DataScientist Society

雲雀丘学園DX【情報Ⅱ】第6回全国指導力向上研修会

"情報Ⅱで学んだ内容が実社会でどのように使われているか"のご紹介

「クラスタリング」の実務応用事例

一般社団法人 データサイエンティスト協会 学生委員会

2025年2月12日

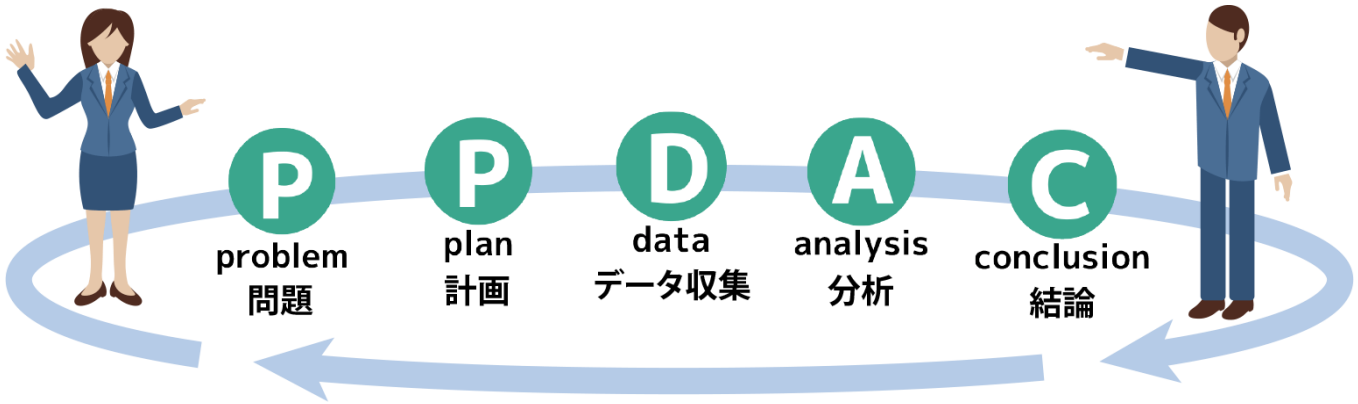
アウトライン

1. データサイエンスの実社会での使われ方｜PPDACサイクル
2. Problem｜問題の設定
3. Plan
    1. As-IsとTo-Beの設定、分析方針策定
    2. 要因分析
4. Data｜データの取得
5. Analysis
    1. データ可視化
    2. 階層的クラスタリング
    3. 非階層的クラスタリング
    4. 解釈のための可視化
    5. 解釈
6. Conclusion｜結論
7. Appendix

1. 実社会での使われ方｜PPDACサイクル

データサイエンスはあくまでも”手段”
実社会のどのような課題を解決するかの設計が大事！

データサイエンスを活用した現場改善は、
教育現場でも使用されるPPDACサイクルに近いプロセスで進めることが多いため、
今回はPPDACサイクルに則った生命保険会社の営業でのデータサイエンス活用事例を紹介します。



出典：総務省統計局ホームページ
https://www.stat.go.jp/dstart/point/seminar1/01.html

一般社団法人
データサイエンティスト協会

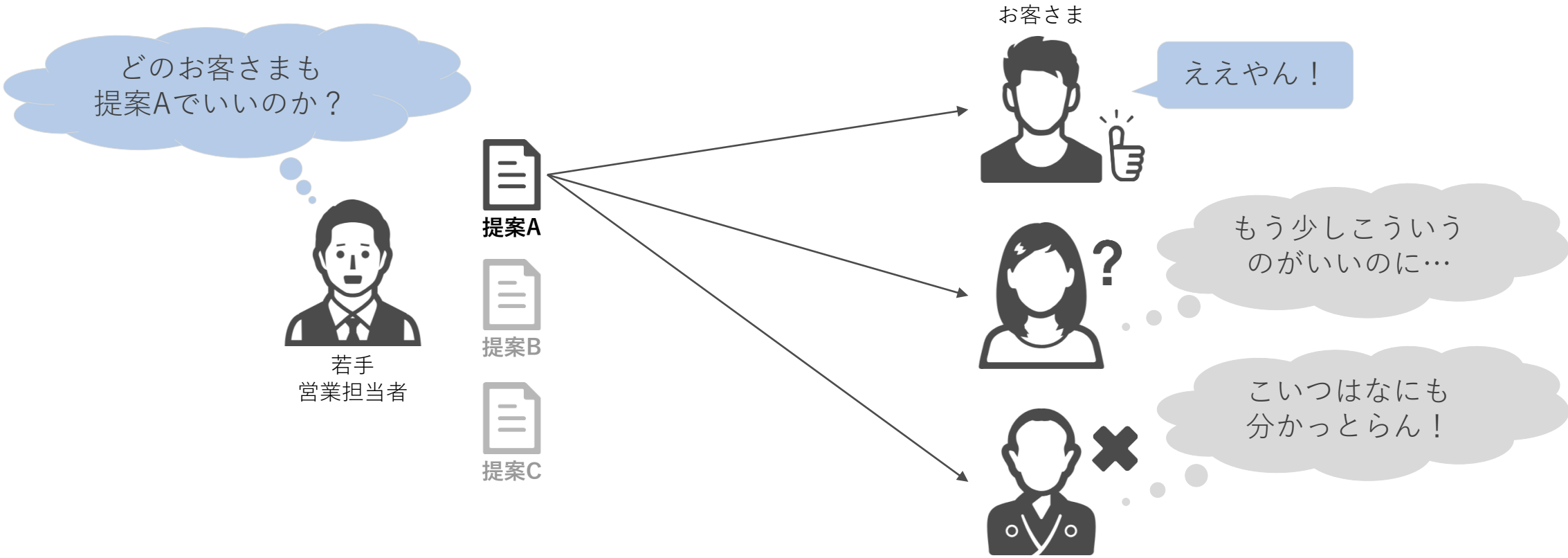
ビジネススキル向上のための課題解決型人材コンテスト2024（応募期間：～6/7まで）

データ活用に必要なスキルはAI/機械学習などのテクノロジーにとどまらず、ビジネスオーナーの課題を適切に理解して実行可能な解決策を提示することにあると言われています。当協会はそれに応える試みとして、実課題と実データを用いてアクションの提案までを行う分析コンテストを実施いたします。他業界のメンバーとチームを組み3か月間、メンターのサポートを受けながら分析プロジェクトを推進することで、ビジネススキルを身につけていただける機会です。分析スキルをさらに業務で活かして行きたいとお考えの方は是非ご参加ください。

出典：データサイエンティスト協会ホームページ
https://www.datascientist.or.jp/news/n-news/post-2890/

2. Problem｜問題の設定

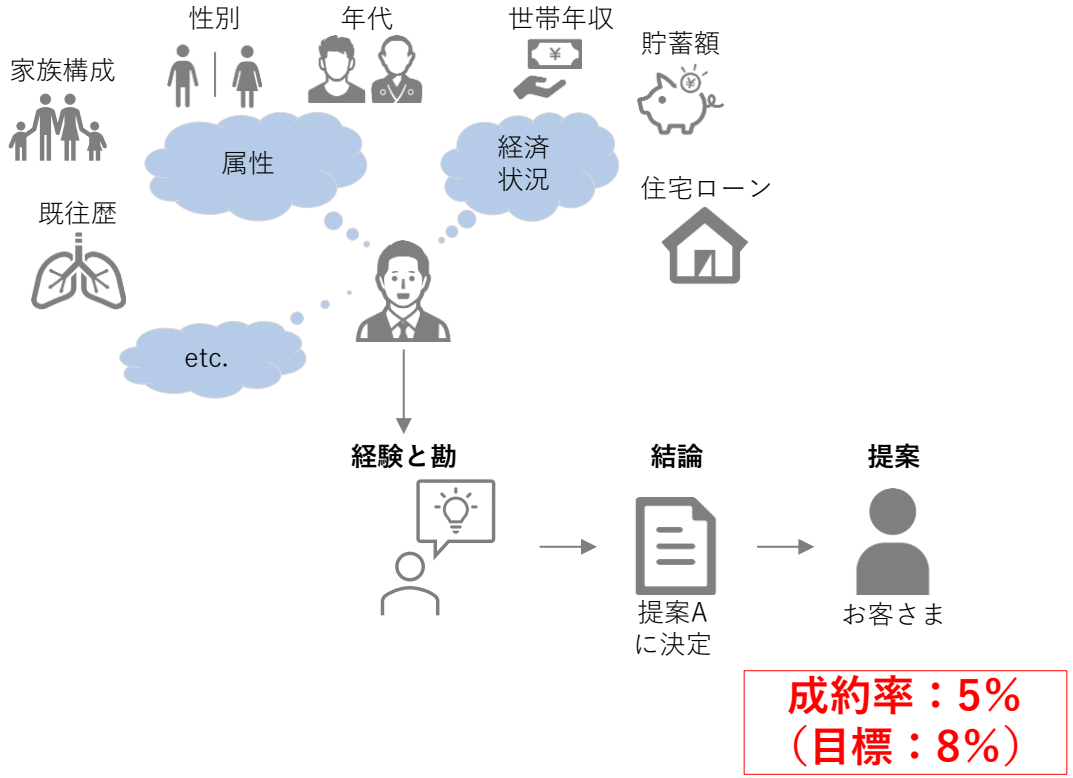
- **問題**：生命保険会社の営業では、お客さまのニーズによって提案内容を変えることが重要。しかし、特に若手営業担当者はお客さまのニーズを考え提案内容を変えることに慣れていないので、お客さまが求める提案が出来ず成約率が低くなってしまっている。



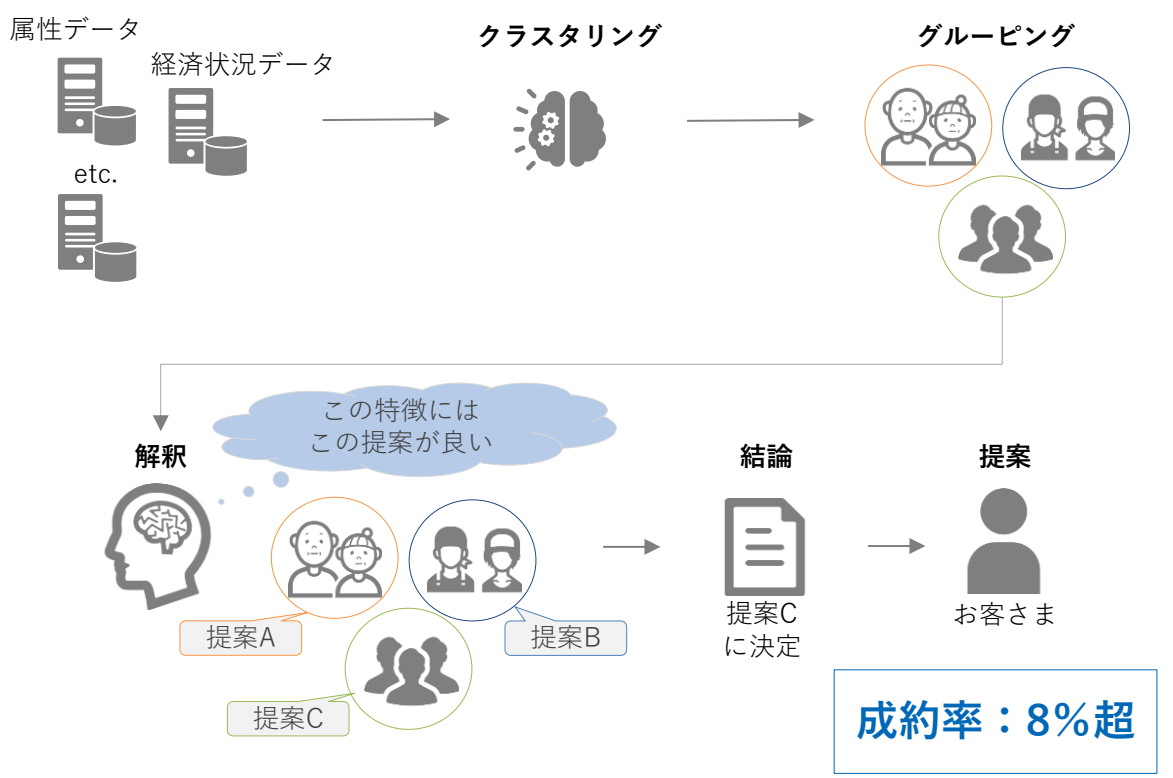
3-1. Plan｜As-IsとTo-Beの設定、分析方針の策定

- **As-Is**：お客さまのニーズを元に提案内容を変えることが難しいため、成約率が低い。
- **To-Be**：お客さまをグルーピングし、グループごとに提案内容を考え変えることで、成約率を向上させる。

As-Is（現状）

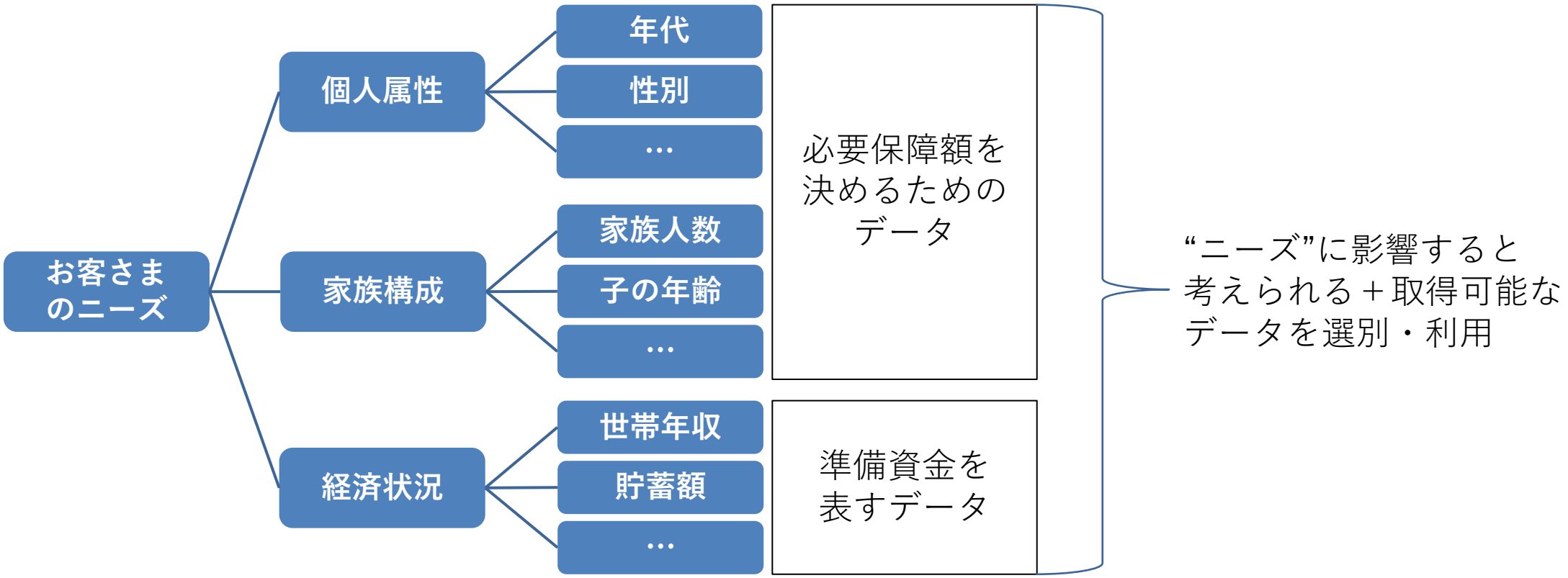


To-Be（あるべき姿）



3-2. Plan｜要因分析

- お客さまのニーズに影響する要因を分解して、使用するデータを洗い出す。



4. Data｜データの収集

- お客さまをグルーピングするために、お客さまの属性・行動履歴などのデータを利用。
  ※クラスタリングは教師なし学習のため、今回は目的変数なしのデータ。

説明変数

お客さまID	性別（1：男性、2：女性、3：法人）	年齢(歳)	役職（1：経営層、2：管理職、3：一般社員、4：専業主婦）	家族人数(人)	子どもの平均年齢(歳) ※子どもいない場合も0歳	世帯収入(万円)	貯蓄額(万円)	投資経験（0：なし、1：あり）
0	2	28	3	2	0	354	283	1
1	1	35	3	2	0	616	439	0
2	2	23	2	1	0	305	199	0
3	3	46	1	1	0	630	349	1
4	1	41	3	3	8	552	308	1
5	2	26	3	3	0	320	234	1

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5. Analysis

- 取得したデータの全体像を把握してから詳細データの確認を行う。
  その後、クラスタリングを実施し、お客さまをグルーピング。

全体

演習①　データ全体の俯瞰
- データは何行×何列あるか確認しよう。
- データの最初の5行の値を確認して、どのような列名があるかを確認しよう。
- 欠損値が無いかを確認しよう。
- 要約統計量を確認しよう。

演習②　各説明変数の分布の確認
- 連続値の説明変数についてはヒストグラムで、離散値の説明変数については各要素数で分布を確認しよう。

演習③　説明変数同士の関係の確認
- 各説明変数同士の関係性を相関係数で確認しよう。

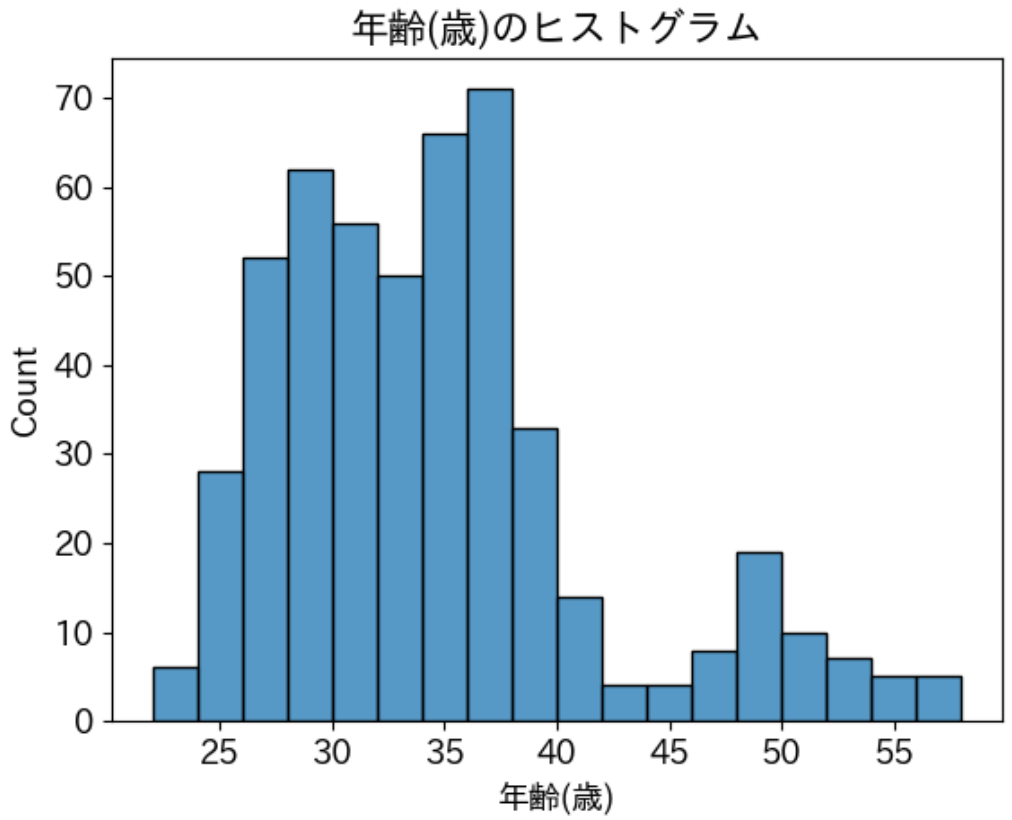
演習④　階層的クラスタリングと非階層的クラスタリングの実施
- 階層的クラスタリングと非階層的クラスタリングを実施し、何グループに分けられそうかを考えよう。

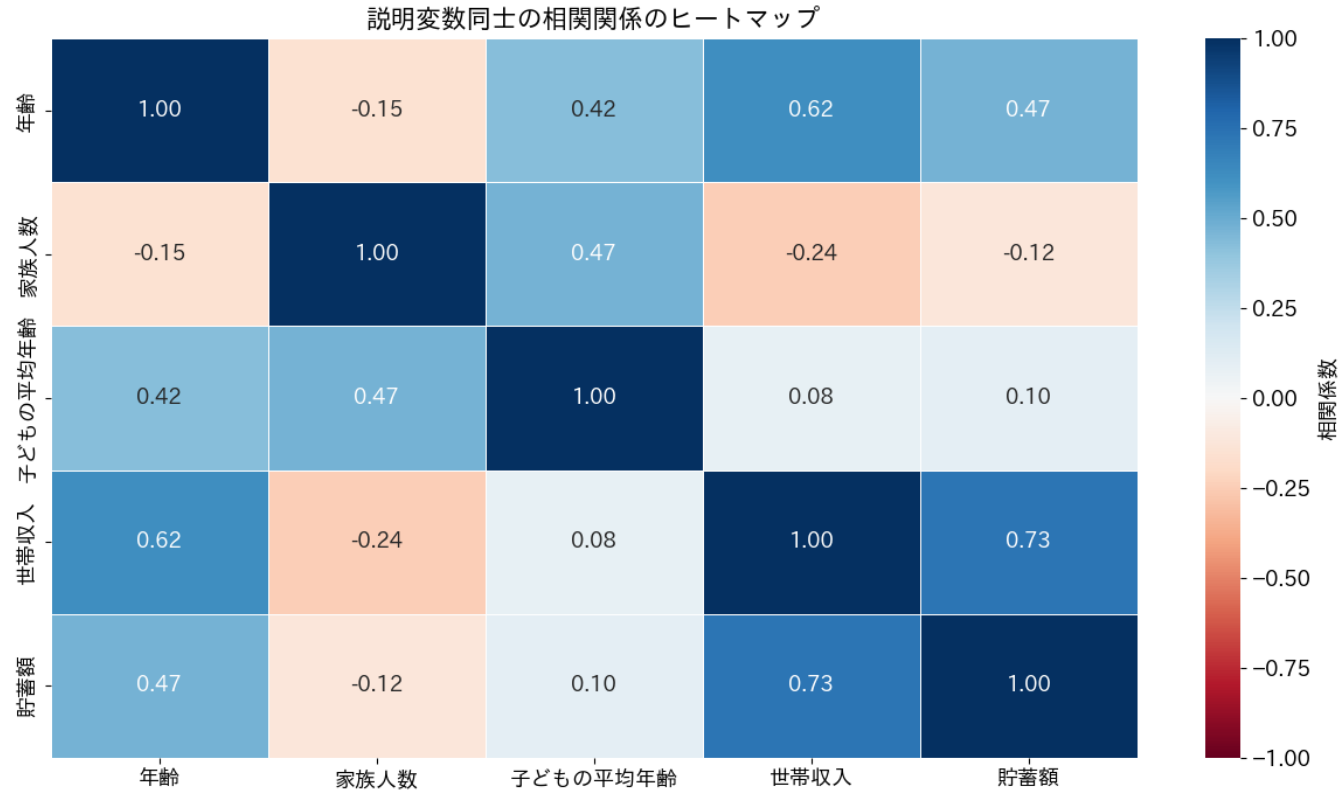
演習Advanced　グループごとの特徴の可視化と解釈
- グループごとにどのような提案が良いかを考えるために、各グループの特徴を解釈してみよう。

詳細

5-1. Analysis｜データ可視化

- 各説明変数の分布を確認。
- 各説明変数同士の関係性を散布図・ヒートマップで確認。

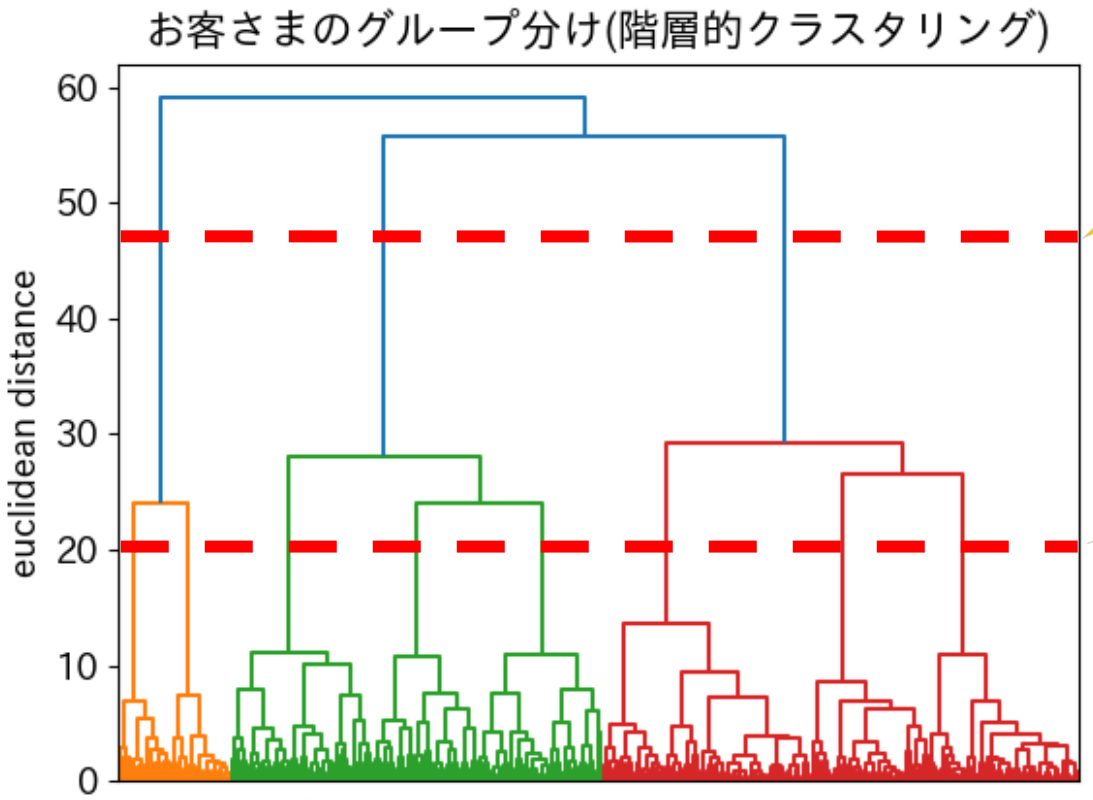




※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5-2. Analysis｜階層的クラスタリング

- 各説明変数のスケールを揃えるため標準化を実施したのち、階層的クラスタリングを実施。
- デンドログラムで何グループに分けるのが良さそうか確認。（人によってブレがある。）
- 実務では業務要件や制約に応じて、グループ数を決定することもある。



3つにグルーピングするのが良さそう？

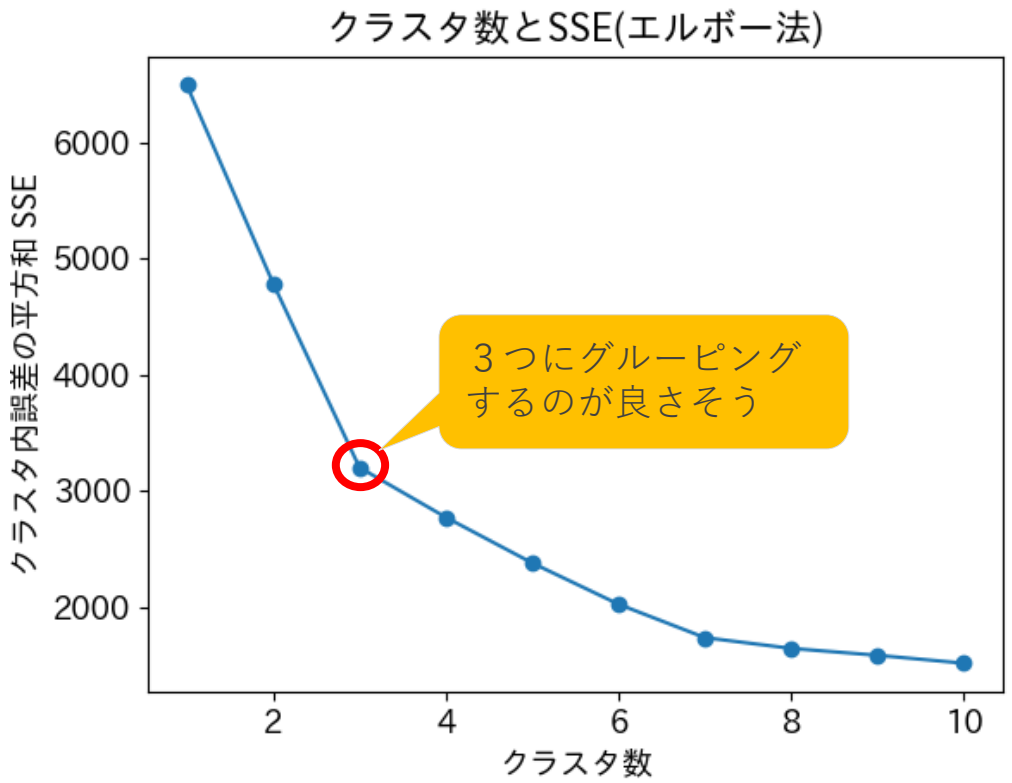
8つが良さそう？と考える人もいるかも

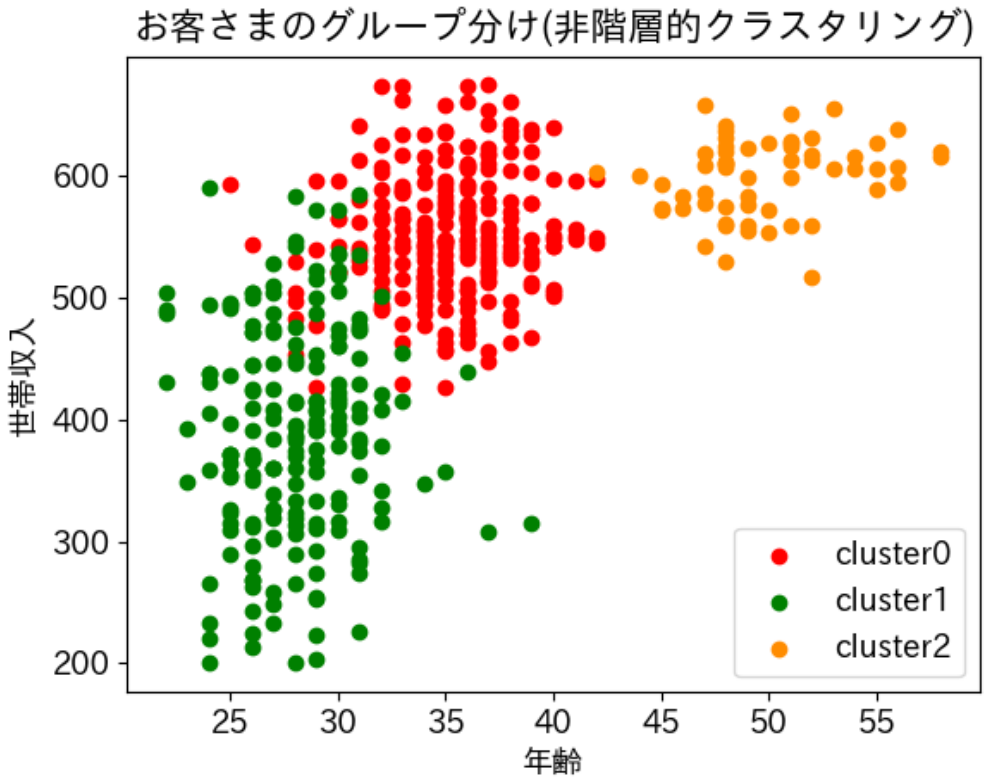
特にこれが正解などは無く、
実務では、業務要件に応じて
グループ数を決定することもある

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5-3. Analysis｜非階層的クラスタリング

- 階層的クラスタリングと同じく標準化を実施したのち、非階層的クラスタリングを実施。
- エルボー法で何グループに分けるのが良さそうか確認。（先ほど同様、正解はない。）
- 今回は3グループに分けるのが良さそうと判断し、各グループの解釈を進めていく。

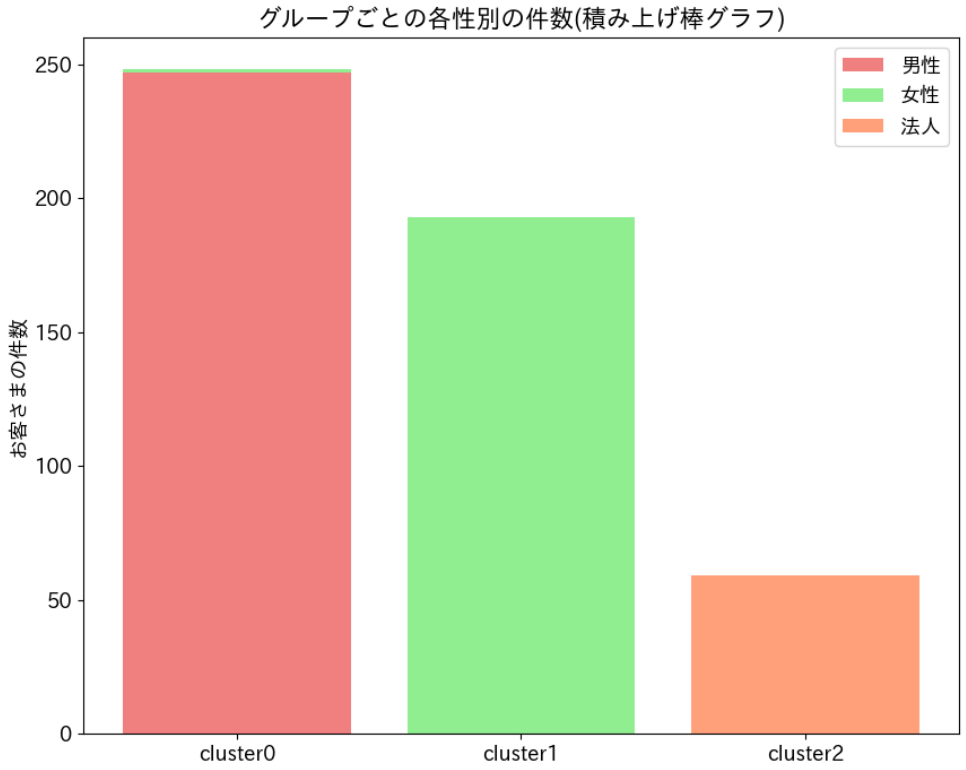


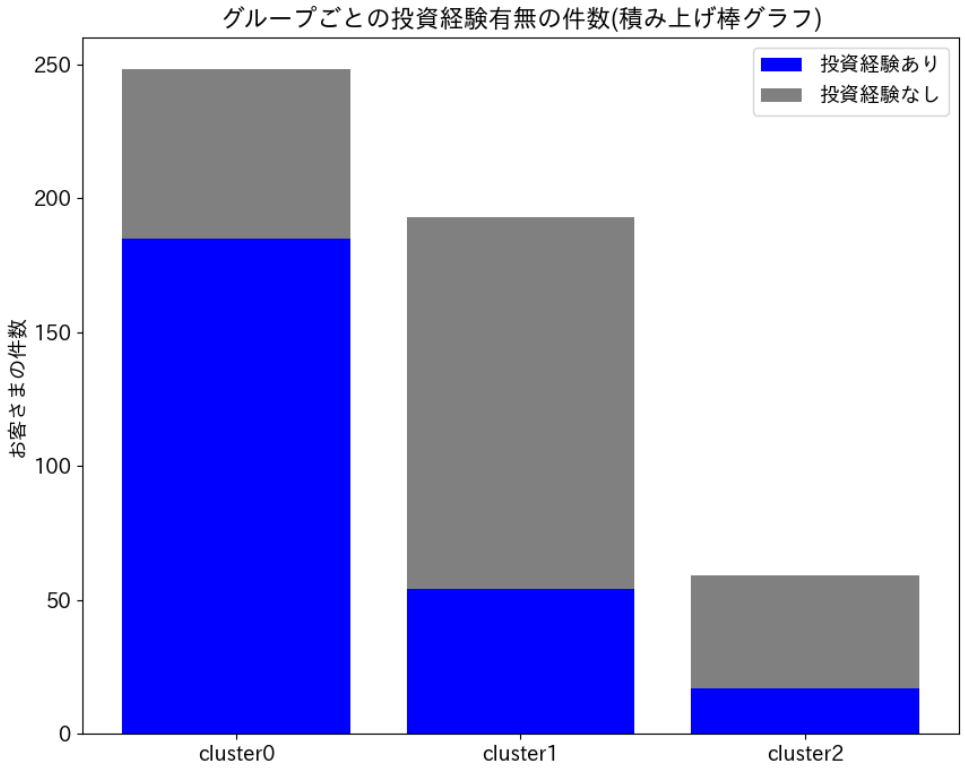


※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5-4. Analysis｜解釈のための可視化

- 各グループの特徴を可視化結果から把握し、特徴に合わせた提案内容の検討に活用。
- 可視化から読み取れる特徴としては、cluster0は男性がほとんどで投資経験ありの人が多い。





※あくまでダミーデータのため、現実の傾向を反映するものではありません。

5-5. Analysis｜解釈

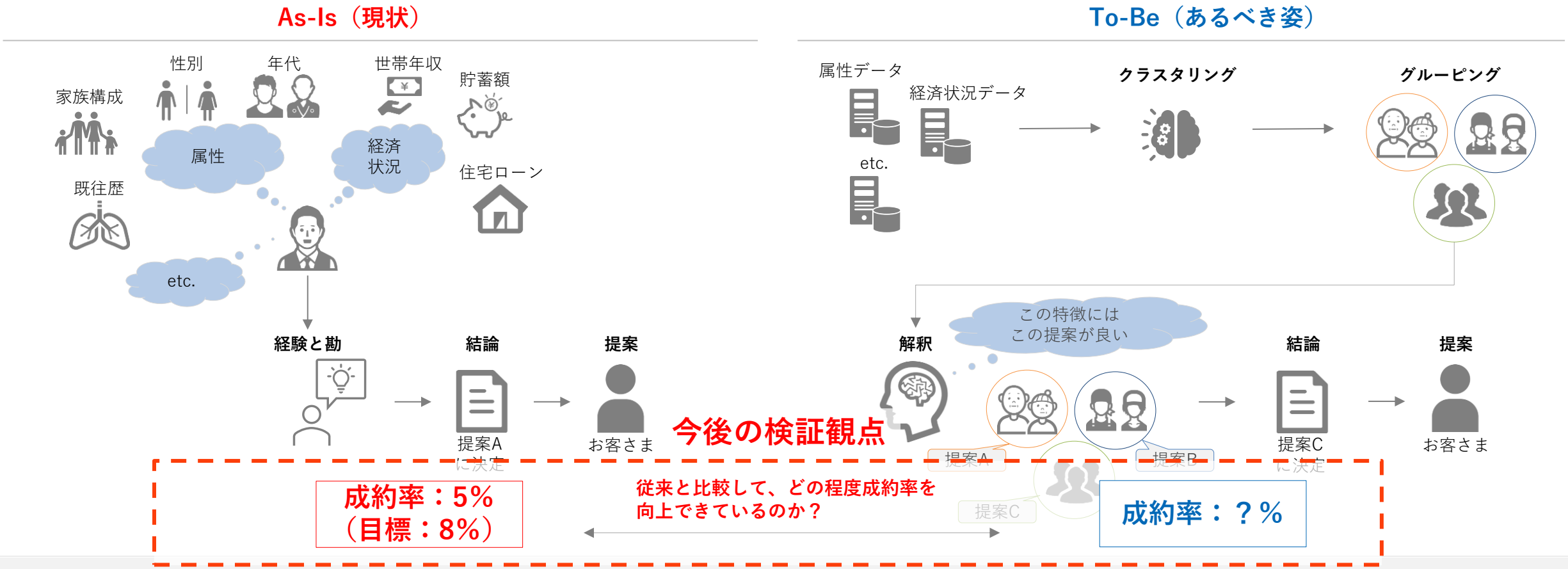
- 各グループの特徴からお客さま像を解釈し、特徴に合わせた提案内容を検討。
  ※提案内容は、過去のデータ（お客さまの属性・提案内容・成約率など）も勘案し検討。

	グループ名（解釈）	特徴	提案内容
cluster0	中年サラリーマン層	30代のサラリーマンが多く、投資経験ありの人が多い	【提案A】 自身の老後に備えるための資金形成を目的とした変額保険
cluster1	若年ママ層	20代の女性が多く、子どもが0～5歳と幼い	【提案B】 子どもの将来や自身の生活習慣病に備える保険
cluster2	中高年経営者	社長などの経営者が多く、高齢の方が多い	【提案C】 自身が病気で不在になった場合に、会社の借入金などの返済に備える大型保険

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

6. Conclusion｜結論

- **As-Is**：お客さまのニーズを元に提案内容を変えることが難しいため、成約率が低い。
- **To-Be**：お客さまをグルーピングし、グループごとに提案内容を考え変えることで、成約率を向上させる。



以上、ご意見とアンケートのご回答
よろしくお願いします。

アンケートURL：https://forms.office.com/r/04hcQ9SVqq

ZoomチャットにURLをお送りいたします。
匿名回答で2分ほどで終わりますので、
ご回答のほどよろしくお願いいたします！

7. Appendix

演習回答

演習①

・データは何行×何列あるか確認しよう。

```
[4]  print(df.shape)
```

```
(500, 9)
```

500行×9列

・データの最初の5行の値を確認して、どのような列名があるかを確認しよう。

```
[5]  df.head()
```

	お客さまID	性別(1：男性、2：女性、3：法人)	年齢（歳）	役職(1：経営層、2：管理職、3：一般社員、4：専業主婦)	家族人数（人）	子どもの平均年齢(歳)※子どもいない場合も0歳	世帯収入（万円）	貯蓄額（万円）	投資経験(0：なし、1：あり)
0	0	2	30	4	5	10	379	299	1
1	1	1	35	4	2	0	535	329	1
2	2	1	29	4	2	0	427	313	1
3	3	2	30	4	2	0	469	333	0
4	4	2	28	4	4	8	584	366	0

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

演習回答

演習①

・欠損値が無いかを確認しよう。

```
[6]  df.isnull().sum()
```

	0
お客さまID	0
性別(1：男性、2：女性、3：法人)	0
年齢(歳)	0
役職(1：経営層、2：管理職、3：一般社員、4：専業主婦)	0
家族人数(人)	0
子どもの平均年齢(歳)※子どもいない場合も0歳	0
世帯収入(万円)	0
貯蓄額(万円)	0
投資経験(0：なし、1：あり)	0

各列で欠損値は無し。

・要約統計量を確認しよう。

```
df.describe()
```

	お客さまID	性別(1：男性、2：女性、3：法人)	年齢(歳)	役職(1：経営層、2：管理職、3：一般社員、4：専業主婦)	家族人数(人)	子どもの平均年齢(歳)※子どもいない場合も0歳	世帯収入(万円)	貯蓄額(万円)	投資経験(0：なし、1：あり)
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	249.500000	1.624000	34.118000	2.768000	2.564000	3.380000	495.014000	347.452000	0.512000
std	144.481833	0.686707	7.280803	0.977798	1.156104	6.863659	107.818828	75.919347	0.500357
min	0.000000	1.000000	22.000000	1.000000	1.000000	0.000000	200.000000	115.000000	0.000000
25%	124.750000	1.000000	29.000000	2.000000	2.000000	0.000000	421.750000	298.000000	0.000000
50%	249.500000	2.000000	33.000000	3.000000	2.000000	0.000000	521.000000	352.000000	1.000000
75%	374.250000	2.000000	37.000000	4.000000	3.000000	5.000000	574.000000	393.250000	1.000000
max	499.000000	3.000000	58.000000	4.000000	7.000000	36.000000	675.000000	603.000000	1.000000

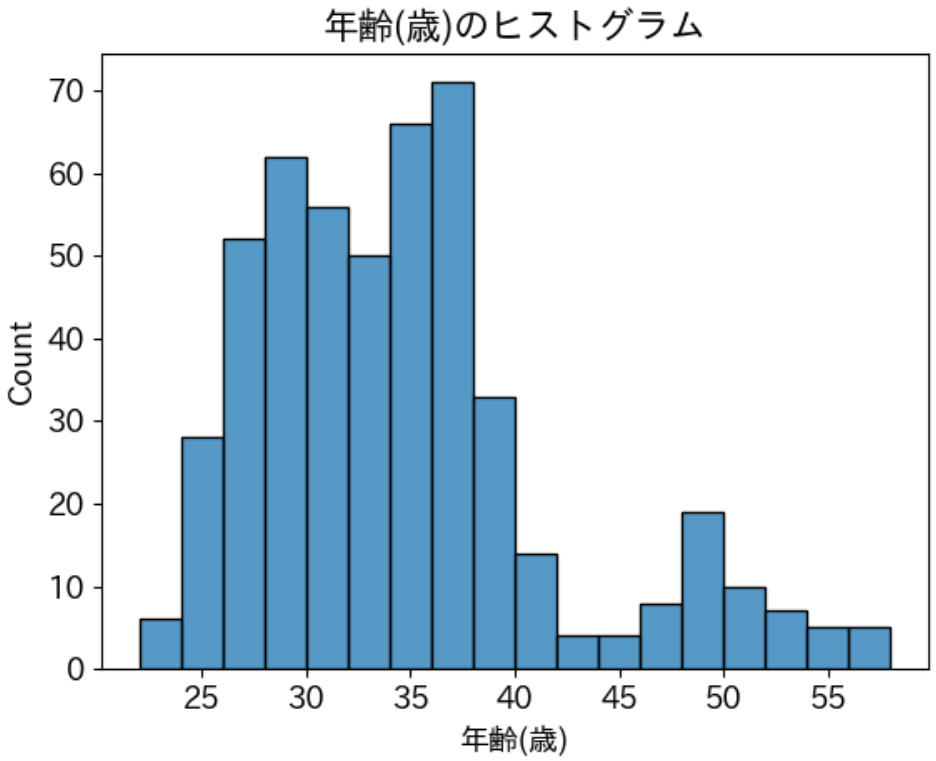
min,maxで明らかにおかしい値が無い事を確認。

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

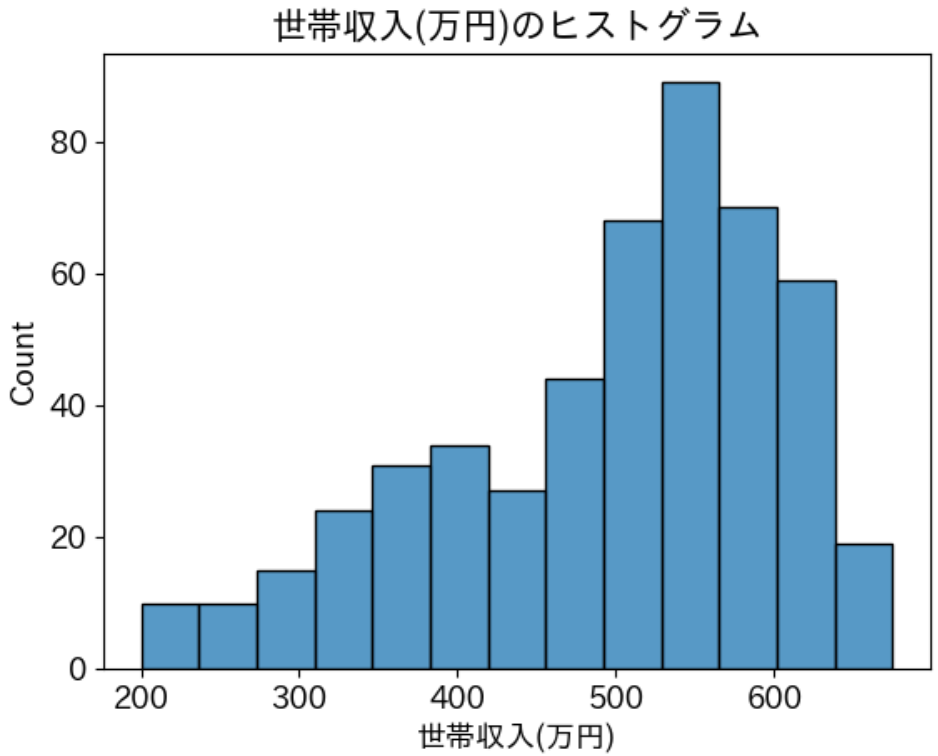
演習回答

演習②

・＜１変量分布＞各説明変数の分布はどのようになっているか？ヒストグラムで確認しよう。



20代後半～30代が多い。40代以上は件数が少ない。



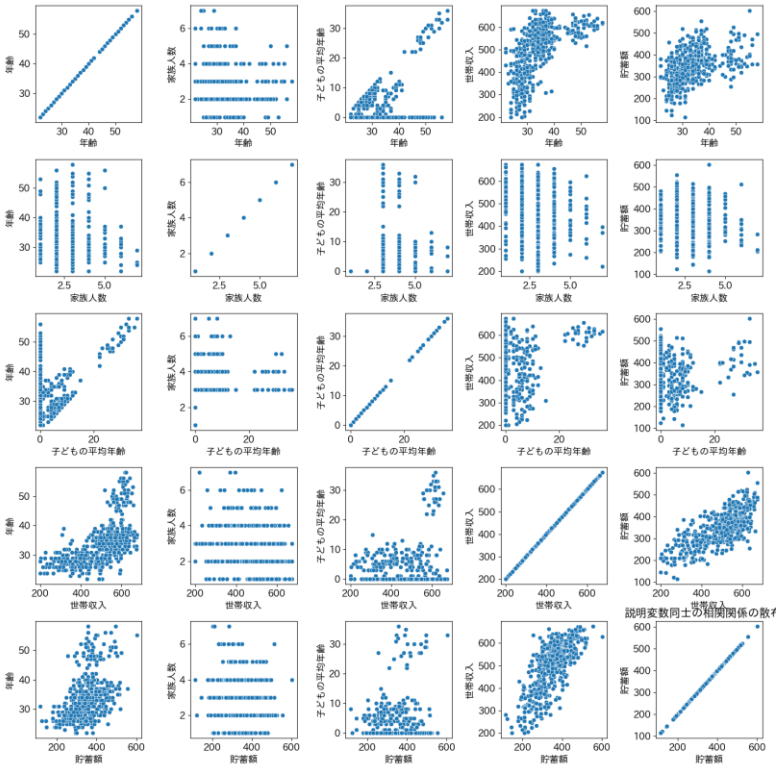
500～600万円が多い。

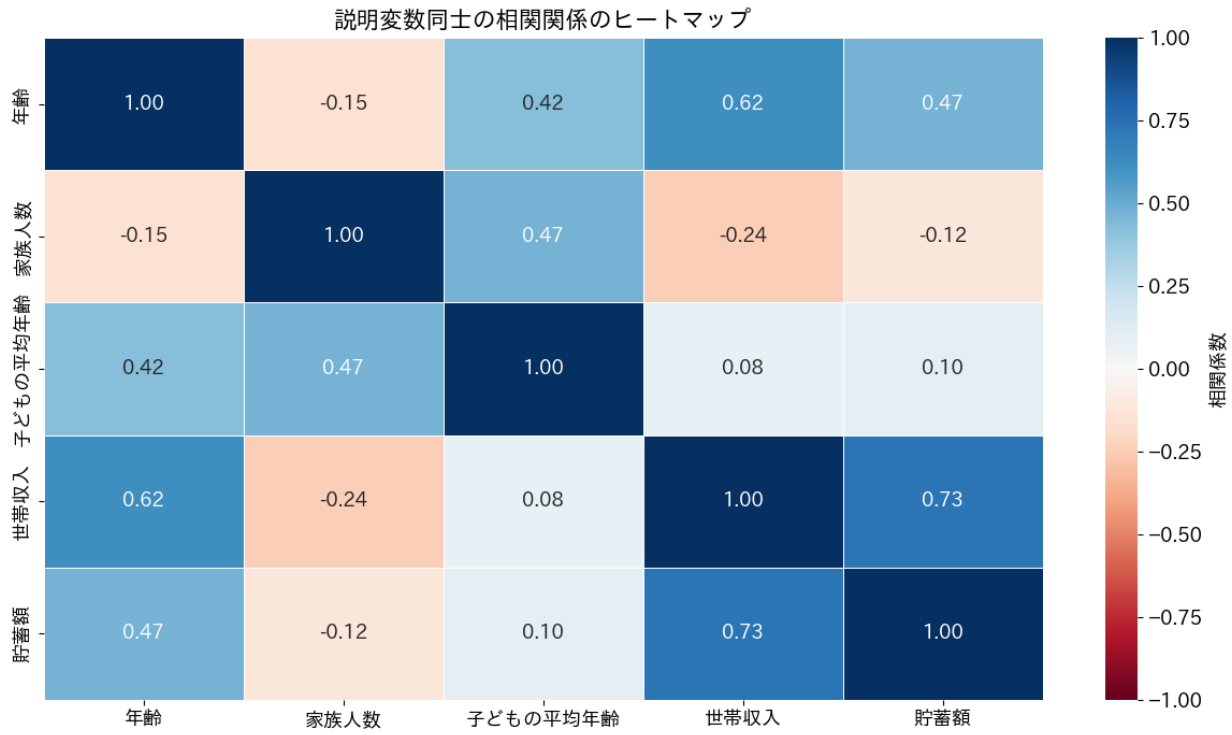
※あくまでダミーデータのため、現実の傾向を反映するものではありません。

演習回答

演習③

・各説明変数の関係を散布図とヒートマップで確認しよう。





散布図と相関係数は必ずセットで確認する。その理由はhttps://bellcurve.jp/statistics/course/9591.html

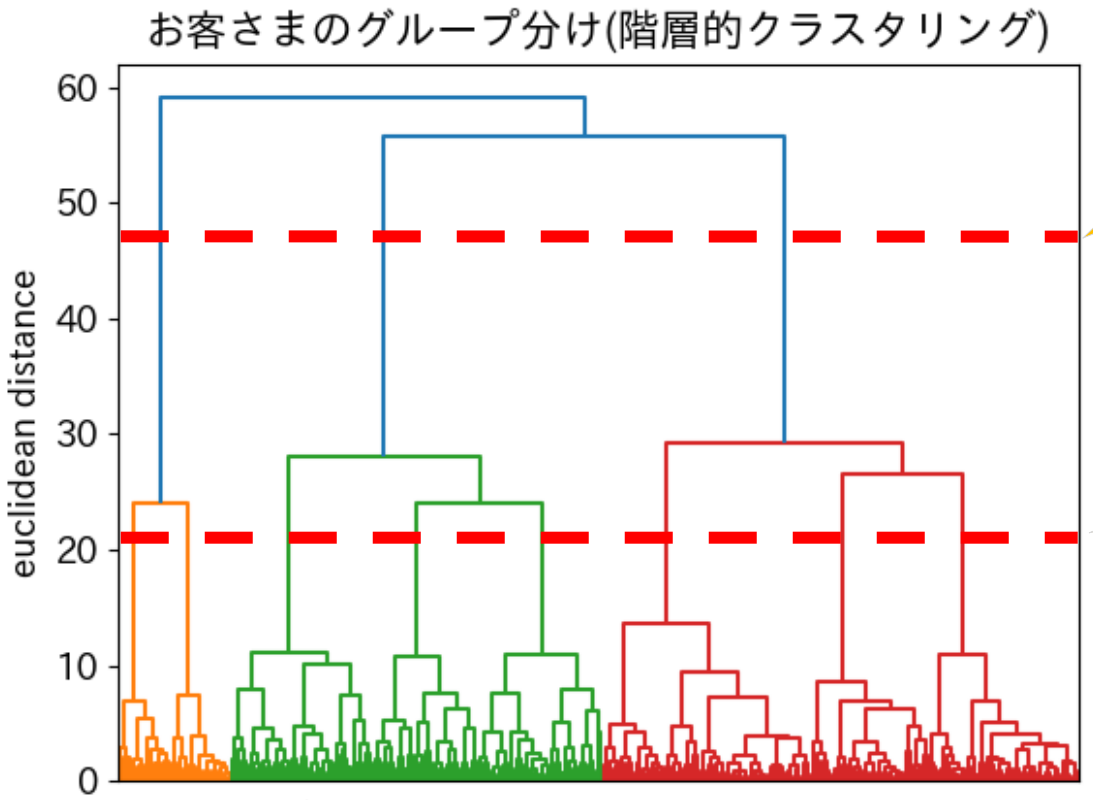
世帯年収と貯蓄額に正の相関あり、年齢と世帯収入に正の相関あり。いずれも直感通りの関係性。

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

演習回答

演習④

・各説明変数のスケールを揃えてから、階層的クラスタリングを実施し、何グループに分ければよさそうか考えてみよう。



3つにグルーピングするのが良さそう？

8つが良さそう？と考える人もいるかも

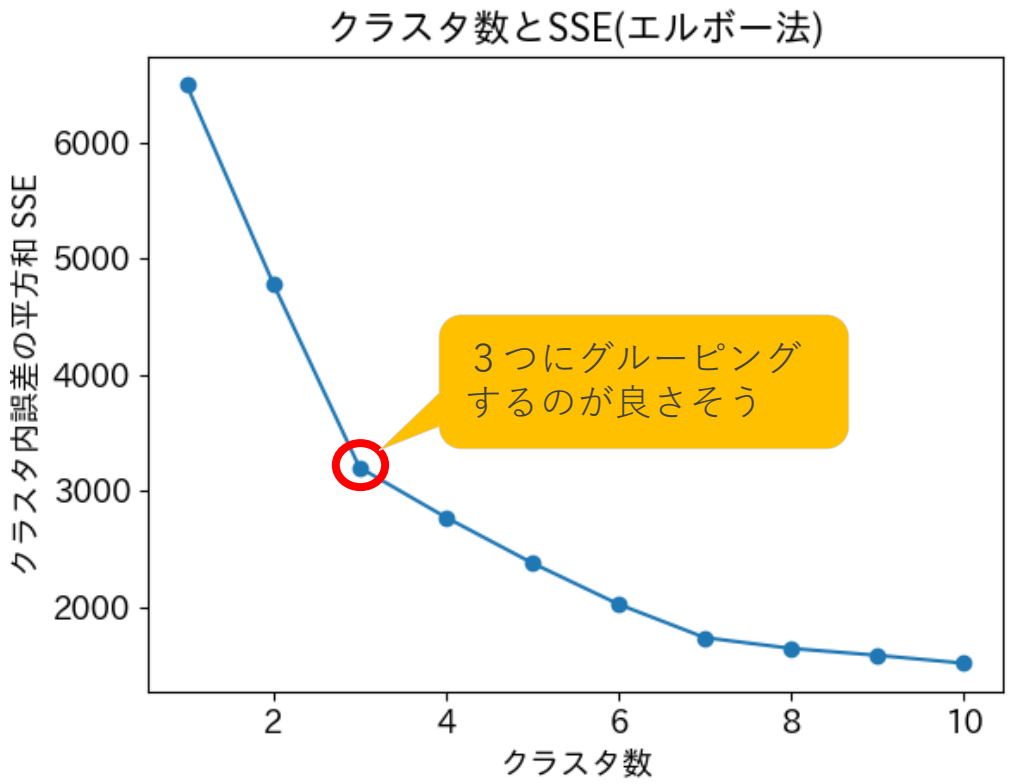
特にこれが正解などは無く、
実務では、業務要件に応じて
グループ数を決定することもある

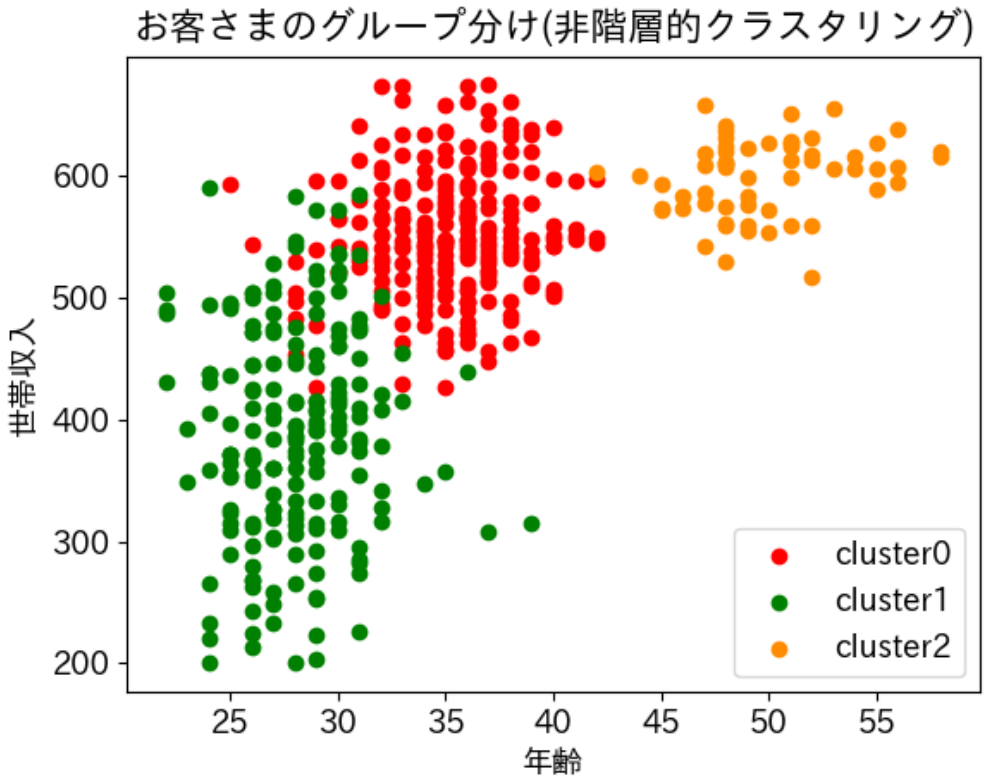
※あくまでダミーデータのため、現実の傾向を反映するものではありません。

演習回答

演習④

・各説明変数のスケールを揃えてから、非階層的クラスタリングを実施し、何グループに分ければよさそうか考えてみよう。



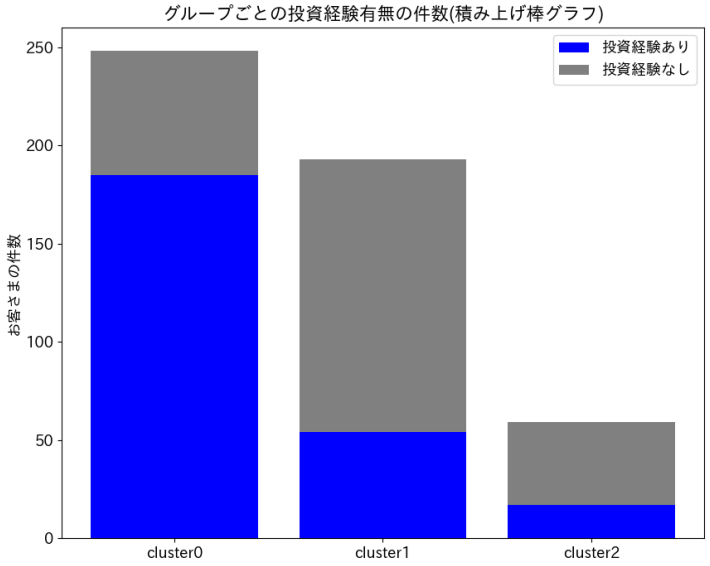


※あくまでダミーデータのため、現実の傾向を反映するものではありません。

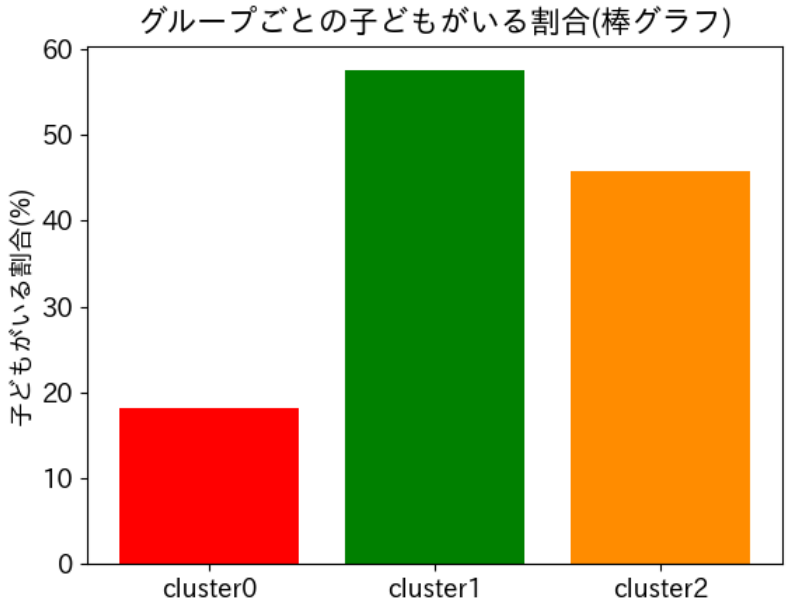
演習回答

演習Advanced

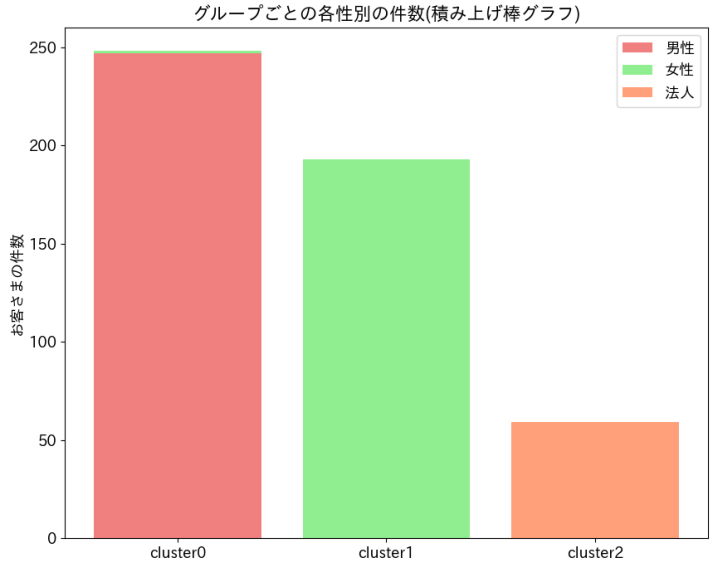
・非階層的クラスタリングで分けた３グループごとの特徴を可視化から解釈してみよう。



cluster0は投資経験ありの人が多い。



cluster1は子どもがいる人が多い。



cluster2は法人のお客さま。

※あくまでダミーデータのため、現実の傾向を反映するものではありません。

演習回答

演習Advanced

・解釈した結果から各グループへの提案内容を考えてみよう。

- cluster0への提案：自身の老後に備えるための資金形成を目的とした変額保険
  （投資経験ありの人が多いため、為替などのリスクを理解したうえでリターンが大きい保険を提案）

- cluster1への提案：子どもの将来や自身の生活習慣病に備える保険
  （自身の万が一にも備えつつ子どもの養育費にも備える保険を提案）

- cluster2への提案：自身が病気で不在になった場合に、会社の借入金などの返済に備える大型保険
  （経営者の不在は会社にとって大きな損失となるため保険金が大きい大型契約を提案）

※あくまでダミーデータのため、現実の傾向を反映するものではありません。