RAPPORT DE PROJET

Projet : Température Terrestre
Cursus : Data Analyst — Bootcamp Mars 2025
Auteur : Nicolas Quéquet
Groupe : Arthur Roger, Solen Billot, Yann Dumeur, Nicolas Quequet
Niveau de difficulté : 06/10

Objectif du projet :
Constater le réchauffement (et le dérèglement) climatique global à l'échelle de la planète sur les derniers siècles et dernières décennies.

- Analyse au niveau mondial
- Analyse par zone géographique
- Comparaison avec des phases d'évolution de température antérieure à notre époque

Données utilisées :
NASA GISTEMP — https://data.giss.nasa.gov/gistemp/
OWID CO_2 — https://github.com/owid/co2-data

Dossier partagé du groupe (Google Drive) :
https://drive.google.com/drive/folders/1wfae_AZAEwtWBaZWDHvzw1XYdq_0D5EJ

Température Terrestre — Rapport d'Exploration et Prétraitement

Introduction

Ce projet porte sur l'analyse des températures terrestres et du réchauffement climatique. Notre objectif, dans un premier temps, est d'étudier l'évolution des températures à travers le temps depuis les deux derniers siècles, puis d'en déterminer certains facteurs.

Pour ce faire, nous disposons de plusieurs jeux de données :

- Évolution des anomalies de températures par rapport à une moyenne de référence ; plusieurs jeux de données observées par différents capteurs, sur des périodes de temps différentes et avec différentes moyennes de référence, à l'échelle mondiale et par régions du monde. Ces données proviennent du GISS (Goddard Institute for Space Studies) de la NASA (National Aeronautics and Space Administration).
- Évolution de la consommation et production d'énergie à travers le temps. Ces données proviennent de l'organisation Our World in Data.
- Évolution des émissions de gaz à effet de serre à travers le temps. Ces données proviennent de l'organisation Our World in Data.

Pour cette première partie de projet, mon travail a été de me focaliser sur ce dernier jeu de données. Ma mission était de comprendre, nettoyer et analyser le fichier intitulé owid-co2-data.csv, contenant de nombreuses informations sur les émissions de différents gaz à effet de serre.

Compréhension du jeu de données

La première étape fut de comprendre ce jeu de données.
Il s'agit d'un fichier CSV contenant 50191 entrées et 79 colonnes.

Chaque entrée correspond à l'observation des différentes données pour une année et une zone spécifique.
Vous trouverez des informations sur les sources des données dans la partie Souces à la fin du rapport.

Dans la partie suivante, nous décrirons les colonnes du jeu de données. Certaines colonnes sont très similaires, elles suivent la même logique ; nous ne les détaillerons pas forcément toutes individuellement mais nous expliquerons les logiques identiques.

Colonne 'country'

Il s'agit de la première colonne. C'est une colonne qui ne contient aucune valeur manquante, et qui est de type 'object' (chaîne de caractères).

Contrairement à ce que son nom indique, country n'est pas forcément un nom de pays.
Il peut aussi s'agir d'une zone plus large.

Outre les noms de pays, parmi ces zones, nous pouvons également trouver :

- le monde entier : World
- des continents : Africa, Asia, Europe, North America, South America, Oceania
- les mêmes continents avec le suffixe (GCP) : cela signifie que l'on prend les données de l'organisation scientifique Global Carbon Project qui peuvent inclure des sources de données différentes et des ajustements méthodologiques, par exemple en incluant des sources d'émissions supplémentaires.
- des unions de plusieurs pays : European Union (de différentes tailles selon l'époque, par exemple Europe des 27, Europe des 28), les pays de l'OCDE (Organisation de coopération et de développement économiques) ou non-OCDE, afin de voir la différence des tendances entre les pays développés et pays émergents.
- Dans la même idée, des zones correspondant à des taux de développement ou de richesses des pays :'High-income countries', 'Low-income countries'...
- des sources particulières d'émissions : International aviation / International shipping (transport maritime) / International transport (tout type de transport), ou encore Kuwaiti Oil Fires qui est un évènement exceptionnel (incendies de puits de pétrole au Koweït) ayant eu un impact climatique significatif.

Vous trouverez en annexe les différentes valeurs qu'il peut y avoir dans cette colonne, il y a 255 valeurs distinctes.

L'une des difficultés sera de distinguer les zones que nous voulons réellement étudier.

Le ‘country’ correspondant au monde entier sera étudié en premier ; nous voulons avant tout étudier l’impact des émissions de gaz à effet de serre à l’échelle mondiale sur le réchauffement climatique.
Si besoin, nous pourrons aller un peu plus dans le détail, en faisant une analyse par continents par exemple, ou par richesses de pays.

En revanche, pour les événements exceptionnels, il ne nous paraît pas pertinent de les étudier dans le cadre de ce projet. En effet, il ne faut pas oublier que le projet porte essentiellement sur l’étude des évolutions des températures et du dérèglement climatique, et non pas sur les émissions de gaz à effet de serre dans le détail. Les données sur les émissions de gaz nous serviront à expliquer l’un des facteurs du dérèglement, mais il ne sera pas nécessaire d’aller à ce point dans le détail.

Colonne ‘year’

Cette colonne correspond à l’année de l’observation. Elle est de type ‘int64’ et ne possède aucune valeur manquante.

Elle s’étend sur une large période allant de 1750 à 2023.
Compte tenu de l’étendue de la période observée, nous manquons de données pour un grand nombre de colonnes sur une large période.
Par exemple, nous verrons plus loin que certaines colonnes n’ont des données qu’à partir de 1850, d’autres à partir de 1965.

Colonne ‘iso_code’

Il s’agit du code iso 3166-1 d’un pays, composé de 3 lettres. Cette colonne est de type ‘object’ et possède 15,8% de valeurs manquantes.
Cela est logique par rapport au fait que la colonne ‘country’ ne contient pas que des noms de pays. En effet, la colonne ‘ico_code’ n’est remplie que lorsque ‘country’ correspond véritablement à un pays, puisqu’elle sert à identifier le pays. Ainsi, il nous sera facile d’identifier et de séparer les lignes pour lesquelles le ‘country’ n’est pas un pays : ces lignes auront un iso_code manquant (valeur ‘nan’).

A partir de maintenant, toutes les colonnes qui suivront seront de type ‘float64’ et contiendront des valeurs manquantes (NaN).

Colonne ‘population’

Il s’agit du nombre d’habitants pour la zone observée, à la date correspondante.
Il peut paraître surprenant au premier abord que cette colonne soit de type float64. En effet, il s’agit d’un nombre d’habitants ; nous ne pouvons pas compter une partie décimale d’un habitant (il ne peut pas y avoir 0.3 habitant, ce serait très étrange).
Cela s’explique par le fait que cette colonne contient des NaN, et que NaN est de type float.

Nous expliquerons le traitement des NaN plus tard dans le rapport ; nous convertirons la colonne ‘population’ en ‘int’ après avoir remplacé les NaN.
Cela a peu d’importance puisque dans tous les cas, la colonne reste numérique, mais c’est surtout pour une question de logique.

Pour le country ‘world’, nous n’avons des données que tous les 10 ans entre 1750 et 1950. Pour d’autres valeurs de country comme les continents, nous avons des données tous les 10 ans jusqu’en 1800.

Colonne ‘gdp’

Il s’agit du Produit Intérieur Brut basé sur la valeur de 2011 du dollar international.

Colonnes liées aux valeurs absolues des émissions de co2

Nous allons ici décrire un ensemble de colonnes suivant la même logique.

Il s’agit de la quantité annuelle d’émissions de CO2 selon certains critères, exprimée en million de tonnes. Voici les colonnes concernées :
co2 : émission totale de CO2, excluant les émissions causées par les activités humaines liées à l’utilisation des terres (land-use change). Nous pouvons établir la relation mathématique suivante, basée sur les colonnes décrites ci-dessous :
co2 = cement_co2 + coal_co2 + flaring_co2 + gas_co2 + oil_co2 + other_industry_co2

land_use_change_co2 : émissions causées par l’utilisation des terres (changement d’affectation des terres, UTCATF en français et LUC en anglais). Les modifications de terres liées aux activités humaines sont par exemple causées par : l’agriculture, la déforestation, la reforestation. Cela peut être négatif si, par exemple, grâce à la reforestation, la zone a absorbé plus de CO2 qu’elle n’en a émis.

co2_including_luc : émission totale de CO2 incluant les émissions causées par les activités humaines liées au changement d’affectation des terres. Concrètement, cette colonne peut être calculée comme suit : *co2_including_luc = co2 + land_use_change_co2*

cement_co2 : émission de CO2 liée à la fabrication du ciment.

coal_co2 : émission de CO2 causé par l’exploitation du charbon.

flaring_co2 : émission de CO2 causée par le torchage de gaz naturel. Il s’agit de la combustion volontaire de gaz qu’une société ne veut pas stocker ni transporter (pour des réductions de coût), ce qui crée un gaspillage de gaz et émet du CO2. Cela arrive fréquemment avec le gaz naturel qui s’échappe lors de l’extraction du pétrole par exemple.

gas_co2 : émission de CO2 causée par la combustion de gaz naturel (pour la production d’électricité, l’industrie, le chauffage domestique, certains transports).

oil_co2 : émission de CO2 causée par l'exploitation du pétrole.

other_industry_co2 : émission de CO2 causée par les autres industries non mentionnées dans les autres colonnes. Elle inclut la métallurgie, la chimie, le verre, le textile et d'autres secteurs industriels.

consumption_co2 : émission de CO2 relative à la consommation de la zone étudiée. Contrairement aux émissions de production, qui comptabilisent ce qu'un pays émet en produisant de l'énergie et des biens, **consumption_co2** attribue les émissions aux consommateurs finaux. Par exemple, un pays A produit (et émet du CO2 en produisant) un produit qui sera exporté et consommé dans un pays B ; les émissions de production sont réattribuées au pays consommateur. Nous n'avons des valeurs que de 1990 à 2014 pour cette colonne.

trade_co2 : émission de co2 relative aux échanges internationaux (import export). Comme vu précédemment, un pays peut être consommateur de nombreux produits provenant d'un autre pays qui a émis du co2 en le produisant. trade_co2 est la différence entre consumption_co2 (la quantité de CO2 émise par la consommation du pays) et co2 (la quantité de co2 produite par le pays).
Une valeur positive signifie que le pays consomme davantage de produits générant des émissions de co2 venant d'autres pays, qu'il n'en produit lui-même. Une valeur négative signifie que le pays consomme moins de produits générant des émissions de co2, que ce qu'il produit lui-même.

On peut établir la relation suivante :

trade_co2 = consumption_co2 - co2

Nous n'avons des valeurs que de 1990 à 2014 pour cette colonne.

Colonnes liées aux valeurs absolues des émissions d'autres gaz à effet de serre

methane : émission de méthane (CH4), incluant celles issues de l'utilisation des terres, exprimée en million de tonnes équivalent CO2 sur une période de 100 années. On considère que sur une période de 100 ans, une tonne de méthane correspond à entre 27 et 30 tonnes de CO2, selon le rapport AR6 du GIEC.

nitrous_oxide : émission de protoxyde d'azote (oxyde nitreux, N2O), incluant les émissions issues de l'utilisation des terres, exprimée en million de tonnes équivalent CO2 sur une période de 100 années. Sur 100 ans, on estime qu'une tonne de protoxyde d'azote équivaut environ à 265 tonnes de CO2, concernant l'impact sur le réchauffement climatique.

total_ghg : émission totale de gaz à effet de serre incluant celles issues des utilisations de terres, ce qui inclut le CO2, le méthane, le protoxyde d'azote, et d'autres gaz non détaillés dans le jeu de données. Exprimée en million de tonnes équivalent CO2 sur 100 ans.

total_ghg_excluding_lucf : même chose que total_ghg mais en excluant cette fois les émissions issues des modifications de terres liées aux activités humaines.

Émissions des gaz par habitants

Les colonnes suivantes suivent la même logique de calcul : elles se basent sur une colonne correspondant à une émission de gaz, divisée par la colonne 'population'. Ainsi, il s'agit de la quantité d'émissions de gaz par habitant pour la zone observée, mesurée cette fois en tonnes par habitant.

Voici ci-dessous les colonnes concernées, leur nom étant pour la plupart clair, nous ne détaillerons que la première ou celles dont le nom est moins clair.

co2_per_capita : quantité totale de co2, excluant les modifications de terres (land-use change), par habitant, exprimée en tonnes par habitant. Correspond à la formule :
*co2_per_capita = co2 / population * 1 000 000*

cement_co2_per_capita
co2_including_luc_per_capita
coal_co2_per_capita
consumption_co2_per_capita
flaring_co2_per_capita
gas_co2_per_capita
ghg_per_capita = *total_ghg / population * 1 000 000*
ghg_excluding_lucf_per_capita = *total_ghg_excluding_lucf / population * 1 000 000*
land_use_change_co2_per_capita
methane_per_capita
nitrous_oxide_per_capita
oil_co2_per_capita
other_co2_per_capita = *other_industry_co2 / population * 1 000 000*

Émissions des gaz par PIB

Ces colonnes suivent le même principe que les colonnes ayant le suffixe "per_capita", sauf que cette fois, c'est par PIB (gdp).
Ces colonnes sont exprimées en kilogrammes par dollar du PIB (basée sur la valeur dollar international 2011).

Elles sont calculées comme suit :

co2_per_gdp = *co2 / gdp * 1 000 000 000*
co2_including_luc_per_gdp = *co2_including_luc / gdp * 1 000 000 000*
consumption_co2_per_gdp = *consumption_co2 / gdp * 1 000 000 000*

Croissance des émissions de CO2

Concernant les colonnes ‘co2’ et ‘co2_including_luc’, nous avons deux autres colonnes qui y sont liées et qui montrent la croissance des émissions.

co2_growth_abs : croissance absolue des émissions de co2 par rapport à l’année précédente, en million de tonnes. Cela correspond à la différence entre les émissions de cette année et celles de l’année précédente, pour le même ‘country’. Cela se traduit mathématiquement par la formule :

$$co2\_growth\_abs(n) = co2(n) - co2(n-1)$$

où n signifie l’année de l’observation, et n-1 l’année précédente.

co2_growth_prct : croissance en pourcentage par rapport à l’année précédente.
Cela se traduit mathématiquement par la formule :

$$co2\_growth\_prct\ (n) = (co2(n) - co2(n-1)) / co2(n-1) * 100$$

Nous retrouvons exactement la même logique avec la colonne co2_including_luc pour les colonnes suivantes :
co2_including_luc_growth_abs
co2_including_luc_growth_prct

Colonnes d’émissions cumulées

Les colonnes suivantes calculent le cumul des émissions de co2 pour chacune des colonnes d’émission correspondantes. Nous mettrons la formule mathématique pour la première colonne uniquement, la logique est la même pour toutes les autres. Ces colonnes s’expriment en million de tonnes.
cumulative_co2 : calculé par la formule suivante :

$$cumulative\_co2(n) = cumulative\_co2(n-1) + co2(n)$$

où n est l’année de l’observation, et n-1 l’année précédente.

La même logique s’applique pour les colonnes ci-dessous :
cumulative_cement_co2
cumulative_co2_including_luc
cumulative_coal_co2
cumulative_flaring_co2
cumulative_gas_co2
cumulative_luc_co2
cumulative_oil_co2
cumulative_other_co2 : Pour celle-ci, on se base sur la colonne other_industry_co2.

Colonnes liées à la consommation d'énergie

primary_energy_consumption : consommation d'énergie primaire (énergie extraite de la nature, avant transformation en électricité, essence, chaleur…), mesurée en terawatt-heures.

Les deux colonnes suivantes représentent la consommation d'énergie par habitant et par PIB, en suivant les mêmes logiques de calcul que précédemment.

energy_per_capita : exprimée en kilowatt-heures par habitant.
energy_per_gdp : exprimée en kilowatt-heures par dollar international (2011).

Enfin, nous avons également deux colonnes qui représentent les émissions de co2 (et co_including_luc) par unité d'énergie, exprimées en kilogrammes par kilowatt-heures.

co2_per_unit_energy : la formule mathématique est la suivante :

co2_per_unit_energy = co2 / primary_energy_consumption

co2_including_luc_per_unit_energy : reprend la même logique en divisant co2_including_luc par primary_energy_consumption.

Changements de température par gaz à effet de serre

En plus des émissions de gaz, nous avons quatre colonnes qui représentent, en °C, les changements de température mesurés à la surface, pour chaque type de gaz à effet de serre. Nous avons ainsi une colonne pour le dioxyde de carbone (CO2), une colonne pour le méthane (CH4), une colonne pour le protoxyde d'azote (N2O), et enfin, une colonne pour l'ensemble des gaz à effet de serre, incluant les trois précédents :

temperature_change_from_ch4
temperature_change_from_co2
temperature_change_from_n2o
temperature_change_from_ghg

Colonnes de pourcentage d'émissions de gaz par rapport à l'émission mondiale

Enfin, nous allons expliquer le dernier groupe de colonnes.
Ces colonnes ont toutes le mot "share" (en préfixe de la colonne, sauf pour le trade_co2 où c'est en suffixe).
Elles s'expriment toutes en pourcentage.
Ces colonnes représentent, pour chacune des valeurs représentées (émissions de gaz, cumul des émissions d'un gaz, ou encore changement de température), la proportion représentée par la zone observée (country) par rapport au reste du monde, pour l'année de l'observation.
Par exemple, si le country est France et l'année est 2020, ces colonnes représentent la proportion des émissions (ou du changement de température causé par les gaz à effet de serre) de la France par rapport au reste du monde sur l'année 2020.

Ces colonnes étant nombreuses, et leurs noms explicites, nous n'allons pas les détailler individuellement puisqu'elles suivent toutes le même principe. Nous allons simplement les lister ci-dessous :

share_global_cement_co2
share_global_co2
share_global_co2_including_luc
share_global_coal_co2
share_global_cumulative_cement_co2
share_global_cumulative_co2
share_global_cumulative_co2_including_luc
share_global_cumulative_coal_co2
share_global_cumulative_flaring_co2
share_global_cumulative_gas_co2
share_global_cumulative_luc_co2
share_global_cumulative_oil_co2
share_global_cumulative_other_co2
share_global_flaring_co2
share_global_gas_co2
share_global_luc_co2
share_global_oil_co2
share_global_other_co2
trade_co2_share
share_of_temperature_change_from_ghg

Etant donné que ces pourcentages se font par rapport aux émissions globales mondiales, notons que lorsque le country est 'world', ces données ne sont pas remplies.

Nettoyage du jeu de données

Notre nettoyage va essentiellement s'articuler autour de deux principaux objectifs :

- Gérer les valeurs manquantes (NaN) qui sont très nombreuses
- Diviser le csv en plusieurs csv plus concis, plus ciblés et donc plus simples à analyser.

En effet, à travers l'étude des colonnes, nous avons pu voir que le csv originel est un peu un "fourre-tout". Nous avons énormément de données sur énormément de choses différentes qui ne peuvent pas être toujours comparées entre elles. A cause de la longue période étudiée, du grand nombre de données diverses, et de la multitude de zones (country) différentes (pays, alliance de pays, autres…), nous avons aussi une très grande proportion de valeurs manquantes.
Il sera donc très intéressant de diviser ce csv en plusieurs jeux de données plus ciblés qui ne contiendront plus de valeurs manquantes, quitte à réduire la période de temps étudiée.

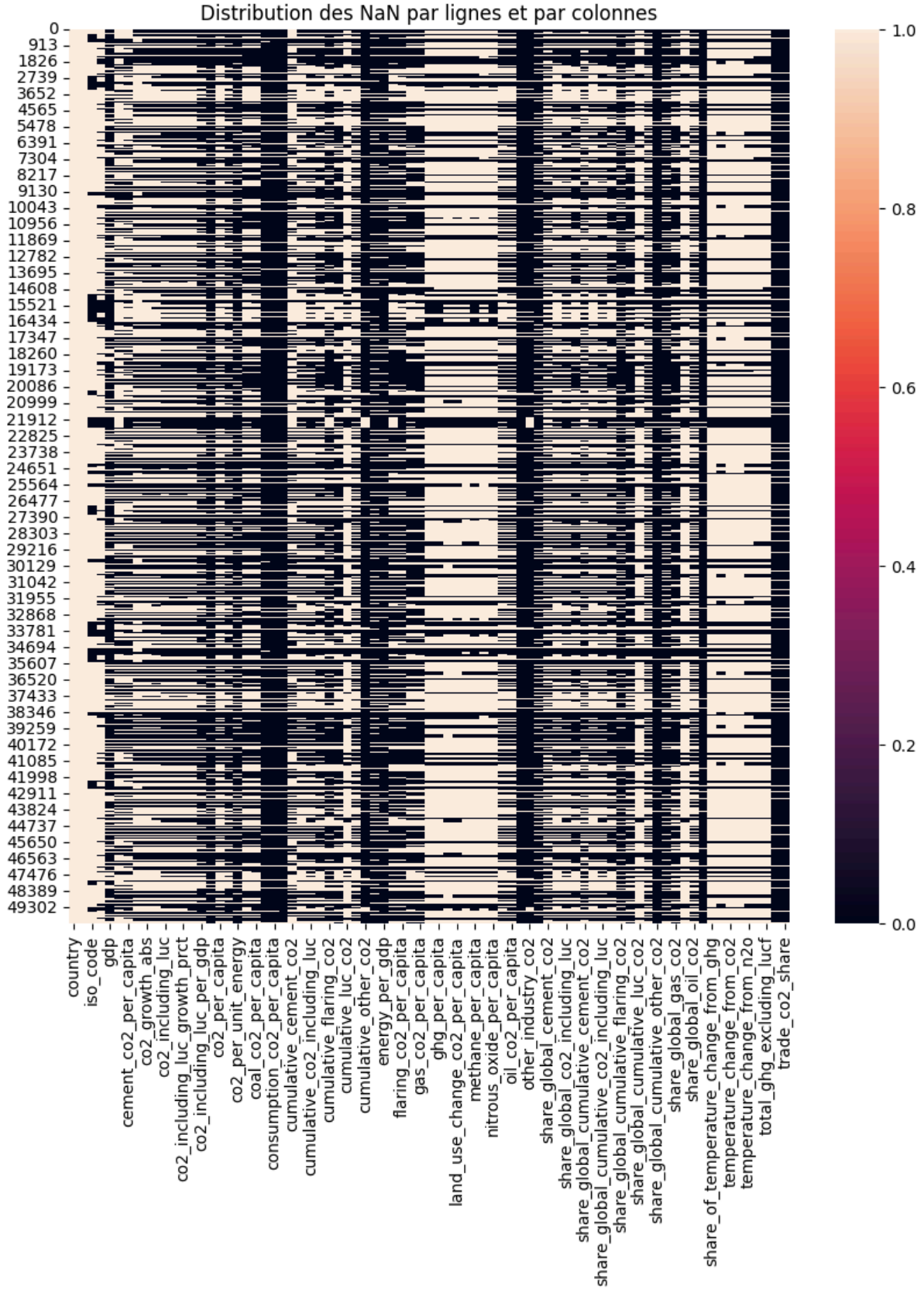
Répartition des valeurs manquantes (NaN)

Voici une première analyse des valeurs manquantes, qui indique la proportion de valeurs manquantes pour chaque colonne, triée par ordre décroissant.
Par souci de clarté du rapport, nous n'allons présenter que les 10 colonnes ayant la plus grande proportion de NaN, mais l'intégralité de l'analyse est trouvable en annexe.

share_global_cumulative_other_co2 : 95.8%
share_global_other_co2 : 95.8%
other_co2_per_capita : 95.07%
cumulative_other_co2 : 93.62%
other_industry_co2 : 93.62%
consumption_co2_per_gdp : 91.15%
consumption_co2_per_capita : 91.03%
trade_co2 : 90.96%
trade_co2_share : 90.96%
consumption_co2 : 90.31%

Nous constatons que les 10 colonnes ayant le plus de valeurs manquantes ont plus de 90% de valeurs manquantes, ce qui est énorme.

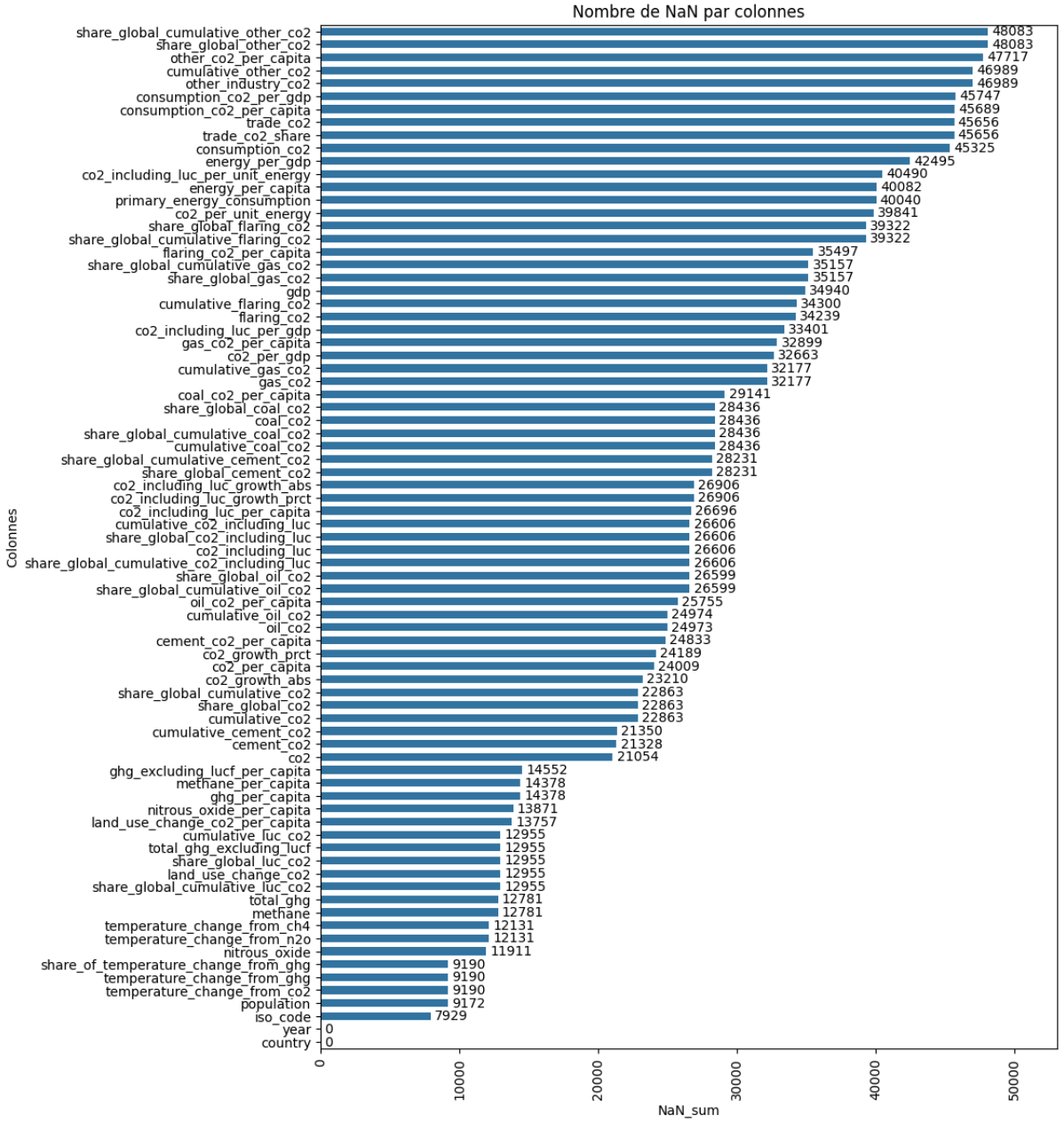
Nous pouvons visualiser la répartition des valeurs manquantes par colonnes au travers de la heatmap ci-dessous. Les parties noires indiquent les valeurs manquantes. Sur l'axe des abscisses nous avons nos 79 colonnes, et sur l'axe des ordonnées, toutes les lignes du dataset.



Il y a plusieurs choses intéressantes à retirer de cette visualisation.

Certaines colonnes semblent être groupées entre elles, dans le sens où si une colonne a une valeur manquante, les autres colonnes du même groupe en ont également, et vice-versa. Autrement dit, ces colonnes suivent la même logique de remplissage ; les données proviennent certainement des mêmes sources.

Nous pouvons également visualiser le nombre de valeurs manquantes au travers d'un diagramme à barres.



Ce diagramme nous montre, de façon plus visuelle, la répartition des NaN. Surtout, nous avons ici la valeur exacte du nombre de valeurs manquantes par colonnes, ce qui nous permet une analyse plus précise : certaines colonnes ont exactement le même nombre de

valeurs manquantes. Ce fait, couplé à la visualisation de la heatmap, nous conforte dans l'idée que certaines colonnes sont étroitement liées.

Afin de faciliter le découpage du dataframe en plusieurs autres dataframe, ainsi que le nettoyage des valeurs manquantes, il serait intéressant d'identifier ces groupes de colonnes.

Nous définirons ici "groupe de colonnes" de la sorte : un groupe de colonnes est constitué de plusieurs colonnes qui, pour chaque ligne, ont toutes des NaN, ou bien n'ont aucun NaN. Ainsi, ces colonnes sont liées entre elles par le fait qu'elles sont toutes renseignées pour les mêmes lignes.

De par leur définition, les colonnes d'un même groupe ont forcément le même nombre de valeurs manquantes. Toutefois, deux colonnes ayant exactement le même nombre de valeurs manquantes ne font pas forcément partie du même groupe : en effet, deux colonnes peuvent avoir le même nombre de NaN, mais répartis sur des lignes différentes.

Pour pouvoir identifier nos groupes de colonnes, nous rassemblons dans un premier temps celles qui ont le même nombre de NaN ; puis nous vérifions, au sein de chacun de ces groupes, si chaque ligne suit la règle établie : n'avoir que des NaN, ou n'avoir aucun NaN.

Ce travail nous permet d'établir les groupes de colonnes suivants :

Voici donc les groupes ayant des colonnes cohérentes entre elles, à savoir que si l'une est remplie, les autres du même groupe le sont également :

['year', 'country'] : aucun NaN

['share_of_temperature_change_from_ghg', 'temperature_change_from_ghg', 'temperature_change_from_co2'] : 9190 NaN

['temperature_change_from_ch4', 'temperature_change_from_n2o'] : 12131 NaN

['total_ghg', 'methane'] : 12781 NaN

['cumulative_luc_co2', 'share_global_luc_co2', 'land_use_change_co2', 'share_global_cumulative_luc_co2'] : 12955 NaN

['methane_per_capita', 'ghg_per_capita'] : 14383 NaN

['share_global_cumulative_co2', 'share_global_co2', 'cumulative_co2'] : 22863 NaN

['share_global_oil_co2', 'share_global_cumulative_oil_co2'] : 26599 NaN

['cumulative_co2_including_luc', 'share_global_co2_including_luc', 'co2_including_luc', 'share_global_cumulative_co2_including_luc'] : 26606 NaN

['co2_including_luc_growth_abs', 'co2_including_luc_growth_prct'] : 26906 NaN

['share_global_cumulative_cement_co2', 'share_global_cement_co2'] : 28231 NaN

['share_global_coal_co2', 'coal_co2', 'share_global_cumulative_coal_co2', 'cumulative_coal_co2'] : 28436 NaN

['cumulative_gas_co2', 'gas_co2'] : 32177 NaN

['share_global_cumulative_gas_co2', 'share_global_gas_co2'] : 35157 NaN

['share_global_flaring_co2', 'share_global_cumulative_flaring_co2'] : 39322 NaN

['trade_co2', 'trade_co2_share'] : 45656 NaN

['cumulative_other_co2', 'other_industry_co2'] : 46989 NaN

['share_global_cumulative_other_co2', 'share_global_other_co2'] : 48083 NaN

Notons que ce travail nous a permis de constater que la colonne ‘total_ghg_excluding_lucf’, qui possède aussi 12955 valeurs manquantes, n’est pas incluse dans le groupe des 12955, car ses valeurs manquantes ne sont pas toujours sur les mêmes lignes que les autres colonnes du groupe.

Ces groupes de colonnes liées entre elles pourraient nous donner un aperçu des différents dataframes que nous pourrions créer à partir du dataframe original. Toutefois, un groupe de colonnes ne correspond pas forcément à un dataframe cohérent.
Voici plusieurs exemples :

- nous avons des colonnes “temperature_change” dans des groupes de colonnes différents. Pourtant, il serait logique que ces colonnes soient toutes dans le même dataframe (puisqu’elles montrent toutes un changement de température causé par un GES (gaz à effet de serre) en °C).
- il serait plus logique que des colonnes comme ‘methane’ et ‘methane_per_capita’ soient dans le même dataframe, même si elles n’ont pas le même nombre de NaN, parce que ces colonnes sont étroitement liées : la deuxième peut être calculée par rapport à la première et à la colonne ‘population’.

Cela pourra toutefois nous aider à nettoyer les colonnes et à savoir lesquelles conserver pour nos dataframes, nous pourrons nous y référer plus tard, mais nous ne pouvons pas nous baser uniquement là-dessus pour diviser notre dataframe original en plusieurs dataframe plus ciblés. Pour cela, nous allons plutôt privilégier une approche par zone, c’est-à-dire par ‘country’, comme nous allons le voir dans la section suivante.

Séparation des dataframe par country et iso_code

Nous l'avions évoqué plus en amont : nous avons 255 valeurs distinctes de country, et 219 valeurs distinctes de iso_code.

La colonne 'country' est composée de nombreuses informations qui ne sont pas forcément des pays.

Une première séparation logique est de séparer le dataframe en deux dataframes : l'un contenant les véritables pays, et l'autre contenant les autres valeurs de 'country'.

Notre supposition est que les entrées pour lesquelles 'country' n'est pas un véritable pays ont un iso_code null. Pour vérifier cela, nous récupérons toutes les lignes qui ont un iso_code null, et nous en affichons les valeurs distinctes de 'country', ce qui nous donne la liste que vous trouverez en annexe "Valeurs de 'country' n'ayant pas d'iso-code".

A partir de ces valeurs, nous avons décidé de créer plusieurs dataframes :

- **owid-co2-world** : un dataframe contenant toutes les lignes pour lesquelles le country est 'world' : en effet, notre étude porte sur les températures terrestres, à l'échelle mondiale. Le monde sera donc notre premier et principal axe d'analyse avant d'aller plus dans le détail. Nous verrons par la suite que ce dataframe sera lui-même subdivisé en plusieurs dataframes plus ciblés et plus proches de certains groupes de colonnes vus précédemment.
- **owid-co2-continent** : un dataframe contenant les lignes liées à des continents. Nous avons constaté que les continents apparaissent plusieurs fois, parfois avec la mention (GCP). Nous verrons plus tard comment nous avons traité cela.

Nous pourrions également créer au moins deux autres dataframes plus ciblés :
- **owid-co2-international-shipping-aviation** : contient toutes les lignes dont country vaut "international shipping" ou "international aviation", afin de voir la part d'émissions de GES dus à l'aviation et au transport maritime. A vrai dire, nous l'avons créé par acquis de conscience, mais nous ne pensons pas qu'il nous sera utile.
- **owid-co2-international-transport** : même chose que précédemment mais pour tout type de transport. A noter que les données ne sont pas remplies sur les mêmes périodes que pour shipping et aviation, d'où le fait d'en faire un à part. Il est peu probable que nous l'utilisions.
- **owid-co2-per-development** : contient toutes les lignes où la valeur de country correspond à un ensemble de pays par richesse ('high-income countries', 'low-income countries', etc…). Cela pourra nous être utile si nous souhaitons faire une analyse des émissions de GES pour voir la responsabilité des pays causant le plus le dérèglement climatique par rapport aux régions du monde qui en souffrent le plus.

Enfin, nous avons le **owid-co2-iso-code**, qui contient les données du dataframe original pour lesquelles la colonne iso-code est remplie, ce qui correspond donc à de vrais pays.

Nous ne l'utiliserons que si nous voulons réellement cibler certains pays pour voir leur empreinte carbone.

Note : essentiellement par manque de temps, nous n'avons pas pu nettoyer tous les jeux de données présentés ci-dessus. Nous nous sommes focalisés sur les dataset correspondant au monde entier et à celui des continents, pour le moment, mais n'avons pas terminé le nettoyage de celui sur les continents au moment de la rédaction du rapport, il nous restera à appliquer les mêmes transformations que pour celui concernant le monde entier en ce qui concerne le traitement des valeurs manquantes.

Nettoyage du jeu de données 'owid-co2-world'

Dans cette partie, nous allons parler des transformations appliquées au jeu de données 'World', créé à partir des lignes pour lesquelles la valeur de la colonne 'country' vaut 'world'.

Nous verrons que nous avons subdivisé ce dataset par plusieurs autres dataset plus ciblés au niveau des données mesurées.

Puisque cela ne concerne que le country 'world', les colonnes iso_code (toujours vide) et country (une seule valeur) ne sont plus pertinentes, nous les avons donc supprimées.

Nous avons ensuite repris l'étude des valeurs manquantes pour ce dataset.

Nettoyage de la population

Nous avons constaté qu'entre 1750 et 1950, la population n'avait des valeurs que tous les 10 ans, et nous avions des NaN entre ces années.

	year	population
0	1750	753279296.0
1	1751	NaN
2	1752	NaN
3	1753	NaN
4	1754	NaN
5	1755	NaN
6	1756	NaN
7	1757	NaN
8	1758	NaN
9	1759	NaN
10	1760	788254976.0
11	1761	NaN

Afin de remplacer ces NaN par des valeurs cohérentes, nous avons décidé d'utiliser une interpolation quadratique, avec la méthode interpolate en utilisant la méthode 'spline' d'ordre 2.
Cette interpolation permet de remplir les valeurs manquantes avec des valeurs situées entre deux intervalles remplis, en suivant une progression en forme de courbe très proche de la réalité (plus proche qu'une interpolation linéaire qui ferait simplement une augmentation proportionnelle). Nous avons ensuite converti la colonne 'population' en int, puisqu'il n'y a pas de raisons que ce soit un float, maintenant qu'il n'y a plus de NaN.

Voici un exemple du remplissage pour les premières valeurs :

	year	population
0	1750	753279296
1	1751	756867523
2	1752	760435604
3	1753	763983539
4	1754	767511327
5	1755	771018968
6	1756	774506463
7	1757	777973811
8	1758	781421012
9	1759	784848067
10	1760	788254976
11	1761	791641737

On constate bien que les lignes qui étaient remplies sur l'image précédente (1750 et 1760) ont toujours les mêmes valeurs, et que les NaN ont été remplies par des valeurs qui suivent une progression cohérente entre les deux intervalles.

Nettoyage des colonnes 'per_capita'

Puisque les colonnes ayant pour suffixe 'per_capita' se basent sur la population, il y avait également des NaN pour toutes les lignes où la population était vide.

Etant donné que nous connaissons la formule mathématique permettant de calculer ces colonnes à partir de la colonne de base et la population, nous avons pu facilement remplacer les valeurs manquantes en refaisant les calculs (qui sont présentés dans la partie précédente sur l'explication des colonnes).

Nettoyage de la colonne 'gdp'

La colonne 'gdp' suivait le même principe que 'population', à savoir, des valeurs manquantes entre des intervalles. La différence était que les valeurs remplies n'étaient pas forcément tous les 10 ans. La première valeur apparaît pour l'année 1820, puis nous avons des données tous les 20 ou 30 ans jusqu'à 1940, puis tous les 10 ans jusqu'à 2010. A partir de 2015, nous avons des données chaque année.

	year	population	gdp
70	1820	1065623616	1.175114e+12
100	1850	1287033856	1.546684e+12
120	1870	1346763136	1.963043e+12
150	1900	1670635648	3.503708e+12
170	1920	1927857152	4.824949e+12
190	1940	2328460032	7.646890e+12
200	1950	2493092843	8.461552e+12
210	1960	3015470890	1.333808e+13
220	1970	3694683801	2.194193e+13
230	1980	4447606208	3.198261e+13
240	1990	5327803075	4.303361e+13
250	2000	6171703018	5.989767e+13
260	2010	7021732131	8.980771e+13
265	2015	7470491904	1.068718e+14
266	2016	7558554580	1.104072e+14

Pour combler les trous, nous avons utilisé la même méthode d'interpolation quadratique que pour la population, afin d'avoir des valeurs cohérentes et une courbe lisse.

Nous avons rempli les valeurs manquantes des premières années par des 0. Notons que cela ne concerne que les années avant 1820. Nous n'avons pas vraiment d'intérêt à faire des analyses poussées sur la période allant de 1750 à 1820, voilà pourquoi nous remplaçons les NaN par des 0 sans chercher à faire quelque chose de plus cohérent.

Note : de la même manière que nous avons nettoyé les colonnes 'per_capita' en refaisant nous-même le calcul, nous avons pu le faire pour les colonnes 'per_gdp' en effectuant une division par la colonne 'gdp'. Vous retrouverez les formules de calcul correspondantes dans la partie précédente sur la description des colonnes.

Division par années d'apparition des données

Pour les autres colonnes qui avaient des valeurs manquantes, nous avons regardé à partir de quelle année nous avions des données. Ces informations, couplées aux groupes de colonnes présentés plus tôt dans le rapport, nous ont permis de faire une subdivision du jeu de données owid-co2-world en plusieurs autres jeux de données.

Ainsi, nous avons constaté que les colonnes liées aux énergies avaient des informations à partir de 1965 uniquement. Il nous a donc semblé pertinent de créer un csv concernant uniquement ces données pour la période 1965-2023, et regroupant ainsi les colonnes suivantes : ['year', 'population', 'gdp', 'co2', 'co2_including_luc', 'co2_including_luc_per_unit_energy', 'co2_per_unit_energy', 'energy_per_capita', 'energy_per_gdp', 'primary_energy_consumption']
Nous avons appelé ce jeu de données nettoyé : **owid_co2_per_energy_world.csv**

Il sera sans doute intéressant de le comparer avec le dataset relatif aux énergies également fourni par l'organisme OWID, dont nous avons parlé en introduction et qui est nettoyé et analysé par une autre personne de l'équipe.

Nous avons également remarqué que beaucoup de colonnes liées aux émissions de gaz et aux changements de température avaient des données à partir de 1850. Nous avons décidé de les regrouper dans un même csv.
Pour de rares colonnes parmi celles-ci, il y avait encore des NaN après 1850, sur une période allant jusqu'à 30 ans. Nous avons décidé de remplacer ces NaN par des 0, parce que c'était déjà ce qui avait été fait de base dans le dataframe pour certaines colonnes telles que cement_co2.

Une fois nettoyé, nous avons pu créer le csv correspondant intitulé : **owid_ghg_world.csv**
Ce CSV contient toutes les informations sur les différentes émissions de gaz à effet de serre (avec sources d'émissions de co2 telles que vues dans la partie sur l'explication des colonnes), ainsi que les changements de température liés à ces émissions, sur une période allant de 1850 à 2023.

Les colonnes de ce dataset sont les suivantes :
['year', 'population', 'gdp', 'cement_co2', 'cement_co2_per_capita',
'co2', 'co2_growth_abs', 'co2_growth_prct', 'co2_including_luc',
'co2_including_luc_growth_abs', 'co2_including_luc_growth_prct',
'co2_including_luc_per_capita', 'co2_including_luc_per_gdp',
'co2_per_capita', 'co2_per_gdp', 'coal_co2', 'coal_co2_per_capita',
'cumulative_cement_co2', 'cumulative_co2',
'cumulative_co2_including_luc', 'cumulative_coal_co2',
'cumulative_gas_co2', 'cumulative_luc_co2', 'cumulative_oil_co2',
'cumulative_other_co2', 'gas_co2', 'gas_co2_per_capita',
'ghg_excluding_lucf_per_capita', 'ghg_per_capita',
'land_use_change_co2', 'land_use_change_co2_per_capita', 'methane',
'methane_per_capita', 'nitrous_oxide', 'nitrous_oxide_per_capita',
'oil_co2', 'oil_co2_per_capita', 'other_co2_per_capita',
'other_industry_co2', 'temperature_change_from_ch4',
'temperature_change_from_co2', 'temperature_change_from_ghg',
'temperature_change_from_n2o', 'total_ghg', 'total_ghg_excluding_lucf']

Nettoyage final du owid-co2-world

Maintenant que nous avons nettoyé et divisé owid-co2-world en d'autres dataset plus ciblés, nous avons choisi de ne garder que les colonnes ayant des informations pertinentes sur la période de base (1750 à 2023).
Ainsi, le jeu de données **owid-co2-world** ne contiendra que les informations d'émission de co2, par rapport à la population et au PIB, sur la période 1750 à 2023.
Les colonnes de ce dataset sont donc : ['year', 'population', 'gdp', 'co2', 'co2_growth_abs', 'co2_growth_prct', 'cumulative_co2', 'co2_per_capita', 'co2_per_gdp']

Nettoyage du jeu de données 'owid-co2-continent'

Merge des données GCP

Comme vu lors des parties précédentes traitant de la colonne 'country', pour les continents, nous avons plusieurs valeurs possibles.
Par exemple, il y a des lignes ayant la valeur "Africa" et d'autres "Africa (GCP)".

GCP est le sigle de Global Carbon Project, une organisation scientifique qui collecte et analyse les données sur les émissions de carbone à l'échelle mondiale. Leurs données sont observées et calculées différemment et peuvent donc différer par rapport aux autres observations.
Le 'country' "Africa (GCP)" n'est donc pas une zone en soi, mais une manière différente d'obtenir les données pour la zone "Africa".
En explorant les données pour lesquelles 'country' contient le sigle (GCP), nous constatons que ces données n'ont que la colonne co2 qui est remplie (en plus de 'country' et 'year' qui servent d'identification d'observations), et la colonne consumption_co2 pour la période 1990-2022 uniquement.

Il nous paraît alors pertinent de faire en sorte que ces données ne soient, non pas des lignes supplémentaires dans le dataset, mais des colonnes supplémentaires pour la zone concernée.
Ainsi, nous n'aurons pas une ligne "Africa" et une ligne "Africa (GCP)", mais plutôt une seule ligne "Africa" avec une colonne "co2" et une colonne "co2 (GCP)", et l'équivalent pour consumption_co2.

Nous avons donc, dans un premier temps, créé deux dataframes (un contenant les entrées ayant le sigle "(GCP)" et l'autre non), puis après avoir renommé les country "(GCP)" en leur penchant sans (GCP) et après avoir renommé les colonnes, nous avons mergé ces dataframes.
Notons que la plupart du temps, les valeurs de co2 et de co2 (GCP) sont identiques, mais peuvent différencier pour les continents Asie, Amérique du Nord et Océanie.

Différences des moyennes de 'co2' et 'co2 (GCP)' par 'country'

country	co2	co2 (GCP)	diff
Asia	2195.718529	1912.027588	283.690942
North America	2077.141954	2044.512590	32.629364
Oceania	80.858496	80.127157	0.731339
Africa	302.276845	302.274460	0.002385
Europe	2000.723000	2000.722982	0.000018
South America	257.489578	257.489567	0.000011

Différences des sommes de ‘co2’ et ‘co2 (GCP)’ par ‘country’

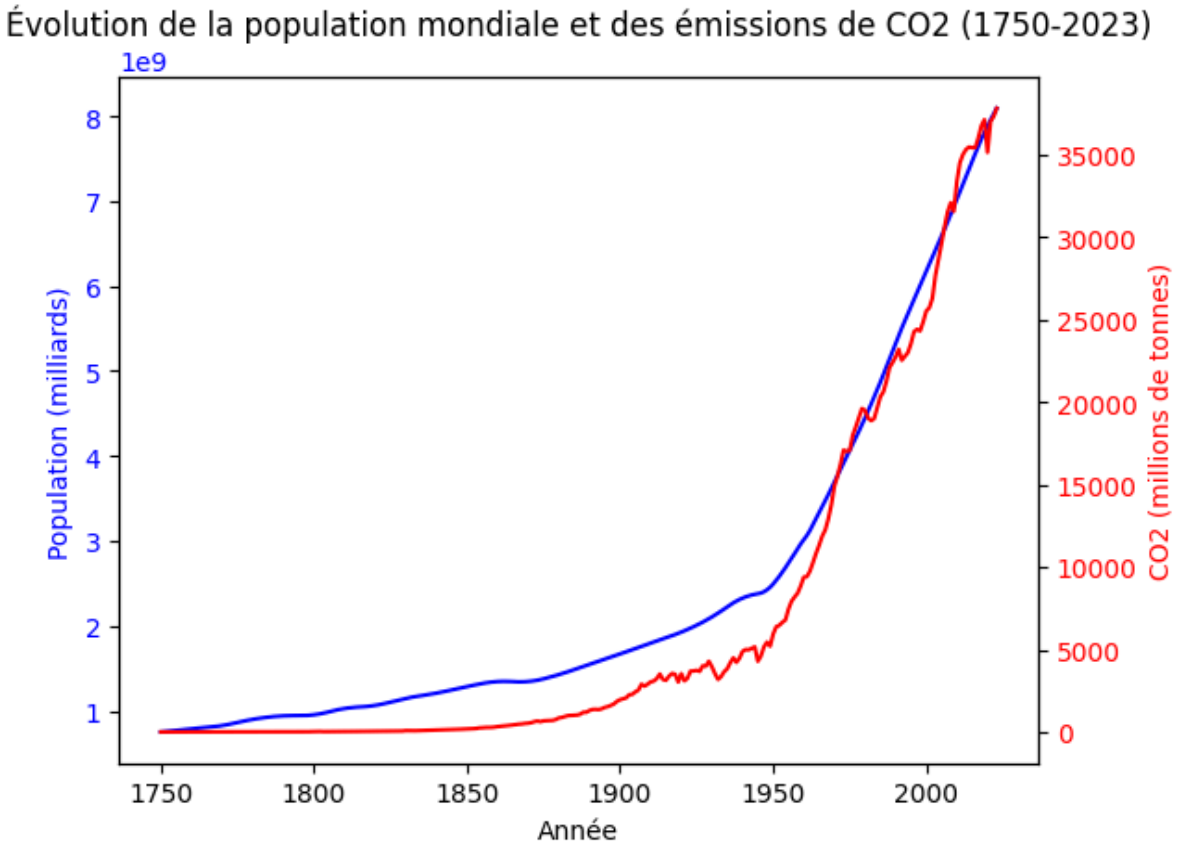
country	co2	co2 (GCP)	diff
Asia	601626.877	523895.559	77731.318
North America	496436.927	488638.509	7798.418
Oceania	22155.228	21954.841	200.387
Africa	52596.171	52595.756	0.415
Europe	548198.102	548198.097	0.005
South America	46348.124	46348.122	0.002
Asia (excl. China and India)	266225.323	266225.323	0.000
Europe (excl. EU-27)	249747.627	249747.627	0.000
European Union (28)	378228.177	378228.177	0.000
European Union (27)	298450.476	298450.476	0.000
Europe (excl. EU-28)	169969.916	169969.916	0.000
North America (excl. USA)	64584.216	64584.216	0.000

La différence est très notable pour l’Asie, qui est le continent émettant le plus de co2.
Cette différence peut s’expliquer par le fait que l’Asie contient des pays très fermés, comme la Chine par exemple, dont nous ne sommes pas sûr de la fiabilité des données.

Analyse et visualisation du jeu de données

A partir du dataset owid-co2-world

Evolution de la population et des émissions de CO2 mondiales

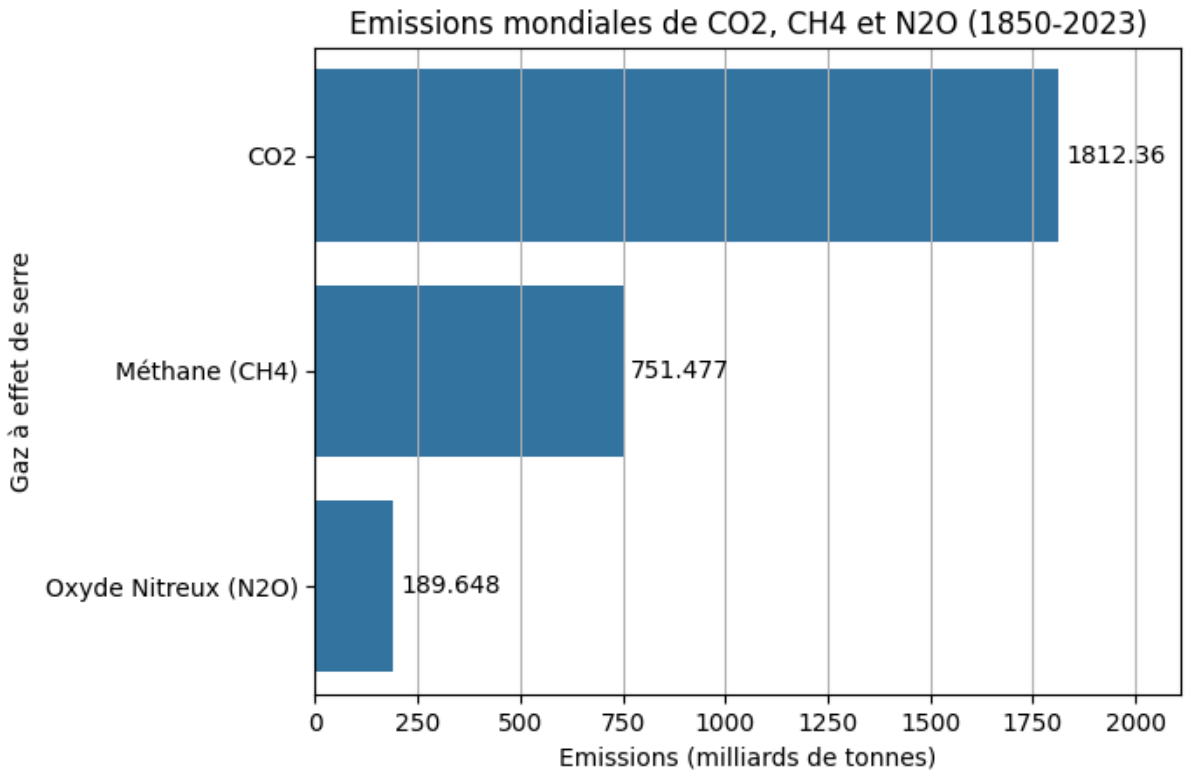


Ce graphique, créé à partir de notre dataset owid-co2-world, montre l'évolution des émissions de CO2 en million de tonnes (en rouge) ainsi que l'évolution de la population en milliards d'habitants (en bleu).
On remarque une certaine corrélation entre les courbes, qui augmentent toutes deux de façon presque exponentielle.

L'augmentation de la population et la quantité d'émissions de CO2 explosent à partir de 1950, connaissant une hausse extrêmement importante.

A partir du dataset owid-ghg-world

Cumul des émissions de gaz à effet de serre

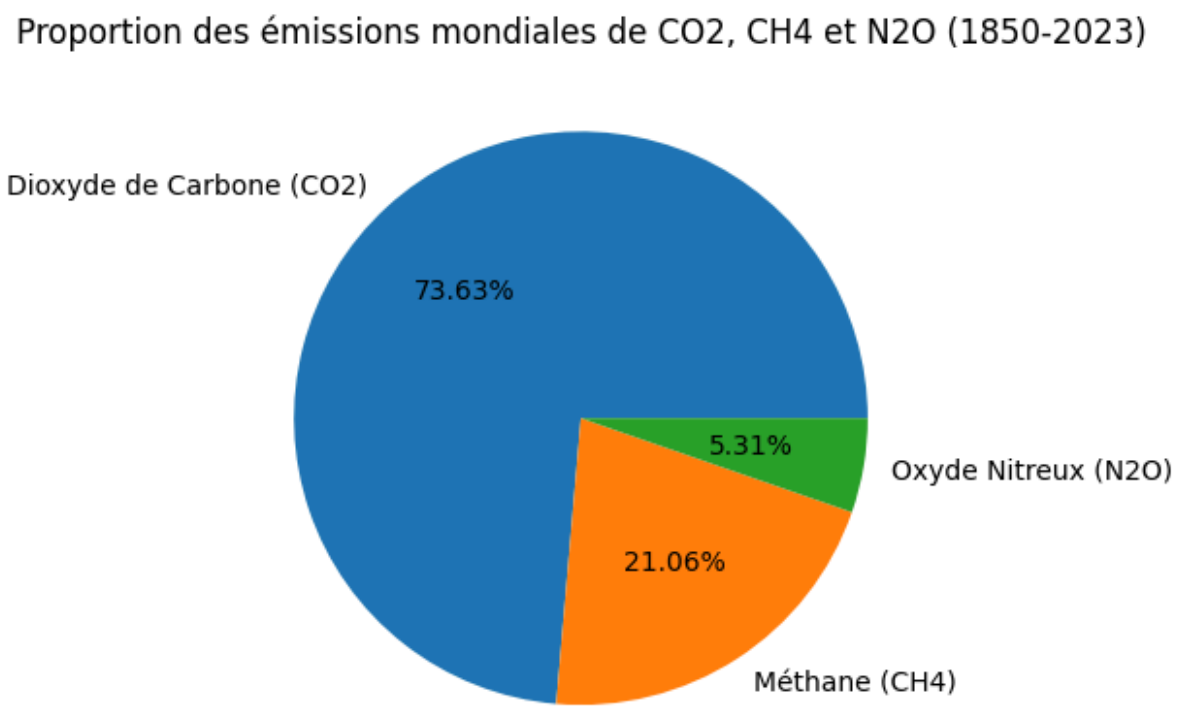


Notre dataset owid-ghg-world contient des informations détaillées sur les émissions de différents gaz à effet de serre sur la période 1850-2023.

Ce graphique représente le cumul des principaux gaz à effet de serre : dioxyde de carbone (CO2), méthane (CH4) et oxyde nitreux (N2O, aussi appelé protoxyde d'azote), dans le monde, de 1850 à 2023. L'unité est en milliards de tonnes.

Les valeurs d'émissions du méthane et de l'oxyde nitreux sont exprimées en équivalent CO2 sur une période de 100 ans.

L'on constate que les émissions de CO2 sont très largement supérieures aux émissions des deux autres gaz (plus du double des émissions de méthane). Il est donc pertinent de s'intéresser en premier lieu à comment réduire les émissions de ce gaz, puisque c'est là qu'on a le plus à gagner en matière de réduction d'émissions de gaz à effet de serre.



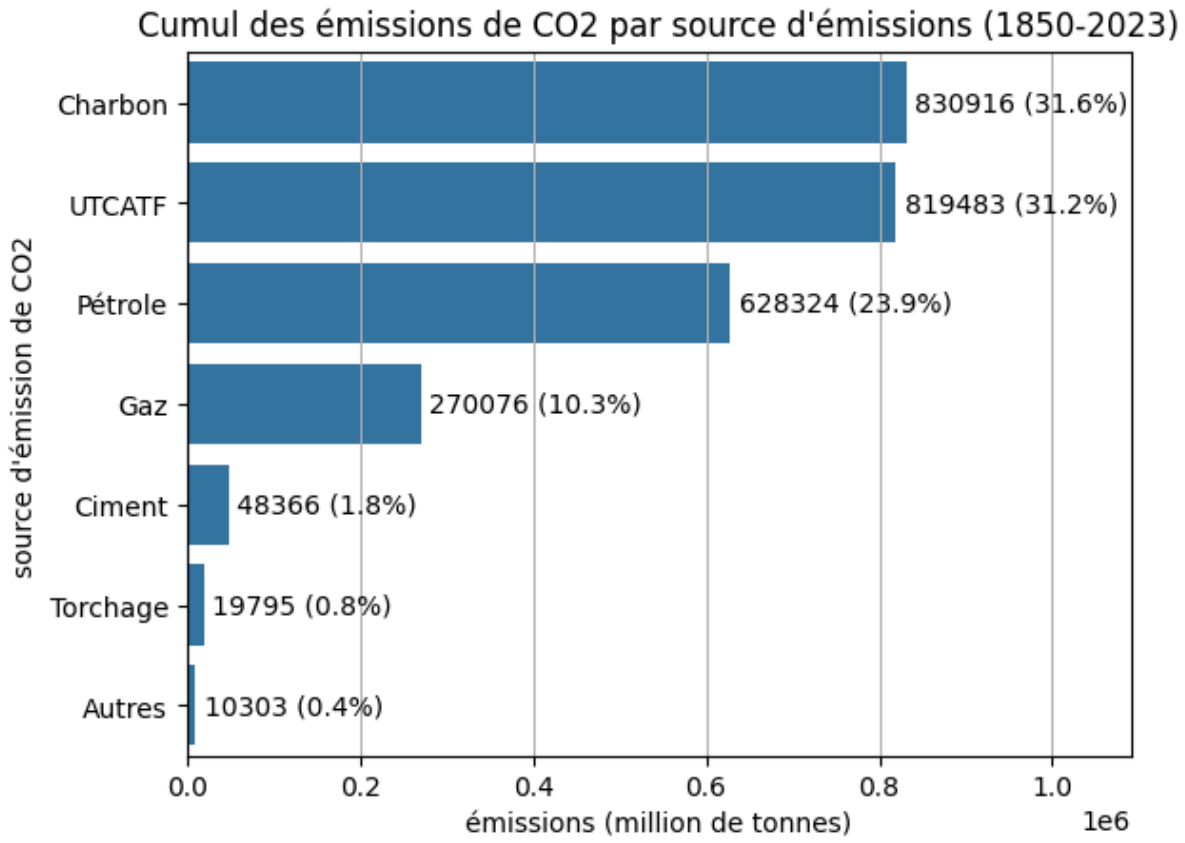
Ce graphique nous montre la proportion des émissions des gaz à effet de serre disponibles dans les données, à savoir : dioxyde de carbone, méthane et oxyde nitreux.
Il reprend les mêmes données que le graphique précédent, en faisant une représentation par pourcentage pour mieux visualiser la part d'émissions de chaque gaz.

Nous pouvons constater que les émissions de CO2 représentent à elles-seules près de trois quart des émissions, tandis que celles de méthane sont de plus d'un cinquième.

Nous aurions aimé voir également la proportion des autres gaz à effet de serre, pour lesquels les données ne sont pas présentes dans notre jeu de données. Nous pensions pouvoir récupérer ces données grâce à la colonne "total_ghg", en faisant la différence entre cette colonne et les trois gaz présentés ci-dessus. Hélas, cela n'est pas possible compte tenu d'une erreur que nous avons détectée et qui est détaillée un peu plus loin dans le rapport.

Puisque les émissions de CO2 sont très majoritaires par rapport aux émissions des autres GES, il est intéressant d'analyser les différentes sources d'émissions de CO2, ce que nous allons faire dans le prochain graphique.

Cumul des émissions de CO2 par source d’émissions



Ce graphique représente le cumul des émissions de CO2 depuis 1850 jusqu’à 2023, par source d’émissions, en million de tonnes.

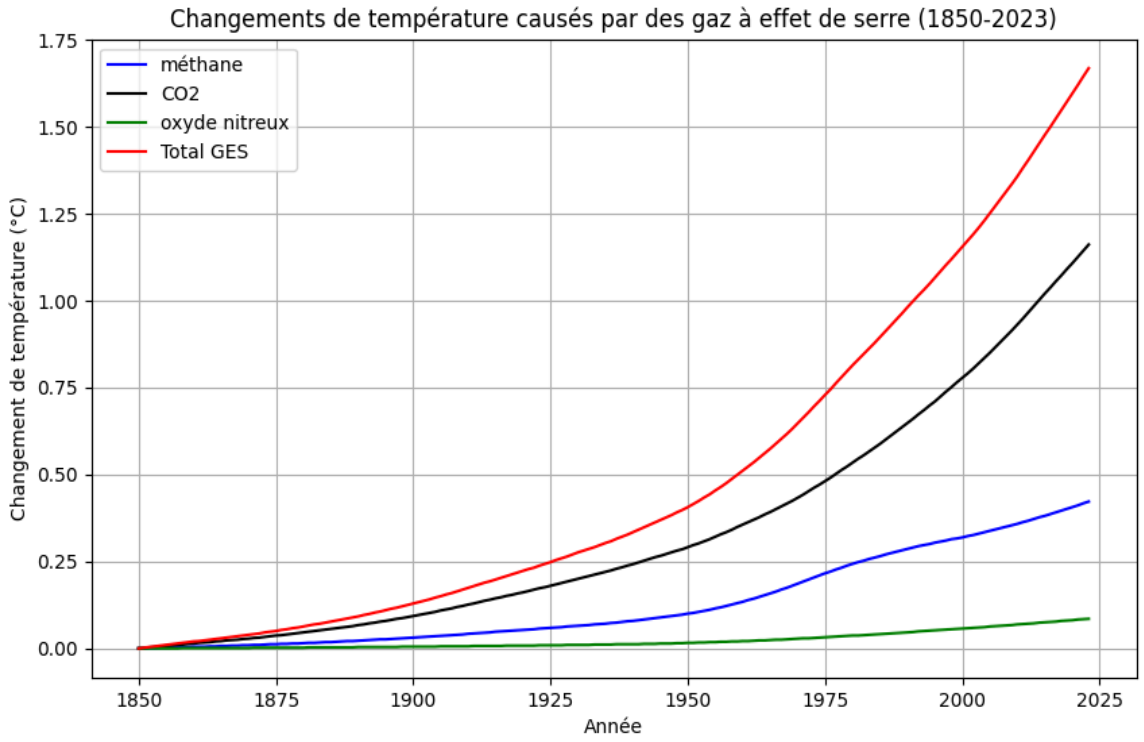
Les principales sources d’émissions de CO2 sont l’exploitation du charbon (notamment dans les centrales thermiques à charbon pour produire de l’électricité), et l’UTCATF (Utilisation des terres, changement d'affectation des terres et foresterie) correspondant au secteur agricole et à la déforestation. Ces deux sources ont émis plus de 800 milliards de tonnes de CO2, soit près d’un tiers des émissions de CO2 pour chacune.

Ensuite, nous avons l’exploitation du pétrole (utilisé comme carburant pour les transports terrestres, maritimes et aériens, par exemple), qui a émis 628 milliards de tonnes, ce qui représente presque un quart des émissions totales de CO2.

Dans une moindre mesure, nous retrouvons l’émission de CO2 dû à la combustion de gaz naturel (pour le chauffage domestique par exemple), avec une valeur de 270 milliards de tonnes, ce qui représente 10% des émissions de CO2.

Enfin, la fabrication du ciment, le torchage de gaz naturel et les autres industries représentent une petite proportion par rapport aux autres sources d’émission (3% à eux trois).

Evolution des changements de température causés par les différents gaz à effet de serre



Ce graphique montre l'évolution des changements de température (en °C) causés par divers gaz à effet de serre, sur la période 1850 à 2023, à l'échelle mondiale.

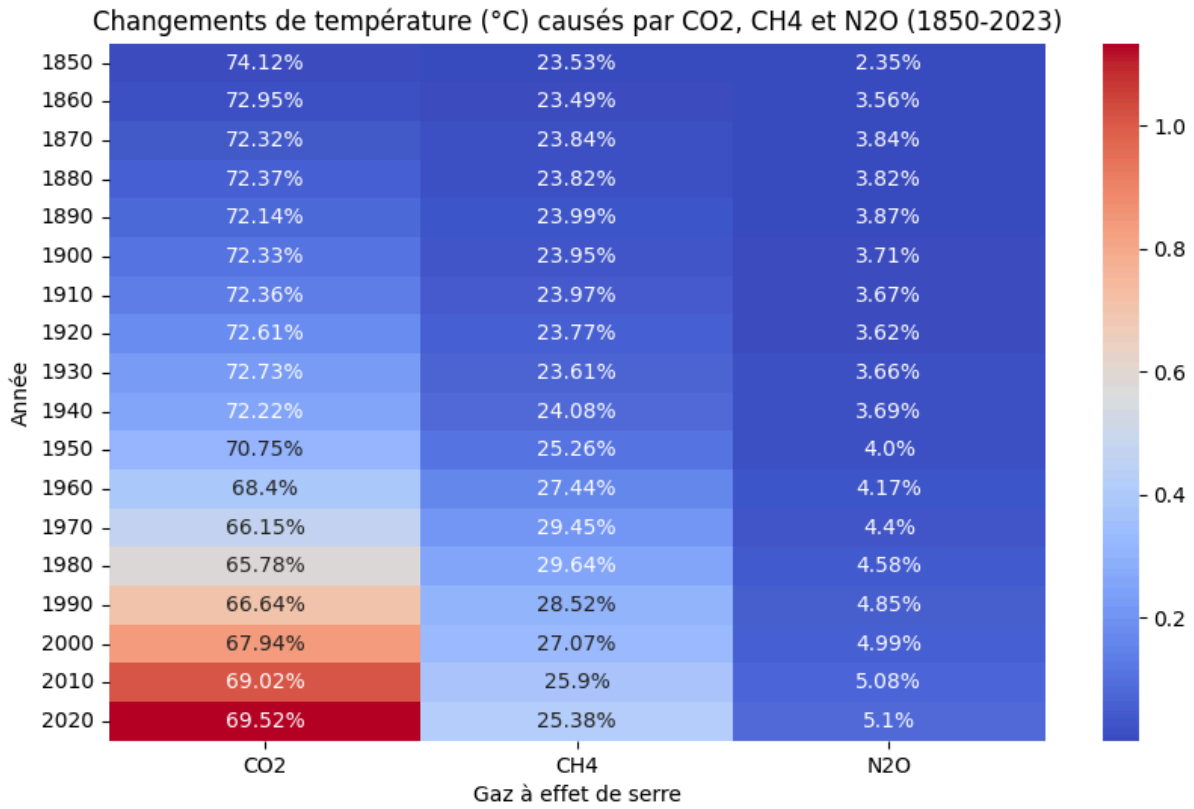
La courbe rouge montre les changements de température causés par l'ensemble des gaz à effet de serre. La courbe noire représente les changements de température causés par les émissions de CO2. La bleue, par les émissions de méthane. Et enfin, la verte par les émissions d'oxyde nitreux (ou protoxyde d'azote).

On remarque que la modification de température, à l'échelle mondiale, causée par l'ensemble des GES est en croissance exponentielle depuis le début du XXéme siècle. Avant 1950, les GES étaient responsables d'une hausse des températures de 0.5 °C. Environ 75 ans plus tard, ils sont responsables d'une hausse de près de 1.7 °C.

Le dioxyde de carbone (CO2) est le principal responsable de la hausse des températures. A lui seul, il représente aujourd'hui une hausse de 1.25 °C, ce qui est plus du double du changement de températures causé par le méthane (CH4), qui est le deuxième GES ayant le plus d'impact sur le climat.
Enfin, le protoxyde d'azote (N2O) est celui générant le moins de hausse de température. Les modifications de température causées par le N2O sont toutefois en légère hausse depuis les années 70.

L'évolution des changements de température provoquée par ces trois types de gaz est cohérente avec la répartition des émissions étudiée plus tôt ; à savoir que le CO2, gaz le plus émis, est aussi celui qui est le plus responsable du réchauffement climatique.

Proportion des gaz à effet de serre dans les changements de température par décennies

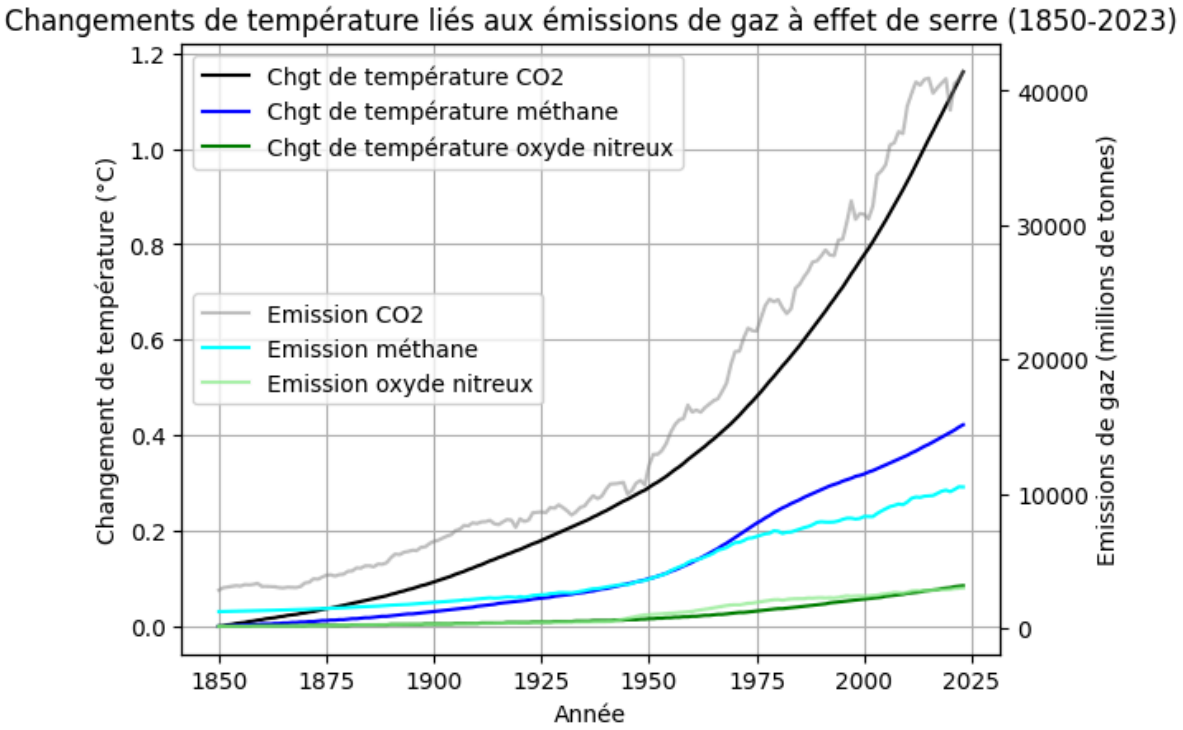


Ce graphique montre l'évolution des changements de température (en °C) sur la période 1850-2023, causés par chacun des trois gaz à effet de serre étudiés, avec la part de chacun pour chaque décennie.

Nous constatons plusieurs choses : les changements de température causés par les émissions de CO2 se sont grandement accentués depuis 1950 (inférieur à 0.5 °C avant 1950, supérieur à 1°C à partir de 2020). Cela est cohérent avec le fait que les quantités d'émission de CO2 augmentent de façon exponentielle depuis les années 1950, comme vus lors de nos précédents graphiques.

En revanche, la part du CO2 dans le réchauffement des températures diminue par rapport aux autres gaz, alors que la part de méthane (CH4) et de protoxyde d'azote (N2O) ne fait qu'augmenter. Cela ne signifie pas pour autant que les émissions de CO2 se réduisent ou ont moins d'impact sur le réchauffement, notre premier graphique a clairement montré que les émissions de CO2 étaient toujours en hausse. Au contraire, cela signifie que les émissions de méthane et de protoxyde d'azote sont en hausse à tel point qu'elles ont de plus en plus d'impact sur le réchauffement. Notons également que ces deux gaz ont un impact plus important sur le réchauffement climatique (raison pour laquelle leurs émissions sont converties en équivalent CO2 afin d'avoir une base de comparaison cohérente).

Evolution des changements de température relatifs aux émissions de gaz à effet de serre



Ce dernier graphique met en relation les changements de température causés par les trois principaux gaz à effet de serre étudiés, et les émissions de ces derniers, sur la période allant de 1850 à 2023.

Nous constatons un lien très fort entre l'augmentation des émissions de ces gaz, et l'augmentation des températures causée par ces gaz, à tel point que les deux courbes liées à l'oxyde nitreux sont quasiment superposées, tandis que les deux courbes liées au CO2 sont presque parallèles (la courbe des changements de température est beaucoup plus lisse, probablement qu'un traitement effectué en amont a permis de lisser les données).

Les courbes liées au méthane ont une évolution étonnante puisqu'elles s'inversent dans le temps. On remarque que l'augmentation des températures dues au méthane croît plus rapidement que les émissions de ce gaz.

Notons par ailleurs que, bien que ces données proviennent du même csv, les sources mêmes de ces données peuvent différer. Ainsi, les émissions de CO2 ne proviennent pas de la même source que les émissions des deux autres gaz et que les changements de température associés.

Ce dernier graphique est un récapitulatif de ce que nous avons pu observer lors des graphiques précédents, et met en lien la corrélation entre l'augmentation des émissions de gaz à effet de serre, et l'augmentation des températures.

Erreur dans les données d'émissions de gaz à effet de serre

Nous avons détecté une erreur à propos des colonnes total_ghg et total_ghg_excluding_lucf.

Normalement, on devrait s'attendre à ce que ces colonnes qui représentent le total d'émissions de gaz à effet de serre soient supérieures à la somme des colonnes des principaux gaz (dioxyde de carbone, méthane et oxyde nitreux).

On devrait ainsi avoir les formules suivantes :

total_ghg > co2_including_luc + methane + nitrous_oxide
total_ghg_excluding_lucf > co2 + methane + nitrous_oxide

Pourtant, nous nous sommes rendus compte que c'était l'inverse que nous observions.

	gaz	cumul
0	CO2_luc	2627266.546
1	CO2	1807783.284
2	methane	751477.115
3	nitrous_oxide	189648.129
4	total_ghg_luc	3522160.094
5	total_ghg	2210784.774
6	other_ghg_including_luc	-46231.696
7	other_ghg	-538123.754

Dans ce tableau, nous avons calculé la somme des émissions de chaque gaz de 1850 à 2023, pour les quatre premières lignes.
La ligne d'index 4 'total_ghg_luc' correspond à la somme des valeurs de la colonne 'total_ghg' pour la même période. L'index 5, 'total_ghg', correspond à la somme des valeurs de 'total_ghg_excluding_lucf', pour garder une nomenclature cohérente avec les colonnes 'co2' et 'co2_including_luc' (ici renommé en 'CO2_luc').

Les deux dernières lignes correspondent aux formules suivantes :

other_ghg_including_luc = total_ghg_luc - CO2_luc - methane - nitrous_oxide
other_ghg = total_ghg - CO2 - methane - nitrous_oxide

Le tout est exprimé en million de tonnes.

On constate une très grande différence, alors que ces valeurs devraient être positives. En effet, les colonnes liées à “total_ghg” sont supposées contenir non seulement les trois gaz principaux, mais également d’autres GES aux proportions plus petites et qui ne sont pas présentées dans le jeu de données.

Une explication probable se trouve au niveau des sources des données. Les colonnes relatives à “total_ghg” trouvent leurs sources du rapport “Jones et al. - National contributions to climate change (2024)”, tandis que les autres données d’émissions proviennent du Global Carbon Budget (2024).

Il est probable que ces deux sources n’utilisent pas les mêmes méthodologies pour récupérer les données.
Nous ne pourrons donc malheureusement pas comparer ces données entre elles.

Conclusion

Le fichier csv owid-co2-data offre une mine d’informations pour comprendre l’un des facteurs du réchauffement climatique, à savoir l’émission de gaz à effet de serre.

De par son volume de données, il est complexe et long à analyser. Contraint par le temps, nous devons cibler ce que nous voulons exactement faire ressortir de ce jeu de données, et c’est ce que nous avons essayé de faire dans cette première partie.

Nous avons ainsi décidé de nous focaliser dans un premier temps sur les données à échelle mondiale, pour observer les émissions de gaz dans le monde entier et constater leur impact sur les changements de température.

Nous avons pu ainsi remarquer la croissance exponentielle des émissions de CO2 depuis ces deux derniers siècle, ainsi que la terrifiante augmentation des températures qui en est la conséquence directe. En outre, nous avons constaté également une augmentation des émissions de méthane et de protoxyde d’azote, au pouvoir réchauffant supérieur à celui du CO2, si bien que la part de CO2 sur l’impact climatique diminue par rapport aux deux autres malgré la hausse des émissions globales.

Il sera intéressant de croiser ces données avec les jeux de données analysés par les autres membres de l’équipe. Nous pourrons notamment voir quelle est la part des changements de température liés aux gaz à effet de serre par rapport aux anomalies de température observées ces dernières décennies, et comparer les émissions de gaz avec les productions et consommations d’énergie.

Nous aimerions également avoir le temps de faire des analyses par zones géographiques, notamment en finalisant le nettoyage de notre csv par continents, et pourquoi pas même par pays, afin de croiser cela avec les données géographiques des autres dataset. En effet, il serait intéressant de visualiser les zones les plus concernées par le réchauffement climatique, et celles qui en sont le plus responsable.

Sources

owid-co2-data.csv provient de https://github.com/owid/co2-data?tab=readme-ov-file de l'organisation Our World in Data. https://ourworldindata.org
Les données proviennent de ces sources :

- Jones et al. - National contributions to climate change (2024) [https://zenodo.org/records/7636699/latest]
- Global Carbon Budget (2024) [https://globalcarbonbudget.org/]
- Bolt and van Zanden - Maddison Project Database 2023 [https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2023]
- U.S. Energy Information Administration - International Energy Data (2023) [https://www.eia.gov/opendata/bulkfiles.php]; Energy Institute - Statistical Review of World Energy (2024) [https://www.energyinst.org/statistical-review/]
- Population based on various sources (2024) [https://ourworldindata.org/population-sources]

Informations complètes ici :
https://github.com/owid/co2-data/blob/master/owid-co2-codebook.csv

équivalences CO2 :
https://www.connaissancedesenergies.org/questions-et-reponses-energies/gaz-effet-de-serre-quest-ce-que-l-equivalent-co2

Annexes

Valeurs possibles pour ‘country’

'Afghanistan' 'Africa' 'Africa (GCP)' 'Albania' 'Algeria' 'Andorra'
'Angola' 'Anguilla' 'Antarctica' 'Antigua and Barbuda' 'Argentina'
'Armenia' 'Aruba' 'Asia' 'Asia (GCP)' 'Asia (excl. China and India)'
'Australia' 'Austria' 'Azerbaijan' 'Bahamas' 'Bahrain' 'Bangladesh'
'Barbados' 'Belarus' 'Belgium' 'Belize' 'Benin' 'Bermuda' 'Bhutan'
'Bolivia' 'Bonaire Sint Eustatius and Saba' 'Bosnia and Herzegovina'
'Botswana' 'Brazil' 'British Virgin Islands' 'Brunei' 'Bulgaria'
'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon' 'Canada' 'Cape Verde'
'Central African Republic' 'Central America (GCP)' 'Chad' 'Chile' 'China'
'Christmas Island' 'Colombia' 'Comoros' 'Congo' 'Cook Islands'
'Costa Rica' "Cote d'Ivoire" 'Croatia' 'Cuba' 'Curacao' 'Cyprus'
'Czechia' 'Democratic Republic of Congo' 'Denmark' 'Djibouti' 'Dominica'
'Dominican Republic' 'East Timor' 'Ecuador' 'Egypt' 'El Salvador'
'Equatorial Guinea' 'Eritrea' 'Estonia' 'Eswatini' 'Ethiopia' 'Europe'
'Europe (GCP)' 'Europe (excl. EU-27)' 'Europe (excl. EU-28)'
'European Union (27)' 'European Union (28)' 'Faroe Islands' 'Fiji'
'Finland' 'France' 'French Polynesia' 'Gabon' 'Gambia' 'Georgia'
'Germany' 'Ghana' 'Greece' 'Greenland' 'Grenada' 'Guatemala' 'Guinea'
'Guinea-Bissau' 'Guyana' 'Haiti' 'High-income countries' 'Honduras'
'Hong Kong' 'Hungary' 'Iceland' 'India' 'Indonesia'
'International aviation' 'International shipping'
'International transport' 'Iran' 'Iraq' 'Ireland' 'Israel' 'Italy'
'Jamaica' 'Japan' 'Jordan' 'Kazakhstan' 'Kenya' 'Kiribati' 'Kosovo'
'Kuwait' 'Kuwaiti Oil Fires' 'Kuwaiti Oil Fires (GCP)' 'Kyrgyzstan'
'Laos' 'Latvia' 'Least developed countries (Jones et al.)' 'Lebanon'
'Lesotho' 'Liberia' 'Libya' 'Liechtenstein' 'Lithuania'
'Low-income countries' 'Lower-middle-income countries' 'Luxembourg'
'Macao' 'Madagascar' 'Malawi' 'Malaysia' 'Maldives' 'Mali' 'Malta'
'Marshall Islands' 'Mauritania' 'Mauritius' 'Mexico'
'Micronesia (country)' 'Middle East (GCP)' 'Moldova' 'Monaco' 'Mongolia'
'Montenegro' 'Montserrat' 'Morocco' 'Mozambique' 'Myanmar' 'Namibia'
'Nauru' 'Nepal' 'Netherlands' 'New Caledonia' 'New Zealand' 'Nicaragua'
'Niger' 'Nigeria' 'Niue' 'Non-OECD (GCP)' 'North America'
'North America (GCP)' 'North America (excl. USA)' 'North Korea'
'North Macedonia' 'Norway' 'OECD (GCP)' 'OECD (Jones et al.)' 'Oceania'
'Oceania (GCP)' 'Oman' 'Pakistan' 'Palau' 'Palestine' 'Panama'
'Papua New Guinea' 'Paraguay' 'Peru' 'Philippines' 'Poland' 'Portugal'
'Qatar' 'Romania' 'Russia' 'Rwanda' 'Ryukyu Islands'
'Ryukyu Islands (GCP)' 'Saint Helena' 'Saint Kitts and Nevis'
'Saint Lucia' 'Saint Pierre and Miquelon'
'Saint Vincent and the Grenadines' 'Samoa' 'San Marino'
'Sao Tome and Principe' 'Saudi Arabia' 'Senegal' 'Serbia' 'Seychelles'
'Sierra Leone' 'Singapore' 'Sint Maarten (Dutch part)' 'Slovakia'

'Slovenia' 'Solomon Islands' 'Somalia' 'South Africa' 'South America'
'South America (GCP)' 'South Korea' 'South Sudan' 'Spain' 'Sri Lanka'
'Sudan' 'Suriname' 'Sweden' 'Switzerland' 'Syria' 'Taiwan' 'Tajikistan'
'Tanzania' 'Thailand' 'Togo' 'Tonga' 'Trinidad and Tobago' 'Tunisia'
'Turkey' 'Turkmenistan' 'Turks and Caicos Islands' 'Tuvalu' 'Uganda'
'Ukraine' 'United Arab Emirates' 'United Kingdom' 'United States'
'Upper-middle-income countries' 'Uruguay' 'Uzbekistan' 'Vanuatu'
'Vatican' 'Venezuela' 'Vietnam' 'Wallis and Futuna' 'World' 'Yemen'
'Zambia' 'Zimbabwe'

Valeurs possibles pour ‘iso_code’

'AFG' nan 'ALB' 'DZA' 'AND' 'AGO' 'AIA' 'ATA' 'ATG' 'ARG' 'ARM' 'ABW'
'AUS' 'AUT' 'AZE' 'BHS' 'BHR' 'BGD' 'BRB' 'BLR' 'BEL' 'BLZ' 'BEN' 'BMU'
'BTN' 'BOL' 'BES' 'BIH' 'BWA' 'BRA' 'VGB' 'BRN' 'BGR' 'BFA' 'BDI' 'KHM'
'CMR' 'CAN' 'CPV' 'CAF' 'TCD' 'CHL' 'CHN' 'CXR' 'COL' 'COM' 'COG' 'COK'
'CRI' 'CIV' 'HRV' 'CUB' 'CUW' 'CYP' 'CZE' 'COD' 'DNK' 'DJI' 'DMA' 'DOM'
'TLS' 'ECU' 'EGY' 'SLV' 'GNQ' 'ERI' 'EST' 'SWZ' 'ETH' 'FRO' 'FJI' 'FIN'
'FRA' 'PYF' 'GAB' 'GMB' 'GEO' 'DEU' 'GHA' 'GRC' 'GRL' 'GRD' 'GTM' 'GIN'
'GNB' 'GUY' 'HTI' 'HND' 'HKG' 'HUN' 'ISL' 'IND' 'IDN' 'IRN' 'IRQ' 'IRL'
'ISR' 'ITA' 'JAM' 'JPN' 'JOR' 'KAZ' 'KEN' 'KIR' 'KWT' 'KGZ' 'LAO' 'LVA'
'LBN' 'LSO' 'LBR' 'LBY' 'LIE' 'LTU' 'LUX' 'MAC' 'MDG' 'MWI' 'MYS' 'MDV'
'MLI' 'MLT' 'MHL' 'MRT' 'MUS' 'MEX' 'FSM' 'MDA' 'MCO' 'MNG' 'MNE' 'MSR'
'MAR' 'MOZ' 'MMR' 'NAM' 'NRU' 'NPL' 'NLD' 'NCL' 'NZL' 'NIC' 'NER' 'NGA'
'NIU' 'PRK' 'MKD' 'NOR' 'OMN' 'PAK' 'PLW' 'PSE' 'PAN' 'PNG' 'PRY' 'PER'
'PHL' 'POL' 'PRT' 'QAT' 'ROU' 'RUS' 'RWA' 'SHN' 'KNA' 'LCA' 'SPM' 'VCT'
'WSM' 'SMR' 'STP' 'SAU' 'SEN' 'SRB' 'SYC' 'SLE' 'SGP' 'SXM' 'SVK' 'SVN'
'SLB' 'SOM' 'ZAF' 'KOR' 'SSD' 'ESP' 'LKA' 'SDN' 'SUR' 'SWE' 'CHE' 'SYR'
'TWN' 'TJK' 'TZA' 'THA' 'TGO' 'TON' 'TTO' 'TUN' 'TUR' 'TKM' 'TCA' 'TUV'
'UGA' 'UKR' 'ARE' 'GBR' 'USA' 'URY' 'UZB' 'VUT' 'VAT' 'VEN' 'VNM' 'WLF'
'YEM' 'ZMB' 'ZWE'

Valeurs de ‘country’ n’ayant pas d’iso-code

'Africa' 'Africa (GCP)' 'Asia' 'Asia (GCP)'
'Asia (excl. China and India)' 'Central America (GCP)' 'Europe'
'Europe (GCP)' 'Europe (excl. EU-27)' 'Europe (excl. EU-28)'
'European Union (27)' 'European Union (28)' 'High-income countries'
'International aviation' 'International shipping'
'International transport' 'Kosovo' 'Kuwaiti Oil Fires'
'Kuwaiti Oil Fires (GCP)' 'Least developed countries (Jones et al.)'
'Low-income countries' 'Lower-middle-income countries'
'Middle East (GCP)' 'Non-OECD (GCP)' 'North America'
'North America (GCP)' 'North America (excl. USA)' 'OECD (GCP)'
'OECD (Jones et al.)' 'Oceania' 'Oceania (GCP)' 'Ryukyu Islands'

'Ryukyu Islands (GCP)' 'South America' 'South America (GCP)'
'Upper-middle-income countries' 'World'

Proportion de valeurs manquantes par colonnes

share_global_cumulative_other_co2 : 95.8%
share_global_other_co2 : 95.8%
other_co2_per_capita : 95.07%
cumulative_other_co2 : 93.62%
other_industry_co2 : 93.62%
consumption_co2_per_gdp : 91.15%
consumption_co2_per_capita : 91.03%
trade_co2 : 90.96%
trade_co2_share : 90.96%
consumption_co2 : 90.31%
energy_per_gdp : 84.67%
co2_including_luc_per_unit_energy : 80.67%
energy_per_capita : 79.86%
primary_energy_consumption : 79.78%
co2_per_unit_energy : 79.38%
share_global_flaring_co2 : 78.34%
share_global_cumulative_flaring_co2 : 78.34%
flaring_co2_per_capita : 70.72%
share_global_cumulative_gas_co2 : 70.05%
share_global_gas_co2 : 70.05%
gdp : 69.61%
cumulative_flaring_co2 : 68.34%
flaring_co2 : 68.22%
co2_including_luc_per_gdp : 66.55%
gas_co2_per_capita : 65.55%
co2_per_gdp : 65.08%
cumulative_gas_co2 : 64.11%
gas_co2 : 64.11%
coal_co2_per_capita : 58.06%
share_global_coal_co2 : 56.66%
coal_co2 : 56.66%
share_global_cumulative_coal_co2 : 56.66%
cumulative_coal_co2 : 56.66%
share_global_cumulative_cement_co2 : 56.25%
share_global_cement_co2 : 56.25%
co2_including_luc_growth_abs : 53.61%
co2_including_luc_growth_prct : 53.61%
co2_including_luc_per_capita : 53.19%
cumulative_co2_including_luc : 53.01%
share_global_co2_including_luc : 53.01%
co2_including_luc : 53.01%

share_global_cumulative_co2_including_luc : 53.01%
share_global_oil_co2 : 53.0%
share_global_cumulative_oil_co2 : 53.0%
oil_co2_per_capita : 51.31%
cumulative_oil_co2 : 49.76%
oil_co2 : 49.76%
cement_co2_per_capita : 49.48%
co2_growth_prct : 48.19%
co2_per_capita : 47.84%
co2_growth_abs : 46.24%
share_global_cumulative_co2 : 45.55%
share_global_co2 : 45.55%
cumulative_co2 : 45.55%
cumulative_cement_co2 : 42.54%
cement_co2 : 42.49%
co2 : 41.95%
ghg_excluding_lucf_per_capita : 28.99%
methane_per_capita : 28.65%
ghg_per_capita : 28.65%
nitrous_oxide_per_capita : 27.64%
land_use_change_co2_per_capita : 27.41%
cumulative_luc_co2 : 25.81%
total_ghg_excluding_lucf : 25.81%
share_global_luc_co2 : 25.81%
land_use_change_co2 : 25.81%
share_global_cumulative_luc_co2 : 25.81%
total_ghg : 25.46%
methane : 25.46%
temperature_change_from_ch4 : 24.17%
temperature_change_from_n2o : 24.17%
nitrous_oxide : 23.73%
share_of_temperature_change_from_ghg : 18.31%
temperature_change_from_ghg : 18.31%
temperature_change_from_co2 : 18.31%
population : 18.27%
iso_code : 15.8%
year : 0.0%
country : 0.0%