TP3 Analyse Multivariée en Python

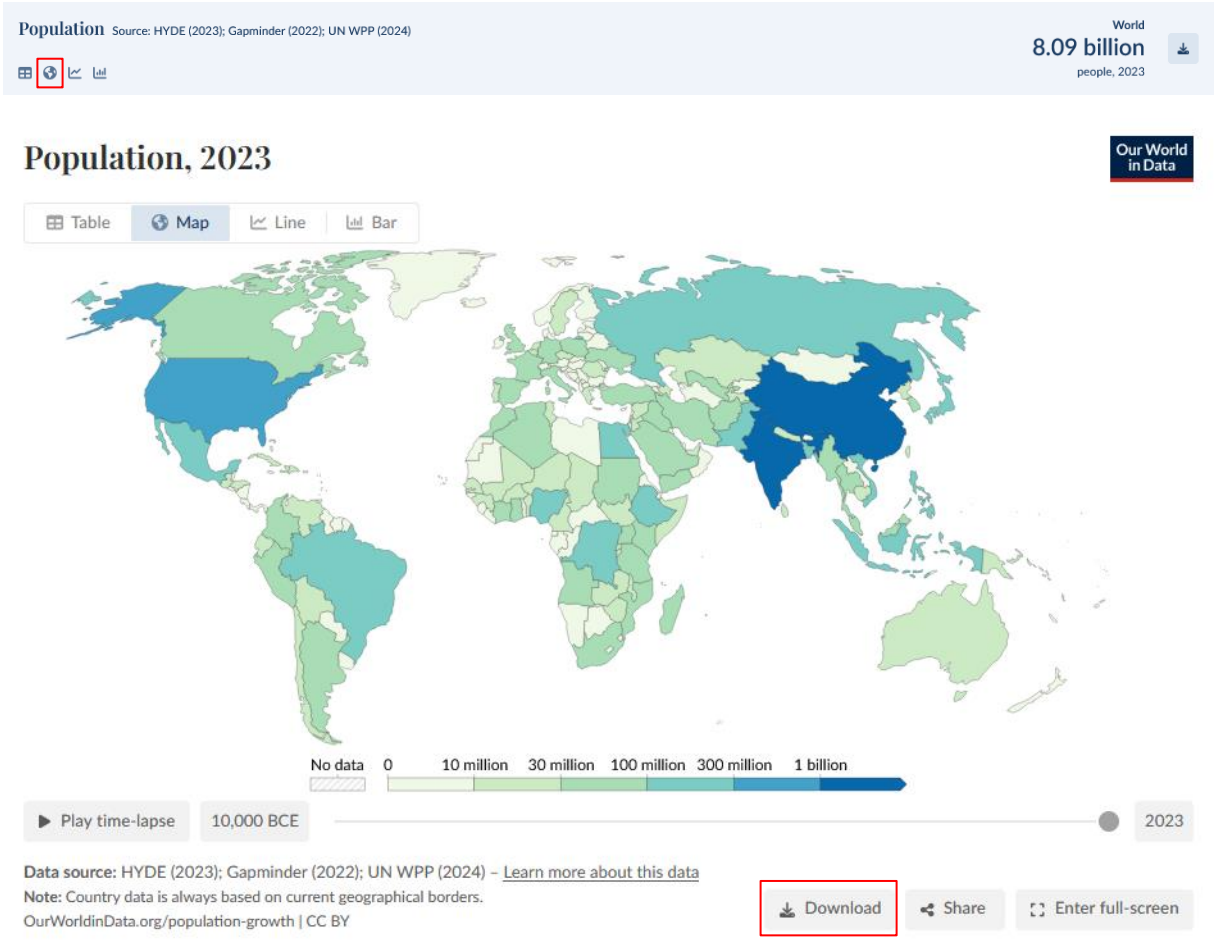
2025-2026

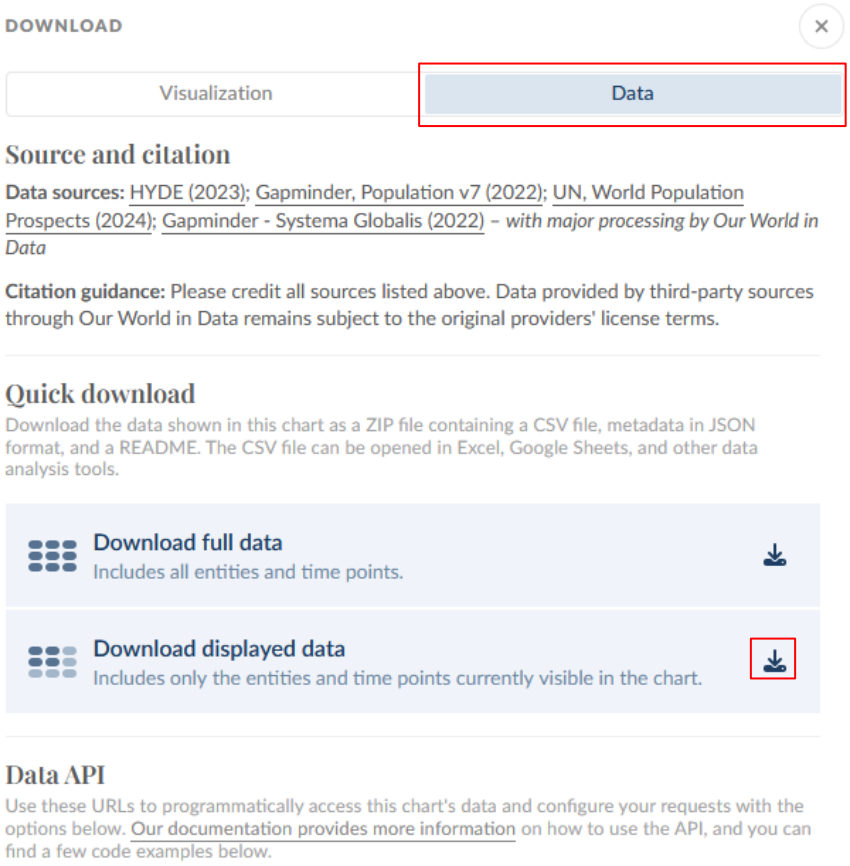
1 Les données

Le site https://ourworldindata.org/search regroupe plusieurs jeux de données à l'échelle mondiale, sur des sujets comme la démographie, la santé, l'énergie, l'environnement, l'agriculture, la pauvreté, l'éducation… Ces jeux de données sont issus de la recherche, ou produits par des institutions nationales ou internationales ou par des organisations non gouvernementales. Le site est maintenu par Global Change Data Lab, un organisme à but non lucratif et par l'université d'Oxford.

Choisissez plusieurs jeux de données (au moins 5) décrivant des variables quantitatives par pays (indiqués par le symbole globe). Pour chaque variable ne prendre qu'une seule date. Réfléchir au choix des dates.

Exemple :





Pour la visualisation en Python des résultats, vous pouvez utiliser les polygones issus de ce dépôt github : https://github.com/datasets/geo-countries/blob/main/data/countries.geojson. Cependant, les pays n'ont pas d'identifiants, il faudra faire une jointure sur le nom des pays, mais ceux-ci certains sont notés différemment entre les deux jeux de données :

```
{{
    "United States of America": "United States",
    "Democratic Republic of the Congo": "Democratic Republic of Congo",
    "Ivory Coast": "Cote d\'Ivoire",
    "Republic of the Congo": "Congo",
    "Czech Republic": "Czechia",
    "The Bahamas": "Bahamas",
    "Guinea Bissau": "Guinea-Bissau",
    "Federated States of Micronesia": "Micronesia (country)",
    "Macedonia": "North Macedonia",
    "eSwatini": "Eswatini",
    "Republic of Serbia": "Serbia",
    "United Republic of Tanzania": "Tanzania",
    "São Tomé and Principe": "Sao Tome and Principe",
    "Cape Verde": "Cabo Verde"
}
```

2 Construction et exploration du jeux de donnée

- Construisez un `GeoDataframe` avec pour géométrie le contour des pays et comme données les variables issues de Our World In Data.
- Certaines variables sont incomplètes pour certains pays (valeurs `NaN`, c'est-à-dire Not a number). Que faire ?
- Produire une carte pour chaque variable.
- La méthode pandas `.describe()` permet d'afficher plusieurs statistiques univariées en lien avec chaque variable : moyenne, écart-type, minimum, maximum, quartiles.
- Utilisez la fonction `pd.plotting.scatter_matrix()` pour visualiser les liens statistiques entre chaque paire de variables. Lesquelles semblent avoir un lien ? Décrire ces liens.
- Calculer la matrice de corrélation et commenter.

3 Analyse en composantes principales

L'objectif de cette partie est de réaliser et d'analyser une ACP.

- Centre-réduire les variables.
- Réaliser l'ACP (en utilisant `sklearn.decomposition.PCA`):
  - Construire un objet `pca` de la classe `PCA`
  - Transformer les données à l'aide de la méthode `.fit_transform()` de l'objet `pca`
  - Remettre les données transformées dans un nouveau `geodataframe`
- Afficher le taux de variance/d'inertie expliqué en fonction du nombre de dimensions de l'ACP (attribut `explained_variance_ratio_` de l'objet `pca`). Rappelez l'intérêt de la réduction de dimension. Si vous utilisiez cette ACP pour réduire le nombre de variables, combien de composantes garderiez-vous ?
- L'attribut `components_` de l'objet `pca` contient pour chaque composante principale les coefficients de la combinaison linéaire des variables initiales qui permet de l'obtenir. (Par exemple (`pca.components_[0]` pour la première composante principale). Comment peut-on interpréter les deux premières composantes ?
- Faire la carte des deux premières dimensions de l'ACP.
- Quelles sont par exemple les valeurs pour la France ?
- Pour obtenir le cercle des corrélations, faire pour chaque variable i une flèche d'origine 0 et de destination la correlation de la variable i avec les 2 premières composantes (qu'on peut obtenir comme `pca.components_[0][i] * np.sqrt(pca.explained_variance_[0])` et `pca.components_[1][i] * np.sqrt(pca.explained_variance_[1])`). Quelles variables sont corrélées à ces composantes principales ? Lesquelles sont corrélées entre elles ?

4 Classification non supervisée

L'objectif de cette partie est de comparer les résultats de plusieurs classifications non supervisées. Vous pouvez effectuer ces regroupements en utilisant soit les variables initiales, soit les premières composantes principales (et comparer les deux si vous avez de l'avance).

- Construire et interpréter un dendrogramme avec le critère de Ward. Ne pas hésiter à faire une très grande figure et à l'enregistrer. Quel pays sont les moins similaires à la France ?
- Effectuer une classification ascendante hiérarchique avec 2 classes, puis avec 5 classes et afficher les cartes correspondantes. Colorier les nuages de points des 2 premières composantes principales en fonction de la classe des pays. Quels pays ont étés regroupés ?
- Bonus : Faire la même chose avec K-means (et bonus DBSCAN), et comparer. (Combien de classes DBSCAN a-t-il trouvé ?)
- Bonus : Evaluer et comparer la qualité de ces regroupements en calculant **Q = Inertie inter classe / inertie totale**
    - Calculer le centre de chaque classe et de l'ensemble des points
    - Calculer l'inertie totale = somme des distances au carré entre les variables des pays et le centre
    - Calculer l'inertie inter classe = somme des distances au carré entre le centre des classes et le centre de l'ensemble des données

5 Classification supervisée

Dans cette partie, on va chercher à utiliser les variables pour prédire le niveau de démocratie des pays. Le jeu de données https://ourworldindata.org/grapher/political-regime classifie les pays en 4 types de régimes, selon les travaux des chercheurs politiques Anna Lührmann, Marcus Tannenberg and Staffan Lindberg :

- Dans **les autocraties fermées**, les citoyens n'ont pas le droit de choisir le chef du gouvernement ou le corps législatif par le biais d'élections multipartites.
- Dans **les autocraties électorales**, les citoyens ont le droit de choisir le chef de l'exécutif et le corps législatif par le biais d'élections multipartites, mais ils ne disposent pas de certaines libertés, telles que la liberté d'association ou d'expression, qui rendent les élections significatives, libres et équitables.
- Dans les **démocraties électorales**, les citoyens ont le droit de participer à des élections significatives, libres et équitables et multipartites.
- Dans les **démocraties libérales**, les citoyens jouissent d'autres droits individuels et minoritaires, sont égaux devant la loi et les actions de l'exécutif sont limitées par le législatif et les tribunaux.

- Faire la carte de cette nouvelle variable, calculer l'effectif de chaque classe, et le mode.
- Comparer les valeurs de chacune de vos variables en fonction des types de régimes. Quels outils vus dans les cours précédent pouvez-vous utiliser ?
- Séparez aléatoirement les pays en deux groupes d'entraînement et de test.
- Entrainez un modèle de forêts aléatoires, et prédire les régimes des pays du jeu de test.
- Calculer la matrice de confusion et les rappels et précisions par classes et globaux. Certains régimes sont-ils bien identifiés ? Est-ce qu'il y a des confusions entre certaines classes ? Afficher la carte des régimes prédits sur le jeu de test et comparer la à la vérité terrain.
- Bonus : Pour chaque classe, faire la carte (sur le jeu de test) des probabilités de cette classe prédites par le modèle.
- Bonus : Faire varier les hyperparamètres du modèle pour voir l'impact sur les résulats de classification.