Université Claude Bernard  Lyon 1

RAPPORT DE PROJET

Data Mining – Master 2 informatique

2023 – 2024

MERCIER LORIS
p1906860

LEFEBVRE Julien
p2105454

Code source : https://github.com/LorisMercier/DataMining

Responsable : Rémy CAZABET

Sommaire

Introduction

Au cours des dernières décennies, les changements climatiques ont suscité des préoccupations croissantes en raison de leurs répercussions à l'échelle planétaire. A travers notre étude utilisant des techniques d'exploration de données, nous nous focaliserons principalement sur l'étude climatique de la planète, mettant en évidence l'impact du réchauffement climatique. Fondé sur le nettoyage et l'enrichissement des données, notre premier volet vise à discerner les tendances climatiques générales à l'échelle mondiale. Les résultats et analyses qui suivent dévoilent une compréhension plus approfondie du climat, englobant ses caractéristiques à l'échelle planétaire jusqu'aux nuances régionales et locales en passant par l'exploration des variations saisonnières présentes au sein de notre belle planète. En examinant enfin quelques températures anormales vis-à-vis des positions géographiques de chaque site, nous tenterons d'amener des éléments de compréhension autres que géographique pour expliquer les différents climats mondiaux. Notre recherche s'aligne et s'appuie alors sur sur les préoccupations mondiales soulevées par le Groupe d'experts Intergouvernemental sur l'Evolution du climat (GIEC).

1. Traitement des données

L'étape de traitement des données revêt d'une importance cruciale dans l'analyse de toute information. Dans notre situation particulière, cette phase s'est avérée particulièrement complexe en raison de la présence significative de données aberrantes et manquantes. Afin de naviguer efficacement à travers ces sources d'information, nous avons structuré notre approche en deux volets distincts : le nettoyage des données et l'ajout d'informations complémentaires

a. Nettoyage des données

Le processus de traitement des données au sein de ce code a été minutieusement conçu afin de garantir la fiabilité et la cohérence des informations contenues dans un jeu de données dédié aux observations météorologiques. Dans cette démarche, une attention particulière a été accordée à la normalisation des données, notamment en raison de la présence d'années telles que "201" au lieu de "2019". Afin d'écarter toute valeur aberrante similaire, des mesures ont été prises pour définir une plage temporelle excluant les années antérieures à 1995 et postérieures à 2019.

Les entrées associées à des jours ayant la valeur 0 ont toutes été éliminées dans le but de supprimer les données invalides ou manquantes, garantissant ainsi la qualité des enregistrements temporels. De plus, les observations de températures égales à -99.0 ont été exclues afin d'éliminer les valeurs aberrantes susceptibles de biaiser l'interprétation des données, assurant ainsi la fiabilité des mesures de température.

Les températures, initialement exprimées en degrés Fahrenheit, ont été systématiquement converties en degrés Celsius. Cette conversion uniformise l'unité de mesure, facilitant ainsi la comparaison et l'analyse ultérieure. Un effort conséquent a été déployé pour normaliser les noms de villes à l'aide d'un dictionnaire de correspondance, garantissant la cohérence des données en standardisant les dénominations des lieux. Cette démarche simplifie l'agrégation et l'interprétation des informations géographiques.

b. Ajout de données

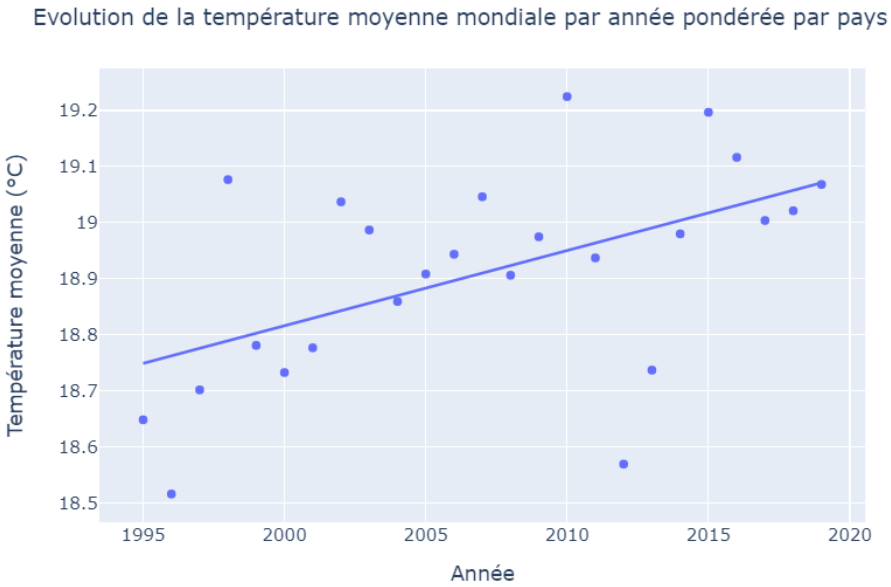
Afin de pallier aux manques de données géographiques nous permettant d'approfondir nos analyses, nous avons fait appel à l'*API Ninja* pour récupérer des informations cruciales telles que la latitude, la longitude et l'altitude pour chacune des 321 villes incluses dans notre jeu de données. L'utilisation de cette API a permis d'automatiser la collecte de données géographiques, offrant ainsi une approche efficace pour enrichir nos enregistrements.

Cependant, une difficulté particulière est survenue lors de la manipulation de villes portant des noms identiques, comme c'est le cas, par exemple, pour Hamilton. Bien que ces villes partagent un nom similaire, elles ne sont pas situées au même endroit géographique. Cette situation a créé des incohérences dans les données, car les informations fournies par l'API ne correspondaient pas toujours à la localisation correcte des villes.

Face à cette complexité, nous avons dû entreprendre une démarche manuelle pour corriger les données. Cela a impliqué la vérification et la mise à jour individuelle des informations de latitude, longitude et altitude pour les villes présentant des discordances afin de garantir la précision et la fiabilité de géolocalisation de chacune d'entre elles. Ce processus de correction manuelle a été essentiel pour pallier les limitations de l'API Ninja et assurer une représentation géographique correcte de toutes les villes incluses dans notre analyse.

2. Résultats et analyses

A. Réchauffement climatique à l'échelle mondiale.

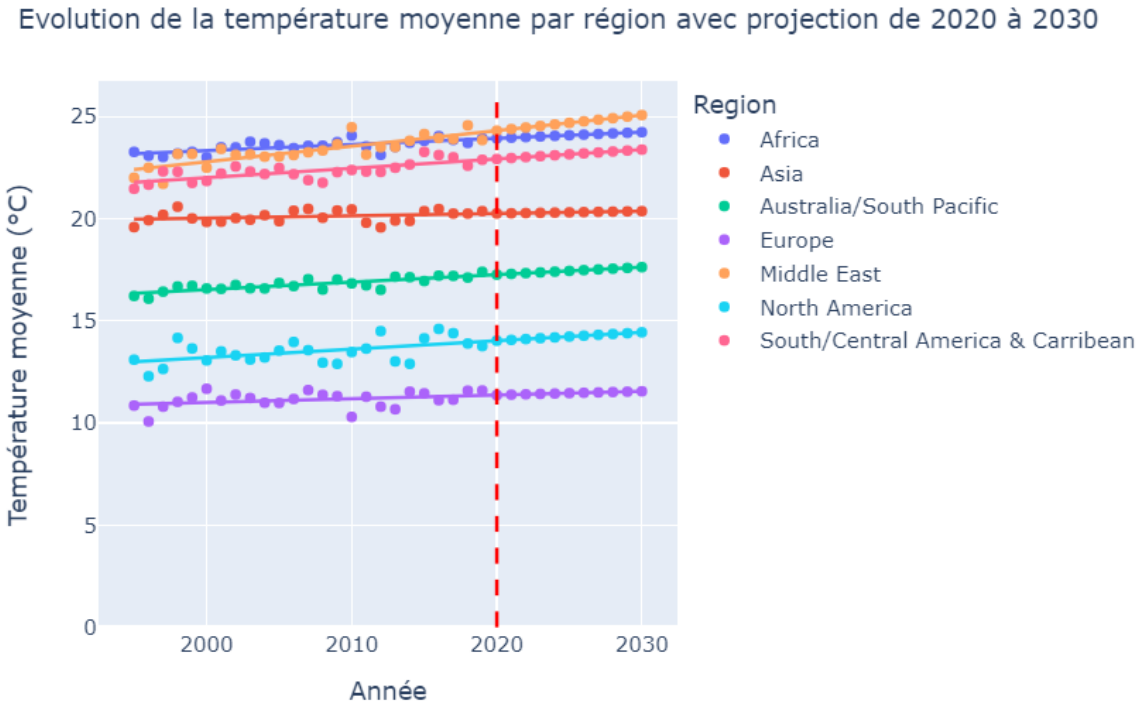


Conformément à ce que nous ressentons tous dans notre quotidien, nous constatons bel et bien sur le graphique ci-dessus une évolution croissante de la température d'année en année. Avec une progression de 0.3°C sur 20 ans, notre modèle linéaire nous estime une température moyenne annuelle de 19.5°C à l'horizon 2050.

Néanmoins, il est important de nuancer cette projection par rapport à la qualité de notre jeu de données. Ces mesures ne contiennent pas les températures des espaces maritimes et zones glaciaires et ne compilent que les données de 125 pays sur les 193 reconnus par l'ONU. Ainsi, de nombreuses aires mondiales ne rentrent pas en compte dans notre prédiction altérant les conclusions de nos résultats.

A noter : Chaque température moyenne annuelle a été calculée à partir des moyennes de chaque pays et non de chaque station météo afin de compenser un déséquilibre dans les données où les USA pèsent 20% des relevés météorologiques de notre dataset.

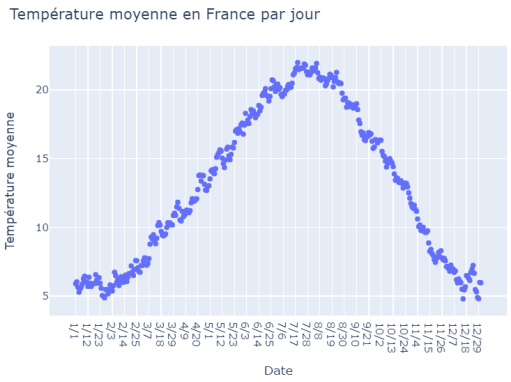
B. Réchauffement climatique à l'échelle régionale

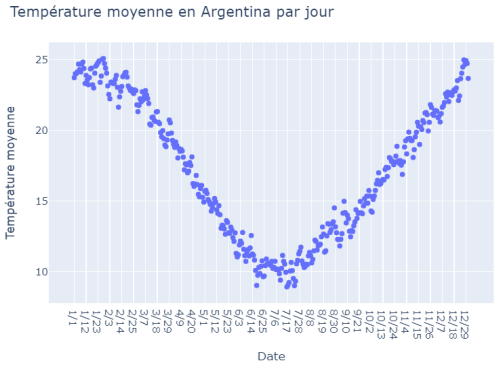


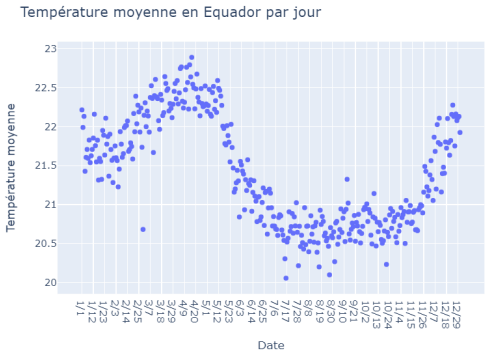
En zoomant notre analyse à l'échelle régionale, nous observons une croissance générale de la température sur l'ensemble des régions mondiales. Néanmoins, certaines parties du monde semblent plus touchées par le réchauffement climatique comme le Moyen-Orient où la température annuelle moyenne connaîtra une augmentation de +3°C en 35 ans là où l'évolution en Amérique centrale ne sera que de 0.5°C supplémentaire.

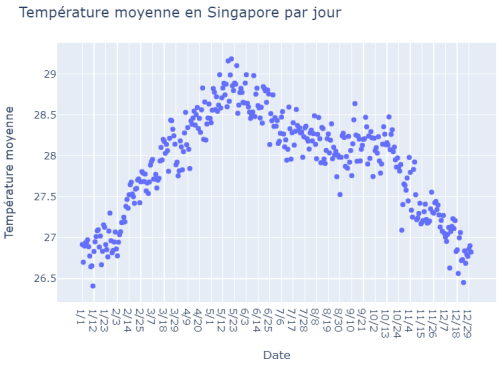
Cette différence pourrait notamment s'expliquer par une sécheresse permanente présente dans ces pays où les vents marins ne peuvent limiter l'effet du réchauffement. Ce raisonnement pourrait également s'appliquer au continent Africain, cependant, nous ne possédons que des données de pays côtiers pour cette partie du monde biaisant ainsi l'interprétation.

C. Détection de saisonnalités

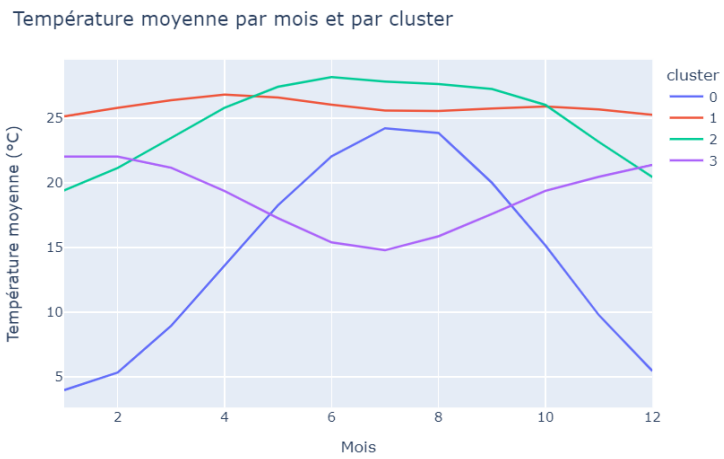


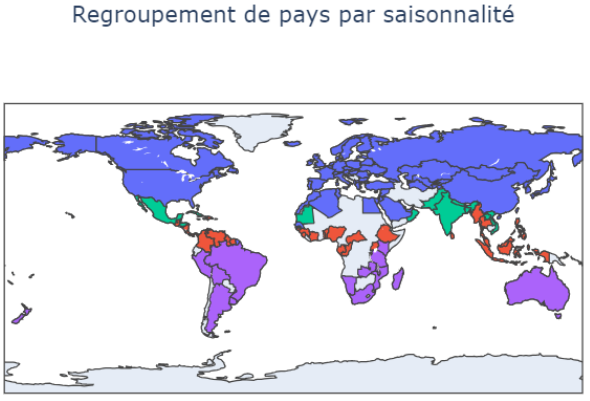






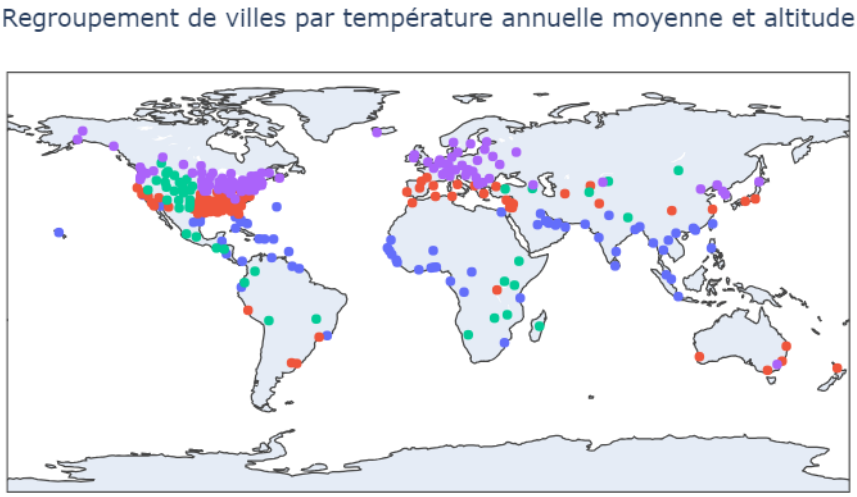
Comme nous l'observons ci-dessus, la temporalité des températures diffère d'un pays à l'autre. Il serait alors intéressant de chercher à regrouper les pays par saisonnalité afin de trouver les pays partageant ces mêmes propriétés. Réalisons cette opération à travers une méthode de clustering telle que K-means !





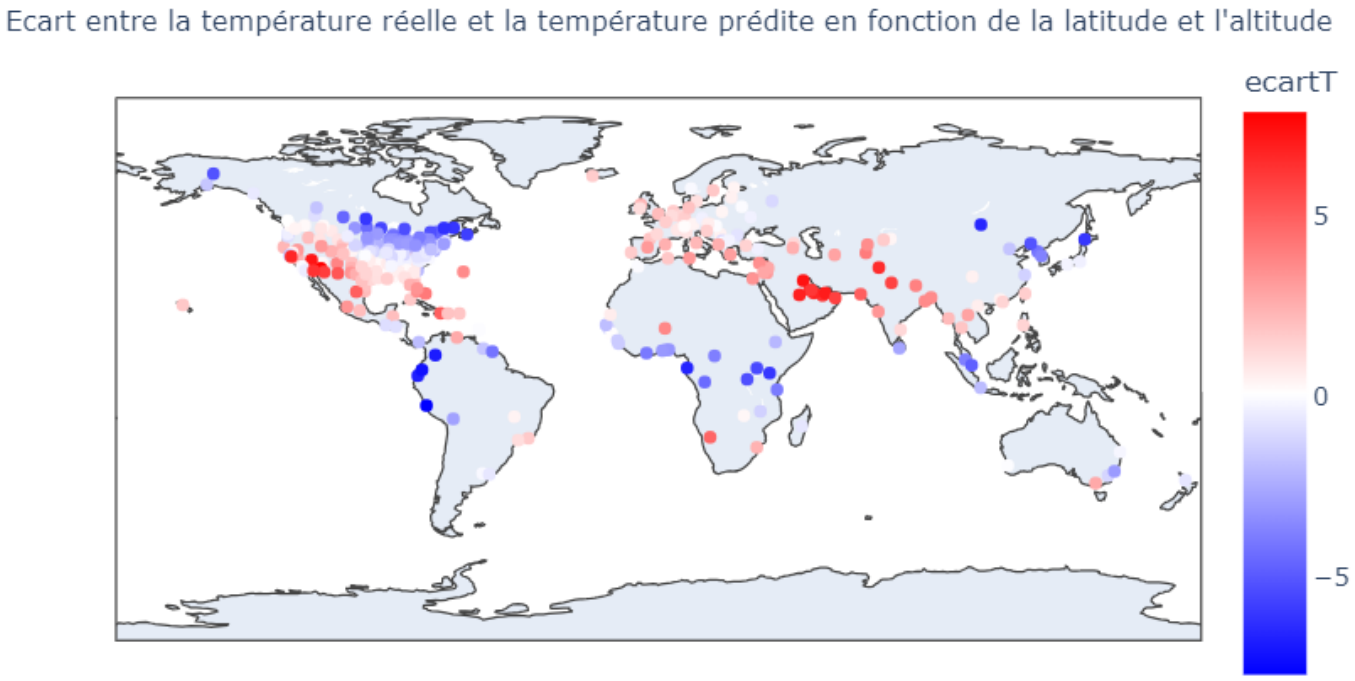
Nous retrouvons ainsi le monde découpé selon 4 grandes régions. Les clusters 0 et 3 regroupent les pays aux 4 saisons classiques que nous connaissons : Printemps, Été, Automne, Hiver mais inversé temporellement en fonction de l'hémisphère. Le cluster 2 correspond également à ce schéma mais avec des pays de l'hémisphère nord ayant des températures globalement plus élevées. Le cluster 1 regroupe quant à lui les pays situés sur l'axe de l'équateur ayant un climat tropical très chaud alternant saisons des pluies et saisons du beau temps. Nous repérons ainsi 2 pics de températures sur la période d'avril et d'octobre.

D. Détection de villes aux mêmes températures annuelles



La carte ci-dessus illustre les villes ayant la même température annuelle moyenne en fonction de leur altitude. Nous mettons ainsi en évidence l'impact de l'altitude dans les températures d'un pays.

E. Détection de températures anormales en fonction de la localisation



Une question reste en suspens depuis le début de notre analyse : les températures moyennes que nous traitons sont-elles en adéquation avec leur position géographique ? Les critères géographiques bruts tels que la localisation ou l'altitude sont-ils suffisants pour analyser un schéma de température mondiale ?

Pour répondre à cette question, nous allons analyser les températures en fonction de la latitude et de l'altitude d'une station météo. A l'aide d'une régression linéaire nous pouvons rapidement prévoir une température théorique ne se basant que sur des critères géographiques pures. Nous pouvons ainsi créer la carte ci-dessus mettant en lumière les écarts de températures entre la valeur supposée par rapport à la position géographique et la température réelle mesurée.

Ainsi, nous voyons par exemple que la température annuelle moyenne théorique de Paris prenant en compte que sa géographie devrait être de 10.57°C au lieu de 12.2°C. D'autres villes comme celles se situant dans le golfe persique atteignent des écarts supérieurs à 5 degrés dans les températures mesurées. Comment peux-t-on alors expliquer ces écarts ?

En réalité, de nombreux facteurs autre que la position géographique pure entrent en jeu pour caractériser et exprimer la température d'une zone terrestre. Les courants marins, l'humidité ou encore la pression atmosphérique en sont trois illustrations. Le réchauffement de l'Europe par rapport à sa position spatiale s'explique par exemple par le phénomène du *Gulf Stream*, un courant océanique amenant une masse d'air plus chaude dans notre région. Les températures extrêmes du Golfe persique s'expliquent quant à elle par une absence d'humidité rendant le climat sec et aride, renforçant encore plus la chaleur présente. Cette explication se retrouve aussi en partie dans le désert américain des États de l'Ouest où l'absence de vents marins entraîne une concentration de la chaleur.

D'autres villes ont en revanche un climat bien plus faible que leur position laisserait supposer. Les courants marins en sont là encore la principale raison comme nous le voyons avec les villes côtières proches de l'équateur majoritairement plus froides que la théorie. Les courants atmosphériques glaciaux provenant de l'arctique expliquent également les fraîches températures canadiennes mesurées.

Conclusion

En résumé, notre étude portant sur l'analyse du climat mondial, guidée par un traitement méticuleux des données nous a permis d'obtenir certaines tendances significatives. Le nettoyage et l'enrichissement des données ont permis d'obtenir un ensemble robuste, révélant une augmentation linéaire de la température moyenne mondiale de 0.3°C par décennie, avec une projection de 19.5°C d'ici 2050. Cependant, des nuances sont nécessaires en raison des limites de notre jeu de données.

À l'échelle régionale, des disparités émergent, soulignant la sensibilité variable de différentes régions au réchauffement climatique. L'analyse des saisons, la détection de villes aux températures similaires, et l'évaluation des températures en fonction de la localisation mettent en lumière la complexité des influences climatiques locales, dépassant la simple géographie.

En conclusion, cette étude offre une vision éclairante du réchauffement climatique, soulignant l'importance d'une approche complète. Ces résultats fournissent des bases solides pour comprendre les variations climatiques mondiales, tout en soulignant la nécessité de prendre en compte la diversité des facteurs influençant les températures régionales. Ces éléments sont cruciaux pour orienter des actions informées face aux défis urgents du changement climatique.