Covid19 data anaysis

2025-02-25

COVID analysis

This document describes the work done on COVID cases dataset for CU Boulder MDS 2025.

0. Structure of the document

- **Introduction**
- **Research Question**
- **Methodology Description**
- **Methodology Implementation**
- **Conclusions**
- **Bias Report**

1. Introduction

COVID pandemic was a global challenge that put to the test the scientific community around the world. Although is not the first pandemic in the world, the globalization made the population to move around the globe several times a day, helping the propagation of viruses and increasing mutation of the virus, making it more problematic to fight. Is not the first time that there are curfews, quarentines and not even the first time the humanity developed a vaccine in record time. During this pandemic more tan 700 million people got infected with 7 million dead worldwide. "https://www.worldometers.info/coronavirus/" This is mortality rate of 1% (reported), bigger than Seasonal Influenza (flu) which is 0.1%, which is around 500.000 dead depending on the season. Even being a tragedy, our perception is that COVID was extremely dangerous and most infected people died, this 1% looks small. Therefore, the question arises, how deadly was the COVID-19.

2. Research question

To address this question, a more formal one need to be stated in order to perform a quantitative research: "Is the infection rate, meaning, the speed of virus spread (virosity), a good estimator of death rate?"

3. Methodology description

- Analyze the evolution of quantity of cases (trajectory) per country around the world
- Correlate this trajectory with the quantity of deaths per country.
- Data will be normalized by population
- Create a model to estimate the number of deaths based on number of cases.

4. Methodology implementation

Datasources Data comes from following repository: "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/"

Files are: - **Global confirmed cases** "time_series_covid19_confirmed_global.csv"

- **Global deaths**: "time_series_covid19_deaths_global.csv"
- **Global population**: "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv" #### Data loading

- Defines the external data sources used for the analysis.
- Loads worldwide COVID-19 case and death data from an online repository.
- Reads population data to allow comparisons between countries.

DATA PRE-PROCESSING Convert from wide to long format

- Converts data from a format where each day is a separate column into a format where each row represents a date.
- Ensures that cases and deaths are structured correctly for analysis.

COMPUTE INFECTED PER MILLION

- Joins COVID case and death data with population data to calculate how many people per million were infected in each country.
- Filters out missing values to keep only reliable data.

INSIGHTS

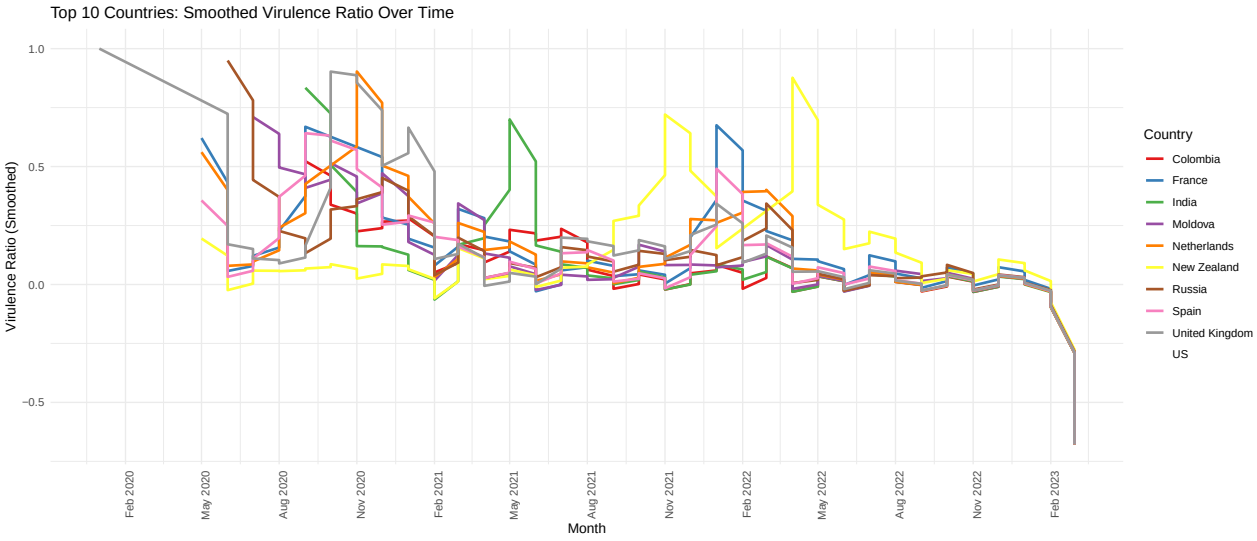
- Defines a formula to measure how fast the virus spreads over time.

$$\text{Spread Rate} = \frac{\text{Final Cases}}{\text{Average Cases Per Month}}$$

- Calculates the total number of cases per country and the average number of cases per month. **Computing Monthly Cases and Deaths**
- Groups cases and deaths by month instead of daily counts. This helps smooth out daily fluctuations and provides a clearer trend.
- Computing Virulence Ratio – Compares the number of cases each month to the previous month to measure how fast the infection rate is changing. – Normalizes this ratio to avoid extreme values that might distort the analysis.

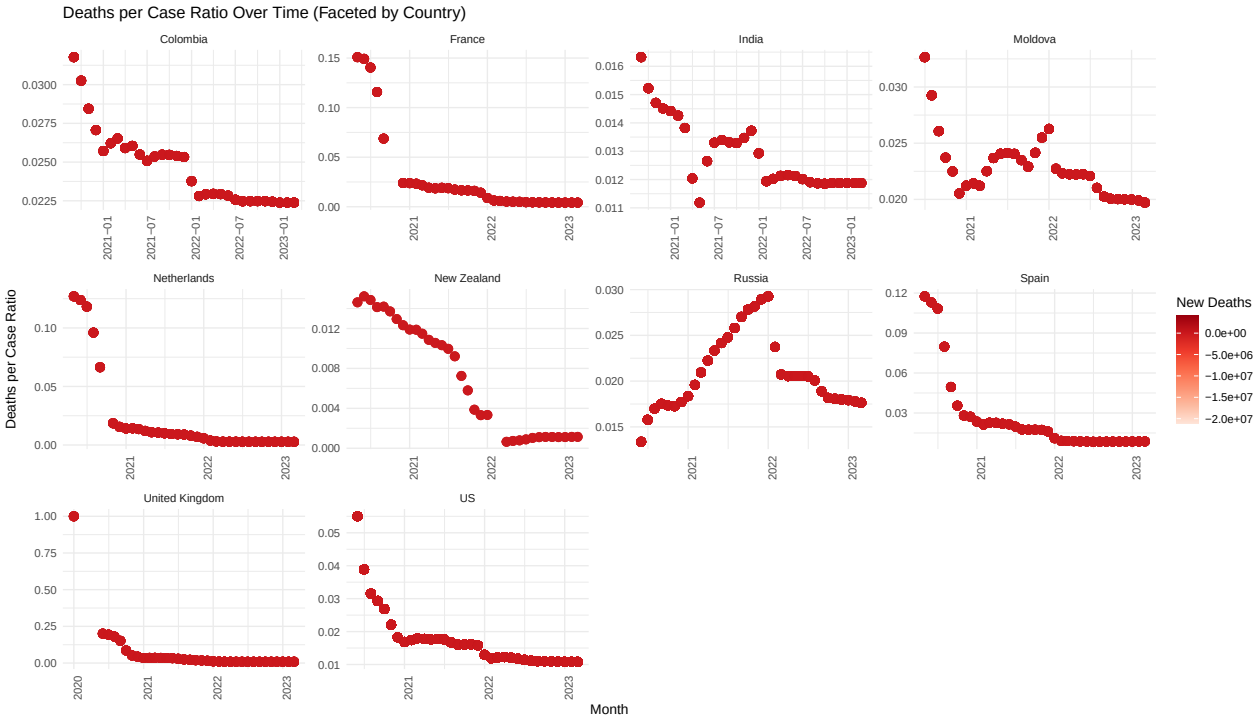
VISUALIZE VIRUS SPREAD TRAJECTORIES

- Selects the 10 most affected countries based on total cases per million.
- Creates a smoothed trend line to show how quickly the virus spread in these top 10 countries.



PLOT THE DEATHS TRAJECTORY FOR THE TOP 10 COUNTRIES

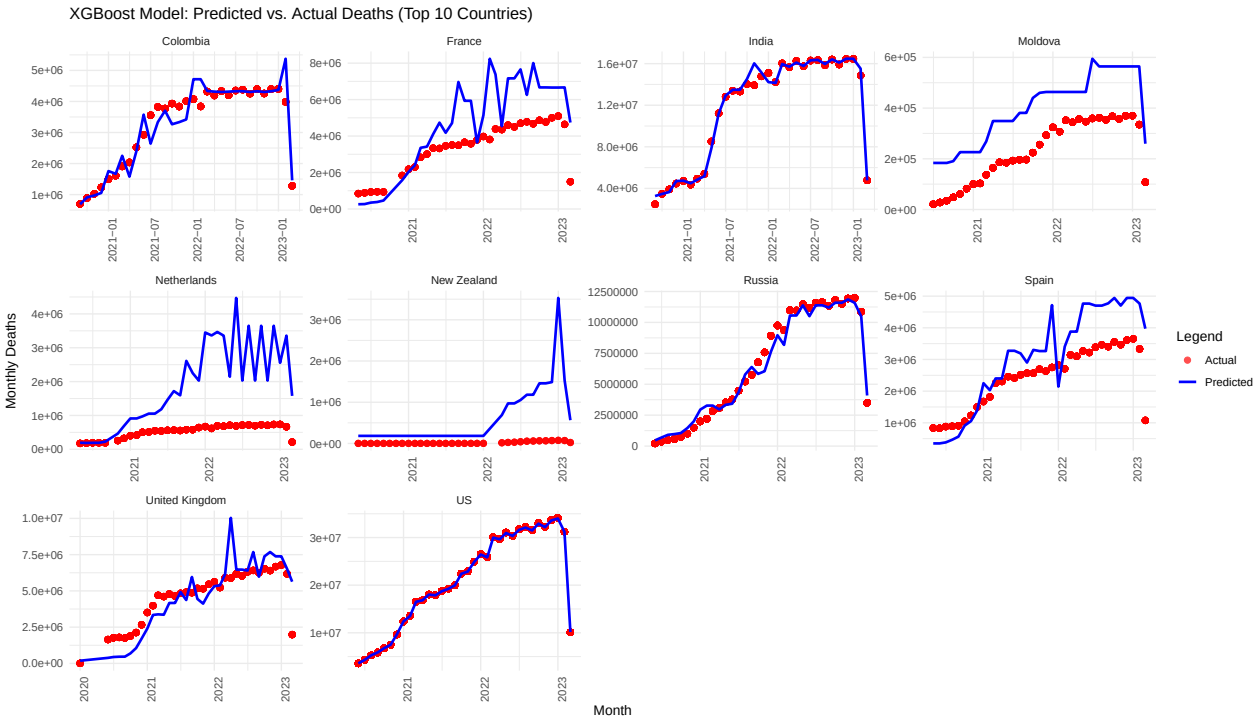
- Computes the ratio of deaths to cases over time.
- Colors the points based on the number of deaths per month.
- Uses separate small charts (facets) for each country to make comparisons easier.



MODELIZATION

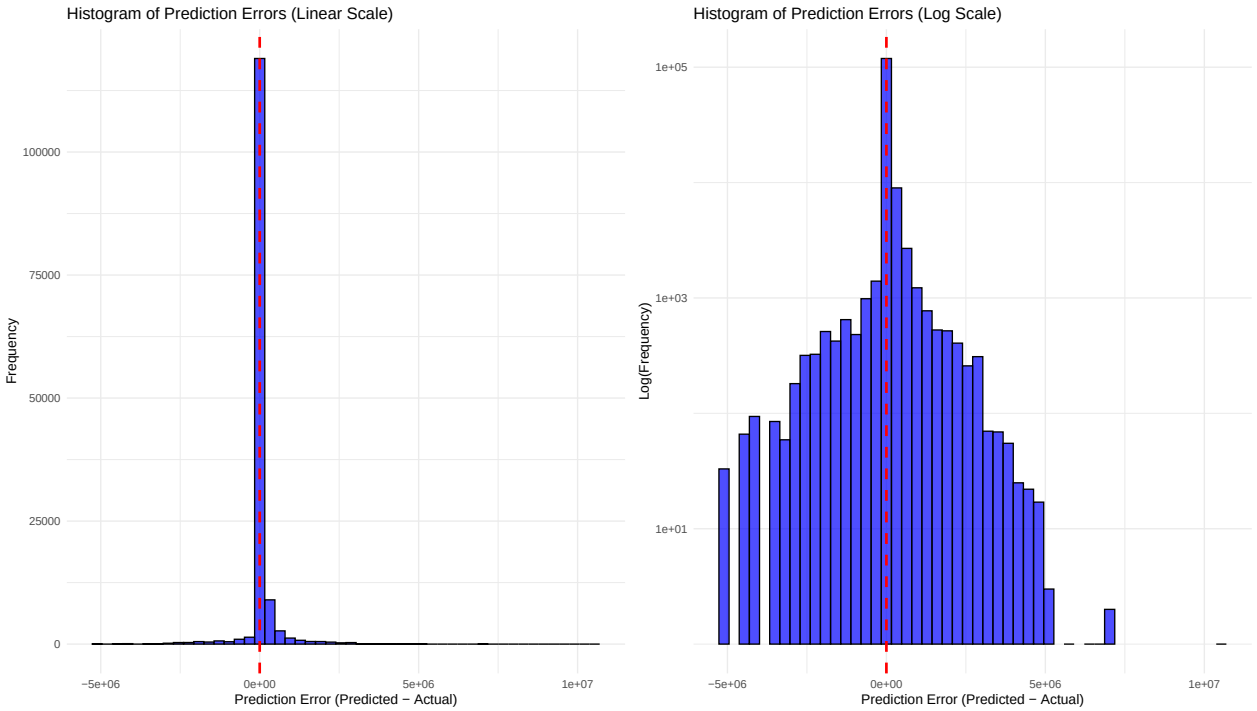
- Prepares the data for machine learning by normalizing values.
- Trains an XGBoost model (a powerful predictive algorithm) to estimate deaths based on new cases.
- Predicts deaths for each country and compares them to actual recorded deaths.
- Shows the top ten countries to observe the model behavior.

For solving the clear not linear correlation between deaths and virulence, an **XGBoost** algorithm will be used to allow different countries to be modeled differently (taking advantage of how decision trees work).



MODEL EVALUATION

- Calculates the difference (error) between predicted and actual deaths.
- Creates histograms to visualize how well the model performed.
- One histogram shows errors in normal scale, while the other shows errors using a log scale to highlight small deviations.



5. CONCLUSION

This study analyzed the relationship between COVID-19 infection rates and mortality rates across different countries. By leveraging global datasets on confirmed cases, deaths, and population sizes, we investigated whether the speed of viral spread (virulence) serves as a reliable predictor of mortality.

Key Findings

1. Virus Spread and Mortality:
   - The infection trajectory varied significantly across countries, with some experiencing rapid early outbreaks, while others showed slower but sustained case increases.
   - The correlation between infection rate and deaths was not linear, highlighting the complexity of pandemic dynamics.
2. Machine Learning Model:
   - An XGBoost model was trained to predict monthly deaths based on the number of new cases.
   - The model performed well for some countries, capturing trends accurately, but struggled with others, indicating that additional factors (e.g., healthcare infrastructure, policy interventions) influence outcomes.
3. Model Limitations & Biases:
   - Data Quality Issues: Some countries underreported cases and deaths due to limited testing or political reasons.
   - Normalization Challenges: Adjusting deaths per million people assumes uniform exposure, which might not always hold.
   - Time Lag Effect: The direct correlation between new cases and deaths might be misleading, as fatalities typically lag infections by several weeks.

Final Thoughts While the study provides insights into COVID-19 spread and mortality, further refinements—such as incorporating time-lagged features and healthcare-related variables—could improve predictive accuracy. Future research should explore alternative modeling techniques and validate results using additional epidemiological data.