

Base de données

à quoi ça doit ressembler ?

Benjamin Louis

15/10/2019 (MàJ: 18/11/2021)

Disclaimer !

Dans cette présentation :

Base de données \neq SQL

Base de données = Tableau de données

Beaucoup du contenu de cette présentation provient de :

<https://thinkr.fr/base-de-donnees-reussie/>

by



Tableau de données, kezako ?

id	variable_1	variable_2	variable_3	variable_4
1				
2				
3		x		
4				
5				
6				
7				

Ligne :

- **1 ligne = 1 observation/individu statistique** (une personne, une parcelle, ...).
- Si répétition --> 1 ligne par répétition. Il est toujours bon d'avoir une colonne avec un identifiant unique pour chaque observation.

Colonnes : **1 colonne = 1 variable** (caractéristique mesurable, e.g. sexe, age, rendement, ...)

x : donnée = valeur de la variable pour une observation

UN SEUL TABLEAU !

1 colonne = 1 variable / 1 variable = 1 colonne



id	Femme	Homme
1	oui	non
2	non	oui
3	oui	non
4	oui	non
5	oui	non



id	sexe
1	Femme
2	Homme
3	Femme
4	Femme
5	Femme

1 colonne = 1 variable et 1 variable = 1 colonne



id	type_sport
1	tennis; football
2	escalade
3	
4	aucun sport
5	football; pétanque



id	sport	escalade	football	pétanque	tennis
1	oui	non	oui	non	oui
2	oui	oui	non	non	non
3					
4	non	non	non	non	non
5	oui	non	oui	oui	non

1 colonne = 1 variable et 1 variable = 1 colonne



id	sport	escalade	football	pétanque	tennis
1	oui		x		x
2	oui	x			
3					
4	non				
5	oui		x	x	



id	sport	escalade	football	pétanque	tennis
1	oui	non	oui	non	oui
2	oui	oui	non	non	non
3					
4	non	non	non	non	non
5	oui	non	oui	oui	non

Bonnes pratiques

Un seul en-tête !

✗ NON !

parcelle	Expérimentation			Climat		Mesure	
	repetition	type_sol	traitement	temperature	pluie	matiere_seche	rendement
1	1	b	T1	27	102	43.20745	85.03678
2	1	c	T1	15	110	34.59607	81.80376
3	1	b	T1	34	135	36.75857	84.97888
4	1	a	T1	32	126	49.68792	69.91752
5	1	a	T1	23	173	41.05897	68.79525

✓ OUI !

parcelle	repetition	type_sol	traitement	temperature	pluie	matiere_seche	rendement
1	1	b	T1	27	102	43.20745	85.03678
2	1	c	T1	15	110	34.59607	81.80376
3	1	b	T1	34	135	36.75857	84.97888
4	1	a	T1	32	126	49.68792	69.91752
5	1	a	T1	23	173	41.05897	68.79525

Variable qualitative \neq variable quantitative

 **NON !**

id	sexe
1	1
2	2
3	1
4	1
5	1

 **OUI !**

id	sexe
1	Femme
2	Homme
3	Femme
4	Femme
5	Femme

Un style (couleur, italique, ...) n'est pas une variable !

 **NON !**

id	score
1	8
2	0
3	6
4	1
5	9

 **OUI !**

id	score	diagnostic
1	8	bon
2	0	mauvais
3	6	moyen
4	1	mauvais
5	9	bon

On ne met que les données brutes !



id	age
1	31
2	33
3	37
4	18
5	66
Moyenne	37



id	age
1	31
2	33
3	37
4	18
5	66

Attention aux variables numériques !



id	age
1	31 ans
2	33
3	37
4	08/10/2001
5	66ans



id	age
1	31
2	33
3	37
4	18
5	66

Modalités homogènes !



id	sexe
1	Femme
2	Homme
3	Fille
4	Féminin
5	F



id	sexe
1	Femme
2	Homme
3	Femme
4	Femme
5	Femme

Dates homogènes (et notations universelles) !



id	date
1	08/10/2019
2	1er septembre 1995
3	12 dec 2008
4	01-05-08
5	07/07/07



id	date
1	08/10/2019
2	01/09/1995
3	12/12/2008
4	01/05/2008
5	07/07/2007

Valeurs manquantes homogènes !

 **NON !**

id	age
1	25
2	NA
3	
4	33
5	-

 **OUI !**

id	age
1	25
2	NA
3	NA
4	33
5	NA

Décimales homogènes !

 **NON !**

id	taux
1	25,30
2	
3	12.46
4	33.23
5	5,89

 **OUI !**

id	taux
1	25,30
2	
3	12,46
4	33,23
5	5,89

Les commentaires dans une colonne séparée !



id	sexe
1	Femme
2	Homme
3	Femme (à vérifier)
4	Femme
5	Femme



id	sexe	commentaires
1	Femme	
2	Homme	
3	Femme	à vérifier
4	Femme	
5	Femme	

Base anonyme !



prenom	nom	score	diagnostic
Martine	Louis	8	bon
Alain	Maringot	0	mauvais
Elsa	Durand	6	moyen
Henriette	Laplace	1	mauvais
Paulette	Reynaud	9	bon



id	score	diagnostic
1	8	bon
2	0	mauvais
3	6	moyen
4	1	mauvais
5	9	bon

Noms des colonnes ! (et des modalités)

- **Intelligibles :**
 - ❌ `variable_1`, `tdcs`, `v2`
 - ✅ `age`, `rendement`, `temperature`
- **Attention à la longueur :**
 - ❌ `fiche_35_session_3_jour4_par_temps_de_pluie_score`
 - ✅ `score_f35s3j4`
- **Metadonnées :** fichier supplémentaire qui définit les colonnes (description, unité, description modalités, ...)
- **Format homogène, pas d'espace, pas d'accent :**
 - ❌ `Rendement (Q_ha)`, `sexe`, `scoreJour1`, `taux_carbone`
 - ✅ `rendement`, `sexe`, `score_j1`, `taux_carbone`

cf package `{janitor}` et <http://stat405.had.co.nz/r-style.html>

- **Enlever les colonnes inutiles** (colonnes vides, colonnes ajoutées par logiciel de sondage, ...)