

COVID-19 Vaccines Analysis

Project Documentation

Shivashankar.M
AU211521243145

PROBLEM STATEMENT

The goal of this project is to analyze COVID-19 vaccination data to gain insights into the distribution and progress of vaccinations across different countries.

Specifically, we aim to:

- Examine the total number of vaccinations over time.
- Analyze daily vaccination rates.
- Investigate the vaccination progress in specific countries.
- Identify the top countries with the highest total vaccinations and vaccinations per hundred.
- Explore the number of people vaccinated and fully vaccinated in select countries.
- Calculate and visualize key statistical measures such as median, variance, standard deviation, skewness, and kurtosis.
- Develop a machine learning model to find patterns and calculate feature importance in the dataset.

DESIGN THINKING PROCESS

This project follows a structured data analysis approach:

1. Data Collection: We obtained the COVID-19 vaccination dataset from a reliable source.
2. Data Preprocessing: Data cleaning was conducted to remove missing values. The 'date' column was formatted to the datetime data type. Additionally, unnecessary columns like 'source_name' and 'source_website' were removed.
3. Data Exploration and Visualization: We explored the data by creating various visualizations, including line plots for total and daily vaccinations over time, bar plots to compare different aspects of vaccination across countries, and histograms to analyze statistical properties of the data.
4. Statistical Analysis: Key statistics, such as median, variance, standard deviation, skewness, and kurtosis, were calculated and visualized.
5. Model Development: A machine learning model is developed to find hidden patterns in the dataset.

PHASES OF DEVELOPMENT

1. Data Collection and Loading

We imported the necessary packages, including NumPy, Pandas, Seaborn, and Matplotlib. The dataset was loaded from a CSV file into a Pandas DataFrame.

2. Data Exploration

We displayed the initial rows of the dataset and conducted an overview of the data by checking its dimensions, summary statistics, and the presence of missing values.

3. Data Preprocessing

Missing values were removed to ensure data quality. The 'date' column was converted to a datetime data type, and unnecessary columns were dropped.

4. Data Visualization

We created a series of visualizations to explore the data, such as line plots for vaccinations over time, bar plots for comparisons between countries, and bar plots for vaccinations per hundred.

5. Statistical Analysis

We calculated and visualized key statistical measures to better understand the data. These measures include median, variance, standard deviation, skewness, and kurtosis.

6. Model Development

A XGBoost model is developed based on the given data to find patterns in the dataset and to find the important features in the dataset.

DATASET

The dataset used for this analysis is the COVID-19 vaccination dataset, which includes information on various aspects of vaccinations in different countries, such as total vaccinations, daily vaccinations, people vaccinated, and people fully vaccinated.

Link to the dataset: <https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>

KEY FINDINGS AND RECOMMENDATIONS

Based on the analysis of the COVID-19 vaccination dataset, here are the key findings and recommendations:

1. Total Vaccinations Over Time: Total vaccinations increased over time, showing the progress of global vaccination efforts. Countries experienced fluctuations in daily vaccination rates.
2. Top Countries with Total Vaccinations: We identified the top countries with the highest total vaccinations. The United States, China, India, and Brazil were among the top contributors.
3. Vaccinations Per Hundred: Some countries achieved high vaccinations per hundred people, indicating their success in reaching a substantial portion of their population.
4. People Vaccinated vs. People Fully Vaccinated: We compared the number of people vaccinated to the number of people fully vaccinated. Some countries had a significant gap between these two categories.
5. Statistical Analysis: We provided statistical measures like median, variance, standard deviation, skewness, and kurtosis to describe the data's properties.

Based on these findings, we recommend monitoring and adjusting vaccination strategies in countries where there is a significant gap between the number of people vaccinated and fully vaccinated. Additionally, countries with lower vaccination rates should focus on increasing their vaccination efforts.