The Effects of Covid-19 from 2020 to 2022

Stat 3355.001 Project

Group 24/Team DCB

Roopali Kallem and Kenan Stredic

November 22, 2022

1 Introduction

The purpose of this project was to use R to analyze large Covid-19 datasets to find any trends and statistics in the data. To complete this analysis, we used packages such as ggplot2, gridExtra, dplyr, plotrix, scales, and more to create visuals from this data, to sort, arrange and filter the data. We wanted to observe the trends of Covid-19 deaths, cases, and vaccinations along with any factors that would affect these trends.

Using various data sets from the World Health Organization (WHO), we gained a better understanding of the effects of the pandemic in various countries between the years 2020 and 2022. Along with this, we used flight data during 2020 to observe how air traffic affected the pandemic in the USA. This includes the WHO Covid-19 Cases dataset, the WHO Covid-19 Vaccination dataset, and the Air Traffic 2020 Dataset. These datasets were chosen because of their clarity and overlapping information. We had 252, 537 total observations and 35 total variables between these datasets. These variables include the date (Date_reported), country (Country), new and cumulative cases (New_cases, Cumulative_cases), new and cumulative cases (New_deaths, Cumulative_deaths), vaccinations (Total_vaccinations), the percent of baseline (PercentOfBaseline), and numerous others. The percent of baseline refers to the proportion of flights on this date as compared to the average number of flights on the same day of the week. We will use these variables to compare and contrast different countries' Covid-19 trends in their cases, deaths, and vaccinations from 2020 to 2022 along with comparing the air traffic of the USA to its total Covid-19 cases in 2020.

Having lived through the Covid-19 Pandemic, there was little time to process how the world and the United States were affected. The overall spread of Covid-19 drastically affected the daily lives of billions of people in numerous economic sectors. Over the last 3 years, there were about 8 major variants and multiple countries imposing complete lockdowns. Therefore, in this project, we aim to analyze datasets regarding the number of confirmed cases and deaths in

the United States and compare them with that of other countries. From this, We can discover which countries were most impacted by the pandemic to get a better understanding of the long-lasting effects of Covid-19.

2 Data Cleaning

We read all the .csv files into 3 different data frames and looked at the data types of each variable in the data frames. We found that the date columns are being read in as type characters, so proceeded with converting them into type date. We initially factored the date column in each data frame then strptime it and finally inserted it into the respective date column on the data frames as a date type. Then rechecked the data type of each variable to confirm and then moved on to removing the columns that were unnecessary for our analysis purpose. For example, for Air traffic data we removed aggregation method, version, airport name, and many other columns.

Then checked for any null values in all the data frames, and found no null values or unknown characters in WHO covid19 cases and Air Traffic, However, we had a few null values in our WHO vaccination data. We looked at each option for data cleaning and discussed each of them. We did not want to remove the rows with null values as each row represented a country and that means causing data loss. So we decided on finding the mean of those columns, as all the columns with null values are numerical and not categorical, and replaced the null entries in those columns with the mean. Then rechecked if we missed out on any other null values before proceeding to the main analysis part.

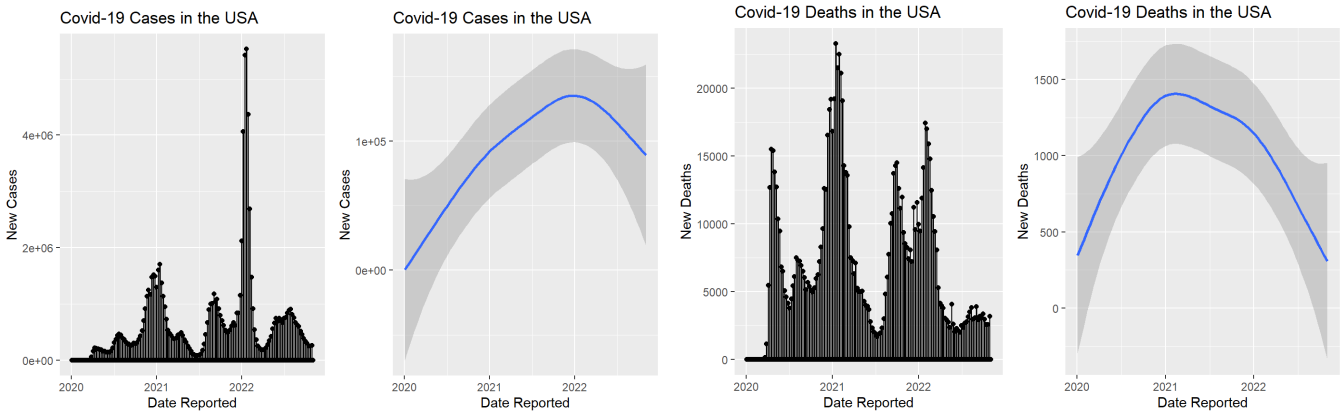
Then we wanted to look more into just the top 5 countries that had the highest covid cases and chose to proceed with the following 5 countries: USA, India, Germany, France, and Brazil. Thus we further sorted and subsetted the data into 5 different data frames for further use in the Analysis and visualization.

3 Analysis

3.1 Between 2020 and 2022, is there any pattern between the number of cases and the number of deaths?

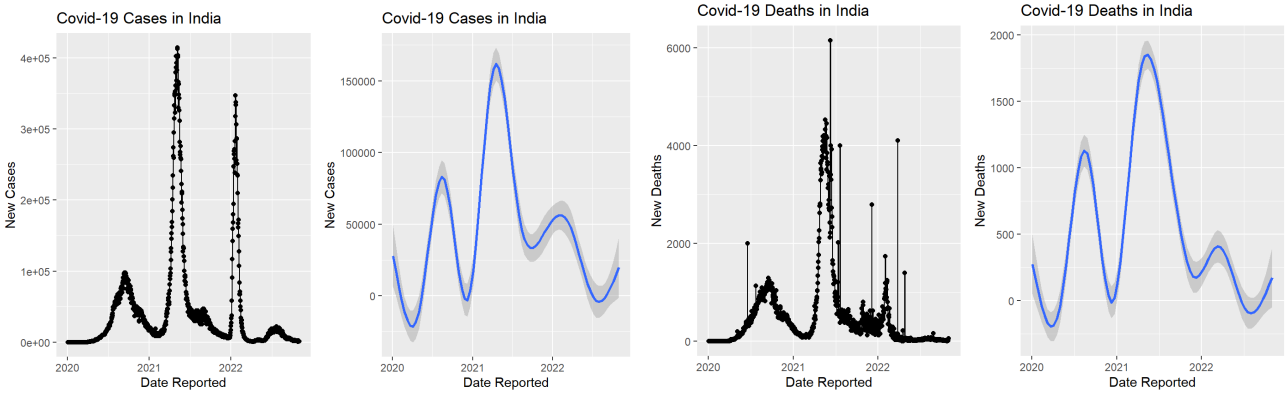
In this section, we wanted to see if there were any trends between the number of new Covid-19 cases and deaths during the pandemic. To do this, we observed the total number of new Covid-19 cases and deaths in the USA, India, Germany, France, and Brazil. For each country, we used a line graph with points and a smooth curve graph that has the dates reported as the X variable and the number of new cases or deaths as the Y variable to look at the peaks and dips in the new cases and deaths from 2020 to 2022.

3.1.1 USA Covid-19 Cases and Deaths from 2020 to 2022



From the line graphs of the Covid-19 cases, we can see that the USA's new Covid-19 cases hit their first peak around the beginning of 2021. However, the number of new cases hit its highest peak at the beginning of 2022. We believe this is because the USA struggled to contain the virus, as they had weaker lockdown regulations than the other countries. During this time, many US citizens were against quarantine and wanted to quickly return to their ways of life before the outbreak of Covid-19. This resulted in citizens attempting to reopen businesses earlier than recommended, which caused the Covid-19 cases to once again rise. This is supported by the line graph of new deaths in the USA, as the deaths first peaked around the start of the pandemic, decreased until hitting their second peak at the beginning of 2021, and had their final peak around 2022. We believe this final peak in deaths was lower than the second peak because of the widespread use of vaccines. After the beginning of 2022, the cases and deaths in the USA both steadily decreased.

3.1.2 India Covid-19 Cases and Deaths from 2020 to 2022

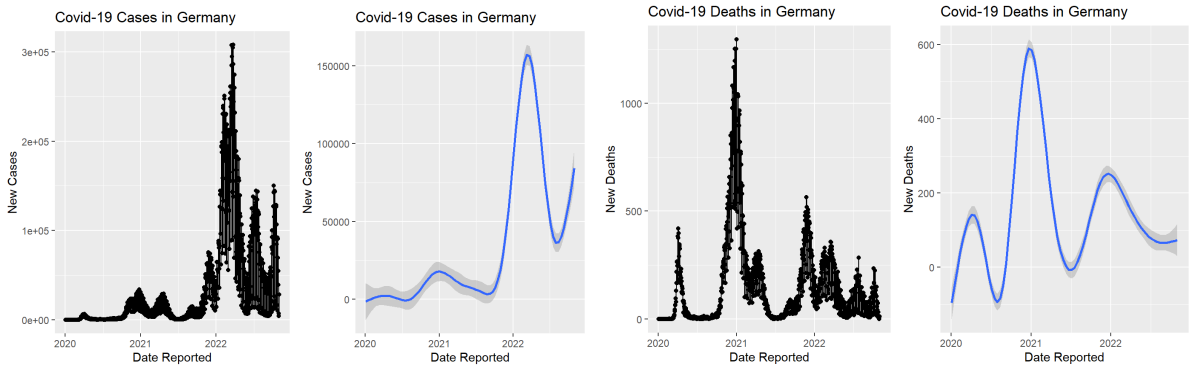


Im

From the line graphs of the Covid-19 cases, we can see that India's new Covid-19 cases hit their first peak around mid-2020. Its highest peak in cases began at the beginning of 2021,

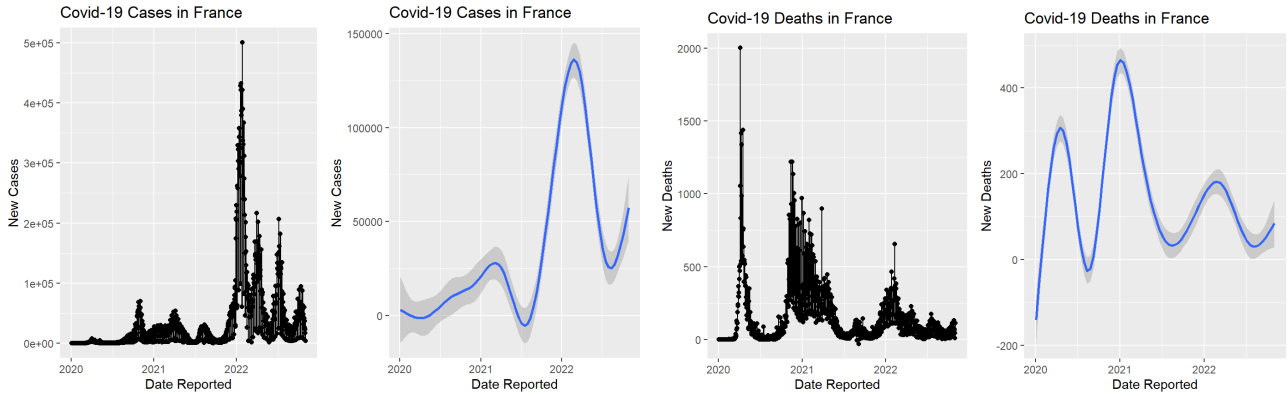
after which it decreased rapidly before hitting another peak at the beginning of 2022. We can see that compared to the USA, India had numerous dips in the cases throughout the 3 years. While the USA constantly increased in cases until 2022, India was able to stop the rise of cases at different points of the pandemic. We believe this is because India had stronger lockdown regulations than the USA, which allowed them to temporarily contain the spread of the virus during the pandemic. This is supported by the line graphs of the new deaths in India, as the deaths hit their first peak around mid-2020, hit their second peak around the beginning of 2021, and decreased rapidly before hitting a smaller peak at the beginning of 2022. We believe the use of the vaccines caused the third peak to be smaller than the previous two peaks.

3.1.3 Germany Covid-19 Cases and Deaths from 2020 to 2022



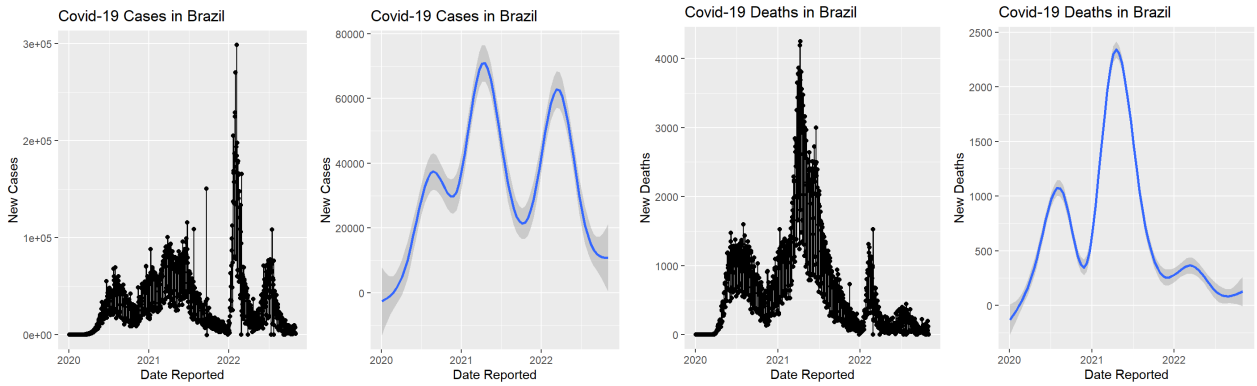
From the line graphs of the Covid-19 cases, we can see that Germany's new Covid-19 cases hit their first peak around the beginning of 2021. Similar to the USA, the cases hit their highest peak at the beginning of 2022. This differed from India and USA, as Germany had consistently lower peaks during 2020 and 2021. We believe this rapid peak in 2022 was caused by the spread of the Omicron Covid-19 variant in Germany. Although Germany had used stricter lockdown procedures at the beginning of the pandemic, they were unable to contain the Omicron variant. This is supported by the line graphs of the new deaths in Germany, as the deaths first peaked around mid-2020, hit their highest peak during the beginning of 2021, and hit their final peak at the beginning of 2022. Even though Germany had more cases during the beginning of 2022, we believe that the introduction of vaccines caused the number of deaths during the beginning of 2022 to be less than the deaths at the beginning of 2021.

3.1.4 France Covid-19 Cases and Deaths from 2020 to 2022



From the line graphs of the Covid-19 cases, we can see that France's new Covid-19 cases hit their first major peak at the end of 2020. Similar to Germany, the cases hit their highest peak at the beginning of 2022. This differed from India and the USA, as France had consistently lower peaks during 2020 and 2021. We believe this rapid peak in 2022 was also caused by the spread of the Omicron Covid-19 variant in France. Although France had stronger lockdown regulations at the start of the pandemic, they were unable to contain the Omicron variant. This is supported by the line graphs of the new deaths in France, as the deaths first peaked around mid-2020, hit their highest peak during the beginning of 2021, and hit their final peak at the start of 2022. Similar to Germany, we believe the introduction of vaccines caused the number of deaths at the beginning of 2022 to be less than the number of deaths at the beginning of 2021. We also believe that Germany and France's similar trends could be caused by their proximity since they are both in Europe.

3.1.5 Brazil Covid-19 Cases and Deaths from 2020 to 2022



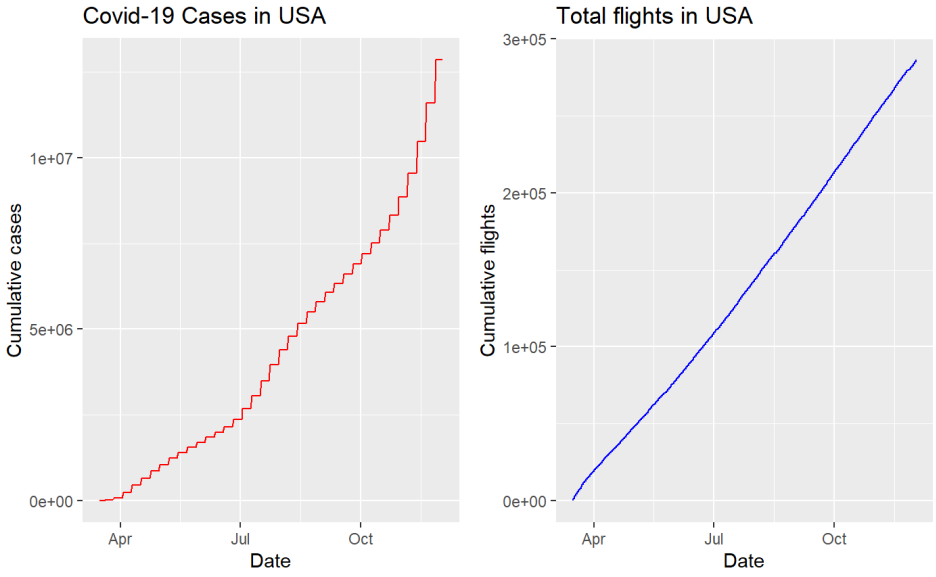
From the line graphs of the Covid-19 cases, we can see that Brazil's new Covid-19 cases hit their first peak at the end of 2020. As opposed to the other countries, Brazil consistently increase in cases, hitting its highest peak during the middle of 2021, before rapidly dropping during the end of 2021. Then, similar to Germany and France, it had another peak at

the beginning of 2022. We believe that this rapid peak at the beginning of 2022 was caused by the Omicron variant. Similar to the USA, Brazil struggled to contain the virus at the beginning of the pandemic, which was due to Brazil's weaker lockdown procedures. Even though they were able to temporarily contain the Covid-19 virus at the end of 2021, they were unable to contain the Omicron variant at the beginning of 2022. This is supported by the line graphs of the new deaths in Brazil, as the deaths first peaked at the end of 2020, hit their highest peak during the middle of 2021, and hit another peak at the beginning of 2022. We believe the introduction of vaccines caused the deaths during the beginning of 2022 to be less than the deaths during the middle of 2021.

3.2 Did air traffic play a role in the increase of covid cases at the start of the pandemic?

In this section, we wanted to see if air traffic affected the number of cumulative Covid-19 cases in the USA at the beginning of the pandemic. For each day in 2020, we compared the number of cumulative Covid-19 cases to the number of cumulative flights in the USA. Using a line graph, we had the dates as the X variable and the number of cumulative cases and flights as the Y variable to see if the cases and flights shared a similar trend from the start of the pandemic to the end of 2020.

3.2.1 Air traffic and Covid-19 Cases in the USA from the start of the pandemic to the end of 2020

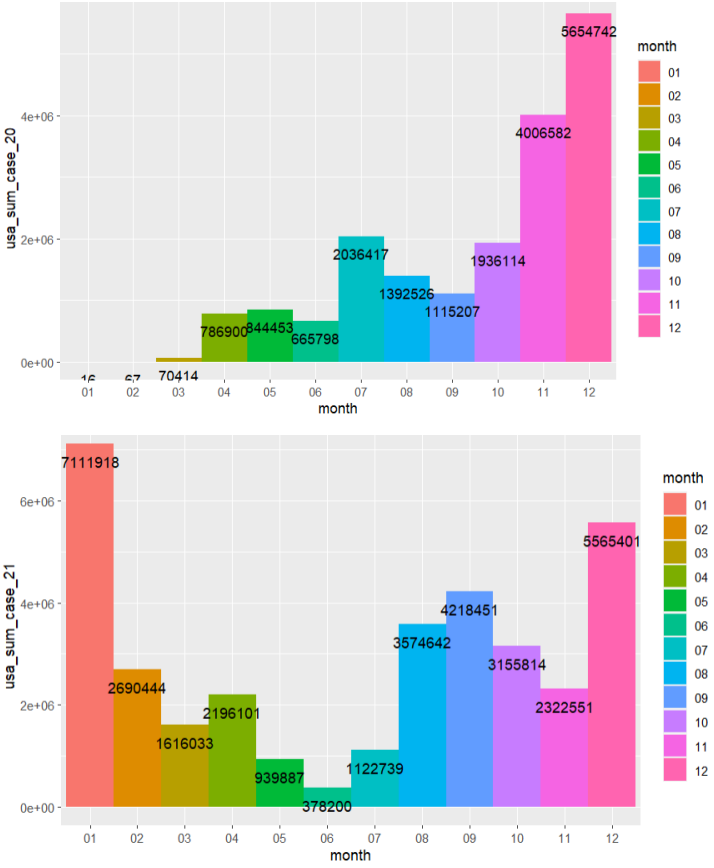


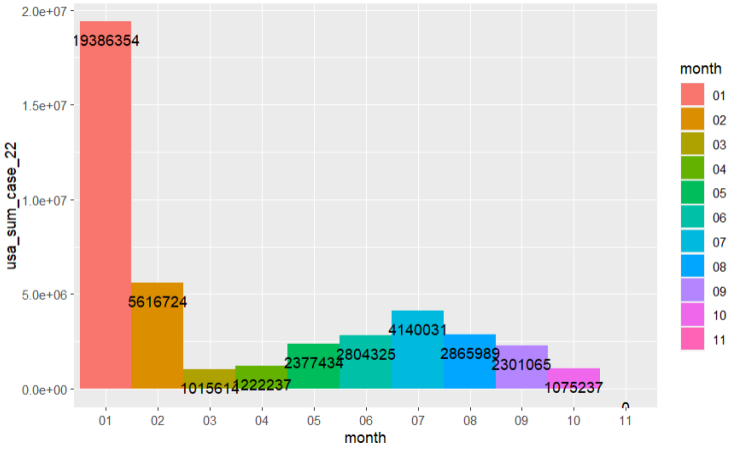
From the line graphs of the Covid-19 cases, we can see that the USA's cumulative cases rapidly increased from the start of the pandemic to the end of 2020. Similarly, the USA's cumulative flights rapidly increased from the start of the pandemic to the end of 2020. This

showed that the overall number of Covid-19 cases increased as the number of flights increased in the USA. We believe this is caused by infected US citizens carrying the covid-19 virus to different cities and states, which contributed to the increase in the total number of covid cases throughout the country. By the end of 2020, The US had 12, 854, 494 total confirmed Covid-19 cases and 286, 271 total flights.

3.3 Did covid spread or cases increase in the summer season and festival seasons due to more travel and family gatherings? What was the trend?

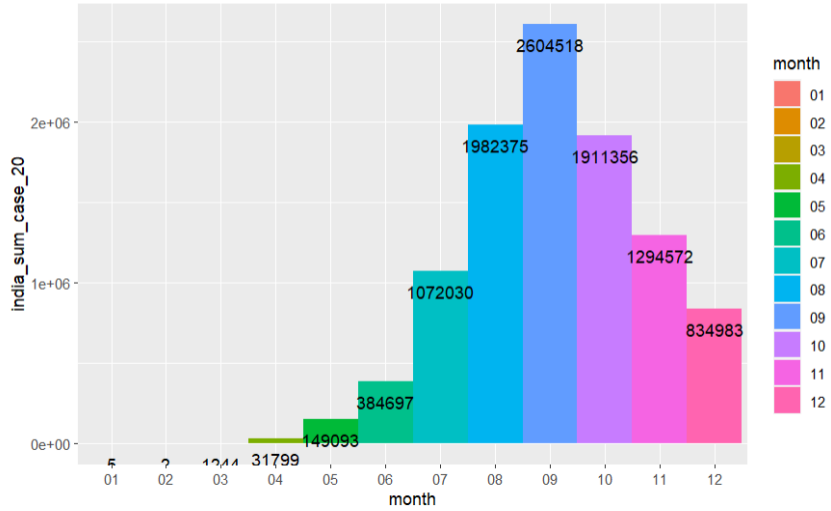
In this section, we wanted to see if the covid cases registered per year were comparatively more during the holiday seasons or not. We anticipated that the cases might have increased due to gatherings, so wanted to see with the freedom to travel and families gathering for festivals, did the cases increase. So we wanted to observe if this is true or not in a few countries and chose to look into the United States of America, India, and Germany.3.3.1 Covid cases distributed by month in years 2020, 2021, and 2022 in the USA

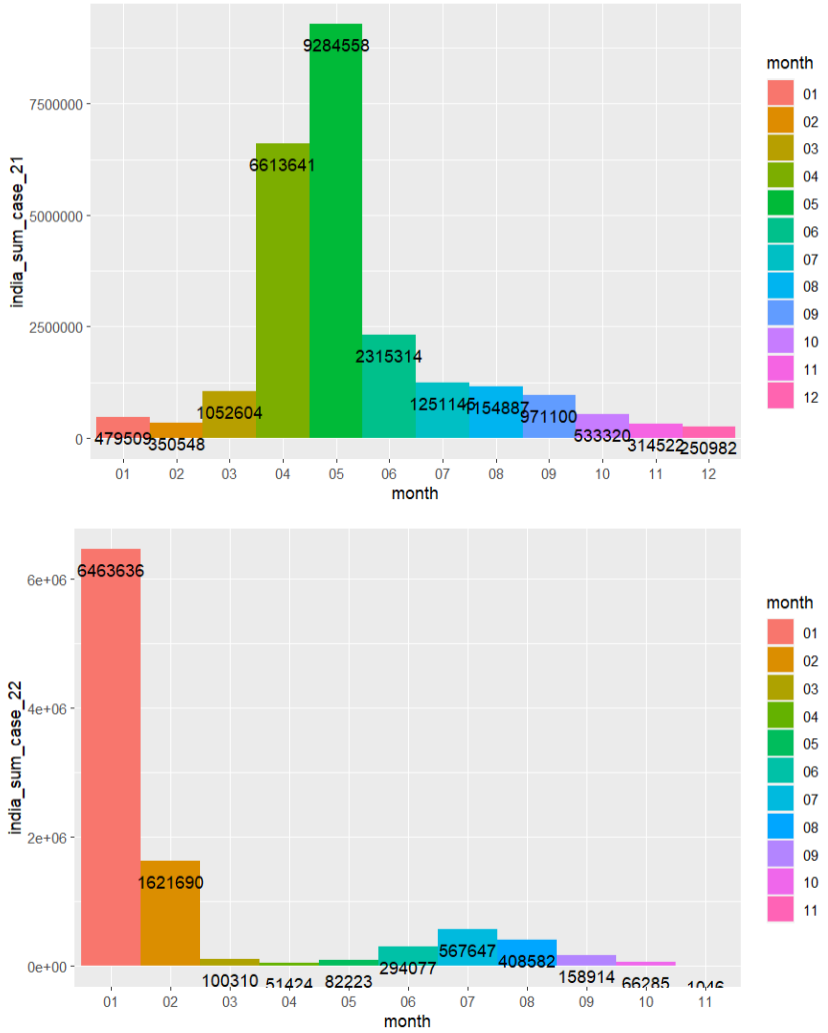




It was only in mid of 2020 did people realize about covid, its symptoms, and its severity, but during November and December of 2020, which is during Thanksgiving and Christmas there are usually a lot of gatherings and shopping, thus increasing the exposure and spread of cases. We see this trend even in 2021 but it is less than before due to all the travel restrictions imposed at the beginning of 2021 to control the spread of the virus and the limits by the stores and all the precautionary restrictions like masks, gloves, and covering being imposed. However, by 2022 we see the spread has decreased drastically. We see that the cases were highest during December and January as people were gathering to celebrate Christmas and new year.

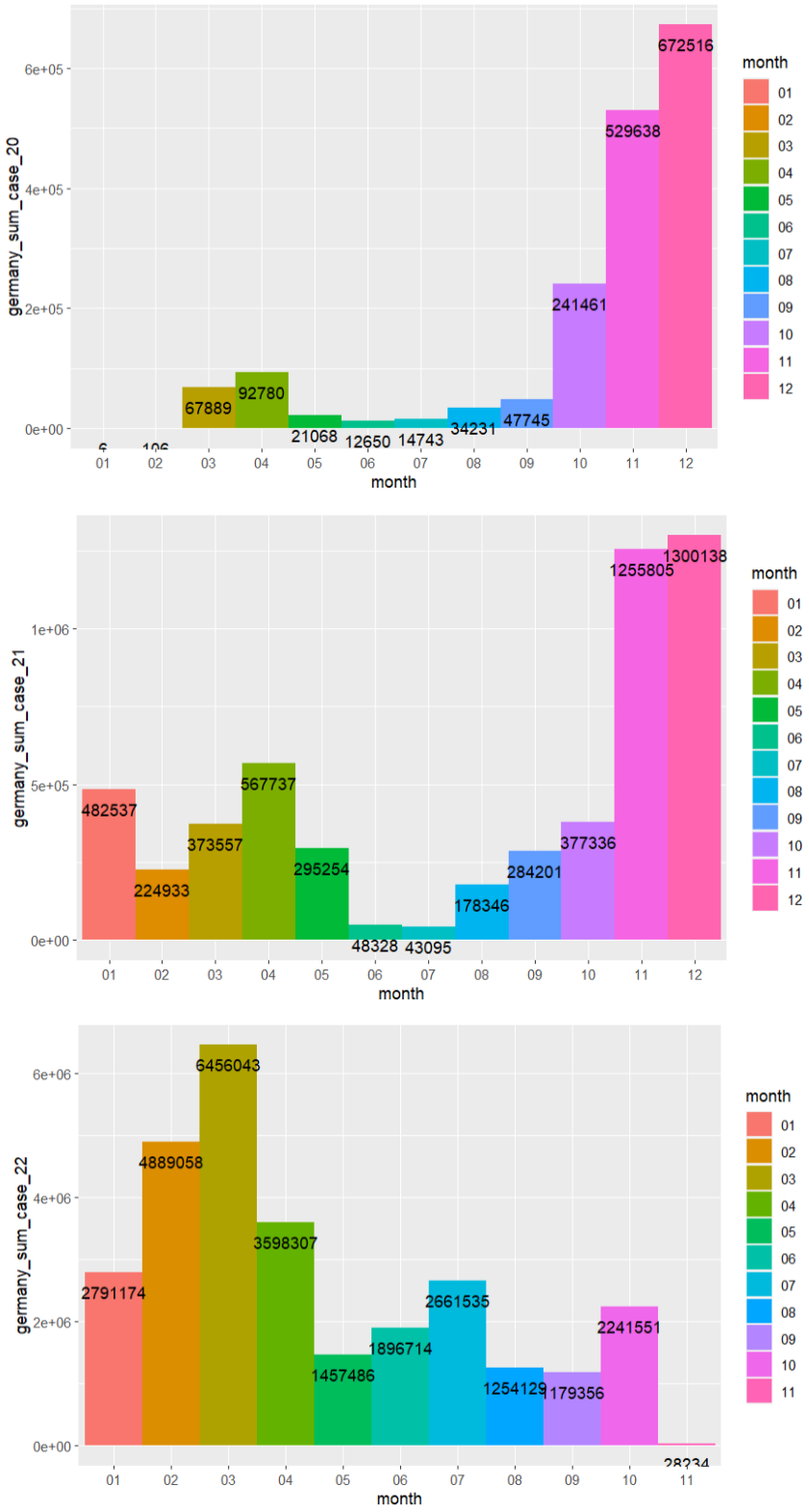
3.3.2 Covid cases distributed by month in years 2020,2021, and 2022 in India





Similarly, It was only in mid of 2020 did people realize about covid, its symptoms, and its severity. However, Students in India have their summer holidays during June and July leading to many people traveling, visiting people, and gatherings increasing the spread of cases in 2020. Then due to the strict restrictions put in place by the end of 2020 till the beginning of 2021 the cases were reduced but then after the restrictions were removed everyone came out to celebrate their festivals that's why we see a rise in April 2021 again as its time for Ugadi(cultural new year). After seeing the spread during Ugadi, restrictions were imposed to control the spread until the end of the year. Then there was a new variant spreading in India during January coinciding with another festival Sankranti(crop harvest festival) and thus increasing the number of cases drastically. We see that during the festivals the spread was about highest in all the above years.

3.3.3 Covid cases distributed by month in years 2020,2021, and 2022 in Germany



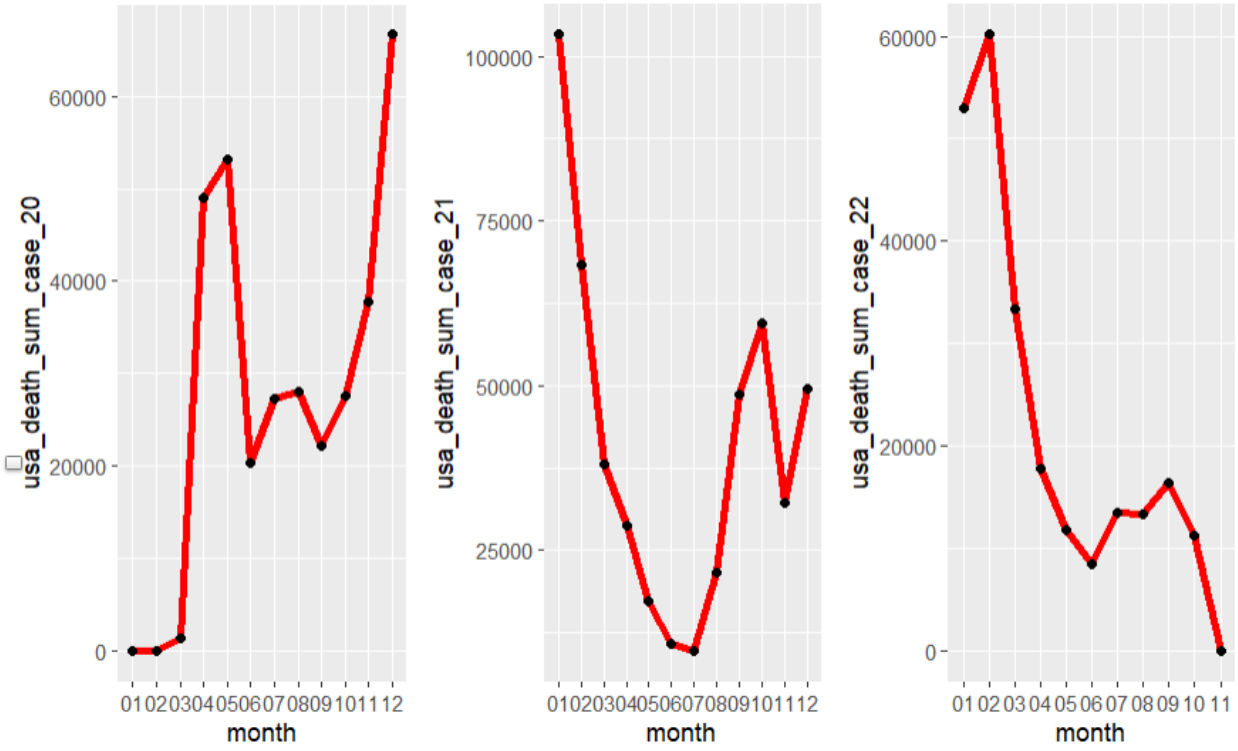
In Germany as well we see during 2020 the highest spread and cases registered were during November and December during Martinmas, and Weihnachten(Christmas). During this

time we see people harvesting crops and celebrating the harvest, shopping for gifts, and meeting up with family to celebrate this increasing spread. The following year the government imposed lockdowns and travel restrictions reducing the spread yet we can notice that in 2021 during April which is their carnival season the cases were more even with all the restrictions, and we see the highest during November and December for the same reason stated above. In 2022 after the removal of the restrictions we see the cases increase due to the spread of the new variant, However in 2020 and 2021 we do see the trend of the high spread of covid-19 during celebrations or holidays.

3.4 Did the total number of vaccines affect the total amount of deaths in a country?

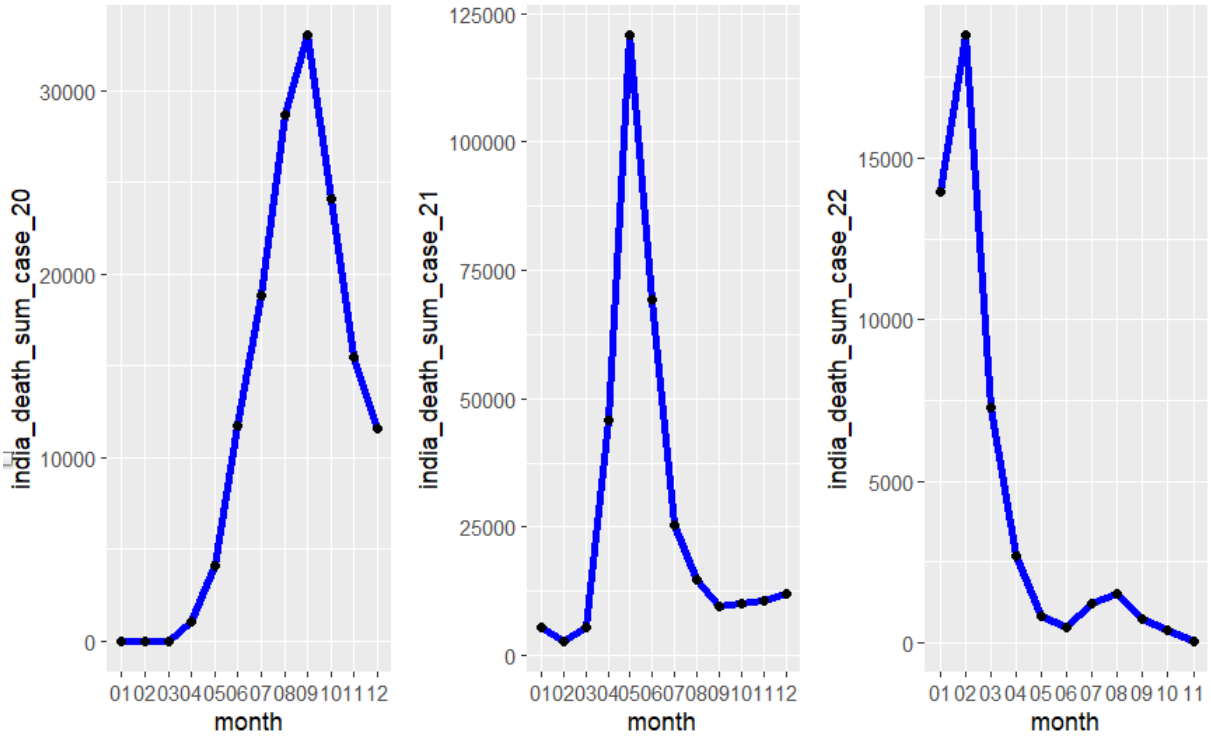
We wanted to see if the number of deaths decreased after the invention and distribution of vaccines so we considered the same three counties as before the USA, India, and Germany to make this observation. We further took note of when the covid vaccines were started distributing in each of these counties and found there were about 629,073,496 vaccines distributed by 2022 in the USA which means only 65% population was vaccinated with at least 1 dose. In India by 2022 2,196,382,882 vaccines were distributed meaning, about 90% population was vaccinated with at least one dose. In Germany by 2022 186,739,714 vaccines were distributed meaning 76% were vaccinated with at least one dose.

3.4.1 Deaths due to covid distributed by month in years 2020,2021, and 2022 in the USA



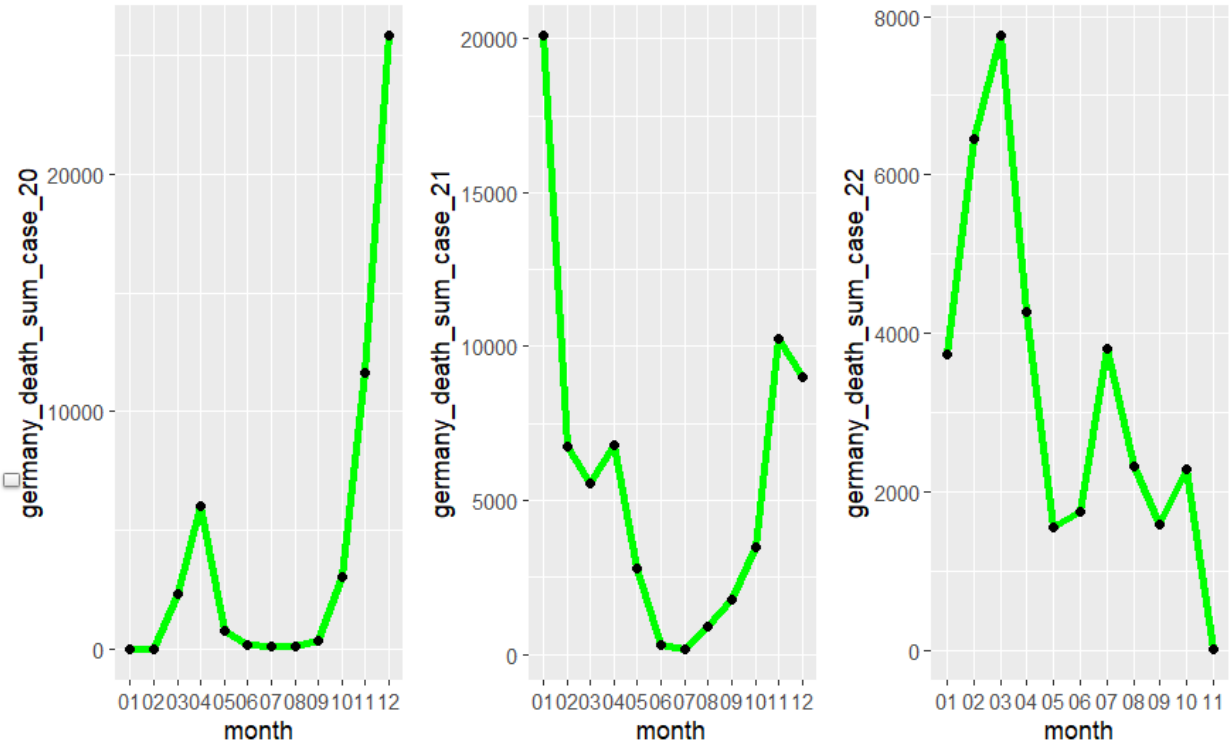
In the USA we see the deaths were at peaks during the end of 2020 and beginning of 2021, but due to the travel impositions and shutting down of all the businesses and universities, we saw the cases decrease, thus decreasing the number of deaths. Then when businesses started back up we saw the deaths increase a little however we had vaccines being distributed by the end of 2021 to the elderly people who were more prone to severe symptoms of the virus.

3.4.2 Deaths due to covid distributed by month in years 2020, 2021, and 2022 in India



In India, we see the deaths were at peaks during the mid of 2020 and it was hard to control so strict lockdowns were imposed and causing a heavy drop in the spread and deaths. However, by the beginning of 2021, the lockdown was removed and there was also a heavy scarcity of necessary medical equipment like masks and oxygen tanks, and more causing the deaths to skyrocket again. Then with the distribution by the end of 2021, the deaths dropped as people were recovering and the spread was also under control. However, at the beginning of 2022, there was a new variant spreading in the country and the existing vaccine was not a cure for the new variant so we see cases and deaths increase a little however after modification the vaccine was a good fit for the new variant as well. Thus we see the cases drop and stay low.

3.4.3 Deaths due to covid distributed by month in years 2020, 2021, and 2022 in Germany



We see the same situation in Germany as it was in India. We see the deaths were at peaks by the end of 2020 and strict lockdowns were imposed and causing a heavy drop in the spread and deaths. However, at the beginning of 2022, there was a new variant spreading in the country and cases were increasing but with the distribution of the vaccines, we see deaths dropped as people were recovering and the spread was also under control. We continue to see the german government mandating masks in public transport to completely reduce the spread and the deaths caused by covid but we do see that ever since the distribution of vaccines the deaths due to covid decreased.

4. Conclusion

After extensive data cleaning and thorough analysis of this data, we discovered numerous interesting things from this data. We have noticed that Air Traffic has been one of the major reasons for the drastic spread of the Covid-19 virus all over the world and every country was affected to a large extent due to this. We have also noticed that in numerous countries there was a trend to notice that the highest spread of covid in a year was usually during their holidays or celebration months. Moreover, we wanted to see if vaccines helped reduce this

spread and deaths caused by the virus and found to a large extent the more percent of the population vaccinated the fewer the deaths due to covid were in those countries. This was just the beginning but for future work, we would like to look deeper into this and see if the countries which were close to each other had any similarities in the overall cases as there are many other public means of transport in countries that people use to travel like trains, buses, and cars, and those aspects were not looked into in this report and we majorly just looked into air traffic. We would also like to see the rate of recovery after vaccination as multiple variants were being spread across the world, some counties had more variants than others so we would like to look at some of those counties and compare the recovery. This is also linked to the country's economic standing thus bringing us to our next question of seeing if a country's economic standing and technology helped implement restrictions better and control covid. We see that in many heavily populated countries the covid cases were high and deaths were high and we believe this might be due to the shortage of masks, gloves, and oxygen tanks those countries face. We wanted to see if countries that didn't have these issues did they lose any fewer people as they were well prepared and had more resources to protect their people. Countries that are doing well have more resources to allocate to the health care sector and vaccinations thus helping more people recover, whereas, on the other hand, low-level countries even struggle to maintain their government and thus have very few resources to spare for the well-being of their people.

5. Appendix

Code for Data Cleaning

```
# importing necessary libraries
library(ggplot2)
library(gridExtra)


# reading the csv files into data frames
world_covid <- read.csv( "WHO C-19 cummulative.csv" )
air_traffic <- read.csv( "c-19 impact on air traffic.csv" )
world_vacc <- read.csv( "vaccination-data.csv" )

# Data Pre-Processing
# converting dates in the data frames from character type to date type.

# Factor Date_reported in world_covid
world_covid_Date_reported_factored <- as.factor( world_covid$Date_reported )

# Convert Date_reported from character to POSIXlt in world_covid
world_covid_Date_reported_POSIXlt <- strptime( world_covid_Date_reported_factored,
                                        format = "%m/%d/%Y" )

# Convert Date_reported from POSIXlt to Date in world_covid
world_covid$Date_reported <- as.Date( world_covid_Date_reported_POSIXlt,
                                format = "%Y-%m-%d" )
str( world_covid ) # printing to see the data type of each variable



# Removing unnecessary columns from world_covid
world_covid <- world_covid[ -c( 2, 4 ) ]
str( world_covid ) # printing again to see if the correct columns are removed and
# to confirm the data type of each variable


# checking for any possible null values
sum( is.na( world_covid$Date_reported ) )
sum( is.na( world_covid$Country ) )
sum( is.na( world_covid$New_cases ) )
sum( is.na( world_covid$Cumulative_cases ) )
sum( is.na( world_covid$New_deaths ) )
sum( is.na( world_covid$Cumulative_deaths ) )
```

```
# converting dates in the data frames from character type to date type.
# Factor Date_reported in air_traffic
air_traffic_Date_factored <- as.factor( air_traffic$Date )

# Convert Date from character to POSIXlt in air_traffic
air_traffic_Date_POSIXlt <- strptime( air_traffic_Date_factored,
                                      format = "%m/%d/%Y" )

# Convert Date from POSIXlt to Date in air_traffic
air_traffic$Date <- as.Date( air_traffic_Date_POSIXlt,
                             format = "%Y-%m-%d" )

str( air_traffic ) # printing to see the data type of each variable


# Removing unnecessary columns from air_traffic
air_traffic <- air_traffic[ -c( 1, 3, 4, 6, 9, 11 ) ]
str( air_traffic )  # printing again to see if the correct columns are removed and
# to confirm the data type of each variable

# checking for any possible null values
sum( is.na( air_traffic$Date ) )
sum( is.na( air_traffic$Country ) )
sum( is.na( air_traffic$City ) )
sum( is.na( air_traffic$State ) )




# Factor Date_updated in world_vacc
world_vacc_Date_Updated_factored <- as.factor( world_vacc$DATE_UPDATED )

# Convert Date_updated from character to POSIXlt in world_vacc
world_vacc_Date_Updated_POSIXlt <- strptime( world_vacc_Date_Updated_factored,
                                         format = "%m/%d/%Y" )

# Convert Date_updated from POSIXlt to Date in world_vacc
world_vacc$DATE_UPDATED <- as.Date( world_vacc_Date_Updated_POSIXlt,
                                    format = "%Y-%m-%d" )


# Factor FIRST_VACCINE_DATE in world_vacc
world_vacc_FIRST_VACCINE_DATE_factored <- as.factor( world_vacc$FIRST_VACCINE_DATE )

# Convert FIRST_VACCINE_DATE from character to POSIXlt in world_vacc
world_vacc_FIRST_VACCINE_DATE_POSIXlt <- strptime(
world_vacc_FIRST_VACCINE_DATE_factored,
                                          format = "%m/%d/%Y" )
```

```
# Convert FIRST_VACCINE_DATE from POSIXlt to Date in world_vacc
world_vacc$FIRST_VACCINE_DATE <- as.Date( world_vacc_FIRST_VACCINE_DATE_POSIXlt,
                                format = "%Y-%m-%d" )

str( world_vacc )  # printing to see the data type of each variable



# Removing unnecessary columns from world_covid
world_vacc <- world_vacc[ -c( 2:4, 12:16 ) ]
str( world_vacc )  # printing again to see if the correct columns are removed and
# to confirm the data type of each variable


# checking for any possible null values
sum( is.na( world_vacc$DATE_UPDATED ) )
sum( is.na( world_vacc$Country ) )
sum( is.na( world_vacc$TOTAL_VACCINATIONS ) )
sum( is.na( world_vacc$PERSONS_VACCINATED_1PLUS_DOSE ) )
sum( is.na( world_vacc$TOTAL_VACCINATIONS_PER100 ) )
sum( is.na( world_vacc$PERSONS_VACCINATED_1PLUS_DOSE_PER100 ) )
sum( is.na( world_vacc$PERSONS_FULLY_VACCINATED ) )
sum( is.na( world_vacc$PERSONS_FULLY_VACCINATED_PER100 ) )



# replacing the Null values with the mean of the column

world_vacc$TOTAL_VACCINATIONS[ is.na( world_vacc$TOTAL_VACCINATIONS ) ] <-
  mean( world_vacc$TOTAL_VACCINATIONS, na.rm = TRUE )

world_vacc$PERSONS_VACCINATED_1PLUS_DOSE[ is.na(
world_vacc$PERSONS_VACCINATED_1PLUS_DOSE ) ] <-
  mean( world_vacc$PERSONS_VACCINATED_1PLUS_DOSE, na.rm = TRUE )

world_vacc$TOTAL_VACCINATIONS_PER100[ is.na( world_vacc$TOTAL_VACCINATIONS_PER100 ) ]
<-
  mean( world_vacc$TOTAL_VACCINATIONS_PER100, na.rm = TRUE )

world_vacc$PERSONS_VACCINATED_1PLUS_DOSE_PER100[ is.na(
world_vacc$PERSONS_VACCINATED_1PLUS_DOSE_PER100 ) ] <-
  mean( world_vacc$PERSONS_VACCINATED_1PLUS_DOSE_PER100, na.rm = TRUE )

world_vacc$PERSONS_FULLY_VACCINATED[ is.na( world_vacc$PERSONS_FULLY_VACCINATED ) ]
<-
  mean( world_vacc$PERSONS_FULLY_VACCINATED, na.rm = TRUE )
```

```
world_vacc$PERSONS_FULLY_VACCINATED_PER100[ is.na(
world_vacc$PERSONS_FULLY_VACCINATED_PER100 ) ] <-
  mean( world_vacc$PERSONS_FULLY_VACCINATED_PER100, na.rm = TRUE )

# rechecking to see if any more null values
sum( is.na( world_vacc$DATE_UPDATED ) )
sum( is.na( world_vacc$Country ) )
sum( is.na( world_vacc$TOTAL_VACCINATIONS ) )
sum( is.na( world_vacc$PERSONS_VACCINATED_1PLUS_DOSE ) )
sum( is.na( world_vacc$TOTAL_VACCINATIONS_PER100 ) )
sum( is.na( world_vacc$PERSONS_VACCINATED_1PLUS_DOSE_PER100 ) )
sum( is.na( world_vacc$PERSONS_FULLY_VACCINATED ) )
sum( is.na( world_vacc$PERSONS_FULLY_VACCINATED_PER100 ) )
```

Code for Data visualization and analysis in Section 3.1 – 3.4

```
# Q1 Between the two years is there any pattern between the number of cases and the number of deaths
# Between the two years is there any pattern between the number of cases and the number of deaths in
the USA
# Instead of using just the smooth line graph, we showed the line and point graph next to the smooth line
graph

usa_covid <- world_covid[ which( world_covid$Country == "United States of America" ), ]

usa_deaths1 <- ggplot( data = usa_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in the USA" )

usa_deaths2 <- ggplot( data = usa_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in the USA" )

usa_cases1 <- ggplot( data = usa_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in the USA" )

usa_cases2 <- ggplot( data = usa_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in the USA" )

grid.arrange( usa_deaths1, usa_deaths2, nrow = 1 )
grid.arrange( usa_cases1, usa_cases2, nrow = 1 )

# Between the two years is there any pattern between the number of cases and the number of deaths in
India
```

```
india_covid <- world_covid[ which( world_covid$Country == "India" ), ]

india_deaths1 <- ggplot( data = india_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in India" )

india_deaths2 <- ggplot( data = india_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in India" )

india_cases1 <- ggplot( data = india_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in India" )

india_cases2 <- ggplot( data = india_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in India" )

grid.arrange( india_deaths1, india_deaths2, nrow = 1 )
grid.arrange( india_cases1, india_cases2, nrow = 1 )

# Between the two years is there any pattern between the number of cases and the number of deaths in
Germany

germany_covid <- world_covid[ which( world_covid$Country == "Germany" ), ]

germany_deaths1 <- ggplot( data = germany_covid, mapping = aes( x = Date_reported, y = New_deaths
) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in Germany" )

germany_deaths2 <- ggplot( data = germany_covid, mapping = aes( x = Date_reported, y = New_deaths
) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in Germany" )

germany_cases1 <- ggplot( data = germany_covid, mapping = aes( x = Date_reported, y = New_cases ) )
+
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in Germany" )

germany_cases2 <- ggplot( data = germany_covid, mapping = aes( x = Date_reported, y = New_cases ) )
+
  geom_smooth() +
```

```
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in Germany" )

grid.arrange( germany_deaths1, germany_deaths2, nrow = 1 )
grid.arrange( germany_cases1, germany_cases2, nrow = 1 )

# Between the two years is there any pattern between the number of cases and the number of deaths in
France

france_covid <- world_covid[ which( world_covid$Country == "France" ), ]

france_deaths1 <- ggplot( data = france_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in France" )

france_deaths2 <- ggplot( data = france_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in France" )

france_cases1 <- ggplot( data = france_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in France" )

france_cases2 <- ggplot( data = france_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in France" )

grid.arrange( france_deaths1, france_deaths2, nrow = 1 )
grid.arrange( france_cases1, france_cases2, nrow = 1 )

# Between the two years is there any pattern between the number of cases and the number of deaths in
Brazil

brazil_covid <- world_covid[ which( world_covid$Country == "Brazil" ), ]

brazil_deaths1 <- ggplot( data = brazil_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in Brazil" )

brazil_deaths2 <- ggplot( data = brazil_covid, mapping = aes( x = Date_reported, y = New_deaths ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Deaths", title = "Covid-19 Deaths in Brazil" )

brazil_cases1 <- ggplot( data = brazil_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_line() +
  geom_point() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in Brazil" )
```

```
brazil_cases2 <- ggplot( data = brazil_covid, mapping = aes( x = Date_reported, y = New_cases ) ) +
  geom_smooth() +
  labs( x = "Date Reported", y = "New Cases", title = "Covid-19 Cases in Brazil" )

grid.arrange( brazil_deaths1, brazil_deaths2, nrow = 1 )
grid.arrange( brazil_cases1, brazil_cases2, nrow = 1 )
```

```
# Q2 Did air traffic play a role in the increase of covid cases at the start of the pandemic (2020)?

library( dplyr )

usa_air_traffic <- air_traffic[ which( air_traffic$Country == 'United States of America (the)' ), ]

usa_air_traffic <- usa_air_traffic %>% arrange( Date )

usa_air_traffic[, "Cumulative_flights" ] <- cumsum( usa_air_traffic$PercentOfBaseline )


usa_20 <- usa_covid %>% filter(between(Date_reported, as.Date('2020-03-16'), as.Date('2020-12-02')))

# Changed from bar plots to line plots

usa_cases <- ggplot( data = usa_20, aes( x = Date_reported, y = Cumulative_cases ) ) +
  geom_line( stat = "identity", color = "red" ) +
  labs( x = "Date", y = "Cumulative cases", title = "Covid-19 Cases in USA" )

usa_traffic <- ggplot( data = usa_air_traffic, aes( x = Date, y = Cumulative_flights ) ) +
  geom_line( stat = "identity", color = "blue" ) +
  labs( x = "Date", y = "Cumulative flights", title = "Total flights in USA" )

grid.arrange( usa_cases, usa_traffic, nrow = 1 )
```

```
# Q3 checking if covid spread or cases increased in the summer season and festival season due to more
travel and family gatherings

# Get the library.
library( plotrix )
# library(scales)

# summer holidays in USA mid-May to about mid-August
```

```
usa_covid$year <- format( usa_covid$Date_reported, "%y" )             # Extract year

usa_covid_20 <- usa_covid[which(usa_covid$year == '20'),]

usa_covid_21 <- usa_covid[ which( usa_covid$year == '21' ), ]

usa_covid_22 <- usa_covid[ which( usa_covid$year == '22' ), ]


usa_covid_20$month <- format( usa_covid_20$Date_reported, "%m" )             # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
usa_covid_sum_20 <- usa_covid_20 %>% group_by( month ) %>% summarise( usa_sum_case_20 =
sum( New_cases ) )

# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( usa_covid_sum_20, aes( x = month, y = usa_sum_case_20, fill = month, width = 1)) + geom_bar(
stat = "identity" ) + geom_text( aes(label = usa_sum_case_20 ), vjust = 2, colour = "black")


usa_covid_21$month <- format( usa_covid_21$Date_reported, "%m" )             # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
usa_covid_sum_21 <- usa_covid_21 %>% group_by( month ) %>% summarise( usa_sum_case_21 =
sum( New_cases ) )


# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( usa_covid_sum_21, aes( x = month, y = usa_sum_case_21, fill = month, width = 1 )) + geom_bar(
stat = "identity" ) + geom_text( aes( label = usa_sum_case_21 ), vjust = 2, colour = "black" )


usa_covid_22$month <- format( usa_covid_22$Date_reported, "%m" )             # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
usa_covid_sum_22 <- usa_covid_22 %>% group_by( month ) %>% summarise( usa_sum_case_22 =
sum( New_cases ) )


# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( usa_covid_sum_22, aes( x = month, y = usa_sum_case_22, fill = month, width = 1 )) + geom_bar(
stat = "identity" ) + geom_text( aes( label = usa_sum_case_22 ), vjust = 2, colour = "black" )
```

```
# summer holidays in India Mid May to July 1st

india_covid$year <- format( india_covid$Date_reported, "%y" )             # Extract year

india_covid_20 <- india_covid[which(india_covid$year == '20'),]

india_covid_21 <- india_covid[ which( india_covid$year == '21' ), ]

india_covid_22 <- india_covid[ which( india_covid$year == '22' ), ]



india_covid_20$month <- format( india_covid_20$Date_reported, "%m" )             # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
india_covid_sum_20 <- india_covid_20 %>% group_by( month ) %>% summarise( india_sum_case_20 =
sum( New_cases ) )


# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( india_covid_sum_20, aes( x = month, y = india_sum_case_20, fill = month, width = 1 )) +
geom_bar( stat = "identity" ) + geom_text( aes( label = india_sum_case_20 ), vjust = 2, colour = "black" )



india_covid_21$month <- format( india_covid_21$Date_reported, "%m" )             # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
india_covid_sum_21 <- india_covid_21 %>% group_by( month ) %>% summarise( india_sum_case_21 =
sum( New_cases ) )


# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( india_covid_sum_21, aes( x = month, y = india_sum_case_21, fill = month, width = 1 )) +
geom_bar( stat = "identity" ) + geom_text( aes( label = india_sum_case_21 ), vjust = 2, colour = "black" )



india_covid_22$month <- format( india_covid_22$Date_reported, "%m" )             # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
india_covid_sum_22 <- india_covid_22 %>% group_by( month ) %>% summarise( india_sum_case_22 =
sum( New_cases ) )
```

```
# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( india_covid_sum_22, aes( x = month, y = india_sum_case_22, fill = month, width = 1 )) +
geom_bar( stat = "identity" ) + geom_text( aes( label = india_sum_case_22 ), vjust = 2, colour = "black" )
```

```
germany_covid$year <- format( germany_covid$Date_reported, "%y" )            # Extract year

germany_covid_20 <- germany_covid[ which( germany_covid$year == '20' ),]

germany_covid_21 <- germany_covid[ which( germany_covid$year == '21' ), ]

germany_covid_22 <- germany_covid[ which( germany_covid$year == '22' ), ]
```

```
germany_covid_20$month <- format( germany_covid_20$Date_reported, "%m" )            # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
germany_covid_sum_20 <- germany_covid_20 %>% group_by( month ) %>% summarise(
germany_sum_case_20 = sum( New_cases ) )
```

```
# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( germany_covid_sum_20, aes( x = month, y = germany_sum_case_20, fill = month, width = 1 )) +
geom_bar( stat = "identity" ) + geom_text( aes( label = germany_sum_case_20 ), vjust = 2, colour =
"black" )
```

```
germany_covid_21$month <- format( germany_covid_21$Date_reported, "%m" )            # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
germany_covid_sum_21 <- germany_covid_21 %>% group_by( month ) %>% summarise(
germany_sum_case_21 = sum( New_cases ) )
```

```
# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( germany_covid_sum_21, aes( x = month, y = germany_sum_case_21, fill = month, width = 1 )) +
geom_bar( stat = "identity" ) + geom_text( aes( label = germany_sum_case_21 ), vjust = 2, colour =
"black" )
```

```
germany_covid_22$month <- format( germany_covid_22$Date_reported, "%m" )            # Extract month

# grouping by month and adding the covid cases to get the total cases registered in that whole month
```

```
germany_covid_sum_22 <- germany_covid_22 %>% group_by( month ) %>% summarise(
germany_sum_case_22 = sum( New_cases ) )


# plotting bar graphs by passing in month as x and cases as y
# setting each bar width to 1 and writing the value in black
ggplot( germany_covid_sum_22, aes( x = month, y = germany_sum_case_22, fill = month, width = 1 )) +
geom_bar( stat = "identity" ) + geom_text( aes( label = germany_sum_case_22 ), vjust = 2, colour =
"black" )




# Q4 Did the total number of vaccines affect the total amount of deaths in a country?

# grouping by month and adding the covid cases to get the total cases registered in that whole month
usa_covid_death_sum_20 <- usa_covid_20 %>% group_by( month ) %>% summarise(
usa_death_sum_case_20 = sum( New_deaths ) )

# grouping by month and adding the covid cases to get the total cases registered in that whole month
usa_covid_death_sum_21 <- usa_covid_21 %>% group_by( month ) %>% summarise(
usa_death_sum_case_21 = sum( New_deaths ) )

# grouping by month and adding the covid cases to get the total cases registered in that whole month
usa_covid_death_sum_22 <- usa_covid_22 %>% group_by( month ) %>% summarise(
usa_death_sum_case_22 = sum( New_deaths ) )


# plotting line graphs by passing in month as x and deaths as y
# setting each bar width to 1 and drawing lines in red for this country
usa_20 <- ggplot( data = usa_covid_death_sum_20, aes( x = month, y = usa_death_sum_case_20, group
= 1 )) + geom_line( color = "red", size = 1.5 ) + geom_point()
usa_21 <- ggplot( data = usa_covid_death_sum_21, aes( x = month, y = usa_death_sum_case_21, group
= 1 )) + geom_line( color = "red", size = 1.5 ) + geom_point()
usa_22 <- ggplot( data = usa_covid_death_sum_22, aes( x = month, y = usa_death_sum_case_22, group
= 1 )) + geom_line( color = "red", size = 1.5 ) + geom_point()

grid.arrange( usa_20, usa_21, usa_22, nrow = 1 )



# grouping by month and adding the covid cases to get the total cases registered in that whole month
india_covid_death_sum_20 <- india_covid_20 %>% group_by( month ) %>% summarise(
india_death_sum_case_20 = sum( New_deaths ) )

# grouping by month and adding the covid cases to get the total cases registered in that whole month
india_covid_death_sum_21 <- india_covid_21 %>% group_by( month ) %>% summarise(
india_death_sum_case_21 = sum( New_deaths ) )
```

```
# grouping by month and adding the covid cases to get the total cases registered in that whole month
india_covid_death_sum_22 <- india_covid_22 %>% group_by( month ) %>% summarise(
india_death_sum_case_22 = sum( New_deaths ) )


# plotting line graphs by passing in month as x and deaths as y
# setting each bar width to 1 and drawing lines in blue for this country
india_20 <- ggplot( data = india_covid_death_sum_20, aes( x = month, y = india_death_sum_case_20,
group = 1 )) + geom_line( color = "blue", size = 1.5 ) + geom_point()
india_21 <- ggplot( data = india_covid_death_sum_21, aes( x = month, y = india_death_sum_case_21,
group = 1 )) + geom_line( color = "blue", size = 1.5 ) + geom_point()
india_22 <- ggplot( data = india_covid_death_sum_22, aes( x = month, y = india_death_sum_case_22,
group = 1 )) + geom_line( color = "blue", size = 1.5 ) + geom_point()

grid.arrange( india_20, india_21, india_22, nrow = 1 )


# grouping by month and adding the covid cases to get the total cases registered in that whole month
germany_covid_death_sum_20 <- germany_covid_20 %>% group_by( month ) %>% summarise(
germany_death_sum_case_20 = sum( New_deaths ) )

# grouping by month and adding the covid cases to get the total cases registered in that whole month
germany_covid_death_sum_21 <- germany_covid_21 %>% group_by( month ) %>% summarise(
germany_death_sum_case_21 = sum( New_deaths ) )

# grouping by month and adding the covid cases to get the total cases registered in that whole month
germany_covid_death_sum_22 <- germany_covid_22 %>% group_by( month ) %>% summarise(
germany_death_sum_case_22 = sum( New_deaths ) )

# plotting line graphs by passing in month as x and deaths as y
# setting each bar width to 1 and drawing lines in green for this country
germany_20 <- ggplot( data = germany_covid_death_sum_20, aes( x = month, y =
germany_death_sum_case_20, group = 1 )) + geom_line( color = "green", size = 1.5 ) + geom_point()
germany_21 <- ggplot( data = germany_covid_death_sum_21, aes( x = month, y =
germany_death_sum_case_21, group = 1 )) + geom_line( color = "green", size = 1.5 ) + geom_point()
germany_22 <- ggplot( data = germany_covid_death_sum_22, aes( x = month, y =
germany_death_sum_case_22, group = 1 )) + geom_line( color = "green", size = 1.5 ) + geom_point()

grid.arrange( germany_20, germany_21, germany_22, nrow = 1 )
```

References

CH, Carriech. “COVID Impact on Airport Traffic-Simple Data Viz✈🛫.” *Kaggle*, 17 Aug. 2021, www.kaggle.com/code/carriech/covid-impact-on-airport-traffic-simple-data-viz/data.

"COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University", https://github.com/CSSEGISandData/COVID-19

“COVID-19 Dataset.” *Kaggle*, 7 Aug. 2020, www.kaggle.com/datasets/imdevskp/corona-virus-report.

“COVID-19 Deaths Dataset.” *Kaggle*, 29 July 2022, www.kaggle.com/datasets/dhruvildave/covid19-deaths-dataset?select=all_weekly_excess_deaths.csv.

World Health Organization (WHO), https://covid19.who.int/data