



高知工科大学 経済・マネジメント学群

計量経済学

7. データの収集・クリーニング

やない ゆう き
矢内 勇生



<https://yukiyanai.github.io>



yanai.yuki@kochi-tech.ac.jp



今日の目標

- 分析対象となるデータの形式について理解する
 - ▶ データの入手法
 - ▶ 理想的なデータの「かたち」：tidy data
 - ▶ データの前処理（Rによる実習）

データ

- データ分析：データが必要！
 - ▶ **どのような**データが必要？
 - ▶ **どうやって**データを集める？

長方形データ (Rectangular Data)

- 最も一般的なのは、長方形データセット
- 各行 (row) が観測単位1つを表す
 - ▶ 例：右の図では、「候補者」が1つの行
- 各列 (column) が1つの変数を表す
- 各セル (cell; 行と列の組) が値 (数値または文字列) を持つ

	A	B	C	D	E	F	G	H	
1	year	ku	kun	party	name	age	status	nocand	w
2	1996	aichi	1	1000	KAWAMURA, TAKASHI	47	2	7	
3	1996	aichi	1	800	IMAEDA, NORIO	72	3	7	
4	1996	aichi	1	1001	SATO, TAISUKE	53	2	7	
5	1996	aichi	1	305	IWANAKA, MIHOKO	43	1	7	
6	1996	aichi	1	1014	ITO, MASAKO	51	1	7	
7	1996	aichi	1	1038	YAMADA, HIROSHIB	51	1	7	
8	1996	aichi	1	1	ASANO, KOSETSU	45	1	7	
9	1996	aichi	2	1000	AOKI, HIROYUKI	51	2	8	
10	1996	aichi	2	800	TANABE, HIROO	71	3	8	
11	1996	aichi	2	1001	FURUKAWA, MOTOHISA	30	1	8	
12	1996	aichi	2	305	ISHIYAMA, JYUNICHI	31	1	8	
13	1996	aichi	2	1003	FUJIWARA, MICHIKO	44	1	8	
14	1996	aichi	2	1014	ISHIKAWA, KAZUMI	61	1	8	
15	1996	aichi	2	1	MURAMATSU, YOICHI	47	1	8	
16	1996	aichi	2	1038	YAMAZAKI, YOSHIKI	43	1	8	
17	1996	aichi	3	1000	YOSHIDA, YUKIHIRO	35	1	7	
18	1996	aichi	3	800	KATAOKA, TAKESHI	46	2	7	
19	1996	aichi	3	1001	KONDO, SHOICHI	38	1	7	
20	1996	aichi	3	305	YANAGIDA, SAEKO	50	1	7	
21	1996	aichi	3	1038	NAKANO, YOKO	54	1	7	
22	1996	aichi	3	1014	OGAWA, OSAMU	35	1	7	
23	1996	aichi	3	1	ATOJI, MASAO	43	1	7	
24	1996	aichi	4	1000	MISAWA, JUN	44	1	6	
25	1996	aichi	4	800	TSUKAMOTO, SABURO	69	3	6	
26	1996	aichi	4	305	SEKO, YUKIKO	49	1	6	
27	1996	aichi	4	1001	TAKAGI, HIROSHI	43	1	6	
28	1996	aichi	4	1038	ITO, TAKAYOSHI	61	1	6	
29	1996	aichi	4	1014	SHIOKAWA, CHIKANAO	40	1	6	

図：浅野・矢内 (2018) の衆院選データ

CSV ファイル

- CSV: Comma Separated Values（カンマ区切りのファイル）
 - ▶ テキストファイル
 - ▶ 汎用性が高い
 - LibreOffice Calc やMS Excel などの表計算ソフトで編集可能
 - すべてのデータ分析ソフト（アプリ）で開ける（計算できる）
 - ▶ すべてのデータセットをCSV形式で保存しよう！
 - 再現性の確保：他人のため、将来のため

CSVファイルの例：hr96-17.csv

- テキストエディタで開いた場合

```
1 year,ku,kun,status,name,party,party_code,previous,w1,votes,age,nocand,rank,vote,
2 1996,aichi,1,1,"KAWAMURA, TAKASHI",NFP,8,2,1,40,47,7,1,66876,346774,49.2,9828097
3 1996,aichi,1,2,"IMAEDA, NORIO",LDP,1,3,0,25.7,72,7,2,42969,346774,49.2,9311555
4 1996,aichi,1,1,"SATO, TAISUKE",DPJ,3,2,0,20.1,53,7,3,33503,346774,49.2,9231284
5 1996,aichi,1,0,"IWANAKA, MIHOKO",JCP,2,0,0,13.3,43,7,4,22209,346774,49.2,2177203
6 1996,aichi,1,0,"ITO, MASAKO",others,100,0,0,0.4,51,7,5,616,346774,49.2,.
7 1996,aichi,1,0,"YAMADA, HIROSHIB",kokuminto,22,0,0,0.3,51,7,6,566,346774,49.2,.
8 1996,aichi,1,0,"ASANO, KOSETSU",independent,99,0,0,0.2,45,7,7,312,346774,49.2,.
9 1996,aichi,2,1,"AOKI, HIROYUKI",NFP,8,1,1,32.9,51,8,1,56101,338310,51.8,12940178
10 1996,aichi,2,2,"TANABE, HIROO",LDP,1,1,0,26.4,71,8,2,44938,338310,51.8,16512426
11 1996,aichi,2,0,"FURUKAWA, MOTOHISA",DPJ,3,0,2,25.7,30,8,3,43804,338310,51.8,11435567
12 1996,aichi,2,0,"ISHIYAMA, JUNICHI",JCP,2,0,0,12.5,31,8,4,21337,338310,51.8,2128510
13 1996,aichi,2,0,"FUJIWARA, MICHIKO",jiyu-rengo,10,0,0,1.6,44,8,5,2670,338310,51.8,3270
14 1996,aichi,2,0,"ISHIKAWA, KAZUMI",others,100,0,0,0.4,61,8,6,701,338310,51.8,.
15 1996,aichi,2,0,"MURAMATSU, YOICHI",independent,99,0,0,0.2,47,8,7,418,338310,51.8,.
16 1996,aichi,2,0,"YAMAZAKI, YOSHIKI",kokuminto,22,0,0,0.2,43,8,8,348,338310,51.8,.
17 1996,aichi,3,0,"YOSHIDA, YUKIHIRO",NFP,8,1,1,32.3,35,7,1,52478,331808,50.4,11245219
18 1996,aichi,3,1,"KATAOKA, TAKESHI",LDP,1,3,0,27.46,7,2,43884,331808,50.4,5365436
19 1996,aichi,3,0,"KONDO, SHOICHI",DPJ,3,1,2,23.6,38,7,3,38351,331808,50.4,11767342
20 1996,aichi,3,0,"YANAGIDA, SAEKO",JCP,2,0,0,16.1,50,7,4,26225,331808,50.4,2110540
21 1996,aichi,3,0,"NAKANO, YOKO",kokuminto,22,0,0,0.5,54,7,5,773,331808,50.4,.
22 1996,aichi,3,0,"OGAWA, OSAMU",others,100,0,0,0.4,35,7,6,722,331808,50.4,.
23 1996,aichi,3,0,"ITO, TADAHICO",independent,99,0,0,0.2,43,7,7,246,331808,50.4,.
24 1996,aichi,4,0,"MISAWA, JUN",NFP,8,1,1,35.7,44,6,1,57361,315704,52,12134215
25 1996,aichi,4,2,"TSUKAMOTO, SABURO",LDP,1,10,0,30,69,6,2,48209,315704,52,13303388
26 1996,aichi,4,0,"SEKO, YUKIKO",JCP,2,1,2,19.3,49,6,3,30976,315704,52,3152034
27 1996,aichi,4,0,"TAKAGI, HIROSHI",DPJ,3,0,0,14.6,43,6,4,23411,315704,52,6461034
28 1996,aichi,4,0,"ITO, TAKAYOSHI",kokuminto,22,0,0,0.2,61,6,5,348,315704,52,.
29 1996,aichi,4,0,"SHIOKAWA, CHIKANAO",others,100,0,0,0.2,40,6,6,243,315704,52,.
30 1996,aichi,5,1,"AKAMATSU, HIROTAKA",DPJ,3,2,1,30.9,48,7,1,48648,319846,50.3,11894801
31 1996,aichi,5,0,"KIMURA, TAKAHIDE",LDP,1,1,2,29.5,41,7,2,46485,319846,50.3,5010946
32 1996,aichi,5,0,"BANNO, YUTAKA",NFP,8,0,0,27.3,35,7,3,43028,319846,50.3,9817826
33 1996,aichi,5,0,"NAGATOMO, TADAHICO",JCP,2,0,0,11.2,37,7,4,17670,319846,50.3,2728093
34 1996,aichi,5,0,"SUZUKI, TSUTOMU",independent,99,0,0,0.5,63,7,5,768,319846,50.3,70000
35 1996,aichi,5,0,"ICHIKAWA, MINORU",kokuminto,22,0,0,0.3,56,7,6,470,319846,50.3,.
36 1996,aichi,5,0,"UCHIYAMA, TAKAO",others,100,0,0,0.2,43,7,7,335,319846,50.3,.
37 1996,aichi,6,1,"KUSAKAWA, SHOZO",NFP,8,8,1,39.7,68,8,1,90812,433930,54.2,11252336
38 1996,aichi,6,0,"ITO, KATSUNDO",LDP,1,0,0,26.1,51,8,2,59631,433930,54.2,10987232
```


CSVファイルの例：hr96-17.csv

- 表計算ソフト（LibreOffice Calc）で開いた場合

	A	B	C	D	E	F	G	H	I	J	K	L
1	year	ku	kun	party	name	age	status	nocand	wl	rank	previous	vote
2	1996	aichi	1	1000	KAWAMURA, TAKASHI	47	2	7	1	1	2	66
3	1996	aichi	1	800	IMAEDA, NORIO	72	3	7	0	2	3	42
4	1996	aichi	1	1001	SATO, TAISUKE	53	2	7	0	3	2	33
5	1996	aichi	1	305	IWANAKA, MIHOKO	43	1	7	0	4	0	22
6	1996	aichi	1	1014	ITO, MASAKO	51	1	7	0	5	0	
7	1996	aichi	1	1038	YAMADA, HIROSHIB	51	1	7	0	6	0	
8	1996	aichi	1	1	ASANO, KOSETSU	45	1	7	0	7	0	
9	1996	aichi	2	1000	AOKI, HIROYUKI	51	2	8	1	1	2	56
10	1996	aichi	2	800	TANABE, HIROO	71	3	8	0	2	1	44
11	1996	aichi	2	1001	FURUKAWA, MOTOHISA	30	1	8	2	3	1	43
12	1996	aichi	2	305	ISHIYAMA, JYUNICHI	31	1	8	0	4	0	21
13	1996	aichi	2	1003	FUJIWARA, MICHICO	44	1	8	0	5	0	2
14	1996	aichi	2	1014	ISHIKAWA, KAZUMI	61	1	8	0	6	0	
15	1996	aichi	2	1	MURAMATSU, YOICHI	47	1	8	0	7	0	
16	1996	aichi	2	1038	YAMAZAKI, YOSHIAKI	43	1	8	0	8	0	
17	1996	aichi	3	1000	YOSHIDA, YUKIHIRO	35	1	7	1	1	1	52
18	1996	aichi	3	800	KATAOKA, TAKESHI	46	2	7	0	2	3	43
19	1996	aichi	3	1001	KONDO, SHOICHI	38	1	7	2	3	1	38

どこで手にいれる？(1) インターネット

- 長方形データがそのまま手に入る場合
 - ▶ 公的機関のウェブサイト
 - World Bank
 - OECD
 - 総務省統計局, etc.
 - ▶ 研究者や大学のウェブサイト
 - ▶ オープンデータアーカイブ
 - Harvard Dataverse
 - ICPSR
 - SSJ, etc.
- データはあるが、長方形でない or ファイルのダウンロードができない場合
 - ▶ 手で数字を入力する
 - ▶ 内容をコピー&ペーストで表計算ソフトに貼り付ける
 - ▶ OutWit Hub などのスクレイピングソフトを使う
 - ▶ R（またはPython）でウェブスクレイピングを実行する

どこで手にいれる？(2) 図書館

- CD-ROM などの（一昔前の）メディアに保存されたデータ
- オンラインデータベースへのアクセス
- 書籍に印刷されたデータ
 - ▶ 手入力
 - ▶ ドキュメントスキャナでスキャン -> OCR -> スクレイピング（RまたはPython）

どこで手にいれる？(3) 購入する

- 販売されているデータもある
- 高額なものが多い：学生が購入するのは現実的ではない
 - ▶ 図書館が購入していないか確認する
 - ▶ ないなら、図書館に購入依頼を出してみる

どこで手にいれる？(4) 作る

- 独自のデータセットを作るのも、研究の一部
 - ▶ 調査、観察、実験によってデータを集める
 - ▶ データソース（新聞やアーカイブ）を読んで情報を集める
- ★ 注意：データを集める過程も再現可能でなければならない
 - データソースを含め、すべてを記録する（秘匿すべき情報は公開前にマスクする。個人情報には慎重に扱う）
 - データを集め**始める前に**コーディングのルールを決め、**文書**にしておく

どのようなデータセットを用意すべきか？

- Rでの分析を円滑に行いたい：データを「良い」形式で用意したい
 - ▶ 回帰分析やデータの可視化に便利な方がよい
 - ▶ 1つの答え：tidy data（整然データ）
 - **tidyverse** の “tidy”

Tidy Data (整然データ)

- Hadley Wickham が提唱
- Tidy data : データの「構造」と「意味」が一致
- Tidy data ではないもの : messy data (雑然データ)
 - ▶ Tidy data を用意したい !

Tidy Data の4条件

1. 1つの列は、1つの変数を表す
2. 1つの行は、1つの観測を表す
3. 1つの表は、1つの観測単位 (unit of observation) を表す
4. 1つのセルは、1つの値を表す

3都市の天気：messy data の例

都市	6	12	18
高知	晴れ	晴れ	くもり
東京	くもり	雨	雨
大阪	雨	晴れ	晴れ

3都市の天気：tidy data の例

都市	時刻	天気
高知	6	晴れ
高知	12	晴れ
高知	18	くもり
東京	6	くもり
東京	12	雨
東京	18	雨
大阪	6	雨
大阪	12	晴れ
大阪	18	晴れ

Tidy vs. Messy Data

- どんなときも、tidy dataの方がmessy dataより優れているわけではない
 - ▶ 人間が読むには、messy dataの方がわかりやすい場合もある：天気の例
- しかし、データ分析においては、tidy dataの方が圧倒的に扱いやすいので、tidy dataを用意すべき

Messy Data の変数と列

時刻			
都市	6	12	18
都市 高知	晴れ	晴れ	くもり
東京	くもり	雨	雨
大阪	雨	晴れ	晴れ
天気			

Tidy Data の変数と列

都市	時刻	天気
高知	6	晴れ
高知	12	晴れ
高知	18	くもり
東京	6	くもり
東京	12	雨
東京	18	雨
大阪	6	雨
大阪	12	晴れ
大阪	18	晴れ

Messy Data の観測と行

都市	6	12	18
高知	晴れ	晴れ	くもり
東京	くもり	雨	雨
大阪	雨	晴れ	晴れ

1つの観測

Tidy Data の観測と行

都市	時刻	天気
高知	6	晴れ
高知	12	晴れ
高知	18	くもり
東京	6	くもり
東京	12	雨
東京	18	雨
大阪	6	雨
大阪	12	晴れ
大阪	18	晴れ

1つの観測

1つの表は、1つの観測単位

- 1つの表（行と列の組み合わせ、すなわちデータセット
[データフレーム]）
 - ▶ 例：1つ1つの観測がすべて個人
 - ▶ 例：1つ1つの観測がすべて市区町村
 - ▶ ダメな例：ある観測は国、ある観測は県、ある観測は個人

複数の観測単位がある messy data

国	大統領制？	都市	人口（100万人）
Japan	No	Tokyo	9.4
Japan	No	Osaka	2.7
Japan	No	Nagoya	2.3
USA	Yes	New York	8.5
USA	Yes	Chicago	2.7
USA	Yes	Los Angeles	3.9

観測単位：国

観測単位：都市

観測単位が1つのtidy data × 2

都市	人口（100万人）	国
Tokyo	9.4	Japan
Osaka	2.7	Japan
Nagoya	2.3	Japan
New York	8.5	USA
Chicago	2.7	USA
Los Angeles	3.9	USA

観測単位：都市

2つの表をつなぐためのキー

国	大統領制？
Japan	No
USA	Yes

観測単位：国

1つのセルが1つの値

Tidy Data

都市	時刻	天気
高知	6	晴れ
高知	12	晴れ
高知	18	くもり
東京	6	くもり
東京	12	雨
東京	18	雨
大阪	6	雨
大阪	12	晴れ
大阪	18	晴れ

Messy Data (ver. 2)

都市	時刻	天気
高知	6 & 12	晴れ
高知	18	くもり
東京	6	くもり
東京	12 & 18	雨
大阪	6	雨
大阪	12 & 18	晴れ

構造と意味の一致

- Tidy data:
 - ▶ 列：変数
 - ▶ 行：観測
 - ▶ セル：値
 - ▶ 表（データセット）：1つの観測単位に基づいて集められた情報
- データ分析：変数間の関係の**意味を調べたい**
- Rでプログラミングするときには、意味ではなく**構造に頼る**必要がある
- 構造と意味が一致：構造を利用して意味を理解することが可能

次回

仮説を立てる