

Atelier de formation à la Gestion-Analyse des données et interprétation des résultats des ESS



Gestion de la base de données



Numsai	exrepond	Chefmena	Regio	Cercle	Village	zoneagroec
11	1	1	2	9	bara-village	1
18	1	1	2	9	outagouna-village	1
19	2	1	2	9	outagouna-village	1
20	1	1	2	9	outagouna-village	1
21	1	1	2	9	bara-village	1
22	2	1	2	9	bara-village	1
23	1	1	2	9	bara-village	1
24	2	1	2	9	bara-village	1
25	1	1	2	9	bara-village	1
28	1	1	2	9	bara-village	1
31	2	1	2	9	bara-village	1
32	1	1	2	9	bara-village	1
33	1	1	2	9	bara-village	1
35	2	1	2	9	bara-village	1

Objectifs

1. Etre capable de **minimiser les erreurs** durant le traitement des données
2. Etre capable **d'éliminer les erreurs** et/ou les données aberrantes lors du traitement de données.
3. **Stocker/enregistrer les données** en toute sécurité

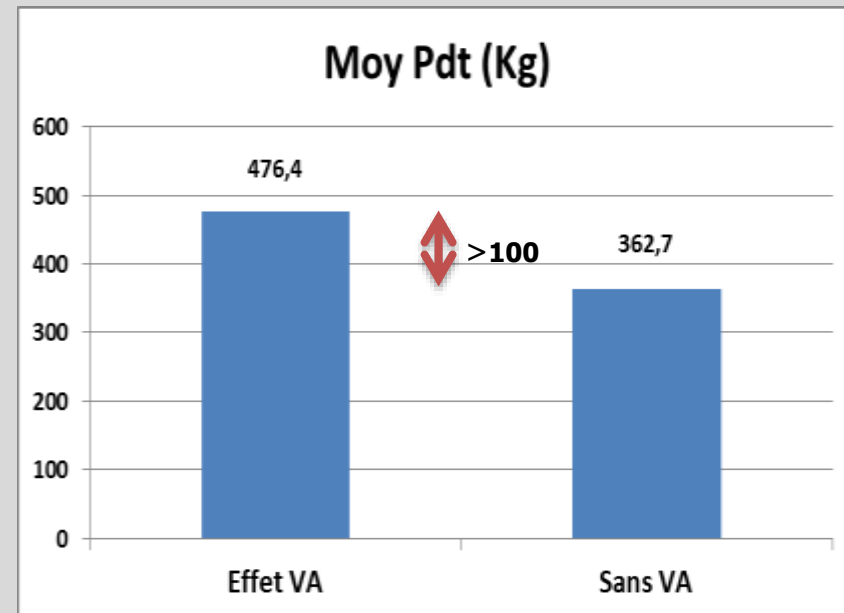
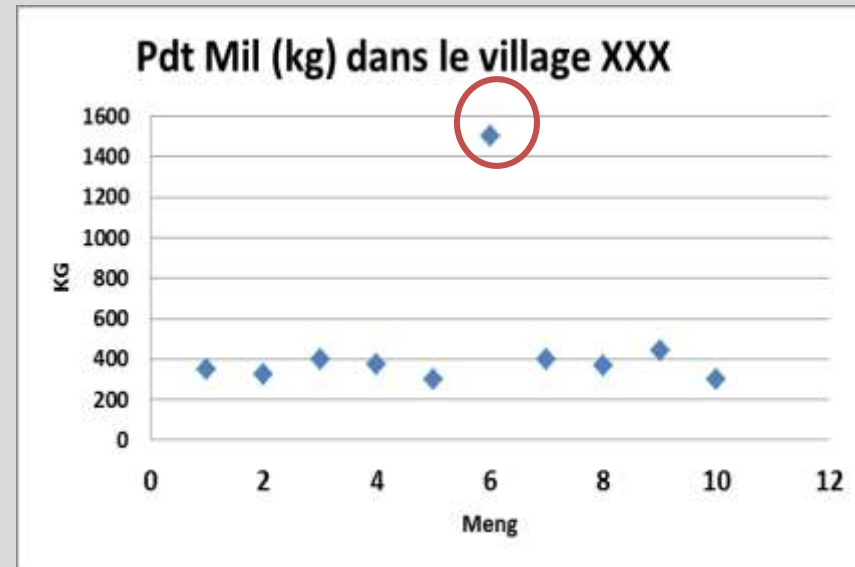


Traitement de données (Etapes)

- 4 étapes pendant lesquelles, les erreurs pourraient se produire/être gérées: Toilettage, Saisie, Nettoyage et Estimation.
 - a) Le **toilettage** des données implique un contrôle préalable avant d'entrer les données (**sur les fiches**)
 - b) **Saisie** des données– entrer les données dans la base de données. Cela peut conduire à l'introduction d'erreurs dans la base de données
 - c) Édition / **nettoyage** - le but est d'éliminer et / ou réduire les erreurs
 - d) **Estimation** – qui permet de repérer les erreurs

Traitement de données – valeurs aberrantes

- **valeur aberrante.** une valeur qui semble "non logique" car très différente des autres valeurs d'un ensemble de données.
 - Il pourrait s'agir d'un cas réel ex. Données d'un ménage très riche (Chef canton).
 - Elle peut être aussi une erreur opérée lors de la collecte ou de la saisie de données;
 - De manière générale, il est préférable de l'exclure en dernier ressort car celle-ci pourrait influencer les résultats de l'analyse des données (Graphique illustratif)

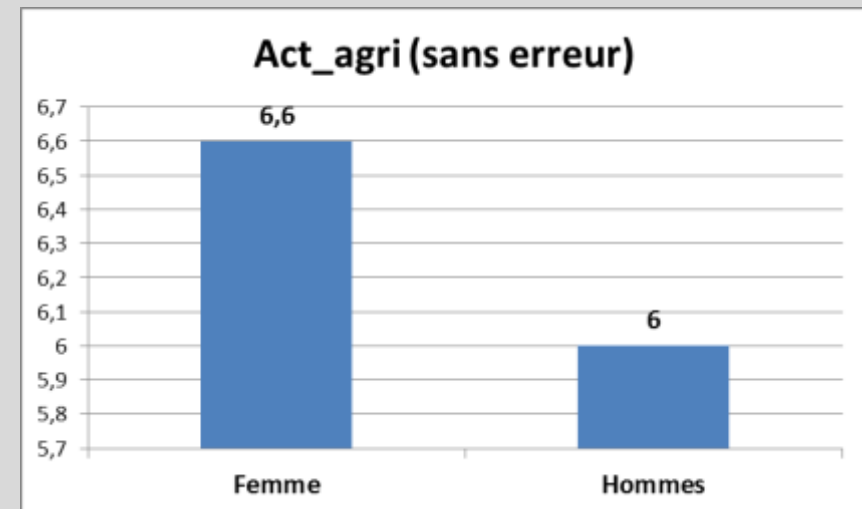
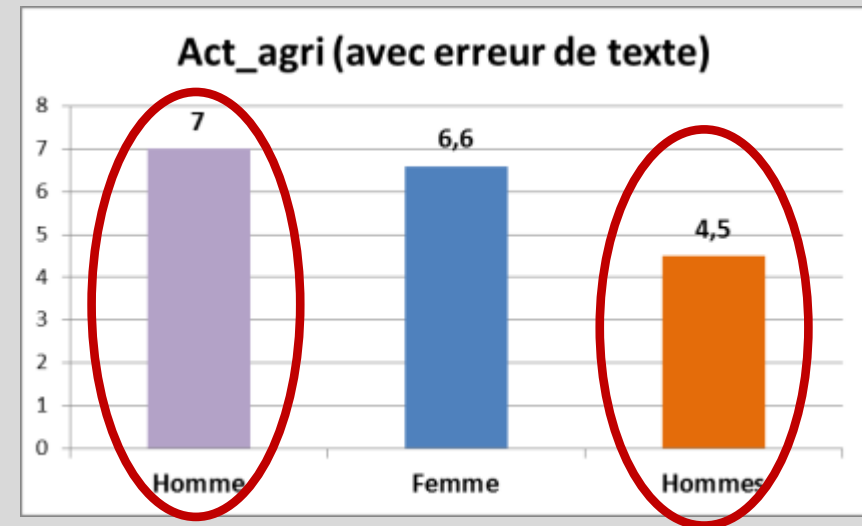


Traitement de données– Erreurs

- **Erreur**: une déviation de précision ou d'exactitude; une faute. Les erreurs peuvent être introduites au cours de la collecte des données ou pendant la saisie de données.

Les erreurs de texte conduisent normalement au fractionnement des données lors de l'analyse. Il est commun avec des noms de lieux, de la variété, etc.

Les erreurs numériques peuvent influencer l'analyse statistique telle que: somme, moyenne, etc.



Traitement de données– Erreurs

- Erreur courante lors de la saisie de données
 - ✓ **Saisie de données dans la « mauvaise cellule » ligne**
 - Cette erreur peut être réduite en cachant le dossier déjà saisie
 - ✓ **Duplication des dossiers – Saisie du même dossier plusieurs fois**
 - Étiqueter chaque questionnaire avec un numéro unique et éviter d'avoir deux ou plusieurs questionnaires avec le même numéro
 - ✓ **Saisie des données dans un mauvais champ (colonne)**
 - Ceci peut être minimisé en validant les champs ou en masquant les autres colonnes ou en utilisant des couleurs différentes pour la colonne

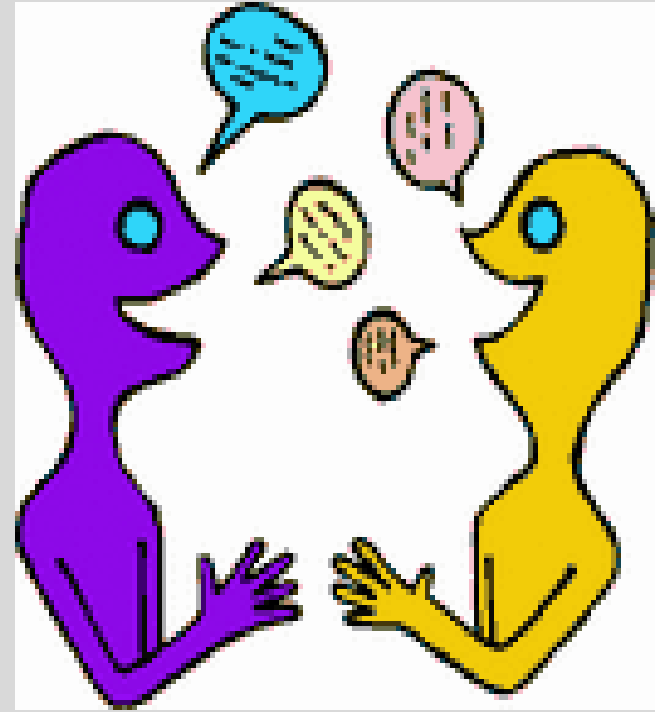
Traitement de données– Erreurs

- Erreur courant durant la saisie de données
 - d) Utilisation de 2 ou plusieurs unités de mesure. Ex. 200 g et 20 kg – très commun dans la saisie des quantités. Utiliser une seule unité: Ex: 200g et 20000g ou 0.2kg et 20kg
 - ✓ Conversion de toutes les unités en une seule et attirer l'attention des enquêteurs sur cet aspect.
 - ✓ Éviter de mélanger les données quantitatives et les textes (ie. 20 kg), entrer comme donnée : 20
 - e) Erreurs de texte - erreurs de frappe commises par l'opérateur de saisie. Celles-ci peuvent être minimisées en : Validant les cellules / Copier et coller / Filtre

Erreur de données

- **Discussion:** Un opérateur de saisie entre les surfaces (Ha) plantées en mil par 10 paysans: **2 - 3,5 - 45 - 2 - 4 - 1 - 1,5 - 10 - 3 et 4,5**. Dans region de GAO, les parcelles paysannes varient normalement entre **1 et 6** Ha:

- a) Par paires: identifier le/les erreurs, et discuter de la possible source d'erreur (des données).
- b) Comment voulez-vous gérer cette erreur?



Nettoyage de données, contrôle et vérification

Bonne trousse à outils et compétences



Im not obsessed.
i'm an editor!

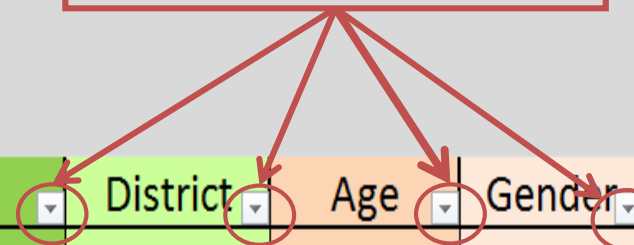
Definitions

- a) **Le nettoyage des données** – Permet de s'assurer qu'il n'y a pas de données incohérentes, des erreurs ou des valeurs aberrantes sous les variables ou des dossiers répétés
- b) **Modification des données** – Procéder à la correction
- c) **Vérification des données**: - Pour établir le bien-fondé, l'exactitude, ou la légitimité d'un ensemble de données. Il peut se faire par la mise en relation avec d'autres preuves (variables).
 - Dans Microsoft Excel, des filtres de données permettent d'effectuer le nettoyage et l'édition (modification) des données.

Nettoyage et modification des données utilisant le filtre

1. Mettre en **surbrillance** toutes les variables
2. Aller au menu de données et cliquez sur **l'icône trier & filtrer**, puis choisir **filtrer** : Un menu déroulant apparaîtra sur le côté droit de chaque case de variable.
3. Cliquez sur le menu déroulant et faire les observations nécessaires

Menu déroulant



Name	District	Age	Gender
Andezu Monica	Arua	28	0
Asibazuyo Christe	Arua	20	0
Bezu Gloria	Arua	21	0
Driciru Magret	Arua	41	0
Agotre Onesmas	Koboko	65	1
Agotre Stephen	Koboko	63	1
Asibazuyo Fotina	Koboko	39	0
Sabo Fred	Koboko	22	1
Drateru Phibi	Moyo	23	0
Madelena Ezaru	Moyo	62	0
Andioku Peter	Moyo	26	1
Elimasla Eidah	Moyo	30	0

4) Une fois que vous avez identifié les données aberrantes ou incohérentes, décochez toutes les cases, puis sélectionnez les données incohérentes ou aberrantes. Cliquez sur OK. Seules les lignes sur lesquelles sont les données incohérentes apparaissent.

5) Vérifiez le numéro de saisie de données correspondant aux données incohérentes ou aberrantes identifiées, et retour à la copie papier du questionnaire et corriger.

6) Lorsque l'incohérence ou la valeur aberrante est inscrite dans le questionnaire, alors consulter l'enquêteur ou le chef d'équipe pour la correction.

Procédure de nettoyage des données

7. Dans un cas de figure où, ni le questionnaire, ni l'enquêteur/chef d'équipe ne peut apporter la correction, alors il est préférable de supprimer la donnée si celle-ci a une incidence sur le résultat final.

Filtre numérique

- **Pour appliquer le filtre numérique, procéder de la manière suivante.**
 1. Cliquer sur l'icone de la bande déroulante d'une colonne de données numériques
 2. Cliquer sur l'option filtre numérique et sélectionner la commande appropriée. Ex: égal à

Note

- Vous pouvez également afficher la cellule "**Egale à la valeur**", "**est supérieur à** "..... Vous pouvez aussi afficher des données - **égal à une valeur** - **inférieur à une valeur** - **entre deux valeurs** - **les valeurs les plus élevées**, etc.
- Cela pourrait être intéressant pour l'identification des valeurs aberrantes, etc.

Filtre Textuel

- **Pour appliquer le filtre de texte, procédez de la manière suivante.**
 1. Cliquer sur l'icone de la bande déroulante d'une colonne de données textuelles
 2. Cliquer sur l'option **filtre textuel** et sélectionner la commande appropriée. Ex: égal à, commençant par, etc.

OU

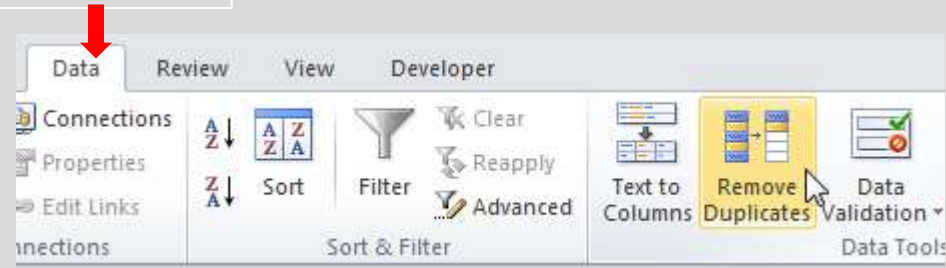
1. Entrez le texte ou le nom dans la cellule "**recherche**". Le résultat s'affichera et cliquer sur OK
- **Note:** Vous pouvez aussi afficher les cellules commençant par des caractères spécifiques, finissant par des caractères spécifiques, etc.

Suppression des doublons

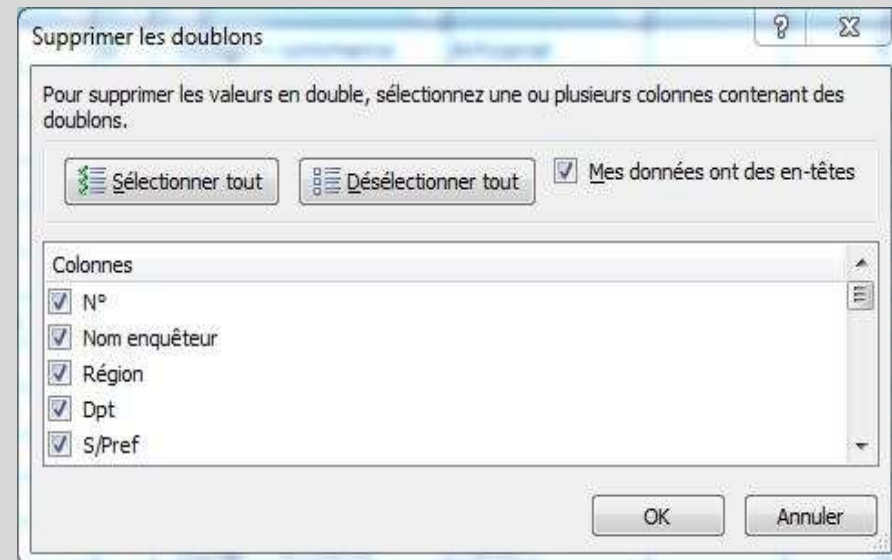
- **Les Doublons** sont des enregistrements entrés plus d'une fois dans la base de données. Ils tendent à influencer les analyses statistiques (Sommes, moyenne, nombre, etc.)

- **Supprimer les doublons**

1. Cliquer sur n'importe quelle cellule du champ de données.
2. Cliquer sur "**Données**" dans la barre d'outil et sur supprimer les doublons.



La boîte de dialogue apparaît avec les niveaux de champs.



3. Laisser toutes les cases cochées et cliquez sur OK

Verification de données

1. Certaines variables sont issues d'autres variables.
Par exemple
 - i. La quantité de semences plantées ou la production est normalement liée à la superficie cultivée et aux cultures plantées. Les variables dérivées telles que la dose de semis (semences plantées/surface) ou le rendement, pourraient fournir une indication utile de la validité de l'ensemble des données recueillies.
 - ii. Les noms des variétés sont normalement liés aux noms des cultures. Regarder les deux variables simultanément aide au nettoyage des données.

Variables dérivées

- **Variables dérivées-**

Sont celles qui sont générées à partir de 2 ou d'un ensemble de variables: exemples

a) $Rdt = Pdt \div Surface$

b) $Dose\ de\ semis = Qté\ de\ semences\ utilisée \div surface\ cultivée$

Etales

1. Insérer une nouvelle colonne après l'une des variables
2. Nommer la variable (Entête de colonne)
3. Appliquer la formule appropriée dans la première cellule en dessous de champ
4. Etendre la formule à toutes les cellules suivantes

Variables dérivées [formules d'erreurs]

a) ##### → Problème de largeur de cellule

- ✓ Elargir ou double clics dans le champ de la colonne

	A2	f _x	15000000
	A	B	C
1	7,500,000		
2	#####		
3	500,000		
4			
5			

b) #NAME? Se produit lorsqu'Excel ne reconnaît pas le texte dans une formule.

- ✓ Corriger la formule en remplaçant SU par SUM

	A4	f _x	=SU(A1:A3)
	A	B	C
1	4		
2	5		
3	3		
4	#NAME?		
5			
6			

Variables dérivées [formules d'erreurs]

c) **#VALUE!**. Se produit lorsque la formule comporte une variable non conforme

- ✓ Changer la cellule (A3) par un nombre, ou utilisez la formule `{=SUM(A1:A3)}` pour ignorer la cellule contenant le texte

c) **#DIV/0!**. Se produit lorsque le diviseur est une cellule comportant la valeur 0 ou vide

- ✓ Eviter l'erreur en utilisant la fonction logique IF `{=IF(A2=0,"",A1/A2)}`
- ✓ Ignore et supprimer les cellule avec #DIV/0!
- ✓ Remplacer la valeur de (A2) par une valeur différente de 0.

A4		fx		=A1+A2+A3	
	A	B	C		
1	4				
2	5				
3	Hi				
4	#VALUE!				
5					
6					

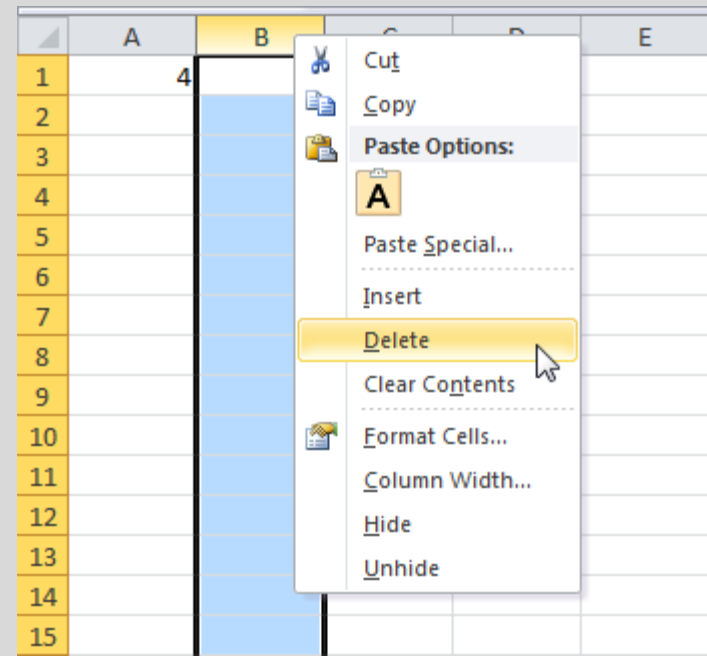
A3		fx		=A1/A2	
	A	B	C		
1	4				
2	0				
3	#DIV/0!				
4					
5					

Variables dérivées [formules d'erreurs]

c) #REF!. Survient quand une formule se réfère à une cellule non valide. Ex. La cellule C1 est liée à cellule A1 et cellule B1.

- Cette Erreur apparait lorsqu'une colonne/ligne contenant une des cellule referante est supprimée
- Pour corriger cette erreur, vous pouvez soit supprimer + # REF! dans la formule de la cellule B1 ou vous pouvez annuler votre action en cliquant sur Annuler dans la barre d'outils (ou appuyez sur Ctrl + Z).

	C1		f_x	=A1+B1
	A	B	C	
1	4	6	10	
2				
3				



Exercise 3.1: Nettoyage et modification des données

- Vous avez reçu la base de données de l'ESS réalisée dans les régions de Gao et Tombouctou (Mali) en Mars 2015.
- 1) Dans cette base de données, identifier les données aberrantes, les erreurs de textes/numériques dans les champs et dossiers. Quelle serait l'impact de ces erreurs sur les champs? Et que ferez-vous pour les minimiser?
 - 2) Existe-t-il des duplications dans les données? Si oui, combien en existe-t-il? Comment allez-vous supprimer les données dupliquées?

Exercice 3.2: Dérivation et vérification des données

- Vous avez reçu la base de données de l'ESS réalisée dans les régions de Gao et Tombouctou (Mali) en Mars 2015.
1. Insérer la variable "Dose de semis" (à une décimale) relative à la **CultprinA (Culture A saison dernière)**. Identifier et expliquer le type de formule d'erreur présent. Suggérer des moyens possibles de corriger les erreurs.
 2. Identifier toute donnée aberrante survenue dans le calcul de la dose de semis. Expliquez comment cela a pu se produire et comment comptez-vous résorber cette situation?
 3. Exemple de vérification
 4. Calculer la "dose de semis" relative à la **cultprinA (Culture A saison prochaine)**. Comparer ces valeurs à celles de 2014 et catégorisez ces changements comme: Positif / négatif / pas de changement dans un autre champs.