Exploitation d'une Base de Données

Cours 3 - *Open data* & visualisation

Anaïs Durand

28 Février 2023

IUT CLERMONT AUVERGNE
Aurillac - Clermont-Ferrand - Le Puy-en-Velay
Montluçon - Moulins - Vichy

Open data

Open data, données ouvertes/libres

- Données *accessibles* et *utilisables* librement
- *Origine* :
  - privée (entreprises, organisations à but non lucratif, ...), ou
  - publique (collectivités, établissements publics, ...)

Open data, données ouvertes/libres

- Données *accessibles* et *utilisables* librement
- *Origine* :
  - privée (entreprises, organisations à but non lucratif, ...), ou
  - publique (collectivités, établissements publics, ...)

Exemples :

- `https://ressources.data.sncf.com/` :
  données de la SNCF (trafic, retards, ...)
- `https://opendata.clermontmetropole.eu/` :
  données de la métropole de Clermont-Ferrand (monuments historiques, équipements sportifs, indice de qualité de l'air, ...)

Intérêts de l'*open data*

- Pour la *recherche* : libre accès aux résultats scientifiques, reproductibilité des expériences, vérification des résultats ...
- Pour la *transparence*/le *contrôle des institutions* : vérifications des dépenses publiques, contrôle de l'action du gouvernement, ...
- Pour l'*intelligence artificielle (IA)* : plus de données = meilleur apprentissage
- ...

Incitations à l'*open data*

- ***Dans le monde :***
  - 2004 : déclaration signées par les pays membres de l'OCDE pour rendre libres les publications scientifiques financées par des fonds publiques.

Incitations à l'*open data*

- *Dans le monde :*
  - 2004 : déclaration signées par les pays membres de l'OCDE pour rendre libres les publications scientifiques financées par des fonds publiques.

- *En Europe :*
  - 2003, 2008, et 2019 : directives du conseil de l'UE pour inciter à l'open data surtout pour les données publiques
  - 2015 : ouverture du portail européen `https://data.europa.eu/`

Incitations à l'*open data*

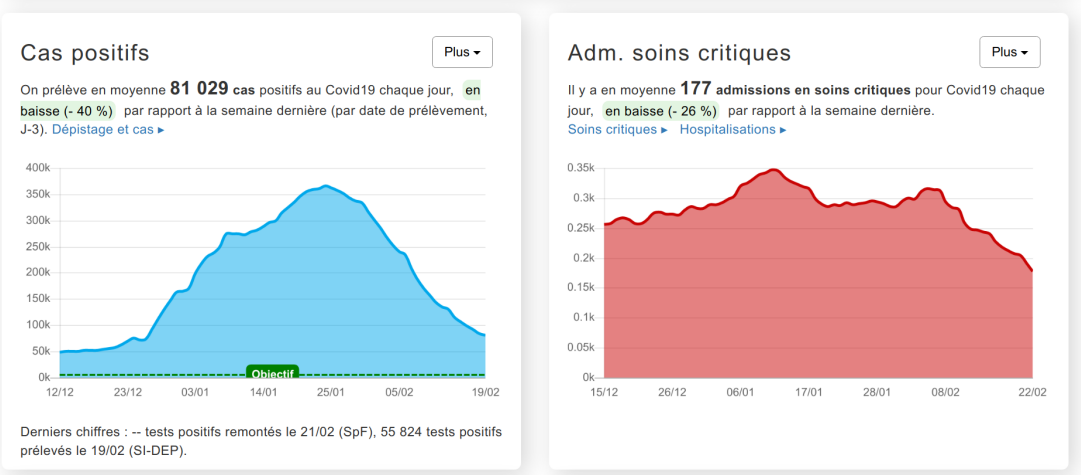
- *En France :*
  - 2011 : ouverture du portail `https://www.data.gouv.fr/`
  - 2014 : mission etalab chargée de promouvoir et d'encadrer l'open data pour les données publiques
  - 2016 : loi pour une République numérique
    $\Rightarrow$ par défaut toute donnée publique doit être ouverte

Incitations à l'*open data*

- *En France :*
  - 2011 : ouverture du portail `https://www.data.gouv.fr/`
  - 2014 : mission etalab chargée de promouvoir et d'encadrer l'open data pour les données publiques
  - 2016 : loi pour une République numérique
    $\Rightarrow$ par défaut toute donnée publique doit être ouverte

- Exemple récent :

  CovidTracker (`https://covidtracker.fr/`)



Utilisation des données ouvertes

Récupérer des données ouvertes

- *Où ?* sur une banque de données en lignes :
  - `https://www.data.gouv.fr/`
  - `https://data.europa.eu/`
  - `https://www.data.gov/`
  - `https://www.kaggle.com/datasets`
  - ...

Récupérer des données ouvertes

- *Où ?* sur une banque de données en lignes :
  - `https://www.data.gouv.fr/`
  - `https://data.europa.eu/`
  - `https://www.data.gov/`
  - `https://www.kaggle.com/datasets`
  - ...

- *Sous quel format ?*
  un ou plusieurs fichiers au format `.csv`, `.json`, `.xls`, ...

Préparation des données

Problèmes possibles :

données non structurées, incomplètes, incohérentes ...

$\Rightarrow$ Besoin de préparer les données avant de les exploiter

- Structuration en tables
- Suppression des doublons
- Suppression ou complétion des données manquantes
- Rectification des incohérences
- ...

Outils utilisés dans ce cours

- SGBD *PostgreSQL*
- Langage de programmation *Python*

Outils utilisés dans ce cours

- SGBD *PostgreSQL*
- Langage de programmation *Python*
  - psycopg2 : connexion à la base PostgreSQL
  - pandas : manipulation des données et interrogation de la base
  - numpy : manipulation des données

pandas

NumPy

Outils utilisés dans ce cours

- SGBD *PostgreSQL*
- Langage de programmation *Python*
  - psycopg2 : connexion à la base PostgreSQL
  - pandas : manipulation des données et interrogation de la base
  - numpy : manipulation des données

- Exemples dans le polycopié de cours + documentation des bibliothèques

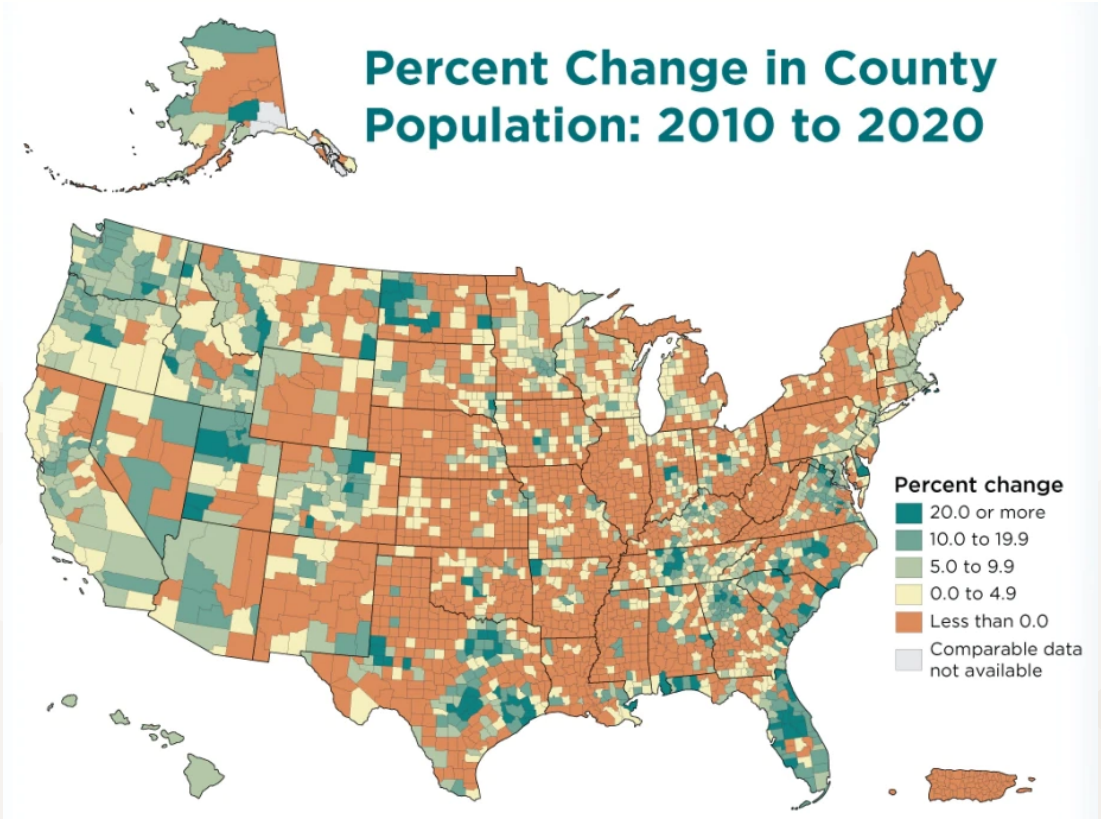
Visualisation

Visualisation de données

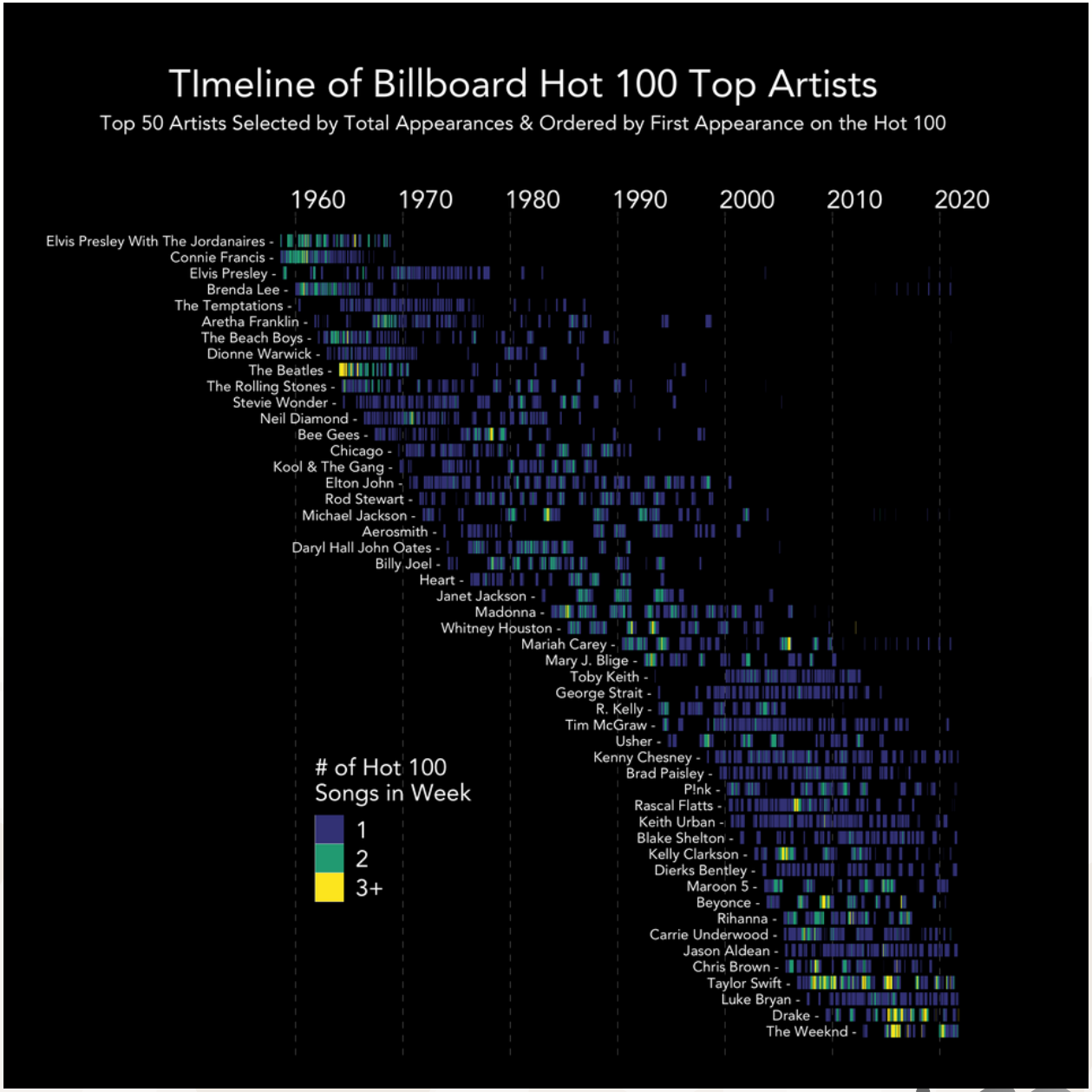
Réprésentation graphique des données

$\Rightarrow$ plus *lisible* que les données brutes

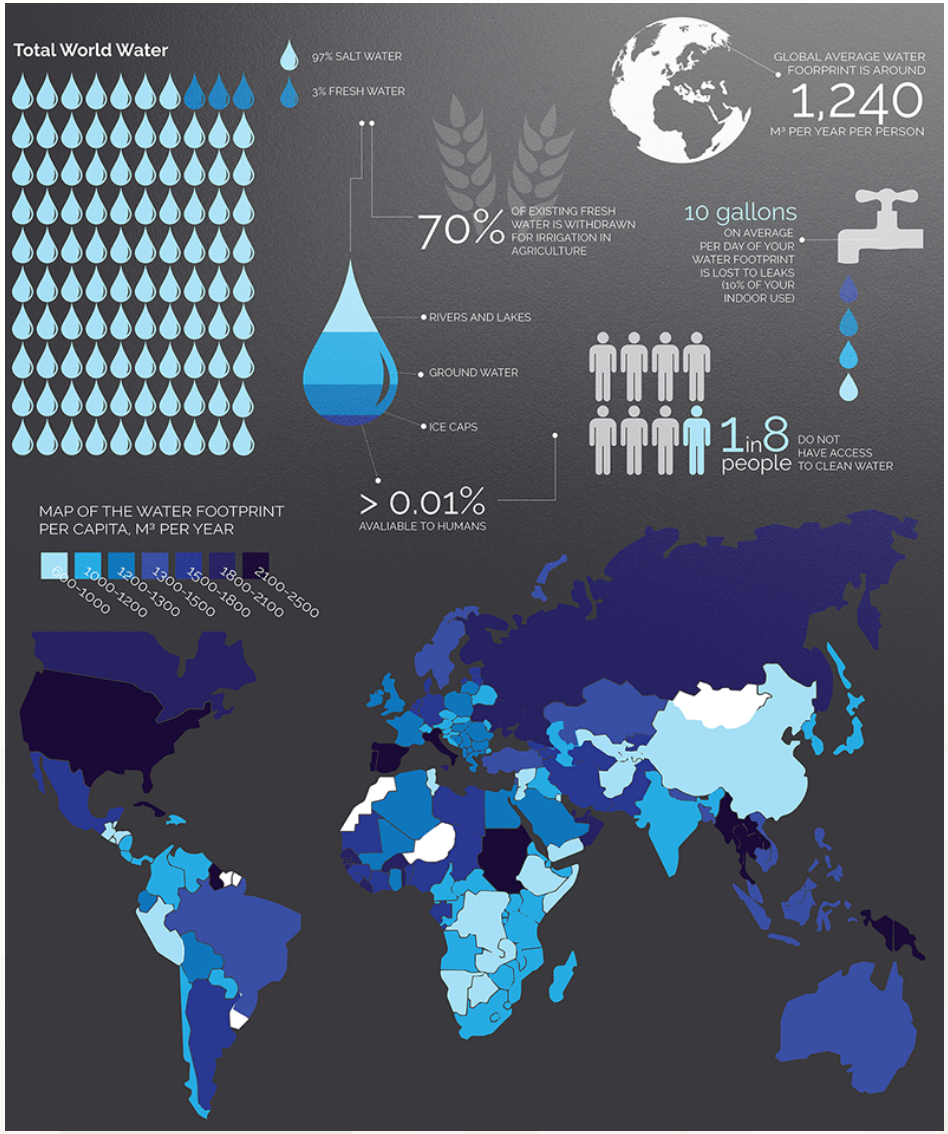
Exemple :



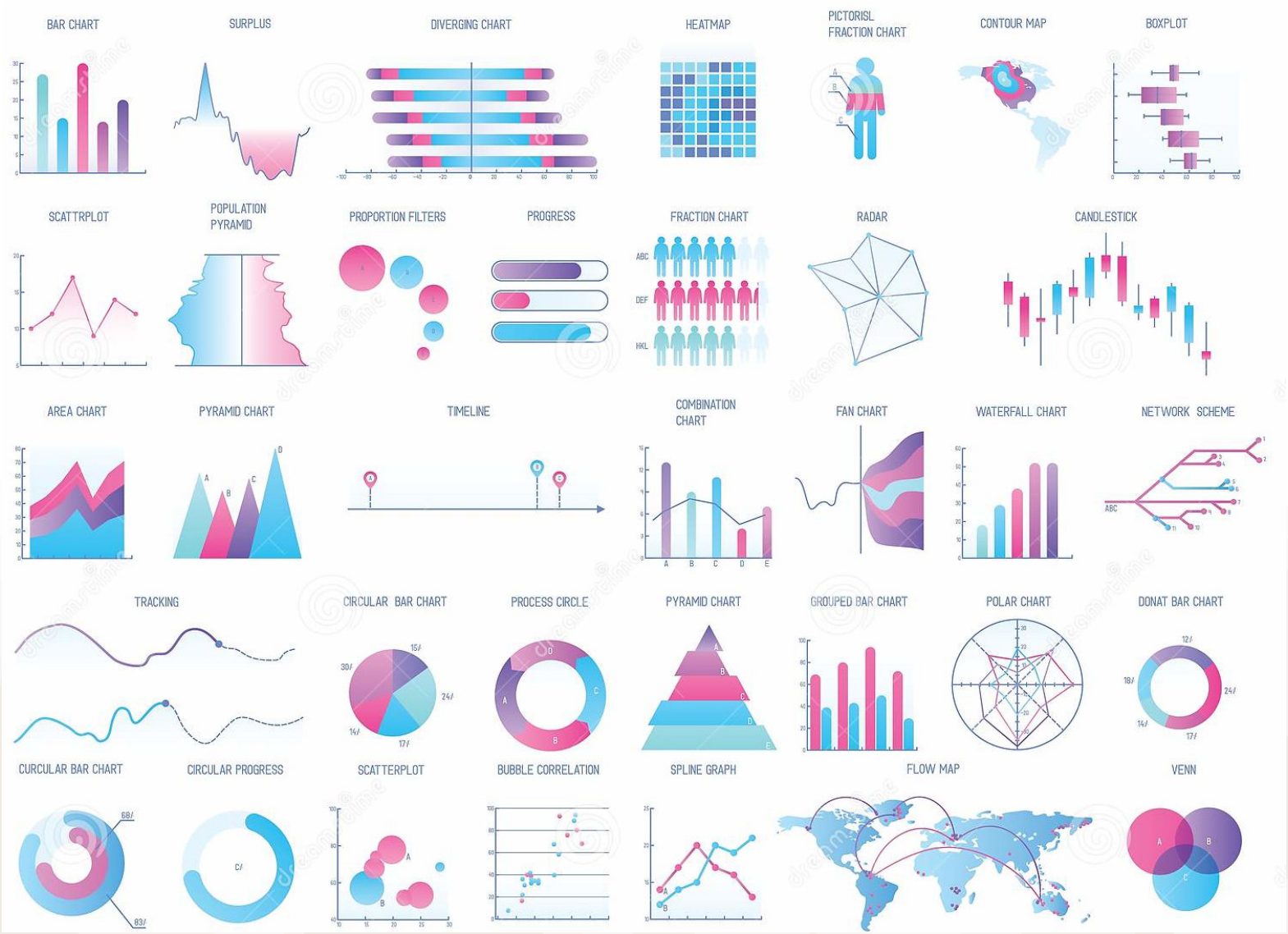
Exemple :



Exemple :


Total World Water
97% SALT WATER
3% FRESH WATER
GLOBAL AVERAGE WATER FOORPRINT IS AROUND
1,240
M³ PER YEAR PER PERSON
70% OF EXISTING FRESH WATER IS WITHDRAWN FOR IRRIGATION IN AGRICULTURE
10 gallons
ON AVERAGE PER DAY OF YOUR WATER FOOTPRINT IS LOST TO LEAKS (10% OF YOUR INDOOR USE)
RIVERS AND LAKES
GROUND WATER
ICE CAPS
1 in 8 people DO NOT HAVE ACCESS TO CLEAN WATER
> 0.01%
AVALIABLE TO HUMANS
MAP OF THE WATER FOOTPRINT PER CAPITA, M³ PER YEAR
0-1000
1000-1200
1200-1300
1300-1500
1500-1800
1800-2100
2100-2500

Choix de la représentation



Choix de la représentation

- Montrer une évolution dans le temps

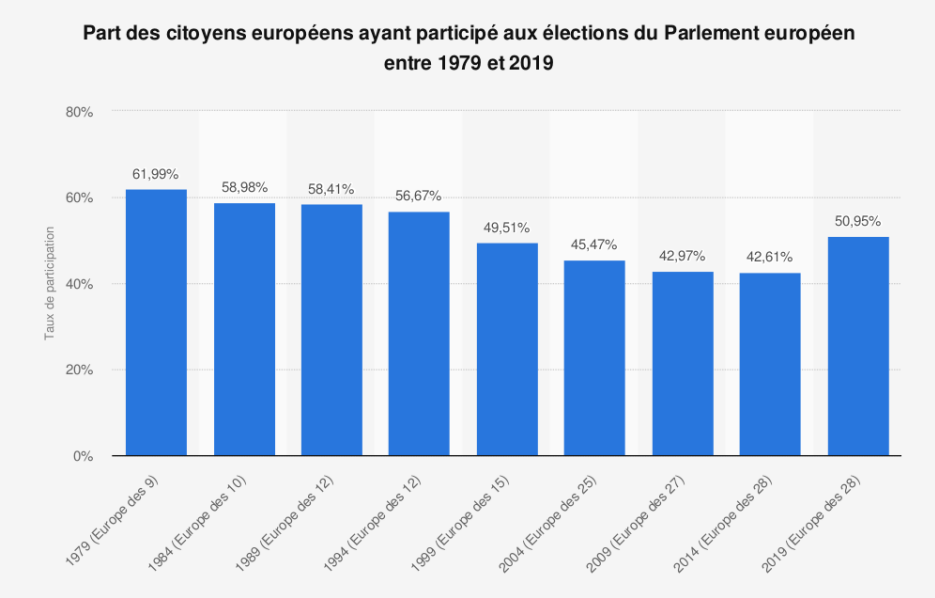
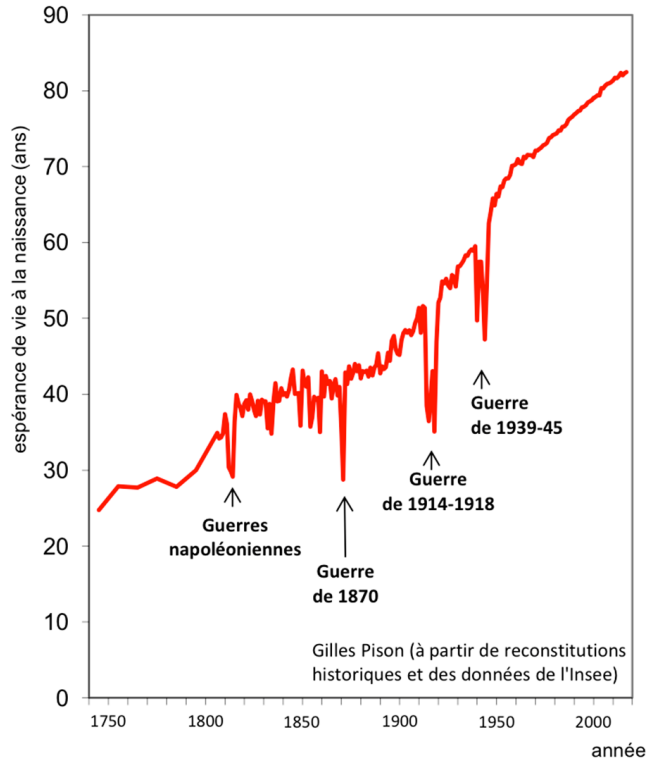
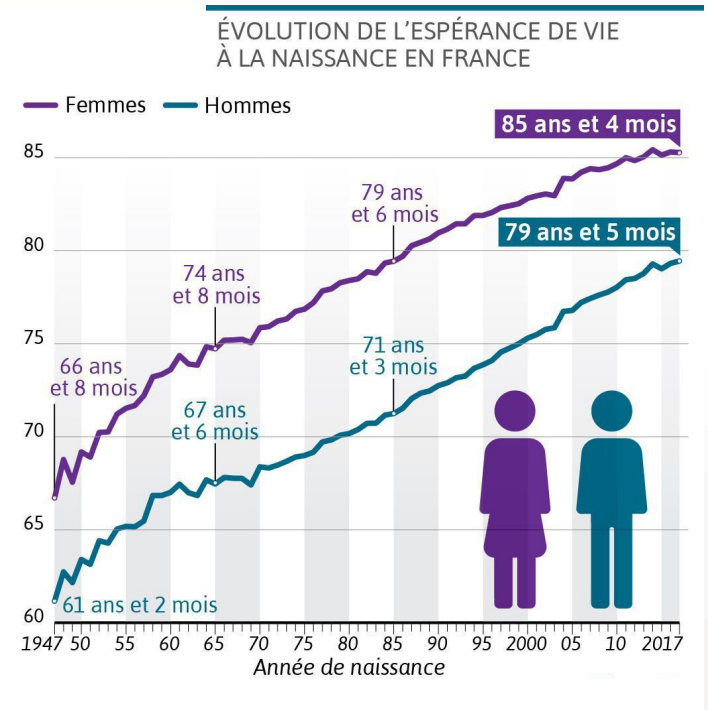


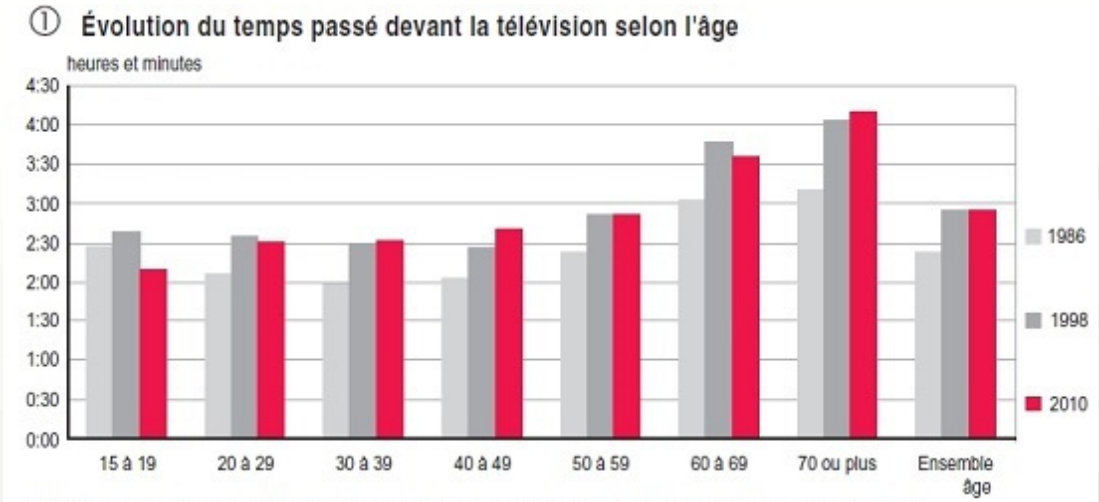
Figure 1. Évolution de l'espérance de vie à la naissance en France de 1740 à 2017



Choix de la représentation

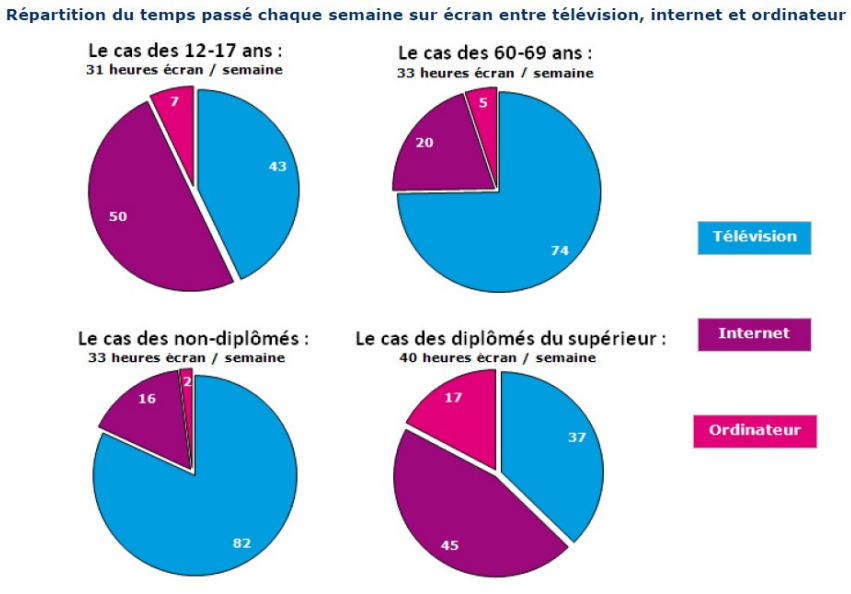
- Comparer des valeurs

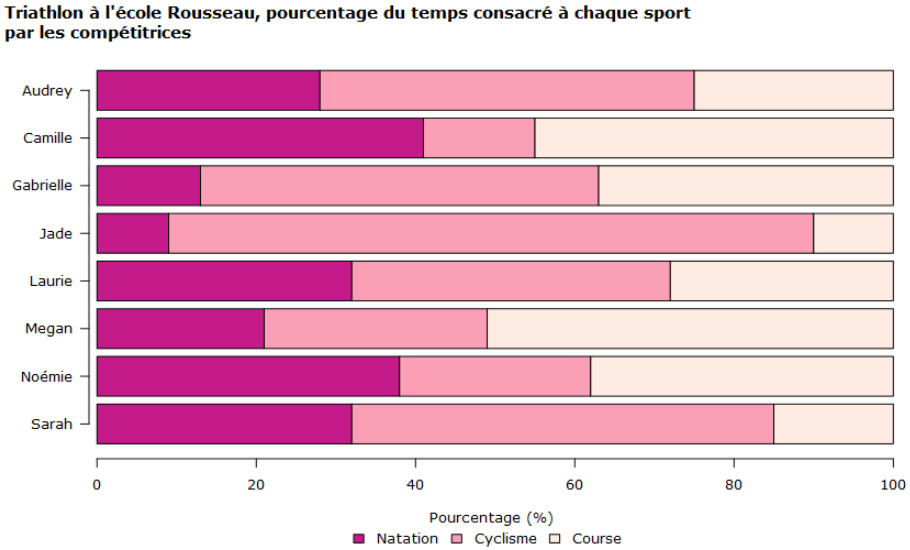




Choix de la représentation

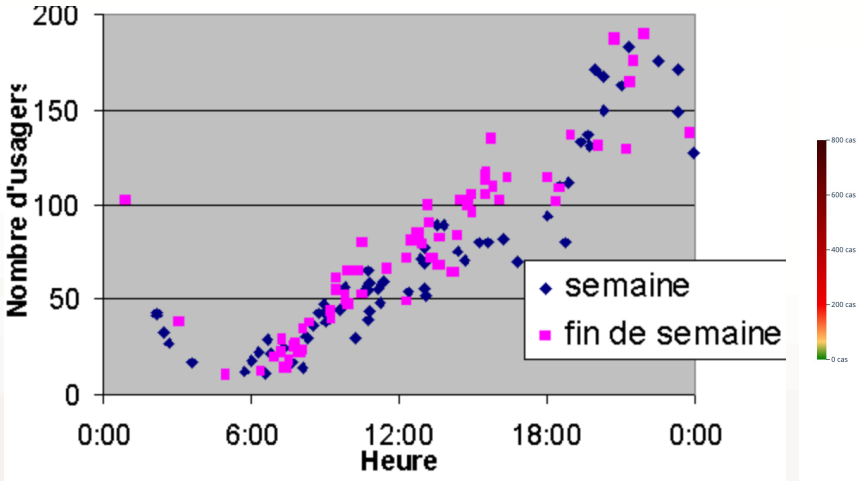
- Montrer une répartition/des proportions

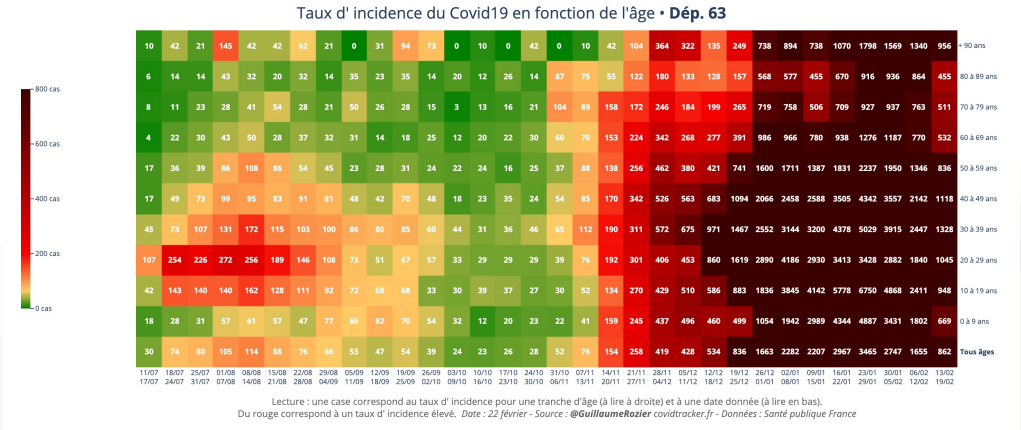




Choix de la représentation

- Montrer une corrélation





Préparer les données à visualiser

- Filtrer les données
  $\Rightarrow$ choix des données à afficher, du format
- Trier les données
  $\Rightarrow$ afficher dans l'ordre le plus lisible (par valeur croissante, ...)
- Calculer des statistiques
  $\Rightarrow$ moyenne, médiane, écart-type ...

Outils utilisés dans ce cours

- SGBD *PostgreSQL*
- Langage de programmation *Python*
  - pandas : manipulation des données, interrogation de la base, affichage
  - matplotlib : création de graphiques

Outils utilisés dans ce cours

- SGBD *PostgreSQL*
- Langage de programmation *Python*
  - pandas : manipulation des données, interrogation de la base, affichage
  - matplotlib : création de graphiques

- Exemples dans le polycopié de cours + documentation des bibliothèques