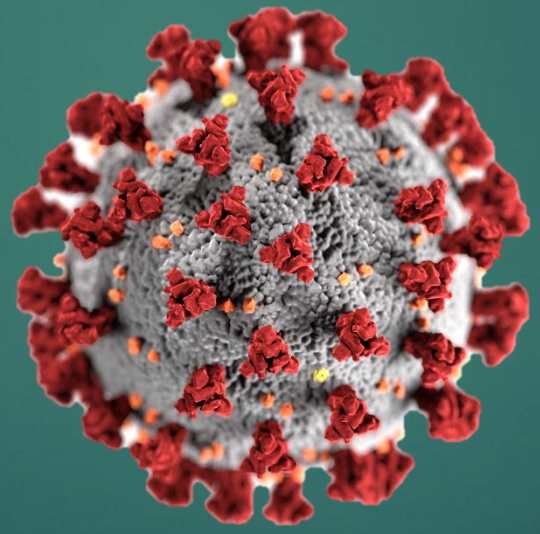
# COVID-19 Cases Forecast



DR. NAWANA COYLE

# Problem

Looking at the past data and the patterns, can we predict new cases and deaths associated with COVID-19 for the next 6 months?

# Why It's Important to Predict Incidents

- According to WHO, there are over 455 millions of confirmed COVID-19 cases and 6.04 million related deaths in the world.
- Forecasting help prepare for the future needs.
- Brainstorm solutions to minimize the cases with the subject matter experts and authorities.
- Educate the public on best practices ways to minimize cases.

# Data for The Project

- Original data for this project came from Johns Hopkins Center for Systems Science and Engineering page in GitHub at https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

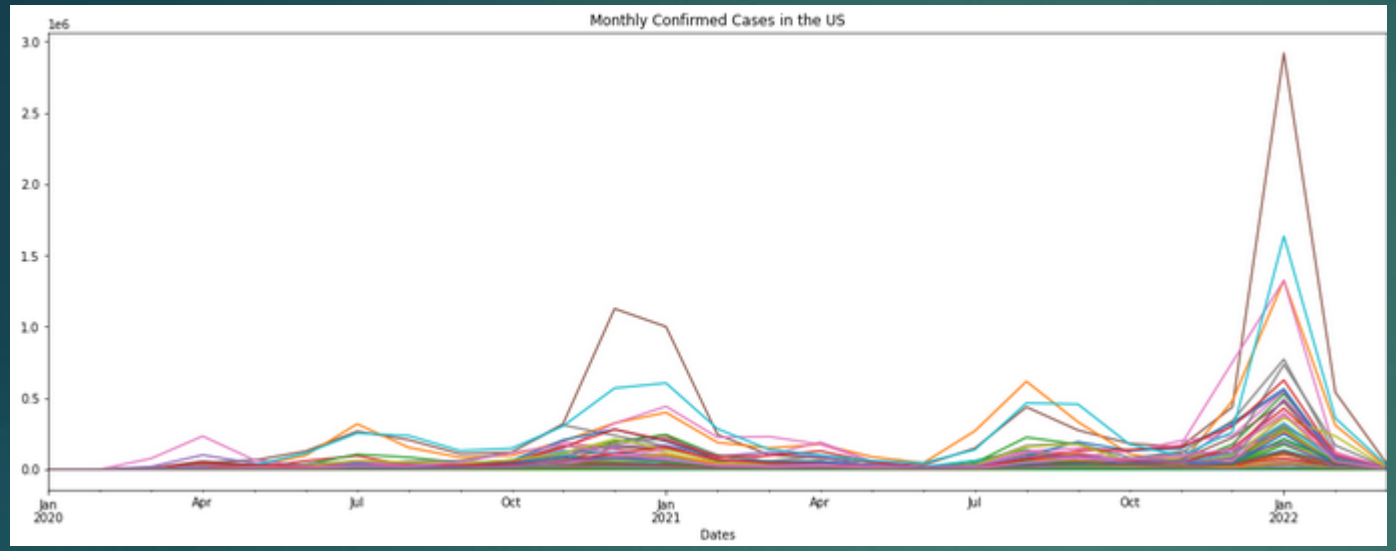
# Data Cleaning and Processing

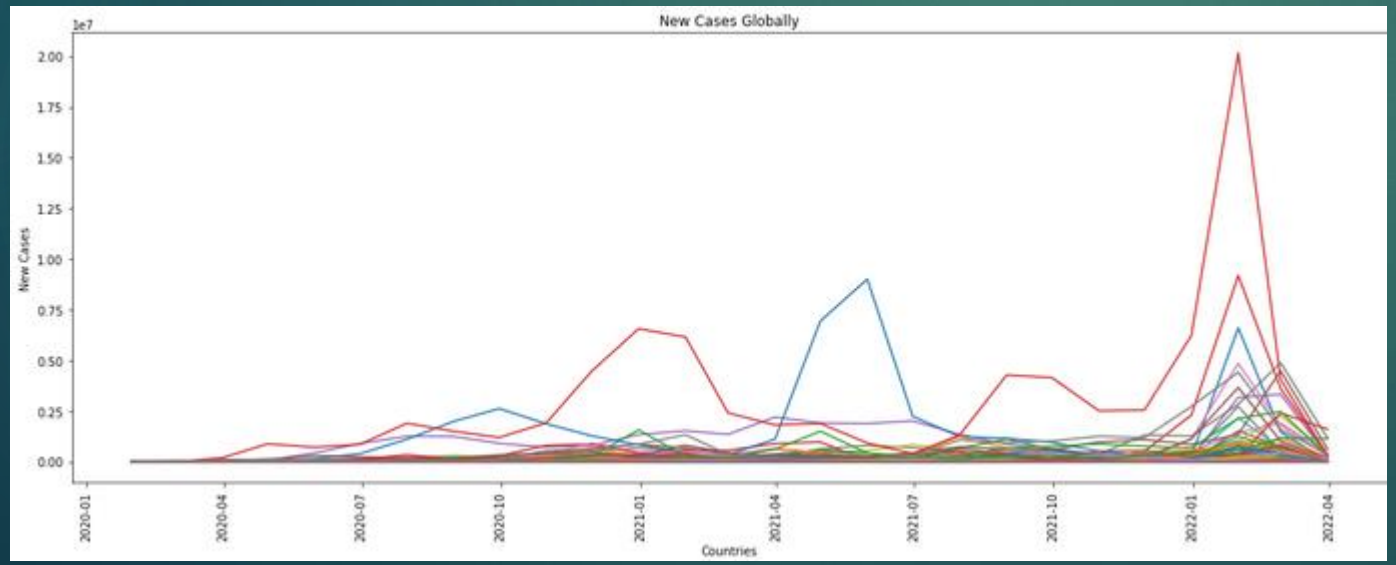
- 5 different files:
  - Confirmed cases in the US US.
  - Deaths in the US
  - Confirmed Cases Globally
  - Deaths in the US
  - Recovered in the US
- Dropped unnecessary columns → for easy data processing
- Renamed columns → for easier access
- Calculated case totals by the state →understand the patterns

- Formatted the date column → Apply models
- Created separate monthly and yearly datasets
- Using .dff() function, isolated the daily occurrences, instead of, up to date occurrences.

| States<br>Dates | Alabama | Alaska |
|---|---|---|
| 2022-03-03 | 762.0 | 0.0 |
| 2022-03-04 | 573.0 | 704.0 |
| 2022-03-05 | 0.0 | 0.0 |
| 2022-03-06 | 0.0 | 0.0 |
| 2022-03-07 | 3710.0 | 467.0 |

# Exploratory Data Analysis (EDA)



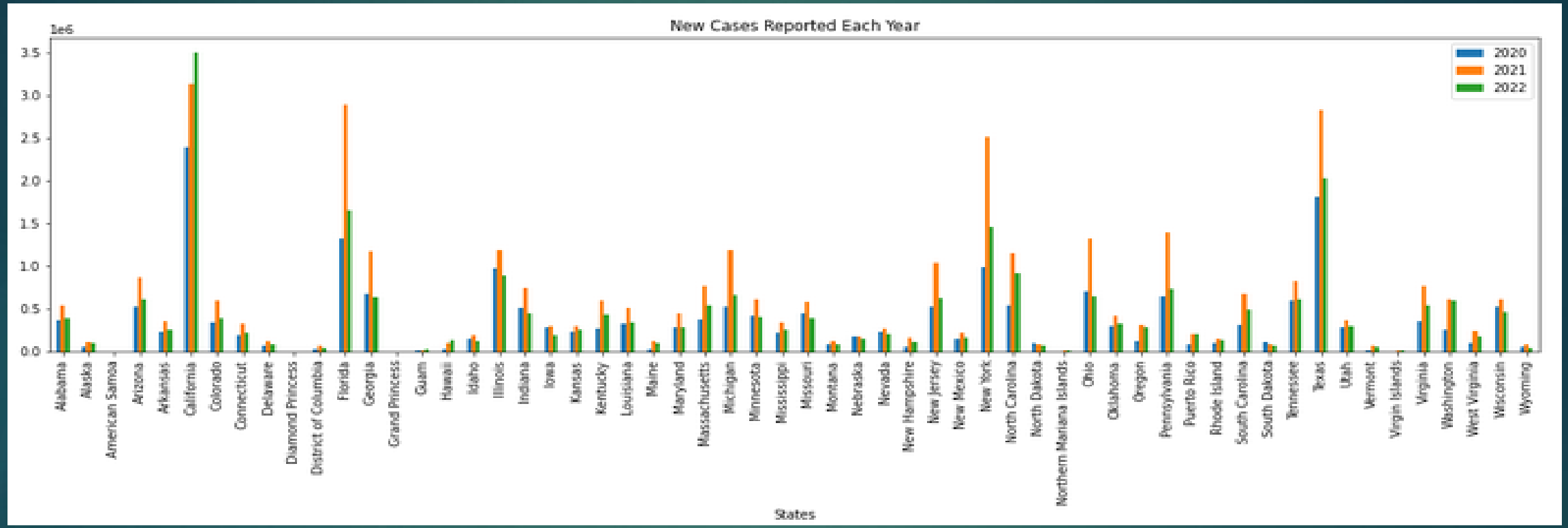

**Observation:**

- Dec, Jan → highest confirmed cases → US and Globally
- Spike in winter 2021> 2020 both in the US and globally

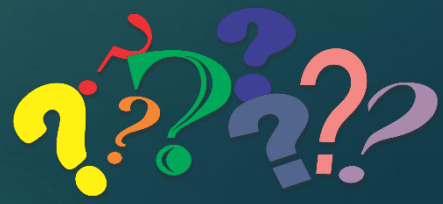
**Explanation:**

- People gather more due to holidays?
- More testing done during holidays?
- Both?

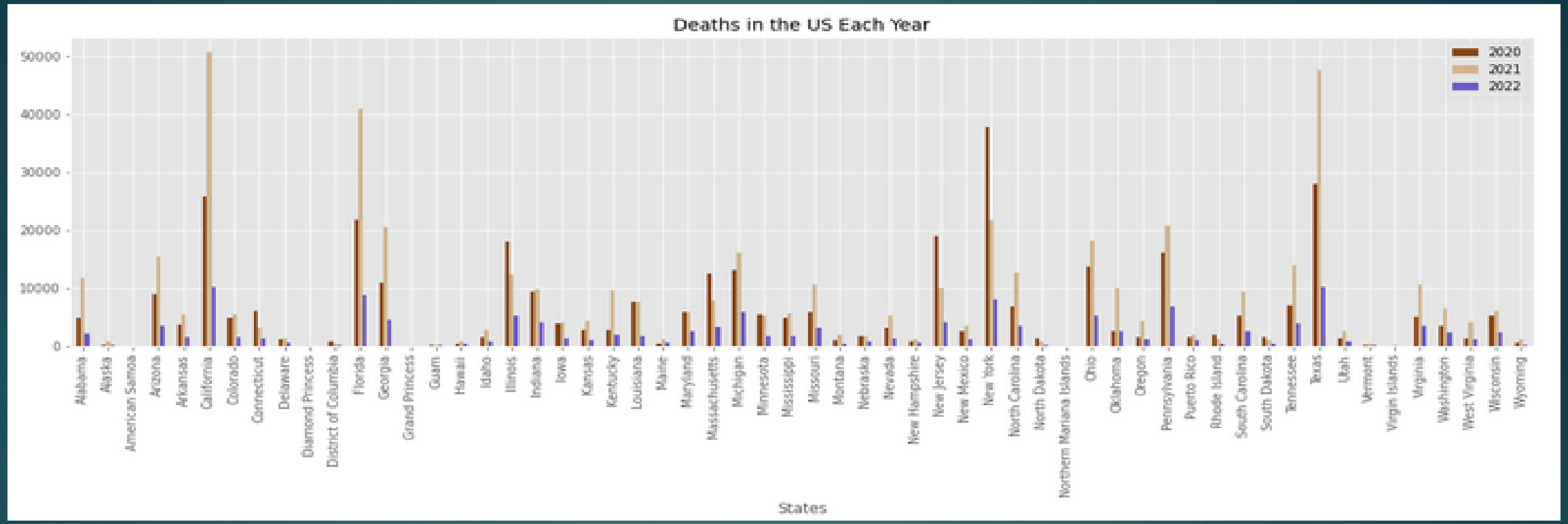
# New Cases in The US Each Year (EDA)




- Confirmed cases US in 2021 >2020
  1. Not a true representation ?
  2. No tests were available at the beginning of the pandemic?
  3. 2021 had a different variant with higher transmissibility?
  4. All?

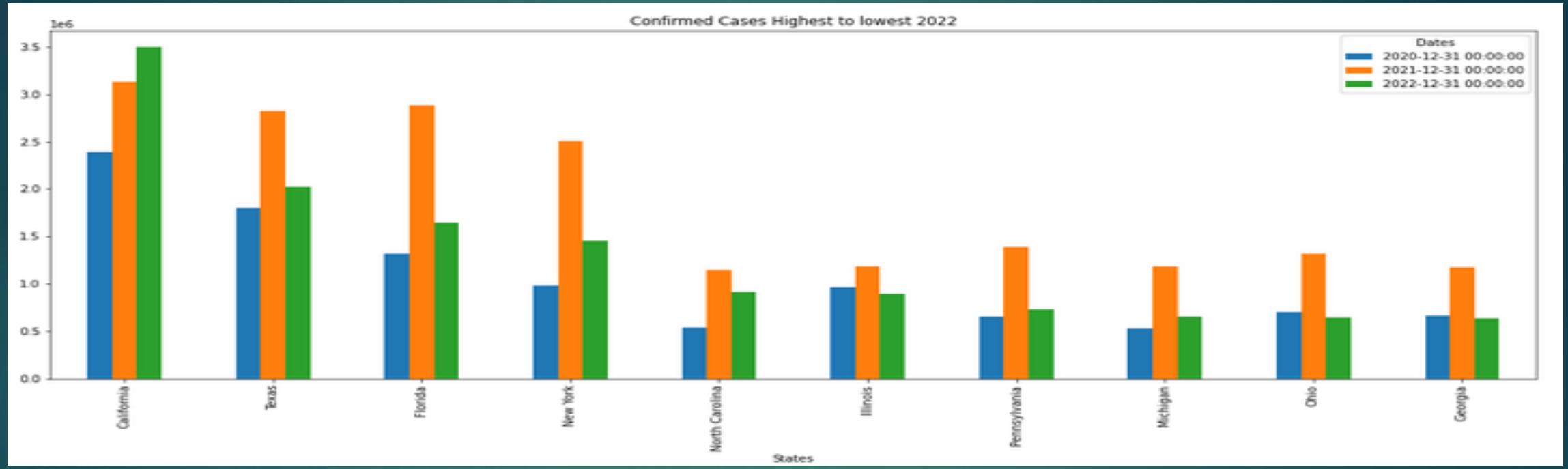


# Deaths in The US Each Year (EDA)



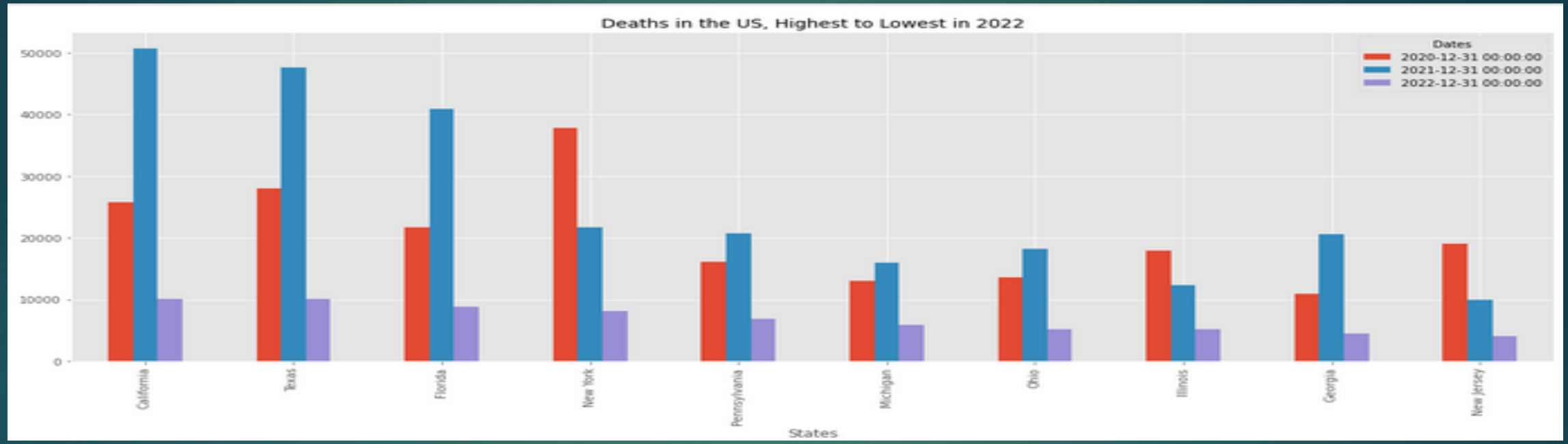

- Deaths caused  in the US in 2021 >2020
- Confirmed cases in many states, especially in CA by march 2022 is already higher than that of 2021 , but low death rate
  - A new variant with high transmissibility and low mortality rate?
  - OR not a data collection error?

# 10 States with The Highest Number of New Cases in the US (EDA)




- CA has the highest confirmed cases, followed by TX and FL respectively, which aligns with the order of population in each state .

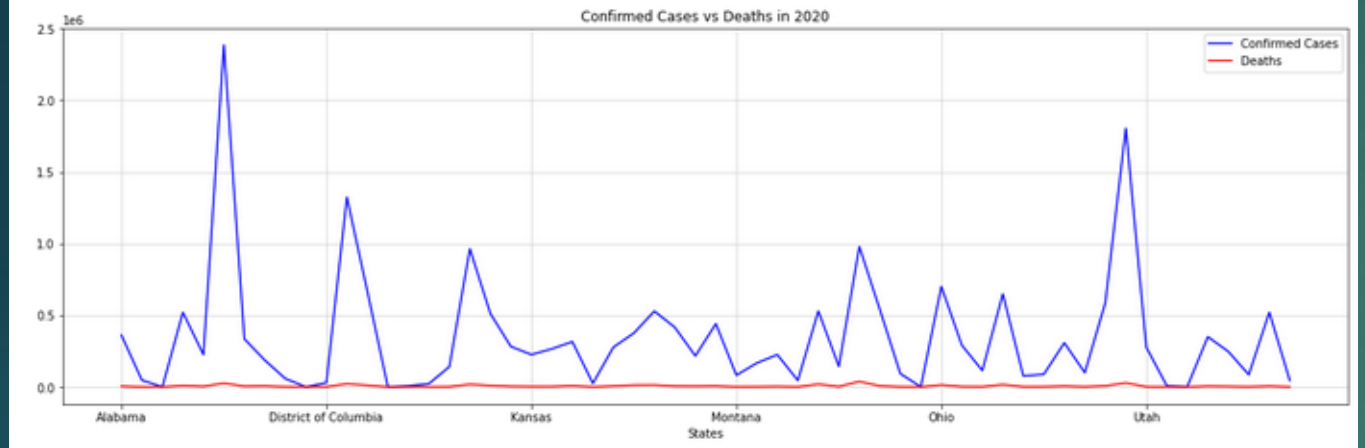
# 10 States with The Highest Number of Deaths in the US (EDA)



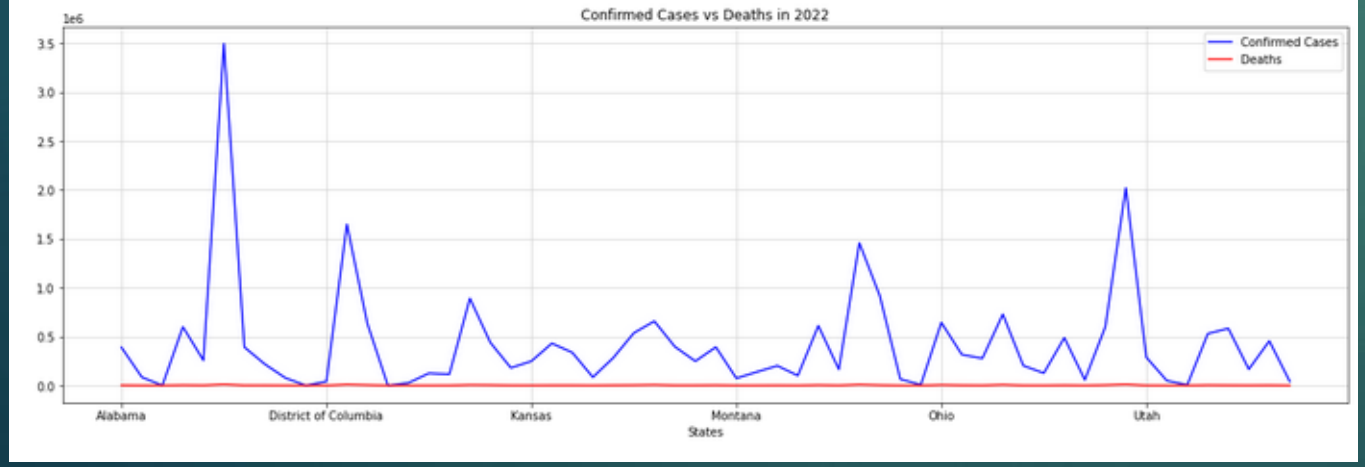

- NY has the 4th highest population and confirmed cases in the US in 2021; yet surprisingly low death rates in 2021, especially given that fact NY had the highest deaths in the entire nation in 2020.
- How did that happen?
  1. Does NY has an extra cautious population? OR
  2. Is it a result of an inaccurate  data reporting?

# Pattern of New Cases in the US (EDA)

```
comparison('2020')
```

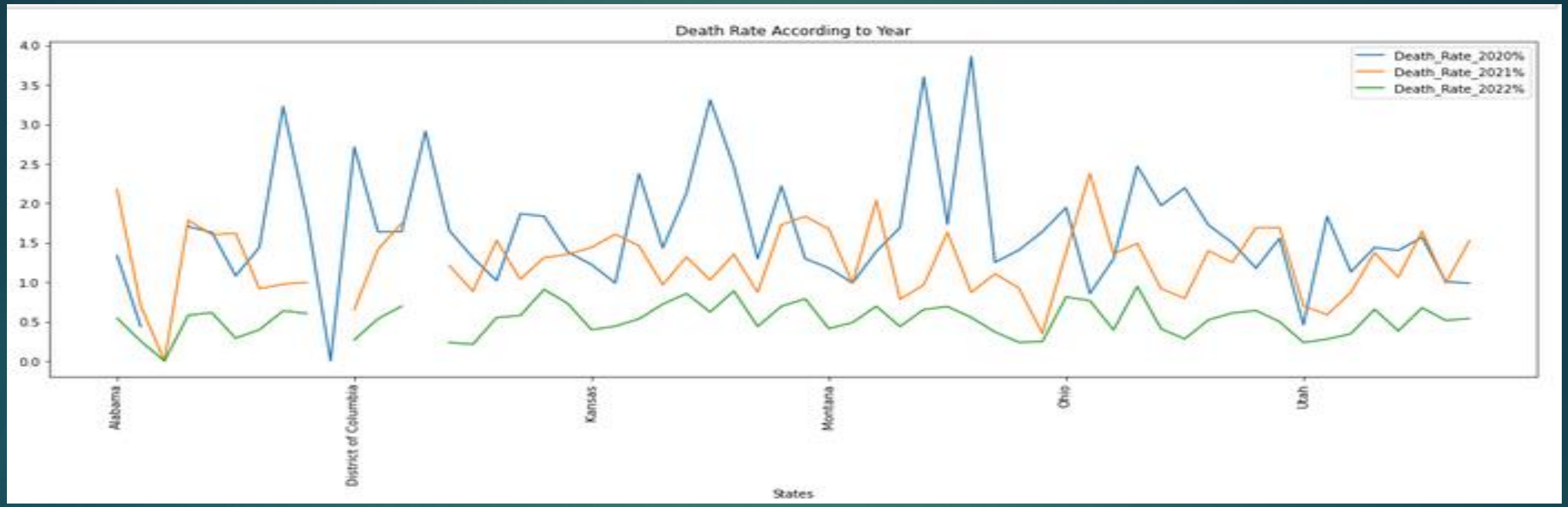



```
comparison('2022')
```




- Comparison of new cases in the US, 2020 vs 2022
- Number of confirmed cases throughout have a similar pattern year to year in each state.

# Death Rate in The Last 3 Yrs (EDA)

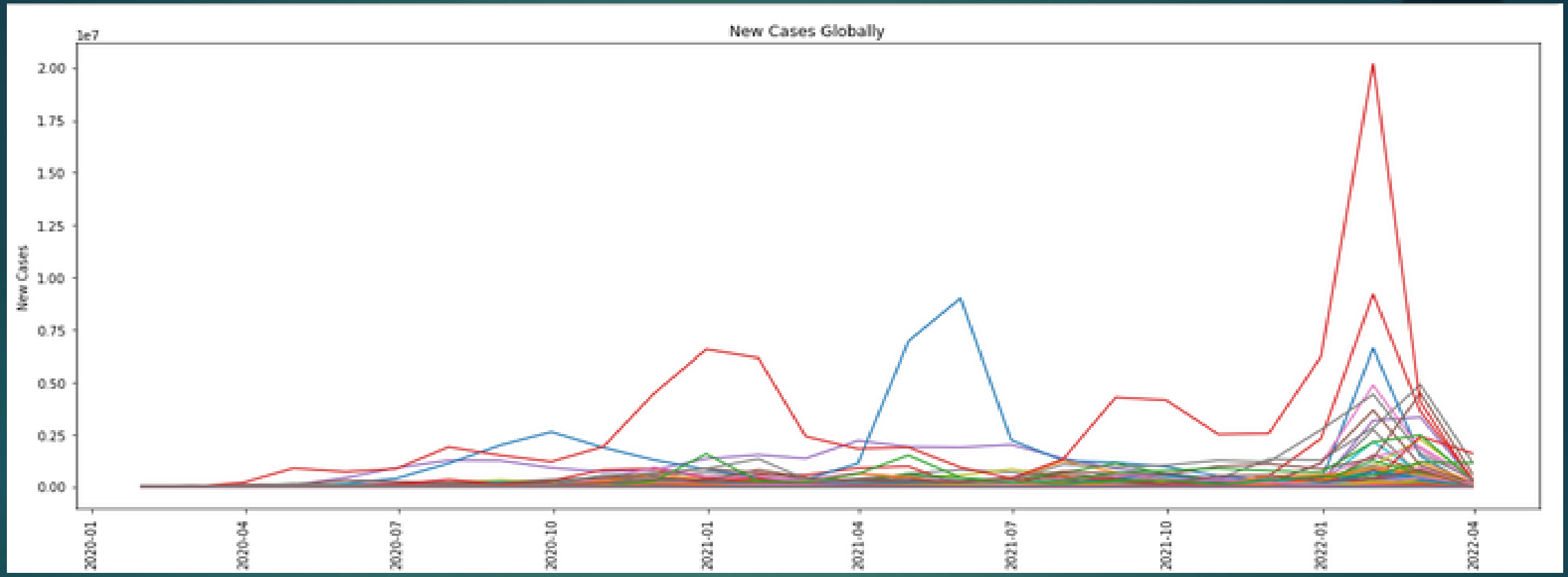



- Overall, death rate throughout the country declined each year.
- How ? → Change in virus variance and its mortality ?

  Better upstanding of the virus and management of patient care ?

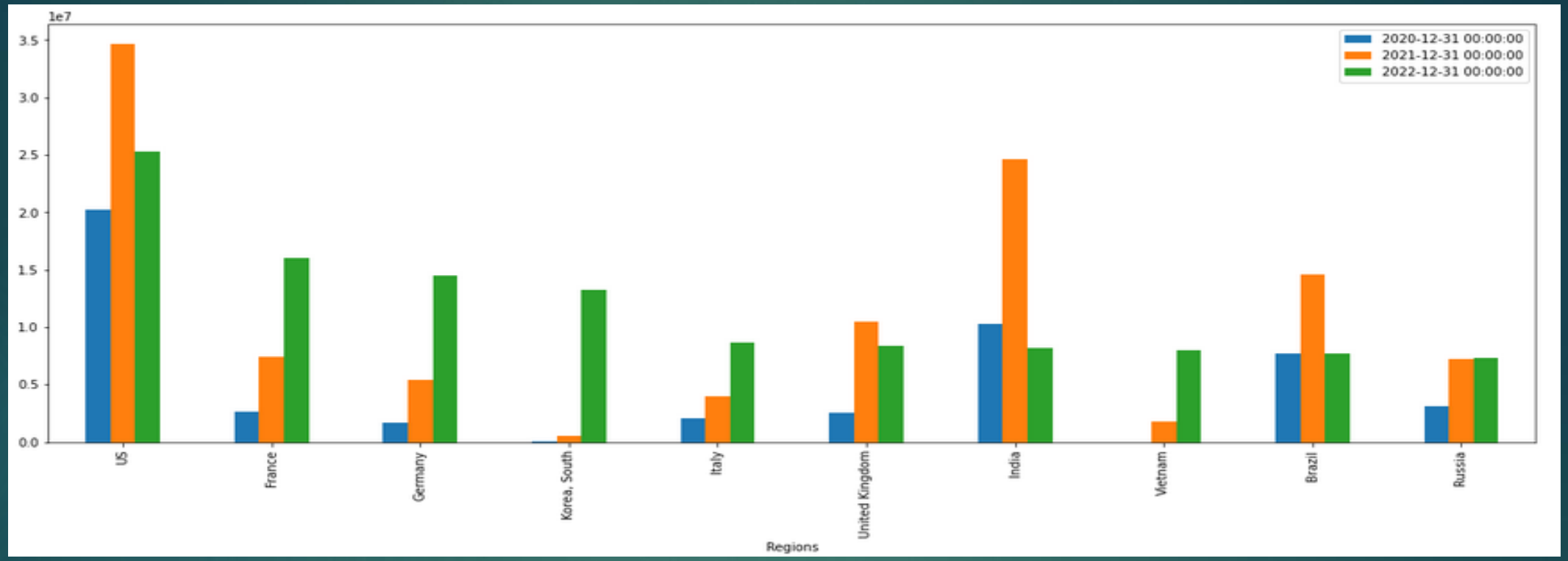
  All?

# Confirmed Cases – Global (EDA)




- Similar to the US, the confirmed cases around the world peaked in December and January of the following year.

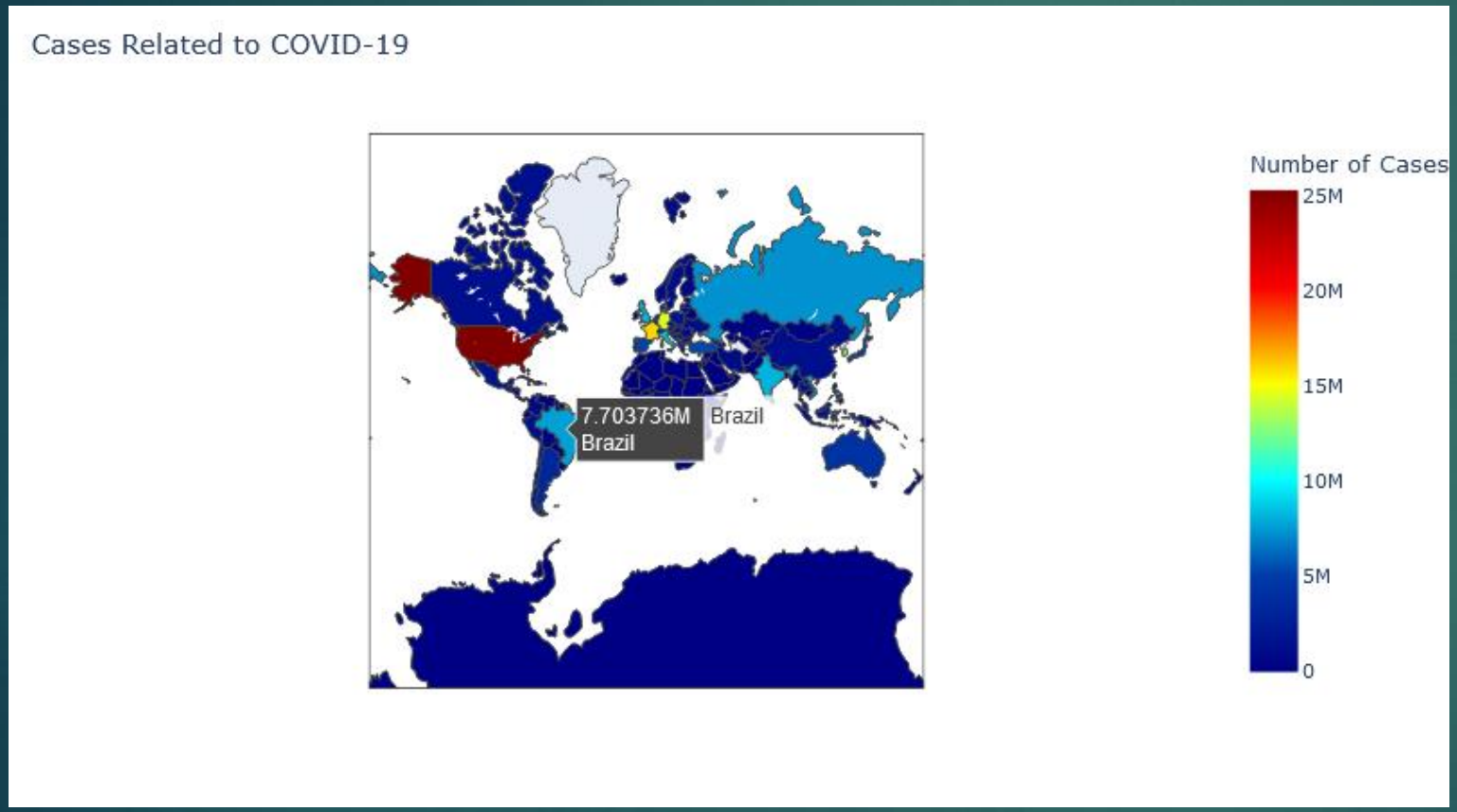
# Confirmed Cases – Global (EDA)




- Despite India and China having higher populations than the US, US still recorded highest number of confirmed cases
- Why? → Inaccurate reporting? OR

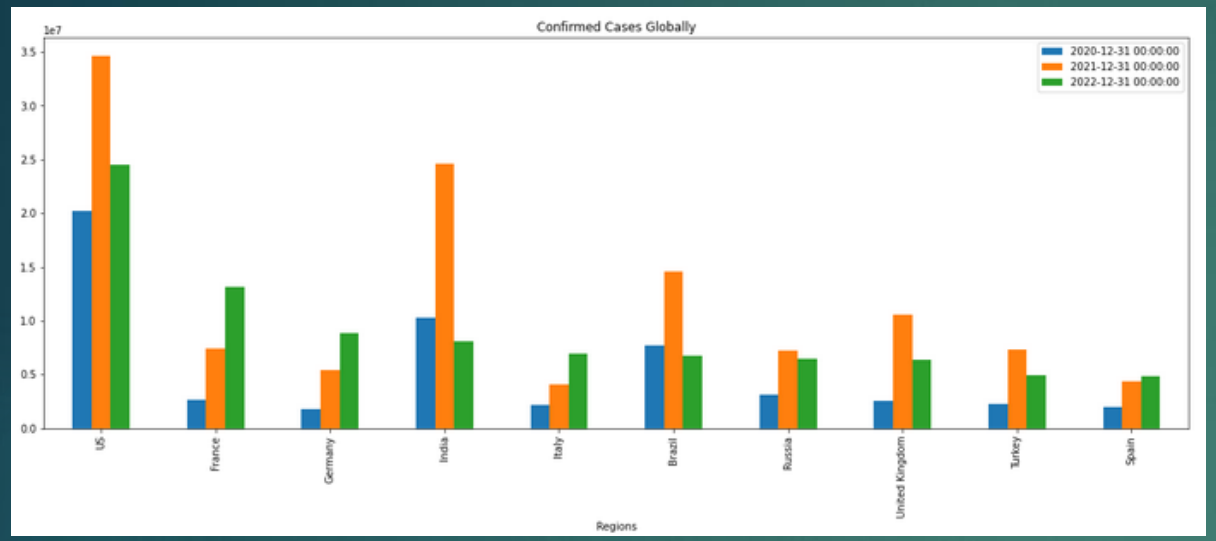
  Not having test kits in the other countries?

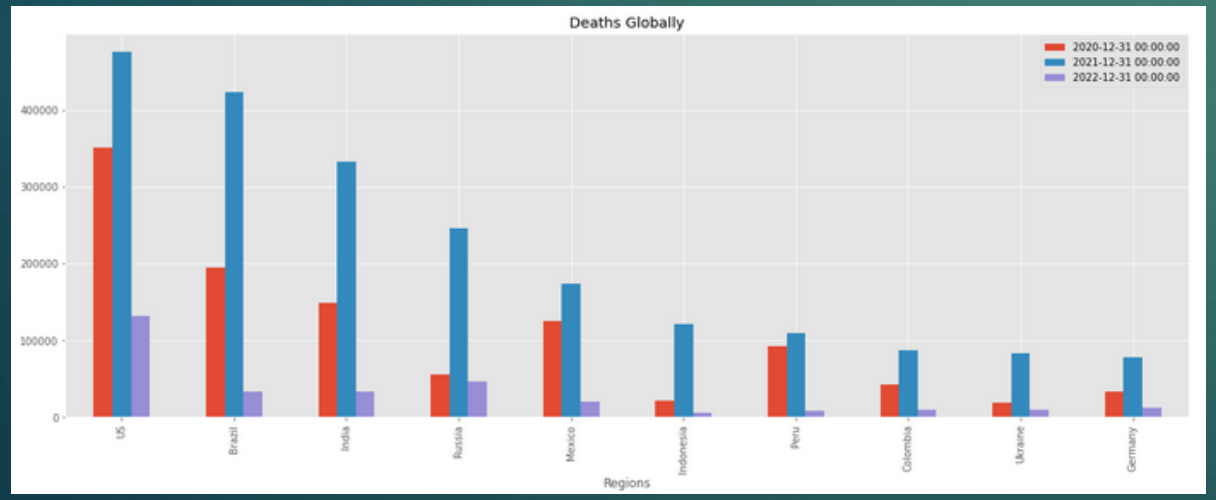
# Creating a Map




- Using ploty python library, data on prevalence for a given time frame was visualized.
- Hovering over an area was able to show the number of cases in that area for the time frame.
- The heatmap results aligns with the results created by the bar chart from high to low cases in 2022

# Confirmed Cases VS Deaths – Global (EDA)




- World Population in 2022,

China > India > US

- Confirmed cases in 2022,

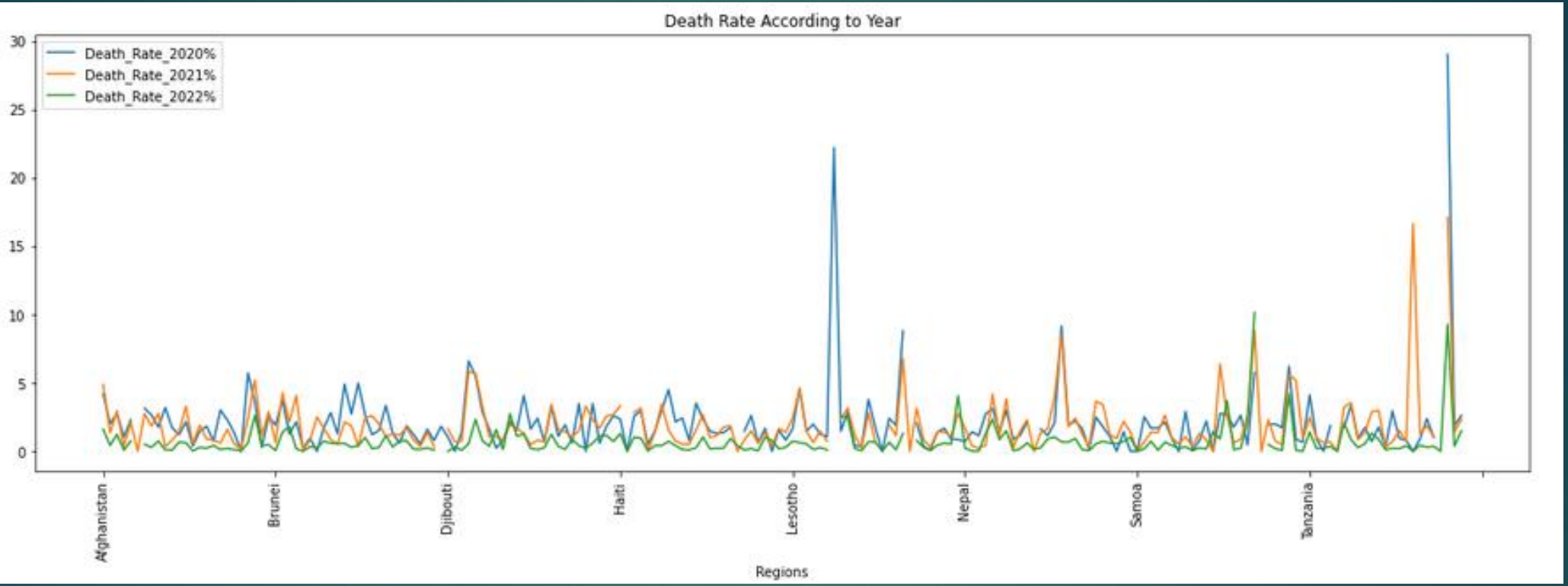
US > France > Germany > India >Italy

- Deaths in 2022,

US > Russia > India > Brazil > Mexico

- Why are the confirmed cases and deaths do not align with world populations in general? Better patient care ? OR Inaccurate tracking?

# Global Death Rate in The Last 3 Yrs




- Global death rate have come down over the last 3 years.
- Even though total number of deaths in almost every country was higher in 2021 compared to 2020, the death rate in 2021 have come down significantly in certain regions.

# Data Modeling

- Understanding the trends and patterns of new cases and help educate the general public, and have necessary supplies and equipment for the healthcare facilities.
- This is a project involving time forecasting future using prior data.
- Models used for the project:
  1. ARIMA model – Autoregressive Integrated Moving Average
  2. SARIMA model – Seasonal ARIMA

# Data Modeling

Success of the project: → Lowest RMSE (root mean square error)

"A low RMSE value indicates that the simulated and observed data are close to each other showing a better accuracy"

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

# Why Choose ARIMA Model

1. Works with data on a time frame.
2. Auto regression : Looks at data from past periods, p, when making predictions
3. Moving average : Future predictions are also made looking at the errors of the prior values, q.
4. Integration of past values and errors of prior values, d.

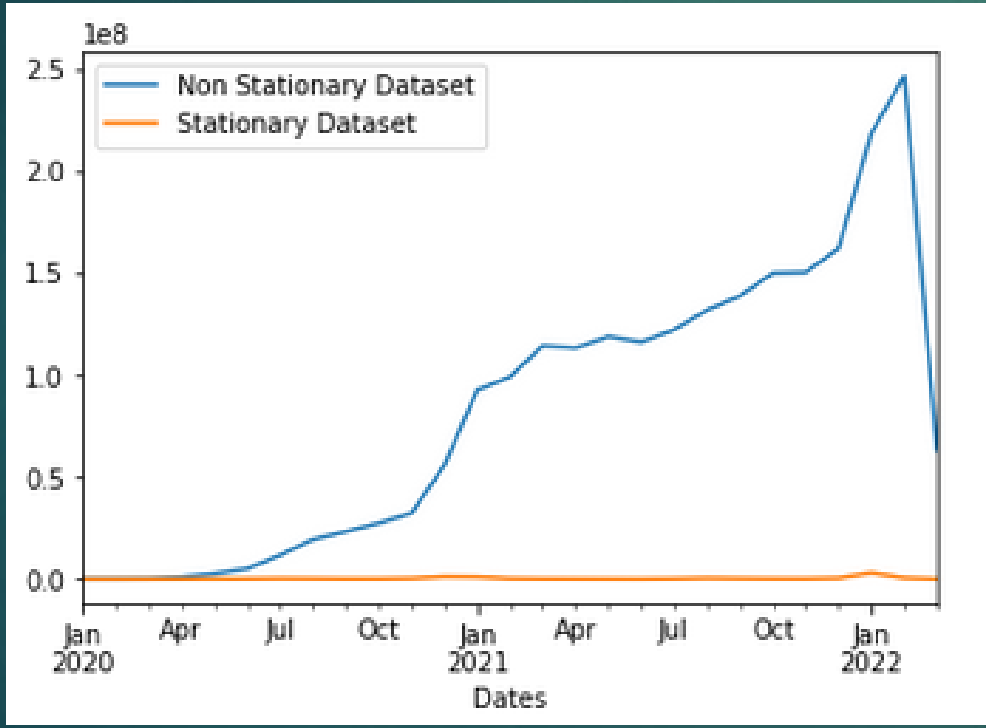
# Why Choose SARIMA Model

1. SARIMA - an extension of ARIMA.
2. It explicitly supports univariate time series data with a seasonal component.
3. ARIMA does not support seasonal data, SARIMA does.
4. Auto regression : Looks at data from past periods, p, when making predictions
5. Moving average : Future predictions are also made looking at the errors of the prior values, q.
6. Integration of past values and errors of prior values, d.

S

# Implementing ARIMA

- Checking for Stationary and Non Stationary Data




- Data showing cases of just individual months – stationary
- Showing data up to a specific month - non-stationary
- the value obtained through the Dicky Fuller test confirms this.

Value < 0.05 is stationary

```
print(adfuller(monthly_US['California'].dropna())[1])
```

```
0.0021476882056157493
```

```
print(adfuller(total_monthly_US['California'].dropna())[1]
```

```
0.931366959949502
```

# Implementing ARIMA

Goal: Find the best p, d, q order which gives the lowest RMSE.

1. Split data to train (70% of total data), test (30% of total data) datasets.

2. Give a range of value combinations for p, d, q and try different combinations using itertools .

```
p = np.arange(0,8)

d = np.arange(0,2)

q = np.arange(0,8)

Pdq_combos = list(itertools.product(p,d,q))
```

3. Try each p,d,q combination and save the RMSE associated with each combination to a file.

```
rmse = np.sqrt(mean_squared_error(test_data, predictions)
```

# Turning ARIMA Continues

4. Find the order with the lowest RMSE

   results['RMSE'].idxmin(), results['RMSE'].min()

5. Training the data using the order for p,d,q which gives the lowest RMSE.

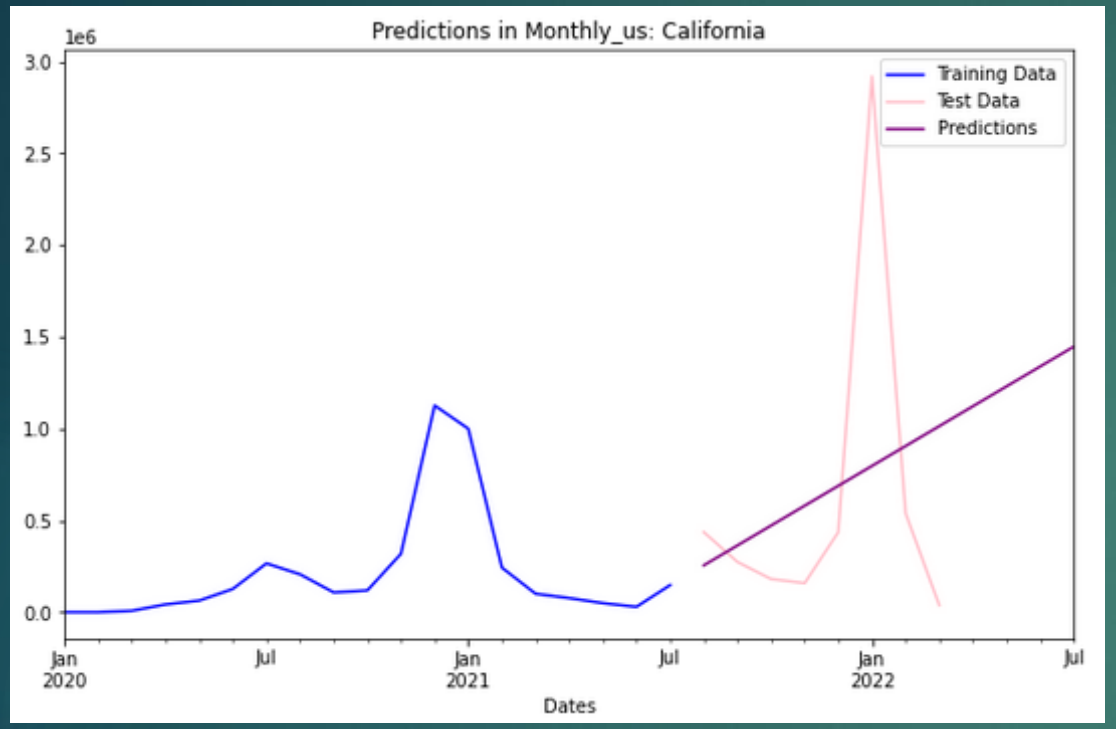
   model = ARIMA(train_data, order = p,d,q).fit()

6. Make predictions using this new model with the parameters

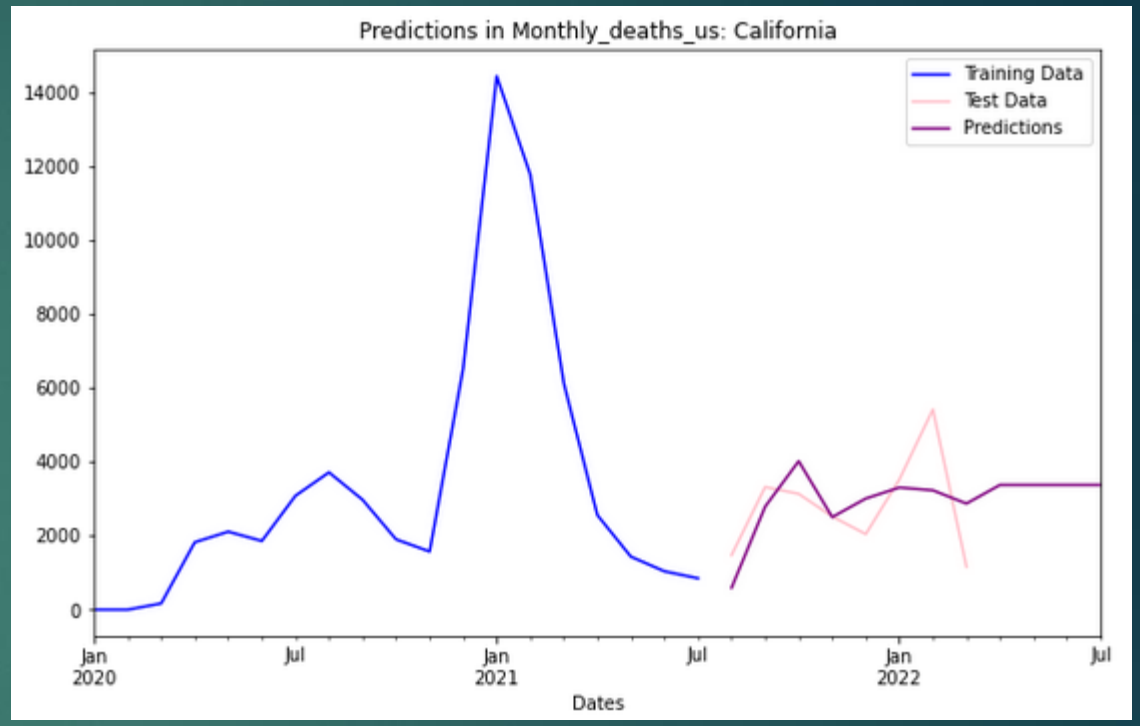
   predictions = model.predict(start = .., end = ..)

7. Visualize the training data, test data and predictions in a plot.

# Results of ARIMA Model



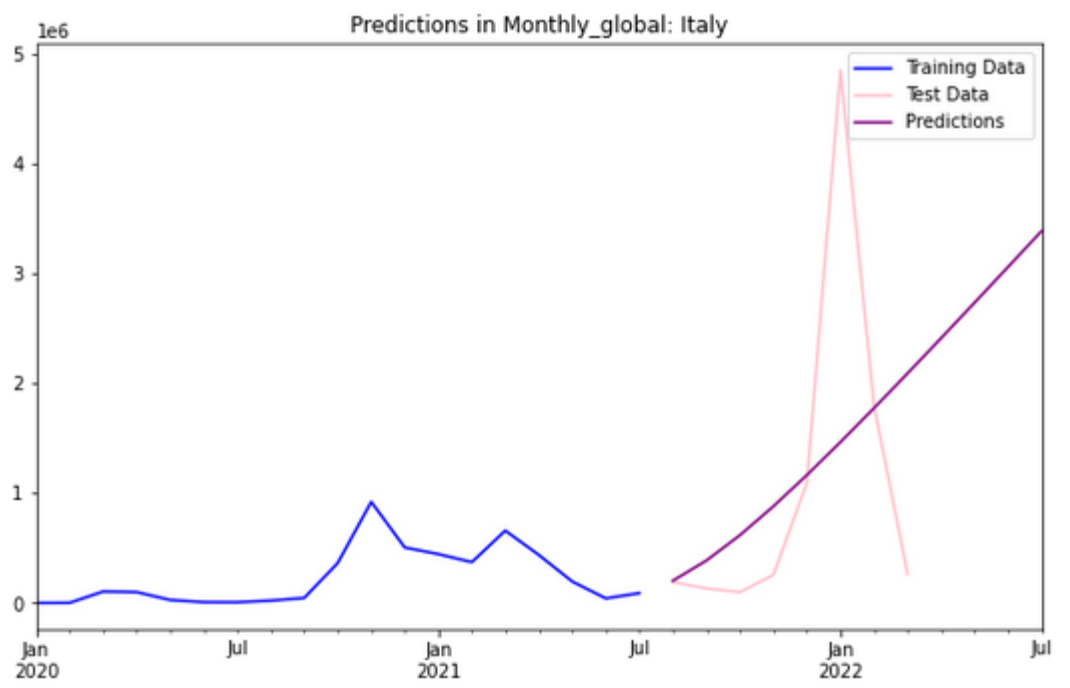

Order = (1, 2, 0)
The predictions are done on monthly confirmed cases in California.

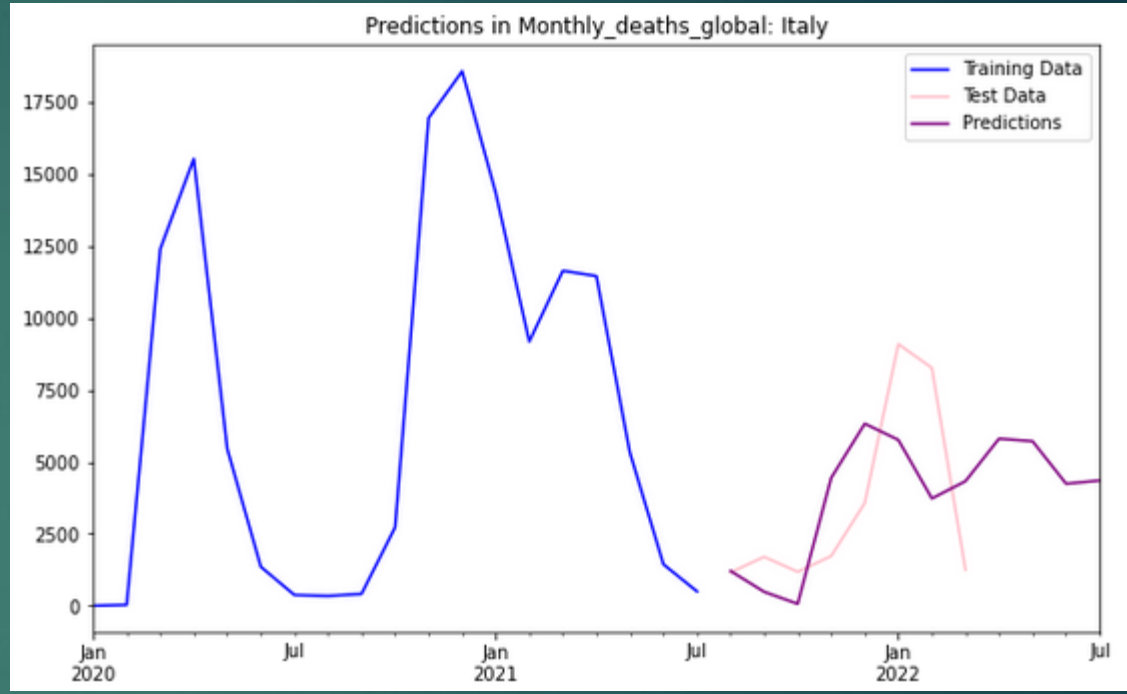



Order = (0, 0, 8)
Predictions are done on monthly deaths in California.

Predictions made on US deaths for the 3 months into the future is more accurate than the new cases for the next 3 months.

# Results of ARIMA Model




Order = (1, 2, 4)
The predictions are done on monthly confirmed cases in Italy.

Order = (3, 1, 7)
Predictions are done on monthly deaths in Italy.

Similar to the predictions made on US, global deaths for the 3 months into the future is more accurate than the new cases for the next 3 months.

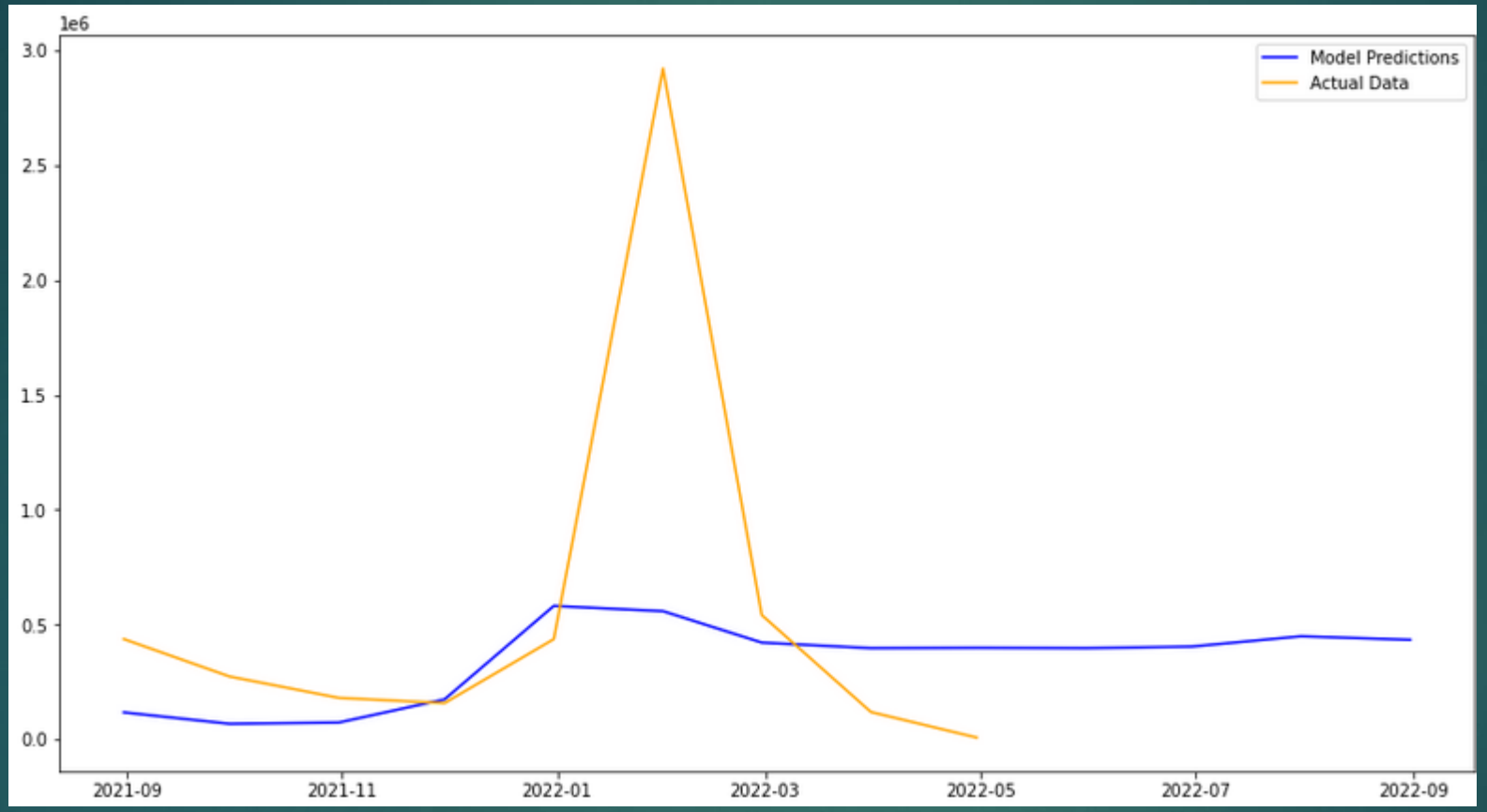
# Evaluating The Model

- The model was successful in forecasting the deaths data set better than the prevalence dataset.
- This could be due to the significant spike in new cases in the first quarter of 2022 which the model was not trained on.
- Extrapolation possibly caused the error in forecasting new cases

# SARIMA Model

1. Works with data on a time frame.
2. Auto regression : Looks at data from past periods, p, when making predictions
3. Moving average : Future predictions are also made looking at the errors of the prior values, q.
4. Integration of past values and errors of prior values, d.

# SARIMA Results




Similar to ARIMA model, forecasting using a model that has not seen the spike in data (extrapolation) could be the reason for results.

# Summary of The Project

- **Problem**
  - Looking at given clinical data, can we forecast the new cases of COVID-19 and related deaths?

- **Findings**
  - **New cases of COVID-19 infections and related deaths seem to spike at the end of each year, leading to January and February of the next year.**
    - **Could it be due to more testing, more gatherings or both?**
  - **The mortality rate related to COVID-19 of seems to be coming down in a consistent rate each year.**
    - **Is it because of new variances with lower mortality rate or are the healthcare providers have better understanding on treating patients, or both?**

- **US has the 3rd largest population in the world, less than quarter of the population of India or China. Yet, the US has the highest number of reported new cases and deaths.**
  - **Is it is a true representation or error in data gathering?**

- **Results**
  - **Both ARIMA and SARIMA were able to forecast deaths related to COVID-19 for both the US population and the world accurately.**
  - **However, the new instances were not forecasted accurately**
    - **There was a large spike in new cases in at the beginning of 2022. The model has not been trained on such data. Therefore, it could not predict the spike : Extrapolation**

# Areas for Improvement

- Finding a better way train data to avoid errors caused by extrapolation – Model was trained on one set of data which had a consistent pattern. Then it was tested on an inconsistent dataset which had an unexpectedly spiked dataset which would result in errors in making predictions. – Avoid extrapolation

- Conduct further research to understand why there is such a small ratio gap in total populations and new reported cases of COVID-19 in the US compared to India and China.

- Gather more information as to the percentage of tests administered in each country to understand the contradictory results in instances of COVID-19 cases.

- Gather more insight into what criteria were used when classifying deaths related to COVID-19.

  E.g.: If a patient which a history of chronic heart failure was hospitalized for coronary embolism and he/she gets routinely checked for COVID-19 gets positive results, and even though he/she dies from coronary embolism, could it be classified as a COVID-19 related death? If so, do all countries have the same guidelines for classifications under COVID-19?

# Thank you!