COVID-19 GLOBAL SPREAD ANALYSIS REPORT

Executive Summary

A novel coronavirus designated 2019-nCoV was first identified in Wuhan, China, which marked the beginning of a global pandemic that ravaged the entire world. Initially characterized by pneumonia cases with no clear existing treatment or vaccines, the virus quickly showed evidence of high human-to-human transmission, which escalated rapidly in transmission rates by mid-January 2020.

Key Findings:

- **Total countries affected: 187**
- **Pandemic duration: 7 months (January – July 2020)**
- **Highest WHO Region impacted: Americas (54% of confirmed cases)**
- **Global Recovery Rate: 29%**
- **Global Active Cases: 19%**
- **Global Death Cases recorded: 2%**
- **Estimated Total Number of Infections:**
- **Highest Transmission rate recorded: 50% at January 28th 2020**
- **Most severely affected country: United States of America**

Introduction

In late 2019, the world was confronted with an unprecedented health challenge with the emergence of a new coronavirus. This strain of the virus was first identified in Wuhan, the capital of China's Hubei province. The virus quickly spread at an alarming rate across countries, challenging global health infrastructures and fundamentally altering the way of life of the entire world.

Objectives of the Analysis

The primary objectives of this analysis were to:

a) Identify trends and patterns in the covid-19 virus spread.
b) Uncover critical factors influencing virus transmission.
c) Provide actionable insights for mitigation strategies.
d) Develop predictive models to understand potential future scenarios.

Data Sources and Methodology

a) **Primary Source:** The datasets used in this analysis was obtained from www.kaggle.com
b) **Geographical Coverage:** 188 countries
c) **Time period:** January 22nd, 2020 – July 27th, 2020 (7 months)

Data Preprocessing Techniques

The research team implemented rigorous data preparation methods to ensure analysis accuracy and data validity. The following was implemented:

a) Address missing values
b) Standardize date field formats for time series analysis
c) Calculate derived metrics:
    - **Estimated Total Infection:** the proportion of the population that has been confirmed infected. This helped us to assess how widespread the virus was within a community and inform public health decisions.

    - **Transmission Rate (%)**: refers to how many people on average one infected person will transmit the virus to.

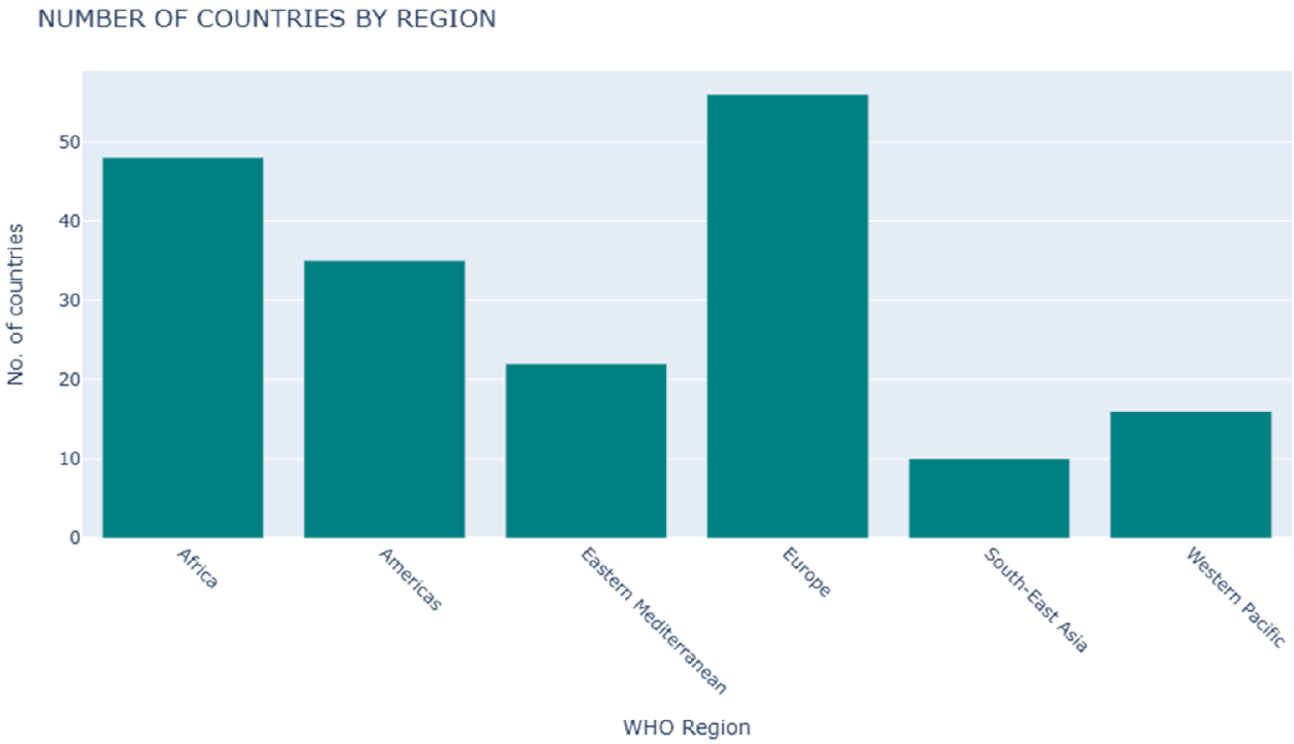
d) Other columns were renamed:
    - **Death / 100 Cases -> Case Fertility Rate (%)**:
    - **Recovered / 100 Cases -> Case Recovery Rate (%)**

Dataset Composition

a) **Country-wise dataset**
    - Daily and weekly case numbers per country
    - Key variables including confirmed, death, and recovered cases.
    - Regional classification
b) **Day-wise dataset**
    - Global daily case recordings
    - Comprehensive time-series data
    - Aggregated global statistics such as confirmed, death cases, etc.

Exploratory Data Analysis (EDA)

A. Country Demographics and Affected Regions

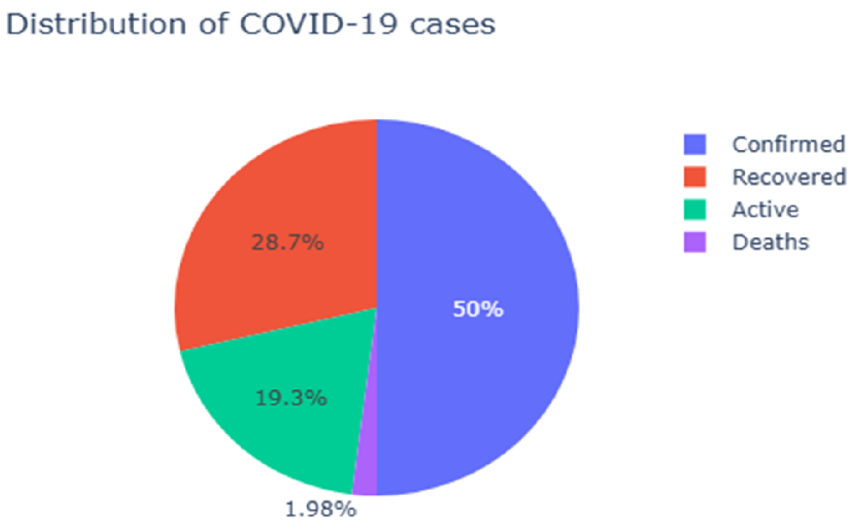


Insights:

The analysis revealed that 6 regions were affected by the Covid-19 virus. Europe emerged as the most extensively affected region, with 56 countries reporting covid-19 cases. This widespread impact can be attributed to some critical factors such as:

a) High population density in European countries
b) Extensive international travel which probably introduced the virus through multiple entry points e.g. flight, sea travel etc.
c) Advanced testing capabilities which led to more cases being reported.

B. Total Case Breakdown



Key Insights and Statistics:

a) **Total Recoveries: 28% of total cases**
b) **Confirmed Cases: 50% of total cases recorded.**

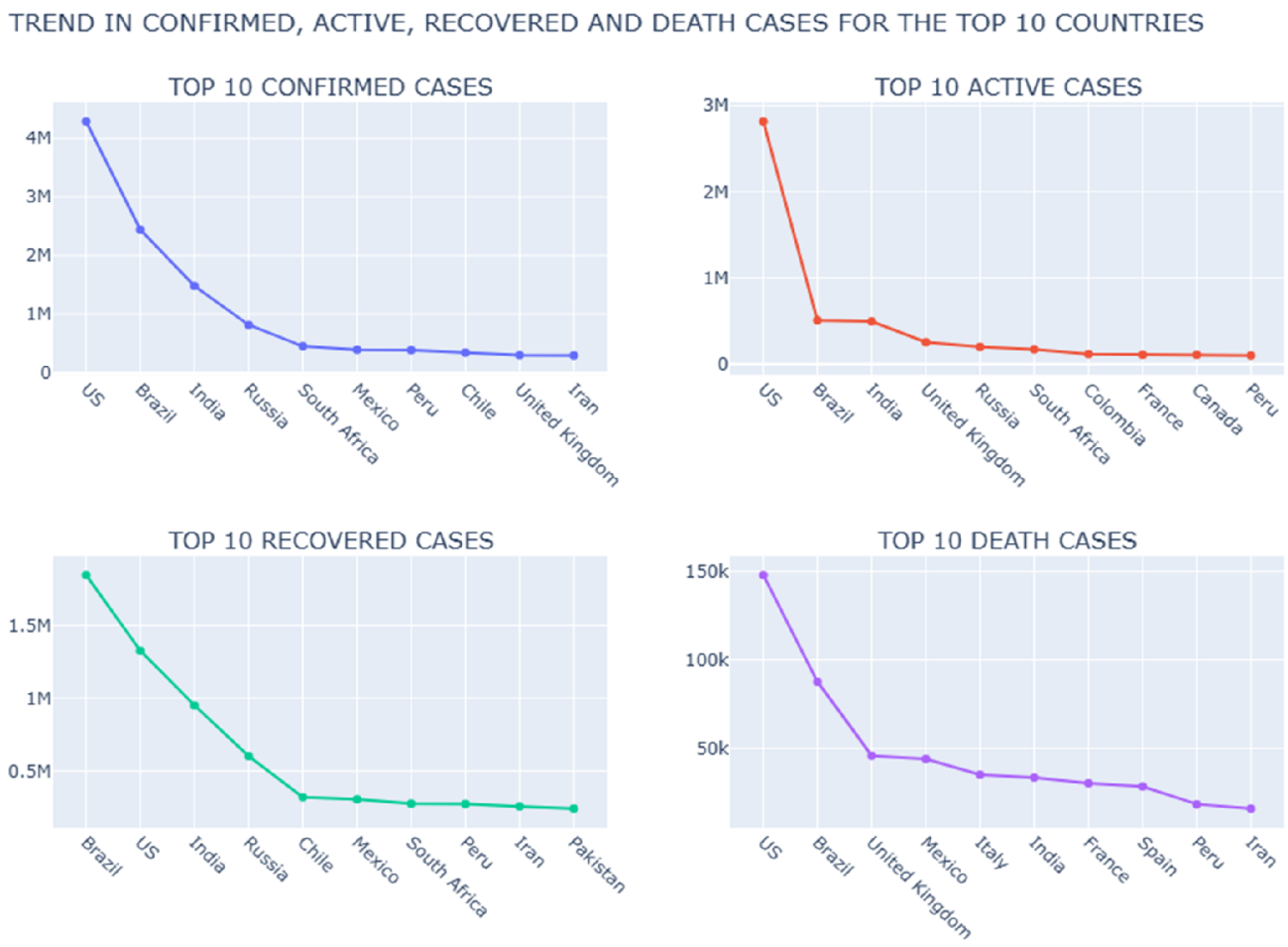
The relatively low recovery rate of 28% exposes some critical challenges:

a) **Strain on the healthcare system:** the low recovery percentage suggests a significant pressure on global healthcare systems, with limited capacity to effectively treat and rehabilitate patients.

b) **Virus complexity:** the low recovery rate shows the severity of the virus which is potentially due to:
    - Severe respiratory complications
    - Complex immune system interactions
    - Limited understanding of effective treatment protocols

This percentage highlights the need for robust preventive measure and comprehensive tracking systems in each country.

C. Top 10 countries by number of cases



The United States of America stood out with the highest number of confirmed cases, active cases and death cases. Brazil on the other hand, showed a significant recovery rate (1.8 million recoveries).

Insights:

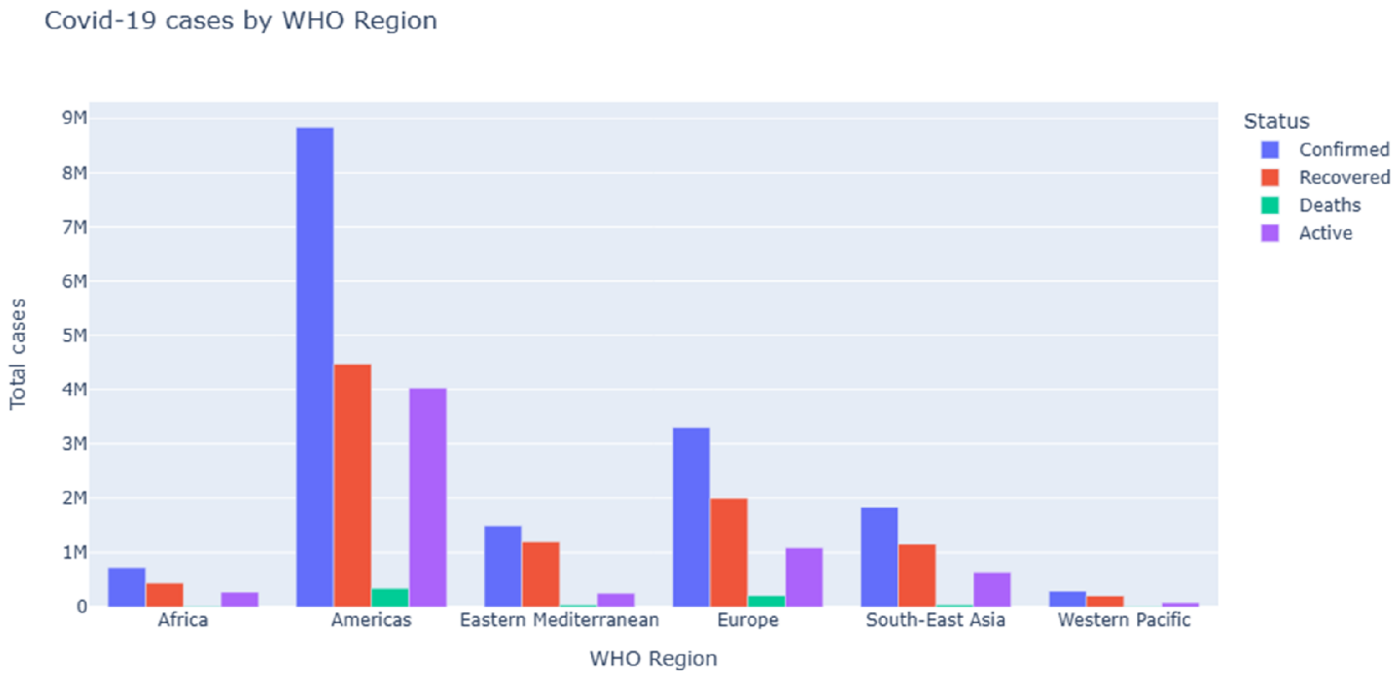
The USA's high case numbers can be attributed to:

a) Initial delays in implementing comprehensive preventive measures
b) High population density and passenger in-bound to the region.

Brazil's notable recovery rate could suggest:

a) Potentially more coordinated healthcare response
b) Effective implementation of treatment protocols
c) They utilized more testing and early intervention strategies
d) Possibility of a younger population (younger people are more likely to recover quickly than the elderly because of their strong immune system.

D. Total Cases by Region



Breakdown of the chart above:

Americas Region

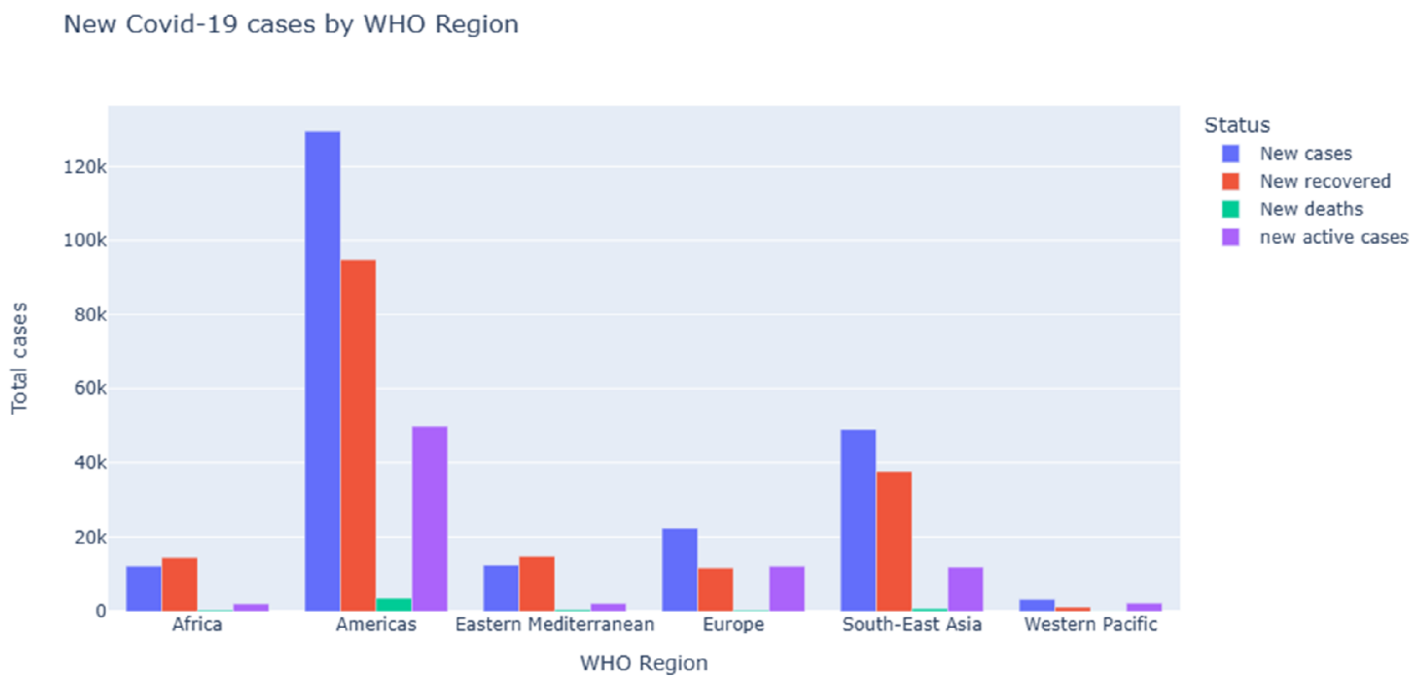
a) 54% of confirmed cases
b) 48% of recoveries
c) 63% of active cases

Insights:

These statistics suggests that there are challenges in controlling the virus transmission despite significant recovery efforts. This is probably due to the following:

a) Overwhelmed medical infrastructure
b) Dense urban population
c) Limited initial quarantine effectiveness
d) Varied compliance with preventive measures e.g. lockdown

E. Rise in new cases



The analysis of new case trends reveals a slight progression in the pandemic.

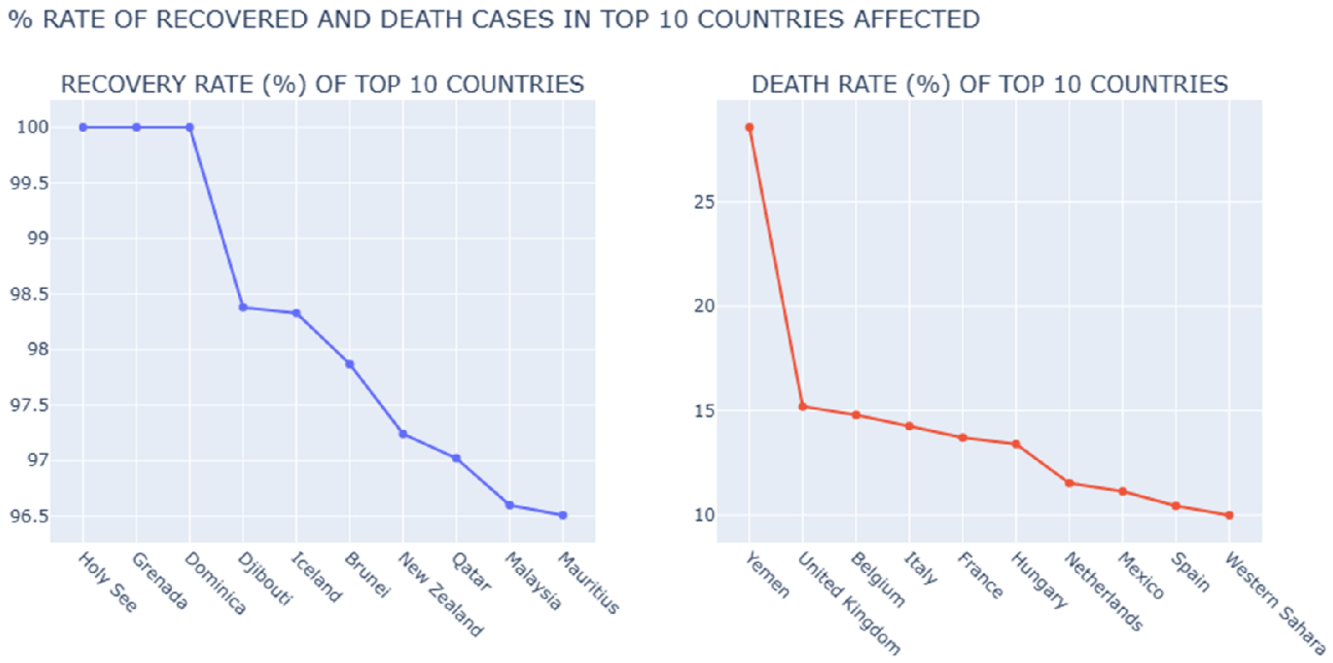
Americas continue to lead in case numbers

a) This time around, there is a decrease by 1% in active cases for Americas
b) South East Asia shows a significant increase in active cases.

For South East Asia, this could suggest:

a) Emerging hotspots
b) Potential new transmission waves
c) Evolving viral variants
d) Possible relaxation of preventive measures

F. Recovery Rate vs Death Rate



From the above:

a) Least affected countries showed a remarkable recovery rate.
b) Yemen emerged with the highest death rate -29%

Insight into Yemen's death rate

This high death rate may stem from:

a) Ongoing humanitarian crisis
b) Severely compromised healthcare infrastructure
c) Limited resources for pandemic response
d) Present political instability

However, the variations in recovery rate emphasize the importance of:

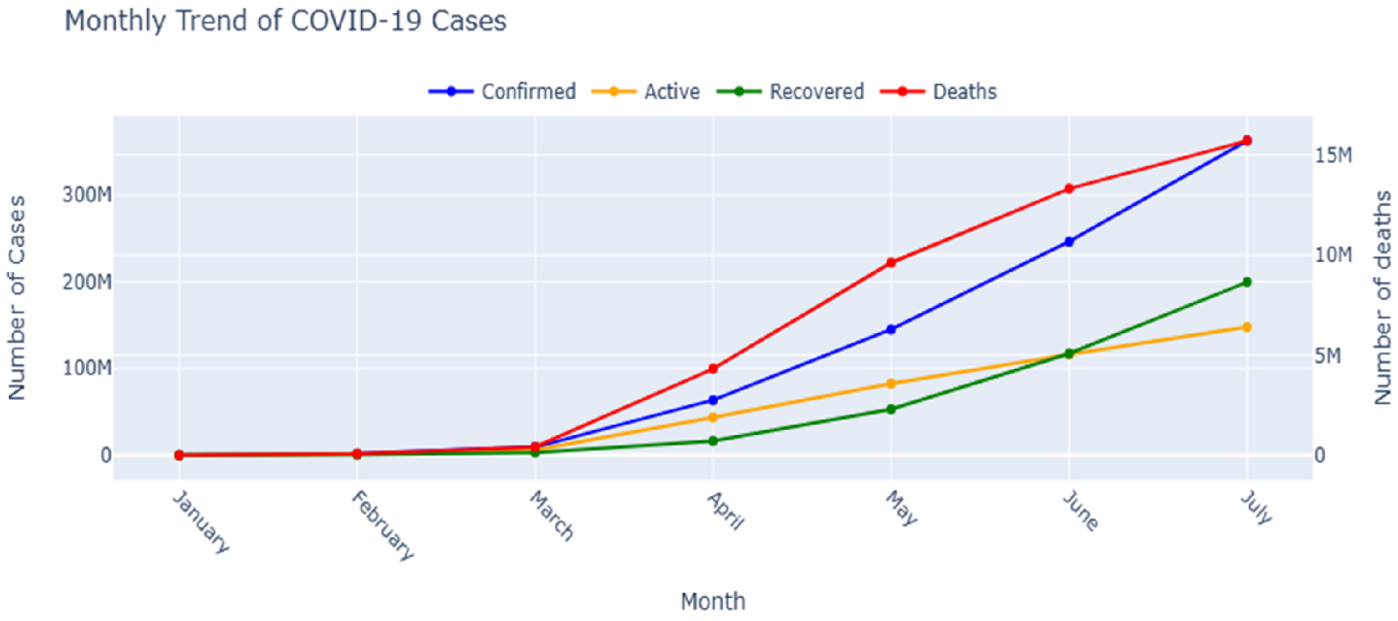
a) Robust healthcare systems
b) Early intervention strategies
c) Comprehensive medical resources and effective government response
d) Low population / in-bound travelers into those countries, which helps minimize the transmission rate.

Time Series Analysis

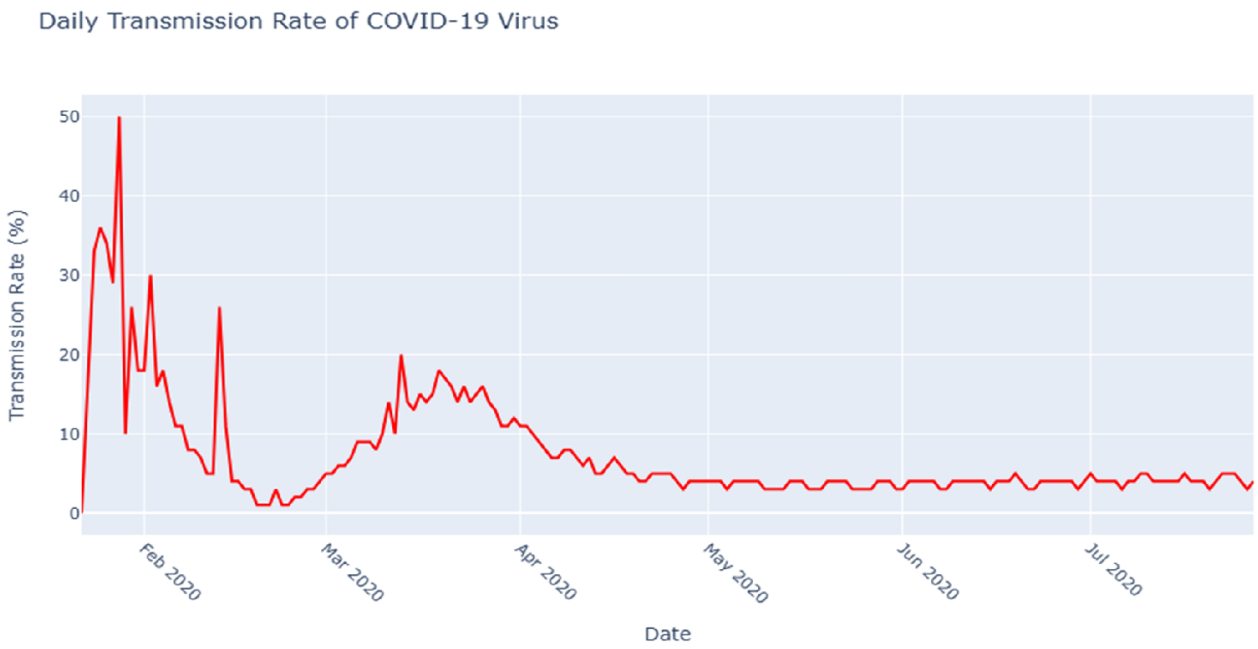
Note that the no. of deaths depicted in the graph by the red marker has its separate axes (secondary axis) on the right, so as not to confuse it as belonging to the primary axis.

- **Monthly Trend**

Over the months, the number of confirmed cases rose progressively at an alarming rate. This was also the same for the number of recoveries, death and active cases.



- **Trends in Transmission Rate**



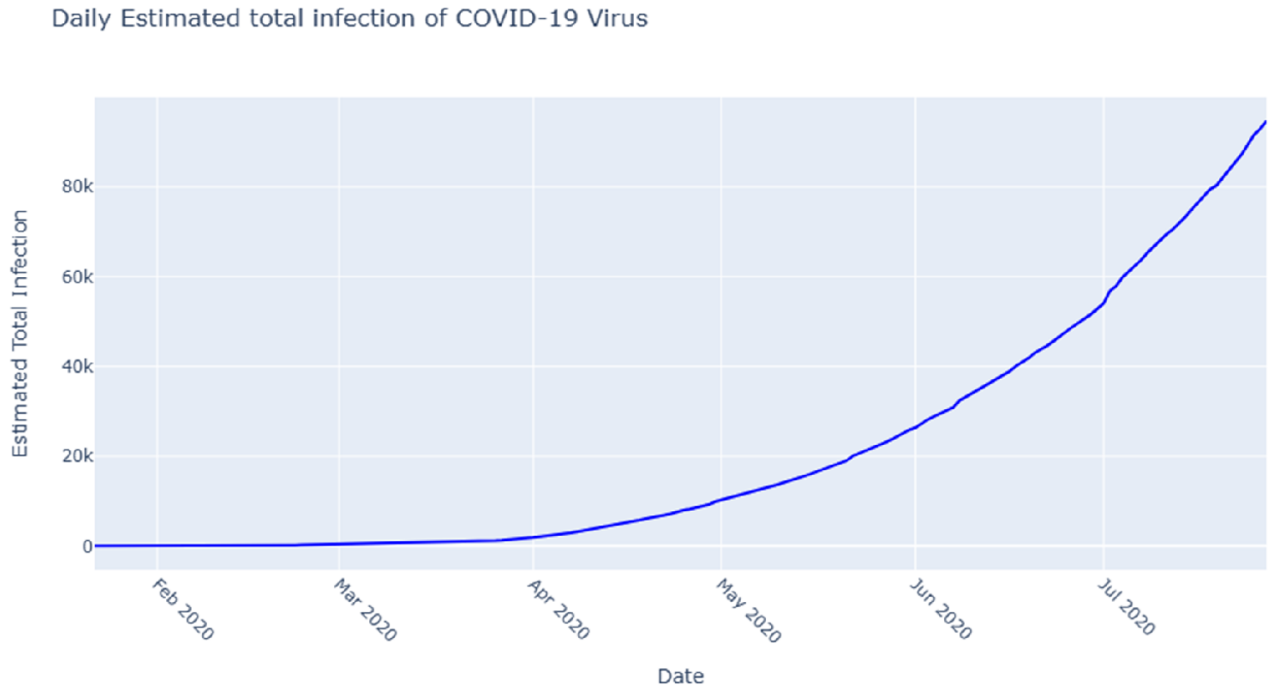
Insights:

From the graph, we can see that transmission rate rose up to 50% in January 28th. This figure shows that 50% of all active cases recorded that day led to new infections. This means at that this point the virus had spread at an alarming rate. But we see that the following day, it dropped by 40%, which means the number of new infections was reduced significantly. The transmission rate rose and fell and reached its final peak at 26% before dropping off again. From March 19th, we can see a progressive decline in the transmission rate, which meant that there were lesser number of new infections recorded as a result of the reduced transmission rate.

Implications:

The insights above highlight the difficulty in immediately controlling the rapid spread of the virus, as shown by the alarming 50% increase in transmission rate. However, the subsequent drops and the eventual sustained decline suggest that the interventions put in place were gradually gaining traction in curbing the transmission.

- **Trends in Estimated Total Infection**



Insights:

a) There was a strong positive correlation between estimated total infections and confirmed/active cases which suggests that as the total estimated infections rise, the number of confirmed and active cases increases in a nearly proportional manner.

Implication:

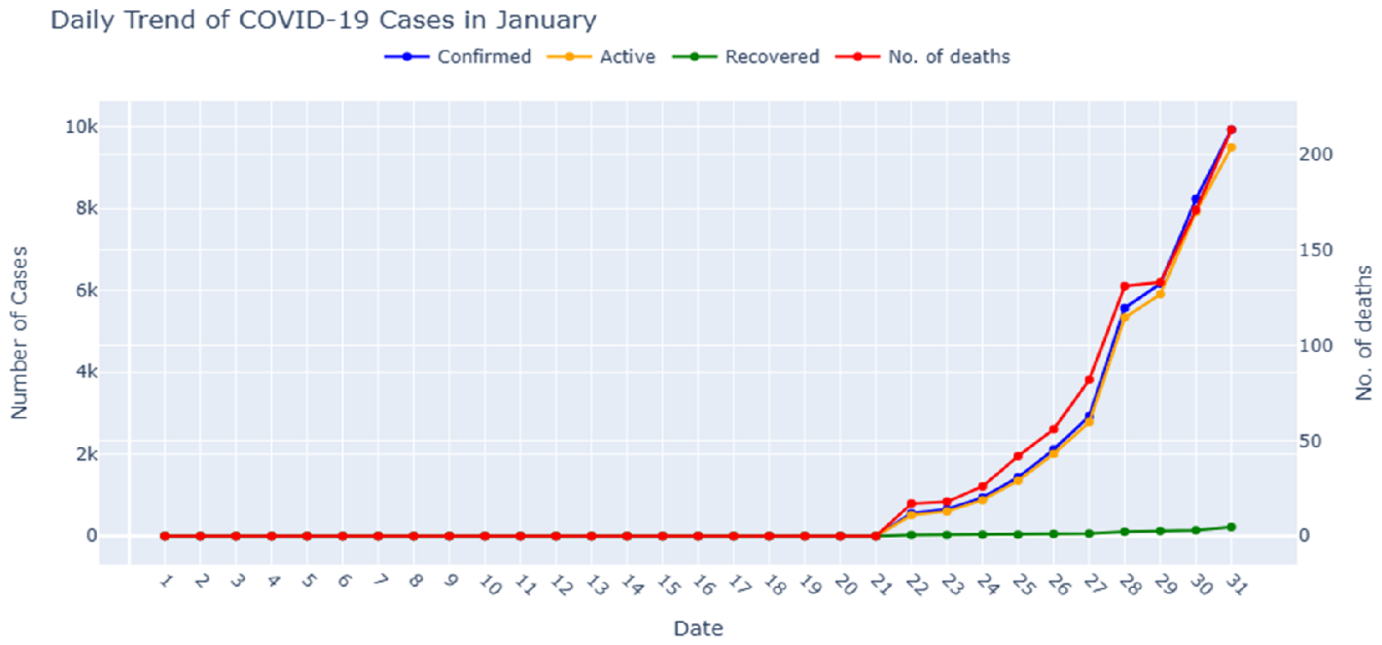
a) This tells us that that the estimated total infection is a good means for tracking the overall progression of the pandemic. Closely monitoring the estimated total infections can provide an early warning signal for potential surges in confirmed and active cases.

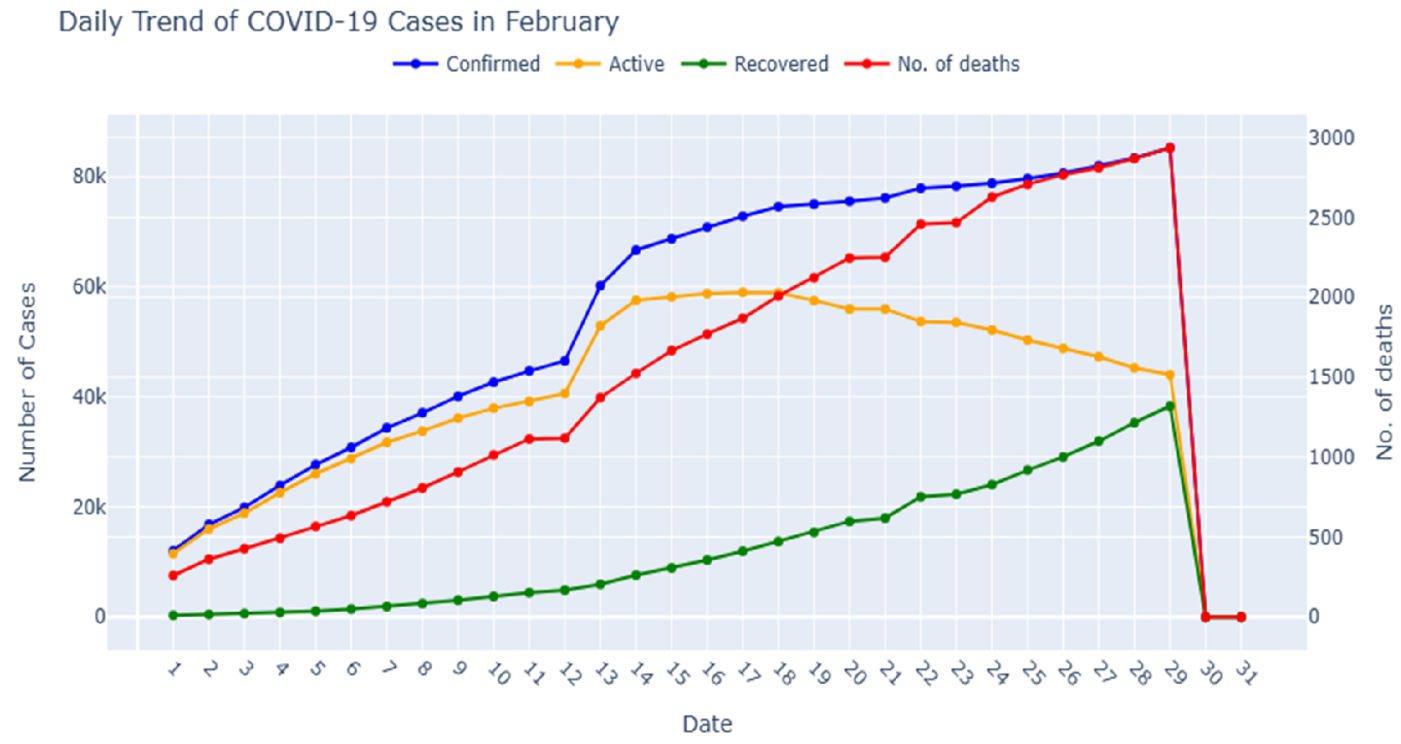
Daily Trend

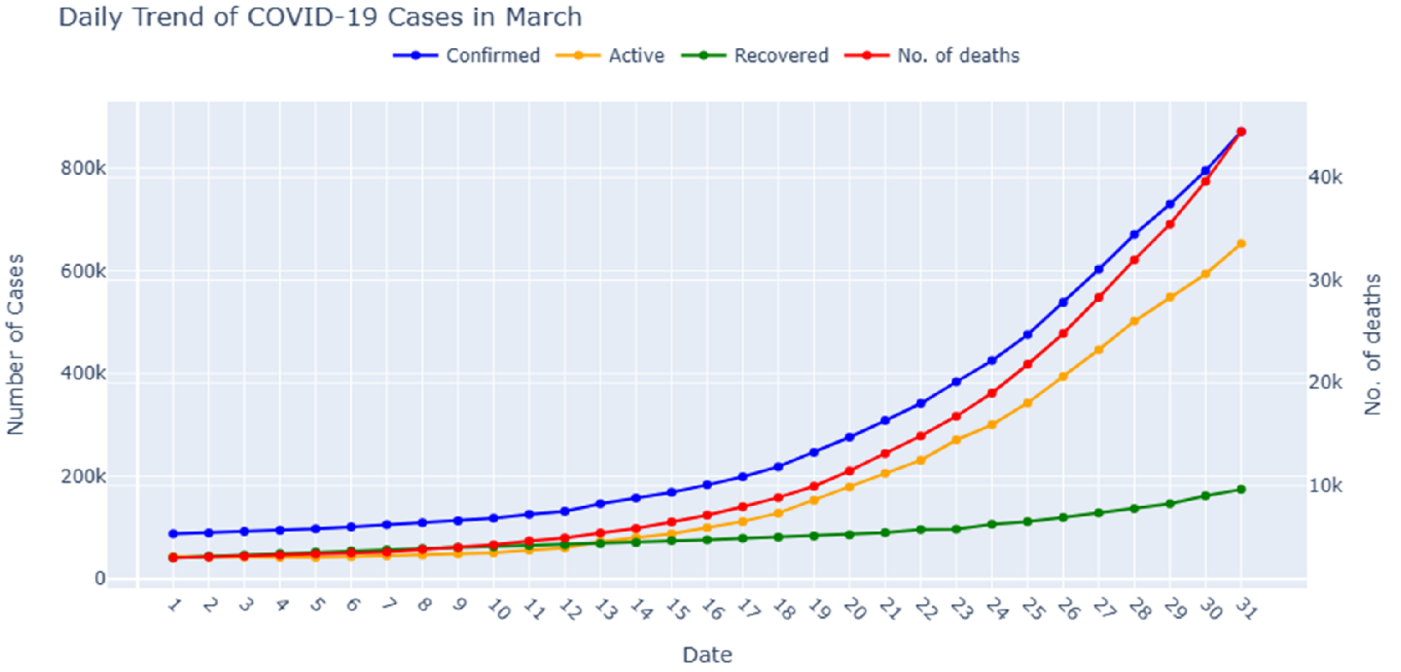
Significant observations include:

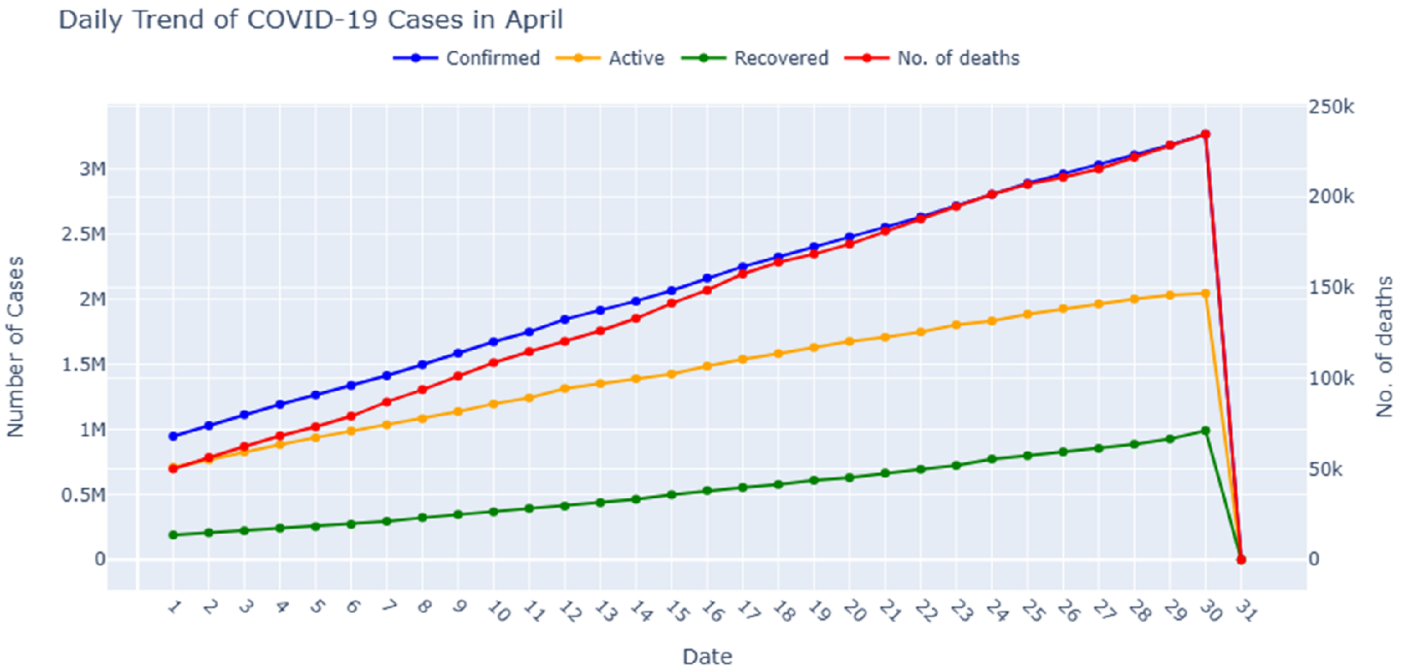
a) Spike in monthly case numbers

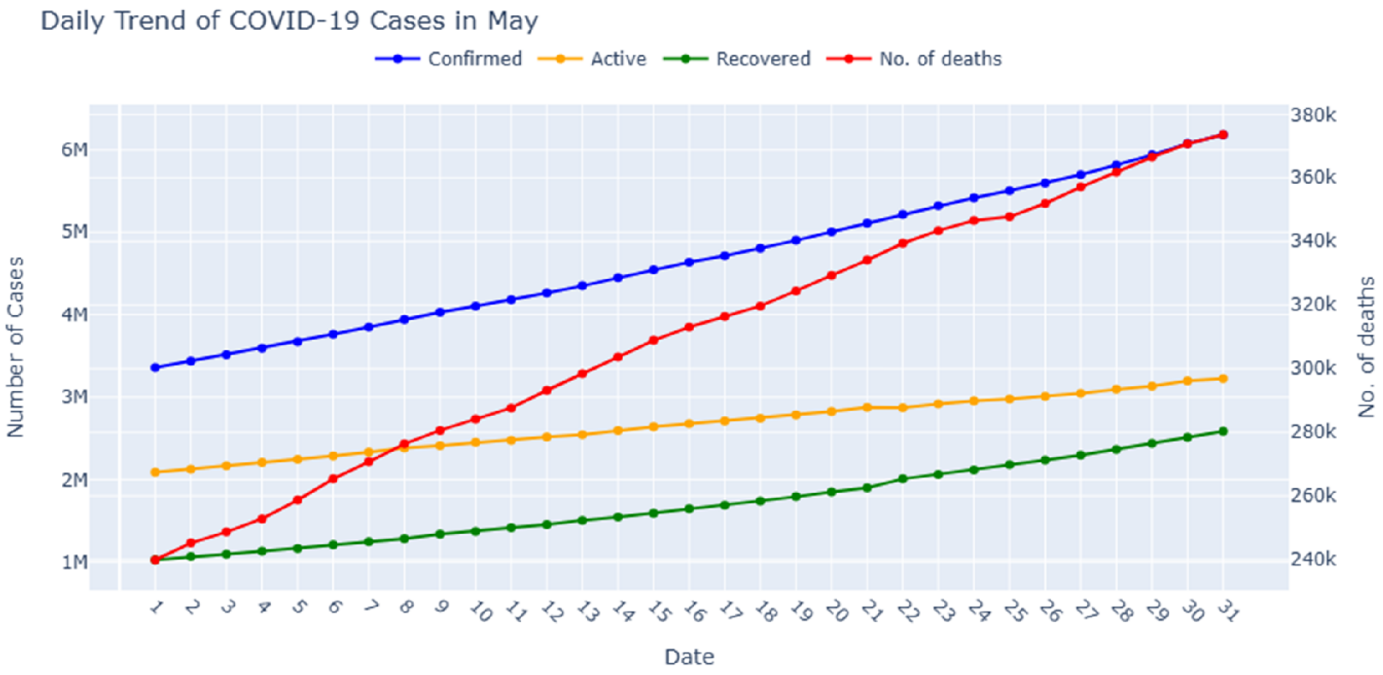
b) Increasing transmission rates

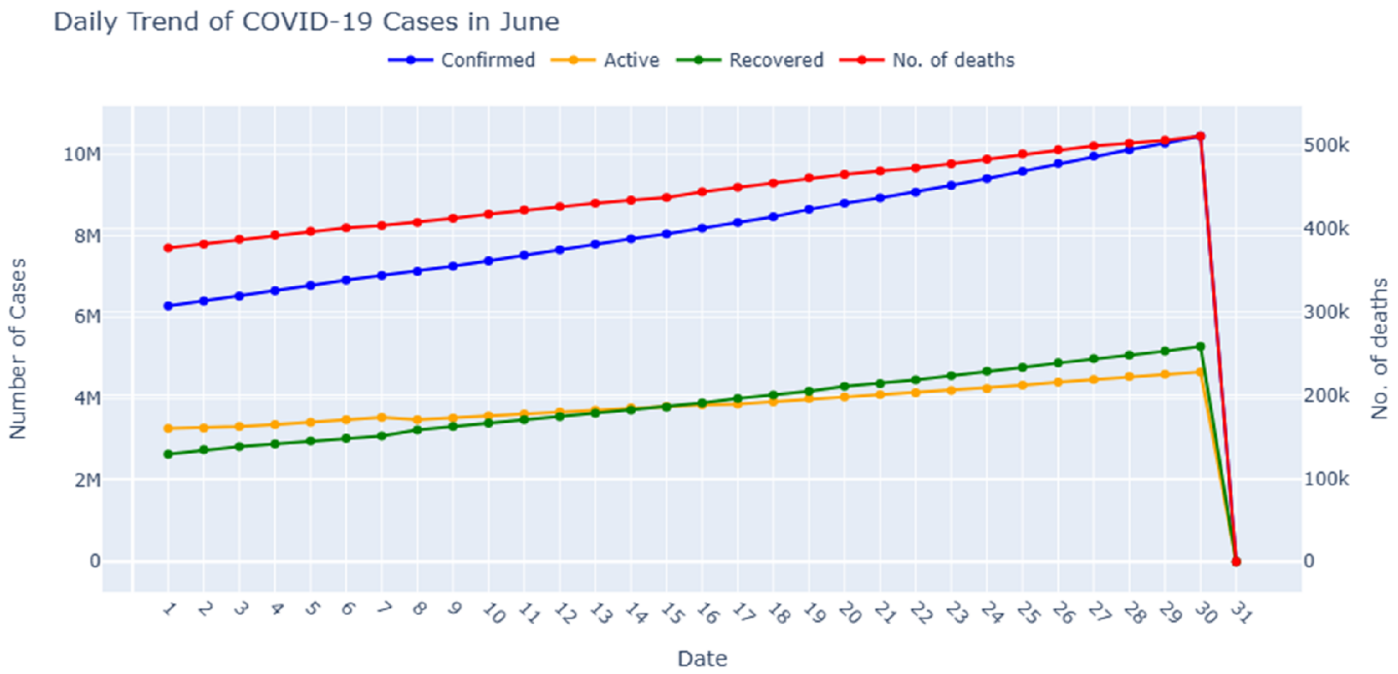


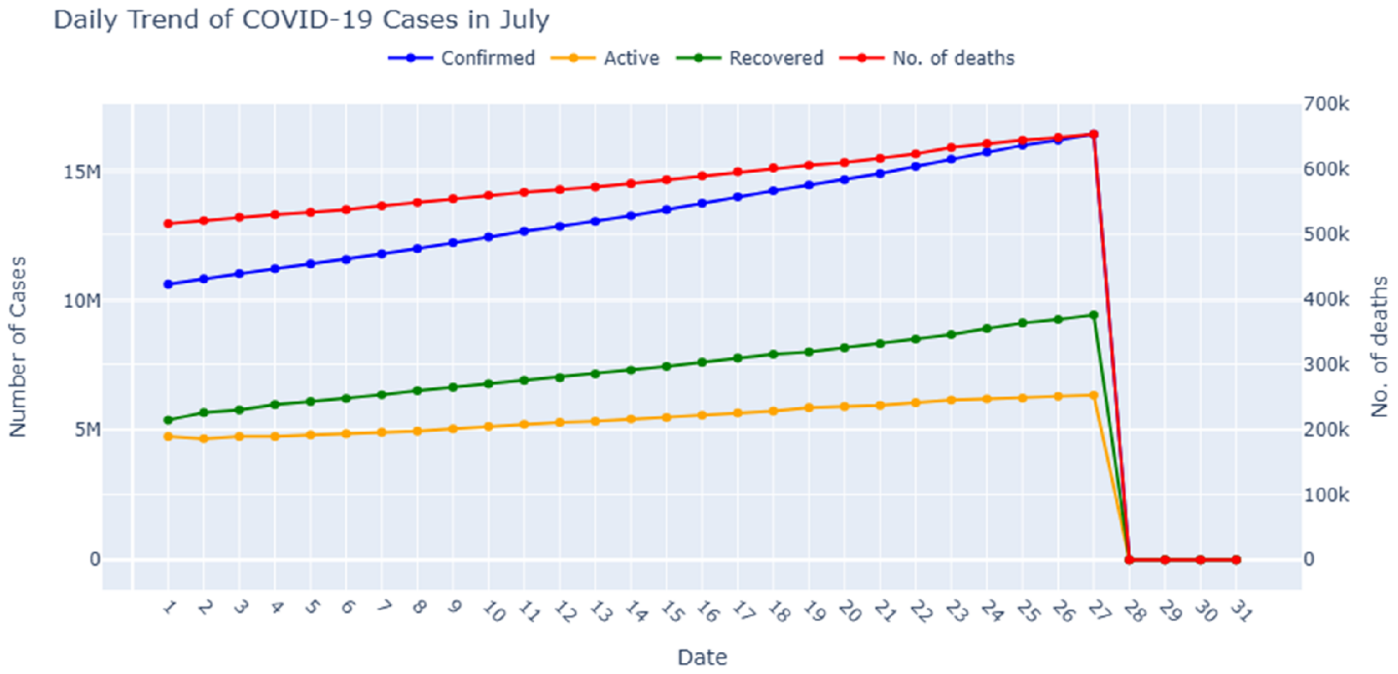

Daily Trend of COVID-19 Cases in February
Confirmed
Active
Recovered
No. of deaths
Number of Cases
No. of deaths
Date


Daily Trend of COVID-19 Cases in March
Confirmed
Active
Recovered
No. of deaths
Number of Cases
No. of deaths
Date


Daily Trend of COVID-19 Cases in April
Confirmed
Active
Recovered
No. of deaths
Number of Cases
No. of deaths
Date


Daily Trend of COVID-19 Cases in May
Confirmed
Active
Recovered
No. of deaths
Number of Cases
No. of deaths
Date


Daily Trend of COVID-19 Cases in June
Confirmed
Active
Recovered
No. of deaths
Number of Cases
No. of deaths
Date

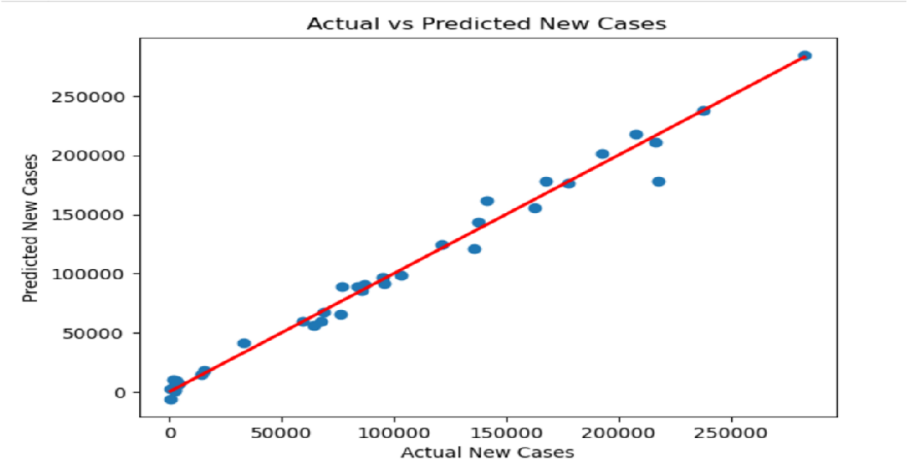

Daily Trend of COVID-19 Cases in July
Confirmed
Active
Recovered
No. of deaths
Number of Cases
No. of deaths
Date

Predictive Modeling

Model Architecture

To further analyze the factors contributing to the spread of the virus, we developed a **Logistic Regression Model and a Random Forest Regression Model** to predict the rate of spread of the virus based on the relevant features identified in the EDA phase.

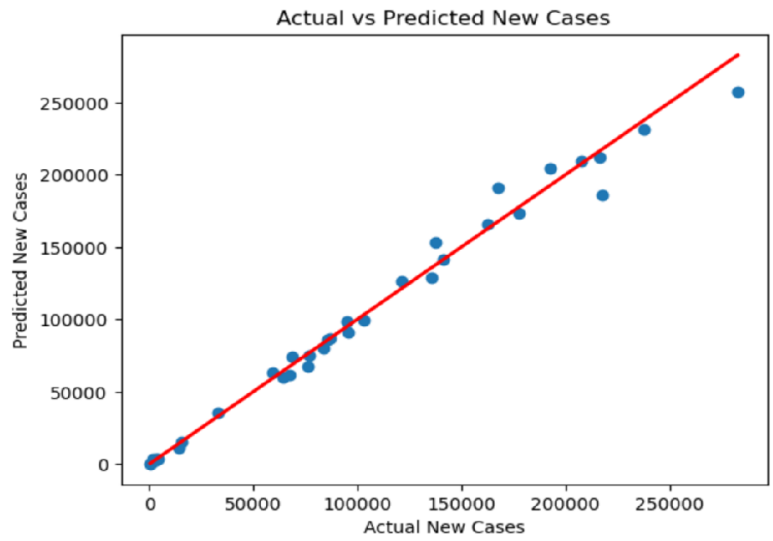
A. Model Performance – Linear Regression



a) **Prediction Accuracy / R- Squared value** = 0.99

b) **Mean Squared Error (MSE)** = 90,256,812.65

B. Random Forest Regression



a) **Prediction Accuracy / R-Squared value** = 0.99

b) **Mean Squared Error:** 78,013,767.21

Understanding the Results of Our COVID-19 Predictions from our models

In our analysis of COVID-19 new cases using a linear regression model and the random forest regressor, we obtained two important results: **The R-squared value and the Mean Squared Error (MSE).**

- **R-Squared Value = 0.99:** this value means that our model explains 99% of the changes in the number of new cases. In other words, our predictions are very close to the actual numbers most of the time. This is an excellent result, indicating that we have a strong model that captures the trends in COVID-19 spread effectively.

- **Mean Squared Error (MSE) = 90,256,812.65:** While our model fits the overall data well (as shown by the high R-squared value), this high MSE suggests that there are some individual cases where our predictions might be quite different from what actually happened. Since COVID-19 is constantly changing, it's important to keep updating our model with new data to ensure that it remains accurate over time.

Recommendations

Based on our analysis and the insights we've gathered; we propose the following strategies to enhance our response to COVID-19 and future pandemics:

1. Improved Global Coordination and Collaboration -

a) **Standardized International Protocols:** Establish clear and consistent pandemic response protocols that countries can adopt. This will ensure a unified approach during health crises.

b) **Build Robust Communication Networks:** Create strong global health communication networks that facilitate timely information sharing among nations, allowing for swift action and coordination.

2. Healthcare System Preparedness -

A resilient healthcare system is vital for managing outbreaks. We suggest:

a) **Investment in Medical Infrastructure:** Allocate resources to build scalable medical facilities that can adapt to surges in patient numbers, ensuring that healthcare systems can respond effectively during crises.

b) **Flexible Resource Allocation:** Develop strategies for reallocating healthcare resources dynamically based on real-time needs, ensuring that critical areas receive support when necessary.

3. Advanced Monitoring Systems -

To stay ahead of potential outbreaks, we need to enhance our monitoring capabilities:

a) **Real-Time Disease Tracking:** Implement global disease tracking systems that provide real-time data on infection rates and trends, enabling proactive responses to emerging threats.

b) **Early Warning Detection Systems:** Invest in epidemiological detection systems that can identify potential outbreaks at their onset, allowing for rapid intervention.

4. Public Health Education -

Empowering the public with knowledge is essential for effective pandemic management. This can be done through:

a) **Comprehensive Awareness Campaigns:** Launch public awareness campaigns that inform communities about preventive measures, vaccination benefits, and health guidelines.

b) **Clear Communication Strategies**: Develop clear and accessible health communication strategies that reach diverse populations, ensuring everyone understands the importance of following health directives.

5. Research and Development -

Ongoing research is critical for improving our understanding of infectious diseases:

a) **Accelerate Research on Viral Transmission**: Prioritize research initiatives focused on understanding how viruses spread, which will inform better prevention strategies.

b) **Support Rapid Development Initiatives:** Encourage and fund initiatives aimed at speeding up the development of vaccines and treatments, ensuring we are prepared for future outbreaks.

By implementing these recommendations, we can enhance our preparedness for COVID-19 and other potential pandemics, ultimately protecting public health and saving lives.

Conclusion

The insights that have been derived from this analysis provides a crucial framework for understanding and preparing for future global health emergencies, which highlights the importance of proactive, comprehensive, and collaborative approaches to emerging health threats.

Reported by:

Mercy E. Festus

3MTT Cohort 2 Data Science Fellow,

Kaduna State.