RAPPORT
ASTHMAWATCH

Abdou Wahab DOSSOU
Belkis ABDELATIF
Dylan PIN
Hiba AZZOUZI
Saad OMAR ABDILLAHI

M1 Data Science en santé
2025

1 Introduction

L’asthme est une maladie respiratoire qui se manifeste à travers des crises durant lesquelles il existe une gêne respiratoire (dyspnée) et des sifflements. Cette maladie touche plus de 4 millions de personnes en France. L’asthme sévère concerne environ 5 % des patients et est associé à près de 900 décès par an. Parmi les facteurs environnementaux influençant son évolution, l'exposition aux polluants atmosphériques et aux pollens est reconnue pour jouer un rôle clé dans les exacerbations aiguës de l’asthme. (1)

Les pollens sont des déclencheurs majeurs des crises d'asthme chez les personnes sensibilisées. Plusieurs études ont démontré une corrélation entre la concentration de certains pollens dans l'air et l'augmentation des visites aux urgences pour asthme. (2)

Les polluants atmosphériques sont reconnus comme des facteurs aggravants de l'asthme. Leur inhalation provoque une inflammation des voies respiratoires, augmentant la fréquence et la sévérité des crises. (3) Une étude menée par le Stockholm Environment Institute (SEI) a estimé que la pollution de l’air est responsable de plusieurs millions de visites aux urgences pour asthme chaque année à travers le monde. (4)

Ce projet a pour objectif de fournir une analyse indicative des relations entre les concentrations de pollens, les niveaux de pollution atmosphérique et les taux de passage aux urgences pour asthme en France. En utilisant des données à l'échelle départementale et communale, nous cherchons à offrir une vue d'ensemble des tendances observées entre ces variables. Bien que l'objectif ne soit pas de démontrer des corrélations précises par manque de données brutes, cette analyse vise à sensibiliser sur l'impact environnemental sur la santé respiratoire et à informer les décideurs et le public sur les risques potentiels associés aux facteurs environnementaux.

2 Outils et données exploités

Dash : framework Python permettant de créer des applications web interactives, souvent utilisées pour le développement de tableaux de bord.

Plotly : bibliothèque de visualisation de données en Python qui permet de créer des graphiques interactifs et dynamiques

BeautifulSoup : bibliothèque Python dédiée à l’analyse de documents HTML et XML, facilitant l'extraction de données structurées à partir de pages web.

Selenium : outil de scraping pour applications web qui permet d'automatiser les interactions avec les sites web.

AWS App Runner : service de déploiement et de gestion d’applications web serverless.

Gitlab : plateforme de gestion de code source et de DevOps qui facilite le contrôle de version et la collaboration.

3 Méthodologie

3.1. Web Scraping

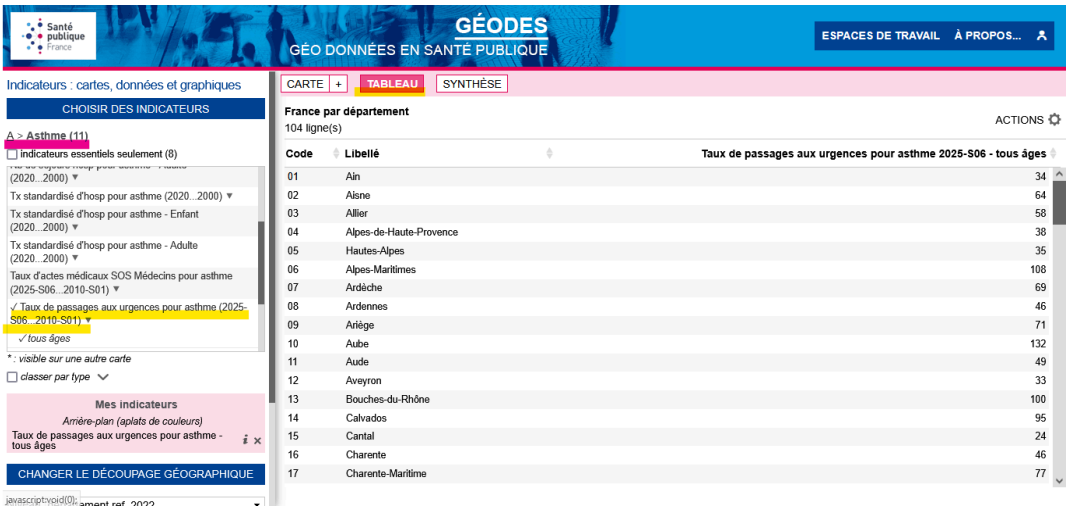
Extraction des données depuis Géodes Santé Publique France

Présentation du site cible :

Le site Géodes est une plateforme interactive qui permet d'explorer des indicateurs de santé publique sous forme de cartes, tableaux et graphiques.

Donnée cible :

Le site met à jour un tableau de taux de passage aux urgences pour asthme par département de façon hebdomadaire, c'est cette donnée là que nous récupérons.



Interactions nécessaires : L'objectif étant de récupérer les valeurs du tableau, la visualisation du tableau nécessite :

- Ouverture de la page
- Effectuer une recherche dans la barre de recherche
- Cliquer sur des éléments (ex: le bouton "OK", le menu déroulant "tous âges", l'onglet "tableau")
- Faire défiler la page pour afficher certains éléments invisibles

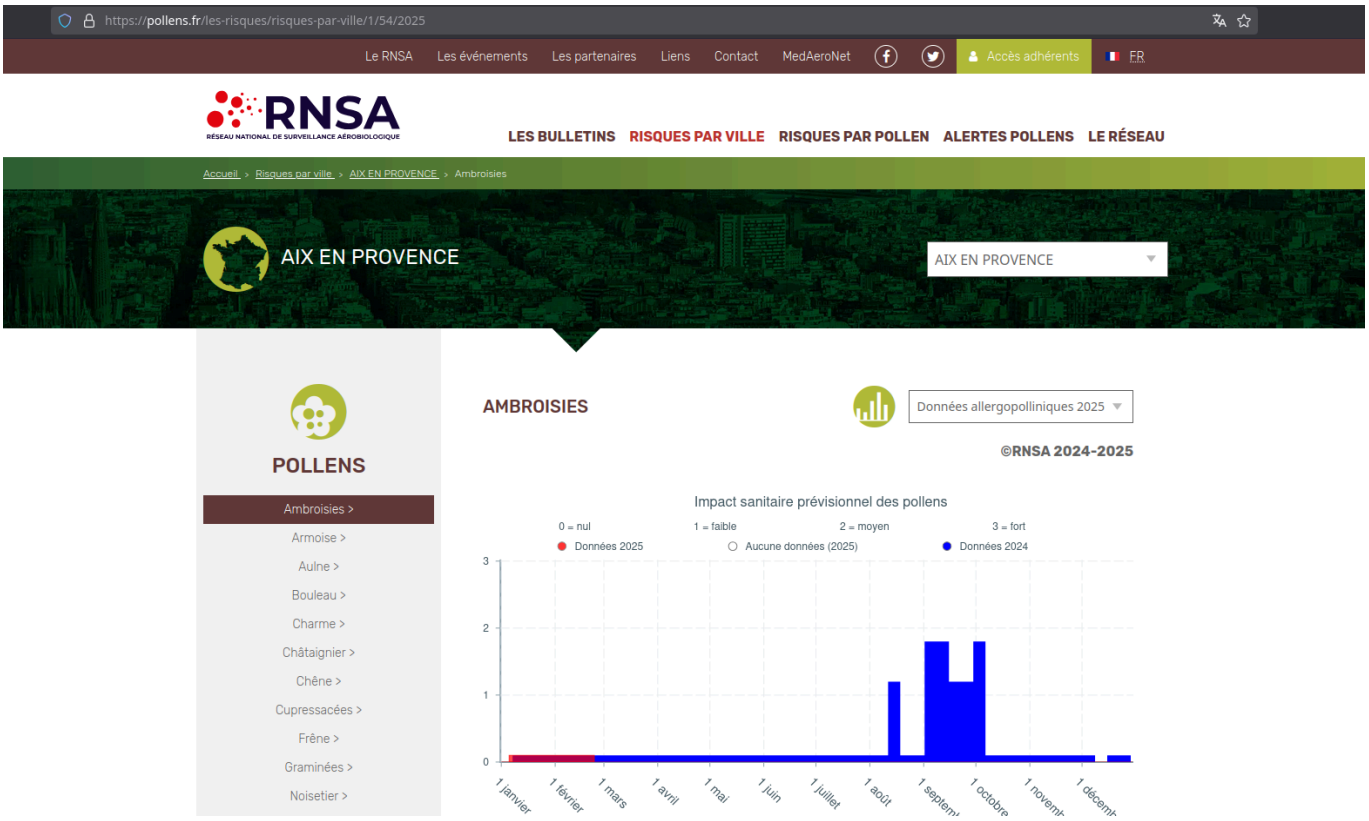
Obstacle rencontré :

Sur Géodes, plusieurs éléments n'ont pas d'ID fixe, donc impossible de les cibler directement avec find_element(By.ID, "ID_element"). Certains éléments changent de classe en fonction de leur état (ex. : déplié ou non). XPath permet de chercher un élément en fonction de son texte, ce qui est crucial pour des boutons comme "Taux de passages aux urgences" qui ne possèdent pas un identifiant stable.

Extraction et mise en forme des données :

Une fois les données hebdomadaires extraites, elles sont intégrées dans un fichier Excel geodes_complet.xlsx.
Comme les mises à jour des données sur Géodes sont hebdomadaires mais sans jour précis connu, il est possible que le script de scrapping sele_ajout_csv.py soit exécuté avant que les nouvelles données ne soient disponibles. Pour éviter d'enregistrer plusieurs fois les mêmes valeurs, le script compare les nouvelles entrées avec celles déjà stockées et ne rajoute pas de ligne inutile si elles sont identiques.

Extraction des données depuis Pollens



Présentation du site cible :

Le site pollens.fr est une plateforme dédiée à la surveillance des risques allergiques liés aux pollens en France. Il fournit des informations en temps réel sur les niveaux de pollen par ville et par type de pollen, ainsi que des données historiques pour permettre une analyse comparative.

Donnée cible :

Les données ciblées incluent les niveaux de risque allergique par type de pollen (graminées, bouleau, etc.) et par ville, ainsi que les données historiques pour l'année en cours et l'année précédente.

Choix de la méthode :

Pour extraire les données, un web scraping est nécessaire, en utilisant des outils comme BeautifulSoup pour parser le HTML et requests pour interagir avec le site.
Interactions nécessaires :
L'extraction des données nécessite les étapes suivantes :

- Accéder à une page avec graphique d'un pollen pour une ville pour une période temporelle pour récupérer les listes des villes et des types de pollen disponibles dans le code html par identification d'un balise cible.
- Construire des URLs spécifiques pour chaque combinaison ville/pollen afin d'accéder aux données détaillées. Le scrapper parcourt toutes les pages identifiées une à une.
- Extraire les données des scripts JavaScript contenant les informations sur les niveaux de pollen (format proche d'un dataFrame pour les objets 'graphData' et 'previousYearGraphData'.
- Structurer et sauvegarder les données dans un fichier CSV pour une utilisation ultérieure.

Extraction et mise en forme des données :

Le script Python développé pour cette tâche effectue les opérations suivantes :

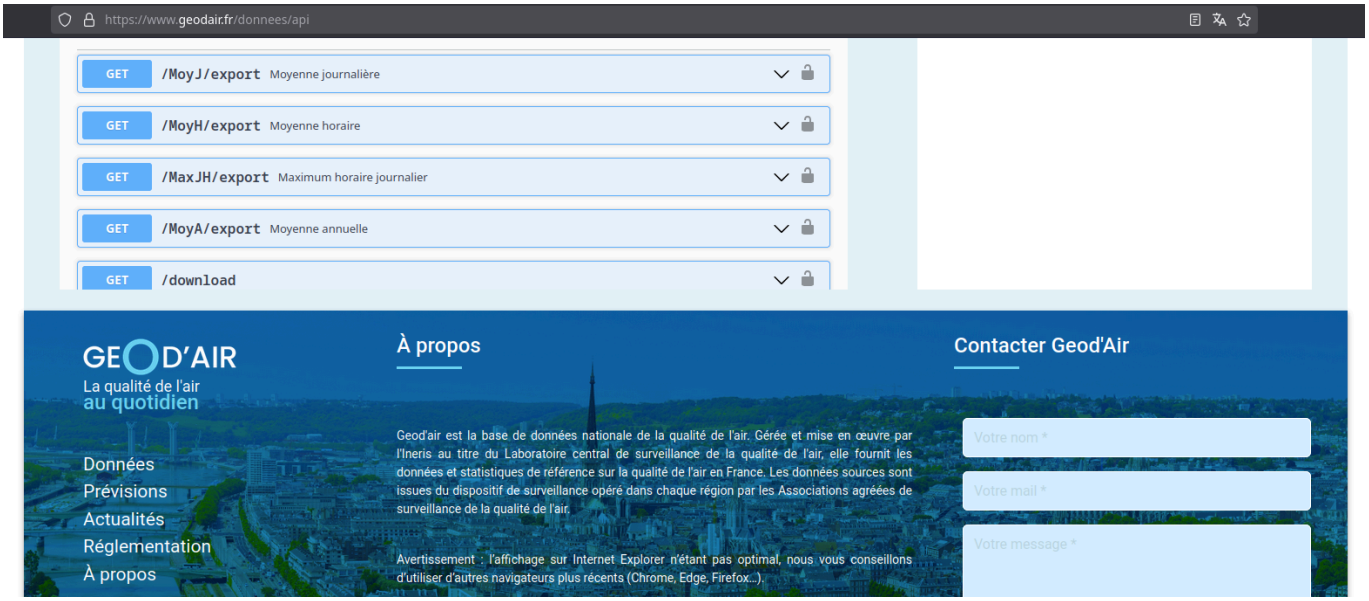
- Récupération des options : Les listes des villes et des types de pollen sont extraites à partir des balises <select> du site.
- Construction des URLs : Pour chaque combinaison ville/pollen, une URL spécifique est générée pour accéder aux données détaillées.
- Extraction des données : Les données sont extraites des scripts JavaScript en utilisant des expressions régulières (regex) pour identifier et parser les objets JSON.
- Sauvegarde des données : Les données extraites sont structurées dans un DataFrame Pandas et sauvegardées dans un fichier CSV (pollen.csv).

NB : le format du fichier n'a pas été modifié car utilisé par la suite dans ce format, il aurait été parcimonieux de normaliser l'ensemble des titres de colonne.

Fonctionnalités clés du script :

- Gestion des erreurs : Le script vérifie les codes de réponse HTTP et gère les erreurs de parsing JSON.
- Extraction des données historiques : Les données de l'année en cours et de l'année précédente sont extraites pour permettre une analyse comparative.
- Automatisation : Le script parcourt automatiquement toutes les combinaisons ville/pollen, ce qui permet une extraction complète des données disponibles.

Extraction des données depuis Geodair



Présentation du site cible :

Le site Geodair est une plateforme dédiée à la surveillance de la qualité de l'air en France. Il fournit des données en temps réel et historiques sur les concentrations de polluants atmosphériques sur l'ensemble du territoire français. Possibilité d'avoir plusieurs stations dans une ville, et plusieurs villes dans un département. Les indicateurs de concentration des polluants varient d'un site de prélèvement à un autre, un site renseigné peut avoir un ou plusieurs polluants suivis.

Donnée cible :

Les données ciblées incluent les concentrations horaires et journalières des polluants atmosphériques, ainsi que les informations géographiques des stations de mesure (coordonnées GPS, communes, départements).

Choix de la méthode :

Geodair met à disposition une API permettant d'accéder aux données de qualité de l'air. Cette API nécessite une clé d'authentification pour interagir avec les endpoints. L'extraction des données est réalisée en utilisant des requêtes HTTP via la bibliothèque requests en Python. Les données sont ensuite structurées et sauvegardées dans des fichiers CSV pour une utilisation ultérieure.
Interactions nécessaires :
L'extraction des données nécessite les étapes suivantes :

1. Authentification : Utilisation d'une clé API pour accéder aux endpoints de l'API Geodair.
2. Récupération des données horaires : Extraction des concentrations horaires des polluants pour la journée en cours.
3. Récupération des données journalières : Extraction des pics journaliers de concentration pour chaque polluant.
4. Agrégation des données hebdomadaires : Calcul des maximums hebdomadaires à partir des données journalières.
5. Enrichissement des données : Ajout des informations géographiques (communes, départements, coordonnées GPS) aux données de pollution.

Obstacles rencontrés :

Authentification : La nécessité d'une clé API pour accéder aux données limite l'automatisation et nécessite une gestion sécurisée des identifiants.
Données dynamiques : Les données sont générées en temps réel, ce qui peut entraîner des délais dans la disponibilité des fichiers à télécharger.
Structure des données : Les fichiers CSV retournés par l'API nécessitent un nettoyage et une structuration pour être utilisables (nettoyage des en-têtes, gestion des doublons, conversion des dates).

Extraction et mise en forme des données :

Le script Python développé pour cette tâche effectue les opérations suivantes :

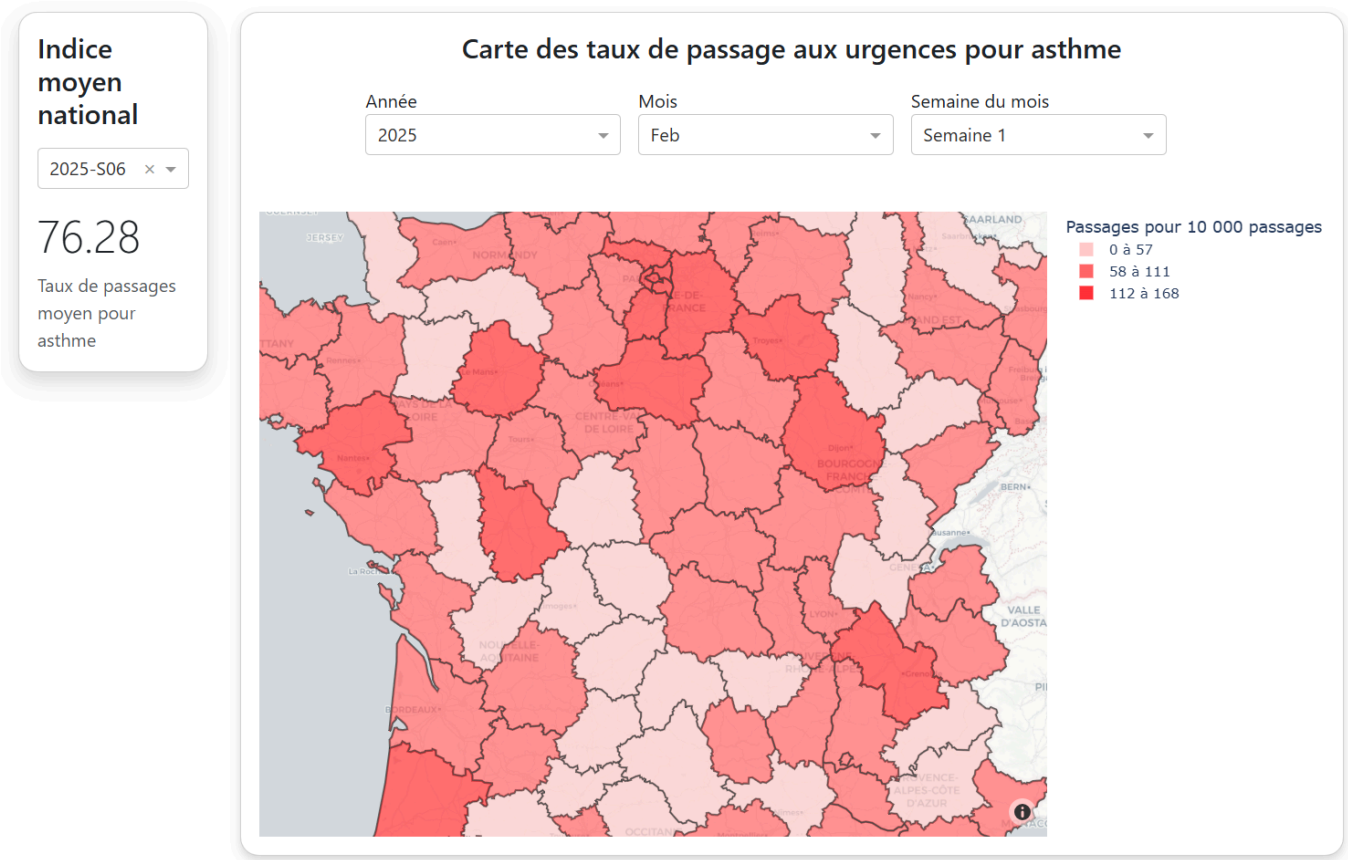
1. Récupération des données journalières : Les pics journaliers de concentration sont extraits pour chaque polluant et ajoutés à un fichier CSV historique (geodair_max_daily.csv).
2. Agrégation des données hebdomadaires : Les maximales hebdomadaires sont calculées à partir des données journalières et sauvegardées dans un fichier CSV (geodair_max_weekly.csv).
3. Enrichissement des données : Les informations géographiques des stations de mesure sont ajoutées aux fichiers de données pour permettre une analyse spatiale.
4. Nettoyage et structuration : Les fichiers CSV sont nettoyés (suppression des doublons, conversion des dates, réorganisation des colonnes) pour garantir leur qualité et leur cohérence.

Fonctionnalités clés du script :

- Gestion des erreurs : Le script vérifie les codes de réponse HTTP et gère les erreurs de téléchargement ou de parsing des données.
- Automatisation : Le script peut être exécuté de manière récurrente pour mettre à jour les données quotidiennement ou hebdomadairement.
- Calcul de l'IQA : Un Indice de Qualité de l'Air (IQA) est calculé pour chaque station de mesure en fonction des concentrations de polluants.

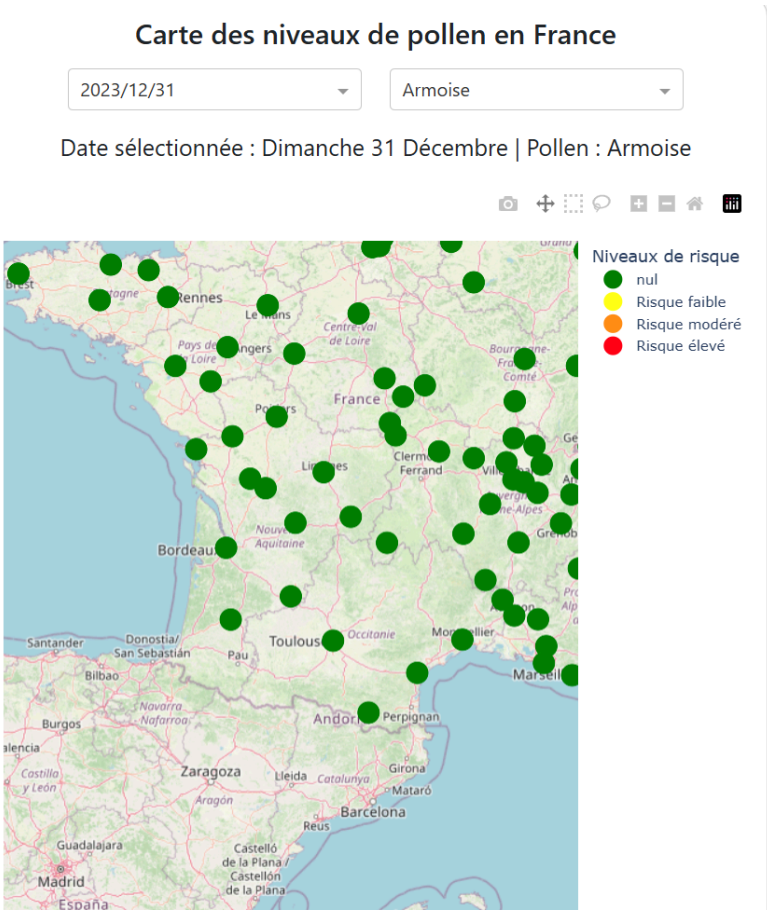
3.2. Tableau de bord et visualisation

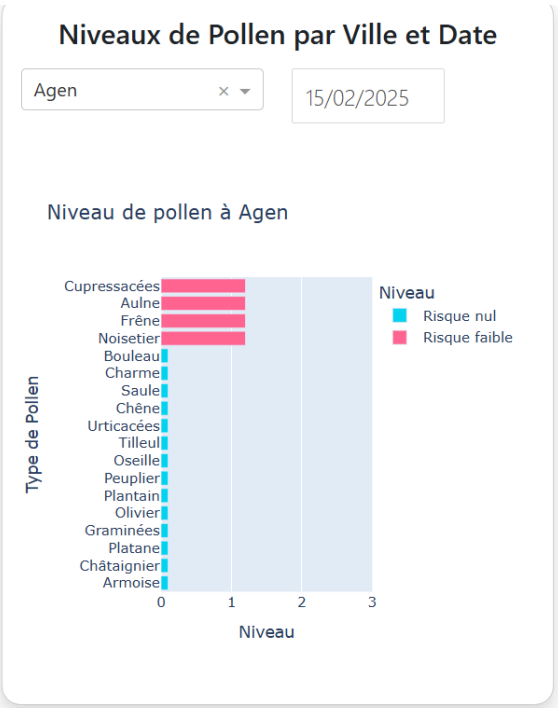
Overview



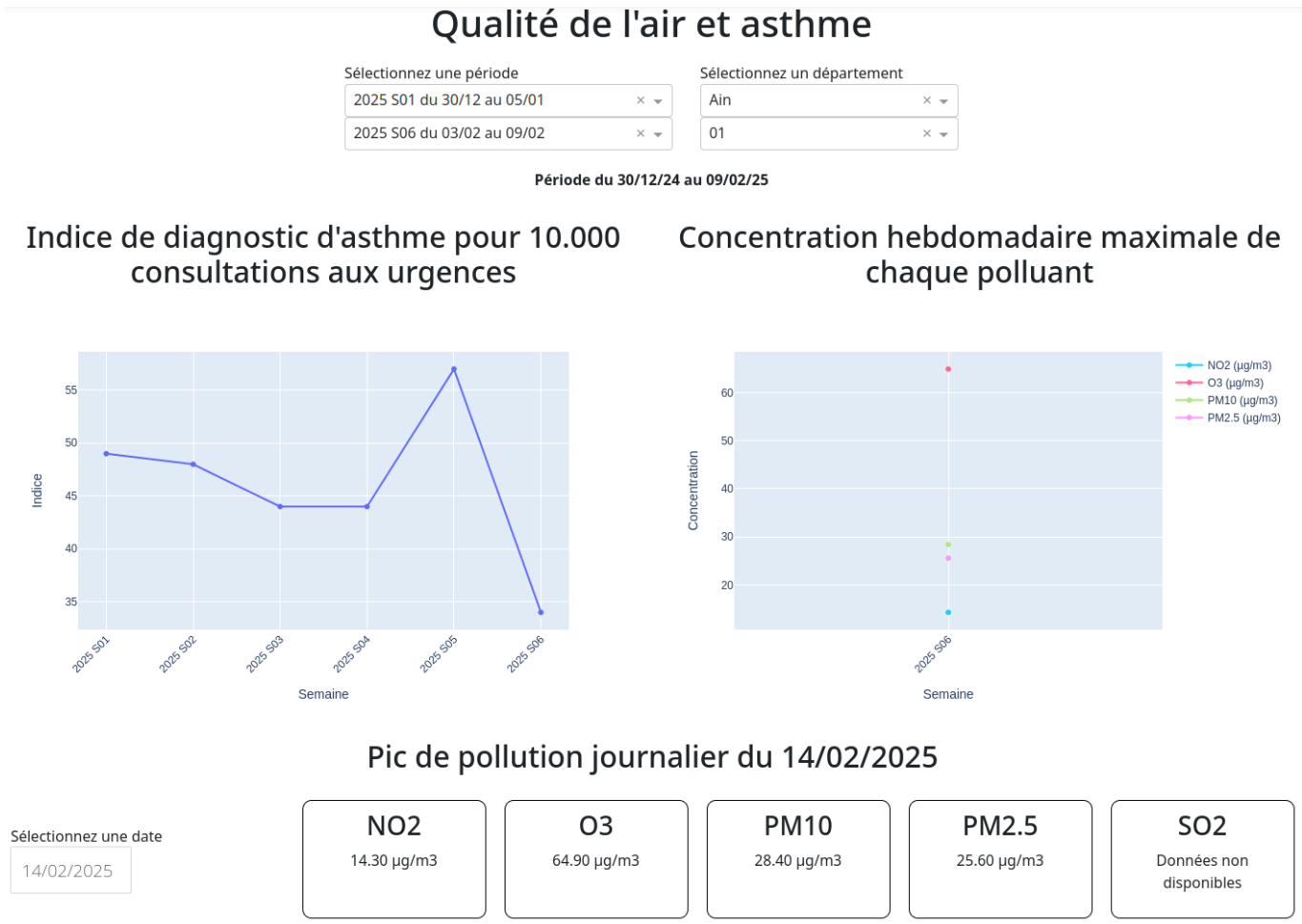
Le Dashboard offre une vue d'ensemble de l'indice de passages aux urgences pour asthme, à partir des données hebdomadaires de Géodes, couvrant la période de 2010 à la semaine la plus récente. Une carte interactive permet de visualiser cet indice par département, en filtrant par semaine avec des mises à jour hebdomadaires. Il calcule également un indice moyen pour toute la métropole française (faisant l'hypothèse de faciliter d'interprétation que le nombre de passages aux urgences par département et par période est équivalente) et propose un classement des départements selon le taux de passages aux urgences pour asthme, ajusté sur 10 000 passages d'urgence. Cela permet de repérer les zones les plus et les moins touchées par les exacerbations d'asthme dans le contexte des urgences. Les valeurs présentées sont des indices de diagnostic d'asthme pour 10.000 passages aux urgences.

Pollen





La section Pollen du dashboard offre une carte interactive des risques liés aux pollens en filtrant par jour. De plus, un barplot affiche les pollens les plus répandus par ville par jour. De cette manière, si un utilisateur constate sur la carte de l'asthme un taux élevé de passages aux urgences pour asthme dans une zone, il peut alors consulter la section Pollen pour vérifier si les villes de cette zone présentent également un risque accru lié aux pollens et identifier le type de pollen dominant. Démontrer une corrélation statistique précise entre les chiffres disponibles pour l'asthme et le pollen serait délicat pour plus raisons. (cf. Limites et perspectives)



La section Polluants du Dashboard fournit une analyse des concentrations de divers polluants atmosphériques et leur relation avec les exacerbations d'asthme observées aux urgences. Un graphique d'évolution hebdomadaire de l'indice de diagnostic d'asthme pour 10 000 passages aux urgences, permettant de suivre l'évolution des cas en fonction du temps. Un graphique des concentrations hebdomadaires maximales des principaux polluants (SO_2, NO_2, O_3, PM10, PM2.5), facilitant l'identification des tendances et des pics de pollution. En complément, le Dashboard affiche un pic de pollution journalière, qui renseigne sur les niveaux de pollution mesurés à une date donnée. Les utilisateurs peuvent sélectionner une journée spécifique pour consulter les valeurs relevées de NO_2, O_3, PM10, PM2.5, ainsi que les données de SO_2 lorsqu'elles sont disponibles.

L'utilisateur sélectionne un département, soit à l'aide du menu déroulant par le nom du département qui est listé, soit à l'aide du code de département (les deux div sont liées et interagissent). La sélection d'un département et d'une période active la représentation graphique, avec à gauche les données d'indice asthme/urgence par département pour une période donnée et à droite les pics de concentrations de polluant (max de chaque polluant pour toutes les stations de prélèvements de la zone départementale confondue pour une période donnée ; le choix de la ville et d'une station spécifique est en cours de développement) et d'indice de qualité de l'air (en cours de développement). Une période pré-sélectionnée de la première semaine de l'année en cours à la semaine la plus proche dont nous disposons des données pour Géodes facilite l'utilisation. L'utilisateur peut modifier cette plage temporelle à sa guise et remonter à la première semaine des données de Géodes. La sélection d'une date dans une div en bas à gauche permet d'activer la visualisation des indicateurs de concentration de polluant à date et de l'IQA (en cours de développement).

3.3. Déploiement et automatisation

Pour garantir la robustesse et l'efficacité de notre solution, nous avons mis en place plusieurs étapes d'automatisation et de déploiement, tant pour la collecte de données que pour la mise en production de notre dashboard.

Sauvegarde des Données sur AWS S3

- Mise en place du bucket S3 : Nous avons créé un bucket S3 dédié afin de sauvegarder nos fichiers de données au format CSV et XLSX. Ce bucket permet une gestion centralisée et sécurisée des données.
- Mise à jour des scripts : Les scripts de scrapping ont été modifiés pour écrire directement dans les fichiers stockés sur le bucket S3, en générant notamment les fichiers pollen.csv et geodes_complet.xlsx. Cela garantit une sauvegarde continue et facilite l'accès aux données depuis différentes instances.

Automatisation des Scripts de Scrapping via une Instance EC2

- Connexion et installation : Une instance EC2 d'AWS a été configurée pour héberger et automatiser nos scripts de scrapping. Après connexion à l'instance via SSH depuis le terminal, nous avons installé les dépendances nécessaires au bon fonctionnement des scripts.
- Déploiement des scripts : Les scripts ont été transférés sur l'instance EC2, puis nous avons automatisé leur exécution en configurant un cron job (via crontab -e). Ce mécanisme permet de lancer périodiquement le processus de scrapping, garantissant ainsi l'actualisation régulière des données.

Déploiement de l'Application Dashboard

- Adaptation du code : Pour la mise en production du dashboard, les fonctions de lecture des données locales ont été remplacées par des fonctions capables de récupérer les fichiers depuis le bucket S3. Cette modification assure une cohérence entre les environnements de développement et de production.

Création de l'image Docker

Un Dockerfile a été rédigé afin de construire une image Docker de l'application. La commande suivante a permis de créer l'image : docker build -t mon-dashboard .
Pour tester en local, nous avons lancé l'image avec : docker run -p 8050:8050 mon-dashboard

- Le port 8050 est exposé conformément à la configuration de l'application.
- Déploiement sur AWS App Runner :
  Pour déployer l'application sur AWS, nous avons utilisé App Runner en suivant ces étapes :
- Accès à AWS App Runner : Dans la console AWS (région eu-west-1), nous avons recherché App Runner et cliqué sur "Create service".
- Sélection de la source du conteneur :
  1. Choix de Container registry comme source.
  2. Sélection d'Amazon ECR et choix du dépôt mon-dashboard avec le tag latest.
- Configuration du service :
  1. Définition du port de l'application sur 8050, correspondant à l'exposition du port dans le Dockerfile.
  2. Ajout des variables d'environnement nécessaires à la configuration de l'application (par exemple, les paramètres AWS).
  3. Configuration des permissions en attachant un rôle IAM doté de la politique AmazonS3ReadOnlyAccess, permettant à l'application d'accéder au bucket S3, même si celui-ci se trouve dans une autre région (eu-north-1).
- Lancement du déploiement :
  Une fois ces configurations validées, le déploiement a été lancé via App Runner.

Intégration CI/CD avec GitLab

Configuration du pipeline GitLab :
Pour automatiser le processus de déploiement des nouvelles versions de notre application, nous avons mis en place un pipeline CI/CD via GitLab. Le fichier .gitlab-ci.yml a été configuré pour :

- Déclencher automatiquement la construction de l'image Docker.
- Lancer des tests et valider la nouvelle version.
- Déployer l'image mise à jour sur AWS via App Runner.

Gestion des variables d'environnement :
Les variables d'environnement nécessaires au bon fonctionnement de l'application (comme les clés d'accès et les paramètres de configuration) ont été définies directement dans l'interface GitLab, garantissant ainsi leur sécurité et leur mise à jour centralisée.

4 Limites et perspectives

Limites

L'absence de données brutes détaillées a limité notre capacité à calculer des indicateurs statistiques précis, comme la moyenne ou le total des passages aux urgences pour asthme. Nous ne disposons que du taux de passage aux urgences pour asthme sur 10 000 passages. Bien que cela restreigne notre analyse, cette donnée permet d'obtenir une vue d'ensemble utile pour observer les tendances hebdomadaires par département.

L'objectif initial était de démontrer une corrélation statistique précise entre les taux de passage aux urgences pour asthme et les concentrations de pollens, plusieurs facteurs rendent cette analyse complexe. Parmi ceux-ci, il y a la difficulté de paralléliser les données au niveau temporel (par jour pour le pollen et par semaine pour l'asthme), la prise en compte des effets du vent qui peuvent moduler la propagation des pollens, ainsi que la différence de niveau géographique des données : les informations sur les pollens sont disponibles à l'échelle communale, tandis que les données sur les passages aux urgences pour asthme le sont à l'échelle départementale.

Les mêmes obstacles surviennent pour l'analyse des polluants. Cependant, dans le cadre de ce projet, nous avons choisi de prendre en compte le pic de pollution pour chaque jour, puis de calculer le pic des pics pour la semaine, afin d'obtenir une vue d'ensemble des périodes de forte pollution. Ce traitement nous a permis de tenter de corréler ces pics de pollution avec les taux de passage aux urgences pour asthme, tout en restant conscient des limites de cette analyse.

La visualisation des polluants (page Polluant) est à ce jour limitée par l'historique issu de Géodair, les multiples utilisations et essaie de modification ont conduit à une perte de donnée, de plus nous sommes limité à un usage de 15 requête par heure sur l'API. Nous comptons rattraper l'historique et le mettre en ligne au fur et à mesure. La représentation des pics de polluants par département/période ainsi que les pics de concentration de polluants à date sont fortement impactés (veuillez choisir une date entre le 14 février 2025 et la veille de votre consultation du tableau de bord).

Perspectives

Nous envisageons de développer un modèle prédictif visant à estimer le taux de passages aux urgences pour asthme en fonction des concentrations de polluants. Ce modèle permettrait de mieux anticiper les périodes à risque en lien avec les variations de la pollution atmosphérique et d'offrir un outil supplémentaire pour la gestion des risques sanitaires.

Nous envisageons de mettre à jour la page de visualisation des polluants en présentant les données de qualité de l'air, indicateur basé sur les concentrations de polluants atmosphériques.

A travers la modélisation et le traitement des données de pollution, nous voulons aider à sensibiliser le publique au risques environnementaux comme la pollution, donner des outils d'aide à la décision (déménagement dans une zone à risque si allergie, variation des concentrations des polluants dans le temps et sur le territoire…) pour les personnes souffrantes d'asthme, leur entourage, les autorités publiques et les professionnels qui travaillent sur cette thématique.

5 Bibliographie

1. Inserm. Asthme : Inserm, La science pour la santé. Disponible sur: https://www.inserm.fr/dossier/asthme/

2. Ito K, Weinberger KR, Robinson GS, Sheffield PE, Lall R, Mathes R, et al. The associations between daily spring pollen counts, over-the-counter allergy medication sales, and asthma syndrome emergency department visits in New York City, 2002-2012. Environ Health Glob Access Sci Source. 27 août 2015;14:71.

3. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. The Lancet. 3 mai 2014;383(9928):1581‑92.

4. SEI. Air pollution leads to millions of emergency room visits for asthma attacks worldwide. Disponible sur: https://www.sei.org/about-sei/press-room/air-pollution-leads-millions-emergency-room-visits-asthma-attacks-worldwide/