



www.datalgo.fr

Audit du fichier « TESTSOC » comportant 2693 enregistrements

Analyse réalisée par Datalgo

Contenu :

- Préambule
- Synthèse graphique
- Analyse détaillée champ par champ
- Conclusions

Préambule :

- ❑ Cet audit de fichier a pour objet de réaliser une analyse détaillée de chaque champ afin de vous permettre d'effectuer des ajustements sur vos données : corrections, fusions, suppressions, ajouts...

Ce rapport présente en amont une synthèse graphique de votre fichier permettant de comprendre immédiatement les points forts et points faibles de votre base de données.

Pour chaque champ, Datalgo réalise un comptage et une analyse des contenus.

- Les comptages sur le format, les doublons et la qualité permettent de connaître avec exactitude les paramètres de chaque champ.
- Un feu tricolore mesure simplement l'état de vos données : tout va bien (vert), améliorations possibles (orange) ou danger (rouge).
- Quand cela est nécessaire, Datalgo vous signale que votre fichier devrait éventuellement faire l'objet d'une déclaration à la CNIL.
- Des commentaires ajoutent une explication aux différents comptages.
- Souvent Datalgo vous fournit des conseils pour optimiser la qualité de vos données, voir réduire les coûts d'exploitation.
- Dans certains cas, Datalgo vous propose de réaliser des traitements spécifiques qui vous aideront à effectuer les bons correctifs sur votre base de données.
Pour obtenir de plus amples informations sur ces traitements suggérés, vous pouvez lire leur description complète sur le site www.datalgo.fr/traitement.

- ❑ Pour chaque champ, quels sont les comptages et analyses réalisés dans le rapport d'audit ?

FORMAT :

- Nombre d'enregistrements renseignés (exhaustivité) : détermine la proportion d'enregistrements renseignés par opposition à des enregistrements restés vides de toute information.
- Nombre d'enregistrements numériques : établit un ratio entre les informations numériques ou textuelles. Cela permet dans bien des cas de détecter des erreurs de saisie : texte dans un champ numérique et réciproquement.
- Ratio Loi de Benford : cette loi permet de vérifier qu'un champ de données numériques contient bien des valeurs numériques aléatoires (chiffres d'affaire, métrages, quantités...) Fréquemment utilisée pour l'audit comptable, cette loi démontre qu'un ratio supérieur à 10% est peut-être le signe qu'il y a des erreurs ou même des irrégularités dans vos données.
- Largeur du champ : précise le nombre maximum de caractères contenus dans ce champ. Cette information permet souvent de formater au plus juste un système de gestion de base de données afin d'optimiser la taille des fichiers... et la rapidité des sauvegardes.

DOUBLONS STRICTS :

- Nombre d'enregistrements strictement uniques : affiche 100% lorsqu'il n'y a aucun doublon.
- Nombre de doublons stricts : précise le nombre de doublons stricts (c'est-à-dire contenant des valeurs absolument identiques).

- Nombre total d'enregistrements affectés par les doublons : calcule le nombre de fiches concernées par ces doublons dans votre base de données.

- Nombre maximum d'enregistrements répétés : précise s'il s'agit de doublons (2 fiches identiques), triplons (3 fiches identiques), quadruplons (4 fiches identiques), etc.

DOUBLONS APPROCHANTS :

- Nombre de doublons approchants : précise le nombre de doublons dont les valeurs sont ressemblantes mais pas forcément identiques.

- Nombre total d'enregistrements affectés par les doublons : calcule le nombre de fiches concernées par ces doublons approchants dans votre base de données.

- Nombre maximum d'enregistrements répétés : précise s'il s'agit de doublons (2 fiches ressemblantes), triplons (3 fiches), quadruplons (4 fiches), etc.

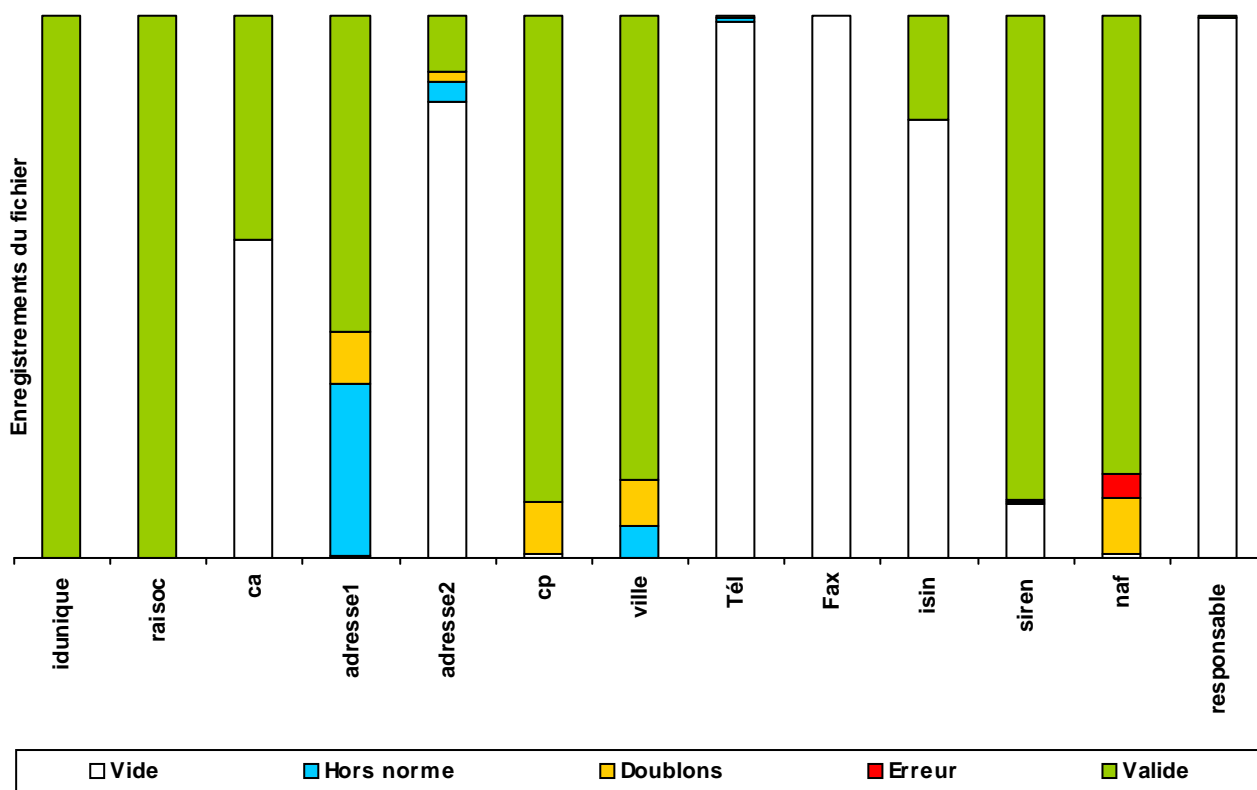
QUALITÉ :

- Nombre d'enregistrements hors normes : certaines données doivent être normalisées, soit pour être présentées correctement et facilement déduplicables, soit pour réduire les coûts d'affranchissement lorsqu'il s'agit d'adresses postales.

- Nombre d'erreurs (code ou donnée erronée) : pour certains types de données, il est possible de vérifier le contenu via un algorithme (mauvais numéro SIREN, TVA Intracommunautaire, numéro de sécurité sociale, nomenclature d'activité, etc).

Synthèse graphique :

Contenu des champs pour les 2693 enregistrements du fichier.



Ce graphique détaille visuellement les éventuels problèmes de contenus dans votre base de données. Chaque champ est matérialisé par une barre verticale contenant les 2693 enregistrements du fichier «TESTSOC».

Vide :

Représente les enregistrements restés vides, donc non renseignés. La valeur zéro n'est pas considérée comme une valeur vide.

Hors norme :

Certaines informations doivent généralement répondre à un format de données prédéfini (par exemple, les SIREN sont sur 9 chiffres, les adresses ne doivent pas dépasser 38 caractères par lignes, etc). Normaliser les contenus facilite la déduplication, abaisse les coûts d'affranchissement (pour les adresses), rend les tris et sélections plus performants, permet une validation plus sûre des données.

Doublons :

Repérage des données en double, ou approchantes. Sur certains champs comme les contacts ou les noms de société, il est important de repérer les données semblables qui pourraient constituer des doublons, triplons, etc.

Erreur :

Pour certains types de données, il est possible de vérifier le contenu via un algorithme. Ce marqueur permet donc d'appréhender le nombre d'erreurs de saisie possible (mauvais numéro SIREN, TVA Intracommunautaire, numéro de sécurité sociale, nomenclature d'activité, etc)

Valide :

Cette dernière visualisation distingue les informations qui a priori ne présentent pas de problèmes de normalisation, de doublonnages ou d'erreurs flagrantes. Il n'en reste pas moins qu'une information, même normalisée peut être fausse ! Par exemple, une adresse peut être aux normes postales sans pour autant aboutir chez le bon correspondant.

Analyse détaillée des 13 champs de la table TESTSOC

Nom du champ : {idunique}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	2 693	100,00%
	Nombre d'enregistrements numériques :	2 693	100,00%
	Largeur du champ (nombre de caractères maximum) :	5	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2693 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	2 693	100,00%
	Nombre de doublons stricts :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :		
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Qualité →	Nombre d'enregistrements hors normes :		
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires



Mesure de l'exhaustivité des données (100,00%)

Ce champ est rempli à 100%.



Clef unique (100,00%)

Ce champ contient une "clef", c'est-à-dire un code identifiant unique pour chaque enregistrement.



Doublons stricts (0,00%)

Ce champ ne contient aucun doublon.



Format du champ (numérique ou texte) (0,00%)

Champ 100% numérique : les valeurs enregistrées sont toutes des chiffres (sauf erreur de conversion possible, par exemple sous Excel).

Nom du champ : {raisoc}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	2 692	99,96%
	Nombre d'enregistrements numériques :	1	0,04%
	Largeur du champ (nombre de caractères maximum) :	60	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2692 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	2 692	100,00%
	Nombre de doublons stricts :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	1	0,04%
	Nombre total d'enregistrements affectés par les doublons :	2	0,07%
	Nombre maximum d'enregistrements répétés :	2	
Qualité →	Nombre d'enregistrements hors normes :		
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires



Mesure de l'exhaustivité des données (99,96%)

Il manque quelques valeurs à ce champ pour être complètement renseigné.



Clef unique (99,96%)

Il existe quelques doublons dans ce champ.

• Suggestions :

Si les enregistrements de ce champ doivent être uniques, vous devez corriger les doublons ou compléter les valeurs vides.



Doublons stricts (0,00%)

Ce champ ne contient aucun doublon.



Format du champ (numérique ou texte) (99,96%)

Ce champ contient quelques valeurs numériques parmi des valeurs habituellement textuelles.

• Suggestions :

Assurez-vous qu'il soit normal que quelques chiffres se soient glissés dans ce champ contenant habituellement du texte. Un simple tri sur ce champ permettra de mettre en exergue les chiffres.



Doublons approchants des noms de société (0,04%)

Vous avez quelques doublons de société dont le nom est ressemblant.

- **Suggestions :**

Datalgo peut vous sélectionner la liste des sociétés susceptibles d'être en doublons dès lors que leur raison sociale est ressemblante.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage des noms de société => détection des doublons possibles lorsque des raisons sociales de société sont très ressemblantes.
- Dédoublonnage par similitude => cet algorithme complexe analyse deux mots ou groupes de mots et attribue un pourcentage (score) de similitude entre les deux permettant de retrouver des doublons aux orthographes différentes au sein d'un même fichier.

Nom du champ : {ca}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	1 116	41,44%
	Nombre d'enregistrements numériques :	1 116	41,44%
	Ratio Loi de Benford	10,22%	
	Largeur du champ (nombre de caractères maximum) :	9	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (1116 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	1 084	97,13%
	Nombre de doublons stricts :	6	0,54%
	Nombre total d'enregistrements affectés par les doublons :	32	2,87%
	Nombre maximum d'enregistrements répétés :	22	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :		
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Qualité →	Nombre d'enregistrements hors normes :		
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires

Mesure de l'exhaustivité des données (41,44%)

Ce champ est faiblement complété.

Doublons stricts (0,54%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• **Suggestions :**

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.

Format du champ (numérique ou texte) (0,00%)

Champ 100% numérique : les valeurs enregistrées sont toutes des chiffres (sauf erreur de conversion possible, par exemple sous Excel).



Loi de Benford sur les valeurs numériques (10,22%)

Selon la loi de Benford, ce champ contient des valeurs numériques répétées. S'il s'agit d'informations financières, c'est peut-être le signe qu'il y a des erreurs ou même des irrégularités.

- **Suggestions :**

Si ce champ doit être aléatoire, une vérification du contenu de ce champ s'impose : ne manque-t-il pas des données ? Certaines informations ont-elles été comptabilisées plusieurs fois ?

Nom du champ : {adresse1}

	Détail de l'analyse	Valeurs	% du nb total (2693 regist.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	2 684	99,67%
	Nombre d'enregistrements numériques :	0	0,00%
	Largeur du champ (nombre de caractères maximum) :	38	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2684 regist.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	2 105	78,43%
	Nombre de doublons stricts :	211	7,86%
	Nombre total d'enregistrements affectés par les doublons :	579	21,57%
	Nombre maximum d'enregistrements répétés :	13	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	263	9,80%
	Nombre total d'enregistrements affectés par les doublons :	718	26,75%
	Nombre maximum d'enregistrements répétés :	14	
Qualité →	Nombre d'enregistrements hors normes :	852	31,74%
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires



Recommandation : Des adresses en double ont été détectées. Un dédoublonnage par match-code et/ou similarité avec les noms de société pourrait permettre d'affiner la détection de réels doublons sur ce fichier.



Mesure de l'exhaustivité des données (99,67%)

Il manque quelques valeurs à ce champ pour être complètement renseigné.



Doublons stricts (7,86%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• Suggestions :

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.



Normalisation des voies postales (31,74%)

Vous avez un nombre important d'adresses non normalisées.

- **Suggestions :**

Faites normaliser vos adresses par Datalgo, et bénéficiez auprès de La Poste de tarifs préférentiels pour vos envois en nombre.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Mise aux normes des voies postales et villes => réduisez les coûts de routage en utilisant la normalisation des adresses postales : simplification des libellés de voie, suppression des éléments parasites, respect du nombre de caractères par lignes...



Doublons approchants des adresses (9,80%)

Vous avez quelques doublons d'adresses dont le nom est ressemblant.

- **Suggestions :**

Datalgo peut vous sélectionner la liste des adresses susceptibles d'être en doublons car ressemblantes.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage des adresses => détection des doublons à partir de la ressemblance des adresses postales au sein d'un même fichier.
- Dédoublonnage par match-code => repérage des fiches en double d'une base de données en créant un code de rapprochement à partir de plusieurs champs d'une même fiche.

Nom du champ : {adresse2}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	423	15,71%
	Nombre d'enregistrements numériques :	0	0,00%
	Largeur du champ (nombre de caractères maximum) :	32	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (423 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	330	78,01%
	Nombre de doublons stricts :	41	9,69%
	Nombre total d'enregistrements affectés par les doublons :	93	21,99%
	Nombre maximum d'enregistrements répétés :	5	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	45	10,64%
	Nombre total d'enregistrements affectés par les doublons :	104	24,59%
	Nombre maximum d'enregistrements répétés :	5	
Qualité →	Nombre d'enregistrements hors normes :	97	22,93%
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires



Mesure de l'exhaustivité des données (15,71%)

Ce champ est faiblement complété.



Doublons stricts (9,69%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• Suggestions :

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.



Normalisation des voies postales (22,93%)

Vous avez un nombre important d'adresses non normalisées.

• Suggestions :

Faites normaliser vos adresses par Datalgo, et bénéficiez auprès de La Poste de tarifs préférentiels pour vos envois en nombre.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Mise aux normes des voies postales et villes => réduisez les coûts de routage en utilisant la normalisation des adresses postales : simplification des libellés de voie, suppression des éléments parasites, respect du nombre de caractères par lignes...



Doublons approchants des adresses (10,64%)

Vous avez un nombre important de doublons d'adresses dont le nom est ressemblant.

- **Suggestions :**

Datalgo peut vous sélectionner la liste des adresses susceptibles d'être en doublons car ressemblantes.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage des adresses => détection des doublons à partir de la ressemblance des adresses postales au sein d'un même fichier.
- Dédoublonnage par match-code => repérage des fiches en double d'une base de données en créant un code de rapprochement à partir de plusieurs champs d'une même fiche.

Nom du champ : {cp}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	2 678	99,44%
	Nombre d'enregistrements numériques :	2 662	98,85%
	Largeur du champ (nombre de caractères maximum) :	6	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2678 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	817	30,51%
	Nombre de doublons stricts :	263	9,82%
	Nombre total d'enregistrements affectés par les doublons :	1 861	69,49%
	Nombre maximum d'enregistrements répétés :	349	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :		
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Qualité →	Nombre d'enregistrements hors normes :		
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires



Recommandation : Attention ce fichier a probablement été constitué à partir de données venant d'Excel ce qui a eu pour effet de convertir les codes postaux en valeurs numériques. Ainsi, certains codes postaux commençant par zéro ont été mis sur 4 caractères (exemple 06100 => 6100). Un correctif devra être appliqué.



Mesure de l'exhaustivité des données (99,44%)

Il manque quelques valeurs à ce champ pour être complètement renseigné.



Doublons stricts (9,82%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• Suggestions :

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublement par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.



Format du champ (numérique ou texte) (0,60%)

Ce champ contient quelques valeurs non numériques.

• Suggestions :

Assurez-vous qu'il est normal que quelques lettres se soient glissées dans ce champ habituellement numérique.

Nom du champ : {ville}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	2 691	99,93%
	Nombre d'enregistrements numériques :	0	0,00%
	Largeur du champ (nombre de caractères maximum) :	31	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2691 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	635	23,60%
	Nombre de doublons stricts :	238	8,84%
	Nombre total d'enregistrements affectés par les doublons :	2 056	76,40%
	Nombre maximum d'enregistrements répétés :	900	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	237	8,81%
	Nombre total d'enregistrements affectés par les doublons :	2 083	77,41%
	Nombre maximum d'enregistrements répétés :	900	
Qualité →	Nombre d'enregistrements hors normes :	153	5,69%
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires



Mesure de l'exhaustivité des données (99,93%)

Il manque quelques valeurs à ce champ pour être complètement renseigné.



Doublons stricts (8,84%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• Suggestions :

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.



Normalisation des villes (5,69%)

Vous avez quelques villes non normalisées (caractères parasites, signes diacritiques, abréviations, etc).

- **Suggestions :**

Faites normaliser les coordonnées de vos contacts, et bénéficiez auprès de La Poste de tarifs préférentiels pour vos envois en nombre.


Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Mise aux normes des voies postales et villes => réduisez les coûts de routage en utilisant la normalisation des adresses postales : simplification des libellés de voie, suppression des éléments parasites, respect du nombre de caractères par lignes...

Nom du champ : {Tél}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	33	1,23%
	Nombre d'enregistrements numériques :	33	1,23%
	Largeur du champ (nombre de caractères maximum) :	14	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (33 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	31	93,94%
	Nombre de doublons stricts :	1	3,03%
	Nombre total d'enregistrements affectés par les doublons :	2	6,06%
	Nombre maximum d'enregistrements répétés :	2	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	2	6,06%
	Nombre total d'enregistrements affectés par les doublons :	4	12,12%
	Nombre maximum d'enregistrements répétés :	2	
Qualité →	Nombre d'enregistrements hors normes :	27	81,82%
	Nombre d'erreurs (code ou donnée erronée) :	1	3,03%

Commentaires

 **Recommandation :** La plupart des téléphones sont enregistrés avec un espace entre les groupes de 2 chiffres ce qui explique le faible score de téléphones normés (on préférera enregistrer les téléphones sans espace ou ponctuation pour favoriser un rapprochement ou un dédoublonnage des données).

Mesure de l'exhaustivité des données (1,23%)

Ce champ est très faiblement complété.

• **Suggestions :**

Dans certains cas, il vaut mieux supprimer un champ très faiblement renseigné.

Doublons stricts (3,03%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• **Suggestions :**

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.

Format du champ (numérique ou texte) (0,00%)

Champ 100% numérique : les valeurs enregistrées sont toutes des chiffres (sauf erreur de conversion possible, par exemple sous Excel).



Erreur de numérotation téléphonique (3,03%)

Vous avez quelques numéros de téléphones erronés.

- **Suggestions :**

Datalgo peut vous fournir la liste des numéros de téléphone erronés à corriger en comparant le numéro du département avec le préfixe téléphonique.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Préfixes téléphoniques français => vérification des deux ou quatre premiers chiffres de la numérotation téléphonique française en corrélation avec le département.



Numéros de téléphone hors norme (81,82%)

Vos numéros de téléphone ne sont pas tous normalisés.

- **Suggestions :**

Vous pouvez demander à Datalgo de normaliser vos numéros de téléphone en retirant les espaces ou la ponctuation.



Doublons de numéros de téléphone (6,06%)

Vous avez quelques téléphones qui apparaissent en double après normalisation.

- **Suggestions :**

Vous pouvez demander à Datalgo de détecter les doublons de numéros de téléphone.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.

Nom du champ : {Fax}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	3	0,11%
	Nombre d'enregistrements numériques :	2	0,07%
	Largeur du champ (nombre de caractères maximum) :	16	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (3 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	3	100,00%
	Nombre de doublons stricts :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Qualité →	Nombre d'enregistrements hors normes :	2	66,67%
	Nombre d'erreurs (code ou donnée erronée) :	1	33,33%

Commentaires



Mesure de l'exhaustivité des données (0,11%)

Ce champ est très faiblement complété.

• Suggestions :

Dans certains cas, il vaut mieux supprimer un champ très faiblement renseigné.



Doublons stricts (0,00%)

Ce champ ne contient aucun doublon.



Format du champ (numérique ou texte) (33,33%)

Ce champ contient des valeurs numériques et textuelles.

• Suggestions :

Assurez-vous qu'il soit normal d'avoir dans le même champ des valeurs sous forme de texte et de chiffres.



Erreur de numérotation téléphonique (33,33%)

Vous avez des numéros de téléphones erronés.

- **Suggestions :**

Datalgo peut vous fournir la liste des numéros de téléphone erronés à corriger en comparant le numéro du département avec le préfixe téléphonique.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Préfixes téléphoniques français => vérification des deux ou quatre premiers chiffres de la numérotation téléphonique française en corrélation avec le département.



Numéros de téléphone hors norme (66,67%)

Vos numéros de téléphone ne sont pas tous normalisés.

- **Suggestions :**

Vous pouvez demander à Datalgo de normaliser vos numéros de téléphone en retirant les espaces ou la ponctuation.



Doublons de numéros de téléphone (0,00%)

Vous n'avez pas téléphones en double.

Nom du champ : {isin}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	512	19,01%
	Nombre d'enregistrements numériques :	0	0,00%
	Largeur du champ (nombre de caractères maximum) :	12	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (512 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	512	100,00%
	Nombre de doublons stricts :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Qualité →	Nombre d'enregistrements hors normes :	0	0,00%
	Nombre d'erreurs (code ou donnée erronée) :	0	0,00%

Commentaires

Mesure de l'exhaustivité des données (19,01%)

Ce champ est faiblement complété.

Doublons stricts (0,00%)

Ce champ ne contient aucun doublon.

Erreur de code ISIN (0,00%)

L'algorithme de vérification des codes Isin ne détecte aucune erreur sur vos codes ISIN.

ISIN hors norme (0,00%)

Tous vos codes ISIN corrects sont normalisés.

Doublons de code ISIN (0,00%)

Vous n'avez pas de codes ISIN en double.

Nom du champ : {siren}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format	Nombre d'enregistrements renseignés (exhaustivité) :	2 424	90,01%
	Nombre d'enregistrements numériques :	2 418	89,79%
	Largeur du champ (nombre de caractères maximum) :	9	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2424 enregistr.)
Doublons stricts <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	2 412	99,50%
	Nombre de doublons stricts :	6	0,25%
	Nombre total d'enregistrements affectés par les doublons :	12	0,50%
	Nombre maximum d'enregistrements répétés :	2	
Doublons approchants <small>Données ressemblantes</small>	Nombre de doublons approchants :	6	0,25%
	Nombre total d'enregistrements affectés par les doublons :	12	0,50%
	Nombre maximum d'enregistrements répétés :	2	
Qualité	Nombre d'enregistrements hors normes :	0	0,00%
	Nombre d'erreurs (code ou donnée erronée) :	13	0,54%

Commentaires



Recommandation : Une incohérence est détectée : des doublons de Siren existent alors qu'il n'y a pas de doublons de codes ISIN.



Mesure de l'exhaustivité des données (90,01%)

Il manque quelques valeurs à ce champ pour être complètement renseigné.



Doublons stricts (0,25%)

Ce champ contient quelques doublons stricts (c'est-à-dire des valeurs strictement identiques).

• Suggestions :

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.



Format du champ (numérique ou texte) (0,25%)

Ce champ contient quelques valeurs non numériques.

• Suggestions :

Assurez-vous qu'il est normal que quelques lettres se soient glissées dans ce champ habituellement numérique.



Erreur de code SIREN (0,54%)

Vous avez quelques codes SIREN faux.

- **Suggestions :**

Datalgo peut vous fournir la liste exhaustive des codes SIREN erronés à corriger.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- SIREN (identifiant des sociétés françaises) => algorithme de vérification des codes SIREN pour les sociétés françaises.



SIREN hors norme (0,00%)

Tous vos codes SIREN corrects sont normalisés.

- **Suggestions :**

Vous pouvez faire calculer par déduction les numéros de TVA Intracommunautaire à partir du numéro SIREN. Ainsi, votre base de données sera à la norme européenne en vigueur depuis janvier 1993.

- Converti un code SIREN en code TVA Intracommunautaire => conversion des codes SIREN en code TVA intracommunautaire pour les sociétés françaises.



Doublons de code SIREN (0,25%)

Vous avez quelques codes SIREN qui après normalisation apparaissent en double.

- **Suggestions :**

Vous pouvez demander à Datalgo de détecter les doublons de code SIREN.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.

Nom du champ : {naf}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	2 670	99,15%
	Nombre d'enregistrements numériques :	2 670	99,15%
	Ratio Loi de Benford	99,19%	
	Largeur du champ (nombre de caractères maximum) :	4	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (2670 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	132	4,94%
	Nombre de doublons stricts :	277	10,37%
	Nombre total d'enregistrements affectés par les doublons :	2 538	95,06%
	Nombre maximum d'enregistrements répétés :	348	
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	277	10,37%
	Nombre total d'enregistrements affectés par les doublons :	2 538	95,06%
	Nombre maximum d'enregistrements répétés :	348	
Qualité →	Nombre d'enregistrements hors normes :	0	0,00%
	Nombre d'erreurs (code ou donnée erronée) :	116	4,34%

Commentaires



Mesure de l'exhaustivité des données (99,15%)

Il manque quelques valeurs à ce champ pour être complètement renseigné.



Doublons stricts (10,37%)

Ce champ contient un nombre important de doublons stricts (c'est-à-dire des valeurs strictement identiques).

• Suggestions :

Si cela est nécessaire, Datalgo peut vous sélectionner la liste des données en double.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- Dédoublonnage par contenu identique => détection des enregistrements strictement identiques au sein d'un même fichier.



Format du champ (numérique ou texte) (0,00%)

Champ 100% numérique : les valeurs enregistrées sont toutes des chiffres (sauf erreur de conversion possible, par exemple sous Excel).



Loi de Benford sur les valeurs numériques (99,19%)

Selon la loi de Benford, ce champ contient des valeurs numériques répétées. S'il s'agit d'informations financières, c'est peut-être le signe qu'il y a des erreurs ou même des irrégularités.

- **Suggestions :**

Si ce champ doit être aléatoire, une vérification du contenu de ce champ s'impose : ne manque-t-il pas des données ? Certaines informations ont-elles été comptabilisées plusieurs fois ?



Erreur de code NAF (4,34%)

Vous avez quelques codes NAF faux.

- **Suggestions :**

Datalgo peut vous fournir la liste exhaustive des codes NAF erronés à corriger.

Datalgo vous propose pour cela d'effectuer les traitements suivants :

- NAF (nomenclature d'activité française) => vérification de la segmentation par activité des sociétés selon le code NAF officiel en France.



NAF hors norme (0,00%)

Tous vos codes NAF corrects sont normalisés.

- **Suggestions :**

Vous pouvez demander à Datalgo d'ajouter dans votre fichier les segmentations des libellés NAF en toutes lettres.

- Affichage et segmentation des libellés NAF => ajout des libellés pour les codes de la Nomenclature d'Activité Française : division, groupe et classe.

Nom du champ : {responsable}

	Détail de l'analyse	Valeurs	% du nb total (2693 enregistr.)
Format →	Nombre d'enregistrements renseignés (exhaustivité) :	5	0,19%
	Nombre d'enregistrements numériques :	0	0,00%
	Largeur du champ (nombre de caractères maximum) :	18	
	Détail de l'analyse	Nombre d'enregist.	% du nb renseignés (5 enregistr.)
Doublons stricts → <small>Données rigoureusement identiques</small>	Nombre d'enregistrements strictement uniques :	5	100,00%
	Nombre de doublons stricts :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Doublons approchants → <small>Données ressemblantes</small>	Nombre de doublons approchants :	0	0,00%
	Nombre total d'enregistrements affectés par les doublons :		
	Nombre maximum d'enregistrements répétés :		
Qualité →	Nombre d'enregistrements hors normes :		
	Nombre d'erreurs (code ou donnée erronée) :		

Commentaires

Protection des données personnelles : ce champ contient des données permettant l'identification d'une personne physique. En France, vous devez déclarer ce fichier à la CNIL.



Mesure de l'exhaustivité des données (0,19%)

Ce champ est très faiblement complété.

• Suggestions :

Dans certains cas, il vaut mieux supprimer un champ très faiblement renseigné.



Doublons stricts (0,00%)

Ce champ ne contient aucun doublon.

Conclusions :

En analysant pour chaque champ les "alertes" (feux tricolores), cet audit de fichier fait apparaître :



17 feux verts : ce qui signifie des annotations tout à fait positives pour les données concernées.



30 feux oranges : des améliorations peuvent être réalisées sur certains champs.



11 feux rouges : des erreurs sont disséminées dans le fichier. Un nettoyage précis des données semble nécessaire. Pour chaque champ problématique, Datalgo vous a proposé des traitements correctifs.

131 erreurs sont totalisées dans l'ensemble des champs et enregistrements du fichier.

CNIL : vous avez au moins un champ relatif à des données personnelles. Pour la France, vous devez déclarer ce fichier auprès de la CNIL.

En savoir plus > <http://www.datalgo.com/flash-20041015-CNIL-loi-8-aout-2004.htm>



Recommandations générales et conclusions :

Un certain nombre de doublons existent dans ce fichier. Parallèlement, un travail de normalisation de certaines données conduirait à faciliter le repérage des doublons et à améliorer l'adressage postal.

o o o

Après cet audit, Datalgo se tient à votre disposition pour effectuer les traitements nécessaires au nettoyage de votre fichier.

Fait le jeudi 17 mars 2005