DATAFRAME TUTO

Maxime & Lucas

I. CREATION ET MANIPULATION D'UNE DATAFRAME

Creation du dataframe (est un tableau qui ressemble a une matrice mais dans lequel on peut avoir differents types de tableaux de donnees avec des numeriques et des caracteres)

```
Nom <- c("Aline","Bertrand","Charlie","Adrien") #Creation d'un vecteur de chaines de caracteres
Rang <- matrix(c(1,3,2,4), nrow=4,ncol=1) #Création d'une matrice de 4 lignes et une colonne
Sexe <- c("F","H","H","H") #chaines de caractere et on attribue a Sexe les variables qualitatives
Sex <- factor(Sexe) #on crée le facteur Sex comptenant les valeurs de notre vecteur Sexe
mondata <- data.frame(Nom,Rang,Sex) #création du dataframe a partir des deux vecteurs et de la matrice

str(mondata)#affichage des types de nos differentes variables
```

```
## 'data.frame':    4 obs. of  3 variables:
##  $ Nom : Factor w/ 4 levels "Adrien","Aline",..: 2 3 4 1
##  $ Rang: num  1 3 2 4
##  $ Sex : Factor w/ 2 levels "F","H": 1 2 2 2
```

```
View(mondata)#visualisation en tableau du dataframe crée
summary (mondata) #Aperçu generale sur les stats du dataframe
```

```
##        Nom         Rang      Sex
##  Adrien  :1   Min.   :1.00   F:1
##  Aline   :1   1st Qu.:1.75   H:3
##  Bertrand:1   Median :2.50
##  Charlie :1   Mean   :2.50
##               3rd Qu.:3.25
##               Max.   :4.00
```

II.TRAITEMENT DES DONNEES

1. IMPORT DATAFRAME

```
#Import du fichier csv
data<- read.csv("C:/Users/allak/Desktop/PSB Cours/Mes cours/Maths pour le Big Data et programmation R/j
```

2. STRUCTURE DES DONNEES

```
str(data)
```

```
## 'data.frame':    1058 obs. of  12 variables:
##  $ genre    : Factor w/ 2 levels "F","H": 2 2 2 2 1 1 2 2 2 2 ...
```

```
##  $ age      : int  32 32 33 33 34 34 35 35 36 36 ...
##  $ poids    : Factor w/ 307 levels "105","108","115",..: 298 298 295 295 293 293 289 289 283 283 ...
##  $ taille   : int  186 186 185 185 184 184 183 183 182 182 ...
##  $ caucasien: logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ Cpulm    : Factor w/ 104 levels "1,31","1,32",..: 51 51 51 51 51 51 51 51 51 51 ...
##  $ fumeur   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ sportif  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ urbain   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ obesite  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ amiante  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Malade   : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
head(data)#Les 6 premieres lignes y compris les titres
```

```
##   genre age poids taille caucasien Cpulm fumeur sportif urbain obesite amiante
## 1     H  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 2     H  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 3     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 4     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 5     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 6     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
##   Malade
## 1  FALSE
## 2  FALSE
## 3  FALSE
## 4  FALSE
## 5  FALSE
## 6  FALSE
```

```
data[1050 : 1058,] #extraction des 8 dernieres lignes
```

```
##      genre age poids taille caucasien Cpulm fumeur sportif urbain obesite
## 1050     F  49    90    183      TRUE   1,9   TRUE    TRUE   TRUE   FALSE
## 1051     F  49    79    173      TRUE   1,9   TRUE    TRUE   TRUE   FALSE
## 1052     H  49    92    183      TRUE   1,9   TRUE   FALSE   TRUE   FALSE
## 1053     H  49    92    183      TRUE   1,9   TRUE    TRUE   TRUE   FALSE
## 1054     H  74  63,8    184      TRUE  1,94   TRUE    TRUE  FALSE   FALSE
## 1055     F  74  65,5    178      TRUE  1,92   TRUE   FALSE  FALSE   FALSE
## 1056     F  74    70    185      TRUE  2,11   TRUE    TRUE   TRUE   FALSE
## 1057     F  74  66,4    178     FALSE  2,19   TRUE   FALSE   TRUE   FALSE
## 1058     F  62  73,4    179      TRUE   2,1   TRUE   FALSE  FALSE   FALSE
##      amiante Malade
## 1050   FALSE   TRUE
## 1051   FALSE   TRUE
## 1052   FALSE   TRUE
## 1053   FALSE   TRUE
## 1054   FALSE   TRUE
## 1055   FALSE   TRUE
## 1056   FALSE   TRUE
## 1057   FALSE   TRUE
## 1058    TRUE   TRUE
```

3. RESUME STATISTIQUES ET TABLE DE CONTINGENCE

```r
summary(data)#permet d'obtenir un resume statistiques
```

```
##  genre        age             poids          taille      caucasien
##  F:511   Min.   : 32.0   88     : 48   Min.   :147.0   Mode :logical
##  H:547   1st Qu.: 61.0   93     : 26   1st Qu.:173.0   FALSE:194
##          Median : 66.0   92     : 24   Median :178.0   TRUE :864
##          Mean   : 63.2   80     : 20   Mean   :176.9
##          3rd Qu.: 69.0   82     : 17   3rd Qu.:182.0
##          Max.   :110.0   85     : 13   Max.   :192.0
##                          (Other):910
##      Cpulm       fumeur          sportif          urbain          obesite
##  2      :183   Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  1,9    : 76   FALSE:668       FALSE:379       FALSE:233       FALSE:1041
##  1,95   : 23   TRUE :390       TRUE :679       TRUE :825       TRUE :17
##  2,04   : 22
##  2,07   : 21
##  2,05   : 20
##  (Other):713
##   amiante          Malade
##  Mode :logical   Mode :logical
##  FALSE:1046      FALSE:636
##  TRUE :12        TRUE :422
##
##
##
##
```

```r
mean(data$age)#moyenne age
```

```
## [1] 63.1966
```

```r
table(data$Malade.data$fumeur) #table de contingence
```

```
## < table of extent 0 >
```

4. COPIE DES COLONNES D'UN DATAFRAME ET MODIFICATIONS

```r
attach(data)#copie de notre dataframe
head(age)
```

```
## [1] 32 32 33 33 34 34
```

```r
age[4]<-34 #modification de l'element  la 4eme position
head(age)
```

```
## [1] 32 32 33 34 34 34
```

```r
head(data)#base de donnes initiale non modifie
```

```
##   genre age poids taille caucasien Cpulm fumeur sportif urbain obesite amiante
## 1     H  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 2     H  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 3     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 4     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 5     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 6     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
##   Malade
## 1  FALSE
## 2  FALSE
## 3  FALSE
## 4  FALSE
## 5  FALSE
## 6  FALSE
```

```
head(age)
```

```
## [1] 32 32 33 34 34 34
```

5. MODIFICATION D'UNE VALEUR DE LA TABLE DE DONNEES

```
head(data)
```

```
##   genre age poids taille caucasien Cpulm fumeur sportif urbain obesite amiante
## 1     H  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 2     H  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 3     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 4     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 5     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 6     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
##   Malade
## 1  FALSE
## 2  FALSE
## 3  FALSE
## 4  FALSE
## 5  FALSE
## 6  FALSE
```

```
data[2,]$genre<-"F" #modification 2eme ligne et 1ere colonne en Femme
data[1,1]<-"F" #modification 1ere ligne et 1ere colonne en Femme
head(data)
```

```
##   genre age poids taille caucasien Cpulm fumeur sportif urbain obesite amiante
## 1     F  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 2     F  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 3     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 4     H  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 5     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 6     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
##   Malade
## 1  FALSE
## 2  FALSE
```

```
## 3  FALSE
## 4  FALSE
## 5  FALSE
## 6  FALSE
```

```
data[1:6,1]<-"F"
head(data)#modification des 6premieres lignes de la 1ere colonne
```

```
##   genre age poids taille caucasien Cpulm fumeur sportif urbain obesite amiante
## 1     F  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 2     F  32    88    186      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 3     F  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 4     F  33    87    185      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 5     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
## 6     F  34    86    184      TRUE     2  FALSE    TRUE   TRUE   FALSE   FALSE
##   Malade
## 1  FALSE
## 2  FALSE
## 3  FALSE
## 4  FALSE
## 5  FALSE
## 6  FALSE
```

6. AJOUT D'UNE OU PLUSIEURS COLONNES A LA TABLE DE DONNEES

```
str(data$poids)
```

```
##  Factor w/ 307 levels "105","108","115",..: 298 298 295 295 293 293 289 289 283 283 ...
```

```
new_poids <- as.numeric(as.factor(data$poids))#Changement du caratere 'factor' de la variable poids en
new_taille = data$taille/100 #Creation d'une nouvelle variable qui divise la taille en 100
IMC = new_poids/(new_taille^2)#Creation d'une variable "Indice de Masse Corporelle" qui divise le poids
new_data<-cbind(data,new_poids, new_taille,IMC) #on cre des nouvelles colonnes avec les nouvelles varia
```

7. RENOMMER UNE COLONNE

```
colnames(new_data) #affiche tous les noms des colonnes
```

```
##  [1] "genre"      "age"        "poids"      "taille"     "caucasien"
##  [6] "Cpulm"      "fumeur"     "sportif"    "urbain"     "obesite"
## [11] "amiante"    "Malade"     "new_poids"  "new_taille" "IMC"
```

```
colnames(new_data)[14]<- "nouvelle_taille" # On a chang le nom de la colonne 'new_taille' par 'nouvelle
colnames(new_data)[15]<- "Indice de Masse Corporelle" #pareillement pour 'IMC' par 'Indice de masse cor
colnames(new_data)
```

```
##  [1] "genre"                     "age"
##  [3] "poids"                     "taille"
##  [5] "caucasien"                 "Cpulm"
##  [7] "fumeur"                    "sportif"
```

```
##  [9] "urbain"                    "obesite"
## [11] "amiante"                   "Malade"
## [13] "new_poids"                 "nouvelle_taille"
## [15] "Indice de Masse Corporelle"
```

8. SUPPRESSION D'UNE COLONNE

```
#on va creer une colonne nomm 'Tuto' vide NA et ensuite la supprimer
Tuto = age/2
new_data2 = cbind(new_data,Tuto) #on a ajouter une colonne dans notre nouvelle dataframe
colnames(new_data2)# Tuto est bien parmi les noms des colonnes
```

```
##  [1] "genre"                     "age"
##  [3] "poids"                     "taille"
##  [5] "caucasien"                 "Cpulm"
##  [7] "fumeur"                    "sportif"
##  [9] "urbain"                    "obesite"
## [11] "amiante"                   "Malade"
## [13] "new_poids"                 "nouvelle_taille"
## [15] "Indice de Masse Corporelle" "Tuto"
```

```
new_data2[,16]<-NULL #Suppresion de la 16eme colonne 'Tuto'
colnames(new_data2)
```

```
##  [1] "genre"                     "age"
##  [3] "poids"                     "taille"
##  [5] "caucasien"                 "Cpulm"
##  [7] "fumeur"                    "sportif"
##  [9] "urbain"                    "obesite"
## [11] "amiante"                   "Malade"
## [13] "new_poids"                 "nouvelle_taille"
## [15] "Indice de Masse Corporelle"
```