# A COMPREHENSIVE DATA-DRIVEN STUDY OF THE COVID-19 PANDEMIC

Big Data Final Report

Authors:
Weiyi Luo (WL3398)
Jia Yang (JY5081)
Geng Niu (GN2279)

2025-12-9

# 1. Executive Summary

This project analyzes the global progression of COVID-19 using the large-scale OWID dataset, which includes over 379k records covering cases, deaths, vaccinations, testing, demographics, and government policies. To process this high-volume data, we built an end-to-end big data pipeline using Hadoop HDFS, Hive, and Spark, and organized it into a multi-layer warehouse (ODS, DWD, DWS) for raw ingestion, standardized cleaning, and analytical aggregation. Python and Spark were then used to visualize the DWS outputs.

Our analysis focuses on eight key research questions, including global and continental trends, country-level infection and mortality differences, vaccination impact, testing capacity, socioeconomic variance, and policy effectiveness. For each question, we designed targeted Hive and Spark workflows and generated visualizations to support data-driven insights. The results reveal clear continental outbreak patterns, strong correlations between vaccination rates and reduced mortality, and notable lag effects between government policy strictness and transmission rates.

This report summarizes the data engineering pipeline, analytical methods, visual findings, and key observations, along with technical challenges, lessons learned, and future improvements. Overall, the project demonstrates how distributed data tools can generate meaningful and scalable insights from complex public health datasets.

**All the code and datasets can be found on Github: https://github.com/ByFan-coder/COVID-19-Big-Data-Analytics-and-Visualization**

# 2. Problem Statement and Objectives

## 2.1 Background

The OWID COVID-19 dataset provides detailed, daily global records across cases, deaths, vaccinations, testing, and policy responses. After examining the scale and richness of this data, we identified several key patterns and questions worth exploring, which directly shaped the eight research questions of this project.

## 2.2 Research Questions

Our project aims to answer eight core questions originally proposed in our project plan:

1. How have global confirmed cases and deaths evolved over time across continents?
2. Which countries experienced the highest infection rates and mortality ratios?
3. How do vaccination rates correlate with changes in confirmed cases and deaths?
4. Which regions show the fastest vaccination progress and strongest recovery patterns?
5. What are the time-series patterns of daily new cases, deaths, and vaccinations?
6. How have testing capacities affected case detection and positivity rates?
7. What differences exist between high-income and low-income countries in pandemic outcomes?
8. How effective were government policies in limiting transmission, particularly with lag effects?

These questions guided the design of our data warehouse, analytical tables, and visualization workflows.

## 2.3 Project Objectives

To address the above questions, this project has three main objectives:

- **Build a scalable big data processing pipeline** using Hadoop HDFS, Hive, and Spark to handle the OWID dataset (379k+ rows, 61 variables).
- **Create a structured data warehouse (ODS → DWD → DWS)** to clean raw data, standardize inconsistent fields, and generate analytical summary tables tailored to each research question.
- **Conduct comprehensive data analyses and visualizations** to identify trends, correlations, and cross-regional differences, providing interpretable insights into the evolution of COVID-19 and the effectiveness of global responses.

# 3. Dataset Description

## 3.1 Data Source

Our analysis is based on the *Our World in Data (OWID) COVID-19 Dataset*, an open-access global dataset that compiles daily statistics on the pandemic from national health agencies, the WHO, and other international sources. The dataset contains more than 379k+ rows and 61 variables, covering nearly every country and region from early 2020 to 2024.

## 3.2 Dataset Structure

Although the dataset contains many fields, our analysis focuses primarily on the following categories:

- **Epidemiological Indicators:**

  total_cases, new_cases, total_deaths, new_deaths, new_cases_per_million.

- **Vaccination Data:**

  people_vaccinated, people_fully_vaccinated,

  people_vaccinated_per_hundred, total_vaccinations.

- **Testing & Positivity Statistics:**

  total_tests, new_tests, positive_rate, tests_per_case.

- **Policy Response Indicators:**

  stringency_index, reproduction_rate.

- **Demographic & Socioeconomic Variables:**

  population, median_age, gdp_per_capita, human_development_index.

- **Geographic Fields:**

  iso_code, location, continent, date.

### Dataset Structure Overview

| Category | Key Variables | Purpose |
|---|---|---|
| Epidemiological Indicators | total_cases, new_cases, total_deaths, new_deaths, new_cases_per_million | Track outbreak severity and daily |

| | | |
|---|---|---|
| | | transmission levels. |
| 💉 Vaccination Data | people_vaccinated, people_fully_vaccinated, people_vaccinated_per_hundred, total_vaccinations | Measure vaccination rollout and evaluate its impact on cases and mortality. |
| 🔬 Testing & Positivity Statistics | total_tests, new_tests, positive_rate, tests_per_case | Assess testing capacity and infection detection accuracy. |
| 🏛️ Policy Response Indicators | stringency_index, reproduction_rate | Analyze government restrictions and their effect on transmission. |
| 👥 Demographic & Socioeconomic Variables | population, median_age, gdp_per_capita, human_development_index | Explain cross-country differences in infection, mortality, and recovery. |
| 🌍 Geographic Fields | iso_code, location, continent, date | Identify regions and align all time-series records. |

These variables directly support our eight research questions related to trends, vaccination, testing, policy effectiveness, and cross-country differences.

## 3.3 Data Challenges

The OWID dataset, while comprehensive, contains several challenges that require careful preprocessing. Many fields include missing or partially reported values, especially in testing, hospitalization, and policy indicators. Reporting standards also vary across countries, leading to inconsistencies in case definitions, testing frequency, and vaccination updates. Additionally, some dates contain sudden spikes caused by backlogged reporting rather than real-time changes. These issues make data cleaning, type standardization, and validation essential before conducting reliable analysis.

# 4. System Architecture and Technology Stack

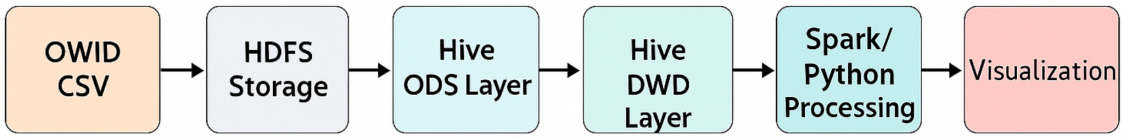
## 4.1 Why Big Data Tools (Spark & Hive)

The OWID dataset contains over 379k+ records and multiple high-dimensional variables, making distributed processing essential for efficient analysis. Hive provides a structured data warehouse environment that supports our multi-layer design (ODS → DWD → DWS), enabling standardized cleaning and analytical table construction. Spark complements this by offering fast in-memory computation for large-scale aggregations and serving as the primary engine for data visualization and exploratory analysis.

## 4.2 System Architecture Overview

Our system follows a full big data pipeline designed to manage ingestion, transformation, storage, and analysis. The workflow is structured as:

**OWID CSV → HDFS Storage → Hive ODS Layer → Hive DWD Layer → Hive DWS Analytical Tables → Spark/Python Processing → Visualization**




This architecture ensures that raw data is first captured and preserved, then cleaned and normalized, and finally aggregated into optimized analytical datasets that support our eight research questions.

## 4.3 Technology Stack

We used a combination of distributed computing and analytical tools, including:

- **Hadoop HDFS** for scalable data storage and file management.
- **Apache Hive** for schema definition, data transformation, partitioning, and data warehouse operations.
- **Apache Spark** for high-performance computation, analytical queries, and integration with Python for visualization.

- **Python (Pandas, Matplotlib, Seaborn)** for downstream analysis and figure generation.
- **Parquet Format** for efficient storage and optimized querying within the DWD and DWS layers.

This stack enables reliable data processing at scale while maintaining flexibility for analytical and visualization tasks.

# 5. Data Preprocessing & Warehouse Design (ODS → DWD → DWS)

A structured data warehouse design was implemented to ensure consistent data ingestion, cleaning, transformation, and analytical preparation. The pipeline follows a three-layer architecture—ODS, DWD, and DWS—allowing raw data to be standardized and optimized for downstream analysis and visualization.

## 5.1 ODS Layer (Raw Data)

The Operational Data Store (ODS) stores the dataset in its raw form immediately after being loaded into Hive.

- All fields are stored as **STRING** to preserve the original OWID schema.
- No cleaning, formatting, or type conversion is performed at this stage.
- Data is loaded into Hive as an **external table** to maintain direct referencing of the raw files in HDFS.

The ODS layer acts as the immutable baseline for subsequent transformations.

## 5.2 DWD Layer (Cleansed Data)

The Data Warehouse Detail (DWD) layer contains the cleaned and standardized version of the dataset.

- All numeric fields are **converted from STRING to DOUBLE** using explicit CAST operations.
- Records are partitioned by **continent** and **year**, improving query efficiency for regional and temporal analysis.
- Date values are standardized, and rows with incomplete or missing essential fields are filtered out.

- The table is stored in **Parquet format**, enabling efficient columnar storage, compression, and optimized Spark and Hive querying.

This layer ensures schema consistency and prepares the dataset for analytical aggregation.

### 5.3 DWS Layer (Analytical Tables)

The Data Warehouse Summary (DWS) layer contains four analytical tables specifically designed to address the eight research questions in our project. Each table aggregates and restructures the DWD data to support targeted analyses.

1. **dws_time_series_trend**
    - Provides continent-level daily aggregates of cases, deaths, vaccinations, and 7-day rolling averages for temporal trend analysis.
2. **dws_country_risk_profile**
    - Summarizes country-level total cases, deaths, infection rates, mortality ratios, and socioeconomic indicators for cross-country comparisons.
3. **dws_vaccine_testing_effect**
    - Combines vaccination coverage, testing metrics, positivity rate, and derived recovery indicators to evaluate vaccination and testing impact.
4. **dws_policy_effectiveness**
    - Links policy strictness (stringency index) with case trends and includes 7-day lag variables to assess the delayed effect of government interventions.

Together, these DWS tables serve as the analytical backbone of the project, enabling efficient Spark processing and visualization for each research question.

## 6. Hive Processing Pipeline

The Hive processing pipeline handles the backend data transformation from raw OWID records in HDFS to structured analytical tables used for Spark visualization.

### 6.1 ODS Table Creation

We first created an external ODS table (ods_covid_raw) that loads the raw CSV from HDFS with all fields stored as **STRING**. No cleaning or type conversion is performed in this stage; it serves as the direct copy of the original dataset.

### 6.2 DWD Layer Construction

The DWD table (dwd_covid_clean_full) contains the cleaned dataset. In this step:

- All numerical fields were **CAST from STRING to DOUBLE**.
- Rows with missing country, date, or continent were removed.
- A **year** column was extracted for partitioning.
- The table was stored in **Parquet** and partitioned by **continent** and **year**.

This layer standardizes the dataset and prepares it for analytical queries.

### 6.3 DWS Analytical Tables

Using the DWD layer, we generated four DWS tables to support analysis for the eight research questions:

1. **dws_time_series_trend** – daily continent-level totals and 7-day averages
2. **dws_country_risk_profile** – country-level infection and mortality metrics
3. **dws_vaccine_testing_effect** – vaccination, testing, positivity, and recovery indicators
4. **dws_policy_effectiveness** – policy strictness with 7-day lagged case trends

These tables aggregate the data into formats required for Spark visualization.

### 6.4 Export for Visualization

Each DWS table was exported using INSERT OVERWRITE DIRECTORY, producing CSV files that were later loaded into Spark and Python for plotting.

## 7. Spark Analytics Pipeline

The Spark analytics pipeline is responsible for transforming the DWS outputs into visual insights. After exporting the analytical tables from Hive, Spark and Python

were used to conduct additional preprocessing and create figures for each research question.

## 7.1 Data Loading and Preparation

We loaded all four DWS tables (time_series_trend, country_risk_profile, vaccine_testing_effect, and policy_effectiveness) into Spark and converted them to Pandas DataFrames when needed. Basic preprocessing was performed, including:

- Converting date strings to datetime
- Filtering missing or invalid values
- Sorting data chronologically
- Selecting specific continents or countries for comparison

These steps ensured consistent input for visualization.

## 7.2 Analytical Processing

Spark was used to create subsets of the data for each research question, such as:

- Continental and global time-series segments
- Top-10 countries by infection or mortality rate
- Vaccination and testing metrics for scatter analysis
- Policy strictness with lagged case values

This allowed lightweight analytical computation before visualization.

## 7.3 Visualization Workflow

Using matplotlib, we generated line charts, bar charts, and scatter plots based on Spark-processed data. These figures illustrate:

- Global and continental COVID-19 trends
- Country-level infection and mortality comparisons
- The relationship between vaccination/testing and case changes
- The lagged impact of government policy strictness

All visualizations directly correspond to the eight research questions and serve as the basis for the final analytical results.

# 8. Results and Analysis

## 8.1 RQ1: How did COVID-19 cases and deaths evolve across continents over time?

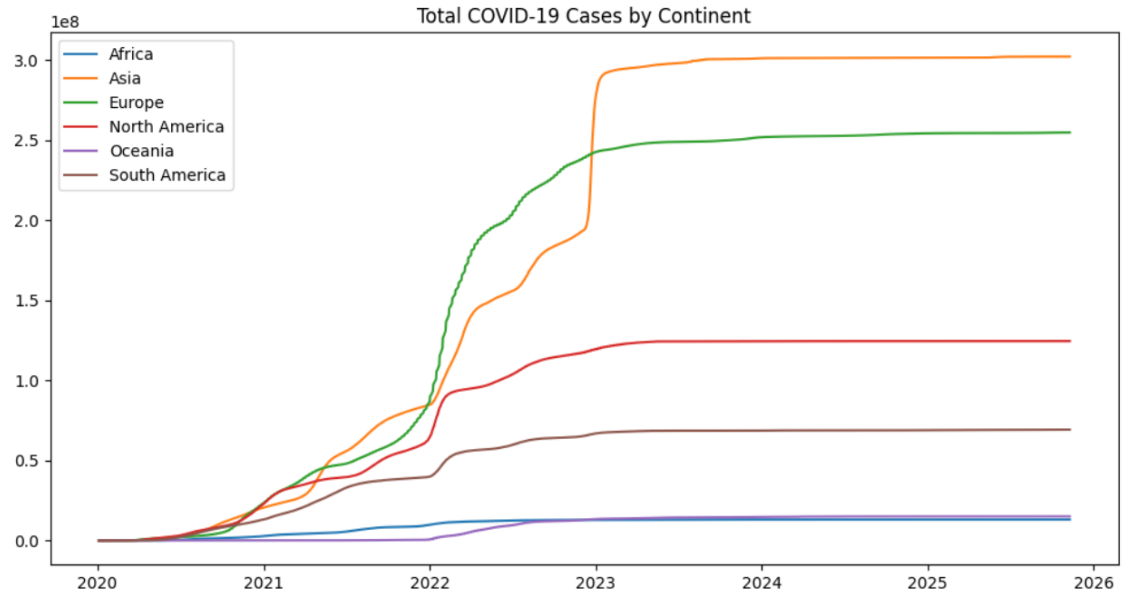
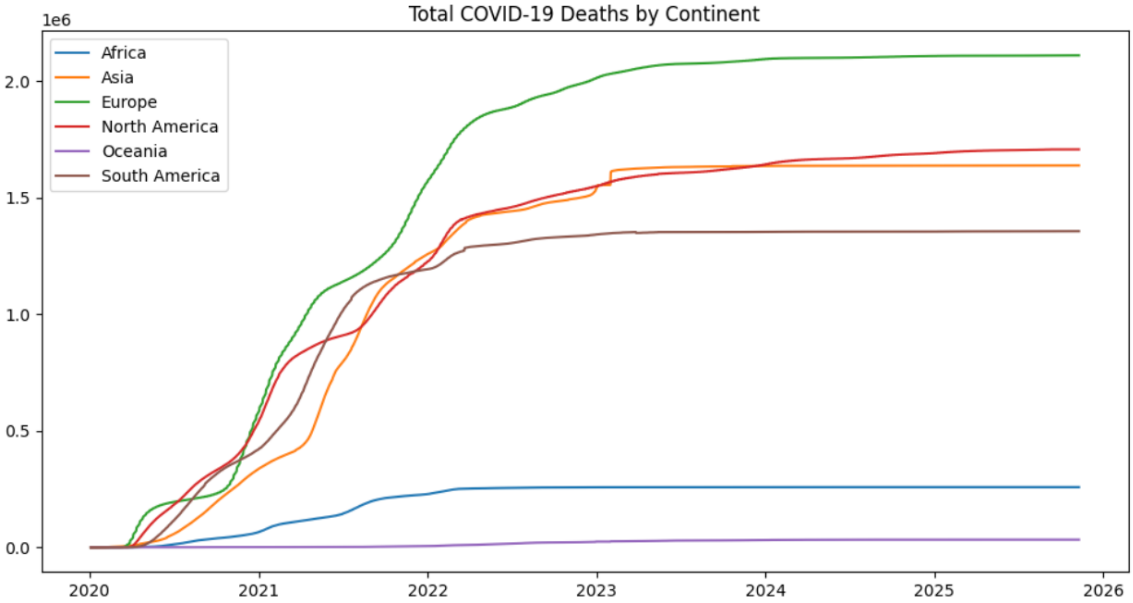


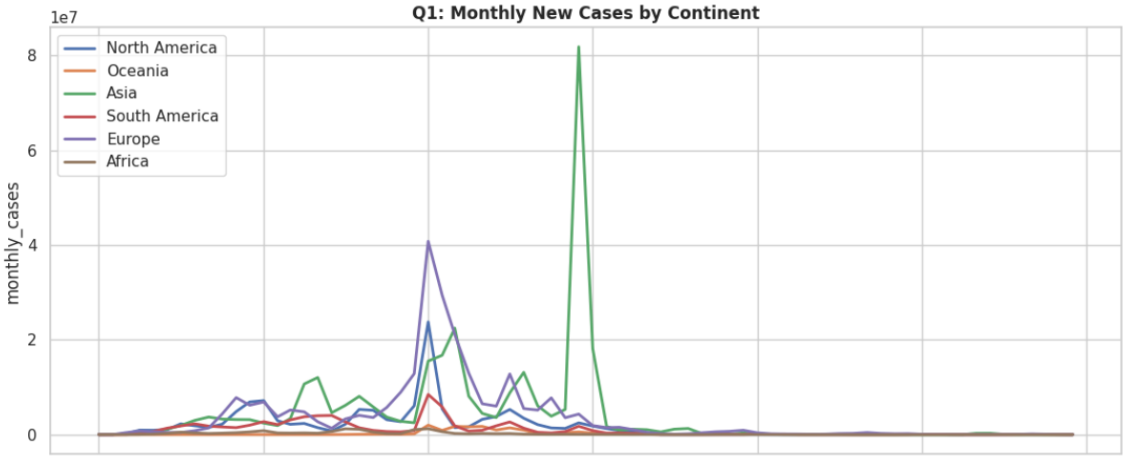

**Figure 1 (Hive): Total COVID-19 Cases by Continent**

*Europe and Asia show the fastest and highest cumulative case growth, while Africa and Oceania remain near zero.*



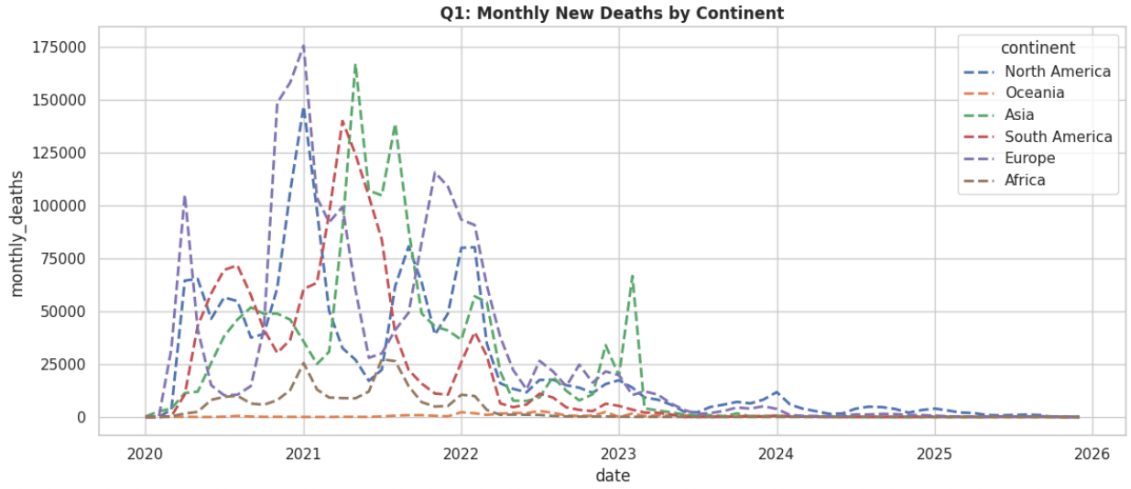
## Figure 2 (Hive): Total COVID-19 Deaths by Continent

*Europe records the highest cumulative deaths, followed by Asia, with Africa and Oceania showing minimal mortality.*




## Figure 3 (Spark): Monthly New COVID-19 Cases by Continent

*Asia and Europe display the strongest monthly case spikes, particularly during late 2021–2022, while Africa and Oceania remain low throughout.*




## Figure 4 (Spark): Monthly New COVID-19 Deaths by Continent

*Monthly deaths peak alongside case surges, with Europe and Asia showing the largest spikes before all regions decline sharply after 2022.*

### Summary

Across continents, Europe and Asia experienced the most severe and sustained outbreaks in both cases and deaths. Monthly trends reveal clear global waves, followed by a universal decline after 2022 as vaccination and immunity increased.

## 8.2 RQ2: Which countries experienced the highest infection rates and mortality ratios?

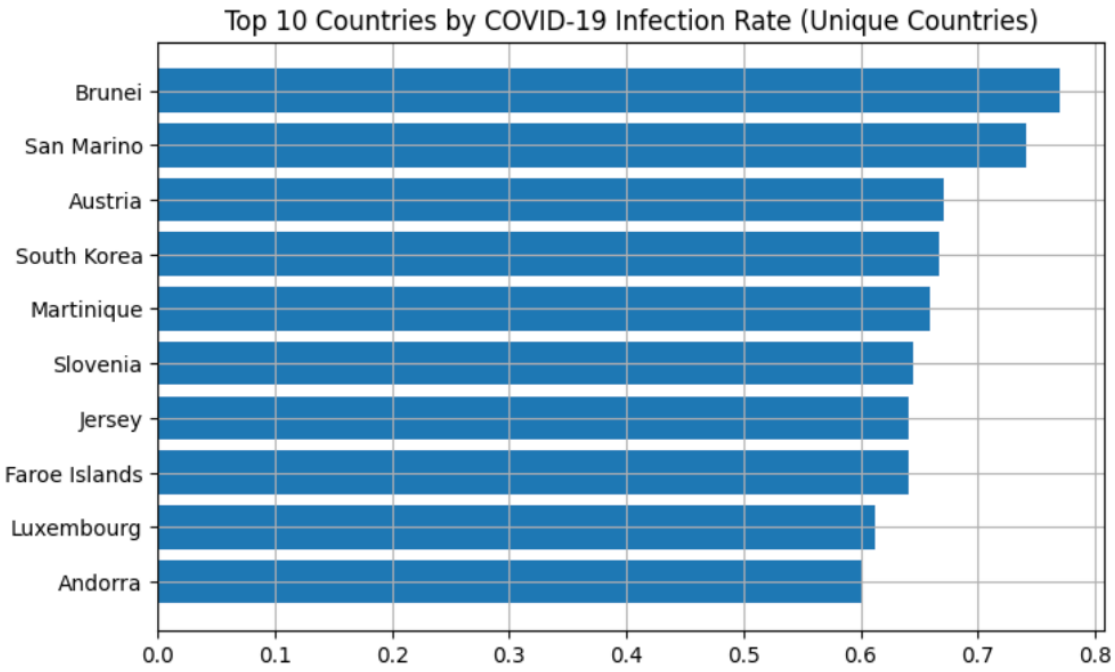
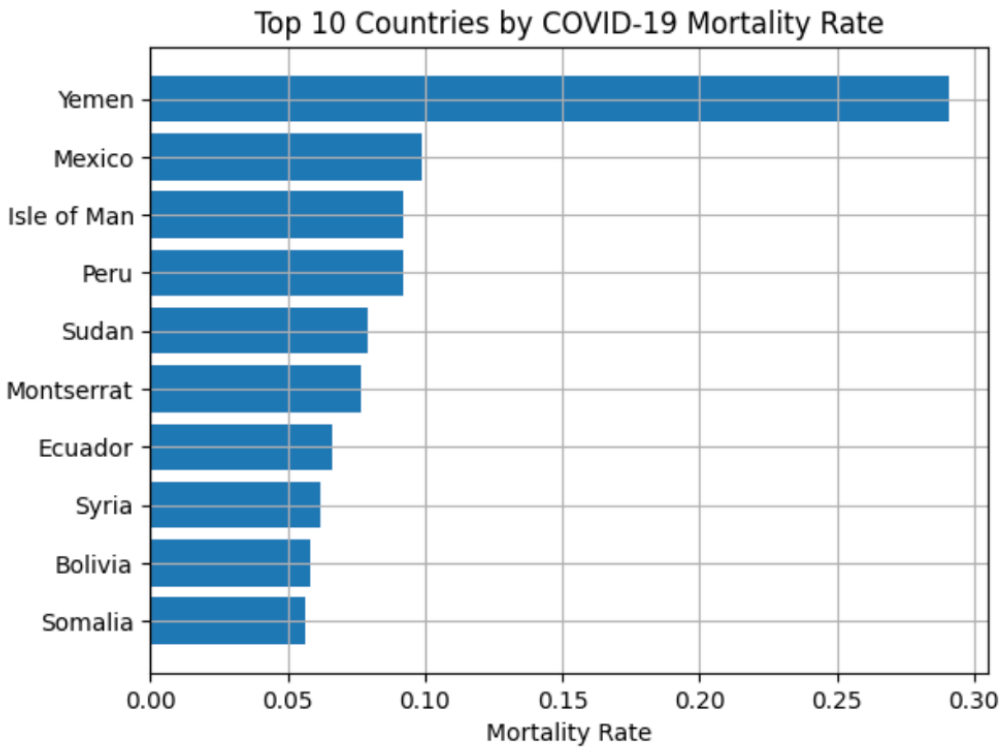


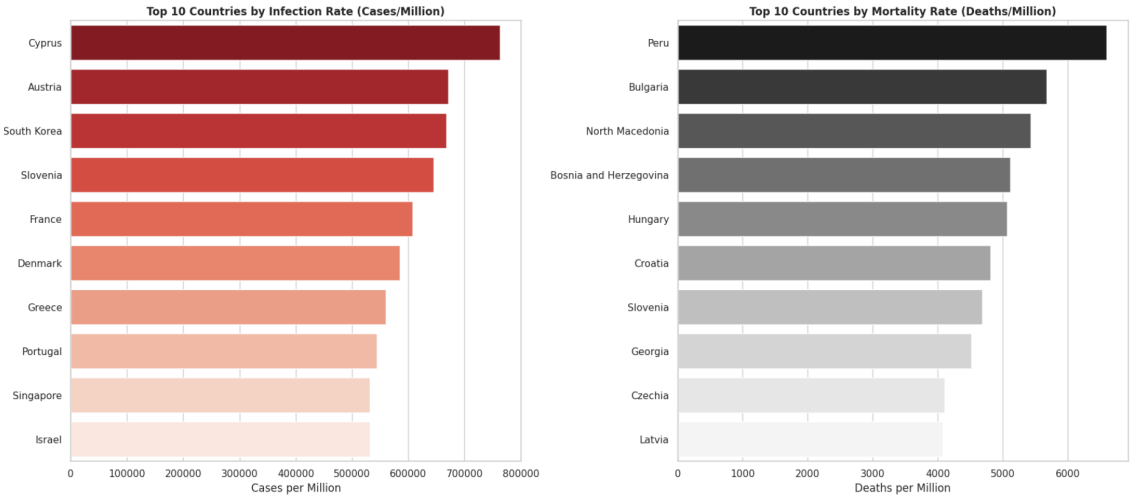

**Figure 5 (Hive): Top 10 Countries by COVID-19 Infection Rate**

*Countries such as Brunei, San Marino, and Austria show the highest infection rates, with several exceeding 60–75% of their populations.*




## Figure 6 (Hive): Top 10 Countries by COVID-19 Mortality Rate

*Yemen, Mexico, and Isle of Man exhibit the highest mortality ratios, with Yemen's fatality rate rising above 29%.*




## Figure 7 (Spark): Top 10 Countries by Infection Rate and Mortality rate (Per Million)

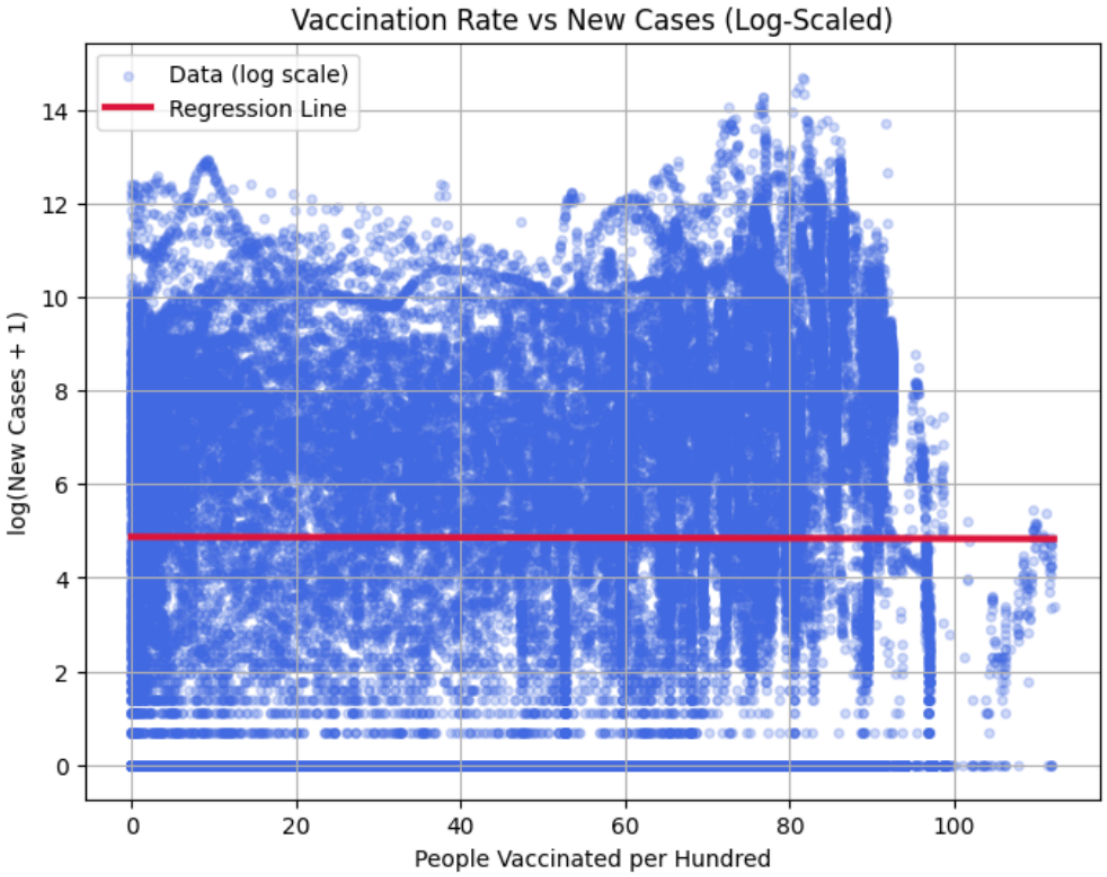
*Left: Cyprus, Austria, and South Korea show the highest infection rates per million, exceeding 600,000–750,000 cases per million.*

*Right: Peru, Bulgaria, and North Macedonia rank at the top in mortality rate per million, with Peru surpassing 6,000 deaths per million—reflecting one of the most severe fatality burdens globally*

### Summary

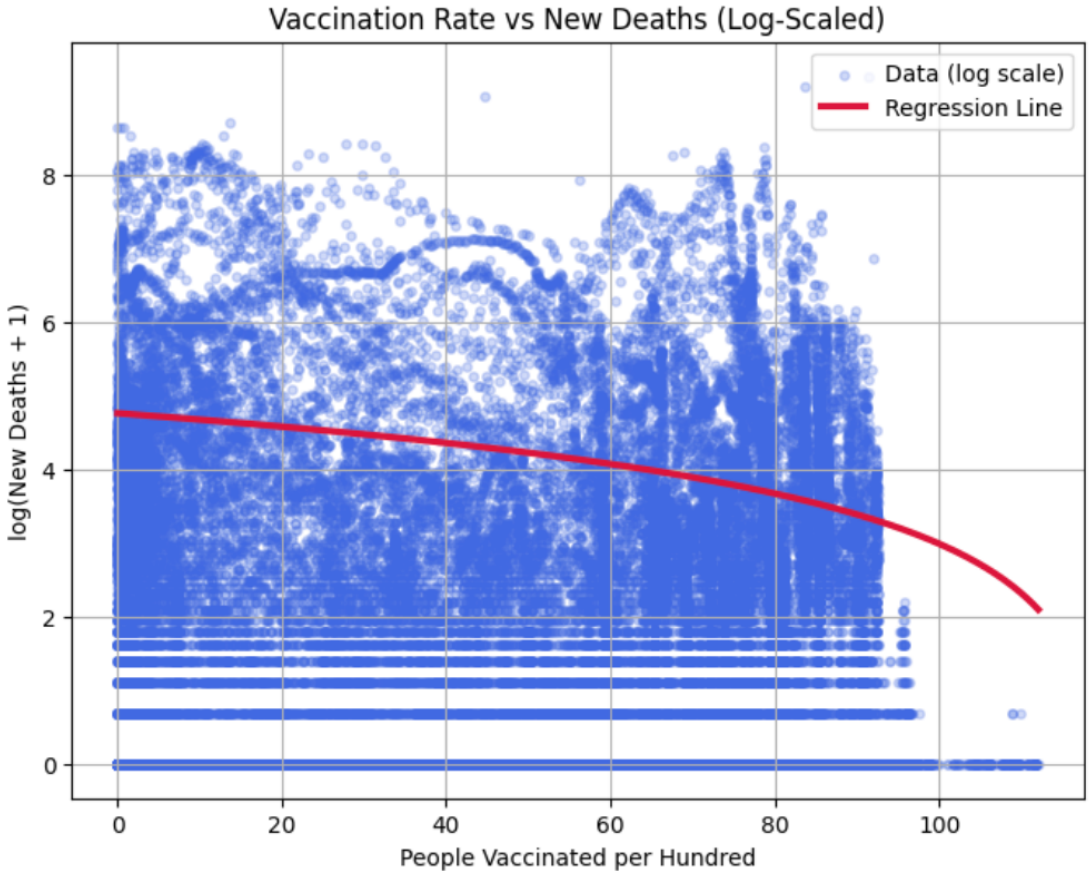
Small, high-testing countries (e.g., Brunei, San Marino) show the highest infection rates overall, while low-resource countries such as Yemen display extreme mortality ratios. When population filters are applied, larger countries like Austria and South Korea emerge as high-burden regions, illustrating both demographic and reporting differences.

## 8.3 RQ3: How do vaccination rates relate to new COVID-19 cases and deaths?



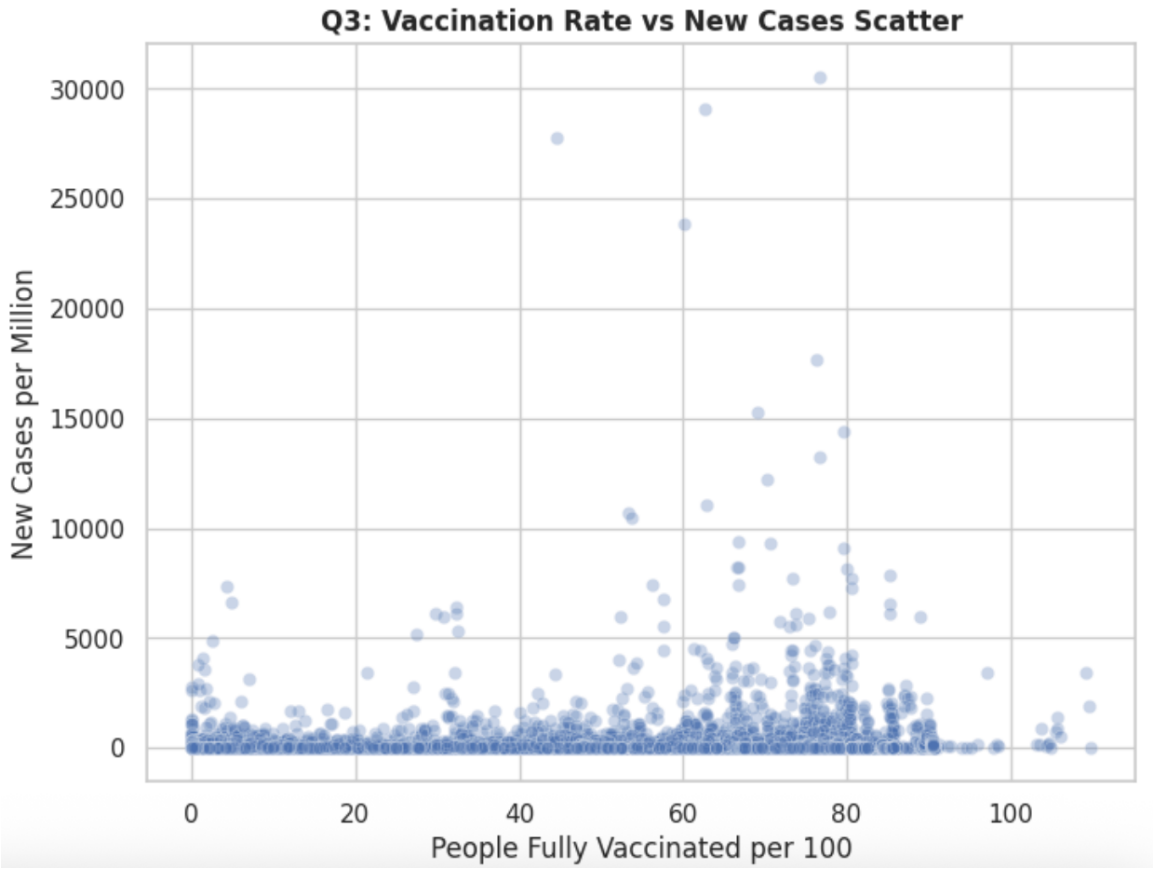
## Figure 8 (Hive): Vaccination Rate vs New Cases (Log Scale)

*New cases remain widely dispersed across vaccination levels, and the nearly flat regression line suggests a weak direct relationship.*




## Figure 9 (Hive): Vaccination Rate vs New Deaths (Log Scale)

*Deaths show a clearer downward trend, with higher vaccination rates generally associated with fewer new deaths.*



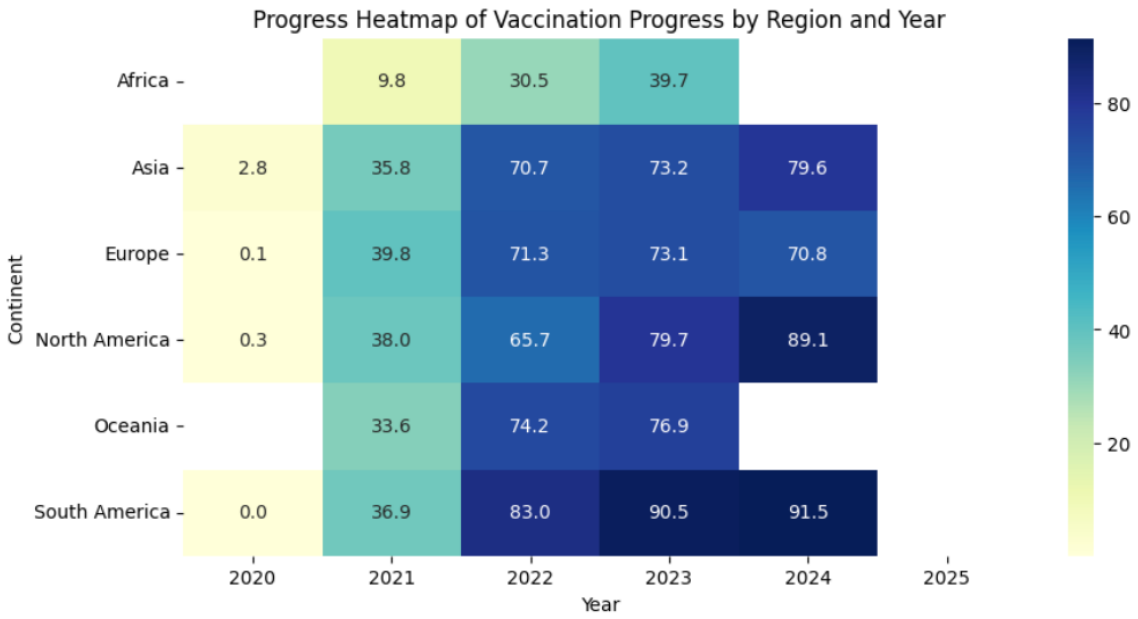

## Figure 10 (Spark): Vaccination Rate vs New Cases (Linear Scale)

*Most high-vaccination countries cluster near the bottom, indicating lower new-case counts compared to low-vaccination regions*.

## Summary

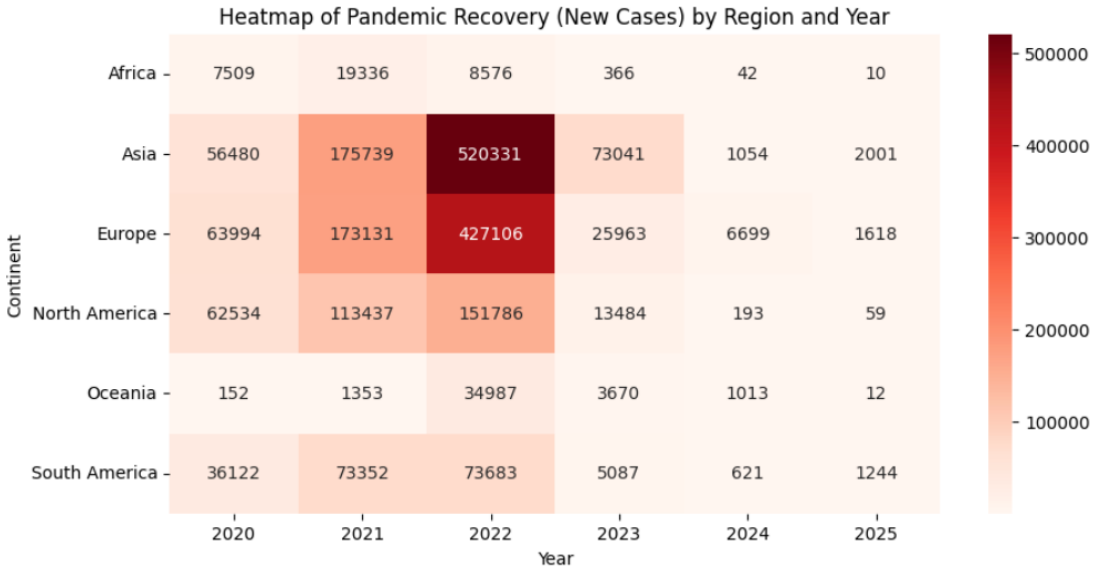
Vaccination levels show limited correlation with new cases but a noticeable negative relationship with new deaths. This indicates that vaccination primarily reduces disease severity rather than preventing all transmission.

## 8.4 RQ4: How do vaccination progress and pandemic recovery differ across regions?



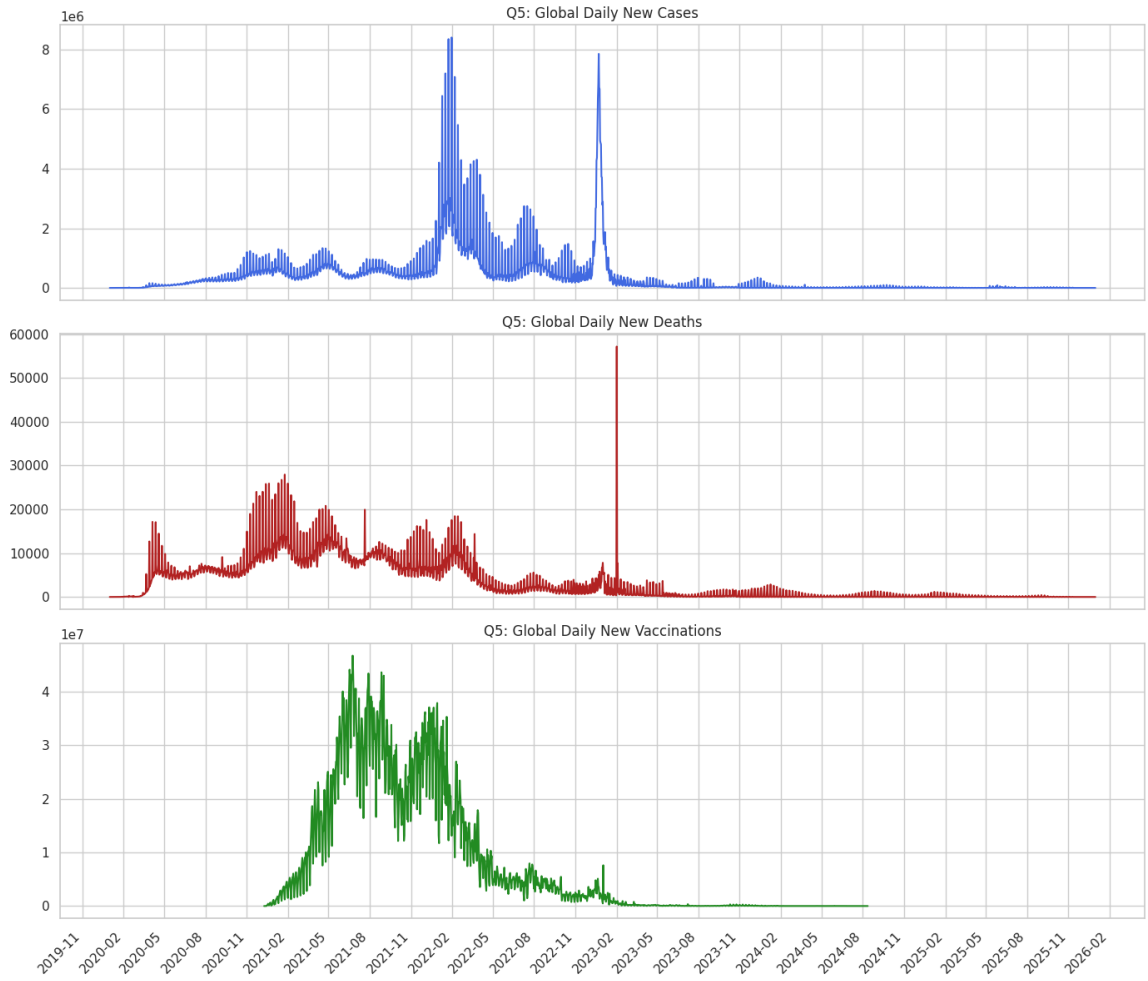

### Figure 11 (Hive): Vaccination Progress Heatmap by Region and Year

*Vaccination levels increase sharply from 2021 onward, with South America, Europe, and North America reaching the highest coverage by 2023–2024.*



## Figure 12 (Hive): Pandemic Recovery Heatmap (New Cases) by Region and Year

*New cases drop dramatically across all regions by 2023–2025, with Africa showing the lowest reported case levels.*




## Figure 13 (Spark): Global Daily New Cases, Deaths, and Vaccinations Over Time

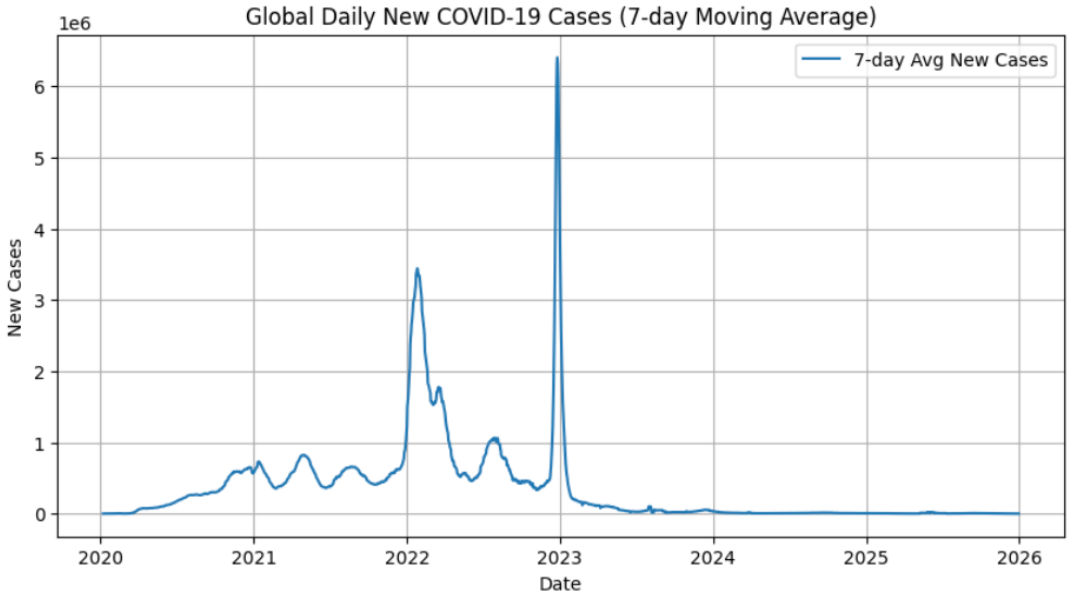
*Global daily new cases and deaths show multiple major waves between 2020 and 2023, with sharp peaks during late 2021–2022 before dropping to near-zero levels by 2024.*

*Daily vaccination counts rise rapidly in 2021, reach a global peak during mid-2021 to early-2022, and then decline steadily as global vaccination campaigns near completion.*

## Summary

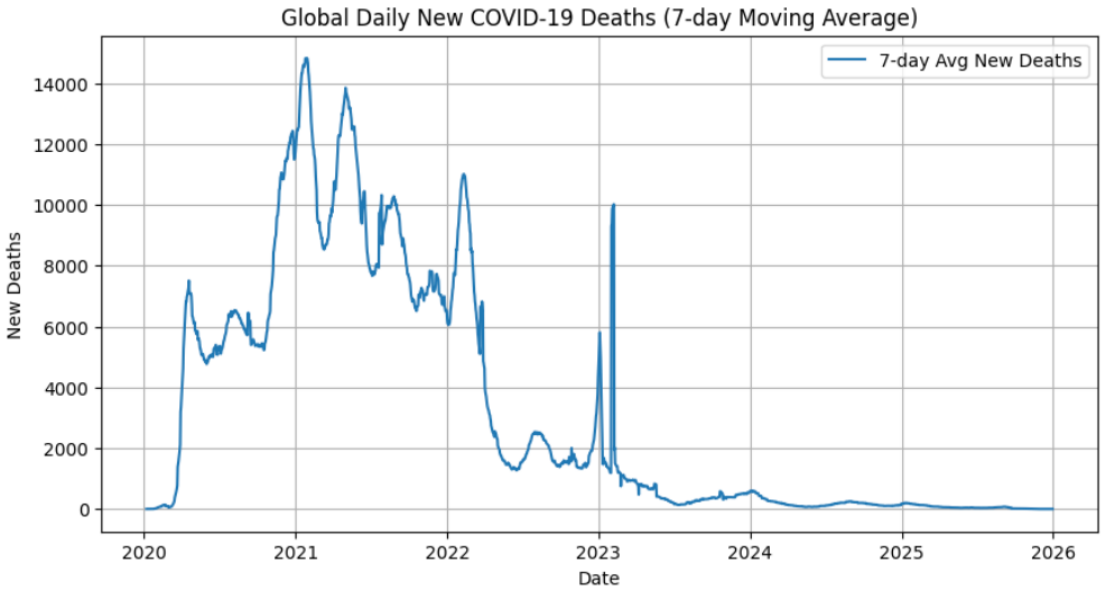
Vaccination progress varies widely across regions, with South America, Europe, and North America leading global rollout. Recovery patterns show sharp case declines after 2022, but Africa's low reported cases are likely influenced by limited testing rather than faster recovery.

## 8.5 RQ5: What do the global time-series patterns of new cases, deaths, and vaccinations look like?



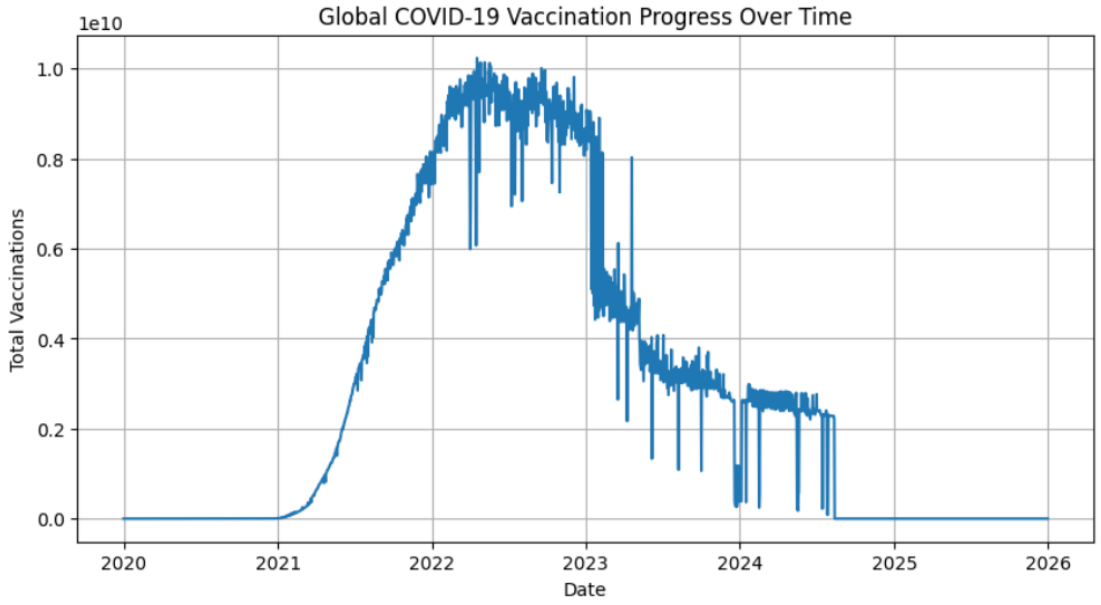

### Figure 14 (Hive): Global Daily New COVID-19 Cases (7-Day Moving Average)

New cases show multiple global waves, peaking sharply in early 2023 before rapidly declining toward near-zero levels.*



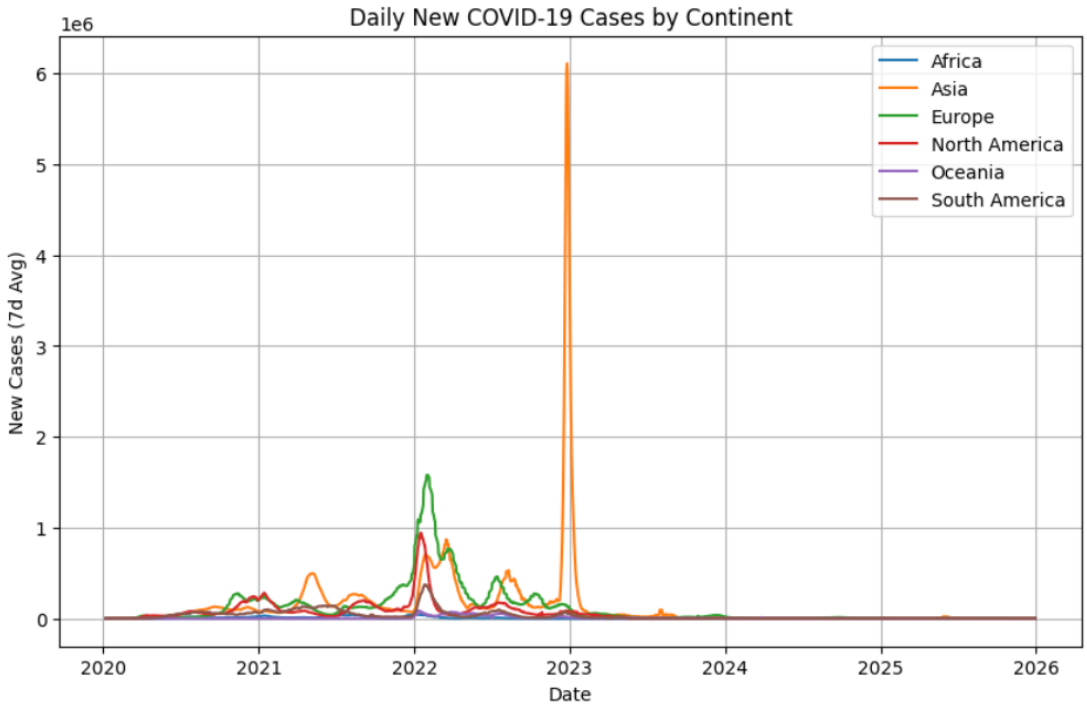

## Figure 15 (Hive): Global Daily New COVID-19 Deaths (7-Day Moving Average)

*Death trends closely follow case waves, reaching high levels during 2021–2022 and decreasing significantly after early 2023.*



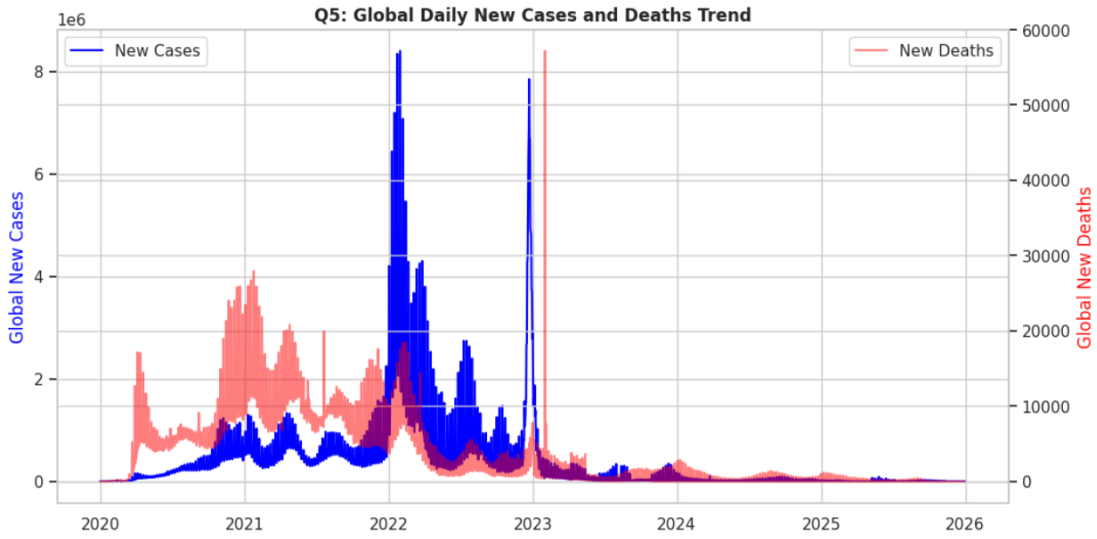

## Figure 16 (Hive): Global COVID-19 Vaccination Progress Over Time

*Vaccinations rise rapidly throughout 2021–2022, plateauing at high levels before gradually declining as global rollout completes.*




## Figure 17 (Hive): Daily New COVID-19 Cases by Continent

*Asia and Europe show the most prominent waves, while Africa and Oceania remain consistently low across the entire timeline.*



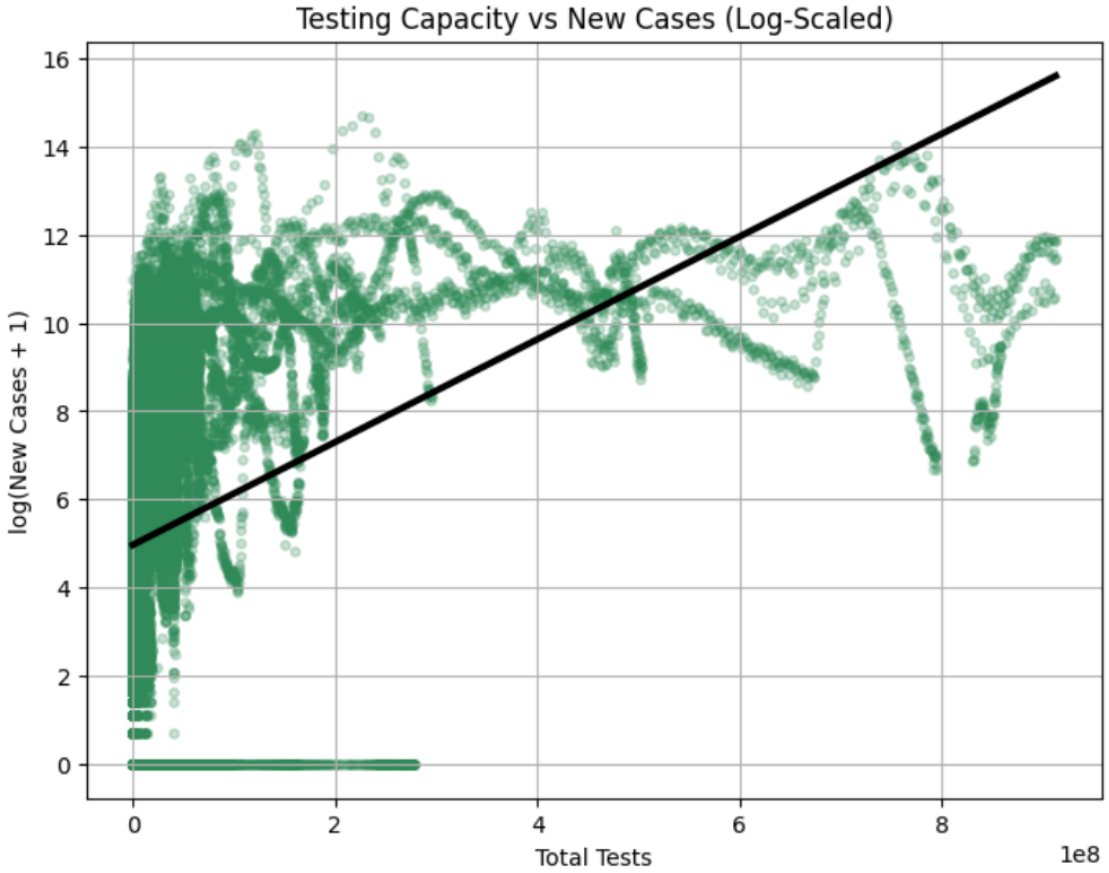

## Figure 18 (Spark): Global Daily New Cases and Deaths Trend

*Spark visualization confirms synchronized peaks in cases and deaths, with both metrics collapsing to minimal levels after 2023.*

### Summary

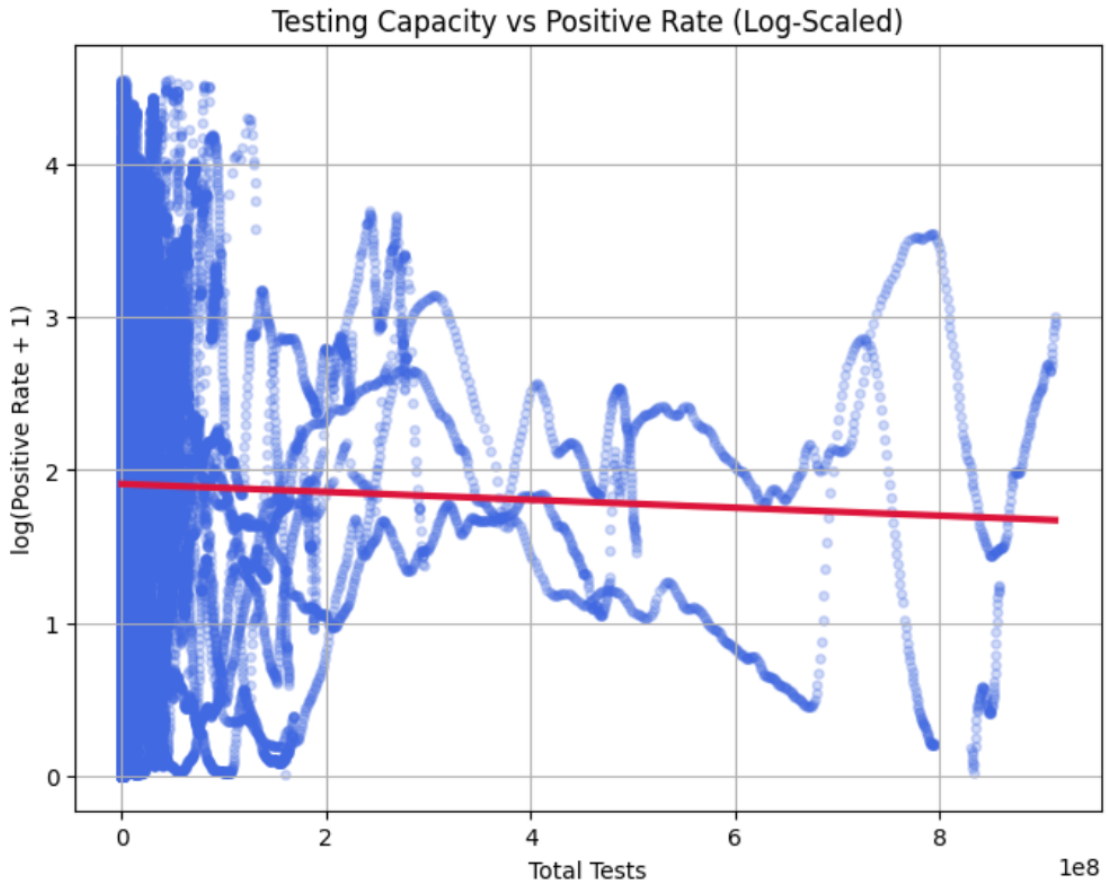
Global time-series patterns reveal several major pandemic waves between 2020 and 2023, followed by a sharp worldwide decline in both cases and deaths. Vaccination rollout aligns with the reduction of severe outcomes, marking a clear turning point in the global pandemic trajectory.

## 8.6 RQ6: How has testing capacity influenced case detection and positivity rates over time?



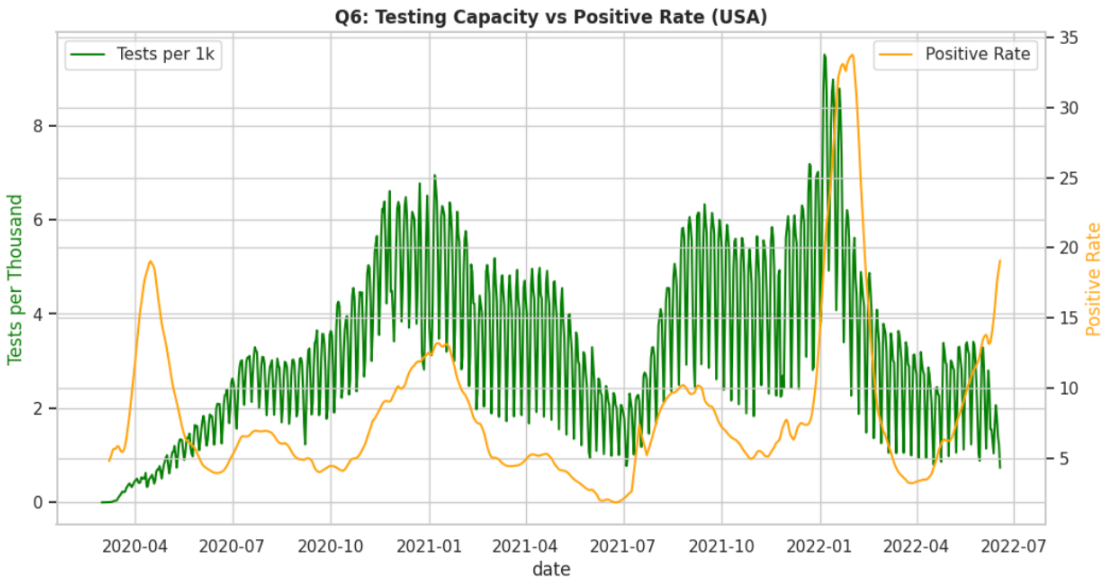

### Figure 19 (Hive): Testing Capacity vs New Cases (Log-Scaled)

*Higher testing volumes are associated with higher detected case counts, as shown by the upward-sloping regression line.*




## Figure 20 (Hive): Testing Capacity vs Positive Rate (Log-Scaled)

*Positive rates generally decrease as total tests increase, indicating expanded testing captures more mild and asymptomatic cases.*



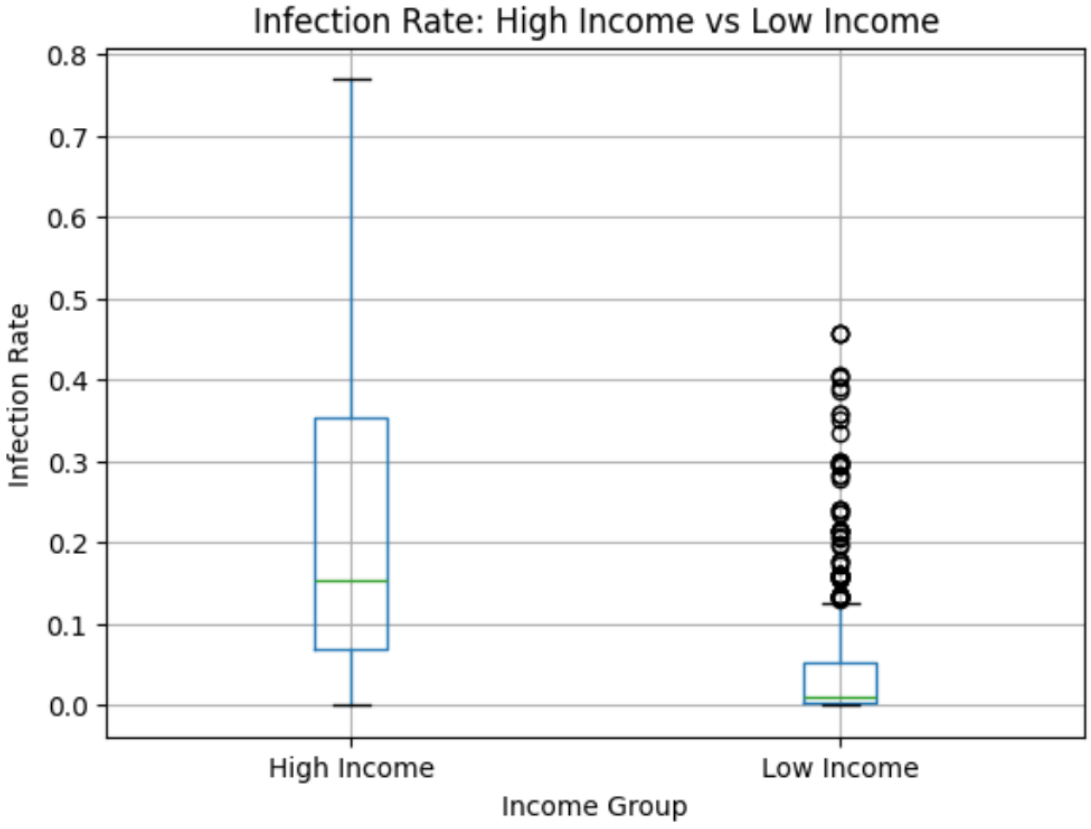
### Figure 21 (Spark): Testing Capacity vs Positive Rate Over Time (USA)

*In the United States, periods of increased testing correspond to noticeable drops in positive rates, especially during major waves.*

### Summary

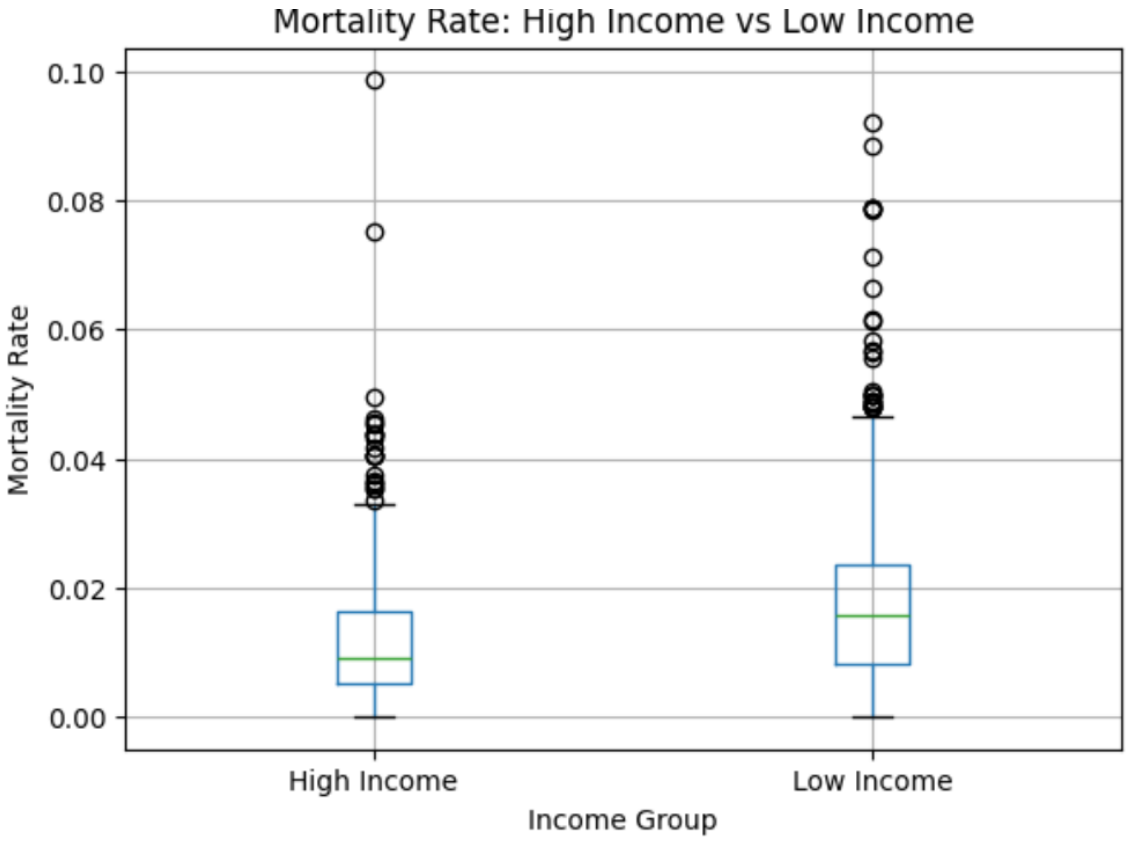
Greater testing capacity leads to more detected cases while reducing the overall positivity rate, reflecting improved detection of mild infections. Spark's time-series view further validates this inverse relationship within a single-country context.

## 8.7 RQ7: What are the differences in COVID-19 outcomes between high-income and low-income countries?



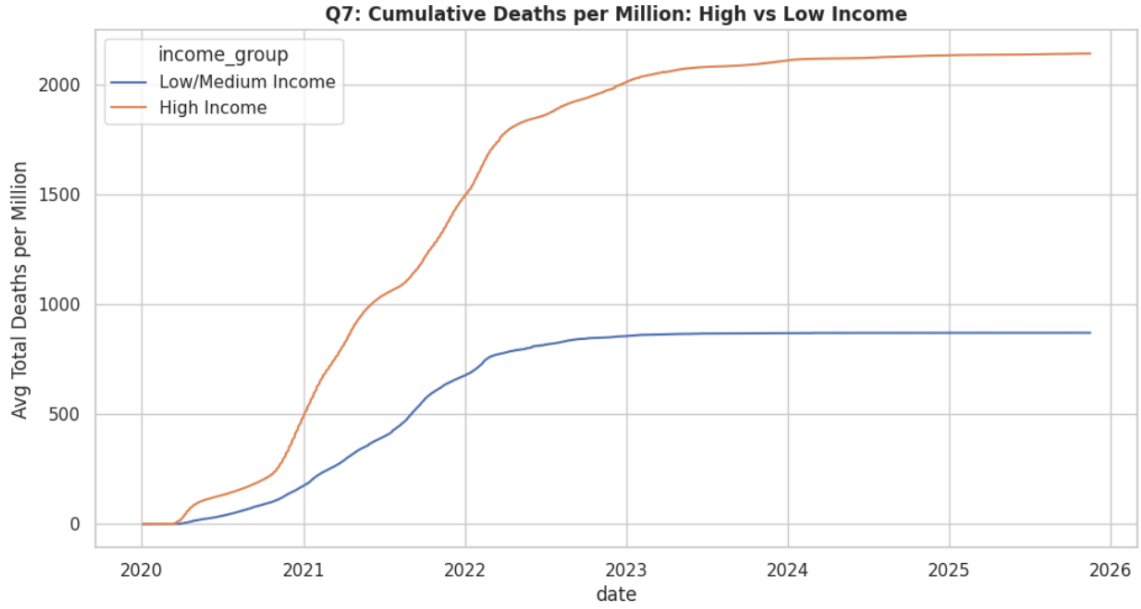

### Figure 24 (Hive): Infection Rate — High Income vs Low Income

*High-income countries show significantly higher infection rates, largely due to extensive testing and more complete reporting.*




## Figure 25 (Hive): Mortality Rate — High Income vs Low Income

*Low-income countries display wider variability in mortality rates, with many outliers indicating inconsistent detection and limited healthcare capacity.*



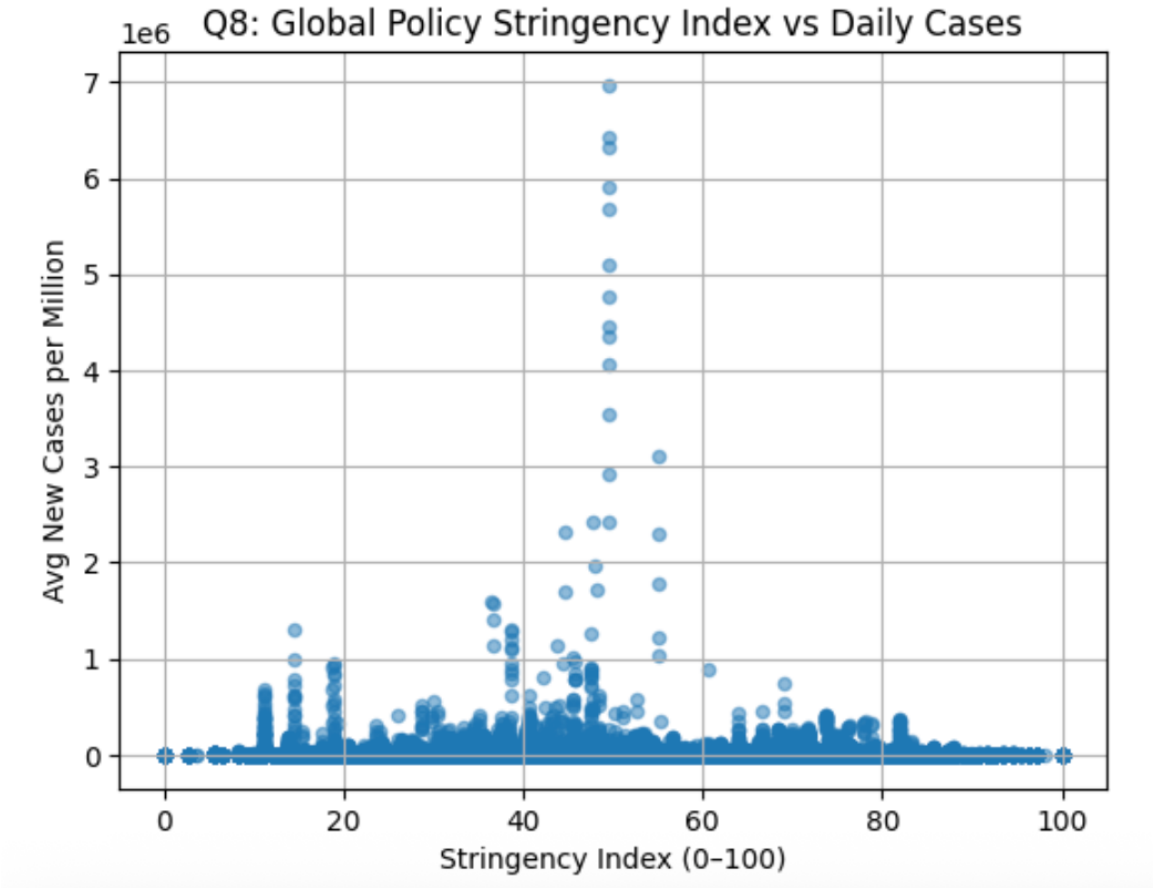
### Figure 26 (Spark): Cumulative Deaths per Million — High vs Low Income

*High-income countries report much higher cumulative deaths per million, reflecting more accurate case recording and older population demographics.*

### Summary

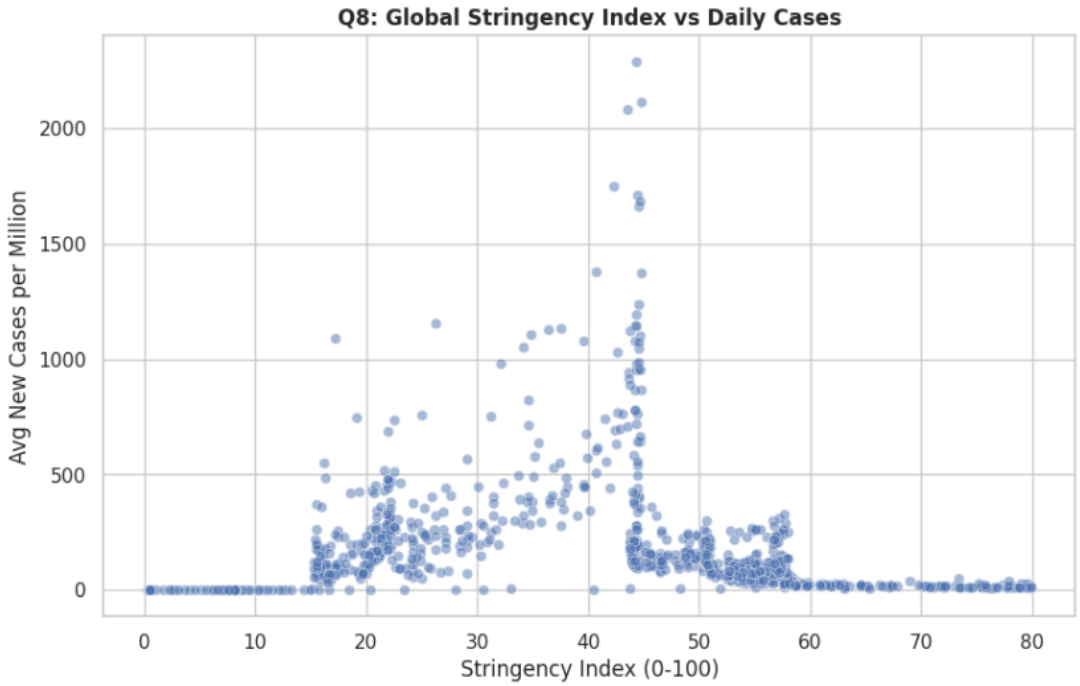
High-income countries show higher reported infection rates due to better testing, while low-income countries exhibit greater mortality variability driven by limited healthcare and underreporting. Overall, outcome differences highlight strong socioeconomic disparities in pandemic impact and detection accuracy.

## 8.8 RQ8: How is policy strictness related to changes in daily COVID-19 cases?




### Figure 27 (Hive): Global Policy Stringency Index vs Daily Cases

*Daily new cases show a wide spread across moderate stringency levels (30–60), while the highest stringency (>80) corresponds to consistently lower case counts.*




### Figure 28 (Spark): Global Stringency Index vs Daily Cases (Filtered)

*Spark visualization confirms that higher policy strictness aligns with reduced case levels, with most high-stringency points clustering near zero.*

### Summary

Across both Hive and Spark views, stronger policy measures tend to correlate with lower daily case counts. The relationship is not perfectly linear, but high stringency consistently appears alongside suppressed transmission levels.

## 9. Technical Challenges

1. **Hive dynamic partition issues** — Inserting data into DWD/DWS tables frequently failed due to partition configuration and inconsistent types.
2. **Large amounts of missing data** — OWID had incomplete vaccination, testing, and policy fields, requiring extra filtering and cleaning.
3. **Spark performance limits** — Window functions and large DataFrames caused slowdowns, so more computation was pushed into Hive.

## 10. Changes in Technology

1. **More processing moved to Hive** — We originally planned to calculate most metrics in Spark, but shifted aggregations and window logic into Hive for efficiency.
2. **Simplified analytical scope** — Some planned predictive or advanced analyses were removed due to data quality and time constraints.
3. **Visualization approach adjusted** — Instead of generating all figures in Spark, some plots were produced using Python for better flexibility.

## 11. Lessons Learned

1. **Data quality matters more than expected** — Missing values, inconsistent reporting, and noisy records had a greater impact on analysis than the tools themselves.
2. **Pre-aggregation improves performance** — Moving heavy computations (rolling averages, lag features) into Hive significantly reduced Spark runtime.
3. **Visualization requires clean data** — Even small formatting issues (dates, nulls, inconsistent columns) can break plots or distort patterns.
4. **Pipeline design evolves during development** — The final ODS → DWD → DWS structure was not fully planned at the start but became essential for clarity and workflow.

## 12. Future Improvements

1. **Incorporate predictive modeling** — Applying time-series forecasting or machine learning could extend descriptive results into actionable predictions.
2. **Build automated ETL workflows** — Scheduling daily or weekly updates with Airflow or Oozie would make the pipeline more production-ready.
3. **Integrate more granular datasets** — Adding mobility, hospitalization, or variant genomic data would deepen analysis and reduce ambiguity.
4. **Improve data validation** — More automated checks for anomalies, outliers, or missing values would make preprocessing more robust.

5. **Enhance dashboard-style visualization** — Creating an interactive dashboard (e.g., Power BI, Tableau, or Plotly) would improve usability for non-technical users.

## 13. References


1. OWID/covid-19-data:

https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

2. COVID-19 Our World in Data (OWID):https://docs.owid.io/projects/etl/api/covid/

3. OWID/covid-19 Raw Data:https://github.com/owid/covid-19-data/tree/master/public/data