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

#3 クラスター分析による分類（2）

クラスター分析の応用 と 階層的クラスタリング

2022年10月26日

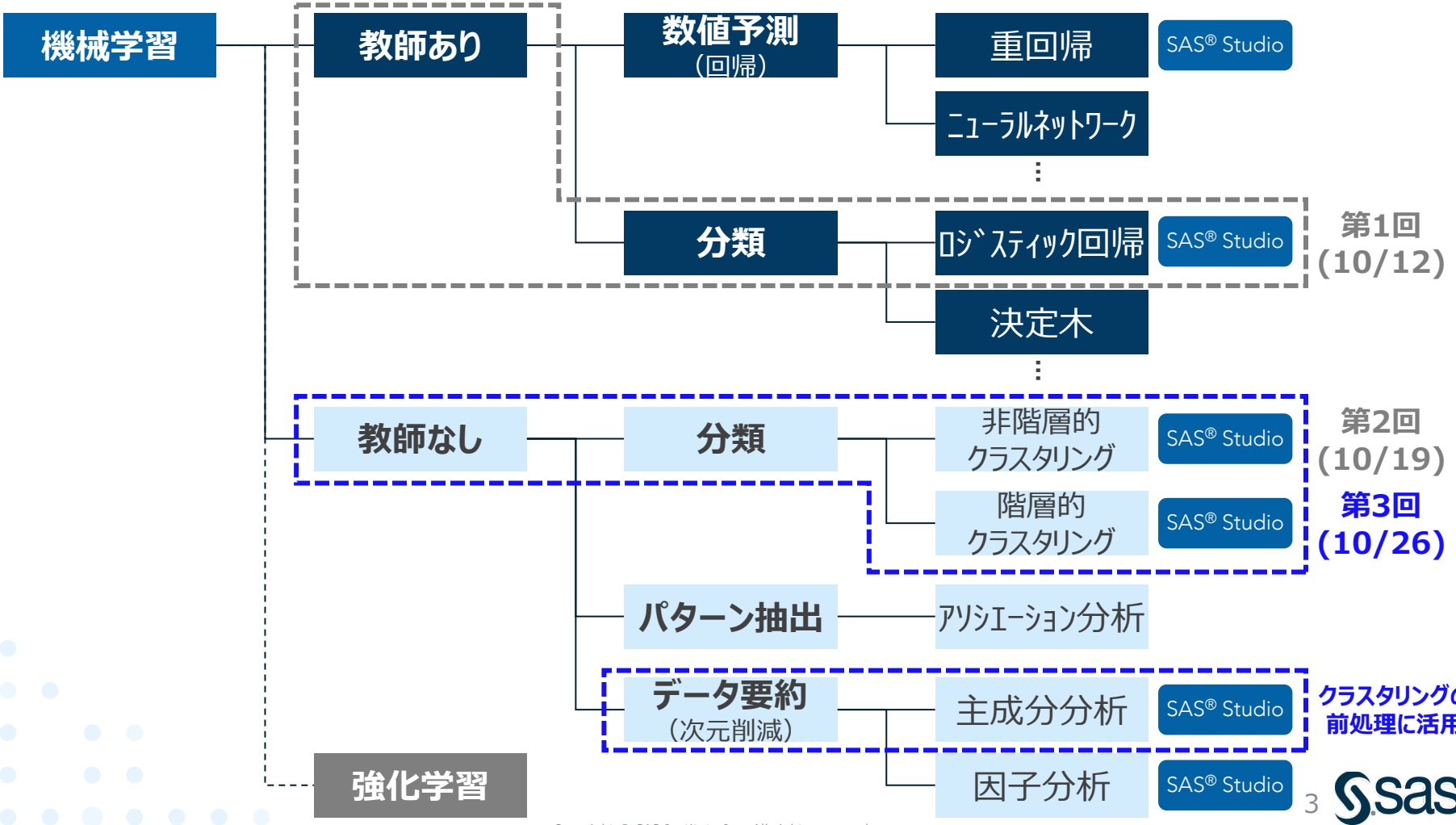
Agenda

- **クラスター分析の応用（他の分析手法との組み合わせ）**
  - 主成分分析により説明変数を要約する
  - 主成分軸でクラスター分析を行う
- **クラスター分析による分類（2）：階層的クラスタリング**
  - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
  - 樹形図（デンドログラム）とクラスタ数の検討
  - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する
- **今後のデータサイエンス学習に向けたスキルアップ**
  - データサイエンティストに求められるスキル
  - SAS内サンプルデータの紹介と使い方
  - オープンデータの紹介

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



Agenda

- **クラスター分析の応用（他の分析手法との組み合わせ）**
  - 主成分分析により説明変数を要約する
  - 主成分軸でクラスター分析を行う

- **クラスター分析による分類（2）：階層的クラスタリング**
  - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
  - 樹形図（デンドログラム）とクラスタ数の検討
  - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する

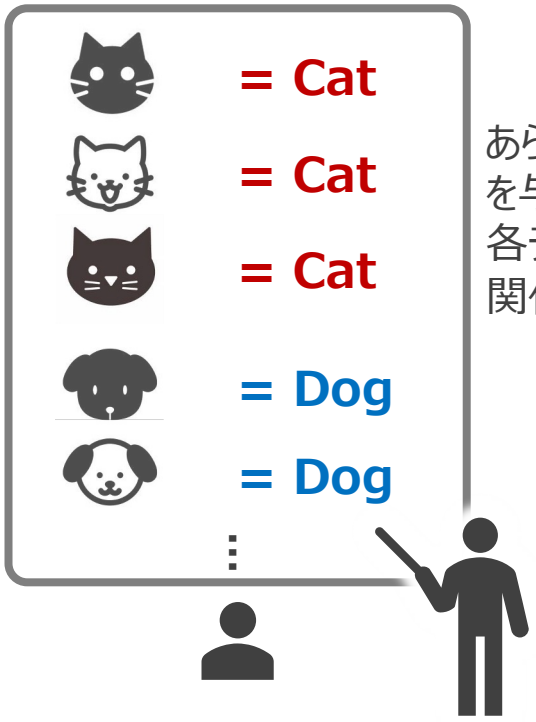
- **今後のデータサイエンス学習に向けたスキルアップ**
  - データサイエンティストに求められるスキル
  - SAS内サンプルデータの紹介と使い方
  - オープンデータの紹介

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

再掲

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

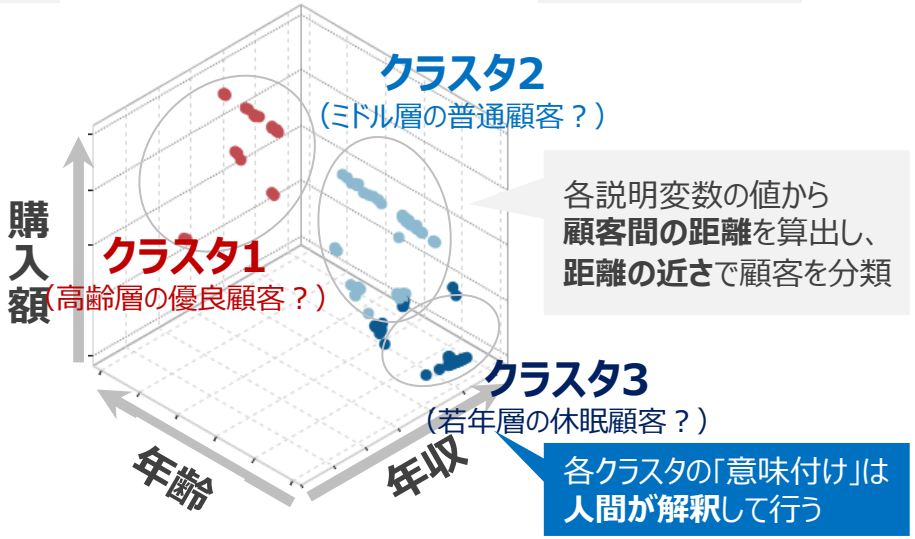
データ例

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	…
⋮	⋮	⋮	⋮	⋮	⋮	…

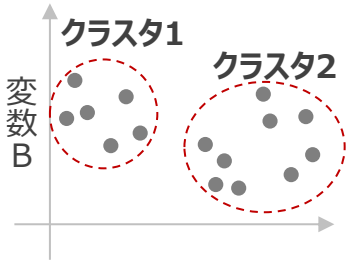
説明変数

※目的変数は無し

クラスタリング



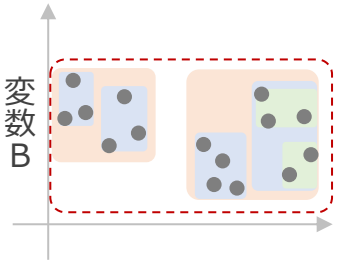
非階層的クラスタリング

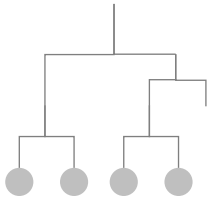


主な手法

- **k-means法**（k平均法）
- 混合ガウス

階層的クラスタリング





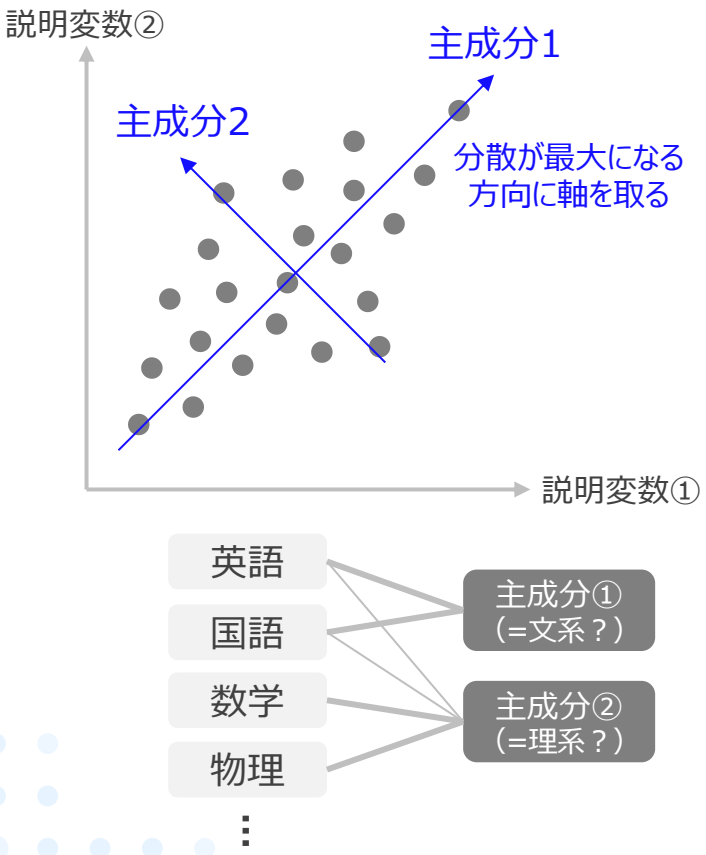
主な手法

- 最短距離法
- 最長距離法
- 群平均法
- ウォード法

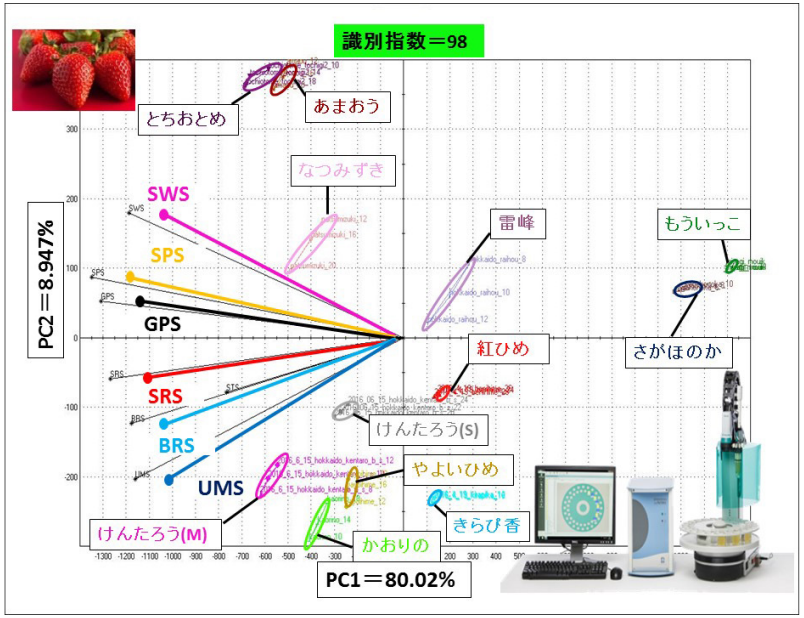
主成分分析の概要

- 主成分分析は、多数の説明変数が存在する場合に、（それらの分散構造を考慮して）**変数を合成**していくことで、より少ない変数**（＝主成分）**でデータを説明しようとするアプローチ
- アンケート調査や官能評価でよく用いられるほか、**分析前の次元削減**としても活用される

▼主成分分析のイメージ



▼官能評価における活用例



Source: https://www.nodai.ac.jp/research/teacher-column/22913/

SAS® Studio

SAS Studio での実装方法

- 主成分分析
- 主成分に対するk-meansクラスタリング

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

※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

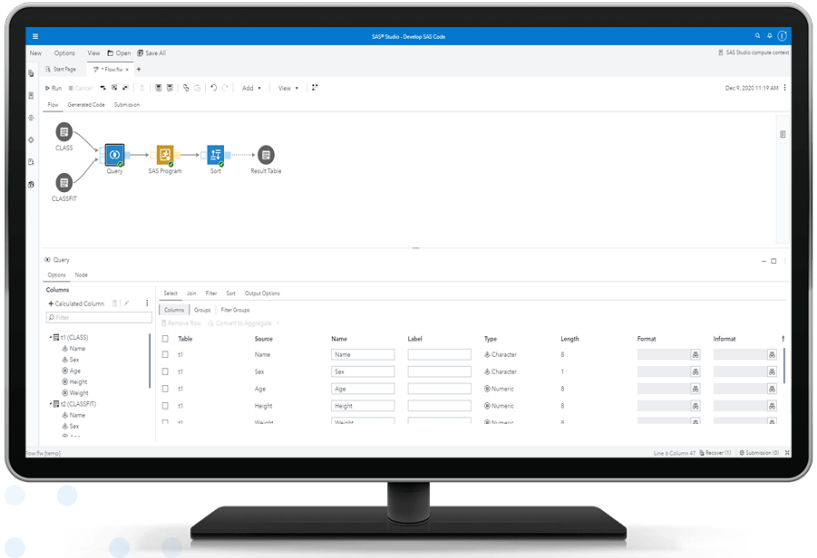
予測(分析)したい対象

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。

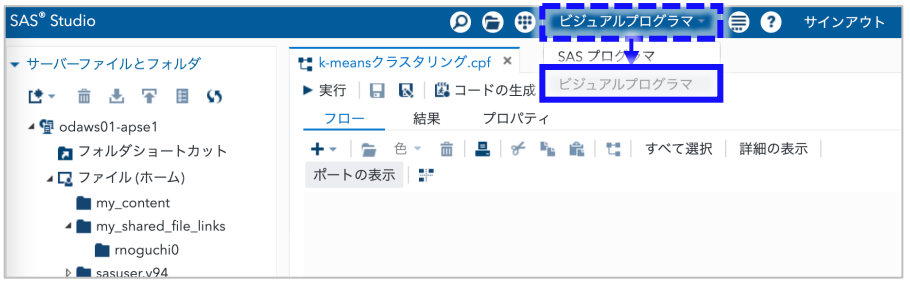
▼SAS Studio 画面イメージ



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）

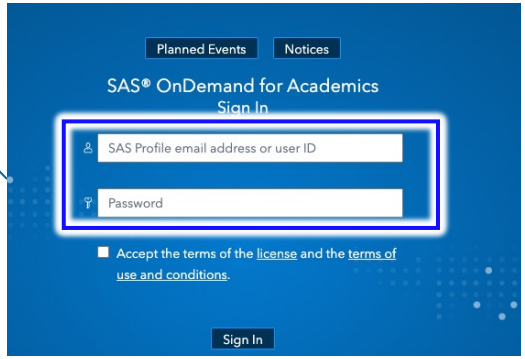


再掲

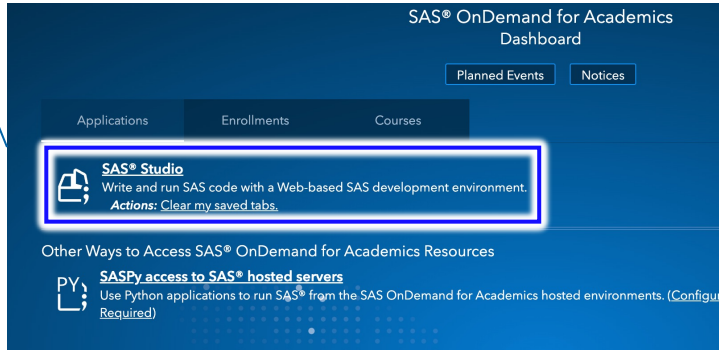
参考：SAS Studio 起動方法

- 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

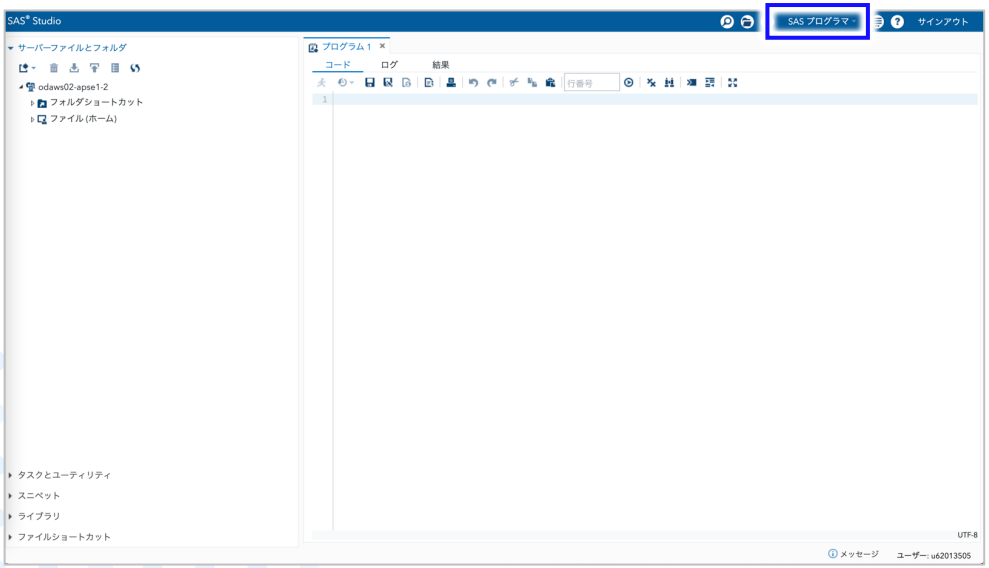
1 **SAS OnDemand for Academics** にログイン
https://welcome.oda.sas.com/login



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



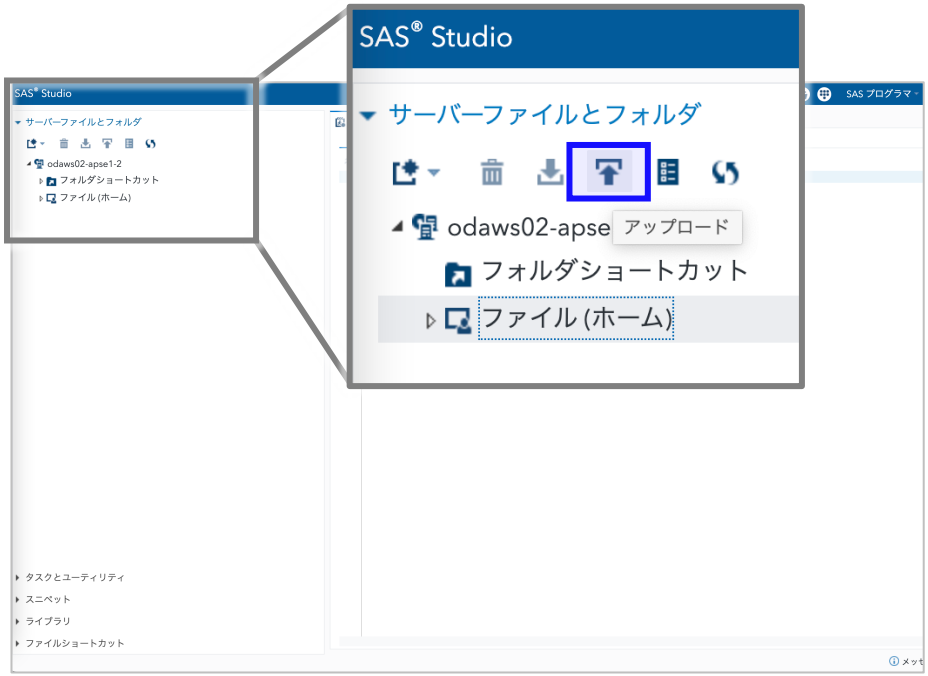
3 SAS Studioが起動／右上よりビジュアルプログラマを選択



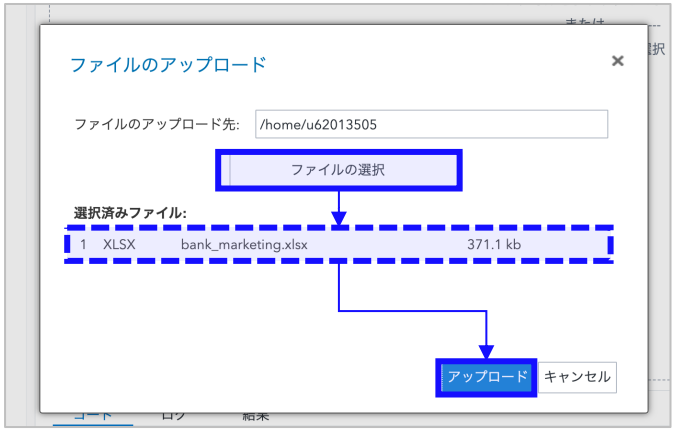
再掲 SAS® Studio

データの読み込み（1/2）

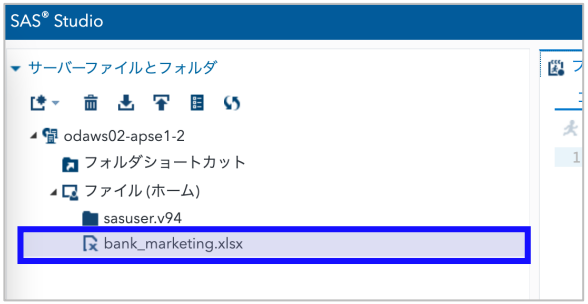
①左パネル内の「アップロード」アイコン をクリック



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

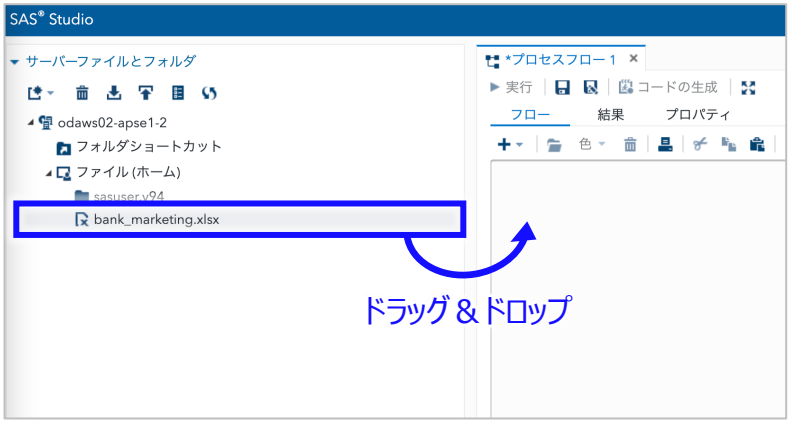


④左パネル内にファイルがアップロードされていることを確認

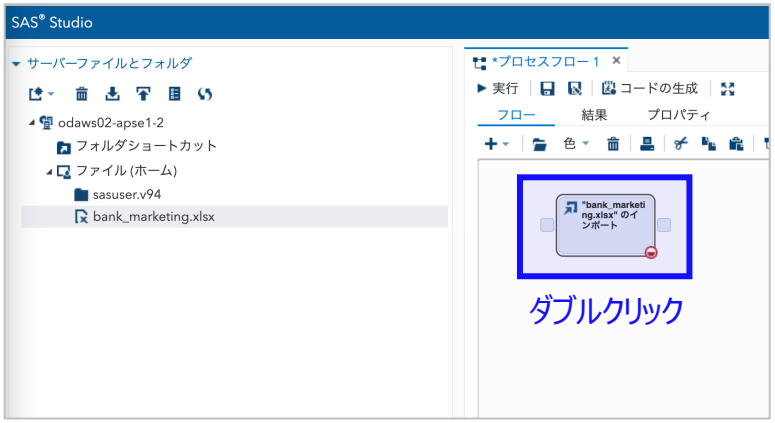


データの読み込み（2/2）

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

再掲 SAS® Studio

読み込んだデータの確認

データ概要の確認

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

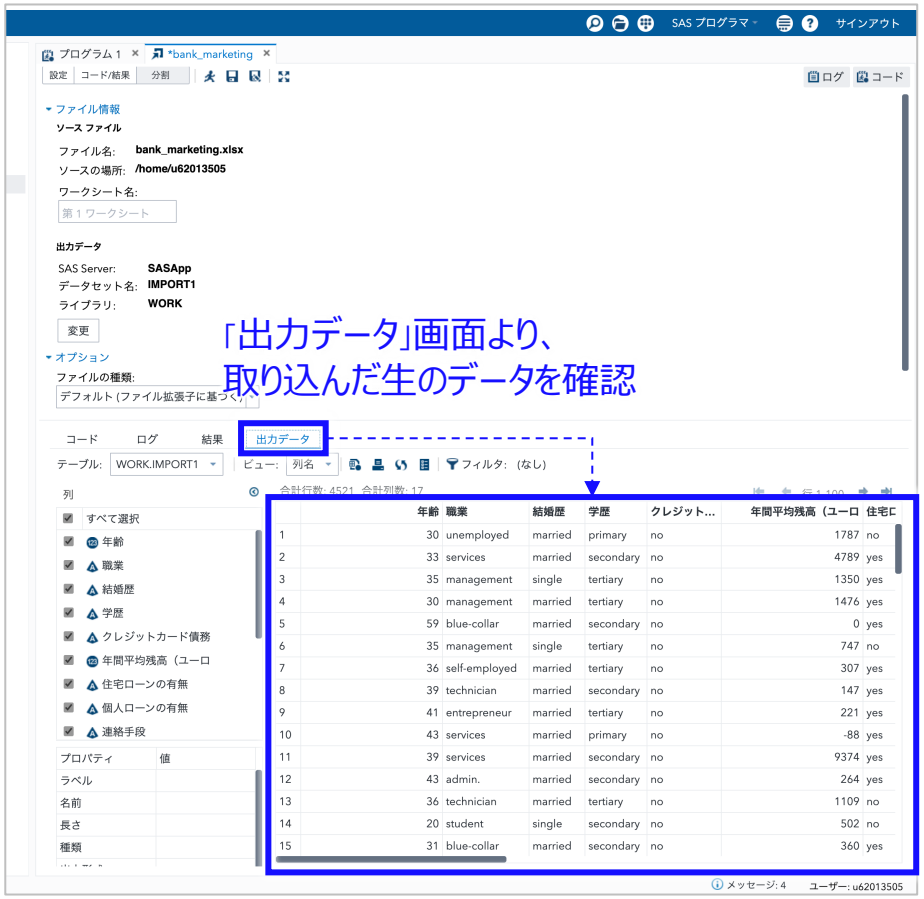
データセットのページサイズ	131072
データセットのページ数	5
データページの先頭	1
ページごとの最大OBS数	1090
先頭ページのOBS数	1052
データセットの修復数	0
ファイル名	/saswork/SAS_work71F80001F3FA_odaws01-apse1-2.oda.sas.com/SAS_workC7860001F3FA_odaws01-apse1-2.oda.sas.com/import1.sas7bdat
作成したリリース	9.0401M6
作成したホスト	Linux
I ノード番号	33850
アクセス権限	rw-r--r--
所有者名	u62013505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

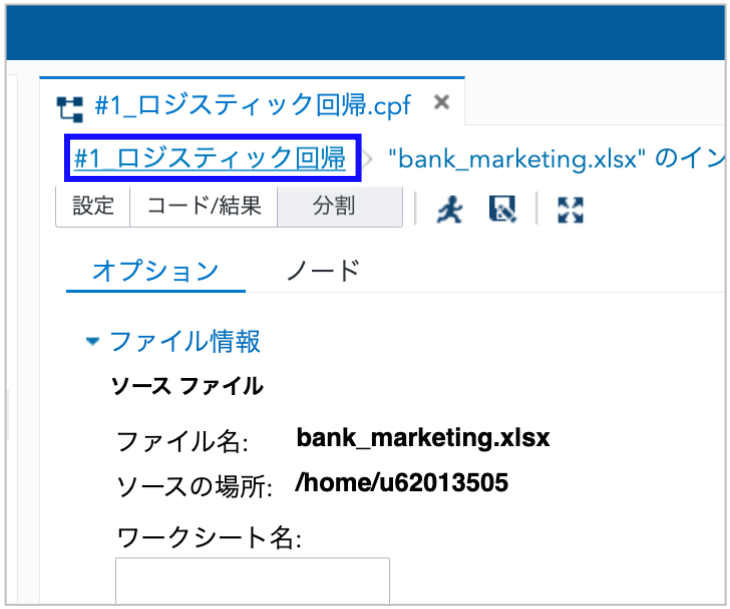


「出力データ」画面より、
取り込んだ生のデータを確認

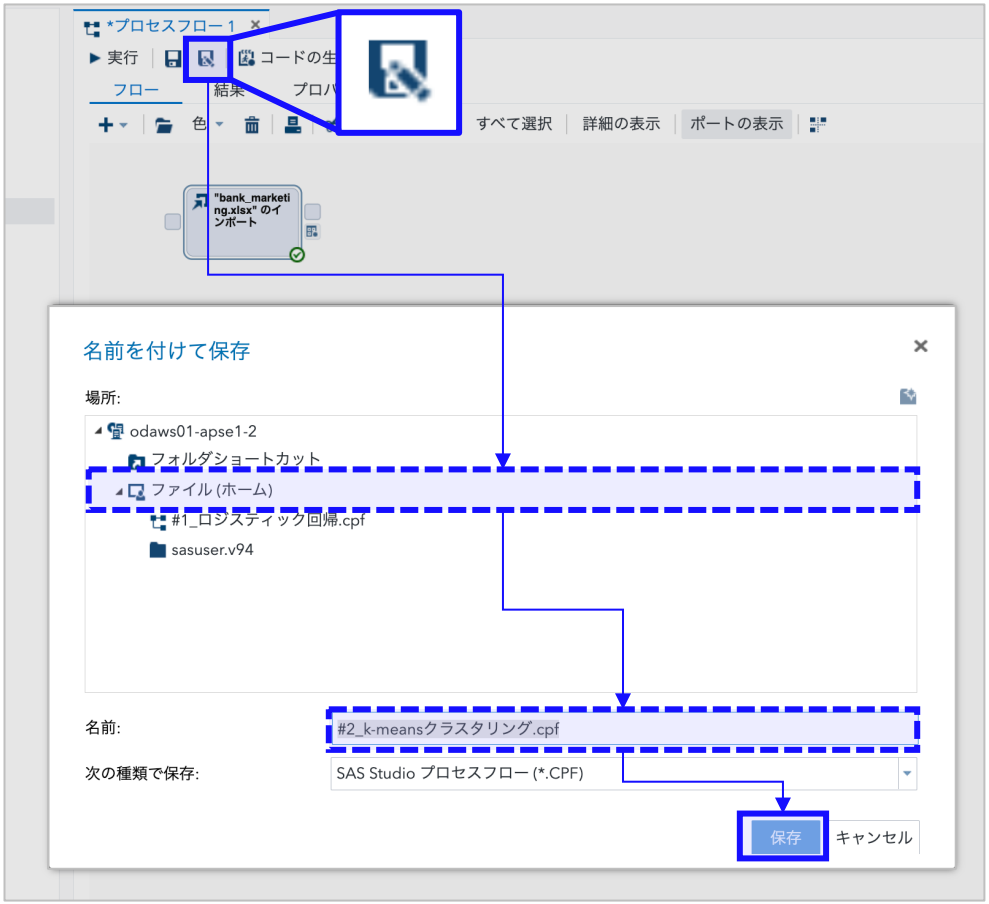
	年齢	職業	結婚歴	学歴	クレジット…	年間平均残高（ユーロ	住宅ロ
1	30	unemployed	married	primary	no	1787	no
2	33	services	married	secondary	no	4789	yes
3	35	management	single	tertiary	no	1350	yes
4	30	management	married	tertiary	no	1476	yes
5	59	blue-collar	married	secondary	no	0	yes
6	35	management	single	tertiary	no	747	no
7	36	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

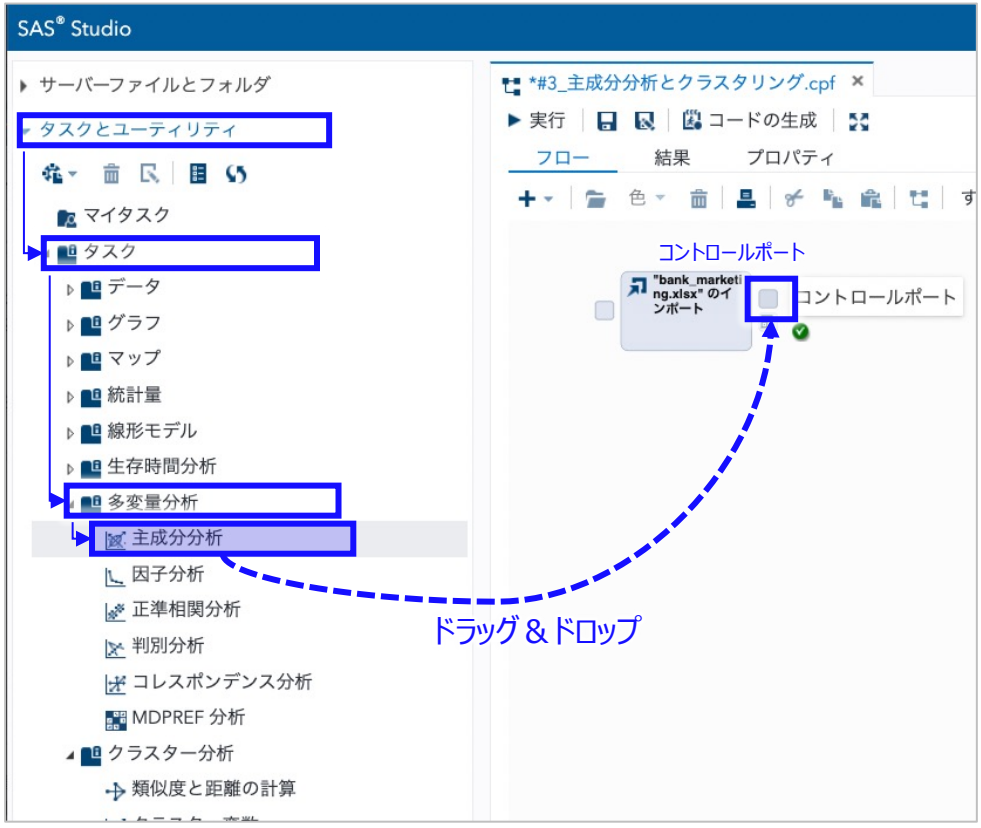


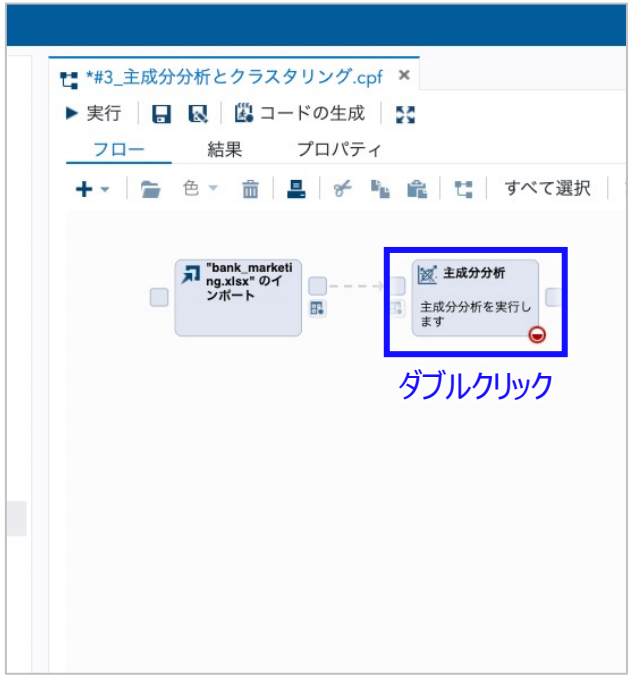
主成分分析 – 実行方法（1/2）ノードの設置

①左パネルより、[タスクとユーティリティ]→[タスク]
→[多変量解析]→[**主成分分析**]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 主成分分析ノードが
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

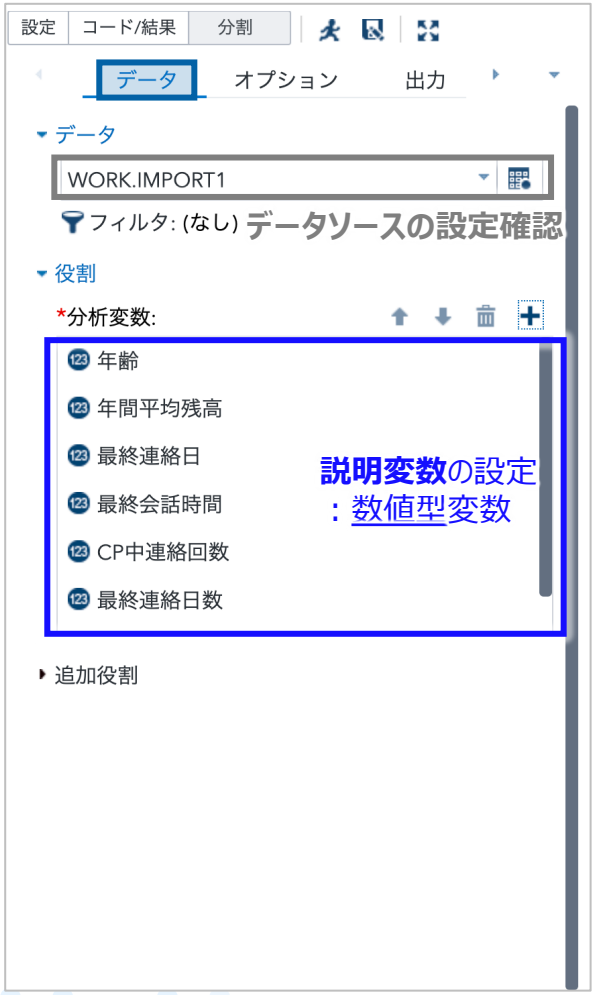




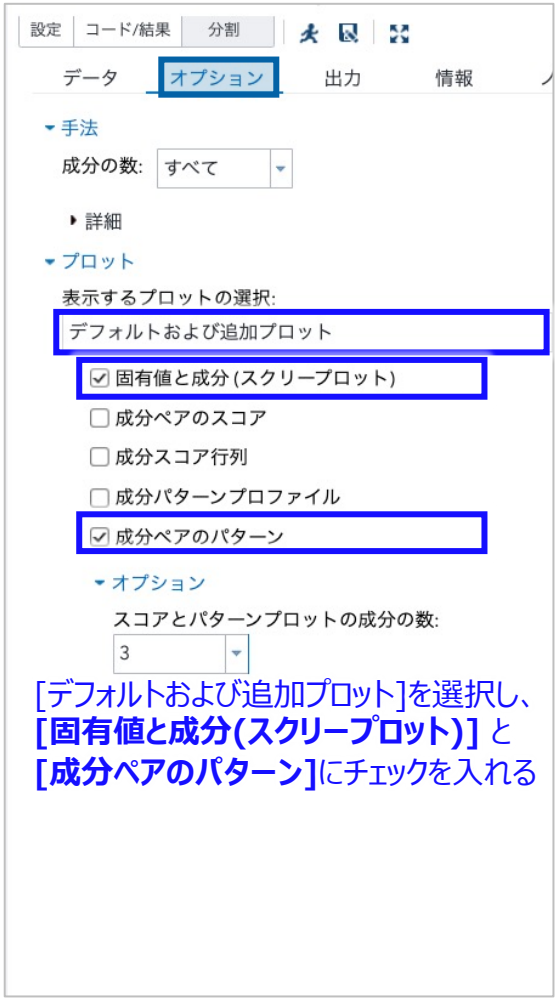
SAS® Studio

主成分分析 – 実行方法（2/2）説明変数・オプション・出力

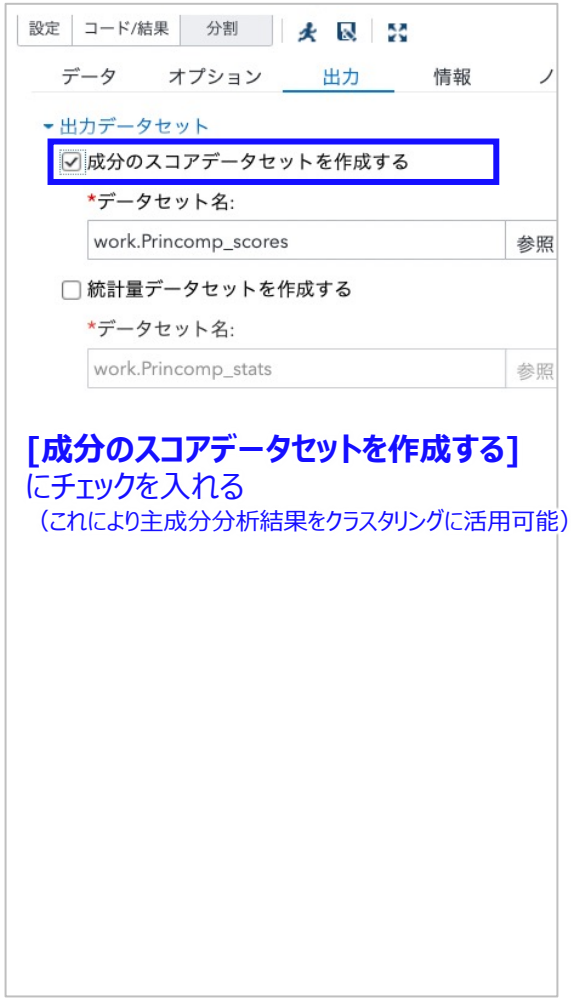
[データ]の設定（説明変数）



[オプション]の設定（各種出力）

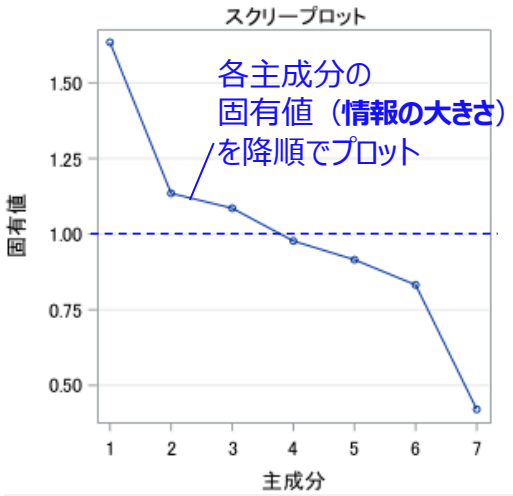


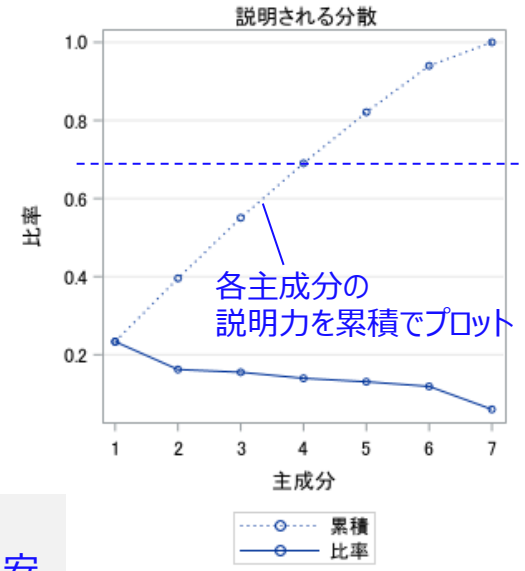
[出力]の設定（分析結果の二次利用）



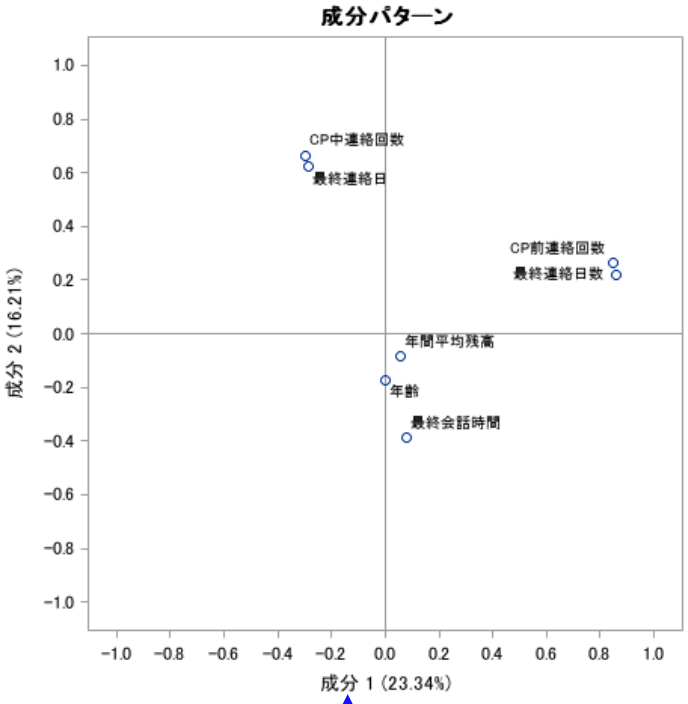
主成分分析 – 実行結果（主成分分析の出力）

- 主成分分析では、まずスクリープロットと累積寄与率のグラフから、最適な主成分数を検討する
  → 今回の分析では、固有値の値と簡単のため、**主成分数＝2とする**
- 各主成分に対する変数寄与度から、各主成分軸の意味を検討する





- 固有値１以上
- 累積寄与率70-80%以上 が目安



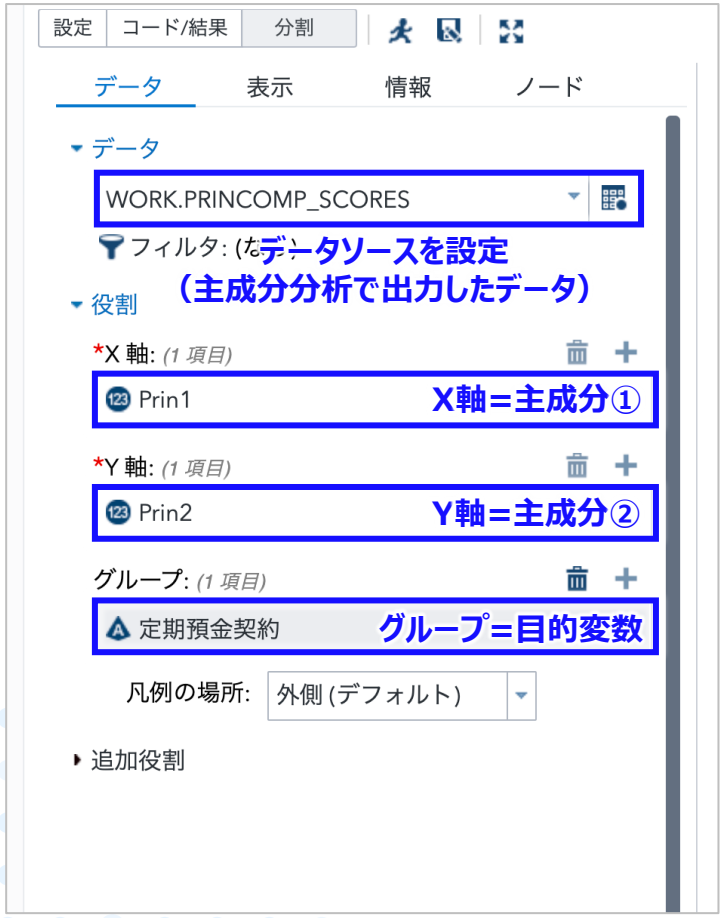
固有ベクトル

		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
年齢	年齢	0.000508	-.165040	0.673358	0.178661	0.693790	-.077848	0.005079
年間平均残高	年間平均残高	0.042574	-.080778	0.694411	0.078178	-.706073	0.066114	0.020836
最終連絡日	最終連絡日	-.223982	0.584159	0.028128	0.409640	-.067857	-.658714	0.038555
最終会話時間	最終会話時間	0.062709	-.361582	-.230598	0.876043	-.065448	0.200538	0.014183
CP中連絡回数	CP中連絡回数	-.233727	0.624431	0.099304	0.143047	0.095881	0.717913	0.023258
最終連絡日数	最終連絡日数	0.672063	0.205337	-.002164	0.025903	0.039382	0.007083	0.709853
CP前連絡回数	CP前連絡回数	0.661657	0.249352	0.023455	0.074696	0.021999	0.000220	-.702439

主成分分析 – 実行結果（主成分分析結果の可視化）

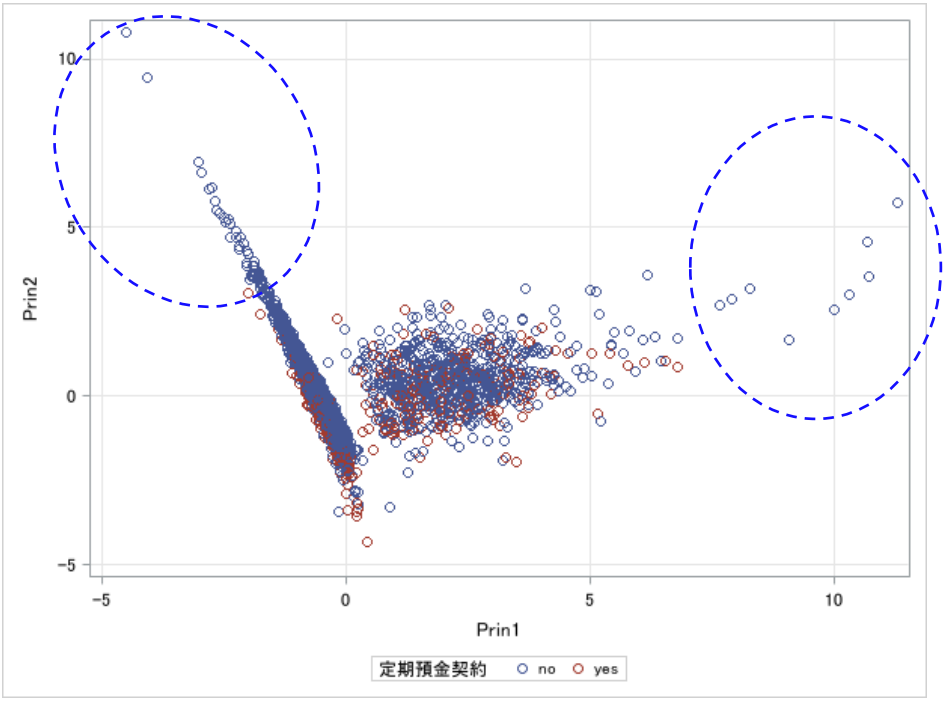
- 「**散布図**」ノードを活用して、各主成分軸をX軸、Y軸にとり、目的変数で色分け表示することで、主成分軸における各データポイントの位置付けと、目的変数との関係性が観察できる

▼散布図の設定



▼散布図の出力結果

主成分①、②が大きい範囲では、契約者はほとんどいない

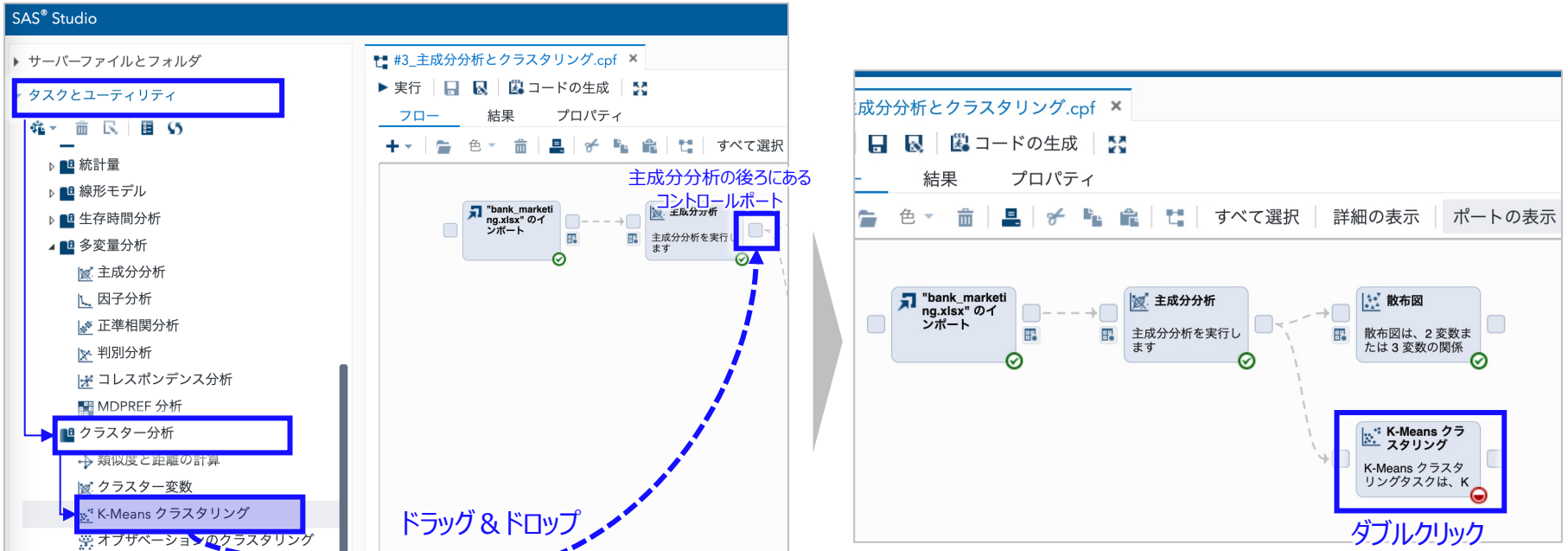


主成分分析結果のクラスタリング – 実行方法（1/2）ノードの設置

①前回と同様に、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**主成分分析の後の**
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

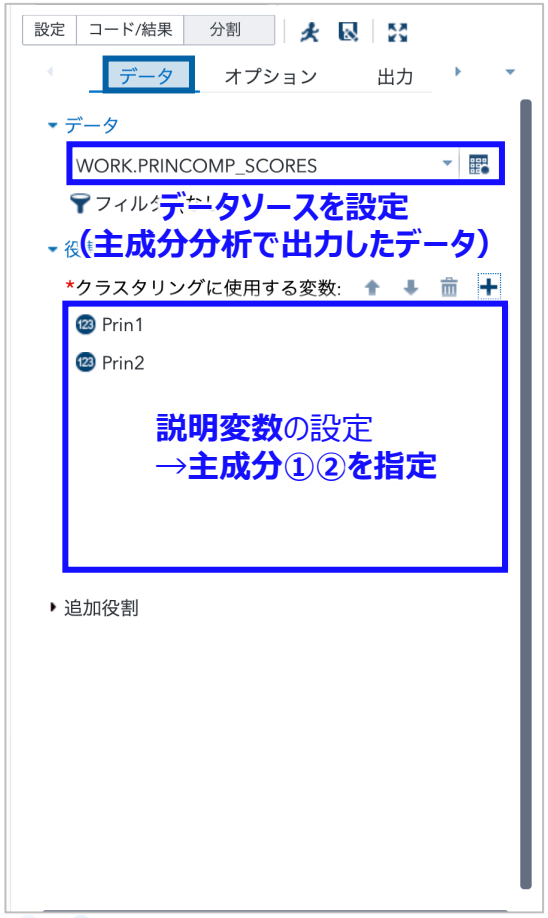
③プロセスフロー上に K-Means クラスタリングノードが
生成されるのでダブルクリックして詳細設定画面を開く



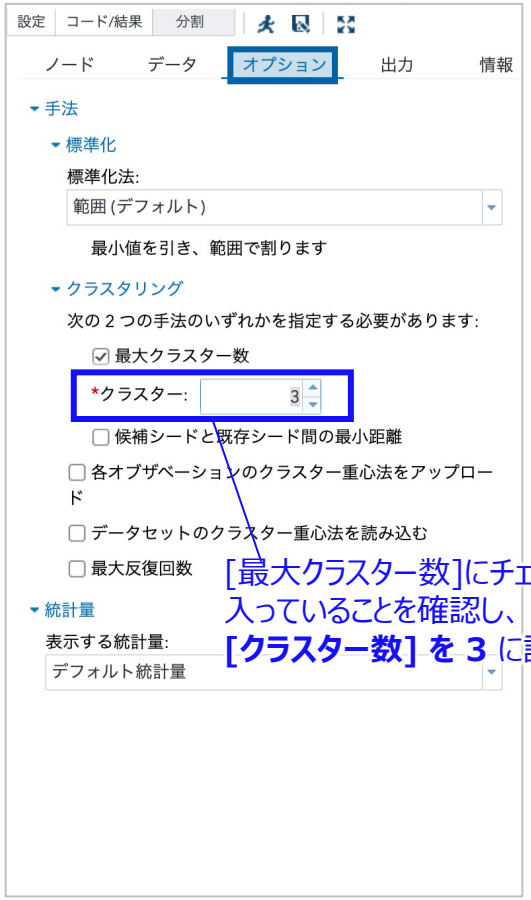
SAS® Studio

主成分分析結果のクラスタリング – 実行方法（2/2）説明変数・オプション

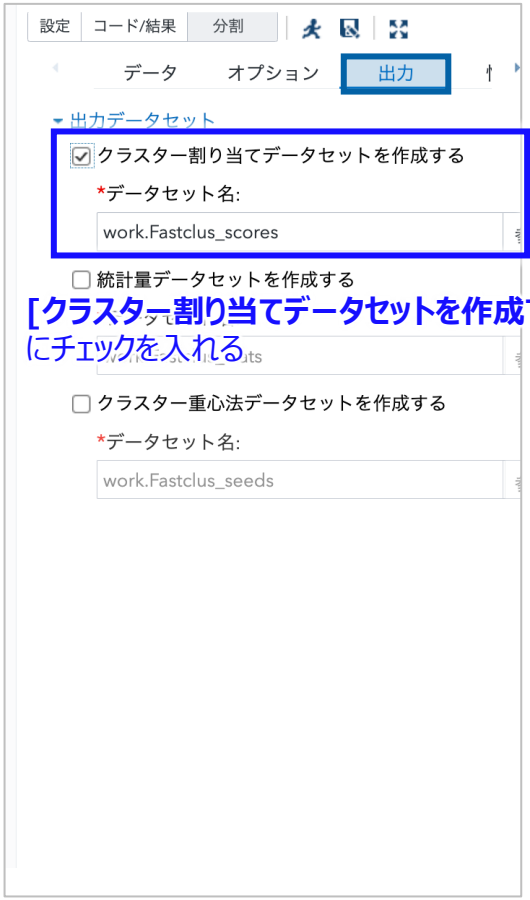
[データ]の設定（説明変数）



[オプション]の設定（各種出力）



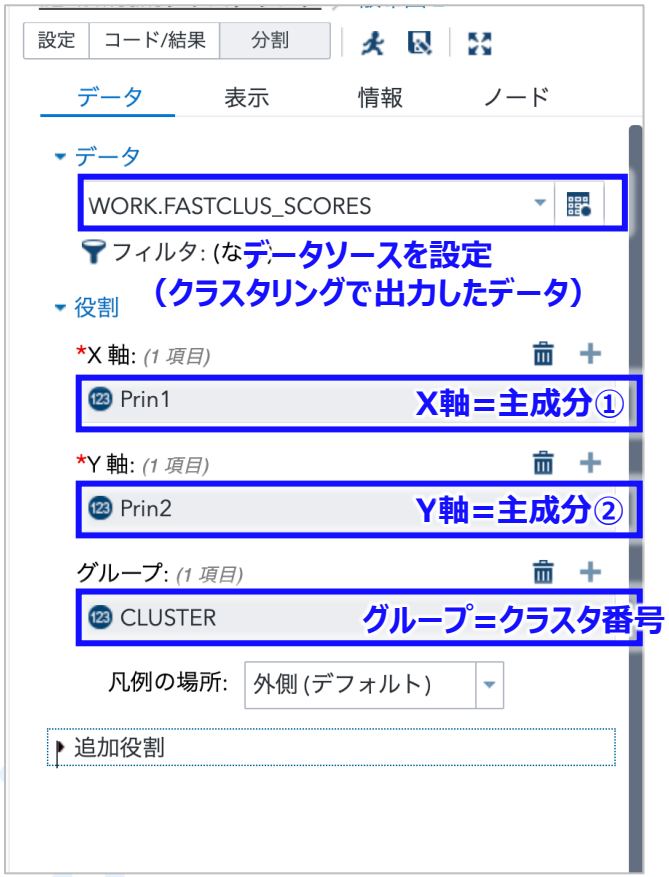
[出力]の設定（分析結果の二次利用）



主成分分析結果のクラスタリング – 実行結果（クラスタリングの可視化）

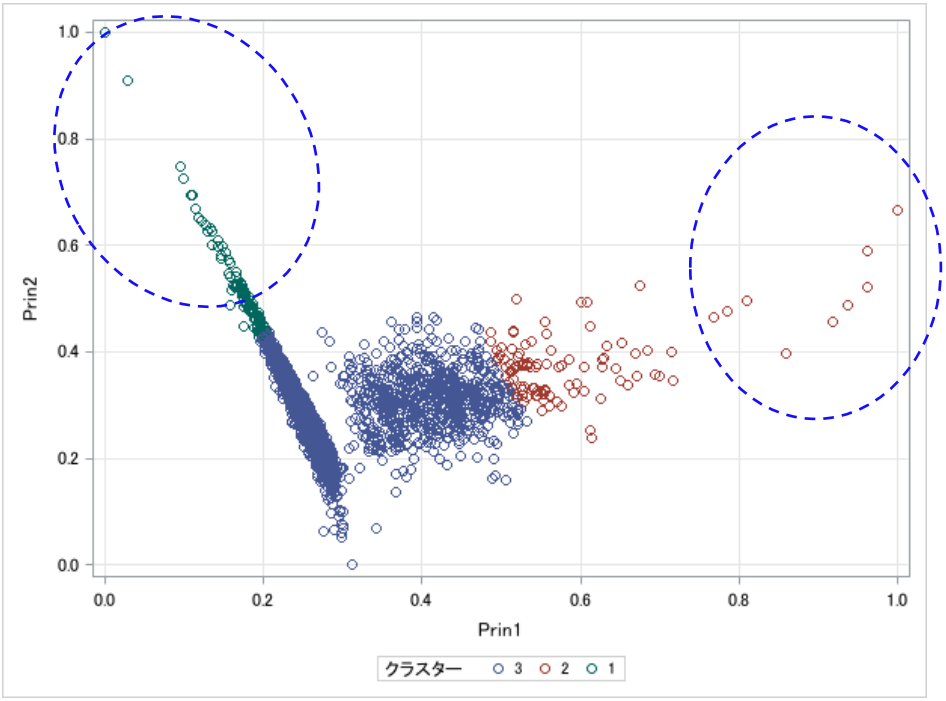
- 「**散布図**」ノードを活用して、各主成分軸をX軸、Y軸にとり、**クラスタ番号で色分け**表示することで、主成分軸における各データポイントの位置付けと、各クラスタとの関係性が観察できる

▼散布図の設定



▼散布図の出力結果

主成分①、②が大きい範囲では、契約者はほとんどいなかったが、クラスタリングではこの傾向をある程度捉えた分類が行われている



Agenda

- クラスター分析の応用（他の分析手法との組み合わせ）
  - 主成分分析により説明変数を要約する
  - 主成分軸でクラスター分析を行う

- **クラスター分析による分類（2）：階層的クラスタリング**
  - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
  - 樹形図（デンドログラム）とクラスタ数の検討
  - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する

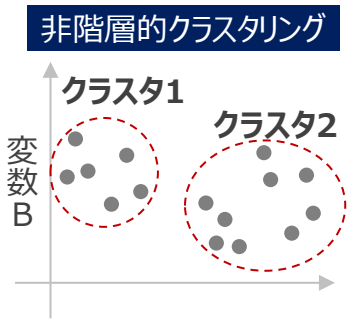
- 今後のデータサイエンス学習に向けたスキルアップ
  - データサイエンティストに求められるスキル
  - SAS内サンプルデータの紹介と使い方
  - オープンデータの紹介

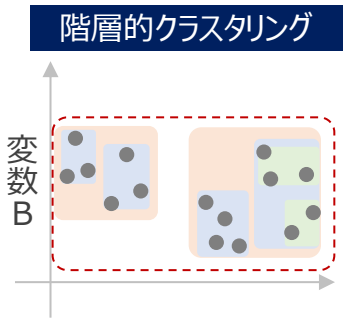
再掲

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

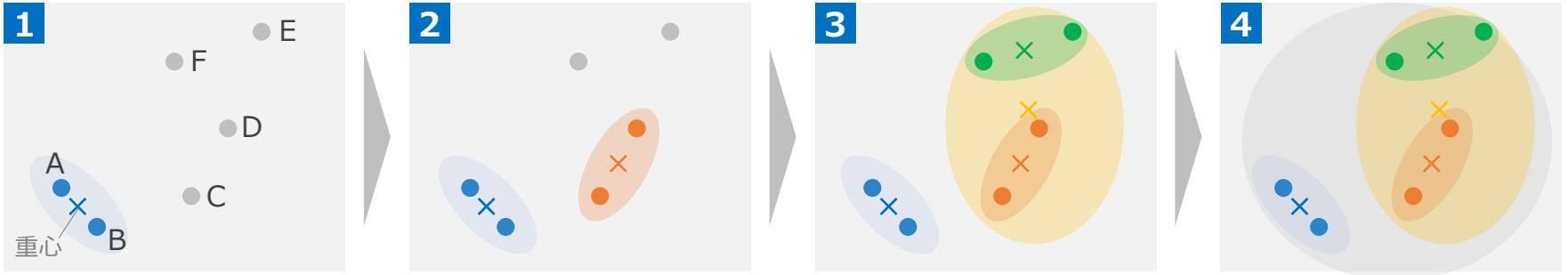
手法の分類	手法		
非階層的クラスタリング	▪ k-means法（k平均法）	クラスタ内データの平均値をクラスタ重心として、距離に基づき、事前に設定したクラスタ数k個に分割	SAS® Studio
	▪ その他	混合ガウス法、超体積法など	第2回で説明
凝集型	**似ている（≒距離の近い）データ／クラスタ同士を逐次まとめる（ボトムアップアプローチ）**		
	▪ ウォード法	クラスタ内のデータの平方和を最小にするように併合	SAS® Studio
	▪ 最短距離法（最近隣法）	距離の近いデータから順番に併合	本日ご説明
	▪ 最長距離法（最遠隣法）	距離の遠いデータから順番に併合	
	▪ 重心法	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	
分割型	**似ていないデータ／クラスタ同士を逐次分離させる（トップダウンアプローチ）**		
	▪ Diana法		





代表的な**階層的**クラスタリング：
凝集型階層クラスタリング

- 凝集型階層クラスタリングは、距離に応じて小さいクラスタを束ねて階層的に分類する手法
- クラスタ数は自動的に決定してくれる他、分類過程を可視化した**樹形図（デンドログラム）**も同時に出力されるので、結果の解釈やクラスタ数の決定に役立つ



最も距離の近い点を同一クラスタとしてまとめ、重心を計算

同様に、重心と残りの点の中で最も距離の近い点同士をまとめ、重心を計算

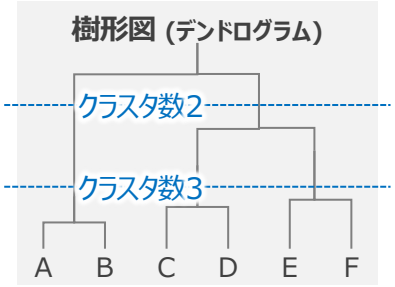
②を繰り返し、クラスタの重心同士が最近傍となった場合は、クラスタ同士をまとめる

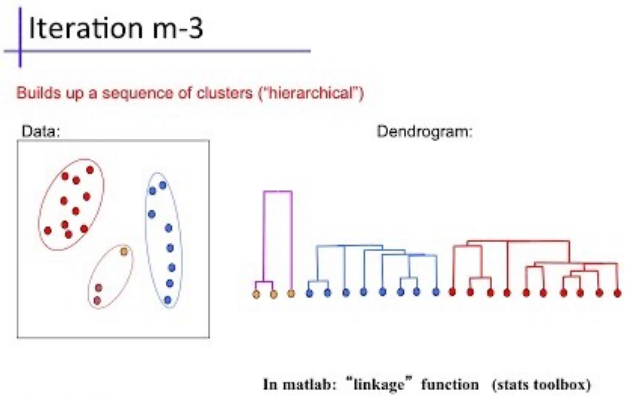
最後にクラスタ全体をまとめる

凝集型階層的クラスタリング
(agglomerative hierarchical clustering)

メリット
- クラスタ数は自動決定
- 樹形図により**分類過程が可視化**されることで、妥当なクラスター数を人が判断可能

デメリット
- 計算量が膨大
- データ量が多い場合、樹形図が複雑となり、解釈が困難になる





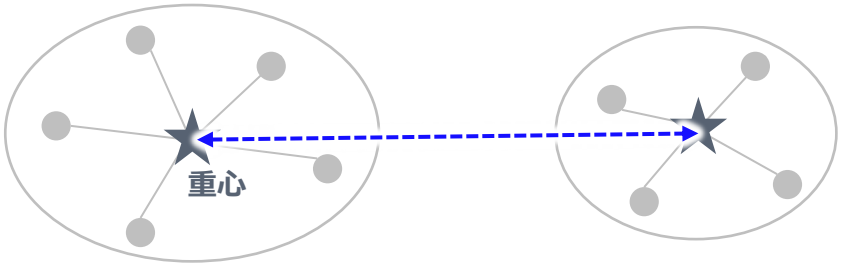
出典：https://youtu.be/OcoE7JlbXvY

「近い」の評価尺度バリエーション

- クラスタ間の「近さ」を測る指標には様々あるが、一概にどれが良いとは言えないため、**複数試して比較**するのが一般的である。ただし、一般には、群平均法やWard法（次頁）が頻用される
- 最短距離／最長距離法は、計算量が少なくて済む反面、1点の影響を大きく受けやすい

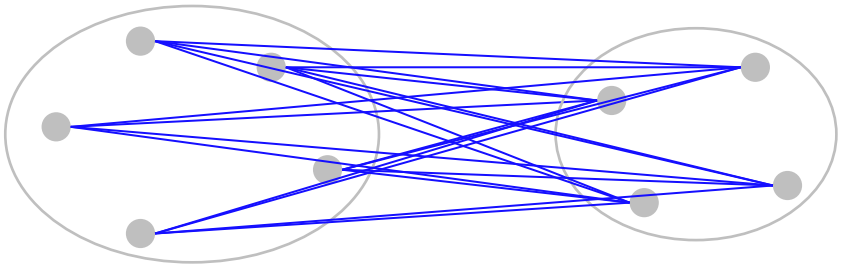
重心法

重心間の距離が近いクラスタを結合



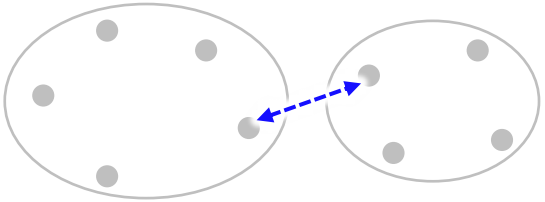
群平均法

クラスタ間で全データ間の距離を算出し、
その**平均値**が近いクラスタを結合



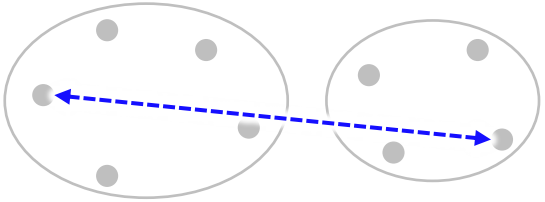
最短距離法

2つのクラスタ間で**最近傍**のデータを
クラスタ間距離として採用



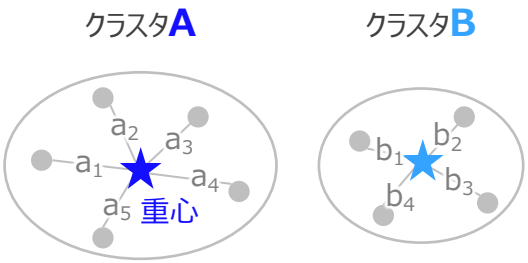
最長距離法

2つのクラスタ間で**最遠方**のデータを
クラスタ間距離として採用

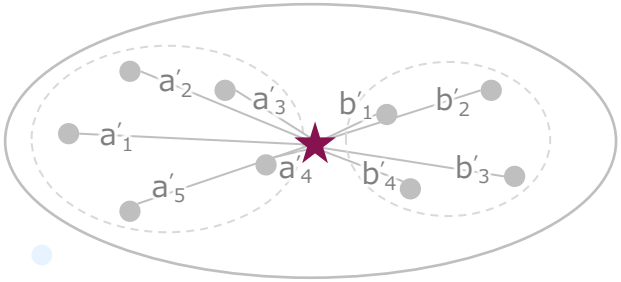


Ward法の考え方

- Ward法*は最もよく用いられる手法であり、計算量は多いが、各データ点とクラスタ重心との関係性まで評価しているため、他手法に比べ、**分類感度が高い**とされる
  *米国の統計学者Joe H. Ward, Jr.が1963年に発表した論文にちなむ



A, Bの結合を仮定した場合のクラスタAB



1 「クラスタ重心」と、「当該クラスタ内の各データ」との距離の総和（二乗和）をクラスタごとに算出

クラスタAの場合

$$A = a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2$$

クラスタBの場合

$$B = b_1^2 + b_2^2 + b_3^2 + b_4^2$$

2 注目する2つのクラスタを結合した場合を仮定し、「結合後のクラスタ重心」と「当該クラスタ内の各データ」との距離の総和（二乗和）を算出

$$AB = a'_1{}^2 + a'_2{}^2 + a'_3{}^2 + a'_4{}^2 + a'_5{}^2 + b'_1{}^2 + b'_2{}^2 + b'_3{}^2 + b'_4{}^2$$

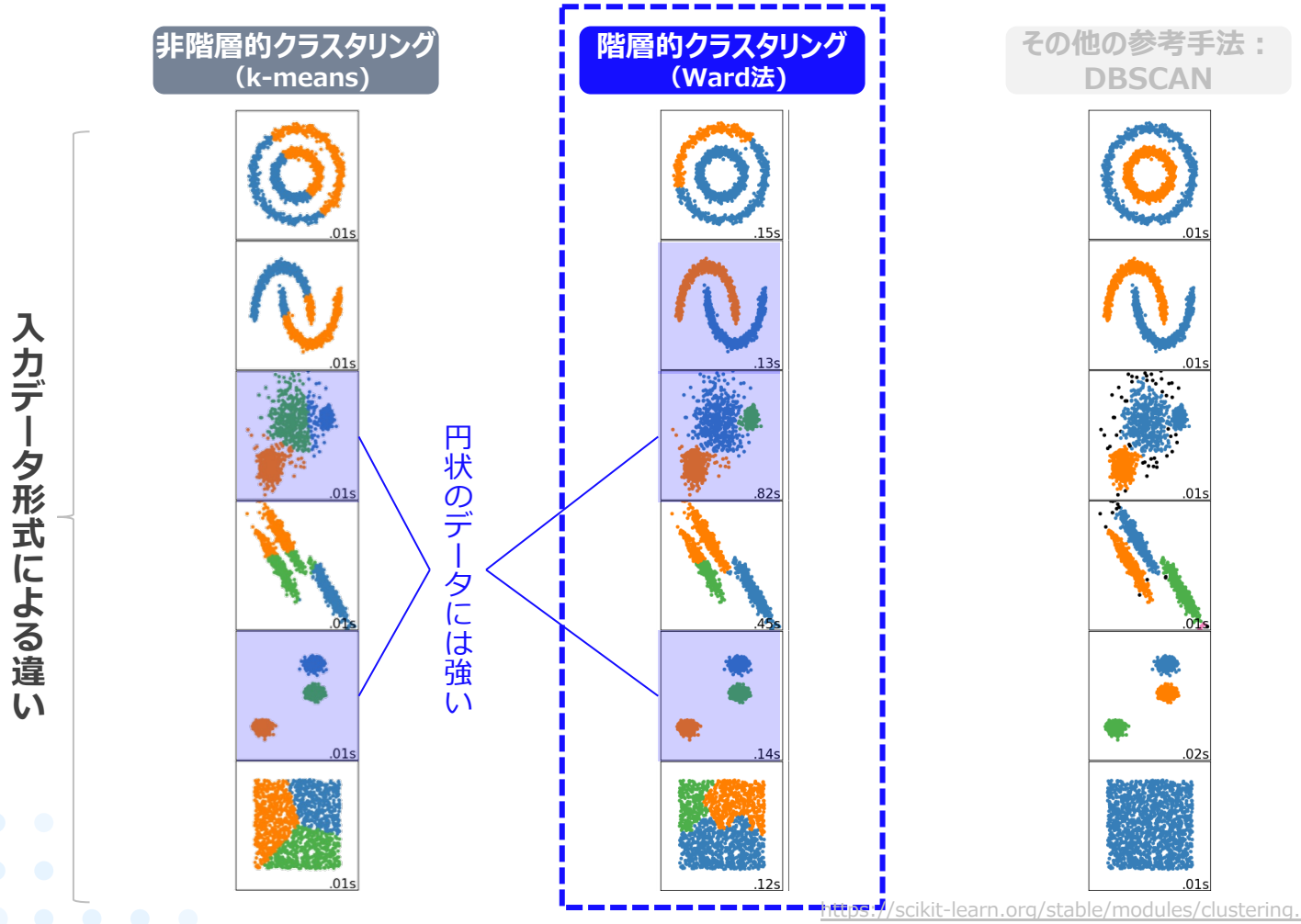
3 1 と 2 の差、つまり、**AB** − (**A**+**B**) が**最小**となるクラスター結合を採用
（結合前後でクラスタ内のばらつきに変化なし→統合してもOKと判定）

※近くにあり、ばらつきの小さいクラスタ同士が結合しやすい

再掲

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



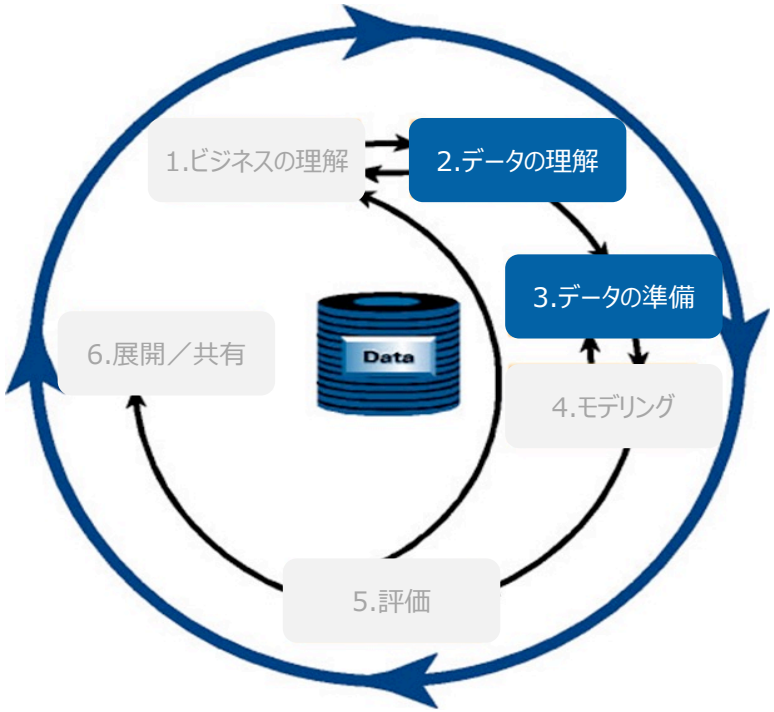
https://scikit-learn.org/stable/modules/clustering.html

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

使用データ（e-Statについて）

- 政府が公開する政府統計のオープンデータ ”e-Stat” のデータを活用する
- 今回扱うデータの他にも、様々な統計データが公開されているので、企業内のデータと組み合わせることで、さらなる付加価値を生む可能性がある

Source: https://www.e-stat.go.jp/

使用データ

- 今回は、このうち、5年に1度実施している「全国消費実態調査」（現在の名称は「全国家計構造調査」）のデータを用いて、**都道府県別の消費動向から、類似の都道府県をグルーピング**することを考える

政府統計名	全国家計構造調査（旧全国消費実態調査）	詳細
提供統計名	平成26年全国消費実態調査	
提供分類1	全国	
提供分類2	家計収支に関する結果	
提供分類3	総世帯	

表番号	統計表	調査年月	公開（更新）日	表示・ダウンロード
フロー編				
42	年間収入階級・年間収入十分位階級別１世帯当たり１か月間の収入と支出			
	総世帯	2014年	2015-12-16	EXCEL DB
	勤労者世帯	2014年	2015-12-16	EXCEL DB
43	世帯主の年齢階級別１世帯当たり１か月間の収入と支出			
	総世帯・勤労者世帯	2014年	2015-12-16	EXCEL DB
44	住居の所有関係別１世帯当たり１か月間の収入と支出			
	総世帯・勤労者世帯	2014年	2015-12-16	EXCEL DB
45	資産の種類・資産額階級別１世帯当たり１か月間の収入と支出（純資産）			
	総世帯	2014年	2016-03-25	EXCEL DB
	勤労者世帯	2014年	2016-03-25	EXCEL DB
	資産の種類・資産額階級別１世帯当たり１か月間の収入と支出（総資産）			
	総世帯	2014年	2016-03-25	EXCEL DB
	勤労者世帯	2014年	2016-03-25	EXCEL DB
地域編				
13	地域別１世帯当たり１か月間の収入と支出			
	総世帯	2014年	2015-12-16	EXCEL DB
	勤労者世帯	2014年	2015-12-16	EXCEL DB

Source: https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00200564&tstat=000001073908&cycle=0&tclass1=000001073965&tclass2=000001074840&tclass3=000001077457&tclass4val=0

データの概要（加工前）

- e-Statより素データをダウンロードして開くと、開始行や開始列がずれていたり、空白行があったりと、加工が必要な形式であることがわかる
- 今回は、本データから都道府県別の消費細目データ部分を抽出し、加工済のデータを用いる

第 13 表　地域別 1
Table 13.　Monthly Receipts and Di

総世帯
Total Households

都道府県データ

収支項目	都道府県 Prefectures					都道府県（続き）									
	北海道	青森県	岩手県	宮城県	秋田県	山形県	福島県	茨城県	栃木県	群馬県	埼玉県	千葉県	東京都	神奈川県	新潟県
	Hokkaido	Aomori -ken	Iwate -ken	Miyagi -ken	Akita -ken	Yamagata -ken	Fukushima -ken	Ibaraki -ken	Tochigi -ken	Gumma -ken	Saitama -ken	Chiba -ken	Tokyo -to	Kanagawa -ken	Niigata -ken
集計世帯数	2,164	739	746	789	770	757	877	1,424	877	905	2,878	2,759	2,563	2,671	
世帯数分布（抽出率調整）	2,374,215	476,970	507,292	923,420	375,617	379,431	757,313	1,076,891	722,506	727,684	2,850,919	2,467,193	6,157,574	3,918,553	849,
（1万分比）	459	92	98	178	73	73	146	208	140	141	551	477	1,190	757	
世帯人員（人）	2.21	2.52	2.54	2.49	2.53	2.73	2.59	2.63	2.57	2.63	2.49	2.43	2.14	2.29	2.
18歳未満人員（人）	0.36	0.40	0.37	0.38	0.35	0.40	0.43	0.43	0.42	0.44	0.42	0.39	0.35	0.35	0.
65歳以上人員（人）	0.64	0.76	0.82	0.77	0.80	0.83	0.77	0.68	0.72	0.71	0.65	0.69	0.58	0.61	0.
うち無職者人員（人）	0.54	0.58	0.60	0.65	0.64	0.65	0.64	0.53	0.56	0.55	0.52	0.55	0.42	0.49	0.
有業人員（人）	0.97	1.25	1.29	1.12	1.29	1.44	1.22	1.29	1.23	1.30	1.17	1.11	1.07	1.09	1.
世帯主の年齢（歳）	59.2	58.9	60.2	59.1	59.6	56.3	57.9	57.3	58.1	58.2	56.3	57.6	57.1	56.1	5
世帯主の性別 男（人）	0.721	0.752	0.744	0.727	0.789	0.801	0.785	0.843	0.820	0.786	0.795	0.808	0.742	0.784	0.
女（人）	0.279	0.248	0.256	0.273	0.211	0.199	0.215	0.157	0.180	0.214	0.205	0.192	0.258	0.216	0.
持ち家率（現住居）（%）	75.7	77.2	80.8	73.2	80.6	80.8	75.9	83.6	83.9	82.5	77.4	80.9	67.6	71.7	8
うち住宅ローン保有率（%）	22.4	16.1	20.1	19.3	16.9	22.0	17.1	21.9	23.9	23.2	25.3	24.8	23.3	23.5	2
家賃・地代を支払っている世帯の割合（%）	25.5	20.9	20.7	24.0	19.8	18.4	21.2	16.4	18.1	15.8	21.7	19.6	34.3	26.4	1
現住居の延べ床面積（㎡）	107.1	126.9	134.2	117.6	141.7	148.9	124.6	120.5	125.9	123.9	96.7	98.2	77.3	96.7	13
自動車保有台数（千世帯当たり）	1,074	1,390	1,459	1,326	1,544	1,751	1,603	1,697	1,666	1,697	996	1,020	490	731	1,
自動車保有率（%）	77.5	83.4	82.0	82.0	88.8	92.5	90.9	90.8	92.9	91.7	73.0	74.9	43.6	61.8	8
年間収入（千円）	4,592	4,451	4,993	4,994	5,153	5,821	5,085	5,676	5,576	5,449	5,678	5,752	6,004	5,762	5,
消費支出	231,757	215,712	239,630	251,694	229,200	266,205	245,341	269,082	275,745	255,178	265,501	269,903	278,284	269,216	251
食料	54,281	55,180	57,514	59,052	59,146	63,042	57,891	62,433	63,866	61,858	64,632	66,536	68,380	67,197	64,
穀類	5,147	4,726	5,013	5,048	5,170	5,726	5,549	5,465	5,516	5,901	5,512	5,393	4,899	5,244	6,
米	2,113	1,742	2,004	2,090	2,550	2,443	2,528	2,237	2,146	2,359	1,814	1,796	1,369	1,575	3,
パン	1,760	1,617	1,633	1,632	1,296	1,627	1,619	1,781	1,779	2,000	2,130	2,175	2,128	2,197	1,
麺類	1,017	1,163	1,111	1,114	1,173	1,352	1,167	1,168	1,277	1,270	1,253	1,113	1,090	1,161	1,
他の穀類	258	204	264	213	152	304	235	280	315	272	315	309	313	310	
魚介類	5,137	5,543	5,651	5,223	5,828	5,390	5,047	5,255	5,286	5,239	4,979	5,401	4,886	5,123	5,
生鮮魚介	3,073	3,379	3,453	3,[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3,101	3,080	3,021	3,351	3,019	3,138	3,
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

消費の細目

消費の細目に関わるデータを活用

データの概要（加工後）

- e-Statより素データをダウンロードして開くと、開始行や開始列がずれていたり、空白行があったりと、加工が必要な形式であることがわかる
- 今回は、本データから都道府県別の消費細目データ部分を抽出し、加工済のデータを用いる

予測(分析)対象を説明するための変数

都道府県

説明変数

#	都道府県	食料	住居	光熱・水道	家具・家事用品	被服及び履物	保健医療	交通・通信	教育	教養娯楽	読書	聴視・観覧	旅行	スポーツ	月謝	会費・つきあい費
1	北海道	54281	17491	19520	9089	10208	9234	35627	5630	23323	3894	2106	6506	2492	1430	1011
2	青森県	55180	14357	22420	9162	8972	9936	33039	4880	16564	3256	2084	3839	1153	777	1079
3	岩手県	57514	14782	21267	8973	8288	10273	42912	5805	20278	3534	2428	6326	1447	1219	1458
4	宮城県	59052	16353	20331	9700	10640	10851	40742	7331	23394	3241	2322	7734	2172	1587	1168
5	秋田県	59146	12875	22394	9108	7467	9214	37645	4472	21037	3349	2403	6672	1617	979	1191
6	山形県	63042	13186	24030	11771	9486	10407	40365	6667	24577	3706	2227	5438	1659	1333	2230
7	福島県	57891	12579	20102	9633	8949	9613	44195	4614	21216	3637	1921	6701	1743	1432	1027
8	茨城県	62433	17292	20330	9186	10645	10668	44764	10113	26592	3638	2451	7254	2639	2041	1134
9	栃木県	63866	18994	19997	9914	10252	12956	47208	6889	27561	3564	2723	7996	2908	2400	996
10	群馬県	61858	15629	18305	9605	9682	11270	43782	7539	25659	3767	2473	6800	2837	1965	931
11	埼玉県	64632	20131	17747	8544	11403	11616	40152	12953	29055	3800	3115	8887	2699	2705	632
12	千葉県	66536	17887	18039	8820	11826	11859	39048	12165	30385	4186	3512	10407	2889	2800	684
13	東京都	68380	33295	16315	8691	12404	12151	33118	11060	32038	4180	3882	14361	3061	2964	808
14	神奈川県	67197	22708	16957	8783	11591	11443	38440	11004	31833	4159	3607	10790	2874	2894	889
15	新潟県	64400	15713	21881	8900	9077	10628	38983	6736	23878	3605	2535	7327	1955	1715	1032
16	富山県	67635	14518	21894	10624	9387	10776	51532	7879	27246	4003	4101	5869	2139	2433	1373
17	石川県	66478	17678	18423	8733	9512	11135	42087	7993	27548	4220	3207	8516	1898	1966	1402
18	福井県	67429	12168	20741	9034	10204	11287	45576	9585	27984	3482	4031	8052	1863	2080	1539
19	山梨県	57641	17234	18209	7890	9429	10280	39392	9066	25849	3531	3699	6404	1877	2212	1132
20	長野県	62406	21145	21350	9866	9375	11987	42846	8047	27147	4035	2910	7265	2537	1830	1224
21	岐阜県	61939	12754	19952	9042	9942	10463	41580	[illegible]	25777	3627	2610	6754	2280	2056	1306
22	静岡県	62396	15048	18407	9012	9985	11488	[illegible]	[illegible]	28082	3714	2936	8355	2350	2259	1048
23	愛知県	64248	21485	17573	9010	11051	11880	[illegible]	[illegible]	28967	4041	3305	8547	2948	2655	818
24	三重県	63275	12856	19237	9036	11444	12889	[illegible]	[illegible]	28462	3556	3578	8754	2773	2385	1173
25	滋賀県	63385	16479	18807	9587	10034	10818	[illegible]	[illegible]	26609	3280	2845	8377	1848	2052	1486
26	京都府	65337	13829	17928	8409	12630	9239	36643	[illegible]	26012	3984	3109	8260	1855	2174	985
27	大阪府	62386	18778	16292	7230	9898	10782	31046	10348	25016	3744	3264	7492	2465	2250	658
28	兵庫県	63620	19262	16725	8281	10712	10926	36040	9806	27000	3827	3023	8397	2546	2548	823
29	奈良県	66408	17630	19784	9875	11068	12405	42593	14481	27121	3849	3065	9467	2170	2609	986
30	和歌山県	58010	10696	17125	8152	9250	8326	36333	6001	23890	3376	2656	4995	2098	1787	888
31	鳥取県	58027	13626	18488	8143	9050	10320	41570	4966	24212	3198	3787	7077	1775	1865	904
32	島根県	59223	11926	19494	8915	8767	11814	40722	3866	23678	3446	3538	6335	1619	1512	1656
33	岡山県	58368	13776	18306	8286	9846	10347	38978	8451	24914	3052	2796	6682	2430	1789	931
34	広島県	58058	17721	17128	9180	9622	11195	38580	8773	24978	3308	2660	7944	1918	2004	997
35	山口県	55832	18576	16610	9381	8003	10961	35524	5193	23931	3557	2873	6234	1514	1798	840
36	徳島県	55896	16389	18015	8680	9656	10261	38507	6659	23923	3439	3762	7064	2039	1810	997
37	香川県	57352	15438	17319	8338	8754	11070	40876	6059	25565	3476	2814	6197	2689	2176	942
38	愛媛県	55531	13489	17201	8171	8284	9224	32679	7901	19353	2872	2450	5539	1732	1774	937
39	高知県	54971	14463	16479	7609	7510	10329	32613	5206	20184	3273	2447	5228	1722	1167	867
40	福岡県	54633	18999	16314	8029	9823	10405	36057	7360	24134	3073	2790	9478	2489	1809	784
41	佐賀県	57104	13214	17556	8682	9647	11281	40406	6975	24864	3354	2813	6326	2333	1889	1189
42	長崎県	51798	18624	16853	7291	7934	10115	34480	6345	19631	2687	2421	7528	1383	1517	1013
43	熊本県	55006	11286	16802	8254	10041	11155	34633	6967	21046	2889	2309	5544	1688	2016	682
44	大分県	53558	14707	15685	8558	10853	11677	36458	3243	22105	3021	3223	5354	2139	1327	1153
45	宮崎県	53347	15963	15828	8228	8386	9476	36294	6276	21314	2565	3061	6208	2410	1413	1210
46	鹿児島県	50294	14792	15496	7800	7857	10022	39992	5063	18721	2593	2002	5533	2053	1160	1488
47	沖縄県	48770	22616	17251	6750	5010	8088	28055	5169	16217	2500	1492	3913	1767	1700	970

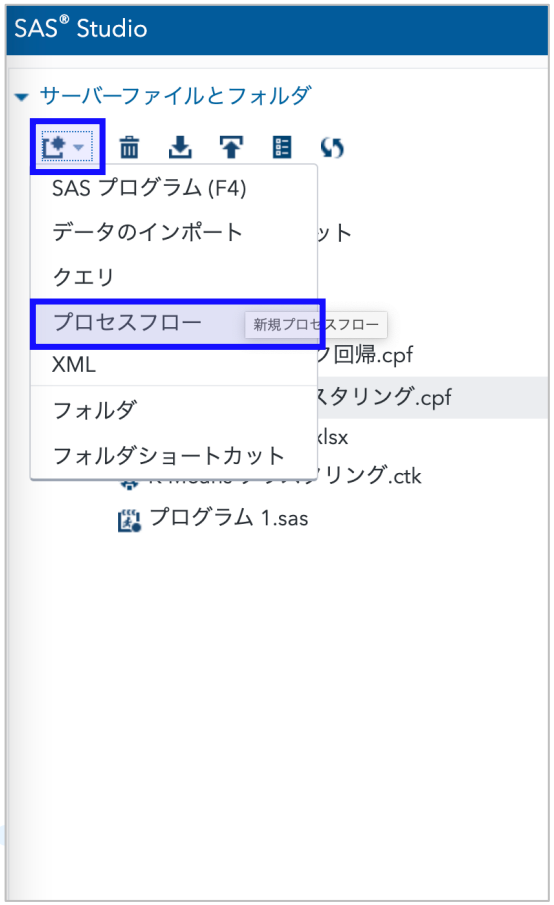
SAS® Studio

SAS Studio での実装方法

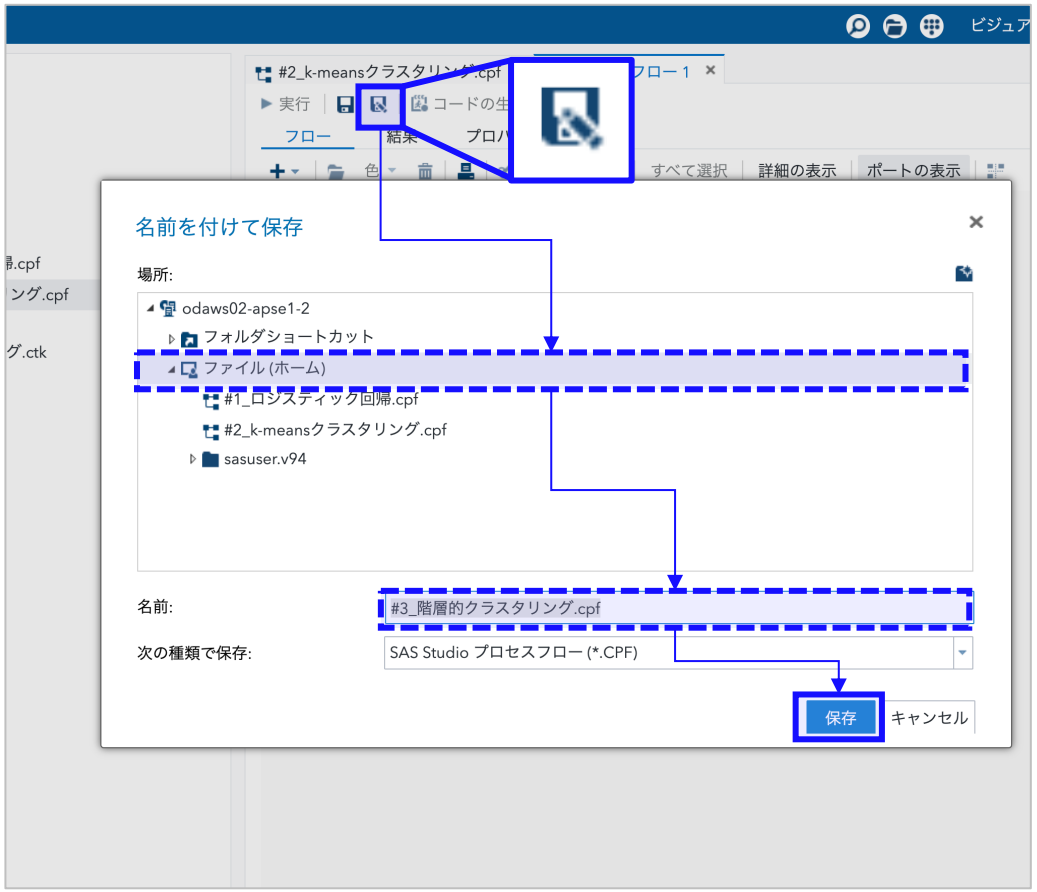
- データの読み込み
- 階層的クラスタリング（Ward法）
- 標準化したクラスタリング

新規プロセスフローの作成と保存

①左上メニューの アイコンをクリックし、
[プロセスフロー] を選択

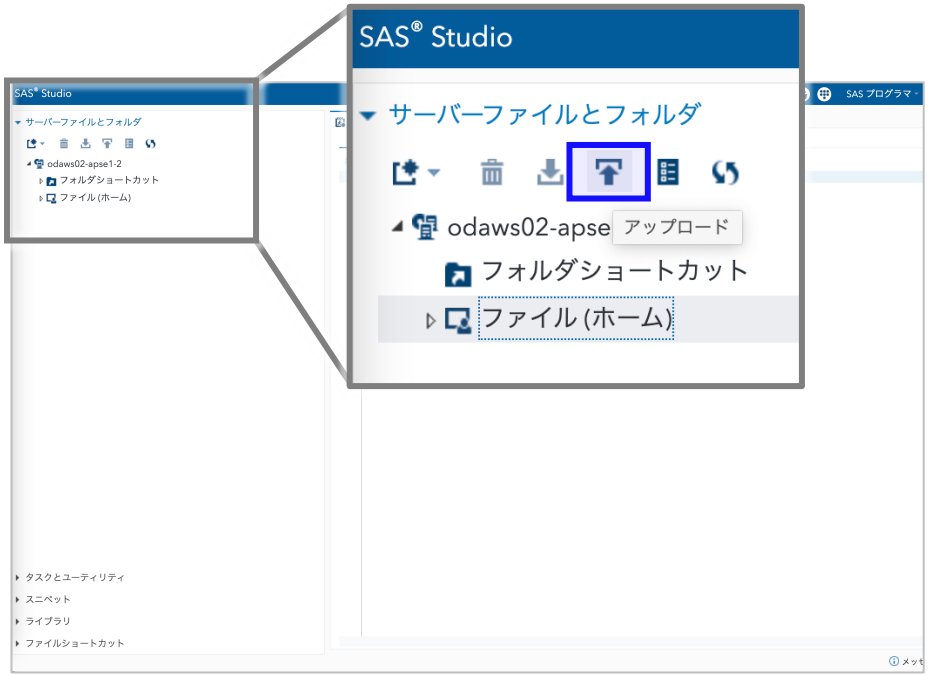


②新規のプロセスフローが作成されるので、
「**名前を付けてプロセスフローを保存**」アイコンをクリックし、
保存場所、ファイル名を指定して保存ボタン

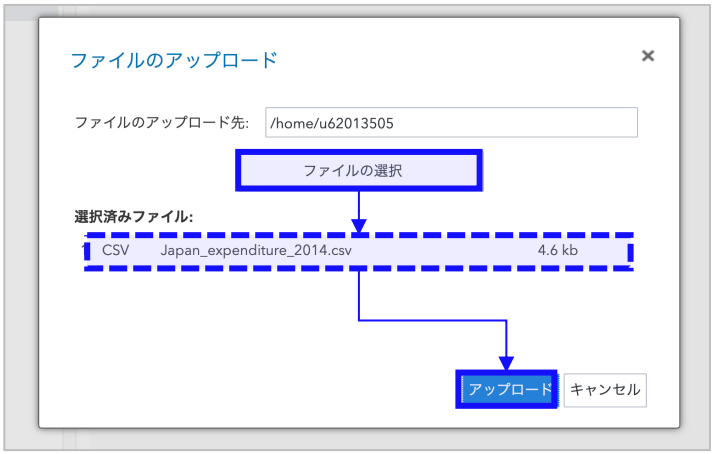


データの読み込み（1/2）

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“Japan_expenditure_2014.csv”を選択し、OKボタン
③「アップロード」ボタンをクリック

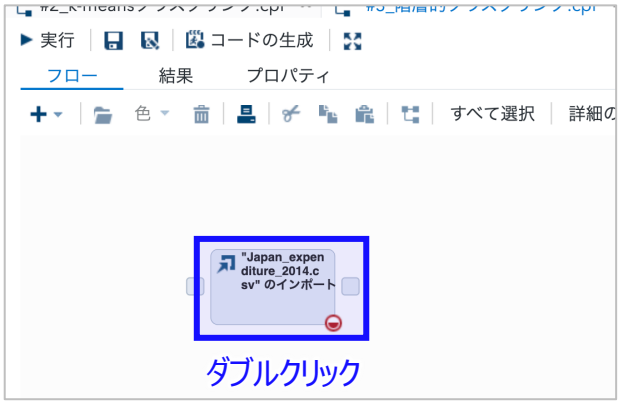


④左パネル内にファイルがアップロードされていることを確認

データの読み込み（2/2）

①左パネル内の “Japan_expenditure_2014.csv” を選択し、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）

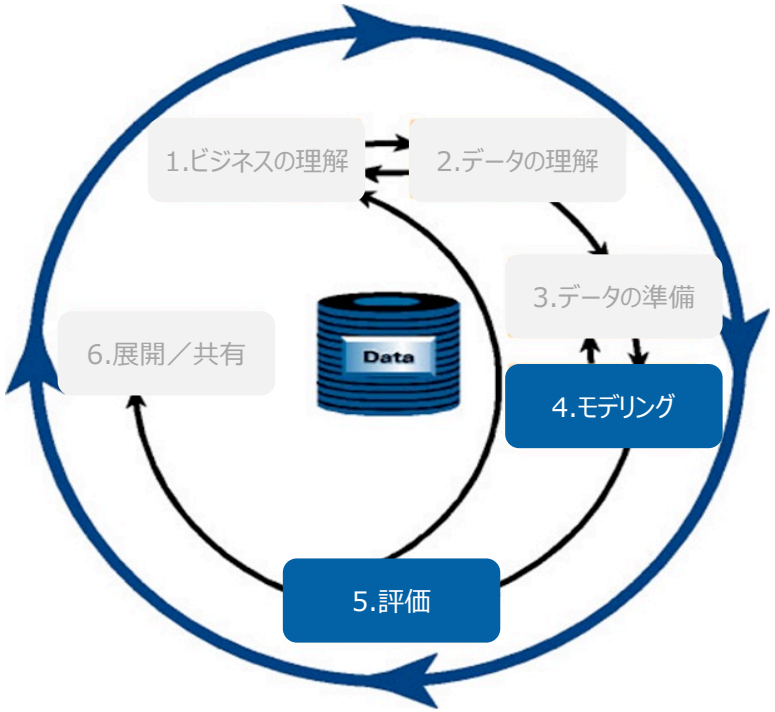
④「結果」のタブ画面に読み込んだデータの概要が出力

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)

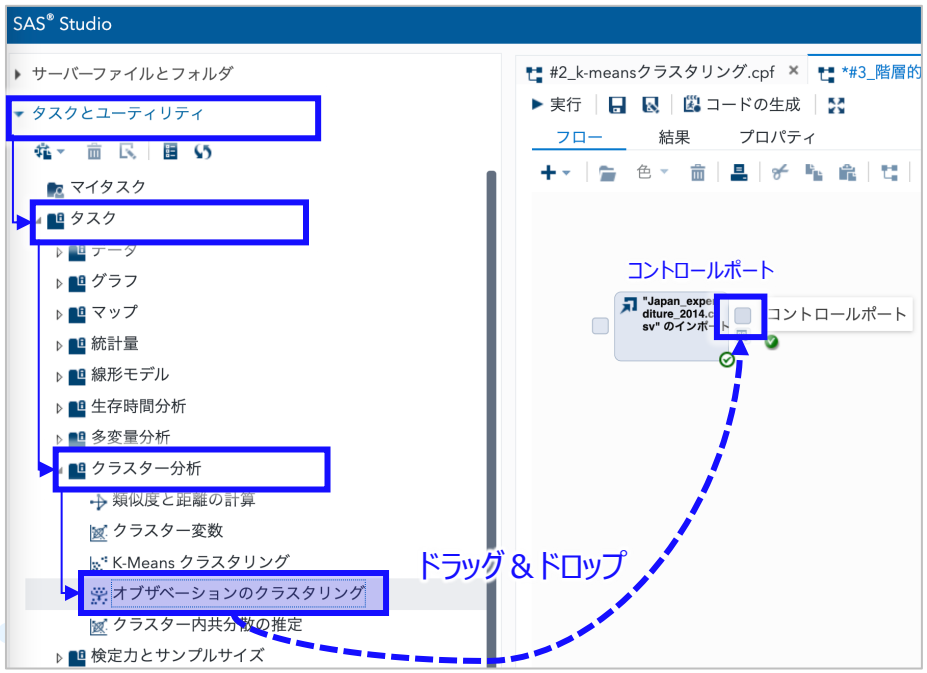
1.ビジネスの理解	•ビジネス、データマイニング目標の決定 •プロジェクトの立ち上げ
2.データの理解	•データの収集 •データの調査 •データ品質の検証
3.データの準備	•データの選択や除外 •データのクリーニング •データの構築や統合
4.モデル作成	•モデリング手法の選択 •モデルの作成 •モデルの評価
5.評価	•データマイニングの結果の評価 •プロセスの見直し •実行可能なアクションリストの作成
6.展開／共有	•業務への導入計画 •モニタリング、メンテナンスの計画

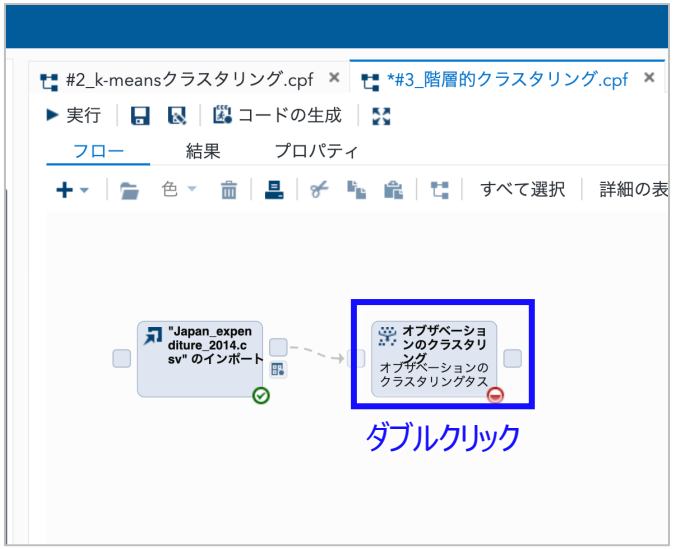
階層的クラスタリング（ward法）– 実行方法（1/2）ノードの設置

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]→[**オブザーベーションのクラスタリング**]
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 オブザーベーションの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

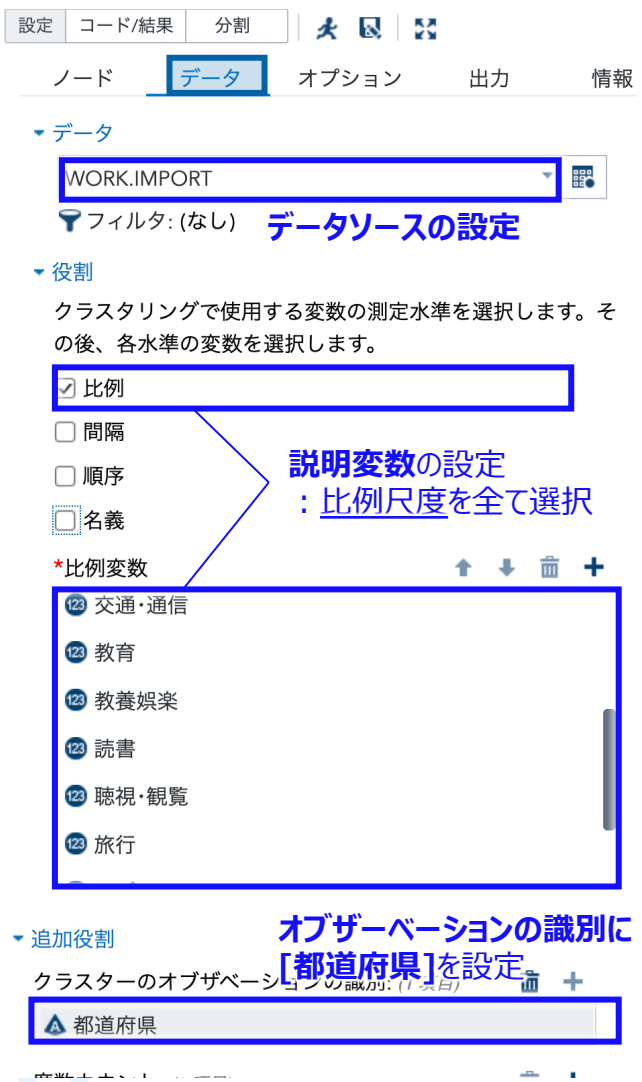




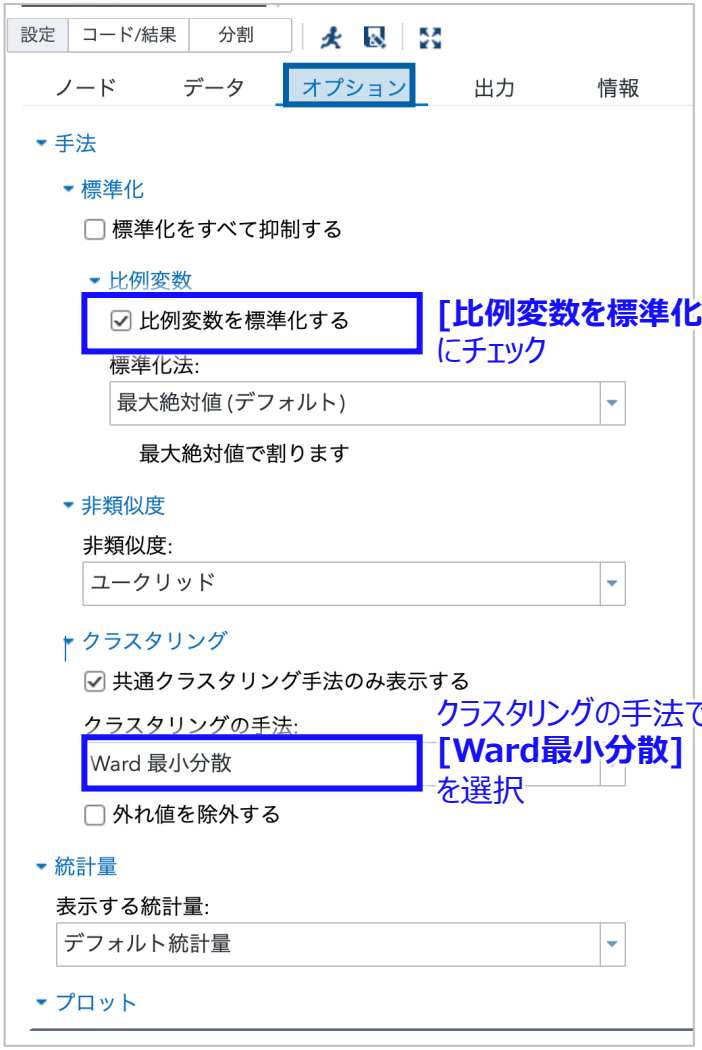
SAS® Studio

階層的クラスタリング（ward法）– 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）



再掲

参考：変数の尺度（名義尺度・順序尺度・間隔尺度・比例尺度）

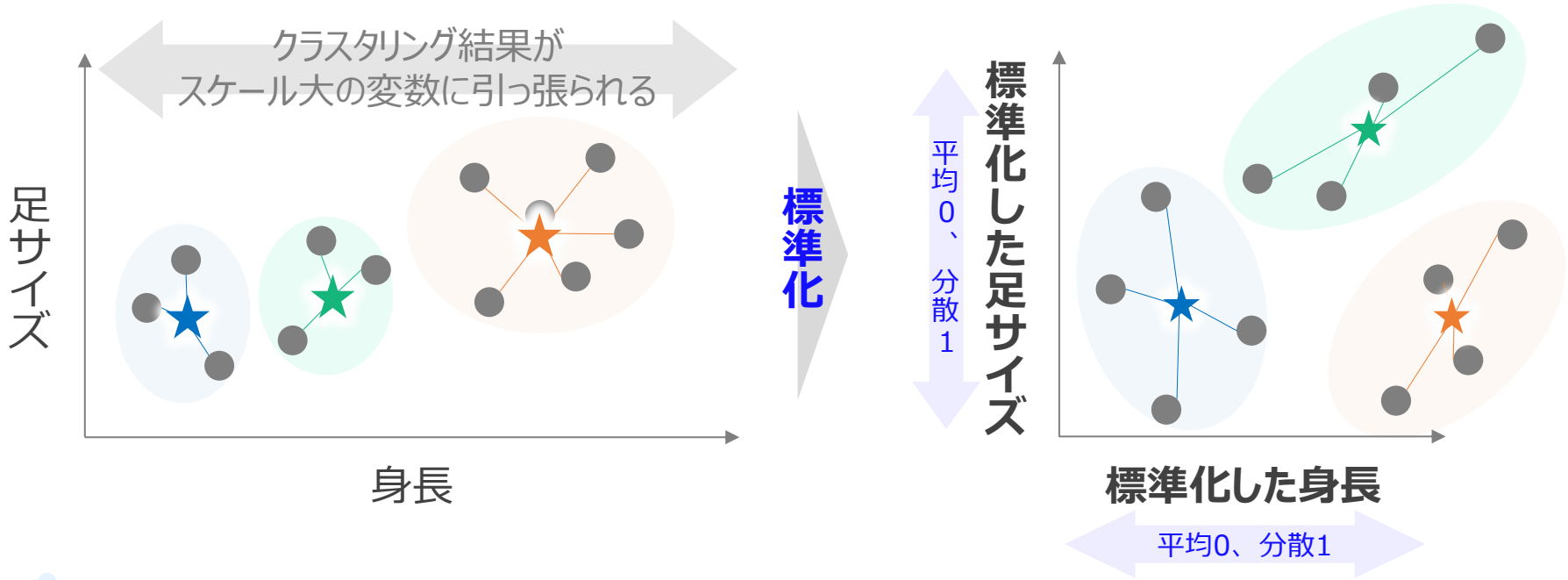
- 変数の種類は大きく「**質的データ**」と「**量的データ**」に分けられ、それぞれの特性に合わせて扱う必要がある

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	●　●　●

再掲

クラスタリングにおける変数スケールの影響と標準化

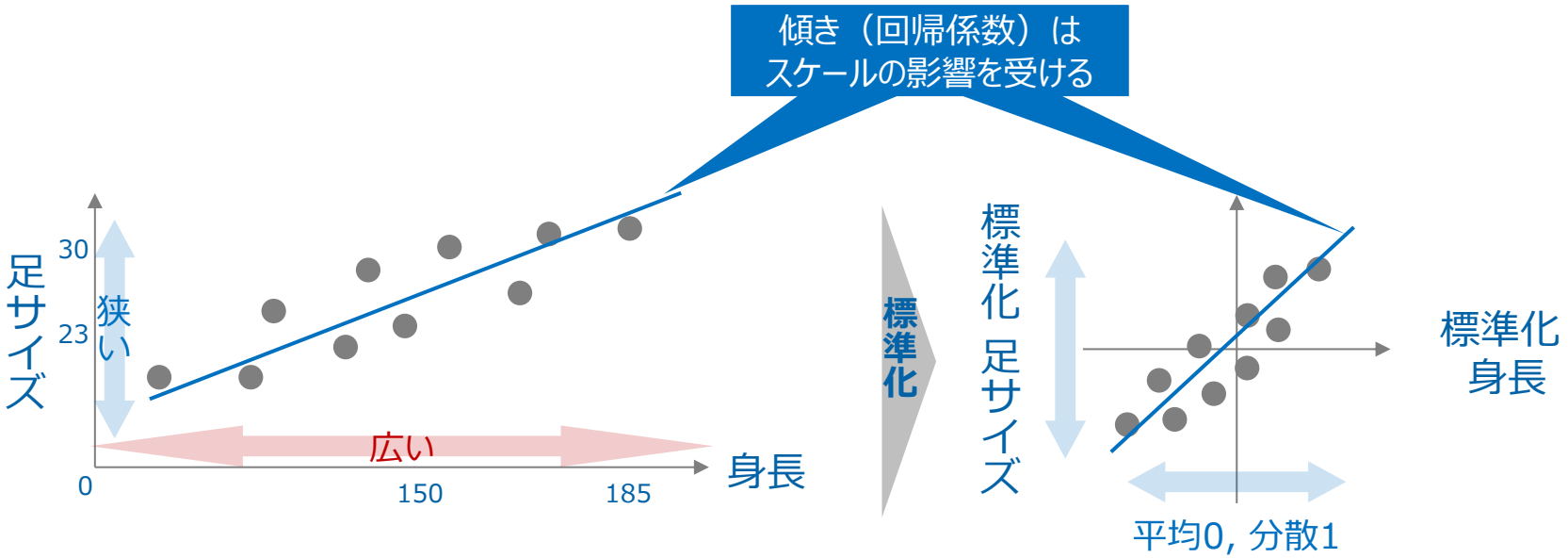
- k-means法などの「距離」に基づくによるクラスタリング手法は、データの「**スケール**」に大きく影響を受ける。このため、必要に応じて、「**標準化**」の処理を行なった上でクラスタリングを行う必要がある



再掲

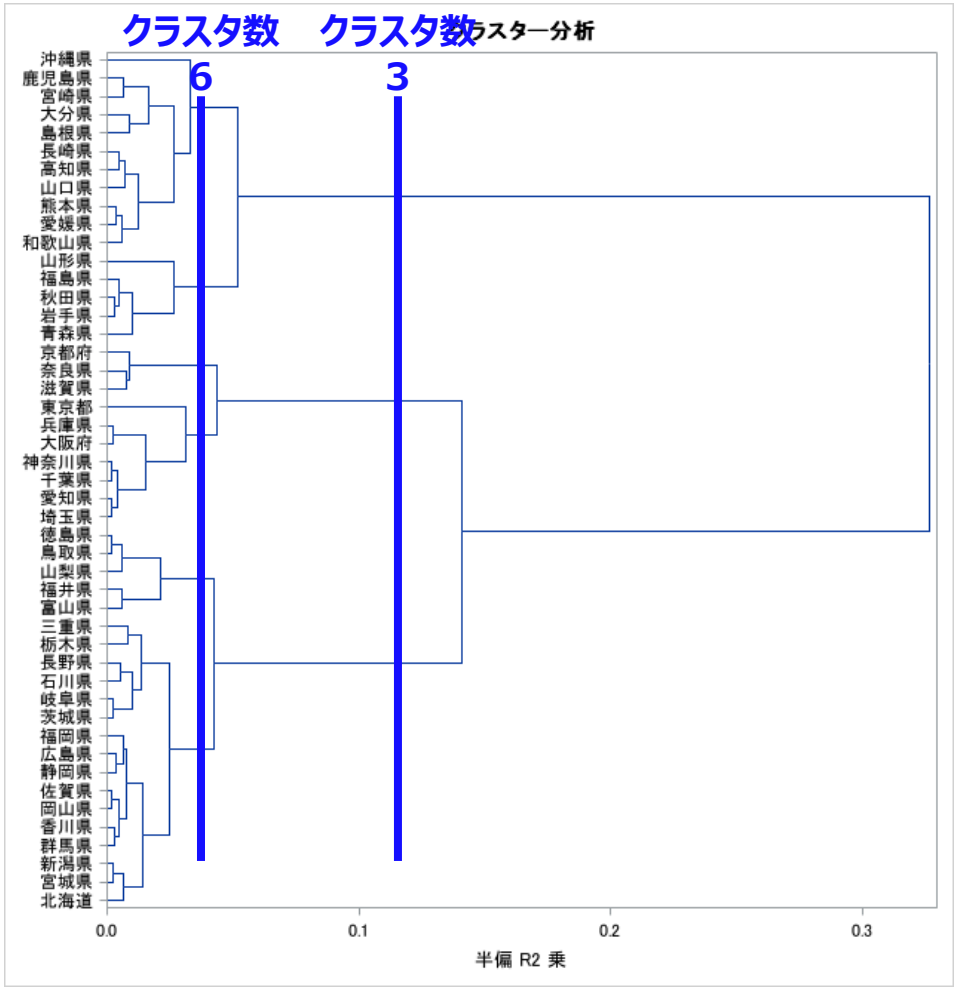
（参考）回帰分析における標準化の有効性

- 機械学習では、各変数間でスケール（値範囲）が大きく異なると、計算に時間がかかったり、回帰係数などのパラメータの直接比較が困難になるため、**スケールを揃える**ことが有効
- 特に、各変数を**平均0, 分散1に変換する「標準化」**を用いることが多い



階層的クラスタリング（Ward法） – 実行結果

- Ward法の結果、47都道府県が階層的にクラスタリングされた。Ward法では、全体的にバランスよく、分類が行われていることが確認できる
- 任意の場所で区切ることで、最適なクラスタ数の検討が可能



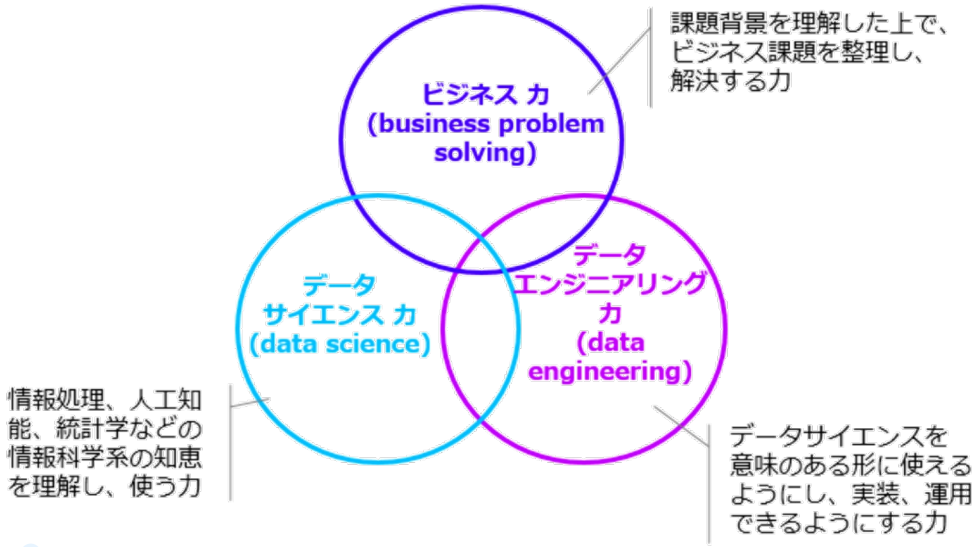
Agenda

- クラスター分析の応用（他の分析手法との組み合わせ）
  - 主成分分析により説明変数を要約する
  - 主成分軸でクラスター分析を行う
- クラスター分析による分類（2）：階層的クラスタリング
  - 階層的クラスタリング（群平均法、重心法、Ward法）のしくみ
  - 樹形図（デンドログラム）とクラスタ数の検討
  - 都道府県データを用いて階層的クラスタリングにより類似地域を分析する
- **今後のデータサイエンス学習に向けたスキルアップ**
  - データサイエンティストに求められるスキル
  - SAS内サンプルデータの紹介と使い方
  - オープンデータの紹介

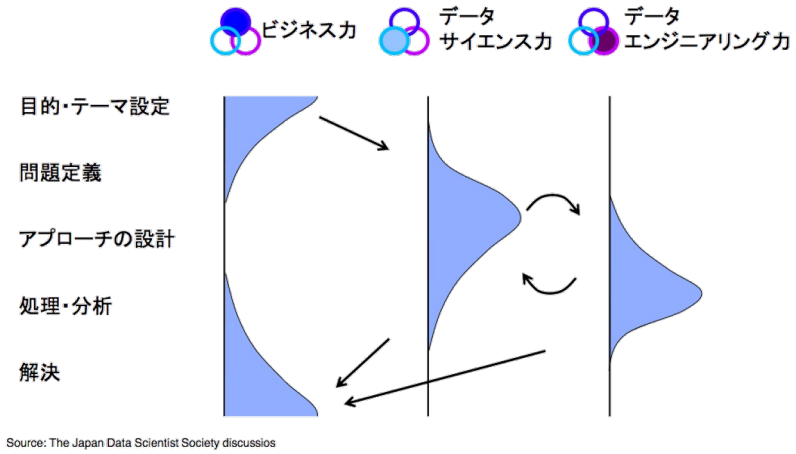
データサイエンティストに求められる知識・スキルセット

- データサイエンティスト協会が定義するデータサイエンティスト：
「データサイエンティストとは、**データサイエンス力、データエンジニアリング力**をベースにデータから価値を創出し、**ビジネス**課題に答えを出すプロフェッショナル」
- これら3スキルはどれも不可欠で、分析フェーズによって中心となるスキルが変化する、としている

一般社団法人
データサイエンティスト協会



課題解決の各フェーズで要求されるスキルセットのイメージ

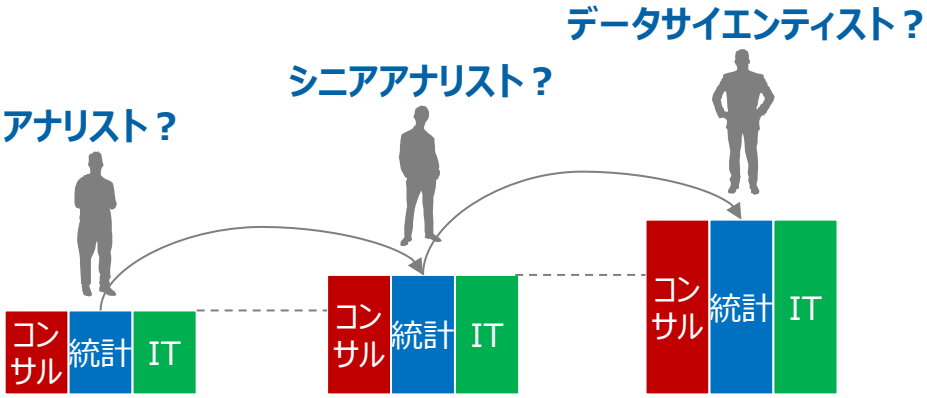


出展：データサイエンティスト協会資料
http://www.datascientist.or.jp/news/2014/pdf/1210.pdf

データサイエンティストの育成モデルの例

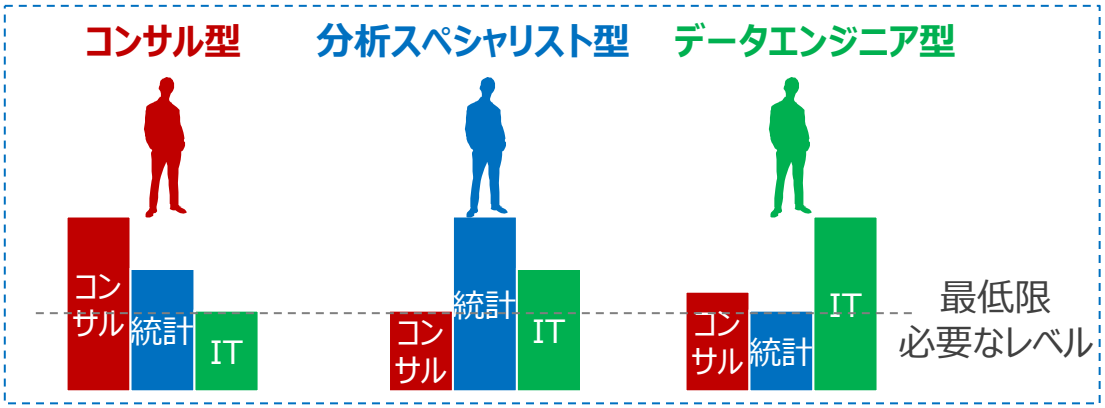
- 理想的には、すべてのスキルを持つ人材を育成できればよいが、元々のバックグラウンドや経験値を生かした育成を行うのが現実的
- チームでスキルを補い合い、プロジェクトフェーズに応じて、役割分担や協業体制が必要

全方位育成モデル



現実的には・・・

バックグランドベース育成モデル



チームでスキルを補い合う

参考：データサイエンティスト協会が定義するスキルレベル

- バックグランドベース育成モデルを実現する上でも、各メンバーは、各スキルについて最低限のレベルを保有する必要がある。データサイエンティスト協会が定義する「**独り立ちレベル**」がまず目標

レベル定義	コンサルカ（ビジネス力）	統計力（データサイエンス力）	IT力（データエンジニアリング力）
業界代表	▪ **組織や市場全体にインパクト**を出せる ▪ 対象とする事業全体、産業領域における課題の切り分け、テーマ、論点の明確化ができる	▪ **新しいアルゴリズムや分析手法の開発**ができる ▪ 複数のパラメータやアルゴリズムの選択など、**適切な分析アプローチの設定**ができる	▪ **複数のデータソースを統合したデータシステム**、もしくはデータプロダクトの構築、全体最適化ができる
棟梁	▪ 分析を通じ**オペレーション上の革新が実現**できる ▪ **仮説や可視化された問題がない中で、適切に問題を定義**し、解き、価値を見出すことができる ▪ 特定の課題領域において、課題と取組のテーマを構造的に整理し、見極めるべき論点をクリアにできる ▪ 組織全体を見渡して、必要なデータの当たりをつけることができる	▪ 多変量解析の概念を理解し、活用することができる ▪ **機械学習、自然言語、画像処理のアルゴリズム**を理解し、適切に活用、問題解決することができる ▪ **モデルを構築**できる	▪ 分析に必要なデータフォーマット、取得蓄積仕様等を設計できる ▪ 問題設定に応じた新規データマート設計ができる ▪ **構造化データ／非構造化データ**を問わず、分析システムを設計できる ▪ **構築したモデルを実装**できる ▪ データ分析を作ったシステムを自身で構築できる

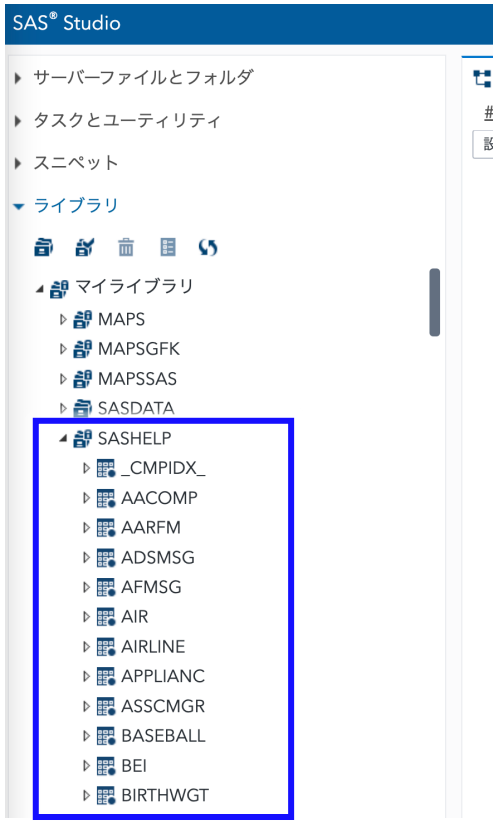
全てのスキルで「独り立ちレベル」以上の習得が理想

独り立ち	▪ **仮説や既知の問題が与えられた中で、最適解・最大解を見出す**ことができる ▪ 扱っている課題領域で新規の課題を切り分け、構造化できる ▪ 当該プロジェクト・サービスを超えて、必要なデータの当たりをつけることができる	▪ **SPSS/SAS/R等が使える**。指示されなくてもサンプル抽出ができるとともに内容を確認できる ▪ **データクレンジング、分布、単回帰やP値の概念を理解し、活用**することができる	▪ **大規模のファイルや、データベースにアクセスし大量の構造化データを処理**することができる
見習い	▪ ビジネスにおける論理とデータの重要性を認識している ▪ **仮説や既知の問題が与えられた中で、必要なデータに当たりをつけて、データを用いて改善**することができる ▪ 扱っている課題領域における基本的な課題の枠組みが理解できる	▪ 基本統計量（平均、中央値など）の知識を有し、**指示されればデータの抽出、グラフ作成を正しく行う**ことができる	▪ 一般的なアクセス解析システムを使うことができる ▪ **抽出されたデータサブセットに対し、ExcelやAccess等の統合環境を用い**、目的に応じた処理をすることができる
未経験	▪ ビジネスは勘と経験だけで回すものと思っている ▪ 課題を解決する際に、そもそも定量化する意識がない	▪ 基本統計量の意味を正しく理解していない ▪ 指数を指数で割り算したりする ▪ 「平均年収」をそのまま鵜呑みする ▪ グラフ・チャートの使い方が不適切	▪ レポートされてくる数値サマリに目は通すが、特に記憶には残らない ▪ アクセス解析システムを使っていない ▪ ExcelやAccessは数字しか入れない

SASHELPデータ（サンプルデータ）の活用

- SAS Studioでは、デフォルトで様々なサンプルデータ（**SASHELP Data Sets**）が格納されており、分析のトレーニングなどに有効活用できる

- **左パネルより、**
  [ライブラリ]→[マイライブラリ]→[SASHELP]を選択



▼データセットの例

BASEBALL	1986年のメジャーリーガーの成績データ
BMIMEN	年齢とBMIに関するデータ
BMT	骨髄移植患者の生存期間データ
BIRTHWGT	2003年の乳児死亡率に関するデータ
FAILURE	機械の不具合に関するデータ
DEMOGRAPHICS	各国の人口などに関するデータ
JUNKMAIL	迷惑メールデータ
ORSALES	売上に関するデータ

データセットの一覧と詳細は、下記リンクを参照（英文）
https://support.sas.com/documentation/tools/sashelpug.pdf

オープンデータの候補リスト (2022年4月現在)

★…オススメもしくは、よく使われている

カテゴリ	#	サイト名	概要	ﾃﾞｰﾀ数	ﾏｰｹ	製造	医療	URL
政府系	★1	政府統計e-stat	各府省の公表データを1つにまとめたサイト。 また、これを分析-readyな形式に加工して扱いやすくした、「教育用標準データセット (SSDSE)」というサイトもある。	約700件	●	● ※統計調査のみ	●	https://www.e-stat.go.jp/ ▼教育用標準データセットSSDSE https://www.nstac.go.jp/use/literacy/ssdse/
	2	データカタログサイト	二次利用が可能な公共データの横断的検索が可能なデータカタログサイト。ただし、PDFやHTMLなどの未整形なデータがほとんど。	約2.5万件	●		●	https://www.data.go.jp/
	3	観光統計データ	日本政府観光局が運営する日本の観光統計データ。	不明	●			https://statistics.jnto.go.jp/
大学系	★4	UCI Machine Learning Repository	米カルフォルニア大学アーバイン校による公開データセットで、非常に有名で、利用者が多い。	約600件	●	●	●	http://archive.ics.uci.edu/ml/index.php
	5	Harvard Dataverse	米ハーバード大学による公開データセット。主に論文で公開された、様々な分野のデータが揃っている。	約15万件	●	●	●	https://dataverse.harvard.edu/
民間系	6	AWS パブリックデータセット	米Amazon Web Service社による公開データセット。画像、ゲノム、テキストなど非構造化データが多い。	約300件	●		●	https://registry.opendata.aws/
	7	Tableau Public Sample Data Set	米Tableau社による公開データセット。 アメリカ国内の公共系データが多い。	約30件	●		●	https://public.tableau.com/s/resources
	8	Google Dataset Search	米Google社が提供している、データセットの検索エンジン。一部、日本語でも検索可能。 ただし独自データではなく、単なるWeb上の寄せ集め。	不明	●	●	●	https://datasetsearch.research.google.com/ ▼使い方を解説しているサイト https://atmarkit.itmedia.co.jp/ait/articles/2007/15/news021.html
	9	Microsoft Research Open Data	米Microsoft社が提供しているデータセットで、同社の研究部門が研究に用いたデータを公開。 テキスト、画像系などの非構造化データが多い。	約100件	●		●	https://msropendata.com/
	10	Yahoo Webscope Datasets	米Yahoo社が提供しているデータセットで、同社のサイトなどで収集されたマーケティングデータが中心。 **※ただし、非営利団体の研究目的でのみ利用可能**	約70件	●			https://webscope.sandbox.yahoo.com/
	11	日経平均プロファイル	日本経済新聞が公開するデータセットで、日経平均、日経アジア指数などが利用可能。	約30件	●			https://indexes.nikkei.co.jp/nkave/index?type=download

オープンデータの候補リスト (2022年4月現在)

★…オススメもしくは、よく使われている

カテゴリ	#	サイト名	概要	ﾃﾞｰﾀ数	ﾏｰｹ	製造	医療	URL
分析コンペサイト	★12	Kaggle	言わずと知れた、米国の分析コンペティションサイト。コンペ用の様々なデータセットが公開されており、中には民間企業から提供受けたリアルなデータもある。（コンペ終了後、非公開にされているデータもあり）分析コードも公開されているため、自学習におすすめ。	不明	●	●	●	https://www.kaggle.com/
	13	SIGNATE	日本版Kaggleともいうべき、日本の分析コンペティションサイト。企業から提供を受けたリアルなデータが大半。**※ただし、基本的にはコンペ目的以外での利用を禁止しているため、要注意。**	不明	●	●	●	https://signate.jp
学術研究向け	14	NDBオープンデータ	厚労省が提供する、匿名化したレセプト情報・特定健診等情報に関するデータセット。	約300件			●	https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000177182.html
	15	国立情報学研究データリポジトリ	国立情報学研究所（NII）が公開するデータセット。民間企業や大学等から提供を受けたリアルデータ（楽天の購買データ、アットホームの不動産データ、Yahoo!知恵袋データ、など）が公開されている。**※ただし、非営利団体の研究目的でのみ利用可能**	約20件	●		●	https://www.nii.ac.jp/dsc/idr/datalist.html

その他（付属データ/個人サイト等）	★16	Python scikit-learn内のデータセット	Pythonのscikit-learnライブラリに付属しているデータセット。ボストンの住宅価格や、アヤメ品種、糖尿病患者や乳がんのデータなど。
			▼参考：Pythonでの読込み例
			`from sklearn.datasets import load_boston`
			`boston = load_boston()`
			`df = pd.DataFrame(boston.data,columns=boston.feature_names)`
				14件	●		●	▼わかりやすくまとめているサイト
								https://zenn.dev/nekoallergy/articles/scikit-learn-datasets

	17	松原望先生（東大名誉教授）の個人サイト	松原先生が公開する様々なデータ（Webから収集？）。少し古いデータが多いものの、Excelによる分析手法とともに掲載されているので、自学習用に向いている。	約90件	●		●	https://www.bayesco.org/top/datasite
	18	データで学ぶ!統計活用授業のための教材サイト	統計教育推進委員会が公開する様々なデータ。やや古めで、項目数もそれほど多くないため、手元での簡単な分析学習に使うイメージ。	26件	●		●	https://estat.sci.kagoshima-u.ac.jp/data/

まとめ

- 主成分分析とクラスター分析の組み合わせ
  - **主成分分析を行うことで、多数の説明変数を要約**することができた
  - 主成分分析結果に対してクラスター分析を適用することで、別観点の知見を得ることができた

- 階層的クラスタリングによるデータ分類
  - **階層的クラスタリング（群平均法、重心法、Ward法）のしくみ**について学習した
  - 各手法を都道府県データに適用し、**類似の都道府県をグルーピング**することができた
  - デンドログラムを観察することで、**最適なクラスタ数を検討**することができた

- 今後のデータサイエンス学習に向けて
  - 実践的なスキルを鍛えるには、座学だけでなく、実際のデータを触ってみることが一番
  - 様々なオープンデータを活用して、スキルアップを目指す
  - データサイエンスの知識だけでなく、「**ビジネス力**」「**IT力**」も極めて重要

アンケートのお願い・ご質問

10月26日 機械学習によるビッグデータ分析の手法-3

今後の参考にさせていただくため、ぜひともアンケートにご協力をお願いします。

・無記名　・所要時間目安: １～３分

アンケートURL

https://sas.qualtrics.com/jfe/form/SV_8pqzZmIQg9SpJ42

・本日のアーカイブは、2022年10月31日～2023年3月31日迄視聴できます。

・本日の内容に関するご質問は、以下宛にご連絡ください。

que@datascience.co.jp

ご視聴ありがとうございました。

End of File