データサイエンス講座
第2回 機械学習その１

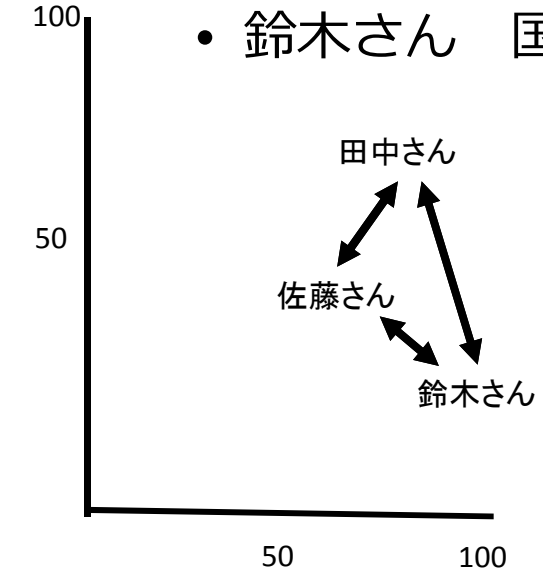
- クラスタリング分析
- 主成分分析
- 因子分析
- アソシエーション分析

クラスタリング分析

- クラスタリング分析でできること：
  - 教師なし学習
  - 似たもの同士をグルーピングする
  - グルーピングすることにより、どのアイテムとアイテムが“似ている”を把握することができる

- “似ている”ものの定義
  - お互いの“距離”が近い
  - 距離　＝　ユークリッド距離

  - ユークリッド距離の例：
    - 田中さん　国語　80点　数学　60点
    - 佐藤さん　国語　70点　数学　40点
    - 鈴木さん　国語　90点　数学　30点

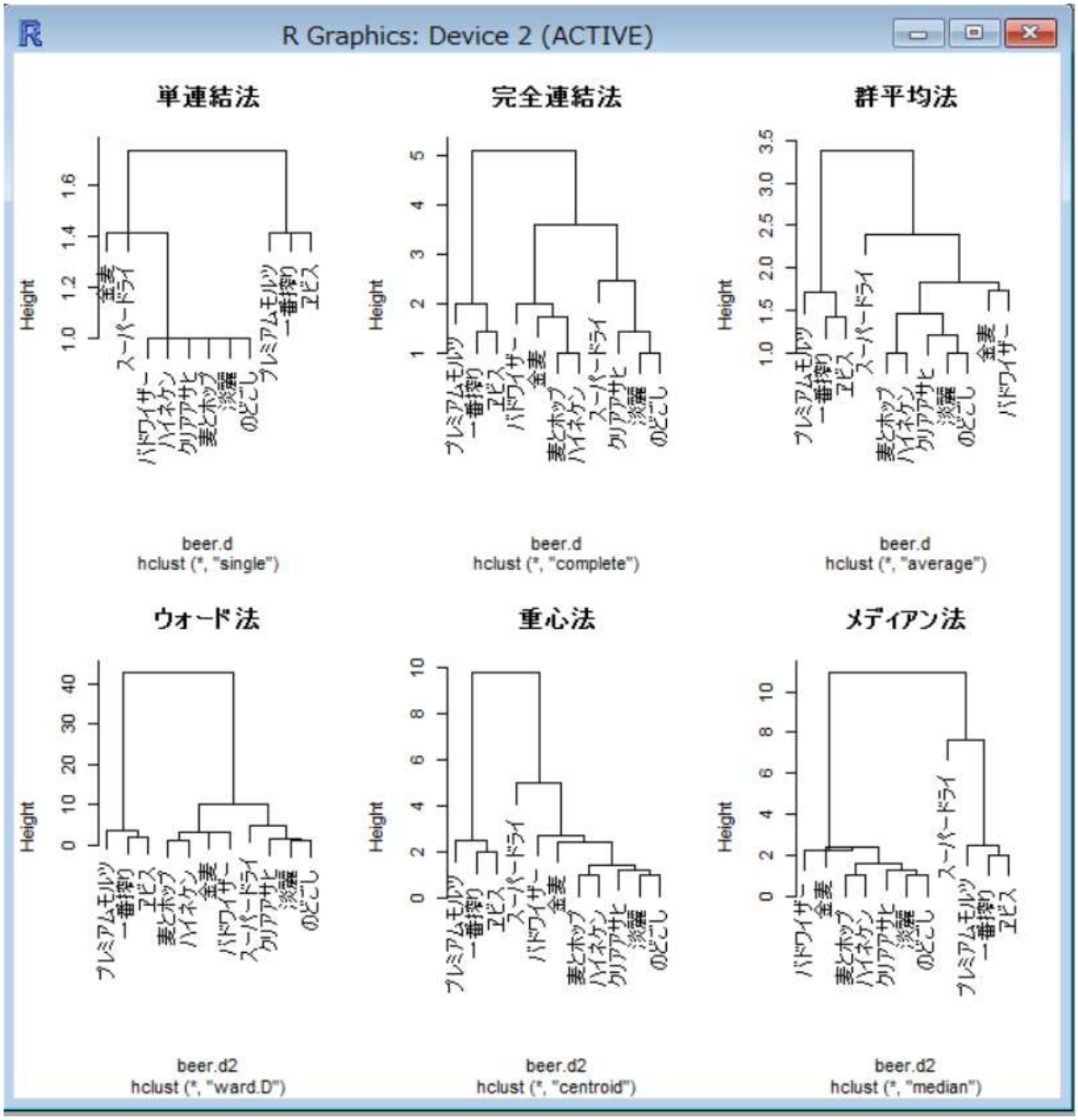


距離を求める方法

田中さんと佐藤さんの距離
＝√（田中さん国語80点 – 佐藤さん国語70点)^2
　+（田中さん数学60点 – 佐藤さん数学40点)^2
=√ 100 + 400 = 22.36
田中さんと鈴木さんの距離 31.6
佐藤さんと鈴木さんの距離 22.36

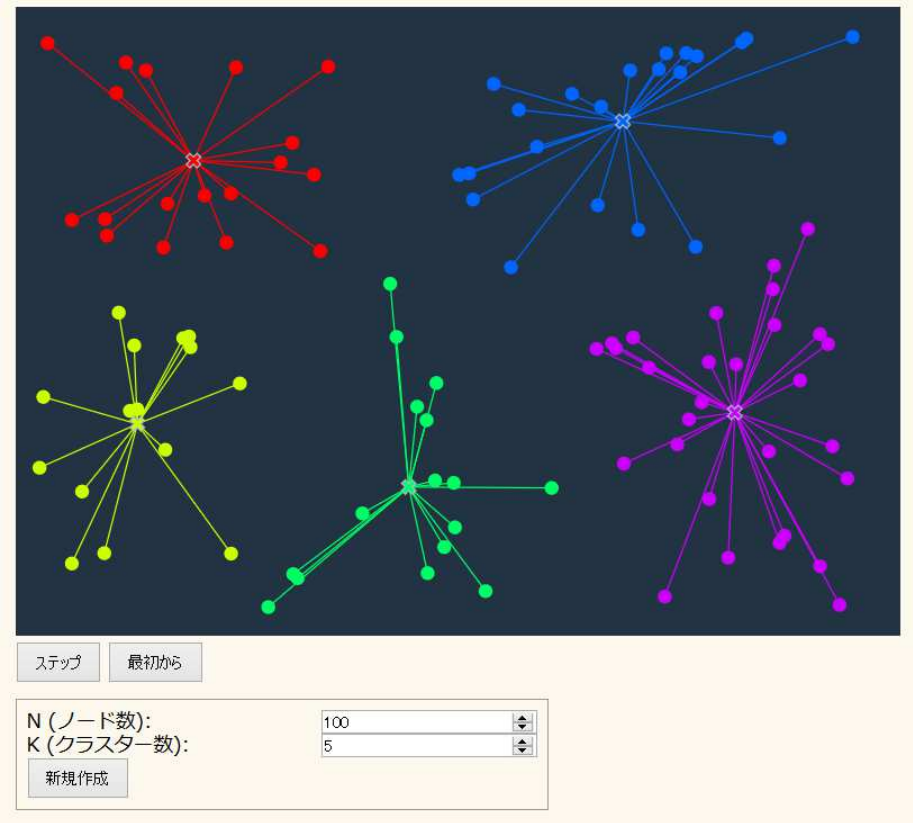
クラスタリング分析

- height = 結合されたクラスタリング間の距離
- 手法によってことなるものの、
  - プレミアムモルツと一番搾りは同じクラスタ
  - スーパードライは上記とは別のクラスタ
  - ２つくらいに分類できそう



k平均法

- k平均法とは？
  - 非階層クラスタリング
  - クラスタの数をkとしてクラスタをする方法
  - クラスタリング方法
    - 初期化：データをランダムにk個に分類し、クラスタの重心を求める
    - クラスタの決定：あるデータに対し、クラスタの重心の中で最も近いクラスタがデータの属するクラスタとする。
    - クラスタの中心の再計算：新しく属したクラスタについて重心を再計算し、収束するまで続ける。
    - ビジュアル化
    - http://tech.nitoyon.com/ja/blog/2013/11/07/k-means/



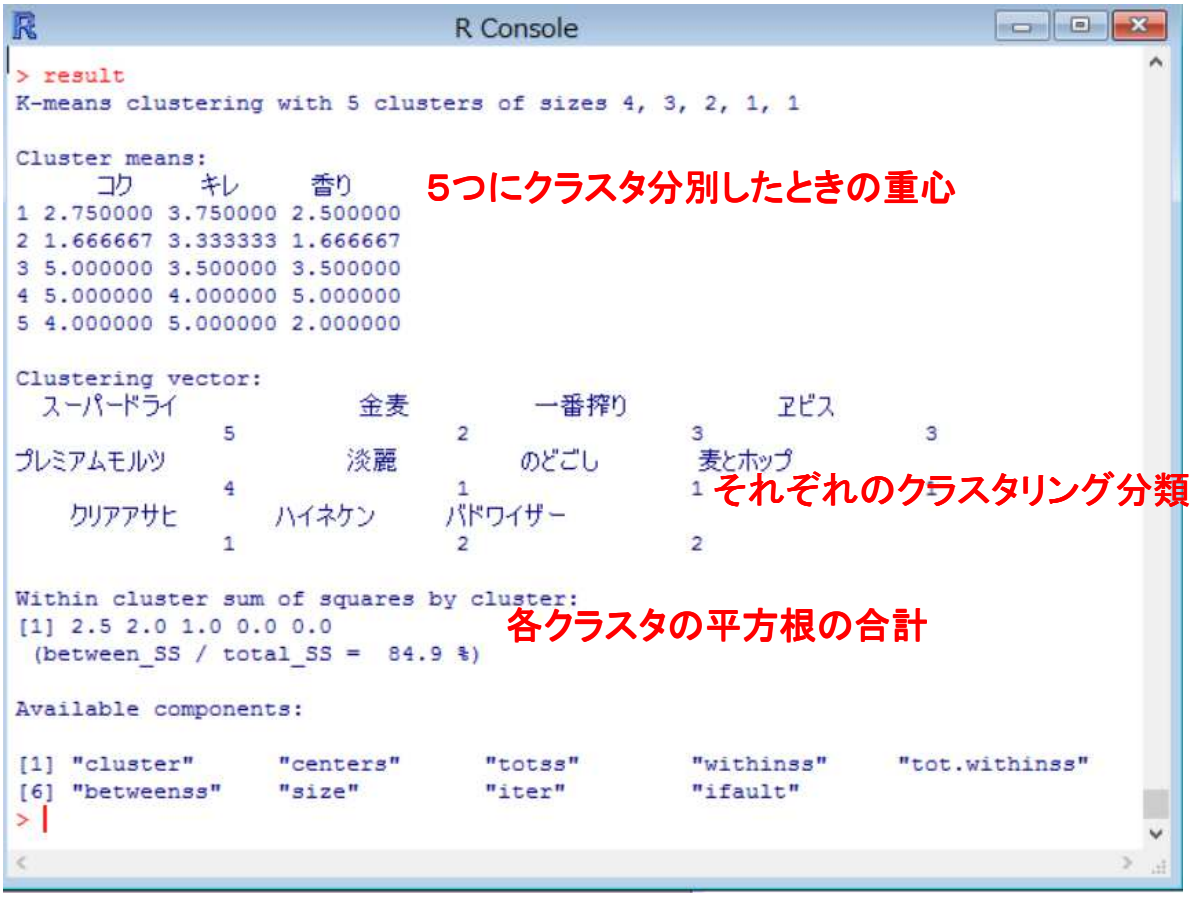
k 平均法

RによるK平均法

- kmeans(x, k, nstart=5, iter.max=10, algorithm=c("Hartigan-Wong")
  - x → データセット
  - k → クラスタの数
  - nstart -> 初期値に試すデータの数
  - iter.max → 計算回数上限
  - algorithm → 計算アルゴリズム

データのセット

- x = as.matrix(beer)
- result = kmeans(x,5)
- result



主成分分析

- 主成分分析とは？
  - Wikipediaによれば、「直交回転を用いて変数間に相関がある元の観測値を、相関の無い主成分とよばれる値に変換するための数学的な手続きのこと」
  - ざっくりとした全体像
    - 19世紀のフランス印象派
    - 風景を細部まで写実するのではなく、対象全体から水、光などを浮きだたせる手法

出所：大原美術館　クロード・モネ　睡蓮
http://www.ohara.or.jp/201001/jp/C/C3a26.html

  - 主成分分析のアプローチ
    - たくさんの情報のなかから、水、光などの重要な部分を浮き出すアプローチ（＝次元削減）

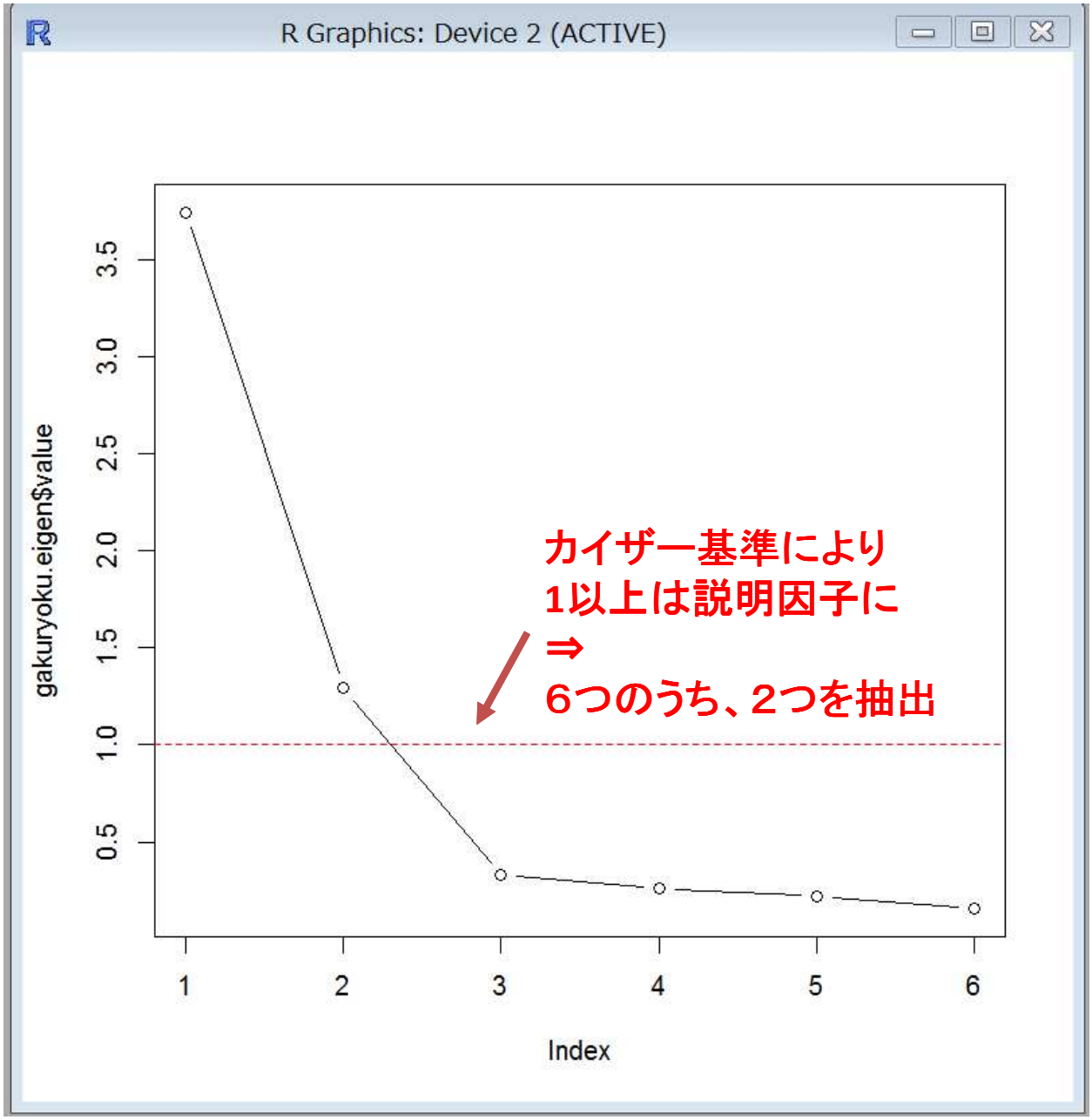
主成分分析

◘主成分分析の流れ

1. データを入力する
2. 要素間の相関行列（近さ）を求める
3. 相関行列から固有値と固有ベクトルを求める
4. 成分をプロットする
5. 主成分と主成分得点を求める
6. 分析結果を検討する

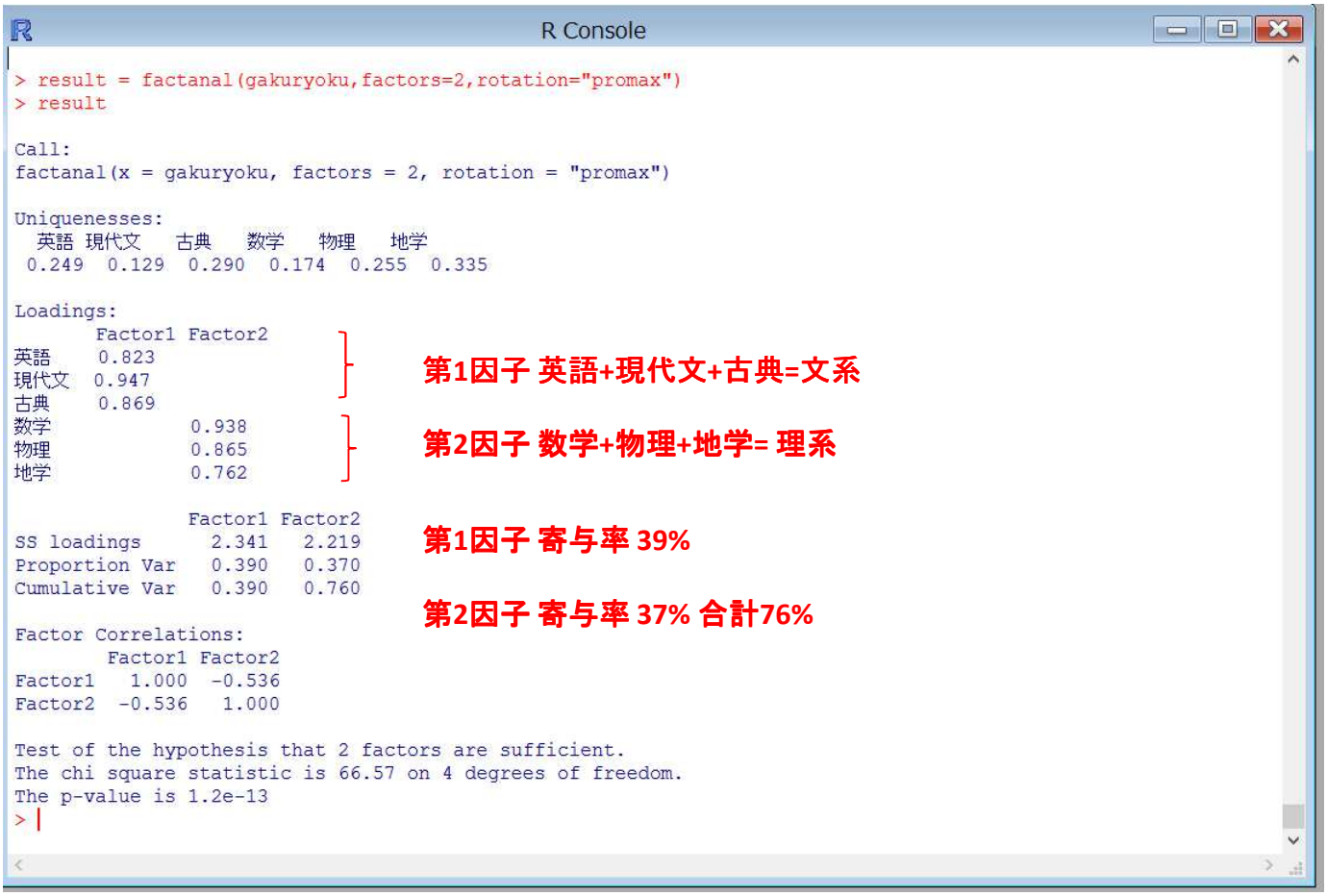
因子分析

- 主成分分析と同様スクリープロットで、主要な因子の説明度（固有値）を図示する。
  - plot(gakuryoku.eigen$value, type="b") ; abline(h=1, col="red",lty=2)



因子分析

- 因子分析
  - result = factanal(gakuryoku, factor=2)
  - result



アソシエーション分析

- コンビニPOSデータを関連分析したい場合
  - そのままR、Excelでは分析できない（トランザクション方式）

ID	項目（アイテム）
000000001	調理パン、調理パン、缶コーヒー
000000002	タバコ、缶コーヒー
000000003	雑誌、お菓子
000000004	おにぎり、タバコ、お惣菜
000000005	缶コーヒー、お菓子
…	

- 一般的な解決方法
  - アイテムをすべて列挙する

ID	項目（アイテム）			
	缶コーヒー	タバコ	お菓子	…
000000001	1			
000000002	1	1		
000000003			1	
000000004		1		
000000005	1		1	
…				

  - 問題点：アイテム数が増えると計算量は膨大に

アソシエーション分析

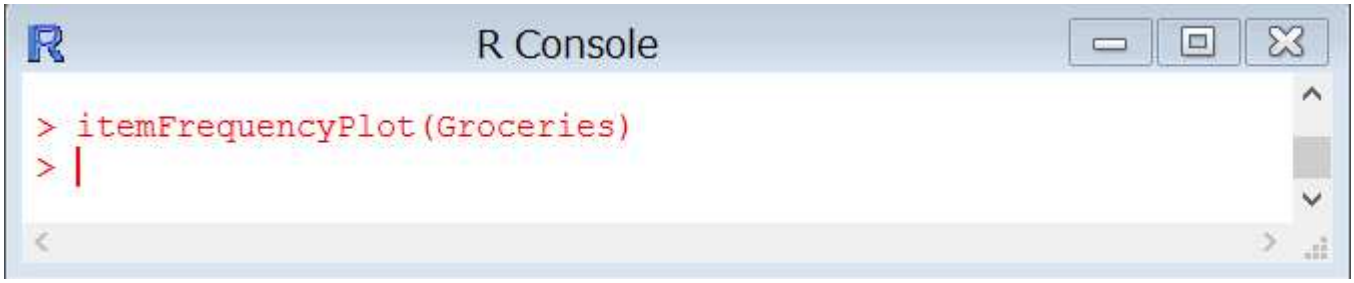
◘アソシエーション分析の考え方

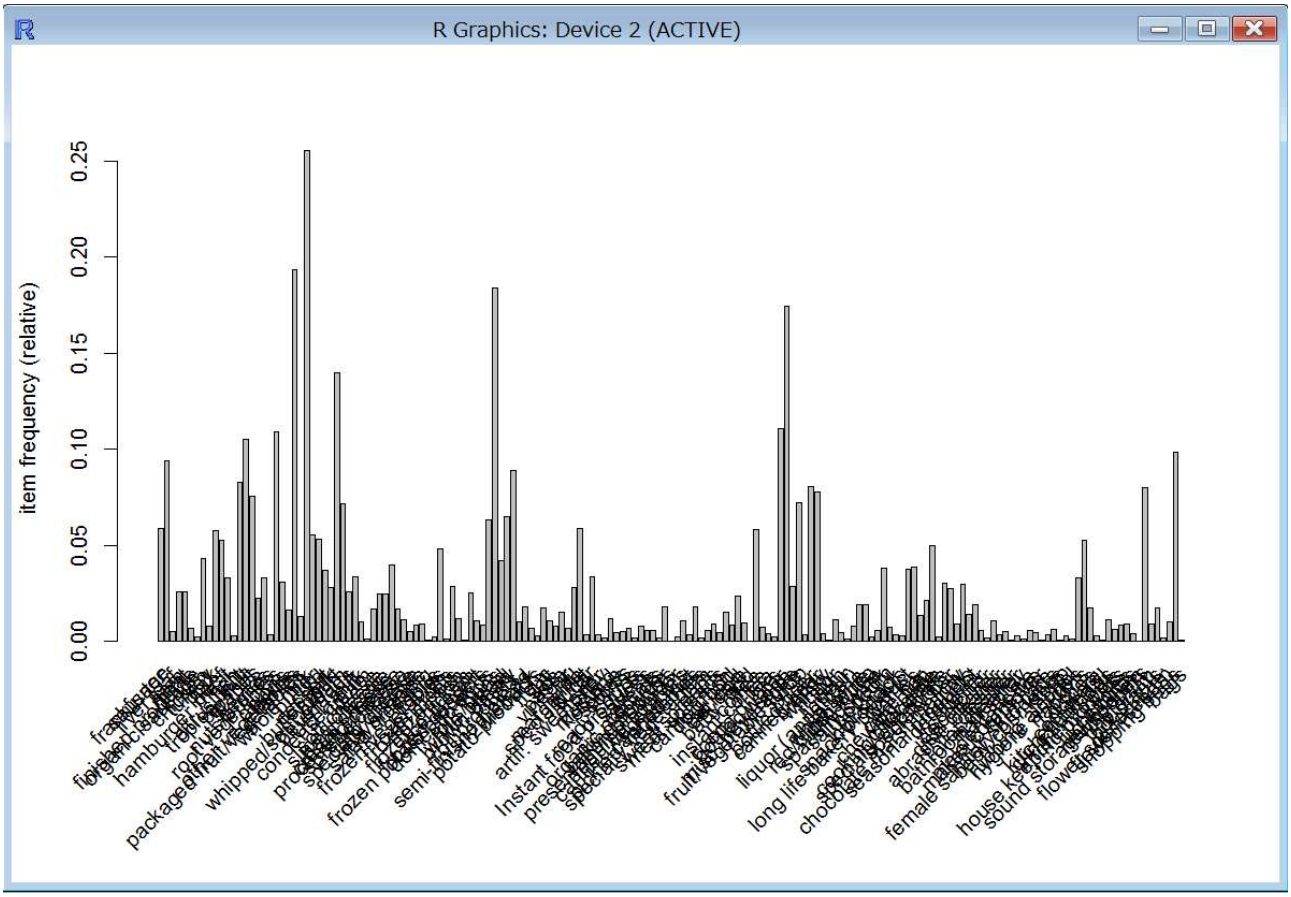
アソシエーション・ルール			支持度 (support)	確信度 (confidence)	リフト (lift)
LHS(条件部)		RHS(結論部)			
タバコ	⇒	缶コーヒー	0.60	1.00	1.25
タバコ	⇒	お菓子	0.50	0.80	1.60
缶コーヒー	⇒	お菓子	0.45	0.80	1.78
…					

- 関連するルールを作る
- 条件部(LHS：Left Hand Side)と結論部(RHS: Right Hand Side)があり、条件と結論が対応
- ルール１：常に1対1とは限らない。たとえば、たばこと缶コーヒーを買っている人は（条件部）、お菓子も買っている（結論部）というケースもありうる
- ルール２：一方向であること。たとえば、たばこ（条件部）⇒缶コーヒー（結論部）と缶コーヒー（条件部）⇒たばこ（結論部）、同じ、たばこ、缶コーヒーを買ったとしても、別モノとして扱う。

アソシエーション分析

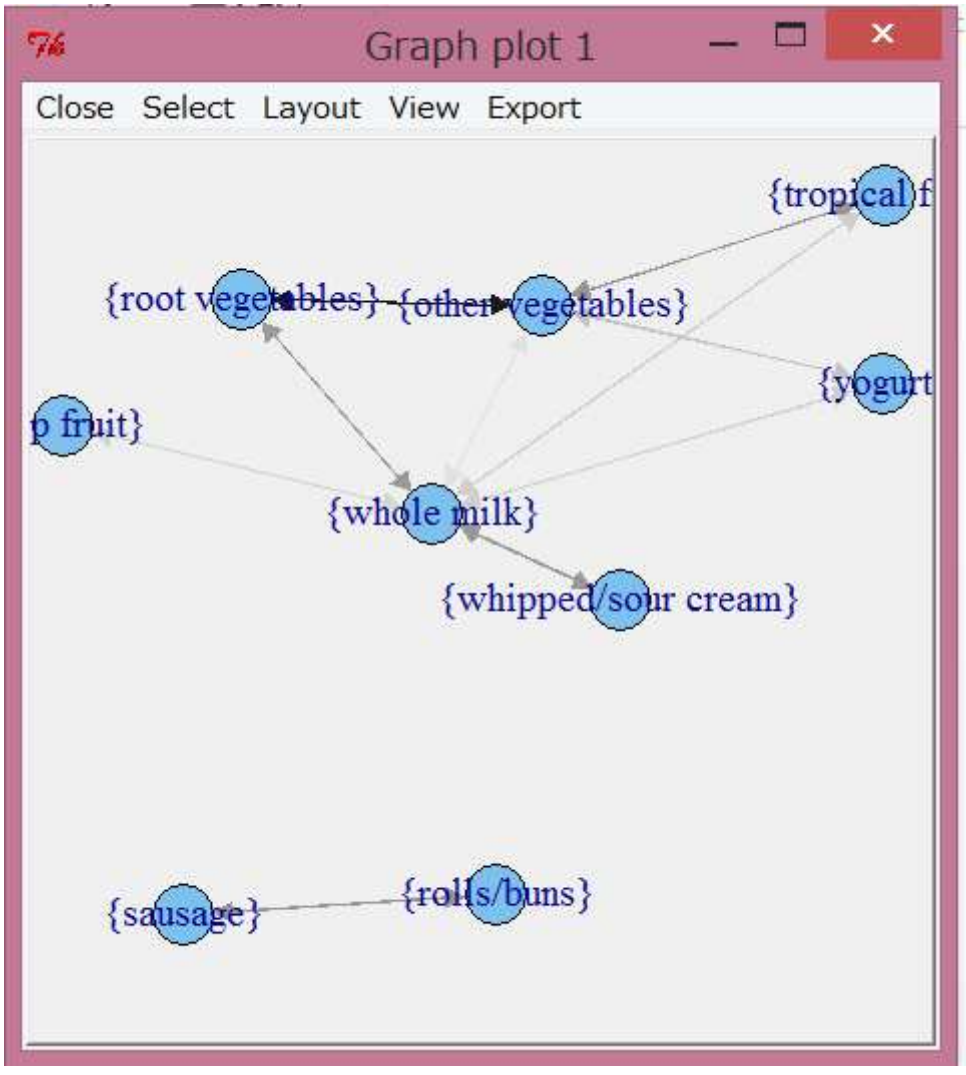
- 頻度をプロット
  - itemFrequencyPlot(Groceries)





アソシエーション分析

- インタラクティブグラフでビジュアル化
  - install.packages(“arulesViz”)
  - library(arulesViz)
  - gruleX = apriori(Groceries, p=list(support=0.03, confidence=0.05, ext=TRUE))
  - gruleX2 = subset(gruleX, subset=(lift>=1.5))
  - plot(gruleX2, method="graph",interactive=TRUE)

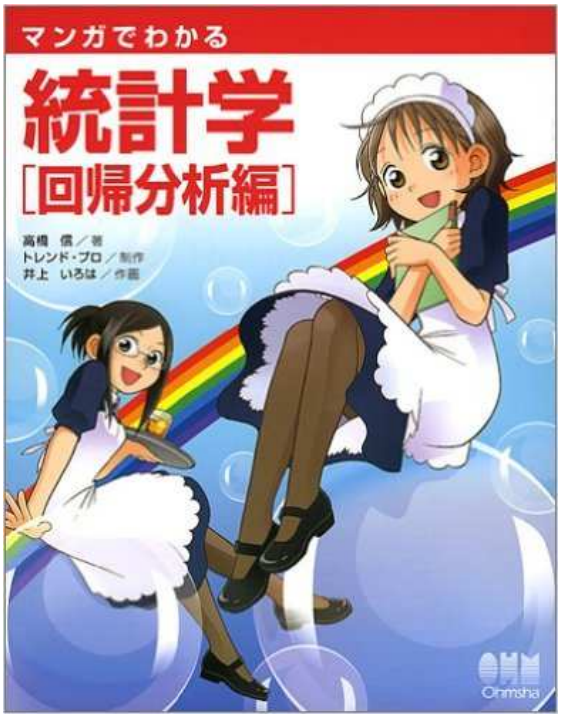


アソシエーション分析

- 演習問題
  - 次のサンプルをもとにアソシエーション分析をしてみましょう

  - data1 = list(c("パン","牛乳","ハム","果物"),c("パン","オムツ","ビール","ハム"),c("ソーセー ジ","ビール","オムツ"),c("弁当","ビール","オムツ","タバコ"),c("弁当","ビール","オレンジジュース","果物"))
  - data.tran = as(data1,"transactions")
  - as(data.tran,"matrix")
  - as(data.tran,"data.frame")

おススメ書籍



「マンガでわかる統計学
回帰分析編」

2005年9月

高橋　信　トレンドプロ（著）

オーム社

　統計学同様に回帰分析、重回帰分析、ロジスティクス回帰まで踏み込んで解説。同様に因子分析もおススメ