# IBM SPSS Data Preparation

*Améliorer la préparation des données pour obtenir des résultats plus précis*

## Points essentiels

- Identification des situations, des variables et des valeurs suspectes ou non valides.
- Visualisation des structures de données manquantes.
- Résumé des distributions des variables.
- Préparer plus rapidement et avec une meilleure qualité les données à analyser.

Tous les analystes doivent préparer leurs données avant de les étudier. IBM SPSS Statistics inclut des outils de préparation des données, mais il est parfois nécessaire d'utiliser des techniques spécifiques. IBM SPSS Data Preparation vous permet d'identifier facilement les situations, les variables et les valeurs suspectes ou non valides, de visualiser les structures de données manquantes, de résumer les distributions des variables et de travailler plus précisément avec des algorithmes conçus pour les attributs nominaux. Cette approche rationalise le processus de préparation des données pour que vous soyez prêt plus tôt pour l'analyse et que vous obteniez des conclusions plus précises. Choisissez entre une procédure de préparation entièrement automatisée pour obtenir plus rapidement vos résultats, ou une autre méthode qui vous aidera à traiter les jeux de données les plus complexes.

SPSS Data Preparation peut être installé en tant que logiciel client autonome. Toutefois, pour des performances et une évolutivité optimales, une version serveur est également disponible.

## Choisissez parmi les options de préparation des données

### La procédure de validation des données

La validation des données est généralement un processus manuel. Vous pouvez par exemple réaliser une analyse de fréquence sur vos données, imprimer les fréquences, identifier ce qui doit être fait et identifier les cas concernés. Ce processus demande beaucoup de temps et, comme chaque analyste dans votre organisation est susceptible d'appliquer une méthode légèrement différente, il est difficile d'assurer la cohérence entre les projets.

Pour éliminer les vérifications manuelles, utilisez plutôt la procédure Validate Data. Elle vous permet d'appliquer des règles de vérification en fonction du niveau de chaque variable (catégorielle ou continue). Par exemple, pour analyser les données d'une enquête qui contient des variables sur une échelle Likert de 5 points, vous pouvez utiliser la procédure Validate Data pour appliquer une règle de vérification et marquer tous les cas qui ont des valeurs en dehors de cette échelle. Vous pouvez recevoir un rapport sur les cas non valides et un résumé des violations des règles et du nombre de cas concernés. Vous pouvez aussi spécifier des règles de validation pour des variables isolées (par exemple, des vérifications de la fourchette de valeurs) et effectuer des vérifications croisées (par exemple, « sexe masculin ET enceinte »).

Ces vérifications vous aideront à définir la validité des données et à éliminer ou corriger vous-même les cas suspects avant l'analyse.

## Préparation des données en une seule étape automatique

La préparation manuelle des données est un processus complexe qui peut représenter 40 à 90 % du temps qu'un analyste consacre à un projet. Si vous avez besoin rapidement des résultats, la procédure Automated Data Preparation (ADP) vous aidera à détecter et corriger les problèmes de qualité et à compléter les valeurs manquantes en une seule étape. La fonction ADP fournit un rapport clair avec des recommandations complètes et des éléments de visualisation qui vous aident à déterminer quelles données utiliser dans l'analyse.

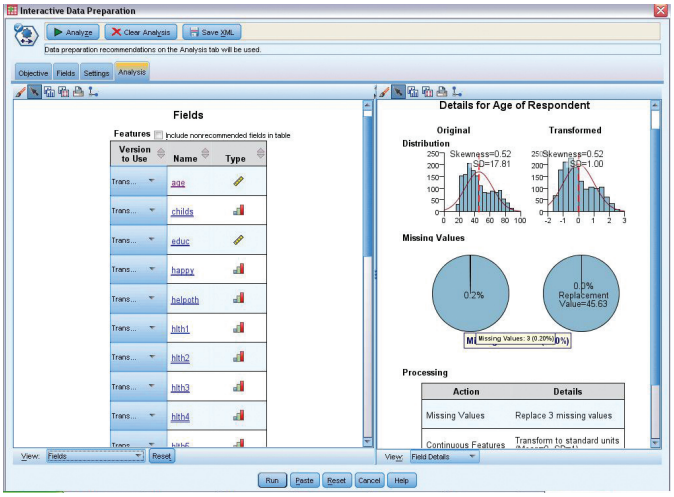


*Figure 1 :* La fonction Automated Data Preparation émet des recommandations et permet à l'utilisateur d'approfondir et d'examiner ces recommandations.

## La procédure de détection des anomalies

Pour éviter que les valeurs extrêmes ne distordent les analyses, vous pouvez utiliser la procédure Anomaly Detection. Elle recherche les cas inhabituels en fonction de leur écart par rapport à des cas similaires, et fournit les causes de l'écart. Vous pouvez alors marquer les valeurs extrêmes à l'aide d'une nouvelle variable. Une fois que vous avez identifié ces cas, vous pouvez les étudier de plus près et déterminer s'ils doivent être inclus ou non dans votre analyse.

## Regroupement optimal

Pour utiliser des algorithmes conçus pour des attributs nominaux (comme les modèles naïfs de Bayes et les modèles LOGIT), vous devez discrétiser vos variables d'échelle avant de créer le modèle. Si vous ne le faites pas, les algorithmes comme la régression logistique multinomiale demanderont beaucoup de temps, et ils peuvent ne pas converger, en particulier si votre jeu de données est très volumineux. De plus, les résultats peuvent être difficiles à lire ou à interpréter.

La procédure de regroupement optimal (Optimal Binning) vous permet de déterminer les niveaux limites qui vous aideront à atteindre les meilleurs résultats possibles pour des algorithmes conçus pour des attributs nominaux.

Avec cette procédure, vous avez le choix entre trois types de regroupement pour le prétraitement des données avant la génération du modèle.

- Non supervisé : crée des regroupements avec un décompte égal.
- Supervisé : Prend en compte la variable cible pour déterminer les niveaux limites. Cette méthode est plus précise que la précédente, mais elle demande plus de ressources de calcul.
- Approche hybride : Combine les approches non supervisée et supervisée. Elle est particulièrement utile si vous avez un grand nombre de valeurs distinctes.

Pour partager et réutiliser efficacement les ressources, pour les protéger selon les exigences de conformité internes et externes et publier les résultats afin qu'un plus grand nombre d'utilisateurs professionnels puisse consulter les résultats et interagir avec ceux-ci, envisagez d'enrichir votre logiciel IBM SPSS Statistics avec IBM SPSS Collaboration and Deployment Services. Vous trouverez des informations supplémentaires sur ces fonctionnalités sur le site suivant : **ibm.com**/spss/cds

Notre suite de logiciels statistiques est maintenant disponible en trois éditions : IBM SPSS Statistics Standard, IBM SPSS Statistics Professional et IBM SPSS Statistics Premium. En regroupant les fonctionnalités essentielles, ces éditions offrent un moyen efficace de garantir que toute votre équipe ou département dispose des fonctions nécessaires pour effectuer les analyses qui contribuent au succès de votre organisation.

## Caractéristiques

### Automated Data Preparation

Recommande des étapes pour accélérer la génération du modèle et améliorer le pouvoir prédictif.

- Déterminer l'objectif : équilibrer la vitesse et l'exactitude, optimiser avec un objectif de vitesse, optimiser avec un objectif d'exactitude, ou personnaliser l'analyse
- Préparer les dates et heures pour la modélisation :
  - Calculer le temps écoulé par rapport à une date de référence
  - Calculer le temps écoulé par rapport à une heure de référence
  - Extraire les éléments temporels cycliques
- Exclure les champs d'entrée de mauvaise qualité :
  - Exclure les champs avec trop de valeurs manquantes
  - Exclure les champs nominaux avec trop de catégories uniques
  - Exclure les champs catégoriels avec trop de valeurs dans une seule catégorie
- Ajuster les niveaux de mesure :
  - Ajuster les niveaux de mesure des champs numériques
- Préparer les champs pour améliorer la qualité des données :
  - Traitement des valeurs extrêmes
  - Remplacer les valeurs manquantes
  - Réorganiser les champs nominaux
- Redimensionner les champs
  - Pondération de l'analyse
- Champs de nature continu
- Champs cible continus
- Transformer les champs
  - En utilisant à la fois des champs de saisies catégorielles et continues
- Réaliser la sélection et la construction des fonctions
- Nommer les champs :
  - Champs transformés et construits
  - Durées calculées
  - Eléments temporels cycliques extraits
- Appliquer les transformations aux données

## Validate data

Utilisez la procédure Validate Data pour valider les données dans le fichier de travail : Vérifications de base : spécifiez les vérifications de base à appliquer aux variables et aux valeurs de votre fichier.

- Par exemple, création d'un rapport qui identifie les variables avec un pourcentage élevé de valeurs manquantes ou les valeurs vides :
  - Pourcentage maximal de valeurs manquantes
  - Pourcentage maximal de cas dans une seule catégorie
  - Pourcentage maximal de cas avec une seule occurrence
  - Coefficient minimal de variation
  - Ecart-type minimal
  - Marquage des identificateurs incomplets
  - Marquage des identificateurs en double
  - Marquage des valeurs vides
- Règles standard : Décrire les données, consulter les règles pour les variables uniques et appliquer ces règles aux variables d'analyse :
  - Description des données :
    - Distribution : Affiche un diagramme à barres en taille réduite pour les variables catégorielles ou un histogramme pour les variables à échelle
    - Les valeurs minimales et maximales des données sont affichées
  - Règles pour des variables uniques :
    - Appliquer les règles aux variables individuelles pour identifier les valeurs manquantes ou non valides, comme les valeurs hors limites
    - Il est aussi possible de définir ses propres règles pour les variables uniques
- Règles personnalisées : Définir des expressions de règle sur variables croisées pour déterminer les réponses non logiques (« sexe masculin ET enceinte » par exemple)
- Résultat : rapports décrivant les données non valides :
  - Rapport indiquant les références des cas, avec une liste des violations des règles de validation pour chaque cas :
    - Spécifier le nombre minimal de violations nécessaires pour qu'un cas soit inclus dans le rapport
    - Spécifier le nombre maximal de cas dans le rapport
  - Rapport sur les règles de validation standard :
    - Résume les violations par variable d'analyse
    - Résume les violations par règles
    - Affiche des statistiques descriptives
- Sauvegarder : Permet de sauvegarder les variables qui présentent des violations des règles, et de les utiliser pour nettoyer les données et filtrer les cas incorrects :
  - Résumé des variables :
    - Indicateur de cas vide
    - Indicateur d'identificateur dupliqué
    - Indicateur d'identificateur incomplet
    - Violation de règles de validation (nombre total)
  - Variables indicateurs enregistrant toutes les violations des règles de validation

## Identifier les cas inhabituels

La procédure Anomaly Detection recherche les cas inhabituels en fonction de leur écart par rapport à des cas similaires, et fournit les causes de l'écart :

- Spécifier les variables à utiliser par la procédure avec la sous-commande VARIABLES. Spécifier les variables catégorielles, continues et ID (pour identifier les cas) et lister les variables exclues de l'analyse.
- La sous-commande HANDLEMISSING permet d'indiquer les méthodes de traitement des valeurs manquantes dans cette procédure :
  - Appliquer le traitement des valeurs manquantes. Si cette option est sélectionnée, les moyennes générales sont substituées aux valeurs manquantes des variables continues, et les catégories manquantes des variables catégorielles sont combinées et traitées comme une catégorie valide. Les variables traitées sont ensuite utilisées dans l'analyse. Si cette option n'est pas sélectionnée, les cas à valeurs manquantes sont exclus de l'analyse.
  - Créer une variable supplémentaire appelée « Missing Proportion Variable » et l'utiliser dans l'analyse. Si cette option est retenue, une variable supplémentaire appelée Missing Proportion Variable est créée. Elle représente la proportion de variables manquantes dans chaque enregistrement, et elle est utilisée dans l'analyse. Si cette option n'est pas retenue, la variable n'est pas créée.
- La sous-commande CRITERIA permet d'indiquer les paramètres suivants :
  - Nombre minimal et maximal dans chaque groupe.
  - Pondération d'ajustement au niveau de la mesure
  - Nombre de causes dans la liste des anomalies
  - Pourcentage de cas considérés comme des anomalies et inclus dans la liste des anomalies
  - Nombre de cas considérés comme des anomalies et inclus dans la liste des anomalies
  - Limite de l'index des anomalies, permettant de déterminer si un cas est considéré comme anomalie ou non
- Sauvegarder les variables supplémentaires dans un fichier de travail avec la sous-commande SAVE :
  - Index des anomalies
  - ID du groupe d'homologues
  - Taille du groupe d'homologues
  - Taille du groupe d'homologues en pourcentage
  - La variable, associée à une cause
  - La mesure de l'impact de la variable, associée à une cause
  - La valeur de la variable, associée à une cause
  - La valeur normale, associée à une cause
- Exporter le modèle dans un nom de fichier spécifié en XML avec la sous-commande OUTFILE
- Commander l'affichage des résultats en sortie avec la sous-commande PRINT
- Vous pouvez imprimer :
  - Le résumé du traitement des cas
  - La liste de l'index des anomalies, la liste des ID homologues des anomalies et la liste des causes des anomalies
  - La table des valeurs normales des variables continues, si des variables continues ont été utilisées dans l'analyse, et les variables catégorielles
  - Les valeurs normales des variables, si des variables catégorielles ont été utilisées dans l'analyse
  - Le résumé de l'index des anomalies
  - La table de résumé des causes pour chaque cause :
    - Supprimer tous les résultats affichés, sauf la table des notes et les avertissements

## Optimal Binning

Pré-traitement des données à l'aide de la fonction de regroupement optimal (Optimal Binning), qui catégorise une ou plusieurs variables continues en répartissant les valeurs de chaque variable dans des regroupements. Cette procédure est utile pour réduire le nombre de valeurs des variables d'entrée du regroupement, ce qui permet d'améliorer considérablement la performance des algorithmes. Lors de l'utilisation de certaines méthodes de regroupement optimal, une variable guide vous aide à déterminer les valeurs limites, et donc à maximiser les relations entre la variable guide et la variable regroupée.

- Sélectionner une des méthodes suivantes :
  - Regroupement non supervisé, avec l'algorithme des fréquences égales. Cette méthode utilise l'algorithme dit « des fréquences égales » pour isoler les variables d'entrée du regroupement. Il n'y a pas besoin de variable guide.
  - Regroupement supervisé avec l'algorithme MDLP (Minimal Description Length Principle). Cette méthode isole les variables d'entrée du regroupement avec l'algorithme MDLP sans prétraitement. Il est adapté aux jeux de données avec un faible nombre de cas. Une variable guide est nécessaire.
  - Regroupement hybride MDLP. Il comprend un prétraitement par l'algorithme des fréquences égales, puis l'algorithme MDLP. Cette méthode est adaptée aux jeux de données avec un grand nombre de cas. Une variable guide est nécessaire.
- Indiquer les critères suivants :
  - Mode de définition de la limite minimale de chaque variable d'entrée du regroupement
  - Mode de définition de la limite maximale de chaque variable d'entrée du regroupement
  - Mode de définition de la limite minimale d'un intervalle
  - Décision de forcer la fusion des regroupements faiblement peuplés
  - Décision de traiter les valeurs manquantes par une suppression à partir d'une liste ou par paires
- Sauvegarder les éléments suivants :
  - Nouvelles variables contenant les valeurs regroupées
  - Syntaxe vers un fichier de syntaxe SPSS Statistics Base
- Commander l'affichage des résultats avec la sous-commande PRINT. Vous pouvez imprimer :
  - Les ensembles de limites des variables d'entrée du regroupement
  - Les informations de description de toutes les variables d'entrée du regroupement
  - Entropie du modèle pour les variables regroupées

## Configuration système requise

La configuration dépend de la plateforme. Pour plus d'informations, consultez **ibm.com**/spss/requirements

## À propos d'IBM Business Analytics

Les logiciels IBM Business Analytics permettent aux décideurs de disposer des précieux éclairages dont ils ont besoin pour améliorer les performances métier. IBM propose à cet effet une gamme complète et unifiée d'applications d'aide à la décision, d'analyse prédictive avancée, de pilotage de la stratégie et des performances financières, de gouvernance, de gestion du risque et de la conformité et d'applications analytiques.

Avec les logiciels IBM, les entreprises peuvent non seulement détecter les tendances, les schémas récurrents et les anomalies, comparer des scénarios de simulation, prédire les menaces et opportunités potentielles mais aussi planifier, élaborer les budgets et prévoir les ressources nécessaires. Grâce aux puissantes fonctions analytiques dont ils disposent, nos clients dans le monde entier sont à même de mieux comprendre, anticiper et maîtriser leurs résultats métier.

## Pour plus d'informations

Pour plus d'informations, visitez le site :
**ibm.com**/business-analytics/fr

## Nous contacter

Pour demander à être appelé ou pour poser une question, accédez au site **ibm.com**/business-analytics/fr
Un représentant IBM vous répondra sous deux jours ouvrés.

**Compagnie IBM France**
17 Avenue de l'Europe
92 275 Bois-Colombes Cedex

La page d'accueil d'IBM est accessible à l'adresse suivante :
**ibm.com**



Pensez à recycler ce document

YTD03016-FRFR-04