**STEPS TO COLLECT AND PREPROCESS DATA**

**1. Data Collection:**

a. Data Sources:

Official Health Organizations: Websites of health organizations like the World Health Organization (WHO) or the Centers for Disease Control and Prevention (CDC) often provide reliable data.

Government Health Portals: Government health websites of different countries might provide region-specific data.

Kaggle Datasets: Kaggle is a platform where you can find a variety of datasets, including COVID-19 data.

**b. Data Variables**:

Vaccine Type: Specify which vaccines you are interested in (Pfizer, Moderna, AstraZeneca, etc.).

Vaccination Rates: Number of people vaccinated, both first and second doses.

Location Data: Breakdown by countries, states, or regions.

Time Period: Daily, weekly, or monthly data, depending on your analysis requirements.

2. Data Preprocessing:

a. Data Cleaning:

Handling Missing Values: Check for and handle missing or null values in the dataset. This might involve interpolation, removal, or replacement with appropriate values.

Data Validation: Ensure the data values fall within expected ranges.

**b. Data Transformation:**

Normalization/Standardization: Scale numerical features if necessary.

Encoding Categorical Variables: Convert categorical variables (like vaccine types) into numerical representations for analysis.

**c. Feature Engineering:**

Derive New Features: Create new features if necessary. For example, you might calculate vaccination rates based on the available data.

Date-Time Features: Extract features like day of the week, month, or year from the date if your dataset includes timestamps.

**d. Data Aggregation:**

Aggregate Data: Depending on your analysis, you might need to aggregate data at different levels (daily to weekly, country level to continent level).

**e. Data Visualization:**

Explore Data: Visualize the data using charts (line charts, bar charts, heatmaps) to understand patterns and trends.

Outlier Detection: Identify and handle outliers if they exist in the data.

**f. Data Splitting:**

Training and Testing Sets: If you're building predictive models, split the data into training and testing sets.

**g. Data Saving:**

Save Processed Data: Save the cleaned and preprocessed data into a separate file for further analysis. This ensures you can work from the same cleaned dataset for future analysis without repeating preprocessing steps.

**Python code**

import pandas as pd

# Load the data from CSV file

##### data = pd.read\_csv(' country\_vaccinations.csv')

# Data Cleaning

data = data.dropna()

# Drop rows with missing values

# Data Transformation

# Assuming 'Vaccine\_Type' is categorical, encode it into numerical values

data['Vaccine\_Type'] = pd.Categorical(data['Vaccine\_Type']).codes

# Data Aggregation (Example: Sum the number of vaccinated people per vaccine type)

aggregated\_data=data.groupby('Vaccine\_Type')['Number\_of\_Vaccinated'].sum().reset\_index()

# Data Saving (Optional)

aggregated\_data.to\_csv('aggregated\_vaccine\_data.csv', index=False)

# Print the preprocessed data

print(aggregated\_data)

