第 31 回 Tokyo.R

はじめよう多変量解析

~ 主成分分析編 ~

@sanoche16

About me

About me

- 佐野宏喜、 @sanoche16
- 現在の地位はフリーター (システムエンジニア)
  注) ニートではありません！！
- PHP, Python, Linux, Java, Ruby, assembler
- 商学部出身
- 最近、機械学習の勉強を始めました！
- 修行が終わったら起業します！

agenda

agenda

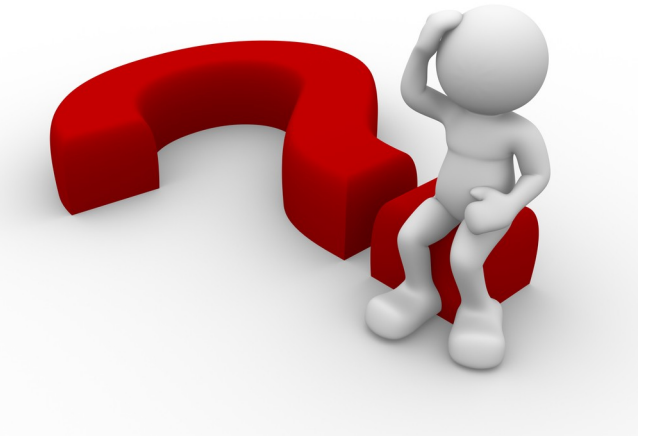
1、多変量解析とは

2、主成分分析の簡単なお話

3、２次元から多次元に

4、量的データから質的データに

1、多変量解析とは



多変量解析とは？

「多数の変量を持つデータを用いた分析」

1、多変量解析とは

例えば・・・

変量

	広告費	社員数	会員数	売上
A 社	12 億円	2000 人	100 万人	200 億円
B 社	2 億円	1200 人	150 万人	750 億円
C 社	10 億円	800 人	60 万人	600 億円
D 社	8 億円	1000 人	200 万人	?

データ

1、多変量解析とは

例えば・・・

変量

	広告費	社員数	会員数	売上
A 社	12 億円	2000 人	100 万人	200 億円
B 社	2 億円	1200 人	150 万人	750 億円
C 社	10 億円	800 人	60 万人	600 億円
D 社	8 億円	1000 人	200 万人	?

データ

D 社の売上はいくらと予想出来るか！？

1、多変量解析とは

例えば・・・

変量

データ

	広告費	社員数	会員数	売上
A 社	12 億円	2000 人	100 万人	200 億円
B 社	2 億円	1200 人	150 万人	750 億円
C 社	10 億円	800 人	60 万人	500 億円
D 社	8 億円	1000 人	200 万人	**550** 億円

1、多変量解析とは

例えば・・・

変量

データ

	広告費	社員数	会員数	売上
A 社	12 億円	2000 人	100 万人	200 億円
B 社	2 億円	1200 人	150 万人	750 億円
C 社	10 億円	800 人	60 万人	500 億円
D 社	8 億円	1000 人	200 万人	**550** 億円

予測出来た！

1、多変量解析とは

例えば・・・

変量

	国語	数学	社会	理科
A 君	82 点	68 点	92 点	76 点
B 君	76 点	98 点	58 点	62 点
C 君	80 点	92 点	72 点	86 点
D 君	86 点	74 点	82 点	90 点

データ

1、多変量解析とは

例えば・・・

変量

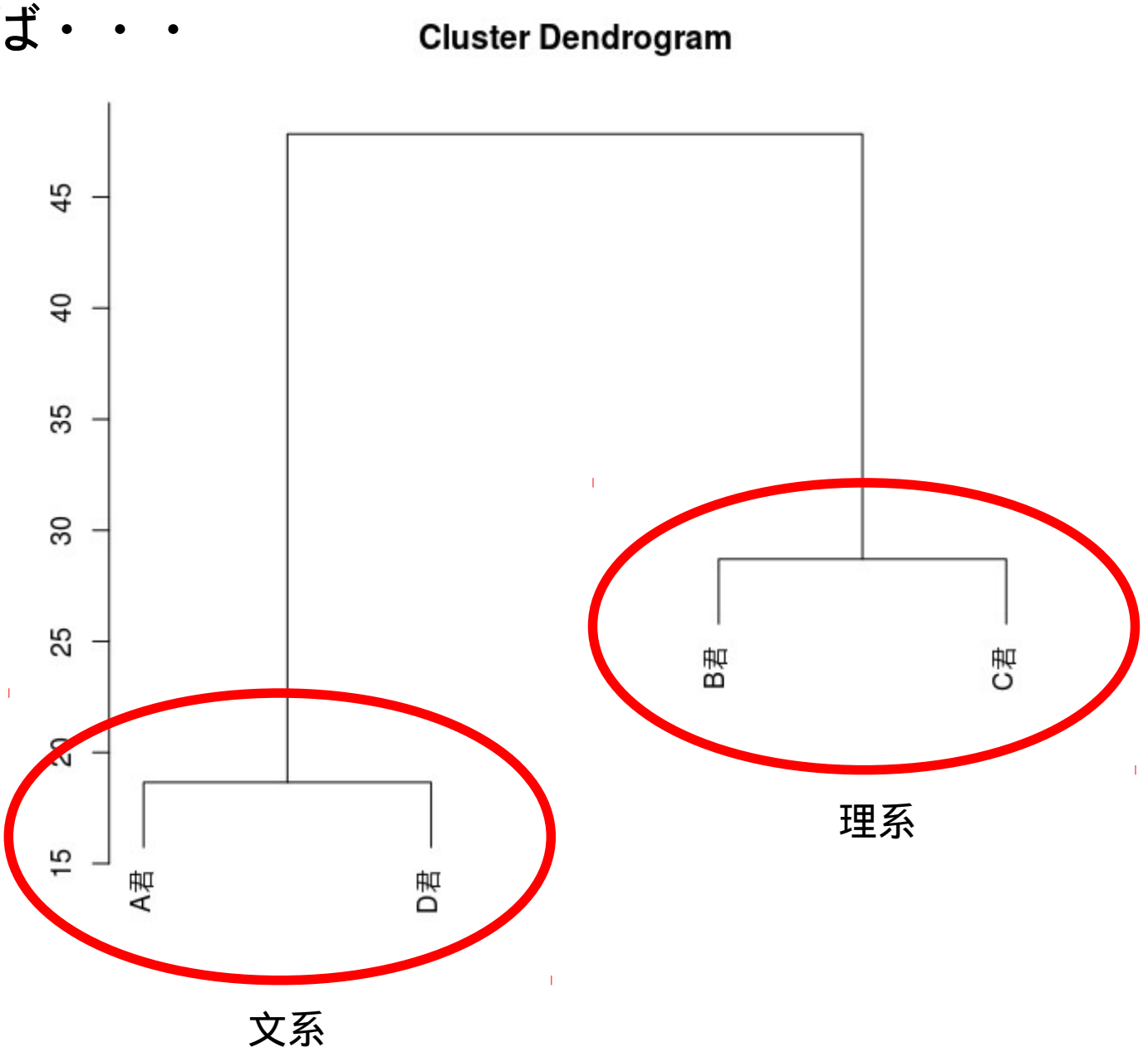
	国語	数学	社会	理科
A 君	82 点	68 点	92 点	76 点
B 君	76 点	98 点	58 点	62 点
C 君	80 点	92 点	72 点	86 点
D 君	86 点	74 点	82 点	90 点

データ

2 つのタイプに分けたい！

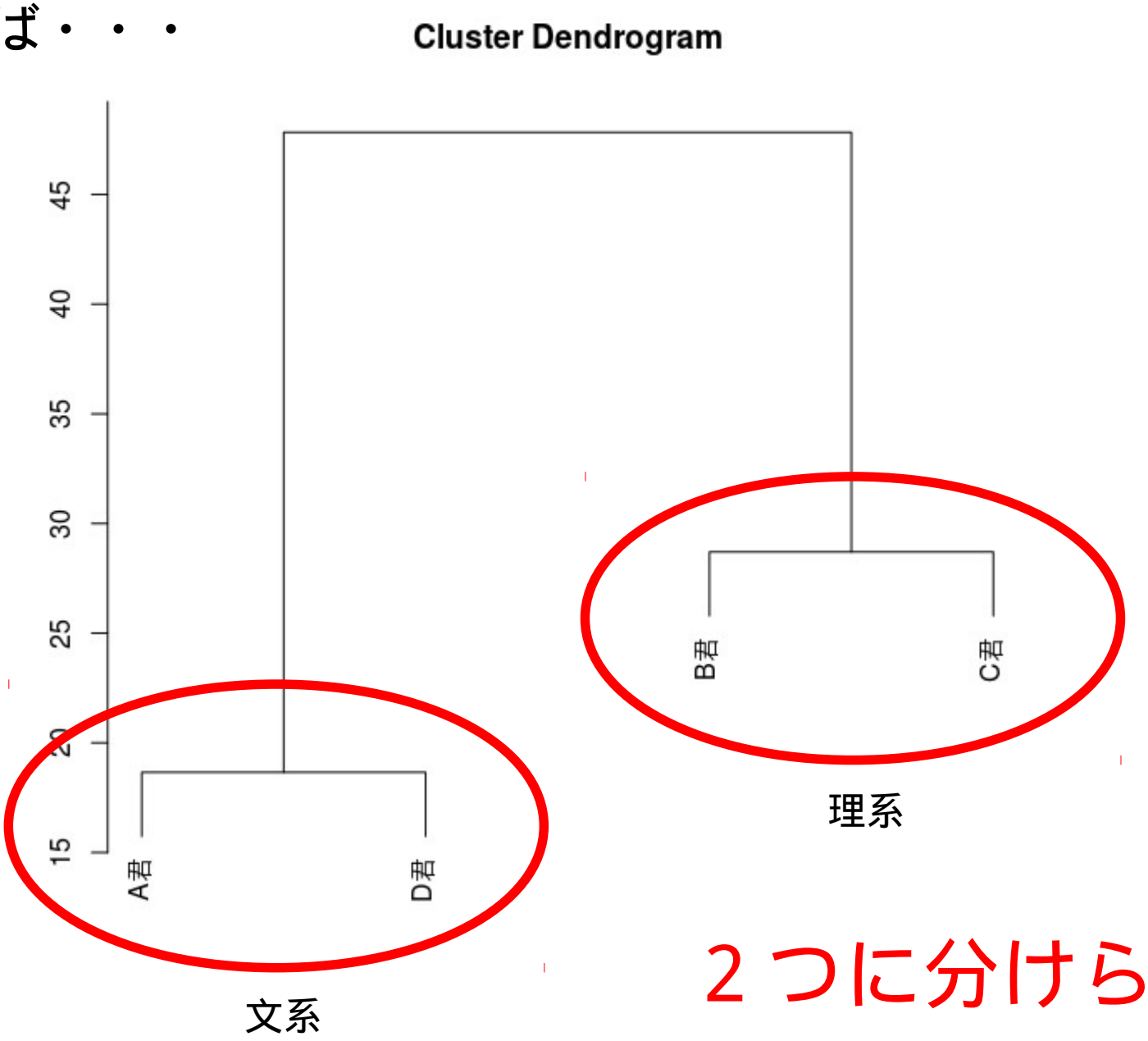
1、多変量解析とは

例えば・・・



1、多変量解析とは

例えば・・・



2つに分けられた！

1、多変量解析とは

多変量解析を行うには？

　１変量・２変量の分析が出来なければいけない！

必要な知識

　平均・分散・共分散・相関係数・行列演算

　たったこれだけ！！

1、多変量解析とは

多変量解析を行うには？

　１変量・２変量の分析が出来なければいけない！

必要な知識

　平均・分散・共分散・相関係数・行列演算

　たったこれだけ！！

とは言え必要な知識

　微分積分・分布（正規分布など）

1、多変量解析とは

多変量解析を行うには？

　１変量・２変量の分析が出来なければいけない！

必要な知識

　平均・分散・共分散・相関係数・行列演算

　たったこれだけ！！

とは言え必要な知識

　微分積分・分布（正規分布）

注）数式を使って統計学を学びましょう。

多変量解析の例

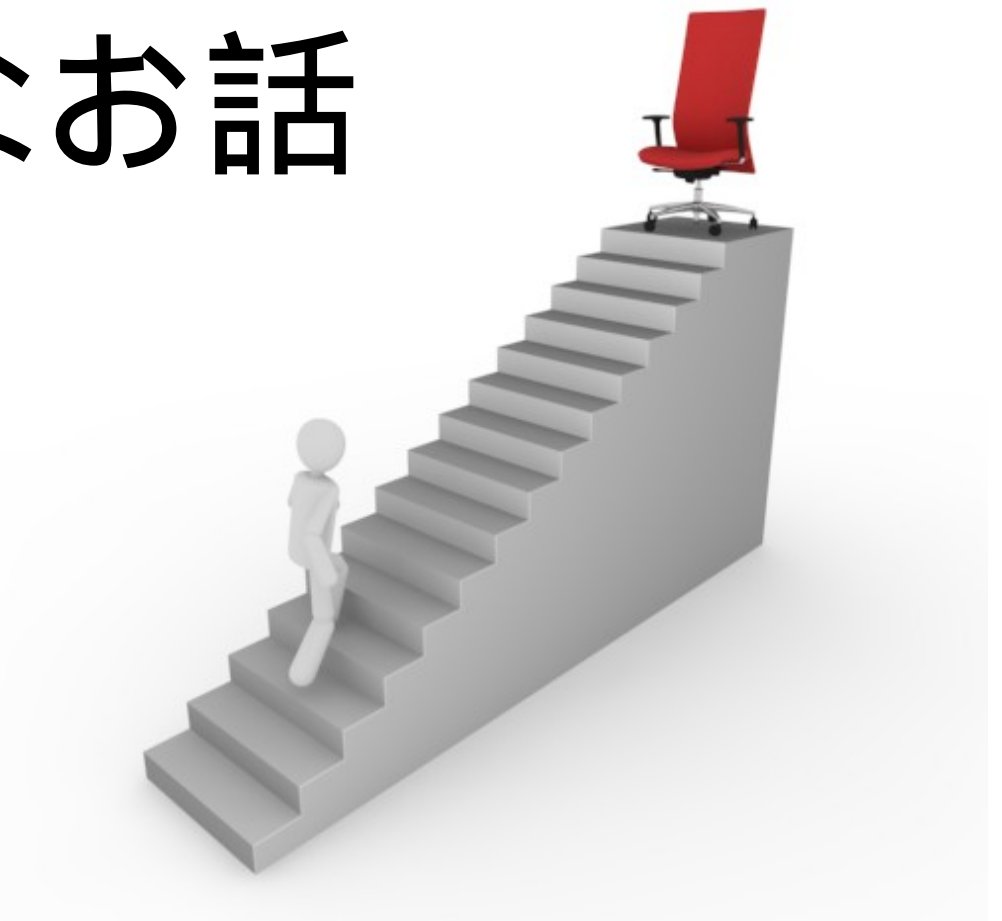
回帰分析・主成分分析・因子分析・判別分析・・・

多変量解析の例

回帰分析・主成分分析・因子分析・判別分析・・・

本日はこれ

2、主成分分析の簡単なお話



以下の８社を企業の規模順に並べたいとする

注）単位は十億円

	時価総額	純資産
ガンホー	1,267	32
マツモトキヨシ	137	137
旭化成	952	824
キリン	1662	1278
アオキ	139	111
資生堂	601	304
第一生命	1412	1649
シャープ	629	135

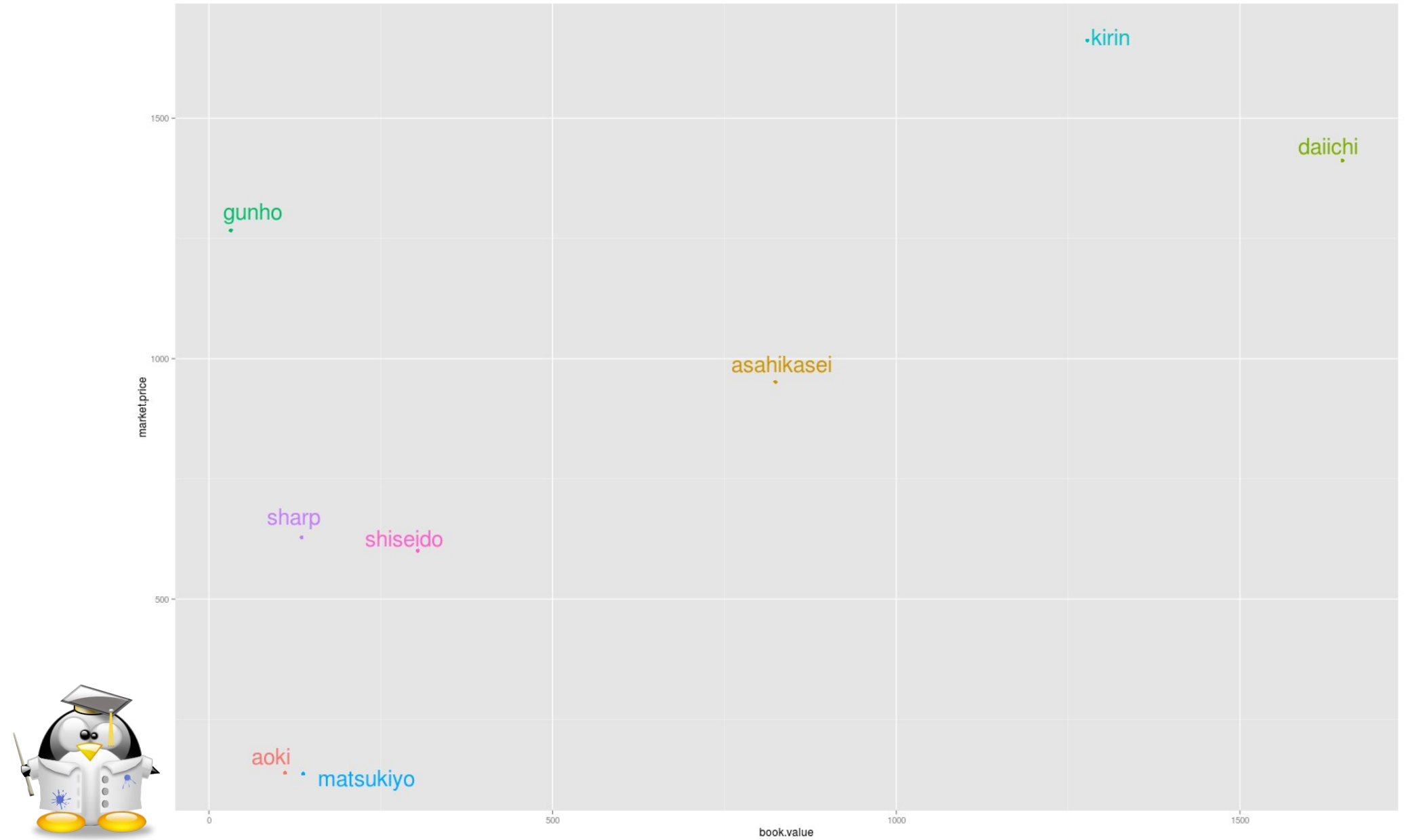
以下の８社を企業の規模順に並べたいとする

注）単位は十億円

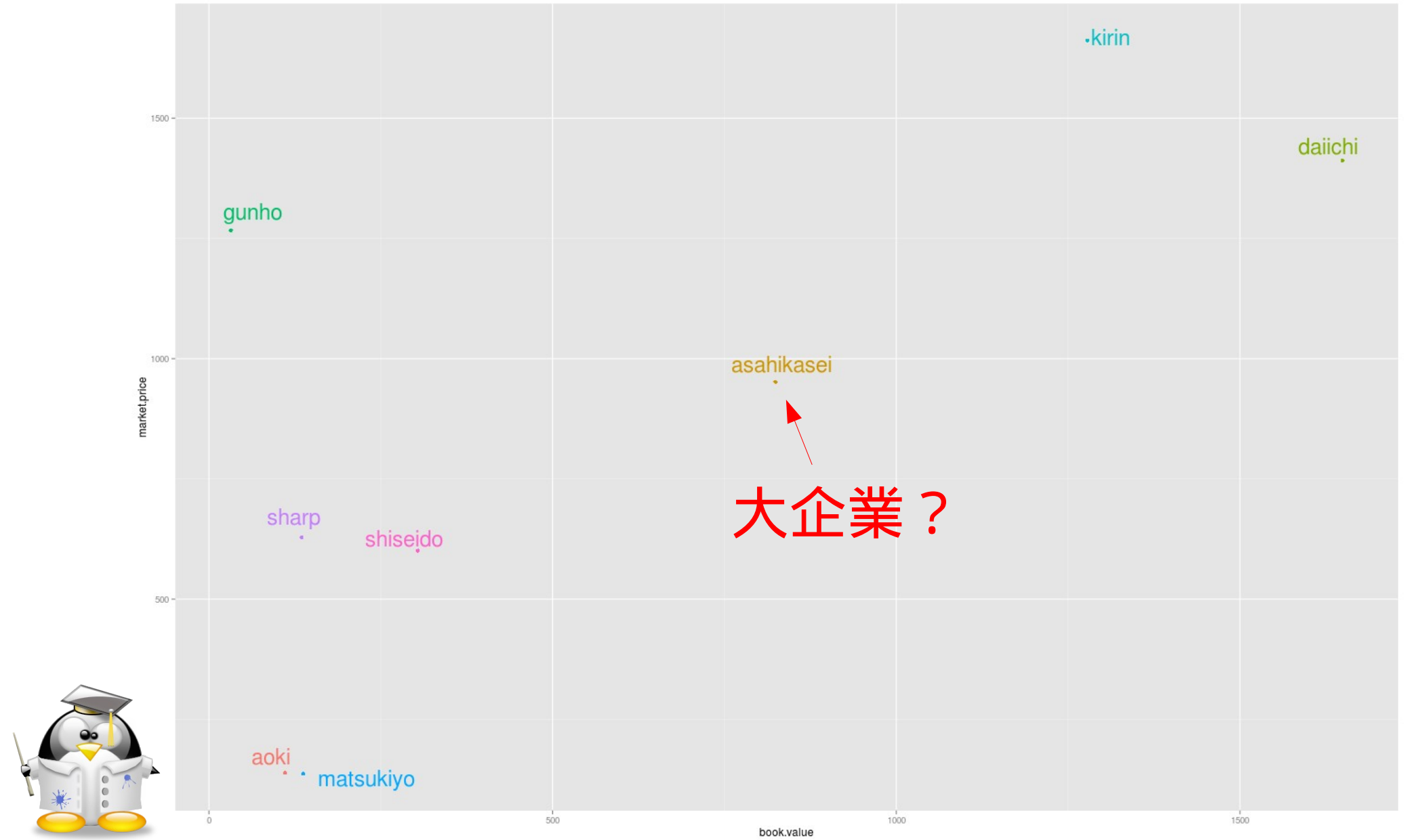
	時価総額	純資産
ガンホー	1,267	32
マツモトキヨシ	137	137
旭化成	952	824
キリン	1662	1278
アオキ	139	111
資生堂	601	304
第一生命	1412	1649
シャープ	629	135

どれが大企業か？？

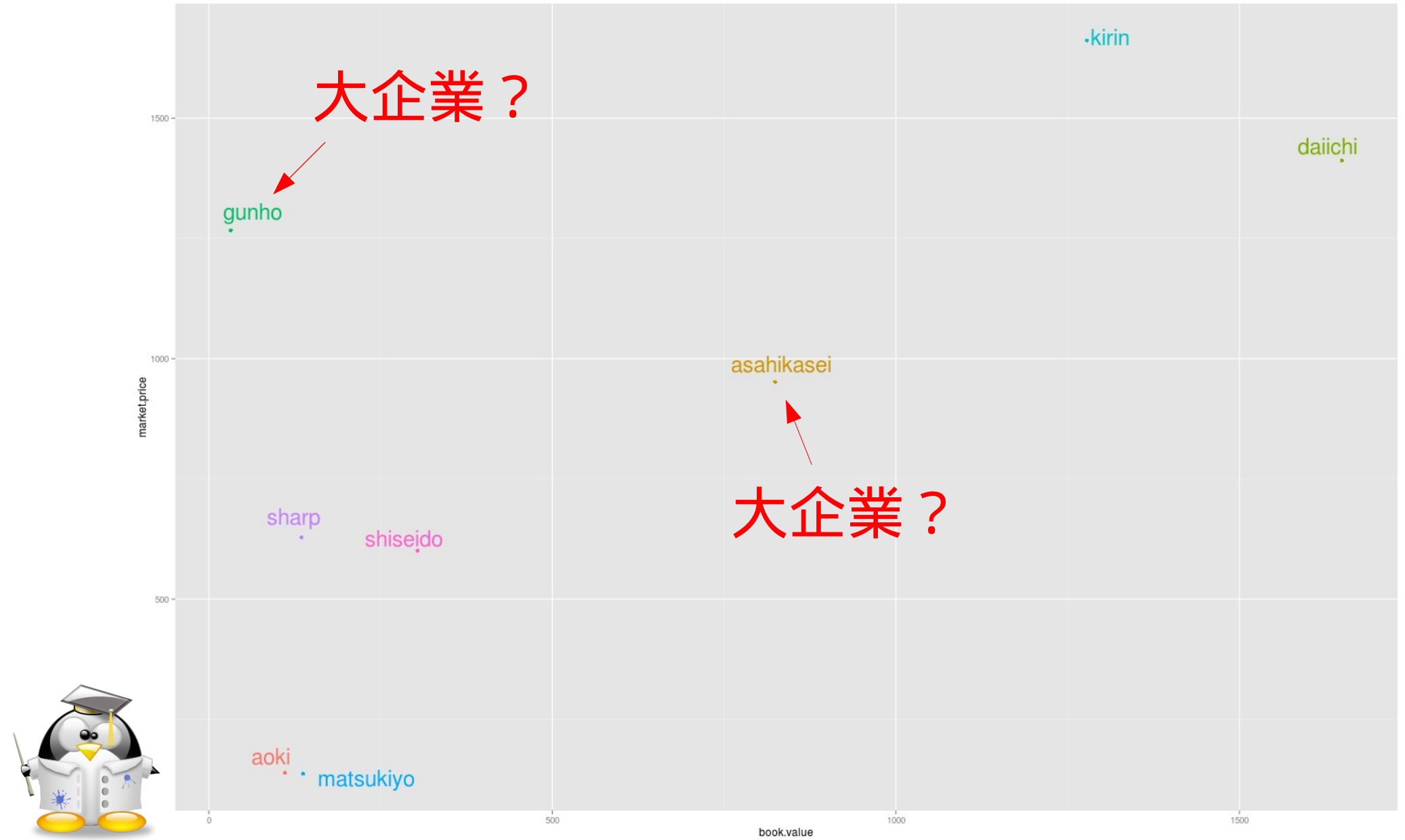
とりあえずプロットする



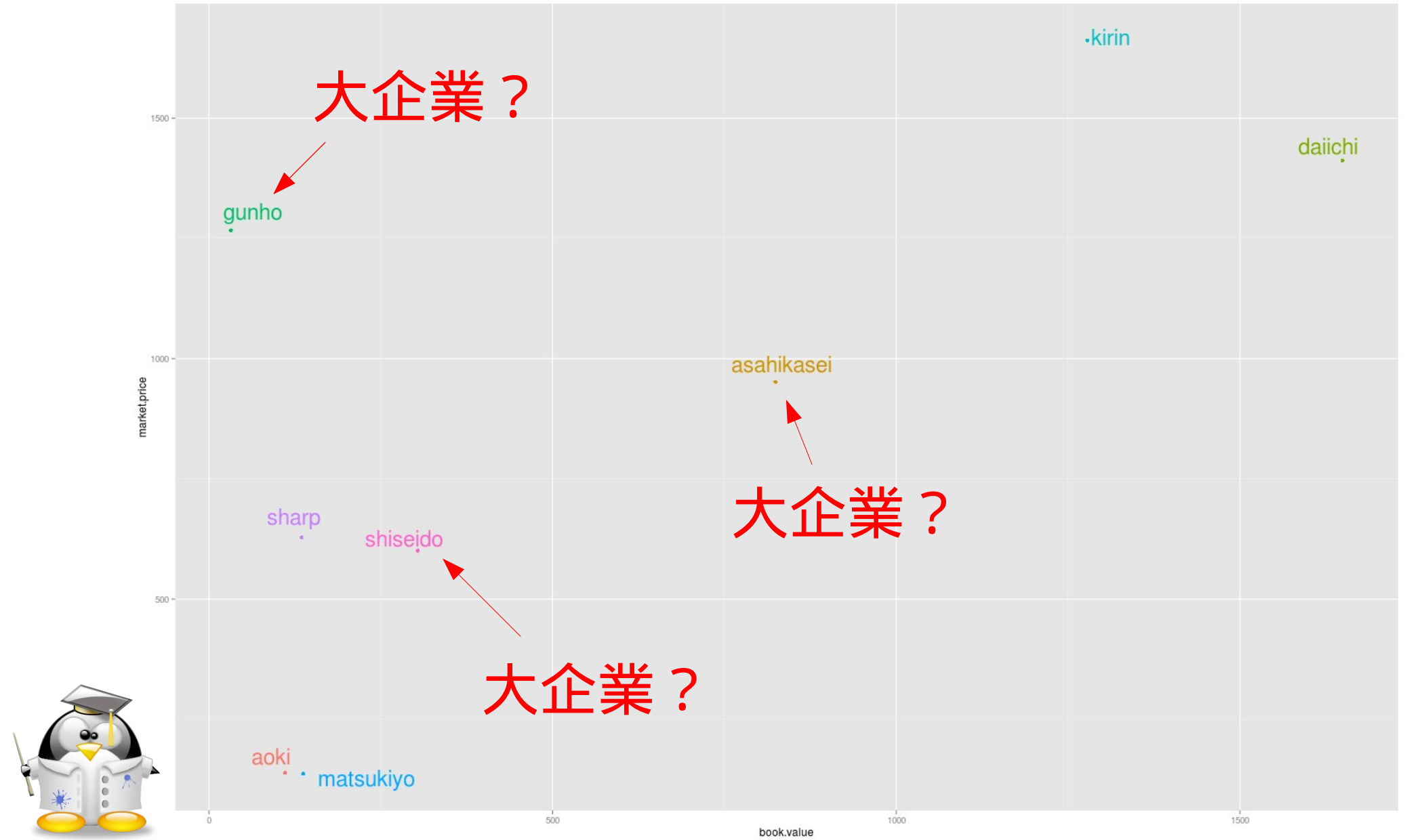
とりあえずプロットする



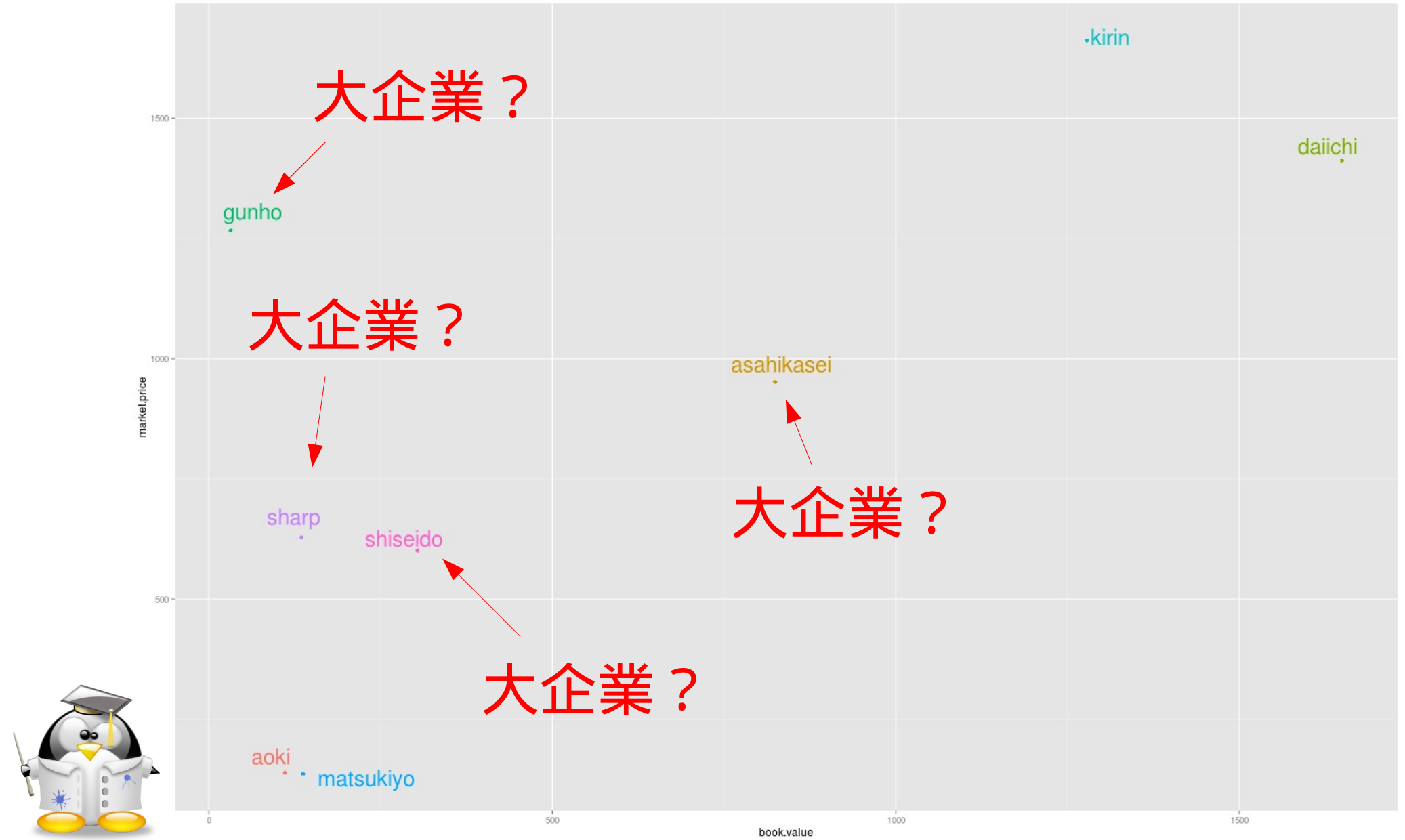
とりあえずプロットする



とりあえずプロットする



とりあえずプロットする



2 次元だと分かりにくい！！

2次元だと分かりにくい！！

出来れば得点をつけて1列に並べたい！！

2 次元だと分かりにくい！！

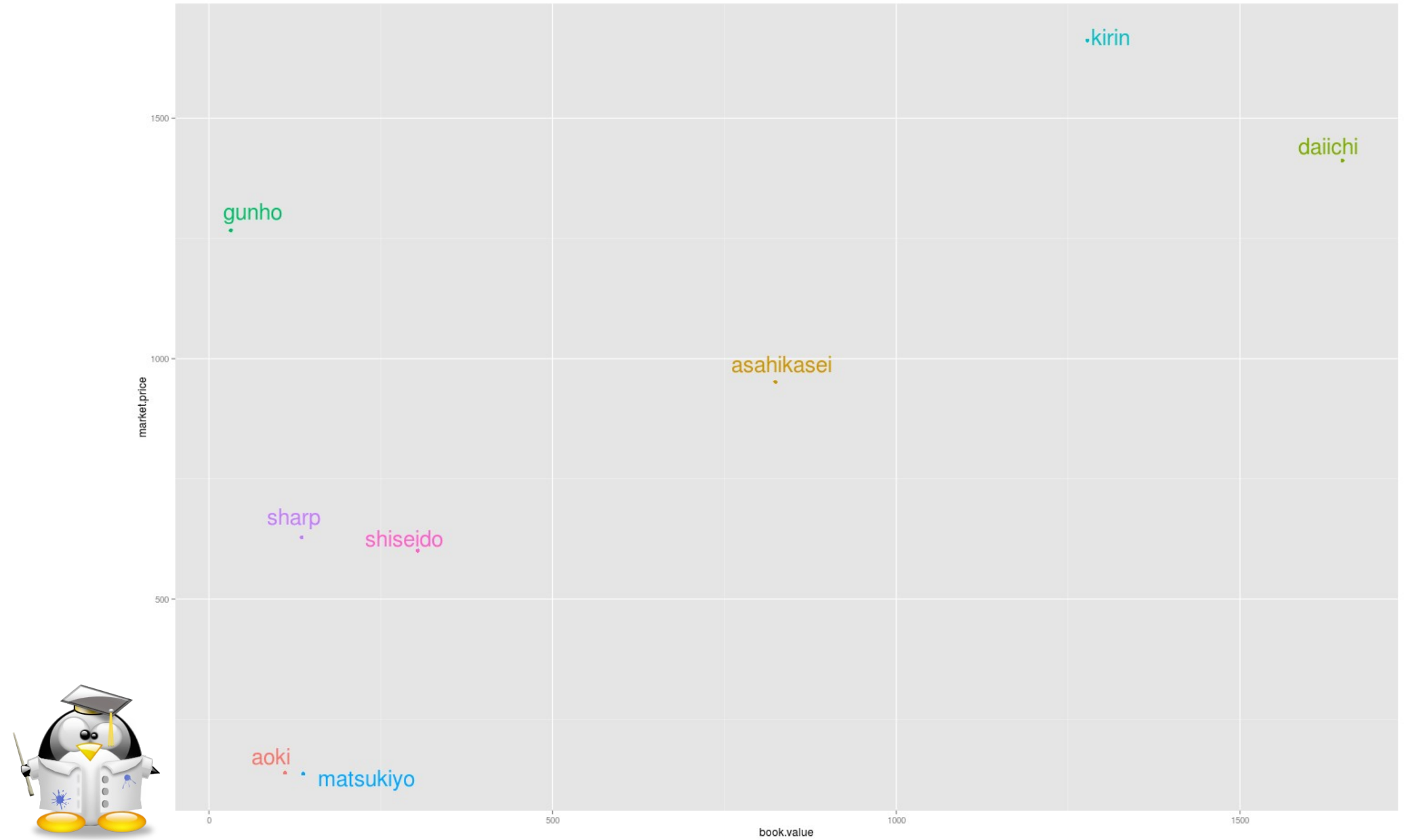
出来れば得点をつけて 1 列に並べたい！！

=> 得点をつける方法
を考える

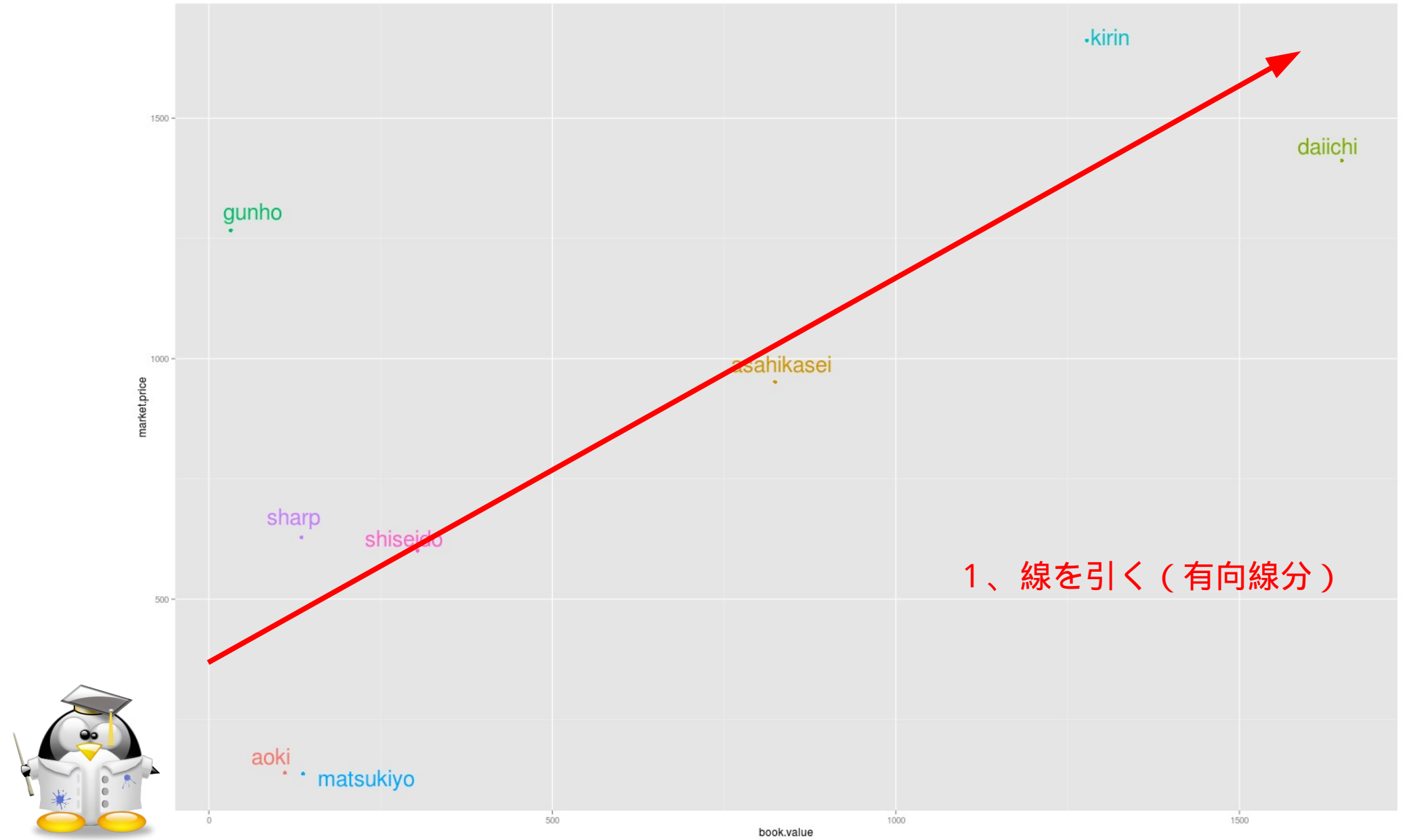
2、主成分分析の簡単なお話

TRY IT!!

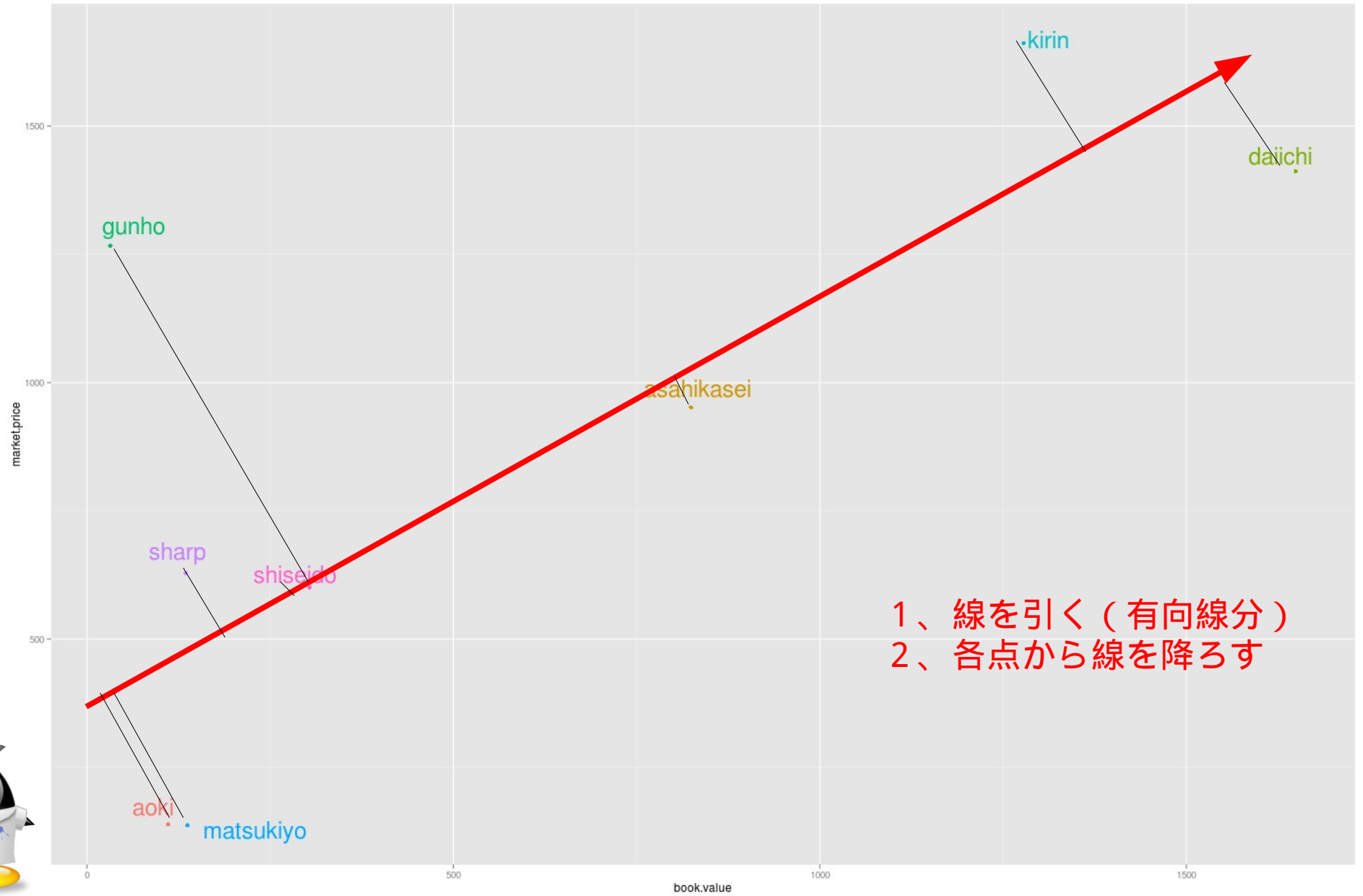
もう一度眺めてみる


kirin
daiichi
gunho
asahikasei
sharp
shiseido
aoki
matsukiyo
1500
1000
500
market.price
0
500
1000
1500
book.value

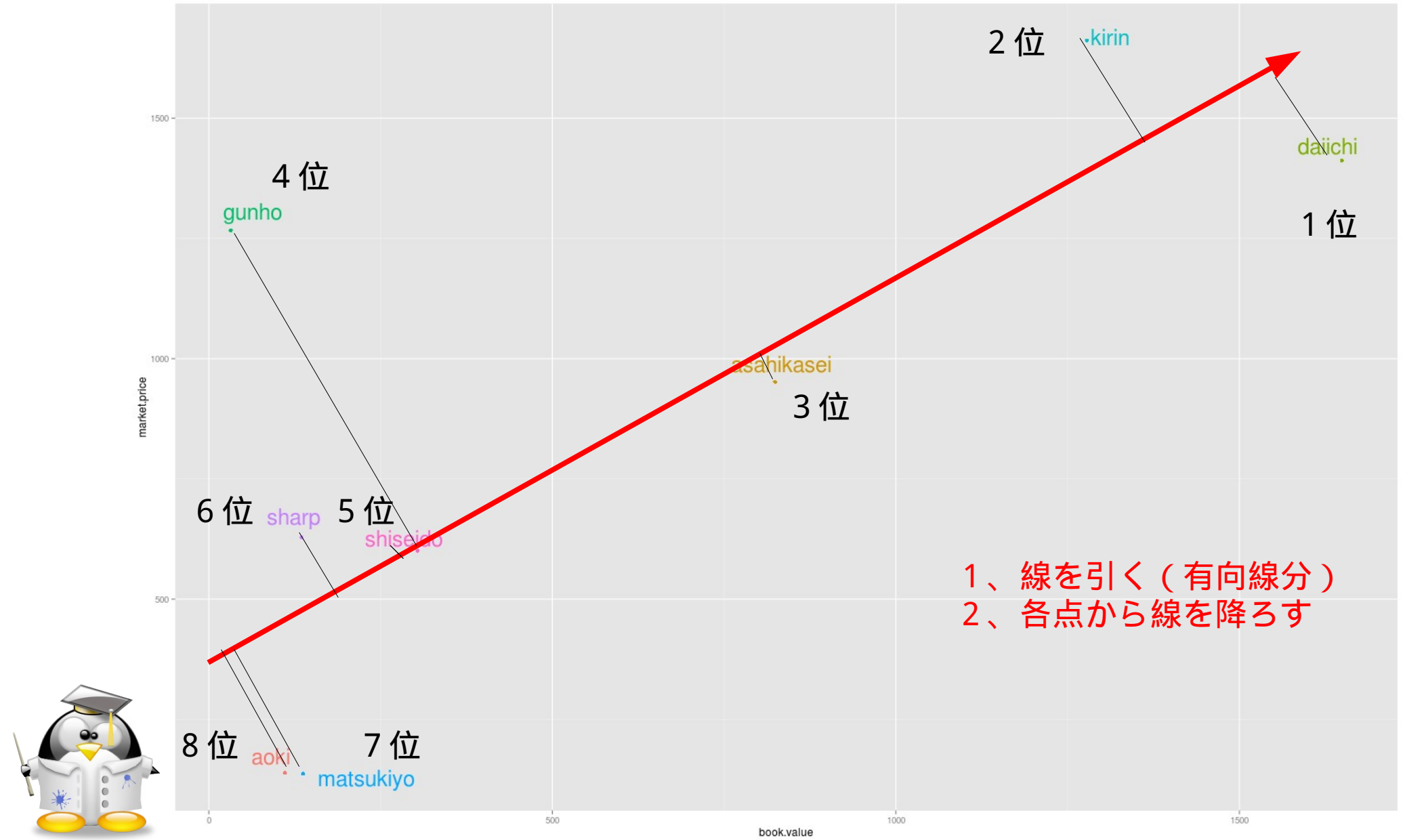
もう一度眺めてみる


kirin
daiichi
gunho
asahikasei
sharp
shiseido
aoki
matsukiyo
1、線を引く（有向線分）
market.price
book.value
1500
1000
500
0
500
1000
1500

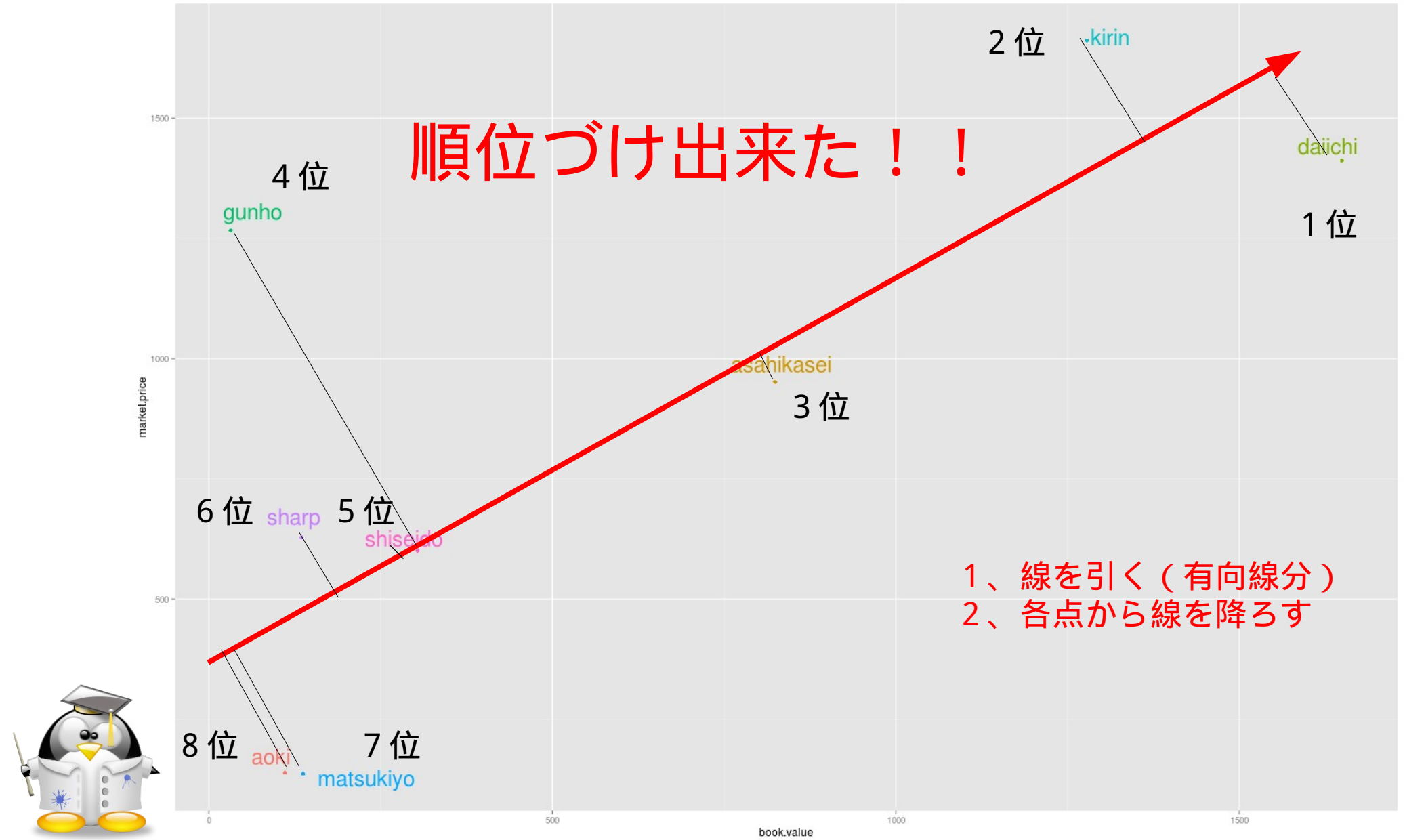
もう一度眺めてみる


kirin
daiichi
gunho
asahikasei
sharp
shiseido
aoki
matsukiyo
1500
1000
500
market.price
0
500
1000
1500
book.value
1、線を引く（有向線分）
2、各点から線を降ろす

もう一度眺めてみる


2 位
kirin
daiichi
1 位
4 位
gunho
asahikasei
3 位
6 位
sharp
5 位
shiseido
8 位
aoki
7 位
matsukiyo
1、線を引く（有向線分）
2、各点から線を降ろす
market.price
book.value
1500
1000
500
0
500
1000
1500

もう一度眺めてみる


順位づけ出来た！！
2 位
kirin
daiichi
1 位
4 位
gunho
asahikasei
3 位
6 位
sharp
5 位
shiseido
8 位
aoki
7 位
matsukiyo
1、線を引く（有向線分）
2、各点から線を降ろす
market.price
book.value
1500
1000
500
0
500
1000
1500

~まとめ~

・主成分分析（２次元の場合）とは？

２次元データ（時価総額と純資産）を変換して１次元（企業規模を表す得点）データに置き換えること

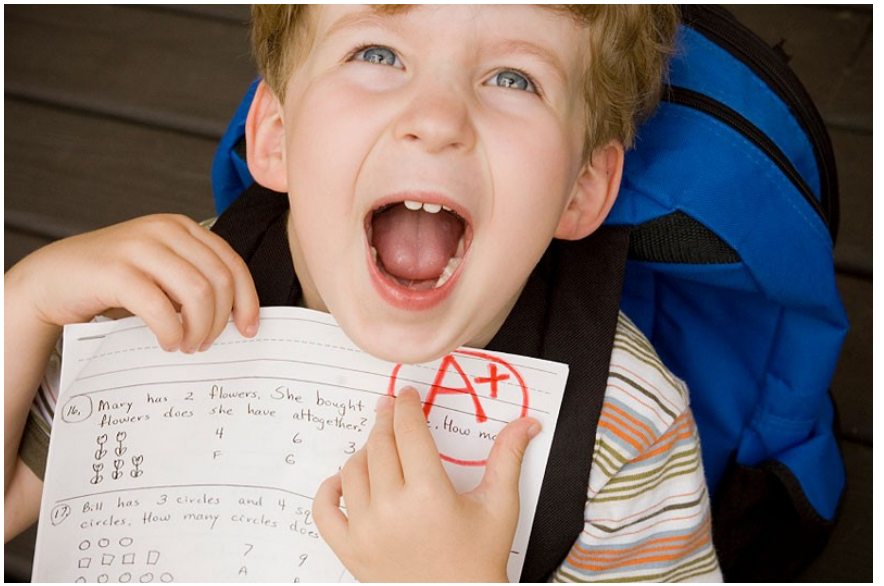
~まとめ~

・主成分分析（２次元の場合）とは？

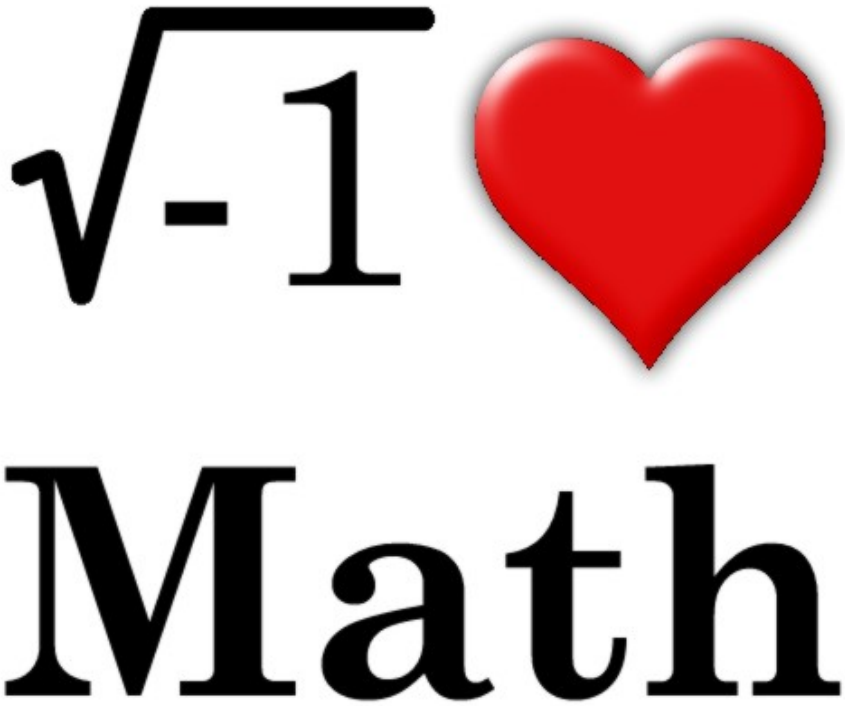
２次元データ（時価総額と純資産）を変換して１次元（企業規模を表す得点）データに置き換えること

分かりやすい！

超便利！！



みんな大好きな数学のお話



そもそも問題は。。。

「企業の規模を時価総額と純資産の両方を考慮して評価したい」 => 重み付けして考える

企業規模を z とおく。

時価総額を x1、純資産を x2 とおいて

$$z = a_1 x_1 + a_2 x_2$$

という式を作り上げればよい

例えば $z = x_1 + x_2$

	時価総額 (x1)	純資産 (x2)	x1	x2	z
ガンホー	1,267	32	1,267	32	1299
マツモトキヨシ	137	137	137	137	274
旭化成	952	824	952	824	1776
キリン	1662	1278	1662	1278	2940
アオキ	139	111	139	111	250
資生堂	601	304	601	304	905
第一生命	1412	1649	1412	1649	3061
シャープ	629	135	629	135	764

例えば $z = x_1 + 2x_2$

	時価総額 (x1)	純資産 (x2)	x1	2 × x2	z
ガンホー	1,267	32	1,267	64	1331
マツモトキヨシ	137	137	137	274	411
旭化成	952	824	952	1648	2600
キリン	1662	1278	1662	2556	4218
アオキ	139	111	139	222	361
資生堂	601	304	601	608	1209
第一生命	1412	1649	1412	3298	4710
シャープ	629	135	629	270	899

例えば $z = 3x_1 + x_2$

	時価総額 (x1)	純資産 (x2)	3 × x1	x2	z
ガンホー	1,267	32	3809	32	3833
マツモトキヨシ	137	137	411	137	548
旭化成	952	824	2856	824	3680
キリン	1662	1278	4986	1278	6264
アオキ	139	111	417	111	528
資生堂	601	304	1803	304	2107
第一生命	1412	1649	4236	1649	5885
シャープ	629	135	1887	135	2022

一般化

	時価総額 (x1)	純資産 (x2)	a1 × x1	a2 × x2	z
ガンホー	1,267	32	1267a1	32a2	1267a1 + 32 a2
マツモトキヨシ	137	137	137a1	137a2	137a1 + 137a2
旭化成	952	824	952a1	824a2	952a1 + 824a2
キリン	1662	1278	1662a1	1278a2	1662a1 + 1278a2
アオキ	139	111	139a1	111a2	139a1 + 111a2
資生堂	601	304	601a1	304a2	601a1 + 304a2
第一生命	1412	1649	1412a1	1649a2	1412a1 + 1649a2
シャープ	629	135	629a1	135a2	629a1 + 135a2

一般化

	時価総額 (x1)	純資産 (x2)	a1 × x1	a2 × x2	z
ガンホー	1,267	32	1267a1	32a2	1267a1 + 32 a2
マツモトキヨシ	137	137	137a1	137a2	137a1 + 137a2
旭化成	952	824	952a1	824a2	952a1 + 824a2
キリン	1662	1278	1662a1	1278a2	1662a1 + 1278a2
アオキ	139	111	139a1	111a2	139a1 + 111a2
資生堂	601	304	601a1	304a2	601a1 + 304a2
第一生命	1412	1649	1412a1	1649a2	1412a1 + 1649a2
シャープ	629	135	629a1	135a2	629a1 + 135a2

$$z = a_1 x_1 + a_2 x_2$$ とおく

どうやって a1, a2 を決めればよいか？？

一般化

	時価総額 (x1)	純資産 (x2)	a1 × x1	a2 × x2	z
ガンホー	1,267	32	1267a1	32a2	1267a1 + 32 a2
マツモトキヨシ	137	137	137a1	137a2	137a1 + 137a2
旭化成	952	824	952a1	824a2	952a1 + 824a2
キリン	1662	1278	1662a1	1278a2	1662a1 + 1278a2
アオキ	139	111	139a1	111a2	139a1 + 111a2
資生堂	601	304	601a1	304a2	601a1 + 304a2
第一生命	1412	1649	1412a1	1649a2	1412a1 + 1649a2
シャープ	629	135	629a1	135a2	629a1 + 135a2

z で会社の規模を判断したい
=>z が最もバラつくように a1, a2 決める
=>z の分散を最大化するように a1, a2 を決める！

一般化

z の平均: $849.875a_1 + 558.75a_2$

z の分散:

$$\frac{1}{7}\{(417.125a_1 - 526.75a_2)^2 + (-712.875a_1 - 421.75a_2)^2 + (102.125a_1 + 265.25a_2)^2$$

$$+(812.125a_1 + 719.25a_2)^2 + (-710.875a_1 - 447.75a_2)^2 + (-248.875a_1 + -254.75a_2)^2$$

$$+(562.125a_1 + 1090.25)^2 + (-220.875a_1 - 423.75a_2)^2\}$$

$$= 326316a_1^2 + 382372a_2^2 - 2 \times 254327a_1a_2$$

x1 の分散　x2 の分散　　X1, x2 の共分散

z の分散を最大化させるような a1, a2 を決める

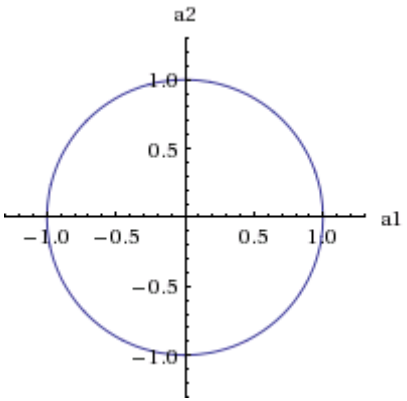
一般化

z の分散：$326316a_1^2 + 382372a_2^2 - 2 \times 254327a_1a_2$

これだけでは a1, a2 は決まらない (当たり前)

=> a1 と a2 の関係を決める必要がある

=> $a_1^2 + a_2^2 = 1$ を制約条件式とする



※ これ以外の制約式を使った事がある方がいればぜひ教えてください！

一般化

要するに

$$a_1^2 + a_2^2 = 1$$

の元で

$$326316a_1^2 + 382372a_2^2 - 2 \times 254327a_1a_2$$

を最大化

=> 条件付き極値問題に帰着出来た！

注）理系数学を学んでいない方はこの辺りから少々難しくなってくるかも知れませんが、実際やってみるととても簡単なお話です。

一般化

ラグランジュの乗数法を使って解く！

$$a_1^2 + a_2^2 = 1$$

$$326316a_1^2 + 382372a_2^2 - 2 \times 254327a_1a_2$$

から

$$g(a_1, a_2) = 326316a_1^2 + 382372a_2^2 - 2 \times 254327a_1a_2$$

$$f(a_1, a_2) = a_1^2 + a_2^2 - 1 = 0$$

とおくと

$$g_{a_1} = 2 \times 326316a_1 - 2 \times 254327a_2$$

$$g_{a_2} = 2 \times 382372a_2 - 2 \times 254327a_1$$

$$f_{a_1} = 2a_1, f_{a_2} = 2a_2$$

一般化

よって

$$326316a_1 - 254327a_2 - \lambda a_1 = 0$$

$$-254327a_1 + 382372a_2 - \lambda a_2 = 0$$

を解けばよい！

行列を用いて表現すると

$$\begin{bmatrix} 326316 & -254327 \\ -254327 & 382372 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

分散共分散行列

一般化

よって

$$326316a_1 - 254327a_2 - \lambda a_1 = 0$$

$$-254327a_1 + 382372a_2 - \lambda a_2 = 0$$

を解けばよい！

行列を用いて表現すると

$$\begin{bmatrix} 326316 & -254327 \\ -254327 & 382372 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

分散共分散行列

固有値問題になった！

一般化

ところで

$$326316a_1 - 254327a_2 - \lambda a_1 = 0$$

$$-254327a_1 + 382372a_2 - \lambda a_2 = 0$$

を上式 ×a1、下式 ×a2 をして足してみると

$$326316a_1^2 + 382372a_2^2 - 2 \times 254327a_1a_2 = \lambda$$

z の分散になった！

λ は z の分散だった！

一般化

R を使って解いてみる

```
> x1 <-c(1267,137,952,1662,139,601,1412,629)
> x2 <-c(32, 137, 824, 1278, 111, 304, 1649, 135)
> data <- data.frame(x1, x2)
> eigen(var(data))
$values
[1] 610211.2  98476.9

$vectors
           [,1]       [,2]
[1,] 0.6672553 -0.7448291
[2,] 0.7448291  0.6672553
```

解けた！

一般化

λ = 610211.2, 98476.9 と

2 つ出てくるが、分散 (λ) の大きい方を選べばよい

よって、 λ = 610211.2 のとき、

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.67 \\ 0.74 \end{bmatrix}$$

となる！

一般化

λ = 610211.2, 98476.9 と

2 つ出てくるが、分散 (λ) の大きい方を選べばよい

よって、 λ = 610211.2 のとき、

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.67 \\ 0.74 \end{bmatrix}$$

となる！

(* ￣∇￣ *) ｴﾍﾍ

実際に使ってみる

	時価総額 (x1)	純資産 (x2)	0.67 × x1	0.74 × x2	z
ガンホー	1,267	32	848.89	23.68	872.57
マツモトキヨシ	137	137	91.79	101.38	193.17
旭化成	952	824	637.84	609.76	1247.60
キリン	1662	1278	1113.54	945.72	2059.26
アオキ	139	111	93.13	82.14	175.27
資生堂	601	304	402.67	224.96	627.63
第一生命	1412	1649	946.04	1220.26	2166.30
シャープ	629	135	421.43	99.90	521.33

実際に使ってみる

	時価総額 (x1)	純資産 (x2)	0.67 × x1	0.74 × x2	z
ガンホー	1,267	32	848.89	23.68	872.57
マツモトキヨシ	137	137	91.79	101.38	193.17
旭化成	952	824	637.84	609.76	1247.60
キリン	1662	1278	1113.54	945.72	2059.26
アオキ	139	111	93.13	82.14	175.27
資生堂	601	304	402.67	224.96	627.63
第一生命	1412	1649	946.04	1220.26	2166.30
シャープ	629	135	421.43	99.90	521.33

主成分得点が出せた！

さらに上へ

もうひとつの

$\lambda = 98476.9$

と

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -0.74 \\ 0.67 \end{bmatrix}$$

とはなんだろうか？？

比べてみる

・固有値

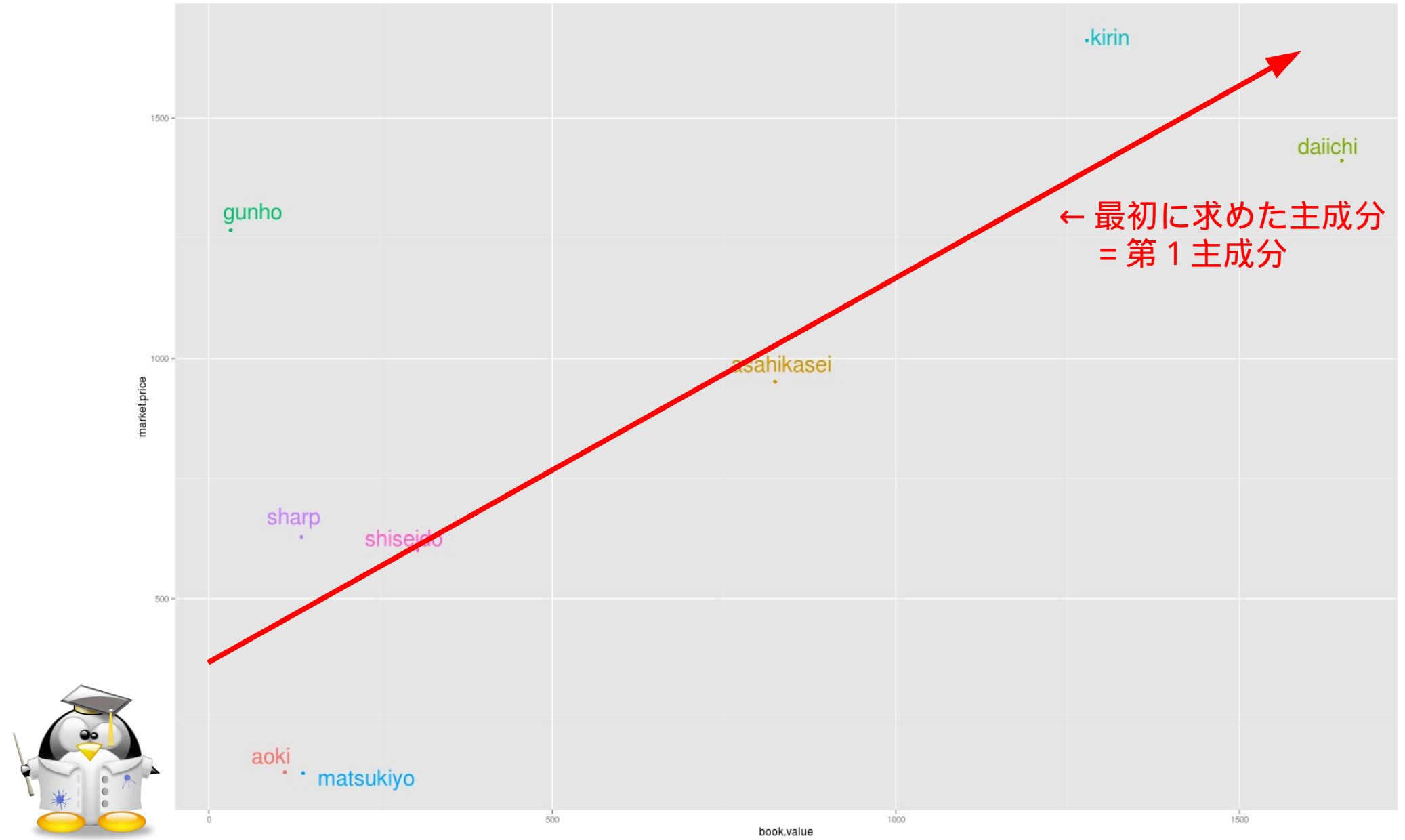
$\lambda = 610211.2$　　$\lambda = 98476.9$

・固有ベクトル

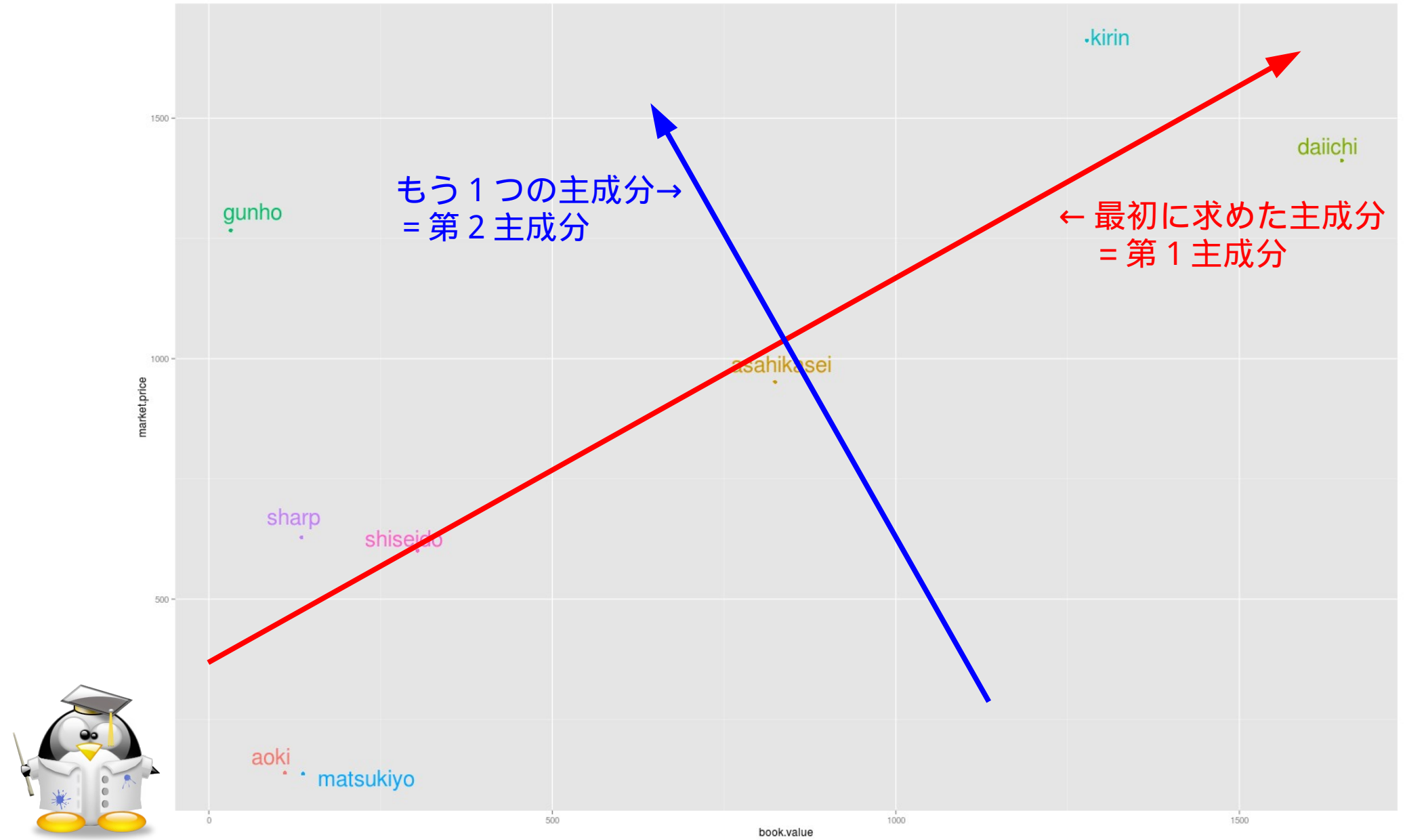
$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -0.74 \\ 0.67 \end{bmatrix} \qquad \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0.67 \\ 0.74 \end{bmatrix}$$

直行している！！

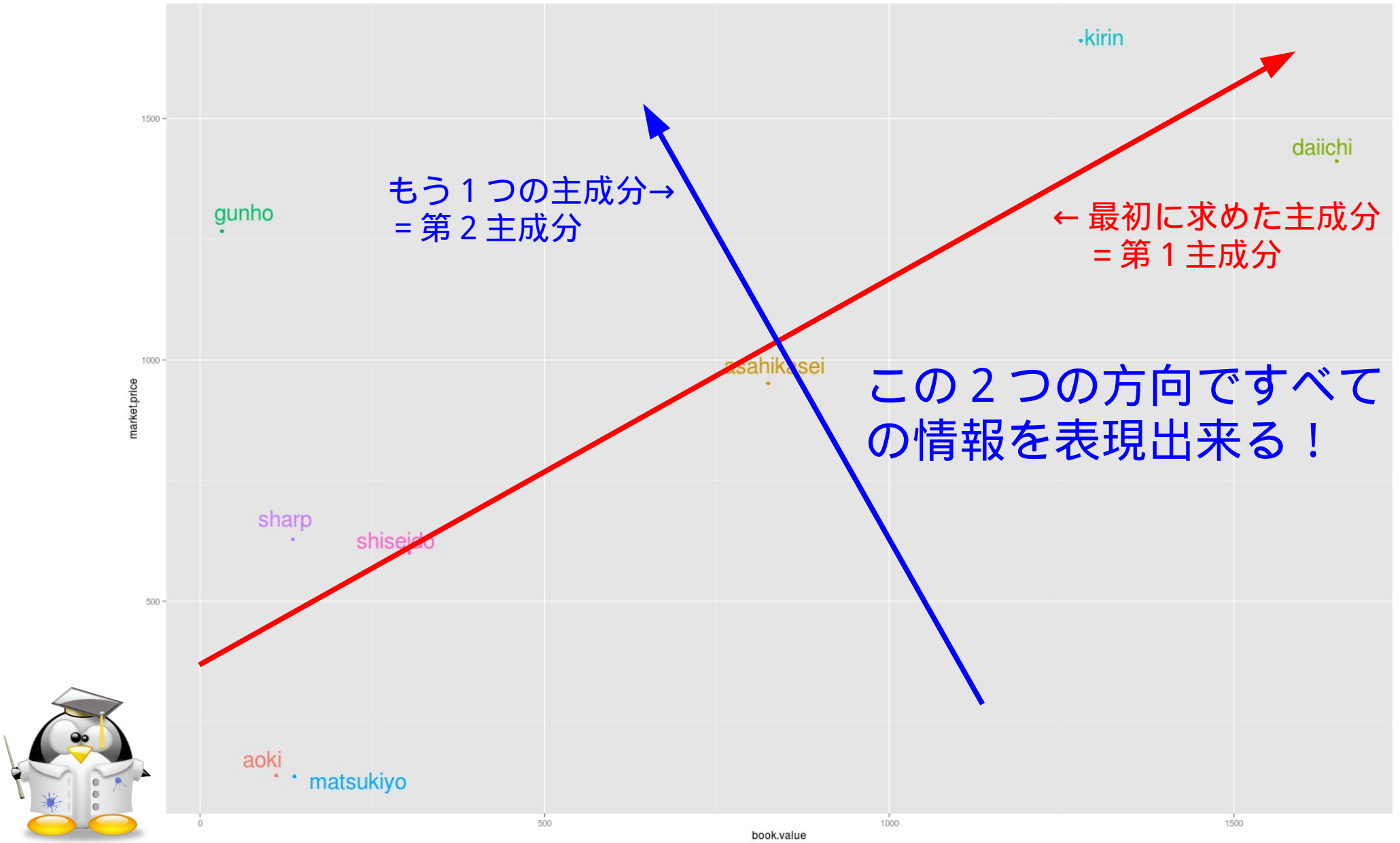
もう一度眺めてみる


kirin
daiichi
gunho
asahikasei
sharp
shiseido
aoki
matsukiyo
← 最初に求めた主成分
= 第１主成分
market.price
book.value
1500
1000
500
0
500
1000
1500

もう一度眺めてみる

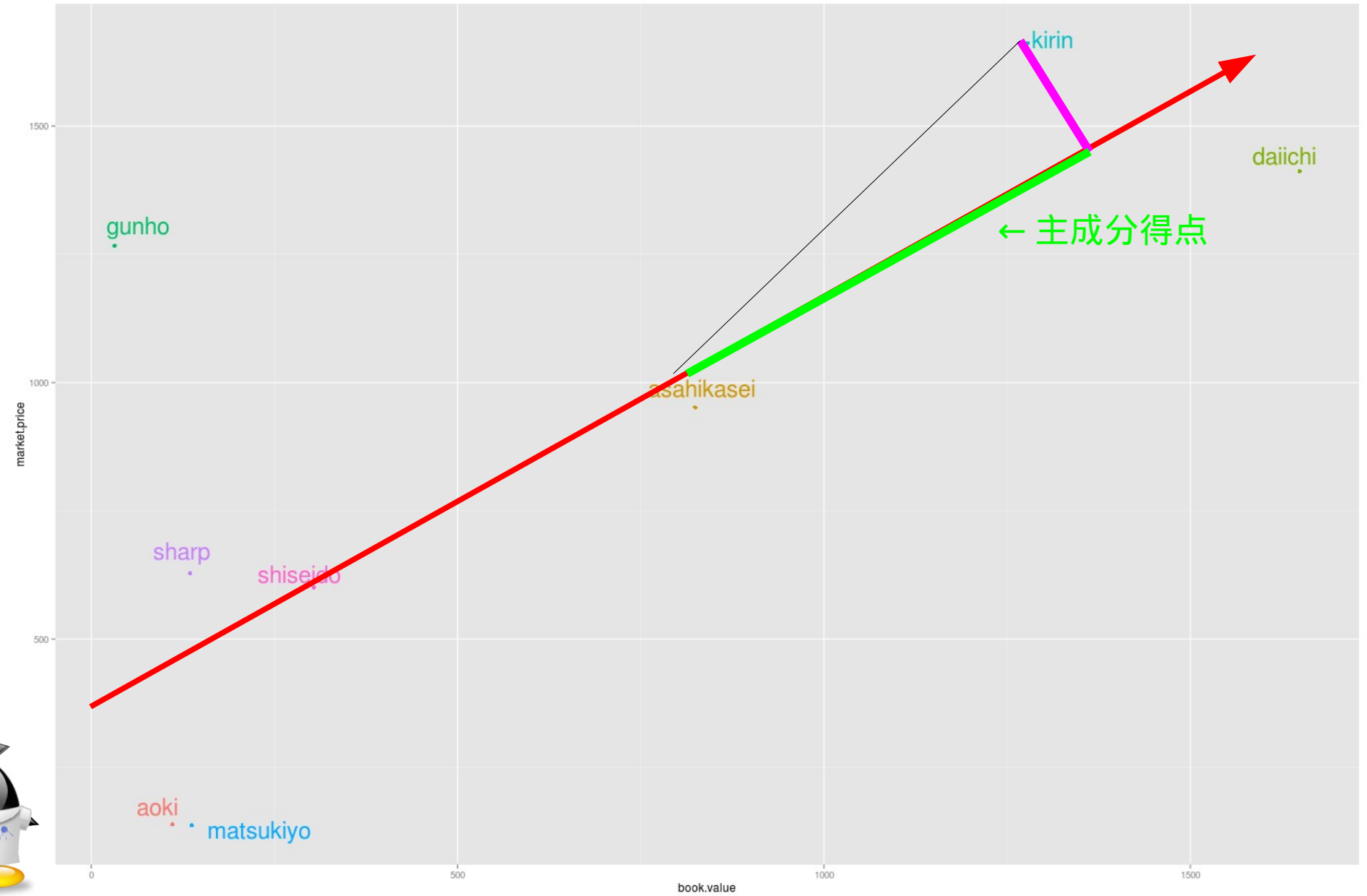

kirin
daiichi
gunho
もう１つの主成分→
＝第２主成分
← 最初に求めた主成分
＝第１主成分
asahikasei
sharp
shiseido
aoki
matsukiyo
1500
1000
500
market.price
0
500
1000
1500
book.value

もう一度眺めてみる

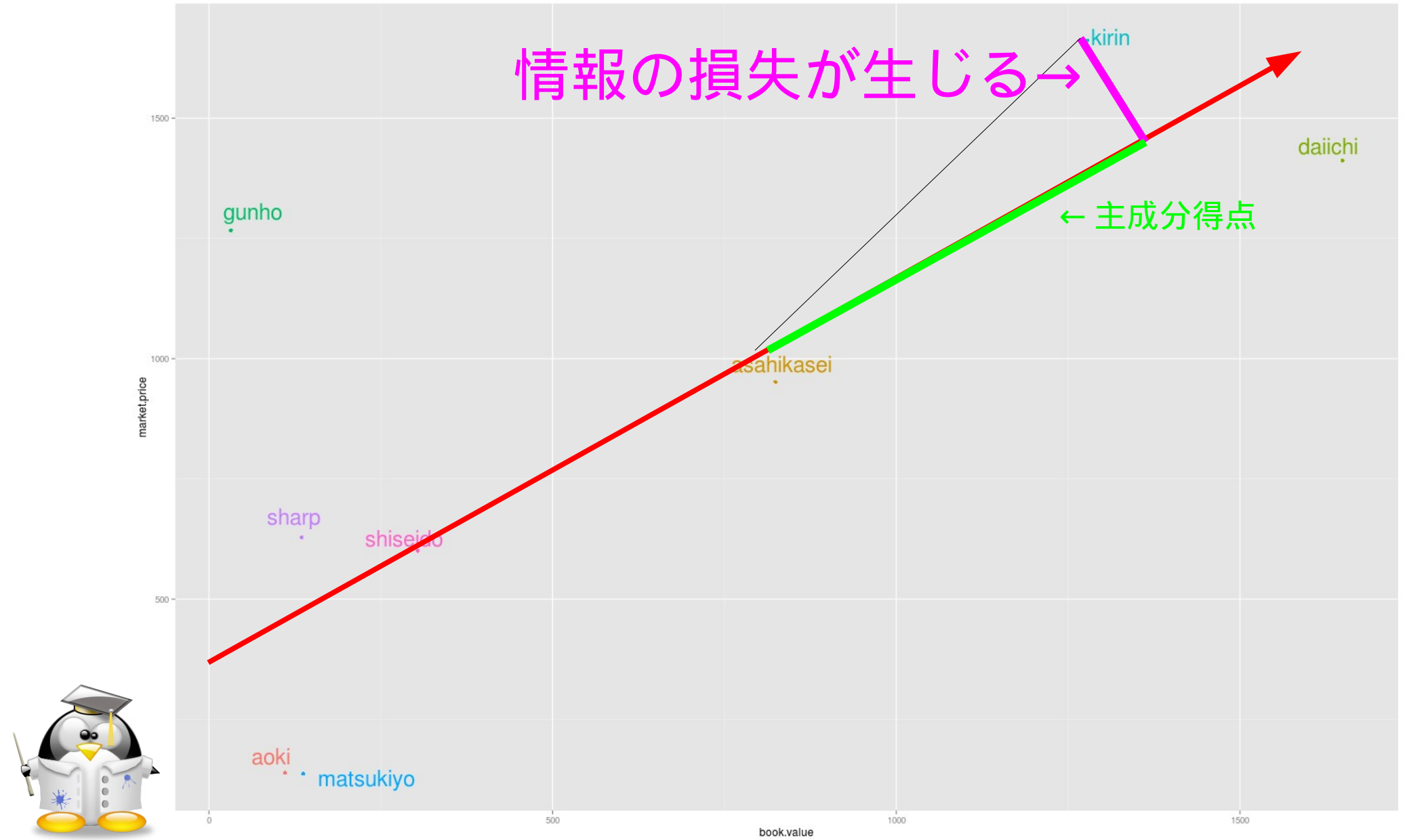

kirin
daiichi
gunho
もう１つの主成分→
＝第２主成分
← 最初に求めた主成分
＝第１主成分
asahikasei
この２つの方向ですべての情報を表現出来る！
sharp
shiseido
aoki
matsukiyo
market.price
book.value
1500
1000
500
0
500
1000
1500

2、主成分分析の簡単なお話

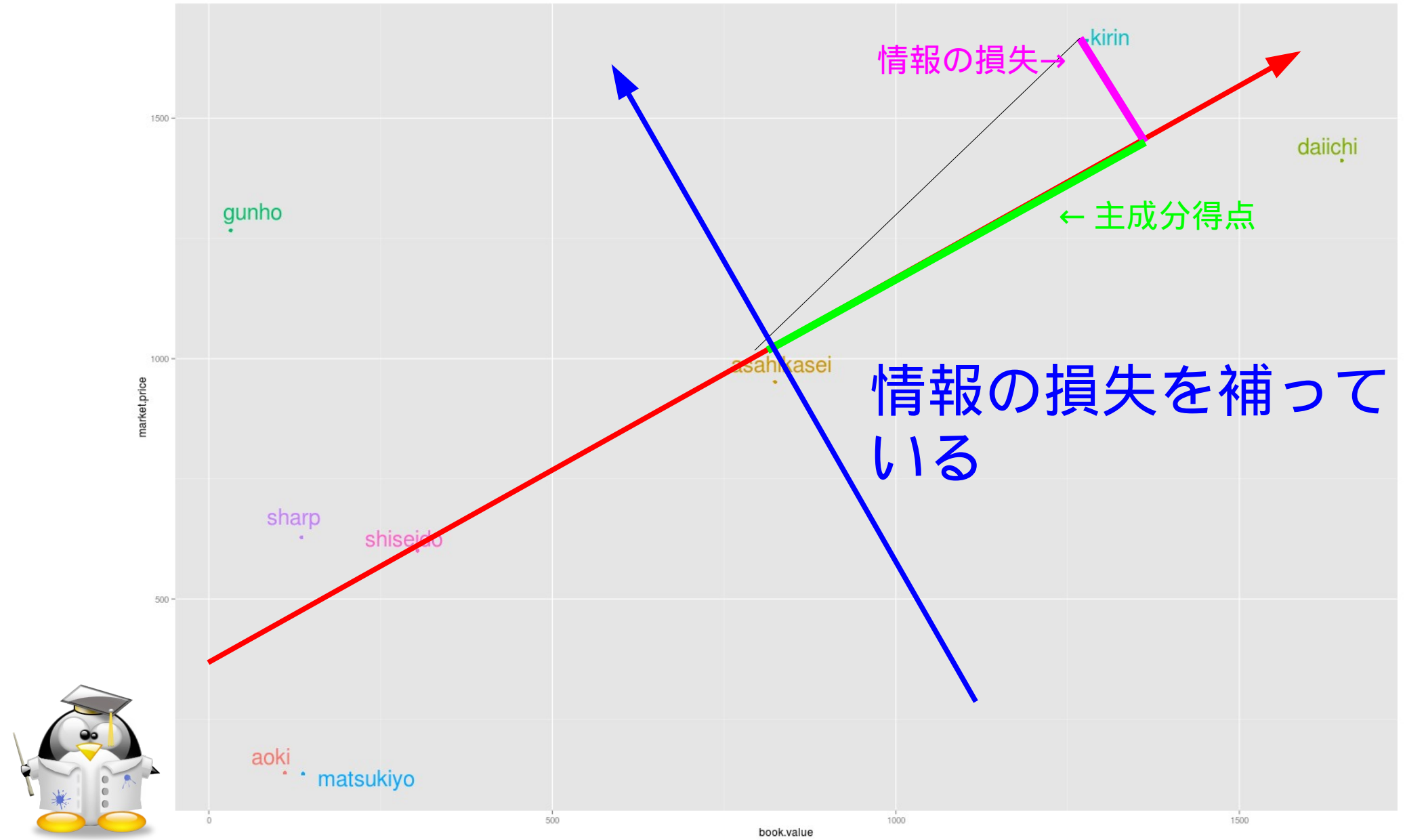
1本だけだと



1本だけだと


情報の損失が生じる→
kirin
daiichi
gunho
← 主成分得点
asahikasei
sharp
shiseido
aoki
matsukiyo
market.price
book.value
1500
1000
500
0
500
1000
1500

もう１本引くことで


情報の損失→
kirin
daiichi
gunho
← 主成分得点
asahikasei
情報の損失を補っている
sharp
shiseido
aoki
matsukiyo
market.price
book.value
1500
1000
500
0
500
1000
1500

ふと疑問

第１主成分だけでどれくらい
表現出来ているのか？

寄与率

・λは１つの主成分得点の分散を表している

・すべての主成分の分散により、すべてのデータの分散が表現できる

ので

第１主成分の寄与率（どれくらい説明出来ているか）は

$$\frac{\lambda_1}{\lambda_1 + \lambda_2}$$

で表せる！

今回は

・2 つの λ は 610211.2 と 98476.9 なので

第 1 主成分の寄与率 $= \dfrac{610211.2}{610211.2 + 98476.9} = 86.1\%$

第 2 主成分の寄与率 $= \dfrac{98476.9}{610211.2 + 98476.9} = 13.9\%$

主成分の解釈

・ところで主成分とは？？

・数学的な解釈はここまでなのであとは勘と経験で解釈する

$$z_1 = 0.67x_1 + 0.74x_2$$

$$z_2 = -0.74x_1 + 0.67x_2$$

z1 は時価総額・純資産共に高ければ高いほど良い

=> 企業の規模を表す（はず）

z2 は時価総額が低いほどよく、純資産が高いほどよい

=> 企業への期待の少なさを表す（はず）

R を用いて

~ R でやってみる ~

・ prcomp 関数にデータフレームを入れれば良い

```
> market.price<-c(1267,137,952,1662,139,601,1412,629)
> book.value<-c(32, 137, 824, 1278, 111, 304, 1649, 135)
> labs<-c("gunho", "matsukiyo", "asahikasei", "kirin", "aoki", "shiseido", "daiichi",
"sharp")
> data <- data.frame(market.price, book.value, row.names=labs)
> pca <- prcomp(data)
> pca
Standard deviations:
[1] 781.1601 313.8103

Rotation:
                   PC1        PC2
market.price -0.6672553 -0.7448291
book.value   -0.7448291  0.6672553
```

第 1 主成分　第 2 主成分

注) データを標準化して分析したい場合は
prcomp 関数の引数 scale に T を指定する

R を用いて

個々の主成分得点は x にアクセス

	PC1	PC2
gunho	114.0099	-662.16355
matsukiyo	789.8013	249.55516
asahikasei	-265.7094	100.92378
kirin	-1077.6130	-124.97101
aoki	807.8323	230.71686
shiseido	355.8084	15.38607
daiichi	-1187.1308	308.78797
sharp	463.0013	-118.23528

注）各々で平均が 0 になるよう調整されている

R を用いて

寄与率は要約でみる（λ（分散）の平方根、寄与率、累積寄与率を表示）

```
> summary(pca)
Importance of components:
                            PC1     PC2
Standard deviation      781.160 313.810
Proportion of Variance    0.861   0.139
Cumulative Proportion     0.861   1.000
```

Rを用いて

寄与率は要約でみる（λ（分散）の平方根、寄与率、累積寄与率を表示）

```
> summary(pca)
Importance of components:
                             PC1     PC2
Standard deviation     781.160 313.810
Proportion of Variance   0.861   0.139
Cumulative Proportion    0.861   1.000
```

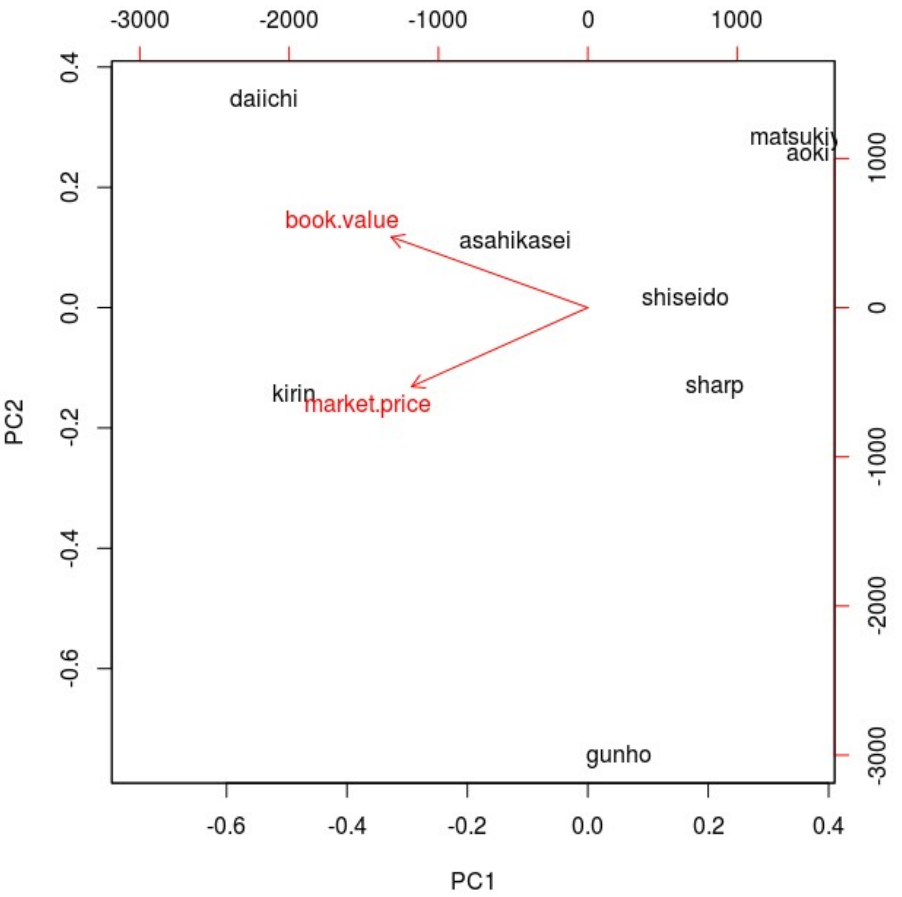
出来た！

R を用いて

可視化は biplot を利用するとよい

```
> biplot(pca)
```

主成分の 2 軸にそってデータを plot してくれる！



3、２次元から
多次元へ

多次元の主成分分析

- 基本的に 2 次元の主成分分析と変わらない
- 次元（軸）の数だけ主成分が出てくる
- 主成分得点の形:

$$z_1 = a_{11}x_1 + a_{12}x_2 \cdots a_{1n}x_n$$

$$z_2 = a_{21}x_1 + a_{22}x_2 \cdots a_{2n}x_n$$

$$\cdots$$

$$z_n = a_{n1}x_1 + a_{n2}x_2 \cdots a_{nn}x_n$$

実際にやってみる

6/1 に yahoo より作成

	安打	本塁打	打点	三振	四球
松本	27	3	9	19	4
山崎	10	1	6	8	1
田村	23	4	15	19	11
ブランコ	64	21	58	45	29
中村	54	7	25	23	19
後藤	7	2	4	7	2
荒波	35	0	5	26	8
鶴岡	19	0	11	13	7

データの作成

```
> anda<-c(27,10,23,64,54,7,35,19)
> honruida<-c(3,1,4,21,7,2,0,0)
> daten<-c(9,6,15,58,25,4,5,11)
> sanshin<-c(19, 8, 19, 45, 23, 7, 26, 13)
> shikyu<-c(4, 1, 11, 29, 19, 2, 8, 7)
> data <- data.frame(anda, honruida, daten, sanshin, shikyu, row.names=c("松本",
 "山崎", "田村", "ブランコ", "中村", "後藤", "荒波" ,"鶴岡"))
> data
         anda honruida daten sanshin shikyu
松本       27        3     9      19      4
山崎       10        1     6       8      1
田村       23        4    15      19     11
ブランコ   64       21    58      45     29
中村       54        7    25      23     19
後藤        7        2     4       7      2
荒波       35        0     5      26      8
鶴岡       19        0    11      13      7
```

結果

```
> pca <- prcomp(data, scale.=TRUE)
> pca
Standard deviations:
[1] 2.13758020 0.53075484 0.32502347 0.18494237 0.09594912

Rotation:
                PC1         PC2         PC3         PC4           PC5
anda      0.4387867  0.58619452  0.33480489 -0.5792111  0.1275241356
honruida 0.4425334 -0.57258817 -0.18286015 -0.4565481 -0.4841827800
daten     0.4537569 -0.43070140  0.09751365  0.1321312  0.7626489207
sanshin   0.4427319  0.37385746 -0.76934362  0.2689409 -0.0005057743
shikyu    0.4579559  0.05697088  0.50305997  0.6052193 -0.4094764930
> summary(pca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     2.1376 0.53075 0.32502 0.18494 0.09595
Proportion of Variance 0.9139 0.05634 0.02113 0.00684 0.00184
Cumulative Proportion  0.9139 0.97019 0.99132 0.99816 1.00000
```

結果

```
> pca <- prcomp(data, scale.=TRUE)
> pca
Standard deviations:
[1] 2.13758020 0.53075484 0.32502347 0.18494237 0.09594912

Rotation:
                PC1         PC2         PC3        PC4           PC5
anda      0.4387867  0.58619452  0.33480489 -0.5792111  0.1275241356
honruida 0.4425334 -0.57258817 -0.18286015 -0.4565481 -0.4841827800
daten     0.4537569 -0.43070140  0.09751365  0.1321312  0.7626489207
sanshin   0.4427319  0.37385746 -0.76934362  0.2689409 -0.0005057743
shikyu    0.4579559  0.05697088  0.50305997  0.6052193 -0.4094764930
> summary(pca)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     2.1376 0.53075 0.32502 0.18494 0.09595
Proportion of Variance 0.9139 0.05634 0.02113 0.00684 0.00184
Cumulative Proportion  0.9139 0.97019 0.99132 0.99816 1.00000
```

寄与率が 99% なので第３主成分まで考えてみる

結果

・第１主成分

z = 0.44× 安打 + 0.44× 本塁打 + 0.45× 打点 + 0.44× 三振 + 0.46× 四球

=> どれだけ試合に出場しているか

・第２主成分

z = 0.59× 安打－ 0.57× 本塁打－ 0.43× 打点 + 0.37× 三振 + 0.06× 四球

=> 短打力（逆は長打力）

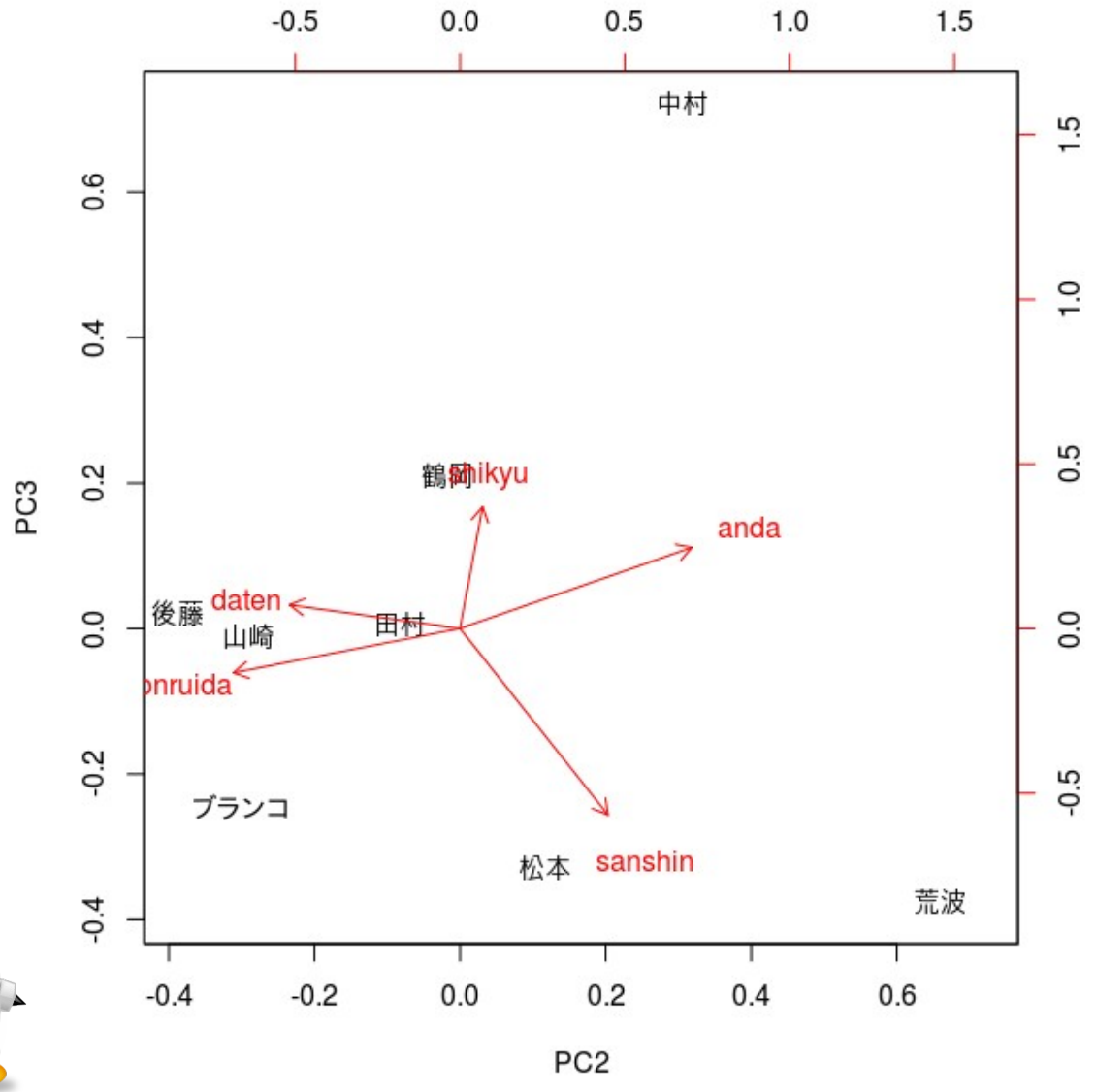
・第３主成分

z = 0.33× 安打－ 0.18× 本塁打 + 0.10× 打点－ 0.77× 三振 + 0.50× 四球

=> 逃げる力（三振をとにかく回避）

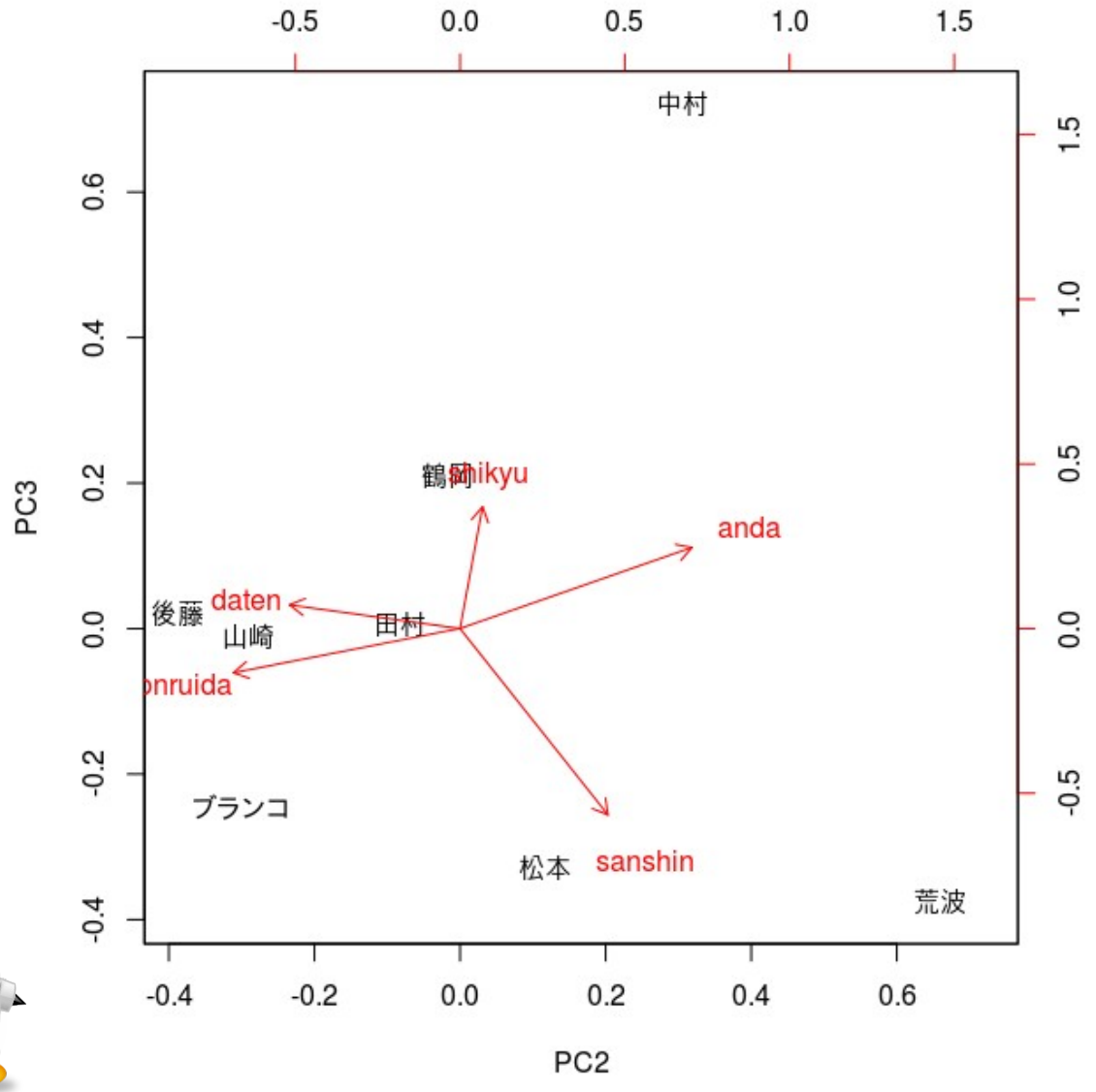
可視化

```
> biplot(pca, c=2:3)
```



可視化

```
> biplot(pca, c=2:3)
```



実は中村とブランコは
タイプが全然違う！！

４、量的データから
質的データに

これまでのもの

・データはすべて量的データであった

（例：売上、時価総額、点数、打数・・・）

・質的データ（ R の factor ）の分析は出来ないか？？

（例：美味しさ、清潔さ、香り、コク・・・）

=> 質的データも量的データに変換して考える

例えば

・アンケートをとる

（問）運転をしますか？

（１）する　（２）しない

（問）甘いものは好きですか？

（１）大好き　（２）好き　（３）好きではない

例えば

・アンケートをとる

（問）運転をしますか？

（１）する　（２）しない

=> １点、０点

（問）甘いものは好きですか？

（１）大好き　（２）好き　（３）好きではない

=> １点、０点、-1 点

それぞれを得点化する！

4、量的データから質的データに

とても良い記事

連載 Excelビジネス統計～アンケートの設計と分析～

【主成分分析最終回】缶コーヒー総合力1位はどれだ?「コク」「香り」「酸味」の主成分得点を求め、散布図を描いて解釈する

第13回　データを要約する主成分分析（4）散布図を描いてマーケティング解釈を行う

末吉　正成 [著]　2013/03/14 08:00

B! 7　ツイート 35　いいね! 30　+1 0　バックナンバー　印刷用を表示

マーケティングにはプル型、プッシュ型2つのリサーチが欠かせません。プッシュ型リサーチの代表「アンケート」を効果的に活用するため、ビジネス統計を学びましょう。

Excel ビジネス統計

1 2 3 4 5

今回は「相関係数行列」と「固有値と固有ベクトルの求めかた」をやります

いよいよ主成分分析シリーズも最終回。前回は、手順2「相関係数行列」と手順3「固有値と固有ベクトル」を求め、手順4「分析の精度を確かめる」を行いました。今回は、**「主成分得点を求め、散布図を描いて分析結果を解釈する」**を行います。

1.変数の単位を揃えるためにデータを基準化する

15	基準化			
16	缶コーヒー名	コク	香り	酸味
17	Sマルタ	-0.116248	1.2456822	1.5275252
18	モーニングS	-1.278724	-1.245682	0.0727393
19	BOSS	1.0462287	-0.415227	0.8001323
20	FIRE	1.0462287	0.4152274	-0.654654
21	サンタマルタ	1.0462287	1.2456822	1.5275252
22	BLACK無糖	0.4649906	-0.415227	-0.654654
23	UCCB	-1.278724	1.2456822	-1.382047
24	ジョージアB	-1.278724	-1.245682	-1.382047
25	ROOT	-0.697486	-1.245682	0.0727393
26	WANDA無糖	1.0462287	0.4152274	0.0727393
27	平均	-1.11E-16	1.665E-17	8.882E-17
28	標準偏差	1	1	1

データはすべて架空です。実在のものとは全く関係ありません。

Markezineの Excel ビジネス統計

4、量的データから質的データに

ということで拝借

	コク	香り	酸味
S マルタ	-0.116248	1.2456822	1.5275252
モーニング S	-1.278724	-1.245682	0.0727393
BOSS	1.0462287	-0.415227	0.8001323
FIRE	1.0462287	0.4152274	-0.654654
サンタマルタ	1.0462287	1.2456822	1.5275252
BLACK 無糖	0.4649906	-0.415227	-0.654654
UCCB	-1.278724	1.2456822	-1.382047
ジョージア B	-1.278724	-1.245682	-1.382047
ROOT	-0.697486	-1.245682	0.0727393
WANDA	1.0462287	0.4152274	0.0727393

データの作成

```
> KOKU<-c(-0.116248, -1.278724, 1.0462287, 1.0462287, 1.0462287, 0.4649906, -1.2
78724, -1.278724,-0.697486,1.0462287)
> KAORI <- c(1.2456822, -1.245682, -0.415227, 0.4152274, 1.2456822, -0.415227, 1
.2456822, -1.245682, -1.245682, 0.4152274)
> SANMI <- c(1.5275252, 0.0727393, 0.8001323, -0.654654, 1.5275252, -0.654654, -
1.382047, -1.382047, 0.0727393, 0.0727393)
> data <- data.frame(KOKU, KAORI, SANMI, row.names=c("Sマルタ", "モ―ニングS", "
BOSS", "FIRE", "サンタマルタ", "BLACK無糖", "UCCB", "ジョ―ジアB", "ROOT", "WAND
A"))
> data
                KOKU      KAORI      SANMI
Sマルタ       -0.1162480  1.2456822  1.5275252
モ―ニングS   -1.2787240 -1.2456820  0.0727393
BOSS           1.0462287 -0.4152270  0.8001323
FIRE           1.0462287  0.4152274 -0.6546540
サンタマルタ   1.0462287  1.2456822  1.5275252
BLACK無糖      0.4649906 -0.4152270 -0.6546540
UCCB          -1.2787240  1.2456822 -1.3820470
ジョ―ジアB   -1.2787240 -1.2456820 -1.3820470
ROOT          -0.6974860 -1.2456820  0.0727393
WANDA          1.0462287  0.4152274  0.0727393
```

結果

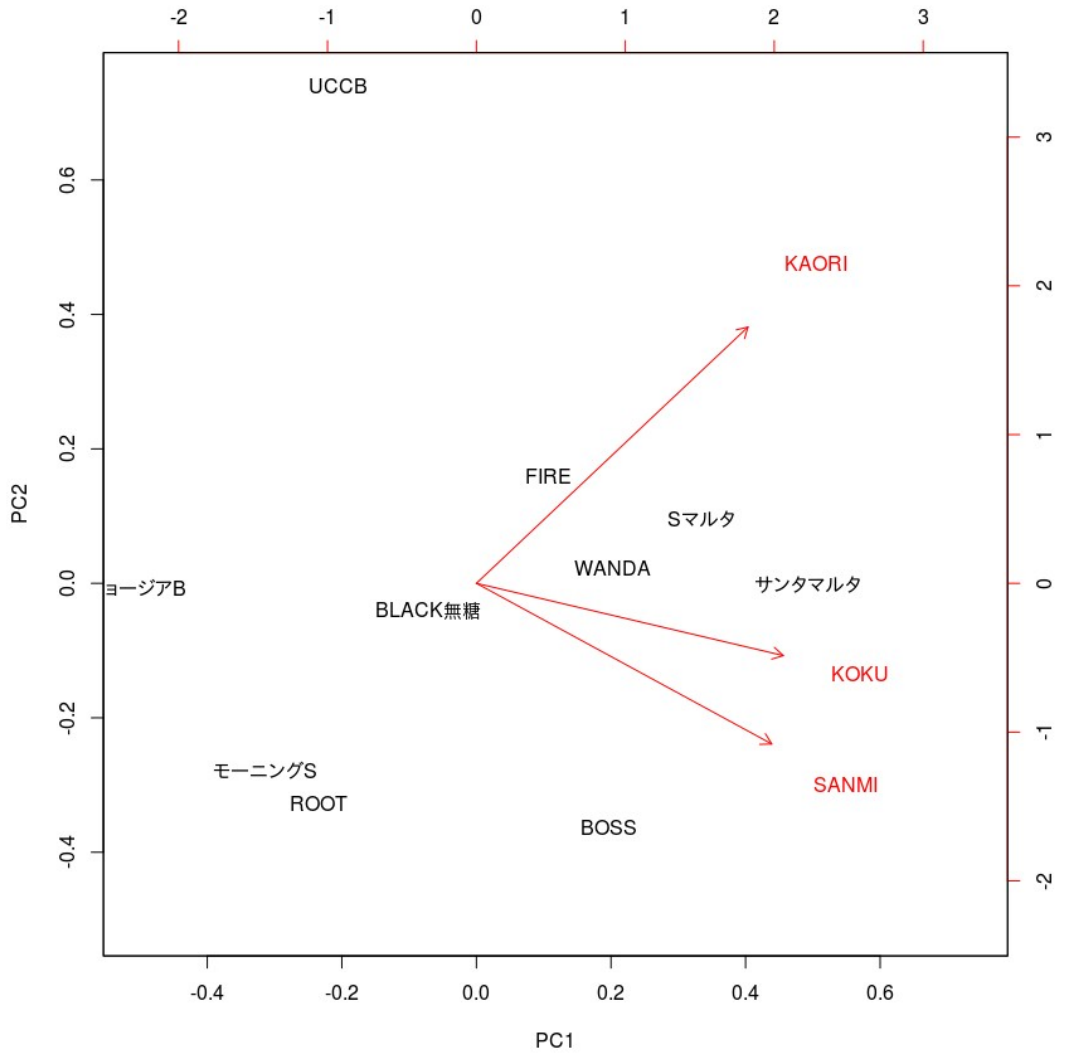
```
> pca <- prcomp(data, scale.=TRUE)
> pca
Standard deviations:
[1] 1.3407225 0.8263832 0.7208008

Rotation:
            PC1        PC2        PC3
KOKU  0.6074840 -0.2324076 -0.7595722
KAORI 0.5376966  0.8241644  0.1778634
SANMI 0.5846756 -0.5164686  0.6256315
> summary(prcomp(data.frame(anda, honruida, daten, sanshin, shikyu), scale.=TRUE
))
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     2.1376 0.53075 0.32502 0.18494 0.09595
Proportion of Variance 0.9139 0.05634 0.02113 0.00684 0.00184
Cumulative Proportion  0.9139 0.97019 0.99132 0.99816 1.00000
```

4、量的データから質的データに

結果

```
> biplot(pca)
```



Thank you

ご清聴ありがとうございました！