分析大会について

分析大会概要

新型コロナウィルスに関する二種類のデータ(集計データ、個票データ)を用いて今年度の勉強会で学んできたデータラングリングと可視化を行います。分析に利用する環境は限定しません。

集計データによる分析

まずは厚生労働省オープンデータを用いて、日頃、ニュースなどで目にするグラフを作成してみます。

個票データによる分析

Covid19 Japanのデータを用いて、集計データでは分からない傾向などを見るためのグラフを作成してみます。どのような傾向を見るかは後述のチーム単位で決めます。

実施概要

分析大会は1月回と2月回の2回に渡り実施しますので、基本的に両回への参加をお願いします。

日程

1. 集計データを用いた可視化（1月回の午前）
2. 個票データを用いた分析（1月回の午後ならびに2月回の午前）
3. 発表会（2月回の午後）

チーム単位での実施

分からないことや方針などを相談しながら進められるようにするために数人単位のチーム分けを行います。各チームには「ファシリテーター（支援者）」としてのリーダを配置しますが、意思決定は基本的にメンバー間で行ってください。なお、勉強会の時間外で任意に打ち合わせなどを行うことは自由です。

分析大会用データ

新型コロナウィルスに関するデータは以下の二つを用います。どちらもオンラインで直接取得することが可能です。

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

取得タイミングによってはデータがそろっていない場合もありますので、適宜、フィルタリングしてください。なお、Covid19 Japanには集計データもありますが、少し厄介な形式なのでチャレンジしてみたい方のみとします。

データの概略

厚生労働省オープンデータ(公開／集計／公式)

日本国の公式データ。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。基本的に前日までのデータとなります。

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 329 rows

Previous 1 2 3 4 5 6 … 33 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
重症者数	

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

2020/12/10からファイル構成が変更されています。読み込むには`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。データはCovid19 Japanのサイトでなく[GitHub](https://github.com/reustle/covid19japan-data)で公開されています。

```
library(tidyverse)
library(jsonlite)
"https://raw.githubusercontent.com/reustle/covid19japan-data/master/" %>%
  paste0("docs/patient_data/latest.json") %>%
  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](https://github.com/reustle/covid19japan-data/blob/master/docs/patient_data/README.md)。

各データの特徴と注意点

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

基本的には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。ただし、バイヤーによっては下ごしらえのみの場合もありアレンジに余地があるものも。

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・時期や品質が異なり、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
- 本資料でのコードはGoogle Colabで動作することを確認済

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
  - データフレームは`readr::write_excel_csv(df, filepath)`で書き出し
- 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ので必要に応じて加工してください(推定人口はH30年時点のもので単位は千人)
- 個票を集計しただけでは因果関係は分からない

Enjoy!