IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

FISE - A2

UE BC : A Journey To A Data Scientist

Descriptive Statistics (Data understanding et Data preparation)

Groupe n°4

Version N°1

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Table des matières

I. Introduction…… 4
II. Compréhension des données……6
III. Préparation des données……..15
IV. Proposition pour l'apprentissage automatique…..……17
V. Conclusion et perspectives……18
VI. Annexe ……..……19

Liste des figures

Figure1:Carte de pollution de la France 2018……..4
Figure2:Prévalence moyenne par département pour les pathologies liées à la pollution en 2018……..9
Figure3:Top 10 des polluants ayant le plus d'occurrence dans le dataset en 2010……..10
Figure4:Top 10 des départements par quantité émise de CO2 (biomasse et non biomasse)……....11
Figure5:Évolution temporelle des émissions de CO2 entre 2004 et 2017 ……..12
Figure6:Corrélation spearman entre quelques polluants et le cancer....12
Figure7:Corrélation Pearson entre quelques polluants et le cancer…..13
Figure8:P-valeur pour les tests de corrélations Spearman entre les polluants et le Cancer…….. 14
Figure9:P-valeur pour les tests de corrélations Pearson entre les polluants et le Cancer……...14
Figure10:Corrélations entre quelques polluants……....15
Figure11:Extrait du jeu de données combiné : Prévalence des pathologies et quantités de polluants par département……....16

I. Introduction

I.1 Présentation du problème métier

La pollution de l'air et de l'eau en France représente un défi majeur pour la santé publique et l'économie. Avec des impacts sanitaires significatifs, tels qu'environ **40 000 décès prématurés** et **20 000 hospitalisations annuelles**, la pollution affecte directement la qualité de vie et l'espérance de vie des Français, particulièrement dans les zones urbaines. **35 % des Français** sont exposés à des niveaux de pollution atmosphérique dépassant les recommandations de l'OMS, et **3 millions** subissent les effets de la pollution de l'eau potable.

Les principaux polluants, comme les **particules fines (PM2.5)**, le **dioxyde d'azote (NO2)** et l'ozone, sont liés à des maladies graves, notamment l'asthme, les maladies cardiovasculaires et le cancer. Sur le plan économique, la pollution coûte environ **3,8 milliards d'euros par an**, avec des pertes importantes en productivité et des charges élevées pour le système de santé.

Ces impacts sanitaires et économiques nécessitent des solutions basées sur des données pour cibler les zones les plus touchées et prioriser les interventions de santé publique. La **figure 1** ci-dessous illustre les niveaux de pollution dans les différentes régions françaises, mettant en lumière les disparités géographiques.

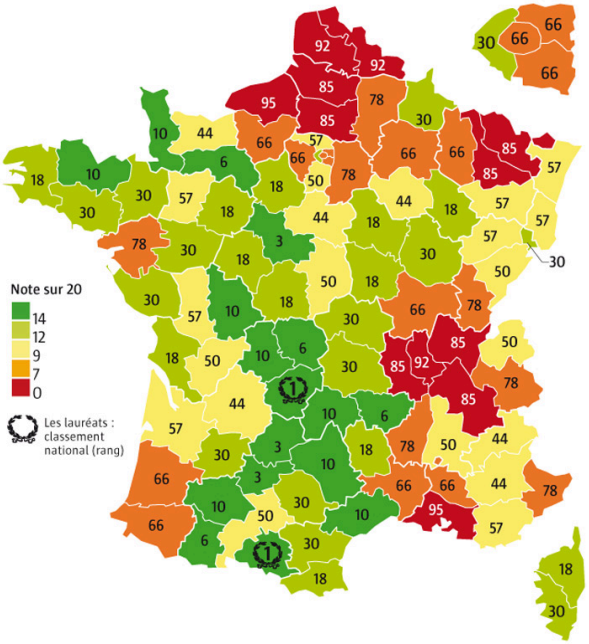

figure n°1: Carte de pollution de la France 2018

source : https://altoservices.fr/carte-de-france-de-pollution/

I.2 Problématique générale

Comment la pollution de l'air et de l'eau influence-t-elle l'incidence des maladies respiratoires, cardiovasculaires et des cancers dans les régions les plus touchées ? L'objectif est d'aider les agences régionales à mieux comprendre ces liens et à ajuster leurs stratégies de sensibilisation et d'allocation de ressources médicales en fonction des spécificités de chaque département.

I.3 Problématique Data Science

Pour répondre à cette problématique, l'approche de data science se concentre sur l'analyse des corrélations entre les indicateurs de pollution et la prévalence des maladies, tout en prenant en compte les facteurs démographiques et géographiques. Les objectifs analytiques incluent :

1. **Corrélations** : Identifier les polluants ayant la plus forte corrélation avec des conditions de santé spécifiques.
2. **Classification** : Classer les régions en fonction des risques sanitaires liés à la pollution.
3. **Prédictions** : Construire des modèles prédictifs pour estimer la prévalence des maladies à partir des données sur la pollution.

I.4 Résultats attendus

Les résultats de cette analyse permettront de :

- Identifier les **polluants critiques** liés à des conditions de santé spécifiques.
- Localiser les **points chauds géographiques** pour maximiser l'efficacité des interventions.
- Générer des **modèles prédictifs** fiables pour estimer la prévalence des maladies et anticiper les besoins en ressources médicales.

II. Compréhension des données

II.1 Description des données disponibles

Pour répondre à la problématique posée, deux ensembles de données pertinents ont été identifiés :

II.1.a. Premier jeu de données : *effectifs.csv*

Taille du dataset :4 636 800 lignes et 16 colonnes.

Description : Ce jeu de données contient des informations détaillées sur les effectifs de patients pris en charge par l'Assurance Maladie en France depuis 2015. Il fournit une répartition départementale des patients selon les pathologies, les traitements chroniques ou les épisodes de soins. Les catégories médicales incluent notamment les maladies cardio-neurovasculaires, les maladies respiratoires chroniques, les traitements du risque vasculaire, le diabète, les cancers, et les hospitalisations liées à la Covid-19.

Modalités des variables (types et descriptions):

- **Variables numériques** :
  - `Ntop` : Ratio (effectif des patients atteints par la pathologie).
  - `Npop` : Ratio (taille de la population).
  - `prev` : Ratio (prévalence de la pathologie en pourcentage).
  - `tri` : Ratio (valeur calculée ou score).
- **Variables catégoriques** :
  - `patho_niv1` : Nominal (type de pathologie, par ex. : maladies cardio-neurovasculaires, cancers, etc.).
  - `patho_niv2` et `patho_niv3` : Nominal (sous-catégories de pathologies offrant plus de détails).
  - `cla_age_5` : Nominal (catégorie d'âge).
  - `sexe` et `libelle_sexe` : Nominal (genre des patients).
  - `région` et `dept` : Nominal (zones géographiques).
  - `libelle_classe_age` : Nominal (description des classes d'âge).
  - `Niveau prioritaire` : Nominal (catégorie de priorité).

Quantification des données manquantes:

Certaines variables présentent un pourcentage significatif de valeurs manquantes, comme indiqué dans le tableau ci-dessous :

Variable	Nombre de valeurs manquantes
année	0
patho_niv2	483 840
patho_niv3	1 048 320
prev	1 238 024
région	0
dept	0

Ce tableau met en évidence le nombre de valeurs manquantes dans chaque colonne du dataset. Les variables *patho_niv 2*, *patho_niv 3* et *prev* présentent des valeurs manquantes importantes, ce qui pourrait influencer les analyses ultérieures.

Statistiques descriptives:

- **Variables numériques** :
  Les principales statistiques pour les variables pertinentes sont résumées dans le tableau suivant :

Statistique \ Colonne	Ntop	Npop	prev
Nombre de valeurs non manquantes	3 398 776	3 398 776	3 398 776
Moyenne	5 527.4	130 880	6.21
Écart-type	158 991.2	1 146 238	15.96

- **Variables catégoriques** :

Les distributions des valeurs catégoriques sont résumées ci-dessous :

Statistique \ Colonne	patho_niv1	dept	région
Nombre de valeurs uniques	18	102	19
Valeur la plus fréquente	Maladies cardio-neurovasculaires	999	76
Fréquence	660 507	587 455	376 196

II.1.b. Deuxième jeu de données : *registre-français-des-émissions-polluantes-émissions.csv*

Taille du dataset :138 726 lignes et 16 colonnes.

Description : Ce jeu de données contient des informations sur les émissions polluantes déclarées par des établissements en France. Il inclut des détails sur les types de polluants, les quantités émises, les milieux impactés (air, eau, sol), ainsi que des informations géographiques et industrielles sur les établissements.

Modalités des variables (types et descriptions):

- **Variables numériques** :
  - `Quantité` : Ratio (quantité de polluants émis).
  - `Coordonnées` : Ratio (position géographique).
- **Variables catégoriques** :
  - `Milieu` : Nominal (environnement affecté : air, eau, sol).
  - `Polluant` : Nominal (type de polluant, ex. : CO2, SO2, etc.).
  - `Région` et `Département` : Nominal (zones géographiques).
  - `Nom Etablissement` : Nominal (identité de l'émetteur).

- `Année Emission` : Intervalle (année de l'enregistrement).

Quantification des données manquantes:

Les variables présentant des valeurs manquantes les plus significatives sont listées ci-dessous :

Variable	Nombre de valeurs manquantes
Année Emission	0
Département	520
Région	520
Unité	88 885

Ces manques, particulièrement pour les données géographiques, peuvent limiter certaines analyses spatiales.

Statistiques descriptives:

- **Variables numériques** :

Variable	Nombre de valeurs non manquantes	Moyenne	Ecart type
Quantite	138 726	6 334 483	120 083 369

- **Variables catégoriques** :
  Les distributions des variables catégoriques clés sont résumées ci-dessous :

Variable	La valeur la plus fréquente	Fréquence
Milieu	Air	86 546
Polluant	CO2 total	16 116
Région	Rhône-Alpes	15 941

II.2 Visualisation des données

II.2.a. Premier jeu de données : *effectifs.csv*

Visualisation de la prévalence moyenne par département :

Le graphique montre le **Top 10 des départements** en France avec la plus haute **prévalence moyenne de cancers,Maladies Cardio Neurovasculaires et Maladies respiratoires chroniques (%).**

- **Département 48 (Lozère)** : Il enregistre la prévalence la plus élevée, dépassant 4% pour le Cancer et plus que 6% pour les maladies cardio neurovasculaires.
- **Départements 23, 5, 4, 15** : Ces départements suivent de près avec des prévalences autour de 3,8%, suggérant une situation sanitaire préoccupante liée à l'incidence du cancer.
- **Départements 2A,2B,36** : Ces départements suivent de près avec des prévalences autour de 5,9%, suggérant une situation sanitaire préoccupante liée à l'incidence des maladies cardio neurovasculaire.
- **Département 974 (Département de la Réunion) et 08(Ardennes)** : Ils enregistrent la prévalence la plus élevée, dépassant 8.5% pour les maladies respiratoires chroniques

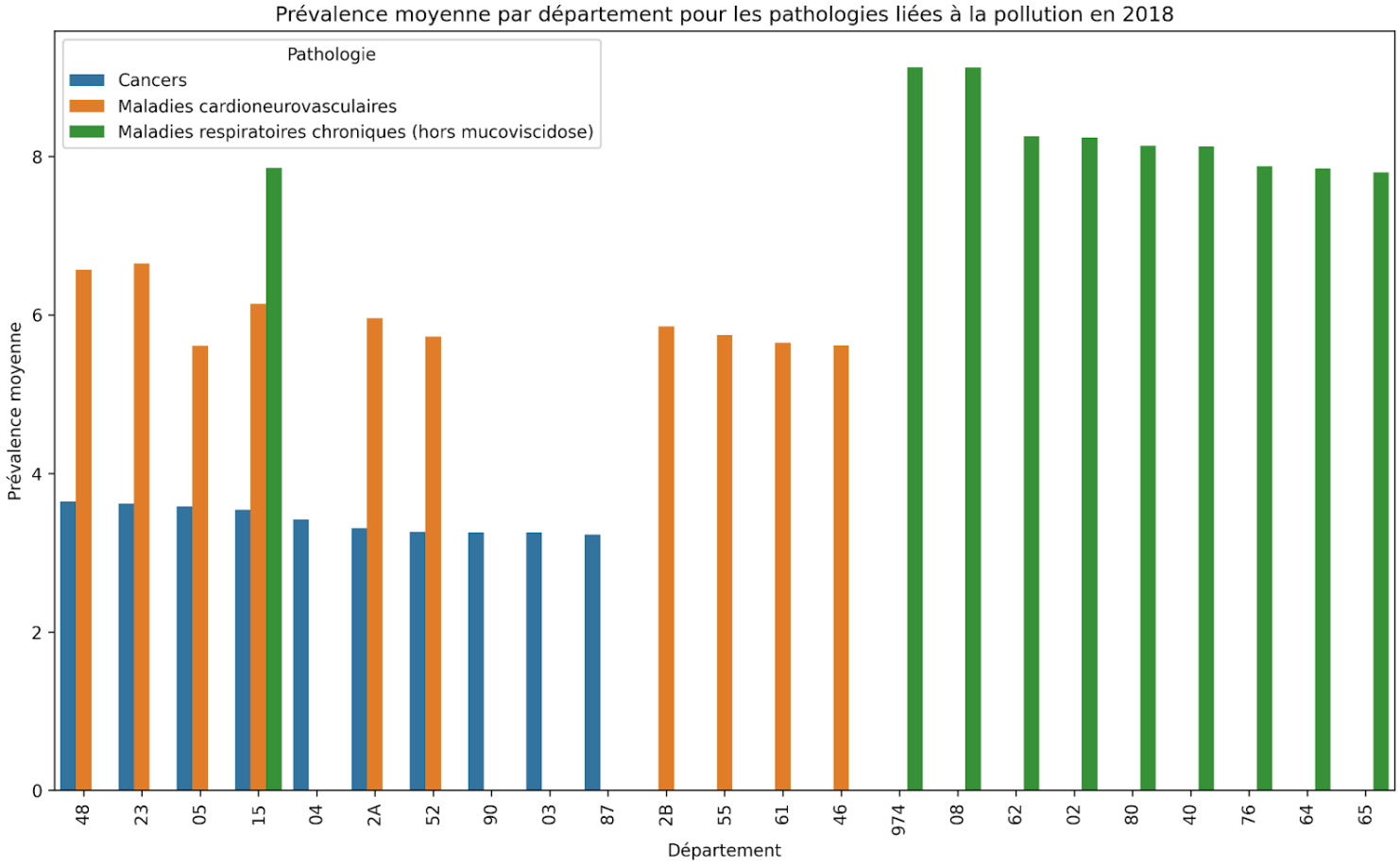

figure n°2 : Prévalence moyenne par département pour les pathologies liées à la pollution en 2018.

II.2.b. Deuxième jeu de données : *registre-français-des-émissions-polluantes-émissions.csv*

Statistiques sur la quantité émise par type de polluant:
La figure met en évidence les trois polluants les plus fréquents en France :

- **Dioxyde de carbone (CO2) d'origine biomasse :** Principal polluant, il provient de la combustion ou décomposition de matières organiques, notamment dans le cadre des activités énergétiques et agricoles.
- **Ammoniac (NH3) :** Émis principalement par le secteur agricole, à travers les déjections animales et les engrais, il contribue à la formation de particules fines, nocives pour la santé.
- **Dioxyde de carbone (CO2) d'origine non biomasse :** Issu de la combustion d'énergies fossiles et de processus industriels, il reflète l'impact des activités humaines sur les émissions.

Ces polluants illustrent les principales sources de pollution en France, combinant activités agricoles, industrielles et énergétiques, et soulignent la nécessité d'une gestion ciblée.

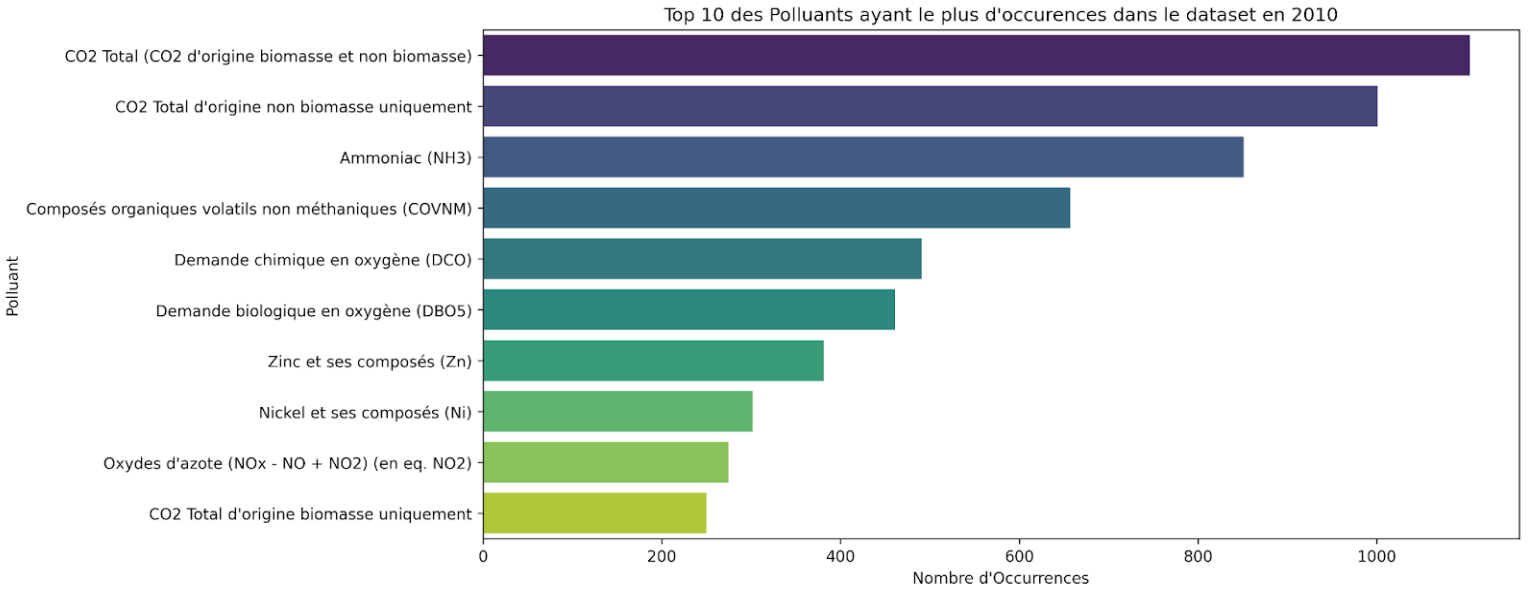

Figure n°3 : Top 10 des polluants ayant le plus d'occurrence dans le dataset en 2010.

<u>Émissions totales par département:</u>

Ce graphique montre les départements les plus émetteurs de CO2 en 2016, avec une prédominance des Bouches-du-Rhône et de La Réunion, reflétant une forte activité industrielle ou énergétique. Dans des départements comme la Corse-du-Sud et la Haute-Corse, les émissions de CO2 biomasse et non biomasse sont similaires, indiquant des sources mixtes. En revanche, des départements comme le Cher affichent une dominance des émissions d'origine non biomasse, souvent liées à l'industrie. Ces disparités régionales soulignent la nécessité d'interventions ciblées, notamment dans les départements ultramarins encore fortement dépendants des combustibles fossiles.

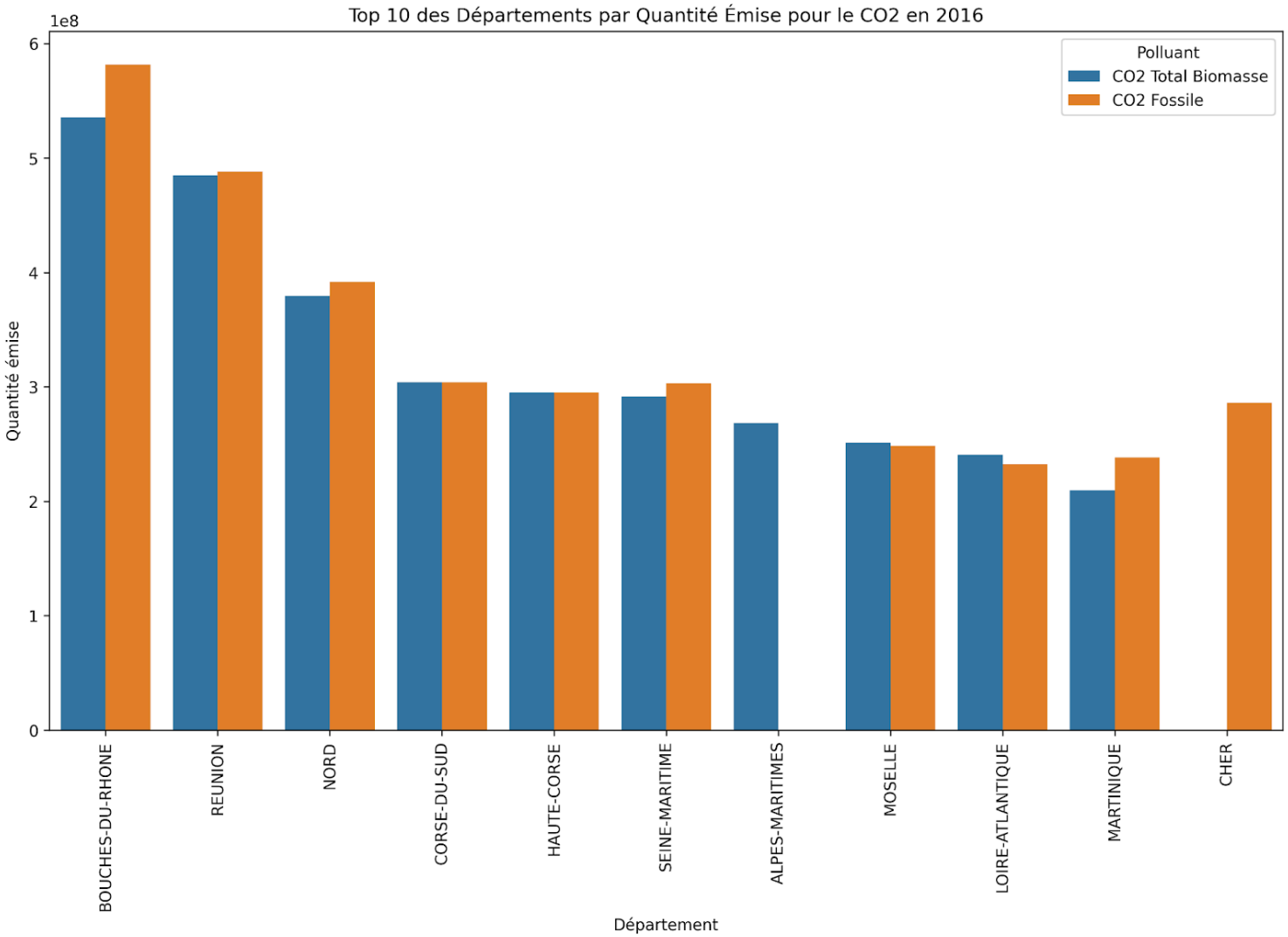

Figure n°4 : Top 10 des départements par quantité émise de CO2 (biomasse et non biomasse).

<u>Évolution temporelle des émissions de CO2 :</u>

Le graphique illustre l'évolution des émissions de CO_2 (biomasse et non-biomasse) entre 2004 et 2017. Un pic significatif est observé en 2014, atteignant un sommet pour les émissions totales et celles issues uniquement de la biomasse. Les différences entre ces deux types d'émissions restent faibles sur toute la période. Après 2014, les niveaux d'émissions se stabilisent, suggérant une diminution des fluctuations observées précédemment.

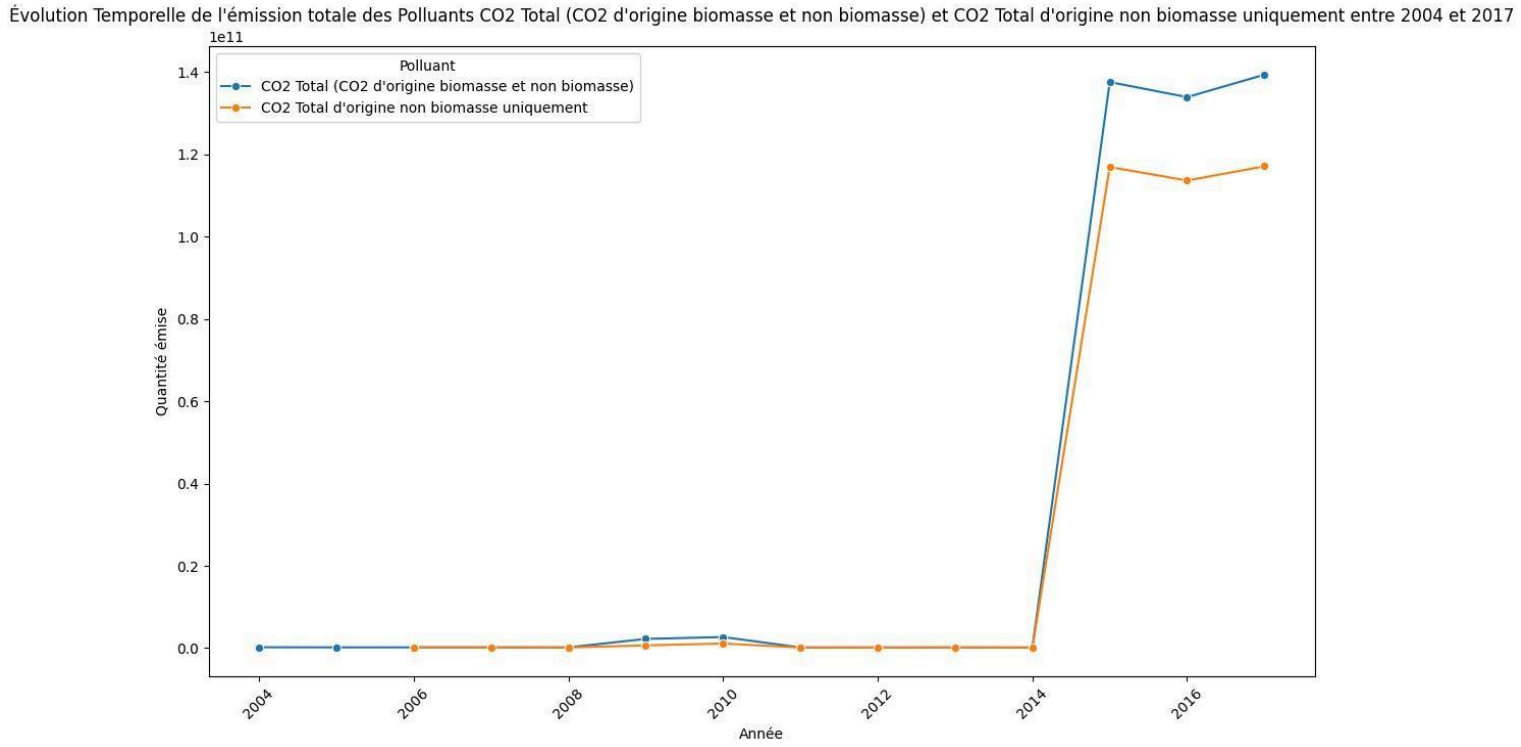

figure n°5 : Évolution temporelle des émissions de CO2 entre 2004 et 2017.

II.3 Corrélations entre variables

II.3.a Corrélations avec les pathologies

Les corrélations entre les polluants et la prévalence des pathologies ont été explorées à l'aide des coefficients de Spearman et de Pearson

- **Spearman**: Le polluant ayant la corrélation la plus élevée avec le cancer est l'hexachlorobutadiène (0,24).

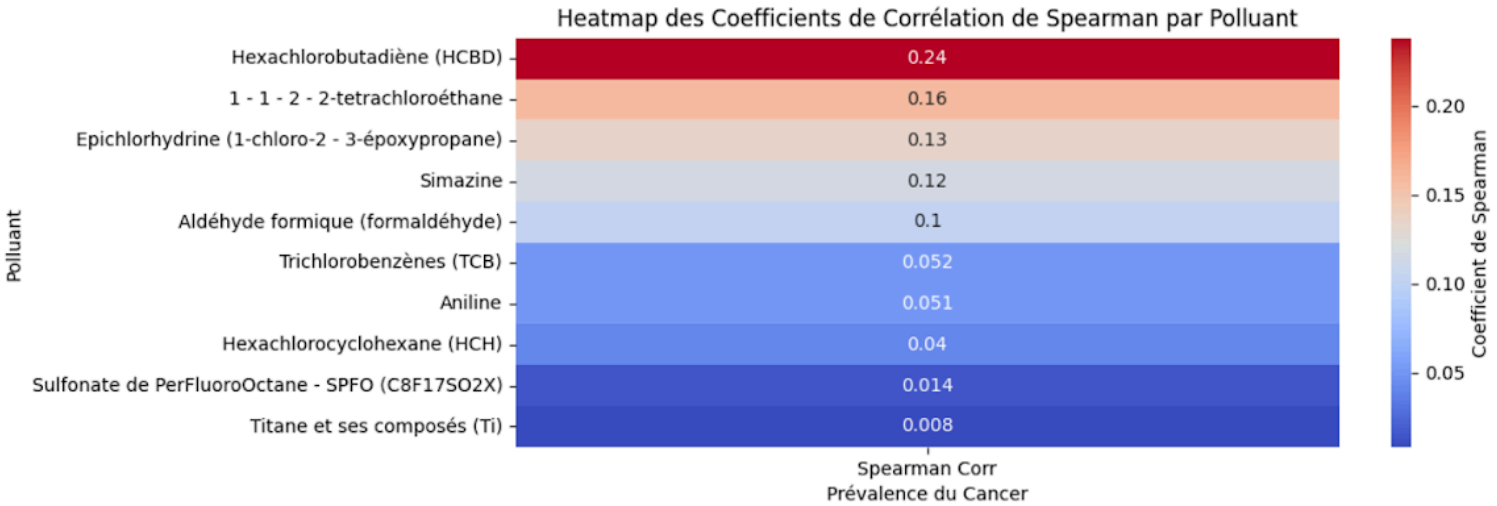

figure n°6 : Corrélation spearman entre quelques polluants et le cancer

- **Pearson**: Le polluant ayant la corrélation la plus élevée avec le cancer est le 1,1,2,2-tétrachloroéthane (0,18).

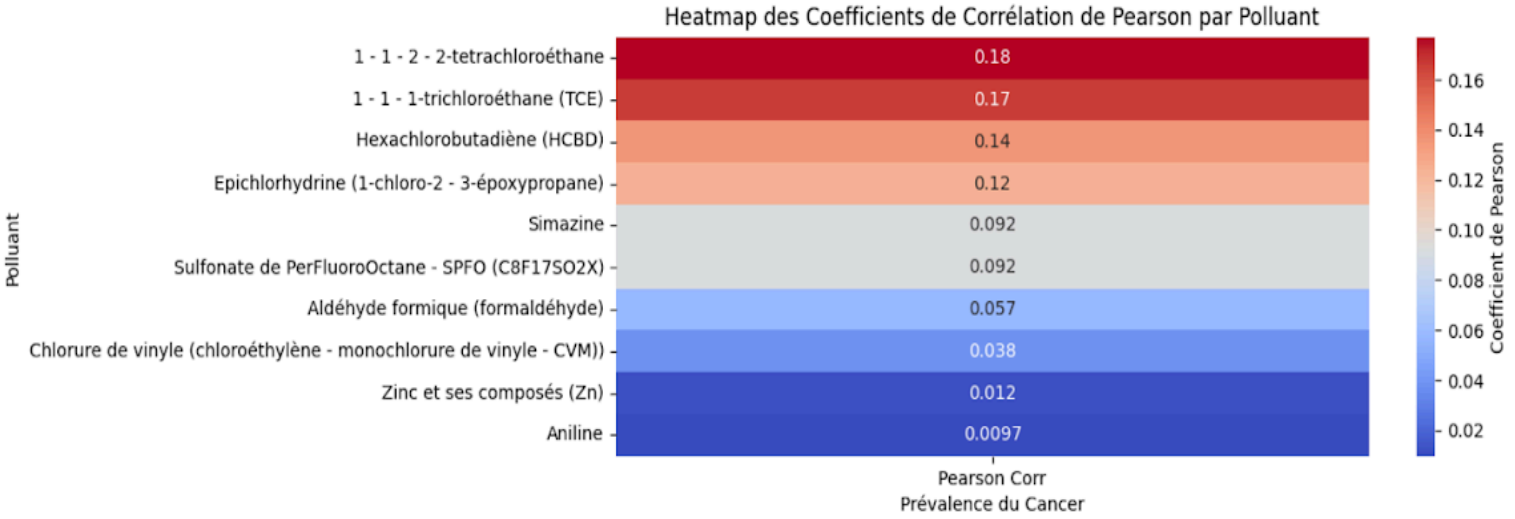

figure n°7 : Corrélation Pearson entre quelques polluants et le cancer

- **Les corrélations entre les quantités de polluants et la prévalence** sont faibles pour les deux métriques (Pearson et Spearman). Pour Pearson, la valeur maximale est de 0,18 pour le 1,1,2,2-tétrachloroéthane. Pour Spearman, la valeur maximale est de 0,24 pour l'hexachlorobutadiène. De plus, **leurs p-valeurs élevées confirment que ces corrélations ne sont pas statistiquement significatives.**

Polluant	**P-valeur pour le test de Spearman**
Simazine	0.245485
Sulfonate de PerFluoroOctane - SPFO	0.891780
1 - 1 - 2 - 2 - tétrachloroéthane	0.115070
Aldéhyde formique	0.312199
Trichlorobenzène	0.605733
Aniline	0.609376
Hexachlorocyclohexane	0.693994
Epichlorhydrine	0.217087
Titane et ses composés	0.936462

figure n°8 : P-valeur pour les tests de corrélations Spearman entre les polluants et le Cancer

Polluant	P-valeur pour le test de Pearson
Simazine	0.358520
Sulfonate de PerFluoroOctane - SPFO	0.358566
1 - 1 - 2 - 2 - tétrachloroéthane	0.076603
Aldéhyde formique	0.572476
1 - 1 - 1 - trichloroéthane	0.097126
Aniline	0.923236
Hexachlorobutadiène	0.173800
Epichlorhydrine	0.217087
Zinc et ses composés	0.908540
Chlorure de Vinyle	0.704282

figure n°9 :P-valeur pour les tests de corrélations Pearson entre les polluants et le Cancer

II.3.b Corrélations entre les polluants

De nombreuses relations fortes ont été identifiées entre certains polluants. Le tableau suivant présente des exemples de corrélations élevées, avec des p-valeurs très faibles, indiquant que ces corrélations sont hautement significatives.

Polluant 1	Polluant 2	Coefficient de corrélation de Pearson	P-valeur du test de Pearson
Sulfure d'hydrogène (H2S)	Crésol (mélange d'isomères)	0.998731	2.563471e-130
Demande biologique en oxygène (DBO5)	Demande chimique en oxygène (DCO)	0.999069	5.693526e-137
Chrome et ses composés (Cr)	Aluminium et ses composés (Al)	0.998811	1.022277e-131
Chrome et ses composés (Cr)	Fer et ses composés (Fe)	0.998195	9.541423e-123
Chrome et ses composés (Cr)	1 - 2-dichloroéthane (DCE - chlorure d'éthylène)	0.996569	5.892691e-109
Chrome et ses composés (Cr)	Chlorures (Cl total)	0.998395	2.874750e-125
Demande chimique en oxygène (DCO)	Demande biologique en oxygène (DBO5)	0.999069	5.693526e-137

figure n°10 : Corrélations entre quelques polluants.

III. Préparation des données

L'objectif de cette étape est de préparer un ensemble de données propre et structuré, permettant d'analyser la relation entre les quantités de polluants émis et la prévalence d'une pathologie spécifique, telle que le cancer. Les étapes incluent la fusion des données, la réduction des variables corrélées et la préparation pour les modèles de machine learning.

III.1. Fusion des données

Pour établir une base de données cohérente, nous avons suivi les étapes suivantes :

III.1.a. Création des variables pour les polluants

À partir des données sur la pollution, nous avons créé un jeu de données contenant des variables représentant les quantités totales de chaque polluant émis dans un département donné pour une année spécifique.

Structure du tableau :

- Département
- Année
- Quantité totale émise de chaque polluant (par exemple : CO_2, NO_2, particules fines, etc.).

III.1.b. Agrégation des prévalences pathologiques

À partir des données sur les pathologies, nous avons extrait un tableau contenant les prévalences moyennes des pathologies pour chaque département et chaque année.

Structure du tableau :

- Département
- Année
- Prévalence moyenne d'une pathologie donnée (exemple : cancer).

III.1.c. Concaténation des deux jeux de données

Les deux ensembles ont été fusionnés en utilisant **l'année** et **le département** comme clés de jonction.

➔ Les données combinées regroupent les émissions de polluants en 2010 et les prévalences pathologiques en 2022, tenant compte d'un délai estimé de 10 ans pour l'apparition des symptômes liés à la pollution.
Voici un extrait des 5 premières lignes du jeu de données final :

	prev	Departement	Ammoniac (NH3)	CO2 Total (CO2 d'origine biomasse et non biomasse)	CO2 Total d'origine non biomasse uniquement	Sulfure d'hydrogène (H2S)	Demande biologique en oxygène (DBO5)	Phénols (Ctotal)	Cadmium et ses composés (Cd)	Méthane (CH4)	...
0	2.718709	AIN	158600.00	673200.0	595600.0	0.0	464700.0	0.0	18.30000	1214000.0	...
1	2.707403	AISNE	132904.63	881900.0	641200.0	0.0	102300.0	0.0	86.46000	854000.0	...
2	3.272814	ALLIER	64237.00	503500.0	432500.0	0.0	351100.0	24.5	2.15755	1857000.0	...
3	3.393718	ALPES-DE-HAUTE-PROVENCE	0.00	76000.0	76000.0	0.0	68500.0	592.0	0.00000	0.0	...
4	3.573994	HAUTES-ALPES	0.00	0.0	0.0	0.0	0.0	0.0	0.00000	628000.0	...

figure n°11: Extrait du jeu de données combiné : Prévalence des pathologies et quantités de polluants par département

III.2. Réduction des dimensions

Pour éviter les problèmes liés à des variables redondantes ou fortement corrélées, nous avons appliqué une méthodologie rigoureuse en deux étapes :

III.2.a. Analyse des corrélations binaires

Nous avons calculé les coefficients de corrélation entre les variables représentant les quantités de polluants.

Les variables présentant des corrélations supérieures à 0,85 ont été identifiées, et certaines d'entre elles ont été supprimées pour réduire la redondance.

Cette étape a permis de réduire le nombre de variables de **101 à 49**.

III.2.b. Analyse de multicolinéarité avec le VIF (Variance Inflation Factor)

Nous avons calculé le VIF pour chaque variable restante. Les variables ayant un VIF supérieur à 10, indiquant une multicolinéarité élevée, ont été supprimées.

Résultat final : Réduction à **32 variables explicatives**, éliminant ainsi les variables inutiles tout en conservant l'information pertinente.

III.3. Préparation pour les modèles de machine learning

III.3.a. Division des données

Nous avons divisé l'ensemble de données final en deux parties :

- **Ensemble d'entraînement (80 %)** : Utilisé pour construire les modèles prédictifs.
- **Ensemble de test (20 %)** : Utilisé pour évaluer la performance des modèles.

III.3.b. Normalisation des données

Les variables explicatives (quantités de polluants) ont été normalisées afin d'éviter qu'un polluant avec des valeurs naturellement élevées (par exemple : CO_2) influence de manière disproportionnée les modèles.

IV. Proposition pour l'apprentissage automatique

IV.1. Formulation des tâches de machine learning

IV.1.a. Exploration des Relations Non Linéaires

Bien que des modèles linéaires puissent être efficaces, il est important d'explorer des relations non linéaires entre la pollution et les pathologies. Des techniques comme les arbres de décision ou les modèles basés sur les forêts aléatoires (Random Forest) pourraient mieux capturer ces relations complexes.

IV.1.b. Modélisation Spatio-Temporelle

Pour mieux comprendre l'impact de la pollution au fil du temps, il serait pertinent d'intégrer une dimension temporelle dans les analyses. Des modèles spatio-temporels pourraient être utilisés pour prendre en compte les variations de pollution et de prévalence des maladies sur plusieurs années, et ainsi mieux évaluer l'évolution des risques.

IV.1.c. Analyse de la Causalité

Bien que la corrélation fournisse des insights importants, une analyse plus poussée sur les liens causals (par exemple, via des diagrammes causaux dirigés ou des modèles de régression instrumentale) permettrait d'identifier les facteurs sous-jacents qui contribuent réellement à la prévalence des maladies liées à la pollution.

IV.1.d. Clustering

Identifier les départements les plus affectés par les effets de la pollution afin de cibler les interventions des Agences Régionales de Santé (ARS).

IV.2. Algorithmes recommandés et Métriques d'évaluation

Les premières tentatives (non optimisées) avec certains modèles de régression montrent que l'utilisation de modèles linéaires n'est pas très efficace pour décrire le problème (c'est-à-dire établir les liens entre les polluants et la prévalence). Il semble donc préférable, à priori, d'explorer des algorithmes non linéaires tels que le Random Forest Regressor. Une première tentative (non optimisée) avec ce modèle indique que 22 % de la prévalence pourrait être expliquée par les polluants.

Pour aller plus loin, il serait intéressant d'essayer des algorithmes comme Gradient Boosting ou XGBoost, qui sont également adaptés aux relations complexes et non linéaires, tout en optimisant les paramètres du Random Forest pour maximiser ses performances.

Pour évaluer les performances des modèles, nous utiliserons des métriques telles que :

- ***Mean Absolute Error (MAE) :*** La moyenne des écarts absolus entre les valeurs réelles et prédites.
- ***Mean Squared Error (MSE) :*** La moyenne des carrés des écarts entre les valeurs réelles et prédites.
- ***Cross-Validation :*** Divise les données en plusieurs sous-ensembles pour évaluer la robustesse du modèle sur différents jeux de données.
- ***Robustesse du Modèle : Durée d'entraînement et de prédiction :*** Pour évaluer l'efficacité computationnelle.

V. Conclusion et perspectives

Cette étude a permis de mettre en lumière des relations significatives entre les variables issues des deux jeux de données étudiés : les données sur les pathologies et celles sur la

pollution de l'air et de l'eau. La combinaison de ces informations a révélé des corrélations intéressantes et pertinentes.
Le premier jeu de données couvre les pathologies par région et tranche d'âge entre 2016 et 2022. Il inclut des variables telles que l'année, la région, la classe d'âge, le sexe, ainsi que des détails spécifiques sur les types de pathologies (notamment les cancers et les maladies respiratoires).
Le second jeu de données, portant sur la pollution de l'air et de l'eau entre 2004 et 2017, fournit des informations sur les polluants, les quantités émises et les zones géographiques concernées (départements et régions).
Grâce à une analyse statistique approfondie, nous avons mis en évidence plusieurs corrélations importantes. Par exemple, les zones présentant une forte concentration de certains polluants, comme le dioxyde de soufre ou les particules fines, affichent une prévalence plus élevée de maladies respiratoires et de cancers. Les départements les plus touchés par la pollution montrent une distribution plus marquée de ces pathologies, renforçant l'hypothèse d'un lien significatif entre pollution et santé.
Ces résultats contribuent à répondre à notre problématique métier et constituent une base solide pour atteindre les objectifs définis. Les prochaines étapes consisteront à approfondir l’analyse des relations identifiées, en se concentrant sur les variables les plus pertinentes pour l’étude.
Enfin, pour aller au-delà de l’analyse descriptive, nous avons proposé de développer des modèles de machine learning, tels que le Random Forest Regressor, afin de prédire le nombre de cas de pathologies par région à partir des données des années précédentes.
Ces modèles permettront d’affiner les prévisions et d’éclairer les prises de décision en matière de santé publique.