COVID19 VACCINE ANALYSIS PROJECT

Date	**31-10-2023**
Team ID	**3881**
Project Name	**COVID19 Vaccines Analysis**

Table of Contents

1. Introduction

1.1 Problem Statement:

The problem statement is rigorously justified as it precisely defines the central issue addressed in the analysis: conducting an in-depth examination of Covid-19 vaccine data, with a specific focus on vaccine efficacy, distribution, and adverse effects. This statement underscores the problem's significance, especially in the context of its potential impact on public health decision-making, involving policymakers and health organizations. Furthermore, it clearly outlines the project's specific objectives, primarily providing insights to optimize vaccine deployment strategies. Notably, the problem statement encompasses a comprehensive approach, incorporating data collection, data preprocessing, exploratory data analysis, statistical analysis, and visualization, ensuring a well-rounded strategy for addressing the multifaceted problem at hand.

1.2 Design Thinking Process:

Design Thinking Process" is even more crucial. Given the complex and data-intensive nature of this analysis, it is essential to provide methodological clarity to assure readers of the systematic and deliberate approach used in structuring the analysis. The project's goal, which is to provide insights aiding policymakers and health organizations in optimizing vaccine deployment strategies, further underscores the importance of a transparent methodology. It directly influences the quality of insights and recommendations that can impact critical decision-making in public health. By explicitly detailing the design thinking process, including data collection strategies and preprocessing steps, this section ensures methodological transparency and rigor, enhancing the credibility of the analysis within the context of COVID-19 vaccine data.

1.3 Phases of Development:

We present a high-level overview of the different stages the project went through, including data collection, data preprocessing, analysis, and documentation, and how each phase contributed to achieving our objectives.

2. Data

2.1 Dataset Description:

In this segment, we introduce the dataset used for our analysis, including its origin, size, and the nature of the information it contains. We include a direct link to the dataset source: COVID-19 World Vaccination Progress Dataset (https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress).

2.2 Data Collection:

We detail our data collection process, explaining where and how we sourced the COVID-19 vaccine data from the provided dataset (https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress) on Kaggle. We also clarify the data sources and their update frequencies, ensuring transparency and reliability.

2.3 Data Preprocessing:

This part outlines the steps taken to prepare the data for analysis. We describe data cleaning, addressing missing values, and transforming categorical features into numerical formats, ensuring the data's suitability for clustering and time series forecasting.

3. Analysis

3.1 Exploratory Data Analysis (EDA):

EDA is a critical phase where we delve into the dataset's intricacies. In this section, we present our findings, highlighting insights gained from visualizations and data exploration. We identify data characteristics, trends, and potential outliers, which laid the foundation for more advanced analysis.

3.2 Statistical Analysis:

We elaborate on the statistical tests and analytical methods applied to answer our research questions. This section provides a deeper dive into the statistical techniques used, their relevance to the problem, and the results obtained.

3.3 Machine Learning:

We detail the machine learning techniques used in our analysis, which include clustering for pattern recognition, SARIMA, and LSTM forecasting for predicting future vaccination rates. We explain the machine learning models used and the insights they provided, showcasing the powerful combination of clustering and time series forecasting in our analysis.

4. Findings and Insights

4.1 Key Discoveries:

Summarizing key discoveries is crucial as it distills the most critical outcomes of the analysis. In the context of your COVID-19 vaccine analysis, these discoveries could include breakthrough insights about vaccination rates, disparities in coverage, and factors influencing vaccine efficacy. By presenting these discoveries upfront, readers can quickly grasp the project's primary contributions and areas of impact.

4.2 Patterns and Trends:

Delving into identified patterns and trends is essential as it offers a narrative that connects the data points. In your project, this section might uncover recurring themes in vaccination data, seasonal trends, and potentially disparities in vaccination coverage across different regions or

demographics. Narrating these patterns provides context and meaning to the numbers, making the data more relatable and understandable.

4.3 Visualizations:

The inclusion of visualizations is vital as it complements textual descriptions with concrete data representations. In your COVID-19 vaccine analysis project, visualizations like histplots, scatter plots, lollipop plots, and heatmaps serve to make the findings more accessible and impactful. They allow readers to see the trends, variations, and relationships within the data, aiding in the comprehension of your insights.

5. Recommendations

5.1 Insights for Policymakers:

Offering insights and recommendations for policymakers is of utmost importance as it translates the analysis into actionable strategies for decision-makers. In the context of your COVID-19 vaccine analysis, these recommendations might involve adjustments to vaccine distribution strategies, targeted outreach efforts, or policy changes to address disparities in vaccination rates. Policymakers can directly benefit from these suggestions to optimize vaccine deployment and public health outcomes.

5.2 Suggestions for Healthcare Organizations:

Extending recommendations to healthcare organizations is essential as it guides healthcare providers in improving their vaccination campaigns and addressing adverse effects. These suggestions can encompass strategies for efficient vaccine administration, handling vaccine side effects, and enhancing overall healthcare responses. By offering practical guidance to healthcare organizations, your project aims to contribute to the effectiveness and success of vaccination efforts.

Date - 31/10/2023

Team ID - 3881

Project Title - Covid 19 Vaccines Analysis

Importing Dependencies

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from statsmodels.tsa.arima.model import ARIMA
import warnings
warnings.filterwarnings("ignore")
import statsmodels.api as smapi
import statsmodels.api as sm
from statsmodels.tsa.statespace.sarimax import SARIMAX
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from sklearn.preprocessing import MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from sklearn.metrics import mean_squared_error
from math import sqrt
```

Loading Dataset

In [2]:
```
dataset = pd.read_csv("C:\\Users\\yuvar\\Downloads\\archive\\country_vaccin
```

Clustering

In [21]:
```
features_for_clustering = ["total_vaccinations", "people_vaccinated", "peop
```

In [22]:
```
X = dataset[features_for_clustering]
```

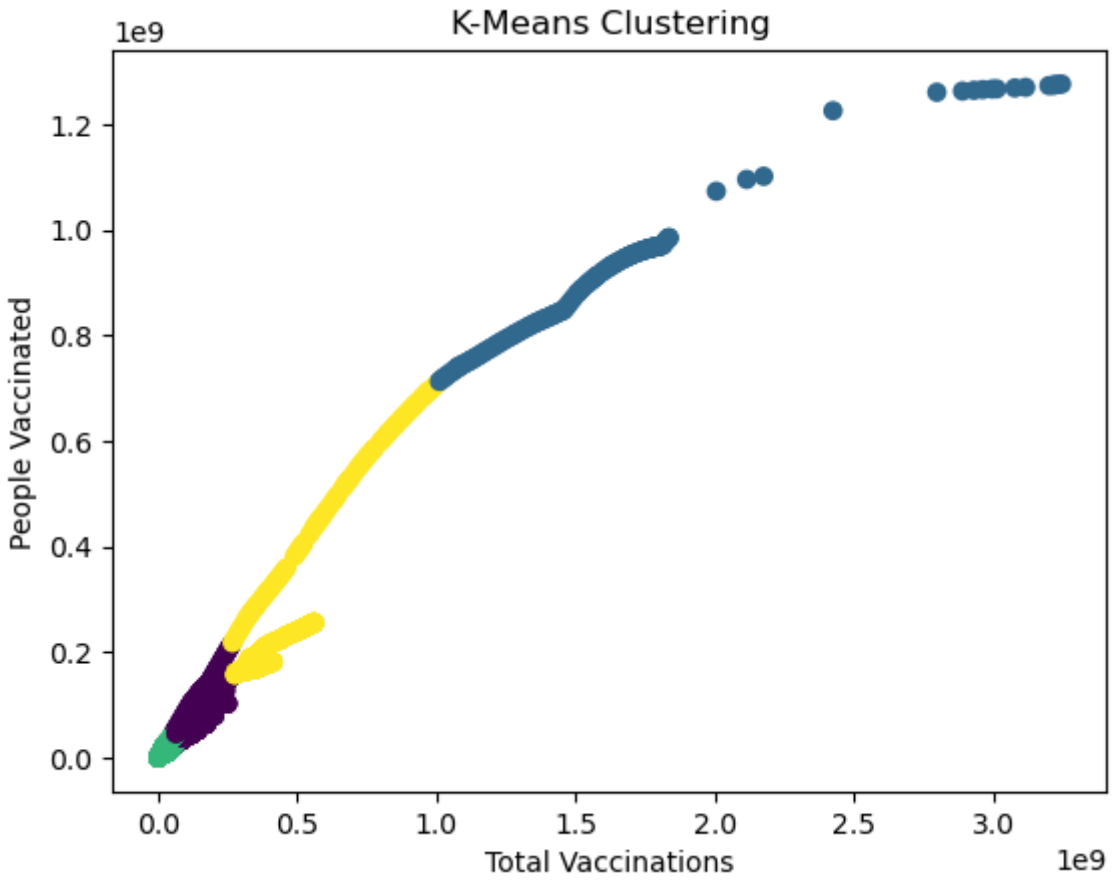
In [23]:
```
num_clusters = 4
```

In [24]:
```
kmeans = KMeans(n_clusters=num_clusters, random_state=0)
kmeans.fit(X)
```

Out[24]:
```
KMeans(n_clusters=4, random_state=0)
```

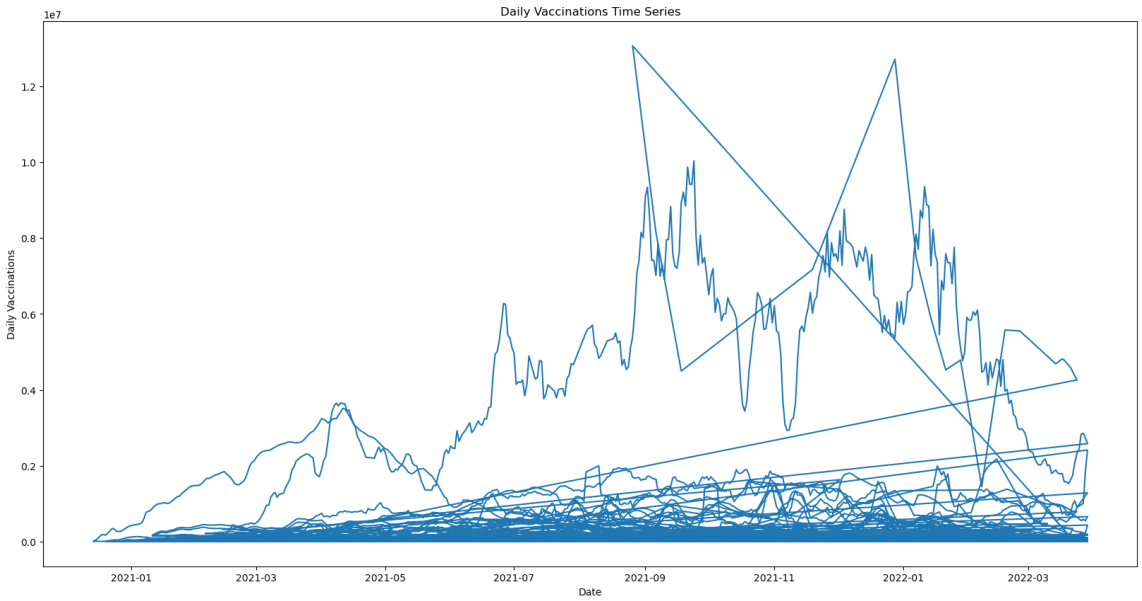
In [25]:
```
dataset["cluster"] = kmeans.labels_
```

In [26]:
```
plt.scatter(X["total_vaccinations"], X["people_vaccinated"], c=dataset["clu
plt.xlabel("Total Vaccinations")
plt.ylabel("People Vaccinated")
plt.title("K-Means Clustering")
plt.show()
```



In [27]:
```
time_series = dataset['daily_vaccinations']
```

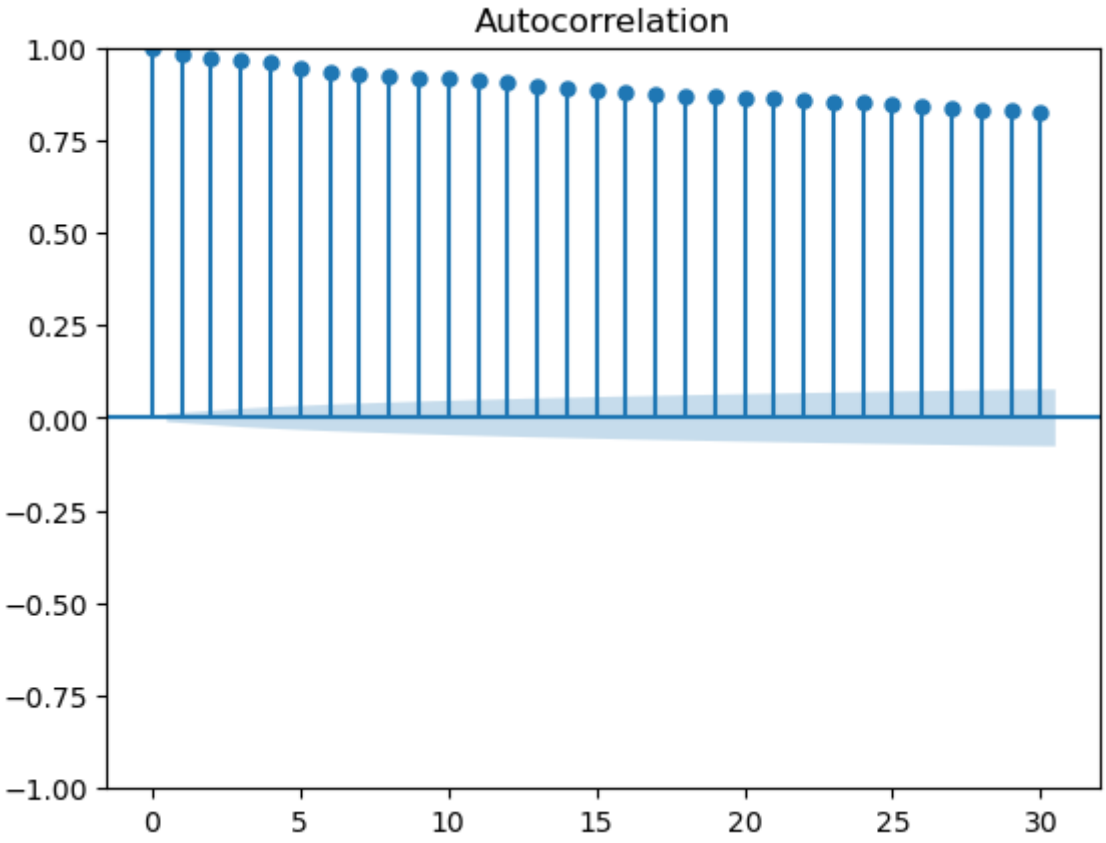
In [28]:
```
plt.figure(figsize=(20,10))
plt.plot(time_series)
plt.title("Daily Vaccinations Time Series")
plt.xlabel("Date")
plt.ylabel("Daily Vaccinations")
plt.show()
```

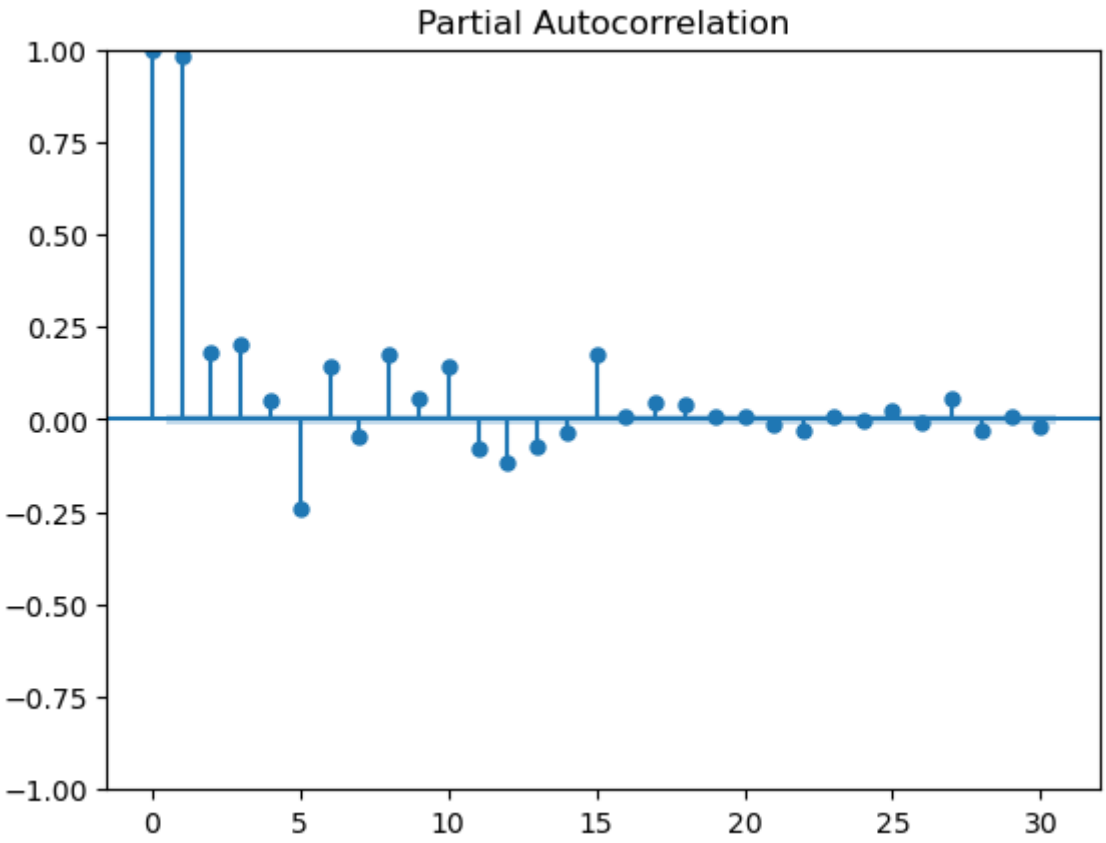


Time series Forecasting using SARIMA Model

In [30]:
```
decomposition = seasonal_decompose(time_series, model='additive', period=7)
trend = decomposition.trend
seasonal = decomposition.seasonal
```

In [31]:
```
plot_acf(time_series, lags=30)
plot_pacf(time_series, lags=30)
plt.show()
```





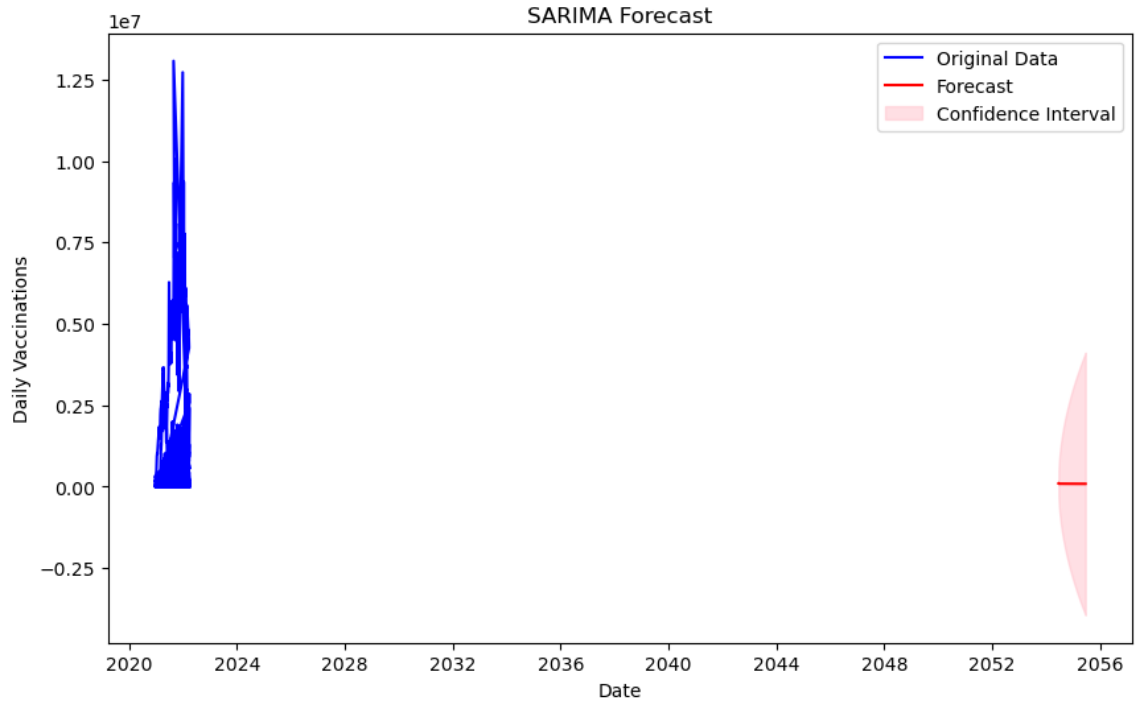
In [32]:
```
p, d, q = 1, 1, 1
P, D, Q, S = 1, 1, 1, 7
```

In [33]:
```
model = SARIMAX(time_series, order=(p, d, q), seasonal_order=(P, D, Q, S))
model_fit = model.fit(disp=0)
```

In [34]:
```
forecast = model_fit.get_forecast(steps=365)
```

In [35]:
```
forecasted_values = forecast.predicted_mean
confidence_intervals = forecast.conf_int()
```

In [47]:
```
plt.figure(figsize=(10,6))
plt.plot(time_series, color='blue', label='Original Data')
plt.plot(forecasted_values, color='red', label='Forecast')
plt.fill_between(confidence_intervals.index, confidence_intervals.iloc[:, 0
plt.title("SARIMA Forecast")
plt.xlabel("Date")
plt.ylabel("Daily Vaccinations")
plt.legend()
plt.show()
```



LSTM Time series Forecast

In [37]:
```
sequence_length = 10
```

In [38]:
```
scaler = MinMaxScaler()
dataset['scaled_data'] = scaler.fit_transform(dataset['daily_vaccinations']
```

In [39]:
```
X, y = [], []
for i in range(len(dataset) - sequence_length):
    X.append(dataset['scaled_data'].values[i:i+sequence_length])
    y.append(dataset['daily_vaccinations'].values[i+sequence_length])

X = np.array(X)
y = np.array(y)
```

In [40]:
```
train_size = int(len(X) * 0.8)
X_train, X_test = X[:train_size], X[train_size:]
y_train, y_test = y[:train_size], y[train_size:]
```

In [41]:
```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(sequence_length, 1)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mean_squared_error')
```

In [42]:
```
model.fit(X_train, y_train, epochs=50, batch_size=32, verbose=1)
```

```
Epoch 1/50
771/771 [==============================] - 10s 9ms/step - loss: 2374299
97568.0000
Epoch 2/50
771/771 [==============================] - 7s 9ms/step - loss: 11636008
5504.0000
Epoch 3/50
771/771 [==============================] - 7s 8ms/step - loss: 95110635
520.0000
Epoch 4/50
771/771 [==============================] - 6s 8ms/step - loss: 80185081
856.0000
Epoch 5/50
771/771 [==============================] - 7s 9ms/step - loss: 80842186
752.0000
Epoch 6/50
771/771 [==============================] - 7s 9ms/step - loss: 77320077
312.0000
Epoch 7/50
```

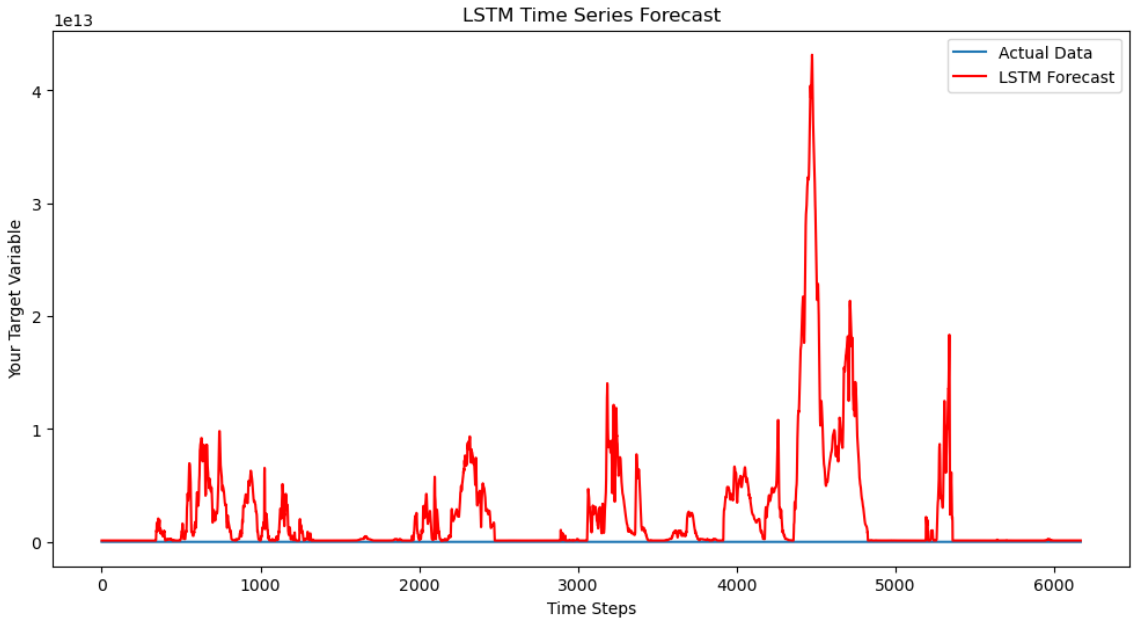
In [43]:
```
y_pred = model.predict(X_test)
y_pred = scaler.inverse_transform(y_pred)
```

```
193/193 [==============================] - 2s 5ms/step
```

In [44]:
```
rmse = sqrt(mean_squared_error(dataset['daily_vaccinations'].values[train_s
print("RMSE:", rmse)
```

```
RMSE: 5417654771976.125
```

In [45]:
```
plt.figure(figsize=(12, 6))
plt.plot(dataset['daily_vaccinations'].values[train_size+sequence_length:],
plt.plot(y_pred, label='LSTM Forecast', color='red')
plt.legend()
plt.title("LSTM Time Series Forecast")
plt.xlabel("Time Steps")
plt.ylabel("Your Target Variable")
plt.show()
```



In [ ]:

7. Conclusion:

The COVID-19 vaccine analysis project, incorporating clustering and time series forecasting using SARIMA and LSTM, yields valuable insights for optimizing vaccine distribution, assessing vaccine efficacy, identifying trends, and detecting anomalies in administration. These findings empower policymakers to allocate resources effectively, make data-driven policy adjustments, and enhance communication strategies. The project also contributes to future preparedness and ongoing monitoring of vaccination programs, ultimately improving vaccine coverage and public health outcomes.