# Fa/Fi1 - Saé 105 - Traiter les données

---

## 1. Contexte

Vous avez accès à un fichier compilant tous les mouvements de produits pharmaceutiques pour différents services d'un hôpital. Ce fichier est au format ".csv". Il prend en compte diverses informations qui sont présentées en colonnes dans le csv.

## 2. Objectifs

Il s'agit de "traiter ces données" pour en extraire des informations. Réaliser une application permettant, à l'aide d'une interface graphique, à partir de ce fichier, ou d'un autre fichier au même format, choisi par l'utilisateur, de traiter puis d'afficher les données sous forme de graphiques, et d'exporter ces représentations graphiques dans un fichier au format PDF.

## 3. Les ressources

- Le jeu de données (à récupérer sur Moodle)
- La documentation officielle python
- La documentation en ligne pour les bibliothèques
    - https://pandas.pydata.org/docs/getting_started/index.html
    - https://pandas.pydata.org/docs/user_guide/index.html
    - https://matplotlib.org/stable/tutorials/introductory/pyplot.html (intro à matplotlib)
    - https://matplotlib.org/stable/contents.html (documentation complète)
    - https://matplotlib.org/stable/gallery/lines_bars_and_markers/horizontal_barchart_distribution.html#sphx-glr-gallery-lines-bars-and-markers-horizontal-barchart-distribution-py
    - https://matplotlib.org/stable/plot_types/basic/stem.html#sphx-glr-plot-types-basic-stem-py

Exemple de jeu de données ouvert avec Pandas :

| | PRCLEUNIK | DATEMVT | HEUREMVT | SENSMVT | TYPEMVT | SERVICE | MAGASIN | QUANTITE | VALHT | B_URGENT |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 5926 | 2020-06-03 | 0823 | 2 | 7 | 1602 | 433 | 20 | 0.7000 | 1.0 |
| 1 | 3716 | 2020-06-03 | 0823 | 2 | 7 | 1602 | 433 | 10 | 5.1970 | 1.0 |
| 2 | 29207 | 2020-06-03 | 0833 | 2 | 7 | 1901 | 433 | 12 | 10.7988 | 1.0 |
| 3 | 2034 | 2020-06-03 | 0833 | 2 | 7 | 1901 | 433 | 20 | 43.0020 | 1.0 |
| 4 | 11675 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 56 | 1.0024 | 0.0 |
| 5 | 2386 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 24 | 0.2880 | 0.0 |
| 6 | 1855 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 10 | 0.3700 | 0.0 |
| 7 | 1856 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 20 | 0.5980 | 0.0 |
| 8 | 14522 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 20 | 0.4000 | 0.0 |
| 9 | 3460 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 40 | 1.4000 | 0.0 |
| 10 | 13473 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 10 | 3.5990 | 0.0 |
| 11 | 8749 | 2020-03-06 | 1224 | 2 | 7 | 5121 | 433 | 10 | 1.1880 | 0.0 |

# 4. Travail à réaliser et à produire : l'analyse de données sur les données contenues dans le fichier CSV.

## Les séries de graphiques possibles :

À chaque fois on vous propose de faire un graphique parmi 4 proposés. Ils sont présentés en général du plus facile au plus difficile.

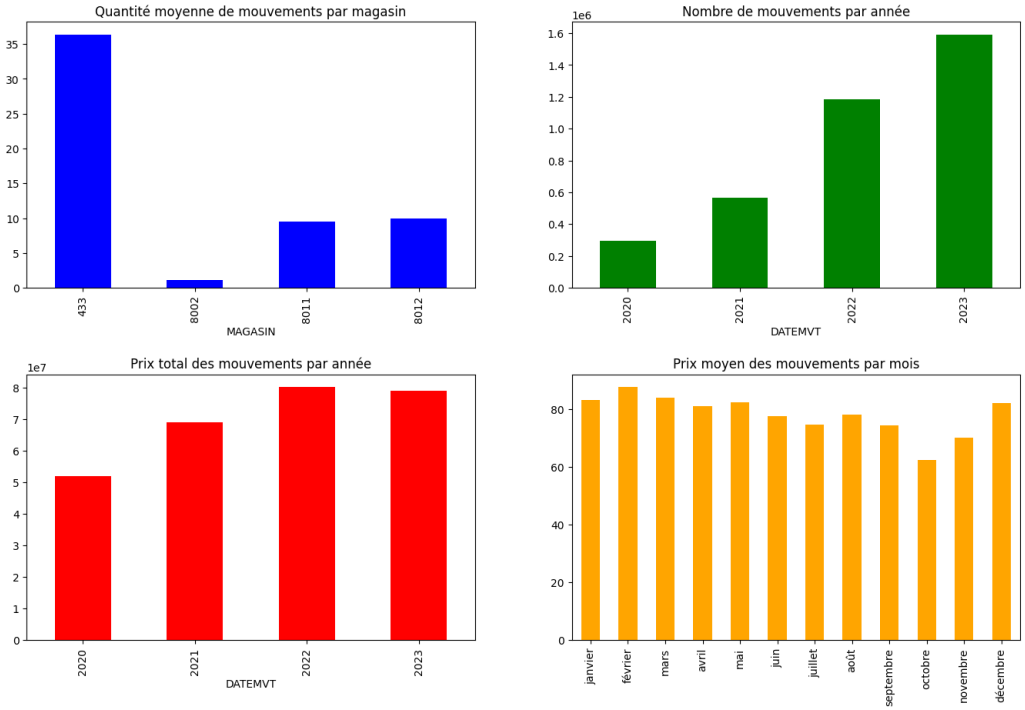
## Vous réaliserez les niveaux suivants dans cet ordre :

### Niveau 0 : ouverture du fichier

- L'ouverture et la lecture du fichier CSV choisi par l'utilisateur
- À la fois en utilisant la bibliothèque Pandas et la bibliothèque standard de Python. Vous ouvrirez le fichier principal directement à partir du fichier zip sans passer par un fichier csv sur le disque dur.
- Si un certain nombre de lignes sont incorrectement formatées pour vos analyses, vous pouvez les laisser de côté à condition de **les compter** et de vérifier qu'elles sont en **proportion réduite** (inférieure à 1%) pour ne pas fausser les statistiques que vous allez faire.

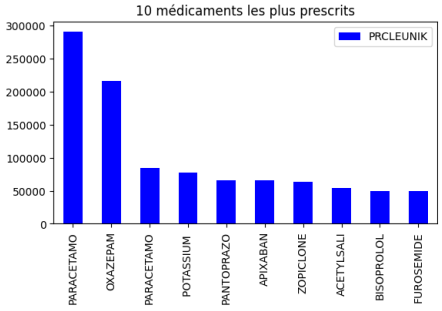
### Niveau 1 : un graphique parmi les 4 suivants avec matplotlib et datetime, sans utiliser Pandas

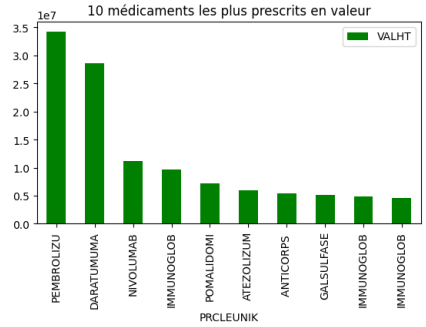
1. La quantité moyenne des mouvements de médicaments par magasin
2. Le nombre de mouvements par année
3. Le prix total des mouvements par année
4. Le prix moyen des mouvements par mois

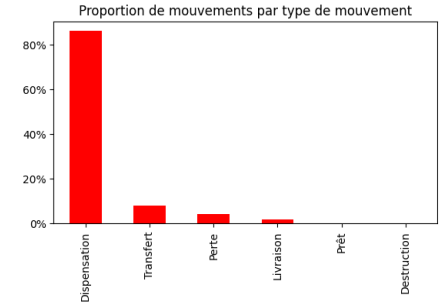


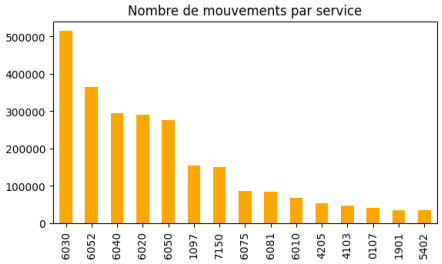
**Niveau 2 : un graphique parmi les 4 suivants en utilisant Pandas, sans matplotlib**

1. Les 10 médicaments les plus prescrits en quantité
2. Les 10 médicaments les plus prescrits en valeurs marchande
3. La proportion de mouvements par type
4. Le nombre de mouvements par service pour les 15 principaux services (sauf -1)



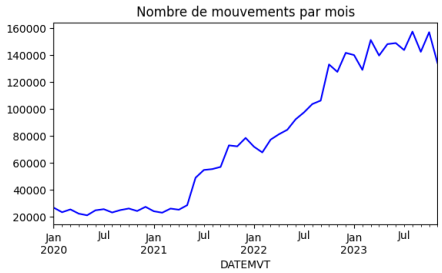



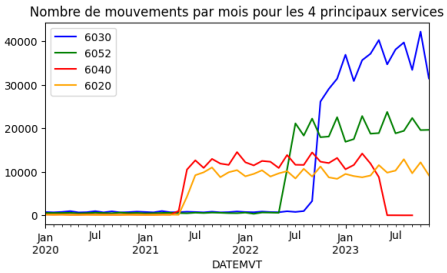



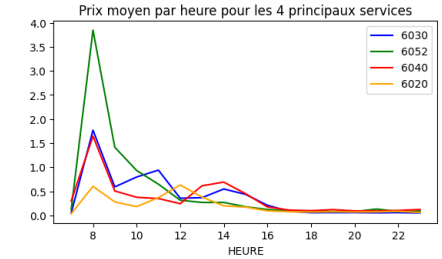



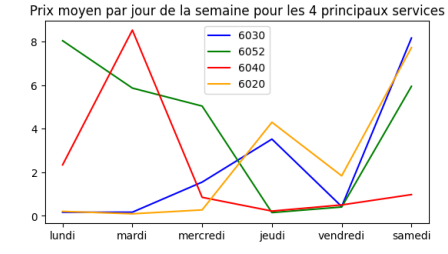

**Niveau 3 : un graphique parmi les 3 suivants avec matplotlib, et/ou Pandas**

1. La courbe du nombre de mouvements par mois
2. Le nombre de mouvements par mois pour les 4 principaux services
3. Le prix moyen par heure de la journée (de 7h à 23h) pour les 4 principaux services
4. Le prix moyen par jour de la semaine pour les 4 principaux services



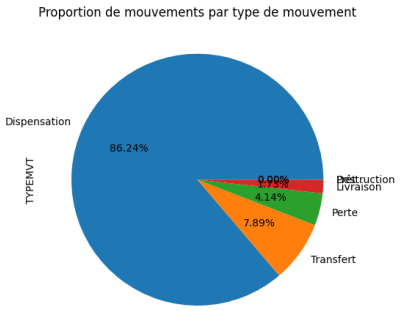



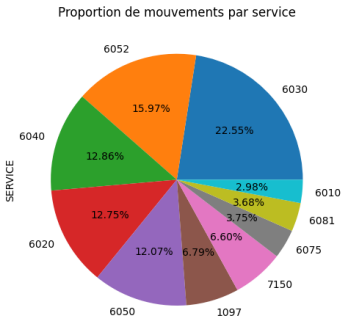



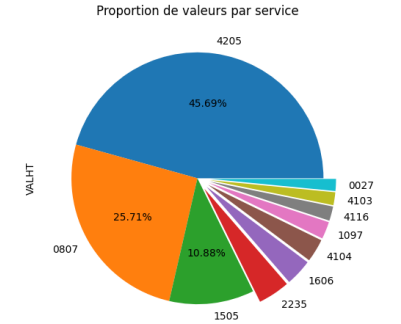



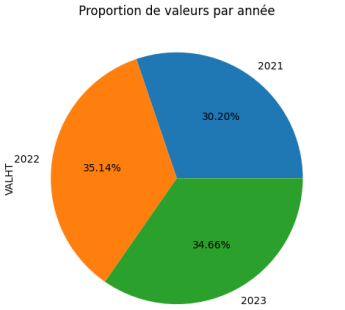
## Niveau 4 : un graphique parmi les 3 suivants sans utiliser Pandas

1. Proportion de mouvements par type de mouvement
2. Proportion de mouvements par service
3. Proportion du total de valeur par type service
4. Proportion du total de valeur par année



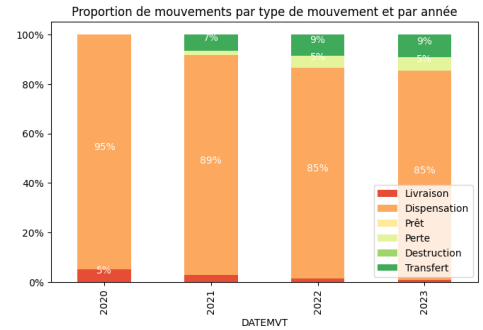



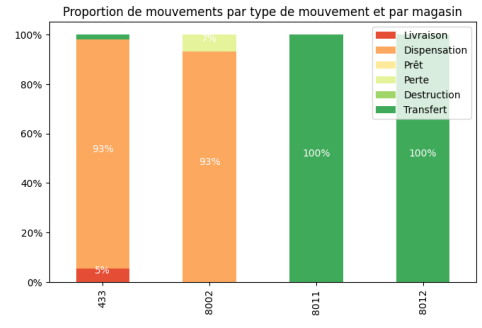



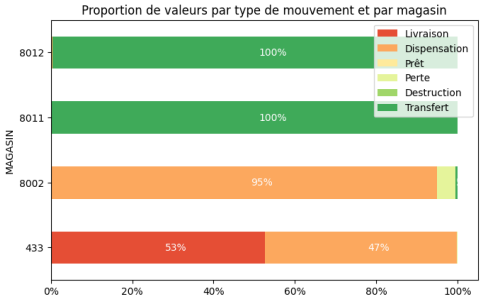



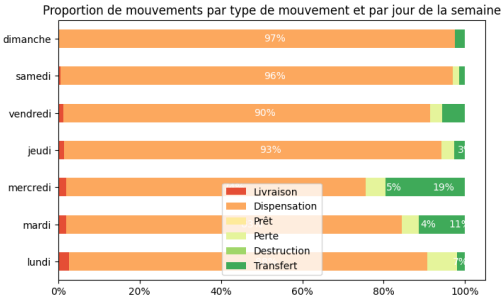

## Niveau 5 : un graphique parmi les 4 suivants, en barres empilées

1. La proportion du nombre de mouvements par type de mouvement pour chaque année
2. La proportion du nombre de mouvements par type de mouvement pour chaque magasin
3. La proportion de la valeur des mouvements par type de mouvement pour chaque magasin
4. La proportion de ces valeurs par type de mouvement pour chaque jour de la semaine



### Vous réaliserez ensuite les améliorations parmi :

**Amélioration 1 : un graphique de votre choix**

- Réaliser un graphique qui illustre un traitement de données que vous choisissez (différent des propositions ci-dessus). C'est une création personnelle.

**Amélioration 2 : une interface graphique**

- Réaliser une interface graphique pour sélectionner le fichier de données, sélectionner le type de graphique (et éventuellement la catégorie, l'unité de temps, bref les options de ce graphique), afficher le tableau de données après analyse, afficher le graphique sélectionné.

**Amélioration 3 : un pdf (après avoir réalisé l'amélioration 2)**

- Ajouter à l'interface l'export des graphiques au format PDF

# 5. L'évaluation

- L'évaluation se fera par une présentation en binôme de 5 minutes suivie de 5 minutes de questions techniques.
- À partir d'un fichier csv fourni, vous montrerez le résultat de votre travail en présentant les analyses demandées
- L'aspect technique y sera principalement jugé
- Montrez ce que vous savez faire !

# 6. Organisation

Plusieurs conseils sur votre façon de vous organiser :

- **Fonctions** : faites des fonctions associées aux différentes tâches que vous effectuez. Par exemple : ouverture du csv avec retour d'une dataframe, ou d'un dictionnaire de données, une fonction pour la création de chaque graphique etc.
- **Données** : séparez les données en plusieurs ensembles (par échantillons aléatoires, ou en séparant par années etc.). Idéalement on peut envisager 3 ensembles de données : entraînement (pour mettre au point les graphiques), tests (après la première étape) et validation.
- **Tests** : à chaque modification d'une fonctionnalité, il faut prévoir un peu de temps pour vérifier que les autres fonctionnalités sont encore actives et correctes.