# Analyzing Countries and USA COVID-19 data to investigate rise of COVID

Harsha, Joshi

New York University, hj2468@nyu.edu

Supreeta Anand, Byatnal

New York University, sab9920@nyu.edu

This study presents a comprehensive analysis of COVID-19 trends, leveraging three distinct datasets to elucidate the pandemic's progression both globally and within the United States. The first dataset, 'country_wise_latest' data encompasses a range of metrics including confirmed cases, deaths, recoveries, active cases, and new incidents, providing a snapshot of the pandemic's impact in various regions alongside WHO regional classifications. The second dataset, 'full_grouped' data,' offers a more granular view, detailing daily changes in COVID-19 statistics across different countries from January to July 2020. This temporal resolution allows for an in-depth examination of the pandemic's evolution on a global scale. Lastly, the 'USA_County_wise' data dataset narrows the focus to the United States, dissecting the pandemic's trajectory at the state and county levels with detailed geographical and temporal data[1] Through a series of data visualizations, this paper reveals global trends in COVID-19 spread and response, subsequently narrowing its focus to the U.S. context. The analysis not only highlights the dynamic nature of the pandemic but also underscores regional disparities and response effectiveness. This comprehensive approach aims to provide valuable insights for policymakers and health professionals in strategizing more effective responses to current and future public health crises.

**Additional Keywords and Phrases:** Data Visualization, Predictive Analytics, Linear Regression, Random Forest Algorithm

## 1 INTRODUCTION

The world has never seen anything like the COVID-19 pandemic, which was brought on by the new coronavirus SARS-CoV-2 and affected everyday life, economy, and public health. Since its discovery in late 2019, this pandemic has grown to rank among the most serious worldwide health emergencies in recent memory. It is more important than ever to comprehend how it spreads, how it affects healthcare systems, and how successful control methods are. Here's a more thorough account of COVID-19:

**SARS-CoV-2 Emergence:** Wuhan, China, reported cases of an unexplained respiratory disease in December 2019. Because of its genetic resemblance to the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), scientists were able to identify the causative agent as a novel coronavirus very quickly and gave it the name SARS-CoV-2. Global health entered a new era as a result of this[4].

**Never before Spread:** COVID-19 spread quickly throughout continents as a result of SARS-CoV-2's high contagiousness. It had reached pandemic proportions by early 2020, and in March 2020, the World Health Organization (WHO) formally declared it to be a global epidemic.

**Healthcare Overwhelm:** Hospitals struggled with a surge in COVID-19 cases, pushing healthcare systems around the world to breaking point. The spike in demand for hospital beds, ventilators, and personal protective equipment (PPE) underscores the vital need for resource allocation based on data.

**Public Health Measures:** To stop the virus's transmission, governments and health authorities put in place a number of public health measures, such as mask laws, lockdowns, and social distance. The goal of these actions was to "flatten the curve" and keep the healthcare systems from overburdening itself.

**Vaccination Efforts:** The creation and dissemination of COVID-19 vaccinations became a ray of hope as the pandemic deepened. Herd immunity was the goal of mass vaccination campaigns, which also attempted to lessen the virus's effects.

**International Cooperation:** Scientists, researchers, and the medical community came together on a never-before-seen scale in response to the pandemic.

In addition to its social value, this analysis provides personal satisfaction through improving the lives of others. It's a wonderful opportunity to make people's lives better and increase public awareness of health issues. This task is both professionally and intellectually exciting, since it offers doors to the expanding field of healthcare analytics and related fields. In the end, the project is a desirable topic for discussion since it promises both significant society influence and personal development.[3]

## 1.1 **Objectives and Goals**

COVID-19's Lasting Impact: Analyzing Early Data

Following our earlier discussion, we will delve into the initial stages of the COVID-19 pandemic, examining its significant impact on humanity and its ongoing influence on our lives. To gain deeper understanding, we set the following objectives:

- Understanding the Pandemic's Trajectory: Through analysis of confirmed cases and deaths, we aim to grasp the pandemic's intensity and evolution over time. These statistics offer raw insight into the situation's severity.
- Location-specific Insights: Analyzing data on a country-by-country basis will reveal which regions were most affected, which managed to control the spread, and what global trends emerged.
- Comparing with the USA: We will examine the USA's unique journey with the virus, comparing its statistics to global figures to assess its performance. Was the situation better, worse, or comparable?
- Data Visualization: Numbers are crucial, but visual aids enhance storytelling. We will utilize various chart types, including bar charts, line charts, maps, tree maps, and pie charts, to simplify data interpretation and identify patterns with greater ease.
- Predictive Analysis: By employing simple linear regression and random forest algorithms, we will attempt to predict future trends based on past data.

This comprehensive approach will provide valuable insights into the COVID-19 pandemic's early stages, its impact on different regions and the US, and potential future scenarios.

### 1.2 Dataset Description

- The COVID-19 Dataset is a comprehensive collection of data related to the COVID-19 pandemic. It provides detailed information on COVID-19 cases, deaths, recoveries, and other relevant statistics across various countries and regions.[2] This dataset is a valuable resource for researchers, data analysts, and anyone interested in studying the impact of the pandemic.

- The dataset aggregates data from multiple reliable sources, including government health agencies, international organizations, and reputable news outlets. It undergoes regular updates to ensure the latest information is available for analysis.

- Link to the datasets : https://www.kaggle.com/datasets/imdevskp/corona-virus-report/data

- **country_wise_latest.csv**:
This dataset provides the latest available country-level statistics for COVID-19 cases, deaths, recoveries, and other relevant information. It is designed to offer a snapshot of the pandemic's status for each country.
The dataset includes key metrics such as total cases, total deaths, total recoveries, and active cases for each country. It also provides per capita statistics and information on critical cases. The data set has 187 records

- **usa_county_wise.csv**:
This dataset focuses specifically on the United States and offers day-to-day COVID-19 data at the county level. It provides detailed information on cases, deaths, and recoveries for various U.S. counties.
The dataset includes data for individual counties within the United States, enabling users to analyze and track the pandemic's impact at a more localized level.
The data set has 627920 records.

- **full_grouped.csv:**
This dataset focuses on all the countries and offers day-to-day COVID-19 data at the country level. It provides detailed information on total cases, total deaths, new cases, new deaths, total recovered etc.
The dataset includes data for individual countries, enabling us to analyze and track the global effect of pandemic on daily basis from January 2020 to July 2020.
The data set has 35156 records.

## 2 METHODOLOGY

### 2.1 Data Pre-processing

- **Missing Values** - We analyzed all the three datasets and observed “full_grouped” and “country_wise_latest” don’t have any missing values but “use_county_wise” had two columns “FIPS” and “Admin2” which had a lot of null values. We decided to delete these two columns as they were not part of out analysis and each row represented data from a certain day and deleting rows with null values would have led to a lot of critical data loss.

## 2.2 **Data Visualization**

This section uses data visualization to discuss the COVID-19 scenario in the world and then discuss scenarios specific to United States

### *2.2.1. COVID-19 Across the World.*

- **Country Level Analysis:** Below (see figure 1) represents relationship between the countries and total death due to covid-19 parallel to deaths/100 cases. This visualization provides an understanding of how well countries were in terms of handling covid.

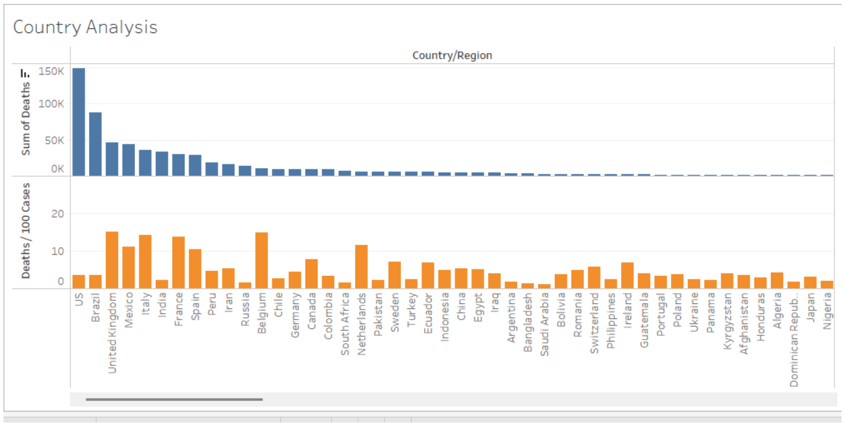

Figure 1: Bar graph of Countries vs Total Deaths and Death/100 cases

- **Country level recovery:** Below (see figure 2) represents the number of covid recoveries in each country. It provides us an idea about how well each country has handled covid.

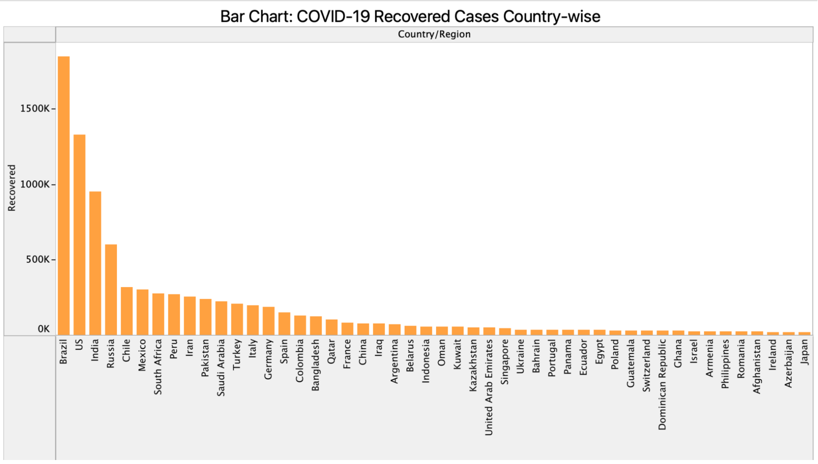

Figure 2: Comparision of each country to see where they stand with their recovery rate

- **Weekly new cases, recoveries and deaths:** This image (see below fig 3) is a table chart displaying weekly data on new COVID-19 cases, recoveries, and deaths in the United States, starting from January 19, 2020. This shows the rate of increase in the these attributes week by week in the US

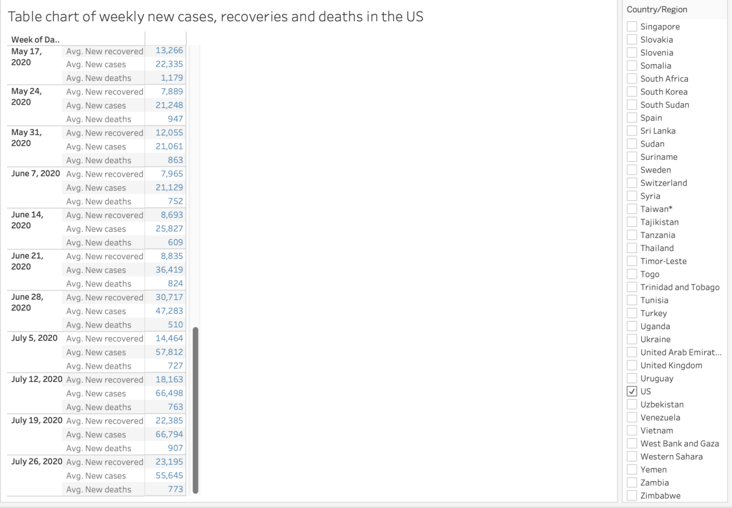

Table chart of weekly new cases, recoveries and deaths in the US

| Week of Da.. | | |
|---|---|---|
| May 17, 2020 | Avg. New recovered | 13,266 |
| | Avg. New cases | 22,335 |
| | Avg. New deaths | 1,179 |
| May 24, 2020 | Avg. New recovered | 7,889 |
| | Avg. New cases | 21,248 |
| | Avg. New deaths | 947 |
| May 31, 2020 | Avg. New recovered | 12,055 |
| | Avg. New cases | 21,061 |
| | Avg. New deaths | 863 |
| June 7, 2020 | Avg. New recovered | 7,965 |
| | Avg. New cases | 21,129 |
| | Avg. New deaths | 752 |
| June 14, 2020 | Avg. New recovered | 8,693 |
| | Avg. New cases | 25,827 |
| | Avg. New deaths | 609 |
| June 21, 2020 | Avg. New recovered | 8,835 |
| | Avg. New cases | 36,419 |
| | Avg. New deaths | 824 |
| June 28, 2020 | Avg. New recovered | 30,717 |
| | Avg. New cases | 47,283 |
| | Avg. New deaths | 510 |
| July 5, 2020 | Avg. New recovered | 14,464 |
| | Avg. New cases | 57,812 |
| | Avg. New deaths | 727 |
| July 12, 2020 | Avg. New recovered | 18,163 |
| | Avg. New cases | 66,498 |
| | Avg. New deaths | 763 |
| July 19, 2020 | Avg. New recovered | 22,385 |
| | Avg. New cases | 66,794 |
| | Avg. New deaths | 907 |
| July 26, 2020 | Avg. New recovered | 23,195 |
| | Avg. New cases | 55,645 |
| | Avg. New deaths | 773 |

Country/Region

□ Singapore
□ Slovakia
□ Slovenia
□ Somalia
□ South Africa
□ South Korea
□ South Sudan
□ Spain
□ Sri Lanka
□ Sudan
□ Suriname
□ Sweden
□ Switzerland
□ Syria
□ Taiwan*
□ Tajikistan
□ Tanzania
□ Thailand
□ Timor-Leste
□ Togo
□ Trinidad and Tobago
□ Tunisia
□ Turkey
□ Uganda
□ Ukraine
□ United Arab Emirat...
□ United Kingdom
□ Uruguay
☑ US
□ Uzbekistan
□ Venezuela
□ Vietnam
□ West Bank and Gaza
□ Western Sahara
□ Yemen
□ Zambia
□ Zimbabwe

Figure 3: Table Chart of weekly new cases, recoveries and deaths in the US

*2.2.2. COVID-19 USA Stats.*

- **USA Map based on total deaths:** Below heat map (see figure 4) representing different states of USA and color spectrum signifying the total no. of deaths which happened in each state due to covid.

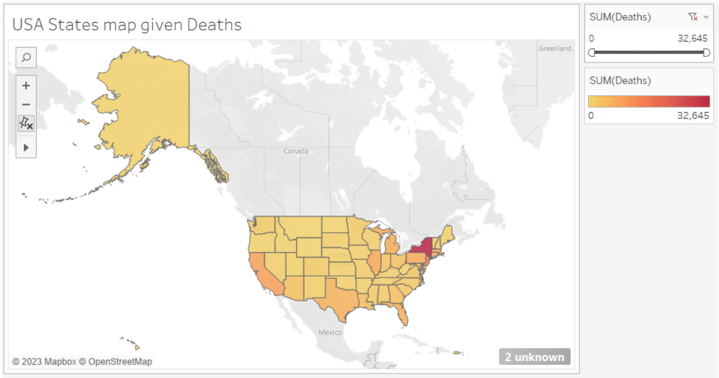

Figure 4: Heat Map of USA based on total deaths

- **Top 5 states of USA based on deaths:** Below map (see figure 5) shows five states of USA. This map signifies the top 5 states in terms of deaths dur to covid-19

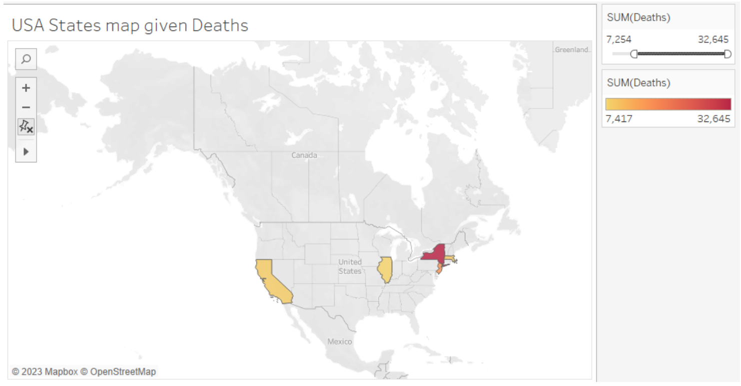

Figure 5: Top five USA states based on total deaths

- **USA Stats from Jan to July :** A line chart drawing parallels of how the confirmed cases and deaths are progressing in USA starting January to May (see figure 6).

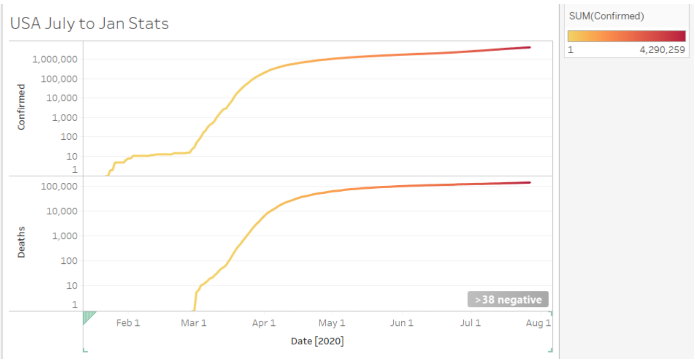

Figure 6: Statistics showcasing Confirmed Cases and Deaths from January to July

- **Breakdown of global Confirmed cases:** This pie chart visually represents which WHO region has how many number of confirmed COVID cases. Tooltip also shows the number of deaths and number of people who have recovered in that region ( see figure 7).

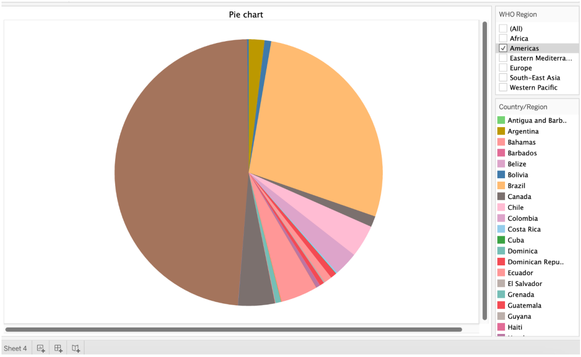

Figure 7: Pie chart

Creating data visualizations for both USA and global COVID-19 data is invaluable in understanding and managing the pandemic. By visually representing the data, it becomes easier to grasp complex trends and patterns, such as the spread of the virus, effectiveness of public health measures, and regions most affected. These visualizations serve as crucial tools for policymakers and health officials, aiding in making informed decisions about lockdowns, resource allocation, and vaccine distribution. For the general public, these visual representations demystify the data, fostering a greater understanding and awareness of the pandemic's status both locally and worldwide. Moreover, in an educational context, they provide a clear and concise way to communicate the impacts and progression of COVID-19, making it an essential component in public health communication and education strategies.

## 2.3 **Predictive Analytics**

In this segment, we describe the methods and strategies employed for predictive analysis, aimed at forecasting scenarios related to COVID-19. We employ two popular machine learning algorithms: Linear Regression and Random Forest

2.3.1. Covid-19 Confirmed Cases Prediction on Countries level:

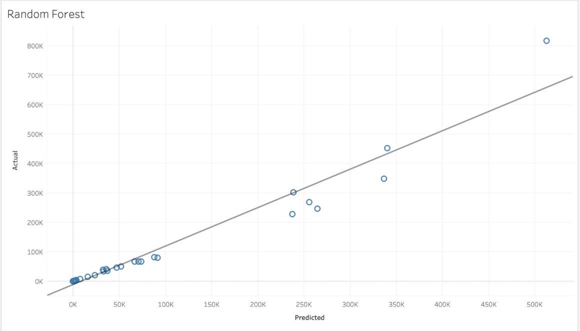

Figure 8 : Random Forest to predict the number of confirmed cases

Model Selection:

- Random Forest: Random Forest is a versatile algorithm that can be applied to various types of tasks, including both regression and classification. This flexibility allows us to adapt the algorithm to different problem domains and research questions.

Model Evaluation:

- Data: The dataset is loaded from the specified file path using the Pandas library. The 'Country/Region' and 'WHO Region' columns are dropped as they were not relevant for our regression task. We handled potential issues with the data by replacing infinities with NaN values. The missing values in the 'Deaths / 100 Recovered' column are imputed with the median value.

- Split: The data is divided into training and testing sets, allocating 80% of the data for training and reserving 20% for testing. This splitting strategy helps evaluate the model's performance on unseen data.

- Training: Random Forest Regressor is trained with 100 estimators on the training data.

- Metrics: To assess the model's predictive performance, we utilized the following regression evaluation metrics:

    - **Root Mean Square Error (RMSE)**: RMSE is calculated, which measures the average deviation of the model's predictions from the actual values. A lower RMSE indicates better predictive accuracy.

    - **R-squared (R^2) Score**: The R^2 score is computed, which quantifies the proportion of variance in the target variable explained by the model. A higher R^2 score signifies a better fit to the data.

The scatter plot visualizes the relationship between the predicted and actual values from a statistical model. Data points represent individual observations, and the line of best fit drawn through these points reflects the model's predictions. A positive slope of the line indicates a direct relationship where increases in the independent variable (predicted values) correspond to increases in the dependent variable (actual values). This graph is an essential tool for evaluating the accuracy of the model's predictions and can be instrumental for informed decision-making in fields such as public health, where it can predict outcomes like new cases or deaths based on current trends. Such insights are vital for planning and resource allocation in healthcare systems.

2.3.2. Predict new deaths for USA region :

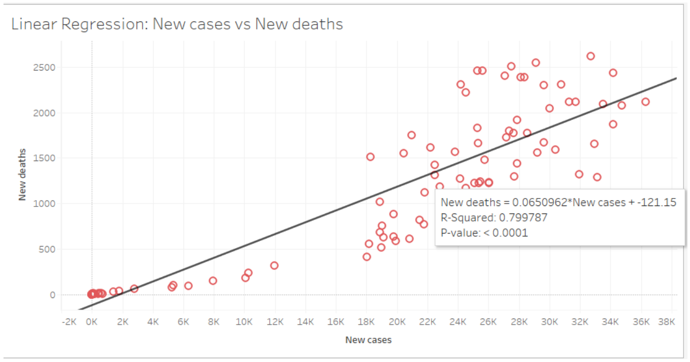

Figure 9: Linear Regression of New Deaths vs New Cases

Model Selection:

- Linear Regression: Use linear regression to predict the new no. of deaths which can happen in the US based on the new cases.

Model Evaluation:

- Evaluation Metrics: Assess the trend line and make observations on the slop, linear equation line, p-value and R-squared value

The line of best fit drawn through the data points represents the predicted average number of new deaths given any number of new cases. The slope of the line indicates the rate at which new deaths increase for an increase in new cases. If the line has a positive slope, as it appears in this graph, it suggests that higher numbers of new cases are associated with higher numbers of new deaths. This type of analysis is crucial for health planning and resource allocation, as it can help anticipate the demand for healthcare services based on the incidence of new cases..

# 3 DISCUSSION & RESULTS

## 3.1. Data Visualizations

### *3.1.1. COVID-19 Across the World*

- **Country Level Analysis:** The chart allows for a comparison between the absolute number of deaths and the death rate relative to the number of cases, which can provide insights into the severity of the pandemic's impact and the effectiveness of each country's response. For instance, a high death rate might indicate overwhelmed healthcare systems, higher population vulnerability, or possible issues with case reporting and data accuracy.

  Here we observe that while US has much high number of total cases which also makes sense purely due to much larger population, the deaths per 100 cases in US is much lower compared to several other developed countries. We are comparing US with developed countries to make sure we compare only the actually comparable healthcare systems to eliminate differences due to possible lack of good healthcare system. European Nations in particular have a much higher deaths per 100 cases, which can be thought of as a consequence of the phase differences in the spread of pandemic. Europe was hit in the earlier stages of Coronavirus pandemic, a time when very few possible support treatments were known. However, the lessons learnt there and preparedness could have helped to maintain low death rate in US. Also, these differences can be due to several other differences in quality and protocols of healthcare system, public policies, social awareness, etc. However, what we definitely do observe is a contrasting difference

- **Country level recovery:** In this visualization, we can clearly see that the majority of other countries have a relatively smaller number of recoveries, suggesting either fewer COVID-19 cases or differences in healthcare systems, reporting practices, or population sizes

  There is one country with a notably higher number of recoveries, which could indicate a larger number of reported cases or more effective reporting mechanisms. Different countries have different methodologies for reporting cases and recoveries. Brazil may have a broader definition of recovery, or a more aggressive approach to testing and reporting recoveries. The virus may have spread earlier in Brazil or peaked at different times compared to the US, affecting the number of recoveries reported at any given time.

- **Weekly new cases, recoveries and deaths :** This visulatization helps us see the progress of covid in the US, understand which week did it peak. There is a clear upward trend in the average number of new cases reported weekly, starting from zero cases in the week of February 23, 2020, to 25,471 new cases in the week of May 3, 2020. This indicates that the virus was spreading rapidly during these initial months.

The number of recoveries also shows an increasing trend, which is expected as the number of reported cases rises. Notably, the average number of new recoveries has a significant increase in the week of April 26, 2020, jumping to 10,716 from the previous week's average of 5,076.
The average weekly deaths due to COVID-19 show fluctuations but generally increase over time, peaking at 2,238 in the week of April 19, 2020. There is a slight decrease in the average number of deaths in the following weeks.
The data could be indicative of the spread of COVID-19 during the early phase of the pandemic in the United States, reflecting the initial impact on the health care system and the response to the virus.

*3.1.2.COVID-19 USA Stats.*

- **USA Map based on total deaths:** The color scale on the right indicates the range of the number of deaths, with lighter shades representing lower numbers of deaths and darker shades representing higher numbers. The darkest shade, the state of New York, indicates the highest number of deaths in this representation, while states with lighter shades have fewer deaths.

- **Top 5 states of USA based on deaths:** The map and the filter shows the states of NY, NJ, MA, IL and CA are the worst COVID hit states when it comes to the total number of deaths. This can be explained due to the high population densities in these areas, the presence of large urban work centres in these states, international travel and also early outbreaks in these states.

- **USA Stats from Jan to July :** It can be observed from this visualization that as the no. confirmed cases by end of July is still increasing drastically but on the contrary the no. of death have seemed to have a smaller growth slope. This can be because maybe the country started taking better medical precautions and found better way to conbat the virus.

  Also, the gap that we see in the chart corresponding to the deaths in the starting range can be because of the phase difference. New cases started to come up in January but it must have taken 2-3 weeks for doctors to treat them, get better or succumb to the disease.

- **Breakdown of global confirmed covid cases :** In this visualization we see that the pie chart is divided based on the number of confirmed cases. We can see that even wrt all WHO regions, US has the maximum number of covid cases. This could be because of multiple reasons like US having a good testing procedure in place, more international interaction, not having stricter rules in place at the time, etc.

**3.2. Predictive Analytics**

3.2.1 Covid-19 Confirmed Cases Prediction on Countries level:

This section is a discussion on the predictive analysis we did on the no. of confirmed cases employing random forest regressor

- **Random Forest :**

  - The model is quite effective in predicting the confirmed COVID-19 cases, as indicated by the high R^2 score. However, there is still an average error as indicated by the RMSE, which may be significant depending on the context (for instance, in smaller countries or regions, an average error of 53,825 might be very high).

  - RMSE: 53825.84653347116
    R^2 Score: 0.8913851893663205

  - Given the high R^2 score, the model's predictions can be considered reliable for the dataset it was trained on. It may be useful for forecasting confirmed cases or understanding which features are most predictive of COVID-19 spread.
  - The scatter plot and the linear relationship it shows between actual and predicted values suggests that the model has a good fit for the data it was tested on.
  - The scatter plot indicates that most predictions are close to the actual numbers, as the points generally cluster around the line of perfect prediction. This visual confirmation of the model's high R^2 value suggests that the model is generally making predictions that are in the right direction.
  - The results, particularly the RMSE, should be considered in the context of the range and distribution of confirmed cases in the dataset. For large countries with millions of cases, an RMSE of 53,825 might be relatively small, while for countries with fewer cases, this would be a substantial error.
  - Outliers or anomalies could be affecting the RMSE. An examination of such points could offer insights into the data or suggest areas where the model might be improved.

3.2.2 Predict new deaths for USA region :

This section is a discussion on the predictive analysis we did on the no. of new deaths based on the number of new cases, employing Linear Regression

- **Linear Regression:** The Trend Line (Black Line) through the data points is the linear regression trend line, which represents the average relationship between 'New cases' and 'New deaths'. The equation for this line is given by New deaths = 0.0650962*New cases - 121.15.

  Regression Equation:

  - Slope (0.0650962): For every additional new case, there is an average increase of approximately 0.0650962 new deaths. This is the slope of the line and indicates a positive relationship between 'New cases' and 'New deaths'.

  - Y-intercept (-121.15): Theoretically, when there are zero new cases, there would be -121.15 new deaths. Since negative deaths are not possible, this may suggest that the model isn't perfectly calibrated at lower values or that there are other factors influencing the number of new deaths.

  - R-Squared (0.799787): This value indicates that approximately 79.98% of the variability in 'New deaths' can be explained by the number of 'New cases'. This is a measure of the model's fit; a value closer to 1 indicates a better fit.

  - P-value (< 0.0001): This indicates that the relationship between 'New cases' and 'New deaths' is statistically significant, with a p-value much less than 0.05, which is a common threshold for significance.

  Overall, the graph suggests that there is a significant positive relationship between 'New cases' and 'New deaths', as shown by the positive slope of the trend line and the strong R-squared value. However, the negative y-intercept might indicate some issues with the model at lower case counts or could be a result of data recording practices.

  The graph can be used to understand the severity of the outbreak over time and to make informed decisions about public health interventions. The relationship between cases and deaths can also provide insight into factors such as the mortality rate, healthcare system capacity, and effectiveness of treatment protocols during the pandemic.

## 4 CONCLUSION

The comprehensive visualizations and data analysis presented in our research paper offer a profound understanding of the COVID-19 pandemic's impact. The heat map of the United States vividly illustrates the varying severity of the pandemic across states, emphasizing regional disparities likely influenced by factors such as population density and healthcare infrastructure. Map which ighlights the top 5 states with the highest COVID-19 death tolls, directs attention to areas that have borne the brunt of the crisis, prompting further investigation into the causes. The temporal analysis of tracking cases and deaths from January to July provides a crucial timeline of the pandemic's evolution, potentially revealing connections to policy changes and vaccination campaigns. On the global scale, bar graph comparing countries based on total deaths and deaths per 100 cases offers insights into international responses to the pandemic, enabling the identification of countries that effectively managed the virus. Along with this, out predictive analysis using Linear regression and Random Forest gives a insight over how the future trends of covid-19 after July'2020 might look and it shows we need to be careful as there is a good chance covid getting worse. Overall, these visualizations emphasize the importance of recognizing regional disparities, the necessity for continuous monitoring, and the potential lessons to be gleaned from nations that successfully navigated the pandemic, contributing to our knowledge of pandemic management for the future.[5]

## 6 REFERENCES

1. Worldometer - COVID-19 Statistics : Worldometer. (2020). Covid-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus
2. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. (2021). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. GitHub. https://github.com/CSSEGISandData/COVID-19
3. The First Three Months of COVID-19: Epidemiological Evidence for Two SARS-CoV-2 Strains Spreading and Implications for Prevention Strategies (The American Journal of Tropical Medicine and Hygiene, 2020) by Paolo Emilio Puddu et al. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9572891/
4. M.K. Sharmal, Analysis of the COVID-19 pandemic: lessons towards a more effective response to public health emergencies (Globalization and Health, 2022) https://pubmed.ncbi.nlm.nih.gov/33654660/
5. S. Khan, Investigation of COVID-19 and scientific analysis big data analytics with the help of machine learning (National Institutes of Health, 2020) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9069062/