Data visualisation

Master 1 IDSM

Iuliia TKACHENKO, Laboratoire LIRIS

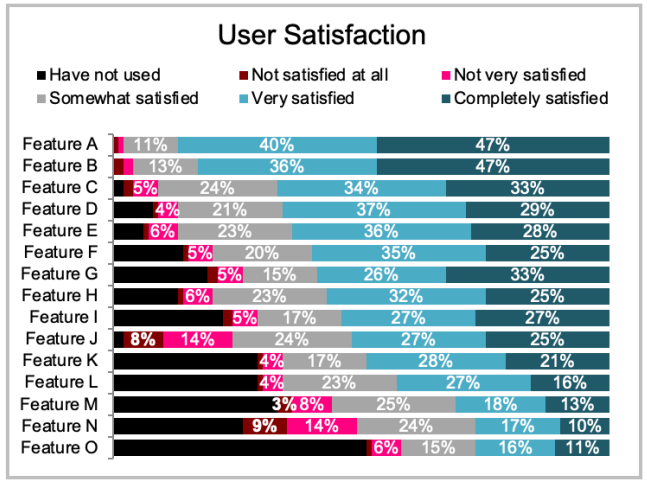
iuliia.tkachenko@univ-lyon2.fr

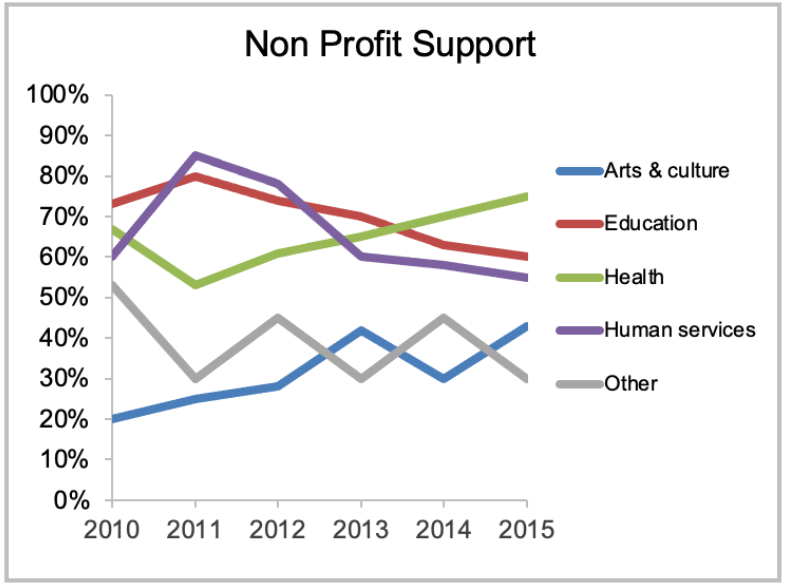
LIRIS

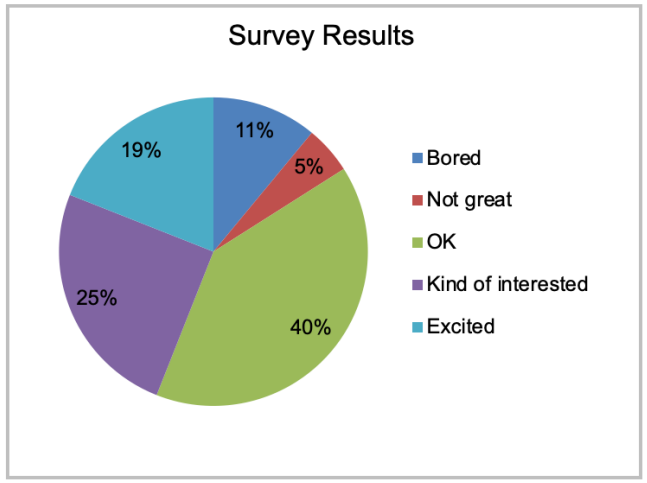
- **L**aboratoire d'**I**nfo**R**matique en **I**mage et **S**ystèmes d'information
- 14 équipes de recherche sont structurées en 6 pôles de compétences
- L'équipe **IMAGINE** réunit 15 enseignants-chercheurs.
- Le métier de base est l'analyse et le traitement des médias visuels.

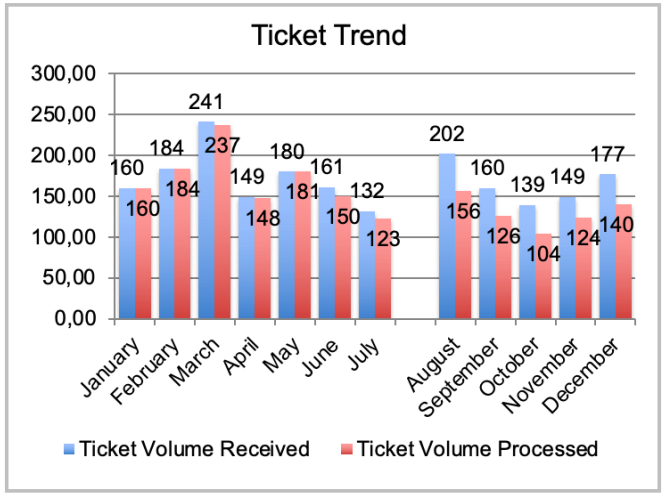
Motivation

- Les mauvais graphiques sont partout







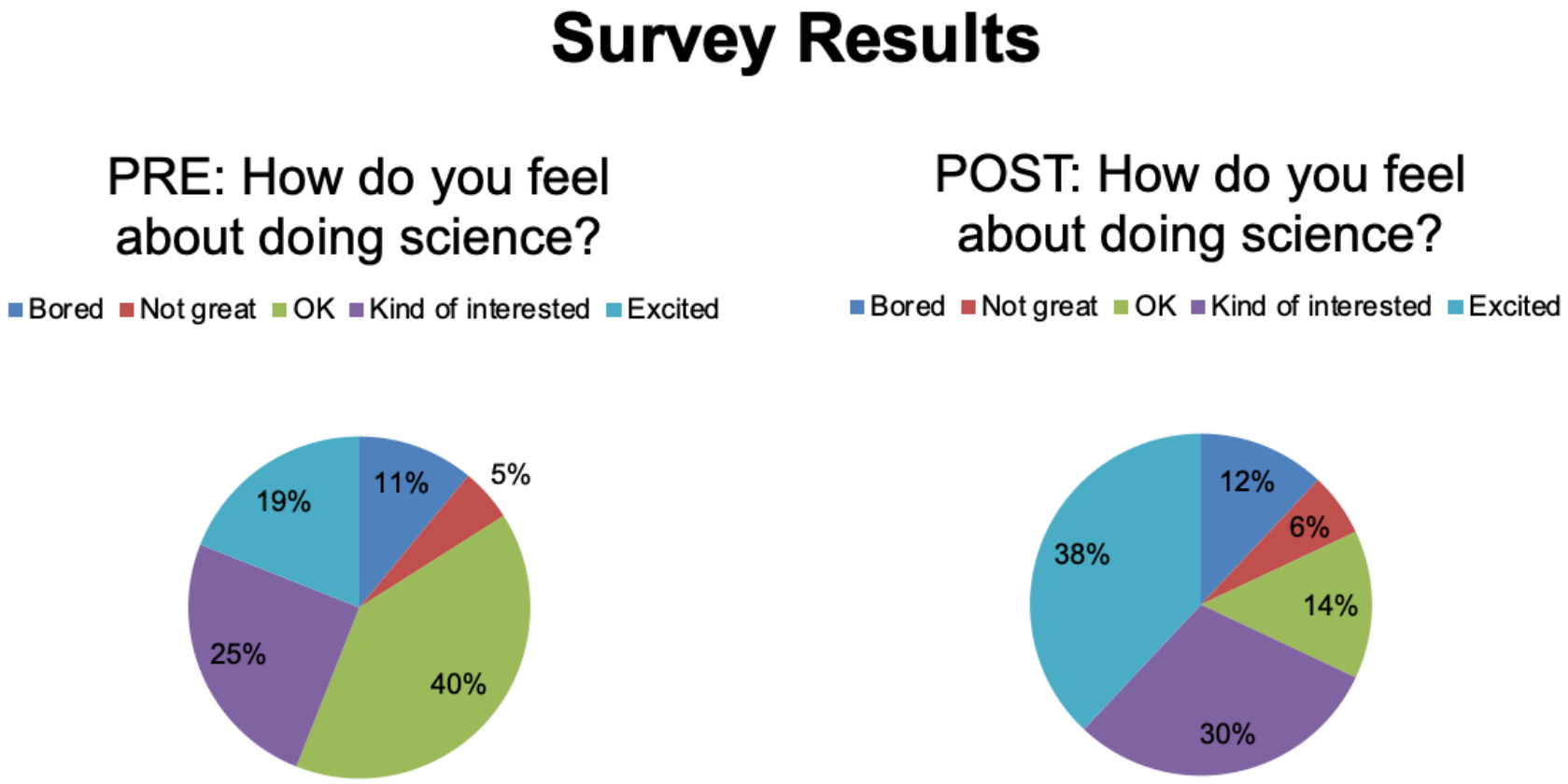


Ici et dans la suite de cours : les visualisations du livre « Storytelling with data: a data visualization guide for business professionals », Cole Nussbaumer Knaflic

Création de « bonne » visualisation

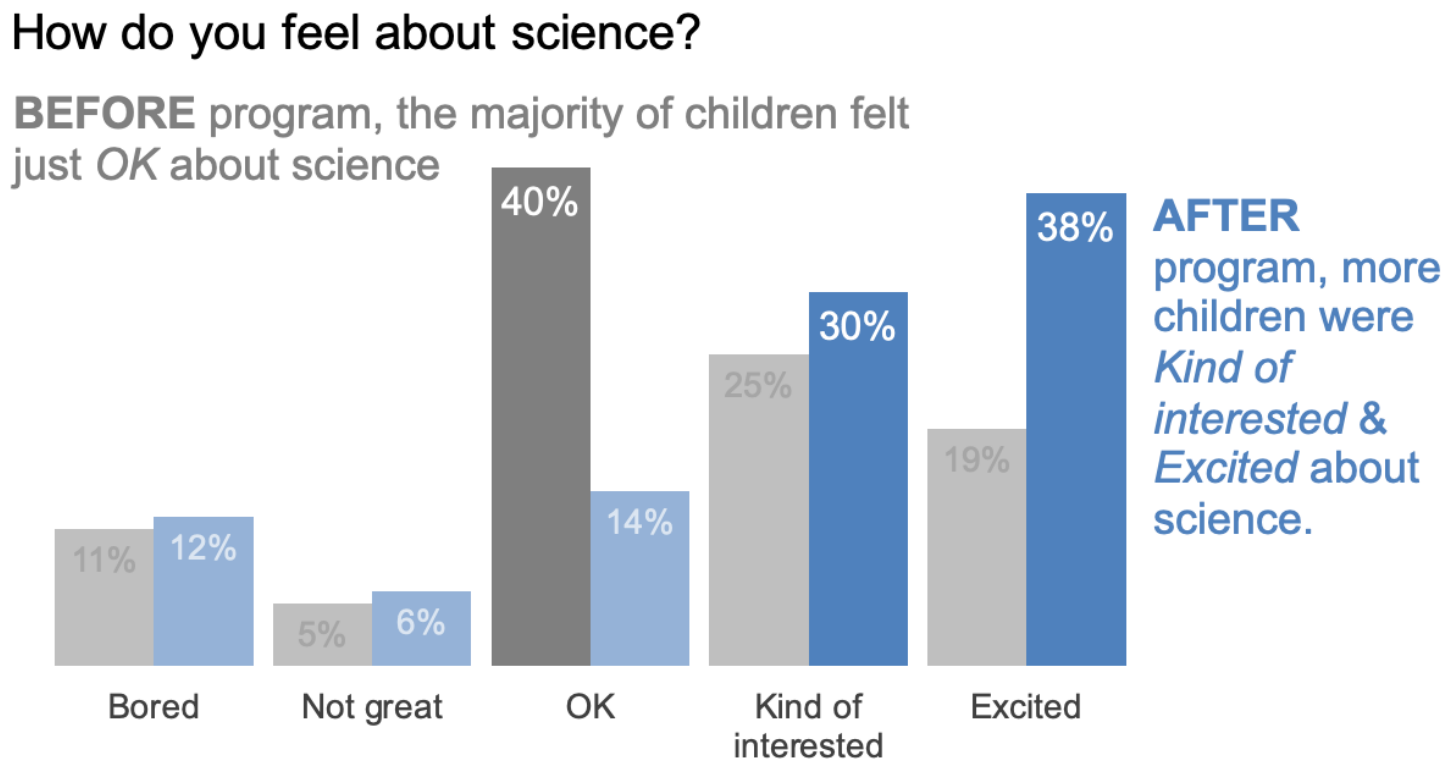
- Composant artistique
  - Conception visuel
  - Création de quelque chose beau et irrésistible
- Composant scientifique et mathématique
  - Être en mesure de fournir un bon aperçu

Exemple de « mauvaise » visualisation



Exemple de « bonne » visualisation

Pilot program was a success



Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

Organisation du cours

- Volume horaire :
  - Séance = Cours Magistral + Travaux pratique
  - 6 séances de 3h
  - Du 18 mars au 22 mars de 17h à 20h (heur France)
  - Le 23 mars de 9h à 12h

- Evaluations :
  - Contrôle continue **mercredi 20 mars**
  - Projet en binôme : la date du rendu **7 avril 2024 à 23h55**

Plan du cours

- L’intérêt de la visualisation de données
- L’importance du contexte
  - Qui, Quoi, Comment
  - La création de « storyboard»
- Le choix un visuel efficace
  - Les types de graphiques
  - Un tableau de bord

Statistique vs Visualisation

- Les mesures statistiques peuvent être utiles pour tenter comprendre rapidement des données.

I	
X	Y
10	8,04
8	6,95
13	7,58
9	8,81
11	8,33
14	9,96
6	7,24
4	4,26
12	10,84
7	4,82
5	5,68

II	
X	Y
10	9,14
8	8,14
13	8,74
9	8,77
11	9,26
14	8,1
6	6,13
4	3,1
12	9,13
7	7,26
5	4,74

III	
X	Y
10	7,46
8	6,77
13	12,74
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73

IV	
X	Y
8	6,58
8	5,76
8	7,71
8	8,84
8	8,47
8	7,04
8	5,25
19	12,5
8	5,56
8	7,91
8	6,89

Statistique vs Visualisation

- Les mesures statistiques peuvent être utiles pour tenter comprendre rapidement des données.

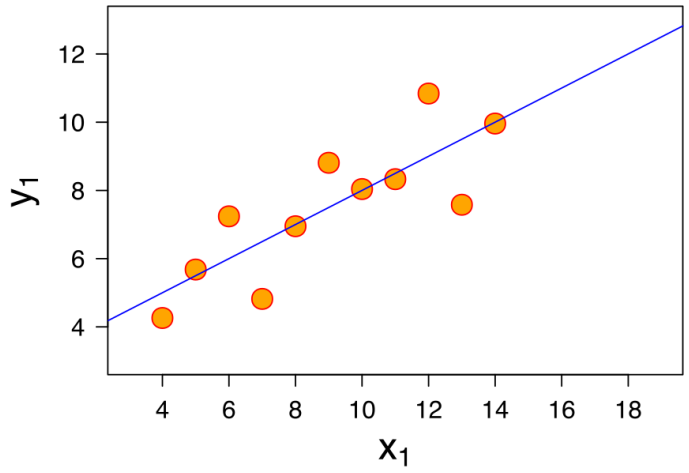
Quartet d'Anscombe

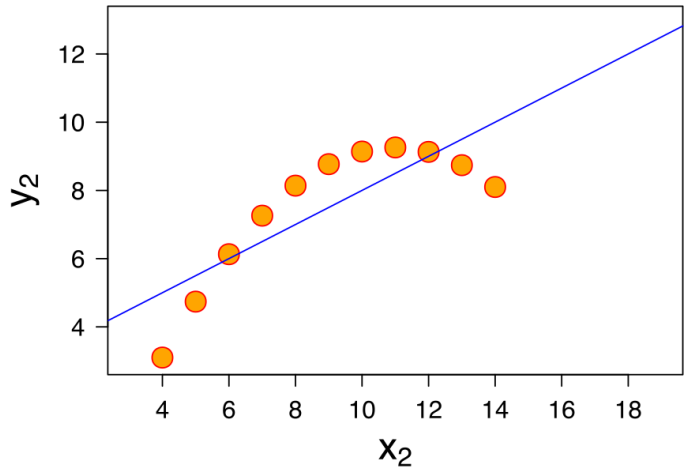
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

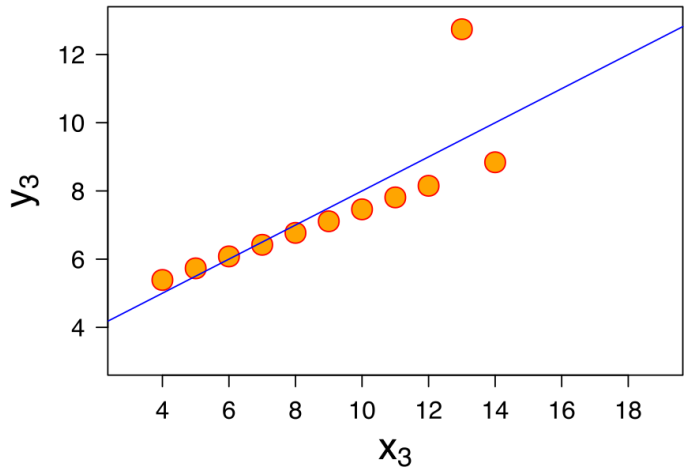
Moyenne	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
Écart-type	3,16	1,94	3,16	1,94	3,16	1,94	3,16	1,94

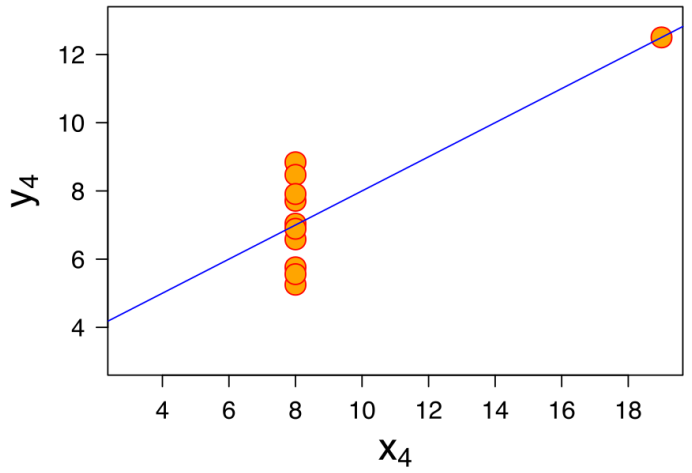
Est-ce que les données sont similaires ?

Quartet d'Anscombe

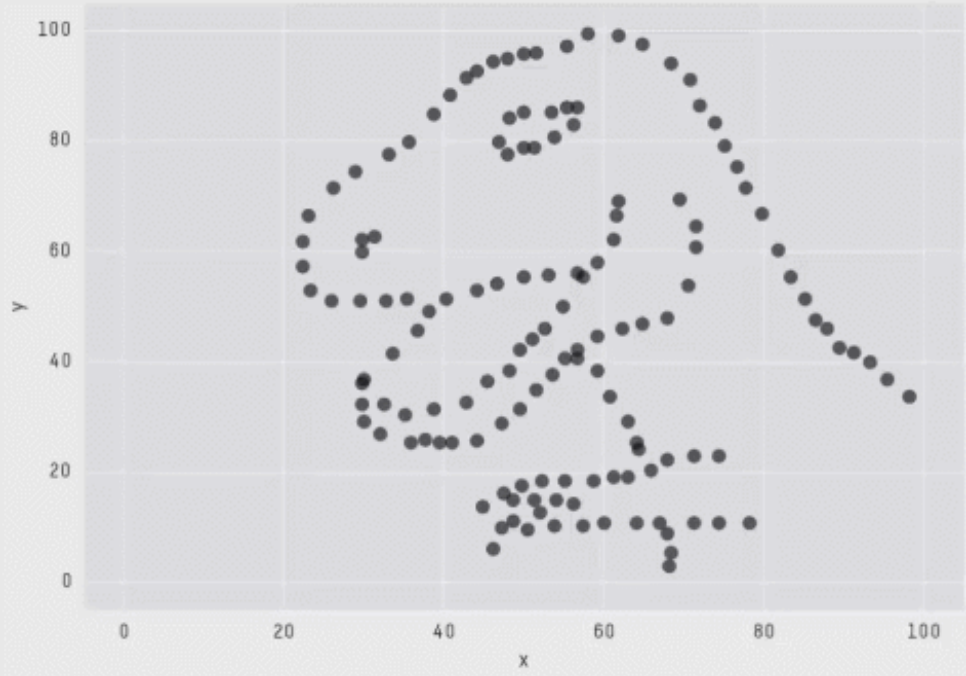








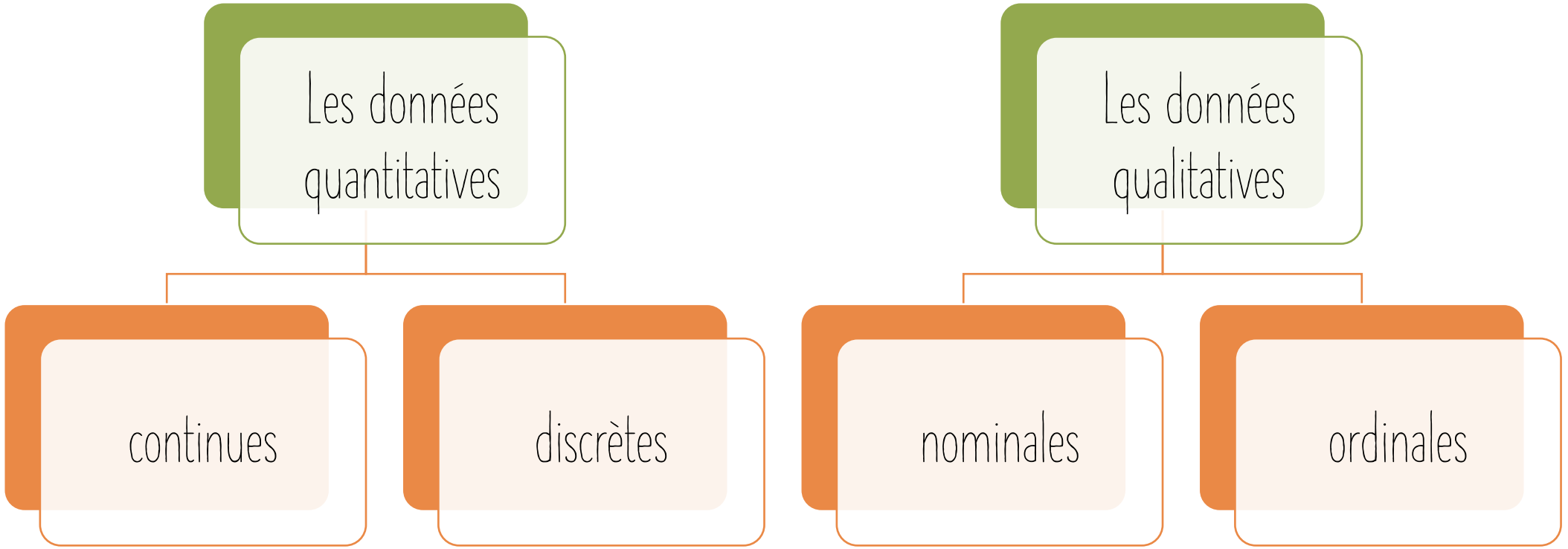
Datasaurus dataset



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Source: https://jumpingrivers.github.io/datasauRus/

Les types des données



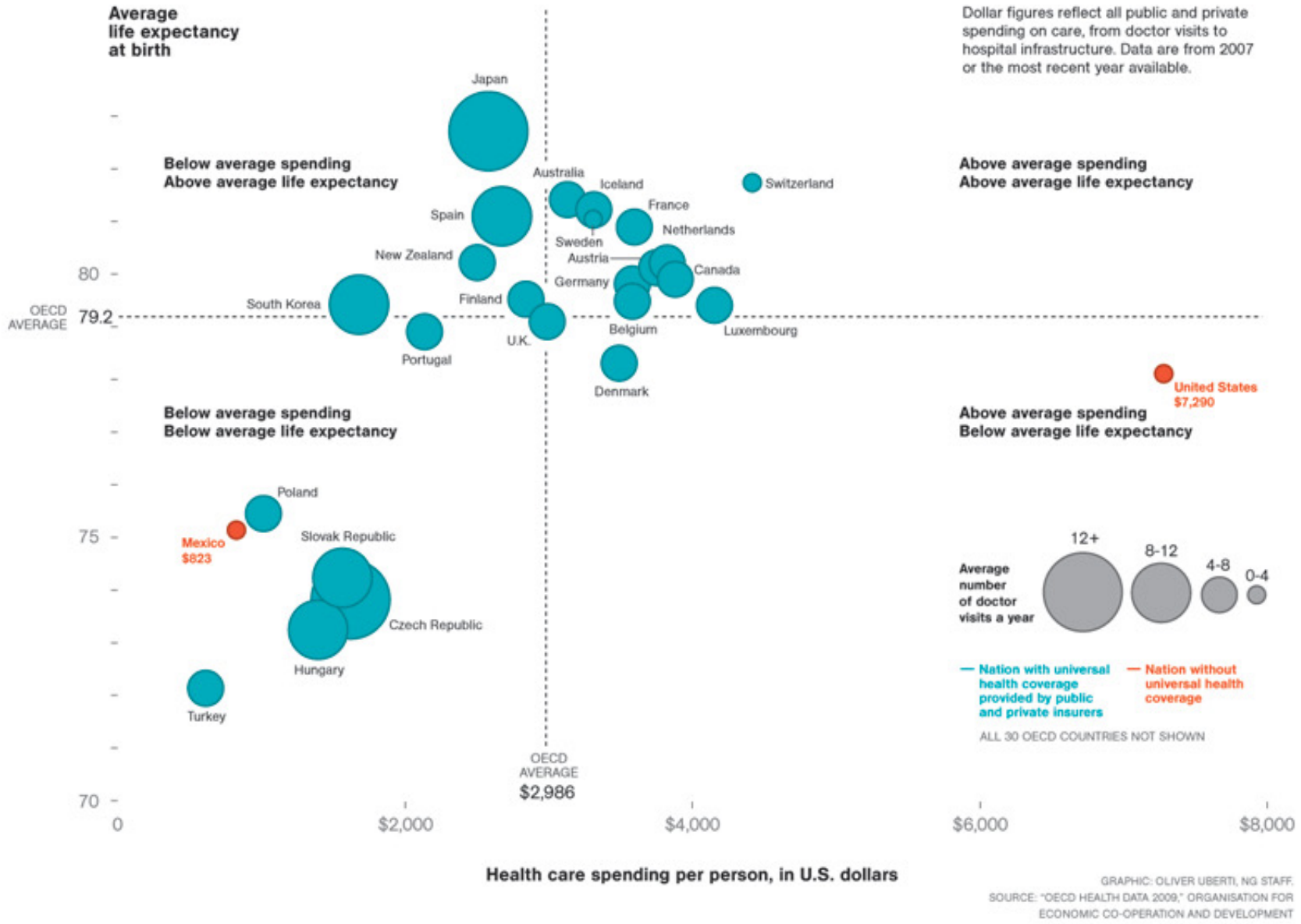
Exemples par type de données

Données nominales	Données ordinales	Données discrètes	Données continues
Couleur des cheveux (Blond, Roux, Brun, Noir, etc.) État matrimonial (Célibataire, Veuf, Marié) Nationalité (indienne, allemande, américaine) Sexe (Homme, Femme, Autres) Couleur des yeux (noir, marron, etc.)	Lorsque les entreprises demandent des retours, de l'expérience ou de la satisfaction sur une échelle de 1 à 10 Notes alphabétiques à l'examen (A, B, C, D, etc.) Classement des personnes dans une compétition (Premier, Deuxième, Troisième, etc.) Statut économique (élevé, moyen et faible) Niveau d'éducation (Supérieur, Secondaire, Primaire)	Nombre total d'élèves présents dans une classe Coût d'un téléphone portable Nombre d'employés dans une entreprise Le nombre total de joueurs ayant participé à une compétition Jours dans une semaine	Taille d'une personne Vitesse d'un véhicule "Le temps pris" pour terminer le travail Fréquence Wi-Fi Prix de la part de marché

Exercice 1

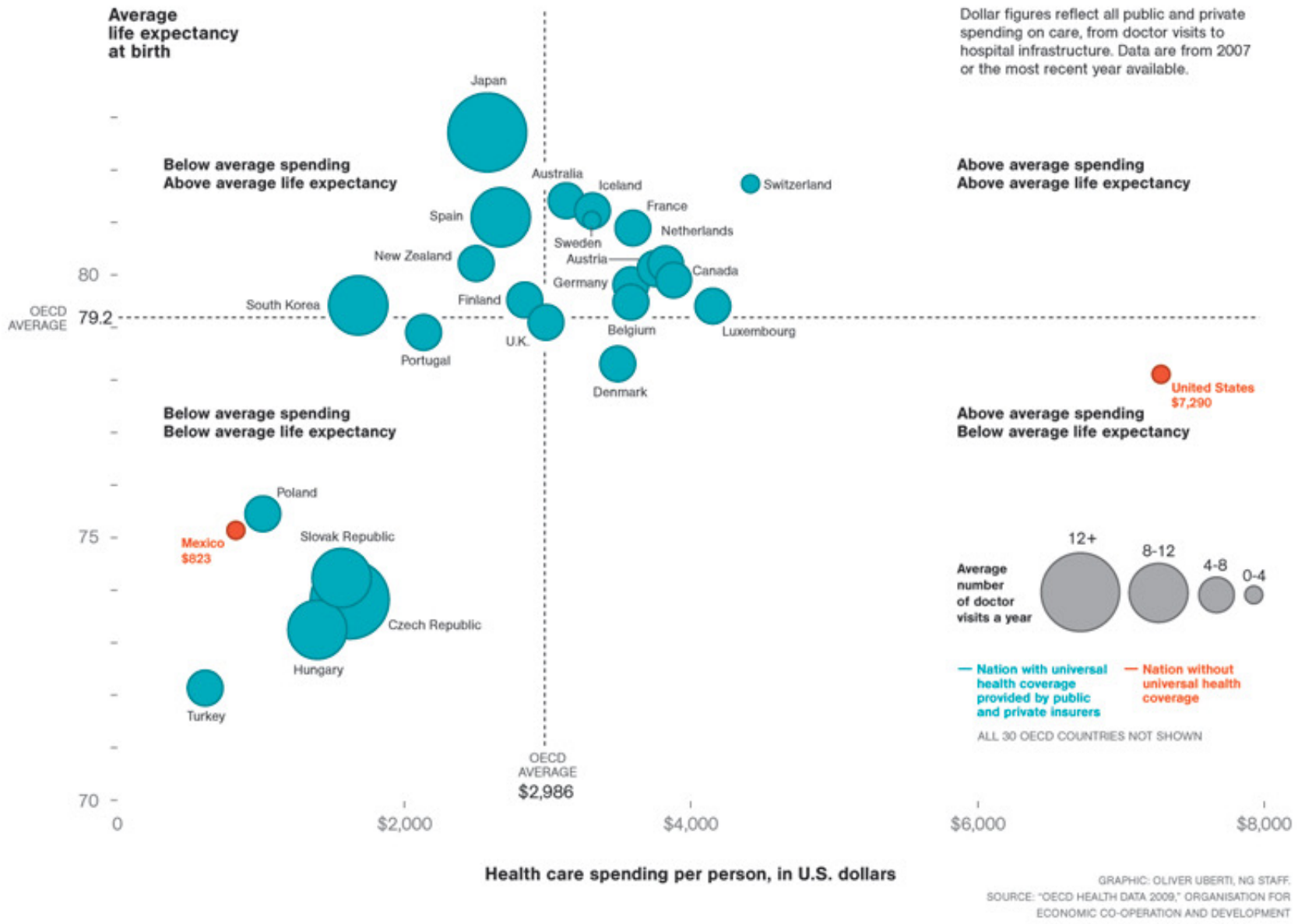
continues | discrètes | nominales | ordinales

Variable	Type de données
Visites chez le médecin par an	
Espérance de vie	
Dépenses par personne	
A des soins de santé universels	



Exercice 1 correction

Variable	Type de données
Visites chez le médecin par an	Ordinale
Espérance de vie	Continue
Dépenses par personne	Continue
A des soins de santé universels	Nominale



6 leçons de storytelling

1. L’importance du contexte.
2. Choisir un visuel efficace.
3. Eliminer la surcharge.
4. Attirer l’attention du public
5. Penser comme un designer.
6. Raconter une histoire.

« Storytelling with data: a data visualization guide for business professionals », Cole Nussbaumer Knaflic

Importance du contexte

Analyse exploratoire et explicative

L'**analyse exploratoire** est une familiarisation avec les données pour décider des éléments intéressant à signaler à d'autre personnes.

L'**analyse explicative** a pour le but d'expliquer un point précis, de raconter une histoire particulière - qui portera sur les deux-trois point clés.

5 étapes d’analyse des données

- Extraire - Obtenez les données à partir d'une feuille de calcul, de SQL, du Web, etc.
- Nettoyer - Nous pourrions utiliser des visuels exploratoires.
- Explorer - Nous utilisons des visuels exploratoires.
- Analyser - Nous pourrions utiliser des visuels exploratoires ou explicatifs.
- Partager - C'est ici que vivent les visuels explicatifs.

Trois questions clés

- À qui vous adressez-vous ?
- Que voulez-vous que votre public sache ou fasse ?
- Comment utiliser les données pour transmettre votre message ?

Qui

- Éviter de cibler un public trop large

- Réfléchir à la relation que vous entretenez avec votre public

Quoi

- Il faut toujours attendre quelque chose de vos auditeurs :
  - Qu’ils retiennent un point
  - Qu’ils entreprennent une action
- Vous êtes la personne qui connait le mieux les données
  - vous êtes le mieux placé pour interpréter les résultats
- Il est important de réfléchir au ton général que vous voulez instaurer dès le début.

Comment

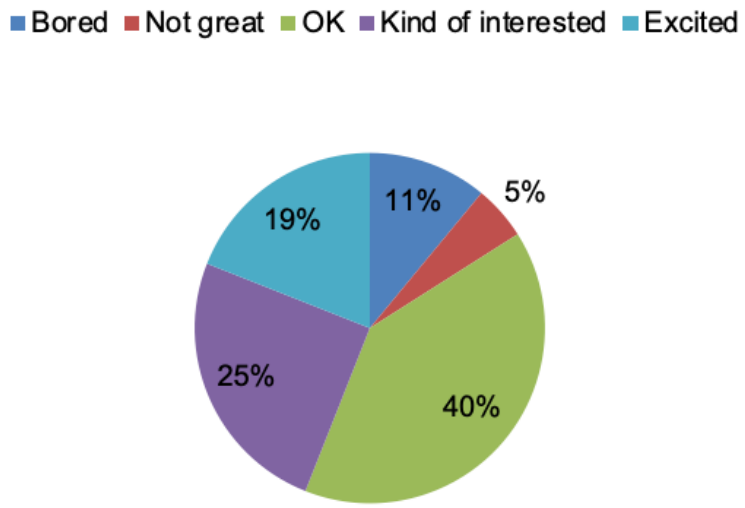
- Quelles sont, parmi les informations disponibles, celles qui vont vous permettre d’atteindre votre objectif ?

- Évitez de montrer seulement les données favorables
  - Manque de confiance à l’orateur

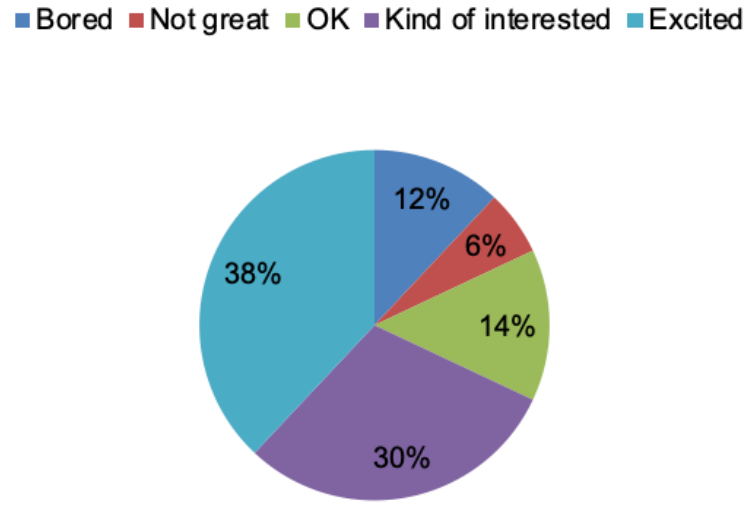
Exemple

Survey Results

PRE: How do you feel about doing science?

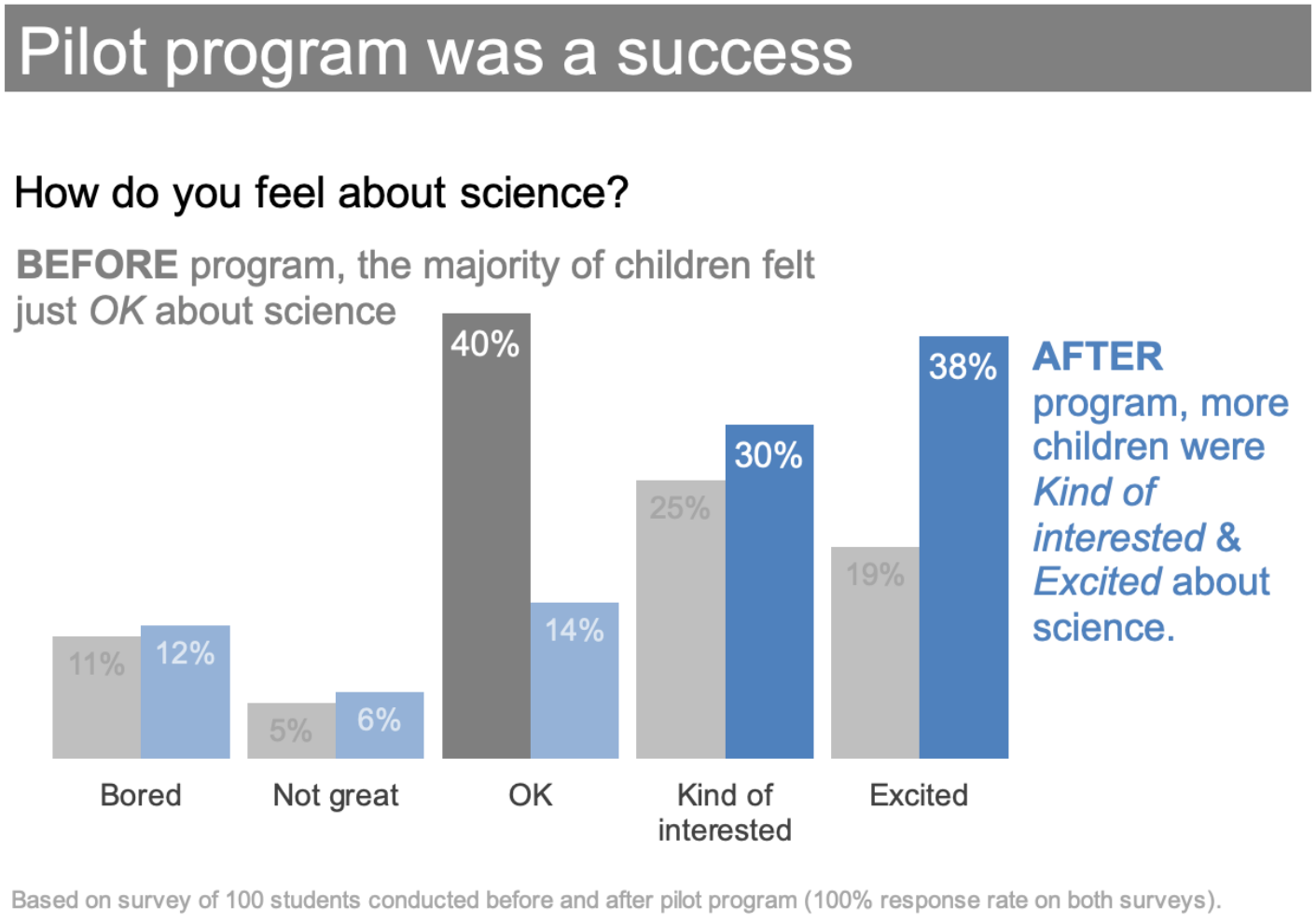


POST: How do you feel about doing science?



Exemple

- Qui : le comité budgétaire
- Quoi : succès ⟹ approuvez le budget
- Comment : démonter les effets positifs du programme pilote



Centrer le contexte

But : résumer les principaux éléments de ma communication en un seul paragraphe (une seule phrase)

- Le récit de 3 minutes
  - Transmettre le message au public en 3 minutes
- L'idée phare
  - Exprimer une seule opinion
  - Indiquer ce qui est en jeu
  - S'énoncer en une phrase complète

Le storyboard

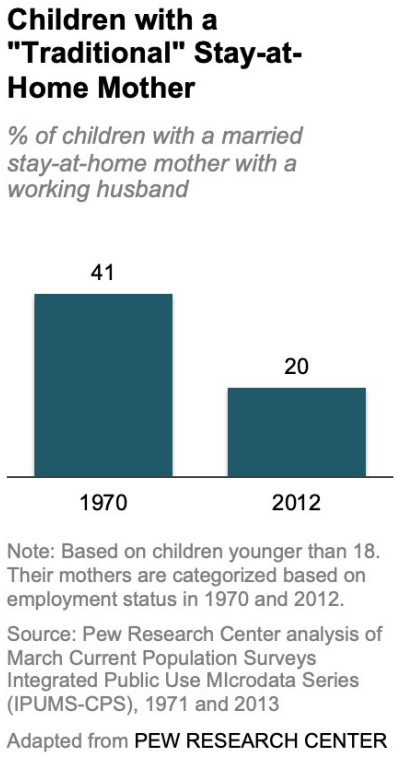
Le storyboard est un scénario sous forme d'images qui vous permet de préciser la structure de votre communication.

<u>Problème :</u> Les enfants ont une mauvaise opinion des sciences	Montrer le problème : présenter les notes des élèves durant l'année	Pistes pour résoudre le problème + projet pilote
Décrire le projet pilote : objectifs, organisation, …	Présenter les résultats de l'enquête avant/après pour démontrer la réussite du projet	Recommandation: Succès du projet pilote Développons le programme Nous avons besoin de €

Choisir un visuel

Seulement le texte

- Si votre objectif est de montrer une ou deux valeurs, utiliser les chiffres



Les tableaux

- Les tableaux permettent de communiquer avec un public mélangé, donc chaque membre va examiner la ligne qui l'intéresse.
- Les éléments graphiques doivent se fondre dans l'arrière-plan pour laisser la place centrale aux données.

Bordure épaisses

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Bordures claires

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Bordure minimales

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

La carte thermique

- La carte thermique permet de conserver le niveau de détail d'un tableau tout en y associant des indications visuelles.

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

LOW-HIGH

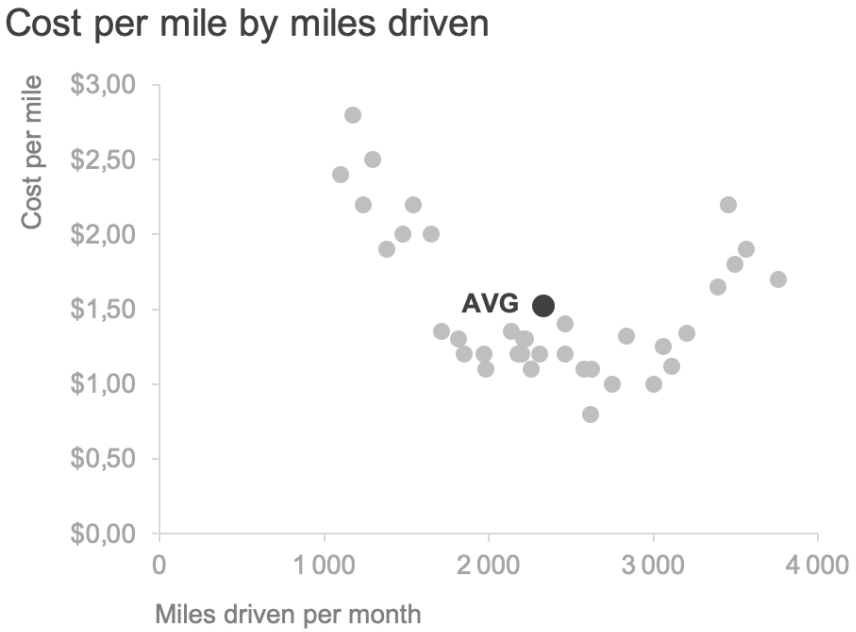
	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

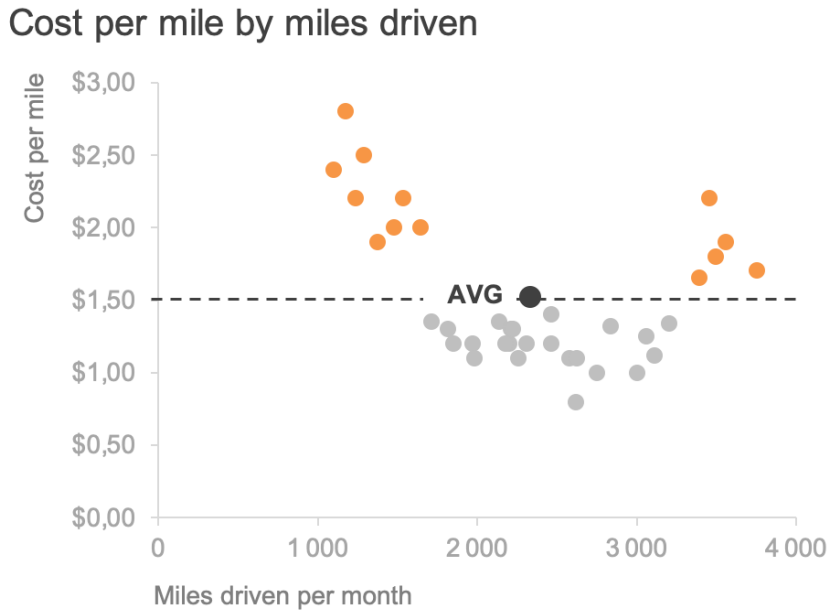
Les graphiques

- Utilise notre système visuel
- Accélère le traitement de l'information
- Les catégories de graphiques les plus utilisés :
  - Les points
  - Les courbes
  - Les diagrammes en bâtons
  - Les aires

Le nuage de points

Pour monter la relation entre deux variables



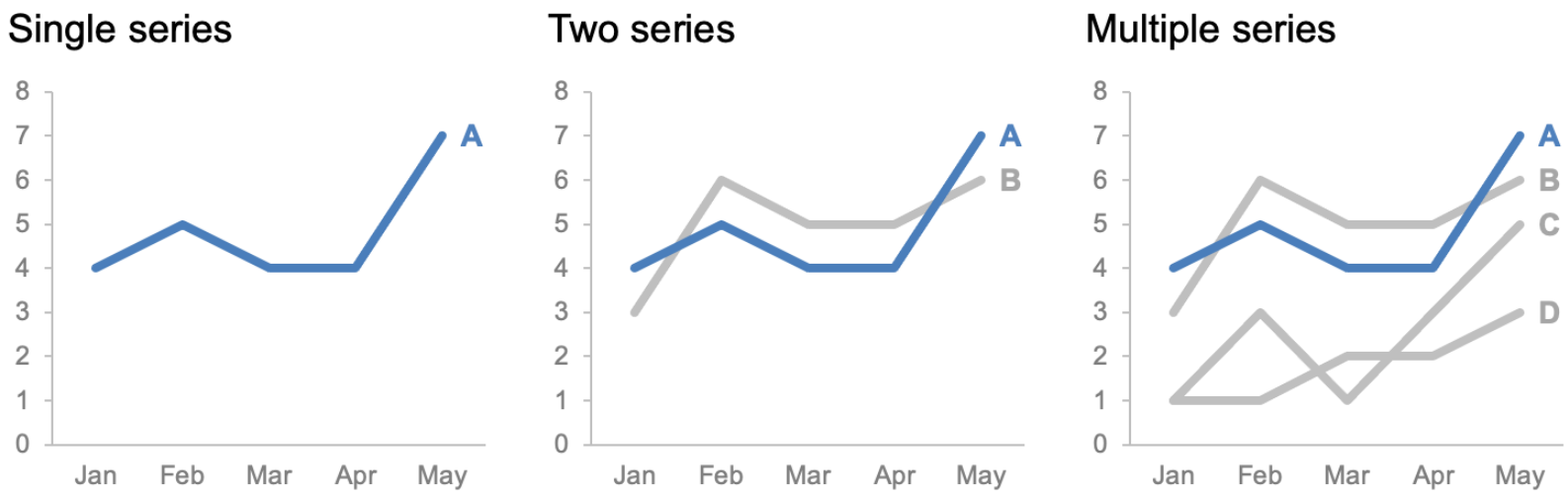


Les courbes

- Généralement réservées aux données continues
- Les deux types de courbes principalement utilisés :
  - La courbe classique
  - Le graphique en pente

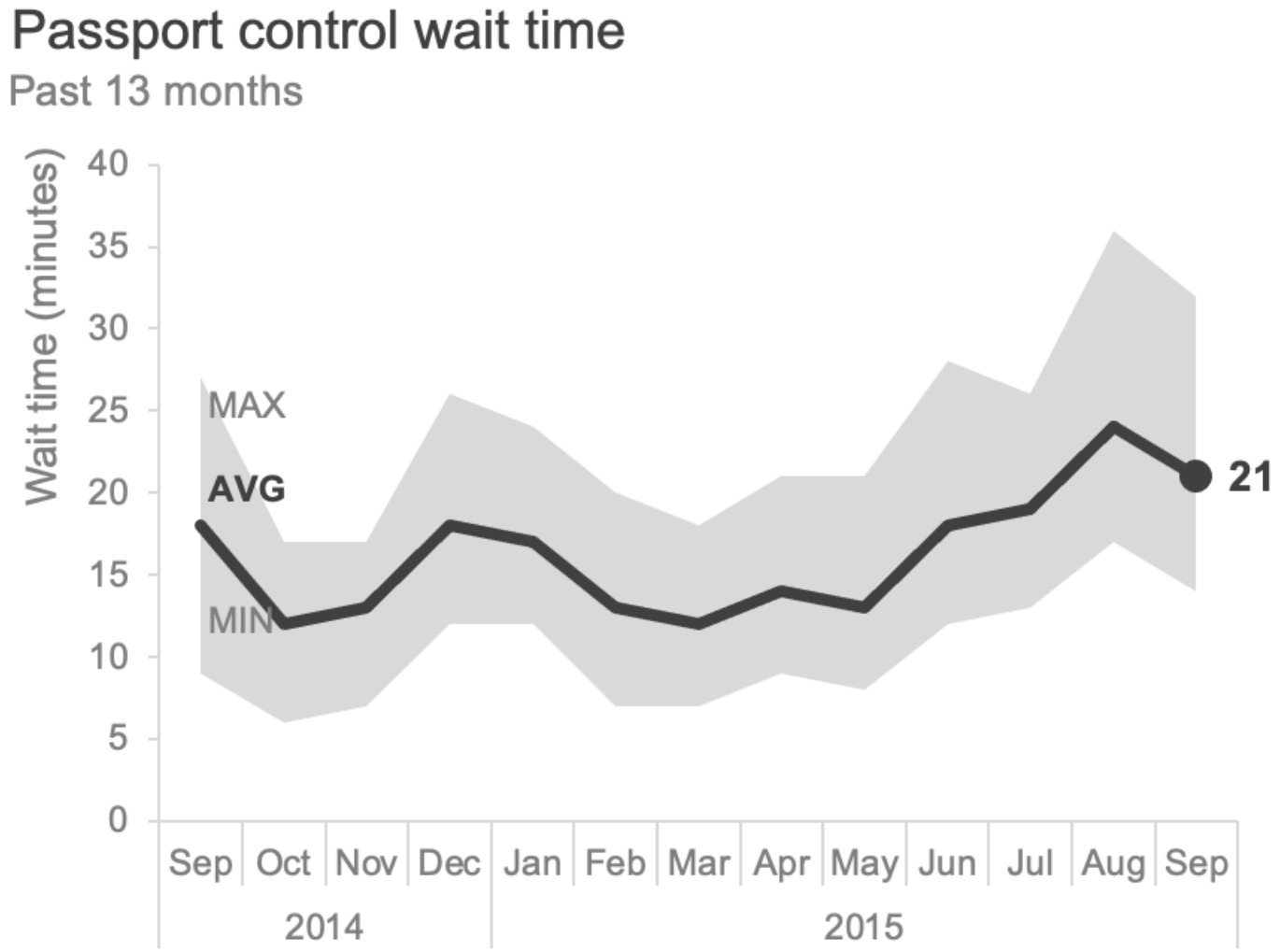
La courbes classique (1/2)

- Possible de représenter sur un même graphique une, deux ou plusieurs séries de données



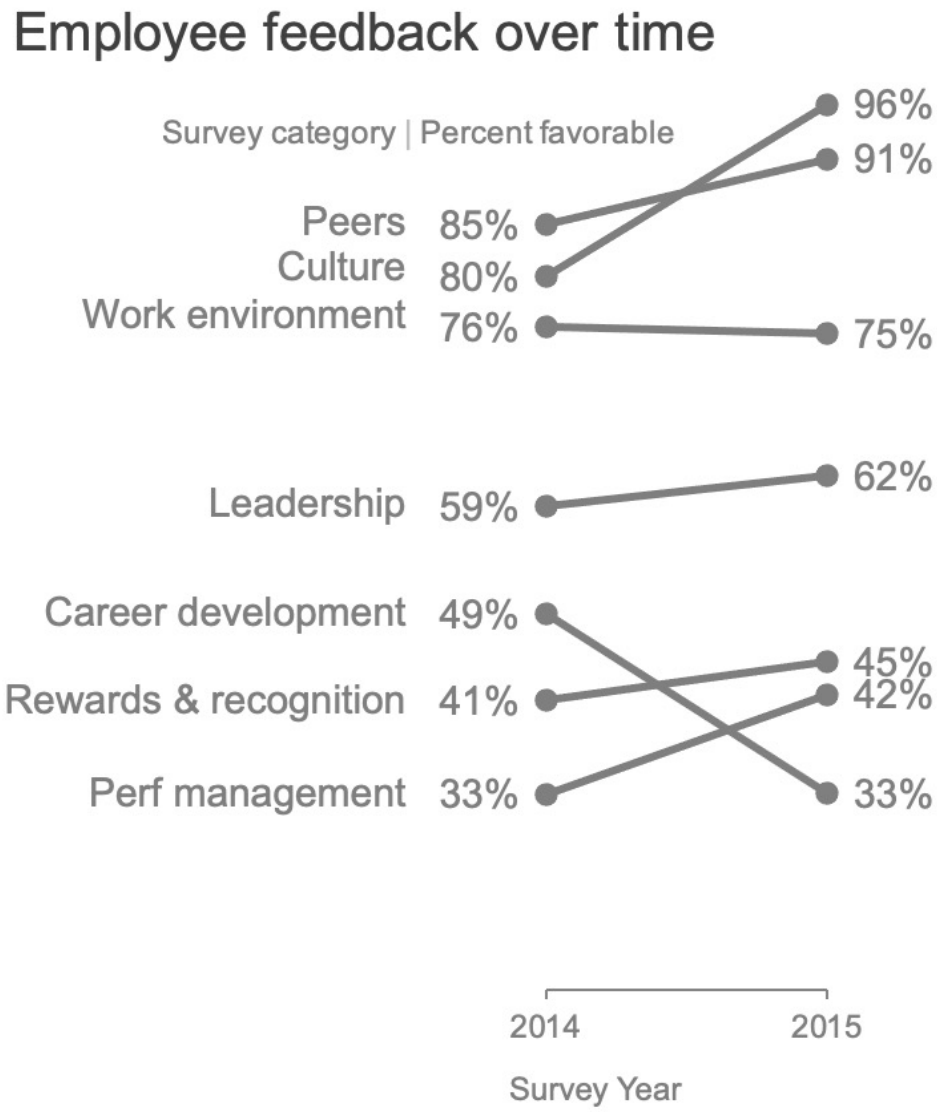
La courbes classique (2/2)

- Vous pouvez indiquer l'intervalle de confiance directement sur le graphique.



Le graphique en pente (1/2)

- Utilisé pour montrer rapidement les variations ou changements relatifs entre les deux états pour plusieurs catégories.

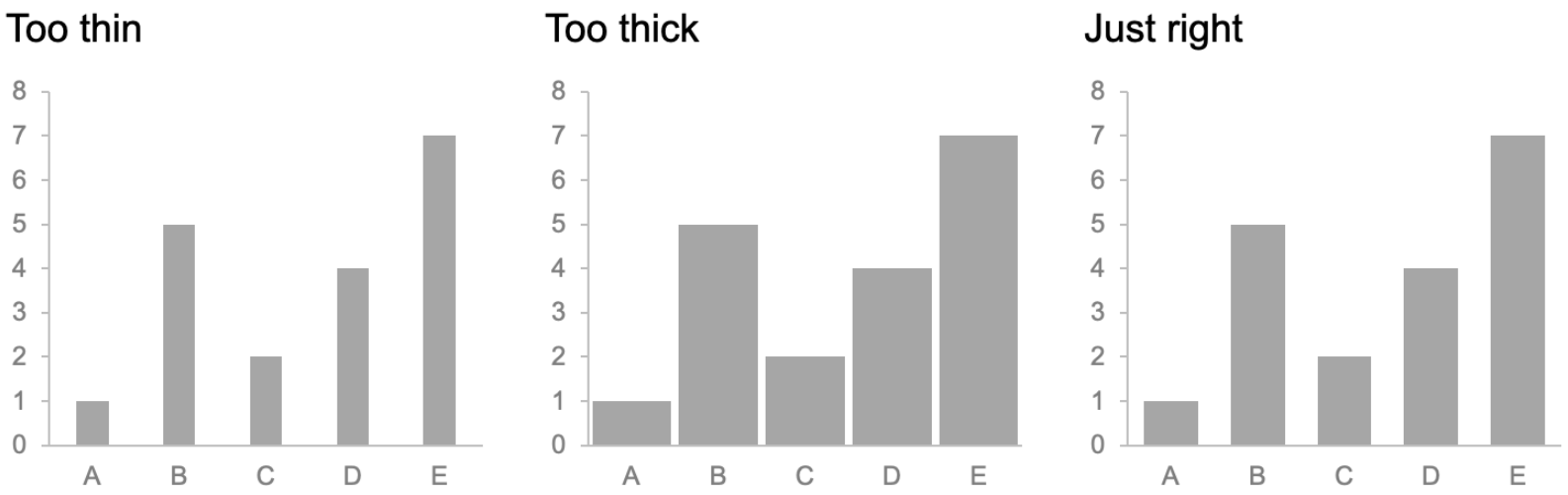


Le graphique en pente (2/2)

- La couleur peut attirer l'attention sur la catégorie en baisse.

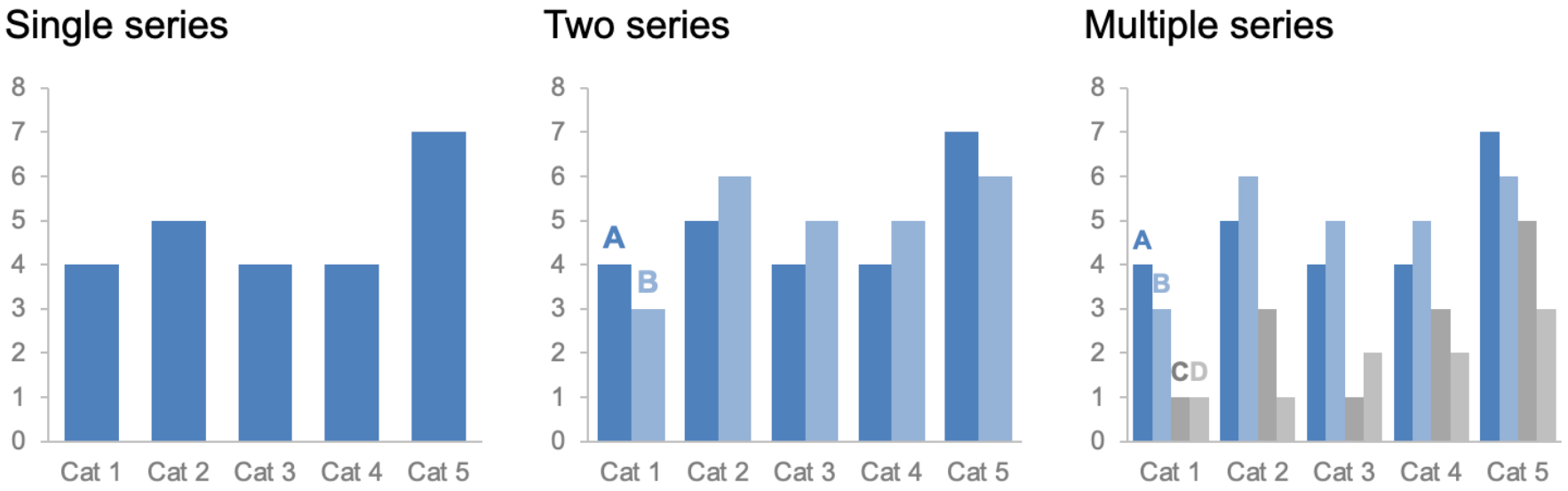
Les diagrammes en bâtons

- Demandent moins d'effort de compréhension de la part de l'assistance.
- Bien adaptés à notre vision
- Il est nécessaire que la base des bâtons soit fixée à zéro, pour éviter une comparaison visuelle erronée.



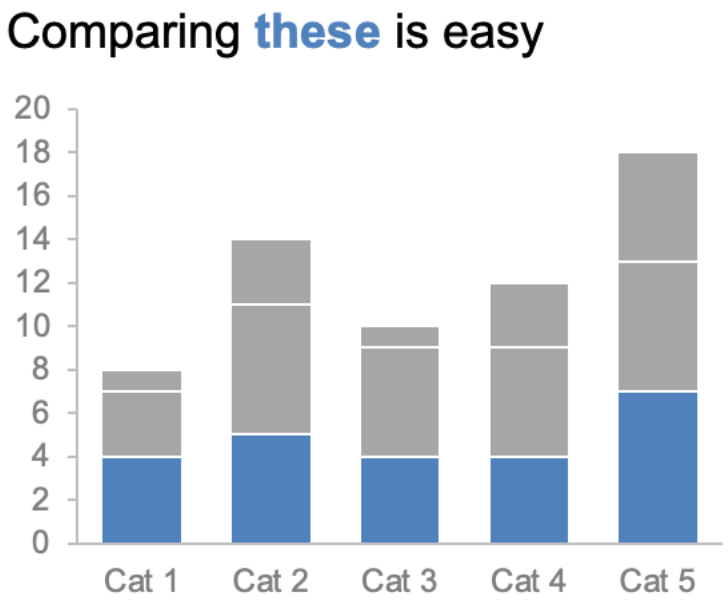
Le diagramme vertical

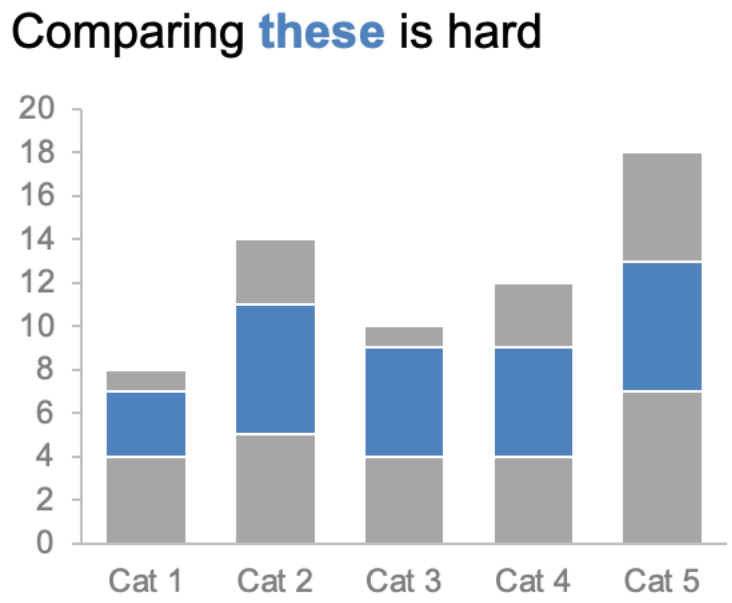
- Peuvent présenter une, deux ou plusieurs séries de données
- Pensez aux espaces entre les séries
- L'ordre des catégories est important pour les comparaisons souhaitées



Le diagramme vertical empilé

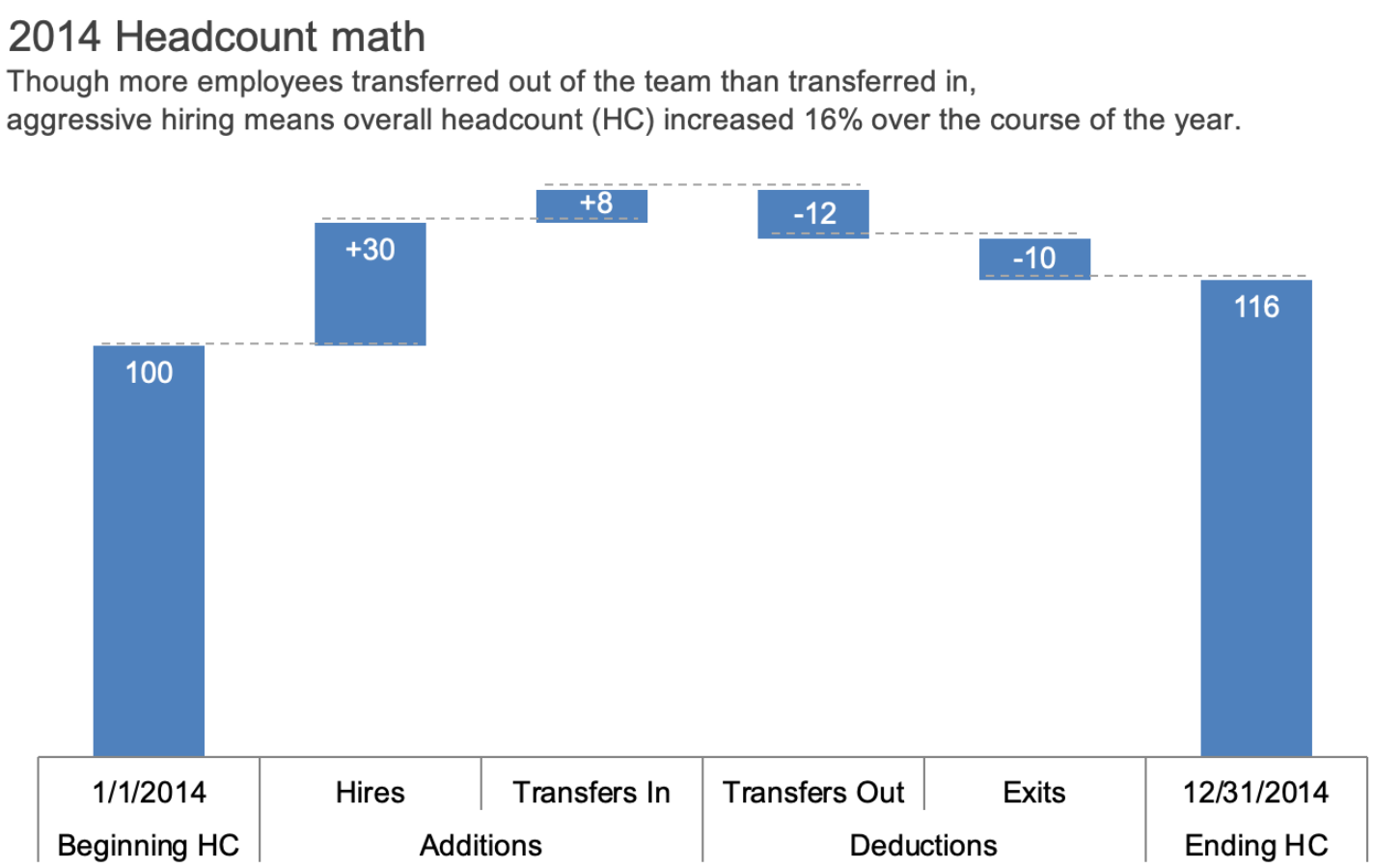
- Le but est de pouvoir comparer les différentes catégories en termes d'effectifs totaux et de sous-composantes.
- Peut être difficile à comparer à cause de couleurs choisis.





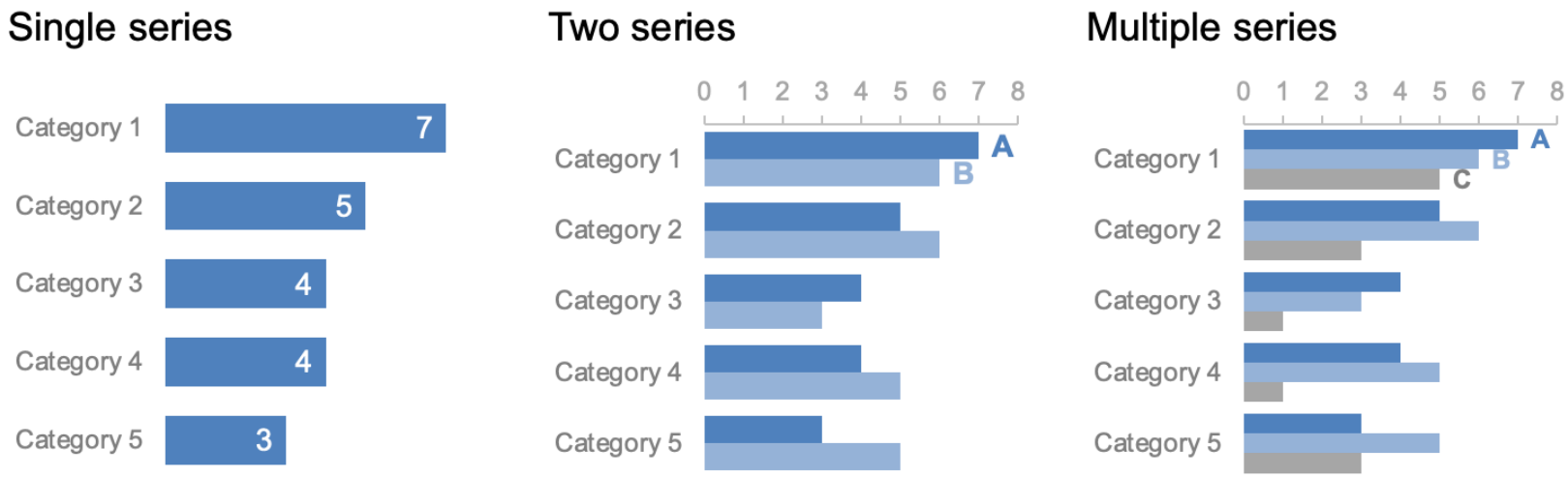
Le diagramme en cascade

- Ces sont des diagrammes empilés où on a dissocié les différents éléments dans l’espace pour pouvoir se concentrer sur un seul d’entre eux



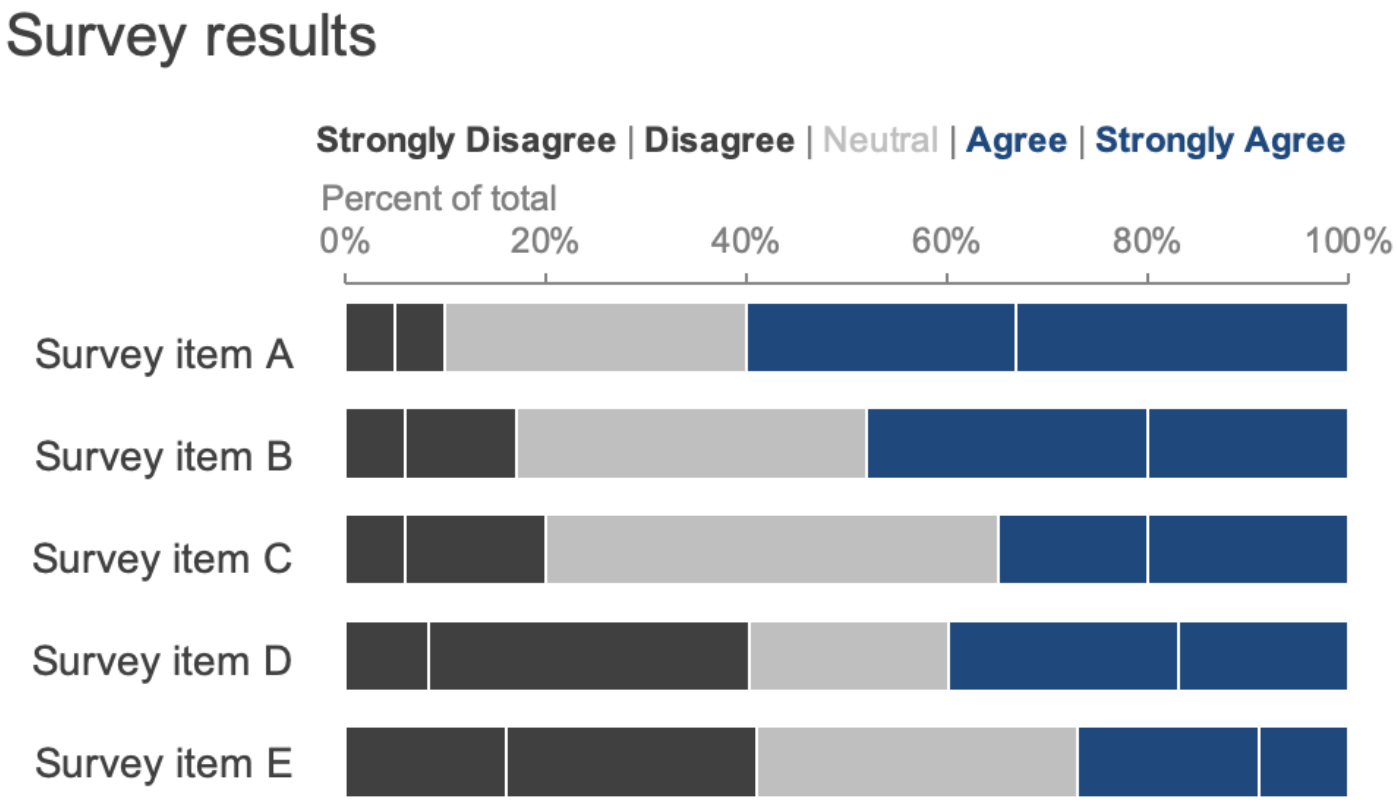
Le diagramme horizontal

- Il est extrêmement facile à lire.
- Il est utile si les intitulés des catégories sont longs.
- Il est dans le « bon » sens de lecture.



Le diagramme horizontal empilé

- Servent à présenter le total et les sous-composantes de plusieurs catégories.
- Peuvent indiquer soit les valeurs réelles, soit les pourcentages avec un total à 100%.



Les aires

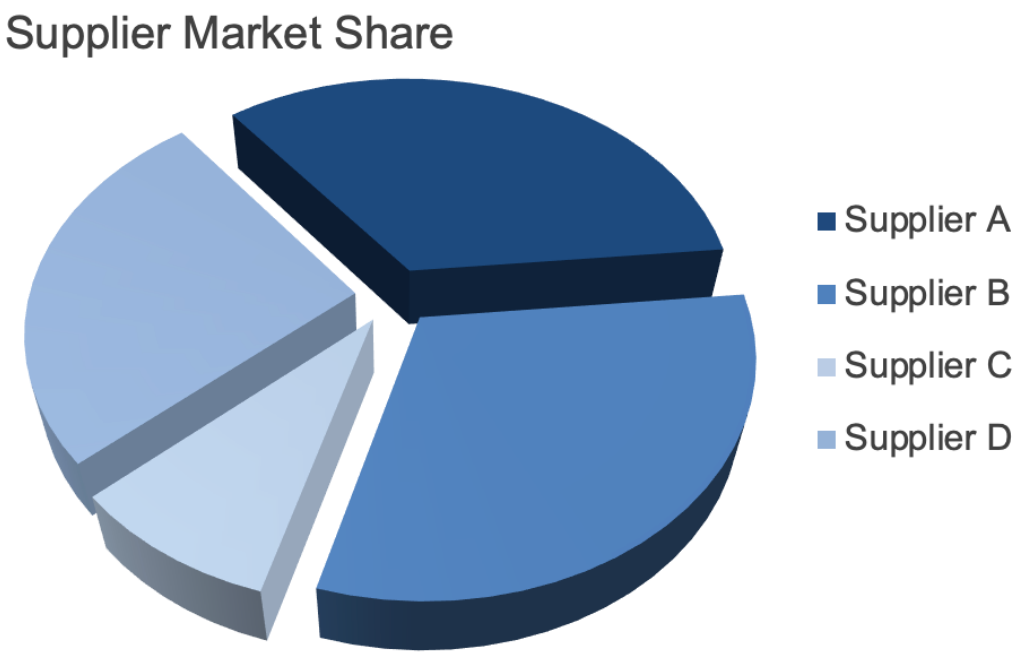
- L’œil humain n’est pas fait pour attribuer une valeur exacte à une surface.
- Cela rend les graphiques représentant des aires plus difficiles à lire.
- Utile s’il faut visualiser des nombres aux ordres de grandeur très différents.

Graphiques à éviter (1/5)

- Les diagrammes à secteurs ou en anneau
- L’effet 3D
- Le second axe des ordonnées

Graphiques à éviter (2/5)

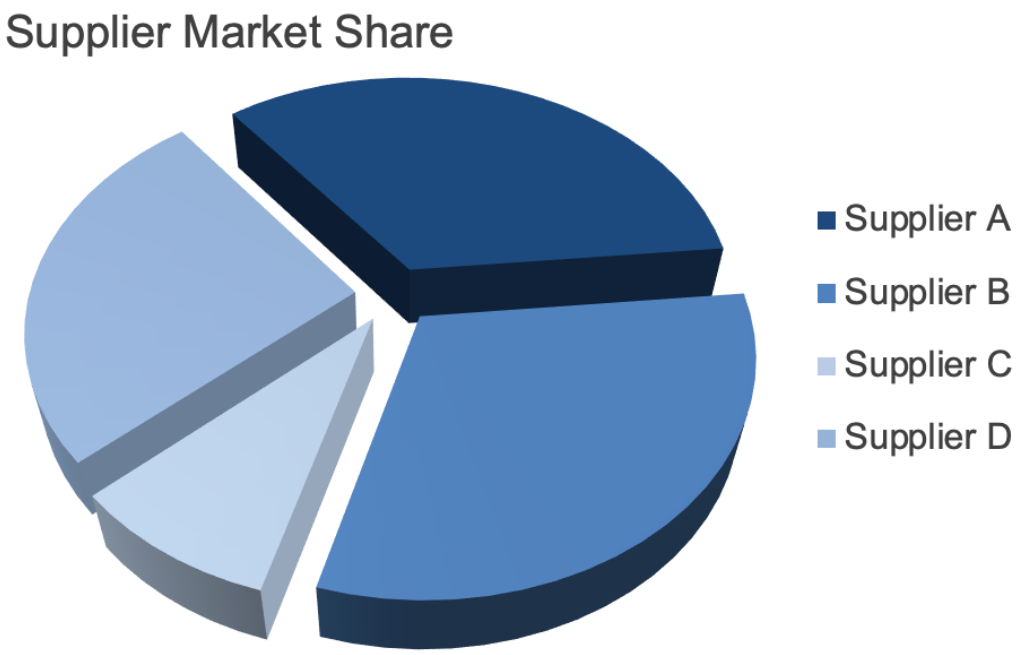
Les diagrammes à secteurs

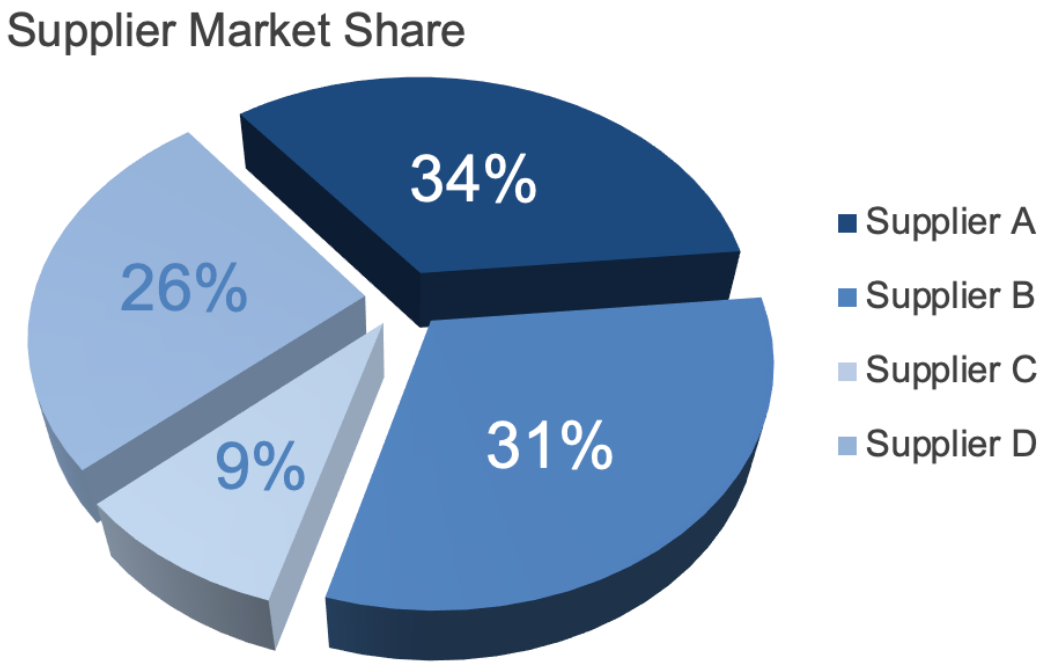


- Quel est le fournisseur le plus important d'après ce visuel ?
- Si vous deviez estimer la part que celui-ci représente sur le marché global, quel pourcentage indiqueriez-vous ?

Graphiques à éviter (3/5)

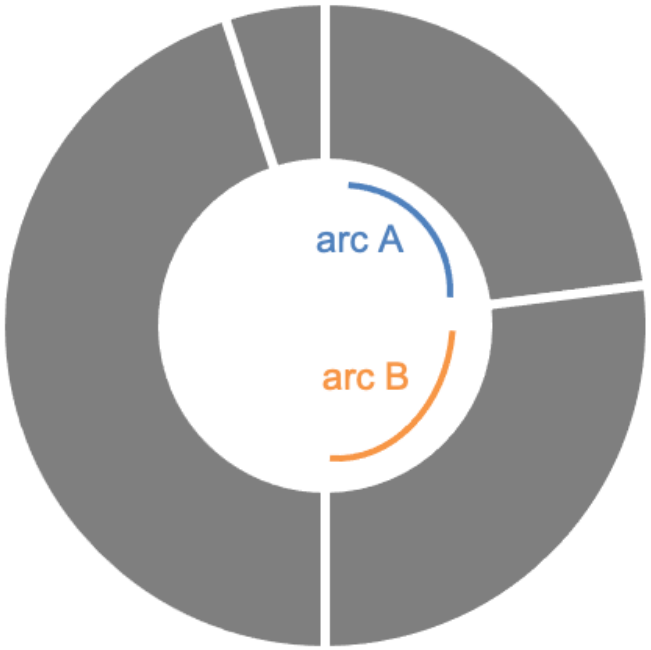
Les diagrammes à secteurs



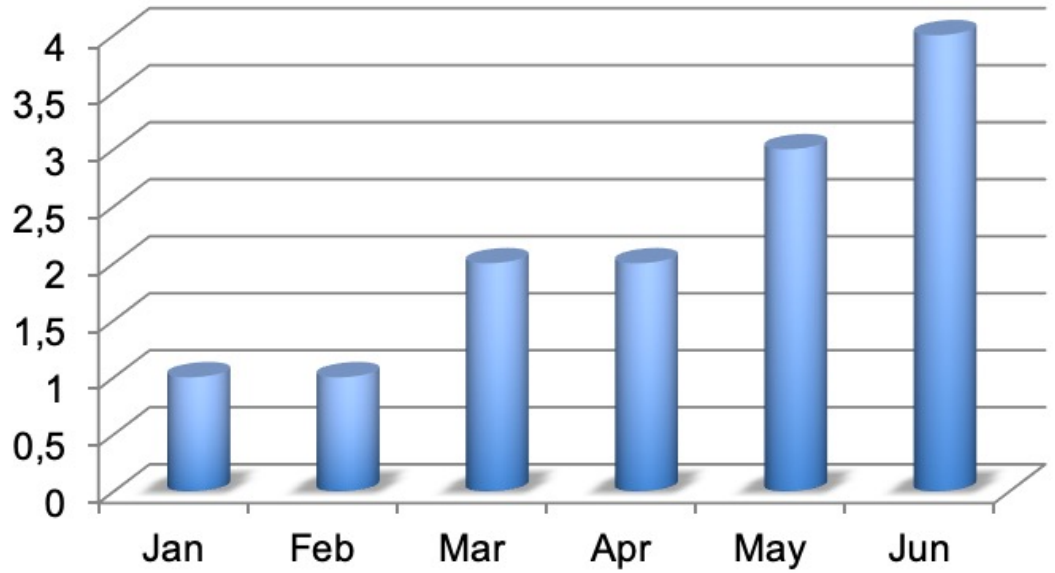


Graphiques à éviter (4/5)

Le diagramme en anneau

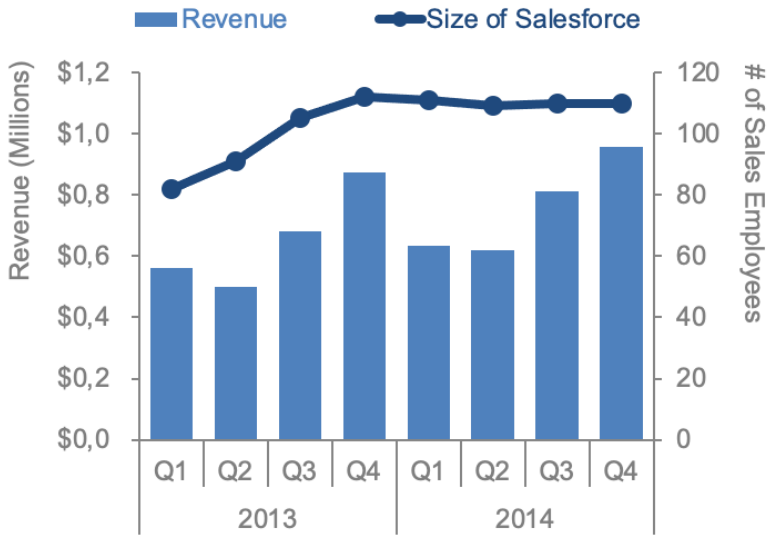


Nombre de problèmes

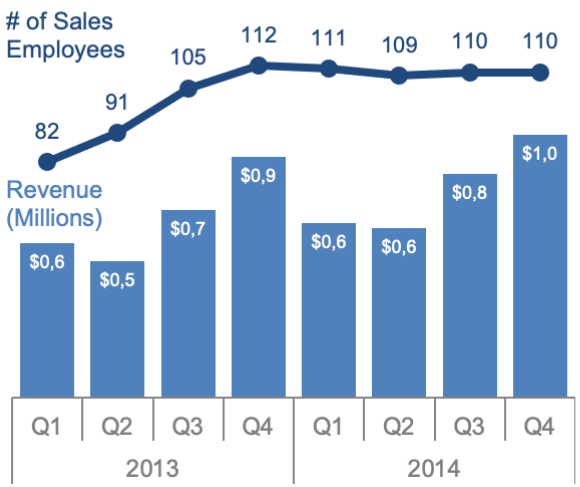


Graphiques à éviter (5/5)

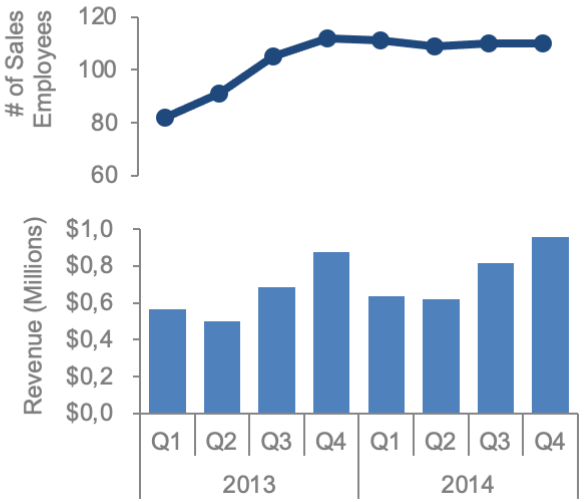
Secondary y-axis



Alternative 1: **label directly**



Alternative 2: **pull apart vertically**



Bilan

- Ces sont des types de graphiques les plus fréquemment utilisés.
- Ce n’est pas une liste exhaustive.
- Sélectionnez avant tout un type de graphique qui vous permet de faire passer clairement votre message au public.

Exercice 2

Imaginez que nous vendons des cookies et que nous gardons une trace du nombre de chaque type vendu.

Quel graphique choisir pour tracer le nombre de chaque type de cookie vendu ? Pourquoi ?

Cookie Type	Sold
Mint Cookies	309
Chocolate Chip Cookies	245
Peanut Butter Cookies	210
Toffee Cookies	140

Exercice 2 correction

Imaginez que nous vendons des cookies et que nous gardons une trace du nombre de chaque type vendu.

Quel graphique choisir pour tracer le nombre de chaque type de cookie vendu ? Pourquoi ?

Cookie Type	Sold
Mint Cookies	309
Chocolate Chip Cookies	245
Peanut Butter Cookies	210
Toffee Cookies	140

Diagramme verticale

Utilisation de « dashboard »

- Démontrer plusieurs aspects de données sur un seul graphique
- Quand nous avons plusieurs variables à prendre en compte
- Utilisation de « visual encodings » (encodages visuels)

Hans Rosling examines global trends in health and wealth

Encodages visuels

- Axe X
- Axe Y
- Taille
- Forme
- Texture
- Angle
- Longueur

Hans Rosling examines global trends in health and wealth

Exercice 3 : Encodages visuels

Vidéo Youtube : **Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four**

Encodages visuels	Donnés
Axe X	
Axe Y	
Taille	
Couleur	
Mouvement	
Type de graphique	

Exercice 3 correction

Encodages visuels	Donnés
Axe X	Revenu
Axe Y	Durée de vie
Taille	Population
Couleur	Pays (géographie)
Mouvement	Temps
Type de graphique	Nuage de points

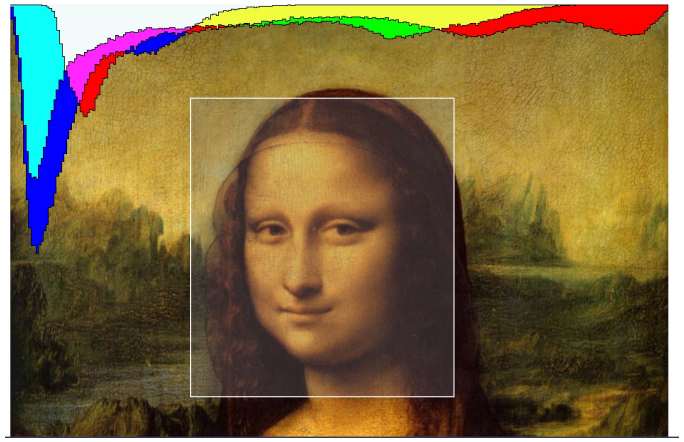
Logiciels

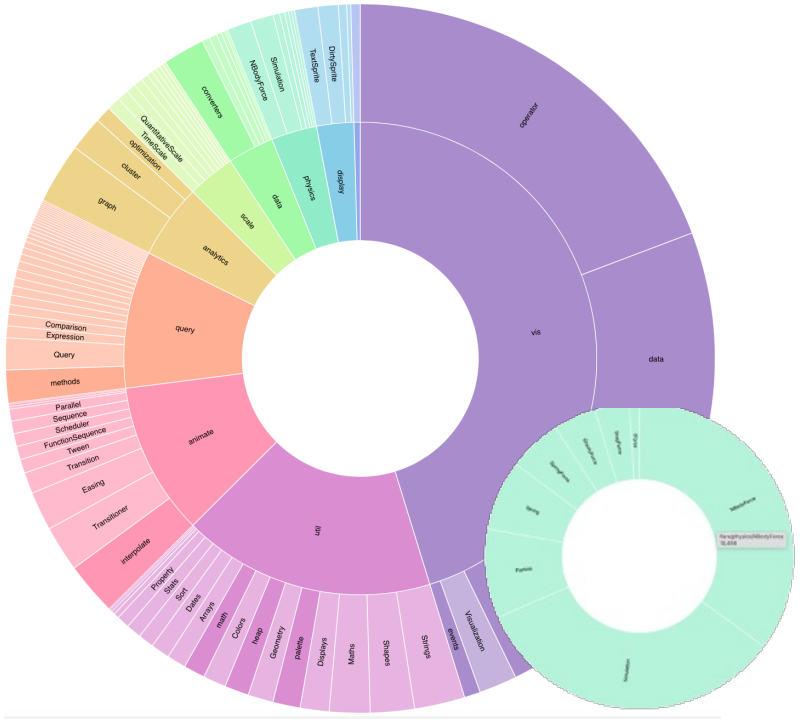
Logiciels de traitements de données

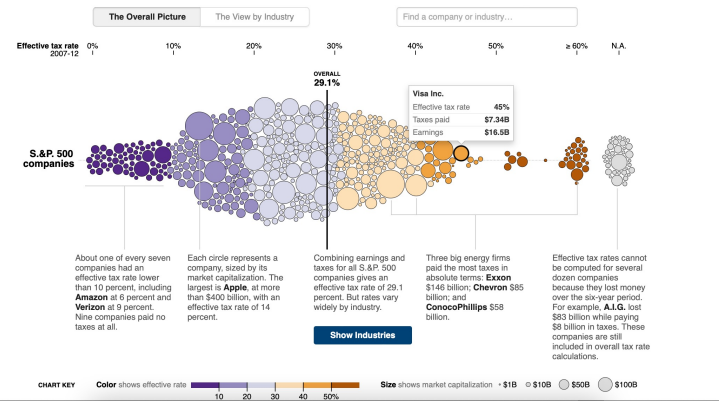
- Microsoft Excel
- Libre Office Calc
- Tableu Software

Bibliothèque D3.js

- D3.js

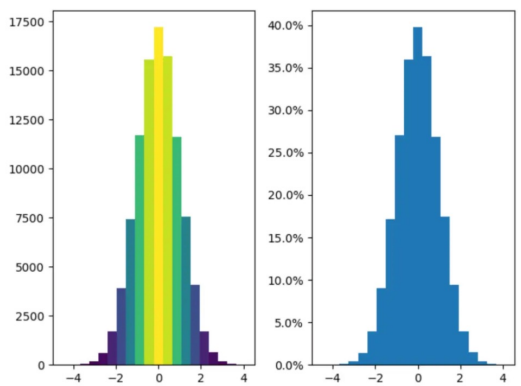


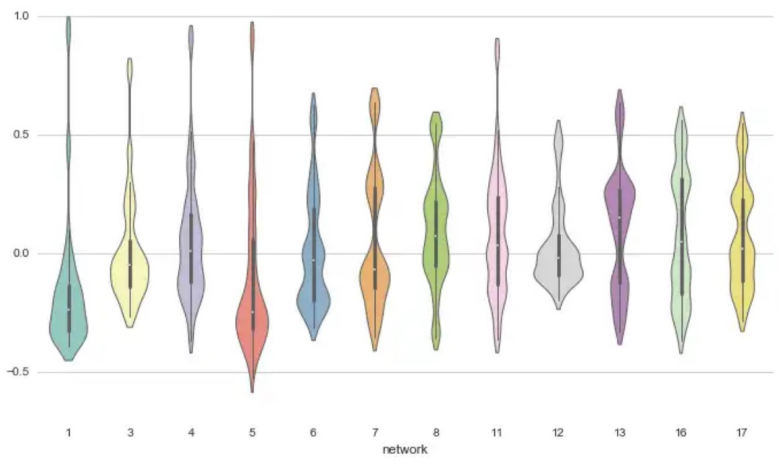


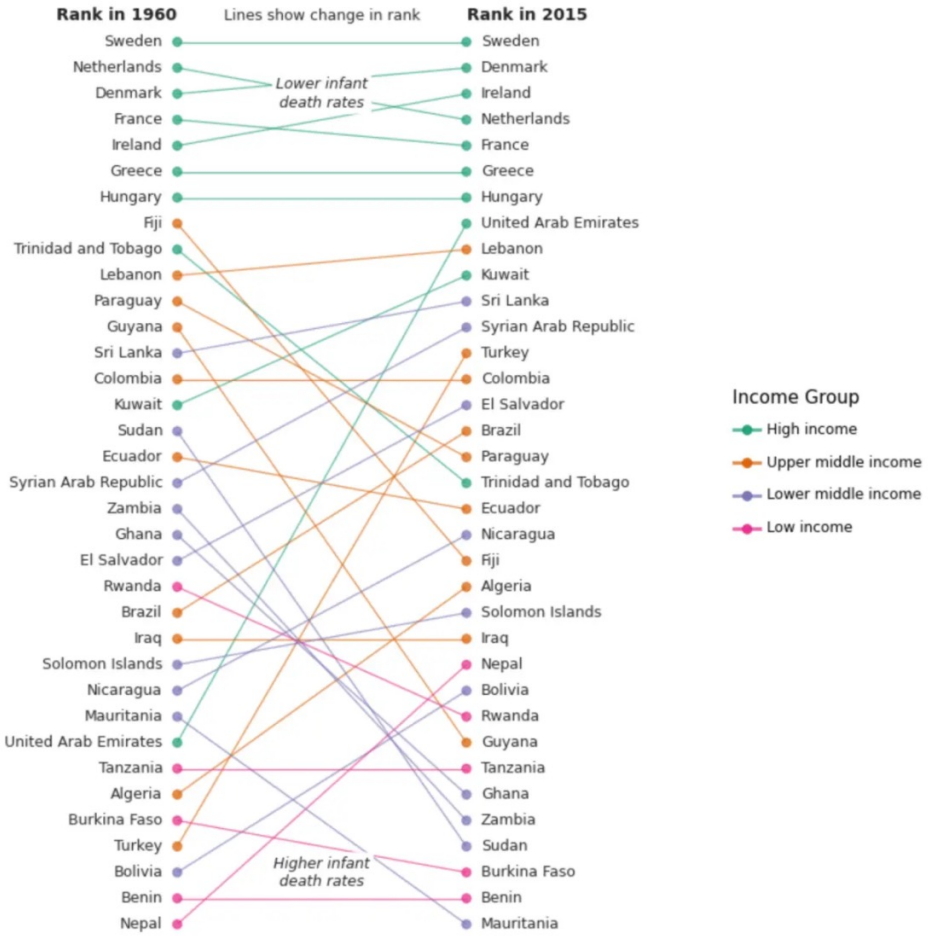


Bibliothèques de Python

- Matplotlib
- Seaborn
- Plotnine(ggplot)
- Bokeh
- pygal
- Plotly
- geoplotlib
- Gleam
- missingno
- Leather
- Altair
- Folium







Travaux pratique

Exercice 1 : Analyse exploratoire et explicative

Exercice 2 : Types de données

Exercice 3 : Comparaison des visualisations des données

Exercice 4 : Comparaison des graphiques

Exercice 5 : Installer Python + savoir programmer en Python pour jeudi 🙃

La page du cours :

https://perso.liris.cnrs.fr/itkachenko/pages/idsm_dataviz.html