

データ作業、 AIに任せられること・ 任せられないこと

生成AI時代のデータクレンジング・名寄せ・個人情報の考え方を、
サンプルデータでわかりやすく解説

—— AIを安心して使うための7つのチェックポイント

表記ゆれの統一や重複の洗い出しは、生成AIでかなりの部分ができるようになりました。では、データの整備はもうAIに任せればよいのでしょうか。本資料では、AIが得意な「作業」と、人が決めなければならない「判断」を、架空のサンプルデータで切り分けます。個人情報をAIに渡すときの注意点、AIに入れる前のデータ整備、AI時代の不正対策、人とAIの役割分担まで、データに詳しくない方にも読めるようにまとめました。

合同会社オンエーケー（OnAK）

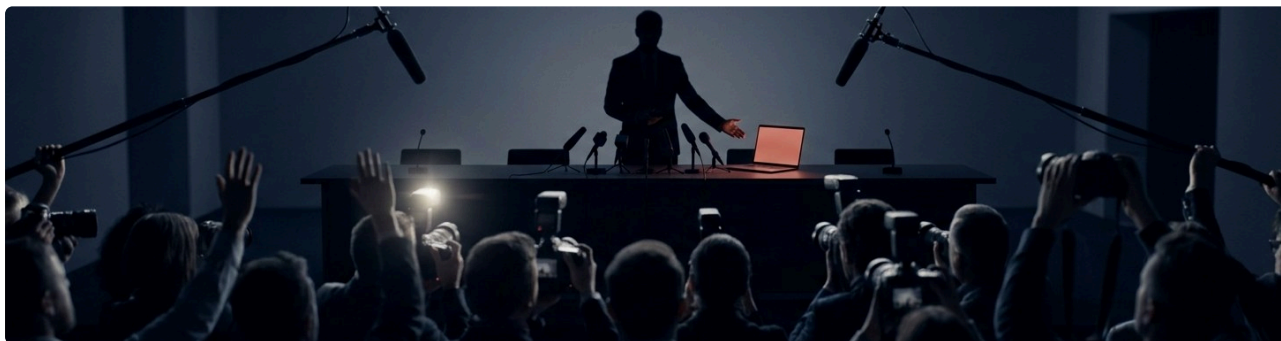
データクレンジング・名寄せ・データ突合の専門サービス

「データ整理は、もうAIでいいのでは？」

ChatGPTをはじめとする生成AIに、表計算のデータを貼り付けて「表記ゆれを直して」「重複をまとめて」と頼む。数年前には考えられなかったことが、いまは誰でもできます。ちょっとした整理なら、AIのほうが速く正確なことも少なくありません。

AIを使い始めた現場で起きていること（よくある例）

- ・ AIに整えさせたら、見た目はきれいなのに、中身が少しずつ違っていた
- ・ 同じ指示を出したのに、昨日と今日で結果が変わった
- ・ 「この人をなぜ対象外にしたのか」と聞かれ、AIがそうしたから、としか答えられない
- ・ 顧客リストをAIに貼り付けてよいのか、社内でルールが決まっていない



「AIがやりました。私たちは悪くありません」—— この説明は、通用しません。 ※イメージ（架空の場面です）

AIが変えたのは、データ整備の「作業」の部分です。逆に言えば、何を同じとみなすか、どこまでの誤りを許すか、結果をどう説明するか——という「判断」と「責任」の重みは、AIの登場でむしろ増えています。

本資料は、第1弾『データクレンジング入門』の続編です。「AIに任せてよいこと」と「任せてはいけないこと」の境目を、サンプルデータでお見せします。

この資料で使う言葉

ハルシネーション	AIが、事実ではない内容をもっともらしく作り出してしまうこと。
名寄せ（なよせ）	バラバラに登録された「同じ人」のデータを見つけて、1人分にまとめること。
仮名化（かめいか）	氏名などを記号やIDに置き換え、そのデータだけでは個人がわからないようにすること。
推論と学習	推論＝入力をもとにAIが回答を作ること。学習＝入力をAIの改良の材料にすること。

AIに任せられること——「作業」と「候補出し」

まず、AIが得意なことから見ていきましょう。共通するのは、**ルールがはっきりしていて、結果を人が確かめやすい作業**です。

作業	AIの向き・不向き	ポイント
全角・半角、スペース、記号の統一	◎ 得意	ルールを先に決めて伝えれば、ほぼ正確。エクセルの関数より柔軟
会社名の「株式会社」「(株)」「株」の統一	◎ 得意	そろえ方（前株・後株を含めるかなど）を指示しておく
住所の書き方をそろえる	○ おおむね得意	「一丁目2番3号→1-2-3」は得意。ただし町名や番地を 補ってしまうことがある （次ページ）
重複していそうな行の候補出し	○ おおむね得意	「太郎／太朗」のような揺れも拾える。 候補として出させる のが正しい使い方
自由記述の分類・要約	○ おおむね得意	アンケートの感想を「価格」「品質」などに分ける作業は大きな時短に
エクセル関数・マクロの作成の手伝い	◎ 得意	作った関数は、少ない件数で動きを確かめてから使う
同じ人かどうかの最終判定	× 任せない	判定の基準は事業ごとの判断。説明責任が伴う（4ページ）
個人情報を含む全件の処理	△ 条件次第	サービスの設定・契約・社内ルールの確認が先（5ページ）

数千件規模の一般的なデータを、一般的な生成AIで処理した場合の目安です。サービスや指示の仕方によって結果は変わります。

ひとことで言うと… AIは「優秀で手の速いアルバイト」。決められた作業はどんどん進めてくれますが、何を同じとみなすか、どこまで直してよいかを決めるのは、仕事を頼む側です。

大切なのは、AIに「答え」ではなく「下書き」や「候補」を作らせるという考え方です。候補を人が確かめる形にしておけば、AIの速さを生かしながら、誤りを最終結果に持ち込まずに済みます。

AIに表記ゆれを直させると、何が起きるか

ある会員データ（架空）の住所・氏名・会社名を、生成AIに「書き方をそろえて」と指示して整えさせた結果の一部です。

No	元のデータ	AIが整えた結果	何が問題か
1	練馬区〇〇一丁目2番3号	練馬区〇〇1-2-3	正しく整えられている
2	府中市〇〇町1-5	東京都府中市〇〇町1-5	都道府県がない元データに「東京都」を補った。広島県府中市の可能性もある
3	練馬区〇〇2-	練馬区〇〇2-1-1	番地の欠けを、もっともらしい数字で埋めてしまった（ハルシネーション）
4	高橋 一郎	高橋 一郎	異体字を統一。照合用ならよいが、宛名に使うと失礼になることも
5	(株)オンエーケー／(有)オンエーケー	オンエーケー	法人格を消したため、別の2社が1社に見えるようになった
6	090-1234-5678	09012345678	そろえ方は正しいが、先頭の0が消える形式で保存すると「9012345678」に

緑＝正しい、黄＝使い道によっては問題、赤＝誤り。6件中、そのまま使えるのは1件だけです。

AIの誤りが見つけにくい3つの理由

(1) もっともらしい。 No.3のように、AIの誤りは「ありそうな値」の形をしています。入力ミスと違い、目で見ても違和感がありません。

(2) 何を変えたか報告しない。 AIは頼まれない限り、「どの行の、どこを、なぜ変えたか」を記録しません。数万件になると、元データと見比べるのは現実的ではありません。

(3) 毎回同じとは限らない。 同じデータに同じ指示を出しても、結果が少し変わることがあります。キャンペーンの抽選のように「同じ処理をすれば同じ結果になる」ことが求められる場面では、大きな弱点です。

ご担当者のチェックポイント AIで整えたデータは、「変更前と変更後の件数」「変更した行の一覧」が出せる形で処理していますか？ 出せないなら、結果を説明できません。

「同じ人かどうか」は、AIに決めさせない

まず、クイズです。次の4件は、ある会員データの一部です（架空のデータ）。

No	氏名	住所	電話番号	生年月日
F	山田 太郎	練馬区〇〇1-2-3	03-1234-5678	1958/6/10
G	山田 太郎	練馬区〇〇1-2-3	03-1234-5678	1988/2/14
H	佐藤 花子	練馬区〇〇4-5-6 〇〇マンション201	090-1111-2222	1985/4/1
I	佐藤 花子	練馬区〇〇4-5-6 〇〇マンション305	080-3333-4444	1962/11/20

Q FとG、HとIは、それぞれ「同じ人」でしょうか？ ☐ 同じ人 ☐ 別の人

回答欄
回答欄の質問の順番は回別・H（(と意中03.04日自主) 左隣の回答欄は逆の回別・F（「Y」の順）からス：と答
▲ 答えは逆さまに書いてあります

同じデータを生成AIに渡して「同じ人をまとめて」と頼むと、AIはF・Gも、H・Iも「同一人物」として1人にまとめました。AIは「似ているものをまとめる」のは得意ですが、どの項目の違いを重く見るべきかは、事業の目的によって変わるため、AIには判断できません。

場面	F・Gの扱い	理由
商品券の「1人〇口まで」	別人として扱う	親子それぞれに申込の権利がある
「1世帯1通」のDM	同じ世帯としてまとめる	同じ住所に2通届くのは無駄
会員の購買分析	別人として扱う	年代が違えば、好みも買い方も違う

同じデータでも、正解は場面ごとに変わります。これを決めるのは、事業の目的を知っている人です。

「AIがそう判断しました」は、説明になりません。落選した応募者や、ポイントが付かなかった会員から問い合わせがあったとき、答えるのは発注者であり、その委託先です。判定の基準を人が決め、言葉で残し、AIはその基準に沿って候補を出す——この順番を守ることが、AI時代の名寄せの基本です。

個人情報とAI —— ルールは「育っている途中」

顧客リストや応募者データを、AIに入れてよいのか。ルールが「ない」わけではありません。むしろ、法律、国のガイドライン、Pマーク、ISMSと、よりどころは複数あります。ただ、そのどれもが**生成AI**の登場より前に作られたか、生成AIへの対応を始めたばかりです。

よりどころ	もともと決めていること	生成AIについて示していること	まだ判断が残る部分
個人情報保護法 個人情報保護委員会	個人情報の取得・利用・提供のルール	2023年の注意喚起で、個人データを入れるならAI事業者が 学習に使わないこと等を十分に確認 するよう求めた	AIへの入力が「委託」「第三者提供」「提供に当たらない」のどれか。専門家の間でも見解が分かれる
AI事業者ガイドライン 総務省・経済産業省	AIを開発・提供・利用する事業者の心構えと取り組み	AIを使う企業にも、利用ルールの整備、重要な判断への人の関与、記録の確認を求める	「どこまでやれば十分か」は示されない。2024年の初版から毎年改訂中
Pマーク JIS Q 15001・JIPDEC	個人情報を守る社内の仕組みの認証	JIPDECは、利用目的・第三者提供・海外への移転を確かめ、入力は慎重にと呼びかけている	AIサービスを「委託先」として管理するのか、何を確認すれば足りるのか
ISMS ISO/IEC 27001	情報セキュリティ全般の仕組みの認証	2022年版でクラウドサービス利用の管理が加わり、AIサービスもその一つとして扱える	学習への利用や誤った出力など、AI特有の問題は直接は扱わない
AIMS ISO/IEC 42001	AIそのものを管理する仕組みの認証	2023年に国際規格、2025年にJIS化。国内でも認証が始まった	取得企業はまだ少なく、取引条件として定着していない

出典：個人情報保護委員会、総務省・経済産業省、JIPDEC、ISMS-AC等の公表資料（2026年9月時点）をもとに当社で整理。

つまり、物差しは複数あるのに、「この使い方ならOK、ここから先はNG」という線は、どの物差しにもはっきりとは引かれていません。しかも、物差し同士はつながっていません。Pマークを持っていても、それだけでAIの使い方が保証されるわけではなく、ISMSも同じです。

ひとことで言うと… 交通ルールで言えば、「安全運転しなさい」という原則はあるのに、制限速度の標識がまだ立っていない道路。何キロまで出してよいかは、運転する側が考えて決めるしかありません。

なぜ「判断ライン」が引きにくいのか

(1) 同じ行為でも、契約と設定で結論が変わる

「顧客リストをAIに入れる」という同じ行為でも、法人向けか個人向けか、学習に使われる設定か、委託契約があるか、サーバーが海外か——で、問題のない使い方にも、避けるべき使い方にもなります。

(2) AIの変化に、ルールが追いつかない

AI事業者ガイドラインは2024年4月の初版から毎年改訂され、2026年にはAIエージェント（自分で判断して操作するAI）への対応が加わりました。2026年7月成立の改正個人情報保護法も、AI開発のためのデータ利用を条件付きで広げましたが、細かな条件は施行（公布から2年以内）までに規則で決まります。

(3) 「適法」と「信頼される」の間に、すき間がある

法律上は問題がなくても、会員や応募者が「自分のデータがAIに使われた」と知れば不信感を持つかもしれません。どこまで許すかは、顧客や取引先との約束の問題でもあります。

判断ラインの例——迷いどころはどこにあるか

使い方	判断の方向	迷いどころ
架空のデータで、関数やマクロの作り方を相談する	問題になりにくい	本物のデータが紛れ込まないように運用で防げるか
氏名をIDに置き換えた照合用データを、学習に使われない法人向けAIで処理する	比較的安全	IDと氏名の対応表を、AIとは別にきちんと管理できているか
顧客リストをそのまま、法人向けAIで処理する	条件次第	「委託」と整理する考え方が有力だが、契約・監督・記録が前提。Pマーク取得企業なら委託先管理の手続きも必要
クライアントから預かったデータを、AIで処理する	条件次第	委託契約にAI利用の定めがなければ、まずクライアントの了承を得る
顧客リストを、個人向けAIに貼り付ける	避けるべき	学習に使われる設定のまま入力されるおそれ。会社として管理できない

一般的な考え方の整理です。実際の判断は、データの内容、契約、利用するサービスの規約によって変わります。

ルールは成熟していくのか——待つより、自社の線を引く

方向としては、ルールは確実に具体的になっていきます。改正個人情報保護法の規則づくり、AI事業者ガイドラインの毎年の改訂、AIMS認証の広がり、そしてPマークやISMSの審査でのAIの扱いの具体化。数年のうちに、「標識」は少しずつ立っていくでしょう。

ただし、AIの進化も同じ速さで進みます。ルールが「完成」するのを待っていると、何も使えないまま取り残されるか、気づかないうちに線を越えてしまうか、どちらかになりがちです。

いま必要なのは、「自社の判断ライン」を言葉にして持つことです。

- ① どの使い方を認め、どれを認めないかを書き出す（前ページの表のような形で）
- ② なぜそう判断したかの理由を、法律・ガイドライン・顧客との約束に照らして残す
- ③ ルールやサービスの規約が変わったら、見直す日を決めておく

同じAIでも、契約プランで「学習」の扱いが違う

サービス	個人向けプラン（無料・有料）	法人向けプラン・API（有料）
ChatGPT	初期設定では学習に利用される。設定でオフにできる	初期設定で学習に利用されない
Gemini	アクティビティの保存がオンだと、会話が学習や人によるレビューに使われることがある。設定でオフにできる	Google Workspace版・有料APIは学習に利用されない（無料APIは改善に利用）
Claude	利用者の選択による。オンにすると学習に利用され、保存期間も長くなる	学習に利用されない（契約で明記）

2026年8月時点の各社の公表情報をもとに整理。設定や規約は変わることがあるため、必ず最新の内容をご確認ください。

渡し方の基本 会社として契約したサービスだけを使い、社員個人のアカウントは使わない。AIに見せるのは「作業に必要な最小限」。氏名をIDに置き換え（仮名化）、照合に必要な列だけを渡す。関数づくりの相談は架空のデータで。海外で扱われる場合や、預かったデータの場合は、契約も確認する。

※一般的な注意点の整理であり、法的な助言ではありません。具体的な判断は、弁護士などの専門家にご確認ください。

AIに入れる前に、データを整える

AIで顧客分析をしたい、AIのチャットボットに会員情報を参照させたい、AIで需要を予測したい。こうした「AIの活用」は、どれも元のデータが正しいことが前提です。データの世界には、「**ゴミを入れれば、ゴミが出てくる（Garbage In, Garbage Out）**」という言葉があります。AIは、入ってきたデータが重複していても、それを疑わずに分析します。

架空の会員データ1万件で、名寄せの前と後を比べてみましょう。

項目	名寄せ前	名寄せ後	何が起きていたか
会員数	10,000人	8,200人	1,800件が同じ人の重複登録だった
1人あたりの年間購入回数	2.4回	2.9回	1人の購入が複数の会員に分かれていた
優良顧客（年10回以上）	120人	210人	90人の上客が「普通の会員」に埋もれていた
DMの送付数	10,000通	8,200通	1通100円なら約18万円の差

数値は説明のための架空の例です。重複の割合はデータの集め方によって大きく変わります。

名寄せ前のデータをそのままAIに分析させると、「優良顧客は120人」という誤った前提で施策が組み立てられます。AIの分析が高度になるほど、**土台のデータの誤りは見えにくくなり、影響は大きくなります。**

STEP 1

点検する

重複・表記ゆれ・欠けがどれくらいあるかを、まず数字で把握する。

STEP 2

整える

正規化と名寄せで、比べられる・数えられる状態にする。判定の基準は人が決める。

STEP 3

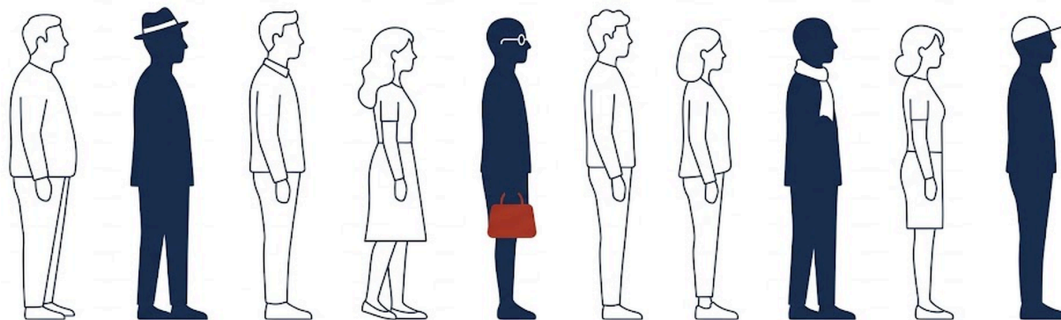
AIで活用する

整ったデータで分析や自動化を行う。結果に違和感があれば、データに戻って確かめる。

ご担当者のチェックポイント AI活用の提案を受けたら、「元のデータの重複や表記ゆれは、誰が、いつ点検しますか？」と聞いてみてください。この工程が計画に入っていないなら要注意です。

不正する側も、AIを使い始めている

キャンペーンの重複応募や、ポイントの不正取得。AIは不正を見つける側の道具であると同時に、**不正をする側の道具**にもなっています。名前や住所の書き方を少しずつ変えた申込データを大量に作ることは、AIを使えば簡単です。



帽子を変えても、同じ人。——書き方を変えた重複応募は、こう見えています。 ※イメージ

実際のデータでは、次のように見えます（架空のデータです）。

No	氏名	住所	メールアドレス	応募時刻
J	鈴木 一郎	〇〇市△△2-8-1	s.ichiro01@example.jp	23:41
K	鈴木 市郎	〇〇市△△二丁目8番1号	s.ichiro02@example.jp	23:42
L	スズキ イチロウ	〇〇市△△2-8-1-101	suzuki.i03@example.jp	23:44
M	鈴木 一朗	〇〇市△△2丁目8-1	ichiro.s04@example.jp	23:45

1件ずつ見ると、どれも普通の応募です。名前の書き方はすべて違い、メールアドレスも別々です。

4件を並べると、**住所の番地が同じ、メールアドレスに連番がある、数分おきに応募している**という共通点が見えてきます。これは、表記ゆれを「直す」だけでは見つかりません。どんな組み合わせが不自然かという、**不正の手口についての知識**が必要です。

AI時代の不正対策で大切なこと

ルールは見破られる前提で。ルールは知られればすり抜けられます。判定の観点は公開せず、応募データの傾向を見ながら見直し続けます。

AIは「怪しい候補」を出す役。不自然なパターン探しはAIの得意分野。ただし、応募を無効にする判断は、根拠を確かめて人が行います。

人とAIの役割分担——人が決め、AIが作業し、人が確かめる

工程	AIの役割	人の役割
目的を決める	—	何のためのデータか、どんな失敗を避けたいかを定める
ルールを決める	ルール案の下書き、過去の例の整理	そろえ方、同じ人とみなす基準、許せる誤りの範囲を決めて文書にする
整える	表記の統一、住所の書き方そろえ	ルールどおりに処理されたか、件数で確かめる
候補を出す	重複・同一世帯・不正の疑いがある候補の抽出	候補の出し方に偏りや漏れがないか確かめる
判定する	判断材料の整理	候補を確認し、まとめる・まとめないを決める
検品する	集計・照合の手伝い	抜き取り確認と、処理前後の件数の突き合わせ
説明する	報告書の下書き	判断の根拠と経過を、わかる言葉で説明する

黄色が「人が決める・人が責任を持つ」部分です。

AIの出力を「検品」する方法

AIの結果を全件確かめるのは大変です。次の3つを組み合わせます。

CHECK 1

件数を突き合わせる

処理前の件数＝残した件数＋まとめた件数＋除外した件数、になっているか。

CHECK 2

抜き取りで確かめる

無作為に選んだ一定数（たとえば200件）を人が見て、誤りの割合を出す。

CHECK 3

迷う例を重点的に

同姓同名、同じ建物、法人格違いなど、間違いやすい行を優先して確かめる。

AIを安全に使う3つの原則

- ① AIには「答え」ではなく「下書き・候補」を作らせる
- ② 何を変えたかを記録し、同じ処理をすれば同じ結果になるようにする
- ③ AIに渡す個人情報は必要最小限にする

AIを使う前・委託する前に確認したい7つのポイント

社内でAIを使う場合も、AIを使う委託先に任せる場合も、同じ観点で確認できます。各項目に「聞き方」の例を添えました。

1 目的と判断の基準が、言葉になっているか

何を同じとみなすか、どこまでの誤りを許すかが文書になっていて、発注者と委託先で共有されているか。

▶ 聞き方の例：「同じ人とみなす基準を、文書で見せてもらえますか？」

2 個人情報をどのAIに、どこまで渡すか決まっているか

学習利用・保存場所・利用目的を確認した会社契約のサービスだけを使い（社員個人のアカウントは使わない）、渡す項目を最小限にしているか。

▶ 聞き方の例：「AIを使う場合、どのサービスに、どの項目を渡しますか？」

3 AIの出力を検品しているか

件数の突き合わせと抜き取り確認を行い、誤りの割合を把握しているか。

▶ 聞き方の例：「AIで処理した結果の誤りは、どうやって見つけていますか？」

4 同じ人かどうかの最終判断を、誰がしているか

AIの候補をそのまま採用せず、迷う例（4ページのF・G、H・I）を人が確かめているか。

▶ 聞き方の例：「同名の親子や、同じ建物の同姓同名はどう扱いますか？」

5 処理の経過を記録し、再現できるか

何件を変更・統合・除外したかを記録し、同じ処理をすれば同じ結果になるか。

▶ 聞き方の例：「もう一度同じ処理をしたら、同じ結果になりますか？」

6 不正の手口の変化に合わせて、見直しているか

AIを使った新しい不正の傾向をふまえ、判定の観点を更新し続けているか。

▶ 聞き方の例：「最近の不正応募には、どんな傾向がありますか？」

7 誤りが見つかったときの責任と手順が決まっているか

AIによる誤りも含め、誰が・どこまで・どの手順で直すかが事前に決まっているか。

▶ 聞き方の例：「AIの処理で誤りがあった場合、どう対応しますか？」

まとめ——AIは道具、決めるのは人

生成AIによって、表記ゆれの統一や重複候補の洗い出しといった「作業」は、速く、安くなりました。一方で、何を同じとみなすかという判断、個人情報の扱い、結果についての説明責任は、AIに任せることができません。

AIを使うかどうかではなく、AIの結果を確かめ、根拠をもって説明できる体制があるかどうか。それが、AI時代のデータ整備で問われることです。

「決まっていなかった」は、通用しません

こう言っても……	通用しない理由
「国のルールが、まだ決まっていなかったの で」	個人情報保護法は、AIにも今のまま適用される
「わが社に、AI利用の規定はなかったの で」	ルールを整えること自体が、法律の求める安全管理の一部。規定がないことが不備になる
「社員が勝手に入力したので」	従業員を監督する義務は、会社にある
「委託先に任せていたので」	委託先を監督する義務は、発注者にある
「AIがそう判断したので」	その基準を決めたのは誰か、と必ず聞かれる

ご質問・ご相談の窓口

表記ゆれ・重複・欠けなど、お手元のデータで気になっていることを、**まずは一行でお送りください**。「AIに入れる前に、今のデータの状態を知りたい」「委託先の処理が適切か確かめたい」という段階のご相談も歓迎しています。

合同会社オンエーカー（OnAK）は、データクレンジング・名寄せ・データ突合の専門会社です。行政・金融・流通分野の大規模なデータ処理の経験をもとに、AIを道具として使いながら、データの品質と判断の根拠を設計・検証します。

お問い合わせ：onak.co.jp のフォーム（ご用件「名寄せ・住所の照合についての質問」）

Mail：kano.toshiyuki@onak.co.jp ／ 第1弾『データクレンジング入門』も onak.co.jp で公開中