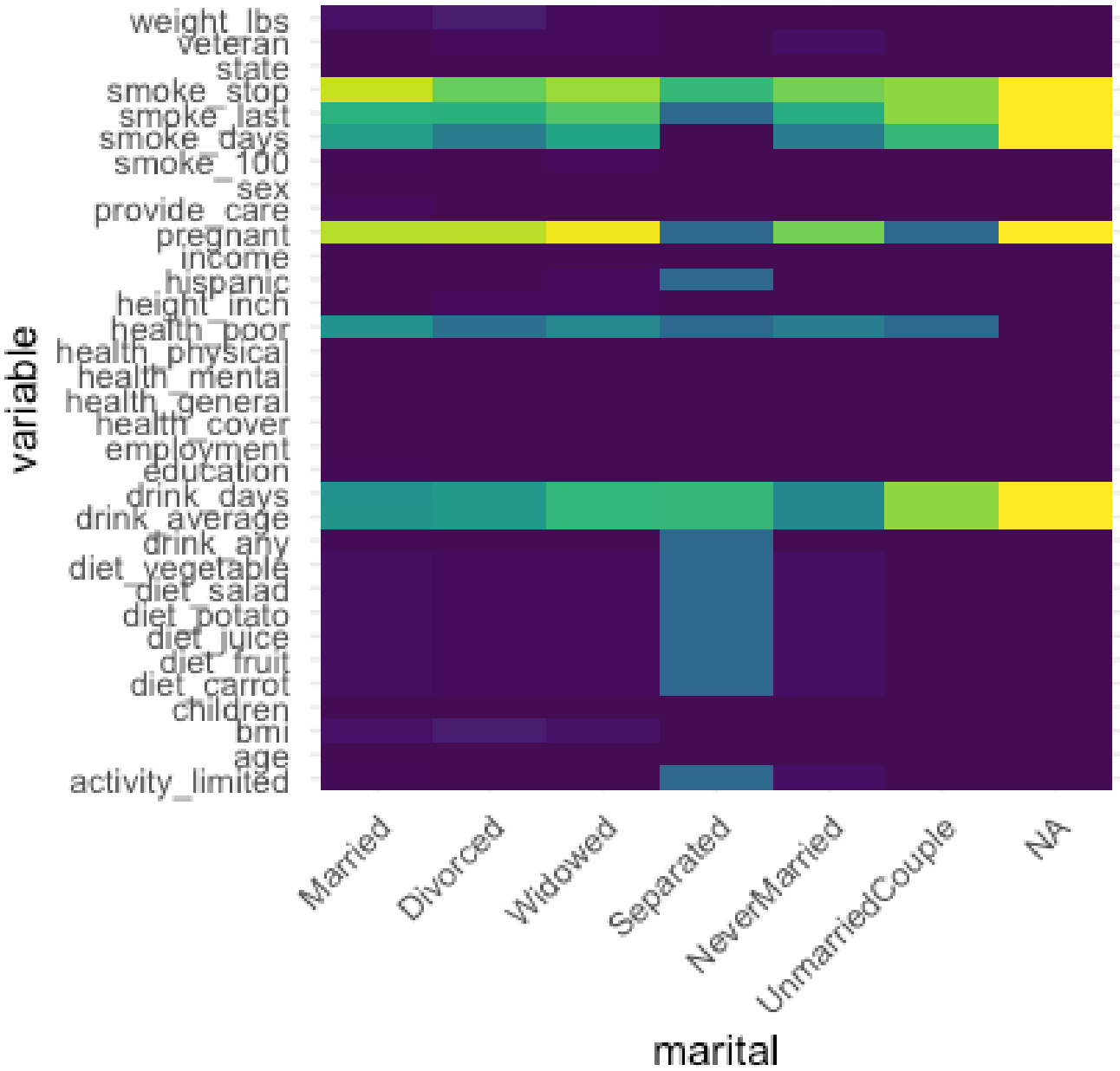
- Visualisez les données manquantes



- Corrélation entre les cas qui présentent des données manquantes pour les variables de logement et de revenu.
  - Certains ID présentent plus de données manquantes. Quelles sont leurs caractéristiques ?

- Cette visualisation peut être difficile avec des centaines de variables ; il vaut mieux analyser certains groupes de variables sensibles.

- Nombreuses visualisations possibles :



Voir https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html pour d'autres exemples utiles.

2) Avoir un aperçu des données agrégées

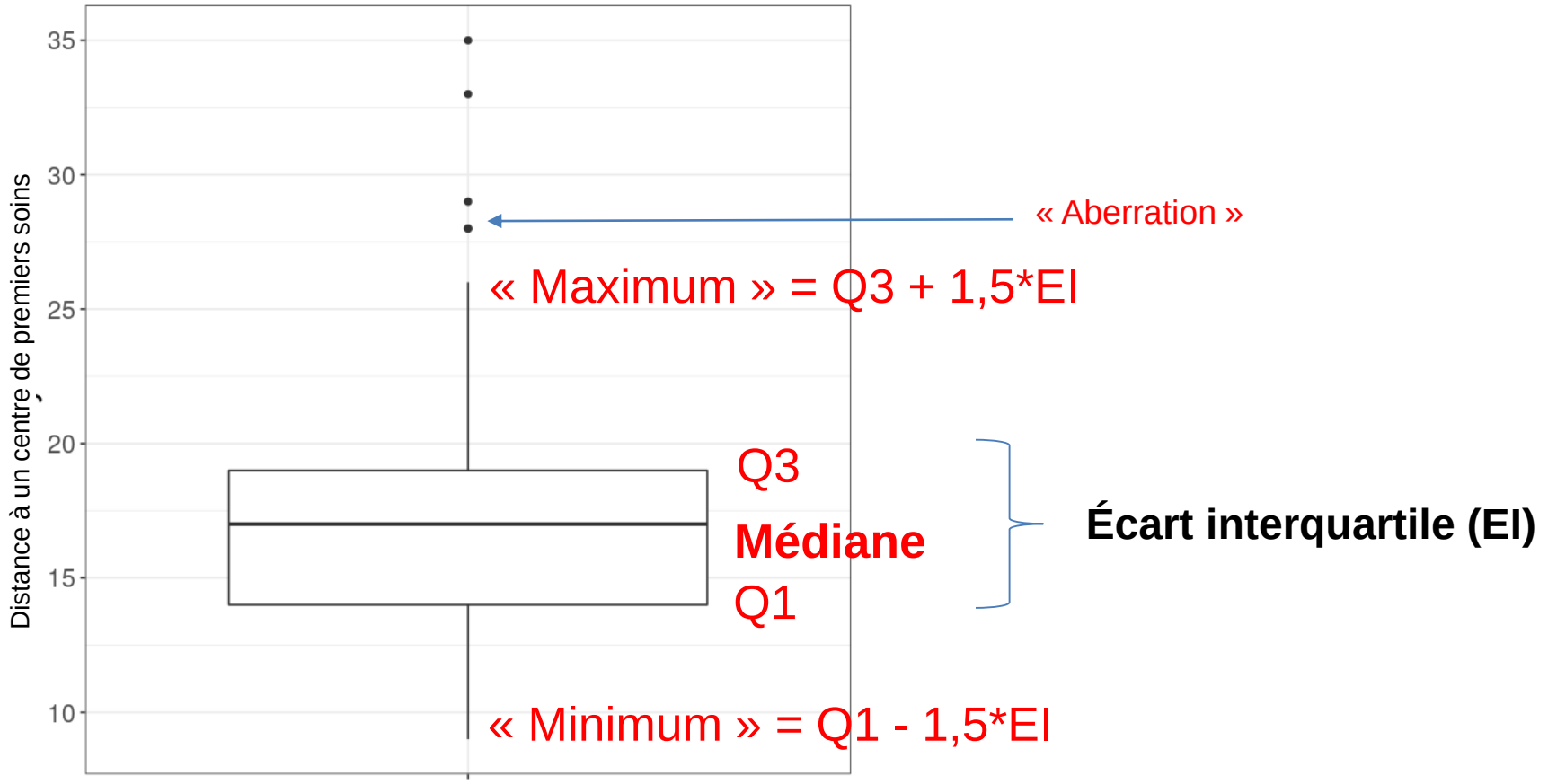
Tableaux de synthèse : synthèse en cinq nombres

	Distance à un centre de premiers soins
Minimum	9
1er quartile (Q1)	14
Médiane	17
3e quartile (Q3)	19
Max.	35
Moyenne	16.9

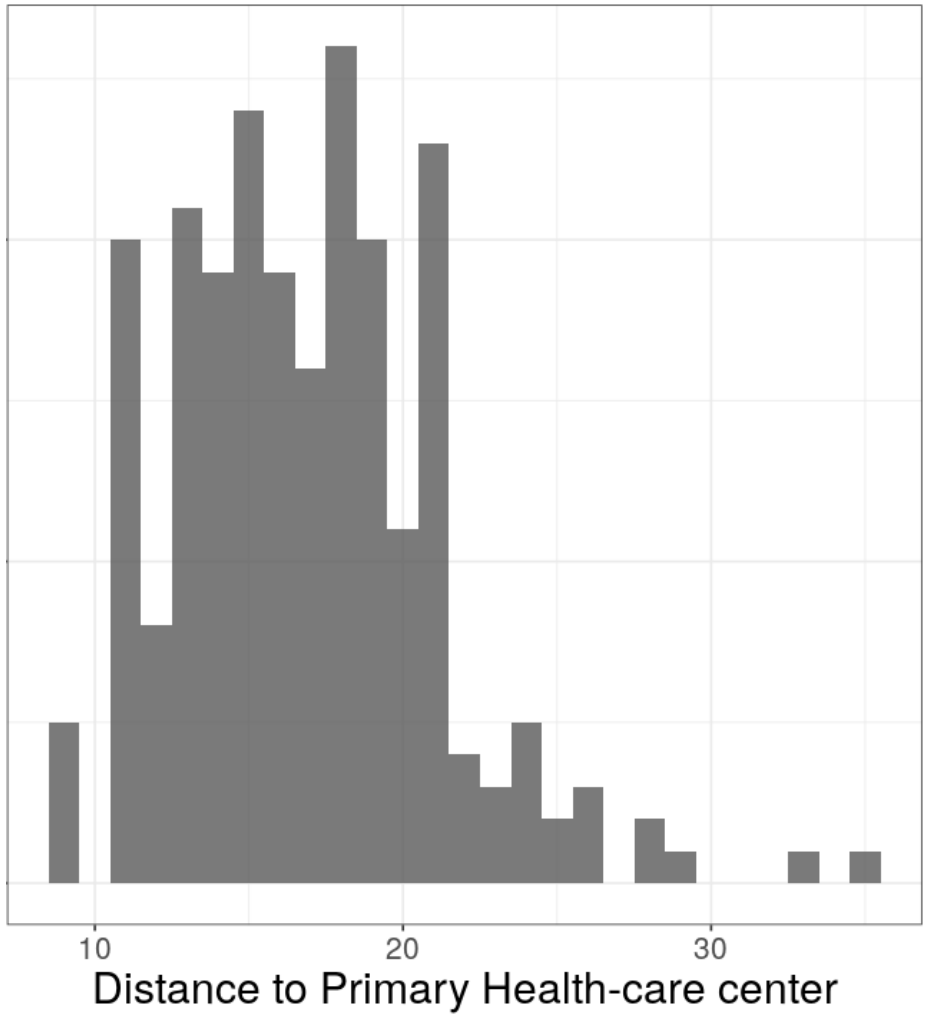
- Ces données sont-elles cohérentes ?
- Comment sont-elles par rapport aux vagues précédentes de l'enquête ?
- Comparaison avec d'autres données de référence ?

Données adaptées du jeu de données « mpg » disponible dans le paquet ggplot2 dans R. Les variables ont été réétiquetées pour les besoins de cette présentation (« cty », « hwy »). n = 234.

Représentations unidimensionnelles : boîte à moustaches

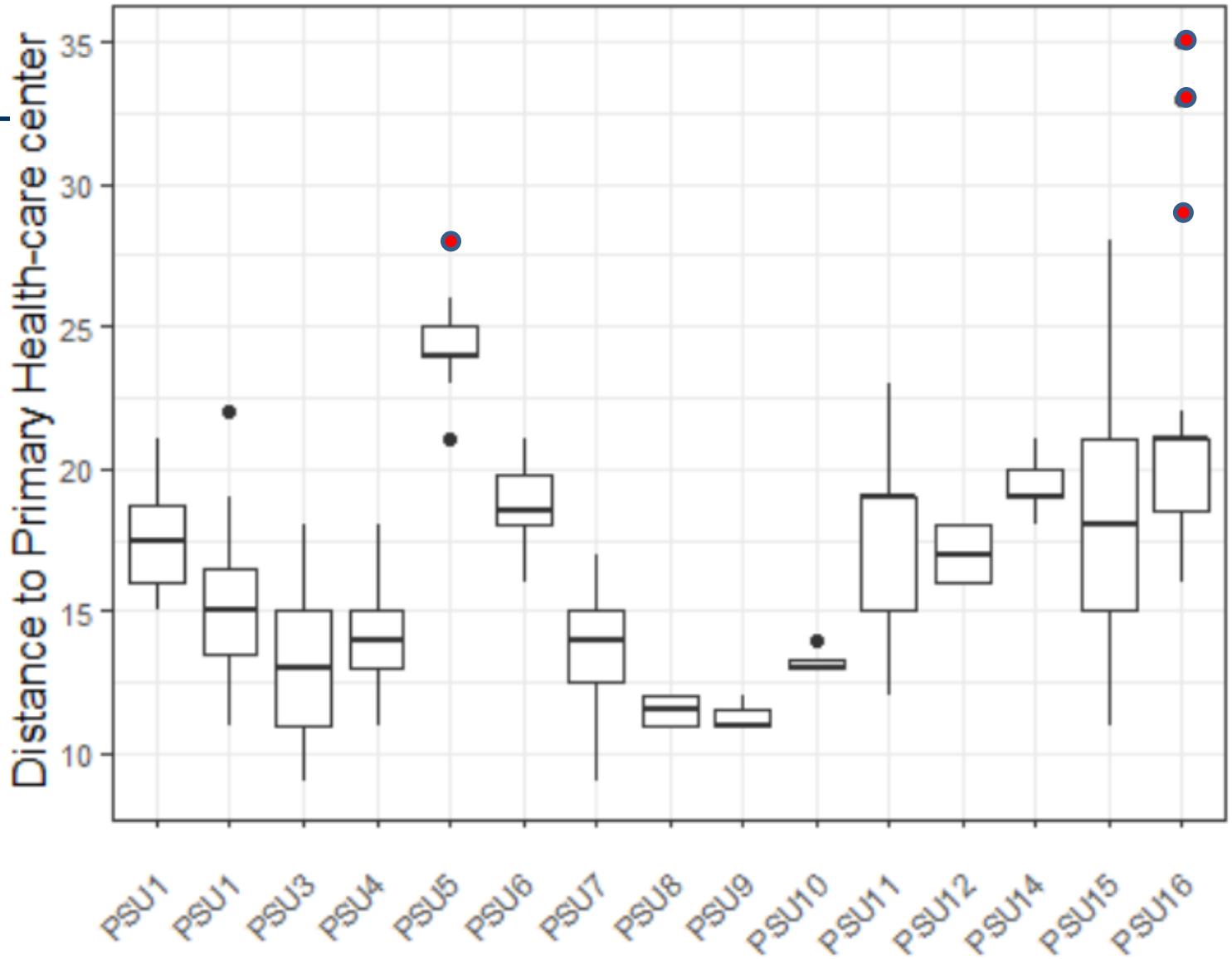


Représentations unidimensionnelles : Histogrammes ou courbes de densité



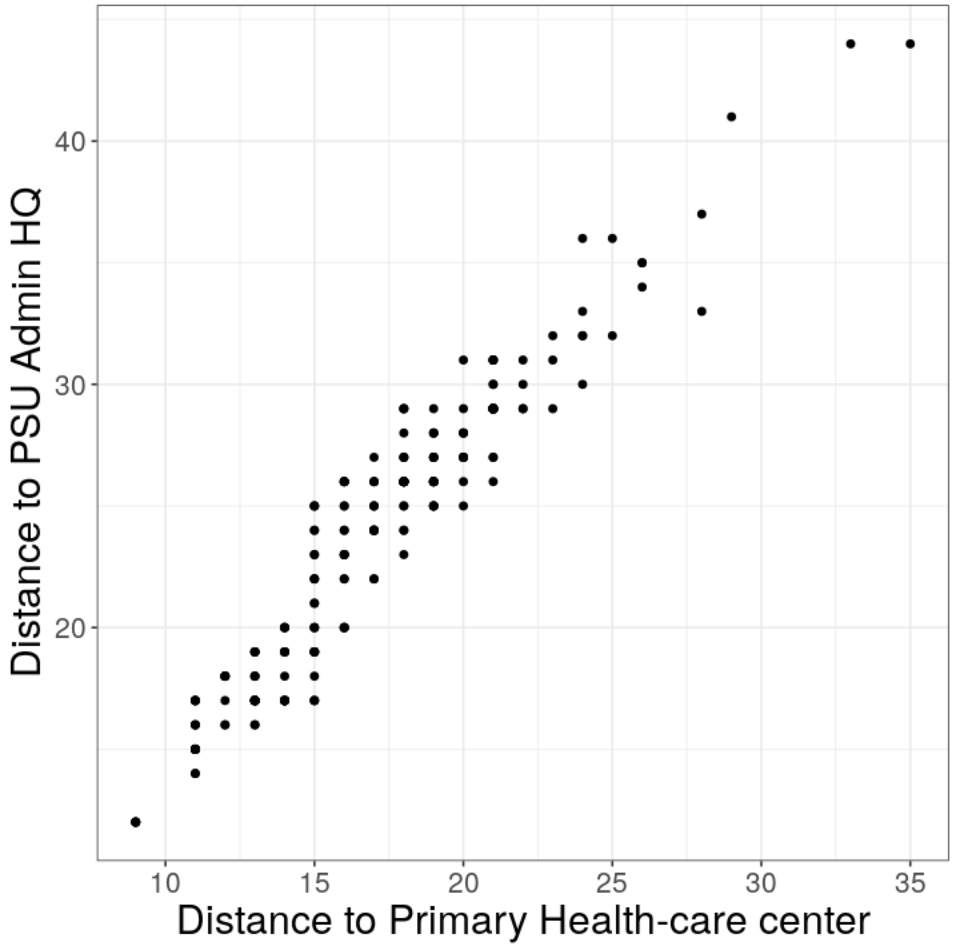
Si vous souhaitez également voir la forme de la répartition

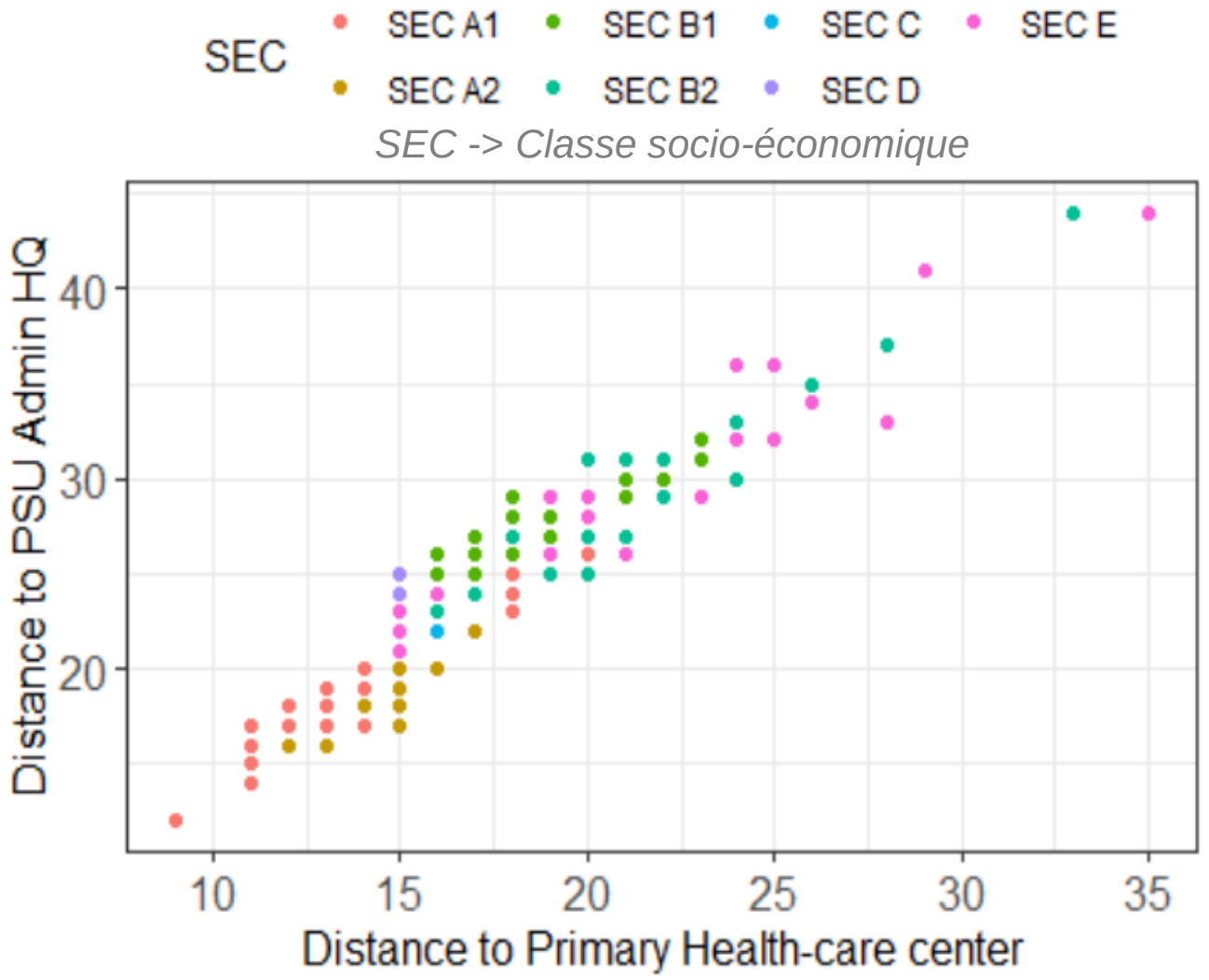
D’où viennent ces valeurs très élevées ?

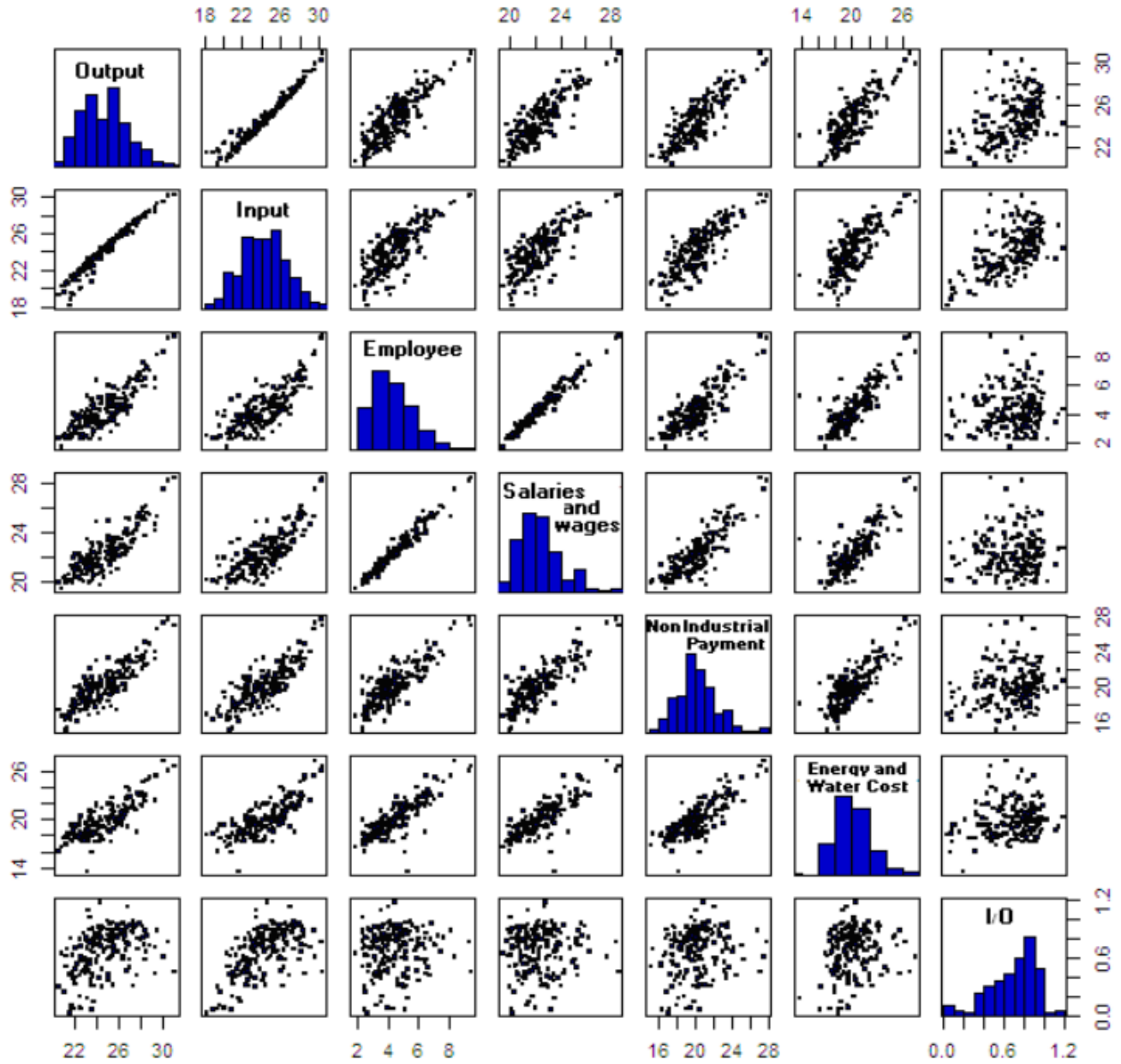

Distance to Primary Health-care center
35
30
25
20
15
10
PSU1
PSU1
PSU3
PSU4
PSU5
PSU6
PSU7
PSU8
PSU9
PSU10
PSU11
PSU12
PSU14
PSU15
PSU16

Représentations bidimensionnelles : nuage de points

La couleur peut apporter de l'information !







Représentation conjointe de répartitions uni- et bidimensionnelles

Source : Ghahroodi et al 2015

Logiciels de visualisation

- Les logiciels de statistiques standard (SAS, STATA, SPSS, par ex.) peuvent produire ce type de résultat.

- Les logiciels libres comme R ont également beaucoup apporté, par ex. le paquet ggplot2.

- On peut également avoir recours à des visualisation Web interactives. Les graphiques des diapos précédentes ont été réalisés à l'aide de : https://shiny.gmw.rug.nl/ggplotgui/

Pour les variables catégorielles

- Les variables catégorielles sont souvent employées pour diviser les bases de données à des fins d'analyse. Mais il est aussi nécessaire de les analyser indépendamment.

- Synthétisez/représentez une répartition par catégorie (histogrammes). Comparez le résultat avec les vagues précédentes ou des données externes, si possible. Y a-t-il des éléments surprenants ? Par ex. le pourcentage des diplômés du supérieur > pourcentage des sans diplôme.

- Des tableaux de contingence sont employés avec plusieurs variables catégorielles
  - classes socio-économiques x catégories de revenu
  - Y a-t-il des combinaisons incohérentes ? par ex. des cas qui sont à la fois « plus haut niveau de diplôme » = licence, et « études en cours » = études secondaires supérieures (cycle court).

Vérifications de cohérence

- Certaines vérifications sont automatiquement intégrées dans les instruments informatisés.
  - Vérifications des plages, par ex. enfants mineurs < 18 ans
  - Type de réponse, par ex. le nombre d'enfants doit être un nombre entier, une question ouverte sur la profession doit contenir des valeurs texte, etc.
  - Vérifications logiques (par ex. nombre d'années de mariage < âge)

- Mais il est impossible d'intégrer toutes les vérifications possibles. Il est indispensable d'effectuer des vérifications de cohérence après la collecte des données.

Table 10.1: Range restriction rules for inconsistent and extreme values in the student file

Sequence	Description	SAS Code
1	Invalidate if number for an individual's weight is negative.	if (WB151Q01HA < 0) then WB151Q01HA=.I;
2	Invalidate if number for an individual's height is negative.	if (WB152Q01HA < 0) then WB152Q01HA=.I;
3	Invalidate if number of class periods per week in test language lessons (ST059Q01TA) is greater than 40.	if (ST059Q01TA > 40) then ST059Q01TA =.I;
4	Invalidate if number of class periods per week in maths (ST059Q02TA) is greater than 40.	if (ST059Q02TA > 40) then ST059Q02TA =.I;
5	Invalidate if number of class periods per week in science (ST059Q03TA) is greater than 40.	if (ST059Q03TA > 40) then ST059Q03TA =.I;
6	Invalidate if number of <class periods> per week in foreign language is greater than 40.	if (ST059Q04HA > 40) then ST059Q04HA= .I;
7	Invalidate if number of total class periods in a week (ST060Q01NA) is greater than 120 or less than 10	if (ST060Q01NA > 120 or ST060Q01NA < 10) and NOT MISSING(ST060Q01NA) then ST060Q01NA =.I;

Exemple de contrôles de cohérence du programme PISA (Programme for International Student Assessment) (2018).

https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-10-Data-Management.pdf

Les activités sont nombreuses...

1) Encodage des réponses ouvertes

2) Préparation des données

3) Synthèse et visualisation

4) Édition des données

5) Imputation et pondération (voir module 2)

6) Contrôle de la divulgation

7) Traitement final, documentation et diffusion

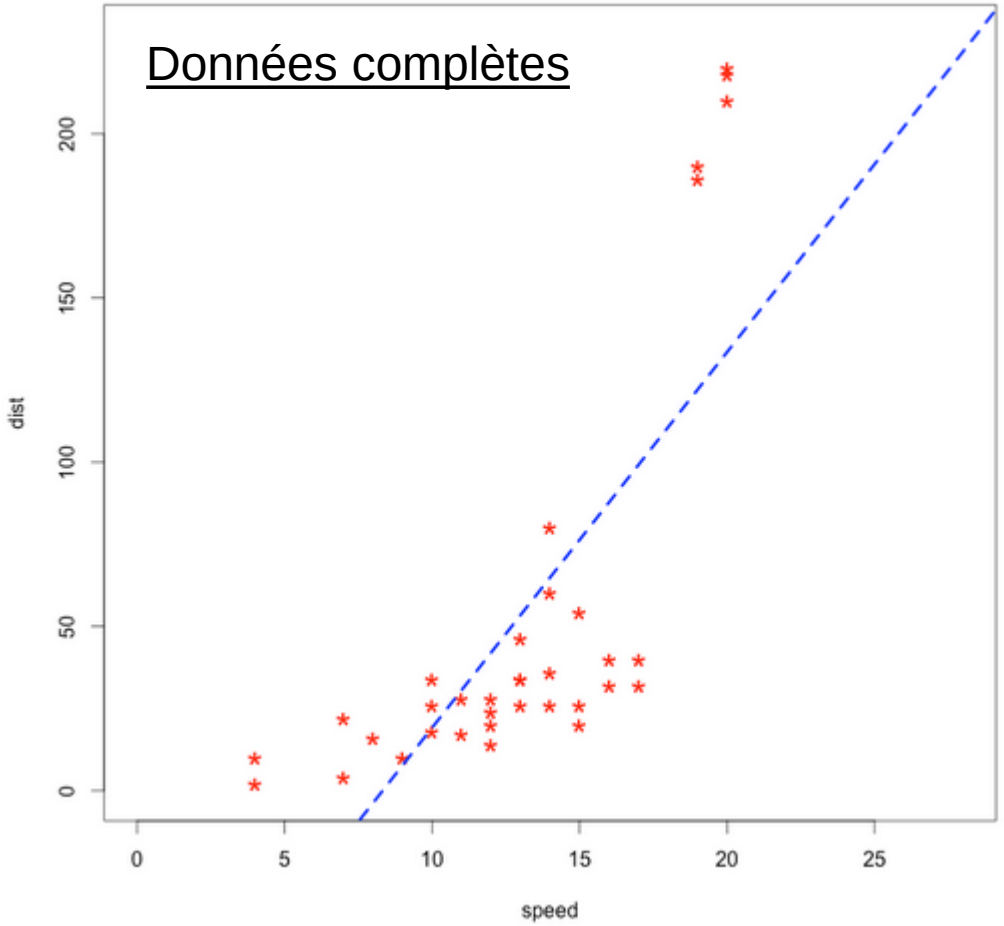
3) Édition des données

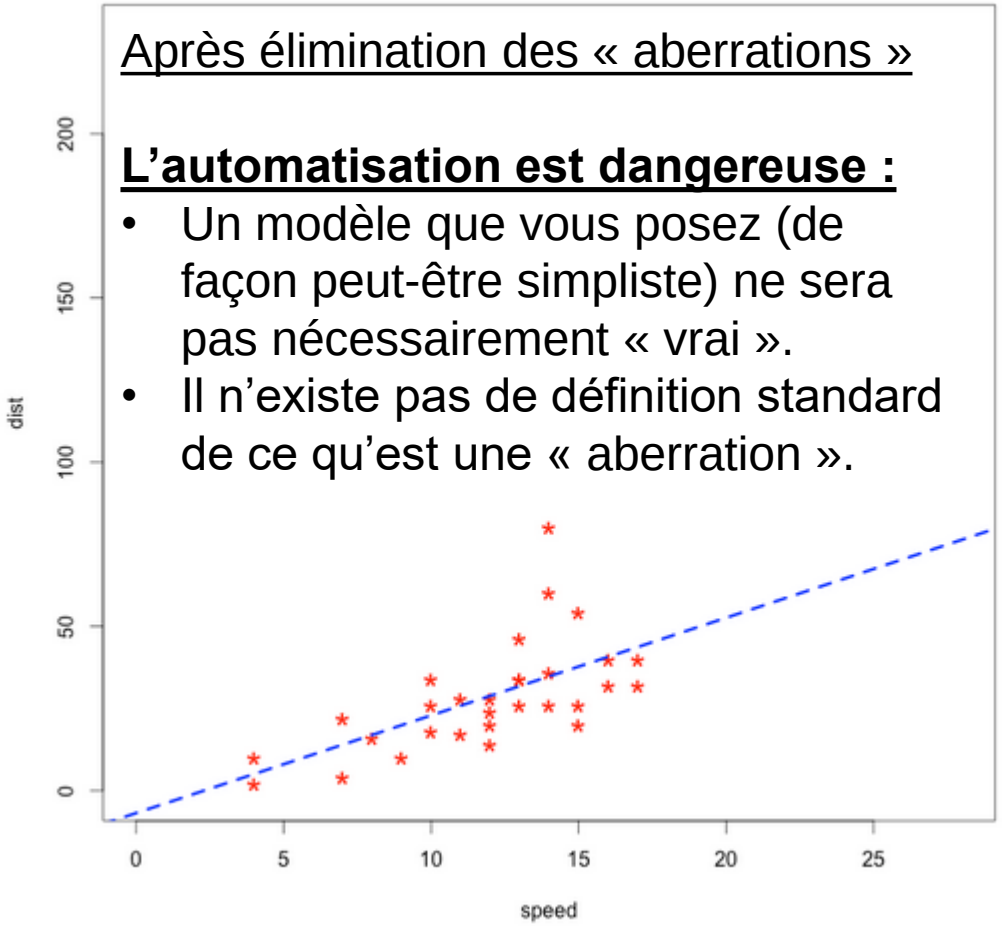
- Que faire des valeurs « extrêmes » ou incohérentes vues plus haut ?

- Sont-elles plausibles ?
  - Aberrations « représentatives » (extrêmes mais valides) / aberrations « non représentatives » (erreurs)

- « Numériquement éloignées du reste des données. » Mais de quelles données ?
  - Nombreux algorithmes

- Sur un plan plus formel : données ne correspondant à aucun modèle.





L'automatisation est dangereuse :

- Un modèle que vous posez (de façon peut-être simpliste) ne sera pas nécessairement « vrai ».
- Il n'existe pas de définition standard de ce qu'est une « aberration ».

Source du graphique : r-statistics.co/Outlier-Treatment-With-R.html

Détection des aberrations - questions pratiques

1. Il n'existe pas de définition claire de ce qu'est une aberration.
   - Aguinis et al (2013) : revue de la littérature avec 46 sources méthodologiques, et 232 articles de publications scientifiques du secteur.

     Résultat : 14 définitions différentes des aberrations, 39 techniques de détection, et 20 manières différentes de traiter les aberrations détectées.
   - Quelle que soit la définition choisie, faites preuve de cohérence.

2) Répartitions asymétriques
   - Il est toujours possible d'utiliser une boîte à moustaches, mais en appliquant la transformée de Box-Cox aux données.

Détection des aberrations - questions pratiques

3) Beaucoup de zéros.

- Décrivez/tracez sans les zéros.
- Pour de très rares évènements, les procédés habituels de détection des aberrations ne sont pas adaptés.

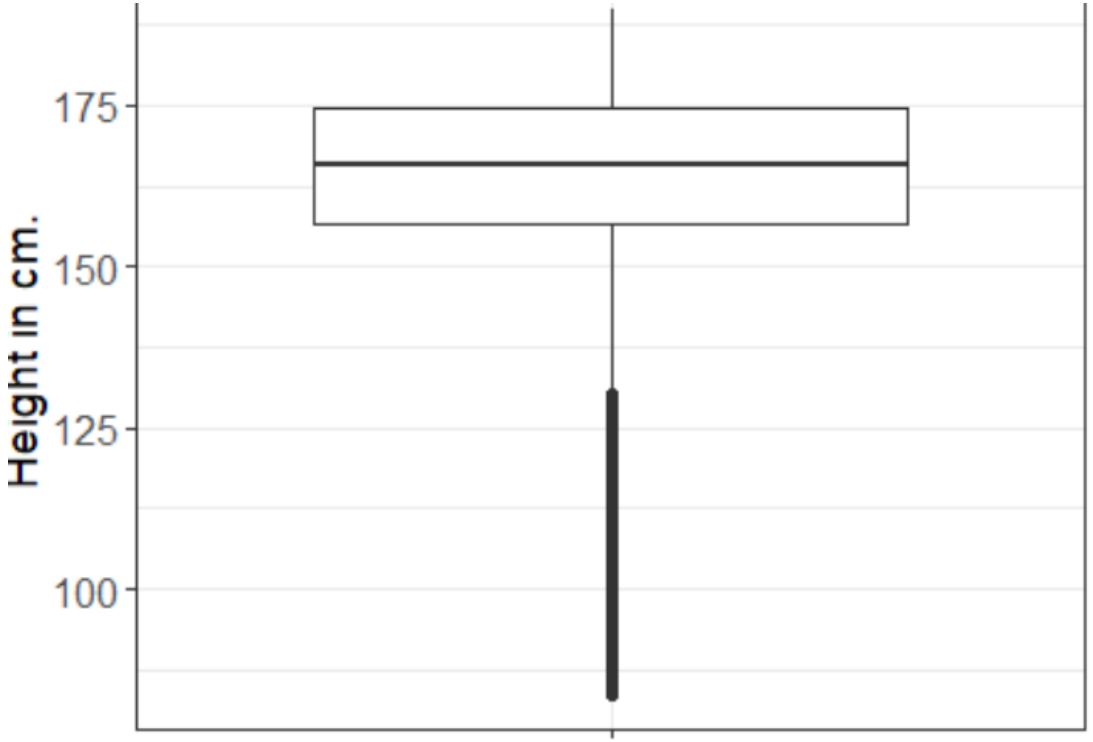
4) Non pondérée ou pondérée ?

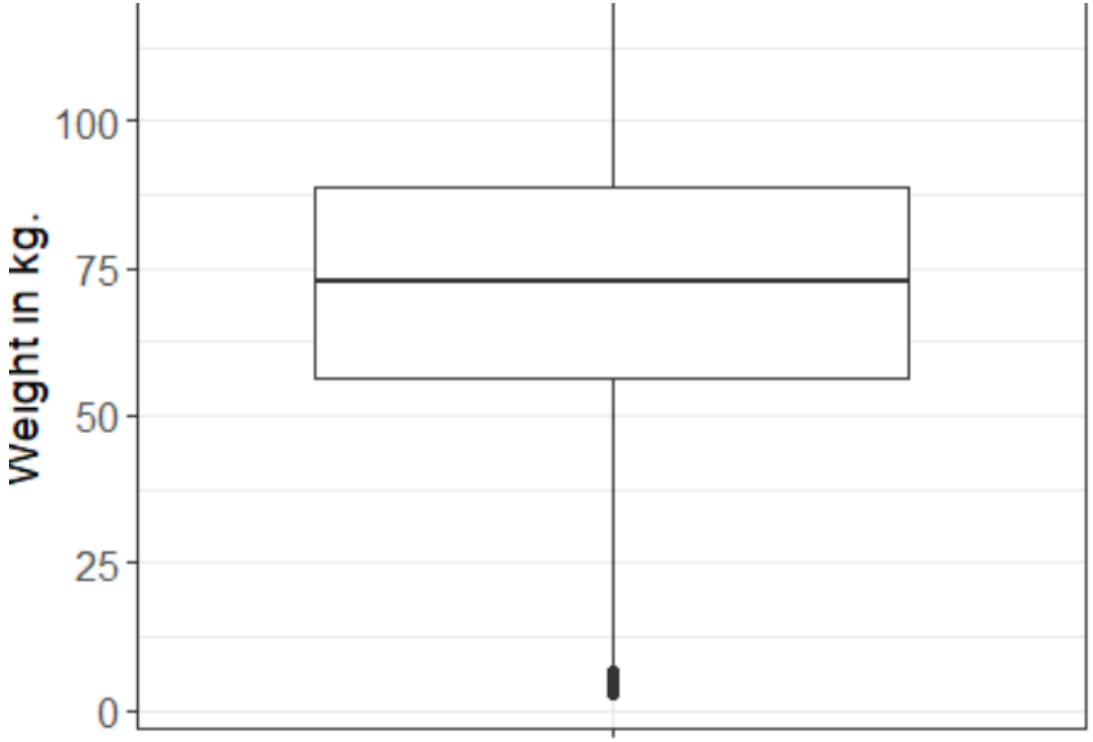
- Commencez par une détection non pondérée, mais n'écartez pas pour autant la détection pondérée

5) Erreur courante : élimination des aberrations. Nouvelle analyse : de nouvelles aberrations apparaissent. Élimination à nouveau. -> L'édition de données est par nature une source d'erreur.

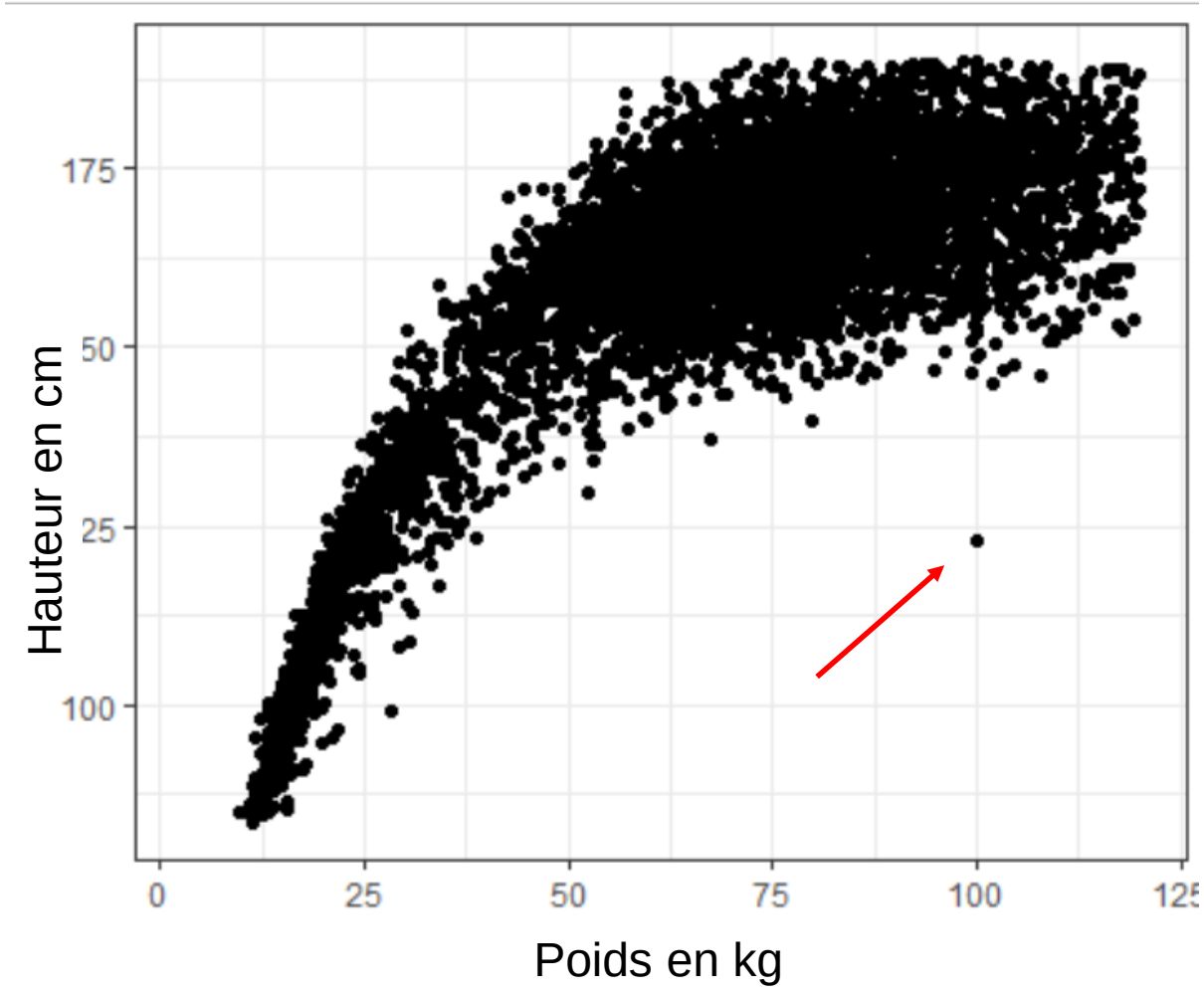
6) Ne vous arrêtez pas à une analyse unidimensionnelle...

Boîtes à moustaches unidimensionnelles





Nuage de points bidimensionnel



Aberrations multidimensionnelles

- Étant donné que la plupart des enquêtes collectent plusieurs (centaines/milliers de) variables, on peut en théorie réaliser des visualisations 3D, 4D, etc.
  - Il faut une solution automatisée.

- Plusieurs algorithmes multidimensionnels de détection des aberrations existent, comme Epidemic, BACON-EEM
  - Ils recourent à des estimations robustes pour éviter que le « centre » des données ne soit déformé par les valeurs extrêmes.
  - Ils prennent également en compte les poids d'échantillonnage.

Pour plus de détails, voir Filzmoser et al (2016), et Todorov et al (2009)

Que faire des aberrations ou des valeurs qui ne passent pas les vérifications de cohérence ?

1 : les conserver

Si les valeurs proviennent d'une erreur de saisie avérée, faut-il les conserver ? Cela peut nuire à l'analyse.

Dans certains cas, il est possible d'appeler le répondant pour confirmation.

Pour les études en panel, il peut être utile de vérifier les anciennes valeurs fournies par le répondant pour les mêmes variables.

Que faire des aberrations ou des valeurs qui ne passent pas les vérifications de cohérence ?

1 : les conserver

2 : les supprimer

- Les méthodes de détection automatique sont nécessaires pour la plupart des enquêtes, mais la suppression automatique en fonction d'un seuil donné n'est pas une bonne idée.
- Il faut examiner les aberrations.
- Les aberrations constituent souvent une part informative des données...

Un leçon tirée dans un autre contexte

Pourquoi n'ont-ils pas découvert le phénomène plus tôt ? Malheureusement, le logiciel d'analyse de données TOMS avait été programmé pour détecter et écarter les points trop éloignés des mesures attendues. Les mesures initiales qui auraient dû alerter ont donc été tout simplement ignorées. En bref, l'équipe TOMS n'a pas détecté la baisse de la couche d'ozone plus tôt parce qu'elle était beaucoup plus grave que ce que les scientifiques imaginaient.

https://earthobservatory.nasa.gov/features/RemoteSensingAtmosphere/remote_sensing5.php

Que faire des aberrations ou des valeurs qui ne passent pas les vérifications de cohérence ?

1 : les conserver

2 : les supprimer

3 : winsorisation/ajustement statistique

4 : imputation

winsorisation/ajustement statistique

- Les valeurs dépassant/ou pas une certaine limite sont remplacées par cette valeur limite.

- Rien n’est gratuit : vous réduirez peut-être la variance, mais vous risquez d’augmenter le biais (voir Module 1 pour les concepts de biais et de variance).

- Pour les répartitions asymétriques, on peut avoir recours à des procédures comme la modélisation de la queue de Pareto. Les valeurs dépassant un seuil prédéfini de distribution sont remplacées par une valeur prédite.

Que faire des aberrations ou des valeurs qui ne passent pas les vérifications de cohérence ?

1 : les conserver

2 : les supprimer

3 : winsorisation

4 : imputation

Les activités sont nombreuses...

1) Encodage des réponses ouvertes

2) Préparation des données

3) Synthèse et visualisation

4) Édition des données

5) Imputation et pondération (voir Module 2 pour la pondération)

6) Contrôle de la divulgation

7) Traitement final, documentation et diffusion

FIN DE LA VIDÉO 2