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.

Table 1: Sample entries of csse covid-19 time series 03/21/2020

Province.State	Country.Region	X3.16.20	X3.17.20	X3.18.20	X3.19.20	X3.20.20	X3.21.20
	Thailand	147	177	212	272	322	411
	Japan	825	878	889	924	963	1007
	Singapore	243	266	313	345	385	432
	Nepal	1	1	1	1	1	1
	Malaysia	566	673	790	900	1030	1183

```
path <- getwd()
covid_Confirmed_ts <- "data/covid_19_2020_03_21_Confirmed_TS.csv"
covid_Deaths_ts <- "data/covid_19_2020_03_21_Deaths_TS.csv"
covid_Recovered_ts <- "data/covid_19_2020_03_21_Recovered_TS.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))
r_ts <- read.csv(paste(path, covid_Recovered_ts, sep = "/"), header = TRUE)
r_ts_col_count <- length(colnames(r_ts))
```

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

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.

I noticed that the data on 03/10/2020 have different format for Province.State column. They added entries to track the sum for each state. If we blindly sum up all rows, we will end up over counting the data.

So, I am going to wrangle the data as following:

1. Process the data before and on 03/09/2020 the same way since for those days, we do not have extra entries.
2. Process the data on 03/10/2020 separately. I am using the new entries torwards the total count. This includes the Province.State fields without any comma or with the word "Princess" which come from the two cruise ships.

```
US_c_ts_0309 <- c_ts[1:52] %>% filter(Country.Region == "US") %>%
    gather(date, value, 5:52) %>%
    group_by(Country.Region, date) %>%
    summarize(total = sum(value)) %>%
    ungroup() %>%
    mutate(Date_temp = str_replace(date, "X", "")) %>%
    mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y")))

delta <- c_ts_col_count - 53

US_c_ts_after_0309 <- c_ts[, c(1,2,3,4, 53:(53+delta))] %>% filter(Country.Region == "US") %>%
    filter(!str_detect(Province.State, ",") | str_detect(Province.State, "Princess")) %>%
    gather(date, value, 5:(5+delta)) %>%
    group_by(Country.Region, date) %>%
    summarize(total = sum(value)) %>%
    ungroup() %>%
    mutate(Date_temp = str_replace(date, "X", "")) %>%
```

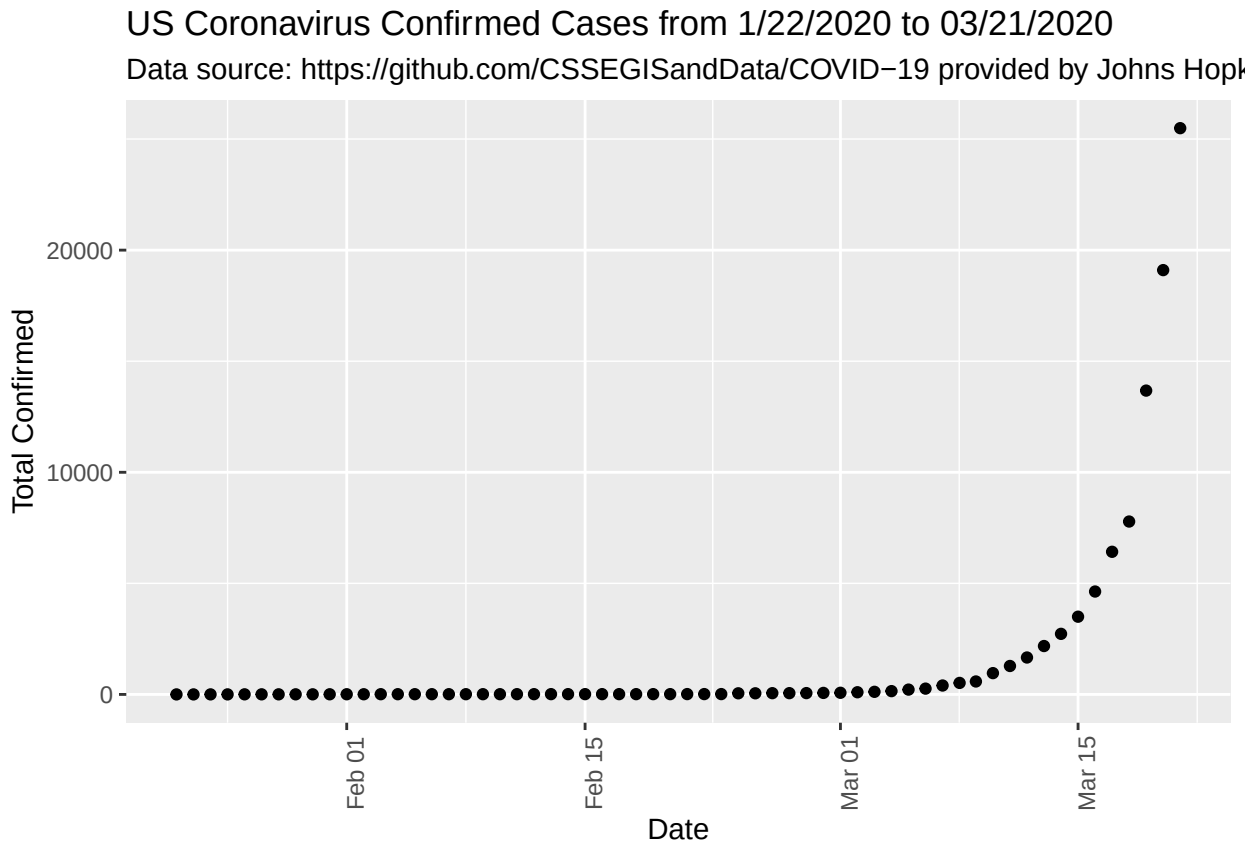
```
    mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y")))

US_c_ts_after_0309
```

```
## # A tibble: 12 x 5
##    Country.Region date     total Date_temp Date
##    <fct>          <chr>    <int> <chr>     <dttm>
##  1 US             X3.10.20   959 3.10.20   2020-03-10 00:00:00
##  2 US             X3.11.20  1281 3.11.20   2020-03-11 00:00:00
##  3 US             X3.12.20  1663 3.12.20   2020-03-12 00:00:00
##  4 US             X3.13.20  2179 3.13.20   2020-03-13 00:00:00
##  5 US             X3.14.20  2727 3.14.20   2020-03-14 00:00:00
##  6 US             X3.15.20  3499 3.15.20   2020-03-15 00:00:00
##  7 US             X3.16.20  4632 3.16.20   2020-03-16 00:00:00
##  8 US             X3.17.20  6421 3.17.20   2020-03-17 00:00:00
##  9 US             X3.18.20  7783 3.18.20   2020-03-18 00:00:00
## 10 US             X3.19.20 13677 3.19.20   2020-03-19 00:00:00
## 11 US             X3.20.20 19100 3.20.20   2020-03-20 00:00:00
## 12 US             X3.21.20 25489 3.21.20   2020-03-21 00:00:00
```

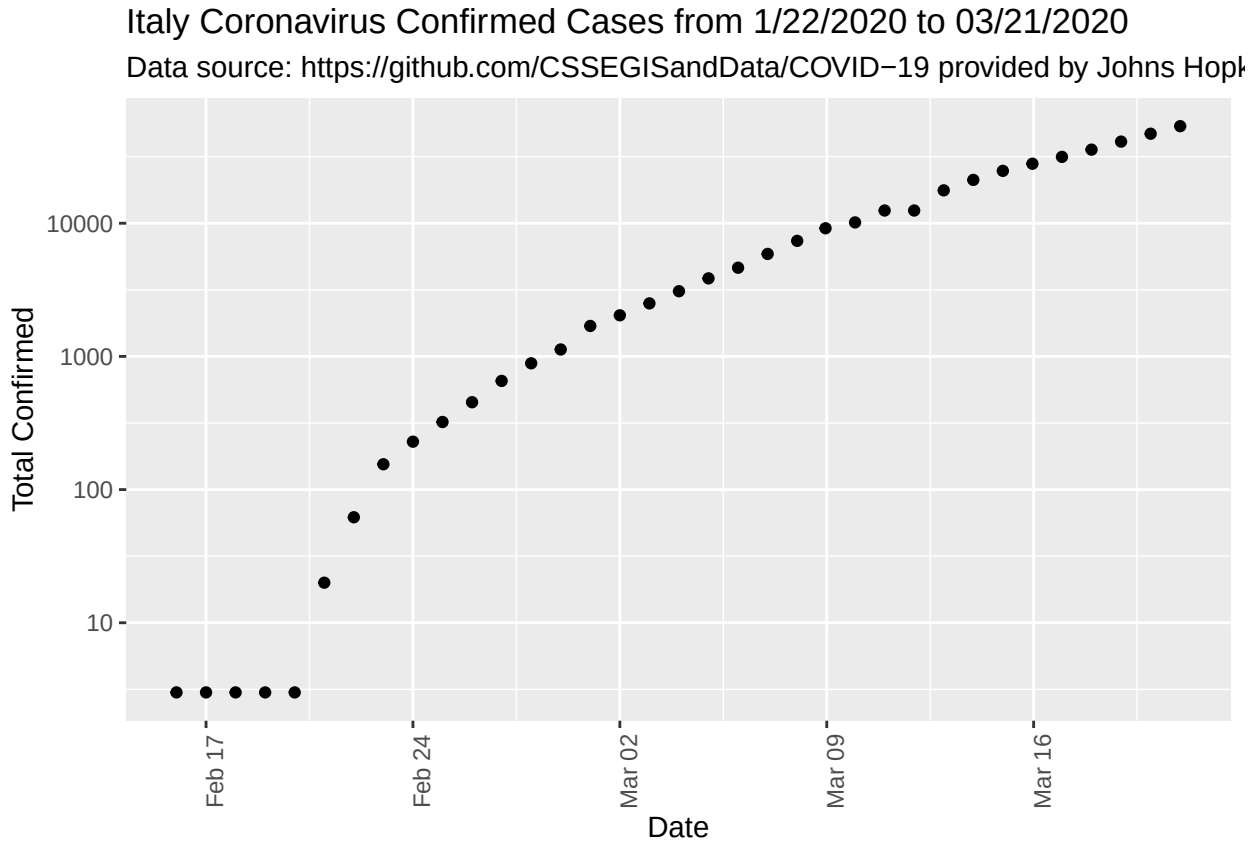
```
US_c_ts <- rbind(US_c_ts_0309, US_c_ts_after_0309) %>% select(Country.Region, Date, total)
today <- as.Date(max(US_c_ts$Date))
US_c_ts
```

```
## # A tibble: 60 x 3
##    Country.Region Date                total
##    <fct>          <dttm>              <int>
##  1 US             2020-01-22 00:00:00     1
##  2 US             2020-01-23 00:00:00     1
##  3 US             2020-01-24 00:00:00     2
##  4 US             2020-01-25 00:00:00     2
##  5 US             2020-01-26 00:00:00     5
##  6 US             2020-01-27 00:00:00     5
##  7 US             2020-01-28 00:00:00     5
##  8 US             2020-01-29 00:00:00     5
##  9 US             2020-01-30 00:00:00     5
## 10 US             2020-01-31 00:00:00     7
## # ... with 50 more rows
```

```
US_c_ts %>%
    ggplot(aes(Date, total)) +
    geom_point() +
    theme(axis.text.x = element_text(angle = 90)) +
    ylab("Total Confirmed") +
    xlab("Date") +
    labs(title = "US Coronavirus Confirmed Cases from 1/22/2020 to 03/21/2020",
         subtitle = "Data source: https://github.com/CSSEGISandData/COVID-19 provided by Johns Hopkins U
```




```
Italy_c_ts <- c_ts %>% filter(Country.Region == "Italy")
Italy_c_ts <- Italy_c_ts %>% gather(date, value, 5:c_ts_col_count) %>%
  mutate(Date_temp = str_replace(date, "X", "")) %>%
  mutate(Date = as.POSIXct(strptime(Date_temp, "%m.%d.%y"))) %>%
  select(Country.Region, Date, total = value)
# Italy_c_ts
Italy_c_ts %>% filter(Date > "2020-02-15") %>%
  ggplot(aes(Date, total)) +
    geom_point() +
    theme(axis.text.x = element_text(angle = 90)) +
    ylab("Total Confirmed") +
    xlab("Date") +
    scale_y_log10() +
    labs(title = "Italy Coronavirus Confirmed Cases from 1/22/2020 to 03/21/2020",
      subtitle = "Data source: https://github.com/CSSEGISandData/COVID-19 provided by Johns Hopkins U")
```




```
md_padded <- function(mydate){
  paste(str_pad(month(mydate), width = 2, side = "left", pad = "0"), "-",
        str_pad(day(mydate), width = 2, side = "left", pad = "0"), sep = ""
        )
}

max_date_shown_Italy <- today - 5
max_date_shown_Italy
```

```
## [1] "2020-03-16"
```

```
US_day_0 <- as.Date("2020-03-04")
Italy_day_0 <- as.Date("2020-02-22")
days_diff <- as.integer(US_day_0 - Italy_day_0)
Italy_c_ts <- Italy_c_ts %>% mutate(Day = as.Date(Date) - Italy_day_0) %>% filter(Date <= max_date_shown
US_c_ts <- US_c_ts %>% mutate(Day = as.Date(Date) - US_day_0)

stop_1 <- as.integer(as.Date("2020-03-13") - US_day_0)
stop_1
```

```
## [1] 9
```

```
stop_2 <- as.integer(today - US_day_0)
stop_2
```

```
## [1] 17
```

```
stop_3 <- as.integer(max_date_shown_Italy - Italy_day_0)
stop_3
```

```
## [1] 23
```

```
day_0_label <- paste("0\nUS", md_padded(US_day_0), "\nItaly", md_padded(Italy_day_0))
day_0_label
```

```
## [1] "0\nUS 03-04 \nItaly 02-22"
```

```
day_school_closed_label <- paste(as.character(stop_1), "\nUS 03-13\nGA school\nclosed", sep = "")
day_school_closed_label
```

```
## [1] "9\nUS 03-13\nGA school\nclosed"
```

```
day_today_label <- paste(as.character(stop_2), "\nUS", md_padded(today), "\nItaly", md_padded(today - 1
day_today_label
```

```
## [1] "17 \nUS 03-21 \nItaly 03-10"
```

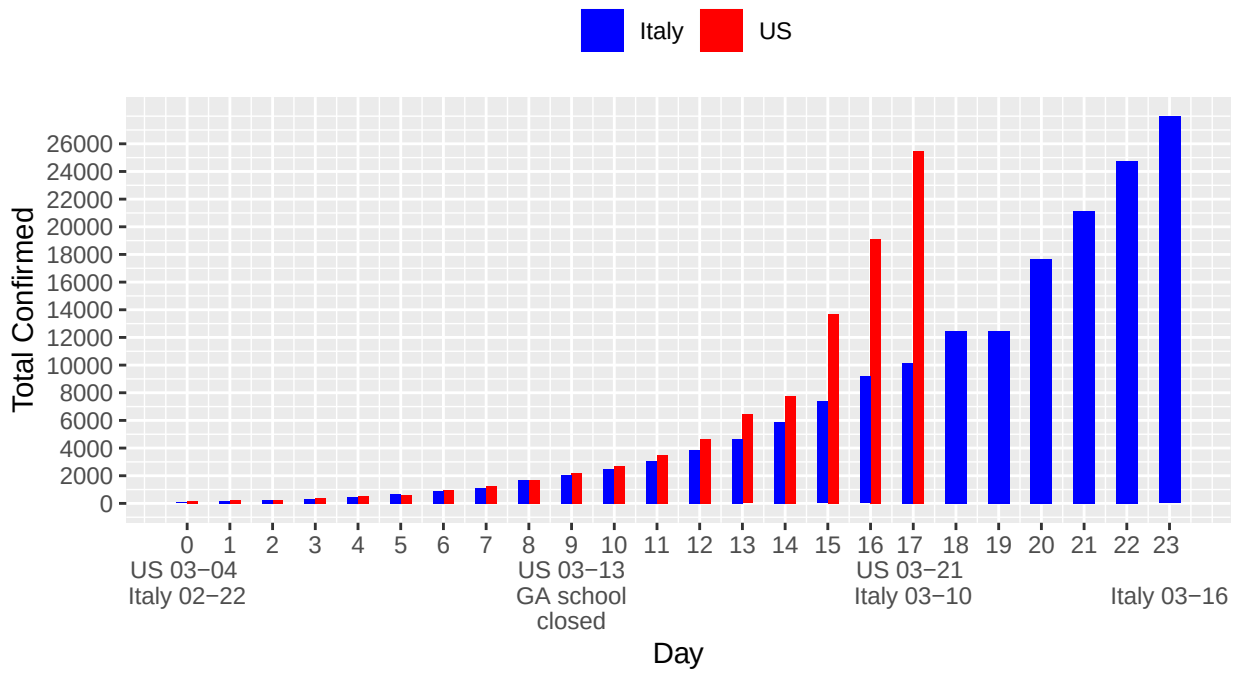
```
day_last_label <- paste(as.character(stop_3), "\n\nItaly", md_padded(max_date_shown_Italy))
day_last_label
```

```
## [1] "23 \n\nItaly 03-16"
```

```
y_limit <- Italy_c_ts[Italy_c_ts$Date == (Italy_day_0 + stop_3), "total"]
y_limit
```

```
## [1] 27980
```

```
# pdf_filename <- paste("US-Italy-", as.character(today), ".pdf", sep = "")
# pdf(pdf_filename, width = 8, height = 6)

rbind(Italy_c_ts, US_c_ts) %>% filter(Day >= 0) %>%
  ggplot(aes(Day, total, fill = Country.Region)) +
  geom_bar(stat = "identity", width = 0.5, position = "dodge") +
  theme(legend.position = "top", legend.title = element_blank()) +
  scale_fill_manual(values = c("blue", "red")) +
  scale_x_continuous(breaks = seq(0, stop_3),
      labels = c(day_0_label, 1:(stop_1 - 1), day_school_closed_label,
                 (stop_1 + 1):(stop_2 - 1), day_today_label,
                 (stop_2 + 1):(stop_3 - 1), day_last_label)) +
  scale_y_continuous(breaks = seq(0, y_limit, 2000)) +
  labs(title = "Covid-19 Confirmed cases Italy vs US (Year 2020)",
       subtitle = "Trying to see if the growth patterns are similar between Italy and US. \nDay 0 were
       caption = "Data Source: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data
       y = "Total Confirmed")
```



ɔurce: https://github.com/CSSEGISandData/COVID−19/tree/master/csse_covid_19_data/csse_covid_19_time_series

```
# dev.off()
```

```
c_by_country <- c_ts %>%
  filter(!Country.Region == "US" | (Country.Region == "US" & !str_detect(Province.State, ","))) %>%
  select(Country.Region, num = c(c_ts_col_count)) %>%
  group_by(Country.Region) %>% summarize(Confirmed = sum(num))

d_by_country <- d_ts %>%
  filter(!Country.Region == "US" | (Country.Region == "US" & !str_detect(Province.State, ","))) %>%
  select(Country.Region, num = c(d_ts_col_count)) %>%
  group_by(Country.Region) %>% summarize(Deaths = sum(num))

r_by_country <- r_ts %>%
  filter(!Country.Region == "US" | (Country.Region == "US" & !str_detect(Province.State, ","))) %>%
  select(Country.Region, num = c(r_ts_col_count)) %>%
  group_by(Country.Region) %>% summarize(Recovered = sum(num))

all_by_country <-
  inner_join(inner_join(c_by_country, d_by_country, by = "Country.Region"),
             r_by_country, by = "Country.Region") %>%
  mutate(Country.Region = as.character(Country.Region))
```

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

all <- left_join(all_by_country, wppd, by = "Country.Region")

# sapply(all, function(col) {sum(is.na(col))})
# all[is.na(all$Density),]
# five countries do not get a matching density
# There is a row for Cape Verde and Cabo Verde, which are the same country
all$Density[all$Country.Region == "Cape Verde"] <-
  wppd$Density[wppd$Country.Region == "Cabo Verde"]

all$Population[all$Country.Region == "Cape Verde"] <-
  wppd$Population[wppd$Country.Region == "Cabo Verde"]

# 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

# There is no entries for Kosovo in world population data. 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$Density[str_detect(all$Country.Region, "Kosovo")] <-1810366/10890
all$Population[str_detect(all$Country.Region, "Kosovo")] <- 1810366

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 = "World Covid-19 Summary 03/21/2020. C/Population, D/Population, R/Population are per 10,000
  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

Table 2: World Covid-19 Summary 03/21/2020. C/Population, D/Population, R/Population are per 10,000 people

Country.Region	Confirmed	Deaths	Recovered	Density	Population	Median_Age	D/C %	C/Population	D/Population	D/C % by Density
China	81305	3259	71857	153	1.4e+09	38	4.01	0.56	0.02	0.03
Italy	53578	4825	6072	206	6.0e+07	47	9.01	8.86	0.80	0.04
US	25489	307	0	36	3.3e+08	38	1.20	0.77	0.01	0.03
Spain	25374	1375	2125	94	4.7e+07	45	5.42	5.43	0.29	0.06
Germany	22213	84	233	240	8.4e+07	46	0.38	2.65	0.01	0.00
Iran	20610	1556	7635	52	8.4e+07	32	7.55	2.45	0.19	0.15
France	14431	562	12	119	6.5e+07	42	3.89	2.21	0.09	0.03
Korea, South	8799	102	1540	527	5.1e+07	44	1.16	1.72	0.02	0.00
Switzerland	6575	75	15	219	8.7e+06	43	1.14	7.60	0.09	0.01
United Kingdom	5067	234	67	281	6.8e+07	40	4.62	0.75	0.03	0.02
Netherlands	3640	137	2	508	1.7e+07	43	3.76	2.12	0.08	0.01
Belgium	2815	67	263	383	1.2e+07	42	2.38	2.43	0.06	0.01
Austria	2814	8	9	109	9.0e+06	43	0.28	3.12	0.01	0.00
Norway	2118	7	1	15	5.4e+06	40	0.33	3.91	0.01	0.02
Sweden	1763	20	16	25	1.0e+07	41	1.13	1.75	0.02	0.05

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.