PROJET FIL ROUGE

Antibiotiques et Antibiorésistance

Rapport de data visualisation et prétraitement des données

Wilfried YA

Adresse e-mail: wilfriedya@hotmail.fr

TABLE DES MATIERES

1. INTRODUCTION

1.1 Rappel du sujet

Dans le cadre du projet Data Analyst, notre groupe a choisi d'étudier la consommation d'antibiotiques en France à partir des données publiques OpenMedic mises à disposition par la Caisse Nationale d'Assurance Maladie (CNAM).

Ces données permettent d'analyser les remboursements de médicaments délivrés en pharmacie de ville, selon différents critères : région, sexe, âge, classe thérapeutique, prescripteur, et année de délivrance.

Les antibiotiques, identifiés par le code ATC commençant par "J01", constituent une catégorie médicamenteuse stratégique pour les politiques de santé publique. Leur suivi est crucial, car leur usage excessif ou inapproprié favorise l'émergence de résistances bactériennes, un enjeu mondial majeur reconnu par l'OMS et Santé publique France.

1.2 Position au sein du groupe

Mon rôle dans ce projet s'est construit à la suite de plusieurs réunions avec mon groupe. Ensemble, nous avons défini différents axes d'étude complémentaires autour de la thématique des médicaments et de la santé publique.

À l'issue de ces échanges, j'ai choisi de concentrer mon travail sur un axe analytique spécifique : l'évolution temporelle et régionale de la consommation d'antibiotiques entre 2019 et 2024.

Ce choix s'inscrit dans une approche cohérente avec le projet collectif, tout en me permettant de mobiliser mes compétences personnelles en exploration, nettoyage et structuration de données afin d'obtenir un jeu de données propre, homogène et exploitable pour des analyses approfondies.

1.3 Contexte général

L'antibiorésistance représente aujourd'hui l'un des défis sanitaires les plus pressants à l'échelle mondiale. L'usage inapproprié des antibiotiques, prescriptions inutiles, durées trop courtes, automédication, contribue à l'émergence de bactéries résistantes, réduisant l'efficacité de nombreux traitements.

En France, malgré les campagnes de sensibilisation, la consommation d'antibiotiques demeure élevée par rapport à la moyenne européenne.

Dans ce contexte, analyser les données OpenMedic permet de mieux comprendre les comportements de prescription et de consommation au fil du temps, de mesurer les efforts réalisés et d'identifier les zones ou catégories de population nécessitant une attention particulière.

1.4 Problématique

Comment la consommation d'antibiotiques a-t-elle évolué en France entre 2019 et 2024 ?
Quelles sont les tendances temporelles, régionales et pharmacologiques observées ?
Et dans quelle mesure ces évolutions traduisent-elles les efforts de maîtrise de la prescription et de lutte contre l'antibiorésistance ?
Ces questions guident l'ensemble de mon travail, depuis la préparation des données brutes jusqu'à la mise en place des indicateurs nécessaires à la visualisation et à l'interprétation.

1.5 Principaux objectifs

Les objectifs de mon étude sont à la fois techniques et analytiques :

- Nettoyer et harmoniser les fichiers OpenMedic annuels (2019–2024) afin d'obtenir une base de données consolidée et cohérente.
- Filtrer et isoler les lignes relatives aux antibiotiques (ATC J01).
- Analyser les tendances d'évolution de la consommation à travers le temps et le territoire.
- Identifier les classes thérapeutiques les plus prescrites et mesurer leur évolution.
- Fournir une base de données propre pour les visualisations et futures modélisations (corrélations, séries temporelles, etc.).

Enfin, ce travail s'inscrit dans un projet complet en deux volets : une première phase consacrée à l'exploration et au nettoyage des données, suivie d'une deuxième étape dédiée à la visualisation et à l'interprétation des résultats (réalisée dans le notebook *02_visualisation.ipynb*). Cette continuité permet de passer d'une base brute à une lecture analytique claire et exploitable.

1.6 Exemples de KPI / Indicateurs étudiés

Pour répondre à la problématique posée, plusieurs indicateurs clés (KPI) ont été définis afin d'examiner les différentes dimensions de la consommation d'antibiotiques en France sur la période 2019 - 2024.
Ces indicateurs couvrent à la fois les aspects temporels, régionaux, sociodémographiques et médicaux de l'étude.

L'analyse débute par une vision d'ensemble de l'évolution annuelle du montant total remboursé pour les antibiotiques, permettant d'observer les grandes tendances au fil du temps.
Cette première approche est ensuite affinée par la répartition du montant remboursé selon le sexe, afin d'identifier d'éventuelles différences de consommation entre les hommes et les femmes sur la même période.

Une dimension territoriale est ensuite explorée à travers la répartition des montants remboursés par région de résidence, mettant en évidence les disparités géographiques et les spécificités locales. Cette approche régionale est complétée par une analyse croisée entre région et tranche d'âge, afin de relier les comportements de consommation aux profils démographiques.

L'étude inclut également une exploration détaillée des tranches d'âge et du sexe, qui permet d'identifier les groupes de population les plus consommateurs d'antibiotiques, ainsi que les éventuelles différences d'exposition entre hommes et femmes.

Sur le plan pharmaceutique, deux analyses de classement complètent cette approche :

- le Top 10 des sous-groupes d'antibiotiques les plus remboursés, qui met en lumière les classes thérapeutiques dominantes au sein du code ATC J01 ;
- le Top 10 des spécialités pharmaceutiques les plus remboursées, qui offre une vision précise des produits contribuant le plus à la dépense nationale.

Enfin, une dernière série d'indicateurs porte sur le rôle des prescripteurs dans la dynamique globale :

- le Top 10 des montants remboursés par type de prescripteur, identifiant les professions médicales les plus contributrices ;
- la part en pourcentage des principaux prescripteurs dans le montant total remboursé, permettant d'évaluer la concentration ou la diversité des origines de prescription.

Ces différents KPI offrent une lecture multidimensionnelle de la consommation d'antibiotiques : ils permettent d'analyser à la fois l'évolution dans le temps, les disparités territoriales et démographiques, ainsi que les tendances thérapeutiques et prescriptrices.

Ils constituent ainsi la base d'un tableau de bord analytique complet, essentiel pour comprendre les dynamiques de consommation et alimenter la réflexion sur la maîtrise de l'usage des antibiotiques en France.

1.7 Contexte métier et enjeux

Ce projet s'inscrit dans une démarche de valorisation des données de santé.
L'exploitation des bases publiques comme OpenMedic permet de fournir aux acteurs institutionnels (CNAM, Santé Publique France, Ministère de la Santé) des éléments chiffrés fiables pour orienter les campagnes de prévention et les stratégies de santé publique.

Du point de vue d'un Data Analyst, cette étude illustre la manière dont les données peuvent être nettoyées, structurées et transformées en connaissances exploitables.
Elle met en lumière la chaîne complète d'un projet data, depuis la collecte et la préparation jusqu'à la production de visualisations interprétables.
La phase de visualisation occupe une place centrale : elle permet de rendre les données accessibles, lisibles et exploitables par les décideurs, favorisant une meilleure compréhension des dynamiques de consommation d'antibiotiques en France.

2. COMPREHENSION ET ANALYSE DU JEU DE DONNEES

2.1 Cadre et périmètre de l'étude

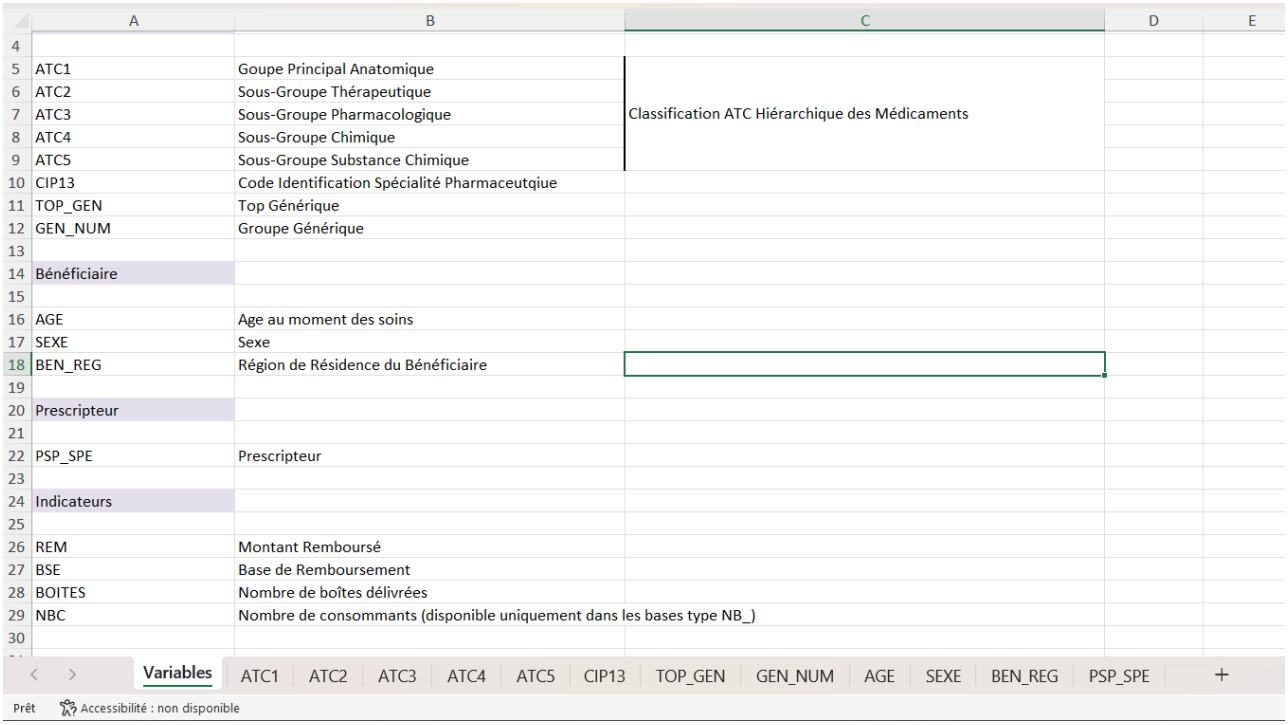
Le jeu de données utilisé dans le cadre de ce projet provient de la base OpenMedic publiée par la Caisse Nationale d'Assurance Maladie (CNAM). Cette base recense l'ensemble des dépenses de médicaments délivrés en pharmacie de ville, tous régimes confondus (régime général, agricole, indépendants, etc.).

Les données sont disponibles publiquement sur data.gouv.fr et mises à jour chaque année. Mon travail se concentre sur la période 2019 à 2024, afin d'observer les tendances récentes, y compris les années marquées par la pandémie de COVID-19. Cette période est particulièrement intéressante, car elle inclut à la fois une phase de forte activité médicale (pré-2020), un changement brutal de comportement durant la crise sanitaire, puis un retour progressif à la normale.

2.2 Présentation du jeu de données OpenMedic

Chaque fichier OpenMedic annuel contient plusieurs dizaines de colonnes décrivant les caractéristiques des médicaments, les profils de patients, les prescripteurs, et les indicateurs de consommation.

Les principales variables utilisées dans ce projet sont les suivantes :



	A	B	C	D	E
4					
5	ATC1	Goupe Principal Anatomique			
6	ATC2	Sous-Groupe Thérapeutique			
7	ATC3	Sous-Groupe Pharmacologique	Classification ATC Hiérarchique des Médicaments		
8	ATC4	Sous-Groupe Chimique			
9	ATC5	Sous-Groupe Substance Chimique			
10	CIP13	Code Identification Spécialité Pharmaceutqiue			
11	TOP_GEN	Top Générique			
12	GEN_NUM	Groupe Générique			
13					
14	Bénéficiaire				
15					
16	AGE	Age au moment des soins			
17	SEXE	Sexe			
18	BEN_REG	Région de Résidence du Bénéficiaire			
19					
20	Prescripteur				
21					
22	PSP_SPE	Prescripteur			
23					
24	Indicateurs				
25					
26	REM	Montant Remboursé			
27	BSE	Base de Remboursement			
28	BOITES	Nombre de boîtes délivrées			
29	NBC	Nombre de consommants (disponible uniquement dans les bases type NB_)			
30					

Figure 1 : Descriptif des variables open medic

Ces variables permettent d'explorer les dimensions temporelle, géographique, démographique et thérapeutique de la consommation d'antibiotiques.

2.3 Choix du périmètre analytique

L'ensemble des fichiers OpenMedic contient des millions d'enregistrements par an. Pour cibler la problématique du projet, seuls les médicaments dont le code ATC commence par "J01" ont été conservés, correspondant à la catégorie des antibiotiques à usage systémique.
Ce filtrage a permis de concentrer l'analyse sur les produits pertinents, d'alléger le volume de données, et de faciliter la mise en place de visualisations exploitables.
Ce choix s'appuie sur la classification officielle de l'Organisation mondiale de la santé (OMS), qui regroupe sous le code J01 les principales familles d'antibiotiques.

2.4 Structure et volume des données

Après concaténation des cinq fichiers annuels (2019 à 2024), le jeu de données consolidé atteint une volumétrie de plusieurs millions de lignes, représentant des centaines de milliers de références médicamenteuses et de situations de remboursement.
Les données sont de type mixte :

- catégorielles (sexe, région, tranche d'âge, code ATC, type de prescripteur),
- numériques (montants, volumes délivrés),
- et alphanumériques (libellés, codes CIP ou ATC).

La structure tabulaire permet une exploitation flexible sous Python/Pandas, en facilitant les agrégations et les regroupements par année, par région ou par sous-groupe thérapeutique.

2.5 Fiabilité et pertinence des données

Les données OpenMedic sont issues directement du Système National des Données de Santé (SNDS), ce qui leur confère un haut niveau de fiabilité.
Cependant, plusieurs particularités doivent être prises en compte :

- Les données couvrent uniquement les ventes en pharmacie de ville (hors hôpital).
- Les régimes spéciaux peuvent être sous-représentés ou agrégés.
- Certaines variables changent légèrement de nom d'une année à l'autre (ex. SEXE / sexe).
- Les montants négatifs ou manquants peuvent apparaître lors de régularisations de remboursement.

Ces éléments justifient les étapes de nettoyage, normalisation et harmonisation décrites dans la partie suivante.

2.6 Limites et précautions d'interprétation

Plusieurs limites doivent être mentionnées pour encadrer l'analyse :

- Les données OpenMedic ne permettent pas d'identifier les pathologies associées ni les causes médicales de prescription.
- Les volumes et montants ne tiennent pas compte des automédications ni des ventes non remboursées.
- Les années 2020–2021 sont atypiques à cause de la pandémie de COVID-19, qui a entraîné une forte baisse de certaines prescriptions infectieuses.
- Les comparaisons régionales nécessitent une normalisation par la population (via les données INSEE) pour éviter les biais liés à la taille des territoires.

Malgré ces limites, le jeu de données constitue une base solide pour une analyse descriptive et comparative de la consommation d'antibiotiques sur le territoire national.

3. PREPARATION ET NETTOYAGE DES DONNEES

3.1 Cadre et périmètre de l'étude

L'étape de préparation débute par l'importation des bibliothèques Python nécessaires au traitement des données, pandas, numpy, et matplotlib.pyplot constituent le socle de l'analyse, permettant respectivement la manipulation tabulaire, la gestion des valeurs numériques et les visualisations basiques.

Une attention particulière a été portée à la configuration des paramètres d'importation (pd.options.display, encodage latin1, séparateurs décimaux et de milliers) afin d'assurer la compatibilité avec les fichiers OpenMedic, qui utilisent des conventions de formatage françaises.

```
import pandas as pd
from IPython.display import display
import numpy as np
import matplotlib.pyplot as plt
```

Figure 2 : Importation des bibliothèques principales (pandas, numpy, matplotlib) et paramétrage initial du notebook.

3.2 Définition du typage des variables (dtypes)

Les fichiers OpenMedic contiennent de nombreuses colonnes de nature différente : certaines numériques, d'autres textuelles ou codées.

Pour éviter les erreurs de lecture (comme les zéros supprimés dans les codes ATC ou CIP), un dictionnaire de typage (d_types) a été défini avant le chargement.

Ce typage garantit une interprétation correcte des données et facilite les opérations de regroupement et de filtrage.

```
# conversion des types de données pour certaines colonnes
d_types={
            "ATC3" : "str",
            "L_ATC3" : "str",
            "ATC4": "str",
            "L_ATC4" : "str",
            "ATC5" : "str",
            "L_ATC5" : "str",
            "CIP13" : "str",
            "L_CIP13" : "str",
            "age":"str",
            "sexe":"str",
            "SEXE":"str",
            "BEN_REG":"str",
            "PSP_SPE":"str",
            "REM":"float",
            "BSE":"float",
            "BOITES":"int"}
```

Figure 3 : Exemple du dictionnaire de typage utilisé pour forcer le format des colonnes lors de l'import.

3.3 Lecture des fichiers annuels (2019 - 2024)

Chaque fichier OpenMedic annuel (2019, 2020, 2021, 2022, 2023 et 2024) a été importé séparément à l'aide de la fonction pd.read_csv().

Les paramètres utilisés (encoding='latin1', decimal=',', thousands='.') assurent une lecture correcte des montants remboursés, souvent mal interprétés sans configuration adaptée.

```
# Chargement des données OPEN MEDIC de 2019 à 2024

# 2024
df_24 = pd.read_csv("C:\\Users\\wilfr\\002-Projet_fil_rouge\\01_data_brute\\OPEN_MEDIC_2024.zip", compression="zip",
# Ajout d'une colonne pour l'année d'exercice 2024
df_24["Annee"] = str(2024)
# 2023
df_23 = pd.read_csv("C:\\Users\\wilfr\\002-Projet_fil_rouge\\01_data_brute\\OPEN_MEDIC_2023.zip", compression="zip",
# Ajout d'une colonne pour l'année d'exercice 2023
df_23["Annee"] = str(2023)
# 2022
df_22 = pd.read_csv("C:\\Users\\wilfr\\002-Projet_fil_rouge\\01_data_brute\\OPEN_MEDIC_2022.zip", compression="zip",
# Ajout d'une colonne pour l'année d'exercice 2022
df_22["Annee"] = str(2022)
# 2021
df_21 = pd.read_csv("C:\\Users\\wilfr\\002-Projet_fil_rouge\\01_data_brute\\OPEN_MEDIC_2021.zip",compression="zip",
# Ajout d'une colonne pour l'année d'exercice 2021
df_21["Annee"] = (2021)
# 2020
df_20 = pd.read_csv("C:\\Users\\wilfr\\002-Projet_fil_rouge\\01_data_brute\\OPEN_MEDIC_2020.zip", compression="zip",
# Ajout d'une colonne pour l'année d'exercice 2020
df_20["Annee"] = (2020)
# 2019
df_19 = pd.read_csv("C:\\Users\\wilfr\\002-Projet_fil_rouge\\01_data_brute\\OPEN_MEDIC_2019.zip", compression="zip",
# Ajout d'une colonne pour l'année d'exercice 2019
df_19["Annee"] = (2019)
```

Figure 4 : Lecture d'un fichier OpenMedic annuel avec gestion des séparateurs et encodage français.

```
, sep=";", encoding="latin1", dtype=d_types, thousands='.', decimal=',' )

, sep=";", encoding="latin1", dtype=d_types, thousands='.', decimal=',' )

, sep=";", encoding="latin1", dtype=d_types, thousands='.', decimal=',' )

sep=";", encoding="latin1", dtype=d_types, thousands='.', decimal=',' )

, sep=";", encoding="latin1", dtype=d_types, thousands='.', decimal=',' )

, sep=";", encoding="latin1", dtype=d_types, thousands='.', decimal=',' )
```

Python

Figure 5 : Suite lecture d'un fichier OpenMedic annuel avec gestion des séparateurs et encodage français.

3.4 Harmonisation des noms de colonnes

Lors de la concaténation, certaines incohérences de noms de variables ont été détectées (ex. SEXE au lieu de sexe dans le fichier 2019).

Ces différences ont été normalisées à l'aide de la méthode rename() afin de garantir l'uniformité du schéma de données.

Cette étape, bien que mineure, est essentielle pour éviter la création de doublons de colonnes lors de futures manipulations.

```
81]
# Renommer la colonne "SEXE" open_medic_2019 afin de l'harmoniser avec les autres années,
# eviter les problèmes lors de la concaténation des dataframes (NaN)

df_19 = df_19.rename(columns={'SEXE':'sexe'})
```

Figure 6 : Uniformisation de la nomenclature des colonnes (exemple : "SEXE" → "sexe").

3.5 Ajout de la colonne "Année" et concaténation

Une fois chaque fichier chargé, une colonne « Annee » a été ajoutée pour identifier l'année d'origine.

L'ensemble des DataFrames a ensuite été fusionné en un seul jeu de données grâce à pd.concat(), permettant une vue globale de la période 2019 - 2024.

```
82]
# concaténer tous les dataframes

df = pd.concat([df_19, df_20, df_21, df_22, df_23, df_24])
df.head()
```

Figure 7 : Fusion des fichiers annuels après ajout d'une colonne "Année" pour faciliter l'analyse temporelle.

3.6 Limites et précautions d'interprétation

Avant le nettoyage, une vérification globale du jeu de données a été effectuée grâce aux méthodes df.info(), df.describe() et df.head().

Ces fonctions permettent de connaître :

- le nombre total de lignes et colonnes,
- la présence de valeurs manquantes,
- et les types de données chargés.

```
<class 'pandas.core.frame.DataFrame'>
Index: 11192560 entries, 0 to 1916884
Data columns (total 22 columns):
 #   Column   Dtype
---  ------   -----
 0   ATC1     object
 1   l_ATC1   object
 2   ATC2     object
 3   L_ATC2   object
 4   ATC3     object
 5   L_ATC3   object
 6   ATC4     object
 7   L_ATC4   object
 8   ATC5     object
 9   L_ATC5   object
 10  CIP13    object
 11  l_cip13  object
 12  TOP_GEN  object
 13  GEN_NUM  int64
 14  age      object
 15  sexe     object
 16  BEN_REG  object
 17  PSP_SPE  object
 18  BOITES   int32
 19  REM      float64
 20  BSE      float64
 21  Annee    object
dtypes: float64(2), int32(1), int64(1), object(18)
memory usage: 1.9+ GB
```

Figure 8 : Aperçu général de la structure et du volume des données avant nettoyage.

3.7 Ciblage des antibiotiques (ATC J01)

Pour restreindre l'analyse à la thématique du projet, un filtrage spécifique a été appliqué : seules les lignes dont le code ATC commence par "J01" (antibiotiques à usage systémique) ont été conservées.

Ce filtrage, opéré à l'aide de df[df['ATC2'].str.startswith('J01')], permet de créer un sous-ensemble df_2 exclusivement dédié aux antibiotiques.

```
# Sélectionner uniquement les lignes du DataFrame df dont la colonne 'ATC2'
# commence par 'J01' (codes correspondant aux antibiotiques selon la classification ATC),
# puis créer une copie indépendante du sous-ensemble obtenu pour éviter les avertissements
# liés aux modifications sur une vue de DataFrame.

df_2 = df[df['ATC2'].str.startswith('J01')].copy()
df_2.head()
```

Figure 9 : Sélection des enregistrements correspondant aux antibiotiques (codes ATC commençant par "J01").

3.8 Réduction du schéma et normalisation

Certaines colonnes jugées peu utiles à l'analyse (identifiants techniques, codes internes, doublons) ont été supprimées.

En parallèle, plusieurs libellés ont été **normalisés** (via str.capitalize() ou str.lower()) afin d'améliorer la lisibilité et d'éviter les doublons dus à des variations d'écriture.

Deux colonnes explicatives ont également été ajoutées pour renforcer la compréhension métier :

- Region_residence (copie explicite du code régional),
- Prescripteur (libellé issu du code prescripteur).

```
# transformer la première lettre en majuscule et le reste en minuscule

# Pour la colonne des sous-groupes pharmacologiques
df_3["Libelle_sous-groupe_pharmacologique"] = (
    df_3["Libelle_sous-groupe_pharmacologique"]
    .str.capitalize()   # transforme la première lettre en majuscule et le reste en minuscule
)

# Pour la colonne des spécialités pharmaceutiques
df_3["Libelle_identification_pharmaceutique"] = (
    df_3["Libelle_identification_pharmaceutique"]
    .str.capitalize()
)
```

Figure 10 : Uniformisation des libellés et ajout de variables explicites pour la lisibilité métier.

```
# créer une copie de la colonne "region_residence" et l'insérer à l'index 8
df_3.insert(8, "Region_residence", df_3["Code_region_residence"])

# créer une copie de la colonne "prescripteur" et l'insérer à l'index 10
df_3.insert(10, "Prescripteur", df_3["Code_prescripteur"])

df_3.head()
```

Figure 11 : Suite uniformisation des libellés et ajout de variables explicites pour la lisibilité métier.

3.9 Contrôle qualité : valeurs manquantes, doublons et cohérence

Un contrôle de qualité a ensuite été mené pour s'assurer de la cohérence des données :

- Détection et comptage des valeurs manquantes (isnull().sum()),
- Vérification de la présence de doublons (duplicated().sum()),
- Contrôle de la cohérence des valeurs numériques (montants négatifs, anomalies).

Les lignes erronées ou incohérentes ont été supprimées pour garantir la fiabilité des indicateurs à venir.

```
[100]
# Vérifier les valeurs manquantes dans df_3 par colonne

df_3.isnull().sum()
```

```
Sous-groupe_pharmacologique              0
Libelle_sous-groupe_pharmacologique      0
Code_identification_pharmaceutique       0
Libelle_identification_pharmaceutique    0
Top_Generique                            0
Tranche_age_soins                        0
Sexe                                     0
Code_region_residence                    0
Region_residence                         0
Code_prescripteur                        0
Prescripteur                             0
nb_boites_delivrees                      0
Montant_rembourse                        0
Base_remboursement                       0
Annee                                    0
dtype: int64
```

Figure 12 : Évaluation de la complétude et de la cohérence des données avant export.

3.10 Ré indexation et export final

Une fois les données nettoyées et homogénéisées, l'index du DataFrame a été réinitialisé (reset_index(drop=True)), puis le jeu final a été exporté au format CSV sous le nom open_medic_cleaned.csv.

Ce fichier consolidé sert désormais de base d'analyse principale pour la phase suivante de visualisation.

```
# J'exporte le DataFrame final

df_3.to_csv(
    "open_medic_cleaned.csv",
    sep=";",
    index=False,
    encoding="utf-8-sig"
)

print(" Le fichier open_medic_cleaned.csv a été créé avec succès.")
```

```
Le fichier open_medic_cleaned.csv a été créé avec succès.
```

Figure 13 : Exportation du jeu de données final prêt pour la phase de visualisation.

3.11 Bilan de la phase de préparation

Cette phase de préparation constitue une étape cruciale du projet.

Elle a permis de transformer un ensemble de fichiers hétérogènes en un jeu de données unifié, propre et cohérent, prêt à être exploité pour des analyses visuelles et statistiques.

Le processus appliqué — typage, filtrage, harmonisation, contrôle qualité et export — illustre la rigueur nécessaire à tout pipeline de data cleaning dans un projet d'analyse.

Cette rigueur garantit la fiabilité des résultats et renforce la crédibilité des visualisations qui en découlent.

4. ANALYSE DESCRIPTIVE ET VISUALISATION DES DONNEES

4.1 Objectif de la phase de visualisation

Après la préparation et le nettoyage du jeu de données OpenMedic (2019 - 2024), cette phase a pour but de valoriser les informations contenues dans les données à travers une série de visualisations.

Les graphiques ont été réalisés à l'aide des bibliothèques Matplotlib et Seaborn, permettant de représenter les évolutions temporelles, les disparités régionales et les différences sociodémographiques dans la consommation d'antibiotiques.

Les analyses visuelles réalisées permettent :

- d'identifier les tendances globales de consommation d'antibiotiques,
- d'observer les variations selon le sexe, l'âge et la région,
- et de déterminer les classes thérapeutiques et prescripteurs les plus influents.

4.2 Évolution temporelle de la consommation d'antibiotiques

La première série de visualisations met en évidence l'évolution du montant total remboursé pour les antibiotiques entre 2019 et 2024.

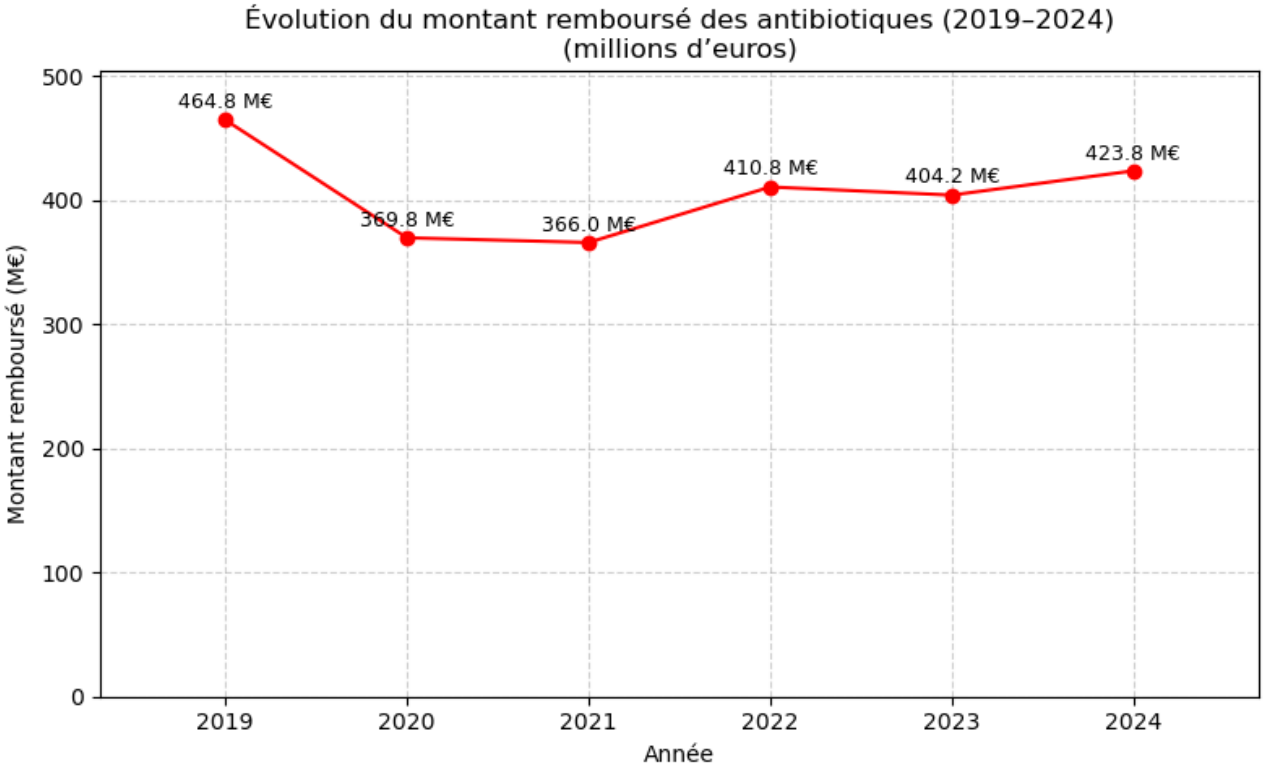


Figure 14 : Tendance globale de la consommation d'antibiotiques en France entre 2019 et 2024.

L'analyse montre une diminution marquée en 2020, correspondant à la période de confinement liée à la crise du COVID-19, suivie d'une reprise progressive à partir de 2021.

Cette évolution traduit l'impact direct des restrictions sanitaires sur la propagation des infections bactériennes et la baisse temporaire des prescriptions.

4.3 Différences selon le sexe

La consommation a ensuite été analysée par sexe afin de détecter d'éventuelles disparités dans les remboursements.

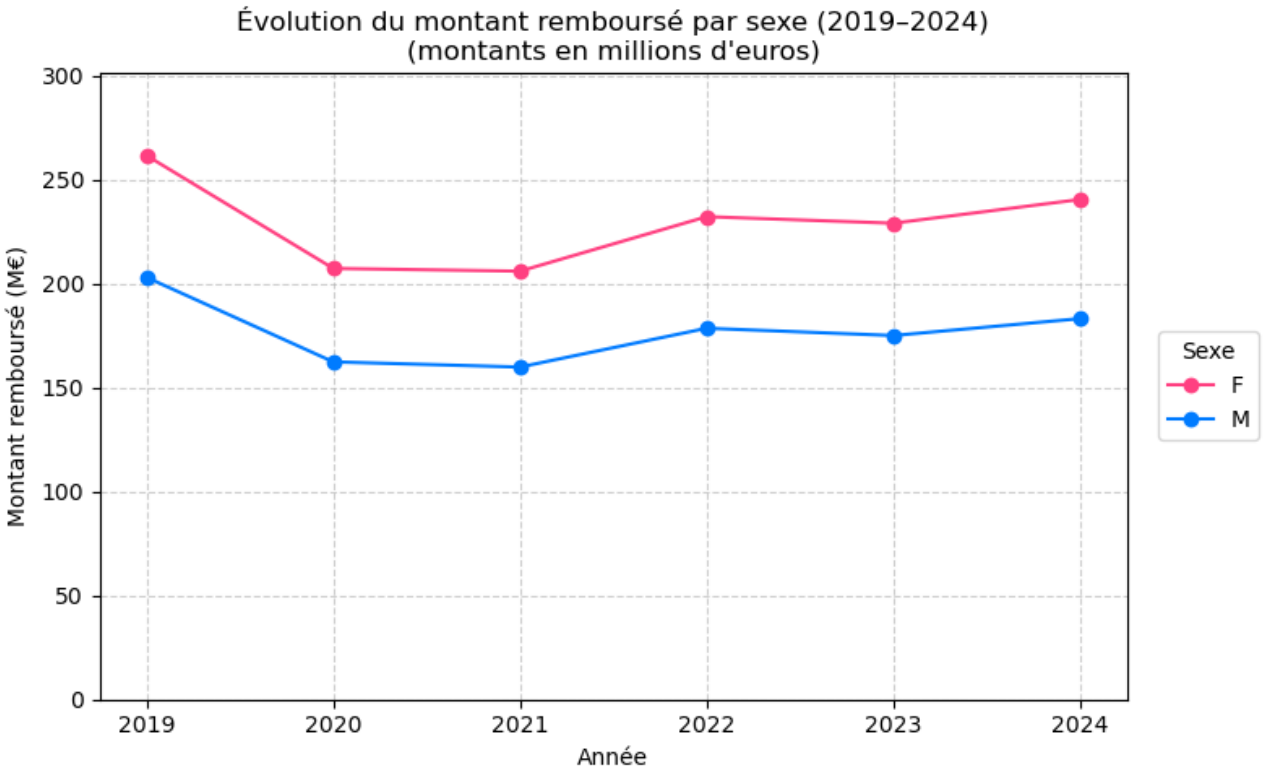


Figure 15 : Comparaison de l'évolution du montant remboursé entre les femmes et les hommes.

Les résultats révèlent que les femmes présentent systématiquement des montants remboursés supérieurs à ceux des hommes sur l'ensemble de la période.

Cette tendance peut s'expliquer par des facteurs démographiques (espérance de vie plus longue, suivi médical plus fréquent) ou épidémiologiques (plus grande prévalence de certaines infections bactériennes dans la population féminine).

4.4 Disparités régionales

Une répartition des montants remboursés a ensuite été réalisée par région de résidence.

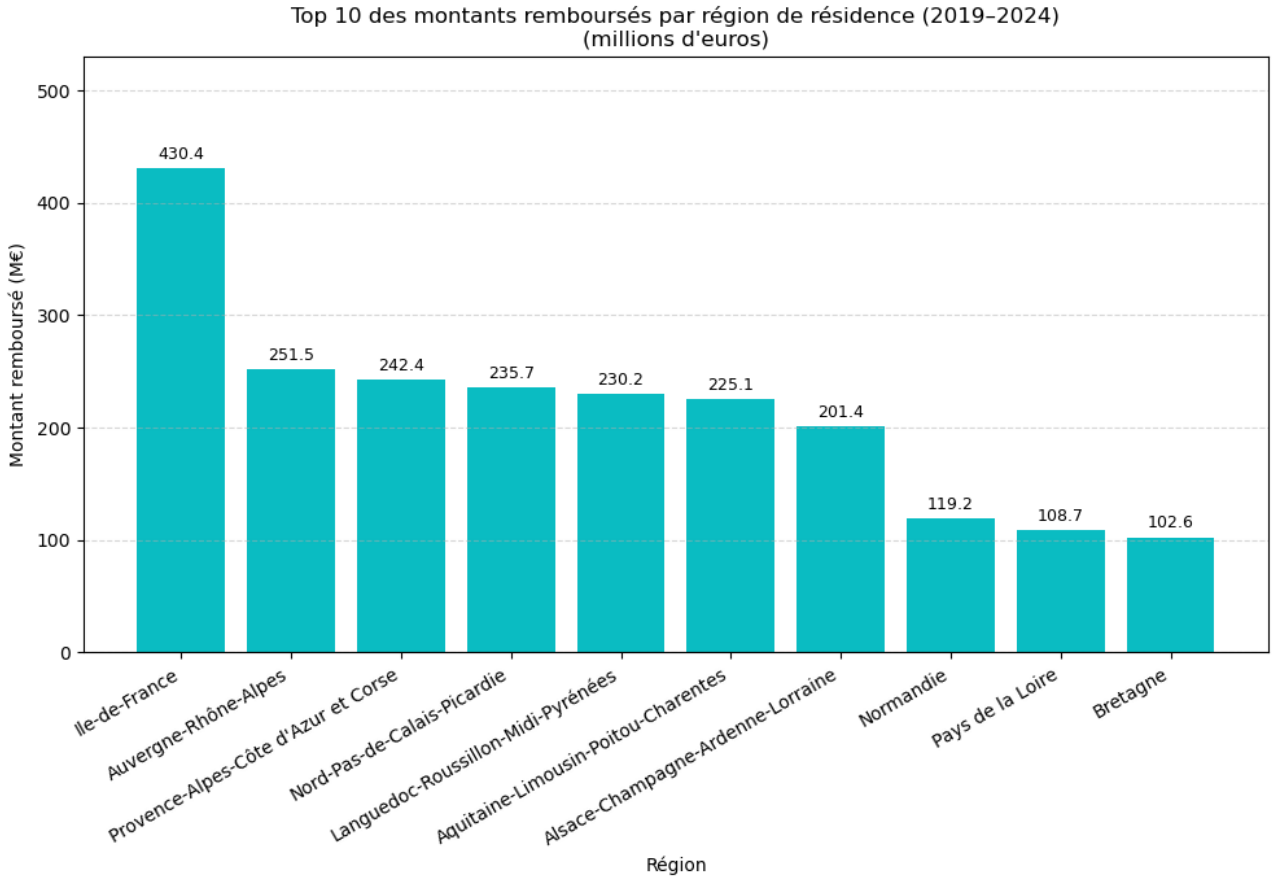


Figure 16 : Top 10 des montants remboursés pour les antibiotiques par région de résidence, exprimés en millions €.

L'analyse régionale met en évidence des écarts significatifs entre les territoires.

L'Île-de-France se distingue très nettement avec plus de 430 millions d'euros remboursés, soit près du double du montant observé dans la deuxième région du classement, Auvergne–Rhône-Alpes (251 M€).

Viennent ensuite Provence-Alpes-Côte d'Azur et Corse, Nord–Pas-de-Calais–Picardie, et Languedoc–Roussillon–Midi-Pyrénées, dont les montants oscillent autour de 230 à 240 millions d'euros.

Les régions de l'Ouest comme la Bretagne, les Pays de la Loire ou la Normandie affichent quant à elles des niveaux de remboursement plus faibles (autour de 100 à 120 millions d'euros).

Ces disparités peuvent s'expliquer par plusieurs facteurs :

- la densité de population (l'Île-de-France concentre près de 20 % des assurés français),
- la structure démographique (poids relatif des seniors ou des jeunes enfants, plus souvent consommateurs d'antibiotiques),
- et la densité médicale, plus élevée dans certaines régions métropolitaines.

Cette hiérarchie reflète donc davantage la taille et la structure démographique des régions qu'une surconsommation proprement dite.

Pour une interprétation plus fine, il serait pertinent de ramener ces montants à la population régionale (par habitant), ce qui permettrait d'identifier les véritables différences de comportements de prescription.

4.5 Répartition par sexe

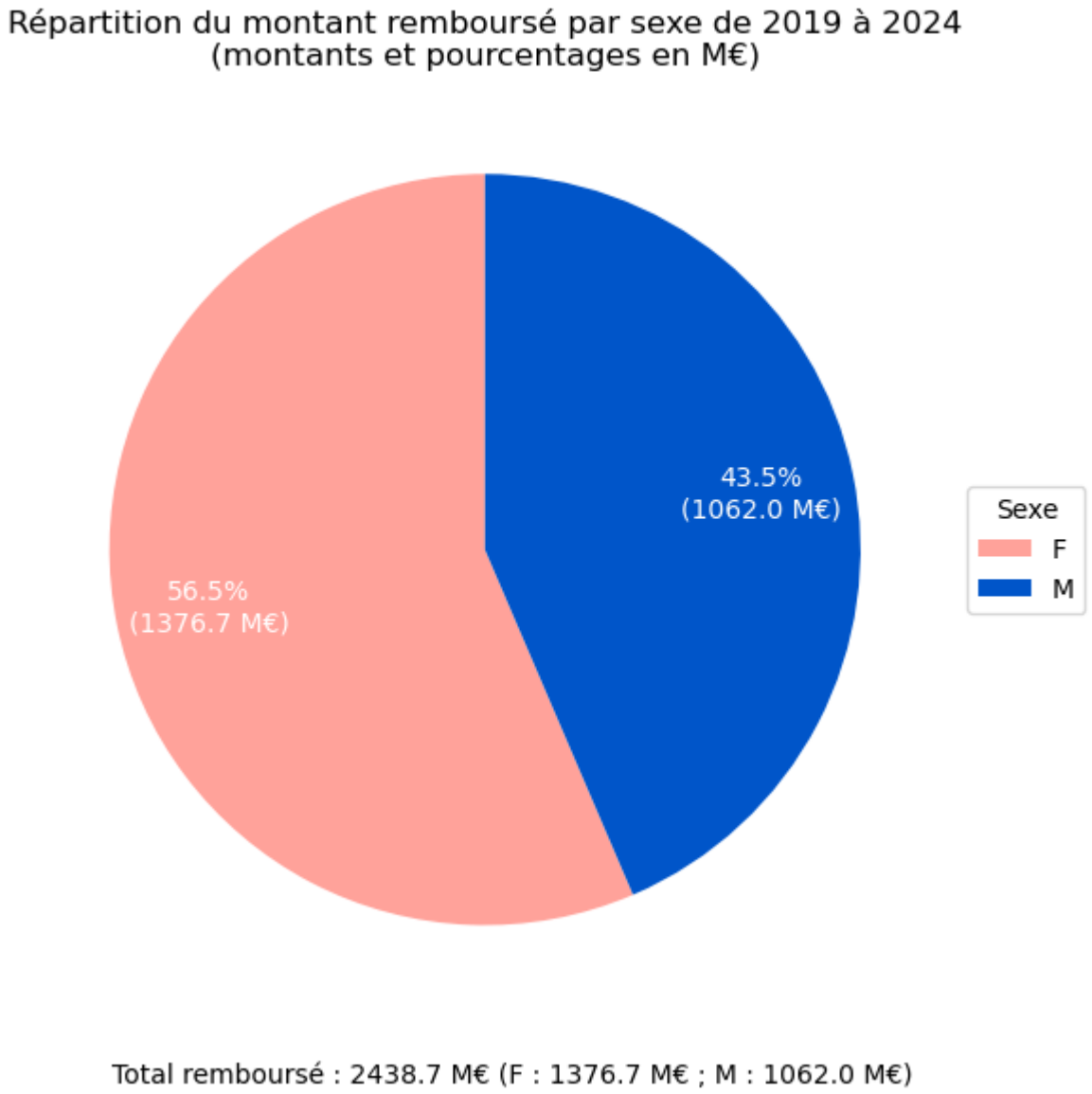


Figure 17 : Répartition du montant total remboursé pour les antibiotiques selon le sexe.

Sur la période 2019 - 2024, la consommation féminine d'antibiotiques représente 56,5 % du montant total remboursé, contre 43,5 % pour les hommes.

En valeur absolue, cela correspond à 1,38 milliard d'euros pour les femmes et 1,06 milliard pour les hommes.

Cette différence, relativement stable d'une année sur l'autre, s'explique notamment par une fréquence de consultation médicale plus élevée chez les femmes, une espérance de vie plus longue, ainsi qu'un suivi médical spécifique (gynécologique ou périnatal) impliquant des prescriptions plus régulières.

Ces résultats confirment la tendance déjà observée dans d'autres études pharmaco-épidémiologiques : les femmes présentent globalement un taux de prescription plus élevé pour plusieurs classes de médicaments, dont les antibiotiques.

4.6 Consommation par tranche d'âge et sexe

Une analyse croisée des tranches d'âge et du sexe a permis de mieux comprendre les profils des patients consommateurs d'antibiotiques.

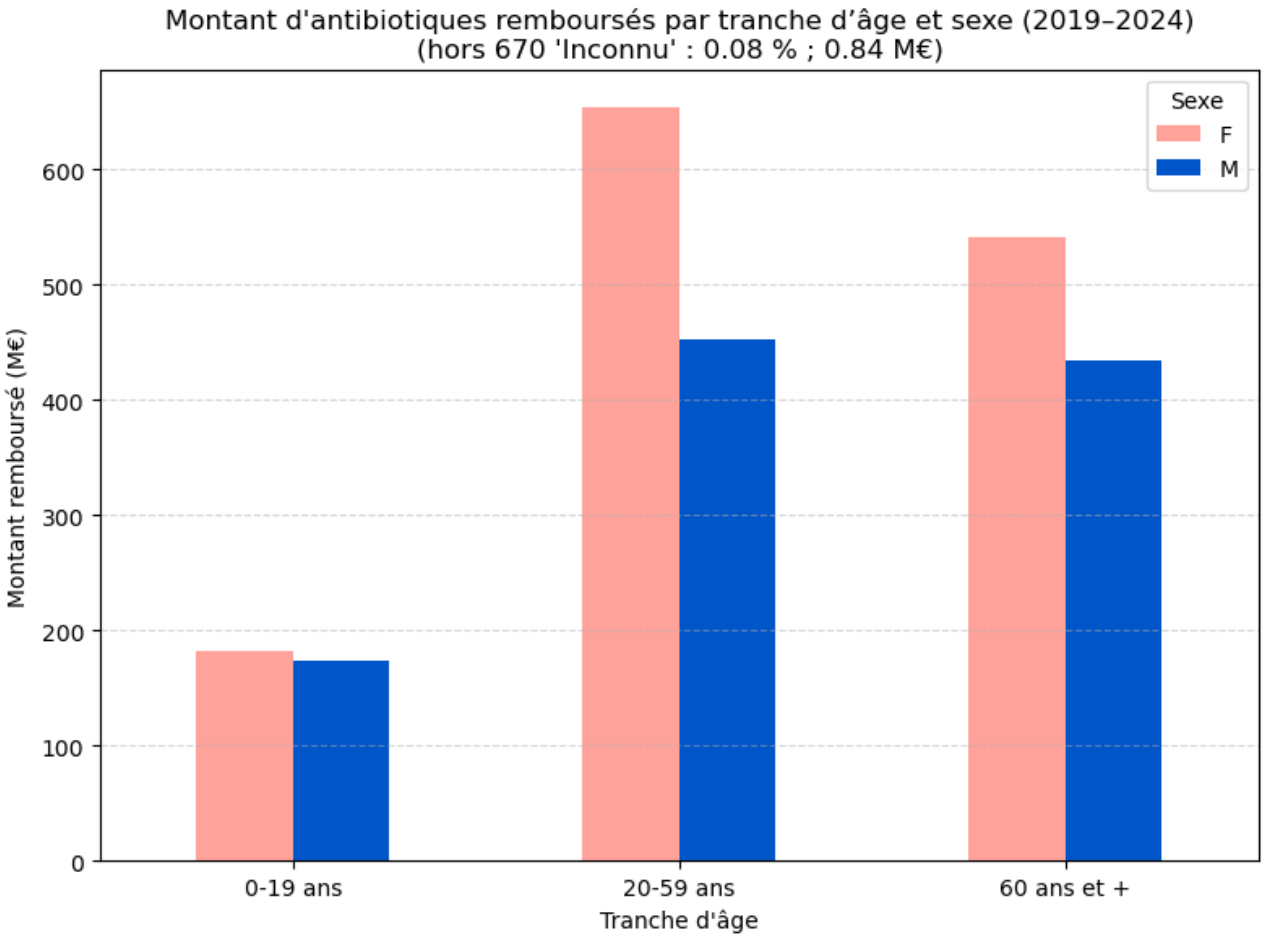


Figure 18 : Montant d'antibiotiques remboursés par tranche d'âge et sexe.

L'analyse par tranche d'âge montre que la consommation d'antibiotiques est nettement plus élevée chez les adultes âgés de 20 à 59 ans, représentant la part la plus importante des remboursements.

Les personnes de 60 ans et plus arrivent en deuxième position, avec un niveau de remboursement significatif, tandis que les jeunes de moins de 20 ans affichent une consommation plus modérée.

Dans toutes les tranches d'âge, les femmes présentent des montants supérieurs à ceux des hommes, ce qui confirme les écarts déjà observés dans la répartition globale.

Ces résultats suggèrent que la consommation d'antibiotiques suit à la fois une logique d'activité professionnelle (adultes actifs) et une logique de vulnérabilité physiologique (seniors).

4.7 Croisement région, tranche d'âge

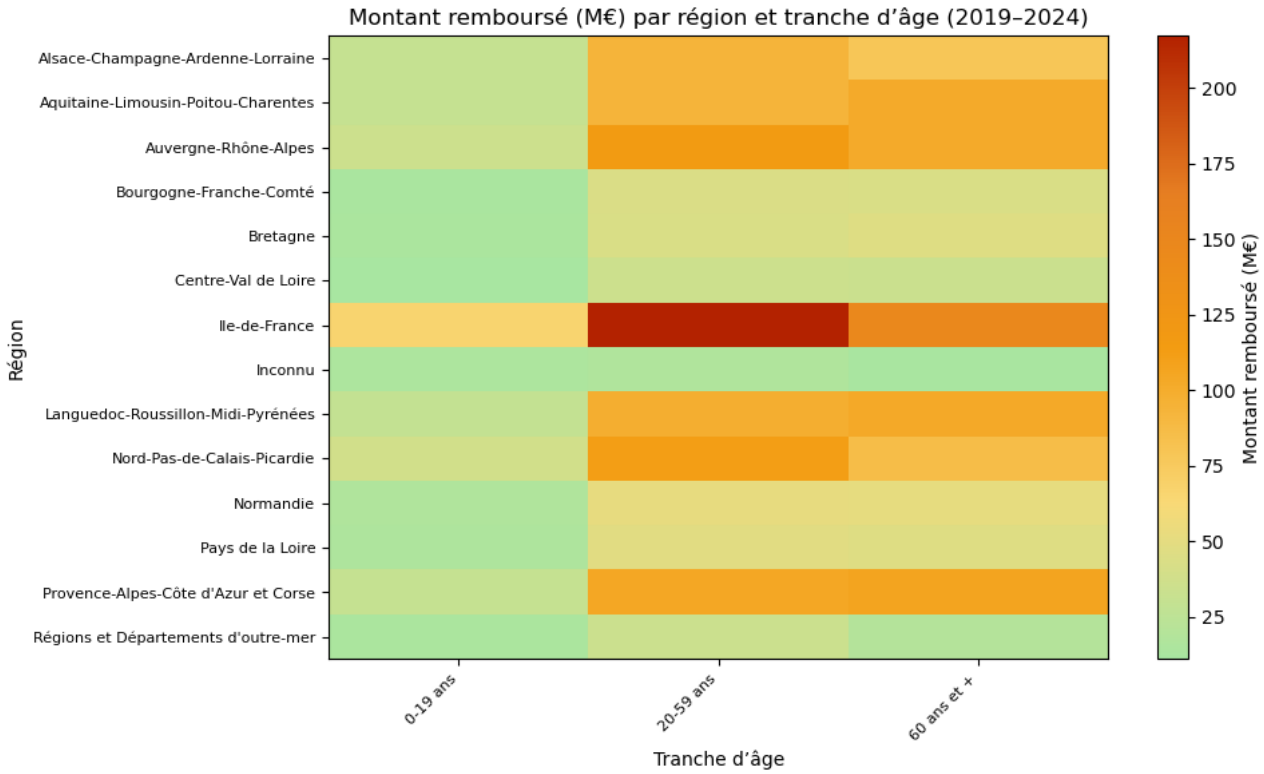


Figure 19 : Heatmap représentant la répartition du montant remboursé par région et tranche d'âge.

La carte thermique met en évidence une forte concentration des remboursements en Île-de-France, principalement dans la tranche d'âge 20 - 59 ans, qui constitue le cœur de la population active.

Les régions Auvergne–Rhône-Alpes et Provence-Alpes-Côte d'Azur et Corse affichent également des montants importants, surtout dans les tranches adultes et seniors.

À l'inverse, les régions moins peuplées (Bretagne, Centre-Val de Loire, Bourgogne-Franche-Comté) présentent des volumes nettement inférieurs.

Cette représentation permet de visualiser la corrélation entre densité de population et volume de remboursement, tout en soulignant la contribution plus marquée des seniors dans les zones rurales.

4.8 Sous-classes d'antibiotiques les plus remboursées

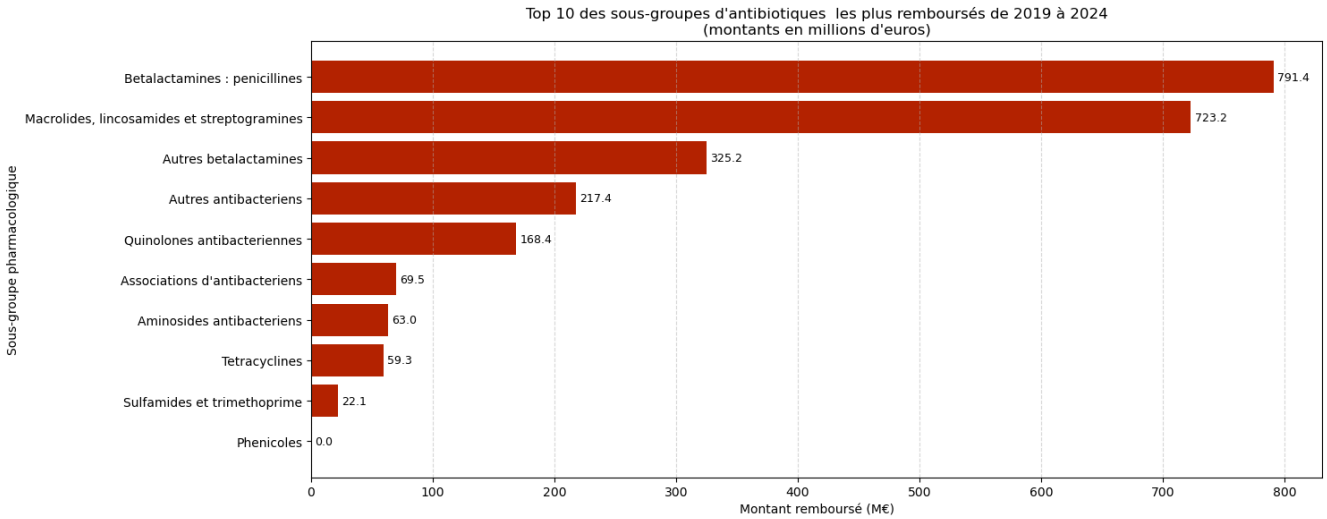


Figure 20 : Principaux sous-groupes d'antibiotiques selon le montant remboursé (en millions d'euros).

Le classement met en avant deux familles largement dominantes :

- les bêta-lactamines : pénicillines (J01C) avec 791 millions d'euros remboursés,
- et les macrolides, lincosamides et streptogramines (J01F) avec 723 millions d'euros.

Ces deux classes représentent à elles seules près des deux tiers du total remboursé.

Elles sont suivies des autres bêta-lactamines (J01D), puis des quinolones antibactériennes (J01M) et des autres antibactériens (J01X).

Cette hiérarchie reflète les pratiques de prescription de première intention, où les pénicillines restent la classe la plus utilisée dans le traitement des infections respiratoires ou ORL.

4.9 Spécialités pharmaceutiques les plus remboursées

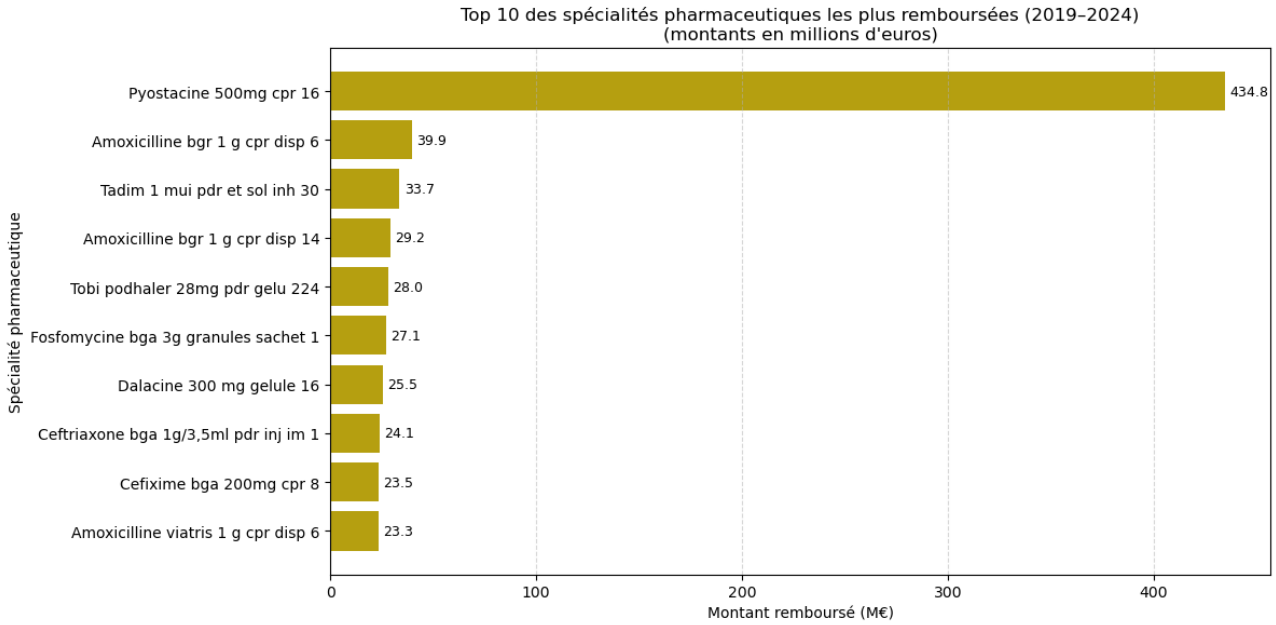


Figure 21 : Médicaments antibiotiques les plus remboursés sur la période 2019 - 2024.

Le graphique révèle la forte domination de la spécialité Pyostacine 500 mg comprimé, avec plus de 434 millions d'euros remboursés, soit un écart considérable par rapport à la deuxième position.

Cette spécialité est suivie par plusieurs produits à base d'amoxicilline, comme *Amoxicilline bgr 1 g*, *Tadim*, ou *Ceftriaxone*, dont les montants varient entre 23 et 40 millions d'euros.

Cette distribution montre une forte concentration des remboursements sur un petit nombre de produits, confirmant la dépendance du marché antibiotique français à quelques molécules clés, notamment l'amoxicilline et ses associations.

.

4.10 Analyse par type de prescripteur

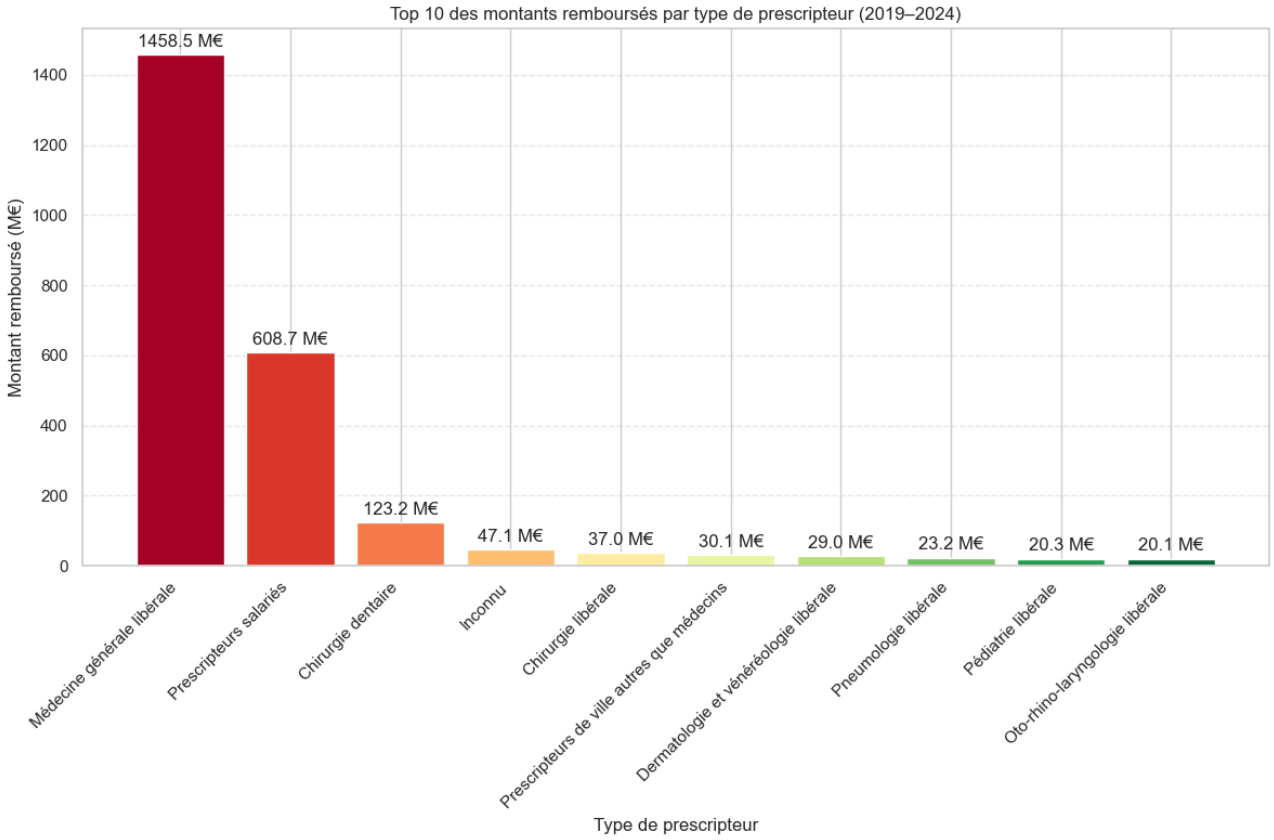


Figure 22 : Répartition des montants remboursés selon la catégorie de prescripteur.

Les médecins généralistes libéraux dominent largement le classement, avec 1,46 milliard d'euros remboursés, soit plus du double des prescripteurs salariés (608 M€).

Les chirurgiens-dentistes arrivent en troisième position, avec 123 M€, suivis de plusieurs spécialités à contribution marginale (pédiatrie, pneumologie, dermatologie, ORL).

Cette distribution met en évidence le rôle central de la médecine générale dans la prescription d'antibiotiques en France, reflet d'un système de soins reposant sur le premier recours.

4.11 Répartition relative des prescripteurs

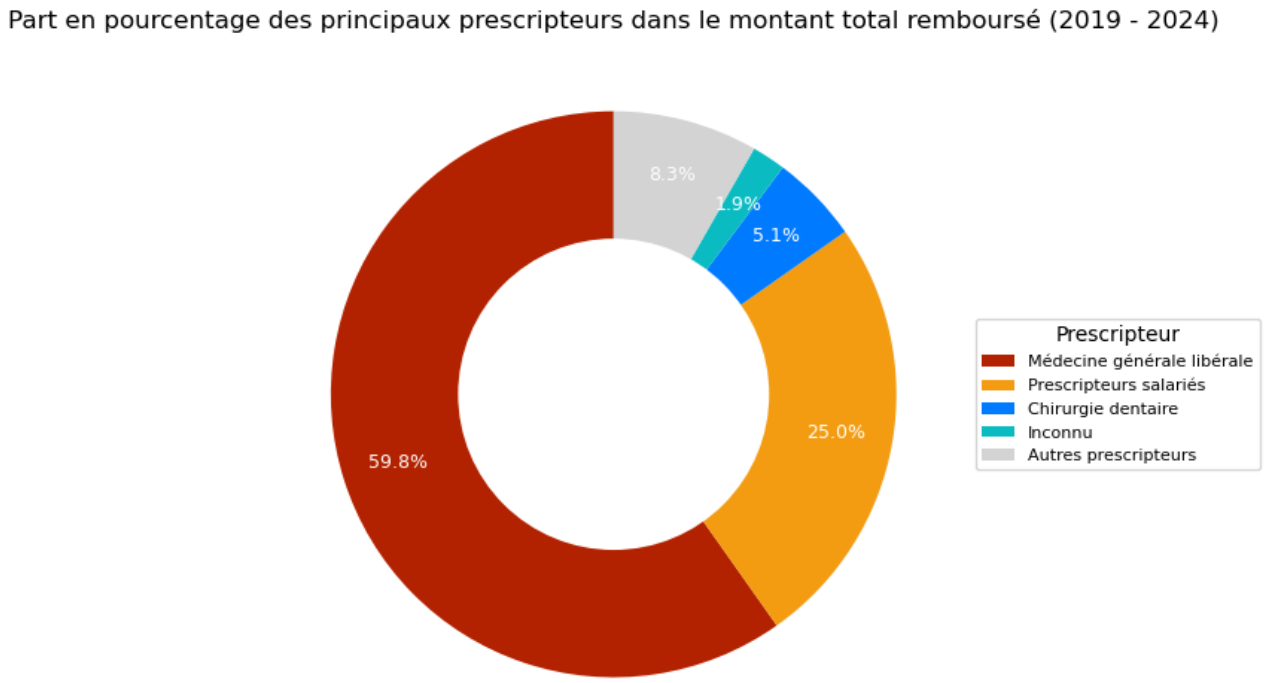


Figure 23 : Répartition des principaux types de prescripteurs en pourcentage du montant global remboursé.

Le graphique en anneau illustre visuellement la concentration du volume de prescription :

- les médecins généralistes libéraux représentent près de 60 % du total,
- les prescripteurs salariés environ 25 %,
- et les chirurgiens-dentistes un peu plus de 5 %.

Les autres catégories (pédiatres, pneumologues, dermatologues, etc.) ne dépassent pas individuellement les 2 %.

Cette structure confirme que la prescription d'antibiotiques reste très concentrée sur la médecine de ville, principalement libérale, ce qui justifie les efforts de sensibilisation ciblant cette profession dans les politiques de santé publique.

4.12 Synthèse générale des analyses visuelles

L'ensemble des visualisations réalisées entre 2019 et 2024 met en évidence une tendance générale stable, marquée par des différences significatives selon les populations et les territoires.
Sur le plan global, les montants remboursés pour les antibiotiques restent élevés, traduisant une consommation encore importante malgré les efforts de maîtrise engagés depuis plusieurs années.
L'année 2020 constitue toutefois un point de rupture notable, lié à la crise du COVID-19, le confinement et la réduction des consultations ont entraîné une baisse temporaire des prescriptions. Dès 2021, la consommation reprend, témoignant d'un retour progressif à la normale dans les comportements de soins.
Du point de vue sociodémographique, les femmes demeurent les plus consommatrices, avec plus de la moitié des montants remboursés. Cette surreprésentation est observée dans toutes les tranches d'âge, mais elle est particulièrement marquée entre 20 et 59 ans, période où les consultations médicales sont les plus fréquentes.
Les personnes âgées de 60 ans et plus constituent le second groupe majeur de consommateurs, ce qui s'explique par une plus grande fragilité immunitaire et la présence de comorbidités chroniques nécessitant un suivi antibiotique régulier.
Sur le plan territorial, l'Île-de-France domine largement le classement des montants remboursés, devant l'Auvergne-Rhône-Alpes et la Provence-Alpes-Côte d'Azur et Corse. Ces écarts sont étroitement corrélés à la densité de population et à la concentration urbaine.
Les régions de l'Ouest affichent des montants plus faibles, mais cela ne traduit pas nécessairement une meilleure maîtrise, pour évaluer les disparités réelles, une normalisation par habitant serait indispensable.
Concernant les familles d'antibiotiques, les pénicillines et macrolides constituent les classes dominantes, confirmant leur place centrale dans la médecine générale pour le traitement des infections respiratoires et ORL.
La concentration du remboursement sur un petit nombre de molécules, dont la Pyostacine et l'Amoxicilline, illustre une dépendance thérapeutique forte du système de soins français à certaines spécialités bien établies.
Enfin, l'analyse des prescripteurs montre une domination écrasante de la médecine générale libérale, responsable de près de 60 % des prescriptions d'antibiotiques.
Ce résultat confirme le rôle pivot du médecin généraliste dans la chaîne de prescription, mais

souligne également la nécessité de poursuivre les efforts de sensibilisation à l'usage raisonné des antibiotiques au sein de cette profession.

En résumé, ces analyses visuelles offrent une lecture complète et nuancée de la consommation d'antibiotiques en France, une tendance globalement stable, des disparités territoriales et démographiques marquées, une forte concentration thérapeutique, et un poids décisif de la médecine générale dans la dynamique de prescription.

Elles constituent une base solide pour des analyses futures plus avancées, notamment en modélisation prédictive (prévisions de consommation), en analyse saisonnière, ou en études d'impact des politiques publiques.

5. CONCLUSION

Ce premier volet du projet consacré à l'analyse de la consommation d'antibiotiques en France à partir des données OpenMedic (2019 - 2024) a permis de parcourir l'ensemble du cycle de préparation et d'exploration des données en environnement Python et Pandas.

L'objectif principal était de transformer un jeu de données brut, volumineux et hétérogène en une base propre, homogène et exploitable, apte à servir de fondement aux visualisations et analyses ultérieures.

Pour y parvenir, un travail rigoureux de nettoyage, filtrage et harmonisation a été mené, typage des colonnes, suppression des valeurs incohérentes, normalisation des variables et regroupement des fichiers annuels en un seul ensemble cohérent couvrant six années.

Une fois cette phase de preprocessing achevée, le jeu de données consolidé a permis de produire plusieurs analyses descriptives et graphiques mettant en lumière :

- l'évolution annuelle des montants remboursés pour les antibiotiques,
- les différences selon le sexe, l'âge et la région,
- la répartition des classes thérapeutiques et des spécialités pharmaceutiques les plus prescrites,
- et la contribution respective des différents types de prescripteurs.

Ces visualisations ont offert une première lecture claire et documentée de la dynamique de consommation des antibiotiques en France.

Elles ont aussi permis de confirmer certaines tendances connues, comme la prépondérance féminine, la domination de l'Île-de-France et la place centrale des médecins généralistes, tout en apportant un regard chiffré et actualisé sur la période post-COVID.

Sur le plan méthodologique, ce projet a représenté une mise en pratique complète des compétences fondamentales du Data Analyst, collecte, nettoyage, structuration, agrégation et visualisation des données à l'aide de Python (Pandas, Matplotlib, Seaborn).

Cette première étape constitue ainsi le socle technique et analytique du projet global.

Elle prépare directement la suite du travail, qui consistera à exploiter le fichier nettoyé "open_medic_cleaned.csv" dans Power BI, afin de concevoir un dashboard interactif permettant une exploration dynamique des données et une communication visuelle plus accessible pour le grand public.

6. BIBLIOGRAPHIE ET SOURCES

- CNAM, OpenMedic
- Santé publique France – Dossier antibiorésistance
- OMS, Global AMR/GLASS
- INSEE, Données démographiques
- Documentation Pandas / Matplotlib / NumPy