

汎用視覚化データ活用環境 (TRAD)

データサイエンスコンソーシアム, 慶應義塾大学

柴田 里程

```
> kyphosis
  Kyphosis Age Number Start
1  FALSE  71      3      5
2  FALSE 158      3     14
3   TRUE 128      4      5
4  FALSE   2      5      1
5  FALSE   1      4     15
6  FALSE   1      2     16
7  FALSE  61      2     17
8  FALSE  37      3     16
9  FALSE 113      2     16
10  TRUE  59      6     12
11  TRUE  82      5     14
12  FALSE 148      3     16
13  FALSE  18      5      2
14  FALSE   1      4     12
15  FALSE 168      3     18
16  FALSE   1      3     16
17  FALSE  78      6     15
18  FALSE 175      5     13
19  FALSE  80      5     16
20  FALSE  27      4      9
21  FALSE  22      2     16
22  TRUE 105      6      5
23  TRUE  96      3     12
24  FALSE 131      2      3
25  TRUE  15      7      2
26  FALSE   9      5     13
27  FALSE   8      3      6
28  FALSE 100      3     14
29  FALSE   4      3     16
30  FALSE 151      2     16
31  FALSE  31      3     16
```

データ全体を見渡せない

説明不足

イメージが湧かない



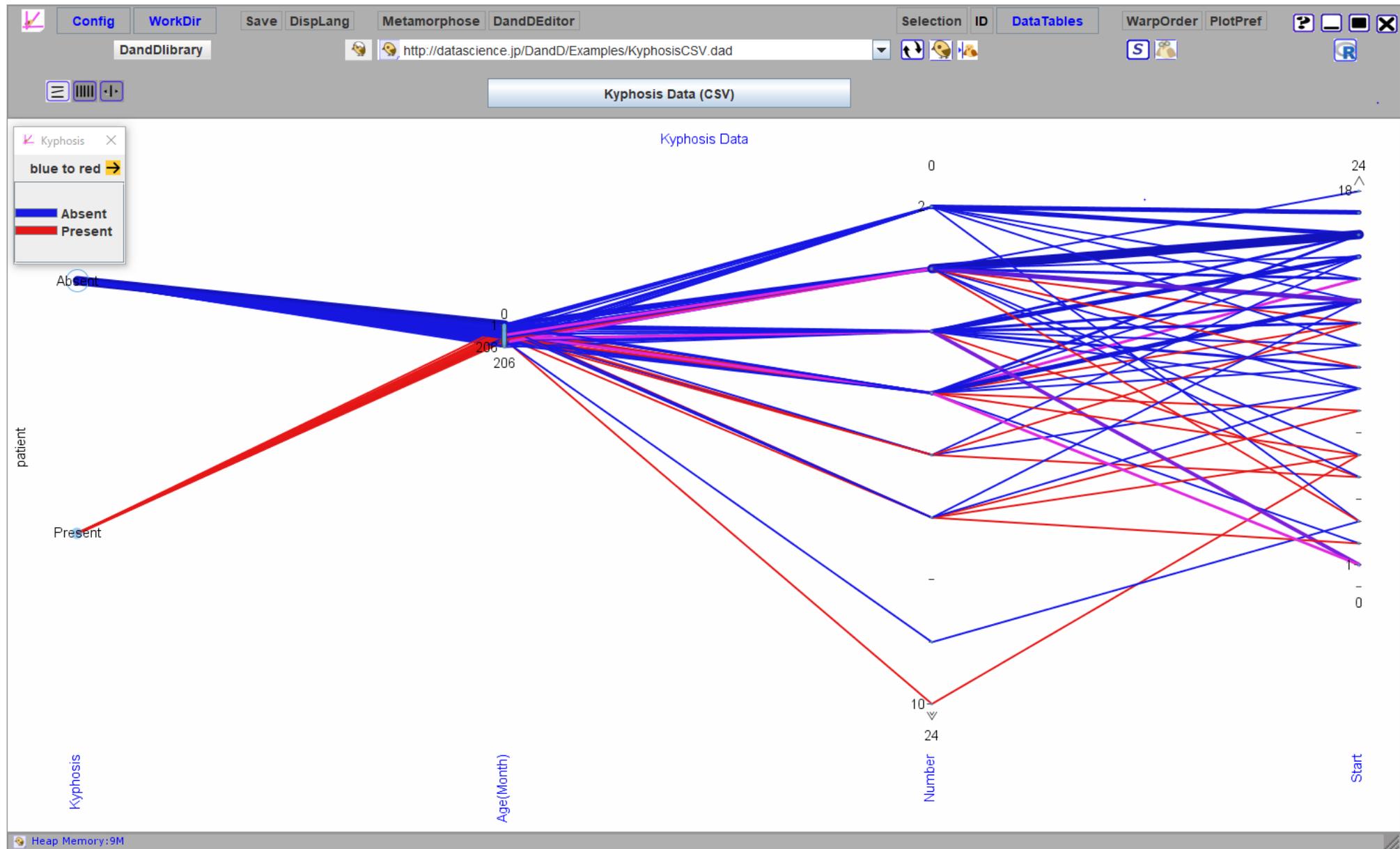
データを眺めることは
単調, 退屈

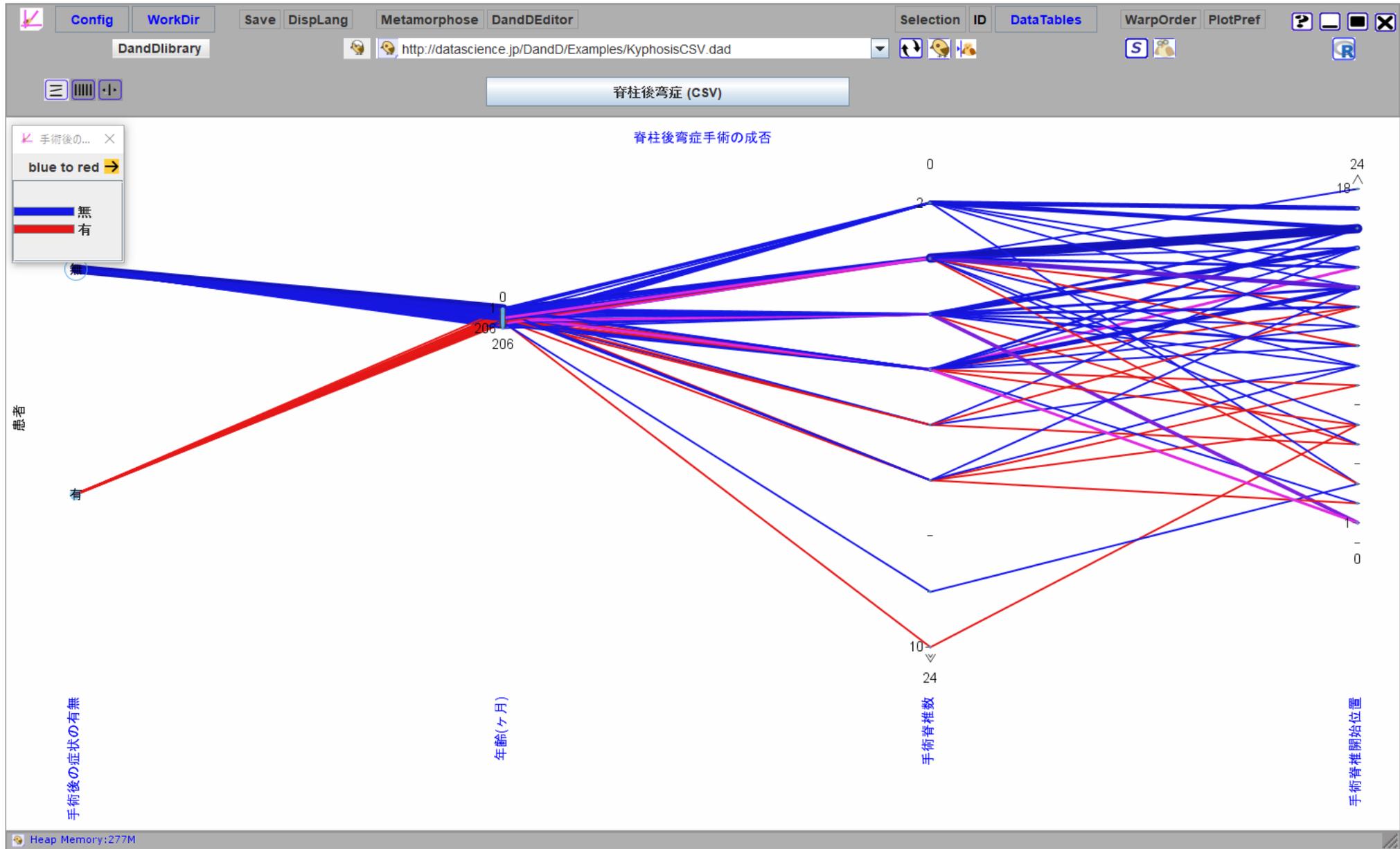
手っ取り早く何か定まった
方法を適用してしまいたい

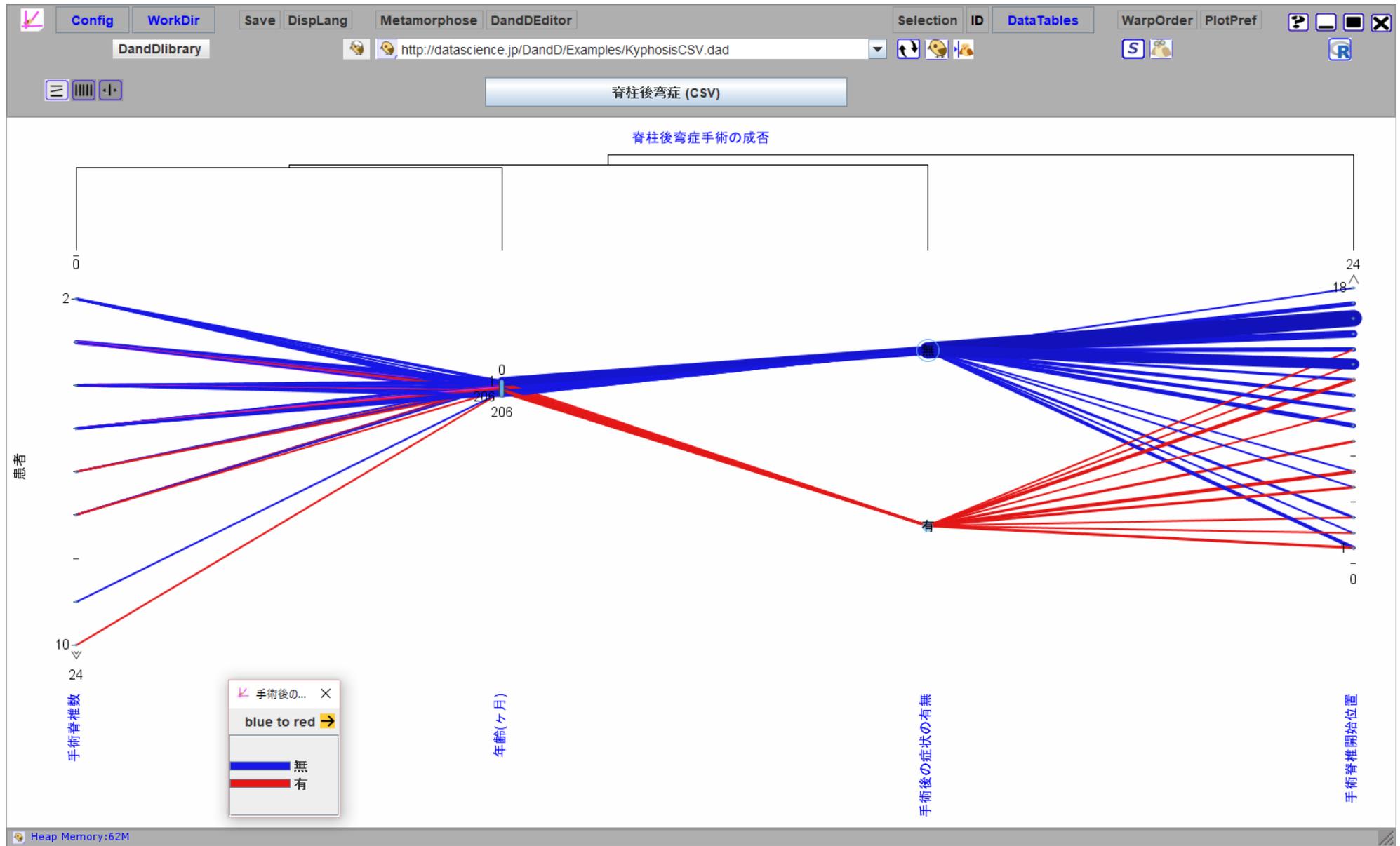
統計学: 方法の学問

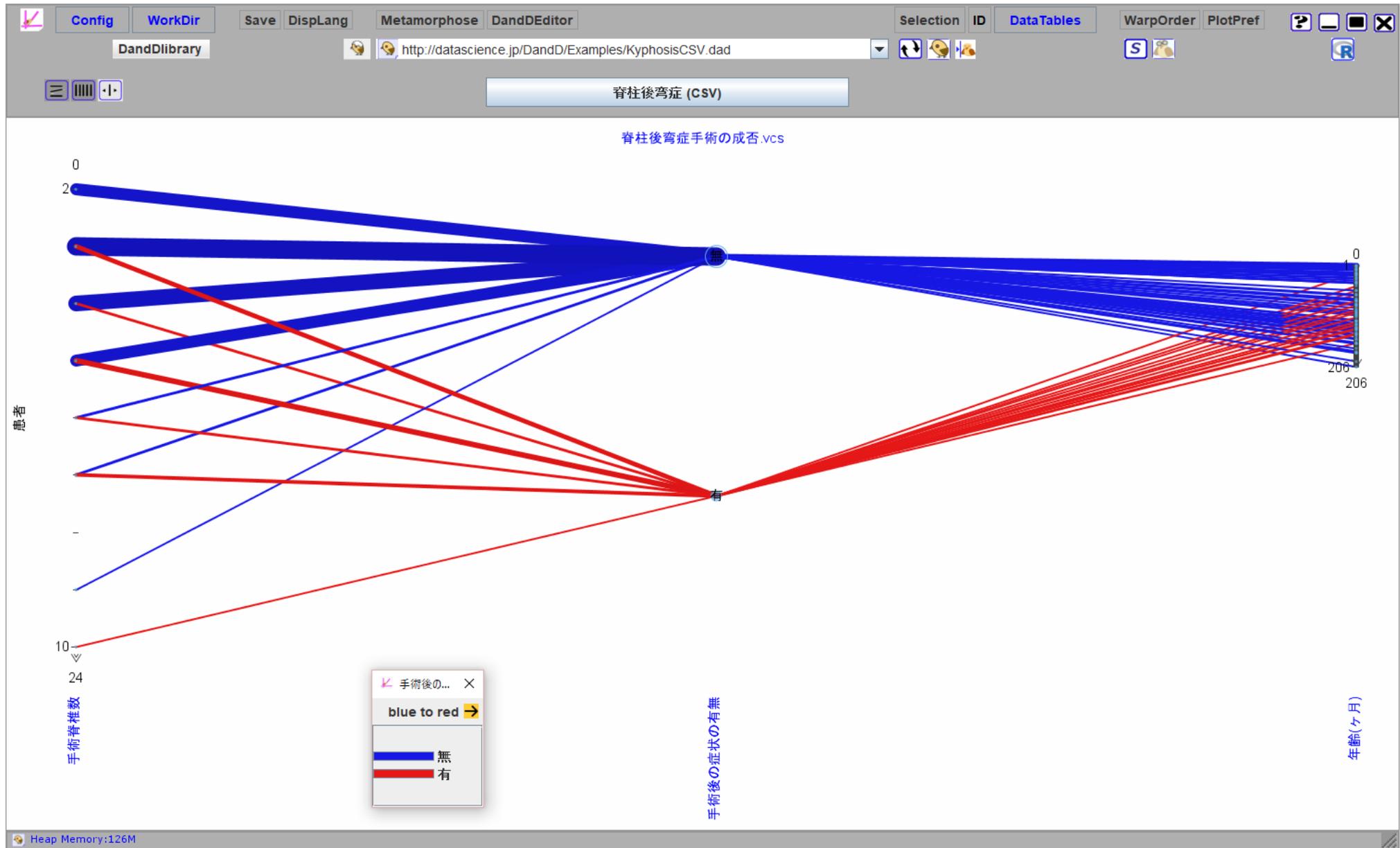
視覚化データ解析環境

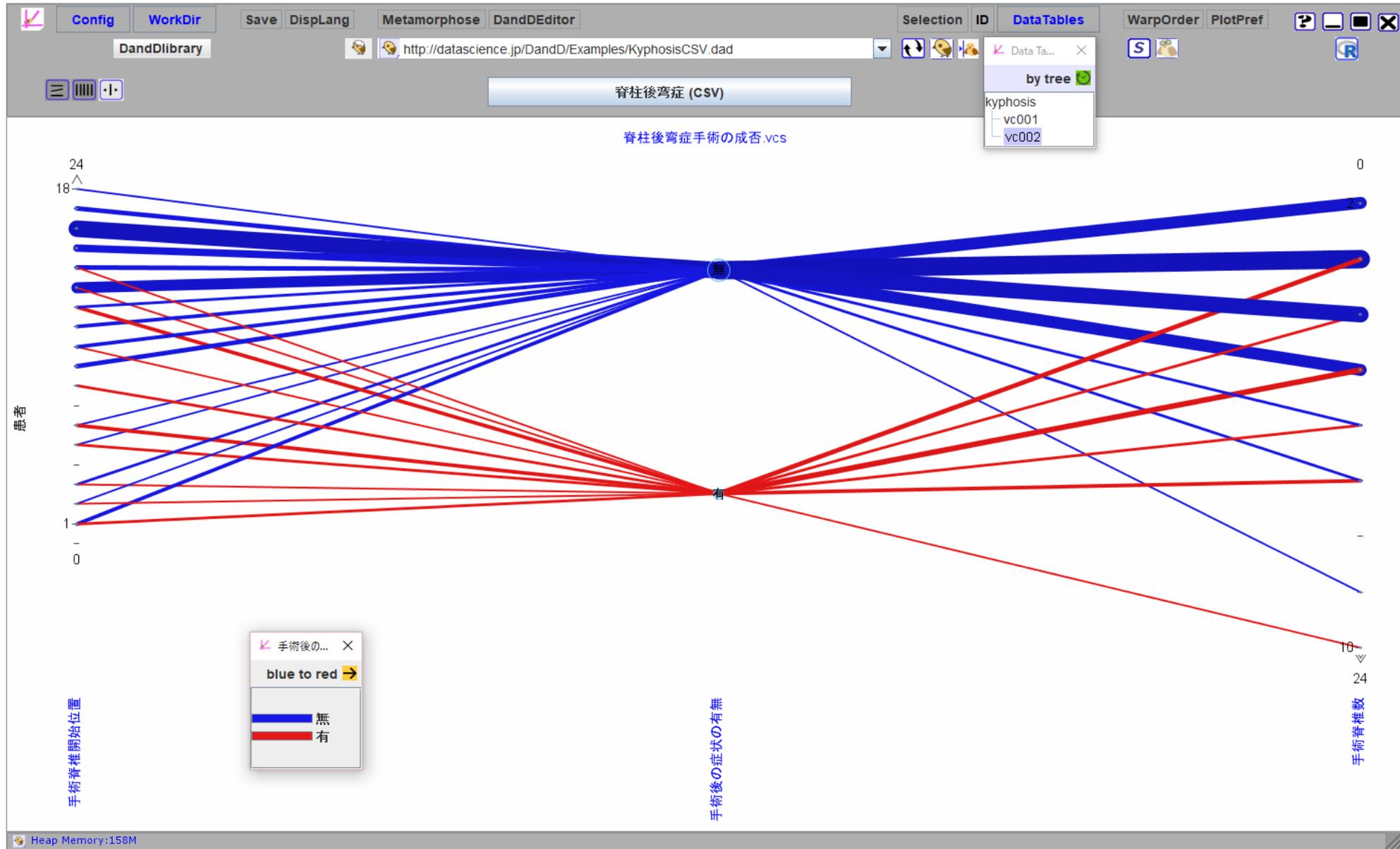
- 必要十分な視覚表現
 - どんな型の変量が混在している多変量データでも一定の形式で視覚表現
 - 特定の用途, モデルに依存しない汎用な視覚表現
- 念入りなヒューマンインタフェース
 - 直感にあったGUI
 - やり直しが容易
 - カスタマイズが容易
- R とのシームレスな連携
 - 役割分担
 - 継ぎ目を意識しないですむインターフェース
- 多国語対応

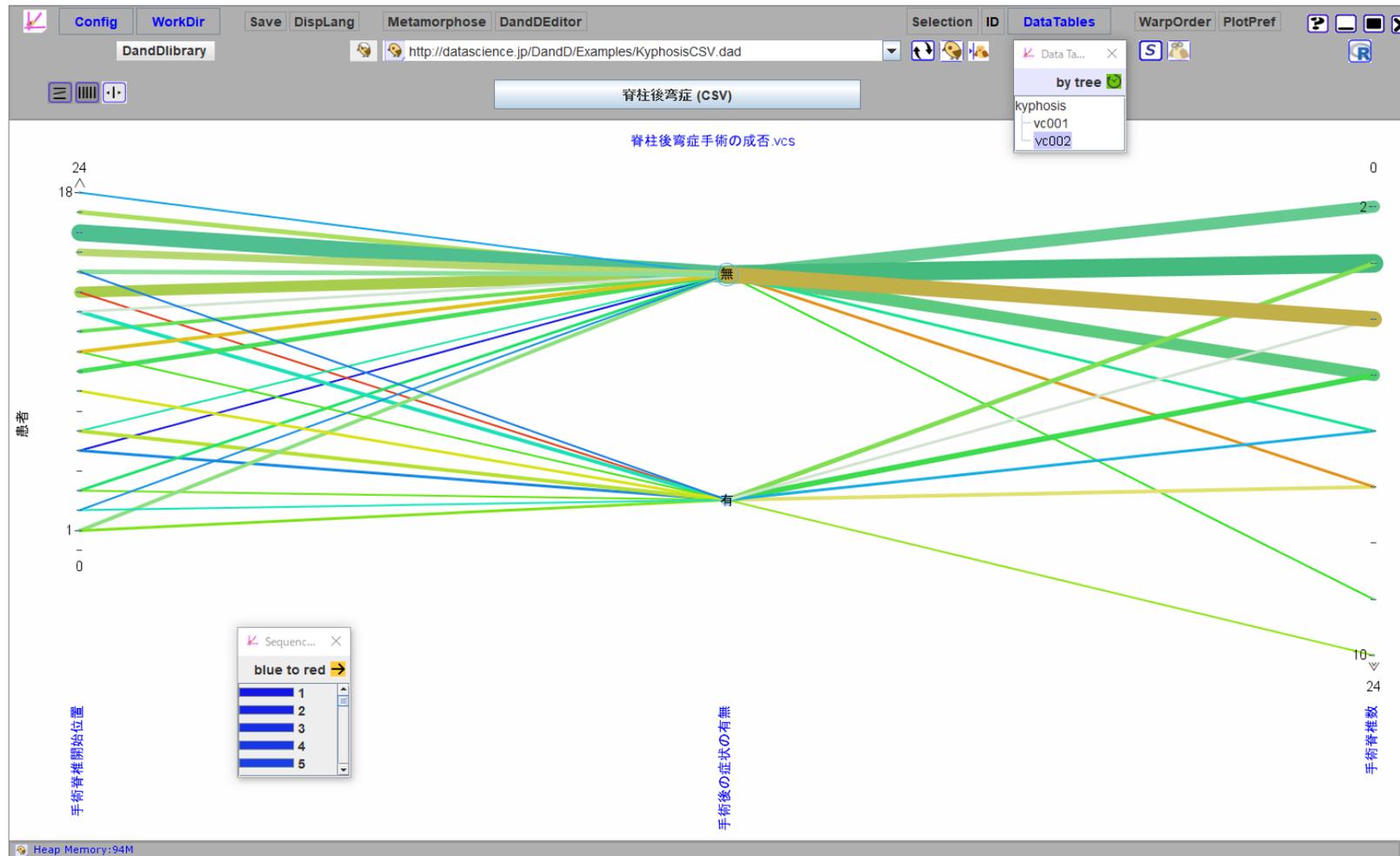




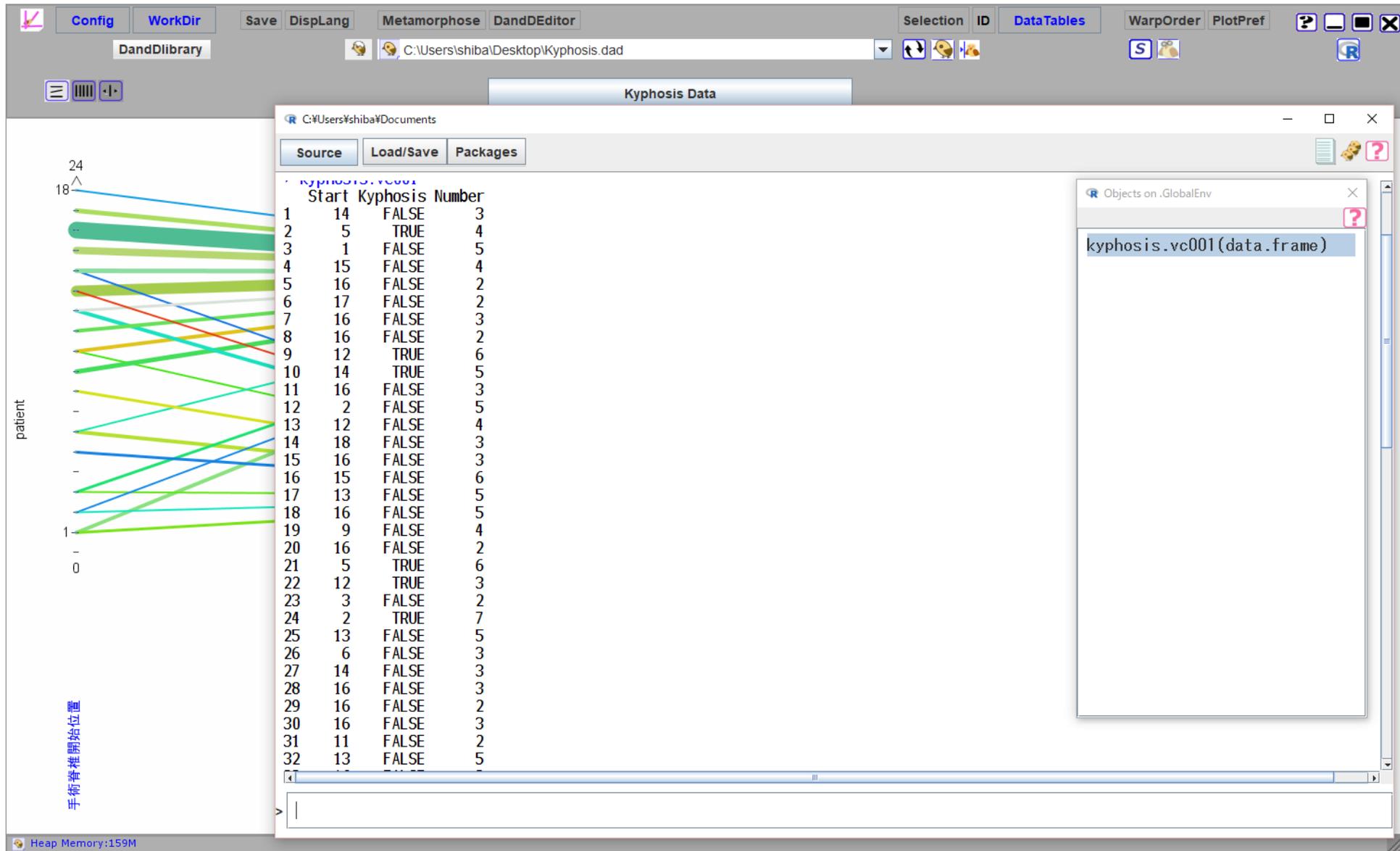








月齢は、手術の成否にほとんど無関係
 手術した脊柱の箇所、個数で成否は左右される
 箇所が、15-18 なら必ず成功している
 脊柱2個なら必ず成功している



Config WorkDir Save DispLang Metamorphose DandEditor Selection ID DataTables WarpOrder PlotPref

DandLibrary C:\Users\shiba\Desktop\Kyphosis.dad

R Work

C:\Users\shiba\Documents

Source Load/Save Packages

Row	Start	Target?	Number
07	13	FALSE	4
68	11	FALSE	4
69	16	FALSE	5
70	14	FALSE	5
71	12	FALSE	4
72	16	FALSE	4
73	10	FALSE	4
74	15	FALSE	3
75	15	FALSE	4
76	13	TRUE	3
77	13	FALSE	7
78	13	FALSE	2
79	6	TRUE	7
80	13	FALSE	4

Target?

1

18

Start

```

> glm.result=glm(Kyphosis~., family="binomial")
Error in terms.formula(formula, data = data) :
  '.' が式中にありますが、'data' 引数がありません
> glm.result=glm(Kyphosis~., family="binomial", data=kyphosis.vc001)
> glm.result

Call: glm(formula = Kyphosis ~ ., family = "binomial", data = kyphosis.vc001)

Coefficients:
(Intercept)      Start      Number
   -0.8332    -0.1914    0.3343

Degrees of Freedom: 79 Total (i.e. Null); 77 Residual
Null Deviance: 82.76
Residual Deviance: 63.82 AIC: 69.82
>
>
> kyphosis.vc001=cbind(kyphosis.vc001, fitted(glm.result))
> txp(kyphosis.vc001)

```

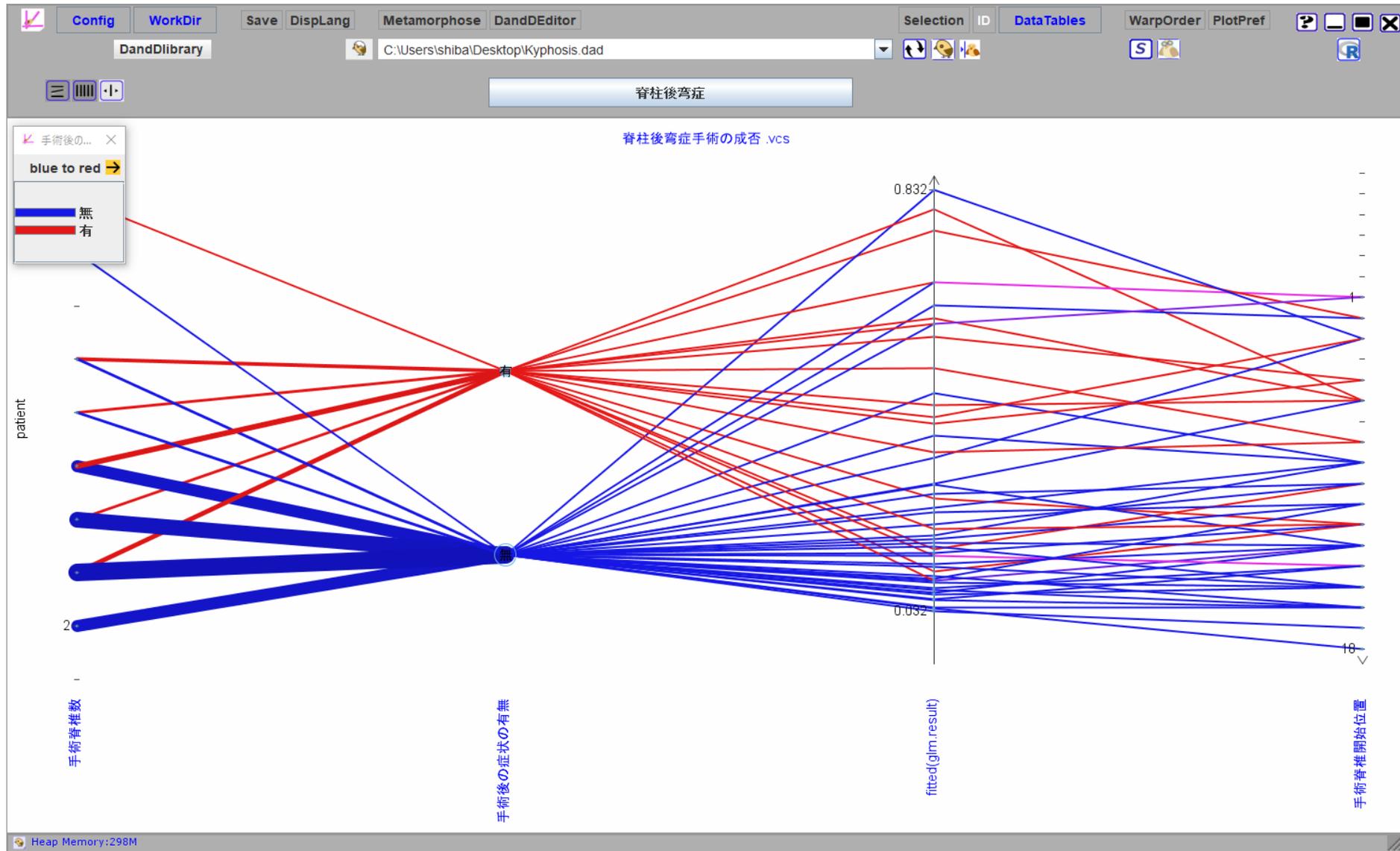
Objects on .GlobalEnv

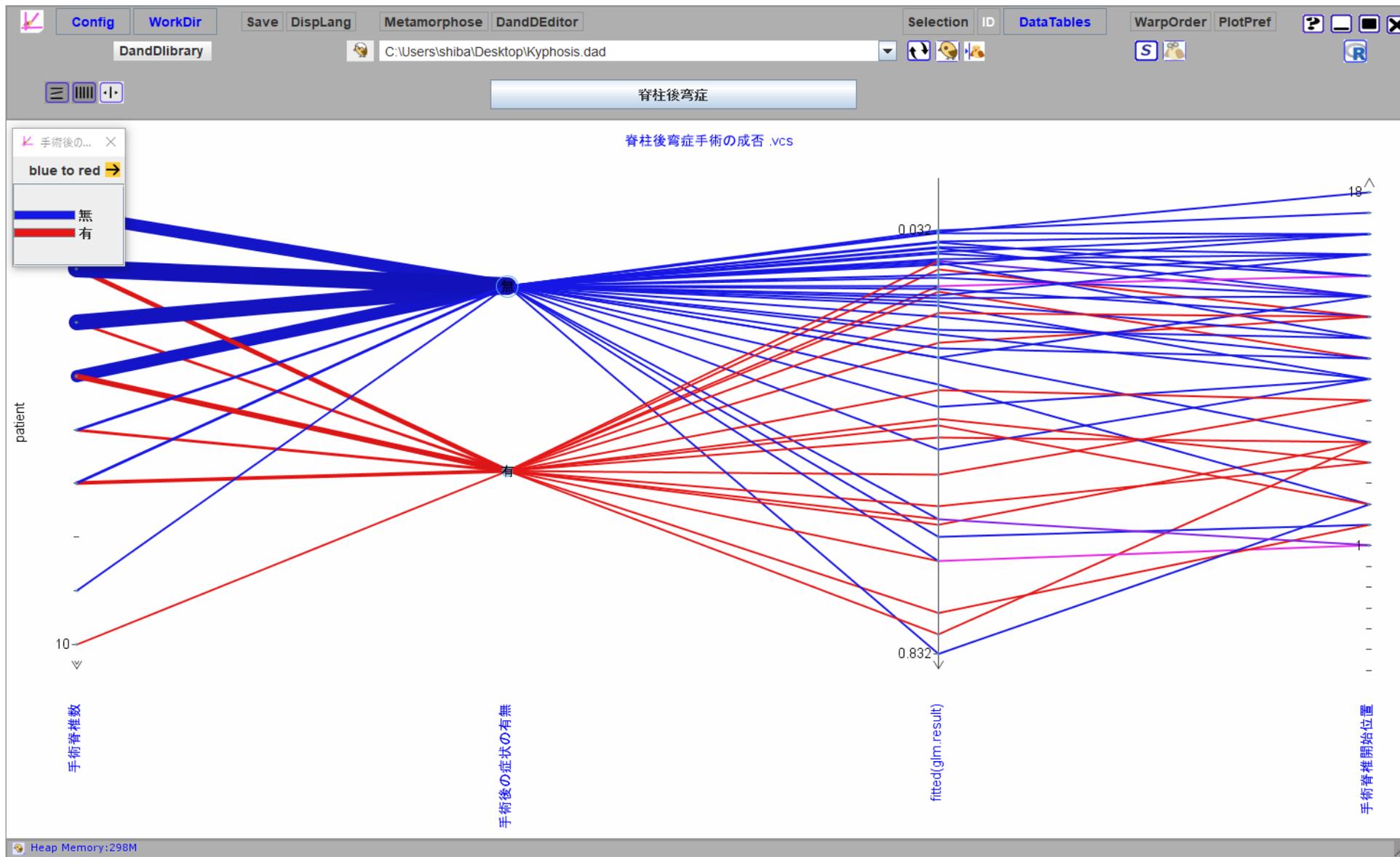
```

glm.result (glm,lm)
kypho.R (data.frame)
kyphosis.vc001 (data.frame)

```

Heap Memory:148M





Config WorkDir Save DispLang Metamorphose DandDEditor Selection ID DataTables WarpOrder PlotPref

DandDlibrary C:\Users\shiba\Desktop\Kyphosis.dad

R Work

Source Load/Save Packages

69	10	FALSE	3
70	14	FALSE	5
71	12	FALSE	4
72	16	FALSE	4
73	10	FALSE	4
74	15	FALSE	3
75	15	FALSE	4
76	13	TRUE	3
77	13	FALSE	7
78	13	FALSE	2
79	6	TRUE	7
80	13	FALSE	4

```

> glm.result=glm(Kyphosis~., family="b
Error in terms.formula(formula, data =
'.' が式中にありますが、'data' 引数
> glm.result=glm(Kyphosis~., family="b
> glm.result

Call: glm(formula = Kyphosis ~ ., fan

Coefficients:
(Intercept)      Start      Number
      -0.8332      -0.1914      0.3343

Degrees of Freedom: 79 Total (i.e. Null
Null Deviance:      82.76
Residual Deviance: 63.82      AIC: 6

>
>
> kyphosis.vc001=cbind(kyphosis.vc001,
> txp(kyphosis.vc001)
>

```

R object editor: txp.data.frame

```

1 function(x, new=FALSE) {
2 x=deparse(substitute(x))
3 attr(x, "new")=new
4 assign(".Out", x, pos=1)
5 }

```

Write on txp.data.frame

Heap Memory:282M

Config WorkDir Save DispLang Metamorphose DandDEditor Selection ID DataTables WarpOrder PlotPref

DandDlibrary C:\Users\shiba\Desktop\Kypnosis.dad

R Work

C:\Users\shiba\Documents

Source Load/Save Packages

Installed packages

Attached	Package	Description
<input type="checkbox"/>	Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript)
<input type="checkbox"/>	JGR	Java GUI for R
<input type="checkbox"/>	JavaGD	Java Graphics Device
<input type="checkbox"/>	KernSmooth	Functions for Kernel Smoothing Supporting Wand & Jones (1995)
<input type="checkbox"/>	MASS	Support Functions and Datasets for Venables and Ripley's MASS
<input type="checkbox"/>	Matrix	Sparse and Dense Matrix Classes and Methods
<input checked="" type="checkbox"/>	base	The R Base Package
<input type="checkbox"/>	boot	Bootstrap Functions (Originally by Angelo Canty for S)
<input type="checkbox"/>	class	Functions for Classification
<input type="checkbox"/>	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.
<input type="checkbox"/>	codetools	Code Analysis Tools for R
<input type="checkbox"/>	compiler	The R Compiler Package
<input checked="" type="checkbox"/>	datasets	The R Datasets Package
<input type="checkbox"/>	foreign	Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...
<input checked="" type="checkbox"/>	grDevices	The R Graphics Devices and Support for Colours and Fonts
<input checked="" type="checkbox"/>	graphics	The R Graphics Package
<input type="checkbox"/>	grid	The Grid Graphics Package
<input type="checkbox"/>	lattice	Trellis Graphics for R
<input checked="" type="checkbox"/>	methods	Formal Methods and Classes
<input type="checkbox"/>	mgcv	Mixed GAM Computation Vehicle with Automatic Smoothness Estimation
<input type="checkbox"/>	nlme	Linear and Nonlinear Mixed Effects Models
<input type="checkbox"/>	nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models
<input type="checkbox"/>	parallel	Support for Parallel computation in R
<input type="checkbox"/>	rJava	Low-Level R to Java Interface
<input type="checkbox"/>	rpart	Recursive Partitioning and Regression Trees
<input type="checkbox"/>	spatial	Functions for Kriging and Point Pattern Analysis
<input type="checkbox"/>	survival	Regression Survival Functions and Classes

OK No

Heap Memory: 282M

基本

- DandDインスタンス
 - データテーブル
 - データベクトル
 - 値
- データソース
 - テキスト, CSV, Excel (複数シート)
 - RDB (SQL アクセス)
- 視覚化インタフェース
 - TextilePlot
 - 位置, 尺度の変換のみ
 - ノット (他と直交する変量)
- 軸(Warp)のクラスタリング
 - 軸どうしの距離を水平性規準からのずれで定めている
 - ノットは除外して考えたほうがよい
- 『データ分析とデータサイエンス』 近代科学社, 2015 にも例がある

設計のポイント

- 型の役割
- 欠損値の扱い (NA, NaN, \cdot , ..., -)
- 不定形 (コメント, 不完全な行, ...)
- マニピュレーション (複数回答, 階層構造をもつラベル, インデントで区別されたあたい, 基数系)
- ブラウジング
- データベクトルの区分
 - ID
 - Main
 - Aux
- 実用性
 - 規模とスピード
 - 安定性
- 楽しさ, 美しさ

型の役割

- よくある値の型

- 論理, 整数, 実数, 文字, 文字列
- 論理, 数, 文字列 (R)
- 非数値 (カテゴリーカル), 数値

- TRADにおける値の型

- Measurement
- Cardinal
- Ordinal
- Frequency
- Mark
- Ordered Mark
- Logical
- Time
- Elapsed Time

Numerical data		Non-numerical data		
Continuous	Discrete	Ordered	Unordered	Logical

- Measurement

- Ordinal
- Cardinal $\hat{\wedge}$

Config WorkDir Save DispLang Metamorphose DandEditor Selection ID DataTables WarpOrder PlotPref

DandLibrary C:\Users\shiba\Desktop\Kyphosis.dad

Kyphosis Data

脊柱後彎症手術の有無

patient

無

有

手術後の症状の有無

年齢(ヶ月)

手術脊椎数

手術脊椎開始位置

Attributes of Data Vectors

Id	Kyphosis	Age	Number	Start
ShortName(AlphaNumeric)	Kyphosis	Age	Number	Start
LongName	手術後の症状の有無	年齢	手術脊椎数	手術脊椎開始位置
Unit		ヶ月		
DataType	Logical	Ordinal	Cardinal	Ordinal
Code	2			
Pad NA				

ShortNames by a String (Ctrl+V & Return) Kyphosis Age Number Start

LongNames by a String (Ctrl+V & Return) 手術後の症状の有無 年齢 手術脊椎数 手術脊椎開始位置

Units by a String (Ctrl+V & Return) ヶ月

Ok No

0 2 18 24

10 24

0

Heap Memory:9M

Config WorkDir Save DispLang Metamorphose DandDEditor Selection ID DataTables WarpOrder PlotPref

DandDlibrary C:\Users\shiba\Desktop\Kyphosis.dad

Kyphosis Data

脊柱後弯症手術の成否

patient

無

有

手術後の症状の有無

年齢(ヶ月)

手術脊椎数

手術脊椎開始位置

Attributes of Data Vectors

	Kyphosis	Age	Number	Start
Id	Kyphosis	Age	Number	Start
ShortName(AlphaNumeric)	Kyphosis	Age	Number	Start
LongName	手術後の症状の有無	年齢	手術脊椎数	手術脊椎開始位置
Unit		ヶ月		
Data Type	Logical	Measurement	Measurement	Measurement
Code	2			
Pad NA				

ShortNames by a String (Ctrl+V & Return) Kyphosis Age Number Start

LongNames by a String (Ctrl+V & Return) 手術後の症状の有無 年齢 手術脊椎数 手術脊椎開始位置

Units by a String (Ctrl+V & Return) ヶ月

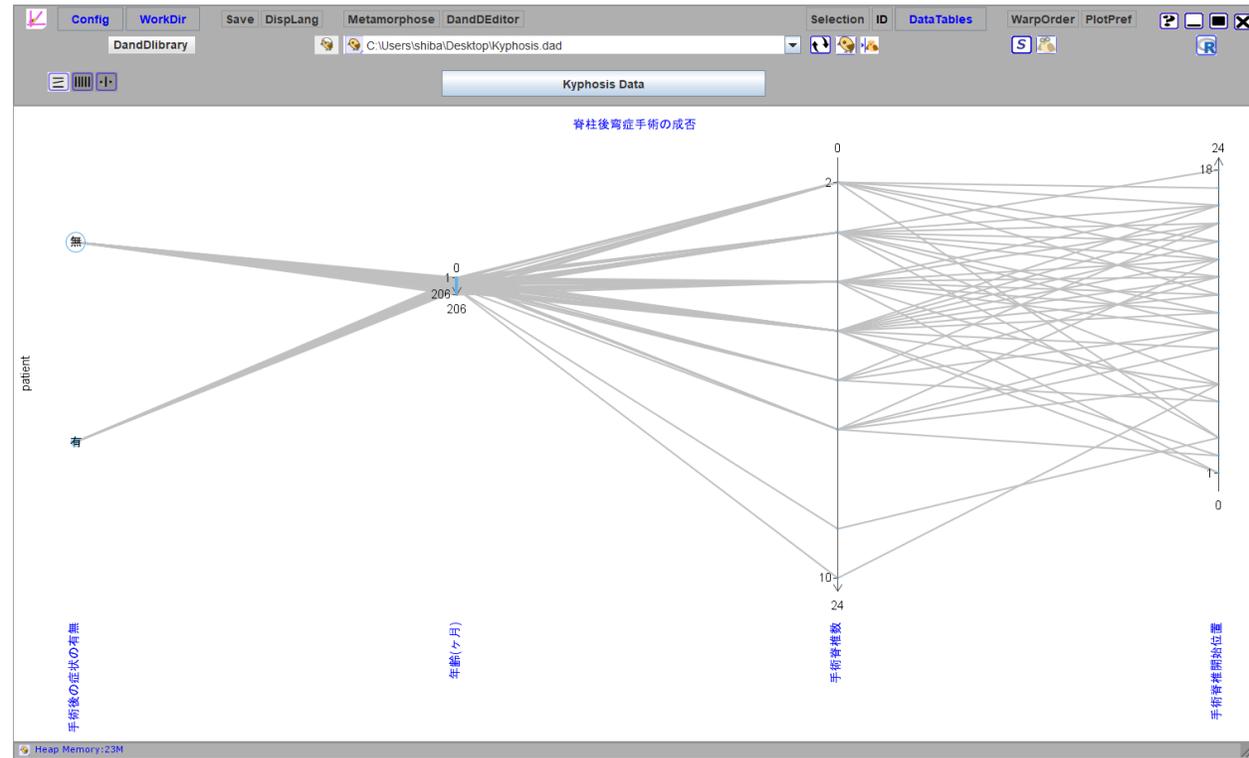
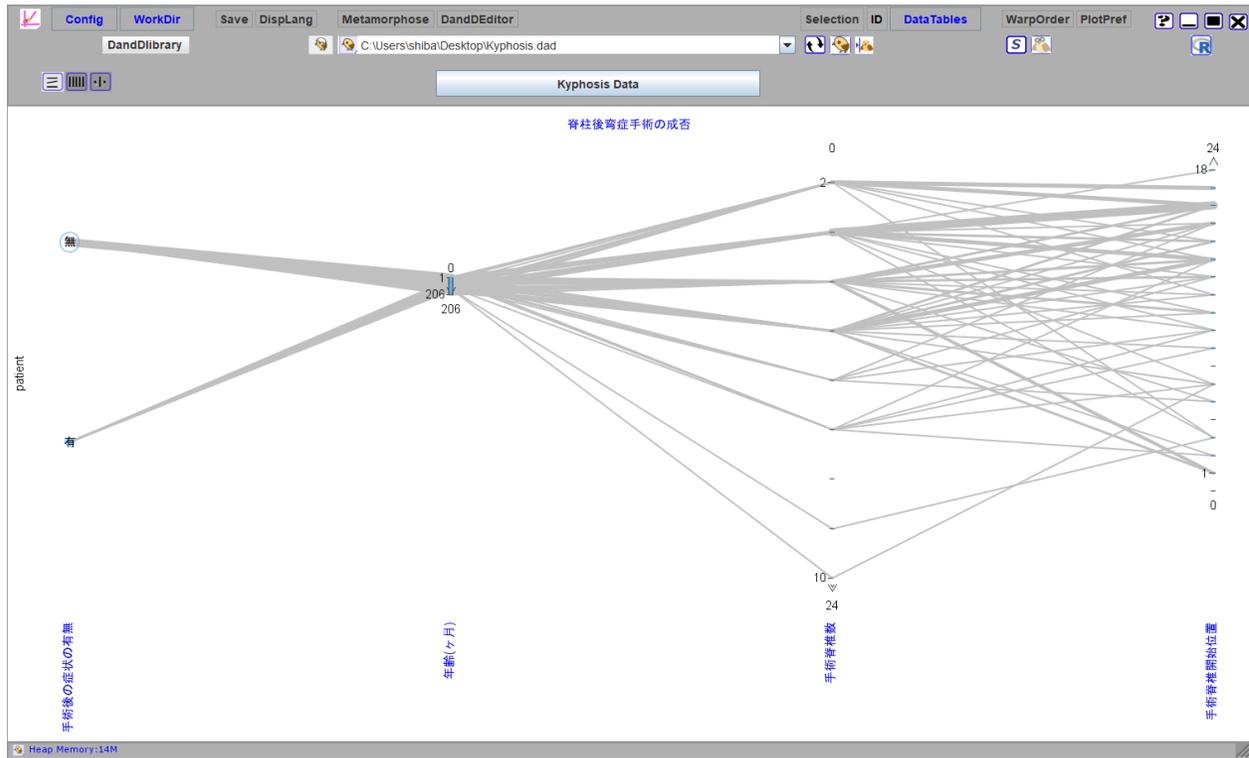
Ok No

0 2 18 24

0 10 24

0 1

Heap Memory:23M



Attributes of Data Vectors

Id	Kyphosis	Age	Number	Start
ShortName(AlphaNumeric)	Kyphosis	Age	Number	Start
LongName	手術後の症状の有無	年齢	手術脊椎数	手術脊椎開始位置
Unit		ヶ月		
DataType	Logical	Ordinal	Cardinal	Ordinal
Code	2			
Pad NA				

ShortNames by a String (Ctrl+V & Return) Kyphosis Age Number Start

LongNames by a String (Ctrl+V & Return) 手術後の症状の有無 年齢 手術脊椎数 手術脊椎開始位置

Units by a String (Ctrl+V & Return) ヶ月

OK No

Attributes of Data Vectors

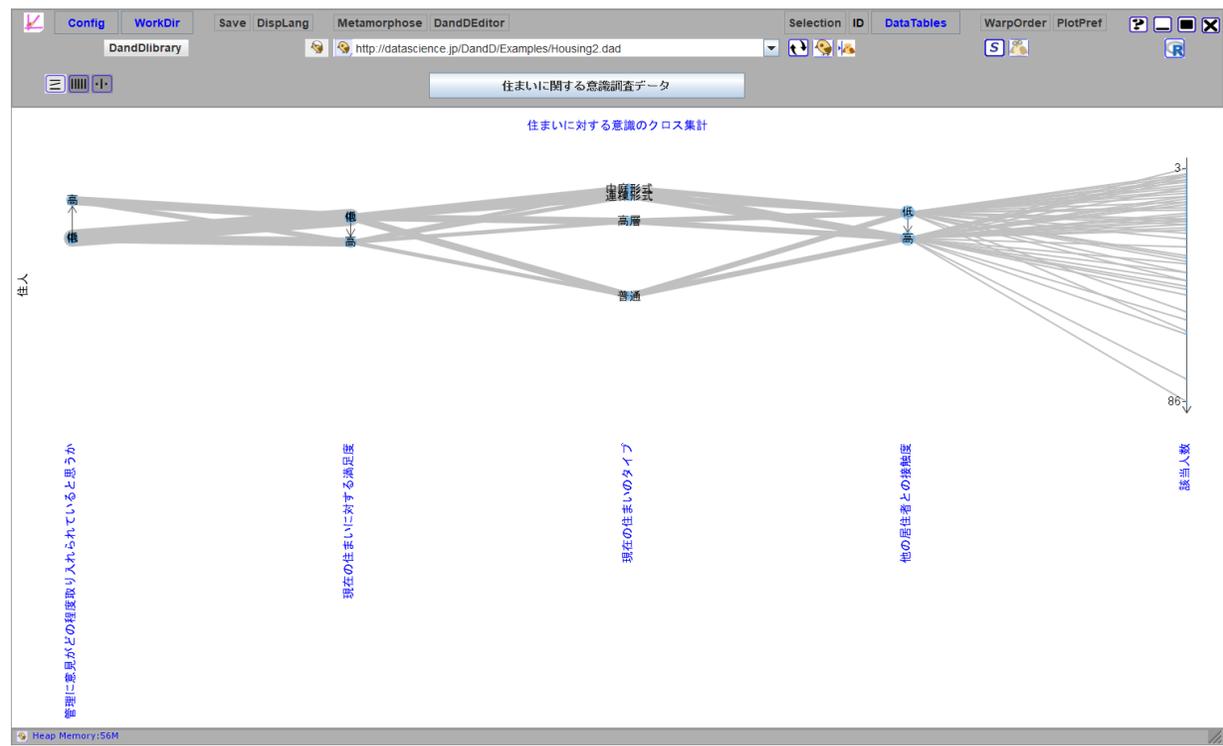
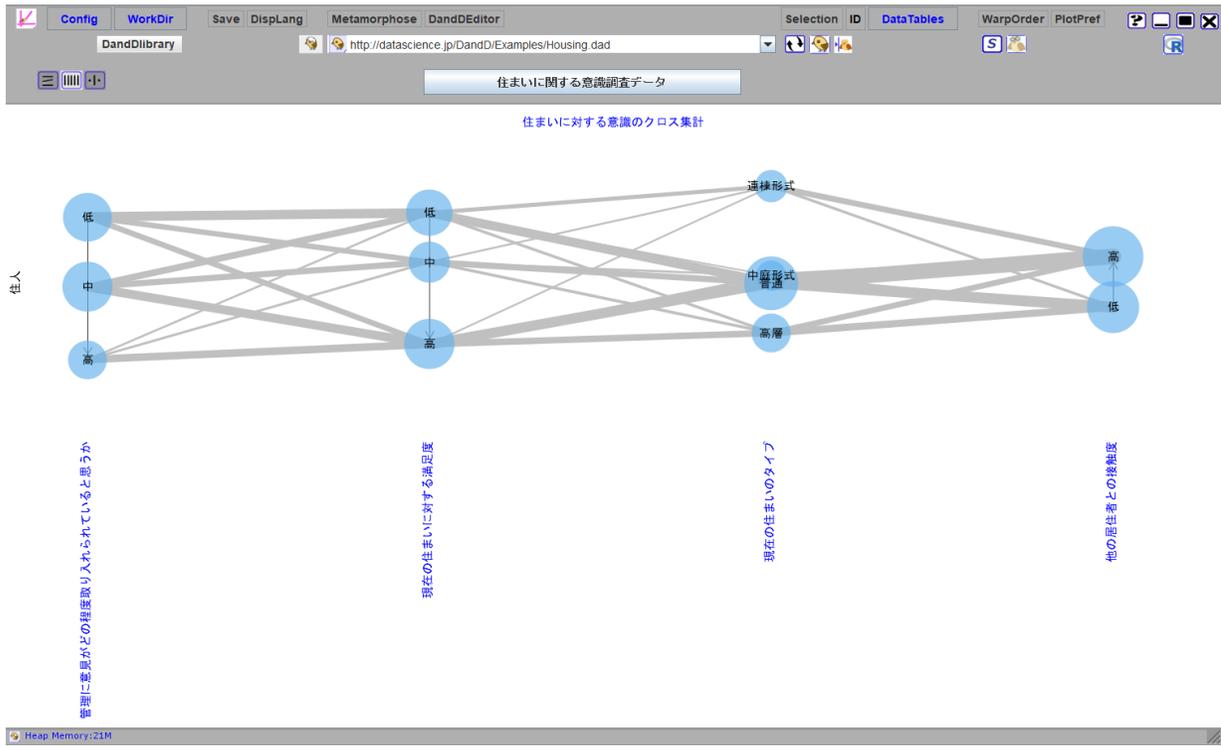
Id	Kyphosis	Age	Number	Start
ShortName(AlphaNumeric)	Kyphosis	Age	Number	Start
LongName	手術後の症状の有無	年齢	手術脊椎数	手術脊椎開始位置
Unit		ヶ月		
DataType	Logical	Measurement	Measurement	Measurement
Code	2			
Pad NA				

ShortNames by a String (Ctrl+V & Return) Kyphosis Age Number Start

LongNames by a String (Ctrl+V & Return) 手術後の症状の有無 年齢 手術脊椎数 手術脊椎開始位置

Units by a String (Ctrl+V & Return) ヶ月

OK No



Attributes of Data Vectors

influence	satisfaction	type	contact	n.person
influence	satisfaction	type	contact	nperson
Influence	Satisfaction	Type of housing	Contact level	Number of persons
Ordered Mark	Ordered Mark	Mark	Ordered Mark	Frequency
3	3	4	2	

ShortNames by a String (Ctrl+V & Return) influence satisfaction type contact nperson

LongNames by a String (Ctrl+V & Return) e Satisfaction Type of housing Contact level Number of persons

Units by a String (Ctrl+V & Return)

Ok No

Attributes of Data Vectors

influence	satisfaction	type	contact	n.person
influence	satisfaction	type	contact	nperson
管理に意見がどの程度取り入れられていると思うか	現在の住まいに対する満足度	現在の住まいのタイプ	他の居住者との接触度	該当人数
Ordered Mark	Ordered Mark	Mark	Ordered Mark	Measurement
3	3	4	2	

ShortNames by a String (Ctrl+V & Return) influence satisfaction type contact nperson

LongNames by a String (Ctrl+V & Return) いに対する満足度 現在の住まいのタイプ 他の居住者との接触度 該当人数

Units by a String (Ctrl+V & Return)

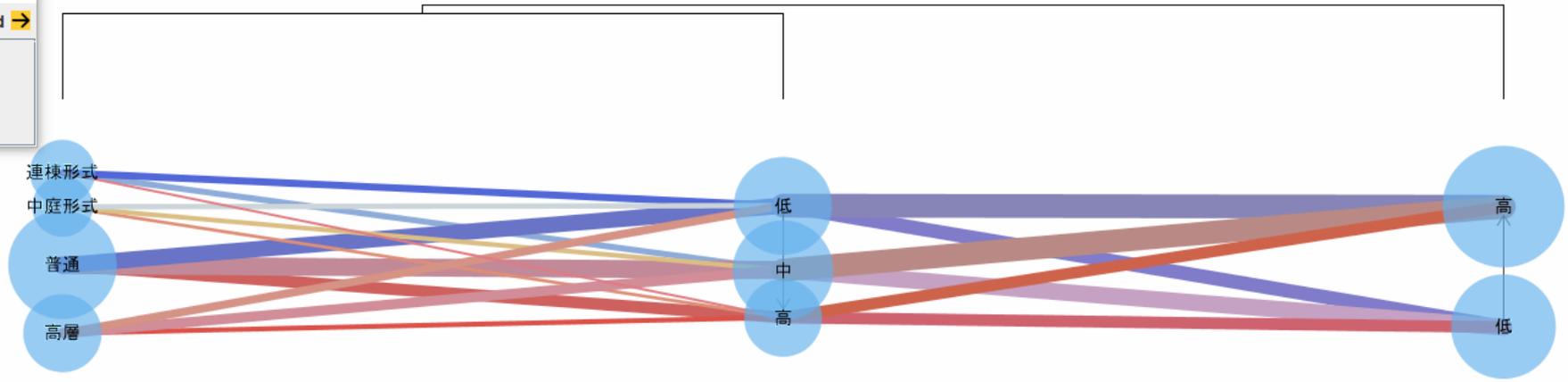
Ok No

住まいに関する意識調査データ

住まいに対する意識のクロス集計 .vcs

現在の住... X
 blue to red →
 低
 中
 高

住人



現在の住まいのタイプ

Wefts colored by X

- Sequence No.
- 現在の住まいに対する満足度
- 管理に意見がどの程度取り入れられているか

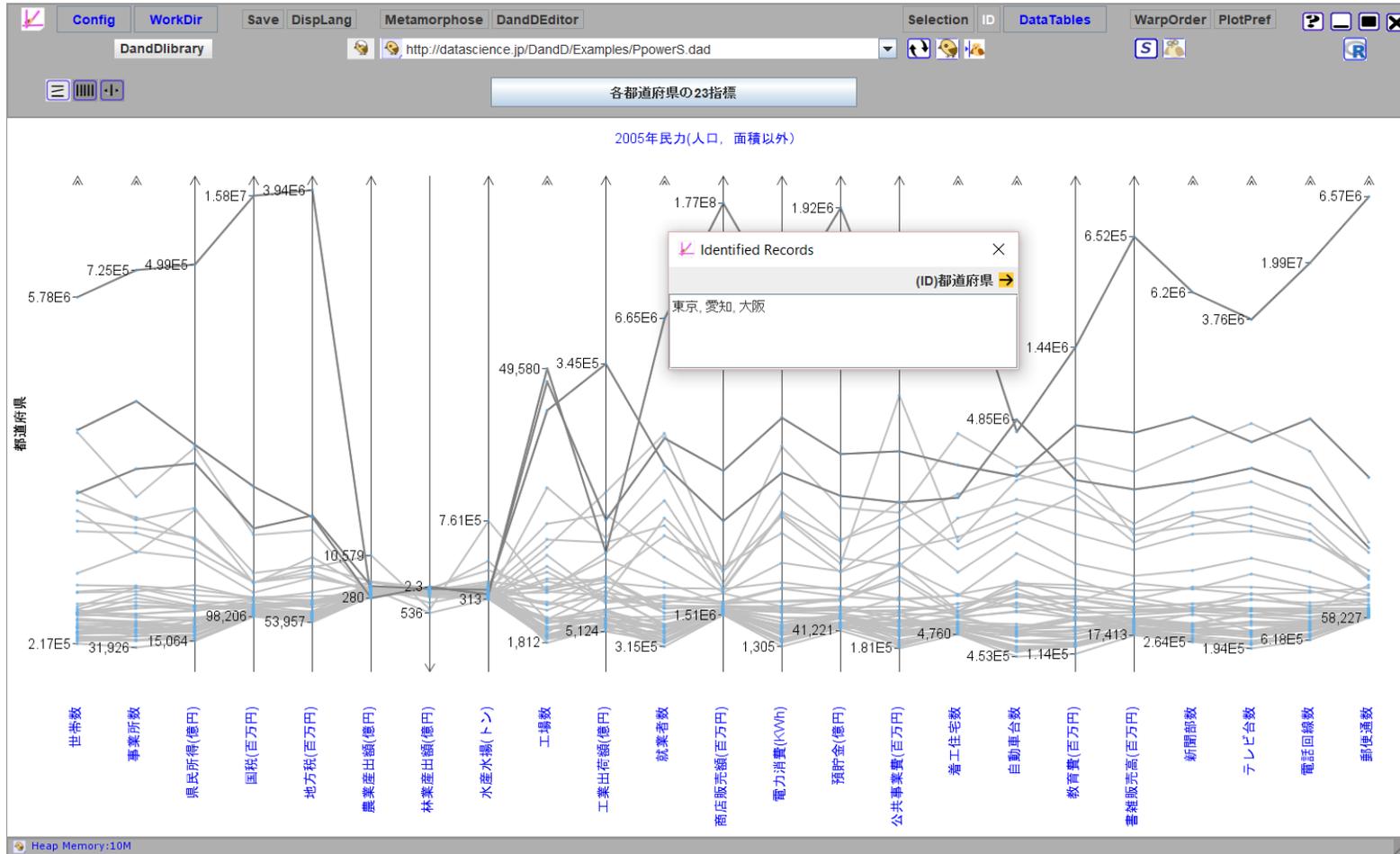
No Color

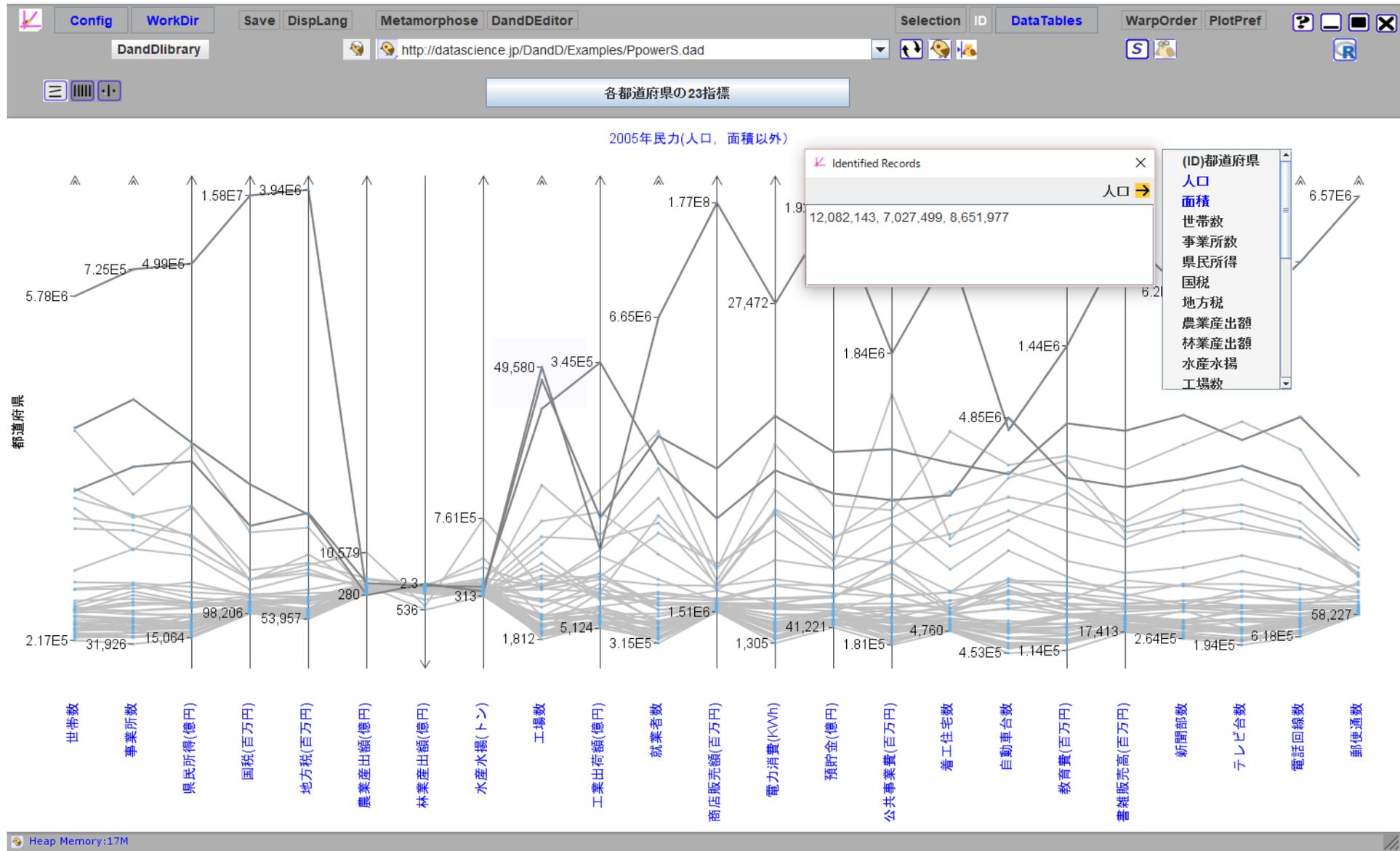
管理に意見がどの程度取り入れられているか

コペンハーゲンのある地域で1960年から1968年に渡り 賃貸住宅の居住者1680人について意識調査を行った結果である。管理についての意見がどの程度聞き入れられているか、その住宅に満足しているかどうか、居住している住宅のタイプ、隣人と接触があるかどうかによって居住者が分類され、その 該当数が分割表の形にまとめられている。「現在の住まいのタイプ」の TowerBlocks は高層住宅、Apartments は中低層住宅、AtriumHouses は中庭つきのマンション、TerracedHouses は連棟式のいわゆるテラスハウスである。

他の居住者との接触度

ブラウジング





データベクトルの区分

- ID
 - 一意なので, 視覚表示に用いると混乱
- Main
 - 視覚表示に用いるデータベクトル
- Aux
 - Mainから削除されたが補助的に用いるデータベクトル
 - 色分け
 - 必要に応じて Main に戻す

規模とスピード

- 数十万記録, 数万変量
 - ほとんど待ち時間なし
- 適切なタイミングでのガベージコレクション
- 細部までチューニング
- 使いやすさと負荷のぎりぎりのバランス

TRAD (TextilePlot, R and DandD)

- データサイエンスコンソーシアムの長年の蓄積と大勢の努力の産物
- データサイエンスの基礎理論と実践の反映
- 業務使用に耐えるだけの完成度

- データサイエンスの健全な発展を願って無償で公開 (Windows, Mac)
 - <http://datascience.jp>
 - 本プレゼンテーションもアップしておきます
 - 日々アップデートしていますが、なにか問題があれば query@datascience.jp までご一報ください。