Exploratory Data Analysis using R

MSc Apostolos Kakampakos AM: 03400133
email: apostolos.kakampakos@gmail.com

13 August 2022

Abstract

This presentation is about the covid19_vaccine dataset, located in the coronavirus library of R. We will present various statistics about the vaccination programs of many countries during the covid 19 pandemic crisis. The presentation was made for the class "Programming Tools and Technologies for Data Science, 2021-2022" of the Postgraduade Program: DSML, NTUA

National Technical
University of Athens

Exploring our dataset

First we will import the following libraries used in the analysis of the dataset. The library coronavirus contains the dataset with vaccination data from all over the world.

```
#libraries
library(coronavirus)
library(data.table)
library(ggplot2)
library(gridExtra)
```

It should also be noted that the command 'update_dataset(silence = FALSE)' was used for the library coronavirus, so as to have as close as possible, up to date data.

Libraries like data.table and ggplot2 will be used throughout our analysis. Let us now use a data table structure to explore our data. The start and end dates are:

```
data(covid19_vaccine)

covid19_vaccine <- as.data.table(covid19_vaccine)

start_date = covid19_vaccine[,min(date)]
end_date = covid19_vaccine[,max(date)]

print(start_date)
```

```
## [1] "2020-12-14"
```

```
print(end_date)
```

```
## [1] "2022-01-19"
```

It seems our dataset spans from December 2020 to January 2022. The year of 2021 will be the main year of analysis, which is to be expected with the covid-19 shots.

Let us access our unprocessed dataset:

```
head(covid19_vaccine)
```

```
##    country_region       date doses_admin people_partially_vaccinated
## 1:         Canada 2020-12-14           5                           0
## 2:          World 2020-12-14           5                           0
## 3:         Canada 2020-12-15         723                           0
## 4:          China 2020-12-15     1500000                           0
## 5:         Russia 2020-12-15       28500                       28500
## 6:          World 2020-12-15     1529223                       28500
##    people_fully_vaccinated report_date_string uid province_state iso2 iso3
## 1:                       0         2020-12-14 124           <NA>   CA  CAN
## 2:                       0         2020-12-14  NA           <NA> <NA> <NA>
## 3:                       0         2020-12-15 124           <NA>   CA  CAN
## 4:                       0         2020-12-15 156           <NA>   CN  CHN
## 5:                       0         2020-12-15 643           <NA>   RU  RUS
## 6:                       0         2020-12-15  NA           <NA> <NA> <NA>
##    code3 fips      lat     long combined_key population continent_name
## 1:   124 <NA> 60.00000 -95.0000       Canada   37855702  North America
## 2:    NA <NA>       NA       NA         <NA>         NA           <NA>
## 3:   124 <NA> 60.00000 -95.0000       Canada   37855702  North America
## 4:   156 <NA> 35.86170 104.1954        China 1404676330           Asia
## 5:   643 <NA> 61.52401 105.3188       Russia  145934460         Europe
## 6:    NA <NA>       NA       NA         <NA>         NA           <NA>
##    continent_code
## 1:             NA
## 2:           <NA>
## 3:             NA
## 4:             AS
## 5:             EU
## 6:           <NA>
```

As we can see from the first 6 entries we have various empty and uninteresting values. Also we have many columns that are simply not needed in this analysis.

```
# remove various uninteresting columns...
covid19_vaccine[ ,':='(uid = NULL, province_state = NULL,
                       iso2 = NULL, iso3 = NULL,
                       code3 = NULL, fips = NULL,
                       lat = NULL, long = NULL,
                       continent_code = NULL,
                       report_date_string = NULL)]
```

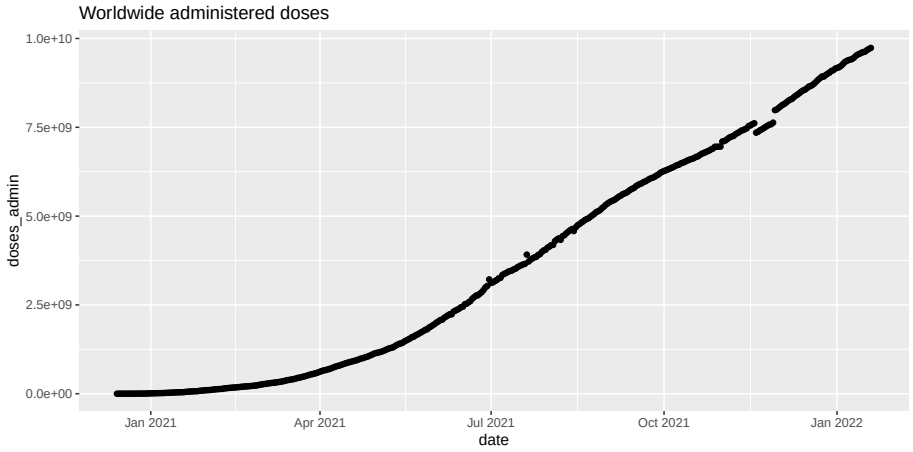
Therefore our dataset has the form:

```
head(covid19_vaccine)
```

```
##    country_region       date doses_admin people_partially_vaccinated
## 1:         Canada 2020-12-14           5                           0
## 2:          World 2020-12-14           5                           0
## 3:         Canada 2020-12-15         723                           0
## 4:          China 2020-12-15     1500000                           0
## 5:         Russia 2020-12-15       28500                       28500
## 6:          World 2020-12-15     1529223                       28500
##    people_fully_vaccinated combined_key population continent_name
## 1:                       0       Canada   37855702  North America
## 2:                       0         <NA>         NA           <NA>
## 3:                       0       Canada   37855702  North America
## 4:                       0        China 1404676330           Asia
## 5:                       0       Russia  145934460         Europe
## 6:                       0         <NA>         NA           <NA>
```

As we can see from above, doses were first administered in Canada. Interestingly the next day China administered 1.5 million doses! Furthermore, Russia started with almost 30 thousand administered doses. Curiously, as we can see from above, Russia also classified them as already partially vaccinated, while Canada and China didn't. This means that, there are different methods of calculating the above statistics depending the country! Now for the worldwide administration of doses we have the following graph:

```
ggplot(covid19_vaccine[country_region == 'World'],
       aes(x = date, y = doses_admin)) +
       geom_point() + ggtitle("Worldwide administered doses")
```



Clean up

Finally, before we move to the analysis, we need to deal with values that are simply missing from our dataset. First we will find the world population, using all the available data we have from this dataset.

```
world_population = na.omit(unique(covid19_vaccine,
                    by = c('country_region')))[, sum(population)]
cat("World's population is approximately", world_population)
```

```
## World's population is approximately 7508793688
```

So, we will replace the information of 7.5 billion to our dataset, which for the world had population registered as NA. Although now the world population is accounted to be close to 7.9 billion, we will keep the value present in the countries of our dataset.

```
covid19_vaccine[country_region == 'World',
                population := world_population]
```

The only thing that is left, is to remove any NA rows from our dataset, that are in the columns people_partially_vaccinated and people_fully_vaccinated.

```
covid19_vaccine <-covid19_vaccine[!is.na(covid19_vaccine$
                    people_partially_vaccinated|
                     covid19_vaccine$people_fully_vaccinated),]
```

Now the only thing left is to clean our dataset from information on provinces. This will be done using the combined_key column, which only lists the country. First we will locate all instances where the country_region is equal to the combined_key and keep them. Then remove the combined_key column.

```
covid19_vaccine<-covid19_vaccine[covid19_vaccine$country_region
                                 == covid19_vaccine$combined_key|
                                   covid19_vaccine$country_region
                                      == 'World',]
covid19_vaccine[ ,':='(combined_key=NULL)]
```

Having adequately cleaned our dataset, we can begin analysing and plotting.

Analyzing our dataset

creating data tables

As we can see from the previous part, our data is subject to noise and human error, as well as different strategies of counting vaccinations.

Now we will begin to analyze our dataset and extract as much information as we can. We will begin by creating two distinct data tables for fully vaccinated and partly vaccinated populations. The idea is to represent how the percentage of partly and fully vaccinated populations change over time.

First for the partially vaccinated:

```
partially_vaccinated_ratio <- covid19_vaccine[,
                    !c("people_fully_vaccinated", "doses_admin")]
partially_vaccinated_ratio[,
  people_partially_vaccinated :=
       round(100*people_partially_vaccinated/population,digits=1)]
colnames(partially_vaccinated_ratio) <- c("country", "date",
                           "percentage", "population", "continent")
```

Secondly for the fully vaccinated:

```
fully_vaccinated_ratio <- covid19_vaccine[,
                    !c("people_partially_vaccinated", "doses_admin")]
fully_vaccinated_ratio[,
  people_fully_vaccinated :=
       round(100*people_fully_vaccinated/population,digits=1)]
colnames(fully_vaccinated_ratio) <- c("country", "date",
                      "percentage", "population", "continent")
```

And finally for the doses administered:

```
doses_admin_ratio <- covid19_vaccine[,
                    !c("people_partially_vaccinated",
                       "people_fully_vaccinated")]
doses_admin_ratio[,doses_admin :=
       round(100*doses_admin/population,digits=1)]
doses_admin_ratio <- doses_admin_ratio[,!c("population")]
colnames(doses_admin_ratio) <- c("country", "date",
                                 "percentage", "continent")
```

plotting cumulative vaccination rates

Let us see how the worldwide vaccination program evolved during 2021.

```
ggplot() +
  geom_line(data=fully_vaccinated_ratio[country == 'World'],
            aes(date, percentage, colour='fully vaccinated')) +
  geom_line(data=partially_vaccinated_ratio[country == 'World'],
            aes(date, percentage, colour='partially vaccinated'))+
  scale_color_manual(name = "status",
            values = c("fully vaccinated" = "darkblue",
                               "partially vaccinated" = "red")) +
  ggtitle("Worldwide growth of vaccinations")
```

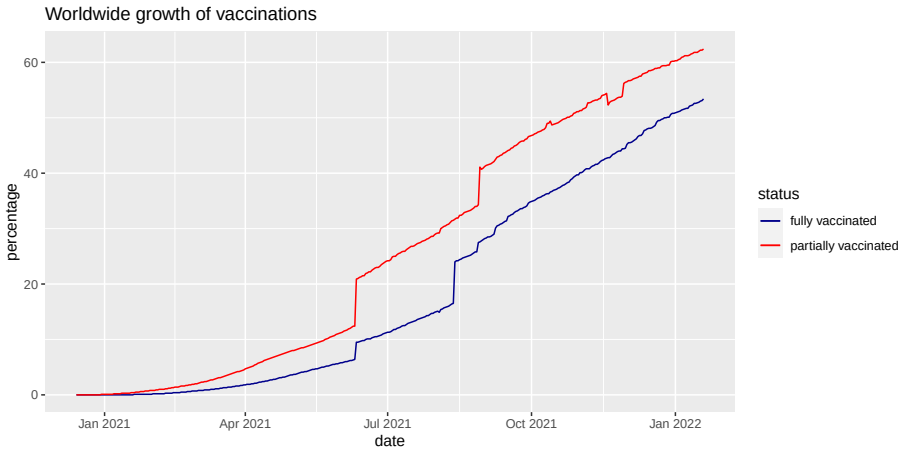


Figure 1: Evolution of worldwide vaccination program

As we can clearly see from figure 1, world vaccinations have been steadily increasing throughout 2021, with an expected lead from partial vaccinations.

Two things are to be noted:

- The fact that the above graph is of cumulative quantities, hiding daily vaccination rates.
- The difference between partial and fully vaccinated populations changes over time, which in it self holds potentially interesting info.

We will explore both observations.

Daily vaccinations

Our objective now is to extract daily vaccination rates from the cumulative percentages. This is easily done, as we can express each day's percentage as the cumulative of this day minus the previous day.

First for the partially vaccinated:

```
all = unique(partially_vaccinated_ratio$country)
for (cntry in all)
{
partially_vaccinated_ratio[country == cntry,
  daily := c(percentage[1], diff(percentage))]
}
```

Secondly for the fully vaccinated:

```
all = unique(fully_vaccinated_ratio$country)
for (cntry in all)
{
fully_vaccinated_ratio[country == cntry,
  daily := c(percentage[1], diff(percentage))]
}
```

Let us plot the daily vaccinations worldwide.

```
dt_partial <- partially_vaccinated_ratio[country=='World']
dt_full <- fully_vaccinated_ratio[country=='World']
plot1 <- ggplot(data = dt_full, aes(x = date, y = daily)) +
  geom_area(fill="#1FD371", alpha=1) +
  geom_line(color="#1FD371", size=0.6) +
  labs(x = 'Time period', y = 'Percentage (%)',
       title = 'Daily increase of fully vaccinated')
plot2 <- ggplot(data = dt_partial, aes(x = date, y = daily)) +
  geom_area(fill="#73CDFA", alpha=1) +
  geom_line(color="#73CDFA", size=0.6)+
  labs(x = 'Time period', y ='',
       title = 'Daily increase of partially vaccinated')
grid.arrange(plot1, plot2, ncol =2)
```

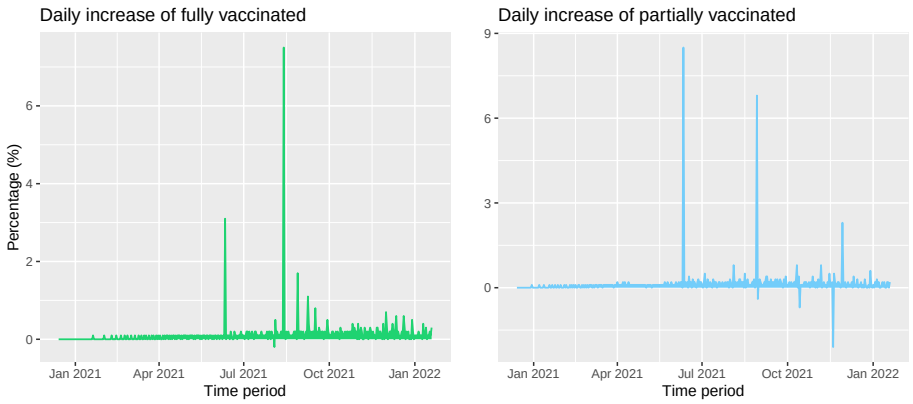


Figure 2: Daily vaccinations worldwide

As we can see from figure 2, we have mostly small (positive) updates each day. Interestingly, some times we have a decrease in vaccinations! Obviously, this means that due to human error, miscalculation, etc some previous vaccination reports were reconcidered and a correct reduced update was given.

We can also find the difference of vaccination status worldwide:

```
dif <- covid19_vaccine[country_region == 'World'][,difference :=
       people_partially_vaccinated - people_fully_vaccinated]
ggplot(dif, aes(x = date, y = difference)) +geom_point() +
  ggtitle("Worldwide difference of partially and fully vaccinated")
```

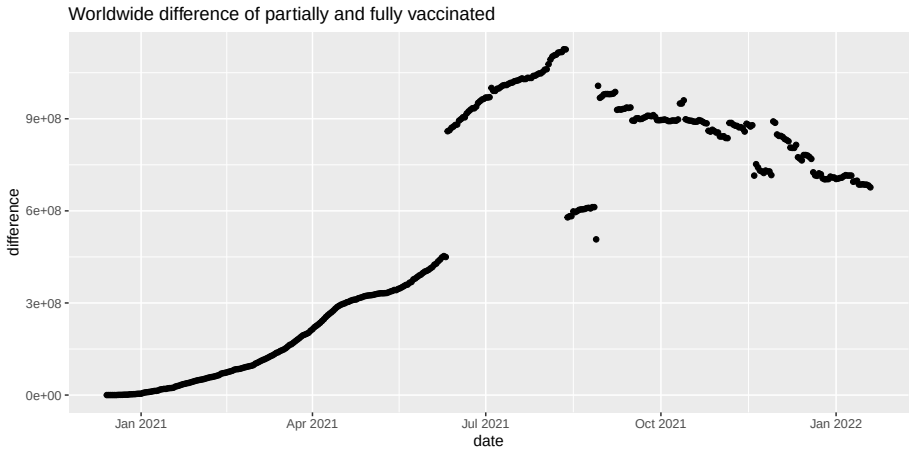


Figure 3: Vacination status difference worldwide

Figure 3 shows us that the difference in partially and fully vaccinated populations reaches close to a billion! Furthermore, one can see that the difference amplifies during the summer months and then slowly declines as we enter the winter season. There are multiple reasons for this. From the slow decrease of lockdown measures, to the vaccination offer bonuses to younger populations. Unfortunately this dataset, doesn't offer the age groups, to further explore how the vaccination program success differed with different age groups.

Today's stats

Besides, the evolution of the vaccinations, we can analyze current information about the vaccination programs of different countries.

Let us present the first 10 countries with the highest vaccination percentage in the world.

```
dt<-fully_vaccinated_ratio[date==end_date,.(percentage=percentage),
                     by = .(country)][order(-percentage)]
ggplot(data = dt[1:10], aes(x = reorder(unlist(country),
                           unlist(percentage)),y = percentage)) +
  geom_bar(stat="identity", position=position_dodge(),
           color = rgb(70, 130,180, maxColorValue = 255),
           fill = rgb(70,130,180, maxColorValue = 255))+
  labs(title = paste('Top 10 highest vaccinated countries until '
                     , end_date),x = 'Countries',y = 'Percentage')+
  theme(axis.text.x = element_text(angle = 90))+coord_flip()
```

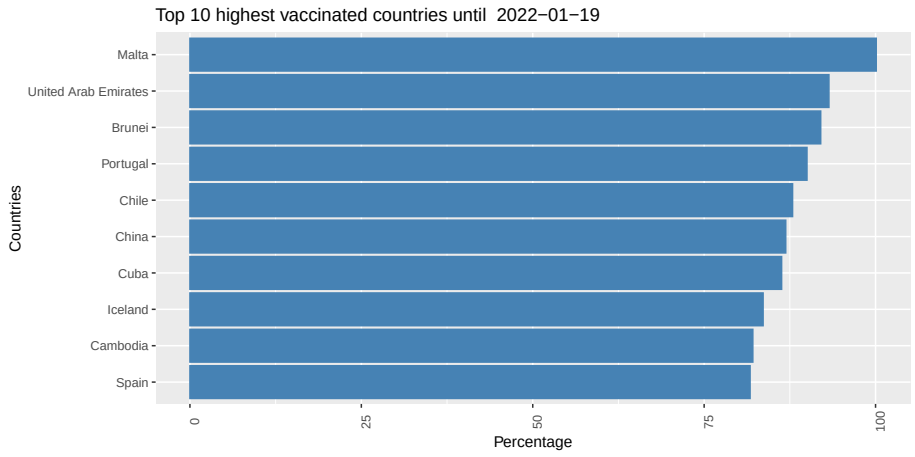


Figure 4: Top 10 vaccinated countries today

We can also find how the fully vaccinated percentages are distributed in the European continent.

```
dt1 <- fully_vaccinated_ratio[continent=='Europe'][date == end_date]
ggplot(data=dt1[order(-percentage)], aes(x=reorder(country,
 -percentage), y=percentage)) +
 geom_bar(stat="identity", position=position_dodge(),
  fill = "#FF6666", width=1) + labs(title =
   paste('Percentage of Fully vaccinated in Europe by country, date: ',
         end_date), x = 'Europe',y = 'Percentage (%)') +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

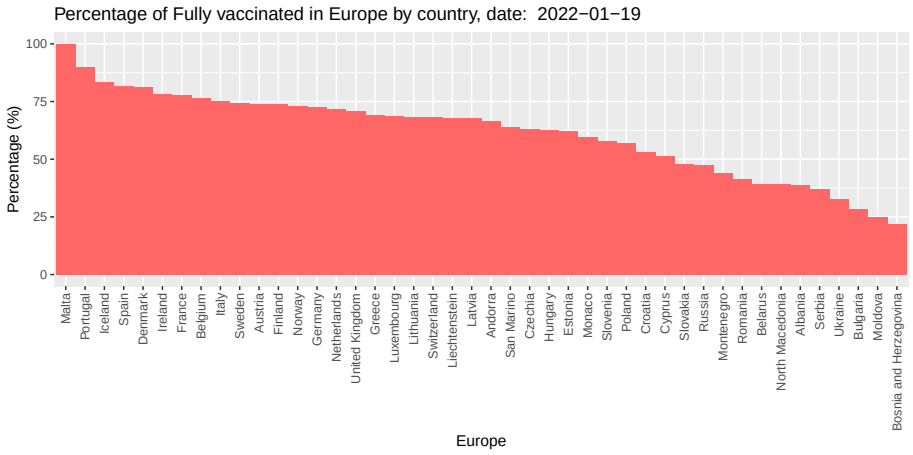


Figure 5: European distribution of vaccinations

The above figure 5, although it informs us about the successes of the individual countries programs of vaccinations, it certainly is misleading. Malta's 100% is nowhere near the ~50% of Russia's population. Therefore we can search for the contribution of each European country in the vaccination of the continent.

First we can calculate Europe's total population

```
european_population = unique(dt1,
                     by = c('country'))[, sum(population)]
cat("Europe's population is approximately", european_population)
```

```
## Europe's population is approximately 748501150
```

Then we can simply readjust our previous calculation, so as to show the contribution of each country to the vaccination of the continent.

```
dt1 <- dt1[, contribution:= percentage*population/european_population]
ggplot(data=dt1[order(-contribution)], aes(x=reorder(country,
 -contribution), y=contribution)) +
 geom_bar(stat="identity", position=position_dodge(),
  fill = "#FF6666", width=1) + labs(title =
   paste('Contribution of each country in Europe, date: ',
         end_date), x = 'Europe',y = 'Percentage (%)') +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

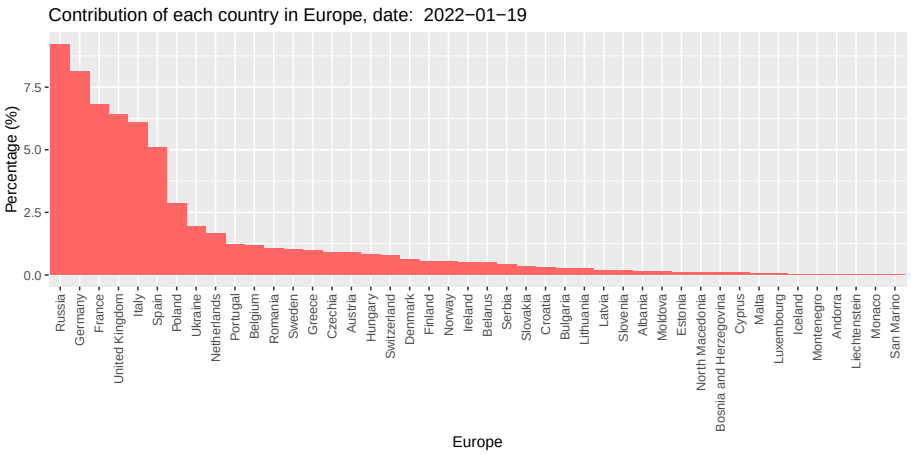


Figure 6: European distribution of vaccinations

```
european_vaccinee_percentage = unique(dt1,
                    by = c('country'))[, sum(contribution)]
cat("European population fully vaccinated"
    , round(european_vaccinee_percentage, digits=1), "%")
```

```
## European population fully vaccinated 62.5 %
```

Figure 6, clearly shows that countries like Russia and Germany lead the way in the vaccination of the European continent. In terms of global population, Europe's vaccination contribution is:

```
european_vaccinee_global <-
 european_vaccinee_percentage*european_population/world_population
cat("Europe's contribution to the world"
    , round(european_vaccinee_global, digits=1), "%")
```

```
## Europe's contribution to the world 6.2 %
```

Therefore, an important information would be how the contribution of all continents are in terms of global population.

```
contributions = c()
populations = c()
continents =  na.exclude(unique(fully_vaccinated_ratio[
  date == end_date]$continent))
for (cont in continents)
{
  dt2 <- fully_vaccinated_ratio[continent==cont][date == end_date]
  dt2 <- dt2[, contribution:= percentage*population/world_population]
  continent_percentage = unique(dt2,
                    by = c('country'))[, sum(contribution)]
  continent_population = unique(dt2,
                    by = c('country'))[, sum(population)]
  contributions <- c(contributions, continent_percentage)
  populations <- c(populations, continent_population)
}
world <- data.table(continent = continents,
                    contribution = contributions,
                    population = populations)
#ploting contributions...
ggplot(data=world[order(-contribution)], aes(x=reorder(continent,
 -contribution), y=contribution)) +
 geom_bar(stat="identity", position=position_dodge(),
  fill = "#0000FF", width=1) + labs(title =
   paste('Contribution of each continent, date: ',
         end_date), x = 'Continents',y = 'Percentage (%)') +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

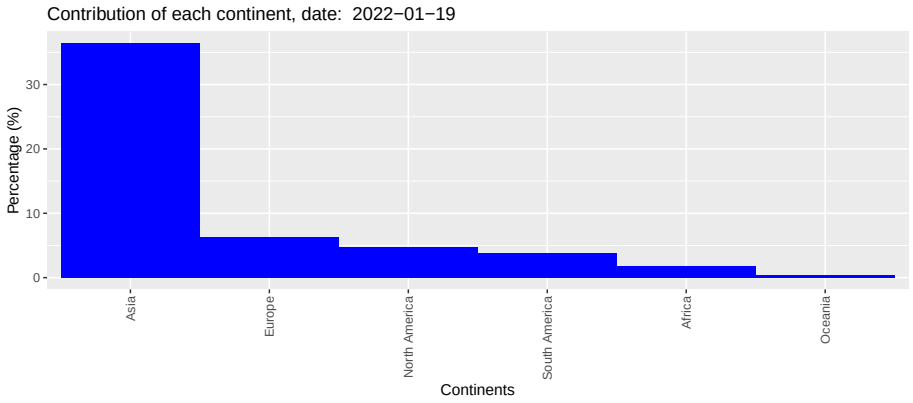


Figure 7: Continents distribution of vaccinations

As we can see from Figure 7 the greatest contribution to the world's vaccinations has clearly been Asia, with Europe a distant second. Obviously, a part of this is explained simply from population densities. To get a complimentary view, we will also plot the vaccination rate of each continent in terms of it's own population.

```
world[, percentage := contributions*world_population/population]
ggplot(data=world[order(-percentage)], aes(x=reorder(continent,
 -percentage), y=percentage)) +
 geom_bar(stat="identity", position=position_dodge(),
  fill = "#0000BB", width=0.8) + labs(title =
   paste('Vaccination percentage of each continent, date: ',
         end_date), x = 'Continents',y = 'Percentage (%)') +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

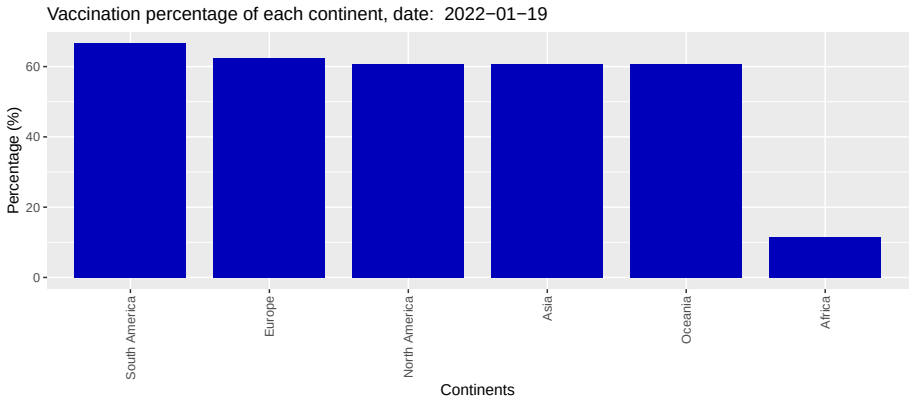


Figure 8: Continent percentages of vaccinations

Impressively, all continents but Africa are pretty close in their vaccination rates. Due to many factors, Africa, which contains a huge percentage of the world's population, has simply the worst vaccination rate by far. This is to be expected, since most developing countries are in Africa.

Summary

As we have seen in this presentation of the covid_19 vaccination dataset, world vaccinations with the covid 19 shots have been steadily increasing throughout 2021. The most contributing continent in worldwide vaccinations has clearly been Asia. Africa is the notable exception, with it having the lowest vaccination rates in the world. In terms of Europe, countries like Malta had the greatest success in their vaccination programs, with the North European countries as well as the Iberian continent contributing the most in terms of vaccinations, not including Russia, which although having relatively low rates has the greatest contributions, simply because of population size. Further analysis could be made, if the dataset provided vaccination rates for different population brackets.

This presentation was made using the R language and R markdown.