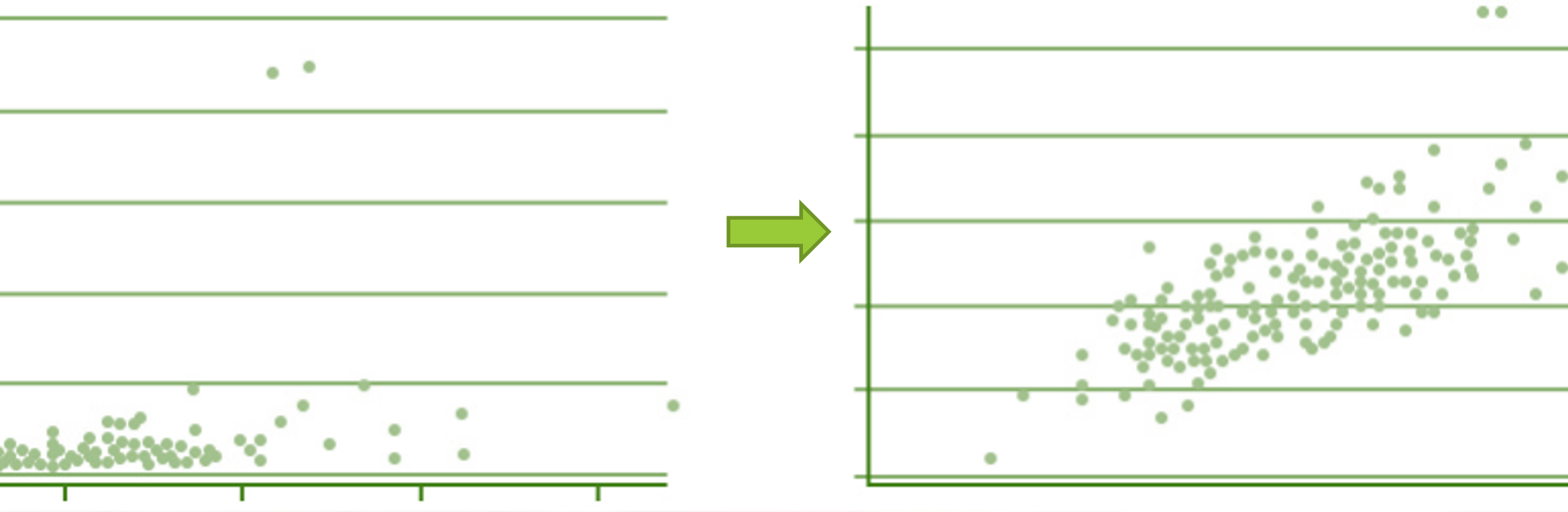


10. La dimensionnalité et les transformations de données

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)

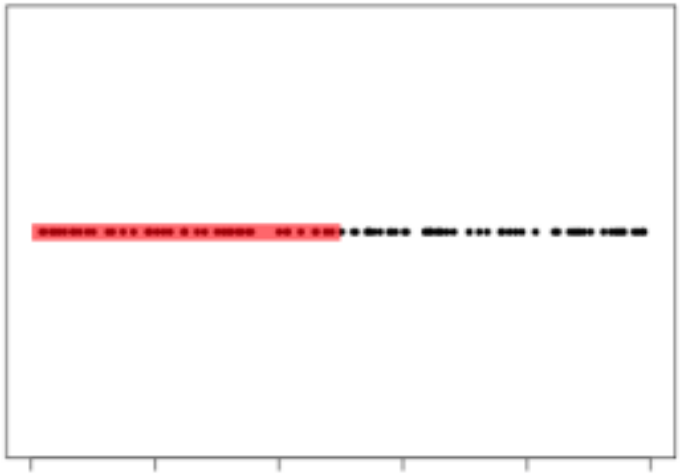
Exemples :

- 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**

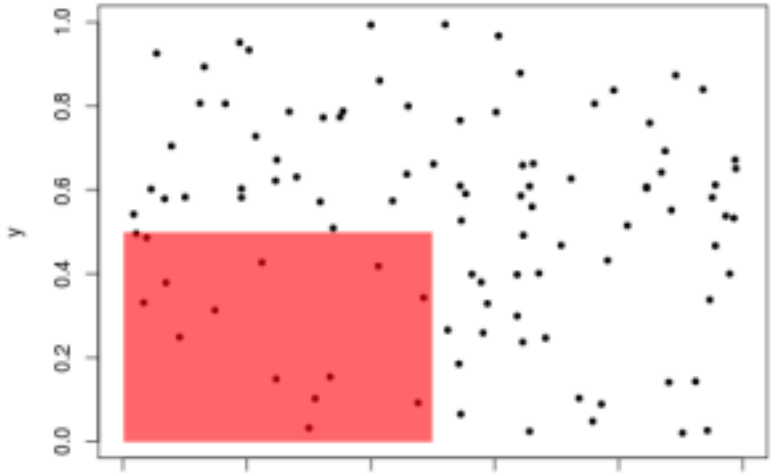
[http://simplystatistics.org]

Le fléau de la dimensionnalité

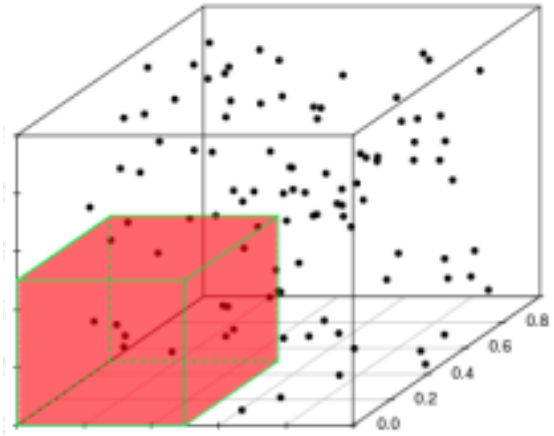
42% des données sont capturées



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

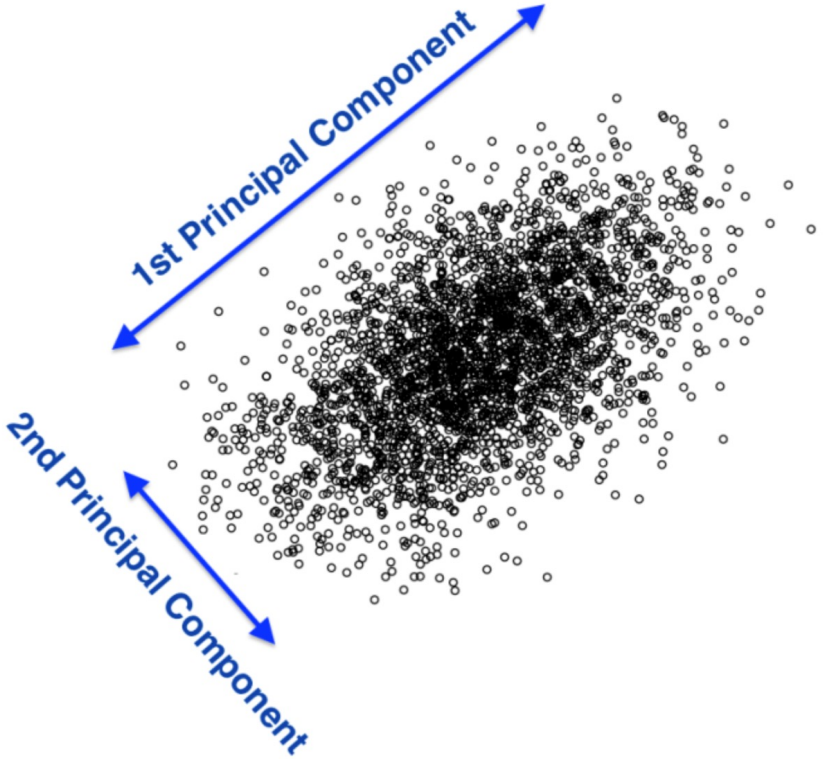
[A. Ng, K. Soo, *Numsense !*]

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 !*]

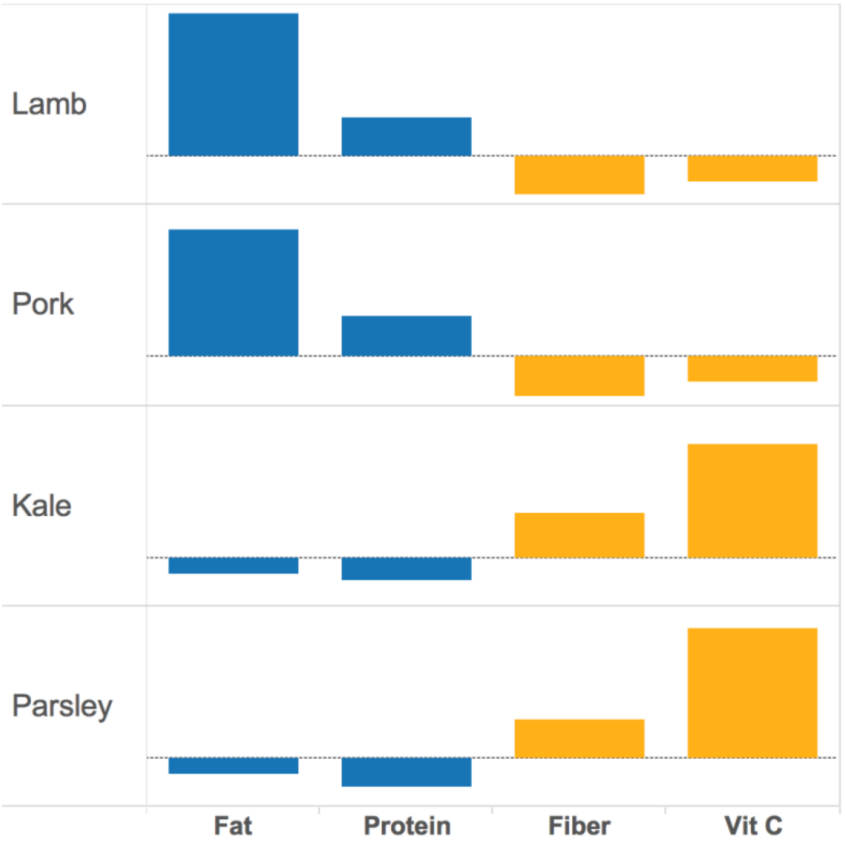
La réduction de dimension : ACP

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*.

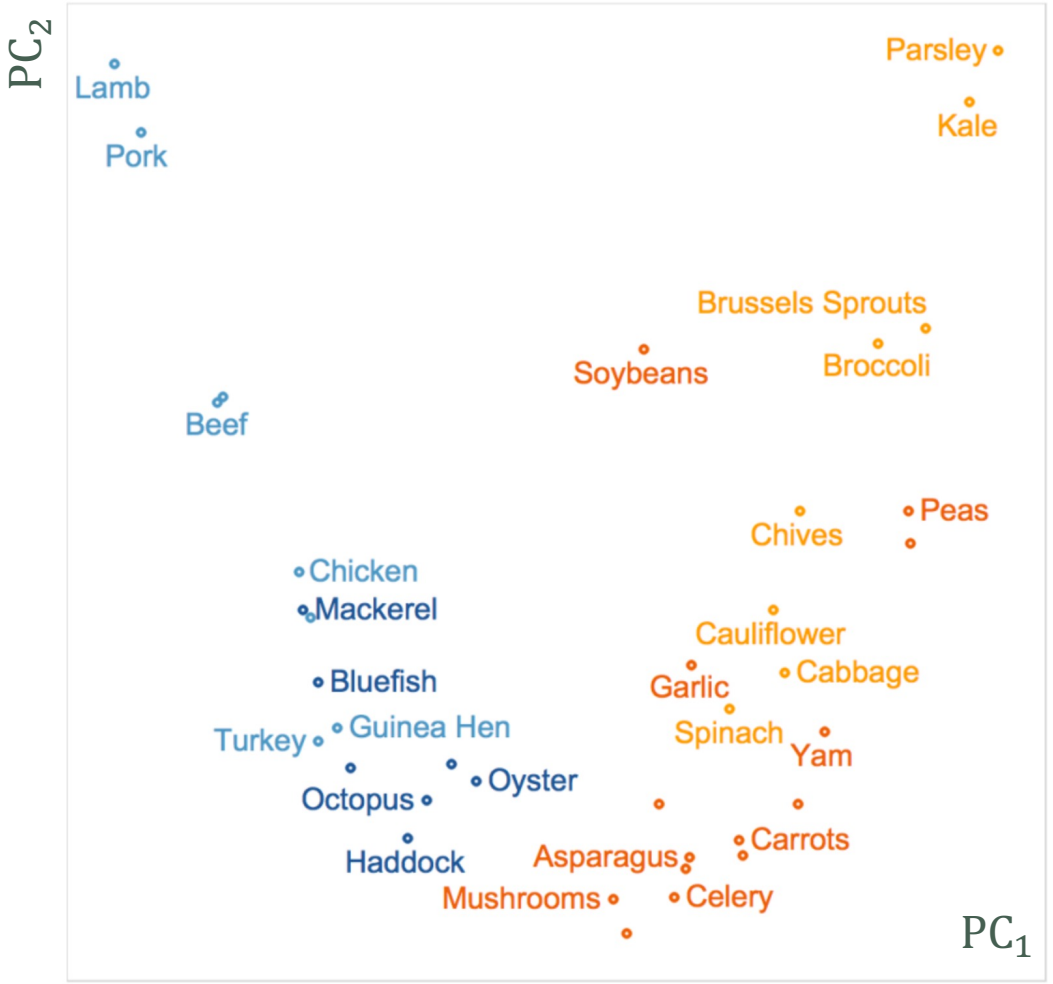
Dans un ensemble de données plus vaste, les corrélations sont $r = 0.56$ et $r = 0.57$

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



$$PC_1 = -0.45 \times \text{Fat} - 0.55 \times \text{Protein} + 0.55 \times \text{Fiber} + 0.44 \times \text{Vitamin C}$$
$$PC_2 = 0.66 \times \text{Fat} + 0.21 \times \text{Protein} + 0.19 \times \text{Fiber} + 0.70 \times \text{Vitamin C}$$

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**.

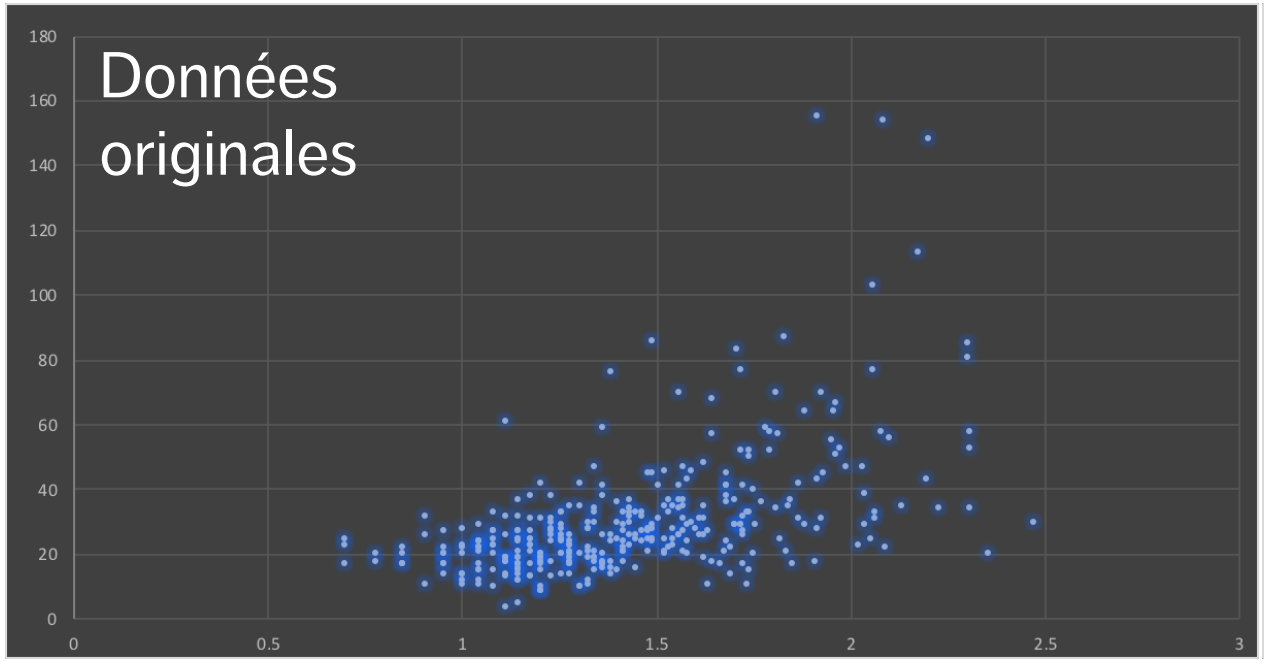
Les transformations communes

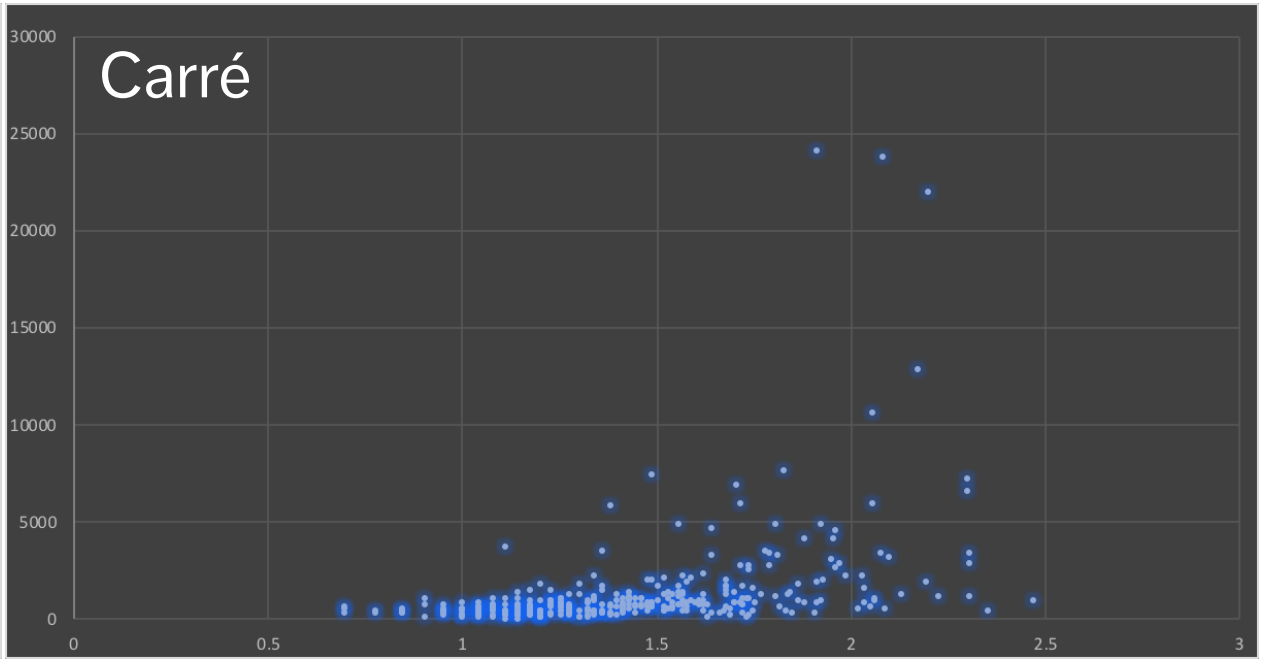
Dans le contexte de l'analyse des données, les transformations sont **monotones :**

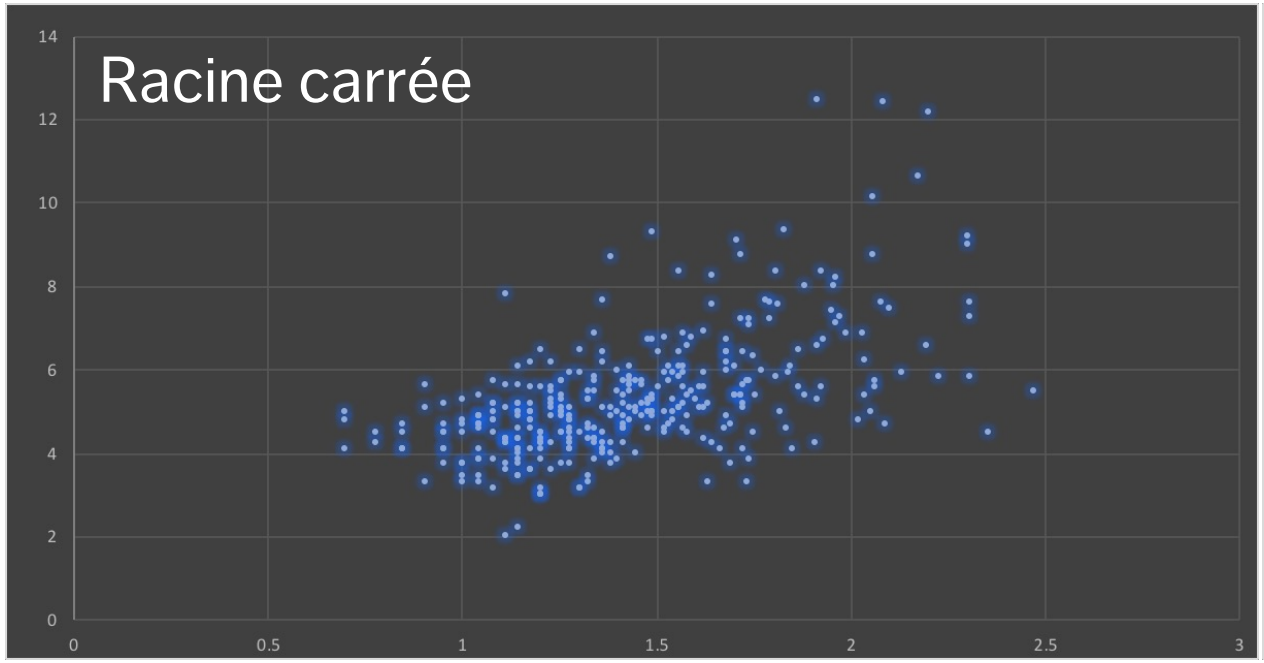
- logarithmique
- racine carrée, inverse, puissance :
- exponentielle
- Box-Cox, etc.

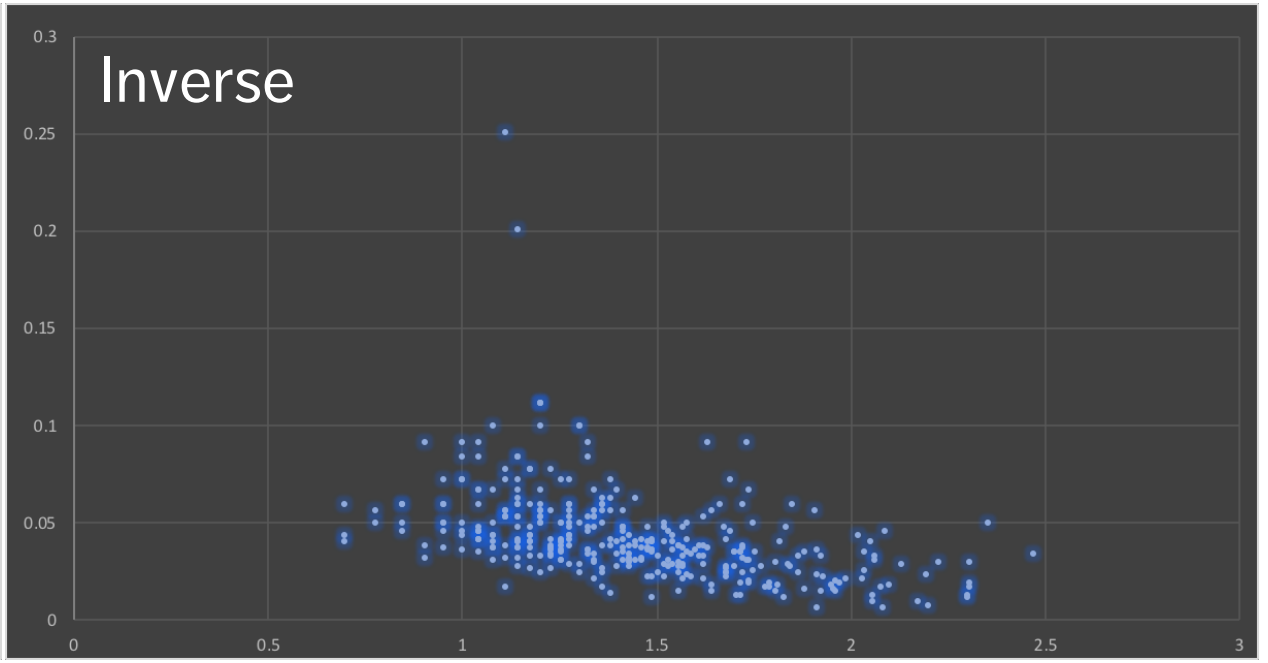
Les transformations sur les prédicteurs X peuvent atteindre la linéarité, mais à un prix (les corrélations ne sont pas préservées, par exemple).

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


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

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

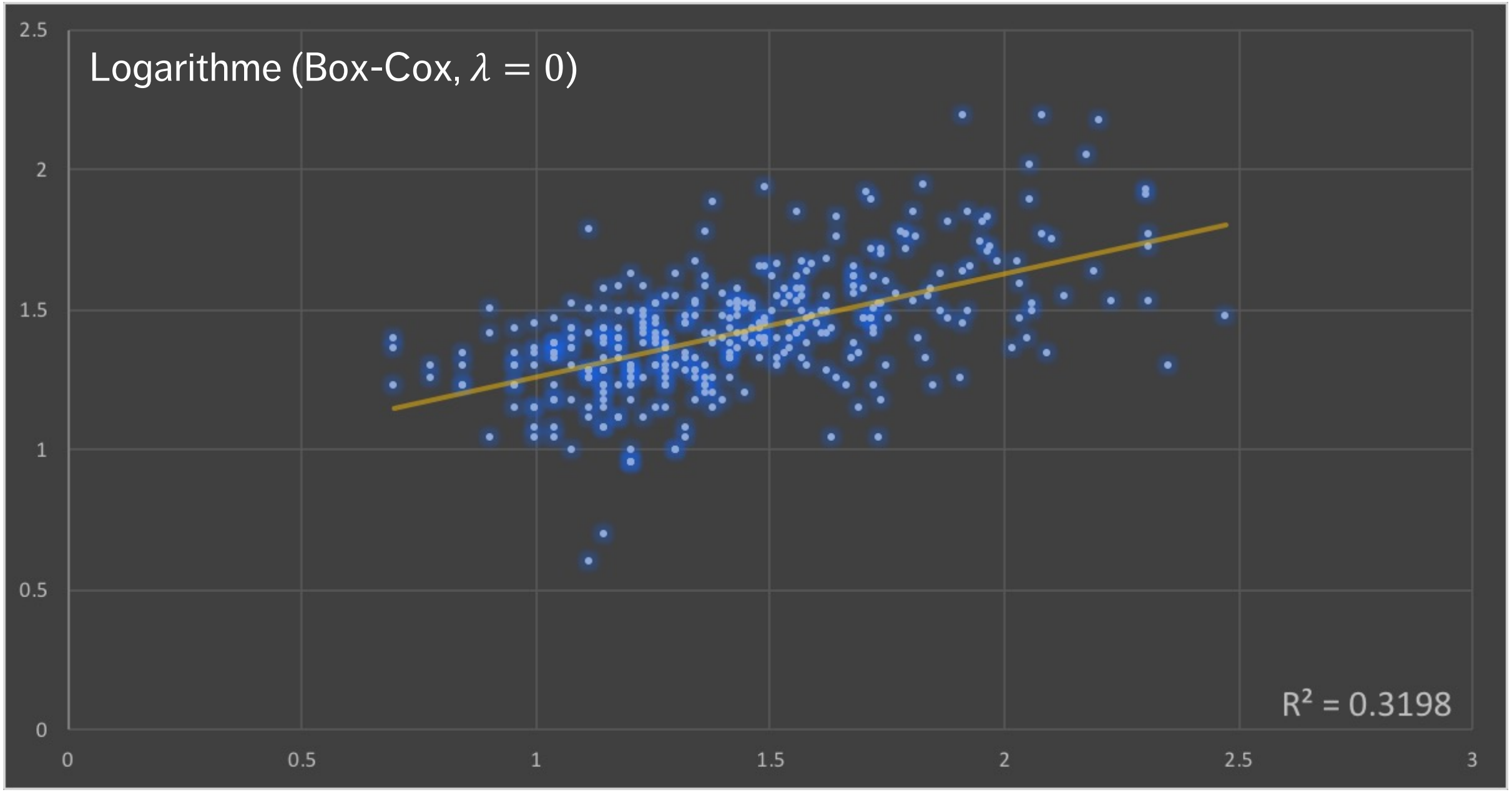
La transformation de Box-Cox

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 suggérées

La dimensionnalité et les transformations de données

Data Understanding, Data Analysis, Data Science

Data Preparation

Data Transformations

- Common Transformations
- Box-Cox Transformations
- Scaling
- Discretizing
- Creating Variables

***Feature Selection and Dimension Reduction** (advanced)

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.