Clean Air

— Batita Yahya —
— Schneider Aymeric —

1. Contexte

Dans le contexte d'une France industrielle, de nombreuses molécules identifiées comme étant polluantes sont rejetées dans l'atmosphère et sont respirées quotidiennement par tous. Certes, des efforts sont mis en œuvre pour limiter cette pollution mais elle est toujours présente et est connue pour causer des troubles de la santé. Nous cherchons ici à trouver des corrélations entre la présence de molécules polluantes avec les cas de maladie respiratoires chroniques déclarés en France.

2. Sources de Data

Nous avons utilisé plusieurs sources d'informations pour les combiner entre elles. elle sont toute de 2019 :

Les concentrations moyennes annuelles de différents agents polluants

Qui sont prélevées un peu partout en France dans environ 400 stations de mesures.

La prévalence des maladies chroniques respiratoires

Ceci représente les chances d'attraper dite maladie par rapport à la moyenne nationale (un score élevé indique que la population est plus touchée que la normale) ces informations sont données par département. Ce sont les maladies chroniques qui sont prises en compte ici, c'est à dire les maladies nécessitant un traitement long terme (ex: asthme).

La carte des frontières des départements français

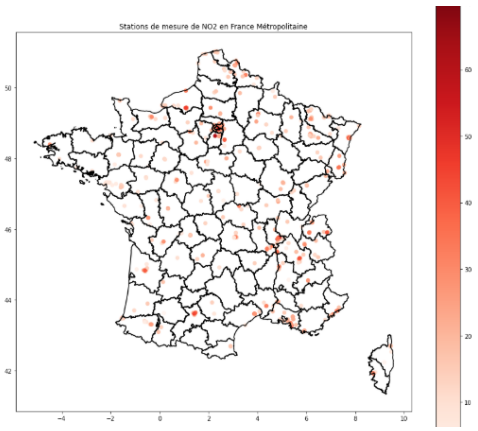
Utilisée pour produire les graphes et certains calculs

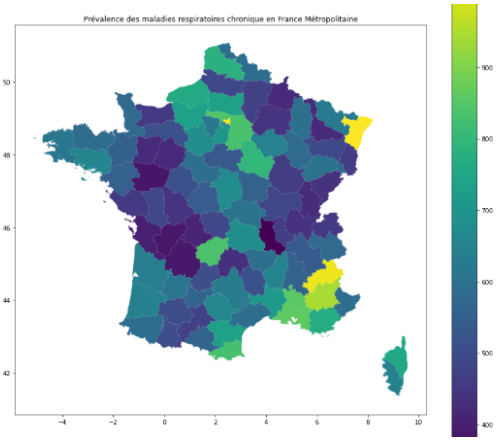
Il est à noter que nous avons aussi demandé l'avis d'une étudiante en 5ème année de médecine qui nous a aidé à peaufiner notre modèle et à débattre des résultats obtenus.

3. Preprocessing

a. Mise en forme

Nous avons utilisé un programme Sparks pour mettre en forme les datasets téléchargés au préalable. Puis nous les avons chargés dans un notebook Jupyter sous la forme de Dataframes Pandas. Nous avons ensuite dû parser les positions GPS afin d'obtenir une GeoDataframe de la bibliothèque GeoPandas. On peut alors visualiser les stations de mesures ainsi que la prévalence des maladies :

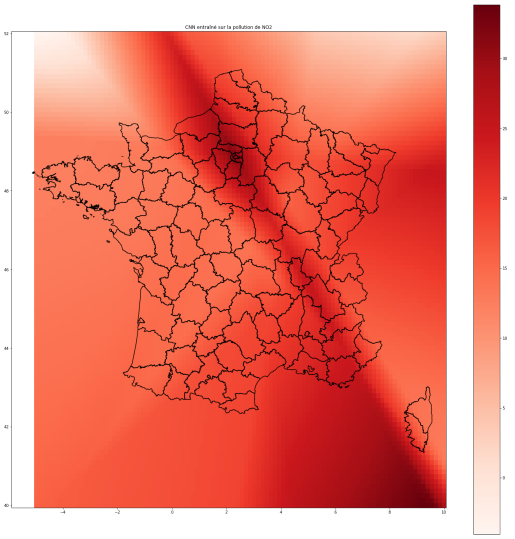




b. Scores de départements

Étant donné que les informations sur les maladies respiratoires sont données par département, nous devons donner un "score de pollution" à ces derniers afin de pouvoir comparer.

Nous avons d'abord tenté d'entraîner un CNN avec les données des stations qui sont connues ceci dans l'espoir qu'il pourrait nous donner une bonne estimation de la pollution à n'importe quel point de la France. Cependant nous n'avions de toute évidence pas assez de data, le CNN ne converge pas tout le temps vers le même résultat et de toute façon voici ce que cela donne :



Nous avons donc employé une méthode plus simple et plus approximative. Pour calculer le score de pollution d'un département :

Soit le département contient une ou plusieurs stations de mesure, alors nous en faisons la moyenne.

Dans le cas où il n'y a pas de stations (ce qui est possible) alors nous cherchons les 3 stations les plus proches du centre du département et faisons la moyenne de ces dernières.

Selon cette logique, on obtient le tableau suivant :

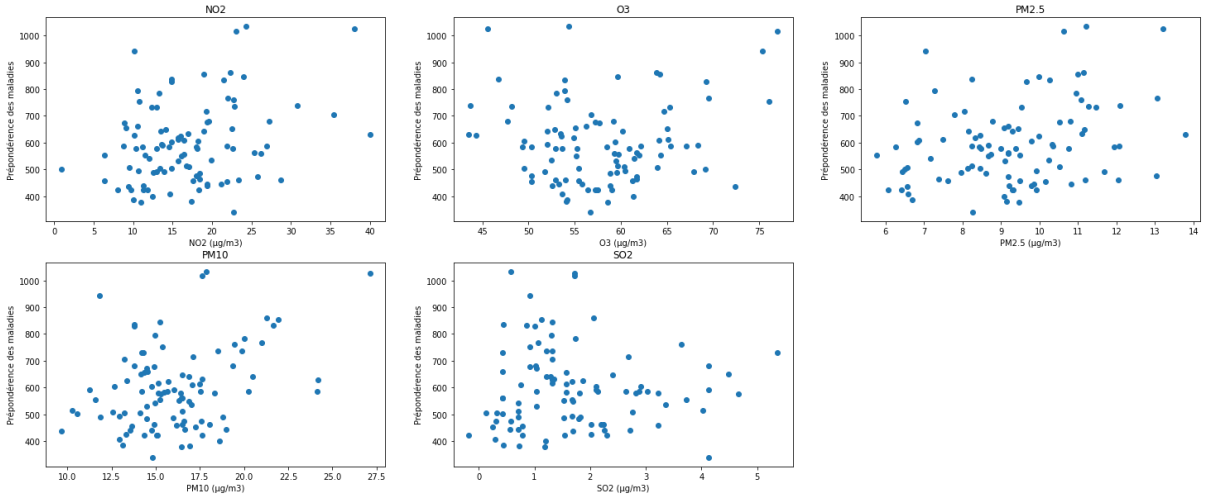
	code	nom	geometry	disease	NO2	O3	PM2.5	PM10	SO2
0	01	Ain	POLYGON ((4.78021 46.17668, 4.78024 46.18905, ...	440	11.309661	52.565980	9.205202	13.531373	2.711290
1	02	Aisne	POLYGON ((3.17296 50.01131, 3.17382 50.01186, ...	475	17.966328	50.336674	13.036675	17.564406	0.311225
2	03	Allier	POLYGON ((3.03207 46.79491, 3.03424 46.79080, ...	603	14.849769	59.426090	6.817638	12.670063	2.914878
3	04	Alpes-de-Haute-Provence	POLYGON ((5.67604 44.19143, 5.67817 44.19051, ...	943	10.177062	75.392374	7.035978	11.810604	0.923570
4	05	Hautes-Alpes	POLYGON ((6.26057 45.12685, 6.26417 45.12641, ...	1016	22.992953	76.979542	10.633081	17.641022	1.716219
..	...	...	...	...	...	...	...	...	...
91	91	Essonne	POLYGON ((2.22656 48.77610, 2.22866 48.77451, ...	705	35.398132	56.800330	7.781828	13.214885	1.318656
92	92	Hauts-de-Seine	POLYGON ((2.29097 48.95097, 2.29162 48.95077, ...	737	30.804433	43.674274	12.087967	19.889085	1.216686
93	93	Seine-Saint-Denis	POLYGON ((2.55306 49.00982, 2.55814 49.01201, ...	1026	38.064120	45.586676	13.203023	27.125083	1.715832
94	94	Val-de-Marne	POLYGON ((2.33190 48.81701, 2.33371 48.81677, ...	680	27.284078	47.702316	10.797380	19.379992	1.023450
95	95	Val-d'Oise	POLYGON ((2.59052 49.07965, 2.59013 49.07786, ...	845	23.961001	59.650674	9.989229	15.259122	1.318656

Nous avons ainsi plusieurs scores pour chaque département, nous pouvons à présent y appliquer divers modèles d'analyse mathématique pour en tirer des informations utiles.

4. Résultats

a. Les gaz suspects

Tout d'abord nous avons produit un nuage de points pour chaque gaz. Les points représentent chacun un département et ont pour abscisse le score environnemental et pour ordonnée la prépondérance aux maladies respiratoires chronique.



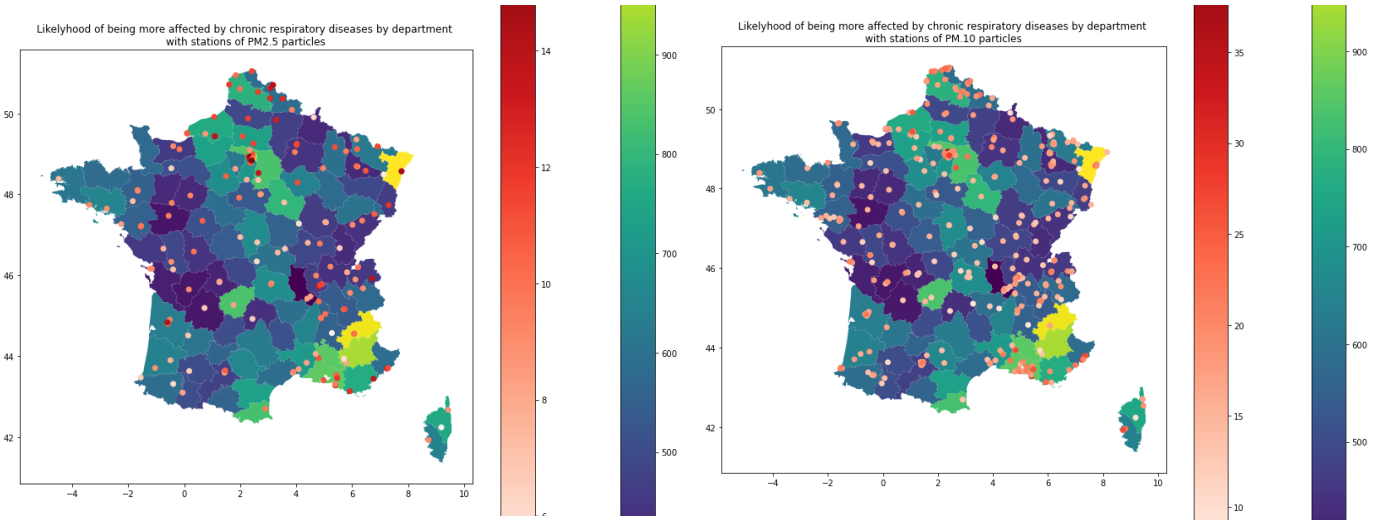
On voit déjà que certains gaz suivent mieux une droite que d'autres, par exemple le graphe PM10 (Petites molécules de moins de 10 micromètres) est plus proche d'une droite que le graphe SO2 (dioxyde de soufre).
Les Dataframes pandas ont une fonction très utile, corr(), qui permet de calculer la matrice de corrélation directement, elle donne :

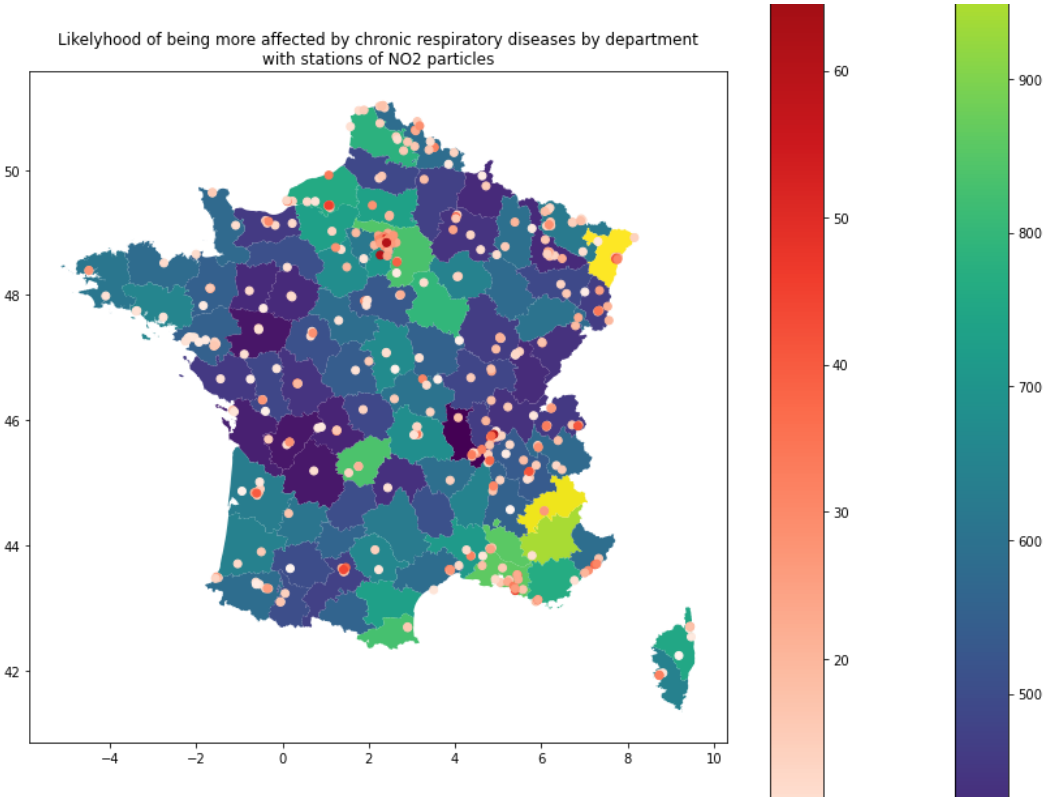
```
3  franceData.corr()
✓  0.4s
```

	disease	NO2	O3	PM2.5	PM10	SO2
disease	1.000000	0.295928	0.091996	0.289416	0.338127	-0.002247
NO2	0.295928	1.000000	-0.289428	0.587708	0.615162	0.088869
O3	0.091996	-0.289428	1.000000	-0.383225	-0.310568	-0.008858
PM2.5	0.289416	0.587708	-0.383225	1.000000	0.698226	0.082152
PM10	0.338127	0.615162	-0.310568	0.698226	1.000000	-0.069480
SO2	-0.002247	0.088869	-0.008858	0.082152	-0.069480	1.000000

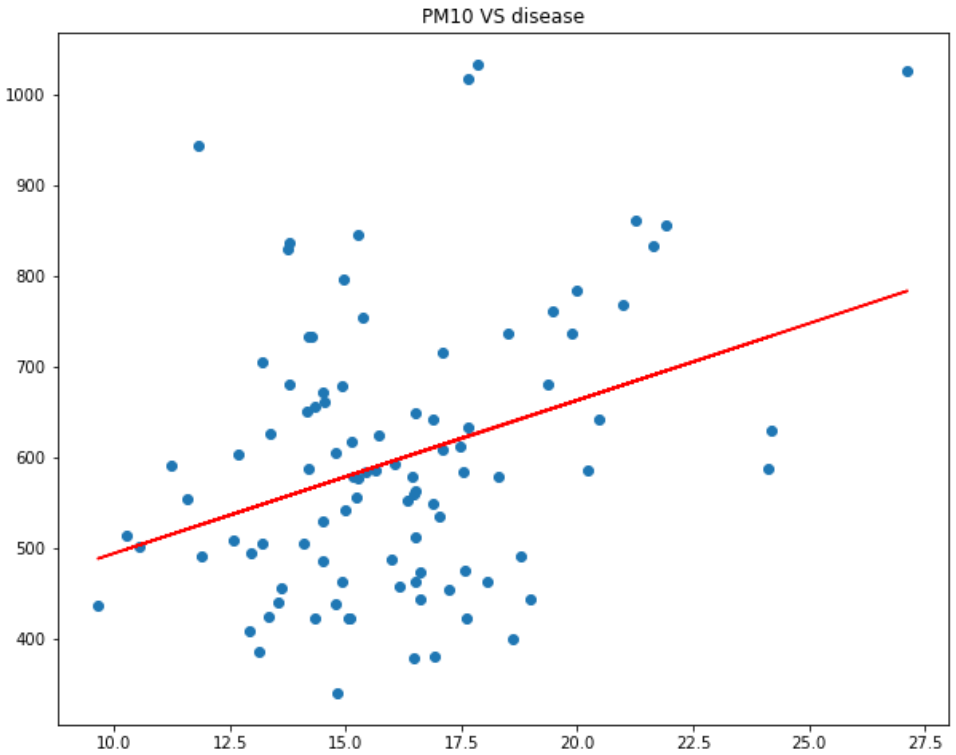
Nous nous intéressons à la première colonne ou la première ligne, on y voit qu'il ressort deux catégories de molécules, celles avec un taux de corrélation non négligeable (entre 29 et 34%) et celles qui sont presque pas du tout corrélées (0 et 9%).

Ce tableau suggère fortement de se concentrer sur les molécules NO2 et les petites molécules PM10 et PM2.5. Voici ce que l'on obtient si on superpose les deux sur une carte :


Likelyhood of being more affected by chronic respiratory diseases by department
with stations of PM2.5 particles
Likelyhood of being more affected by chronic respiratory diseases by department
with stations of PM.10 particles


Likelyhood of being more affected by chronic respiratory diseases by department
with stations of NO2 particles

Une régression linéaire nous donne (exemple de PM10) :



Arrivé à ces résultats, nous avons songé à appliquer un modèle de clustering étant donné que le sujet du projet porte en partie sur le machine learning. Cependant nous avons jugé cela non nécessaire car peu de plus value serait ajoutée et ce ne serait qu'une démonstration de notre capacité à en faire.

b. Comparaison aux articles médicaux

Lorsque l'on cherche à se renseigner sur l'impact de la pollution atmosphérique sur la santé, de nombreux articles ressortent. Nous avons sélectionné ces deux-ci issus de santepubliquefrance.fr :

Pollution atmosphérique : quels sont les risques ?

Pollution de l'air ambiant : nouvelles estimations de son impact sur la santé des Français

Ces deux articles sont clairs : les particules les plus nocives pour la santé sont le NO2 et les PM10 et 2.5. Plus de 40 000 morts par an sont attribuées à ces molécules. L'ozone (O3) est cité une fois de manière anecdotique et le SO2 est introuvable.
Ceci confirme bien nos propres conclusions et semble vérifier notre modèle.

5. Conclusion et débats

Grâce à des données publiques et gratuites, nous avons pu calculer des scores de pollution et pour chaque département et les comparer aux prépondérances de maladie respiratoires chroniques. Nous en avons fait ressortir le lien entre ces dernières et certaines des particules polluantes présentes dans l'atmosphère et ceci s'est vérifié à travers des articles de la santé publique.
Cependant notre modèle n'est pas infaillible. Déjà les données atmosphériques étaient peu nombreuses, en tout cas largement insuffisantes pour y appliquer des modèles de machine learning.
De plus, l'étudiante en médecine nous a fait remarquer que les données que nous avons obtenues sur les maladies respiratoires comprennent la BPCO, une maladie attribuée aux fumeurs. Elle nous avait conseillé de nous concentrer sur les maladies comme l'asthme et en particulier chez les jeunes enfants. Mais nous n'avions malheureusement pas accès à ces données aussi précises.