# Introduction à la Science des données
# Visualisation des données

V. Monbet

[1] Université de Rennes/UFR Mathématiques

# Outline

Introduction

Différents types de données

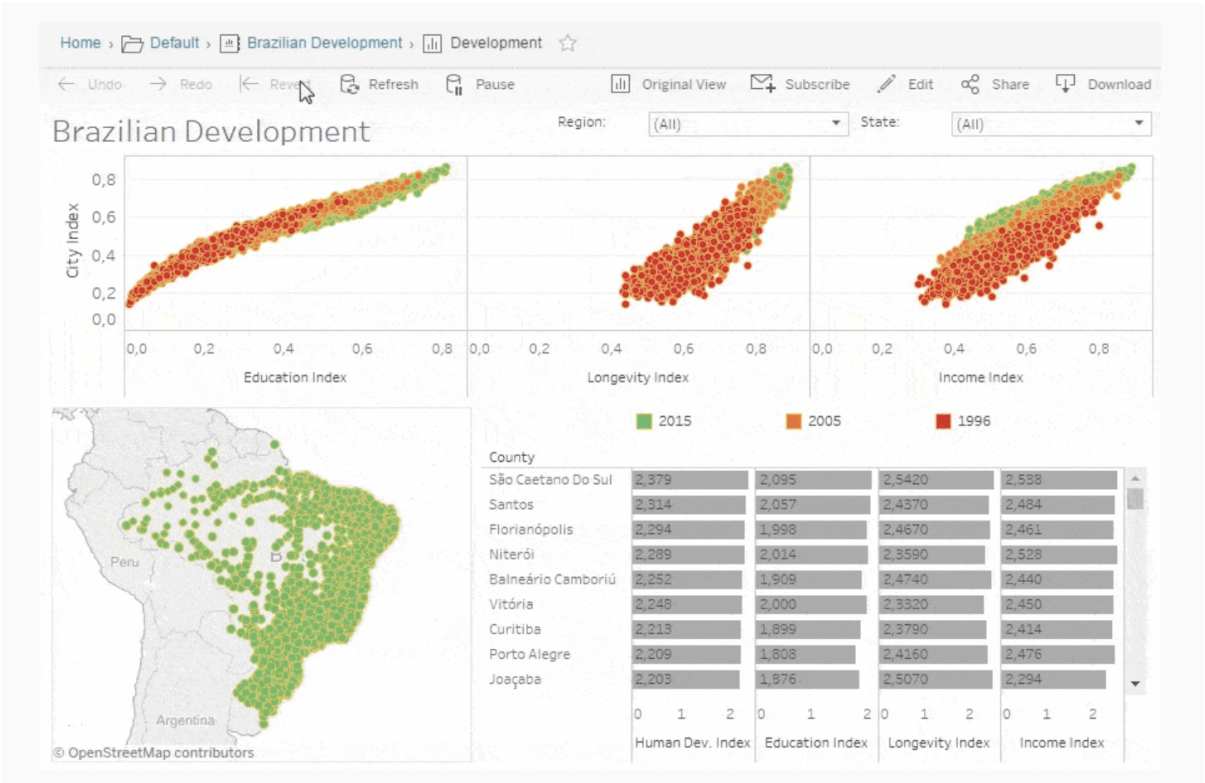
Différents graphiques

Données multivariées

# Qu’est ce que la visualisation des données ?

Dans notre monde de plus en plus axé sur les données, il est plus important que jamais de disposer de moyens accessibles pour **visualiser et comprendre les données**.

La visualisation des données est la représentation graphique des informations et des données. En utilisant des éléments visuels tels que **des diagrammes, des graphiques, des cartes et des réseaux**, les outils de visualisation des données offrent un moyen accessible de voir et de comprendre les tendances, les valeurs aberrantes et les schémas dans les données.



La *Data viz* aide à comprendre les informations beaucoup plus rapidement et à reconnaître des schémas qui seraient autrement difficiles à voir avec des données textuelles.

Dans le monde du Big Data, les outils et technologies de visualisation des données sont essentiels pour analyser des quantités massives d'informations et prendre des décisions fondées sur les données.




Prenons 4 minutes pour regarder "The famous GapminderVideo, Hans Rosling : 200 Countries, 200 Years, 4 Minutes"

https://www.youtube.com/watch?v=jbkSRLYSojo

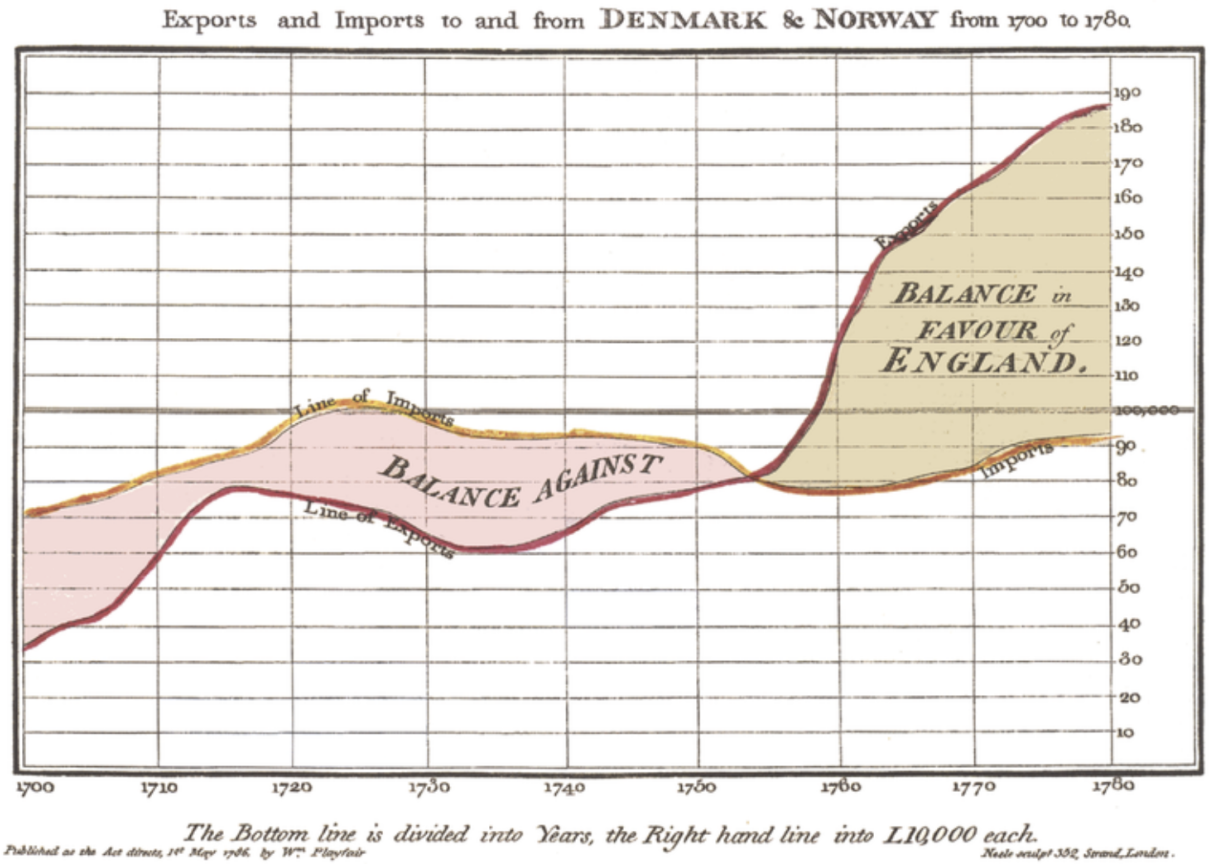
Une visualisation de données efficace doit être informative, efficiente, attrayante et, dans certains cas, interactive et prédictive. Pittenturf explique les critères de base auxquels une visualisation de données doit répondre pour être efficace (Pittenturf 2018) :

| Critères | Description |
|---|---|
| Informative | La visualisation doit être capable de transmettre au lecteur les informations souhaitées à partir des données. |
| Efficace | La visualisation ne doit pas être ambiguë. |
| Attrayante | La visualisation doit être captivante et visuellement agréable. |
| Interactive et prédictive (facultatif) | Les visualisations peuvent contenir des variables et des filtres avec lesquels les utilisateurs peuvent interagir pour prédire les résultats de différents scénarios. |

# DataViz, un outil moderne ?

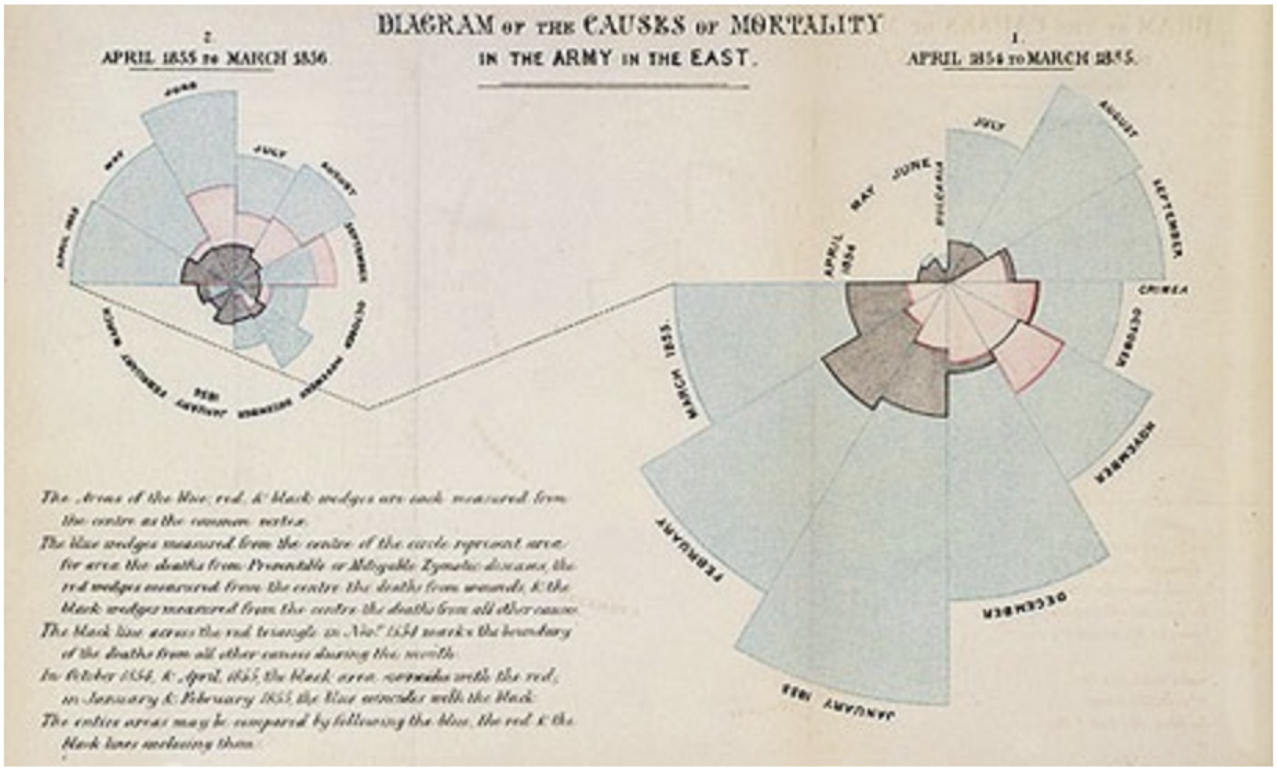
Compte tenu de l'explosion récente de la disponibilité des données et des outils de visualisation, il serait naturel de supposer que les graphiques statistiques et les visualisations de données sont des développements relativement modernes. Cependant, la visualisation des données n'est pas un produit moderne : elle s'est développée au fil du temps pour intégrer les outils que nous utilisons aujourd'hui et les tendances que nous prévoyons.

Par exemple, au 18eme siècle

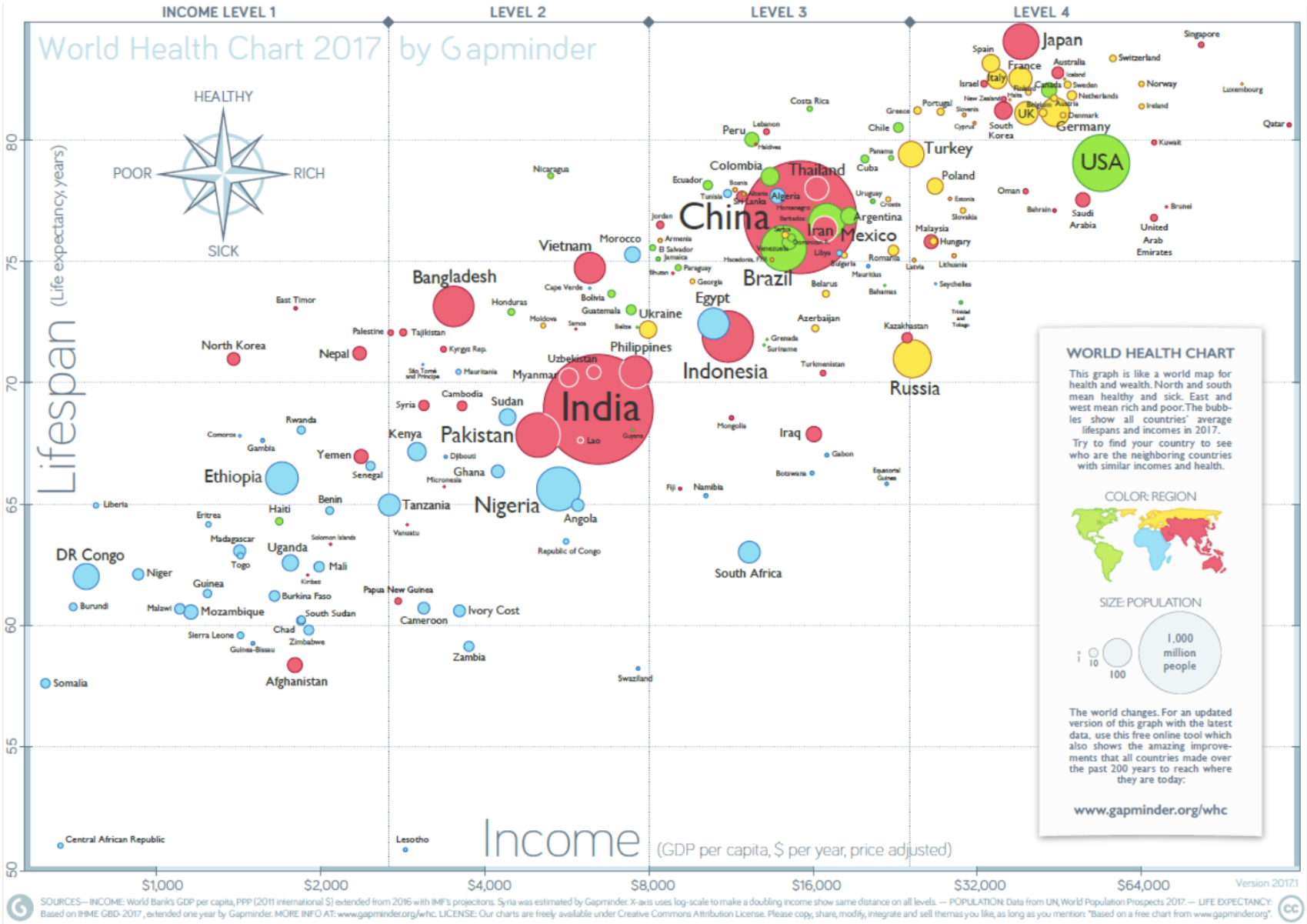


# Florence Nightingale (1820–1910)

Florence Nightingale est célèbre pour son travail d'infirmière pendant la guerre de Crimée, mais elle était aussi une journaliste de données. Elle s'est rendu compte que les soldats mouraient à cause du manque d'hygiène et de la malnutrition. Elle a donc tenu des registres méticuleux des décès dans les hôpitaux et a visualisé les données. Ses diagrammes de coxcomb ou de rose l'ont aidée à lutter pour l'amélioration des conditions hospitalières et, en fin de compte, à sauver des vies.



# Aujourd'hui : graphiques interactifs



# Outline

Introduction

Différents types de données

Différents graphiques

Données multivariées

# Différents types de données

Il existe différents types de données qui seront illustrées par différents types de graphiques.

- **données structurées** : sondage, fichier de notes, liste de préférence, etc
- séries temporelles : évolution de la température moyenne sur terre, série financière, nombre de malades atteints de la grippe par semaine, etc
- données géographiques : nombre d'espèces de plantes relevées en différents sites, indice de pauvreté dans le monde, position d'un téléphone portable, etc
- **données non structurées** : langage naturel, chants d"oiseaux, imgaes, vidéos

# Données structurées

Les **données structurées**, généralement organisées en lignes et en colonnes, qui incluent des nombres et des mots, tels que des noms, des dates et des informations de carte de crédit...

Santé publique
Le fichier ci-dessous reporte des statistiques de santé publique pour les pays d'Afrique (année ?)

| Pays | Médecins (1000 hab) | PIB Hab | Pop ($10^6$ hab) | % pop urb | % dép. santé pub | % tot dép. santé | IDH [1] | Educ. | Santé ($10^3$ cas) | VIH |
|---|---|---|---|---|---|---|---|---|---|---|
| South Africa | 0,73 | 7392,87 | 50,8 | 62,22 | 3,99 | 8,50 | 0,64 | 0,66 | 0,55 | 6400 |
| Algeria | 1,21 | 4473,49 | 36,0 | 67,53 | 3,59 | 5,12 | 0,73 | 0,49 | 0,77 | 7 |
| Angola | 0,14 | 3886,48 | 21,2 | 40,10 | 2,06 | 3,39 | 0,51 | 0,30 | 0,39 | 260 |
| Benin | 0,06 | 732,95 | 9,5 | 41,85 | 2,69 | 4,95 | 0,47 | 0,27 | 0,54 | 64 |
| . | | | . | | | . | | | . | |
| . | | | . | | | . | | | . | |
| . | | | . | | | . | | | . | |

1. IDH : indice de développement humain

# Données structurées

Les **données structurées**, généralement organisées en lignes et en colonnes, qui incluent des nombres et des mots, tels que des noms, des dates et des informations de carte de crédit...

Données bancaires : informations sur des prêts (République Tcheque, 1999)

| loan id | account id | date | amount | duration | payments | status |
|---|---|---|---|---|---|---|
| 5314 | 1787 | 930705 | 96396 | 12 | 8033.00 | B |
| 5316 | 1801 | 930711 | 165960 | 36 | 4610.00 | A |
| 6863 | 9188 | 930728 | 127080 | 60 | 2118.00 | A |
| 5325 | 1843 | 930803 | 105804 | 36 | 2939.00 | A |
| 7240 | 11013 | 930906 | 274740 | 60 | 4579.00 | A |
| ⋮ | | | | | ⋮ | |

Données médicales
Données en écologie
Fichier de carte fidélité dans le commerce

# Données non structurées

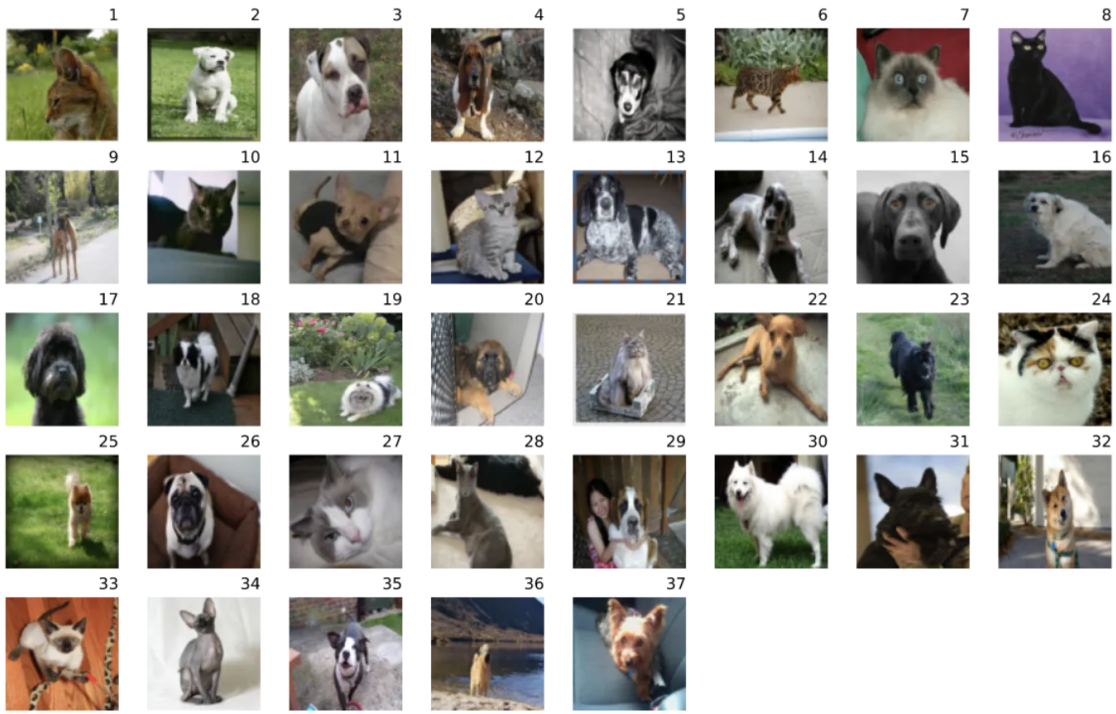
Les données non structurées, qui ne sont pas organisées et qui incluent le texte des fichiers et documents, les données mobiles et issues des réseaux sociaux, le contenu des sites web et les vidéos.

**détection d'attaques sur un réseau**
Avec l'évolution des réseaux et des technologies, les cyberattaques deviennent de plus en plus sophistiquées. La science des données peut aider à les détecter efficacement.
Ces données sont par nature non structurées, mais on peut les structurer en relevant certaines caractéristiques des connections : durée de la connection, heure de la connexion, temps entre deux connections, etc.

# Données non structurées

Les données non structurées, qui ne sont pas organisées et qui incluent le texte des fichiers et documents, les données mobiles et issues des réseaux sociaux, le contenu des sites web et les vidéos.

**Identification de la race d'un chat ou d'un chien à partir d'une image**
Les images sont toutes différentes par divers aspects : fond, position de l'animal, intensité de lumière, etc.
On ne peut pas les comparer ou les agréger directement.



# Données non structurées

**Détecter le niveau d'anxiété ou de stress à partir de tweets**
L'information est contenue dans le texte du tweet. Bien que le texte s'appuie sur une structure grammaticale, on ne peut pas considérer qu'il est structuré (au sens de la science des données).

| | subreddit | post_id | sentence_range | text | id | label | confidence | social_timestamp | social_karma |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ptsd | 8601tu | (15, 20) | He said he had not felt that way before, sugge... | 33181 | 1 | 0.8 | 1521614353 | 5 |
| 1 | assistance | 8lbrx9 | (0, 5) | Hey there r/assistance, Not sure if this is th... | 2606 | 0 | 1.0 | 1527009817 | 4 |
| 2 | ptsd | 9ch1zh | (15, 20) | My mom then hit me with the newspaper and it s... | 38816 | 1 | 0.8 | 1535935605 | 2 |
| 3 | relationships | 7rorpp | [5, 10] | until i met my new boyfriend, he is amazing, h... | 239 | 1 | 0.6 | 1516429555 | 0 |

# Données non structurées

**Identification de chants d'oiseaux**
La présence d'oiseaux, en nombre et en variété d'espèces, est un marqueur de biodiversité de de bonne santé de l'environnement. Il est souvent difficile d'observer tous les oiseaux d'un site donné. Il est plus facile de les enregistrer.

Pourquoi dit-on que les données sont non structurées ?

Bruant Maritime



Chevalier semipalmé



Sterne de Forster



# Outline

Introduction

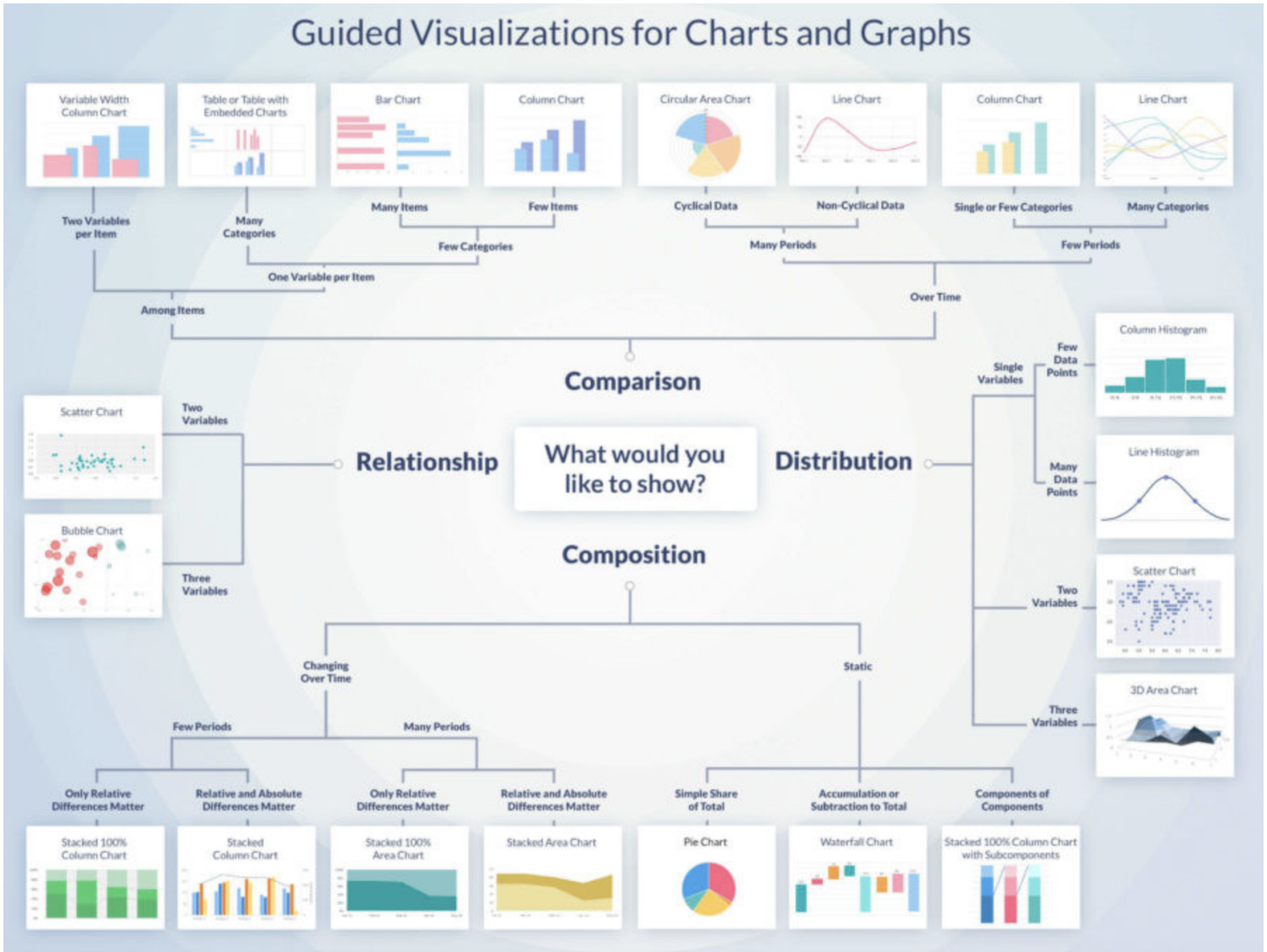
Différents types de données

Différents graphiques

Données multivariées

# Différents graphiques

Selon le type de données que l'on doit visualiser et selon le message qu'on veut faire passer, on va utiliser différents types de graphiques.




source

# Comment choisir le bon graphique ?

La visualisation des données, comme tout autre outil, peuvent être très utiles lorsqu'ils sont bien utilisés. Pour tirer le meilleur parti de vos données, vous devez les associer au bon type de graphique. Comment y parvenir ? En vous posant les questions suivantes :

1. Qui est mon public ?
2. Quelles sont les connaissances que je souhaite apporter à mes lecteurs ?
3. Quelle doit être la portée de mon axe ?
4. Dois-je afficher les valeurs dans le temps ou parmi les groupes ?
5. De quelles informations ai-je besoin concernant le nombre de catégories ?
6. De combien de points de données ai-je besoin pour chaque catégorie ?

# Comparer des données

Diagramme en batons horizontal
est un bon graphique pour afficher et comparer le rang des valeurs et se concentrer sur les extrêmes.

On préfère

- des noms courts pour les catégories
- faible nombre de catégories (idéalement moins de 7)

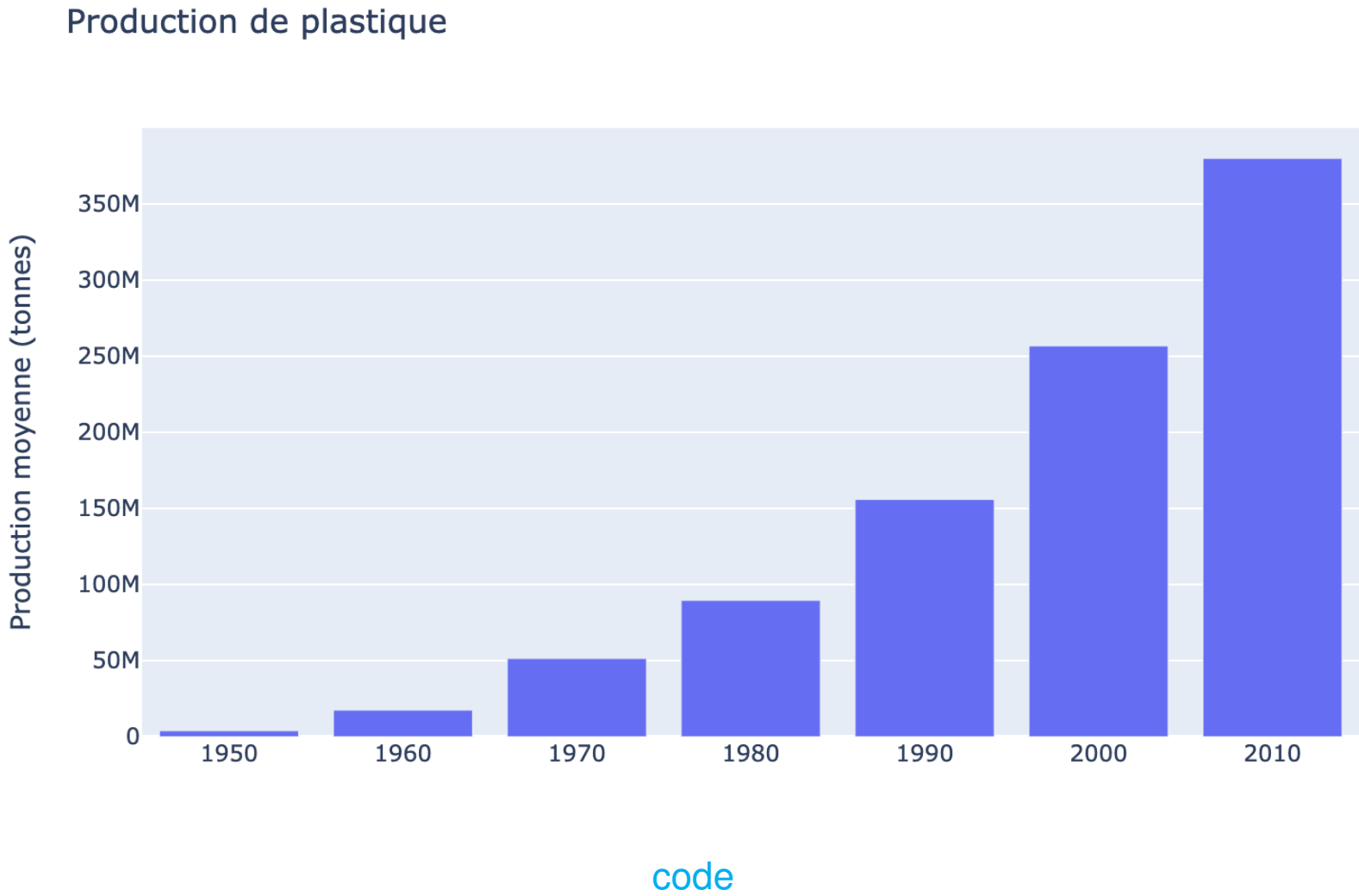
Exemples : Chiffre d'affaires, chiffre d'affaires par année, etc.

# Exemple

On considère la production mondiale de plastique par tranche de 10 ans.

| | Year | Annual plastic production between 1950 and 2019 | decade |
|---|---|---|---|
| **decade** | | | |
| **1950.0** | 1954.500000 | 3.900000e+06 | 1950.0 |
| **1960.0** | 1964.500000 | 1.750000e+07 | 1960.0 |
| **1970.0** | 1974.555556 | 5.133333e+07 | 1970.0 |
| **1980.0** | 1984.500000 | 8.950000e+07 | 1980.0 |
| **1990.0** | 1994.500000 | 1.558000e+08 | 1990.0 |
| **2000.0** | 2004.500000 | 2.566000e+08 | 2000.0 |
| **2010.0** | 2014.500000 | 3.796904e+08 | 2010.0 |

# Diagramme en batons horizontal

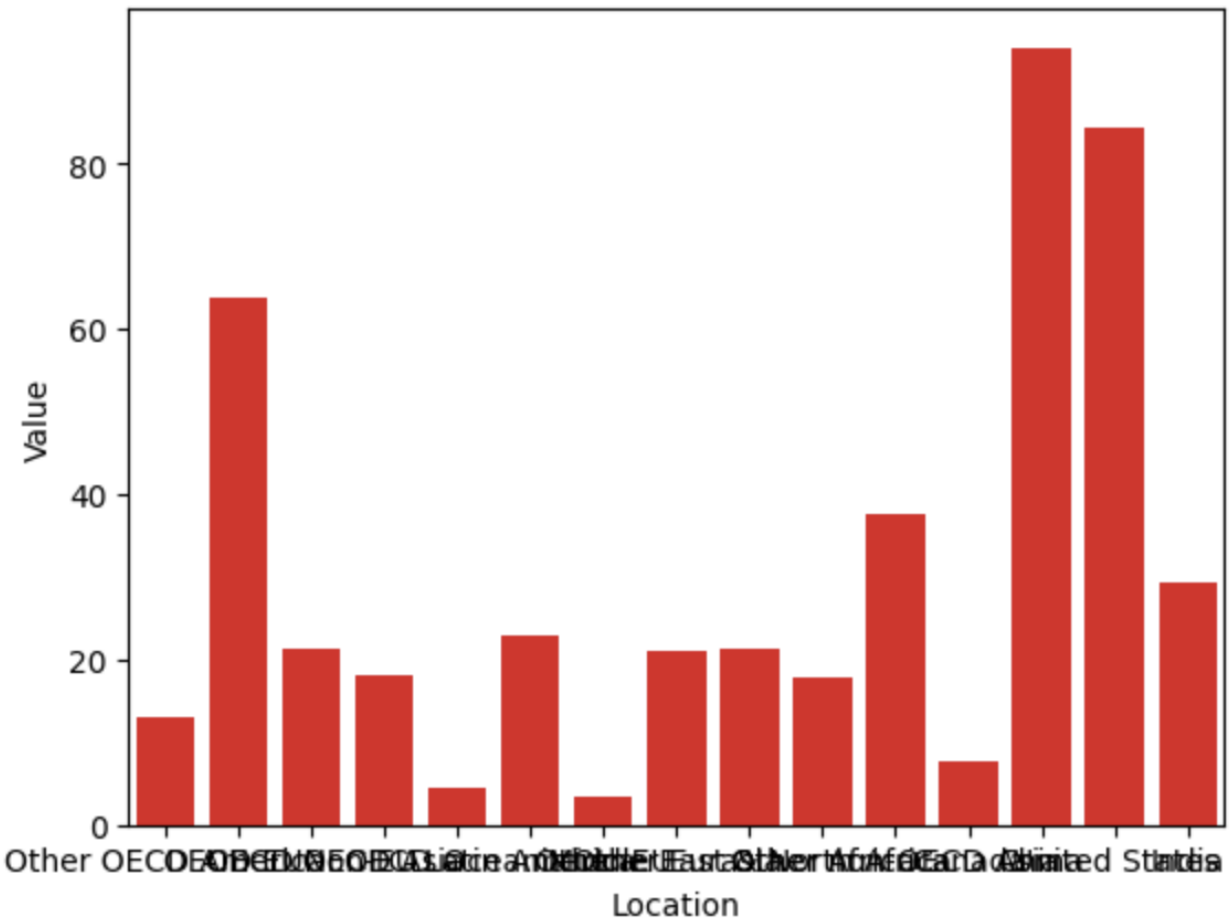


# Utilisation de plastique par région du monde (2019)

Données de l'OCDE

| Location | Value |
|---|---|
| Other OECD America | 13.1231 |
| OECD EU | 63.6734 |
| OECD Non-EU | 21.3707 |
| OECD Asia | 18.2069 |
| OECD Oceania | 4.3848 |
| Latin America | 22.8634 |
| Other EU | 3.3225 |
| Other Eurasia | 20.9538 |
| Middle East & North Africa | 21.1704 |
| Other Africa | 17.8515 |
| Other non-OECD Asia | 37.641 |
| Canada | 7.539 |
| China | 94.0061 |
| United States | 84.3059 |
| India | 29.3334 |

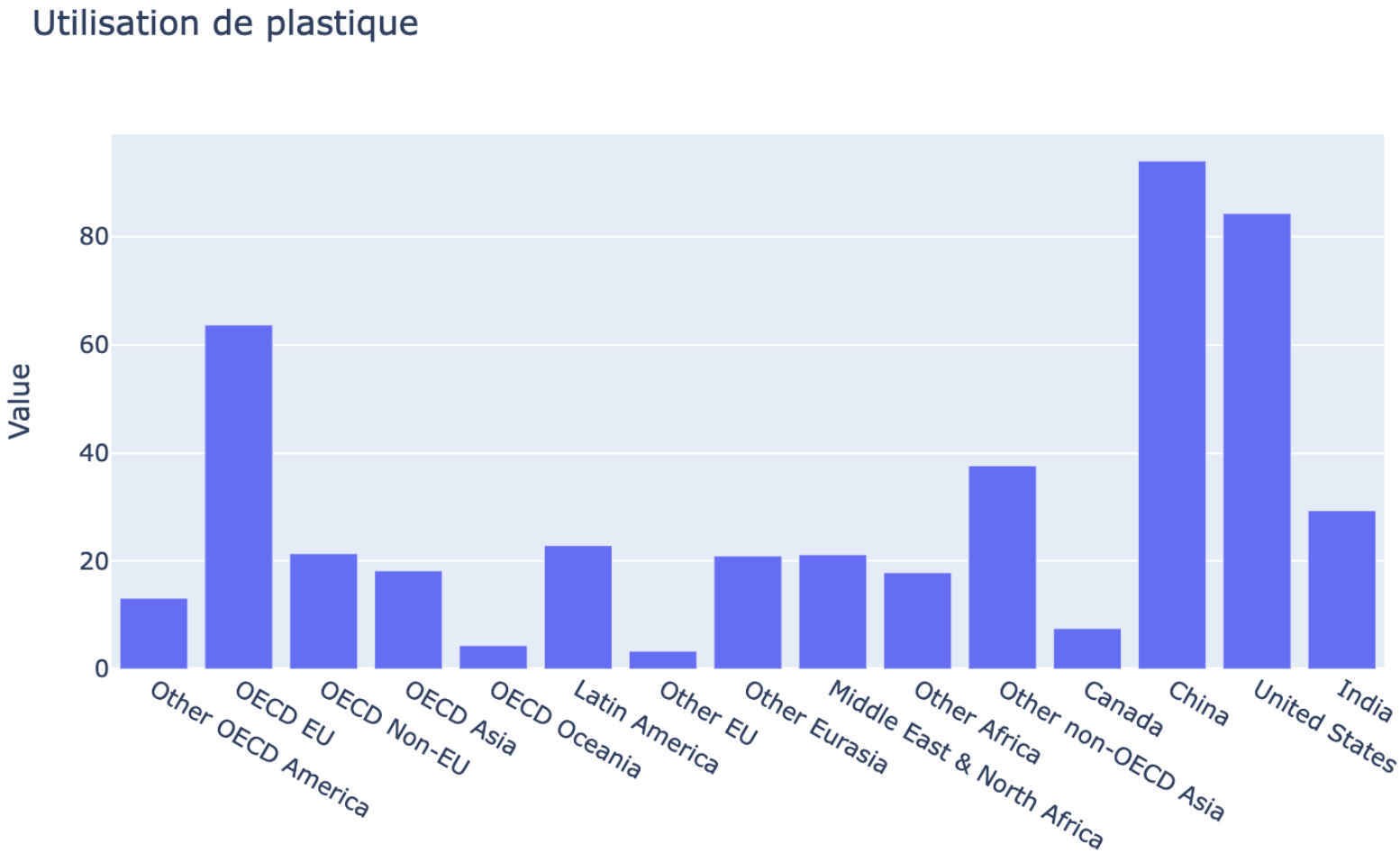
# Un premier diagramme, avec seaborn




code

Quels sont les 3 ou 4 problèmes de cette figure ?

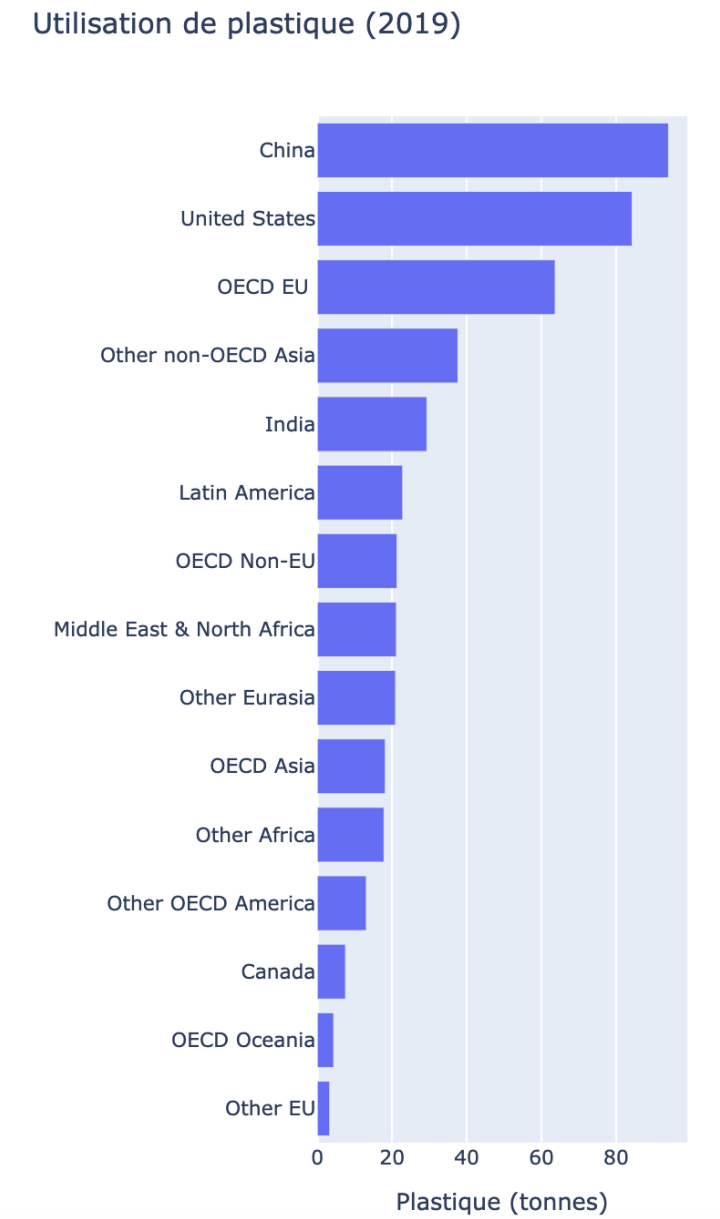
# Un premier diagramme, avec plotly




code

Que peut-on encore améliorer pour faciliter la lecture et compléter les informations disponibles ?

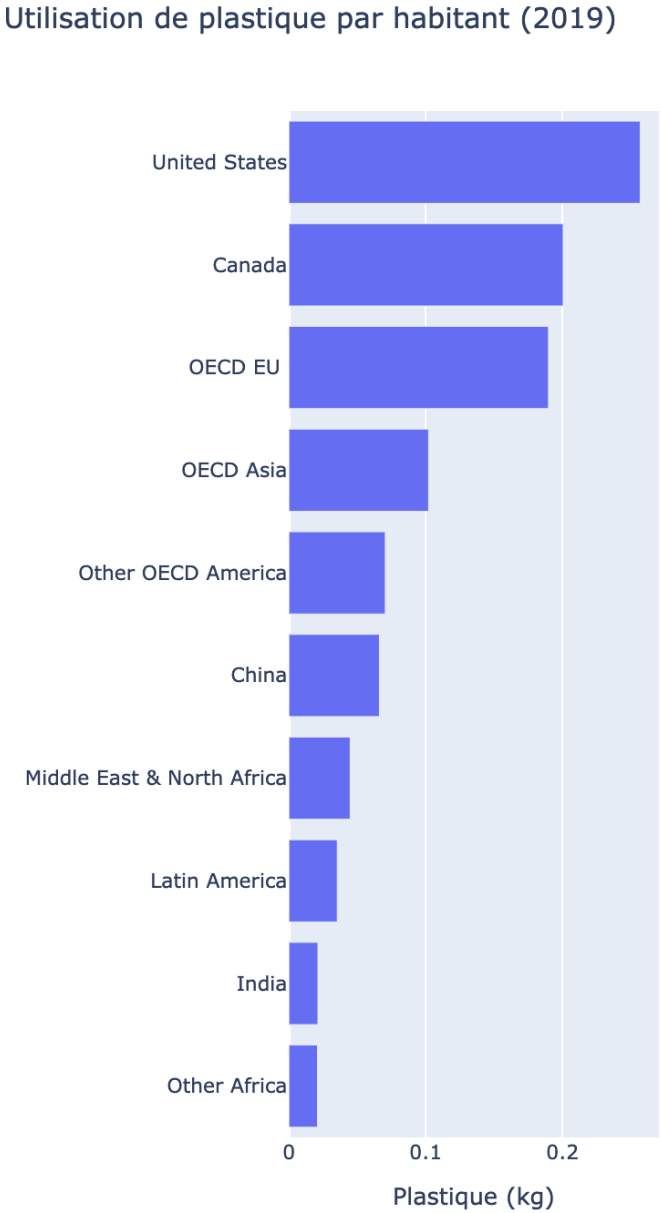
# Un second diagramme, avec plotly




- Le diagramme en barres verticale permet une lecture facile des titres de catégories (ici régions du monde).
- Ordonner les valeurs permet d'identifier très vite les gros utilisateurs de plastique et les petits utilisateurs.
- On voit aussi plus facilement que la Chine et les Etats-Unis utilisent presque la même quantité de plastique.
- Les légendes d'axe et leur unité de mesure sont importantes !
- Il manque encore un titre indiquant de à quoi on s'intéresse et l'année des données.

code

# Un second diagramme par habitant, avec plotly




- Il faut bien sûr aussi faire attention aux chiffres qu'on interprête et ce qu'on leur fait dire...

code

# Comparer des données

Le diagramme en batons vertical est utilisé si

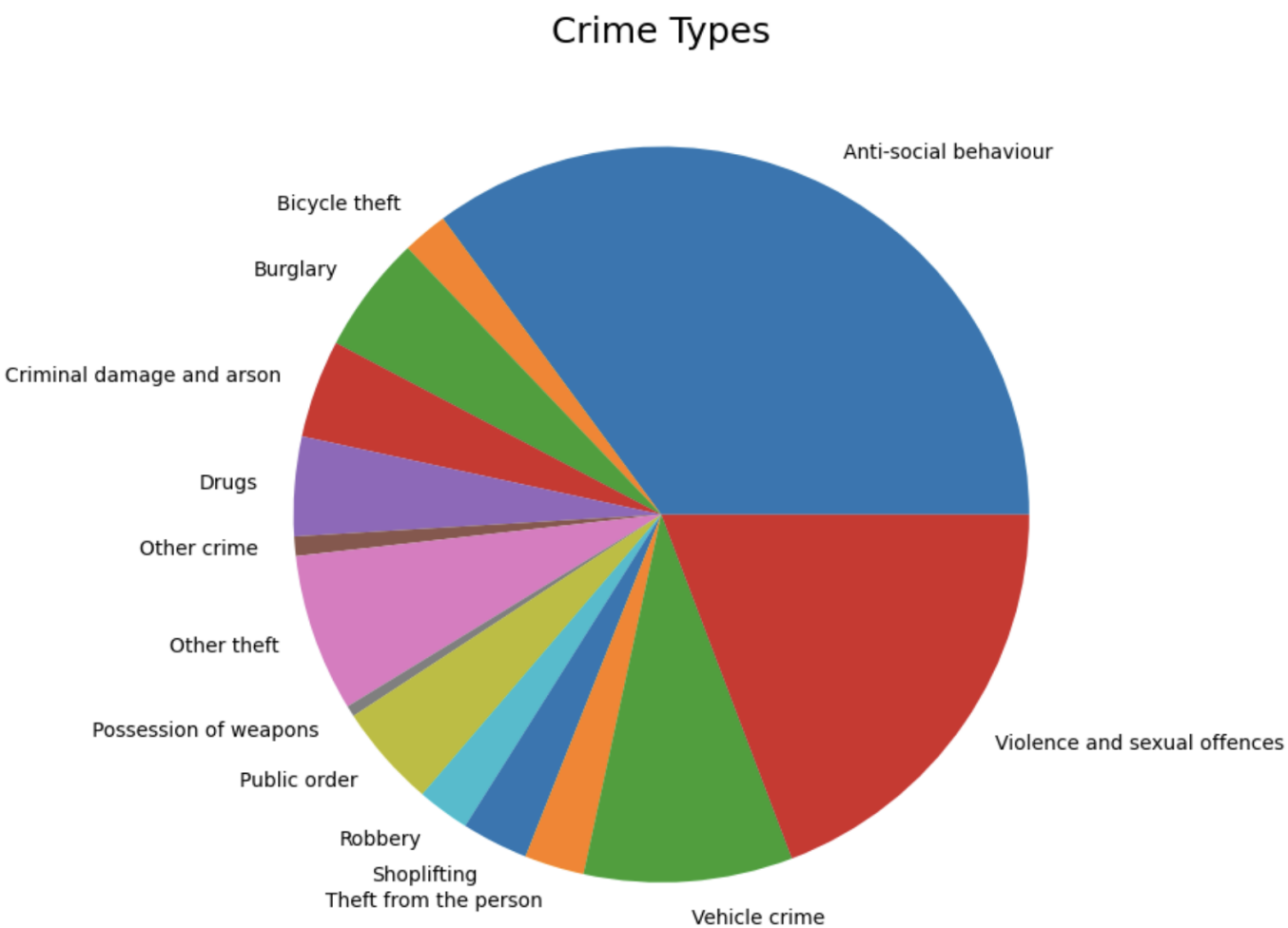
- on veut savoir si les valeurs des données ont atteint un objectif particulier
- les éléments du graphique comportent plus de sept mais moins de 15 catégories,
- on a besoin d'afficher des nombres négatifs
- les étiquettes des catégories de données sont longues.

Exemple : Visiteurs du site web par pays, Clients gagnés par rôle, etc.

# Visualiser de compositions

Le graphique usuel pour visualiser est le *"pie-chart"* (ie le camembert).

On considère des données de criminalité à Londres en 2020.

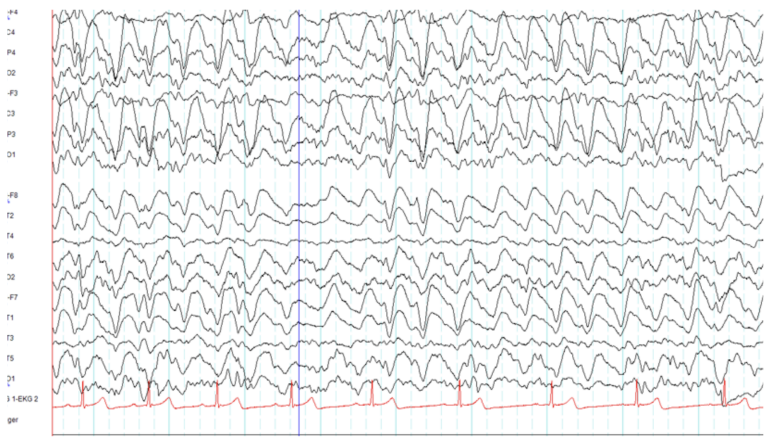


# Suivre une évolution (séries temporelles)

Exemples de séries temporelles

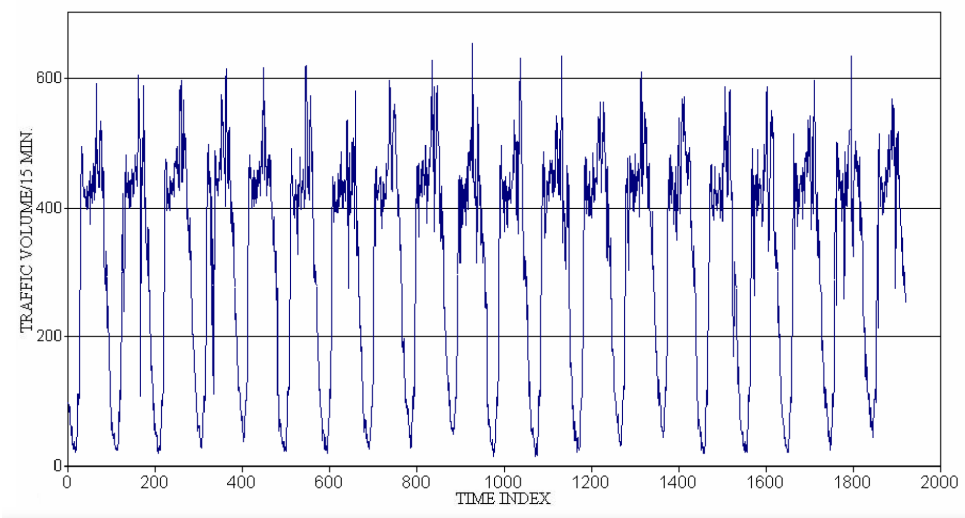
- Activité électrique dans le cerveau, Battements de cœur par minute.
- Mesures de précipitations, température, direction de vent
- Cours des actions, des crypto-monnaies
- Nombre de taches solaires, nombre d'animaux dans un environnement
- Ventes au détail annuelles, abonnés mensuels.
- Trafic ferrovière, routier

*Activité électrique dans le cerveau*



This graph of electrical activity in the brain is known as an EEG. Here, the patterns of spikes show that the person was asleep and likely dreaming.

*Trafic routier, Dublin*



# Données météorologiques

Données SYNOP : 13 ans de mesures météorologiques au pas de temps de 3h.
Ici, deux sites : Brest et Rennes.

| | ID OMM station | Date | Pression au niveau mer | Variation de pression en 3 heures | Type de tendance barométrique | Direction du vent moyen 10 mn | Vitesse du vent moyen 10 mn | Température | Point de rosée | Humidité | ... | Altitude | communes (name) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1584** | 7110 | 2017-12-03T10:00:00+01:00 | 103060.0 | 110.0 | 3.0 | 280.0 | 1.7 | 280.95 | 280.35 | 96.0 | ... | 94 | Guipavas |
| **1585** | 7110 | 2010-04-24T02:00:00+02:00 | 101570.0 | 10.0 | 0.0 | 100.0 | 1.5 | 278.05 | 276.35 | 89.0 | ... | 94 | Guipavas |
| **1586** | 7110 | 2017-12-04T10:00:00+01:00 | 103690.0 | 160.0 | 2.0 | 280.0 | 1.6 | 282.05 | 281.55 | 97.0 | ... | 94 | Guipavas |
| **1587** | 7110 | 2017-12-10T10:00:00+01:00 | NaN | -30.0 | 8.0 | 270.0 | 8.7 | NaN | NaN | NaN | ... | 94 | Guipavas |
| **1588** | 7110 | 2010-02-22T07:00:00+01:00 | 97610.0 | 10.0 | 3.0 | 250.0 | 6.2 | 282.95 | 282.65 | 98.0 | ... | 94 | Guipavas |

5 rows × 82 columns

On va s'intéresser à la température pour avec pour objectif de déterminer si le changement climatique est visible à cette échelle.

code

# Données météorologiques, Brest-Guipavas

Relever au moins trois "difficultés" dans ces données qui doivent nous conduire à préparer les données avant de les analyser.

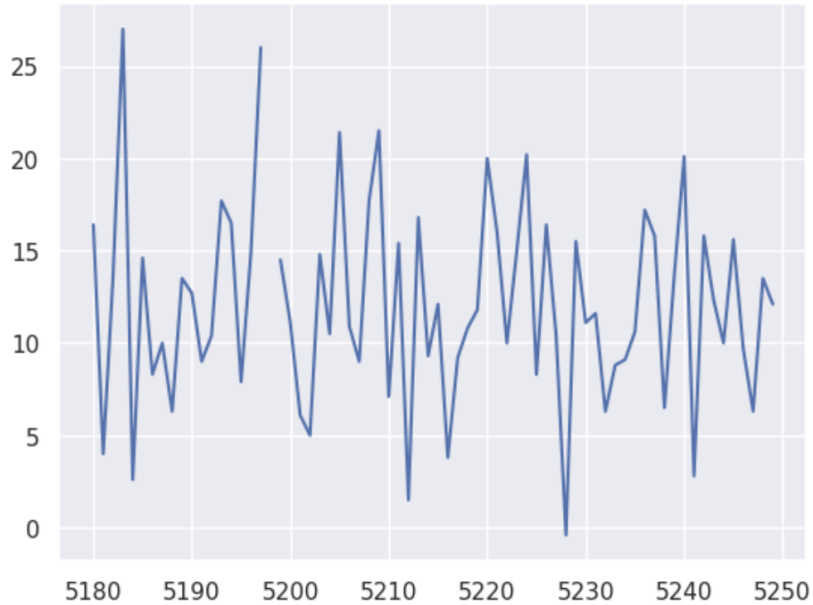
| | ID OMM station | Date | Pression au niveau mer | Variation de pression en 3 heures | Type de tendance barométrique | Direction du vent moyen 10 mn | Vitesse du vent moyen 10 mn | Température | Point de rosée |
|---|---|---|---|---|---|---|---|---|---|
| **1584** | 7110 | 2017-12-03T10:00:00+01:00 | 103060.0 | 110.0 | 3.0 | 280.0 | 1.7 | 280.95 | 280.35 |
| **1585** | 7110 | 2010-04-24T02:00:00+02:00 | 101570.0 | 10.0 | 0.0 | 100.0 | 1.5 | 278.05 | 276.35 |
| **1586** | 7110 | 2017-12-04T10:00:00+01:00 | 103690.0 | 160.0 | 2.0 | 280.0 | 1.6 | 282.05 | 281.55 |
| **1587** | 7110 | 2017-12-10T10:00:00+01:00 | NaN | -30.0 | 8.0 | 270.0 | 8.7 | NaN | NaN |
| **1588** | 7110 | 2010-02-22T07:00:00+01:00 | 97610.0 | 10.0 | 3.0 | 250.0 | 6.2 | 282.95 | 282.65 |

5 rows × 82 columns

# Température à Brest et Rennes

La première étape consiste à préparer les données.

1. Les données ne sont pas ordonnées par date. Le champ "date" est une chaine de caractère. On créé une nouvelle variable date (voir code ).
2. La température est enregistrée en degrés Kelvin. Pour faciliter la lecture on la convertit en degrés Celsius.
3. La température présente des données manquantes (NaN) qui peuvent ou doivent être imputées.

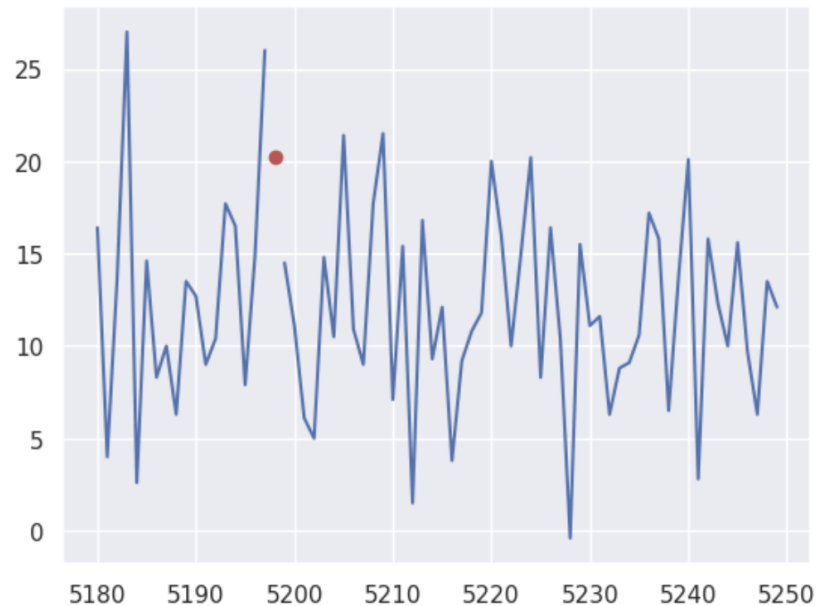



Dans les séries temporelles, si les données manquantes sont isolées, le plus simple est de réaliser une interpolation linéaire.

Les données manquantes (NaN) doivent être imputées pour calculer des statistiques avec pandas [2]. Par exemple, si des points de la séries sont NaN, la moyenne est NaN...

Ici, on choisit d'imputer les données manquantes.

Dans les séries temporelles, si les données manquantes sont isolées, le plus simple est de réaliser une interpolation linéaire.

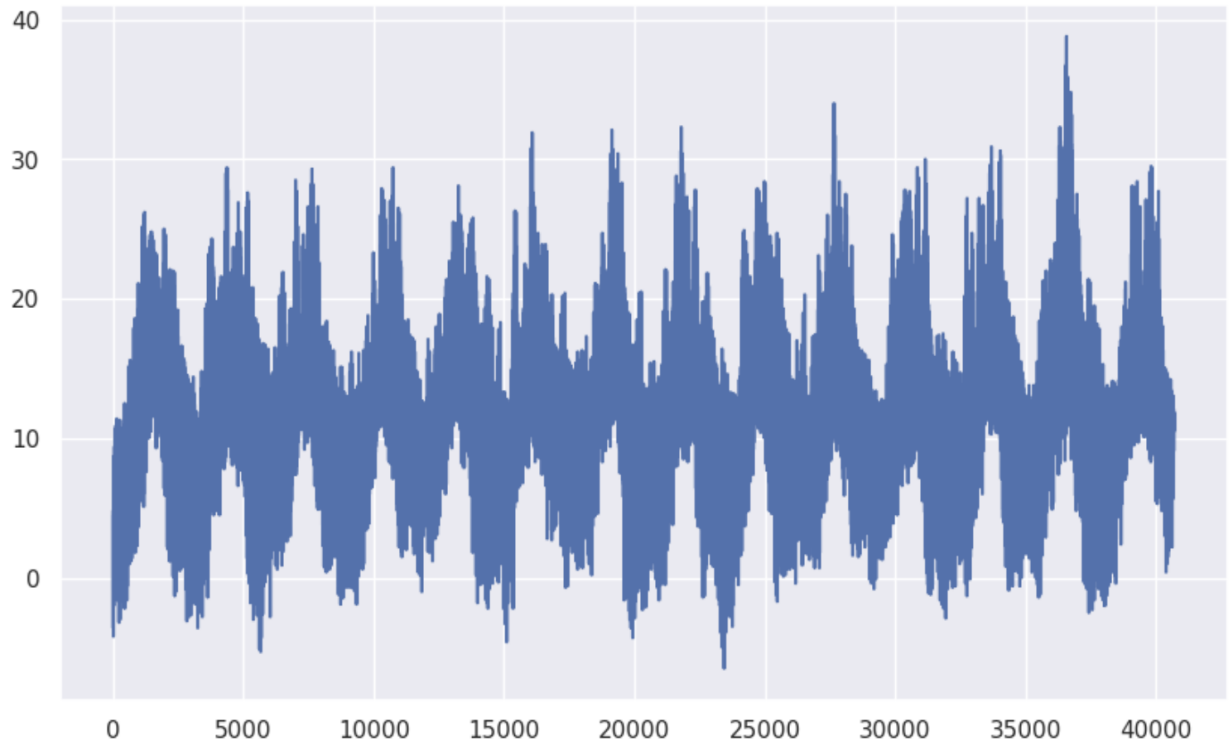



2. il est bien sûr possible de recoder les fonctions pour omettre les données manquantes dans les calculs

# Premier graphique

Dans la figure ci-dessous, relever

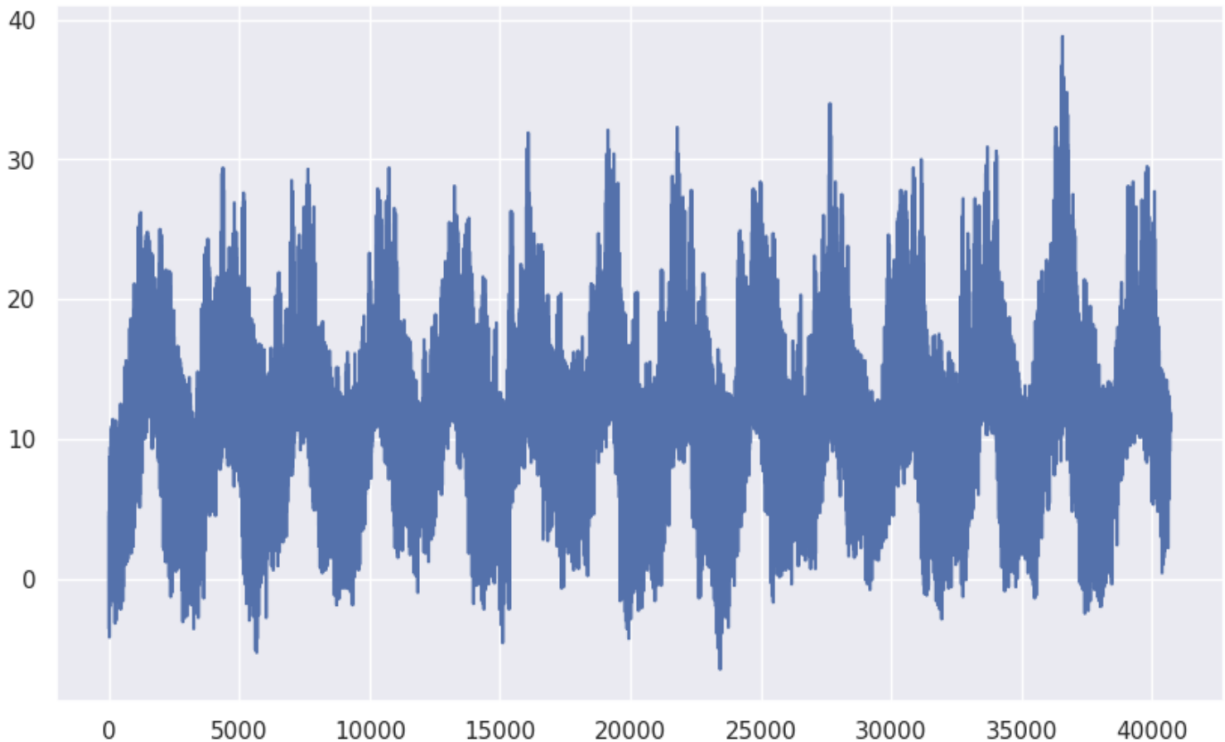
- les points forts  et
  ✓
  ✓
  ✓
- les points faibles
  ✗
  ✗
  ✗



# Premier graphique

Dans la figure ci-dessous,

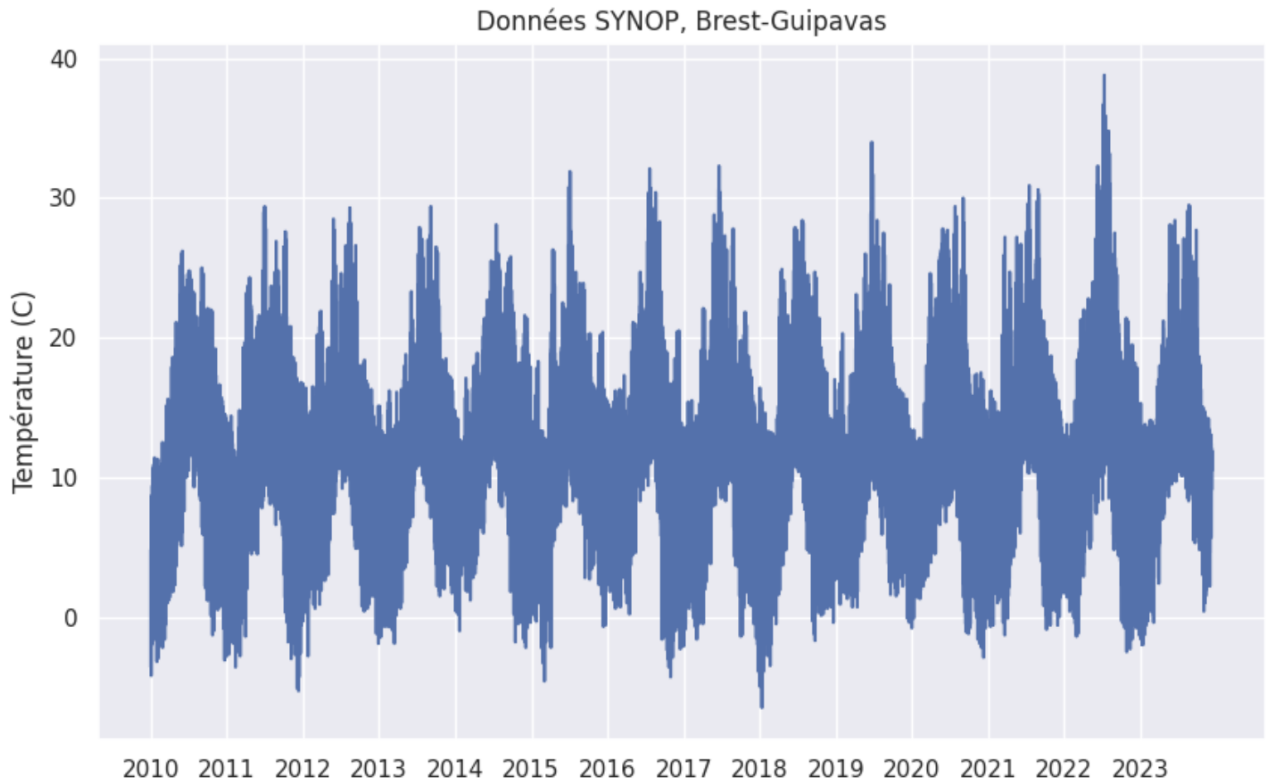
- les points forts  et
  - ✓ On voit tous les points et notamment la saisonnalité
  - ✓ Le fond est joli et aide à lire les valeurs (notamment sur l'axe des ordonnées)
  - ✓
- les points faibles
  - ✗ Absence de titre et de légende d'axes
  - ✗ L'axe des abscisse n'est pas clair : on aimerait voir une date
  - ✗ Si on veut mettre en évidence un changement climatique, il y a trop d'information sur la courbe : la variabilité journalière et annuelle masque la tendance long terme.



# Second graphique avec des titres

- les points forts  et
  - ✓ On voit tous les points et notamment la saisonnalité
  - ✓ Le fond est joli et aide à lire les valeurs (notamment sur l'axe des ordonnées)
  - ✓ Le titre et les légendes d'axes permettent d'interpréter le graphique.
  - ✓ L'unité de la température est indiquée.
- les points faibles

  ✗ Si on veut mettre en évidence un changement climatique, il y a trop d'information sur la courbe : la variabilité journalière et annuelle masque la tendance long terme.



# Pour mettre en évidence la tendance long terme

Pour mettre en évidence une tendance à long terme, on peut calculer différentes types de moyennes.

- **Moyennes annuelles**

  Soit $x_{ij}$ la température du temps $i$ de l'année $j$,

  $$\mu_{an,j} = \frac{1}{n_{an,j}} \sum_{i=1}^{n_{an,j}} x_{ij}$$

- **Moyenne mobile** sur une fenêtre de largeur $n_{window}$

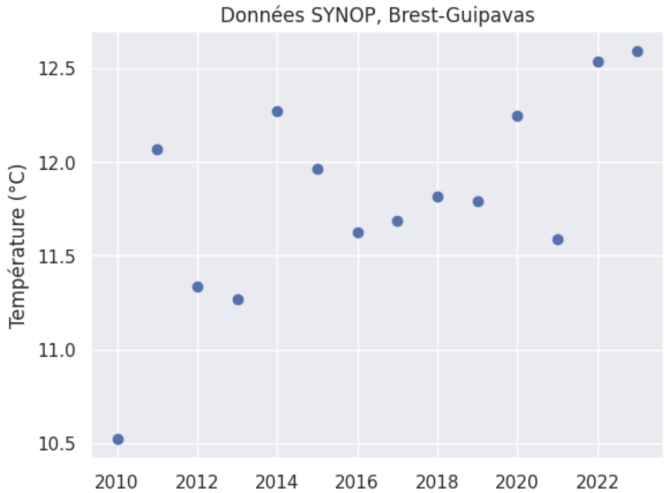
  Soit $x_i$ la température du temps $i$,

  $$\mu_i = \frac{1}{n_{window}} \sum_{i=k-n_{window}/2}^{k+n_{window}/2} x_i$$

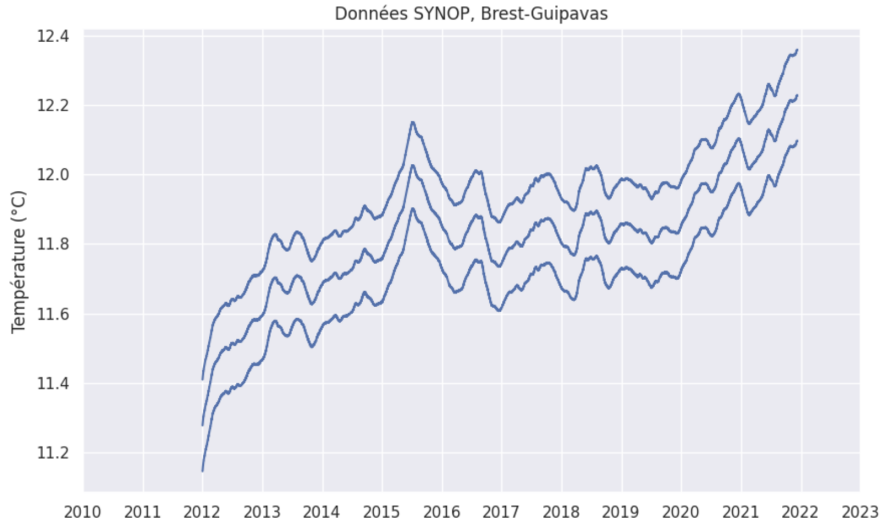
  Il faut choisir la largeur de fenêtre.

# Choisir la "meilleure" figure

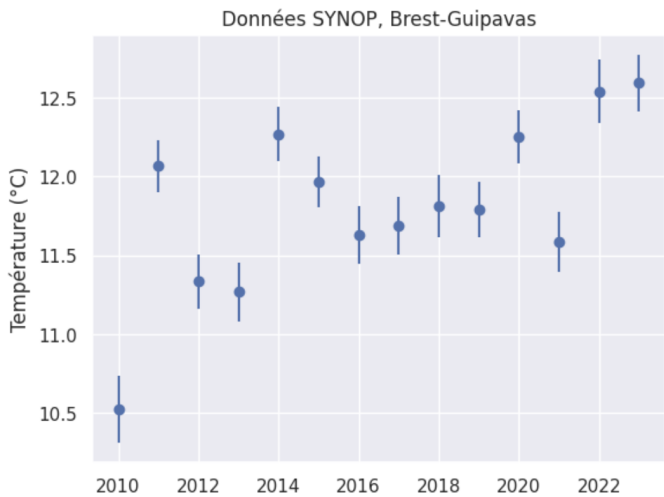
(a)



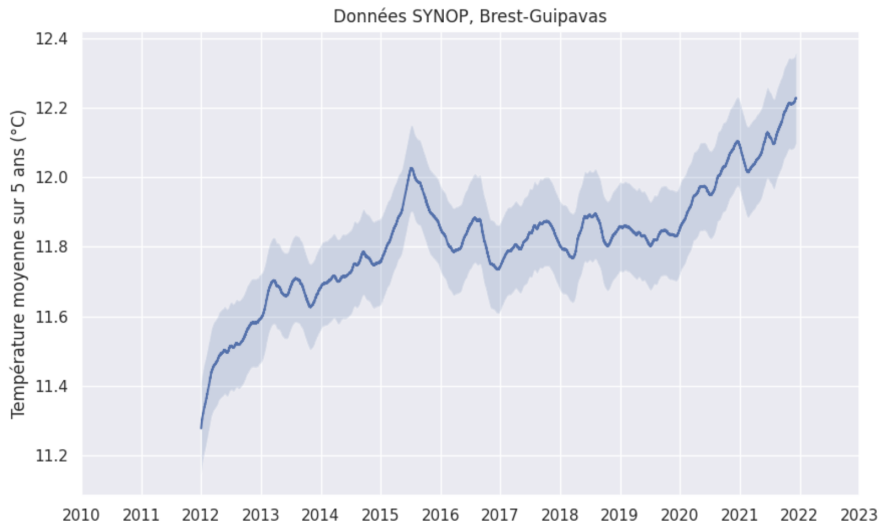

(c)




(b)




(d)



# Choisir la "meilleure" figure

Il faut d'abord définir "meilleure"...

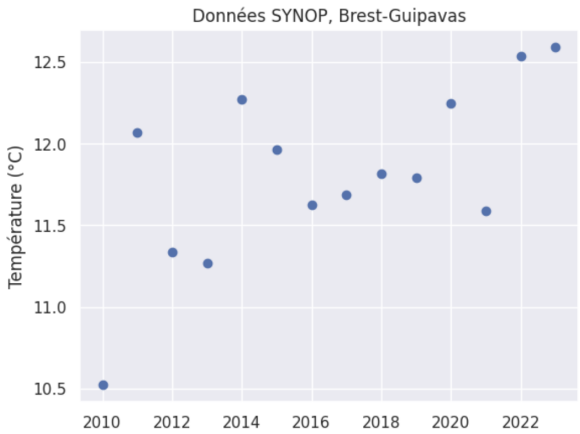
- Facilité de lecture
- Bonne illustration du phénomène d'intérêt (ici le changement climatique)
- qualité des titres et légendes
- prise en compte de la variabilité de l'estimateur

# Choisir la "meilleure" figure

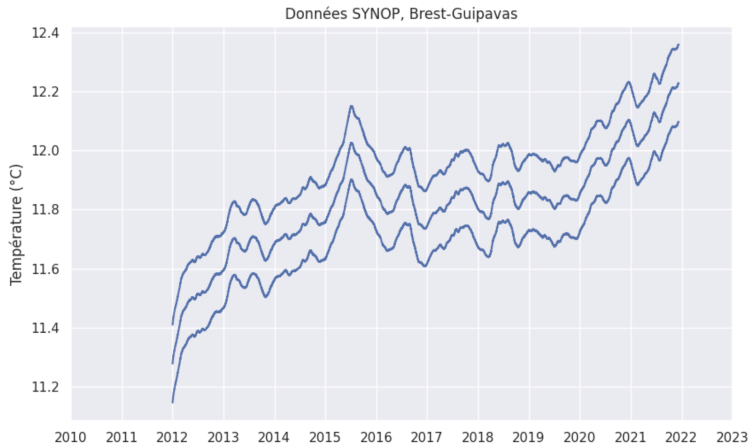
La figure (d) montre une moyenne mobile, avec une fenètre de 5 ans, ainsi que l'intervalle de confiance à 95% associé. L'intervalle de confiance donne une estimation de la variabilité de la moyenne.

La figure (d) montrer une tendance croissante de la température moyenne sur 5 ans.

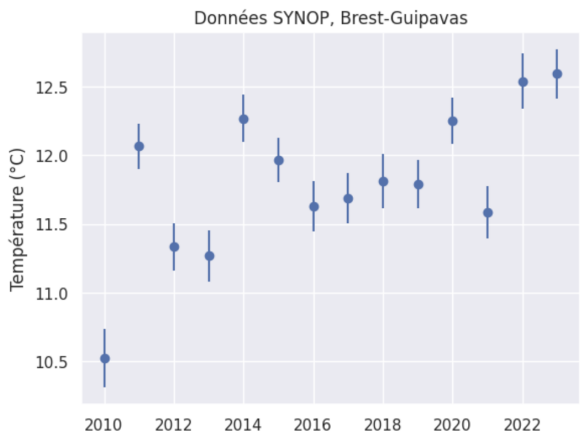
(a)



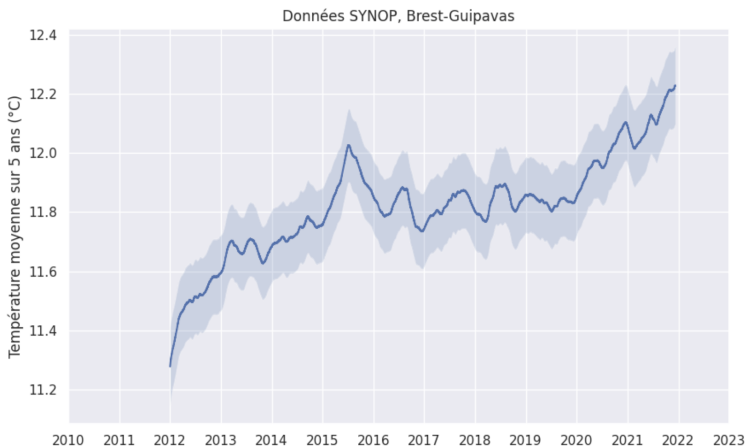

(c)




(b)



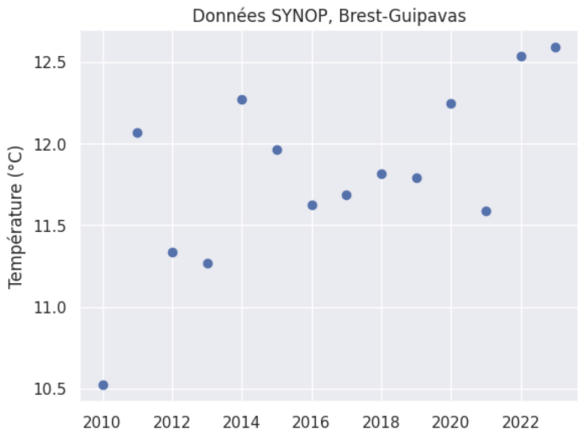

✓ (d)



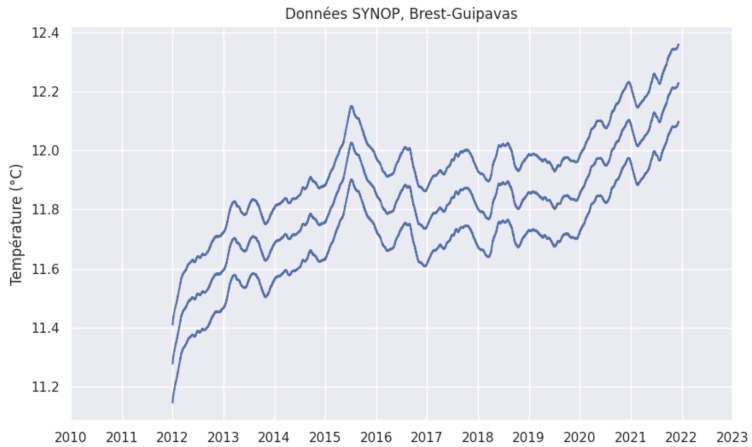
# Choisir la "meilleure" figure

La figure (d) montre une moyenne mobile, avec une fenètre de 5 ans, ainsi que l'intervalle de confiance à 95% associé. L'intervalle de confiance donne une estimation de la variabilité de la moyenne.
La figure (d) montrer une tendance croissante de la température moyenne sur 5 ans.

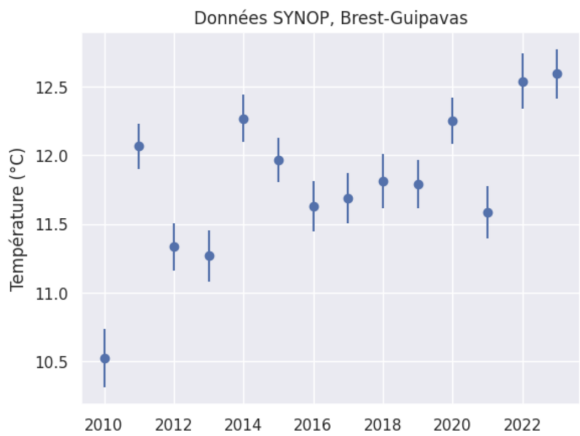
(a)



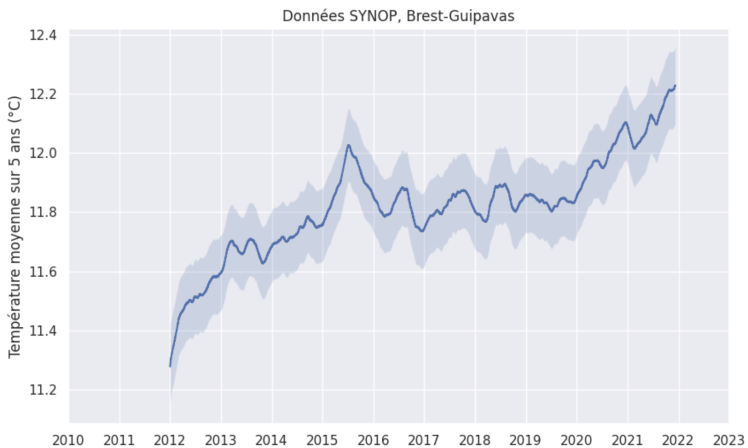

(c)




(b)

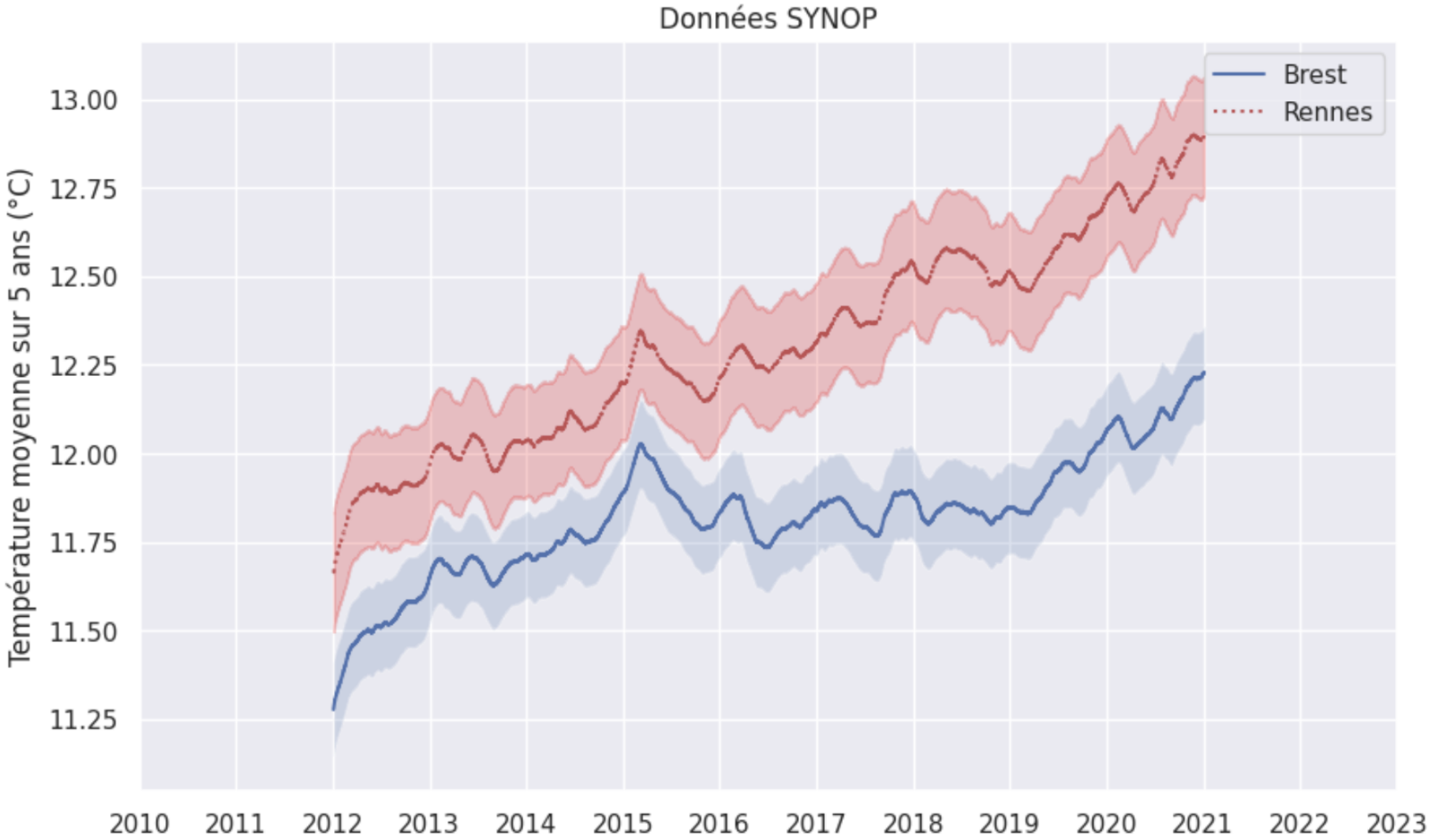



✓ (d)



# Comparaison de Brest et Rennes

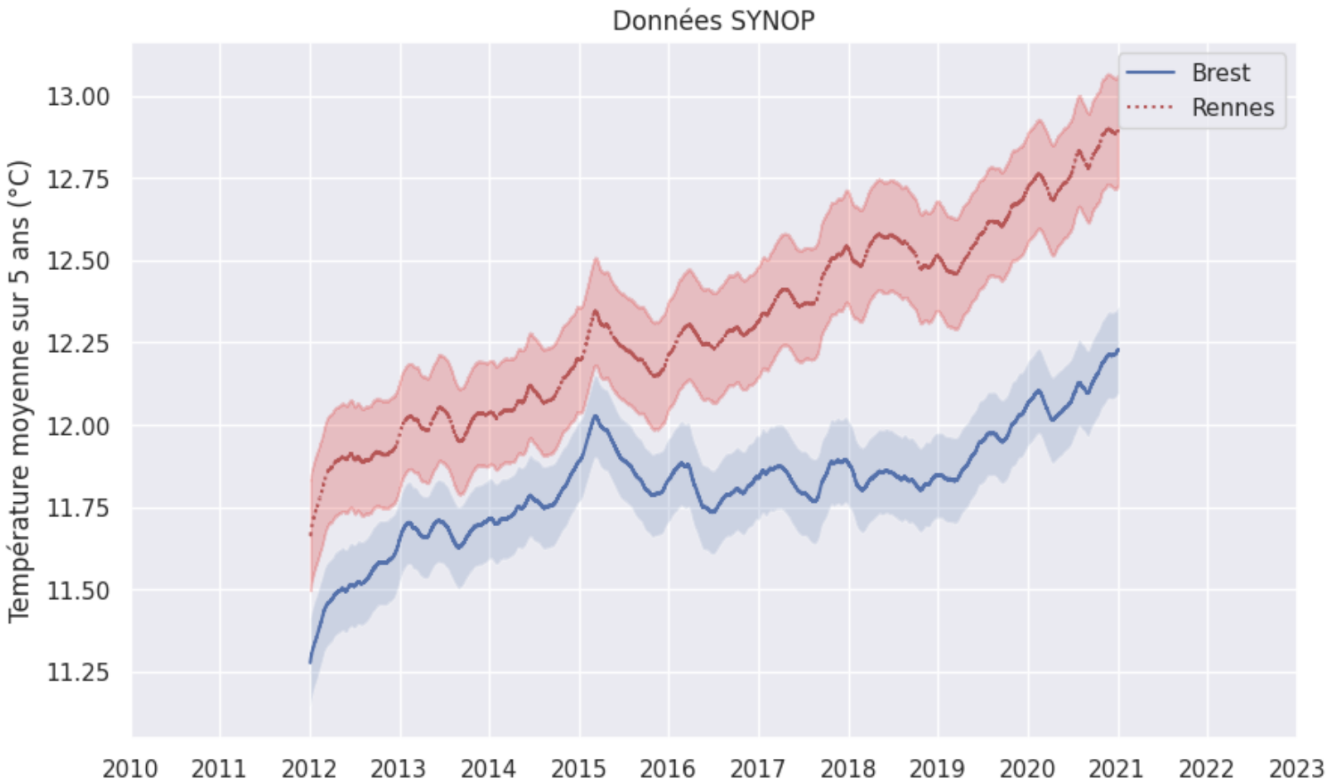
Commenter la figure ci-dessous.



# Comparaison de Brest et Rennes

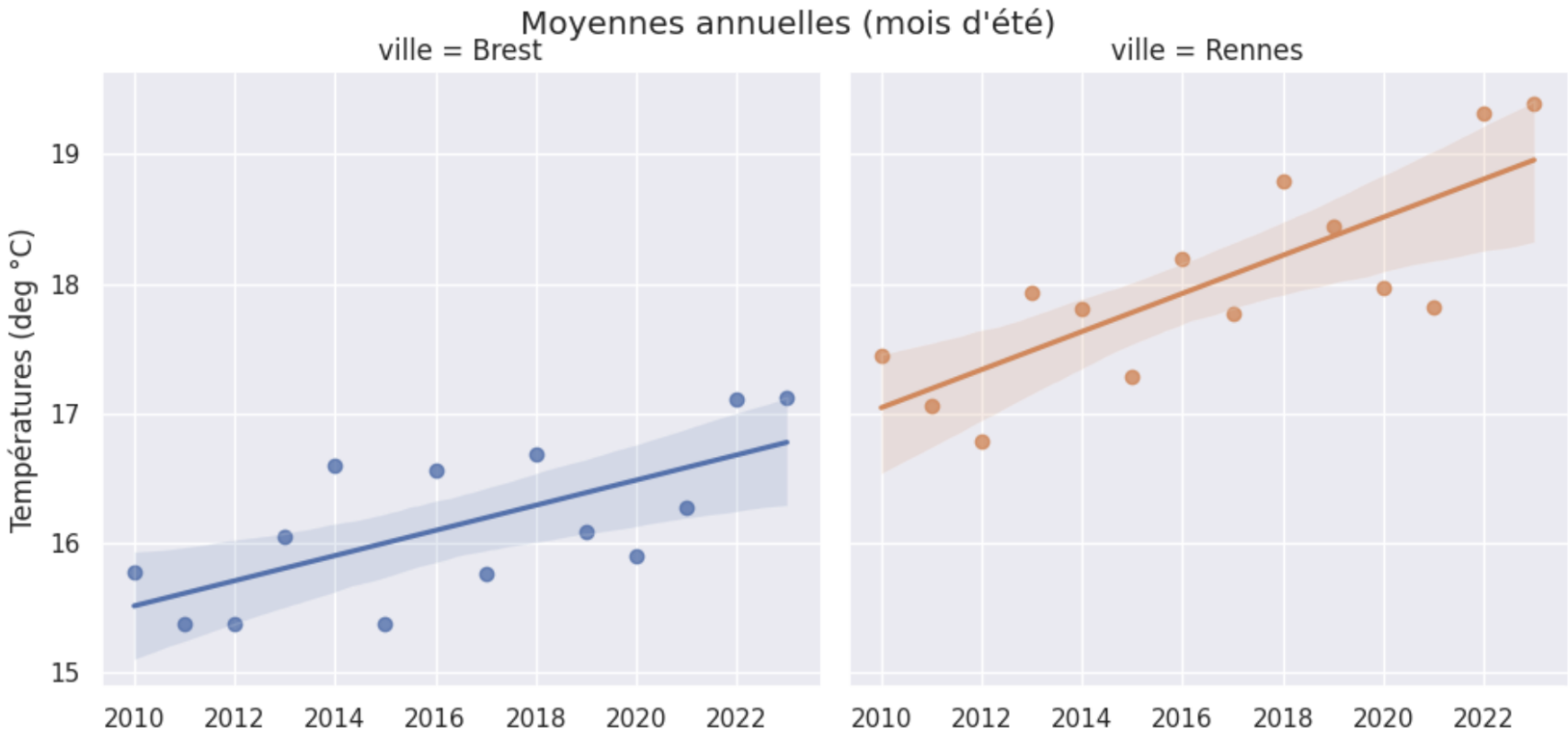
On observe

- A Brest et Rennes la température moyenne est gloablement croissante.
- Les températures de Rennes sont plus chaudes et leur variabilité est plus importante.
- L'année 2025 semble marquer un pic de température moyenne. Mais, attention (!!!),
  - on regarde ici un lissage sur une fenêtre de 5 ans,
  - un réchauffement annuel peut être dû à un hiver doux et/ou un été chaud...



# Zoom sur les mois d'été à Brest et Rennes

La figure présente des moyennes annuelles calculées sur les mois d'été (juin à septembre) ainsi qu'une droite de régression et son intervalle de confiance.
On observe un été 2015 plutôt froid (bien en dessous de la moyenne) .
On remarque que les température à Rennes semble augmenter plus vite qu'à Brest. Le climat de Brest est stabilisé par la proximité de l'océan.



# Données spatiales ou géographiques

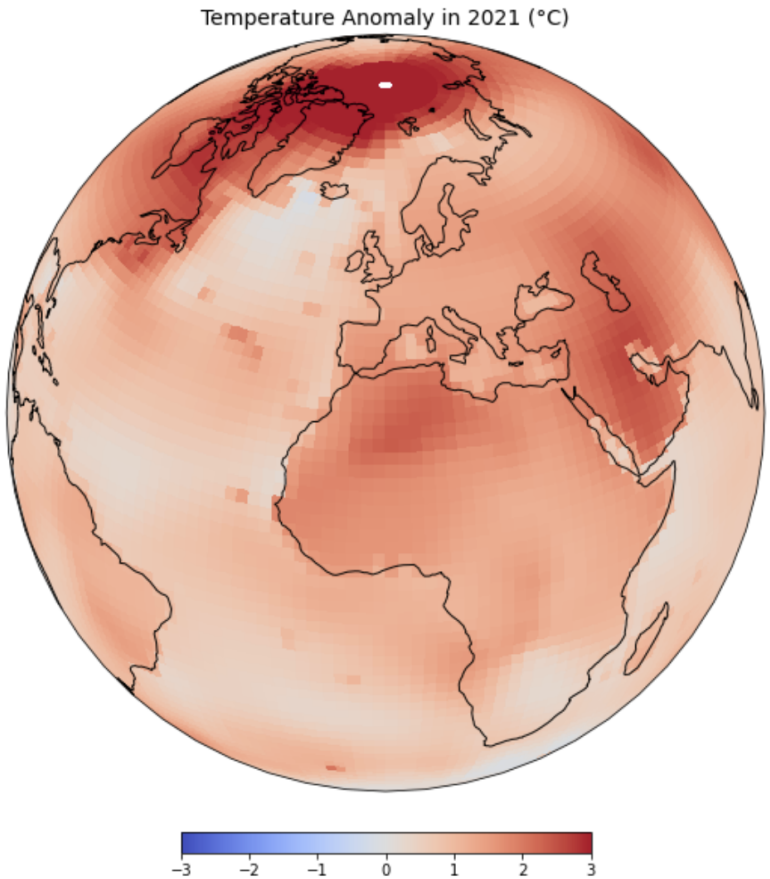
Il existe plusieurs types de données géographiques.

- **Données matricielles (*Raster data*)** : Elles conviennent aux informations continues sans frontières fixes, représentées sous la forme d'une grille de cellules dont les valeurs indiquent les caractéristiques observées. Elles sont souvent contrôlées à intervalles réguliers et interpolées pour créer une surface continue.

- **Données vectorielles (*Vector data*)** : Elles utilisent des points, des lignes et des polygones pour représenter les propriétés spatiales, notamment les points d'intérêt, les réseaux de transport, les limites administratives et les parcelles de terrain. Elles sont souvent utilisées pour des données discrètes avec des positions précises ou des contraintes strictes.

# Données matricielles

Par exemple, la figure ci-dessous est obtenue à partir de données grillées (gridded data) d'**anomalies de température mensuelle** en 2021 en comparaison des températures mensuelles de 1880 au présent.
La couleur rouge montre que l'année 2021 a été plus chaude en moyenne que la moyenne calculée sur les données de 1880 à aujourd'hui.



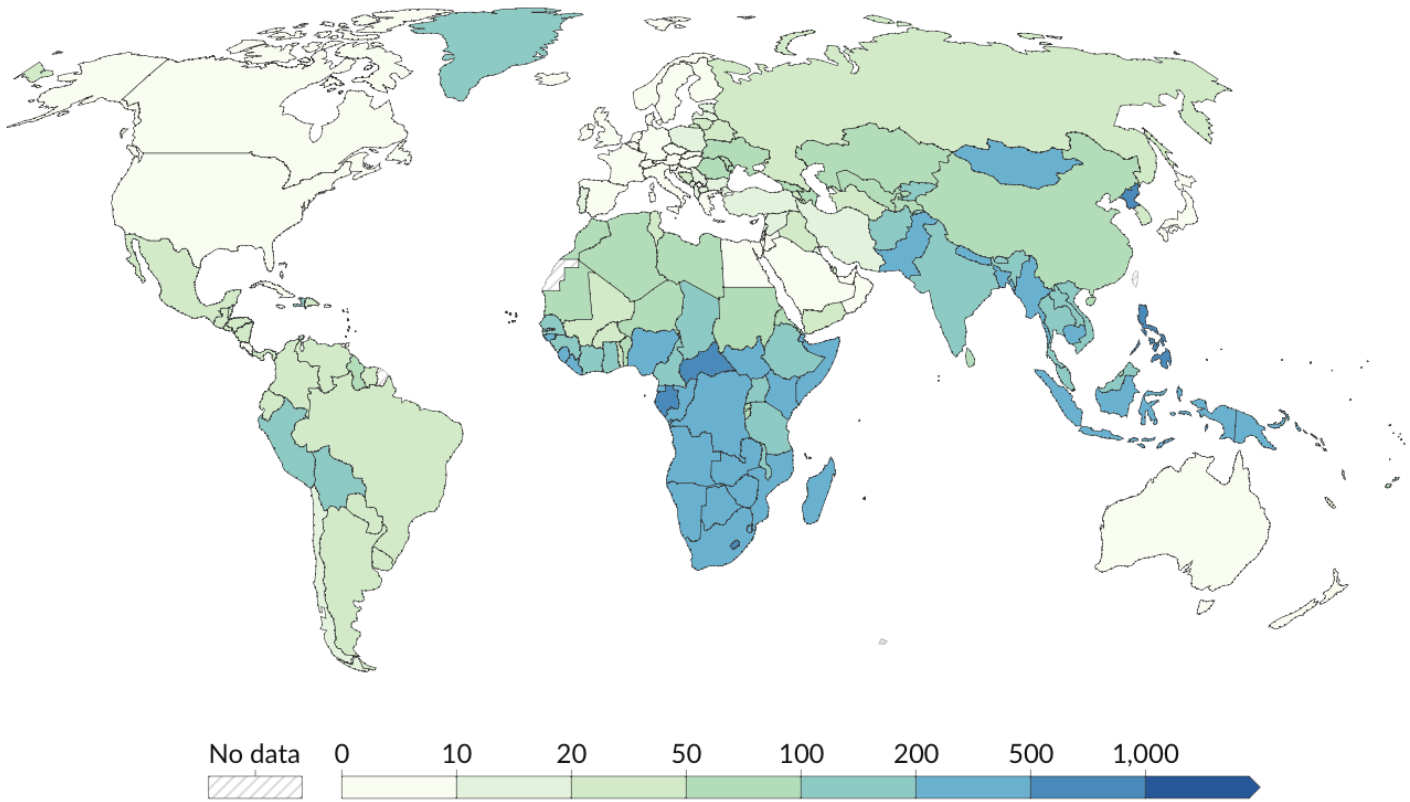

code

# Données vectorielles

Données de surcôte (SHOM)
ou cholera :

# Données vectorielles

Cette carte illustre le Taux estimé de nouveaux cas de tuberculose pour 100 000 personnes. Ce taux comprend à la fois les nouvelles infections et les infections latentes réactivées.



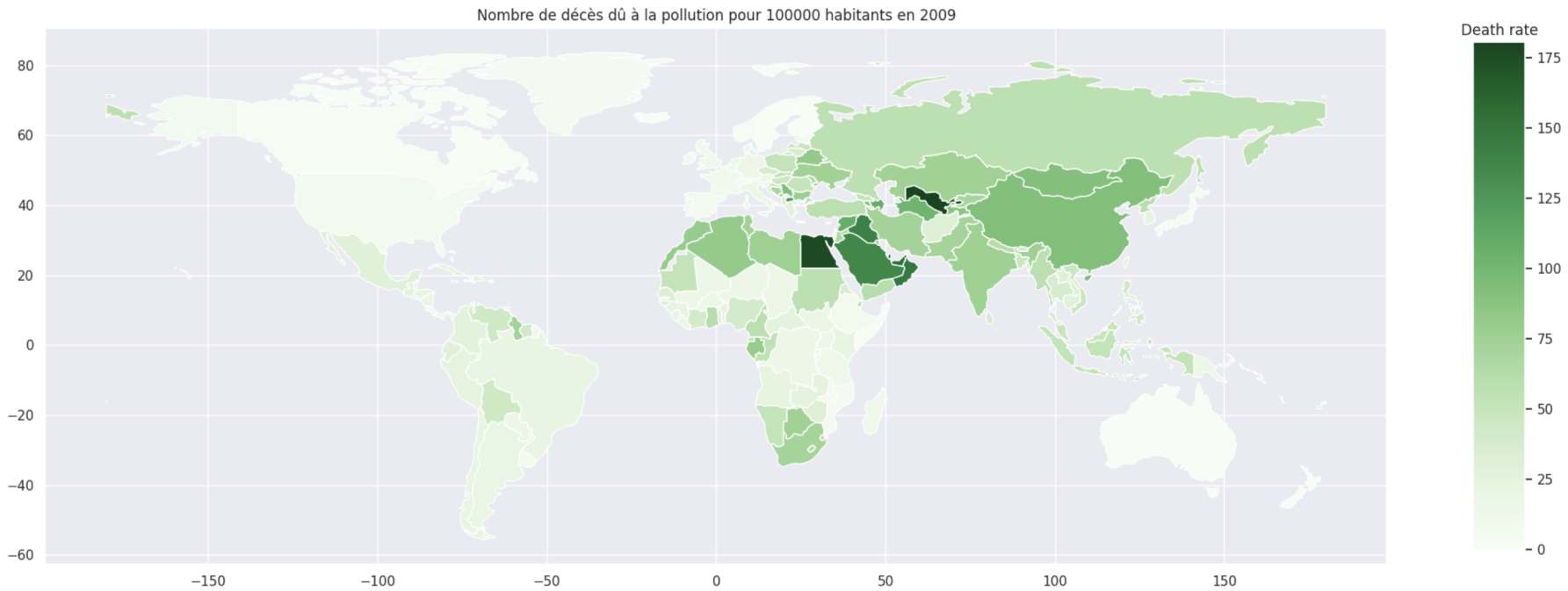

code

# Données géographiques : *choropleth*

Une **carte choroplèthe** est un type de carte dans laquelle différentes zones administratives sont colorées (ou ombrées) en fonction de l'ampleur de leur valeur numérique.

# Données géographiques

Sous python le module geopanda permet de réaliser ce type de carte.
Avec plotly, on obtient une version interactive, utile par exemple pour les sites web.




code

# Outline

Introduction

Différents types de données

Différents graphiques

Données multivariées

# Données multivariées

En science des données, on s'intéresse en général à plusieurs variables et les graphiques peuvent aider à mettre en évidence des liens entre les variables.

L'exemple le plus simple est le nuage de points bivarié (*scatter plot*). Le graphique ci-dessous croise les informations de deux **variables quantitatives** : l'espérance de vie en 1800 et l'espérance de vie en 2020 des pays du monde.




code

# Un peu de vocabulaire

Les données d'espérance de vie se présentent comme suit

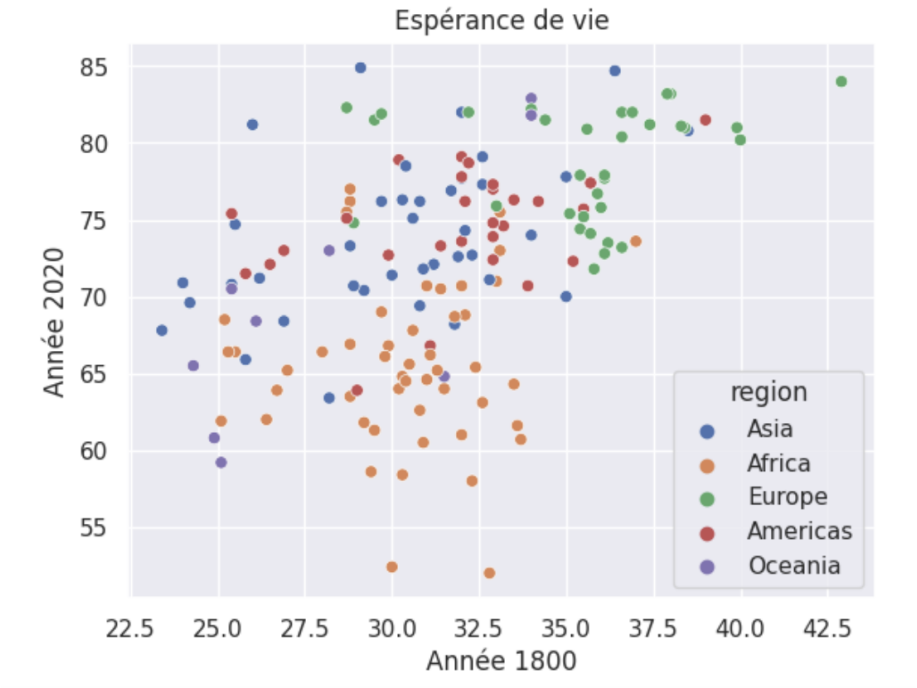
| | country | 1800 | 1801 | 1802 | 1803 | 1804 | 1805 | 1806 | 1807 | 1808 | ... | alpha-2 | alpha-3 | country-code | iso_3166-2 | region | sub-region | intermediate-region |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 28.2 | 28.2 | 28.2 | 28.2 | 28.2 | 28.2 | 28.1 | 28.1 | 28.1 | ... | AF | AFG | 4 | ISO 3166-2:AF | Asia | Southern Asia | NaN |
| 1 | Angola | 27.0 | 27.0 | 27.0 | 27.0 | 27.0 | 27.0 | 27.0 | 27.0 | 27.0 | ... | AO | AGO | 24 | ISO 3166-2:AO | Africa | Sub-Saharan Africa | Middle Africa |
| 2 | Albania | 35.4 | 35.4 | 35.4 | 35.4 | 35.4 | 35.4 | 35.4 | 35.4 | 35.4 | ... | AL | ALB | 8 | ISO 3166-2:AL | Europe | Southern Europe | NaN |
| 3 | Andorra | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | AD | AND | 20 | ISO 3166-2:AD | Europe | Southern Europe | NaN |
| 4 | Argentina | 33.2 | 33.2 | 33.2 | 33.2 | 33.2 | 33.2 | 33.2 | 33.2 | 33.2 | ... | AR | ARG | 32 | ISO 3166-2:AR | Americas | Latin America and the Caribbean | South America |

Chaque pays (on parle aussi d'individu ou d'observation) est décrit par plusieurs variables. Par convention, les **observations** sont en ligne et les **variables** en colonne. Elles forment une **table de données**.

**Variable quantitative** : définie sur une partie ou l'ensemble de $\mathbb{R}$.
**Variable qualitative** : définie sur une ensemble discret.

# Espérance de vie

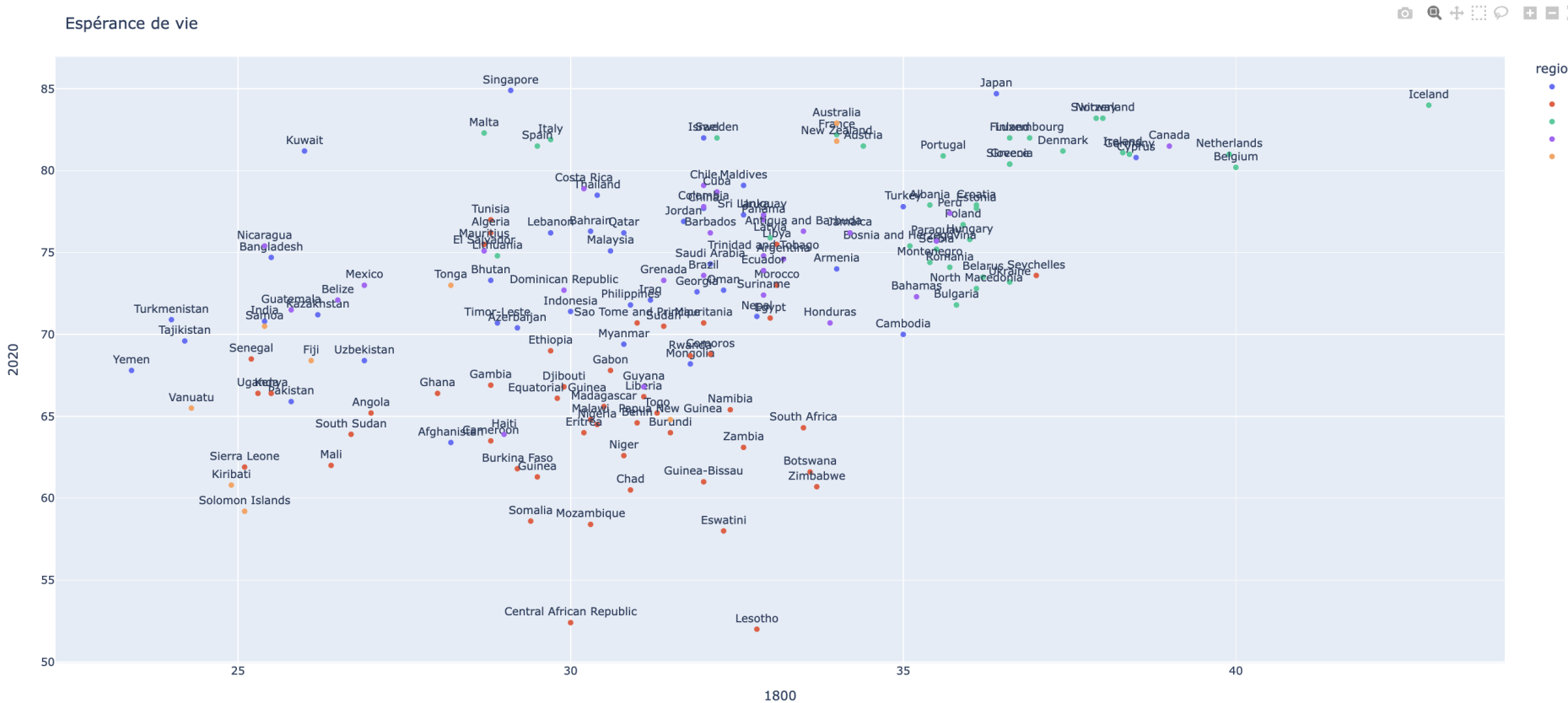
Pour enrichir l'information, on peut ajouter l'information d'une **variable qualitative** : la région du monde. On utilise alors la couleur ou le symbole pour différencier les classes. On représente ainsi les informations provenant de 3 variables.




code

# Espérance de vie

On peut encore ajouter le label de chaque point pour améliorer le pouvoir informatif du graphique.



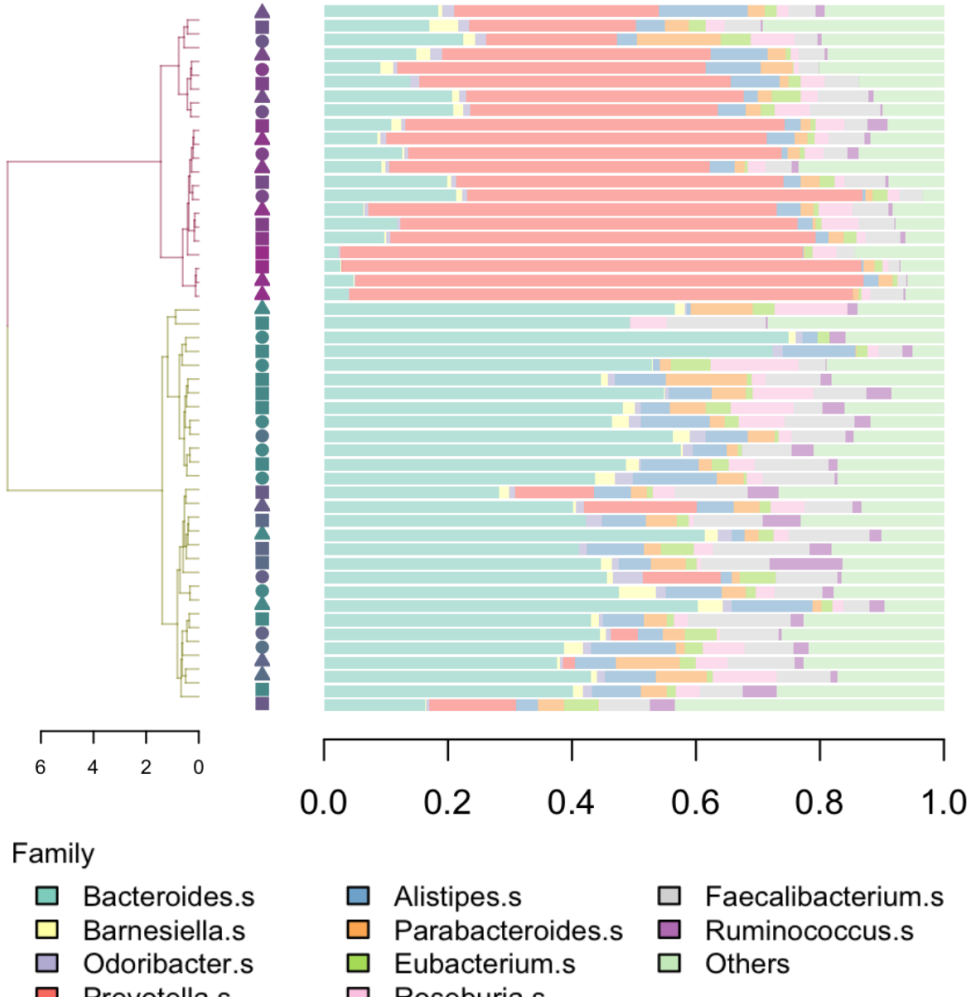

code

# Données de composition multivariées

Le diagramme en batons vertical s'adapte aussi aux données multivariées sous la forme de *stacked barplot*.

Par exemple les composition décrivent un individu par plusieurs variables de comptage.

Le graphique ci-dessous illustre la composition du microbiote intestinal de jeunes gens sportifs de haut niveau (vélo ou foot) et non sportifs.



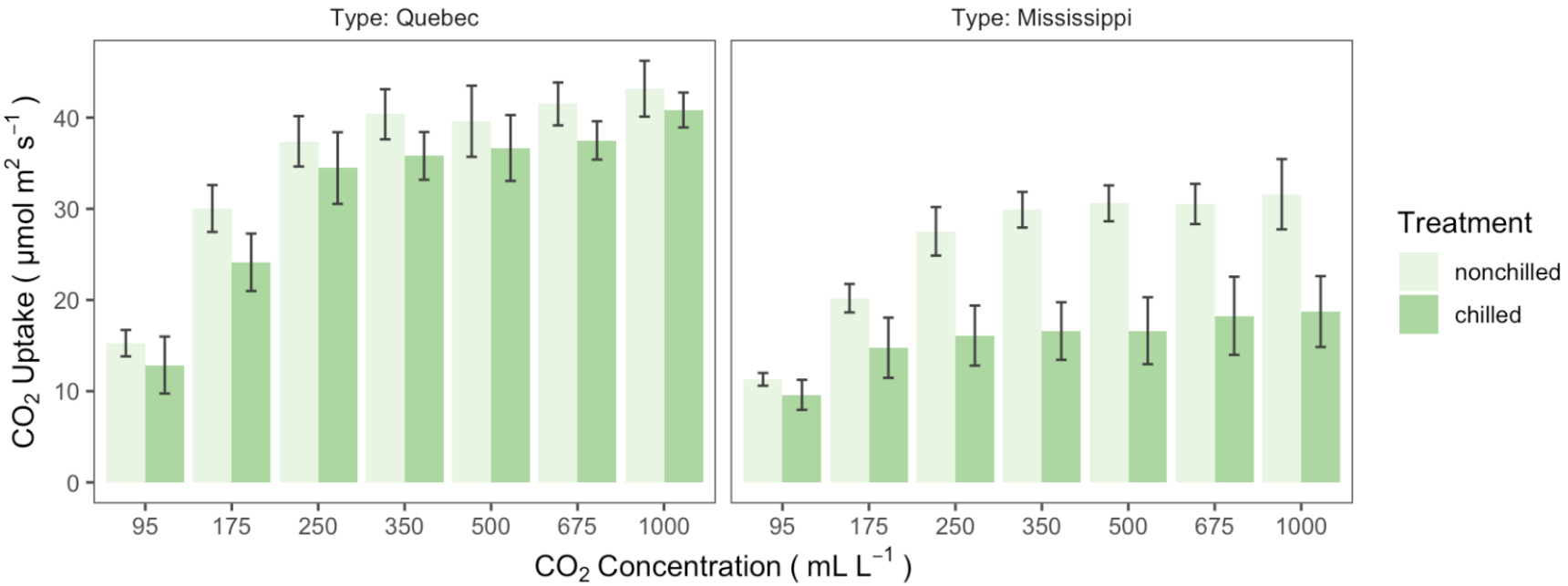
# Diagramme en batons pour analyser le résultat d'une expérience

Nous allons utiliser les données d'une expérience qui a mesuré la tolérance au froid de l'espèce de graminée Echinochloa crus-galli.



Les facteurs étudiés sont le Type, les niveaux 'Québec' et 'Mississippi' donnant l'origine de la plante, le Traitement, 'non réfrigéré' et 'réfrigéré', et la concentration ambiante de CO2 ($mL/L$).
La variable réponse est l'absorption de CO2 ($\mu mol/m^2 s$).
source

# Diagramme en batons pour analyser le résultat d'une expérience




source

# Données textuelles

Les données textuelles sont des données qui regroupent des textes, par exemple un ensemble de tweets. Chaque individu est un tweet. Pour obtenir une table de données, on extrait les (racines des) mots importants des tweets et on compte leur nombre d'occurrences [3]. On obtient ainsi des données de composition.

3. il est nécessaire de nettoyer le text au préalable en retirant les mots de liaison, la ponctuation, etc

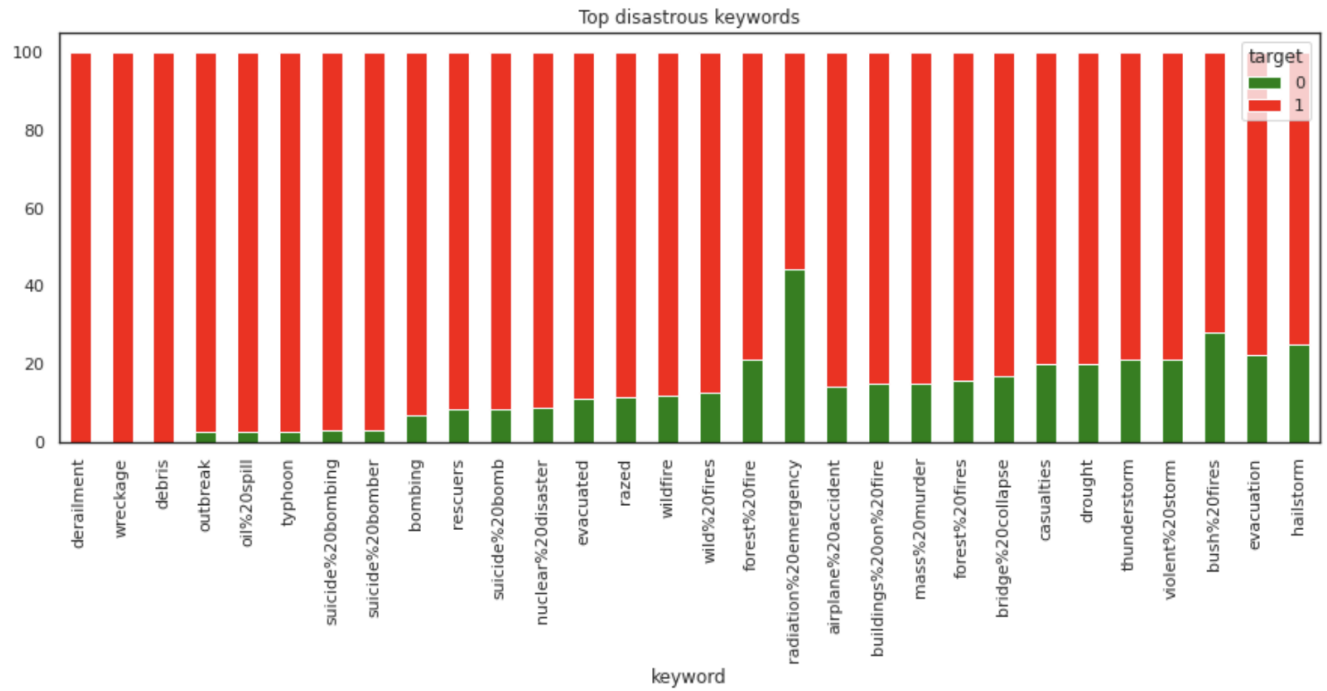
# Catastrophe ?

Twitter est devenu un canal de communication important en cas d'urgence. L'omniprésence des smartphones permet aux gens d'annoncer en temps réel une situation d'urgence qu'ils observent. C'est pourquoi de plus en plus d'organismes s'intéressent à la surveillance programmatique de Twitter (organisations de secours en cas de catastrophe et agences de presse).

Cependant, il n'est pas toujours évident de savoir si les mots d'une personne annoncent réellement une catastrophe.

Dans la compétition kaggle, on propose de construire un (ou des) modèle(s) permettant de prédire si un tweet annonce une catastrophe ou non.

Le graphique illustre le nombre de tweets par mot clé.



# Données multivariées

Les données de composition sont un peu particulières :

- ce sont des données de comptage ;
- les variables sont liées entre elles par le total des comptages ;
- les diagrammes en barres illustrent les comptages mais ne permettent pas de mettre en évidence des relations entre les variables, contrairement aux nuages de points.

Dans un cadre plus général, il n'est pas immédiate de représenter des données de plus de 3 variables : 2 variables quantitatives et une variable qualitative.

# Exemple de données multivariées

On s'intéresse de nouveau à des statistiques mondiales à l'échelle des pays (source : data.world).

Ces données portent principalement sur les indices de développement humain et des indices [4]. l'indice de développement humain mesure la qualité de vie (santé, niveau d'étude, etc).

| | Total in km2 | Water in km2 | Water % | HDI | %HDI Growth | IMF Forecast GDP(Nominal) | World Bank Forecast GDP(Nominal) | UN Forecast GDP(Nominal) | IMF Forecast GDP(PPP) | World Bank Forecast GDP(PPP) | CIA Forecast GDP(PPP) | Internet Users | Population 2022 | Population 2023 | Population %Change |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Russia** | 17098246.0 | 0.957900 | 4.2 | 0.822 | 0.29 | 1862470.0 | 2240422.0 | 1778782.0 | 5056479.0 | 5326855.0 | 3875690.0 | 0.898616 | 1.447133e+08 | 1.444444e+08 | -0.19 |
| **Canada** | 9984670.0 | 0.910747 | 8.9 | 0.936 | 0.25 | 2117805.0 | 2139840.0 | 1988336.0 | 2378973.0 | 2273489.0 | 1742790.0 | 0.875438 | 3.845433e+07 | 3.878129e+07 | 0.85 |
| **Brazil** | 8510346.0 | 0.994133 | 0.6 | 0.754 | 0.38 | 2126809.0 | 1920096.0 | 1608981.0 | 4101022.0 | 3837261.0 | 2989430.0 | 0.763784 | 2.153135e+08 | 2.164224e+08 | 0.52 |
| **India** | 3287263.0 | 0.904458 | 9.6 | 0.633 | 0.88 | 3732224.0 | 3385090.0 | 3201471.0 | 13119622.0 | 11874583.0 | 8443360.0 | 0.616851 | 1.417173e+09 | 1.428628e+09 | 0.81 |
| **Argentina** | 2796427.0 | 0.978638 | 1.6 | 0.842 | 0.09 | 621833.0 | 632770.0 | 487227.0 | 1239515.0 | 1225435.0 | 893310.0 | 0.874400 | 4.551032e+07 | 4.577388e+07 | 0.58 |

Comment représenter ces données sur un seul graphique pour visualiser les pays qui se ressemblent ?

4. Nominal Gross Domestic Product (Nominal GDP) is the total market value of all goods and services produced in a country's economy over a given period.
PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates.

# Outline

Introduction

Différents types de données

Différents graphiques

Données multivariées

# Take home messages

- Les graphiques sont un outil très puissant pour aider à la compréhension et l'interprétation d'un jeu de données.
  L'analyse graphique est souvent la première étape d'un projet de sciences des données.

- Le graphique doit être clair et lisible, accompagné de légendes, titre, etc.

- Différents types de graphiques sont associés à différents types de données.
  Diagramme pour aider au choix du bon graphique

- Au delà de 3 variables, on a recours à des techniques de réduction de données ou de clustering (voir les prochains cours !).