TITLE: COVID-19 VACCINE ANALYSIS

Step 1: Problem Definition

- The objective of this project is to conduct a comprehensive analysis of Covid-19 vaccine data, with a primary 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 data collection, data preprocessing, exploratory data analysis (EDA), statistical analysis, visualization, and the formulation of actionable recommendations.

Step 2: Data Collection

- We will gather Covid-19 vaccine data from reliable sources, including health organizations (e.g., WHO, CDC), government databases, and peer-reviewed research publications.
- The dataset located at (https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress) will serve as a primary source.
- Data is collected daily from **Our World in Data** GitHub repository for covid-19, merged and uploaded. Country level vaccination data is gathered and assembled in one single file.
- Then, this data file is merged with locations data file to include vaccination sources information. A second file, with manufacturers information, is included.

Step 3: Data Preprocessing

- Cleaning and preprocessing the data are essential steps in preparing it for analysis.
- This involves addressing issues such as duplicate records, inconsistent formatting, handling missing values, and converting categorical features into numerical representations.

Step 4: Data Exploration

- Perform exploratory data analysis (EDA) to understand the data's distribution, correlations, and trends.
- In this phase, we will dive into the dataset to gain a deeper understanding of its characteristics. EDA will involve generating statistical summaries, visualizing data distributions, and identifying trends and outliers.
- Key areas of exploration include vaccine distribution across regions, vaccination rates over time, and potential anomalies.
- Visualize the data to gain insights into vaccine distribution and adverse effects.

Step 5: Feature Engineering

- Create relevant features or transformations that can help in clustering or time series forecasting.
- Consider aggregating data at different levels (e.g., country, region) for a more granular analysis.

Step 6: Clustering Analysis

- Utilize advanced clustering techniques such as K-means, hierarchical clustering, or DBSCAN to group countries or regions based on vaccination progress and adverse effects patterns.
- Determine the optimal number of clusters using techniques like the elbow method or silhouette score.

Step 7: Time Series Forecasting

- For time series forecasting, choose an appropriate algorithm such as ARIMA, Prophet, or LSTM.
- Split the data into training and testing sets and train the model to predict future vaccination rates or adverse effects.

Step 8: Model Evaluation

- Assess the performance of your clustering or time series forecasting model using appropriate evaluation metrics.
- Visualize the results to communicate insights effectively.

Step 9: Documentation and Reporting

Create a comprehensive document that includes:

- Introduction and problem statement.
- Data description and preprocessing details.
- Clustering or time series forecasting methodology and results.
- Interpretation of findings and actionable recommendations.
- Visualizations, charts, and graphs for better understanding.
- Include code snippets, algorithms used, and parameters tuned in the document.