実務直結! 分析カ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10月19日



Copyright © SAS Institute Inc. All rights reserved.

### 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





#### 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





# 教師あり学習と教師なし学習



教師あり学習



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





Copyright © SAS Institute Inc. All rights reserved.

クラスタリング手法の種類



U

- ・ 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	SAS <sup>®</sup> Studio
変 数 B	■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	<ul> <li>・ウォード法</li> </ul>	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 <sup>®</sup> 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
数 B	■ 最長距離法(最遠隣法)	距離の遠いデータから順番に併合	(10/26)
	■重心法	クラスタ重心からの距離に基づき併合	SAS <sup>®</sup> Studio
<ul> <li></li></ul>	■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使	样合 SAS <sup>®</sup> Studio
(dendrogram)	<ul> <li>■その他</li> </ul>	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: <a href="https://www.youtube.com/watch?v=BVFG7fd1H30">https://www.youtube.com/watch?v=BVFG7fd1H30</a>



# 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



Copyright © SAS Institute Inc. All rights reserved.



#### 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



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	<ul> <li>モデリング手法の選択</li> <li>モデルの作成</li> <li>モデルの評価</li> </ul>
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

10



#### 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	ultivariate Number of Instances: 452		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:

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]
 bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
 bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
 bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
 bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

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







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、

			クレジット 債務不履行	カード テの有無	年間平	均残高 -□)			最終連約 会話時間	各時の (秒)	キャンペーン 連絡回	·中の 最終 数 糸	≷連絡からの ≩過日数	キャンペーン前の 前 直絡回数	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	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	P		23	feb	141	2	176	3 failure	
36	self-employed	married	tertiary	no	307	yes	≣₩ B	日変数	14	may	341	1	330	2 other	目的変数
39	technician	married	secondary	no	147	yes	0/0-3	JEESA	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	8 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
					101				20		400	2			
56	i technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0 unknov	5、1119 44 #
37	admin.	single	tertiary	no	2317	yer	則(分	物》对家	20	apr	114	1	152	2 failure	
25	5 blue-collar	single	primary	no	-221	説日	lt る	ための変	数 23	may	250	1	-1	0 unknow	ったい 対象
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1 other	no
38	8 management	divorced	unknown	no	0	yes	no	cellular	18	nov	96	2	-1	0 unknow	
42	2 management	divorced	tertiary	no	16	no	no	cellular	19	nov	140	3	-1	0 unknow	<b>NDd</b>
44	services	single	secondary	no	106	C <b>∂<sub>l</sub>9</b> /righ	t 🗗 🛛 AS I	ns <b>unkกดเพท</b> .แ r	ights re <b>1</b> 2	jun	109	2	-1	0 unknown	no
4.4		the second second			0.2				-	4.1	105	2	-	0	

4,521人



- 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

▼GUI画面への変更方法 (ビジュアルプログラマ)





### 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





### 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

ファイルのアップロード	
ファイルのアップロード先: /home/u62013505	
ファイルの選択	
選択済みファイル:	
1 XLSX bank_marketing.xlsx	371.1 kb
	アップロード キャンセル

#### 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

SAS <sup>®</sup> Studio	
<ul> <li>サーバーファイルとフォルダ</li> </ul>	
は→ 竜 玉 平 圃 55	
⊿ 🛱 odaws02-apse1-2	×
🔁 フォルダショートカット	1
🔺 📮 ファイル (ホーム)	
sasuser.v94	
🔀 bank_marketing.xlsx	





#### 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



#### 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)



#### 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



#### ④「結果」のタブ画面に読み込んだデータの概要が出力

ファイル名: bank_ma	arketing.xlsx			
ソースの場所: /home/ut	52013505			
ワークシート名:				
第1ワークシート				
W1 2 2 2 1				
7 1 7 7 7	***	= <i>h</i>		
1-k 12	商業 17月	5-9		
6 P P 🗄 🖶 🖂	· 22			
<ul> <li>目次</li> </ul>				
		V		
		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	インデックス数	0
	作成日時	2022/08/08 09:34:47	オブザベーションのバッファ長	120
	更新日時	2022/08/08 09:34:47	削除済みオブザベーション数	0
	保護		圧縮済み	NO
	データセットタイプ		ソート済み	NO
	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)		
		THE ADDRESS OF THE ADDRESS		
データセットのページ	191079	エンシンバスト国連回報		
データセットのページ	5			
	1			
エーダベージの先調				
テーダページの先頭 ページごとの最大OBt	D1X 1000			





### 読み込んだデータの確認

#### データ概要の確認

新し		ブラウ	リザタブ	<u>ç</u> f	る				
	□ □	2 B	A 8						
						CONT	ENTS プロシミ	ĴŦ	1
			データセット名 V	VORK.IMPC	ORT1			オブザベーション数	4521
					CON	TENTS 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	ータイプ	DATA					変数の数	17
	エンジ	<u>ب</u>	V9					127777	
	作成日	诗	2022/08/08 09:34:4	7					
	更新日	痔	2022/08/08 09:34:4	7				УЈ БХС1ТБХ	で唯記!
	保護	Hw L A /-						(L'und" A/	
	7-9	275917						(Lツクテータケ	「竹り屋本」
	ラヘル	±19	SOLADIS VIS 64		V0C CA			IV IACA	
	7-7	- K	Utf & Unicode (UTE		100_04	, ALFHA_I	H004, LINO	1/1/1/14	
			and official (011-	~/					
					エンシ	ノン/ホスト	司連情報		
データセットのペ	ページサイズ	131072							
データセットのペ	ページ数	5							
データページの先	頭	1							
ページごとの最大	COBS数	1090							
先頭ページのOBS	S数	1052							
データセットの修	復数	0							
ファイル名		/saswork/S	AS_work71F80001F3F	A_odaws	01-aps	e1-2.oda.sa	as.com/SAS	_workC7860001F3FA_odaws01-aps	e1-2.oda.sas.com/import1.sas7bdat
作成したリリース	L .	9.0401M6							
作成したホスト		Linux							
ノノード番号		33850							
アクセス権限		rw-rr							
所有者名		U62013505							
ファイルサイズム	15 × 17		<u> </u>	<b>#11</b> 7	7	tr=x	-		
77 IN IN IN	~ <u>`</u>	עניצ,	ワーク	$\pm 7$	τľ	目記			
				変数	と属性!	リスト (アル	ファベット	順)	
		# 3	定数	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	数值	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	数值	8	BEST.		キャンペーン前の連絡回数	
		5 :	ウ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 {	主宅ローンの有無	文字	3	\$3.	\$3.	住宅ローンの有無	
		8 f	個人ローンの有無	文字	3	\$3.	\$3.	個人ローンの有無	
		16 1	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 4	学歴	文字	9	\$9.	\$9.	学歴	
		17 5	2期預金契約有無	文字	3	\$3.	\$3.	定期預金契約有無	
		6 4	F回半均残局(ユーロ	奴他	8	BEST.		平岡平均残局(ユーロ)	
		1 4	+郡	致1直 数/庙	8	BESI.		牛部	
		14 1	RRNE船からの胚週日 島線道絃口	<u>奴</u> [[] 数值	б Р	BEST		取た連絡からの枢辺口奴 最終連絡口	
		12 4	8475,428日日 最終連絡時の会話時間	数值	о Р	BEST.		AX7 <aeatu 最終連絡時の会話時間(秒)</aeatu 	
		14 1		が開	3	\$3	60	BATS ABOTO PU OF TA BORD PU (127)	
		11 4	曾怒運怒日	× -+- •			L 1997	16於14約日	
		11 J	最終連絡月 吉婚歴	文字	8	\$8.	\$3. \$8.	取於連結月 結婚歷	
		11 J 3 A 2 I	長終連絡月 吉婚歴 載業	文子 文字 文字	8 13	\$8. \$13.	\$8. \$13.	最終運輸月 結婚歷 職業	

					00	🖻 SAS プログラ	र 🗧 🖨 🕐 मनः	シアウト
ログラム1 × <b>ス</b> *bank_marketin	g ×						dii	
	10. 25							<u>*:</u> I
イル情報								
スノアイル - イリター bank marketing view								
-スの場所: /home/u62013505								
-クシート名:								
1 ワークシート								
F-y								
-タセット名: IMPORT1								
プラリ: WORK								
<u>ب</u>	セカデー	_力 ) 面	iran ⊢/	า				
			іше.	<u>л</u>				
イルの種類:	D17 (1	ビ生の	デース	ったん	在言刃			
	Xリノン/い	こエッ	ノーン	ግሮዝ	王可心			
オルト(ノアイル拡張士に基づく)	•							
ノオルト(ノアイル拡張士に基づく?								
リード ログ 結果	▲ 当力データ				-1			
ード ログ 結果 -ブル: WORK.IMPORT1 ▼   1	 出力データ ビュー: 列名 ▼ 配	≞ 0 ≣	<b>マ</b> フィルタ: (1	<b>なし</b> )	-			
マオルド(ファイル監要子に乗うく) ード ログ 結果 ブル: WORK.IMPORT1 ▼   1	<ul> <li>□</li> <li>□<td>- <b>묘 (5 団</b>) 승당列数:17</td><td><b>♀</b>フィルタ: (≀</td><td>なし)</td><td>- 1 1 1</td><td></td><td><b>抽 条</b> 符1100 1</td><td></td></li></ul>	- <b>묘 (5 団</b> ) 승당列数:17	<b>♀</b> フィルタ: (≀	なし)	- 1 1 1		<b>抽 条</b> 符1100 1	
<ul> <li>マオルト(ブアイル血液子に盛りく)</li> <li>ード ログ 結果</li> <li>ブル: WORKJMPORT1 ~ 1</li> <li>すべて選択</li> </ul>	▲ 出力データ ビュー: 列名 → ■ ③ 合計行数:4521		♥ フィルタ: (≀ 職業	なし) 結婚歴	- 1 ↓ 学歴	クレジット	<u>)</u> 年間平均残高(ユーロ	• •
<ul> <li>マネルト(ファイル Might Line DC)</li> <li>ード ログ 結果</li> <li>ブル: WORK IMPORT1 *</li> <li>すべて 選択</li> <li>③ 年齢</li> </ul>	▲ 出力データ ビュー: 列名 → ■ ③ 合計行数: 4521 1	· <b>문 · 5 国</b> 승計列数· 17 年齢 30	♥ フィルタ: (* 職業 unemployed	なし) 結婚歴 married	・ 学歴 primary	<b>クレジット</b> no	た。た。行き1000日 年間平均残高(ユーロ 1787	◆ → 住宅⊑ no
<ul> <li>マオルド(ジアイル血液子に盛りく)</li> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 ▼</li> <li>すべて選択</li> <li>● 年齢</li> <li>▲ 職業</li> </ul>	■カデータ	▲ 5 国 合計列数·17 年齢 30 33	♥ フィルタ: (? 職業 unemployed services	なし) 結婚歴 married married	学歴 primary secondary	<b>クレジット</b> no no	た 年間平均残高 (ユーロ 1787 4789	● ● 住宅口 no yes
<ul> <li>マオルド(ジアイル処装すに盛りく)</li> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 ▼</li> <li>すべて選択</li> <li>● 年齢</li> <li>▲ 職業</li> <li>▲ 結婚歴</li> </ul>	■出力データ ビュー: 列名 ∨ □ ビュー: 列名 ∨ □ 1 2 3	<ul> <li>・</li> <li>・</li></ul>	♥フィルタ: () 職業 unemployed services management	なし) 結婚歴 married married single	学歴 primary secondary tertiary	<b>クレジット</b> no no no	た た た 100 J 年間平均残高 (ユーロ 1787 4789 1350	住宅口 no yes yes
<ul> <li>マオルド(ジアイル鉱装子に盛りく)</li> <li>ード ログ 結果</li> <li>ブル: WORK.IMPORT1 ▼</li> <li>すべて選択</li> <li>② 年齢</li> <li>▲ 親端歴</li> <li>▲ 結婚歴</li> <li>▲ 学歴</li> </ul>	出力データ     ビュー: 列名 ▼ □     □	ユ い 国 合計列数・17 年齢 30 33 35 30     33     35     30     3     3     3     5     30     3     3     3     5     3	♥ フィルタ: (? 職業 unemployed services management management	なし) 結婚歴 married married single married	学歴 primary secondary tertiary tertiary	<b>クレジット</b> no no no no	<b>注 名 日 100 1</b> 年間平均残高(ユーロ 1787 4789 1350 1476	住宅に no yes yes yes
<ul> <li>マオルド(ファイル血蛋子に盛りく)</li> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 -</li> <li>すべて選択</li> <li>② 年齢</li> <li>▲ 職業</li> <li>▲ 精緻歴</li> <li>▲ 学歴</li> <li>▲ クレジットカード債務</li> </ul>	★計算数:4521 ○ 本計算数:4521 1 2 3 4 5	<ul> <li>・</li> <li>・</li></ul>	♥ フィルタ: () 職業 unemployed services management management blue-collar	結婚歴 married married single married married	学歴 primary secondary tertiary tertiary secondary	<b>クレジット</b> no	<del>体 個 日 1000 [ </del>	住宅に no yes yes yes yes
<ul> <li>マオルド(ファイル鉱会子に盛っく)</li> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 ~</li> <li>すべて選択</li> <li>③ 年齢</li> <li>▲ 職業</li> <li>▲ 結婚歴</li> <li>▲ クレジットカード債務</li> <li>④ 年間平均残高(ユーロ</li> </ul>	<ul> <li>出力データ</li> <li>● ● ● ●</li> <li>ビュー: 列名 ●</li> <li>●</li> <li>●&lt;</li></ul>	<ul> <li>▲ 5 目</li> <li>合計列時·17</li> <li>年齢</li> <li>30</li> <li>33</li> <li>35</li> <li>30</li> <li>59</li> <li>35</li> </ul>	マフィルタ: () 職業 unemployed services management blue-collar management	結婚歴 married married single married single	学歴 primary secondary tertiary tertiary secondary tertiary	<b>クレジット</b> no 1000000000000000000000000000000000000	体 を E + 100 - 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	住宅口 no yes yes yes yes no
<ul> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 ~</li> <li>すべて選択</li> <li>③ 年齢</li> <li>▲ 職業</li> <li>▲ 結婚歴</li> <li>▲ クレジットカード債務</li> <li>③ 年間平均残高(ユーロ 人 住宅ローンの有無</li> </ul>	<ul> <li>出力データ</li> <li>→</li> <li>ジョー: 列名 →</li> <li>○</li> <li>○<td><ul> <li>▲ 5 目</li> <li>合計列数·17</li> <li>年齢</li> <li>30</li> <li>33</li> <li>35</li> <li>30</li> <li>59</li> <li>35</li> <li>36</li> <li>20</li> </ul></td><td>マフィルタ: (? 職業 unemployed services management blue-collar management self-employed</td><td>なし) 結婚歴 married married married single married married</td><td>学歴 primary secondary tertiary tertiary tertiary tertiary</td><td>クレジット no no no no no no no no  </td><td><del>体 を 日1000</del> 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 307 317</td><td>住宅に no yes yes yes no yes</td></li></ul>	<ul> <li>▲ 5 目</li> <li>合計列数·17</li> <li>年齢</li> <li>30</li> <li>33</li> <li>35</li> <li>30</li> <li>59</li> <li>35</li> <li>36</li> <li>20</li> </ul>	マフィルタ: (? 職業 unemployed services management blue-collar management self-employed	なし) 結婚歴 married married married single married married	学歴 primary secondary tertiary tertiary tertiary tertiary	クレジット no no no no no no no no  	<del>体 を 日1000</del> 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 307 317	住宅に no yes yes yes no yes
<ul> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 ●</li> <li>すべて選択</li> <li>③ 年齢</li> <li>▲ 職業</li> <li>▲ 結婚歴</li> <li>学歴</li> <li>△ クレジットカード債務</li> <li>③ 年間平均残高(ユーロ</li> <li>▲ 信和ローンの有無</li> <li>▲ 個人ローンの有無</li> </ul>	■出力データ	4 5 日 合計列数: 17 年齢 30 33 35 30 30 30 30 30 30 30 30 30 30 30 30 30	♥フィルタ: (2 職業 unemployed services management blue-collar management self-employed technician	結婚歴 married married single married single married married married	学歴 primary secondary tertiary tertiary tertiary tertiary secondary	クレジット no 100 100 100 100 100 100 100 100 100 10	<b>体 を 日 100</b> 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 307 147 307 202	住宅口 no yes yes yes no yes yes
<ul> <li>ード ログ 結果</li> <li>ブル: WORK.IMPORT1 ●</li> <li>すべて選択</li> <li>③ 年齢</li> <li>▲ 職業</li> <li>▲ 結婚歴</li> <li>◆ 学歴</li> <li>▲ クレジットカード債務</li> <li>③ 年間平均残高(ユーロ</li> <li>▲ 信和一シの有無</li> <li>▲ 調約ローンの有無</li> <li>▲ 連絡手段</li> </ul>	出力データ →     ビュー: 列名 → □ 日 → □	会計列数 17 年齢 30 33 35 30 599 35 36 39 39 41	♥フィルタ: () 職業 unemployed services management management blue-collar management self-employed technician entrepreneur considen	結婚歴 married married single married married married married married		クレジット   ハロ   ハロ   ハロ   ハロ   ハロ   ハロ   ハロ   ハロ	年 を ま 100 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 307 747 307 147 6	住宅に no yes yes yes no yes yes yes yes
<ul> <li>マオルド(ファイル鉱会主に盛っく)</li> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 *</li> <li>すべて選択</li> <li>● 年齢</li> <li>▲ 職業</li> <li>▲ 結婚歴</li> <li>◇ 学歴</li> <li>▲ クレジットカード債務</li> <li>● 年間平均残高(ユーロ</li> <li>▲ 住宅ローンの有無</li> <li>▲ 通知手段</li> <li>● 四ちょ</li> </ul>	* 出力データ ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●	二 上 い 日 会計別称・17 年齢 30 33 35 30 59 35 36 39 41 43 900	■ マフィルタ: () 職業 unemployed services management blue-collar management self-employed technician entrepreneur services services	結婚歴 married married single married married married married married	学歴 primary secondary tertiary tertiary tertiary tertiary tertiary tertiary secondary tertiary	クレジット 内の れの れの れの れの れの れの れの れの れの れ	注 を 日 1000 「 年間平均残高 (ユーロ 1787 4789 1350 1476 0 0 747 307 147 47 221 -88	defection no yes yes yes yes yes yes yes
<ul> <li>マオルド(フアイル鉱装于に盛くく)</li> <li>ード ログ 結果</li> <li>ブル: WORKIMPORT1 ~</li> <li>すべて選択</li> <li>② 年齢</li> <li>▲ 韓歴</li> <li>▲ 今班歴</li> <li>▲ 今班歴</li> <li>▲ 今レジットカード債務</li> <li>③ 年間平均残高(二一口</li> <li>▲ 住宅ローンの有無</li> <li>▲ 個人ローンの有無</li> <li>▲ 連絡手段</li> <li>コパティ</li> <li>値</li> </ul>	<ul> <li>● 出力データ ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●</li></ul>	二 上 い 日 二 二 二 二 二 二 二 二 二 二 二 二 二	マフィルタ: () 職業 unemployed services management management bluecollar management self-employed technican entrepreneur services services services	なし) 結婚歴 married single married married married married married married married	学歴 primary secondary tertiary tertiary tertiary tertiary tertiary tertiary primary secondary secondary	クレジット つの 10 つの	年間平均残高(ユーロ 1787 1789 1350 1476 0 747 0 747 1476 1476 147 147 221 147 221 -88 9374	teren no yes yes yes yes yes yes yes
<ul> <li>マオルド(フテイル鉱装于に盛っく)</li> <li>ード ログ 結果</li> <li>・ブル: WORK.IMPORT1 ~</li> <li>すべて選択</li> <li>② 年齢</li> <li>▲ 聴歴</li> <li>▲ 特趣歴</li> <li>▲ クレジットカード債務</li> <li>③ 年間平均残高(ユーロ</li> <li>▲ 住宅ローンの有無</li> <li>▲ 個人ローンの有無</li> <li>▲ 連絡手段</li> <li>コパティ 値</li> <li>パレー</li> </ul>	★出力データ ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●	<ul> <li>二</li> <li>公計判許・17</li> <li>年齢</li> <li>30</li> <li>33</li> <li>355</li> <li>30</li> <li>59</li> <li>355</li> <li>36</li> <li>36</li> <li>39</li> <li>41</li> <li>43</li> <li>39</li> <li>43</li> <li>22</li> </ul>	マフィルタ: ()	なし) 結婚歴 married single married single married married married married married married	学歴 primary secondary tertiary tertiary tertiary tertiary secondary tertiary primary secondary secondary	クレジット つの 10 つの	体 を (王・100 ) 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 307 747 307 147 307 147 221 48 88 9374 224	住宅に no yes yes yes yes yes yes yes
<ul> <li>レストレンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーンシストレーシンシン、シーシンシンシンシンシンシンシンシンシンシンシンシンシンシンシンシン</li></ul>	当力データ → → → → → → → → → → → → → → → → → → →	<ul> <li>二 4、 10</li> <li>(1)</li> <li>(1)&lt;</li></ul>	♥フィルタ: () 職業 unemployed services management blue-collar management self-employed technician entrepreneur services services admin. technician technician	結婚歴 married married single married single married married married married married married married	PE           primary           secondary           tertiary           tertiary           tertiary           tertiary           tertiary           tertiary           tertiary           secondary           tertiary           secondary           tertiary           secondary           secondary           tertiary           secondary           tertiary	クレジット つの の つの の の の つの の の の つの の の の の つの の の の の の の つの の の の の の の の の の の の の の の の の の の	使 を E 100 日 年間平均残高 (ユーロ 1787 4789 1350 1476 0 747 307 147 307 147 221 221 221 221 221 224 224 224	住宅に no yes yes yes yes yes yes yes yes yes no
レド         ログ         結果           ・ブル:         WORK.IMPORT1 ◆         1           すべて選択         (************************************	<ul> <li>出力データ</li> <li>● ● ● ●</li> <li>ビュー: 列名 ●</li> <li>● ● ● ●</li> <li>● ● ● ●</li> <li>● ● ● ●</li> <li>● ● ● ●</li> <li>● ● ●</li> <li>● ●</li> <li>●</li>     &lt;</ul>	全計2時 17 年齢 300 333 355 300 599 355 366 399 411 433 399 413 399 411 433 309 433 309 433 309 433 309 433 309 433 309 433 309 433 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 435 309 437 438 439 439 439 439 439 439 439 439	♥フィルク: () ■## Winemployed services management blue-collar management self-employed technician entrepreneur services admin. technician student	結婚歴 married single married married married married married married married married married single		クレジット   ハの ハの 、 、 、 、 、 、 、 、 、 、 、 、 、 、 、 、 、 、 、	生 を 日 1000 ゴ 年間平均残高(ユーロ 1787 1350 1476 0 1476 0 1476 307 147 147 147 147 147 147 147 14	住宅に no yes yes yes yes yes yes yes yes yes no no

生データの確認





# 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

- <u>#1_ロジスティック回帰</u> > "bank_marketing.xlsx" のイン
設定 コード/結果 分割 🏒 😡 🚼
オプション  ノード
▼ファイル情報
ソース ファイル
ファイル名: bank_marketing.xlsx
ソースの場所: <b>/home/u62013505</b>
ワークシート名:

#### 「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





### 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







- 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	右上部分とは対称	左下部分 <sup>435</sup> 応	-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							

「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





#### 相関分析の出力 – 実行結果 (散布図行列)





# (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



	全行りの女性別の目生														
	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								

甘大的北西约纮斗旦

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





28 S.Sas



Copyright © SAS Institute Inc. All rights reserved.





Korrelation: 0.74, 0.82, 0.75, 0.72, 0.69

Copyright © SAS Institute Inc. All rights reserved.



# 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



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	約右は、 「最終す	にまたシ	べっぷれ	の連絡に	J叙小を、 い相関	1.00000			
	年齢 年間平均残高 年間平均残高 最終 最終 会話時間 最終会話時間 最終会話時間 CP中連絡回数 CP中連絡回数 CP前連絡回数 CP前連絡回数	年齢           年齢         1.00000           年齢         1.00000           年間平均残高         0.16845           年間平均残高         0.05207           最終連絡日         -0.03633           最終会話時間         -0.06583           CP中連絡回数         -0.052072           最終連絡日数         -0.05503           星終連絡日数         -0.05503           CP中連絡回数         -0.05072           最終連絡日数         -0.011	年齢         年間平均残高           年齢         1.00000         0.16845           年齢         0.16845         1.00000           年間平均残高         0.16845         1.00000           年間平均残高         0.16845         1.00000           最終連絡日         -0.05207         -0.03858           最終全話時間         -0.03633         -0.12007           最終会話時間         -0.06583         -0.02804           CP中連絡回数         -0.05072         0.01352           最終連絡日数         -0.05072         0.01352           日本経連絡回数         -0.0115         2           日本経連絡回数         -0.0115         2           日本経連絡回数         -0.0115         2           日本経連絡回数         -0.0115         2	年齢         年間平均残高         最終連絡日           年齢         1.00000         0.16845         -0.05207           年齢         1.00000         0.16845         -0.05207           年間平均残高         0.16845         1.00000         -0.03858           最終連絡日         -0.05207         -0.03858         1.00000           最終連絡日         -0.05207         -0.03858         -0.03610           日本         -0.06583         -0.02804         0.13780           CP中連絡回数         -0.05072         0.01352         -0.03734           最終連絡日数         -0.01152         -0.03734           日本         -0.01152         -0.03734           日本         -0.01152         -0.03734	年齢         年間平均残高         最終連絡日         最終会話時間           年齢         1.00000         0.16845         -0.05207         -0.03633           年間平均残高 年間平均残高 年間平均残高         0.16845         1.00000         -0.03858         -0.12007           最終連絡日 最終連絡日         0.05207         -0.03858         1.00000         0.03610         0.03610           最終連絡日 最終連絡日         0.05207         -0.02804         0.13780         0.23432           日本経連絡日数 最終連絡日数 最終連絡日数 最終連絡日数 CP前連絡回数 CP前連絡回数         0.05072         0.01352         -0.03734         -0.15489           日本経連絡日数 最終連絡日数 最終連絡日数         -0.011:         契約者は、「たたや>ペーンシー中           日本経連絡日数 最終連絡日数         -0.011:         契約者は、「たたや>ペーンシー中	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           年齢         1.00000         0.16845         -0.05207         -0.03633         -0.06583           年齢         1.00000         0.16845         -0.05207         -0.03633         -0.06583           年間平均残高 年間平均残高 年間平均残高         0.16845         1.00000         -0.03858         -0.12007         -0.02804           最終連絡日 最終連絡日         -0.05207         -0.03858         1.00000         0.03610         0.13780           最終会話時間 最終会話時間         -0.05633         -0.12007         0.03610         1.00000         0.23432           CP中連絡回数 長終連絡日数 最終連絡日数         -0.06583         -0.02804         0.13780         0.23432         1.00000           最終連絡日数 長終連絡日数 CP前連絡回数 CP前連絡回数         -0.05072         0.01352         -0.03734         -0.15489         -0.08488           CP前連絡回数 CP前連絡回数         -0.0111         契約者は、「またたンペーンシャの連絡回         -0.0548         -0.0548         -0.0548	年齢         年間平均残高         最終連絡日         最終金話時間         CP中連絡回数         最終連絡日数           年齢         1.0000         0.16845         -0.05207         -0.03633         -0.06583         0.05072           年齢         1.0000         0.16845         -0.05207         -0.03633         -0.02804         0.01352           年間平均残高 年間平均残高 年間平均残高         0.16845         1.00000         -0.03858         -0.12007         -0.02804         0.01352           最終連絡日 最終連絡日         -0.05207         -0.03858         1.00000         0.03610         0.13780         -0.03734           最終主絡日 最終主路日 最終主給日数         -0.06583         -0.12007         0.03610         1.00000         0.23432         -0.15489           CP中連絡回数 長終連絡日数         -0.05072         -0.03734         -0.15489         -0.08488         1.00000           最終連絡日数 最終連絡日数         -0.05072         -0.01152         -0.03734         -0.15489         -0.08488         1.00000           CP前連絡回数 CP前連絡回数         -0.0111         契約者は、「もたた」         -0.0548         1.0000         -0.08488         1.00000			

				Pearson の	相関係数, N = 4	000		
		年齡	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	(参考)相関	係数の目安
未契約者	年間平均残高 年間平均残高	0.07291	1.00000	-0.00539	-0.00858	-0.00762	相関係数相	関の強弱
	最終連絡日 最終連絡日	-0.01165	-0.00539	1.00000	-0.03679	0.16339	-1.0 5	亀い相関
	最終会話時間 最終会話時間	-0.01836	-0.00858	-0.03679	1.00000	-0.09611	191 	や相関あり
	CP中連絡回数 CP中連絡回数	0.00446	-0.00762	0.16339	-0.09611	1.00000	-0.2	弱い相関
	最終連絡日数 最終連絡日数	-0.02733	0.00694	-0.10334	0.00179	-0.08961	اعا 0	54111111111111111111111111111111111111
	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





### 散布図行列 (層別) との比較



未契約者 Pearson @擱懸 N=4000

-0.01165

-0.00539

1.00000

-0.03679

0.16339

-0.10334

-0.05957

最終連絡日 最終会話時間 CP中連絡回数

-0.01836

-0.00858

-0.03679

1.00000

-0.09611

0.00179

0.00563

0.00446

-0.00762

0.16339

-0.09611

1.00000

-0.08961

-0.05856

最終連絡日数 CP前連絡回数

-0.00819

0.02507

0.00563

-0.05856

0.58368

1.00000

-0.02733

0.00694

-0.10334

0.00179

-0.08961

1.00000

0.58368

(参考)相	<b>1関</b> 係数の目安
相関係数	相関の強弱
(絶対値) <b>1.0</b>	<b>油山相関</b>
0.7	やや相関あり
0.4	弱い相関
	ほぼ相関なし
·	

31 **SSAS** 

主刀女力	
突剂	白

rearson vrip所数, 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			

Copyright © SAS Institute Inc. All rights reserved.

年齡

年齢 年間平均残高

年間平均残高 最終連絡日

最終連絡日 最終会話時間

最終会話時間 CP中連絡回数

CP中連絡回数 最終連絡日数

最終連絡日数

CP前連絡回数

CP前連絡回数

年齢

1.00000

0.07291

-0.01165

-0.01836

0.00446

-0.02733

-0.00819

年間平均残高

1.00000

-0.00539

-0.00858

-0.00762

0.00694

0.02507





#### 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。



## 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



# 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 差分 比率		
質的データ	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	<ul> <li>●性別、血液型、顧客ID</li> <li>●作業者、個品ID、</li> <li>良品/不良品</li> </ul>	(ACB)(A-B)(A/B)  ※集計によるカウントのみ可能		
(カテゴリーデータ)	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	•		
<b>量的データ</b> (数量データ)	間隔尺度	数値演算可能だが、 <b>値の差</b> 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	•年齡、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、 <b>値の比</b> 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	<ul> <li>●身長、売上金額</li> <li>●寸法、圧力、作業時間、</li> <li>絶対温度</li> </ul>	• • •		



### 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





Copyright © SAS Institute Inc. All rights reserved.



### 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







# 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

再揭



#### ▼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)



### 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 (DBScanae) 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









- 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



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



ダブルクリック





#### 非階層的クラスタリング(k-means) - S

#### [データ]の設定(説明変数・目的変数)



実行方法(2/2)説明変数・オプション
[オプション]の設定(各種出力)
設定 コード/結果 分割 🖌 😡 🔛
データ オプション 出力 情報 ノード
▼手法
▼標準化
標準化法:
範囲 (デフォルト)
最小値を引き、範囲で割ります
▼ クラスタリング
次の 2 つの手法のいずれかを指定する必要があります:
☑ 最大クラスター数 [最大クラスター数]にチェックが
*クラスター: 10 ↓ 人っていることを確認し、 [クラスター数]を10 に設定
□ 候補シードと既存シード間の最小距離
<ul> <li>各オブザベーションのクラスター重心法をアップロー</li> <li>ド</li> </ul>
🗌 データセットのクラスター重心法を読み込む
□ 最大反復回数
▼統計量
表示する統計量:
デフォルト統計量

4

各クラスタ

の標準偏差

### 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

			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	0.014285714
2	0.2784474446	0.0629810558	0.4825136612	0.0907255767	0.0240046838	0.3600212438	0.100655737
3	0.5049847212	0.0648683352	0.2758441558	0.0754145226	0.0277498012	0.0154950554	0.010753246
4	0.29960784	つ個のク	ラスタ数	C3(211839)	か結果	の解釈	2310133333
5	0.3860294118	0.6093106135	- 73万日月?	マキキシビをお	一市六七、日本	<b>告任</b> 00000000	0.000000000
6	0.3120204604	0.0617715665	人之間で	ノイオ。1王义上	」用人して	表告43381864	0.108405797
7	0.2184991446	0.05822224	フラスタ	数を減ら	.029014 == 3	0/0-387271	0.006935123
8	0.3158757436	0.0641564482	0.7179026217	0.0765807884	0.0401375831	0.0180579322	0.010337078
9	0.2979302832	0.0613663152	0.8728395062	0.0446994495	0.4799697657	0.0000000000	0.000000000
10	0.3231707317	0.0731065649	0.4471544715	0.0885750963	0.0243902439	0.2246587603	0.468292682

#### 入力した説明変数

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

<ul> <li>ノード データ オブション 出力 情報</li> <li>・手法</li> <li>・標準化 標準化法:</li> <li>範囲(デフォルト)</li> <li>・ 最小値を引き、範囲で割ります</li> <li>・ クラスタリング</li> <li>次の2つの手法のいずれかを指定する必要があります:</li> <li>・ クラスターツ</li> <li>・ 金 長大クラスター数 [最大クラスター数]にチェック</li> <li>入っていることを確認し、 [クラスター: 3]</li> <li>・ (つうスター数]を3 に設立</li> <li>○ 候補シードと既存シード間の最小距離</li> <li>○ 各オブザベーションのクラスター重心法をアップロード</li> <li>ドータセットのクラスター重心法を読み込む</li> <li>□ 最大反復回数</li> <li>・ 統計量</li> <li>表示する統計量:</li> </ul>	設定 コード/結果 分割 🥢 🛃 🔀	
<ul> <li>・ 手法</li> <li>・ 標準化</li> <li>標準化法:</li> <li>範囲(デフォルト)</li> <li>最小値を引き、範囲で割ります</li> <li>・ クラスタリング</li> <li>次の 2 つの手法のいずれかを指定する必要があります:</li> <li>・ クラスター数</li> <li>「最大クラスター数</li> <li>「最大クラスター数</li> <li>「最大クラスター数</li> <li>「最大クラスター数</li> <li>「最大クラスター数</li> <li>「日本のののです」</li> <li>(ご会び)</li> <li>(こ会び)</li> <li>(こ会び)<th>ノード データ オプション 出力 情報</th><th></th></li></ul>	ノード データ オプション 出力 情報	
<ul> <li>- 標準化</li> <li>標準化法:</li> <li>範囲(デフォルト)</li> <li>最小値を引き、範囲で割ります</li> <li>- クラスタリング</li> <li>次の 2 つの手法のいずれかを指定する必要があります:</li> <li>シ 最大クラスター数</li> <li>(最大クラスター数</li> <li>(最大クラスター数)</li> <li>(アクラスターご)</li> <li>(アクラスターご)</li> <li>(マクラスターを)</li> <li>(ロージョンのクラスター重心法をデップロード</li> <li>データセットのクラスター重心法を読み込む</li> <li>(日本の中国)</li> <li>(日本の中国)&lt;</li></ul>	▼ 手法	
標準化法:          範囲(デフォルト)       ・         最小値を引き、範囲で割ります       ・         クラスタリング       次の2つの手法のいずれかを指定する必要があります:         ・       会長大クラスター数       [最大クラスター数][Cチェック         ・       クラスター:       3         ・       クラスター会]       を 3         ・       会社 ブザベーションのクラスター重心法を読み込む       こ設立         ・       データセットのクラスター重心法を読み込む       ・         ・       データセットのクラスター重心法を読み込む       ・         ・       新計量	▼標準化	
<ul> <li>範囲(デフォルト)</li> <li>最小値を引き、範囲で割ります</li> <li>クラスタリング</li> <li>次の2つの手法のいずれかを指定する必要があります:</li> <li>● 最大クラスター数</li> <li>「最大クラスター数</li> <li>「最大クラスター数</li> <li>「最大クラスター数</li> <li>「日本クラスター数]にチェック</li> <li>入っていることを確認し、</li> <li>[クラスター数]を3</li> <li>(こ設立)</li> <li>(こ会)</li> <li< td=""><td>標準化法:</td><td></td></li<></ul>	標準化法:	
最小値を引き、範囲で割ります - クラスタリング 次の2つの手法のいずれかを指定する必要があります: ✓ 最大クラスター数 [最大クラスター数]にチェック 、 (フラスター数] にチェック へつていることを確認し、 [クラスター数] を 3 に設立 ○ 候補シードと既存シード間の最小距離 ○ 各オブザベーションのクラスター重心法をアップロード ○ データセットのクラスター重心法を読み込む ○ 最大反復回数 - 統計量 表示する統計量:	範囲(デフォルト)	
<ul> <li>クラスタリング</li> <li>次の2つの手法のいずれかを指定する必要があります:         <ul> <li>● 最大クラスター数</li> <li>● ポクラスター:</li> <li>● 日本の中には、</li> <li></li></ul></li></ul>	最小値を引き、範囲で割ります	
次の2つの手法のいずれかを指定する必要があります: ● 最大クラスター数 [最大クラスター数]にチェック へっていることを確認し、 [クラスター: 3] ● [日大クラスター数]にチェック へっていることを確認し、 [クラスター数]を3 (に設立 ● 候補シードと既存シード間の最小距離 ● 各オブザベーションのクラスター重心法をアップロード ■ データセットのクラスター重心法を読み込む ■ 最大反復回数 ● 統計量 表示する統計量:	<ul> <li></li></ul>	
<ul> <li>● 最大クラスター数         「最大クラスター数]にチェック 入っていることを確認し、         (アラスター: 3)         ○ 候補シードと既存シード間の最小距離         ○ 各オブザベーションのクラスター重心法をアップロード         ○ データセットのクラスター重心法を読み込む         ○ 最大反復回数         ◆統計量         表示する統計量:         ○ デフォルト統計量         ○ マーム・         ○ ポーム・         ○ 第二日の第二日の第二日の第二日の第二日の第二日の第二日の第二日の第二日の第二日の</li></ul>	次の 2 つの手法のいずれかを指定する必要があります:	
*クラスター:       3       入っていることを確認し、         (クラスター数]を3       こ設気         (候補シードと既存シード間の最小距離)       各オブザベーションのクラスター重心法をアップロード         データセットのクラスター重心法を読み込む       最大反復回数         *統計量       表示する統計量:	◎ 最大クラスター数 [最大クラスター数]にチ	<b>Eック</b>
<ul> <li>【クラスター数】を3 (こ設気</li> <li>○ 候補シードと既存シード間の最小距離</li> <li>○ 各オブザベーションのクラスター重心法をアップロード</li> <li>○ データセットのクラスター重心法を読み込む</li> <li>○ 最大反復回数</li> <li>◆統計量</li> <li>表示する統計量:</li> </ul>	*クラスター: 3 入っていることを確認し、	
<ul> <li>▲ KKHP + CUKHP + HIGORAGE HE</li> <li>▲ Aオブザベーションのクラスター重心法をアップロード</li> <li>■ データセットのクラスター重心法を読み込む</li> <li>■ 最大反復回数</li> <li>◆ 統計量</li> <li>表示する統計量:</li> </ul>		設定
<ul> <li>□ オイブリバージョンのグブスター重心法をデックプロード</li> <li>□ データセットのクラスター重心法を読み込む</li> <li>□ 最大反復回数</li> <li>◆統計量</li> <li>表示する統計量:</li> </ul>		
<ul> <li>□ データセットのクラスター重心法を読み込む</li> <li>□ 最大反復回数</li> <li>◆ 統計量</li> <li>表示する統計量:</li> </ul>	「日本ノノリベーションのグノベター重心法をノックロー	
<ul> <li>□ 最大反復回数</li> <li>◆統計量</li> <li>表示する統計量:</li> </ul>	□ データセットのクラスター重心法を読み込む	
<ul> <li>◆統計量</li> <li>表示する統計量:</li> <li>デフォルト統計量</li> </ul>	□ 最大反復回数	
★航計重 表示する統計量: デフォルト統計量		
表示∮る統計重: デフォルト統計量		
	衣示9 る統訂軍: デフォルト統計量	



#### 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、

年齢が若め

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	t		フラスタ2は	C	P中連絡I	回数、最終

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	<b>/</b> ー標準偏差			
クラスター	年齡	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



Copyright © SAS Institute Inc. All rights reserved.

# (参考) クラスター番号のデータ化と追加分析 (1/3)

SAS<sup>®</sup> Studio

*#2_k-meansクラスタリンク.cpt ×				
<u>ŧ2_k-meansクラスタリング</u> > K-Means クラス	タリング			
安定 コード/結果 分割 🛃 🛃				
データ オプション 出力	] 出力の設定	定画面を開く	ログ 結野	₹
▼ 出力データセット	I	<b>6</b> P G d	k 🔒 🗗 🛪 🖂	
📝 クラスター割り当てデータセットを作成	する	▶目次		
*データセット名:			年齢	0.155
work.Fastclus_scores	参照		年間平均残高	0.040
	27/11		最終連絡日	0.2749
	<b>「てデータセッ</b> 」	トを作成する]	最終会詰時間 CP中連終回数	0.0860
* <sup>データセット</sup>			最終連絡日数	0.114
work.Fastclus_stats	参照		CP前連絡回数	0.067
work.Fastclus_stats	参照る		CP前連絡回数 OVER-ALL	0.067
work.Fastclus_stats <ul> <li>クラスター重心法データセットを作成す</li> <li>*データセット名:</li> </ul>	参照る		CP前連絡回数 OVER-ALL	0.067
work.Fastclus_stats クラスター重心法データセットを作成す  *データセット名:  work.Fastclus_seeds	参照 る 参照		CP前連絡回数 OVER-ALL	0.067
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照		CP前連絡回数 OVER-ALL	0.067 0.136
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照		CP前連絡回数 OVER-ALL	<ul> <li>0.067</li> <li>0.136</li> <li>すべて</li> <li>3.2</li> </ul>
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照		CP前連絡回数 OVER-ALL	: 0.067 0.136 すべ 3 え
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照		CP前連絡回数 OVER-ALL WARNING: 上記	: 0.067 0.136 すべ 3 2 の2値は
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照		CP前連絡回数 OVER-ALL WARNING: 上記	: 0.067 0.136 すべ: 3: の2値は
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照	2	CP前連絡回数 OVER-ALL WARNING: 上記	: 0.067 0.136 すべ: 3.3 の2値は
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照	クラスタ	CP前連絡回数 OVER-ALL WARNING: 上記	: 0.067 0.136 すべて 3 え の2値は
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照	ク ラ ス タ ー 全	CP前連絡回数 OVER-ALL WARNING:上記	: 0.067 0.136 すべつ 3 2 の2値は 最終
work.Fastclus_stats つ クラスター重心法データセットを作成す *データセット名: work.Fastclus_seeds	参照 る 参照	7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	CP前連絡回数           OVER-ALL           WARNING:上記           第           年間平均残高           062           0.0632798060	<ul> <li>0.067</li> <li>0.136</li> <li>すべて</li> <li>32</li> <li>の2値は2</li> <li>最終</li> <li>0.76586</li> </ul>



# (参考) クラスター番号のデータ化と追加分析 (2/3)

列		0	合計行数: 4521 合言	†列数: 19		<b>₩  ←</b> 4	⊤1-100 <b>→ →</b>	
 / すべ	て選択		CP前連絡回数	前回C	定期	CLUSTER	D	
- · •	年齢		0	unknown	no	1	0.233	
	漁業		0.16	failure	no	3	0.180	
	<sup>帆木</sup>		0.04	failure	no	3	0.224	
	当时		0	unknown	no	2	0.339	
	子座		0	unknown	no	2	0.29	
			0.12	failure	no	1	0.23	
23 -	中间半均残高		0.08	other	no	3	0.172	
🗹 🛕 1	住宅ローン		0	unknown	no	2	0.186	
🖉 🛕 1	固人ローン		0	unknown	no	2	0.137	
🖌 💧	連絡手段		0.08	failure	no	3	0.22	
🖌 🔞 i	最終連絡日		0	unknown	no	1	0.190	
🖊 💧	最終連絡月		0	unknown	no	2	0.212	
🖌 🔞 🖞	最終会話時間		0	unknown	no	2	0.141	
<ul> <li>20</li> </ul>	CP中連絡回数		0	unknown	yes	1	0.34	
🗸 🔞 j	最終連絡日数		0.04	failure	no	1	0.340	
/ 🔞	CP前連絡回数		0	unknown	no	1	0.166	
🗸 🚺	前回CP結果		0	unknown	no	1	0.271	
プロパテ	ィー値		0.08	failure	no	1	0.240	
ラベル			0	unknown	no	1	0.218	
前			0.04	other	no	3	0.201	
[z			0	unknown	no	1	0.218	
類			0	unknown	no	1	0.185	
力形式			0	unknown	no	2	0.07	
力形式			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	tailure	no	2	0.420	
						_	/1	





## (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

#2_k-meansクラスタリング     棒グラフ       設定     コード/結果     分割       データ     表示       情報     ノード
<ul> <li>データ</li> </ul>
WORK.FASTCLUS_SCORES
<sup>▼フィルタ:(なし)</sup> デークを変更
<ul> <li>         ・<sup>グラフの方向</sup>         (クラスターノードで出力したデータ)         <ul> <li></li></ul></li></ul>
○ 横方向
▼ 役割
*カテゴリ: (1 項目)   💼 🕂
CLUSTER
サブカテゴリ: (1 項目)
▲ 定期預金契約
▼ オプション:
グループ化バーの表示:
● クラスター (横方向)
○ 積み上げ
凡例の場所: 外側 (デフォルト) 🗸
メジャー: 度数カウント (デフォルト) 🔹
▶追加役割







### 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



Copyright © SAS Institute Inc. All rights reserved.



#### 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

			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 クラス	近い 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	終会話時間	CP中連絡回数	最終連絡日数	CP前連絡回数		
1	0.25	23959022	0.0835507389	0.72453	318352	0.25	529710261	0.0815828041	0.0207341754	0.0216693419		
2	0.29	67011129	0.0988696446	0.18738	373874	0.16	611507455	0.0420094007	0.1081081081	0.1332046332		
		74208145	0 1121606234	0.50188	303419	0.15	566901639	0.0408026756	0.1301544832	0.0871794872		
3	<sup>0.46</sup> クラ 年齢	i スタ3は 層が高	は め				$\overline{\ }$	クラスタ1 CP中連約	。 は 各回数が特(	こ多い		
3	<sup>0.46</sup> クラ 年齢	ラスタ3は 層が高	0.1121000204 な					クラスタ1 CP中連約	.は 洛回数が特(	こ多い		
3	0.46 クラ 年齢	iスタ3に 層が高	ひつうスターの男 クラスターの男 シードから オブザベーション	<b>長約</b> 半径	最も	近い	177.9-	クラスタ1 CP中連約	は 格回数が特( の者は全般	こ多い 役的に最終	終会話時間が	長め
3 : クラスター	0.46 クラ 年齢	マスタ3に アスタ3に 層が高 <sup>RMS</sup> 標準偏差	クラスターの要 シードから オブザベーション までの最大距離	<b>E約</b> 半径 超える	最も クラス	近い ター	クラスタ- 重心間の距離	クラスタ1 CP中連約	は 路回数が特( 内者は全般	こ多い 役的に最終	終会話時間が	長め
3 クラスター 1	0.46 クラ 年齢 <sup>度数</sup>	マスタ3に 層が高 <sup>RMS</sup> 標準偏差 0.0921	クラスターの要 シードから オブザベーション までの最大距離 0.9973	<b>E約</b> 半径 超える	最も クラス	近い ター 3	クラスタ- 重心間の距離 0.471	クラスタ1 CP中連約	は <sup>路回数が特()</sup> り者は全般	こ多い 役的に最終	終会話時間が	長め
3 クラスター 1 2	0.46 クラ 年齢 <sup>度数</sup> 1691 436	スタ3に 層が高 <sup>RMS</sup> 標準偏差 0.0921 0.1139	クラスターの要 シードから オブザベーション までの最大距離 0.9973 0.9584	<b>E約</b> 半径 超える	最もクラス	近い ター 3 3	クラスタ- 重心間の距離 0.471 0.338	クラスタ1 CP中連約	は 路回数が特( り者は全般	こ多い 役的に最終	終会話時間が	長め
3 クラスター 1 2 3	0.46 クラ 年齢 <sup>度数</sup> 1691 436 1873	スタ3は 層が高 際MS 標準偏差 0.0921 0.1139 0.0932	クラスターの要 シードから オブザベーション までの最大距離 0.9973 0.9584 0.9748	<ul> <li></li> <li>半径</li> <li>超える</li> <li>i</li> </ul>	最も クラス	近い ター 3 3 2	クラスタ- 重心間の距離 0.471 0.338 0.338	クラスタ1 CP中連約 マロクロン の の	は 路回数が特( <b> う者は全</b> 般	こ多い 役的に最終	終会話時間が	長め
3 クラスター 1 2 3	0.46 クラ 年齢 1691 436 1873	スタ3に 層が高 際MS 標準偏差 0.0921 0.1139 0.0932	クラスターの要 シードから オブザベーション までの最大距離 0.9973 0.9584 0.9748	<ul> <li></li> <li>半径</li> <li>超える</li> <li></li> </ul>	最も クラス クラス	近い ター 3 3 2 、ター	クラスタ- 重心間の距離 0.471 0.338 0.338 平均	クラスタ1 CP中連約 マロック マロック マロック マロック マロック マロック マロック マロック	は <sup>路回数が特()</sup> り者は全般	こ多い 役的に最終	終会話時間が	長め
3 クラスター 1 2 3 クラスター	0.46 クラ 年齢 <sup>度数</sup> 1691 436 1873	マスタ3に アMS 標準偏差 0.0921 0.1139 0.0932 年齢	クラスターの要 シードから オブザベーション までの最大距離 0.9973 0.9584 0.9748	<ul> <li>E約</li> <li>半径</li> <li>超える</li> <li></li> <li></li></ul>	ま し し し し し し し し し し し し し	近い ター 3 2 くター 最新	クラスタ- 重心間の距離 0.471 0.338 0.338 平均 終会話時間	クラスタ1 CP中連約 マロクロン の の CP中連絡回数	は 洛回数が特( り者は全般 <sub>最終連絡日数</sub>	こ多い 役的に最新 CP前連絡回数	終会話時間が	長め
3 クラスター 1 2 3 クラスター 1	0.46 クラ 年齢 1691 436 1873	スタ3は 層が高 標準偏差 0.0921 0.1139 0.0932 年齢 48185742	the resource of the res	<ul> <li>半径</li> <li>超える</li> <li>日</li> <li>日<!--</td--><td>最も クラス 連絡日 358072</td><td>近い ター 3 3 2 <b>メター</b> 4 のの</td><td>クラスタ- 重心間の距離 0.471 0.338 0.338 平均 終会話時間 591620318</td><td>クラスタ1 CP中連約 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2</td><td>は 各回数が特に <b>う者は全</b>所 <sup>最終連絡日数</sup> 0.0169732909</td><td>こ多い 役的に最新 CP前連絡回数 0.0084210526</td><td>終会話時間が</td><td>長め</td></li></ul>	最も クラス 連絡日 358072	近い ター 3 3 2 <b>メター</b> 4 のの	クラスタ- 重心間の距離 0.471 0.338 0.338 平均 終会話時間 591620318	クラスタ1 CP中連約 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	は 各回数が特に <b>う者は全</b> 所 <sup>最終連絡日数</sup> 0.0169732909	こ多い 役的に最新 CP前連絡回数 0.0084210526	終会話時間が	長め
3 クラスター 1 2 3 クラスター 1 2	0.46 クラ 年齢 1691 436 1873 0.31 0.28	スタ3に 層が高 標準偏差 0.0921 0.1139 0.0932 年齢 48185742 88196632	クラスターの要 シードから オブザベーション までの最大距離 0.9973 0.9584 0.9748 年間平均残高 0.0632917973 0.0639327358	<ul> <li>E約     <li>半径     <li>超える</li> <li>6     <li>0.76398     <li>0.34113</li> </li></li></li></li></ul>	最も クラス <b>クラス</b> 連絡日 358072 314985	近い ター 3 3 2 くター・ 長 # 0.06	クラスタ- 重心間の距離 0.471 0.338 0.338 平均 終全話時間 591620318 737437327	クラスタ1 CP中連約 シング の の の の の の の の の の の の の の の の の の の	は 各回数が特( う者は全分 最終連絡日数 0.0169732909 0.3079075835	に多い ひのに最新 CP前連絡回数 0.0084210526 0.1282568807	終会話時間が	長め

## 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





### 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







#### 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



### アンケートのお願い・ご質問

### <u>10月19日 機械学習によるビッグデータ分析の手法-2</u>

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



https://sas.qualtrics.com/jfe/form/SV\_1C6i14BbnisRffM



- ・本日のアーカイブは、2022年10月24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**





Copyright © SAS Institute Inc. All rights reserved.