Rapport d'utilisation
Outils de nettoyage de données Excel en VBA

AJAX-RICAUD Lenny, GALMICHE Tom,
HAMMOUCH Siham, RAMANANTSALAMA Antsaniaina Lalatiana

Avril 2025

Introduction

Le nettoyage et la préparation des données constituent une étape fondamentale et souvent chronophage dans tout projet d'analyse de données. La qualité des données brutes est rarement parfaite : elles peuvent contenir des erreurs, des incohérences, des valeurs manquantes ou des doublons qui, s'ils ne sont pas traités correctement, peuvent fausser les analyses et conduire à des conclusions erronées.

Au cours de divers projets impliquant la manipulation de jeux de données, nous avons été confrontés de manière récurrente à ces défis. Les tâches de nettoyage manuel se sont avérées non seulement fastidieuses et répétitives, mais aussi sources d'erreurs potentielles. Divers problèmes ont été rencontrés tels que la gestion des cellules vides, la standardisation des formats textuels ou l'élimination des enregistrements redondants. Nous avons constaté un manque d'outil intégré, simple et efficace permettant de centraliser ces opérations courantes directement dans l'environnement Excel.

C'est dans ce contexte qu'est né le projet de développer un outil de nettoyage de données en `Visual Basic for Applications (VBA)` pour Excel. Notre motivation principale était de créer une solution pratique et unifiée, capable de fournir une large gamme de réponses aux problèmes de nettoyage les plus fréquents. L'objectif était de concevoir une interface qui permette aux utilisateurs, même sans expertise approfondie en programmation, de réaliser des opérations de nettoyage complexes en seulement quelques clics, via un menu interactif et intuitif.

Ce rapport détaille les fonctionnalités clés de notre outil et explique comment il simplifie et accélère le processus de préparation des données dans Excel. Dans un premier temps, nous présenterons les différentes fonctionnalités intégrées à notre projet, parmi lesquelles la configuration initiale des données, la sélection des colonnes à conserver, la gestion des valeurs manquantes ou encore l'harmonisation des données textuelles. Cette première partie vise à détailler les outils de nettoyage automatisé que notre solution propose. Dans un second temps, nous aborderons l'utilisation et l'interface utilisateur, qui constituent en quelque sorte un mode d'emploi guidé de notre outil, montrant comment l'utilisateur peut accéder aux différentes fonctions via un formulaire interactif intégré à Excel.

1 Fonctionnalités principales

Notre outil VBA pour Excel intègre plusieurs modules de nettoyage accessibles depuis une interface utilisateur (UserForm). Voici les fonctionnalités majeures implémentées :

1.1 Configuration initiale et informations sur les données

Dès que l'utilisateur a déposé sa base de données sur la feuille Excel, l'interface de notre outil le guide pour fournir des informations essentielles. Dans un premier temps, il est invité à renseigner les dimensions de son jeu de données, c'est-à-dire le nombre total de lignes et le nombre total de colonnes que sa base de données occupe sur la feuille à l'aide de zones de texte *(TextBox)* intégrées dans l'UserForm. Ces informations permettent à l'outil de délimiter précisément la plage de données à traiter et seront nécessaires pour les opérations qui suivent. Même si l'utilisateur se trompe dans les chiffres, les dimensions correctes, déjà calculées en amont, seront prises en compte pour la suite.

Ensuite, l'utilisateur se voit offrir la possibilité d'effectuer une première action de nettoyage : la suppression des lignes identiques (doublons complets). Il peut choisir d'éliminer ou non les enregistrements redondants avant même de passer aux nettoyages plus spécifiques.

Enfin, l'utilisateur doit indiquer si son jeu de données comporte une ligne d'en-tête *(header)* contenant les noms des colonnes. Cette spécification se fait simplement en cochant "oui" dans l'interface. La présence ou l'absence d'un en-tête est une information cruciale qui conditionne la manière dont l'outil identifiera et présentera les colonnes dans les étapes ultérieures.

Ainsi, ces renseignements et cette action permettent de configurer l'outil pour une interaction adaptée au jeu de données présent et d'assurer une base plus saine pour les traitements à venir.

1.2 Choix des colonnes à conserver

Avant de procéder au nettoyage spécifique au contenu de chaque colonne (valeurs manquantes, formatage), il est souvent judicieux de s'assurer que le jeu de données ne contient que les informations nécessaires à l'analyse ou au traitement envisagé. Les fichiers sources peuvent en effet contenir de nombreuses colonnes superflues qui alourdissent la manipulation et l'interprétation.

Notre outil intègre une fonctionnalité permettant à l'utilisateur de sélectionner les colonnes qu'il souhaite conserver. Typiquement, via une *ListBox* (une boîte de dialogue listant les en-têtes de colonnes à sélectionner), l'utilisateur peut sélectionner une ou plusieurs colonnes parmi celles présentes dans sa plage de données active. Pour faciliter l'identification des colonnes à traiter, l'interface est conçue pour être la plus intuitive possible. Si le jeu de données sélectionné par l'utilisateur possède une ligne d'en-tête, l'outil affiche directement les noms des colonnes issus de cette ligne d'en-tête dans les options de sélection. Cela permet une visualisation et une identification immédiate des données. En revanche, si aucune ligne d'en-tête n'est présente, l'interface propose alors de numéroter les colonnes. De plus, grâce à la conception de notre UserForm, l'utilisateur conserve généralement une visibilité sur son jeu de données en arrière-plan.

Une fois la sélection des colonnes à conserver validée, l'outil procède à la suppression de toutes les autres colonnes non sélectionnées. Cette opération s'effectue en quelques clics et permet de :

- Simplifier radicalement la structure du jeu de données en ne gardant que l'essentiel.
- Éliminer simplement et rapidement les colonnes jugées inutiles, redondantes ou non pertinentes pour la tâche en cours.
- Préparer un tableau de données plus ciblé et plus léger pour les étapes ultérieures de nettoyage et d'analyse.

Cette étape initiale de filtrage structurel contribue grandement à rendre le travail sur les données plus efficace et plus clair dès le départ.

1.3 Gestion des valeurs manquantes

Les valeurs manquantes (cellules vides) constituent un problème fréquent susceptible de fausser les calculs et les analyses. Notre outil aborde ce défi en proposant une approche adaptative et

ciblée. La première étape cruciale pour l'utilisateur est de spécifier le type de données contenu dans la colonne qu'il souhaite traiter à l'aide d'une liste déroulante *(ListBox)*.

En effet, les stratégies de gestion des valeurs manquantes pertinentes diffèrent selon que les données sont quantitatives, qualitatives ou textuelles. L'interface guide l'utilisateur dans ce choix initial, car les opérations disponibles et leurs implications varient en fonction de la nature des données.

Pour les données quantitatives : Lorsque la colonne sélectionnée contient des données quantitatives, l'outil propose des options spécifiques. Nous avons également tenu compte du fait que ces données peuvent être continues ou discrètes ainsi le traitement sera différent en fonction de la nature des données. Pour traiter les cellules vides dans une colonne numérique, l'utilisateur dispose des choix suivants :

- **Remplacer par la moyenne :** Calculer la moyenne arithmétique des valeurs présentes dans la colonne et l'utiliser comme valeur de remplacement.
- **Remplacer par le mode :** Permet de remplacer les valeurs manquantes par le mode dans le cas discret (pas de chiffres après la virgule).
- **Remplacer par la moyenne + un bruit aléatoire :** Imputer les valeurs manquantes en utilisant la moyenne de la colonne additionnée d'un multiple aléatoire de l'écart-type ($Moyenne + n \times \sigma$). Le facteur n est généré aléatoirement, typiquement dans l'intervalle [-2, 2]. Cette technique plus avancée vise à simuler une distribution, évitant ainsi une concentration des valeurs sur la moyenne.
- **Remplacer par une valeur personnalisée :** Permettre à l'utilisateur de saisir une valeur numérique spécifique qui sera utilisée pour combler toutes les valeurs manquantes dans la colonne.
- **Supprimer les lignes :** Éliminer complètement de la base de données toutes les lignes pour lesquelles une valeur est manquante dans la colonne numérique sélectionnée. Cette option est radicale mais peut être nécessaire si la donnée est indispensable.
- **Ne rien faire :** Laisser les valeurs manquantes telles quelles dans cette colonne spécifique, sans modification. L'utilisateur choisit également cette option si la colonne ne présente aucune valeur manquante.

Pour les données qualitatives : Si la colonne contient des données qualitatives (représentant des catégories, comme "Oui/Non", "Pays", etc.), les options de traitement sont adaptées à cette nature non numérique :

- **Remplacer par le mode :** Identifier la modalité la plus fréquente dans la colonne et l'utiliser pour remplacer les éléments manquants.
- **Remplacer par une valeur personnalisée :** Laisser l'utilisateur définir une catégorie ou un texte spécifique (par exemple, "Inconnu", "Non Renseigné", "Autre") comme substitut pour les valeurs manquantes.
- **Supprimer les lignes :** Supprimer entièrement les lignes où l'information est manquante dans la colonne considérée.
- **Ne rien faire :** Conserver les cellules vides ou les indicateurs de valeur manquante sans y toucher. S'il n'y a pas de valeurs manquantes, l'utilisateur coche cette option.

Pour les données textuelles : Pour les colonnes contenant du texte libre (comme des commentaires, des descriptions), où les approches numériques sont inapplicables et où la notion de catégorie est moins stricte que pour les données qualitatives, les options proposées sont similaires à ces dernières :

- **Remplacer par une valeur personnalisée :** Permettre à l'utilisateur de spécifier lui-même un texte de remplacement.

- **Supprimer les lignes :** Éliminer les lignes pour lesquelles le contenu est manquant dans la colonne sélectionnée.
- **Ne rien faire :** Ne pas modifier les lignes où les cellules textuelles sont vides ou contenant des indicateurs de manque. Ou bien laisser la colonne telle qu'elle car elle ne contient pas de valeurs manquantes.

Dans chaque scénario, l'utilisateur interagit avec l'outil en sélectionnant d'abord la colonne à traiter, puis en indiquant son type de données, et enfin en choisissant l'action de gestion des valeurs manquantes souhaitée parmi les options pertinentes proposées. L'outil exécute ensuite automatiquement l'opération sur la plage concernée.

1.4 Harmonisation des données textuelles

Pour pallier à l'hétérogénéité des données textuelles, notre outil propose un processus d'harmonisation spécifique appliqué aux colonnes textuelles sélectionnées par l'utilisateur.

Cette colonne doit être de type "Chaîne de caractères" pour qu'il puisse faire ces changements.

L'utilisateur peut choisir parmi les opérations de standardisation suivantes :

- **Conversion en minuscules :** L'intégralité du contenu textuel de chaque cellule de la sélection est systématiquement convertie en lettres minuscules.
- **Remplacement des délimiteurs courants par des espaces :** Certains caractères souvent utilisés comme séparateurs ou faisant partie de la ponctuation (tels que les points et les tirets) sont remplacés par un caractère espace unique. L'objectif est de délimiter les mots de manière uniforme.
- **Normalisation des espaces :** Après l'étape précédente, tous les espaces multiples consécutifs sont réduits à un seul espace. De plus, les espaces superflus en début et en fin de chaîne de caractères sont supprimés (fonction équivalente à TRIM).
- **Suppression des accents :** Tous les caractères accentués présents dans le texte (comme é, è, à, ç, ù, î, ô, etc.) sont remplacés par leur équivalent alphabétique non accentué (respectivement e, e, a, c, u, i, o, etc.). Cela neutralise les différences liées à l'utilisation ou non des accents.

L'application rigoureuse de cette séquence de transformations sur les colonnes désignées permet d'obtenir des données textuelles fortement standardisées. Elle élimine un grand nombre de variations qui pourraient empêcher Excel de reconnaître comme identiques des entrées équivalentes. Cette uniformité améliore considérablement la fiabilité des analyses et des manipulations de données ultérieures.

2 Utilisation et interface utilisateur

L'un des objectifs clés du projet était la simplicité d'utilisation. L'outil est conçu pour s'intégrer de manière claire et intuitive dans l'environnement Excel. L'accès aux différentes fonctionnalités se fait typiquement via :

— Un bouton *Lancer le nettoyage* pour démarrer le processus.
— Une succession d'*UserForms* qui guide l'utilisateur à travers les étapes.

Conçu pour la simplicité, l'outil s'intègre à Excel et guide l'utilisateur à travers une interface progressive et intuitive :

1. **Configuration initiale :** L'utilisateur définit d'abord les dimensions de ses données et la présence d'un en-tête via une interface, avec une option pour une suppression rapide des doublons complets.

2. **Sélection des colonnes à conserver :** L'utilisateur sélectionne ensuite les colonnes à conserver.
3. **Traitement des colonnes :** Pour un nettoyage ciblé, il choisit la colonne à traiter (où un aperçu est fourni) avant de passer à l'opération. L'utilisateur sélectionne ensuite l'action de nettoyage souhaitée et configure les options spécifiques proposées par l'interface et ce pour chaque colonne.
4. **Exécution :** Une fois les choix validés, l'outil exécute automatiquement la tâche de nettoyage demandée sur les données concernées.

Cette approche permet de réaliser des opérations de nettoyage complexes en quelques clics, sans nécessiter de formules Excel complexes ou de scripts VBA de la part de l'utilisateur.

À l'issue de ces modifications, l'utilisateur pourra enregistrer son jeu de données nettoyé dans son `Explorateur de fichiers`. Le contenu de la feuille pourra donc être effacé pour pouvoir laisser place à un nouveau jeu de données à traiter.

Conclusion

L'outil de nettoyage de données Excel développé en VBA répond directement aux défis rencontrés lors de la préparation des données. En centralisant des fonctionnalités essentielles telles que la gestion des valeurs manquantes, l'harmonisation textuelle et la suppression des doublons au sein d'une même interface, il offre un gain de temps significatif et réduit le risque d'erreurs manuelles.

Ce projet démontre l'intérêt d'automatiser les tâches répétitives de nettoyage de données directement dans l'environnement de travail. Il fournit une solution pratique et efficace pour améliorer la qualité des données et, par conséquent, la fiabilité des analyses réalisées sous Excel.

En plus de rendre les manipulations plus rapides, notre outil a été pensé pour être utilisé facilement par tout le monde, même sans connaissances en programmation. Grâce à une interface simple et des menus clairs, l'utilisateur est guidé pas à pas dans le nettoyage de ses données, sans avoir besoin de savoir écrire du code. Il permet aussi de s'adapter à différentes situations et est personnalisable en tenant de compte de la nature des données. Cette flexibilité rend l'outil utile dans beaucoup de contextes différents.

Des évolutions futures pourraient inclure des fonctions de validation de données plus avancées, comme la gestion de formats de date par exemple. On pourrait aussi imaginer un bouton de résumé automatique du nettoyage effectué, ou encore des boutons *Undo* qui permettraient à l'utilisateur de revenir en arrière et d'annuler les modifications récentes.

En résumé, cet outil rend le nettoyage de données plus accessible, plus rapide et accessible à tous, quels que soient leurs niveaux techniques. Dans un cadre professionnel, il peut aussi permettre à chacun d'être plus autonome dans son travail, tout en gagnant un maximum de temps sur des tâches souvent longues et répétitives.