Analyse de données médico-administratives avec R: Exploration des données Open MEDIC et Open DAMIR

Nolwenn Le Meur - EHESP

Sequence 2 - Mai-Juin 2023

Pour ces exercices je vous propose de travailler directement sous le format Rmarkdown.

I. Exploration des données Open MEDIC

L'offre de données Open Medic est constituée d'un ensemble de bases annuelles, portant sur l'usage du médicament, délivré en pharmacie de ville (2018-2021). Toutes les données sont extraites du système national interrégimes de l'Assurance Maladie (Sniiram). Les données sur le médicament sont restituées au travers de la classification ATC.

Vous disposez d'un jeux de données Open_Medic de 2021 qui permet d'étudier les dépenses annuelles de médicaments (montants remboursés - REM - et remboursables - BSE) ainsi que le nombre de boîtes délivrées, en fonction d'éléments descriptifs sur les bénéficiaires (tranche d'âge, sexe, région de résidence selon la nouvelle nomenclature Insee) ou de l'information sur la spécialité du prescripteur.

Le traitement des données a été opéré de manière à garantir la confidentialité des informations sur les bénéficiaires ainsi que sur les professionnels de santé. Notamment, certaines modalités ont été floutées (par les valeurs inconnues 9,99,999,etc..) lorsque le seuil critique de 10 bénéficiaires n'était pas respecté.

source: http://open-data-assurance-maladie.ameli.fr/medicaments/index.php

1. Lecture des données Open MEDIC

Vous disposez du fichier "OPEN_MEDIC_2021.CSV" pour étudier la consommation médicamenteuse dans votre région. https://assurance-maladie.ameli.fr/etudes-et-donnees/open-medic-base-complete-depenses-medicaments-2021

a. Dans RStudio, créez un nouveau projet R et fichier Rmarkdown du nom de *"openMedic.Rmd"*.

b. Utilisez tour à tour les fonctions *read.csv()* et *fread()* de la library *data.table* pour lire les données du fichier "OPEN_MEDIC_2021.CSV". Faites en sorte de créer 2 objets R: *med2021_csv* et *med2021*, respectivement. Quelles différences faites vous?

c. Supprimez l'objet *med2021_csv* créé par la fonction *read.csv()* en utilisant la fonction *rm()*.

d. Observez le type de l'objet *med2021* issu de la fonction *fread()* en utilisant la fonction *str()*. Est-ce que les différentes variables sont au bon format?

Les lignes de commandes ci-dessous peuvent vous être utiles pour corriger quelques problèmes d'importation de format de données

d. Utilisez la fonction *summary()* pour un rapide résumé statistique des variables. Que remarquez vous?

e. Pour régulation (erreur de remboursements), l'assurance maladie enregistre des retraits de boîtes d'où les valeurs négatives qui doivent être supprimées

J'ai 1820538 enregistrements

Après nettoyage des boîtes en erreur j'ai 1820538enregistrements

2. Description de l'âge des consommants

a. Documentez la variable **age** en la transformant au format *factor* et en ajoutant les étiquettes (labels): "0-19 ANS","20-59 ANS","60 ANS ET +" et "AGE INCONNU".

b. Que representent les graphiques issus des 2 commandes qui suivent? Quelles sont les différences entre les 2 approches?

La fonction ci-dessous dénombre les patients dans chaque classe d'âge et dans chaque région.

```
# Frequence absolue
med2021 %>% group_by(BEN_REG) %>% count(age)
```

c. Complétez la fonction pour ajouter la fréquence relative au tableau

d. Dénombrez pour chaque sexe, les patients dans chaque classe d'âge (fréquence absolue et relative)

e. Dénombrez pour chaque groupe d'âge, les patients de chaque sexe (fréquence absolue et relative)

3. Décrire la prescription d'anxiolytiques dans votre région.

Votre objectif est de décrire le niveau de prescription d'anxiolytiques (ATC3 = N05B) par les médecins généralistes libéraux dans votre région.

a. Sélectionnez les données de votre région (exemple Bretagne code 53) en utilisant l'indexation ou la fonction *filter()*

b. Sélectionnez le sous-ensemble de prescription d'anxiolytiques (ATC3 = N05B)

c. Sélectionnez le sous-ensemble de médicament prescrit par des médecins généralistes libéraux (PSP_SPE == 1)

d. Combinez l'ensemble des commandes a-b-c en une seule ligne

e. Pour chaque sexe, calculez le nombre de boîtes prescriptes par classe d'âge

f. Utilisez la library *gtsummary* pour faire un tableau qui résume le niveau de prescriptions des médecins généralistes statistique chez les hommes et les femmes et par classe d'âge. Essayez d'afficher, les moyennes, les écart-types et intervalle de confiance à la place des médianes et intervalles inter-quartiles.

4. Niveau de consommations par âge et par sexe

a. Utilisez la librarie *ggplot2* et la fonction *geom_bar()* pour représenter vos résultats. Utilisez la page d'aide pour embellir votre graphique

b. Refaites votre graphique avec les barres adjacentes

c. Refaites votre graphique avec des proportions

5. Evolution des ventes

Vous désirez visualiser l'évolution des ventes sur plusieurs années. Votre data manager vous demande pour cela de vérifier son code avant de récupérer les données simplifiées pour les années 2018 à 2021.

a. Votre data manager a utilisé une boucle *for()* et sauvegardé les données dans l'objet R *openMed*. Que fait cette boucle *for()*.

```
openMed <- c()
for(i in 2018:2021){
    filename2open <- paste("../openMedic/OPEN_MEDIC_", i, ".CSV",sep="")
    temp <- fread(filename2open, header=T, dec = ",")
    temp <- temp[, -seq(2,12,2)]
    if(i==2019){
      colnames(temp)[10] <- "sexe"
    }
    temp <- cbind(temp,"ANNEE"=i)
  openMed <- rbind(openMed, temp)
}
# verification
# table(openMed$ANNE)
```

b. Que fait le code chunck suivant et pourquoi dans cet ordre?

```
openMed53 <- openMed %>% filter(BEN_REG == 53 & ATC3 == "N05B" & PSP_SPE%in%c(1,17))
openMed53$age <- factor(openMed53$age,
                        labels=c("0-19 ANS","20-59 ANS","60 ANS ET +"))
openMed53$sexe <- factor(openMed53$sexe, labels=c("Homme","Femme"))
# save(openMed53, file="openMed53.Rdata")
```

c. Calculez le nombre de boîtes prescriptes par classe d'âge, par année, par sexe et par spécialité de prescripteur avec la fonction *summarise()* de *dplyr*.

d. Représentez graphiquement l'évolution des ventes au cours des 3 années par classe d'âge, par sexe et par spécialité de prescripteur en utilisant la librairie *ggplot2*

II. Exploration des données Open DAMIR

1. Acte d'echographie ccam à l'hospital en janvier 2021

Vous vous intérrogez sur l'activité des services d'échographie en établissement hospitalier. Pour cette analyse vous disposez des fichiers open DAMIR mis à disposition mensuellement par l'assurance maladie sur le site open data du gourvernement (sources: https://www.data.gouv.fr/fr/datasets/open-damir-base-complete-sur-les-depenses-dassurance-maladie-inter-regimes/)

Dans un premier temps vous découvrez la structure des fichiers. a. Lire les 10 premières lignes du fichier *"A202101.csv"* grâce à la fonction *fread()* et observer sa structure.

a.Vous n'avez pas besoin de l'ensemble des variables. Relire le fichier (ensemble des lignes) en sélectionnant les variables:

- AGE_BEN_SNDS
- BEN_RES_REG
- BEN_SEX_COD
- ETE_CAT_SNDS
- ETE_REG_COD
- PRS_ACT_QTE
- PRS_NAT
- SOI_ANN
- SOI_MOI

c. Selectionnez la sous-population des prescriptions hospitalières relatives (ETE_CAT_SNDS: "1101","1102") en utilisant *filter()* de la library *dplyr*

d. Sélectionnez la sous-population des échographies réalisées en 2021 (PRS_NAT: 1324) en utilisant *filter()* de la library *dplyr*

e. Dénombrer les actes d'échographie par région

2. Evolution du volume et du taux d'actes d'echographie ccam à l'hospital pour le premier trimestre 2021

a. Vous souhaitez les mêmes statistiques pour les 3 premiers mois de l'année 2021. Que fait la code chunk ci-dessous?

```
hecho <- c()
filepath <- c("../openDamir/A202101.csv.gz",
              "../openDamir/A202102.csv.gz", "../openDamir/A202103.csv.gz")
for (i in 1:3){
  m <- fread(filepath[i], select = c("AGE_BEN_SNDS", "BEN_RES_REG", "BEN_SEX_COD",
                                     "ETE_CAT_SNDS", "ETE_REG_COD", "PRS_ACT_QTE",
                                   "PRS_NAT","SOI_ANN","SOI_MOI"))
  hosp <- m %>% filter(ETE_CAT_SNDS%in%c("1101","1102"))
  temp <- hosp %>% filter(PRS_NAT==1324, SOI_ANN==2021)
  hecho <- rbind(hecho, temp)
}
hecho$SOI_MOI <- factor(hecho$SOI_MOI, labels=c("JAN", "FEV", "MAR"))
hecho$BEN_SEX_COD <- factor(hecho$BEN_SEX_COD, labels=c("Hommes", "Femmes"))
save(hecho, file="hecho.RData")
```

b.Pour optimiser le temps de traitement, nous vous avons générer la base *hecho*. Charger la base *hecho* dans votre environnement avec la fonction *load()*

b. Résumer le nombre d'échographies prescrites par les établissements par région et par mois en utilisant les fonctions *group_by()* et *summarise()* de la library dplyr

c. Résumer le nombre d'échographies prescrites par les établissements par region, par mois et par sexe en utilisant les fonctions *group_by()* et *summarise()* de la library dplyr

d. Visualiser ces données grâce aux méthodes de la librarie *ggplot2*.

e. Vous souhaitez standardiser les données par rapport à la taille de la population par région. Vous disposez du fichier *pop-reg.csv* dans ce but. Calculez les taux d'échographie par région, par sexe et par chaque mois (jouer avec les variables qualitatives pour obtenir des regards différents sur la distribution des données).