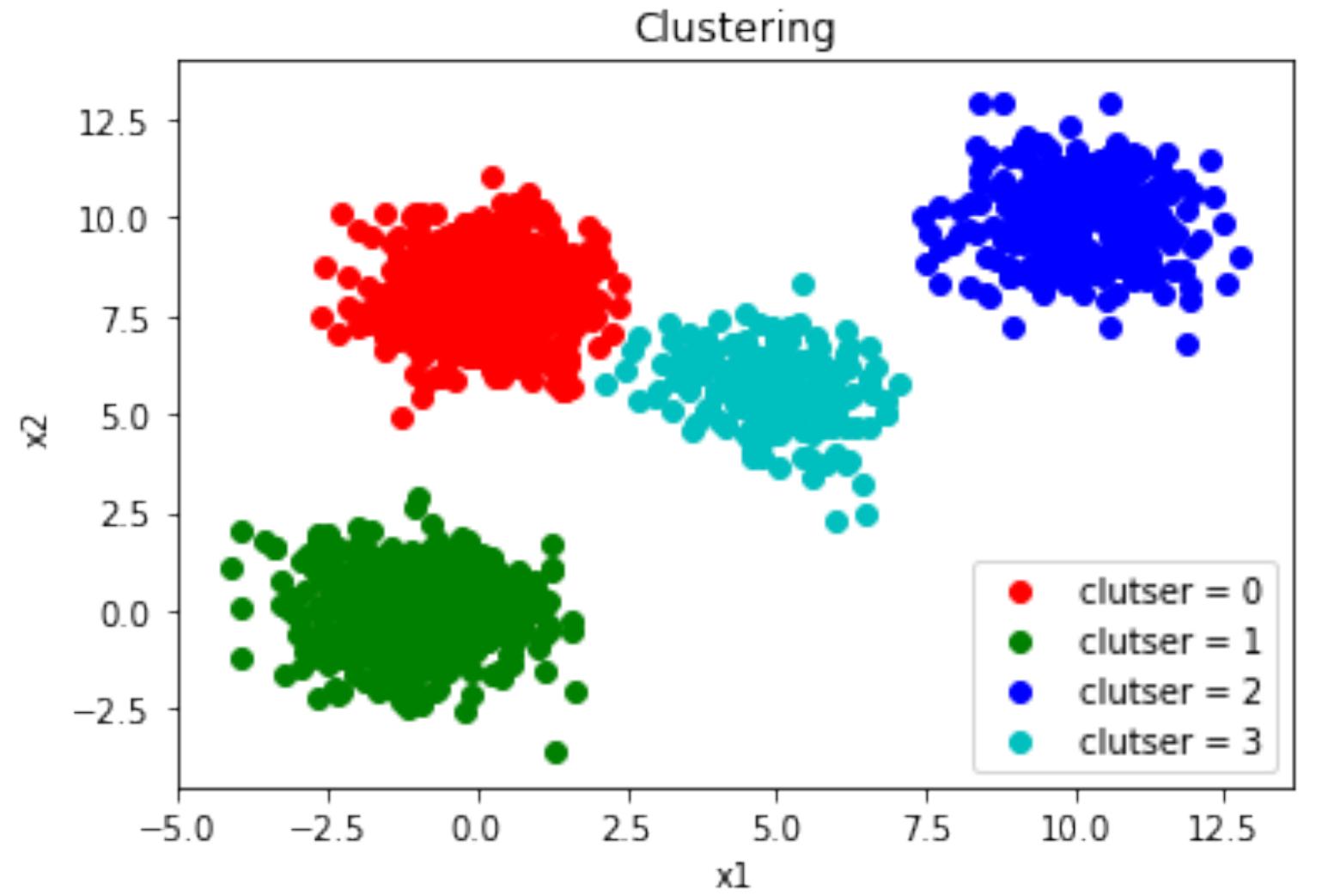


インテグレーションステップ

中原 2017/12/23

1. 課題



顧客のクラスタリング

2. なぜその課題を選んだのか

現在、旅行会社のオンライン販売部門で
メールマーケティングを担当。

メールマーケティングの状況は
業務量は「①メルマガ > ②ターゲティングメール」だが
販売額は「①メルマガ < ②ターゲティングメール」である

	①メルマガ	②ターゲティングメール
配信対象	顧客全体	特定顧客
配信回数	週3回	毎日
業務量(入稿)	月に12回(3回×4週)	月に2回(1回×2種類)
販売額	47%	53%

→閲覧履歴や購入履歴を元にしたオートメーションメールを増やし効率よく販売を増やしたい
→そのためには顧客像や行動パターンが明確になっていた方が打ち手を考えやすいため

例：出張利用者は宿泊翌月までにリピートしないと2度と利用しないと分析
→宿泊翌日に送るお帰りなさいメールで、
翌月泊まってCPNを訴求したところ、リピータの販売額が○%増加した

そこで、今回学んだ機械学習を活かして、
今まで気がついていなかった顧客像を見つだし、
今までにないビジネスインパクトが出したいと考えたため

3. データの収集方法



Personalize Expedia Hotel Searches - ICDM 2013

Learning to rank hotels to maximize purchases
\$25,000 · 337 teams · 4 years ago

→ KaggleのExpediaのデータセットを使用
(検索結果最適化のコンペに使用するデータ)

■ データ内容 (Rows: 6,622,629、Columns: 54)

<基本情報>

srch_id
date_time
site_id
visitor_location_country_id
visitor_hist_starrating
visitor_hist_adr_usd

site_id

Home Vacation Packages Hotels Cars Flights Cruises Things to Do DEALS

PLAN YOUR TRIP ON EXPEDIA

Flight Hotel Flight + Hotel Flight + Car Flight + Hotel + Car Hotel + Car

CHOOSE FROM MORE THAN 140,000 HOTELS WORLDWIDE

Hotel

Find hotels near:
A city, airport or attraction

What City?
New York (and vicinity), New York, United States or America

Check-in: 10/18/2013 Check-out: 10/20/2013 Rooms: 1

srch_destination_id srch_room_count

srch_booking_window srch_length_of_stay

srch_adults_count srch_children_count

BEST PRICE GUARANTEE

SEARCH FOR HOTELS

<検索条件>

srch_destination_id
srch_length_of_stay
srch_booking_window
srch_adults_count
srch_children_count
srch_room_count
srch_saturday_night_bool
src_query_affinity_score
orig_destination_distance
random_bool

<検索結果(施設情報)>

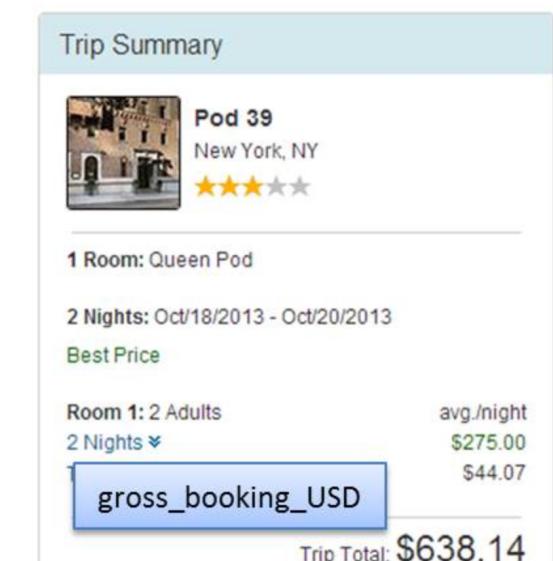
prop_country_id
prop_id
prop_starrating
prop_review_score
prop_brand_bool
prop_location_score1
prop_location_score2
prop_log_historical_price
position
price_usd
promotion_flag

<検索結果(競合状況)>

comp1_rate
comp1_inv
comp1_rate_percent_diff
comp2_rate
comp2_inv
comp2_rate_percent_diff
comp3_rate
comp3_inv
comp3_rate_percent_diff
~
comp8_rate
comp8_inv
comp8_rate_percent_diff

<行動履歴(クリックと購入)>

click_bool
gross_bookings_use
booking_bool



4. 使用したデータについて

■ 使用したデータ (Rows: 118,762、Columns: 27)

□ Rows

→ 「データ構造」から、検索IDをユニークにする必要があると判断

→ 購入がない検索IDは存在せず、かつ検索IDにつき購入は1回のみだったため、購入有データのみ使用

→ 販売ボリュームからサイトID(Expedia.com)、居住国ID(アメリカ)、宿泊国ID(アメリカ)に絞った

<データ構造>

	srch_id	date_time	site_id	visitor_location_country_id	visitor_hist_adr_usd	prop_country_id	prop_id	prop_starring	prop_review_score	...	comp6_rate_percent_diff	co_diff	click_bool	gross_bookings_usd	booking_bool
0	1	2013-04-04 08:32:15	12	187	NaN	219	893	3	検索結果1(prop_idが異なる)	...	NaN	False	NaN	False	
1	1	2013-04-04 08:32:15	12	187	NaN	219	10404	4	検索結果2(prop_idが異なる)	...	NaN	False	NaN	False	
2	1	2013-04-04 08:32:15	12	187	NaN	219	21315	3	検索結果3(prop_idが異なる)	...	NaN	False	NaN	False	

□ Columns

→ ほとんど欠損値の変数は除外。クラスタリングで重要度が低いと思われる変数は一旦、除外

<基本情報>

- ・その他：居住地と宿泊地の距離、居住国(id)、サイト(id)、競合サイトの価格、評価

<検索履歴>

- ・検索条件：大人人数、子供人数、部屋数、泊数、検索日時、リードタイム
- ・検索結果：クリックしたホテルの価格、評価、地域(id)、国(id)

<購入履歴>

- ・検索結果のホテルの購入有無、合計代金
- ・過去購入したホテルの平均価格、平均評価

5. 初期仮説・分析アプローチ

検索日時があるのでビジネスとレジャーはクラスタリングできそう。
レジャーのクラスタリングがポイント。

- ① 基礎分析
- ② 欠損値対応&変数作成
- ③ **Kmeans++** でクラスタリングし、エルボー分析、シルエット分析 で変数選択(加減)
- ④ **決定木、ランダムフォレスト** で、
クラスタリングに使用した変数と重要度を確認、各クラスタごとの特徴(分散状況)を把握
- ⑤ クラスタの特徴が明確になりそうな変数作成
- ⑥ ②～⑤を繰り返し
- ⑦ 分けたクラスタの特徴からペルソナ設定
- ⑧ 打ち手の検討と、ABテスト設計

6. 工夫したこと

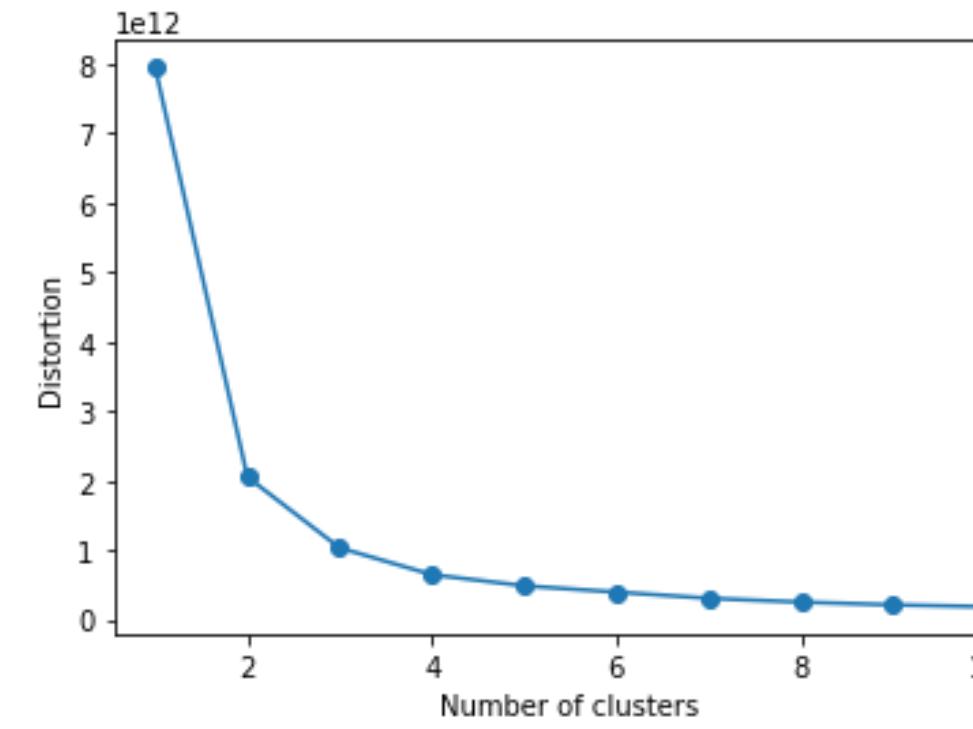
①変数作成

- ・検索日時から、検索曜日、検索時間帯、検索月、検索日を作成
 - ・検索日時とリードタイムから、宿泊曜日、宿泊月、宿泊日を作成
 - ・子供人数と大人人数から、ファミリーフラグ、ビジネスフラグを作成
- ※クラスタリングに役に立たなくても、特徴把握で役に立つものも多かった
※ちなみに他にも大量に変数作成したがほとんど失敗
(ハイシーズンフラグ、人気地域ランク、価格UP率、当日予約・前日予約フラグ…)

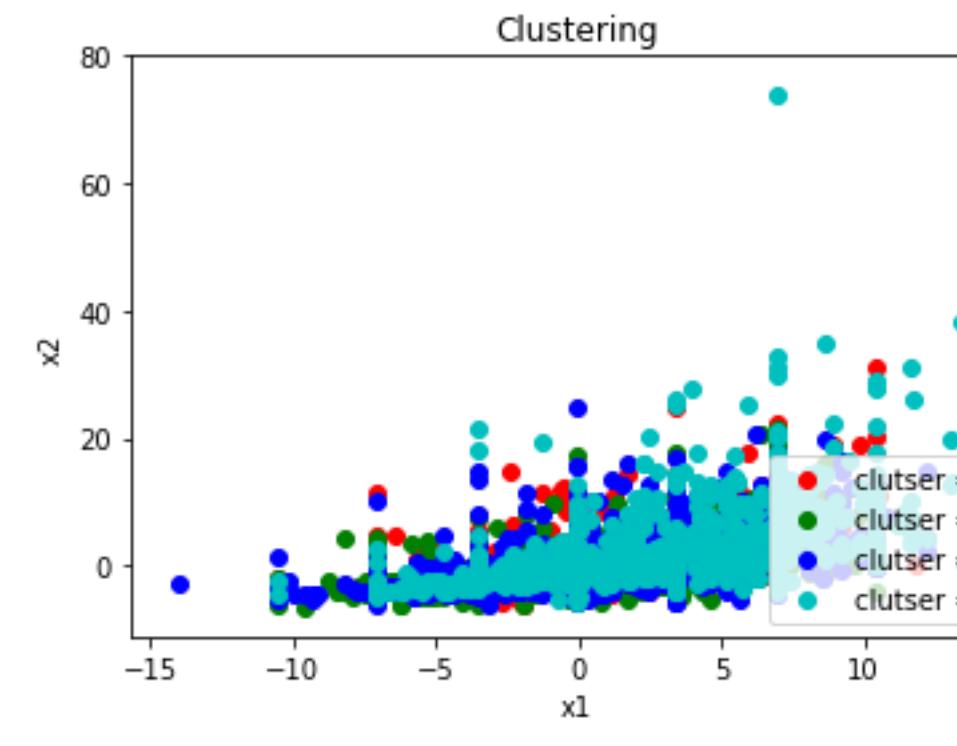
②クラスタリングの評価

「うまく分かれているか」 「クラスタのボリュームの偏り」 がわかりやすい
シルエット分析を使用したいが、計算コストが高さが課題だった(計算中にフリーズしてしまう)
データをランダムサンプリングする仕様にすることで解決

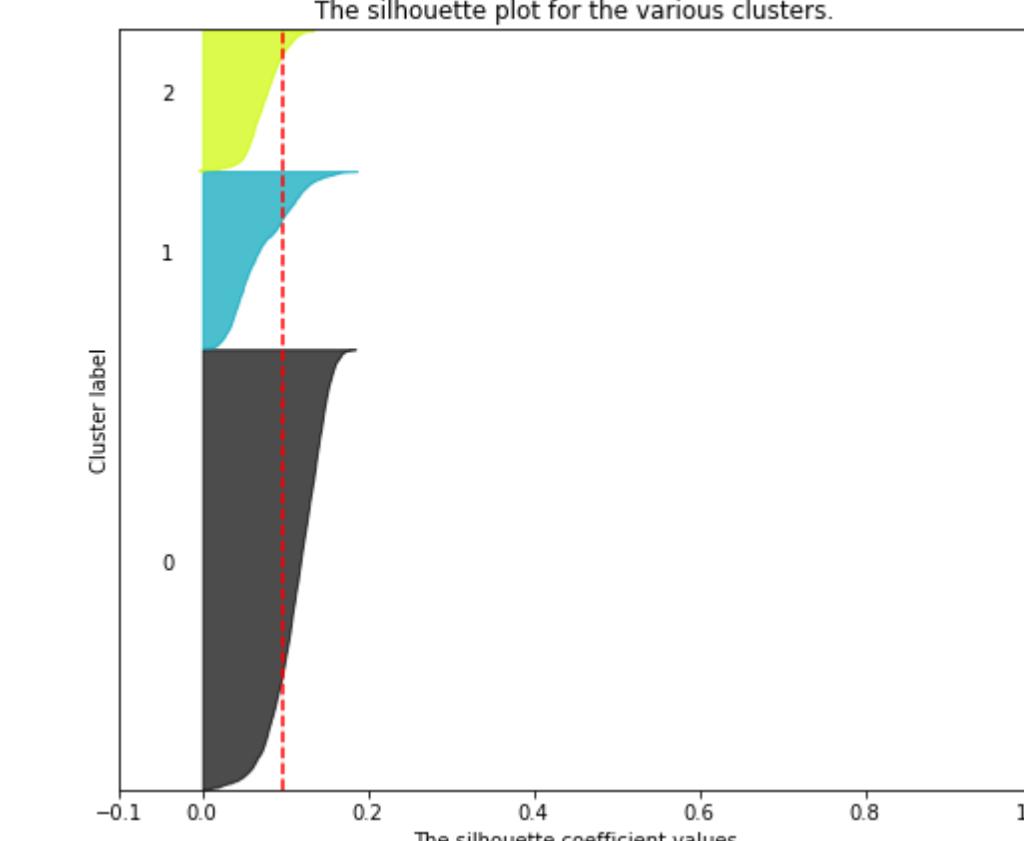
エルボー分析



散布図を色分け



シルエット分析



7. 難しかったこと

- ・解釈性があるデータが少なく、
結果から打ち手のイメージが膨らみにくかった
例：ボリュームからサイト、国を判断。距離から地域を判断
- ・連續値をフラグにした方がいいケース（人数）と
そのままにした方がいいケース（価格、距離、泊数）があり、
試行錯誤に時間を要した

8. 分析結果・考察①

エルボー分析

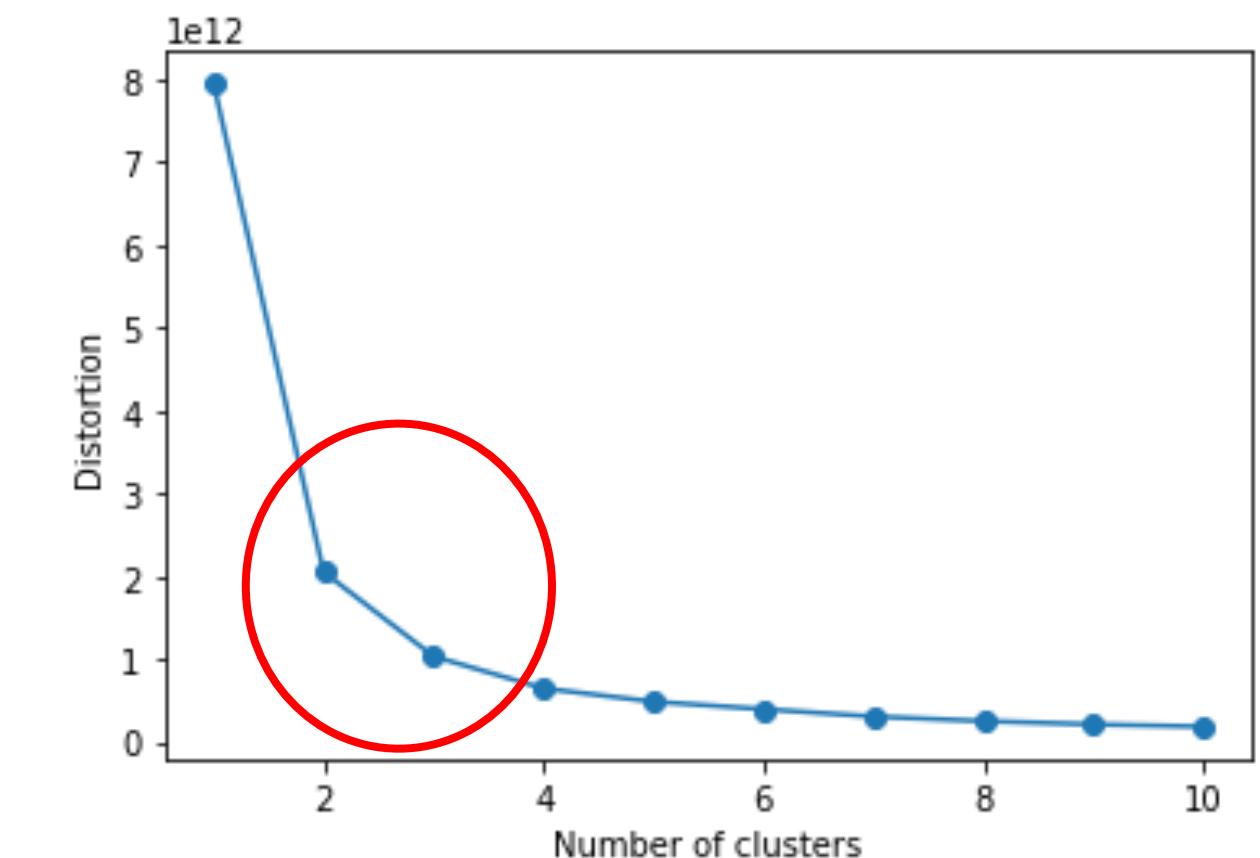
クラスタ数は、2か3が良さそう

シルエット分析

念のため、クラスタ数2～5で試し、最終的には4で決定

※2、3にはすぐに分けられたが、クラスタのボリュームに偏りがあったため、
4めのクラスタリングにチャレンジしたが、なかなかうまく行かなかったが、
試行錯誤を繰り返し、なんとか綺麗にクラスタリングに成功

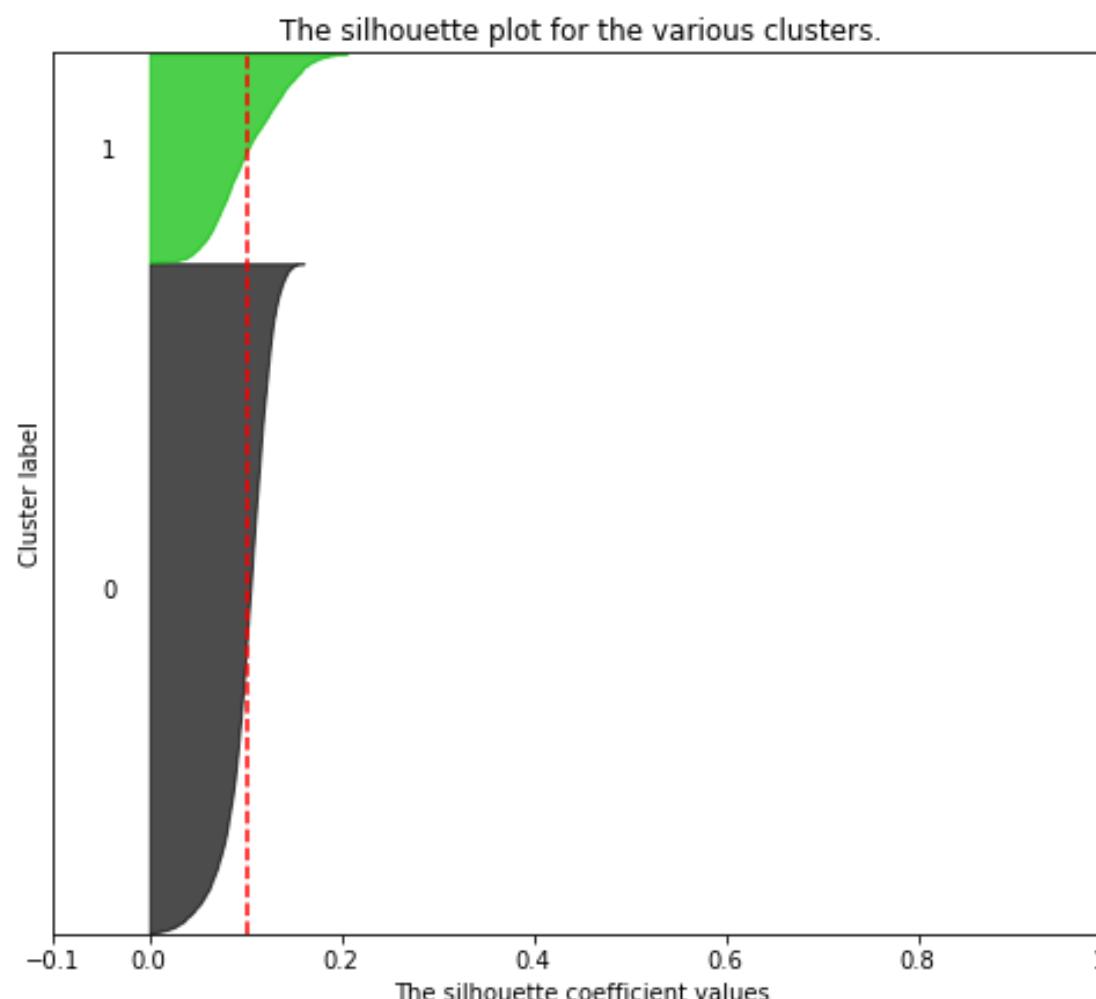
<エルボー分析>



<エルボー分析>

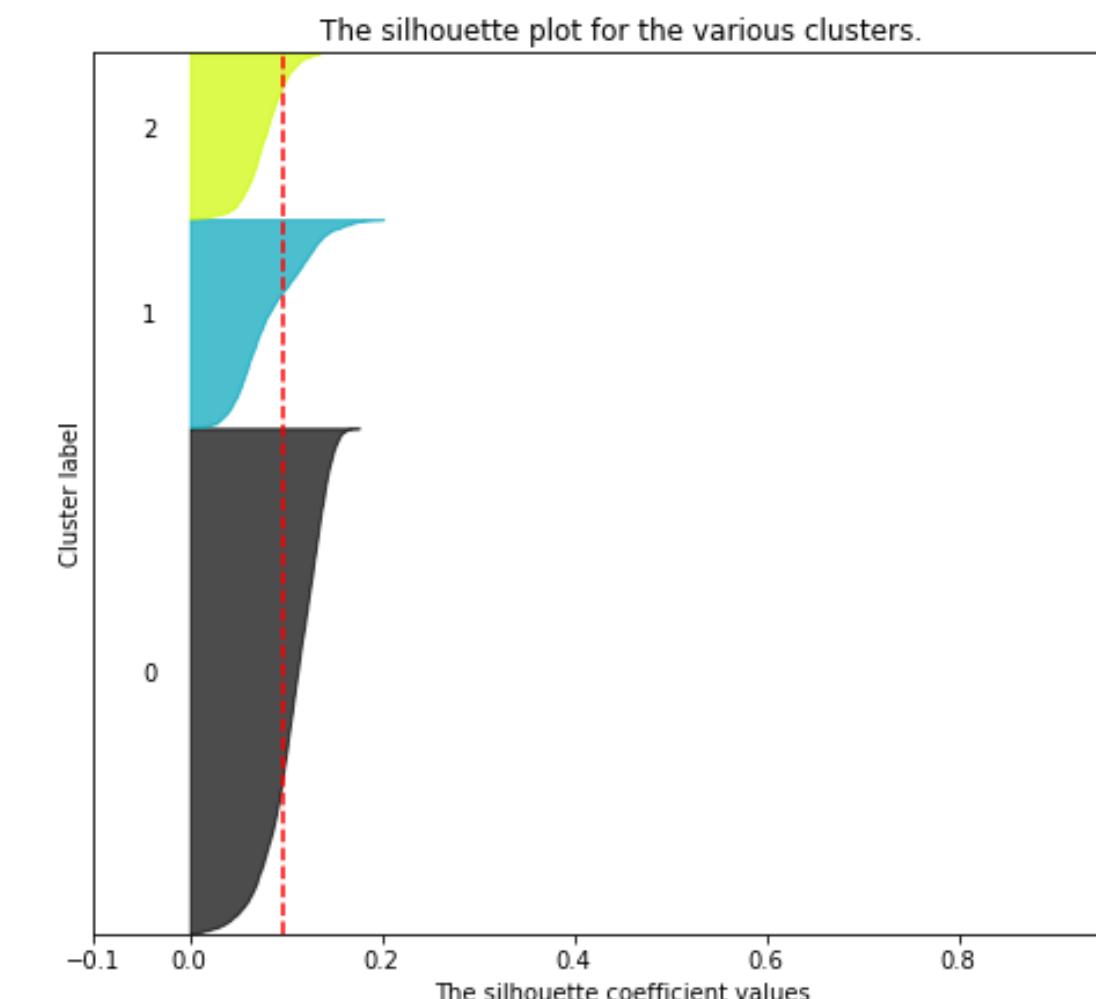
Cluster数 : 2

平均シルエットスコア : 0.10149



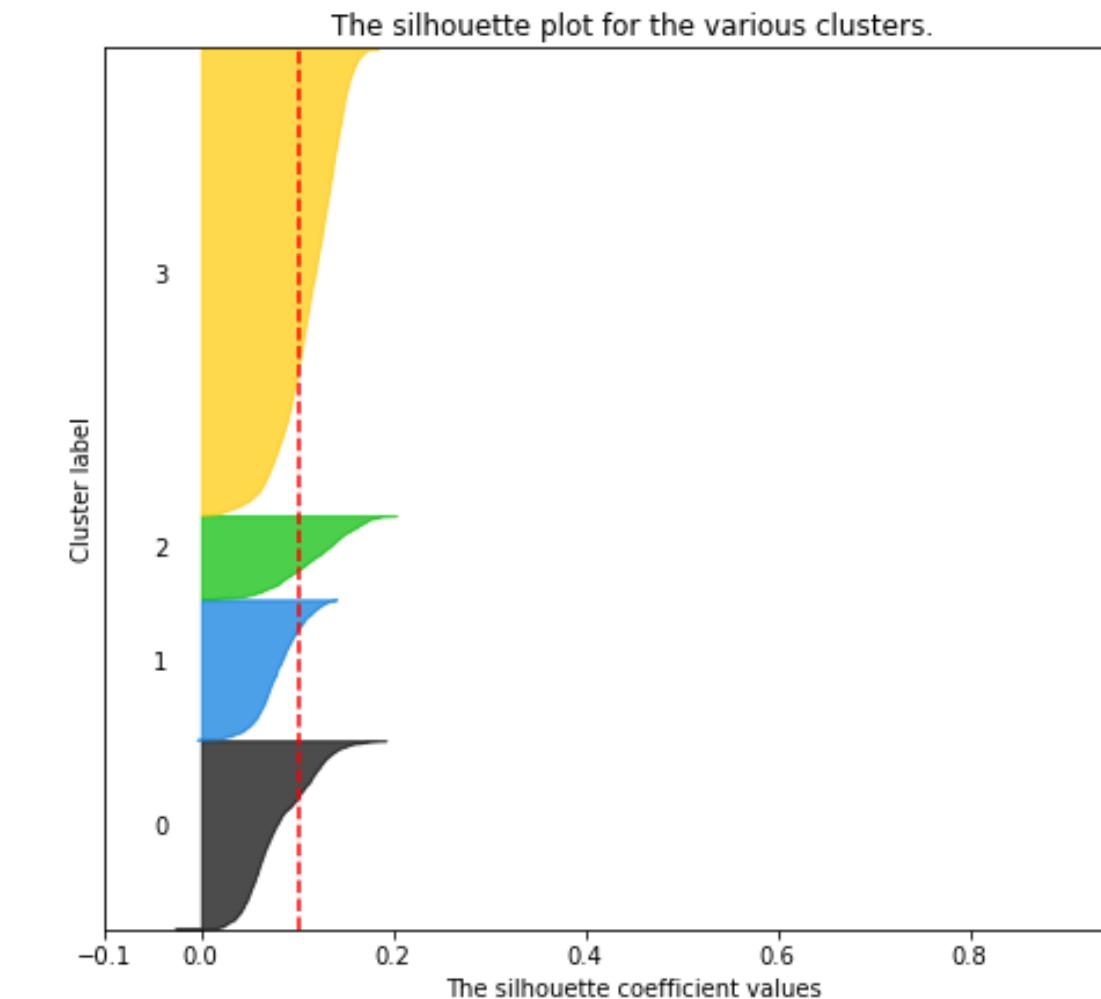
Cluster数 : 3

平均シルエットスコア : 0.09681



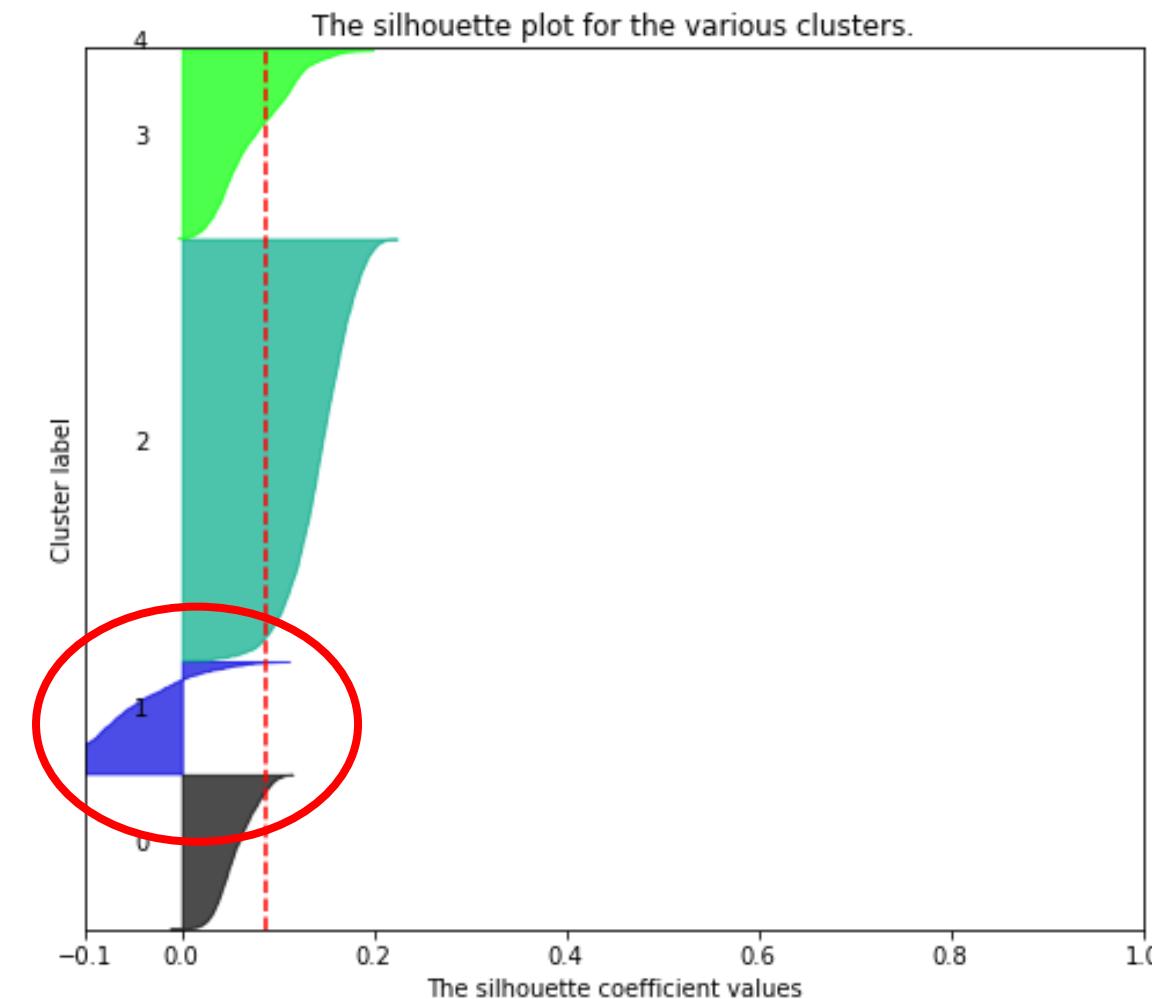
Cluster数 : 4

平均シルエットスコア : 0.10163



Cluster数 : 5

平均シルエットスコア : 0.08709

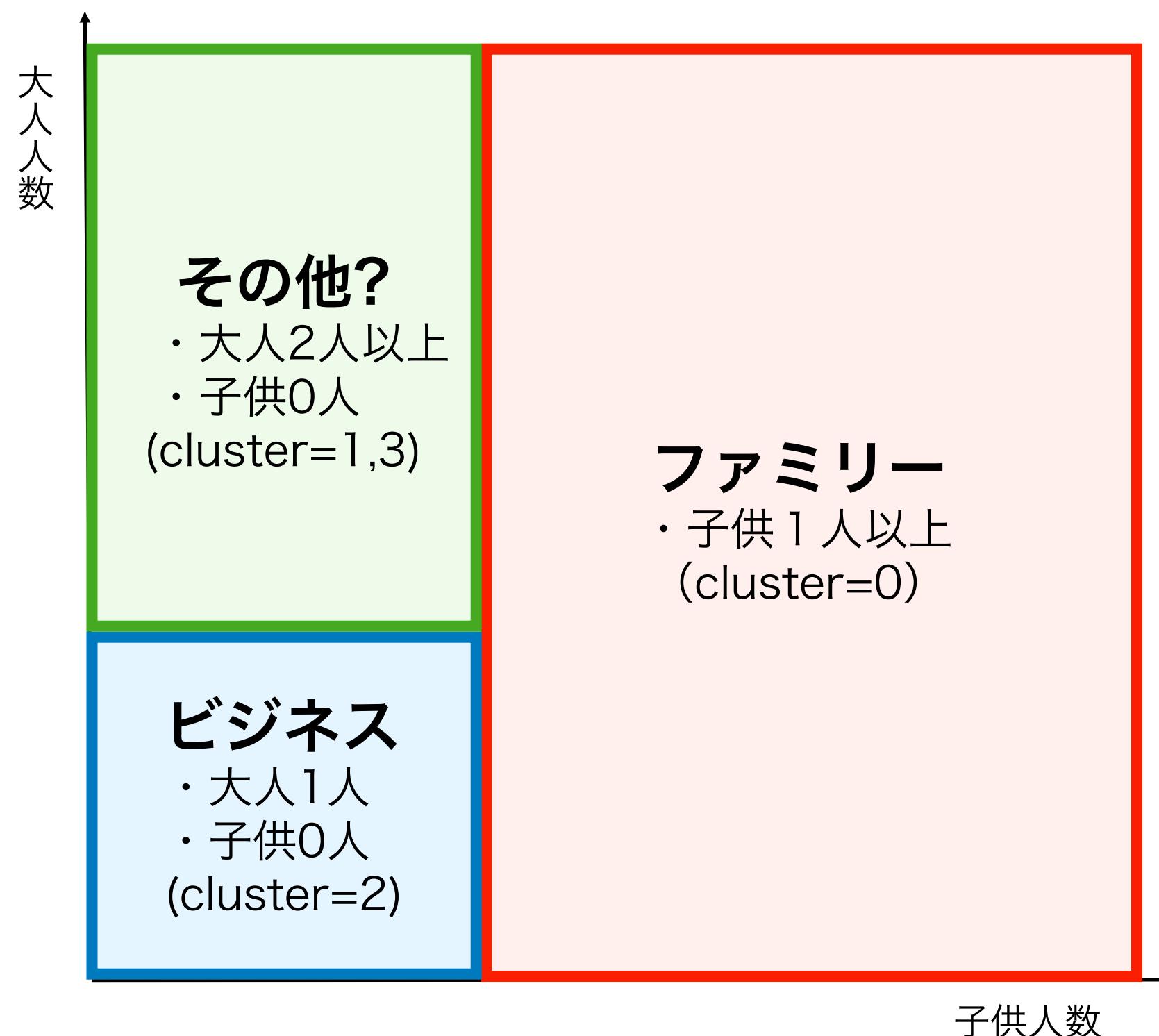


8. 分析結果・考察②

cluster0、cluster2、cluster1・3を子供人数、大人人数で分類している

→ cluster0は、ファミリー
→ cluster3は、ビジネス

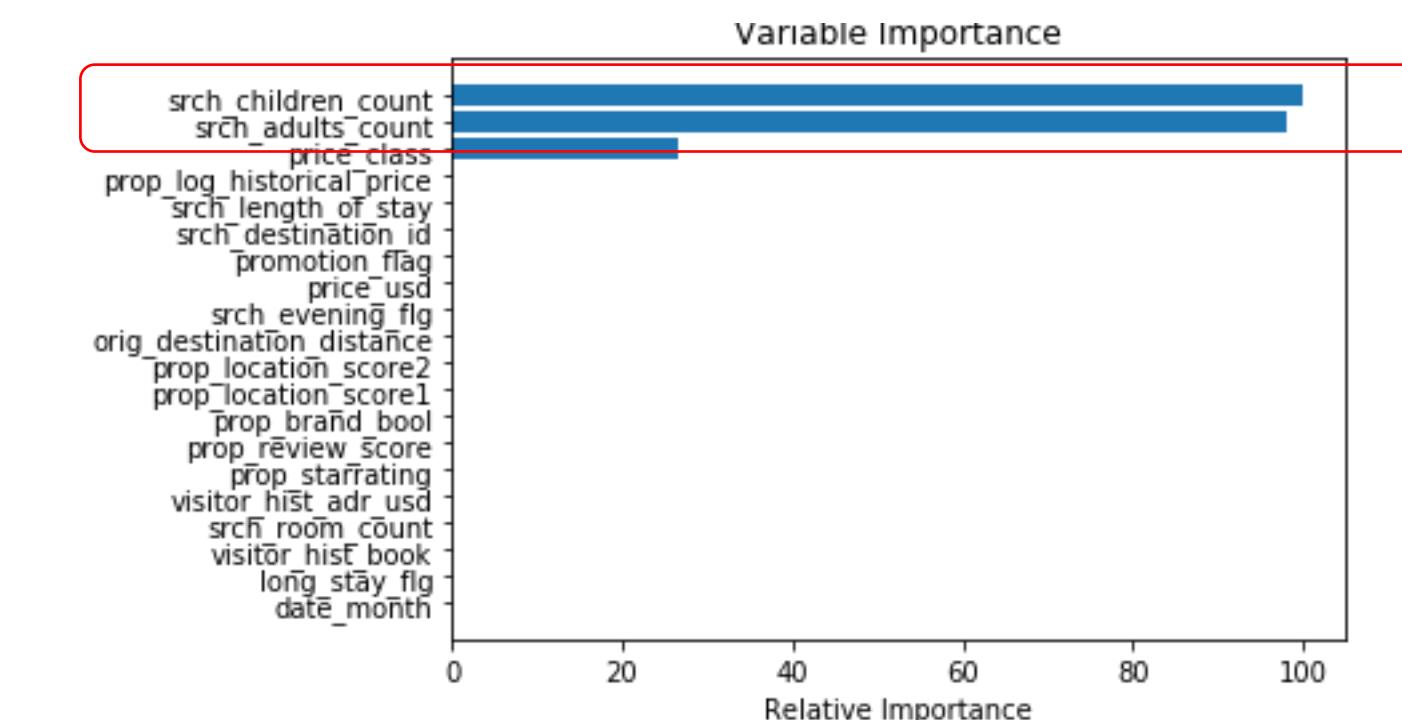
→ cluster1、3の分類は？



<分類するために重要な変数>

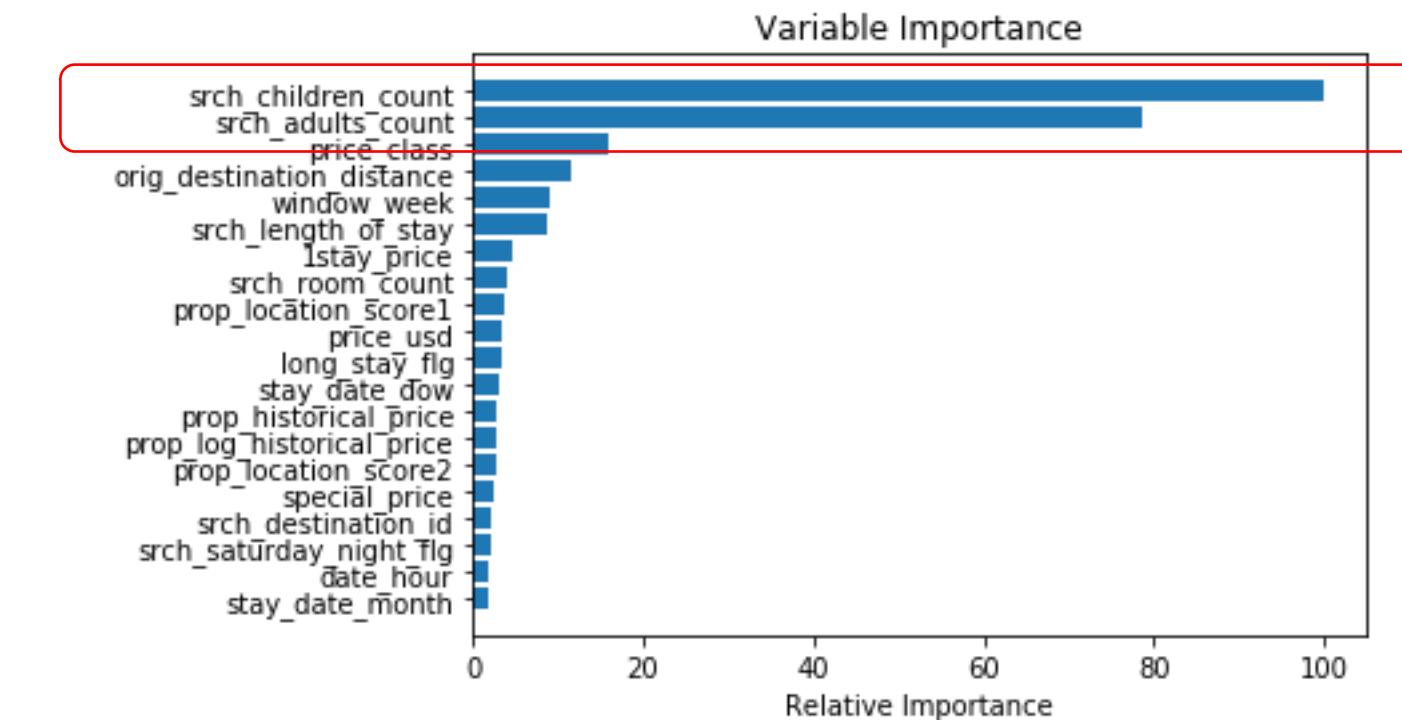
決定木

→ 子供人数、大人人数が重要



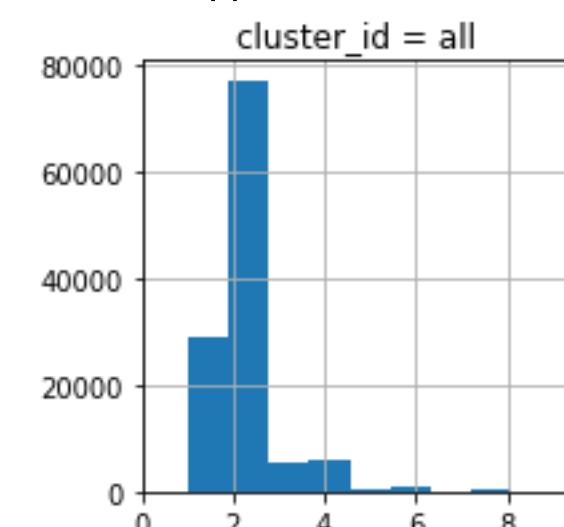
ランダムフォレスト

→ 子供人数、大人人数が重要

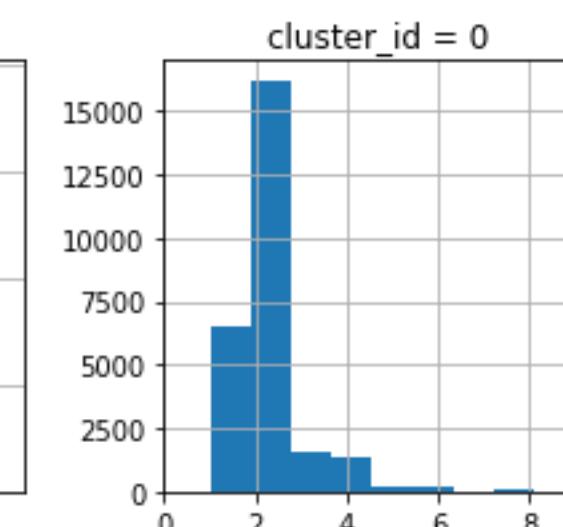


<大人人数>

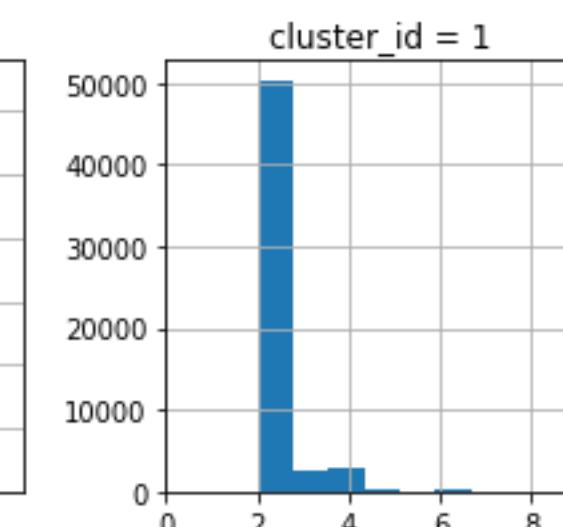
全体



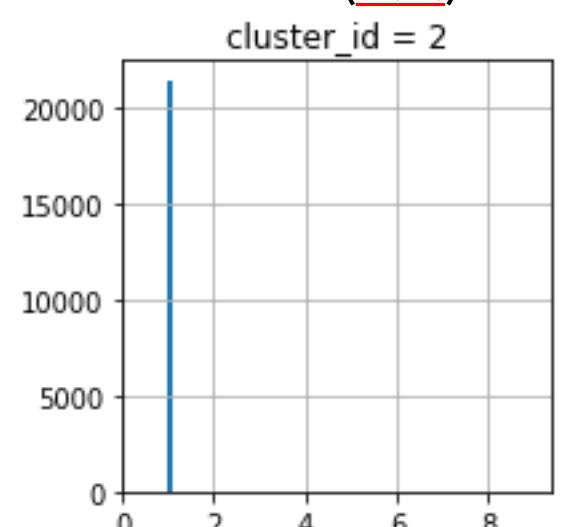
ファミリー



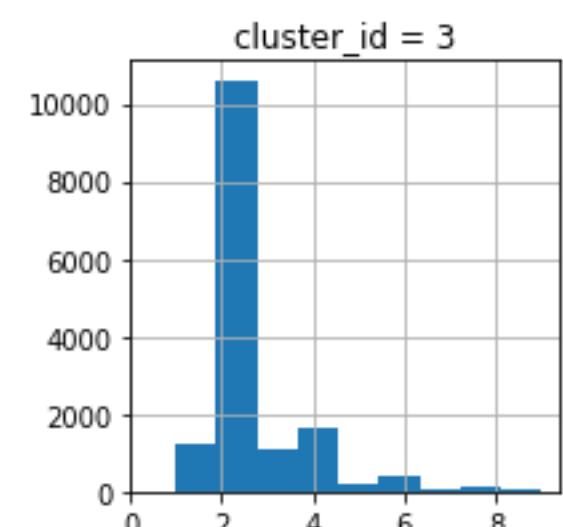
?



ビジネス(1人)

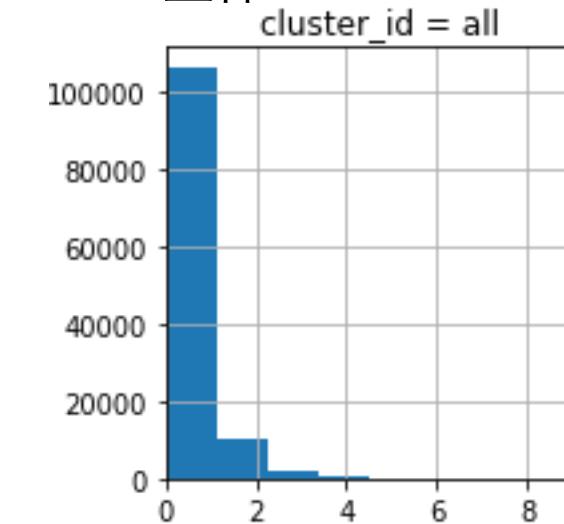


?



<子供人数>

全体



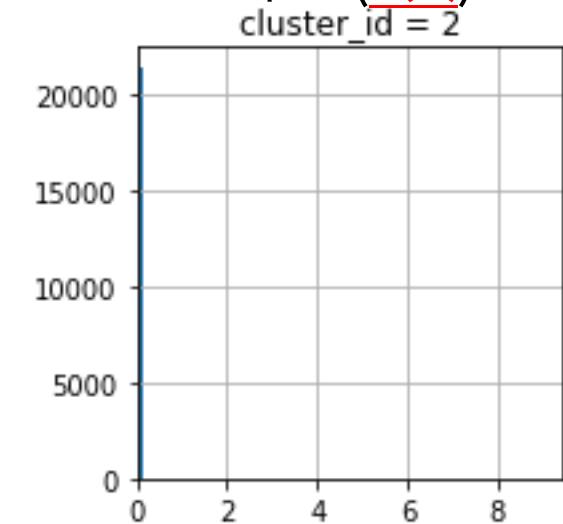
ファミリー(1人以上)



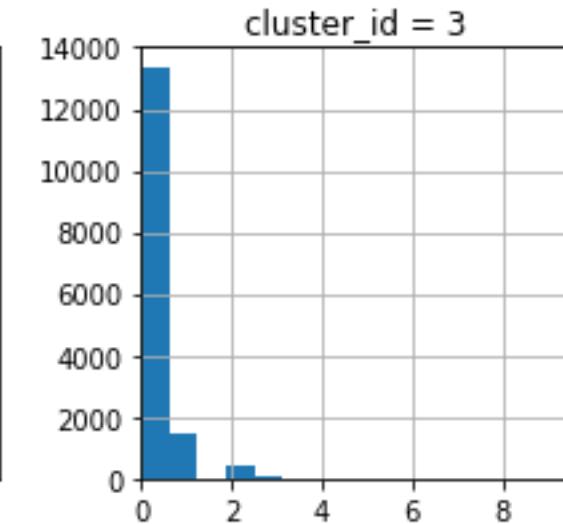
?



ビジネス(0人)



?



8. 分析結果・考察③

cluster1、3は、合計代金、距離(居住地と宿泊地)、泊数で分類している

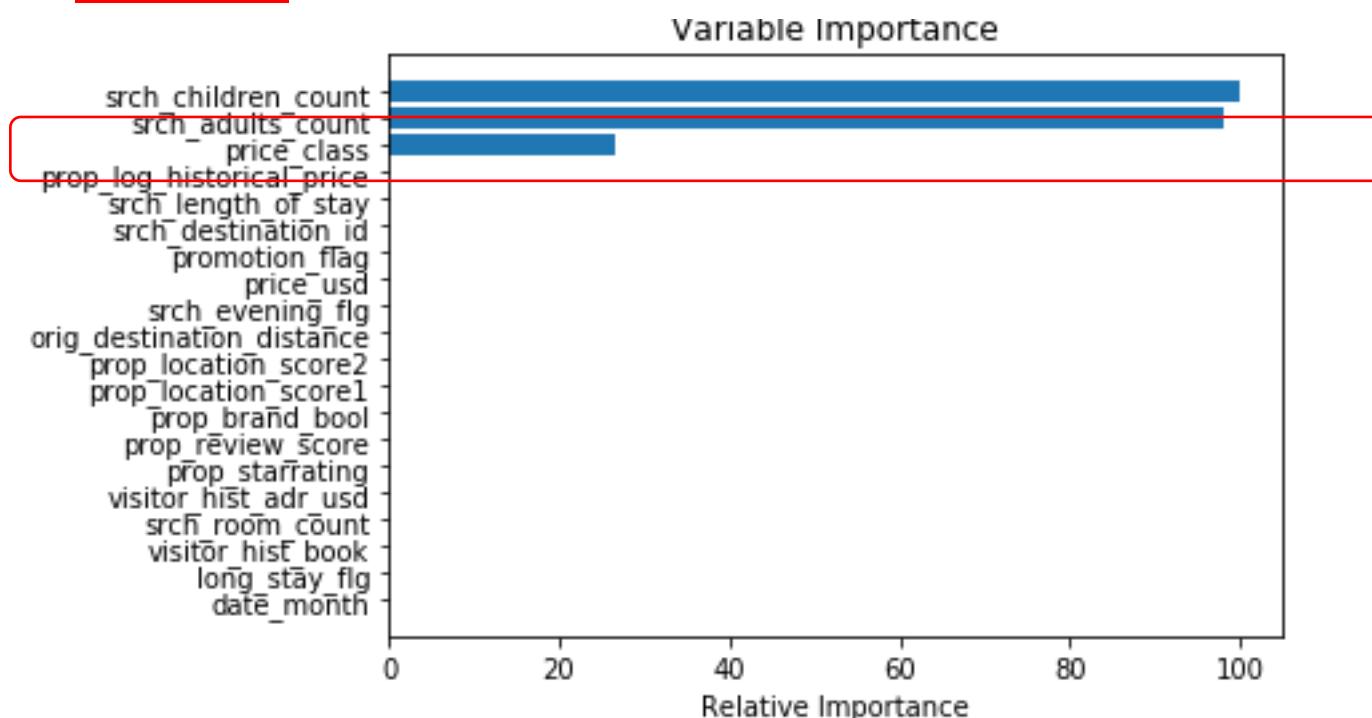
→cluster1は、**リゾート**(高価格、遠距離、泊数多)

→cluster3は、**おでかけ**(低価格、近距離、泊数少)

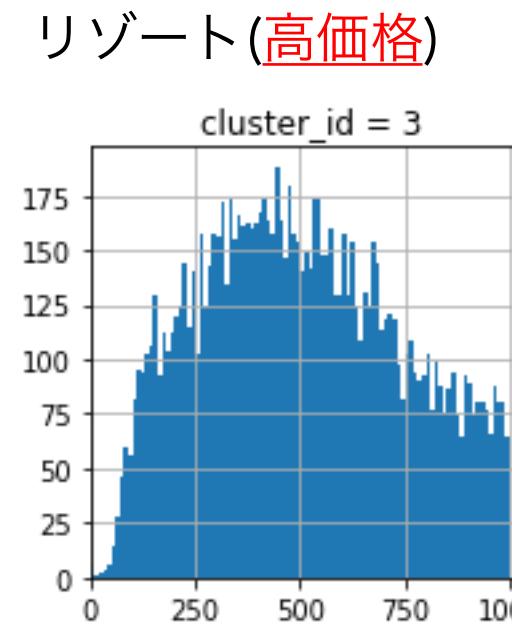
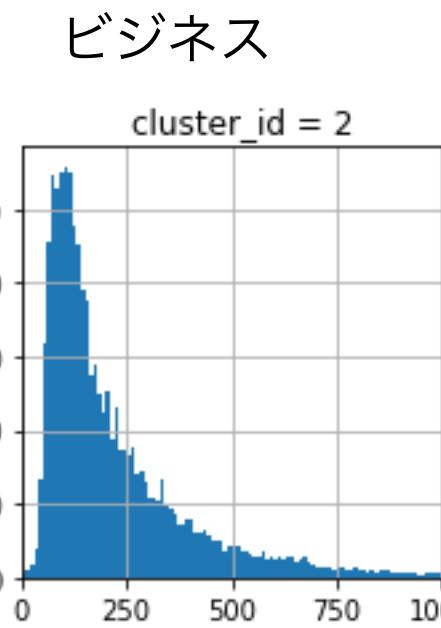
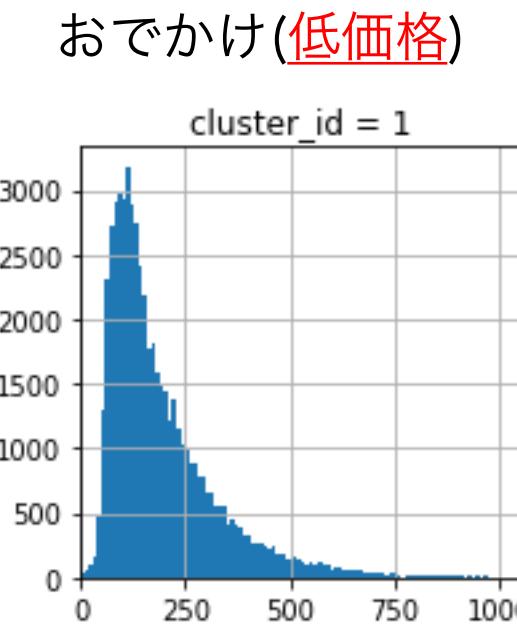
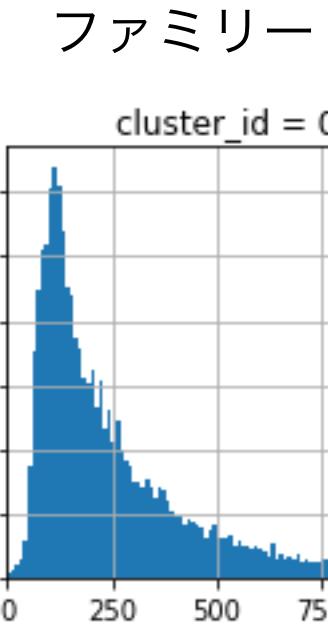
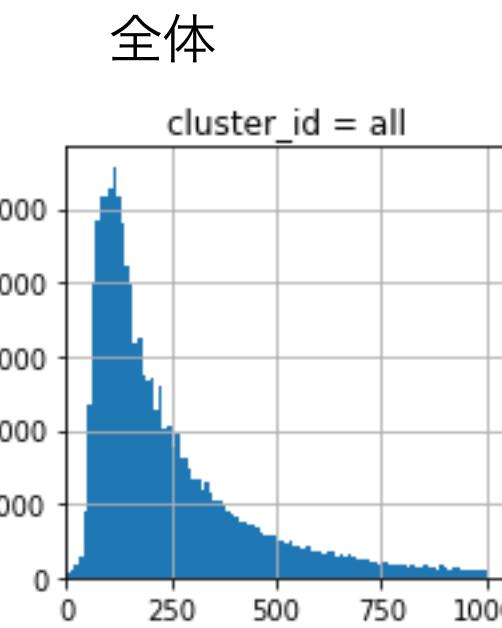
<分類するために重要な変数>

決定木

→合計代金が重要

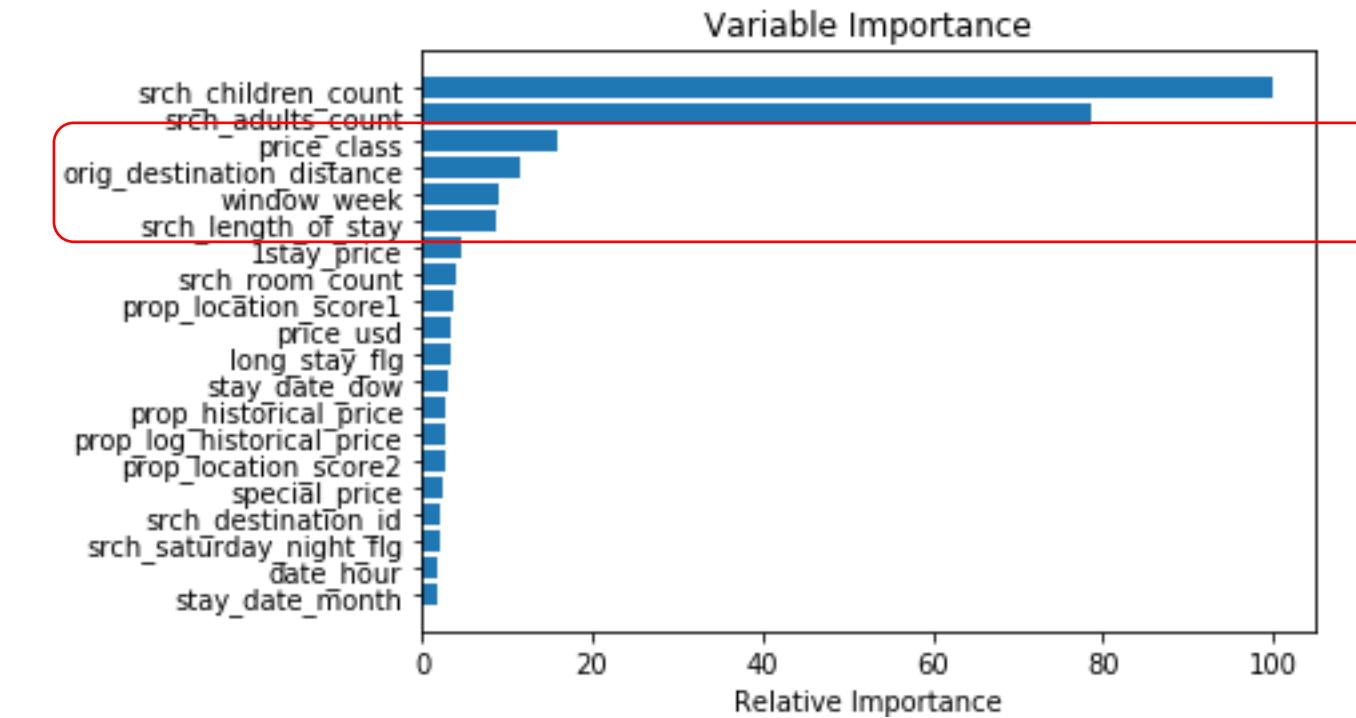


<合計代金>

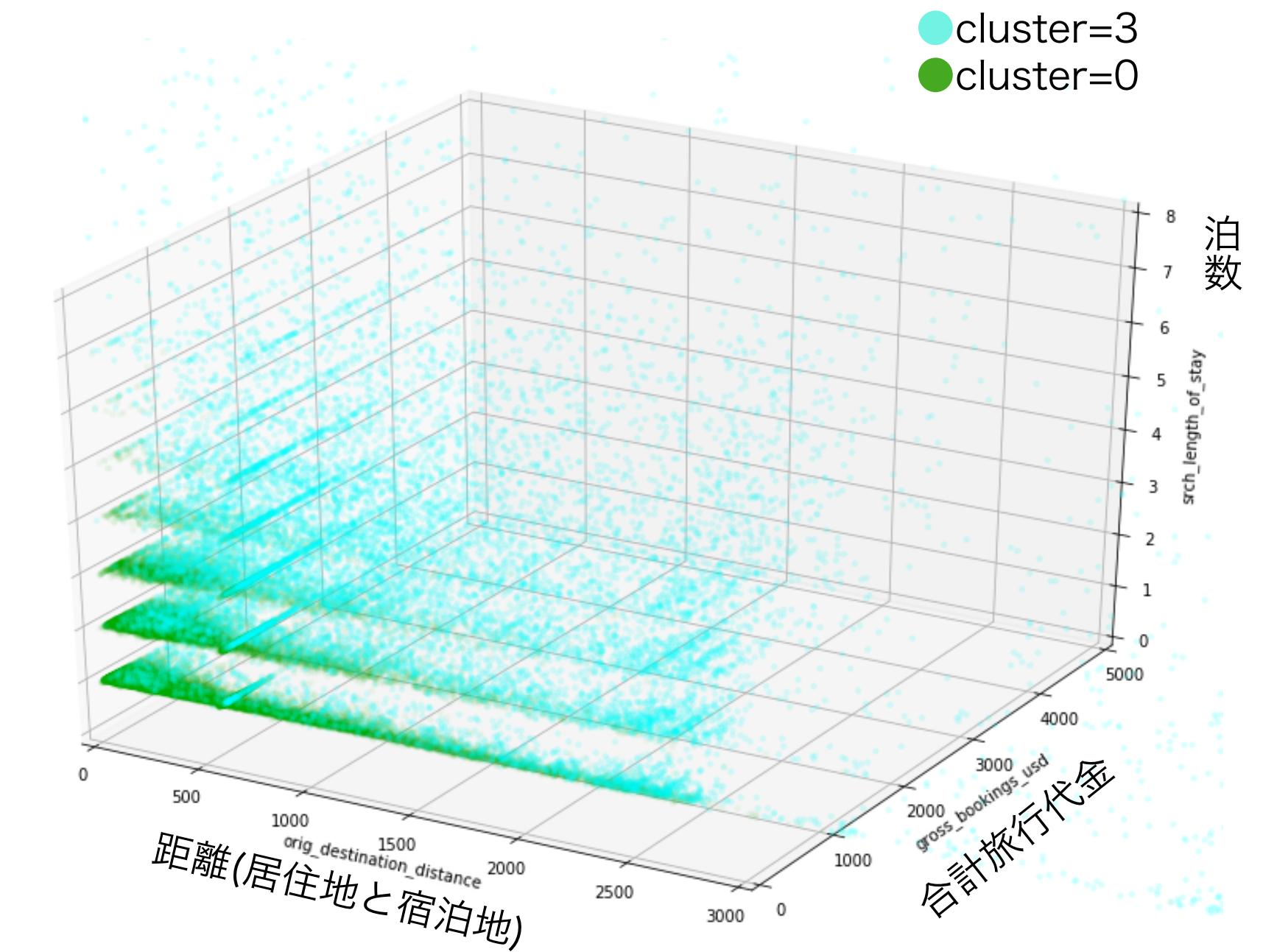


ランダムフォレスト

→合計代金が重要、あとは、距離(居住地と宿泊地)、泊数

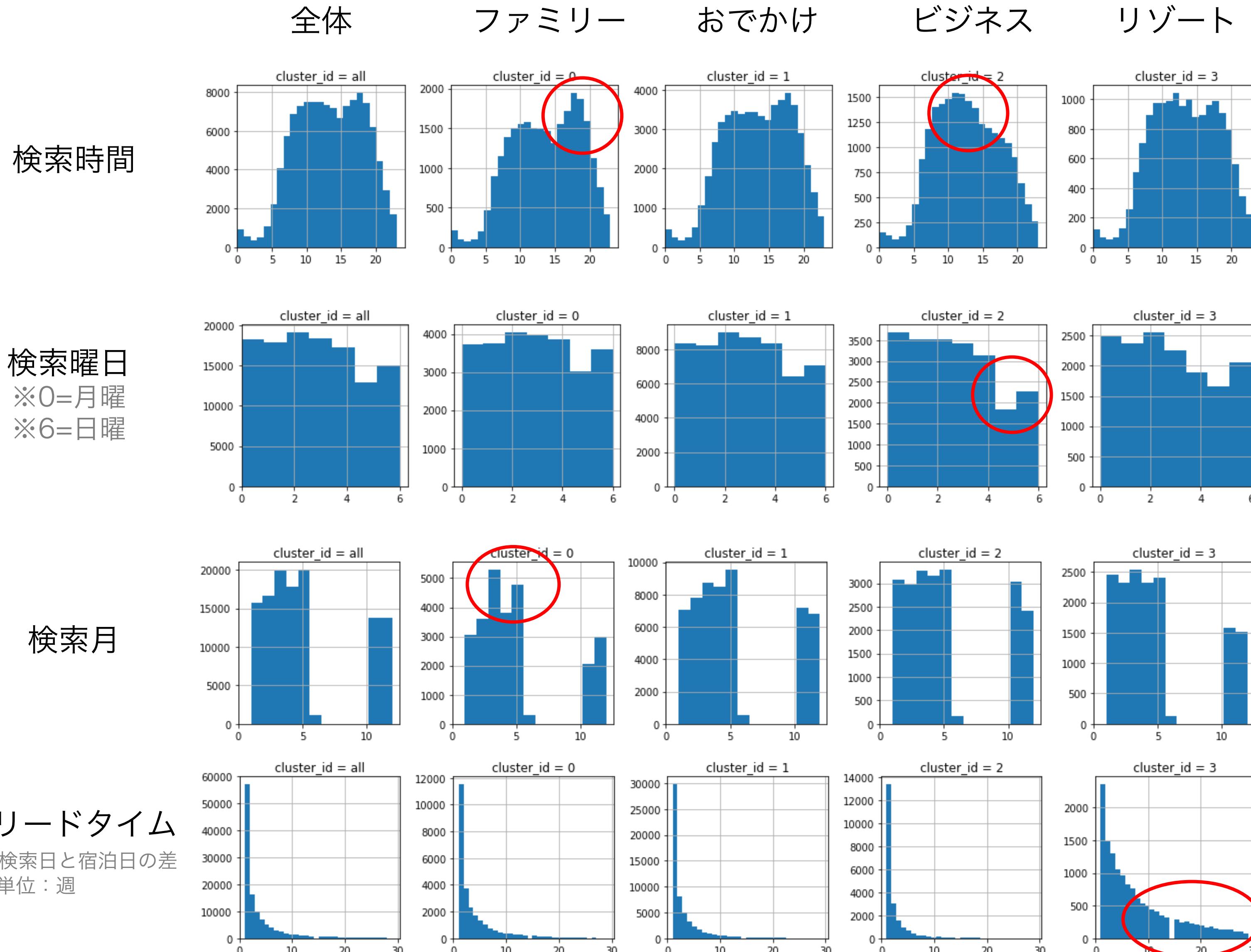


<合計代金、距離、泊数>



8. 分析結果・考察④

<クラスタごとの特徴>



<ペルソナ>

①ファミリー (子供1名以上)

[件数シェア : 22%、販売額シェア : 20%]

→夕食の時間帯や土日に家族と話しながら検討

→長期休暇（春休み、夏休み）の家族旅行

②おでかけ (低価格、近距離、泊数少)

[件数シェア : 47%、販売額シェア : 31%]

→日帰りで行ける場所でゆったり過ごす

③ビジネス (部屋あたり大人1名、子供0名)

[件数シェア : 18%、販売額シェア : 14%]

→平日朝昼にメールを見ながらスケジューリング

④リゾート (高価格、遠距離、泊数多)

[件数シェア : 13%、販売額シェア : 35%]

→(隠れ家高級ホテルのいい部屋は早く埋まるので)

早めに予約

<考察>

→販売額シェアから考えると

④リゾートの打ち手を考えたいところだが…

※超高額予約が販売額を引き上げている点は懸念

※④以外は、特に、国名、地域名、ホテル名等のマスター情報がないと打ち手が考えにくい

9. 当該結果のビジネスインパクト

ビジネス利用者に対し、
履歴に基づくレコメンドメールを
毎週月曜朝8時に送信すれば、
対象者の販売額が6%増加する
(24万件・販売額55億円)

<効果想定
配信数：420万通

クリック率：3%
CVR：3%

メール経由予約件数：12万件

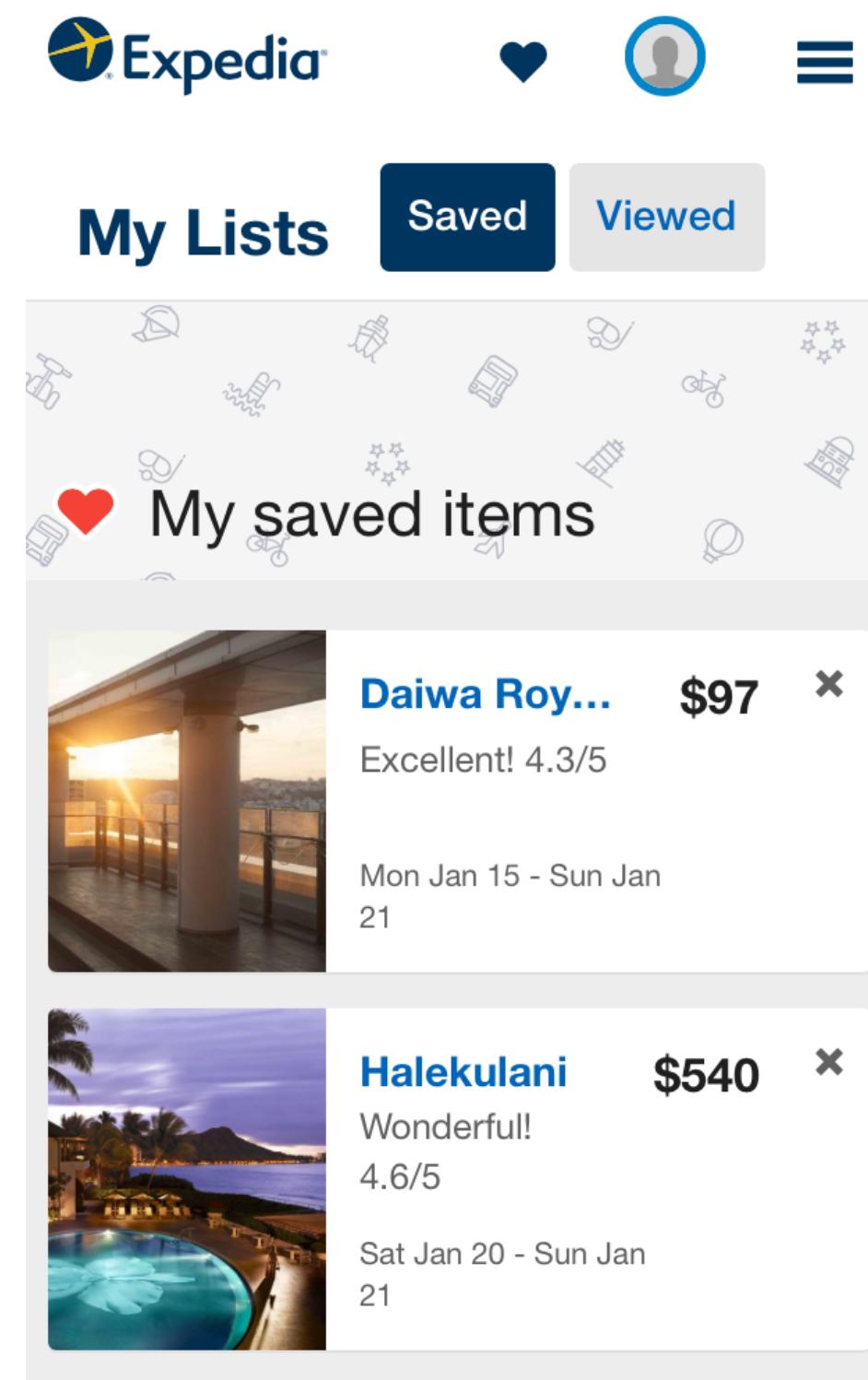
増加予約件数：週に6万件 → 月に24万件

※50%が自然流入の転換として算出

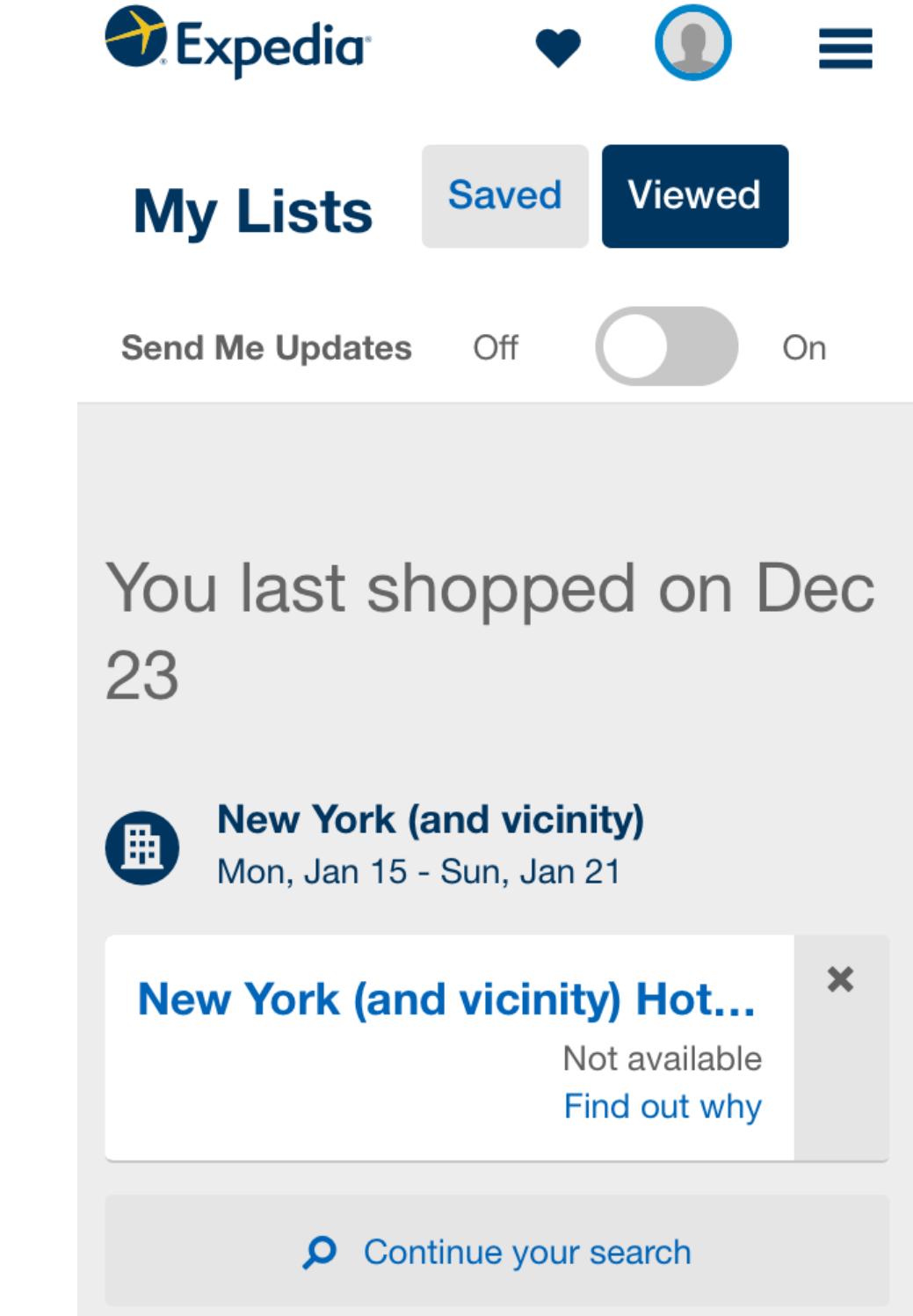
※ビジネス利用の販売が月に350万件・805億円として算出
(Expediaの売上、事業別収益比率、コンペデータを基に)

<履歴に基づくレコメンドのイメージ>

お気に入り検索条件



前回検索条件



※ビジネス利用者は検索履歴or購入履歴が

「部屋あたり大人1名、子供0名」を想定(1年以内の履歴)

※履歴に基づくレコメンドは、お気に入り検索条件 or

前回検索条件を想定(会員ページにすでにあるコンテンツ)

※月曜朝8時は、週初め・出社時のメールチェック時にメールを見るタイミング

※毎週開催しているフラッシュセール(72時間限定)をこのタイミングに合わせるのも良いのでは

10. ビジネスインパクトを調査する方法

- ・ **ビジネス利用者を2つに分けてABテスト**
(顧客idを奇数偶数で分ける)
- ・ **指標：配信対象者の販売額**
※配信日から1週間
※メール経由以外の購入も含む
(メール以外の購入が減る可能性もあるため)

10. 今後さらに実施してみたいこと

- ・**クラスタリングのアルゴリズム変更**

相関性の強い変数がある程度出やすいらしい
ユーグリット距離出なくマハラノビス距離を使用してみたい

- ・**同じ手法でもう少しリッチなデータセットで分析したい**

(idのマスタがある情報で分析、もしくは地域やホテルをクラスタリングして特徴を把握できれば…)