National Technical Univerisy of Athens
Data Science and Machine Learning

Programming Tools and Technologies for Data Science

Exploratory Data Analysis using R

Nanos Georgios
MSc student
03400144
nanosgiwrgos1997@gmail.com

"Information is the oil of the 21st century, and analytics is the combustion engine."
– Peter Sondergaard

Aim of this Assignment

The purpose of this paper is to research and analyze covid19_vaccine dataset which contains a daily summary of the COVID-19 vaccination by country and province. The raw data is being pulled from the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) Coronavirus repository is being updated regularly. The results of this assignment are exported by using the covid19_vaccine dataset [1] which was last updated on **23-1-2022**. The format of the data is shaped by 18 variables but, our variables of interest are the country and consequently the continent, the date, the people who are fully and partially vaccinated and the population of each country.

Preprocessing Data Analysis

Firstly, we import some useful libraries that we are going to use through our analysis.

```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(RColorBrewer)
```

We read the data into R and save the result into a dataframe.

```
data <- read.table('./covid19_vaccine.csv', header = T, sep = ',')
```

We check the names of the data and their type.

```
tibble(names(data), sapply(data, class))
```

R output

A tibble: 18 × 2

	‘names(data)‘	‘sapply(data, class)‘
	<chr>	<chr>
1	country_region	factor
2	date	factor
3	doses_admin	numeric
4	people_partially_vaccinated	numeric
5	people_fully_vaccinated	numeric
6	report_date_string	factor
7	uid	integer
8	province_state	factor
9	iso2	factor
10	iso3	factor
11	code3	integer
12	fips	logical
13	lat	numeric
14	long	numeric
15	combined_key	factor
16	population	integer
17	continent_name	factor
18	continent_code	factor

We also remove the information on provinces and check whether there are any missing values in our dataset.

```
data = subset(data, select = -province_state)
cat('Number of missing values are:', sum(is.na(data)) + sum(is.null(data
   )))
cat('The number of countries is:', length(unique(data$country_region)) -
    1)
```

R output

Number of missing values are: 844297
The number of countries is: 190

We then calculate the ratio of the partially and the fully vaccinated people and add for every single day and country and add those two columns in the dataframe.

```
ratio_partially <- data[, 'people_partially_vaccinated']/data[, '
   population'] * 100
ratio_fully <- data[, 'people_fully_vaccinated']/data[, 'population'] *
   100

data$partially_vaccinated_ratio <- as.numeric(format(round(ratio_
   partially, 2), nsmall = 2, format = "f"))
data$fully_vaccinated_ratio <- as.numeric(format(round(ratio_fully, 2),
   nsmall = 2, format = "f"))
```

Exploratory Data Analysis

To start with, let's observe our country's (Greece) fully vaccinated ratio.

```
data %>%
  filter(country_region %in% c('Greece')) %>%
  ggplot() +
  geom_smooth(mapping = aes(x = date, y = fully_vaccinated_ratio, color
     = country_region)) +
  labs(x = "Date",
       y = "Fully vaccinated ratio",
       title = "People Fully Vaccinated Ratio in Greece")
```

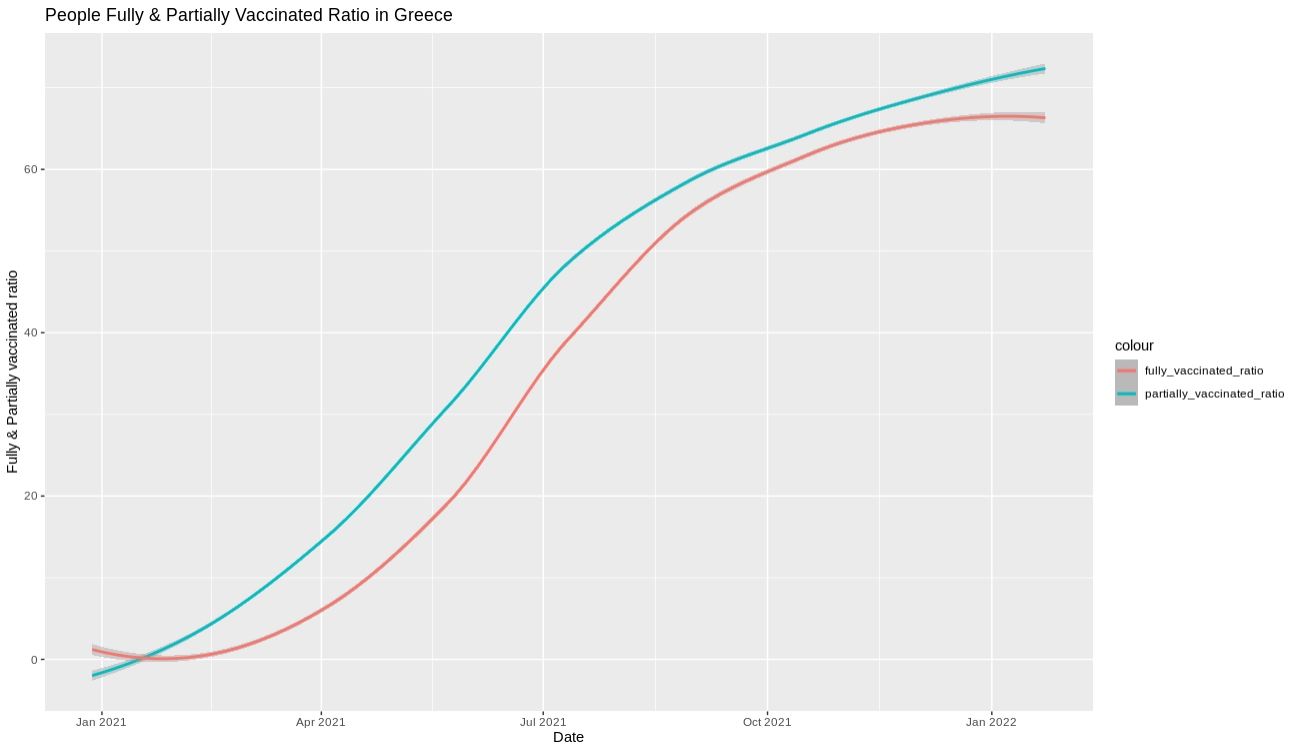

Figure 1: People Fully Vaccinated Ratio in Greece since January 2021 until 23-01-2022

We can observe that in Greece until January 2022 a little more than 60% of the total population. It would be much more interesting to compare Greece with countries with a similar population. We calculate the total population of Greece and take the mean value so that we can compare Greece 's fully vaccinated ratio with other countries of Europe that have approximately ±10% of Greece's population.

```
gr_popul = mean(data[ grep("Greece", data$country_region) , ]$population
   )
likegr_eu = data[data$population >= gr_popul*0.9 & data$population <= gr
   _popul*1.1 & data$continent_name =="Europe",]
```

We created a dataframe which contains the data of "similar" countries with Greece and plot the fully vacinated ratio of those countries.

```
likegr_eu %>%
  filter(country_region %in% (unique(likegr_eu$country_region))) %>%
  ggplot() +
  geom_smooth(mapping = aes(x = date, y = fully_vaccinated_ratio, color
     = country_region)) +
  labs(x = "Date",
       y = "Fully vaccinated ratio",
       title = "People Fully Vaccinated Ratio in Europe's countries that
           have approximately the same population with Greece")+ theme(
          plot.title = element_text(size=10))
```

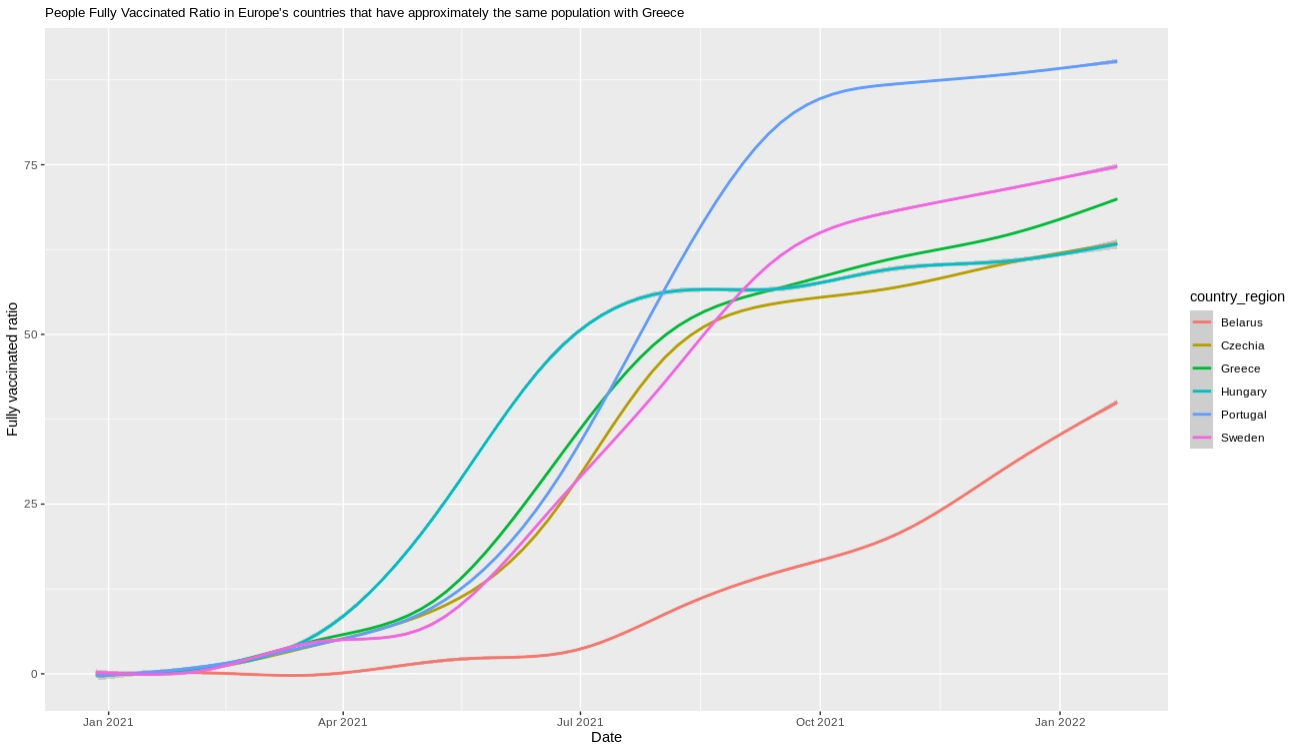

Figure 2: People Fully Vaccinated Ratio in Europe's countries that have approximately the same population with Greece since January 2021 until 23-01-2022

Above we compare the vaccination ratio of Greece with other countries of Europe that have approximately the same population with Greece ($\pm$ 10%). We can easily observe that Greece is theoretically above the average in comparison with other countries with comparable population. Portugal and Sweden are have the highest ratios of all with their ratios reaching approximately 81% and 75% respectively. Our next step would be to find out which countries of Europe seem to have the highest fully vaccinated ratio, which we will consider high as 75% and more.

```
eu = data[data$fully_vaccinated_ratio>75 & data$fully_vaccinated_ratio<1
   00 &data$continent_name == "Europe", ]
eu = eu[!is.na(eu$country_region), ]
mycolors = c(brewer.pal(name="Dark2", n = length(unique(eu$country_
   region))%/%6), brewer.pal(name="Paired", n = length(unique(eu$country
   _region)) - length(unique(eu$country_region))%/%6))
ggplot(eu, aes(x=date, y=fully_vaccinated_ratio, group=1, color =
   country_region ))+
  geom_line() +
  scale_color_manual(values = mycolors)+
  labs(x = "Date", y = "Fully vaccinated ratio", title = "Countries with
      fully vaccinated more than 75% ratio per date in Europe") + theme(
     plot.title = element_text(size=10)) + theme_minimal()
```

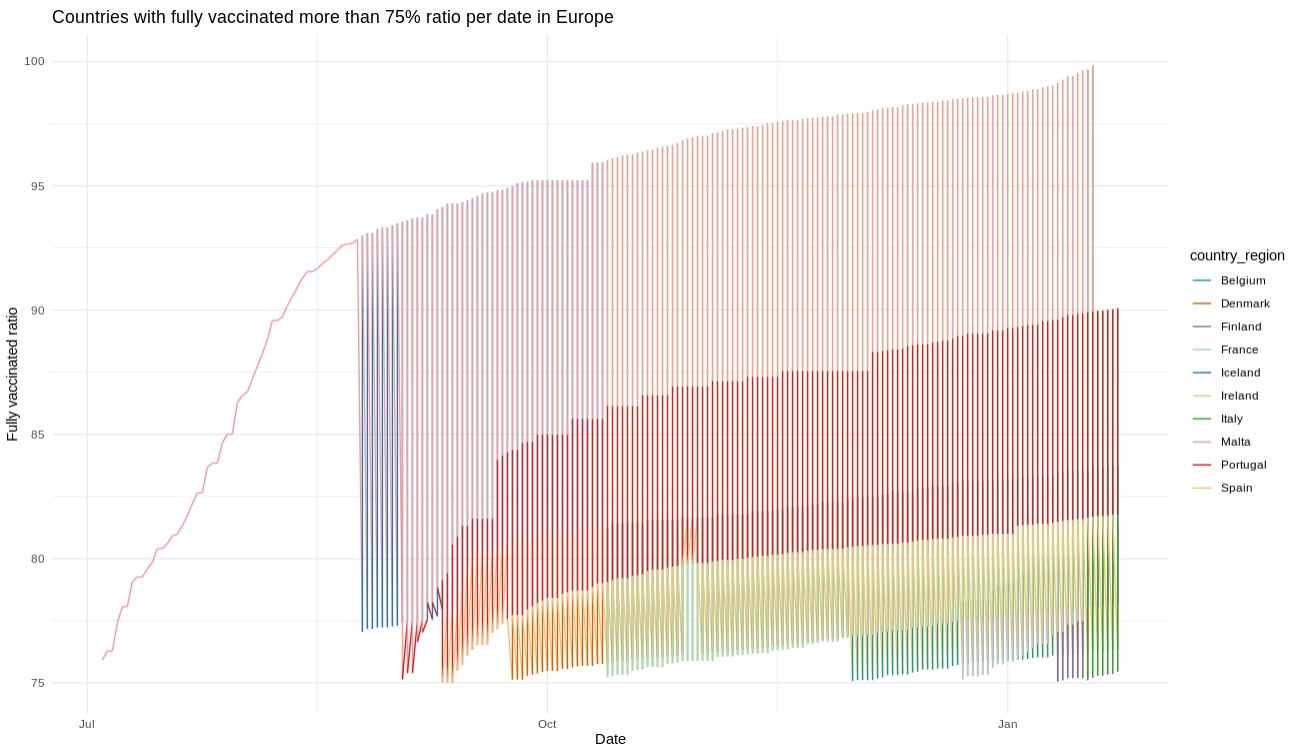

Figure 3: Countries with ratio of fully vaccinated more than 75% ratio per date in Europe since January 2021 until 23-01-2022

In figure 3 we see that in Europe Malta in is the 1st country that reaches a vaccination rate of over 75% first with a big difference, about 2.5 months before the rest start, surpassing Iceland and Portugal and Denmark which also reach a percentage vaccination above 75% before October. Since we have taken a taste about Europe's stats of fully vaccinated ratios per country, for our next step we perform some calculations to extract data and results for each continent.

```
asia = 0
asia_pop = 0
africa = 0
africa_pop = 0
europe = 0
europe_pop = 0
north_a = 0
north_a_pop = 0
south_a = 0
south_a_pop = 0
oceania = 0
oceania_pop = 0
for (country in unique(data$country_region)) {
  if(country!='World'){
    eachcountry = data[data$country_region == country, ]

    if(!is.na(eachcountry$continent_name)){
      vac = eachcountry$people_fully_vaccinated[length(eachcountry$
         people_fully_vaccinated)]
      for (vac in rev(eachcountry$people_fully_vaccinated)) {
        if(!is.na(vac))
          break
```

```
      }
      if(!is.na(vac)){
        if(eachcountry$continent_name == 'Asia'){
          asia = asia + vac
          asia_pop = asia_pop + max(eachcountry$population,na.rm = TRUE)
        }
        if(eachcountry$continent_name == 'Europe'){
          europe = europe + vac
          europe_pop = europe_pop + max(eachcountry$population,na.rm =
             TRUE)
        }
        if(eachcountry$continent_name == 'Africa'){
          africa = africa + vac
          africa_pop = africa_pop + max(eachcountry$population,na.rm =
             TRUE)
        }
        if(eachcountry$continent_name == 'North America'){
          north_a = north_a + vac
          north_a_pop = north_a_pop + max(eachcountry$population,na.rm =
              TRUE)
        }
        if(eachcountry$continent_name == 'South America'){
          south_a = south_a + vac
          south_a_pop = south_a_pop + max(eachcountry$population,na.rm =
              TRUE)
        }
        if(eachcountry$continent_name == 'Oceania'){
          oceania = oceania + vac
          oceania_pop = oceania_pop + max(eachcountry$population,na.rm =
              TRUE)
        }
      }
    }
  }
}
for (max_date_asia in as.character(rev(data[data$continent_name == "Asia
   ", ]$date))) {
  if(!is.na(max_date_asia))
    break
}
for (max_date_europe in as.character(rev(data[data$continent_name == "
   Europe", ]$date))) {
  if(!is.na(max_date_europe))
    break
}
for (max_date_africa in as.character(rev(data[data$continent_name == "
   Africa", ]$date))) {
  if(!is.na(max_date_africa))
    break
}

for (max_date_north_a in as.character(rev(data[data$continent_name == "
   North America", ]$date))) {
  if(!is.na(max_date_north_a))
```

```
    break
}
for (max_date_south_a in as.character(rev(data[data$continent_name == "
   South America", ]$date))) {
  if(!is.na(max_date_south_a))
    break
}
for (max_date_oceania in as.character(rev(data[data$continent_name == "
   Oceania", ]$date))) {
  if(!is.na(max_date_oceania))
    break
}
asia_ratio_fully_vaccinated = asia/asia_pop
africa_ratio_fully_vaccinated = africa/africa_pop
europe_ratio_fully_vaccinated = europe/europe_pop
north_a_ratio_fully_vaccinated = north_a/north_a_pop
south_a_ratio_fully_vaccinated = south_a/south_a_pop
oceania_ratio_fully_vaccinated = oceania/oceania_pop
continent_ratios = c(asia_ratio_fully_vaccinated, africa_ratio_fully_
   vaccinated, europe_ratio_fully_vaccinated, north_a_ratio_fully_
   vaccinated, south_a_ratio_fully_vaccinated,oceania_ratio_fully_
   vaccinated)
continents = c("Asia", "Africa", "Europe", "North America", "South
   America", "Oceania")
continent.data = data.frame(continent_ratios, continents)

ggplot(continent.data, aes(x=continents, y=continent_ratios, fill=
   continents)) +
  geom_bar(stat="identity", color="black", position=position_dodge(0.9),
      size=1) +
  geom_text(aes(label=sprintf("%0.4f", round(continent_ratios, digits =
     4))), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5) +
  scale_fill_brewer(palette="Paired") +
  labs(x = "Continets", y = "Fully vaccinated ratio", title = "People
     Fully Vaccinated Ratio per continent") + theme(plot.title = element
     _text(size=15)) + theme_minimal() + scale_x_discrete(guide = guide_
     axis(angle = 90))
```

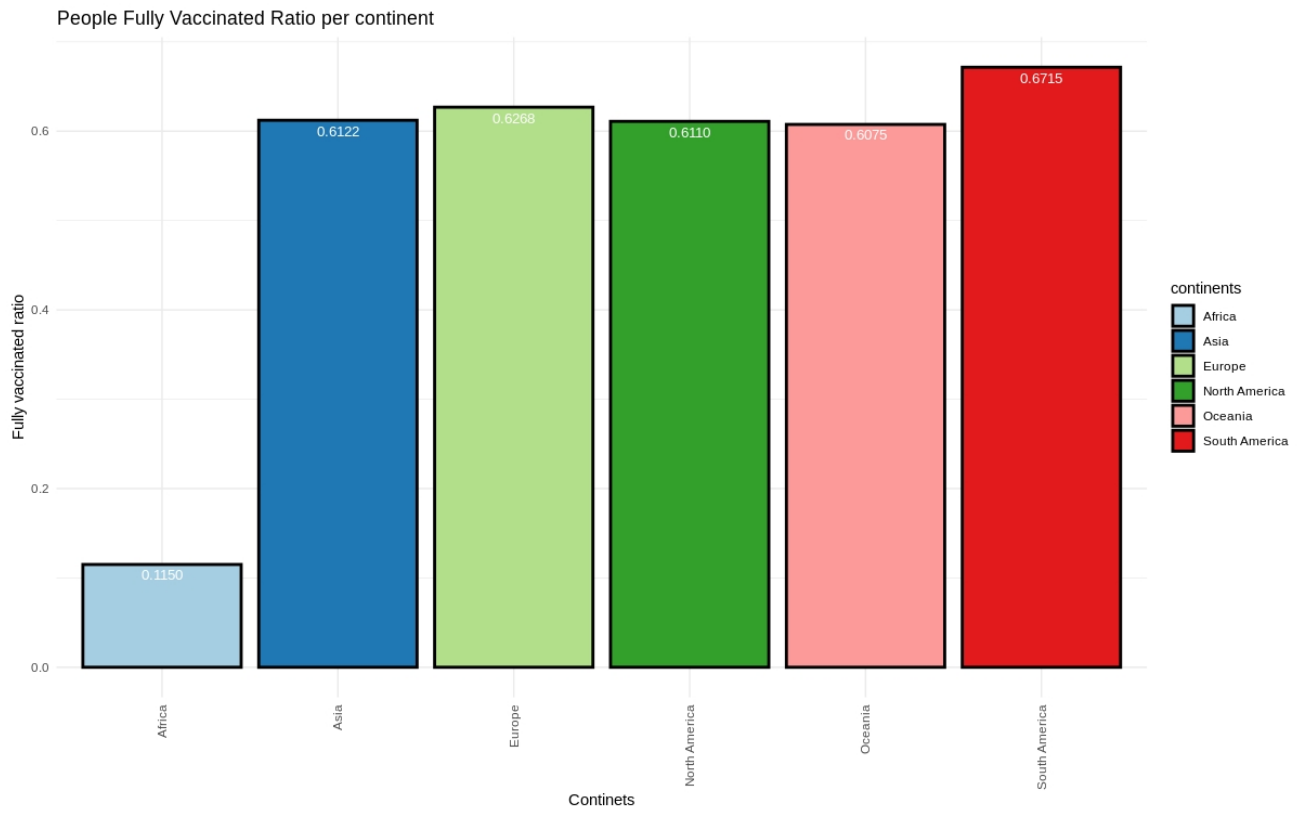

Figure 4: People Fully Vaccinated Ratio per continent until 23-01-2022

As we can observe South America has the highest ratio of fully vaccinated people, which seems quite surprising. Also, it is quite interesting that Europe stands third between the six continents. Africa's vaccination rate is worryingly low (little more than 10%), which is something that has to be taken seriously into consideration by the World Health Organization, in order to limit the spread of covid19.

```
sa = data[data$continent_name=="South America", ][!is.na(data[data$
    continent_name=="South America", ]$country_region), ]
sa = sa[sa$people_fully_vaccinated>0 & !is.na(sa$people_fully_vaccinated
    ), ]
ggplot(sa, aes(x=country_region, y=(people_fully_vaccinated), fill=
    country_region)) +
  geom_bar(stat="identity", color="black", position=position_dodge(1),
      size=.5) +
  scale_fill_brewer(palette="Paired") +
  labs(x = "South America Countries", y = "People Fully Vaccinated",
      title = "People Fully Vaccinated in South America") + theme(plot.
      title = element_text(size=10)) + theme_minimal() +
  scale_x_discrete(guide = guide_axis(angle = 90))

  ggplot(sa, aes(x=country_region, y=(fully_vaccinated_ratio), fill=
      country_region)) +
  geom_bar(stat="identity", color="black", position=position_dodge(1),
      size=.5) +
  scale_fill_brewer(palette="Paired") +
  labs(x = "South America Countries", y = "Fully vaccinated ratio",
      title = "People Fully Vaccinated Ratio in South America") + theme(
      plot.title = element_text(size=10)) + theme_minimal() +
```

```
  scale_x_discrete(guide = guide_axis(angle = 90))
```

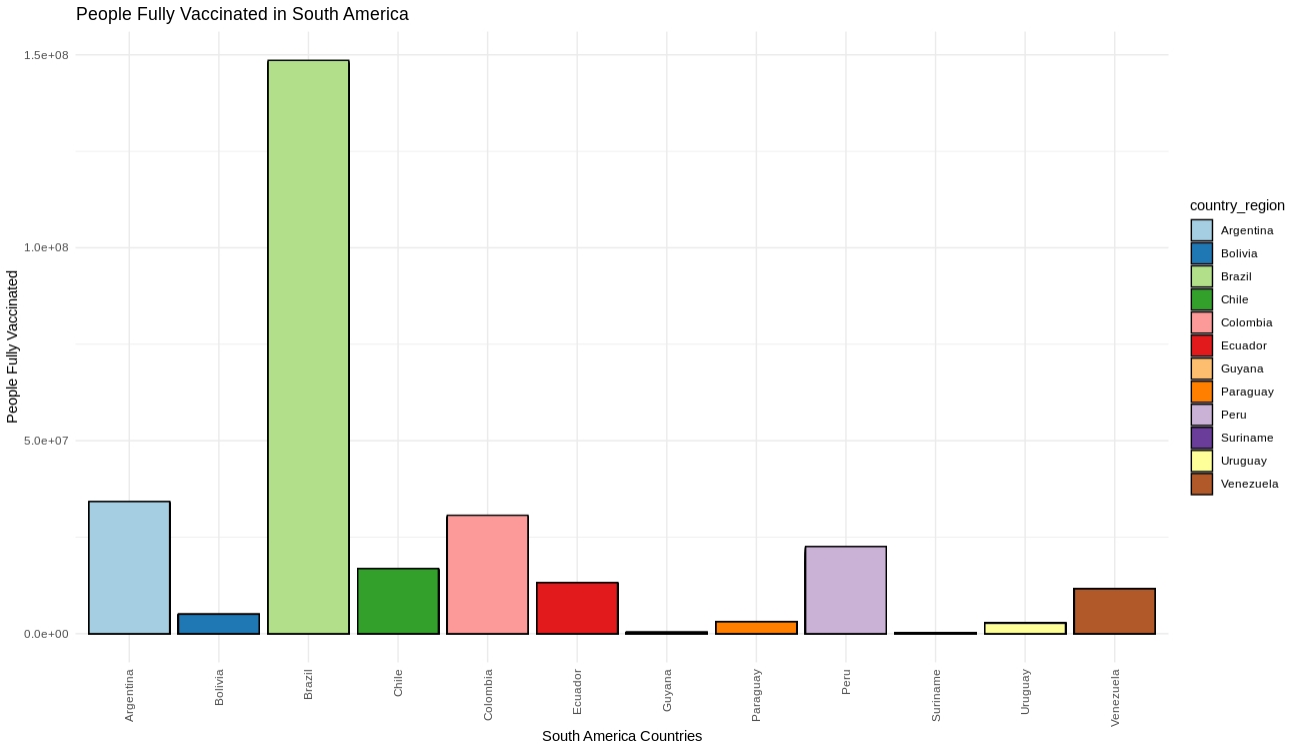

Figure 5: People Fully Vaccinated in South America until 23-01-2022

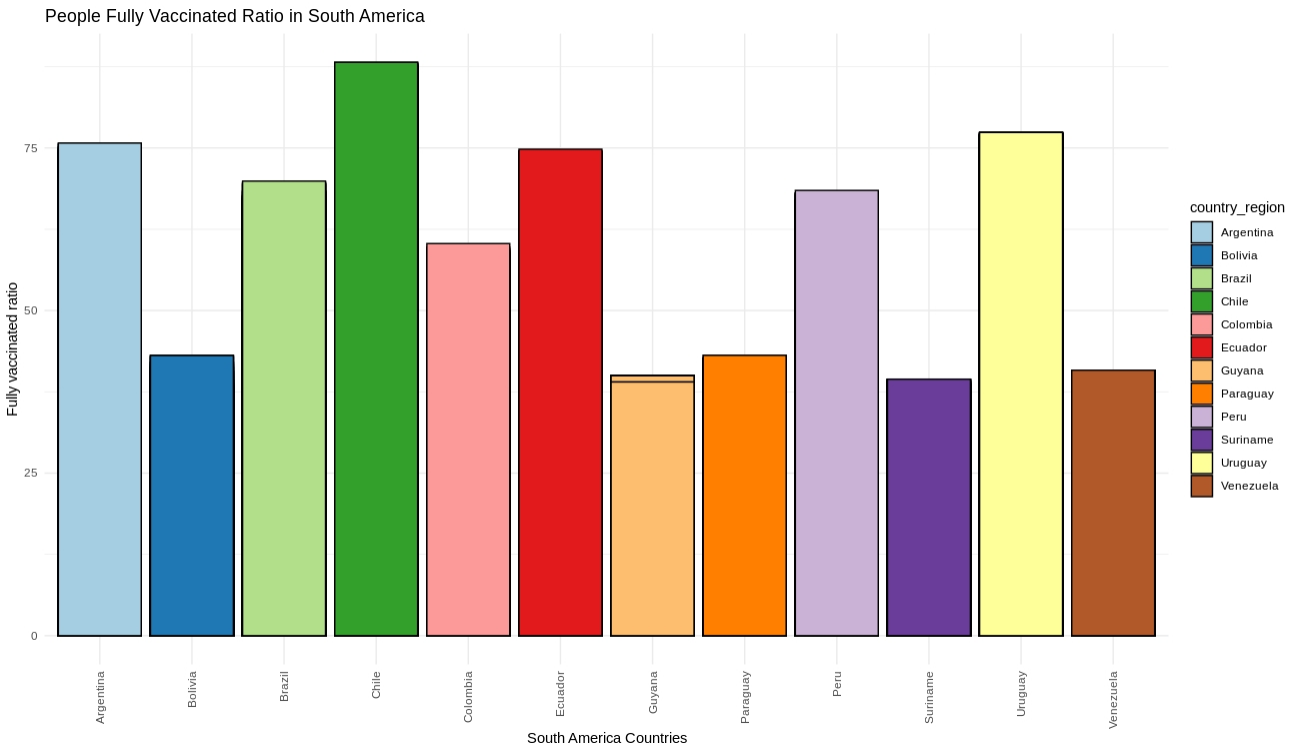

Figure 6: People Fully Vaccinated Ratio in South America until 23-01-2022

As we can see Chile has the highest ratio in South America with more than 85% in total. We

also see that Argentina, Ecuador and Uruguay have also a high ratio up to 75%. Furthermore, we observe that even though Brazil has a total ratio below than 70% and also seems to be fifth since the day we are doing this survey, in general and as we show in the figure above it has the most fully vaccinated people, approximately 150 millions. Thus, we may conclude that Brazil is the main reason that South America leads worldwide in terms of fully vaccinated ratio as we show in the figure above. Also for our next step of the analysis, we make further observations globally, by grouping the countries that have the largest populations on earth.

```
most_pop = data[data$population>max(data$population,na.rm = TRUE)*0.09 &
    data$population<max(data$popula
most_pop = most_pop[!is.na(most_pop$country_region), ]
ggplot(most_pop, aes(x=country_region, y=(fully_vaccinated_ratio), fill=
   country_region)) +
  geom_bar(stat="identity", color="black", position=position_dodge(1),
     size=.5) +
  scale_fill_brewer(palette="Paired") +
  labs(x = "Most popululated countries worldwide", y = "Fully vaccinated
      ratio", title = "People Fully Vaccinated Ratio in the countries
     with the largest population") + theme(plot.title = element_text(
     size=10)) + theme_minimal() +
  scale_x_discrete(guide = guide_axis(angle = 90))
```

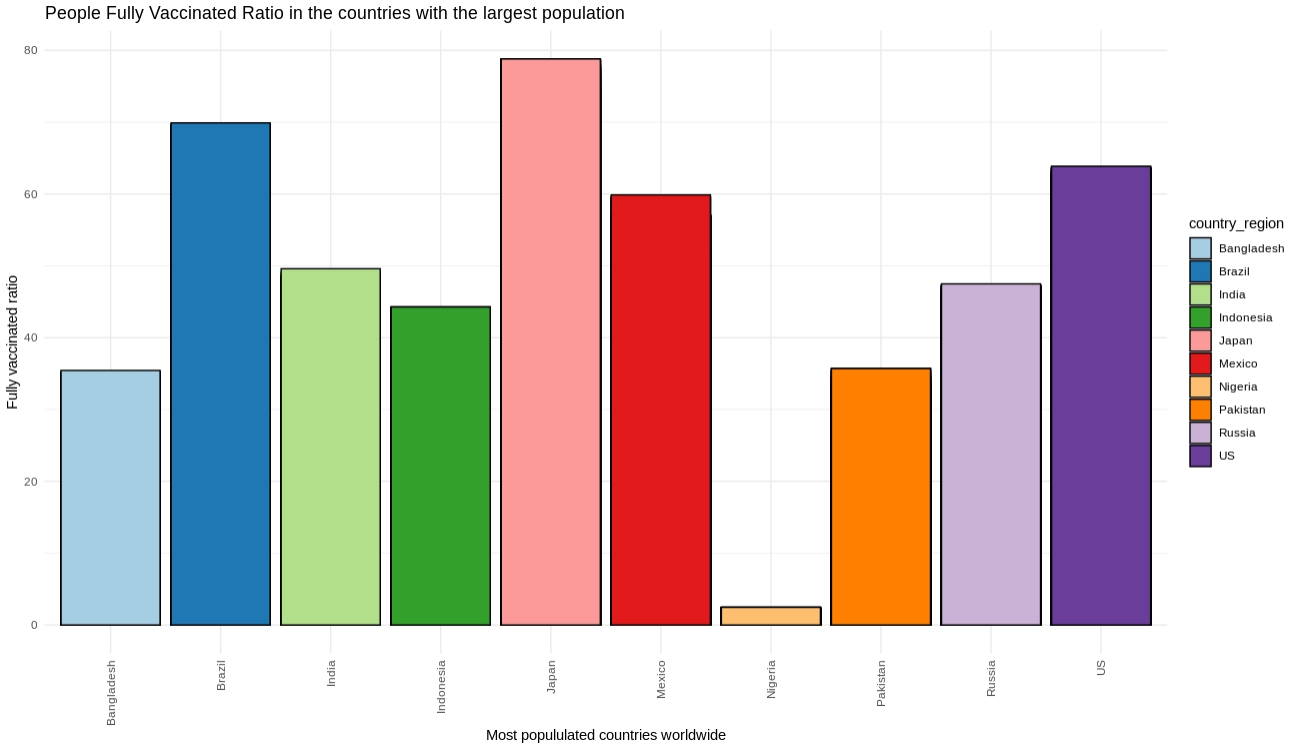

Figure 7: People Fully Vaccinated Ratio globally for the countries with the largest population, until 23-01-2022

As we can observe in the figure 7 Japan has the highest ratio of fully vaccinated people. Except of Brazil USA and Mexico which exceed the threshold of 60% the others the other highly populated countries seem to have rates below 50%. This may be due in part to the fact that these countries are quite densely populated and are generally considered poor.

```
most_ratio = data[data$fully_vaccinated_ratio>82 & data$fully_vaccinated
   _ratio<100, ]
```

```
most_ratio = most_ratio[!is.na(most_ratio$country_region), ]
ggplot(most_ratio, aes(x=country_region, y=(fully_vaccinated_ratio),
   fill=country_region)) +
  geom_bar(stat="identity", color="black", position=position_dodge(1),
     size=.5) +
  scale_fill_brewer(palette="Paired") +
  labs(x = "Countries", y = "Fully vaccinated ratio", title = "People
     Fully Vaccinated Ratio in countries with the highest ratios
     globally") + theme(plot.title = element_text(size=10)) + theme_
     minimal() +
  scale_x_discrete(guide = guide_axis(angle = 90))
```

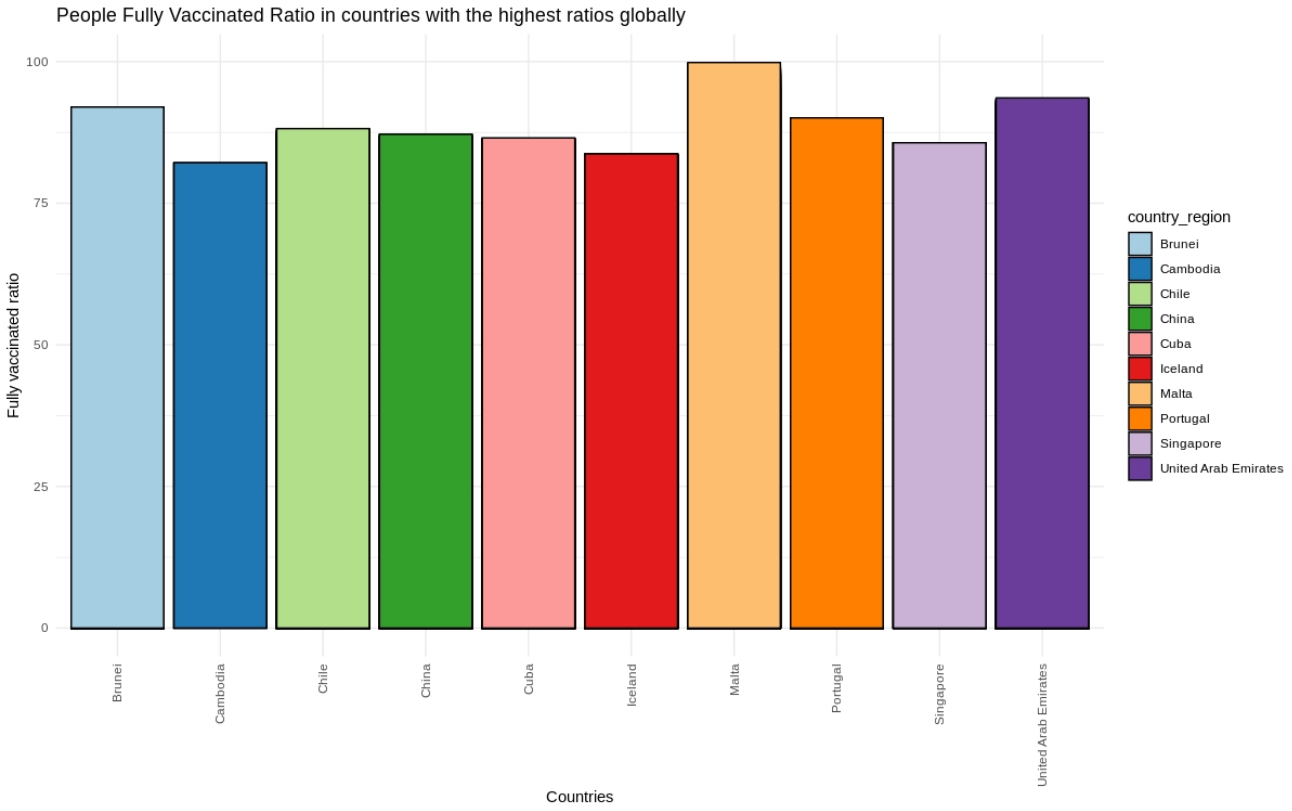

Figure 8: People Fully Vaccinated Ratio in countries with the highest ratios globally, until 23-01-2022

Globally, Malta with population around 500 thousand achieves the highest percentage of vaccinated people by reaching very close to 100%. It is followed by the United Arab Emirates while it is quite impressive that the third country with the highest percentage is Brunei with a population of relatively small number around 400 thousand inhabitants. The countries that exceeded the barrier of 10 million inhabitants and have high vaccination rates (higher than 75%) are Cambodia, Chile, China, Cuba and Portugal.

```
ggplot(most_ratio, aes(x=country_region, y=(partially_vaccinated_ratio),
    fill=country_region)) +
  geom_bar(stat="identity", color="black", position=position_dodge(1),
     size=.5) +
  scale_fill_brewer(palette="Paired") +
  labs(x = "Countries", y = "Fully vaccinated ratio", title = "People
     Partially Vaccinated Ratio in countries with the highest ratios
```

```
    globally") + theme(plot.title = element_text(size=10)) + theme_
    minimal() +
  scale_x_discrete(guide = guide_axis(angle = 90))
```

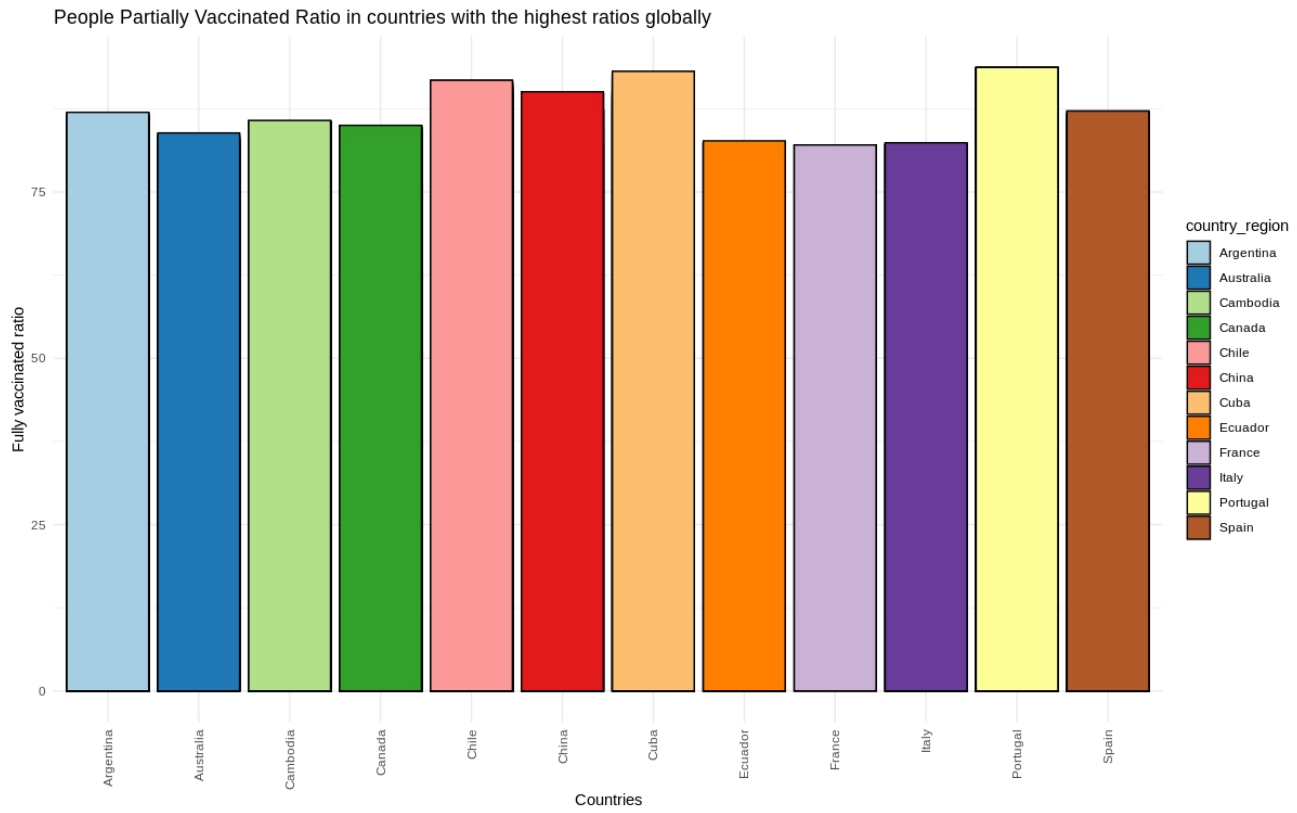

Figure 9: People Partially Vaccinated Ratio in countries with the highest ratios globally, until 23-01-2022

We observe that the countries that have the highest rates as we expected differ from those that have the highest rates of fully vaccinated. As of the date of our analysis, we observe that Portugal has the highest rates of single-dose vaccination, which is not particularly impressive because it is, as we have seen, higher than the highly vaccinated countries in Europe 3. High percentages are presented by Argentina, Canada, France, Italy and Spain, which obviously can be seen to be late in reaching high vaccination rates, since they do not appear in the previous charts concerning rates of fully vaccinated.

```
mycolors = c(brewer.pal(name="Dark2", n = length(unique(most_ratio$
   country_region))%/%3), brewer.pal(name="Paired", n = length(unique(
   most_ratio$country_region)) - length(unique(most_ratio$country_region
   ))%/%3))

ggplot(most_ratio, aes(x=date, y=partially_vaccinated_ratio, group=1,
   color = country_region ))+
  geom_line() +
  scale_color_manual(values = mycolors)+
  labs(x = "Countries with the highest partial vaccinated ratio globally
      per date", y = "Partially vaccinated ratio", title = "People
     Partially Vaccinated Ratio in countries with the highests ratios
     globally") + theme(plot.title = element_text(size=10)) + theme_
     minimal()
```

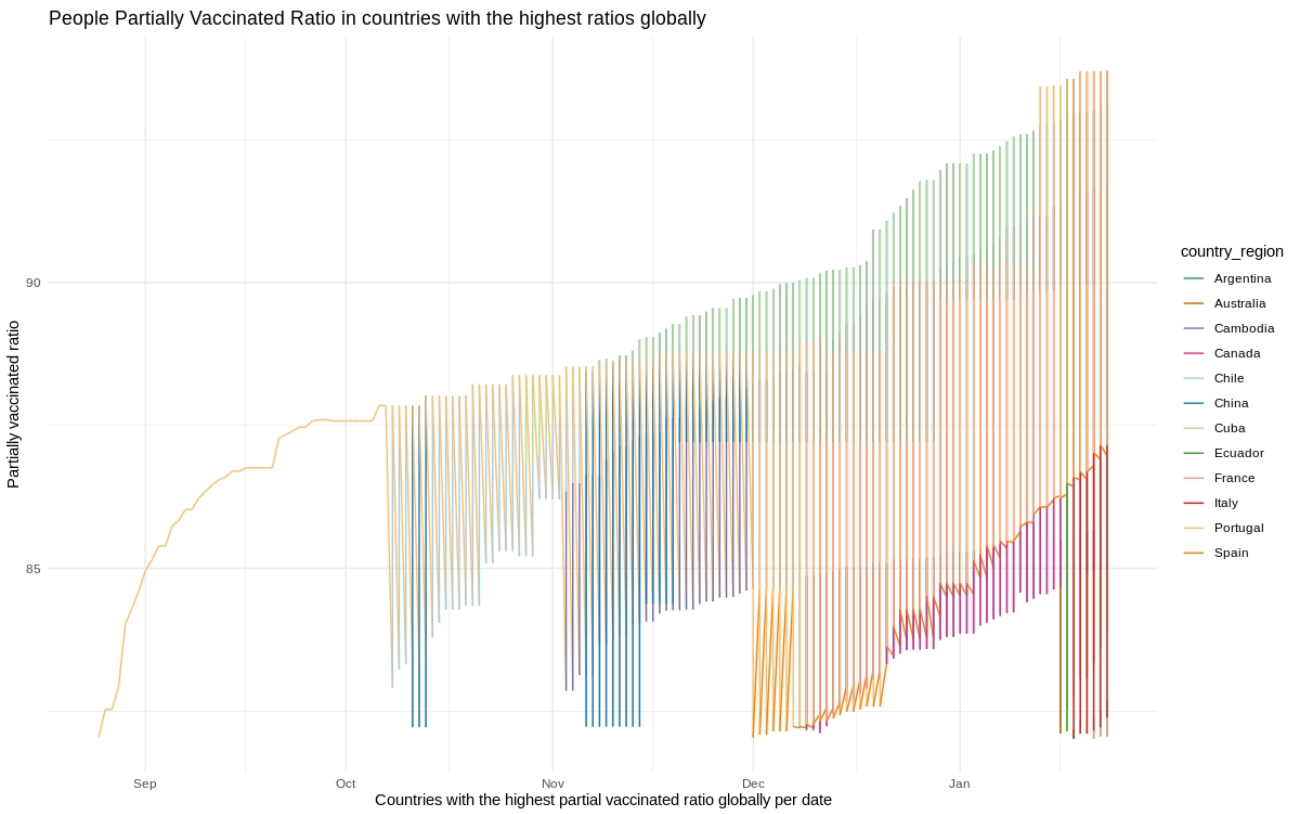

Figure 10: People Partially Vaccinated Ratio in countries with the highests ratios globally per day, until 23-01-2022

As we saw above in figures 2 and 3, Portugal has one of the highest vaccination rates in Europe. The chart of some vaccination rates also shows that Portugal started vaccination quite early compared to other countries with high rates. Specifically, Portugal started 1 month before China some of its vaccinations. The people of Portugal trusted the vaccine fairly quickly. At the same time we find that most mobility starts between November 2021 and becomes more intense in December.

References

[1] Rami Krispin https://github.com/RamiKrispin/coronavirus. R: A language and environment for statistical computing. https://cran.r-project.org/web/packages/coronavirus/coronavirus.pdf, 10 2015.