2019 - 2020 Coronavirus outbreak

Linlin Sun

3/10/2020

Background

The 2019-2020 coronavirus outbreak is an ongoing global outbreak of coronavirus disease 2019 that has been declared a Public Health Emergency of International Concern. It is caused by the SARS-CoV-2 coronavirus, first identified in Wuhan, Hubei, China. Over 100 countries and territories have been affected at the beginning of March 2020 with major outbreaks in central China, South Korea, Italy, Iran, France, and Germany.

Background of the author.

As a newbie in the data science world, I would like to use the coronavirus data to practice my data wrangling skills. As a Chinese-American, I have spent more than half of my life here in US than in China. I have paid attention to the news in China while Wuhan and other places in China were experiencing the covid-19 spreading. Now, I am experiencing covid-19 spread in US. Beside trying my best to do the social distancing, I would like to make my small portion of contribution to the covid-19 analysis.

I am hoping we will get through this soon and wish the best for everyone!

Data files

I have been referring to Johns Hopkins CSSE Coronavirus

https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf for affected cases. From their website, they provided a github link containing the data they are using to populate the website.

I got the world population information from https://www.worldometers.info/world-population/population-by-country/

Including the library.

```
library(tidyverse)
library(lubridate)
library(matrixStats)
library(kableExtra)
options(digits=2)
```

Loading the data.

```
path <- getwd()
covid_Confirmed_ts <- "data/time_series_covid_19_confirmed_global.csv"
covid_Deaths_ts <- "data/time_series_covid_19_deaths_global.csv"

c_ts <- read.csv(paste(path, covid_Confirmed_ts, sep = "/"), header = TRUE)
c_ts_col_count <- length(colnames(c_ts))
d_ts <- read.csv(paste(path, covid_Deaths_ts, sep = "/"), header = TRUE)
d_ts_col_count <- length(colnames(d_ts))
```

Here is what the time series data look like. I am showing only partial columns due to space issue.

\begin{table}

\caption{Sample entries of time_series_covid_19_confirmed_global.csv}

Province.State	Country.Region	X3.19.20	X3.20.20	X3.21.20	X3.22.20	X3.23.20	X3.24.20
	Afghanistan	22	24	24	40	40	74
	Albania	64	70	76	89	104	123
	Algeria	87	90	139	201	230	264
	Andorra	53	75	88	113	133	164
	Angola	0	1	2	2	3	3

\end{table}

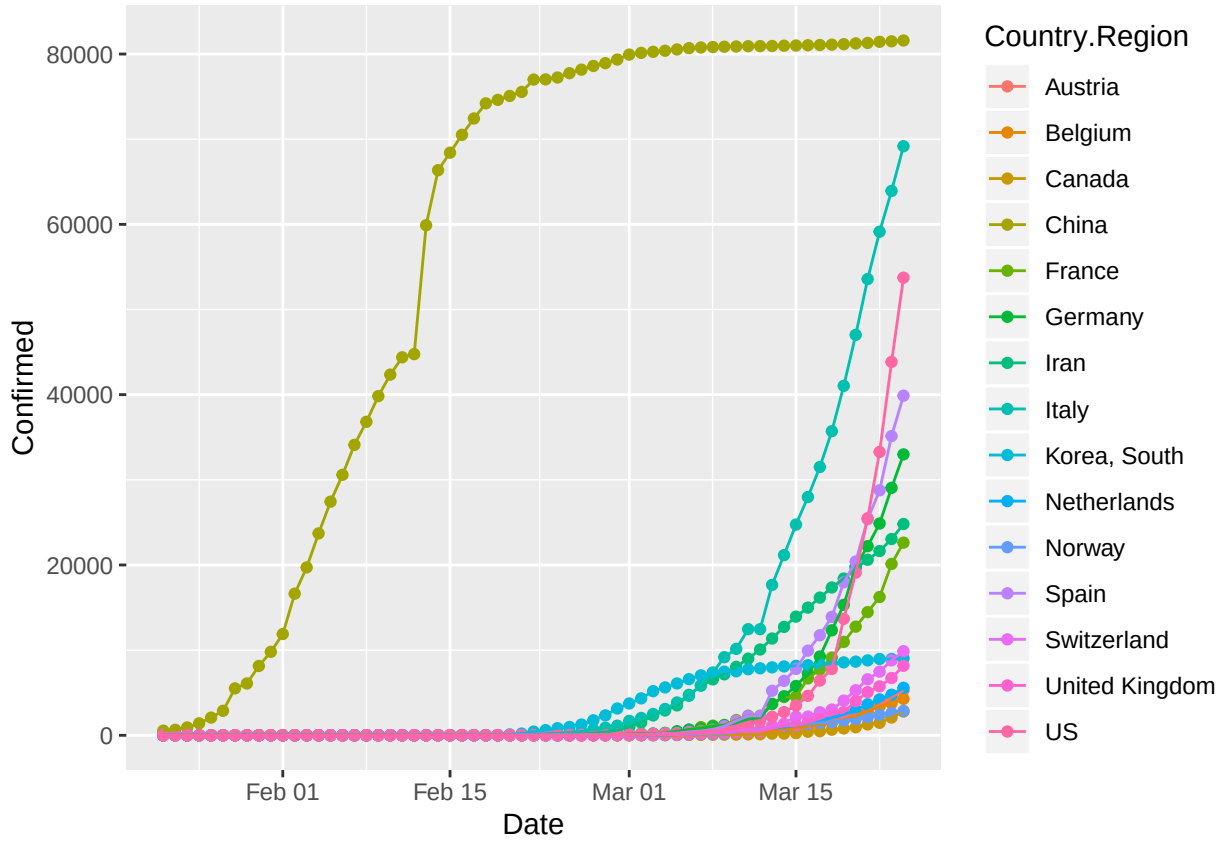
I would like to just focus on US case growth. I woould like to have a plot to simply show how the total confirmed US cases grows each day.

```
confirmed <- c_ts %>% select(-Province.State, -Lat, -Long) %>%
    gather(Date_temp, value, -Country.Region) %>%
    group_by(Country.Region, Date_temp) %>%
    summarize(Confirmed = sum(value)) %>%
    ungroup() %>%
    mutate(Date_temp = str_replace(Date_temp, "X", "")) %>%
    mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y")))

deaths <- d_ts %>% select(-Province.State, -Lat, -Long) %>%
    gather(Date_temp, value, -Country.Region) %>%
    group_by(Country.Region, Date_temp) %>%
    summarize(Deaths = sum(value)) %>%
    ungroup() %>%
    mutate(Date_temp = str_replace(Date_temp, "X", "")) %>%
    mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y")))

confirmed_deaths <- cbind(confirmed[, c(1, 4)], confirmed[, 3], deaths[3])

str(confirmed_deaths)
```

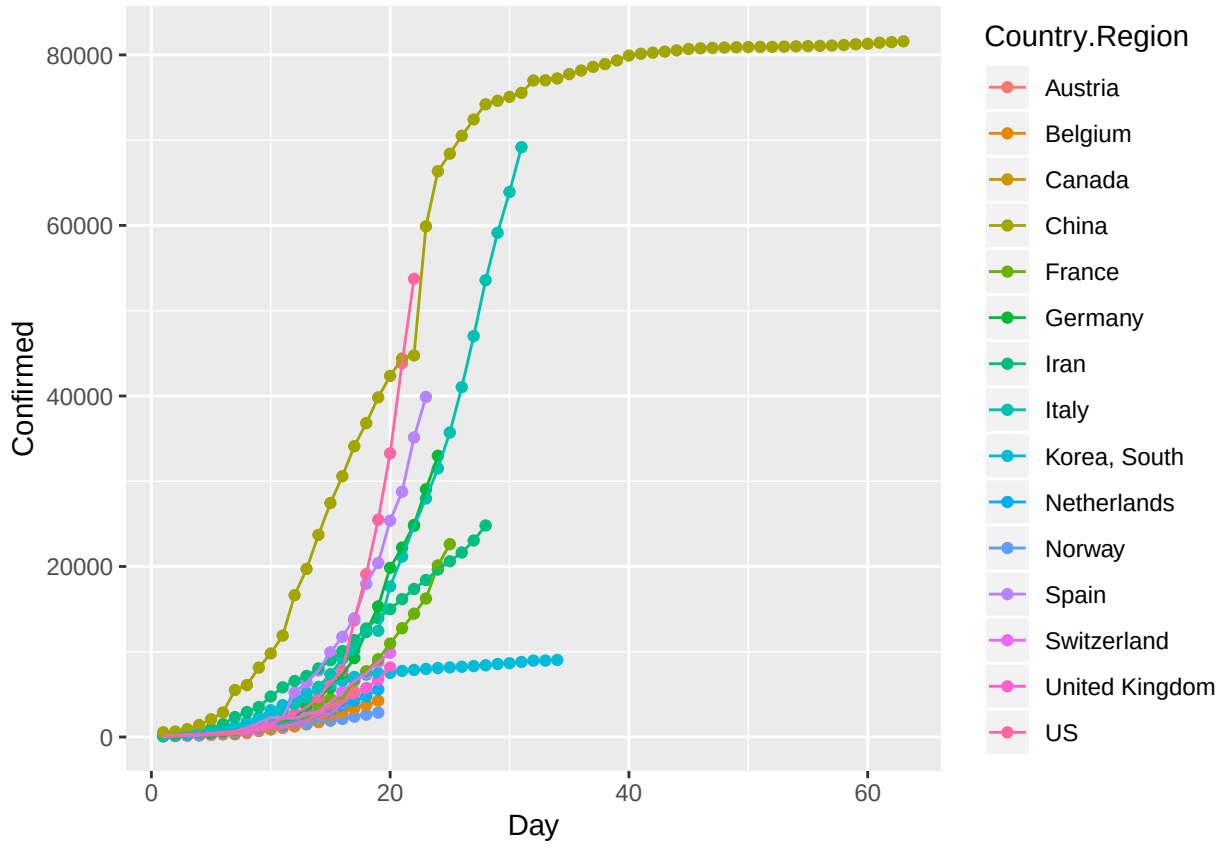
```
## 'data.frame':    10710 obs. of  4 variables:
##  $ Country.Region: Factor w/ 170 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Date          : POSIXct, format: "2020-01-22" "2020-01-23" ...
##  $ Confirmed     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths        : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
today <- max(confirmed$Date)
```

```
select_top <- 15
top_confirmed <- confirmed %>% filter(Date == today) %>%
    arrange(desc(Confirmed)) %>%
    top_n(select_top, Confirmed) %>%
    mutate(Country.Region = as.character(Country.Region)) %>%
    .$Country.Region

# confirmed %>% filter(Country.Region == "United Kingdom")

confirmed %>% filter(Country.Region %in% top_confirmed) %>%
    ggplot(aes(Date, Confirmed, color = Country.Region)) +
    geom_point() +
    geom_line()
```




```
# Day1 plot
confirmed_day1 <- confirmed %>% filter(Confirmed >= 100) %>% group_by(Country.Region) %>%
    arrange(Confirmed) %>% mutate(Day = row_number()) %>% ungroup()

confirmed_day1 %>% filter(Country.Region %in% top_confirmed) %>%
    ggplot(aes(Day, Confirmed, color = Country.Region)) +
    geom_point() +
    geom_line()
```




```
source("WorldPopulation.R")
```

```
## Loading required package: xml2
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:purrr':
##
##     pluck
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
wp <- getWorldPopulation()
```

```
wppd <- wp %>%
    mutate(Density = str_replace_all(Density, ",", "")) %>%
    mutate(Density = as.numeric(Density)) %>%
    mutate(Population = str_replace_all(Population, ",", "")) %>%
    mutate(Population = as.numeric(Population)) %>%
    select(Country.Region, Density, Population, Median_Age)
```

Table 1: World Covid-19 Summary 2020-03-24 C/Population, D/Population, R/Population are per 10,000 people

Country.Region	Date	Confirmed	Deaths	Density	Population	Median_Age	D/C %	C/Population	D/Population	D/C % by Density
China	2020-03-24	81591	3281	153	1.4e+09	38	4.02	0.57	0.02	0.03
Italy	2020-03-24	69176	6820	206	6.0e+07	47	9.86	11.44	1.13	0.05
US	2020-03-24	53740	706	36	3.3e+08	38	1.31	1.62	0.02	0.04
Spain	2020-03-24	39885	2808	94	4.7e+07	45	7.04	8.53	0.60	0.07
Germany	2020-03-24	32986	157	240	8.4e+07	46	0.48	3.94	0.02	0.00
Iran	2020-03-24	24811	1934	52	8.4e+07	32	7.79	2.95	0.23	0.15
France	2020-03-24	22622	1102	119	6.5e+07	42	4.87	3.47	0.17	0.04
Switzerland	2020-03-24	9877	122	219	8.7e+06	43	1.24	11.41	0.14	0.01
Korea, South	2020-03-24	9037	120	527	5.1e+07	44	1.33	1.76	0.02	0.00
United Kingdom	2020-03-24	8164	423	281	6.8e+07	40	5.18	1.20	0.06	0.02
Netherlands	2020-03-24	5580	277	508	1.7e+07	43	4.96	3.26	0.16	0.01
Austria	2020-03-24	5283	28	109	9.0e+06	43	0.53	5.87	0.03	0.00
Belgium	2020-03-24	4269	122	383	1.2e+07	42	2.86	3.68	0.11	0.01
Norway	2020-03-24	2863	12	15	5.4e+06	40	0.42	5.28	0.02	0.03
Canada	2020-03-24	2790	26	4	3.8e+07	41	0.93	0.74	0.01	0.23

```
confirmed_deaths <- confirmed_deaths %>% mutate(Country.Region = as.character(Country.Region))

all <- left_join(confirmed_deaths %>% filter(Date == today), wppd, by = "Country.Region")

# sapply(all, function(col) {sum(is.na(col))})
unique(all[is.na(all$Density),]$Country.Region)
```

```
## [1] "Congo (Brazzaville)" "Congo (Kinshasa)"    "Diamond Princess"
```

```
# Three countries do not get a matching density

# There are two Congo entries in covid-19 data, one congo entry from the world population data
all$Density[str_detect(all$Country.Region, "Congo")] <-
    wppd$Density[wppd$Country.Region == "Congo"]

all$Population[str_detect(all$Country.Region, "Congo")] <-
    wppd$Population[wppd$Country.Region == "Congo"]/2

# Cruise ship is not a real country.
all$Density[str_detect(all$Country.Region, "Cruise Ship")] <- 10
all$Population[str_detect(all$Country.Region, "Cruise Ship")] <-2000
```

```
all <- all %>%
    mutate("D/C %" = (Deaths/Confirmed)*100) %>%
    mutate("C/Population" = (Confirmed/Population)*10000,
           "D/Population" = (Deaths/Population)*10000) %>%
    mutate("D/C % by Density" = `D/C %`/Density) %>%
    arrange(desc(Confirmed)) %>% slice(1:15)

knitr::kable(all,
  caption = paste("World Covid-19 Summary", today, "C/Population, D/Population, R/Population are per 10
  format="latex", booktabs=TRUE) %>%
  kable_styling(latex_options="scale_down")
```

Below are my personal opinion from this table. I could be wrong. Please let me know what you think. I will really appreciate it.

On websites, we normally see the numbers for Confirmed, Deaths, Recovered. Some places might show the Death rate which is calculated as Deaths/Confirmed.

I added population, density and median age of the countries to the table. I calculated the number of cases (Confirmed, Deaths) divided by population multipled by 10,000. So C/Population and D/Population columns shows number of cases per 10,000 people.

"D/C by Density" is calculated by Deaths/Confirmed divided by country density. (Density here is number of people per square km). So this rate removes the density factor. If a country has higher density, that makes the virus to be transmitted more easily.

1. If not considering the population and density factors, China has the highest confirmed case. Italy has the highest Deaths/Confirmed rate (D/C %).
2. Adding consideration of the countries' population, Italy has the highest Confirmed and Death cases per 10,000 people.
3. Adding the consideration of population density, Iran has relative high Deaths/Confirmed by density. Italy is not on the top of this list.
4. For most of the countries which have higher confirmed cases, most of them have D/C % by Density within 0.05. This means to me that the virus transmission seems to be similar across all countries.
5. There can be other factors that contribute to D/C % by Density. If a country has more people older than a certain age, it will be more affected since covid-19 has much worse impact on older people. Italy is a great exmaple. Better medical facilities will have positive effect on this.