SASお客様向けwebinarシリーズ

機械学習によるビッグデータ分析の手法

#2 クラスター分析による分類（1）
非階層的クラスタリング

2021年12月8日

Agenda

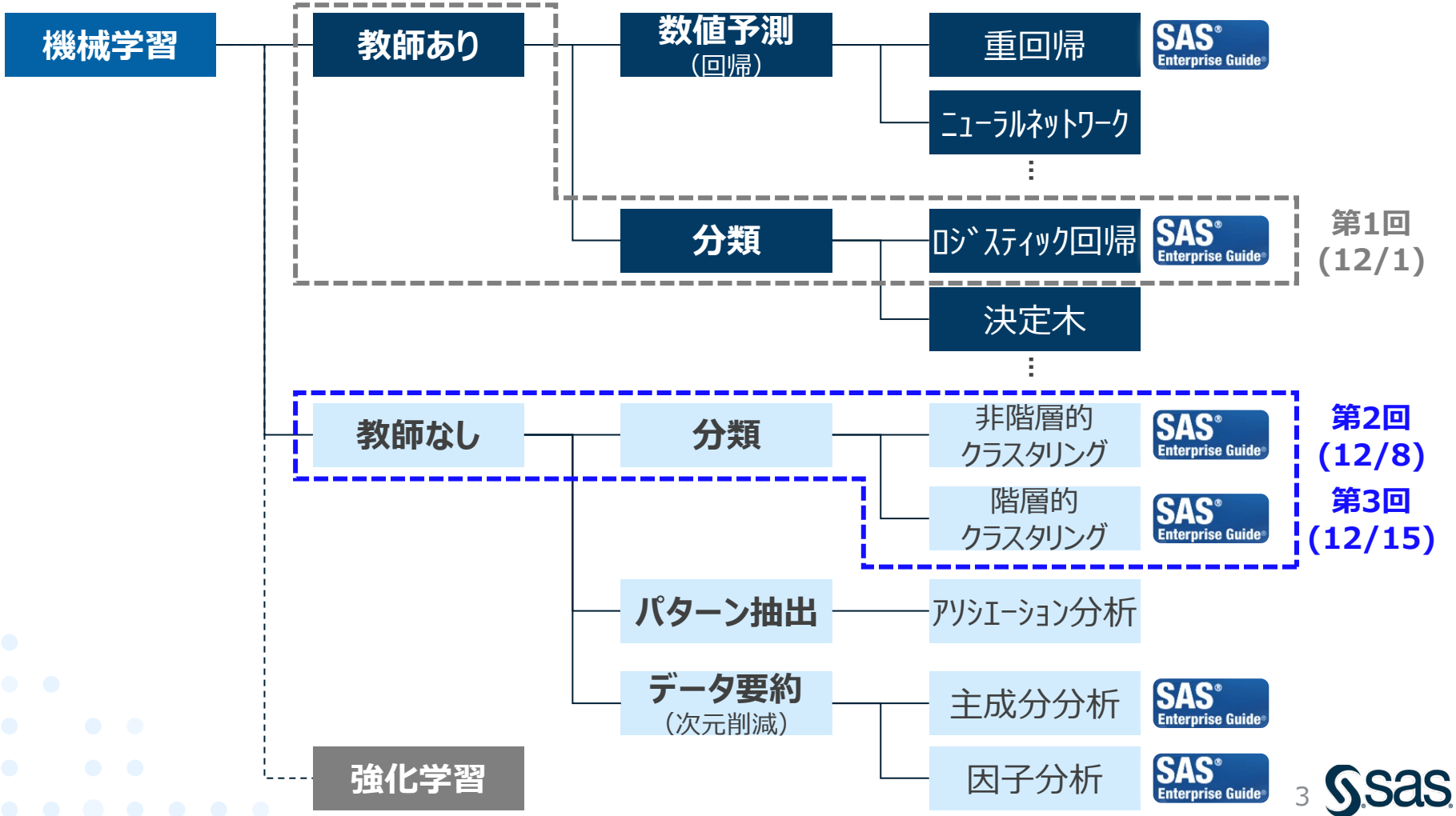
- 相関行列によるデータ観察
  - 相関関係の全体把握
  - 散布図行列との同時活用

- クラスター分析による分類（1）：非階層的クラスタリング
  - 教師なし学習とクラスタリング
  - 非階層的クラスタリング（k-means法）のしくみ
  - クラスタ数設定の考え方
  - 各クラスタの解釈方法
  - 顧客データを用いて非階層的クラスタリングにより類似顧客をグルーピングする

再掲

代表的な機械学習手法

- 機械学習手法は、教師あり、教師なし、強化学習に大別される
- なかでも、**教師あり分類、教師なし分類**は極めて基本的かつ頻用される手法である

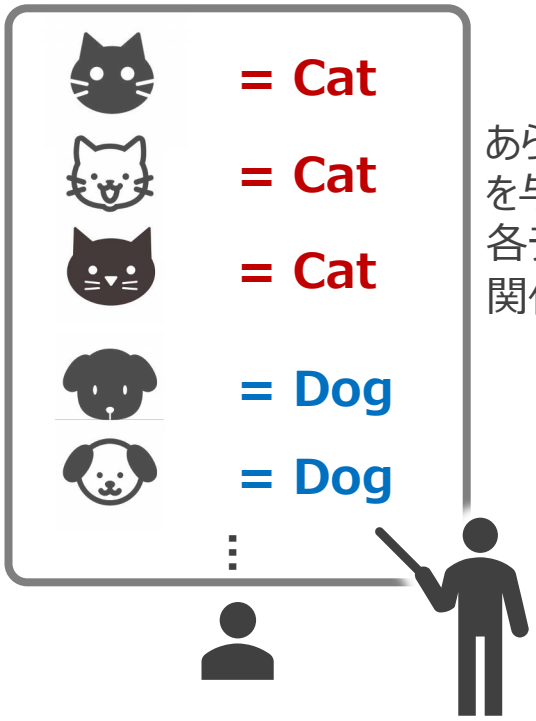


再掲

教師あり学習 と 教師なし学習

教師あり学習

学習



あらかじめ「正解」を与えて、
各データと正解の関係を学習させる

推論

教師なし学習

学習

「正解」を与えずに、
各データのパターン（距離の近さ、頻出の組み合わせなど）を
学習する

推論

※分類されたグループの意味づけは人が行う

教師なし学習のイメージ（クラスタリング）

- 各データ間の距離に基づき、近接データ（＝類似度が高いデータ）同士のグループ（**クラスタ**）を作り、データを分類する手法
- **学習データなし**でデータを大きく層別したい場合に有効

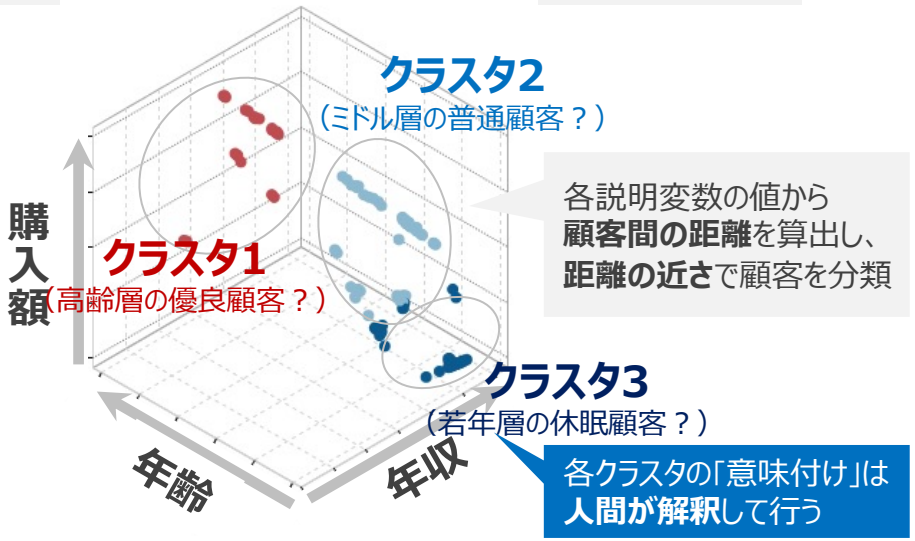
データ例

顧客ID	名前	年齢	年収	購入額	購入有無	…
0001	xx	25	300万	35,000	購入	…
0002	xx	35	600万	68,000	購入	…
0003	xx	18	120万	0	非購入	…
0004	xx	42	820万	85,000	購入	…
⋮	⋮	⋮	⋮	⋮	⋮	…

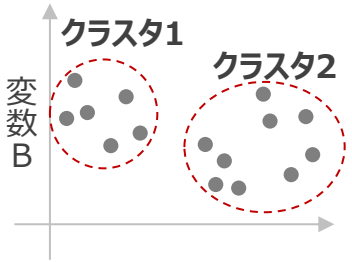
説明変数

※目的変数は無し

クラスタリング



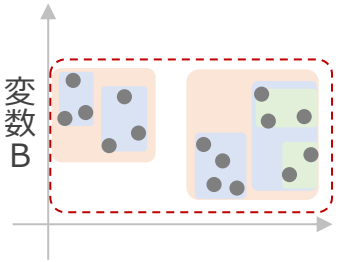
非階層的クラスタリング

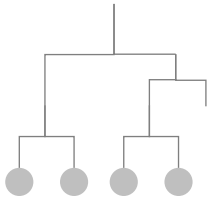


主な手法

- **k-means法**（k平均法）
- 混合ガウス

階層的クラスタリング



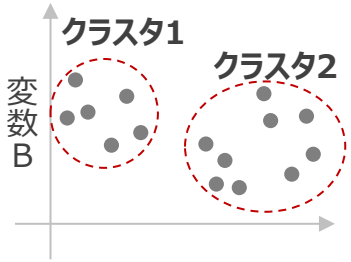
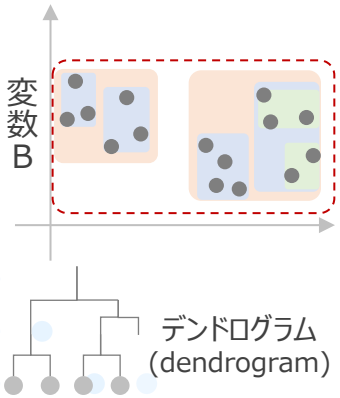


主な手法

- 最短距離法
- 最長距離法
- 群平均法
- ウォード法

クラスタリング手法の種類

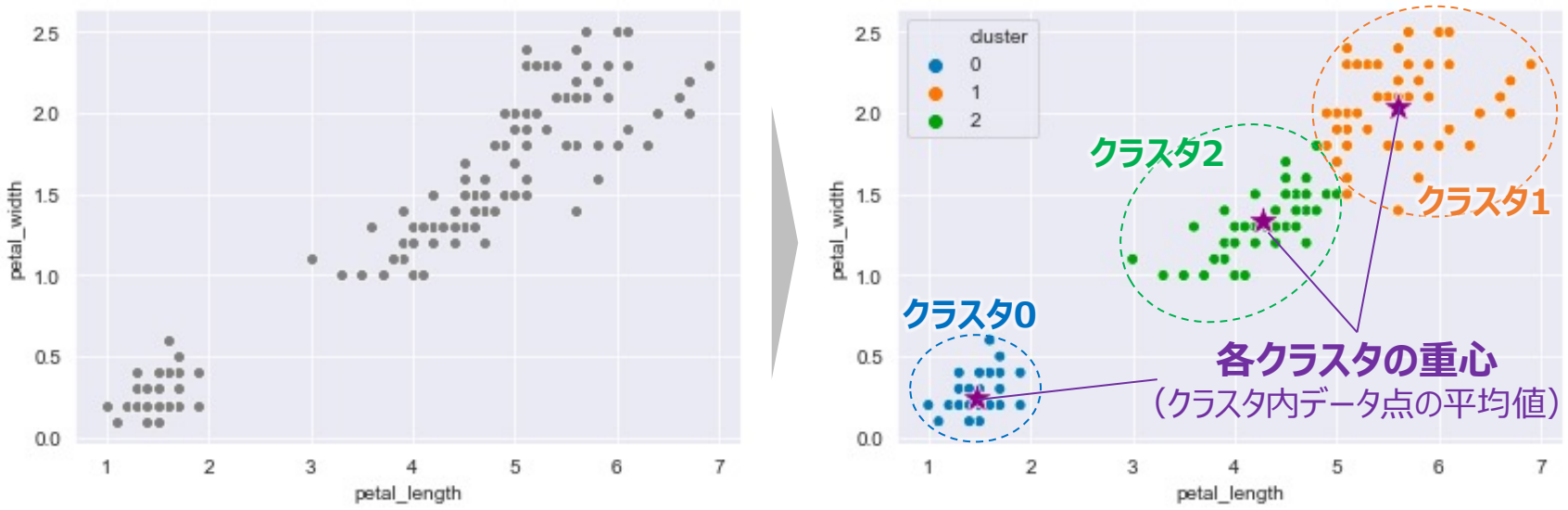
- クラスタリング手法は、「**非階層的**」と「**階層的**」に大別される
- 階層的クラスタリングはさらに 凝集型 と 分割型 があり、凝集型が用いられるのが一般的

手法の分類	手法		
非階層的クラスタリング 	▪ k-means法（k平均法）	クラスタ内データの平均値をクラスタ重心として、距離に基づき、事前に設定したクラスタ数k個に分割	SAS Enterprise Guide
	▪ その他	混合ガウス法、超体積法など	**本日ご説明**
階層的クラスタリング 	▪ ウォード法	クラスタ内のデータの平方和を最小にするように併合	SAS Enterprise Guide
	▪ 最短距離法（最近隣法）	距離の近いデータから順番に併合	**第3回 (12/15)**
	▪ 最長距離法（最遠隣法）	距離の遠いデータから順番に併合	
	▪ 重心法	クラスタ重心からの距離に基づき併合	SAS Enterprise Guide
	▪ 群平均法	各クラスタ同士で全データの距離の平均を基準に併合	SAS Enterprise Guide
	▪ その他	メディアン法、可変法	

非階層クラスタリング：k-means法

- クラスタリング手法の中で代表的かつ最もシンプルな手法が「**k-means法**」であり、各クラスタ内の**データ平均値（means）**を重心として、**k個のクラスター**に分類することができる

▼2次元のk-meansクラスタリング例



▼分類結果の特徴
- 教師なしのため、各クラスタの意味解釈は人が行う
- 円状（球状）のクラスタになりやすい
- クラスタサイズ（クラスタ内のデータ数）が同程度になりやすい

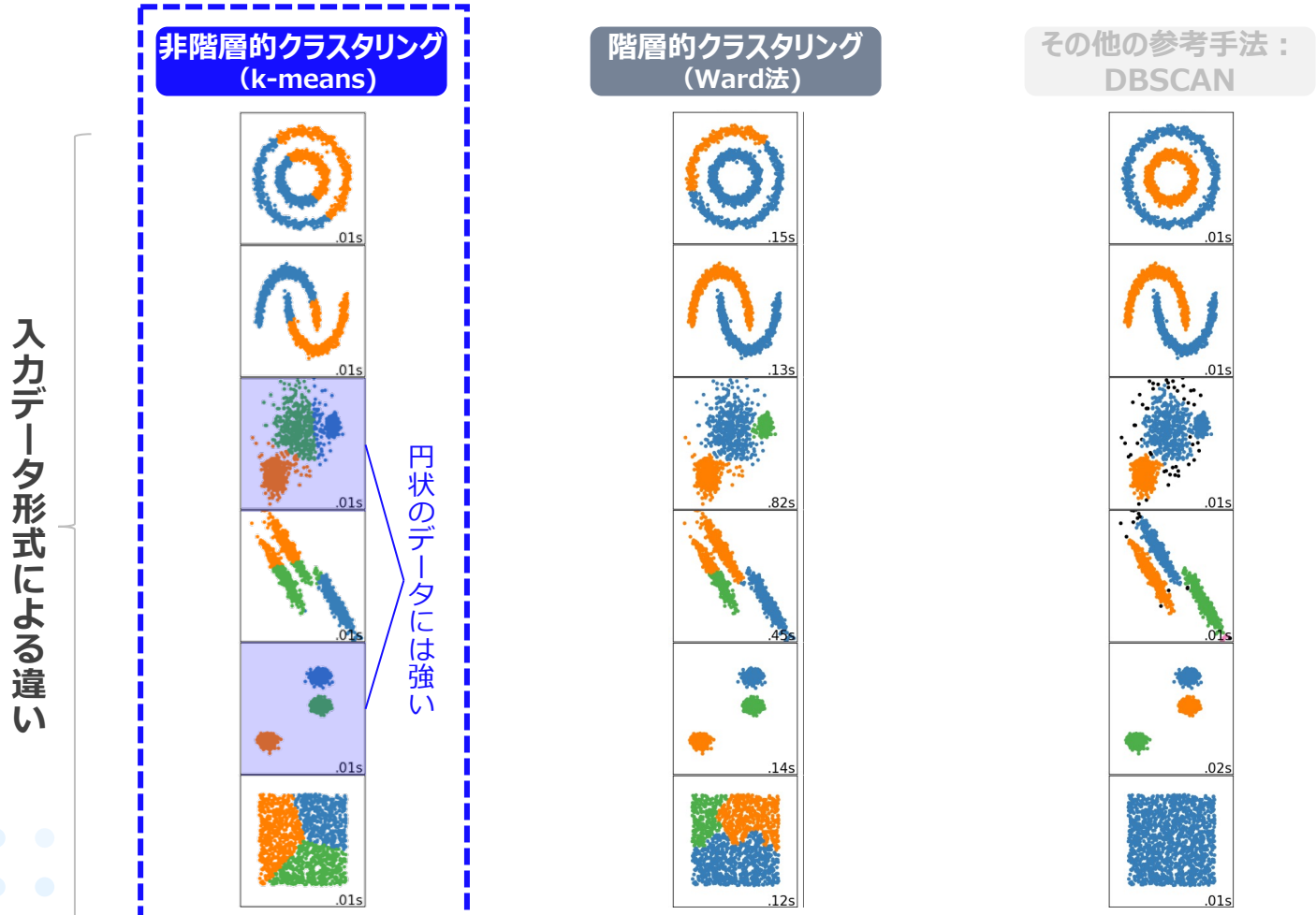
▼アルゴリズムの特徴
- クラスタ数を事前に明示的に決める必要がある
- 距離依存のため、データのスケールによって結果が変わる
- 初期値（初期重心）に大きく依存

参考：k-means法のイメージ（動画）

Source: https://www.youtube.com/watch?v=BVFG7fd1H30

参考：クラスタリング手法における分類結果の比較

- クラスタリング手法によって得意なデータパターンは異なり、様々な手法を試しながら、最適な手法を選択することが望ましい。中でも、k-meansは「重心からの距離」を用いて分類するため、円状のデータには強いが、楕円状や曲線状のデータは苦手



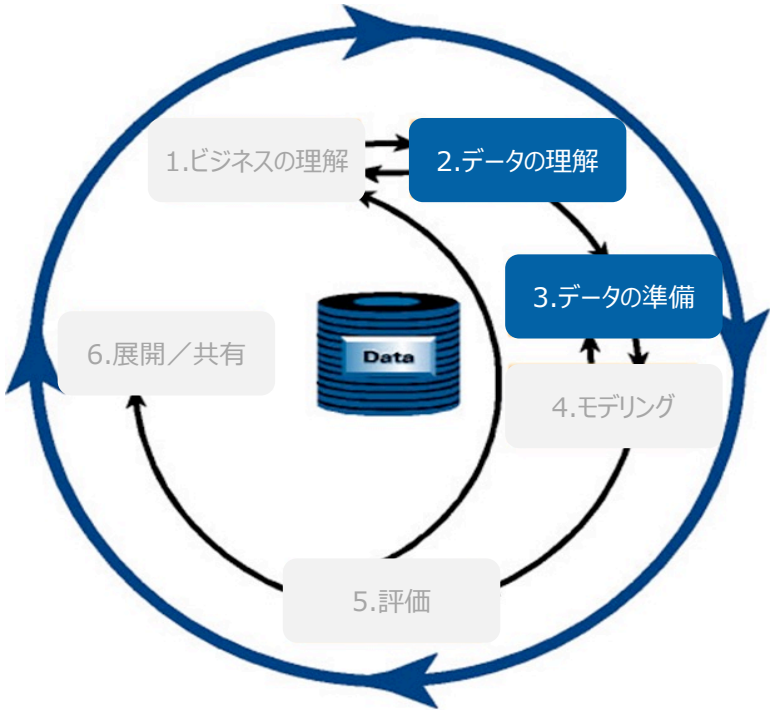
https://scikit-learn.org/stable/modules/clustering.html

再掲

ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)

1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル作成	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

再掲

使用データ

- **UCI Machine Learning Repository**では様々な分野のデータが公開
- 今回は、**銀行のマーケティングデータ**を活用し、分析を行う

Bank Marketing Data Set

Download: Data Folder, Data Set Description

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Characteristics:	Multivariate	**Number of Instances:**	45211	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	17	**Date Donated**	2012-02-14
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	1577437

Source:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was require
('yes') or not ('no') subscribed.

There are four datasets:
1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

https://archive.ics.uci.edu/ml/datasets/bank+marketing

再掲

データの概要

- 4,521人分の顧客について、顧客情報や営業アプローチ状況、最終的な狙いである「定期預金の契約有無」に関する情報（計17列）が格納されている

予測(分析)対象を説明するための変数

予測(分析)したい対象

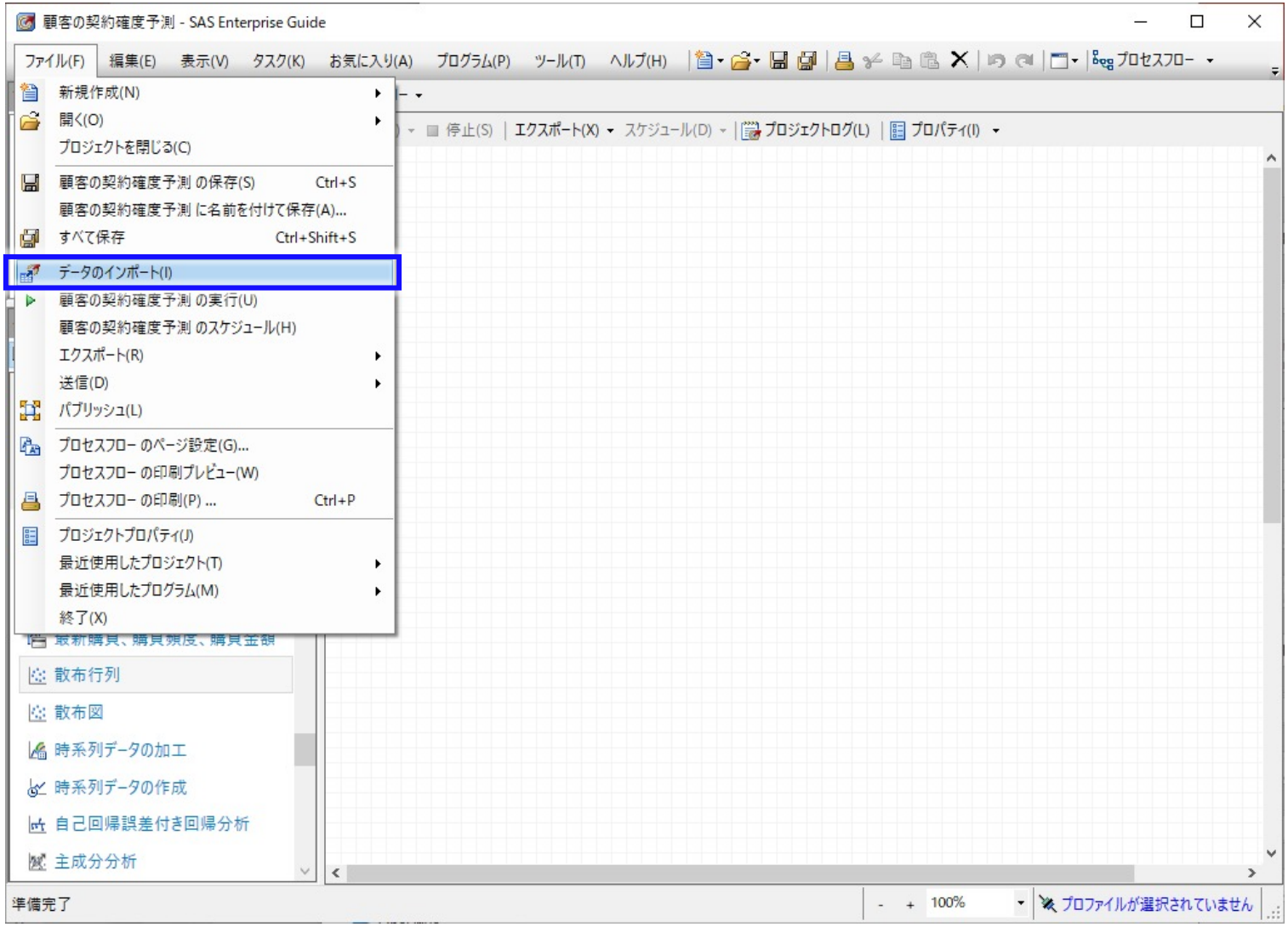
年齢	職業	結婚歴	学歴	クレジットカード債務不履行有無	年間平均残高（ユーロ）	住宅ローンの有無	個人ローンの有無	連絡手段	最終連絡日	最終連絡月	最終連絡時の会話時間（秒）	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数	前回キャンペーンの結果	定期預金契約有無
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes		cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	ye		nknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	y		llular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no		cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no		cellular	30	apr	261	1	-1	0	unknown	ye
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no
38	management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no
42	management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no
44	services	single	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no
44	entrepreneur	married	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no
26	housemaid	married	tertiary	no	543	no	no	cellular	30	jan	169	3	-1	0	unknown	no

説明変数

不使用

再掲

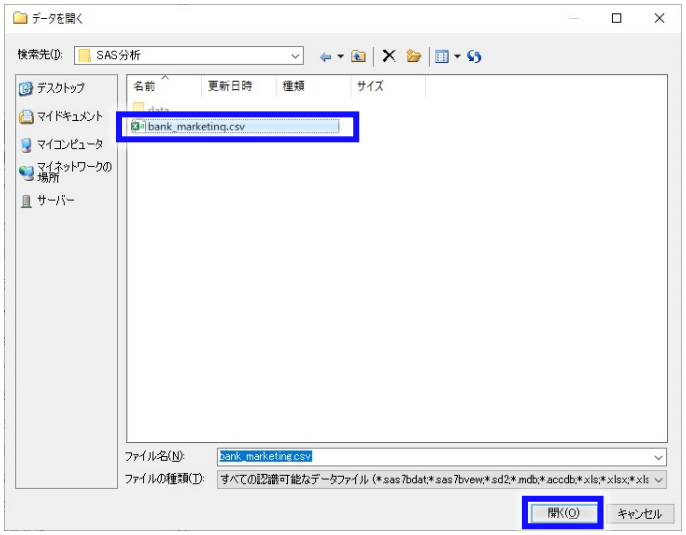
データの読み込み（1/2）

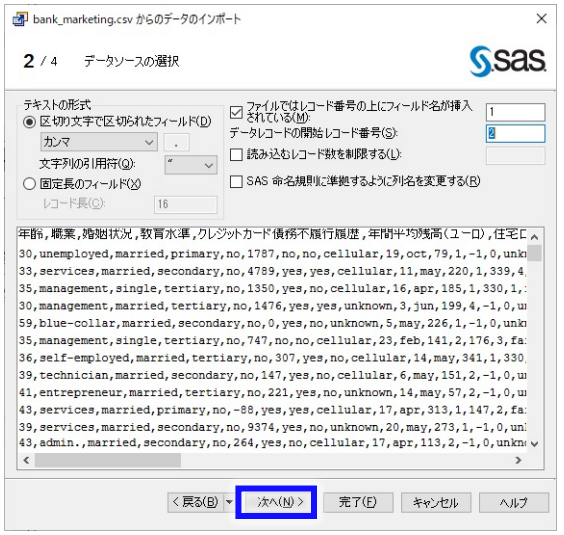


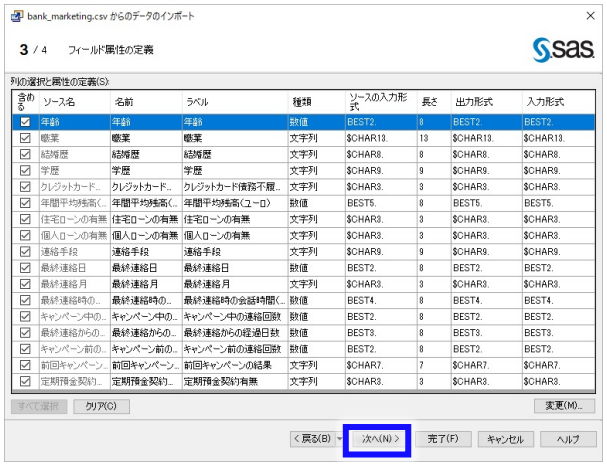
再掲

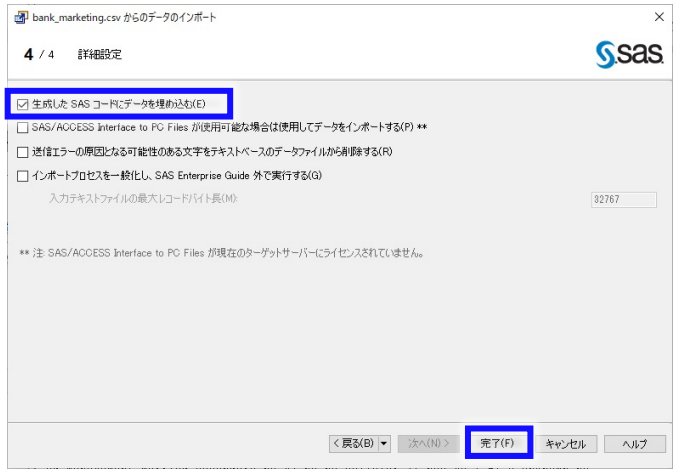
SAS Enterprise Guide

データの読み込み（2/2）





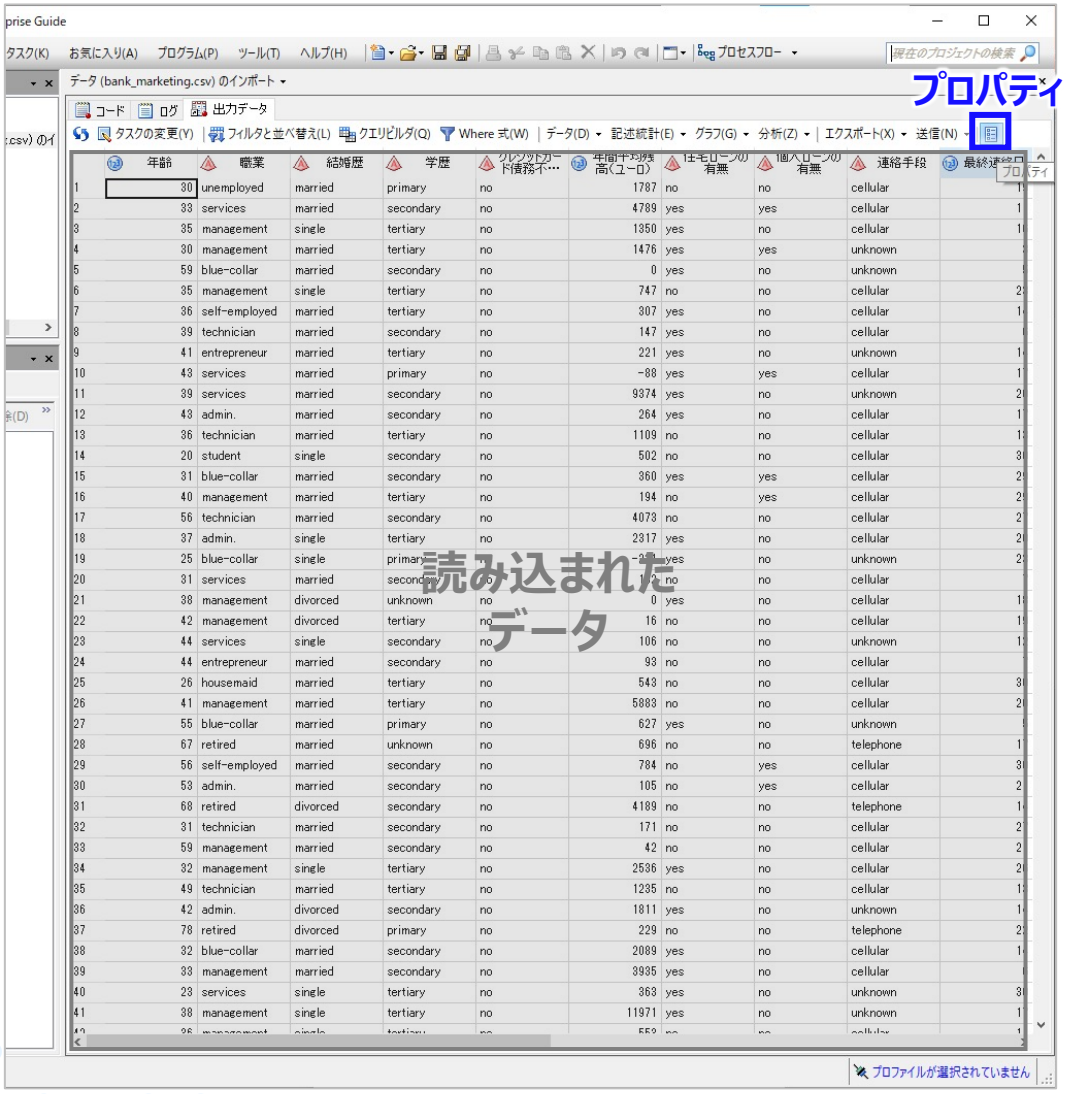


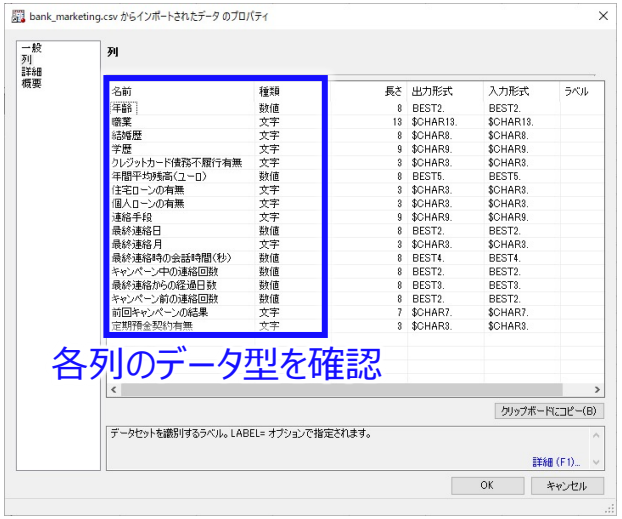


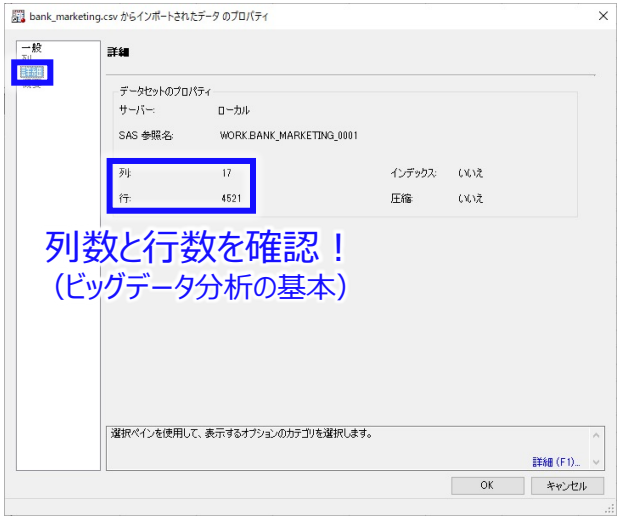
再掲

SAS Enterprise Guide

読み込んだデータの確認







再掲

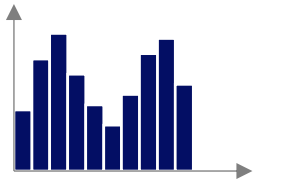
データの特徴の捉え方

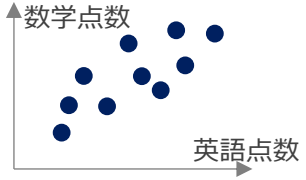
- ビッグデータでは個々のデータをくまなく見るのは難しいため、**グラフ**（ヒストグラムや散布図）や**要約統計量**（平均値や標準偏差） を用いて全体傾向を把握する

グラフ化

視覚的にデータの特徴や傾向を把握

ID	英語	数学	…
1	55	48	
2	70	47	
3	66	44	
:	:	:	
18	65	55	
19	57	51	
20	68	43	





散布図

：

数値化（データ要約）

データの特徴を示す値に要約し
比較可能な客観的傾向を掴む
（**要約統計量**）

ID	英語	数学	…
1	55	48	
2	70	47	
3	66	44	
:	:	:	
18	65	55	
19	57	51	
20	68	43	

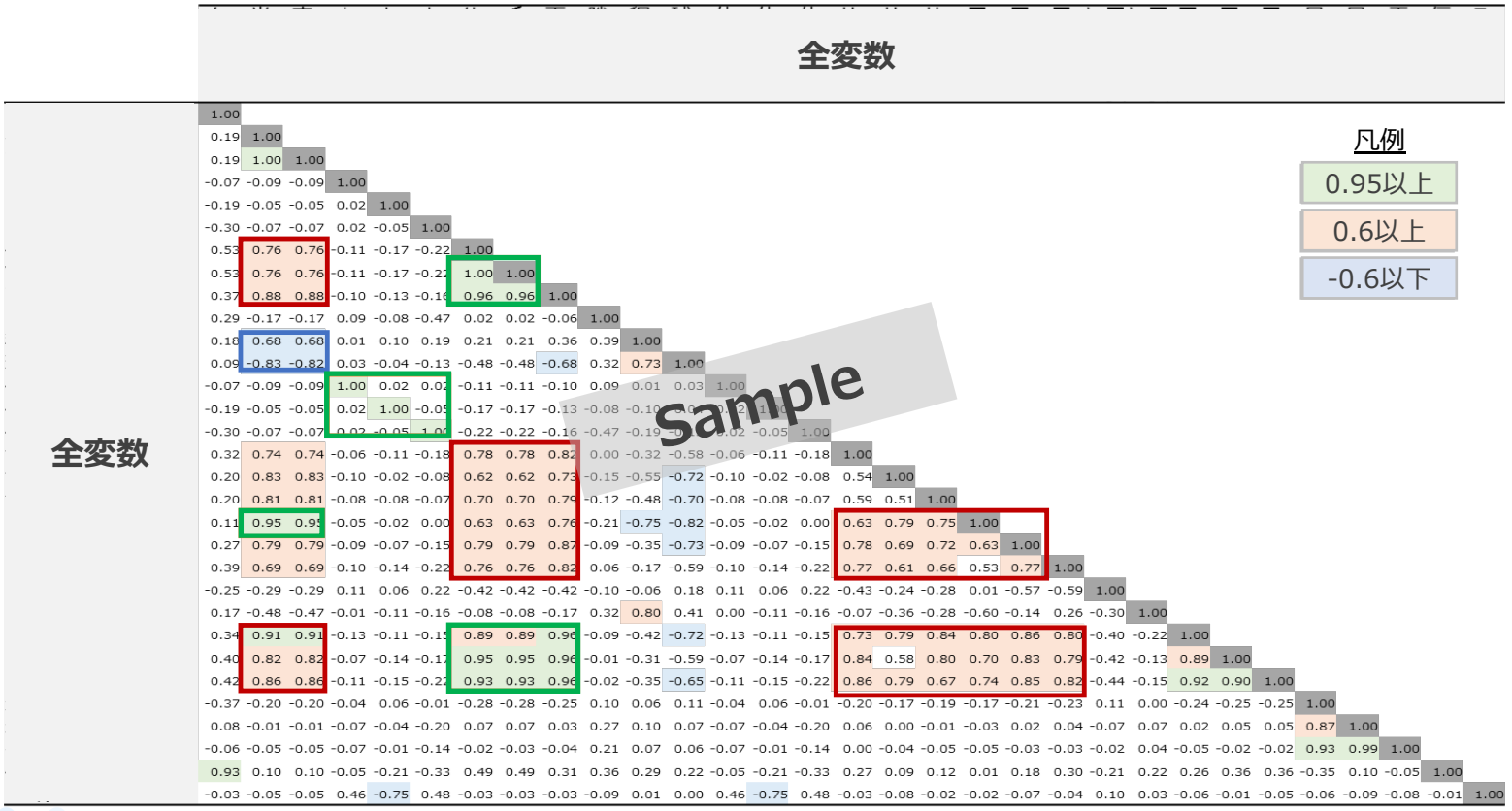
- 平均値＝ＸＸＸ
- 中央値＝ＸＸＸ
- 標準偏差＝ＸＸＸ

：

再掲

相関行列

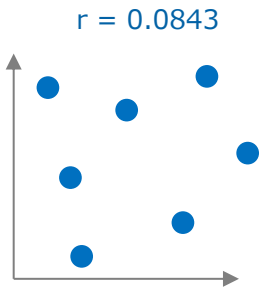
- 事前に各変数間の相関係数を総当たりで調べておくと、後々の結果解釈に役立つ（**相関行列**）
- また、**共線性が高い変数**（相関の高い）が複数混ざっていると、その変数の影響を強く受け、偏った分析結果になることがある。この場合、共線性が高い変数は除外することが有効



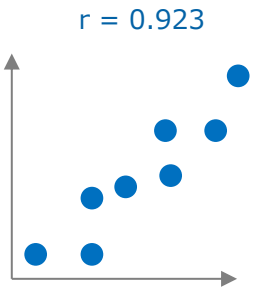
相関係数について

- 相関係数r (correlation coefficient) とは、2つの変数間の相関の度合いを表す指標
- -1 ≦ r ≦ 1 の値を取り、正の場合は正相関、負の場合は負相関、0の場合は無相関

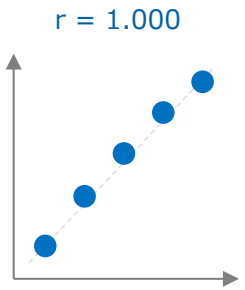
正の相関



かなり低い正の相関

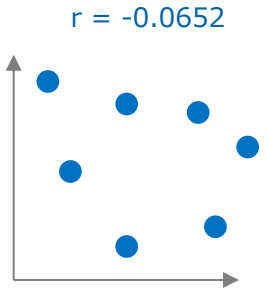


強い正の相関

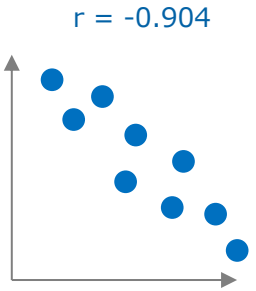


完全相関

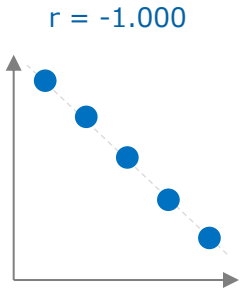
負の相関



かなり低い負の相関



強い負の相関



完全相関

（参考）**相関**係数の目安

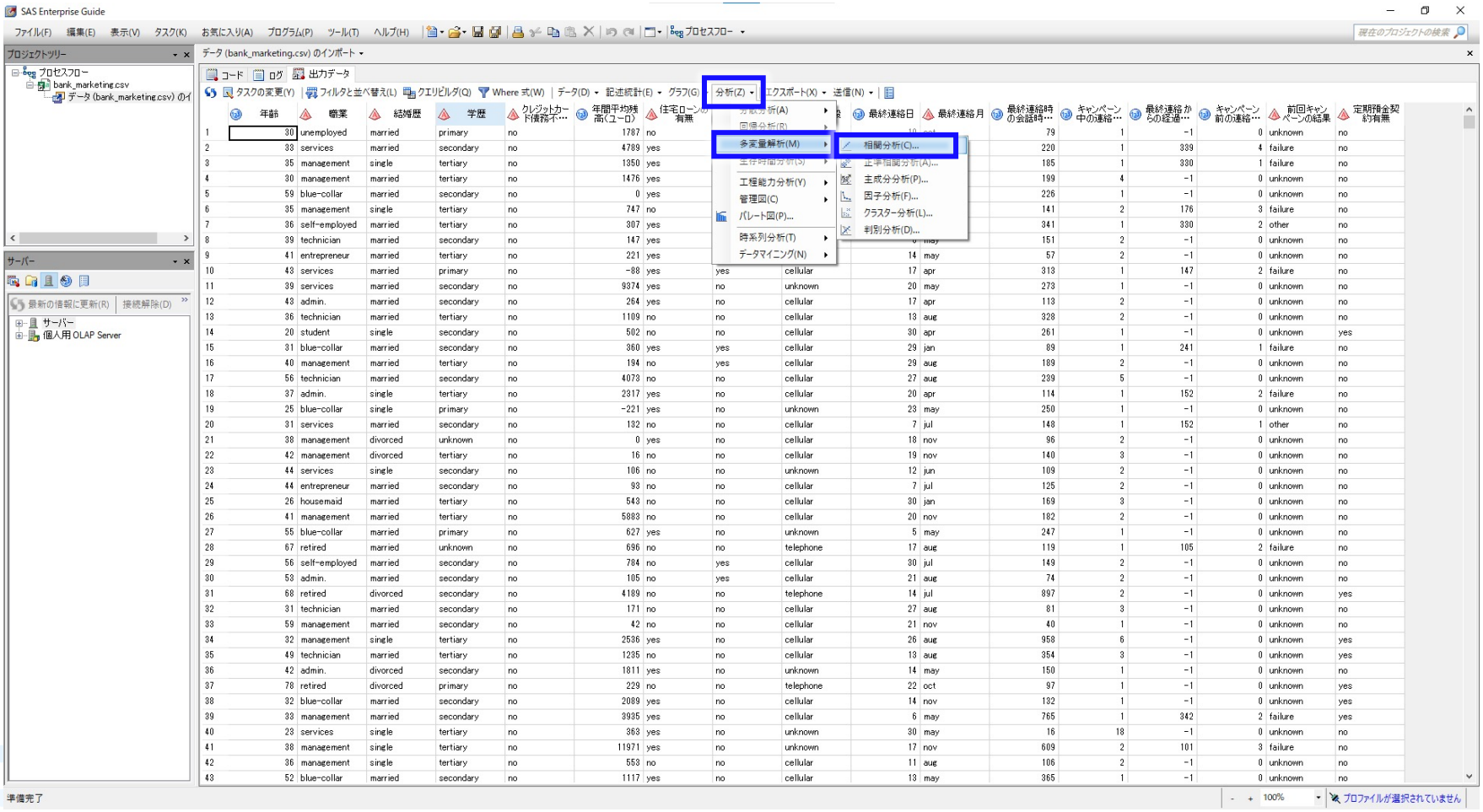
相関係数（絶対値）	相関の強弱
1.0	**強い相関**
0.7	**やや相関あり**
0.4	**弱い相関**
0.2	**ほぼ相関なし**
0	

SAS Enterprise Guide での実装方法

- 相関分析
- グループ変数を設定した相関分析

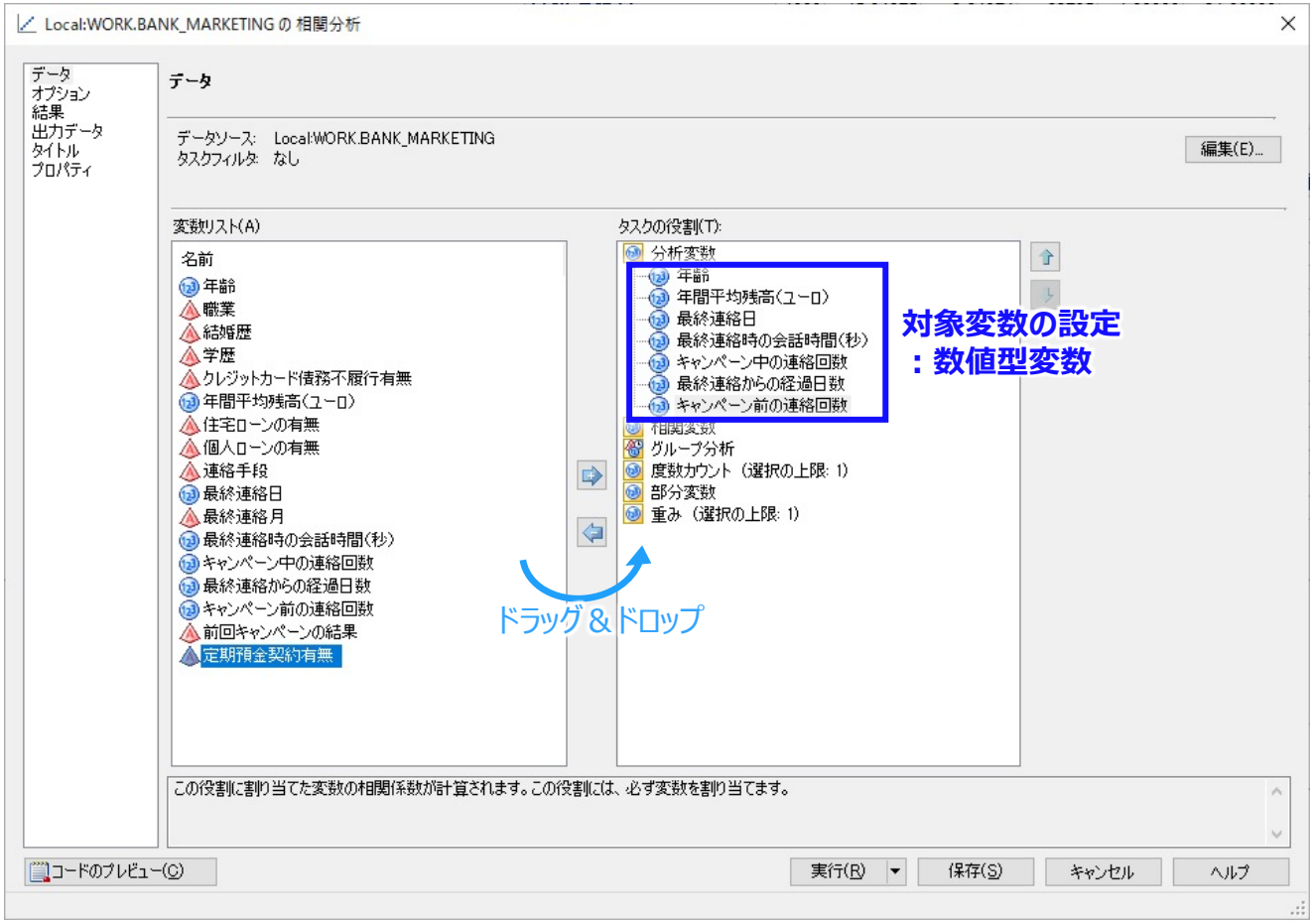
相関分析の出力 – 実行方法（1/2）

[分析] → [多変量解析] → [相関分析] をクリック



相関分析の出力 – 実行方法（2/2）

左側の [変数リスト] より、数値型の変数を選択し、
右側の [分析変数] にドラッグ＆ドロップ



相関分析の出力 – 実行結果（1/3）

相関分析

CORR プロシジャ

7 変数：年齢　年間平均残高(ユーロ)　最終連絡日　最終連絡時の会話時間(秒)　キャンペーン中の連絡回数　最終連絡からの経過日数　キャンペーン前の連絡回数

単純統計量

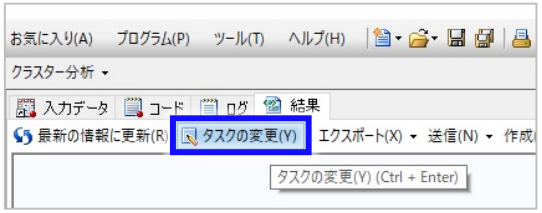
変数	N	平均	標準偏差	合計	最小値	最大値
年齢	4521	41.17010	10.57621	186130	19.00000	87.00000
年間平均残高(ユーロ)	4521	1423	3010	6431836	-3313	71188
最終連絡日	4521	15.91528	8.24767	71953	1.00000	31.00000
最終連絡時の会話時間(秒)	4521	263.96129	259.85663	1193369	4.00000	3025
キャンペーン中の連絡回数	4521	2.79363	3.10981	12630	1.00000	50.00000
最終連絡からの経過日数	4521	39.76664	100.12112	179785	-1.00000	871.00000
キャンペーン前の連絡回数	4521	0.54258	1.69356	2453	0	25.00000

Pearson の相関係数, N = 4521
H0: Rho=0 に対する Prob > |r|

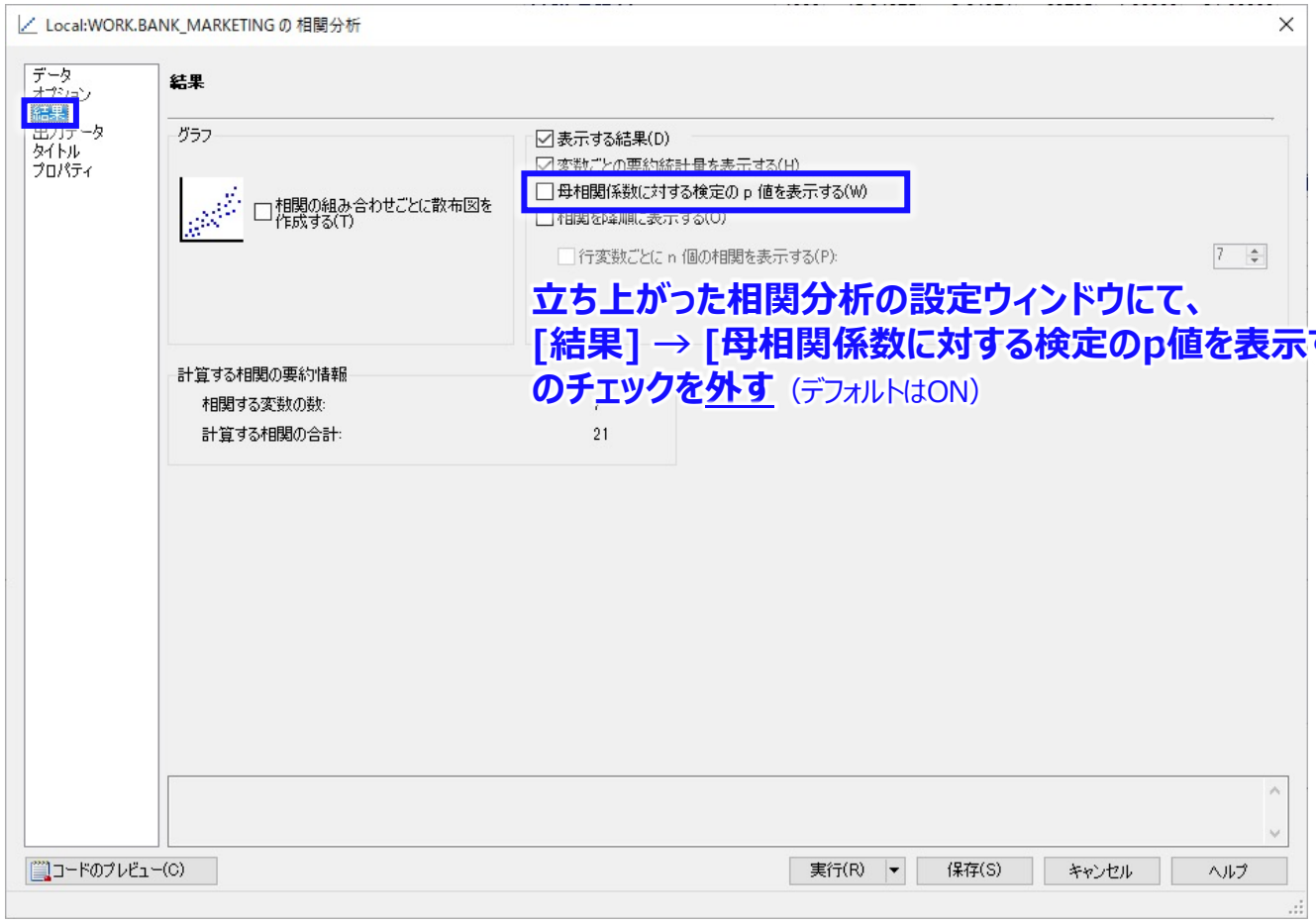
	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
年齢	1.00000	0.08382	-0.01785	-0.00237	-0.00515	-0.00889	-0.00351
		<.0001	0.2301	0.8736	0.7293	0.5500	0.8134
年間平均残高(ユーロ)	0.08382	1.00000	-0.00868	-0.01595	-0.00998	0.00944	0.02620
	<.0001		0.5597	0.2836	0.5025	0.5259	0.0782
最終連絡日	-0.01785	-0.00868	1.00000	-0.02463	0.16071	-0.09435	-0.05911
	0.2301	0.5597		0.0978	<.0001	<.0001	<.0001
最終連絡時の会話時間(秒)	-0.00237	-0.01595	-0.02463	1.00000	-0.06838	0.01038	0.01808
	0.8736	0.2836	0.0978		<.0001	0.4853	0.2242
キャンペーン中の連絡回数	-0.00515	-0.00998	0.16071	-0.06838	1.00000	-0.09314	-0.06783
	0.7293	0.5025	<.0001	<.0001		<.0001	<.0001
最終連絡からの経過日数	-0.00889	0.00944	-0.09435	0.01038	-0.09314	1.00000	0.57756
	0.5500	0.5259	<.0001	0.4853	<.0001		<.0001
キャンペーン前の連絡回数	-0.00351	0.02620	-0.05911	0.01808	-0.06783	0.57756	1.00000
	0.8134	0.0782	<.0001	0.2242	<.0001	<.0001	

P値が有益な情報となることもあるが、全体俯瞰したい時には、情報量が多い
→ P値の省略が可能（次頁）

相関分析の出力 – 実行結果（2/3）



分析結果の画面上において
上部メニュー [タスクの変更] をクリック



立ち上がった相関分析の設定ウィンドウにて、
[結果] → [母相関係数に対する検定のp値を表示する]
のチェックを外す（デフォルトはON）

相関分析の出力 – 実行結果（3/3）

相関分析

CORR プロシジャ

7 変数：年齢　年間平均残高(ユーロ)　最終連絡日　最終連絡時の会話時間(秒)　キャンペーン中の連絡回数　最終連絡からの経過日数　キャンペーン前の連絡回数

単純統計量

変数	N	平均	標準偏差	合計	最小値	最大値
年齢	4521	41.17010	10.57621	186130	19.00000	87.00000
年間平均残高(ユーロ)	4521	1423	3010	6431836	-3313	71188
最終連絡日	4521	15.91528	8.24767	71953	1.00000	31.00000
最終連絡時の会話時間(秒)	4521	263.96129	259.85663	1193369	4.00000	3025
キャンペーン中の連絡回数	4521	2.79363	3.10981	12630	1.00000	50.00000
最終連絡からの経過日数	4521	39.76664	100.12112	179785	-1.00000	871.00000
キャンペーン前の連絡回数	4521	0.54258	1.69356	2453	0	25.00000

Pearson の相関係数, N = 4521

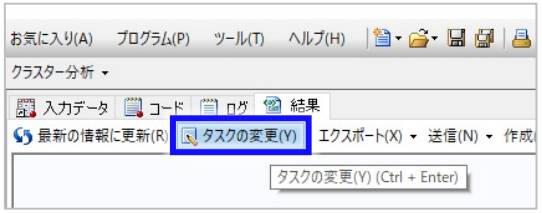
	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
年齢	1.00000	0.08382	-0.01785	-0.00237	-0.00515	-0.00889	-0.00351
年間平均残高(ユーロ)	0.08382	1.00000	-0.00868	-0.01595	0.00998	0.00944	0.02620
最終連絡日	-0.01785	-0.00868	1.00000	-0.02463	0.16071	-0.09435	-0.05911
最終連絡時の会話時間(秒)	-0.00237	-0.01595	-0.02463	1.00000	-0.06838	0.01038	0.01808
キャンペーン中の連絡回数	-0.00515	-0.00998	0.16071	-0.06838	1.00000	-0.09314	-0.06783
最終連絡からの経過日数	-0.00889	0.00944	-0.09435	0.01038	-0.09314	1.00000	0.57756
キャンペーン前の連絡回数	-0.00351	0.02620	-0.05911	0.01808	-0.06783	0.57756	1.00000

右上部分と左下部分は対称的

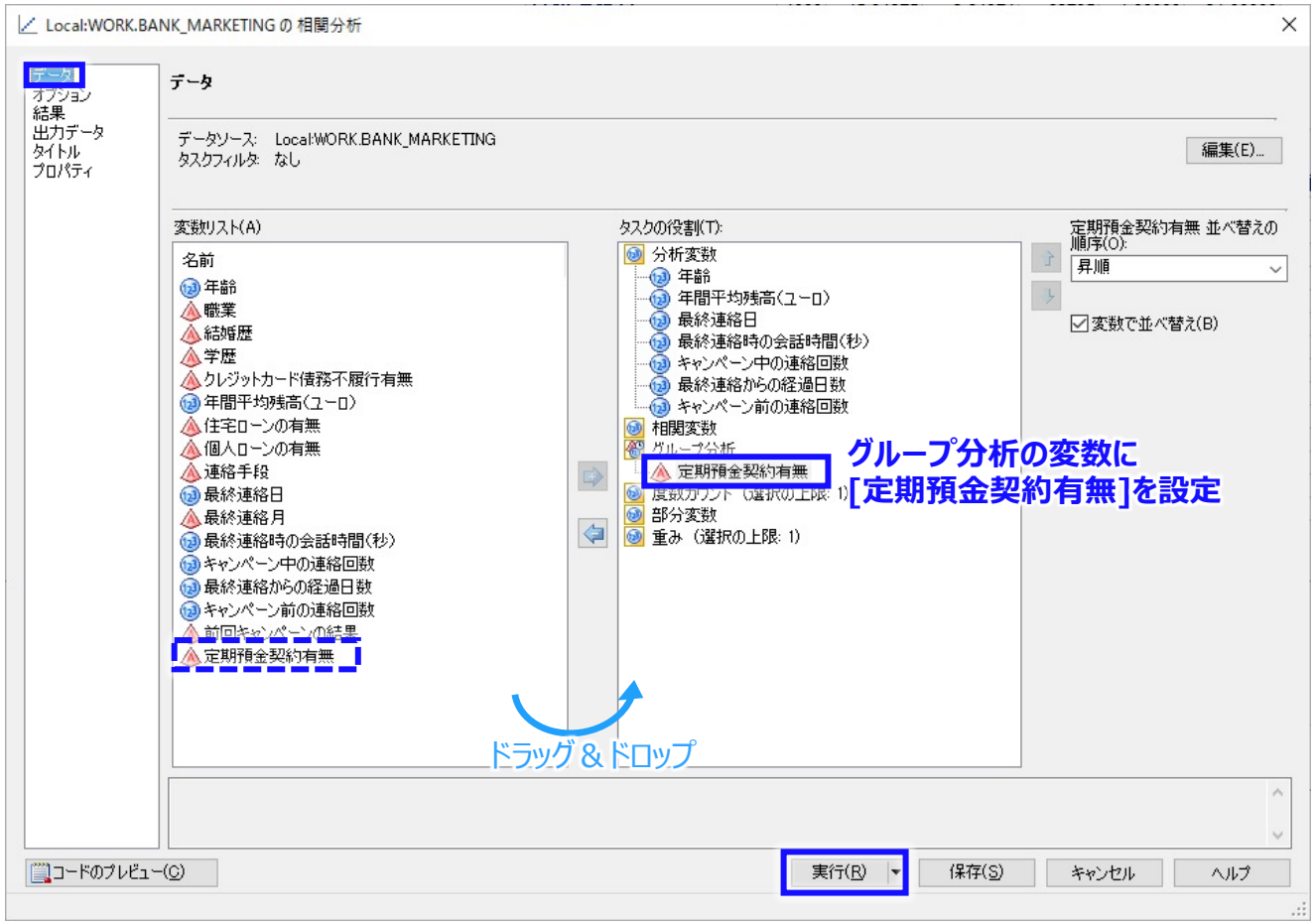
（参考）**相関**係数の目安

相関係数(絶対値)	相関の強弱
1.0	強い相関
0.7	やや相関あり
0.4	弱い相関
0.2	ほぼ相関なし
0	

相関分析の出力：グループ変数の設定 – 実行方法



分析結果の画面上において
上部メニュー [タスクの変更] をクリック



グループ分析の変数に
[定期預金契約有無]を設定

相関分析の出力：グループ変数の設定 – 実行結果

契約者

相関分析

CORR プロシジャ

定期預金契約有無=yes

7 変数：年齢　年間平均残高(ユーロ)　最終連絡日　最終連絡時の会話時間(秒)　キャンペーン中の連絡回数　最終連絡からの経過日数　キャンペーン前の連絡回数

単純統計量

変数	N	平均	標準偏差	合計	最小値	最大値
年齢	521	42.49136	13.11577	22138	19.00000	87.00000
年間平均残高(ユーロ)	521	1572	2444	818989	-1206	26965
最終連絡日	521	15.65835	8.23515	8158	1.00000	31.00000
最終連絡時の会話時間(秒)	521	552.74280	390.32580	287979	30.00000	2769
キャンペーン中の連絡回数	521	2.26679	2.09207	1181	1.00000	24.00000
最終連絡からの経過日数	521	68.63916	121.96306	35761	-1.00000	804.00000
キャンペーン前の連絡回数	521	1.09021	2.05537	568.00000	0	14.00000

Pearson の相関係数, N = 521

	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
年齢	1.00000	[illegible]	[illegible]	[illegible]	[illegible]	0.05072	-0.01192
年間平均残高(ユーロ)	0.16845	[illegible]	[illegible]	[illegible]	[illegible]04	0.01352	0.02050
最終連絡日	-0.05207	[illegible]	[illegible]	[illegible]	[illegible]80	-0.03734	-0.05123
最終連絡時の会話時間(秒)	-0.03633	-0.12007	0.03610	1.00000	0.23432	-0.15489	-0.15549
キャンペーン中の連絡回数	-0.06583	-0.02804	0.13780	0.23432	1.00000	-0.08488	-0.09863
最終連絡からの経過日数	0.05072	0.01352	-0.03734	-0.15489	-0.08488	1.00000	0.51823
キャンペーン前の連絡回数	-0.01192	0.02050	-0.05123	-0.15549	-0.09863	0.51823	1.00000

契約者は、「キャンペーン中の連絡回数」と、「最終連絡時の会話時間」に弱い相関

（参考）相関係数の目安

相関係数(絶対値)	相関の強弱
1.0	
	強い相関
0.7	
	やや相関あり
0.4	
	弱い相関
0.2	
	ほぼ相関なし
0	

未契約者

定期預金契約有無=no

7 変数：年齢　年間平均残高(ユーロ)　最終連絡日　最終連絡時の会話時間(秒)　キャンペーン中の連絡回数　最終連絡からの経過日数　キャンペーン前の連絡回数

単純統計量

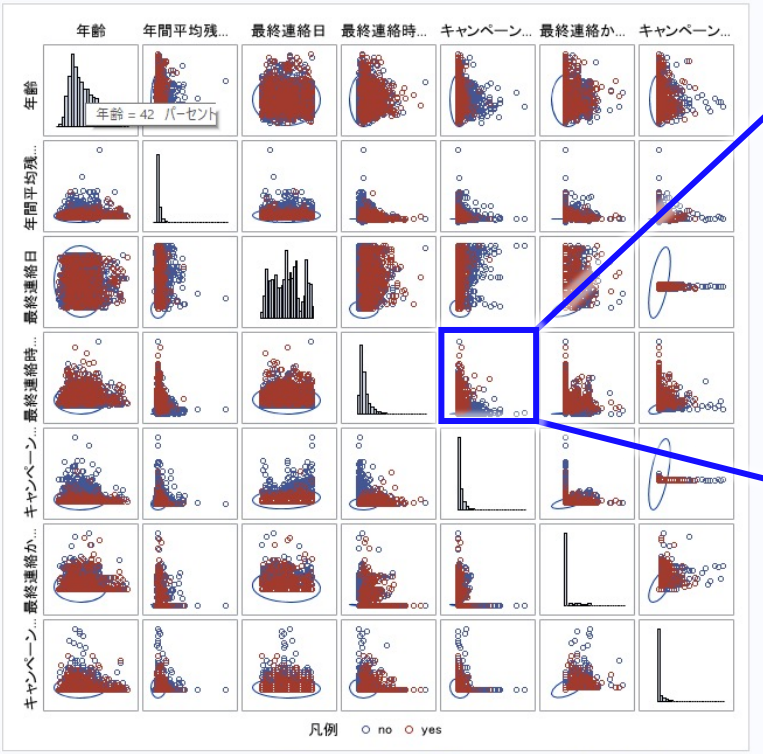
変数	N	平均	標準偏差	合計	最小値	最大値
年齢	4000	40.99800	10.18840	163992	19.00000	86.00000
年間平均残高(ユーロ)	4000	1403	3075	5612847	-3313	71188
最終連絡日	4000	15.94875	8.24974	63795	1.00000	31.00000
最終連絡時の会話時間(秒)	4000	226.34750	210.31363	905390	4.00000	3025
キャンペーン中の連絡回数	4000	2.86225	3.21261	11449	1.00000	50.00000
最終連絡からの経過日数	4000	36.00600	96.29766	144024	-1.00000	871.00000
キャンペーン前の連絡回数	4000	0.47125	1.62737	1885	0	25.00000

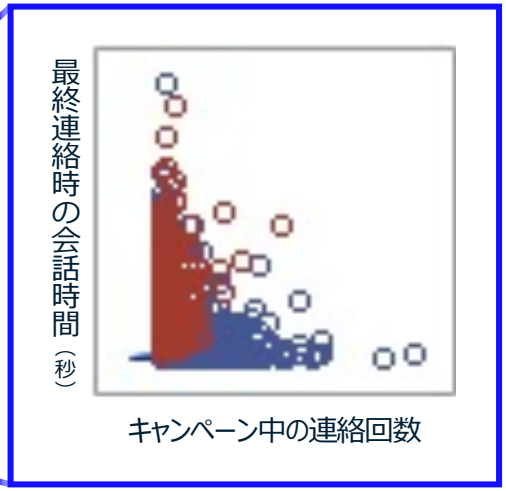
Pearson の相関係数, N = 4000

	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
年齢	1.00000	0.07291	-0.01165	-0.01836	0.00446	-0.02733	-0.00819
年間平均残高(ユーロ)	0.07291	1.00000	-0.00539	-0.00858	-0.00762	0.00694	0.02507
最終連絡日	-0.01165	-0.00539	1.00000	-0.03679	0.16339	-0.10334	-0.05957
最終連絡時の会話時間(秒)	-0.01836	-0.00858	-0.03679	1.00000	-0.09611	0.00179	0.00563
キャンペーン中の連絡回数	0.00446	-0.00762	0.16339	-0.09611	1.00000	-0.08961	-0.05856
最終連絡からの経過日数	-0.02733	0.00694	-0.10334	0.00179	-0.08961	1.00000	0.58368
キャンペーン前の連絡回数	-0.00819	0.02507	-0.05957	0.00563	-0.05856	0.58368	1.00000

散布図行列（層別）との比較

散布図行列　※前回出力した散布図





散布図と相関係数を併せて確認することで
より深い洞察や、外れ値の発見につながる

契約者

Pearson の相関係数, N = 521

	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
年齢	1.00000	0.16845	-0.05207	-0.03633	-0.06583	0.05072	-0.01192
年間平均残高(ユーロ)	0.16845	1.00000	-0.03858	-0.12007	-0.02804	0.01352	0.02050
最終連絡日	-0.05207	-0.03858	1.00000	0.03610	0.13780	-0.03734	-0.05123
最終連絡時の会話時間(秒)	-0.03633	-0.12007	0.03610	1.00000	0.23432	-0.15489	-0.15549
キャンペーン中の連絡回数	-0.06583	-0.02804	0.13780	0.23432	1.00000	-0.08488	-0.09863
最終連絡からの経過日数	0.05072	0.01352	-0.03734	-0.15489	-0.08488	1.00000	0.51823
キャンペーン前の連絡回数	-0.01192	0.02050	-0.05123	-0.15549	-0.09863	0.51823	1.00000

未契約者

Pearson の相関係数, N = 4000

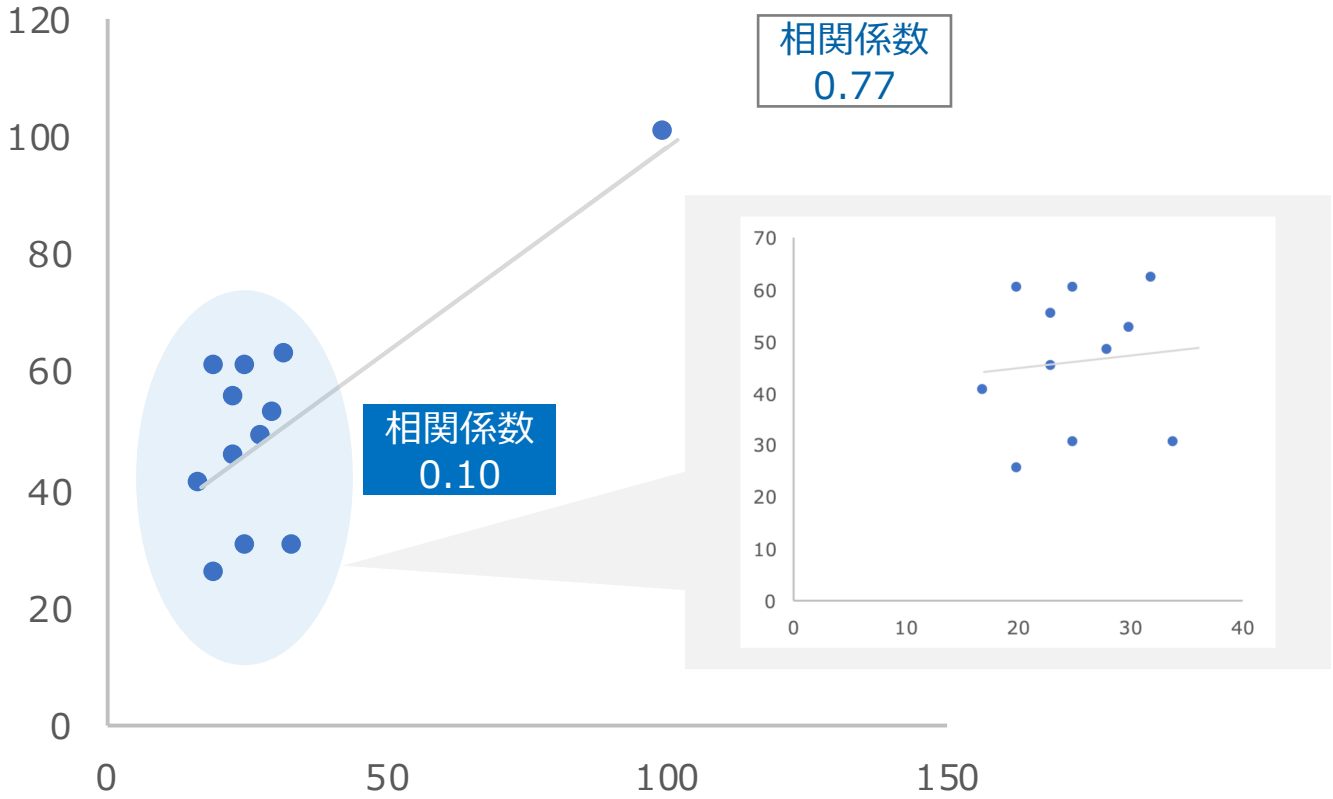
	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
年齢	1.00000	0.07291	-0.01165	-0.01836	0.00446	-0.02733	-0.00819
年間平均残高(ユーロ)	0.07291	1.00000	-0.00539	-0.00858	-0.00762	0.00694	0.02507
最終連絡日	-0.01165	-0.00539	1.00000	-0.03679	0.16339	-0.10334	-0.05957
最終連絡時の会話時間(秒)	-0.01836	-0.00858	-0.03679	1.00000	-0.09611	0.00179	0.00563
キャンペーン中の連絡回数	0.00446	-0.00762	0.16339	-0.09611	1.00000	-0.08961	-0.05856
最終連絡からの経過日数	-0.02733	0.00694	-0.10334	0.00179	-0.08961	1.00000	0.58368
キャンペーン前の連絡回数	-0.00819	0.02507	-0.05957	0.00563	-0.05856	0.58368	1.00000

（参考）**相関**係数の目安

相関係数（絶対値）	相関の強弱
1.0	
	強い相関
0.7	
	やや相関あり
0.4	
	弱い相関
0.2	
	ほぼ相関なし
0	

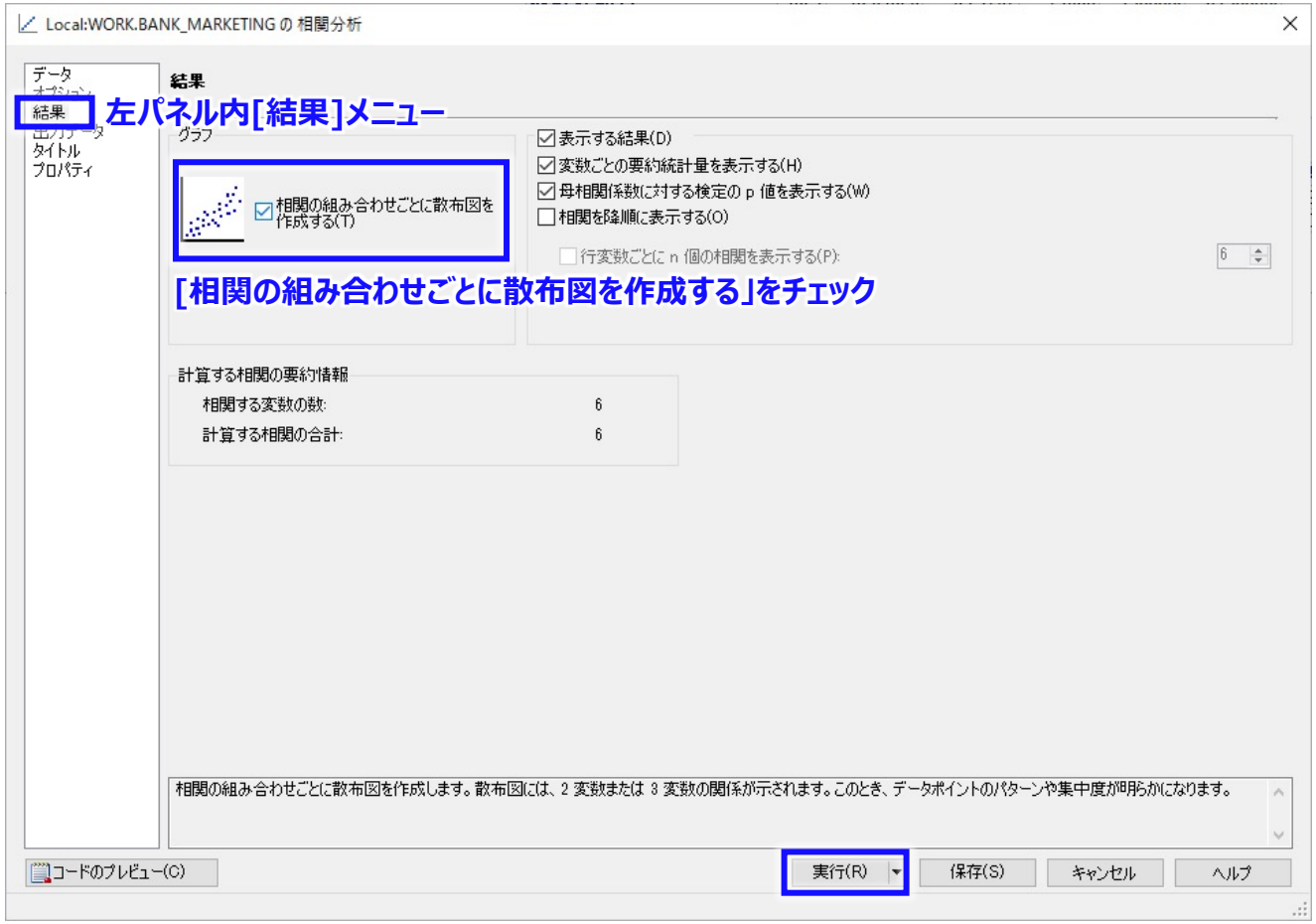
（参考）外れ値の影響例

- 相関係数は外れ値の影響を大きく受けるため、数字だけに惑わされぬよう、散布図の確認も併せて行うことが重要である



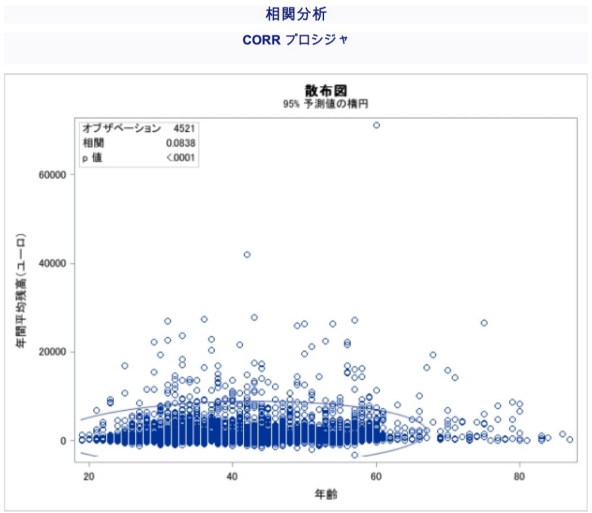
（参考）相関分析の出力：散布図の同時出力 – 実行方法

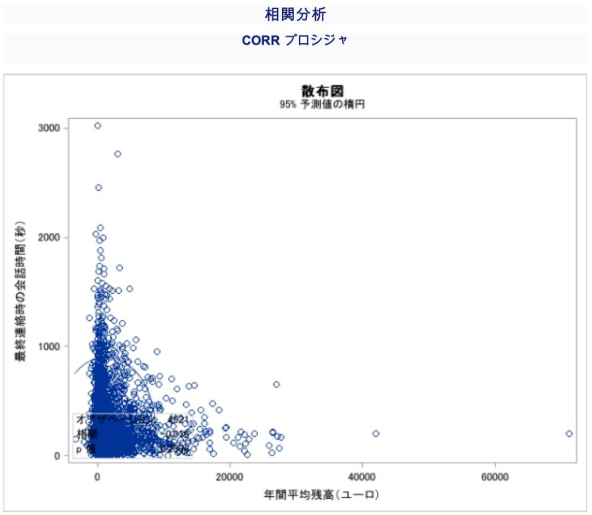
- 相関分析と同時に、各々の散布図を結果に含めて出力することも可能
- ただし、散布図行列形式ではなく、個別にグラフが分かれて出力されるため、閲覧性は低い

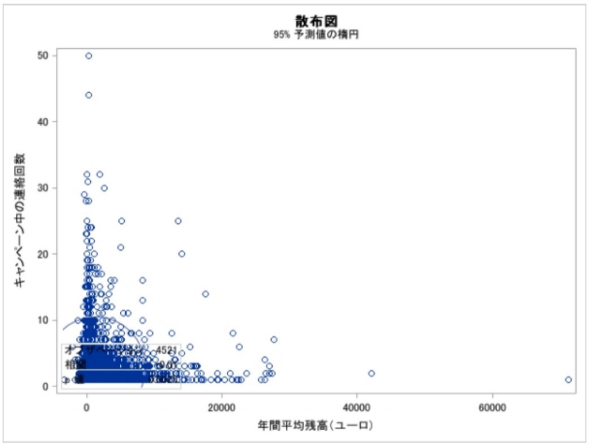


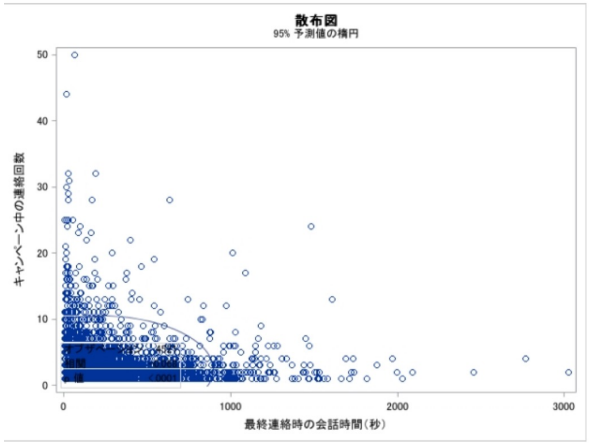
（参考）相関分析の出力：散布図の同時出力 – 実行結果

- 相関分析と同時に、各々の散布図を結果に含めて出力することも可能
- ただし、散布図行列形式ではなく、個別にグラフが分かれて出力されるため、閲覧性は低い









相関分析の注意点：「アイスクリーム売上」と「溺死件数」の関係

あなたは、あるシンクタンクの社員として働いている。

今回、とある省庁から、様々な消費者データと社会データについての調査を任された。

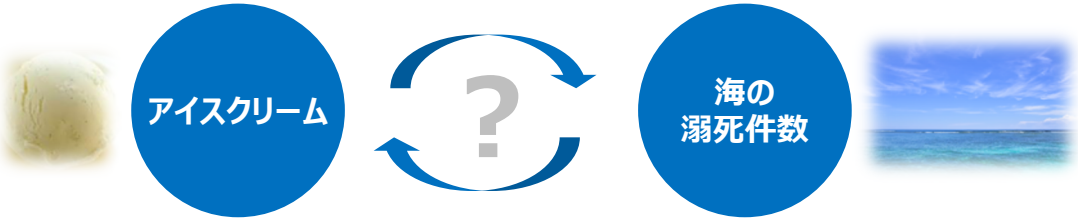
調査の結果、

「**アイスクリームが売れると、海の溺死件数が増える**」

という衝撃的なデータが得られた。

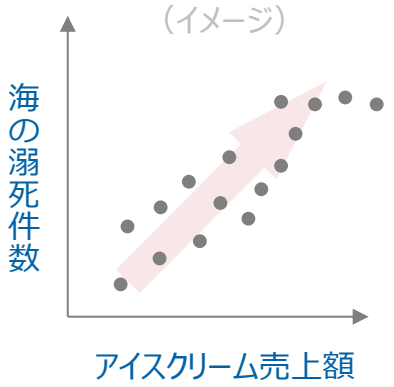
これが事実なら、即刻、**アイスクリームの販売に規制をかけるべき**である。

あなたの見解は？



https://shop17-makeshop.akamaized.net/shopimages/takanashi/0000000007922.jpg

https://upload.wikimedia.org/wikipedia/commons/thumb/7/7a/Caribbean_sea_-_Morrocoy_National_Park_-_Playa_escondida.jpg/1600px-Caribbean_sea_-_Morrocoy_National_Park_-_Playa_escondida.jpg



相関分析の注意点：相関と因果の違い

- 相関が高くても（連動しているように見えても）、必ずしも**因果があるとは限らない**
- このケースでは、両者の間に気温（季節）という**潜伏変数**が介在しており、
  これが両者に影響を与えることで**見かけの相関（疑似相関）**となって現れた可能性が高い

https://shop17-makeshop.akamaized.net/shopimages/takanashi/0000000007922.jpg

https://upload.wikimedia.org/wikipedia/commons/thumb/7/7a/Caribbean_sea_-_Morrocoy_National_Park_-_Playa_escondida.jpg/1600px-Caribbean_sea_-_Morrocoy_National_Park_-_Playa_escondida.jpg

· 夏になればアイスクリーム売上は増加
· 猛暑の夏であれば、さらに売れる

· 夏になれば海水浴客も多くなり、溺死件数も増加
· 猛暑の夏であれば、さらに海水浴客は増加

https://torange.biz/jp/sun-27344

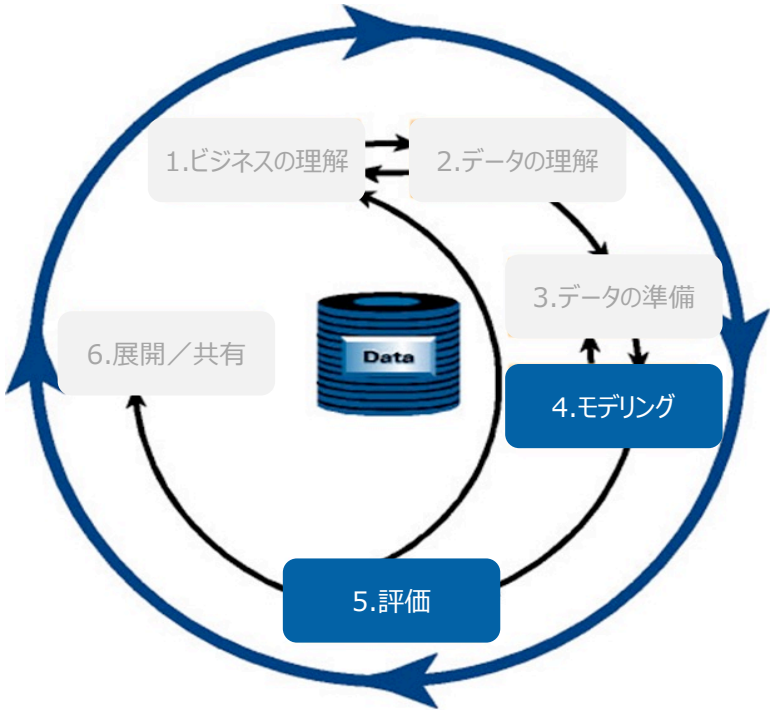
潜伏変数

再掲

ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)

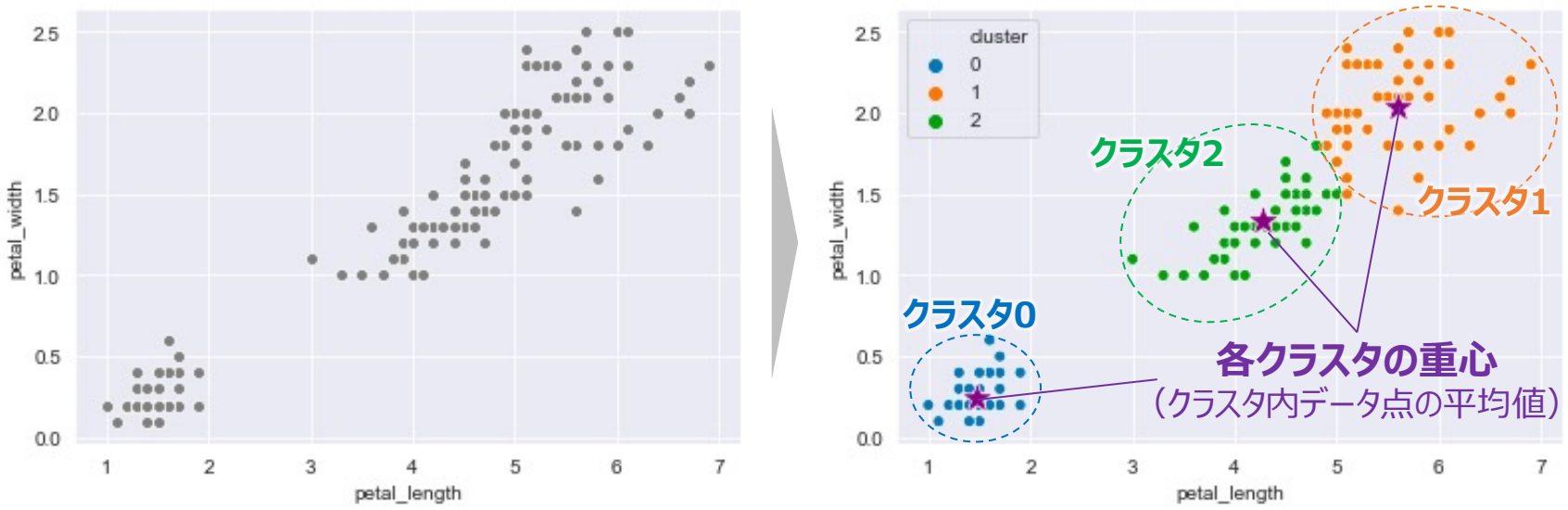
1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル作成	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

再掲

非階層クラスタリング：k-means法

- クラスタリング手法の中で代表的かつ最もシンプルな手法が「**k-means法**」であり、各クラスタ内の**データ平均値（means）**を重心として、**k個のクラスター**に分類することができる

▼2次元のk-meansクラスタリング例

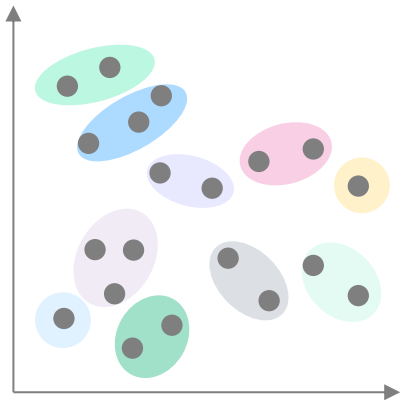


▼分類結果の特徴
- 教師なしのため、各クラスタの意味解釈は人が行う
- 円状（球状）のクラスタになりやすい
- クラスタサイズ（クラスタ内のデータ数）が同程度になりやすい

▼アルゴリズムの特徴
- クラスタ数を事前に明示的に決める必要がある
- 距離依存のため、データのスケールによって結果が変わる
- 初期値（初期重心）に大きく依存 ※後述

クラスタ数設定の考え方

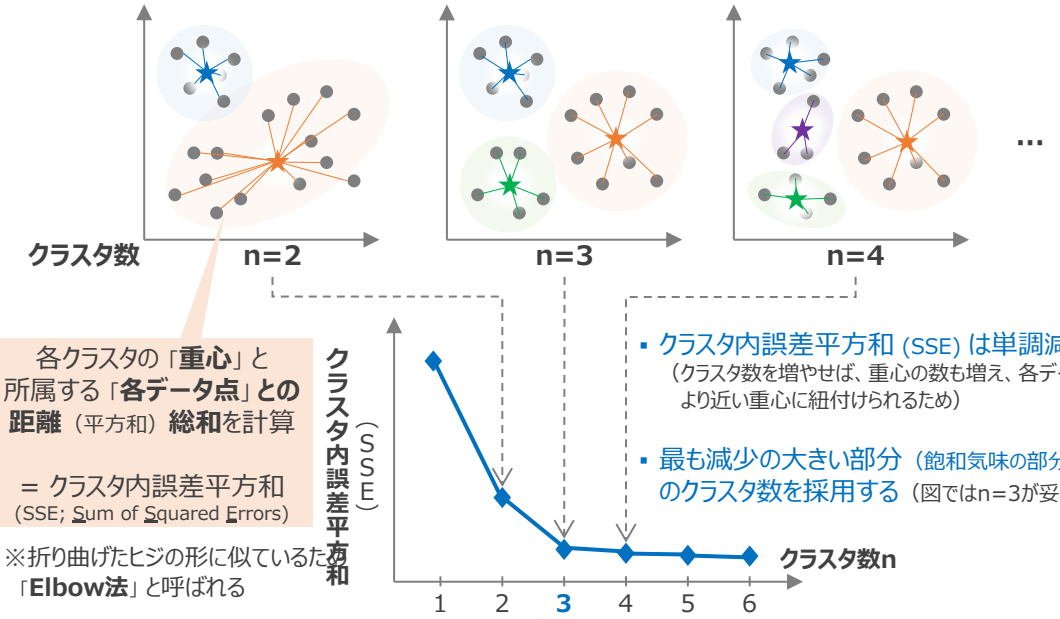
- クラスタ数を客観的に評価する「Elbow法」などの手法もあるが、一般的には、まずは**人間が解釈可能なレベルの3~5個程度**から着手してみることが賢明
- 階層的クラスタリングや、自動的にクラスタ数を決めてくれる手法（DBScanなど）を活用する



細かくクラスタを分けすぎても、
解釈（=各クラスタの意義付け）が困難

人間が解釈しやすい3~5個程度から始めてみることが有効

参考：Elbow法による客観的なクラスタ数の評価法

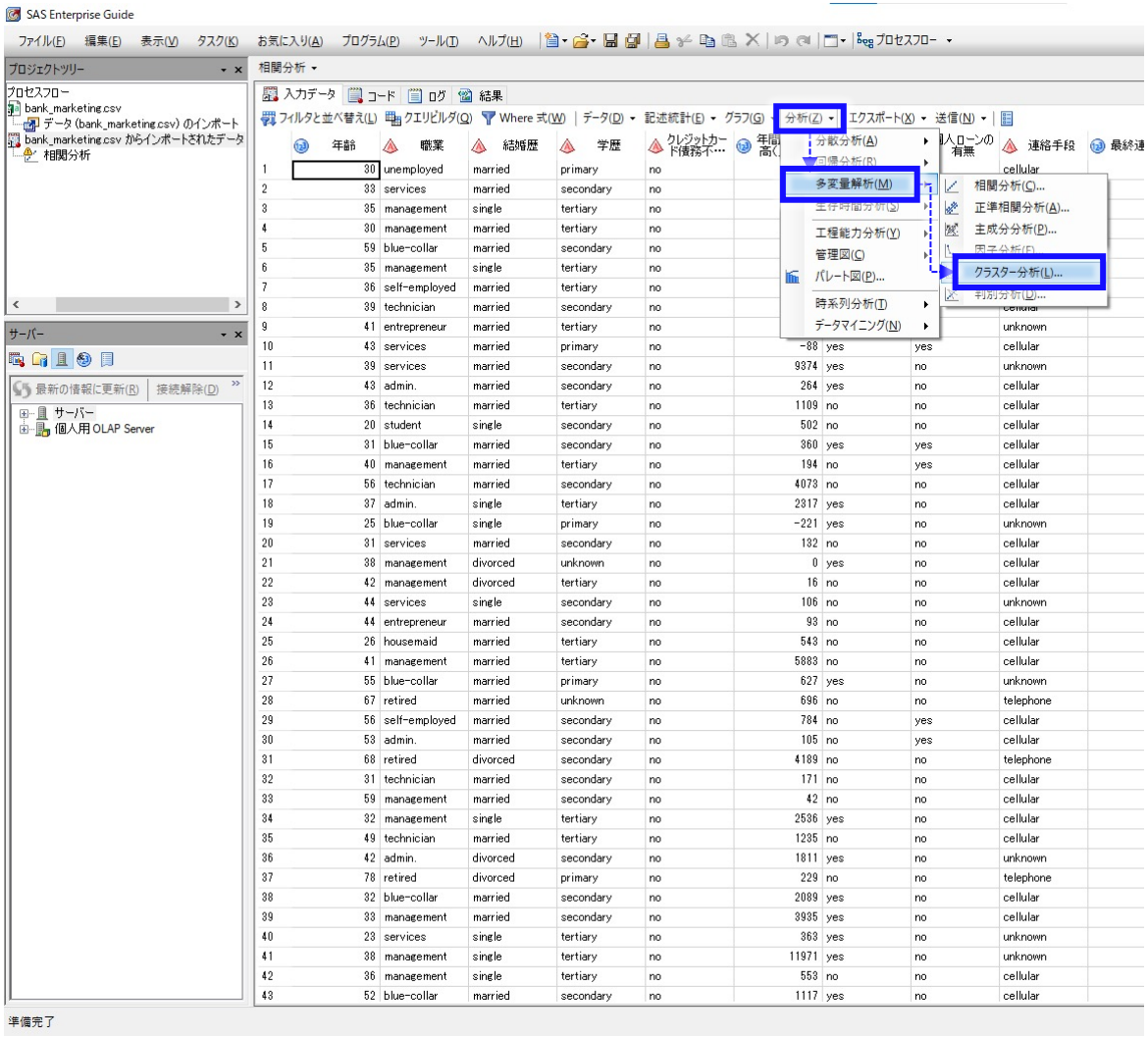


SAS Enterprise Guide での実装方法

- 非階層的クラスタリング（k-means）
- クラスタ数の変更
- クラスタリング結果の解釈
- クラスタ番号の出力と追加分析
- グループ変数の設定

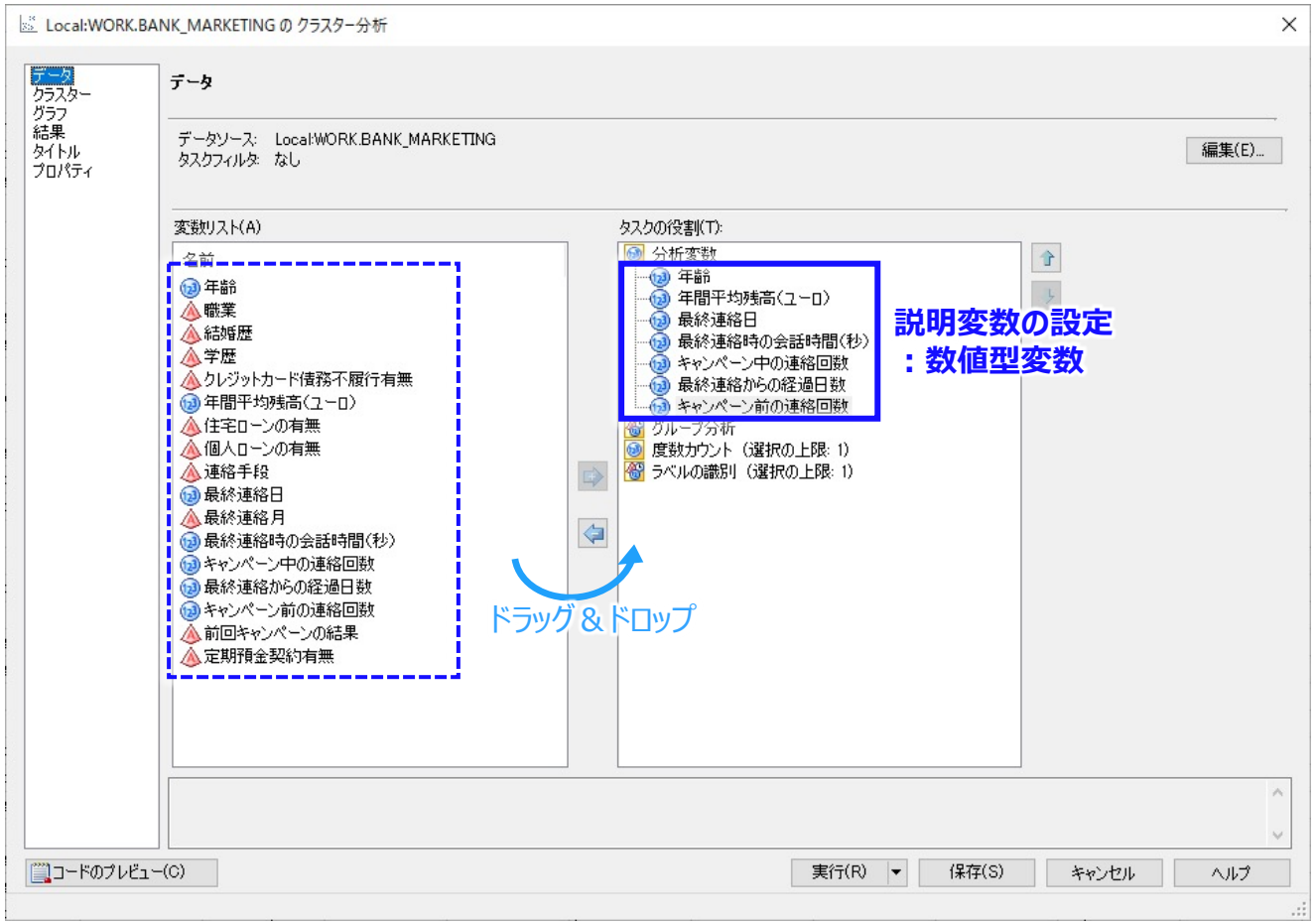
非階層的クラスタリング（k-means） – 実行方法（1/3）

[分析] → [多変量解析] → [クラスター分析] をクリック



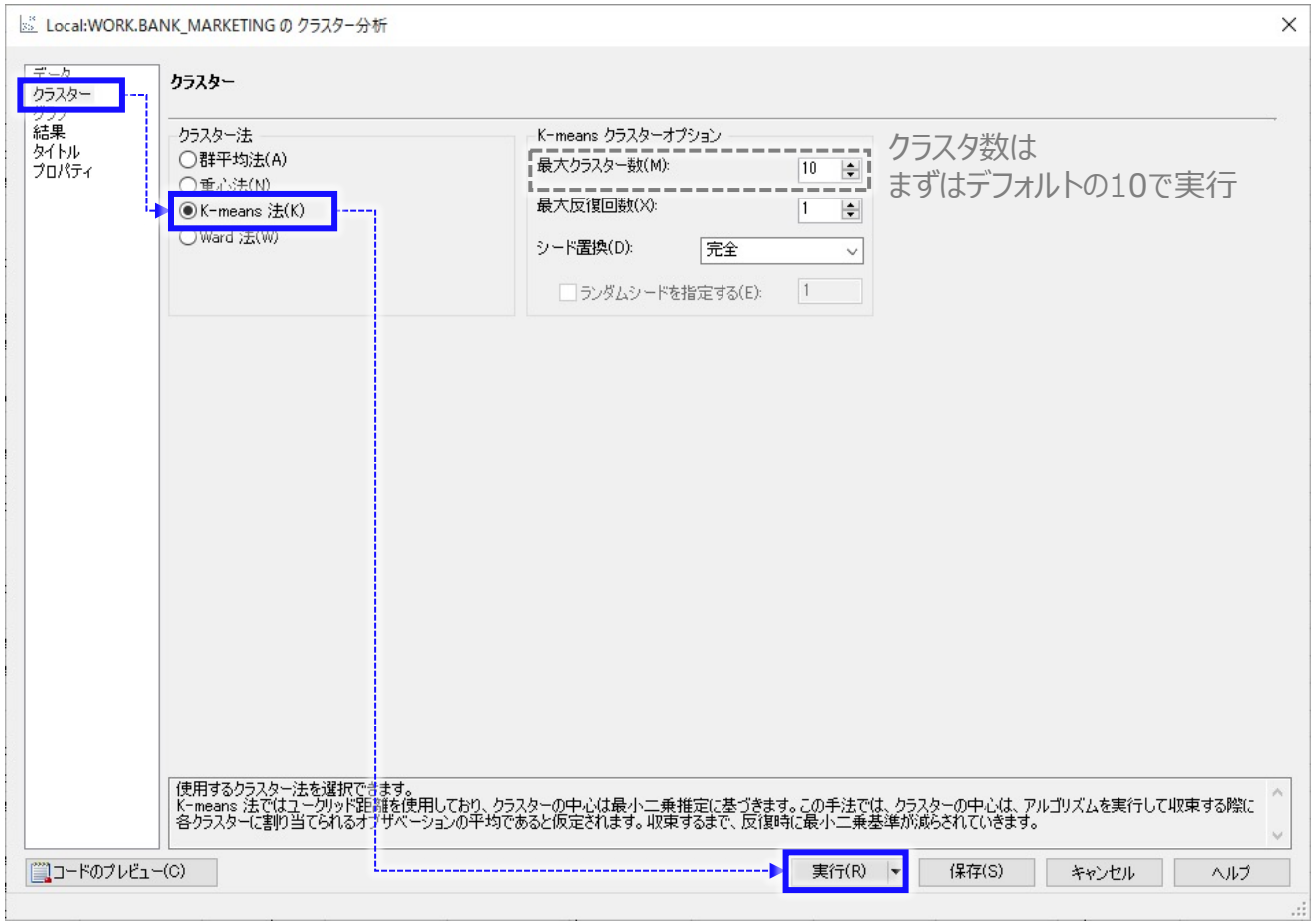
非階層的クラスタリング（k-means） – 実行方法（2/3）

- 左側の [変数リスト] より、数値型の変数を選択
- 右側の [分析変数] にドラッグ＆ドロップ



非階層的クラスタリング（k-means） – 実行方法（2/4）

- 左パネルから [クラスター]メニュー をクリック
- クラスター法から[K-means法]を選択し、実行ボタン



非階層的クラスタリング（k-means） – 実行結果

- 教師なし学習のクラスタリングでは、各クラスタの特徴は人間が解釈を行う必要がある。具体的には、各クラスタにおける説明変数の値傾向（=**クラスタ内での平均値**）を確認していく
- ただし、クラスタ数が多すぎると、分析結果が複雑化し、解釈が非常に難しくなる

入力した説明変数

各クラスターの**平均値**

クラスター平均

クラスター	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
1	41.15393	1007.35032	15.58245	269.03432	2.73107	42.75973	0.56511
2	43.50000	19079.50000	16.16667	202.83333	3.33333	-1.00000	0.00000
3	41.85714	10945.50000	16.00000	217.53571	2.53571	29.51786	0.32143
4	60.00000	71188.00000	6.00000	265.20945	2.78253	-1.00000	0.00000
5	43.41311	5668.45014	16.48148	265.20945	2.78253	43.19943	0.74074
6	43.00000	22247.22222	12.44444	236.44828	3.44828	127.88889	1.11111
7	42.00000	42045.00000	8.00000	256.49595	2.94327	-1.00000	0.00000
8	49.37500	26762.75000	14.37500	201.12500	2.71875	9.62500	0.37500
9	45.65517	15188.89655	17.00000	236.44828	3.44828	41.06897	1.03448
10	40.33874	-81.05754	16.53566	256.49595	2.94327	32.18720	0.43355

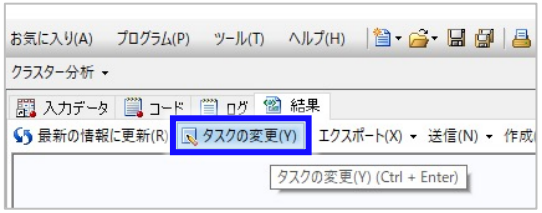
10個のクラスタ数では、分析結果の解釈や
クラスタ間の特徴比較が困難
→ **クラスタ数を減らして再実行**

各クラスターの**標準偏差**

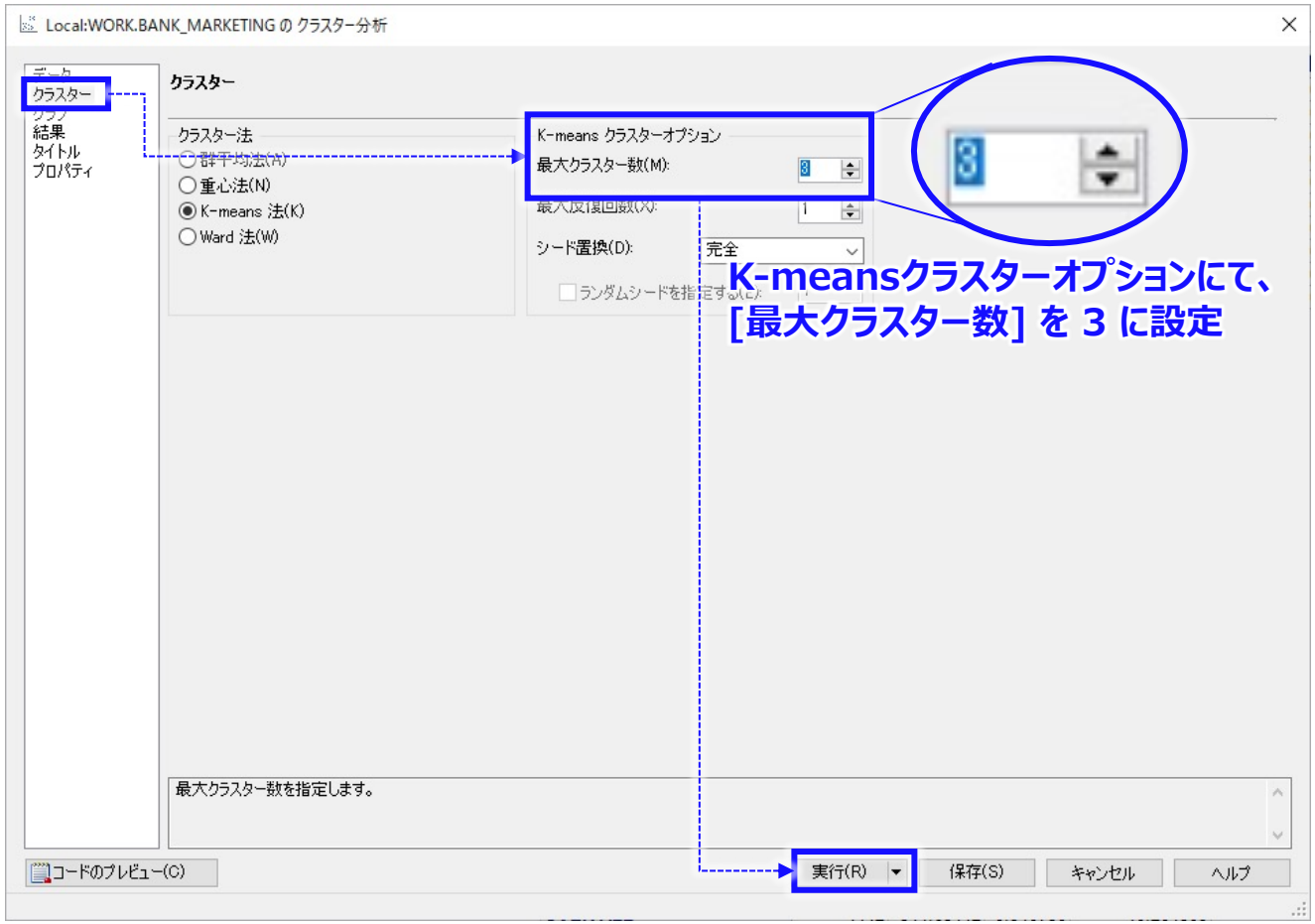
クラスター標準偏差

クラスター	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
1	10.624434	914.738296	8.158936	266.086384	3.141028	104.412924	1.759779
2	13.982131	1001.617043	9.703951	137.781590	5.240865	0.000000	0.000000
3	9.541162	1331.760407	7.783432	186.779984	1.672932	75.061185	0.833550
4	.	.	.	.	.	.	.
5	11.775895	1453.982754	7.924218	221.608726	2.706275	99.942127	1.985237
6	11.079260	739.135099	6.912147	72.348117	2.603417	166.575392	1.536591
7	.	.	.	.	.	.	.
8	13.595561	627.239701	8.158037	195.038778	1.982062	30.052038	1.060660
9	12.809806	1235.598737	8.375133	158.657220	5.467999	78.994544	2.306555
10	9.911711	252.414859	8.524424	261.749378	3.120457	90.531743	1.445706

非階層的クラスタリング（k-means）：クラスタ数変更 – 実行方法



分析結果の画面上において
上部メニュー [タスクの変更] をクリック



K-meansクラスターオプションにて、
[最大クラスター数] を 3 に設定

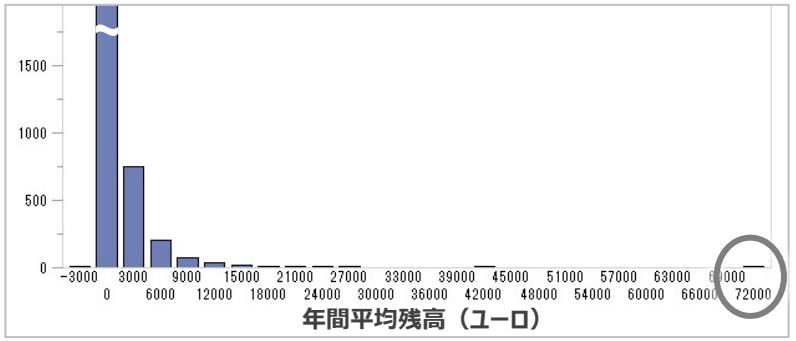
非階層的クラスタリング（k-means）：クラスタ数変更 – 実行結果

▼各クラスタの概要

各クラスタに所属するデータの数

クラスターの要約

クラスター	度数	RMS標準偏差	シードからオブザベーションまでの最大距離	半径超える	最も近いクラスター	クラスター重心間の距離
1	97	2133.8	24363.6		3	14502.2
2	1	.	0		1	55589.9
3	4423	658.0	8220.7		1	14502.2



必要に応じて、
2段階クラスタリングも有効

クラスタ2は、[年間平均残高]の外れ値が検出されている
→ 外れ値を除いて再実行 or クラスタ数を4つ以上にして再実行

▼各クラスタにおける説明変数の値傾向（平均値）

クラスター平均

クラスター	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
1	43.85567	15598.10309	15.71134	202.88660	2.74227	38.32990	0.58763
2	60.00000	71188.00000	6.00000	205.00000	1.00000	-1.00000	0.00000
3	41.10694	1096.00543	15.92200	265.31404	2.79516	39.80737	0.54171

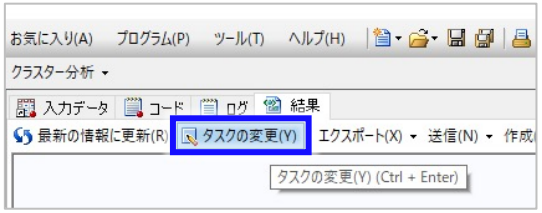
クラスタ1：残高が多いが、最後のコミュニケーションが希薄
クラスタ3：残高が少ないが、最後のコミュニケーションが濃密

▼各クラスタにおける説明変数の値傾向（標準偏差）

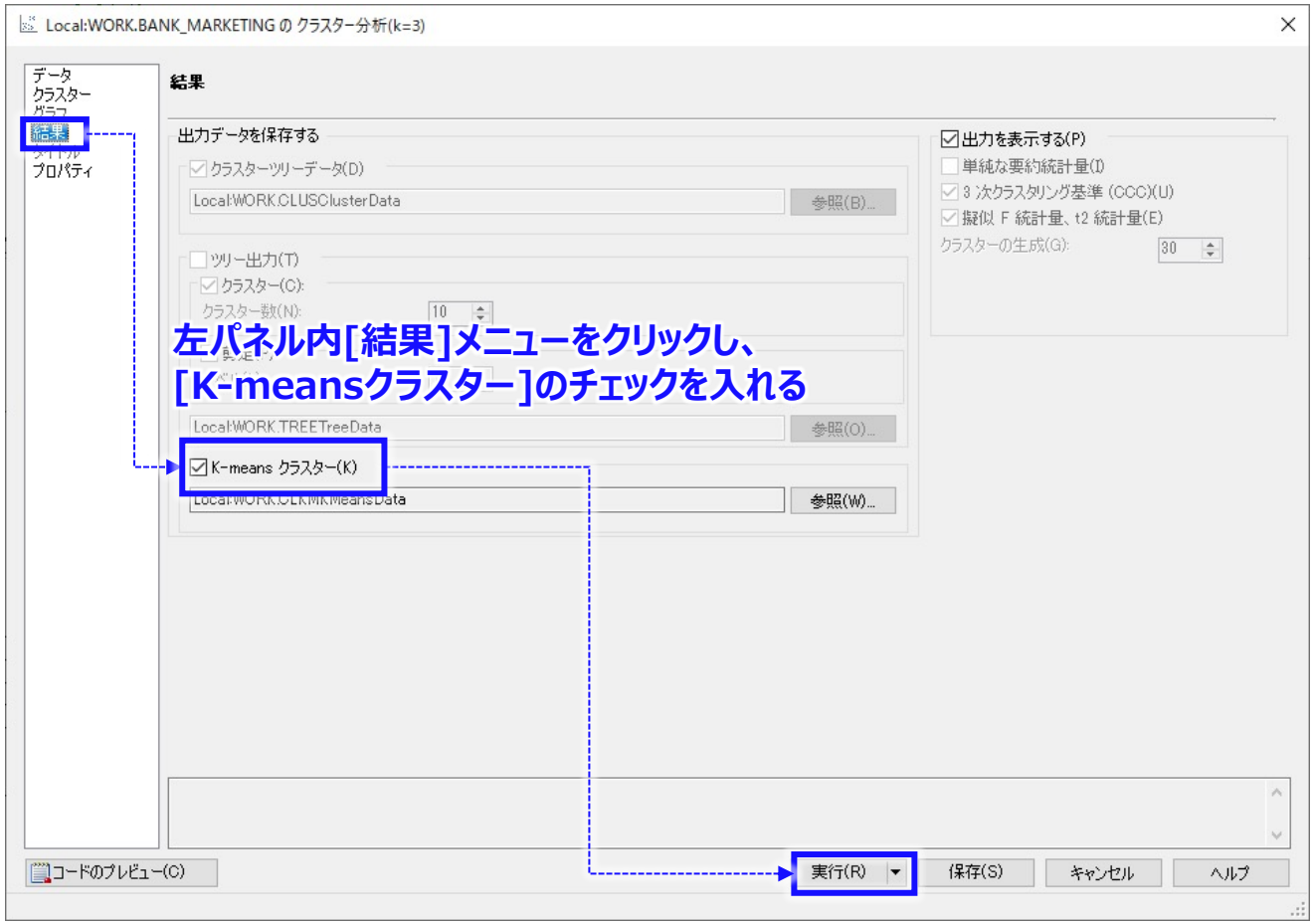
クラスター標準偏差

クラスター	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
1	11.316461	5642.461735	8.111153	157.542645	3.485919	88.519621	1.512122
2	.	.	.	.	.	.	.
3	10.550480	1718.206591	8.251065	261.531447	3.101714	100.378596	1.697635

（参考）クラスター番号のデータ化 と 追加分析（1/3）



分析結果の画面上において
上部メニュー [タスクの変更] をクリック



（参考）クラスター番号のデータ化 と 追加分析（2/3）

[出力データ]タブをクリックすると
[CLUSTER]カラムが挿入されている

入力データ | コード | ログ | 出力データ | 結果

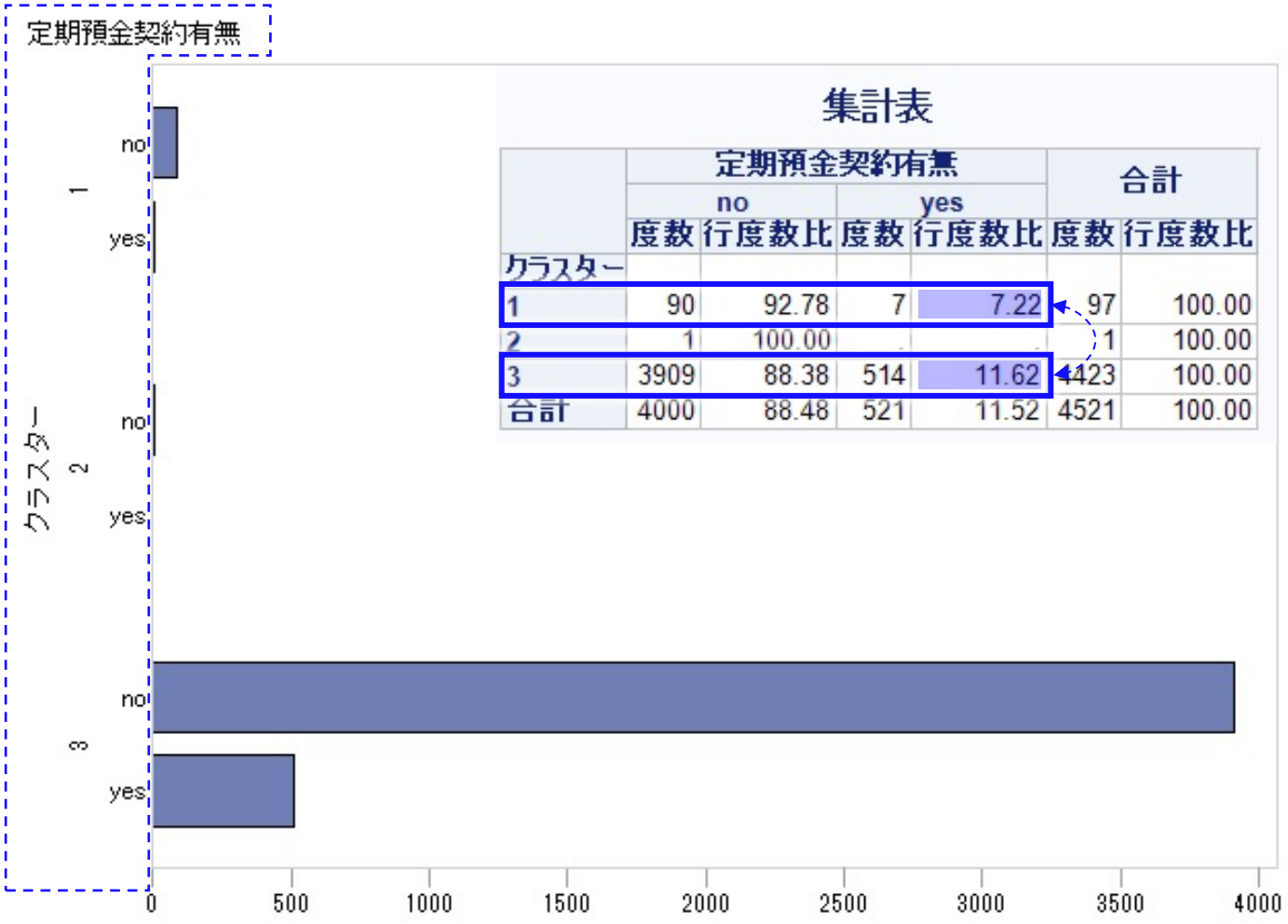
タスクの変更(Y) | フィルタと並べ替え(L) | クエリビルダ(Q) | Where 式(W) | データ(D) | 記述統計(E) | グラフ(G) | 分析(Z) | エクスポート(X) | 送信(N)

	学歴	クレジットカード債務不…	年間平均残高(ユーロ)	住宅ローンの有無	個人ローンの有無	連絡手段	最終連絡日	最終連絡月	最終連絡時の会話時…	キャンペーン中の連絡…	最終連絡からの経過…	キャンペーン前の連絡…	前回キャンペーンの結果	定期預金契約有無	CLUSTER	DISTANCE
1	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no	3	656.63940173
2	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no	3	3642.9642192
3	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no	3	356.87202241
4	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no	3	327.10671707
5	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no	3	1160.2056178
6	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no	3	451.11164555
7	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no	3	903.02117766
8	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no	3	1018.9044937
9	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no	3	961.28605959
10	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no	3	1252.1927541
11	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no	3	8215.4561176
12	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no	3	908.36809445
13	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no	3	90.173090227
14	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes	3	658.40906959
15	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no	3	842.38629128
16	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no	3	968.57427861
17	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no	3	2914.8085107
18	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no	3	1173.5487031
19	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no	3	1380.4419663
20	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no	3	1039.4669453
21	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0	unknown	no	3	1171.608764
22	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0	unknown	no	3	1150.1797634
23	secondary	no	106	no	no	unknown	12	jun	109	2	-1	0	unknown	no	3	1064.92077
24	secondary	no	93	no	no	cellular	7	jul	125	2	-1	0	unknown	no	3	1075.6045291
25	tertiary	no	543	no	no	cellular	30	jan	169	3	-1	0	unknown	no	3	624.74149137
26	tertiary	no	5883	no	no	cellular	20	nov	182	2	-1	0	unknown	no	3	4725.2542419
27	primary	no	627	yes	no	unknown	5	may	247	1	-1	0	unknown	no	3	533.79651123
28	unknown	no	696	no	no	telephone	17	aug	119	1	105	2	failure	no	3	490.18959856
29	secondary	no	784	no	yes	cellular	30	jul	149	2	-1	0	unknown	no	3	394.80371599
30	secondary	no	105	no	yes	cellular	21	aug	74	2	-1	0	unknown	no	3	1071.6552078
31	secondary	no	4189	no	no	telephone	14	jul	897	2	-1	0	unknown	yes	3	3095.9542112
32	secondary	no	171	no	no	cellular	27	aug	81	3	-1	0	unknown	no	3	1005.5608977
33	secondary	no	42	no	no	cellular	21	nov	40	1	-1	0	unknown	no	3	1139.9531316
34	tertiary	no	2536	yes	no	cellular	26	aug	958	6	-1	0	unknown	yes	3	1542.5002593
35	tertiary	no	1235	no	no	cellular	13	aug	354	3	-1	0	unknown	yes	3	124.44802746
36	secondary	no	1811	yes	no	unknown	14	may	150	1	-1	0	unknown	no	3	663.64503127
37	primary	no	229	no	no	telephone	22	oct	97	1	-1	0	unknown	yes	3	946.3063197
38	secondary	no	2089	yes	no	cellular	14	nov	132	1	-1	0	unknown	yes	3	940.72240681
39	secondary	no	3935	yes	no	cellular	6	may	765	1	342	2	failure	yes	3	2837.2176662

（参考）クラスター番号のデータ化 と 追加分析（3/3）

- クラスター番号をデータ化することで、様々な切り口で層別した追加分析が可能となる
- 例えば、クラスターごとに「定期預金契約有無」の傾向を確認することができる

クラスターごとの定期預金契約有無

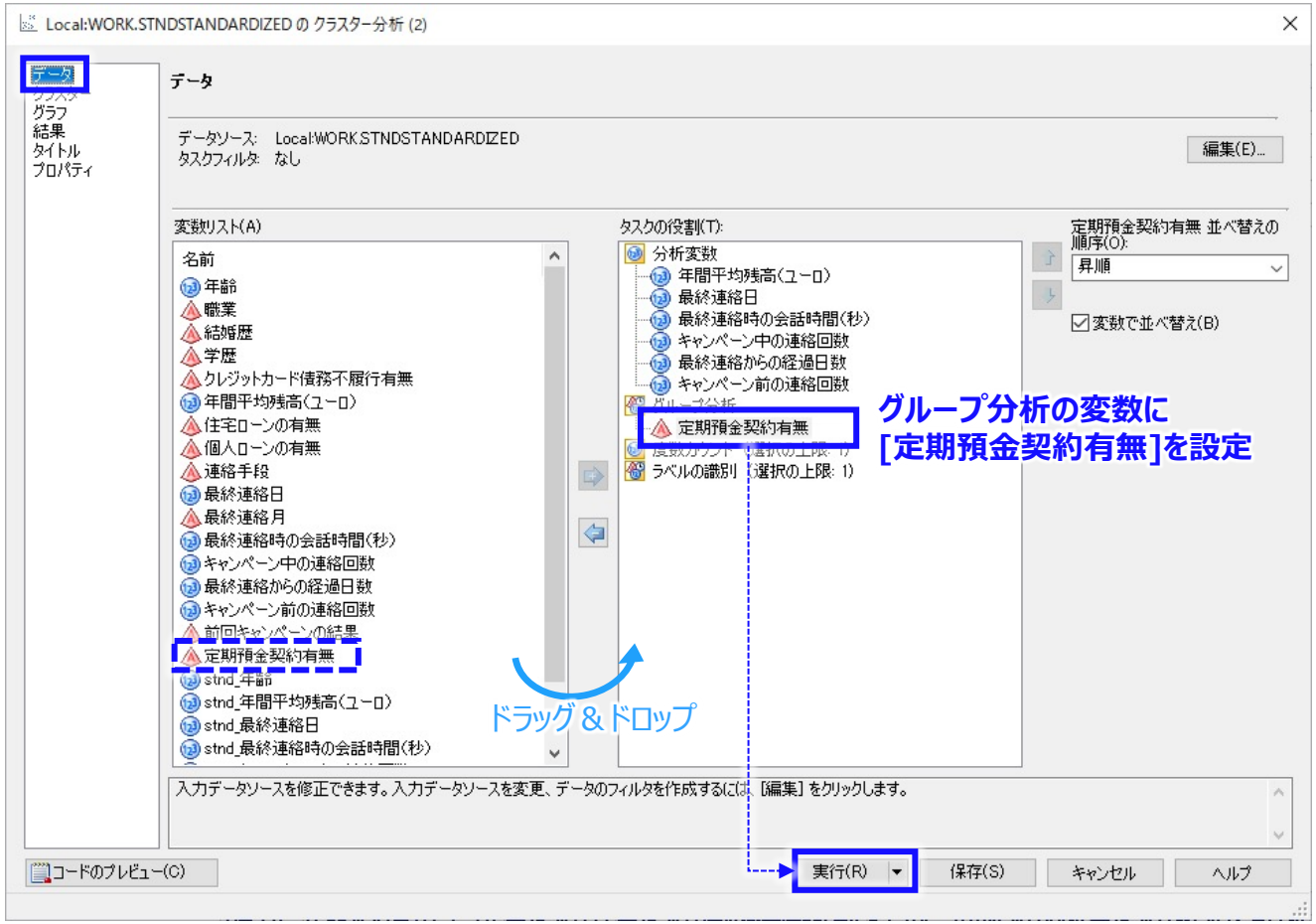


集計表

	定期預金契約有無				合計	
	no		yes			
	度数	行度数比	度数	行度数比	度数	行度数比
クラスター						
1	90	92.78	7	7.22	97	100.00
2	1	100.00	.	.	1	100.00
3	3909	88.38	514	11.62	4423	100.00
合計	4000	88.48	521	11.52	4521	100.00

非階層的クラスタリング（k-means）：グループ変数 – 実行方法

- 例えば「契約者」と「未契約者」で明確にグループを分けて、各々のグループ内でクラスタリングを実行したい場合は、**「グループ変数」を設定してクラスタリングを行う**



非階層的クラスタリング（k-means）：グループ変数 – 実行結果

- グループ変数を活用することで、「契約者」の中でのパターン分析、「未契約者」の中でのパターン分析をそれぞれ行うことができる
- また、異なるグループ間でのクラスタ特性を比較することで新たな洞察を得られる可能性がある

契約者

クラスターの要約

クラスター	度数	RMS標準偏差	シードからオブザベーションまでの最大距離	半径超える	最も近いクラスター	クラスター重心間の距離
1	487	512.9	3765.9		3	6238.5
2	2	2046.9	7658.8		3	15799.9
3	32	877.6	5452.5		1	6238.5

クラスター平均

クラスター	年齢	年間平均残高(ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回数	最終連絡からの経過日数	キャンペーン前の連絡回数
1	42.09651	1104.28542	15.60780	561.64476	2.27105	69.17659	1.08624
2	49.50000	2[illegible]41.00000	12.50000	451.50000	1.500[illegible]	-1.00000	0.00000
3	48.06250	7[illegible]41.25000	16.62500	423.59375	2.250[illegible]	64.81250	1.21875

年齢や平均残高などの顧客属性は類似しているが、顧客接点の深さ・多さには明確な差がある

未契約者

クラスターの要約

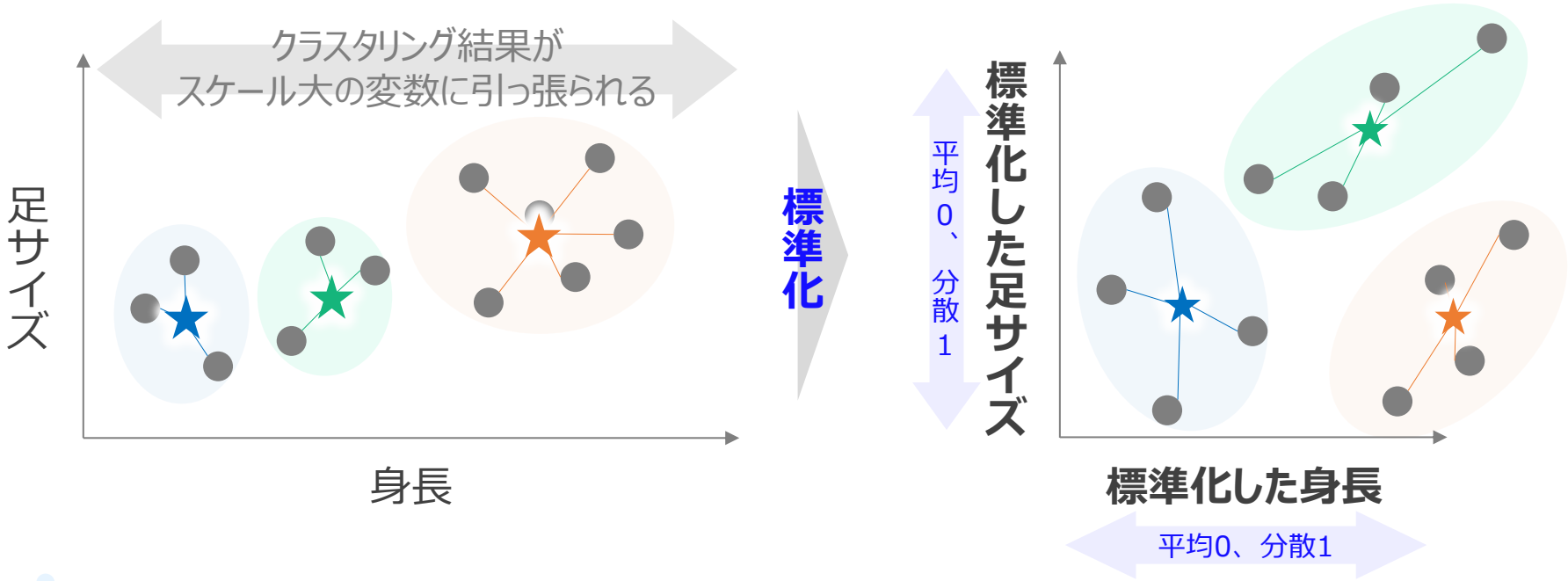
クラスター	度数	RMS標準偏差	オブザ[illegible]までの	ードから[illegible]ーション[illegible]最大距離	半径超える	最も近いクラスター	クラスター重心間の距離
1	90	2126.3		24400.2		3	14601.2
2	1	.		0		1	55529.7
3	3909	651.0		8257.0		1	14601.2

クラスター平均

クラスター	年齢	年間平均残[illegible](ユーロ)	最終連絡日	最終連絡時の会話時間(秒)	キャンペーン中の連絡回[illegible]	最終連絡からの経過日数	キャンペーン前の連絡回数
1	43.10000	1[illegible]358.34444	15.81111	193.06667	2.844[illegible]	37.03333	0.61111
2	60.00000	7[illegible]88.00000	6.00000	205.00000	1.000[illegible]	-1.00000	0.00000
3	40.94474	1057.15221	15.95446	227.11921	2.86314	35.99181	0.46815

（参考）クラスタリングにおける変数スケールの影響と標準化

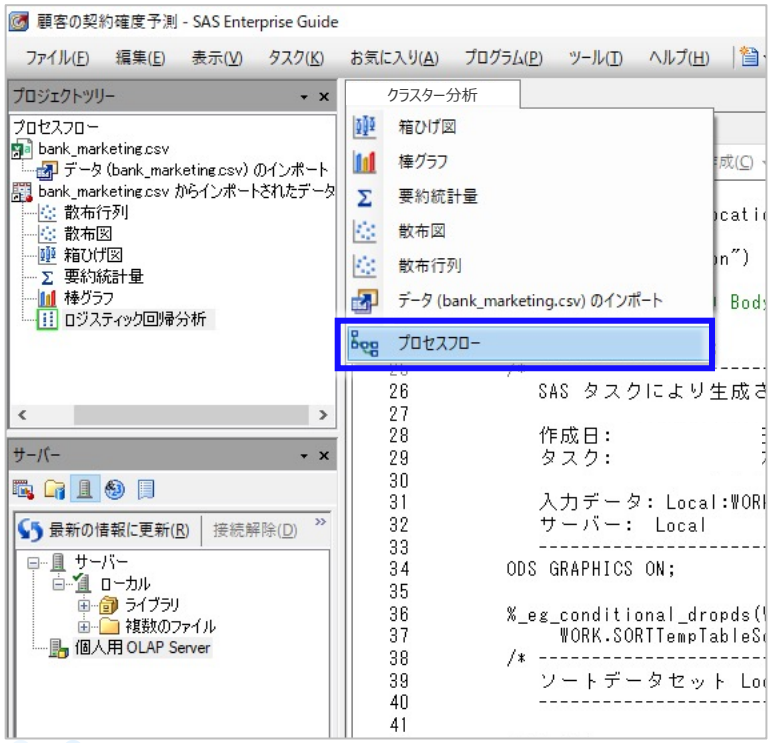
- k-means法などの「距離」に基づくによるクラスタリング手法は、データの「**スケール**」に大きく影響を受ける。このため、必要に応じて、「**標準化**」の処理を行なった上でクラスタリングを行う必要がある

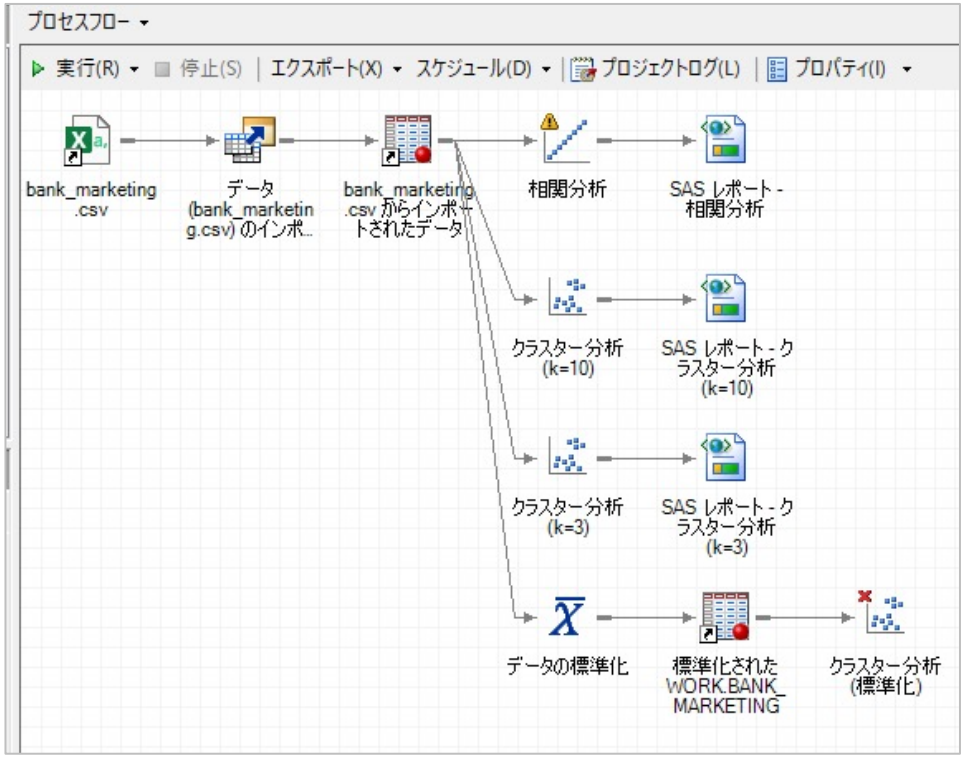


※第3回にて取り扱う予定

（参考）プロセスフローでの確認

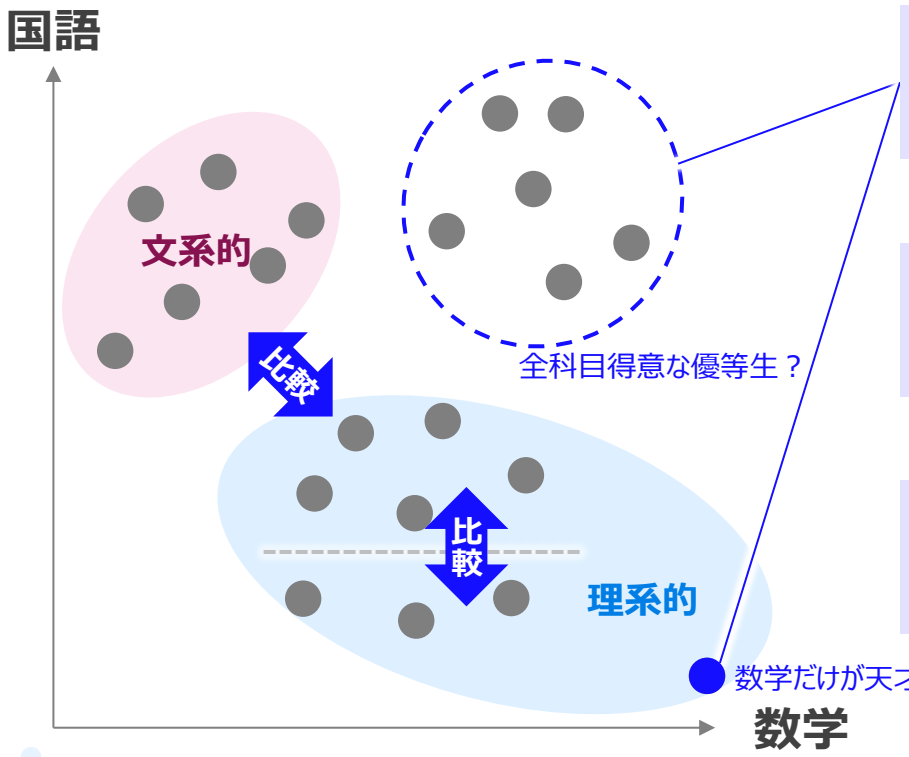
- SAS Enterprise Guideでは、「**プロセスフロー**」によりデータ加工やグラフ観察、モデル構築などの一連の分析プロセスを俯瞰的／反復的に確認可能





非階層クラスタリングの活用方法のまとめ

- クラスタリングでは、ある程度の「似たもの同士」がより分けられるため、同一クラスタ内でも存在する差異を分析したり、異なるクラスタ間での特徴の違いを分析することが有効である
- 一方、教師なし学習という特性から、「意外性」のあるグループやデータを見つけられることもある



①**意外性**のあるグループ（クラスタ）や
仲間外れのデータポイントを見つける

②**混在要因**をより分ける
→ クラスタ「**内**」での分析

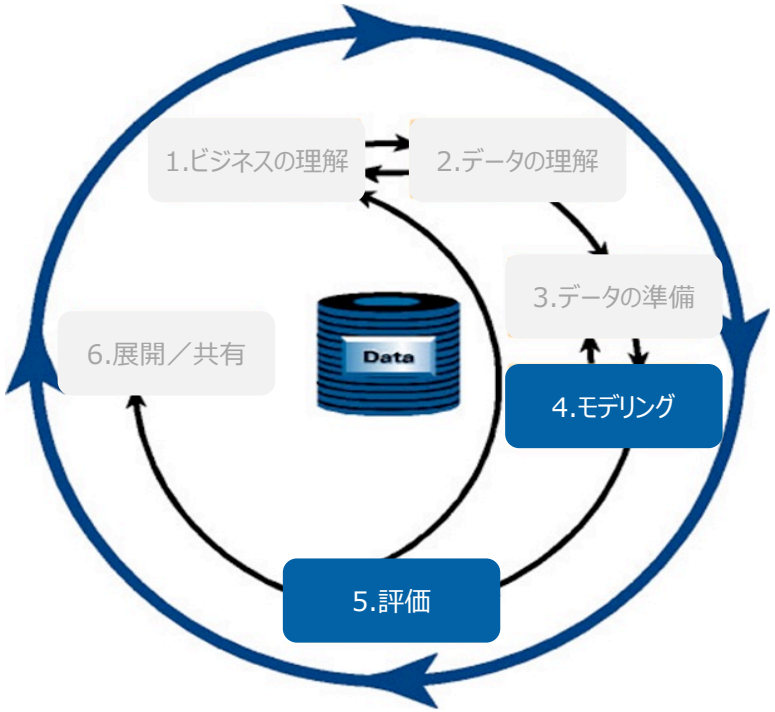
③各**クラスタの特徴**を比較する
→ クラスタ「**間**」での分析（目的変数をクラスタ番号とする）

再掲

ビッグデータ分析の進め方

- データマイニングの進め方に関する方法論「**CRISP-DM**」に基づいて、分析と評価を繰り返して試行錯誤しながら進めるのが一般的である

CRISP-DM: データマイニング方法論



(**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)

1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル作成	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

まとめ

- 相関行列によるデータ観察
  - 相関分析を行うことにより、**変数間の関係性を全体把握**した
  - **目的変数別（グループ変数設定）に相関分析**を行うことで、異なるグループ間（契約者／未契約者）で、変数の相関性に違いを見出した

- クラスター分析による分類（1）：非階層的クラスタリング
  - 非階層的クラスタリング（k-means法）を適用することで、**類似の顧客をグルーピング**した
  - **クラスタ数をチューニング**することで、解釈しやすい結果を得た
  - 各クラスタにおける**説明変数の値傾向を確認することで、各クラスタの特徴を把握**した
  - **クラスター番号を出力して元データに紐づけ**ることで、様々な観点で追加分析が行えた
  - **目的変数別（グループ変数設定）にクラスター分析**を行い、異なるグループ間（契約者／未契約者）におけるクラスタ特性を比較することで、新たな洞察を得た

End of File