COVID-19 by Our World in Data

untitled

25/02/2021

Contents

Introduction

Our World in Data COVID-19 dataset contains up-to-date data on confirmed cases, deaths, hospitalizations, testing, and vaccinations as well as other variables of potential interest. The data set was derived from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University and can be found here: https://github.com/CSSEGISandData/COVID-19. Analyzing this data set will give us some incredible insights into how the covid-19 virus has affected several countries across the globe. We will explore such insights through exploratory data analysis right after we clean and preprocess the data.

Data Structure

In this data set we have a mix of character and numeric variables. It has 59 variables and 70,645 obervations.

```
## ’data.frame’:    70851 obs. of  59 variables:
##  $ iso_code                           : chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ continent                          : chr  "Asia" "Asia" "Asia" "Asia" ...
##  $ location                           : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan
##  $ date                               : chr  "2020-02-24" "2020-02-25" "2020-02-26" "2020-02-27" .
##  $ total_cases                        : num  1 1 1 1 1 1 1 1 2 4 ...
##  $ new_cases                          : num  1 0 0 0 0 0 0 0 1 2 ...
##  $ new_cases_smoothed                 : num  NA NA NA NA NA 0.143 0.143 0 0.143 0.429 ...
##  $ total_deaths                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_deaths                         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_deaths_smoothed                : num  NA NA NA NA NA 0 0 0 0 0 ...
##  $ total_cases_per_million            : num  0.026 0.026 0.026 0.026 0.026 0.026 0.026 0.026 0.051
##  $ new_cases_per_million              : num  0.026 0 0 0 0 0 0 0 0.026 0.051 ...
##  $ new_cases_smoothed_per_million     : num  NA NA NA NA NA 0.004 0.004 0 0.004 0.011 ...
##  $ total_deaths_per_million           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_deaths_per_million             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_deaths_smoothed_per_million    : num  NA NA NA NA NA 0 0 0 0 0 ...
##  $ reproduction_rate                  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ icu_patients                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ icu_patients_per_million           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ hosp_patients                      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ hosp_patients_per_million          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ weekly_icu_admissions              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ weekly_icu_admissions_per_million  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ weekly_hosp_admissions             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ weekly_hosp_admissions_per_million : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_tests                          : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_tests                        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_tests_per_thousand           : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_tests_per_thousand             : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_tests_smoothed                 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_tests_smoothed_per_thousand    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ positive_rate                      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ tests_per_case                     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ tests_units                        : chr  "" "" "" "" ...
##  $ total_vaccinations                 : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ people_vaccinated                  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ people_fully_vaccinated            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_vaccinations                     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_vaccinations_smoothed            : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ total_vaccinations_per_hundred       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ people_vaccinated_per_hundred        : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ people_fully_vaccinated_per_hundred  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_vaccinations_smoothed_per_million: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ stringency_index                     : num  8.33 8.33 8.33 8.33 8.33 ...
##  $ population                           : num  38928341 38928341 38928341 38928341 38928341 ...
##  $ population_density                   : num  54.4 54.4 54.4 54.4 54.4 ...
##  $ median_age                           : num  18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 18.6 ...
##  $ aged_65_older                        : num  2.58 2.58 2.58 2.58 2.58 ...
##  $ aged_70_older                        : num  1.34 1.34 1.34 1.34 1.34 ...
##  $ gdp_per_capita                       : num  1804 1804 1804 1804 1804 ...
##  $ extreme_poverty                      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ cardiovasc_death_rate                : num  597 597 597 597 597 ...
##  $ diabetes_prevalence                  : num  9.59 9.59 9.59 9.59 9.59 9.59 9.59 9.59 9.59 9.59 ...
##  $ female_smokers                       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ male_smokers                         : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ handwashing_facilities               : num  37.7 37.7 37.7 37.7 37.7 ...
##  $ hospital_beds_per_thousand           : num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
##  $ life_expectancy                      : num  64.8 64.8 64.8 64.8 64.8 ...
##  $ human_development_index              : num  0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511 0.511
```

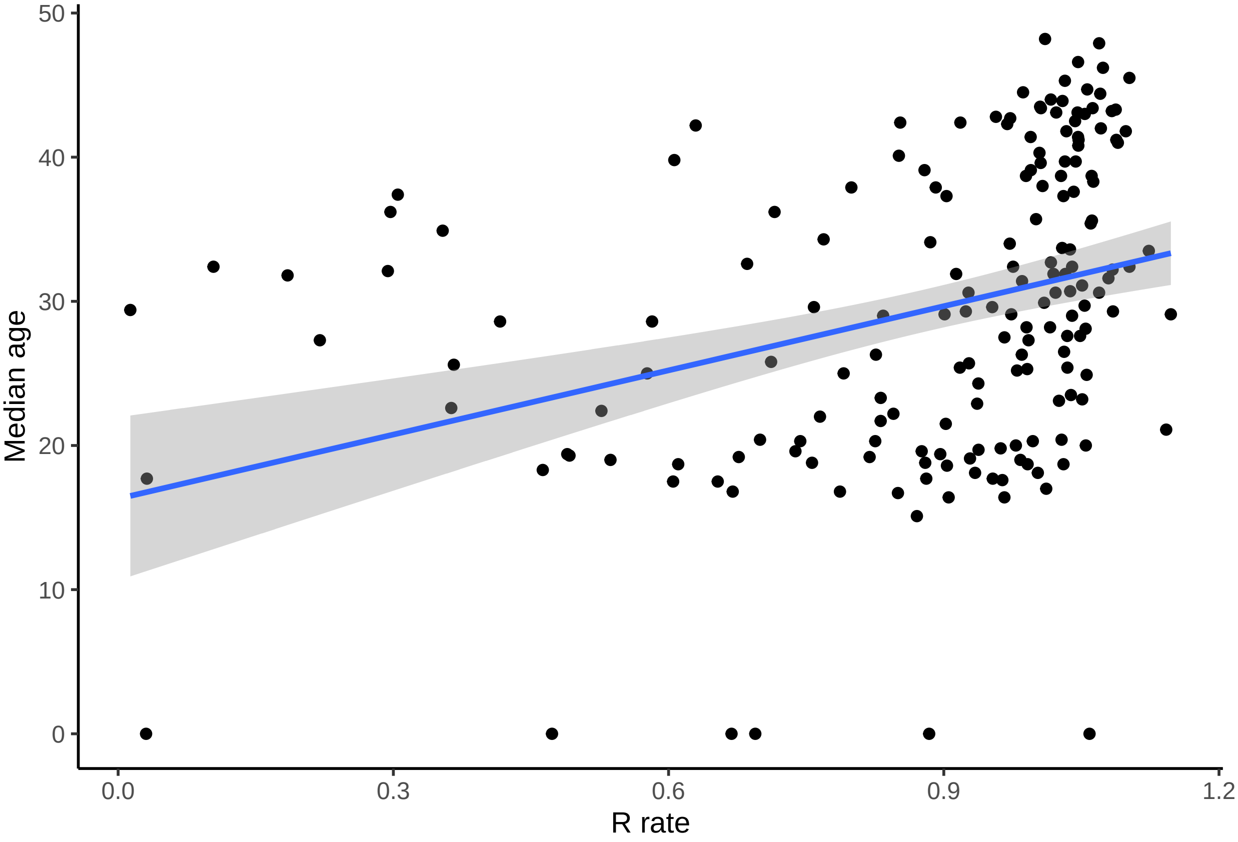
Data Preprocessing

Tests units is a character variable in our dataset which shows if samples were tested and if other processes were followed. This variable is not of interest in the exploratory data analysis I am about to undertake, therefore it would be removed. Additionally, there are 1,685,771 missing values which are all numeric, these will be replaced with 0.

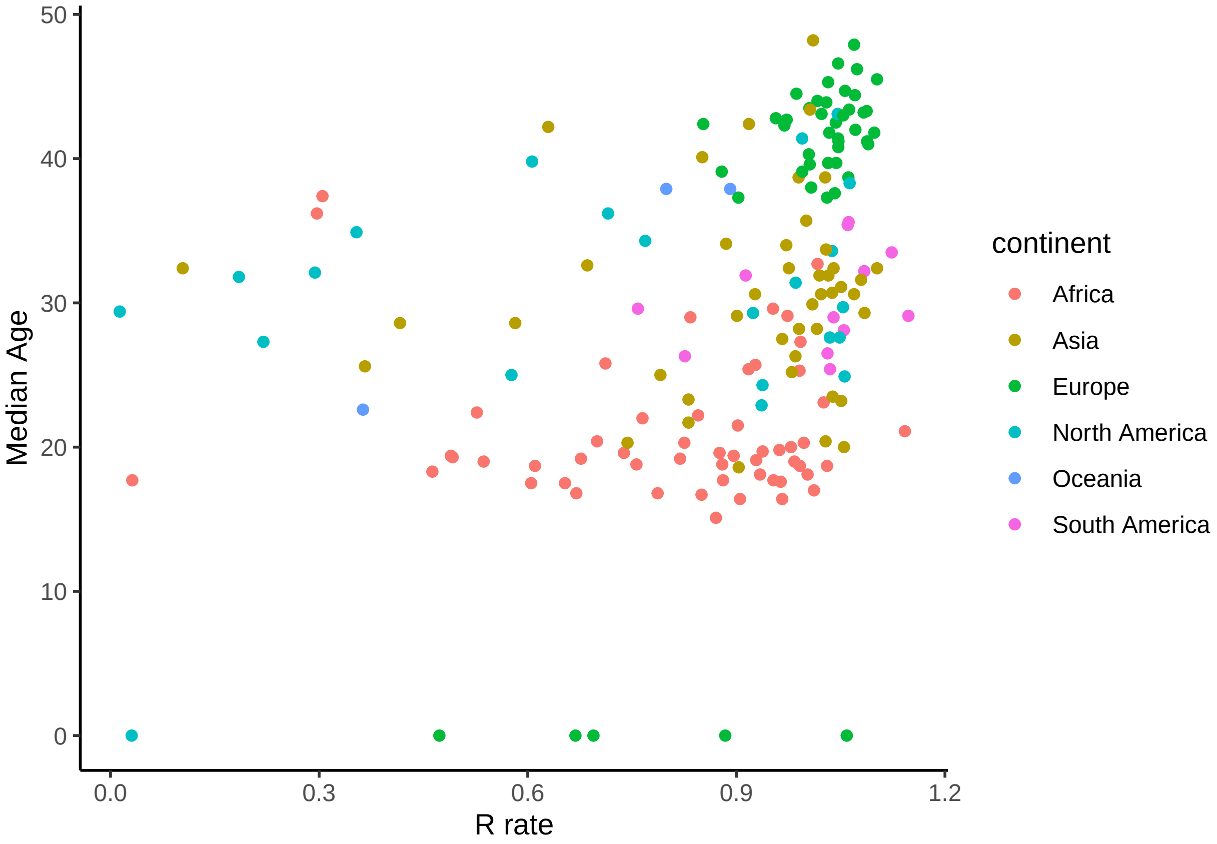
Exploratory Data Analysis

A. Relationship between Population Median Age and R-rate

The R rate (reproduction rate) is a way of rating coronavirus's ability to spread. R is the number of people that one infected person will pass on a virus to, on average. So if a countries R rate is 2 it means, on average, one person will spread the virus to 2 others. For this analysis, I would like to find out if youthful populations spread the virus at a higher rate. Recent reporting in advanced countries like the USA seems to suggest that youthful elements in a society spread the virus rapidly due to exuberance and opposition to restrictions. Thus, we will investigate if countries with a lower median age have a higher R rate. We would begin by exploring the relationship between these two variables by plotting the median age and reproduction rate of all countries. We will also utilize the geom_smooth function to aid us to detect the relationship within the data.



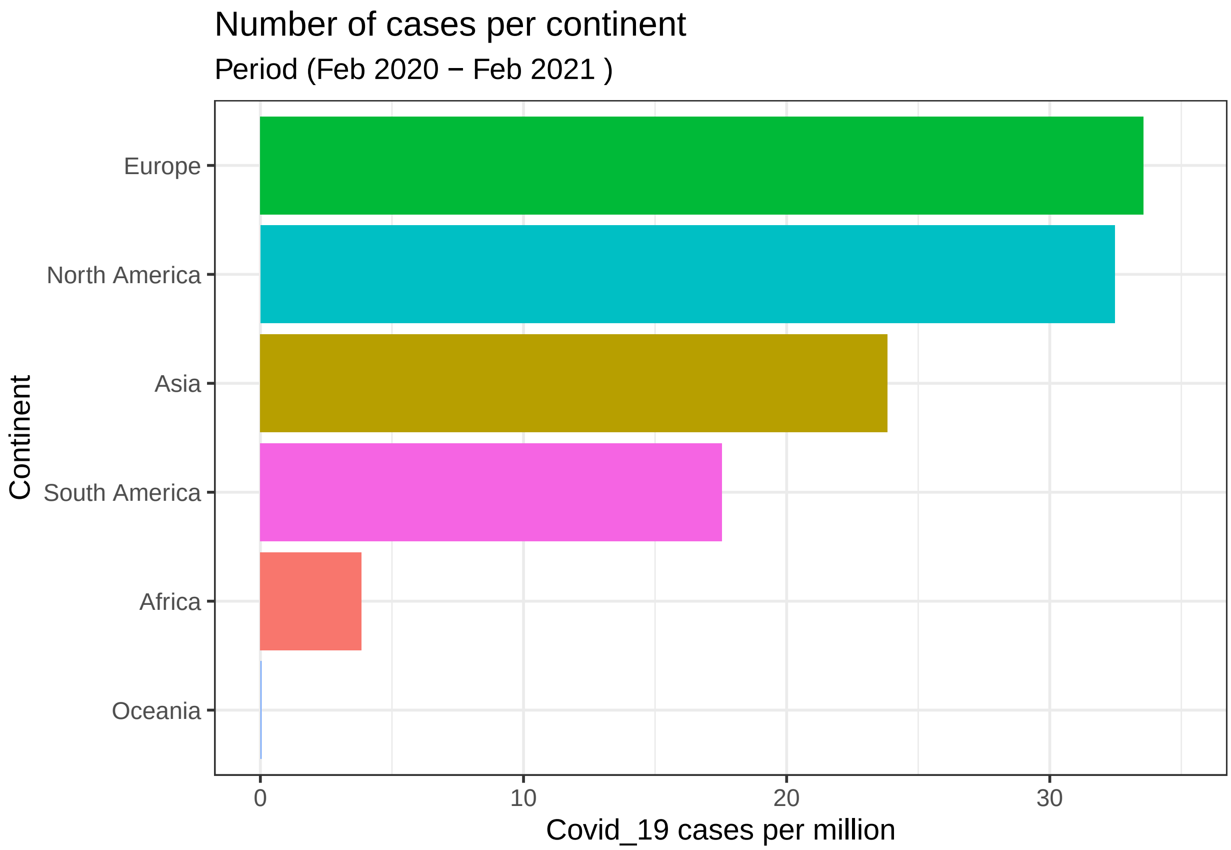
From the diagram above we can see a positive relationship between the median age and the R-rate. This implies that countries with a more aged population tend to spread the virus at a higher rate, holding all other variables constant. However, we can still unearth more insights about our data and get an intrinsic understanding of things. Lets's highlight the continent these data points can be found.



From the figure above, we can immediately point out some interesting clusters. Most European countries tend to have a higher median age in the range of 40 to 50 as well as a high R rate concentrated around 1. The African continent on the other hand has countries with the lowest median age around 20 years with an R-rate spread out from 0.4 to 1.1. This does not necessarily imply that the older median age of European countries solely accounts for the high r-rate as there are other variables at play. In the same light, I can not conclude that a more youthful population gives rise to more COVID spread.

B. Cases per Continent

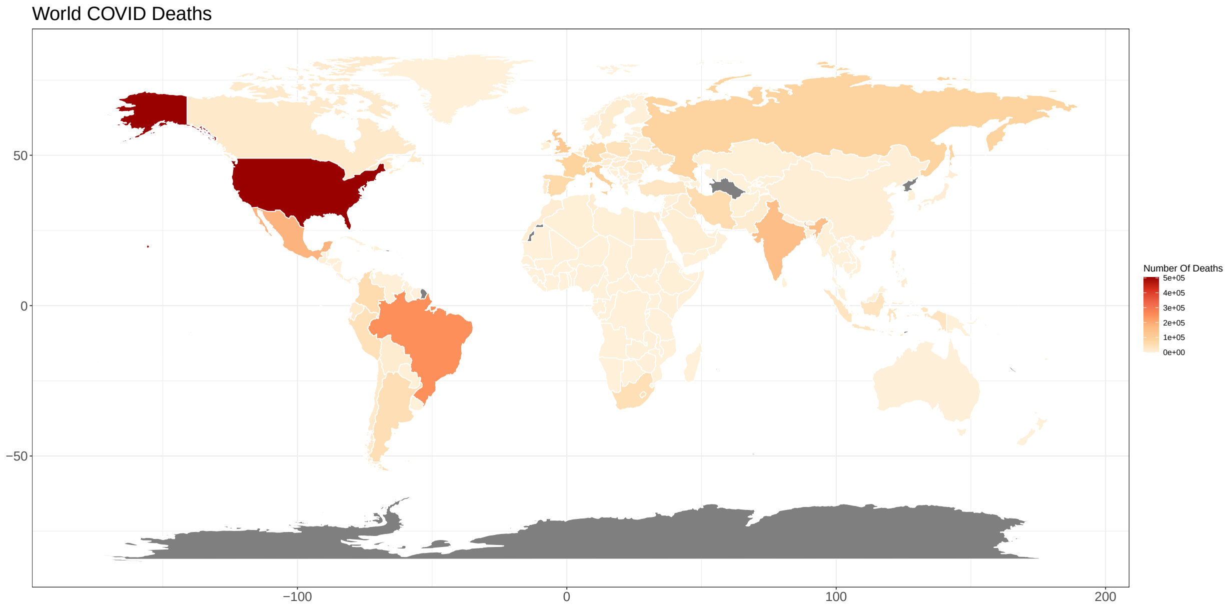
we will now explore the total numer of cases per continent. Previously we visualized the varying R-rates of counties in differnt continents. Lets see if continents with high R-rates also leads to high cases as that would intuitively be the case.



As expected we can see that the continents that contain countries with high R rates take the lead in COVID cases. Europe and North America have some of the highest cases of about thirty million. Africa on the other hand has around four million cases in total, with Oceania, which consists of Australia and other islands scattered throughout most of the Pacific Ocean, having barely any cases compared to the rest of the continents.

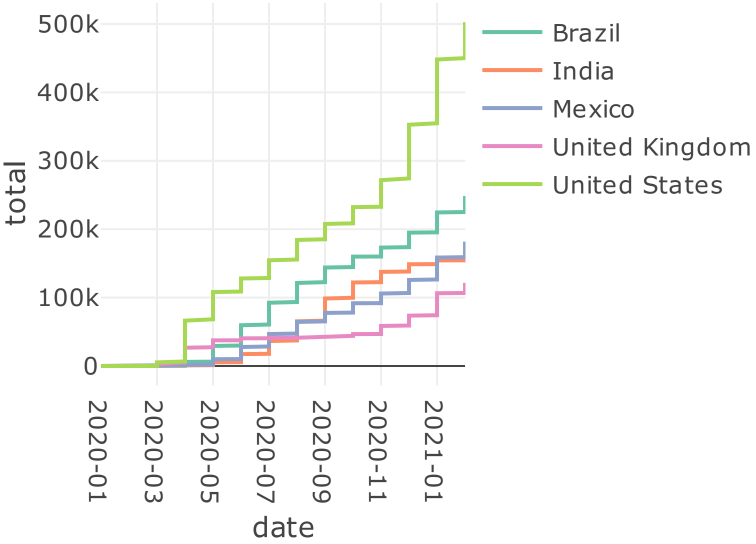
C. World Map of Coronavirus Death

A dark but necessary topic we must address is that of death. The coronavirus pandemic has cost us millions of lives around the world. To truly visualize this we will build a map showing the total number of deaths per country. It is important to note that North Korea and Turkmenistan have not released any data on covid cases and death. Data on French Guiana was also not provided in our data set.



D. Time Series of Countries with the Most Deaths

From the previous plot, we saw the number of COVID deaths per country. Now we will analyze the top 5 countries with the highest deaths. We will do this by performing a time series analysis to understand the rise in deaths since the inception of this pandemic to February 2021.


500k
400k
300k
200k
100k
total
date
Brazil
India
Mexico
United Kingdom
United States