

DATAFRAME TUTO

Maxime & Lucas

I. Créer une dataframe R vs Python

Comment créer un tableau qui contient différents type de données.

```
Rang <- matrix(c(1,3,2,4), nrow=4, ncol=1)
Nom <- c("Aline", "Bertrand", "Charlie", "Adrien")
Age <- c(15,20,58,32)
Sexe <- c("F", "H", "H", "H")
Sex <- factor(Sexe)
df <- data.frame(Rang, Nom, Sex, Age)
```

df

	Rang <dbl>	Nom <chr>	Sex <fctr>	Age <dbl>
	1	Aline	F	15
	3	Bertrand	H	20
	2	Charlie	H	58
	4	Adrien	H	32

4 rows

```
import pandas as pd

Nom=["Aline", "Bertrand", "Charlie", "Adrien"]
Sexe=["F", "H", "H", "H"]
Age=[15,20,58,32]
df=pd.DataFrame({"Nom":Nom, "Sexe":Sexe, "Age":Age})
```

head(py\$df)

	Nom <chr>	Sexe <chr>	Age <dbl>
1	Aline	F	15
2	Bertrand	H	20
3	Charlie	H	58
4	Adrien	H	32

4 rows

II.Traitement de données

1. Import de Dataframe

```
data<-read.csv("donnees.csv",header=TRUE,encoding="UTF-8")
```

```
data=pd.read_csv("donnees.csv",encoding="UTF-8")
```

genre	age	poids	taille	caucasien	Cpulm	fumeur	sportif	urbain	obesite	amiante	Malade
H	32	88	186	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
H	32	88	186	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
H	33	87	185	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
H	33	87	185	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
F	34	86	184	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
F	34	86	184	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE

2. Structure de données

```
str(data)
```

```
## 'data.frame':  1058 obs. of  12 variables:
## $ genre      : chr  "H" "H" "H" "H" ...
## $ age        : int   32 32 33 33 34 34 35 35 36 36 ...
## $ poids      : chr  "88" "88" "87" "87" ...
## $ taille     : int   186 186 185 185 184 184 183 183 182 182 ...
## $ caucasien  : logi   TRUE TRUE TRUE TRUE TRUE TRUE ...
## $ Cpulm      : chr   "2" "2" "2" "2" ...
## $ 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 ...
```

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

	genre <chr>	a... <int>	poids <chr>	taille <int>	caucasien <lg>	Cpulm <chr>
1050	F	49	90	183	TRUE	1,9
1051	F	49	79	173	TRUE	1,9
1052	H	49	92	183	TRUE	1,9
1053	H	49	92	183	TRUE	1,9
1054	H	74	63,8	184	TRUE	1,94
1055	F	74	65,5	178	TRUE	1,92
1056	F	74	70	185	TRUE	2,11
1057	F	74	66,4	178	FALSE	2,19
1058	F	62	73,4	179	TRUE	2,1

9 rows | 1-7 of 13 columns

```
data.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 1058 entries, 0 to 1057
## Data columns (total 12 columns):
## #  Column      Non-Null Count  Dtype
## ---  ---
## 0  genre      1058 non-null  object
## 1  age        1058 non-null  int64
## 2  poids      1058 non-null  object
## 3  taille     1058 non-null  int64
## 4  caucasien  1058 non-null  bool
## 5  Cpulm      1058 non-null  object
## 6  fumeur     1058 non-null  bool
## 7  sportif    1058 non-null  bool
## 8  urbain     1058 non-null  bool
## 9  obesite    1058 non-null  bool
## 10 amiante  1058 non-null  bool
## 11 Malade  1058 non-null  bool
## dtypes: bool(7), int64(2), object(3)
## memory usage: 48.7+ KB
```

py\$data[1050:1058,]

	genre <chr>	a... <dbl>	poids <chr>	taille <dbl>	caucasien <lg>	Cpulm <chr>
1050	F	49	90	183	TRUE	1,9
1051	F	49	79	173	TRUE	1,9
1052	H	49	92	183	TRUE	1,9
1053	H	49	92	183	TRUE	1,9
1054	H	74	63,8	184	TRUE	1,94
1055	F	74	65,5	178	TRUE	1,92
1056	F	74	70	185	TRUE	2,11
1057	F	74	66,4	178	FALSE	2,19
1058	F	62	73,4	179	TRUE	2,1

9 rows | 1-7 of 13 columns

rgzvrvtetrtbzetrtznet

3. Résumé statistiques

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

##	genre	age	poids	taille
----	-------	-----	-------	--------

```
## Length:1058      Min.   : 32.0      Length:1058      Min.   :147.0
## Class :character  1st Qu.: 61.0      Class :character  1st Qu.:173.0
## Mode :character   Median : 66.0      Mode :character   Median :178.0
##                Mean  : 63.2      Mean :176.9
##                3rd Qu.: 69.0      3rd Qu.:182.0
##                Max.   :110.0      Max.   :192.0
## caucasien        Cpulm          fumeur        sportif
## Mode :logical    Length:1058      Mode :logical     Mode :logical
## FALSE:194        Class :character  FALSE:668         FALSE:379
## TRUE :864         Mode :character   TRUE :390         TRUE :679
##
##
##
## urbain          obeseite        amiante        Malade
## Mode :logical    Mode :logical    Mode :logical     Mode :logical
## FALSE:233        FALSE:1041       FALSE:1046        FALSE:636
## TRUE :825         TRUE :17         TRUE :12          TRUE :422
##
##
##
```

```
data.describe()
```

```
##          age      taille
## count 1058.000000 1058.000000
## mean   63.196597  176.942344
## std     9.452020   5.954215
## min    32.000000  147.000000
## 25%    61.000000  173.000000
## 50%    66.000000  178.000000
## 75%    69.000000  182.000000
## max    110.000000 192.000000
```

```
data["sportif"].value_counts()
```

```
## True      679
## False     379
## Name: sportif, dtype: int64
```

```
def count(val):
    for i in val :
        print(data[i].value_counts())

count(["fumeur","sportif"])
```

```
## False      668
## True       390
## Name: fumeur, dtype: int64
## True       679
## False      379
## Name: sportif, dtype: int64
```

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

```
data[2,$genre<-"F"
data[1,1]<-"F"
```

```
data.loc[1,"genre"] = "F"
data.iloc[0,0] = "F"
```

```
head(data)
```

```
head(py$data)
```

	genre	age	poids	taille	caucasien	Cpulm	
	<chr>	<int>	<chr>	<int>	<lg>	<chr>	
1	F	32	88	186	TRUE	2	
2	F	32	88	186	TRUE	2	
3	H	33	87	185	TRUE	2	
4	H	33	87	185	TRUE	2	
5	F	34	86	184	TRUE	2	
6	F	34	86	184	TRUE	2	

6 rows | 1-7 of 13 columns

	genre	age	poids	taille	caucasien	Cpulm	
	<chr>	<dbl>	<chr>	<dbl>	<lg>	<chr>	
1	F	32	88	186	TRUE	2	
2	F	32	88	186	TRUE	2	
3	H	33	87	185	TRUE	2	
4	H	33	87	185	TRUE	2	
5	F	34	86	184	TRUE	2	
6	F	34	86	184	TRUE	2	

6 rows | 1-7 of 13 columns

commentaire sur les

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

```
taille_m = data$taille/100
n_data<-cbind(data,taille_m)
#head(n_data)
```

```
taille_m = data["taille"]/100
data["taille_m"] = taille_m
#data.head()
```

genre	age	poids	taille	caucasien	Cpulm	fumeur	sportif	urbain	obesite	amiante	Malade	taille_m
F	32	88	186	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	1.86
F	32	88	186	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	1.86
H	33	87	185	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	1.85
H	33	87	185	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	1.85
F	34	86	184	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	1.84
F	34	86	184	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	1.84

6. SUPPRESSION D'UNE COLONNE

```
n_data <-subset(n_data, select=c(taille_m))
```

```
data = data.drop(["taille_m"], axis=1)
```

genre	age	poids	taille	caucasien	Cpulm	fumeur	sportif	urbain	obesite	amiante	Malade
F	32	88	186	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
F	32	88	186	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
H	33	87	185	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
H	33	87	185	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
F	34	86	184	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
F	34	86	184	TRUE	2	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE