Programming for Data Science - R assignment

Konstantinos Papadakis
Student of MSc Data Science and Machine Learning (03400149)
k.i.papadakis@gmail.com

26 January 2022

Abstract

In this document we explore the progress of vaccinations against COVID19 around the world, from 14 December 2020 to 23 January 2022.

Data overview

We begin by loading our data.

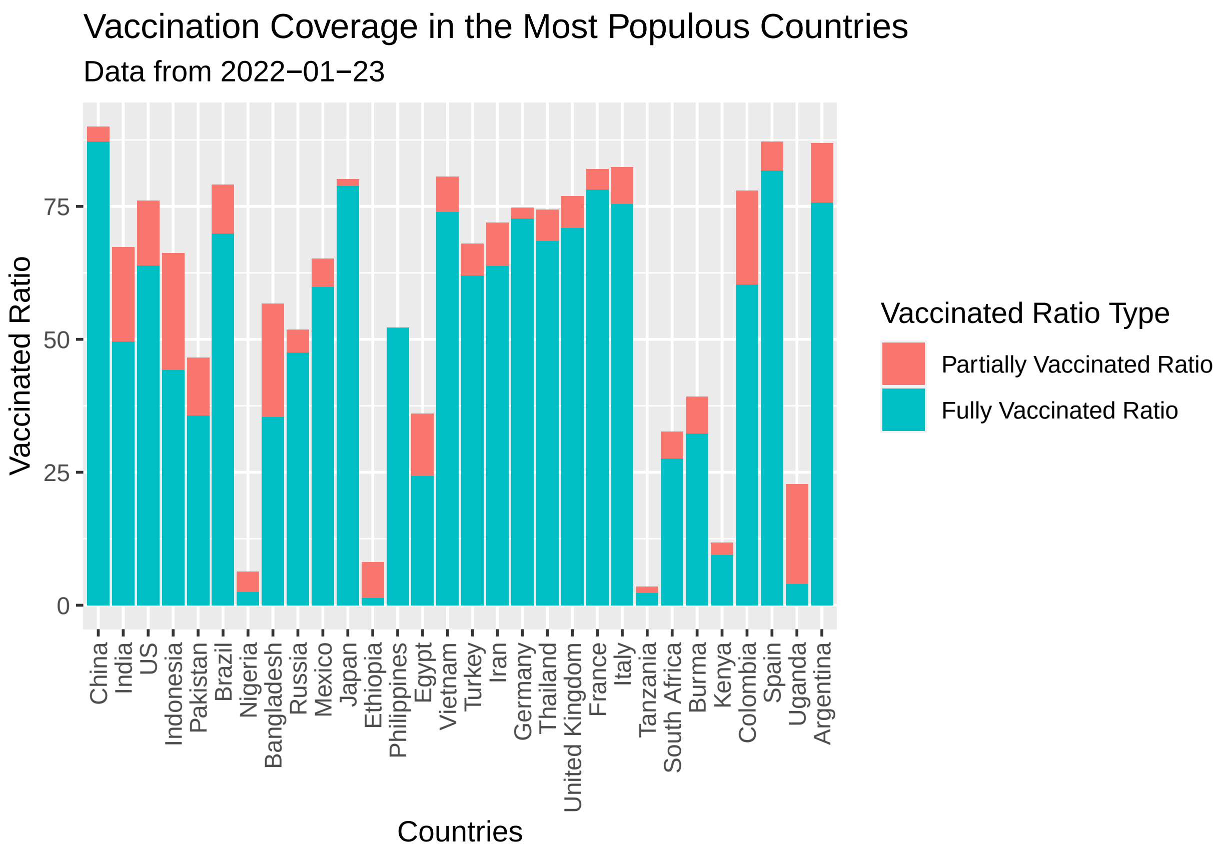
```r
library(data.table)
library(ggplot2)
library(ggrepel)
library(maps)

Sys.setlocale("LC_ALL", "English")

# Load the data
coronavirus::update_dataset(silence = TRUE)
vax <- as.data.table(coronavirus::covid19_vaccine)
# Drop the provinces
vax <- vax[is.na(vax$province), ]
# Add vaccination ratios percentages
vax$fully_vaccinated_ratio <- round(vax$people_fully_vaccinated / vax$population,
                                    digits = 4) * 100
vax$partially_vaccinated_ratio <- round(vax$people_partially_vaccinated / vax$population,
                                        digits = 4) * 100
```

We will keep a slice of our data from the most recent day (2022-01-23), and we will use it as our main source for time-independent results.

```r
# vax_now<- vax[, .SD[which.max(.SD$date)], by = country_region]
# length(unique(vax_now$date)) == 1  returns TRUE, but it is possible to return FALSE.
# Thus, we will find the most recent day over the entire dataset,
# and pick the corresponding entries.
vax_now <- vax[date == max(date)]
```

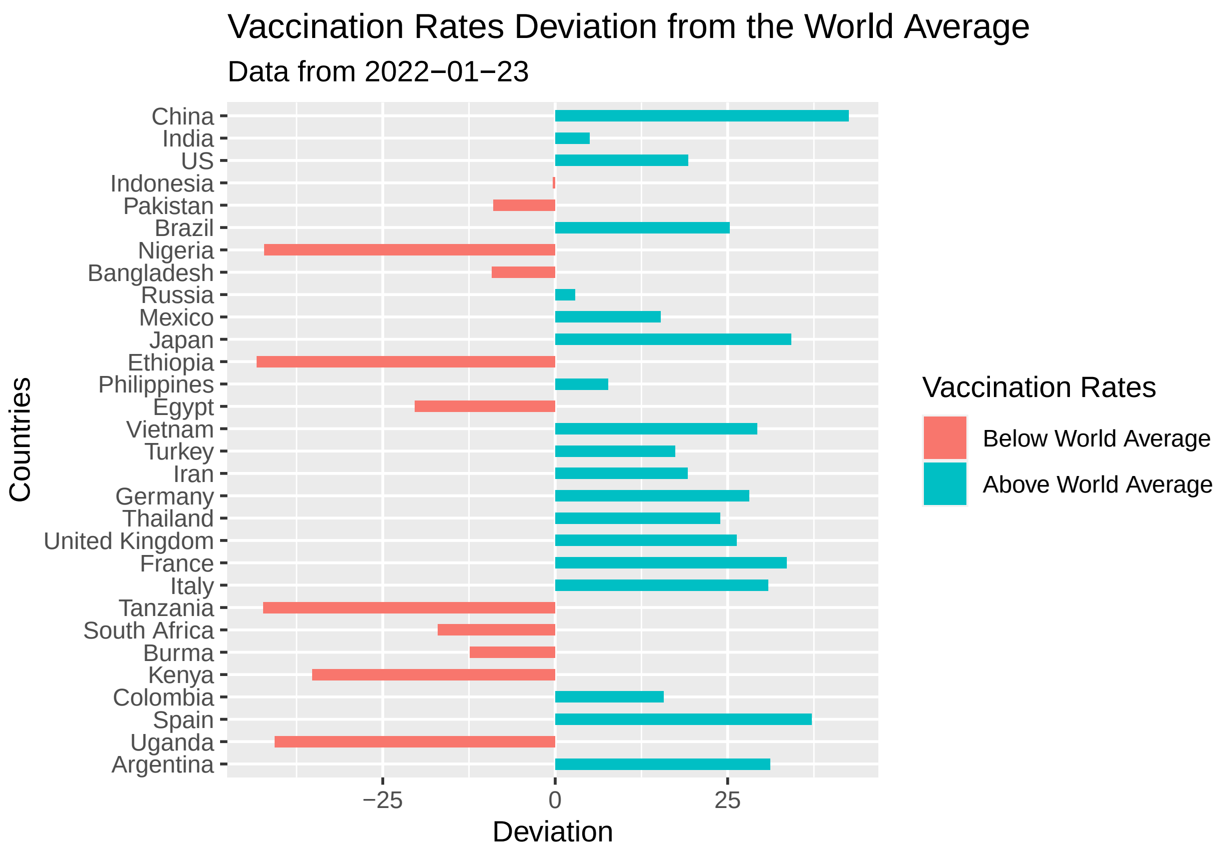
We proceed to plot the Vaccination rates among the world's 30 most populous countries. The results show that there is tremendous inequality in vaccinations between countries. Poorer countries cannot afford the vaccines and the richer countries are not helping them enough.

```r
# Plot Partially and Fully Vaccinated rates for the most populous countries
t <- na.omit(vax_now, cols = c("partially_vaccinated_ratio", "fully_vaccinated_ratio"))
t <- t[order(-population)[1:30],
       .(country_region, population, partially_vaccinated_ratio, fully_vaccinated_ratio)]
# Melt the partially_vaccinated_ratio and the fully_vaccinated_ratio,
# so that we can plot them together easier.
t <- melt(t, measure.vars = c("partially_vaccinated_ratio", "fully_vaccinated_ratio"),
          variable.name = "vaccinated_ratio_type", value.name = "vaccinated_ratio")
ggplot(data = t,
       aes(x = reorder(country_region, -population), y = vaccinated_ratio,
           fill = vaccinated_ratio_type)) +
  geom_bar(stat = "identity", position = "identity") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = "Vaccination Coverage in the Most Populous Countries",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "Countries", y = "Vaccinated Ratio") +
  scale_fill_discrete(name = "Vaccinated Ratio Type",
                      labels = c("Partially Vaccinated Ratio", "Fully Vaccinated Ratio"))
```



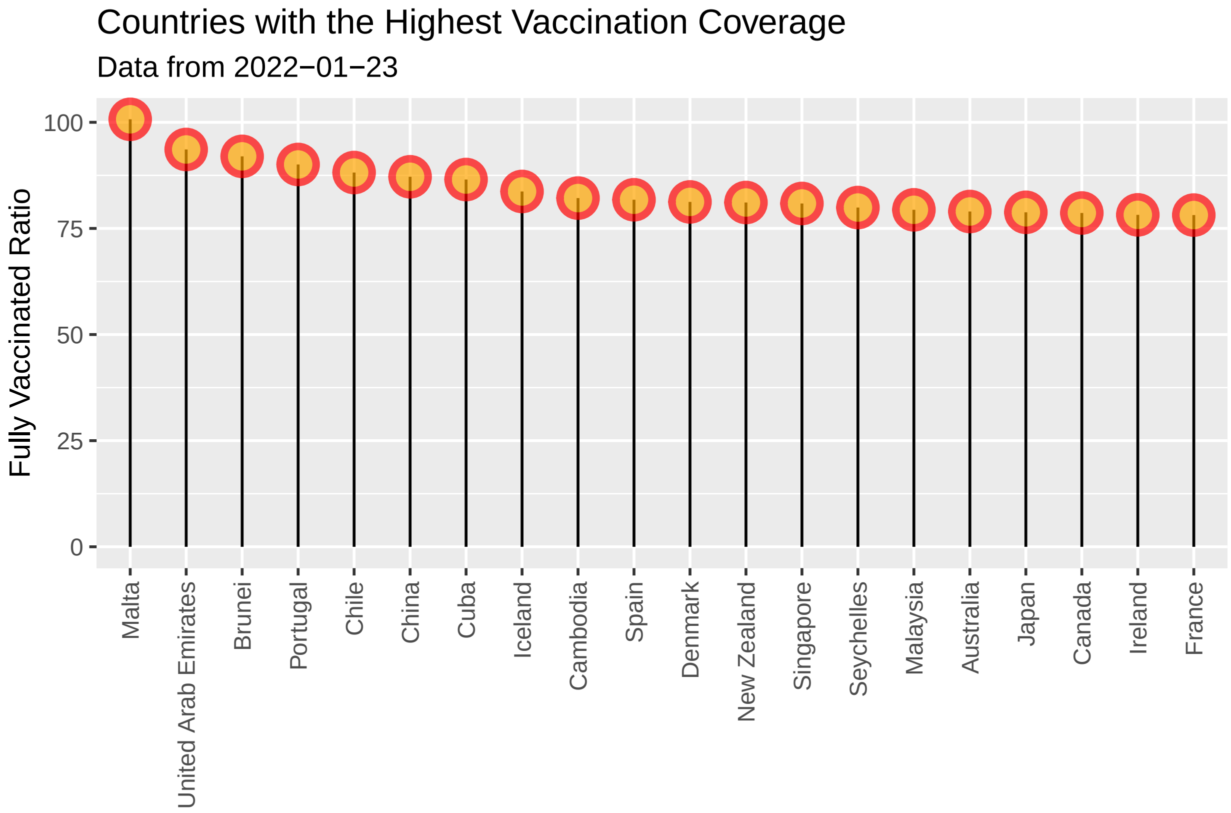
And we can also compare them to the world average.

```
# Deviation from the World Average Histogram for the 30 most populous countries
mean_partial <- vax_now[, mean(fully_vaccinated_ratio, na.rm = TRUE)]
t <- na.omit(vax_now, cols = c("fully_vaccinated_ratio", "fully_vaccinated_ratio"))
t$above_below <- t$fully_vaccinated_ratio > mean_partial
t$deviation <- t$fully_vaccinated_ratio - mean_partial

ggplot(data = t[order(-population)[1:30]],
       aes(x = reorder(country_region, population), y = deviation, fill = above_below)) +
  geom_bar(stat = 'identity', width = .5) +
  scale_fill_discrete(name = "Vaccination Rates",
                    labels = c("Below World Average", "Above World Average")) +
  labs(title = "Vaccination Rates Deviation from the World Average",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "Countries",
       y = "Deviation"
  ) +
  coord_flip()
```



Below we can see the 20 countries with the highest vaccination rates. All of them are developed countries.

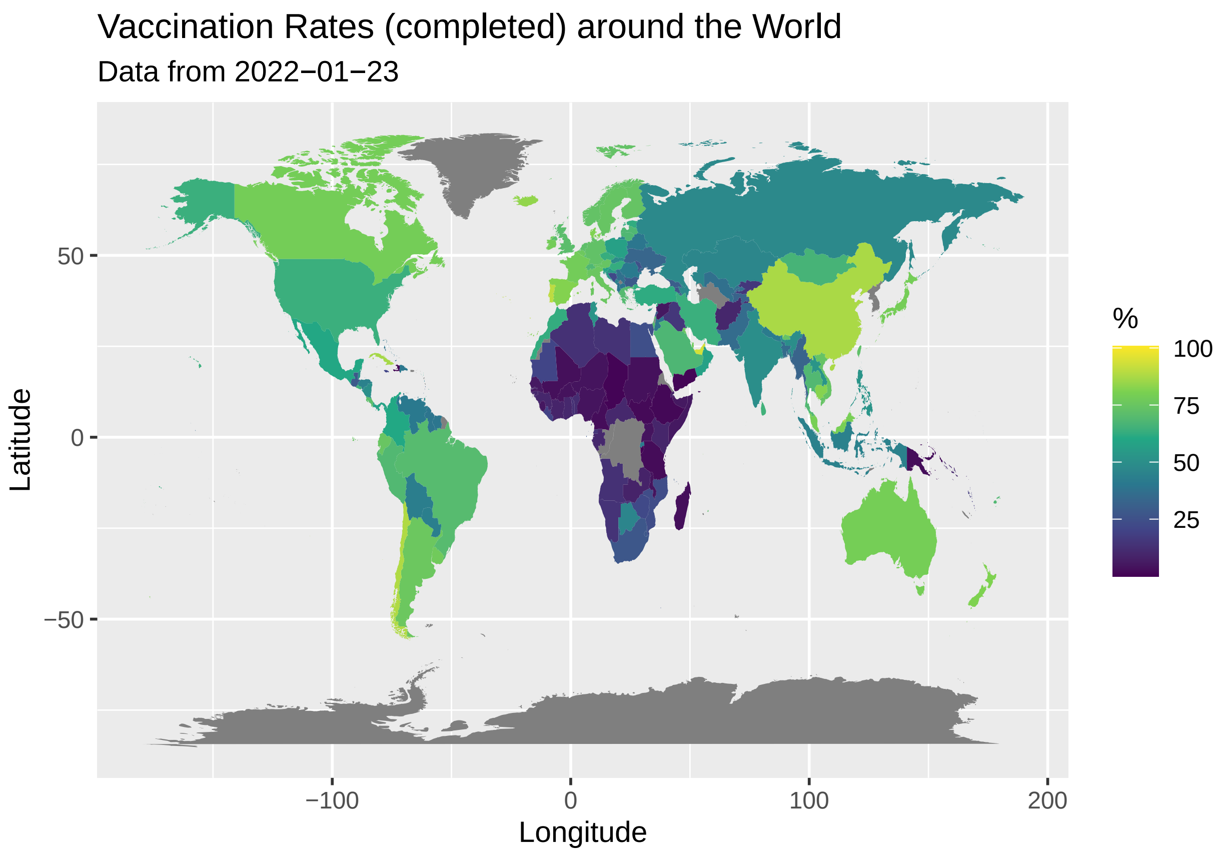
```
# Lolipop plot for the top 20 countries with regard to vaccination rate
t <- na.omit(vax_now, cols = c("partially_vaccinated_ratio", "fully_vaccinated_ratio"))
t <- t[order(-fully_vaccinated_ratio)[1:20]]
ggplot(t, aes(reorder(country_region, -fully_vaccinated_ratio), fully_vaccinated_ratio)) +
  geom_segment(aes(x=reorder(country_region, -fully_vaccinated_ratio),
                   xend=reorder(country_region, population),
                   y=0, yend=fully_vaccinated_ratio)) +
  geom_point(size=5, color="red",
             fill=alpha("orange", 0.3), alpha=0.7, shape=21, stroke=2) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = "Countries with the Highest Vaccination Coverage",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "", y = "Fully Vaccinated Ratio")
```



Map plots

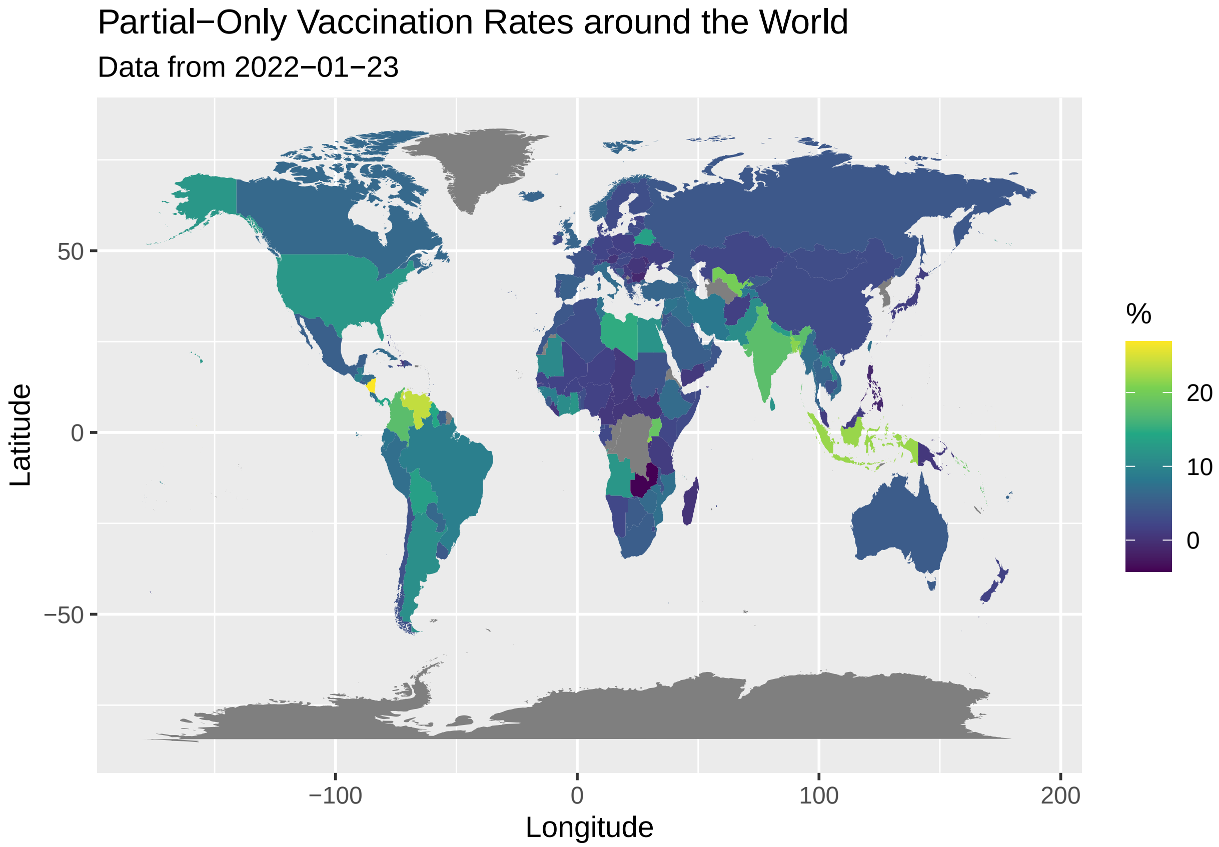
The data can be displayed on a world map which can provide additional insight. It is clear that Africa has very low vaccination coverage, while in the Western World and China most people are vaccinated.

```r
# Make a world map data.table
world_map <- as.data.table(map_data("world"))
# Convert to iso3 to use it as the key for joining
world_map$iso3 <- iso.alpha(world_map$region, 3)
vax_map <- vax_now[world_map, on = .(iso3)]  # left join

# Map-plot Fully Vaccinated Rates
ggplot(vax_map, aes(i.long, i.lat, group = group)) +
  geom_polygon(aes(fill = fully_vaccinated_ratio)) +
  scale_fill_viridis_c(option = "viridis", name = "%") +
  labs(title = "Vaccination Rates (completed) around the World",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "Longitude", y = "Latitude")
```



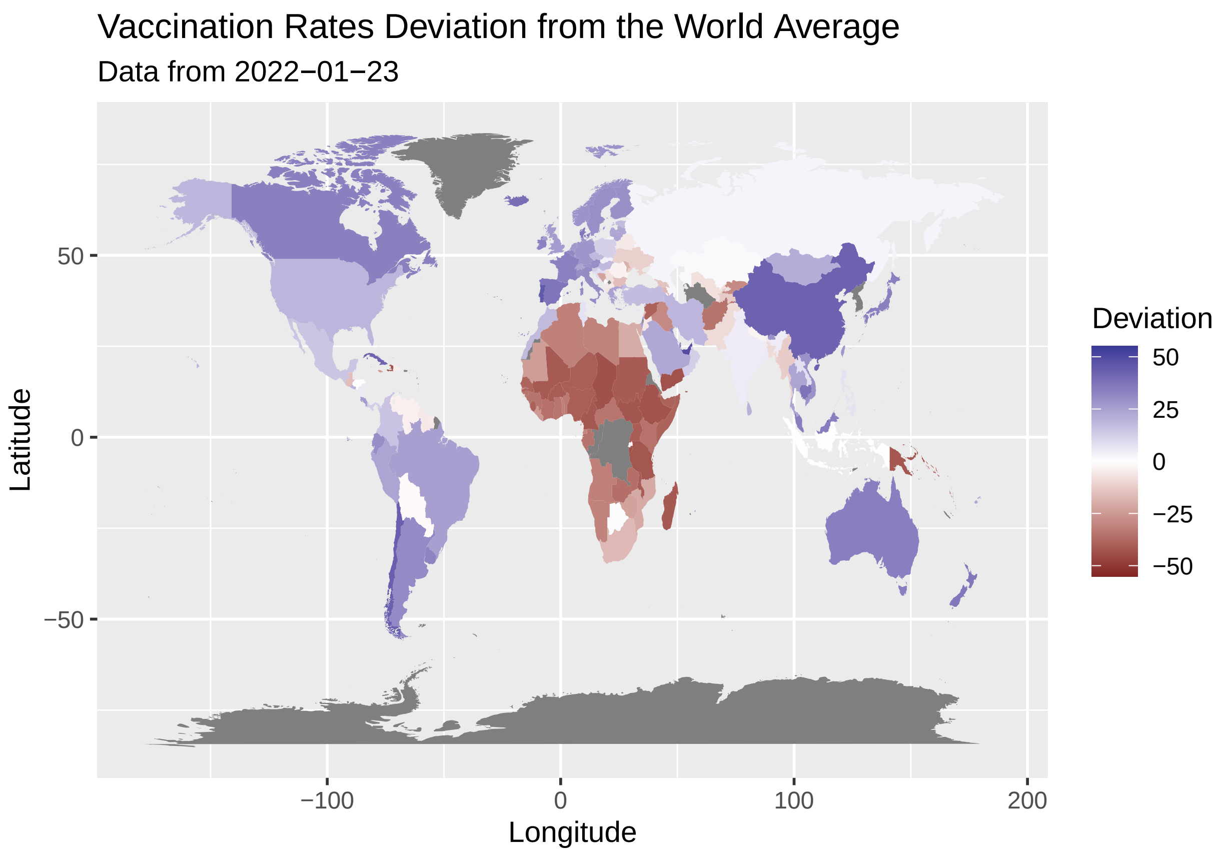
Some countries have many partial vaccinations, because there's not enough doses to fully vaccinate everyone.

```
# Map-plot Partial-Only Vaccination Rates
ggplot(vax_map, aes(i.long, i.lat, group = group)) +
  geom_polygon(aes(fill = partially_vaccinated_ratio - fully_vaccinated_ratio)) +
  scale_fill_viridis_c(option = "viridis", name = "%") +
  labs(title = "Partial-Only Vaccination Rates around the World",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "Longitude", y = "Latitude")
```



Finally, in the map below, we can observe the deviations from the world average.

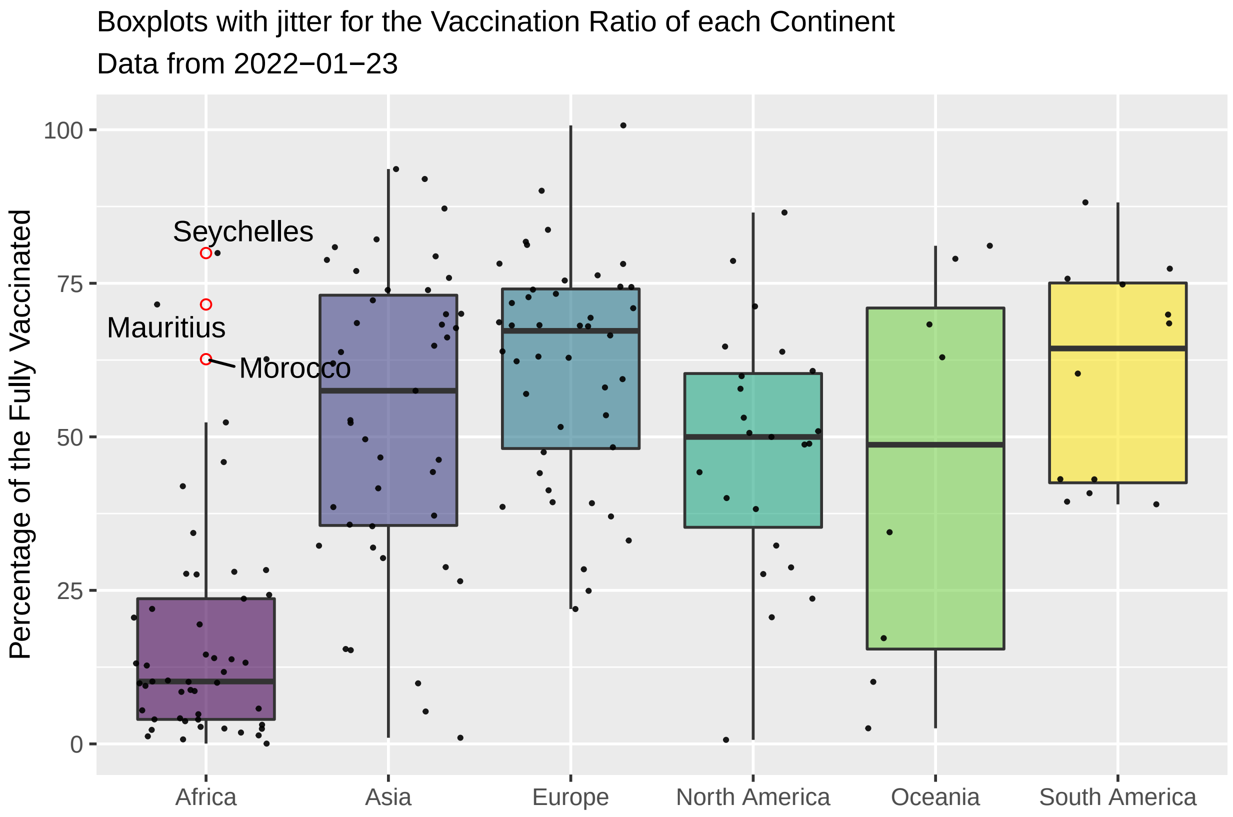
```r
# Deviation from the World Average Map
mean_partial <- vax_now[, mean(fully_vaccinated_ratio, na.rm = TRUE)]
t <- na.omit(vax_now, cols = c("fully_vaccinated_ratio", "fully_vaccinated_ratio"))
t$above_below <- t$fully_vaccinated_ratio > mean_partial
t$deviation <- t$fully_vaccinated_ratio - mean_partial
ggplot(t[world_map, on = .(iso3)], aes(i.long, i.lat, group = group)) +
  geom_polygon(aes(fill = deviation)) +
  # scale_fill_distiller(type = "div", name = "Deviation") +
  scale_fill_gradient2(midpoint = 0, limits = c(-55, 55), name = "Deviation") +
  labs(title = "Vaccination Rates Deviation from the World Average",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "Longitude", y = "Latitude")
```



Grouping by Continent

With a boxplot we can see how the vaccination rates are distributed on each continent. In the boxplots below, the outliers are annotated explicitly. Europe has very high variance, since Western Europe has much higher vaccine rates than Easter Europe. Asia also has high variance, due to the sheer size of the continent.

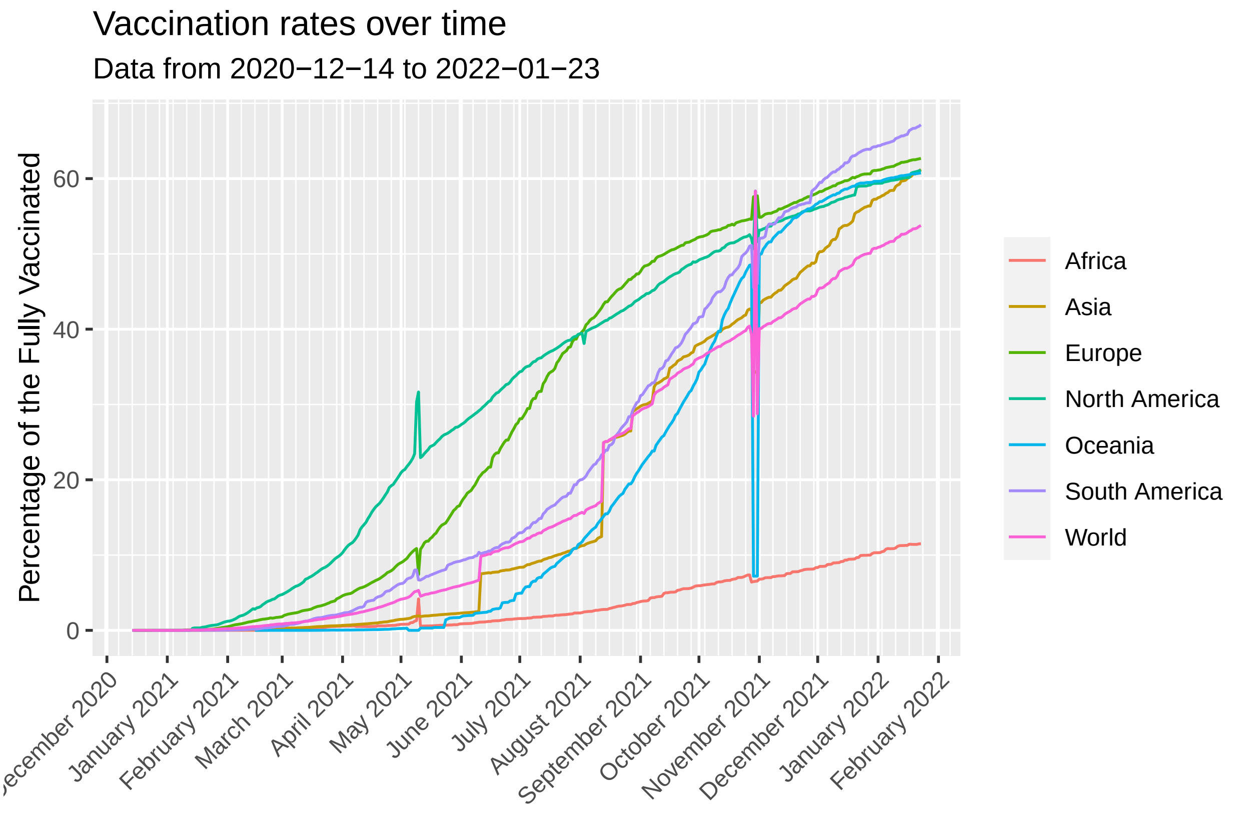
```
# Box Plots per continent
t <- na.omit(vax_now, cols = c("continent_name", "fully_vaccinated_ratio"))

# We add a column that keeps track of outliers for each continentm
# so that we can use it to annotate them in the boxplots.
is.outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}
t[ , outlier := ifelse(is.outlier(fully_vaccinated_ratio), country_region, NA_character_) ,
     by = continent_name]

ggplot(t,
       aes(x = continent_name, y = fully_vaccinated_ratio, fill = continent_name)) +
  geom_boxplot(na.rm = TRUE, outlier.shape = 1, outlier.color = "red") +
  scale_fill_viridis_d(alpha=0.6) +
  geom_jitter(na.rm = TRUE, color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  labs(title = "Boxplots with jitter for the Vaccination Ratio of each Continent",
       subtitle = sprintf("Data from %s", vax_now[1, date]),
       x = "", y = "Percentage of the Fully Vaccinated") +
  geom_text_repel(aes(label = outlier), na.rm = TRUE, show.legend = FALSE)
```



By utilizing a time series plot, we can see how the vaccinations progressed on each continent. Note that the curves are not strictly increasing. This could be due to wrong/missing reports, or maybe some countries stop considering people vaccinated after a certain amount of time has passed since they received the last vaccine dose.

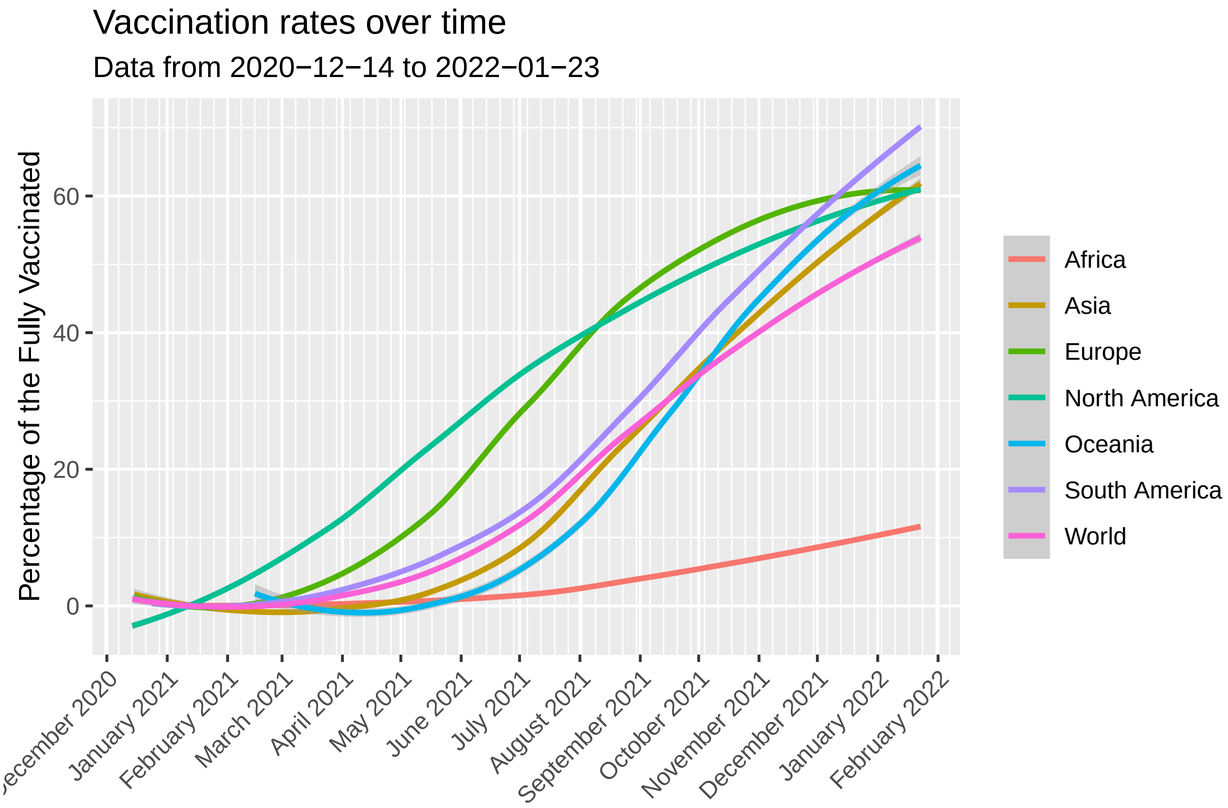
```
# World-wide and per continent daily time series
t <- na.omit(vax, cols = c("fully_vaccinated_ratio", "date",
                          "country_region", "continent_name"))
continent_avg <- t[ ,
  .(mean_fully_vaccinated_ratio = weighted.mean(fully_vaccinated_ratio, population)),
  by = .(date, continent_name)
]
world_avg <- t[ ,
  .(continent_name = "World",
    mean_fully_vaccinated_ratio = weighted.mean(fully_vaccinated_ratio, population)),
  by = date
]
continent_avg <- rbind(continent_avg, world_avg)[order(date, continent_name)]
```

```
ggplot(continent_avg, aes(date, mean_fully_vaccinated_ratio,
                          colour=continent_name, group = continent_name)) +
  geom_line() +
  scale_x_date(date_breaks = "1 month", date_minor_breaks = "1 week",
               date_labels = "%B %Y") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) +
  labs(title = "Vaccination rates over time",
       subtitle = sprintf("Data from %s to %s",
                          min(continent_avg$date), max(continent_avg$date)),
       x = "", y = "Percentage of the Fully Vaccinated") +
  scale_color_discrete(name = "")
```



Finally, we can create a less noisy view by smoothening the curves. Initially, Europe and North America had very high rates compared to the rest of the world, but now the other continents, except Africa, have caught up. It is worth noting that South America is now the leader in vaccinations.

```
ggplot(continent_avg, aes(date, mean_fully_vaccinated_ratio,
                          colour=continent_name, group = continent_name)) +
  geom_smooth(method = "loess", formula = y ~ x) +
  scale_x_date(date_breaks = "1 month", date_minor_breaks = "1 week",
               date_labels = "%B %Y") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1)) +
  labs(title = "Vaccination rates over time",
       subtitle = sprintf("Data from %s to %s",
                          min(continent_avg$date), max(continent_avg$date)),
       x = "", y = "Percentage of the Fully Vaccinated") +
  scale_color_discrete(name = "")
```


Vaccination rates over time
Data from 2020−12−14 to 2022−01−23
Percentage of the Fully Vaccinated
ecember 2020
January 2021
February 2021
March 2021
April 2021
May 2021
June 2021
July 2021
August 2021
September 2021
October 2021
November 2021
December 2021
January 2022
February 2022
Africa
Asia
Europe
North America
Oceania
South America
World