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

FISE - A2

UE BC : A Journey To A Data Scientist

Descriptive Statistics

Groupe n°4
Version N°0

Sommaire

I. Introduction

La pollution de l'air et de l'eau en France représente un défi crucial pour la santé publique, avec des impacts significatifs sur les maladies respiratoires et cardiovasculaires. Voici un résumé des points essentiels :

Impacts sanitaires

Mortalité et morbidité

- Environ 40 000 décès prématurés par an sont attribuables à la pollution de l'air.
- Près de 20 000 hospitalisations annuelles sont liées à la pollution atmosphérique.

Maladies spécifiques

- 12% des cas d'asthme infantile sont attribués à la pollution de l'air.
- 23% des décès par maladies cardiovasculaires chez les plus de 65 ans sont liés à l'exposition chronique à la pollution atmosphérique.
- 2 500 cas de cancers du poumon par an sont causés par la pollution de l'air.

Exposition de la population

- 35% des Français sont exposés à des niveaux de pollution atmosphérique dépassant les recommandations de l'OMS.
- Environ 3 millions de personnes sont affectées par la pollution de l'eau potable.
- Un Français sur quatre est exposé à des résidus de pesticides dans l'eau du robinet.

Impacts économiques et sociaux

- Les coûts socio-économiques de la pollution de l'air s'élèvent à 3,8 milliards d'euros par an.
- Plus de 100 000 journées de travail sont perdues annuellement à cause des maladies liées à la pollution.
- L'espérance de vie est réduite en moyenne de 8 mois dans les zones urbaines les plus touchées.

Polluants principaux et leurs effets

- Particules fines (PM2.5) : pénètrent profondément dans l'appareil respiratoire, pouvant causer des maladies cardiovasculaires et des cancers.
- Dioxyde d'azote (NO2) : irritations respiratoires, aggravation de l'asthme.

- Ozone : gêne respiratoire, toux, irritations oculaires.

La pollution reste un enjeu majeur en France, malgré les améliorations observées ces dernières années. Des efforts continus sont nécessaires pour réduire l'exposition de la population et améliorer la santé publique.

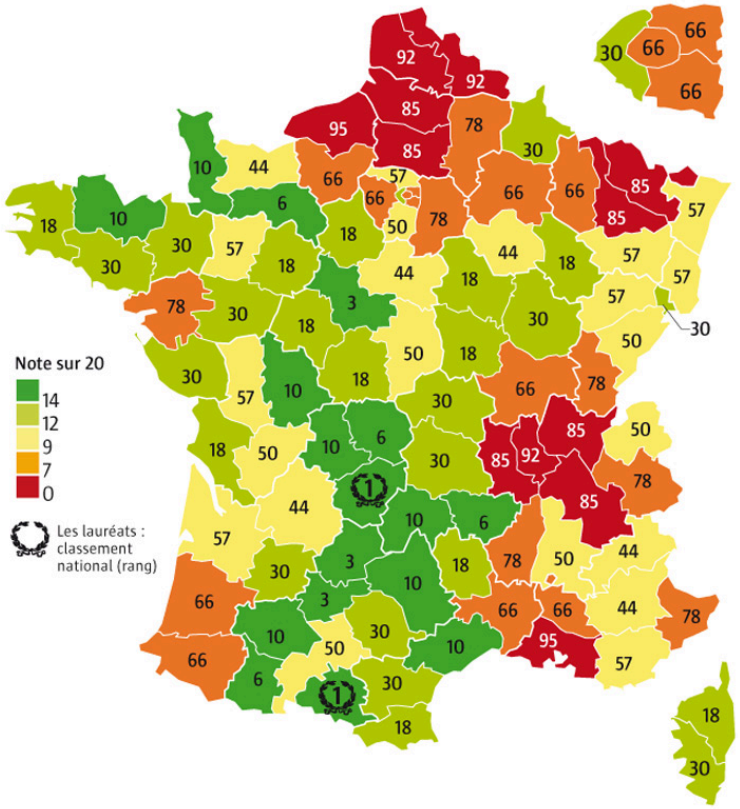

figure n°1: Carte de pollution de la France 2018

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

Problématique

Comment la pollution de l'air et de l'eau influence-t-elle l'incidence des maladies respiratoires et cardiovasculaires et les cancers dans les régions les plus touchées par la pollution ?

II. Analyse des données

1. Collecte des données :

Pour répondre à la problématique, nous avons identifié deux ensembles de données pertinents :

a. Premier jeu de données: "effectifs.csv"
Ce jeu de données fournit des informations détaillées sur les effectifs de patients pris en charge par l'Assurance Maladie en France depuis 2015. Il présente une répartition départementale des patients par pathologie, traitement chronique ou épisode de soins. Les données couvrent un large éventail de catégories médicales, allant des maladies cardio-neurovasculaires aux hospitalisations liées à la Covid-19.

Les catégories sont notamment :

maladies cardio-neurovasculaires
maladies respiratoires chroniques
traitements du risque vasculaire
diabète
cancers
etc …

Les types de variables sont :

annee : Interval
patho_niv1 : Nominal
top : Nominal
cla_age_5 : Nominal
sexe : Nominal
region : Nominal
dept : Nominal
Ntop : Ratio
Npop : Ratio
prev : Ratio
Niveau prioritaire : Nominal
libelle_classe_age : Nominal
libelle_sexe : Nominal
tri : Ratio

b. Deuxième jeu de données: "registre-français-des-émission-polluantes-emissions.csv"
Ce dataset contient des informations sur les émissions polluantes de divers établissements en France. Il inclut des détails sur le type de polluant, la quantité émise, le milieu affecté (air,

eau, sol), ainsi que des informations géographiques et industrielles sur les établissements émetteurs.

Les types de variables sont

Identifiant: Un numéro unique pour chaque observation.

Nom Etablissement: Le nom de l'établissement émetteur.

Annee Emission: L'année au cours de laquelle l'émission a eu lieu.

Milieu: Le milieu dans lequel l'émission s'est produite (air, eau, sol ?).

Polluant: Le type de polluant émis.

Quantité: La quantité de polluant émise.

Unité: L'unité de mesure de la quantité (kg, tonnes, etc.).

Numéro Siret de l'Etablissement: Le numéro de SIREN de l'établissement, un identifiant juridique français.

Libellé APE: Le code APE, qui identifie l'activité principale exercée par l'établissement.

Adresse, Code Postal, Commune, Departement, Region, Coordonnées: Les informations géographiques de l'établissement.

Code INSEE: Le code INSEE de la commune, un identifiant géographique français.

Identifiant : Nominal

Nom Etablissement : Nominal

Annee Emission : Interval

Milieu : Nominal

Polluant : Nominal

Quantité : Ratio

Unité : Nominal

Numéro Siret de l'Etablissement : Nominal

Libellé APE : Nominal

Adresse : Nominal

Code Postal : Nominal

Commune : Nominal

Departement : Nominal

Region : Nominal

Coordonnées : Ratio

Code INSEE : Nominal

2. Gestion des données manquantes :

a. Premier jeu de données: “effectifs.csv”
A première vue, le dataset contient deux colonnes nommée patho_niv 2 et patho_niv 3 qui sont des sous-catégories des pathologies contenant plusieurs valeurs manquantes donc on a décidé d’ignorer ces colonnes et les effacer. La variable de prévalence des maladies, qui est l’une des variables clés , est manquante dans plusieurs lignes, avec aucune méthode qui permet de les prédire efficacement , on a enfin décidé d’effacer les lignes en question.

b. Deuxième jeu de données: “registre-français-des-émission-polluantes-emissions.csv”

La gestion des valeurs manquantes pour la variable "Unité" de chaque polluant s’effectue en deux étapes. Dans un premier temps, pour chaque type de polluant dans la liste “`polluants`”, les valeurs manquantes de la colonne "Unité" sont comblées par la valeur la plus fréquente parmi les unités associées à ce polluant, si celle-ci est définie. Cette approche permet de conserver la cohérence des données en utilisant des valeurs déjà présentes dans le jeu de données. Ensuite, pour les polluants restants (liste “`polluants_restants`”), la valeur "kg/an" est imputée par défaut pour les unités manquantes, afin d’assurer une uniformité des unités pour ces polluants. Cette méthode offre une stratégie à la fois adaptée aux polluants bien représentés et cohérente pour les autres cas.

III. Représentation des données

Pour une bonne compréhension des données , on a tracé différentes courbes

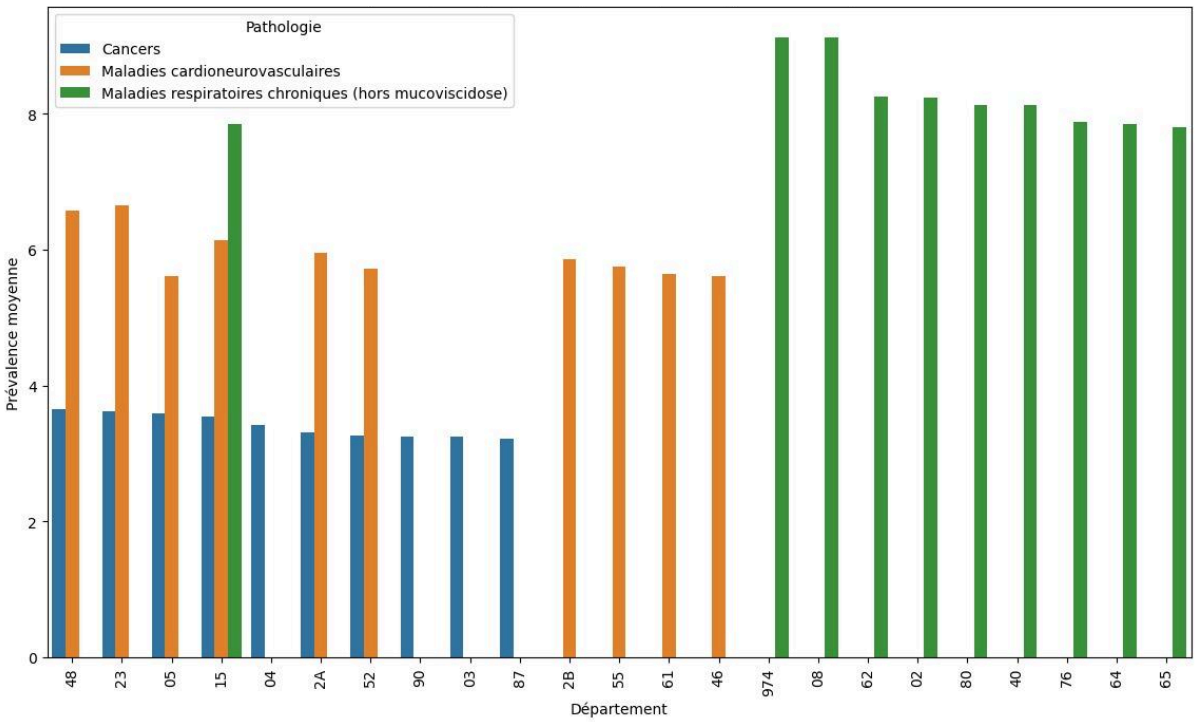

figure n°1: Prévalence moyenne par département pour les pathologies liées à la pollution en 2018

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

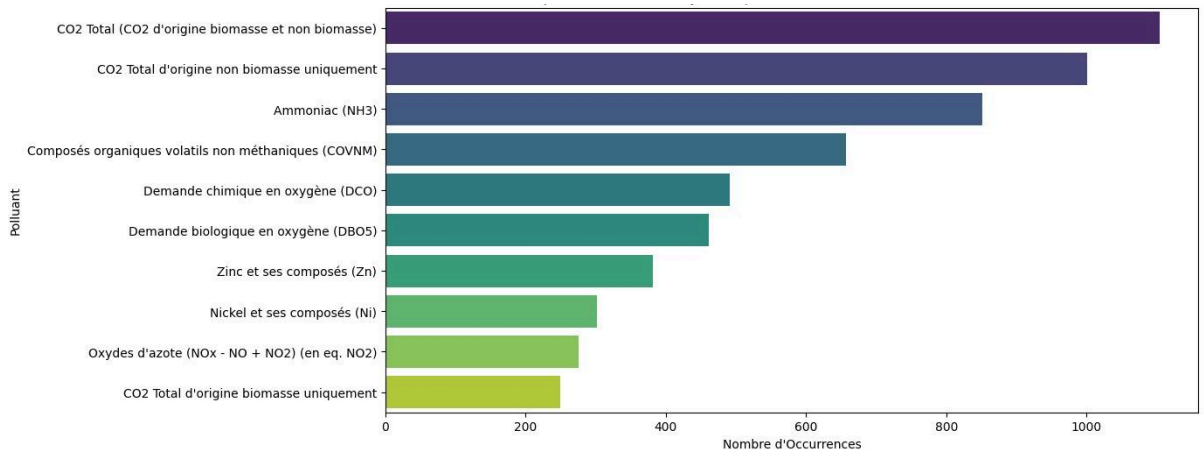

figure n°2: Top 10 des polluants ayant le plus d'occurrence dans le dataset en 2010

Dioxyde de carbone (CO2) d'origine biomasse

Le CO2 d'origine biomasse est le principal polluant en France. Il provient de la combustion ou de la décomposition de matières organiques.

Ammoniac (NH3)

L'ammoniac se classe en deuxième position. Il est principalement émis par l'agriculture, notamment à travers les déjections animales et l'utilisation d'engrais.

Dioxyde de carbone (CO2) d'origine non biomasse

En troisième position, le CO2 d'origine non biomasse est émis par la combustion d'énergies fossiles et les processus industriels.

Ces trois polluants illustrent les principales sources de pollution en France.

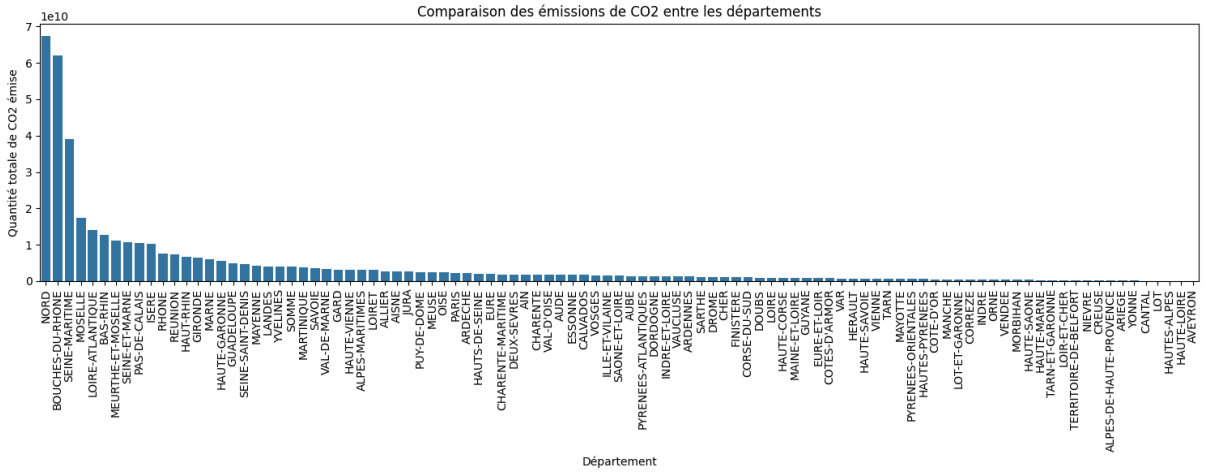

Figure 3 : Émissions de CO2 par département depuis 2004

Le graphique présente une répartition très hétérogène des émissions de CO2 entre les différents départements français. On observe une décroissance rapide des émissions à partir du Nord, suggérant une forte concentration de sources d'émissions dans les premiers départements classés.

Le département avec les émissions de CO2 les plus élevées est le Nord (59) suivie du Rhône (13) , ce qui peut expliquer le nombre de pathologies observées dans ce département.

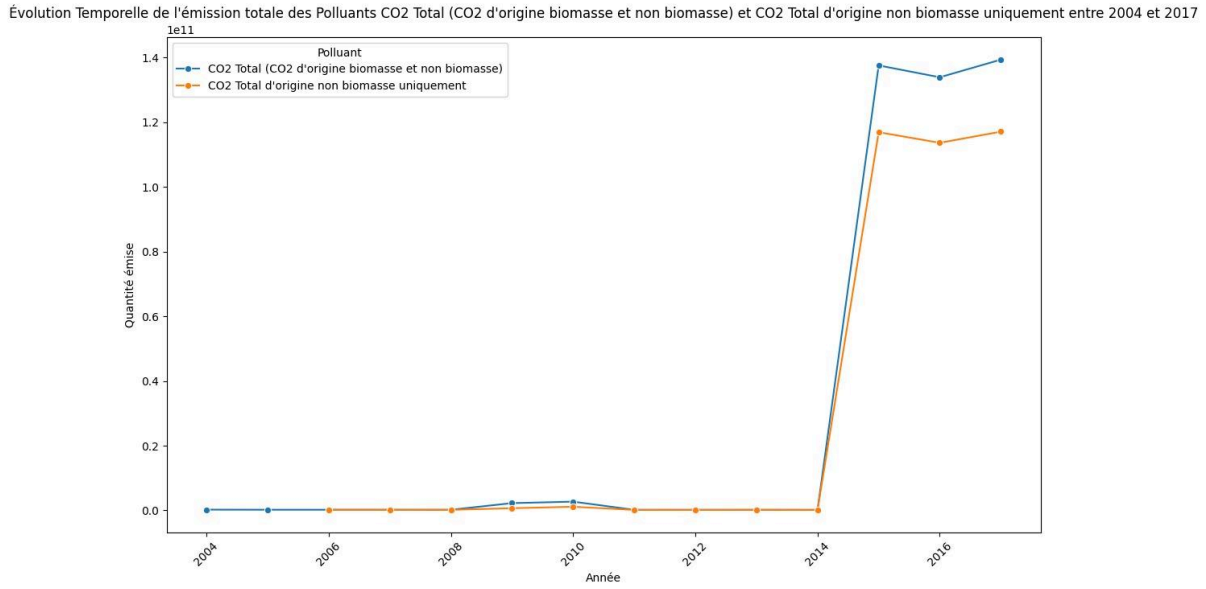

Figure 4 : Évolution temporelle de l'émission de CO2 entre 2004 et 2017

Le graphique présente l'évolution des émissions de CO_2 provenant de la biomasse et non-biomasse ainsi que du CO_2 issu uniquement de la biomasse sur la période de 2004 à

2017. On observe une tendance générale de fluctuations pour les deux types de CO_2, avec un pic notable en 2014. Ce pic représente un sommet dans les émissions, tant pour le CO_2 total (biomasse et non-biomasse) que pour le CO_2 uniquement issu de la biomasse. La différence entre les niveaux de CO_2 total et ceux provenant uniquement de la biomasse reste non significative tout au long de la période. Après 2014, les niveaux des deux types d'émissions restent quasiment stables.

IV. Etude de Corrélation

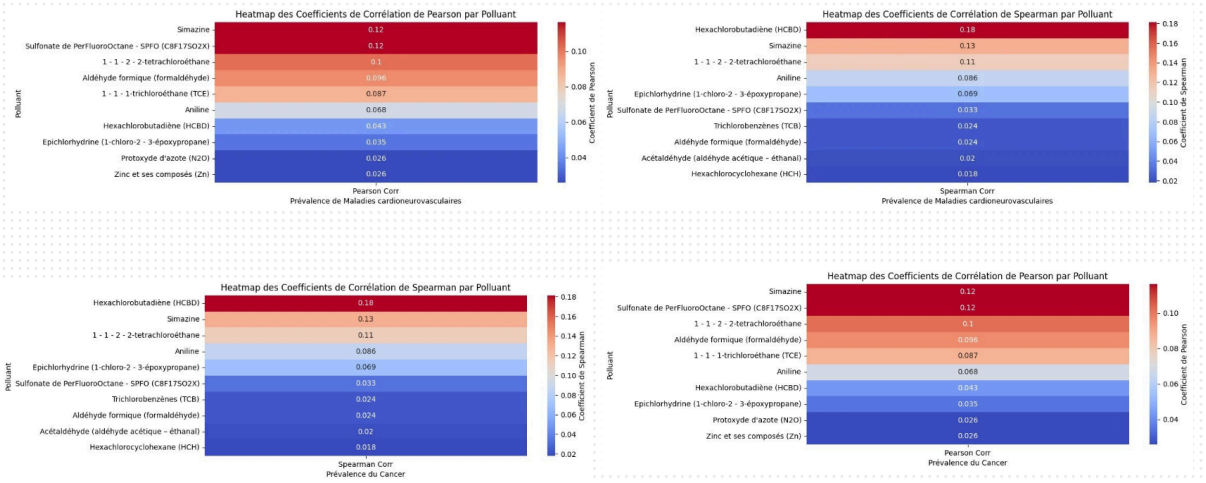

Figure 5 : Corrélation entre quelques polluants et maladies cardioneurovasculaires /Cancer

Ces graphes montrent le degré de corrélation entre quelques polluants et la prévalence du cancer / maladies cardio neurovasculaires selon Spearman et Person

	Polluant 1	Polluant 2	Coefficient de corrélation
0	CO2 Total (CO2 d'origine biomasse et non bioma...	CO2 Total d'origine biomasse uniquement	0.994467
1	Sulfure d'hydrogène (H2S)	Crésol (mélange d'isomères)	0.998752
2	Demande biologique en oxygène (DBO5)	Demande chimique en oxygène (DCO)	0.999323
3	Nickel et ses composés (Ni)	Plomb et ses composés (Pb)	0.990198
4	Chrome et ses composés (Cr)	Aluminium et ses composés (Al)	0.999864
5	Chrome et ses composés (Cr)	Fer et ses composés (Fe)	0.999764
6	Chrome et ses composés (Cr)	1 - 2-dichloroéthane (DCE - chlorure d'éthylène)	0.999164
7	Chrome et ses composés (Cr)	Chloroforme (trichlorométhane)	0.999223
8	Chrome et ses composés (Cr)	Chlorures (Cl total)	0.999722
9	Chrome et ses composés (Cr)	Nonylphénols	0.993997
10	Chrome et ses composés (Cr)	Octylphénols	0.994667
11	Aluminium et ses composés (Al)	Fer et ses composés (Fe)	0.999834
12	Aluminium et ses composés (Al)	1 - 2-dichloroéthane (DCE - chlorure d'éthylène)	0.999466
13	Aluminium et ses composés (Al)	Chloroforme (trichlorométhane)	0.999370
14	Aluminium et ses composés (Al)	Chlorures (Cl total)	0.999937
15	Aluminium et ses composés (Al)	Nonylphénols	0.993801
16	Aluminium et ses composés (Al)	Octylphénols	0.994723
17	Hexachlorocyclohexane (HCH)	Trichlorobenzènes (TCB)	0.995812
18	Hexachlorocyclohexane (HCH)	Trifluorure d'azote (NF3)	0.999181

Figure 6 :Corrélation entre les différents polluants

Ce tableau permet de lire le degré de corrélation entre quelques polluants. Pour alléger la lecture on a seulement pris quelques exemples. Les corrélations entre le reste des variables est dans le notebook et on peut choisir le seuil de corrélation

Pour un seuil de 0.9, on a plus que 400 couples de polluants.On peut aisément dire que nos données sont très corrélées.

V. conclusion

Cette étude a permis de mettre en évidence des relations intéressantes entre les variables issues des deux jeux de données, à savoir les données de pathologies et celles sur la pollution de l'air et de l'eau. En combinant ces deux ensembles d’informations, nous avons observé plusieurs corrélations importantes.

Le premier jeu de données, couvrant les pathologies par région et tranche d'âge entre 2016 et 2022, inclut des variables telles que l'année, la région, la classe d'âge, le sexe, et des informations spécifiques sur les types de pathologies (comme les cancers ou les maladies respiratoires).

Le second jeu de données, relatif à la pollution de l'air et de l'eau entre 2004 et 2017, contient des informations sur les polluants, les quantités émises et les zones géographiques associées (départements et régions).

L’analyse statistique a révélé plusieurs corrélations pertinentes. Par exemple, les zones avec une forte concentration de polluants spécifiques, comme le dioxyde de soufre ou les particules fines, présentent également une prévalence plus élevée de maladies respiratoires et de cancers. Les départements les plus touchés par la pollution montrent une distribution plus marquée des pathologies.

6.Annexe :

Le lien pour le Notebook :
Projet DS.ipynb