

BONNES PRATIQUES POUR LE NETTOYAGE DES DONNEES

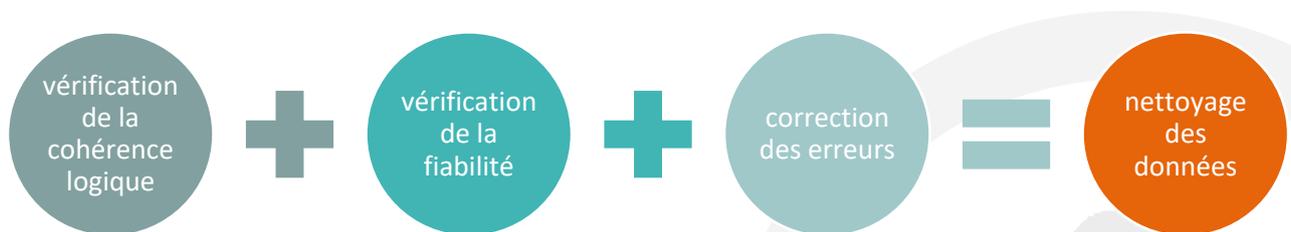
Table des matières

I. Étapes de base pour le nettoyage de vos données	1
II. Comment nettoyer vos données dans Excel	2
II.1. Débarrassez-vous des espaces supplémentaires	2
II.2. Sélectionner et traiter toutes les cellules vides (données manquantes)	3
II.3. Convertir les nombres stockés sous forme de texte en nombres	4
II.4. Supprimer les doublons	5
II.5. Surligner et corriger les erreurs	7
II.6. Changer le texte en minuscules, majuscules ou nom propre	8
II.7. Utilisez Convertir pour séparer les données dans Excel	9
II.8. Supprimer tout le formatage	11
II.9. Utilisez "trouver et remplacer" pour nettoyer les données dans Excel	12

Le processus de nettoyage des données peut être effectué à l'aide de différentes techniques et logiciels (Excel, STATA, SPSS, etc.). Ce tutoriel se concentre sur les principes généraux du nettoyage des données, en zoomant sur la façon d'implémenter le nettoyage des données dans Excel, car c'est l'outil le plus commun et polyvalent utilisé par Tdh.

I. Étapes de base pour le nettoyage de vos données¹

Le nettoyage de vos bases de données vous assure d'avoir toujours une base de données à jour qui peut être analysée sans risque de produire de la désinformation pour ceux qui prendront ensuite les décisions. Le nettoyage des bases de données est avant tout logique, l'analyse de la cohérence des données et la triangulation avec d'autres informations qui peuvent être disponibles.



 Rappelez-vous que la meilleure façon de réduire la charge de travail de nettoyage des données est de planifier soigneusement votre collecte de données dès le début², y compris

¹ Source : <https://support.office.com/fr-fr/article/les-dix-meilleures-solutions-pour-nettoyer-vos-donn%c3%a9es-2844b620-677c-47a7-ac3e-c2e157d1db19?ui=fr-FR&rs=fr-FR&ad=FR>

² <https://www.mdc-toolkit.org/fr/design-your-forms/>

l'analyse que vous voudrez faire lors de la planification, et de la tester avec de fausses données avant le déploiement !

Lors du nettoyage de nos bases de données, nous recherchons les incohérences et les erreurs suivantes :

- Pourcentages (ou total des pourcentages) > ou < de 100%.
- Sommes qui ne correspondent pas (c.-à-d. nombre total de membres du ménage différent de la somme des hommes et des femmes dans le ménage...)
- Caractères spéciaux qui ont été transformés lors de l'exportation
- Mauvaise interprétation des questions par les recenseurs ou les personnes interrogées
- "0" au lieu d'une cellule vide ou "N.A." (c.-à-d. pour une question numérique qui n'a pas reçu de réponse)
- Cellules vides qui doivent être remplies
- Erreurs de frappe ("-5" personnes dans le ménage)
- Problèmes d'unités (âge en mois/années, mètres/pieds, etc.), date ou format de cellule
- Etc etc. etc.

Il y a 4 étapes de base recommandées pour nettoyer vos données :

 Avant, pendant et après chacune de ces étapes, vous devez vérifier (visuellement ou à l'aide de filtres, etc.) que vous n'avez effectué aucun changement auquel vous ne vous attendiez pas !

1. Créez une copie des données originales dans un classeur séparé.

 **Que se passe-t-il si vous ne créez pas une copie des données d'origine et que vous supprimez par inadvertance une colonne contenant des données (ou un autre élément) dans votre base de données ?**

Eh bien, si elle n'est plus sur Kobo, ces données sont perdues à jamais et il n'y a aucun moyen de les récupérer. C'est pourquoi il est très important de créer une copie des données originales dans un classeur séparé !

2. Assurez-vous que les données sont présentées sous forme de tableaux contenant des lignes et des colonnes avec : des données similaires dans chaque colonne, toutes les colonnes et lignes visibles, aucune cellule fusionnée, aucune réponse multiple dans une cellule et aucune ligne vide dans l'intervalle.

3. Exécutez des tâches qui ne nécessitent pas de manipulation préalable des colonnes, telles que l'élimination des espaces supplémentaires ou l'utilisation de la boîte de dialogue "Rechercher et remplacer".

4. Ensuite, exécutez les tâches qui nécessitent une manipulation de colonne.

II. Comment nettoyer vos données dans Excel³ ⁴

II.1. Débarrassez-vous des espaces supplémentaires

Pour vous débarrasser des espaces supplémentaires, vous pouvez utiliser la **fonction SUPPRESPEACE dans Excel**. La fonction TRIM supprime tous les espaces du texte, sauf les espaces simples entre les mots. La syntaxe de la fonction est **SUPPRESPEACE(texte)**, où (texte) est un argument obligatoire de la fonction et fait référence au texte dont vous voulez supprimer les espaces.

³ Sources : <https://trumpexcel.com/clean-data-in-excel/> et <https://www.youtube.com/watch?v=e0TfIbZZXPeA>

⁴ La version Excel utilisée pour ce tutoriel est Excel 2013. De plus, si vous avez des questions ou des problèmes, n'hésitez pas à chercher votre question sur le Web, car il existe une multitude de sources susceptibles de vous aider.

Tout d'abord, tapez votre formule à côté de votre première entrée de texte. Faites ensuite glisser la formule vers le bas pour couvrir toutes vos entrées de texte comme ci-dessous. Toutes vos entrées de texte seront alors nettoyées des espaces supplémentaires.

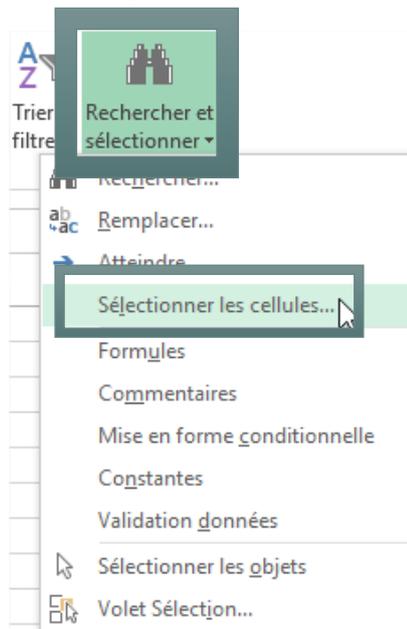
	A	B
1	How to clean your data	=SUPPRESPE(A1)
2	How to clean your data	
3	How to clean your data	
4	How to clean your data	
5	How to clean your data	
6	How to clean your data	
7		

	A	B
1	How to clean your data	How to clean your data
2	How to clean your data	How to clean your data
3	How to clean your data	How to clean your data
4	How to clean your data	How to clean your data
5	How to clean your data	How to clean your data
6	How to clean your data	How to clean your data
7		

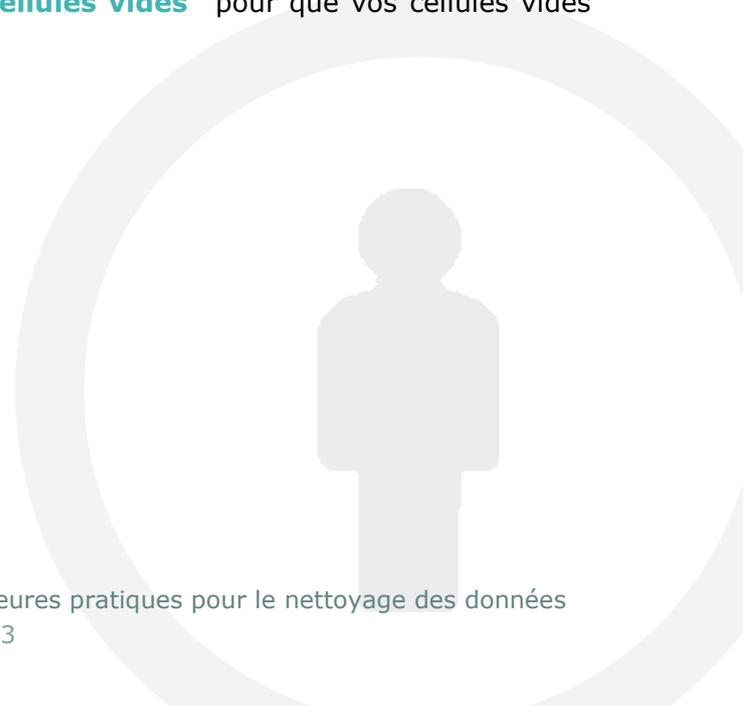
II.2. Sélectionner et traiter toutes les cellules vides (données manquantes)

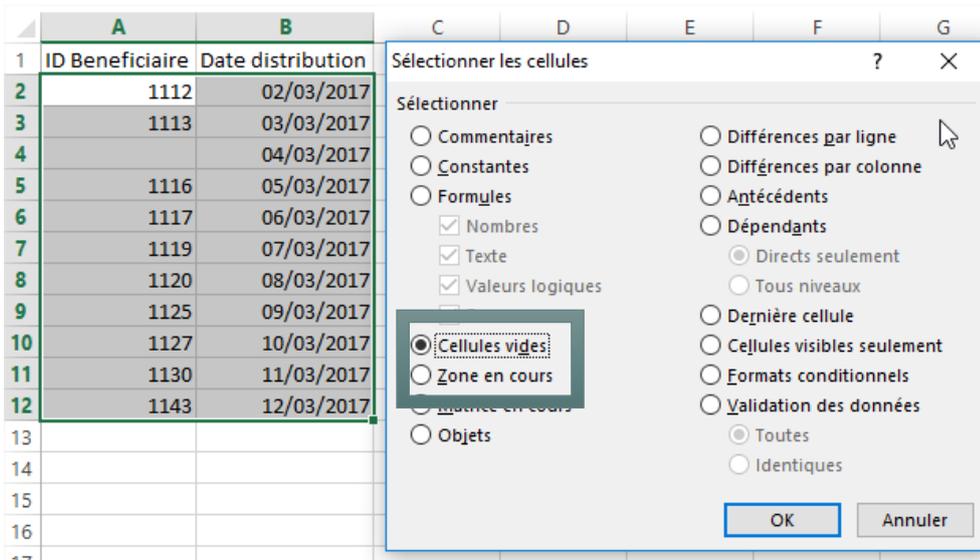
Vous voudrez peut-être détecter toutes les cellules vides, car elles pourraient représenter des données manquantes, et les remplacer par "Données manquantes", par exemple. Pour ce faire, sélectionnez l'ensemble des données, cliquez sur "**Rechercher et sélectionner**", puis sur "**Sélectionner les cellules...**" où une boîte de dialogue Sélectionner les cellules s'ouvre.

	A	B
1	ID Beneficiaire	Date distribution
2	1112	02/03/2017
3	1113	03/03/2017
4		04/03/2017
5	1116	05/03/2017
6	1117	06/03/2017
7	1119	07/03/2017
8	1120	08/03/2017
9	1125	09/03/2017
10	1127	10/03/2017
11	1130	11/03/2017
12	1143	12/03/2017



Dans la boîte de dialogue, sélectionnez l'option "**Cellules vides**" pour que vos cellules vides apparaissent en gris.





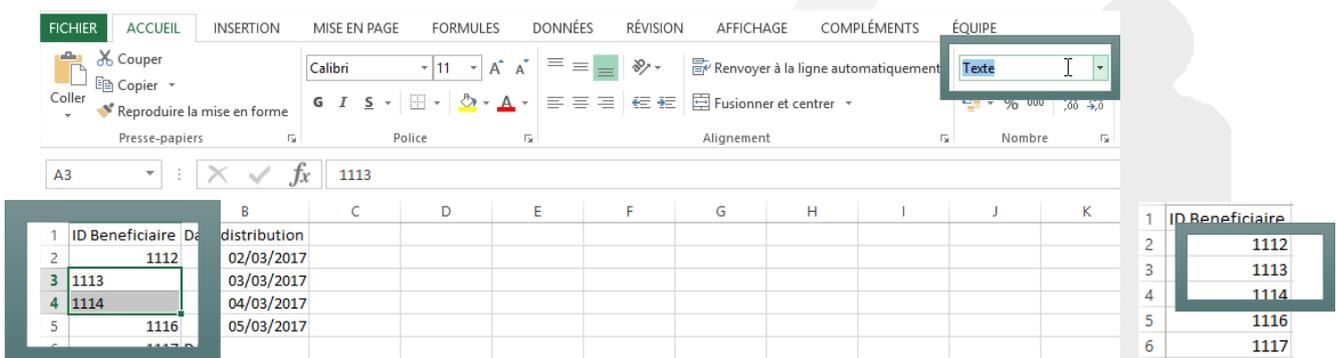
Toutes les cellules vides de votre ensemble de données seront sélectionnées en même temps. La première cellule sélectionnée sera vide (et non pas gris clair comme les autres) car cette cellule est la **cellule active**. Pour saisir D/M dans toutes les cellules vides (= données manquantes), tapez ce texte dans la cellule active et appuyez sur **Ctrl+Enter**.

	A	B		A	B
1	ID Beneficiaire	Date distribution		ID Beneficiaire	Date distribution
2	1112	02/03/2017		1112	02/03/2017
3	1113	03/03/2017		1113	03/03/2017
4		04/03/2017		D/M	04/03/2017
5	1116	05/03/2017		1116	05/03/2017
6	1117			1117	D/M
7	1119	07/03/2017		1119	07/03/2017
8	1120	08/03/2017		1120	08/03/2017
9	1125	09/03/2017		1125	09/03/2017
10	1127			1127	D/M
11	1130	11/03/2017		1130	11/03/2017
12	1143	12/03/2017		1143	12/03/2017

II.3. Convertir les numéros stockés sous forme de texte en numéros

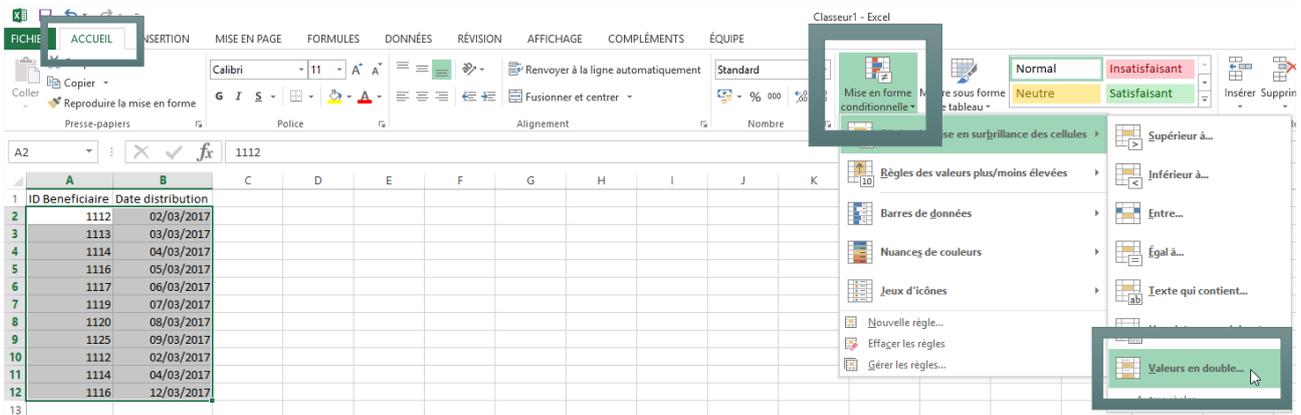
Pour convertir un nombre stocké sous forme de texte en nombre, allez dans la **"Format de nombre"** et tapez **"Standard"**. Ceci transformera tous les nombres stockés sous forme de texte en nombres (c.-à-d. qu'ils sont alignés à droite de la cellule).

💡 Les nombres sont toujours alignés à droite de la cellule, alors que le texte est aligné à gauche de la cellule.

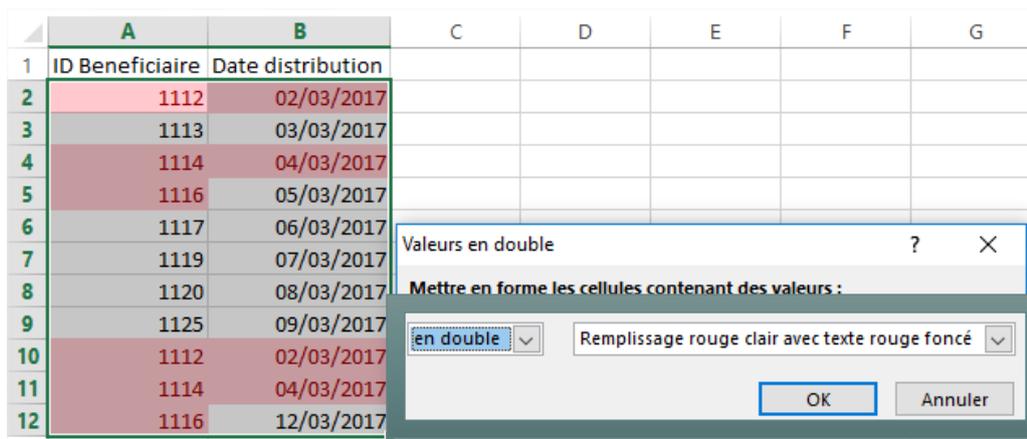


II.4. Supprimer les doublons

Vous devez d'abord trouver où se situent les doublons. Sélectionnez l'ensemble des données, allez dans "Accueil" sur "Mise en forme conditionnelle", puis sur "Règles de mise en surbrillance des cellules" et cliquez sur "Valeurs en doubles...".



Excel vous donne alors la possibilité de surligner les doublons en rouge clair, où vous devrez cliquer sur "OK" pour approuver.



Il mettra alors en surbrillance toutes les valeurs dupliées.

💡 Dans l'exemple ci-dessous, vous pouvez voir que l'ID du bénéficiaire 1116 a été mis en évidence uniquement dans la colonne "ID du bénéficiaire" et non dans la colonne "Date de distribution". Cela s'explique par le fait qu'il ne s'agit pas de vrais doublons, car les dates de distribution des deux entrées sont différentes : il s'agit donc de deux entrées de données différentes.

	A	B
1	ID Beneficiaire	Date distribution
2	1112	02/03/2017
3	1113	03/03/2017
4	1114	04/03/2017
5	1116	05/03/2017
6	1117	06/03/2017
7	1119	07/03/2017
8	1120	08/03/2017
9	1125	09/03/2017
10	1112	02/03/2017
11	1114	04/03/2017
12	1116	12/03/2017

Ensuite, allez dans **"Données"** et cliquez sur **"Supprimer les doublons"** pour supprimer les vrais doublons.

The screenshot shows the Excel ribbon with the 'DONNÉES' tab selected. The 'Supprimer les doublons' button is highlighted with a green box. A tooltip for this button is visible, stating: 'Supprimer les doublons. Supprimer les lignes en double dans une feuille de données. Vous pouvez sélectionner les colonnes dans lesquelles vous souhaitez vérifier la présence d'informations en double.'

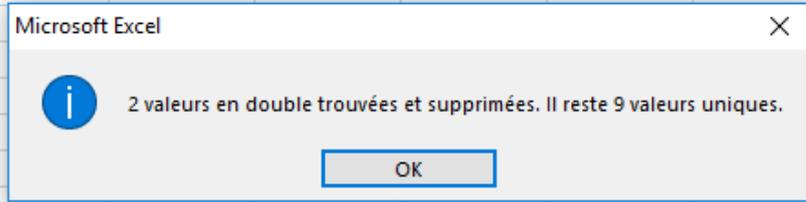
Excel vous demandera alors de sélectionner les colonnes qui contiennent les doublons. **Sélectionner tout** et cliquez sur **"OK"**.

The screenshot shows the 'Supprimer les doublons' dialog box. The 'Sélectionner tout' button is highlighted with a green box. Below it, the 'Colonnes' section has two checkboxes selected: 'ID Beneficiaire' and 'Date distribution'. The 'OK' button is also highlighted with a green box.

Il vous informe ensuite du nombre de doublons supprimés et du nombre d'entrées de données (valeurs uniques) qui ont été conservées.

Comme vous pouvez le voir, la "fausse" duplication 1116 a été conservée.

	A	B	C	D	E	F	G	H
1	ID Beneficiaire	Date distribution						
2	1112	02/03/2017						
3	1113	03/03/2017						
4	1114	04/03/2017						
5	1116	05/03/2017						
6	1117	06/03/2017						
7	1119	07/03/2017						
8	1120	08/03/2017						
9	1125	09/03/2017						
10	1116	12/03/2017						
11								
12								

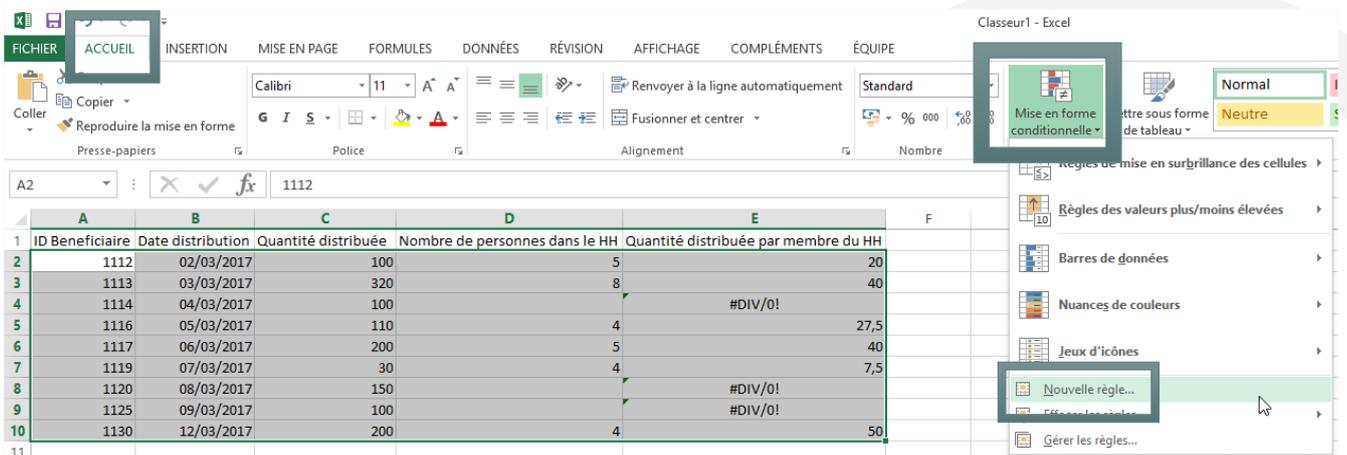


II.5. Surligner et corriger les erreurs

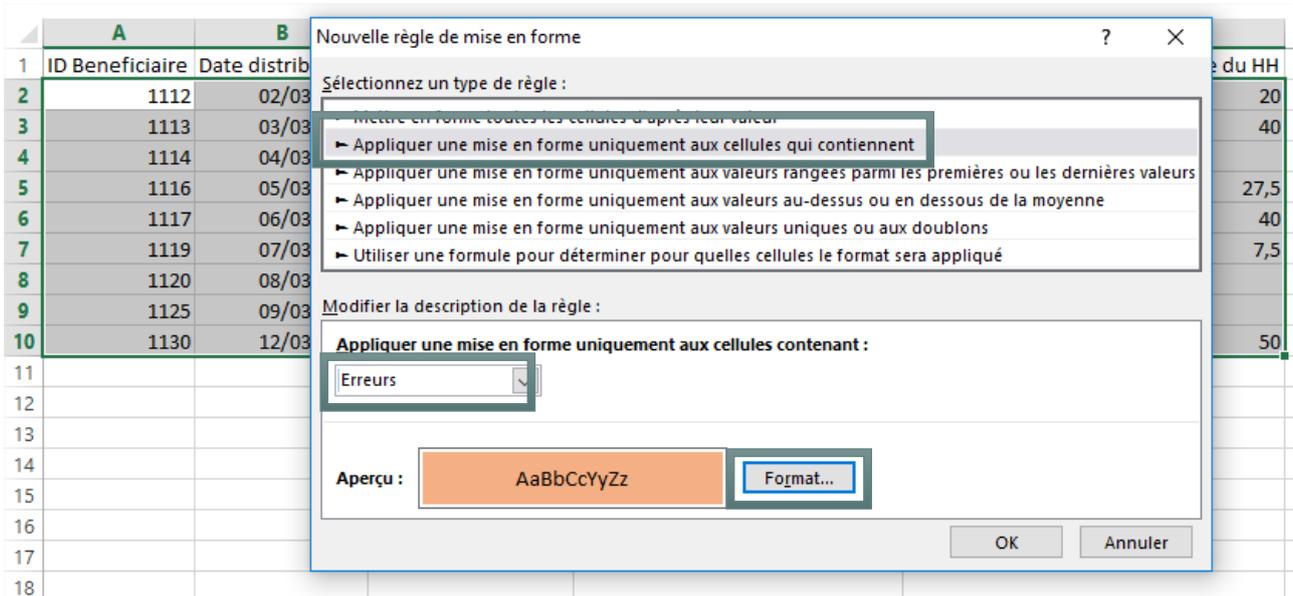
Dans le tableau ci-dessous, un indicateur appelé "Quantité distribuée par membre du ménage" a été calculé à partir des indicateurs "Quantité distribuée" et "Nombre de personnes en ménage". Cependant, certaines valeurs comportent des erreurs (car il y a des valeurs manquantes dans l'indicateur "nombre de personnes dans ménage").

	A	B	C	D	E
1	ID Beneficiaire	Date distribution	Quantité distribuée	Nombre de personnes dans le HH	Quantité distribuée par membre du HH
2	1112	02/03/2017	100	5	20
3	1113	03/03/2017	320	8	40
4	1114	04/03/2017	100		#DIV/0!
5	1116	05/03/2017	110	4	7,5
6	1117	06/03/2017	200	5	40
7	1119	07/03/2017	30	4	7,5
8	1120	08/03/2017	150		#DIV/0!
9	1125	09/03/2017	100		#DIV/0!
10	1130	12/03/2017	200	4	50

Pour détecter ces erreurs, allez dans "Accueil", "Mise en forme conditionnelle" et cliquez sur "Nouvelle règle".



Dans la boîte de dialogue "**Nouvelle règle de mise en forme**" qui s'ouvre, cliquez sur "**Appliquer une mise en forme uniquement aux cellules qui contiennent**", puis sélectionnez "**Erreurs**" et choisissez la couleur du "**Format**".



Excel mettra alors en surbrillance toutes les cellules contenant une erreur. Vous pouvez donc maintenant filtrer les colonnes associées par couleur pour travailler directement sur les cellules présentant des erreurs.

	A	B	C	D	E
1	ID Beneficiaire	Date distribution	Quantité distribuée	Nombre de personnes dans le HH	Quantité distribuée par membre du HH
2	1112	02/03/2017	100	5	20
3	1113	03/03/2017	320	8	40
4	1114	04/03/2017	100		#DIV/0!
5	1116	05/03/2017	110	4	27,5
6	1117	06/03/2017	200	5	40
7	1119	07/03/2017	30	4	7,5
8	1120	08/03/2017	150		#DIV/0!
9	1125	09/03/2017	100		#DIV/0!
10	1130	12/03/2017	200	4	50
11					

II.6. Changer le texte en minuscules, majuscules ou nom propre.

Dans l'exemple ci-dessous, les noms sont écrits de différentes façons. Vous pouvez les harmoniser en utilisant les fonctions MINUSCULE/MAJUSCULE/NOMPROPRE.

	A	B	C	D
1	Nom	Minuscule	Majuscule	Nom propre
2	JANE FONDA			
3	Joe strike			
4	Martha Hoey			
5	Phil COLSON			

Si vous voulez qu'elles soient entièrement écrites en minuscules, vous pouvez utiliser la fonction **MINUSCULE(text)**, où (text) fait référence au texte à modifier. Une fois que vous avez saisi

voire fonction pour une saisie de données, faites glisser la formule vers le bas pour couvrir toutes les saisies de données que vous voulez modifier.

	A	B	C	D
1	Nom	Minuscule	Majuscule	Nom propre
2	JANE FONDA	=MINUSCULE(A2)		
3	Joe strike			
4	Martha Hoey			
5	Phil COLSON			
6				

	A	B	C	D
1	Nom	Minuscule	Majuscule	Nom propre
2	JANE FONDA	jane fonda		
3	Joe strike	joe strike		
4	Martha Hoey	martha hoey		
5	Phil COLSON	phil colson		
6				

Si vous voulez qu'elles soient entièrement écrites en majuscules, vous pouvez utiliser la fonction **MAJUSCULE(texte)**, où (texte) fait référence au texte à modifier. Une fois que vous avez saisi votre fonction pour une saisie de données, faites glisser la formule vers le bas pour couvrir toutes les saisies de données que vous voulez modifier.

	A	B	C	D
1	Nom	Minuscule	Majuscule	Nom propre
2	JANE FONDA	jane fonda	=MAJUSCULE(A2)	
3	Joe strike	joe strike		
4	Martha Hoey	martha hoey		
5	Phil COLSON	phil colson		
6				

	A	B	C	D
1	Nom	Minuscule	Majuscule	Nom propre
2	JANE FONDA	jane fonda	JANE FONDA	
3	Joe strike	joe strike	JOE STRIKE	
4	Martha Hoey	martha hoey	MARTHA HOEY	
5	Phil COLSON	phil colson	PHIL COLSON	
6				

Si vous voulez que les premières lettres du prénom et du nom de famille soient écrites en majuscules et les autres en minuscules, vous pouvez utiliser la fonction **NOMPROPRE(text)**, où (text) renvoie au texte à modifier. Une fois que vous avez saisi votre fonction pour une saisie de données, faites glisser la formule vers le bas pour couvrir toutes les saisies de données que vous voulez modifier.

	A	B	C	D	E
1	Nom	Minuscule	Majuscule	Nom propre	
2	JANE FONDA	jane fonda	JANE FONDA	=NOMPROPRE(A2)	
3	Joe strike	joe strike	JOE STRIKE		
4	Martha Hoey	martha hoey	MARTHA HOEY		
5	Phil COLSON	phil colson	PHIL COLSON		
6					

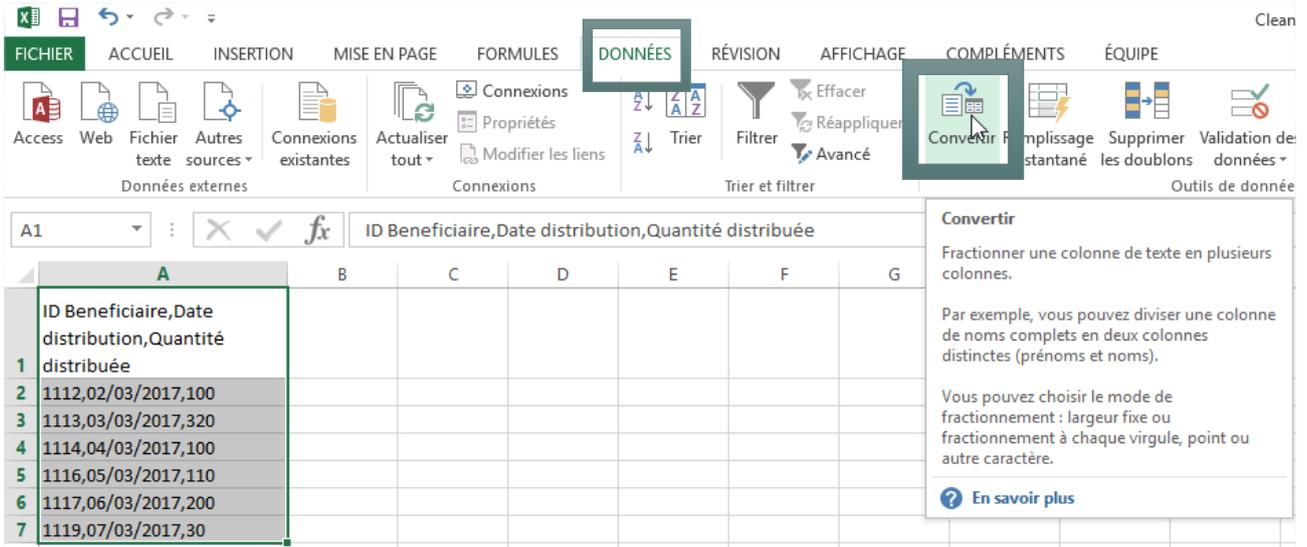
	A	B	C	D
1	Nom	Minuscule	Majuscule	Nom propre
2	JANE FONDA	jane fonda	JANE FONDA	Jane Fonda
3	Joe strike	joe strike	JOE STRIKE	Joe Strike
4	Martha Hoey	martha hoey	MARTHA HOEY	Martha Hoey
5	Phil COLSON	phil colson	PHIL COLSON	Phil Colson
6				

II.7. Utilisez Convertir pour séparer les données dans Excel

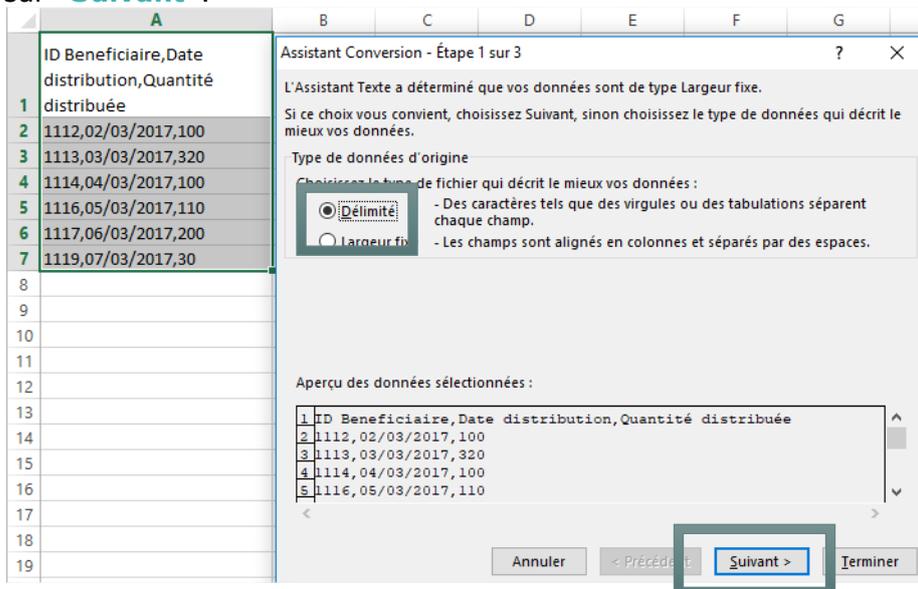
Il peut arriver que vous obteniez une version .csv de la base de données, où toutes les données sont concentrées dans une seule colonne (séparée par une virgule, une tabulation, etc.) comme dans l'exemple ci-dessous.

	A
	ID Beneficiaire,Date distribution,Quantité distribuée
1	
2	1112,02/03/2017,100
3	1113,03/03/2017,320
4	1114,04/03/2017,100
5	1116,05/03/2017,110
6	1117,06/03/2017,200
7	1119,07/03/2017,30
8	

Pour ce faire, allez dans **"Données"** et cliquez sur **"Convertir"**.

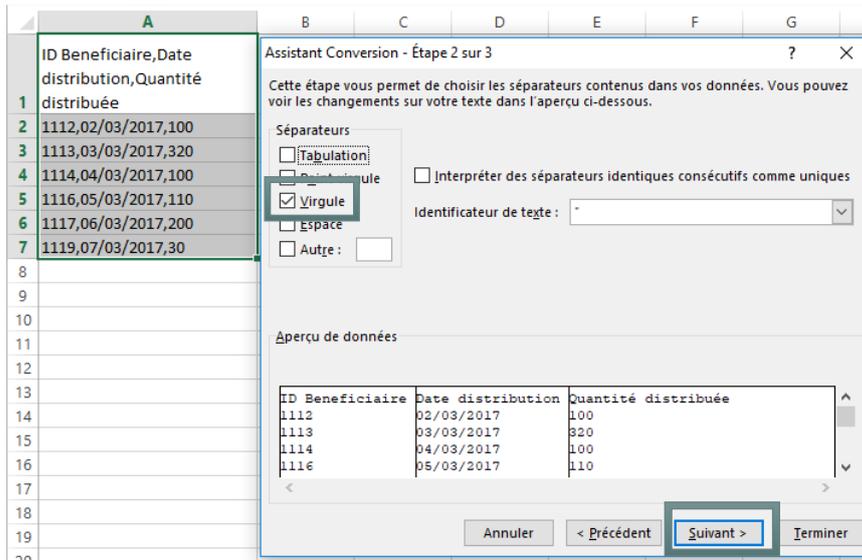


Dans la boîte de dialogue **"Assistant Conversion"** qui s'ouvre, sélectionnez **"Délimité"**, puis cliquez sur **"Suivant"**.

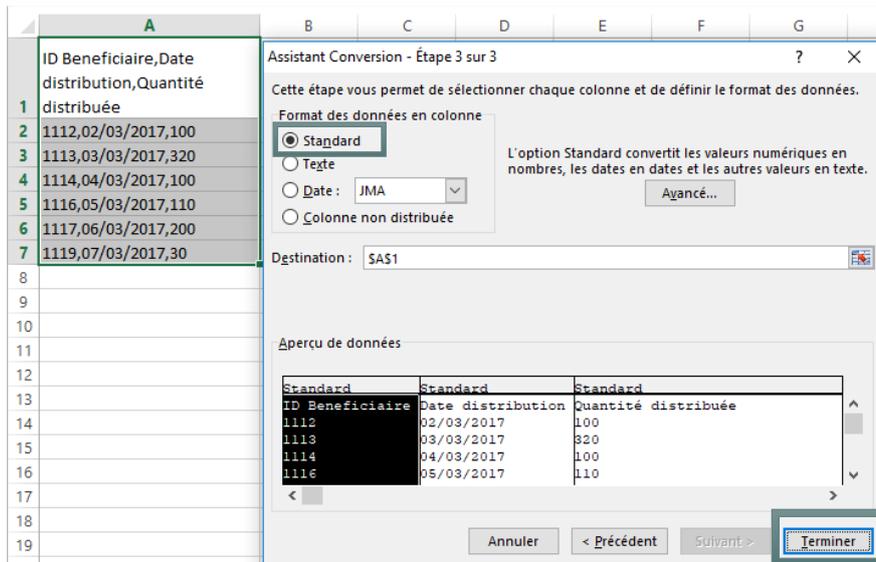


Excel vous demandera ensuite de quelle manière il doit séparer vos données. Comme une virgule sépare vos différentes entrées de données dans notre exemple, sélectionnez **"Virgule"** et cliquez ensuite sur **"Suivant"**. D'autres options peuvent être **"Tabulation"**, **"Point-virgule"**, **"Espace"** et **"Autre"**.





Il vous demandera ensuite dans quel format les données doivent être séparées. Sélectionnez **"Standard"**, puis cliquez sur **"Terminer"**.

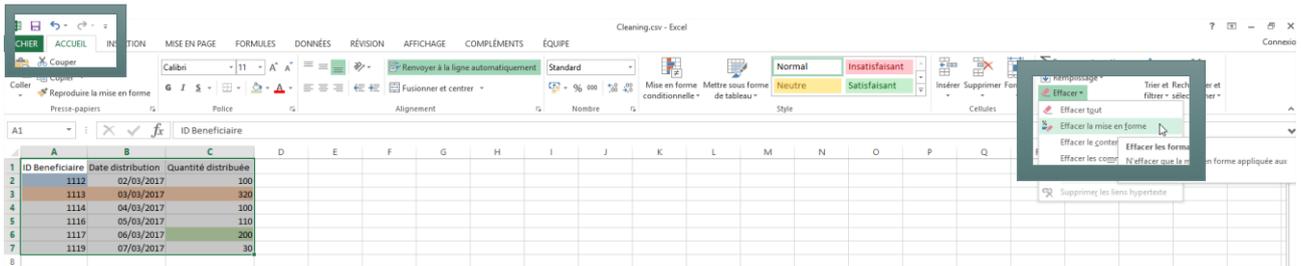


Vos entrées de données sont maintenant séparées dans différentes colonnes d'Excel.

	A	B	C
1	ID Beneficiaire	Date distribution	Quantité distribuée
2	1112	02/03/2017	100
3	1113	03/03/2017	320
4	1114	04/03/2017	100
5	1116	05/03/2017	110
6	1117	06/03/2017	200
7	1119	07/03/2017	30

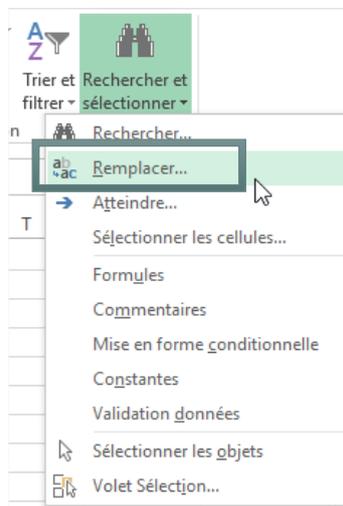
II.8. Supprimer tout le formatage

Si vous avez encore beaucoup de formatage conditionnel dans vos données et que vous souhaitez vous en débarrasser car il n'est plus utile, vous pouvez d'abord sélectionner toutes vos données, allez dans **"Accueil"**, puis **"Effacer"** et cliquez sur **"Effacer la mise en forme"**.

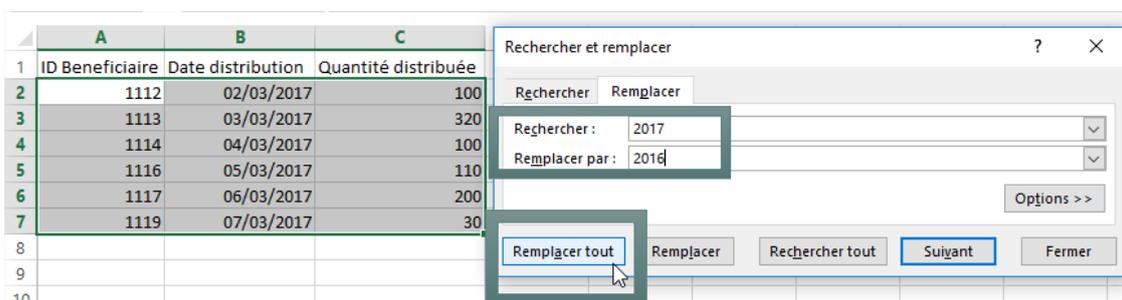


II.9. Utilisez "trouver et remplacer" pour nettoyer les données dans Excel

Si vous avez les mêmes entrées de données dans votre ensemble de données mais écrites de différentes manières (c'est-à-dire avec différentes orthographes, etc.), l'option **"Rechercher et remplacer"** dans Excel peut devenir très utile. Pour l'utiliser, trouvez d'abord toutes les données à remplacer en allant dans **"Accueil"**, dans **"Rechercher et sélectionner"** et cliquez ensuite sur **"Rechercher tout"**. Vérifiez que vous ne remplacerez pas des données provenant d'autres colonnes auxquelles vous ne vous attendriez pas. Ensuite, sélectionnez toutes vos données à nettoyer, allez dans **"Accueil"**, dans **"Rechercher et sélectionner"** et cliquez sur **"Remplacer"**.



Dans l'assistant **"Rechercher et remplacer"** qui s'ouvre, entrez les valeurs que vous voulez remplacer dans **"Rechercher"**, entrez les valeurs par lesquelles vous voulez les remplacer dans **"Remplacer par"**, puis cliquez sur **"Remplacer tout"**.



Excel vous indiquera alors combien de valeurs ont été remplacées/modifiées.

	A	B	C
1	ID Beneficiaire	Date distribution	Quantité distribuée
2	1112	02/03/2016	100
3	1113	03/03/2016	320
4	1114	04/03/2016	100
5	1116	05/03/2016	110
6	1117	06/03/2016	200
7	1119	07/03/2016	30

Rechercher et remplacer

Rechercher Remplacer

Rechercher : 2017

Remplacer par : 2016

Options >>

Remplacer tout Remplacer Rechercher tout Suivant Fermer

Microsoft Excel

Cette action est terminée, 6 remplacements ont été effectués.

OK

 Notez bien sûr que ce tutoriel n'est pas complet car le nettoyage des données peut prendre tellement de formes !

