Final project dataset report

Danyu Zhang, Limingrui Wan and Daniel Alonso

November 16th, 2020

The COVID-19 dataset

Coronavirus disease 2019 (COVID-19) is a contagious respiratory and vascular disease caused by severe acute respiratory syndrome coronavirus 2. COVID-19 mainly spreads through the air when people are near each other long enough, primarily via small droplets or aerosols, as an infected person breathes, coughs, sneezes, sings, or speaks. There are currently no proven vaccines or specific treatments for COVID-19, though several are in development.

We have chosen this dataset as it not just fits the criteria but it allows us to also include interesting demographic information about each country like life expectancy, median age, human development index, among other things. These variables could provide quite interesting insight in the context of the COVID-19 pandemic.

Variables

The dataset contains the following variables:

- continent: Continent of the geographical location
- location: Geographical location, country
- total_cases: Total confirmed cases of COVID-19 at the location
- new_cases: New confirmed cases of COVID-19 at the location
- new_cases_smoothed: New confirmed cases of COVID-19 (7-day smoothed), the average of new cases during 7 days
- total_deaths: Total deaths attributed to COVID-19 of the location
- new_deaths: New deaths attributed to COVID-19 of the region
- new_deaths_smoothed: New deaths attributed to COVID-19 (7-day smoothed), the average of new deaths during 7 days
- total_cases_per_million: Total confirmed cases of COVID-19 per 1,000,000 people of the location
- new_cases_per_million: New confirmed cases of COVID-19 per 1,000,000 people of the location
- new_cases_smoothed_per_million: New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people, the average of new cases per million during 7 days
- total_deaths_per_million: Total deaths attributed to COVID-19 per 1,000,000 people of the location
- new_deaths_per_million: New deaths attributed to COVID-19 per 1,000,000 people
- stringency_index: Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)
- population: Population of the location in 2020

- population_density: Number of people divided by land area, measured in square kilometers
- median_age: Median age of the population of the location, UN projection for 2020

- aged_65_older: Share of the population that is 65 years and older in 2015 of the location
- aged_70_older: Share of the population that is 70 years and older in 2015 of the location

- gdp_per_capita: Gross domestic product at purchasing power parity of the location (constant 2011 international dollars), most recent year available

- extreme_poverty: Share of the population living in extreme poverty of the location
- cardiovasc_death_rate: Death rate from cardiovascular disease in 2017 of each location (annual number of deaths per 100,000 people)

- diabetes_prevalence: Diabetes prevalence of each location (% of population aged 20 to 79) in 2017
- hospital_beds_per_thousand: Hospital beds per 1,000 people of the location, most recent year available since 2010
- life_expectancy: Life expectancy at birth in 2019 of each location
- human_development_index: Summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living

Simple summary statistics

```
library(dplyr)
df <- read.csv('./data/data.csv')
```

```
cols = colnames(df)[colnames(df) != 'location' &
                    colnames(df) != 'continent' &
                    colnames(df) != 'development']
sapply(df %>% dplyr::select(cols), quantile, na.rm = TRUE)
```

```
##         X total_cases new_cases new_cases_smoothed total_deaths new_deaths
## 0%      0           1         0              0.000          1.0          0
## 25%    45        4866         2              8.286         87.5          0
## 50%    90       21793       132            188.000        497.0          1
## 75%   135      114270      1080           1361.107       2255.5         19
## 100%  180     9291245     83883          83817.286     231551.0        555
##       new_deaths_smoothed total_cases_per_million new_cases_per_million
## 0%                0.00000                   3.299                 0.000
## 25%               0.10725                 617.926                 0.349
## 50%               2.07150                4037.204                13.017
## 75%              17.03575               12099.998               125.104
## 100%            830.85700               63262.797              2523.739
##       new_cases_smoothed_per_million total_deaths_per_million
## 0%                           0.00000                   0.0840
## 25%                          1.47525                  16.3485
## 50%                         20.65150                  64.7930
## 75%                        137.88025                 197.7290
## 100%                      1152.20100                1053.9610
##       new_deaths_per_million stringency_index population population_density
## 0%                     0.000            5.56      38137             1.9800
## 25%                    0.000           43.52    2722291            36.0660
## 50%                    0.078           56.48    9660350            82.6000
## 75%                    1.542           66.67   31255435           206.7125
## 100%                  21.010           87.04 1439323774          7915.7310
##       median_age aged_65_older aged_70_older gdp_per_capita extreme_poverty
## 0%         15.10        1.1440        0.5260        661.240             0.1
## 25%        21.65        3.4215        2.0330       3823.194             0.6
## 50%        29.50        6.2240        3.5190      11840.846             2.2
## 75%        38.70       14.0530        8.6605      26382.287            22.5
## 100%       48.20       27.0490       18.4930     116935.600            77.6
```

```
##      cardiovasc_death_rate diabetes_prevalence hospital_beds_per_thousand
## 0%                 79.3700              0.9900                      0.1000
## 25%               170.6675              5.1525                      1.3000
## 50%               243.9640              7.1100                      2.3200
## 75%               329.7885             10.0800                      3.8505
## 100%              724.4170             30.5300                     13.0500
##      life_expectancy human_development_index
## 0%             53.28                   0.354
## 25%            67.13                   0.588
## 50%            74.25                   0.741
## 75%            77.91                   0.825
## 100%           84.63                   0.953
```