

# Big Data

# Open Data

MMI 2 – TP#7 S4

Danielo **JEAN-LOUIS**

# Graphiques / Datavisualisation

- plot / graph / chart en anglais
- Permettent d'illustrer des données sous forme graphique
- Utilisés par les data-analystes, l'équipe marketing...
- Peut rendre ses idées, rapports très clairs... ou les rendre incompréhensibles

# Graphiques / Datavisualisation

- Utilisés :
  - Phase d'exploration
  - Phase d'interprétation

*Un excellent graphique est celui qui donne au spectateur le plus grand nombre d'idées avec le moins d'encre possible, dans le plus petit espace.*

---

Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

Edward R. Tufte, Professeur de statistiques à l'université de Yale

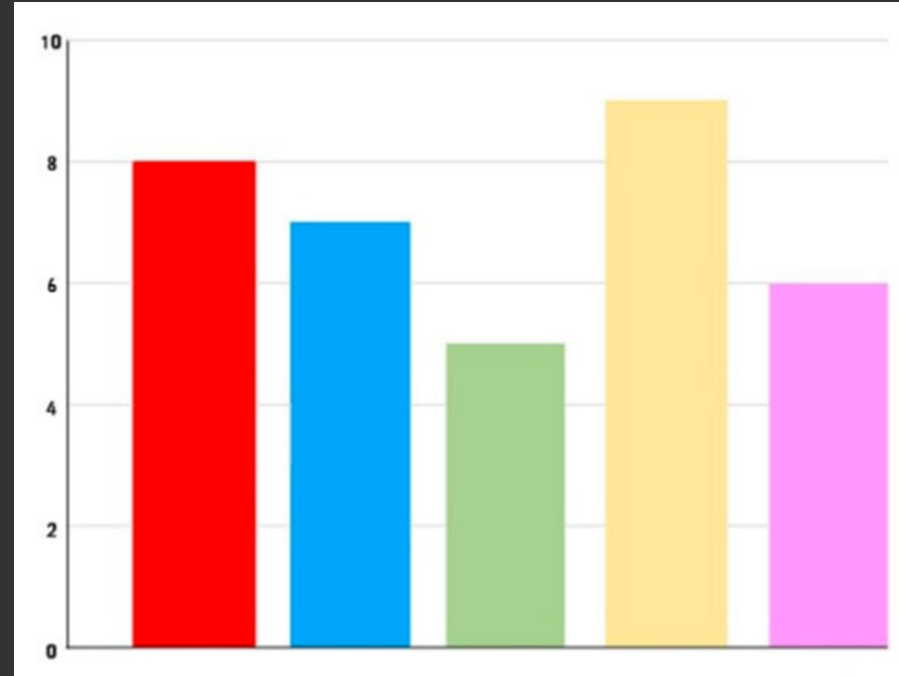
# Pratiquons ! - Graphiques

Pré-requis :

- Avoir la ressource `ressources/graphiques.jpg`
- A télécharger ici :  
<https://downgit.github.io/#/home?url=https://github.com/DanYellow/cours/tree/main/big-data-s4/travaux-pratiques/numero-7>

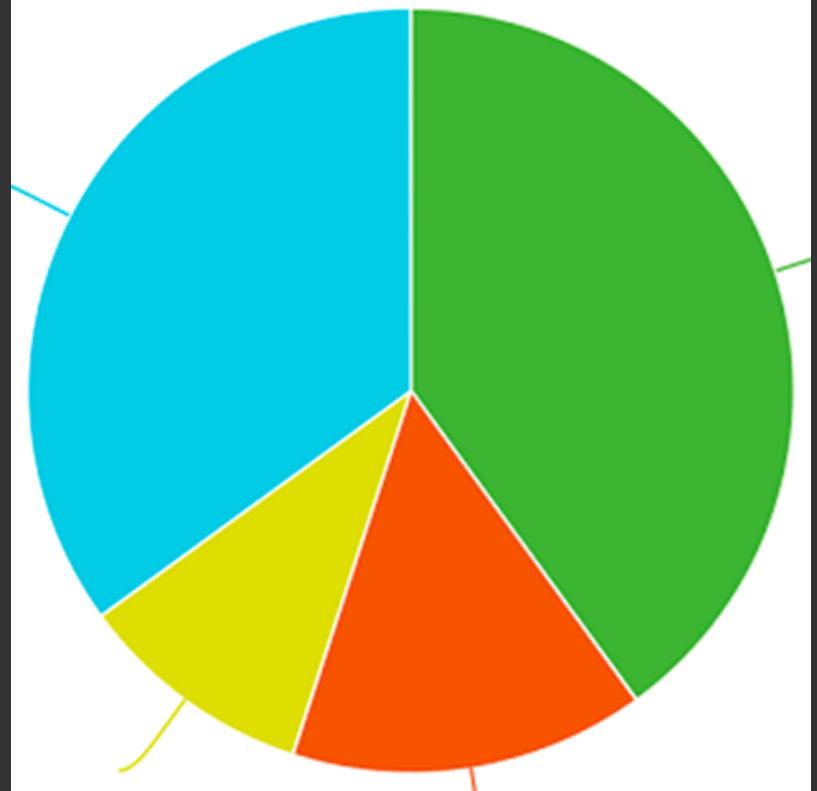
# Diagrammes en bâtons (Bar chart)

- Très facile à lire
- Permet de comparer des données
- Existe aussi sous la forme horizontale, groupée ou encore groupée et empilée



# Diagramme circulaire (Pie chart)

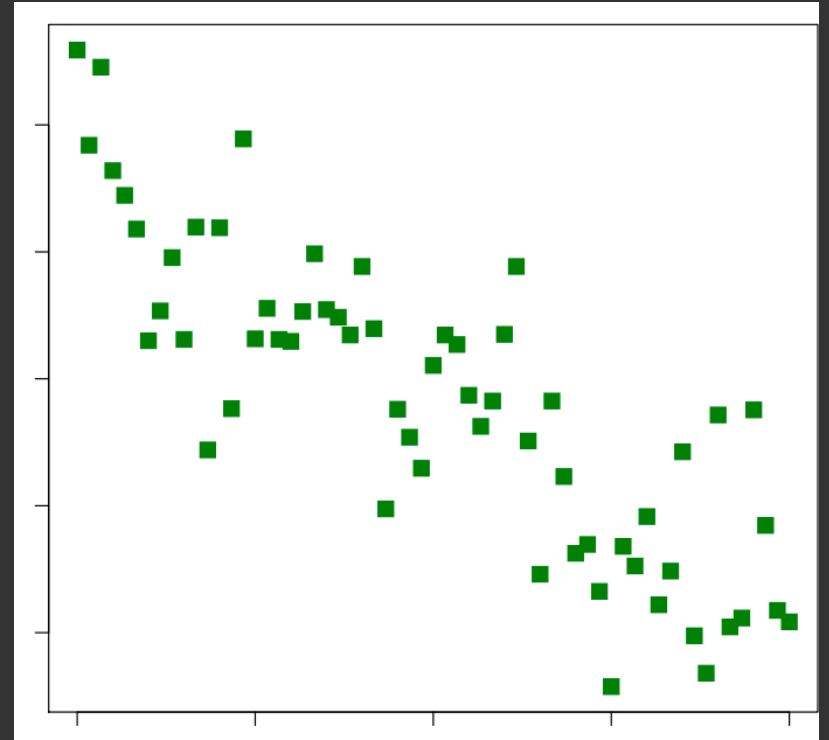
- Appelé aussi "camembert"
- Peut devenir très vite illisible
- Limiter le nombre d'entrées
- Permet de représenter la composition/répartition de qqchose
- Peut être représenté sous forme de bâtons





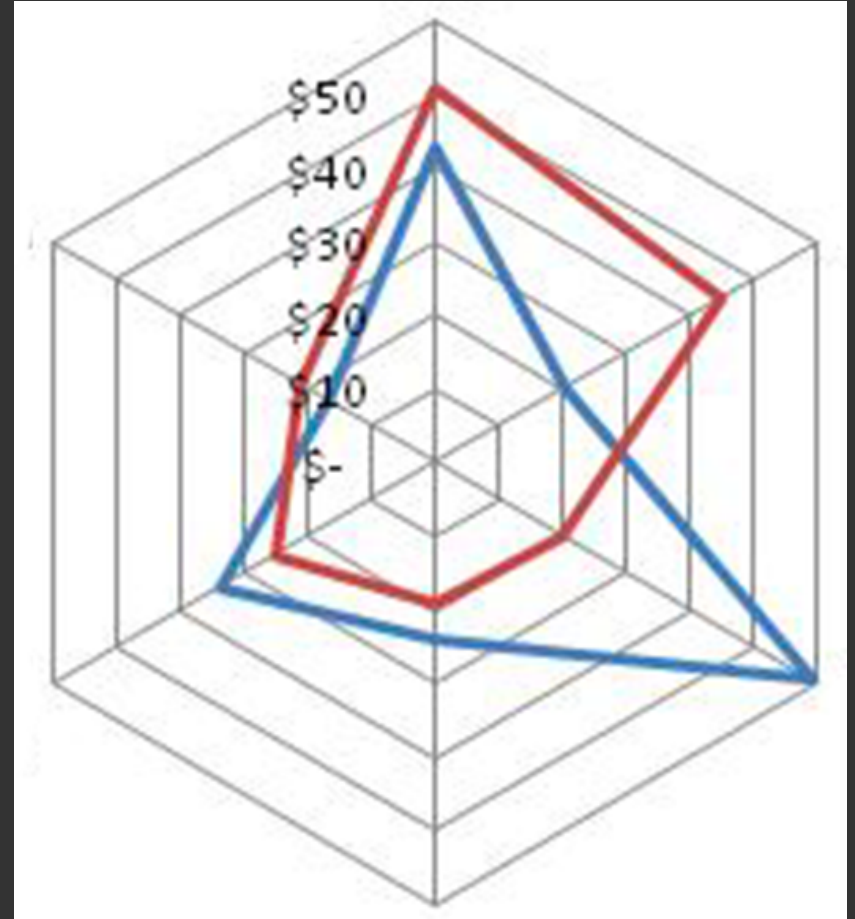
# Nuage de points (scatter plot)

- Peut représenter une corrélation entre deux variables
- Les deux variables doivent être numériques
- Souvent utilisé avec une régression linéaire



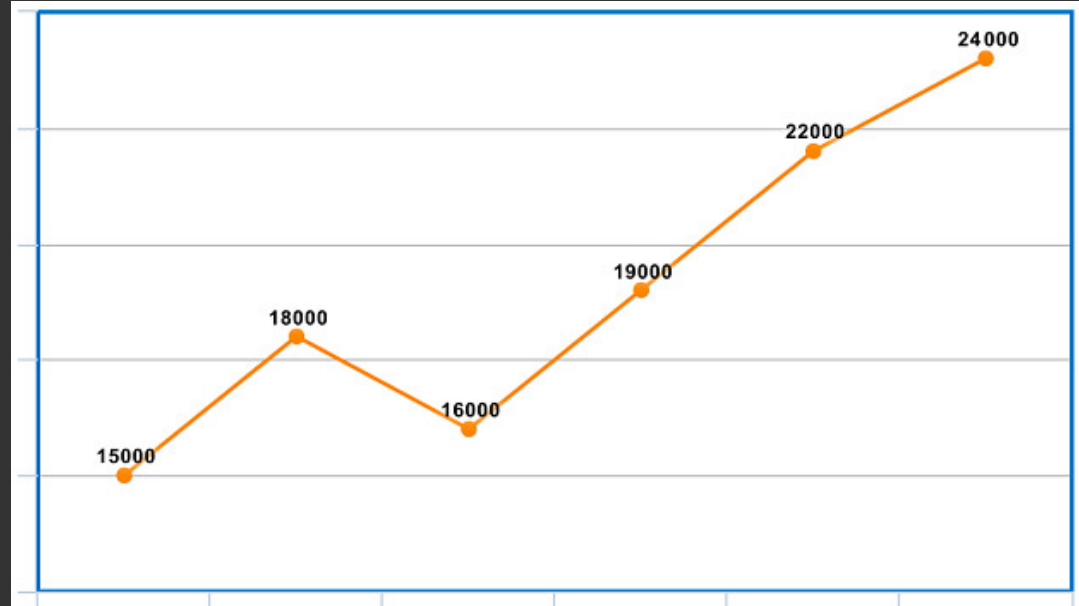
# Diagramme de Kiviat (Radar chart)

- Nécessite au moins 3 variables
- Permet de visualiser la composition/répartition de plusieurs variables avec de multiples comparaisons
- Idéal pour montrer/comparer des performances



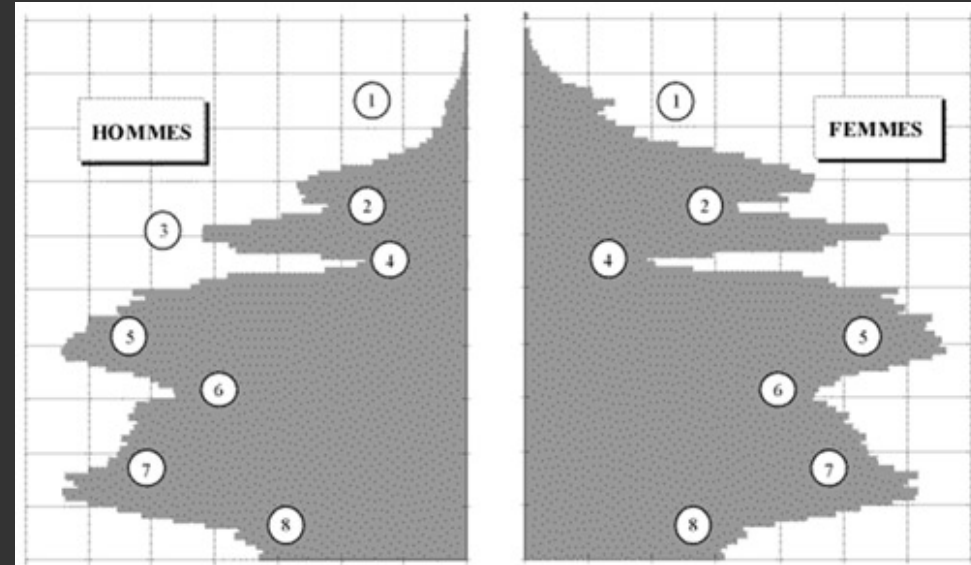
# Courbe (line chart)

- Facile à lire
- Visualise l'évolution d'une catégorie
- Adapté aux données temporelles
- Possibilité d'avoir plusieurs courbes sur le même graphique



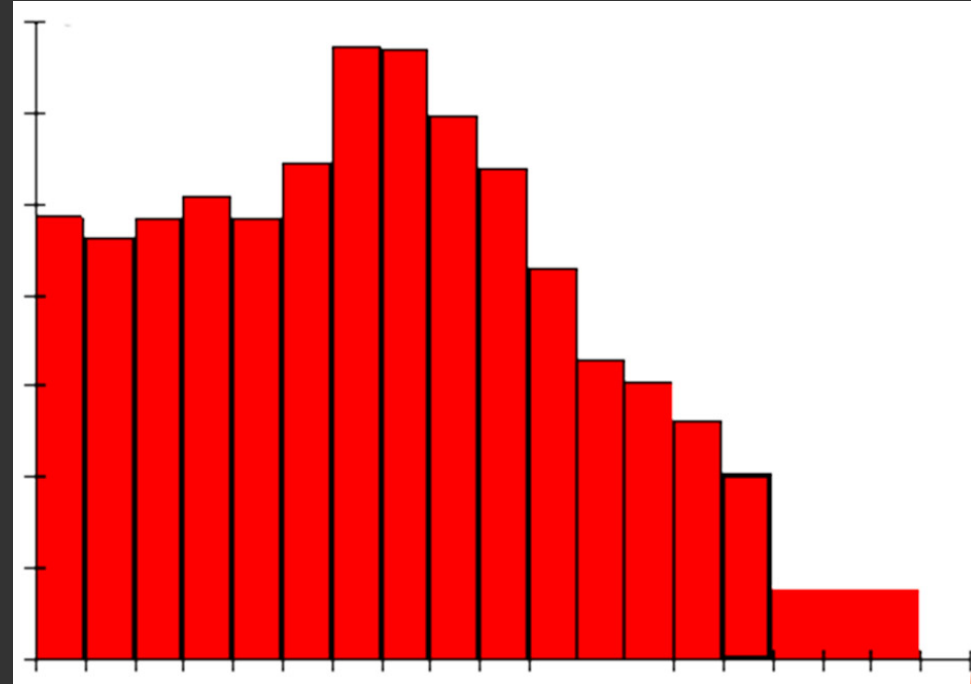
# Pyramide (des âges) (Population pyramid)

- Montre souvent la distribution de l'âge dans différents groupes
- Forme de toupie en coupe
- Données doivent être continues



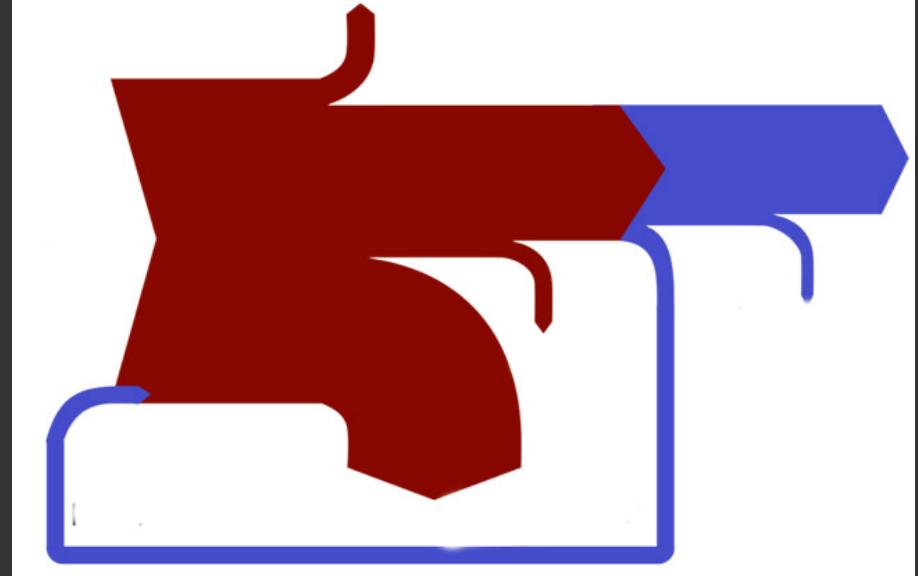
# Histogramme (histogram)

- A ne pas confondre avec le diagramme en bâtons
- Montre la fréquence de distribution d'une valeur
- Les données doivent être continues



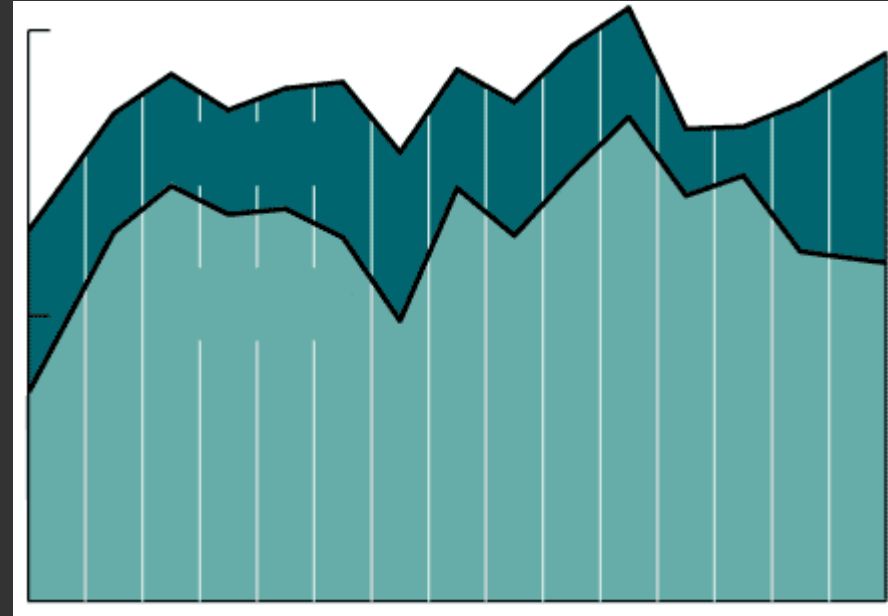
# Diagramme de Sankey (Sankey diagram)

- Flèches sont proportionnelles au flux
- Initialement utilisé dans le domaine de l'énergie
- Visualise le processus d'un flux



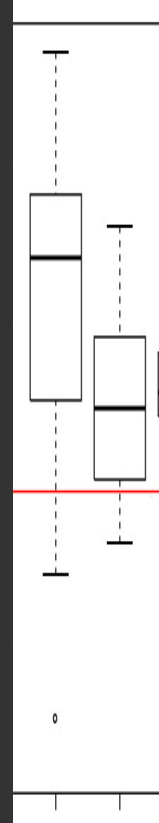
# Graphique en aires (Area chart)

- Attention au contraste des couleurs
- Met en évidence l'amplitude de la variation entre plusieurs éléments



# Boite à moustaches / boîte de Tukey (box plot)

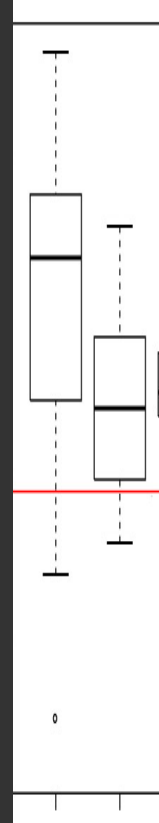
- Nécessite des notions en statistiques pour le lire
- Version "évoluée" : violins-plots



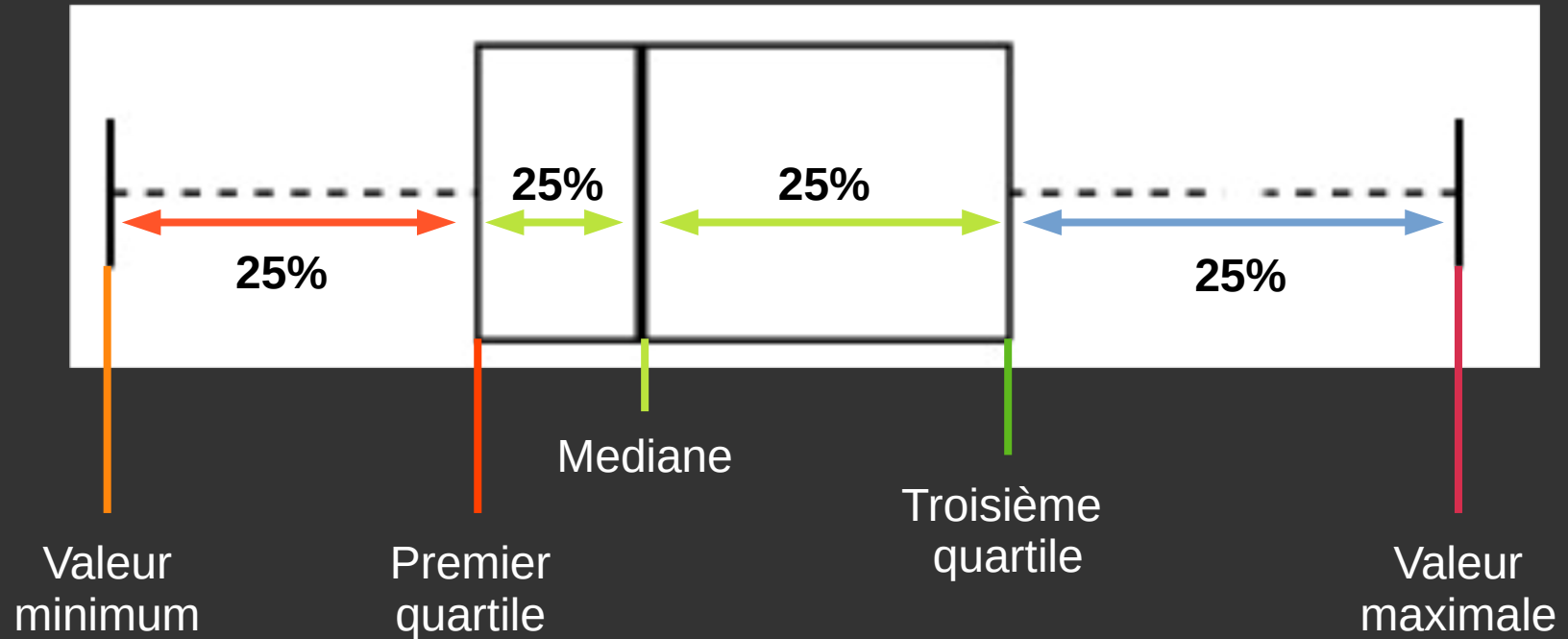


# Boite à moustaches / boîte de Tukey (box plot)

- Nécessite des notions en statistiques pour le lire
- Version "évoluée" : violins-plots
- Les points représentent les données aberrantes (outliers en anglais)



# Boite à moustaches – en détails

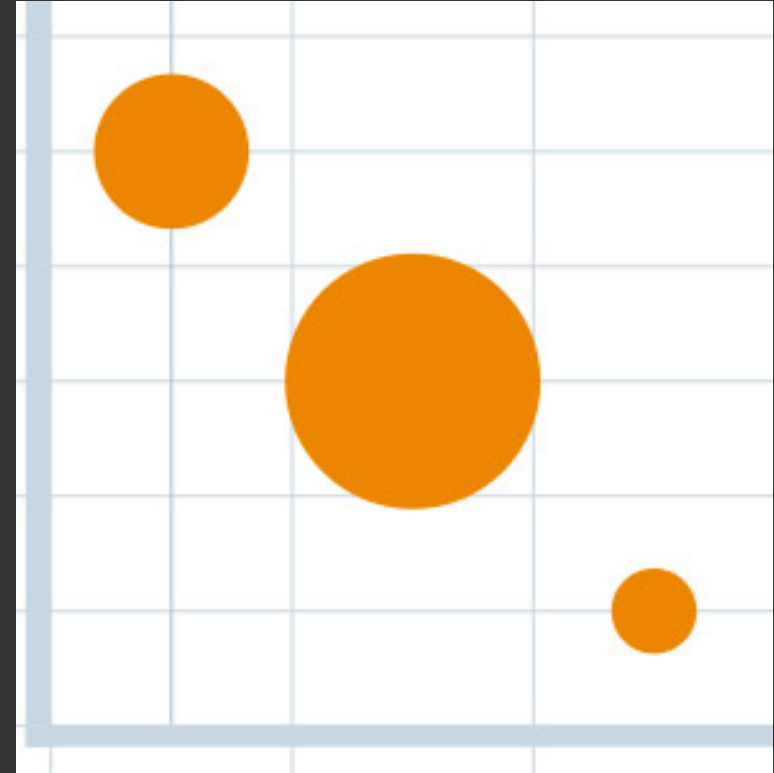


## Sources :

- <https://www.youtube.com/watch?v=oRMzUlhoy6E>
- <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51> - anglais

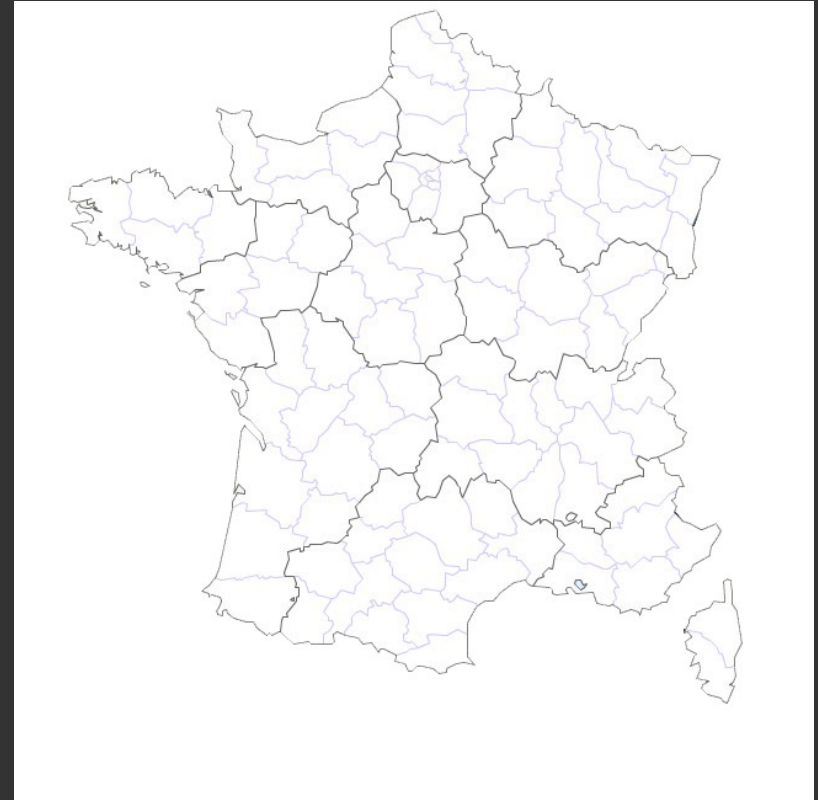
# Graphique à bulles (Bubble chart)

- Visualisation 3D
  - Possibilité d'utiliser trois catégories
- "Cousin" du nuage de points
- La taille de la bulle est proportionnelle à la valeur qu'elle représente

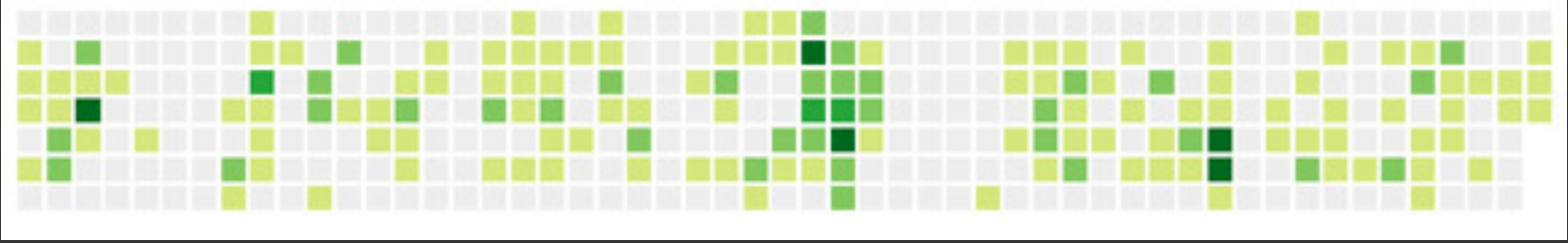


# Carte (map)

- Permet de faire une comparaison géographique
- Doit être couplé avec un autre graphique :
  - carte de chaleur (heatmap)
  - graphique à bulles
  - carte choroplèthe



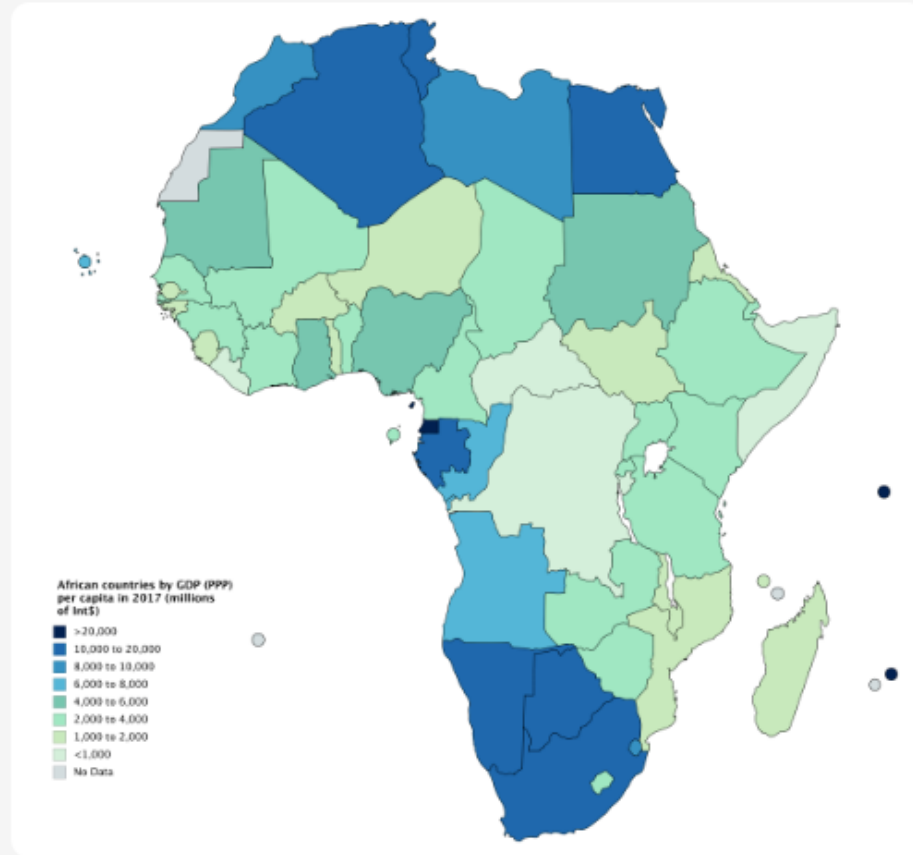
# Carte de chaleur (heatmap)



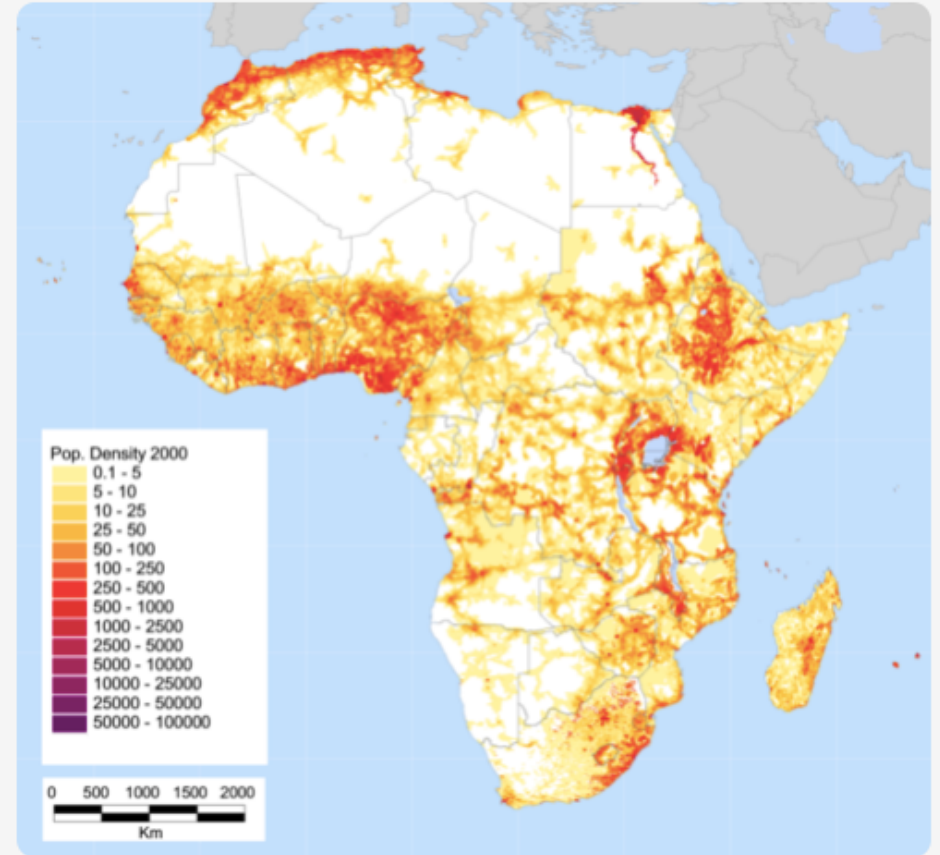
- Montre une relation entre deux variables
- Joue sur l'intensité des couleurs (attention aux couleurs choisies) pour représenter la donnée
- Peut prendre plusieurs formes
- Ne pas confondre avec la "carte choroplèthe"

# Différence entre carte de chaleur et carte chloroplète

Choropleth

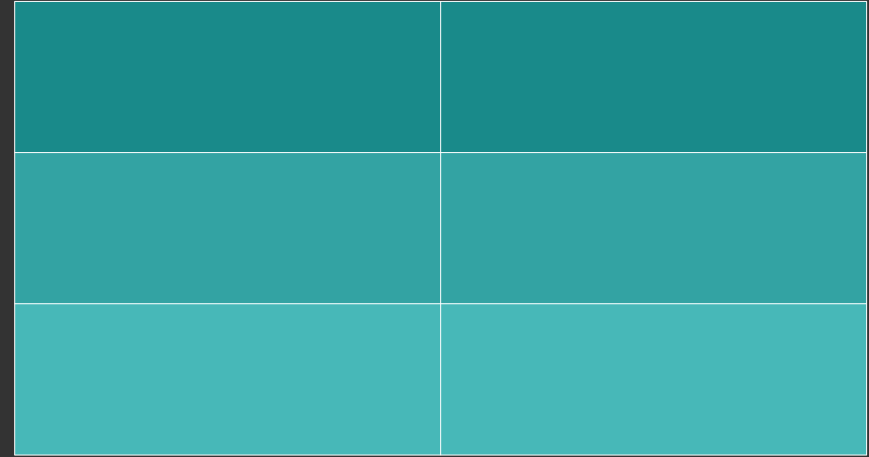


Heatmap



# Tableau

- Représente la donnée sous forme de colonnes et de lignes
- Adapté pour un gros volume de colonnes
- Possibilité d'utiliser des nuances de couleurs pour rendre le tout plus clair



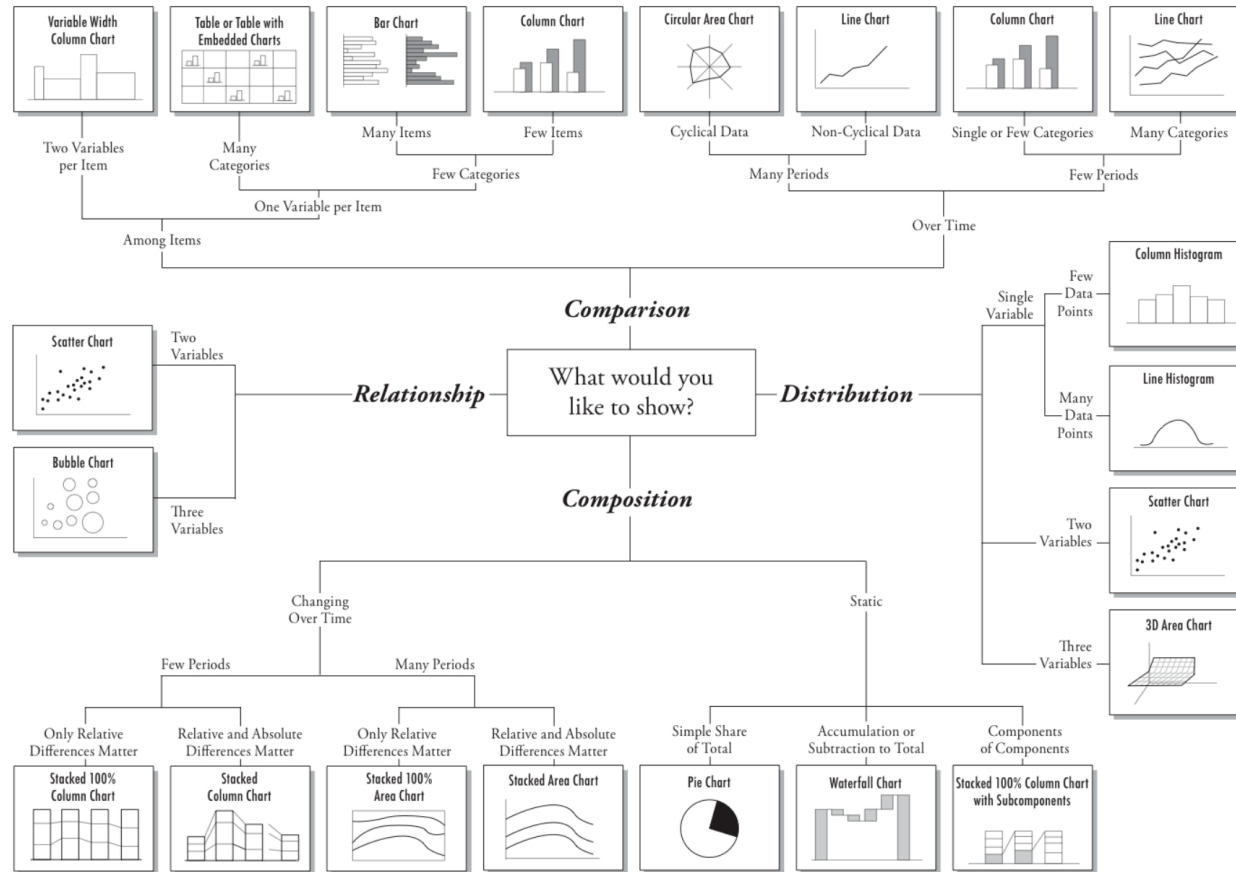

# En résumé (non exhaustif)

Type de graphique	Quand l'utiliser
Courbe	Montrer l'évolution d'une catégorie (souvent en fonction du temps)
Bâtons	Comparer des données
Histogramme	Montrer la fréquence de valeurs d'une catégorie
Camembert	Montrer la composition/répartition d'une catégorie
Radar / Kiviat	Comparer la composition de multiples variables dans plusieurs catégories
Boite à moustache	Comparer/montrer la répartition des valeurs <b>Attention : nécessite des notions en statistiques pour le lire</b>
Carte	Montrer des données à un niveau géographique <b>Attention : nécessite d'être combiné avec un autre type de graphique</b>



# En résumé

## Chart Suggestions—A Thought-Starter



# En résumé

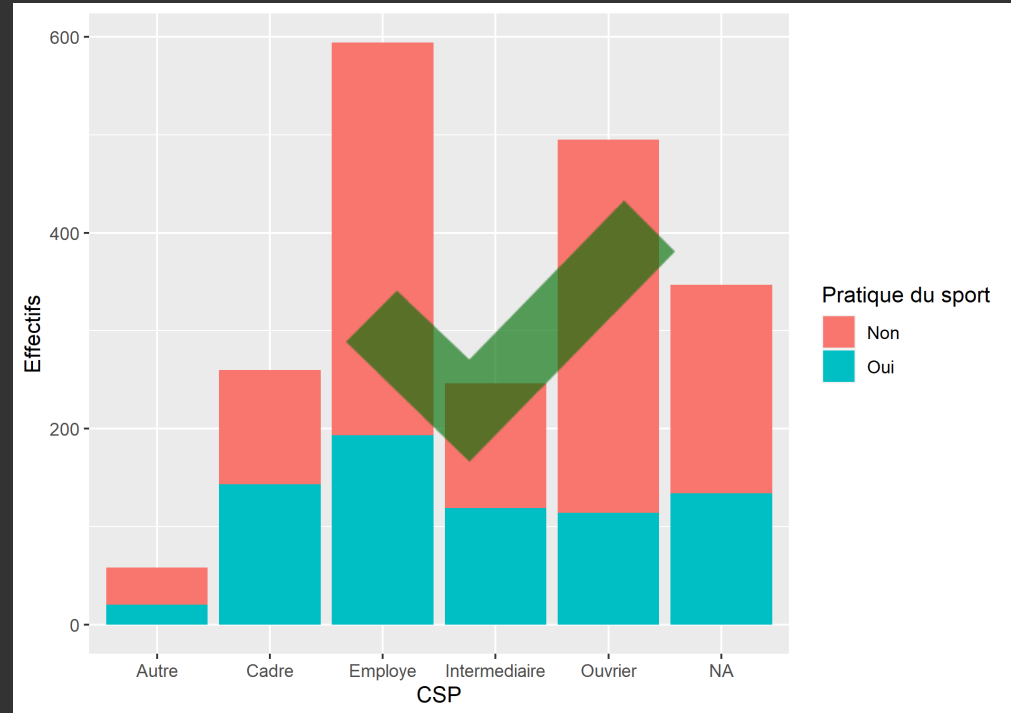
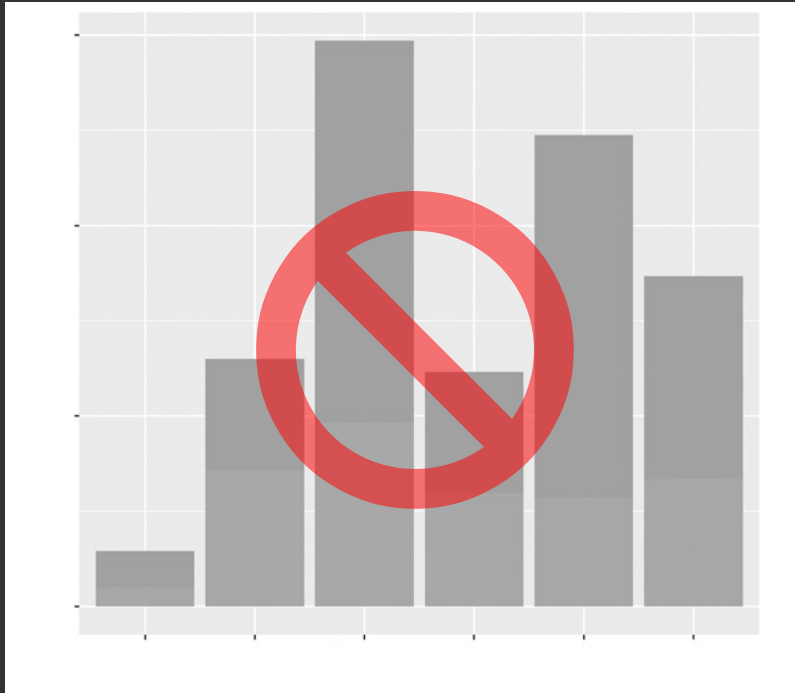
- Plus d'exemples et de cas d'utilisation :
  - Quand utiliser quel graphique :
    - <https://www.data-to-viz.com/> - anglais
    - <http://www.infographicsblog.com/chart-suggestions-a-thought-starter-andrew-abela/>  
- anglais / Source du graphique de la slide précédente
    - <https://gramener.github.io/visual-vocabulary-vega/#>
  - Exemple de graphiques : <https://python-graph-gallery.com/> - anglais

# Données

- **Indispensables à la création d'un graphique**
- Le format changera dépendamment du type de graphique
- Attention aux corrélations de données
  - Exemples de corrélations absurdes :  
<http://www.tylervigen.com/spurious-correlations>

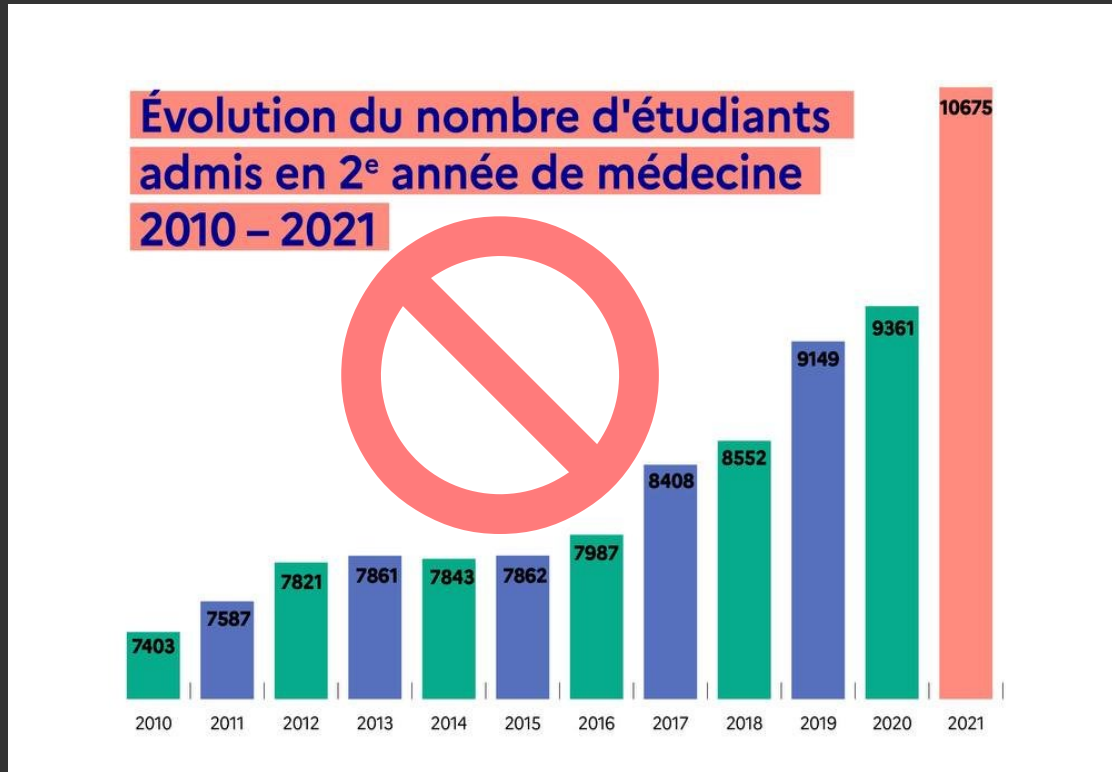
# Graphiques – Avec de la clarté c'est mieux

- Il est important de mettre des légendes et couleurs



# Graphiques – Avec de la clarté c'est mieux

- Et de penser aux échelles



Le graphique ci-contre donne l'impression qu'il y a eu un bon énorme entre 2020 et 2021. En fait l'augmentation n'est que de 1 314 étudiants.

Source(s) :

- <https://twitter.com/vidalfrederique/status/1390735956014256131>

# matplotlib

- Intégrée par défaut dans Anaconda / Google Colab → pas besoin de pip
- Permet de réaliser des graphiques
- Design pas forcément très attrayant
  - Utilisation de seaborn pour le design et plus de choix de graphiques

# seaborn

- Rajoute de nouveaux graphiques à matplotlib
  - Ainsi que de nouvelles options
- Basé sur matplotlib
  - Nécessite matplotlib

Source(s) :

- <https://seaborn.pydata.org/>

# Pratiquons ! - Graphiques

Pré-requis :

- Avoir la ressource `ressources/graphiques.jpg`
- A télécharger ici :  
<https://downgit.github.io/#/home?url=https://github.com/DanYellow/cours/tree/main/big-data-s4/travaux-pratiques/numero-7>



**Questions ?**

