**IBM Phase 1 Project**

**COVID-19 Vaccine Data Analysis**

**Phase 1: Problem Definition and Design Thinking.**

**Project summated by** : VIJAYARATHANA S.P.

**Topic/ Position** : COVID-19 Vaccine Data Analysis/ Phase 1.

**Organization/ Company** : Naan Mudhalvan/IBM Applied Data Science.



**Problem Definition**

The problem at hand involves conducting an extensive analysis of COVID-19 vaccine data with a focus on vaccine efficacy, distribution, and adverse effects. The ultimate goal is to provide valuable insights that can aid policymakers and health organizations in optimizing vaccine deployment strategies. This multifaceted project encompasses various stages, including data collection, data preprocessing, exploratory data analysis, statistical analysis, and visualization.

**Design Thinking**

1. **Data Collection**

We will begin by gathering COVID-19 vaccine data from Kaggle sources, including health organizations, government databases, and relevant research publications. These sources will provide us with a comprehensive and reliable dataset for our analysis.

**Objective**:

Collect Covid-19 vaccine data from credible and authoritative sources, including health organizations, government databases, and research publications.

**Methodology:**

* Identify and access relevant data sources.
* Extract and download the required data, ensuring it covers vaccine efficacy, distribution, and adverse effects.
* Verify the data for accuracy, consistency, and completeness.

**Dataset Link:**[**https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress**](https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress)

**The data 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 vaccin ations** - 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 vaccination 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 vaccination 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;

**Dataset From Kaggle:**

****

1. **Data Preprocessing**

Once we have collected the data, we will embark on the data preprocessing phase. This will involve cleaning the data to remove any inconsistencies or errors, handling missing values through imputation techniques, and converting categorical features into numerical representations to make them suitable for analysis.

**Objective:** Clean and preprocess the collected data to make it suitable for analysis.

**Methodology:**

* Handle missing values by imputation or elimination.
* Standardize data formats and units.
* Convert categorical features into numerical representations using encoding techniques.
* Check for and handle any data anomalies.

1. **Exploratory Data Analysis (EDA)**

EDA will be a critical step in understanding the dataset's characteristics. During this phase, we will explore the data, identify trends, patterns, and potential outliers. EDA will provide essential insights into the distribution of vaccine-related variables.

**Objective:** Explore the data to gain insights into its characteristics, identify trends, and detect outliers.

**Methodology:**

* Perform summary statistics to describe the dataset.
* Create visualizations (e.g., histograms, scatter plots) to identify patterns and trends.
* Use box plots and scatter plots to identify outliers and anomalies.
* Determine correlations between variables.

1. **Statistical Analysis**

We will conduct various statistical tests to analyze vaccine efficacy, adverse effects, and distribution across different populations. These tests will help us quantify the impact of vaccines and provide a rigorous basis for our recommendations.

**Objective:** Perform statistical tests to analyze vaccine efficacy, adverse effects, and distribution across different populations.

**Methodology:**

* Calculate vaccine efficacy rates and confidence intervals.
* Conduct hypothesis tests to compare vaccine efficacy across different groups.
* Analyze adverse effects by demographic variables.
* Assess the geographical distribution of vaccines.

**Visualization**

To effectively communicate our findings, we will create a range of visualizations such as bar plots, line charts, heatmaps, and more. Visualizations will play a crucial role in presenting key insights and making complex data accessible to stakeholders.

**Objective:** Create informative visualizations to present key findings and insights.

**Methodology:**

* Generate various types of plots and charts (e.g., bar plots, line charts, heatmaps) to visualize data trends.
* Utilize interactive visualization tools for better exploration.
* Ensure visualizations are clear, concise, and relevant to the analysis.
* Insights and Recommendations

1. **Insights and Recommendations**

Based on the comprehensive analysis, we will extract actionable insights and formulate recommendations that can be used by policymakers and health organizations. These insights will help optimize vaccine deployment strategies, enhance distribution efficiency, and mitigate adverse effects.

**Objective**: Provide actionable insights and recommendations based on the analysis to assist policymakers and health organizations.

**Methodology:**

* Summarize key findings and trends.
* Translate data insights into actionable recommendations for vaccine deployment strategies.
* Communicate results through reports and presentations.