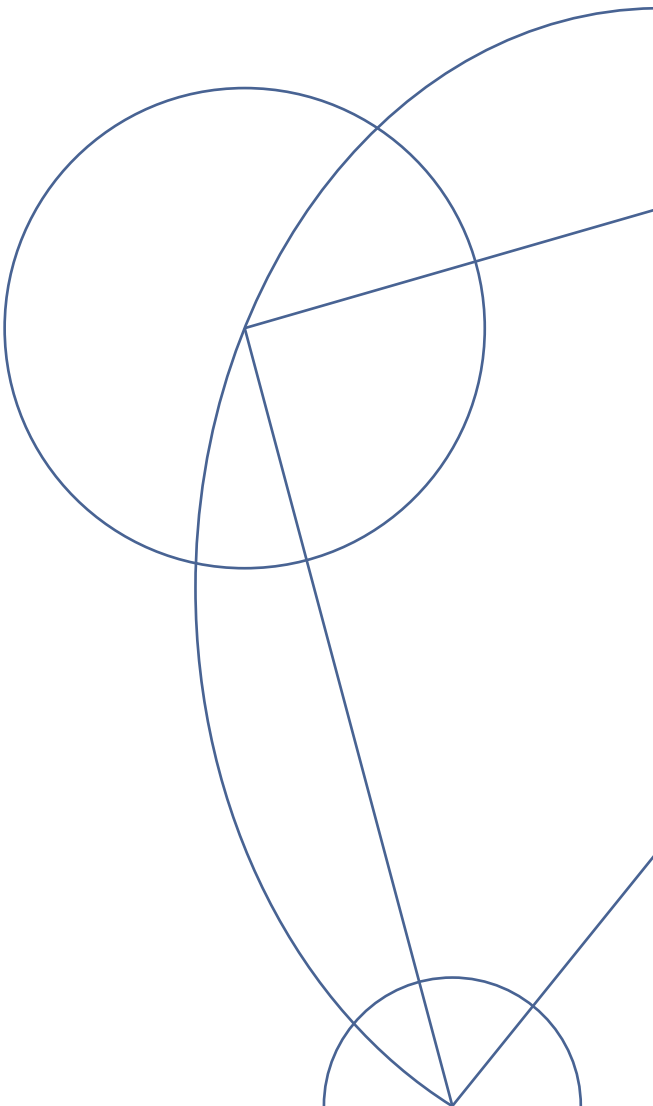
National Technical University of Athens
School of Electrical and Computer Engineering
Inter-Faculty Postgraduate Studies Programme
Data Science and Machine Learning

Exploratory Data Analysis using R: Coronavirus Vaccinations

An assignment written in partial fulfillment of the requirements for the completion of the Programming Tools and Technologies for Data Science core course of the Data Science & Machine Learning post-graduate studies programme.

Instructor
Prof. D. Fouskakis

Written by
Spyridon Rigas (MSc - 03400154)
spiridonrigas@mail.ntua.gr



January 11, 2022

Table of Contents

List of Figures

1 Introduction

The novel coronavirus (Covid-19) that was first reported at the end of 2019 has impacted almost every aspect of life as we know it. The collection and analysis of reliable data has been one of the most crucial and yet challenging tasks during the ongoing efforts to understanding the virus, as well as how effective the measures taken to combat the pandemic are. The present report is but a small contribution towards this direction: it aims to display some key findings and conclusions drawn by analyzing data concerning vaccination ratios for populations around the globe, on a global, continental, as well as regional (country) level. The studied dataset was `covid19_vaccine`, which is part of the `coronavirus` package in R [1] and contains data that is being pulled daily from the Covid-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [2].

2 Data Preprocessing

The cleaning and preprocessing procedures were a prerequisite for the analysis of the available data. For this purpose, once they were loaded using the `data.table` library, a separate data table containing world data was constructed as follows: since the total world population was unavailable, the population for each country was extracted using a custom function and afterwards the world population was defined as the sum of these extracted values. The alternative of using online data [3] was dismissed, due to the fact that the `coronavirus` package does not include data for the whole world, thus the total normalization of the data would end up being incorrect. It is important noting at this point that the last update on the dataset was performed on January 08, 2022, which means that the dataset used for the analysis contains data for dates up to and including January 07, 2022.

The next step was the introduction of two new variables: `partially_vaccinated_ratio` and `fully_vaccinated_ratio`, which correspond to the number of individuals per country to whom at least one and all required vaccination doses, respectively, have been administered, divided by the total population of said country. These two variables are the main focus of the analysis, therefore the data cleaning phase included the removal of all rows for which the entries corresponding to these variables were `NA`. All rows including `NA` values for the `population` variable were also removed, since these mainly involved world data. In this way, all information regarding Provinces/States was indirectly removed, while the world data was omitted from the main data table and was stored in a different one.

Finally, the continent names for Kosovo and Sudan were manually fixed and North and South America were merged into a single continent, named 'Americas'. After the end of this procedure, the dataset contained vaccination information for a total of 181 countries, excluding world data, none of which corresponded to `NA` or `NULL` values. The R Code developed for the aforementioned steps is presented below (R Code Snippet 1).

```
1  # Required libraries
2  library(data.table); library(ggplot2); library(coronavirus)
3
4  # Uncomment to update the coronavirus dataset
5  # Last updated: 08/01/2022
6  #update_dataset(silence = FALSE)
7
8  # Create the data.table from the vaccine dataset
9  vaccs <- setDT(copy(covid19_vaccine))
10
11 # Custom function for what follows
12 custom_max <- function(x) ifelse( !all(is.na(x)), max(x, na.rm=T), NA)
13
```

```
# calculate the total population of all countries present in the
# dataset and sum it to get the world population
world_pop <- vaccs[country_region!='World',
  list(mpop=custom_max(population)),by=country_region][,sum(mpop)]

# Update the population column for World data
vaccs[country_region=='World',population:=world_pop]

# Creates the two required new columns by updating the data.table
vaccs[, `:=`(fully_vaccinated_ratio=
  round((100*people_fully_vaccinated/population),digits=1),
  partially_vaccinated_ratio=
  round((100*people_partially_vaccinated/population),digits=1))]

# Create a new datatable only for world data
world_dt <- vaccs[country_region=='World',
  .(fully_vaccinated_ratio,partially_vaccinated_ratio),by=date]

# Clear all na values from the vaccs dt, including World data
vaccs <- na.omit(vaccs, cols=c("fully_vaccinated_ratio",
  "partially_vaccinated_ratio", "population"))

# Keep only relevant columns
vaccs <- vaccs[country_region!='World'
  & fully_vaccinated_ratio <= 100 & partially_vaccinated_ratio <= 100,
  .(country_region,people_partially_vaccinated,people_fully_vaccinated,
  lat,long,population,continent_name,fully_vaccinated_ratio,
  partially_vaccinated_ratio),by=date]

# Fix these NA values
vaccs[country_region=='Kosovo',continent_name:='Europe']
vaccs[country_region=='Sudan',continent_name:='Africa']

# At this point there are no NA values in the data table.
# it can be seen by running vaccs[is.na(any_column_name)]
# it is an empty datatable

# Also switch to "Americas" instead of North and South
vaccs[continent_name=='South America' | continent_name=='North America',
  continent_name:='Americas']
```

R Code Snippet 1: Data preprocessing.

3 Data Analysis

With the dataset cleaned and split into two data tables including country and world data, the process of data analysis began. As mentioned in the Introduction, this process was divided into three parts: analysis on a worldwide, a continental and a regional level, where the regional level mainly consisted of statistics and graphs concerning Greece. This is also the order in which the findings are presented in what follows.

3.1 Worldwide Level

The first step of the worldwide analysis was the depiction of both vaccination ratios on a spatial scale using the most recent data for all countries. To accomplish this, two heatmaps of the world were constructed using the `maps` library, one for the case of partially vaccinated population ratios (see Fig. 1) and one for the case of fully vaccinated population ratios (see Fig. 2).

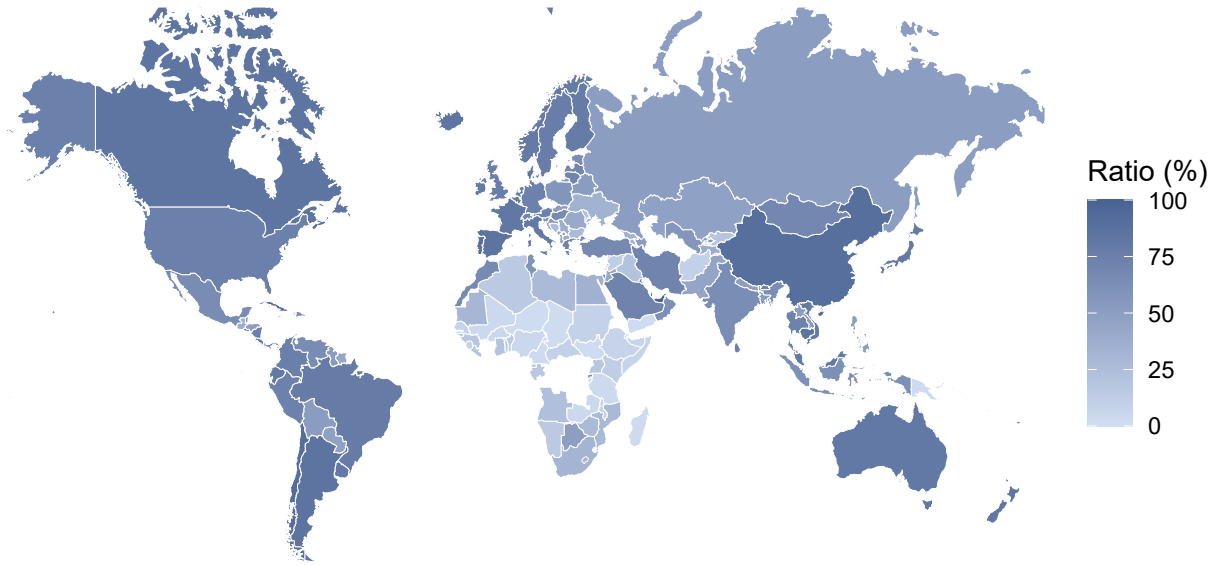


Figure 1: World heatmap corresponding to the partially vaccinated population ratio.

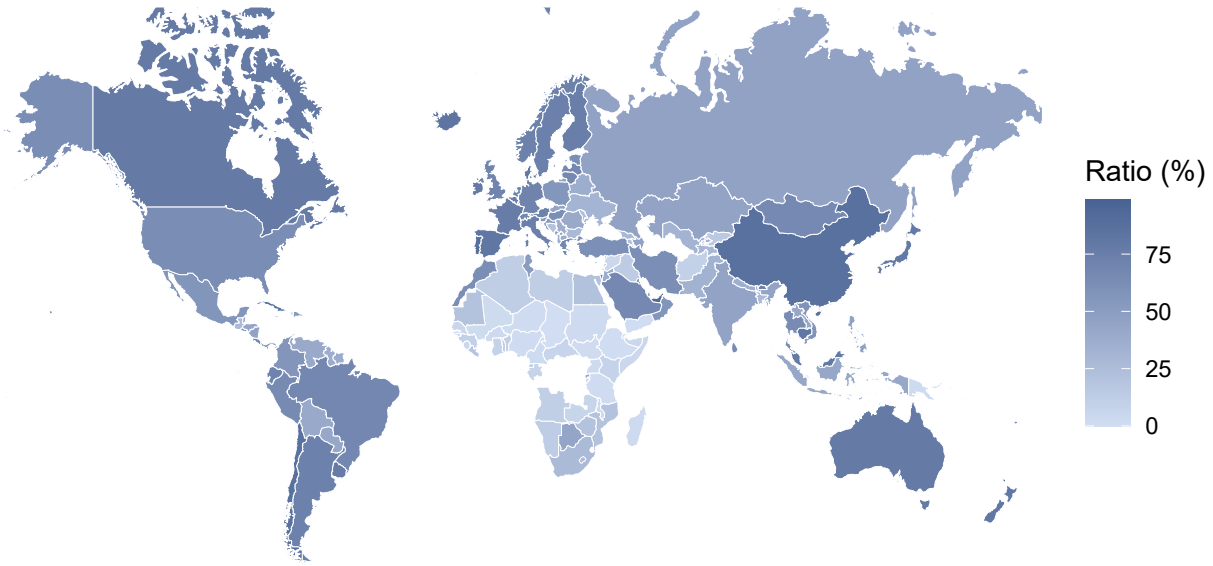


Figure 2: World heatmap corresponding to the fully vaccinated population ratio.

Comparison of the two heatmaps indicates that their relative intensities are similar, even though the heatmap of Fig. 1 includes higher absolute intensities (reaching up to 100%), which is to be expected since the second heatmap is a subset of the first (at least one dose has been administered to fully vaccinated individuals, while the inverse is not true by default). This points to the fact that there are practically no countries with large populations of individuals who refused the administering of both

vaccine doses (which would create heterogeneities between the two heatmaps). One very important observation is that a clear bias appears to exist as far as the continents are concerned, with Africa showing much lower vaccination ratios compared to the others. Closing the discussion on the two heatmaps, it is worth noting that the blank spaces correspond to countries for which there were no available data. The R Code developed for the generation of the heatmaps is presented below (R Code Snippet 2).

```
# Get info for final available date of each country
curr_dt <- vaccs[order(-date)]
curr_dt <- unique(curr_dt, by = "country_region")

library(maps)

# These are the names recognized by the library
curr_dt$country_region[curr_dt$country_region == "US"] <- "USA"
curr_dt$country_region[curr_dt$country_region == "United Kingdom"] <- "UK"

world <- map_data('world')

# Create the input for the worldmap - fully vac
wdt <- world[world$region %in% curr_dt$country_region,]
wdt$value <- curr_dt$fully_vaccinated_ratio[match(wdt$region,
  curr_dt$country_region)]

# Plot the world map
plot1 <- ggplot(wdt, aes(x=long, y=lat, group = group, fill = value)) +
  geom_polygon(colour = "white", size = 0.05) +
  theme_bw()  +
  scale_fill_continuous(low = "#d1def2", high = "#486393") +
  labs(fill = "Ratio (%)" , x="", y="") +
  scale_y_continuous(breaks=c()) +
  scale_x_continuous(breaks=c()) +
  coord_map(xlim = c(-170, 170), ylim = c(-50, 100)) +
  theme(panel.border =  element_blank())

# Create the input for the worldmap - partially vac
wdt <- world[world$region %in% curr_dt$country_region,]
wdt$value <- curr_dt$partially_vaccinated_ratio[match(wdt$region,
  curr_dt$country_region)]

# Plot the world map
plot2 <- ggplot(wdt, aes(x=long, y=lat, group = group, fill = value)) +
  geom_polygon(colour = "white", size = 0.05) + theme_bw() +
  scale_fill_continuous(low = "#d1def2", high = "#486393") +
  labs(fill = "Ratio (%)" , x="", y="") +
  scale_y_continuous(breaks=c()) + scale_x_continuous(breaks=c()) +
  coord_map(xlim = c(-170, 170), ylim = c(-50, 100)) +
  theme(panel.border =  element_blank())
```

R Code Snippet 2: World map generation.

The natural follow up to a spatial analysis was a time-series depiction of the world data, an objective for which the data table including the world data was utilized. The corresponding graph is depicted in Fig. 3, where the dashed (solid) line corresponds to partially (fully) vaccinated population ratios, while the relevant code is presented in R Code Snippet 3.

The first thing one notices in these graphs is the fact that there are small time periods for which the vaccination ratios seem to be declining, with the most prevalent one corresponding to the partially vaccinated population ratio during late November of 2021. Since these ratios are cumulative quantities,

such declines should not be present in their time-series graphs. Perhaps the most reasonable explanation is that the data for these dates contained typographic mistakes or were not pulled properly due to bugs in the API responsible for the process. However, another explanation could be the retraction of previously submitted data due to wrong reports, such as double-counting of administered doses.

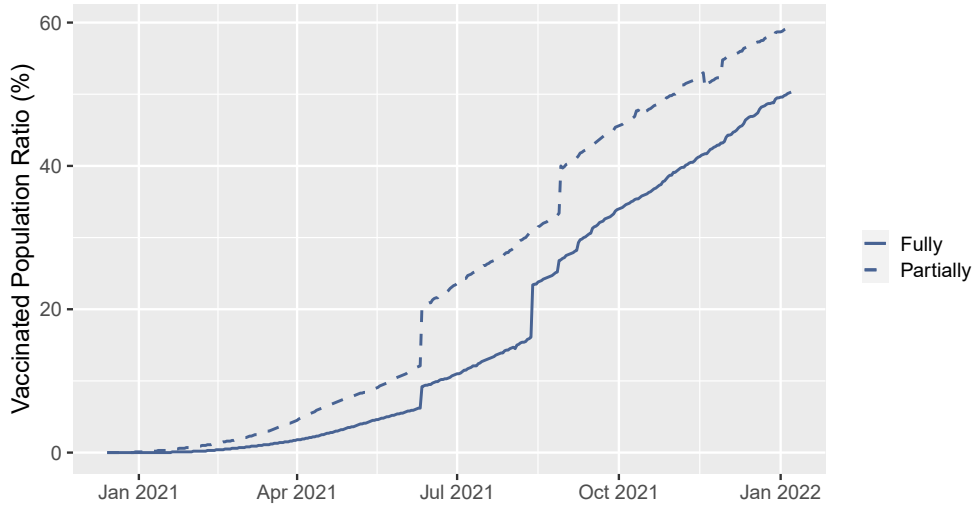


Figure 3: World data time series for the partially (solid) and fully (dashed) vaccinated population ratio.

Another interesting feature is the hysteresis effect[1] that is present in the time-series graphs, i.e. the fact that one of the two curves appears as a displacement of the other on the time axis (especially during the Summer of 2021). This can be attributed to the fact that full vaccination requires two doses (with the exception of the Janssen Ad26.COV2.S COVID-19 vaccine [5]) which cannot be administered simultaneously. The time period required for the administering of the second dose, following the first, varies from vaccine to vaccine, which implies that a fixed time period for all vaccines would enhance this hysteresis phenomenon even further.

Finally, it is important to comment on the sharp increases present in these graphs, most notably during the beginning and the end of summer. This observation can be linked with three possible facts: firstly, vaccination was a prerequisite for inter- or intra-country traveling during the Summer vacations in many countries as a safety measure to prevent the spread of the virus [6]. Secondly, due to limited availability of vaccines, as well as medical staff and resources, many countries (including Greece) opted for an age-dependent vaccination plan, which meant that different age groups couldn't receive vaccine doses from the moment the vaccines were available. Last but not least, the spikes can be due to the introduction of new countries in the `coronavirus` dataset, all ratios of which are normalized to the total population corresponding to the most recent entry (mostly January 07, 2022) of each country.

```
# World level time-series
plot3 <- ggplot(world_dt, aes(x=date)) +
  geom_line(aes(y = fully_vaccinated_ratio,
  linetype = "Fully"), size=.9, color="#486393") +
  geom_line(aes(y = partially_vaccinated_ratio,
  linetype = "Partially"), size=.9, color="#486393") +
  labs(x="",y="Vaccinated Population Ratio (%)") +
  scale_linetype_manual(name = "", values=c("Fully"="solid",
  "Partially"="dashed")) + theme_grey(base_size = 16)
```

R Code Snippet 3: Time series for world data.

[1] In Solid State Physics, hysteresis is connected to the effect of demagnetization of materials [4], however here the term is borrowed owing only to the similarities between the time-series graph and the so-called hysteresis loop.

Before moving to the continental level analysis, some basic aggregations were performed in order to determine the top-10 countries as far as the fully vaccinated population ratio is concerned and also the bottom-10 countries as far as having received at least one vaccine dose (partially vaccinated population ratio) is concerned. The results are presented in the Bar charts of Fig. 4 and Fig. 5, respectively, where the colors are assigned to the countries depending on the continent to which they belong.

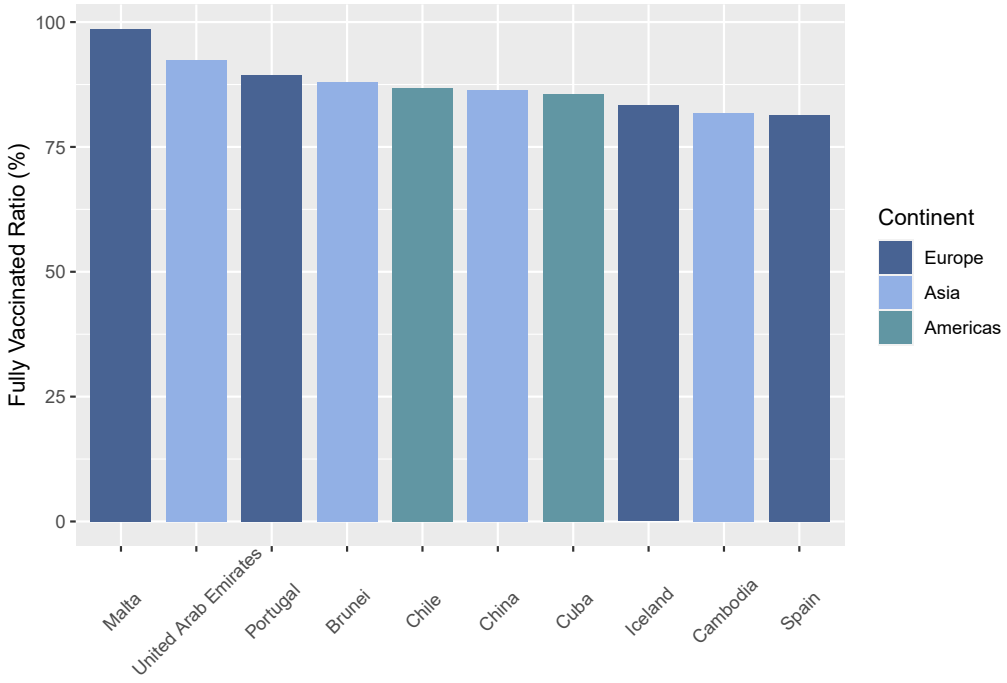


Figure 4: Bar chart of top countries by fully vaccinated population ratio (colors assigned by continent).

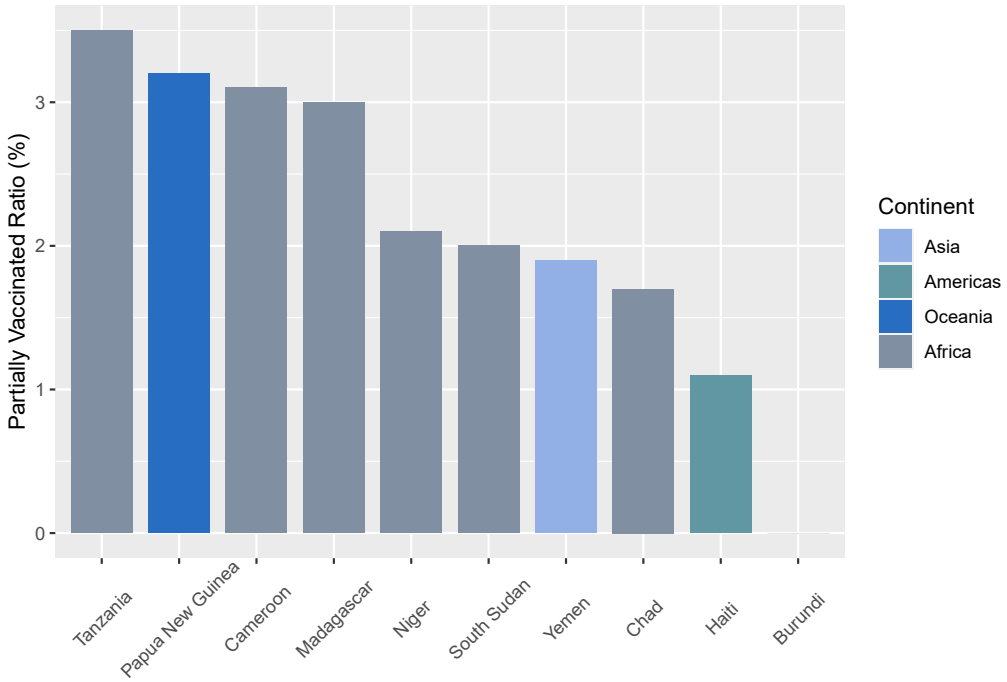


Figure 5: Bar chart of bottom countries by partially vaccinated population ratio (colors assigned by continent).

The previously mentioned bias as far as vaccinations per continent are concerned is illuminated by these bar charts as well. More specifically, 8 out of the 10 countries that have achieved the highest ratios for full vaccination are in Europe and Asia. On the contrary, 70% of the countries that are in the bottom-10 bar chart are located in Africa. While the former group of countries score as high as $\sim 90\%$ on full vaccination ratios, the latter group has a high of $\sim 3.5\%$ as far as individuals who have received at least 1 dose of a vaccine are concerned (obviously, the numbers are even lower for the ratio of the fully vaccinated population).

Another important remark is that, apart from Cuba, which demonstrated an exemplary management of the pandemic [7], or countries like Spain and Portugal, which were hit very hard during the first wave of the pandemic [8], the Bar Chart of Fig. 4 includes some of the countries with the highest GDP [9] or HDI [10] indices. On the other hand, most of the countries in the Bar Chart of Fig. 5 are poor countries [11], plagued by war [12].

These observations illustrate the highly unequal distribution of vaccines among continents (Africa versus the others), as well as the inequalities in vaccinations that occur as an immediate consequence of the lack of primary healthcare resources and medical staff in underdeveloped countries [13]. The R Code developed for the creation of the bar charts is presented below (R Code Snippet 4).

```
# Bar plots with color coding depending on continent
color_mapping <- c("Europe" = "#486393", "Asia" = "#92b0e5",
  "Americas" = "#6196a3", "Oceania" = "#276dc2", "Africa" = "#818fa3")

top_full_vac <- head(curr_dt[order(-fully_vaccinated_ratio)],n=10)
bot_par_vac <- head(curr_dt[order(partially_vaccinated_ratio)],n=10)

plot4 <- ggplot(top_full_vac, aes(x=reorder(country_region,
  -fully_vaccinated_ratio), y=fully_vaccinated_ratio,
  fill=continent_name)) + geom_bar(stat="identity", width=.8) +
  scale_fill_manual(values = color_mapping) +
  labs(fill = "Continent" , x="", y="Fully Vaccinated Ratio (%)") +
  theme(axis.text.x = element_text(angle=45, vjust = 0.6))

plot5 <- ggplot(bot_par_vac, aes(x=reorder(country_region,
  -partially_vaccinated_ratio), y=partially_vaccinated_ratio,
  fill=continent_name)) + geom_bar(stat="identity", width=.8) +
  scale_fill_manual(values = color_mapping) +
  labs(fill = "Continent" , x="", y="Partially Vaccinated Ratio (%)") +
  theme(axis.text.x = element_text(angle=45, vjust = 0.6))
```

R Code Snippet 4: Bar plots for highest and lowest country ratios.

3.2 Continental Level

In view of the high continental bias present in the distribution of vaccines, the continental level analysis of the available data began with an attempt to provide more robust, quantitative results to advocate in favor of said bias. In this direction, the exact ratio of the partially vaccinated population was calculated for each continent and presented in lollipop charts (see Fig. 6). In order to provide more depth as far as the time scale is concerned, the data were split into three seasons: the first spanning from the first day for which vaccination data were available, until April 28, 2021, the second spanning from April 29, 2021 until September 14, 2021 and the third spanning from September 15, 2021 until January 07, 2022. The aforementioned ratios were thus calculated at the end of each of these seasons, with the relevant R Code depicted below (R Code Snippet 5).

```r
# Continent total population, as judged by the curr_dt table
cont_pops <- curr_dt[,list("populations"=sum(population)),by=continent_name]

# First season results
first_season <- vaccs[date <= as.Date('2021-4-28')]
end_of_first <- unique(first_season[order(-date)], by = "country_region")
continental_first <- end_of_first[,
  list("partial_sum"=sum(people_partially_vaccinated)), by = .(continent_name)]
results_first <- merge(continental_first, cont_pops, all=FALSE)
results_first$partial_ratio =
  round(100*results_first$partial_sum/results_first$populations, digits=1)

# Second season results
second_season <- vaccs[date > as.Date('2021-4-28') &
  date <= as.Date('2021-9-14')]
end_of_second <- unique(second_season[order(-date)], by = "country_region")
continental_second <- end_of_second[,
  list("partial_sum"=sum(people_partially_vaccinated)), by = .(continent_name)]
results_second <- merge(continental_second, cont_pops, all=FALSE)
results_second$partial_ratio =
  round(100*results_second$partial_sum/results_second$populations, digits=1)

# Third season results
third_season <- vaccs[date > as.Date('2021-9-14')]
end_of_third <- unique(third_season[order(-date)], by = "country_region")
continental_third <- end_of_third[,
  list("partial_sum"=sum(people_partially_vaccinated)), by = .(continent_name)]
results_third <- merge(continental_third, cont_pops, all=FALSE)
results_third$partial_ratio =
  round(100*results_third$partial_sum/results_third$populations, digits=1)

library(gridExtra)

# Lollipop Charts
plot6 <- ggplot(results_first, aes(x=continent_name, y=partial_ratio)) +
  geom_point(size=3, color="#486393") +
  geom_segment(aes(x=continent_name, xend=continent_name,
  y=0, yend=partial_ratio), color="#486393") + ylim(0, 75) +
  labs(x="First Season",y="Partially Vaccinated Population Ratio (%)") +
  theme(axis.text.x = element_text(angle=45, vjust=0.6))

plot7 <- ggplot(results_second, aes(x=continent_name, y=partial_ratio)) +
  geom_point(size=3, color="#486393") +
  geom_segment(aes(x=continent_name, xend=continent_name,
  y=0, yend=partial_ratio), color="#486393") +
  labs(x="Second Season",y="") + ylim(0, 75) +
  theme(axis.text.x = element_text(angle=45, vjust=0.6))

plot8 <- ggplot(results_third, aes(x=continent_name, y=partial_ratio)) +
  geom_point(size=3, color="#486393") +
  geom_segment(aes(x=continent_name, xend=continent_name,
  y=0, yend=partial_ratio), color="#486393") +
  labs(x="Third Season",y="") + ylim(0, 75) +
  theme(axis.text.x = element_text(angle=45, vjust=0.6))

grid.arrange(plot6,plot7,plot8, nrow = 1, top='')
```

R Code Snippet 5: Lollipop charts for continents per season.

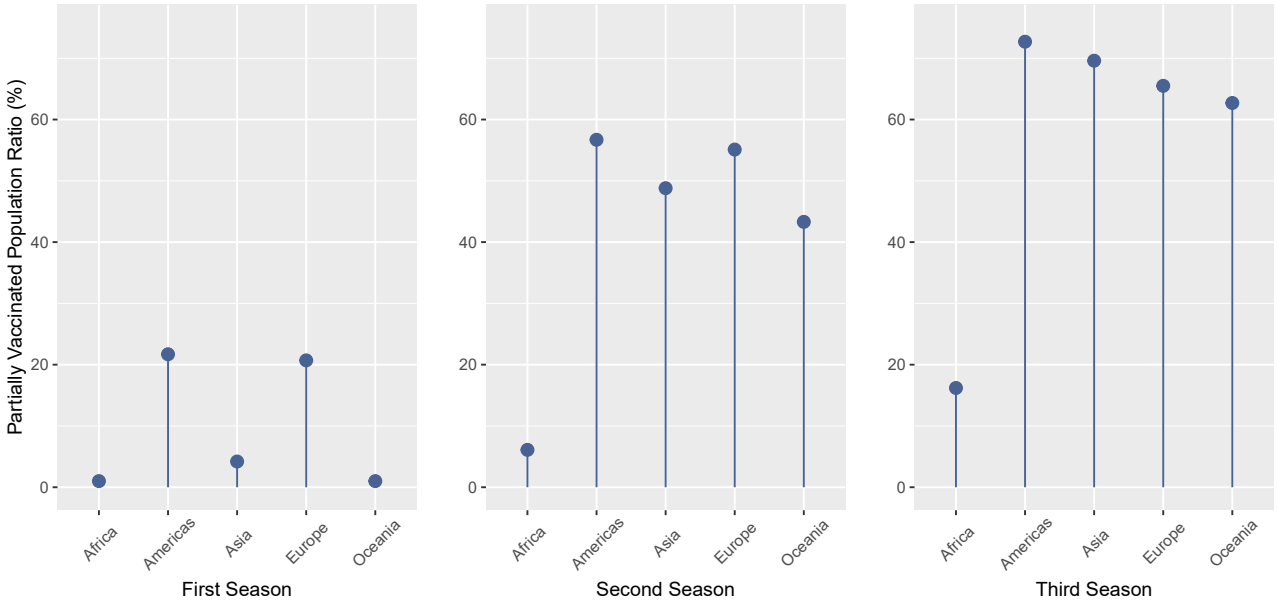


Figure 6: Lollipop charts depicting the total ratios for partially vaccinated populations on the continent level for three different time seasons.

Evidently, America and Europe surpass the 20% mark from as early as the first season, while Africa has not managed to do so by the end of the third season, when all other continents have achieved partial vaccination ratios higher than 60%. It is also worth noting that, even though Asia had a very low vaccination ratio at the end of the first season[2], it managed to become the second most vaccinated continent (as far as the partial ratio is concerned) by the end of the third season. Finally, it appears that the highest vaccination boost happened during the second season, since the relative increase from season two to season three is not as high as the one from season one to season two (with the exception of Africa, which hopefully means that the aforementioned continental bias may be somewhat reduced in the near future).

Figure 7: Europe heatmap corresponding to the partially vaccinated population ratio.

[2] Note that this may be due to the fact that Russia, a country that developed a vaccine relatively early and achieved good vaccination ratios, is considered as a European country for the purposes of the present analysis.

As part of the continental level analysis, a heatmap similar to the ones shown in Figs. 1 and 2 was constructed, this time focusing exclusively on Europe (see Fig. 7). The relevant R code is omitted, since it is almost identical to the one presented in R Code Snippets 1 and 2. Since Figures 1 and 2 indicated that the relative ratios do not differ qualitatively from country to country on large scales, Fig. 7 only depicts the partially vaccinated population ratios for European countries. Even though Europe was confirmed to be a continent with relatively high vaccination ratios, these ratios are by no means uniform across European countries. The above heatmap indicates that, while the countries in Western Europe show partial vaccination ratios as high as $\sim$ 90%, the mean ratio for countries in the Balkan region, as well as Eastern Europe, is far lower. Thankfully, Greece appears to be an exception to this rule of thumb.

3.3 Regional Level: Greece

Picking up the thread from the continental level analysis' final remarks, the final stage of the analysis focused on Greece. By developing the code shown below (R Code Snippet 6), a time-series graph was constructed for the vaccination data corresponding to Greece. This graph is depicted in Fig. 8, alongside the corresponding graphs for the world and European time series, which act as a reference for where Greece stands in a global and European scale, respectively.

```
# Time-series for Greece, with World and Europe for comparison
greece_dt <- vaccs[country_region=='Greece',list("date"=date,
  "fully_gr"=fully_vaccinated_ratio,
  "partially_gr"=partially_vaccinated_ratio)]

europe_ts <- vaccs[continent_name=='Europe',
  list("part_eu"=sum(people_partially_vaccinated),
  "full_eu"=sum(people_fully_vaccinated)), by = .(date)]

# Create ratios for Europe
eur_pop <- cont_pops[continent_name=='Europe']$populations
europe_ts$fully_eu <- round(100*europe_ts$full_eu/eur_pop,digits=1)
europe_ts$partially_eu <- round(100*europe_ts$part_eu/eur_pop,digits=1)

active_dt <- merge(greece_dt,europe_ts,by="date")
active_dt <- merge(active_dt,world_dt,by="date")

# Check for bad values to avoid what happened with World series
for (i in 1:(nrow(active_dt)-1)) {
  if (any(active_dt[i+1]-active_dt[i] < 0)) {
      active_dt[i+1] <- active_dt[i]
  }
}

plot10 <- ggplot(active_dt, aes(x=date)) +
  geom_line(aes(y = fully_gr, color = "Greece", linetype="Fully"), size=.9) +
  geom_line(aes(y = fully_eu, color = "Europe", linetype="Fully"), size=.9) +
  geom_line(aes(y = fully_vaccinated_ratio, color = "World", linetype="Fully"),
  size=.9) + geom_line(aes(y = partially_gr, color = "Greece",
  linetype="Partially"), size=.9) + geom_line(aes(y = partially_eu,
  color = "Europe", linetype="Partially"), size=.9) +
  geom_line(aes(y = partially_vaccinated_ratio, color = "World",
  linetype="Partially"), size=.9) +
  labs(x="",y="Vaccinated Population Ratio (%)") +
  scale_color_manual(name = "",
  values = c("Greece" = "#486393", "Europe" = "#6196a3",
  "World" = "#818fa3")) + scale_linetype_manual(name = "",
```

```
  values=c("Fully"="solid", "Partially"="dashed")) + ylim(0, 75) +
  theme_grey(base_size = 16)
```

R Code Snippet 6: Time series for Greece's data.

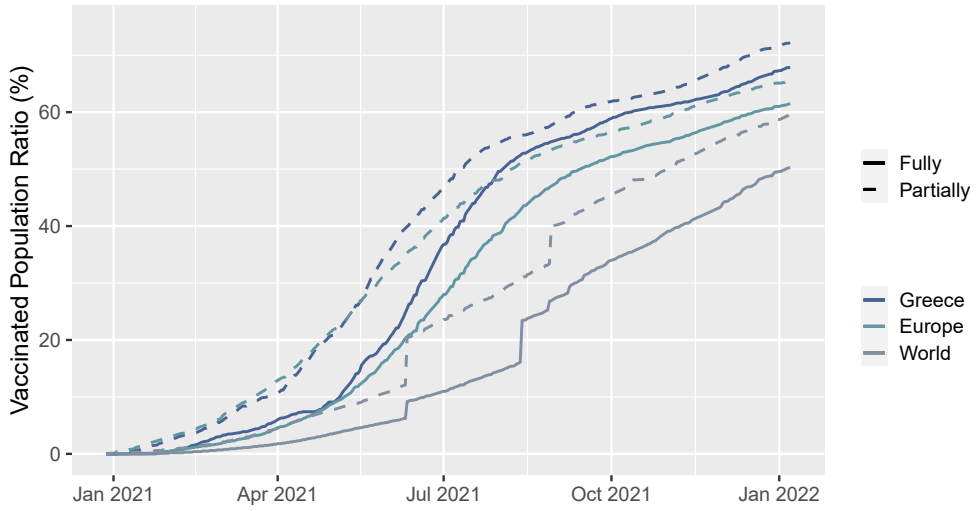


Figure 8: Greek data time series for the partially (solid) and fully (dashed) vaccinated population ratio. World and European data provided for reference.

In contrast to the graph of Fig. 3, the graphs shown in 8 do not show any declines. This is due to the fact that, after observing this behaviour in the time series for the world data, the data table was further processed: if any row contained ratios that were lower compared to the previous one, these ratios were set equal to the previous day's ratios (as if no data had been retrieved for this day)[3].

An interesting observation is that the hysteresis loop observed in the graph of Fig. 3 is even more clear in the time series for Greece. However, much more importantly, Greece appears to have vaccination ratios that are higher compared to the mean ratios for the world or for Europe. In fact, this has been the case ever since the beginning of the vaccination process. Furthermore, as of July 2021, Greece's fully vaccinated population ratio managed to surpass the mean European partially vaccinated population ratio (as far as the mean world partially vaccinated population ratio is concerned, this has been the case ever since the beginning of the vaccination process as well).

The next part of the regional level analysis was a more direct comparison of Greece with other countries. Some simple aggregations showed that on a global scale Greece ranks at 52/181 for partially vaccinated population ratio and at 44/181 for fully vaccinated population ratio. On the European scale, Greece ranks at 19/45 for the partially vaccinated population ratio and 18/45 for the fully vaccinated population ratio. In order to get a better understanding on how Greece ranks within Europe, which is closely related to the primary observations made based on Fig. 7, diverging-bars plots were generated centered around Greece, through the code shown below (R Code Snippet 7).

```
# Greece's World level ranking
tot_ct <- curr_dt[,.N]

rank_part <- curr_dt[order(
  -partially_vaccinated_ratio)][country_region=='Greece',which=TRUE]

rank_full <- curr_dt[order(
```

[3] The reason why this wasn't done during the preprocessing stage is that we wanted to illustrate this flaw of the dataset within this report. Thanks to this observation, a relevant issue has been raised in the package's GitHub page.

```
  -fully_vaccinated_ratio)][country_region=='Greece',which=TRUE]

# Greece's European level ranking
europe_curr <- curr_dt[continent_name=='Europe']
tot_ct <- europe_curr[,.N]

rank_part <- europe_curr[order(
  -partially_vaccinated_ratio)][country_region=='Greece',which=TRUE]

rank_full <- europe_curr[order(
  -fully_vaccinated_ratio)][country_region=='Greece',which=TRUE]

# Diverging plots for Greece compared to other european countries
gr_full <- europe_curr[country_region=='Greece']$fully_vaccinated_ratio
europe_curr$fully <- round((europe_curr$fully_vaccinated_ratio -
  gr_full)/gr_full,digits=1)
europe_curr$fully_type <- ifelse(europe_curr$fully < 0, "below", "above")

active_dt <- europe_curr[order(fully)]
active_dt$country_region <- factor(active_dt$country_region,
  levels = active_dt$country_region)

plot11 <- ggplot(active_dt, aes(x=country_region, y=fully, label=fully)) +
  geom_bar(stat='identity', aes(fill=fully_type), width=.5) +
  scale_fill_manual(name="Percentage (%) of",
  labels = c("Higher Ratio", "Lower Ratio"),
  values = c("above"="#486393", "below"="#6196a3")) +
  labs(x="",y="(b)") + coord_flip() + theme_grey(base_size = 16)

gr_partial <- europe_curr[country_region=='Greece']$partially_vaccinated_ratio
europe_curr$partially <- round((europe_curr$partially_vaccinated_ratio -
  gr_partial)/gr_partial,digits=1)
europe_curr$partially_type <- ifelse(europe_curr$partially < 0,
  "below", "above")

active_dt <- europe_curr[order(partially)]
active_dt$country_region <- factor(active_dt$country_region,
  levels = active_dt$country_region)

plot12 <- ggplot(active_dt, aes(x=country_region, y=partially,
  label=partially)) + geom_bar(stat='identity',
  aes(fill=partially_type), width=.5) + coord_flip() +
  scale_fill_manual(name="Percentage (%) of",
  labels = c("Higher Ratio", "Lower Ratio"),
  values = c("above"="#486393", "below"="#6196a3")) +
  labs(x="",y="(a)") + theme_grey(base_size = 16)
```

R Code Snippet 7: Diverging-bars plots for Europe centered around Greece.

It is highlighted at this point that the diverging-bars plots generated by this code and shown in Fig. 9 do not correspond to usual diverging-bars plots, where the divergence is estimated with respect to standardized parameters [14]. Here, the divergence of country A from Greece is calculated as

$$\alpha = \frac{\text{ratio}\,(\text{A}) - \text{ratio}\,(\text{Greece})}{\text{ratio}\,(\text{Greece})},$$

which means that α is simply the percentage by which the ratio of country A is better/worse compared to Greece's. In other words, the fact that $\alpha = 0.5$ for the fully vaccinated population ratio of Malta

means that Malta has a 50% higher ratio compared to Greece's, while the fact that $\alpha = -0.7$ for the partially vaccinated population ratio of Moldova means that Moldova has a 70% lower ratio compared to Greece's.

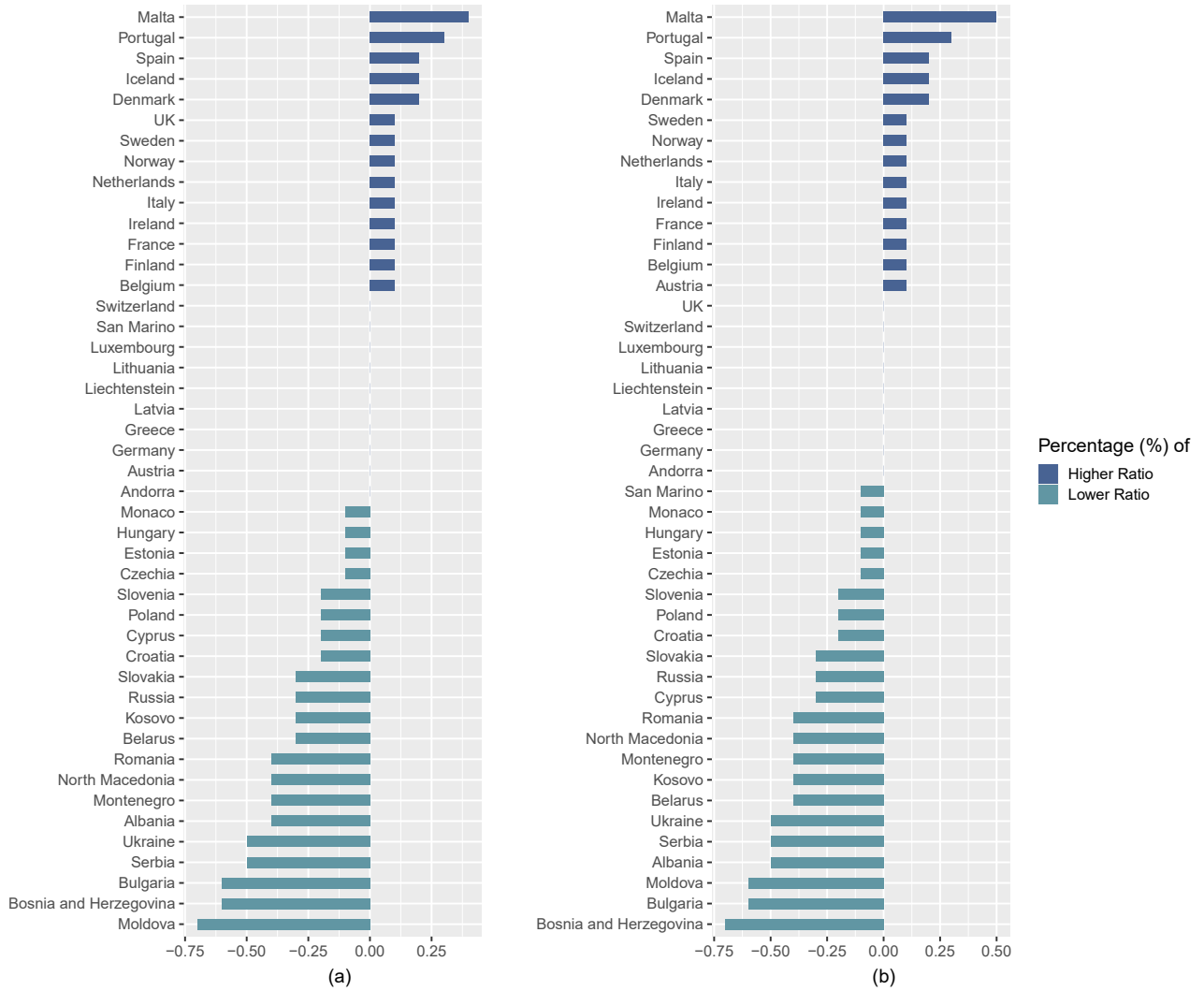


Figure 9: Diverging-bars plots for (a) partially and (b) fully vaccinated population ratios with respect to Greece.

The results of Fig. 9 confirm the observation made based on the European heatmap of Fig. 7: there is a relatively large inhomogeneity with regards to the vaccination ratios across Europe, with Western Europe scoring much higher compared to Eastern Europe and the Balkan region.

The final part of the regional level analysis and by extension the entirety of the present analysis was once again a comparison of Greece with other European countries. This time, however, the comparison was made with European countries that had up to 10% higher or lower total population compared to Greece. The fact that the main parameters of interest for this analysis have been `partially_vaccinated_ratio` and `fully_vaccinated_ratio` has silently removed the population parameter from various discussions. However, a country's population is one of the first parameters that are taken into account in numerous comparisons. For this reason, the R code shown below (R Code Snippet 8) was developed in order to identify the European countries that fall under the 10% criterion and create the corresponding graphs.

```
# Compare with other European countries of similar population (10% margin)
greek_ppl <- curr_dt$population[curr_dt$country_region=='Greece']
```

```
3
4  sim_ct <- vaccs[continent_name == 'Europe' & population < 1.1*greek_ppl &
5    population > 0.9*greek_ppl]
6
7  print(unique(sim_ct$country_region))
8
9  similar_ppl <- sim_ct[country_region=='Greece',list(date=date,
10   full_gr=fully_vaccinated_ratio,part_gr=partially_vaccinated_ratio)]
11 extra_ct <- sim_ct[country_region=='Portugal',list(date=date,
12   full_pr=fully_vaccinated_ratio,part_pr=partially_vaccinated_ratio)]
13
14 similar_ppl <- merge(similar_ppl,extra_ct,by="date")
15 extra_ct <- sim_ct[country_region=='Sweden',list(date=date,
16   full_se=fully_vaccinated_ratio,part_se=partially_vaccinated_ratio)]
17
18 similar_ppl <- merge(similar_ppl,extra_ct,by="date")
19 extra_ct <- sim_ct[country_region=='Belarus',list(date=date,
20   full_be=fully_vaccinated_ratio,part_be=partially_vaccinated_ratio)]
21
22 similar_ppl <- merge(similar_ppl,extra_ct,by="date")
23 extra_ct <- sim_ct[country_region=='Czechia',list(date=date,
24   full_cz=fully_vaccinated_ratio,part_cz=partially_vaccinated_ratio)]
25
26 similar_ppl <- merge(similar_ppl,extra_ct,by="date")
27 extra_ct <- sim_ct[country_region=='Hungary',list(date=date,
28   full_hg=fully_vaccinated_ratio,part_hg=partially_vaccinated_ratio)]
29
30 similar_ppl <- merge(similar_ppl,extra_ct,by="date")
31
32 # Check for bad values to avoid what happened with World series
33 for (i in 1:(nrow(similar_ppl)-1)) {
34   if (any(similar_ppl[i+1]-similar_ppl[i] < 0)) {
35       similar_ppl[i+1] <- similar_ppl[i]
36   }
37 }
38
39 cmap <- c("Greece" = "#486393", "Portugal" = "#276dc2",
40   "Sweden" = "#6196a3", "Belarus" = "#818fa3",
41   "Czechia" = "#225a63", "Hungary" = "#92b0e5")
42
43 plot13 <- ggplot(similar_ppl, aes(x=date)) +
44   geom_line(aes(y = part_gr), size=.9, color = "#486393") +
45   geom_line(aes(y = part_pr), size=.9, color = "#276dc2") +
46   geom_line(aes(y = part_se), size=.9, color = "#6196a3") +
47   geom_line(aes(y = part_be), size=.9, color = "#818fa3") +
48   geom_line(aes(y = part_cz), size=.9, color = "#225a63") +
49   geom_line(aes(y = part_hg), size=.9, color = "#92b0e5") +
50   labs(x="",y="Partially Vaccinated Population Ratio (%)") +
51   theme_grey(base_size = 16)
52
53 plot14 <- ggplot(similar_ppl, aes(x=date)) +
54   geom_line(aes(y = full_gr, color = "Greece"), size=.9) +
55   geom_line(aes(y = full_pr, color = "Portugal"), size=.9) +
56   geom_line(aes(y = full_se, color = "Sweden"), size=.9) +
57   geom_line(aes(y = full_be, color = "Belarus"), size=.9) +
58   geom_line(aes(y = full_cz, color = "Czechia"), size=.9) +
59   geom_line(aes(y = full_hg, color = "Hungary"), size=.9) +
60   labs(x="",y="Fully Vaccinated Population Ratio (%)") +
```

```
  scale_color_manual(name = "", values = cmap) +
  theme_grey(base_size = 16)
```

R Code Snippet 8: Countries with ±10% population compared to Greece.

Portugal, Sweden, Belarus, Czechia and Hungary were identified as the 5 European countries with a population within the 10% margin compared to Greece's population. The time-series graphs depicting all 6 countries can be seen in Fig. 10, with the partially vaccinated population ratio on the left (a) and the fully vaccinated population ratio on the right (b).

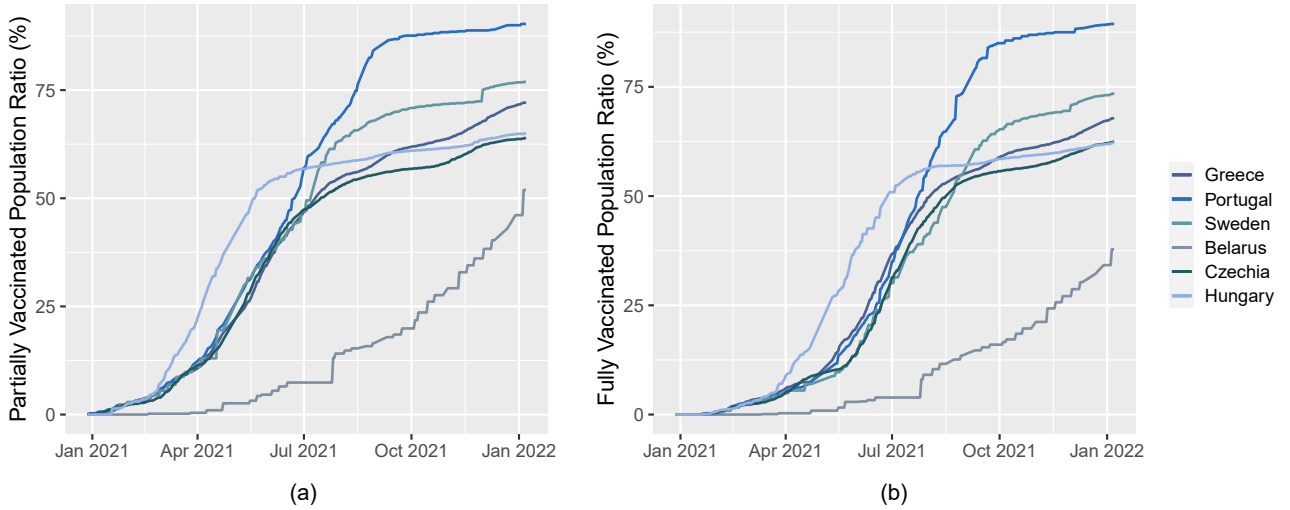


Figure 10: Time series for (a) the partially and (b) the fully vaccinated population ratio of Greece and European countries with populations in the range ±10% compared to Greece.

Excluding Belarus, for which data were not retrieved on a regular basis (as indicated by the fact that the corresponding time-series curve does not appear as smooth as the ones for the other countries), Greece's ratios are neither on top, neither at the bottom, compared to the other 4 countries. In any case, Greece's ratios are close to Sweden's, Czechia's and Hungary's in both cases.

4 Remarks & Outlook

Summarizing the main conclusions drawn from the present analysis, both vaccination ratios on a global scale have fortunately surpassed the 50% mark, however there is still room for improvement. Unfortunately, a high bias appears to exist when it comes to the continental level, since Africa has not even managed to surpass the 20% mark for either vaccination ratio, while all other continents score higher than 60%. On the European scale, despite Europe having good mean scores, there is a high inhomogeneity across its countries. Greece stands somewhere in the middle, with most Western European countries scoring higher and most Eastern European countries and Balkan states scoring lower. Finally, in all time-series plots a hysteresis effect can be seen, which can be attributed to the time interval required between consecutive vaccinations.

Closing the present analysis, it is noted that since the pandemic appears to be far from over, the existing data need to be continuously re-evaluated, while newer data are simultaneously collected. In light of the Omicron variant, the vaccination ratios might change significantly, so the development of a machine learning model to predict future vaccination ratios per country might be a good direction for future analysis. Another such direction, as a continuation of the present analysis, might be the inclusion of booster vaccine doses in the dataset and consequent analysis thereof.

References

[1] R. Krispin, "Coronavirus," *GitHub repository*, 2020.

[2] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," *Lancet Inf Dis.*, vol. 20, no. 5, pp. 533–534, 2020. DOI: 10.1016/S1473-3099(20)30120-1.

[3] *World population data*. [Online]. Available: https://www.worldometers.info/world-population/.

[4] *Magnetic hysteresis*. [Online]. Available: https://www.nde-ed.org/Physics/Magnetism/Demagnetization.xhtml.

[5] *Vaccine by johnson & johnson*. [Online]. Available: https://www.who.int/news-room/feature-stories/detail/the-j-j-covid-19-vaccine-what-you-need-to-know.

[6] *Travelling during the summer of 2021*. [Online]. Available: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2021/6/q-and-a-on-vaccination-and-travel-this-summer.

[7] *Cuba better placed than many nations to fight pandemic*. [Online]. Available: https://www.aa.com.tr/en/americas/cuba-better-placed-than-many-nations-to-fight-pandemic/1824197.

[8] *Vaccine champions spain, portugal focus on the reluctant few*. [Online]. Available: https://www.usnews.com/news/health-news/articles/2021-12-01/portugal-tightens-restrictions-despite-virus-vaccine-success.

[9] *Gdp per country*. [Online]. Available: https://www.worldometers.info/gdp/gdp-by-country/.

[10] *Ranking of countries with the highest human development index (hdi) value in 2019*. [Online]. Available: https://www.statista.com/statistics/264630/countries-with-the-highest-human-development-index-ranking/.

[11] *Poorest countries in the world 2021*. [Online]. Available: https://worldpopulationreview.com/country-rankings/poorest-countries-in-the-world.

[12] *Yemeni civil war*. [Online]. Available: https://en.wikipedia.org/wiki/Yemeni_Civil_War_(2014%E2%80%93present).

[13] *Unequal vaccine distribution self-defeating, world health organization chief tells economic and social council's special ministerial meeting*. [Online]. Available: https://www.un.org/press/en/2021/ecosoc7039.doc.htm.

[14] *Common ggplot visualizations*. [Online]. Available: http://www.math.ntua.gr/~fouskakis/Programming_R/Slides/5.pdf.