Covid-19

Data Visualization

Vita Lanko

v.lanko@student.vu.nl

Section 1: Introduction

Dataset Information

The dataset I have chosen is the "COVID-19 Global Pandemic Data."

This dataset is maintained by the World Health Organization (WHO) and various governmental health agencies around the world.

The dataset has been continually updated since the start of the COVID-19 pandemic in December 2019.

Section 2: Dataset Overview

Key Features

Field name	Description
Date_reported	Date of reporting to WHO
Country	Country, territory, area
New_cases	New confirmed cases. Calculated by subtracting previous cumulative case count from current cumulative cases count.
Cumulative_cases	Cumulative confirmed cases reported to WHO to date.
New_deaths	New confirmed deaths. Calculated by subtracting previous cumulative deaths from current cumulative deaths.
Cumulative_deaths	Cumulative confirmed deaths reported to WHO to date.

Features used from additional vaccination dataset:

COUNTRY	Country, territory, area
TOTAL_VACCINATIONS	Cumulative total vaccine doses administered
PERSONS_VACCINATED_1PLUS_DOSE	Cumulative number of persons vaccinated with at least one dose
PERSONS_LAST_DOSE	Cumulative number of persons vaccinated with a complete primary series
PERSONS_BOOSTER_ADD_DOSE	Cumulative number of persons vaccinated with at least one booster or additional dose

Data Quality

- The dataset is regularly updated and is considered complete.
- It contains global data, broken down by country and sometimes by region.
- All values are recorded in consistent units (e.g., cases, deaths).
- No significant quality problems are present in the dataset. * Negative values in cases and deaths can sometimes appear when a country sends a correction to the ECDC, because it has overestimated the number of cases/deaths.

Key Information

Key Feature	Minimum Value	Maximum Value	Average Value	Most Common Value
Cumulative Cases	0	103 436 829	1 464 717	0
Cumulative Deaths	0	1 127 152	18 002	0
New Cases	-8 261	6 966 046	2 395	0
New Deaths	-43 206	43 206	22	0
Date reported	2020-01-03	2023-09-21	NaN	NaN

Key Feature	Minimum Value	Maximum Value	Average Value	Most Common Value
TOTAL_VACCINATIONS	117	3 516 881 000	59 232 850	NaN
PERSONS_VACCINATED_1PLUS_DOSE	0	1 318 027 000	24 420 416	NaN
PERSONS_LAST_DOSE	0	1 284 480 000	22 499 493	NaN
PERSONS_BOOSTER_ADD_DOSE	0	834 060 100	11 644 230	0

Section 3: Charts

In this section, we will present five charts that show relationships between different features in the COVID-19 dataset. Each chart will provide insights into potential correlations and patterns related to the pandemic.

Chart 1: Relationship between Total Confirmed Cases and Total Deaths

This scatterplot illustrates the relationship between cumulative COVID-19 cases and cumulative deaths in different countries. A positive correlation suggests higher cases lead to more deaths. We can see that depending on the country we have different correlations which suggests that the spread of COVID-19 and related statistics vary.
Possible Causes/Consequences: Higher case numbers may result from a lack of preventative measures or the presence of more infectious variants.

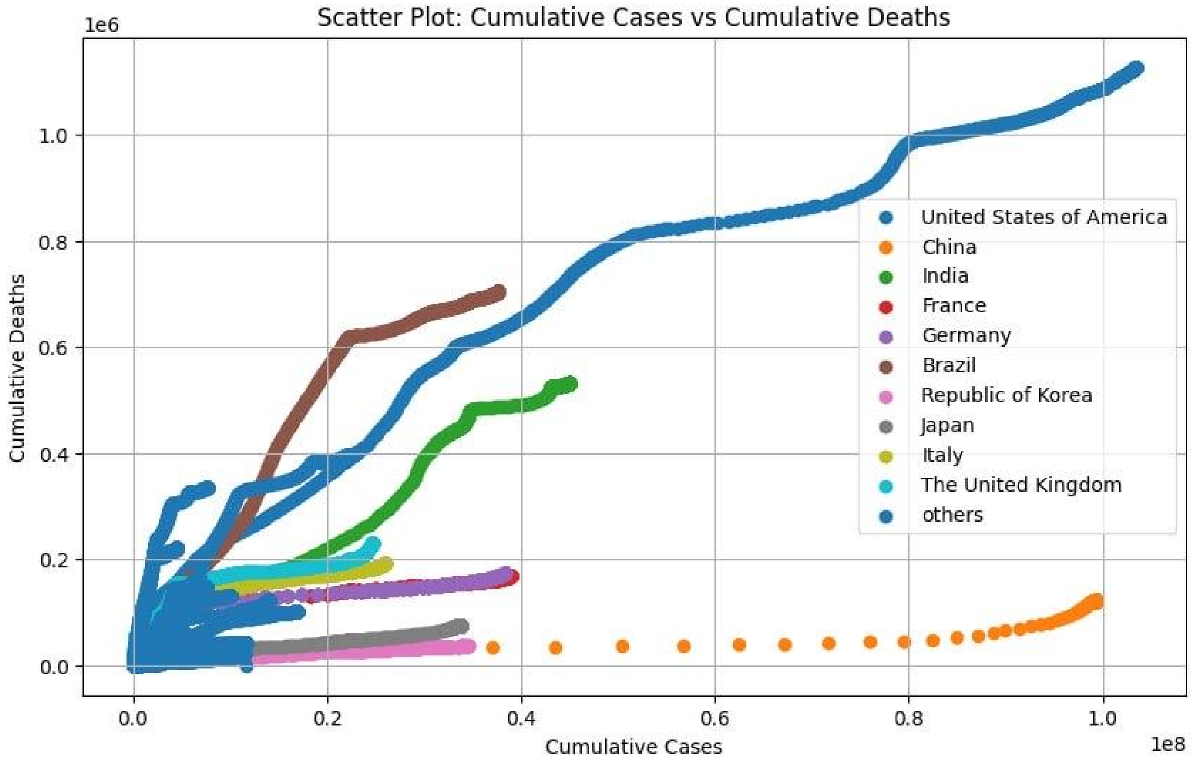


Chart 2: Relationship between New Cases and New Deaths

This scatterplot shows the relationship between the number of new COVID-19 cases and the number of new deaths, providing insights into the immediate impact of the virus.
Possible Causes/Consequences: An increase in new cases may lead to a surge in new deaths. Understanding this connection can aid in healthcare preparedness and response planning. But we can observe that this graph does not go up or down so we can conclude that currently there is no positive neither negative correlation between new cases and new deaths.

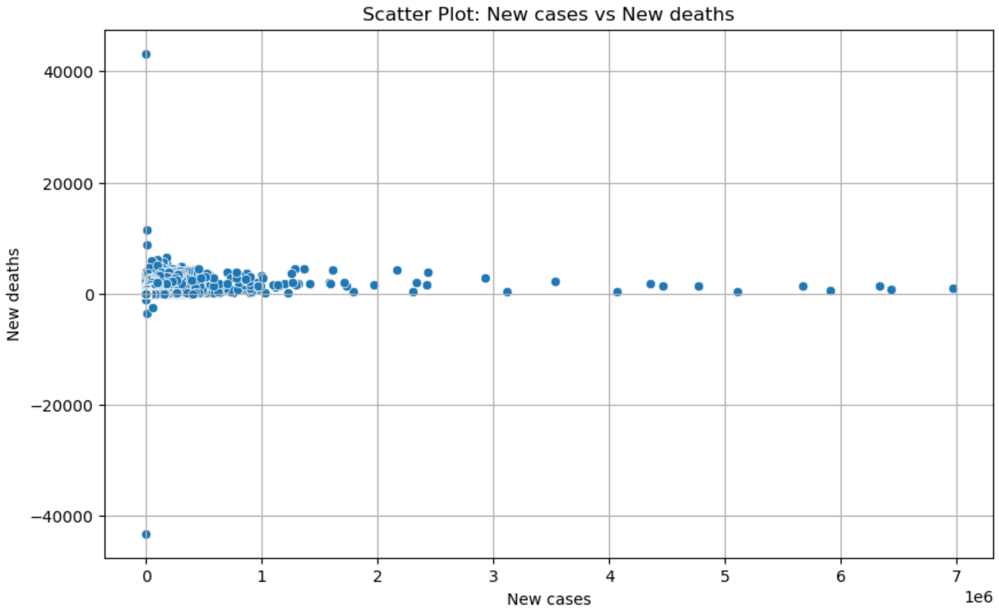


Chart 3: Top 5 countries with the highest COVID-19 vaccinations numbers
This bar chart provides a quick overview of which countries have administered the most vaccine doses. This information is crucial for tracking global vaccination progress and identifying regions where vaccination campaigns are most successful. We used the dataset for vaccinations to show this plot; it is described in the additional datasets section.
Possible Causes/Consequences: Countries with high vaccination numbers may have robust healthcare infrastructure, efficient vaccination distribution, or successful public health campaigns. Conversely, low vaccination numbers may indicate challenges in vaccine access or vaccine hesitancy.

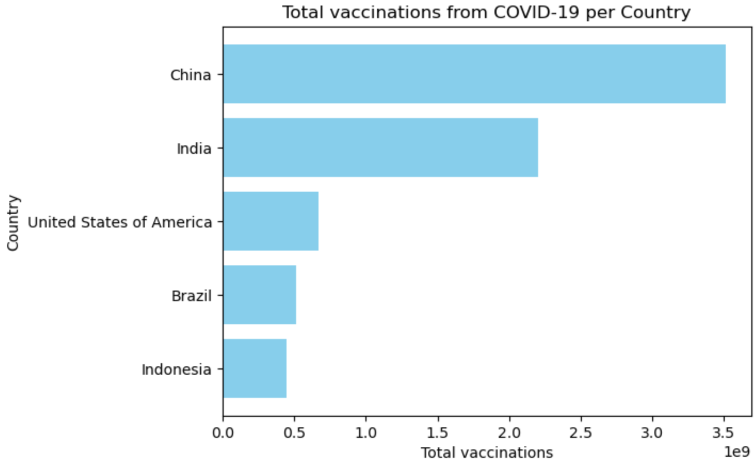


Chart 4: Top 5 countries with the highest cumulative COVID-19 cases
The bar plot provides a concise and visually informative representation of the pandemic's impact on different countries. Healthcare professionals and researchers can use this plot to assess the severity of the pandemic in different regions and make informed decisions regarding resource allocation, travel restrictions, and public health interventions.
Possible Causes/Consequences: Countries with high cumulative COVID-19 cases may have larger populations, higher population density, or delayed implementation of public health measures. Understanding which countries are most affected is essential for targeted response efforts.

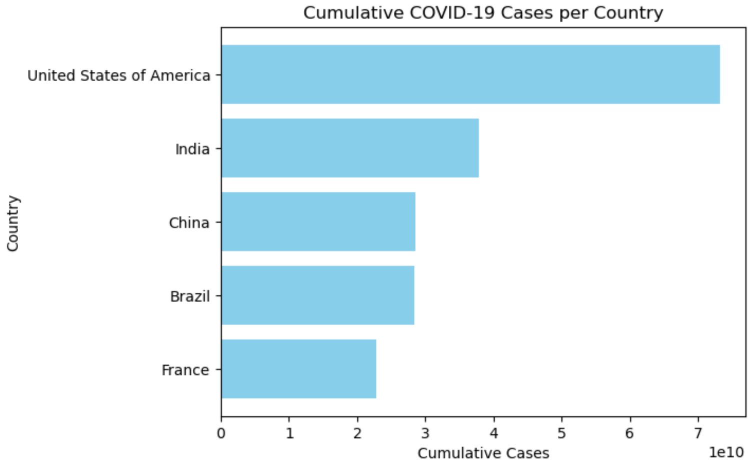
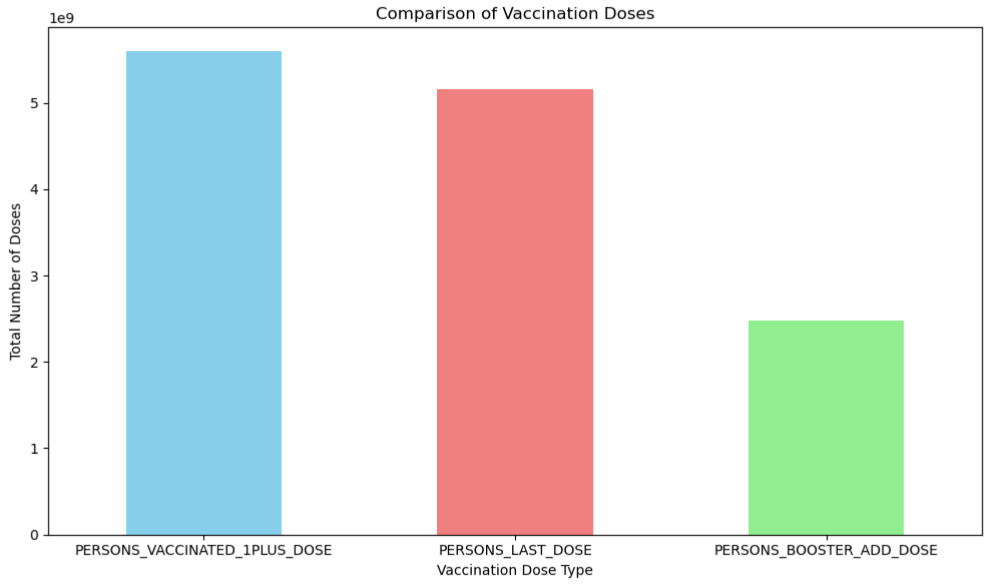


Chart 5: People vaccinated with at least one dose vs vaccinated with a complete primary series vs vaccinated with at least one booster or additional dose

This bar chart provides insights into the vaccination progress and the proportion of the population receiving various levels of vaccine protection. We used the dataset for vaccinations to show this plot; it is described in the additional datasets section.
Possible Causes/Consequences: Variations in these vaccination categories may reflect differences in vaccination rollout strategies, eligibility criteria for boosters, and vaccine availability. Monitoring these numbers helps gauge the level of population immunity and the need for booster campaigns.



Section 4: Problem Description

Problem Statement

Problem Description: Predicting COVID-19 Case Surges for Public Health Preparedness
The problem at hand involves using the COVID-19 dataset to predict potential surges in new cases and deaths due to the virus.

The target group for this prediction includes public health officials, policymakers, and healthcare professionals who are responsible for pandemic management and resource allocation.

Target Values:

- New Cases
- New Deaths

Relevant Columns:

- Date of Observation
- Cumulative Cases
- Cumulative Deaths

Current Approach:
Currently, public health officials rely on historical data and basic statistical models to anticipate trends and surges in COVID-19 cases and deaths. These models often involve simple time-series analysis or tracking of the reproduction number (R0). While these methods provide some insights, they have limitations in terms of accuracy, speed, and adaptability to changing conditions. Moreover, they may not fully leverage the richness of the available data.
Improvement with Data Science Model:
A data science model can significantly enhance the accuracy, speed, and adaptability of predicting COVID-19 case surges:

- Accuracy: Data science models can incorporate a wide range of variables, including vaccination rates, mobility data, and public health measures. By analyzing these factors comprehensively, the model can make more accurate predictions.

- Speed: Data science models can operate in near real-time, providing timely information for decision-making. Rapid predictions allow for proactive measures to be taken to mitigate the impact of surges.
- Consistency: Machine learning models can provide consistent predictions regardless of data volume. They can handle large datasets and identify patterns that may be missed by manual analysis.
- Explainability: Modern data science models offer transparency and interpretability. Decision-makers can understand the factors contributing to a surge prediction, allowing for informed actions.
- Lower Cost: Efficient resource allocation is crucial during a pandemic. Predictive models can optimize resource distribution, potentially reducing unnecessary costs.
- Better Data Utilization: Data science models can make use of complex data sources, such as genomic sequencing, healthcare capacity, and demographic information, providing a holistic view of the pandemic's dynamics.

By implementing a data science model for predicting COVID-19 case surges, public health officials can make informed decisions regarding healthcare resource allocation, vaccination campaigns, and public health interventions. This proactive approach can help in saving lives and reducing the strain on healthcare systems during the ongoing pandemic and future public health crises.

Section 5: References

Related Work

- Authors: Arnaout R, Arnaout R.
- Year: 2022
- Title: Visualizing omicron: COVID-19 deaths vs. cases over time.
- Journal: PLoS One
- DOI: 10.1371/journal.pone.0265233
- Main Result: This paper presents a visual analysis of the relationship between COVID-19 deaths and cases over time, with a specific focus on the omicron variant. It provides insights into how the spread of the virus and its impact on mortality have evolved, which can be relevant for understanding pandemic dynamics.

- Authors: Dong, E., Du, H., & Gardner, L.
- Year: 2020
- Title: An interactive web-based dashboard to track COVID-19 in real-time.
- Journal: The Lancet Infectious Diseases
- DOI: 10.1016/S1473-3099(20)30120-1
- Main Result: This paper introduces an interactive web-based dashboard for real-time tracking of COVID-19 cases worldwide. The dashboard provides essential data visualizations and has been widely used for monitoring the global pandemic's progress.

- Authors: Kucharski, A. J., Russell, T. W., & Eggo, R. M.
- Year: 2020
- Title: Early dynamics of transmission and control of COVID-19: A mathematical modelling study.
- Journal: The Lancet Infectious Diseases
- DOI: https://doi.org/10.1016/S1473-3099(20)30144-4

- Main Result: This paper presents a mathematical modeling study that analyzes the early dynamics of COVID-19 transmission and the potential impact of control measures. It offers insights into the effectiveness of various intervention strategies.

- Authors: JHU Coronavirus Resource Center
- Year: Ongoing
- Title: COVID-19 Dashboard by the Johns Hopkins University
- Website: https://coronavirus.jhu.edu/map.html
- Main Result: The Johns Hopkins University provides an authoritative and frequently updated COVID-19 dashboard that includes various visualizations, such as maps and graphs, to track the spread of the virus globally. This resource has been widely cited and used for COVID-19 data visualization.

- Authors: CDC COVID-19 Response Team
- Year: Ongoing
- Title: COVID Data Tracker by the Centers for Disease Control and Prevention (CDC)
- Website: https://covid.cdc.gov/covid-data-tracker
- Main Result: The CDC's COVID Data Tracker offers a wide range of data visualizations, including case counts, hospitalizations, and vaccinations. This resource is crucial for tracking the pandemic in the United States and demonstrates the use of visualizations for public health communication.

Alternative Datasets

Vaccination data

https://covid19.who.int/who-data/vaccination-data.csv

Latest reported counts of cases and deaths

https://covid19.who.int/WHO-COVID-19-global-table-data.csv

Commonalities:

- Source: All three datasets are provided by the World Health Organization (WHO) and are related to COVID-19 data, which makes them reliable sources of information.
- COVID-19 Data: They all contain data related to the COVID-19 pandemic, including information on cases, deaths, vaccinations, and other relevant statistics.
- Global Coverage: Each dataset covers COVID-19 data on a global scale, providing information about multiple countries and regions.

Differences:

https://covid19.who.int/who-data/vaccination-data.csv:

- Scope: This dataset primarily focuses on COVID-19 vaccination data, including the number of vaccine doses administered and the number of people vaccinated.

- Size: The dataset is smaller in size compared to the others since it primarily deals with vaccination data, which is a subset of overall COVID-19 data.
- Quality: The quality of the dataset is good, as it comes directly from the WHO. Vaccination data tends to be more accurate and consistent.

https://covid19.who.int/WHO-COVID-19-global-table-data.csv:

- Scope: This dataset provides a tabular summary of various COVID-19 statistics, including cases, deaths, and more. It offers a condensed view of the global COVID-19 situation.
- Size: The dataset is moderate in size, as it contains summarized data for multiple countries and regions.
- Quality: The data quality is reliable since it originates from the WHO, but it might be less detailed than the global data.

https://covid19.who.int/WHO-COVID-19-global-data.csv:

- Scope: This dataset is likely the most comprehensive and contains a wide range of COVID-19 data, including cases, deaths, testing, and more. It provides detailed information on the pandemic.
- Size: It is larger in size compared to the other datasets.
- Quality: The dataset's quality is high, as it is directly maintained by the WHO and offers a detailed view of the global COVID-19 situation.

Data Science Library

To perform data analysis and modeling on this dataset, we will utilize Python libraries such as pandas, matplotlib, and seaborn.

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns
```