**Executive Summary**

On December 31, 2019, The World Health Organization China Country Office was informed of a number of cases of pneumonia of unknown etiology (unknown cause) detected in Wuhan, Hubei Province. All cases were connected to the Huanan Seafood Wholesale Market in Wuhan. (Source: <https://www.cdc.gov/museum/timeline/covid19.html>) Within a month, the virus was seen in many parts of the world including the United States of America. By March 11, 2020, the World Health Organization declared COVID-19 a pandemic. Today, the United States has the highest cumulative number of confirmed COVID-19 cases in the world at approximately 47.42 million cases, followed by India with 34.44 million cases (Source: <https://ourworldindata.org/covid-cases>).

This report captures the extent and spread of COVID-19 in the US and the corresponding government response. It studies how key intervention measures have affected the trends in the data related to various attributes of the virus. The project is based on our analysis of data from an open-source dataset provided by Our World in Data.

Our primary findings were as follows:

* Actions such as social distancing, mask mandates and stay at home orders were immediately taken to flatten the curve of the virus and it had a positive effect on containing the virus. The US initially saw a very slow growth in cases, but then failed to use that phase to figure out more targeted and state-level interventions.
* By spring of 2020, Covid-19 cases were spiking, and hospital systems were at risk of being overwhelmed by patients with life-threatening symptoms in the ICU. The US saw increasing death tolls.
* After the vaccines were available starting December 12, 2020, we could see a steep fall in daily new cases as well as hospitalizations and death, demonstrating that this was an effective method to control the virus.
* The US saw the first case of the Delta variant of COVID in March 2021 and started seeing a growth in covid cases again.
* Interventions such as masks mandate, social distancing and lockdowns were very effective in containing the virus but considering the pre-lockdown conditions such as population density, hospital occupancy and planning next steps accordingly is equally important.

**Research Question**

How has COVID-19 spread since the first case was discovered? How has the trend changed since the introduction of key intervention measures such as social distancing, mask requirements, vaccinations, and more?

**Literature Review**

The novel coronavirus (COVID-19) has been classified as a pandemic by WHO due to its world-wide spread, and countries around the world have adopted various intervention measures to contain it. As we started getting more information on the characteristics of the virus such as transmissibility and reproduction rate, the intervention measures evolved. Initially, countries adopted measures such as social distancing and lockdowns to contain and control virus transmission within communities. In their study on factors affecting COVID-19 spread, S. Roy and P. Gosh (2020) <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241165> found that post-lockdown Covid spread and death rates were highly influenced by their pre-lockdown counterparts. Population density, testing rate, airport traffic, and high age groups were discovered to be significant, while ethnicity, gender, healthcare index, homelessness and GDP had little or no impact on pandemic spread and mortality.

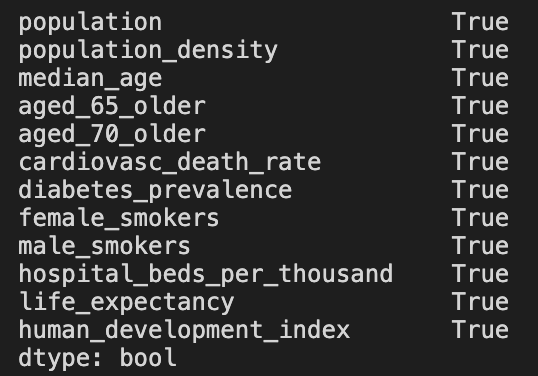
Mask mandates were another measure that was adopted. The paper Cooper, Daniel H. and Garga, Vaishali and Luengo-Prado, Maria Jose and Tang, Jenny, The Roles of Mobility and Masks in the Spread of COVID-19 (December, 2020) (Available at SSRN: <https://ssrn.com/abstract=3753011>) analyzes US state level data to estimate effects of mobility, and mask mandates on COVID-19 cases. The research found a strong, positive correlation between lagged mobility and growth rate of COVID-19 cases. They also found that mask mandates are effective in reducing at least half of the effects of increased mobility on growth rates of COVID-19. In addition, they estimate that total infections in the United States would have been 46.5 to 66.2 percent lower than they were on November 15, 2020 if mobility had remained fixed at its May 15, 2020 level.

Vaccines were available by the end of the year 2020. As the number of people vaccinated increased, the number of new COVID-19 cases decreased. However, the first case of a new variant of COVID-19, called the Delta variant, was identified in the United States in March 2021 which was followed by another steep rise in COVID-19 cases.

**Data Source and Methodology**

We utilized the COVID-19 Dataset provided by Our World in Data to conduct our analysis. The data contains information about vaccinations, tests & positivity rates, hospitalization & ICU numbers, confirmed cases, confirmed deaths, and more in relation to the COVID-19 pandemic from around the world since its beginning. The data is compiled from a number of sources, including the Center for Systems Science and Engineering at John Hopkins University, European Centre for Disease Prevention and Control, various governmental sources, official reports, and more. A full description of the data sources included in this dataset can be found at the official [Github Repository](https://github.com/owid/covid-19-data/tree/master/public/data/) for the dataset. While some variables are updated daily, others are updated weekly or periodically depending on the availability of data from the official sources.

**Data Cleaning**

We narrowed the provided dataset of 134,670 observations of 67 features to a set of 669 observations of 39 features to focus on the most relevant variables for our analysis. To do this, we first filtered the data to only look at the records for the United States. We then removed columns which we knew were not relevant such as smoothed case counts, continent, GDP, and more. The next step was to check the values contained in the remaining columns by calculating their descriptive statistics. In doing this, we found that there were 12 columns for which the range of values was 0:  


The range of these columns being 0 indicated that there was some constant value for every observation in these columns. However, the only one of these columns that a constant value made sense for was hospital\_beds\_per\_thousand, so we dropped the remaining columns, leaving us with 39 features in our dataset.

We then looked at the null values contained in the data, and found that most columns did, in fact, contain null values. However, due to the nature of the data, we concluded that the null values for most columns were actually instances where no data was recorded for that particular measure. In these cases, we replaced the null value with a 0. However, we also found that some columns had only null values and therefore contained no meaningful information for our analysis. In this case, the column was dropped. The only column that did not fit either of these cases was the reproduction\_rate column which measures how fast the virus is spreading. For this column, a null value indicated that we did not have enough data to calculate the correct rate, which was not the same as saying that the reproduction rate was 0. For this reason, we left the null values in this column. At this point we were ready to proceed with our analysis.

**Analysis**

We decided to examine the changes in trends given the implementation of various intervention measures. Specifically, we examined when social distancing measures were implemented, stay at home orders were issued, mask mandates were announced, vaccinations became available, and the Delta variant appeared. To do this, we had to find specific dates for these milestones:

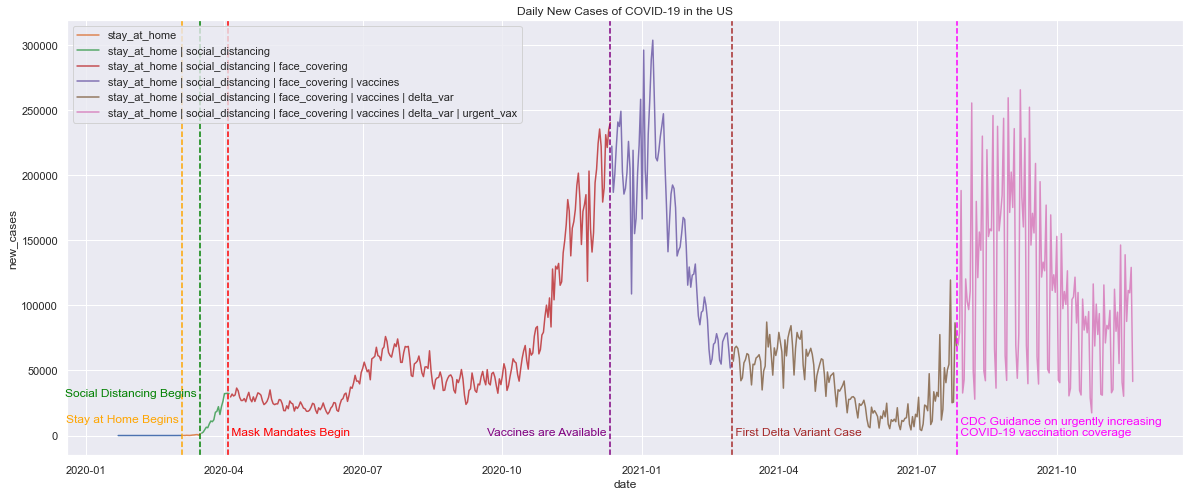
* According to the [Kaiser Family Foundation](https://www.kff.org/policy-watch/stay-at-home-orders-to-fight-covid19/), the first stay at home order was announced in King County in Washington state on March 4, 2020.
* According to [NPR](https://www.npr.org/2020/03/16/816658125/white-house-announces-new-social-distancing-guidelines-around-coronavirus), the White House announced social distancing guidelines on March 16, 2020.
* According to [Wikipedia](https://en.wikipedia.org/wiki/Face_masks_during_the_COVID-19_pandemic_in_the_United_States#Timeline), the CDC issued the first federal guidance recommending non-medical face coverings to be worn on April 3, 2020.
* According to the [FDA](https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-frequently-asked-questions#:~:text=On%20December%2011%2C%202020,%29%20of%20a%20vaccine.), the first Emergency Use Authorization for a COVID vaccine was granted on December 11, 2020 for the Pfizer-BioNTech Vaccine.
* According to the [The NY Times](https://www.nytimes.com/2021/06/22/health/delta-variant-covid.html), the Delta variant was first identified in the United States in March 2021.
* On July 27, 2021 the [CDC](https://www.cdc.gov/coronavirus/2019-ncov/variants/delta-variant.html) released updated guidance on the need for urgently increasing COVID-19 vaccination coverage and a recommendation for everyone in areas of substantial or high transmission to wear a mask in public indoor places, even if they are fully vaccinated.

We then created new columns in our dataset to indicate which of these measures had been implemented for each observation. At this point, we were ready to begin our analysis. We decided to explore the following relationships in our data:

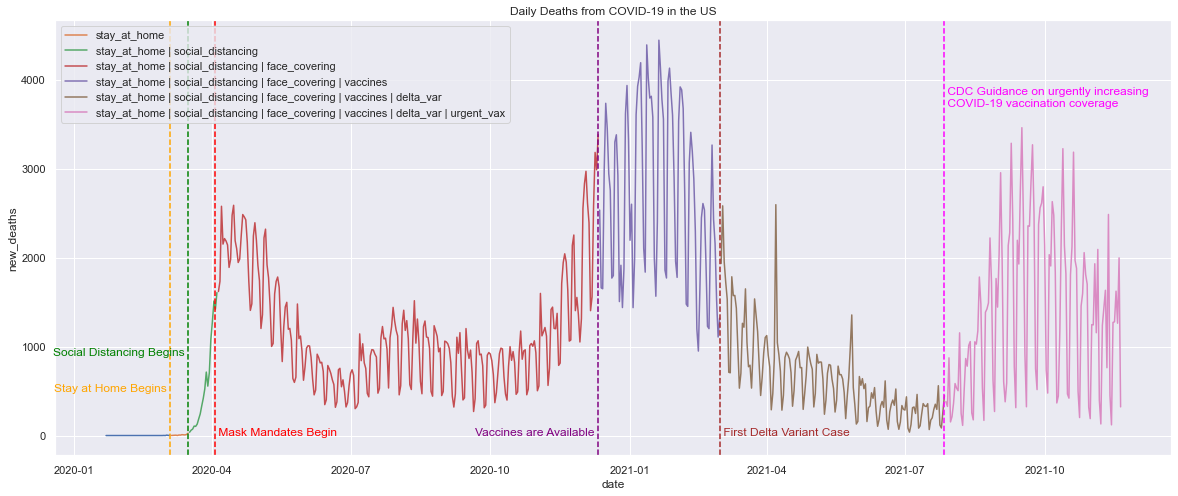
* Daily New Cases over time
* Daily Deaths over time
* Daily ICU Patients over time
* Daily Hospitalizations over time
* Reproduction Rate over time
* Daily New Cases vs Number of Fully Vaccinated People
* Daily New Cases vs Number of Vaccinated People (at least 1 dose)
* Test Positivity Rate vs Number of Fully Vaccinated People
* Daily New Cases vs Policy Stringency Index
* Overall correlation of all variables in the dataset

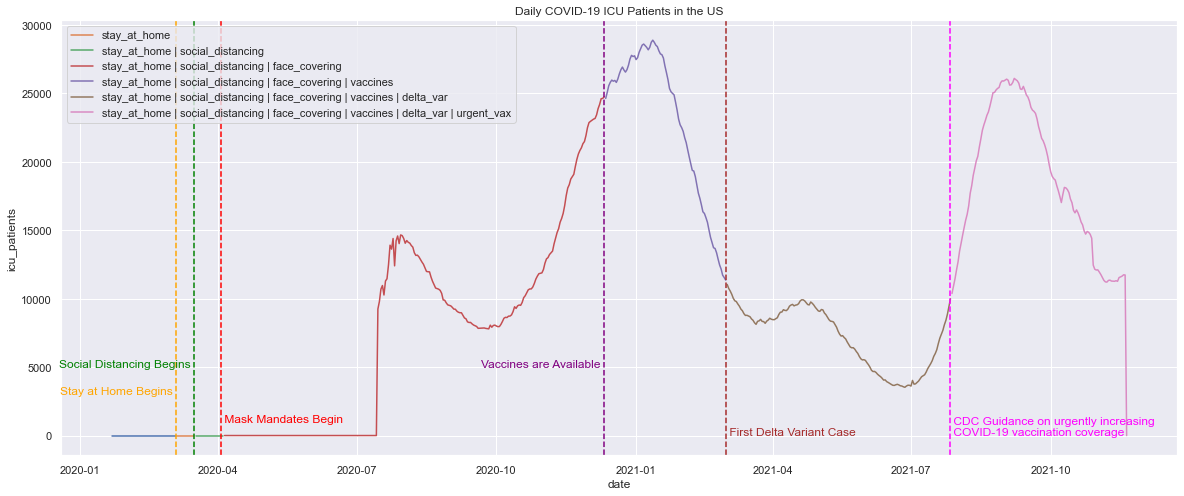
**Results**

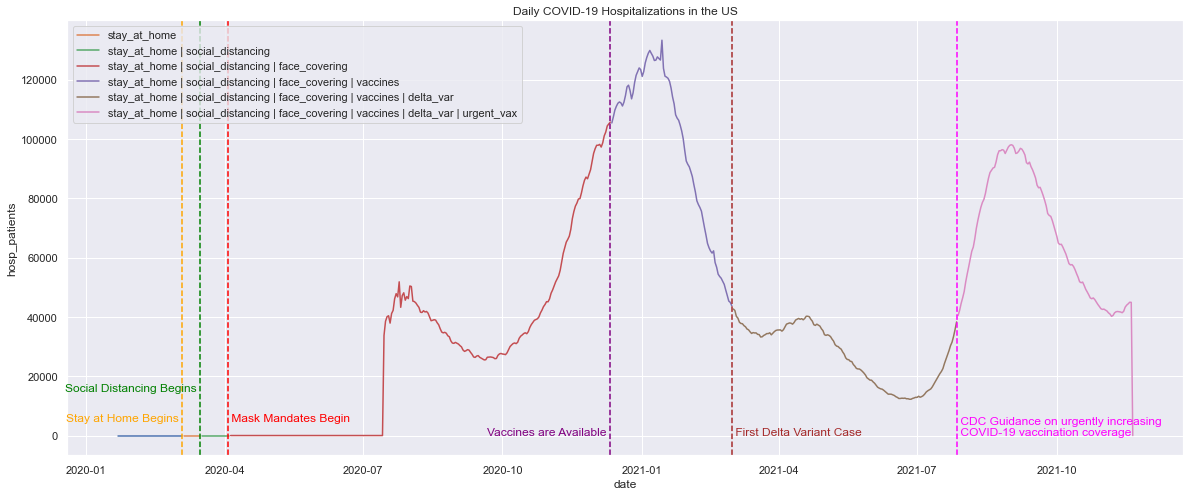
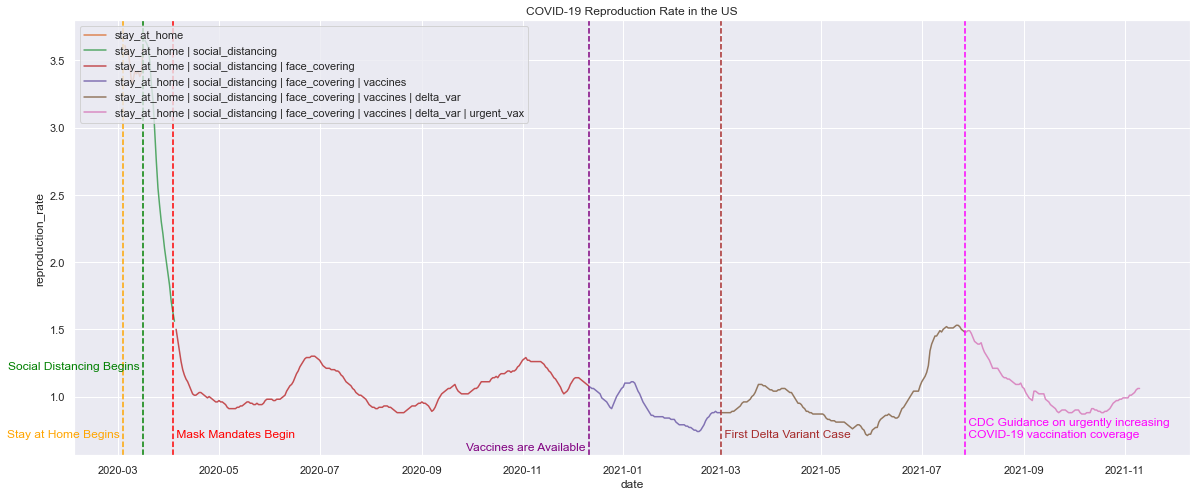
From the data we collected, we used line plots to plot daily COVID-19 data on new cases, deaths, ICU patients, hospitalizations, and reproduction rate.



The plot above shows the trend in daily COVID-19 cases from January 2020 to present day. We can see that the intervention measures were taken immediately after a rise in Covid-19 cases were seen in the US. The yellow vertical line shows the date when stay at home order was introduced, the green vertical line represents the date when social distancing began, and the red vertical line shows the date mask mandates began. The brown vertical line shows when the Delta variant appeared, and the pink vertical line shows when the CDC improved its guidance regarding vaccinations. We can see a downward trend in new Covid cases soon after the interventions were introduced. According to CDC, COVID-19 symptoms may appear 2-14 days after exposure to the virus, which explains the lag in decrease in COVID-19 cases after the interventions were introduced. We saw another rise in COVID-19 cases around June/July 2021. This is when CDC updated guidance on the need for urgently increasing COVID-19 vaccination coverage and a recommendation for everyone in areas of substantial or high transmission to wear a mask in public indoor places, even if they are fully vaccinated. Soon after that, we saw a decrease in COVID-19 cases again.

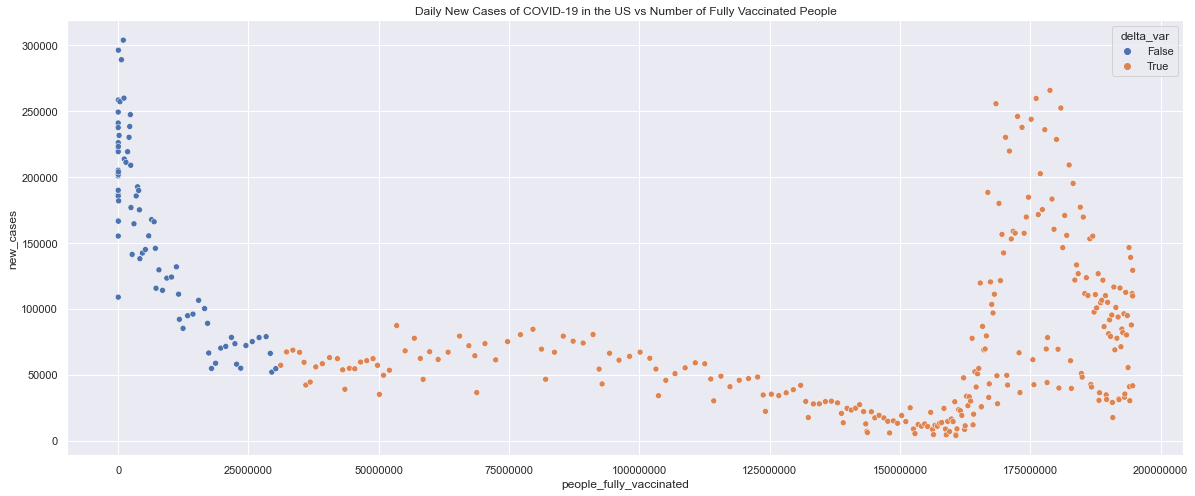
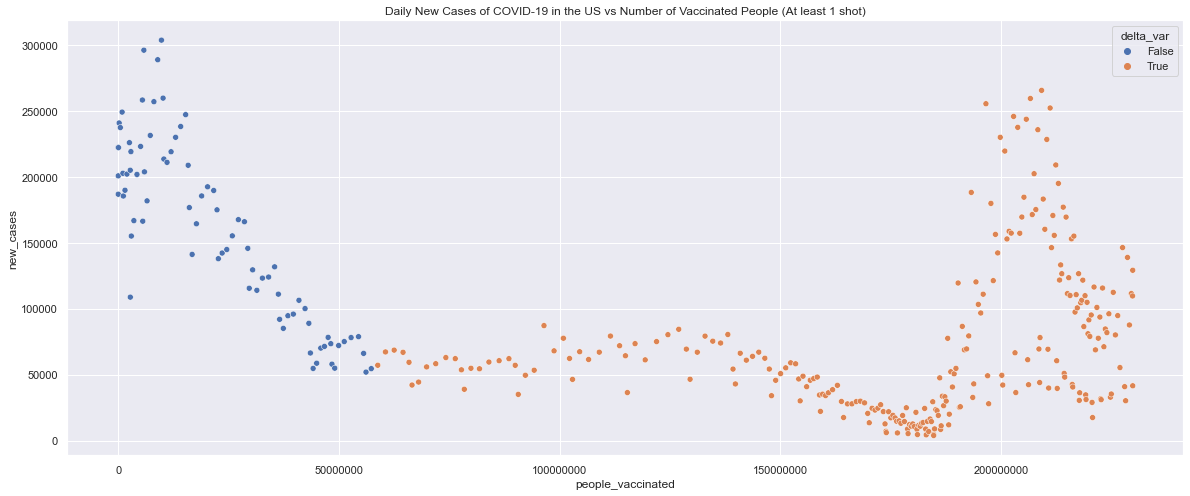
We saw a similar trend in data for new deaths, new ICU patients and new hospitalizations. The graphs are shown below. 

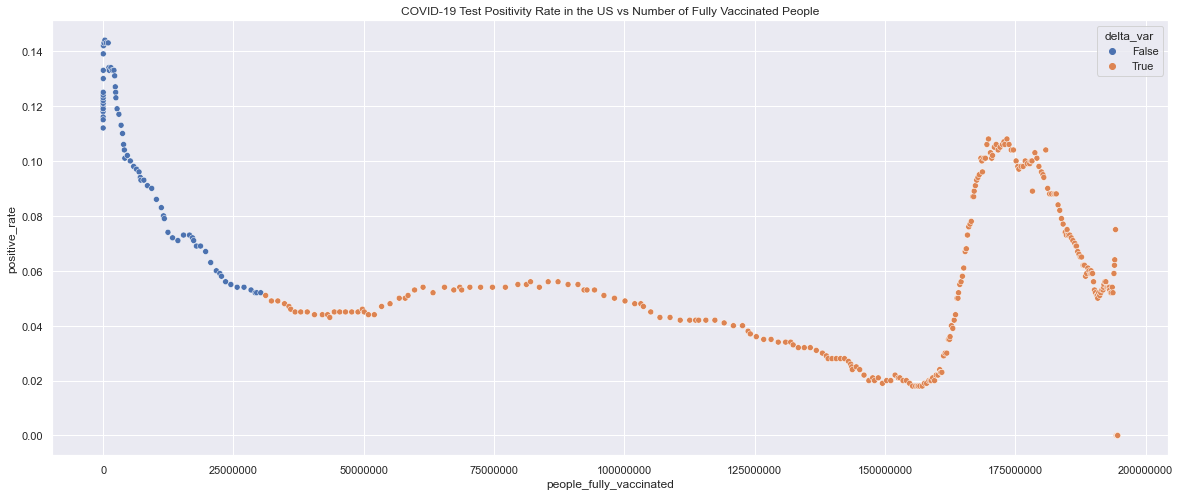


The above visualizations clearly demonstrate the effectiveness of the various intervention measures introduced in controlling the spread of the virus.

Next, we made scatter plots to visualize number of vaccinated people against daily new cases. We saw that as the number of fully vaccinated people increased, the daily new COVID-19 cases decreased until the new COVID-19 variant was found. We have two different scatter plots; one showing new cases against fully vaccinated people while the other shows new cases against all vaccinated people (full and partial vaccinations). The scatter plot looks similar in both cases.

These negative trend in these visuals clearly displays the effectiveness of the vaccines in curbing the spread of the virus, at least until the Delta variant arrived. We also examined the relationship between Covid test positivity rates and the number of fully vaccinated people, and found a negative correlation here as well (displayed below). 

**Recommendations**

We found that interventions such as masks mandates, social distancing and lockdowns were very effective in containing the virus but considering the pre-lockdown conditions such as population density, hospital occupancy and planning next steps accordingly is equally important.

For further research, we could also look at state-level data. Data on which state had stricter policies, higher population density, more hospitals with greater availability of ICU rooms and equipment, state-level lock down policies and timelines facilitates in-depth study on this topic.

**Appendix**

The description of each variable used in the dataset are as follows:

|  |  |
| --- | --- |
| **Confirmed cases** |  |
| Variable | Description |
| total\_cases | Total confirmed cases of COVID-19 |
| new\_cases | New confirmed cases of COVID-19 |
| new\_cases\_smoothed | New confirmed cases of COVID-19 (7-day smoothed) |
| total\_cases\_per\_million | Total confirmed cases of COVID-19 per 1,000,000 people |
| new\_cases\_per\_million | New confirmed cases of COVID-19 per 1,000,000 people |
| new\_cases\_smoothed\_per\_million | New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people |
| **Confirmed deaths** |  |
| Variable | Description |
| total\_deaths | Total deaths attributed to COVID-19 |
| new\_deaths | New deaths attributed to COVID-19 |
| new\_deaths\_smoothed | New deaths attributed to COVID-19 (7-day smoothed) |
| total\_deaths\_per\_million | Total deaths attributed to COVID-19 per 1,000,000 people |
| new\_deaths\_per\_million | New deaths attributed to COVID-19 per 1,000,000 people |
| new\_deaths\_smoothed\_per\_million | New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people |
| **Hospital & ICU** |  |
| Variable | Description |
| icu\_patients | Number of COVID-19 patients in intensive care units (ICUs) on a given day |
| icu\_patients\_per\_million | Number of COVID-19 patients in intensive care units (ICUs) on a given day per 1,000,000 people |
| hosp\_patients | Number of COVID-19 patients in hospital on a given day |
| hosp\_patients\_per\_million | Number of COVID-19 patients in hospital on a given day per 1,000,000 people |
| weekly\_icu\_admissions | Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week |
| weekly\_icu\_admissions\_per\_million | Number of COVID-19 patients newly admitted to intensive care units (ICUs) in a given week per 1,000,000 people |
| weekly\_hosp\_admissions | Number of COVID-19 patients newly admitted to hospitals in a given week |
| weekly\_hosp\_admissions\_per\_million | Number of COVID-19 patients newly admitted to hospitals in a given week per 1,000,000 people |
| **Policy responses** |  |
| Variable | Description |
| stringency\_index | Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) |
| Reproduction rate |  |
| Variable | Description |
| reproduction\_rate | Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database |
| Tests & positivity |  |
| Variable | Description |
| total\_tests | Total tests for COVID-19 |
| new\_tests | New tests for COVID-19 (only calculated for consecutive days) |
| total\_tests\_per\_thousand | Total tests for COVID-19 per 1,000 people |
| new\_tests\_per\_thousand | New tests for COVID-19 per 1,000 people |
| new\_tests\_smoothed | New tests for COVID-19 (7-day smoothed). For countries that don't report testing data on a daily basis, we assume that testing changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window |
| new\_tests\_smoothed\_per\_thousand | New tests for COVID-19 (7-day smoothed) per 1,000 people |
| positive\_rate | The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests\_per\_case) |
| tests\_per\_case | Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive\_rate) |
| tests\_units | Units used by the location to report its testing data |
| **Vaccinations** |  |
| Variable | Description |
| total\_vaccinations | Total number of COVID-19 vaccination doses administered |
| people\_vaccinated | Total number of people who received at least one vaccine dose |
| people\_fully\_vaccinated | Total number of people who received all doses prescribed by the vaccination protocol |
| total\_boosters | Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol) |
| new\_vaccinations | New COVID-19 vaccination doses administered (only calculated for consecutive days) |
| new\_vaccinations\_smoothed | New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window |
| total\_vaccinations\_per\_hundred | Total number of COVID-19 vaccination doses administered per 100 people in the total population |
| people\_vaccinated\_per\_hundred | Total number of people who received at least one vaccine dose per 100 people in the total population |
| people\_fully\_vaccinated\_per\_hundred | Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population |
| total\_boosters\_per\_hundred | Total number of COVID-19 vaccination booster doses administered per 100 people in the total population |
| new\_vaccinations\_smoothed\_per\_million | New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population |
| new\_people\_vaccinated\_smoothed | Daily number of people receiving their first vaccine dose (7-day smoothed) |
| new\_people\_vaccinated\_smoothed\_per\_hundred | Daily number of people receiving their first vaccine dose (7-day smoothed) per 100 people in the total population |
| **Others** |  |
| Variable | Description |
| iso\_code | ISO 3166-1 alpha-3 – three-letter country codes |
| continent | Continent of the geographical location |
| location | Geographical location |
| date | Date of observation |
| population | Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population\_latest.csv for full list of sources |
| population\_density | Number of people divided by land area, measured in square kilometers, most recent year available |
| median\_age | Median age of the population, UN projection for 2020 |
| aged\_65\_older | Share of the population that is 65 years and older, most recent year available |
| aged\_70\_older | Share of the population that is 70 years and older in 2015 |
| gdp\_per\_capita | Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available |
| extreme\_poverty | Share of the population living in extreme poverty, most recent year available since 2010 |
| cardiovasc\_death\_rate | Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people) |
| diabetes\_prevalence | Diabetes prevalence (% of population aged 20 to 79) in 2017 |
| female\_smokers | Share of women who smoke, most recent year available |
| male\_smokers | Share of men who smoke, most recent year available |
| handwashing\_facilities | Share of the population with basic handwashing facilities on premises, most recent year available |
| hospital\_beds\_per\_thousand | Hospital beds per 1,000 people, most recent year available since 2010 |
| life\_expectancy | Life expectancy at birth in 2019 |
| human\_development\_index | A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506 |

Source:<https://github.com/owid/covid-19-data/tree/master/public/data#readme>