

英国通販の顧客別購入額と90日以内の再購入

要旨

英国の通販事業者の取引明細を用いて顧客別購入額の分布を集計した。対象は数量と単価が正で取消番号ではない明細だ。顧客番号のある4,338顧客の上位434顧客が購入額の61.4%を占めた。重複行の処理と最大顧客の除外後も集中は残った。最初に観測された購入から90日後まで記録を確認できる3,370顧客では、90日以内の再購入割合は47.5%だった。単一事業者の記述的な分析であり、一般市場の推定や施策の因果効果は対象としていない。

1. 目的と対象

購入額の集中が少数の大口顧客や重複明細だけで説明できるかを調べた。再購入については観測できる期間の差を抑えるため90日間の集計を併記した。使用した原表はUCIのOnline Retail。2010年12月1日08:26から2011年12月9日12:50までの541,909行を含む。単価の通貨は英ポンド。卸売顧客を含む英国の単一通販事業者の記録だ。

2. データの定義と処理

購入明細はQuantity > 0、UnitPrice > 0、InvoiceNoがCで始まらない行とした。530,104行・19,960受注が該当した。行の購入額は数量と単価の積とし、対象行の合計は10,666,684.544ポンドだった。小数桁は原表の計算値を記録しており、会計帳簿の精度を示す値ではない。

取消番号の行は9,288行。数量が負で単価が正の取消明細の金額の絶対値は896,812.49ポンドだった。取消元の受注との対応付けは実行していない。これを当期の返品率や当期販売分の取消率とは呼ばない。単価が正の全行を符号付きで足すと9,769,872.054ポンドになるが、これも会計上の売上とは断定しない。

顧客番号は135,080行で欠損し、商品説明は1,454行で欠損した。配布ページの欠損に関する表示とは一致しない。今回は実ファイルの検査を優先し、顧客番号を補完せず顧客別の集計から除外した。購入額全体の集計には顧客番号のない正の購入明細も残した。全列が一致する重複は5,268行あり、主分析では残した。同じ商品の正当な分割入力か重複記録かを確定できないためだ。

3. 集計方法

顧客別の購入額を降順に並べた。上位の顧客数は顧客数の10%を切り上げた整数とし、上位の顧客数の購入額合計を顧客番号がある購入額合計で割った。分母は8,911,407.904ポンドで、対象購入額全体の83.5%に相当する。

条件を変えた確認として、全列一致の重複を1行にした場合と最大購入額の顧客を除いた場合を再集計した。最大顧客の除外後は残った顧客数から上位の顧客数を計算し直した。これらは誤記を特定する検定ではない。

再購入は顧客番号と受注番号で受注をまとめ、各受注の最も早い日時を使った。顧客の最初の購入時刻より後で90日以内に別受注があれば再購入とした。同じ時刻の分割受注は再購入と数えない。原表の最終日時まで90日を確保できない顧客はこの集計から除外した。初回観測の前にも購

入した可能性があるため、新規獲得日とは扱わない。

月次の補助集計は2011年1月から11月までとした。2011年12月は9日途中で終了し、2010年12月と同じ長さではない。月次値は季節調整していない。

顧客数は異なるCustomerIDの数で、実人数を確認した数ではない。法人や卸売顧客を含む。90日集計の初回観測月別表では2011年9月が9月10日12時50分までの顧客に限られる。この月を全月の新規獲得数として使わない。

4. 結果

集計条件	上位約1割の購入額割合
主分析・4,338顧客中434顧客	61.4%
全列一致の重複を1行にする	61.5%
最大顧客を除外して再計算	60.2%

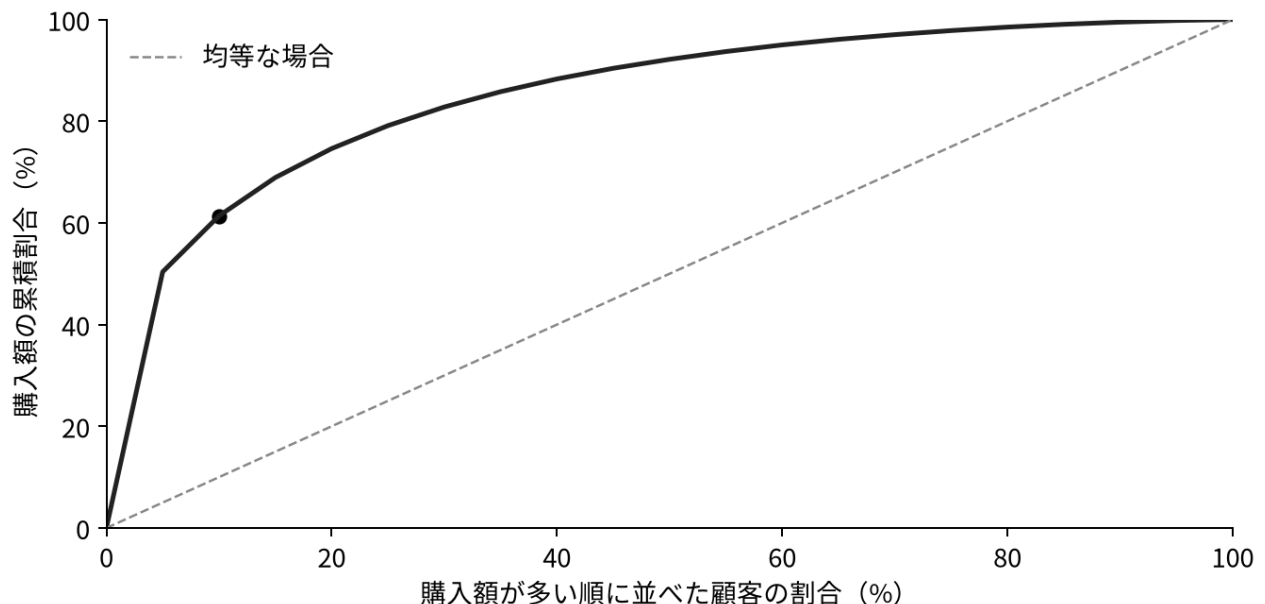


図1 顧客番号がある購入明細の累積購入額割合

顧客番号のある顧客のうち期間内に2受注以上がある割合は65.6%だった。90日を確保できる3,370顧客では1,601顧客が定義上の再購入に該当し、47.5%となった。対象顧客と観測期間が違うため両者の差を改善幅として扱わない。初回観測月別の顧客数と割合は補助表に保存した。

5. 検証と限界

原ZIPのSHA-256は取得時の記録と一致した。明細件数と列構成に想定外の差はなく、数量・単価は有限値だ。行ごとの乗算を通常のループでも足し、集計ライブラリの合計と照合した。月別金額の合計と同じ期間の明細合計も一致する。顧客番号がある明細では1受注番号に複数顧客が混在していないことも確認した。

対象は公開された一事業者の過去の記録で、現在の日本の通販市場を代表しない。無作為標本ではないため母集団への信頼区間や有意差検定を付けていない。顧客番号がない取引の分布は不明。

原価・税・配送費・利益・広告接触は未収録。購入額の集中を利益や顧客への働きかけの効果に置き換えない。匿名化済みの識別子も記事と集計表には掲載していない。

6. 再現資料

analyze.pyが原ZIPを直接読み、results.jsonとmonthly.csv、observed_cohorts_90d.csv、concentration.csv、図1を生成する。原表は変更しない。Python 3.12.14、pandas 2.2.3、NumPy 2.3.5で実行した。金額の照合許容誤差は相対誤差で1e-10、月次照合は1e-12とした。計算には乱数を使っていない。

原ZIPのSHA-256は f5385cbb54bbebf7196389109c6b0621faab0c304e3702548165e71c84aede8b。ダウンロード時点の原本を再利用した分析であり、配布ファイルの変更時には再検査が必要だ。

参考資料

Chen, D. (2015). Online Retail [Dataset]. UCI Machine Learning Repository.
<https://doi.org/10.24432/C5BW33>

配布元と変数定義 <https://archive.ics.uci.edu/dataset/352/online+retail>

利用条件 <https://creativecommons.org/licenses/by/4.0/>。CC BY 4.0に基づき集計・図表化した。査読を受けた論文ではない。分析日 2026年9月7日。