Semaine Data SHS
5 - 9 décembre 2022

Introduction à la dataviz

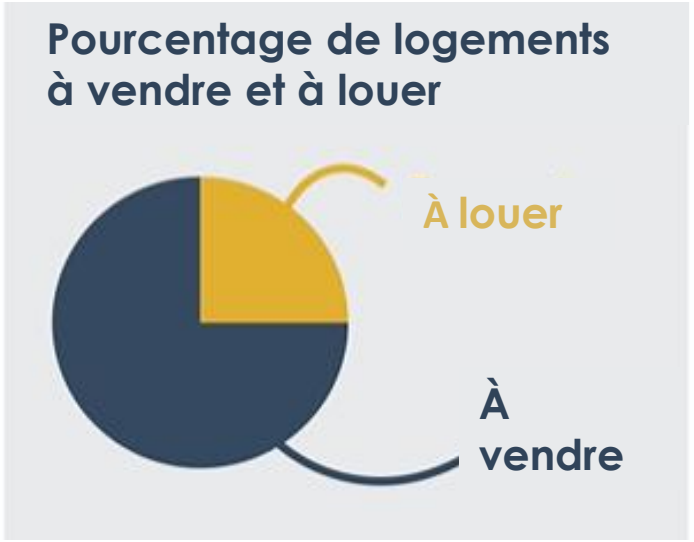
Gaëlle Coz, PUD-UP, 7 décembre 2022

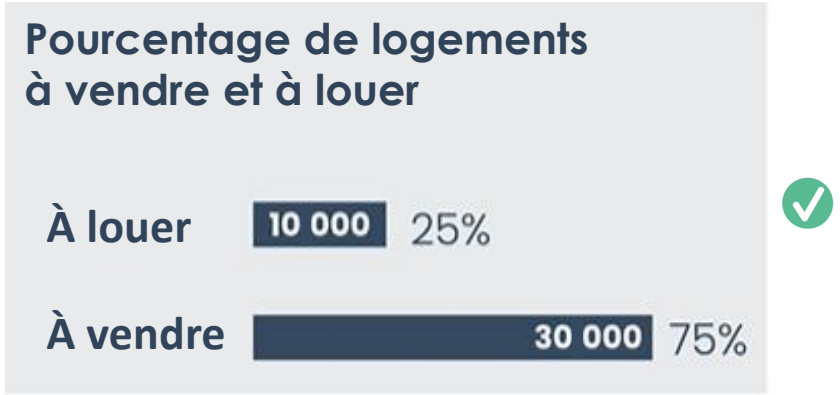
Partie 1 – Quelques règles de visualisation des données

L’excellence graphique

Un concept introduit par Edward Tufte dans son livre « The Visual Display of Quantitative Information » (2001)

Montrer la donnée

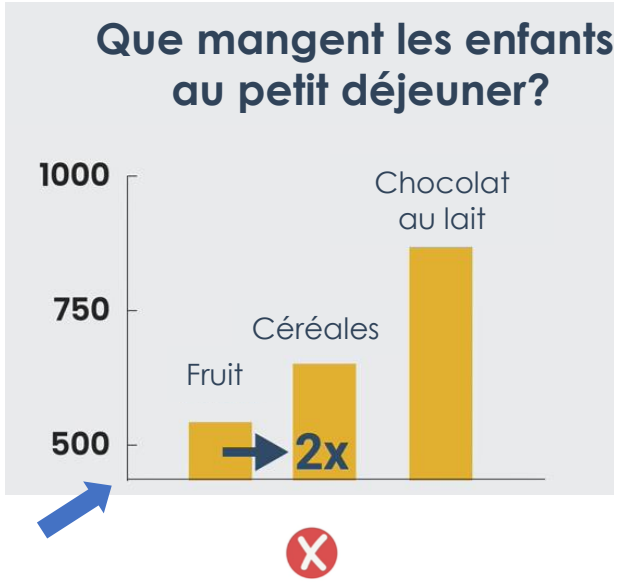


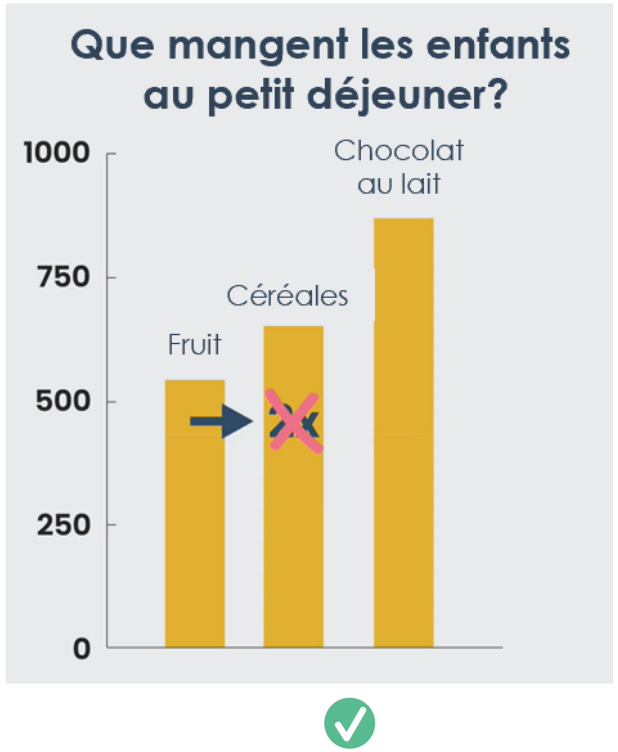


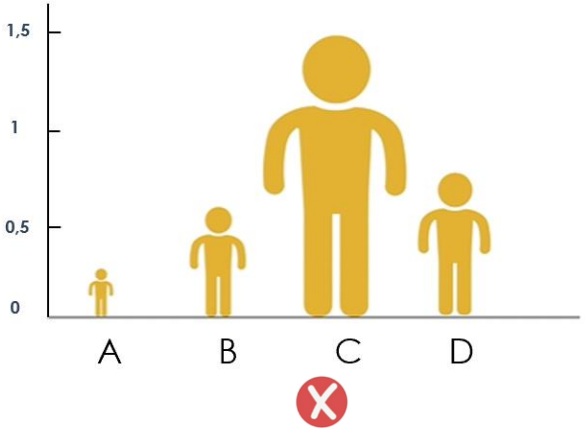
Ne pas cacher les données avec des éléments qui attirent l'attention et qui éloignent de l'information

Ne pas déformer les données

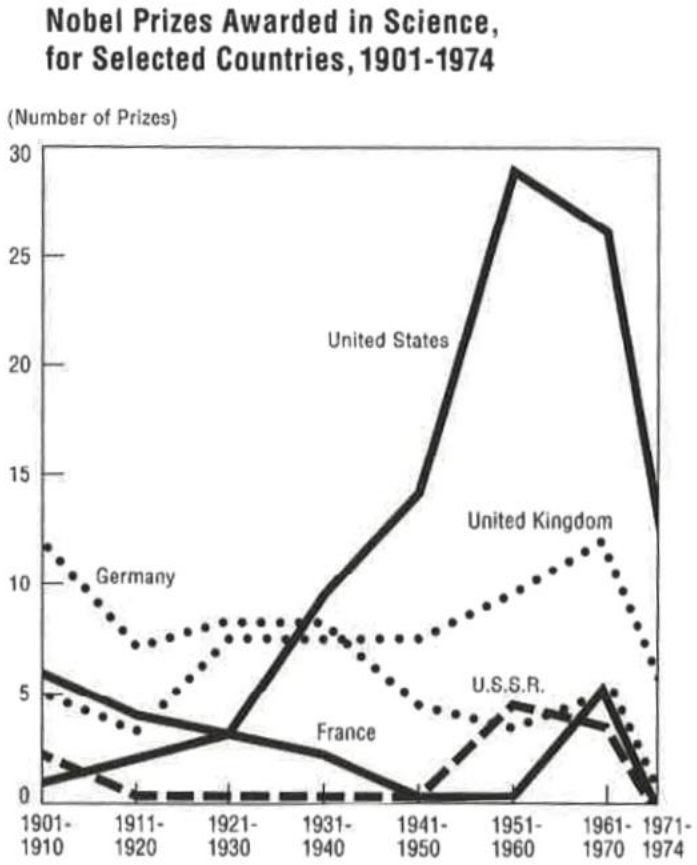
S’assurer que les surfaces qui représentent les nombres sur le graphique soient directement proportionnelles aux valeurs numériques qu’on veut représenter



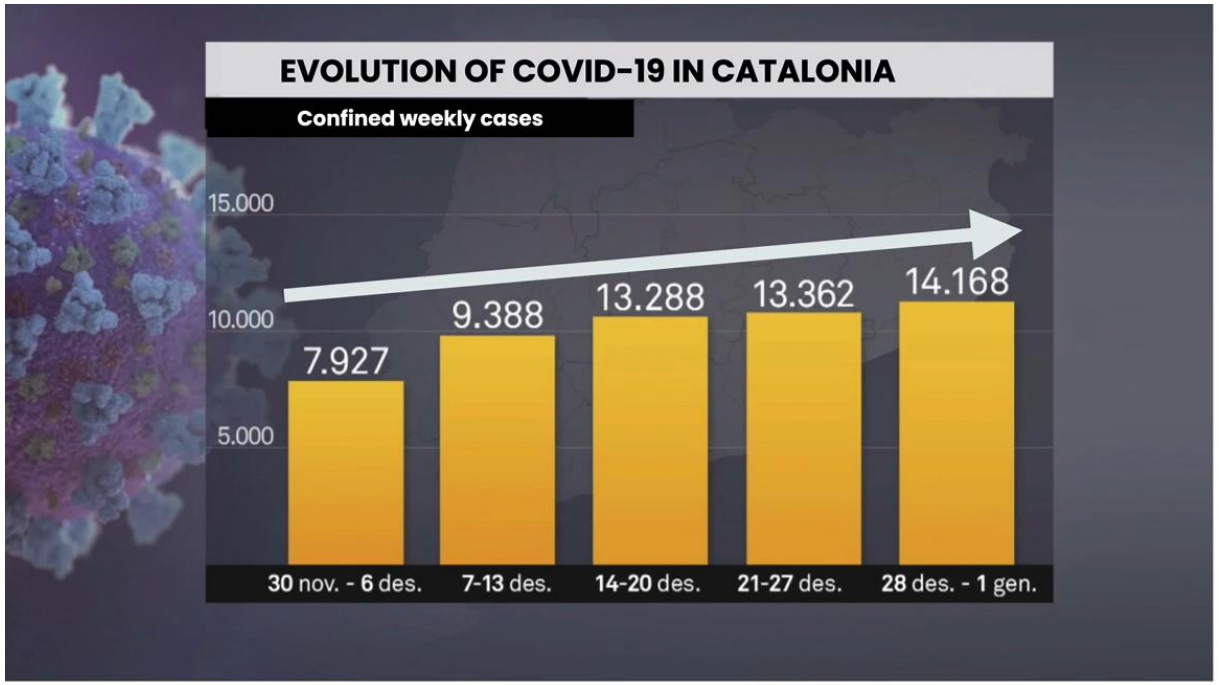




Ne pas varier la construction du graphique



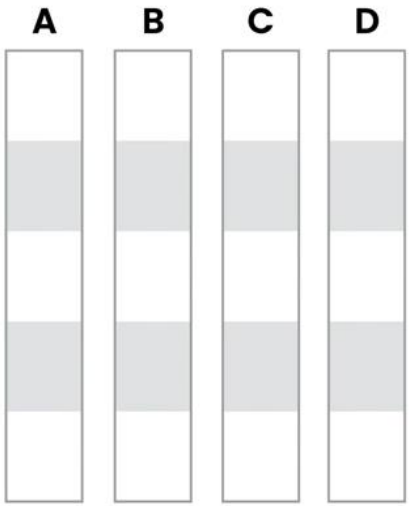
SOURCE: NATIONAL SCIENCE FOUNDATION
VIA THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION, E. TUFTE



SOURCE: TV3 TELEVISION. AIRED: 2 JAN 2021

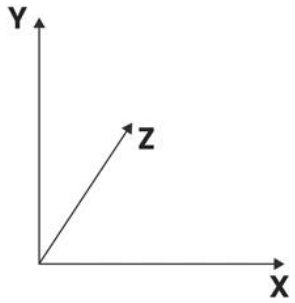
Le nombre de dimensions dans le graphique doit être égal au nombre de dimensions dans les données

Dimensions des données



Nombre de colonnes

Dimensions de la représentation

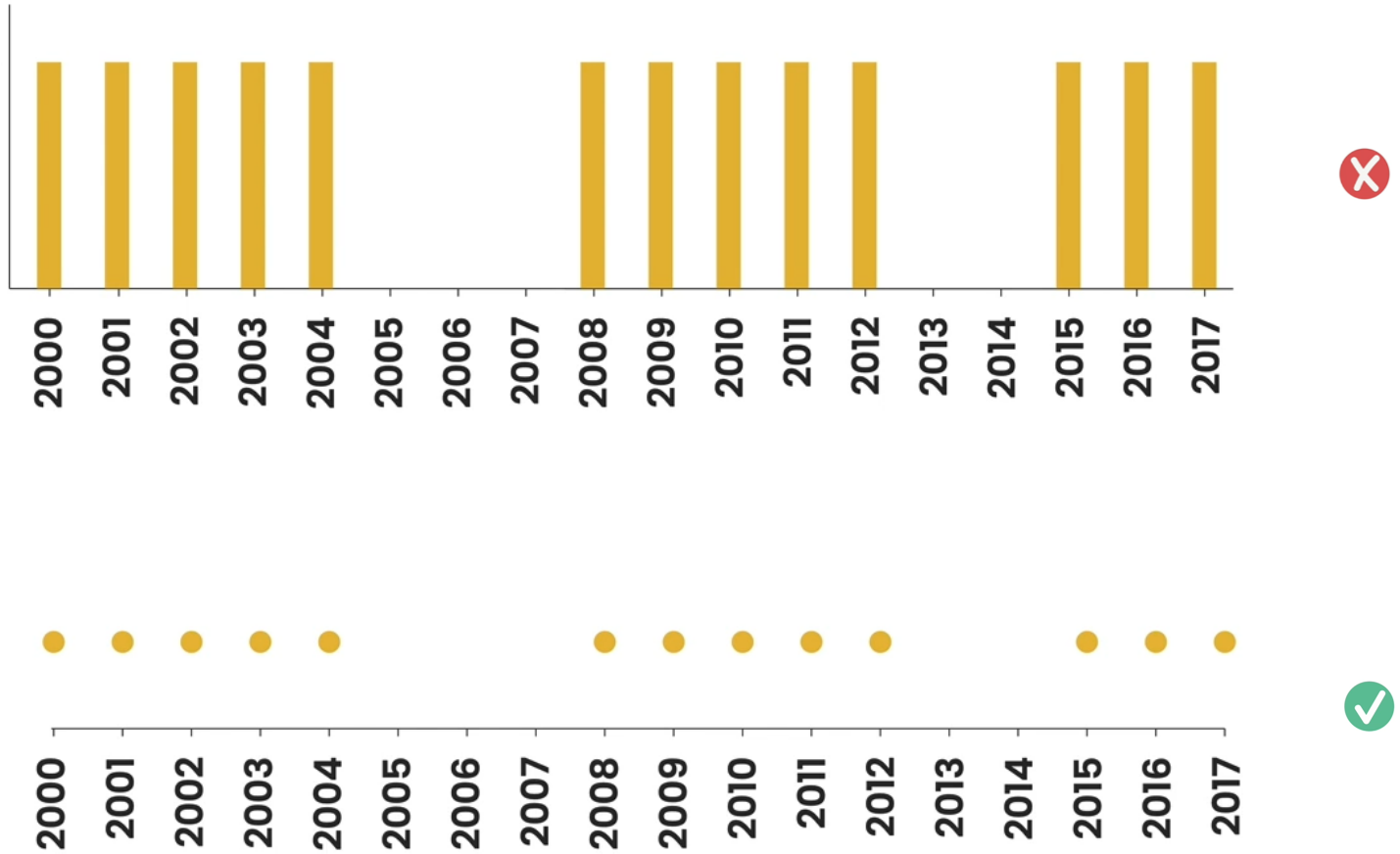


Nombre d’axes + dimensions cachées

1 dimension dans les données

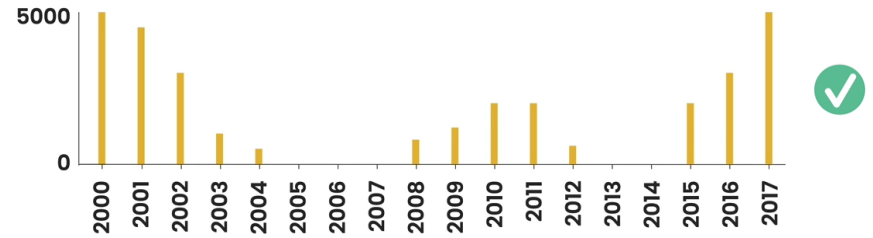
Années où l'entreprise est active

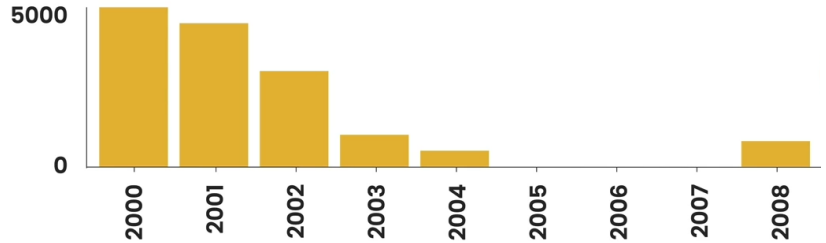
2000
2001
2002
2003
2004
2005
2006
2010
2011
2012
2015
2016
2017

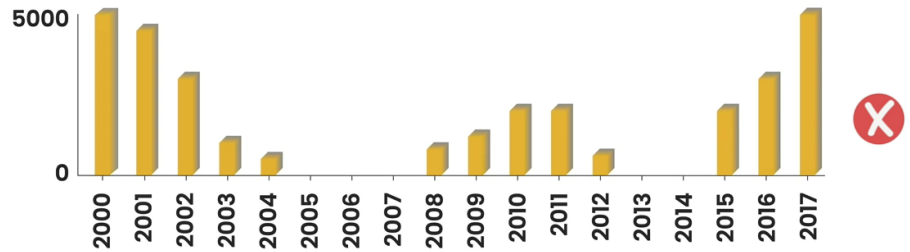


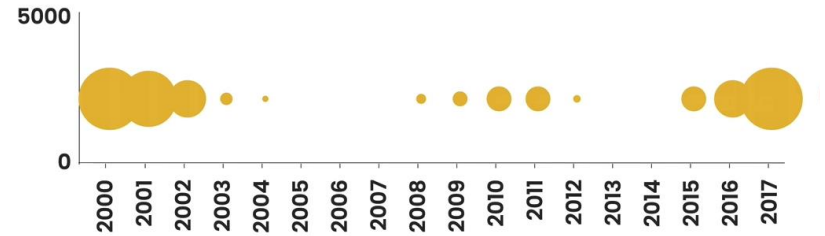
2 dimensions dans les données

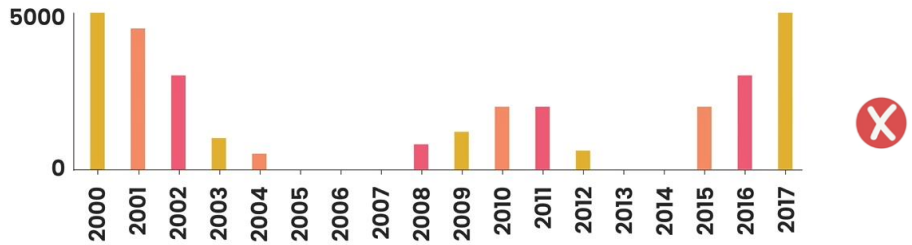
Années actives	Bénéfices
2000	5000
2001	4500
2002	3000
2003	1000
2004	500
2005	800
2006	1200
2010	2000
2011	2000
2012	600
2015	2000
2016	3000
2017	5000



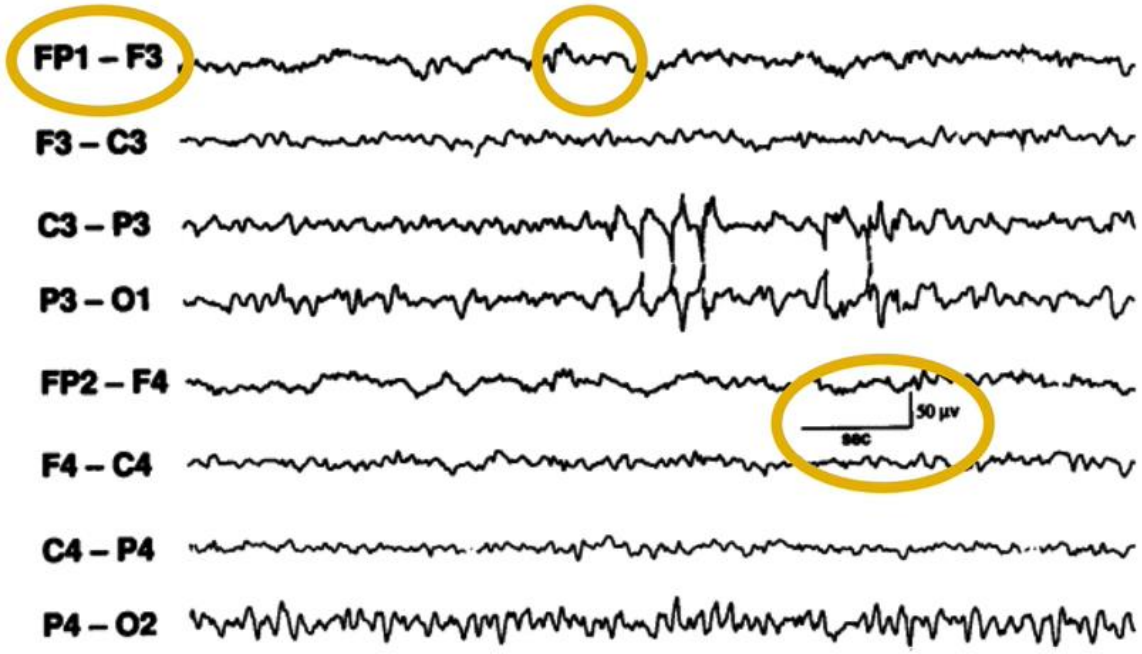




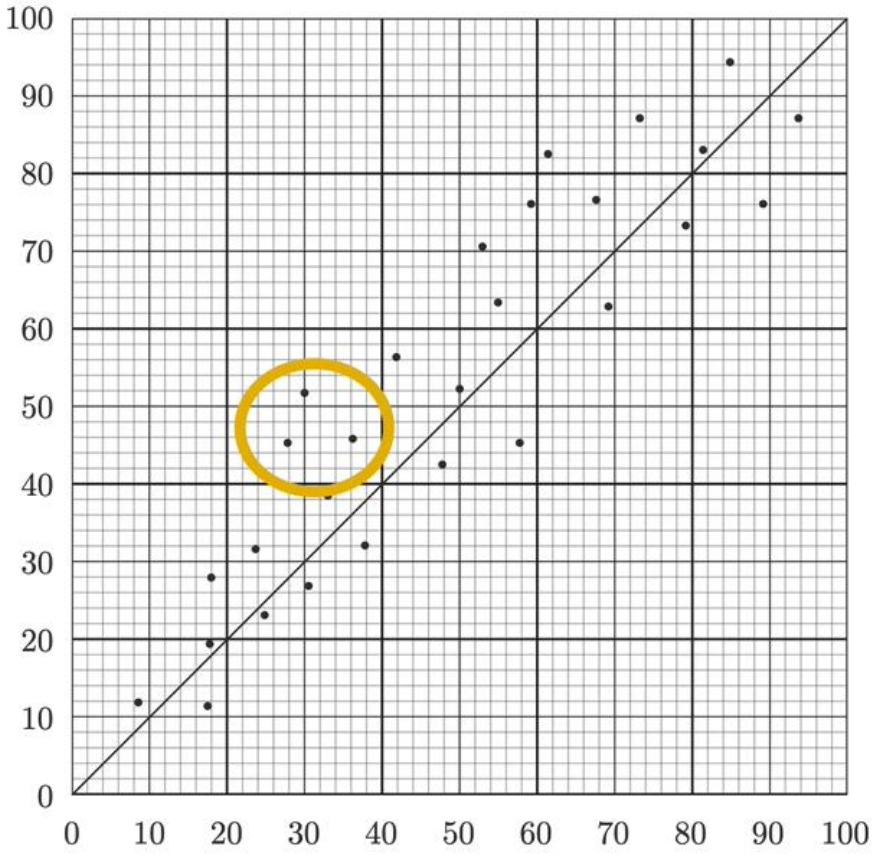




Maximiser le rapport données – encre pour gagner en lisibilité



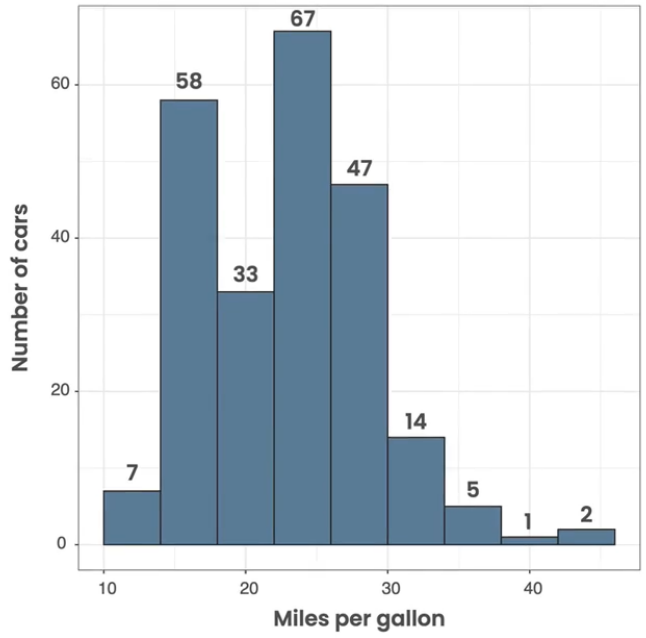
rapport maximal

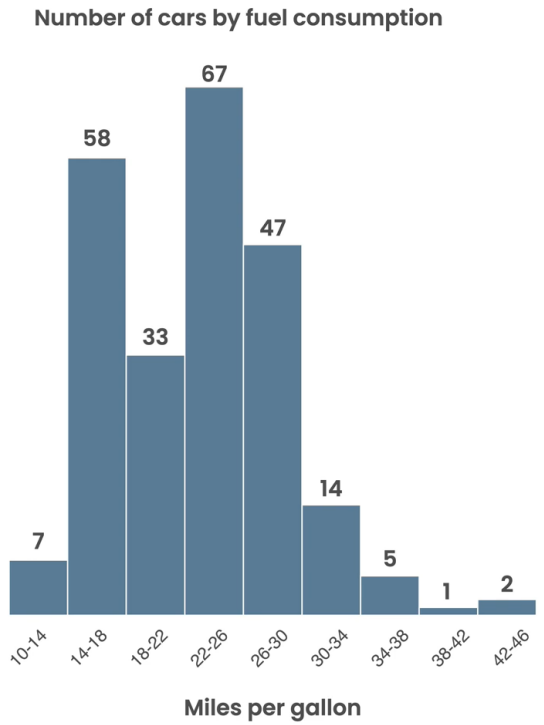


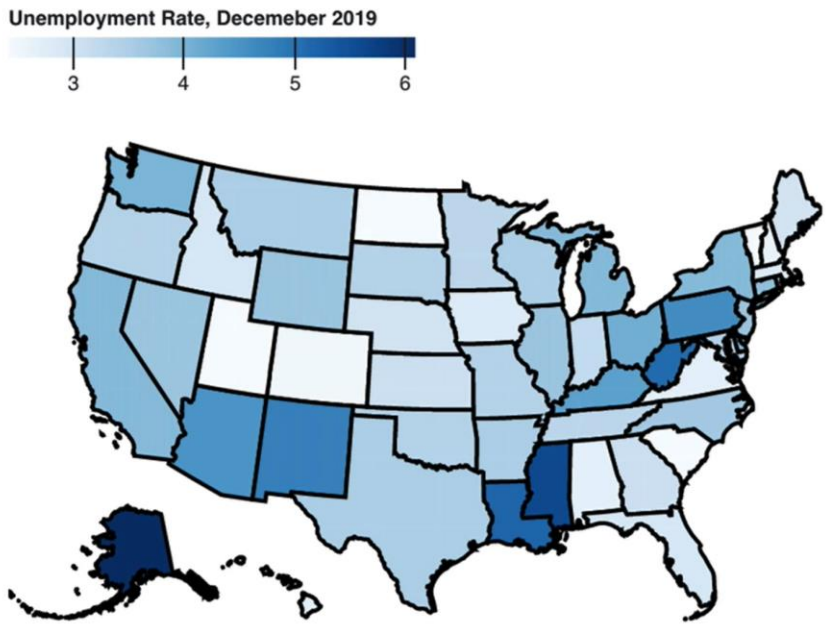
rapport très faible

Maximiser le rapport données – encre en effaçant :

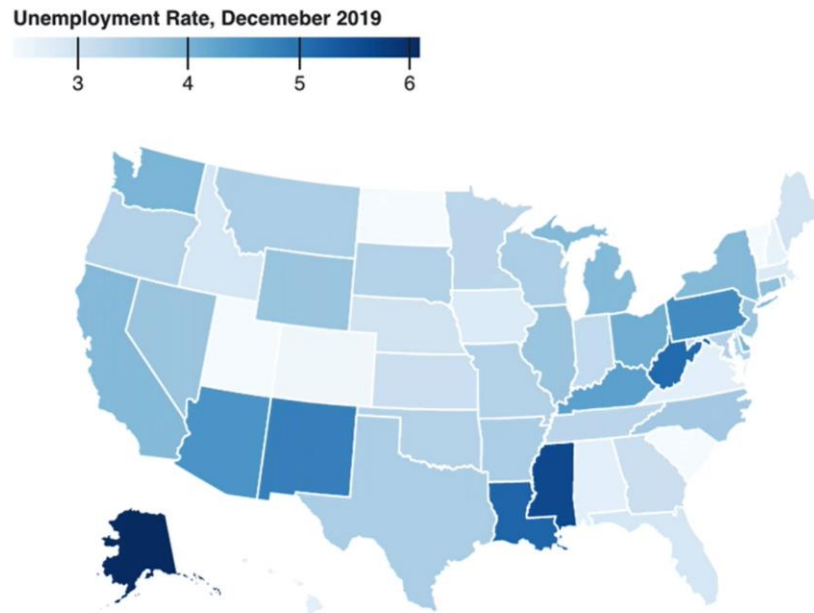
- L’encre qui ne correspond pas à de la donnée
- L’encre qui correspond à de la donnée mais de la donnée redondante







vs.

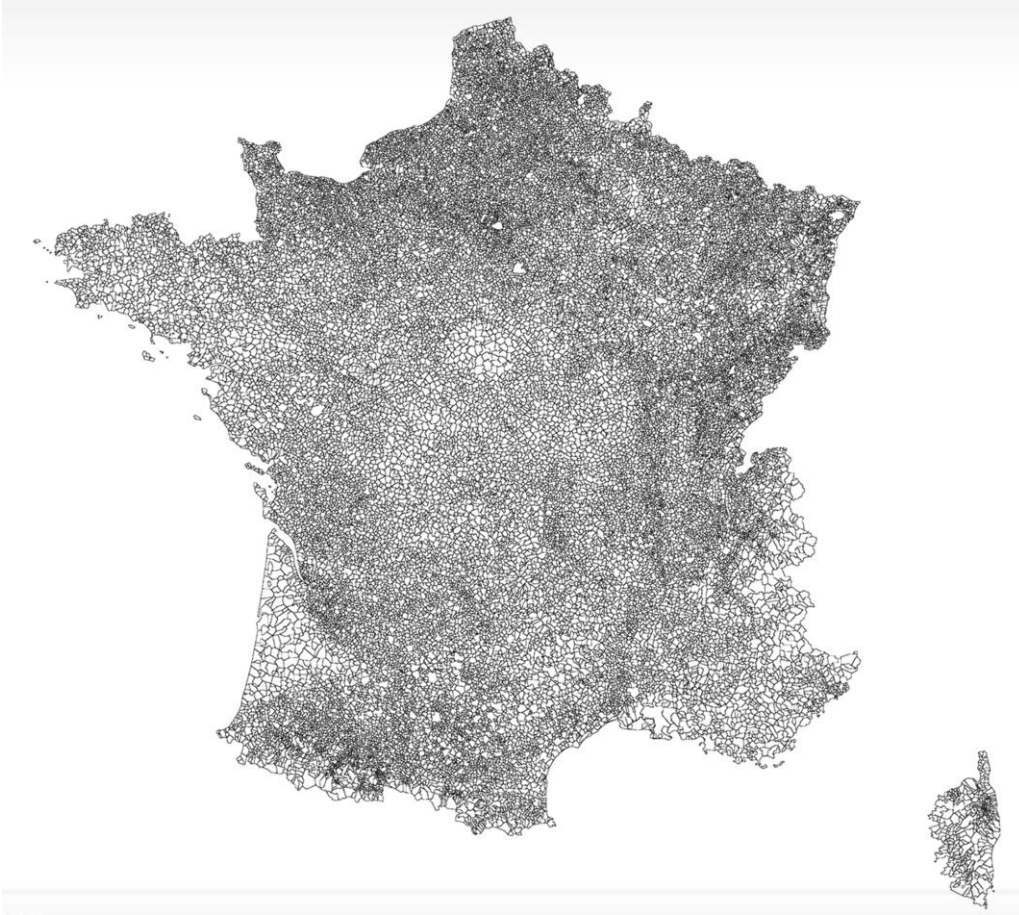


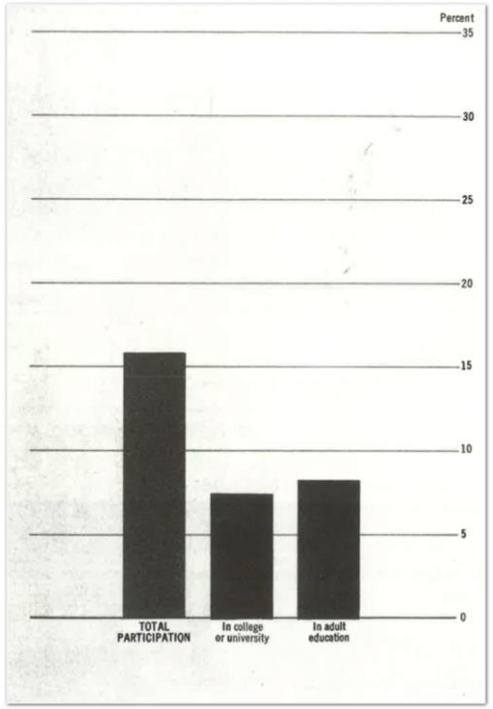
SOURCE: ENRICO BERTINI

Densifier les données

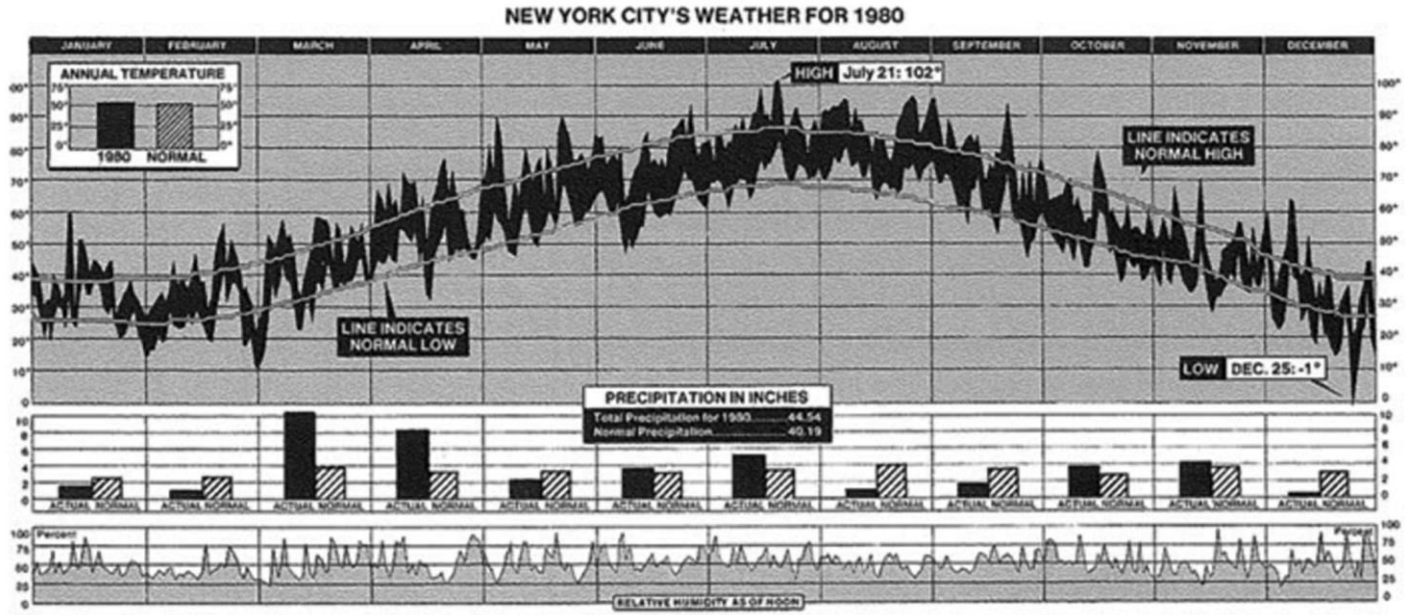
L’œil peut s’accommoder d’un niveau de détails très élevé.

L’objectif est d’en tirer parti et de représenter la plus grande quantité de données dans le plus petit espace possible, dans une limite raisonnable.





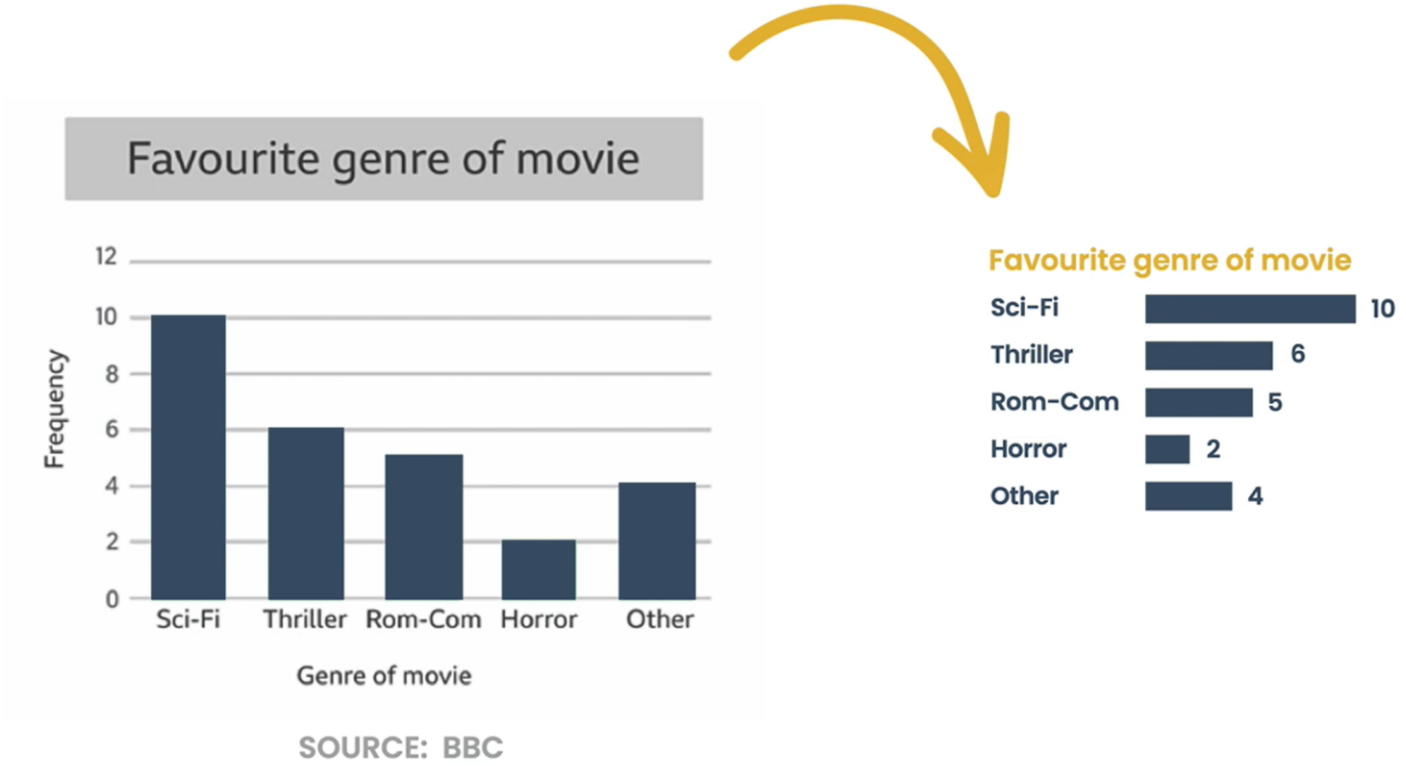
Très faible densité

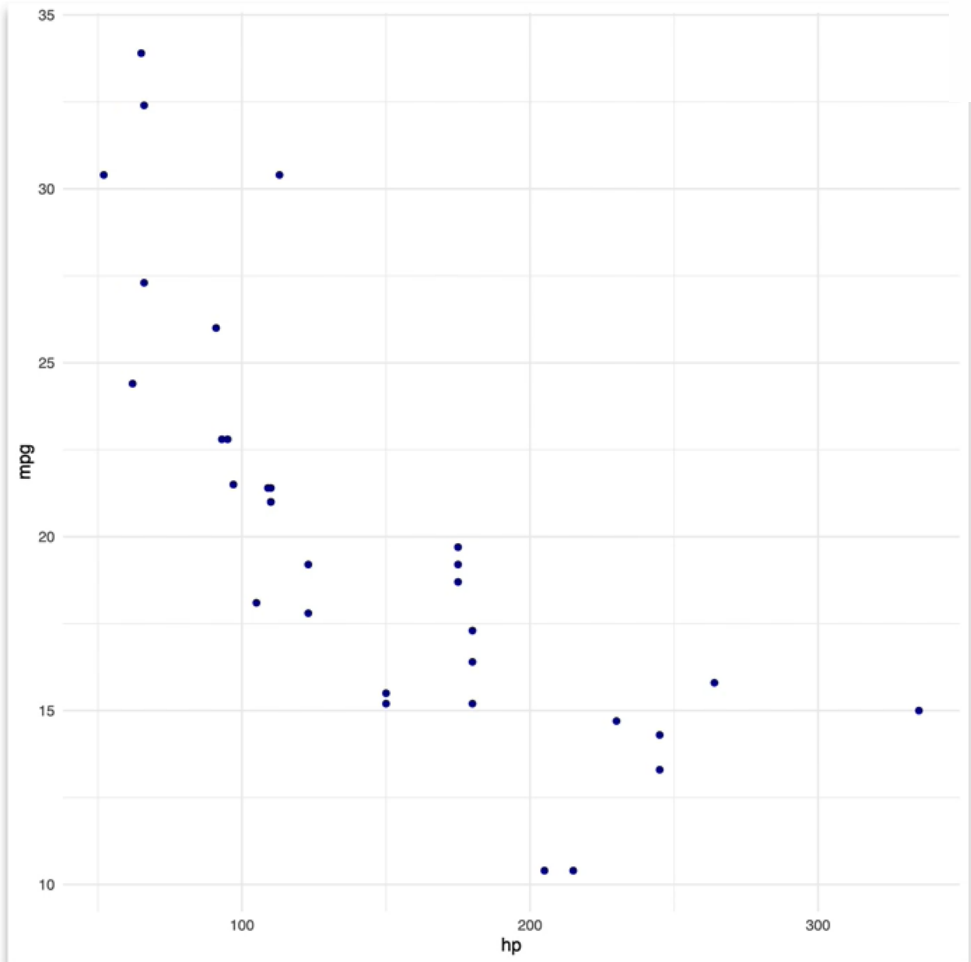


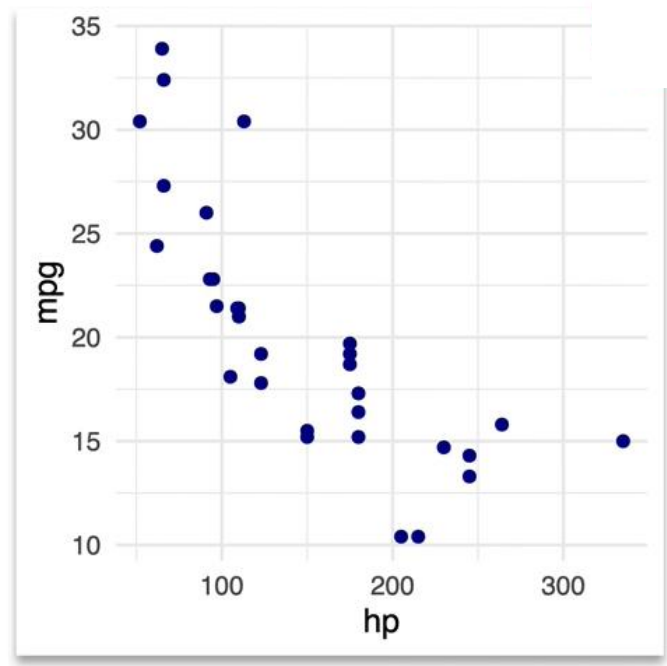
New York Times, January 11, 1981, p. 32.

Très forte densité

Presque tous les graphiques peuvent être réduits sans perte d'information
Plus la densité est faible, moins il est justifié de tracer un graphique, un tableau peut suffire



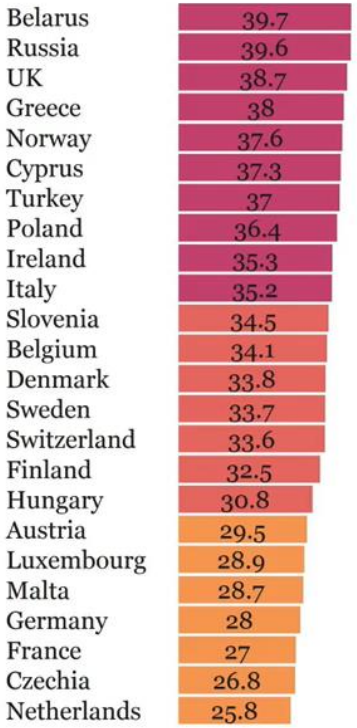

35
30
25
20
15
10
mpg
100
200
300
hp


35
30
25
20
15
10
mpg
100
200
300
hp

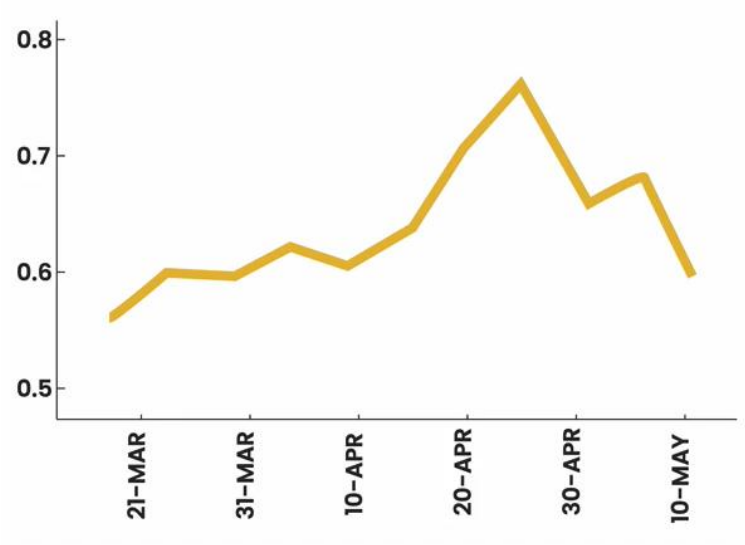
Réfléchir aux proportions du graphique

La plupart des graphiques peuvent être dessinés en format paysage : on a l'habitude de commencer la lecture d'un graphique par l'axe des abscisses

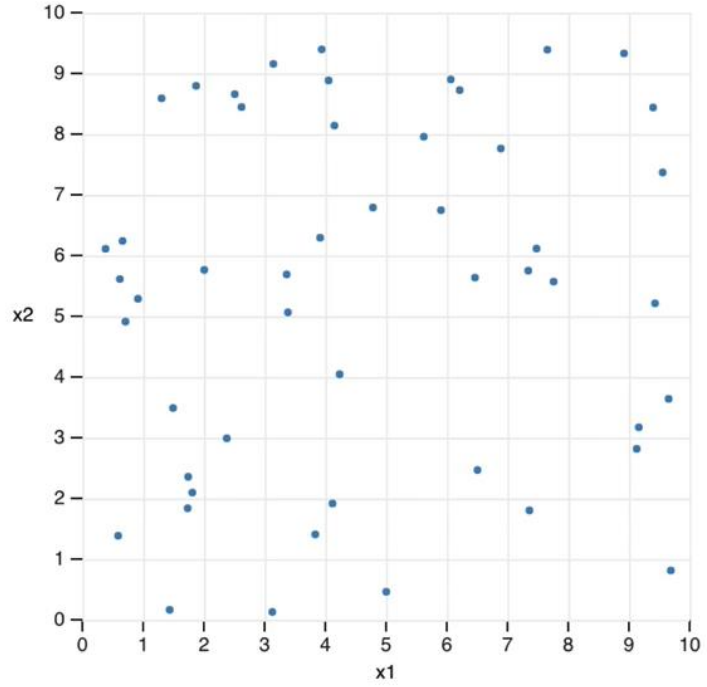
un graphique **étroit** suggère une lecture **verticale**



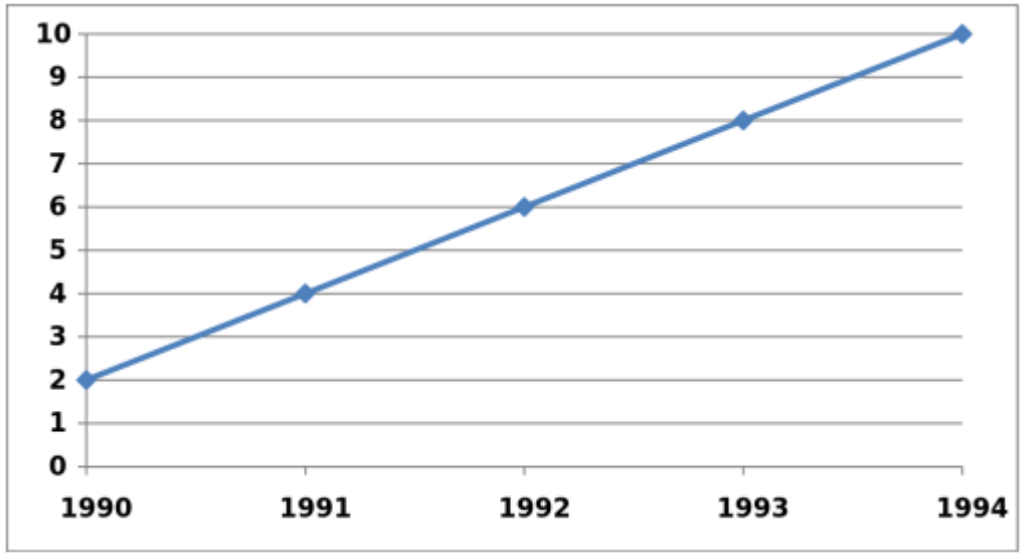
un graphique **large** suggère une lecture **horizontale**



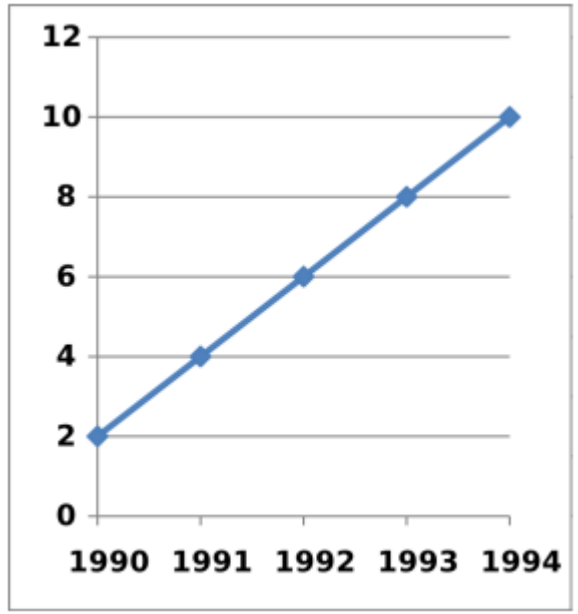
un graphique **carré** suggère moins le sens de lecture



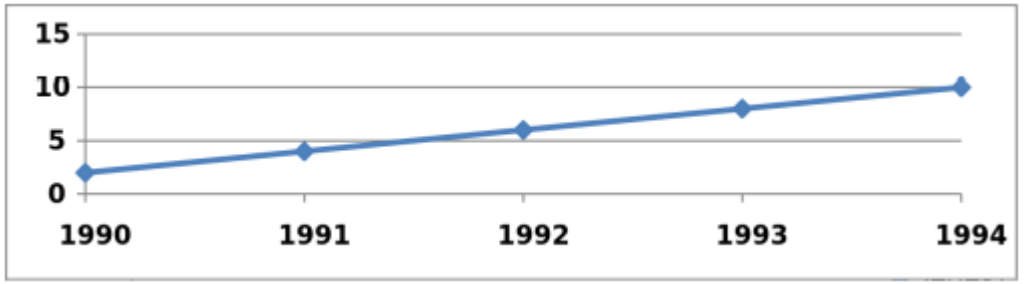
Les mêmes données avec des axes de taille différente



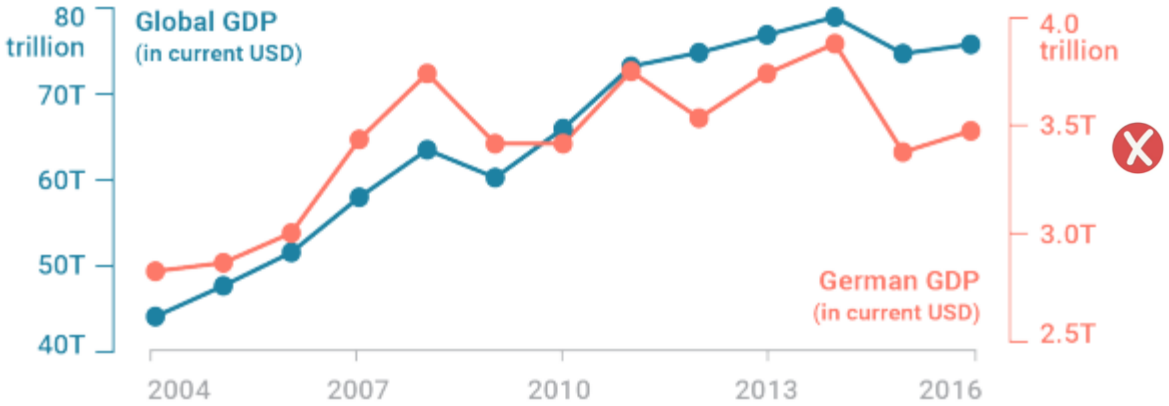
Un axe des abscisses rétréci



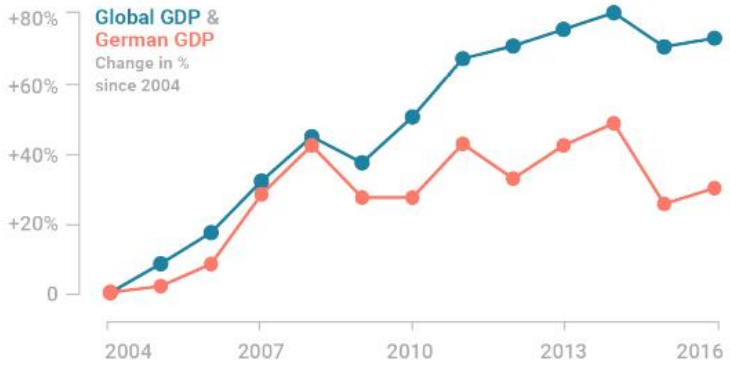
Un axe des ordonnées rétréci



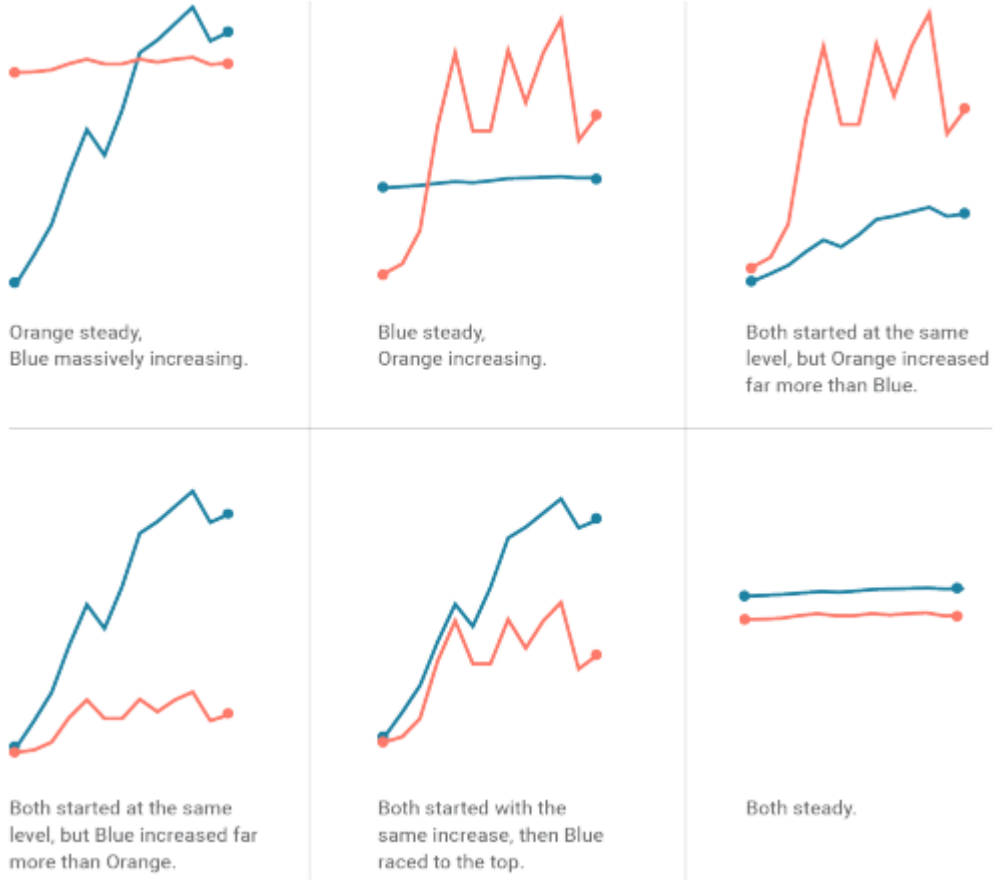
Eviter l'axe double



Solution : Indexes

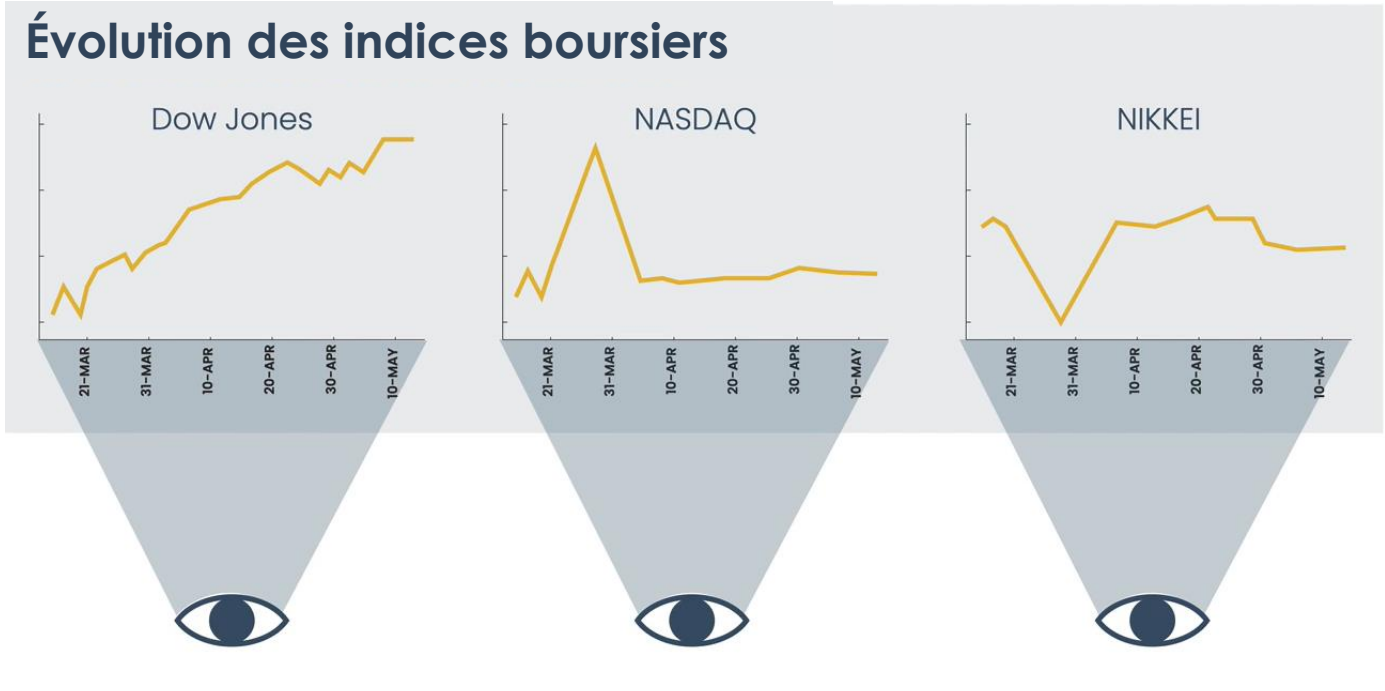


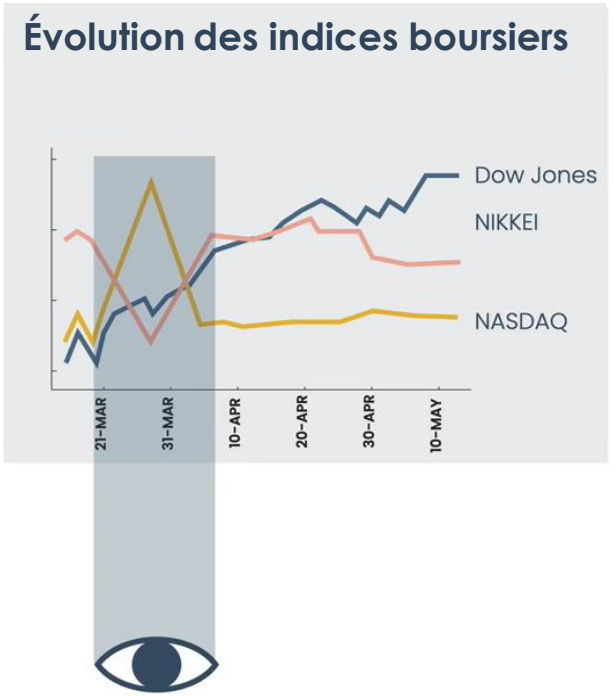
The German GDP and the global GDP **are not** growing at the same rate since 2008 !



Source : Lisa Charlotte Rost, data from World bank

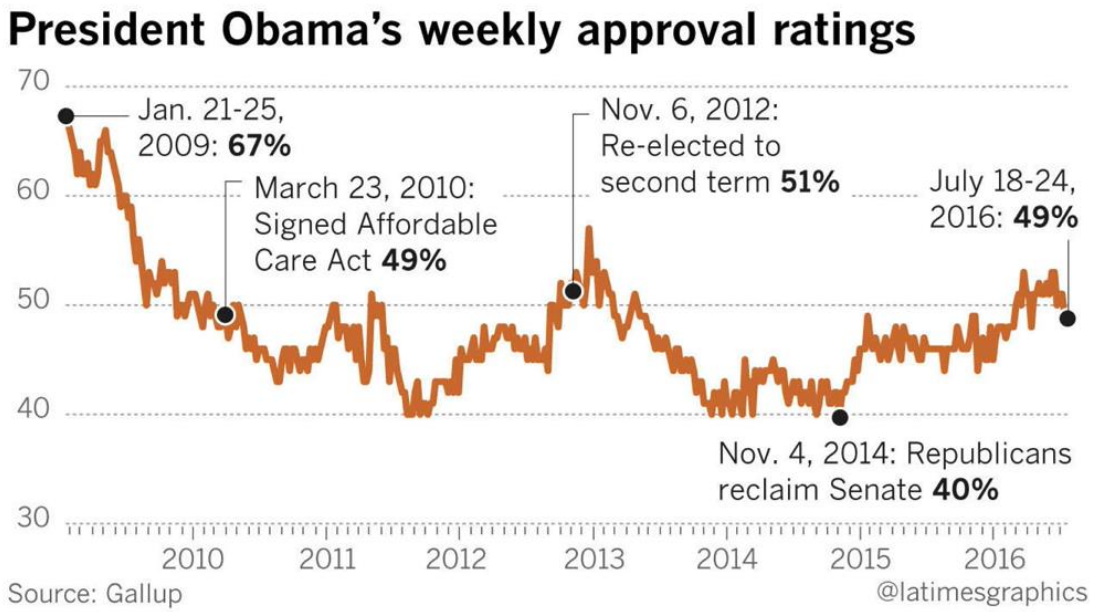
Inviter à comparer les données





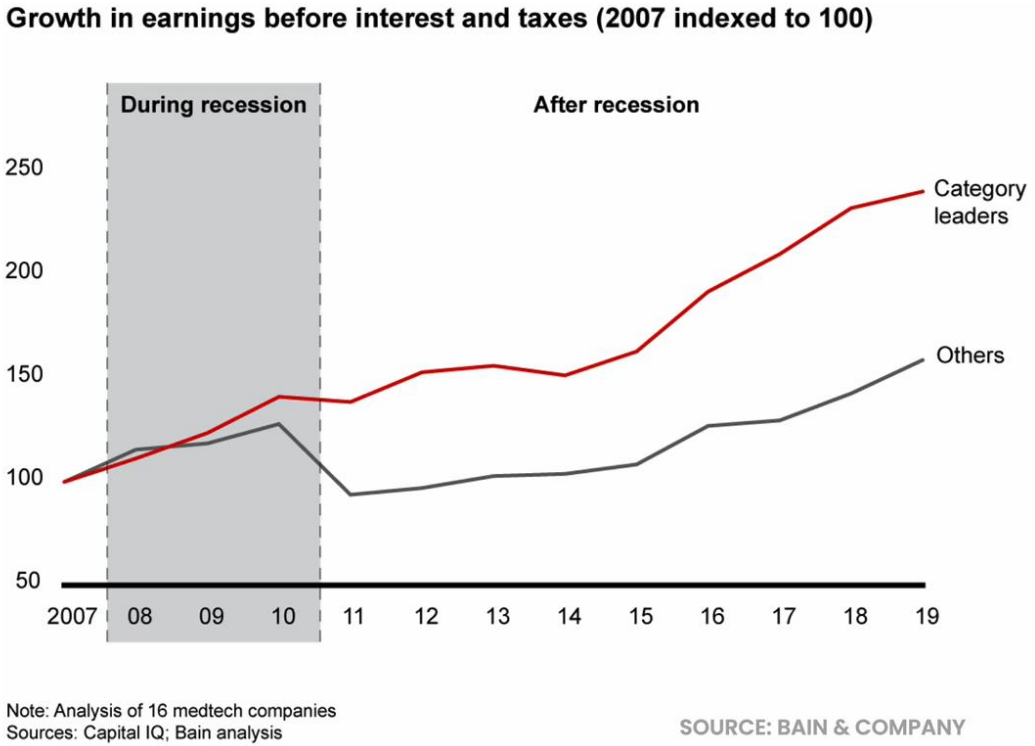
Annoter directement sur le graphique et étiqueter les événements importants des données

Étiqueter des points



SOURCE: THE BALTIMORE SUN

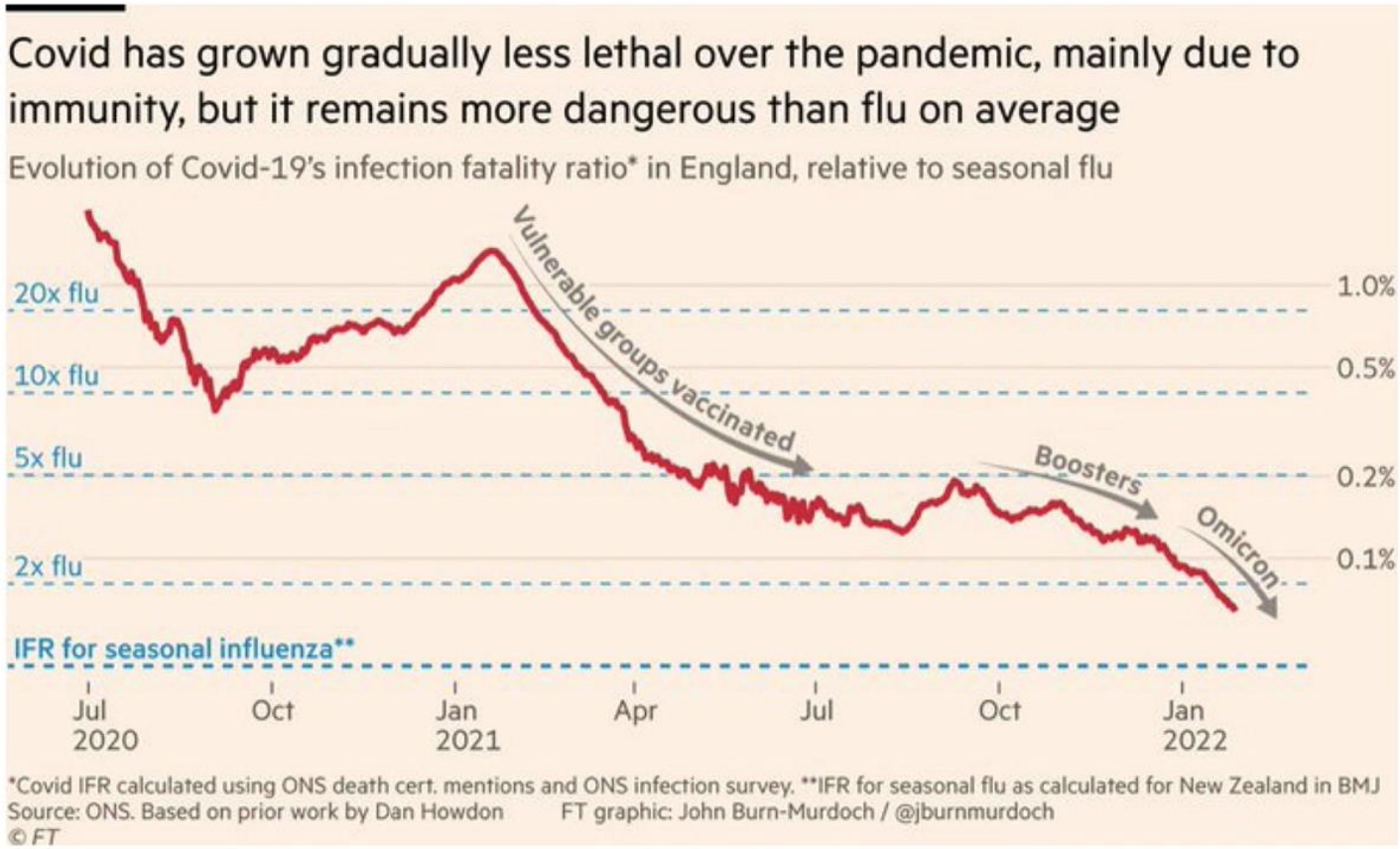
Étiqueter des zones



Note: Analysis of 16 medtech companies
Sources: Capital IQ; Bain analysis

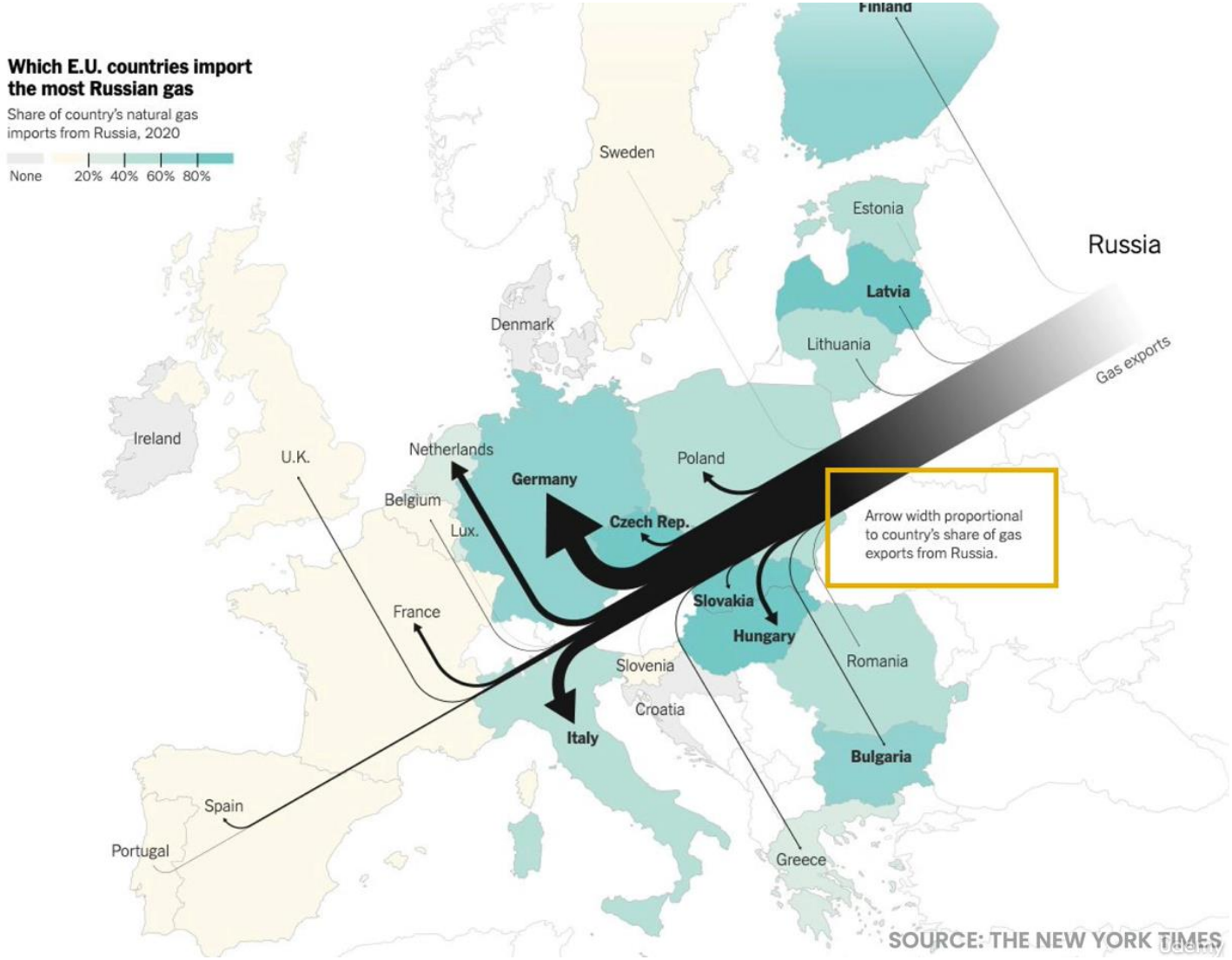
SOURCE: BAIN & COMPANY

Annoter des éléments d'explication



SOURCE: @JBURNMURDOCH

Annoter pour expliquer comment lire le graphique



Partie 2 – Choisir un graphique approprié à ses données

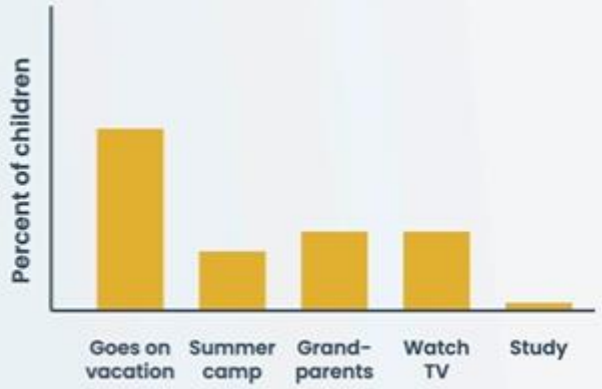
A-t-on vraiment besoin d'un graphique ?

Combien a-t-on de données ? Que peut-on utiliser?

- Une **phrase** suffit si l'on a pas plus de 2 données
- Un **tableau** si les valeurs numériques exactes sont importantes
- Un **graphique** si l'on veut comparer les données et qu'on s'intéresse à la tendance

In the summer holiday period,
43% of children go on vacation,
15% of children enroll in summer camps,
20% spend their time with their grandparents,
20% watch TV all day long and
only a 2% study for next year.

Go on vacation	43 %
Enroll in a summer camp	15 %
Go to grandparent's	20 %
Watch TV	20 %
Study for next year	2 %



Quel est l’objectif?

Distribution

Quand on a beaucoup d’observations sur une variable et que les statistiques de base (moyenne, médiane etc.) ne suffisent pas

Histogramme, diagramme de densité, boîte à moustaches, diagramme en violon, graphique en ligne de crête

---

Relation

Quand on s’intéresse à la relation entre 2 ou 3 variables

Courbe, nuage de points, carte de chaleur

Classement

Quand on veut comparer ou ordonner plusieurs variables

diagramme en barres, parallel plot, nuage de mots

Une partie d’un ensemble

Quand une variable est divisée en catégories et qu’on veut connaître quelle est la catégorie dominante sans forcément avoir besoin de comparer précisément les proportions des différentes catégories

Camembert, dendrogramme, carte proportionnelle

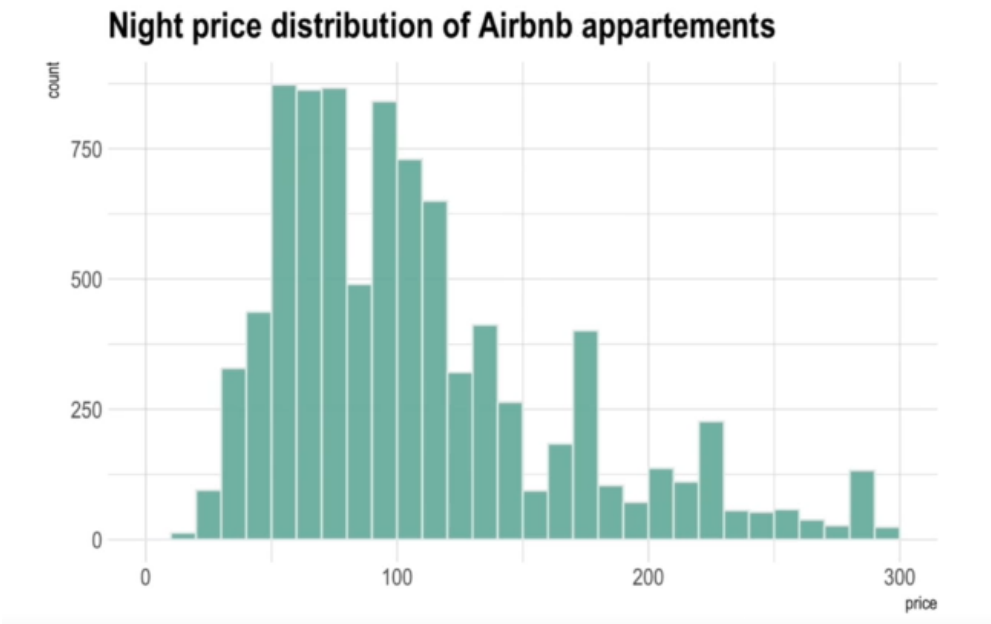
Carte

Dès lors qu’on a une variable géographique

Représenter une distribution

L’histogramme (histogram)

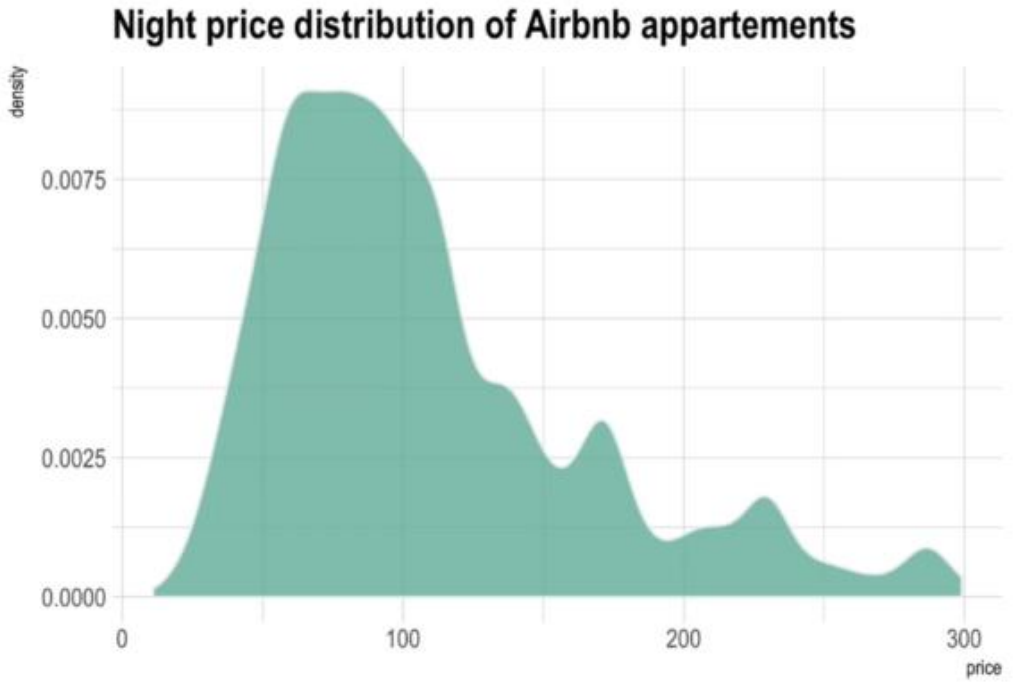
- L’histogramme est une représentation graphique de la distribution d’une variable numérique
- Les valeurs de la variable sont découpées en tranches et la hauteur de la barre représente le nombre d’observations dans la tranche
- Il est conseillé d’essayer plusieurs tailles de tranches avant de choisir la meilleure car le rendu et les conclusions peuvent être très différents

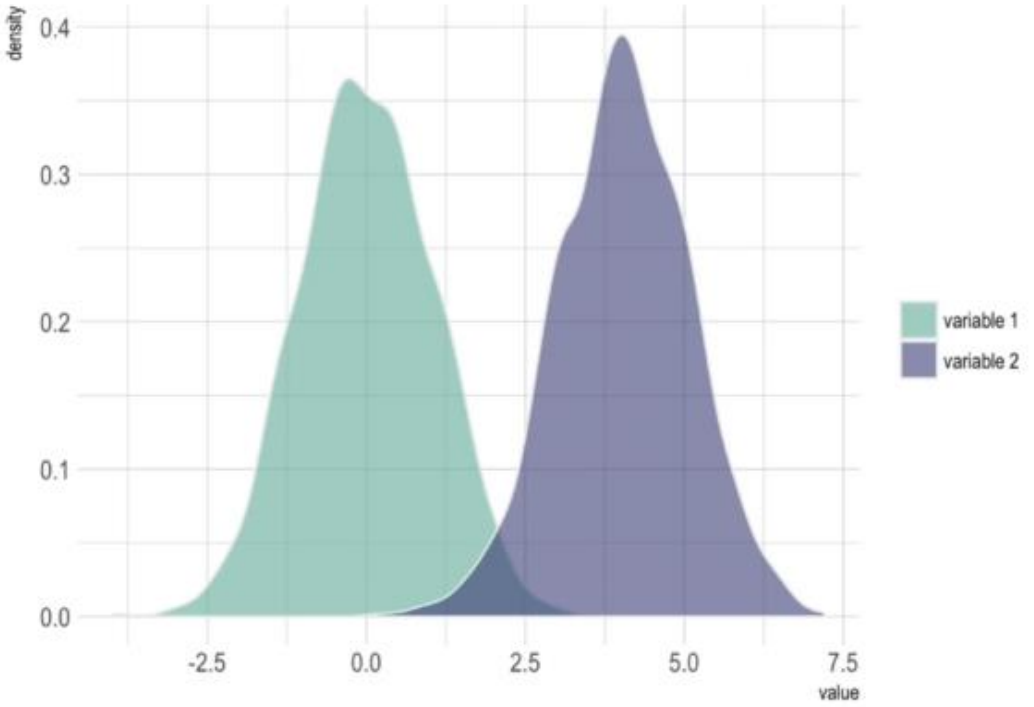


SOURCE: DATA-TO-VIZ

Le diagramme de densité (density plot)

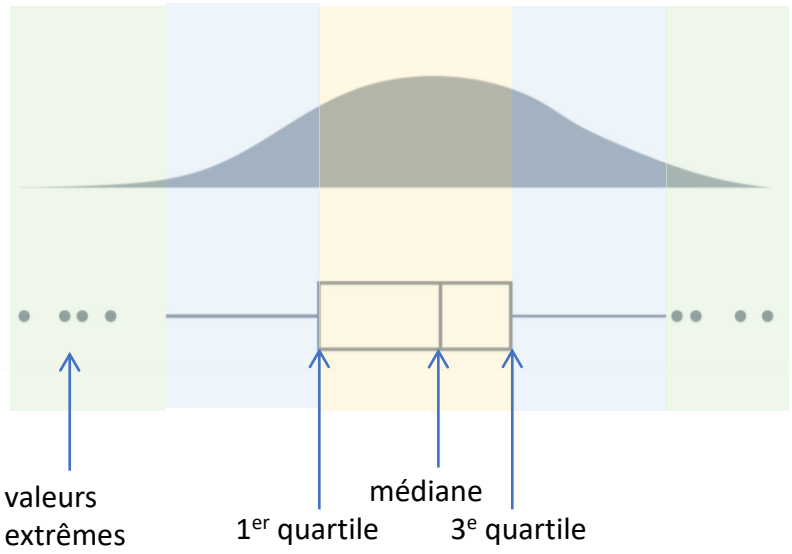
- La courbe de densité est la représentation graphique de la distribution d'une variable numérique **continue**
- Elle peut être un bon moyen de comparer des distributions

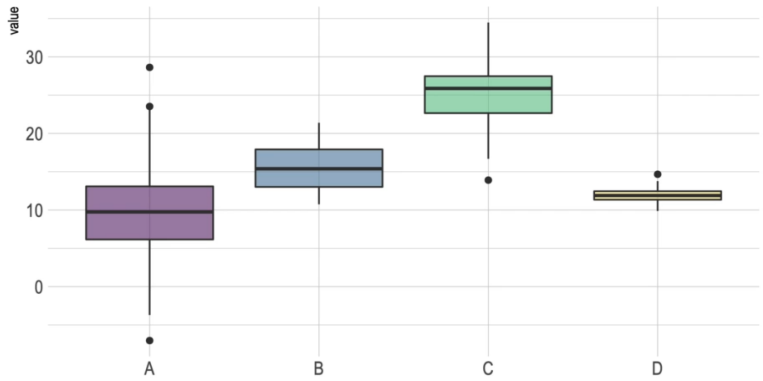


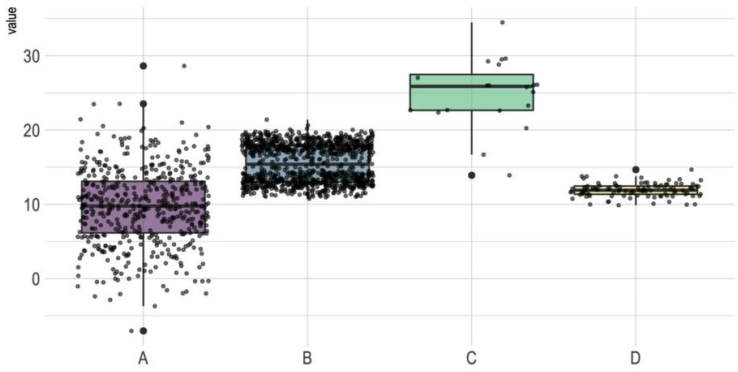


La boîte à moustaches (boxplot)

- La boîte à moustaches permet d'afficher les principales statistiques qui résument la distribution d'une variable numérique
- Elle permet de comparer la distribution d'une variable numérique pour différents groupes

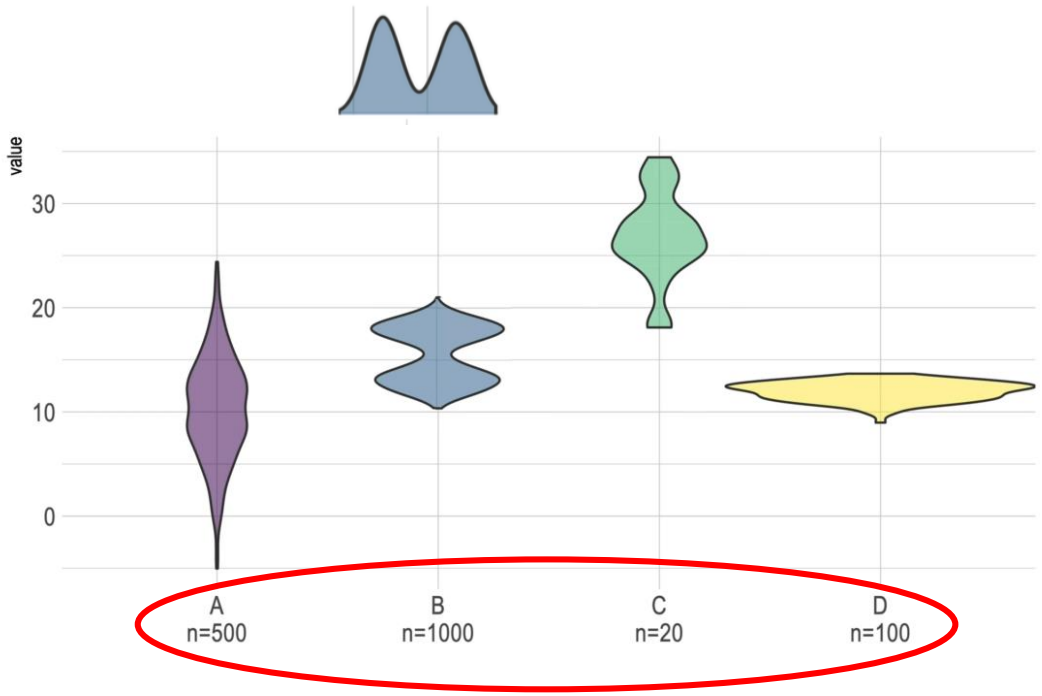


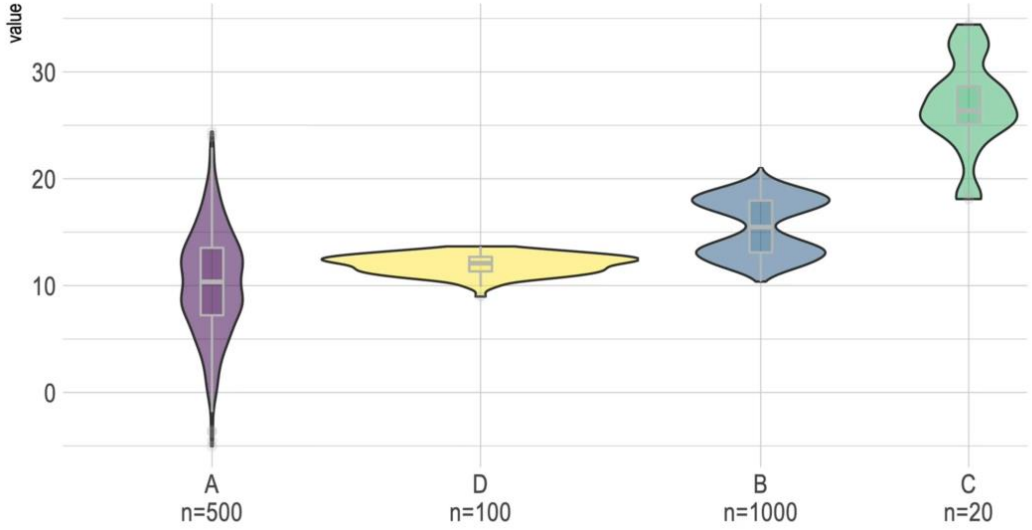




Le diagramme en violon (violin plot)

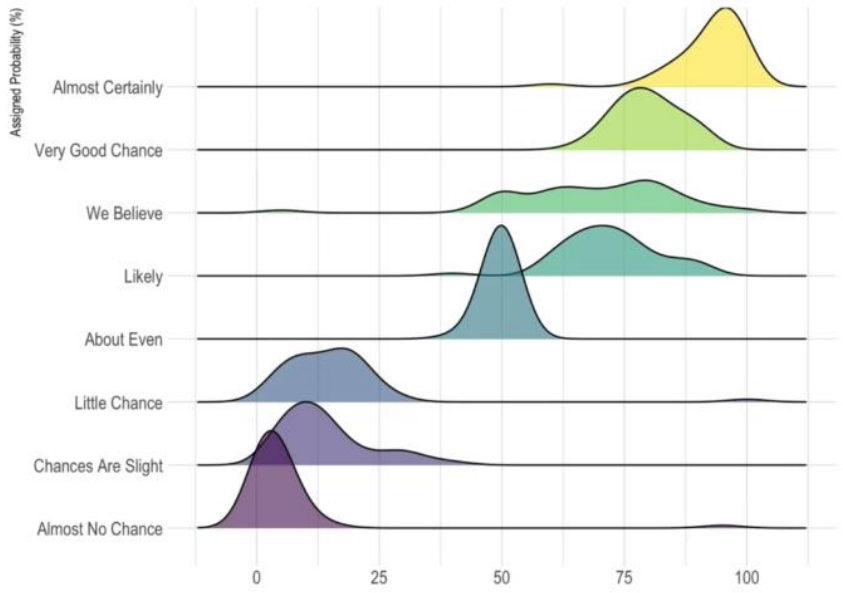
- Un diagramme en violon est un diagramme de densité rendue symétrique
- Il est adapté quand on a un grand volume de données





Le graphique en ligne de crête (ridgeline plot)

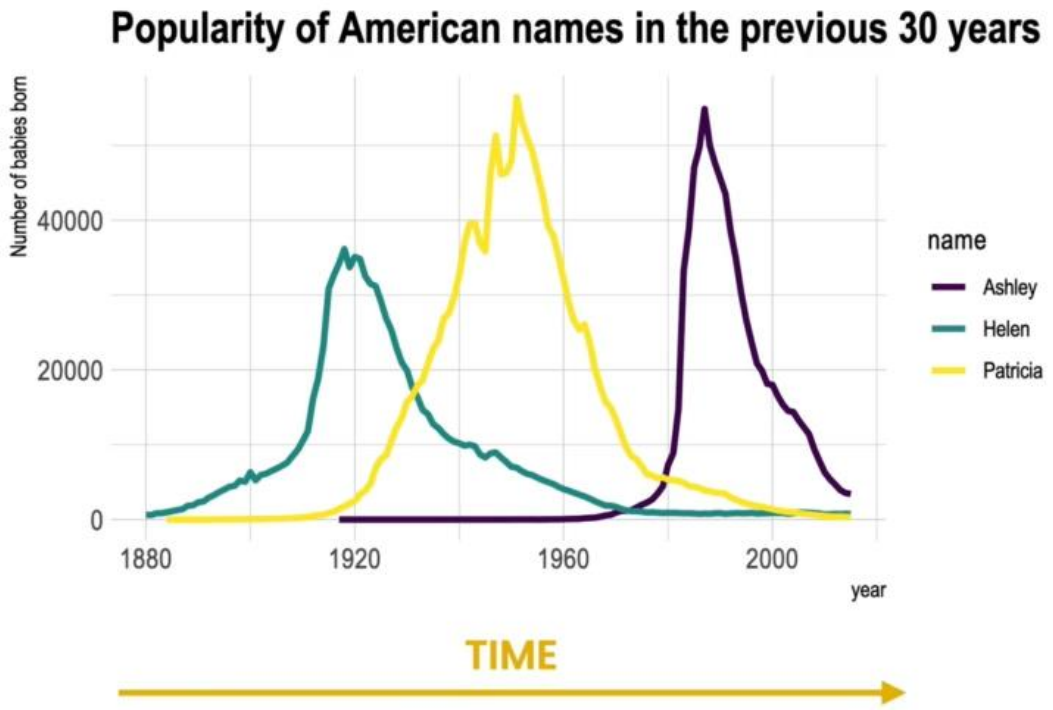
- Il montre la distribution d'une variable numérique pour plusieurs groupes
- L’inconvénient est qu’il ne fonctionne pas si les distributions se chevauchent



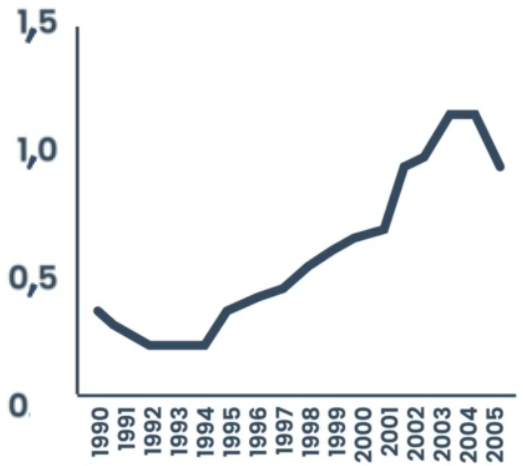
Représenter les relations entre les variables

La courbe (line plot)

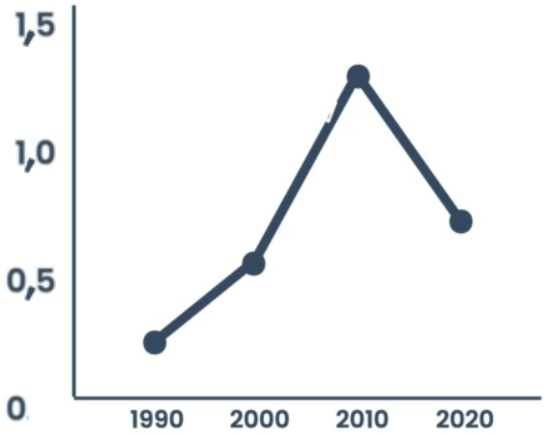
- Elle permet de visualiser l'évolution d'une ou plusieurs variables numériques
- Elle est souvent utilisée pour visualiser une tendance dans les données sur une période de temps
- Attention, **la variable sur l'axe des abscisses doit être numérique et continue**



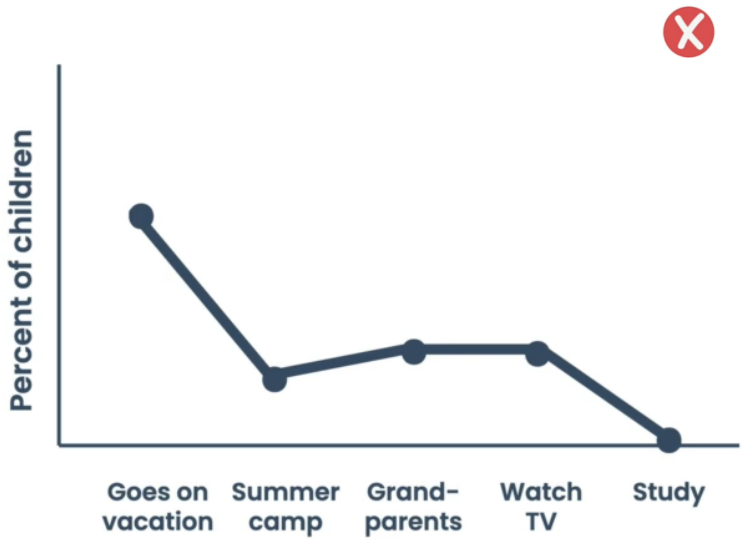
Variable numérique continue



Variable numérique discrète



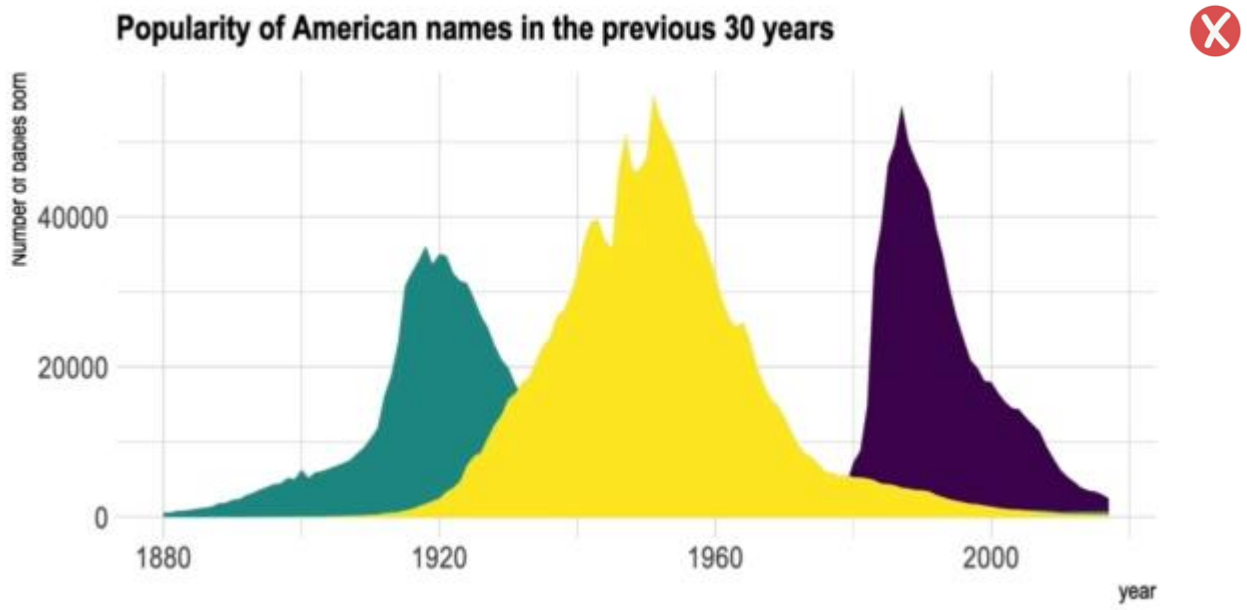
Variable catégorielle



Nuage de points connectés

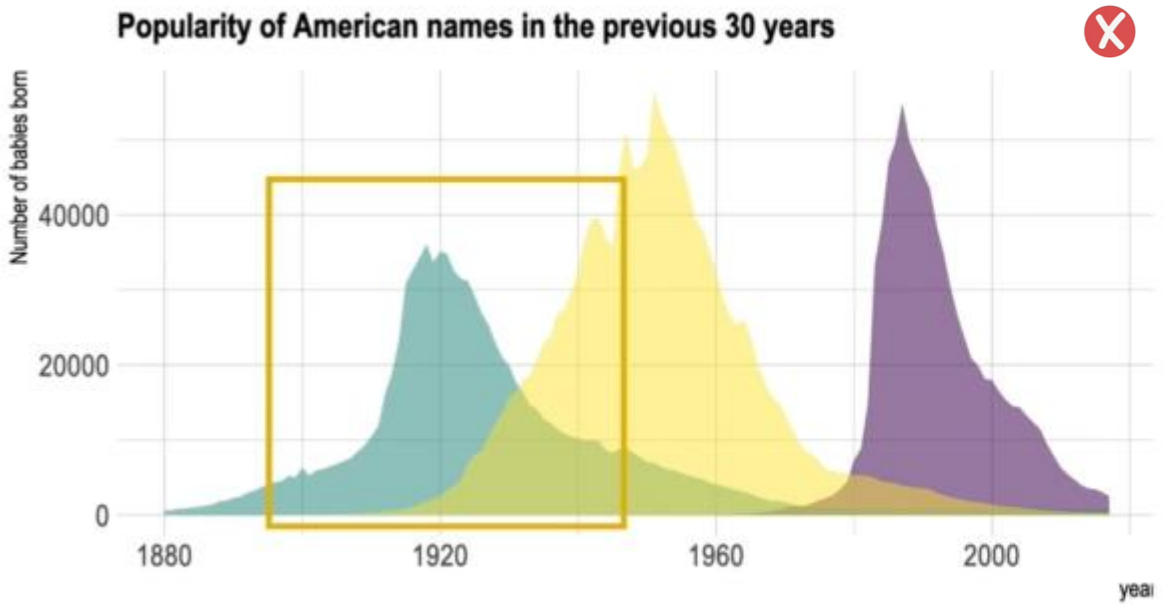
Le graphique à aires (area plot)

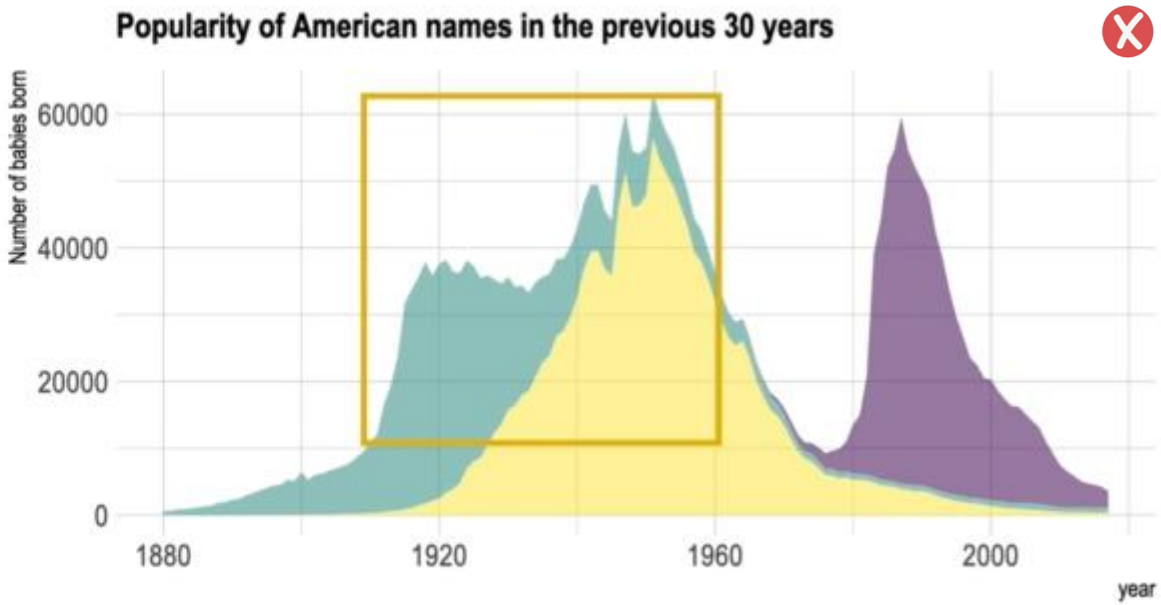
- Approprié pour représenter une distribution mais pas pour représenter une relation entre deux variables
- Ne respecte pas le principe d'un minimum d'encre



Le graphique à aires empilées (stack area plot)

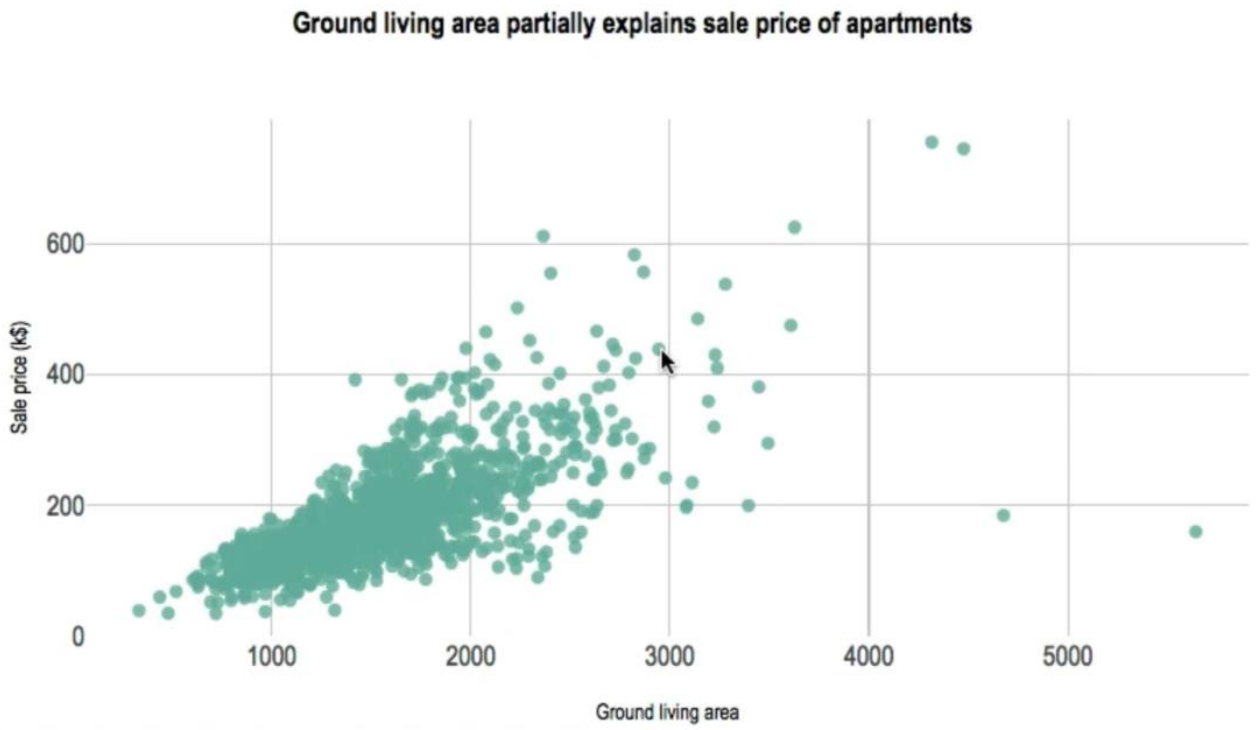
- La distribution de la variable verte ne commence pas sur la ligne horizontale de l'axe. Elle commence précisément là où la variable jaune se termine.
- Donc si l'on veut connaître la distribution de la variable verte on doit soustraire les valeurs vertes des valeurs jaunes, ce qui est impossible à faire à l'œil nu





Le nuage de points (scatter plot)

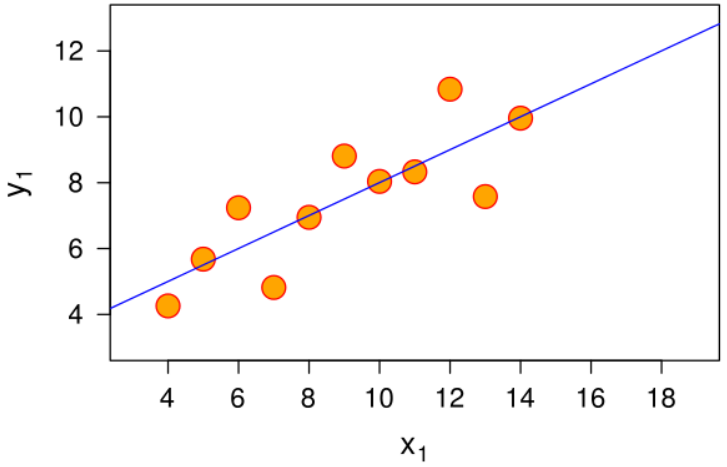
- Affiche la relation entre deux variables numériques à l'aide de simples points
- Pour chacun des points, la valeur de la 1re variable est représentée sur l'axe des abscisses et la valeur de la 2nde sur l'axe des ordonnées
- Très utile pour révéler le type de relation entre deux variables

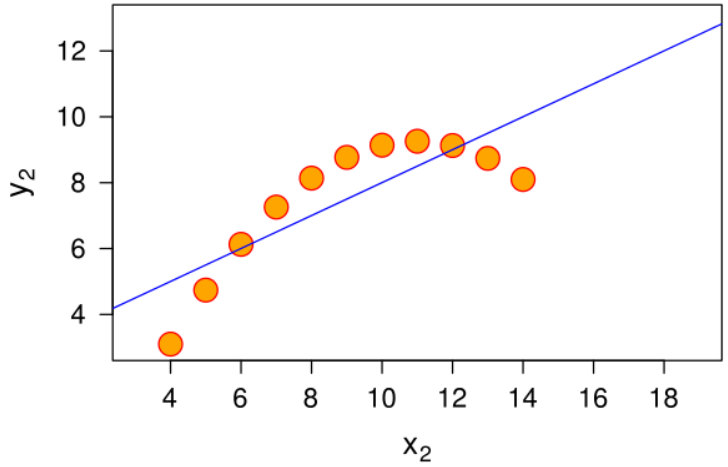


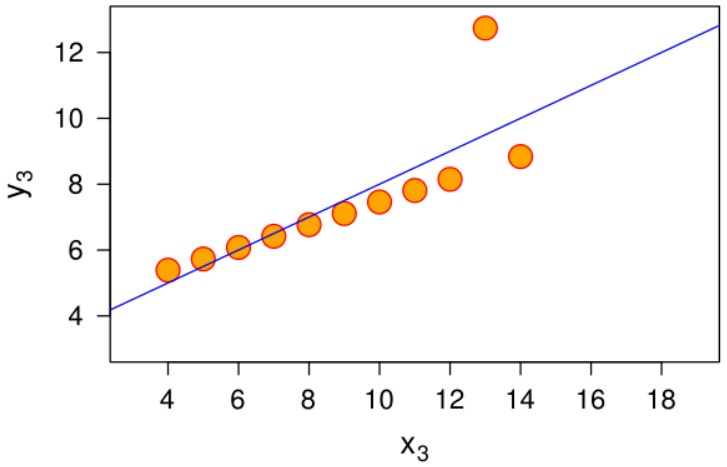
Le quartet d'Anscombe

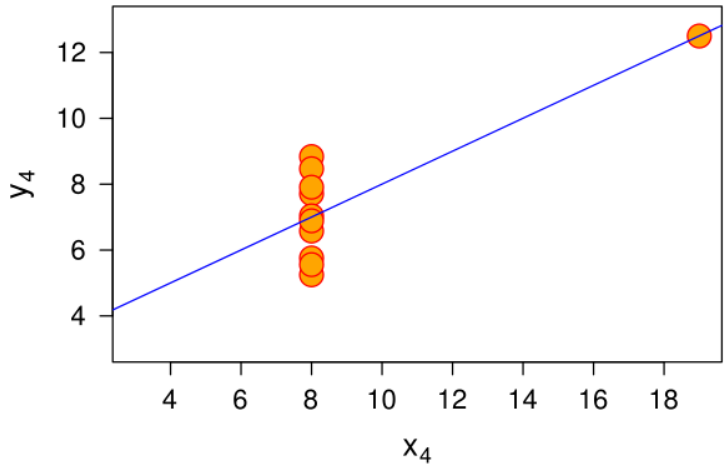
Ces quatre ensembles de données ont des statistiques descriptives simples presque identiques

Mais leurs distributions sont très différentes. On le voit par leur représentation graphique



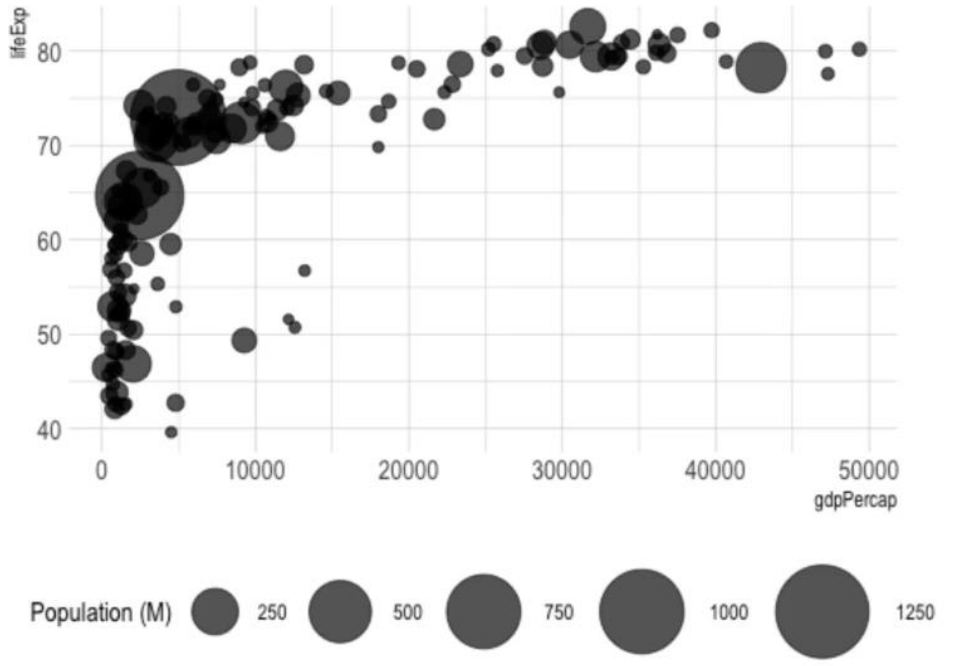






Le graphique à bulles (bubble plot)

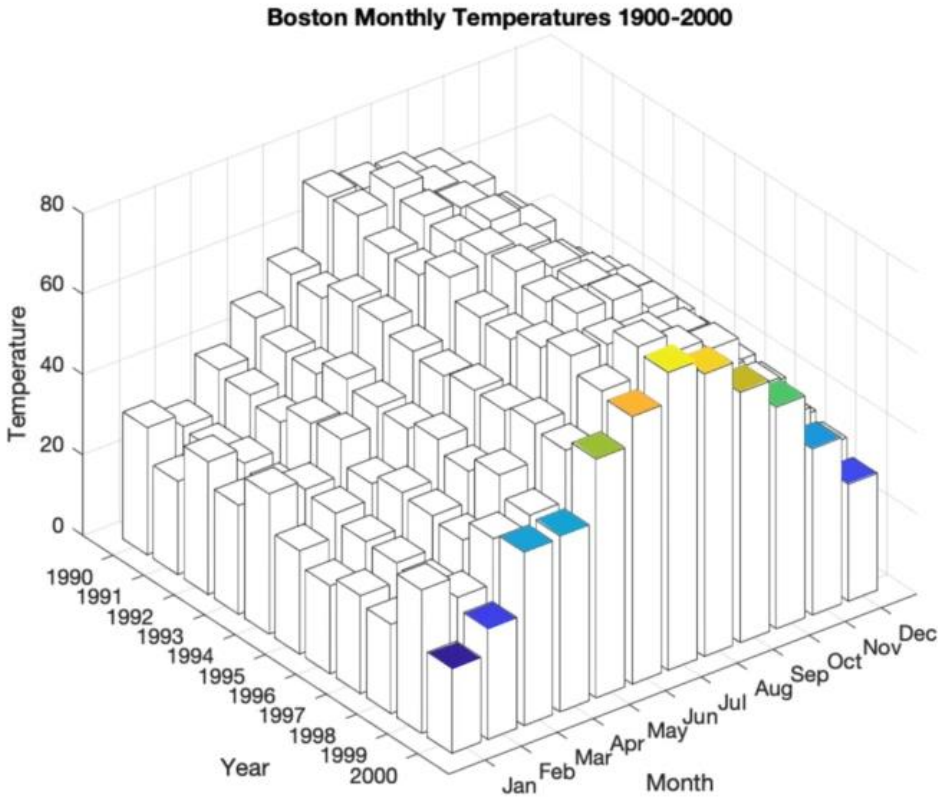
- C’est un nuage de points auquel on ajoute une 3e dimension
- La valeur de cette 3e dimension numérique est représentée par la surface du point
- La 3e variable représente souvent une pondération

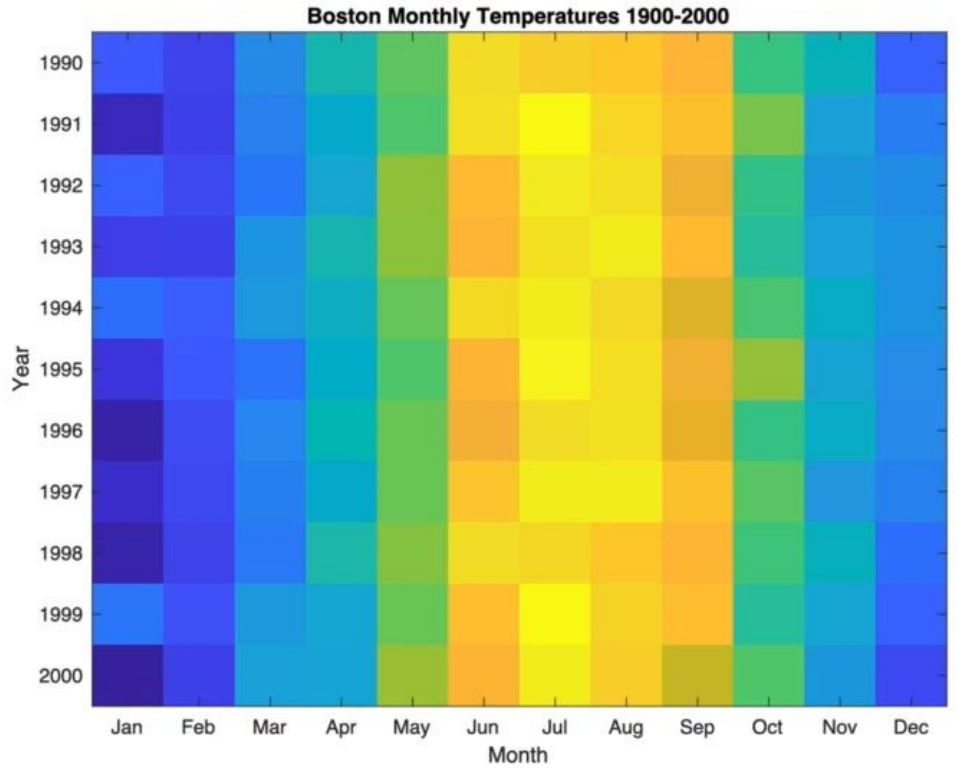


SOURCE: DATA-TO-VIZ

La carte de chaleur (heatmap)

- C’est une représentation graphique de données où les valeurs contenues dans une matrice sont représentées par des couleurs.
- C'est comme un tableau mais sans se soucier des valeurs exactes, ce qui intéresse c’est le modèle
- C’est une bonne alternative à la 3D

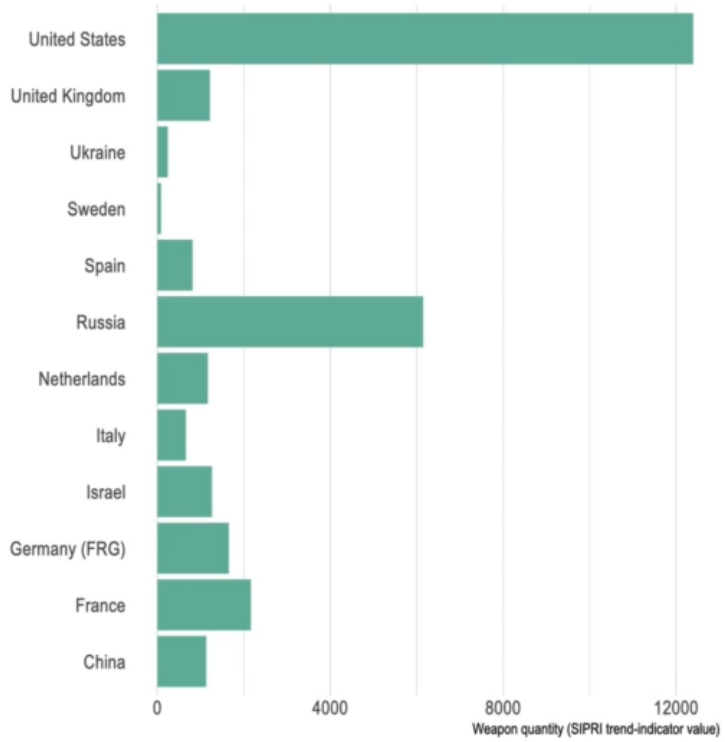


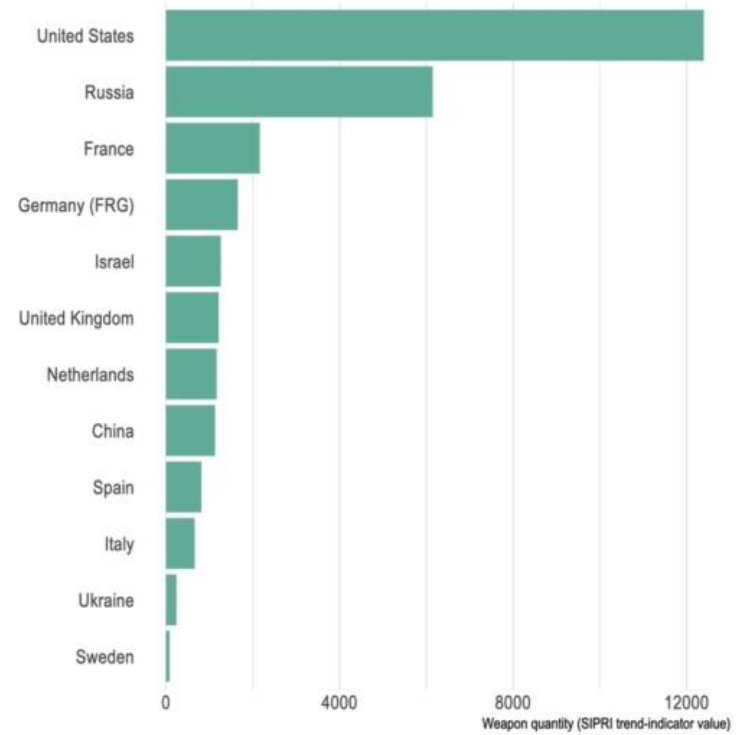


Représenter le classement

Le diagramme en barres

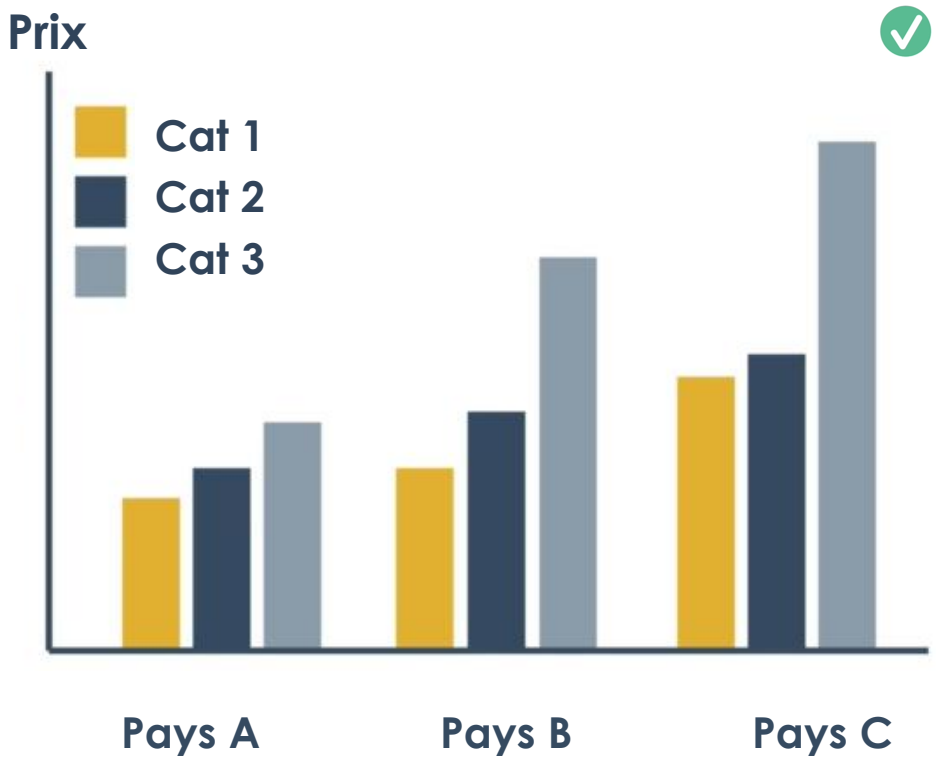
- Il montre la relation entre une variable numérique et une variable catégorielle
- Chaque catégorie est représentée par une barre
- C'est la **longueur** de la barre qui est proportionnelle à la valeur, pas la surface

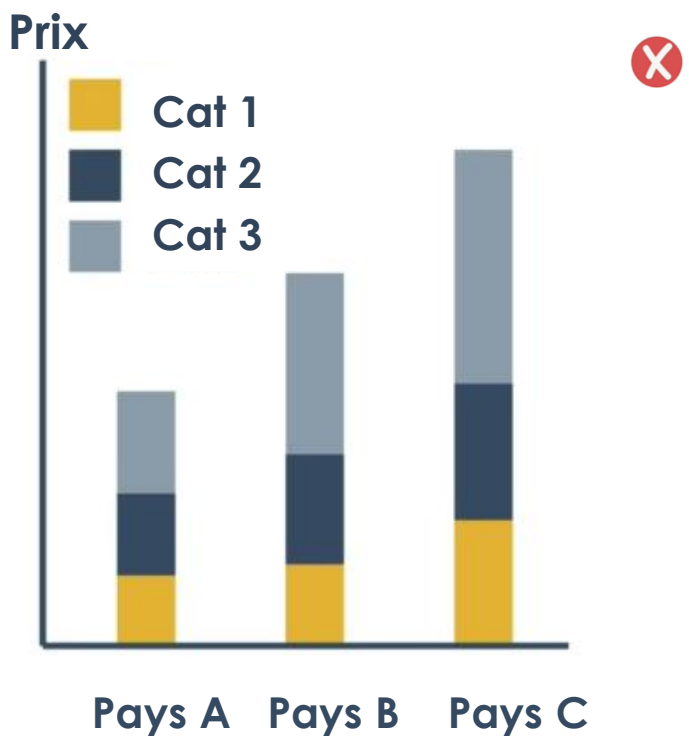




SOURCE: DATA-TO-VIZ

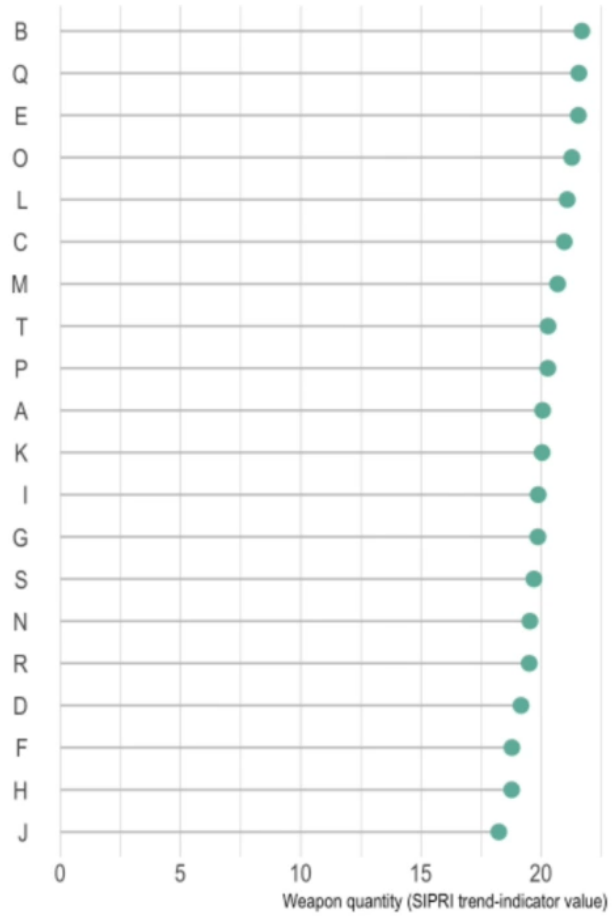
- Si l’on a plusieurs observations pour chacune des variables catégorielles, on doit utiliser un **diagramme à barres groupées**
- Éviter d’utiliser le diagrammes en barres empilées

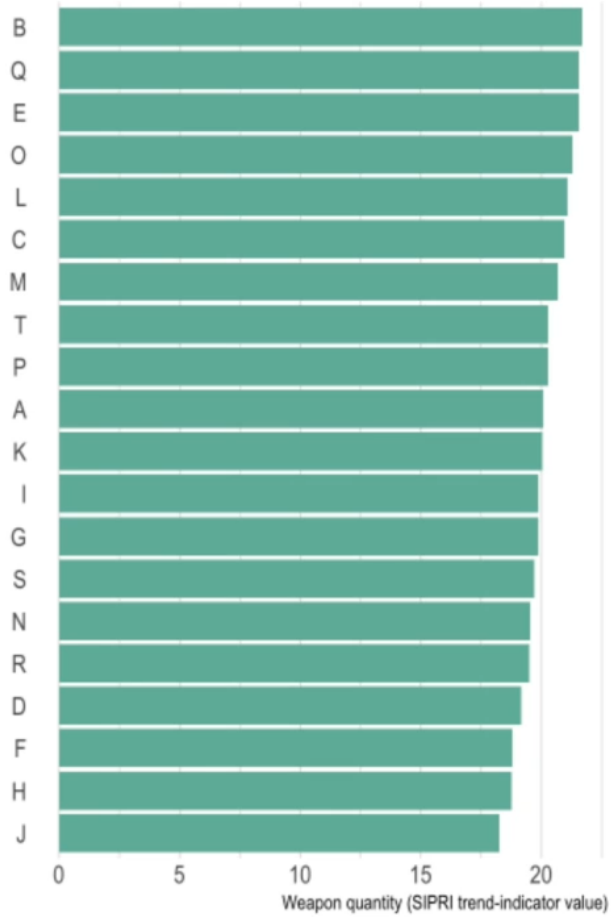




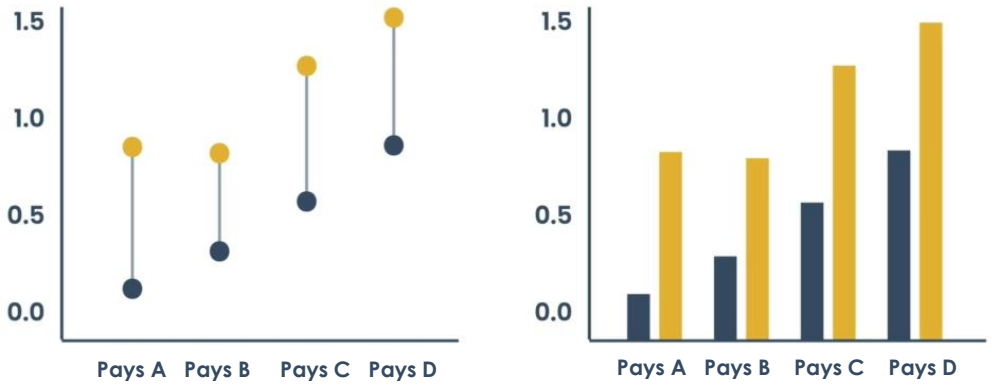
Le diagramme en sucettes

- C’est une bonne alternative au diagramme en barres, surtout quand ce dernier est dense et avec des valeurs très similaires
- Il respecte le principe de l’encre minimale
- L'ajout d'un point incite le lecteur à comparer la position sur une échelle commune, au lieu de comparer les longueurs





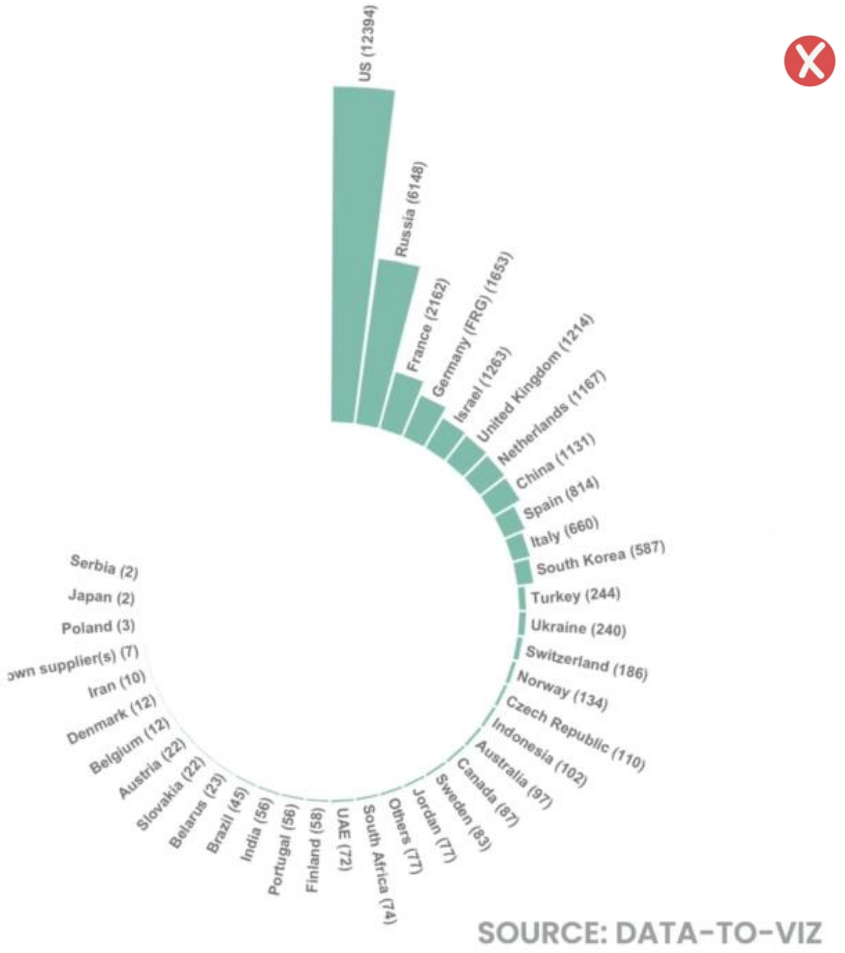
Le diagramme à points de Cleveland



SOURCE: HTTPS://UC-R.GITHUB.IO/

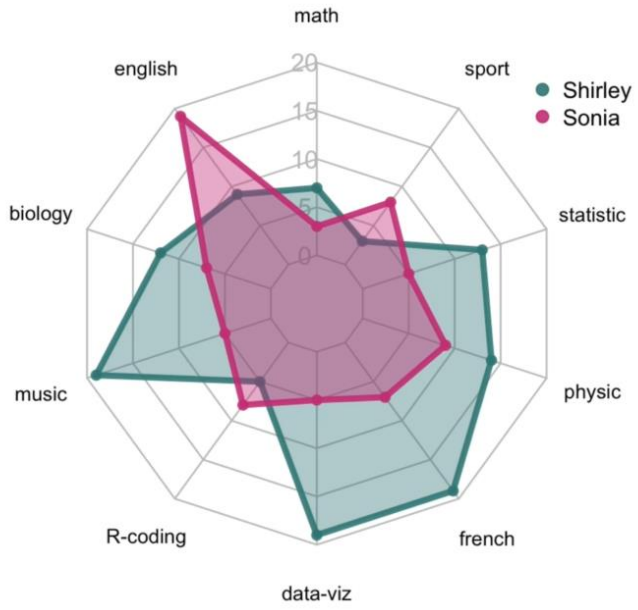
Le diagramme en barres circulaire (circular barchart)

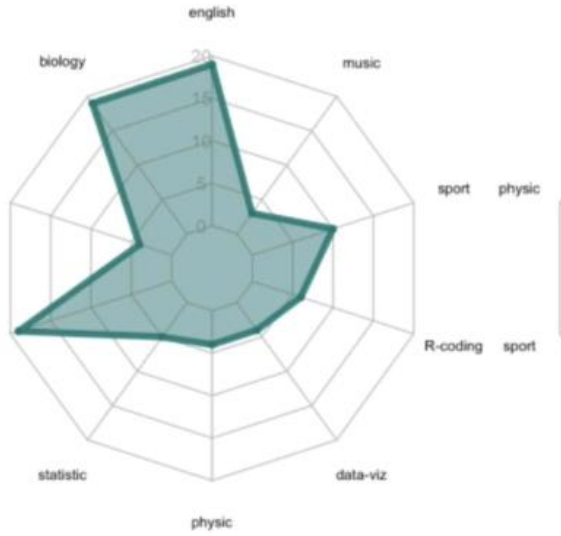
- C’est un diagramme en barres où l'axe des x est circulaire
- Il ne permet pas de comparer correctement les valeurs

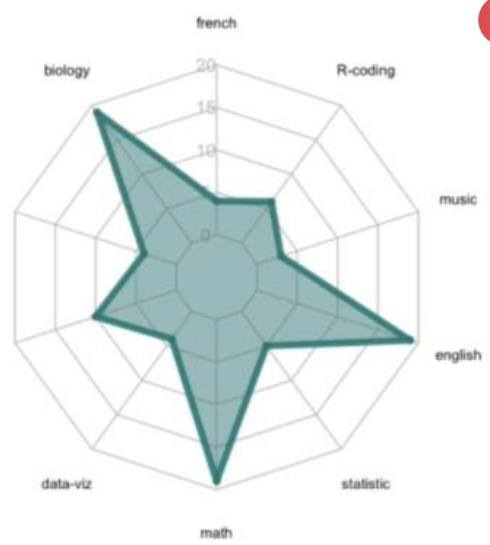


Le diagramme en toile d’araignée (radar plot)

- Le lecteur se concentre sur les surfaces et non sur les points
- La surface varie selon l’ordre des variables







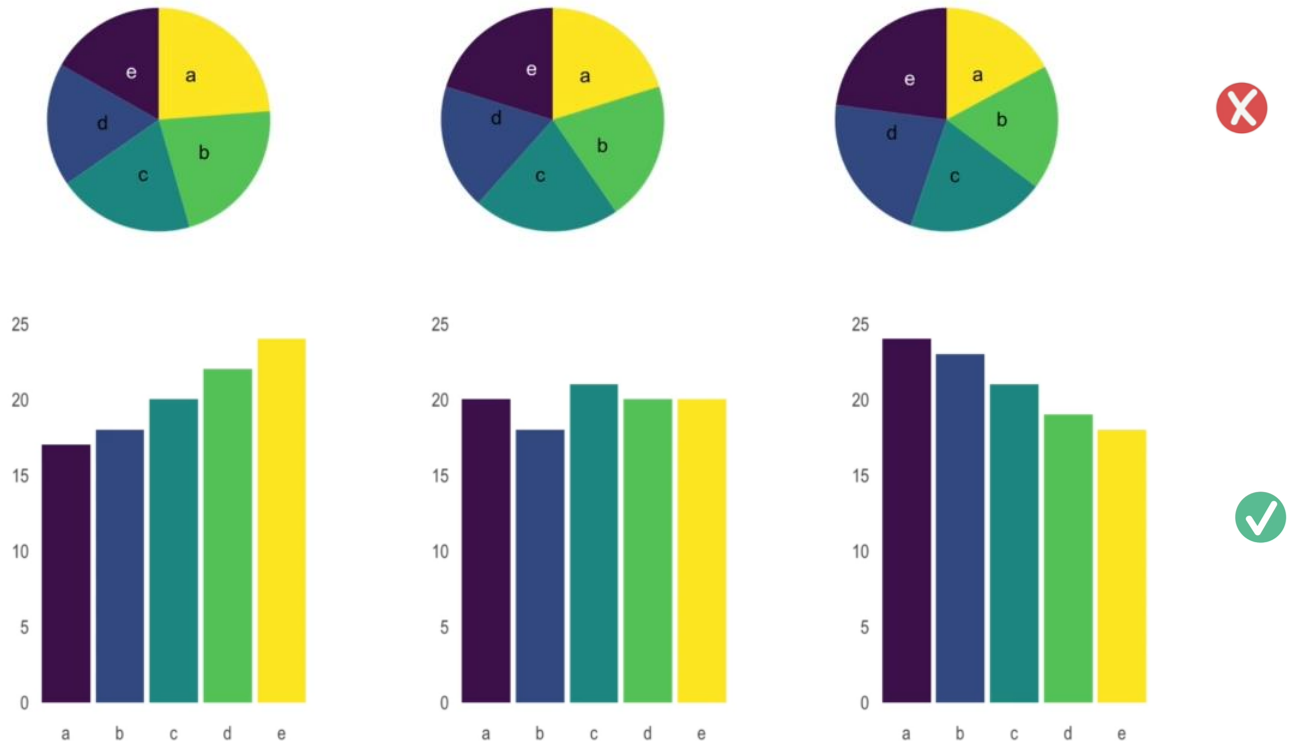
Le nuage de mots (word cloud)

- Il affiche une liste de mots et leur taille de police est proportionnelle au nombre de fois où ils apparaissent dans un texte ou dans une base de données
- Les mots les plus longs apparaissent plus gros puisqu'ils sont composés de plus de lettres
- La perception diffère selon l’orientation des mots

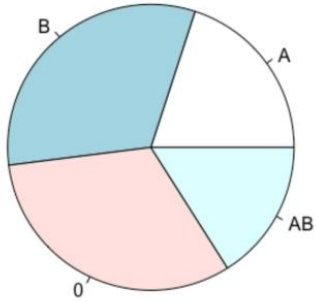
Représenter une part dans un ensemble

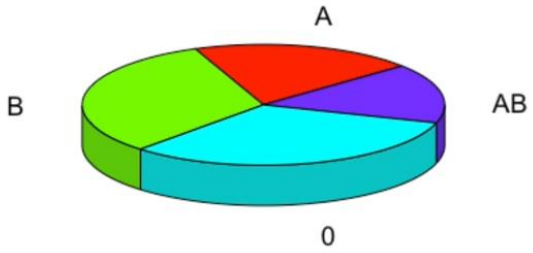
Le camembert (pie chart)

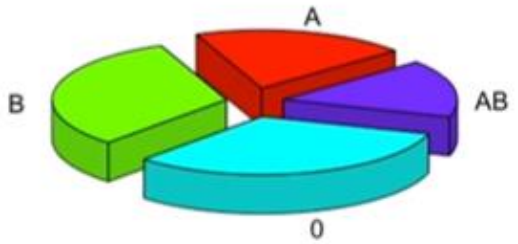
- C’est un cercle divisé en secteurs, tels que chaque secteur représente une proportion de l’ensemble
- Il est souvent utilisé pour montrer une proportion où la somme des secteurs est égale à 100 %
- Il convient surtout si l’on a seulement 2 parts à représenter

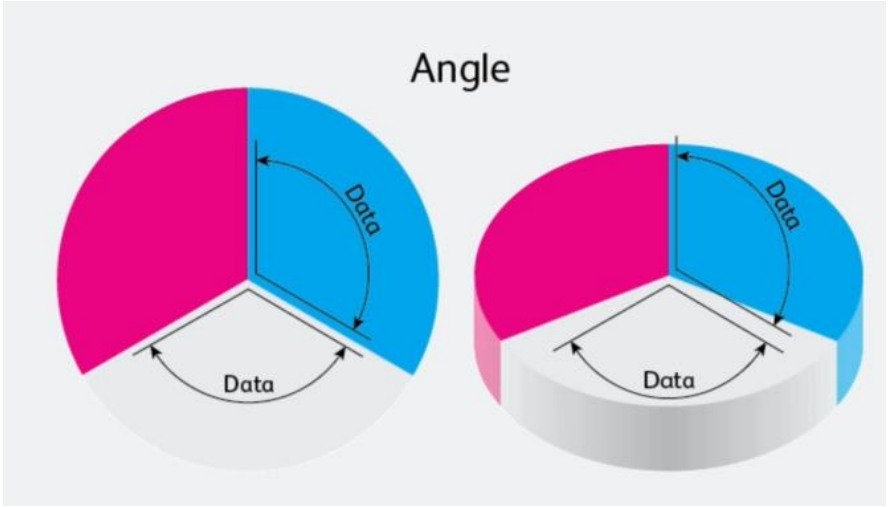


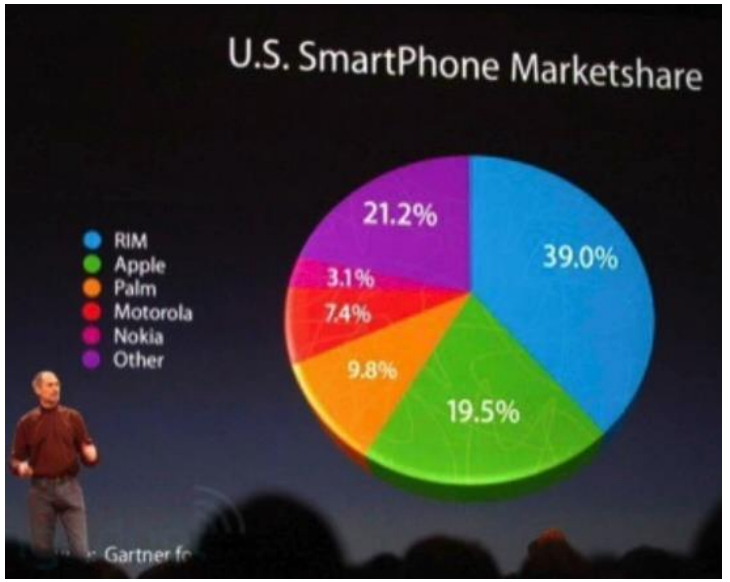
Le camembert en 3D





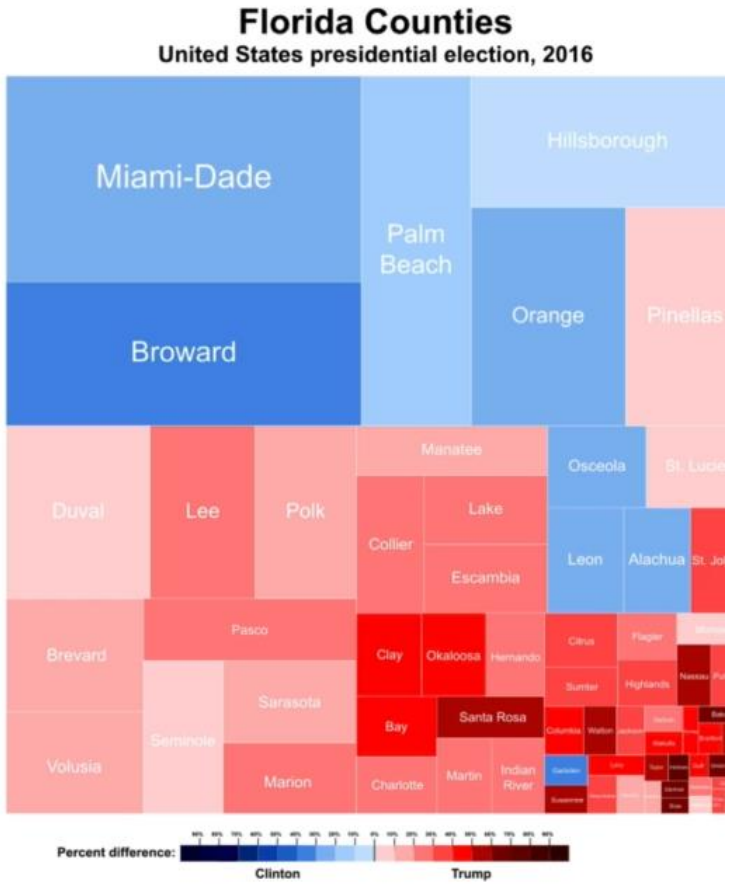






La carte proportionnelle (treemap)

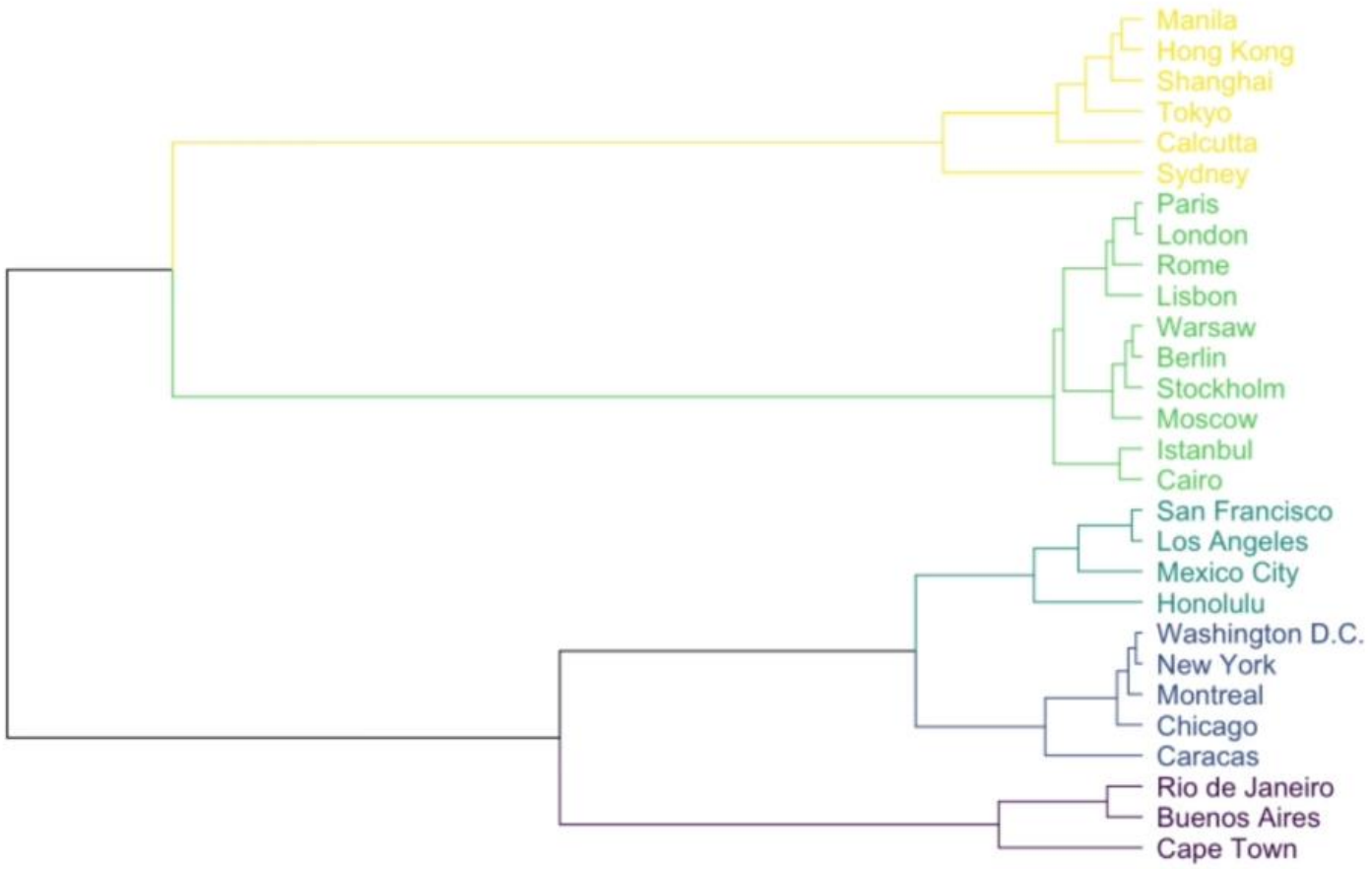
- Elle affiche des données hiérarchiques sous la forme d'un ensemble de rectangles imbriqués
- Chaque groupe est représenté par un rectangle dont la surface est proportionnelle à sa valeur.



SOURCE: ALI ZIFAN, WIKIPEDIA

Le dendrogramme (dendrogram)

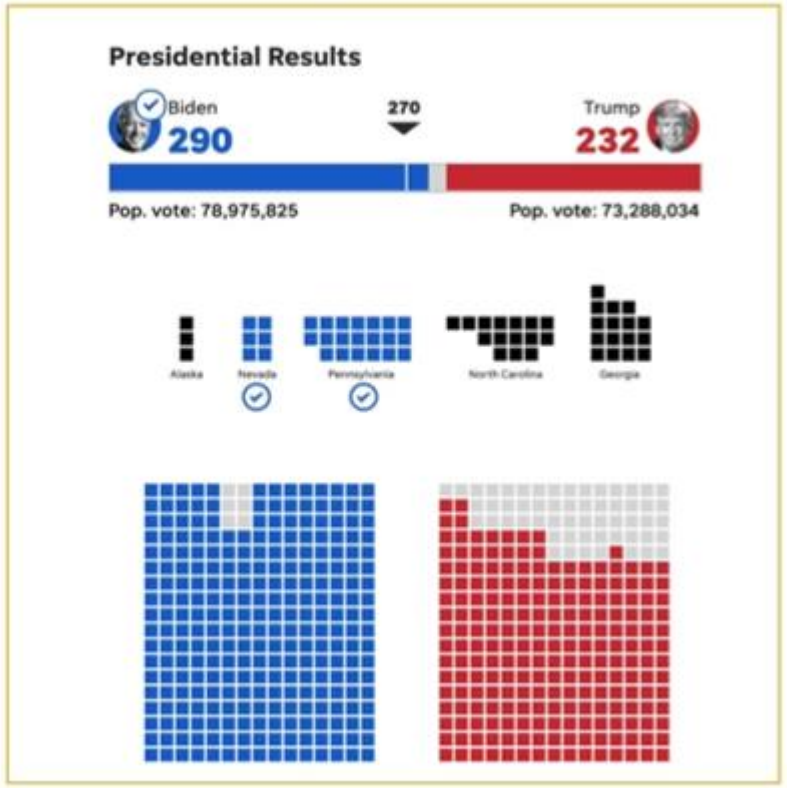
- Un dendrogramme est un diagramme qui représente une relation hiérarchique entre des catégories.
- Chaque groupe est représenté par un rectangle dont la surface est proportionnelle à sa valeur.



Approche moderne mixte

- Utiliser un diagramme en barres empilées quand on a deux catégories est une bonne idée
- Chaque groupe est représenté par un rectangle dont la surface est proportionnelle à sa valeur.

USA TODAY



BBC

Références

- Mastering Data Vizualisation : Theory and Foundations, Clara Granell, 2022
- data to Viz : https://www.data-to-viz.com
- Toulouse dataviz : https://toulouse-dataviz.fr/
- The visual display of quantitative information, E. Tufte, 2001

Merci pour votre attention