acaps

Note technique
Avril 2016

NETTOYAGE DES DONNÉES

TABLE OF CONTENTS

La traduction en français de ce document a été réalisée par CartONG, et a été soutenue par l'Agence Française de Développement (AFD) et le Centre de crise et de soutien du Ministère de l'Europe et des Affaires étrangères (CDCS). Néanmoins, les idées et les opinions présentées dans ce document ne représentent pas nécessairement celles de l'AFD ou du CDCS.

INTRODUCTION

Quel que soit le mode de collecte des données (entretiens en face à face, entretiens téléphoniques, questionnaires auto-administrés, etc.), il y aura un certain niveau d'erreur. Les "données désordonnées" désignent des données remplies d'incohérences. Si certaines de ces incohérences sont justifiées car elles reflètent des différences contextuelles, d'autres sont probablement le résultat d'une erreur de mesure ou de saisie. Il peut s'agir d'erreurs humaines, de systèmes d'enregistrement mal conçus ou simplement d'un contrôle incomplet du format et du type de données importées depuis des sources de données externes. De telles divergences font des ravages lorsque l'on essaie d'effectuer des analyses de données. Avant de traiter les données en vue de leur analyse, il faut veiller à ce que celles-ci soient aussi précises et cohérentes que possible.

Utilisés principalement lorsqu'il s'agit de données stockées dans une base de données, les termes validation des données, nettoyage des données ou traitement des données désignent le processus de détection, de correction, de remplacement, de modification ou de suppression des données douteuses d'un ensemble d'enregistrements, d'une table ou d'une base de données.

Ce document fournit des conseils aux analystes de données pour trouver la bonne stratégie de nettoyage des données lorsqu'ils traitent des données d'évaluation des besoins. Ces conseils s'appliquent à la fois aux données primaires et secondaires, et couvrent les situations où:

- Les données brutes sont générées par les équipes d'évaluation à l'aide d'un questionnaire.
- Les données sont obtenues à partir de sources secondaires (systèmes de suivi des déplacements, données sur la sécurité alimentaire, données de recensement, etc.).
- Les données secondaires sont comparées ou fusionnées avec les données obtenues lors des évaluations sur le terrain.

Ce document complète la note technique d'ACAPS ''How to approach a dataset'' qui détaille spécifiquement les opérations de nettoyage des données primaires saisies dans une feuille de calcul Excel lors des évaluations rapides.

A. PROCESSUS DE NETTOYAGE DES DONNÉES

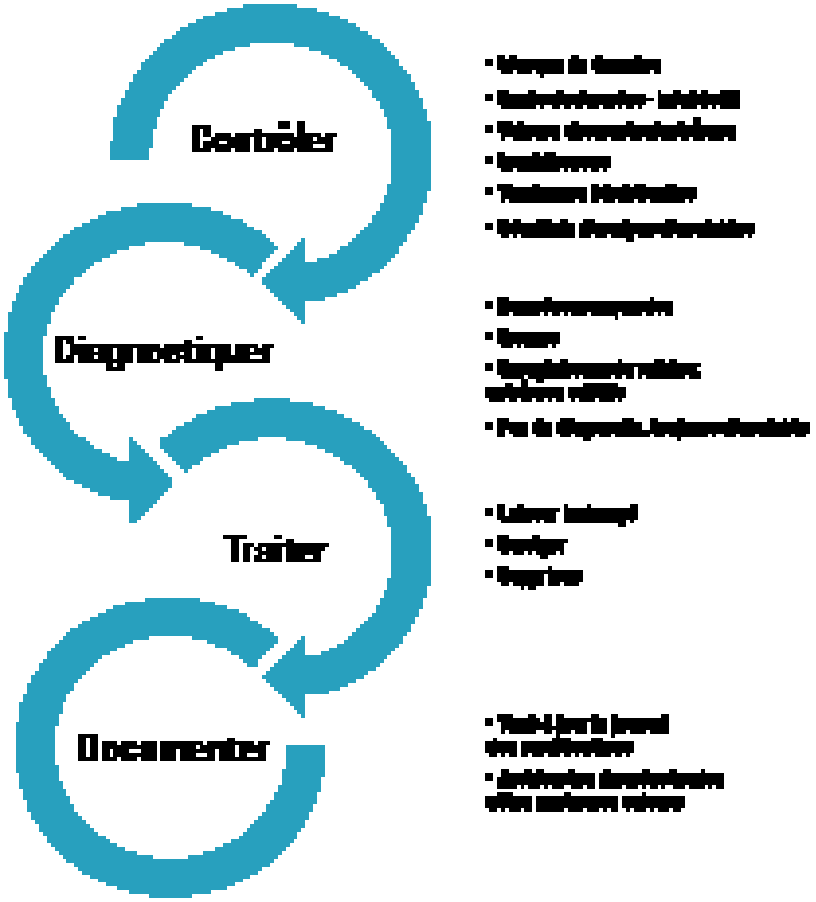
Le nettoyage des données consiste principalement à mettre en œuvre des stratégies de prévention des erreurs avant qu'elles ne se produisent (voir les procédures de contrôle de la qualité des données plus loin dans le document). Cependant, ces stratégies de prévention des erreurs peuvent réduire, sans pour autant éliminer les erreurs courantes. De nombreuses erreurs de données seront détectées de manière fortuite :

- Lors de la collecte ou de la saisie de données
- Lors de la transformation/de l'extraction/du transfert des données
- Lorsque de l'exploration ou de l'analyse des données
- Lors de la soumission d'un rapport pour l'examen par les pairs

Même avec l'application des meilleures stratégies de prévention des erreurs, il sera toujours nécessaire de rechercher, de détecter et de corriger activement et systématiquement les erreurs/problèmes de manière planifiée.

Le nettoyage des données implique des cycles répétés de contrôle, de diagnostic, de traitement et de documentation de ce processus. Au fur et à mesure que des schémas d'erreurs sont identifiés, les procédures de collecte et de saisie des données doivent être adaptées pour corriger ces schémas et réduire les erreurs futures.

Les quatre étapes du nettoyage des données:



Adapté de Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005) et Arthur D. Chapman.

Contrôler consiste à rechercher systématiquement les caractéristiques suspectes dans les questionnaires d'évaluation, les bases de données ou les ensembles de données d'analyse.

Diagnostiquer (identifier la nature des données erronées) et **nettoyer** des données (supprimer, modifier ou maintenir des données en l'état) nécessitent une compréhension approfondie de tous les types et sources d'erreurs possibles lors des processus de collecte et de saisie des données.

Documenter des modifications implique de laisser une trace des erreurs détectées, des modifications et ajouts effectués, et de la vérification des erreurs afin de permettre un retour à la valeur initiale, si nécessaire.

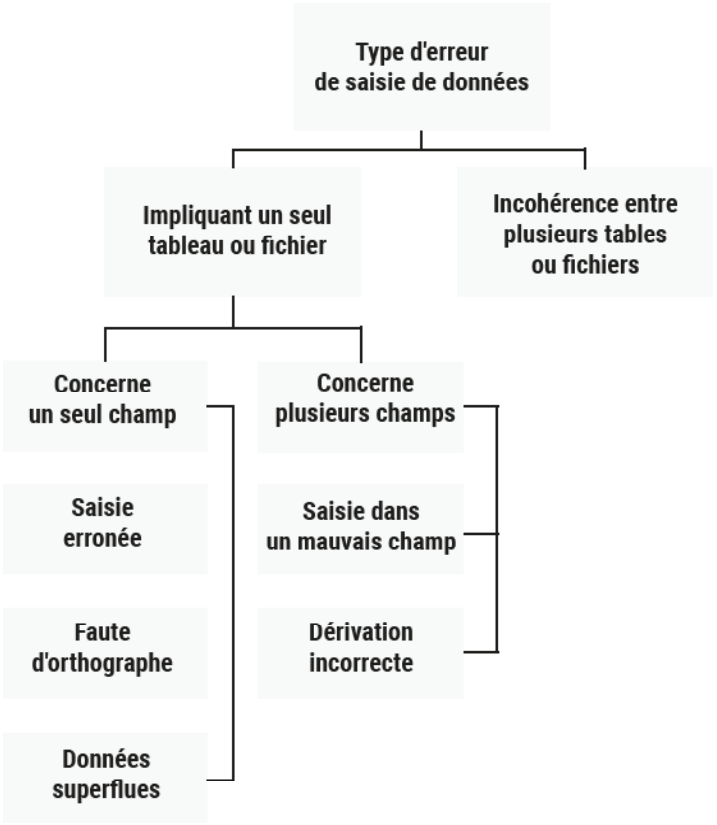
B: SOURCES D'ERREUR

Après la mesure, les données font l'objet d'une séquence d'activités typiques : elles sont saisies dans des bases de données, extraites, transférées vers d'autres tableaux, éditées, sélectionnées, transformées, agrégées, synthétisées et présentées. Il est important d'avoir en tête que des erreurs peuvent se produire à chaque étape du flux de données, y compris pendant le nettoyage des données. De nombreuses sources d'erreurs dans les bases de données entrent dans une ou plusieurs des catégories suivantes:

Erreurs de mesure: Les données sont généralement destinées à mesurer un processus physique, des sujets ou des objets, tels que le temps d'attente au point d'eau, la taille d'une population, l'incidence des maladies, etc. Dans certains cas, ces mesures sont effectuées par des processus manuels qui peuvent comporter des erreurs systématiques ou aléatoires dans leur conception (par exemple, des stratégies d'échantillonnage inappropriées) et leur exécution (par exemple, une mauvaise utilisation des instruments, un biais, etc.) L'identification et la résolution de telles incohérences dépassent le cadre de ce document. Il est recommandé de se référer à la note technique ACAPS ‘‘How sure are you?’’ pour comprendre comment traiter les erreurs de mesure.

Erreur de saisie des données: "La saisie des données" est le processus de transfert des informations du support qui enregistre la réponse (traditionnellement les réponses écrites sur des questionnaires imprimés) vers une application informatique. Du fait des contraintes temporelles, ou du manque de supervision ou de contrôle adéquat, les données sont souvent altérées au moment de la saisie.

Les principales erreurs sont les suivantes:



Adapté de Kim et Al, 2003; Aldo Benini 2013

- Une saisie erronée se produit si, par exemple, l'âge est mal saisi (par exemple, 26 au lieu de 25).
- Les données superflues ajoutent des informations correctes, mais non souhaitées, par exemple le nom et l'intitulé de poste dans un champ réservé au nom.
- La dérivation incorrecte se produit lorsqu'une fonction a été calculée de manière incorrecte pour un champ dérivé (par exemple, une erreur dans l'âge dérivé de la date de naissance).
- Des incohérences entre les tables ou les fichiers se produisent, par exemple lorsque le nombre de sites visités dans la table de la province et le nombre de sites visités dans la table de l'échantillon total ne correspondent pas.

Une grande partie des erreurs de saisie des données peut être évitée en utilisant un formulaire électronique (par exemple, ODK) et une saisie conditionnelle.

Erreurs de traitement: Dans de nombreux contextes, les données brutes sont prétraitées avant d'être saisies dans une base de données. Ce traitement des données est effectué pour diverses raisons : pour réduire la complexité ou le bruit des données brutes, pour agréger les données à un niveau supérieur et, dans certains cas, simplement pour réduire le volume des données stockées. Tous ces processus sont susceptibles de produire des erreurs.

Erreurs d'intégration des données: Il est rare qu'une base de données d'une taille et d'une ancienneté significatives contienne des données provenant d'une seule source, collectées et saisies de la même manière au fil du temps. Très souvent, une base de données contient des informations collectées à partir de sources multiples et par des méthodes multiples au cours du temps. Par exemple, le suivi du nombre de personnes affectées tout au long de la crise, où la définition du terme "affecté" est affinée ou modifiée au fil du temps. En outre, dans la pratique, de nombreuses bases de données évoluent en fusionnant d'autres bases de données préexistantes. Cette tâche de fusion nécessite presque toujours une tentative de résolution des incohérences entre les bases de données impliquant différentes unités de données, périodes de mesure, formats, etc. Toute procédure qui intègre des données provenant de sources multiples peut entraîner des erreurs. La fusion de deux ou plusieurs bases de données permettra à la fois d'identifier des erreurs (lorsqu'il y a des différences entre les deux bases de données) et de créer de nouvelles erreurs (des enregistrements en double). Le tableau 1 ci-dessous illustre certaines des sources et des types d'erreurs possibles au cours d'une grande évaluation, à trois niveaux:

- Lors du remplissage du questionnaire;
- Lors de la saisie des données dans la base de données;
- Lors de l'analyse.

Tableau 1: Sources d'erreurs dans les données

	SOURCES D'ERREURS DANS LES DONNÉES	
ÉTAPE	MANQUE OU EXCÈS DE DONNÉES	VALEURS ABERRANTES ET INCOHÉRENCES
Collecte - Mesure	• Formulaire manquant • Formulaire en double, collecté plusieurs fois • Zone de réponse ou options laissées vides • Plus d'une option sélectionnée alors que ce n'est pas autorisé	• Valeur correcte remplie dans la mauvaise case • Non lisible • Erreur d'écriture • Réponse donnée en dehors de la fourchette acceptable (conditionnelle)
Saisie	• Manque ou excès de données transférées du questionnaire • Format du champ non renseigné • Valeur saisie dans le mauvais champ • Suppression et/ou duplication par inadvertance lors de la manipulation de la base de données	• Valeurs aberrantes et/ou incohérences provenant du questionnaire • Valeur incorrectement saisie, faute d'orthographe • Valeur incorrectement modifiée lors d'un précédent nettoyage des données • Erreur de conversion (programmation)
Traitement et analyse	• Manque ou excès de données extraites de la base de données • Erreur d'extraction, de codage ou de transfert de données • Suppression et/ou duplication par l'analyste	• Valeurs aberrantes et incohérences provenant de la base de données • Erreur d'extraction, de codification ou de transfert de données • Erreurs de tri (feuilles de calcul) • Erreurs de nettoyage des données

Adapté de Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005)

L'inexactitude d'**une seule** mesure et d'**un seul** point de données (une seule valeur) reste acceptable, et peut venir d'une erreur technique inhérente à l'instrument de mesure. Par conséquent, le nettoyage des données devrait se concentrer sur les erreurs qui vont au-delà des petites variations techniques et qui produisent un changement majeur dans l'analyse. De même, et par soucis de contrainte temporelle, considérez l'utilité marginale décroissante d'un nettoyage de plus en plus poussé par rapport à d'autres tâches exigeantes telles que l'analyse, la visualisation et l'interprétation des données. Pour cela :

- Comprendre quand et comment les erreurs sont produites pendant la collecte des données et le flux de travail.
- Garder en tête que les ressources pour le nettoyage des données sont limitées. La priorisation des erreurs liées aux effectifs de la population, à la localisation géographique, aux catégories affectées et à la date est particulièrement importante car celles-ci sont susceptibles de fausser les variables dérivées et l'analyse finale.

Les sections suivantes de ce document proposent une approche étape par étape du nettoyage des données.

C. CHAQUE CHOSE EN SON TEMPS

La première chose à effectuer est de faire une copie des données brutes (données initiales) dans un fichier séparé et de nommer les feuilles de manière appropriée, ou de les enregistrer dans un nouveau fichier.

Conservez TOUJOURS les fichiers sources dans un dossier séparé et changez leur paramétrage en LECTURE SEULE, pour éviter toute modification de l'un des fichiers.

D. CONTRÔLE DES DONNÉES

Pour préparer les données en vue de leur contrôle, il faut mettre de l'ordre dans le jeu de données en les transformant dans un format facile à utiliser.

Dans un jeu de données nettoyé:

- Les polices de caractères ont été harmonisées;
- Le texte est aligné à gauche, les chiffres à droite;
- Chaque variable a été transformée en une colonne et chaque observation en une ligne;
- Il n'y a pas de lignes vides;
- Les en-têtes de colonnes sont clairs et visuellement distincts;
- Les espaces de début de ligne ont été supprimés.

Ensuite, examinez les données pour détecter les erreurs possibles suivantes:

- Incohérences en matière d'orthographe et de formatage: Les variables catégorielles sont-elles écrites correctement? Le format de la date est-il cohérent et harmonisé ? Pour les champs numériques, toutes les valeurs sont-elles des nombres? Etc.
- Manque de données: Certaines questions ont-elles beaucoup moins de réponses que d'autres?
- Excès de données: Y a-t-il des entrées en double ou plus de réponses que celles initialement prévues?
- Valeurs aberrantes/extrêmes: Y a-t-il des valeurs qui se situent si loin de la distribution typique qu'elles semblent potentiellement erronées?
- Schémas remarquables: Existe-t-il des schémas qui suggèrent que le répondant ou l'enquêteur n'a pas répondu ou enregistré les questions honnêtement (par exemple, plusieurs questionnaires avec exactement les mêmes réponses)?
- Analyse des points suspects: Les réponses à certaines questions semblent-elles contre-intuitives ou extrêmement improbables?

Erreurs courantes dans les évaluations des besoins:

- Mauvaise orthographe des noms de lieux, en particulier lors de la traduction entre différents alphabets (par exemple, de l'arabe vers l'anglais).
- Utilisation de formats de date incohérents.
- Les totaux diffèrent des résultats des questions désagrégées (par exemple, le nombre total de membres du ménage ne correspond pas à l'agrégation d'une autre question où il est demandé aux répondants d'énumérer les membres du ménage par âge et par sexe).
- Les valeurs sont en dehors de la fourchette acceptable pour cette question, y compris les valeurs négatives dans les champs qui ne peuvent avoir que des valeurs positives (par exemple, le prix du pain).
- Cause imprécise du manque de données.
- Fusion d'ensembles de données comportant des unités de mesure différentes (par exemple, différentes interprétations du terme "ménage") ou des limites administratives.
- Dans le cas de questions à choix multiples : la sélection de " Autre, veuillez préciser " pour une variable qui est l'une des options à choix multiples.
- Mauvais fonctionnement des sauts de questions.
- Manque général de cohérence dans les réponses fournies par un répondant (par exemple, la réponse aux questions sur les principaux besoins n'est pas conforme aux questions spécifiques au secteur).

Le nettoyage des données peut être partiellement automatisé grâce à des logiciels statistiques. Les outils de statistique descriptive peuvent par exemple être utilisés pendant la phase de contrôle des données pour prédéfinir les attentes, les hypothèses ou les critères concernant les plages normales, les formes de distribution et la force des relations. Cela peut faciliter le repérage de données, de schémas ou de résultats discutables.

Cependant, les méthodes de contrôle ne sont pas seulement statistiques. De nombreuses valeurs aberrantes sont détectées par la perception d'une non-conformité aux attentes définies au préalable ou d'un non-conformité à la norme. Ceci est par exemple basé sur l'expérience de l'analyste, les résultats de l'analyse des données secondaires, les contraintes numériques ou le bon sens (le poids ne peut pas être négatif, les gens ne peuvent pas avoir plus de deux parents, les femmes ne peuvent pas avoir 35 enfants, etc.).

Un problème particulier est celui des valeurs erronées non-aberrantes, c'est-à-dire des valeurs générées par erreur, mais se situant dans la fourchette attendue. Les valeurs erronées non-aberrantes ne sont souvent pas détectées. Les stratégies de détection comprennent:

- La visualisation les données par rapport à d'autres variables, en utilisant des vues multivariées, telles que des diagrammes de dispersion ou des cartes thermiques.
- Des techniques plus avancées et plus exigeantes en ressources : une analyse de régression, des contrôles de cohérence/plausibilité (examen de l'historique de chaque point de données ou comparaison avec un lieu similaire), ou une nouvelle mesure. Cependant, en tenant compte de la contrainte temporelle, un tel examen est rarement réalisable. En revanche, il est possible d'examiner et/ou de remesurer ou d'effectuer une enquête plus approfondie sur un échantillon de valeurs aberrantes afin d'estimer le taux d'erreur.

Les méthodes de détection utiles, des plus simples aux plus complexes, sont les suivantes :

- Contrôle des colonnes après les avoir classées;
- Utilisation des statistiques récapitulatives;
- Saisie validée et/ou doublons des saisies;
- Mise en évidence des variables qui ne sont pas inclues dans la fourchette acceptable et des enregistrements qui ne sont pas cohérents;
- Distribution de fréquences et tableaux croisés;
- Analyse des distributions sous forme de graphiques: diagrammes en boîte, histogrammes et diagrammes de dispersion - à l'aide de logiciels d'analyse visuelle tels que Tableau.
- Graphiques de mesures répétées sur un même individu - par exemple, des courbes de croissance.
- Vérification des questionnaires à l'aide d'algorithmes dédiés.
- Détection statistique des valeurs aberrantes/extrêmes.

- Dans de nombreux cas, sinon la plupart, les données ne peuvent être nettoyées efficacement qu'avec une certaine participation humaine. Connaissez les erreurs courantes et les erreurs à rechercher (et/ou formez les personnes chargées du nettoyage des données).
- L'analyse exploratoire des données et la visualisation des données sont les deux moyens principaux pour détecter les erreurs de données.
- Les différents types d'erreurs nécessitent le recours à des approches de détection différentes - un correcteur orthographique reconnaîtra les variables catégorielles mal orthographiées, tandis que la détection statistique des valeurs aberrantes permet d'identifier les valeurs extrêmes.

E. DIAGNOSTIC DES DONNÉES

L'identification ou la mise en évidence d'une erreur est suivie d'un diagnostic - trouver la cause de cette erreur. Pour comprendre des données discutables, examinez toutes les réponses d'un répondant pour déterminer si les données ont un sens dans le contexte. Il est parfois nécessaire d'examiner un échantillon de réponses de différents répondants, afin d'identifier des problèmes tels qu'une règle de saut de question qui a été mal spécifiée.

Il existe une multitude de diagnostics possibles pour chaque valeur discutable:

- Données manquantes: Des réponses omises par le répondant (non-réponse), des questions sautées par l'enquêteur ou un abandon.
- Erreurs: Les fautes de frappe ou les réponses qui indiquent que la question a été mal comprise.
- Extrêmes validés: Une réponse qui semble élevée mais qui peut être justifiée par d'autres réponses (par exemple, le répondant travaille 60 heures par semaine parce qu'il a un emploi à temps plein et un emploi à temps partiel).
- Normaux validés: Un enregistrement valide.
- Pas de diagnostic, toujours suspect: Faites un choix sur la façon de traiter ces données pendant la phase de traitement.

Certaines valeurs de données sont logiquement ou biologiquement impossibles (par exemple, les personnes de sexe masculin ne peuvent pas être enceintes; le prix du pain ne peut pas être négatif). Des seuils prédéfinis permettent de détecter immédiatement ce type d'erreur. Parfois, la valeur discutable se situe dans la fourchette acceptable, ce qui rend le diagnostic moins évident. Dans ces cas, il est nécessaire d'appliquer une combinaison de procédures de diagnostic:

- **Revenir aux étapes précédentes** du flux de données pour voir si une valeur est toujours la même. Pour cela, il faut avoir accès à des données bien archivées et documentées, avec des justifications pour tout changement effectué à chaque étape.
- **Recherchez des informations** qui pourraient confirmer le véritable statut extrême d'une valeur aberrante. Par exemple, un score très bas pour le poids par rapport à l'âge (avec des Z-scores de -6) peut être dû à des erreurs de mesure de l'âge ou du poids, ou le sujet peut être extrêmement mal nourri, auquel cas les autres variables nutritionnelles devraient également présenter des valeurs extrêmement basses. Ce type de procédure nécessite une connaissance de la cohérence et de la logique entre les variables. Cette connaissance s'acquiert généralement grâce à l'expérience ou aux enseignements tirés de précédents diagnostics. Elle peut être utilisée pour planifier et programmer le nettoyage des données.

- **Collecter des informations supplémentaires**, c'est-à-dire interroger l'enquêteur sur ce qui a pu se passer et, si possible ou nécessaire, répéter la mesure. Ces procédures ne peuvent avoir lieu que si le nettoyage des données commence peu après la collecte des données.

La phase de diagnostic exige beaucoup de travail et les besoins en matière de budget, de logistique, de temps et de personnel sont généralement sous-estimés, voire négligés, au stade de la conception. Moins de ressources sont nécessaires si la saisie des données est conditionnelle (par exemple, au moyen de formulaires électroniques) et si elle commence au début du processus de collecte des données.

- Utilisez le bon sens, l'expérience, la triangulation et les enseignements tirés de précédents diagnostics pour diagnostiquer le type d'erreur.
- Concevez votre questionnaire avec soin pour permettre des vérifications croisées entre les questions.
- Envisagez de collecter des informations supplémentaires auprès de l'enquêteur pour comprendre la cause des erreurs (via des débriefings).

F. TRAITEMENT DES DONNÉES

Après avoir identifié les valeurs manquantes, les erreurs et les valeurs réelles (extrêmes ou normales), les analystes doivent décider de ce qu'il faut faire avec les observations problématiques :

- **Ne rien changer:** La ligne de conduite la plus conservatrice consiste à accepter les données comme une réponse valide et à n'y apporter aucun changement. Plus la taille de l'échantillon est grande, moins une réponse discutable affectera l'analyse ; à l'inverse, plus la taille de l'échantillon est petite, plus la décision est difficile.
- **Corriger les données:** Si l'intention initiale de l'enquêté peut être déterminée, corrigez la réponse (par exemple, après discussion avec l'enquêteur, il est clair que l'enquêté voulait dire le manque de revenu au lieu de trop de revenu).
- **Supprimer les données?** Les données semblent illogiques et la valeur est si éloignée de la norme qu'elle affectera les statistiques descriptives ou inférentielles. Que faire ? Supprimer uniquement cette réponse ou supprimer l'ensemble de l'enregistrement ? N'oubliez pas que chaque fois que des données sont supprimées, il existe un risque de "sélection" consciente ou inconsciente des données pour obtenir les résultats souhaités. Pour comprendre l'impact de la suppression d'une valeur, on peut créer une variable binaire (1 = enregistrement discutable, 0 = ok, non discutable). Cette nouvelle variable peut être utilisée comme un filtre dans les tableaux croisés dynamiques ou directement dans les tables pour comprendre l'impact des données potentiellement erronées sur les résultats finaux.
- **Si le temps et les ressources le permettent, mesurer à nouveau** les valeurs discutables ou erronées.

Il existe quelques règles générales pour appuyer une décision sur la manière de traiter les données:

- Si la personne chargée de la saisie des données a saisi des **valeurs différentes de celles figurant dans le questionnaire**, la valeur doit être modifiée pour correspondre à celle enregistrée dans le formulaire du questionnaire (par exemple, la valeur figurant dans le questionnaire était de 40 000 et l'opérateur de saisie des données a saisi 4 000 - un zéro a été omis).
- Lorsque **les valeurs des variables n'ont pas de sens**, s'il n'y a pas d'erreur de saisie des données, et s'il n'y a pas de notes permettant de déterminer d'où vient l'erreur, **laissez les données telles quelles**. En modifiant la valeur pour obtenir un résultat plus raisonnable, on introduit un biais important et rien ne justifie cette modification. Le cas doit être répertorié comme une valeur aberrante (en utilisant le formatage conditionnel par exemple).
- Lorsque des **cellules vides** sont trouvées ou que l'enregistrement a été demandé même si les informateurs clés ne disposent pas de ce type de données ou que des enregistrements en double ont été saisis, les cas doivent être supprimés du fichier de données.
- Les **valeurs impossibles** ne sont jamais laissées telles quelles, mais doivent être corrigées si une valeur correcte peut être trouvée, sinon elles doivent être supprimées. Pour les variables biologiques continues (par exemple, la taille, le poids, etc.), une certaine variation pour un même individu, ou une petite variation de la mesure peuvent survenir. Si une nouvelle mesure est effectuée très rapidement après la mesure initiale et que les deux valeurs sont suffisamment proches pour être expliquées par la seule variation, prenez la moyenne des deux comme valeur finale.
- Dans le cas des **extrêmes validés** et des **valeurs qui restent discutables** après la phase de diagnostic, l'analyste doit examiner l'influence de ces valeurs, individuellement et en groupe, sur les résultats de l'analyse avant de décider de laisser ou non les données inchangées.
- Pour limiter l'impact des valeurs aberrantes et extrêmes, les analystes peuvent décider de présenter la médiane. Cela est acceptable à condition d'être clairement expliqué dans les conclusions.

- Certains auteurs ont recommandé que les valeurs extrêmes validées restent toujours dans l'analyse. Dans la pratique, de nombreuses exceptions sont faites à cette règle. L'enquêteur peut ne pas vouloir prendre en compte l'effet des valeurs extrêmes validées si elles résultent d'un processus étranger non-anticipé. Cela devient un critère d'exclusion "a posteriori". Les points de données doivent être signalés comme "exclus de l'analyse" dans le chapitre méthodologie du rapport final.

6. VALEURS MANQUANTES

Les valeurs manquantes nécessitent une attention particulière. La première chose à faire est de décider quelles cellules vides doivent être remplies de zéro (parce qu'elles représentent de véritables observations négatives, telles que ‘‘non’’, ‘‘non présent’’, ‘‘option non prise’’, etc.) et lesquelles doivent être laissées vides (si la convention est d'utiliser des blancs pour les valeurs manquantes ou ‘‘N/A’’ pour ‘‘non applicable’’). Certains analystes remplacent les cellules vides par un code de valeur manquante explicite (par exemple, en utilisant ‘‘999’’ pour indiquer un ‘‘ne sait pas’’).

Que faire des cellules qui restent vides ? Les valeurs manquantes peuvent être classées comme aléatoires ou non-aléatoires :

- Des valeurs manquantes aléatoires peuvent survenir parce que, par inadvertance, le sujet n'a pas répondu à certaines questions. L'évaluation peut être trop complexe ou trop longue, ou l'enquêteur peut être fatigué ou ne pas prêter suffisamment attention, et manquer la question. Des valeurs manquantes aléatoires peuvent également survenir en raison d'erreurs de saisie de données. S'il n'y a qu'un petit nombre de valeurs manquantes dans l'ensemble de données (généralement moins de 5 %), il est extrêmement probable qu'il s'agisse d'une valeur manquante aléatoire.
- Les valeurs manquantes non aléatoires peuvent être dues au fait que l'informateur clé a délibérément omis de répondre à certaines questions. Cela se produit par exemple si la question prête à confusion, n'est pas appropriée ou est perçue comme sensible. Les données manquantes sont liées à une ou plusieurs caractéristiques du répondant - par exemple, si les femmes sont plus susceptibles de refuser une question sur le niveau de revenu que les hommes.

L'option par défaut pour traiter les valeurs manquantes est de filtrer et d'exclure ces valeurs de l'analyse:

- **Suppression par liste ou par cas:** Tous les cas pour lesquels des valeurs manquent (par exemple, un répondant) sont exclus. Si une seule variable est analysée, la suppression par liste consiste simplement à analyser les données existantes. Si l'on analyse plusieurs variables, la suppression par liste élimine les cas si une valeur manque pour l'une des variables. L'inconvénient est la perte de données qui se produit lorsque toutes les données sont supprimées pour un seul cas, même si certaines questions ont reçu une réponse.
- **Suppression par paires:** Contrairement à la suppression par liste qui élimine les cas pour lesquels des valeurs manquent pour l'une des variables analysées, la suppression par paires n'élimine de l'analyse que les valeurs manquantes spécifiques (et non le cas, en entier). En d'autres termes, toutes les données disponibles sont inclues. Lorsque vous effectuez une corrélation sur plusieurs variables, cette technique permet d'établir une corrélation bivariée entre tous les points de données disponibles, et n'ignore que les valeurs manquantes si elles existent sur certaines variables. Dans ce cas, la suppression par paires donnera lieu à des tailles d'échantillon différentes pour chaque variable. La suppression par paires est utile lorsque la taille de l'échantillon est faible ou si les valeurs manquantes sont importantes parce qu'il n'y a pas beaucoup de valeurs au départ.

Essayez de réaliser le même test en utilisant les deux méthodes de suppression pour voir comment le résultat change. Notez que dans ces techniques, la "suppression" signifie l'exclusion dans une procédure statistique, et non la suppression (des variables ou des cas) de l'ensemble de données.

Une deuxième option consiste à **supprimer tous les cas ayant des valeurs manquantes.** Ainsi, vous vous retrouvez avec des données complètes pour tous les cas. L'inconvénient de cette approche est que la taille de l'échantillon de données est réduite, ce qui entraîne une perte de puissance statistique et une augmentation de l'erreur d'estimation (intervalles de confiance plus larges). Elle peut également affecter la représentativité d'un échantillon : après avoir retiré les cas avec des valeurs manquantes non-aléatoires d'un petit ensemble de données, la taille de l'échantillon pourrait être insuffisante. En outre, les résultats peuvent être biaisés en cas de valeurs manquantes non-aléatoires. Les caractéristiques des cas avec valeurs manquantes peuvent être différentes de celles des cas sans valeurs manquantes.

Une autre option est **l'imputation**: Il s'agit de remplacer les valeurs manquantes. Cette technique préserve tous les cas en remplaçant les données manquantes par une valeur probable basée sur d'autres informations disponibles. Une procédure simple d'imputation consiste à remplacer la valeur manquante par la moyenne ou la médiane. L'imputation hot-deck remplace les valeurs manquantes par la valeur de cette même variable tirée d'un enregistrement complet pour une personne similaire dans le même ensemble de données. Une fois que toutes les valeurs manquantes ont été remplacées, l'ensemble de données peut alors être analysé en utilisant les techniques standards pour les données complètes. Toutefois, cette méthode peut également biaiser les résultats et les p-valeurs.

Dans certaines conditions, les approches de vraisemblance maximale se sont également avérées efficaces pour traiter les données manquantes. Cette méthode ne remplace aucune donnée, mais utilise plutôt toutes les données disponibles pour les cas spécifiques afin de calculer les estimations du maximum de vraisemblance.

Le détail des aspects techniques, de la pertinence et de la validité de chaque technique dépasse le cadre de ce document. Finalement, le choix de la bonne technique dépend de la quantité de données manquantes, de la raison pour laquelle ces données sont manquantes, des schémas, du caractère aléatoire et de la distribution des valeurs manquantes, des effets des données manquantes et de la façon dont les données seront utilisées pour l'analyse. Il est fortement recommandé de faire appel à un statisticien dans le cas d'un petit ensemble de données avec un grand nombre de valeurs manquantes.

En pratique, pour l'évaluation des besoins avec peu de ressources statistiques, créer une copie de la variable et remplacer les valeurs manquantes par la moyenne ou la médiane peut souvent être suffisant et préférable à la perte de cas dans une analyse multivariée à partir de petits échantillons.

- Il existe plusieurs méthodes pour traiter les données manquantes, notamment la suppression des cas avec des valeurs manquantes, l'imputation et l'approche du maximum de vraisemblance. Cependant, fournir une explication sur la raison pour laquelle les données sont manquantes (par exemple, "les femmes n'ont pas pu être interrogées", "la dernière section du questionnaire n'a pas pu être remplie par manque de temps") peut être beaucoup plus informatif pour l'utilisateur final qu'une multitude de corrections statistiques.
- Créez une variable binaire dont la valeur est 0 pour ceux qui ont répondu à la question et la valeur 1 pour ceux qui n'y ont pas répondu. Utilisez cette variable pour montrer l'impact des différentes méthodes.
- Cherchez une signification aux valeurs manquantes non-aléatoires. Peut-être les personnes interrogées indiquent-elles quelque chose d'important en ne répondant pas à l'une des questions.

H. DOCUMENTATION DES CHANGEMENTS

La documentation des erreurs, des modifications, des ajouts et de la vérification des erreurs est essentielle pour:

- Maintenir la qualité des données
- Éviter la répétition des contrôles d'erreurs par différentes personnes en charge de nettoyer les données
- Récupérer les erreurs de nettoyage des données
- Déterminer la pertinence des données à être utilisées
- Informer les utilisateurs qui ont pu utiliser les données en sachant quelles modifications ont été apportées depuis leur dernier accès aux données

Créez un journal des modifications dans le classeur, dans lequel se trouvent toutes les informations relatives aux champs modifiés. Cela servira de registre montrant toutes les modifications, et permettra de revenir à la valeur d'origine si nécessaire. Dans le journal des modifications, enregistrez les informations suivantes:

- Table (si plusieurs tables sont utilisées)
- Colonne, Ligne
- Date de modification
- Modifié par
- Ancienne valeur
- Nouvelle valeur
- Commentaires / Observations

- Assurez-vous de documenter les étapes et procédures de nettoyage des données qui ont été mises en œuvre ou suivies, par qui, pour quelles questions et combien de réponses ont été affectées.
- Veillez à ce que ces informations soient toujours disponibles lorsque vous partagez l'ensemble de données en interne ou en externe (par exemple, en joignant le journal des modifications dans une feuille de calcul séparée).

I. PROCESSUS D'ADAPTATION

Une fois que les erreurs ont été identifiées, diagnostiquées, traitées et documentées et si la collecte/saisie des données est toujours en cours, la personne chargée du nettoyage des données doit donner des instructions aux enquêteurs ou aux opérateurs de saisie des données pour éviter de nouvelles erreurs, surtout si elles sont identifiées comme non-aléatoires. Le retour d'information permettra de s'assurer que les erreurs courantes ne se répètent pas et d'améliorer la validité de l'évaluation et la précision des résultats. Les principales recommandations ou corrections incluent:

- La révision de la programmation de la saisie des données, des transformations de données et des extractions de données;
- Les corrections de questions dans le formulaire du questionnaire;
- La modification du protocole d'évaluation, de la conception, du calendrier, de la formation des enquêteurs, de la collecte des données et des procédures de contrôle de la qualité;
- Dans des cas extrêmes, il peut être nécessaire de refaire une évaluation sur le terrain (quelques sites) ou de contacter à nouveau les répondants ou les enquêteurs pour demander des informations supplémentaires ou plus de détails, ou pour confirmer certains enregistrements.

- Le nettoyage des données permet souvent de mieux comprendre la nature et la gravité des processus générateurs d'erreurs.
- Identifier les causes fondamentales des erreurs détectées et utiliser ces informations pour améliorer la collecte des données et le processus de saisie des données afin d'éviter que ces erreurs ne se reproduisent.
- Reconsidérer les attentes antérieures et/ou revoir ou mettre à jour les procédures de contrôle de la qualité.

J. RECODAGE DES VARIABLES

Il peut être nécessaire de recoder les variables pour en créer de nouvelles qui répondent aux besoins de l'analyse. Par exemple, le recodage suivant est courant:

- Formatage: date (jour, mois et année), préfixes pour créer un meilleur tri dans les tableaux ;
- Arrondi des variables continues;
- Syntaxe: traduction, style de langue et simplification;
- Recodage d'une variable catégorielle (par exemple, ethnicité, profession, une catégorie "autre", corrections orthographiques, etc.);
- Recodage d'une variable continue (par exemple, l'âge) en une variable catégorielle (par exemple, le groupe d'âge);
- Regroupement des valeurs d'une variable en un nombre réduit de catégories (par exemple, regroupement de tous les problèmes causés par les contraintes d'accès);
- Combinaison de plusieurs variables pour créer une nouvelle variable (par exemple, le score de consommation alimentaire, la construction d'un indice basé sur un ensemble de variables);
- Définition d'un état en fonction de certains seuils (par exemple, population "à risque" ou "à risque aigu");
- Changement du niveau de mesure (par exemple, passage d'une échelle d'intervalles à une échelle ordinale).

Sur le plan conceptuel, il convient d'établir une distinction entre:

- Activités liées au recodage des données qualitatives – par exemple, les réponses aux questions ouvertes.
- Activités qui comprennent la transformation et la dérivation de nouvelles valeurs à partir d'autres, telles que la création de calculs (par exemple, le pourcentage), l'analyse, la fusion, etc. Ici, l'analyste réexprime ce que disent les données (c'est-à-dire qu'il réexprime l'écart en tant que variation en %, moyenne pondérée ou mobile, etc.) Les données sont (normalement) déjà passées par une étape de nettoyage avant d'être transformées.

Pour les deux types, le recodage des variables ou des valeurs peut servir à la fois à nettoyer les données brutes et/ou à transformer les données nettoyées. Cette section se concentre principalement sur le premier cas, plutôt que sur la réexpression des valeurs, qui sera abordée de manière plus approfondie dans un autre chapitre du manuel sur la transformation des données.

Le recodage des variables catégorielles commence par une liste complète de toutes les variantes générées par une variable, ainsi que leurs fréquences. La liste des variantes peut être copiée dans une nouvelle feuille, afin de créer un tableau des variantes et de leurs remplacements souhaités. Conservez TOUJOURS une copie des valeurs originales et essayez différents schémas de recodage avant d'en choisir un définitif.

Il existe trois façons de recoder des données catégorielles: 1. Réduire une variable catégorielle à un nombre réduit de catégories.

Quelle est la profession actuelle du chef de famille?

CATÉGORIES ORIGINALES	CATÉGORIES RECODÉES
Employé du gouvernement	Agriculture
Agriculture et élevage	Non-agriculture
Commerce	
Étudiant	
Etc.	

Les directives pour le regroupement des données sont les suivantes:

- Les variables ordinales doivent être réduites d'une manière qui préserve l'ordre des catégories.
- Ne combinez que les catégories qui vont ensemble. Ne combinez pas deux catégories logiquement distinctes dans le seul but d'éliminer les catégories comportant de petits nombres (par exemple, le manque d'accès dû au manque de revenus et le manque d'accès dû à l'insécurité), car l'interprétation des données devient difficile ou dénuée de sens.
- La manière dont les catégories sont regroupées peut facilement affecter le niveau de significativité des tests statistiques. Les catégories doivent être regroupées pour éviter la critique selon laquelle les données ont été manipulées uniquement pour obtenir un certain résultat. Cela ne signifie pas qu'il faille prendre une décision avant de collecter les données (si c'était le cas, il ne serait pas nécessaire de collecter des catégories distinctes).
- Ne simplifiez pas trop les données. Une réduction inutile du nombre de catégories peut réduire la puissance statistique et masquer les relations dans les données. En règle générale, gardez intactes toutes les catégories qui comprennent 10% ou plus de vos données (ou 5 cas, pour les très petits échantillons).

Décomposition: Plusieurs raisons justifient la décomposition d'une variable catégorielle en plusieurs variables plus petites :

- Les données ont été recueillies d'une manière simple afin d'alléger le travail de collecte des données pour la personne concernée. Par exemple, il est plus facile pour le répondant de fournir une liste de problèmes que d'examiner une longue liste de problèmes.
- Une variable peut contenir plus d'un "concept". Par exemple, considérez la variable ordinale "gravité" ci-dessous:

1. Il n'y a pas de pénurie
2. Quelques personnes sont confrontées à des pénuries
3. De nombreuses personnes sont confrontées à des pénuries
4. Les pénuries touchent tout le monde

Cette variable contient deux concepts, "pénuries" et "nombre de personnes touchées". Il est simple de coder deux nouvelles variables, les pénuries (0 = pas de pénurie, 1 = pénurie) et le nombre de personnes (0 = aucune personne, I= peu de personnes, 2= beaucoup de personnes, 4= toutes les personnes).

La combinaison est le processus inverse de la décomposition, par exemple en combinant "pénuries" et "nombre de personnes" pour obtenir la variable "gravité".

Le recodage des variables peut être fastidieux. L'effort conceptuel nécessaire pour produire un ensemble de catégories recodées de manière significative est souvent sous-estimé. Il faut prendre soin d'évaluer les ensembles de catégories combinés, d'absorber les catégories excessives, incohérentes ou rarement utilisées dans des catégories plus larges, et d'être clair sur la justification du nombre final et du contenu des catégories distinctes. Il faut également être conscient que tout recodage qui réduit le nombre de catégories entraîne une certaine perte d'informations. Comme à toutes les étapes de l'analyse des données, les analystes doivent être attentifs aux erreurs.

Voici quelques conseils de base pour un recodage efficace:

- **Utilisez des noms de variables distincts et faciles à retenir.** N'utilisez jamais le même nom de variable pour désigner à la fois la variable transformée et la variable non transformée. Pour les grands ensembles de données, il est souhaitable d'utiliser une méthode systématique et planifiée pour nommer les variables.
- **Faites attention aux valeurs manquantes.** Lorsque le recodage est effectué, le nombre de cas avec des données manquantes doit être le même qu'avant le recodage. Vérifier que c'est le cas est souvent le premier indice d'une erreur de recodage. Une procédure sûre consiste à commencer le processus de recodage en définissant la nouvelle variable comme manquante pour tous les cas, puis en modifiant les valeurs manquantes uniquement pour ceux qui ont des données sur les variables initiales à recoder. Pour les recodages compliqués, vérifiez quelques valeurs individuelles à la main pour vous assurer qu'elles ont été recodées correctement, et vérifiez la distribution des valeurs.
- **Utilisez des graphiques pour vérifier l'exactitude du recodage.** Le recodage est une traduction systématique des valeurs de données, de sorte que les diagrammes de dispersion des données brutes par rapport aux données recodées devraient montrer des modèles très organisés reflétant le système de recodage. Les histogrammes peuvent montrer si vos données sont maintenant distribuées plus normalement.
- **Utilisez les codes des variables de manière cohérente.** Par exemple, pour les variables dichotomiques « oui/non », utilisez toujours « 0 » = non et « 1 » = oui. Pour les variables polytomiques, faites toujours de « 0 » la catégorie de référence/l'option par défaut.
- **Conservez un enregistrement permanent de votre recodage.** Pour les erreurs de saisie de données, effectuez les modifications directement dans le fichier de données brutes, car il n'est pas nécessaire de conserver les valeurs erronées. Le recodage, en revanche, doit lui être effectué dans un fichier séparé, car il peut être nécessaire de revoir les données brutes. La plupart des logiciels statistiques enregistrent les données dans un fichier spécialement formaté : c'est ce fichier qu'il faut modifier. Les commandes utilisées pour le recodage doivent toutes être placées dans un fichier (généralement, un fichier do) qui peut être exécuté à nouveau. Le fichier do sert également d'enregistrement permanent.

K. PROCÉDURES DE CONTRÔLE DE LA QUALITÉ

Lorsqu'on définit une procédure pour le nettoyage des données, il est utile de considérer les différents types d'erreurs qui peuvent être commises et de planifier à quel moment du flux de données des mesures de contrôle doivent être mises en œuvre. Les meilleures pratiques sont les suivantes:

Rôles et responsabilités: Assurez-vous que le personnel ayant des responsabilités en matière de qualité des données connaît les protocoles de nettoyage (voir l'annexe 1 pour une liste de contrôle complète pour les évaluations des besoins). Les rôles et responsabilités liés à la détection et à la correction des erreurs doivent être clairement définis et communiqués dans le cadre des descriptions de postes (voir l'annexe 2), à chaque étape de la collecte, de la saisie et du traitement des données.

Assurez-vous qu'une deuxième paire d'yeux examine et compare les données originales aux données saisies. Le nettoyage des données doit commencer sur le terrain parallèlement à la collecte des données, car les questionnaires sont examinés quotidiennement par les superviseurs ou les éditeurs du terrain. De même, lors de la saisie des données, les doubles vérifications devraient être obligatoires, notamment lorsque:

- Il existe un processus de traduction au moment de la saisie des données, afin de garantir la cohérence/l'exactitude de la traduction.
- La saisie des données est répartie sur plusieurs sites et la consolidation a lieu dans un autre site.

Au stade de la saisie des données, des procédures de contrôle de la qualité assistées par ordinateur doivent être utilisées. Des fonctionnalités supplémentaires peuvent être ajoutées au logiciel de saisie des données (par exemple: Excel, Survey Gizmo, Access, SPSS, STATA, etc.) pour mettre en évidence les violations des règles (par exemple, via des codes nuls, un formatage conditionnel, etc.) et éviter les erreurs (par exemple grâce à des menus déroulants). La décision d'inclure ces règles dans la base de données doit être pragmatique. Il faut pour cela évaluer les avantages de la détection et de la correction des erreurs par le personnel chargé de la saisie des données, par rapport au temps nécessaire à la mise en place de ces règles et à la réalisation rapide des ajustements nécessaires si la configuration initiale ne fonctionne pas comme prévu.

Cinq types de contrôles peuvent être automatisés:

- Les contrôles d'étendue garantissent que chaque variable de l'enquête ne contient que des données comprises dans un domaine défini de valeurs valides. Les variables catégorielles ne peuvent avoir qu'une seule des valeurs prédéfinies dans le questionnaire (par exemple, le genre ne peut être codé que par "1" pour les hommes ou "2" pour les femmes) ; les variables chronologiques doivent contenir des dates valides et les variables numériques doivent se situer dans les limites des valeurs minimales et maximales définies (par exemple, l'âge doit être compris entre 0 et 120 ans et doit toujours être exprimé sous forme d'années en nombre entier, avec des règles d'arrondi pour les enfants en bas âge).

- La vérification des données de référence est utilisée lorsque les données de deux ou plusieurs champs étroitement liés peuvent être vérifiées par rapport à des tables de référence externes (par exemple, lorsque les valeurs enregistrées pour la taille, le poids et l'âge sont vérifiées par rapport aux tables de référence standard de l'Organisation Mondiale de la Santé).
- Les contrôles de saut de question vérifient si les règles de saut de question ont été suivies de manière cohérente (par exemple, un simple contrôle permet de vérifier que les questions à poser uniquement aux enfants scolarisés ne sont pas enregistrées pour un enfant qui a répondu "non" à une question initiale sur la scolarisation).
- Les contrôles de cohérence vérifient que les valeurs d'une question sont cohérentes avec les valeurs d'une autre question (par exemple, la date de naissance et l'âge d'un individu donné).
- Les contrôles typographiques limitent, par exemple, la transposition de chiffres, comme la saisie de "14" au lieu de "41" dans une entrée numérique. Une telle erreur concernant l'âge peut être détectée par des contrôles de cohérence avec l'état civil ou la situation familiale. Les totaux de contrôle, par exemple, peuvent réduire considérablement les erreurs typographiques.

- Documentez les règles de qualité des données à suivre, les points sur lesquels il faut se concentrer et la manière de résoudre les erreurs/problèmes. Prévoyez des doubles vérifications.
- Communiquez des instructions claires aux enquêteurs, aux chefs d'équipe ainsi qu'au contrôleur de la saisie des données, à toutes les étapes pertinentes du flux de données.
- Veillez à ce que le personnel chargé de la saisie des données soit familier des procédures de remplissage du questionnaire, afin que les erreurs puissent être identifiées rapidement et vérifiées/rectifiées (par exemple, des règles telles que "ne choisir que trois" ou "le total doit être de 100 %").
- Concevoir un plan de nettoyage des données, comprenant:
- Un budget, un calendrier et les besoins en personnel.
- Des outils de contrôle des données.
- Des procédures de diagnostic utilisées pour discerner les erreurs (sur une base périodique et vers la fin de l'évaluation).
- Des instructions ou formations destinées aux enquêteurs et au personnel chargé de la saisie des données concernant les cas de violation du protocole et le contrôle de cohérence.
- Des règles de décision qui seront appliquées lors de la phase d'édition.

L. INTÉGRATION DES DONNÉES

Une autre série de problèmes peut survenir lorsque les jeux de données sont intégrés ou fusionnés avec d'autres données. Les analystes n'ont pas toujours le contrôle du format et du type de données qui sont importées d'une source de données externe, telle qu'une base de données, un fichier texte ou une page Web. Les problèmes les plus courants sont:

- **Les formats:** Les dates sont particulièrement problématiques (26/02/1977, 26 février 1977, 26-02-1977, etc.). Les analystes doivent également être conscients que les différentes applications stockent les dates en interne de différentes manières. Un simple copier-coller d'une application à l'autre entraînera donc des erreurs dans la majorité des cas.
- **Les unités:** Différentes unités de mesure sont utilisées : litre, gallons, gourdes, etc.
- **Les intervalles:** Les intervalles d'âge peuvent différer d'une enquête à l'autre. L'intervalle d'âge ne peut être recréé que si la date de naissance est disponible.
- **L'incohérence:** Lors de la fusion de différentes sources de données, des informations contradictoires peuvent apparaître. Les analystes doivent choisir entre utiliser les deux, utiliser les informations les plus récentes, utiliser les informations provenant de la source la plus fiable, enquêter davantage ou n'utiliser aucune des deux. Les enregistrements en double ne doivent généralement pas être supprimés, mais signalés afin d'être identifiés et exclus de l'analyse dans les cas où les enregistrements en double pourraient biaiser l'analyse. Bien qu'ils semblent être des doublons, dans de nombreux cas, les enregistrements dans les deux bases de données peuvent inclure des informations qui sont uniques à chacune d'entre eux. La suppression de l'un des doublons ("fusionner et purger") n'est donc pas toujours une bonne solution, car elle peut entraîner une perte de données précieuses.
- **L'orthographe:** Les variables catégorielles, et notamment les noms de lieux, peuvent avoir des orthographes différentes.
- **La perte d'éléments de données:** Certains éléments de données, colonnes ou lignes, sont perdus lors de l'extraction, par exemple lors de la récupération de données sur le Web ou de l'extraction d'un fichier PDF (bonne chance!).

- Les données sont sales. Il faut s'en accommoder. Les analystes qui supposent que les données brutes sont propres et qui contournent les contrôles de base vivent dangereusement.
- Vérifiez que la documentation de l'ensemble de données est disponible. Si elle n'est pas disponible (même après demande), NE FAITES PAS CONFIANCE AUX DONNÉES, même si la source est généralement fiable. Commencez à vérifier la qualité.
- Même sous la contrainte temporelle, prenez le temps de passer les données au crible pendant 15 à 30 minutes, en vous concentrant d'abord sur l'orthographe et le formatage des ensembles de données à fusionner, puis examinez les données aberrantes/extrêmes (utilisez des filtres pour une détection visuelle rapide). Si aucune erreur n'est repérée dans cet intervalle de temps, les données sont probablement de bonne qualité et utilisables telles quelles. Si des erreurs sont détectées, il faut alors procéder rigoureusement et méthodiquement au contrôle, au diagnostic et au traitement des données.

M. PRINCIPES CLÉS POUR LE NETTOYAGE DES DONNÉES

Les principes clés du nettoyage des données dans les feuilles de calcul sont les suivants:

1. Créez une copie de sauvegarde des données brutes (originales) dans un classeur séparé.
2. Sauvegardez régulièrement le fichier de travail, aussi bien au moment de la collecte, que du nettoyage ou de l'analyse. Enregistrez les documents avec des noms de fichiers combinant la date et l'heure (les préfixes aammjj-heure permettent de trier les fichiers par ordre de création).
3. Lors de l'intégration ou de la fusion de différents jeux de données, assurez-vous que les données se présentent sous la forme d'un tableau de lignes et de colonnes avec : des données similaires dans chaque colonne, toutes les colonnes et lignes visibles, et aucune ligne vide dans la plage. Vérifiez qu'il n'y a pas de sous-totaux, de totaux ou d'autres enregistrements calculés en bas des colonnes. Les variables calculées peuvent rester à droite des données.
4. Formatez la base de données pour en faciliter la lecture et la navigation : texte aligné à gauche, numéro aligné à droite, titre des variables positionné horizontalement, texte des variables entièrement visible, colonne séparée par des lignes en gras, en-tête avec couleurs de fond, numéros séparés par des virgules tous les 3 chiffres, etc.
5. Commencez par des tâches qui ne nécessitent pas de colonne, comme la vérification de l'orthographe ou l'utilisation de la fonction Rechercher et Remplacer.
6. Ensuite, entreprenez les tâches qui nécessitent une manipulation de colonne. Les étapes générales de la manipulation d'une colonne sont les suivantes:

- Insérez une nouvelle colonne (B) à côté de la colonne originale (A) qui doit être nettoyée.
- Transformez les données de la colonne (B).
- Supprimez la colonne originale (A), ce qui convertit la nouvelle colonne de B en A.

7. Gardez le questionnaire à portée de main. Au fur et à mesure des vérifications, une liste de problèmes sera dressée. Les questionnaires doivent être consultés pour vérifier ou identifier les problèmes.
8. Lorsque vous vérifiez un type de problème pour un lieu donnée ou un informateur clé, vérifiez que les données des autres variables pour ce cas ont été saisies correctement.
9. Regardez les valeurs de toutes les variables et de tous les cas pour ce lieu, cet informateur clé ou cet enquêteur. Il arrive que la personne chargée de la saisie des données saute une variable ou saisisse les valeurs de la variable précédente ou de la variable suivante, et toutes les données qui ont été saisies ensuite ne seront pas correctes.

- Il est essentiel de planifier et de budgétiser le nettoyage des données.
- L'organisation des données permet d'améliorer l'efficacité, par exemple en triant les données par lieu ou les enregistrements par enquêteur.
- Il vaut mieux prévenir que guérir. Il est bien plus efficace de prévenir une erreur que de devoir la trouver et la corriger plus tard.
- La responsabilité de produire des données propres incombe à tous, enquêteurs, détenteurs et utilisateurs.
- La hiérarchisation des priorités réduit les doublons. Concentrez-vous sur les enregistrements pour lesquels des données étendues peuvent être nettoyées à moindre coût ou qui ont le plus de valeur pour les utilisateurs finaux.
- La remontée d'information (feedback) est une voie à double sens : les utilisateurs ou les analystes de données procéderont inévitablement à la détection d'erreurs et devront fournir un retour d'information aux détenteurs/responsables des données. Développez des mécanismes de remontée d'information et encouragez les utilisateurs à signaler les erreurs.
- L'éducation et la formation améliorent les techniques : la mauvaise formation des enquêteurs et des opérateurs de saisie est à l'origine d'une grande partie des erreurs. Formez aux exigences de qualité (lisibilité, etc.) et à la documentation.
- Les processus de nettoyage des données doivent être transparents et bien documentés, avec un bon registre (le journal des modifications) pour limiter les doublons et garantir qu'une fois corrigées, les erreurs ne se reproduisent plus.
- La documentation est la clé d'une bonne qualité des données. Sans une bonne documentation, il est difficile pour les utilisateurs de déterminer si l'utilisation des données est appropriée et il est difficile pour les détenteurs/responsables de savoir quels contrôles de qualité des données ont été effectués et par qui.

N. OUTILS ET TUTORIELS POUR LE NETTOYAGE DES DONNÉES

Les tableurs tels qu'Excel offrent la possibilité de trier facilement les données, de calculer de nouvelles colonnes, de déplacer et de supprimer des colonnes, et d'agréger des données. Pour le nettoyage des données d'évaluation humanitaire, ACAPS a développé une note technique spécifique fournissant une approche étape par étape dans Excel et détaillant les opérations de nettoyage, accompagnée par un classeur de démonstration.

Pour des instructions générales sur la manière d'utiliser les formules, les fonctionnalités ou les options d'Excel pour nettoyer les données, plusieurs notes d'orientation de Microsoft Office sont disponibles:

- Vérification orthographique
- Suppression des lignes en double
- Recherche et remplacement de texte
- Changer la casse d'un texte
- Suppression des espaces et des caractères non imprimables du texte
- Correction des chiffres et des signes numériques
- Correction des dates et des heures
- Fusion et fractionnement des colonnes
- Transformation et réorganisation des colonnes et des lignes
- Rapprocher les données de tables en les joignant ou en les faisant correspondre.
- Fournisseurs tiers

Openrefine (ex-Google Refine) et LODRefine sont des outils puissants pour travailler avec des données désordonnées, pour les nettoyer ou les transformer d'un format à un autre. Des vidéos et des tutoriels sont disponibles pour apprendre les différentes fonctionnalités offertes par ces logiciels. La fonction ''facets'' est particulièrement utile car elle permet de donner très efficacement et rapidement une idée de l'ampleur des variations dans un ensemble de données.

Des tutoriels et des cours détaillés sur le nettoyage des données sont également disponibles sur School of Data:

- http://schoolofdata.org/handbook/recipes/cleaning-data-with-spreadsheets/
- http://schoolofdata.org/handbook/courses/data-cleaning/

Deux outils spécialisés pour réaliser plusieurs de ces tâches sont utilisés par ACAPS. Le premier est Trifacta Wrangler, la nouvelle version de Data Wrangler par le Stanford Visualization Group. Trifacta Wrangler est un outil facile d'utilisation qui peut automatiquement trouver des modèles dans les données en fonction des éléments sélectionnés, et faire automatiquement des suggestions sur ce qu'il faut faire avec ces modèles. Beau et utile. L'autre logiciel phare de nettoyage est Data monarch de Datawatch, qui intègre de nombreuses fonctionnalités de ''wrangling'' (de préparation), de nettoyage et d'enrichissement qui peuvent prendre des heures sur Excel.

0. SOURCES ET LECTURES DE RÉFÉRENCE

- ACAPS. 2013. How to Approach a Dataset – Preparation. Disponible sur: http://www.acaps.org/resourcescats/downloader/how_to_approach_a_dataset_part_1_data_preparation/163/1375434553 Et son manuel annexe, disponible sur: https://www.acaps.org/resourcescats/downloader/how_to_approach_a_dataset_data_management/164
- ACAPS. 2012. Severity Rating, A Data Management Note. http://www.acaps.org/resourcescats/downloader/severity_rating_data_management_note/87/1376302232
- Benini, A. 2011. Efficient Survey Data Entry – A Template for Development NGOs. Friends in Village Development Bangladesh (FIVDB). http://aldo-benini.org/Level2/HumanitData/FIVDB_Benini_Effi cientDataEntry_110314.pdf
- Buchner, D. M. Recherche en médecine physiologique et en réhabilitation. http://c.ymcdn.com/sites/www.physiatry.org/res ource/resmgr/pdfs/pmr-viii.pdf
- Chapman, A. D. 2005. Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data. http://www.gbif.org/orc/?doc_id=1262
- Den Broeck, J. V., Cunningham, S. A., Eeckels, R., Herbst, K. 2005. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. Afrique du Sud, Centre Africain de Recherche sur la Population et la Santé (APHRC)
- Dr. Limaye, N. 2005. Clinical Data Management – Data Cleaning.
- Henning, J. 2009. Data Cleaning. http://blog.vovici.com/blog/bid/19211/Data-Cleaning
- Joint IDP profiling Service (JIPS). Consulté en juillet 2013. Manual Data Entry Staff. http://jet.jips.org/pages/view/toolmap
- Kassoff, M. 2003. Data Cleaning. http://logic.stanford.edu/classes/cs246/lectures/lecture13.pdf
- Kim et Al. 2003. A Taxonomy of Dirty Data. http://sci2s.ugr.es/docencia/m1/KimTaxonomy03.pdf
- Université du Michigan. 2012. Data Cleaning Guidelines (SPSS and STATA). Première édition. https://www.canr.msu.edu/fsg/survey/Data_Cleaning_Guidelines_SPSS_Stata_1stVer.pdf
- Munoz, J. 2005. A Guide for Data Management of Household Surveys. Santiago, Chile, Household Sample Surveys in Developing and Transition Countries. http://unstats.un.org/unsd/hhsurveys/
- Osborne, J. W. 2013. Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data. California, SAGE.
- Psychwiki. Consulté le 7 septembre 2009. Identifying Missing Data. http://www.psychwiki.com/wiki/Identifying_Missing_Data
- Psychwiki. Consulté le 11 septembre 2009. Dealing with Missing Data. http://www.psychwiki.com/wiki/Dealing_with_Missing_Data
- Psychwiki. Consulté le 7 septembre 2009. Missing Values. http://www.psychwiki.com/wiki/Missing_Values
- Sana, M., Weinreb, A. A. 2008. Insiders, Outsiders, and the Editing of Inconsistent Survey Data. Sociological Methods & Research, Volume 36, Number 4, SAGE Publications. http://www.academia.edu/1256179/Insiders_Outsiders_and_the_Editing_of_Inconsistent_Survey_Data
- The Analysis Institute. 2013. Effectively Dealing with Missing Data without Biasing Your Results. http://theanalysisinstitute.com/missing-data-workshop/
- Wikipedia. Consulté le 31 juillet 2013. Nettoyage de données. https://fr.wikipedia.org/wiki/Nettoyage_de_donn%C3%A9es

ANNEXE 1 - LISTE DE CONTRÔLE POUR LE NETTOYAGE DES DONNÉES

Préparer le nettoyage des données

La planification est essentielle. Assurez-vous que les outils, les contacts pour le nettoyage des données et le matériel suivants soient disponibles:

- Les formulaires d'enquête (questionnaires)
- Les contacts des chefs d'équipe ou des enquêteurs, au cas où ils devraient être contactés pour des questions
- La base de données initiale
- Un traducteur, si nécessaire
- Un logiciel d'analyse visuelle (par exemple : tableau public)
- Un logiciel tableur (Excel) ou de base de données (Access, Stata, etc.)
- Certains ajouteraient le café et la musique, ainsi qu'un endroit sans bruit ni nuisance

Identifiez le/la gestionnaire des données. Il/elle sera généralement responsable de la gestion et du stockage des données, ainsi que de la supervision du nettoyage des données, de la centralisation des changements, de la mise à jour et de la maintenance du journal des modifications.

Établir, documenter et communiquer

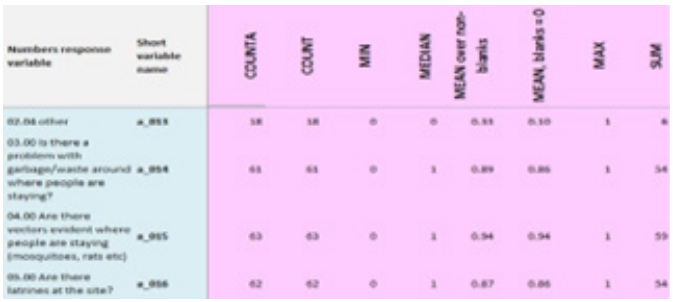
- Formez les opérateurs de saisie de données sur la façon dont le questionnaire est rempli. Expliquez les instructions données aux enquêteurs. Si possible, faites participer le personnel chargé de la saisie des données à la formation des enquêteurs afin de faciliter la communication interne.
- Établissez des règles de décision pour savoir quand modifier une valeur et quand NE PAS le faire.
- Établissez des procédures pour informer sur les données qui ont été modifiées ou qui n'ont pas été collectées, c'est-à-dire "manquantes" ou "non-collectées".
- Expliquez comment utiliser le journal des modifications.
- Communiquez aux opérateurs de saisie de données ou aux analystes les procédures à suivre et les personnes à informer en cas de détection d'erreurs.
- Établissez des canaux de communication pour communiquer les erreurs détectées. La communication écrite est recommandée.
- Pour les évaluations rapides où l'analyse, la cartographie et la visualisation des données coïncident généralement avec la saisie et le nettoyage des données, communiquez régulièrement aux analystes, aux responsables SIG et aux graphistes les parties des ensembles de données qui sont propres et utilisables. Établissez des procédures de rapport claires au cas où des erreurs supplémentaires seraient identifiées. Planifiez avec l'équipe les variables à nettoyer en priorité.

Examiner les dossiers

- Si une stratégie d'échantillonnage a été utilisée, les enregistrements doivent d'abord être vérifiés. Vérifiez si tous les sites ont été saisis, y compris ceux où l'évaluation n'a pas été achevée (cela n'est pas pertinent en cas d'échantillonnage intentionnel). Comparez les enregistrements avec les rapports des équipes d'évaluation sur le terrain, ou avec le fichier de suivi des sites visités.
- Attribuez un identifiant pour chaque site ou ménage et assurez-vous qu'il soit unique.
- Vérifiez régulièrement l'absence de doublons pour chaque enregistrement (lignes dans la base de données). Supprimez tous les entrées vides où il n'y a aucune donnée dans aucune des variables. Assurez-vous d'abord que ces entrées vides doivent bien être supprimés et vérifiez comment cela pourrait affecter les autres données de la ligne.

Contrôler, diagnostiquer et traiter les données

- Tout d'abord, il faut nettoyer les questions filtres, c'est-à-dire celles où l'on demande à la population si elle a fait ou eu une activité particulière en fonction d'une réponse (oui/non) à une question précédente. Dans ce cas, si la réponse est ''oui'', il devrait y avoir des données dans le tableau suivant du questionnaire (ou dans une colonne de la base de données); au contraire, si la réponse est ''non'', il ne devrait y avoir aucune donnée dans ce même tableau (ou dans cette même colonne).
- Vérifiez les règles de saut de question dans le formulaire et exécutez les contrôles dans la base de données pour rechercher les valeurs non valides ou manquantes dans les variables basées sur les règles de saut.
- Nettoyez les questions avec des valeurs de réponse minimales ou maximales (''ne cochez que trois options'', ''quelles sont les trois premières priorités parmi les cinq choix suivants'', etc.)
- Inspectez les variables restantes de manière séquentielle et au fur et à mesure qu'elles sont enregistrées dans le fichier de données. Créez un tableau récapitulatif général des statistiques descriptives, où pour chaque variable sont disponibles le min, le max, la moyenne, la médiane, la somme et le nombre.



Numbers response variable	Short variable name	COUNTA	COUNT	MIN	MEDIAN	MEAN over non-blanks	MEAN, blanks = 0	MAX	SUM
02.04 other	a_053	10	10	0	0	0.33	0.10	1	4
03.00 Is there a problem with garbage/waste around where people are staying?	a_054	61	61	0	1	0.89	0.80	1	54
04.00 Are there vectors evident where people are staying (mosquitoes, rats etc)	a_055	63	63	0	1	0.94	0.94	1	59
05.00 Are there latrines at the site?	a_056	62	62	0	1	0.87	0.66	1	54

Capture d'écran du tableau des statistiques récapitulatives d'Aldo Benini, ACAPS Note technique ''How to approach a dataset, preparation''

- Si la variable est une variable catégorielle/qualitative, vérifiez si l'orthographe est cohérente et effectuez un comptage de fréquence:
- Regardez les effectifs pour chaque catégorie, afin de voir s'ils sont cohérents pour l'échantillon - l'ensemble des données est-il complet?
- Toutes les valeurs doivent avoir des étiquettes si la variable est catégorielle. Vérifiez l'étendue des valeurs.
- Si la variable est une variable continue/d'intervalle, calculez les statistiques descriptives telles que min, max, mode, moyenne et médiane.
- Observez les valeurs minimales et maximales. Sont-elles cohérentes ? Vérifiez surtout si les "0" sont vraiment des "0" et non des valeurs manquantes.
- La moyenne et la médiane sont-elles conformes aux attentes?
- Inspectez les données pour détecter les valeurs manquantes (blancs, codes explicites de valeurs manquantes). Décidez:
- Quelles cellules vides doivent être remplies de zéros - parce qu'elles représentent de véritables observations négatives, telles que : ''non'', ''non-présent'', ''option non-précisée'', etc.).
- Lesquelles laisser en blanc (si la convention est d'utiliser des blancs pour les données manquantes ou non applicables).
- À remplacer par un code de valeur manquante explicite (si nous voulons que toutes les valeurs manquantes soient explicitement codées).
- Vérifiez que dans les variables binaires (oui/non), la valeur positive soit codée "1", et la négative "0".
- Vérifiez la distribution des valeurs (utilisez des diagrammes en boîte si possible). Observez les valeurs extrêmes et comparez-les au questionnaire, même si la valeur est possible et semble raisonnable. S'il s'agit d'une valeur extrême, d'autres variables peuvent également être incorrectes. Recherchez les 5 valeurs les plus petites/grandes.
- Comparez les données entre deux variables ou plus au sein d'un même cas pour vérifier les problèmes de logique. Par exemple, le chef de famille peut-il être âgé de moins de 17 ans? Comparez l'âge avec l'état civil. La personne est-elle trop jeune pour avoir été mariée ? La somme des proportions est-elle égale à 100 %?

- Lorsque des questions portent sur une "unité", les données doivent être normalisées à une unité spécifique, c'est-à-dire lorsqu'une réponse est collectée en utilisant l'unité spécifiée par le répondant. Par exemple, les unités de superficie peuvent être l'acre, l'hectare et le mètre carré. Pour normaliser l'unité de surface, une table de consultation peut être utilisée pour fusionner la valeur de conversion afin de convertir toutes les surfaces en hectares.
- Vérifiez les cohérences au sein d'une série de cas : s'il y a un conjoint, on s'attend à ce qu'il soit d'un sexe différent. L'enfant du chef de ménage ne doit pas être plus âgé que le chef. Le parent du chef de ménage ne peut pas être plus jeune que le chef de ménage.
- Recodez les variables. Remplacer les entrées inutiles (par exemple, les fautes d'orthographe, les descriptions verbales, la catégorie "autres", etc.) par des variantes plus appropriées, de manière cohérente. Les raisons du recodage sont les suivantes: corrections orthographiques, formatage de la date (jour, mois, année), traduction, style et simplification de la langue, regroupement, préfixes pour créer un meilleur tri dans les tableaux, combinaison (pour les variables catégorielles), arrondis (pour les variables continues), et éventuellement d'autres.
- Triez le fichier de différentes manières (par variables individuelles ou groupes de variables) pour voir si des erreurs de données qui n'ont pas été trouvées précédemment peuvent être identifiées.

Considérations finales

- Si les données sont nettoyées par plusieurs personnes, l'étape finale consiste à fusionner toutes les feuilles de calcul pour qu'il n'y ait qu'une seule base de données. Les commentaires ou les journaux de modifications qui sont faits au fur et à mesure du nettoyage doivent être compilés dans un seul document. Les données problématiques doivent être renseignées dans le dossier de documentation.
- Mettez à jour les procédures de nettoyage, le journal des modifications et le fichier de documentation des données (dictionnaire des données) au fur et à mesure de l'avancement du nettoyage. Fournissez des informations en retour aux enquêteurs, aux chefs d'équipe ou aux opérateurs de saisie des données si le processus de collecte et de saisie des données est toujours en cours. Si les mêmes erreurs sont commises par une équipe ou des enquêteurs, veillez à en informer le responsable.
- Soyez préparé. Le nettoyage des données est un processus continu. Certains problèmes ne peuvent être identifiés avant le début de l'analyse. Les erreurs sont découvertes lorsque les données sont manipulées par les analystes, et plusieurs étapes de nettoyage sont généralement nécessaires lorsque des incohérences sont découvertes. Dans les évaluations rapides, il est très fréquent que des erreurs soient même détectées pendant le processus d'examen par les pairs.

ANNEXE 2 – MODÈLES DE DESCRIPTIONS DE POSTE

Cette annexe présente trois descriptions de postes liées à la saisie et au nettoyage des données : Responsable du nettoyage des données, Opérateur de saisie de données et Contrôleur de saisie de données. Les formats word sont disponibles sur http://www.acaps.org/resourcescats/downloader/assessment_team_job_descriptions/97

Titre du poste: Responsable du nettoyage des données

Responsable hiérarchique

Analyste de données

Prérequis

- Expérience en matière d'évaluation et d'enquête.
- Expérience requise dans la saisie de données à grande échelle.

Formation

- Diplôme en statistiques ou en démographie et/ou diplôme en informatique.

Expérience

- 2-3 ans d'expérience dans un institut de statistiques et/ou une expérience professionnelle similaire.
- Expérience avérée en nettoyage de données et en gestion de grands volumes de données quantitatives et qualitatives.
- Expérience avérée en gestion et exploitation de bases de données.

Langue

- Maîtrise de l'anglais écrit et parlé (ou utilisation de la langue de l'enquête).

Compétences

- Professionnalisme;
- Excellentes compétences en matière de communication écrite et orale;
- Bonne connaissance des logiciels de traitement de texte et de données (Word, Excel, PowerPoint, messagerie électronique);
- Compréhension des principes de l'analyse statistique et démographique;
- Compréhension des techniques d'enquêtes;
- Excellentes compétences en matière de rédaction de rapports;
- Solides compétences en dactylographie;
- Solides compétences en matière de relecture;
- Excellente maîtrise des outils informatiques ; Haut niveau de connaissances en informatique;
- Rigueur et précision;
- Capacité avérée à respecter les délais ; Capacité à bien travailler sous pression;
- Bonnes compétences interpersonnelles et capacité à travailler dans un environnement multiculturel. Forte capacité à travailler en équipe;
- Une expérience de travail avec la communauté humanitaire internationale est un avantage.

Description du rôle et des responsabilités:

Sous la supervision de l'Analyste de données, le Responsable du nettoyage des données est chargé de:

- S'assurer que les procédures de vérification, de codage et de saisie des données sont respectées;
- Vérifier la qualité du travail effectué par le personnel chargé de la saisie des données lors du contrôle, du codage et de la saisie des données et fournir toute l'aide et le retour d'information nécessaires pour améliorer la saisie des données et réduire ou prévenir les erreurs;
- Garder une vue d'ensemble documentée du travail quotidien; produire un rapport quotidien sur le nettoyage des données;
- S'assurer que les données brutes ont été saisies avec précision dans un fichier lisible par ordinateur;
- Vérifier que les variables de caractères ne contiennent que des valeurs valides;
- Vérifier que les valeurs numériques sont comprises dans les fourchettes acceptables prédéterminées;
- Rechercher et éliminer les doublons (enregistrements de données en double);
- Vérifier s'il y a des valeurs manquantes pour les variables pour lesquelles des données complètes sont nécessaires;
- Vérifier que certaines valeurs, telles que les ID des personnes interrogées, sont bien uniques;
- Vérifier les valeurs de données invalides et les séquences de dates invalides;
- Suivre les procédures de nettoyage et d'édition des données. Documenter les problèmes de données. Mettre régulièrement à jour la base de données principale avec les dernières modifications.

Coordinateur d'évaluation :

Nom : _____

Poste : _____

Signature : _____

Date : _____

Responsable du nettoyage de données :

Nom : _____

Poste : _____

Signature : _____

Date : _____

Titre du poste: Opérateur de saisie de données

Responsables hiérarchiques

Contrôleur de la saisie des données

Analyste de données

Prérequis

- Expérience en matière d'évaluation et d'enquête.
- Expérience requise en saisie de données à grande échelle.

Formation

- Enseignement secondaire, un diplôme en gestion de l'information/gestion des données est un atout.

Expérience

- 1 à 2 ans d'expérience dans un institut de statistiques et/ou une expérience professionnelle similaire.
- Expérience avérée en saisie et gestion de grands volumes de données quantitatives et qualitatives.
- Expérience avérée en gestion et exploitation de bases de données.

Langue

Maîtrise de l'anglais écrit et parlé (ou utilisation de la langue de l'enquête).

Compétences

- Solides compétences en dactylographie;
- Compétences en saisie de données;
- Solides compétences en matière de relecture;
- Compétences analytiques;
- Excellente maîtrise des outils informatiques;
- Niveau élevé de connaissances en informatique;
- Rigueur et précision;
- Capacité avérée à respecter les délais;
- Bonnes compétences interpersonnelles et capacité à travailler dans un environnement multiculturel;
- Une expérience de travail avec la communauté humanitaire internationale est un avantage.

Description du rôle et responsabilités:

Sous la supervision du Contrôleur de la saisie des données ou de l'Analyste de données, l'Opérateur de la saisie des données est chargé de:

- Vérifier les questionnaires remplis avant la saisie des données;
- Coder des questions ouvertes et semi-fermées;
- Identifier les questionnaires comportant des erreurs et/ou des fautes, lorsque l'identifiant est incorrectement rempli et veiller à ce qu'ils soient corrigés; détecter les formulaires qui doivent être réévalués et organiser la réévaluation;
- Effectuer la saisie des données des questionnaires selon les procédures définies lors de la formation;
- Aider le personnel du site à enregistrer et à gérer avec précision les données collectées;
- Contrôler la qualité des données et éditer les données selon les procédures spécifiées
- Remonter des informations aux responsables des équipes d'Analystes des données et d'évaluation sur les erreurs récurrentes à éviter;
- Tenir un journal des modifications en cas de nettoyage ou d'édition des données;
- Archiver et sauvegarder les données, en utilisant le chemin d'accès spécifié;
- Maintenir et exploiter la base de données;
- Entretenir le matériel informatique du bureau.

Coordinateur d'évaluation :

Nom : _____

Poste : _____

Signature : _____

Date : _____

Opérateur de saisie de données :

Nom : _____

Poste : _____

Signature : _____

Date : _____

Titre du poste: Contrôleur de la saisie des données

Responsable hiérarchique
Analyste de données

Exigences

- Expérience en matière d'évaluation et d'enquête.
- Expérience requise en saisie de données à grande échelle.

Formation

- Diplôme en statistiques ou en démographie et/ou diplôme en informatique.

Expérience

- 3 à 5 ans d'expérience dans un institut de statistiques et/ou une expérience professionnelle pertinente.
- Expérience avérée en saisie et gestion de grands volumes de données quantitatives et qualitatives.
- Expérience avérée en gestion et exploitation de bases de données.

Langue

Maîtrise de l'anglais écrit et parlé (ou utilisation de langues internationales).

Compétences

- Professionnalisme;
- Forte capacité à travailler en équipe;
- Excellentes compétences en matière de communication écrite et orale;
- Capacité à bien travailler sous pression;
- Bonne connaissance des logiciels de traitement de texte et de données (Word, Excel, PowerPoint, messagerie électronique);
- Bonne maîtrise des logiciels de traitement et d'analyse des données : CSPro et SPSS;
- Compréhension des principes de l'analyse statistique et démographique;
- Compréhension des techniques d'enquête;
- Excellentes compétences en matière de rédaction de rapports.
- Solides compétences en dactylographie.
- Solides compétences en matière de relecture;
- Excellente maîtrise des outils informatiques; haut niveau de connaissances en informatique;
- Rigueur et précision;
- Capacité avérée à respecter les délais;
- Bonnes compétences interpersonnelles et capacité à travailler dans un environnement multiculturel;
- Une expérience de travail avec la communauté humanitaire internationale est un avantage.

Description du rôle et responsabilités:

Sous la supervision de l'Analyste de données, le Contrôleur de la saisie des données est chargé de:

- S'assurer que les procédures de vérification, de codage et de saisie des données sont respectées;
- Assurer le suivi du personnel chargé de la saisie des données;
- Vérifier la qualité du travail effectué par le personnel chargé de la saisie des données lors de la vérification, du codage et de la saisie des données, et fournir toute l'aide nécessaire;
- Conserver une vue d'ensemble documentée du travail quotidien; produire un rapport quotidien sur la vérification, le codage et la saisie des données;
- Rédiger des procédures pour le nettoyage et l'édition des données; Superviser le nettoyage des données; Consolider les journaux de modification des données provenant des Opérateurs de saisie. Documenter les problèmes de données; Mettre régulièrement à jour la base de données principale avec les derniers changements;
- Demander les questionnaires et les renvoyer aux archives après la saisie des données;
- Veiller à ce que les documents techniques soient conservés en bon état;
- Veiller au respect des heures de travail, ainsi qu'à l'ordre et à la discipline sur le lieu de travail.

Coordinateur d'évaluation :
Nom : _____
Poste : _____
Signature : _____
Date : _____

Contrôleur de la saisie des données :
Nom : _____
Poste : _____
Signature : _____
Date : _____