National Technical University of Athens

School of Electrical & Computer Engineering

Exploratory Data Analysis using R

Konstantinos Poulinakis, MSc Student of DSML

AM : 03400153 , poulinakis.kon@gmail.com

Project completed as part of the course :

Programming Tools and Technologies for Data Science

January 22, 2022

Introduction

The goal of this paper is to present the insights discovered during the data analysis of the covid19_vaccine dataset available through the R coronavirus package. In the following link you may find information about the dataset. Each section will contain graphs, explanations, conclusions as well as the R code used for the analysis and plotting.

In **section 1** we will introduce the dataset used. We will give some information about it and present the code used for loading and preprocessing. In **section 2** we begin the analysis by exploring the world as a whole. We will see the state of vaccination up until the 6th of January 2022. Plots giving insights about the most vaccinated countries and continents will be shown. In **section 3** we dive deeper, trying to understand why some countries have achieved higher vaccination ratios than others by comparing the vaccination ratios against some important parameters. In **section 4** we uncover the vaccination timeline of 2 different countries by plotting ratios against each month, while also trying to explain some facts through this timeline.

Section 1 : The Dataset

The dataset we have explored is the covid19_vaccination dataset available through the R coronavirus package. The dataset contains information about the covid-19 vaccination status of 191 countries in the world during the pandemic of 2020-2022. The dataset originally contains 141,526 data entries (rows) and 18 features (columns). The dataset timeline begins on 2020-12-14, around the time the first vaccine doses were administered, and has been updated to contain data up to 2022-01-06. Every graph and conclusion in this project depends on data that had been collected up to January 6 of 2022.

The datasets's features include information such as the country's population, fully vaccinated people, partially vaccinated people, doses administered and the reported date. It also contains geographical information such as the continent it belongs to, the country's longitude and latitude. It also contains some other redundant features which we are going to discard eg. iso codes. We also clean the dataset by omitting rows that contain missing values (NA) in any of the remaining columns, in order to avoid problems in the process.

After the cleaning we are left with 45901 rows.

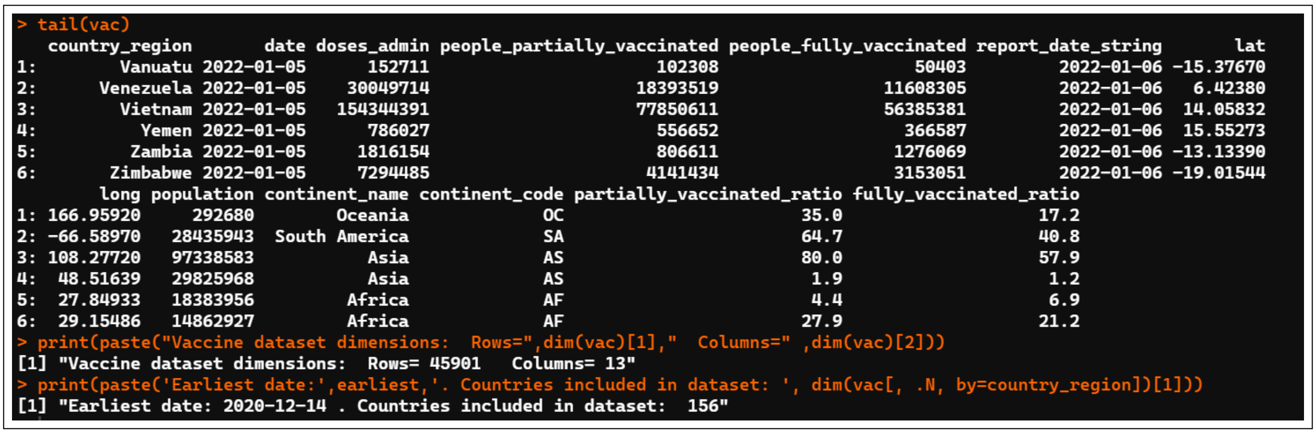
Listing 1: Data Loading

```
library(coronavirus)
library(data.table)
library(ggplot2)
library(dplyr)
library("plyr")
library(tidyverse)

# Read the csv data and print some usefull information
vac = read.csv("␣...␣\\covid19_vaccine.csv",header=TRUE, stringsAsFactors = FALSE)
vac = data.table(vac)
print(paste("Vaccine␣dataset␣dimensions:␣␣Rows=",dim(vac)[1],"Columns=" ,dim(vac)[2]))
earliest = min(vac[, date])
print(paste('Earliest␣date:',earliest,'.Countries␣included␣in␣dataset:',
      dim(vac[, .N, by=country_region])[1]))
```

Listing 2: Data cleaning and adding new column entries

```
# Delete unused columns
vac[, c("code3","fips", "province_state", "iso2", "iso3", "uid",
        "combined_key"):=NULL]
vac = na.omit(vac)
omit = na.omit(vac, invert=FALSE)
#Create new entries
vac[,'part_vac_ratio'] = vac[,'people_part_vac']/vac[,'population']
vac[,'fully_vac_ratio'] = vac[,'people_fully_vac']/vac[,'population']
# Convert to percentages
vac[,'part_vac_ratio'] = vac[, round(part_vac_ratio,3)*100]
vac[,'fully_vac_ratio'] = vac[, round(fully_vac_ratio,3)*100]
```



```
> tail(vac)
   country_region       date doses_admin people_partially_vaccinated people_fully_vaccinated report_date_string       lat
1:        Vanuatu 2022-01-05      152711                      102308                   50403         2022-01-06 -15.37670
2:      Venezuela 2022-01-05    30049714                    18393519                11608305         2022-01-06   6.42380
3:        Vietnam 2022-01-05   154344391                    77850611                56385381         2022-01-06  14.05832
4:          Yemen 2022-01-05      786027                      556652                  366587         2022-01-06  15.55273
5:         Zambia 2022-01-05     1816154                      806611                 1276069         2022-01-06 -13.13390
6:       Zimbabwe 2022-01-05     7294485                     4141434                 3153051         2022-01-06 -19.01544
        long population continent_name continent_code partially_vaccinated_ratio fully_vaccinated_ratio
1: 166.95920     292680        Oceania             OC                       35.0                   17.2
2: -66.58970   28435943  South America             SA                       64.7                   40.8
3: 108.27720   97338583           Asia             AS                       80.0                   57.9
4:  48.51639   29825968           Asia             AS                        1.9                    1.2
5:  27.84933   18383956         Africa             AF                        4.4                    6.9
6:  29.15486   14862927         Africa             AF                       27.9                   21.2
> print(paste("Vaccine dataset dimensions:  Rows=",dim(vac)[1],"  Columns=" ,dim(vac)[2]))
[1] "Vaccine dataset dimensions:  Rows= 45901   Columns= 13"
> print(paste('Earliest date:',earliest,'. Countries included in dataset: ', dim(vac[, .N, by=country_region])[1]))
[1] "Earliest date: 2020-12-14 . Countries included in dataset:  156"
```

Figure 1: Dataset after dropping redundant columns and omitting NA values

On the code above, we created two new column entries, partially_vaccinated_ratio and fully_vaccinated_ratio. These two metrics are derived as the fraction between a country's fully vaccinated residents number and its population . These two variable are going to be used to provide useful insights throughout our analysis.

$$Fully\,vaccinated\,ratio = \frac{people\,fully\,vaccinated}{total\,population}$$

Section 2: The bigger picture

Having pre-processed and cleaned our dataset, it's time to begin the exploratory analysis. We will try and get the bigger picture before diving deeper into the details. Exploring the world vaccination status is our first goal.

We begin with something simple but important. We group our data table by the most recent date (06-01-2022) and order it by the variable fully-vaccinated-ratio. We choose the 25 most vaccinated countries and plot them on a well curated bar plot using ggplot2 library.

Listing 3: Bar plot of the most vaccinated countries using ggplot2

```r
# Group by country, sort by fully vaccinated ratio.
vac_ratios = vac[report_date_string==last_day,
                 .(country_region, continent_name, fully_vaccinated_ratio,
                 partially_vaccinated_ratio, population),
                 keyby=fully_vaccinated_ratio]
vac_ratios = vac_ratios[,-1] # remove dublicate column created by keyby expression
# Grab the 25 most vaccinated.
tail_r = tail(vac_ratios,25)
# Collapse the two variable ratios into a single df
cr_df <- tail_r %>%
  select(country_region, partially_vaccinated_ratio, fully_vaccinated_ratio) %>%
  gather(key="variable", value="value", -country_region)
# Plot
ggplot(cr_df, aes( x=value, y=country_region, fill=variable)) +
  geom_bar(position="dodge",stat="identity", width=0.7) +
  theme_gray()+
  theme(axis.title.y = element_text(size = 11, face="bold"),
        axis.title.x = element_text(size = 11, face="bold"),
        title = element_text(size = 12, face="bold")) +
  scale_fill_manual("Ratio␣type", values = c('navyblue','goldenrod1')) +
  scale_x_continuous(breaks = round(seq(0, 100, by = 10),1)) +
  ggtitle("Top␣25␣Most␣Vaccinated␣Countries")+
  xlab("Continent")+ ylab('Vaccination␣Ratio␣(%)')+
  labs(caption = "(based␣on␣data␣from␣covid19_vaccine␣dataset)",
       plot.caption.position = "bottomright")
```

The resulting bar plot can be seen in figure 2 below.

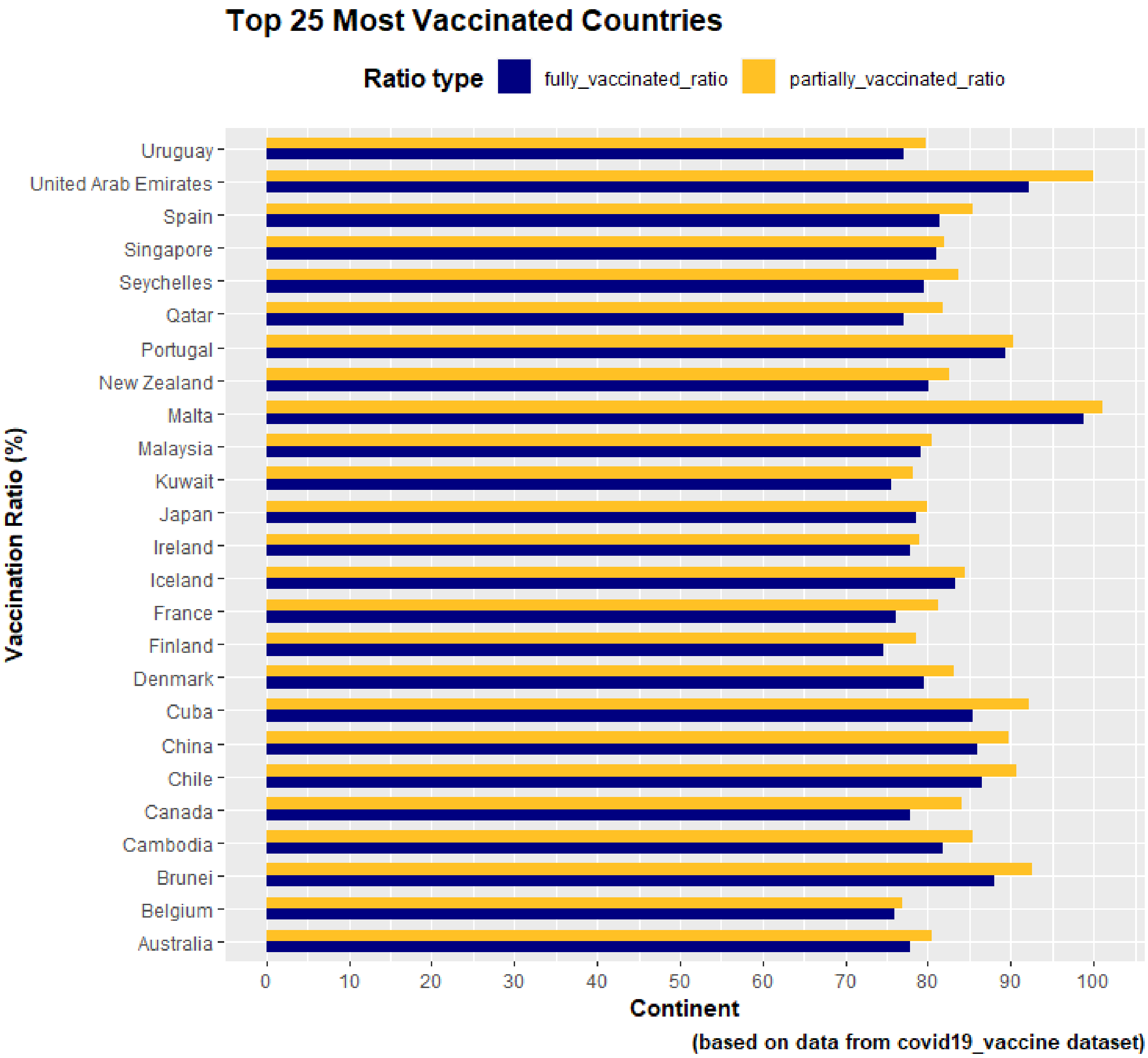


Figure 2: The 25 countries with the highest fully vaccinated ratio. Malta, United Arab Emirates and Portugal forming the top 3.

Malta, UAEB, Portugal, Brunei and Chile make up the top 5 countries. What is very surprising is the complete absence of USA from this top 25 list. China, despite it's tremendous population of 1.5 billion people has made it very high in the list on numher 6.

It is interesting to see how these 25 countries, that have had the most successful vaccine campaigns, are distributed across the continents. Pie chart is the most suitable graph for such a task. Afterwards we also plot the distribution of the 25 least vaccinated countries.

From the first pie chart, fig 3, we observe that the majority of the most vaccinated countries are located in Asia and Europe. Both continents share the top with 36%, meaning 9 countries of these 25 countries belong to each. Following we have Oceania, South and North America with 2 countries each and lastly 1 country from Africa, Seychelles.

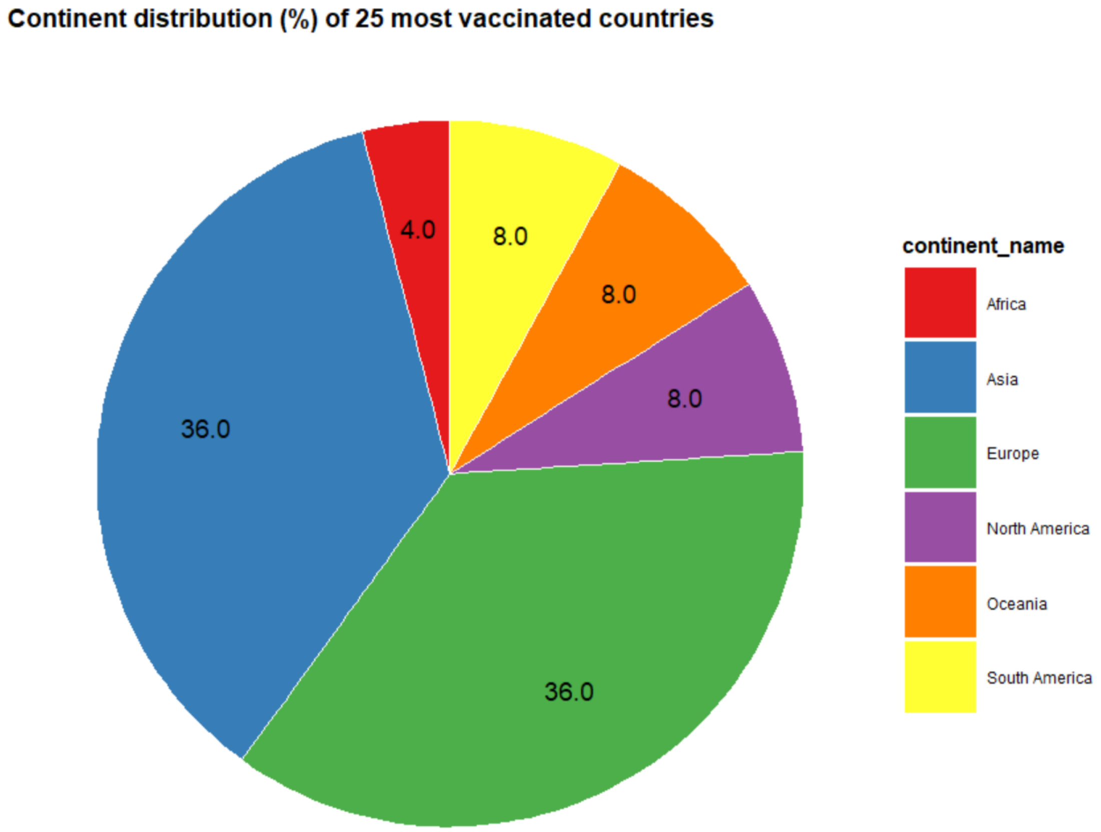


Figure 3: Continent distribution of the 25 most vaccinated countries.

In the second pie chart, figure 4, we observe without surprise that an enormous 84% of the least vaccinated countries are located in Africa. This is followed by Asia's 8%, 2 countries, Oceania's and North America's 1 country, 4%. This comes as no surprise, as it is a commonly discussed issue that poorer countries do not have equal access to vaccines compared to the so called ”developed” countries. A vaccination-ratio regressed on GDP (Gross Domestic Product) would probably result in very similar results.

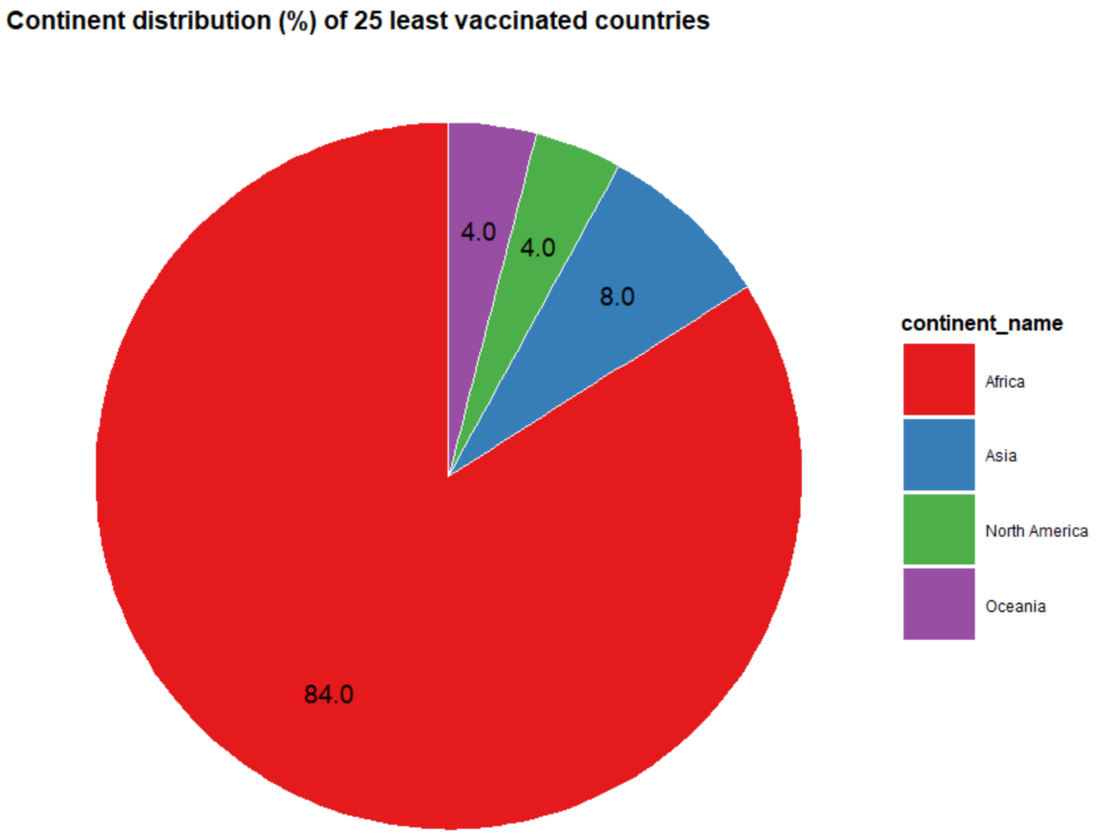


Figure 4: Continent distribution of the 25 least vaccinated countries.

Listing 4: Pie plots with ggplot2

```
make_pie_chart <- function(data, title){
    # Compute the position of labels
    data <- data %>%
    arrange(desc(continent_name)) %>%
    mutate(prop = N / sum(data$N) *100.0) %>%
    mutate(ypos = cumsum(prop)- 0.5*prop )
    # Plot
    ggplot(data, aes( x='', y=prop, fill=continent_name)) +
    geom_bar(stat="identity", width=1, color='white') +
    theme_void()+
    theme(title = element_text(size = 12, face="bold"),
          legend.key.size = unit(1.5, 'cm')) +
    coord_polar("y", start=0) +
    ggtitle(title) +
    geom_text(aes(y = ypos, x=1.2, label = sprintf('%1.1f',prop)),
              color = "white", size=5) +
    scale_fill_brewer(palette="Set1") }

freq = tail_r[, .N, by=continent_name]
make_pie_chart(data=freq, title='Continent␣distribution(%)␣of␣25
most␣vaccinated␣countries')
```

Observation of the above top-25 and bottom-25 pie charts would lead somebody to make the assumption that Europe and Asia are continents with very high ratios whilst Oceania and the American continents aren't so high on the list. One would also conclude that Africa is doing really poorly in terms of vaccination.

However, to properly answer this question we need to study the vaccination ratio of each continent by computing the necessary ratios for each continent specifically. To do that, we calculate the population of each continent from our data, sum up the fully/partially vaccinated people residing in each continent and calculate the ratios. This way we calculate what population fraction of each continent is vaccinated. With the following code we do exactly that and proceed to create a bar plot to convey our results.

Listing 5: Creating continent vaccination ratios

```
c_ratio = vac[report_date_string==last_day,
             .(population,people_fully_vac, people_part_vac),
             by=continent_name]
# Sum up population and vaccinated people for every continent.
c_ratio = c_ratio[, lapply(.SD, sum), by=.(continent_name)]
# Calculate the vaccination ratios as percentages.
c_ratio[,'fully_vac_ratio'] = c_ratio[,'people_fully_vac']/c_ratio[,'population']
c_ratio[,'fully_vac_ratio'] = c_ratio[, round(fully_vac_ratio,3)*100]
c_ratio[,'part_vac_ratio'] = c_ratio[,'people_part_vac']/c_ratio[,'population']
c_ratio[,'part_vac_ratio'] = c_ratio[, round(part_vac_ratio,3)*100]
c_ratio
# Collapse the two variable ratios into a single df
c_df <- c_ratio %>%
  select(continent_name, part_vac_ratio, fully_vac_ratio) %>%
  gather(key="variable", value="value", -continent_name)
```

The bar plot below emphasizes how, sometimes, many graphs and statistics might mislead us into making false assumptions about our world. Judging from the pie charts, one would assume that South America is not a continent with a very successful vaccination campaign as it does not host a big fraction of the top-25 most vaccinated countries. However, from the bar plot in fig 5 we can clearly see that **South America is the most vaccinated continent in terms of** $\frac{vaccinated\ people}{population}$ **!** Europe follows second with Oceania and North America being close thirds. Surprisingly, **Asia falls in the** 4^{th} **place, something one wouldn't expect after studying the pie charts!** This is probably due to the enormous population of some low vaccinated Asian countries, like India and Pakistan, both around the 40% mark. Africa is indeed a continent with very low vaccination rate and Seychelles was an outlier that made it to the top 25 list.

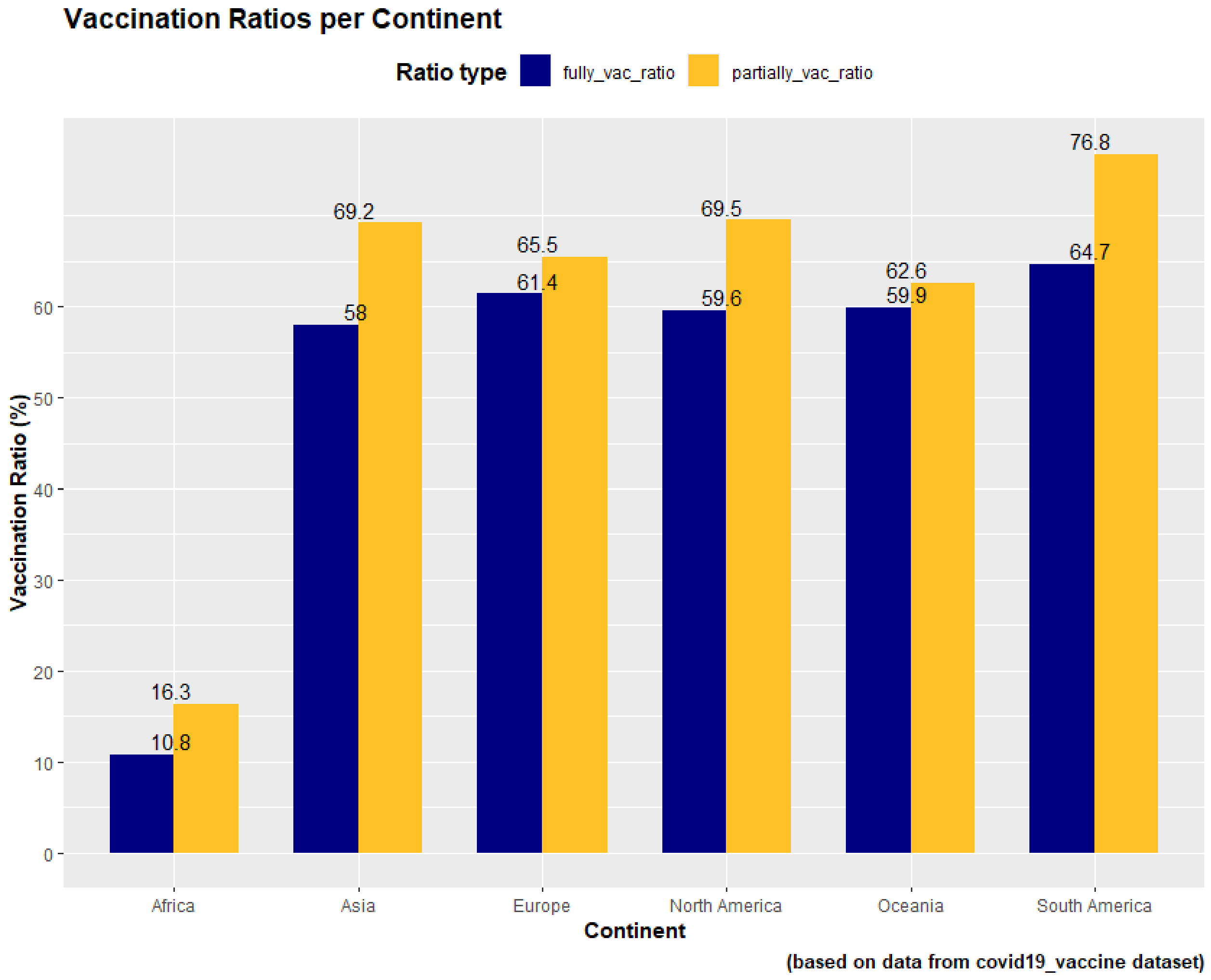


Figure 5: Vaccination ratios of each continent. This offers a more educated look, as the pie chart in 3 could have easily misled somebody. South America is the most vaccinated continent.

Section 3: Vaccination Ratio and Population

As we saw in section 2, the continent a country is located in, can predict its vaccination campaign success up to a point. Most of the European and Latin countries share very high vaccination rates whilst African countries populate the ”bottom of the barrel”. In this section we are going to explore another, possibly, important parameter of vaccination rate. This important parameter is population. As logic suggests, a high population is going to be more difficult to vaccinate than a lower population is. ***Considering the raw amount of available doses needed to vaccinate a country, the specialised stuff (doctors, nurses) that complete the necessary operations, the buildings that need***

to be allocated for this sole purpose and the expenses that arise, one might quickly jump in the conclusion that the lower a country's population is the higher its vaccine ratio is going to be. **In other words, we expect an inverse relationship between population and vaccinated ratio.**

To explore this question we start by gathering the fully vaccinated ratio and the population on the 6^{th} of January 2022 for any country that has a vaccination ratio above 60%. Countries with lower vaccination ratio are excluded as other parameters are probably affecting the vaccination status more than the population does, think of African countries whose economic situation or long wars are disrupting even more vital aspects of their society let alone vaccination campaigns.

In the first graph, fig 6 we present a scatter plot while also plotting the moving average value with a smoothed line. This allows to quickly sense how the vaccination ratio moves while population increases.

By examining figure 6 we see that vaccination ratio tends to be higher on lower populated countries, with population under 20 million. After that population mark we start observing a decline. Although, having excluded China from the graph for graphical reasons we fail to observe the big increase that would happen to this outlier which has around 1.5 billion population and a staggering 86% ratio. So, before hasting into verifying our initial inverse relationship hypothesis it would be useful to get a "second opinion" with a whisker plot (box plot). We plot a whisker plot for the same data in figure 7

On the box plot, green dots represent the observations. Dots that belond in the same 10 million group are grouped together into a box. The box length represents the 1^{st} and 3^{rd} Quantiles (25% and 75%) while the line inside the box represents the median value of the group. The line that streches above and below the box gives us a good sense of the variance the group has, as the upmost and lowest points of the line represent points within 2 standard deviations. Points further than that are ploted as outliers represented with open circles. Malta, which belongs in the first group is such an outlier. Lastly, the width of the box increases the more distributed the values of its group are. If a group has only a single member, then the box converges into a single point. US at 330 million and Japan at 130 million are examples of single point boxes.

So, with all that being said for the box plot (fig. 7), it is obvious that there is big

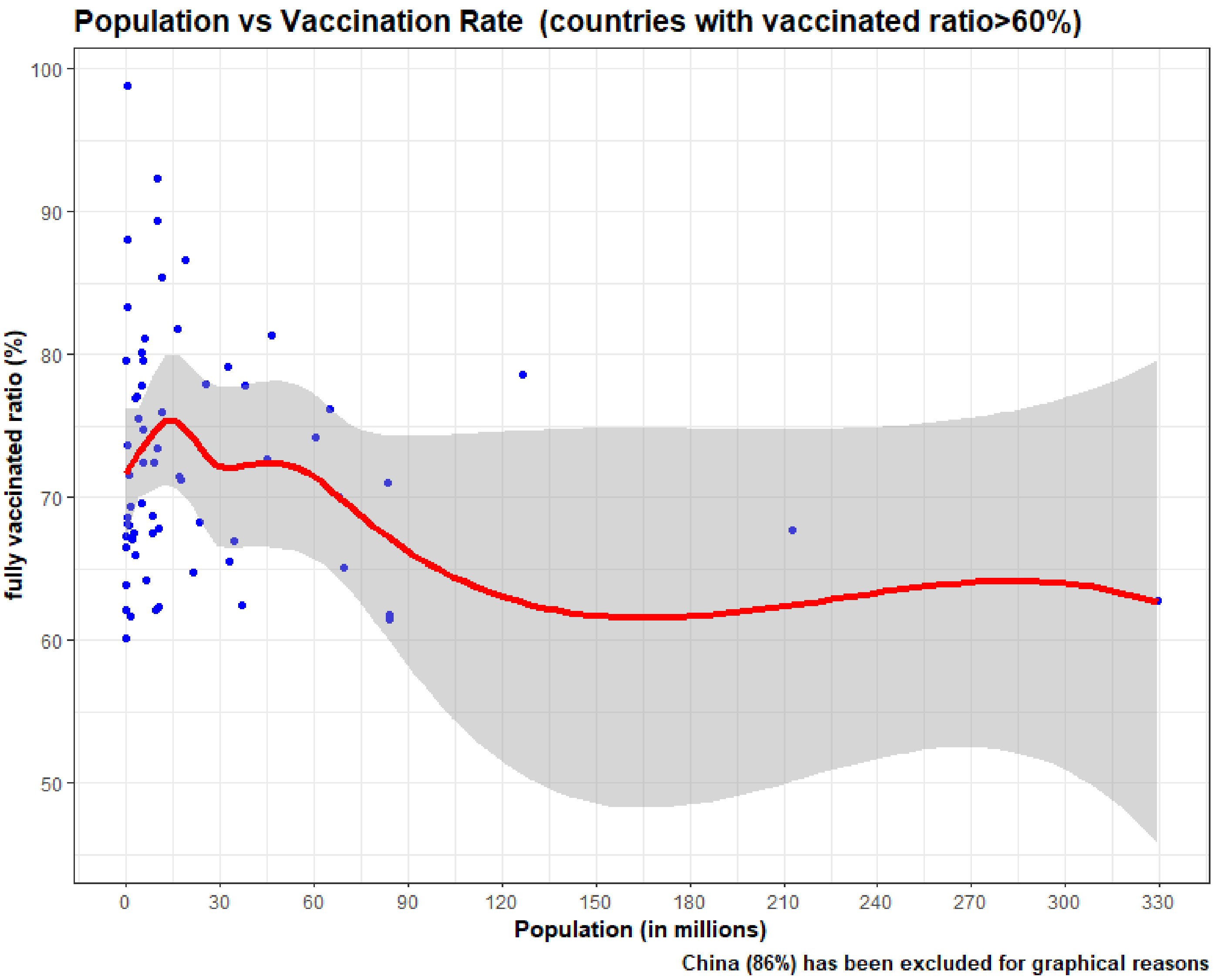


Figure 6: Vaccination ratio vs Population for countries with ratio > 60%. China is excluded for practical reasons.

variance in the vaccination ratio when only population is taken into account. Even though countries below 20 million are well represented in the graph, these groups share very big variance. There are also outlier points which drive the group upwards. I would be hesitant to suggest that a low population is reason for higher vaccination ratio due to the high variance in these groups. Moving from 20 million to 45 million We observe an upward trend, with lower variance,. However, a downward trend begins after 45 million which holds until we reach US at 330 million population, with Japan at 130 million population acting as an outlier in this trend.

To sum up, it does not seem safe to suggest that there is a statistically significant inverse relationship between a country's population and its vaccination ratio. Even though some

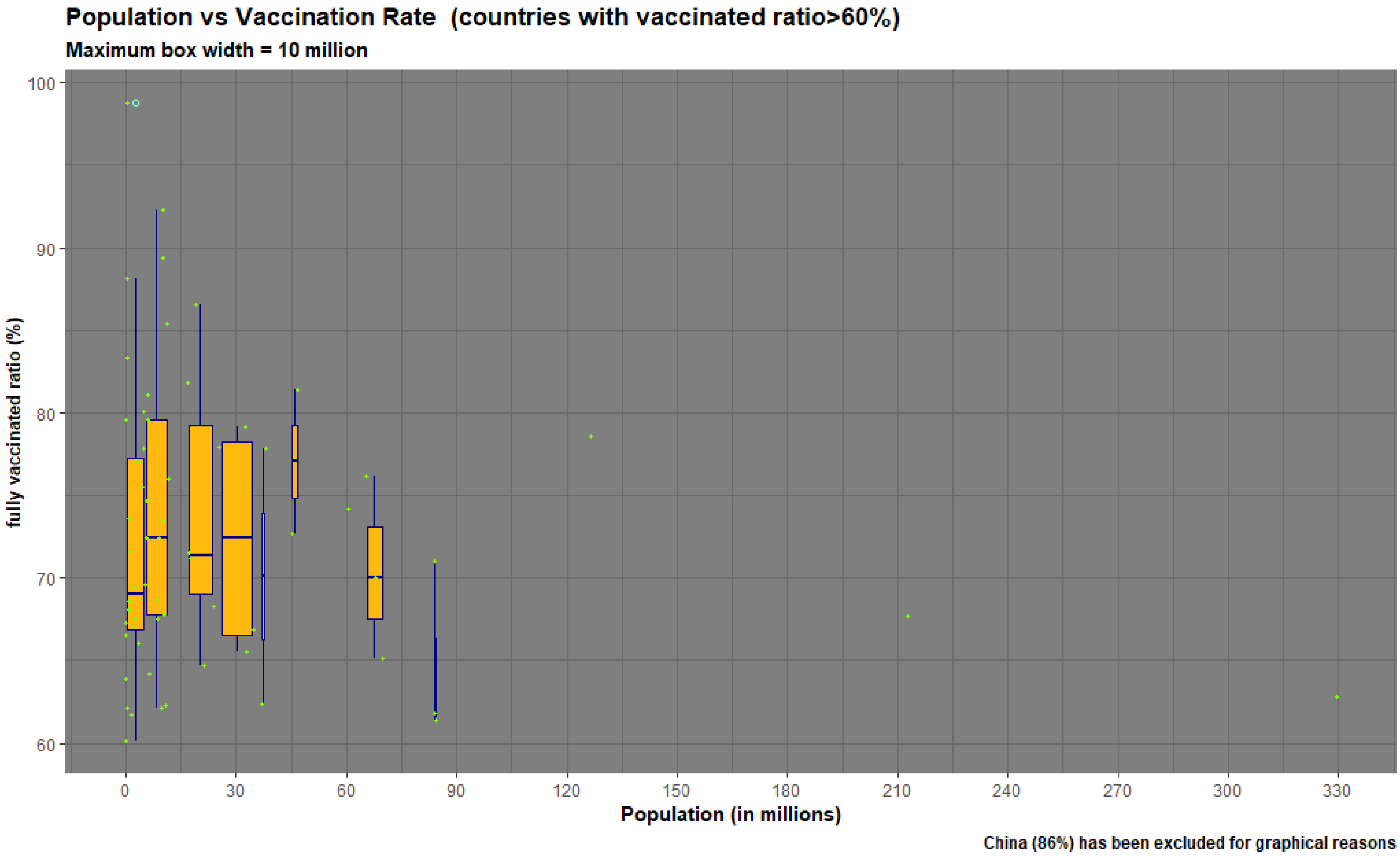


Figure 7: Whisker plot: Vaccination ratio vs Population for countries with ratio > 60%. China is excluded for practical reasons.

ranges eg 20 to 45 million, seem to be benefited by the increase, other ranges, eg 45 million and onwards, seem to be harmed by the increased population. Also the appearance of outliers in those trends eg, Japan, China further deconstructs our initial inverse relationship hypothesis.

To conclude, I would suggest that there is no statistically significant relationship between a country's population and its vaccination ratio. The two variables might as well be independent up to a degree. Other parameters like geographical region and GDP would probably be better predictors than population.

Listing 6: Code comparing population vs ratios.

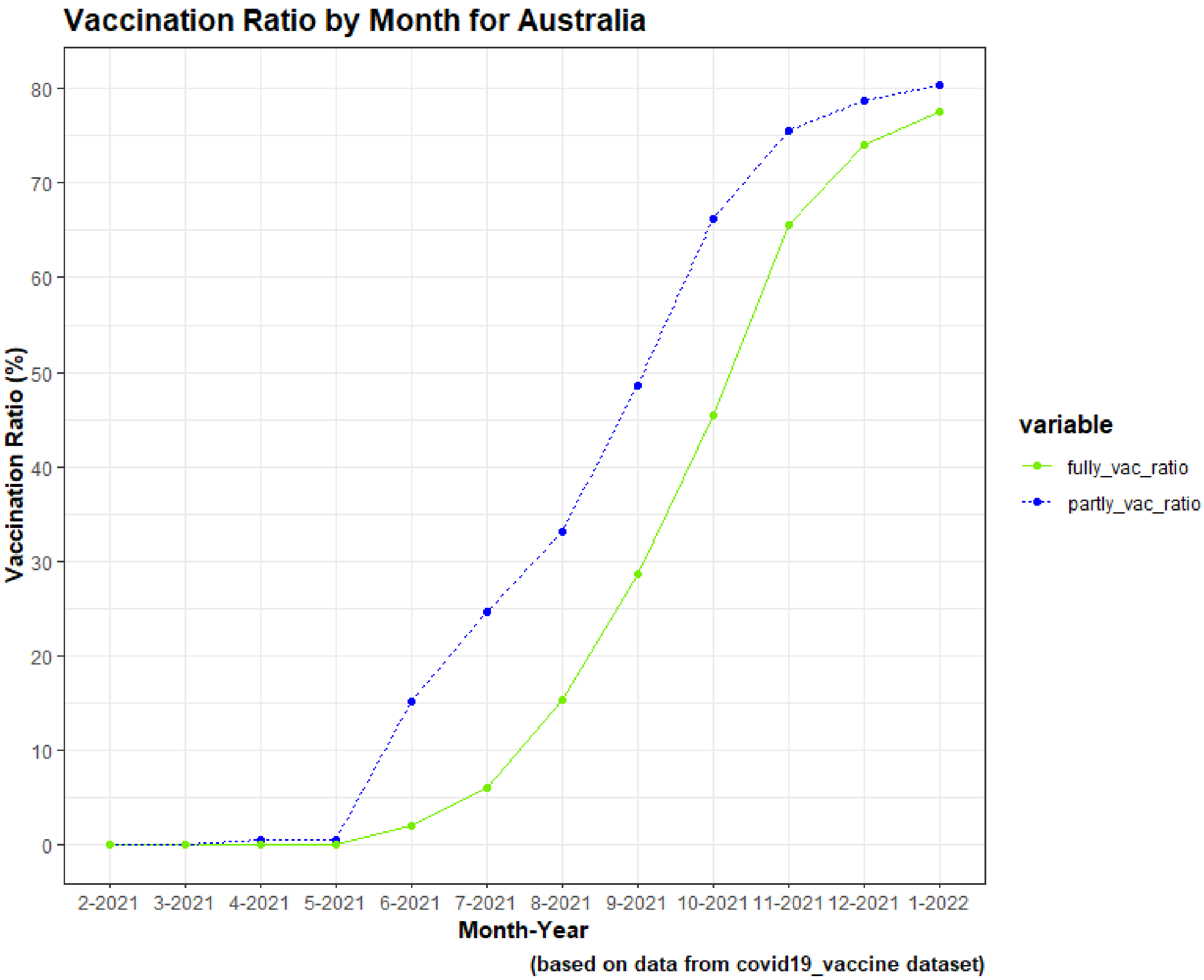
```
# exclude population outlier china and keep ratio>68%
pop = vac[report_date_string==last_day & country_region!='China' & fully_vaccinated_ratio>60,
    .(country_region, fully_vaccinated_ratio, partially_vaccinated_ratio),keyby=(population/10^6)]
# Plot
ggplot(pop, aes(x=(population), y=fully_vaccinated_ratio)) +
  geom_point(color='blue')+
  geom_smooth(color='red', size=1.5) +
  scale_x_continuous(breaks = round(seq(0, 360, by = 30),1)) +
  scale_y_continuous(breaks = round(seq(0, 100, by = 10),1)) +
  theme_bw() +
  theme(axis.title.y = element_text(size = 11, face="bold"),
        axis.title.x = element_text(size = 11, face="bold"),
        title = element_text(size = 12, face="bold")) +
  ggtitle("Population␣vs␣Vaccination␣Rate␣␣(countries␣with␣vaccinated␣ratio>60%)")+
  xlab("Population␣(in␣millions)")+ ylab("fully␣vaccinated␣ratio␣(%)") +
  labs(caption = "China␣(86%)␣has␣been␣excluded␣for␣graphical␣reasons")
## Whisker plot
ggplot(pop, aes(x=(population), y=fully_vaccinated_ratio)) +
  geom_boxplot(aes(group = cut_width(population, 10)),fill = "darkgoldenrod1",
                colour = "navyblue",outlier.colour = "aquamarine",
                outlier.shape = 1)+
  geom_point(size=0.8, color='chartreuse1') +
  scale_x_continuous(breaks = round(seq(0, 360, by = 30),1)) +
  scale_y_continuous(breaks = round(seq(0, 100, by = 10),1)) +
  theme_dark() +
  theme(axis.title.y = element_text(size = 10, face="bold"),
        axis.title.x = element_text(size = 11, face="bold"),
        title = element_text(size = 11, face="bold")) +
  ggtitle("Population␣vs␣Vaccination␣Rate␣␣(countries␣with␣vaccinated␣ratio>60%)")+
  xlab("Population␣(in␣millions)")+ ylab("fully␣vaccinated␣ratio␣(%)") +
  labs(subtitle="Maximum␣box␣width␣=␣10␣million",
  caption = "China␣(86%)␣has␣been␣excluded␣for␣graphical␣reasons")
```

Section 4: Seasonality

In this 4*th* and last section of the analysis we explore another critical relationship, that of time and ratio. We will study how the vaccination ratios have developed over the course of 2021. For this purpose we plot the vaccination ratios at the start of each month and observe through a trend line the movement of the ratio variables.

We explore this relationship for Greece and Australia. Obviously only a positive relationship is to be expected between ratio and time, but by observing the months during which vaccination was more or less active we might come into some interesting conclusions. Greece is doing relatively well and is a a country heavily dependent on tourism during the summer months. Australia has had a great campaign, hitting the top-20 countries, but has introduced some very controversial measures like geo-location with face recognition. We explore possible connections between these facts and vaccination ratios.

Australia and the controversial measures .



(based on data from covid19_vaccine dataset)

Figure 8: Australia's vaccination ratios over the course of time.

Australia, an otherwise very free community, has introduced some very harsh measures to fight the spread of Covid-19. In late August 2021, the state of South Australia launched an app with facial recognition software that Australians subject to mandatory 14-day quarantines can opt to use in lieu of being quarantined at a hotel under police guard. The app randomly prompts users to take a picture of their face and submit geo-location data within 15 minutes of the prompt to prove to the South Australian government that the user is in an approved location. Users who refuse to comply or who fail to respond to a prompt within 15 minutes are checked on by local police and may be subject to fines. **Also, Melbourne was under a very long lockdown that officially ended in mid October**

2021.

After observation of figure 8, we can see that Australia was late compared to other countries to begin vaccinations. Essentially vaccinations began during May. After a slow start we see a rapid increase during August, September and October. During those 3 months alone Australia's ratio jumped from 15% to 65%, an extraordinary 50% of its population got vaccinated in just 3 months. After that vaccination speed seems to plateau. **Those 3 months of rapid vaccination seem to coincide with the ending of Melbourne's long lockdown at late October.** Also the late onset of vaccinations might be a way of justifying the controversial measure of geo-location and face recognition enforced during late August when vaccination ratio was still under 25%.

Listing 7: Plotting ratios over months

```
# Get data grouped by month for australia
seasonality = australia[, .(doses_admin,population,people_partially_vaccinated,
                            people_fully_vaccinated),
                            by=.(Month = paste0(month(date),'-',year(date)))]
seasonality[,'partly_vac_ratio'] = seasonality[,people_partially_vaccinated]/
                                                       seasonality[,'population']
seasonality[,'partly_vac_ratio'] = round(seasonality[,'partly_vac_ratio'],3)*100
seasonality[,'fully_vac_ratio'] = seasonality[,people_fully_vaccinated]/
                                                      seasonality[,'population']
seasonality[,'fully_vac_ratio'] = round(seasonality[,'fully_vac_ratio'],3)*100
# Keep only the first observation from each month
seasonality = seasonality[, .SD[c(1)], by=Month]
# Define level order (ggplot breaks the month order)
level_order <- c('12-2020','1-2021','2-2021','3-2021','4-2021','5-2021','6-2021',
                 '7-2021','8-2021','9-2021','10-2021','11-2021','12-2021','1-2022')
# Collapse the two variable ratios into a single df
df <- seasonality %>%
  select(Month, partly_vac_ratio, fully_vac_ratio) %>%
  gather(key="variable", value="value", -Month)
# Plot
ggplot(df, aes(x=factor(Month, level=level_order), y=value, group=variable))+
  geom_line(aes(color=variable, linetype=variable)) + #plot the line
  geom_point(aes(color=variable),size=1.5) + #plot points
  scale_color_manual(values = c("chartreuse2", "blue1")) + #pick colors
  scale_y_continuous(breaks = round(seq(0, 100, by = 10),1)) +
  ylab('Vaccination␣Ratio␣(%)') + xlab('Month-Year') +
  ggtitle('Vaccination␣Ratio␣by␣Month␣for␣Australia')+
  labs(caption = "(based␣on␣data␣from␣covid19_vaccine␣dataset)") +
  theme_bw() +
  theme(axis.title.y = element_text(size = 11, face="bold"), #theme decoration
        axis.title.x = element_text(size = 11, face="bold"),
        title = element_text(size = 12, face="bold"))
```

Greece and summer tourism.

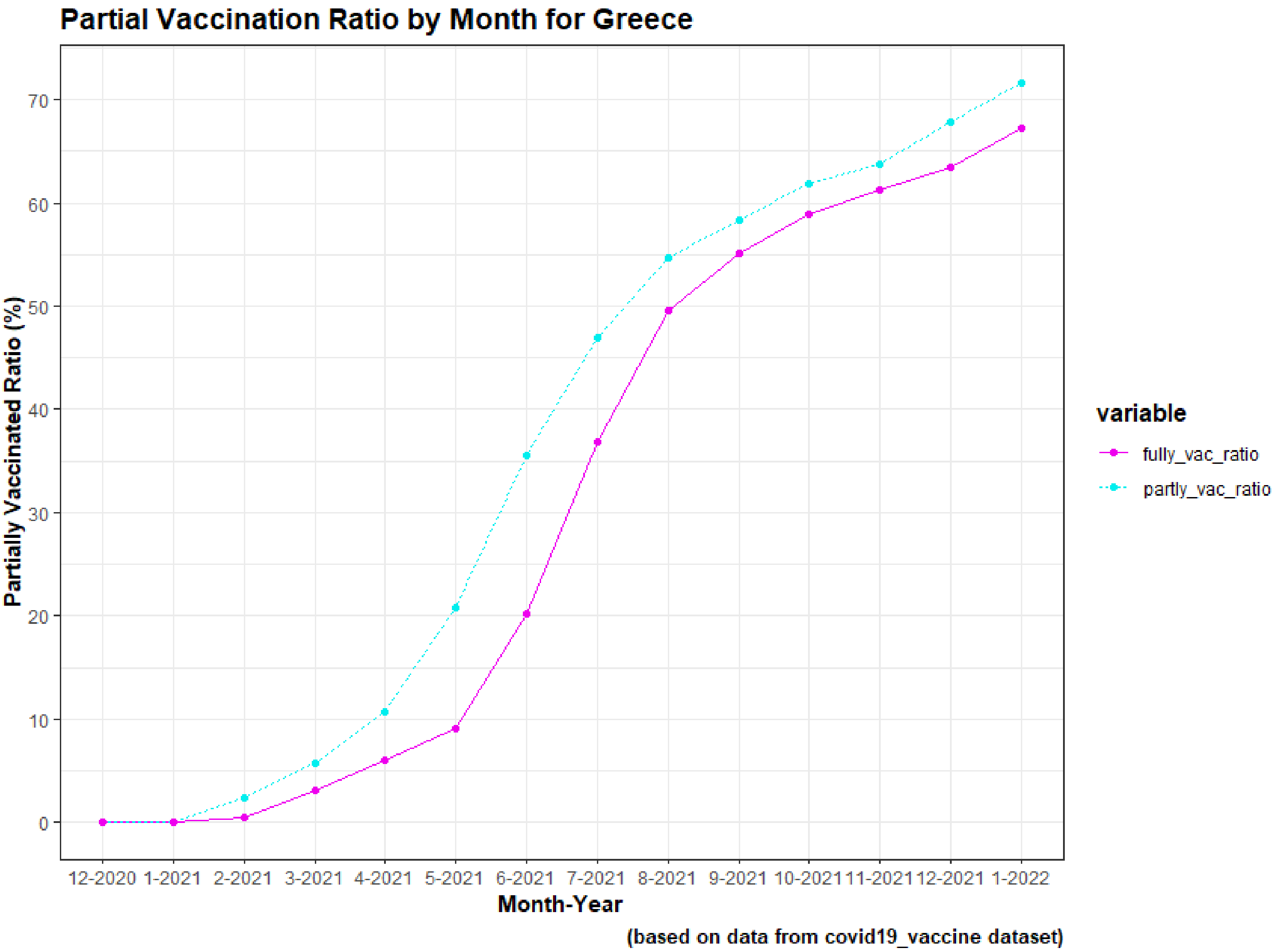


Figure 9: Greece's vaccination ratios over the course of time.

Greece is a very tourism dependent country, especially during summer months. Also, Greeks during the summer tend to travel a lot to the islands and nightlife is very active, especially for the youth. During the month of June the government imposed strict measures against the non-vaccinated, mandating negative rapid or PCR tests in order to enter restaurants, public buildings and travel to any island.

After careful observation of figure 9 we can see that Greece initiated vaccinations during January 2021. The rollout started about 4 months earlier than Australia's. From February to April, we observe a steady increase in vaccination rates. **However, from May to July we observe a very big increase in vaccinations which starts to decelerate by August and onward. This could be explained by the fact that the campaign was very**

active during the summer and a lot of people chose to get vaccinated during that time in order to enjoy their summer vacation without restrictions.

Finally, after 4 months of decreased activity, we observe an acceleration during December. This suggests that people chose to get vaccinated before Christmas holiday either to get protected or to avoid the very strict measures that had been imposed by the time that excluded unvaccinated people from many essential and social activities.

Conclusions

Our exploratory analysis has reached its end. It seems that through a single dataset we are able to conduct analysis that gives answers to many different questions. We were able to explore the world vaccination status, by examining which countries and continents have the highest and lowest vaccination rates. We also tried to answer the important question of how population affects vaccination rates. We started with a hypothesis that stemmed out of some logical thoughts only to find out that our 'logical' hypothesis does not hold so strong after all and might even not apply. Lastly we examined the seasonality of vaccinations and connected the insights gained with some important facts. We observed correlation between the heavy summer tourist season of Greece and increased vaccination rates before and during those months. We also found correlation between the ending of Melbourne's very long lockdown in late October and its very rapid,successful vaccination campaign during the previous 3 months.

It seems that our world can be explained very well through data up to some point. Data analysis, allows us to understand situations and relationships between facts better. It also allows us to validate or debunk our hypothesis, as without the data to support them hypothesis should never be treated as facts even if they do sound logical. Last but not least, a well rounded data analysis prevents us from being misled by statistics and graphs that might only take into account one perspective. This is a big danger with statistics, as it is easy to mislead decisions when analysis is not educated and well rounded.

*"**Without data you're just another person with an opinion.**"*
Edwards Deming, Statistician