**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 planning on next steps after that is equally important.

**Research Question**

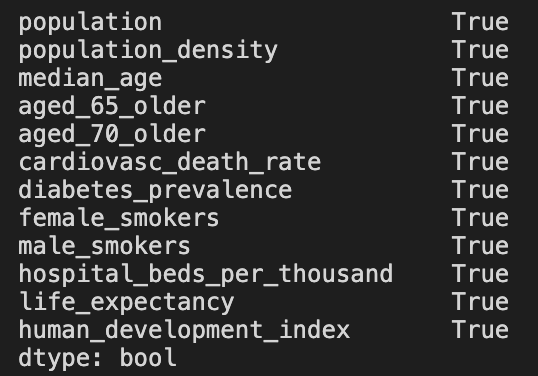
How has COVID-19 spread since the first case was discovered? How has the trend changed since the introduction of key intervention measures such as social distancing, mask requirements, vaccinations, and more?

**Literature Review**

**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 being 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**