Rapport Projet CPES2 2025

Analyse de données sous R

UTHAYAKUMAR Tharushan

tharushan.uthayakumar@hec.edu

Dépôt GitHub associé

23 mai 2025

Résumé

Ce rapport présente deux analyses distinctes réalisées sous R (version 4.5.0) :

— **Partie A** : étude de l'influence de la température moyenne sur le pic journalier de consommation électrique (2012–2025), via une régression linéaire simple, révélant une relation négative forte ($R^2 = 0{,}69$).

— **Partie B** : exploration des données d'effectifs de patients par pathologie, sexe, classe d'âge et territoire (départements, régions), avec visualisations (cartes, heatmaps, évolutions temporelles), afin de décrire la répartition et les tendances des prévalences.

Les méthodologies font appel aux packages `ggplot2`, `dplyr`, `lubridate`, `sf` et `viridis`. Le code est disponible sur le dépôt GitHub ci-dessus.

Table des matières

Introduction

Le présent document propose deux volets d'analyse de données réalisés en langage R. Chaque partie décrit les données, la préparation, la méthodologie suivie, les résultats obtenus et les principales conclusions.

1 Partie A – Pic journalier de la consommation brute d'électricité

1.1 Contexte et objectif

Le premier jeu de données (accessible sur data.gouv.fr) couvre la période 2012–2025 et fournit pour chaque date :

— le pic journalier de consommation électrique en MW,
— la température moyenne observée (Temp_moy, °C),
— la température climatologique de référence (Temp_ref, °C).

L'objectif est de quantifier l'impact de la température moyenne sur la consommation à l'aide d'un modèle de régression linéaire simple.

1.2 Méthodologie

1.2.1 Environnement et packages

Analyse réalisée sous R 4.5.0, avec :

— `ggplot2` pour les visualisations,
— `dplyr` et `lubridate` pour la préparation et la manipulation,
— `scales` pour le formatage des axes.

1.2.2 Préparation des données

Lectures, conversion de la colonne `Date`, calcul de moyennes mensuelles, et vérification des valeurs manquantes.

1.3 Exploration visuelle

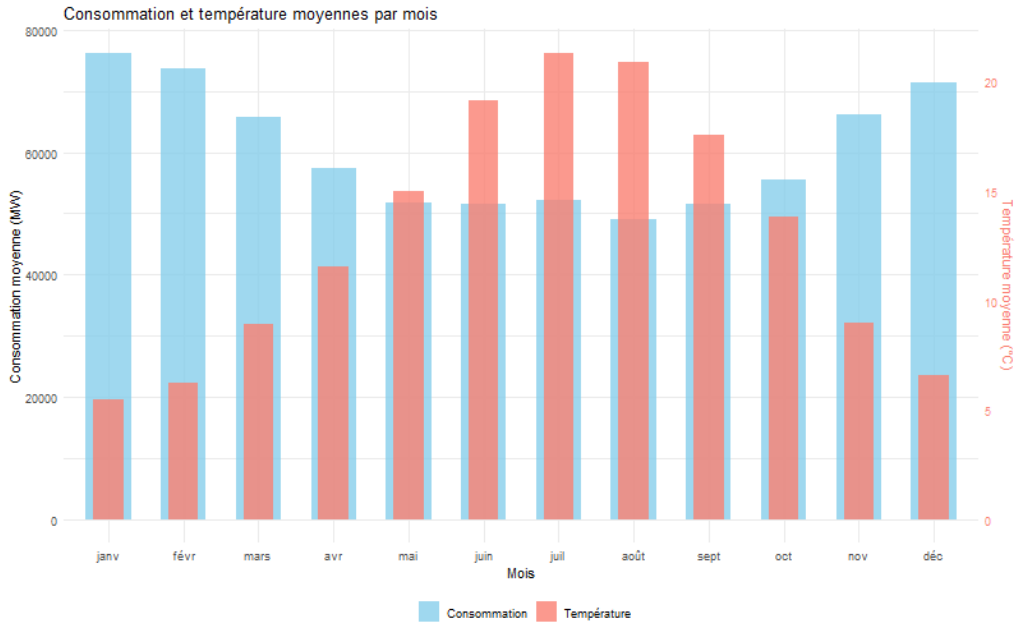

FIGURE 1 – Consommation et température moyennes par mois

— Les mois froids (janv.–fév.) ont une basse température et une haute consommation.
— En été (juil.–août), température élevée et consommation minimale.
— Confirme l'impact direct de la température sur la demande électrique.

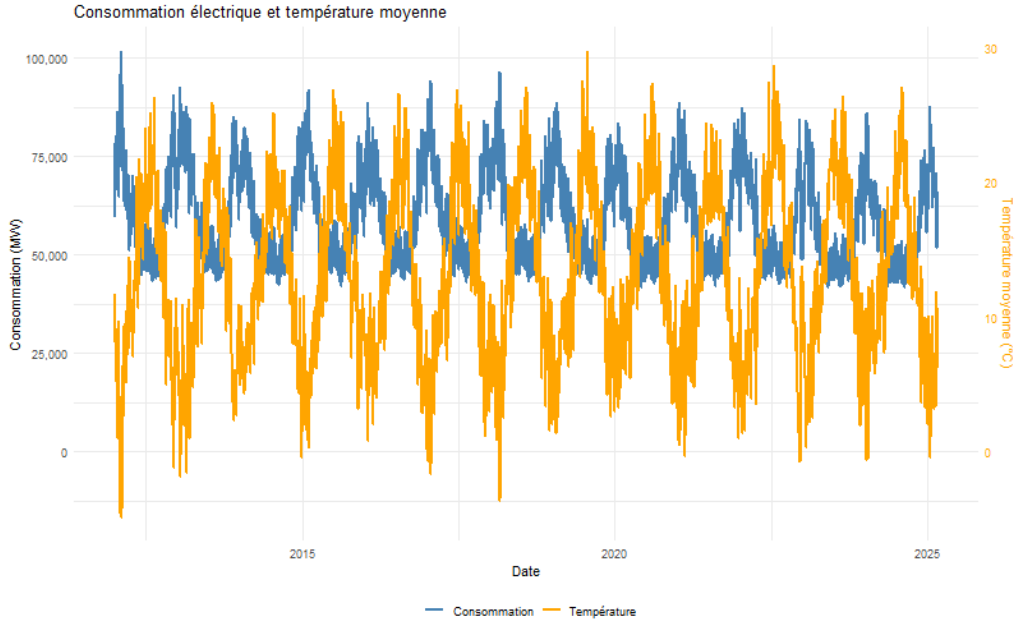

FIGURE 2 – Superposition Consommation & Température (2012–2025)

— Cycles saisonniers opposés : consommation haute quand température basse et vice-versa.
— Calage immédiat, illustrant une dépendance sans délai notable.

1.4 Modèle de régression linéaire

Le modèle ajusté sur $n = 3652$ jours se formule ainsi :

$$\widehat{\text{Conso}} = \beta_0 + \beta_1 \times \text{Temp\_moy} + \varepsilon.$$

Les résultats du modèle sont présentés dans le tableau ci-dessous :

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	79 642.0	209.0	381.4	$< 2 \times 10^{-16}$
Temp_moy	−1 496.4	14.5	−102.9	$< 2 \times 10^{-16}$

TABLE 1 – Coefficients du modèle de régression linéaire simple

Qualité de l'ajustement

- $R^2 = 0.688$, $R^2_{\text{adj}} = 0.688$: la température moyenne explique 68,8 % de la variance du pic de consommation.
- Statistique $F = 1{,}06 \times 10^4$ (ddl : 1 et 3650), $p - valeur < 2{,}2 \times 10^{-16}$: rejet très significatif de l'hypothèse nulle $H_0 : \beta_1 = 0$, ce qui confirme l'existence d'une relation linéaire entre température et consommation.
- Le coefficient $\beta_1 = -1\,496.4$ indique qu'une augmentation de la température moyenne d'un degré Celsius est associée, en moyenne, à une baisse d'environ 1 500 MW du pic journalier de consommation.

1.5 Conclusions Partie A

La température moyenne du jour est un prédicteur fort du pic de consommation, expliquant près de 69 % de sa variabilité. Pour approfondir l'analyse, on pourrait :

- Intégrer des variables calendaires (week-ends, jours fériés).
- Ajouter des indicateurs saisonniers ou des paramètres météorologiques complémentaires (humidité, vent).
- Passer à une régression multiple pour isoler l'impact de la température des autres facteurs.

2 Partie B – Pathologies : effectifs par territoire, âge et sexe

2.1 Contexte et objectif

Le second jeu de données (accessible sur data.ameli.fr) présente pour les années 2019–2022 :

— les effectifs de patients par pathologie,
— par classe d'âge, sexe, département et région,
— avec pour chaque groupe la prévalence en pourcentage.

L'objectif est d'analyser la distribution et l'évolution temporelle des prévalences pour les principales pathologies, en mettant en évidence les disparités selon l'âge, le sexe et le territoire, afin d'orienter d'éventuelles actions de santé publique.

2.2 Méthodologie

— Préparation et nettoyage : filtrage des valeurs manquantes, renommage des variables.
— Cartographie : shapefile des régions, fusion avec la prévalence et rognage sur la métropole.
— Visualisations : barplots, heatmaps, évolutions temporelles, boxplots, scatterplots, pie charts.

2.3 Exploration visuelle

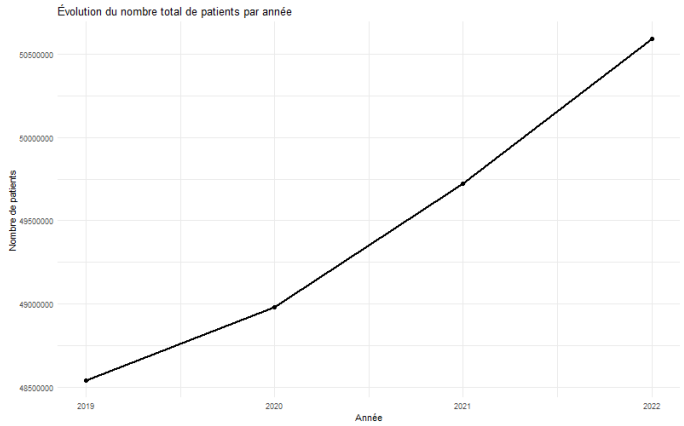

FIGURE 3 – Évolution du nombre total de patients par année

— Une augmentation quasi linéaire du total de patients, passant d'environ 48,5 M en 2019 à 50,5 M en 2022.
— Un signal régulier sans plateau, suggérant une croissance continue de la demande de soins ou de la couverture de données.
— L'absence de point d'inflexion indique aucune rupture majeure (crise sanitaire ou changement de méthodologie).

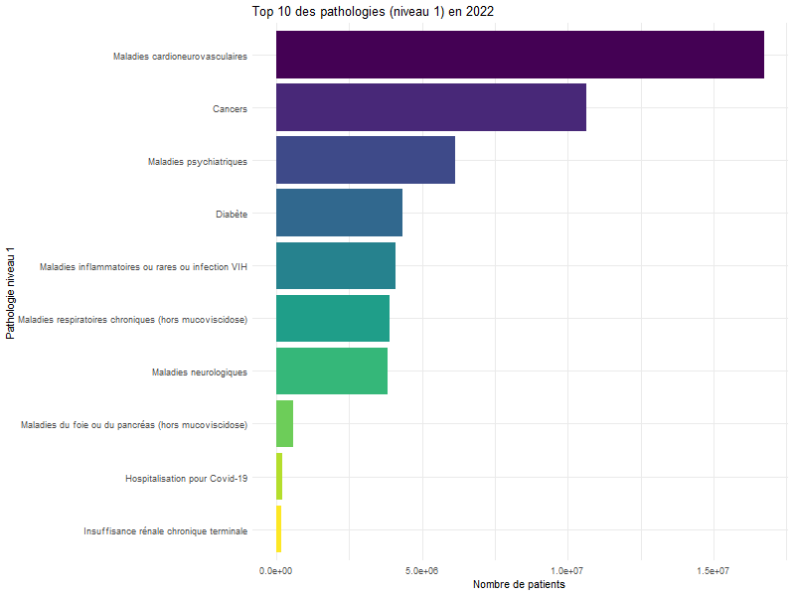

FIGURE 4 – Top 10 des pathologies (niv.1) en 2022

— Les maladies cardiovasculaires en tête (16 M de patients), puis cancers (11 M) et psychiatriques (7 M).
— Diabète et maladies inflammatoires/VIH autour de 4–5 M, position médiane.
— Insuffisance rénale terminale et hospitalisation pour Covid-19 très marginales.

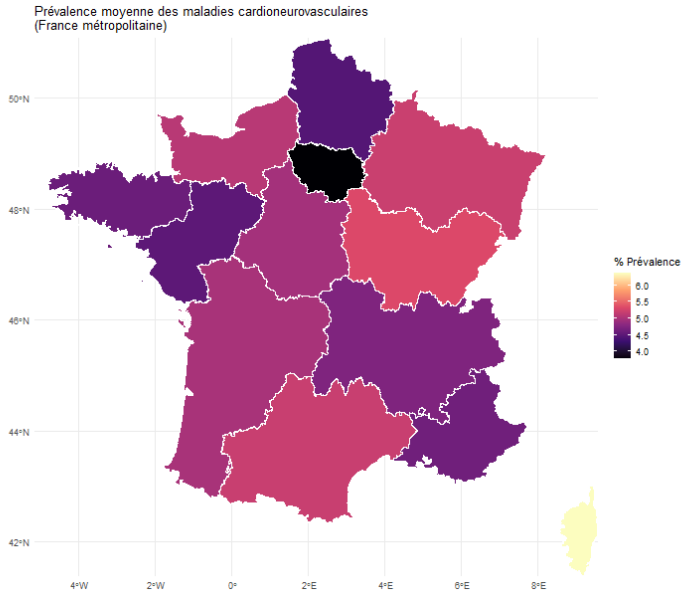

FIGURE 5 – Prévalence des maladies cardiovasculaires (Métropole)

— Île-de-France la plus faible (4 %), Grand Est et Bourgogne-Franche-Comté élevées (5,5 %).
— Arc Est et Sud-Ouest plus exposés, possiblement dû à des facteurs de risque régionaux.
— Rognage sur la métropole améliore la clarté en supprimant les DOM-TOM.

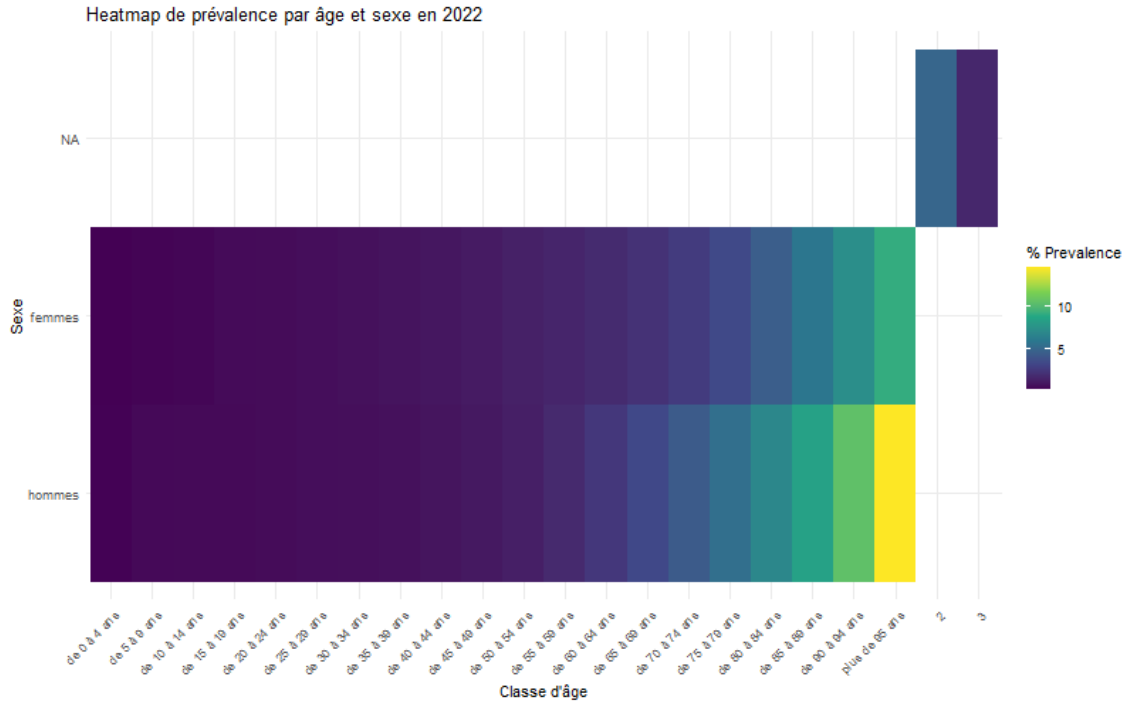

FIGURE 6 – Heatmap prévalence par âge et sexe en 2022

— Croissance exponentielle de la prévalence avec l'âge, de <1 % (<40 ans) à >10 % (>85 ans).
— Légère surreprésentation masculine chez les 65 ans et plus.
— Catégorie « tous sexes » (NA) regroupe des valeurs intermédiaires, moins discriminantes.

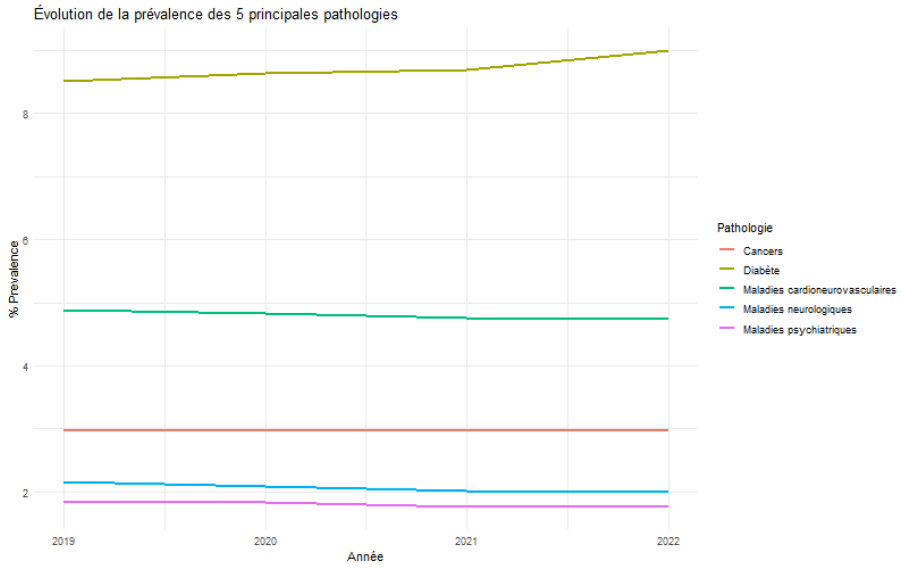

FIGURE 7 – Évolution de la prévalence des 5 pathologies principales

— Diabète en hausse (de 8,5 % à 9 %), indicateur de progression continue.
— Cardioneurovasculaires stables 5 %, cancers (3 %), neurologiques (2 %) et psychiatriques (1,8 %) en légère baisse.
— Différentes dynamiques suggèrent des évolutions distinctes d'incidence ou de prise en charge.

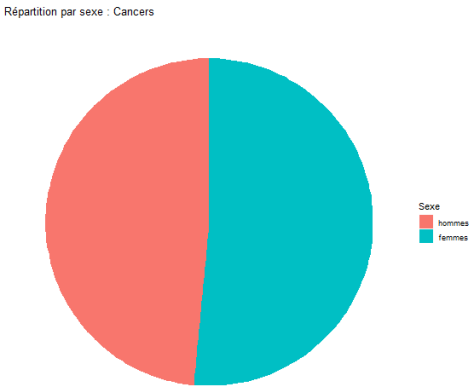

FIGURE 8 – Répartition hommes/femmes pour « Cancers »

— Légère prédominance féminine (52 %) parmi les patients atteints de cancers.
— Répartition presque équilibrée, reflétant la distribution par sexe selon les types de cancer.

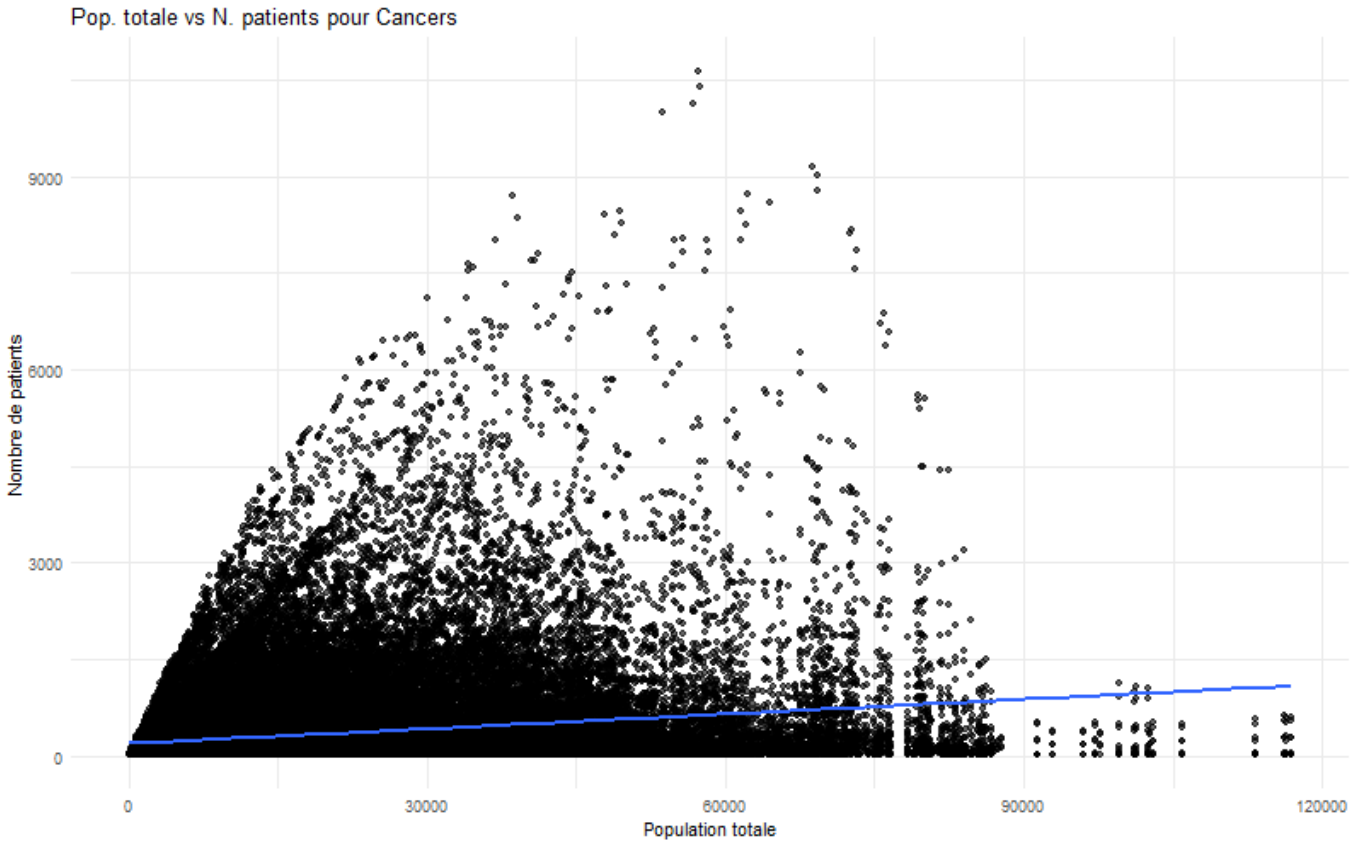

FIGURE 9 – Population vs nombre de patients (Cancers)

— Corrélation positive mais dispersion forte : plus de population, plus de cas, mais variabilité importante.
— Pente de régression faible : +10 000 hab +200 cas, dépendance modérée.
— Territoires très peuplés montrent des prévalences hétérogènes.

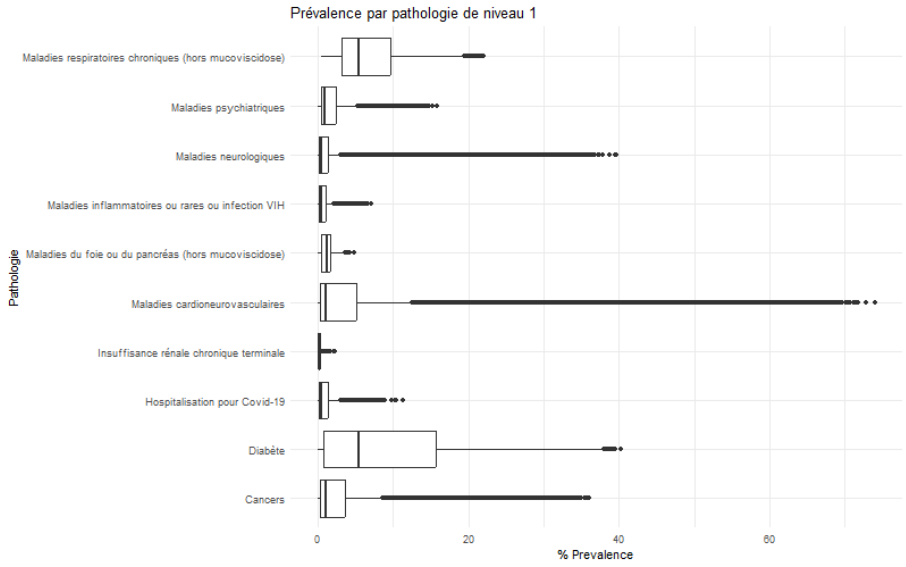

FIGURE 10 – Prévalence par pathologie (niv.1)

— Diabète : Ecart Interquartile large (4–20 %), plus de variabilité départementale.
— Cardioneurovasculaires avec outliers extrêmes (>60 %).
— Insuffisance rénale et Covid-19 très concentrées à faible taux (<5 %).

2.4 Conclusions Partie B

— Les maladies cardioneurovasculaires et les cancers restent les pathologies les plus fréquentes, avec respectivement plus de 16M et 11M de patients en 2022, tandis que le diabète présente une prévalence approchant 9%.
— La cartographie révèle des disparités régionales marquées : certaines zones métropolitaines (Grand Est, Bourgogne-Franche-Comté, Hauts-de-France) enregistrent des prévalences plus élevées que la moyenne nationale.
— La heatmap par âge et sexe montre une progression régulière de la prévalence avec l'âge et une légère prédominance masculine chez les seniors.
— L'évolution temporelle (2019–2022) met en évidence une montée constante du diabète, alors que les autres pathologies se stabilisent ou varient faiblement.
— Les distributions départementales confirment une forte hétérogénéité locale, avec des outliers indiquant des territoires à risque accru.

Conclusions Globales

Cette étude illustre la puissance de R pour l'analyse statistique.

— **Partie A :** la régression simple montre que la température moyenne explique près de 69% de la variabilité du pic de consommation électrique, soulignant l'importance du prévisionnel hivernal.
— **Partie B :** l'exploration des données de santé met en lumière les pathologies prioritaires, leurs disparités géographiques, et leurs évolutions démographiques.

Les approches graphiques (séries temporelles, heatmaps, cartographies) et modélisation fournissent une base solide pour orienter les décisions en politique énergétique et en santé publique.

3 ANNEXE

3.1 Graphiques supplémentaires Partie A

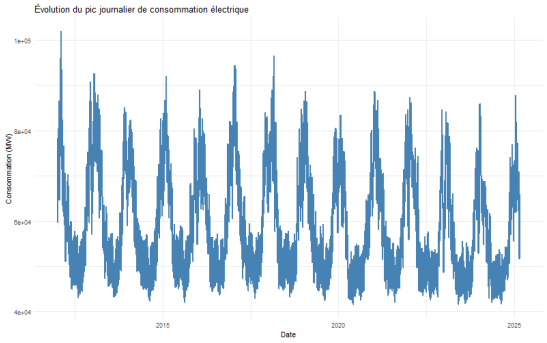

FIGURE 11 – Pic journalier de consommation électrique (2012–2025)

— Cycle annuel marqué : pointes >95 000 MW en hiver, creux 45 000–50 000 MW en été.
— Légère tendance haussière des maxima hivernaux, suggérant une hausse de la demande de chauffage ou d'équipement.

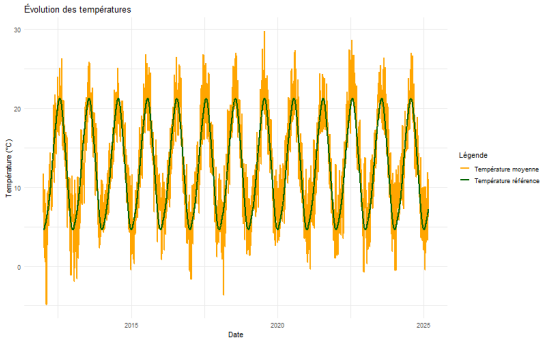

FIGURE 12 – Température moyenne vs référence (2012–2025)

— Température observée (orange) et de référence (verte) très alignées, avec oscillations journalières/saisonnières.
— Max 25–30 °C en été, min parfois <0 °C en hiver ; interannuel stable.

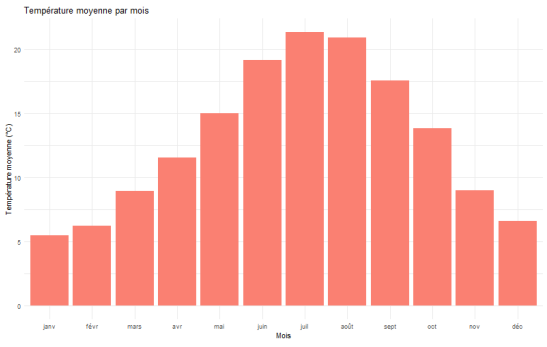

FIGURE 13 – Température moyenne par mois

— Montée linéaire de 5 °C (janv.) à 22 °C (juil.), puis redescente symétrique.
— Forte saisonnalité, avec gradient régulier entre saisons.

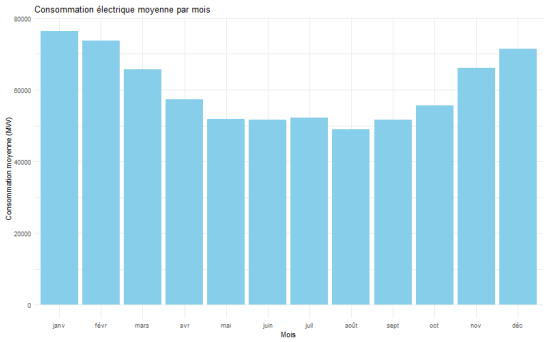

FIGURE 14 – Consommation moyenne par mois

— Profil en « U » : haute consommation en janv.–fév. (75–78 kMW), creux en août (49 kMW), puis remontée.
— Corrélation inverse avec la température, traduisant un usage de chauffage en hiver.

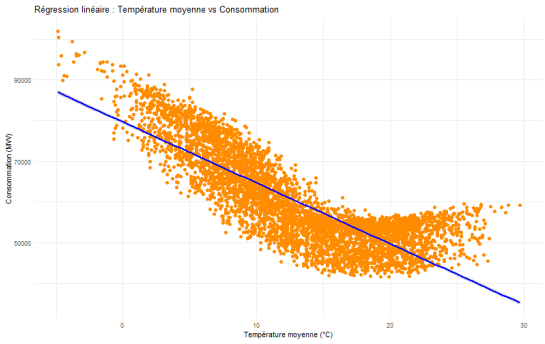

FIGURE 15 – Régression : Température vs consommation

— Relation linéaire négative forte (pente –1 500 MW/°C).
— Dispersion autour de la droite faible, confirmant la fiabilité du modèle.
— 3 652 points journaliers illustrant une robustesse statistique.

3.2 Graphiques supplémentaires Partie B

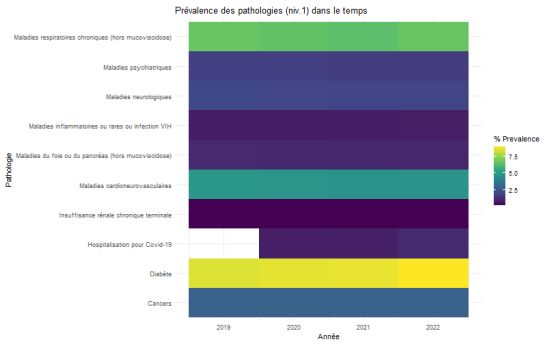

FIGURE 16 – Heatmap temporelle des prévalences (2019–2022)

— Diabète en nette hausse de 2019 à 2022.
— Cancers stables et légèrement croissants, cardiovasculaires constants avec un léger creux en 2021.