RとQuartoではじめるデータサイエンス：データを可視化する《2024》

実習データについて

苅谷 千尋

27, Jun, 2024

Ⅰ. 【お願い】実習用データの再提出の検討

- 6月26日の授業の際に、実習用のデータをもってきてもらいましたが、行政機関が作成したデータ（表）が多く、整然とした（tidyな）データではないものが多かったです（授業の際に気がつけばよかったのですが、すみません）
  - このようなデータは大きな変形が必要であり、また、変形したものの、作図として使える情報量があまり多くありません
  - Rで日本の統計データを効率的に取得しようなど、便利なパッケージはありますが、それでも難しい部類だと思います。
- 6月26日に提出したデータに強い拘りがないのであれば、他のデータ（以下に、私が見つけたサイトを紹介しています）を使うことを検討して下さい
- 次回7月3日の授業内で、一人ひとり、相談に乗りたいと思います

- Google Formsを使ってアンケートをとる予定の受講生は、問題ありません

Ⅱ. tidyなデータとmessyなデータ

1. 整然とした（tidy）データの特徴

- カラム名と値からなる、シンプルな行列からなるデータ（データセット）
- ➡ カラム名（列に一つしかない、あとはデータのみ）を操作するだけなので、dplyrの過程が少なくて済みます

2. 雑然とした（messy）データの特徴

- 同じ列にカラム名が複数ある、カラムを細分化しているものは、雑然とした（messyな）データであり、変形が難しいです
  - 人間が目視することを前提とする、データ上意味のない空白が交じっているものもmessyなデータです
  - エクセルでセルを結合させているようなもの、結合させていなくてもそれに類する空白行をおいているもの
- ➡ 一列に適当な一つのカラム名を作成するプロセスが必要で、dplyrの高度な運用能力が必要です

所属機関種別（男

区分	大学	国立大学法人	公立大学	私立大学	大学共同利用機関
人数	6,985	5,399	231	993	362
年齢別					
29歳以下	1,104	859	35	148	62
30歳以上34歳以下	2,544	1,942	72	393	137
35歳以上39歳以下	1,373	1,055	41	212	65
40歳以上44歳以下	654	509	23	82	40
45歳以上49歳以下	353	263	15	51	24
50歳以上	890	768	30	58	34
不明	67	3	15	49	0

messyデータ（所属機関種別所属者数）

0	1	2	3	4	5	6
_id	日付	曜日	土日祝	小児科	内科	計
1	2018-04-09	月	0	8	6	14
2	2018-04-10	火	0	13	6	19
3	2018-04-11	水	0	17	4	21
4	2018-04-12	木	0	10	9	19
5	2018-04-13	金	0	5	4	9
6	2018-04-14	土	1	14	10	24

tidyデータ（以下の「金沢広域急病センター利用者数」）

3. 注記

- 行政が作成した表（データ）であっても、以下のような表は空白箇所が少なく、dplyrによる加工はそれほど必要ありません
- 総務省：ふるさと納税

令和4年度課税におけるふる
※例年実施している「市町
※新型コロナウイルス感染
(1)「都道府県等に対する寄
「ふるさと納税ワンストッ
「左のうち、申告特例控
(2)「共同募金会、日本赤十
(3)「条例で定めるものに対
(4)「左の3つのうちいずれ
(5)「ふるさと納税に係る寄

都道府県	市区町村	Ⅱ.ふるさと納税ワンストップ特例制度＞	ふるさと納税に係る寄附金税額控除					
			市町村民税			道府県民税		
		. 令和4年度分の課税において申告特例控除が適用されることとなった者（ふるさと納税ワンストップ特例の利用者）に係る申告特例通知書の総件数を記入してください。	人数（人）	寄附金額（円）	控除額（円）※推計値含む	人数（人）	寄附金額（円）	控除額（※推計
北海道	札幌市	192,375	122,918	11,285,467,192	6,638,791,512	122,523	11,223,365,092	1,65
北海道	函館市	12,966	7,580	708,952,300	311,353,672	7,580	708,952,300	20
北海道	小樽市	4,602	3,140	269,437,678	112,522,393	3,140	269,437,678	7
北海道	旭川市	20,212	12,703	1,127,565,409	497,146,043	12,703	1,127,565,409	33
北海道	室蘭市	4,830	2,591	211,429,100	99,060,008	2,591	211,429,100	6
北海道	釧路市	8,530	5,084	430,734,620	195,438,898	5,085	430,735,620	13
北海道	帯広市	11,608	7,684	746,704,400	323,280,221	7,683	746,668,400	21

加工しやすい行政作成の表

III. 作図しやすい情報量の多いデータ

- 複数の意味をもつ、データがそろっている
  - 例：日付; 地域; 年齢; 性別; 人数; 金額
    - 折れ線グラフ： 日付 × 人数；日付 × 金額
    - 散布図：人数 × 金額
    - 棒グラフ：人数；金額

IV. データの入手先（お勧め）

1. ckan

(1) ckan: 金沢市

- tidyなデータ例：
  - 金沢市 投票所別の投票結果
  - 金沢市 違反ごみ件数
  - 金沢広域急病センター利用者数

Note
次回から「金沢広域急病センター利用者数」を例に実演、説明します

Tip
金沢市は山野之義前市長がソフトバンク出身ということもあり、データ活用への意欲、tidyデータへの感覚があるのだと思います。ckanへのデータの登録は、石川県全体（あるいは全国的？）の取り組みのようですが、分析上、意味のあるデータを登録している自治体は少ないです。金沢市は例外的です

Tip
東京都のckanも充実しています。特にcsvファイルで提供されているデータは、tidyデータで、情報量が多く、よいものが多いです

2. 観測・実験装置

- 野生鳥獣の放射線モニタリング調査結果

  福島県では、東京電力福島第一原子力発電所の事故後、原子力災害本部長から福島県知事に対し野生鳥獣（イノシシ、ツキノワグマ、キジ、ヤマドリ、カルガモ、ノウサギ）の肉の摂取及び出荷制限の指示があり、県内で捕獲された野生鳥獣の体内における放射性核種濃度測定調査の結果についてお知らせをしています。

- 生態学的研究の上で便利なオープンデータ集

  - 生態学にかかわるオープンデータのリンクがまとめられています。アカウント登録が必要なサイトがほとんどですが、よいデータが集まっているような印象を受けました。研究分野が近い受講生はアカウント登録をしてデータを入手する価値があるのではないかと思います。

V. 提出先

- 演習：Google Forms

  Warning
  再提出する場合は「授業回を選択してください」を「6月26日」として提出して下さい