# Covid ’19 Vaccination analysis: Data Science Proposal

## Background

COVID-19 outbreak was first reported in Wuhan, China and has spread to more than 50 countries. WHO declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020 Naturally, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate actions to prevent the disease at the community level. Therefore, Corona Tracker was born as the online platform that provides latest and reliable news development, as well as statistics and analysis on COVID-19. This paper is done by the research team in the Corona Tracker community and aims to predict and forecast COVID19 cases, deaths, and recoveries through predictive modelling. The model helps to interpret patterns of public sentiment on disseminating related health information and assess political and economic influence of the spread of the virus.

## Problem Statement Overview

Vaccines save millions of lives each year. The development of safe and effective COVID-19 vaccines is a huge step forward in our global effort to end the pandemic and to get back to doing more of the things we enjoy with the people we love.

Covid ’19 Vaccination analysis dataset gathered has the latest expert information to answer some of the most common questions about COVID-19 vaccines.

Data science focuses on the processes and systems that enable the extraction of knowledge or insights from data. Data Science techniques can be leveraged in this problem statement to answer questions like:

• What vaccines are used and in which countries?

• Which vaccine is used in most of the countries?

• What country has vaccinated a larger percent from its population?

• Represent the progress of India in vaccination.

• Analyze the progress of vaccination country-wise and continent-wise

# Data Science Problem Solving Process

Diagram

Description automatically generated with low confidence

The combination of visual analytics and [data science](https://www.tibco.com/reference-center/what-is-data-science) enables people with little knowledge of statistics, to understand complex scenarios and draw inference about the future, from current events.

The COVID-19 virus has some behavioral attributes and survival strategies that make it difficult to anticipate short- and long-term infection scenarios. In particular, the exponential doubling can turn an initial spark infection into a significant outbreak in a matter of weeks.

To solve Covid ’19 Vaccination analysis problem using Data Science techniques we have to go through below mentioned phases: -

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

**Diagram

Description automatically generated**

## Business Understanding

Using Covid Vaccination dataset we are trying to identify the progress of vaccination across the world

The given problem is clustering problem as we are trying to group the individuals who are vaccinated in a country

The dataset can be used to track COVID-19 vaccination in the World, answer instantly to your questions:

* Which country is using what vaccine?
* In which country the vaccination program is more advanced?
* Where are vaccinated more people per day? But in terms of percent from entire population?

## Data Description

The data (country vaccinations) contains the following information:

**Country**- This is the country for which the vaccination information is provided.

**Country ISO Code** - ISO code for the country.

**Date** - date for the data entry; for some of the dates we have only the daily vaccinations, for others, only the (cumulative) total.

**Total number of vaccinations** - this is the absolute number of total immunizations in the country.

**Total number of people vaccinated** - a person, depending on the immunization scheme, will receive one or more (typically 2) vaccines; at a certain moment, the number of vaccinations might be larger than the number of people.

**Total number of people fully vaccinated** - this is the number of people that received the entire set of immunization according to the immunization scheme (typically 2); at a certain moment in time, there might be a certain number of people that received one vaccine and another number (smaller) of people that received all vaccines in the scheme.

**Daily vaccinations (raw)** - for a certain data entry, the number of vaccination for that date/country.

**Daily vaccinations** - for a certain data entry, the number of vaccination for that date/country.

**Total vaccinations per hundred** - ratio (in percent) between vaccination number and total population up to the date in the country.

**Total number of people vaccinated per hundred** - ratio (in percent) between population immunized and total population up to the date in the country.

**Total number of people fully vaccinated per hundred** - ratio (in percent) between population fully immunized and total population up to the date in the country.

**Number of vaccinations per day** - number of daily vaccinations for that day and country.

**Daily vaccinations per million** - ratio (in ppm) between vaccination number and total population for the current date in the country.

**Vaccines used in the country** - total number of vaccines used in the country (up to date).

**Source name** - source of the information (national authority, international organization, local organization etc.)

**Source website** - website of the source of information.

**Total number of vaccinations** - total number of vaccinations / current time and vaccine type.

## Some Observations from the analysis of data are:-

* The dataset contains 21628 observations and 15 attributes.
* people\_fully\_vaccinated\_per\_hundred and people\_fully\_vaccinated has highest count of nulls followed by daily\_vaccinations\_raw and people\_vaccinated\_per\_hundred
* Observations are recorded from Dec 2020 to May 2021
* Dataset has 214 countries
* Dataset has 45 unique vaccines
* China has maximum number of vaccinations done followed by USA, India, Brazil, and UK

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

## Feature Engineering

In feature engineering, new features are created to extract more information from existing features. These new features may have an improved ability to

explain the variance in the training data and improve model accuracy. Feature engineering is highly influenced by business understanding.

We have used feature engineering here to analyze the data. We have selected relevant features for model building and ignored redundant features.

Also, we have done data manipulation like aggregation to answer business questions

Chart

Description automatically generated

We have identified some redundant features in the given dataset: -

1. **Country** and **iso\_code** are giving same information. So iso\_code can be ignored for modeling.
2. **source\_name** and **source\_website** are giving same information. So source\_website can be ignored for modeling

## Audience

The analysis results of covid vaccination can be used by WHO, Government, Research Organizations, Medical Institutions, Media