RAPPORT DE CONCEPTION TECHNIQUE & ANALYSE DES DONNÉES (STATISTIQUE ET PRÉDICTIVE)

Lounes TAIBI - Kenza Mona EL HITARY - Hope BOME - Aissatou BLONDIN - Khady BA

Objectif et contexte du projet :

Ce projet a pour objectif de simuler, analyser et anticiper la gestion des ressources hospitalières à l'hôpital de la Pitié-Salpêtrière à partir de données synthétiques réalistes. L'enjeu principal est de comprendre l'évolution des admissions, d'évaluer leur impact sur les lits, le personnel et le matériel médical, et de proposer un outil d'aide à la décision capable d'anticiper les situations de tension.

La Pitié-Salpêtrière est un établissement hospitalier de référence, fortement exposé aux pics d'activité saisonniers (hiver, canicules) et aux événements sanitaires exceptionnels, comme la crise du COVID-19. Ces caractéristiques en font un cas d'étude pertinent pour développer des approches de pilotage hospitalier basées sur la data science.

Génération des données synthétiques :

Justification du recours à des données simulées

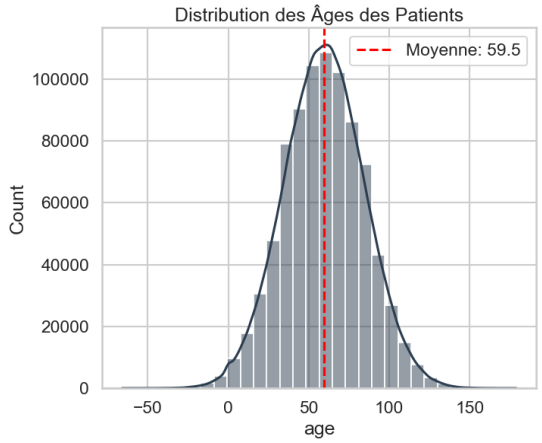
L'utilisation de données réelles hospitalières pose des contraintes fortes en matière de confidentialité, d'accès et d'éthique. Le choix a donc été fait de générer un jeu de données de manière crédible qui permettra de matérialiser le fonctionnement d'un hôpital de grande capacité sur la période 2018–2025.

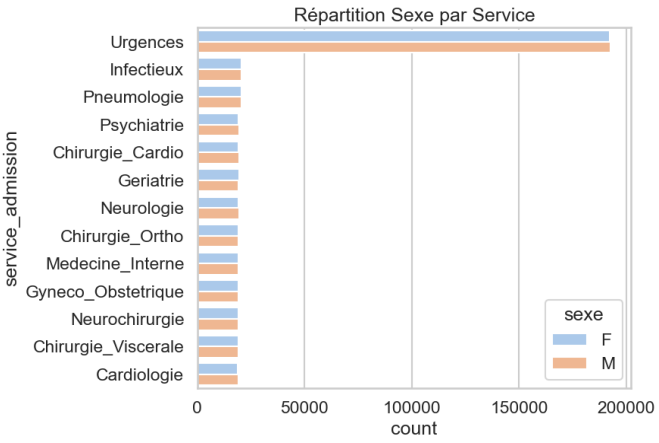
Les admissions quotidiennes sont simulées à l'aide d'une **loi de Poisson**, modèle statistique adapté aux événements discrets et aléatoires dans le temps, dont le paramètre est ajusté par plusieurs composantes explicatives : une saisonnalité annuelle (pics hivernaux), un effet jour de la semaine et des facteurs de rupture liés aux crises sanitaires, notamment le COVID-19. Cette modélisation probabiliste permet de représenter à la fois la variabilité naturelle des flux et les chocs exogènes, tout en conservant des ordres de grandeur réalistes observés dans les hôpitaux français.

La durée d'hospitalisation est quant à elle modélisée à partir d'une **distribution log-normale**, choix justifié par la forte asymétrie des durées de séjour, caractérisée par une majorité de séjours courts et une minorité de séjours longs à fort impact sur les ressources. Les ressources humaines et matérielles sont intégrées de manière conjointe dans le processus de génération : les effectifs présents sont calculés à partir d'un effectif théorique corrigé par un taux d'absentéisme variable (saisonnier et contextuel), avec distinction des équipes jour/nuit. La disponibilité réelle des lits est ensuite déduite de ces effectifs et des taux d'occupation simulés, garantissant une **cohérence métier forte** entre charge de soins, capacité opérationnelle et contraintes organisationnelles.

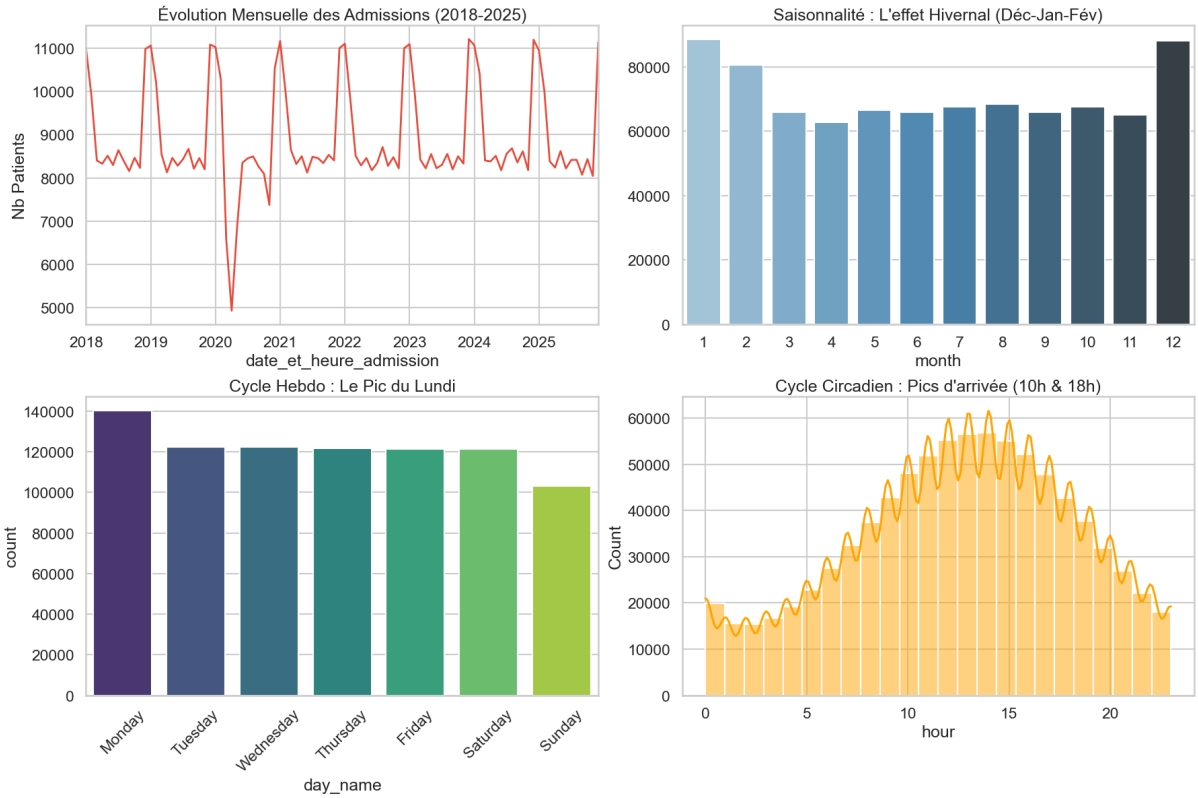
Analyse descriptive et statistique des données

L'analyse descriptive des 852 492 admissions enregistrées entre 2018 et 2025 révèle une activité hospitalière intense caractérisée par une population mature (âge moyen de 59,5 ans) et une prédominance de pathologies respiratoires, premier motif d'admission avec plus de 143 500 cas. La sévérité clinique reste globalement modérée mais significative, comme l'indique un score CCMU moyen de 2,3, tandis que la durée d'hospitalisation de 20,7 heures en moyenne souligne l'efficacité du flux de rotation, malgré des cas complexes pouvant s'étendre sur plusieurs jours. Le service des Urgences s'impose comme la porte d'entrée majeure de l'établissement (45 % des admissions), aboutissant majoritairement à des transferts internes ou externes, ce qui confirme le rôle pivot du plateau technique dans l'orientation rapide des patients au sein du parcours de soins.

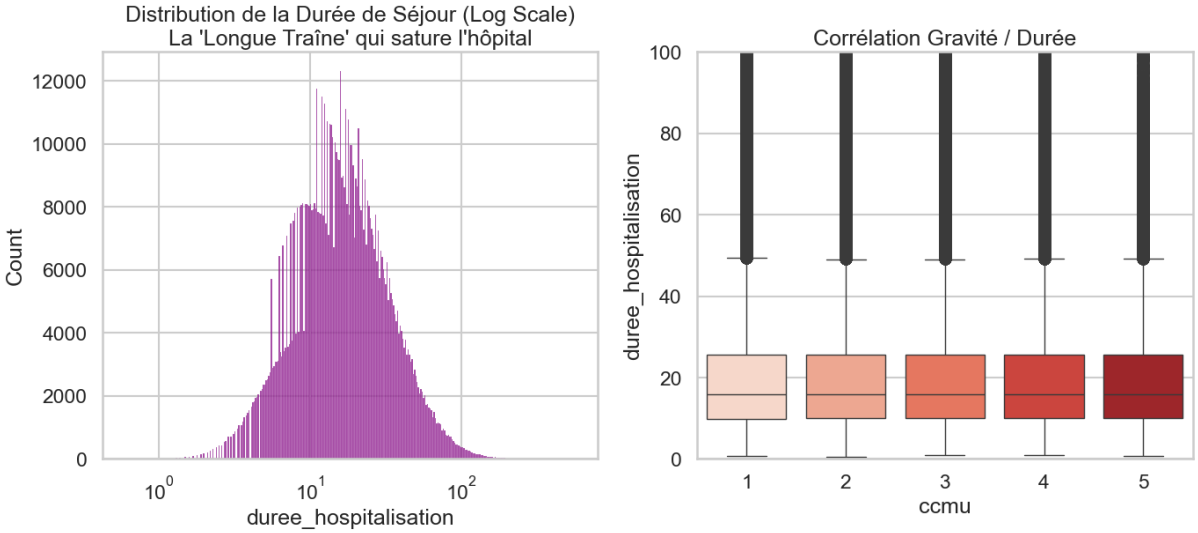




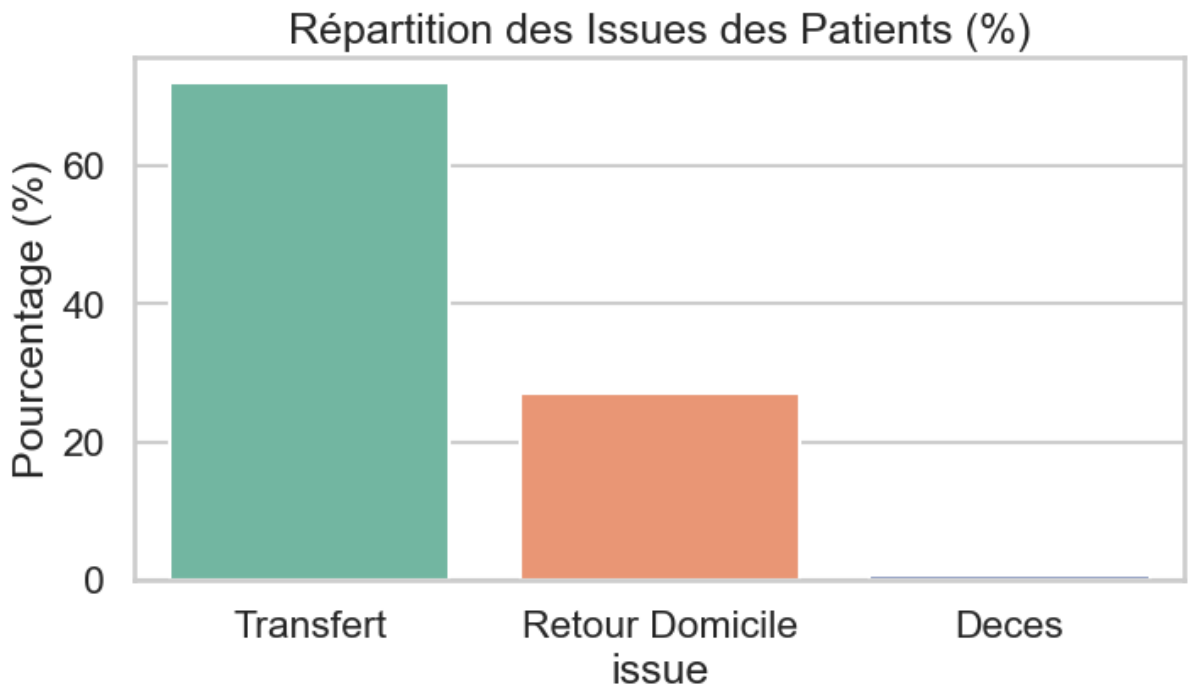
L'analyse temporelle des admissions sur la période 2018-2025 met en évidence une structure d'activité hautement prévisible et cyclique, caractérisée par une stabilité globale des volumes interrompue uniquement par la chute drastique de 2020 (impact Covid). Cette dynamique est régie par une triple saisonnalité : annuelle, avec une nette surcharge hivernale (particulièrement en décembre et janvier) ; hebdomadaire, marquée par un "pic du lundi" significatif contrastant avec une baisse d'activité le week-end ; et circadienne, où les flux se concentrent massivement sur les heures diurnes avec des points culminants autour de 10h et 18h.



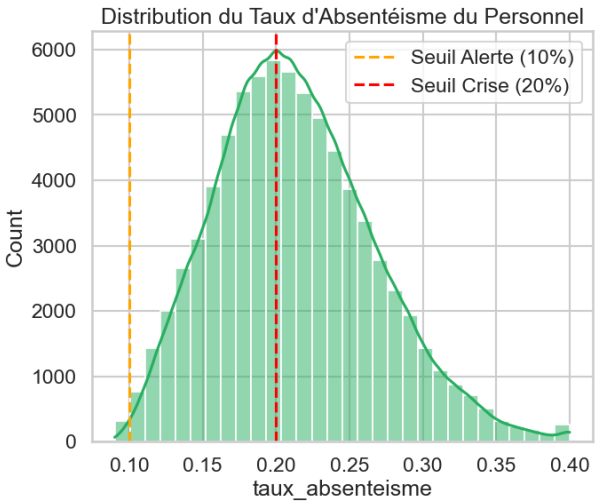
L'analyse croisée des flux et des durées de séjour révèle une activité hospitalière sous tension, caractérisée par une prévisibilité des entrées mais une gestion critique des sorties. Les admissions suivent une **triple cyclicité rigide** : une saisonnalité annuelle marquée par une surcharge hivernale (décembre-janvier), une dynamique hebdomadaire dominée par le "pic du lundi", et un rythme quotidien bimodal concentrant l'essentiel des arrivées à 10h et 18h. Cette pression sur les flux entrants se heurte à une problématique de capacité illustrée par la distribution des durées de séjour : la présence d'une "longue traîne" indique qu'une minorité de patients mobilise durablement les ressources, saturant l'hôpital indépendamment de la gravité immédiate (le score CCMU n'influençant que marginalement la durée médiane), ce qui suggère que les blocages sont moins d'ordre clinique qu'organisationnels ou liés aux difficultés de transfert en aval.

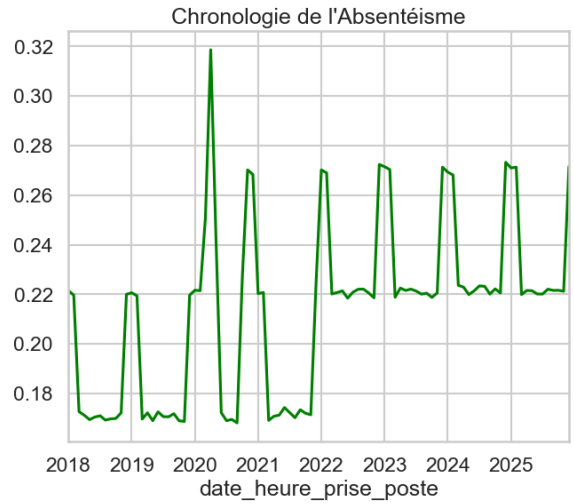


L'examen des données hospitalières révèle une activité structurellement saturée par une **triple cyclicité des admissions** et des goulots d'étranglement logistiques en sortie. Les flux d'entrée présentent une prévisibilité forte, rythmée par une saisonnalité hivernale marquée (pics en décembre et janvier), une pression hebdomadaire culminant le lundi et une dynamique circadienne bimodale centrée sur 10h et 18h. Cette affluence prévisible se heurte cependant à une **inertie de sortie critique** : bien que la gravité clinique (score CCMU) n'impacte que très peu la durée médiane de séjour, la présence d'une « longue traîne » de patients restant hospitalisés sur de longues périodes suggère des blocages organisationnels profonds. Cette saturation est accentuée par une dépendance majeure aux structures d'aval, puisque plus de 70 % des issues de patients se concluent par un transfert, rendant la fluidité globale de l'établissement tributaire de la disponibilité des lits dans les autres unités ou services de soins de suite.

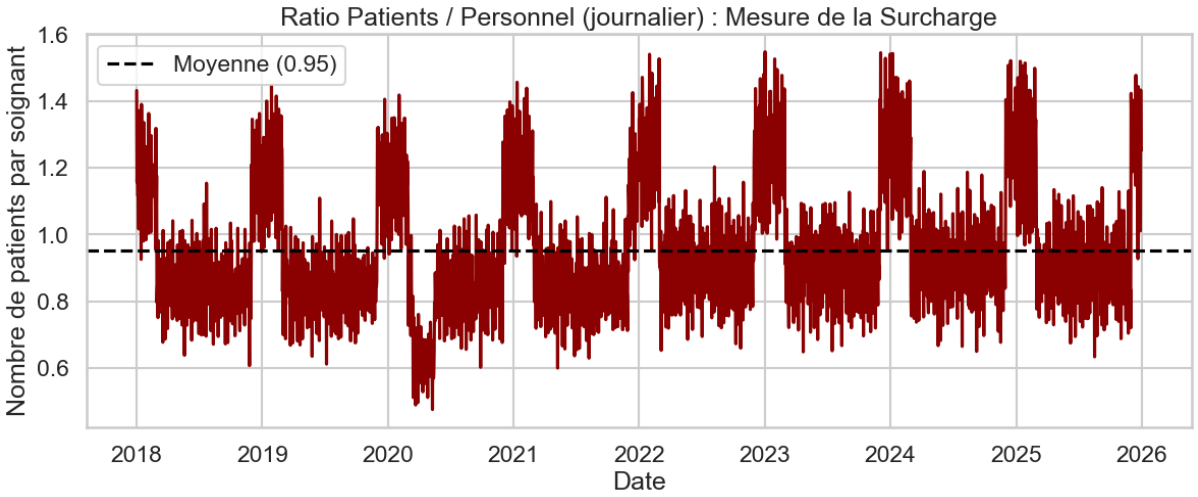


L'examen de la chronologie et de la distribution de l'absentéisme révèle une situation de crise RH structurelle et alarmante pour l'établissement. Le taux d'absentéisme médian se situe autour de **20 %**, atteignant ainsi systématiquement le "seuil de crise" défini, avec une part significative des relevés dépassant même les **30 %**. L'analyse temporelle montre un basculement critique : après une période de relative stabilité (malgré un pic ponctuel majeur en 2020 probablement lié à la pandémie), le plancher de l'absentéisme a subi une **élévation brutale dès le début de l'année 2022**, passant d'un socle de 17 % à plus de 22 % en routine. Cette dégradation persistante, couplée à des pics cycliques de plus en plus hauts, suggère un épuisement chronique des équipes qui menace directement la capacité opérationnelle à répondre aux pics d'admissions identifiés par ailleurs.



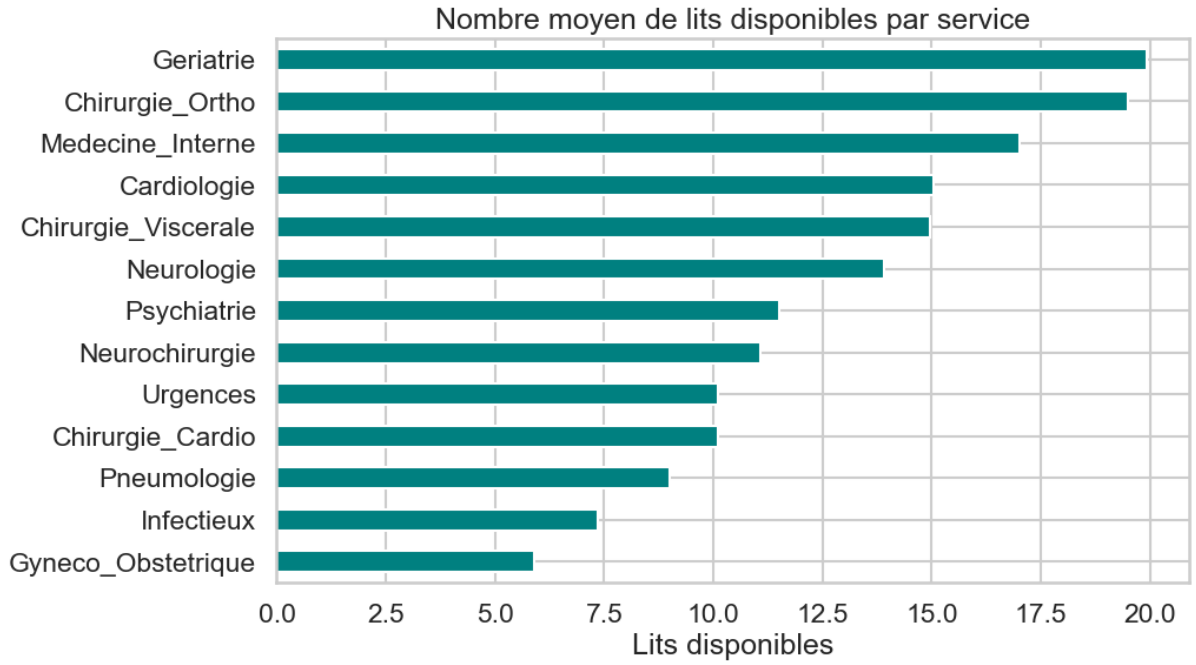


D'ailleurs, ce graphique illustre l'évolution du ratio patients/soignant entre 2018 et début 2026, révélant une surcharge de travail chronique et cyclique au sein de l'établissement. Bien que la moyenne historique s'établisse à 0,95, on observe des pics récurrents dépassant systématiquement le seuil de 1,4, correspondant probablement à des périodes de tension saisonnière ou de crises sanitaires. L'analyse met en évidence une forte volatilité quotidienne ainsi qu'une anomalie notable au premier semestre 2020, où le ratio a chuté sous les 0,6 avant de remonter brutalement. Depuis 2022, les épisodes de surcharge semblent gagner en intensité et en fréquence, suggérant une pression accrue sur le personnel soignant qui opère régulièrement bien au-dessus de la moyenne de référence, ce qui pose des risques directs sur la qualité des soins et l'épuisement professionnel.

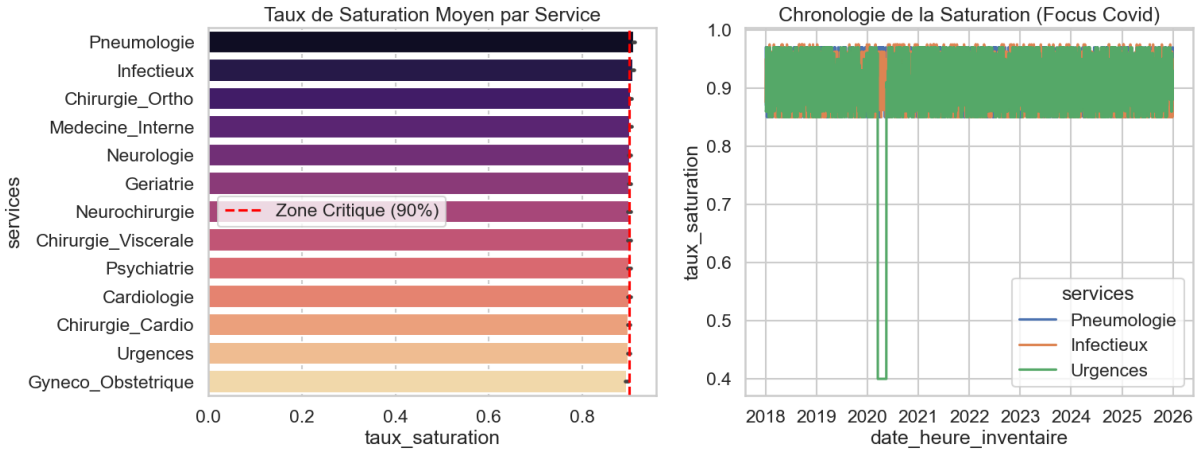


Cependant, l'examen des capacités d'accueil par service révèle une disparité majeure dans la disponibilité des ressources, avec une tension critique concentrée sur les spécialités de court séjour et les urgences. Alors que les services de **Gériatrie** et de **Chirurgie Orthopédique** conservent les moyennes les plus élevées (environ 20 lits disponibles), les secteurs de première ligne comme les **Urgences** ou la **Pneumologie** disposent de marges de manœuvre réduites avec seulement 10 lits ou moins en moyenne. Le point de rupture le plus préoccupant se situe en **Gynécologie-Obstétrique**, qui affiche la disponibilité la plus faible de l'établissement (environ 6 lits), limitant drastiquement la capacité d'absorption de nouveaux flux dans ce service. Cette répartition suggère un risque élevé de saturation rapide lors des pics d'activité circadiens ou hebdomadaires identifiés, particulièrement

pour les services dont le nombre moyen de lits disponibles est inférieur à 10.



L'analyse des capacités d'accueil par service met en lumière une répartition hétérogène des ressources, révélant des vulnérabilités critiques dans les unités de première ligne. Tandis que la **Gériatrie** et la **Chirurgie Orthopédique** disposent des marges de manœuvre les plus larges avec environ **19 à 20 lits disponibles** en moyenne, les services de court séjour et spécialisés sont en situation de tension immédiate. Les **Urgences** et la **Pneumologie** opèrent avec un seuil de sécurité réduit (environ 10 lits), mais le point de rupture le plus préoccupant concerne les services d'**Infectiologie** et de **Gynécologie-Obstétrique**, ce dernier ne disposant que de **6 lits disponibles** en moyenne. Cette faible disponibilité dans des secteurs clés suggère que l'établissement pourrait ne pas être en mesure d'absorber sereinement les pics d'admissions quotidiens ou saisonniers, particulièrement si les sorties et les transferts vers l'aval ne sont pas accélérés en amont.



Donc, l'analyse exploratoire met en évidence une saisonnalité marquée des admissions, une pression structurelle sur les urgences et un rôle déterminant des ressources humaines dans les situations de saturation. Les visualisations temporelles, distributions et heatmaps ont été choisies pour identifier rapidement les périodes critiques, les services sous tension et l'évolution de la gravité clinique (CCMU). Les incohérences détectées dans les données simulées (âges négatifs, valeurs

extrêmes) ont été corrigées lors du preprocessing, assurant la fiabilité des analyses et la préparation des variables pour la modélisation prédictive.

Preprocessing - Feature Engineering

Lors de l'analyse exploratoire, certaines incohérences ont été identifiées, notamment des âges négatifs ou extrêmes.

Le preprocessing repose sur une architecture de traitement hybride qui combine le nettoyage de données structurées et la préparation de séries temporelles, spécifiquement conçue pour l'analyse des flux hospitaliers. La structure utilise des bibliothèques standards de l'écosystème Python (comme `pandas` et `numpy`) pour l'ingénierie des caractéristiques (feature engineering), permettant de transformer des données brutes en indicateurs exploitables, tels que le ratio patients/personnel visible sur votre graphique. La conception privilégie une approche séquentielle où la gestion des valeurs manquantes et la normalisation des types de données (notamment les formats de date) assurent la cohérence nécessaire aux analyses prédictives ultérieures.

Le **feature engineering** a consisté à transformer les données brutes issues de la simulation en variables explicatives directement exploitables par les analyses statistiques et les modèles de machine learning. À partir des horodatages d'admission, plusieurs variables temporelles ont été dérivées, notamment le jour de la semaine, le mois, la saison et des indicateurs binaires identifiant les périodes de crise sanitaire (pré-Covid, vagues Covid, post-Covid). Ces variables permettent de capturer les effets cycliques et les ruptures structurelles observées dans les flux hospitaliers.
Les variables liées aux ressources ont été enrichies par la création de **ratios opérationnels**, tels que le ratio patients/personnel présent, le taux d'occupation des lits et des indicateurs de tension par service. Ces transformations traduisent les contraintes organisationnelles réelles mieux que les valeurs absolues. Les variables catégorielles (service, motif d'admission, type de personnel) ont été encodées de manière adaptée (encodage ordinal ou one-hot selon le cas) afin de préserver l'information métier tout en assurant leur compatibilité avec les modèles de prédiction. Enfin, un nettoyage systématique (suppression des âges négatifs, gestion des valeurs extrêmes) a été appliqué afin de garantir la stabilité et la robustesse des modèles entraînés.

Modèles statistiques et prédictifs

Modélisation statistique

La régression linéaire a été utilisée comme modèle de référence (baseline) afin de disposer d'un point de comparaison simple et interprétable. Elle repose sur l'hypothèse d'une relation linéaire entre une variable explicative basique (ici le jour de la semaine, encodé numériquement) et le flux journalier de patients, selon la forme $Y=aX+b$.

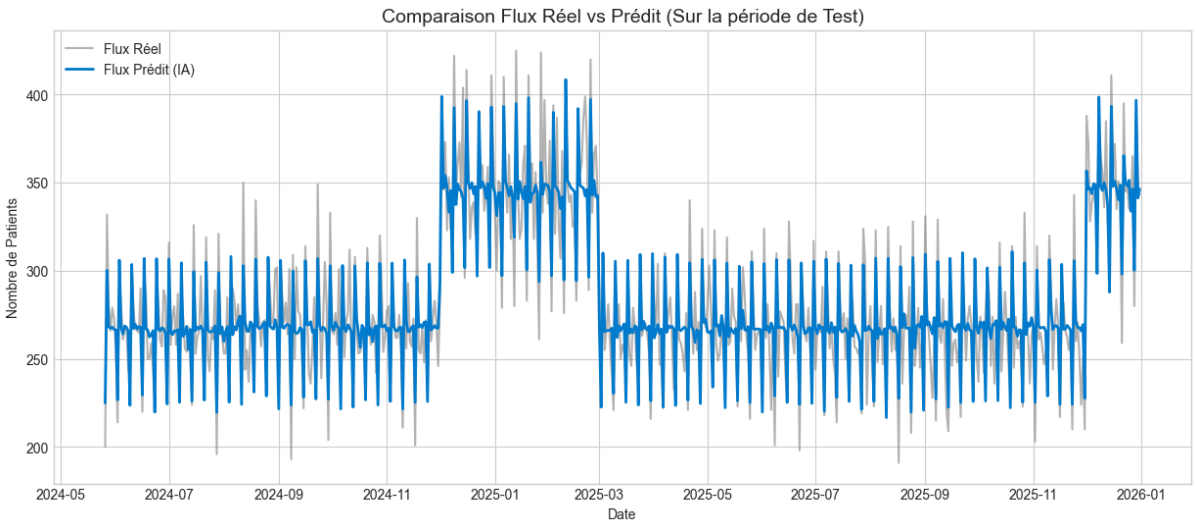
Ce modèle permet d'évaluer rapidement la capacité prédictive minimale atteignable sans ingénierie avancée ni interactions complexes. Les performances obtenues, avec une MAE d'environ 17 patients, montrent une précision limitée mais cohérente avec la simplicité du modèle.

Cependant, ce modèle présente des limites structurelles majeures dans un contexte hospitalier réel. Il est incapable de capturer les non-linéarités et les effets d'interaction entre facteurs calendaires, saisonniers et opérationnels. Par exemple, l'impact combiné d'un lundi hivernal en période de tension sociale (grève ou absentéisme élevé) génère un afflux bien supérieur à la somme de chaque effet pris isolément, phénomène que la régression linéaire ne peut modéliser. Ces limites justifient le recours à des modèles plus avancés, capables d'intégrer des relations complexes et des seuils critiques, comme le gradient boosting.

Modélisation de prédiction

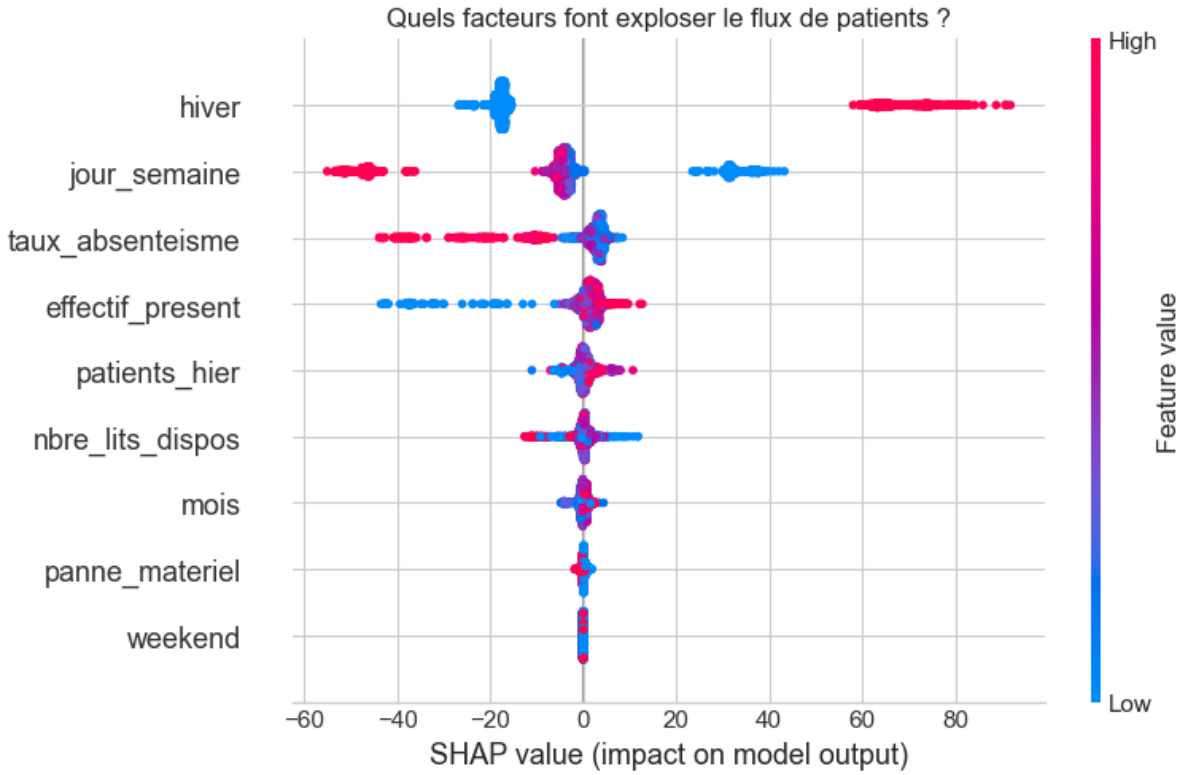
Le modèle de prédiction repose sur un **XGBoost Regressor**, choisi pour sa performance sur les données tabulaires et sa capacité à capturer des **relations non linéaires complexes** entre variables temporelles, opérationnelles et contextuelles. Contrairement aux modèles purement statistiques, XGBoost permet de modéliser des interactions telles que la combinaison d'un effet calendaire (hiver, jour de semaine) avec des contraintes terrain (effectif présent, lits disponibles, pannes matérielles). L'entraînement est réalisé selon un **split temporel strict** (80 % passé / 20 % futur), garantissant l'absence de fuite d'information et reproduisant un cadre réel de prévision opérationnelle.

Les performances du modèle sont évaluées à l'aide de la **MAE** et de la **RMSE**, indicateurs adaptés à une problématique métier où l'erreur doit rester interprétable en nombre de patients. Une erreur moyenne d'environ ±14 patients par jour représente une marge inférieure à 5 % du flux moyen, ce qui est considéré comme robuste pour un usage décisionnel hospitalier. La **validation croisée temporelle** a permis de confirmer la stabilité des performances sur différentes périodes, limitant les risques de sur-apprentissage liés aux pics exceptionnels.



L'interprétabilité du modèle est assurée par l'utilisation des **valeurs SHAP**, qui permettent d'identifier précisément les facteurs déclencheurs des pics d'activité. Cette analyse confirme l'intérêt de l'approche hybride : les variables calendaires expliquent la tendance globale, tandis que les variables terrain (lits, personnel, inertie du jour précédent) jouent un rôle clé dans l'intensité des tensions. Les prédictions sont ensuite traduites en **niveaux d'alerte opérationnels** (Normal, Pré-alerte, Alerte, Critique) à partir de seuils quantiles calculés sur l'historique, transformant ainsi un

modèle de machine learning en **véritable système d'aide à la décision**, directement exploitable via le dashboard.



Prédiction sur 2026

La phase de prospective vise à **simuler l'activité hospitalière future** en l'absence de données réelles disponibles pour 2026. Pour répondre à cette contrainte, une approche de **scénario standard** a été mise en place. Les variables déterministes liées au calendrier (jour de la semaine, week-end, mois, période hivernale) sont générées à partir des dates réelles de janvier et février 2026. Les ressources hospitalières (effectif présent, taux d'absentéisme, nombre de lits disponibles) sont fixées à leur **moyenne historique**, ce qui correspond à une hypothèse de fonctionnement nominal sans crise majeure. Cette hypothèse sert de point de référence et peut être facilement modifiée pour simuler des scénarios dégradés (grève, épidémie, pénurie de lits).

La prédiction repose ensuite sur une **boucle récursive jour par jour**, intégrant une variable d'inertie (patients_hier). La prédiction du flux d'un jour donné est réinjectée comme entrée pour le jour suivant, reproduisant ainsi la dynamique réelle des systèmes hospitaliers où l'activité passée influence directement l'activité future. Cette approche permet de projeter des trajectoires cohérentes dans le temps, mais introduit volontairement une accumulation de l'incertitude, ce qui est réaliste dans un contexte opérationnel. Les flux prédits sont enfin traduits en **niveaux d'alerte opérationnels** à l'aide des seuils définis sur l'historique, permettant d'identifier dès la phase prospective des journées à risque élevé et d'anticiper les actions de pilotage nécessaires.

Ainsi, les prévisions pour le début de l'année 2026 indiquent **une tension hospitalière persistante** dès les premiers jours de janvier, avec des flux journaliers majoritairement situés en zone de **pré-alerte**, traduisant un niveau d'activité supérieur à la normale dans un contexte hivernal. Le 5 janvier se distingue par un dépassement du seuil critique, avec près de 391 patients prédits, suggérant un risque élevé de saturation des urgences et des services d'aval si aucune mesure anticipative n'est prise. À l'inverse, le 4 janvier apparaît comme une journée de respiration relative, probablement liée à un effet week-end, illustrant l'impact combiné des facteurs calendaires et de l'inertie des flux. Ces résultats confirment l'intérêt du modèle pour identifier en amont les journées à risque et ajuster les ressources avant l'apparition effective des tensions.

Dashboard

Architecture et interface utilisateur.

On a implémenté un tableau de bord interactif avec Streamlit, configuré en mode large pour un usage décisionnel. L'interface est enrichie par un CSS personnalisé injecté directement dans l'application afin d'améliorer la lisibilité des KPI (fonds, couleurs, hiérarchisation visuelle) tout en restant compatible avec les modes clair et sombre. La structure repose sur une barre latérale de pilotage (choix de période et de granularité temporelle) et un corps principal organisé en indicateurs synthétiques, graphiques temporels et analyses détaillées par service, garantissant une navigation fluide pour des utilisateurs non techniques.

Logique métier et génération des données.

Les données sont entièrement simulées sur la période 2018–2026 via une fonction mise en cache (@st.cache_data) pour optimiser les performances. Le flux journalier est construit à partir d'une combinaison de tendance longue (croissance annuelle), saisonnalité cyclique (fonction cosinus), choc exogène simulé (COVID 2020–2021) et bruit aléatoire afin de reproduire un comportement réaliste des admissions hospitalières. Une logique de seuils métiers transforme ensuite le flux prédit en niveaux d'alerte (Normal, Pré-alerte, Alerte, Critique). Une seconde fonction traduit ce flux global en impacts opérationnels par service (lits ouverts, fermés pour raisons RH, taux d'occupation, CCMU moyen), intégrant explicitement l'effet de l'absentéisme du personnel sur la capacité réelle.

Calculs dynamiques, visualisation et aide à la décision.

Selon le mode sélectionné (journalier, hebdomadaire ou mensuel), l'application adapte automatiquement les calculs : valeurs exactes au jour le jour ou moyennes quotidiennes sur une période, avec une règle prudente retenant le niveau d'alerte maximal observé. Les indicateurs clés (flux, taux d'occupation, lits disponibles) sont recalculés en temps réel et présentés sous forme de KPI. Les visualisations Plotly permettent de distinguer historique et prédiction, d'annoter la date courante et d'afficher des seuils critiques. Enfin, des onglets analytiques offrent une lecture approfondie de la capacité, de la gravité clinique et des tableaux synthétiques, transformant le tableau de bord en outil opérationnel de pilotage et d'anticipation des tensions hospitalières.