分析用データについて

分析対象データ

今回利用する新型コロナウィルスに関するデータは以下から任意に選択してください。すべてオンラインにてデータが取得できます。

データ名	区分	種別	ダウンロード	備考
厚生労働省オープンデータ	公開	集計	可	項目単位で集計したものを個別ファイルで公開
JAG Japan	公開	個別	可	GIS処理用データ付き
Covid19japan.com	公開	個別	可	GitHubにてJSON形式で公開

その他、任意のデータを用いても構いませんがデータの出典と説明を明記してください。

データ概略

厚生労働省オープンデータ（集計／公開・公式）

河野行政改革担当大臣またはデジタル庁（仮称）に期待。

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 264 rows

Previous 1 2 3 4 5 6 … 27 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各報告日時点の集計値が個別ファイルになっていますので、意味を把握してから分析してください。

データ名	概要
陽性者数	新規に陽性と判断された者の数（除く空港検疫）
PCR検査実施人数	当日と前日の累積人数の差（除く空港検疫）
入院治療等を要する者の数	入院待機中・確認中を除く（除く空港検疫）
退院又は治療解除となった者の数	（除く空港検疫）
死亡者数	（除く空港検疫）
PCR検査の実施件数	暫定数値であり後日変更される可能性あり

JAG Japan （個別／公開・非公式）

日本国内の各サイトから収集・整理して公開しているデータ。

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

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ので、削除しておくことをおすゝめします。また、インスタンスには多数の揺れが含まれている点に注意してください。

なお、読み込み時は以下のオプションを指定しないとWinodws環境ではエラーになります。

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

各列（変量）の定義は こちら で公開されています。

Covid19japan.com（個別／公開・非公式）

Exploratory EDA Salonにあるデータのオリジナル。JSON形式で公開されている。

	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

Covid19japan.com データの注意点

GitHub からjsonliteパッケージを利用して以下のコードで読み込んでください。readr::read_csv関数では正しく読み込めません。

```
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()
```

pathの2行は表示用に分割しています。また、JAG Japanのデータと同様に揺れがあります。

データを扱う上でのポイントなど

- `tidyverse`パッケージを必ずインストールしておいてください
  - `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`関数を用います
  - 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してください
  - 特に文字（chr）型になっている変量には注意してください
- 都道府県の地方区分を使う場合は こちら を使ってください
- 提示コードはGoogle Colabでも動作確認済

おまけ

八地方区分

地方区分	含まれる都道府県
北海道	北海道
東北	青森県・岩手県・秋田県・宮城県・山形県・福島県
関東	茨城県・栃木県・群馬県・埼玉県・千葉県・東京都・神奈川県
中部	山梨県・長野県・新潟県・富山県・石川県・福井県・静岡県・愛知県・岐阜県
近畿	三重県・滋賀県・京都府・大阪府・兵庫県・奈良県・和歌山県
中国	鳥取県・島根県・岡山県・広島県・山口県
四国	香川県・愛媛県・徳島県・高知県
九州	福岡県・佐賀県・長崎県・熊本県・大分県・宮崎県・鹿児島県・沖縄県

地方区分を使った集計事例

一時期、かなり騒がれた「クラスタ」で感染し陽性と判断された割合がどの程度なのかを地方毎にクロス集計してみました。中国・四国地方を除くと結果的にはクラスタ感染を押さえられたと言えそうです。

地方 <fctr>	非クラスタ感染者 <int>	クラスタ感染者 <int>	クラスタ比率[%] <dbl>
北海道地方	2132	190	8.2
東北地方	884	45	4.8
関東地方	44866	654	1.4
中部地方	8682	304	3.4
近畿地方	17146	363	2.1
中国地方	985	172	14.9
四国地方	450	66	12.8
九州地方	9485	497	5.0
NA	1002	21	2.1

9 rows

Enjoy!