Big Data Project Proposal

COVID-19 Big Data Analytics and Visualization

CS - GY 6513-C Big Data
Fall 2025

Authors:

Weiyi Luo (WL3398)

Jia Yang (JY5081)

Geng Niu (GN2279)

Abstract

As part of this project, we intend to analyze the global COVID-19 pandemic using publicly available datasets provided by *Our World in Data.* The report presents detailed steps of data acquisition, preprocessing, hypothesis testing, regression and cluster analysis, and the interpretation of the results. The subsequent sections discuss the problem statement, analytical approach, and key findings, supported by visualizations and quantitative summaries for better clarity and understanding.

Problem Statement and Objectives

People all over the world have been suffering from the covid epidemic since it started. Every year untold numbers of people die under the threat of the epidemic and it has become a problem that we have to take seriously. Despite the availability of rich global datasets, there remains a need for a unified framework to process, visualize, and interpret COVID-19 indicators effectively. Moreover, variations in vaccination rollout, testing rates, and healthcare capacity across regions makes it difficult to evaluate and compare the policies, which leads to challenges in understanding the relationships between measures and outcomes.

This project solves those problems by analyzing the OWID COVID-19 dataset to identify global and regional patterns, quantify correlations among key variables, and help people better understand the situation so that they could respond to it in a positive way.

Inspired by this scenario, we intend to create an analysis to answer

1. How have global COVID-19 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 the strongest pandemic recovery trends?
5. What does the time-series pattern of daily new cases, deaths, and vaccinations look like?
6. How have testing capacities affected case detection and positivity rates over the years?
7. What are the differences in COVID-19 response effectiveness between high-income and low-income countries?
8. Can we visualize and quantify the relationship between government interventions and pandemic control outcomes?

Data Source

Name: Our World in Data – COVID-19 Dataset

Link: https://github.com/owid/covid-19-data/tree/master/public/data

License: Creative Commons BY 4.0 (open access)

Maintained by: Our World in Data (University of Oxford)

1. Overview

The dataset we selected comes from Our World in Data (OWID)—an open-access research initiative that has continuously collected and standardized global COVID-19 data throughout the pandemic. This dataset is both extensive and deep: it integrates health, policy, and demographic information from hundreds of countries into a unified framework.

Rather than focusing on individual countries or metrics, OWID continuously aggregates daily updates from the World Health Organization (WHO), national health agencies, and other sources. This enables us to examine regional variations in pandemic evolution and how government actions, testing capacity, and vaccination rates ultimately shape outcomes.

2. File Information

- **File Name:** owid-covid-data.csv
- **File Size:** ~75 MB (compressed CSV format)
- **Number of Records:** approximately 2.1 million rows and 67 columns
- **Granularity:** Each record corresponds to a unique combination of location (country or region) and date.

3. Dataset Composition

The dataset consolidates multiple thematic data sources into a single table. The key variable categories include:

Category	Example Fields	Description
Geographic & Demographic Info	iso_code, continent, location, population, median_age, gdp_per_capita, life_expectancy	Identifiers and background data for each region.

Epidemiological Indicators	total_cases, new_cases, total_deaths, new_deaths, reproduction_rate	Core COVID-19 case and death statistics, derived from WHO and national health reports.
Testing Data	total_tests, new_tests, positive_rate, tests_per_case	Data on the scale and efficiency of national testing programs.
Vaccination Data	total_vaccinations, people_vaccinated, people_fully_vaccinated, total_boosters	Records from national vaccination reports and WHO updates.
Hospitalization & ICU Data	hosp_patients, icu_patients, weekly_hosp_admissions	Hospital load and critical care statistics from national health agencies.
Excess Mortality	excess_mortality, excess_mortality_cumulative	Measures of deaths above baseline expectations, sourced from Human Mortality Database.
Government Response & Mobility	stringency_index, google_mobility, oxcgrt_policy	Quantitative measures of government interventions, lockdowns, and mobility patterns.
Socioeconomic Indicators	human_development_index, extreme_poverty, hospital_beds_per_thousand	Contextual factors affecting pandemic outcomes.

4. Data Structure

- Primary Keys: iso_code, date
- Temporal Coverage: January 1, 2020 – August 19, 2024
- Frequency: Daily (aggregated from official national and international sources)

- Data Format: CSV (UTF-8 encoded, comma-delimited)

5. Classification and Integration

The dataset can be conceptually divided into the following modules:

1. Epidemiological data – confirmed cases and deaths.
2. Healthcare system data – testing, hospital, ICU capacity.
3. Vaccination data – dose coverage by population share.
4. Policy and behavioral data – government responses, mobility.
5. Demographic and economic indicators – background context for cross-country comparison.

These categories allow integration and correlation analysis across dimensions, such as vaccination rates versus mortality, or government stringency versus case growth.

6. Relevance for Big Data Analysis

Since this dataset combines millions of time-based records and indicators from various fields, it’s a good example for using big data technologies like Spark, Hive, and DWD. The structure makes it easy to handle in a distributed system and useful for analyzing patterns, comparing results, and understanding policy impacts over time.

Methodology and Technology

The project will use a big data processing pipeline built on the Hadoop and Hive ecosystem to analyze COVID-19 data at scale. The methodology follows a multi-layer data warehouse design from ODS to DWD to DWS to ADS to ensure data consistency, scalability, and efficient analytics.

1. Data Ingestion and Storage

COVID-19 datasets will be downloaded from Our World in Data (OWID) and healthdata.gov in CSV format.

The data will be uploaded to the Hadoop Distributed File System (HDFS) for distributed storage.

External tables will be created and data will be loaded in Hive to access the raw data directly from HDFS (in the Operational Data Store (ODS) layer). When importing data into Hive, it will be partitioned by continent, year, and month to improve query performance.

2. Data Cleaning and Transformation

Data will be processed using Spark SQL and HiveQL to handle missing values, format inconsistencies, and duplicate entries.

Columns and all data will be standardized (like date formats, location names).

Cleaned datasets will be stored in the Data Warehouse Detail (DWD) layer.

3. Aggregation and Analytical Processing

The Spark framework will be used for distributed computations and analytical queries.

Aggregation tasks include:

1. Daily and weekly case/death trends by country.
2. Rolling-average infection growth rates.
3. Correlation analysis between vaccination rates and infection trends.
4. Medical system stress analysis by ICU patient ratio vs. medical bed density.
5. Population and mortality analysis by age + chronic disease rate vs mortality rate.
6. etc.

Aggregated data will be written into the Data Warehouse Summary (DWS) layer, ready for reporting and visualization.

4. Visualization and Reporting

The Application Detail Summary (ADS) layer will generate summary tables for visualization in Matplotlib using Python.
Dashboards will display:

1. Total case of affected (country/regional-wise)
2. Vaccination progress (overall and country/regional-wise)
3. Infection trends over time (overall and country/regional-wise)
4. Healthcare Pressure Heatmap (country/regional-wise)
5. etc.

5. Tools and Frameworks

Category	**Technology**
Programming Language	Python in Jupyter Notebook
Big data Framework	Hadoop, Apache Spark
Data Warehouse	Hive
Data Visualization	Matplotlib, Pyecharts

Environment	NYU Dataproc

Reference

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