レビュー・評価と売上の関係

作成者：Shao Chuyan
日付：2025 年 10 月 2 日

1　使用ツール・言語

本分析では以下のツールとプログラミング言語を使用した：

- データ収集・スクレイピング：Python (Selenium, BeautifulSoup)
- データ整理・整形：Python (Pandas)
- データ可視化・統計分析：R
- レポート作成：LaTeX

2　解決したい課題

近年、オンラインショッピングの普及に伴い、価格だけでなくレビューや評価が購買行動に与える影響が注目されている。そこで、「レビュー数や評価が売上にどの程度寄与しているのか」を解明したい。

3　影響要因の仮説

レビュー数が多く、かつ評価が高い商品ほど、売上が増加するのではないか。

4　データの収集法

アメリカの Amazon の 6 つのカテゴリーの売れ筋ランキングページから、商品情報を収集した。具体的には、Python の Selenium を用いてブラウザを自動操作し、ページを順次スクロールして、動的に読み込まれるデータも取得可能とした。各商品ページにアクセスし、BeautifulSoup を用いて HTML を解析し、以下の情報を抽出した：

- 商品名（Product Name）
- ブランド名（Brand）
- 商品 ID（Product ID）
- 価格（Price）
- 評価（Rating）
- レビュー数（Number of Reviews）
- 親カテゴリ・サブカテゴリのランキング（Sales Rank in Parent/Sub Category）
- 発売日（Release Date）
- 原産国（Country of Origin）
- 過去 1 か月の購入数（Purchased in Last Month）

ページ内の表形式データや非表形式データの両方に対応するため、複数の HTML 要素を探索して情報を取得する処理を実装した。また、価格や購入数の表記には「k」「M」などの単位が含まれる場合があるため、数値に変換して統一した。

このようにして取得したデータは Python の Pandas を用いて整理・整形し、最終的に CSV ファイルとして出力した。データ量は各カテゴリで数百件程度であり、分析の対象として十分な規模である。

データは以下の Amazon 売れ筋ランキングページから取得しました：Amazon Best Sellers

5　データの整理・前処理

取得した Amazon の売れ筋ランキングデータは、いくつかの欠損値や形式の不統一が存在したため、分析前にデータを整備した。具体的な処理内容は以下の通りである：

1. **欠損値の除去** - 商品 ID（Product ID）が欠損している行は削除した。
2. **ランキング情報の補完** - 親カテゴリー・サブカテゴリーのランキング（Sales Rank）における欠損値を中央値で補完した。
3. **購入数・価格の補完** - 過去 1 か月の購入数（Purchased in Last Month）と価格（Price）の欠損値をそれぞれ親カテゴリごとの平均値で補完した。
4. **日付形式の統一** - 発売日（Release Date）は、文字列を日付形式に変換し、欠損値は前の行の値で補完した。

5. **カテゴリ・ブランドの欠損値補完** - 親カテゴリ・サブカテゴリ・ブランドの欠損値は「Unknown」で表記。

6. **不要列の削除** - 原産国（Country of Origin）は分析に使用しないため削除した。これにより、欠損値が最小化され、統一された形式で CSV として保存され、Python や Tableau での分析に適した状態となった。

6　データの分析法

分析の結果、以下の点が確認された：

- 評価点数と売上には正の関係が見られる。ただし、評価が最も高い商品が必ずしも最大の売上を示すわけではない。
- レビュー数が多い商品ほど売上も高い傾向がある。ただし、一定数を超えるとその効果は逓減する。
- カテゴリごとに影響の度合いが異なり、Beauty や Clothing ではレビュー効果が大きい一方、Automotive などではブランドや価格といった他要因の影響が強い可能性がある。

7　データ分析結果

7.1　記述的分析

データを見ると、商品の評価 (Rating) は主に 4.4–4.7 の範囲に集中しており、販売数は右に偏っていることが分かる。高販売数の商品は少数である。
Figure 1 に示すように、左側のヒートマップ（Rating vs Purchased in Last Month）から以下が確認できる：

- 評価が高い商品は販売数が比較的高い
- 評価が低い区間では販売数の変動が大きい

右側の散布図（Number of Reviews vs Purchased in Last Month）を見ると、レビュー数が多い商品ほど購入数も多い傾向があり、全体として、評価が高い商品は購入数が多いことが確認できる。

Figure 1: **左**: 評価 (Rating) と販売数の 2 次元ヒートマップ；**右**: レビュー数と購入数の関係（両対数スケール）

7.2 GAM モデルによる適合結果

本分析では、評価 (Rating) と購入数 (Purchased_Last_Month) の関係が非線形の可能性を考慮し、GAM（Generalized Additive Model）を用いた。GAM は線形モデルより柔軟に局所的な非線形関係を捉えられるため、わずかな非線形の影響も可視化・分析できる。特に、平滑曲線は「評価が異なる水準で販売数に与える効果の強弱」を可視化する役割を果たす。

以下のモデルを適合した：

$$\log(\text{Purchased\_Last\_Month}) \sim s(\text{Rating})$$

結果は以下の通りである：

- 平滑項 $s(\text{Rating})$ は有意 (F=3.337, p=0.00119)
- 自由度 $edf \approx 7$ で、Rating と販売数の関係には緩やかな非線形性が存在
- モデルの調整済み決定係数は約 5.8% (Adj R^2=0.0584) であり、評価は有意であるものの、販売数はカテゴリーを含む他の要因にも大きく依存すると考えられる。Figure 2
- 平滑曲線の形状から、低評価区間 (4.2 未満〜4.4 未満) では購入数の増加が急であり、中高評価区間 (4.5〜4.7) では緩やかな増加にとどまることが確認できる。平滑曲線の形状から、評価値が 4.0〜4.2 の区間では一時的に販売数が減少

する傾向が見られる。ただし、この低下はデータの分布や外れ値の影響を受けた可能性が高く、全体傾向としては評価が上がるにつれて販売数が増加することが確認できる。Figure 3

可視化された限界効果曲線においても、全体としては高評価区間で上昇傾向を示すが、増加の勾配は区間によって異なり、「限界効果の強弱」が存在することが確認できる。

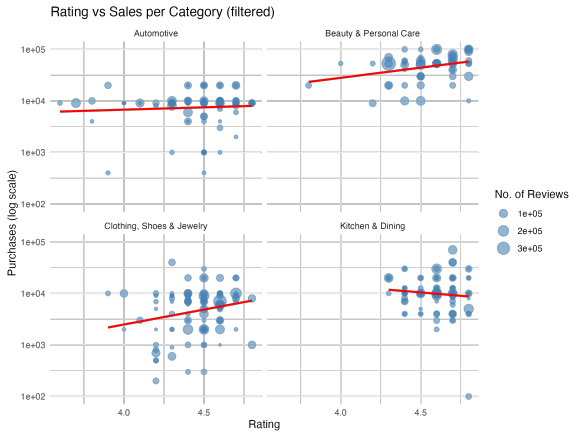

Figure 2: カテゴリ別の評価と購入数の関係（サンプル数 20 以上にフィルタ）

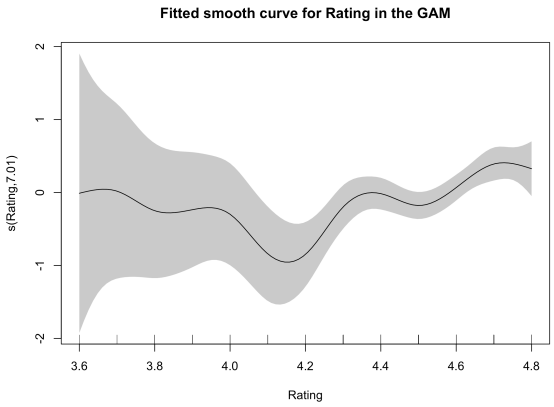

Figure 3: GAM 平滑項 $s(\text{Rating})$ に基づく平滑曲線

8　まとめ

本分析を通じて、レビュー数や評価は消費者の購入意欲に大きな影響を与える主要因であることが確認された。特に GAM モデルの適合結果から、評価と販売数の関係は単純な直線関係ではなく、区間ごとに限界効果の強弱が存在することが明らかになった。これは、単に「評価が高いほど売れる」という一般的な理解を超え、どの水準で評価の影響が強まるかを把握できるという点で実務的に重要である。

今後は、価格やプロモーション情報など他の要因も含めた多変量解析を行うことで、さらに精度の高い販売予測や戦略立案が可能になると考える。

- 評価は販売数に対して正の非線形影響を持ち、高評価ほど効果が大きい
- 平滑曲線から、4.0〜4.2 区間で一時的な減少が確認されたが、全体傾向としては評価が高いほど販売数は増加する
- 限界効果の強弱が確認でき、高評価区間 (4.5〜4.7) の向上は中低評価区間 (4.1〜4.3) よりも販売数への貢献が大きい

- 提案：
  - 高評価商品はマーケティングや在庫の優先配分を行うことで、より効率的に販売数を増加させられる
  - 中低評価商品は製品改善やサービス向上によって評価を引き上げることで、販売数を伸ばせる可能性が高い