PROJET RÉALISÉ PAR L'ÉQUIPE 2

RAPPORT DE GROUPE EN SCIENCES DES DONNÉES 2 + BASES DE DONNÉES

Girondin Audric 22001931 Duckes Jonathan 22001974 Mendil Youcef 201810962
Mohand-Amer Manel 201810962

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Décembre 2022

Soumis comme contribution partielle
pour le cours Science des données 2 et Bases de données

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Signature: _____ Date: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Signature: _____ Date: _____

Remerciements

Nos plus sincères remerciements vont à nos encadrants pédagogiques pour les conseils avisés sur notre travail.

12/09/2022.

Résumé

Table des matières

Chapitre 1

Introduction

Suite à la récente pandémie survenue ces dernieres années à cause du Covid-19, nous nous sommes intéressés au domaine de la santé, plus précisement dans la région d'Outre-mer. Durant cette période de nombreuses personnes sont tombées malades, tous les hopitaux ont été mobilisés ainsi qu'énormément de personnels mais il y a aussi eu une forte quantité de médicaments vendus et du coup remboursés, c'est pourquoi notre objet d'études portera sur :

Quels sont les médicaments les plus ou les mieux remboursés en région d'Outre-mer ?

Il est important de motiver l'importance de ces questions, et leur actualité. Vous pouvez également expliquer quelles données vous pensez utiliser et en quoi elles peuvent permettre d'apporter des éléments de réponses aux questions posées.

1.1 Quelques détails techniques

Rien n'interdit d'inclure une figure tant qu'elle comprend une légende (ce qui est le cas pour la Figure 1.1 ci-après) et qu'elle est référencée quelque part dans le document (ce qui est le cas dans les parenthèses qui précédent).

Figure 1.1: Une légende sous la figure.

Notez que contrairement à une figure, la légende d'un tableau doit être mise **au-dessus** de celui-ci (*e.g.*, voir la Table 7.1 en Annexe).

Chapitre 2

Base de données

2.1 Descriptif des tables

- Lister les tables sélectionnées (vous pouvez en rajouter par rapport à celles initiales) et donner leur url.
- Filtrer les lignes et les colonnes (éventuellement réduire le périmètre du projet) et décrire les critères de sélection (*e.g.*, ne garder que 5 colonnes sur les 15, ne garder que les lignes qui correspondent à une ville en particulier, ...).
- Pour chaque table conservée : préciser le nombre de lignes et de colonnes après filtrage, lister les colonnes et donner pour chacune le type, la signification du champ et des caractéristiques (unique, clés, valeur manquante, ...).

2.2 Modèles MCD et MOD

Ci dessous notre MCD et MOD :

Le MCD,Figure 2.1

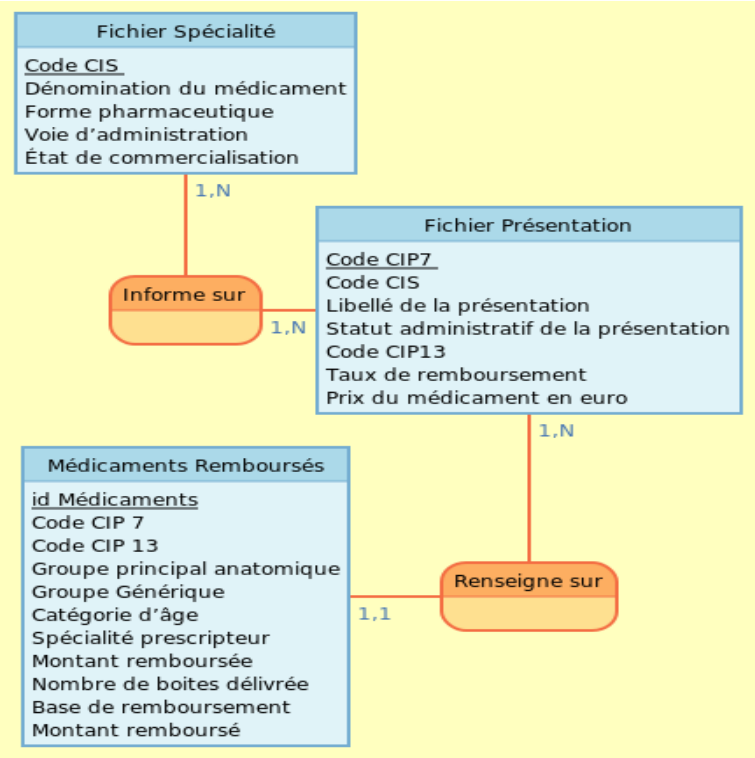


Figure 2.1: LE MCD

Le MOD,Figure 2.2

FICHIER SPÉCIALITÉ (Code CIS,Dénomination du médicament, Forme pharmaceutique, Voie d'administration,État de commercialisation)

FICHIER PRÉSENTATION (Code CIP7, Libellé de la présentation, Statut administratif de la présentation, Code CIP13, Taux de remboursement,Prix du médicament en euro)

Informe_sur (Code_CIS, Code_CIP7)

MÉDICAMENTS REMBOURSÉS (id Médicaments, Groupe principal anatomique, Groupe Générique, Catégorie d'âge, Spécialité prescripteur, Montant remboursée, Nombre de boites délivrée, Base de remboursement, Montant remboursé, *Code CIP7*)

Figure 2.2: LE MOD

2.3 Import des données

- Suppressions de valeurs manquantes
- Suppression de colonnes inutiles pour notre problématique
- Tri par région afin d'étudier uniquement la région Outre-mer

2.4 Requêtes réalisées

Voici les différentes requêtes réalisées au cours de notre projet, voir les figures ci dessous:

- Le montant total vendu par medicament,Figure **??**

```
SELECT pres.libelle_present ,pres.prix_min as 'prix min',pres.prix_max as 'prix max' , om.boites as 'boites vendues',
round(pres.prix_min*om.boites,2) as 'montant total au prix min par medicament'
FROM pres , outremer21a_txt om
WHERE pres.CodeCIP7 = om.CodeCIP7
GROUP BY pres.libelle_present
```

libelle_present	prix min	prix max	boites vendues	montant total au prix min par medicament
plaquette(s) thermoformée(s) PVC PVDC aluminium de...	2.63	3.65	1742	4581.46
plaquette(s) PVC polyéthylène PVDC aluminium de 28...	3.42	4.44	60	205.20
plaquette(s) PVC PVDC aluminium de 28 comprimé(s)	5.69	6.71	184	1046.96
plaquette(s) PVC PVDC aluminium de 28 comprimé(s) ...	5.69	6.71	248	1411.12
1 flacon(s) pressurisé(s) aluminium de 100 g	9.46	10.48	25	236.50
1 plaquette(s) PVC-Aluminium TE (thermo-élastique)...	1.23	2.25	41	50.43
3 plaquette(s) PVC-Aluminium TE (thermo-élastique)...	2.94	3.96	111	326.34
plaquette(s) thermoformée(s) PVC-Aluminium de 30 c...	2.80	3.82	3549	9937.20
plaquette(s) thermoformée(s) PVC-Aluminium de 90 c...	7.44	10.20	495	3682.80

- Lister les médicaments non remboursés

```
SELECT DISTINCT pres.libelle_present
FROM pres
WHERE pres.codecip7 not in (
SELECT DISTINCT om.codecip7
FROM outremer21a_txt om
)
```

libelle_present
1 flacon(s) polyéthylène de 5 ml avec compte-goutt...
plaquette(s) polyamide aluminium PVC-Aluminium de ...
1 ampoule(s) en verre brun de 1 ml
plaquette(s) thermoformée(s) PVC polyéthylène PVDC...
plaquette(s) thermoformée(s) PVC polyéthylène PVDC...
plaquette(s) PVC polyéthylène PVDC aluminium de 30...
plaquette(s) PVC polyéthylène PVDC aluminium de 90...
plaquette(s) thermoformée(s) PVC PVDC aluminium de...
plaquette(s) PVC PVDC aluminium de 90 comprimé(s)
plaquette(s) PVC PVDC aluminium de 28 comprimé(s)
4 flacon(s) en verre - 4 seringue(s) préremplie(s)...

- Les médicaments les plus souvent remboursés

```
SELECT DISTINCT COUNT(om.codecip7) as nb_cip, pres.libelle_present as medicament, om.l_ATC1 as'groupe
anatomique'
FROM outremer21a_txt om, pres
WHERE om.CodeCIP7 = pres.codecip7
GROUP BY om.codecip7
ORDER BY `nb_cip` DESC limit 10
```

nb_cip ▾ 1	medicament	groupe anatomique
19	1 ampoule(s) en verre brun de 2 ml	Système digestif et métabolisme
18	1 ampoule(s) en verre brun de 2 ml	Système digestif et métabolisme
18	1 tube(s) aluminium verni de 30 g	Dermatologie
18	24 sachet(s) polyester aluminium polyéthylène de 1...	Système digestif et métabolisme
18	24 sachet(s) polytéréphtalate (PET) aluminium poly...	Système digestif et métabolisme
18	plaquette(s) polyamide aluminium PVC-Aluminium de ...	Système digestif et métabolisme
17	1 ampoule(s) en verre brun de 2 ml	Système digestif et métabolisme
17	1 flacon(s) en verre brun avec compte-gouttes poly...	Système digestif et métabolisme
17	3 plaquette(s) aluminium PVC PVDC de 21 comprimé(s...	Système génito-urinaire et hormones sexuelles
17	3 plaquette(s) thermoformée(s) PVC-Aluminium de 28...	Système génito-urinaire et hormones sexuelles

- Les moins souvent remboursés

```
SELECT DISTINCT COUNT(om.codecip7) as nb_cip, pres.libelle_present as medicament, om.l_ATC1 as'groupe
anatomique'
FROM outremer21a_txt om, pres
WHERE om.CodeCIP7 = pres.codecip7
GROUP BY om.codecip7
ORDER BY `nb_cip` ASC limit 10
```

nb_cip ▲ 1	medicament	groupe anatomique
1	1 ampoule(s) en verre brun de 2 ml	Système digestif et métabolisme
1	1 cartouche(s) en verre de 0.48 ml avec 6 aiguille...	Système génito-urinaire et hormones sexuelles
1	1 cartouche(s) en verre de 0.5 ml avec 10 aiguille...	Système génito-urinaire et hormones sexuelles
1	1 cartouche(s) en verre de 0.75 ml avec 10 aiguill...	Système génito-urinaire et hormones sexuelles
1	1 cartouche(s) en verre de 1.5 ml dans stylo pré-r...	Hormones systémiques, à l exclusion des hormones s...
1	1 cartouche(s) en verre de 60 dose(s) dans stylo j...	Système digestif et métabolisme
1	1 dispositif intra-utérin conditionné dans une poc...	Système génito-urinaire et hormones sexuelles
1	1 flacon olyéthylène haute densité (PEHD) avec fer...	Système génito-urinaire et hormones sexuelles
1	1 flacon(s) de poudre en verre - 1 ampoule(s) de s...	Anti-infectieux (usage systémique)
1	1 flacon(s) en verre	Anti-infectieux (usage systémique)

- Lister le prix et le montant remboursé de chaque médicament

```
SELECT pres.libelle_present ,pres.prix_min as 'prix min', pres.Prix_max as 'prix max' ,
round((outremer21a_txt.REM/outremer21a_txt.BOITES),2) as 'montant remboursé'
FROM pres , outremer21a_txt
WHERE pres.CodeCIP7 = outremer21a_txt.CodeCIP7
GROUP BY pres.libelle_present
```

libelle_present	prix min	prix max	montant remboursé
plaquette(s) thermoformée(s) PVC PVDC aluminium de...	2.63	3.65	0.48
plaquette(s) PVC polyéthylène PVDC aluminium de 28...	3.42	4.44	3.58
plaquette(s) PVC PVDC aluminium de 28 comprimé(s)	5.69	6.71	2.70
plaquette(s) PVC PVDC aluminium de 28 comprimé(s) ...	5.69	6.71	2.62
1 flacon(s) pressurisé(s) aluminium de 100 g	9.46	10.48	7.49
1 plaquette(s) PVC-Aluminium TE (thermo-élastique)...	1.23	2.25	1.04
3 plaquette(s) PVC-Aluminium TE (thermo-élastique)...	2.94	3.96	2.39
plaquette(s) thermoformée(s) PVC-Aluminium de 30 c...	2.80	3.82	3.17
plaquette(s) thermoformée(s) PVC-Aluminium de 90 c...	7.44	10.20	7.96
plaquette(s) PVC PVDC aluminium de 30 comprimé(s)	8.76	9.78	4.53

- Le Taux de remboursement

```
SELECT pres.libelle_present ,pres.taux_remboursement as 'taux de remboursement theorique' , round((om.rem/om.boites)/ pres.prix_
as 'taux de remboursement au prix min ', round((om.rem/om.boites)/ pres.prix_max ,2) as 'taux de remboursement au prix max '
FROM pres , outremer21a_txt as om
WHERE pres.codecip7 = om.codecip7
AND ((om.rem/om.boites)/pres.prix_min) < 1
AND ((om.rem/om.boites)/pres.prix_max) < 1
GROUP BY pres.libelle_present
```

libelle_present	taux de remboursement theorique	taux de remboursement au prix min	taux de remboursement au prix max
plaquette(s) thermoformée(s) PVC PVDC aluminium de...	0.15	0.18	0.13
plaquette(s) PVC PVDC aluminium de 28 comprimé(s)	0.30	0.47	0.40
plaquette(s) PVC PVDC aluminium de 28 comprimé(s) ...	0.30	0.46	0.39
1 flacon(s) pressurisé(s) aluminium de 100 g	0.65	0.79	0.71
1 plaquette(s) PVC-Aluminium TE (thermo-élastique)...	0.65	0.84	0.46
3 plaquette(s) PVC-Aluminium TE (thermo-élastique)...	0.65	0.81	0.60
plaquette(s) thermoformée(s) PVC-Aluminium de 30 c...	0.65	0.87	0.64
plaquette(s) thermoformée(s) PVC-Aluminium de 90 c...	0.65	0.99	0.72

2.5 Quelques détails techniques

On peut interagir avec une base de données directement depuis RMarkdown. Voilà un exemple:

```
# install.packages("RMySQL")
# install.packages("DBI")
library(DBI)
con <- DBI::dbConnect(RMySQL::MySQL(),
host = "sql11.freemysqlhosting.net",
port = 3306,
username = "sql11522616",
password = "DTqQiaguNA",
dbname = "sql11522616")
```

```
# sql (plutôt que r) chunk avec l'argument connection = con
# Écrire vos requêtes SQL ci-dessous
show tables;
```

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Nous avons utilisés principalement le langage de programmation Rstudio, Wamp mais aussi Excel.

- Rstudio pour les analyses statistiques et la création du rapport à travers RMarkdown
- Wamp afin de se connecter à PhpMyAdmin afin de travailler sur nos différentes requêtes
- Excel pour effectuer le pré-traitement des données
- Whatsapp, une application de messagerie instantanée afin de communiquer sur les avancées

Nous avons travaillé sur 4 ordinateurs differents :

- Swift SF113-31, processeur Intel(R) Pentium(R) CPU N4200 1.10 GHz, Mémoire RAM installée :4,00,Go (3,84,Go utilisable) Type du système : Système d'exploitation 64bits,processeur x64, Windows 10
- Dell XPS 13 7390 2-in-1,processeur Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz, Mémoire RAM installée:16,0 Go (15,8 Go utilisable) Type du système : Système d'exploitation 64 bits, processeur x64, Windows 11

3.2 Description des Données

Comment les données sont-elles stockées? Quelles sont les tailles des fichiers en jeu? Combien y a t-il de fichiers? Combien d'unités statistiques? Combien de variables? etc.

3.3 Nettoyage des données

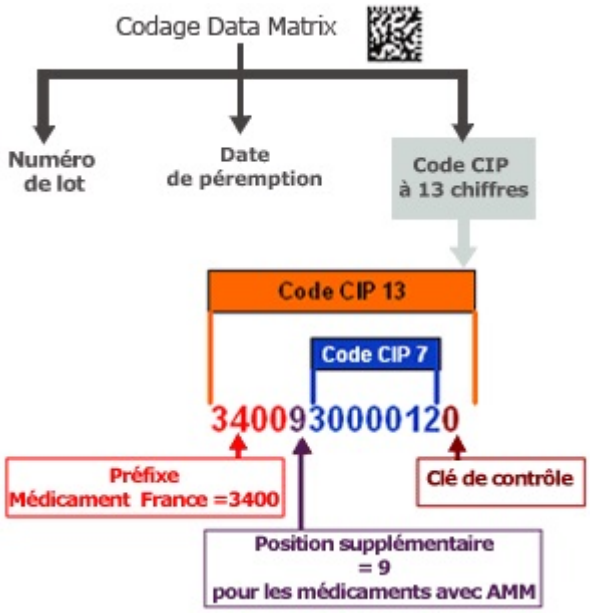
En ce qui concerne notre base de données nous n'avons décidé de supprimer les valeurs manquantes cependant il y avait beaucoup de colonnes inutiles à l'analyse de données, donc nous avons décidés de les supprimer.

3.4 Étapes de Pré-traitements

Quelles transformations avez-vous effectuées sur vos données pour les rendre utilisables? Tout d'abord notre jeu de donnée était composé de plus de d'1 million de lignes.En filtrant par les régions et en ne gardant que la région Outre-Mer cela nous a permis de réduire la base à peu près 45 000 lignes.

A l'aide du logiciel R, nous avons gardé uniquement les lignes des clés existantes dans toutes les tables,nous avons utilisé des jointures internes.

De plus nous avons concaténé deux colonnes afin de créer une clé unique,Figure **??**,ci- dessous :



3.5 Modélisation de la base de données

Proposer un modèle conceptuel des données (UML ou entité relation) sous la forme d'un schéma (*e.g.*, Figure 3.1).

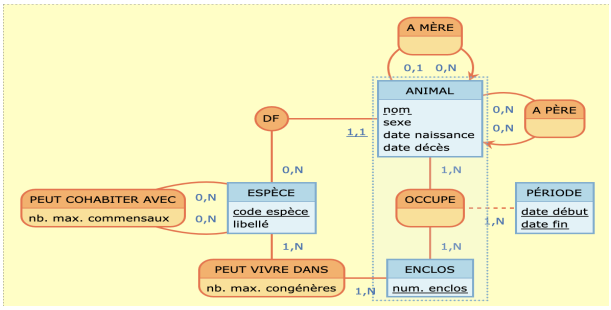


Figure 3.1: Relations.

3.6 Modélisation statistique

Quels outils ou méthodes de statistiques allez-vous utiliser? Donner des équations mathématiques sMil y a lieu et lister les éventuels présupposés («assumptions» en anglais) que vous devez faire sur les données afin d'utiliser ces outils ou méthodes (*e.g.*, normalité, absence de valeurs aberrantes, etc.).

Il est également bon d'indiquer quelles sont les avantages et les limites de ces méthodes.

Vous pourrez consulter avec profit les Chapitre 11–13 du livre sur R utilisé pendant le cours :
http://biostatisticien.eu/springeR/livreR.pdf

Chapitre 4

Analyse Exploratoire des Données

Toute étude impliquant des données doit **obligatoirement** inclure une analyse exploratoire préalable. Celle-ci permet de mieux comprendre l'information contenue dans les données.

Il faut produire de nombreux résumés graphiques (*e.g.*, histogrammes, nuages de points, boxplots, etc.) et numériques (*e.g.*, médiane, moyenne, variance, etc.) et conserver les plus pertinents (les autres pouvant être gardés en Annexe).

Vous pourrez consulter avec profit le Chapitre 9 du livre sur R utilisé pendant le cours :
http://biostatisticien.eu/springeR/livreR.pdf

4.1 Utiliser R

Il est facile d'inclure des codes R dans votre rapport, qui seront exécutés à la volée (*i.e.*, lors de la traduction de votre fichier `Rmd` en fichier `PDF` ou `DOC`). Par exemple:

```
boxplot(cars, col = c("#5975a4", "#cc8963"))
```

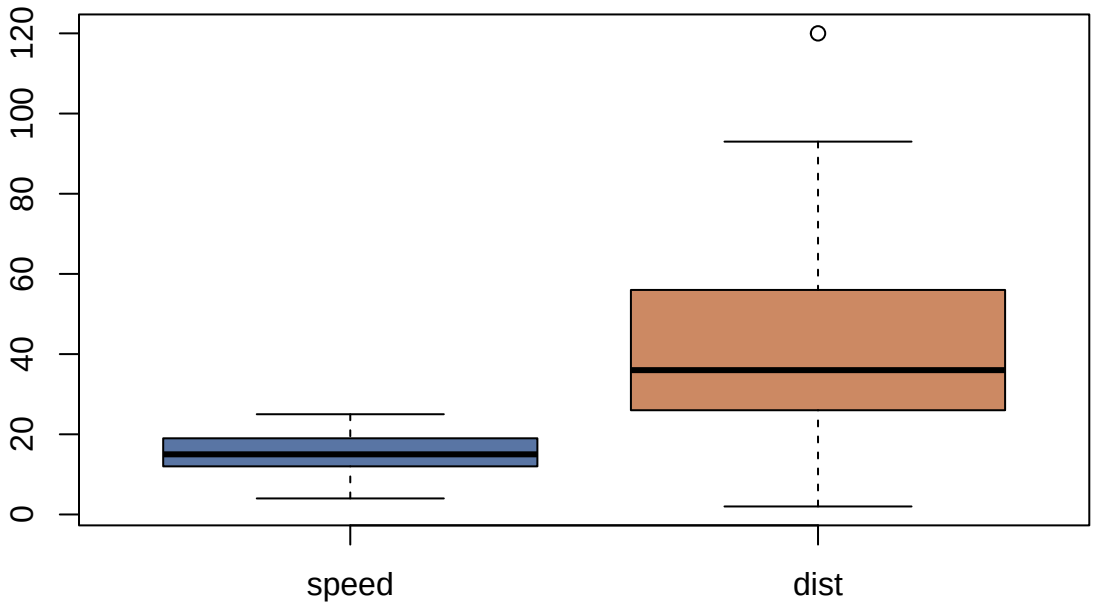


Figure 4.1: Deux boxplots.

```
colMeans(cars)
```

```
## speed  dist
## 15.40 42.98
```

Chapitre 5

Analyse et Résultats

5.1 Un premier modèle

Avoir un modèle très simple est toujours une bonne chose. Cela vous permet de calibrer tout résultat obtenu par la suite avec un modèle plus élaboré. Par ailleurs, il est mieux d'utiliser un modèle simple que l'on maîtrise bien.

Par exemple, si on souhaite expliquer les variations d'une variables réponse Y en fonction d'un certain nombre de prédicteurs $x_1, \ldots, x_p$, on peut utiliser un modèle de régression linéaire simple ($p = 1$) ou multiple ($p > 1$)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i, \qquad i = 1, \ldots, n.$$

où l'on présuppose que les ϵ_i sont i.i.d. $N(0, 1)$ pour tout $i = 1, \ldots, n$ (n étant la taille de l'échantillon).

5.2 Quelques exemples de résultats attendus

- Tests et ou intervalles de confiance pour une moyenne ou une proportion.
- Modèle de régression linéaire simple.
- Anova.

Chapitre 6

Discussion

Placer les résultats que vous avez obtenus dans le chapitre précédent en perspective par rapport au problème étudié.

Chapitre 7

Conclusion et perspectives

Quelles sont les conclusions principales? Quelles sont vos recommandations pour le commanditaire? Quelles analyses subséquentes pourraient être faites dans le futur?

On attend de vous deux types de perspectives : des perspectives à court terme pour améliorer rapidement votre approche et des perspectives à plus long terme qu'elles soient liées à la science des données ou au domaine métier pour lequel vous avez travaillé.

Bibliographie

Annexes

Codes

Ajouter vos codes informatique ici.

Tables

Si vous avez des tableaux supplémentaires, vous pouvez les ajouter ici.

Utiliser https://www.tablesgenerator.com/markdown_tables pour créer des tables Markdown simples, ou bien utiliser LaTeX.

Table 7.1: une légende au-dessus du tableau.

Les tables	sont	cool
col 1 est	alignée à gauche	$1600
col 2 est	centrée	$12
col 3 est	alignée à droite	$1

Aligner les nombres de la troisième colonne sur la droite permet d'afficher les unités au-dessus des unités, les dizaines au-dessus des dizaines, etc. Il faut toujours privilégier cette présentation.