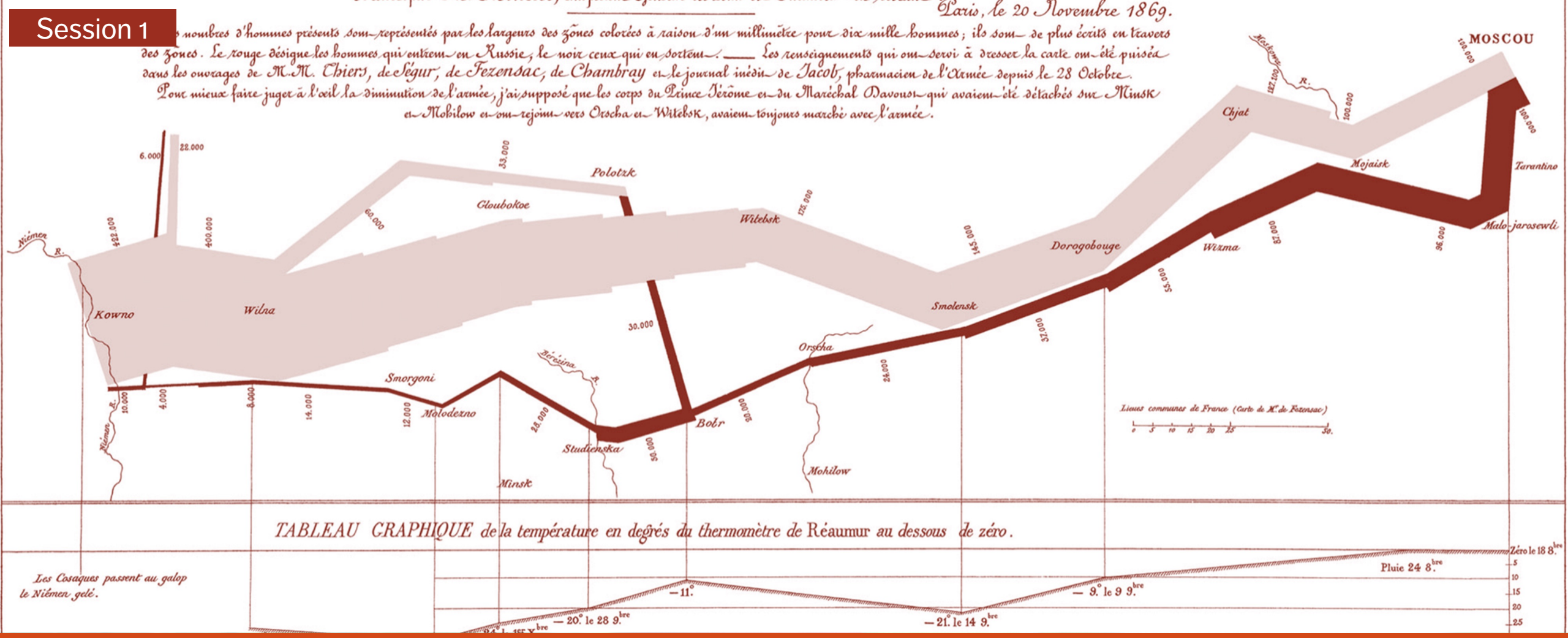
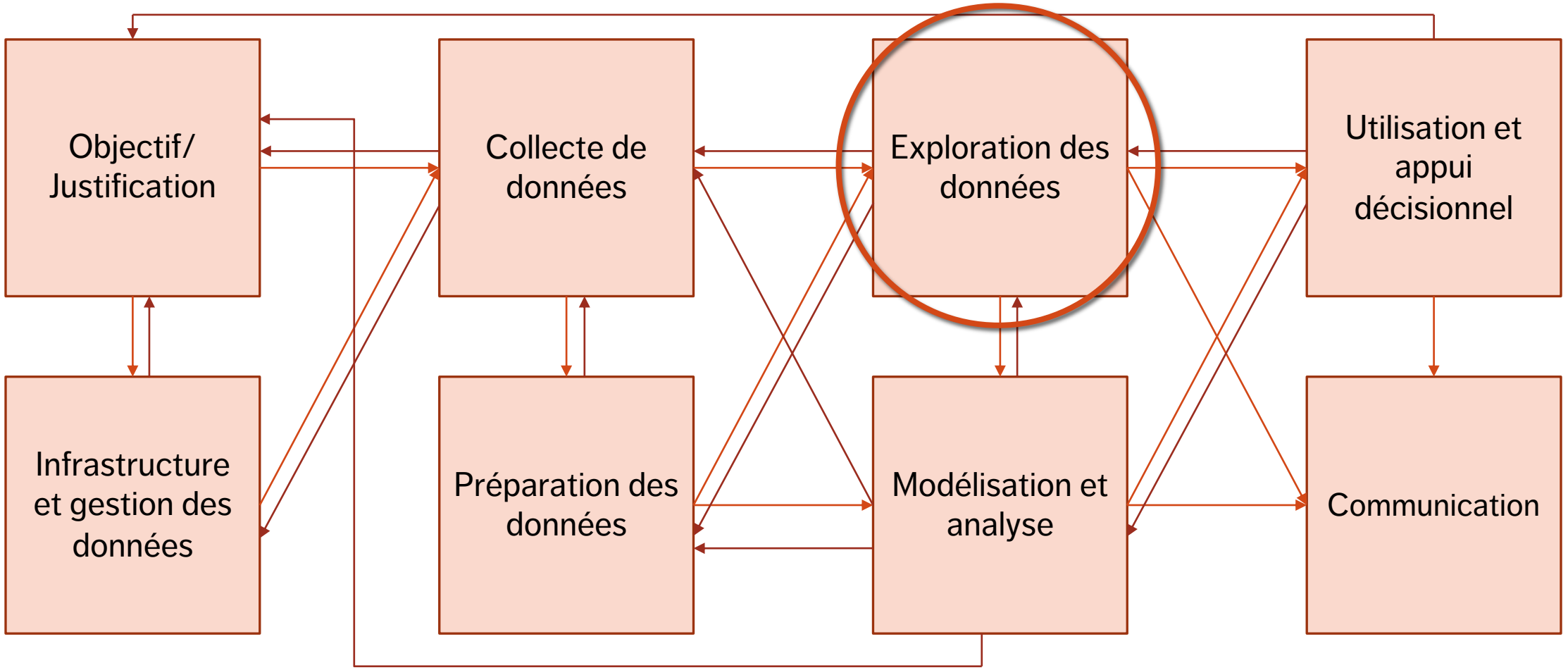
Les concepts de la visualisation des données

LA VISUALISATION DES DONNÉES ET LES TABLEAUX DE BORD



1. L’analyse exploratoire des données

Le processus d'analyse (désordonné)



Quelques questions de base

Quel **système** vos données représentent-elles - **objets**, **attributs**, **relations** ?

Comment représente-t-il ce système - c-à-d quel est le **modèle de données** ?

Qui a créé cet ensemble de données ? Quand ? Dans quel but ?

Supposons un fichier “plat” - que représentent les lignes et les colonnes ?

Disposez-vous de suffisamment d'informations (e.g., des **métadonnées**) pour répondre à ces questions ? Où pouvez-vous trouver plus d'informations ?

Les résumés non visuels

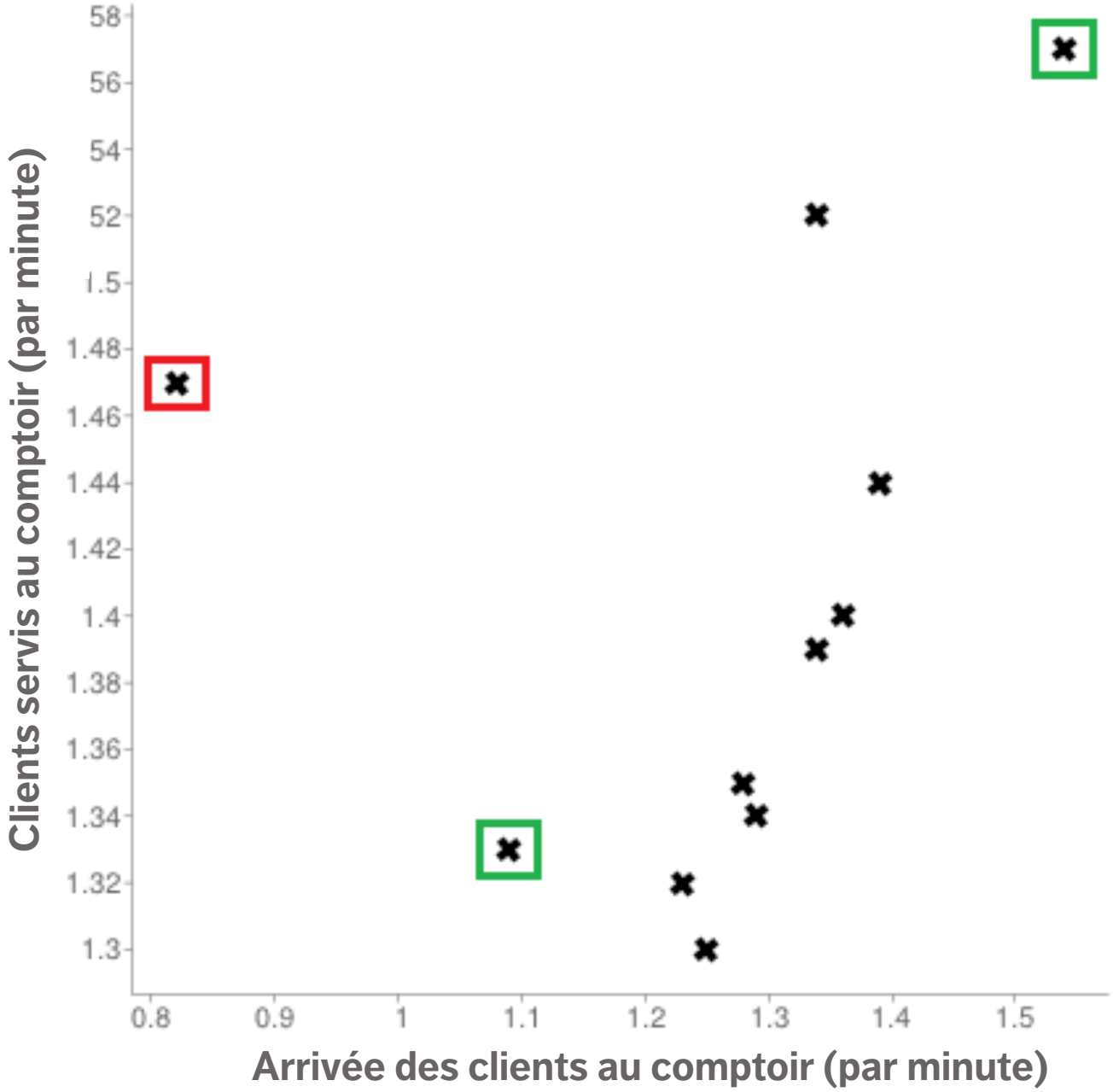
```
       Cl              NO3              NH4
 Min.   :  0.222   Min.   : 0.000   Min.   :    5.00
 1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86
 Median : 32.470   Median : 2.356   Median :  107.36
 Mean   : 42.517   Mean   : 3.121   Mean   :  471.73
 3rd Qu.: 57.750   3rd Qu.: 4.147   3rd Qu.:  244.90
 Max.   :391.500   Max.   :45.650   Max.   :24064.00
 NA's   :16        NA's   :2        NA's   :2

    season
 Length:340
 Class :character      autumn spring summer winter
 Mode  :character          80     84     86     90
```

L’utilisation avant l'analyse

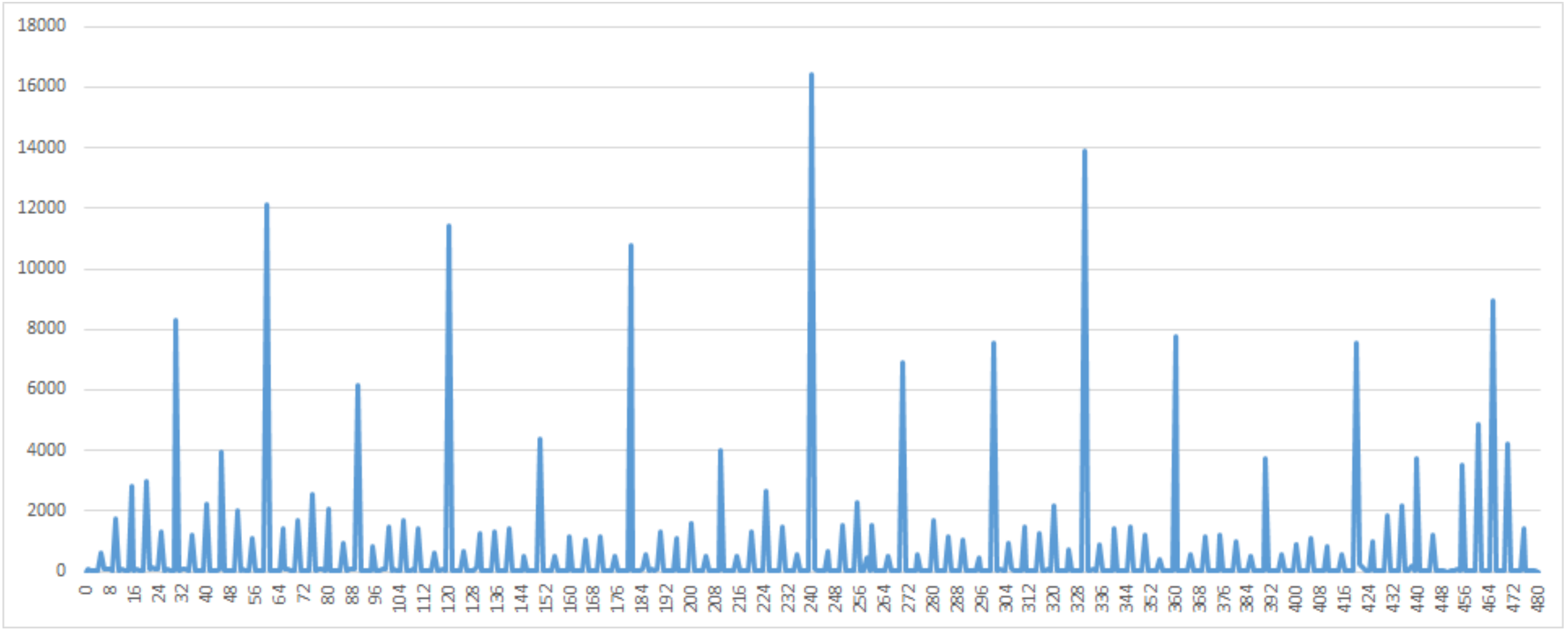
La visualisation des données peut être utilisée pour préparer l'étape de l'analyse :

- **détecter les entrées anormales**
  entrées non valides, valeurs manquantes, valeurs aberrantes
- **façonner les transformations des données**
  “binning”, normalisation, transformations Box-Cox, transformations de type PCA
- **se faire une idée des données**
  analyse des données en tant que forme d'art, analyse exploratoire
- **identification de la structure cachée des données**
  regroupement, associations, modèles informant les étapes suivantes

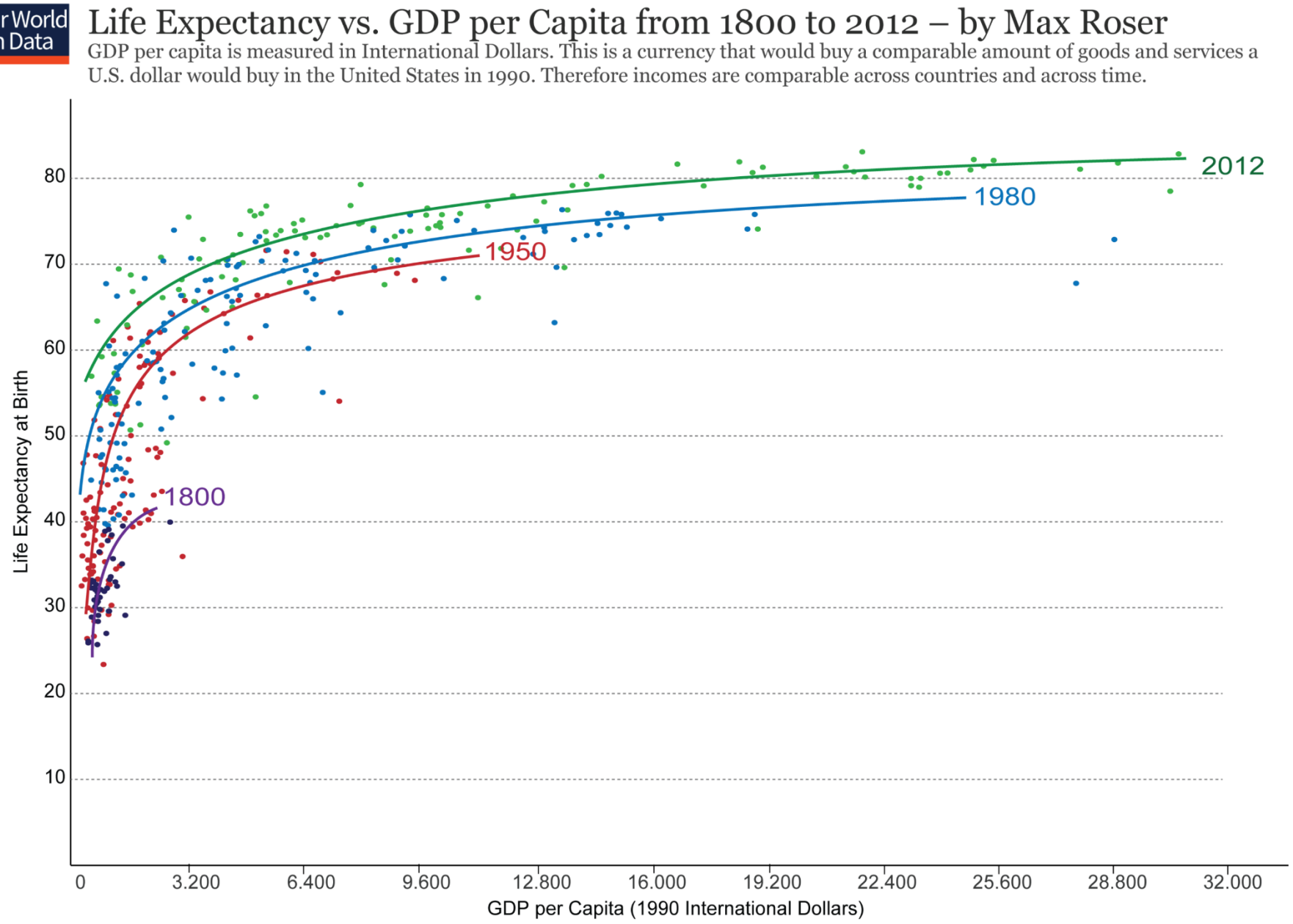
[Ensemble de données personnelles]



[Ensemble de données personnelles]



[https://ourworldindata.org/life-expectancy]



Ce graphique montre la corrélation entre l'espérance de vie et le PIB par habitant.

Les pays dont le PIB est plus élevé ont une espérance de vie plus élevée, en général.

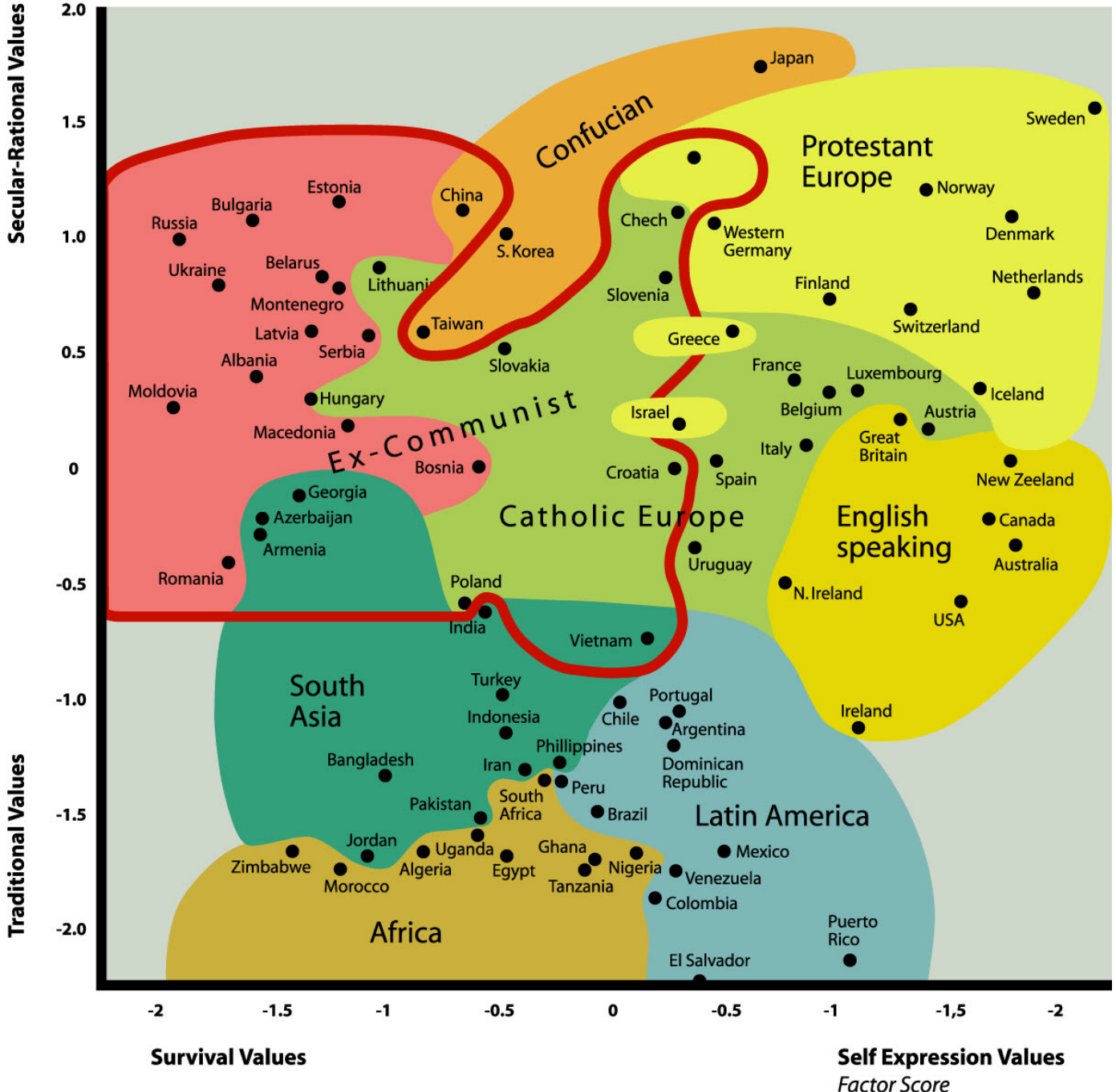
La relation semble suivre une tendance logarithmique : l'augmentation de l'espérance de vie par unité d'augmentation du PIB diminue à mesure que le PIB par habitant augmente.

Data sources: Data on life expectancy are from Gapminder.org; data on GDP per capita are from the 'New Maddison Project Database'.
The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.

Session 1

Carte Culturelle d'Inglehart-Welzel (2010-2014)

[https://en.wikipedia.org/wiki/World_Values_Survey]



Les valeurs traditionnelles
importance de la religion, des liens parents-enfants, de la déférence envers l'autorité et des valeurs familiales traditionnelles.

Valeurs **laïques et rationnelles**
Moins d'importance accordée à la religion, aux valeurs familiales traditionnelles et à l'autorité.

Valeurs de survie
accent mis sur la sécurité économique et physique.

Valeurs d'expression personnelle
priorité élevée à la protection de l'environnement, tolérance croissante à l'égard des étrangers, des gays et lesbiennes et de l'égalité des sexes

Quelques graphiques d'exploration

Texte et tableaux

Cartes à tapis/Lignes de nombres

Histogrammes/Chartes à barres

Boxplots/Graphiques à moustache

Graphiques en ligne

Diagrammes de dispersion

Les lignes de nombres (“rugs”)

Les lacunes dans la ligne numérique : l’**absence** de ces valeurs numériques dans les données.

Rappelez-vous : ceci est (possiblement) différent de l'ordre dans lequel les valeurs apparaissent dans l'ensemble de données – puisqu'il s'agit d'une ligne de nombres, elle montre où les valeurs tombent numériquement.

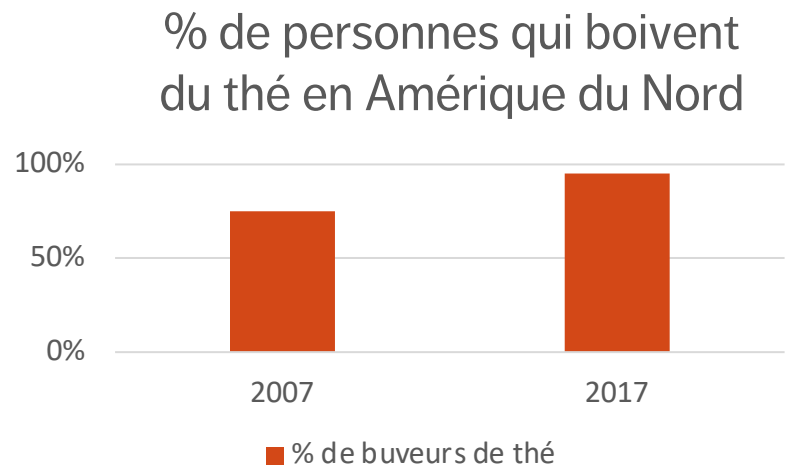
Si certaines valeurs sont **identiques**, elles sont superposées.

Du texte simple

Un ou deux chiffres sur lesquels se concentrer.

Bon pour "planter le décor".

Attire l'attention sur une zone du rapport.



95% de la population boit du thé aujourd'hui, par rapport à 75% en 2007

Les tables et tableaux

Les tableaux interagissent avec notre système **verbal**, ce qui signifie que nous les **lisons** :

- utilisés pour comparer les valeurs
- le public cherche les rangées

Le design de la table doit **se fondre** dans le décor

- les données doivent ressortir, pas les bordures
- tableaux/données denses : utilisez des couleurs de ligne qui alternent

Nom	L'année dernière	Cette année
Bob	20	30
Fred	30	40
George	10	15

Nom	**L'année dernière**	**Cette année**
Bob	20	30
Fred	30	40
George	10	15

Les tables “heatmap”

Tirez parti de la couleur pour transmettre l'ampleur des quantités

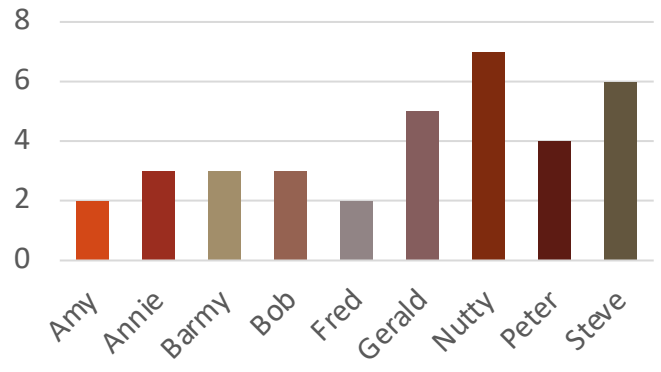
- utilisez la **saturation d'une seule couleur** plutôt que la différenciation (différentes couleurs)
- avec une légende (blanc = faible, bleu = élevé), les chiffres peuvent être supprimés sans altérer le message

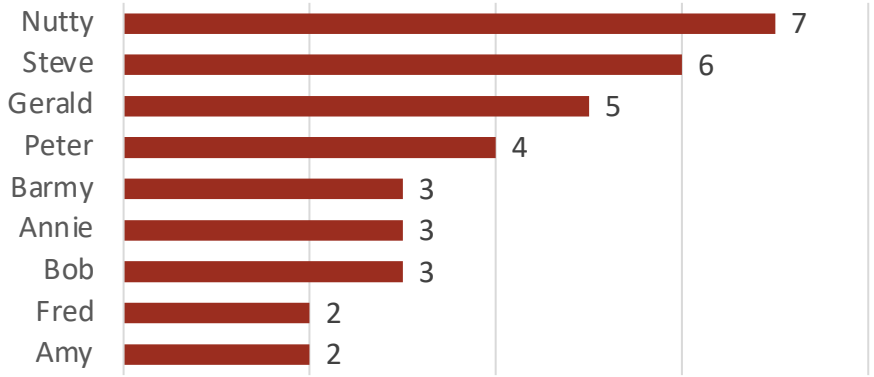
	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40

	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40

	Last Year	This Year	Next Year	Optimum
George				
Peter				
John				
Sandra				

Les diagrammes à barres





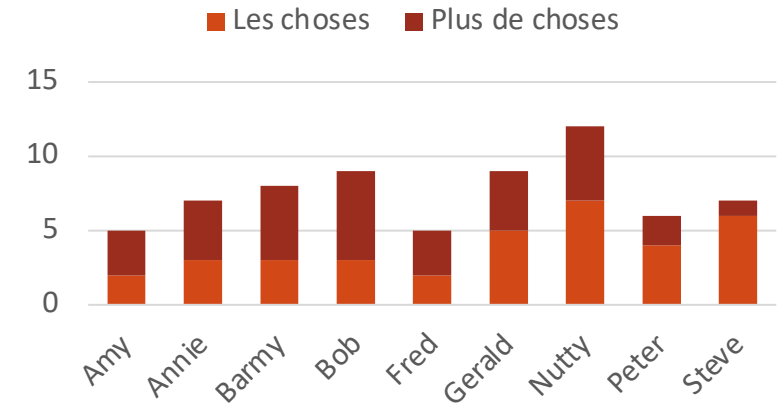
Très polyvalent et utile : utilisez TOUJOURS (?) une ligne de référence au zéro.

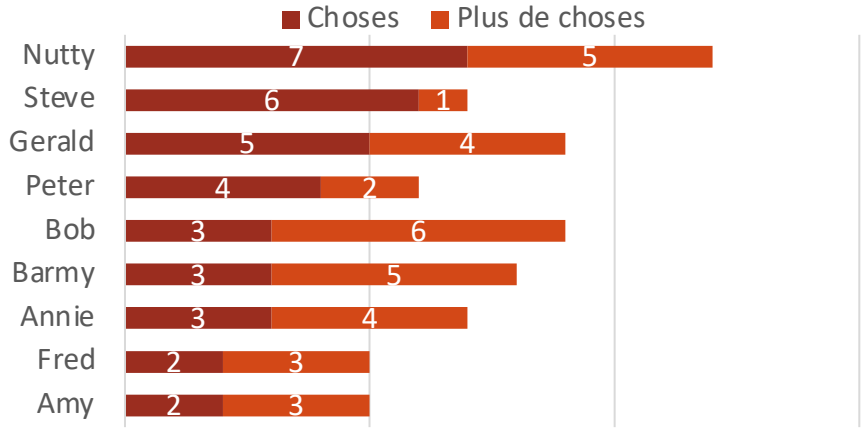
Utilisez soit l'axe du graphique OU des étiquettes pour les données (l'axe pour les déclarations générales, les étiquettes de données pour plus de details).

Les graphiques horizontaux sont apparemment **plus faciles à lire** (selon de nombreuses études).

Pensez à l'ordre des catégories.

Les diagrammes à barres empilées



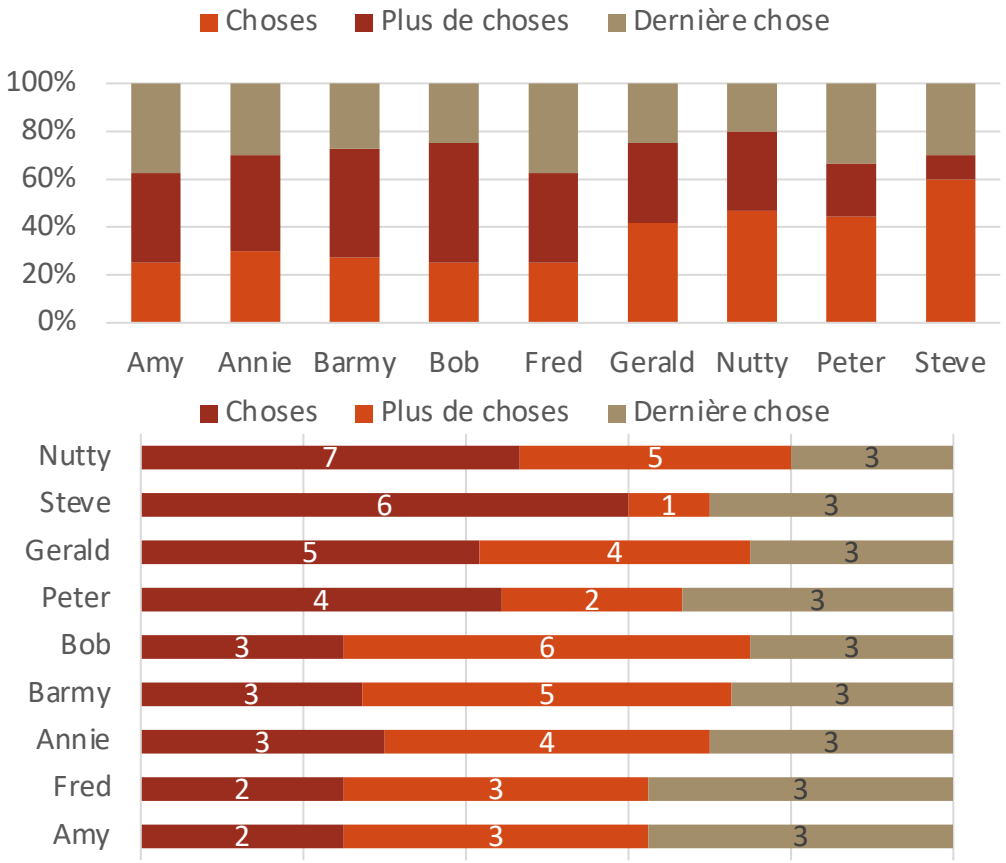


Conçus pour **comparer les totaux**, mais peuvent rapidement devenir “**écrasant**”.

Difficile de trier / ordonner.

Le filtrage est assez compliqué (sur quoi faut-il cliquer et comment le graphique réagit-il lorsque l'on clique sur le filtre ?)

Les diagrammes à barres à 100%



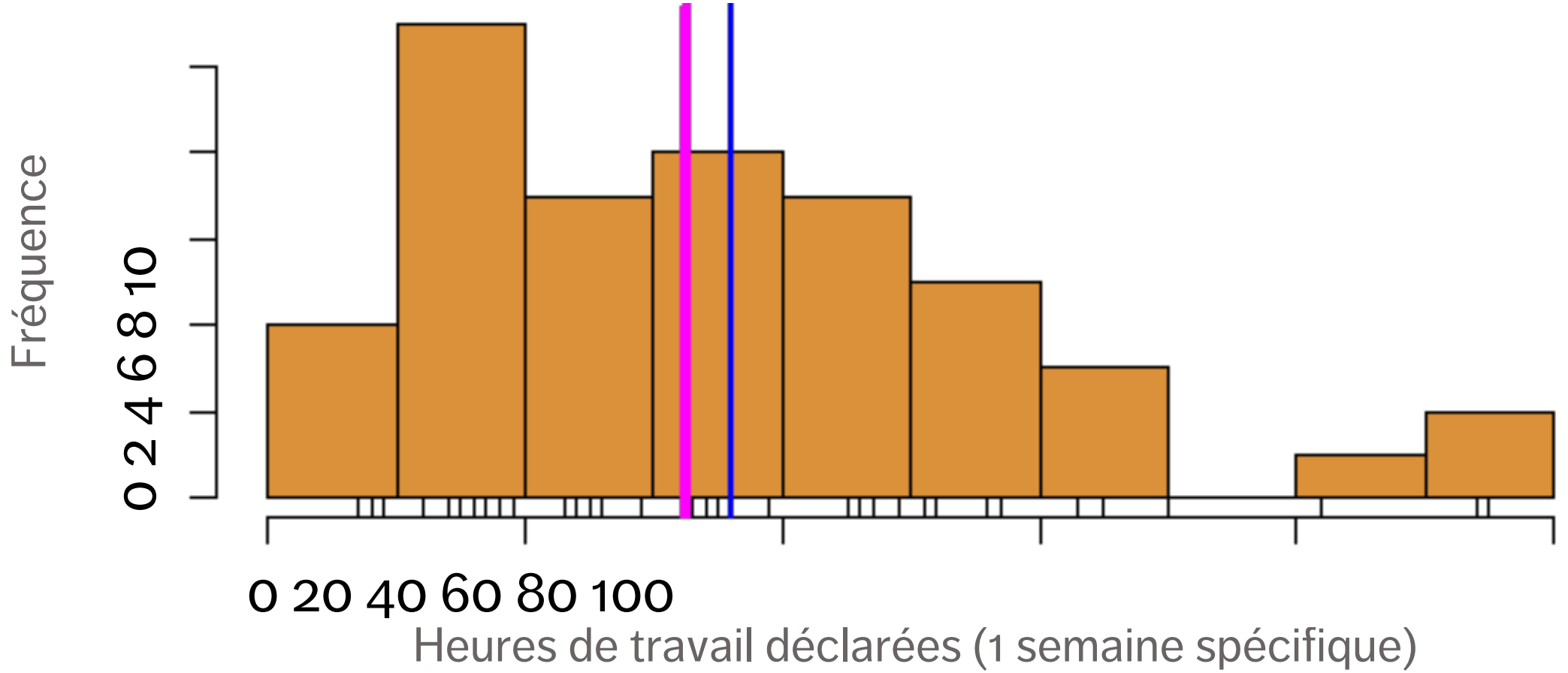
Fonctionnent bien pour visualiser des **portions** d'un tout sur une échelle allant du négatif au positif.

Ligne de base cohérente aux extrêmes.

Aucune mesure relative à l'**ampleur** des données.

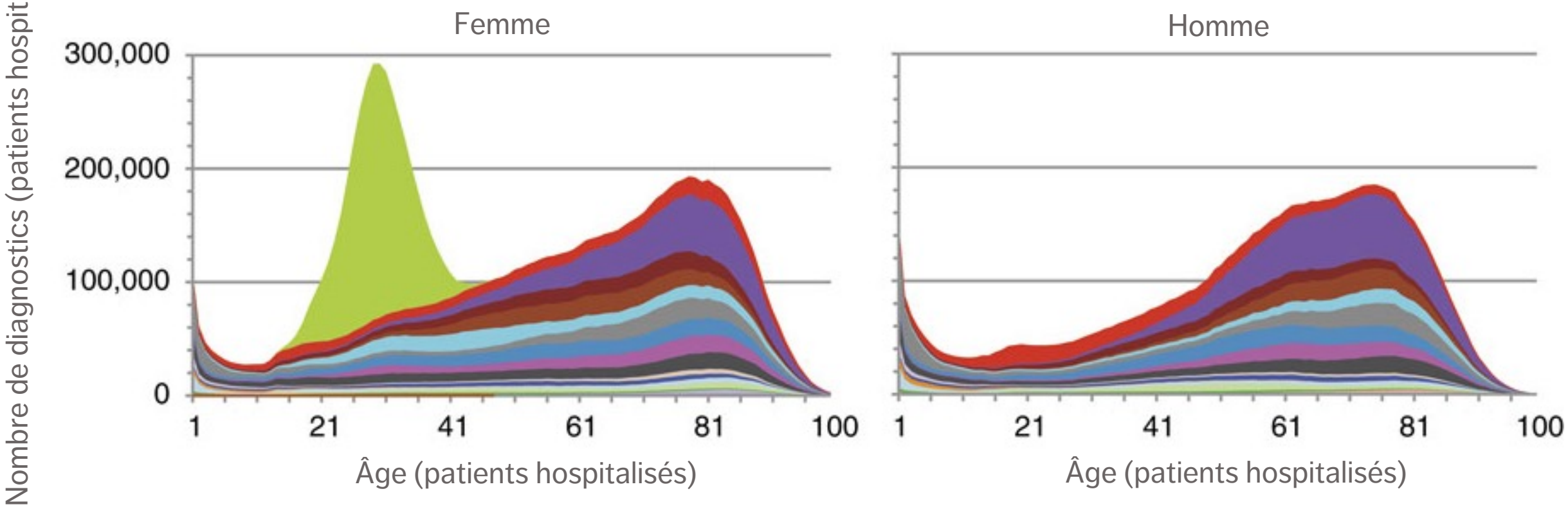
La recherche montre que l'horizontal est plus facile à lire que le vertical.

Les histogrammes



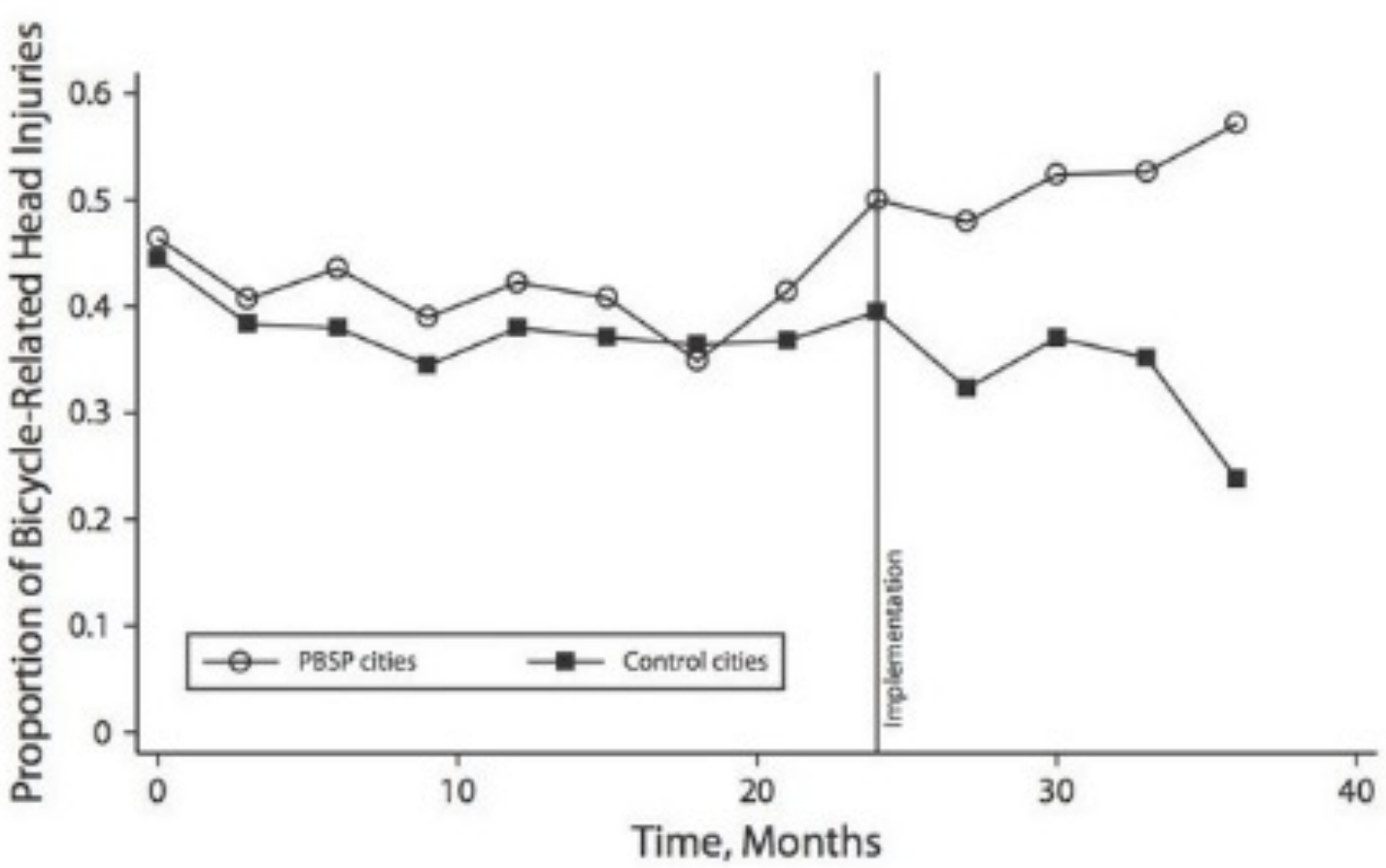
[A.B. Jensen et al., *Danish Medical Study*, Nature Communications].

Les histogrammes empilés



[Graves et al., *Am.J.Phys.Health*, 2014]

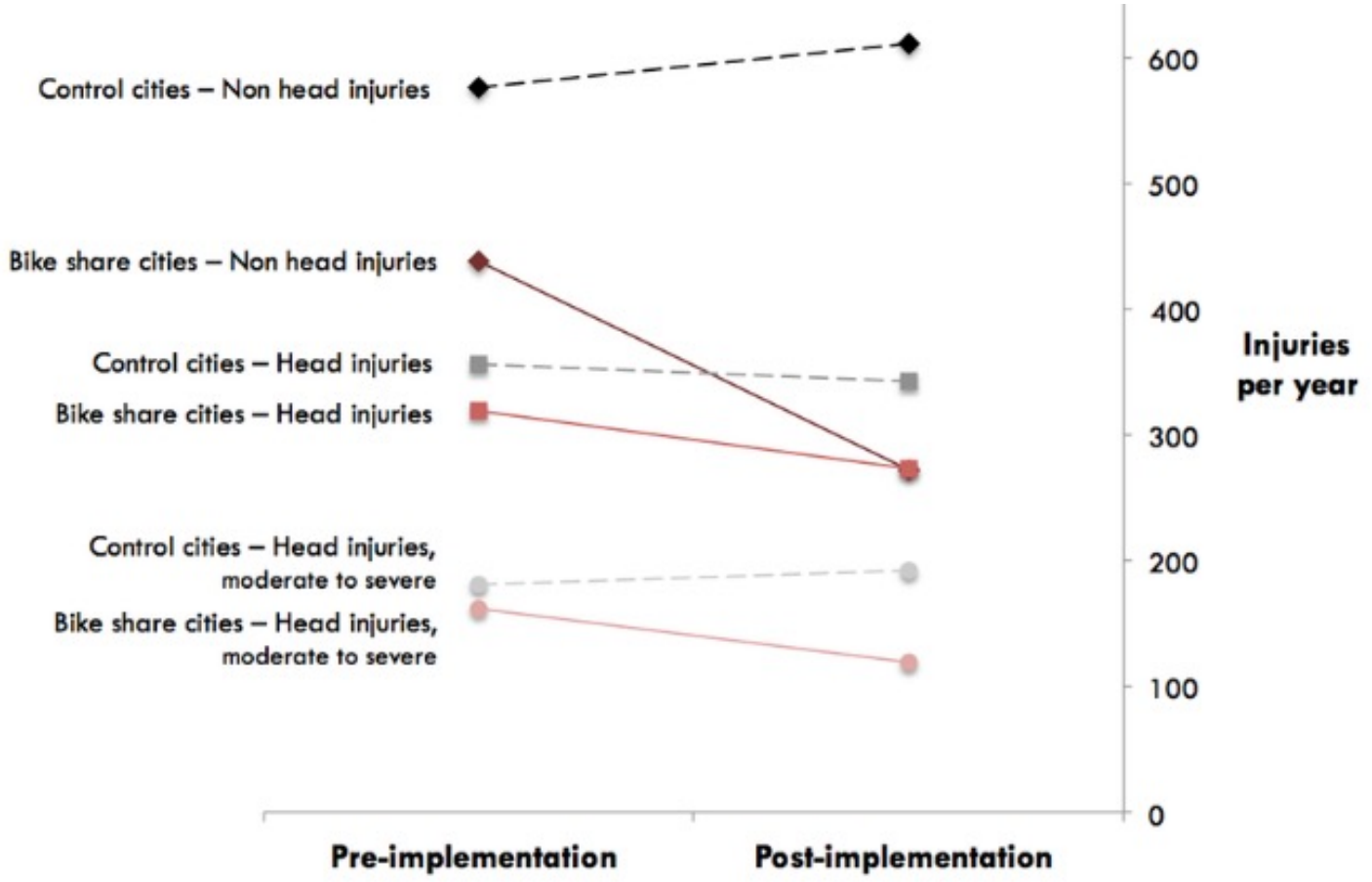
Les graphiques en ligne



Proportion de toutes les blessures liées à la bicyclette qui ont été classées comme des blessures à la tête parmi les villes ayant un programme de vélos en libre-service et les villes témoins, centrées sur la date d'intervention.

[Réfutation de Teschke]

Les graphiques en ligne

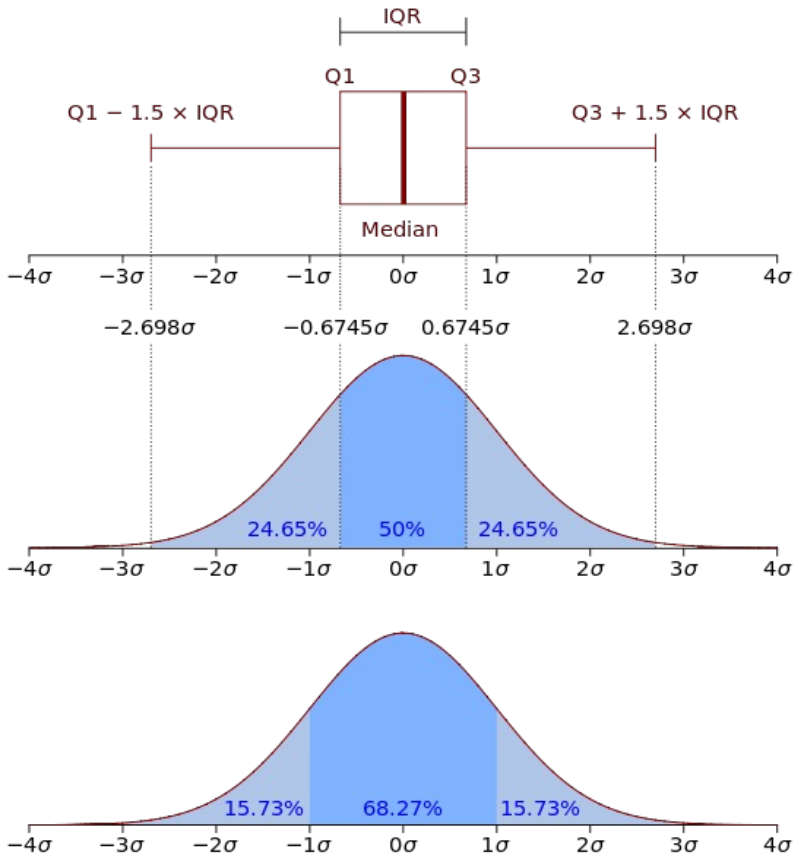


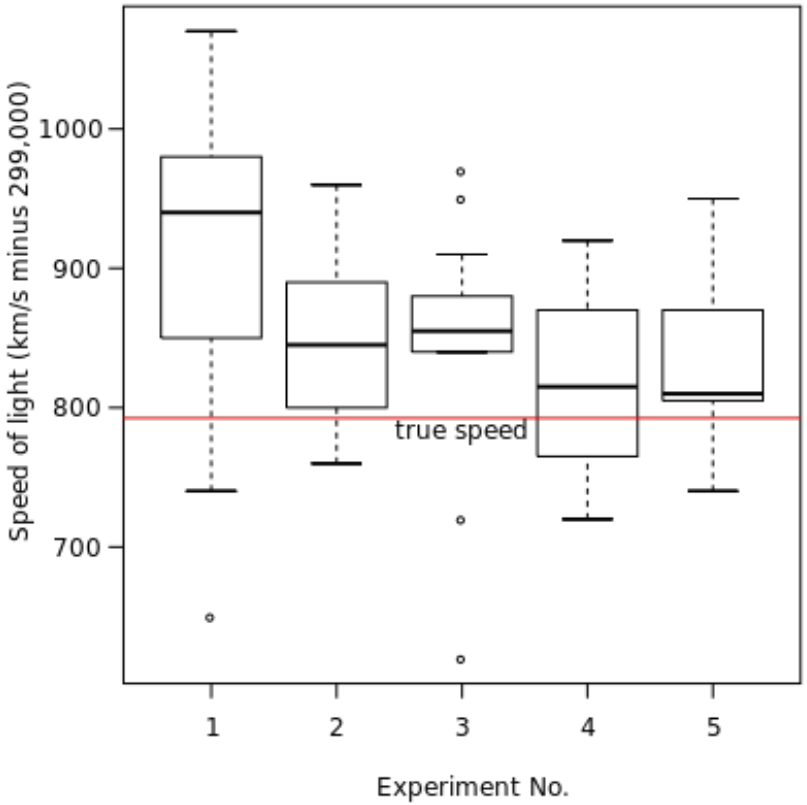
Les données d'une nouvelle étude montrent une baisse de toutes les blessures, y compris les traumatismes crâniens, après la mise en place du système de vélos en libre-service.

Comme les traumatismes crâniens diminuent moins que les autres blessures, ils représentent aujourd'hui une proportion plus importante de toutes les blessures.

[Fichiers de Wikipedia]

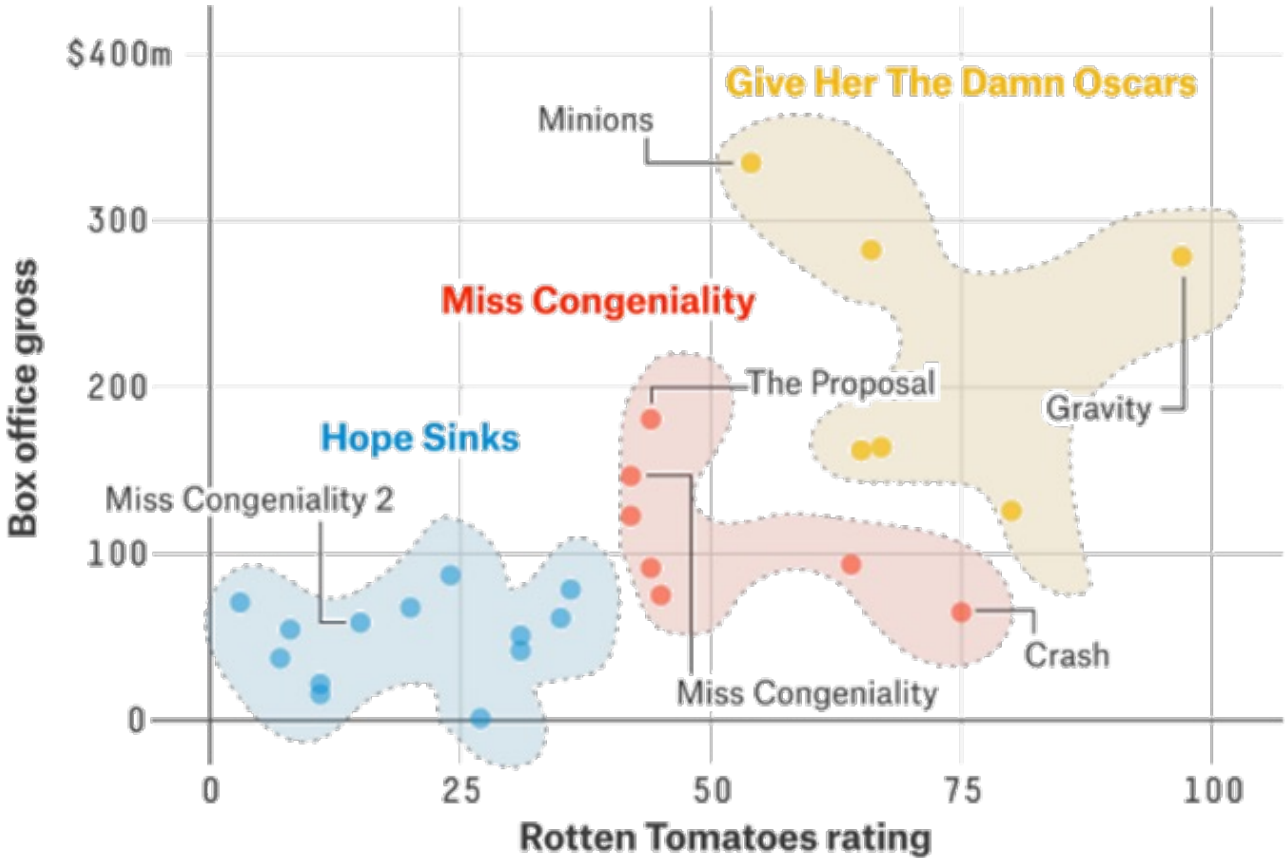
Les bôites à moustaches (“boxplots”)

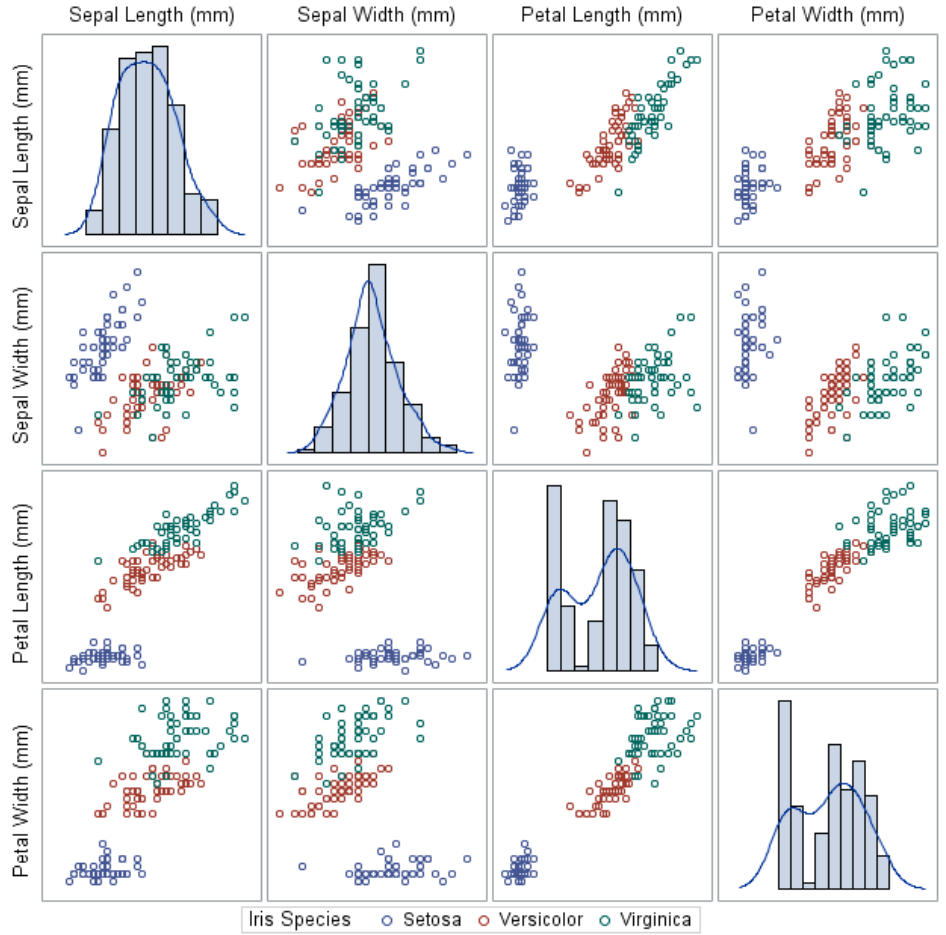




Expérience de Michelson-Morley afin de déterminer la vitesse de la lumière

Les diagrammes de dispersion





Lectures suggérées

L'analyse exploratoire des données

Data Understanding, Data Analysis, Data Science
Data Visualization and Data Exploration

Data and Charts

- Pre-Analysis Uses

The Practice of Data Visualization
Basics of Data Visualization

Data Exploration

Workhorse Data Visualizations

Exercices

L’analyse exploratoire des données

1. Trouvez des exemples de présentations de données que vous considérez comme particulièrement perspicaces et/ou puissantes. Discutez de leurs forces/faiblesses.

2. Trouvez des exemples de présentations de données que vous considérez comme particulièrement trompeuses et/ou inutiles. Discutez de leurs forces/faiblesses.

3. Comment pensez-vous que les nouvelles technologies (par exemple, la réalité virtuelle ou augmentée, l'impression 3D, l'informatique portable) influenceront les présentations de données ?