**Title**:

Covid Analysis

**Problem Statement:**

Analyzing and Forecasting COVID-19 Data: A Comprehensive Study of Vaccination Trends and Time Series Analysis.

**Date**: 16/05/2023

**Objectives**:

a) Exploring COVID-19 Data: Insights into Vaccination Trends by answering the following questions: -

1. Describe the dataset and its columns.

2. Determine the number of persons state-wise vaccinated for the first dose in India.

3. Determine the number of persons state-wise vaccinated for the second dose in India.

4. Calculate the total number of males and females vaccinated.

b) Time Series Analysis for Effective Planning and Forecasting.

**Theory**:

The COVID-19 pandemic has had a significant impact on societies worldwide, and understanding the patterns and trends of the virus is crucial for effective planning and decision-making. This mini-project aims to analyze and forecast COVID-19 data through two main aspects: vaccination trends and time series analysis.

Data analysis and visualization play an essential role in understanding the impact of COVID-19, tracking the spread of the virus, monitoring vaccination efforts, identifying patterns, and informing public health policies. Such projects can involve analysing datasets related to COVID-19 cases, hospitalizations, testing, vaccination rates, and other relevant factors to gain insights and make informed decisions.

1. Vaccination Trends Analysis:

The dataset provides valuable insights into the COVID-19 vaccination process, including the number of individuals vaccinated for the first and second doses in different states of India. By analyzing the dataset, we can determine the state-wise distribution of vaccinations, identify the total number of males and females vaccinated, and assess the progress of the vaccination campaign.

2. Time Series Analysis:

Time series analysis is a powerful tool for understanding the temporal patterns and dynamics of COVID-19 cases. By applying appropriate models, such as SARIMA (Seasonal Autoregressive Integrated Moving Average) or VAR (Vector Autoregression), we can analyze the historical data and forecast the future trends of COVID-19 cases. This enables us to make informed decisions, plan resources, and implement effective strategies to mitigate the spread of the virus.

By combining the analysis of vaccination trends and time series forecasting, this mini-project aims to provide valuable insights into the COVID-19 situation. The results can help healthcare authorities, policymakers, and researchers to better understand the progress of vaccination efforts and make data-driven decisions to combat the pandemic effectively.

Overall, this project serves as a comprehensive study of COVID-19 data, combining insights from vaccination trends and time series analysis. The findings contribute to the broader understanding of the pandemic and support evidence-based decision-making for public health and policy interventions.

**System Architecture:**

No specific system architecture is required for this analysis.

**Methodology:**

**Part 1:**

1. **Describe Datasets:**

covid_vaccine_statewise.csv:

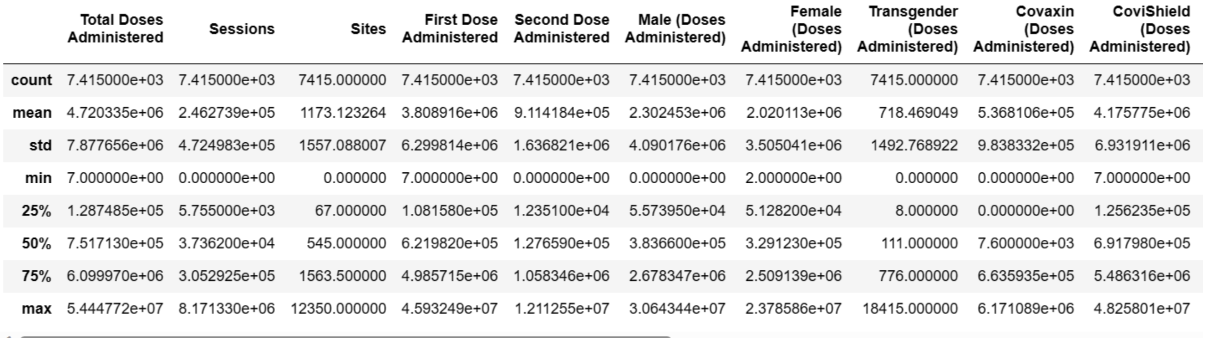
The given dataset represents COVID-19 vaccination data with various columns capturing different aspects of the vaccination process. Here is the theory for the dataset:

- Updated On: This column indicates the date when the data was updated or recorded.
- State: This column represents the states for which the vaccination data is provided. It provides information about the geographical location of the vaccination activities.
- Total Doses Administered: This column represents the total number of vaccine doses administered in a given state. It includes both the first and second doses.
- Sessions: This column denotes the number of sessions conducted for administering the vaccines in a particular state. A session refers to a specific time period or location where vaccinations are carried out.
- Sites: This column indicates the number of vaccination sites where individuals can receive the vaccine.
- First Dose Administered: This column captures the number of individuals who have received the first dose of the COVID-19 vaccine in a specific state.
- Second Dose Administered: This column represents the count of individuals who have received the second dose of the COVID-19 vaccine in a particular state. It indicates the completion of the vaccination process for these individuals.
- Male (Doses Administered), Female (Doses Administered), Transgender (Doses Administered): These columns provide the breakdown of vaccine doses administered based on the gender of the individuals.
- Covaxin (Doses Administered), CoviShield (Doses Administered), Sputnik V (Doses Administered): These columns represent the number of vaccine doses administered for different vaccine types.
- AEFI: This column stands for Adverse Events Following Immunization. It captures the count of adverse events reported after vaccination.
- Age Group Columns: These columns include the breakdown of vaccine doses administered based on different age groups such as 18-44 years, 45-60 years, and 60+ years. It provides insights into the vaccination coverage across different age demographics.
- Gender-Specific Columns: These columns capture the breakdown of vaccinated individuals based on gender.
- Total Individuals Vaccinated: This column represents the total count of individuals who have received at least one dose of the COVID-19 vaccine in a specific state.

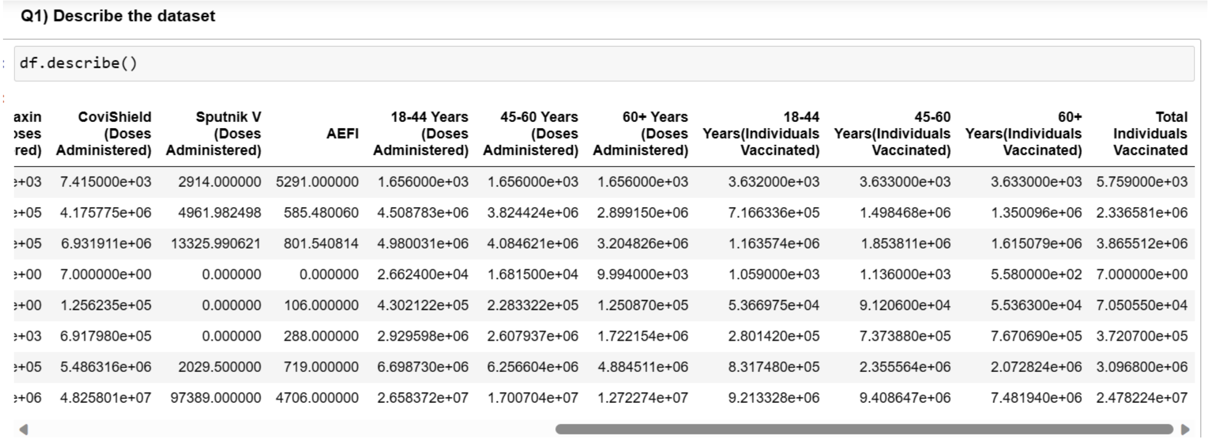
Columns having non-null values less than 30% of the total values in the dataset:

Age Group Columns: The columns related to age groups, such as 18-44 Years (Doses Administered), 45-60 Years (Doses Administered), and 60+ Years (Doses Administered), have a relatively high number of null values. Only 1702 non-null entries are present for these columns.

Gender-Specific Columns: The columns representing gender-specific vaccination data, including Male (Individuals Vaccinated), Female(Individuals Vaccinated), and Transgender(Individuals Vaccinated), have a limited number of non-null values. Only 160 entries are available for these columns.



| | Total Doses Administered | Sessions | Sites | First Dose Administered | Second Dose Administered | Male (Doses Administered) | Female (Doses Administered) | Transgender (Doses Administered) | Covaxin (Doses Administered) | CoviShield (Doses Administered) |
|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 7.415000e+03 | 7.415000e+03 | 7415.000000 | 7.415000e+03 | 7.415000e+03 | 7.415000e+03 | 7.415000e+03 | 7415.000000 | 7.415000e+03 | 7.415000e+03 |
| **mean** | 4.720335e+06 | 2.462739e+05 | 1173.123264 | 3.808916e+06 | 9.114184e+05 | 2.302453e+06 | 2.020113e+06 | 718.469049 | 5.368106e+05 | 4.175775e+06 |
| **std** | 7.877656e+06 | 4.724983e+05 | 1557.088007 | 6.299814e+06 | 1.636821e+06 | 4.090176e+06 | 3.505041e+06 | 1492.768922 | 9.838332e+05 | 6.931911e+06 |
| **min** | 7.000000e+00 | 0.000000e+00 | 0.000000 | 7.000000e+00 | 0.000000e+00 | 0.000000e+00 | 2.000000e+00 | 0.000000 | 0.000000e+00 | 7.000000e+00 |
| **25%** | 1.287485e+05 | 5.755000e+03 | 67.000000 | 1.081580e+05 | 1.235100e+04 | 5.573950e+04 | 5.128200e+04 | 8.000000 | 0.000000e+00 | 1.256235e+05 |
| **50%** | 7.517130e+05 | 3.736200e+04 | 545.000000 | 6.219820e+05 | 1.276590e+05 | 3.836600e+05 | 3.291230e+05 | 111.000000 | 7.600000e+03 | 6.917980e+05 |
| **75%** | 6.099970e+06 | 3.052925e+05 | 1563.500000 | 4.985715e+06 | 1.058346e+06 | 2.678347e+06 | 2.509139e+06 | 776.000000 | 6.635935e+05 | 5.486316e+06 |
| **max** | 5.444772e+07 | 8.171330e+06 | 12350.000000 | 4.593249e+07 | 1.211255e+07 | 3.064344e+07 | 2.378586e+07 | 18415.000000 | 6.171089e+06 | 4.825801e+07 |



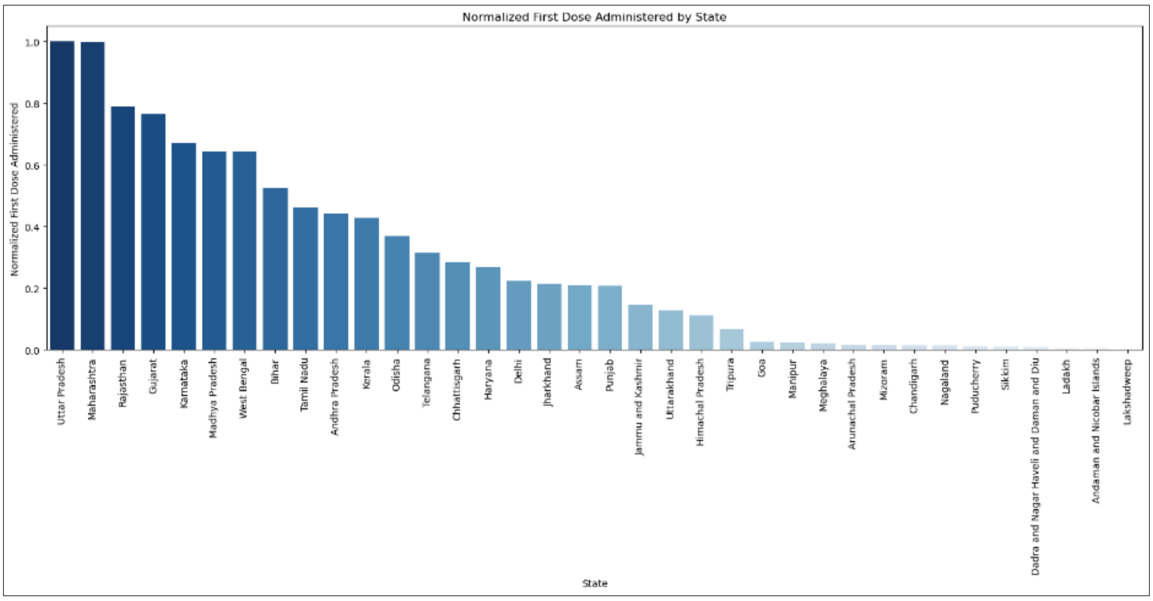
**Q1) Describe the dataset**

```
df.describe()
```

| axin<br>)ses<br>red) | CoviShield (Doses Administered) | Sputnik V (Doses Administered) | AEFI | 18-44 Years (Doses Administered) | 45-60 Years (Doses Administered) | 60+ Years (Doses Administered) | 18-44 Years(Individuals Vaccinated) | 45-60 Years(Individuals Vaccinated) | 60+ Years(Individuals Vaccinated) | Total Individuals Vaccinated |
|---|---|---|---|---|---|---|---|---|---|---|
| e+03 | 7.415000e+03 | 2914.000000 | 5291.000000 | 1.656000e+03 | 1.656000e+03 | 1.656000e+03 | 3.632000e+03 | 3.633000e+03 | 3.633000e+03 | 5.759000e+03 |
| e+05 | 4.175775e+06 | 4961.982498 | 585.480060 | 4.508783e+06 | 3.824424e+06 | 2.899150e+06 | 7.166336e+05 | 1.498468e+06 | 1.350096e+06 | 2.336581e+06 |
| e+05 | 6.931911e+06 | 13325.990621 | 801.540814 | 4.980031e+06 | 4.084621e+06 | 3.204826e+06 | 1.163574e+06 | 1.853811e+06 | 1.615079e+06 | 3.865512e+06 |
| e+00 | 7.000000e+00 | 0.000000 | 0.000000 | 2.662400e+04 | 1.681500e+04 | 9.994000e+03 | 1.059000e+03 | 1.136000e+03 | 5.580000e+02 | 7.000000e+00 |
| e+00 | 1.256235e+05 | 0.000000 | 106.000000 | 4.302122e+05 | 2.283322e+05 | 1.250870e+05 | 5.366975e+04 | 9.120600e+04 | 5.536300e+04 | 7.050550e+04 |
| e+03 | 6.917980e+05 | 0.000000 | 288.000000 | 2.929598e+06 | 2.607937e+06 | 1.722154e+06 | 2.801420e+05 | 7.373880e+05 | 7.670690e+05 | 3.720700e+05 |
| e+05 | 5.486316e+06 | 2029.500000 | 719.000000 | 6.698730e+06 | 6.256604e+06 | 4.884511e+06 | 8.317480e+05 | 2.355564e+06 | 2.072824e+06 | 3.096800e+06 |
| e+06 | 4.825801e+07 | 97389.000000 | 4706.000000 | 2.658372e+07 | 1.700704e+07 | 1.272274e+07 | 9.213328e+06 | 9.408647e+06 | 7.481940e+06 | 2.478224e+07 |

2. **The number of persons state-wise vaccinated for the first dose in India:**

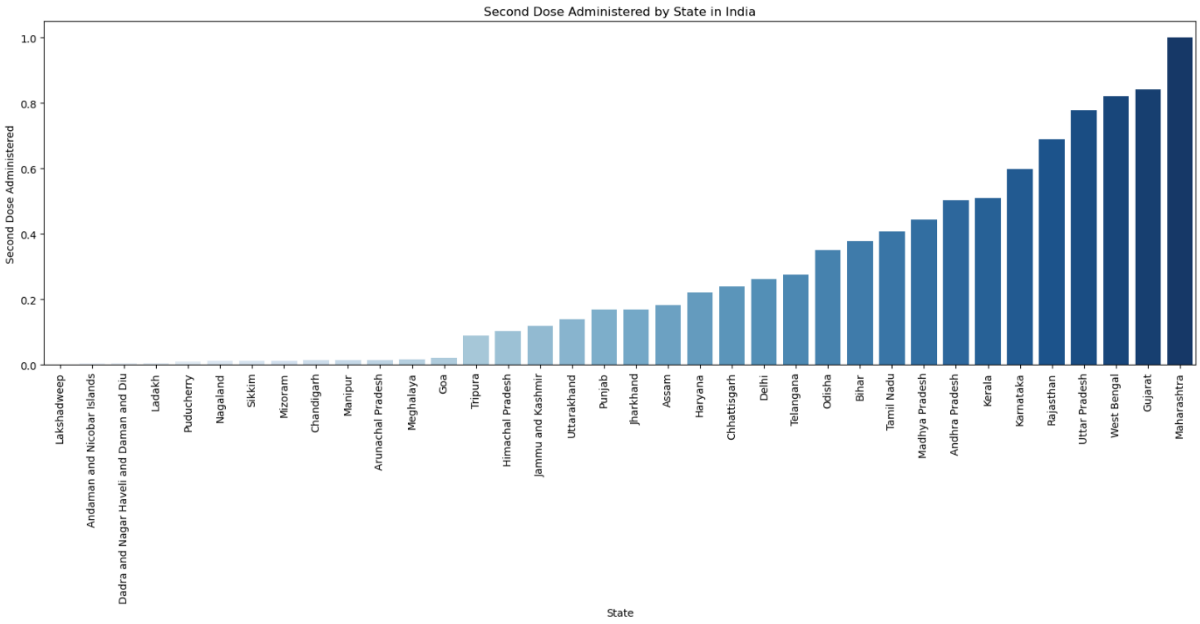
[Diagram: Bar chart showing the top three states with the highest number of first doses administered]



- Utter Pradesh and Maharahstra have the highest first doses administrated.
- Rajasthan and Gujarat have the 3rd and 4th highest first doses administrated.
- The lowest two First Dose Administered states are islands.

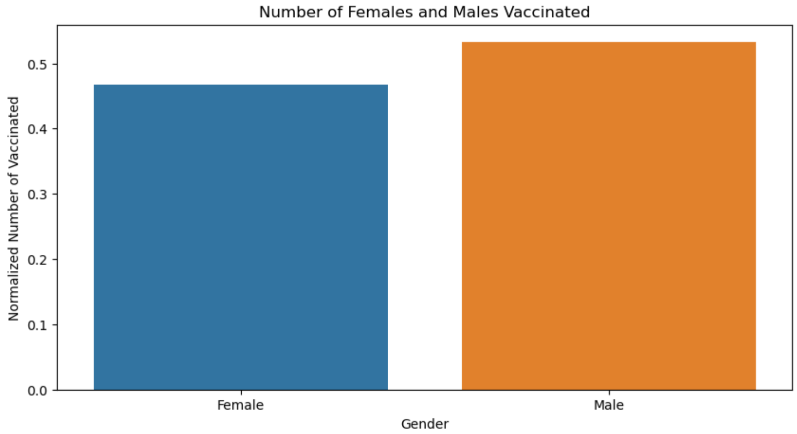
### 3. The number of persons state-wise vaccinated for the second dose in India:

- Maharashtra has the highest first doses administrated.
- Gujarat and West Bengal have the 3rd highest first doses administrated.
- It seems like in Utter Pradesh the first dose administration was more than the second dose administration.



4. **The number of males and females vaccinated:**

- The approximately 6.53% difference between the number of males and females vaccinated suggests a slight gender disparity in vaccination rates.



**Part 2:**

1. **Describe the dataset:**

covid_19_india.csv :

The given dataset contains information about COVID-19 cases in different states/union territories of India. Here is the description of the columns in the dataset:

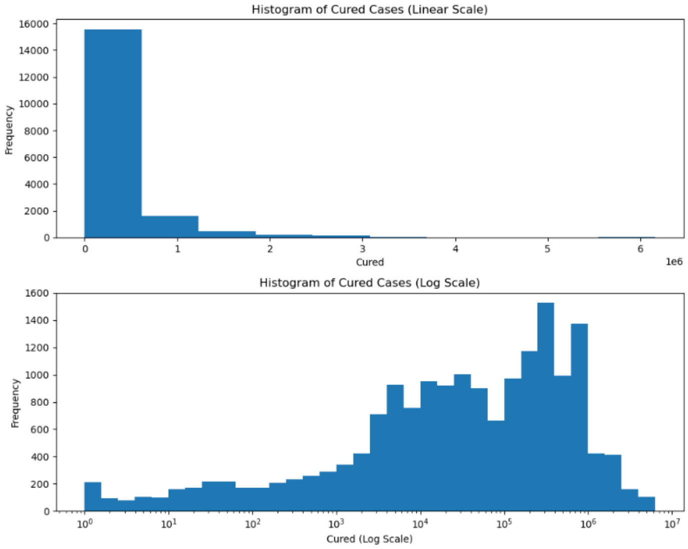
- Date: The date of the recorded data.
- Time: The time of the recorded data.
- State/UnionTerritory: The name of the state or union territory in India.
- ConfirmedIndianNational: The number of confirmed COVID-19 cases among Indian nationals in the respective state/union territory.
- ConfirmedForeignNational: The number of confirmed COVID-19 cases among foreign nationals in the respective state/union territory.
- Cured: The number of individuals who have recovered from COVID-19 in the respective state/union territory.
- Deaths: The number of deaths due to COVID-19 in the respective state/union territory.
- Confirmed: The total number of confirmed COVID-19 cases (including both Indian and foreign nationals) in the respective state/union territory.

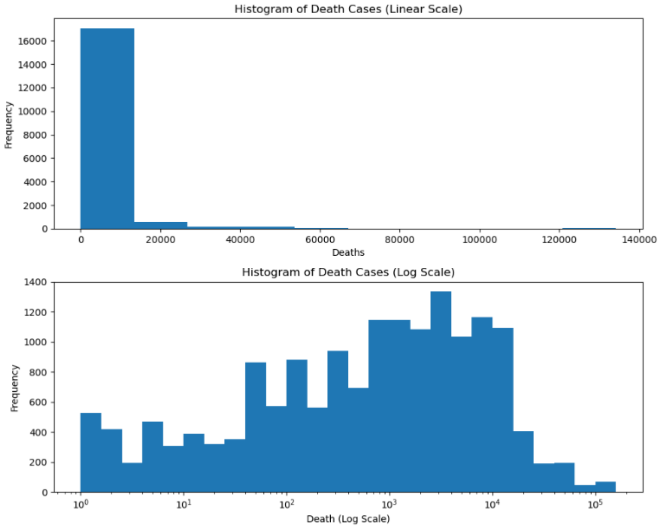
Regarding the data types:

- The columns "Date" and "Time" are represented as objects (likely strings) indicating the date and time values.
- The columns "ConfirmedIndianNational", "ConfirmedForeignNational", "Cured", "Deaths", and "Confirmed" are represented as integers.

The dataset consists of 18,110 entries (rows) and 8 columns. It provides a comprehensive record of COVID-19 cases, recoveries, and deaths in different states and union territories of India.

2. **Frequency Distribution of Cured and Death Cases :**





Both the histogram plots for the "Cured" and "Deaths" variables exhibit a left-skewed distribution. This implies that the frequency of cured individuals and deaths due to COVID-19 is unfortunately quite high. In other words, there is a significant number of instances where individuals have recovered from the virus or, unfortunately, succumbed to it.

Upon closer examination, it is evident that both plots share a similar trend or pattern. This similarity suggests a possible correlation between the number of cured cases and the number of deaths. It could indicate that as the number of individuals who recover from COVID-19 increases, the number of deaths also tends to increase. This finding raises important considerations for understanding the impact of the disease on the population and the effectiveness of medical interventions.

The left-skewed nature of the histograms signifies that the frequency of lower values (fewer cured cases or deaths) is higher compared to higher values. This implies that there may be a concentration of observations towards the lower end of the distribution, indicating a higher prevalence of cases with fewer cured individuals or higher mortality rates.

By analyzing and interpreting these histograms, we gain valuable insights into the distribution and frequency of cured cases and deaths due to COVID-19. This information can contribute to a deeper understanding of the impact of the disease and aid in developing appropriate strategies for prevention, treatment, and public health interventions.

3. **Calculating the ACF and PACF:**

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are statistical tools used in time series analysis to understand the correlation structure within a series.

Autocorrelation Function (ACF):

The Autocorrelation Function measures the correlation between observations in a time series at different lags. It helps identify the presence of any significant autocorrelation, which is the correlation between a variable's current value and its past values. ACF is calculated for a range of lags and plotted as a function of lag.

Key points about ACF:

ACF ranges from -1 to 1, where 1 indicates a perfect positive autocorrelation, -1 indicates a perfect negative autocorrelation, and 0 indicates no autocorrelation.

A positive autocorrelation at lag k suggests that a high value at time t is likely to be followed by a high value at time t + k.

A negative autocorrelation at lag k suggests that a high value at time t is likely to be followed by a low value at time t + k.

ACF can help identify the order of an autoregressive (AR) model by observing when autocorrelations drop off significantly.

Partial Autocorrelation Function (PACF):

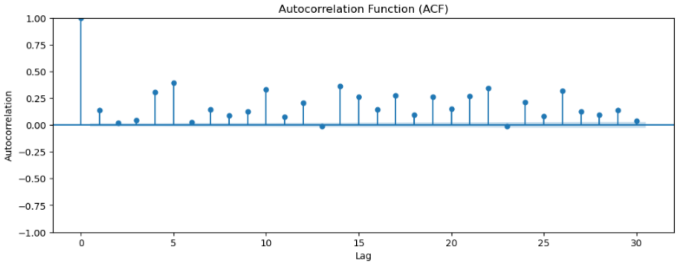
The Partial Autocorrelation Function measures the correlation between observations in a time series at different lags, while controlling for the correlation at shorter lags. It helps identify the direct relationship between an observation and its lags, excluding the influence of intermediate lags.

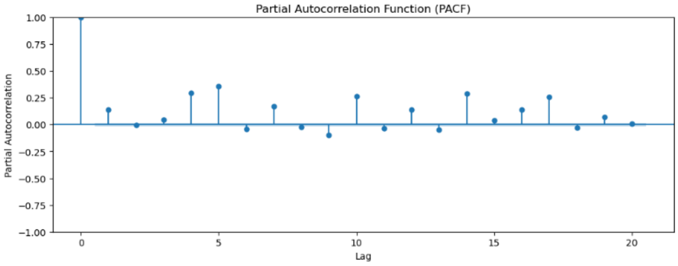
Key points about PACF:

PACF is useful in determining the order of a moving average (MA) or autoregressive moving average (ARMA) model.

The PACF at lag k represents the correlation between the current observation and its kth lag, accounting for the effects of intermediate lags.

A significant PACF value at lag k suggests that there is a direct relationship between the current observation and its kth lag, indicating a potential AR or MA term in the model.





When the lag 0 autocorrelations are significantly high in both the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, it indicates a strong correlation between the current observation and its immediately preceding observation.

This suggests that the number of Cured patients in the current time period is strongly influenced by the number of confirmed cases in the previous time period.

4. **ARIMA (Autoregressive Integrated Moving Average) Model:**

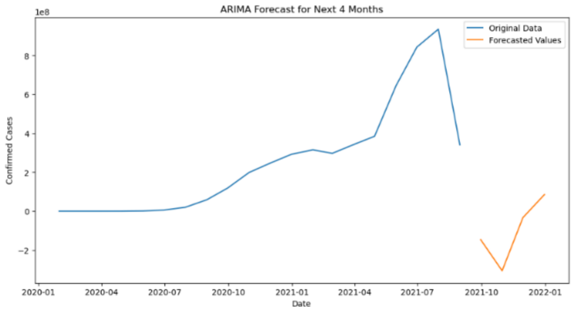
The ARIMA model is a popular time series forecasting method that combines autoregressive (AR), differencing (I), and moving average (MA) components. It is used to analyze and forecast time series data by capturing the underlying patterns and dependencies in the data.

Autoregressive (AR) Component: The AR component refers to the dependency of the current observation on the previous observations in a time series. It assumes that the current value of a variable can be predicted based on its past values. The order of the AR component (denoted as p) determines the number of lagged observations to consider for prediction.

Differencing (I) Component: The differencing component is used to remove the trend and make the time series stationary. Stationarity is important because it allows for more accurate modeling and forecasting. Differencing involves computing the difference between consecutive observations. The order of differencing (denoted as d) determines the number of times differencing is performed.

Moving Average (MA) Component: The MA component models the dependency of the current observation on the past prediction errors. It takes into account the error terms from the previous predictions. The order of the MA component (denoted as q) specifies the number of lagged forecast errors to include in the model.

The ARIMA model is commonly represented as ARIMA(p, d, q), where p, d, and q are the orders of the AR, differencing, and MA components, respectively. By estimating the optimal values of p, d, and q, the ARIMA model can be used to forecast future values based on historical data.



Model Evaluation for ARIMA Model:

1) Mean Absolute Error (MAE): 791140265.2061458

2) Mean Squared Error (MSE): 7.374849450751023e+17

3) Root Mean Squared Error (RMSE): 858769436.5049925

4) Mean Absolute Percentage Error (MAPE): 1.0936231315515619

5) R-squared (R2) Score: -13.23045134565778

The high errors and negative R2 score indicate that the model's predictions deviate significantly from the actual values and fail to explain the variance in the data. the model's performance based on these metrics is not satisfactory. The MAE, MSE, and RMSE values are very high, indicating a significant deviation between the predicted and actual values. The MAPE value is relatively low, suggesting that the percentage deviation is more reasonable.

5. **SARIMAX (Seasonal ARIMA with Exogenous Variables) Model:**

The SARIMAX model is an extension of the ARIMA model that incorporates seasonal components and exogenous variables. It is particularly useful for time series data that exhibit seasonal patterns and are influenced by external factors.

In addition to the AR, I, and MA components of the ARIMA model, SARIMAX includes the following:

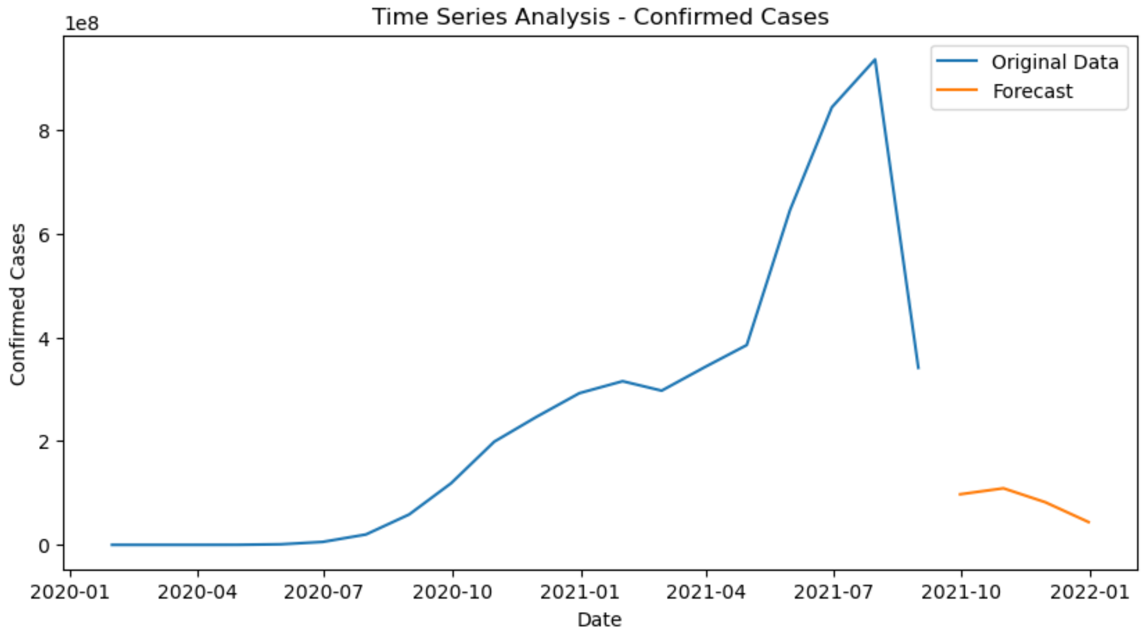
Seasonal AR Component: The seasonal autoregressive component captures the dependence of the current observation on past observations at seasonal lags. It takes into account the seasonal patterns in the data.

Seasonal MA Component: The seasonal moving average component considers the dependency of the current observation on past forecast errors at seasonal lags. It accounts for the seasonal variations in the errors.

Exogenous Variables: SARIMAX allows for the inclusion of exogenous variables that are not part of the time series but may have an impact on the variable being forecasted. These variables can provide additional information and improve the accuracy of the forecast.

The SARIMAX model is typically represented as SARIMAX(p, d, q)(P, D, Q, s), where the uppercase P, D, Q represent the seasonal orders of the AR, differencing, and MA components, respectively, and s denotes the seasonal period.

By incorporating seasonal components and exogenous variables, the SARIMAX model can capture more complex patterns and dependencies in the time series data, leading to more accurate and reliable forecasts.



Model Evaluation for SARIMAX Model:

1) Mean Absolute Error (MAE): 608045567.6980269

2) Mean Squared Error (MSE): 4.137706504420334e+17

3) Root Mean Squared Error (RMSE): 643250068.3575816

4) Mean Absolute Percentage Error (MAPE): 0.8757581665013704

5) R-squared (R2) Score: -6.984085842968504

## 6. **Comparing the model evaluation metrics**

The SARIMAX model outperforms the ARIMA model in terms of all evaluation metrics. It has lower values for MAE, MSE, RMSE, MAPE, and a less negative R-squared score, indicating better accuracy and performance in predicting the target variable.

**Results:**

1) Vaccination Trends Analysis:

- The top three states with the highest number of first doses administered are Uttar Pradesh, Maharashtra, and Rajasthan.
- Maharashtra has the highest number of second doses administered, followed by Gujarat and West Bengal.
- There is a slight gender disparity in vaccination rates, with approximately 6.53% more males vaccinated than females.

2) Time Series Analysis:

- The frequency distribution of cured and death cases exhibits a left-skewed distribution, indicating a higher prevalence of cases with fewer cured individuals or higher mortality rates.
- The ACF and PACF analysis shows a significant autocorrelation between the current number of cured patients and the number of confirmed cases in the previous time period.
- The ARIMA model's performance evaluation indicates high errors and a negative R-squared score, suggesting that the model's predictions deviate significantly from the actual values and fail to explain the variance in the data.
- The SARIMAX model incorporates seasonal components and exogenous variables and can be used to improve the accuracy of the forecast.

**Conclusion:**

In conclusion, this study on COVID-19 data analysis and forecasting provide valuable insights into vaccination trends and time series analysis. The analysis reveals the state-wise distribution of vaccinations in India, highlights a slight gender disparity, and explores correlations within the time series data. Although the ARIMA model showed unsatisfactory results, the SARIMAX model improved forecasting accuracy. These findings contribute to a comprehensive understanding of the pandemic and can inform public health policies and decision-making processes for effective pandemic management.