**COVID-19 Global Vaccination Analysis Project Proposal**

Ryan Weeks, Sarah Yawn, Lisa Hansen

Bellevue University

DSC450 – Applied Data Science

November 2025

<https://github.com/skeew0813/Covid19-Global-Vaccination-Analysis>

**Abstract**

This project aims to analyze global COVID-19 vaccination progress and explore how vaccination rates and manufacturer distributions vary across regions. The primary objective is to prepare and explore data that can later be modeled to uncover vaccination patterns and forecast future trends. Using data from Kaggle’s “COVID World Vaccination Progress” dataset, the initial work will focus on data cleaning, feature engineering, and exploratory visualization to identify early insights and potential inconsistencies.

In the next phase, the team plans to apply statistical and machine learning techniques to develop predictive models that can estimate future vaccination rates, evaluate regional rollout performance, and detect correlations between vaccine types and distribution efficiency. Anticipated challenges include inconsistent country-level reporting, missing values, and potential discrepancies in manufacturer data. The expected outcome is a well-structured analytical framework that enables robust forecasting and supports data-driven decision-making in global public health.

**The Domain**

This study falls within the domain of public health and epidemiological analytics, focusing on global vaccination tracking and distribution during the COVID-19 pandemic. The domain integrates healthcare data management, global health surveillance, and data-driven decision-making for public policy. Insights support understanding of rollout equity, manufacturer influence, and preparedness for future global health crises. The findings are relevant to governments, international health organizations, and data scientists studying the efficiency of pandemic responses.

**The Data**

The primary data source is the “COVID World Vaccination Progress” dataset from Kaggle (Gpreda, 2021), which comprises two files: “country\_vaccinations.csv” and “country\_vaccinations\_by\_manufacturer.csv”. The first provides global, country-level vaccination data, including total doses, daily vaccination rates, and per-hundred metrics. The second details manufacturer-level totals by country and date. Data were imported, cleaned, and feature-engineered in Python using pandas, numpy, and visualization libraries (matplotlib, seaborn). Dates were parsed, missing values reviewed, and continent mapping added to enable regional aggregation. Minor inconsistencies (e.g., micro-territories and unmapped countries) were corrected manually. New features, including monthly and rolling averages, were created to support temporal and geographic analysis.

**Research Questions, Benefits, and Motivation**

This project addresses the following questions:

* Which countries and regions achieved the highest vaccination rates?
* How did daily vaccination trends evolve globally over time?
* What regional disparities exist between continents or economic zones?
* How did manufacturer distribution shape global rollout success?
* Can vaccination rate trends inform future forecasting models?

Analyzing these data provides insight into the effectiveness of international public health coordination and logistical efficiency. Results benefit policymakers and researchers seeking to understand global rollout inequalities and lay the groundwork for predictive modeling of future vaccination campaigns or similar global health responses.

**Method**

The approach combined data wrangling, exploratory data analysis (EDA), and feature engineering to prepare the dataset for downstream modeling. The initial phase involved validating data integrity, handling missing values, parsing date fields, and ensuring consistent naming conventions for countries and manufacturers. Additional features, such as continent mapping and rolling averages, were engineered to enhance temporal and regional analysis.

Exploratory visualizations, including distribution plots, vaccination rates by country, and rolling averages over time, were used to identify disparities in vaccination coverage and potential outliers. These insights informed the analytical direction and provided a foundation for further investigation.

For the upcoming Data Scientist phase, the cleaned and enriched data will serve as the basis for quantitative modeling. Possible next steps include time series forecasting (for example, ARIMA or Prophet models) to project vaccination trends, clustering techniques to identify groups of countries with similar rollout behaviors, and correlation analysis between vaccine types and rollout efficiency. These methods will help quantify the underlying factors influencing vaccination progress and provide actionable insights for public health and policy decision-making.

**Potential Issues**

Anticipated challenges include inconsistent reporting across countries and missing manufacturer data, varying population baselines that affect per-hundred metrics, irregular reporting intervals that introduce noise, political or logistical factors that limit data completeness, and the need for careful treatment of outliers and sudden reporting surges. While these limitations can affect temporal precision, most are mitigated through cleaning, normalization, and rolling-average smoothing.

**Concluding Remarks**

This project aims to establish a robust analytical framework for examining global COVID-19 vaccination progress and regional disparities. Through systematic data wrangling, validation, and exploratory analysis, the goal is to prepare clean, reliable datasets that will support the upcoming data science phase. Future work will focus on applying forecasting models, clustering techniques, and correlation studies to identify underlying factors that influence vaccination success and accessibility. The anticipated outcomes include a clearer understanding of global rollout patterns and insights that can inform future public health planning and policy development.

**References**

Banik, R., Rahman, M., Sikder, T., & Gozal, D. (2021). COVID-19 pandemic and vaccination disparities across countries. *Human Vaccines & Immunotherapeutics*, 17(9), 2843–2850. <https://doi.org/10.1080/21645515.2021.1938493>

Dai, H., Han, X., Zhao, J., & Yuan, J. (2021). The influence of vaccine availability and logistics on COVID-19 vaccination progress. *Nature Communications*, 12(1), 7227. <https://doi.org/10.1038/s41467-021-27560-w>

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5), 533–534. <https://doi.org/10.1016/S1473-3099(20)30120-1>

Gpreda. (2021). *COVID World Vaccination Progress* [Data set]. *Kaggle*. <https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>

Haque, A., Pant, A. B., & Haque, S. (2022). COVID-19 vaccines and global public health policies: A data-driven review. *Frontiers in Public Health*, 10, 864502. <https://doi.org/10.3389/fpubh.2022.864502>

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*, 5(7), 947–953. <https://doi.org/10.1038/s41562-021-01122-8>

Our World in Data. (2025). *Coronavirus (COVID-19) vaccinations*. <https://ourworldindata.org/covid-vaccinations>

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., & Roser, M. (2023). *Tracking COVID-19 vaccination data globally. Our World in Data*. <https://ourworldindata.org/covid-vaccination-tracker>

World Health Organization. (2024). *WHO Coronavirus (COVID-19) Dashboard*. <https://covid19.who.int/>

World Bank. (2023). *Global data on COVID-19 vaccine distribution and access*. <https://datacatalog.worldbank.org/dataset/global-covid-19-vaccine-distribution>