

WHITE PAPER | データ品質の見極め方

データクレンジング入門

データクレンジングとは？表記ゆれ・重複・名寄せを、
サンプルデータでわかりやすく解説
——キャンペーンや顧客管理で失敗しないための7つのチェックポイント

名前や住所の書き方の揺れ、重複登録、「同じ人」の見分け方。データクレンジングと名寄せは、キャンペーンの重複応募対策から、DM・会員管理、顧客分析まで、あらゆるデータ活用の土台です。本資料では、データに詳しくない方にもわかるよう、架空のサンプルデータを使って基礎を説明し、エクセルでできることの範囲や外注を検討する目安、委託先を選ぶときに確認したい観点までまとめました。

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

導入 —— データは、入力のたびに揺れる

名前、住所、電話番号。こうしたデータは、入力する人やタイミングによって書き方が変わります。

「山田太郎」と「山田 太郎」、「1-2-3」と「一丁目2番3号」。人間が見れば同じでも、コンピュータは「一文字でも違えば別のもの」と判断します。

この「揺れ」を放っておくと、「同じ人」を見分けられず、さまざまな場面で問題が起きます。

たとえば、こんなことが起こり得ます

- ・ **キャンペーン・プレミアム商品券（プレミアム付商品券）**：同じ人が書き方を変えて何度も応募し、「1人〇口まで」の制限をすり抜けてしまう
- ・ **DM・会員管理**：同じ人に同じ案内が2通、3通と届き、費用がかさむうえに印象も悪くなる
- ・ **顧客分析**：1人の優良顧客が「3人の別人」に分かれて記録され、本当の上客が見えなくなる
- ・ **逆のケース**：同姓同名の別人を「同じ人」とまとめてしまい、正しい応募者や顧客に不利益が出る

こうした問題を防ぐのが、データを整える**データクレンジング**と、同じ人を見つけてまとめる**名寄せ**です。どちらも目立たない工程ですが、その精度には会社ごとに大きな差があります。

この資料の読み方

本資料は、データの専門家でない方——たとえば、キャンペーンや販促を企画・提案する広告代理店のご担当者——にも読んでいただけるように書いています。専門用語には説明を添え、架空のサンプルデータで「何が起きているか」を具体的にお見せします。

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。
データクレンジング	表記の揺れや入力ミスを直し、データを比べられる形に整えること。名寄せの下ごしらえ。
スコアリング	顧客の買い方に点数をつけ、誰にどんな特典を出すかを決めること。

よくある表記ゆれと、似た言葉との違い

よくある表記ゆれの例

種類	同じものののに、書き方が違う例
氏名	山田 太郎／山田太郎（スペースの有無）、高橋／高橋（異体字）、太郎／太朗（入力ミス）
会社名	株式会社オンエーケー／(株)オンエーケー／(株)オンエーケー／オンエーケー
住所	1-2-3／1丁目2番3号／一丁目2-3、マンション名の有無や略し方
半角・全角	ヤマダ／ヤマダ、A B C／ABC、1 2 3／123
ハイフン・長音	「-」「-」「ー」「ー」など、見た目の似た記号の混在
電話番号	090-1234-5678／09012345678／(090)1234-5678

「表記ゆれ」は「表記揺れ」とも書きます（意味は同じ）。数万件になると、見つけるだけで大仕事です。

データクレンジング・正規化・名寄せとの違い

言葉	意味	この資料での位置づけ
データクレンジング	データの誤り・揺れ・重複・欠けを直し、使える状態に整えること	全体をまとめた呼び方
データクリーニング	ほぼ同じ意味。分析の前処理として、欠損値や外れ値の処理を指して使われることも多い	同じ意味として扱う
正規化（標準化）	書き方を一定のルールにそろえること（全角→半角、「丁目」→ハイフンなど）	クレンジングの中の一工程
名寄せ	同じ人・同じ会社のデータを見つけて、1つにまとめること	整えたあとに行う工程

※データベース設計でいう「正規化」は、表の分け方を整える別の意味の言葉です。

作業の順番は、**正規化で書き方をそろえる → 重複の候補を見つける → 名寄せで1つにまとめる**。書き方がそろっていないまま名寄せをしても、同じ人を見つけられません。ひとことで言えば、クレンジングは「掃除」、正規化は「そろえる」、名寄せは「まとめる」です。

基礎知識① 名寄せとは何か

申込フォームに入力される名前や住所は、人によって、また同じ人でも入力のたびに書き方が変わります。コンピューターは「一文字でも違えば別のもので」と判断するため、そのままでは同じ人を見分けられません。

下の表は、あるキャンペーン（プレミアム商品券の申込）のデータの一部です（架空のデータです）。

No	氏名	住所	電話番号
A	山田 太郎	練馬区〇〇1-2-3	090-1234-5678
B	山田 太郎	練馬区〇〇一丁目2番3号	09012345678
C	ヤマダ タロウ	練馬区〇〇1 - 2 - 3	090-1234-5678

黄色の部分が「書き方の違い」。人間が見ればおそらく同じ人ですが、コンピューターにとっては3人の別人です。

名寄せは、次の3つの段階で「同じ人」を見つけていきます。

STEP 1

完全一致

一文字違わず同じものだけを「同じ人」とする。确实だが、上の例ではA・B・Cは別人扱いになる。

STEP 2

あいまい照合

「太郎／太郎」「1-2-3／一丁目2番3号」「ハイフンの有無」などの違いを吸収して比べる。A・B・Cが「同じ人の候補」になる。

STEP 3

判定ルール設計

「名前と電話番号が一致すれば同じ人」など、どの項目の組み合わせで判定するかを、事業の目的に合わせて決める。

ひとことで言うと… 名寄せは「表記ゆれを直して、同じ人を1人にまとめる作業」。Excelの「重複の削除」ボタンでは、上の3行は1行も消えません。

名寄せの落とし穴——「見落とし」と「まとめすぎ」

名寄せの判定をゆるくすれば重複はたくさん見つかりますが、今度は「別人を同じ人とみなしてしまう」危険が高まります。反対に厳しくすれば安全ですが、重複を見落とします。これが名寄せの一番難しいところです。

次の例を見てください（架空のデータです）。

No	氏名	住所	生年月日
D	佐藤 花子	練馬区〇〇1-2-3 〇〇マンション201	1985/4/1
E	佐藤 花子	練馬区〇〇1-2-3 〇〇マンション305	1962/11/20

名前も建物も同じですが、部屋番号と生年月日が違います。これは同じ建物に住む同姓同名の別人です。

もし「名前と番地が同じなら同じ人」という判定ルールだと、DとEは1人にまとめられ、どちらかの応募が無効になってしまいます。正しく応募した人を落選させることになり、苦情や信頼の低下につながります。

判定の方針	起こりやすいこと	事業への影響
厳しくする	名前の書き方を変えた重複応募（前ページのA・B・C）を見落とす	口数制限のすり抜け・予算の流出
ゆるくする	同姓同名の別人（上のD・E）をまとめてしまう	正しい応募者の不利益・苦情

正解は一つではなく、事業ごとに「どちらの失敗をどこまで許すか」を決める必要があります。大切なのは、その判断基準を発注側と委託先で共有し、あとから確認できる形で残しておくことです。

ご担当者のチェックポイント 委託先に「重複はどうやって見つけていますか？」「同姓同名の別人をまとめてしまわない工夫はありますか？」と聞いてみてください。すぐに具体例で答えられるかどうか、ひとつの目安になります。

エクセルでできること・できないこと

数千件程度で、項目が少なく、1回きりの作業なら、エクセル（Excel）でも十分に整えられることがあります。一方で、エクセルだけでは難しい作業もあります。

エクセルでできること	エクセルだけでは難しいこと
・ TRIM関数で余分なスペースを消す ・ ASC関数／JIS関数で半角・全角をそろえる ・ SUBSTITUTE関数で記号を置き換える ・ 「重複の削除」で、完全に同じ行を消す ・ パワークエリ（Power Query）で置換の手順を記録し、繰り返し使う	・ 「太郎／太朗」のような入力ミス、同じ人の候補として拾う（あいまい照合） ・ 「1-2-3」と「一丁目2番3号」のような住所の書き方をそろえる ・ 同姓同名の別人を見分ける ・ 約105万行を超えるデータや、複数ファイルをまたいだ突き合わせ ・ 各工程で何件を直し、何件をまとめたかを記録して説明する

「重複の削除」は一文字でも違えば別の行として残します。3ページのA・B・Cは、1行も消えません。

ご担当者のチェックポイント 件数が多い、同じ作業がキャンペーンのたびに発生する、結果を説明する責任がある——このどれかに当てはまるなら、エクセルでの手作業から一歩進んだ仕組みを検討する時期です。

基礎知識② 重み付け（スコアリング）とは何か

ポイント施策や会員施策では、お客さまの買い方に応じて点数をつけ（スコアリング）、誰にどんな特典を出すかを決めます。代表的な方法が**RFM分析**です。

R（Recency） 最後買ったのはいつか —— 最近の人ほど高得点

F（Frequency） どれくらいの頻度で買っているか —— 回数が多いほど高得点

M（Monetary） いくら使っているか —— 金額が大きいほど高得点

架空の会員5人で見てみましょう（各項目を1～3点で採点）。

会員	最終購入	購入回数（年）	購入金額（年）	R	F	M	合計	判定
会員1	3日前	24回	12万円	3	3	3	9	優良顧客
会員2	10日前	12回	4万円	3	2	2	7	育成候補
会員3	5か月前	20回	10万円	1	3	3	7	離れかけ
会員4	2日前	2回	3千円	3	1	1	5	新規
会員5	1年前	1回	1千円	1	1	1	3	休眠

会員2と会員3は合計点が同じ7点ですが、中身はまったく違います。会員3は以前は優良顧客でしたが、5か月買っていません。点数の合計だけでは見えないことがあります。

このように、スコアリングは「どう使うか」で結果が大きく変わります。設計のときに押さえておきたい観点は、次のページからの5つです。

ひとこと言うと… スコアリングは「お客さまの通知表」。ただし、何を目的に付けるかで、見るべき教科が変わります。

スコアリング設計で押さえない5つの観点

(1) そもそもRFMが使える商材が確かめる

RFM分析は、住宅・自動車・投資商品のように、めったに買わない高額な商品には向きません（「購入回数」がほぼ全員1回になるためです）。プレミアム商品券やキャンペーンの多くは日常の買い物に結びつくので基本的には使えますが、**最初に一度立ち止まって確かめる**ことが大切です。

(2) 「応募だけする人」を先に分けておく

キャンペーンのたびに応募だけを繰り返し、ふだんは買わない層がいます。たとえば「年12回応募しているが、キャンペーン期間外の購入はゼロ」という会員です。この層を混ぜたまま点数をつけると、全体の数字が歪み、「本当のファン」が見えなくなります。

(3) 目的は一つに絞り、目的に合った分析方法を選ぶ

「もっと買ってもらう」「別の商品も買ってもらう」「離れた人を呼び戻す」「他社から乗り換えてもらう」を一度に狙うと、たいていどれも中途半端になります。さらに、**目的によって使うべき分析方法そのものが違います**。優良顧客をもっと大切にしたいならRFM分析、前ページの会員3のような「離れかけた人の呼び戻し」が目的なら、在籍期間や離脱期間を軸に加えるCPM分析が向いています。

(4) 業界ごとの「節目」を考える

業界には、それぞれ購買の目安となる数字があります。たとえばタバコ業界では「連続3箱」がひとつの節目とされます。こうした節目をふまえて点数を設計すると、スコアが実態に近づきます。

よくある失敗例（抽象化しています）

- ・食品メーカーが嗜好飲料と自社のお菓子を組み合わせで販売しようとした例：組み合わせに無理があり、応募数が通常の3分の1以下に
- ・飲食チェーンが「はしご」来店を促した例：客単価の高い店から低い店へ誘導する形になり、動線も逆になって、反応率は0.1%以下に

(5) 「すぐ上のランク」が見える特典にする

会員ランク制は、「次のランクまであと少し」という目標がお客さまのやる気を高める、というのが業界で広く共有されている考え方です。ポイントは、特典がお金の価値に限らず「**周りから見える**」こと。航空会社の優先搭乗がよい例です。何を「見える特典」にできるかは、クライアントの事業の中に答えがあります。

外注・代行を検討する目安

データクレンジングは社内でも行えますが、次の項目に当てはまる数が多いほど、外部の専門家に任せたほうが確実で、結果的に費用も抑えやすくなります。

- ・件数が数万件を超える
- ・Web・はがき・店頭など、複数の申込経路やシステムのデータをまとめる必要がある
- ・住所や氏名の表記ゆれが多く、エクセルの関数だけでは直しきれない
- ・抽選・当選・ポイント付与など、結果について説明を求められる
- ・同じ作業が、キャンペーンや月次処理のたびに発生する

費用を見積もるときに伝えたいこと

データクレンジングの費用は、件数・項目・求める精度・納期によって大きく変わります。相談の段階で次の情報がそろっていると、見積もりが正確になり、話も早く進みます。

件数	おおよその行数と、今後増える見込み
項目	氏名・住所・電話番号・会社名など、整えたい項目
データの形	Excel・CSV・システムからの出力など
目的	重複応募の排除、DMの送付、顧客分析など
納期・頻度	1回きりか、定期的に発生するか
取り扱い条件	個人情報の受け渡し方法や、作業場所の条件

委託先を選ぶときに確認したい観点は、次のページのチェックリストにまとめました。

委託先を選ぶ前に確認したい7つのポイント

各項目に、委託先への「聞き方」の例を添えました。実績値は、当社代表がこれまでに携わった案件のおおよその数値です（守秘義務のため案件名は伏せています）。

1 二重登録をどこまで見つけられるか

名前の書き方の違い（3ページのA・B・C）まで含めて、重複応募を見つけているか。

実績例：約50万件の申込から約1,500件の二重登録を見つけ、約150万円の予算流出を防止

▶ 聞き方の例：「表記ゆれのある重複は、どうやって見つけていますか？」

2 抽選・選定の結果に偏りがいないか確かめているか

当選者やDMの送付対象が、特定の地域・年代・申込方法に偏っていないかを確認しているか。

▶ 聞き方の例：「抽選や対象者の抽出のあと、結果の偏りをチェックしていますか？」

3 世帯の判定をどう運用しているか

「同じ世帯」の判定基準と、迷うケース（4ページのD・Eのような例）の扱いが決まっているか。

実績例：約1.2万件の同一世帯の候補を検出。プライバシーに配慮して自動ではまとめず、候補リストを作って1件ずつ確認する運用に

▶ 聞き方の例：「同じ世帯かどうか迷ったとき、どうしていますか？」

4 件数が増えても精度を保てるか

数百件のテストでうまくいっても、数十万件で同じ精度が出るとは限らない。テストデータの選び方が重要。

▶ 聞き方の例：「本番の件数で精度をどう確かめますか？」

5 処理の途中経過を説明できるか

各段階で何件を除外し、何件をまとめたかを記録していて、あとから説明できるか。

▶ 聞き方の例：「最初の件数から当選者数まで、途中の件数を出せますか？」

6 不正を見つけた経験があるか

組織的な不正応募や内部不正を見つけた経験と、その仕組みがあるか。

実績例：店頭シール応募キャンペーンでの不正取得や、従業員による大口の不正を発見

7 間違いが見つかったときの対応が決まっているか

誤りが見つかった場合に、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：「当選者に誤りがあったら、どう対応しますか？」

まとめ

キャンペーンの成果も、DMの費用も、顧客分析の正しさも、元になるデータの質で決まります。表記ゆれを直し、同じ人を見つけ、世帯を判定し、結果を確かめる——こうした地道なデータクレンジングの積み重ねが、データ活用の信頼を支えています。

名寄せには必ず「見落とし」と「まとめすぎ」のバランスがあり、スコアリングには目的に合った方法選びが欠かせません。委託先を選ぶときは、仕組みの見た目よりも、**判断の基準と途中の経過を、わかる言葉で説明できるかどうか**に目を向けることをおすすめします。

本資料の7つのチェックポイント

- ① 二重登録の検出 ② 抽選・選定結果の偏りの確認 ③ 世帯判定の運用
- ④ 件数が増えたときの精度 ⑤ 途中経過の記録 ⑥ 不正検知の経験 ⑦ 間違いへの対応

これらを満たすパートナーを選ぶことが、顧客にとって公平で、発注者にとって説明のつくデータ活用につながります。

データ処理の品質について、ご相談ください

合同会社オンエーカー（OnAK）は、データクレンジング・名寄せ・データ突合を専門とする会社です。行政・金融・流通分野の大規模なデータ処理に携わってきた経験をもとに、今のデータの点検から、名寄せ・世帯判定のルール設計、顧客スコアリング、不正の検知まで、データ活用の土台を支えます。

「今のやり方で大丈夫か確認したい」「クライアントへの提案に専門家の目を入れたい」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

Web : onak.co.jp / Mail : kano.toshiyuki@onak.co.jp