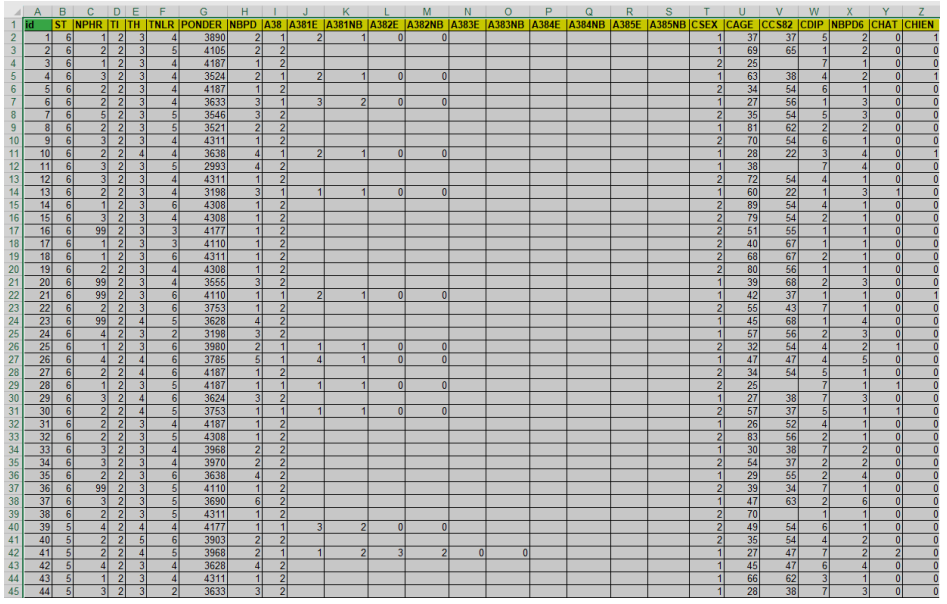
C4. Manipuler les données

Sous les pavés, la plage...

Sous les pavés, la plage...

- Une base de données
  - Variables
  - Observations
- Plage Excel
  - Un bloc de lignes et de colonnes
    - A1:BZ8543
    - = Colonne A, ligne 1 à colonne BZ, ligne 8543
    - Mafeuille!A1:BZ8543
  - Soyons absolu avec la loi du dollar
    - Ex : $A$1:$BZ$8543
- Il faut connaître sa plage



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	id	ST	NPHR	TI	TH	TNLR	PONDER	NBPD	A38	A381E	A381NB	A382E	A382NB	A383E	A383NB	A384E	A384NB	A385E	A385NB	CSEX	CAGE	CCS82	CDIP	NBPD6	CHAT	CHIEN
2	1	6	1	2	3	4	3890	2	1	2	1	0	0							1	37	37	5	2	0	1
3	2	6	2	2	3	5	4105	2	2											1	69	65	1	2	0	0
4	3	6	1	2	3	4	4187	1	2											2	25		7	1	0	0
5	4	6	3	2	3	4	3524	2	1	2	1	0	0							1	63	38	4	2	0	1
6	5	6	2	2	3	4	4187	1	2											2	34	54	6	1	0	0
7	6	6	2	2	3	4	3633	3	1	3	2	0	0							1	27	56	1	3	0	0
8	7	6	5	2	3	5	3546	3	2											2	35	54	5	3	0	0
9	8	6	2	2	3	5	3521	2	2											1	81	62	2	2	0	0
10	9	6	3	2	3	4	4311	1	2											2	70	54	6	1	0	0
11	10	6	2	2	4	4	3638	4	1	2	1	0	0							1	28	22	3	4	0	1
12	11	6	3	2	3	5	2993	4	2											1	38		7	4	0	0
13	12	6	3	2	3	4	4311	1	2											2	72	54	4	1	0	0
14	13	6	2	2	3	4	3198	3	1	1	1	0	0							1	60	22	1	3	1	0
15	14	6	1	2	3	6	4308	1	2											2	89	54	4	1	0	0
16	15	6	3	2	3	4	4308	1	2											2	79	54	2	1	0	0
17	16	6	99	2	3	3	4177	1	2											2	51	55	1	1	0	0
18	17	6	1	2	3	3	4110	1	2											2	40	67	1	1	0	0
19	18	6	1	2	3	6	4311	1	2											2	68	67	2	1	0	0
20	19	6	2	2	3	4	4308	1	2											2	80	56	1	1	0	0
21	20	6	99	2	3	4	3555	3	2											1	39	68	2	3	0	0
22	21	6	99	2	3	6	4110	1	1	2	1	0	0							1	42	37	1	1	0	1
23	22	6	2	2	3	6	3753	1	2											2	55	43	7	1	0	0
24	23	6	99	2	4	5	3628	4	2											1	45	68	1	4	0	0
25	24	6	4	2	3	2	3198	3	2											1	57	56	2	3	0	0
26	25	6	1	2	3	6	3980	2	1	1	1	0	0							2	32	54	4	2	1	0
27	26	6	4	2	4	6	3785	5	1	4	1	0	0							1	47	47	4	5	0	0
28	27	6	2	2	4	6	4187	1	2											2	34	54	5	1	0	0
29	28	6	1	2	3	5	4187	1	1	1	1	0	0							2	25		7	1	1	0
30	29	6	3	2	4	6	3624	3	2											1	27	38	7	3	0	0
31	30	6	2	2	4	5	3753	1	1	1	1	0	0							2	57	37	5	1	1	0
32	31	6	2	2	3	4	4187	1	2											1	26	52	4	1	0	0
33	32	6	2	2	3	5	4308	1	2											2	83	56	2	1	0	0
34	33	6	3	2	3	4	3968	2	2											1	30	38	7	2	0	0
35	34	6	3	2	3	4	3970	2	2											2	54	37	2	2	0	0
36	35	6	2	2	3	6	3638	4	2											1	29	55	2	4	0	0
37	36	6	99	2	3	5	4110	1	2											2	39	34	7	1	0	0
38	37	6	3	2	3	5	3690	6	2											1	47	63	2	6	0	0
39	38	6	2	2	3	5	4311	1	2											2	70		1	1	0	0
40	39	5	4	2	4	4	4177	1	1	3	2	0	0							2	49	54	6	1	0	0
41	40	5	2	2	5	6	3903	2	2											2	35	54	4	2	0	0
42	41	5	2	2	4	5	3968	2	1	1	2	3	2	0	0					1	27	47	7	2	2	0
43	42	5	4	2	3	4	3628	4	2											1	45	47	6	4	0	0
44	43	5	1	2	3	4	4311	1	2											1	66	62	3	1	0	0
45	44	5	3	2	3	2	3633	3	2											1	28	38	7	3	0	0

Trouver le début et fin de fichier

- PC
  - Début
    - Ctrl+début (touche ↖)
  - Fin
    - Ctrl+Fin
- Mac
  - Début
    - Ctrl + fn + ←
  - Fin
    - Ctrl + fn + →

- Si vous utilisez des « volets », le raccourci pour aller au début, ne va pas au début, mais au début du cadran inférieur droit… terminez la navigation avec les touches flèches.
- S'il y a eu des opérations (supprimées) en dessous de la fin du fichier, la fin de fichier n'est pas forcément à la vraie fin. Dans ce cas, réenregistrer le fichier permet de rétablir la vraie fin.

Sélectionner toute la plage de données

- PC
  - Aller au Début
    - Ctrl + ↖
  - Aller à la Fin en sélectionnant
    - Sélection : Touche ⇧ (appelée Shift ou Maj)
    - ⇧ + Ctrl + Fin
- Mac
  - Aller Début
    - Ctrl + fn + ←
  - Aller à la Fin en sélectionnant
    - ⇧ + Ctrl+ fn + →

Aller à la dernière cellule non-vide dans une direction

- PC
  - Ctrl + flèche
  - Avec sélection de la plage :
    - ⇧ + Ctrl + flèche
- Mac
  - Command + flèche
  - Avec sélection de la plage :
    - ⇧ + Command + flèche
- Ces touches sont très pratiques dans deux cas
  - La colonne est pleine (« sans cellules vides »). On va alors du haut en bas facilement et vice versa
  - La colonne est vide. On va alors facilement du haut vers le bas (mais pas l'inverse)

Créer une variable sans peine (par copier coller)

- 1. Créer sa une nouvelle variable après la dernière variable.
  - Donner un nouveau nom à la variable en ligne 1
  - Écrire sa formule en ligne 2
    - Vérifier qu'elle fonctionne
- 2. Copier la formule
  - PC : Ctrl + C
  - Mac : Command + C
- 3. Aller à la fin du fichier
  - Mac : Ctrl + fn + →
  - PC : Ctrl+fin
- 4. Sélectionner la plage de la colonne où coller sa formule
  - PC : ⇧ + Ctrl + ↑
  - Mac : ⇧ + Command + ↑
- 5. Coller
  - PC : Ctrl + V
  - Mac : Command + V

Des variables

- Recommandé
  - Information sur les variables uniquement dans la première ligne
  - Noms de variable courts (8-12 caractères) et mnémotechniques
  - Tous différents
  - De préférence sans accents et sans espaces
  - Bien différencier variables créées, variables originales
  - Éventuellement distinguer variable quantitative et variable qualitative
- À éviter
  - Variables V1, … V150
  - Intitulé complet de la question
  - Informations sur plusieurs lignes sur le contenu de la variable
    - ⇒ Faire une autre table pour les noms de variables
  - Des colonnes avec noms de variable vide

La variable identifiant

- Variable qui permet d'identifier les observations
  - Valeur différente pour chaque ligne / observation
  - Pas de valeur manquante
  - De préférence : Dont l'ordre correspond à l'ordre de la base
  - De préférence en première colonne
- Ce qu'elle permet de faire
  - De (re)trier → revenir à l'ordre initial
  - De fusionner
  - De trouver le bas avec raccourci clavier
- Si elle n'existe pas : la créer
  - Insérer colonne en A1 → IDENT
  - A2 → 1
  - A3 → =A2+1
  - Copier-coller jusqu'à la dernière ligne
  - Copier colonne IDENT collage spécial « valeurs »
- IDENT Insee versus votre ident
  - Parfois il existe plusieurs identifiants
  - Ménage (ex. colonne A) et individu dans le ménage (ex. colonne B)
  - Concaténer → =A2&B2

Les types de variables

- Les variables catégoriques souvent codées avec des chiffres (gain de place, lisibilité dans la plage)
  - Sexe : 1 → Homme, 2 → Femme. PCS : 10, 21, 22, 23, 31, 33, 34 etc.
  - Attention, certaines opérations (somme, moyenne) peuvent ne pas avoir de sens !
- A la différence des logiciels statistiques : mauvaise distinction entre information caractère et numérique
  - '01 → 01 caractère. Aligné à gauche + onglet vert
  - 01 → 1 : numérique. Aligné à droite
- Importation de fichiers (csv) perte de la distinction entre caractère et numérique

Les valeurs manquantes

- La valeur manquante sous Excel est la cellule vide
  - Excel tient compte des cellules vides pour les calculs → Moyenne(Plage) ; Ecartype(Plage)
  - INSEE : non réponse → '9', '99' etc. Question non posée : vide
- Source de nombreux embarras
  - Rend la circulation difficile dans le fichier par raccourci clavier
  - Non conservé dans les formules
    - Soit E5 : cellule vide
    - Si en F5 formule =E5 → 0
  - Solution formule : =SI(E5="";"";E5)

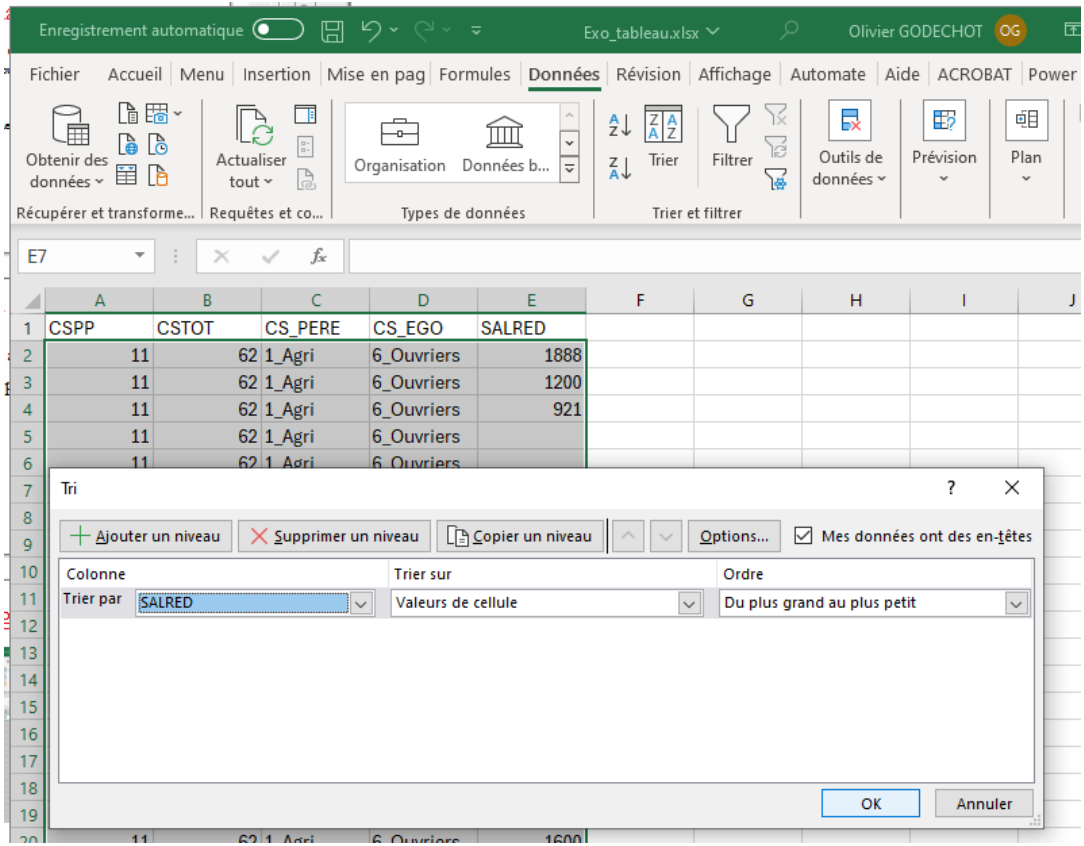
Coder et recoder des variables

- Pourquoi recoder ?
  - Modalités codées avec des chiffres → peu lisible
  - Niveau de détail trop important → regrouper
  - Gestion des non-réponses → supprimer les « 99 », « 98 », etc. qui faussent les moyennes
- Le pain quotidien du travail statistique

- Solution 1.
  Trier + recodage à la main
- Solution 2.
  Rechercher/Remplacer
- Solution 3.
  Formules

Trier + recodage à la main

- Typiquement variable avec de très nombreuses modalités (et éventuellement variation dans les graphies)
- Préalable : Vérifier qu'il existe une clef identification/tri (la créer au besoin)
- Trier la base de données selon l'ordre de la variable (Attention ! Danger !)
  - Sélectionner TOUTE la plage
    - Excel la devine éventuellement avec (« étendre la sélection »), vérifier
  - Menu DONNEES → TRIER



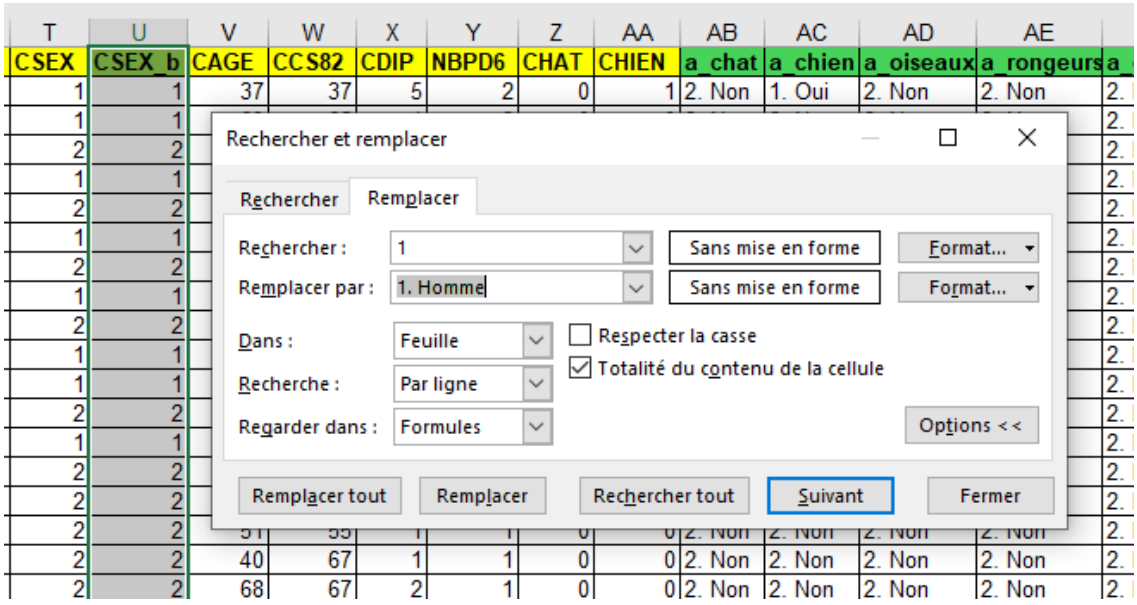
	A	B	C	D	E	F
1	CSPP	CSTOT	CS_PERE	CS_EGO	SALRED	SALRED_t
2	63	56	6_Ouvriers	5_Employés	35	1.<100
3	64	42	6_Ouvriers	4_Prof_Int	40	
4	11	56	1_Agri	5_Employés	50	
5	69	56	6_Ouvriers	5_Employés	50	
6	67	56	6_Ouvriers	5_Employés	60	
7	68	56	6_Ouvriers	5_Employés	60	
8	52	63	5_Employés	6_Ouvriers	64	
9	21	33	2_ArtCom	3_Cadres	64	
10	68	68	6_Ouvriers	6_Ouvriers	70	
11	33	54	3_Cadres	5_Employés	70	
12	37	56	3_Cadres	5_Employés	72	
13	47	68	4_Prof_Int	6_Ouvriers	73	
14	63	62	6_Ouvriers	6_Ouvriers	75	
15	37	56	3_Cadres	5_Employés	80	
16	67	56	6_Ouvriers	5_Employés	80	
17	67	56	6_Ouvriers	5_Employés	80	
18	31	42	3_Cadres	4_Prof_Int	80	
19	46	43	4_Prof_Int	4_Prof_Int	81	
20	53	35	5_Employés	3_Cadres	82	
21	11	67	1_Agri	6_Ouvriers	87	
22	65	56	6_Ouvriers	5_Employés	88	
23	63	68	6_Ouvriers	6_Ouvriers	90	
24	46	56	4_Prof_Int	5_Employés	91	
25	23	67	2_ArtCom	6_Ouvriers	92	
26	63	56	6_Ouvriers	5_Employés	93	
27	23	34	2_ArtCom	3_Cadres	94	
28	63	64	6_Ouvriers	6_Ouvriers	100	

Trier + recodage à la main

- Copier-coller vos modalités par bloc
- Re-trier ensuite par la variable identifiant pour retrouver l'ordre initial

Solution 2. Rechercher/Remplacer

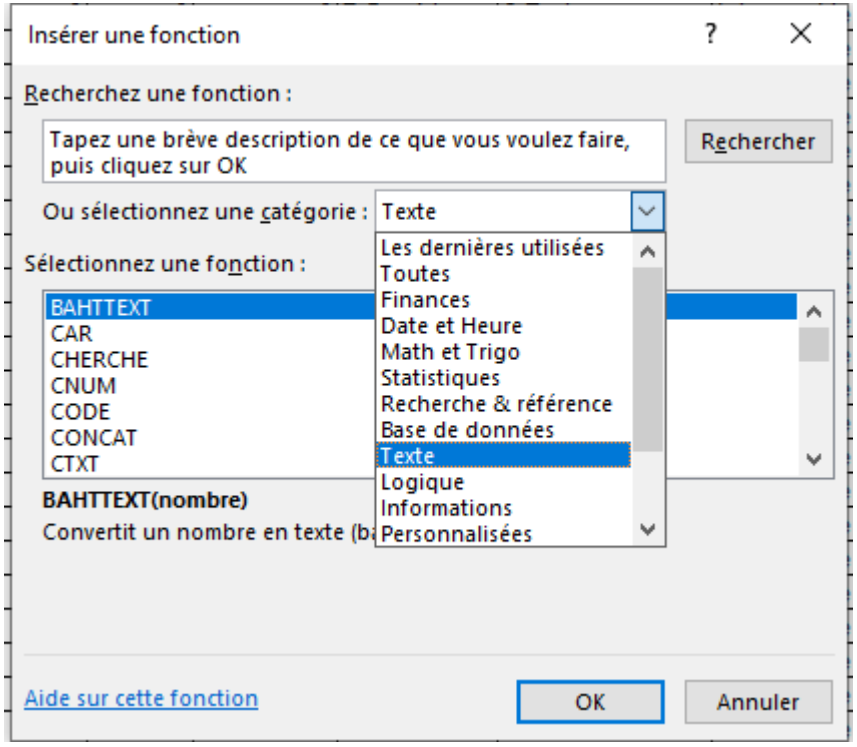
- Dupliquer la variable originale
- Changer son nom
- Sélectionner toute la variable (colonne)
- Rechercher /Remplacer
  - Cocher totalité du contenu de la cellule



T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
CSEX	CSEX_b	CAGE	CCS82	CDIP	NBPD6	CHAT	CHIEN	a_chat	a_chien	a_oiseaux	a_rongeurs
1	1	37	37	5	2	0	1	2. Non	1. Oui	2. Non	2. Non
2	2	40	67	1	1	0	0	2. Non	2. Non	2. Non	2. Non
2	2	68	67	2	1	0	0	2. Non	2. Non	2. Non	2. Non

Solution 3. Formules

- Nombreuses formules sous Excel
- Calcul mathématique : + - / * ^ ( )
  - Ex : IMC=Poids/(Taille en m)²
    - = AR2/(AS2/100)^2
  - $ → colonne ou ligne absolue (vs relative)
- Fonction « SI » /IF
- Opération sur le texte
- Opérateur de fusion « RECHERCHEV » (VLOOKUP)



Formules mathématiques

- Utiles mêmes pour des valeurs catégorielles
  - Ex. CS à 2 chiffres → CS à 1 chiffre
    - =ENT(N2/10)
- Formules mathématiques toutes faites
  - MOYENNE, ECARTYPE
- Formules mathématiques conditionnelles
  - MOYENNE.SI ; SOMME.SI ; NB.SI
- Valeur retardé / Evolution. Formule en Z3
  - Ex Données pays (colonne B) / années
    - =SI(B3=B2;Y2;"")
      → Valeur passée de Y
    - =SI(B3=B2;Y3-Y2;"")
      → Evolution de Y

Fonction SI

- Utile pour des recodages simples (en deux ou trois modalités)

  =SI(D2=1;"1.Homme";"2.Femme")

- On peut empiler les SI

  =SI(D2="";"";SI(D2=1;"1.Homme";"2.Femme"))

- Pratique pour créer des variables dichotomiques (1 ou 0)

  =SI(E2="";"";SI(E2="1.Homme";1;0))

- Attention : Trop de SI → illisible

Fonctions Texte

- Trois premiers caractères de gauche

  =GAUCHE(D2;3)
- Trois premiers caractères de droite

  =DROITE(D2;3)
- Deux caractères à partir du troisième

  =STXT(D2;3;2)

- Rechercher une chaîne de caractères

  =CHERCHE("travail";D2)
  - → Position de la chaîne de caractère « travail » si présente, #Valeur! sinon
- Pour créer une variable dichotomique sur un mot

  =1-ESTERREUR(CHERCHE("travail";D2))

Fonction RECHERCHEV

- Fonction importante pour la fusion de fichiers (cf. supra)
- Peut servir pour recoder une variable avec beaucoup de modalités
- =RECHERCHEV(N2;[pcs2003.xlsx]Pcs!$A$2:$B$44;2;0)

Fichier source

=RECHERCHEV(N2;[pcs2003.xlsx]Pcs!$A$2:$B$44;2;0)

- Champ 1. Colonne où vous avez votre code : N2
- Champ 2. Plage de recherche [pcs2003.xlsx]Pcs!$A$2:$B$44
  - Code recherché dans la première colonne
  - Attention d'avoir les $
- Champ 3 : 2 → Information souhaité dans la deuxième colonne de cette plage
- Champ 4 : 0 → Interdire (0) ou autoriser (1) l'approximation

	A	B
1	PCS 2003	Niveau 3 - Liste des catégories socioprofessionnelles détaillées
2	Code	Libellé
3	11	Agriculteurs sur petite exploitation
4	12	Agriculteurs sur moyenne exploitation
5	13	Agriculteurs sur grande exploitation
6	21	Artisans
7	22	Commerçants et assimilés
8	23	Chefs d'entreprise de 10 salariés ou plus
9	31	Professions libérales
10	33	Cadres de la fonction publique
11	34	Professeurs, professions scientifiques
12	35	Professions de l'information, des arts et des spectacles
13	37	Cadres administratifs et commerciaux d'entreprise
14	38	Ingénieurs et cadres techniques d'entreprise
15	42	Professeurs des écoles, instituteurs et assimilés
16	43	Professions intermédiaires de la santé et du travail social
17	44	Clergé, religieux
18	45	Professions intermédiaires administratives de la fonction publique
19	46	Professions intermédiaires administratives et commerciales des entreprises
20	47	Techniciens
21	48	Contremaîtres, agents de maîtrise
22	52	Employés civils et agents de service de la fonction publique
23	53	Policiers et militaires
24	54	Employés administratifs d'entreprise
25	55	Employés de commerce
26	56	Personnels des services directs aux particuliers
27	62	Ouvriers qualifiés de type industriel
28	63	Ouvriers qualifiés de type artisanal
29	64	Chauffeurs
30	65	Ouvriers qualifiés de la manutention, du magasinage et du transport
31	67	Ouvriers non qualifiés de type industriel
32	68	Ouvriers non qualifiés de type artisanal
33	69	Ouvriers agricoles
34	71	Anciens agriculteurs exploitants
35	72	Anciens artisans, commerçants, chefs d'entreprise
36	74	Anciens cadres
37	75	Anciennes professions intermédiaires
38	77	Anciens employés
39	78	Anciens ouvriers
40	81	Chômeurs n'ayant jamais travaillé
41	83	Militaires du contingent
42	84	Elèves, étudiants
43	85	Personnes diverses sans activité professionnelle de moins de 60 ans (sauf retraités)
44	86	Personnes diverses sans activité professionnelle de 60 ans et plus (sauf retraités)

Combiner des jeux de données

- Fusionner
  - Opération « en largeur »
  - =Rajouter des variables
- Empiler
  - Opération « en longueur »
  - Mettre les plages l'une au dessus de l'autre

Fusionner les données

- RECHERCHEV est le principal opérateur de fusion
- Excel peut fusionner
  - FR : Fichier Réception (où on utilise RECHERCHEV). FV : Fichier Versé dans FR
  - FV: identifiant unique à FR: identifiant unique
  - FV: identifiant unique à FR: identifiant multiple
- Excel ne peut pas fusionner
  - FV: identifiant multiple à FR: identifiant unique
  - FV: identifiant multiple à FR: identifiant multiple

Quelques subtilités pour copier coller

=RECHERCHEV($ED2;menage.xlsx!
$A$1:$JK$4116;COLONNE(B2);0)

- Si valeurs manquantes

=SI(RECHERCHEV($ED2;menage.xlsx!
$A$1:$JK$4116;COLONNE(B2);0)="";"";
RECHERCHEV($ED2;menage.xlsx!
$A$1:$JK$4116;COLONNE(B2);0))

Empiler les données

- Comparer deux enquêtes à deux périodes dans le temps
- Lire la documentation
- Vérifier que les questions et les noms de variables sont les mêmes
- Problème : d'une enquête à l'autre nouvelles variables, anciennes variables
- Vérifier que les noms correspondent et déplacer les colonnes jusqu'à ce soit le cas

Exemple

- =1-(A2=lycee.xlsx!A1)

BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
0	0	0	0	0	0	=1-(BH2=lycee.xlsx!BH1)				0	0	0	0	0	1	1	0	0	1	1
5	C6	D1_1	D1_2	D2_1	D2_2	D3_1	D3_2	D4_1	D4_2	D5	D6	D7	E1	E2	E2_L	E3	E4	E5	E6_L	F1
1	2	140	150	172	184	140	150	172	110	2	1	1	2	0	0	0	22	1	VETERINAIRE	1
2	1	130	140	888	0	161	110	120	0	5	1	2	1	1	1	30	25	1	VETERINAIRE	2
2	2	140	110	185	0	140	120	183	0	5	1	2	1	1	1	30	30	1	CHIRURGIEN	1
2	1	140	183	162	120	183	110	162	120	3	1	2	3	0	0	0	22	2	NC	1
1	1	110	184	150	140	110	181	150	140	5	1	2	3	0	0	0	24	1	ARCHITECTE	1
2	1	150	140	120	162	140	150	120	162	3	1	2	3	0	0	0	22	2	NC	2
2	1	110	183	130	120	162	104	130	184	4	1	2	1	1	1	30	21	2	NC	2

Conserver l'intégrité des données

- Conserver une version originale du fichier non modifié.