実務直結! 分析力向上ウェビナーシリーズ

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

2022年12月09日

Agenda

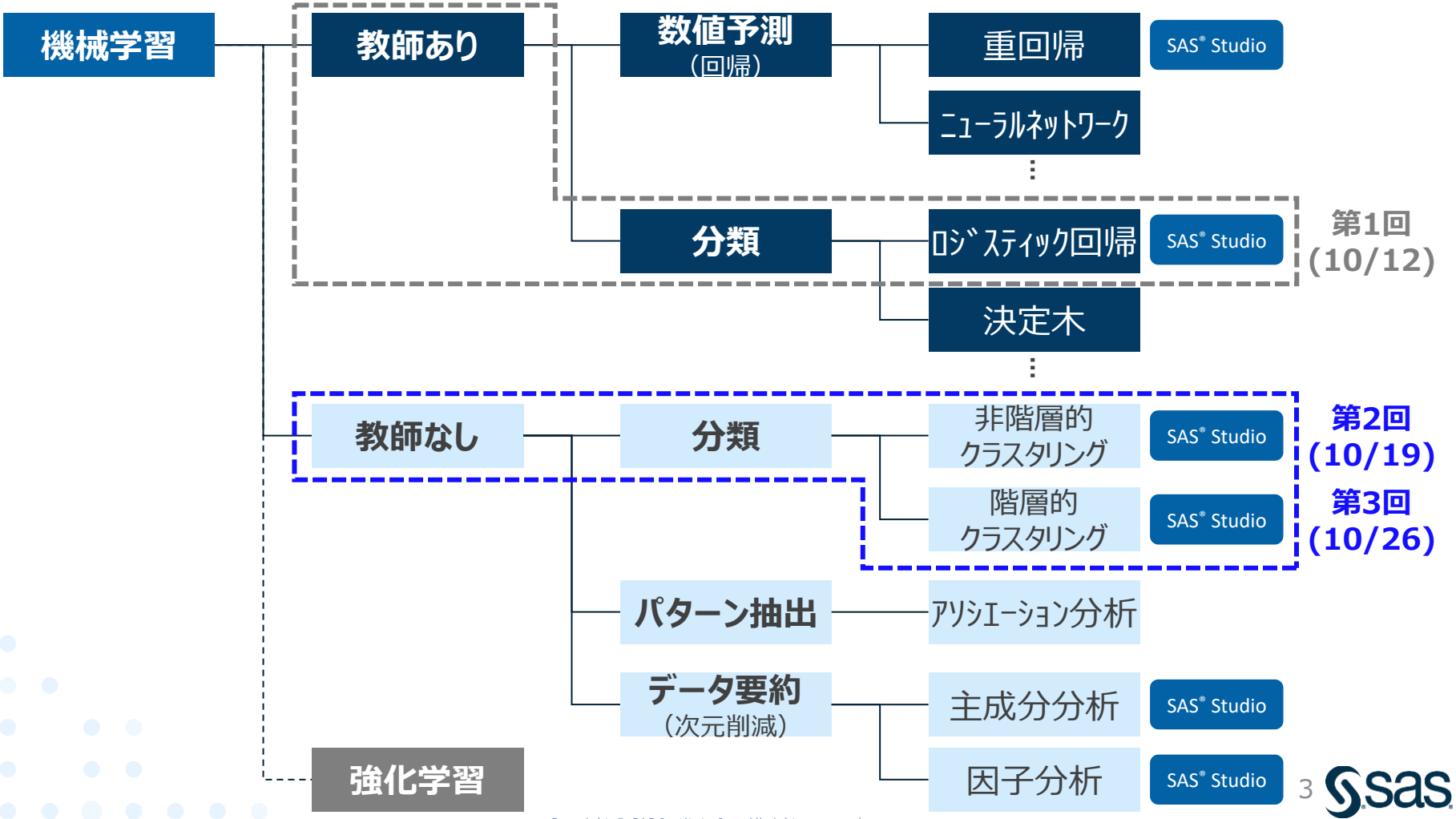
- 相関行列によるデータ観察
  - 相関関係の全体把握
  - 散布図行列との同時活用

- 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
  - グラフを活用した各クラスタの解釈方法
  - クラスタ数設定の考え方
  - 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

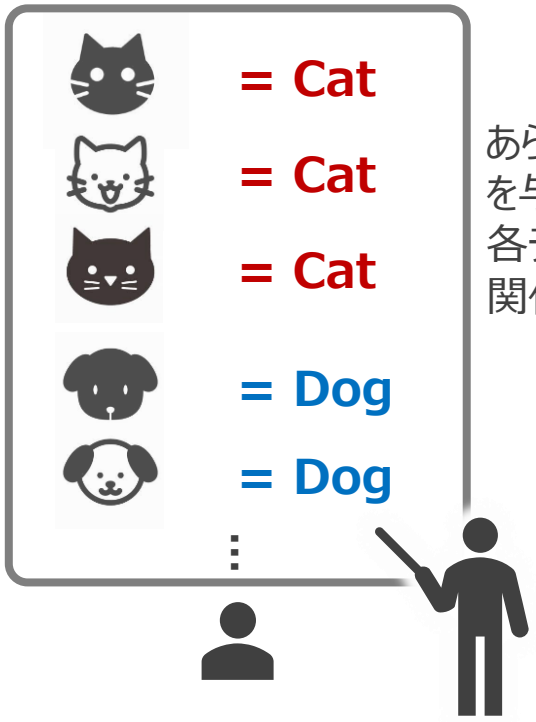


再掲

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

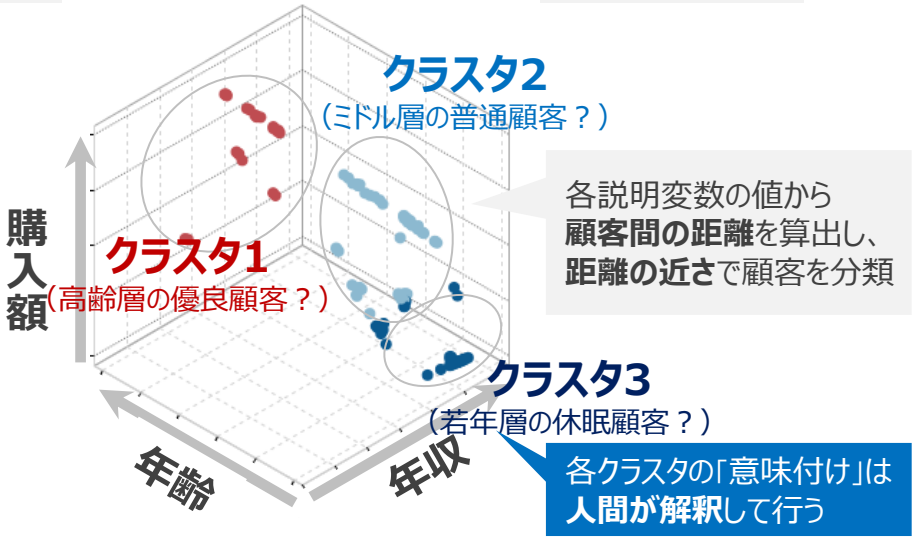
データ例

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	…
⋮	⋮	⋮	⋮	⋮	⋮	…

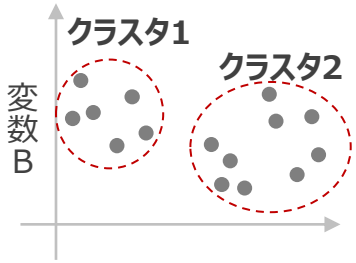
説明変数

※目的変数は無し

クラスタリング



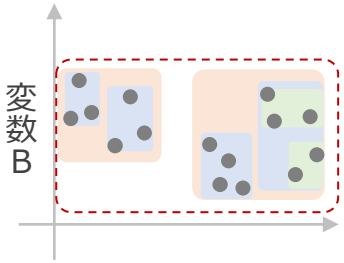
非階層的クラスタリング

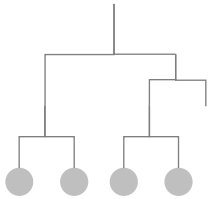


主な手法

- **k-means法**（k平均法）
- 混合ガウス

階層的クラスタリング





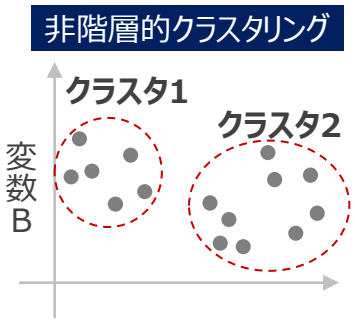
主な手法

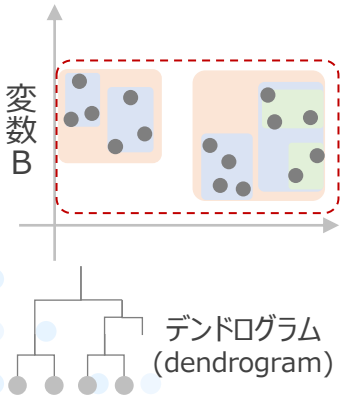
- 最短距離法
- 最長距離法
- 群平均法
- ウォード法

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® Studio
	▪ 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® Studio
	▪ 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 (10/26)**
	▪ 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® Studio
	▪ 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® Studio
	▪ その他	メディアン法、可変法	

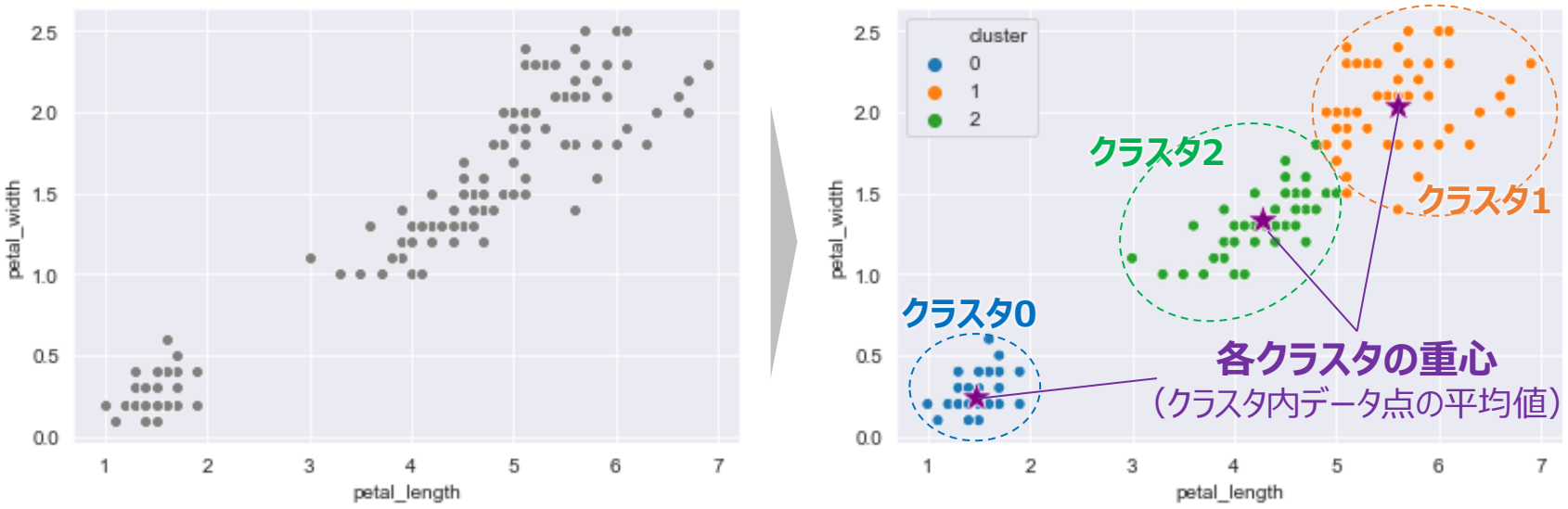




非階層クラスタリング：k-means法

- 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

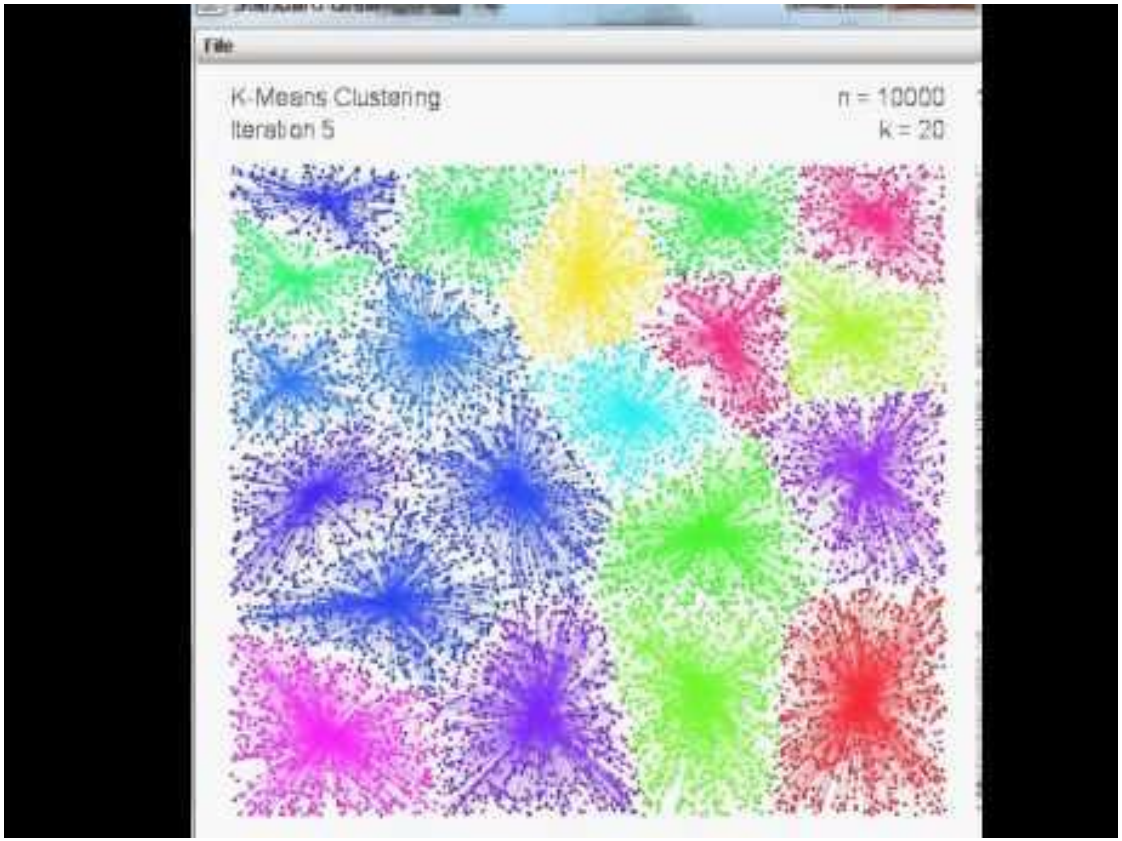
▼2次元のk-meansクラスタリング例



▼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

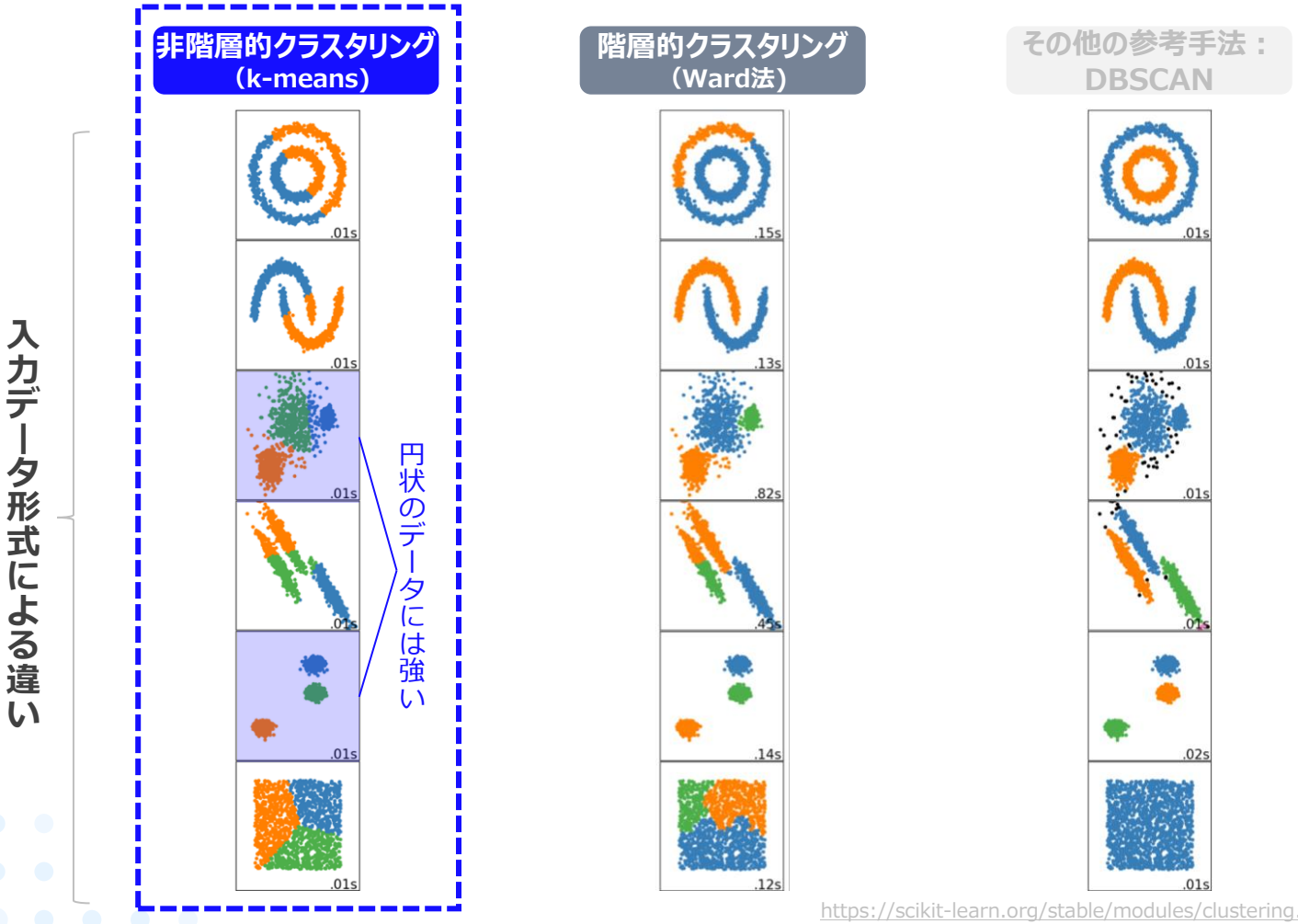
参考：k-means法のイメージ（動画）



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



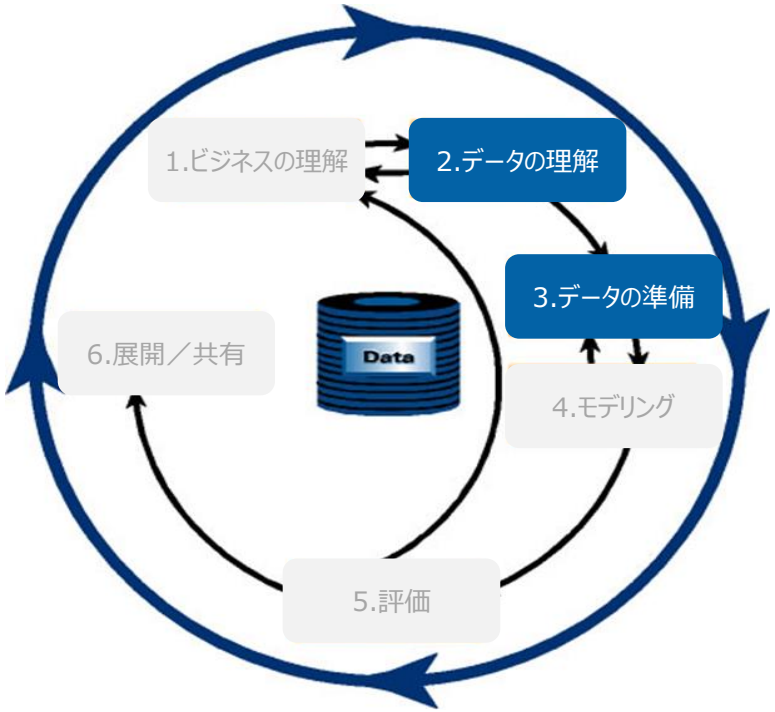
https://scikit-learn.org/stable/modules/clustering.html

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

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

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

※クラウド型のSAS Studio (SAS OnDemand for Academics) において
列名を日本語にする場合、全角6文字以内を推奨

クレジットカード債務不履行の有無
年間平均残高（ユーロ）
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

年齢	職業	結婚歴	学歴	クレカ債務	年間平均残高	住宅ローン	個人ローン	連絡手段	最終連絡日	最終連絡月	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数	前回CP結果	定期預金契約
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			23	feb	141	2	176	3	failure	
36	self-employed	married	tertiary	no	307	yes			14	may	341	1	330	2	other	
39	technician	married	secondary	no	147	yes			6	may	151	2	-1	0	unkno	
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no

4,521人

説明変数

目的変数

予測(分析)対象を説明するための変数

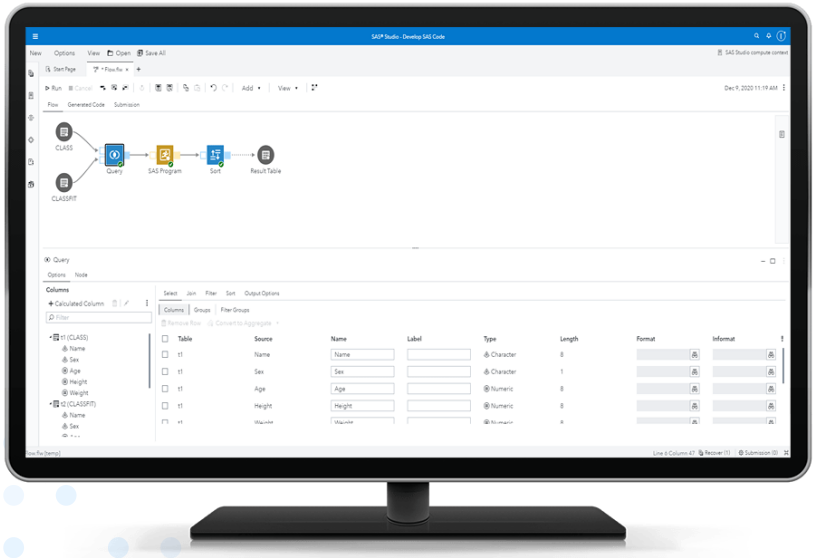
予測(分析)したい対象

再掲

SAS Studio について

- 今年のウェビナーでは、 **SAS Studio** でデモを行います。
- SAS Studio はすべてのSAS製品に付帯しているGUI で、今回は学習用に自宅でもお使い頂けるクラウド型無償版 **SAS OnDemand for Academics** を使っています。
  （※無償版の登録については、SAS からの申込完了メールをご参照ください）
- なお、SAS Studio起動時はコード入力画面となっていますが、画面右上の「SASプログラマ」を「**ビジュアルプログラマ**」に変更するとデモと同様の入力画面となります。

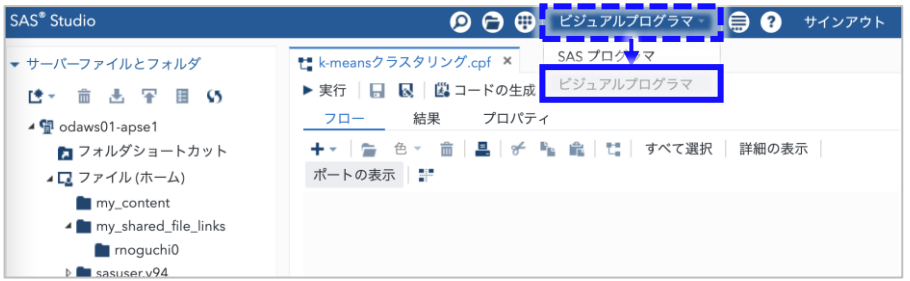
▼SAS Studio 画面イメージ



https://www.sas.com/ja_jp/software/studio.html

▼GUI画面への変更方法

（ビジュアルプログラマ）



再掲

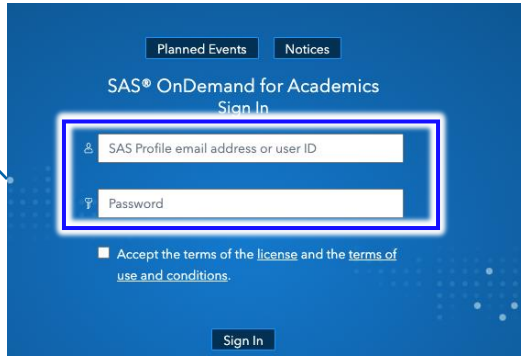
参考：SAS Studio 起動方法

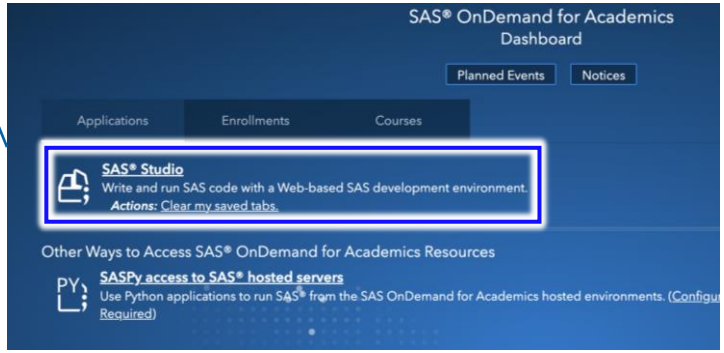
- SAS OnDemamd for Academics にログイン後、Dashboard より SAS Studio を起動
- 起動後、前頁の通り、右上メニューより「ビジュアルプログラマ」を選択

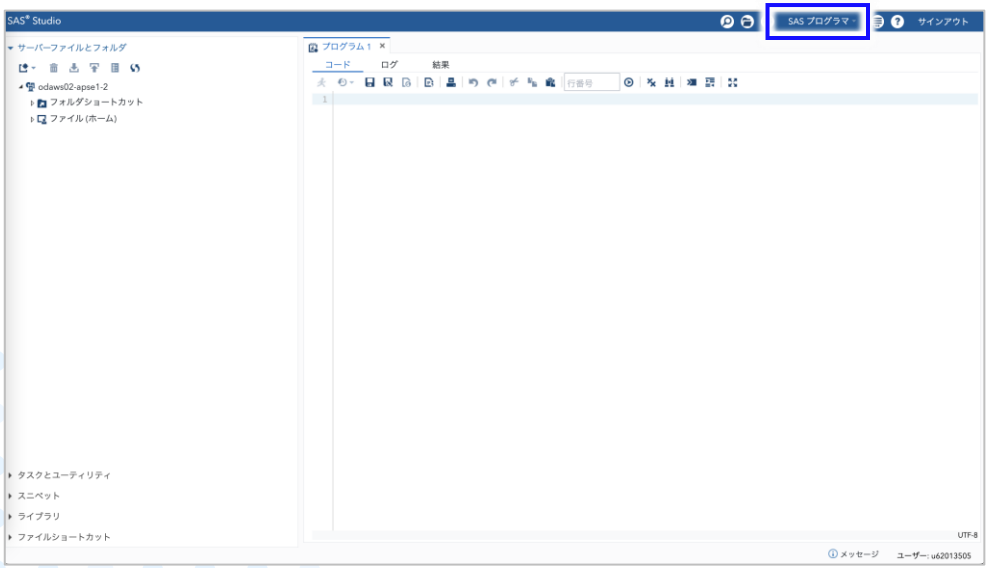
1 **SAS OnDemand for Academics** にログイン
https://welcome.oda.sas.com/login

2 SAS OnDemand for Academics Dashboardより
“SAS® Studio” をクリックして起動

3 SAS Studioが起動／右上よりビジュアルプログラマを選択

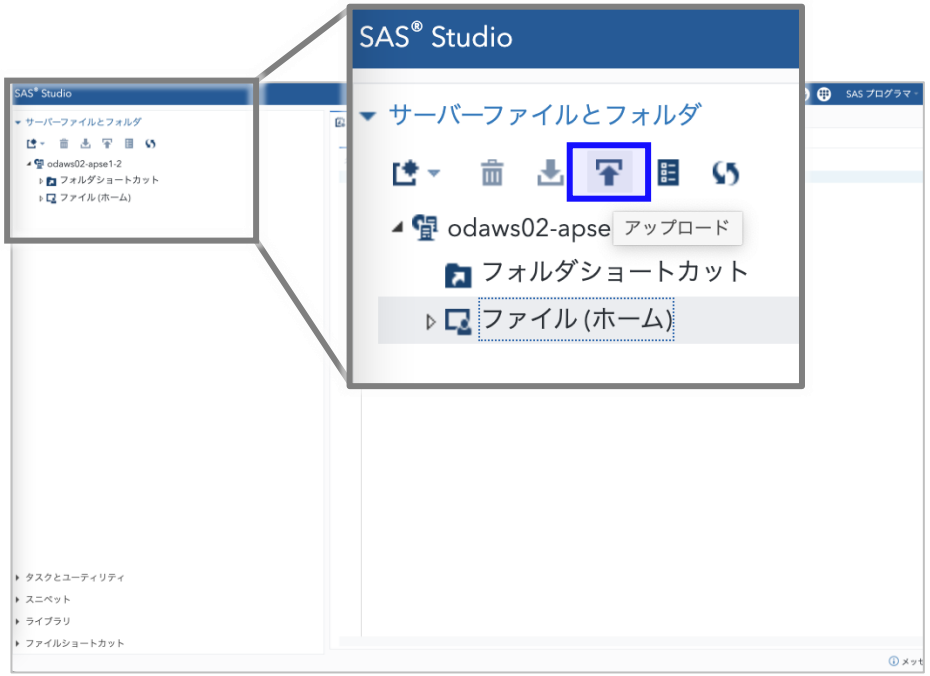




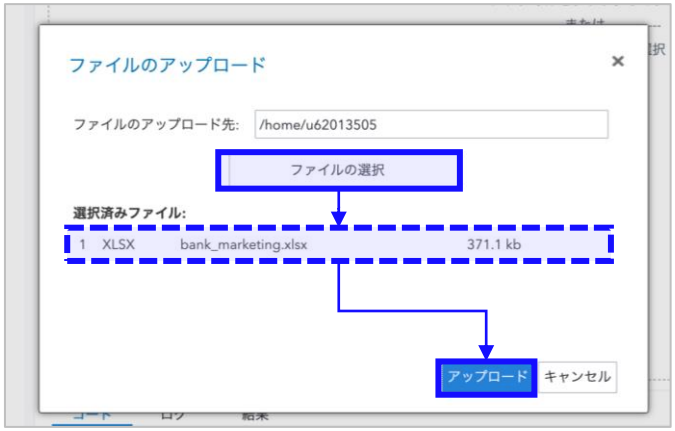


データの読み込み（1/2）

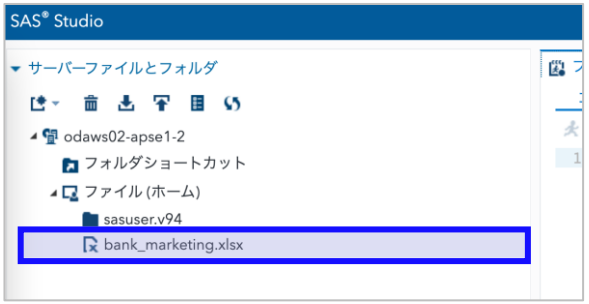
①左パネル内の「アップロード」アイコン をクリック



②「ファイルの選択」ボタンをクリックし、ファイル選択画面で“bank_marketing.xlsx” を選択し、OKボタン
③「アップロード」ボタンをクリック

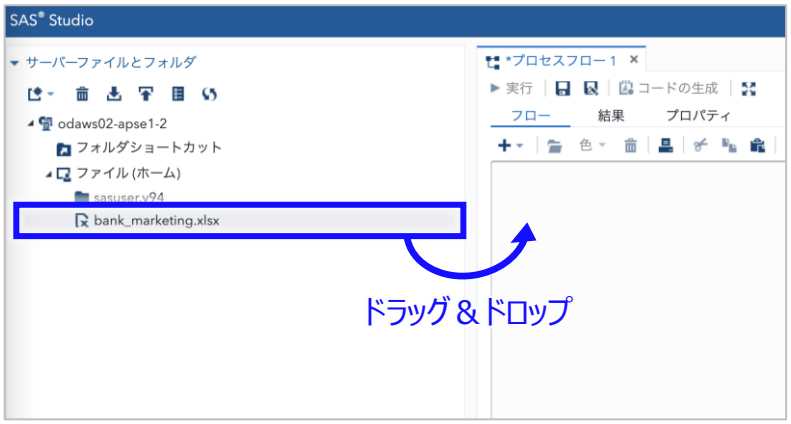


④左パネル内にファイルがアップロードされていることを確認

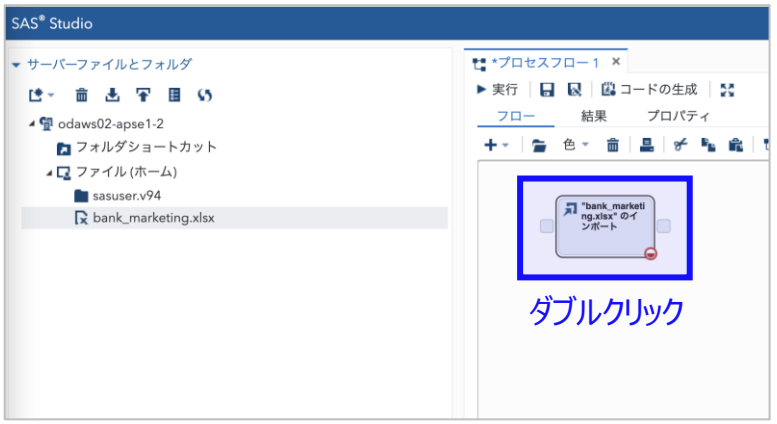


データの読み込み（2/2）

①左パネル内の “bank_marketing.xlsx” を選択し、
画面右側のプログラムエリアにドラッグ＆ドロップ



②右側のプロセスフローにノードが生成されるので、
当該ノードをダブルクリック



③詳細設定画面が開くので、実行ボタンをクリック
（特に各設定は変更不要）

④「結果」のタブ画面に読み込んだデータの概要が出力

読み込んだデータの確認

データ概要の確認

新しいブラウザタブで開く

CONTENTS プロシジャ

データセット名	WORK.IMPORT1	オブザベーション数	4521
メンバータイプ	DATA	変数の数	17
エンジン	V9		
作成日時	2022/08/08 09:34:47		
更新日時	2022/08/08 09:34:47		
保護			
データセットタイプ			
ラベル			
データ表現	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
エンコード	utf-8 Unicode (UTF-8)		

列数と行数を確認！
（ビッグデータ分析の基本）

エンジン/ホスト関連情報

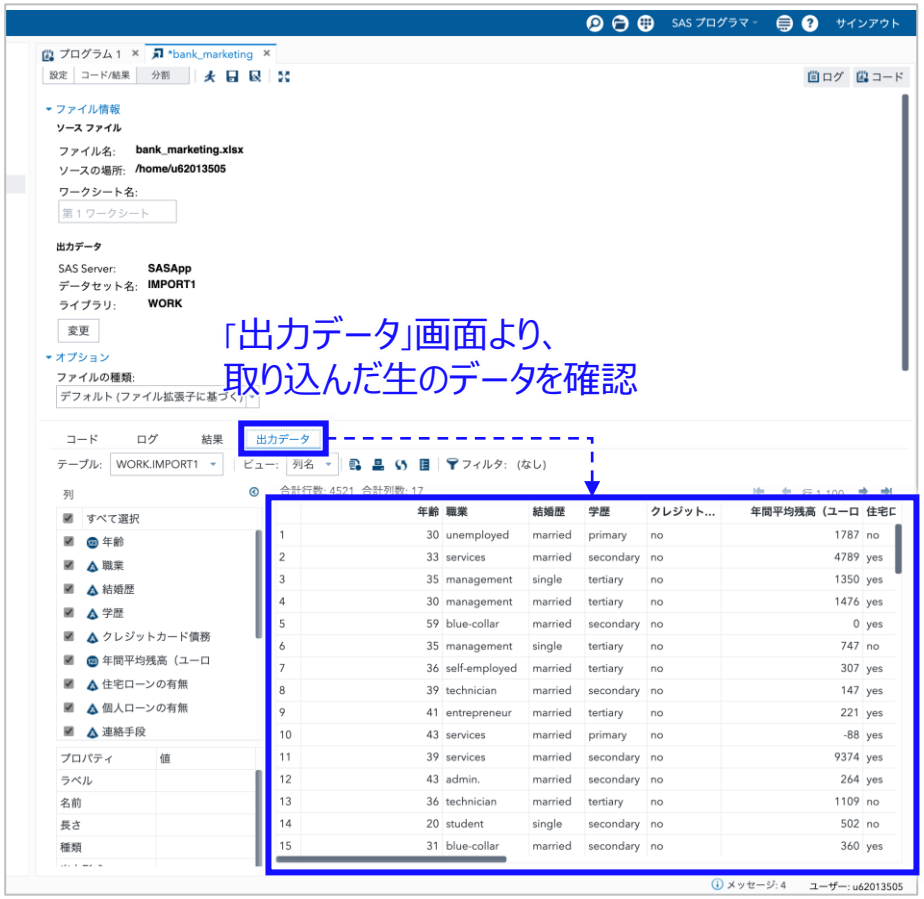
データセットのページサイズ	131072
データセットのページ数	5
データページの先頭	1
ページごとの最大OBS数	1090
先頭ページのOBS数	1052
データセットの修復数	0
ファイル名	/saswork/SAS_work71F80001F3FA_odaws01-apse1-2.oda.sas.com/SAS_workC7860001F3FA_odaws01-apse1-2.oda.sas.com/import1.sas7bdat
作成したリリース	9.0401M6
作成したホスト	Linux
I ノード番号	33850
アクセス権限	rw-r--r--
所有者名	u62013505
ファイルサイズ	768KB
ファイルサイズ (バイ	

各列のデータ型を確認

変数と属性リスト (アルファベット順)

#	変数	タイプ	長さ	出力形式	入力形式	ラベル
13	キャンペーン中の連絡	数値	8	BEST.		キャンペーン中の連絡回数
15	キャンペーン前の連絡	数値	8	BEST.		キャンペーン前の連絡回数
5	クレジットカード債務	文字	3	$3.	$3.	クレジットカード債務不履行有無
7	住宅ローンの有無	文字	3	$3.	$3.	住宅ローンの有無
8	個人ローンの有無	文字	3	$3.	$3.	個人ローンの有無
16	前回キャンペーンの結	文字	7	$7.	$7.	前回キャンペーンの結果
4	学歴	文字	9	$9.	$9.	学歴
17	定期預金契約有無	文字	3	$3.	$3.	定期預金契約有無
6	年間平均残高（ユーロ	数値	8	BEST.		年間平均残高（ユーロ）
1	年齢	数値	8	BEST.		年齢
14	最終連絡からの経過日	数値	8	BEST.		最終連絡からの経過日数
10	最終連絡日	数値	8	BEST.		最終連絡日
12	最終連絡時の会話時間	数値	8	BEST.		最終連絡時の会話時間（秒）
11	最終連絡月	文字	3	$3.	$3.	最終連絡月
3	結婚歴	文字	8	$8.	$8.	結婚歴
2	職業	文字	13	$13.	$13.	職業
9	連絡手段	文字	9	$9.	$9.	連絡手段

生データの確認

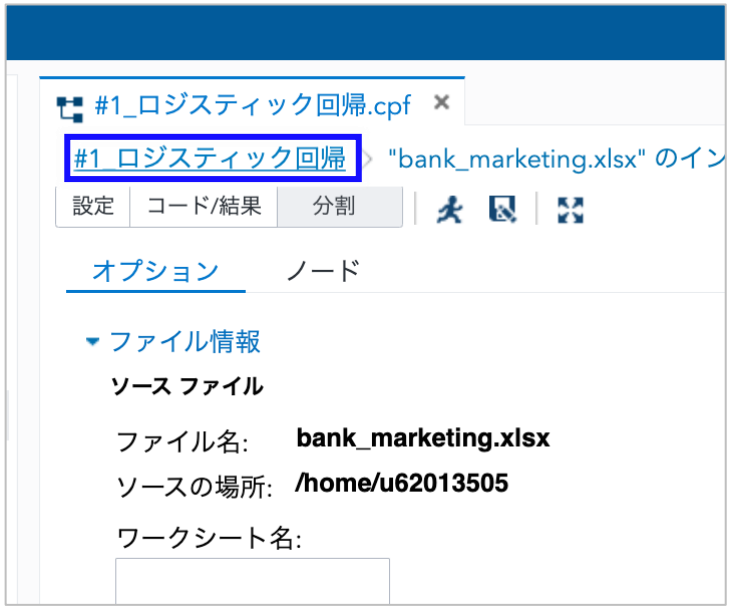


「出力データ」画面より、
取り込んだ生のデータを確認

	年齢	職業	結婚歴	学歴	クレジット...	年間平均残高（ユーロ	住宅ロ
1	30	unemployed	married	primary	no	1787	no
2	33	services	married	secondary	no	4789	yes
3	35	management	single	tertiary	no	1350	yes
4	30	management	married	tertiary	no	1476	yes
5	59	blue-collar	married	secondary	no	0	yes
6	35	management	single	tertiary	no	747	no
7	36	self-employed	married	tertiary	no	307	yes
8	39	technician	married	secondary	no	147	yes
9	41	entrepreneur	married	tertiary	no	221	yes
10	43	services	married	primary	no	-88	yes
11	39	services	married	secondary	no	9374	yes
12	43	admin.	married	secondary	no	264	yes
13	36	technician	married	tertiary	no	1109	no
14	20	student	single	secondary	no	502	no
15	31	blue-collar	married	secondary	no	360	yes

作成したプロセスフローの保存（別名で保存）

プロセスフローをクリックしてプロセスフロー画面に戻る



「名前を付けてプロセスフローを保存」アイコンをクリックし、
保存場所、ファイル名を指定して保存ボタン

再掲

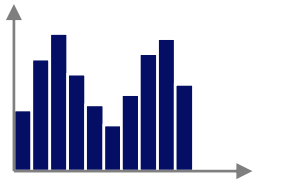
データの特徴の捉え方

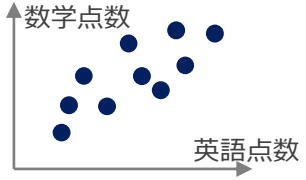
- ビッグデータでは個々のデータをくまなく見るのは難しいため、**グラフ**（ヒストグラムや散布図）や**要約統計量**（平均値や標準偏差） を用いて全体傾向を把握する

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	

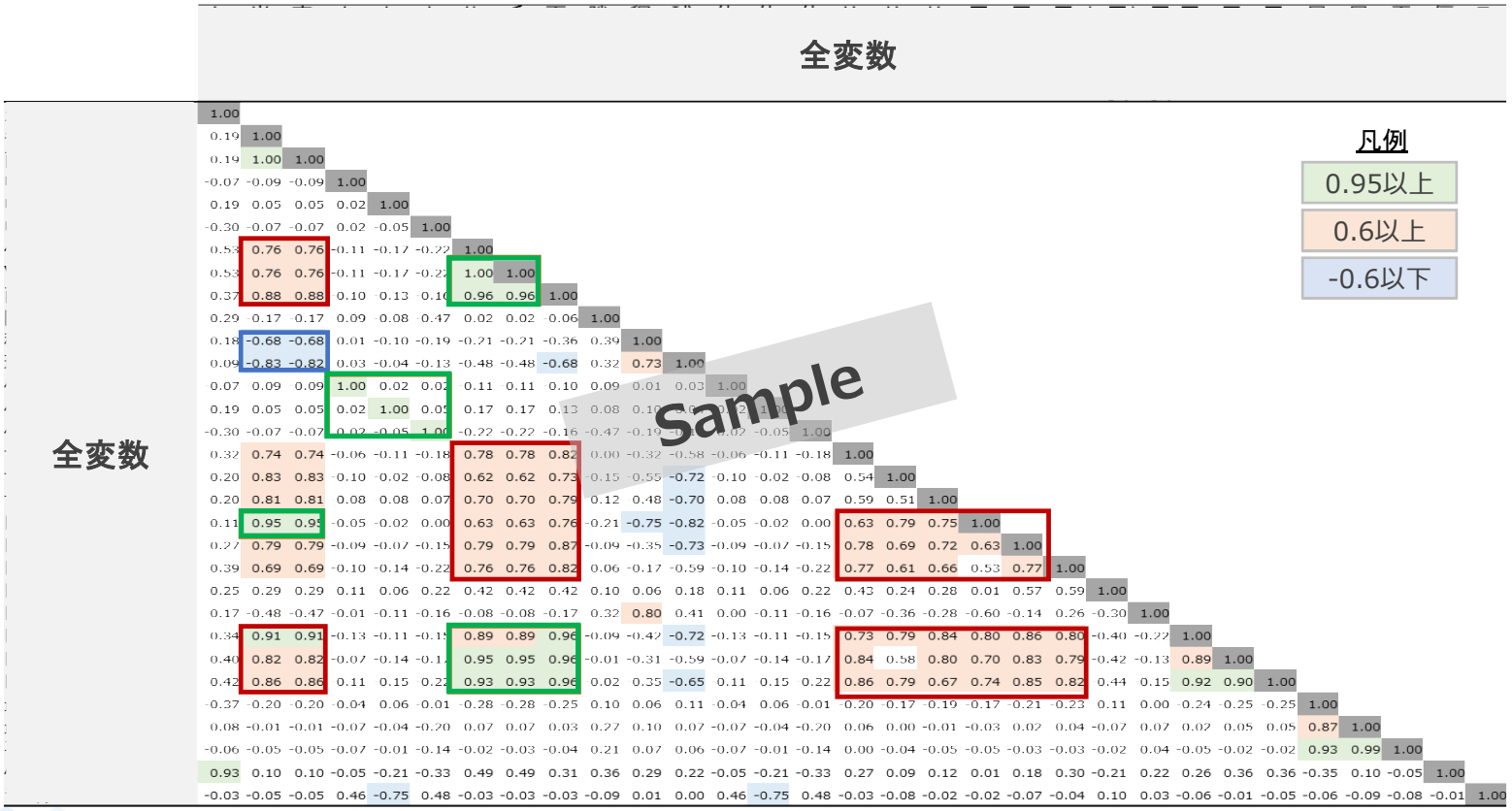
- 平均値＝ＸＸＸ
- 中央値＝ＸＸＸ
- 標準偏差＝ＸＸＸ

・
・
・

再掲

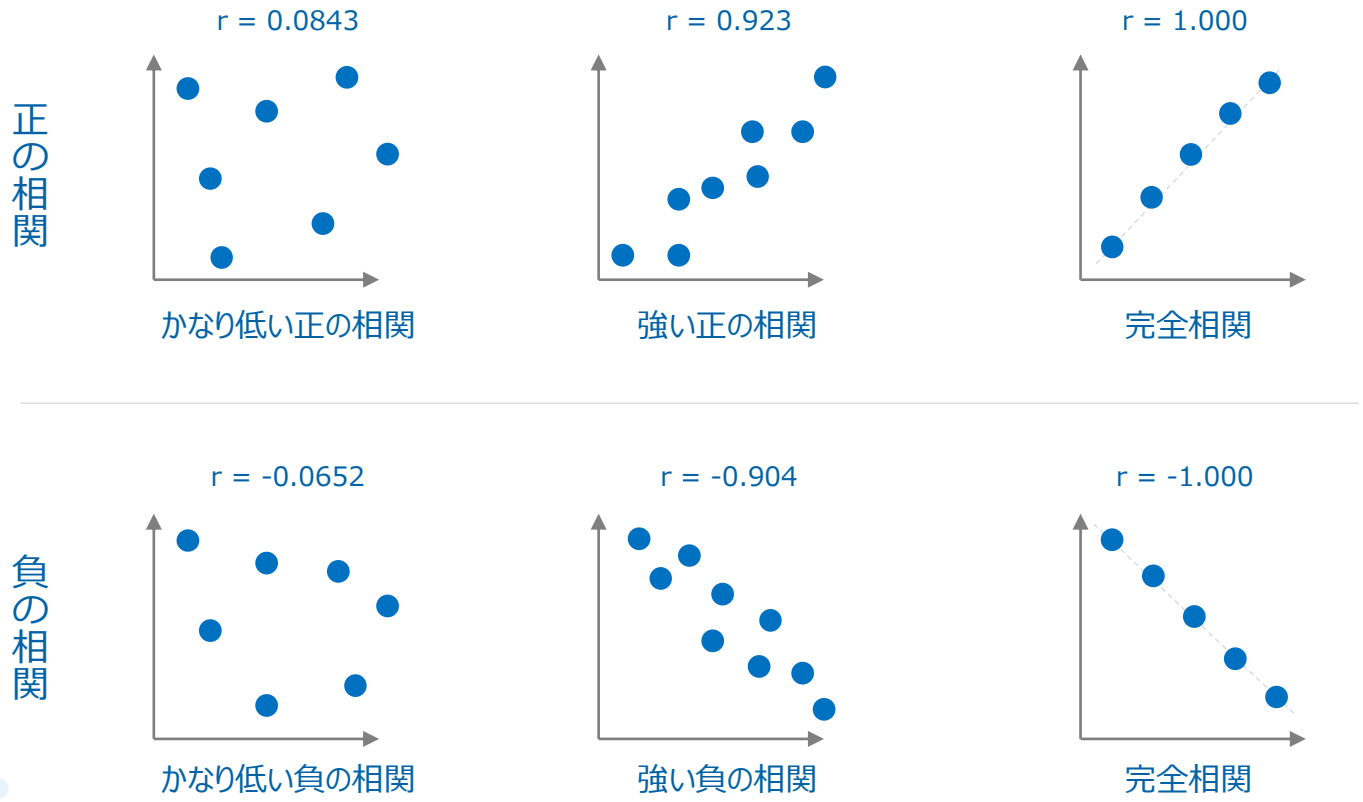
相関行列

- 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



（参考）**相関**係数の目安

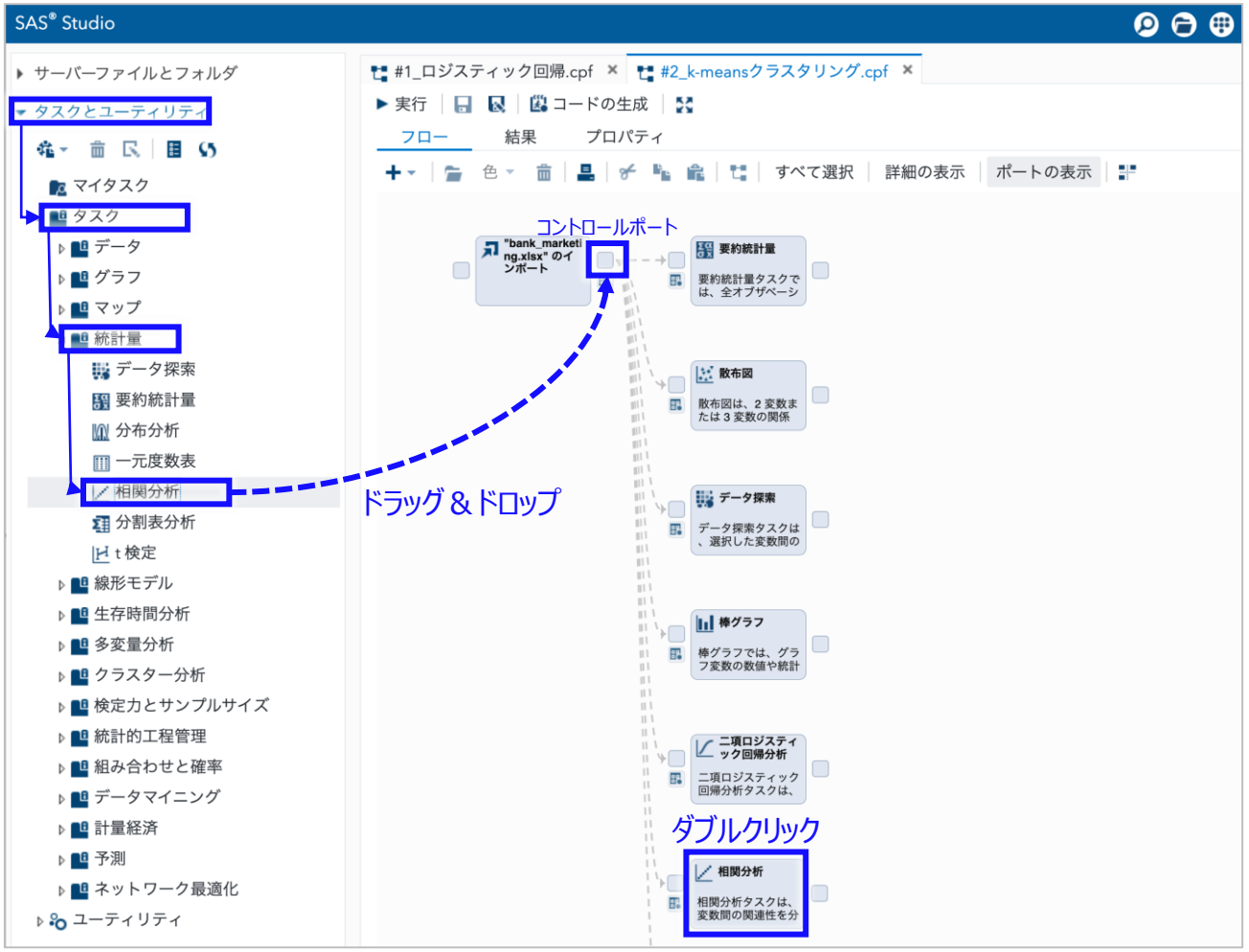
相関係数（絶対値）	相関の強弱
1.0	**強い相関**
0.7	**やや相関あり**
0.4	**弱い相関**
0.2	**ほぼ相関なし**
0	

SAS Studio での実装方法

- 相関分析
- 散布図との比較
- グループ分析を設定した相関分析

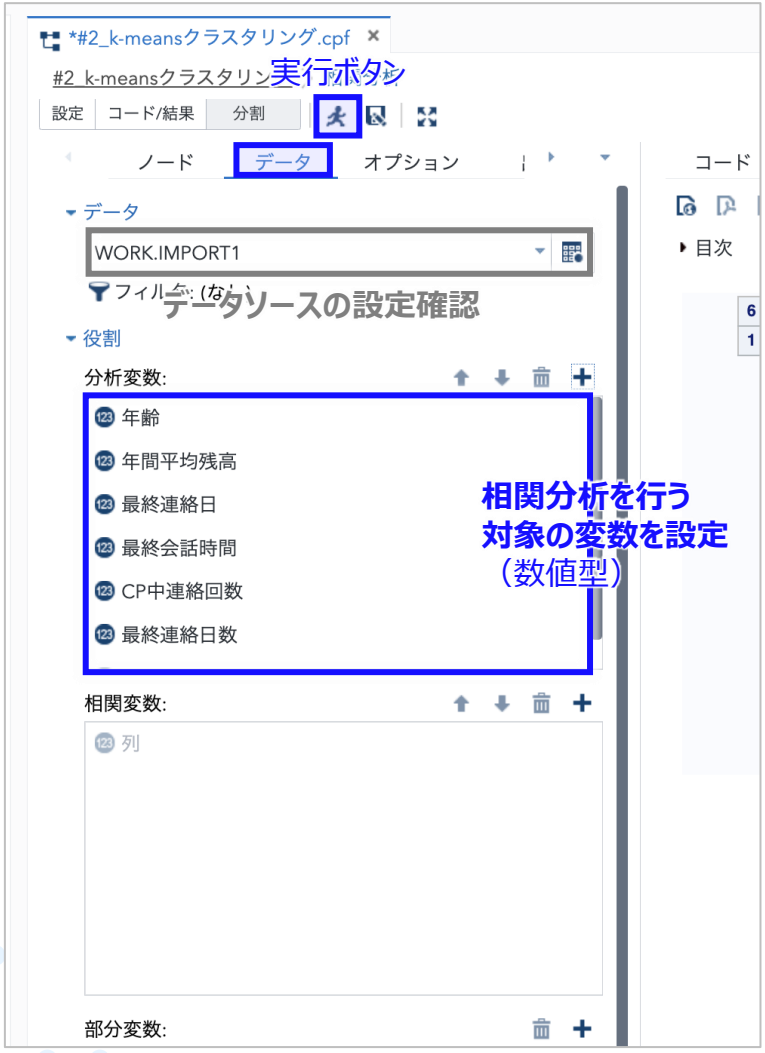
相関分析の出力 – 実行方法（1/2）

①[タスクとユーティリティ]→[タスク]→[統計量]→[**相関分析**] を選択し、
　データインポートノードのコントロールポートに ドラッグ&ドロップ
②生成された [相関分析] ノードをダブルクリックして、詳細設定画面を開く

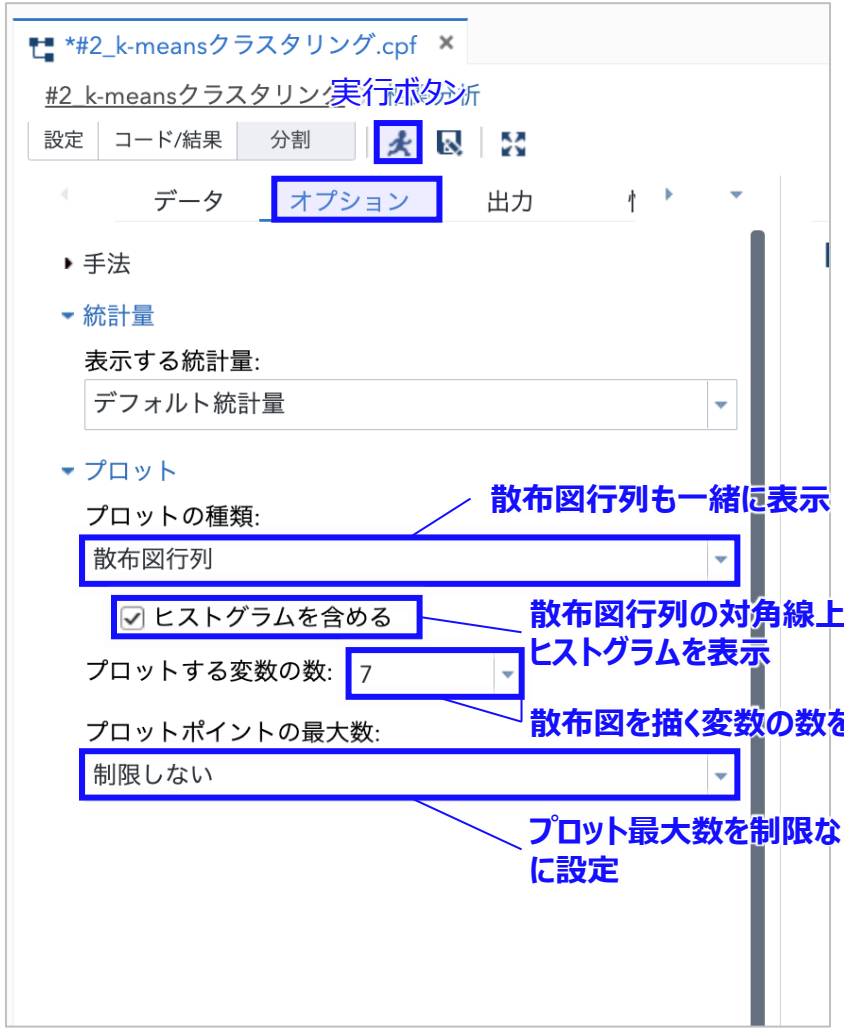


相関分析の出力 – 実行方法（2/2）

[データ]の設定



[オプション]の設定



相関分析の出力 – 実行結果（相関行列）

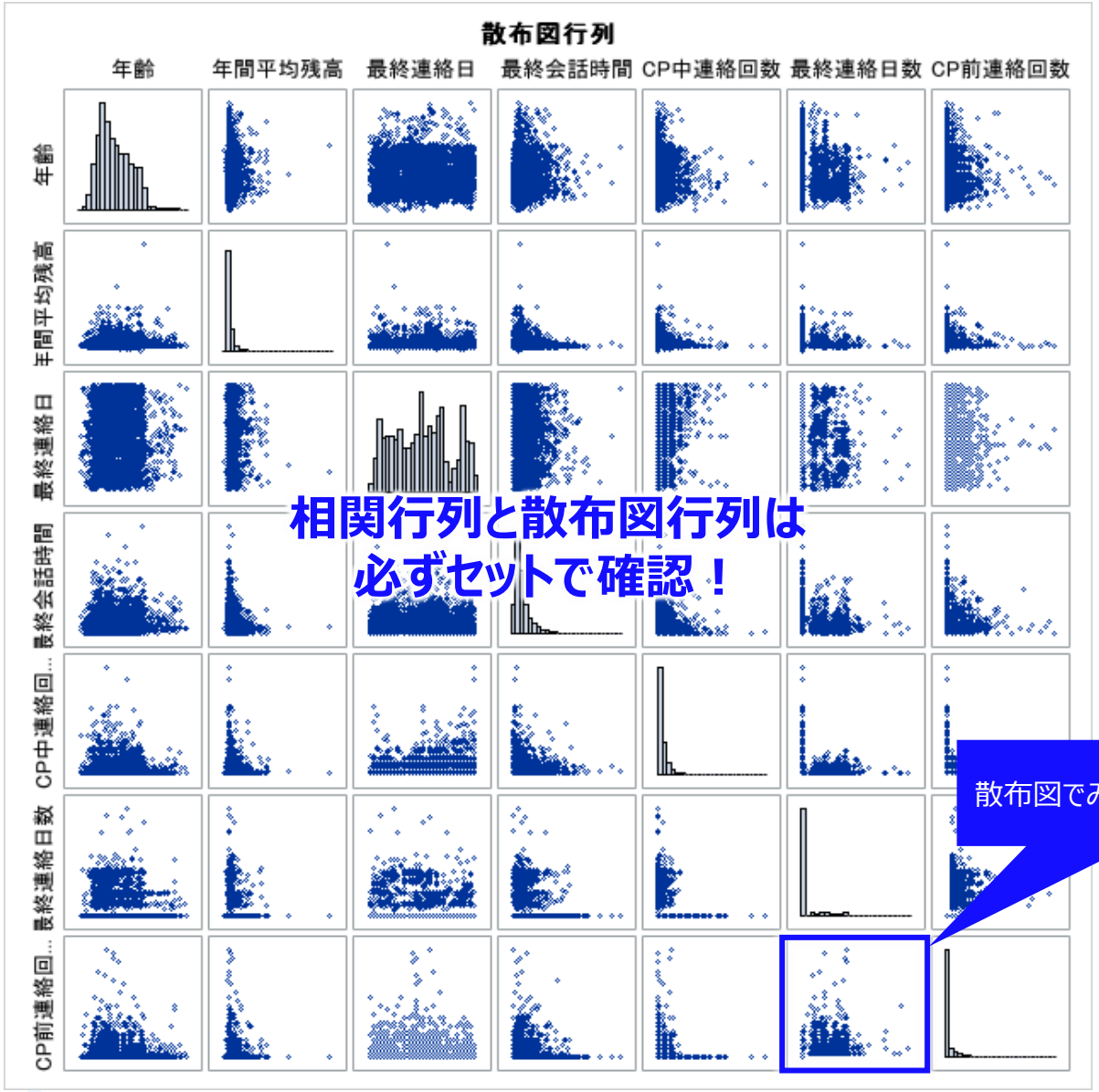
Pearson の相関係数, N = 4521							
	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
年齢 年齢	1.00000	0.08382	-0.01785	-0.00237	-0.00515	-0.00889	-0.00351
年間平均残高 年間平均残高	0.08382	1.00000	-0.00868	-0.01595	-0.00998	0.00944	0.02620
最終連絡日 最終連絡日	-0.01785	-0.00868	1.00000	-0.02463	0.16071	-0.09435	-0.05911
最終会話時間 最終会話時間	-0.00237	-0.01595	-0.02463	1.00000	-0.06838	0.01038	0.01808
CP中連絡回数 CP中連絡回数	-0.00515	-0.00998	0.16071	-0.06838	1.00000	-0.09314	-0.06783
最終連絡日数 最終連絡日数	-0.00889	0.00944	-0.09435	0.01038	-0.09314	1.00000	0.57756
CP前連絡回数 CP前連絡回数	-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

「最終連絡からの経過日数」と「キャンペーン前の連絡回数」とで相関係数が高い

（参考）**相関**係数の目安

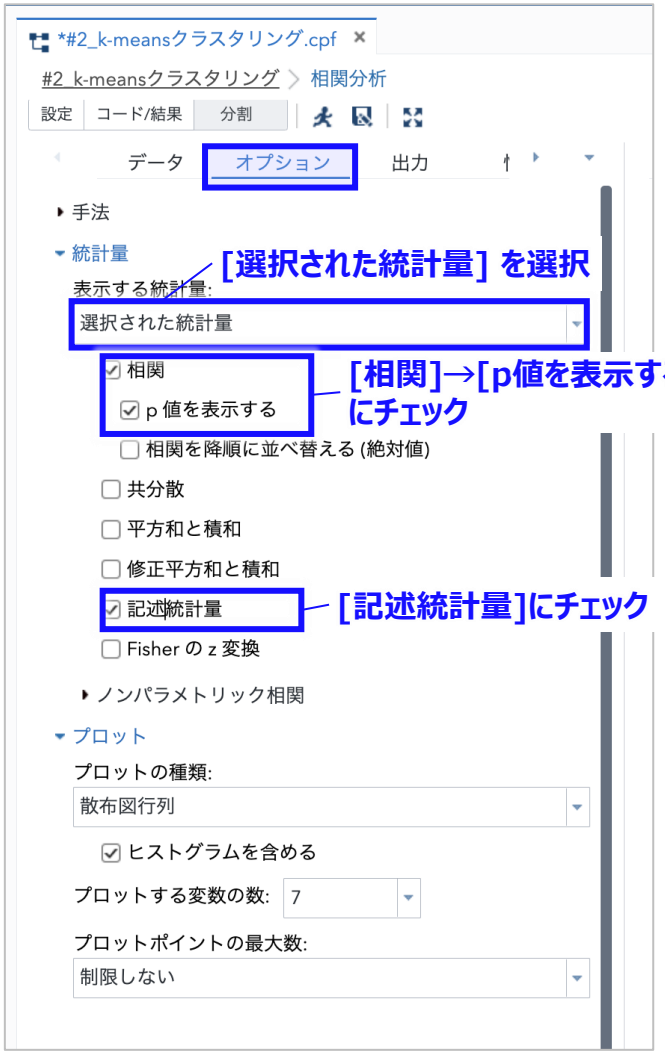
相関分析の出力 – 実行結果（散布図行列）



SAS® Studio

（参考）相関分析の出力：各種統計量・p値の表示

- オプションで追加設定をすることで、基本的な要約統計量や相関係数のp値も同時出力可能



基本的な要約統計量

単純統計量

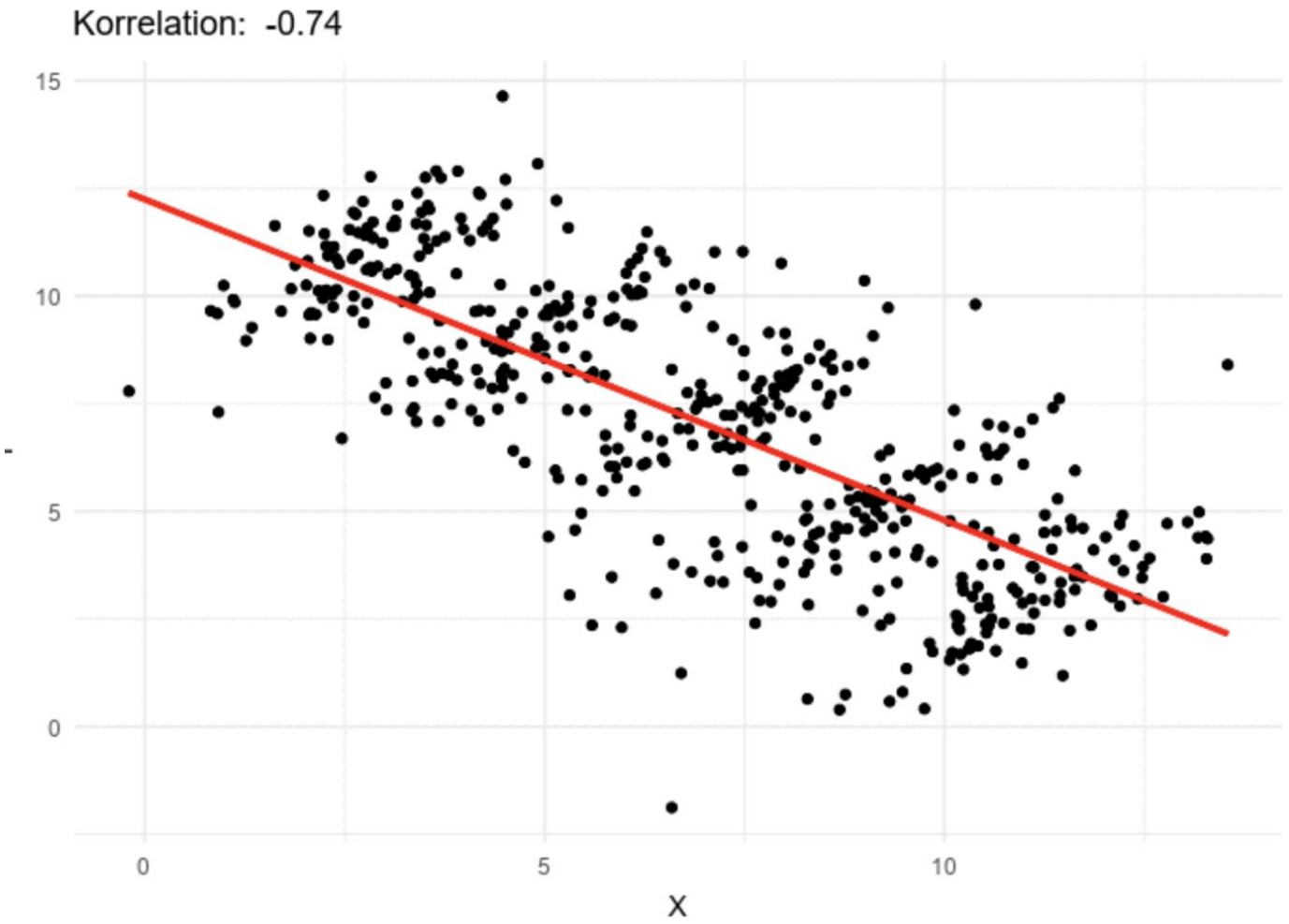
変数	N	平均	標準偏差	合計	最小値	最大値	ラベル
年齢	4521	41.17010	10.57621	186130	19.00000	87.00000	年齢
年間平均残高	4521	1423	3010	6431836	-3313	71188	年間平均残高
最終連絡日	4521	15.91528	8.24767	71953	1.00000	31.00000	最終連絡日
最終会話時間	4521	263.96129	259.85663	1193369	4.00000	3025	最終会話時間
CP中連絡回数	4521	2.79363	3.10981	12630	1.00000	50.00000	CP中連絡回数
最終連絡日数	4521	39.76664	100.12112	179785	-1.00000	871.00000	最終連絡日数
CP前連絡回数	4521	0.54258	1.69356	2453	0	25.00000	CP前連絡回数

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	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
年齢	1.00000	0.08382	-0.01785	-0.00237	-0.00515	-0.00889	-0.00351
年齢		<.0001	0.2301	0.8736	0.7293	0.5500	0.8134
年間平均残高	0.08382	1.00000	-0.00868	-0.01595	-0.00998	0.00944	0.02620
年間平均残高	<.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
最終連絡日	0.2301	0.5597		0.0978	<.0001	<.0001	<.0001
最終会話時間	-0.00237	-0.01595	-0.02463	1.00000	-0.06838	0.01038	0.01808
最終会話時間	0.8736	0.2836	0.0978		<.0001	0.4853	0.2242
CP中連絡回数	-0.00515	-0.00998	0.16071	-0.06838	1.00000	-0.09314	-0.06783
CP中連絡回数	0.7293	0.5025	<.0001	<.0001		<.0001	<.0001
最終連絡日数	-0.00889	0.00944	-0.09435	0.01038	-0.09314	1.00000	0.57756
最終連絡日数	0.5500	0.5259	<.0001	0.4853	<.0001		<.0001
CP前連絡回数	-0.00351	0.02620	-0.05911	0.01808	-0.06783	0.57756	1.00000
CP前連絡回数	0.8134	0.0782	<.0001	0.2242	<.0001	<.0001	

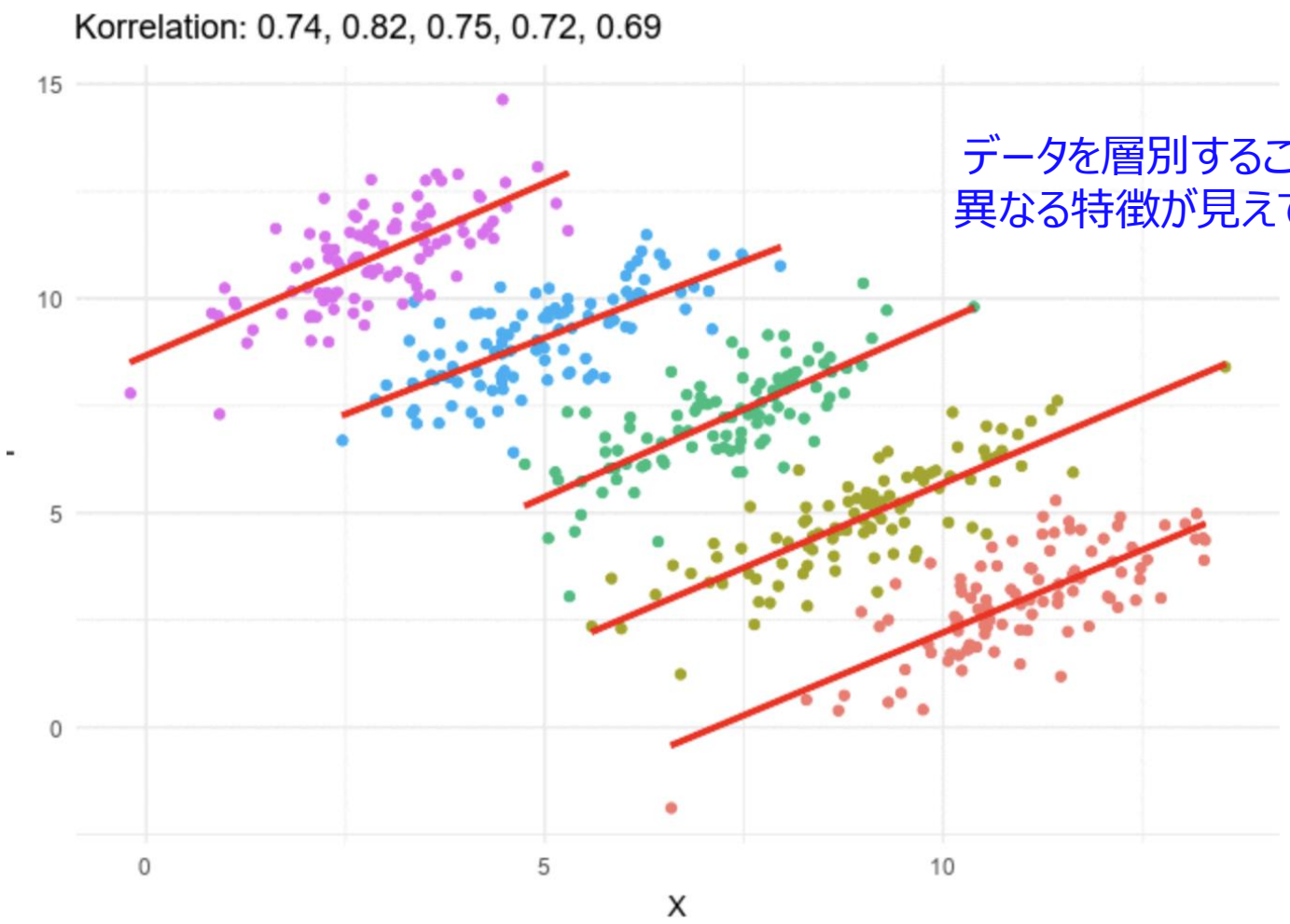
相関係数のp値

データ層別の重要性



出典：シンプソンのパラドックス (Wikipedia) https://w.wiki/qrf

データ層別の重要性



データを層別することで
異なる特徴が見えてくる

出典：シンプソンのパラドックス (Wikipedia)　https://w.wiki/qrf

相関分析の出力：グループ分析 – 実行方法

- グループ分析変数に目的変数を設定することで、目的変数で層別した相関分析が可能

契約者

Pearson の相関係数, N = 521

	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
年齢 年齢	1.00000	0.16845	-0.05207	-0.03633	-0.06583	0.05072	-0.01192
年間平均残高 年間平均残高	0.16845	1.00000	-0.03858	-0.12007	-0.02804	0.01352	0.02050
最終連絡日 最終連絡日	-0.05207	-0.03858	1.00000	0.03610	0.13780	-0.03734	-0.05123
最終会話時間 最終会話時間	-0.03633	-0.12007	0.03610	1.00000	0.23432	-0.15489	-0.15549
CP中連絡回数 CP中連絡回数	-0.06583	-0.02804	0.13780	0.23432	1.00000	-0.08488	-0.09863
最終連絡日数 最終連絡日数	0.05072	0.01352	-0.03734	-0.15489	-0.08488	1.00000	0.51823
CP前連絡回数 CP前連絡回数	-0.011[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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

未契約者

Pearson の相関係数, N = 4000

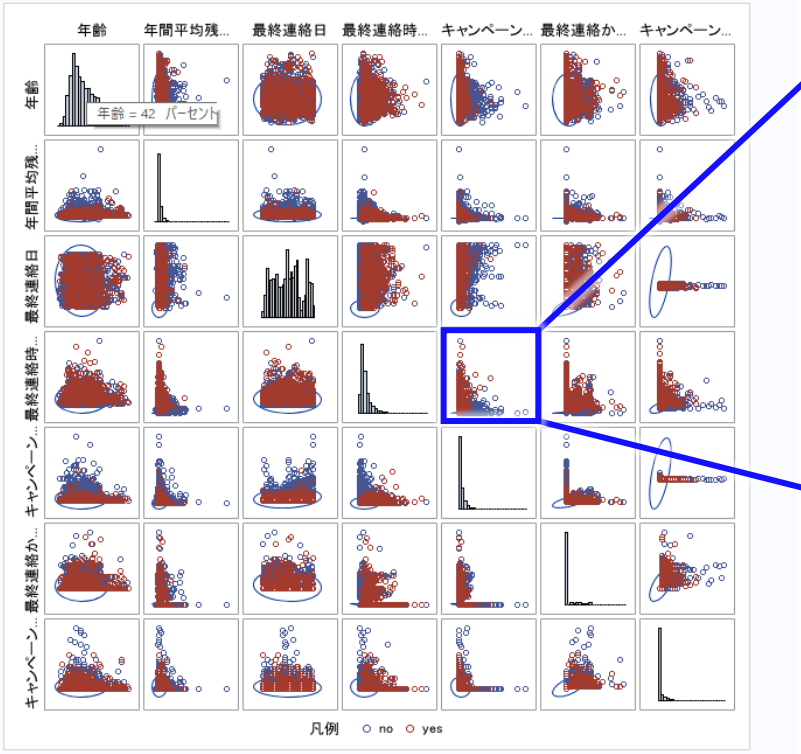
	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
年齢 年齢	1.00000	0.07291	-0.01165	-0.01836	0.00446		
年間平均残高 年間平均残高	0.07291	1.00000	-0.00539	-0.00858	-0.00762		
最終連絡日 最終連絡日	-0.01165	-0.00539	1.00000	-0.03679	0.16339		
最終会話時間 最終会話時間	-0.01836	-0.00858	-0.03679	1.00000	-0.09611		
CP中連絡回数 CP中連絡回数	0.00446	-0.00762	0.16339	-0.09611	1.00000		
最終連絡日数 最終連絡日数	-0.02733	0.00694	-0.10334	0.00179	-0.08961		
CP前連絡回数 CP前連絡回数	-0.00819	0.02507	-0.05957	0.00563	-0.05856	0.58368	1.00000

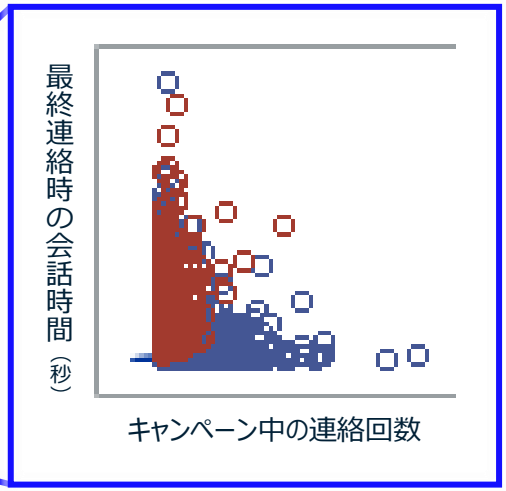
（参考）相関係数の目安

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	

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	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
年齢 年齢	1.00000	0.16845	-0.05207	-0.03633	-0.06583	0.05072	-0.01192
年間平均残高 年間平均残高	0.16845	1.00000	-0.03858	-0.12007	-0.02804	0.01352	0.02050
最終連絡日 最終連絡日	-0.05207	-0.03858	1.00000	0.03610	0.13780	-0.03734	-0.05123
最終会話時間 最終会話時間	-0.03633	-0.12007	0.03610	1.00000	0.23432	-0.15489	-0.15549
CP中連絡回数 CP中連絡回数	-0.06583	-0.02804	0.13780	0.23432	1.00000	-0.08488	-0.09863
最終連絡日数 最終連絡日数	0.05072	0.01352	-0.03734	-0.15489	-0.08488	1.00000	0.51823
CP前連絡回数 CP前連絡回数	-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

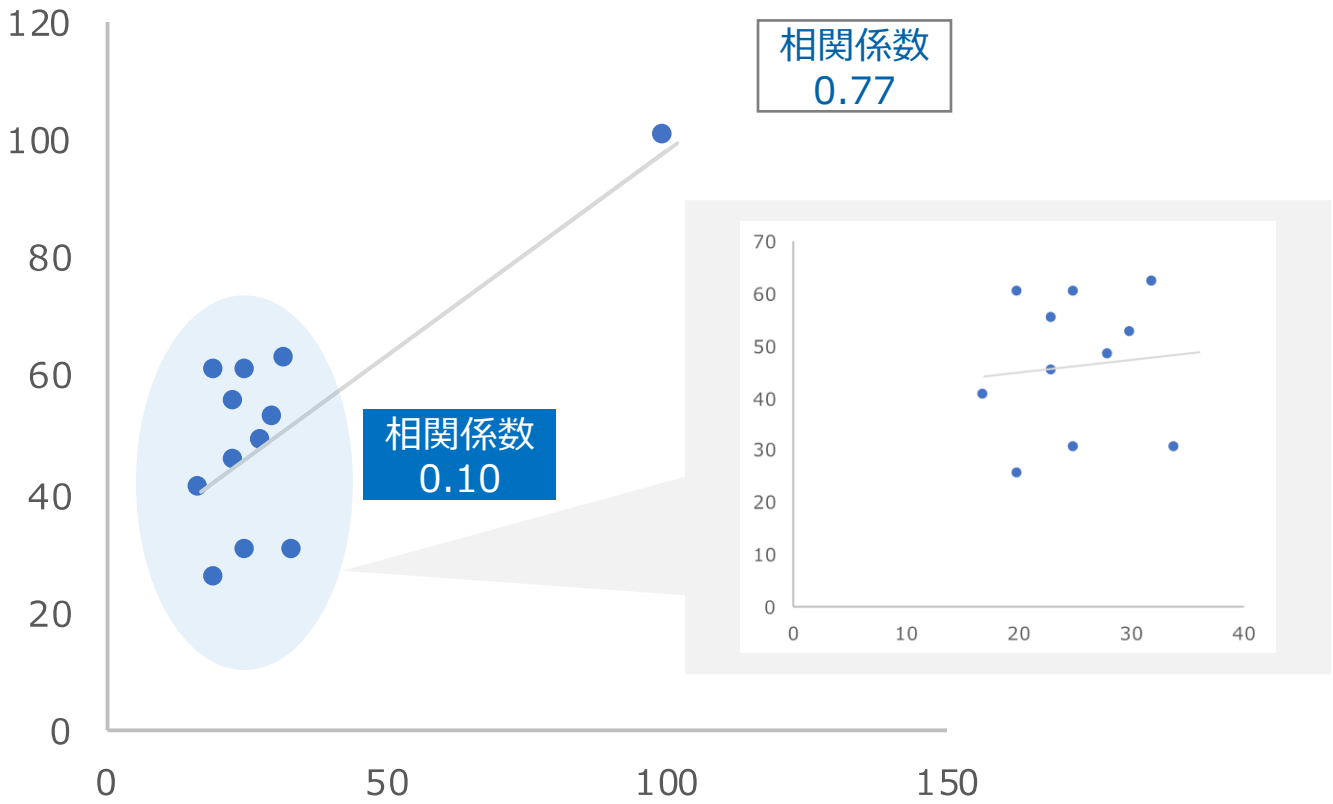
	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
年齢 年齢	1.00000	0.07291	-0.01165	-0.01836	0.00446	-0.02733	-0.00819
年間平均残高 年間平均残高	0.07291	1.00000	-0.00539	-0.00858	-0.00762	0.00694	0.02507
最終連絡日 最終連絡日	-0.01165	-0.00539	1.00000	-0.03679	0.16339	-0.10334	-0.05957
最終会話時間 最終会話時間	-0.01836	-0.00858	-0.03679	1.00000	-0.09611	0.00179	0.00563
CP中連絡回数 CP中連絡回数	0.00446	-0.00762	0.16339	-0.09611	1.00000	-0.08961	-0.05856
最終連絡日数 最終連絡日数	-0.02733	0.00694	-0.10334	0.00179	-0.08961	1.00000	0.58368
CP前連絡回数 CP前連絡回数	-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



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。

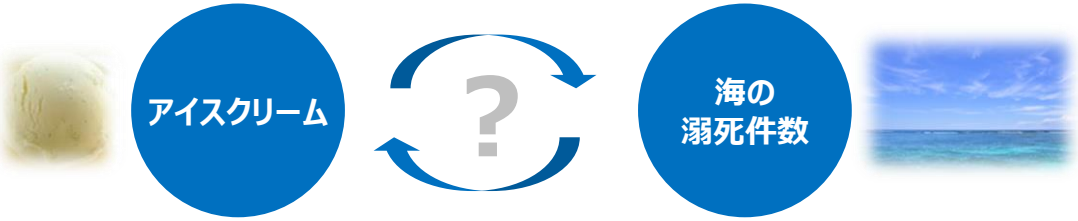
調査の結果、

「**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。

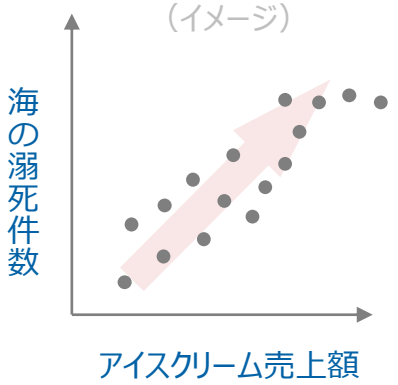
これが事実なら、即刻、**アイスクリームの販売に規制をかけるべき**である。

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

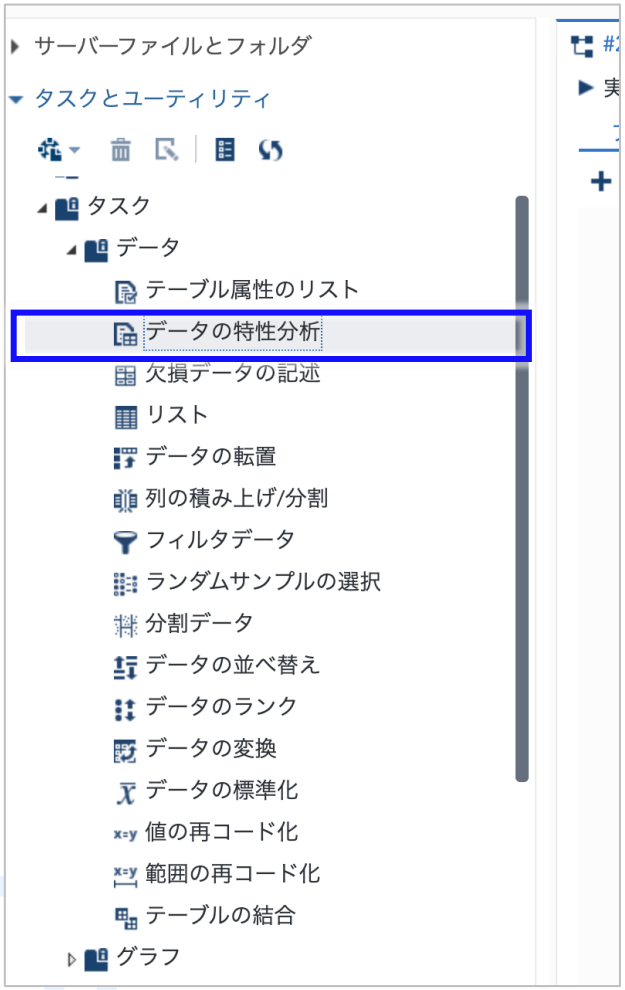
参考：変数の尺度（名義尺度・順序尺度・間隔尺度・比例尺度）

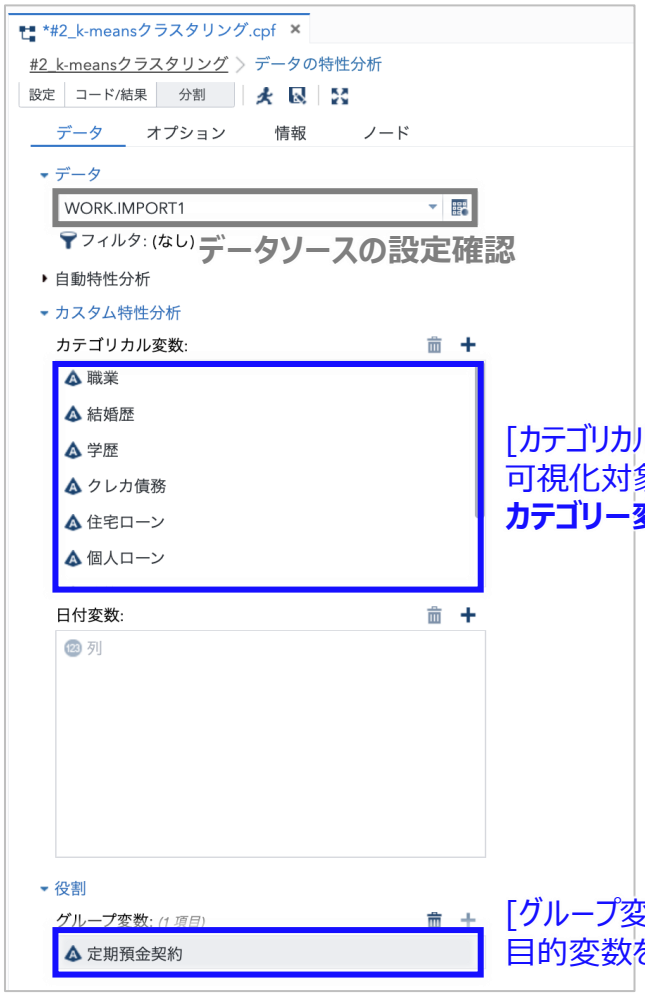
- 変数の種類は大きく「**質的データ**」と「**量的データ**」に分けられ、それぞれの特性に合わせて扱う必要がある

種類	変数の尺度	概要	データの例	扱い方 大小 (A<B) 差分 (A-B) 比率 (A/B)
質的データ (カテゴリーデータ)	名義尺度	単にデータを区別するための分類ラベル。 演算不可で、順序も意味をなさない	▪性別、血液型、顧客ID ▪作業者、個品ID、 良品/不良品	– – – ※集計によるカウントのみ可能
	順序尺度	**順序**（大小関係）にのみ意味がある尺度。 したがって、平均値は意味を持たないが、順序統計量（最大・最小など）は算出可能	▪顧客満足度、震度 ▪不良レベル、工程順序	● – –
量的データ (数量データ)	間隔尺度	数値演算可能だが、**値の差**のみに意味がある尺度。 0はあくまで相対的な位置関係でしかない	▪年齢、西暦、偏差値 ▪温度（℃）、製造日時	● ● –
	比例尺度	数値演算可能で、値の差に加え、**値の比**にも意味がある尺度。 0が「何もない」という絶対的な意味を持つ	▪身長、売上金額 ▪寸法、圧力、作業時間、絶対温度	● ● ●

カテゴリー変数可視化の便利な方法

- データの特性分析ノードを使えば、カテゴリー変数の頻度集計棒グラフを一度に出力可能





[カテゴリカル変数]内で
可視化対象の
カテゴリー変数を選択

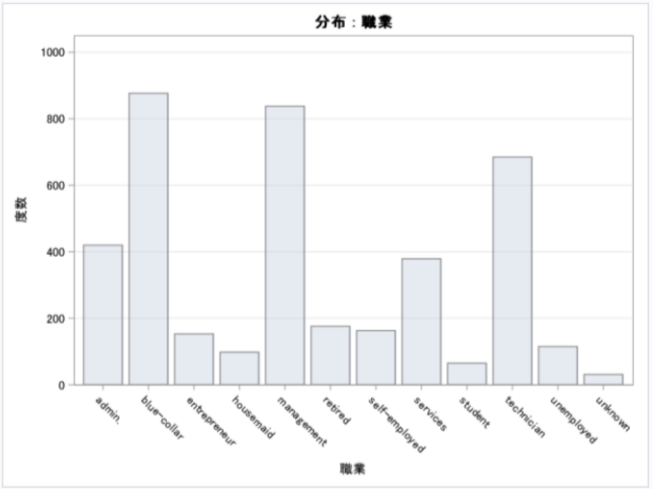
[グループ変数]に
目的変数を設定

カテゴリー変数可視化の便利な方法

- データの特性分析ノードを使えば、カテゴリー変数の頻度集計棒グラフを一度に出力可能

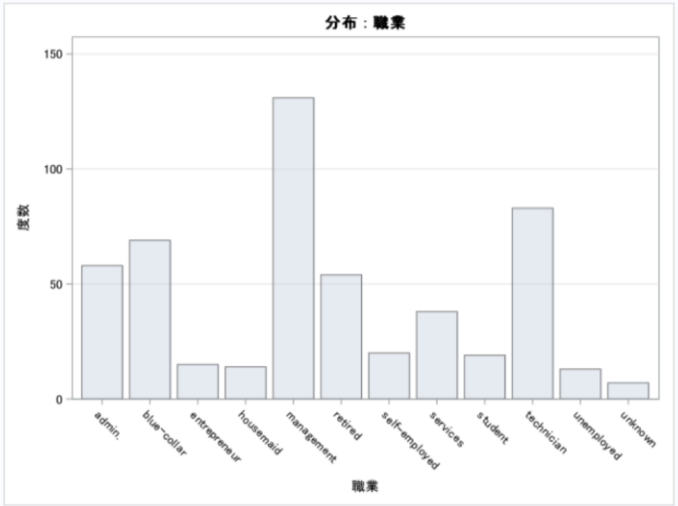
定期預金契約=no

職業	度数	パーセント	累積度数	累積パーセント
admin.	420	10.50	420	10.50
blue-collar	877	21.93	1297	32.43
entrepreneur	153	3.83	1450	36.25
housemaid	98	2.45	1548	38.70
management	838	20.95	2386	59.65
retired	176	4.40	2562	64.05
self-employed	163	4.08	2725	68.13
services	379	9.48	3104	77.60
student	65	1.63	3169	79.23
technician	685	17.13	3854	96.35
unemployed	115	2.88	3969	99.23
unknown	31	0.78	4000	100.00



定期預金契約=yes

職業	度数	パーセント	累積度数	累積パーセント
admin.	58	11.13	58	11.13
blue-collar	69	13.24	127	24.38
entrepreneur	15	2.88	142	27.26
housemaid	14	2.69	156	29.94
management	131	25.14	287	55.09
retired	54	10.36	341	65.45
self-employed	20	3.84	361	69.29
services	38	7.29	399	76.58
student	19	3.65	418	80.23
technician	83	15.93	501	96.16
unemployed	13	2.50	514	98.66
unknown	7	1.34	521	100.00

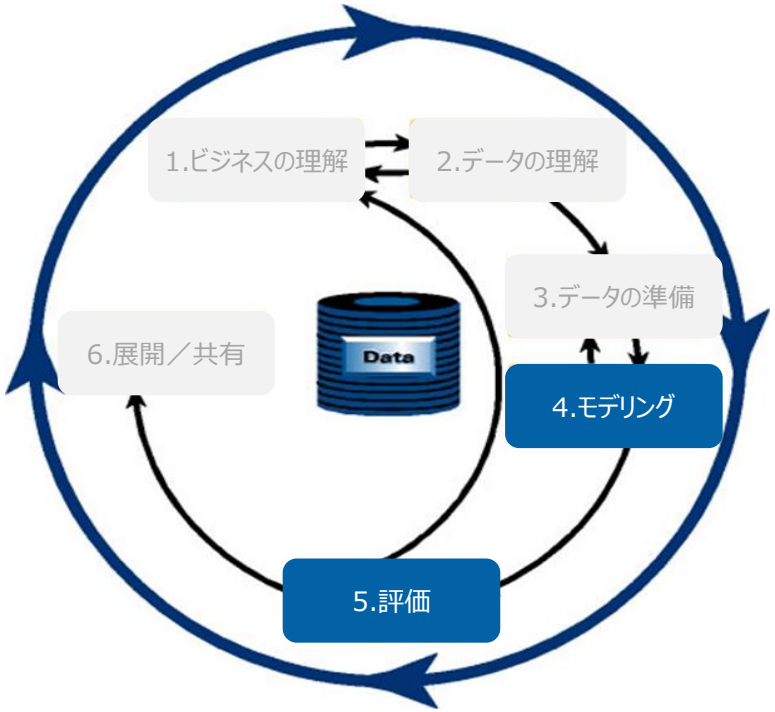


再掲

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)

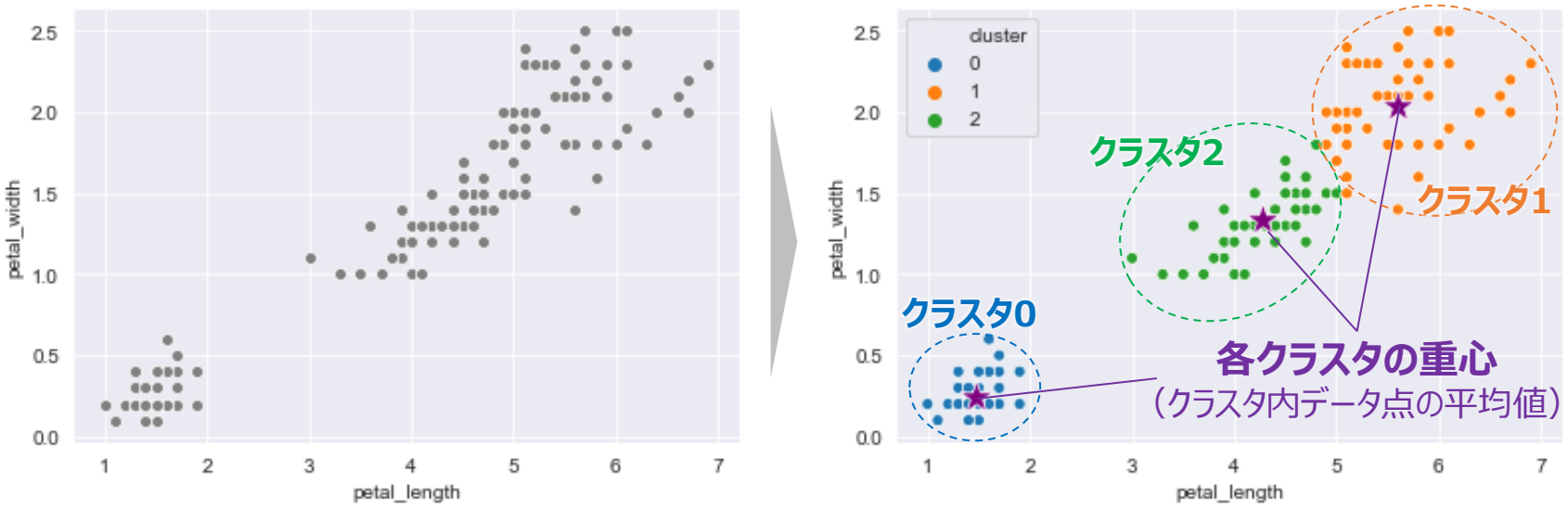
1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル作成	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

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

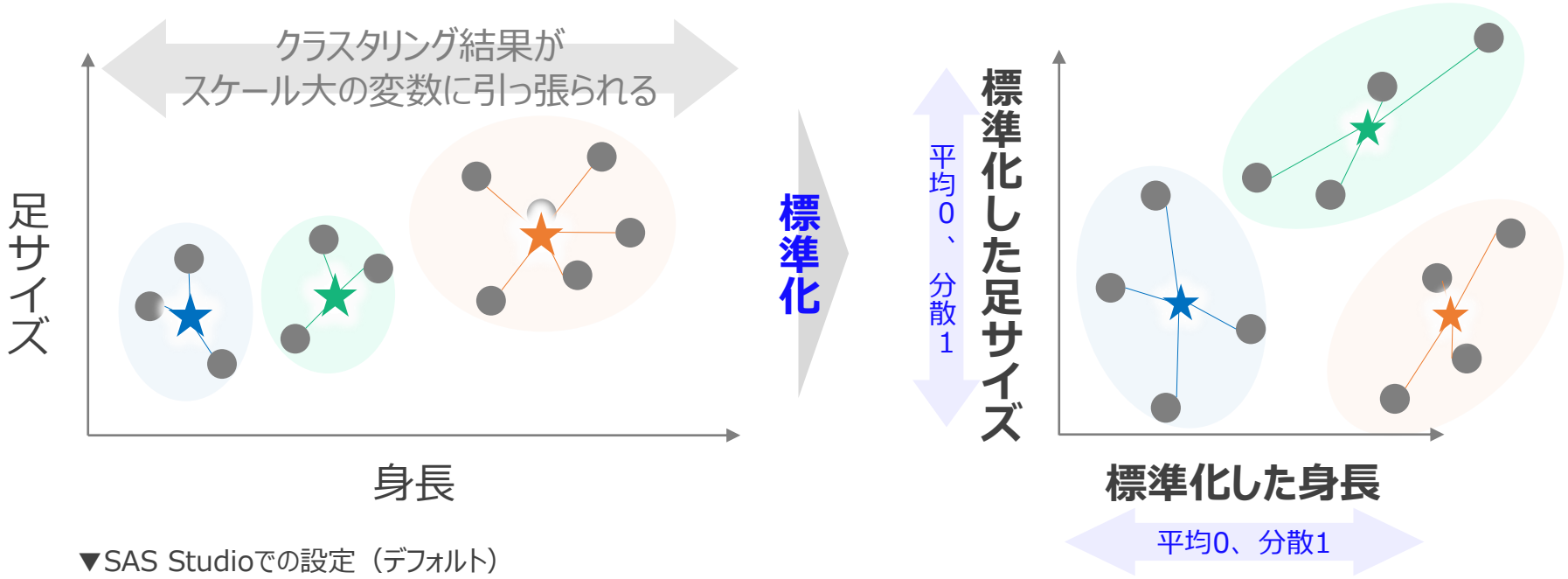


▼分類結果の特徴
- 教師なしのため、各クラスタの意味解釈は人が行う
- 円状（球状）のクラスタになりやすい
- クラスタサイズ（クラスタ内のデータ数）が同程度になりやすい

▼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

（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

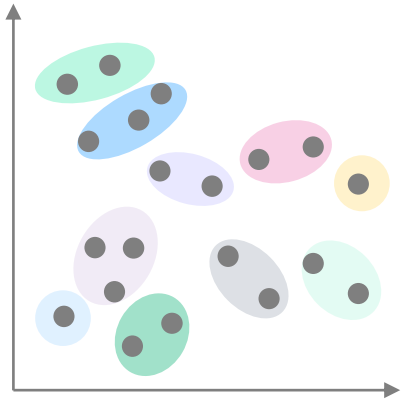
- k-means法などの「距離」に基づくによるクラスタリング手法は、データの「**スケール**」に大きく影響を受ける。このため、必要に応じて、「**標準化**」の処理を行なった上でクラスタリングが必要
- SAS Studioでは、クラスタリング時に、デフォルトで標準化が適用できる（オフにすることも可能）



▼SAS Studioでの設定（デフォルト）

クラスタ数設定の考え方

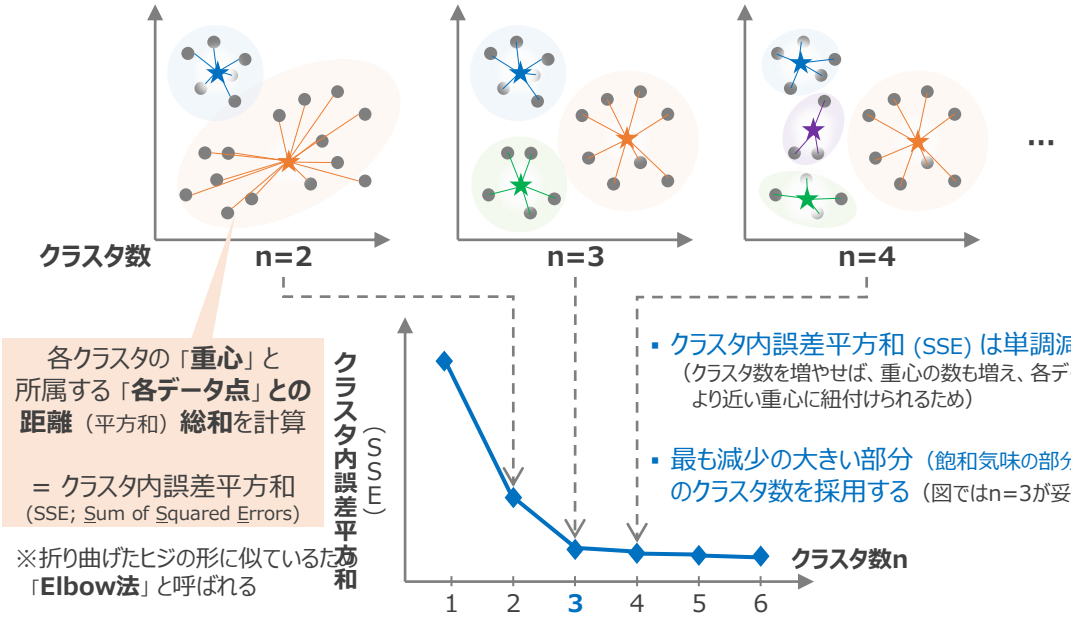
- クラスタ数を客観的に評価する「Elbow法」などの手法もあるが、一般的には、まずは**人間が解釈可能なレベルの3~5個程度**から着手してみることが賢明
- 階層的クラスタリングや、自動的にクラスタ数を決めてくれる手法（DBScanなど）を活用する



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 Studio での実装方法

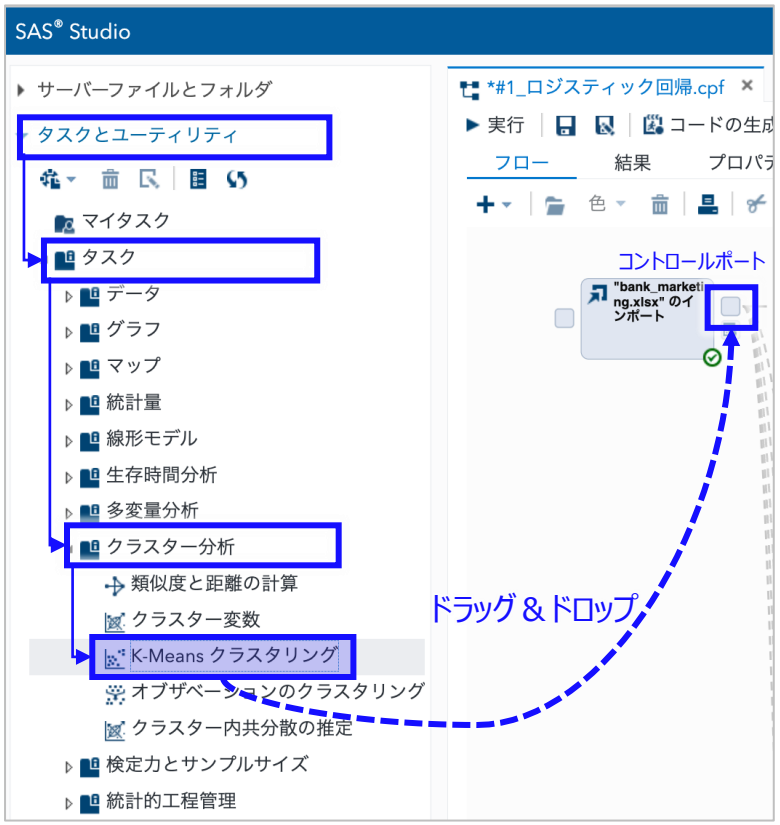
- 非階層的クラスタリング（k-means）
- クラスタ数の設定と変更
- クラスタリング結果の解釈
- クラスタ番号の出力と追加分析
- グループ変数の設定

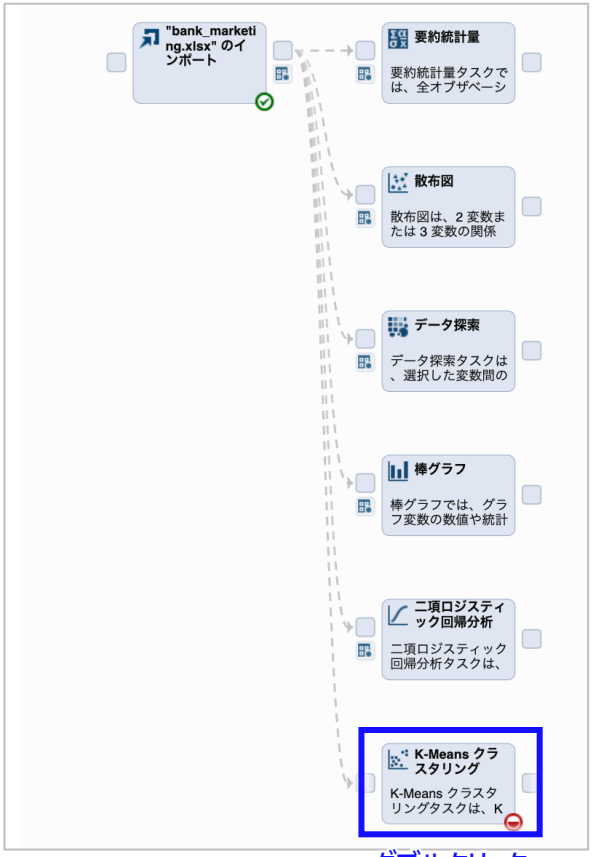
非階層的クラスタリング（k-means）– 実行方法（1/2）ノードの設置

①左パネルより、[タスクとユーティリティ]→[タスク]
→[クラスター分析]→[**K-Meansクラスタリング**]を選択

②右側のプロセスフロー内のインポートノードの
右端の四角 □（コントロールポート）の上へドラッグ＆ドロップ

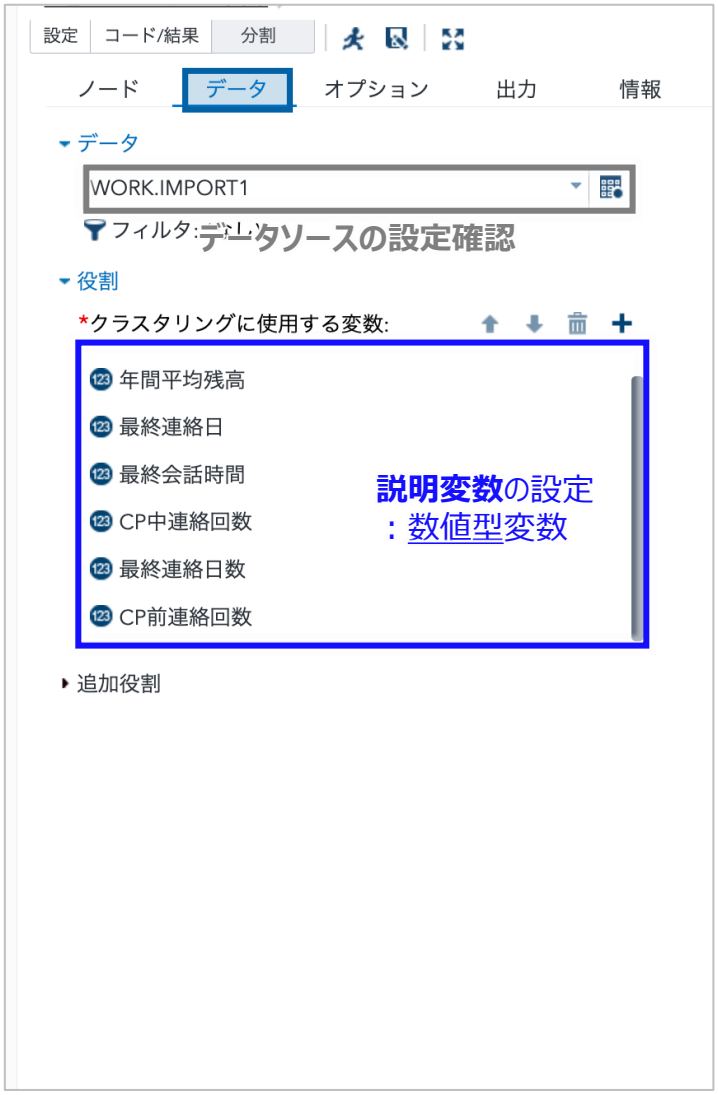
③プロセスフロー上に K-Meansクラスタリング ノードが
生成されるのでダブルクリックして詳細設定画面を開く



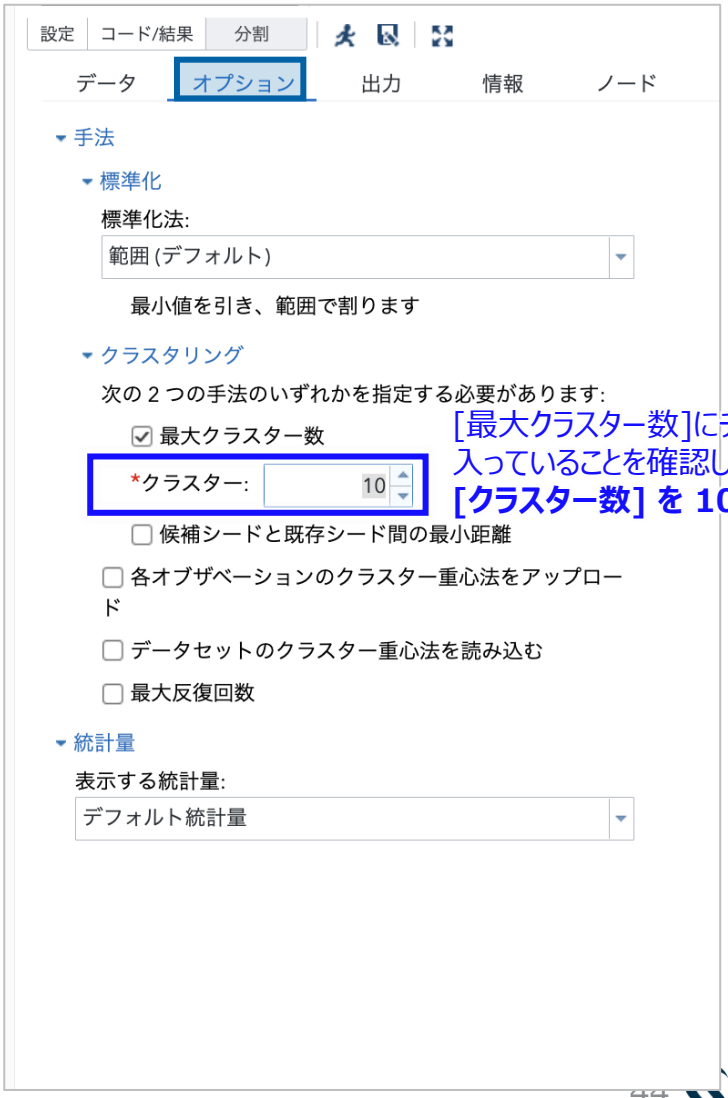


非階層的クラスタリング（k-means） – 実行方法（2/2）説明変数・オプション

[データ]の設定（説明変数・目的変数）



[オプション]の設定（各種出力）



SAS® Studio

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**

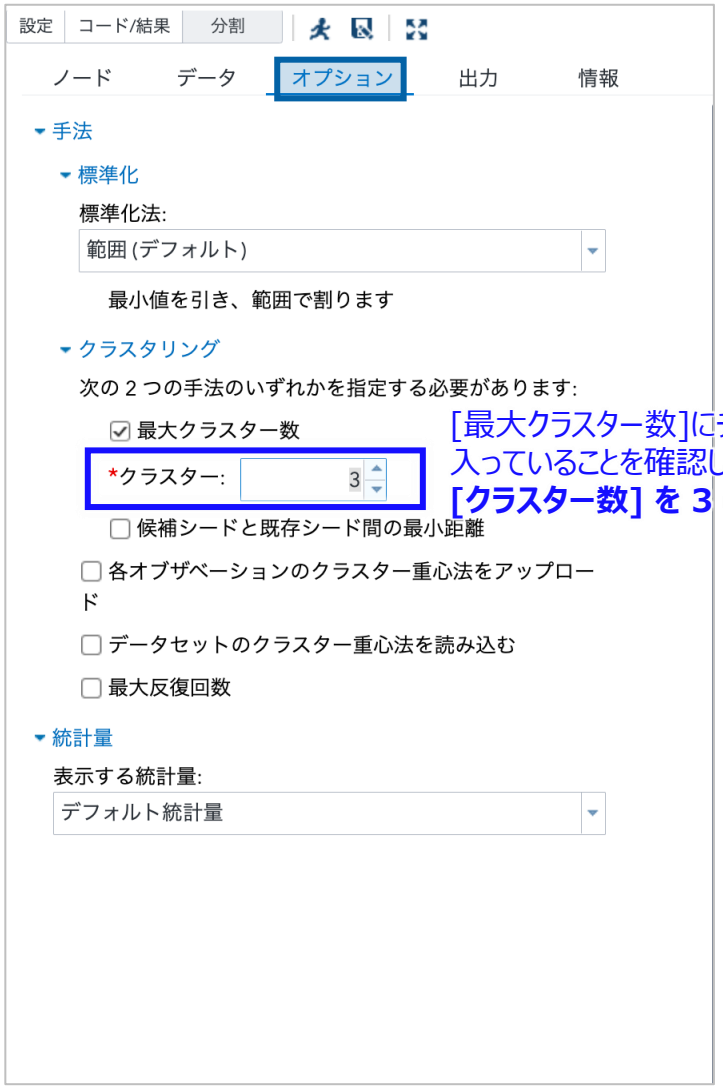
クラスター平均							
クラスター	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
1	0.2930672269	0.0580246286	0.7619047619	0.5061710881	0.0553935860	0.0267447575	0.0142857143
2	0.2784474446	0.0629810558	0.4825136612	0.0907255767	0.0240046838	0.3600212438	0.1006557377
3	0.5049847212	0.0648683352	0.2758441558	0.0754145226	0.0277498012	0.0154950554	0.0107532468
4	0.2996078431	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0101333333
5	0.3860294118	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0000000000
6	0.3120204604	0.0617715665	[illegible]	[illegible]	[illegible]	[illegible]	0.1084057971
7	0.2184991446	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.0069351230
8	0.3158757436	0.0641564482	0.7179026217	0.0765807884	0.0401375831	0.0180579322	0.0103370787
9	0.2979302832	0.0613663152	0.8728395062	0.0446994495	0.4799697657	0.0000000000	0.0000000000
10	0.3231707317	0.0731065649	0.4471544715	0.0885750963	0.0243902439	0.2246587603	0.4682926829

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	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
1	0.1332833775	0.0136267926	0.1462041770	0.1425972931	0.0955384223	0.0688703511	0.0395744560
2	0.1234821079	0.0411433833	0.1166294759	0.0847113930	0.0342251454	0.1070186226	0.0663664562
3	0.1186118589	0.0395683638	0.1347649153	0.0601729245	0.0404789000	0.0497285794	0.0417398457
4	0.1174068148	0.0274275236	0.1282979071	0.1220312667	0.0339841904	0.0585922721	0.0425210042
5	0.1515847656	0.2761522243	0.0816496581	0.0209185053	0.0552655674	0.0000000000	0.0000000000
6	0.1272852521	0.0355211221	0.0696669625	0.0625609989	0.0358977458	0.1163699904	0.0817620401
7	0.0733710808	0.0286117341	0.1212532235	0.0480523917	0.0531900103	0.0377270024	0.0312236852
8	0.1466440802	0.0378032407	0.1618197979	0.0661433663	0.0569974549	0.0579437388	0.0376313479
9	0.1496153680	0.0373506863	0.1630471163	0.0911274848	0.1648636801	0.0000000000	0.0000000000
10	0.1416462697	0.0442681558	0.1277966287	0.0890310869	0.0374142118	0.1097536993	0.2084694515

非階層的クラスタリング（k-means）：クラスタ数変更 – 実行方法



[最大クラスター数]にチェックが入っていることを確認し、
[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	1881	0.0953	0.9847		2	0.4724
2	2105	0.0967	0.9738		3	0.3352
3	535	0.1143	0.9479		2	0.3352

必要に応じて、
2段階クラスタリングも有効

▼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	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
1	0.3107076962	0.0632798060	0.7658692185	0.0818016574	0.0479770856	0.0204026016	0.0108665603
2	0.3486866005	0.0637319316	0.2957086302	0.0895692150	0.0302002036	0.0047004729	0.0034584323
3	0.2907641561	0.0639111035	0.3451713396	0.0871519303	0.0218195689	0.3048379491	0.1315887850

クラスタ3は
年齢が若め

クラスタ2は
最後の会話が長い

CP中連絡回数、最終連絡日数、CP前連絡回数が
クラスタ2＜クラスタ1＜クラスタ3の順

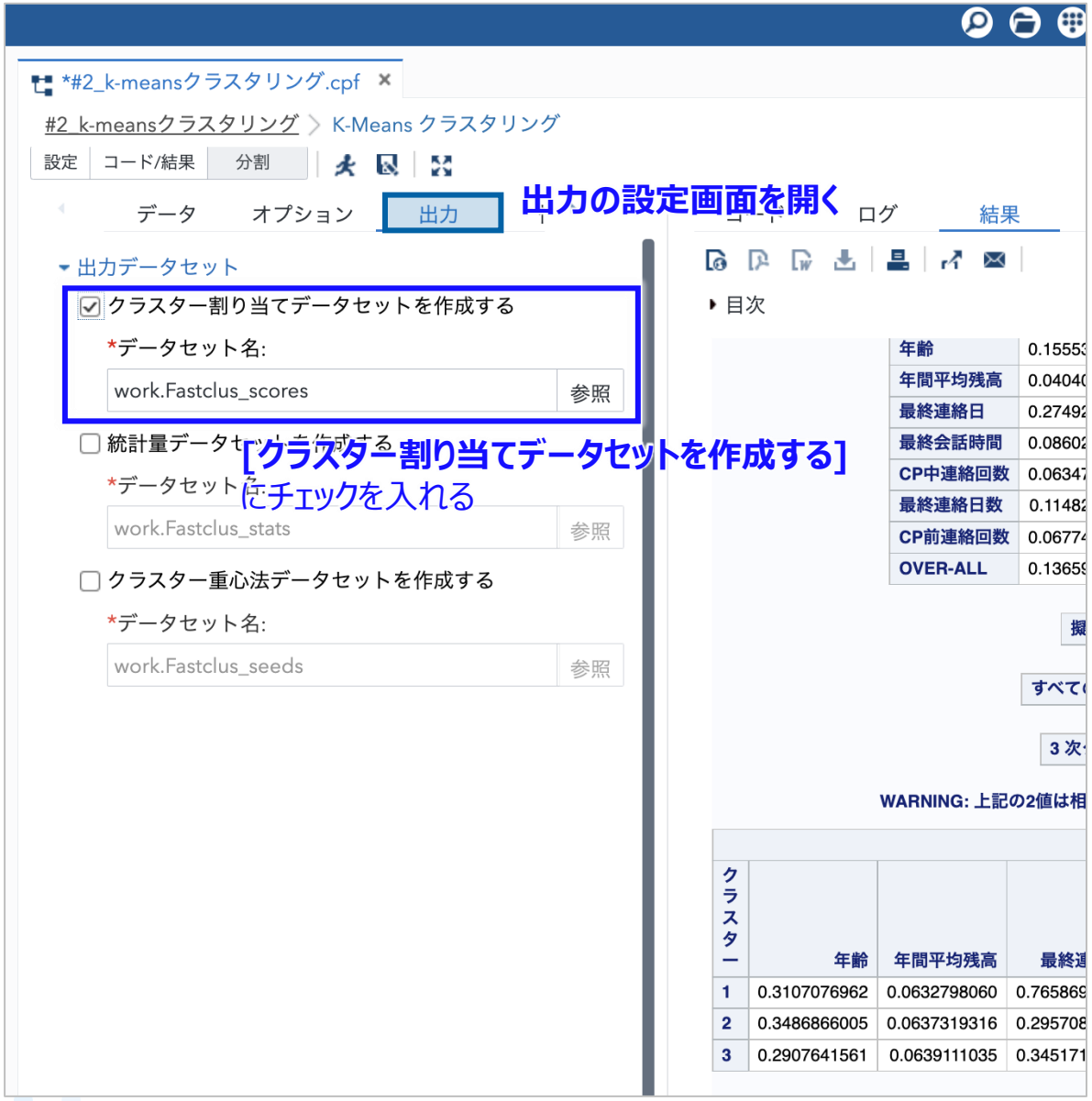
▼各クラスタにおける説明変数の値傾向（標準偏差）

クラスター標準偏差

クラスター	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
1	0.1494764209	0.0363567636	0.1427522516	0.0851396155	0.0791419042	0.0668549819	0.0395526010
2	0.1638910422	0.0440847172	0.1583476932	0.0878893150	0.0506141033	0.0262431917	0.0232405855
3	0.1271532836	0.0387549810	0.1782289260	0.0809842315	0.0333549441	0.1303417255	0.1318504823

SAS® Studio

（参考）クラスター番号のデータ化 と 追加分析（1/3）



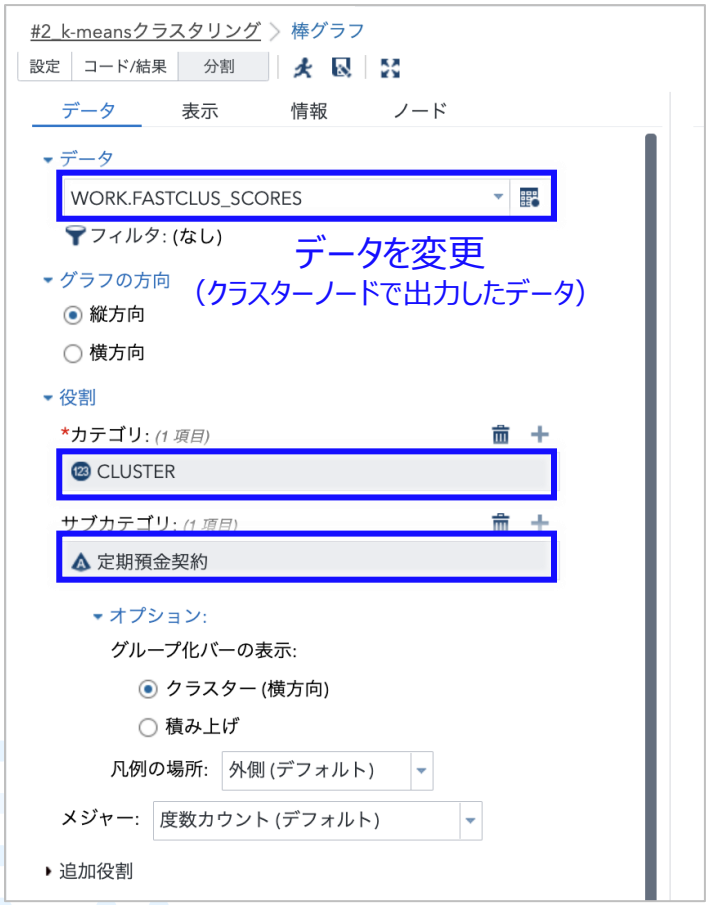
（参考）クラスター番号のデータ化 と 追加分析（2/3）

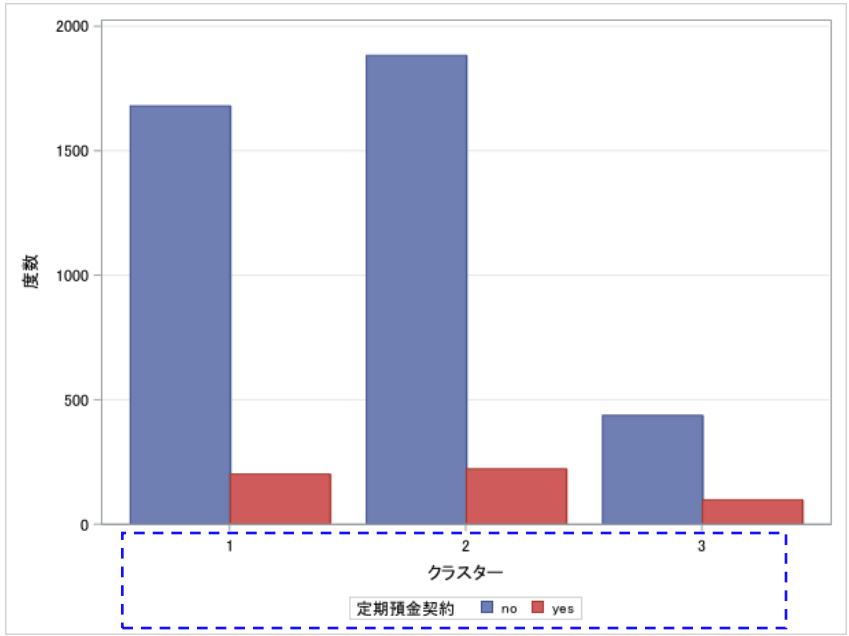
[出力データ]タブをクリックすると
[CLUSTER]カラムが挿入されている

CP前連絡回数	前回C...	定期...	CLUSTER	D
0	unknown	no	1	0.233
0.16	failure	no	3	0.180
0.04	failure	no	3	0.224
0	unknown	no	2	0.339
0	unknown	no	2	0.29
0.12	failure	no	1	0.23
0.08	other	no	3	0.172
0	unknown	no	2	0.186
0	unknown	no	2	0.137
0.08	failure	no	3	0.22
0	unknown	no	1	0.190
0	unknown	no	2	0.212
0	unknown	no	2	0.141
0	unknown	yes	1	0.34
0.04	failure	no	1	0.340
0	unknown	no	1	0.166
0	unknown	no	1	0.271
0.08	failure	no	1	0.240
0	unknown	no	1	0.218
0.04	other	no	3	0.201
0	unknown	no	1	0.218
0	unknown	no	1	0.185
0	unknown	no	2	0.07
0	unknown	no	2	0.141
0	unknown	no	1	0.272
0	unknown	no	1	0.165
0	unknown	no	2	0.253
0.08	failure	no	2	0.420
0	unknown	no	1	0.320

（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

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**

グループ分析の変数に
[定期預金契約有無]を設定

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	178	0.1252	0.9341		3	0.3519
2	148	0.1381	0.9680		3	0.3617
3	195	0.1479	0.8846		1	0.3519

クラスター平均

クラスター	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
1	0.2523959022	0.0835507389	0.7245318352	0.2529710261	0.0815828041	0.0207341754	0.0216693419
2	0.2967011129	0.0988696446	0.1873873874	0.1611507455	0.0420094007	0.1081081081	0.1332046332
3	0.4674208145	0.1121606234	0.5018803419	0.1566901639	0.0408026756	0.1301544832	0.0871794872

クラスタ3は 年齢層が高め

クラスタ1は CP中連絡回数が特に多い

契約者は全般的に最終会話時間が長め

未契約者

クラスターの要約

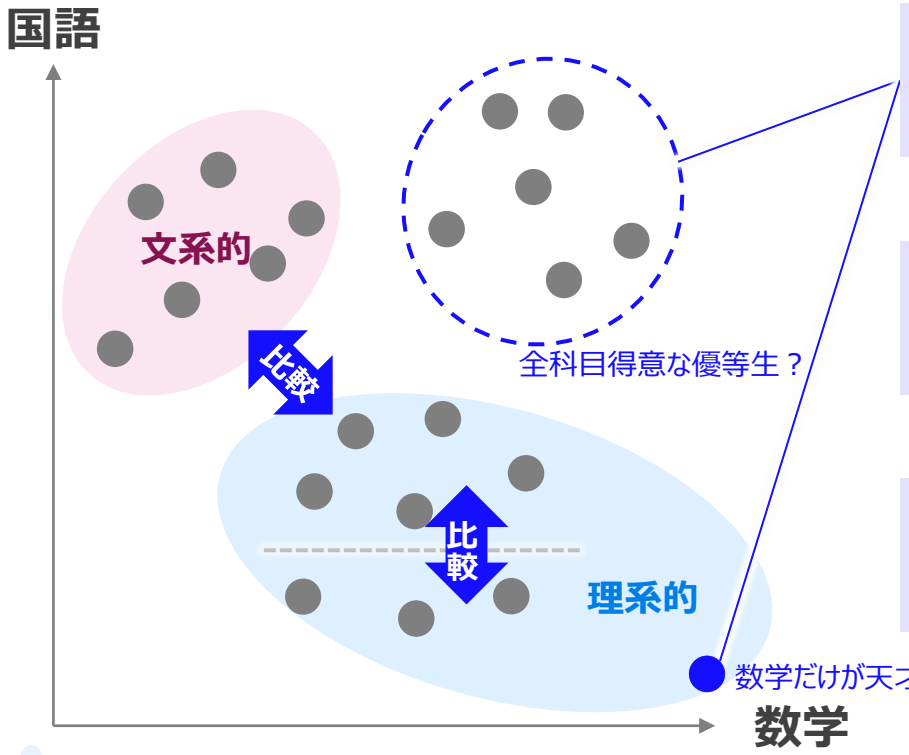
クラスター	度数	RMS 標準偏差	シードから オブザベーション までの最大距離	半径 超える	最も近い クラスター	クラスター 重心間の距離
1	1691	0.0921	0.9973		3	0.4710
2	436	0.1139	0.9584		3	0.3380
3	1873	0.0932	0.9748		2	0.3380

クラスター平均

クラスター	年齢	年間平均残高	最終連絡日	最終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数
1	0.3148185742	0.0632917973	0.7639858072	0.0691620318	0.0497592295	0.0169732909	0.0084210526
2	0.2888196632	0.0639327358	0.3411314985	0.0737437327	0.0234506647	0.3079075835	0.1282568807
3	0.3497222908	0.0631686716	0.2949991102	0.0775746143	0.0307811325	0.0036320087	0.0027976508

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

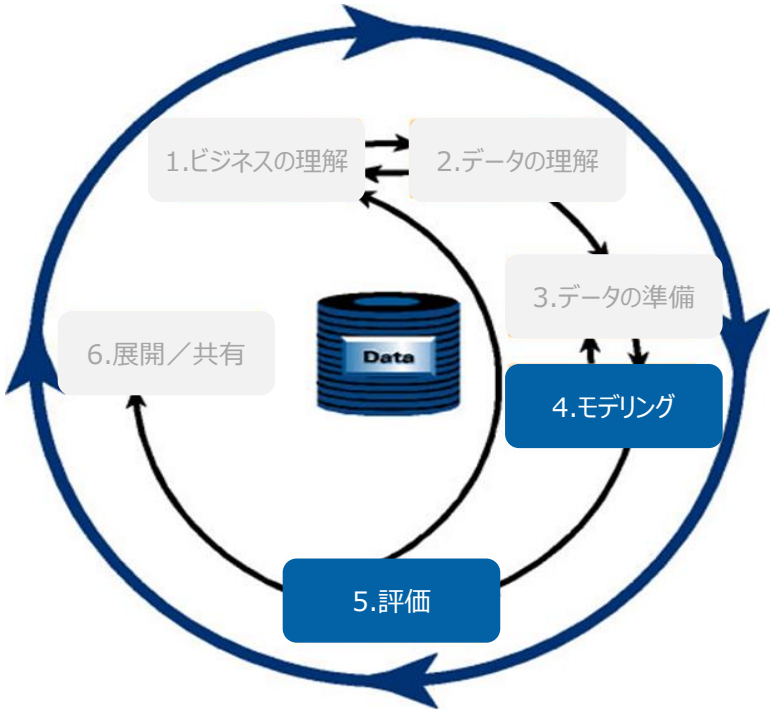
③各**クラスタの特徴**を比較する
→ クラスタ「**間**」での分析（目的変数をクラスタ番号とする）

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、別の視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）の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