UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

Fouille exploratoires des données sur les temps de séjour dans un hôpital

Koussaila HAMMOUCHE
Kamilia LOUALIA
Rachida OUCHENE
Mohamed-Amokrane SELMI
Lylia TOUAZI

Mai 2020

Encadrant : Jean-Michel Fourneau

Année universitaire 2019-2020

Table des matières

1 Introduction

L'analyse de données, connue en anglais comme Data Analysis, et quelques fois en Data Mining au vu de la différence quelques fois confuse, sert à inspecter,nettoyer, trasnformer et modéliser des données dans le but d'extraire des informations pertinentes. Elle est devenue un outil indispensable pour les entreprises et entités manipulant de grandes quantités de données.
L'analyse exploratoire de données a pour objectif d'avoir une vision globale des données et d'obtenir des informations non évidentes ou suspectées grâce à des technologies informatiques (intelligentes ou non) et des méthodes statistiques. [1]
Le personnel en charge de l'exploration a pour rôle de présenter les informations extraites sous forme de structure compréhensibles qu'utiliront ensuite les décisionnaires de l'entreprise pour élaborer les stratégies futures, ou limiter les risques d'erreurs.
Dans cette optique, nous avons été chargé d'effectuer une fouille exploratoire sur les données de l'hôpital de Rambam, situé à Haïfa, en Israël.
Il nous a été demandé d'analyser les phénomènes d'arrivées, de ré-entrance ainsi que les délais des divers workflows de l'hôpital afin de permettre une meilleure affectation de l'hôpital dans ses différents services et départements. Pour se faire, nous avons eu à notre disposition les données de l'hôpital de 2004 à 2007 sur les temps d'arrivées et de séjour de chaque patient, dans les différents services.
L'hôpital étant découpé en trois parties que sont :

- Urgences
- Traitements prévus à l'avance
- Imagerie médicale à rayons X

Après une analyse générale et expérimentales sur les données. Nous avons travaillé sur les services séparément, en priorisant les urgences comme demandé.
L'analyse effectuée est divisée en deux axes principaux :

1. Analyse des arrivées des patients.
2. Analyse des temps d'attente des patients.

Grâce au langage R et aux fonctionnalités qu'il offre pour le traitements des données massives, nous avons pu accomplir les tâches définies sur les données de l'hôpital, nous sommes ensuite arrivés à des déductions suite à la lecture des structures de données, déductions limitées à ce que nos prérogatives.
Le résultat de notre analyse fournit aux décisionnaires et à l'administration de l'hôpital un rapport avec des structures graphiques expliquées et des statistiques compréhensibles. Ces derniers pourront ensuite utiliser notre travail pour définir des changements afin d'améliorer le fonctionnement de l'hôpital.

2 Qu'est ce que l'analyse exploratoire de données ?

2.1 Définition

Exploratory Data Analysis (EDA), en français Analyse Exploratoire de Données (AED) est une approche de la Data Science qui consiste à traiter des ensembles importants de données et de dégager les aspects les plus intéressants de la structure de ceux-ci. L'AED est en effet une branche du Data Analytics qui se concentre tout d'abord sur la découverte de régularité dans les données tels que des dépendances, redondance, groupes homogènes ou corrélation. L'analyste de données doit ensuite modéliser les résultats obtenus pour en tirer des assertions voir des modèles prédictifs qui seront utilisés par l'entreprise ou l'entité en vue d'augmenter le chiffre d'affaire ou d'améliorer la productivité et l'efficacité des tâches planifiées. [2]
L'analyse exploratoire de données utilise notamment des représentations graphiques diverses qui permettent de mieux entrevoir les relations entre les variables, notamment dans ce type d'analyse de données où il n'existe pas d'hypothèses de départ.

2.2 Historique

Bien que l'analyse de données n'est intervenu que récemment dans les systèmes décisionnelles des grandes entreprises, elle est le fruits de plusieurs siècles de développement. La collecte de données remonte à l'antiquité où en Égypte le pharaon Amasis organise le recensement de sa population au ve siècle av. J.-C. [3]
Au cours du 19ème siècle, grâce à la création et évolution des lois statistiques que nous connaissons aujourd'hui (loi binomiale, théorème de Bayes, analyse de la variance, segmentation...etc), les notions requises pour une analyse des données modernes commencent à être maîtrisées. C'est ainsi que Adolphe Quetelet, astronome, statisticien belge, exploite ce qu'il connaît de la loi gaussienne à l'anthropométrie pour examiner la dispersion autour de la moyenne (la variance) des mesures des tailles d'un groupe d'hommes. [4]
Mais c'est John Tukey, mathématicien américain auteur de plusieurs publications sur l'analyse de données et les statistiques, qui encouragea les statisticiens à travailler sur les données calculées et à en extraire des hypothèses ou des assertions. Tukey définissait notamment l'analyse de données comme des procédures d'analyse de données, techniques d'interprétations des résultats de ces procédures grâce à des méthodes de rassemblement, collecte et présentation des données qui permettent de mieux les analyser.[5]
Sa publication du livre "Exploratory Data Analysis" en 1977 posa les bases de la fouille exploratoire de données en suggérant que l'analyse devait, en plus de confirmer des hypothèses, en créer elle même. Et ce en se basant sur les différentes techniques et outils statistiques [6].
L'avènement de l'informatique à la fin du siècle dernier, en plus de l'assurance de Tukey sur l'efficacité de l'EDA, poussa au développement de solutions informatiques qui pourraient aider les statisticiens dans leurs analyses. C'est ainsi que les langages de programmation S-Plus et R sont apparus, qui permettent une meilleure gestion des données et la visualisation de celles ci via des graphiques interactifs. L'augmentation de la puissance de calcul, et le

développement général de l'informatique actuel, permet en plus de mieux gérer des données de plus en plus massives menées par le Big Data.

2.3 Outils et techniques

En plus des outils et lois statistiques déjà connus et utilisés depuis le 20ème siècle. L'informatique propulsa l'analyse de données vers de nouvelles possibilités jusqu'alors impossible. Le langage R est actuellement l'outil le plus utilisé. Créé en 1993, il permet de traiter de gros volumes de données avec rapidité et efficacité, et fournir diverses solutions de représentation, en plus de fonctions statistiques pour quantifier les différentes variables. L'EDA utilise notamment des représentations graphiques : Histogrammes, Diagrammes de Pareto, Nuages de points, Coordonnées parallèles. Ces représentations visuelles, pour certaines interactives, offrent de meilleures possibilités aux analystes pour trouver des relations, corrélations et dépendances entre les variables, certaines insoupçonnées. [2]

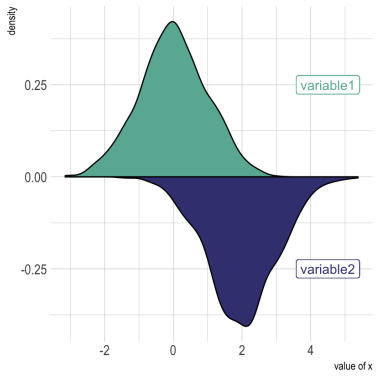

FIGURE 1 – Exemple de graphe généré par le langage R

3 Récupération et nettoyage des données

3.1 Récupération et transfert des données

Les données de l'hôpital nous ont été transmises via Microsoft Access accompagnée d'un document explicatif, elles ont été organisé par mois, chacun ayant une base de données propre, toutes les bases de données contenant les tables suivantes : [7]

- Visit : Cette table contient les informations générales sur chaque patient qui entre à l'hôpital. Chaque enregistrement représente une visite unique avec l'arrivée et la sortie de l'hôpital, un patient peut donc se retrouver plusieurs fois dans cette table.
- Visit details : Cette table représente le passage d'un patient dans un département précis, avec le moment précis de son admission et de sa sortie de ce département.
- Physical details : Cette table contient, en plus des informations contenus dans Visit details, contient des informations sur la localisation physique du patient (unité, pavillon).
- Ward First Procedure : Cette table contient les informations sur les patients se présentant à l'hôpital sur rendez-vous, elle possède notamment des informations temporelles des procédures sur les patients.
- X-Rays visits : Elle contient les informations sur les patients externes se présentant pour une imagerie médicale aux rayons X.

Nous avons transféré les données vers le SGBD MySQL, le langage R contenant un package traitant les bases de données MySQL : RMySQL [8].
Ce package profite d'une documentation riche du fait de sa popularité au sein de la communauté utilisant R.
Nous avons décidé de fusionner les bases de données transmises en une base de données unique contenant les tables mentionnées ci-dessus, R pouvant de ce fait mieux travailler sur les séries temporelles.

3.2 Nettoyage des données

Le fichier de données transmis inclut des tables non mentionnées dans le document explicatif, ses tables n'ont pas été gardés lors du transfert des données.
Les données contiennent également des erreurs (valeurs manquantes ou incorrectes) sur les attributs suivants [7] :

- Sexe : Valeurs manquantes, marqué par "Unknown".
- Date de naissance : Valeur absente, ce qui a pour conséquence d'avoir l'attribut âge à 99.

- Date de sortie : Valeur manquante ayant pour conséquence une durée de séjour négative.
- Incohérence entre date et heure d'arrivée à l'hôpital et celle de début de traitement, avec pour conséquence un temps d'attente dépassant des jours entiers, voir années.

Nous avons nettoyé la base de données en ignorant les valeurs manquantes ou incohérentes, notamment lors de nos analyses utilisant les attributs cités ci-dessus.
Par exemple, nous avons ignoré l'attente dont la durée dépasse 24h lors de nos requêtes SQL. Il n'est pas nécessaire de supprimer les enregistrements car ils contiennent d'autres données utiles pour d'autres recherches.

4 Analyse Génerale

4.1 Répartition dans le département d'imagerie médical

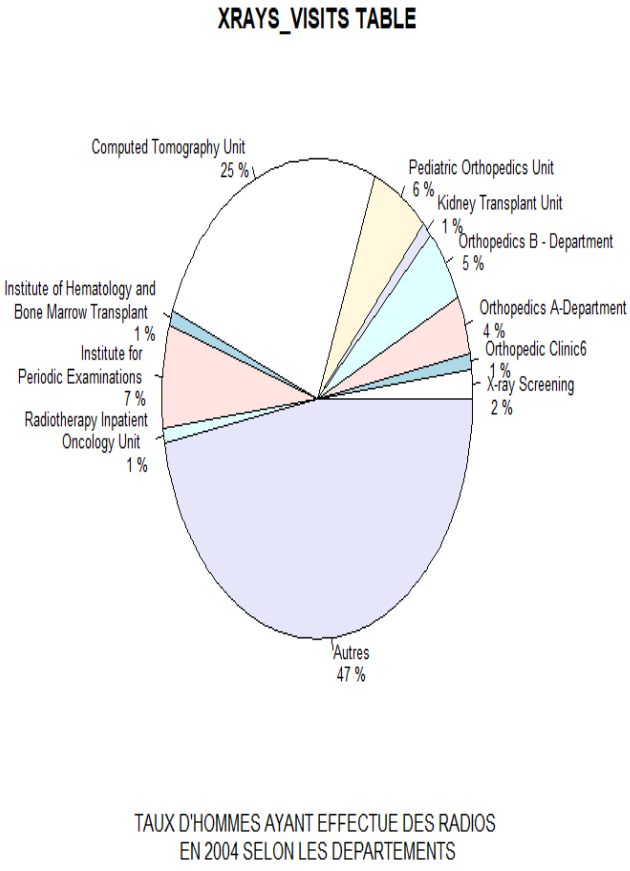

FIGURE 2 – Répartition des patients masculins dans les unités d'imagerie médicale.

Ce diagramme en secteur permet d'observer que le département de tomodensitométrie (scanner) concentre à lui seul 25% du flux de patients. Les 75% restants sont répartis sur les autres unités de façon plus ou moins équitables (maximum 7%).
Les unités les plus fréquentées sont représentées dans le diagramme, les autres sont regroupées dans la partie "autres".

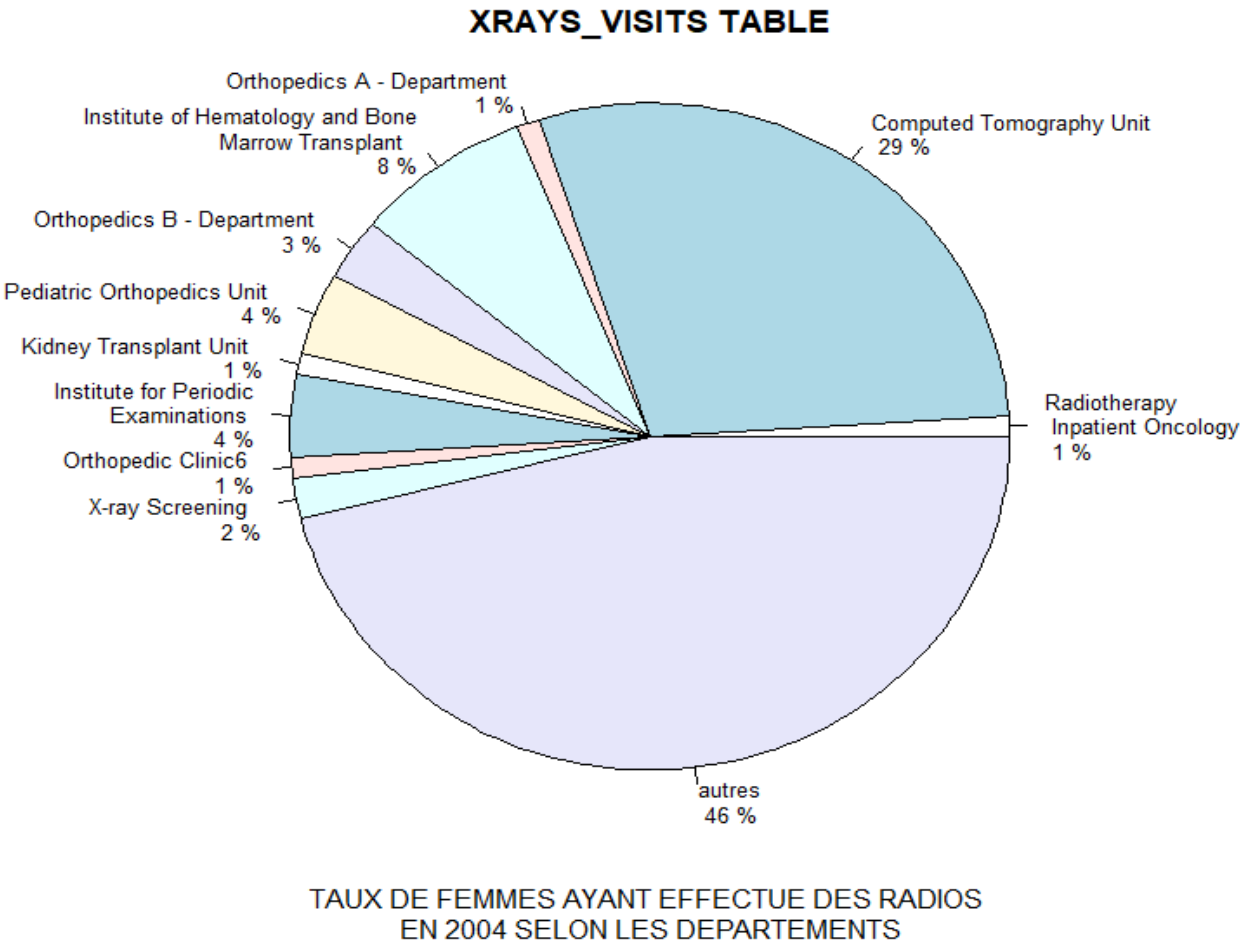

FIGURE 3 – Répartition des femmes dans les unités d'imagerie médicale.

On observe que le département tomodensitométrie est là aussi majoritaire (29%). La répartition est similaire à celle des patients du sexe masculin.

4.2 Répartition dans le département des urgences

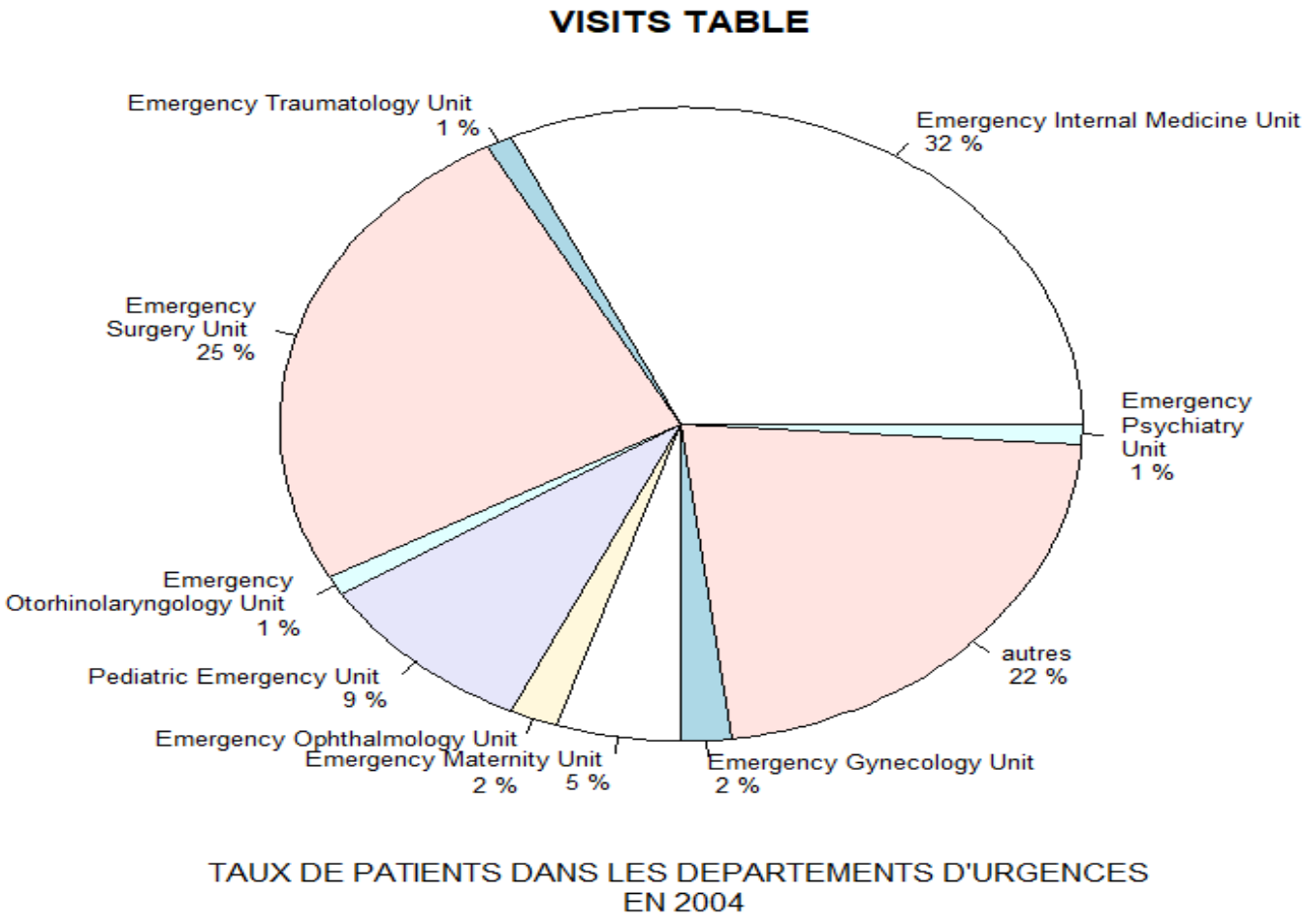

FIGURE 4 – Répartition des patients sur les départements d'urgences.

Ce graphe nous permet de voir la répartition des patients sur les différentes unités des urgences.
On observe que l'unité de médecine interne est la plus fréquentée, suivie de celui de celle de chirurgie. Ces deux unités concentrent à elles seules plus de la moitié des patients dans les urgences.

4.3 Répartition des patients par tranches d'âge

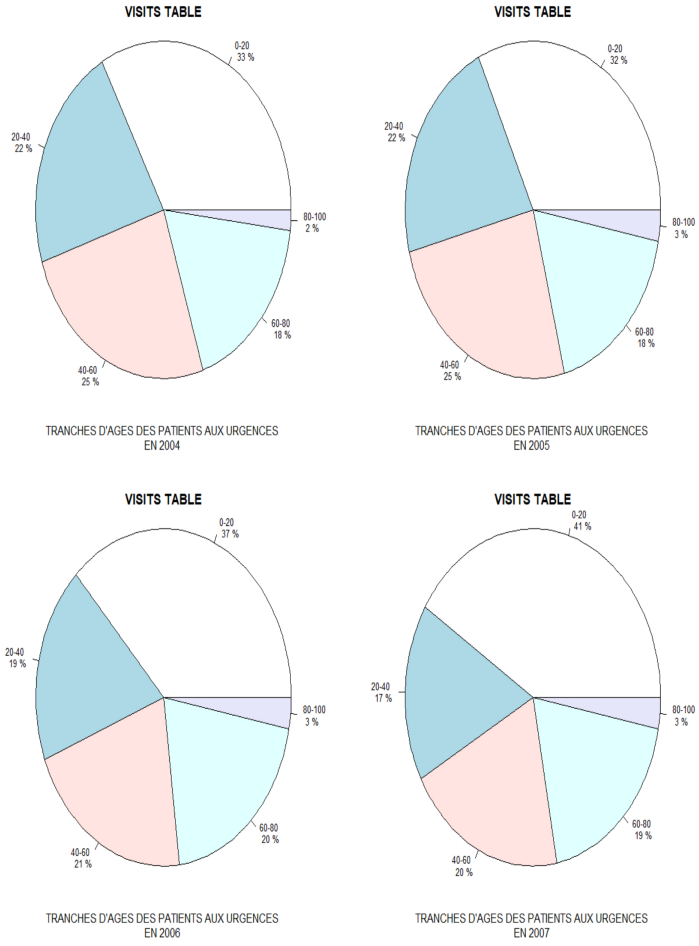

FIGURE 5 – Répartition des patients de l'hôpital par tranches d'âges.

Ces diagrammes camembert permettent de voir la répartition des malades par rapport à leurs âge pour avoir une idée globale sur les patients que reçoit l'hôpital.
On observe que la répartition est homogène.

FIGURE 6 – Répartition des patients dans les services d'imagerie médicale par tranches d'âges.

Ces diagrammes circulaires nous illustrent la répartition des patients par tranches d'âges dans les services de radiographies sur 4 années différentes.
On note que c'est les patients dont l'âge varie entre 40 et 60 ans qui sont le plus concernés par ces unités.

4.4 Répartition selon le sexe des patients.

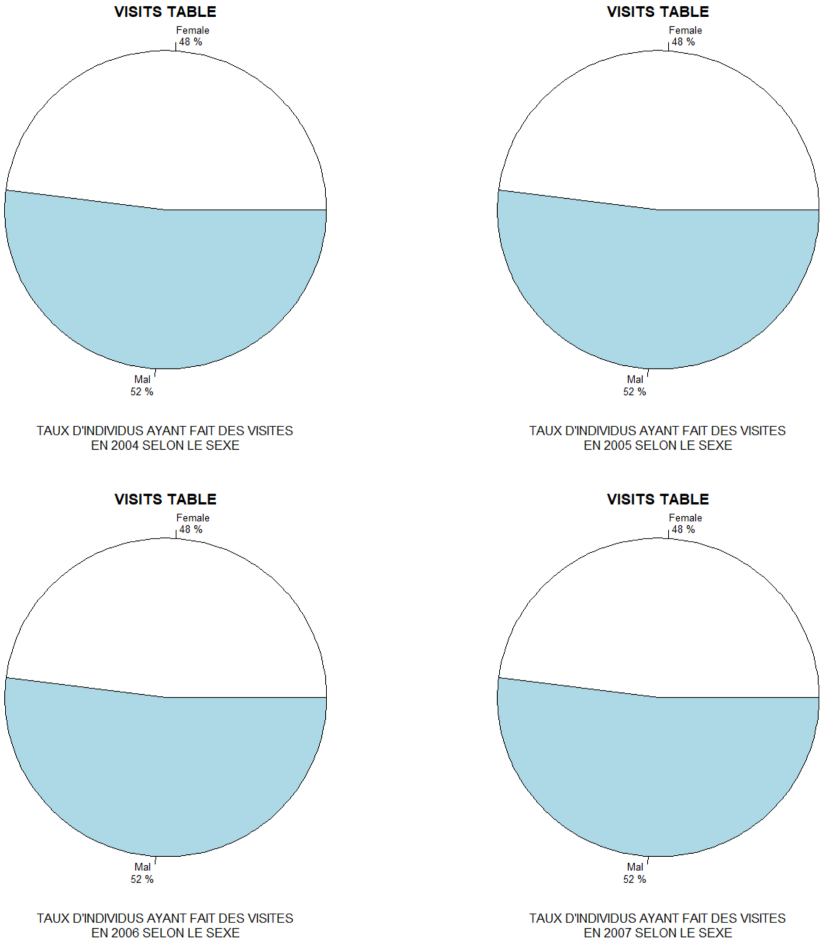

FIGURE 7 – Répartition selon le sexe dans l'ensemble de l'hôpital.

Ces diagrammes circulaires nous montrent la répartition des patients selon leurs sexe sur 4 années différentes au niveau des services d'urgences.
On remarque que c'est quasiment un équillibre parfait entre les deux sexes.

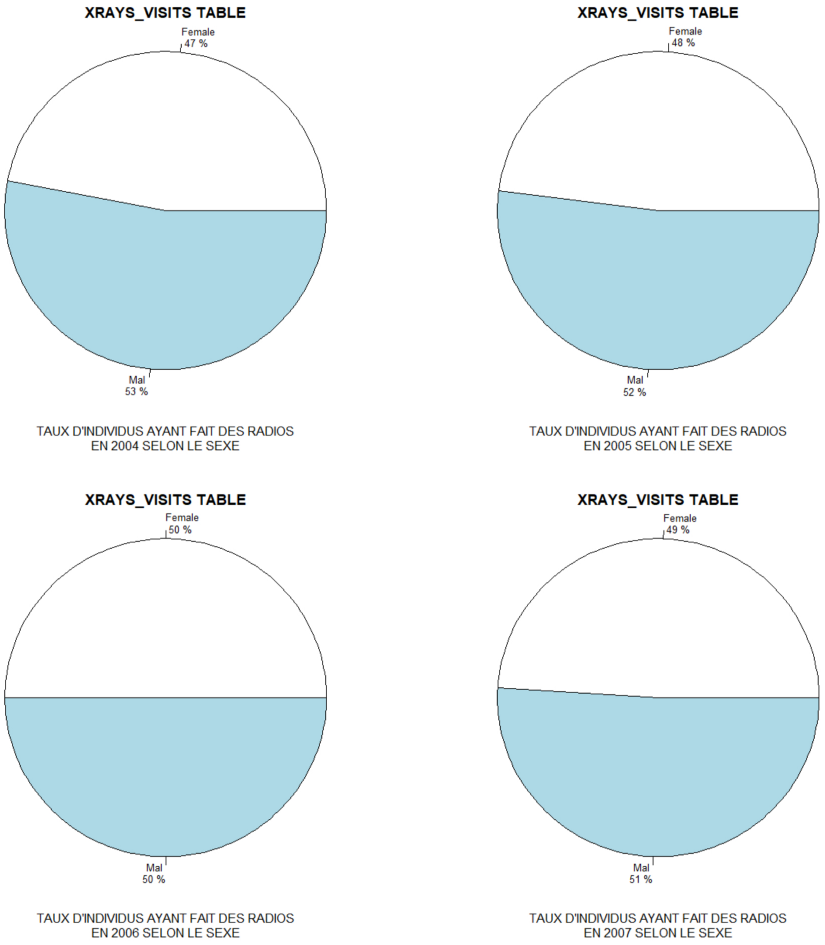

FIGURE 8 – Répartition selon le sexe sur les services de radiographies.

Ces diagrammes circulaires nous montrent la répartition des patients selon leurs sexe sur 4 années différentes au niveau des services de radiographies.
On constate que l'homogénéité reste présente au fil du temps.

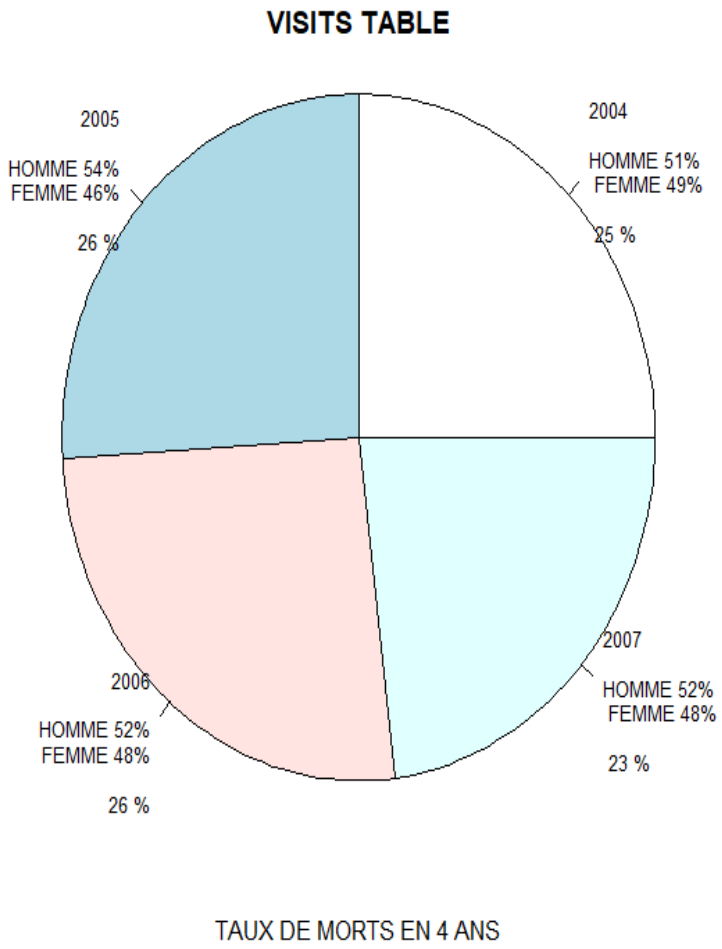

FIGURE 9 – Taux de mortalité.

Ce graphe nous montre le taux de mortalité au fil des 4 ans.
On observe que c'est équilibré entre les années et qu'il n'existe pas de sexe plus touché qu'un autre.

5 Analyse Temporelle du département des urgences

5.1 Analyse des arrivées.

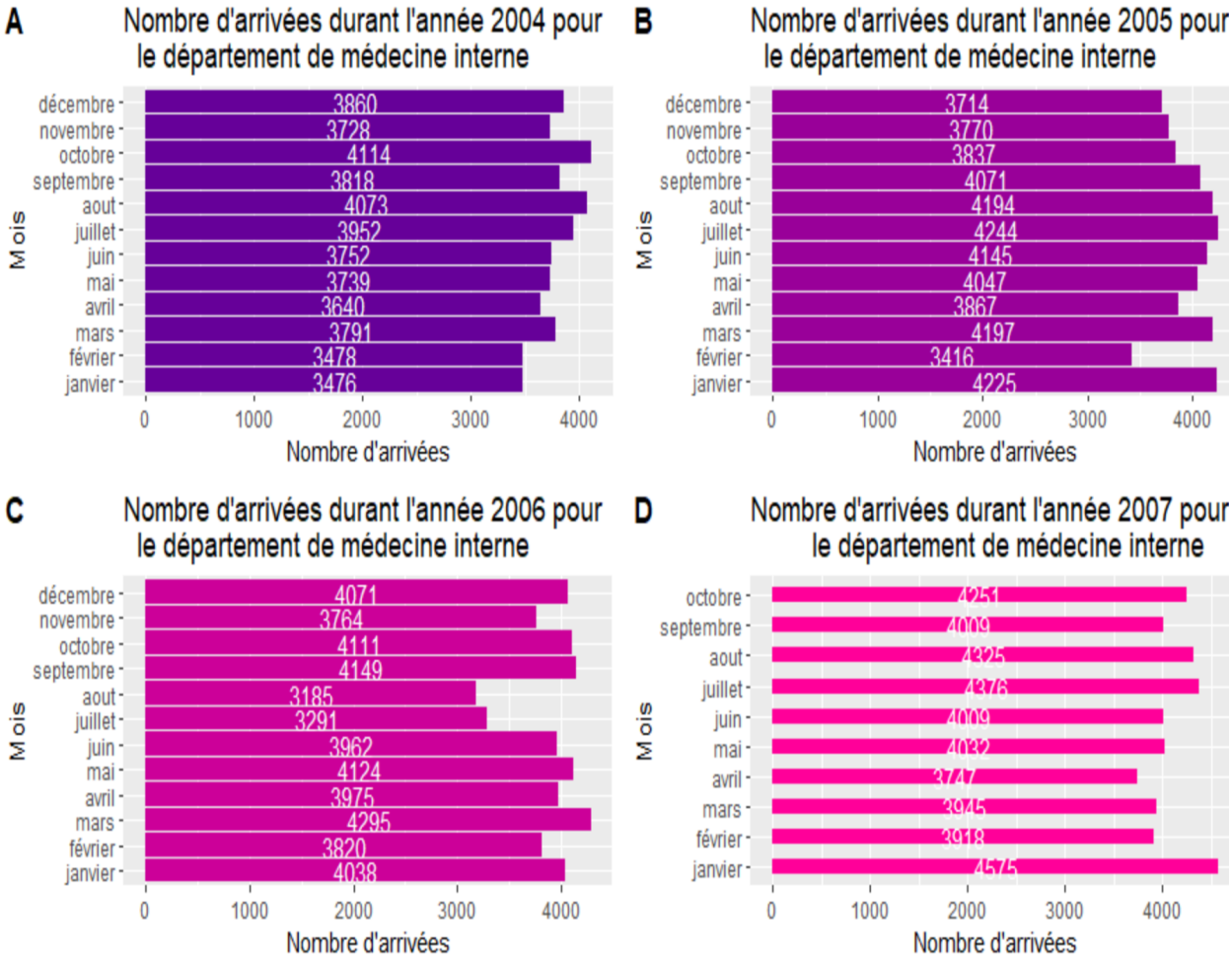

FIGURE 10 – Nombre d'arrivées par annnée à l'unité de médecine interne.

Utilisé pour voir l'évolution de la fréquentation de l'hôpital de 2004 à 2007, ce graphique permet de constater une augmentation modeste du nombre de patients. Avec une progression moyenne de 300 personnes par mois.

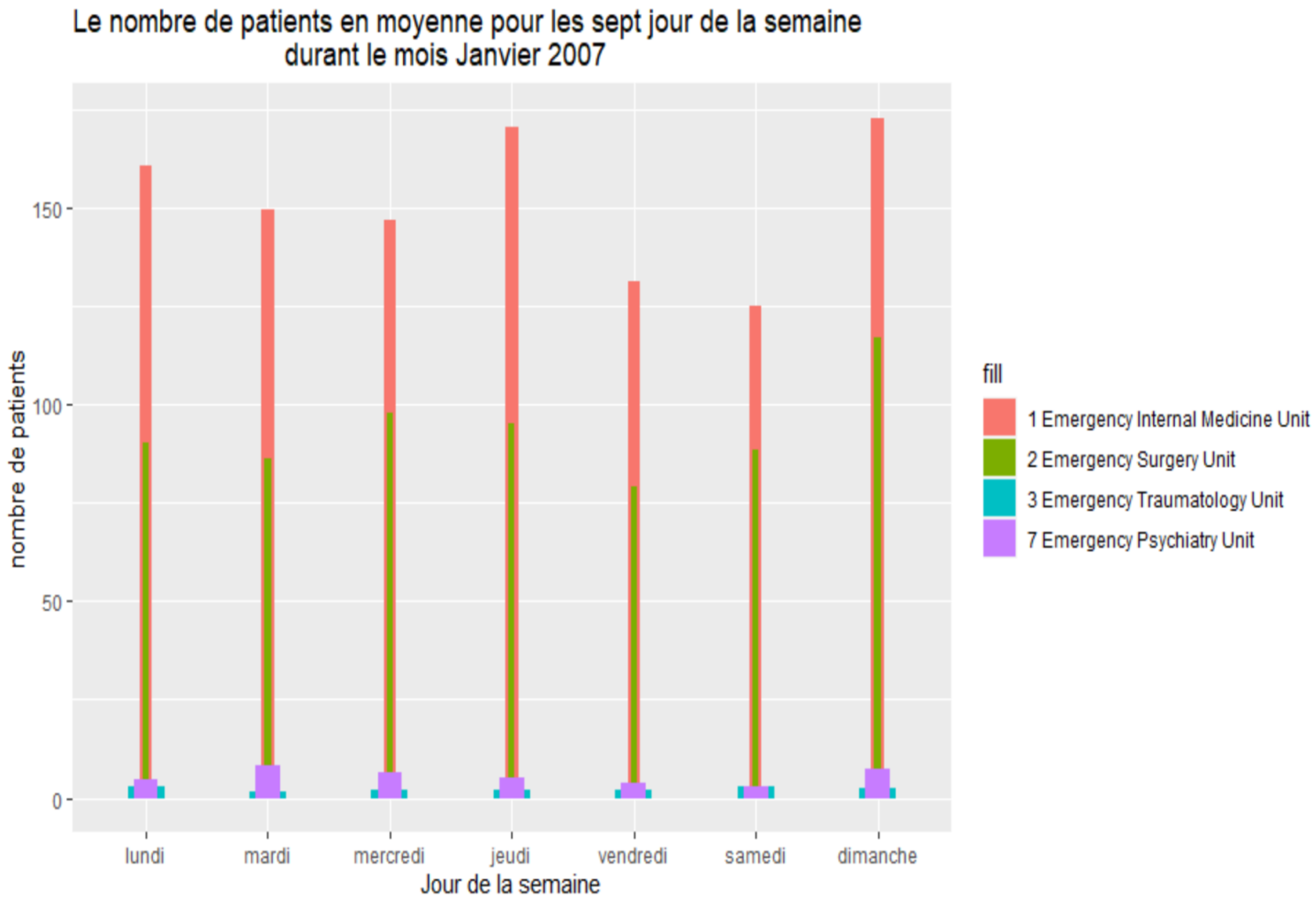

FIGURE 11 – Moyenne des arrivées des unités les plus sollicitées .

On observe via ce diagramme qu'au sein des urgences, les unités de médecine interne et chirurgie concentrent la majorité des patients admis aux urgences.
C'est donc les départements les plus sujets aux débordements.

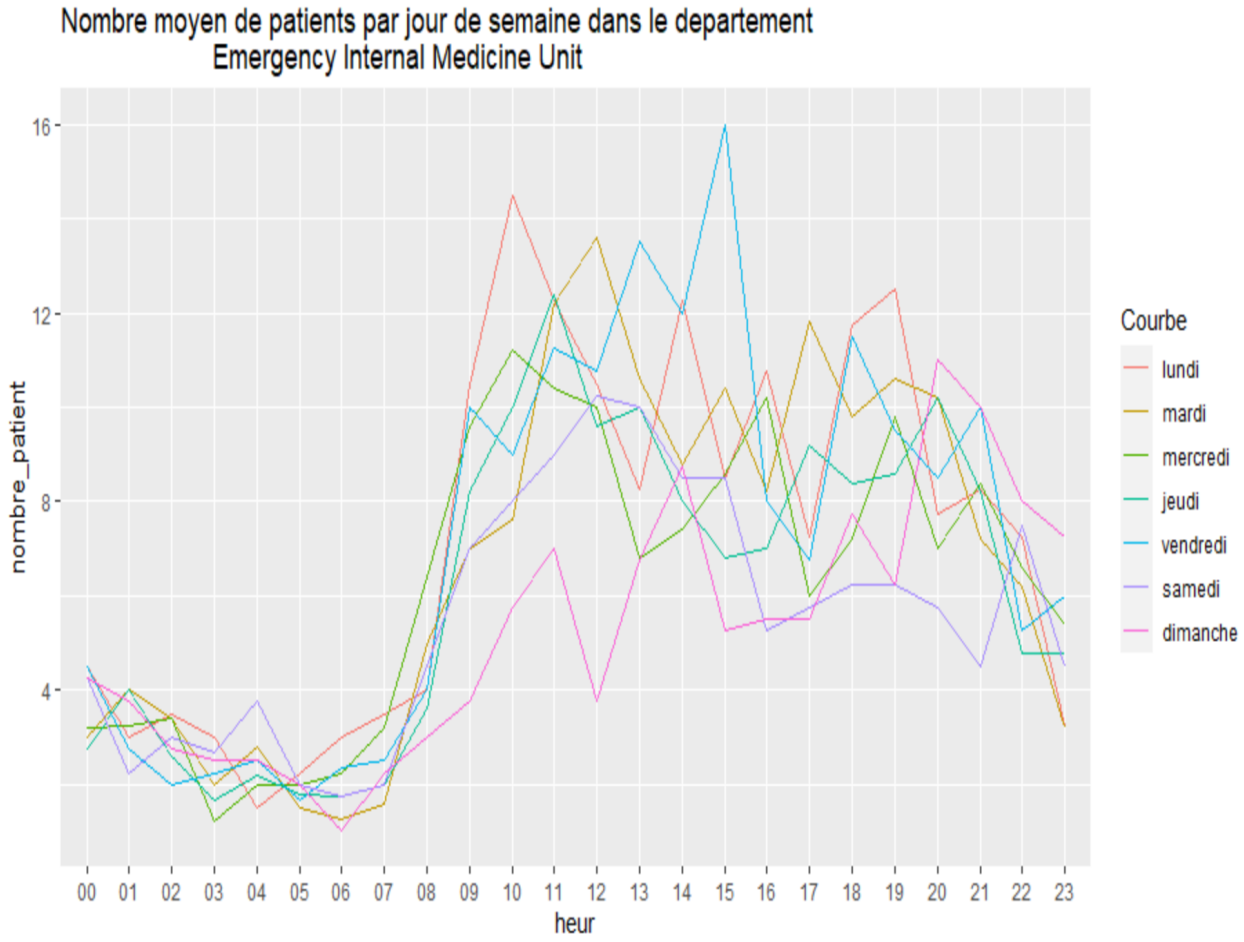

FIGURE 12 – Moyenne des arrivées par heure dans l'unité de médecine interne.

On observe que le flux des arrivées est très proche tout au long de la semaine, avec un pic général entre 10 et 15h.

5.2 Analyse des temps d'attente

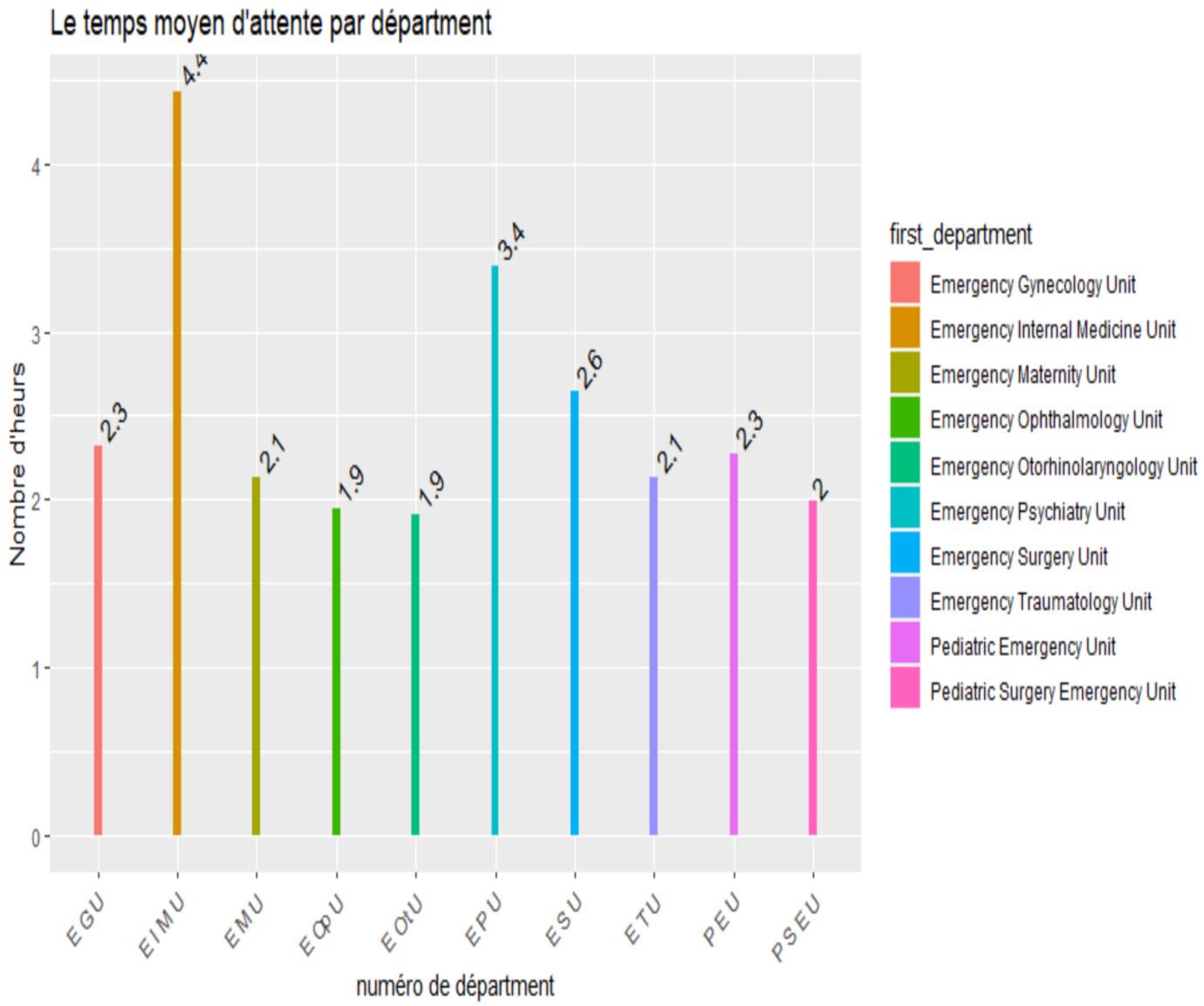

FIGURE 13 – Temps moyen d'attente par département.

Ce diagramme nous permet d'observer un temps d'attente moyen très élevé dans le l'unité de médecine interne qui est aussi celui qui concentre le plus de malades.
L'unité de chirurgie, malgré le 2ème flux de malades des urgences, a un temps moyen acceptable.
L'unité de psychiatrie, 3ème en nombre de malades a le 2ème temps d'attente moyen des urgences.

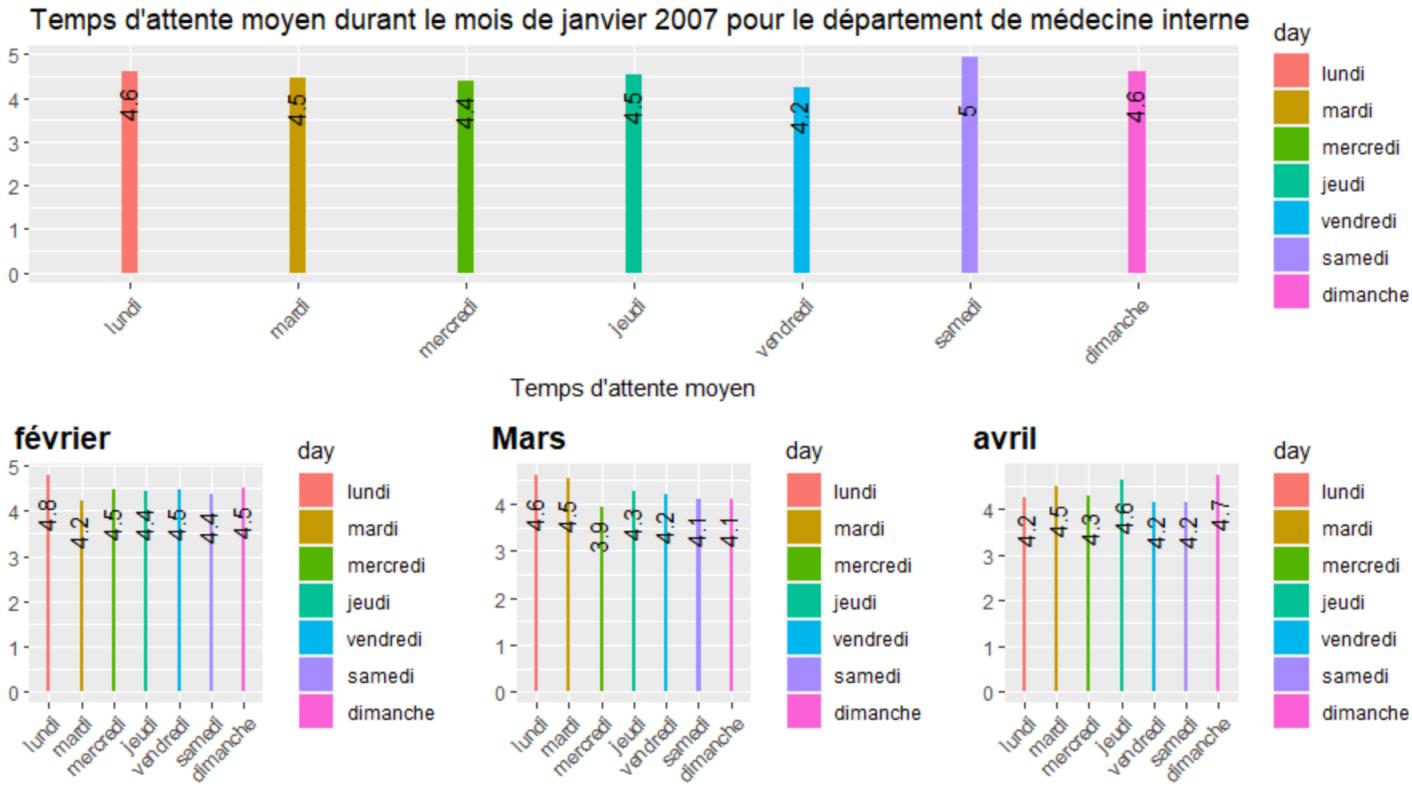

FIGURE 14 – Temps d'attente moyen dans l'unité de médecine interne de janvier à avril 2007.

Le temps d'attente est équillibré tout au long des jours de la semaine, bien qu'il existe une prelative augmentation dimanche et lundi, ce qui est plus compréhensible pour le 1er que le second, nous nous sommes donc penché sur les lundi de ce mois, qui seront l'objet de l'analyse suivante.

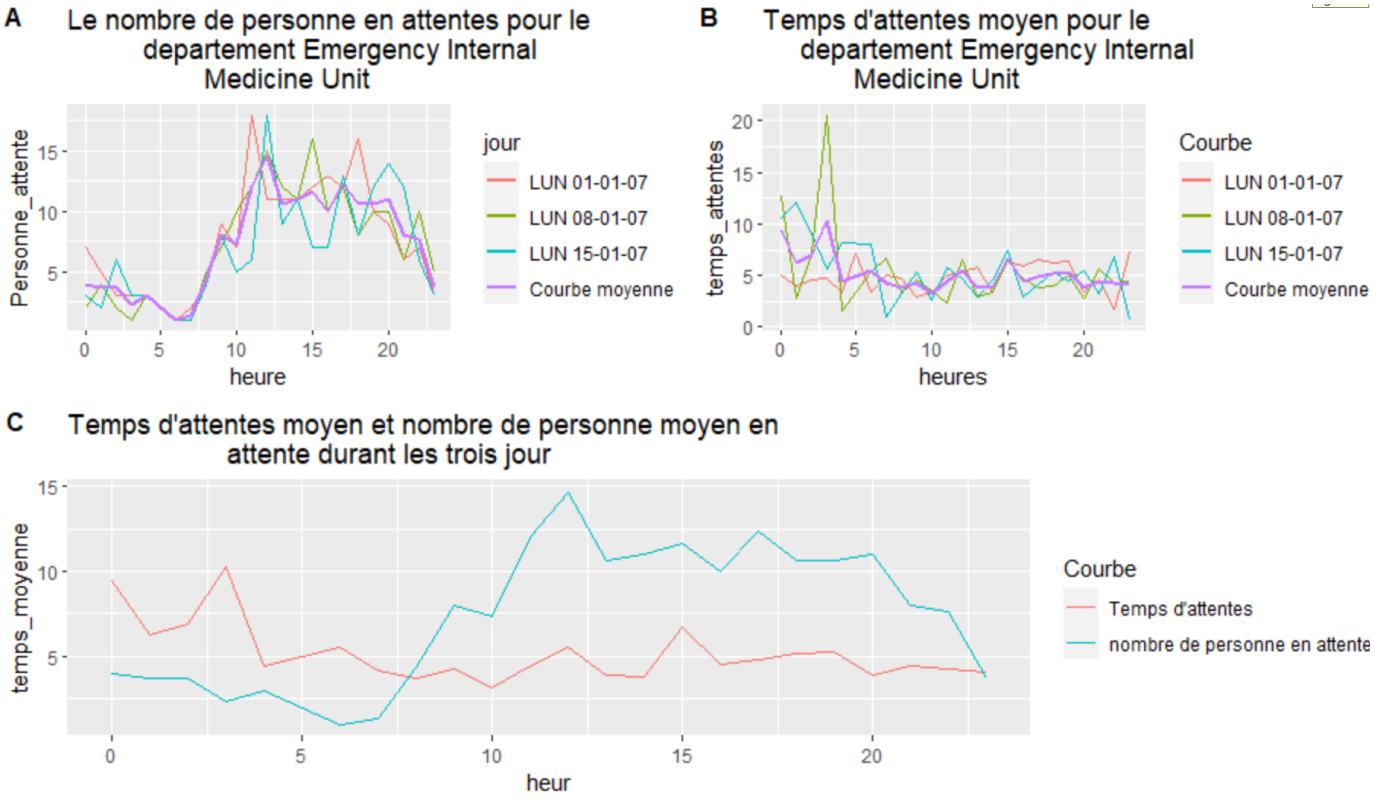

FIGURE 15 – Comparaison entre le temps d'attente moyen et le nombre de personnes en attente dans l'unité de médecine interne.

Dans un premier temps, nous remarquons que les courbes d'arrivées sont similaires entre les différents lundi.
D'un autre côté, l'histogramme de droite démontre que le temps d'attente est plus élevé la nuit, ce qui a déjà été observé globalement précédemment, l'attente peut vite devenir très élevée lors de journées chargées comme le lundi 08.
Le 3ème graphe nous permet d'observer que les temps d'attente diminuent pendant la journée, alors que le nombre de patients augmente.
On déduit donc que le temps d'attente des patients n'est pas forcément lié au nombre de patients.

6 Analyse détaillée

6.1 Courbes et total des arrivées

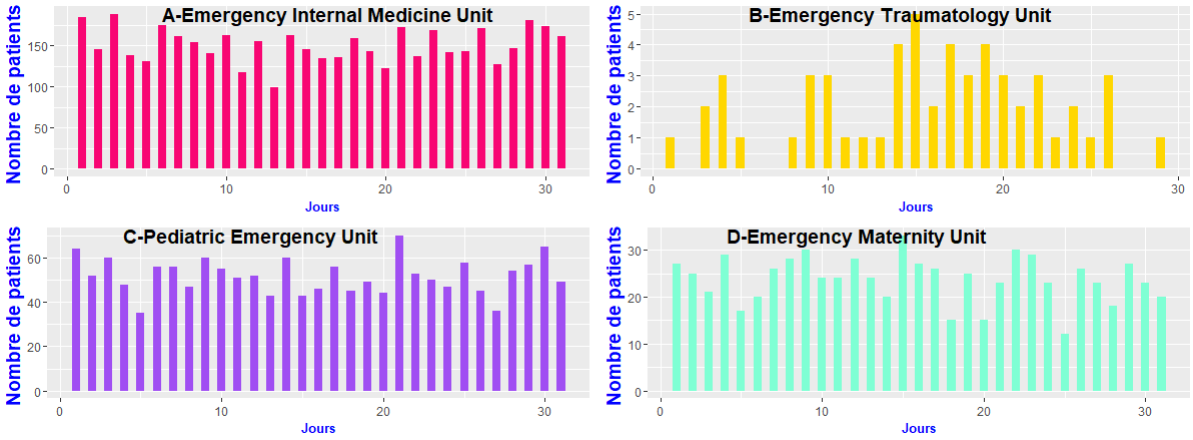

FIGURE 16 – Comparaison des arrivées durant le mois de janvier 2007.

On remarque que le flux des unités A C et D varie de façon similaire malgrè le nombre différent de malades qu'ils acceuillent,
Cependant, on observe que l'unité de traumatologie a une variation irrégulière au cours du mois, avec des journées vides et d'autres plus chargées (proportionnellement à la moyenne de patients de l'unité).

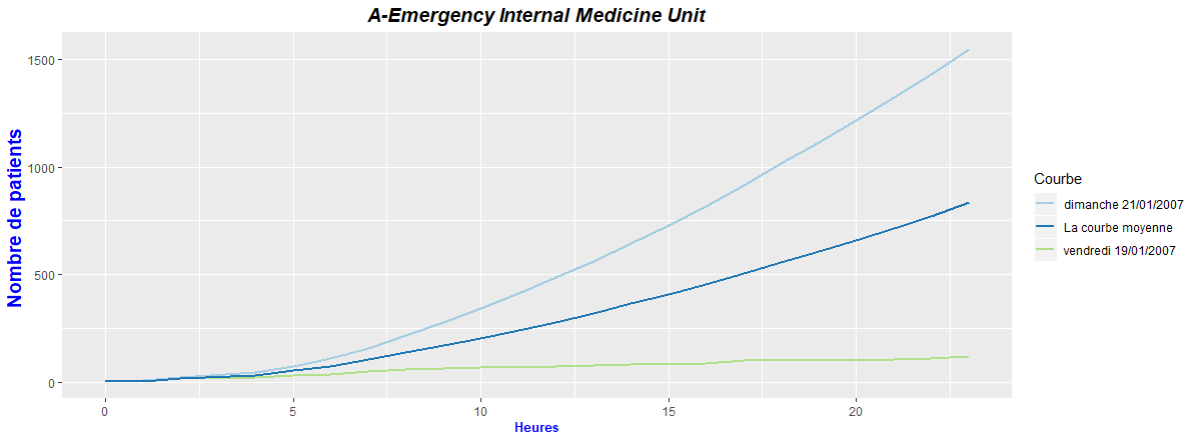

FIGURE 17 – Evolution du total des arrivées les 19/01/007 et 21/01/2007 à l'unité de médecine interne.

Nous remarquons une forte disparité entre le vendredi, premier jour du weekend en Israel, et le dimanche, premier jour de semaine.
On peut logiquement supposé que l'enregistrement informatique des patients du week-end a lieu le dimanche.
La courbe moyenne représenterait le nombre d'entrées plus fidélement pour ces deux jours consécutifs.

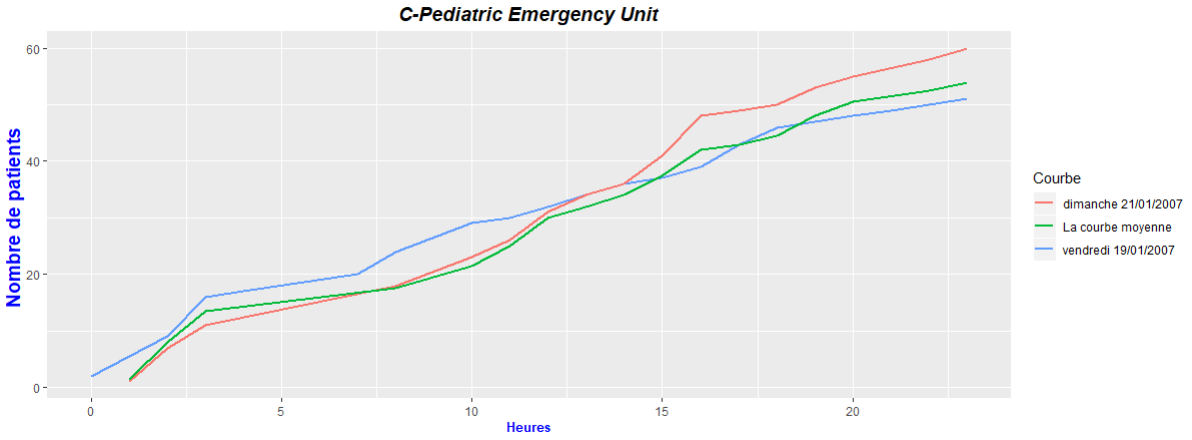

FIGURE 18 – Evolution du total des arrivées les 19/01/007 et 21/01/2007 à l'unité de pédiatrie des urgences.

Ici, le total d'arrivée augmente quasiment de la même façon le vendredi et dimanche. L'évolution est régulière et ne voit pas de soudaine augmentation.

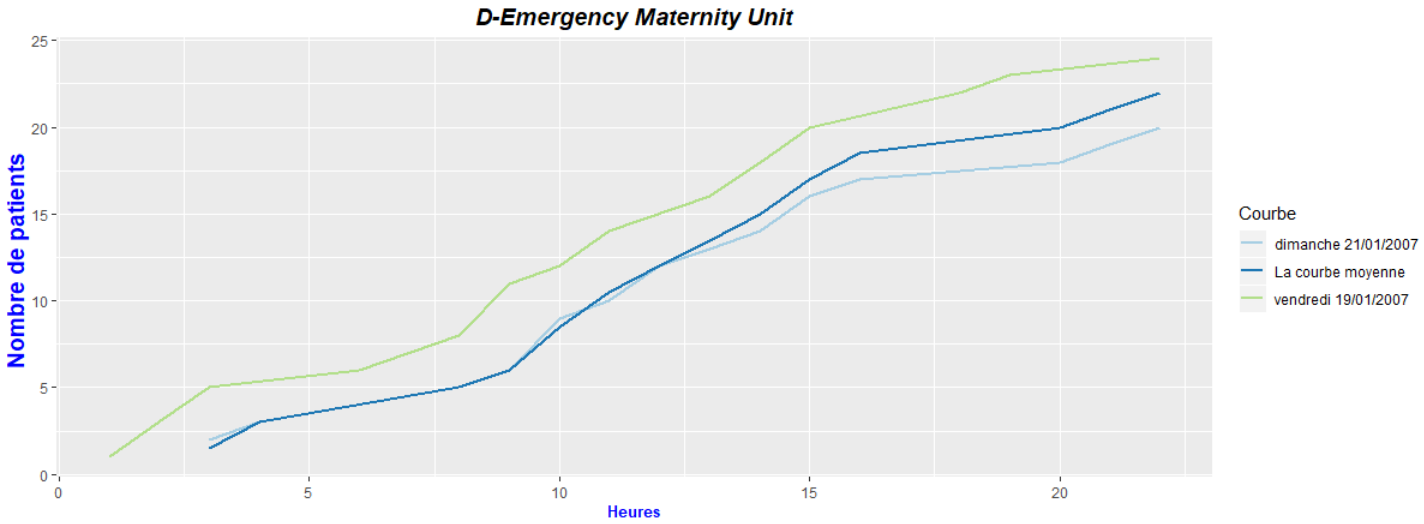

FIGURE 19 – Evolution du total des arrivées les 19/01/007 et 21/01/2007 à l'unité de matérnité des urgences.

Comme lors de la figure précédente, nous constatons une régularité entre les deux jours, et une augmentation stable (légèr crescendo entre 10 et 15h).

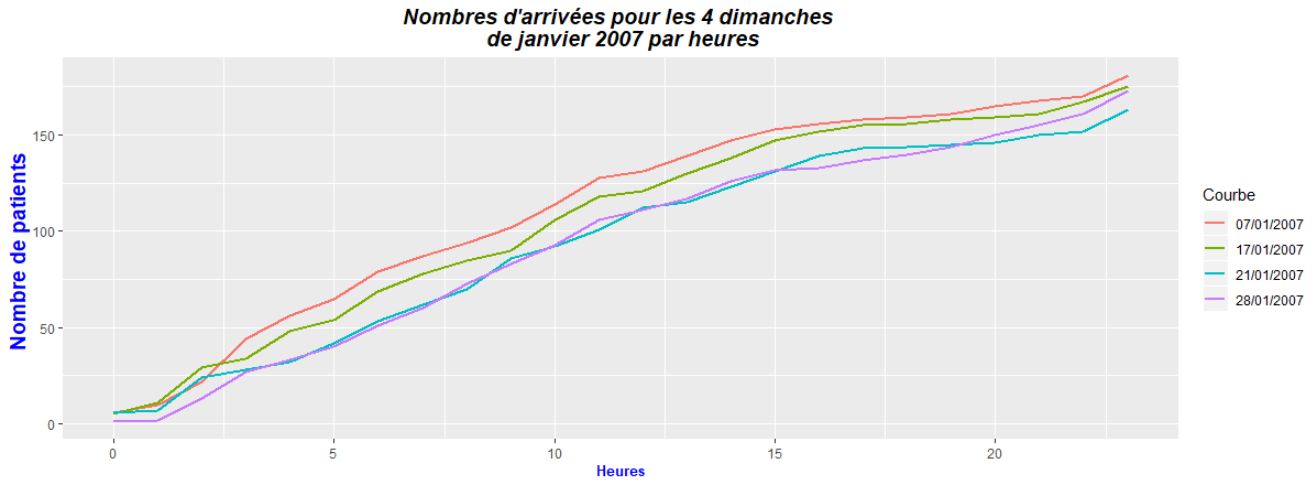

FIGURE 20 – Comparaison des courbes d'arrivées à l'unité de médecine interne entre les 4 dimanche de janvier 2007

Nous constatons que l'évolution temporelle du total des arrivées est similaire entre les 4 dimanche. On peut supposer que la fréquentation est donc régulière et que l'afflux des patients peut être anticipé.

6.2 Corrélations

La corrélation est une quantification de la relation linéaire r entre des variables continues. Le calcul du coefficient de corrélation de Pearson repose sur le calcul de la covariance entre deux variables continues. Le coefficient de corrélation est en fait la standardisation de la covariance. Cette standardisation permet d'obtenir une valeur qui variera toujours entre -1 et +1, peu importe l'échelle de mesure des variables mises en relation [9].
Ce coefficient de corrélation linéaire r peut être interpréter de la maniére suivante [10]. :

- Plus le coefficient est proche de 1, plus la relation linéaire positive entre les variables est forte.
- Plus le coefficient est proche de -1 plus la relation linéaire négative entre les variables est forte.
- Plus le coefficient est proche de 0, plus la relation linéaire entre les variables est faible.

Nous allons donc étudier quelques cas sur les données mises à notre disposition.

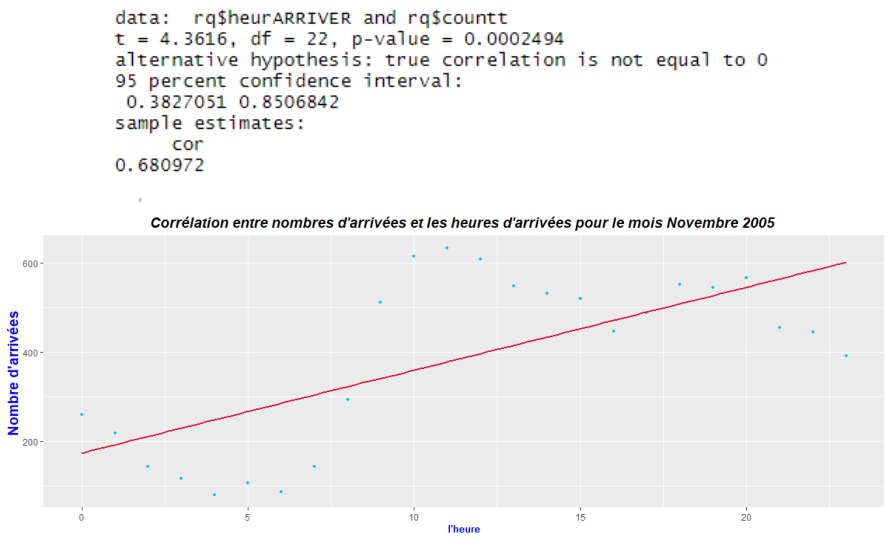

FIGURE 21 – Corrélation entre le nombre d'arrivées et l'heure d'arrivée.

Sur cette figure nous cherchons à trouver une corrélation entre l'heure d'arrivée et le nombre de patients qui arrive à l'hôpital.
Si nous regardons les résultats obtenus grêce a la fonction *"cor.stat"*. nous observons que la corrélation de Pearson est dans l'intervalle **[0.3827051 , 0.8506842]** et que p-Valeur est inférieure (<) au seuil de signification 0.05, le coefficient de corrélation *"cor"* est proche de 1, donc nous pouvons constater que la relation entre ces variables est positive, ce qui indique que lorsque nous avançons dans la journée (heure) le nombre de patients qui arrive augmente. Comme nous pouvons le voir également avec le graphe qui l'accompagne ci-dessus.

```
data:  rq0$jour and rq0$`COUNT(*)`
t = -2.4241, df = 29, p-value = 0.02181
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.66770216 -0.06569366
sample estimates:
       cor
-0.4104787
```

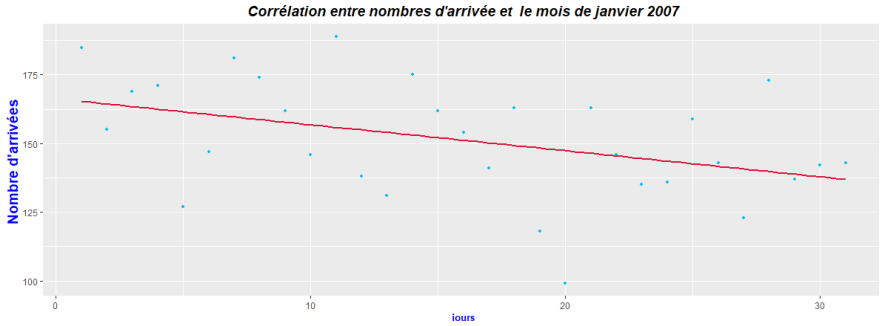

FIGURE 22 – Corrélation entre le nombre d'arrivées et le jour d'arrivée pour le mois de janvier 2007.

Dans cette cette partie nous prenons le nombre d'arrivée avec le jour d'arrivée, en se limitant au mois de janvier 2007.
Les resultats du calcul de *"cor.stat"* nous permettent d'observer que la corrélation de Pearson est dans l'intervalle négative et que p-Valeur<0.05. Le coefficient de corrélation *"cor"* est proche de -1, donc nous pouvons constater que la relation entre ces variables est négative, ce qui indique que lorsque nous avançons (par jour) vers la fin de mois, le nombre de patients qui arrive diminue. Comme nous pouvons le voir encore une fois avec le graphe ci-dessus.

```
data:  rq0$age and rq0$`COUNT(*)`
t = -2.5151, df = 83, p-value = 0.01383
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4535129 -0.0561711
sample estimates:
       cor
-0.2661095
```

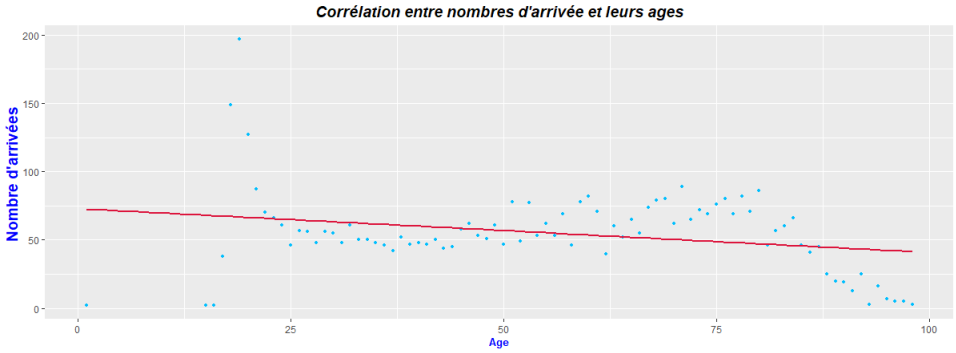

FIGURE 23 – Corrélation entre le nombre d'arrivées et l'âge des patients.

Le nuage de points ci-dessus représente la distribution des patients selon les arrivées à l'unité de médecine interne des urgences et l'âge respectif de ces arrivées pendant le mois de janvier 2007 . Tel que chaque point a pour abscisse x l'age des patients et pour ordonnée y le nombre de patients admis à l'unité. Nous constatons que la corrélation est négative car age et nombre de patients varient dans un sens contraire. En effet, lorsque l'âge augmente, les valeurs de y diminuent. Le resultat du calcul de "cor.test" qui accompagne le graphe confirme nos observations.
On constate donc que le nombre de patients diminue avec l'âge, ce qui avait déjà été entrevu lors de l'analyse générale des données.

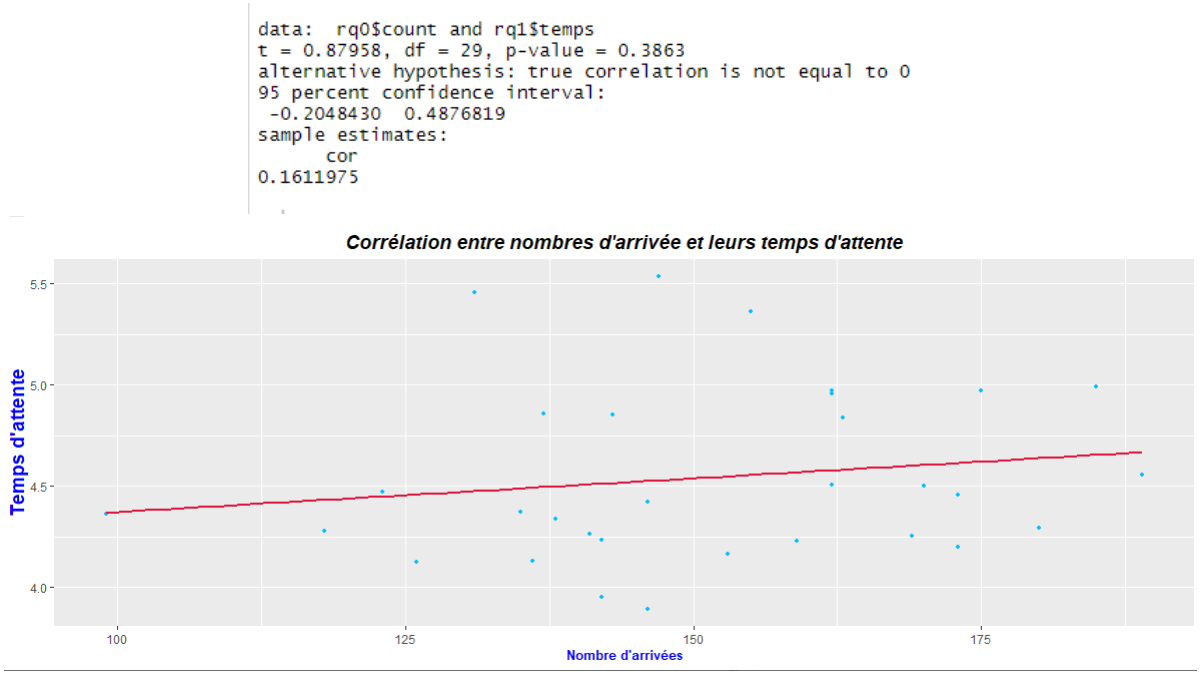

FIGURE 24 – Corrélation entre le nombre d'arrivées et le temps d'attente pendant le mois de janvier 2007 dans l'unité de médecine interne des urgences.

Nous Remarquons sur cette figure une corrélation positive faible entre les deux variables (nombre d'arrivées et temps d'attente).
Ici, lorsque le nombre de patients augmente, les valeurs de temps d'attente augmentent mais faiblement par rapport a l'augmentation des patients.
Ce qui, ajouté aux différentes analyses effectuées, démontre que le temps d'attente élevé dans l'hôpital (notameent les urgences), n'est pas principalement lié à l'afflux de patients.

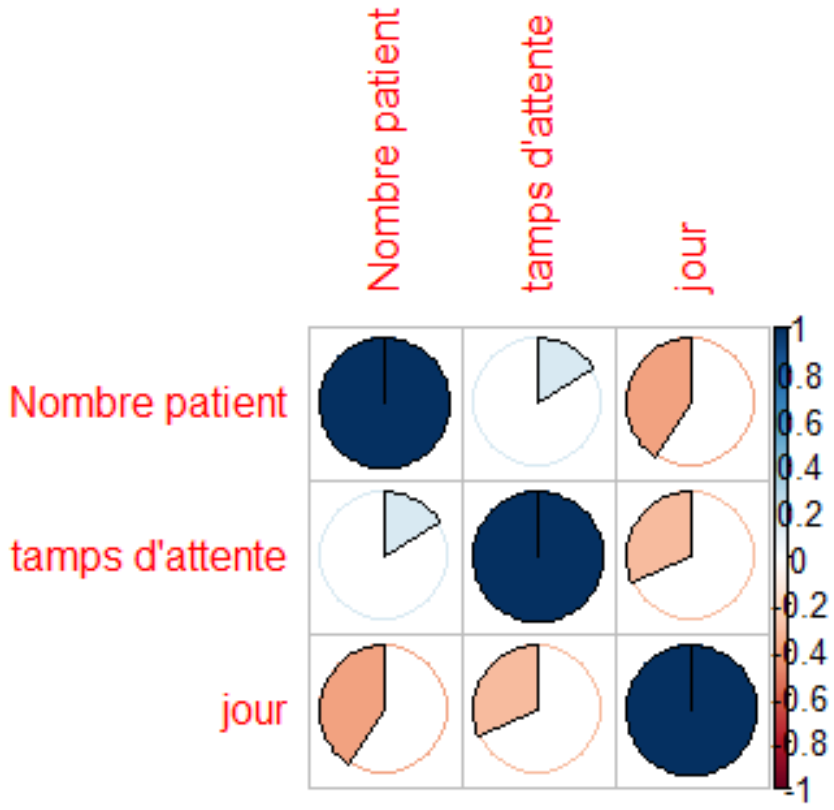

FIGURE 25 – Corrélation entre le nombre d'arrivées, le jour d'arrivée et le temps d'attente pendant le mois de janvier 2007 dans l'unité de médecine interne des urgences.

Cette figure montre les resultats du calcul de corrélation entre le nombre d'arrivée, leurs-temps d'attente et le jour (mois) d'arrivée de ces patients. Nous avons realisé une Visualisation de la matrice de corrélation par un corrélogramme grace a la fonction *"corrplot"* appliquée sur la la matrice de corrélation calculée avec *"cor"*. Tel-que nous pouvons voir qu'il y a une corrélation négative entre les jours et nombres d'arrivées par rapport à la corrélation avecle temps d'attente qui estest moins forte,et celle entre temps d'attentes et nombre d'arrivée qu'est encore moins.

Le département	Le minimum	le premier quartile	la médiane	la moyenne	le troisième quartile	Le maximum
Emergency Internal Medicine Unit	3185	3766	3968	3937	4122	4575
Emergency Surgery Unit	2226	2885	2991	2978	3166	3377
Emergency Traumatology Unit	32	46.5	59	63.43	70	238
Emergency Otorhinolaryngology Unit	36	62.25	69.5	69.54	76	102
Emergency Ophthalmology Unit	164	221	243	240.8	265.5	300
Emergency Psychiatry Unit	123	153.8	168	172.1	188.8	225
Emergency Gynecology Unit	188	235.8	250.5	253.2	269	307
Pediatric Emergency Unit	794	1062	1172	1184	1297	1556
Pediatric Emergency Surgery Unit	1	1	2	3.147	3.75	12
Emergency Maternity Unit t	380	539.2	589	603	661.8	867

FIGURE 26 – Les statistiques sur les arrivées dans les 10 premiers département.

7 Conclusion

Dans le cadre de notre projet de TER, nous avons eu l'opportunité d'avoir une première expérience dans l'analyse de données, notammeent la fouille exploratoire, en travaillant sur des données réelles mises à disposition par l'hôpital de Rambam.
Après avoir découvert l'EDA en faisant une recherche bibliographique et webographique, et transféré puis nettoyé les données vers une solution logicielle adéquate, nous avons donc divisé notre analyse comme suit :

1. Analyse générale
2. Analyse temporelle du département des urgences
3. Analyse détaillée

Ces étapes successives ont eu pour but d'accomplir l'objectif qui nous a été assigné, dont le principal était les phénomènes d'arrivées dans les unités des urgences, qui sont de loin les plus sollicitées dans l'hôpital et les plus à même d'être débordées.
Le resultat de notre travail aspire donc à fournir un premier support d'informations significatives et d'hypothèses pour améliorer l'affectation du personnel soignant ou pour effectuer des modifications logistiques dans l'hôpital.
Ce projet aura été l'occasion pour tous les membres du groupe de prendre en main le langage R et d'acquérir des connaissances aujourd'hui solides nous permettant d'effectuer divers autres projets utilisant cette technologie.
Travailler sur les données d'un hôpital nous aura donné une expérience non négligeable en vue de possibles opportunités de missions/offres sur des projets dans le domaine médical.
Enfin, étant tous les 5 dans un master de Data, cette expérience dans une sous-discipline de la Data Science nous apportera plus de versatilité et de polyvalence en vue de notre carrière.

Références

[1] Shelby Blitz. *What Is Exploratory Data Analysis ?* URL : https://www.sisense.com/blog/exploratory-data-analysis/.

[2] Fabrice Rossi. "Analyse exploratoire de données". In : *Télécom ParisTech* (), p. 5.

[3] Jean-Claude Oriol. *Éléments d'histoire de la statistique*. Statistix. inria-00466297, 2010.

[4] Wikipédia. *Analyse des données*. URL : https://fr.wikipedia.org/wiki/Analyse_des_donn%C3%A9es.

[5] Tukey John W. *The Future of Data Analysis*. 1961.

[6] Tukey John W. *Exploratory Data Analysis*. 1977.

[7] Professor Avishai Mandelbaum et al. "HomeHospital (Rambam) Database Tables and Fields". In : (2013).

[8] CRAN R project. *RMySQL : Database Interface and 'MySQL' Driver for R*. URL : https://cran.r-project.org/web/packages/RMySQL/index.html.

[9] *Corrélation*. URL : http://spss.espaceweb.usherbrooke.ca/pages/stat-inferentielles/correlation.php.

[10] Khan Academy. *Résumé : Coefficient de corrélation*. URL : https://fr.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/scatterplots-and-correlation/a/correlation-coefficient-review?fbclid=IwAR2fO3tLN2wOyEu0EnpNAcXrFd1A6mRZILtcX-34iUrS2nRpn72-BhWi2zc.