Data Analysis Report

Introduction

The COVID-19 pandemic has had a profound impact on the global population, with countries like the United States and India seeing significant case numbers and severe public health consequences. The aim of this report is to analyse the COVID-19 data for the United States and India, to understand key patterns in confirmed cases, recoveries, and deaths. The data utilized for this analysis was sourced from the Johns Hopkins University dataset, a reliable real-time source that provides global COVID-19 case updates. Through this analysis, we aim to provide insights that could aid in managing the pandemic and preparing for future waves.

Data Collection and Description

The dataset used for analysis comes from the Johns Hopkins University COVID-19 repository, which consists of multiple datasets that capture the evolution of the pandemic worldwide. The key datasets include:

- Confirmed Cases: The cumulative number of reported COVID-19 infections.
- Deaths: The total number of confirmed COVID-19 related deaths.
- Recoveries: The number of individuals who have recovered from the virus.

The dataset includes daily updates for each country, with geographical breakdowns down to the state level for the United States and India. The primary focus of this analysis was on daily case counts, deaths, and recoveries.

Data Preprocessing

Data preprocessing was a crucial step in preparing the dataset for analysis. This involved several key steps to ensure that the data was clean, consistent, and ready for modelling:

- **Handling Missing Data:** Incomplete reporting is common in COVID-19 data, especially during periods of inconsistency in data reporting. To address this, we used forward filling to impute missing values, assuming that the last known value would hold until the next available entry.

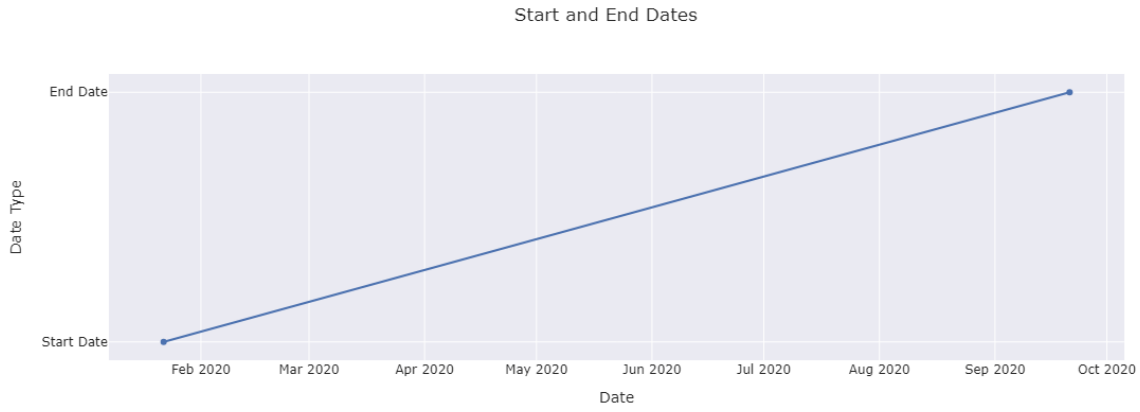
- **Data Transformation:** The data was initially non-stationary, meaning its statistical properties (such as mean and variance) changed over time. We applied:
- **Differencing:** Subtracted previous values from the current values to eliminate trends.
- **Log Transformation:** Applied a logarithmic transformation to smooth the data and make it more robust for modelling.

Exploratory Data Analysis

Exploratory Data Analysis was used to explore the trends in the dataset and identify key patterns:

- **Univariate Analysis:** We analysed the distribution of confirmed cases, deaths, and recoveries for both the United States and India. We found that the distribution was skewed, with frequent low numbers and occasional spikes.
- **Bivariate Analysis:** We analysed the correlation between confirmed cases and deaths, observing a strong positive correlation, as expected, since more cases generally lead to higher mortality.
- **Trend Analysis:** Time series plots revealed that both countries exhibited significant peaks in cases, which were often associated with government policy changes (e.g., lockdowns and restrictions) and the emergence of new virus variants.

Line Plot of Start and End Dates

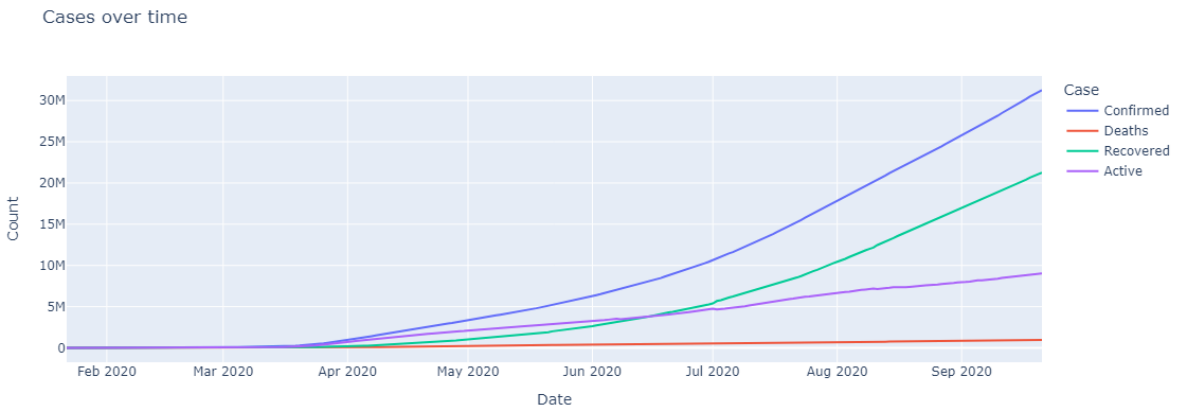


Insights

- The plot visualizes a start and end date over a timeline, with a straight line connecting them.
- It likely represents a time duration for a project, event, or process.

- The linear nature of the plot suggests a steady progression from the start date to the end date.

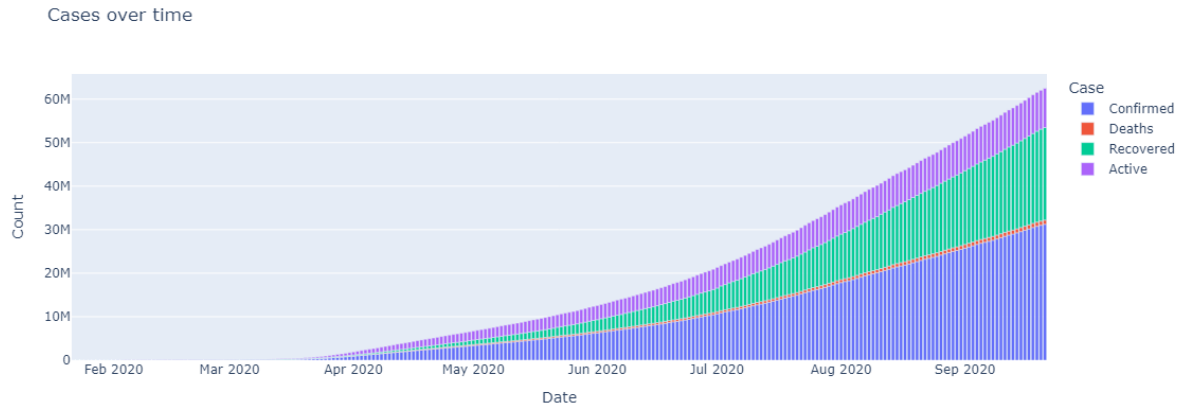
Line Plot of COVID-19 Cases Over Time



Insights

- The "Confirmed" cases show an upward trend over time, indicating the spread of the infection. If the "Recovered" cases also increase significantly, it suggests effective recovery efforts.
- If the "Active" cases start to decline while "Recovered" cases rise, it signals that the outbreak is being controlled. However, if "Deaths" continue to rise, it suggests a higher fatality rate despite recovery efforts.

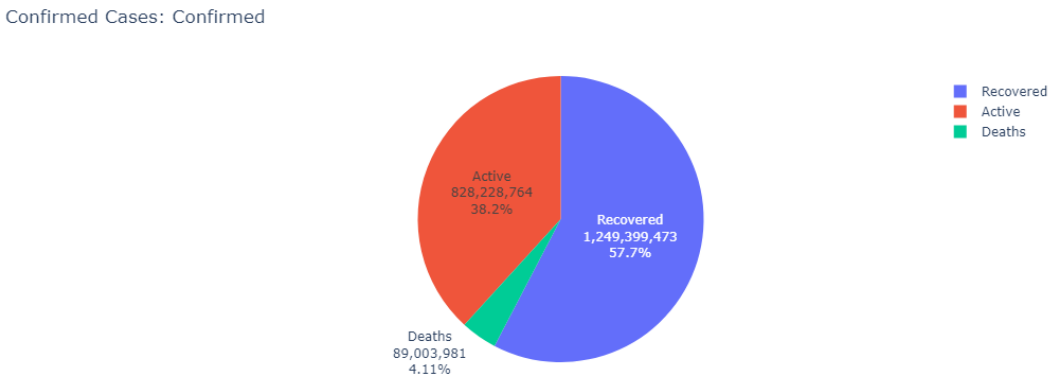
Bar Chart of COVID-19 Cases Over Time



Insights

- The bar chart clearly shows fluctuations in daily cases, making it easier to identify sudden spikes or declines in confirmed, active, recovered, and death counts.
- The use of colour for different case types helps in understanding which category (Confirmed, Deaths, Recovered, or Active) had the most significant changes over time.

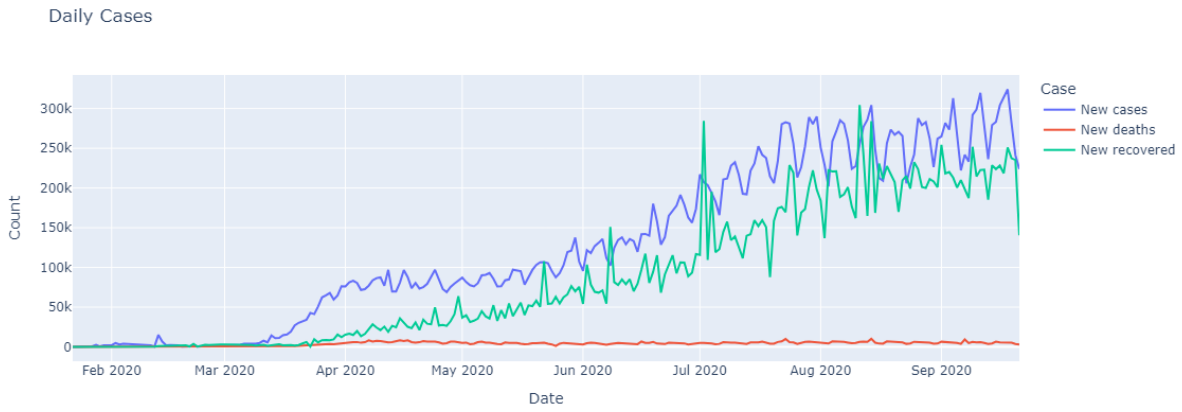
Pie chart showing the Distribution of COVID-19 Cases (Excluding Confirmed Cases)



Insights

- The high recovery rate (57.7%) is positive but still not overwhelming.
- The active cases (38.2%) suggest ongoing infections, requiring continued public health measures.
- The death rate (4.11%), though small, underscores the severity of the disease.

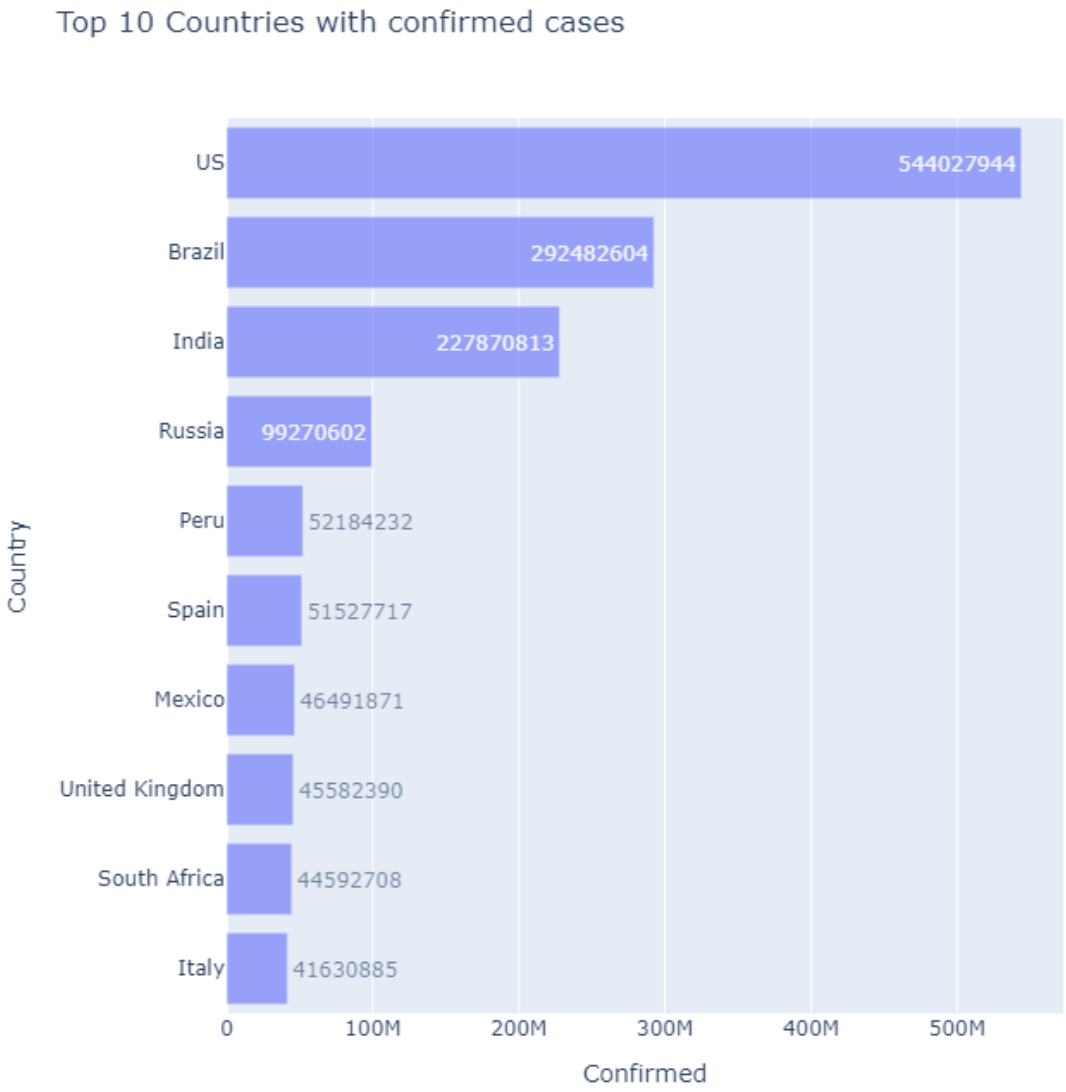
Line Plot of Daily COVID-19 Cases



Insights

- Likely shows an exponential rise initially, peaking at certain points.
- If cases decline later, it suggests containment efforts were effective.
- Typically lags behind new cases (since severe cases take time to result in death).
- A sharp rise in deaths with cases may indicate healthcare system overload.
- If recoveries increase over time, it suggests better treatment & immunity development.
- If recoveries remain low compared to cases, it may indicate longer illness duration or healthcare challenges.

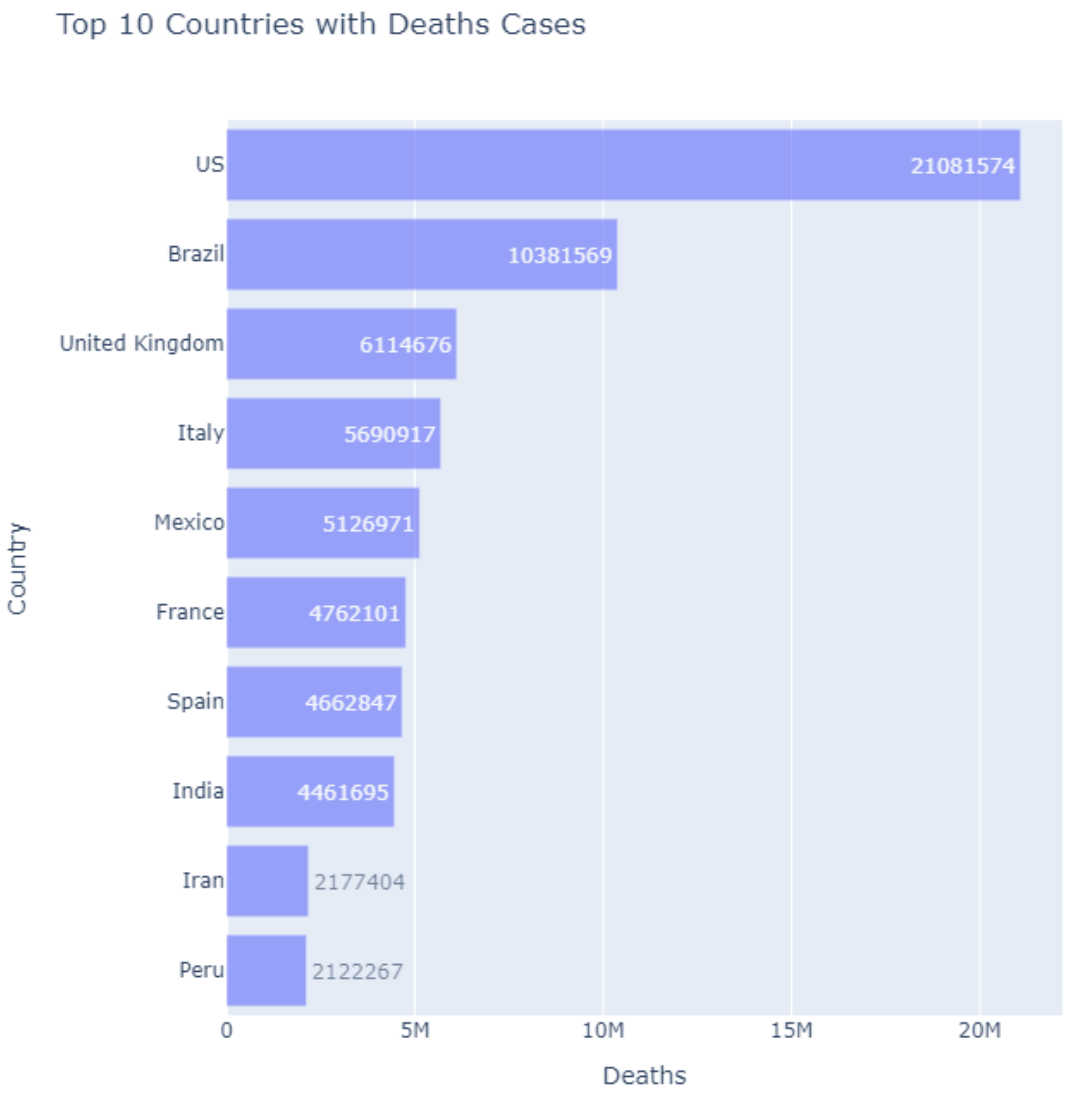
Horizontal Bar Chart of Top 10 Countries by Confirmed Cases



Insights

- US Leads Significantly: The United States has the highest number of confirmed cases (~544 million), which is much higher than any other country.
- Brazil & India Follow: Brazil (~292 million) and India (~228 million) have the second and third highest case counts, but both are significantly lower than the US.

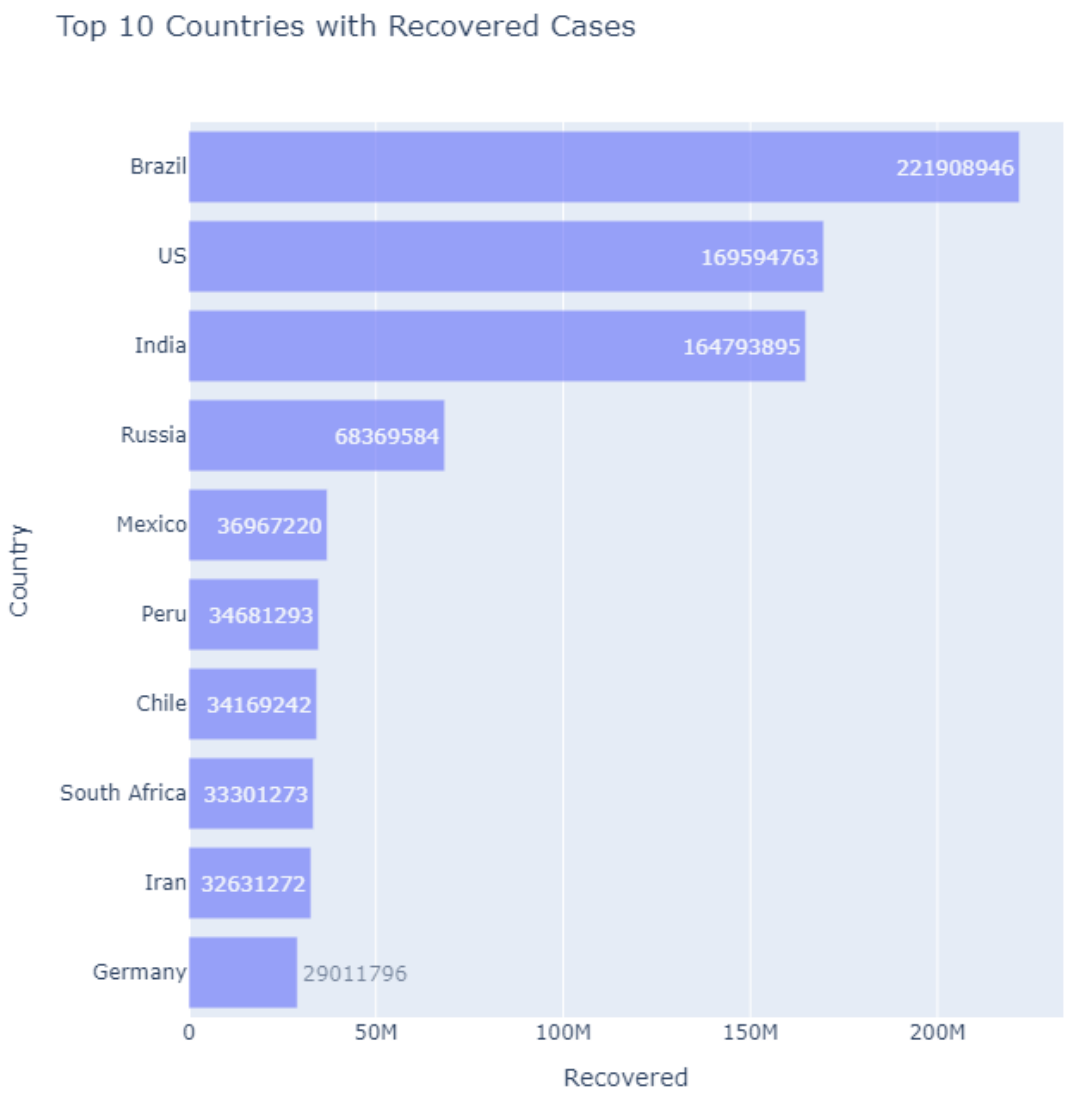
Horizontal Bar Chart of Top 10 Countries by COVID-19 Deaths



Insights

- US Has the Most Deaths: The United States leads with over 21 million deaths, which is significantly higher than other countries.
- High Case-to-Death Ratio in Some Countries: Notably, India had a very high number of confirmed cases (from the previous chart) but ranks lower in deaths, suggesting differences in mortality rates, healthcare responses, or data reporting methods.

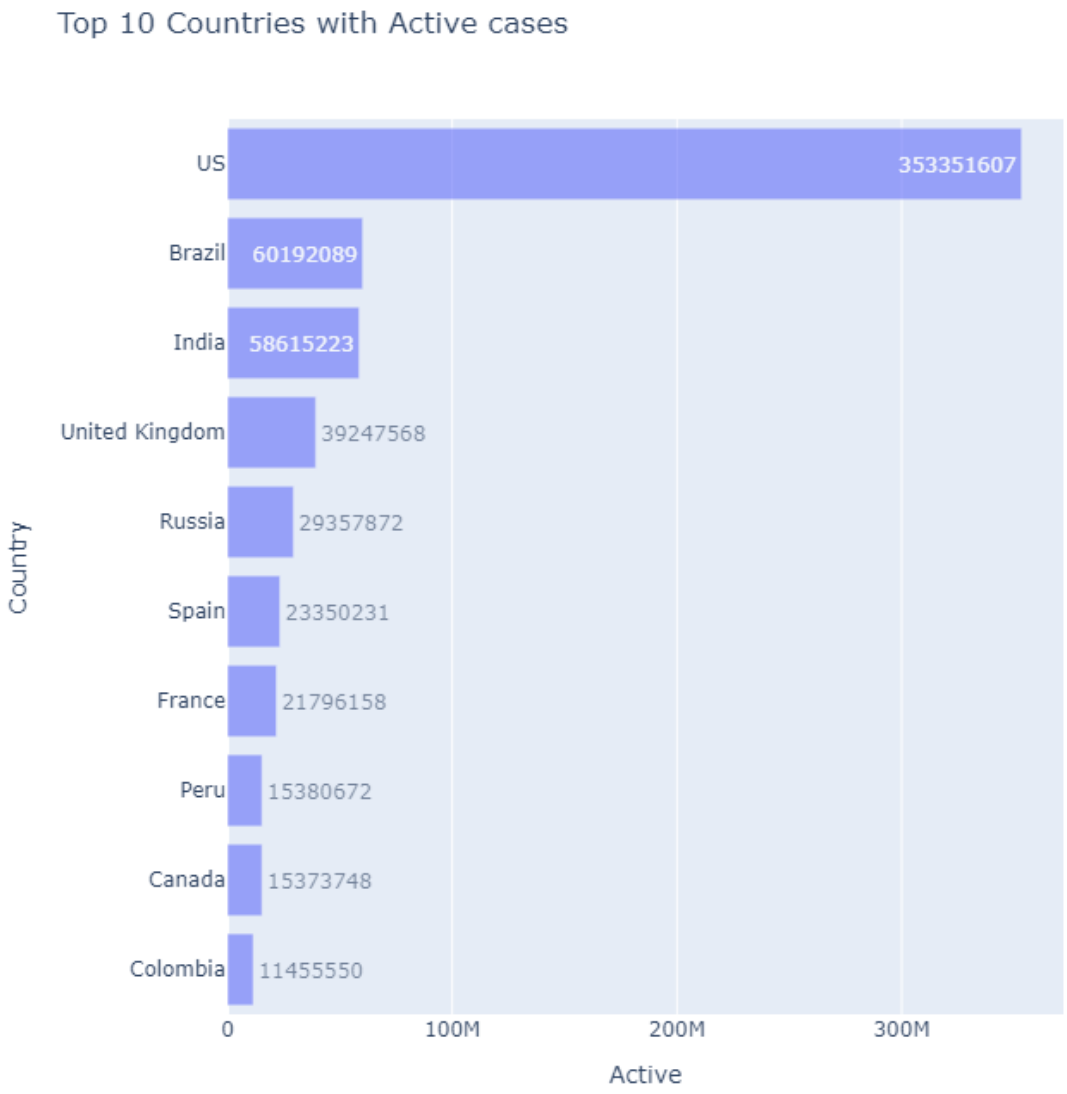
Horizontal Bar Chart of Top 10 Countries by COVID-19 Recovered cases



Insights

- The US had the highest confirmed cases, but its recovery count is slightly lower than Brazil's.
- India also shows a strong recovery rate, with ~165M recovered cases.

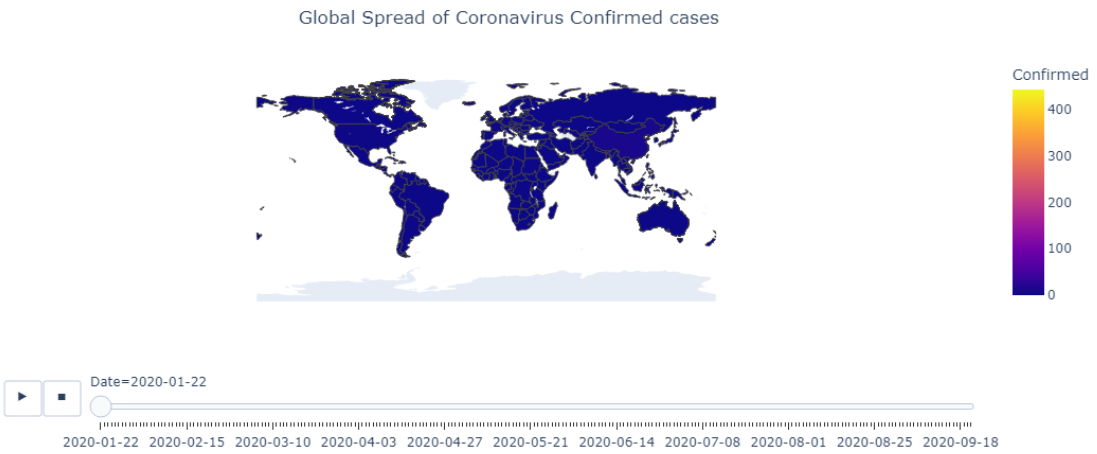
Top Countries with active cases



Insights

- The US has an overwhelming number of active cases, far exceeding other countries.
- Despite high recoveries, both Brazil and India still have a large number of active cases.

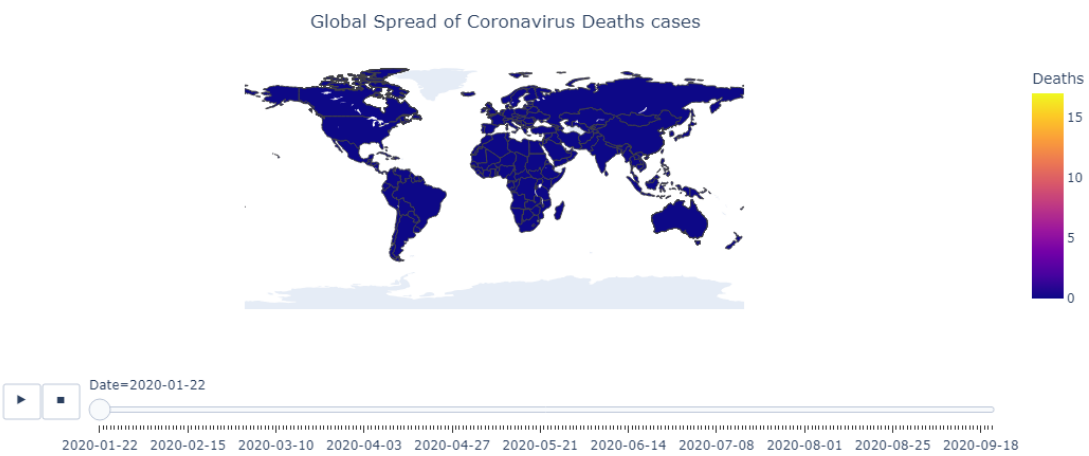
Choropleth Map of Global COVID-19 Confirmed Cases Over Time



Insights

- Above animated map is showing that global spread of confirmed COVID-19 cases over time.
- Each country is represented on the map, and the colour intensity varies corresponding to the number of confirmed cases.
- Based on the 'Date' column, we see the progression of confirmed cases over different dates.
- Higher the cases country is highlighted by yellow shade and lower the cases country is highlighted by blue colour
- From February to march there is gradually increase of confirmed cases in the Us and it became consistent.
- On march onwards there will be gradually increase in the confirmed cases in Brazil and India.
- The moment we play the start button it automatically shows each change till the last date.

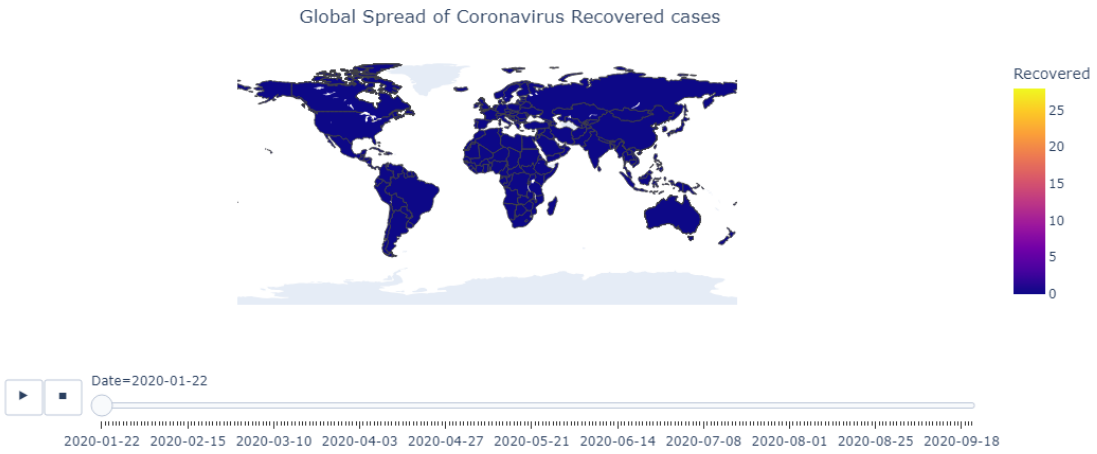
Choropleth Map of Global COVID-19 Deaths Over Time



Insights

- Above animated map is showing that global spread of COVID-19 death cases over time.
- After the February, there was a gradual increase in the death rate in the US, Italy, France, Spain, United Kingdom and in Brazil - till 2020-07-14 there was not much death rate
- From above map we conclude that the US is showing the maximum death rate compared to other country.

Choropleth Map of Global COVID-19 Recoveries Over Time

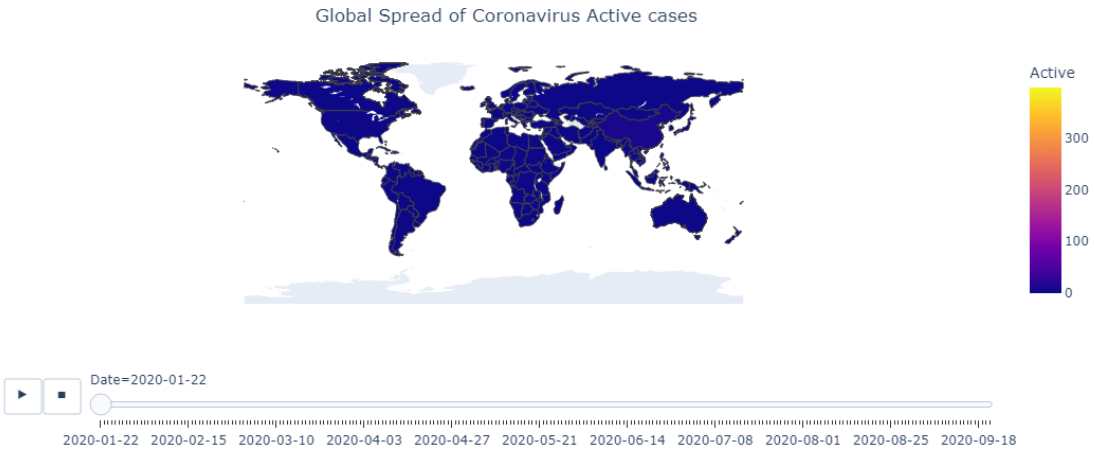


Insights

- Above animated map is showing that global spread of Recovered COVID-19 cases over time.

- Over the duration, there was a gradual increase in the recovery rate in the India.
- US is the 1st country showing more recovery rate up to last of June.

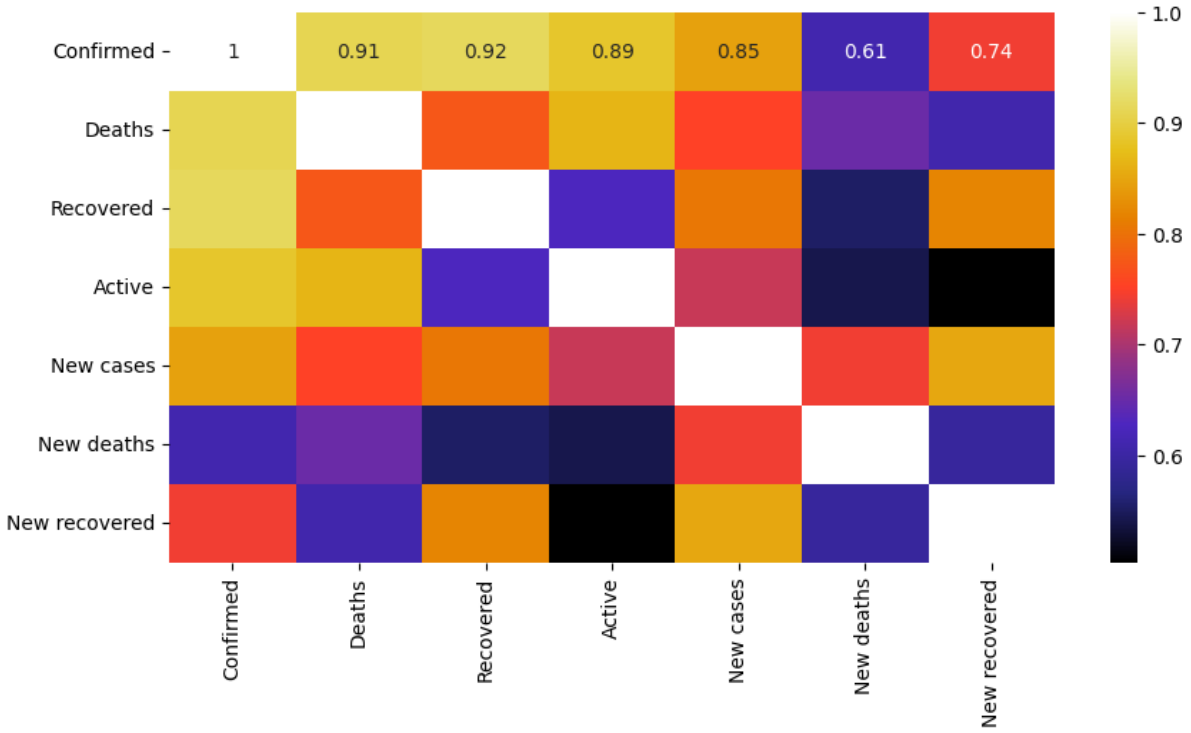
Choropleth Map of Global COVID-19 Active Cases Over Time



Insights

- Above animated map is showing that global spread of COVID-19 Active cases over time.
- At the beginning of the March, Italy is showing the most active cases after sometime there will be gradually decrease in the active cases.
- From 2020-03-10 to 2020-03-22 in US there is gradual increase in the active cases after it become consistent.

Heatmap Showing Correlation Between COVID-19 Metrics



Insights

- Confirmed Cases Strongly Correlate with Deaths & Recoveries
- Active Cases are Strongly Correlated with Confirmed Cases (0.89)
- New Cases Impact Recovery & Death Rates

Time Series Analysis

The time series analysis focused on identifying underlying trends and patterns in the data.

- **Stationarity Check:** We performed the Augmented Dickey-Fuller (ADF) test, which confirmed that the data was non-stationary. This required differencing the data to achieve stationarity before applying forecasting models.
- **Seasonal Decomposition:** We decomposed the time series data into three components: trend, seasonality, and residuals. This helped us understand how seasonality affected case numbers and identify patterns in the residuals (random noise).

Predictive Modelling

To predict future COVID-19 cases, we implemented several time series models:

1. ARIMA (Auto Regressive Integrated Moving Average):

ARIMA was used to model non-seasonal data, with differencing applied to make the data stationary. While ARIMA captured overall trends, it struggled with large surges and seasonal fluctuations. The model's performance showed that it could handle general trends but needed seasonal components to improve accuracy.

2. Auto Regressive (AR):

The Auto Regressive (AR) model, a fundamental time series model that utilizes the relationship between an observation and a number of lagged observations, was used separately. It focuses purely on past values and their direct impact on future values, without involving moving averages. The AR model was effective for short-term forecasting, especially when trends were not heavily seasonal. However, it also struggled with large fluctuations and was less effective than SARIMA in capturing seasonal patterns in COVID-19 cases.

3. SARIMA (Seasonal ARIMA):

SARIMA extended ARIMA by incorporating seasonality, making it more suitable for capturing the periodic spikes observed in COVID-19 cases, especially during waves of infections. The addition of seasonal components allowed the model to capture patterns that ARIMA could not, leading to more accurate predictions.

4. Holt-Winters Exponential Smoothing:

This model was applied to capture both the trend and seasonality in the data. Holt-Winters is particularly useful when there are clear, predictable seasonal patterns, such as the repeated surges in COVID-19 cases observed in both India and the United States.

5. Facebook Prophet:

Prophet is a modern forecasting tool that performs well with irregular data, missing values, and multiple seasonality’s. This model was especially effective at forecasting for both the US and India, as it can handle large datasets and external factors like holidays. It is known for being robust against outliers and irregular fluctuations, which is common in COVID-19 data.

Model Evaluation

The models were evaluated based on Mean Squared Error (MSE), a key metric to measure prediction accuracy:

For India, the SARIMAX model provided the most accurate forecasts with the lowest MSE. This was attributed to its ability to account for seasonal fluctuations in cases, capturing both trends and periodic surges.

For the United States, the Holt-Winters (Triple Exponential Smoothing) model emerged as the best performer. The model effectively captured both the trend and seasonality of the US COVID-19 data, leading to the lowest MSE for predictions.

Conclusion

This analysis provided valuable insights into COVID-19 trends in India and the United States. The SARIMAX model was the best-performing model for India, while Holt-Winters excelled in forecasting the US data. Both models demonstrated superior accuracy in predicting confirmed COVID-19 cases based on MSE, and their findings could help inform future pandemic management strategies.