Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Exploratory and Predictive Data Analysis of COVID-19 Vaccination Data in India using ARIMA model

Dr. V. Sumathy[1], Dr. S.J. Rexline[2], Dr. G. Navamani[3], S. Sureha[4]

[1]Assistant professor, Department of Data Science, Loyola College, Chennai, India

[2]Assistant professor, Department of Computer Science, Loyola College, Chennai, India

[3]Associate professor, Department of Mathematics

Savitha School of Engineering, SIMATS, Chennai, India

[4]PG Research Scholar, Department of Data Science, Loyola College, Chennai

e-mail

Article Info
Page Number: 5585-5601
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

With the increase in Covid-19 patients throughout the world, the advent effect of vaccination plays an important role in improving the population immunity, controlling the effects of virus and reducing the health crisis. The major task in the planning of enhancing the vaccination drive lies in identifying the pattern of people coming forward to vaccinate in the age group expected to expose vaccination. The principal objective of this study is developing a model to predict the demand of vaccines in future pandemic situations in India using the ARIMA statistic model.

Keywords: -component; formatting; style; styling; insert (key words).

Introduction

The novel corona virus epidemic was first reported in January 27, 2020 in India. 34,822,040 people were infected with this disease, 480,860 people lost their lives and 34,258,778 recovered cases. The recovery case is around 98.6 % in India. The death rate in India is 1.4 %. India began administration of COVID-19 vaccines on 16 January 2021. Vaccines are the one of the significant factor to prevent the people from COVID-19 pandemic in India. 60.9 % of the Indian population haven taken the first dose vaccine. 42.5 % of the Indian population were fully vaccinated [11].

It is evidence-based and officially approved by health authorities is generally safe. Vaccination is a collective strategy that needs a high proportion of the population to be vaccinated in order to generate a protective effect [2]. Vaccination against COVID-19 is one of the critical tools to fight the Ongoing pandemic. Globally, it began on 31st December 2020, when WHO issued an Emergency Use Listing (EUL) for the Pfizer vaccine. In India, the Central Drugs Standard Control Organization (CDSCO), a regulatory body, has provided emergency use authorization to Covishield (AstraZeneca's / Serum Institute of India) and Covaxin (Bharat Biotech Limited) on 3rd January 2021[8][10]. Age wise vaccination drive has faced lot of challenges to bring people to get vaccinated.

Related Works

A great effort of research and global coordination, which has resulted in a rapid process of development of vaccines and other medical products considered strategic in the fight against COVID-19, is evident [4],[7]. Considering the initiatives for the rapid development of vaccines, Jason Wang and Petar Radanliev aims at identifying the main factors and innovative environments that are promoting this phenomenon [6][13]. It also seeks to understand the ways and processes that lead to the resolution of problems that plague our society, such as infectious diseases of global relevance, and, based on accumulated learning, to provide new perspectives of pathways and strategies that can be used for new vaccines within the scope of innovation management, especially in pandemic context [1][3]. Hugo Garcia Tonioli Defendi Analyzed the COVID-19 Vaccine Development Process[5]. Shantani Kannan discussed the Role of Artificial Intelligence and Machine Learning Techniques in the Race for COVID-19 Vaccine[12]. Md Mijanur Rahman carried out a study on Artificial Intelligence and Machine Learning Approaches in Confronting the Coronavirus (COVID-19) [9]. L. J. Muhammad designed the Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset [10]. Vikas Chaurasia did time series analysis for prediction COVID-19 pandemic [17]. Kolla Bhanu Prakash Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms and concluded that the Random Forest Regressor and Random Forest Classifier has outperformed other models in terms of CoD and Accuracy [7]. Stephen Wai Hang Kwok said that the level of positive sentiment among the public may not be sufficient to increase vaccination coverage to a level high enough to achieve vaccination-induced herd immunity. Rajani Kumari suggested the model to predict the economic losses in India due to Covid 19 [14].

The paper is organized as follows: Section 2 presents related works; Section 3 discusses the Problem Description and Research framework. Section 4 analyzes the given samples and finally Section 5 contains the Conclusion.

Problem Descriptions and Framework

The main objective of this paper is to create a model that provides the information about the requirement of vaccines against any pandemic situation in India. COVID-19 dataset for corona virus is collected based on confirmed, recovered, and death cases in India from January 2020 to July 2021. The dataset is taken from www.kaggle.com. The csv file downloaded and its details are given below.

State wise Testing Details.csv contains the fields.

Date, State, Total Samples, Negative and Positive. State, Total Doses Administered, Total Sessions Conducted, Total Sites, First Dose Administered, Second Dose Administered, Male (Individuals Vaccinated), Female (Individuals Vaccinated), Transgender (Individuals Vaccinated), Total Covaxin Administered, Total CoviShield Administered, Total Sputnik V Administered, AEFI,18-45 years (Age), 45-60 years (Age),60+ years (Age) and Total Individuals Vaccinated.

Covid_19_india.csv contains the below details.

Date, Time, State/Union Territory, Confirmed Indian National, Confirmed Foreign National, Cured, Deaths and Confirmed

The procedure applied to generate the prediction model is shown in Fig 1.

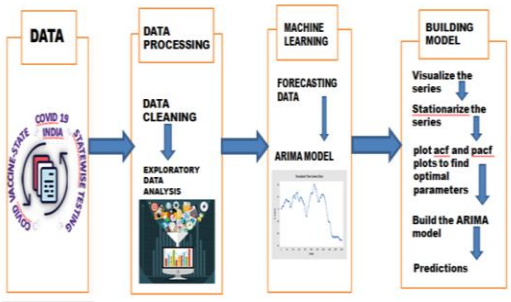


Fig. 1 - Schematic presentation of methodology applid for the study.

The collected data is pre-processed and cleaned to convert the raw data in to efficient data format. The final Data Set after Data pre-processing is shown in figure 2.

Date	State	Sessions	Total doses administered	1st dose administered	2nd dose Administered	Covaxin Administered	Covishield administered	18-44 age	44-60 age	60+ age	Total individuals vaccineted
24/06/2021	Jammu and kashmir	396332	4169680	3546680	623000	87546	4082134	962052	1595923	987144	3546680
1/05/2021	Himachal pradesh	63000	1823400	1562272	261128	2	1823398	92655	800172	669362	1562272
16/03/2021	Ladakh	3700	38926	35573	3353	0	38926	8051	7966	19534	35573
15/03/2021	Gujarat	499226	2772614	2208185	564429	409421	2362193	588839	572477	1044914	2208185
23/06/2021	Mizoram	49289	518600	465500	53100	0	518600	249477	132903	89281	465500

Fig. 2 - Data Set after Data pre-processing

Auto Regressive Integrated Moving Average Model (ARIMA) is used to create a predictive model for the above said pre-processed data. An ARIMA model is a class of statistical models for analyzing and forecasting time series data. A standard notation is used of ARIMA (p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.

d: The number of times that the raw observations are differenced, also called the degree of differencing.

q: The size of the moving average window, also called the order of moving average.

A linear regression model is constructed including the specified number and type of terms, and the data is prepared by a degree of differencing in order to make it stationary, i.e. to remove

trend and seasonal structures that negatively affect the regression model. A value of 0 can be used for a parameter, which indicates to not use that element of the model. This way, the ARIMA model can be configured to perform the function of an ARMA model, and even a simple AR, I, or MA model. Adopting an ARIMA model for a time series assumes that the underlying process that generated the observations is an ARIMA process. This may seem obvious, but helps to motivate the need to confirm the assumptions of the model in the raw observations and in the residual errors of forecasts from the model.

Analysis on Covid-19 India Dataset

From the seaborn package, a barplot has been plotted to visualize the race between the total number of confirmed cases in the State/Union territory and to the cured cases. Figure 3 shows the Race between confirmed cases and to the cured cases. In this below plot RED colour indicates the total number of confirmed cases and GREEN indicates the cured cases. From the below barplot, Maharashtra, Kerala, Karnataka are the states secured the top 3 places in the race between the recovered and the total number of cases

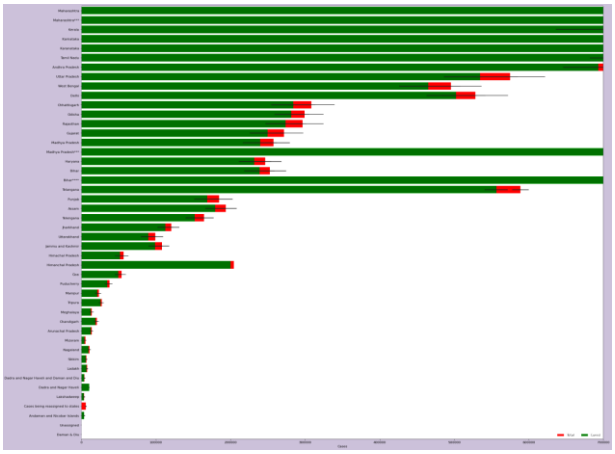
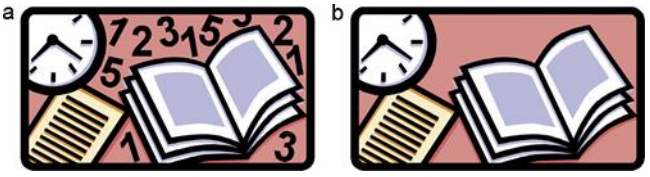


Fig. 3 - Race between confirmed cases and to the cured cases.

Trend of corona virus has been observed using the confirmed cases with the help of a scatter plot. Confirmed cases in India with respective to the date has been visualized and hence an increased trend is observed from MARCH 2020 till JULY 2021. Figure 4 shows the trend of Confirmed Cases in India and Figure 5 shows the cases in India on daily basis.



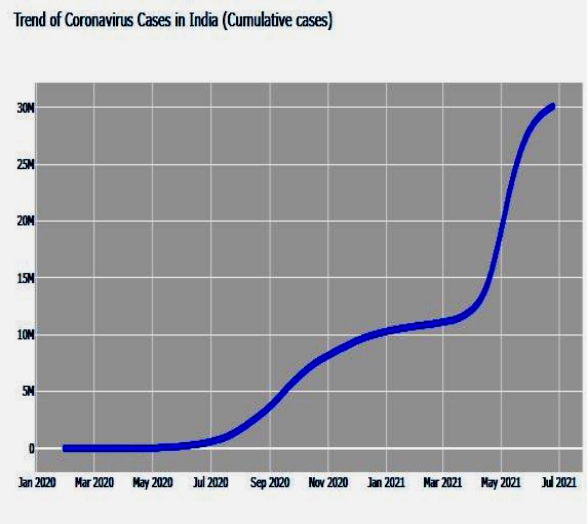


Fig. 4 - Confirmed Cases in India

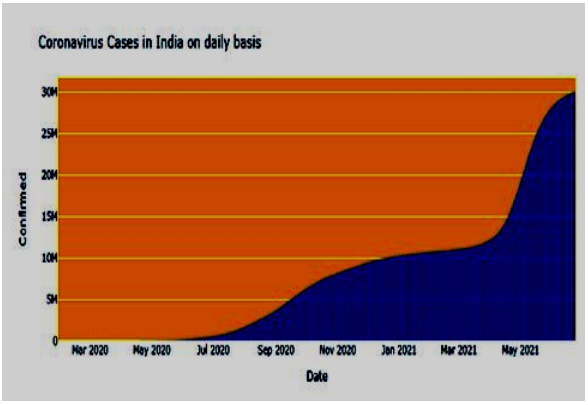


Fig. 5 - Cases in India on daily basis

Trend of death cases has been observed using the number of deaths with the help of a scatter plot. Figure 6 shows the Death Cases in India. Death cases in India with respective to the date has been visualized and hence an increased trend is observed from MARCH 2020 till JULY 2021. Figure 7 shows the Death Cases in India on daily basis.

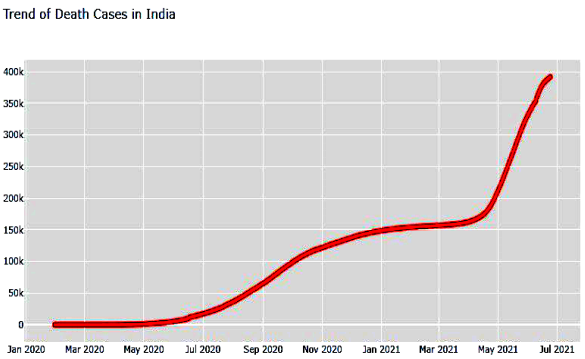


Fig. 6 - Death Cases in India

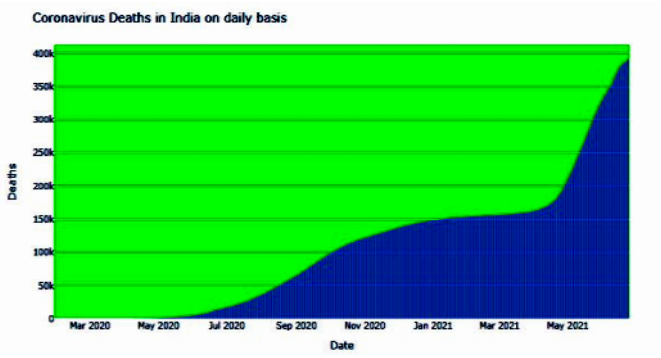


Fig. 7 - Death Cases in India on daily basis

A scatter plot has been plotted to visualize the combined trend of confirmed, recovered and death cases in India with respect to the date starting from JANUARY 2020. Figure 8 shows the confirmed, recovered and death cases in India.

Fig. 8 - confirmed, recovered and death cases in India

Figure 9 Shows the State wise confirmed, death, cured, active cases, Death rate(per 100),Cure rate (per 100) has been visualized in the descending order (i.e Highest number of cases).

	State/UnionTerritory	Confirmed	Deaths	Cured	Active	Death Rate (per 100)	Cure Rate (per 100)
24	Maharashtra	5987587	119303	5753290	124994	1.990000	95.930000
20	Kerala	2863287	12445	2729967	99835	0.440000	96.050000
19	Karnataka	2819465	34287	2668705	116473	1.220000	94.650000
34	Tamil Nadu	2443415	31746	2358785	52884	1.300000	96.540000
1	Andhra Pradesh	1862036	12452	1798380	51204	0.670000	96.580000
39	Uttar Pradesh	1704790	22336	1678788	3666	1.310000	98.470000
41	West Bengal	1487363	17475	1447510	22378	1.170000	97.320000
12	Delhi	1433366	24940	1406629	1797	1.740000	98.130000
8	Chhattisgarh	992074	13407	971057	7610	1.350000	97.880000
32	Rajasthan	951548	8905	940465	2178	0.940000	98.840000
29	Odisha	886946	3717	853012	30217	0.420000	96.170000
14	Gujarat	822758	10040	807911	4807	1.220000	98.200000
23	Madhya Pradesh	789499	8827	779177	1495	1.120000	98.690000
15	Haryana	767900	9314	756426	2160	1.210000	98.510000
4	Bihar	720505	9569	708231	2705	1.330000	98.300000
5	Bihar****	715730	9452	701234	5044	1.320000	97.970000
35	Telangana	616688	3598	596628	16462	0.580000	96.750000
31	Punjab	593572	15923	572008	5641	2.680000	96.370000

Fig. 9 - State wise confirmed, death, cured, active cases.

Fatality ratio is calculated for states using the number of death and confirmed cases. Figure 10 shows the Fatality Ratio among States. Visualization is done using the state/union

territory and the fatality ratio. From the below plot it has been observed, Maharashtra, Punjab, Gujarat are the three states with the highest fatality ratio.

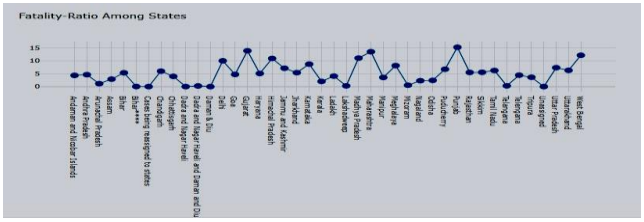


Fig. 10 - Fatality Ratio among States

Fatality ratio in a year is calculated and visualized using the features like Dates and fatality ratio. Figure 11 shows the Fatality Ratio in a Year From the below plot, it is observed that the fatality ratio has increased rapidly during the month of MARCH 2020 and has been quiet steady throughout the months from JULY 2020.

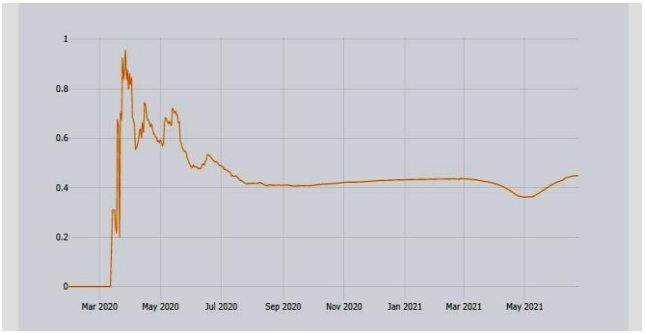


Fig. 11 - Fatality Ratio in a Year

Highest number of confirmed cases with respect to the state/union territory has been analyzed and visualized below. Figure 12 shows the top 10 States with Highest Number of Cases .The states with highest number of confirmed cases are Maharashtra, Kerala, Karnataka, Tamil Nadu, Andhra Pradesh, Uttar Pradesh, West Bengal, Delhi, Chhattisgarh and Rajasthan.

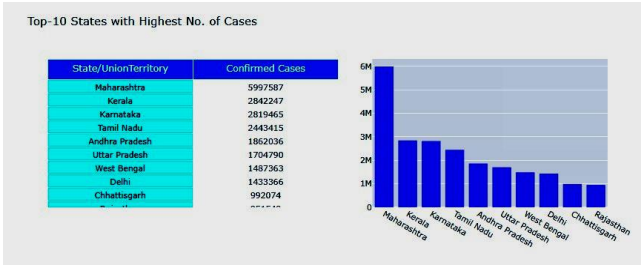


Fig. 12 - Top 10 States with Highest Number of Cases

Highest number of recovered cases with respect to the state/union territory has been analyzed and visualized below. Figure 13 shows the top 10 States with Highest Number of Cured Cases. The states with highest number of cured cases are Maharashtra, Kerala, Karnataka, Tamil Nadu, Andhra Pradesh, Uttar Pradesh, West Bengal, Delhi, Chhattisgarh and Rajasthan.

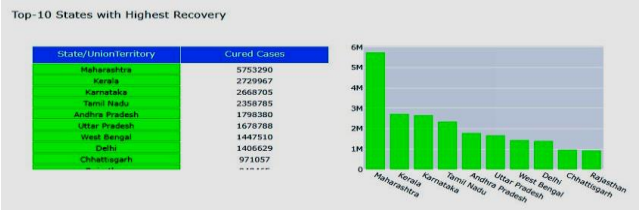


Fig. 13 - Top 10 States with Highest Number of Cured Cases

Highest number of death cases with respect to the state/union territory has been analyzed and visualized below. Figure 14 shows the top 10 States with Highest Number of Death Cases. The top 10 states with highest number of death cases are Maharashtra, Karnataka, Tamil Nadu, Delhi, Uttar Pradesh, West Bengal, Punjab, Chhattisgarh, Andhra Pradesh, and Kerala.

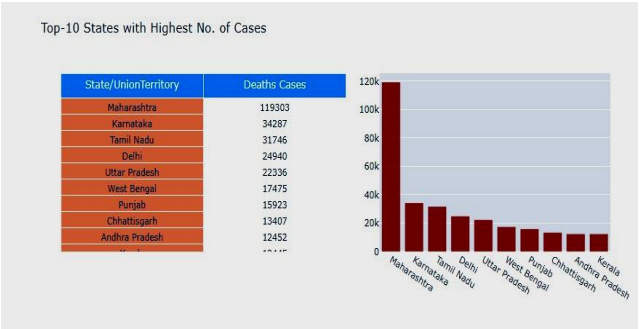


Fig. 14 - Top 10 States with Highest Number of Death Cases

State wise testing details have been analyzed and below 10 states are found with the highest number of sample collected for testing. The states are Uttar Pradesh, Maharashtra, Karnataka, Bihar, Tamil Nadu, Gujarat, Kerala, Andhra Pradesh, Delhi, and Telangana. Figure 15 shows the top 10 States with Highest Number of Sample Collection.

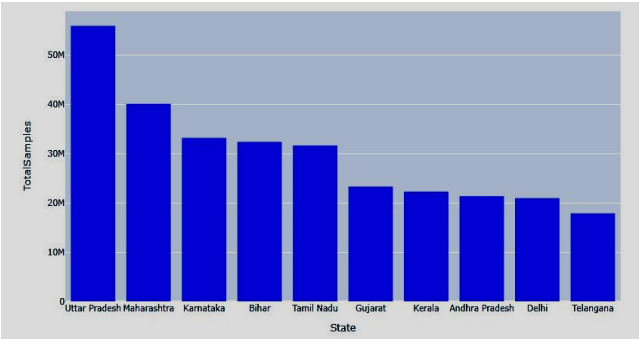


Fig. 15 - Top 10 States with Highest Number of Sample Collection

In order to observe the trend of how INDIA is vaccinating,a line plot has been plotted using the features like Date and Total number of individuals vaccinated. From the below plot, a steady increase in the vaccinating trend has been observed from JANUARY 2020 till JULY 2021. Figure 16 shows the total number of individuals vaccinated.

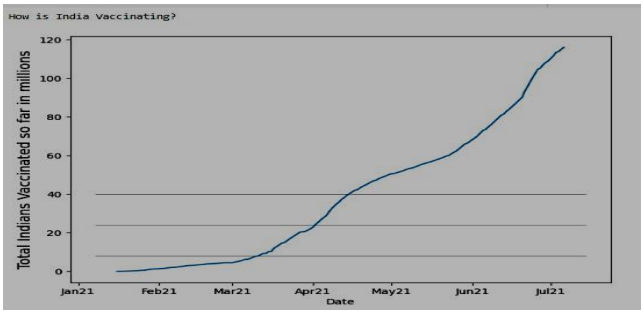


Fig. 16 - Total number of individuals vaccinated

A barplot has been plotted to visualize the top 5 vaccinated states in INDIA. Using the features States and total number of Individuals vaccinated, the top 5 states are Uttar Pradesh, Karnataka, Gujarat, Madhya Pradesh and Haryana. Figure 17 shows the top 5 Vaccinated States in India.

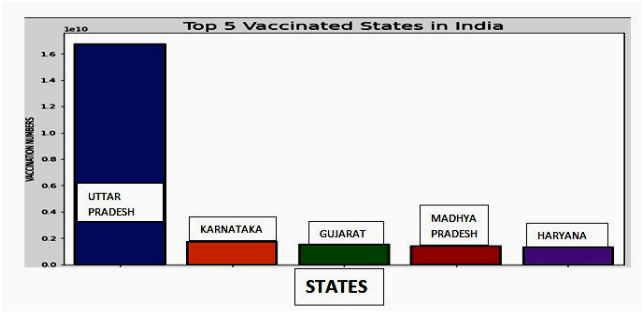


Fig. 17 - Top 5 Vaccinated States in India

A barplot has been plotted to visualize the least 5 vaccinated states in INDIA. Using the features States and total number of Individuals vaccinated, the least 5 states are Lakshadweep, Andaman and Nicobar Islands, Ladakh, Dadra and Nagar Haveli and Daman and Diu, Sikkim. Figure 18 shows Least 5 Vaccinated States in India.

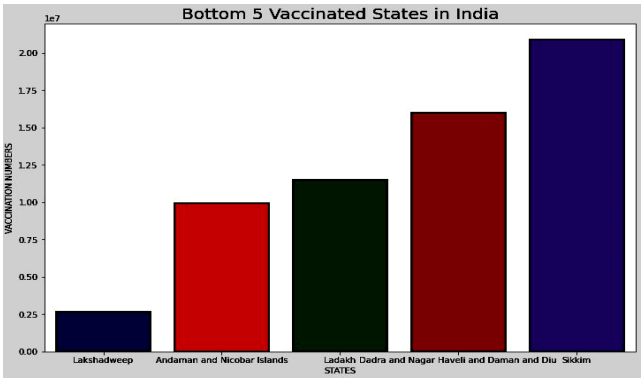


Fig. 18 - Least 5 Vaccinated States in India

Male, female, transgender vaccination ratio are calculated from the feature named INDIVIDUALS_VACCINATED. Using pie plot, gender vaccination ratio has been plotted .The ratio has been observed as follows: Male vaccinated ratio (53.7%), Female vaccinated ratio (46.28%) and Transgender vaccinated ratio (0.0173%). Figure 19 shows the Gender Vaccination Ratio.

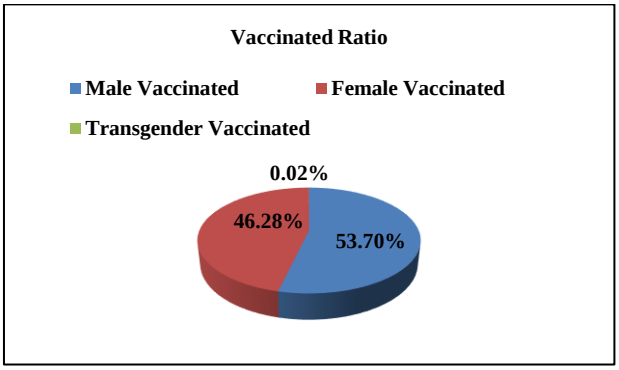


Fig. 19 - Gender Vaccination Ratio

From features (Total_covaxin_administered and Total_covishield_administered), covaxin and covishield administered ratio are 12.4% and 87.6% respectively. Figure 20 shows the covaxin and covishield administered ratio.

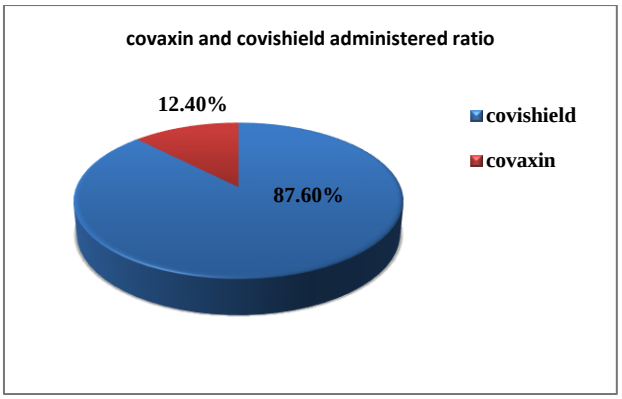


Fig. 20 - covaxin and covishield administered ratio

On analyzing the doses administered and individuals vaccinated, a ratio of 44.9% individuals vaccinated out of 55.1% doses administered. Figure 21 shows the Vaccine Status of States with Highest Number of Cases that representing the doses administered and individual vaccinated ratio. In Kerala, the ratio of Doses Administered and People vaccinated are 44.9% and 55.1% respectively. In Maharashtra, the ratio of Doses Administered and People vaccinated are 35.3% and 64.7%respectively. In Karnataka, the ratio of Doses Administered and People vaccinated are 44.5% and54.5 % respectively. In Tamil Nadu, the ratio of Doses Administered and People vaccinated are 45.2 % and54.5 % respectively.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} -$$

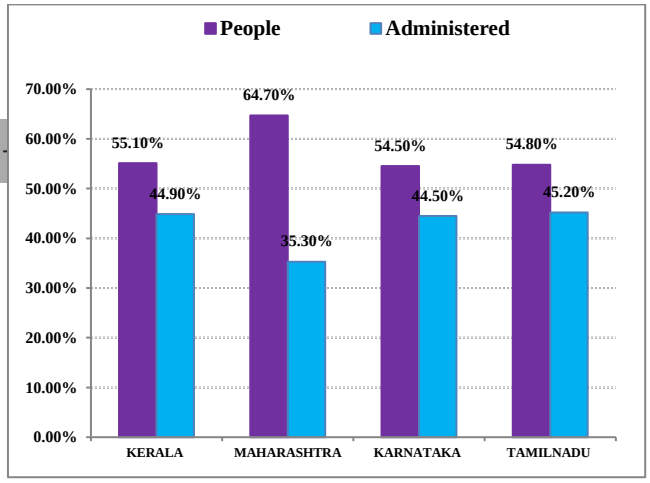
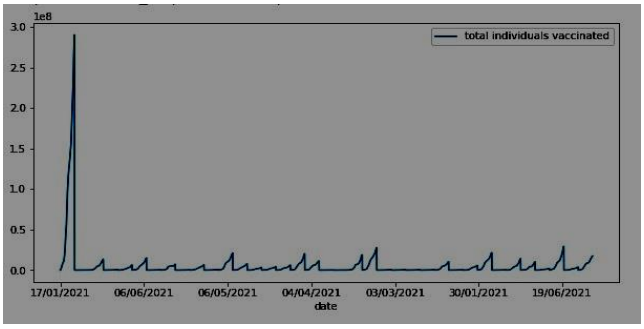


Fig. 21 - Vaccine Status of States with Highest Number of Cases

Arima Model Visualize The Time Series Data:

The data is plotted as a time series with dates along the x-axis and individuals vaccinated on the y axis. We can see that the covid vaccination dataset has a clear trend. This suggests that the time series is not stationary and will require differencing to make it stationarity, at least a difference order of 1.



Augmented Dickey fuller test is used to check the stationarity of the dataset.

Dickey-Fuller Test: A Dickey-Fuller test is a unit root test that tests the mull hypothesis that α=1 in the following model equation. Alpha is the coefficient of the first lag on Y. Null Hypothesis (H0): alpha=1

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e_t$$

where,

- y(t-1) = lag 1 of time series
- delta Y(t-1) = first difference of the series at time (t-1)

Fundamentally, it has a similar null hypothesis as the unit root test. That is, the coefficient of Y(t-1) is 1, implying the presence of a unit root. If not rejected, the series is taken to be non-stationary. The Augmented Dickey-Fuller test evolved based on the above equation and is

one of the most common forms of Unit Root test.

• Augmented Dickey Fuller (Adf) Test: As the name suggest, the ADF test is an 'augmented' version of the Dickey Fuller test. The ADF test expands the Dickey-Fuller test equation to include high order regressive process in the model.

$$y_t = c + \beta t + \alpha y_{t-1} + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} \ldots + \phi_p \Delta Y_{t-p} + e_t$$

If you notice, we have only added more differencing terms, while the rest of the equation remains the same. This adds more thoroughness to the test. The null hypothesis however is still the same as the Dickey Fuller test. A key point to remember here is: Since the null hypothesis assumes the presence of unit root, that is $\alpha=1$, the p-value obtained should be less than the significance level (say 0.05) in order to reject the null hypothesis. Thereby, inferring that the series is stationary.

Checking Stationarity Using Adfuller Test**:** Stationarity of the time series data has been checked using the augmented dickey fuller test.It has been observed that the p value is less than the significance level.Therfeore rejecting the null Hypothesis, and the Data is Stationary.

```
ADF Test Statistic : -7.035524393383377
p-value : 6.030311699701557e-10
#Lags Used : 0
Number of Observations : 6361
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data is stationary
```

P value is less than 0.05 implies there that the time series data is stationary.

Autocorrelation And Partial Autocorrelation: The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF) For example the ACF for a time series ytyt is given by:

Corr(yt,yt−k),k=1,2,....Corr(yt,yt−k),k=1,2,....

This value of k is the time gap being considered and is called the lag. A lag 1 autocorrelation (i.e., k = 1 in the above) is the correlation between values that are one time period apart. More generally, a lag k autocorrelation is the correlation between values that are k time periods apart.

The ACF is a way to measure the linear relationship between an observation at time t and the observations at previous times. If we assume an AR(k) model, then we may wish to only measure the association between ytyt and yt−kyt−k and filter out the linear influence of the random variables that lie in between (i.e., yt−1,yt−2,…,yt−(k−1)yt−1,yt−2,…,yt−(k−1)), which requires a transformation on the time series. Then by calculating the correlation of the transformed time series we obtain the partial autocorrelation function (PACF).

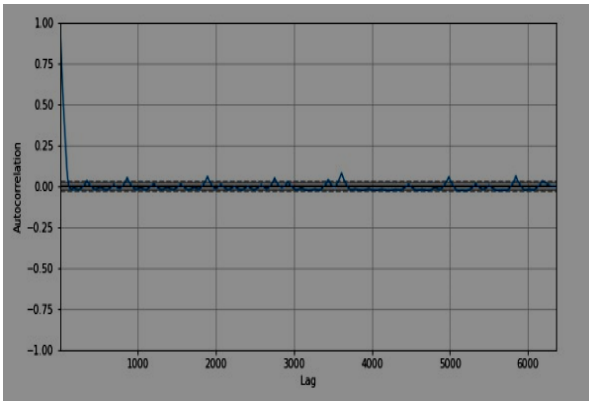


Fig. 22 - ACF PLOT

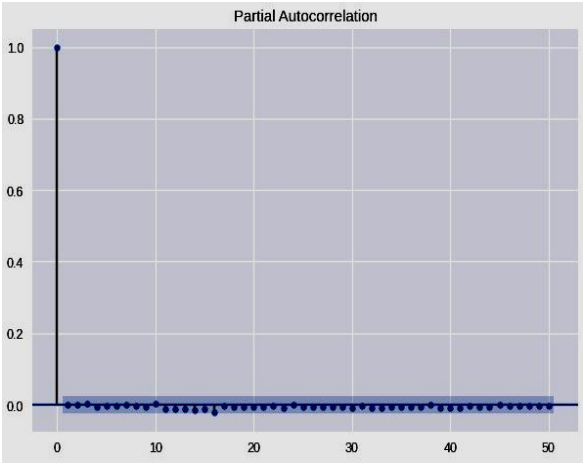


Fig. 23 - PACF PLOT

• MODEL BUILDING: The stats models library provides the capability to fit an ARIMA model. An ARIMA model can be created using the stats models library as follows:

1. Define the model by calling ARIMA() and passing in the p, d, and q parameters.

2. The model is prepared on the training data by calling the fit() function.

3. Predictions can be made by calling the predict() function and specifying the index of the time or times to be predicted.

An ARIMA model is fitted to the entire Covid Vaccination dataset and review the residual errors.An ARIMA (5,0 ,0)model is fitted. This sets the lag value to 5 for autoregression, uses a difference order of 0 because the time series is stationary, and uses a moving average model of 0.

• SUMMARY OF THE MODEL : Finally prints a summary of the fit model. This summarizes the coefficient values used as well as the skill of the fit on the in-sample observations.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:     D.total individuals vaccinated   No. Observations:                 6361
Model:                         ARIMA(5, 1, 0)   Log Likelihood             -105320.734
Method:                            css-mle   S.D. of innovations        3753960.124
Date:                     Wed, 22 Sep 2021   AIC                         210655.467
Time:                             18:18:30   BIC                         210702.773
Sample:                                  1   HQIC                        210671.847
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                          2723.4820   4.68e+04      0.058      0.954   -8.9e+04    9.45e+04
ar.L1.D.total individuals vaccinated     0.0001      0.013      0.012      0.991      -0.024       0.025
ar.L2.D.total individuals vaccinated    -0.0007      0.013     -0.059      0.953      -0.025       0.024
ar.L3.D.total individuals vaccinated     0.0046      0.013      0.365      0.715      -0.020       0.029
ar.L4.D.total individuals vaccinated    -0.0066      0.013     -0.526      0.599      -0.031       0.018
ar.L5.D.total individuals vaccinated    -0.0028      0.013     -0.222      0.824      -0.027       0.022
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            2.3127           -1.8270j            2.9473           -0.1064
AR.2            2.3127           +1.8270j            2.9473            0.1064
AR.3           -1.4562           -2.8307j            3.1833           -0.3256
AR.4           -1.4562           +2.8307j            3.1833            0.3256
AR.5           -4.0831           -0.0000j            4.0831           -0.5000
```

Fig. 24 - Summary of the Model

• Model Validation Using Walk Forward Validation

In time series modeling, the predictions over time become less and less accurate and hence it is a more realistic approach to re-train the model with actual data as it gets available for further predictions. Since training of statistical models are not time consuming, walk-forward validation is the most preferred solution to get most accurate results.

Predicted and estimated values after the walk forward validation are as follows

```
predicted=55114.599750, expected=55127.000000
predicted=55119.296432, expected=55372.000000
predicted=55368.511439, expected=55525.000000
predicted=55523.975047, expected=56300.000000
predicted=56300.569050, expected=56479.000000
predicted=56477.055752, expected=56509.000000
predicted=56510.172965, expected=56489.000000
predicted=56483.748402, expected=57385.000000
predicted=57381.691097, expected=58483.000000
predicted=58481.825806, expected=63319.000000
predicted=63324.625087, expected=67481.000000
predicted=67479.933176, expected=71502.000000
predicted=71515.226184, expected=73602.000000
predicted=73587.199027, expected=73831.000000
predicted=73810.261868, expected=80821.000000
predicted=80799.029368, expected=87233.000000
predicted=87212.379446, expected=94676.000000
predicted=94706.030436, expected=105054.000000
predicted=105044.818179, expected=117031.000000
predicted=117012.233308, expected=122168.000000
predicted=122155.793900, expected=122419.000000
predicted=122395.354093, expected=131756.000000
predicted=131690.438485, expected=139758.000000
predicted=139706.160323, expected=144492.000000
predicted=144533.907741, expected=156035.000000
predicted=156031.748846, expected=166665.000000
predicted=166627.556448, expected=170765.000000
predicted=170783.528167, expected=170856.000000
predicted=170838.517725, expected=181115.000000
predicted=181062.340427, expected=191344.000000
predicted=191314.229031, expected=199452.000000
predicted=199514.398387, expected=206116.000000
```

• Evaluating Forecasts: Forecasts can be evaluated using the test and the predictions. RMSE(Root Mean Squared error) is used to evaluate the forecasts. Test RMSE value of 915383.186 is obtained while evaluating the forecasts of the covid vaccination dataset.

```
Test RMSE: 915383.186
```

The model could use further tuning of the p, d, and maybe even the q parameters.

Arima Model Results

ARIMA MODEL RESULTS	
Model	ARIMA(5, 1, 0)
Method	Css- mle
No of observations	6361
Log likelihood	-105320.734
S.D of Innovations	3753960.124
AIC	210655.467
BIC	210702.773
HQIC	210671.847

	coef	Std err	z	p>|z|	[0.025	0.975]
const	2723.4820	4.68e+04	0.58	0.954	-8.9e+04	9.45e+04
ar.L1.D.total individuals vaccinated	0.0001	0.13	0.12	0.991	-0.024	0.025
ar.L2.D.total individuals vaccinated	-0.0007	0.13	-0.59	0.953	-0.025	0.024

	Real	Imaginary	Modulus	Frequency
AR 1	2.3127	-1.8270j	2.9473	-0.1064
AR 2	2.3127	+1.8270j	2.9473	0.1064
AR 3	-1.4562	-2.8307j	3.1833	-0.3256
AR 4	-1.4562	2.8307j	-3.1833	0.3256

Conclusion

AD fuller test confirms that the time series data under study is stationary which implies it time series with trends or with seasonality are not stationary. We have fit ARIMA (5,1,0) model. This sets the lag value to 5 for auto regression, uses difference order of 1 to make the time series stationary and uses a moving average model of 0.The expected values are compared with rolling forecast predictions and it is seen that most of the trend are in correct scale. The RMSE value of 915383.186 can be used as a point of comparison for other ARIMA configurations. Also AIC and BIC values act as a point of comparison to choose a better ARIMA model. Thus, though the number vaccinated doesn't follow any trend of seasonality the stationarity of data is maintained and a suitable predictive model is fitted.as normal.

References

1. Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, X Patricia B. Munroe, Bina Joe, and X Xi Cheng, "Artificial intelligence and machine learning to fight COVID-19," Physiol Genomics, vol. 52, pp. 200–202, 2020.
2. Arun Velu, and Pawan Whig, "Studying the Impact of the COVID Vaccination on the World Using Data Analytics," Vivekananda Journal of Research, vol. 10, no. 1, pp. 2319–8702, 2021.

3. Himanshu Gupta, Saurav Kumar, Drishti Yadav, Om Prakash Verma, Tarun Kumar, Sharma Chang Wook Ahn, and Jong-Hyun Lee, "Data Analytics and Mathematical Modeling for Simulatinghe Dynamics of COVID-19 Epidemic—A Case Study of India," Mdpi Journal Electronics, 2021.
4. H. Liyanage, S. de Lusignan, S-T. Liaw, C. Kuziemsky, F. Mold, P. Krause, and D. Fleming Jones, "Big Data Usage Patterns in the Health Care Domain: A Use Case Driven Approach Applied to the Assessment of Vaccination Benefits and Risks," IMIA Yearbook of Medical Informatics, 2014.
5. Hugo Garcia Tonioli Defendi, Luciana da Silva Madeira, and Suzana Borschiver, "Analysis of the COVID-19 Vaccine Development Process: an Exploratory Study of Accelerating Factors and Innovative Environments," Journal of Pharmaceutical Innovation, 2021.
6. C. Jason Wang, Y. Ng. Chun, and Robert H. Brook, "Response to COVID-19 in Taiwan Big Data Analytics, New Technology, and Proactive Testing," Journal of American Medical Association, vol. 323, 2020.
7. Kolla Bhanu Prakash, S. Sagar Imambi, Mohammed Ismail, T Pavan Kumar, and YVR Naga Pawan, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," International Journal of Emerging Trends in Engineering Research, vol. 8, no. 5, pp. 2347–3983, 2020.
8. L. J. Muhammad, A. Ebrahem, Algehyn, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay Chakraborty, and A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology DatasetSN Computer Science," vol. 2, no. 11, 2021.
9. Md. Martuza Ahamad, Sakifa Aktar, Md. Jamal Uddin, Md. Rashed-Al-Mahfuz, A.K.M. Azad, Shahadat Uddin, Salem A. Alyami, Iqbal H. Sarker, Pietro Lio, Julian M.W. Quinn, and Mohammad Ali Moni, "Adverse effects of COVID-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and post-vaccination reactogenicity," 2021.
10. Md Mijanur Rahman, Fatema Khatun, Ashik Uzzaman, Sadia Islam Sami, Md Al-Amin Bhuiyan, and Tiong Sieh Kiong, "A Comprehensive Study of Artificial Intelligence and Machine Learning Approaches in Confronting the Coronavirus (COVID-19) Pandemic," International Journal of Health Services, 2021.
11. Merryn Voysey, Sue Ann Costa Clemens, Shabir A Madhi, Lily Y Weckx, Pedro M Folegatti, Parvinder K Aley, Brian Angus, and Vicky L Baillie, "Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil," South Africa, and the UK, vol. 397, 2021.
12. Nasiba M. Abdulkareem, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Dathar A. Hasan, "COVID-19 World Vaccination Progress Using Machine Learning Classification Algorithms," Qubahan Academic Journal, pp. 2709–8206, 2021.
13. Petar Radanliev, David De Roure, and Rob Walton, "Data mining and analysis of scientific research data records on Covid19 mortality, immunity, and vaccine development - In the first wave of the Covid-19 pandemic," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, pp. 1121–1132, 2020.

14. Rajani Kumari, Sandeep Kumar, Ramesh Chandra Paonia, Vijander Singh, Linesh Raja, Vaibhav Bhatnagar, and Pankaj Agarwal, "Analysis and Predictions of Spread, Recovery, and Death Caused by COVID-19 in India," Big Data Mining And Analytics, vol. 4, no. 2, pp 2096–0654, 2021.
15. Samuel Lalmuanawma, Jamal Hussain, and Lalrinfela Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," Elsevier Public Health Emergency Collection, 2020.
16. Sarvam Mittal, "An Exploratory Data Analysis of COVID-19 in India," International journal of Engineering Research & Technology, vol. 9, no. 4, pp. 2278–0181, 2020.
17. Vikas Chaurasi, and Saurabh Pal, "Application of machine learning time series analysis for prediction COVID-19 pandemic," Research on Biomedical Engineering, 2020.