上智大学
2011年度オープンキャンパス 体験授業

統計的データ解析を通して明らかにする世の中のさまざま

理工学部　情報理工学科
加藤　剛（かとう　たけし）

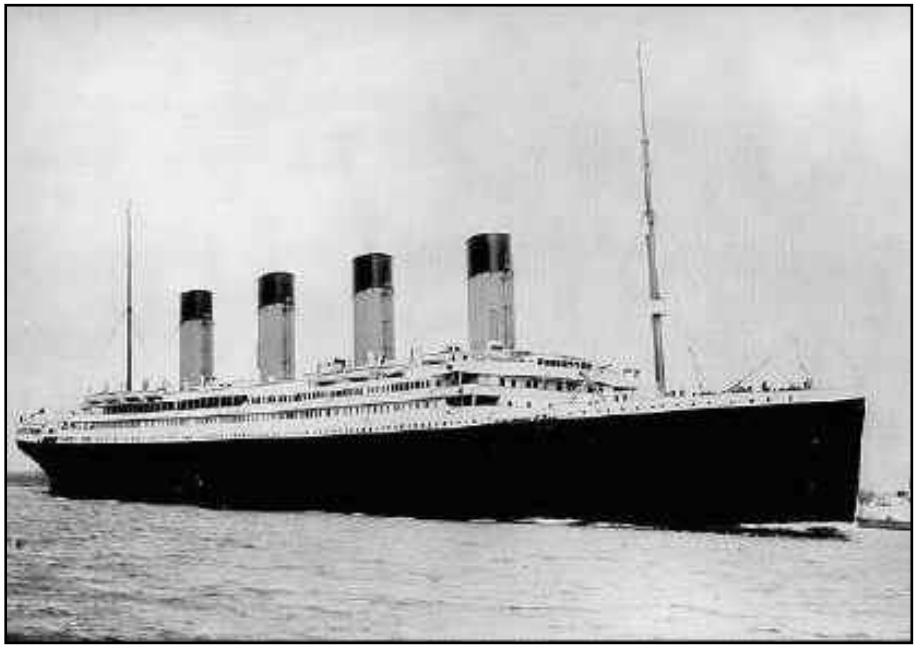
内容

- タイタニック号遭難事故の真相
（モザイクプロット）
- あなたは平均的日本人？
（スピアマンの順位相関係数）
- アメリカ企業産と日本企業産に違いはあったのか
（クラスタ分析）
- 犯罪者が狙ったもの
（因子分析）
- まとめ

タイタニック号遭難事故の真相

モザイクプロット

タイタニック号遭難事故



- 総トン数　46,328トン
- 全長　269.1 m
- 全幅　28.2 m
- 速力　23ノット（42.6km/h)

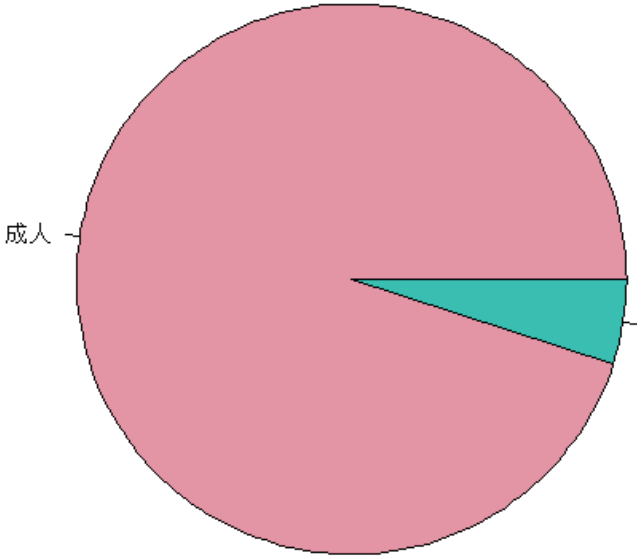
- 1912年4月10日，乗員と乗客2201名を乗せ，処女航海でイギリスのサウサンプトン港を出港．ニューヨークへ向かう
- 4月15日，北大西洋のニューファンドランド沖で氷山に衝突し，沈没

生存の可否に関するデータ

識別番号	船室等級	性別	年齢層	生存の可否
1	2	男	成人	否
2	3	男	未成年	可
3	1	女	成人	否
4	3	男	成人	否
5	乗務員	女	成人	可
6	2	女	成人	可
7	1	男	成人	否
…	…	…	…	…
…	…	…	…	…
2199	1	男	成人	可
2200	乗務員	男	成人	否
2201	3	男	成人	否

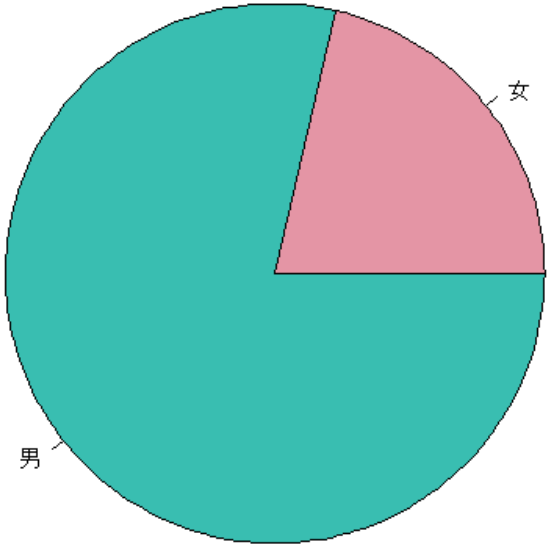
タイタニック号の乗客構成

年齢層



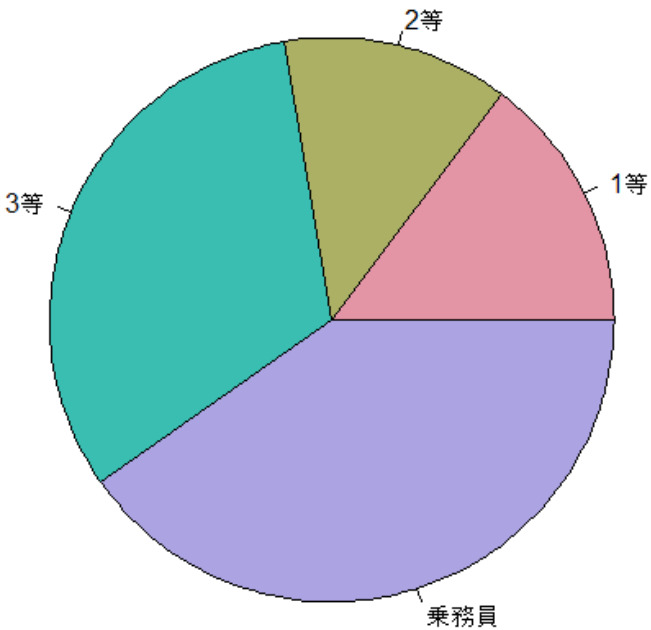
- 緑＝未成年
- 桃＝成人

性別



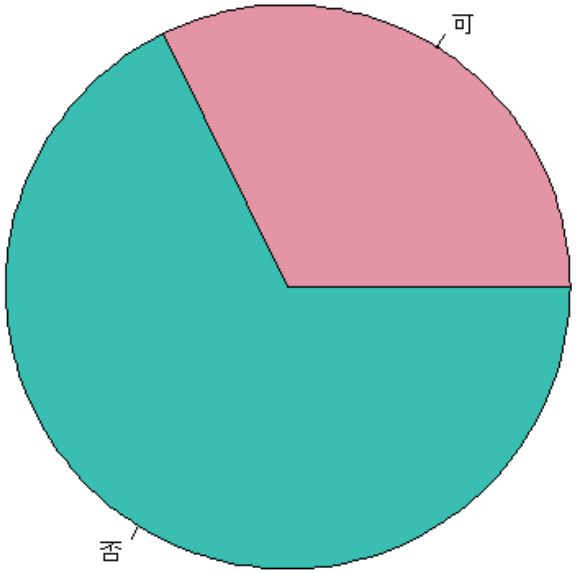
- 緑＝男
- 桃＝女

船室等級



- 桃＝1等
- 黄土＝2等
- 緑＝3等
- 紫＝乗務員

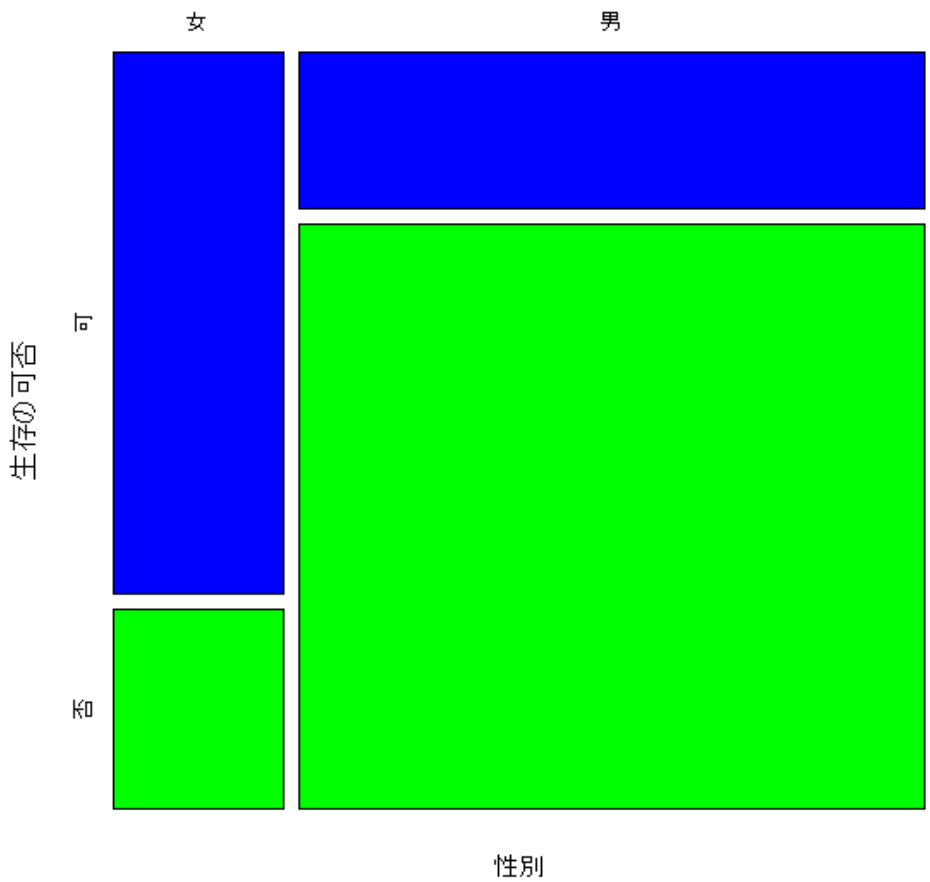
生存の可否



- 桃＝可
- 緑＝否

英国人紳士気質はあったか？

性別と生存の可否の関係を表すモザイクプロット

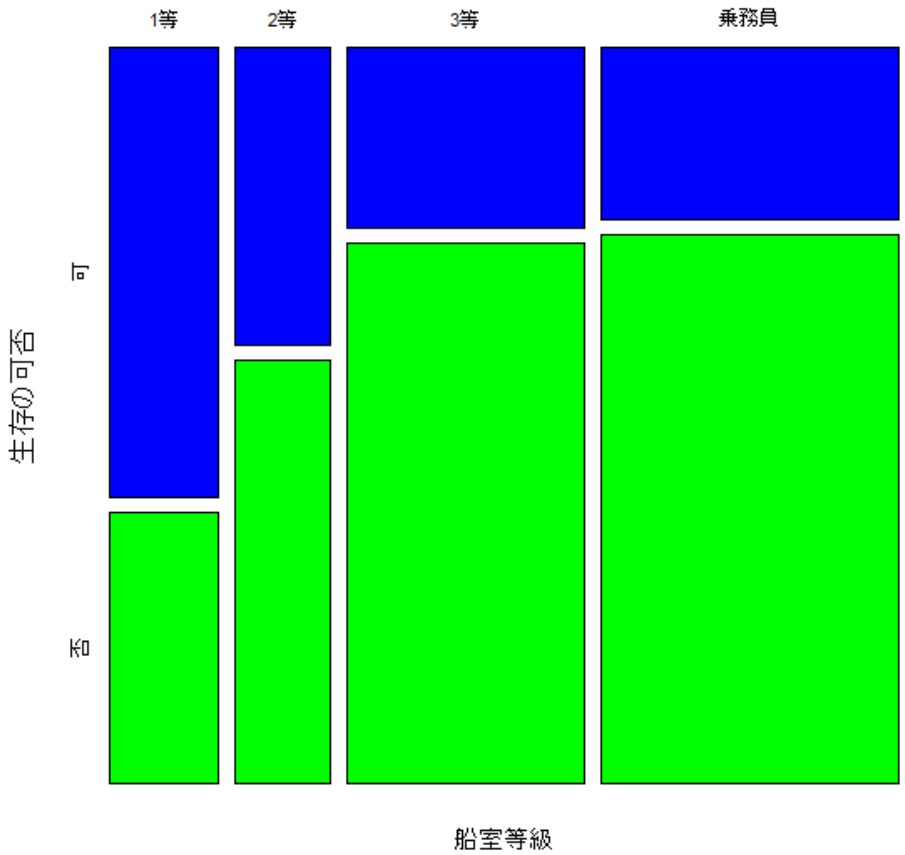


		性別	
		女	男
生存の可否	可	73.2 %	21.2 %
	否	26.8 %	78.8 %

	女	男
乗船者男女比率	21.4 %	78.6 %

表に出てこない真実

船室等級と生存の可否の関係を表すモザイクプロット



		船室等級			
		1等	2等	3等	乗務員
生存の可否	可	62.5 %	41.4 %	25.2 %	24.0 %
	否	37.5 %	58.6 %	74.8 %	76.0 %

	1等	2等	3等	乗務員
船室等級比率	14.8 %	12.9 %	32.1 %	40.2 %

あなたは平均的日本人？

スピアマンの順位相関係数

夢のあるお話

JTB による 2009年発表の海外挙式人気ランキング

第１位
ハワイ

第２位
ミクロネシア

第３位
オーストラリア

第４位　北米またはカリブ海

第５位　バリ島

第６位　フィジー，ニューカレドニア，
またはタヒチ

第７位　ヨーロッパ

授業履修者がつけた海外挙式人気ランキング

場所	JTB順位	A嬢	B氏	C嬢	D嬢	カトウ
ハワイ	1	7	4	2	3	7
ミクロネシア	2	3	1	1	7	4
オーストラリア	3	2	7	4	6	5
北米または カリブ海	4	6	2	3	2	1
バリ島	5	5	6	5	1	6
フィジー，ニューカレドニア，タヒチ	6	1	5	7	5	3
ヨーロッパ	7	4	3	6	4	2

問題

A嬢，B氏，C嬢，D嬢，カトウの5名は，海外挙式場所の好みに関して，一般的な日本人（＝JTB 調査結果）とどれくらい似ているか．あるいは，逆の好みをもっているか

解決の手がかり

スピアマンの順位相関係数

- 順位の付け方の関連性から類似度を測る
- ー1 から 1 までの値をとる

1 に近い	⇒	好みが似ている
0 に近い	⇒	好みにあまり関係はない
ー1 に近い	⇒	好みは正反対

授業履修者がつけた海外挙式人気ランキング

場所	JTB順位	A嬢	B氏	C嬢	D嬢	カトウ
ハワイ	1	7	4	2	3	7
ミクロネシア	2	3	1	1	7	4
オーストラリア	3	2	7	4	6	5
北米または カリブ海	4	6	2	3	2	1
バリ島	5	5	6	5	1	6
フィジー，ニューカレドニア，タヒチ	6	1	5	7	5	3
ヨーロッパ	7	4	3	6	4	2

このデータから，スピアマンの順位相関係数の計算実行

ただいま解析中

```
> cor(honeymoon01[,c("A嬢","B氏","C嬢","D嬢","JTB","カトウ")],
                                use="complete.obs", method="spearman")

           A嬢        B氏        C嬢        D嬢       JTB     カトウ
A嬢  1.0000000 -0.2857143  0.2857143 -0.71428571 -0.3571429  0.21428571
B氏 -0.2857143  1.0000000 -0.2500000 -0.14285714  0.1428571  0.50000000
C嬢  0.2857143 -0.2500000  1.0000000  0.14285714 -0.5357143 -0.21428571
D嬢 -0.7142857 -0.1428571  0.1428571  1.00000000 -0.2142857 -0.07142857
JTB -0.3571429  0.1428571 -0.5357143 -0.21428571  1.0000000 -0.57142857
カトウ 0.2142857  0.5000000 -0.2142857 -0.07142857 -0.5714286  1.00000000

> showData(honeymoon01, placement='-20+200', font=getRcmdr('logFont'),
                                maxwidth=80, maxheight=30)
```

終了

解析結果

JTB公表順位とのスピアマンの順位相関係数

氏名	順位相関係数	一般的な日本人との類似性
A 嬢	−0.36	若干ではあるが，好みは反対
B 氏	0.14	ほとんど関係がない（独自路線）
C 嬢	0.89	好みはかなり似ている
D 嬢	−0.21	ごく若干ではあるが，好みは反対
カトウ	−0.57	中程度の強さで好みは反対

相性診断に使えます

アメリカ企業産と日本企業産に違いはあったのか

クラスタ分析

似たもの同士のグループ分け

興味の対象をじっくり調べるときに，あらかじめ似たもの同士にグループ分けしておくことがある

【例１】犬

小型犬

大型犬

【例2】車

小型車

大型車

問題　次のデータは，5教科の試験に対する受験者7名の得点データである．これらの受験者は，いくつのグループに分けることができるか

	数学	物理	国語	英語	日本史
織田	89	90	67	46	50
羽柴	57	70	80	85	90
松平	80	90	35	40	50
伊達	40	60	50	45	55
上杉	78	85	45	55	60
毛利	55	65	80	75	85
鍋島	90	85	88	92	95

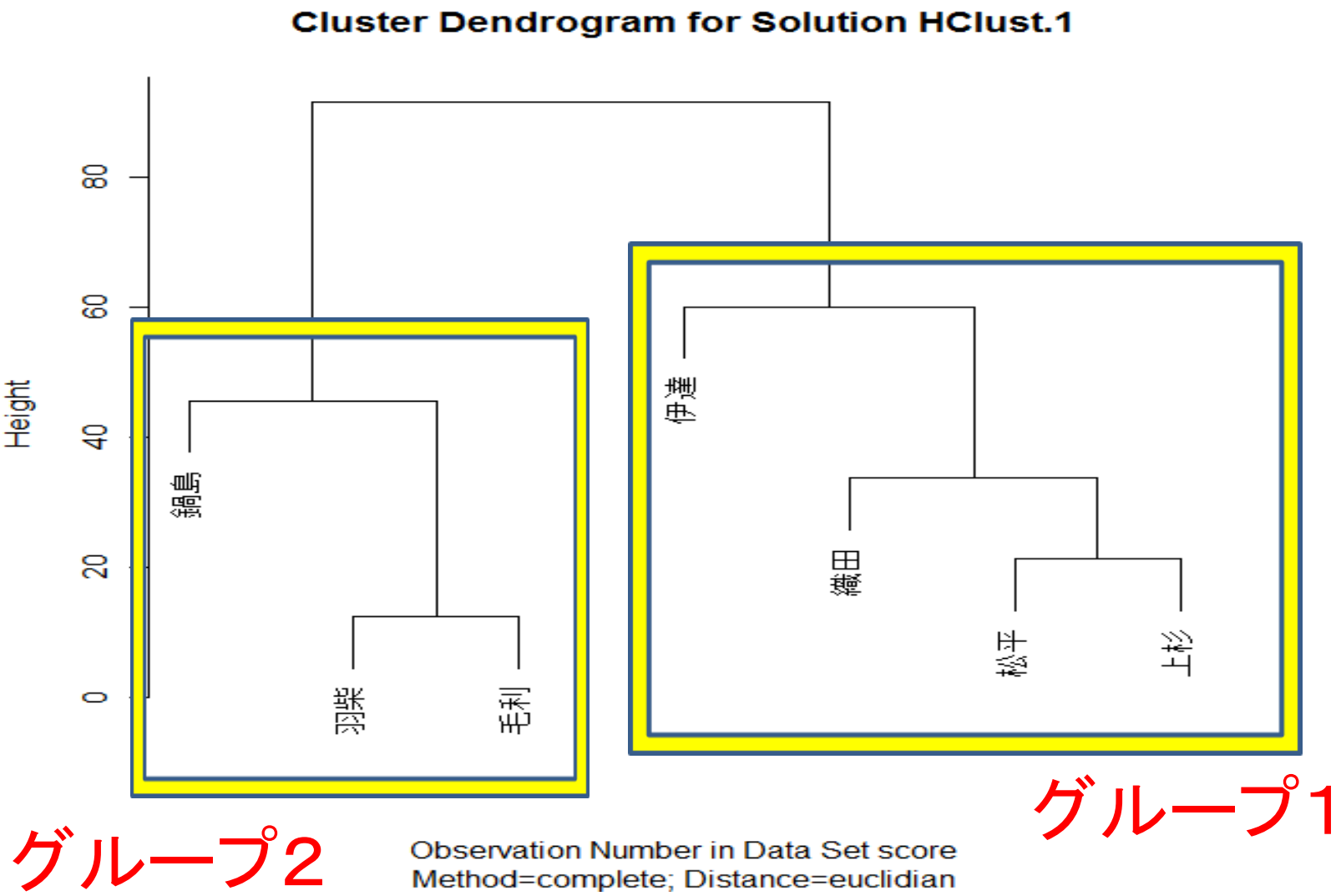
クラスタ分析を使ってグループ分けを試みる

ただいま解析中

```
> remove(.Table)

> score <- read.table("D:/Documents/教材/情報リテラシー（統計処理）
/data/score.txt",
+   header=TRUE, sep="", na.strings="NA", dec=".", strip.white=TRUE)

> showData(score, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80,
+   maxheight=30)

> HClust.1 <- hclust(dist(model.matrix(~-1 + 英語+国語+数学+日本史+物理, score)) ,
+   method= "complete")

> plot(HClust.1, main= "Cluster Dendrogram for Solution HClust.1", xlab=
+   "Observation Number in Data Set score", sub="Method=complete;
+   Distance=euclidian")
```

終了

グループ分けの検討


Cluster Dendrogram for Solution HClust.1
Height
0
20
40
60
80
鍋島
羽柴
毛利
伊達
田邊
松平
上杉
グループ2
グループ1
Observation Number in Data Set score
Method=complete; Distance=euclidian

数値によるクラスタ分析の要約表

グループ1				
英語	国語	数学	日本史	物理
46.50	49.25	71.75	53.75	81.25

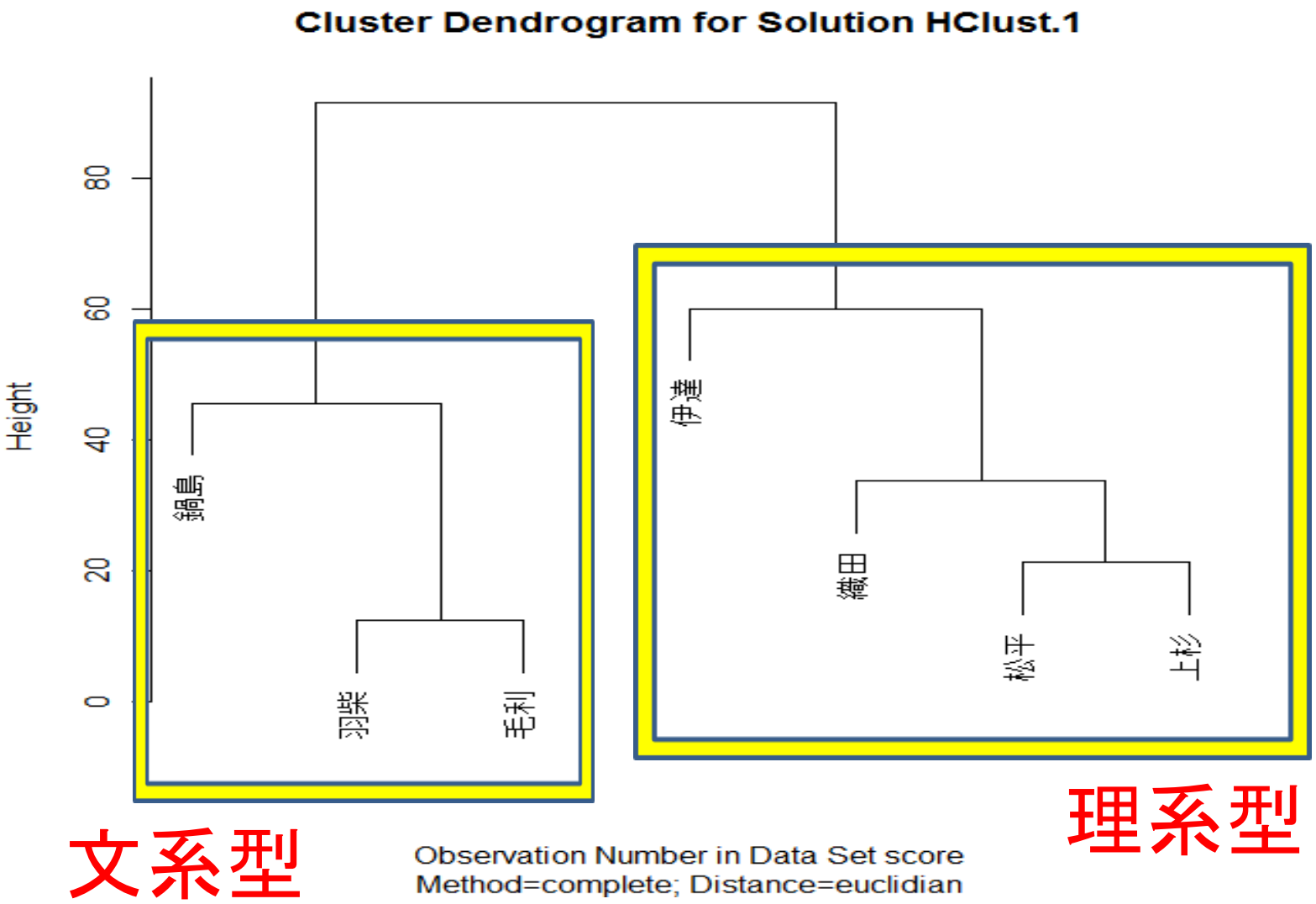
理系科目の平均点が高い

グループ2				
英語	国語	数学	日本史	物理
84.00	82.67	67.33	90.00	73.33

文系科目の平均点が高い

- グループ1 = 理系型
- グループ2 = 文系型

グループ分けの結果


Cluster Dendrogram for Solution HClust.1
Height
0
20
40
60
80
鶴島
羽柴
毛利
伊藤
藤田
松平
上杉
文系型
理系型
Observation Number in Data Set score
Method=complete; Distance=euclidian

日米自動車戦争

- 1970年代以降，日本企業が作る車がアメリカの消費者に好評を博した
- 対照的に，BIG3と呼ばれて世界の自動車市場の覇者として君臨してきた，フォード，GM，クライスラーの勢いに陰りが見え始めた

当時まことしやかに語られたこと

BIG3 が作る車は，なぜ日本車に勝てないのか

- 燃費が悪い
- 図体ばかり大きい
- ガソリンをがぶ飲みする

【参考】

アメリカの戦闘機は，重武装をした重量のある機体を，高出力エンジンで強引に飛ばすと言われていた

噂をデータで確かめる

出典　*Consumer Reports*, April, 1990

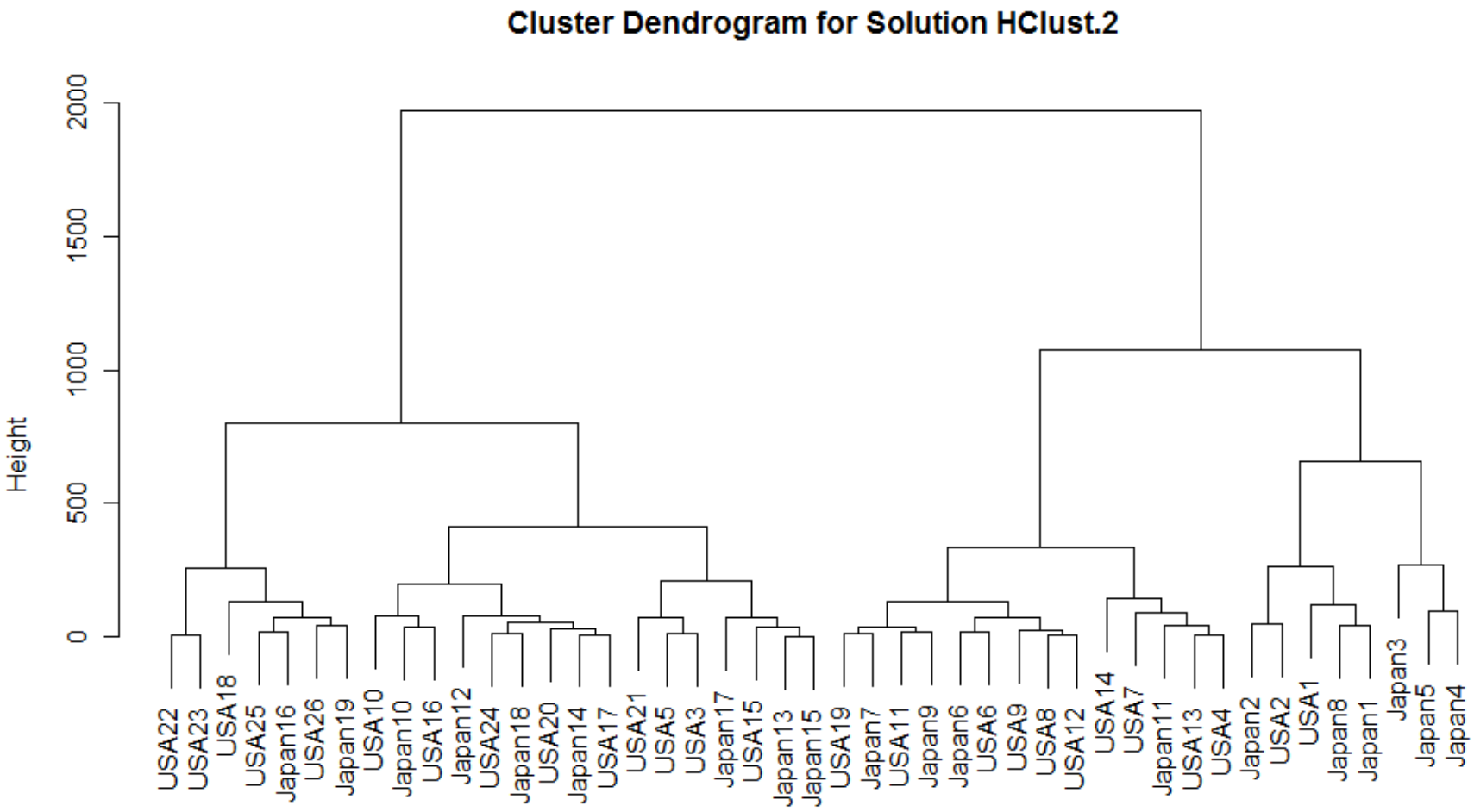
- 日本企業生産の１９車種，アメリカ企業生産の２６車種
- 燃費，重量，排気量の3項目についてのデータ

- 当時の噂通りならば，クラスタ分析を行うと，日本企業生産の車とアメリカ企業生産の車がある程度きれいに分かれることが期待される

解析対象データ

生産企業	燃費	重量	排気量
USA	33	2560	97
USA	33	2345	114
Japan	32	2440	113
Japan	25	2295	109
Japan	34	1900	73
…	…	…	…
…	…	…	…
USA	18	3735	202
USA	18	3665	182
Japan	19	3690	146

解析結果



グループ分け不可能！

犯罪者が狙ったもの

因子分析

犯罪者の心理を探る

アメリカの犯罪データ（全米50州＋1特別区）

人口10万人あたりの1 年間の発生件数

州名	殺人	暴行	強盗	住居侵入窃盗	窃盗	車上荒らし
アラバマ	14.2	25.2	96.8	1135.5	1881.9	280.7
アラスカ	10.8	51.6	96.8	1331.7	3369.8	753.3
アリゾナ	9.5	34.2	138.2	2346.1	4467.4	439.5
アーカンソー	8.8	27.6	83.2	972.6	1862.1	138.4
カリフォルニア	11.5	49.4	187	2139.4	3499.8	663.5
コロラド	6.3	42	170.7	1935.2	3903.2	477.1
…	…	…	…	…	…	…
ウェストバージニア	6	13.2	42.2	597.4	1341.7	163.3
ウィスコンシン	2.8	12.9	52.2	846.9	2614.2	220.7
ワイオミング	5.4	21.9	39.7	811.6	2772.2	282

問題

犯罪者は，何を狙って犯罪を起こしたのか

↕

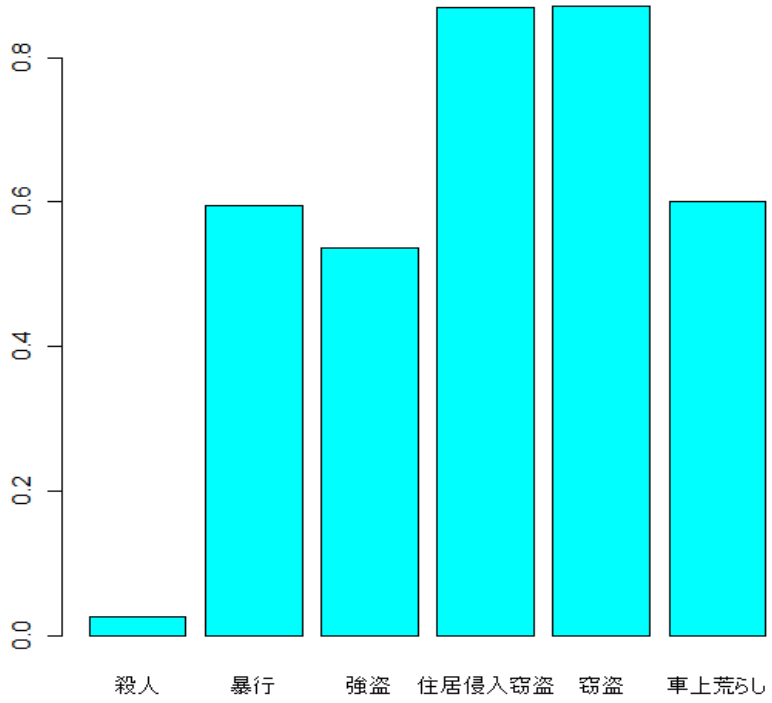
何を狙って起こした犯罪の結果が，このデータなのか

数値の表だけを見ていても，何もわからない

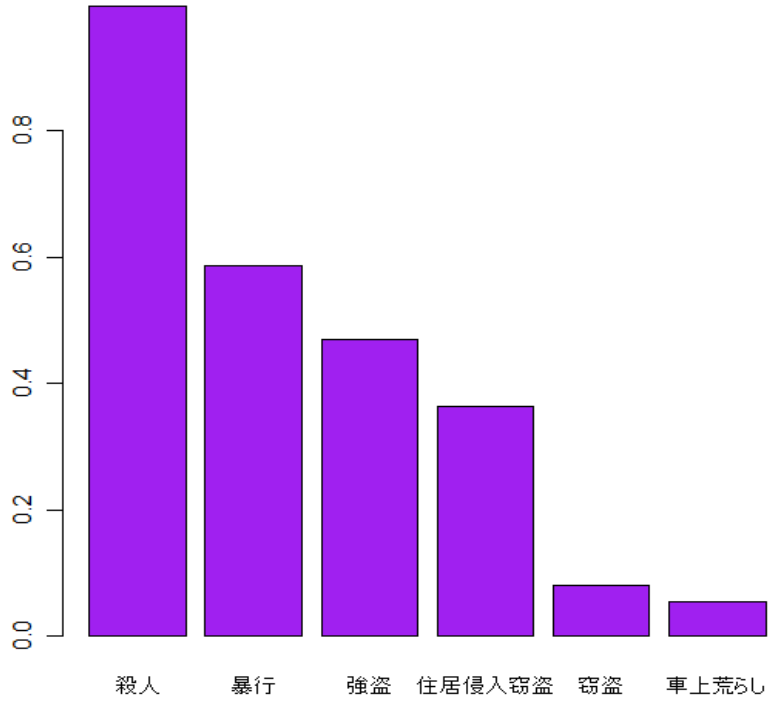
因子分析で犯罪者の心理を探る

因子負荷量の計算

第1因子負荷量

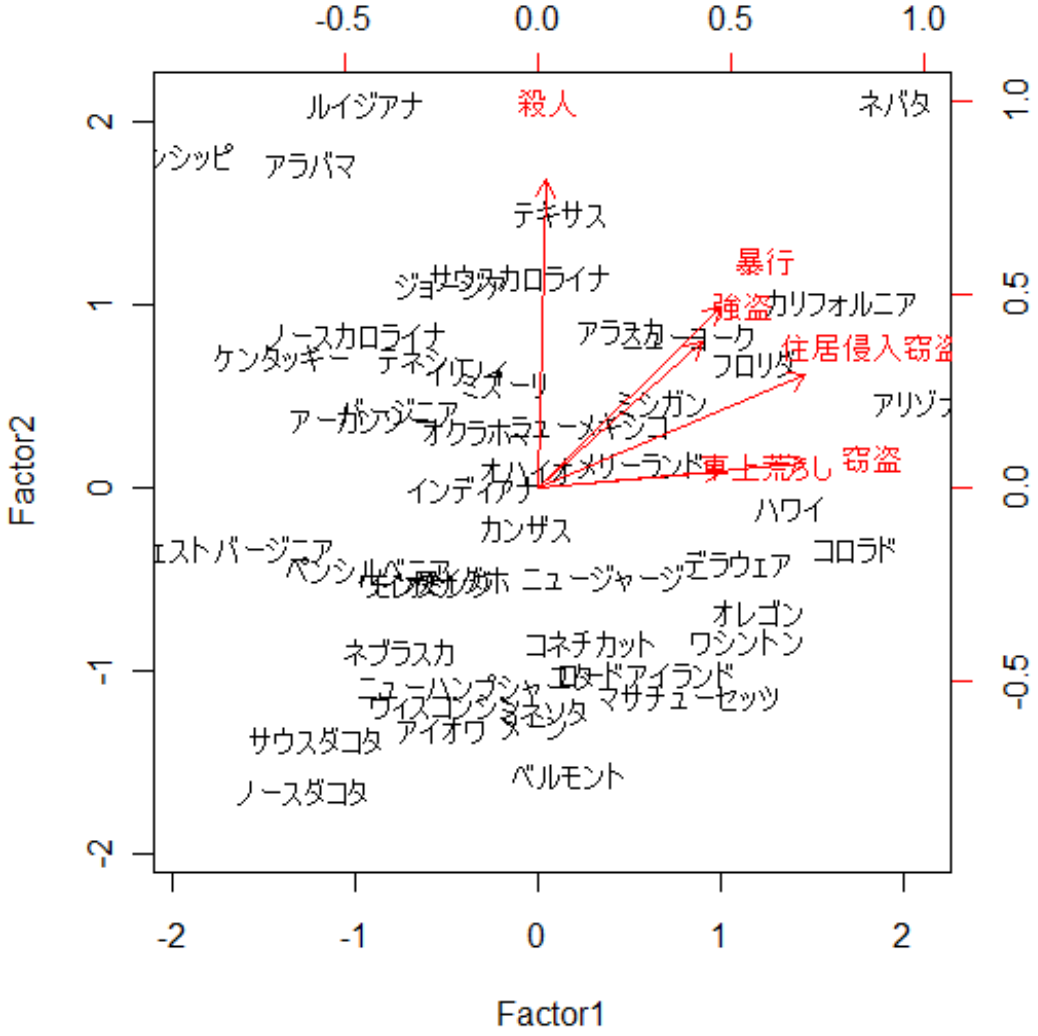


第2因子負荷量

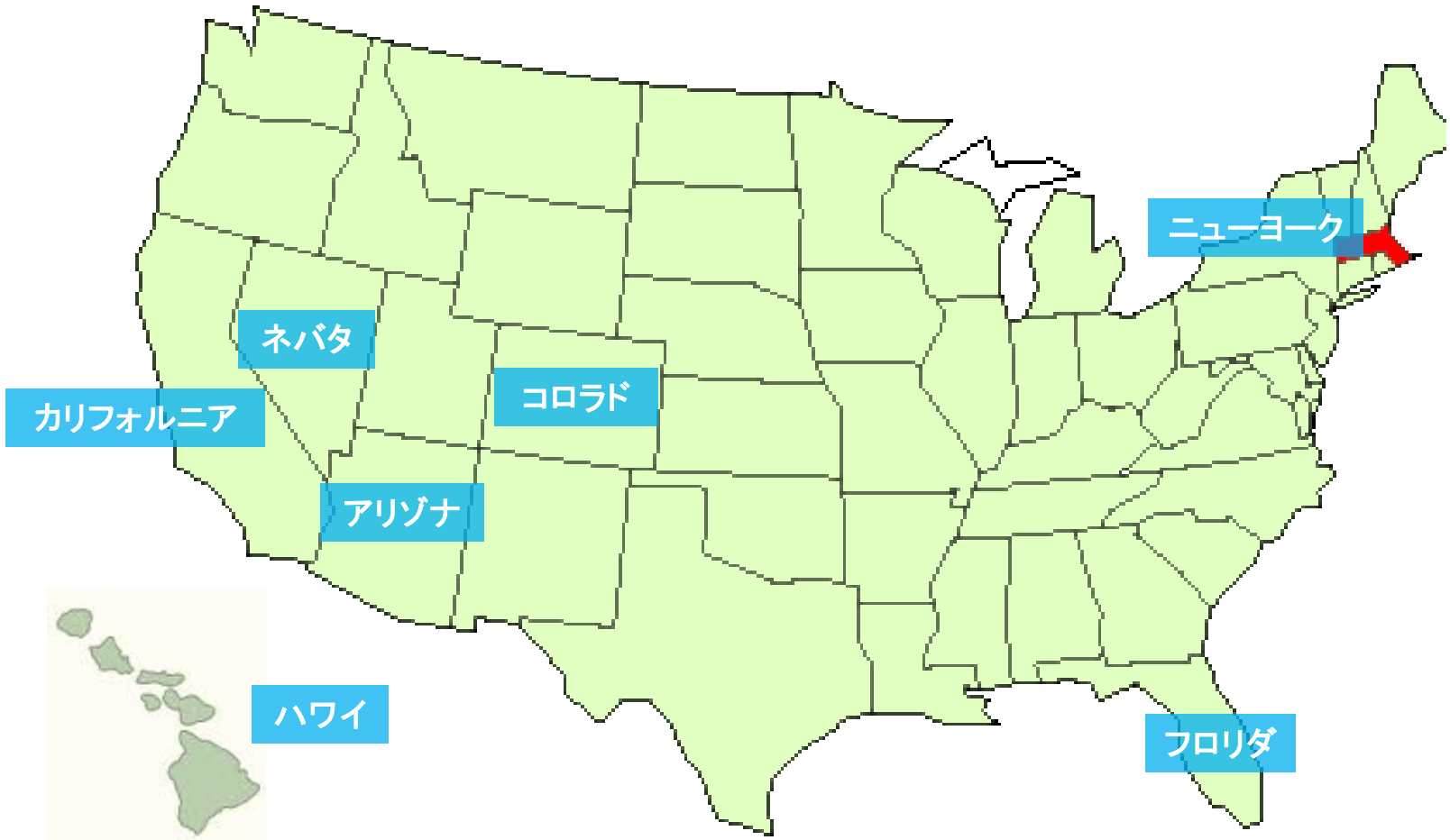


- 第1因子＝金銭目的
- 第2因子＝恨み等による生命剥奪目的

因子負荷量と因子得点のバイプロット

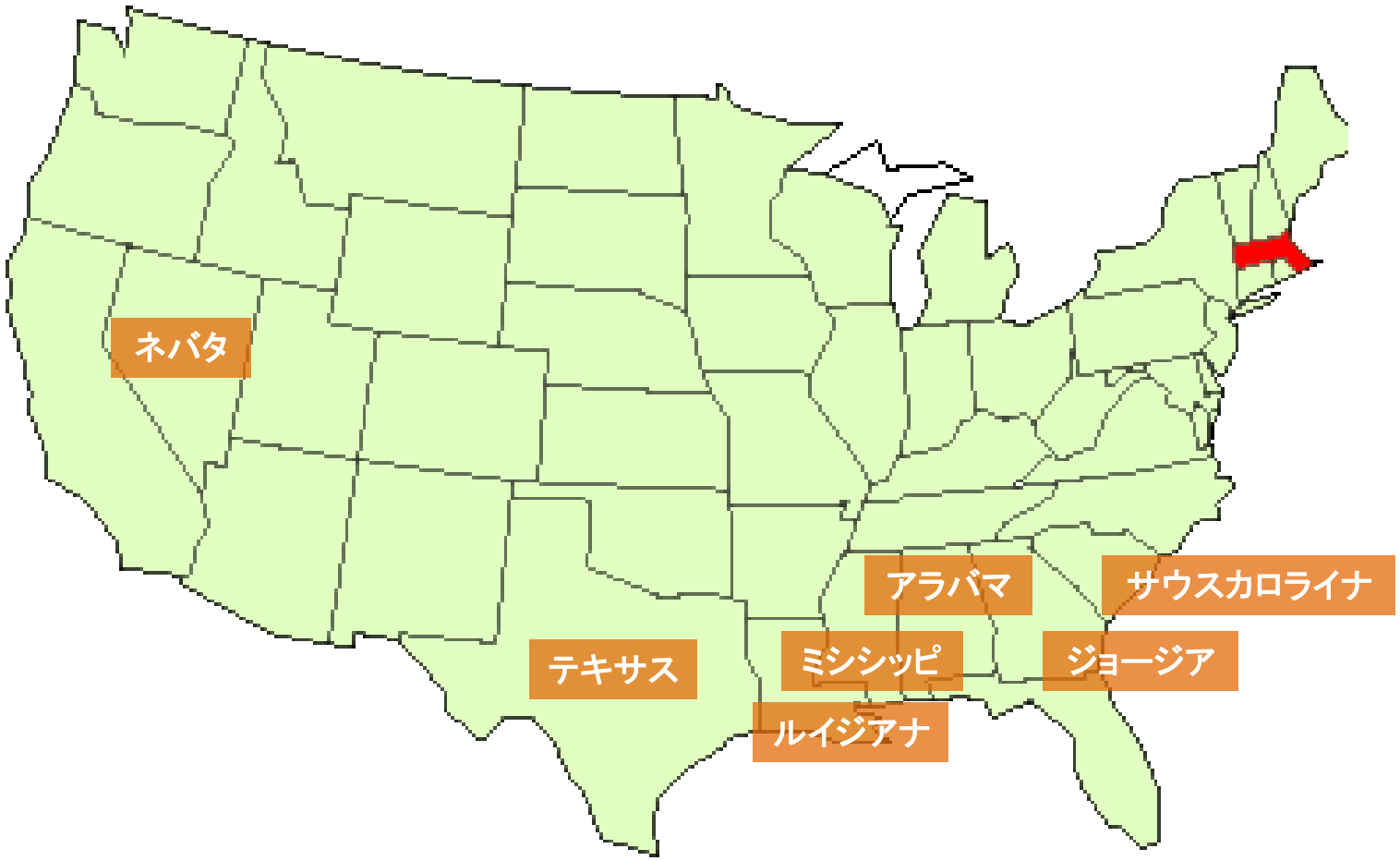


第1因子（金銭目的）の得点が高い州



金銭目的の犯罪は，ほぼ全米にわたっている

第2因子（恨み等による生命剥奪目的）の得点が高い州



南部の州で生命剥奪目的の犯罪が多い

まとめ

- 統計学（統計的データ解析）の応用範囲
  - 歴史，文化人類学（タイタニック号遭難事故の真相）
  - 人類行動学（海外挙式場所の好みの類似性）
  - 経済学（日米自動車戦争の真実）
  - 心理学（犯罪者の行動心理）
  - 数理ファイナンス（金融商品の設計，株価等の予測）
  - 製薬および医療（臨床試験データの解析）
  - 実験計画（最適栽培環境の探索実験）
  - 生物学および生命情報（ゲノム解析）
  - その他多数

- 上智大学で統計的データ解析を学べるところ

経済学部，総合人間科学部など

それぞれの専門分野にやや特化した統計的データ解析が学べます

理工学部情報理工学科

◆ 理工系の十八番である数学を武器にして，確率論，統計学，統計的データ解析を，基礎的な理論からしっかりと学ぶことができます

◆ 統計的データ解析に関する汎用的な知識と技能を活用して，多様な分野に関わることができます

この体験授業を通して,
統計的データ解析で何ができるのかを
直感的に理解し,
進路選択のご参考になれば幸いです

ご参加ならびにご静聴
ありがとうございました