

「データサイエンスセミナー」

一般社団法人データクレイドル

RStudioデータ分析

森 裕 一

岡山理科大学 経営学部 経営学科／統計検定 事業委員

<mori@mgt.ous.ac.jp>

<http://www.mgt.ous.ac.jp/~mori/>

2018/12/19

※本セミナーは倉敷市による「高梁川流域インテリジェントICT実装事業2018」の一環として実施します。



ENJOY SCIENCE!

ボクラ、科学の子。

今日、ここにいる方々

当該分野のエキスパート

データを利用することに関心があり（高く）
ある程度の統計の知識（データ分析の経験）はある

R はトレンドなので（問題解決ツールとして役立つらしいので）、
使えるようになりたい



統計に対する一般的な感覚

業績アップ
問題特定
儲けが倍増

これを使えば、問題が解決する = 統計学は錬金術 **世の中にそんなものない**

でも、むずかしい

= 統計は数学

統計の道具が数学

ソフトさえ使いこなせば

= Rは魔法の道具

出力を解釈するのはあなた

もう少し進むと

失望...
絶望...
断念...

使ってみたけど、何も解決されない

やっぱり、統計はむずかしい



結局、データで解決なんて、できないじゃん

楽器／スポーツ／ダイエット／勉強…… ⇒ **まず手をつける**／そして**本質を理解して**／どれだけ**繰り返す**か ⇒ いっぱい見えてくる



錬金術になりうる！

公式より背景にある**概念**
それはとても**自然**

この**理解**で魔法の道具に



ENJOY SCIENCE

ボクラ、科学の子。

本日の内容

Part1 【入門】

R言語とは

RとRStudio（基本操作とデータ処理の初歩）

Part2 【実践】

Rによるデータ分析 1（傾向の把握と可視化）

Rによるデータ分析 2（相関、回帰、予測）

Rによるデータ分析 3（多変量解析 + a）

まとめ



Part1 【入門】

R言語とは

RとRStudio（基本操作とデータ処理の初歩）

Part2 【実践】

Rによるデータ分析 1（傾向の把握と可視化）

Rによるデータ分析 2（相関、回帰、予測）

Rによるデータ分析 3（多変量解析 + a）

まとめ



統計解析環境 R の特徴

- 統計解析専用ソフト
- フリーでオープンソース
 - フリー（無償）
 - = どこにでもインストールでき，利用しやすい
 - = 広く普及・浸透
 - オープンソース
 - = 数の論理で信頼性
 - = 新しいものや便利なものへの対応が積極的
- どのOSでもOK
 - = Windows, Mac, Linux で動作 = OSを気にしない
- 多くの書籍
 - = 「Rを用いた〇〇」で，すぐに理論理解と分析実施
- 多くの仲間
 - = 活用のコツや新しい取り組みなどの情報交換

関連URL

RjpWiki

<http://www.okadajp.org/RWiki/?RjpWiki>

The R Project for Statistical Computing

<https://www.r-project.org/>

CRAN

<https://cran.r-project.org/>

CRAN Japan mirror

<https://cran.ism.ac.jp/>



統計解析環境 R とビジネス用アプリケーション Excel



Rが得意なこと

- 大量データの処理
- データ解析 (統計解析)
- グラフ…統計グラフ
- GIS (地理情報処理)
- プログラミング (関数作成)
- シミュレーション
- ルーチンワーク処理
- Rからのドキュメント作成 (HTML/PDF/Word…)



Excelが得意なこと

- 集計 (ピボットテーブル)
- ビジネス文書処理
- グラフ…棒, 折れ線, 円などの基本グラフ
- 関数を使ったデータ変換
- フィルター
- データベース関数

Excelはオフィス (ビジネス) ソフト

もちろん統計分析はできる

ただし, ビジネスを前提

簡単なところでは, 散布図でラベル付けができない
多変量解析はできない (それ用のアドインが必要)

ところが, ビジネスシーンで本格的に統計が

⇒Excel2016では, 統計関係が強化

ex) ヒストグラム, 箱ひげ図



Part1 【入門】

R言語とは

RとRStudio (基本操作とデータ処理の初歩)

Part2 【実践】

Rによるデータ分析 1 (傾向の把握と可視化)

Rによるデータ分析 2 (相関、回帰、予測)

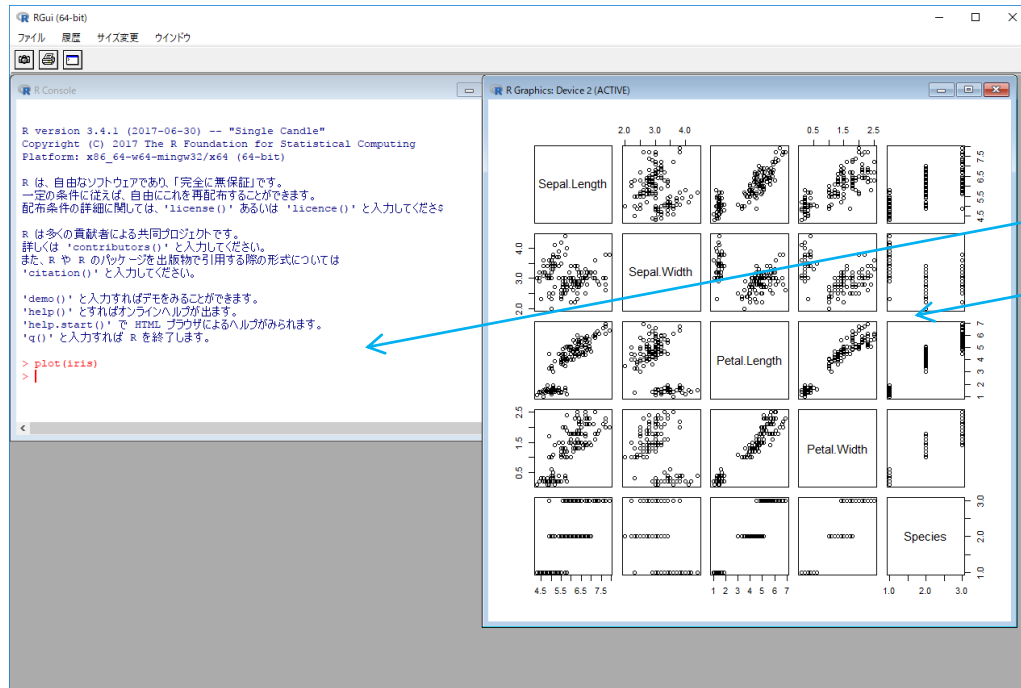
Rによるデータ分析 3 (多変量解析 + α)

まとめ



統計解析環境 R

<https://cran.ism.ac.jp/> などからOSに対応したRをダウンロード ⇒ インストール



Rの操作

基本的に、CUIベース

RのUI

コンソール

命令/計算結果

グラフィックデバイス

グラフ出力

基本的にこの2つ

基本計算

インストールすると最初から入っている関数
基本計算/基本グラフ

パッケージ

ない関数は誰かが作ってくれている。
それがパッケージ。
適宜インストールする。



基本操作

- RはCUI

```
>
```

> (←プロンプトという) 入力待ち状態

- 電卓

```
1+2
```

```
(3 * 4) ^ 2
```

見やすくするために、空白を入れてもよい

- 変数 (オブジェクト)

```
a <- 3 * 4
```

```
A <- 5
```

```
a
```

```
A
```

オブジェクト名は `a`, `b`, `product`, `A1`, `Month_12` のように、英字で始まる英数字からなる文字列。

`sin` や `sqrt` などの予約語、空白や演算記号は使えない。大文字／小文字は区別。

```
sales <- c(1,2,3,4,5)
```

```
sales <- 1:5
```

```
sales2 <- sales
```

```
idx <- "sales"
```

```
idx2 <- c("stock", "sales")
```

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



基本操作

● オブジェクトの操作

```
sales[2]
```

```
[1] 9
```

```
sales * 1.1
```

```
[1] 8.8 9.9 11.0 12.1 13.2
```

● 関数

```
sqrt(2)
```

```
[1] 1.414214
```

```
sqrt(sales)
```

```
[1] 2.828427 3.000000 3.162278 3.316625 3.464102
```

```
mean(sales)
```

```
[1] 10
```

● 練習 1

(1) $\frac{3^2-5}{4}$

(3) {4, 9, 25, 36} の平方根を求めよ。

(2) $a = 2, b = 3, c = 1$ のとき, $\frac{-b+\sqrt{b^2-4ac}}{2a}$ の値

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

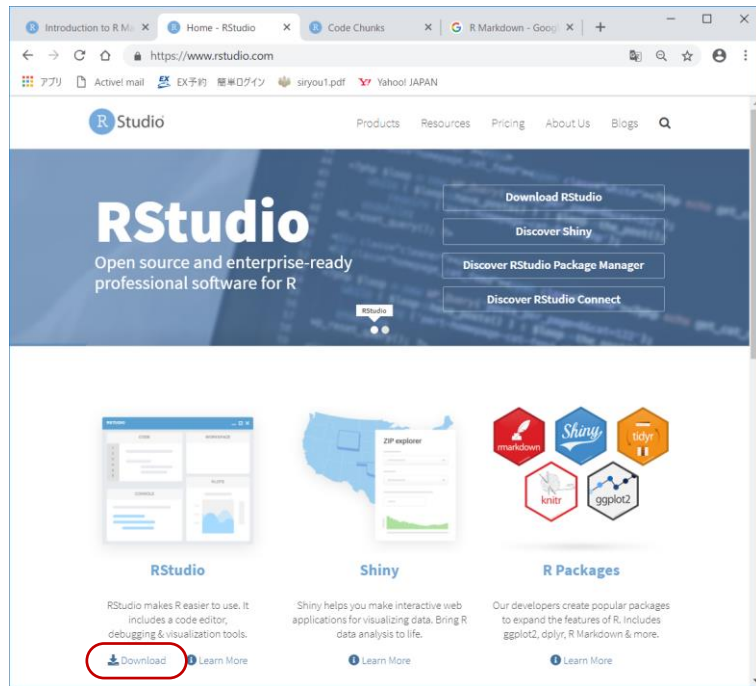
クラスター分析

ドキュメント生成



R Studio

<https://www.rstudio.com/> からダウンロード ⇒ インストール



R Studioとは

Rのための統合開発環境

直感的なユーザインターフェイス

+

強力なコーディングツール

Rstudioのメリット

- ・ 直感的なユーザインターフェイス

Rの中が見える（感じ）

- ・ 強力なコーディング

容易な編集

入力補助

関数や代数などの補完が便利

- ・ グラフのアウトプットのしやすさ

サイズ調整／クリップボード／出力形式png



インタフェース

4つのパネル (ペイン)

ソースエディター／データ表示

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for reading a table, creating a matrix 'mc', and generating histograms for different variables (a, ad, an, b, bd, bn, d, mc, n).
- Environment:** Lists objects in the global environment, including 'a' (120 obs. of 3 variables), 'ad' (60 obs. of 3 variables), 'an' (60 obs. of 3 variables), 'b' (120 obs. of 3 variables), 'bd' (60 obs. of 3 variables), 'bn' (60 obs. of 3 variables), 'd' (120 obs. of 3 variables), 'mc' (240 obs. of 3 variables), and 'n' (120 obs. of 3 variables).
- Console:** Shows the execution of the R code, including the command to read the table and the creation of the matrix 'mc'.
- Histogram:** A plot titled 'Histogram' showing the frequency distribution of 'weight'. The x-axis is labeled 'weight' and ranges from 98.5 to 100.5. The y-axis is labeled 'Frequency' and ranges from 0 to 40. The histogram bars are colored in a gradient from yellow to red.

オブジェクトの管理など
Rの中身が見れるという感じ
コマンドのヒストリーも

コンソール
結果の出力 直接入力も可能

ファイル一覧／グラフ出力／
パッケージ など



初めての...データ分析

- ファイルからのデータを読み込む。

```
dat <- read.csv("c:/R_work/grade1.csv")
```

[Tools]-[Global Options...]で作業ディレクトリを指定すれば、パスは不要。

```
dat <- read.csv("grade1.csv")
```

read.table でも可能。

```
dat <- read.table("c:¥R work/¥grade1.csv", header=T, sep=",")
```

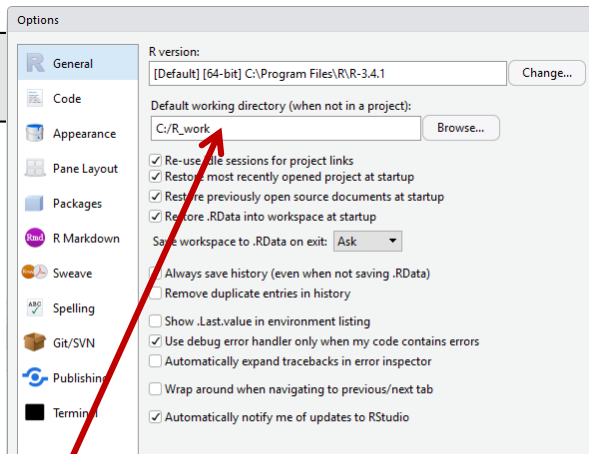
引数 header: 1行目はラベルか, sep: セパレータ (" , ")

```
dat # すべてが表示される
```

```
head(dat) # 最初の6行が表示される head(dat, 6)
```

	NO	Report	Participation	Written	Practice
1	R001	21		24	27
2	R002	19		20	22
3	R003	17		20	14
4	R004	17		24	14
5	R005	26		24	26
6	R006	12		24	22

CSV形式のデータファイルが
C:¥R_work
に入っている場合



ここに作業ディレクトリを指定する

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



ENJOY SCIENCE!

ボクラ、科学の子。

岡山理科大学
OKAYAMA UNIVERSITY OF SCIENCE

初めての...データ分析

- 平均点の計算

```
apply(dat[,2:5], 2, mean)
```

Report	Participation	Written	Practice
20.171053	22.131579	19.434211	7.513158

受講した76人の4つの観点の平均点は、レポート点が20.2点、平常点が22.1点、筆記試験が19.4点、実技試験が7.5点であることがわかる。また、それぞれ満点が30点、30点、30点、10点であるから、得点の取得率が67.3%、76.2%、64.7%、75.5%となり、筆記試験とレポートが他の2つより悪いことがわかる。

- レポート点 (datの2列目) と筆記試験得点 (datの4列目) の散布図を描く。

```
plot(dat[,2], dat[,4])
```

これより、レポート点と筆記試験得点には、正の相関があること、レポート点も筆記試験も成績が悪いグループが存在すること、レポート点は悪かったが、筆記試験をがんばった人が1人いること、逆に、レポート点はよかったのに、筆記試験では点が取れなかった人が7人いることなどがわかる。

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

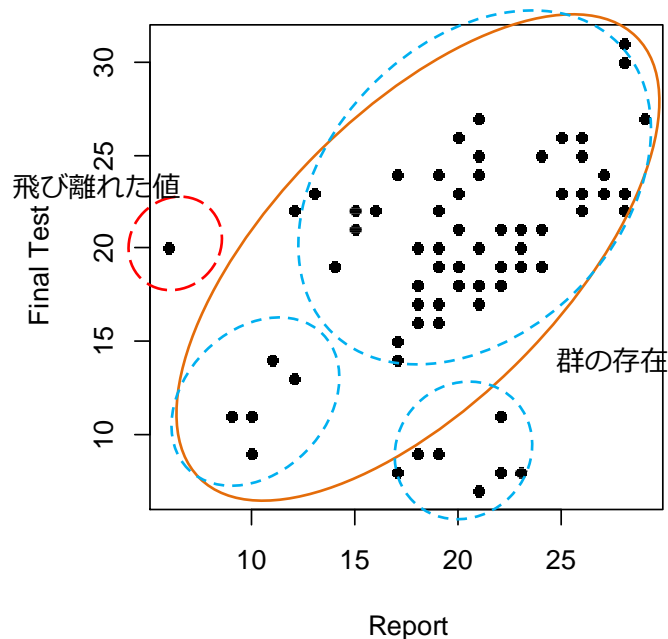
クラスター分析

ドキュメント生成

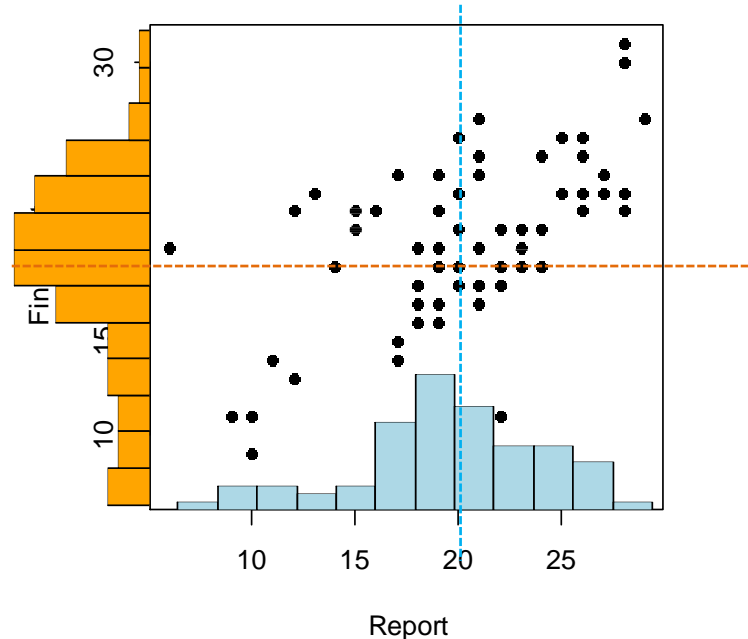


初めての...データ分析

右上がり、直線傾向



x, y それぞれで
平均や度数分布が見える？



初めての...データ分析

● 図の取り込み

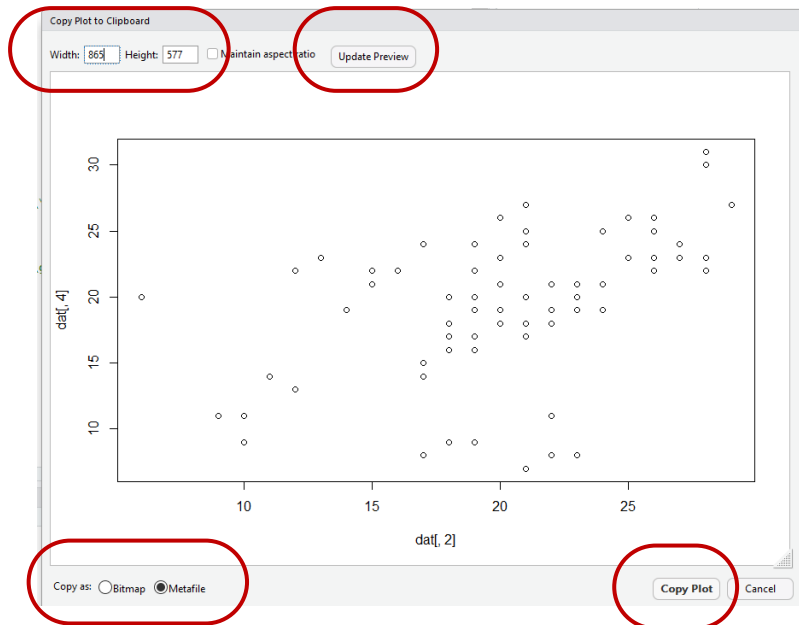
右下のグラフ出力のペインの [Export▼] をクリック。

Image, PDF, Clipboard が選べる。

右は, Copy to Clipboard...

表示されたダイアログボックスで, 必要なら, 縮尺を決める。[Update Preview] ボタンで画像がリサイズされる。

Clipboardの場合は, 形式を選択後, [Copy Plot] で, クリップボードにコピーされる。



練習 2

- (1) `grade1.txt` の各観点の標準偏差 (`sd()` を使う) を求めよ。
- (2) `grade1.txt` の平常点と筆記試験得点の散布図を描け。

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



Part1 【入門】

R言語とは

RとRStudio（基本操作とデータ処理の初歩）

Part2 【実践】

Rによるデータ分析 1（傾向の把握と可視化）

Rによるデータ分析 2（相関、回帰、予測）

Rによるデータ分析 3（多変量解析 + α ）

まとめ



傾向の把握

顧客データから各支店の様相をつかむ。

- データ"customer.csv"の読み込み。

5つの支店の顧客データ : 150顧客 | 6変数 (顧客番号, 支店名, 平均滞在時間, 来店回数, 性別, 購入総額)

```
cst <- read.csv("customer.csv")
```

- 中身の確認。

```
cst      または      head(cst)
```

- とりあえず, 要約。

```
summary(cst)
```

No	Branch	Time	Visit	Sex	Purchase
Min. :11001	岡山 :30	Min. : -1.00	Min. :1.000	女:72	Min. : -210
1st Qu.:12008	岡山南:30	1st Qu.:19.00	1st Qu.:4.000	男:78	1st Qu.:1990
Median :13016	玉島 :30	Median :28.00	Median :5.000		Median :2790
Mean :13016	児島 :30	Mean :27.74	Mean :4.653		Mean :2610
3rd Qu.:14023	倉敷 :30	3rd Qu.:35.00	3rd Qu.:6.000		3rd Qu.:3340
Max. :15030		Max. :60.00	Max. :9.000		Max. :4590

質的変数に対してはカテゴリーとその要素数が, 量的変数に対しては5数要約が表示される。

(Noは量的変数として認識されている。)

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



傾向の把握

- 性別で購入金額に違いはあるか？

```
tapply(cst$Purchase, cst$Sex, mean)
```

女 男

2717.778 2510.513

女性の方が高い。

- 支店別の購入金額は？

```
tapply(cst$Purchase, cst$Branch, mean)
```

岡山 岡山南 玉島 児島 倉敷

2643.333 2583.333 2323.333 2690.000 2810.000

倉敷支店の成績が一番良く、玉島支店が一番低い。

- 男女差は支店別に見ても同じか？

```
tapply(cst$Purchase, list(cst$Sex, cst$Branch), mean)
```

岡山 岡山南 玉島 児島 倉敷

女 2740.000 2606.667 2767.778 2828.095 2590.0

男 2532.857 2567.778 2132.857 2367.778 3002.5

支店別に見ても女性の購入金額の方が全体に高いことがわかる。でも、倉敷支店は男性の方が高い。

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



傾向の把握

- 購入金額の分布は？

```
hist(cst$Purchase)
```

- 箱ひげ図では？

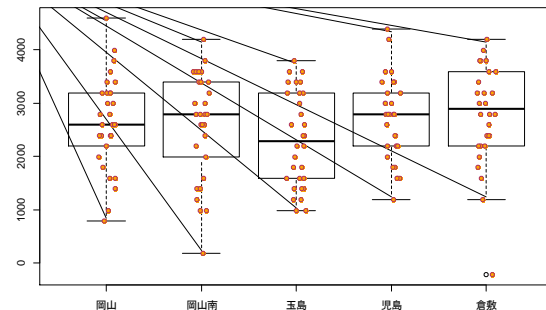
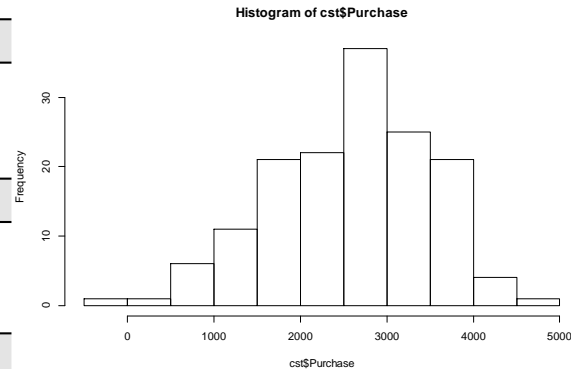
```
boxplot(cst$Purchase)
```

- 支店により購入金額の分布の仕方に差はあるか？

```
boxplot(cst$Purchase~cst$Branch)
```

- ジッターリングで分布の様子をより詳しく

```
stripchart(cst$Purchase~cst$Branch, vertical = TRUE, pch = 21, col = "maroon",  
           bg = "orange", method = "jitter", add = TRUE)
```



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



ENJOY SCIENCE!

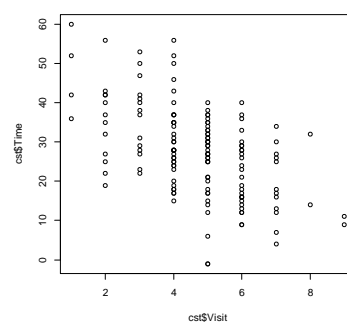
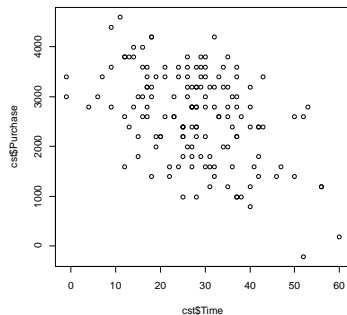
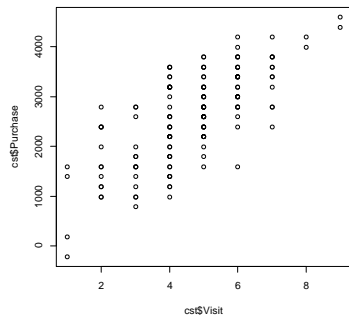
ボクラ、科学の子。

岡山理科大学
OKAYAMA UNIVERSITY OF SCIENCE

傾向の把握

- 平均滞在時間や来店回数は購入金額に関係するか？

```
plot(cst$Visit, cst$Purchase)
plot(cst$Time, cst$Purchase)
plot(cst$Visit, cst$Time)
```



- 相関係数も見ておく。

```
cor(cst[,c(3,4,6)])
```

	Time	Visit	Purchase
Time	1.0000000	-0.5358738	-0.4479815
Visit	-0.5358738	1.0000000	0.7231684
Purchase	-0.4479815	0.7231684	1.0000000

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



傾向の把握

タイタニック号の乗船者の様相をつかむ。

- データ"titanic.csv"の読み込み。
2201人 | 4変数 (乗船クラス, 年齢 (大人か小人か), 性別, 生死)

```
ttn <- read.csv("Titanic.csv")
```

- 中身の確認。

```
head(ttn)
```

- とりあえず, 要約。

```
summary(ttn)
```

Class	Age	Sex	Survive
1等船室:325	子供: 109	女性: 470	死亡:1490
2等船室:285	大人:2092	男性:1731	生存: 711
3等船室:706			
乗組員 :885			

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



傾向の把握

- クロス集計をして、生死の原因を探る。

```
table(ttn$Class, ttn$Survive)
```

	死亡	生存
1等船室	122	203
2等船室	167	118
3等船室	528	178
乗組員	673	212

```
table(ttn$Age, ttn$Survive)
```

	死亡	生存
子供	52	57
大人	1438	654

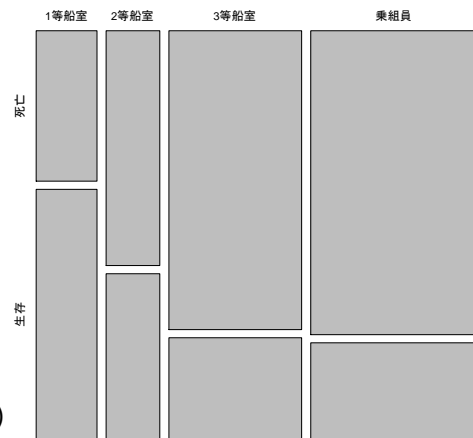
```
table(ttn$Sex, ttn$Survive)
```

	死亡	生存
女性	126	344
男性	1364	367

- モザイクプロット (上記Tableの結果をmosaicplot()の引数に指定する。)

```
mosaicplot(table(ttn$Class, ttn$Survive))
```

table(ttn\$Class, ttn\$Survive)



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



傾向の把握

練習 3

- (1) 最近, 加工品の完成精度が下がっているとの報告があった。
重量にばらつきが出始めているとのこと。
工場では, 2つのマシンを昼夜交代制で24時間稼働させている。
240個の製品を抜き取って重さを測った。
不良品が出るのは, どこに原因があるか調べよ。

場合分け
平均
分散 (標準偏差)
分布
ヒストグラム
箱ひげ図

- (2) 顧客データで, 支店ごとの男女数をクロス表の形で求めよ。

```
MachineID,Period,Weight
A,Day,99.88
A,Day,100.09
A,Day,99.88
A,Day,100.05
:
:
A,Night,99.92
A,Night,100.24
A,Night,99.93
A,Night,100.18
:
:
B,Day,99.37
B,Day,99.38
B,Day,99.50
B,Day,99.67
:
:
B,Night,98.85
B,Night,99.41
B,Night,99.33
B,Night,99.21
:
```

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



可視化 (さまざまなグラフ出力)

描画パラメータの指定

- ヒストグラム (加工品データのヒストグラムで)

```
mcn <- read.csv("machine5.csv")
```

```
a <- mcn[mcn$MachineID=="A",]
```

```
b <- mcn[mcn$MachineID=="B",]
```

```
d <- mcn[mcn$Period=="Day",]
```

```
n <- mcn[mcn$Period=="Night",]
```

```
hist(a$Weight, breaks=seq(98.5,100.5,0.1), border="#990000", col="#99343550", main="Histogram",  
      xlab="weight", ylim=c(0,40))
```

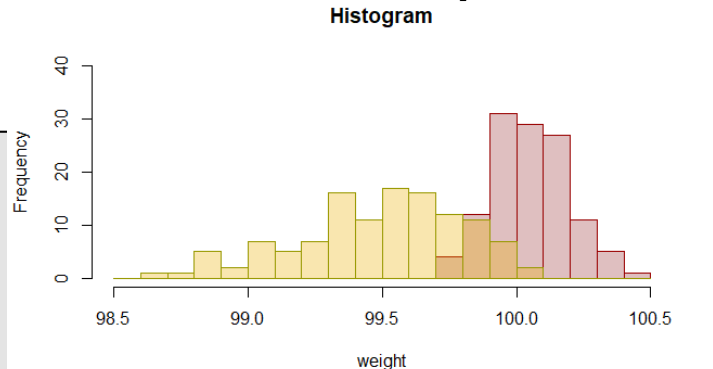
```
hist(b$Weight, breaks=seq(98.5,100.5,0.1), border="#999900", col="#edae0050", add=T)
```

```
hist(d$Weight, breaks=seq(98.5,100.5,0.1), border="#009900", col="#53995250", main="Histogram",  
      xlab="weight", ylim=c(0,40))
```

```
hist(n$Weight, breaks=seq(98.5,100.5,0.1), border="#000099", col="#5399ff50", add=T)
```

場合分けをして、AマシンとBマシン、昼と夜のヒストグラムを色を変えて、重ねて描画することができる。

- 散布図でも棒グラフでも指定ができる。
- Latticeパッケージもグラフ/グラフオプションが豊富。



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成





可視化 (さまざまなグラフ出力)

ggplot2を使ってみよう。

- パッケージのインストール。

`library(***)` とやってエラーが出たら, そのパッケージを取りにいく。

[Tools] – [Install Packages...]

で表示されたダイアログボックスで, パッケージ名を入れる。

- パッケージの読み込み。

```
library(ggplot2)
```

- ggplotの振る舞い

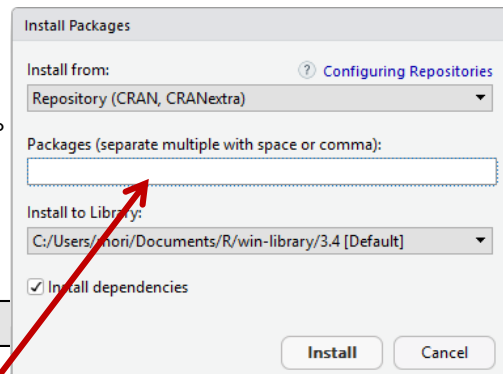
`ggplot()` でキャンバスを準備 (ここで座標の各種設定を行う)

`+geom_**()` 関数でグラフ (レイヤー) を重ね描きしていく。

正確には, データを幾何学的オブジェクト (geometric object) に当てはめて可視化する。

主なgeom :

<code>geom_bar()</code>	棒グラフ
<code>geom_line()</code>	折れ線グラフ
<code>geom_point()</code>	散布図
<code>geom_boxplot()</code>	箱ひげ図



ggplot2と入れて[Install]
[g]と打つだけで, 存在するパッケージ名が補完される

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



ENJOY SCIENCE

ボクラ、科学の子。

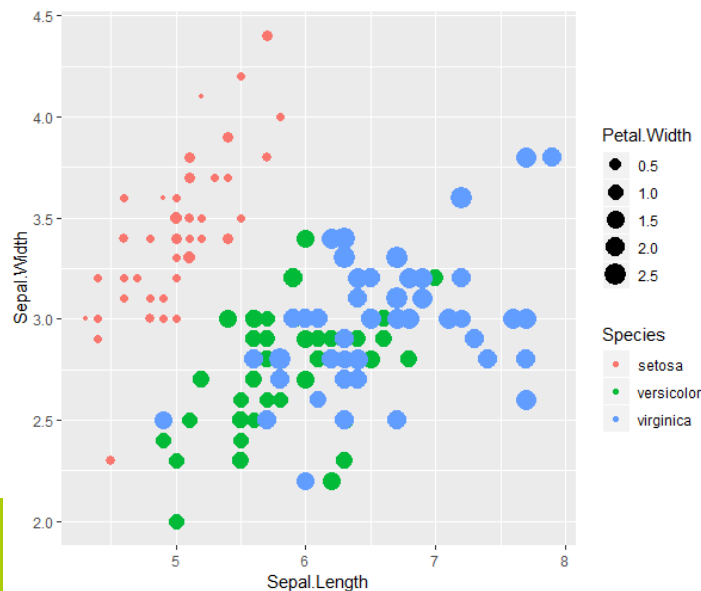
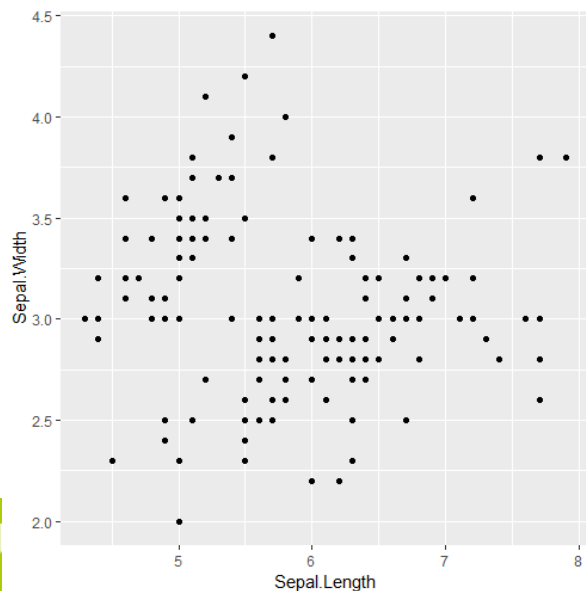
可視化 (さまざまなグラフ出力)

- データは、Rにデフォルトのあやめ"iris"のデータを使う。
150個体 | 5変数 (がくの長さ、がくの幅、花弁の長さ、花弁の幅、品種)
- 普通の散布図

```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width)) + geom_point()
```

- 4つの情報を載せた散布図

```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, colour=Species, size=Petal.Width)) + geom_point()
```



<http://www.ggplot2-exts.org/gallery/>
にはたくさんのサンプルが

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



Part1 【入門】

R言語とは

RとRStudio（基本操作とデータ処理の初歩）

Part2 【実践】

Rによるデータ分析 1（傾向の把握と可視化）

Rによるデータ分析 2（相関、回帰、予測）

Rによるデータ分析 3（多変量解析 + α ）

まとめ



相関と回帰

アイスクリームの売れ行き（1人当たりの支出金額）を決める気象要因をつきとめる。

- データ"IceCream.csv"を読み込む。

(60か月分 | 7変数 (年, 月, 月平均気温 (°C), 降水量の合計(mm), 日照時間(時間), 平均風速(m/s),
アイスクリーム支出金額 (円))

```
ice <- read.csv("IceCream.csv")
```

- 相関係数を求め、支払金額にきいている気象要因を特定する。

```
cor(ice)
```

	Year	Month	Temp	Rain	Sun	Wind	Paid
Year	1.000000e+00	0.00000000	-1.455626e-17	-0.17961282	0.017356740	-0.10339936	0.028964334
Month	0.000000e+00	1.00000000	3.735475e-01	0.07073238	-0.109945984	-0.38729098	0.222026682
Temp	-1.455626e-17	0.37354753	1.000000e+00	0.29671643	-0.144678778	-0.37269477	0.905154304
Rain	-1.796128e-01	0.07073238	2.967164e-01	1.00000000	-0.474815962	0.24365482	0.102697373
Sun	1.735674e-02	-0.10994598	-1.446788e-01	-0.47481596	1.000000000	0.09581258	-0.002532996
Wind	-1.033994e-01	-0.38729098	-3.726948e-01	0.24365482	0.095812580	1.00000000	-0.306280170
Paid	2.896433e-02	0.22202668	9.051543e-01	0.10269737	-0.002532996	-0.30628017	1.000000000

(予想通り) 月平均気温との相関が一番高い。

- この際、散布図行列も描いておく。

```
pairs(ice)
```

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

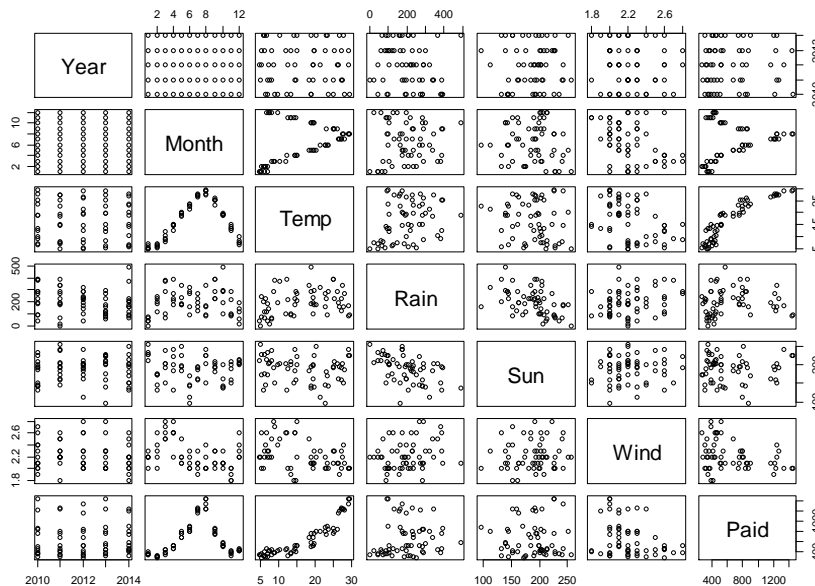
クラスター分析

ドキュメント生成

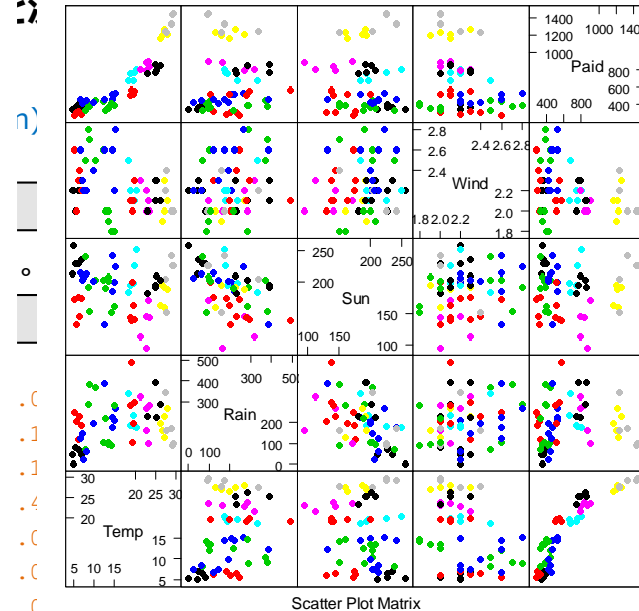


相関と回帰

ア.



イ.



（予想通り）月平均気温との相関が一番高い。

- この際、散布図行列も描いておく。

```
pairs(ice)
```

```
library(lattice)
splom(~ice[3:7], groups = ice$Month, pch=16,
      col=c(1,2,3,4,5,6,7,8,9,10,11,12))
```

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化（グラフ）

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成

回帰と予測

- 「アイスクリームの支出金額」を y , 「月平均基本」を x として, 回帰分析を行う。

```
lm.ice <- lm(Paid~Temp, data=ice)
summary(lm.ice)
```

Residuals:

Min	1Q	Median	3Q	Max
-243.384	-107.689	2.523	121.180	309.985

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	20.332	43.116	0.472	0.639
Temp	37.900	2.337	16.217	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 143.2 on 58 degrees of freedom

Multiple R-squared: 0.8193, Adjusted R-squared: 0.8162

F-statistic: 263 on 1 and 58 DF, p-value: < 2.2e-16

これより, 回帰式は, $y = 20.332 + 37.900x$ となる。

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



回帰と予測

- この回帰直線を表す。(月による違いがありそうなので, Monthでラベルをつけておく。)

```
plot(ice$Temp, ice$Paid)
abline(lm.ice$coef)
text(ice$Temp, ice$Paid, ice$Month, pos=2, col=3)
```

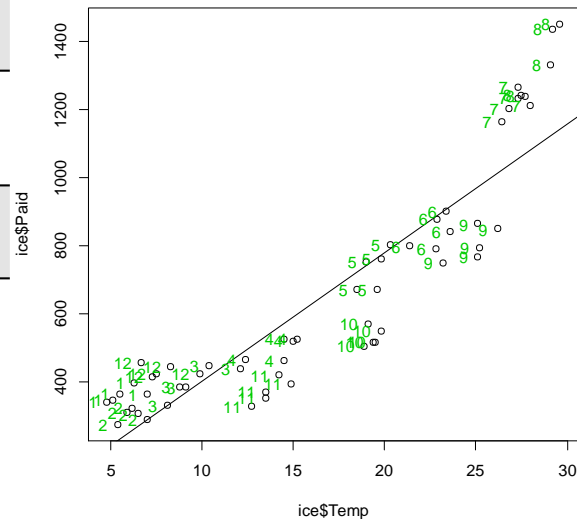
- 月平均気温が17℃の場合の支出金額を予測してみる。

```
x <- c(1, 17)
t(lm.ice$coef) %*% x
```

[,1]

[1,] 664.6334

665円となることが予測される。



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



回帰と予測

練習 4

新しい店舗を建てたい。
最も高い売上げが期待できる店舗が
最初の候補となる。
そのために、既存のチェーン店15店舗
の売上高と6つの調査観点に関する
データ "store.csv" を用いて、
3つの候補の売上げを予測し、
第1候補地を決めよ。

重回帰分析
回帰式
予測

		通行人数	駅からの時間	店舗面積	駐車台数	従業員数	品数	売上高
	branch	numpass	minutes	area	parkcar	numwork	kinds	sales
1	三条	716	16	44	16	7	125	78
2	京都南	2018	30	25	8	3	132	34
3	長岡京	1880	3	68	18	10	110	145
4	生駒	1416	20	30	10	5	70	51
5	高槻	904	10	67	27	10	82	98
6	枚方	1250	2	66	10	10	82	115
7	池田	1039	15	52	15	7	82	75
8	東大阪	2394	1	113	50	20	125	258
9	堺	711	12	30	12	7	102	70
10	八尾	738	10	39	10	7	70	65
11	和歌山	1322	11	60	23	4	72	82
12	宝塚	813	12	34	10	3	97	32
13	西宮	1733	3	96	40	10	145	190
14	西神	1569	5	55	28	10	92	168
15	加古川	1770	6	80	32	8	80	195

支店名	通行人	駅からの時間	店舗面積	駐車台数	従業員数	品数
候補1	1956	3	88	42	10	120
候補2	1300	12	90	45	10	100
候補3	1423	8	42	36	10	90

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



回帰と予測

練習 5

中国地方の最高気温と電力消費量（2017年7～8月）の散布図は右図のとおりで、

相関係数は、0.497

回帰式は、 $y = 139.588 + 23.649x$

である。

が、何となく2つの群が見える。

2つの群に分けて、それぞれで回帰分析を行い、それぞれの相関係数と回帰式を求めよ。

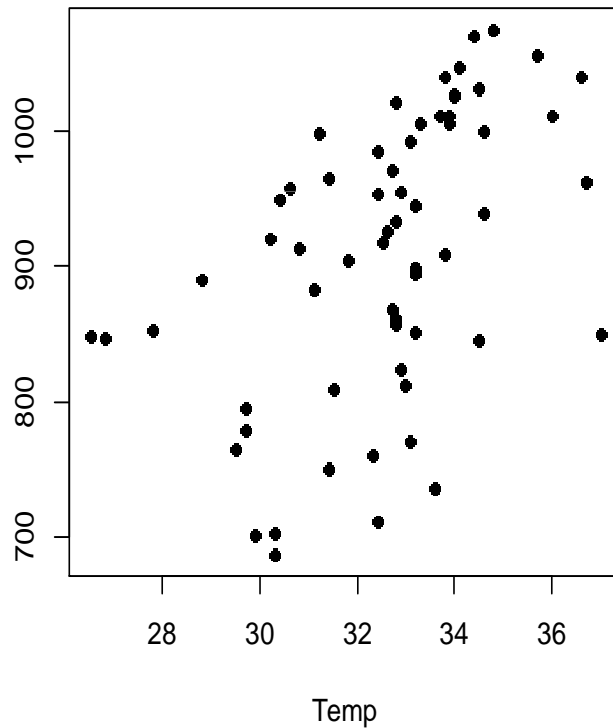
電力…平日と休日では異なるのでは？

- ⇒ HとNで分けて分析
- ⇒ 8月…お盆がある
- ⇒ 新たに分類して回帰

Temp, Power.Max, Day

33.1, 770, H
31.4, 750, H
32.8, 933, N
28.8, 890, N
27.8, 852, N
26.8, 847, N
26.5, 848, N
29.5, 765, H
30.3, 702, H
31.1, 883, N
:
:

Power.Max



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



Part1 【入門】

R言語とは

RとRStudio（基本操作とデータ処理の初歩）

Part2 【実践】

Rによるデータ分析 1（傾向の把握と可視化）

Rによるデータ分析 2（相関、回帰、予測）

Rによるデータ分析 3（多変量解析 + α ）

まとめ



主成分分析

都道府県別アルコール類の年間消費量をもとに、種類に関する都道府県の特徴を明らかにする。

- データ"sake.csv"を読み込む。
47都道府県 | 6変数 (都道府県, 清酒, 焼酎, ビール, 果実酒, ウイスキー)
都道府県名 (第1変数) を行の名前にしておく。

```
sak <- read.csv("sake.csv", row.names="Pref")
```

- 主成分分析を行う。酒類によって量が大きく違うので、標準化して実行する。

```
pca.sak<-prcomp(sak, scale=T)
```

```
pca.sak
```

```
Standard deviations (1, .., p=5):
```

```
[1] 2.1348134 0.5393196 0.3121795 0.1978397 0.1229207
```

```
Rotation (n x k) = (5 x 5):
```

	PC1	PC2	PC3	PC4	PC5
Sake	-0.4440460	-0.48996881	0.4096785	0.6197155	0.10425933
Shochu	-0.4154208	0.83227891	0.3414540	0.1271593	0.04446461
Bear	-0.4575347	-0.23743062	0.2528134	-0.7412063	0.34783235
Wine	-0.4530736	0.05039986	-0.7873631	0.1721826	0.37761969
Whisky	-0.4643652	-0.09126303	-0.1780939	-0.1440474	-0.85062854

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



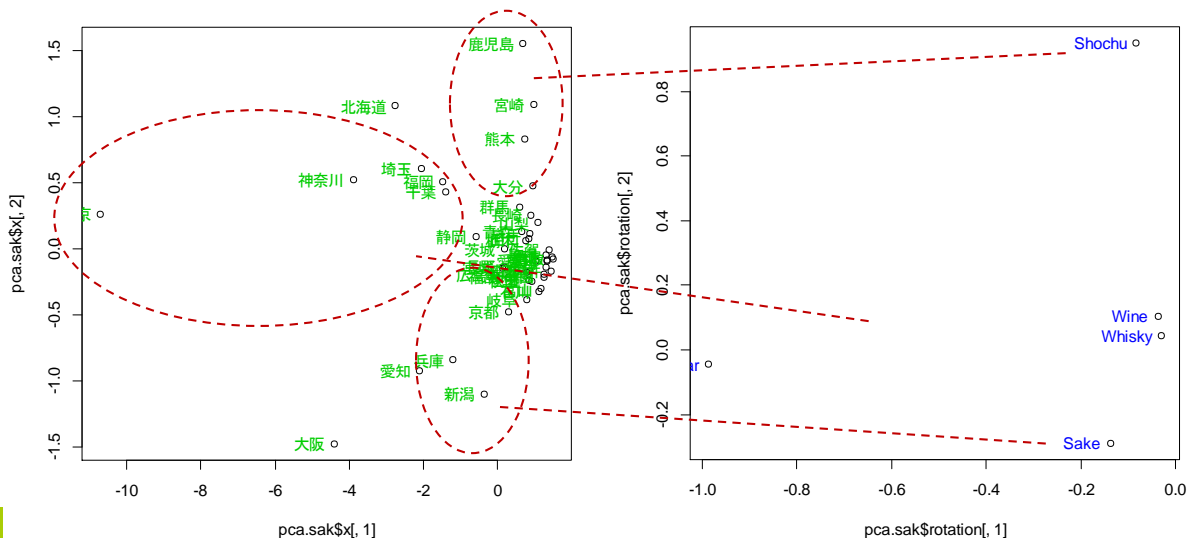
主成分分析

- 主成分得点のグラフを描く。都道府県名をラベルとしてプロットに加えておく。

```
plot(pca.sak$x[,1], pca.sak$x[,2])
text(pca.sak$x[,1], pca.sak$x[,2], rownames(sak), pos=2, col=3)
```

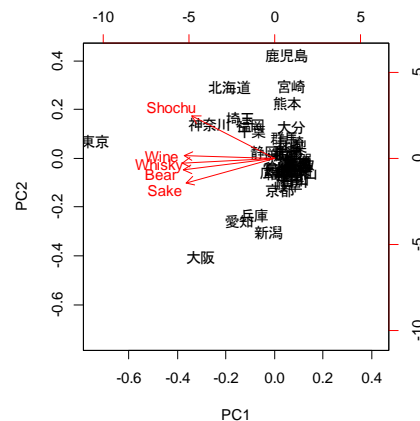
- 負荷量のグラフを描く。変数名でラベル付けしておく。

```
plot(pca.sak$rotation[,1], pca.sak$rotation[,2])
text(pca.sak$rotation[,1], pca.sak$rotation[,2], colnames(sak), pos=2, col=3)
```



- バイプロットもある。

```
biplot(pca.sak)
```



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成

コレスポンデンス分析

麺つゆ各銘柄の特徴をポジショニングし、販売戦略に活かす。

- データ"ndsoup.csv"を読み込む。

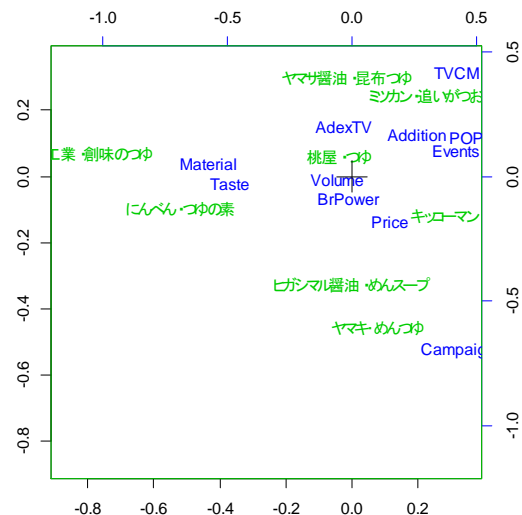
8銘柄 | 12変数 (商品名, 商品の味, テレビCM, ブランド力, 商品の素材・製法, 商品の容量, テレビCM以外の広告, 増量品などキャンペーン品, 希望小売価格, 消費者キャンペーン・イベント, POP等店頭即売物, おまけ・レシピ等)
商品名 (第1変数) を行の名前にしておく。

```
nds <- read.csv("ndsoup.csv", row.names="Brand")
```

- コレスポンデンス分析 (対応分析) を行う。
コレスポンデンス分析は, "MASS"ライブラリーに入っている。
何次元までとるかは, nf=で指定する。

```
library(MASS)
ca.nds <- corresp(nds, nf=2)
biplot(ca.nds, col=c(3, 4))
```

⇒商品の群, 調査項目の群, 両方の位置関係から,
個体のポジションの読み取りを行う。



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



クラスター分析

世界各国の経済指標から、よく似た国どおしをグルーピングする。

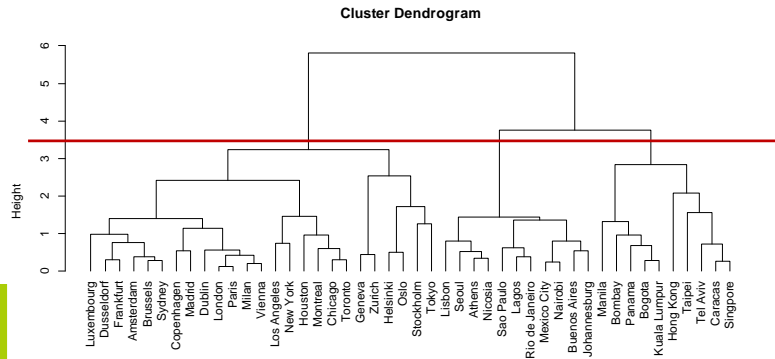
- データ "cityecon2.csv" を読み込む。
46都市 | 4変数 (都市名, 平均労働時間, 物価の指標 (Zurichを100として), 時給の指標 (Zurichを100として))
都市名 (第1変数) を行の名前にしておく。

```
cty <- read.csv("cityecon2.csv", row.names="City")
```

- 階層的クラスター分析を行う。
データは標準化し, 非類似度の計算方法とクラスター間の距離の求め方を指定し, デンドログラムを描く。

```
x=scale(cty)
hc <- hclust(dist(x, method = "euclidean"), "complete")
plot(hc, hang = -1)
```

経済指標の似た任意の数のグループに分けられる。
たとえば, 7グループに分けたい場合,
右図のように赤線のところで切ればよい。



量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



ENJOY SCIENCE!

ボクラ、科学の子。

岡山理科大学
OKAYAMA UNIVERSITY OF SCIENCE

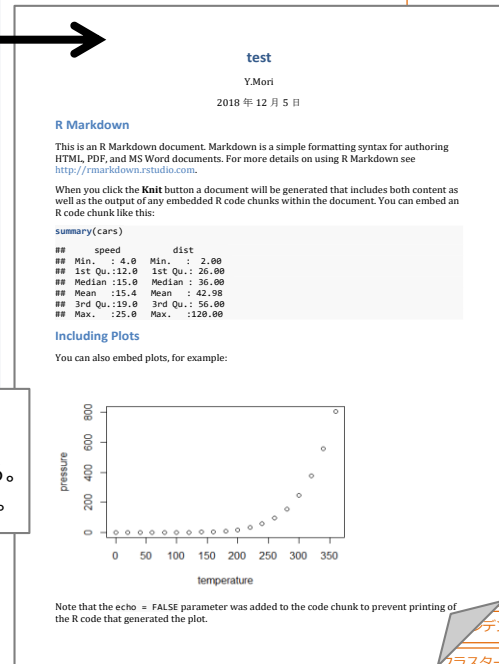
Rからのドキュメント作成

- Rstudioから
 - ・ HTML (Webページ)
 - ・ PDF, Wordなどの文書
 - ・ Beamerなどのプレゼンスライド
 などのドキュメントが生成できる。
- メリット
 - ・ R上で分析からレポート生成まで
 - ・ Office／画像処理ソフトが不要
 - ・ 内容そのままの再現, 配布が簡単など
- パッケージ"rmarkdown"をインストールする。
- 説明は
 - <https://rmarkdown.rstudio.com/>
 - https://kazutan.github.io/kazutanR/Rmd_intro.html
 - などで。

```

1 ---
2 title: "test"
3 author: "Y.Mori"
4 date: "2018年12月5日"
5 output: word_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10
11 ## R Markdown
12
13 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF,
14 and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the "Knit" button a document will be generated that includes both content as well
17 as the output of any embedded R code chunks within the document. You can embed an R code chunk
18 like this:
19
20 ```{r cars}
21 summary(cars)
22 ...
23
24 ## Including Plots
25
26 You can also embed plots, for example:
27
28 ```{r pressure, echo=FALSE}
29 plot(pressure)
30
31 Note that the 'echo = FALSE' parameter was added to the code chunk to prevent printing of the R
32 code that generated the plot.
33
    
```

↑ エディターペイン上で作る。
書式は, 上記のとおりで,
レンダリング (Knit) するとできあがる。
コードは実行される (非実行も指定可)。



量

質

操作

卓

量

ラフ)

表

数

析

析

デンス分析

クラスター分析

ドキュメント生成



ENJOY SCIENCE

ボクラ、科学の子。

関数の作成

- 自分で関数を定義することができる。

```
関数名 <- function (引数1, ..., 引数n) {
  関数本体
}
```

- 例

たとえば、縦の長さや横の長さを入れて、長方形の面積を求める関数は、

```
rec <- function(x, y){
  x * y
}
```

とする。

(コンソール上では、関数の定義が続いている間は、+ で行が繋がっていることが示される。)

縦(x)=3, 横(y)=4 の長方形の面積は、

```
rec(3, 4)
```

とすると、

```
[1] 12
```

となる。

量

質

操作

要約

記述

関連性

予測

層別

分類

比較

関数電卓

基本統計量

可視化 (グラフ)

分割表

相関係数

回帰分析

主成分分析

コレスポンデンス分析

クラスター分析

ドキュメント生成



Part1 【入門】

R言語とは

RとRStudio（基本操作とデータ処理の初歩）

Part2 【実践】

Rによるデータ分析 1（傾向の把握と可視化）

Rによるデータ分析 2（相関、回帰、予測）

Rによるデータ分析 3（多変量解析 + α ）

まとめ



まとめ

- Rは**統計分析**に向いている統計解析環境

Excelではできない**○○分析**などは、絶対 R。

データへのアクセス (あれこれ触る) には断然 R。

何千もの**パッケージ**が分析を助けてくれる。

(四則計算や簡単なグラフ作成, 扱うデータ量が小さい場合は, Excelが優位な場合も)

- Rと**RStudio**

RStudioは**統合環境**

オリジナルRの**使いにくさを解消**

Rの**中身**が見える感じ

入力**補助・補完**／**グラフ出力**などは大変便利

⇒単純なRの操作だけでも, オリジナルのRは使う必要はない。

外との連携が強化 (R MarkdownやShinyなど。こういった特徴の利用へ挑戦！)

- たくさんの**参考URL**や**文献**があり, そして**仲間**がいる！

- ただし, **手法の理解**は必要・・・ですね。



参考URL

- R
RjpWiki <http://www.okadajp.org/RWiki/?RjpWiki>
The R Project for Statistical Computing <https://www.r-project.org/>
CRAN <https://cran.r-project.org/> (CRAN Japan mirror <https://cran.ism.ac.jp/>)
- Rの使い方
<http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html>
<https://stats.biopapyrus.jp/r/#std>
- RStudio
<https://www.rstudio.com/>
- RStudioの使い方
https://kazutan.github.io/JSSP2018_spring/intro_rstudio.html
- 森の教科書
http://mo161.soci.ous.ac.jp/R/DA_R/ Rの基本的な使い方
<http://mo161.soci.ous.ac.jp/@d/indexj.html> 解析ストーリーに基づくデータ分析の学習
(Rのコードがある ⇒ <http://mo161.soci.ous.ac.jp/@d/DoLStat/indexj.html>)
- Rの書籍…たくさん！！ ビジネス分野のものも多い。対象に直結するものが入りやすいと思います。



ご清聴ありがとうございました。
RStudioでのデータ分析, お疲れさまでした。

