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ことを検討したいる受講生は、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」）

III. 作図しやすい情報量の多いデータ

- 複数の意味をもち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。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、意味のあるデータのほとんどは金沢市のものです

Tip
東京都のckanも充実しています。特にcsvファイルはtidyデータで、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

V. 提出先

- 演習：Google Forms