あるホームセンターにおける客層データの統計的分析

2014SS046 三浦和也　2014SS102 穂積祥太

指導教員：松田眞一

1　はじめに

現在あるホームセンターでは，会員データの分析は行われていたが，会員でない客層の行動は把握できていなかった．その結果，どの客層に対してどのような種類の商品が売れているか，購入する時間や時期によっての顧客の買う商品や来店時間の傾向がわからない状態だった．

そのため，このホームセンターでは，会計時に店員によって顧客の性別及び年齢を打たせるように指示した．それにより，顧客の買った商品のレシートデータに，顧客の性別及び年齢を紐づけした新たなデータが作成され，今回の分析ではそのデータを用いている．

本研究では，そのデータを用いて，時間別に見た性別，年代，部門にどのような差があるかを明らかにする事が目的である．

2　データについて

データはあるホームセンターに関して，レジで入力されたレシートデータである．時期としては，2017 年 2 月から 2017 年 8 月までとし，全店舗における性別，年代，購入された商品についてのデータを用いる．商品についてのデータの内容として，商品の部門，購入数，価格，売上高等がある．

3　時間比較について

3.1　分析方法

時間毎の差を見るため，繰り返しのない二元配置の Tukey-Kramer 法を用いた．また，同方法を用いて時間毎に部門別の差の分析も行った．(Crawley[1] 参照)

今回全体的に一番来客数が多くなる時間である 10 時から 12 時の特徴を掴むために主成分分析とクラスター分析を行った．(中村 [2]，田中・脇本 [4] 参照)

3.2　データ加工

実際解析に使用したデータは 5 種類の店舗規模における 2017 年 3 月 1 日から同年 3 月 31 日，2017 年 8 月 1 日から同年 8 月 31 日を時間別の性別，年代，部門の来客数で集計したデータである．ただし，時間については店舗によって営業時間が異なるため，どの店舗でも共通して営業している 10 時から 20 時までを 2 時間毎に 5 分割したもの (10 時から 12 時，12 時から 14 時，…) としている．また今回使用した店舗規模以外にも 2 種類の店舗規模が存在するが，その 2 種類のレシートデータは他の種類の店規模に比べてデータが不足している，もしくはデータが偏っているため今回は除いて集計されている．部門の集計については 1 人が複数の部門を購入する場合もあるため，購入した品物数に関係なく購入した部門の種類で集計を行った．そのため全体の人数は性別，年代と異なり，合計人数は多くなっている．

また，主成分分析とクラスター分析に利用したデータは 10 時から 12 時に来客した人数を 3 月の平日，3 月の休日，8 月の平日，8 月の休日の計 4 つに分け，それをさらに各部門毎に分けたものを使用した．

3.3　時間比較の結果と考察

以下に示す結果は各規模の平均を時間別に比較した時，帰無仮説が成立する p 値である．ただし，実際の検定では 3 月及び 8 月を平日，休日の男性客，女性客に分けた比較を 5 つの規模別，部門別に行っているが，紙面の都合上 5 つの規模の女性客についての結果のみ示す．

表 1　3 月の平日女性の時間比較

	12-14 時	14-16 時	16-18 時	18-20 時
10-12 時	0.0403	0.5054	0.0001	0.0000
12-14 時		0.5450	0.0961	0.0000
14-16 時			0.0046	0.0000
16-18 時				0.0008

表 2　3 月の休日女性の時間比較

	12-14 時	14-16 時	16-18 時	18-20 時
10-12 時	0.3266	0.8383	0.0260	0.0000
12-14 時		0.0573	0.6131	0.0000
14-16 時			0.0033	0.0000
16-18 時				0.0001

表 3　8 月の平日女性の時間比較

	12-14 時	14-16 時	16-18 時	18-20 時
10-12 時	0.0024	0.3711	0.0020	0.0000
12-14 時		0.0933	0.9999	0.0002
14-16 時			0.0784	0.0000
16-18 時				0.0002

表 4　8 月の休日女性の時間比較

	12-14 時	14-16 時	16-18 時	18-20 時
10-12 時	0.0036	0.7976	0.0103	0.0000
12-14 時		0.0333	0.9840	0.0001
14-16 時			0.0900	0.0000
16-18 時				0.0000

共通している事として，18 時から 20 時との比較では単

純な来客数の差が大きいため，棄却されている．来客数に差ができたのは主婦層がいることが考えられる．そのため遅い時間程来客数が減少している．

平日と休日で異なる点として，休日では大型用品や玩具と言った家族での買い物と考えられる商品が棄却されている．また，それらの商品は 14 時から 16 時の時間帯に多い傾向がある．家族での買い物の場合，車の移動が多い．車を出しての移動は午前中より午後の方が多いため，午前でなく午後の特徴として現れたと考えられる．

平日では 10 時から 12 時の午前中に来客数が一番多くなるが，来客数が減っている時間帯である 16 時から 18 時の健康器具に差がない．健康を気にし始める年代は 40 から 60 才の年代であり，働いている年代でもあるため遅い時間になっても来客数の減少は緩やか．そのため差がないと考えられる．

園芸用品に注目するとどちらの月でも棄却はされているが，3 月では 10 時から 12 時の時間帯に特に差ができている．対して 8 月では園芸用品以外にも趣味に関した商品や小物が差を作っている．園芸用品を購入すると考えられる年齢層の比率が 3 月と 8 月で異なるためこのような差を作ったと考えられる．

3.4　主成分分析の結果と考察

分析の結果，寄与率が第 1 主成分で 0.9059，第 2 主成分で 0.0792 と第 2 主成分までで累積寄与率がほぼ 100% まで到達するため，第 2 主成分までを説明する．

第 1 主成分は，3 月か 8 月のどちらの方が来客数が多いかを表す．正に行く程 8 月の来客数が多い部門になり，負に行くほど 3 月の来客数が多い部門となる．

第 2 主成分は，平日か休日のどちらの方が来客数が多いかを表す．正に行くほど平日の来客数が多い部門になり，負に行くほど休日の来客数が多い部門となる．

時間比較では午前中に最も影響の強いと考えられていた園芸の部門だが，植物については 3 月に売れ易いという特徴になっている．またスポーツやレジャーと言った用品については他の部門に比べて 8 月に寄っているため，夏の時期に売れ易いと考えられる．

文具や塗料，園芸などに使用する用品などの業務に使うと考えられる部門については平日に寄っている結果となっている．

逆に休日に寄ると植物やカー用品などの趣味に関した部門が売れているように見える．

中心に近い部門は季節や平日か休日の関係がなく売れている商品と考える事ができる．

3.5　クラスター分析結果と考察

得られたデンドログラムから 5 つの群に分けて考察を行う．

第 1 群 (12，13 部門)

特に 8 月に売れ易い部門.

第 2 群 (4 から 25 部門)

3 月の休日に売れ易い部門.

第 3 群 (16 から 26 部門)

3 月の平日に売れ易い部門.

第 4 群 (22 から 21 部門)

8 月の平日に売れ易い部門.

第 5 群 (11 から 20 部門)

8 月か 3 月や平日か休日に関係なく売れている部門.

以下はクラスター分析を行った結果のデンドログラムである．

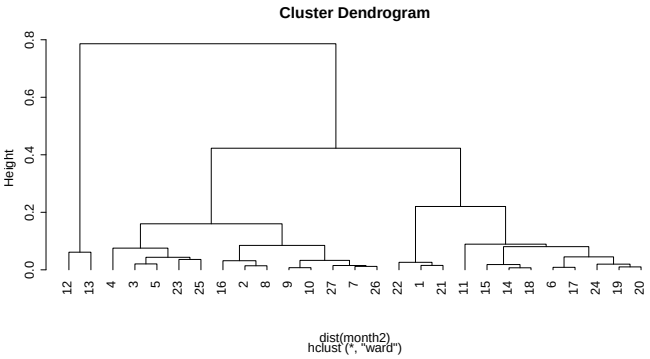

図 1　10 時から 12 時の部門別分析結果

3.6　まとめ

結果的に平日や休日，また月に関係なく 10 時から 12 時の時間に一番来客数が多い傾向にある．また 10 時から 12 時については 8 月の方がスポーツやサイクル・レジャーと言った用品が売れ，逆に 3 月の方が植物が売れる傾向にある．休日ではカー用品や植物と言った趣味に関した用品が売れ，平日では文具や園芸に使用する用品，塗料・補修と言った業務や作業などに使用する用品が売れる傾向がある．ただしグループで見ると園芸に関した用品については 8 月寄りになっている．

12 時から 14 時と 14 時から 16 時は女性では基本的に差ができる事が少なく，できる場合でも休日で 40 から 60 才の割合が高い時である．

16 時以降の女性については主婦層の関係で他の時間との差は大きくなっている．

4　商品の部門別での店舗の分析

4.1　データ加工

データはあるホームセンターに関して，全ての店舗で 2 月 6 日から一週間の期間に購入された品物についてのデータを使用する．それぞれの品物には各商品の分類として部門のうちの 1 つが設定されており，それらのデータを顧客別に集計したものを使用している．そのためデータのレコード 1 単位を 1 顧客とし，各部門の購入数を変数としている．

4.2　分析手法

顧客の分析についてはクラスター分析の内，非階層クラスター分析である k-means 法を用いて研究を行った．k-means 法は，非階層クラスター分析のうち，最もよく使われる手法である．主な特徴として，初期値により計算結果が異なる事が挙げられる．(R サポーターズ [3] 参照)

全データの内から 2 万のデータをランダムに取り出したデータセットを用いて，それで分析を行った．また，今回，非階層クラスター分析を行うにあたって，クラスター数の指定のため gap 統計量を用いている．本研究では R パッケージ cluster 内の clusGap 関数での結果 ¥result で出てきた結果を利用する．結果 ¥result には，設定すべき適切なクラスター数が含まれている．

店舗の分析についてはクラスター分析の内，階層クラスター分析である Ward 法を用いて研究を行った．Ward 法は階層クラスター分析のうち，実用的に優れた方法としてよく利用されている．(田中・脇本 [4] 参照)

4.3　分析手順

各店舗の分析を行う上において，各顧客の分析を購入する部門について行った上で，その分析結果を利用して店舗の分析を行う．本研究では，変数を 14，15，16，17，19 部門の 5 個の部門に絞って分析を行った，14 部門がペット用品，15 部門が消耗品，16 が文房具，17 がキッチン用品，19 が雑貨であり購入する客層が基本的に主婦層に限定できる 5 つの部門であった．

手順としては，顧客に対して k-means 法で分析を行い，それらで分類されたクラスターの顧客がどれだけ各店舗に含まれているかを変数として，店舗に対して Ward 法で分析を行った．本研究では，その分析を 2 回行っている．1 回目の分析では，上記部門を買っていない顧客は全ての変数の値が 0 として分析対象に入っている．2 回目の分析では，上記部門を買っていない顧客は分析対象から外されていて，1 回目の分析とは分析対象が明確に違う．

4.4　全ての部門を含む消耗品部門の分析

初めに，適切なクラスター数を計算する．関数 clusGap により，クラスター数は 6 として分析を行った．

14 部門を 7 個以上買ったクラスター，14 部門を 3-6 個購入した顧客のうち，15 部門よりも 14 部門を多く買った顧客のクラスター，14 部門を 1-2 個購入した顧客のうち，15，16 部門を 3 つ以上購入していない顧客のクラスター，16 部門が中心のクラスター，15 部門が中心のクラスター，17 及び 19 部門を中心に分類されたクラスターの 6 つのクラスターとなった。1 つ目から順に第一，第二，第三，第四，第五，第六クラスターとする．

17，19 部門については，基本的に各点で色々なクラスターに分類された点が混ざっている．これは，14，15，16 部門が 17，19 部門よりも影響が強く，それらが中心に分類された結果であると考える．

第六クラスターは 17，19 部門を中心に分類されたクラスターであると分かる．しかし，他のクラスターは 17，19 部門を中心に分類されていないため，他の各クラスターと比べて分析ではっきりとした結果は出ていない．また，全ての変数が 0 である点は第六クラスターに分類されている．全ての変数が 0 であるという事は今回選んだ消耗品を買っていない，すなわち専門性の高い商品を買っている顧客ということであり，第六クラスターにはそういった顧客も含まれている．

また，今回行った顧客の分析の結果を利用し，店舗の分析を行った．各店舗の顧客について，顧客の分析によって分類された 6 クラスターの割合を各変数とし，店舗に対してクラスター分析を行った．結果は図 2 のようになった．本要旨では概形について解釈を記す．

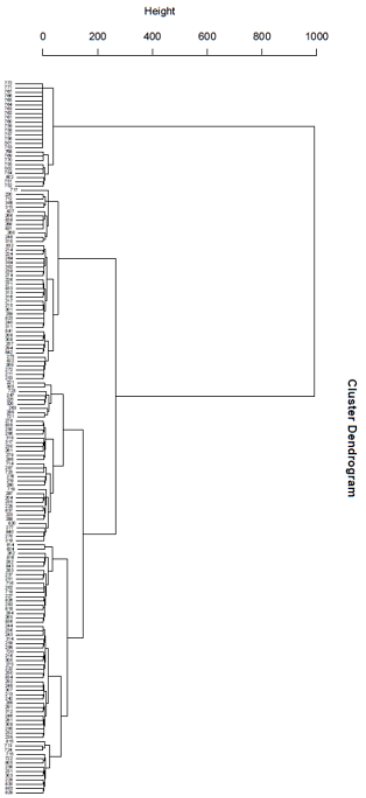

図 2　全ての部門を含む分析における店舗分析の結果

分析結果として，第六クラスターの割合が一番分類に影響し，次に第三クラスターの割合が影響した．第六クラスターは専門品を買った顧客であり，第三クラスターはペットフードを少量買った顧客であるため，店舗はそれらの客の割合を基準とした分類ができた．

4.5　全ての部門を含まない消耗品部門の分析

今節の分析では，今回用いる 5 部門について全ての要素が 0 であるデータを消去してある．

次に適切なクラスター数を計算する．関数 clusGap により，クラスター数は 5 として分析を行った．

顧客に対するクラスター分析の考察を行う．14 部門を

7-10 個以上購入したクラスター，14 部門を 3-6 個購入したクラスター，15 部門を購入している顧客のクラスター，17 部門を購入している顧客のクラスター，14 部門を 1-2 個購入した人のクラスターの 5 つのクラスターに分けられた．それぞれ第一，第二，第三，第四，第五クラスターとする．16 部門を購入している人はやや第三クラスターに，19 部門を購入している人はやや第四クラスターに分類されているようにも見えるが，第三クラスターと第四クラスターどちらのクラスターも存在するものが多く，15，17 部門ほどはっきりと分類されているとは言えない．

前の節と同様に，今回の顧客のクラスター分析の結果を用いて，店舗のクラスター分析を行った．

結果は図 3 のようになった．なお，分析結果からは分析に悪い影響を与えた 500，700 番代の店舗は削除されている．また，全ての部門を含む消耗品部門の分析と同様に，本要旨では概形について解釈を記す．

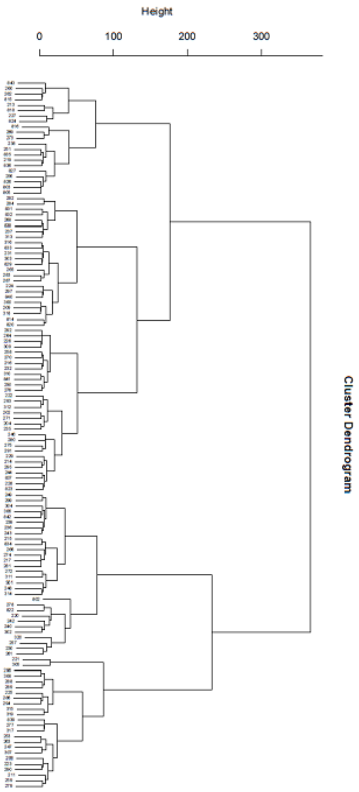

図 3　500，700 番代の店舗を除いた全ての部門を含まない分析における店舗分析の結果

5 つのクラスターに分けて分類を行った．左から順にクラスター番号を第 1，第 2，第 3，第 4，第 5 とする．第 1，第 2，第 3 クラスターと第 4，第 5 クラスターを分けた要因としては，第三クラスターの割合の差が一番大きな要因であった．

次に第 1，第 2，第 3 クラスター内のそれぞれの差として第四，第五クラスターの差が挙げられる．また第 4 クラスターと第 5 クラスターの差は，第五クラスターによって生じていると考えられる．それぞれ，値の範囲クラスターが大きく異なっていた．第三クラスターは 15 部門，第四クラスターは 17 部門，第五クラスターは 14 部門を 1-2 個購入した顧客であるため，店舗はそれらの客の割合を基準とした分類ができた．

4.6　まとめ

今回顧客の分析を行ったが，初めは部門を 5 つに絞らない分析を行い，解釈が出来なかった．その為，部門を 5 種類に絞るなど限定的な分析を行い，解釈が可能な結果が出た．部門を絞らずに行った分析が成功しなかった原因として，データの特徴に焦点を当てると，次のような事が分析の課程で分かった．

1. 商品の特性 (単価，用途，頻度)
2. 顧客の特性 (大量購入，業務用と一般用)

両方の特性が近いデータに限定すれば適切な分析が行われる事がわかり，一部の特性に焦点を当てた前処理では，処理されなかった特性が原因で分析は適切には行われない事がわかった．

今回行えた分析の結果として，両分析において少量ペットフードを買った顧客が店舗の分類に影響した．しかし、全ての部門を含まない分析では顧客の専門性が重視された店舗の分類が行われたのに対し、全ての部門を含まない分析では 15 部門の日用品を買った顧客による分類であったり、17 部門のキッチン用品を買った顧客による分類が行われた．

4.7　今後について

今回は，2 種類のデータに対して，顧客の分類の後，店舗の分類をクラスター分析で行った発展としては，消耗品以外の部門，例えばカー用品であったり，園芸用品であったりにも同様の分析をし，各分析の結果の比較をする事により，新たな発見があると考える．

5　おわりに

今回，当研究室では数年ぶりにあるホームセンターとの共同研究を行った．今回の分析では，何を分析すべきか，どういった分析が必要なのかといったところから始まり，難題でありながらも非常に有意義な時間を過ごせたように感じる．

参考文献

[1] Crawley, M. J. (野間口謙太郎・菊池泰樹 訳) :『統計学 : R を用いた入門書』. 共立出版，2008.

[2] 中村永友 :『R で学ぶデータサイエンス 2　多次元データ解析法』. 共立出版，2009.

[3] R サポーターズ:『パーフェクト R』. 技術評論社, 2017.

[4] 田中豊・脇本和昌 :『多変量解析法』. 株式会社現代数学社，1983.