

RAPPORT PROJET

TEMPÉRATURE TERRESTRE

« *L'été toute l'année* »



Membres du projet :

Guillaume BOGDANOWICZ

Brunel TCHEKELI

Hélène LEBOURG KOULIBALY

Table des matières

Remerciements	3
Introduction.....	4
Méthodologie	4
Résultats attendus.....	5
Cadrage.....	6
Définitions et explications	6
Glossaire	8
Découverte des données et du projet.....	10
Exploration et analyse avec Dataviz'	14
Etudes des corrélations	21
Nettoyage et Pre-processing	23
Modélisation & Machine-learning.....	28
ARBRE DE DECISION :	29
LA FORET ALEATOIRE :	31
REGRESSION LASSO :	31
GRADIENT BOOSTING :	32
Retours d'expérience	37
Conclusion	39
Références, sources et annexes	40



Remerciements

Avant tout, nous souhaitons remercier l'équipe de DataScientest, notre chef de cohorte, Khalil, et plus particulièrement notre tuteur, Tarik ANOUAR, pour son soutien et suivi pédagogique sans faille, qui nous ont permis de relever les enjeux de cette formation et de mener à bien notre projet.



Introduction

Dans le cadre de notre formation Data Analyst en Bootcamp chez Datascientest, nous avons travaillé sur un projet « fil rouge » nous permettant d'aborder toutes les étapes de développement d'une solution data. Nous avons été sélectionnés pour le projet « Température terrestre » avec comme objectif de « *constater le réchauffement (et le dérèglement) climatique global à l'échelle de la planète sur les derniers siècles et dernières décennies* ».

Ce projet nous a permis d'appliquer les compétences apprises au cours de la formation, de pouvoir travailler en équipe et d'être en capacité d'apprendre sur des problématiques métier.

Le réchauffement climatique est un sujet de préoccupation croissant en raison de son impact sur l'environnement, l'économie et la société. Les changements climatiques, en particulier le réchauffement climatique, ont des conséquences néfastes sur la biodiversité, les ressources en eau, l'agriculture, la santé humaine et les infrastructures. Il est donc crucial de surveiller et d'analyser les données de température mondiale pour mieux comprendre les tendances et les variations, et ainsi anticiper et atténuer les effets du changement climatique.

Dans ce contexte, notre projet de data analyse vise à étudier les données de température mondiale afin de fournir des informations pertinentes et actualisées sur les tendances, les variations et les anomalies de température. Nous nous appuierons sur des sources de données fiables et reconnues, telles que les bases de données de la NASA, de Berkeley Earth, certaines données issues de Kaggle et Github (<https://data.giss.nasa.gov/gistemp/> et <https://github.com/owid/co2-data>). Notre objectif est de contribuer à la prise de décision en matière de politiques environnementales et de sensibiliser le public à l'importance de la lutte contre le changement climatique.

Méthodologie

Pour mener à bien notre projet de data analyse, nous suivrons une méthodologie rigoureuse et structurée, qui comprendra les étapes suivantes :

1. Collecte des données : Nous collecterons les données de température mondiale auprès de sources fiables et reconnues, en veillant à couvrir une période suffisamment longue pour



permettre l'analyse des tendances et des variations. Nous nous assurerons également que les données sont complètes, cohérentes et de qualité.

2. Nettoyage et préparation des données : Nous procéderons à un nettoyage et à une préparation minutieuse des données, en éliminant les erreurs, les incohérences et les valeurs manquantes, et en normalisant les données pour faciliter leur analyse et leur comparaison.
3. Analyse exploratoire des données : Nous réaliserons une analyse exploratoire des données pour identifier les tendances, les variations et les anomalies de température, ainsi que les relations entre les différentes variables. Nous utiliserons des techniques de visualisation des données, telles que les graphiques, les cartes et les tableaux, pour faciliter la compréhension et l'interprétation des résultats.
4. Modélisation et prévision : Nous développerons des modèles statistiques et/ou de machine learning pour analyser les relations entre les variables et pour prévoir l'évolution future des températures mondiales. Nous évaluerons la performance de ces modèles en utilisant des indicateurs de qualité, tels que le coefficient de détermination (R^2), l'erreur quadratique moyenne (RMSE) et le coefficient de corrélation (r).
5. Interprétation et communication des résultats : Nous interpréterons les résultats de notre analyse en mettant en évidence les tendances, les variations et les anomalies de température, ainsi que les implications pour l'environnement, l'économie et la société. Nous communiquerons nos résultats de manière claire et accessible, en utilisant des supports visuels et des explications simples, afin de sensibiliser le public et d'informer les décideurs politiques.

Résultats attendus

Notre projet de data analyse des données de température mondiale devrait permettre de répondre aux questions suivantes :

- Quelles sont les tendances à long terme des températures mondiales ? Y a-t-il une augmentation significative des températures au cours des dernières décennies ?
- Quelles sont les variations saisonnières et géographiques des températures ? Existe-t-il des différences marquées entre les régions et les saisons ?
- Quelles sont les anomalies de température les plus importantes et les plus fréquentes ? Peut-on les attribuer à des événements climatiques spécifiques, tels que les phénomènes El Niño et La Niña ?
- Quels sont les facteurs qui influencent les températures mondiales ? Peut-on établir des relations entre les températures et les émissions de gaz à effet de serre ?



Cadrage

Tout d'abord, nous avons observé les données disponibles sur les sites web mentionnés dans la fiche projet (voir Annexes), respectivement la NASA et Github (<https://data.giss.nasa.gov/gistemp/> et <https://github.com/owid/co2-data>).

Une première réunion de cadrage nous a permis d'identifier les étapes du projet et de comprendre ce qui était attendu. Nous nous sommes organisés et partagés les tâches à l'aide d'un diagramme de GANTT (voir Annexes) et de réunions régulières afin de suivre nos avancements et répondre à nos problématiques.

A l'aide des data sets et de recherches parallèles sur le sujet, nous avons établi une présentation du projet permettant de délimiter et de comprendre les enjeux de ce dernier.

L'objectif du projet est de « constater le réchauffement (et le dérèglement) climatique global à l'échelle de la planète sur les derniers siècles et dernières décennies ». Nous avons choisi le cadrage suivant :

1. Découverte donnée et projet
2. Exploration et analyse des données avec Dataviz'
3. Nettoyage et pre-processing
4. Modélisation

Définitions et explications

Changement climatique : définit par les Nations Unies comme « [...] les variations à long terme de la température et des modèles météorologiques. Il peut s'agir de variations naturelles, dues par exemple à celles du cycle solaire ou à des éruptions volcaniques massives. Cependant, depuis les années 1800, les activités humaines constituent la cause principale des changements climatiques, essentiellement en raison de la combustion de combustibles fossiles comme le charbon, le pétrole et le gaz. La combustion



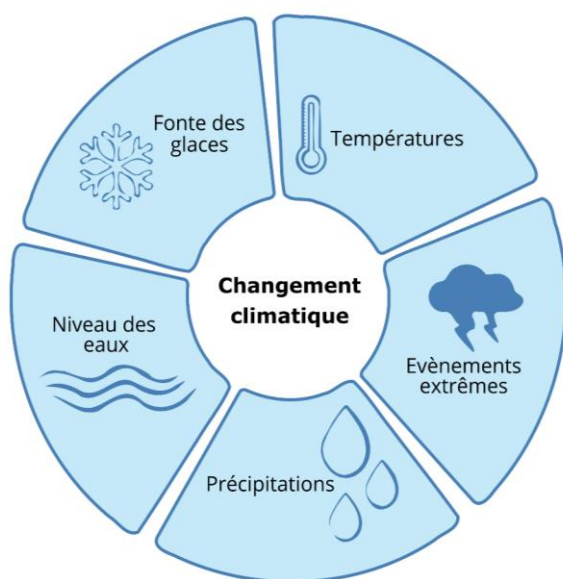
de combustibles fossiles génère des émissions de gaz à effet de serre qui agissent comme une couverture autour de la Terre, emprisonnant la chaleur du soleil et entraînant une hausse des températures¹ ».

Réchauffement climatique : augmentation générale des températures.

Les principales causes du changement climatique : « Les émissions de dioxyde de carbone et de méthane, notamment, sont à l'origine des changements climatiques. Elles résultent par exemple de l'utilisation de carburants pour alimenter les véhicules ou du charbon pour chauffer un bâtiment. Le défrichement des terres et des forêts peut également entraîner la libération de dioxyde de carbone. L'agriculture et les moteurs à combustion constituent une source importante d'émissions de méthane. Les secteurs de l'énergie, de l'industrie, des transports et de la construction ainsi que de l'agriculture et d'autres utilisations des terres figurent parmi les principaux émetteurs² ».

Les principales conséquences du changement climatique : « [...] sécheresses intenses, pénuries d'eau, graves incendies, élévation du niveau de la mer, inondations, fonte des glaces polaires, tempêtes catastrophiques et déclin de la biodiversité³ ».

Les composantes du changement climatique



Office International de l'Eau -

A travers notre
projet nous nous
concentrons sur
une composante :
LES TEMPERATURES !

¹ Source : <https://www.un.org/fr/climatechange/what-is-climate-change>

² <https://www.un.org/fr/climatechange/science/causes-effects-climate-change>

³ <https://www.un.org/fr/climatechange/science/causes-effects-climate-change>



Glossaire

year – Les années.

population – La population totale.

gdp – Produit intérieur brut (PIB).

co2 – Dioxyde de carbone.

cement_co2 – Emission CO₂ du ciment : le ciment est composé de 80 % de calcaire et 20 % d'argile. Les émissions CO₂ du ciment résultent de la décomposition du calcaire par la calcination.

co2_per_capita – CO₂ par habitant.

coal_co2 – CO₂ produit par le charbon : la combustion du charbon émet des particules volatiles dangereuses : benzène et ses dérivés aromatiques.

flaring_co2 - CO₂ produit par le torchage / brûlage : le torchage est une pratique qui consiste à brûler les rejets de gaz naturels associés à l'extraction de pétrole.

gas_co2 - CO₂ produit par l'essence : combustible liquide qui, mélangé à l'air (carburant), est inflammable et peut être utilisé dans un moteur à explosion (i.e. essence, gazole). La combustion de l'essence produit du CO₂.

methane – méthane : gaz composé de molécules de quatre atomes d'hydrogène et d'un atome de carbone. Le méthane est le constituant principal du gaz naturel, combustible d'origine fossile. Les sources naturelles incluent les terres marécageuses, les marais, les termites et les océans. Les sources synthétiques incluent l'exploitation et la brûlure des combustibles fossiles, les processus digestifs chez les ruminants tels que le bétail, les paddys de riz et les sites d'enfouissement des déchets.

nitrous_oxide - Protoxyde d'azote : aussi appelé oxyde nitreux est un composé oxygéné de l'azote.

oil_co2 – CO₂ produit par le pétrole : huile minérale naturelle combustible, hydrocarbure liquide accumulé dans les roches, en gisements. L'exploitation du pétrole est directement liée à la destruction de la forêt et la pollution des nappes phréatiques, mais également par les nombreux déchets d'exploitation ainsi que les produits chimiques libérés pour les différents traitements dans les phases de production de dérivés pétrochimiques.



total_ghg – Total des émissions à effet de serre.

temp_anomaly – Les anomalies de températures.


RMSE – en anglais *root mean square error*, est la racine carrée de la moyenne des erreurs quadratiques. C'est une mesure utilisée pour évaluer les différences entre les valeurs prédites par un modèle de Machine Learning et les valeurs réelles.

MSE – en anglais *mean squared error*, est l'erreur quadratique moyenne qui permet d'évaluer la précision d'une prédiction.



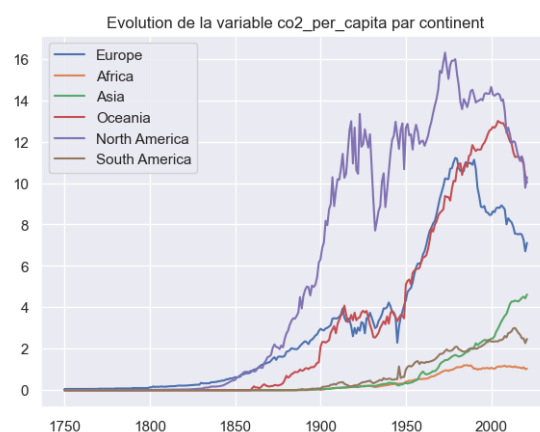
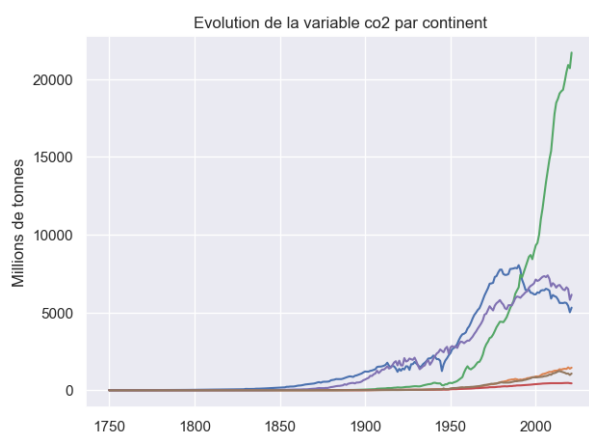
Découverte des données et du projet

Afin de découvrir les données collectées et démarrer l'exploration du projet nous avons suivi les étapes suivantes :

1. Chargement des data sets et des librairies (pandas, NumPy, matplotlib) sur Google COLAB 
2. Création de notebooks de façon partagée
3. Affichage des informations des variables des data sets afin d'inspecter rapidement les données stockées et identifier les types des variables, les valeurs manquantes, les doublons.
4. Choix des variables pertinentes au sujet
5. Première prise en main des données du data set sur les émissions de CO₂ (source : Data on CO₂ and Greenhouse Gas Emissions by Our World in Data) et affichage de quelques graphiques :

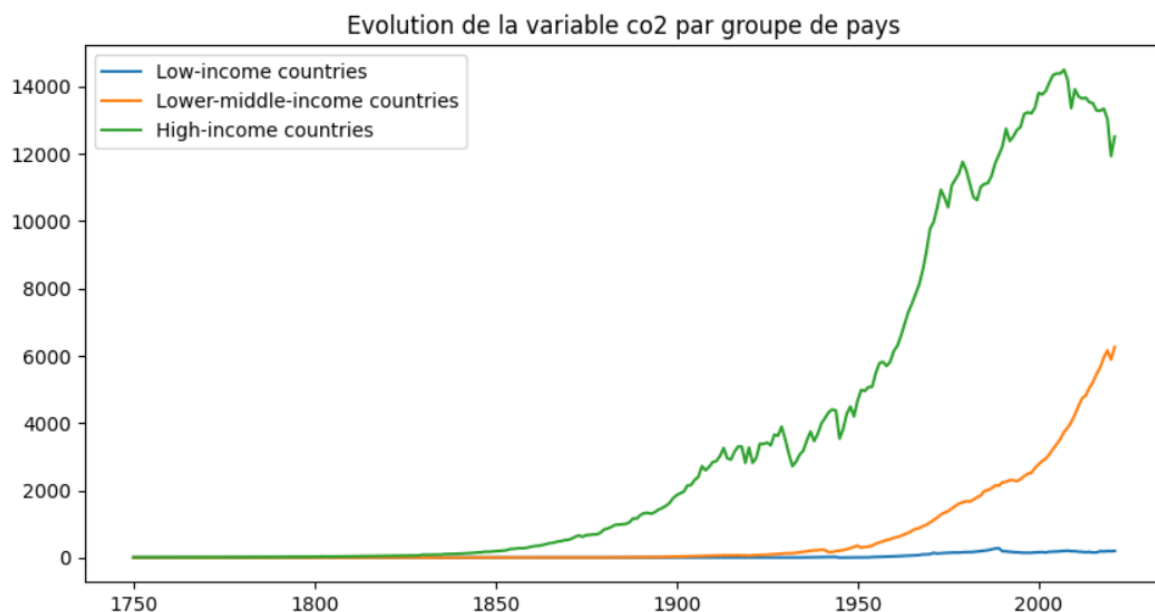
Graphique « Evolution de la variable CO₂ par continent » présente la production de CO₂ des continents. Avec l'Asie en tête depuis le début des années 2000.

Graphique « Evolution de la variable CO₂_per_capita par continent », présente l'évolution de la production de CO₂ par habitant en fonction des six continents. L'Amérique du Nord est le continent ayant la plus forte consommation de CO₂ par habitant de 1800 à 2005.

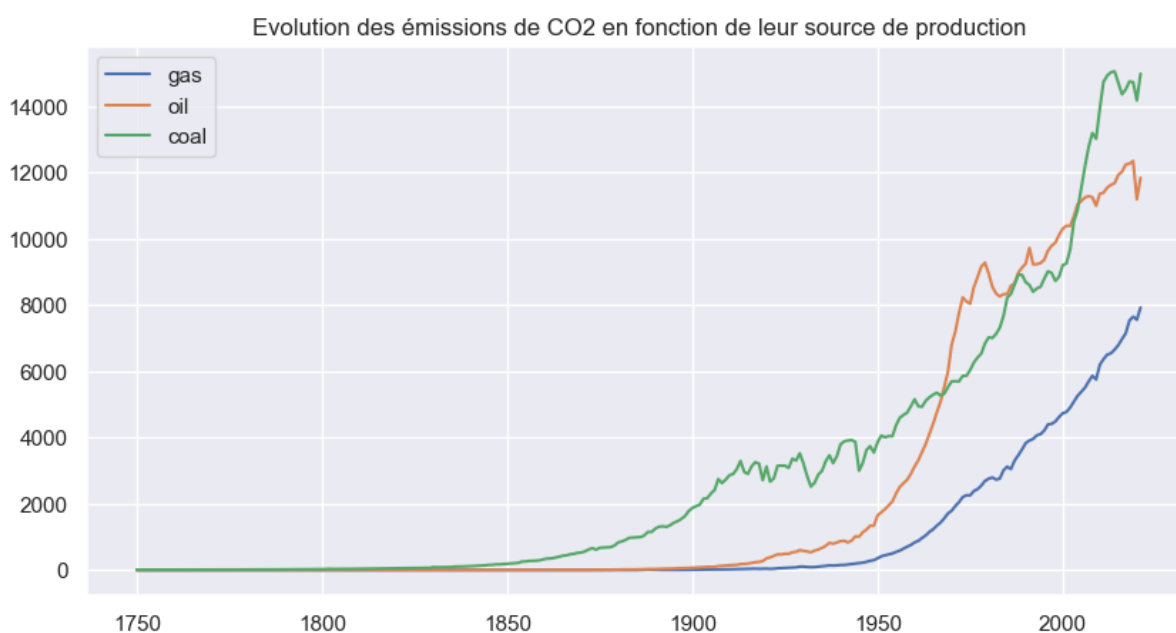


Graphique « Evolution de la variable CO₂ par groupes de pays » illustre la production de CO₂ en fonction des revenus des pays. Il est sans surprise que les pays à revenus élevés sont les pays ayant la production de CO₂ la plus importante.

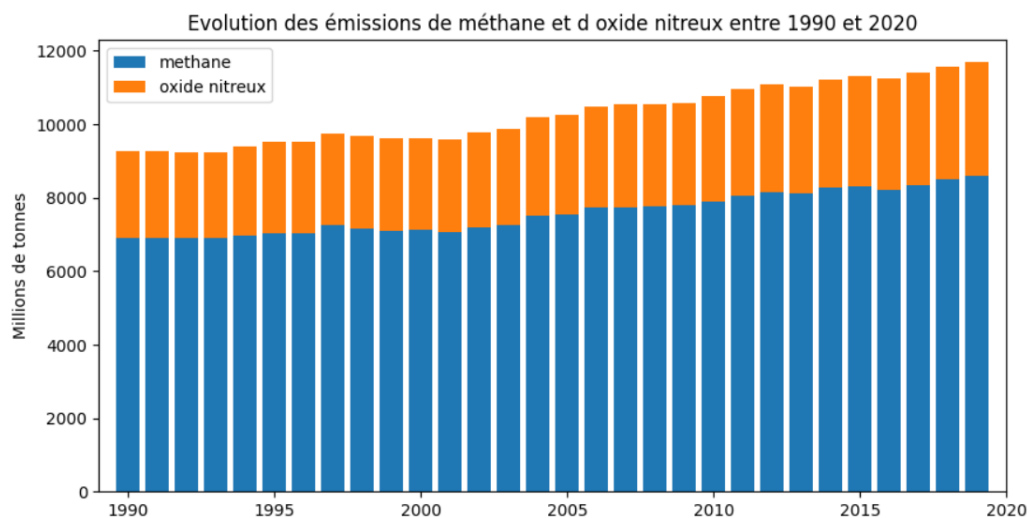




Graphique « Evolution des émissions de CO₂ en fonction de la source de production » observe la production de CO₂ en fonction de l'utilisation de charbon, de pétrole et d'essence.



Graphique « Evolution des émissions de méthane et d'oxyde nitreux entre 1990 et 2020 » montre la prédominance du méthane au fil des ans sur les émissions de CO₂.



Les limites des jeux de données :

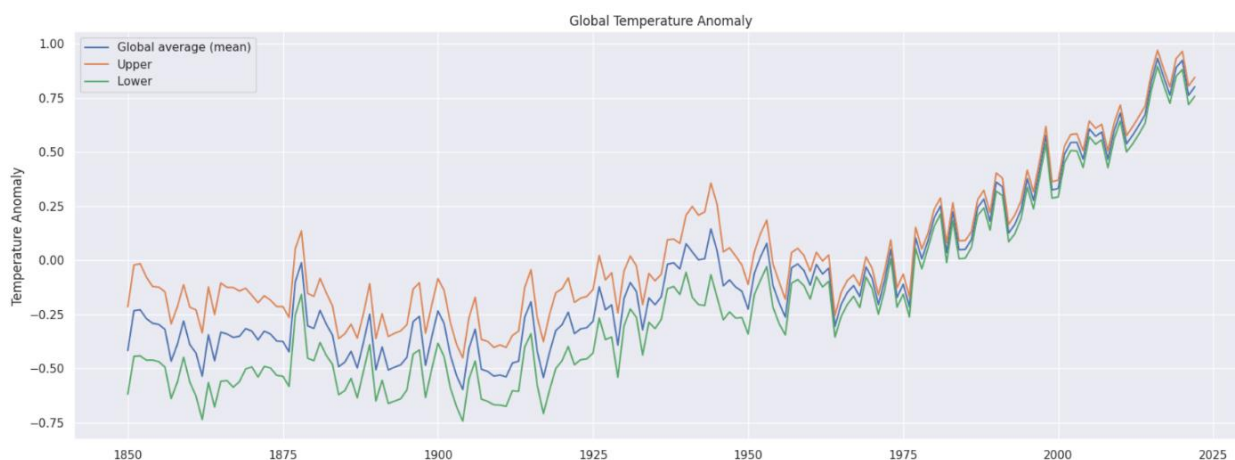
- le jeu de données de la NASA présente les températures globales moyennes par mois de 1880 à 2023
- le jeu de données du site Github présente les niveaux d'émission de CO₂ en fonction d'une multitude de facteurs (année, pays, continent, population, PIB, etc.).
- les informations pour les années les plus anciennes ne sont pas toujours renseignées. Cela s'explique par l'absence de stockage de données et des technologies correspondantes. C'est pourquoi nous avons décidé de commencer l'analyse à partir de 1880 afin d'avoir les données nécessaires à notre analyse.

Afin de compléter notre jeu de données :

- sur le site *Our World in Data* : (<https://ourworldindata.org/grapher/hadcrut-surface-temperature-anomaly>), jeu de données sur les anomalies de températures globales.
- Global temperature anomaly data from NOAA : [Global Time Series | Climate at a Glance | National Centers for Environmental Information \(NCEI\) \(noaa.gov\)](#), compilation par continent des anomalies annuelles de températures.



Nous avons procédé aux mêmes étapes effectuées sur les deux précédents jeux de données et affiché un premier graphique comparant les anomalies de températures maximales, minimales et moyennes.



Cette étape de découverte des données nous a permis de nous familiariser avec les données, de regrouper celles dont nous allons avoir besoin pour la suite du projet et d'avoir une meilleure connaissance du sujet quant aux facteurs liés à la température terrestre et ce que l'on appelle le changement climatique.

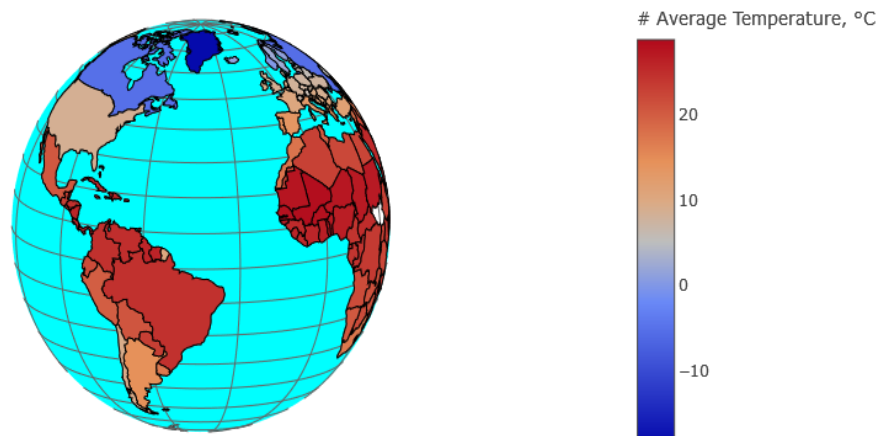


Exploration et analyse avec Dataviz'

Nous pouvons explorer et manipuler nos données. Ceci nous a permis de comprendre davantage le sujet et d'établir les corrélations entre les différentes variables.

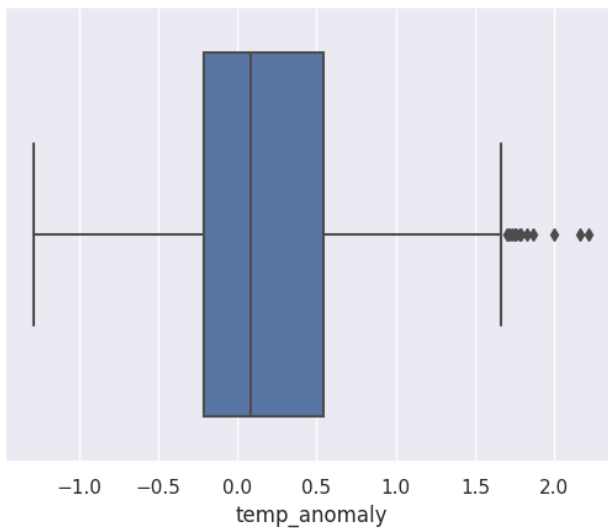
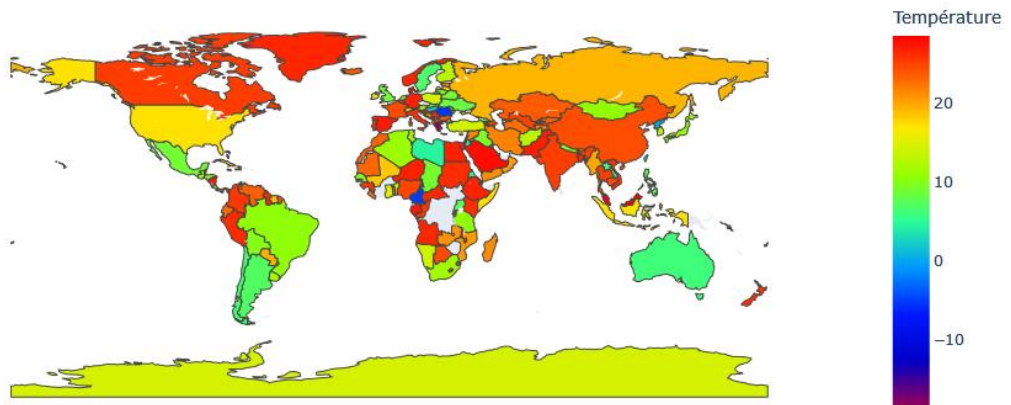
Afin d'avoir un aperçu général, nous avons décidé d'illustrer à travers un globe qui considère les températures terrestres moyennes. Ceci permet de délimiter très clairement les différentes zones climatiques notamment entre l'hémisphère Nord et l'hémisphère Sud.

Average land temperature in countries



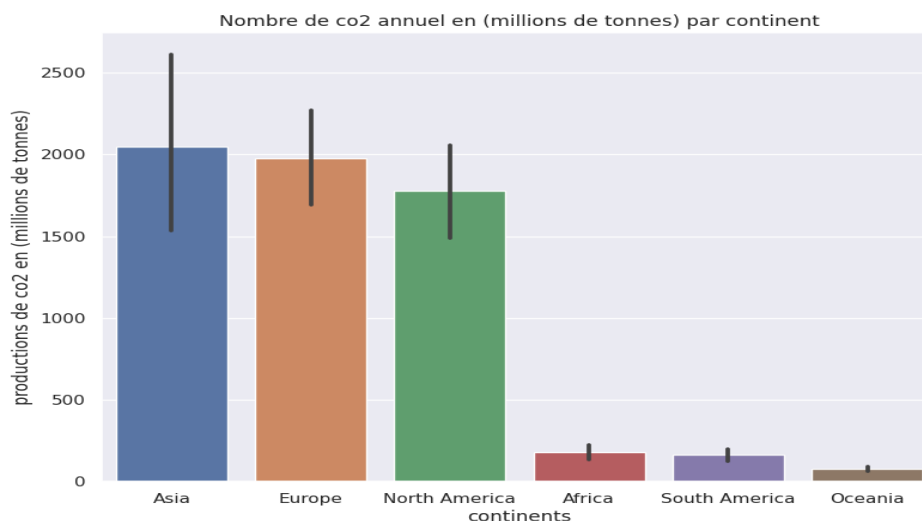
Sur la prochaine carte, il est intéressant de présenter l'évolution des températures, cette fois-ci par pays. En effet, nous pouvons constater que l'évolution des températures entre pays est bien plus disparate notamment sur le continent africain. De plus, nous pouvons observer que l'évolution des températures touche l'hémisphère Nord même si les températures restent inférieures à l'hémisphère Sud.





A l'aide d'une boîte à moustaches, on note que les températures négatives sont beaucoup plus fortes que les températures positives malgré une médiane plus proche des valeurs minimales que maximales. Les températures positives extrêmes sont nombreuses en effet, nous pouvons observer des températures positives extrêmes (outliers). Ces observations nous permettent de valider le changement climatique en tant que réchauffement climatique.

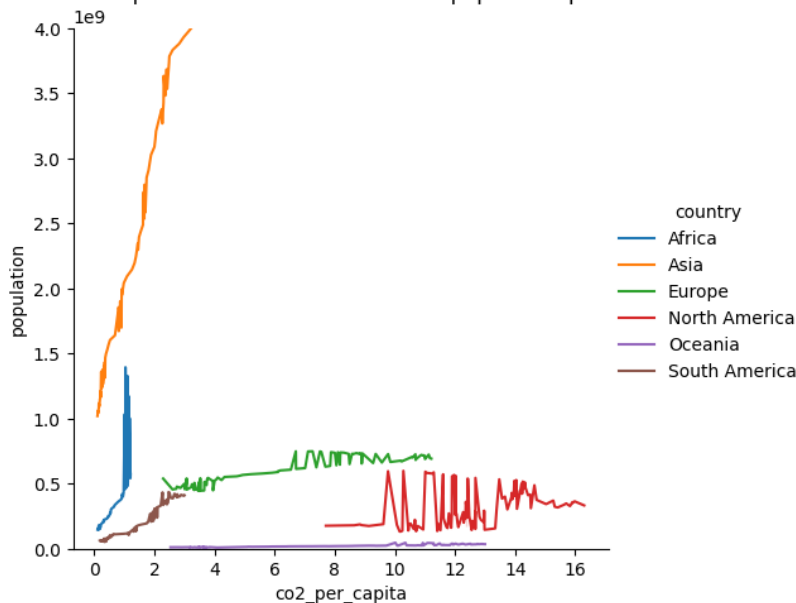
Dans le graphique ci-dessous, nous avons voulu exposer la production de CO₂ par continent afin de mesurer le poids de chaque continent et de pouvoir les comparer entre eux.



Ainsi, dans le graphique ci-après nous pouvons voir très clairement la prédominance de l'Asie, de l'Europe et de l'Amérique du Nord par rapport à l'Afrique, l'Amérique du Sud et l'Océanie. On constate que l'Afrique et l'Océanie produisent environ 10 à 12 fois moins de CO₂ que l'Asie, l'Europe et l'Amérique du Nord. Le premier continent producteur de CO₂ est l'Asie (environ 2000 millions de tonnes) et le dernier continent est l'Océanie (environ 100 millions de tonnes). La différence entre ces deux continents représente environ 1900 millions de tonnes.

Toutes nos visualisations, nous permettent de voir une grande différence entre l'Asie et le reste du monde. Penchons-nous plus en détail à l'aide de combinaisons de différentes variables. Le graphique « Niveau CO₂ par habitant en fonction de la population et par continent » compare trois variables. La population en Amérique du Nord est celle produisant le plus de CO₂ par habitant et se situe en troisième position. En comparaison, la population du continent asiatique, la plus importante, est celle produisant le moins de CO₂ par habitant après l'Afrique.

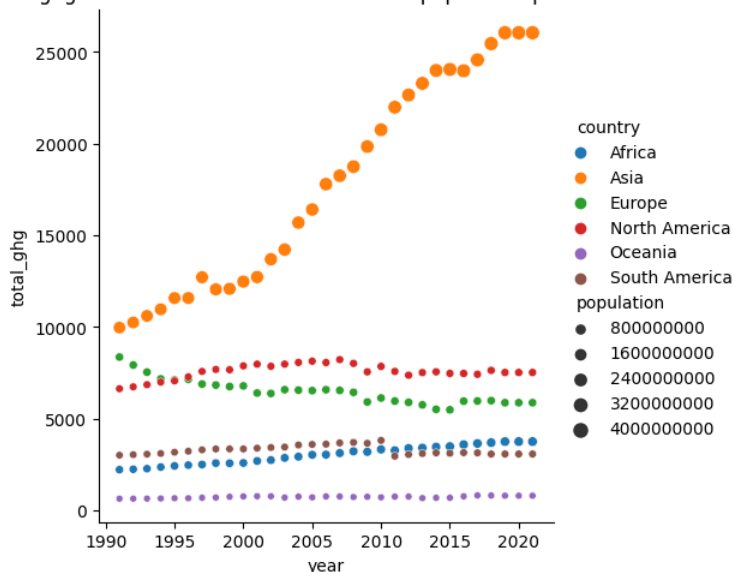
Niveau de Co2 par habitant en fonction de la population par continent



Le graphique suivant prend en compte quatre variables : les continents, les émissions totales de CO₂, les années et la population. Ce graphique est très intéressant car il nous permet de voir l'évolution de la population et des émissions de CO₂ dans le temps. Contrairement au graphique précédent, on peut constater que le poids de la population impact la production totale de CO₂ d'un continent. On constate en Asie une augmentation exponentielle de la population et de la production totale de CO₂ de 1990 à 2021. A contrario, l'Océanie représente la plus faible population ainsi que la plus faible production totale des émissions de CO₂.

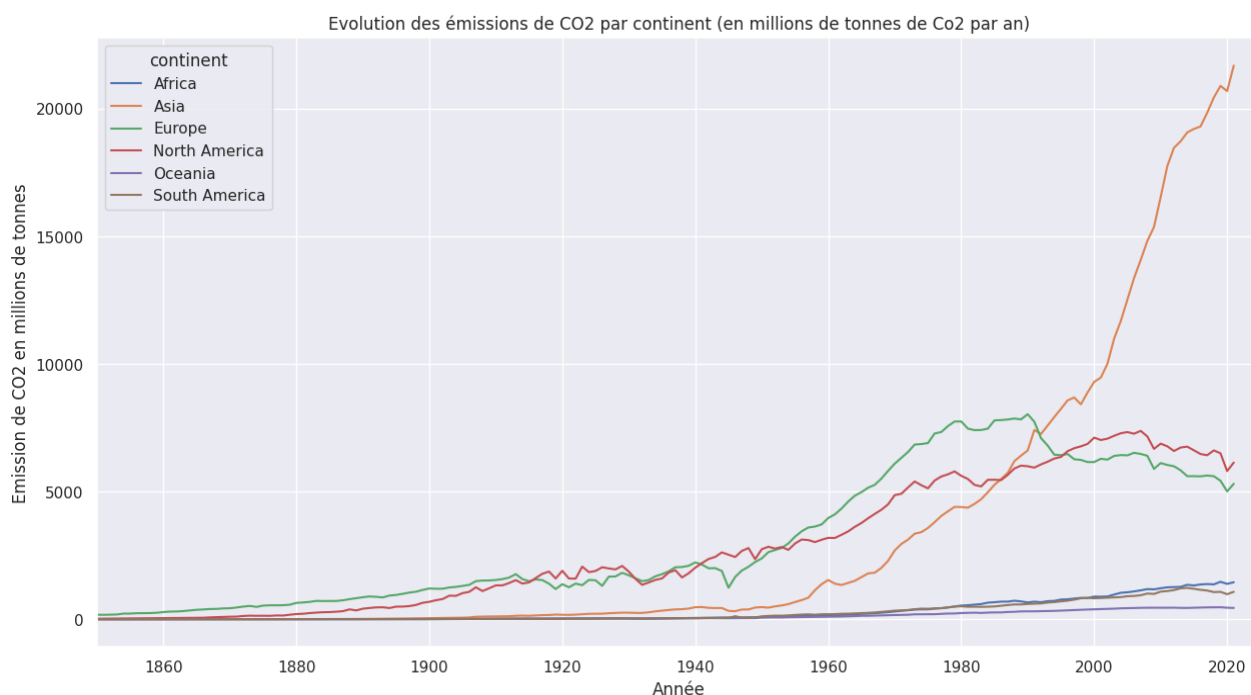


Total ghg de 1909 à 2022 en fonction de la population par continent



Sous un autre format, le constat reste le même et ne s'applique pas uniquement pour la production de CO₂. En effet lorsque l'on fait le total des hydrocarbures (**co2**, **cement_co2**, **coal_co2**, **flaring_co2**, **gas_co2**, **methane**, **nitrous_oxide**, **oil_co2**), le continent asiatique reste le premier producteur et l'Océanie le dernier.

Est-ce que cette production de CO₂ est linéaire dans le temps ? La différence de production de CO₂ entre les continents a-t-elle toujours été aussi disparate ? Dans le graphique suivant, nous pouvons voir



l'évolution des émissions de CO₂ de chaque continent de 1860 à 2020.



Nous pouvons observer une évolution croissante de la production de CO₂ vers 1880 (début de la révolution industrielle) notamment en Europe. La production de CO₂ continue de croître pour l'Amérique du Nord et l'Europe tout le long du XXe siècle et durant les deux guerres mondiales.

Une nouvelle évolution apparaît vers 1960 avec la croissance de la production de CO₂ en Asie. Cette période coïncide avec la décolonisation de l'Afrique et l'ouverture des marchés africains au continent asiatique et aussi l'évolution du multilatéralisme avec l'accroissement des échanges entre l'Asie et le reste du monde.

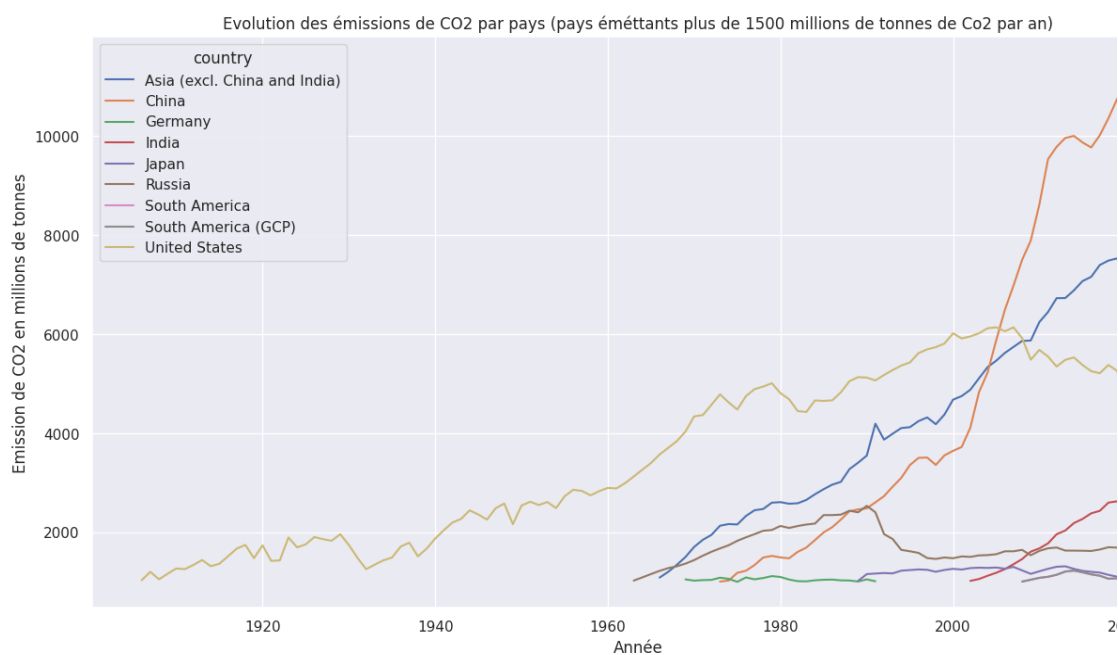
Au début des années 2000, une double évolution émerge. D'une part, une forte croissance de la production de CO₂ en Asie, dépassant de loin tous les autres continents, et d'autre part, une légère décroissance de la production de CO₂ pour l'Amérique du Nord et l'Europe.

Il est important de noter que la majorité de la production de CO₂ en Asie s'est réalisée sur une courte période (1960-2020 environ 60 ans).

Ce graphique ci-dessous nous permet de faire les observations suivantes :

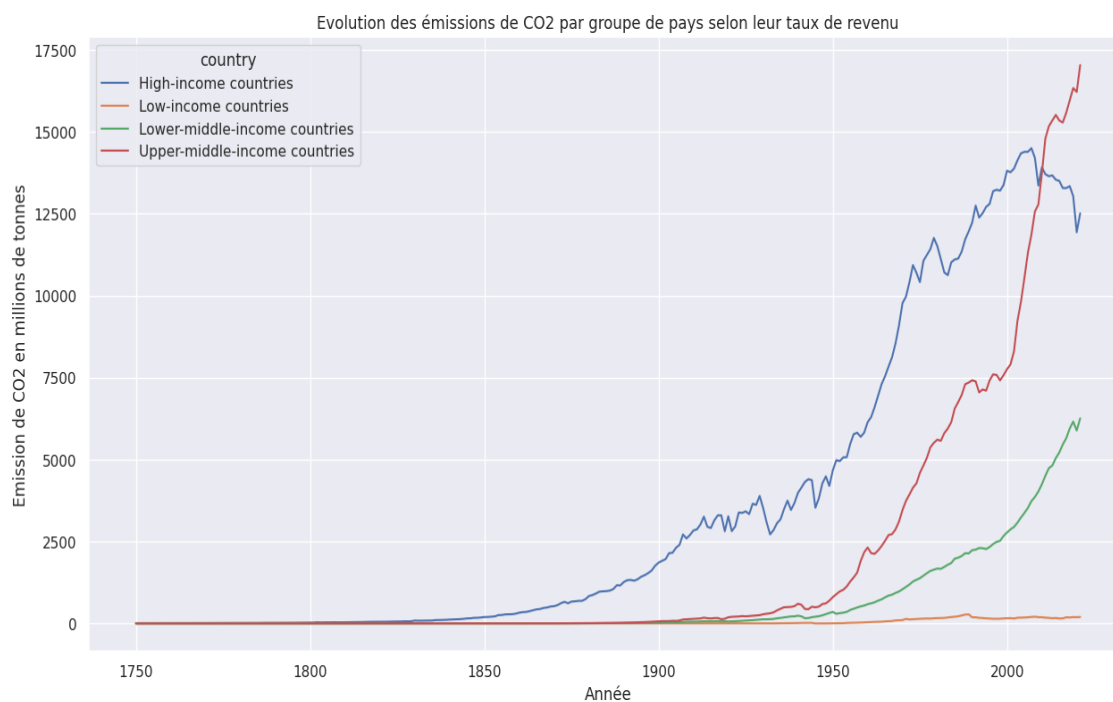
1. En 1910, les Etats-Unis produisent environ 1000 millions de tonnes CO₂ jusqu'à atteindre leur paroxysme dans les années 2005-2007 avant d'amorcer une baisse depuis une décennie (récession et prise de conscience des moyens de production).
2. Depuis le début des années 2000, le continent asiatique (hors Chine et Inde) est le plus grand émetteur de CO₂ au monde
3. La Chine produit deux fois plus de CO₂ que tout le continent asiatique réunit
4. Baisse significative de l'émission du CO₂ pour la Russie correspondant à l'effondrement du bloc de l'URSS à la fin des années 1990
5. Entre 2000 et 2020 la production de CO₂ est en augmentation pour l'Inde





Les événements historiques influent sur la production de CO₂ des pays concernés.

Enfin, il semble sans surprise que la production de CO₂ dépende étroitement de la richesse des pays et des continents étudiés. En effet, les pays à hauts revenus produisent plus de CO₂ que les pays à faibles revenus. Ce graphique nous permet déjà d'établir une relation de cause à effet entre deux facteurs.



L'impact des revenus sur la production de CO₂ n'est effectif qu'à partir des années 1850, où les pays à forts revenus se détachent de manière exponentielle du reste du monde. A préciser qu'une baisse s'effectue depuis les années 2005.

Pour les pays à revenus moyens et faibles, il faudra attendre les années 1950 pour voir une augmentation de leur production de CO₂.

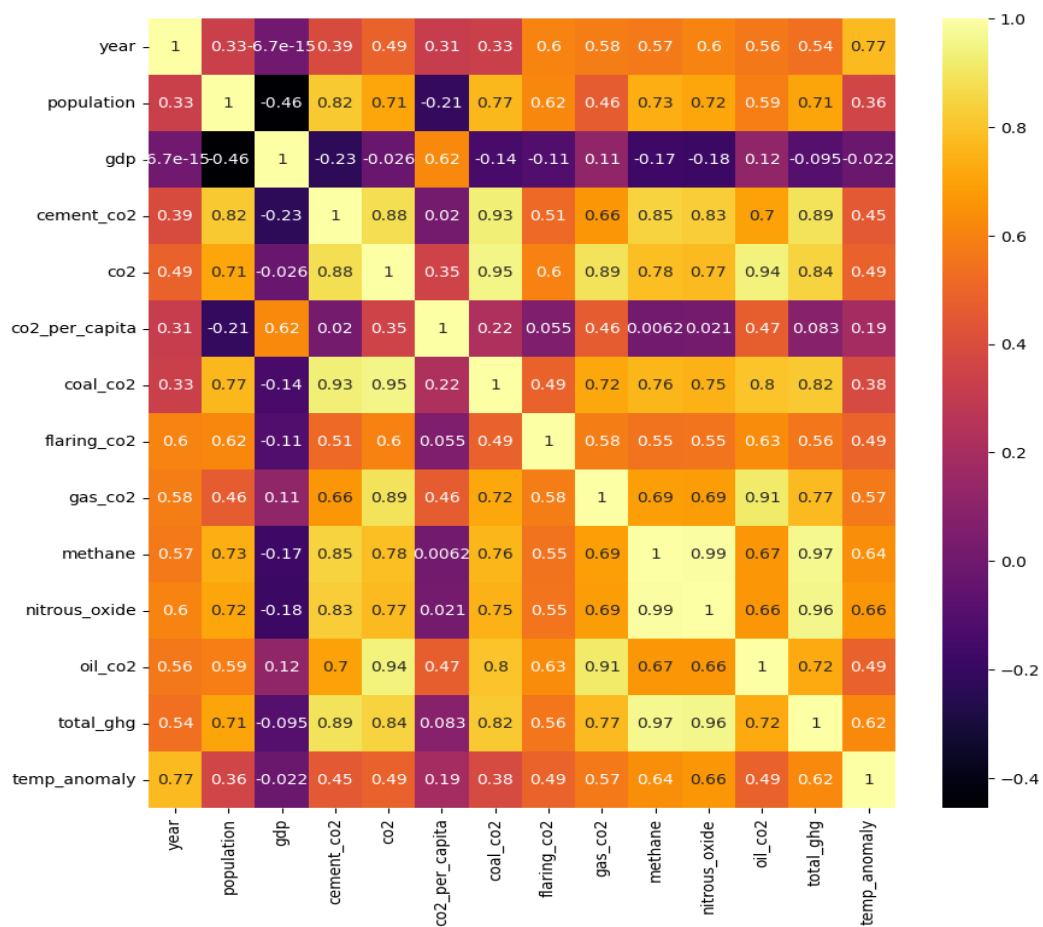
Concernant les pays à faibles revenus, la production de CO₂ reste relativement stable et en dessous des 1000 millions de tonnes.



Etudes des corrélations

Nous allons voir plus en détail les corrélations entre les différentes variables de nos jeux de données afin de comprendre l'évolution des températures terrestres à travers le globe.

Pour cela, nous avons choisi de générer une heatmap représentant les corrélations entre chaque variable numérique de notre jeu de données.



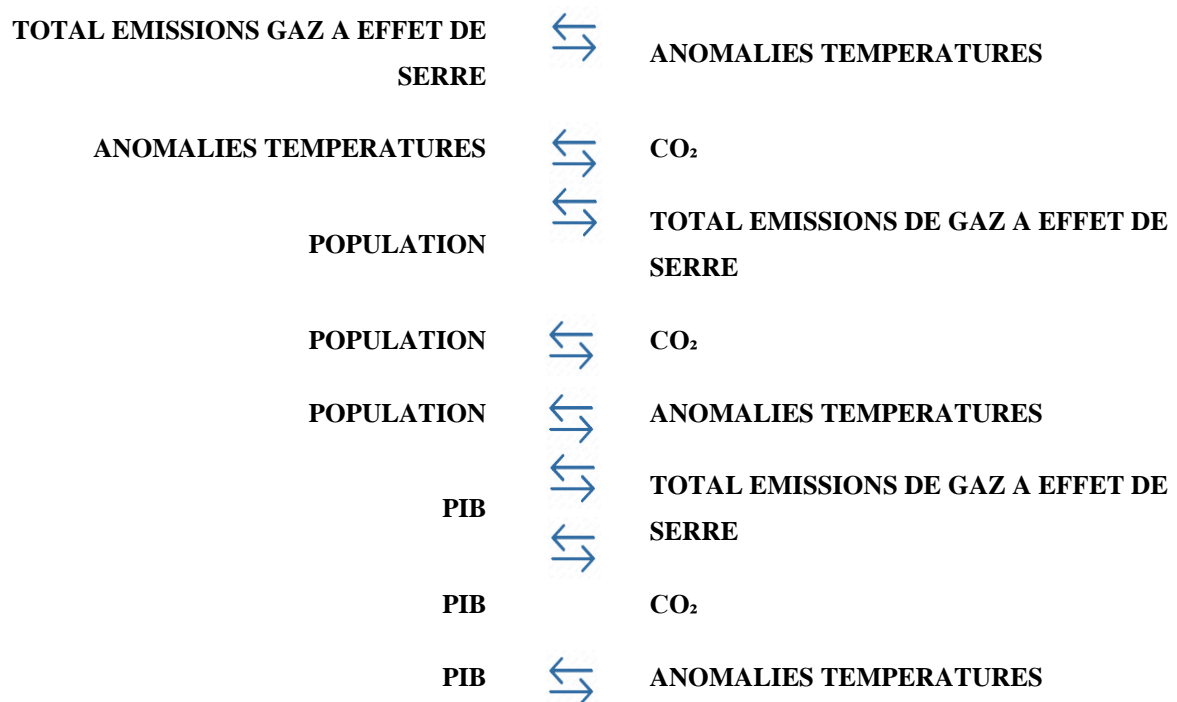
Nous pouvons déjà constater des corrélations notamment entre :

- le méthane et le total des émission de gaz à effet de serre (0.97)
- le protoxyde d'azote et le total des émission de gaz à effet de serre (0.96)
- le CO₂ et le CO₂ produit par charbon (0.95)
- le CO₂ et le CO₂ produit par pétrole (0.94)



- le CO₂ produit par ciment et le CO₂ produit par charbon (0.93)
- le CO₂ produit par ciment et le total des émissions de gaz à effet de serre (0.89)
- le CO₂ et CO₂ produit par l'essence (0.89)
- le CO₂ produit par ciment et le méthane (0.85)
- la population et le CO₂ produit par le ciment (0.82)

Par exemple, le test Pearson réalisé entre le total des émissions de gaz à effet de serre et les anomalies de températures nous permet de conclure une corrélation entre ces deux variables (soit l'hypothèse 1) avec une p-value $1.4115724450324765e-73$ et un coefficient, équivalent à 0.6232510726731024, évaluant l'intensité de la corrélation. A la suite de tests statistiques réalisés nous pouvons observer des corrélations entre les facteurs suivants :



De plus, à l'aide d'un graphique 'quantile-quantile' nous avons pu analyser la distribution des différentes variables selon les pays. Il est intéressant de noter que, pour la variable des anomalies de températures, la distribution est différente selon les pays, ce qui sous-entend que les anomalies de températures ne sont pas homogènes sur le globe même si une linéarité demeure.



Nettoyage et Pre-processing

Le dataframe, à l'exception de la colonne 'iso_code', ne contient aucune valeur manquante. Dans le processus de nettoyage, cette colonne a été supprimée.

Le tableau statistique généré par la fonction 'describe' nous a permis d'inspecter globalement nos données. Nous pouvons noter un écart important entre les troisièmes quartiles et les valeurs maximales, notamment pour la colonne 'CO₂' ou 'méthane'.

Nous pouvons constater de grands écarts entre les valeurs minimales et maximales, ce qui coïncide par exemple avec les importantes disparités entre l'Asie et l'Océanie.

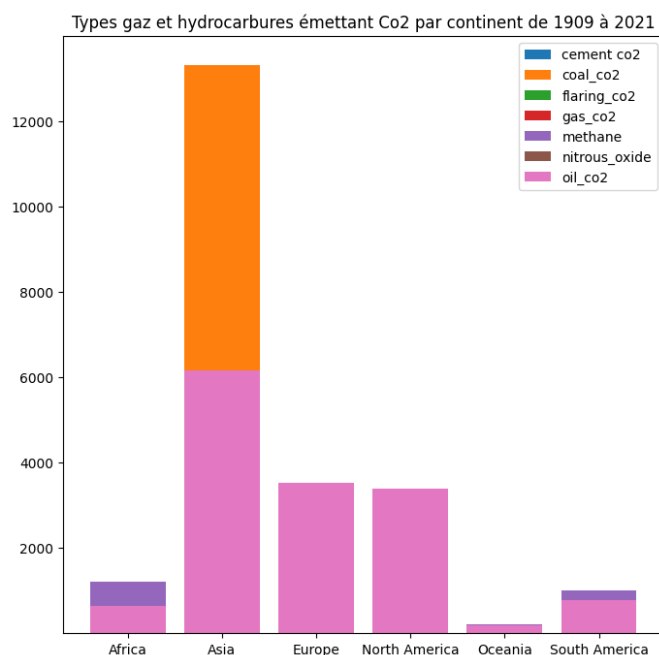
Il est intéressant de noter, qu'il y a une différence importante entre la valeur minimum (0.109) et la valeur maximale (16.295) concernant le CO₂ par habitant. Ce tableau permet de constater de fortes disparités dans nos données que nous avons également pu voir lors de mise en forme des données, notamment entre les continents.

	year	population	gdp	cement_co2	co2	co2_per_capita	coal_co2
count	672.000000	6.720000e+02	6.720000e+02	672.000000	672.000000	672.000000	672.000000
mean	1965.500000	6.625490e+08	4.133415e+11	67.061670	2420.925052	5.075699	1100.131955
std	32.354409	9.305657e+08	3.652822e+11	181.130431	3550.164468	4.824895	1764.845624
min	1910.000000	6.992770e+06	3.329652e+10	0.000000	10.867000	0.109000	9.178000
25%	1937.750000	1.418069e+08	6.759497e+10	1.411250	169.279750	1.013000	65.962000
50%	1965.500000	3.645779e+08	3.225397e+11	13.115500	821.121500	3.114000	302.271000
75%	1965.500001	7.248193e+08	6.732722e+11	53.044250	3600.495000	9.794500	1596.458500
max	1965.500002	4.693332e+09	1.060806e+12	1358.534000	21688.988000	16.295000	11959.011000

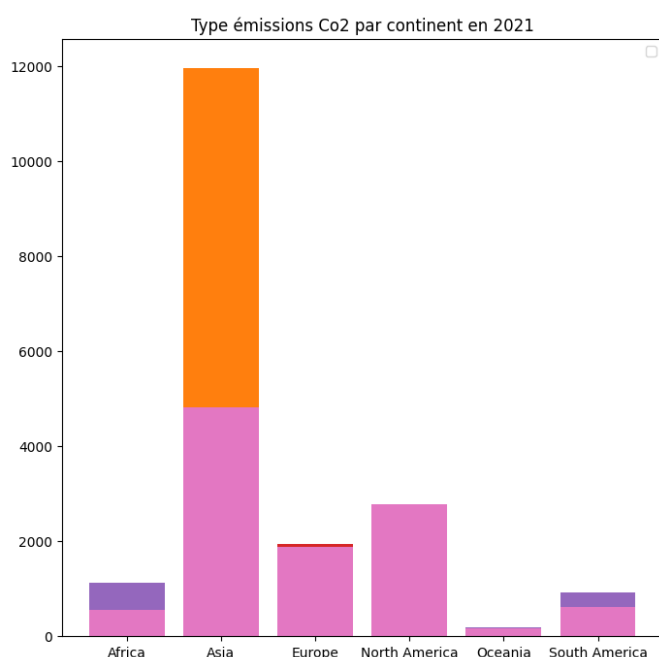
	flaring_co2	gas_co2	methane	nitrous_oxide	oil_co2	total_ghg	temp_anomaly
count	672.000000	672.000000	672.000000	672.000000	672.000000	672.000000	672.000000
mean	27.461726	377.609930	373.655030	136.121190	834.266853	1932.247597	0.202113
std	38.282491	645.793909	753.232621	258.916619	1160.619860	4270.069978	0.583740
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.010000	-1.290000
25%	0.000000	0.930250	4.540000	3.400000	42.272000	26.980000	-0.212500
50%	7.924500	46.611500	80.220000	18.610000	199.254000	183.370000	0.080000
75%	45.944750	305.948500	186.977500	100.575000	1436.954250	757.200000	0.542500
max	230.375000	3242.854000	3877.760000	1337.940000	4806.574000	26048.539000	2.220000



A l'aide d'un histogramme, voici la proportion des différents types de gaz et d'hydrocarbures en fonction des 6 continents. Ce graphique permet d'illustrer la dominance du charbon et du pétrole (et à moindre mesure le méthane) dans la production de CO₂ de 1909 à 2021. De manière évidente, l'Asie reste nettement au-dessus de tous les autres continents avec une forte utilisation du charbon.

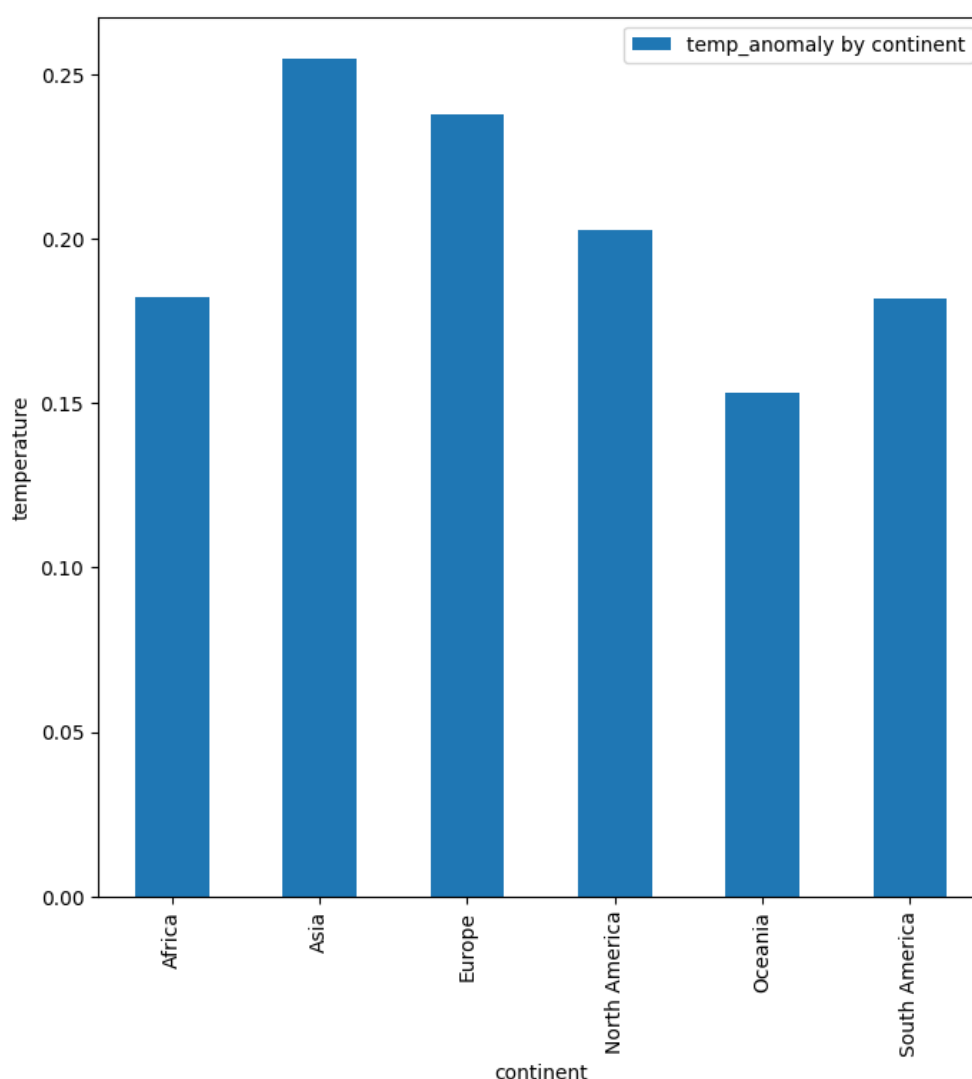


Si on se concentre sur la situation actuelle en prenant uniquement l'année 2021, on constate que le graphique reste relativement le même avec tout de même l'apparition de la production de gaz en Europe.



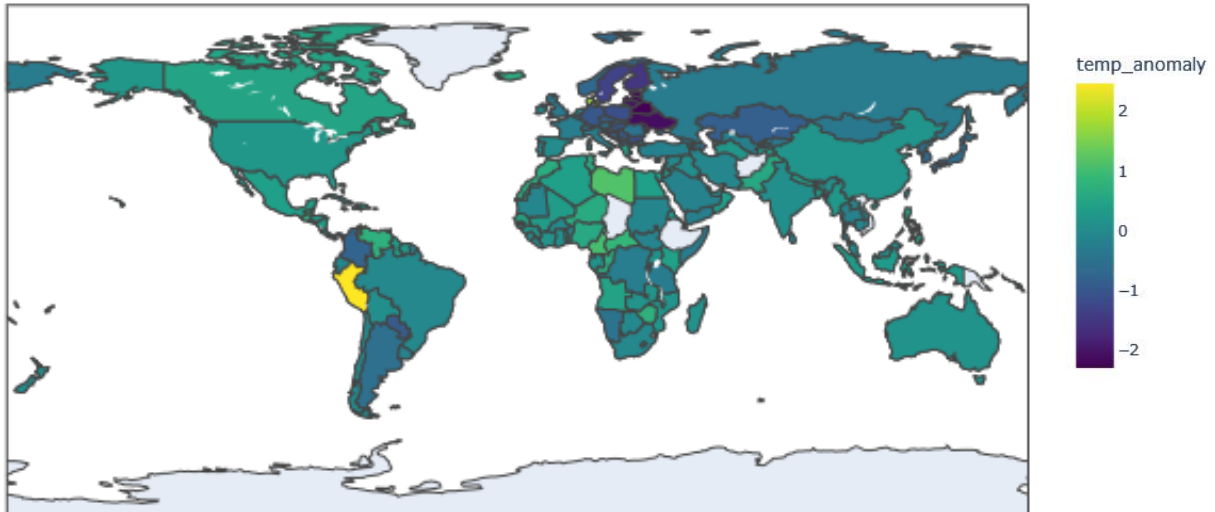
Concernant les anomalies de températures (variable cible), nous avons effectué une moyenne des anomalies par continent car nous savons que la production de CO₂ est extrêmement disparate au fil des années en fonction des continents. Cela nous permet de voir sur toute la période de 1909 à 2021, où se situe chaque continent.

En effet, l'Asie reste en tête mais n'essuie pas une différence extrême avec l'Europe et l'Amérique du Nord. Contrairement aux analyses effectuées sur les variables explicatives à un moment donné, ici nous pouvons voir que les anomalies de températures entre chaque continent sont relativement proportionnelles entre elles. L'Asie enregistre une moyenne au-dessus de 0,25°C contre l'Océanie avec la moyenne la plus faible à environ 0,15°C.



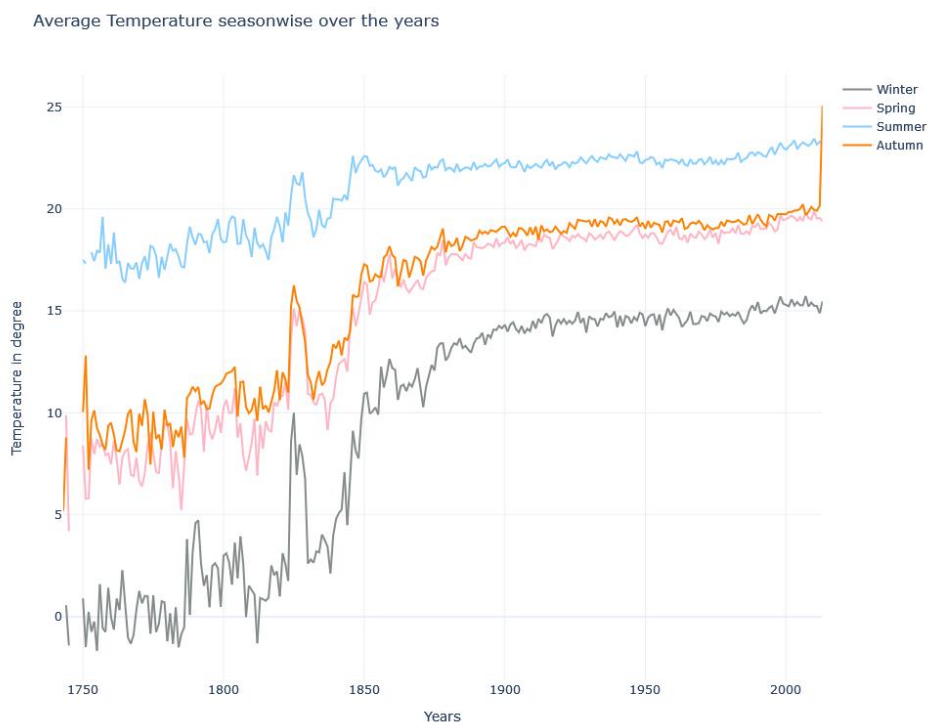
Ci-dessous la carte présente également la moyenne des anomalies de températures. Au même titre que pour l'histogramme, on voit que les continents sont relativement de la même couleur. A noter que la carte prend en compte l'année 1942 tandis que l'histogramme présente les moyennes globales de 1909 à 2021. Il est donc intéressant de noter que cette tendance ne s'applique pas que sur un moment précis

Anomaly des temperature moyenne par pays



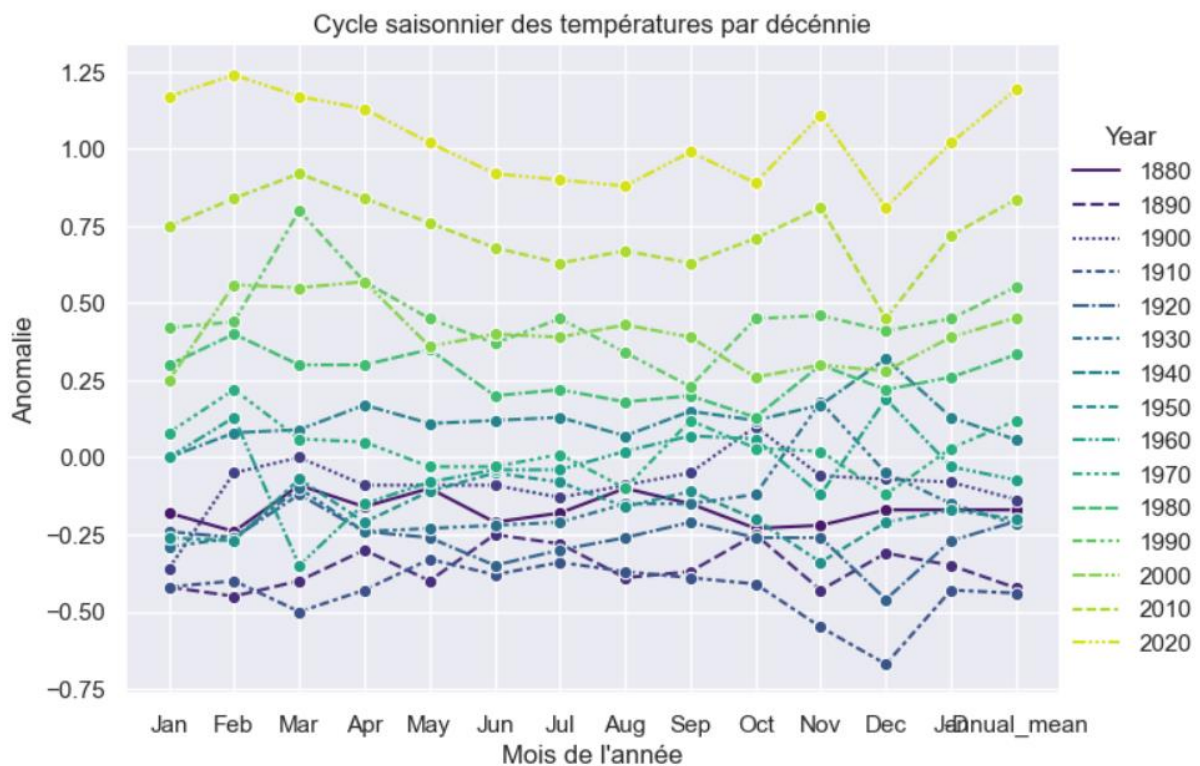
mais sur plusieurs décennies.

Ce graphique permet d'illustrer l'évolution des températures de 1750 à 2000 en fonction des saisons. Ainsi, nous comprenons pourquoi le changement climatique se traduit par un réchauffement climatique. En effet, la saison ayant subi l'augmentation de température la plus importante est l'hiver.



Par exemple, nous pouvons voir avec le graphique « Cycle saisonnier des températures par décennie » que, de 1880 à 2020, la température terrestre moyenne au mois de décembre est passée de -0.70 à +1.20 soit une augmentation de 0.5. Tous les mois de l'année enregistrent une augmentation considérable entre 1880 et 2020.

Notons une stabilité relative entre les années 1880 et 1930 afin d'observer une augmentation rapide à partir de 1980. De plus, l'année 2020 relate des anomalies de températures plus importantes sur les mois de décembre, janvier et février que sur le reste des mois de l'année. La plus haute anomalie de température illustrée sur ce graphique est le mois de février de l'année 1990, avec une anomalie au-dessus de 0,75 contre 0,25 pour le mois de septembre de la même année.



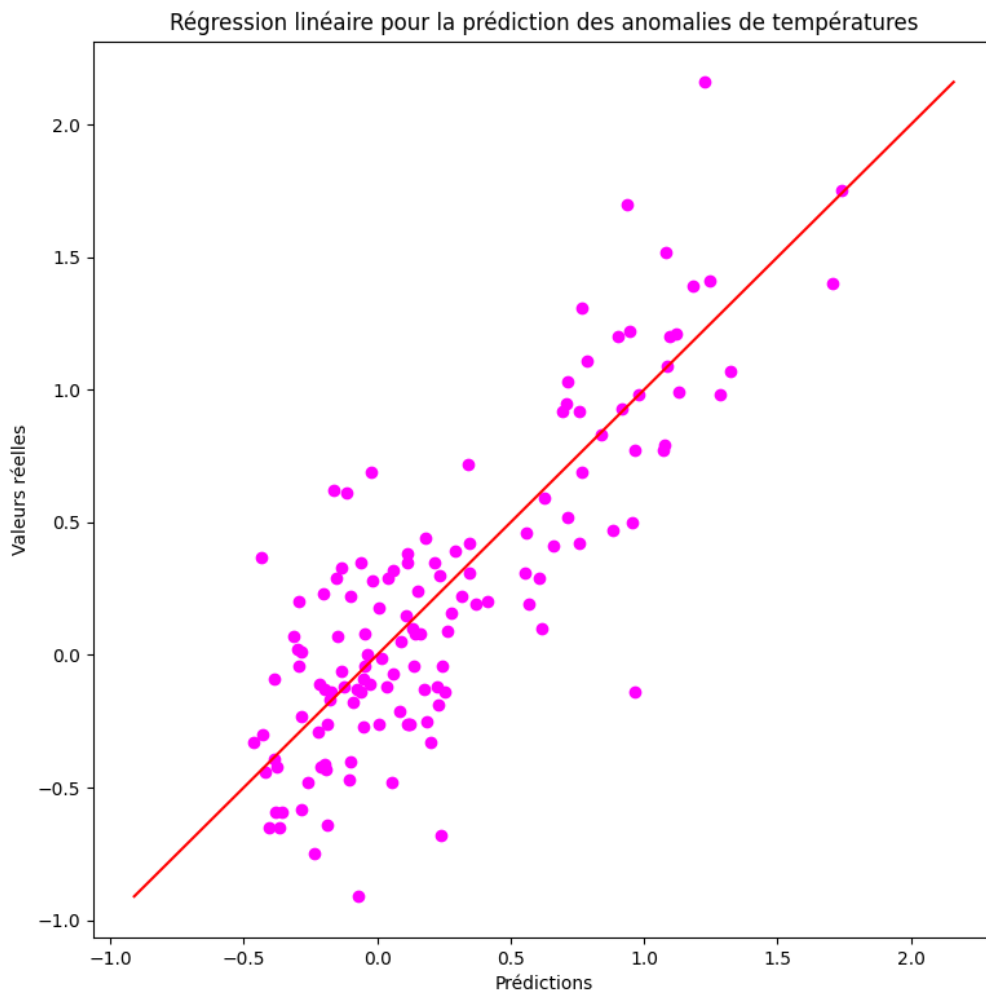
A la fin de cette étape de nettoyage et de pre-processing de nos données, nous pouvons déjà observer que les températures ont une tendance prédominante : augmenter.

Au même titre, les valeurs de nos variables explicatives augmentent également au fil des années.



Dans le respect de l'objectif du projet, nous avons établi la variable « température anomalies » comme la variable cible et toutes les autres variables comme les variables explicatives.

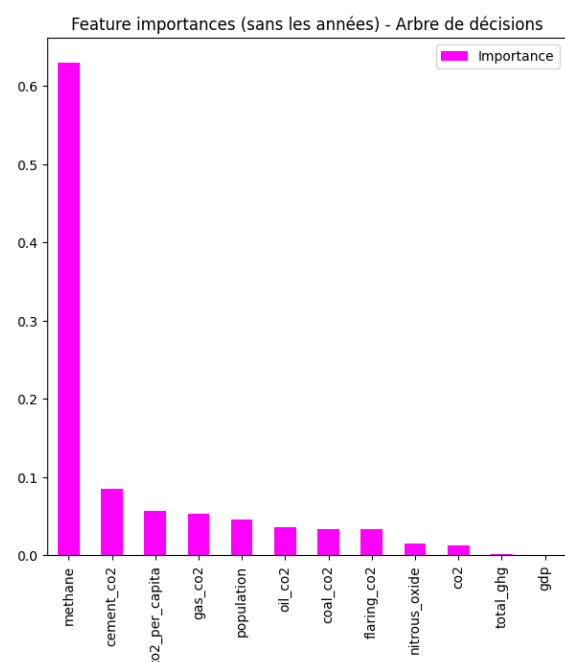
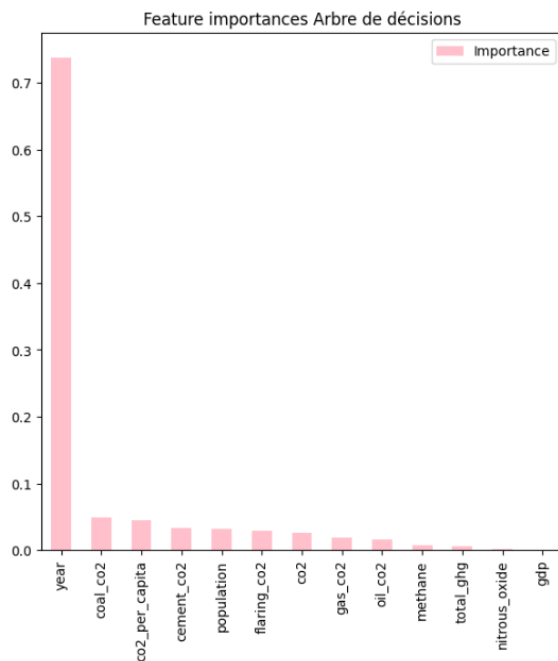
En premier lieu nous nous sommes appuyés sur un modèle de régression linéaire afin d'évaluer une relation avec la variable cible et les variables explicatives et être en capacité, ou non, de prédire y par rapport à nos valeurs réelles (x). Dans ce cas bien précis, nous voyons que le comportement des valeurs pour la prédiction des anomalies de températures suit fonction affine.



Nous avons poursuivi la modélisation de nos données à travers quatre modèles :

ARBRE DE DECISION :

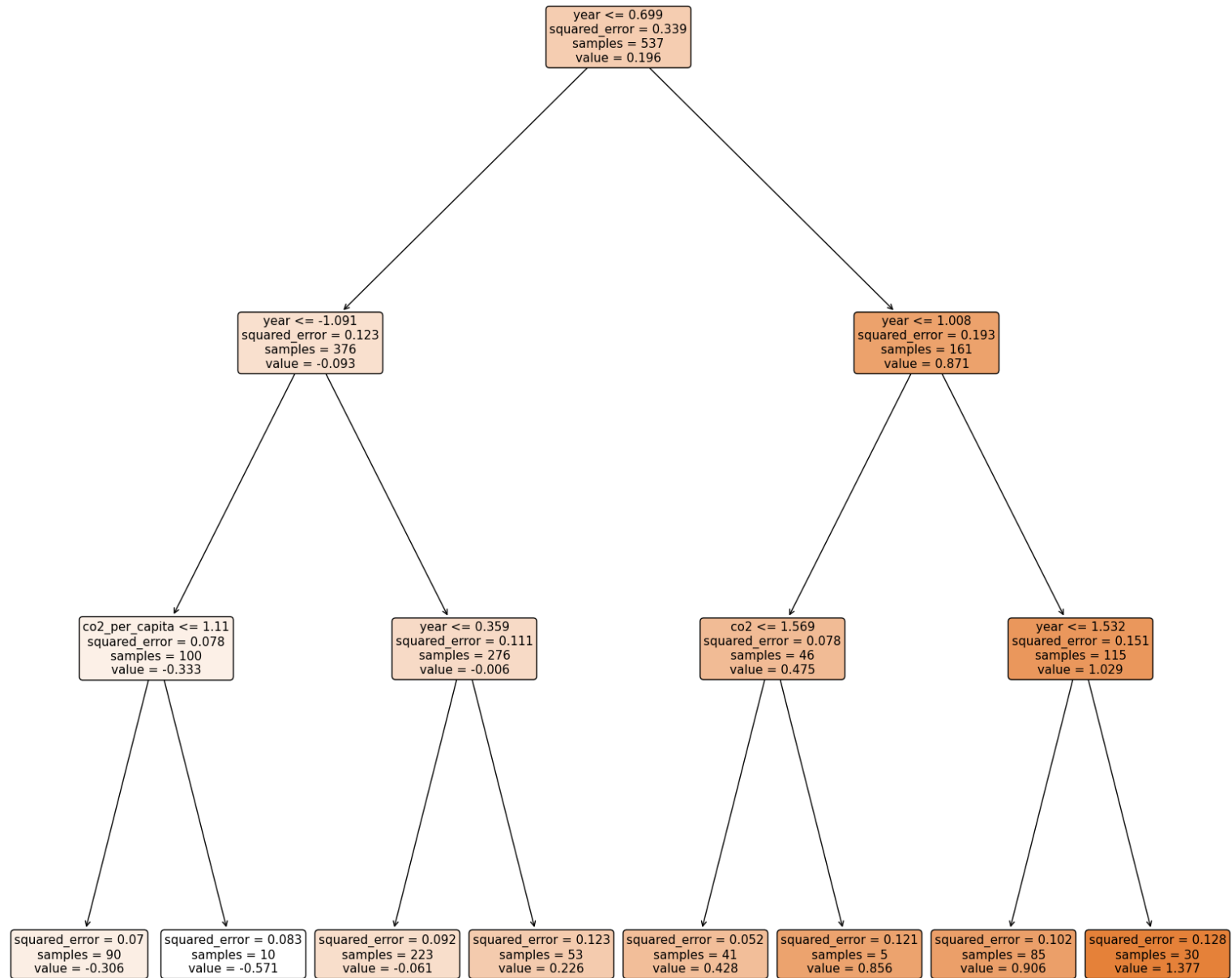
- Score ensemble train = 1.0
- Score ensemble test = 0.5715743416280948
- Graphique du classement de l'importance de toutes les variables explicatives :
variable la plus importante « year ». En retirant la variable « year », le méthane devient la variable la plus importante.



Ci-dessous, le graphique de l'arbre est présenté sur trois niveaux. Nous pouvons également constater à l'aide de ce graphique l'importance de la variable « year » que l'on retrouve jusqu'au troisième nœud. Il est très clair que l'année est décisive dans le choix des possibilités de prédictions.

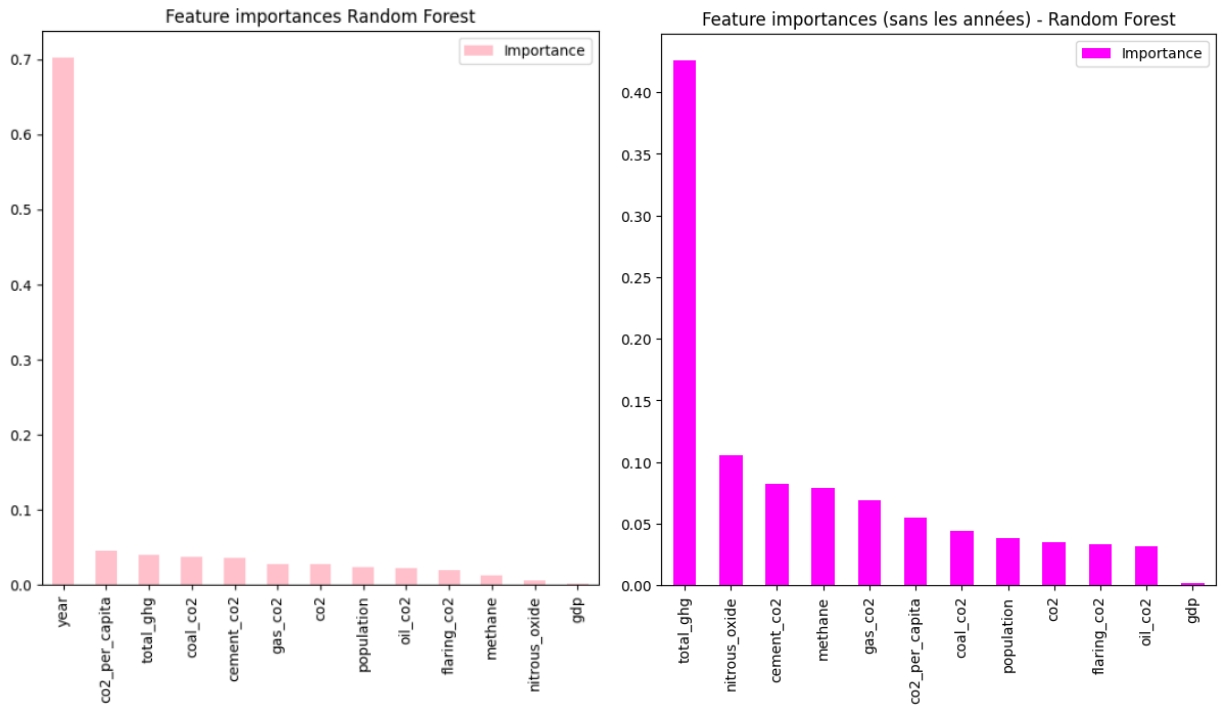
Au niveau du deuxième nœud, les variables CO₂ et CO₂ par habitant sont prises en compte dans le calcul des prédictions. L'intensité de la couleur dépend de la valeur de la MSE, plus elle est élevée plus les couleurs sont intenses. Lorsque la variable CO₂ est prise en compte on remarque une baisse de la précision du modèle par rapport à la variable CO₂ par habitant. Le meilleur résultat s'obtient lorsque les variables « year » et « CO₂ per capita » sont prises en compte avec une MSE égale à 0.083.





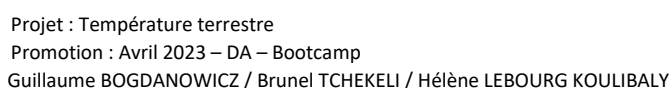
LA FORET ALEATOIRE :

- Score ensemble train = 0.956191893697089
- Score ensemble test = 0.7204673038466303
- Graphique du classement de l'importance des variables explicatives : variable la plus importante « year ». En retirant la variable « year », la production totale de gaz à effet de serre devient la variable la plus importante.

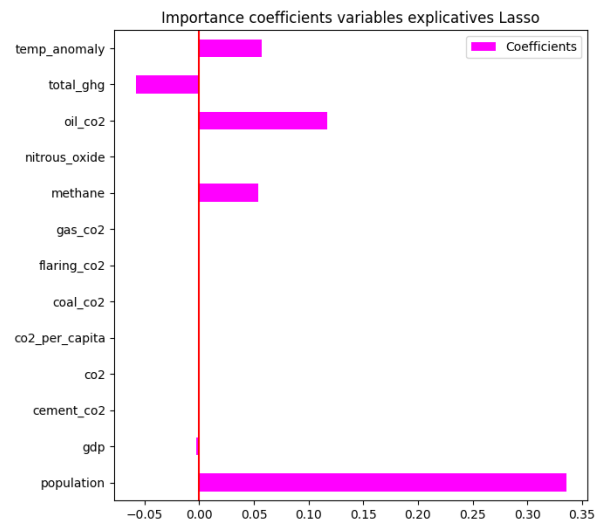
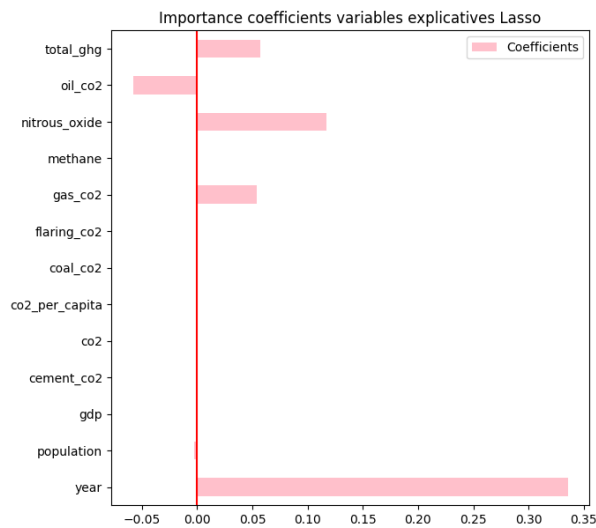


REGRESSION LASSO :

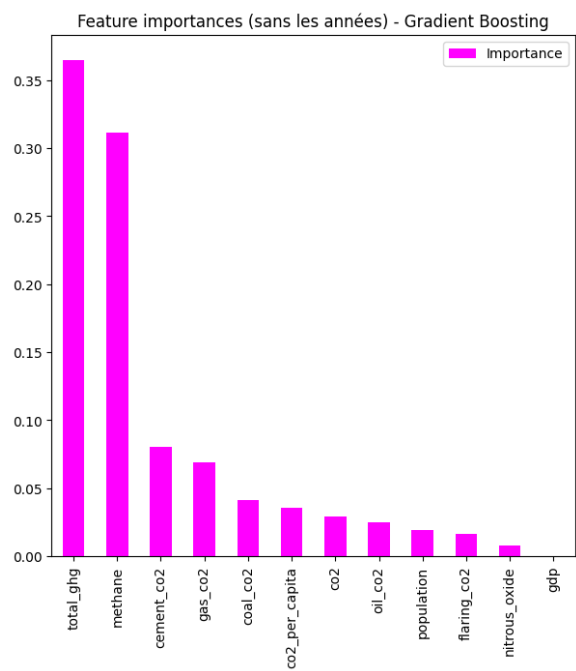
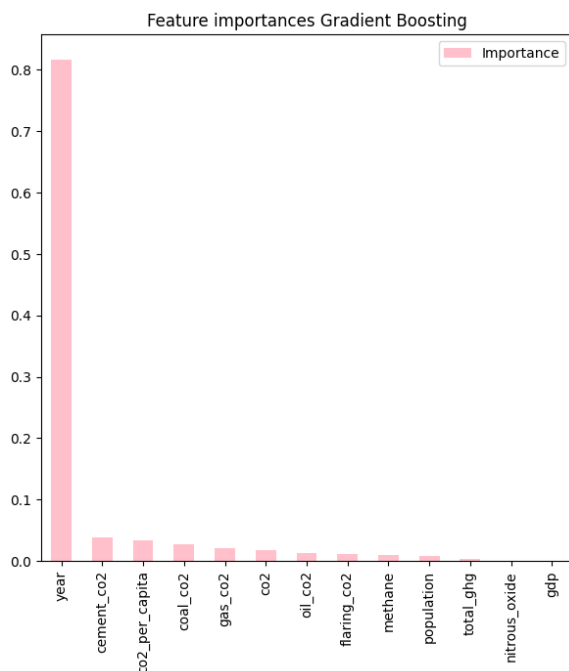
- Score ensemble train = 0.6695170836184214
- Score ensemble test = 0.669806081309698
- Graphique calculant les coefficients des variables explicatives. Coefficient le plus élevé est celui de la variable « year ». En retirant la variable « year », le coefficient de la variable « population » devient le plus élevé.



GRADIENT BOOSTING :



- Score ensemble train = 0.8791403341726561
- Score ensemble test = 0.746568463488195
- Graphique du classement de l'importance des variables explicatives : variable la plus importante « year ». En retirant la variable « year », la production totale de gaz à effet de serre devient la variable la plus importante.



A noter que tous les modèles entraînés font ressortir la variable « années » comme étant la variable la plus importante. En effet, plus les années passent, plus les anomalies de températures augmentent. Cependant, le temps n'en est pas la conséquence. si nous décidons de retirer cette variable nous pouvons noter que les variables ayant le plus d'impact sur les anomalies de températures sont « total_ghg » et « méthane ».

Nous avons choisi d'évaluer ces modèles en fonction de leurs RMSE puisque nos valeurs s'accordent sur un modèle de régression linéaire comme nous l'avons vu précédemment. De plus, cette métrique s'exprimant dans la même unité que notre variable à prédire ('température anomalie'), son résultat est d'autant plus simple à interpréter.

Après l'entraînement de ces modèles et du résultat leurs RMSE), le modèle Gradient Boosting semble être le modèle le plus performant puisque sa RMSE valide l'écart le plus faible.

	MAE train	MAE test	MSE train	MSE test	RMSE train	RMSE test
Decision Tree	0.000000	0.308000	0.000000	0.147930	0.000000	0.384617
Random Forest	0.092556	0.241364	0.014842	0.096519	0.121827	0.310676
Lasso	0.258821	0.260577	0.111965	0.114012	0.334612	0.337657
Gradient Boosting	0.160310	0.227479	0.040946	0.087507	0.202352	0.295816

Au vu des résultats obtenus à la suite de cette étape de modélisation, nous sommes conscients des limites des données prises en compte dans notre projet. Il nous faudrait un grand nombre de caractéristiques afin d'être en capacité d'établir quelles sont les conséquences précises des anomalies de températures.

Dans notre projet, un modèle de machine Learning n'est donc pas forcément nécessaire puisque nos données suivent un modèle linéaire temporel. Ainsi, nous pourrions faire une prédiction pour 2050 à l'aide d'un type de régression linéaire.

Prédictions

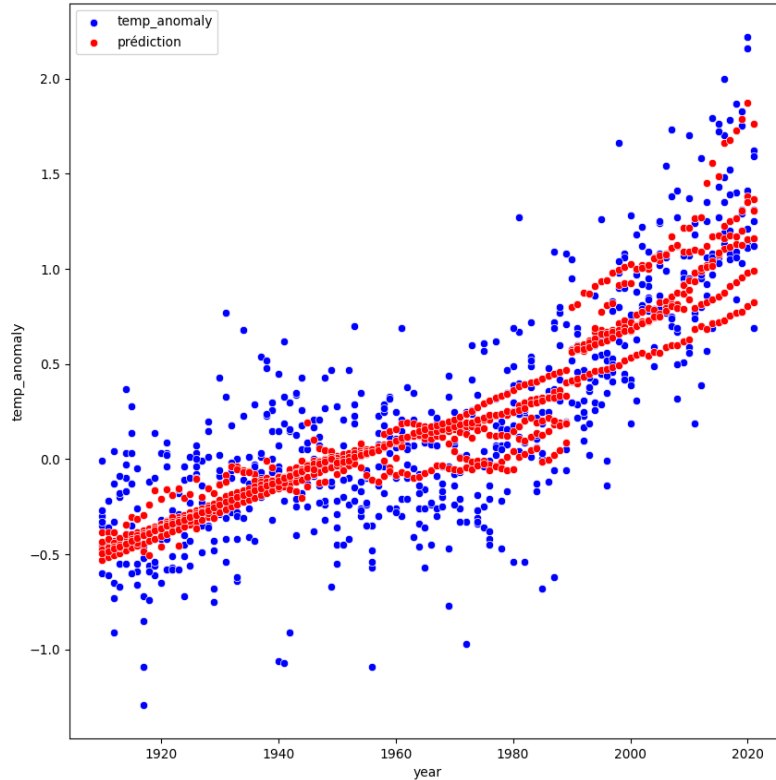
Nous avons décidé de confronter notre modèle de régression linéaire à l'exercice des prédictions d'anomalies de température pour les années futures, entre aujourd'hui et 2060.

(Pour plus de détails sur le code, se reporter au notebook 'Datas prédictions')

Tout d'abord, nous allons afficher les anomalies réelles et celles prédites par le modèle, entre 1910 et 2021

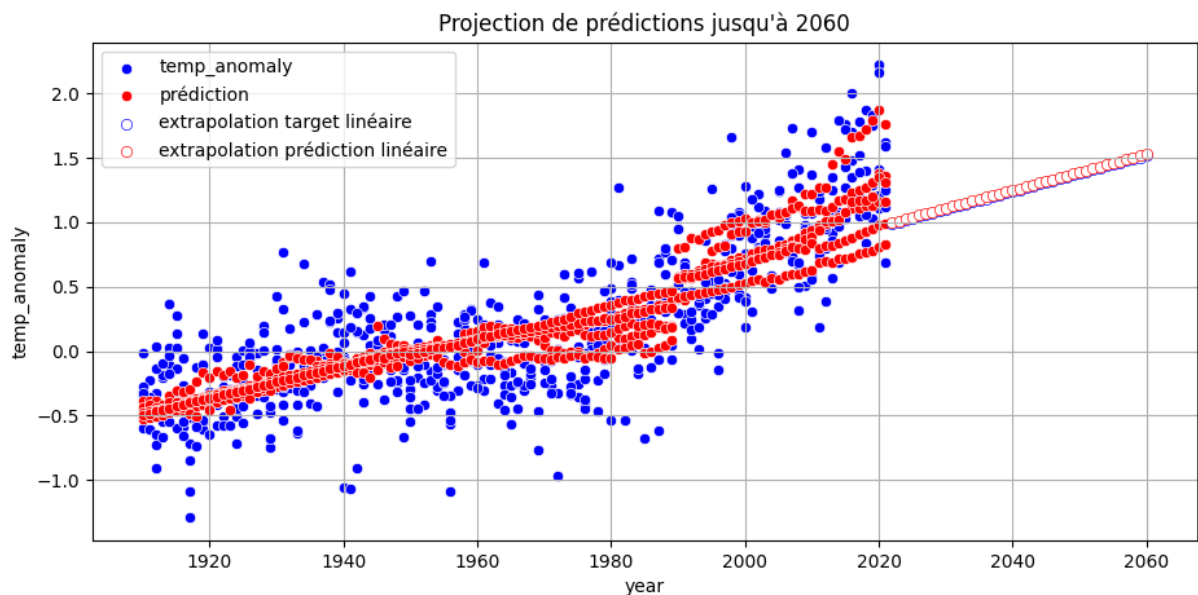


Comparaison entre les anomalies réelles de température et les prédictions du modèle de régression linéaire



On constate un décrochage à la hausse des prédictions, à partir de 1990. Il est probable que certaines variables (méthane, oxyde nitreux) qui n'ont pas de mesures avant cette date, aient une influence ici.

En créant une variable contenant les années futures (soit entre 2022 et 2060), on peut appliquer la fonction issue du modèle de régression et l'appliquer à cette fonction. Nous pouvons ainsi extrapoler les anomalies futures.



Les prédictions suivent bien la tendance générale à la hausse, avec des anomalies de température dépassant un degré et demi d'ici 2060.

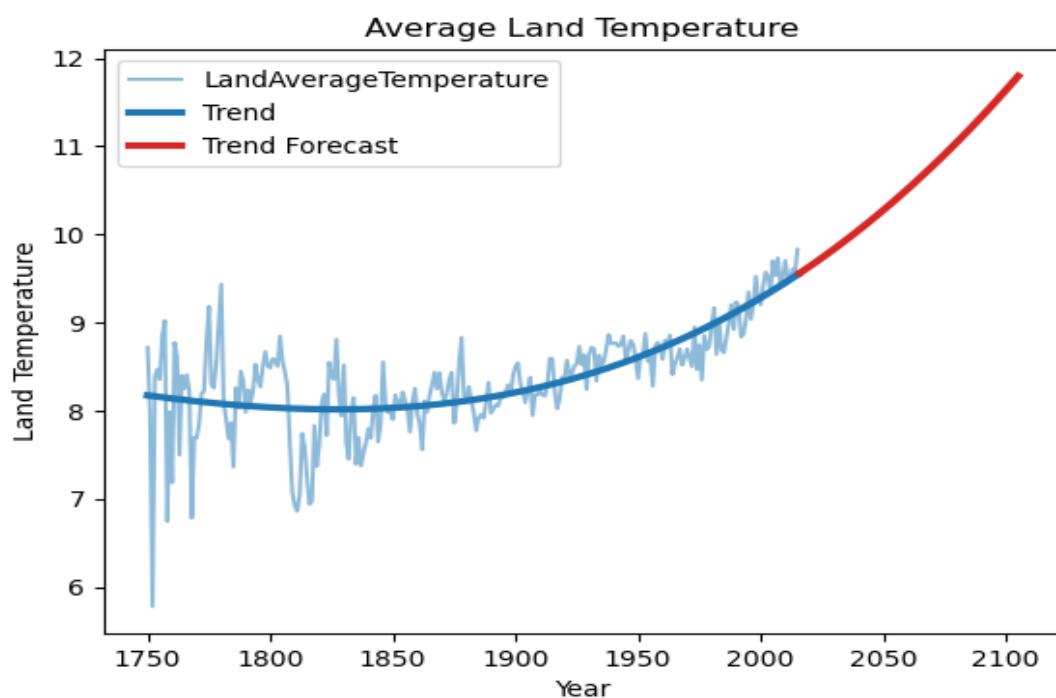
Bien entendu, ce modèle a ses limites ! Avec plus de temps, nous aurions pu étudier des prédictions avec d'autres modèles pour affiner nos observations.

Prédire la température de la France

Après avoir essayé toutes les méthodes ci-dessus, nous nous demandés s'il n'existe pas un modèle plus simple pour prédire la hausse des températures ?

Prédire la température globale nécessite de fournir un jeu de données global de tous les pays au modèle d'analyse. Nous avons dans un premier temps appliqué une régression linéaire sur le jeu de donnée GlobalTemperatures.csv qui reprend les températures globales de la surface terrestre.

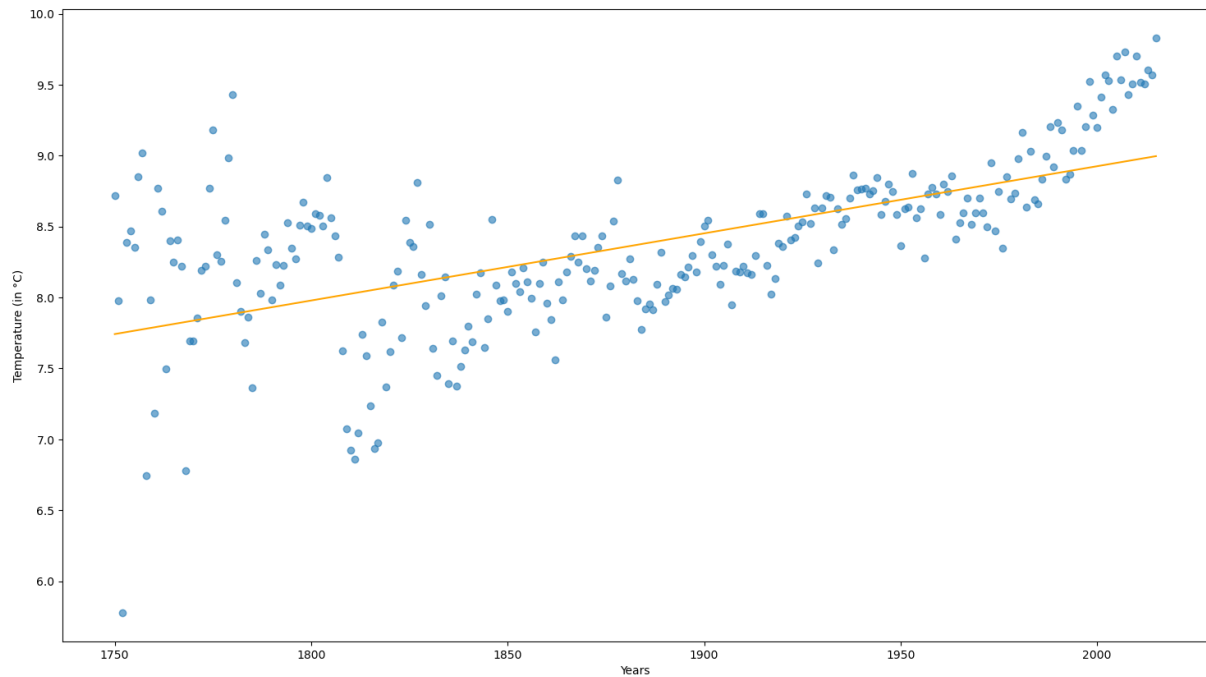
Nous appliquons une technique de diffusion des données (forecasting)



Nous souhaitons nous limiter à la prédiction de la température de la surface terrestre en France. Pour cela, Nous reprenons le data set GlobalLandTemperaturesByCountry.csv et isolons le pays France.

Par la suite nous appliquons simplement des modèles de régression polynomiale sur le jeu de données après avoir appliqué la méthode de forecasting.





Après prédiction sur ce data set, nous faisons l'interprétation suivante :

Chaque année, la température moyenne des surfaces terrestres augmente en moyenne de 0.004°C . Tous les dix ans, la température moyenne des terres augmente en moyenne de 0.047°C . La température moyenne des terres en 2030 sera de 9.06°C et en 2050 de 9.16°C .

Le modèle appliqué sur le jeu de donnée isolé de la France permet de faire les prédictions suivantes :

Coefficients: $1.25376118\text{e-}05$

Mean squared error: 25.13

Variance score : -0.01

- Température prédite pour la France en 2030 : 13.402°
- Température prédite pour la France en 2050 : 13.493°
- Température prédite pour la France en 2100 : 13.722°



Retours d'expérience

Guillaume Bogdanowicz :

« Le projet que nous avons mené dans le cadre de la formation “Data Analyste” m’a permis très rapidement de m’investir dans un cas concret d’étude de données sur un sujet que, au final, je connaissais mal avant cette étude. L’analyse des données, les premiers graphiques réalisés avec Python m’ont permis rapidement d’explorer ce sujet passionnant en mettant à profit les connaissances que j’étais en train d’acquérir. Les graphes m’ont permis de constater rapidement l’évolution des températures, en regard avec de nombreux autres paramètres, et Python et le machine-learning m’ont fait prendre conscience de la part écrasante des activités humaines dans le dérèglement climatique. Je me baserai très probablement sur notre travail pour explorer par moi-même d’autres datasets ! »

Hélène Lebourg-Koulibaly :

« De la prise en main des données jusqu’à l’objectif final, ce projet fut très formateur car cela m’a permis de comprendre davantage ce qui était attendu d’un(e) data analyste.

Dans l’atteinte des objectifs, l’analyse des données et la modélisation ont constitué les parties les plus difficiles pour moi dans le temps imparti. J’ai beaucoup apprécié la partie data visualisation, tant dans la réalisation que dans les résultats obtenus. J’ai trouvé satisfaisant de pouvoir interpréter les données et d’être en mesure d’illustrer ce qu’elles représentent et ce qu’elles veulent dire. Pour finir, le travail en équipe m’a permis de voir d’autres méthodes et d’apprendre davantage sur les outils utilisés. »



Brunel Tchekeli :

« Lorsqu'il nous est demandé de faire un choix de sujet pour notre projet fil rouge, j'avais choisi le projet "température terrestre" comme deuxième choix. Finalement, je suis bien content d'avoir traité ce sujet car c'était un sujet pertinent et au cœur des problématiques actuelles. La curiosité qui m'animait en début de projet était de savoir s'il y a vraiment un réchauffement climatique puisque tout le monde en parle. J'étais curieux de voir comment certains facteurs peuvent influencer le réchauffement climatique et de confirmer ce phénomène social par mes propres analyses. A travers ce projet fil rouge, je me rends compte que pour traiter une problématique projet, il ne faut pas se limiter qu'aux données d'entreprise et que le Data analyste est obligé d'aller puiser des données que ce soit à travers des lectures d'articles ou des data sets pour comparer. Un projet n'est pas que de la visualisation des données ni des analyses statistiques mais aussi être capable d'aller chercher le point de vue des scientifiques en général. Ce projet m'a permis de mettre en pratique les compétences apprises jusque-là mais surtout d'aller loin et d'adapter ses connaissances aux besoins du projet. Maintenant que ce projet est terminé, je souhaiterais explorer plus profondément le sujet à travers d'autres data sets spécifiques. Je tiens à remercier notre mentor projet Tarik avec qui nous avons tenu des réunions fréquemment et qui nous fait des retours chaque semaine sur notre avancée. Un coucou spécial aux membres de projets qui ont été tout aussi efficaces. »



Conclusion

L'élaboration de notre projet nous permet d'avancer l'hypothèse que les anomalies des températures terrestres sont, sans équivoque, liées aux activités humaines et notamment aux émissions de CO₂ résultant des différentes industries et de l'exploitation des ressources pétrolières. Il est donc très probable que, tant que ces facteurs continueront d'augmenter, nous assisterons à une augmentation des anomalies de la température à la hausse.

Afin d'étudier davantage le phénomène du réchauffement climatique il serait intéressant d'avoir d'autres variables explicatives spécifiques. En effet, l'étude approfondie de nos données a relevé certaines limites et, si le temps nous le permettait, il serait très intéressant de connaître les subtilités du fonctionnement du dérèglement climatique terrestre grâce à d'autres données, notamment les températures océaniques ou d'autres données climatiques. Des données issues des moyens de transports utilisés par les hommes (navires, avions, voitures) seraient également utiles. Les activités humaines sont-elles bien les seules responsables du dérèglement que nous constatons ?

Ce projet nous a permis d'étudier une thématique que nous ne connaissions que de manière superficielle (de part sa couverture médiatique notamment) et également à apprendre, grâce à l'analyse des données, à observer visuellement l'évolution du phénomène observé, de poser des hypothèses et, pourquoi pas, d'établir une conclusion. D'autre part, nous avons su identifier les limites de nos ressources et dresser une liste de nos besoins pour une éventuelle analyse plus poussée.

Enfin, ce projet nous a permis de mettre en pratique les connaissances que nous avons acquises au cours de notre formation et d'apprendre concrètement ce qui est attendu d'un Data Analyst. Toutes les techniques que nous avons découvertes, comme la collecte et le nettoyage des données, le processing puis la data visualisation et la modélisation nous ont été indispensables pour mener notre projet à bien et conclure à des observations factuelles vérifiées.



Références, sources et annexes

- NASA - Global Land-Ocean Temperature Index in 0.01 degrees Celsius

https://data.giss.nasa.gov/gistemp/tabledata_v4/GLB.Ts+dSST.txt :

- Our World in Datas - Data on CO2 and Greenhouse Gas Emissions

<https://github.com/owid/co2-data> :

- Our World in Datas - Surface temperature anomaly

<https://ourworldindata.org/grapher/hadcrut-surface-temperature-anomaly> :

- NOAA Global Temperature Anomalies - Graphing Tool

[Global Temperature Anomalies - Graphing Tool | NOAA Climate.gov](#)

- GitHub - Global Land and Ocean-and-Land Temperatures

[GitHub - gindeleo/climate: Data analysis of earth surface temperature](#)

- Wikipédia - Anomalies de températures

[Temperature anomaly - Wikipedia](#)



Diagramme de GANTT

MODÈLE DE DIAGRAMME DE GANTT

Astuce de Smartsheet →

La chronologie visuelle d'un diagramme de Gantt donne un aperçu détaillé des tâches et des dépendances d'un projet.

TITRE DU PROJET	Température Terrestre	
EQUIPE PROJET	Brunel	
	Guillaume	
	Hélène	
	Tous	

NOM DE L'ENTREPRISE	DATA SCIENTEST
DATE	28/04/23

NUMERO	TITRE DE LA TÂCHE	PROPRIÉTAIRE DE LA TÂCHE	DATE LIMITE	TÂCHE TERMINÉE (EN %)	ETAPE 1					ETAPE 2					ETAPE 3					ETAPE 4					ETAPE 5					ETAPE 6					ETAPE 7																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
					SEMAINE 1					SEMAINE 2					SEMAINE 3					SEMAINE 4					SEMAINE 5					SEMAINE 6					SEMAINE 7																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
					L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V	L	M	M	J	V																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																									
1	Découverte données et projet																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																															



Température Terrestre

Cursus concerné : Data Analyst

Niveau de difficulté : 06/10

Description du projet :

Objectif : 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.

La source de données est celle de la NASA.

Ressources à consulter :

- **Données :**
 - <https://data.giss.nasa.gov/gistemp/>
 - <https://github.com/owid/co2-data>

Conditions de validation du projet :

- un **rapport** d'exploration, de data visualisation et de ~~pre-processing~~ des données ;
- un **rapport** de modélisation ;
- un **rapport** final et le **code** associé.

DataScientest.com
Agrément organisme de formation 11755665975
09 80 80 79 49
2 place de Barcelone, 75016 Paris

