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.

12/12/2022.

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 ?

1.1 Quelques détails techniques

la Figure 1.1.

Figure 1.1: Une légende sous la figure.

Chapitre 2

Base de données

2.1 Descriptif des tables

Tables sélectionnées :

Médicament :

ATC1 : Groupe Principal Anatomique CIP13 : Code Identification Spécialité Pharmaceutique GEN_NUM : Groupe Générique

Bénéficiaire :

BEN_REG : Région de Résidence du Bénéficiaire (Filtrée de façon à seulement prendre la région Outre-Mer)

Prescripteur :

PSP_SPE : Prescripteur

Indicateur :

REM : Montant Remboursé BSE : Base de Remboursement BOITES : Nombre de boîtes délivrées

2.2 Modèles MCD et MOD

Ci-dessous notre MCD et MOD :

Le MCD,Figure 3.1

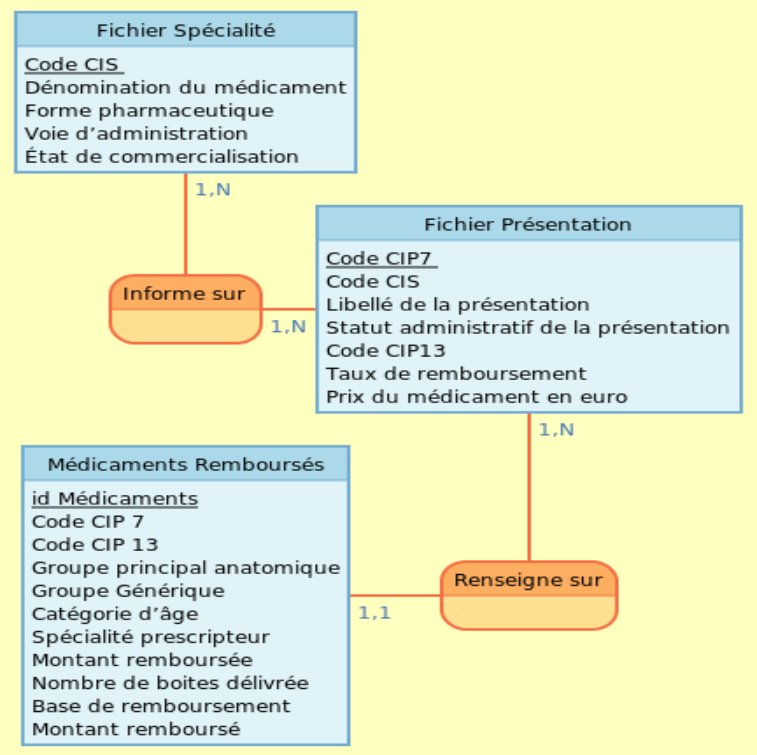


Figure 2.1: LE MCD

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

```
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

- Les 10 médicaments les mieux remboursés

```
SELECT DISTINCT outremer21a_txt.l_cip13, specialite.`Dénomination du médicament`, outremer21a_txt.l_ATC1,
pres.Taux_remboursement, pres.Prix_max, pres.prix_min
FROM outremer21a_txt, pres , specialite
WHERE outremer21a_txt.CodeCIP7=pres.CodeCIP7
AND specialite.Code_CIS=pres.Code_CIS
AND pres.taux_remboursement =
(SELECT MIN(pres.taux_remboursement)
FROM pres, outremer21a_txt , specialite
WHERE outremer21a_txt.CodeCIP7=pres.CodeCIP7
AND specialite.Code_CIS=pres.Code_CIS)
AND pres.Prix_max in
(SELECT pres.Prix_max
FROM pres, outremer21a_txt , specialite
WHERE outremer21a_txt.CodeCIP7=pres.CodeCIP7
AND specialite.Code_CIS=pres.Code_CIS)
ORDER BY pres.Prix_max DESC LIMIT 10
```

l_cip13	Dénomination du médicament	l_ATC1	Taux_remboursement	Prix_max 1	prix_min
MOVENTIG 25MG CPR 30	MOVENTIG 25 mg, comprimé pelliculé	Système digestif et métabolisme	0.15	71.34	70.32
MOVENTIG 125 MG CPR 30	MOVENTIG 12,5 mg, comprimé pelliculé	Système digestif et métabolisme	0.15	71.34	70.32
VITAROS 300 MCG CREME UNIDOSE 4	VITAROS 300 microgrammes, crème	Système génito-urinaire et hormones sexuelles	0.15	36.98	35.96
PROTOPIC 0,03 % POMMADE 1	PROTOPIC 0,03 %, pommade	Dermatologie	0.15	27.19	26.17
KOMBOGLYZE 2,5 MG/1000 MG CPR 60	KOMBOGLYZE 2,5 mg/1000 mg, comprimé pelliculé	Système digestif et métabolisme	0.15	26.90	25.88
ONGLYZA 5 MG CPR 30	ONGLYZA 5 mg, comprimé pelliculé	Système digestif et métabolisme	0.15	26.90	25.88
GALVUS 50MG CPR 60	GALVUS 50 mg, comprimé	Système digestif et métabolisme	0.15	26.55	25.53
EUCREAS 50 MG/1000 MG CPR 60	EUCREAS 50 mg/1000 mg, comprimé pelliculé	Système digestif et métabolisme	0.15	26.55	25.53
PROTOPIC 0,1 % POMMADE 1	PROTOPIC 0,1 %, pommade	Dermatologie	0.15	23.69	22.67
COMBODART 0,5MG/0,4MG GELULE 30	COMBODART 0,5 mg/0,4 mg, gélule	Système génito-urinaire et hormones sexuelles	0.15	20.18	19.16

2.5 Quelques détails techniques

Il est possible d'établir une connection entre Rstudio et PhpMyAdmin en local à l'aide du code suivant:

```
# install.packages("RMySQL")
# install.packages("DBI")
library(DBI)
con <- DBI::dbConnect(RMySQL::MySQL(),
host = "127.0.0.1",
port = 3306,
username = "root",
password = "",
dbname = "projet")
```

Nous avons utilisé cette méthode afin d'importer notre base de données sur Rstudio et créer des requetes ainsi que des graphiques.

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
- DESKTOP-SI1AM2U AMD A9-9425 RADEON R5, 5 COMPUTE CORES 2C + 3G, 3.10GHz, Mémoire RAM : 8,00(7,47 utilisable) Type du système : Système d'exploitation 64 bits, processeur x64, Windows 10
- MacBook Pro(13-inch,2017,Rwo Thunderbolt 3 ports), Processeur : 2,3 GHz Intel Core i5 double coeur, Mémoire 8Go 2133 MHz LPDDR3, macOS Monterey(version 12.6.1)

3.2 Description des Données

Les données sont stockés sur PhpMyAdmin dans la base de données sur un serveur en .sql et sur R elles sont importées sous forme de dataframe. Le fichier comporte 3 tables d'environ 45 000 lignes.

3.3 Nettoyage des données

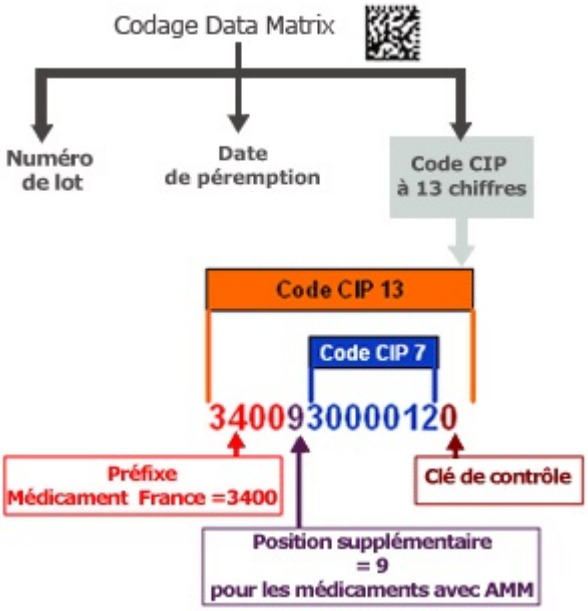
En ce qui concerne notre base de données nous n'avons décidé de supprimer les valeurs manquantes cependant il y avait beaucoup de colonnes inutiles à l'analyse de données, donc nous avons décidés de les supprimer.

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,ci- dessous :



3.5 Modélisation de la base de données

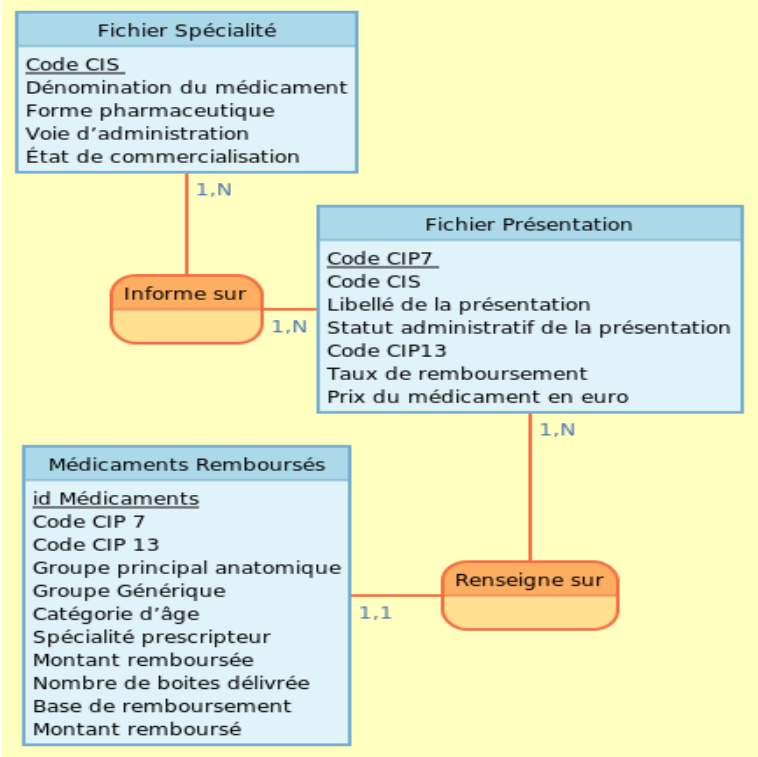


Figure 3.1: LE MCD

3.6 Modélisation statistique

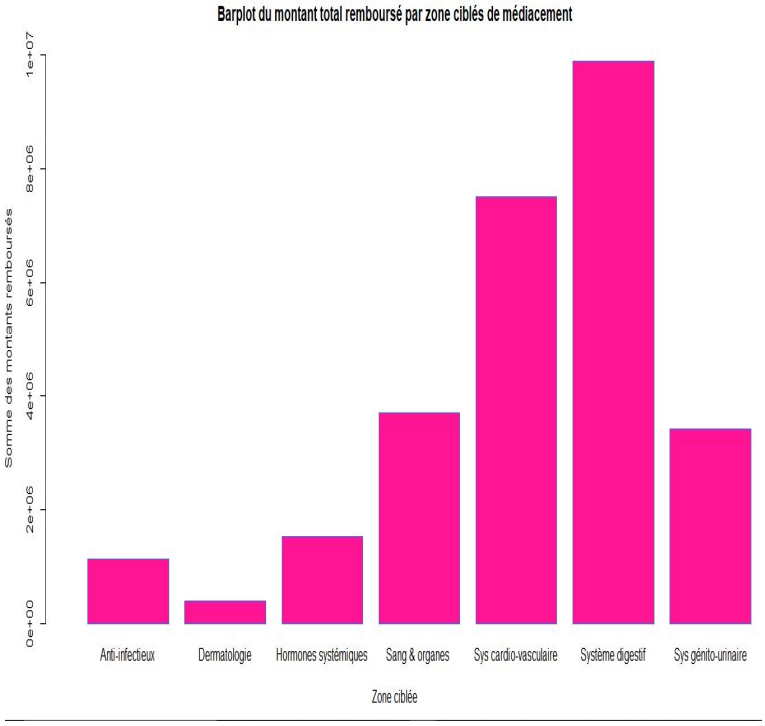
Nous allons utiliser des diagrammes en barre, ainsi que des boxplots afin d'avoir une représentation cohérente avec nos types de variables mais aussi afin d'avoir des informations sur les données que nous étudions et qui sont propre à nos requêtes

Chapitre 4

Analyse Exploratoire des Données

Ici , nous avons des statisques basqiues afin de nous informer sur données de la requete, elles ont été faite à laide de la fonction “summary”

```
> summary(d2)
    l_ATC1          somme_montant_remboursé
Length:7            Min.   : 397021
Class :character    1st Qu.:1319495
Mode  :character    Median :3423084
                    Mean   :3937130
                    3rd Qu.:5603350
                    Max.   :9894112
>
```



4.1 Utiliser R

Nous pouvons utliser R afin de montrer une partie de code utilisé par exemple :

```
l_ATC1<-c("Anti-infectieux","Dermatologie","Hormones systémiques","Sang & organes ","Sys cardio-vasculaire","Systè

somme montant remb<- c(1122911.0,397021.4,1516079.4,3695387.4,7511313.5,9894111.6,3423083.5)

barplot(somme montant remb,main = 'Barplot du montant total remboursé par zone ciblés de médiacement',xlab= "Zone
```

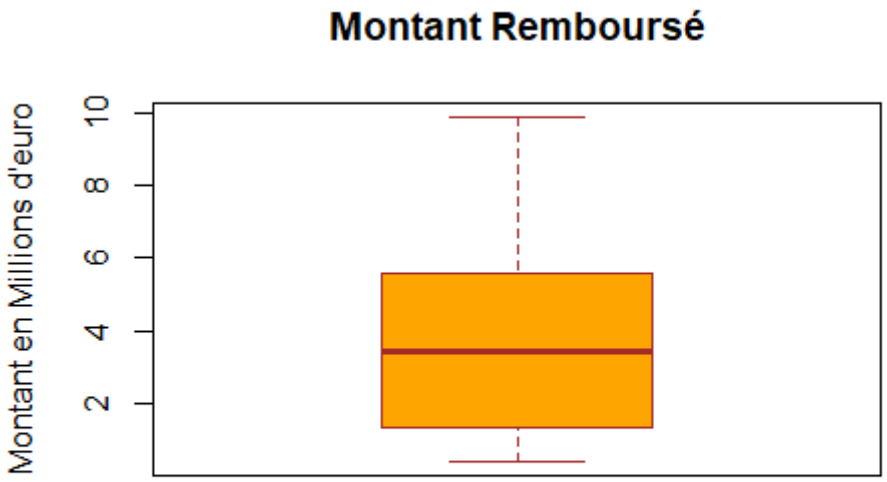
Le code ci-dessus a généré le graphique précédent

Chapitre 5

Analyse et Résultats

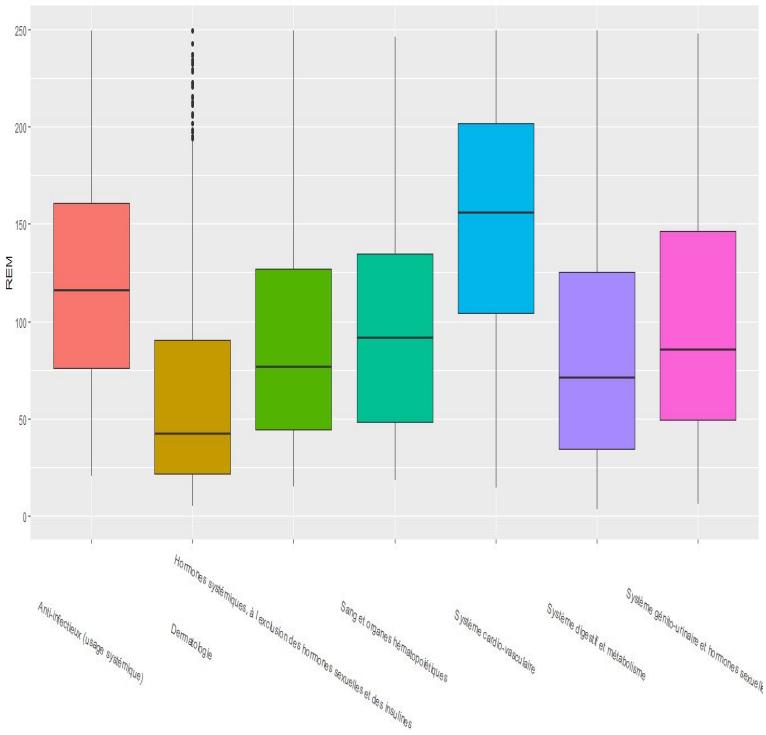
5.1 Un premier modèle

Voici l'un de nos premier modèle afin de déterminer les quartiles, le minimum, le maximum , écart inter-quartile ainsi que la médiane d'une requête.

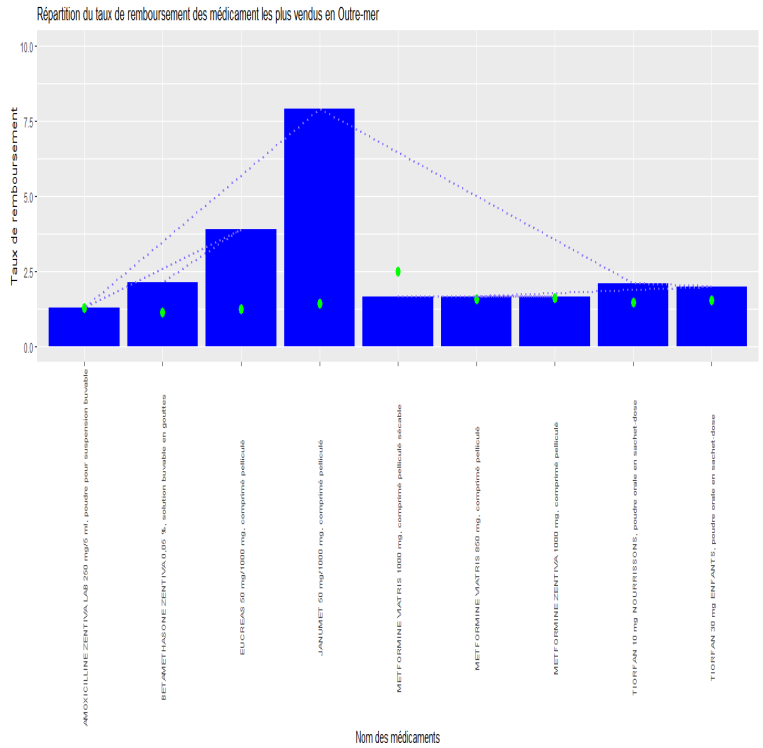


5.2 Quelques exemples de résultats attendus

Boxplot des remboursements par catégorie



Ci-dessous un diragramme représentant le taux de remboursement des médicaments les plus vendus en Outre-mer



Chapitre 6

Discussion

Nous pouvons donc constater que le groupe ciblé, système digestif et métabolisme, est celui qui a le plus grand montant remboursé sur cette année 2021.Cependant nous avons aussi remarqué qu'il y avait une forte relation entre le taux de remboursement et le prix du médicament.

Avec les différentes requetes réalisé, nous avons pu voir aussi l'évolution des prix des médicaments, et les différences de prix selon le type de médicament, le type de zone ciblée ainsi que les différences de remboursement.

Chapitre 7

Conclusion et perspectives

Nous pouvons donc en conclure que à l'aide des plusieurs requêtes et diagrammes que les médicaments les mieux remboursés en 2021 sont ceux traitant le système digestif et le métabolisme. On pourrait aussi dire que le taux de remboursement a une certaine relation avec le prix de vente des médicaments

Nous aurions pu envisager d'étudier cette problématique sur plusieurs années par exemple sur les 5 ou les 10 dernieres années. Mais dans l'avenir nous pouvons penser à faire une projections sur les années futures, des estimations basées sur les années actuelles.

Bibliographie

GroupReportTemplate “Le Logiciel R” par Pierre Lafaye de Micheaux,Remy Drouilhet et Benoit Liquet Cours de

Rstudio dispensé à l’Université Paul-Valery 3 de Montpellier Cours et TD de Programmation Web par Sandra

Bringuay,dispensé à l’Université Paul-Valery 3 de Montpellier R Markdown Cheat Sheet

Annexes

Codes

```
ggplot(df3, aes(x=NomMedoc)) +
  geom_bar(aes(y=coutApresRemboursement), fill='blue', stat="identity") +
  scale_y_continuous(limits = c(0,10))+
  geom_point(aes(y=(NbBoites/10000)), color = rgb(0, 1, 0), pch=16, size=3) +
  geom_path(aes(y=coutApresRemboursement, group=1), colour="slateblue1", lty=3, size=0.9) +
  theme(axis.text.x = element_text(angle=90, vjust=0.6,size=7)) +
  labs(title = "Répartition du taux de remboursement des médicament les plus vendus en Outre-mer",  x = 'Nom des m
```

```
  mutate(class = fct_reorder(l_ATC1, REM, .fun='length' )) %>%
  ggplot( aes(x=l_ATC1, y=REM, fill=l_ATC1)) +
  geom_boxplot() +
  scale_y_continuous(limits = c(0,250))+
  xlab("l_ATC1") +
  theme(legend.position="none",axis.text.x = element_text(angle=-20, vjust=0.6,size=10)) +
  xlab("") +
  xlab("")
```

Tables

Nous n'avons aucun tableaux en supplément à afficher.