分析大会用データについて

分析大会用データ

新型コロナウィルスに関するデータです。すべてオンラインで最新データが取得できます。

データ	区分	種別	形式	言語	DL	備考
厚生労働省オープンデータ	公開	集計	CSV	日本	可	集計データを個別ファイルで公開
Covid19 Japan	公開	個票・集計	JSON	英語	可	GitHubにて
JAG Japan	公開	個票	CSV	日本	可	GIS処理用データ付き

データ概略

厚生労働省オープンデータ（公開／集計／公式）

日本の公式データ。国内事例（チャーター便、空港検疫などを除く）の各報告日時点の集計値。

日付 <chr>	PCR 検査陽性者数(単日) <dbl>
2020/1/16	1
2020/1/17	0
2020/1/18	0
2020/1/19	0
2020/1/20	0
2020/1/21	0
2020/1/22	0
2020/1/23	0
2020/1/24	1
2020/1/25	1

1-10 of 297 rows

Previous 1 2 3 4 5 6 … 30 Next

厚生労働省オープンデータの注意点

厚生労働省のデータはファイルにより単日であったり集計値であったりしますので、サイトの注意書きをよく読んでください。

データ	特記
陽性者数	
PCR検査実施人数	当日と前日の累積人数の差を当日の実施人数として計上
入院治療等を要する者の数	
退院又は治療解除となった者の数	
死亡者数	
PCR検査の実施件数	暫定値であり後日変更される可能性あり

Covid19 Japan（公開／個票・集計／非公式）

Exploratory EDA Salonなどで紹介されている有志によるJSON形式データ。個票データ（下表）と集計データに分かれています。全て英語。

	patientId <chr>	dateAnnounced <chr>	ageBracket <int>	gender <chr>
1	15	2020-01-15	30	M
2	TOK1	2020-01-24	40	M
3	TOK2	2020-01-25	30	F
4	18	2020-01-26	40	M
5	19	2020-01-28	40	M
6	20	2020-01-28	60	M
7	HKD1	2020-01-28	40	F
8	OSK1	2020-01-29	40	F
9	1	2020-01-30	50	M
10	23	2020-01-30	50	M

1-10 of 10,000 rows | 1-5 of 24 columns Previous 1 2 3 4 5 6 … 1000 Next

Covid19 Japan データの注意点

GitHub からjsonliteパッケージを利用して読み込んでください。

```
library(jsonlite)
path <- "https://raw.githubusercontent.com/reustle/covid19japan-data/master/"
path <- paste0(path, "docs/patient_data/")

path %>%
  paste0("latest.json") %>%
  readr::read_lines() %>%
  paste0(path, .) %>%
  jsonlite::fromJSON()
```

個票データへのパスは表示の都合上、分割しています。
各列（変量、フィーチャー）の定義はこちら。

JAG Japan （公開／個票／非公式）

ジャッグジャパンがArcGISプロモーションマップ公開のために作成しているデータを副次的にCSV形式で公開している個票データ。

通し <dbl>	厚労省NO <chr>	無症状病原体保有者 <chr>	国内 <chr>	チャーター便 <chr>	年代 <chr>	性別 <chr>	確定日 <chr>
1	1	*NA*	A-1	*NA*	30	男性	1/15/2020
2	2	*NA*	A-2	*NA*	40	男性	1/24/2020
3	3	*NA*	A-3	*NA*	30	女性	1/25/2020
4	4	*NA*	A-4	*NA*	40	男性	1/26/2020
5	5	*NA*	A-5	*NA*	40	男性	1/28/2020
6	6	*NA*	A-6	*NA*	60	男性	1/28/2020
7	7	*NA*	A-7	*NA*	40	女性	1/28/2020
8	8	*NA*	A-8	*NA*	40	女性	1/29/2020
9	9	*NA*	*NA*	B-1	50	男性	1/30/2020
10	-	チャーター無症状2	*NA*	*NA*	50	女性	1/30/2020

1-10 of 10,000 rows | 1-8 of 54 columns　Previous 1 2 3 4 5 6 … 1000 Next

JAG Japan データの注意点

副次的な公開データなので色々な事情がある模様。特徴的なのはW列（23列）目以降にGIS処理用の変量（フィーチャー）が用意されている点です。これらの変量は分析には必要ありません。

Windows環境ではエラー回避のために下記の guess_maxオプションを指定してください。なお、指定してもGIS関連データの部分でワーニングが出ます。

```
  readr::read_csv(locale = readr::locale(encoding = "UTF-8"), guess_max = 5000)
```

各列（変量、フィーチャー）の定義は こちら。

データを食材に例えると

集計データ

安心の調理済み食材

調理済みなのでアレンジする余地があまりなく、いかに美味しそうに盛り付けるかがポイント。

個票データ

バイヤー厳選食材セット

バイヤーによって食材の産地や種類、収穫方法や品質が異なり、中には調理が厄介な食材が含まれることも。ただ、食材を追加したり調理方法を選ぶことができるので腕を振るえる。

データを扱う上でポイント

- `tidyverse`パッケージを必ずインストール
  - `readr`ならびに`jsonlite`パッケージは`tidyverse`パッケージに含まれます
- CSVの読み込みには`readr::read_csv`関数で
  - ファイルにURLを指定すれば読み込むことができます
  - 文字化けする場合は`locale`オプションを指定してください
  - Warningなどが表示された場合は必ず読んで、確認してください
- 読み込んだデータは各列（変量）のデータ型を必ず確認
  - 特に文字（chr）型になっている変量には注意してください
- 本資料のコードがGoogle Colabで動作することは確認済

分析を行う際の注意点

- 集計データを扱う場合は、集計条件をよく確認
  - 思い込みで扱うと思わぬ落とし穴があります
- 個票データを扱う場合は、各列（変量）の持つ意味をよく確認
  - 個票データは非公式のデータなので、作成者により表記等が変わります
- データがよく分からない場合はCSVファイルに書き出して眺める
  - `readr::write_excel_csv(df, filepath)` で書き出せます
- 都道府県の地方区分などのデータは こちらで公開中
  - Wikipediaと総務省統計局の情報を元に作成してあります
  - 推定人口はH30年時点のもので単位は千人

Enjoy!