National Technical University of Athens

MSc Data Science and Machine Learning

Programming Tools and Technologies for Data Science

Analyzing the Impact of Vaccination against COVID-19

Pavlos Kalfantis
MSc Student
pavloskalfantis@mail.ntua.gr
Student ID: 03400134

January 2022

1 Introduction

December 2021 marked two years since a new disease that came to be known as Covid-19, emerged in Wuhan, China. This new disease, caused by the novel coronavirus SARS-CoV-2, quickly began spreading to other locations worldwide. On March 11, 2020, the World Health Organization declared this outbreak as a global pandemic and governments around the world began imposing strict lockdowns to prevent the uncontrollable spread of the disease. The disruption that the restrictions have caused to the daily life of the world population has been enormous and is very difficult to be quantified. On a personal note, the global pandemic was one of the reasons that led me to move back to my home country of Greece after living and working in the United States for 5 years. Eventually, I ended up going back to school to study Data Science and Machine Learning and complete the exploratory data analysis presented on this report.

Since it became obvious that this disease had serious health effects to the global population, especially the elderly and people with underlying health conditions, the world's biggest Pharmaceutical companies started working on new vaccines that can prevent serious health effects from the disease and most importantly, prevent deaths. Thankfully, several of these companies were able to produce, get approved and distribute vaccines in groundbreaking speeds and on December 2020, the first vaccine dose was administered in the United Kingdom. Throughout 2021, vaccines were made available to an increasing number of people around the world that rushed to get a shot in order to help global life go back to normal.

2 The Dataset

As part of the graduate course 'Programming Tools and Technologies for Data Science', this project consists of Exploratory Data Analysis and Visualizations of Covid-19 vaccinations, using the R programming language. The `coronavirus` package in R contains two datasets, the `coronavirus` dataset which contains information on confirmed Covid-19 cases and deaths of countries around the world, and the `covid19_vaccine` dataset which contains information on vaccine doses administered to people around the world. The second dataset will be the base of the analysis carried on this project, as the progress that countries around the world are making in vaccinating their population is of great interest. Since most of the vaccines available are to be administered in two doses, there are two different variables in our dataset. `people_fully_vaccinated` contains the number of people that have been administered both vaccines for each country and are considered fully vaccinated, whereas `people_partially_vaccinated` contains the number of people that have been administered one dose but are not yet considered fully vaccinated. The data available are given on a daily basis and broken down by country (and province for certain countries).

The first step of the project was to keep only data on the country level and add two new variables, `fully_vaccinated_ratio` and `parially_vaccinated_ratio` that transform the number of people administered vaccines to ratios, based on each country's population. The R code for each step of the project will be shown on the main body of this report as the analysis progresses. Another important note is that since the dataset is updated on a daily basis, the latest date that data were updated for this project was **January 8, 2022**. The code snippet shows how the dataset is updated and how the new columns with the vaccination ratios are created. Finally for the two datasets only information on country level is kept.

```
#Update dataset
library(coronavirus)
detach("package:coronavirus", unload = TRUE)
coronavirus::update_dataset()
data(covid19_vaccine)

#load packages to use
library(coronavirus)
library(data.table)
library(ggplot2)
library(xtable)

data(coronavirus)
data(covid19_vaccine)

#Convert data.frame to data.table
setDT(covid19_vaccine)
setDT(coronavirus)

#Keep information only on country level
covid19_vaccine = covid19_vaccine[is.na(province_state)]
coronavirus = coronavirus[is.na(province)]

#Create columns for vaccination ratios
covid19_vaccine[, fully_vaccinated_ratio :=
                  round(100*people_fully_vaccinated/population,2)]
covid19_vaccine[, partially_vaccinated_ratio :=
                  round(100*people_partially_vaccinated/population,2)]
```

3 Part 1 - Comparing Continents

For the first part in the exploratory data analysis project, the goal is to analyze the vaccination ratios at different continents. We will compare three continents that had early access to vaccines, namely Europe, South America and North America. The code below shows the aggregations of the absolute number of fully and partially vaccinated people and the plot below shows the fully vaccinated people through time.

```
#Part 1
#Aggregations by Continent (Europe, South America and North America)

#Aggregate total vaccinations for the three continents of interest
continents_vax = covid19_vaccine[,
                list(partially_vax=sum(people_partially_vaccinated),
                     fully_vax = sum(people_fully_vaccinated))
                ,by=list(continent_name,date)][continent_name
                      %in% list(’Europe’,’South America’,’North America’)]
```

```
#Plots to show vaccinated people per continent
ggplot(continents_vax, aes(x=date,y=fully_vax, color=continent_name)) +
  geom_line() +
  labs( x = "Date", y = "Fully Vaccinated People (millions)", color='Continent')
```

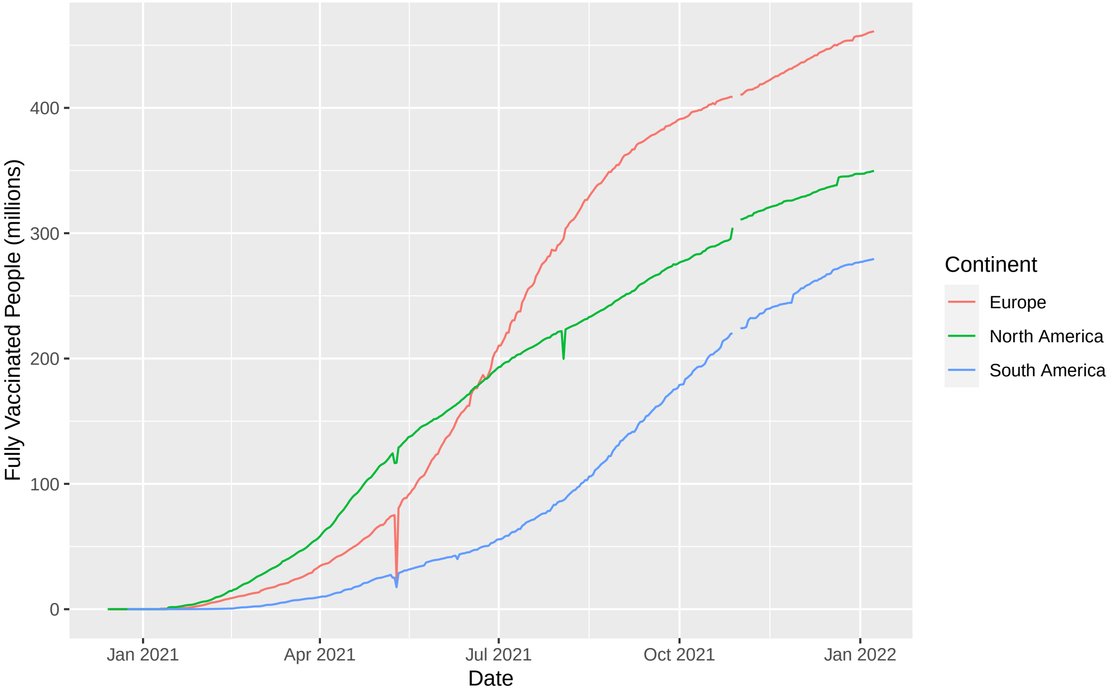

Figure 1: Number of fully vaccinated people per continent

Looking at figure 1 it is obvious that there is some noisy data in our dataset. There are two problems that can be identified with this aggregation. Firstly, the absolute number of fully vaccinated people is sharply dropping at certain dates, which is obviously impossible. Secondly, there are some gaps in the dataset, i.e. dates that total vaccinations are not reported in our data. It also seems that these dates with no reported vaccinations seems to coincide for each continent. Thus, the dataset will be cleaned firstly, by filling `NaN` values with the last available value for each continent and secondly by replacing values that are smaller that the ones on previous date with the value of the previous date (so keeping vaccinated people constant instead of decreasing). Figure 2 shows number of vaccinated people for the updated dataset.

```
#Data Cleaning
#First replace NaN values with previous value
continents_vax = continents_vax[order(continent_name)]
continents_vax[, fully_vax_fixed:= fully_vax[1],.
               (continent_name, cumsum(!is.na(fully_vax)))]
continents_vax[, partially_vax_fixed:= partially_vax[1],.
               (continent_name, cumsum(!is.na(partially_vax)))]
#Second, replace smaller values with previous value
continents_vax[, fully_vax_fixed:= cummax(fully_vax_fixed),
               by=.(continent_name)]
continents_vax[, partially_vax_fixed:= cummax(partially_vax_fixed),
               by=.(continent_name)]

#Plot absolute number of vaccinated people of update dataset
ggplot(continents_vax, aes(x=date,y=fully_vax_fixed, color=continent_name)) +
  geom_line() +
  labs( x = "Date", y = "Fully Vaccinated People (millions)", color=’Continent’)
```

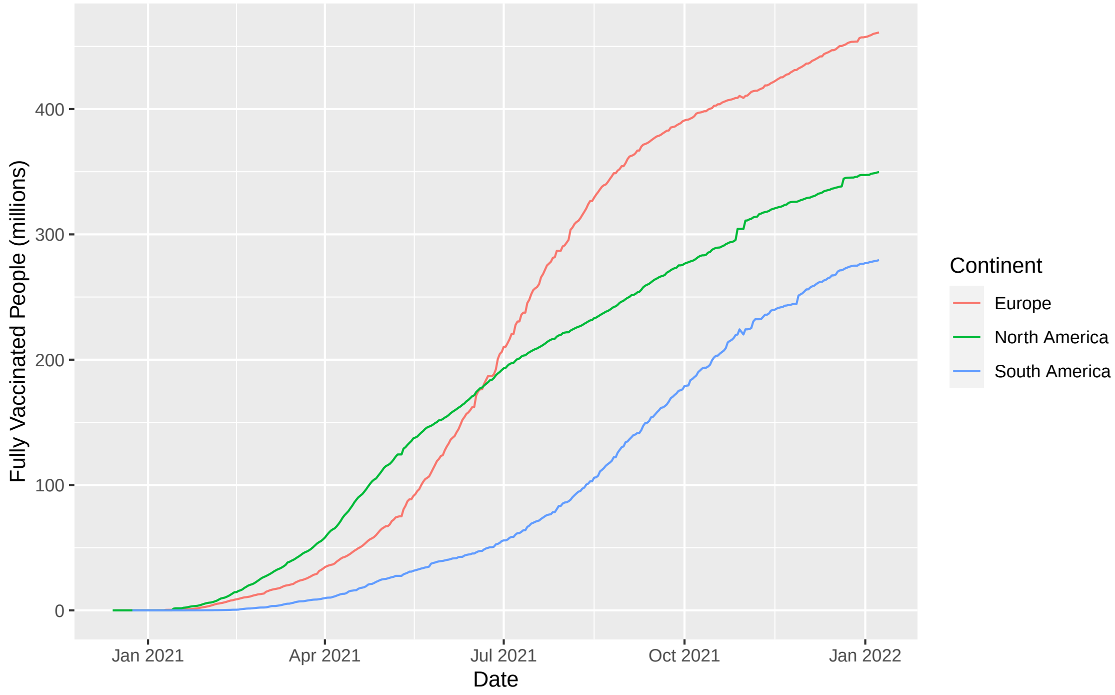

Figure 2: Number of fully vaccinated people per continent - Updated dataset

Now we can create the columns of interest (vaccination ratios) for the cleaned dataset. Since different countries started vaccinated people on different dates, aggregating the population column of the data will have different values per continent depending on the date. That is why when calculating the population of each continent to perform the calculation of vaccination ratios, the max value of aggregated populations will be used. The resulting plots with the two ratios for each continent are presented in figures 3 and 4.

```
#Create columns for vaccination ratios

#First find the population of each continent from the last data point
europe_pop = max(covid19_vaccine[,.(sum(population)),by=
                                   list(continent_name,date)][continent_name==
                                                                ’Europe’]$V1)
south_am_pop = max(covid19_vaccine[,.(sum(population)),by=
                                     list(continent_name,date)][continent_name==
                                                            ’South America’]$V1)
```

```
north_am_pop = max(covid19_vaccine[,.(sum(population)),by=
                                     list(continent_name,date)][continent_name==
                                                           ’North America’]$V1)

#Create vaccination ratio columns

continents_vax[continent_name==’Europe’, ‘:=‘ (
  fully_vaccinated_ratio= round(100*fully_vax_fixed/europe_pop,2),
  partially_vaccinated_ratio= round(100*partially_vax_fixed/europe_pop,2))]

continents_vax[continent_name==’South America’,‘:=‘(
  fully_vaccinated_ratio = round(100*fully_vax_fixed/south_am_pop,2),
  partially_vaccinated_ratio = round(100*partially_vax_fixed/south_am_pop,2))]

continents_vax[continent_name==’North America’,‘:=‘(
  fully_vaccinated_ratio = round(100*fully_vax_fixed/north_am_pop,2),
  partially_vaccinated_ratio = round(100*partially_vax_fixed/north_am_pop,2))]

#Plots

ggplot(continents_vax, aes(x=date,y=fully_vaccinated_ratio,
                           color=continent_name)) +  geom_line() +
  labs(x = "Date", y = "Fully Vaccinated Ratio (%)", color=’Continent’)

ggplot(continents_vax, aes(x=date,y=partially_vaccinated_ratio,
                           color=continent_name)) +  geom_line() +
  labs(x = "Date", y = "Partially Vaccinated Ratio (%)", color=’Continent’)
```

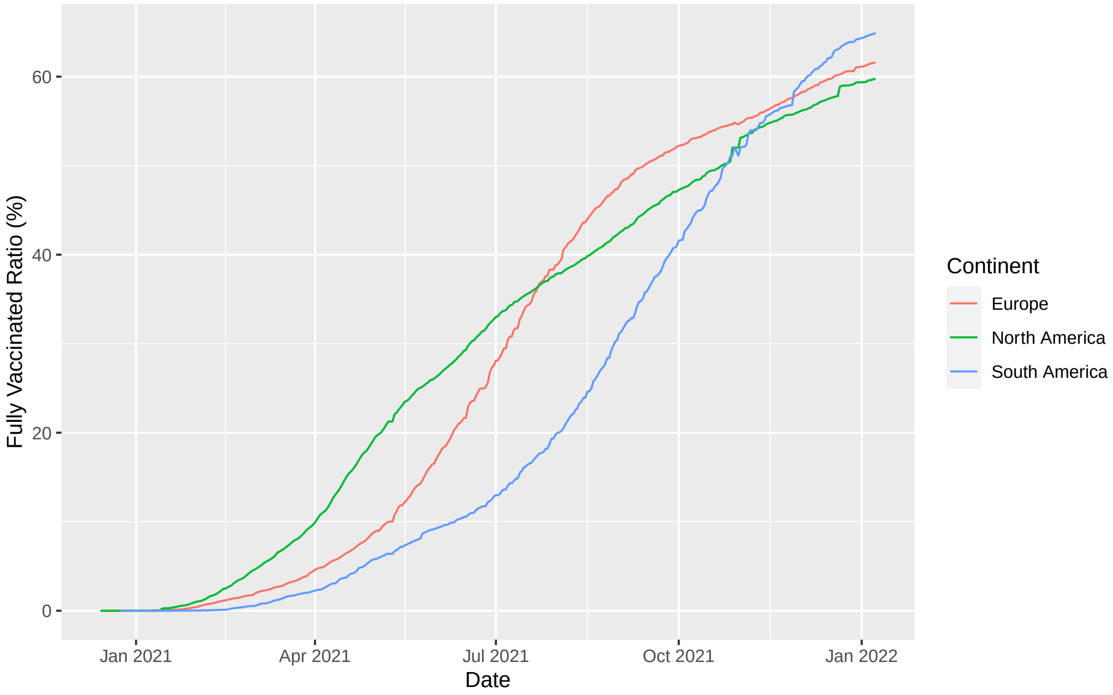

Figure 3: Fully vaccinated ratio per continent

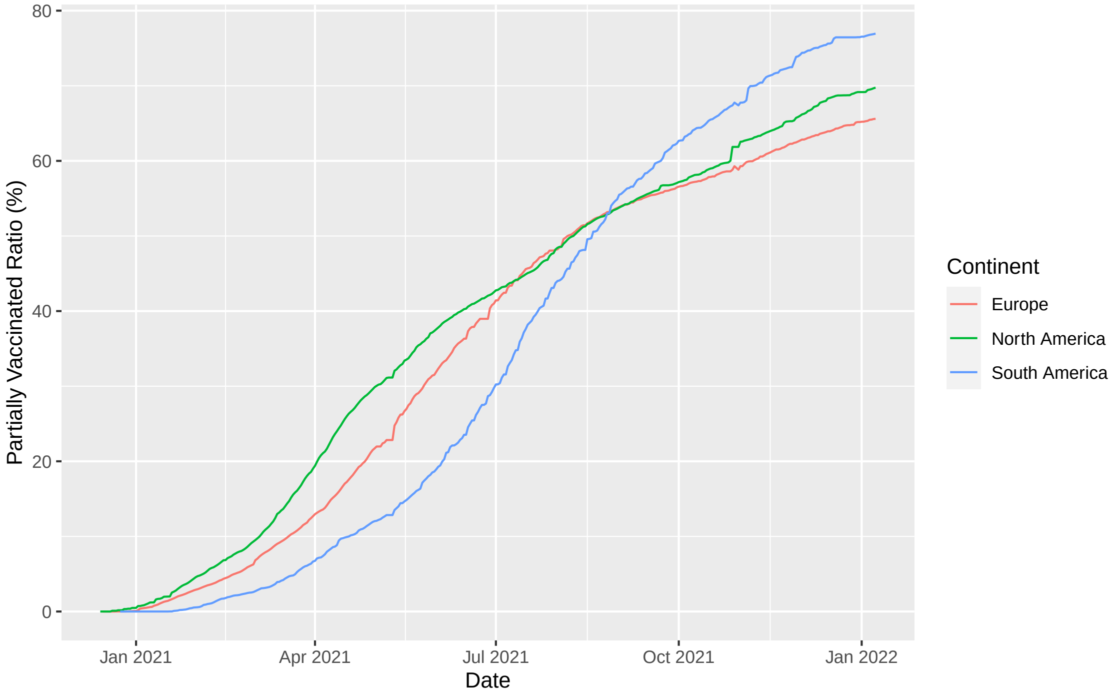

Figure 4: Partially vaccinated ratio per continent

For figures 3 and 4 the same behavior of the population of the three continents can be observed. While vaccines were first made widely available to North America, then Europe and finally South America in early 2021, the coming months South America surpassed the vaccination ratios of the other two continents. People where more willing to get a shot after waiting for it for a longer time. In addition, on all three continents the type of the curve is similar. There is an inflection point after which the vaccination rate is slowing down. Since scientists have estimated that a fully vaccination rate of around 80% is sufficient to end the pandemic and turn it into an endemic illness, the decreasing rate of vaccination in late 2021 and into 2022 shows a potential problem with meeting that goal. What makes things even worse, it has been recently shown that two shots (which for this dataset are sufficient for the population to be considered fully vaccinated) may not be enough after 6 months and that is why in late 2021 most countries began administering a third dose to their population. This dataset is not yet updated with that information to compare the rates with which each continent is getting their booster dose.

4 Part 2 - Comparing European Union Countries

For the second part of this analysis, European countries will be considered. The analysis will consist only of the 27 European Union countries since these countries had access

	Country	Deaths per 100,000	Fully Vaccinated (%)
1	Austria	85	72.91
2	Belgium	77	76.03
3	Bulgaria	345	27.93
4	Croatia	217	52.61
5	Cyprus	44	50.35
6	Czechia	233	62.55
7	Denmark	36	79.94
8	Estonia	130	61.85
9	Finland	19	74.89
10	France	90	76.56
11	Germany	96	71.24
12	Greece	158	68.04
13	Hungary	313	62.17
14	Ireland	75	77.87
15	Italy	107	74.36
16	Latvia	213	66.96
17	Lithuania	210	67.66
18	Luxembourg	68	68.65
19	Malta	61	98.95
20	Netherlands	56	71.54
21	Poland	188	56.00
22	Portugal	119	89.55
23	Romania	225	40.89
24	Slovakia	271	44.66
25	Slovenia	142	57.51
26	Spain	84	81.39
27	Sweden	66	73.58

Table 1: Deaths and Vaccinations of European Union countries

to vaccines at similar dates so their vaccination and death rates can be compared. The second dataset of the coronavirus package will also be analyzed. This data.table contains information on confirmed Covid-19 cases and deaths for each country. End goal will be to show the relation between the vaccination rates of countries to their death rates, in order to confirm the efficacy of vaccines. The same problems with the data explained in part 1 are again addressed before creating plots and tables of interest. The column of interest on the covid19_vaccine dataset is fully_vaccinated_ratio. For the second dataset created (eu_deaths), the most important new column that was creatd is total_deaths_per_100000 (cumulative number of deaths per 100,000 people).

```
#Part 2
#Create data tables for Europe Union for vaccinations and daily deaths
eu_countries = c('Austria', 'Belgium', 'Bulgaria', 'Croatia', 'Cyprus','Czechia',
                 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece',
                 ’Hungary’,’Ireland’, ’Italy’, ’Latvia’, ’Lithuania’,’Luxembourg’,
                 ’Malta’, ’Netherlands’,’Poland’,’Portugal’,’Romania’,’Slovakia’,
                 ’Slovenia’, ’Spain’, ’Sweden’)


eu_vax = covid19_vaccine[country_region %in% eu_countries]
eu_deaths = coronavirus[country %in% eu_countries & type==’death’
                            & date>=’2021-01-01’]

#Keep columns of interest
eu_vax = eu_vax[,.(country_region,date,
           population,fully_vaccinated_ratio,partially_vaccinated_ratio)]
eu_deaths = eu_deaths[,.(date,country,cases,population)]

#calculate cumilative deaths and death ratios
eu_deaths[, daily_deaths_per_100000 := round(100000*cases/population,1)]
eu_deaths[,total_deaths :=cumsum(cases),by=list(country)]
eu_deaths[,total_deaths_per_100000 := round(100000*total_deaths/population,1)]
```

Since the goal is to analyze the effect of vaccinations to the death rates of each country, only deaths after **January 1st, 2021** will be considered (i.e. after vaccines were made available and began being administered). Thus the cumulative number of deaths for each country does not correspond to the whole duration of the pandemic but only after the beginning of 2021. Since all countries are difficult to be plotted on a similar way as was done in part 1 (against the date), bar charts showing the deviations of the two main variables of interest (total deaths per 100,000 and fully vaccinated ratios) on the latest available date of our dataset were created. The data that were used to create the plots are saved on the `latest data.table`. These data points can be also shown on table 1.

```
#Create Data.Table with latest values
latest_vax = eu_vax[date==max(eu_vax$date),.(country_region,
                                             fully_vaccinated_ratio)]
latest_deaths = eu_deaths[date==max(eu_vax$date),.(country,
                                              total_deaths_per_100000)]

latest = merge(latest_deaths,latest_vax, by.x=’country’, by.y=’country_region’)

cor1 = cor(latest$fully_vaccinated_ratio,latest$total_deaths_per_100000)

mean_fully_vax_rate = mean(latest$fully_vaccinated_ratio)
mean_deaths = mean(latest$total_deaths_per_100000)

latest$type_vax <- ifelse(latest$fully_vaccinated_ratio < mean_fully_vax_rate
                          , "below","above")
latest$type_death <- ifelse(latest$total_deaths_per_100000 < mean_deaths
                          , "below","above")

# Diverging Barcharts

latest <- latest[order(latest$fully_vaccinated_ratio), ]
latest$country<- factor(latest$country, levels =latest$country)

ggplot(latest, aes(x=country, y=fully_vaccinated_ratio,
                   label=fully_vaccinated_ratio)) +
  geom_bar(stat=’identity’, aes(fill=type_vax), width=.5) +
  scale_fill_manual(name="Compared to EU average",
                    labels = c("Above Average", "Below Average"),
                    values = c("above"="#00ba38", "below"="#f8766d")) +
  labs(x=’Country’, y=’Ratio (%)’) +
  coord_flip()

latest <- latest[order(-latest$total_deaths_per_100000), ]
latest$country<- factor(latest$country, levels =latest$country)

ggplot(latest, aes(x=country, y=total_deaths_per_100000,
                       label=total_deaths_per_100000)) +
  geom_bar(stat=’identity’, aes(fill=type_death), width=.5) +
  scale_fill_manual(name="Compared to EU average",
                    labels = c("Above Average", "Below Average"),
                    values = c("above"="#f8766d", "below"="#00ba38")) +
  labs(x=’Country’, y=’Deaths per 100,000’) +
  coord_flip()

# Scatterplot

ggplot(latest, aes(x = fully_vaccinated_ratio, y = total_deaths_per_100000)) +
  labs(x=’Ratio (%)’, y=’Deaths per 100,000 after January 1, 2021’)+
  geom_point() +
  geom_smooth(color = "red")
```

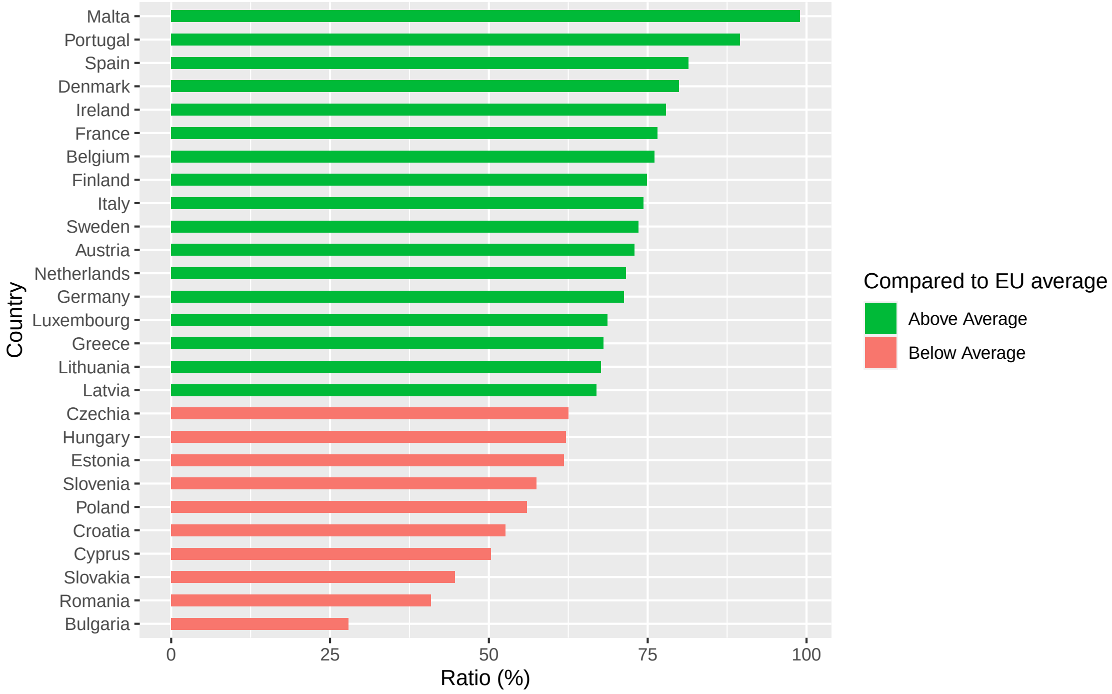

Figure 5: Vaccination Ratios at European Union Countries on January 8, 2022

Figures 5 and 6 show some very interesting results. By plotting the vaccination ratios and deaths of each country, color coded based on whether it’s above or below the EU average, a very important result can be seen. Most of the countries that have below average vaccination rates also have above average death rates (for deaths after January 1st 2021). This is very strong evidence for the efficacy of Covid-19 vaccination. Countries like Bulgaria, Romania, Slovakia and Poland are on the bottom of these two

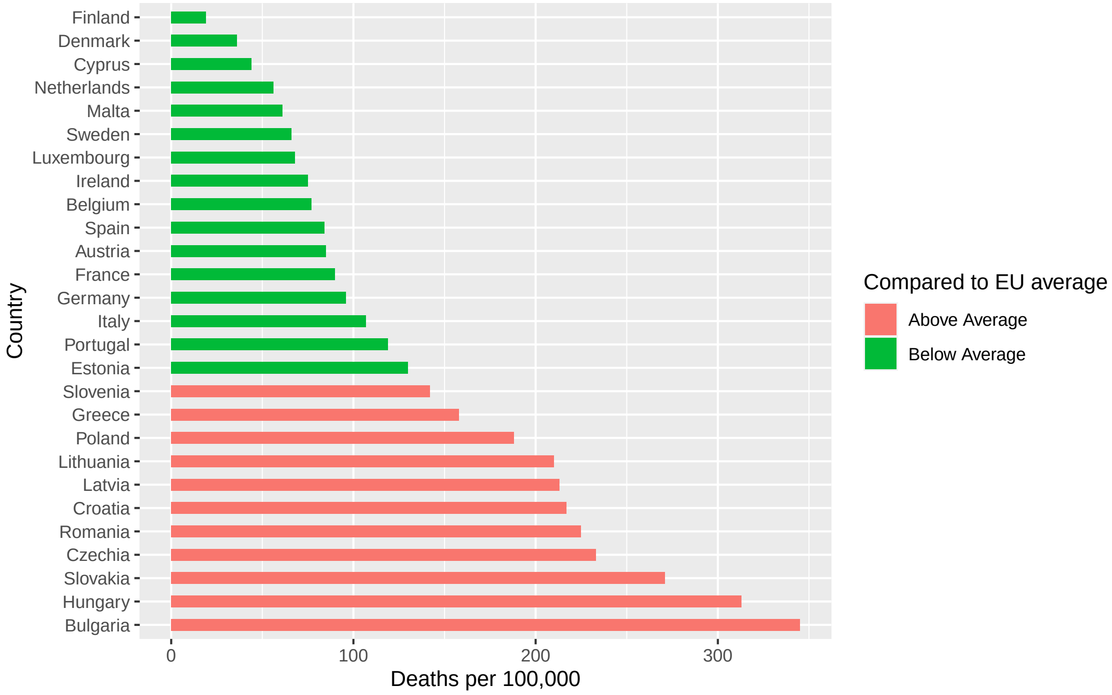

Figure 6: Deaths at European Union Countries after January 1, 2021

figures, whereas countries such as Malta and Denmark at the top. Unfortunately Greece is on the bottom half of both figures, which means that we were not able to vaccinate enough people and had more than the EU average number of deaths in 2021.

Next step is to add a scatter plot that shows the plots the data of Table 1. Figure 7 also has a best fitted line added to the scatter plot of these two values for the 27 data points, where it is easy to see that there is negative correlation between the two variables. The correlation between these two columns is calculated on the script and is equal to **-0.69**.

This part of the analysis was done by considering the total deaths of each country after January 1st, 20201. What is even more interesting is repeating the same analysis, but this time calculating the `total_deaths_per_100000` values of each country not after January 1st 2021, but after a later date. For the last part of this analysis, two different later cutoff dates (deaths after February 1st, 2021 and deaths after October 1st, 2021) were considered and the corresponding scatter plots are shown on figures 8 and 9. February 1st was considered because countries were able to vaccinate most of their vulnerable citizens by that date, whereas by October 1st essentially everybody that wanted to be vaccinated had the opportunity to do so. In addition, the correlation for these two scatter plots decreased to **-0.73** and **-0.76** respectively. This adds to the argument that a high vaccination rate can prevent additional deaths. In other words, countries with high vaccination rates were able to prevent more deaths of their citizens

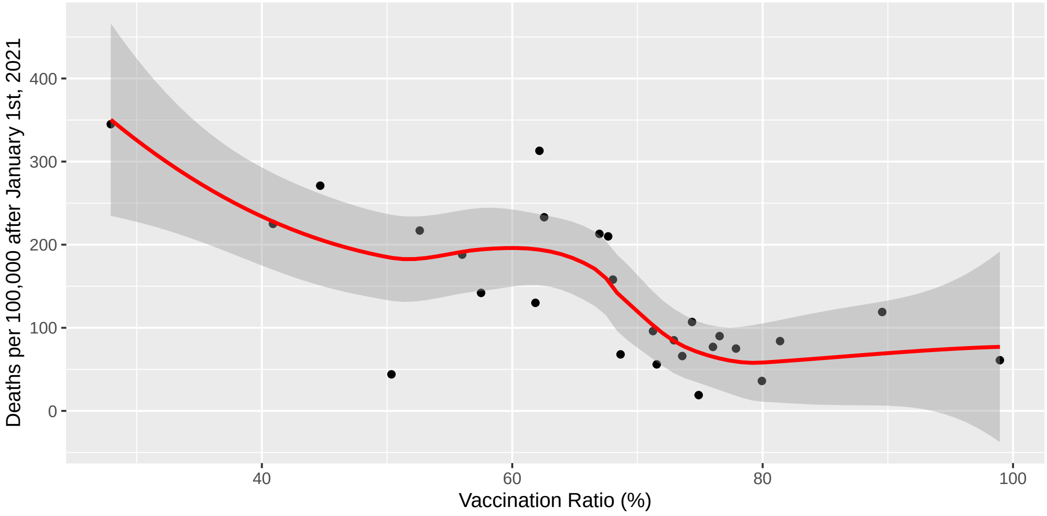

Figure 7: Scatter plot of deaths versus vaccinations for European Union countries

for the latest wave of Covid-19 infections that started on spring and fall of 2021.

```
#Repeat the analysis for Different Dates

#Cutoff Date February 1st

eu_vax = covid19_vaccine[country_region %in% eu_countries]
eu_deaths = coronavirus[country %in% eu_countries & type==’death’
                        & date>=’2021-02-01’]

#calculate cumulative deaths and death ratios
eu_deaths[,total_deaths :=cumsum(cases),by=list(country)]
eu_deaths[,total_deaths_per_100000 := round(100000*total_deaths/population,1)]

#Create data.table with latest values
latest_vax = eu_vax[date==max(eu_vax$date),.(country_region,
                                            fully_vaccinated_ratio)]
latest_deaths = eu_deaths[date==max(eu_vax$date),.(country,
                                                  total_deaths_per_100000)]

latest = merge(latest_deaths,latest_vax, by.x=’country’, by.y=’country_region’)
cor2=cor(latest$fully_vaccinated_ratio,latest$total_deaths_per_100000)

#Scatter Plot

ggplot(latest, aes(x = fully_vaccinated_ratio, y = total_deaths_per_100000)) +
  labs(x=’Vaccination Ratio (%)’, y=’Deaths per 100,000 after February 1st, 2021’)+
  geom_point() +
  geom_smooth(color = "red")

#Cutoff Date October 1st

eu_vax = covid19_vaccine[country_region %in% eu_countries]
eu_deaths = coronavirus[country %in% eu_countries & type==’death’
                        & date>=’2021-10-01’]
```

```
#calculate cumulative deaths and death ratios
eu_deaths[,total_deaths :=cumsum(cases),by=list(country)]
eu_deaths[,total_deaths_per_100000 := round(100000*total_deaths/population,1)]

#Create data.table with latest values
latest_vax = eu_vax[date==max(eu_vax$date),.(country_region,
                                             fully_vaccinated_ratio)]
latest_deaths = eu_deaths[date==max(eu_vax$date),.(country,
                                                   total_deaths_per_100000)]

latest = merge(latest_deaths,latest_vax, by.x=’country’, by.y=’country_region’)
cor3=cor(latest$fully_vaccinated_ratio,latest$total_deaths_per_100000)

#Scatter Plot

ggplot(latest, aes(x = fully_vaccinated_ratio, y = total_deaths_per_100000)) +
  labs(x=’Vaccination Ratio (%)’, y=’Deaths per 100,000 after October 1st, 2021’)+
  geom_point() +
  geom_smooth(color = "red")
```

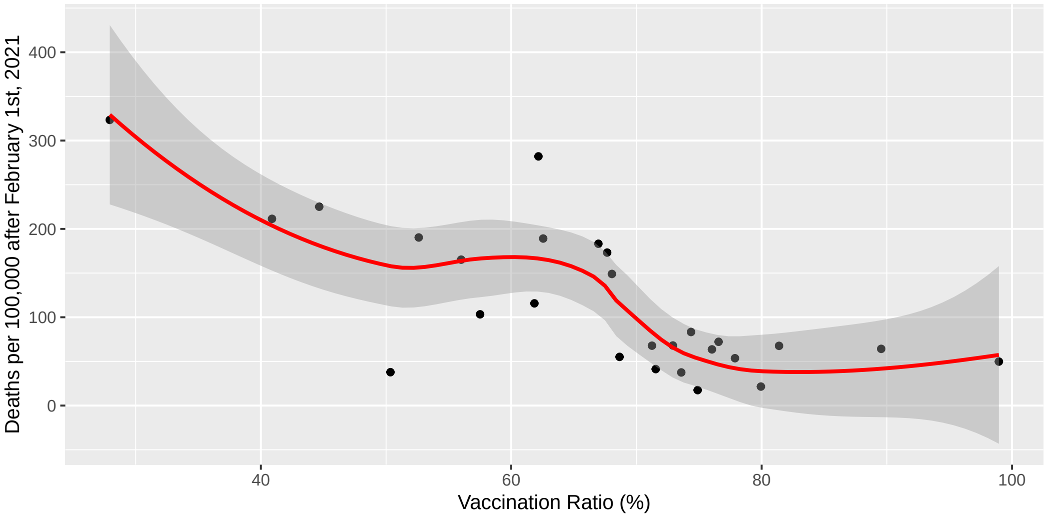

Figure 8: Scatter plot of deaths versus vaccinations for European Union countries

5 Conclusion

For this exploratory data analysis project as part of the graduate course ’Programming Tools and Technologies for Data Science’, the vaccination progress against Covid-19 was analyzed. The data of the `coronavirus` package was analyzed with the R programming language and useful data visualizations were created.

On the first part of the report, results that show the rate of vaccinations at three different continents were presented. The ratio of fully vaccinated people and partially vaccinated people was plotted over time and showed the different rates with which each continent got vaccinated in 2021. South America surpassed Europe and North

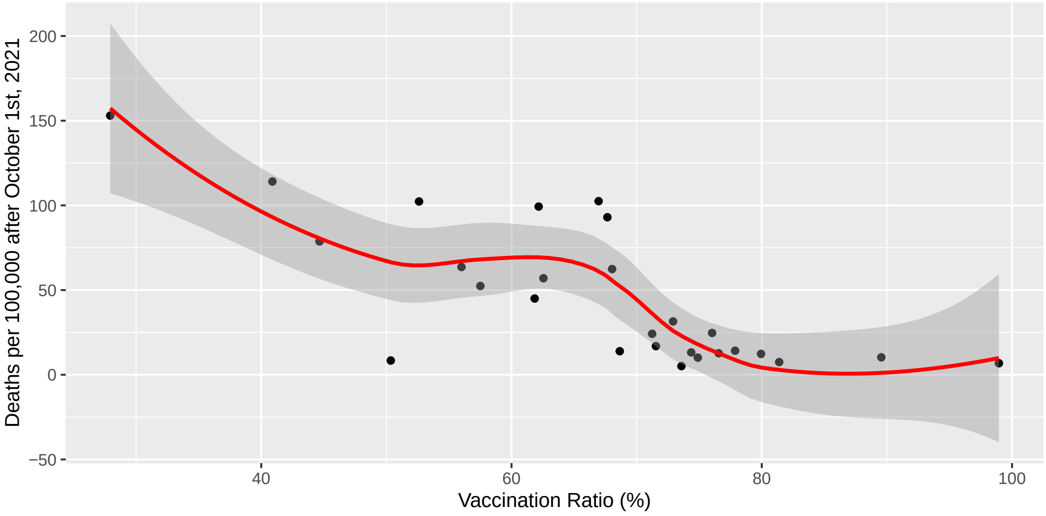

Figure 9: Scatter plot of deaths versus vaccinations for European Union countries

America in vaccination percentages, even though it started the vaccination process at a later date. However, for all three continents the rate of increase has been decreasing over time and it casts doubt to whether these three continents will be able to vaccinate a sufficient amount of people to end the global pandemic.

On the second part of the report, the main interest of the analysis was the European Union. European Union countries were given access to vaccines at similar dates in the beginning of 2021 and started vaccinated their residents with high rates. The analysis compared the latest vaccination rates of each country (on the last date that the dataset was updated) to the number of deaths since the beginning of 2021. Since in 2020 the vaccines were not available, it was not of great interest to include deaths in 2020 in this comparison. The results show that there is a strong negative correlation between the vaccination ratio and the number of deaths of countries in the European Union. The correlation is even stronger when considering deaths starting at a later date, when a sufficient number of people were already vaccinated towards the middle and the end of 2021. It can be seen that countries with higher vaccination rates can now prevent more deaths in the future waves of Covid-19 infections.