Introduction to Data Science Project

Covid Vaccination Drive in India

(Forecasting Next 50 days Vaccination Trend)

Submitted By
Kavinkumar M.
M22AI565
MTECH DCS

OBJECTIVE

In the current situation of India , the daily covid cases are very much reduced. It's commonly observed that nowadays people are ignoring the Covid vaccination. Lot of people are ignoring the Dose 2 or booster doses. We are almost at the End of the Year 2022. There are only less than 50 days remaining to Enter into 2023. It's necessary and useful to do some Analysis to forecast how many people will be fully vaccinated by the end of 2023.

HOW MANY PERCENTAGES OF PEOPLE WILL BE FULLY VACCINATED IN END OF 2022?

To answer the above question , a time series analysis performed to forecast the Vaccination rate per 100 people by the end of 2022

DATA COLLECTION AND PREPARATION

We start by collecting different datasets required for the analysis.

1. **Covid - Vaccination Data**

- https://ourworldindata.org/ collecting this information in the daily basics from the gov of India Twitter handle and stored in the GITHUB repository.We used this information for our analysis. This Dataset has collected the Information up to the current date.

- OWID Covid Vaccination India Dataset https://github.com/owid/covid-19-data/blob/master/public/data/vaccinations/country_data/India.csv

location	date	vaccine	source_url	total_vaccinations	people_vaccinated	people_fully_vaccinated	total_boosters
India	2021-01-15	Covaxin, Oxford/AstraZeneca	https://twitter.com/MoHFW_INDIA/status/1350459...	0	0	0	NaN
India	2021-01-16	Covaxin, Oxford/AstraZeneca	https://twitter.com/MoHFW_INDIA/status/1350459...	191181	191181	0	NaN
India	2021-01-17	Covaxin, Oxford/AstraZeneca	https://twitter.com/MoHFW_INDIA/status/1350815...	224301	224301	0	NaN
India	2021-01-18	Covaxin, Oxford/AstraZeneca	https://www.mohfw.gov.in/	454049	454049	0	NaN
India	2021-01-19	Covaxin, Oxford/AstraZeneca	https://www.mohfw.gov.in/	674835	674835	0	NaN

Covid Vaccination Dataset

Data Preparation:

- After loading the data, we just had to remove/clean some unwanted text present in the github raw dataframe
- There were missing values available in the booster dose feature of the dataset. These missing values are handled by imputation

- There were columns required for data type conversion in the dataset , these conversion handled properly

date	total_vaccinations	people_vaccinated	people_fully_vaccinated	total_boosters
2021-01-15	0	0	0	0
2021-01-16	191181	191181	0	0
2021-01-17	224301	224301	0	0
2021-01-18	454049	454049	0	0
2021-01-19	674835	674835	0	0

Covid Vaccination - Data Preparation

2. Population Of India

- To identify the Population of India in the year 2022, We have gathered the indian population data set for Kaggle
- Indian Population Dataset :

  https://www.kaggle.com/code/mattop/population-of-india-1950-2022-eda/data

- Only the year 2022 Indian Population count is required for our analysis. Remaining all the data available in this dataset is not much useful for our objective this time series analysis.

Data Preparation:

- Based on the above two dataset fully vaccinated per hundred people is calculated.
  Below is a list of features for this Covid Vaccination Time series analysis.

date	total_vaccinations	people_vaccinated	people_fully_vaccinated	total_boosters	people_fully_vaccinated_per_hundred
2021-01-15	0	0	0	0	0.00
2021-01-16	191181	191181	0	0	0.00
2021-01-17	224301	224301	0	0	0.00
2021-01-18	454049	454049	0	0	0.00
2021-01-19	674835	674835	0	0	0.00
...	...	...	...	...	...
2022-11-07	2197295587	1026910508	950285563	220099516	67.06
2022-11-08	2197435518	1026919862	950315704	220199952	67.06
2022-11-09	2197513103	1026925870	950336542	220250691	67.06
2022-11-10	2197646336	1026934310	950369833	220342193	67.06
2022-11-11	2197783974	1026943380	950403228	220437366	67.06

Covid Vaccination Fully Vaccinated People per hundred Individuals

DATA ANALYSIS & VISUALISATION

1. **Tools and Methods Used:**

Libraries and their functions

Pandas

- We used different functions from the pandas library to perform the data preparation
- The **read_csv()** function reads the data from the csv file into python variables so that we can apply different operations on the variable.

Matplotlib

- Used for data visualization

Platform

- Google Collab - To run the python code (RAM - 0.86 GB/ 12.69 GB).
- Google Collab ipython Notebook : https://colab.research.google.com/drive/1w_9omzYnsJe2sSQsTwEhhg5fDwY-ICvP?usp=sharing

2. **Exploratory Data Analysis:**

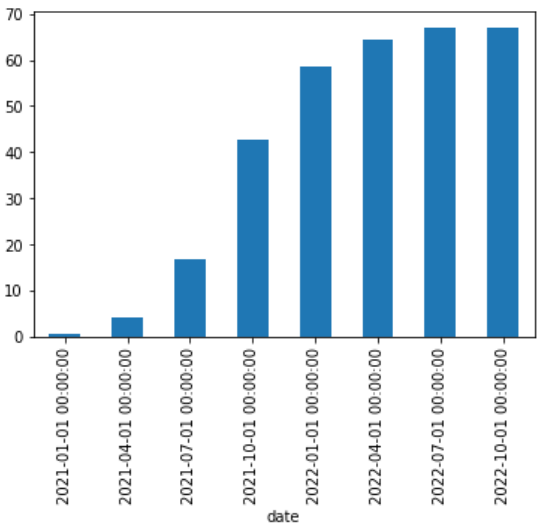
- This dataset contains the historical information of the vaccination information starting from the 01/15/2021 to current date
- Since our objective is forecast the next 50 days trend, so the recent data is very much useful for analysis
- Before that we need to analyze the trend and smoothness of the data.

2.1 Data Trend:

- Time series data is the data collected from various timestamps
- Trend of the time series data is to understand the pattern of the data like upward trend/ downward trend / seasonal trend etc.
- We have plotted the trend of the data using the pandas resample method with Quarterly and monthly rules.
- The resulting plots indicated that India Covid Vaccination data is upward trend data (Data trend is only increasing pattern)
- Also we can observed from the plot, the vaccination trend is not relatively increased in the last few quarters
- Q2,Q3 of Year 2021 and Q1 of Year 2022 results indicating that vaccination trend increased in the Exponential

```
df['people_fully_vaccinated_per_hundred'].resample(rule='QS').max().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc546db7950>
```



2.2 Data Smoothness:

- Time series data usually contains multiple sharp edges because it was connected for multiple series of timestamps. This verification helps to find any seasonality trend available in the data.(seasonal trend is the same pattern follows each season)
- To identify these kinds of patterns we are plotting varios Moving average models.
- Example : Simple moving average
    1. rolling() function will helps to identify find the average vaccination rate with given window period
    2. We are experimenting with multiple rolling windows like 10 days, 30 days, 50 days, 75 days

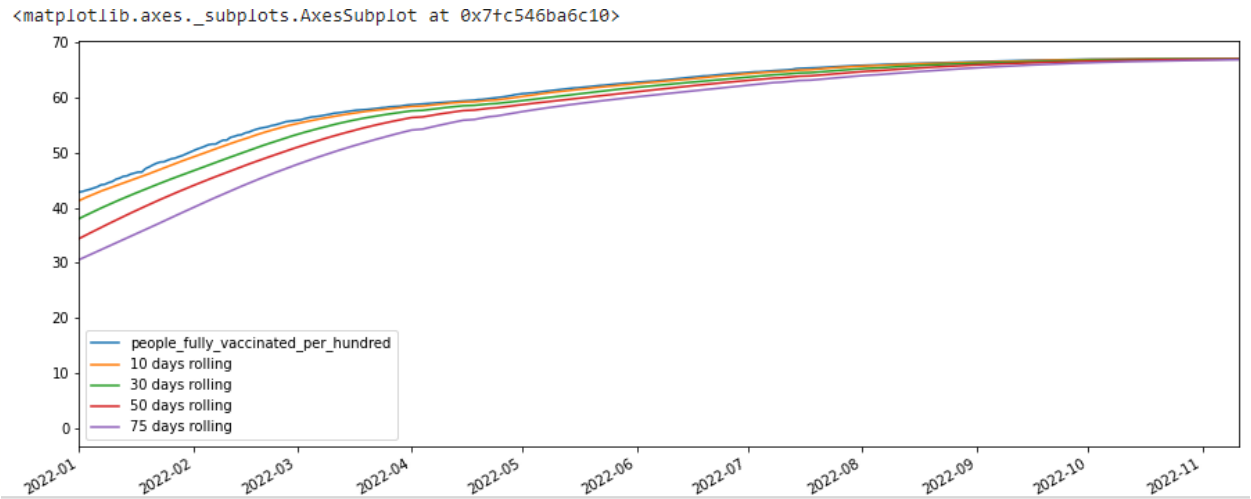
Simple Moving Average

```
[ ]  df_covid['30 days rolling']=df_covid['people_fully_vaccinated_per_hundred'].rolling(30).mean()
```

```
[ ]  df_covid['10 days rolling']=df_covid['people_fully_vaccinated_per_hundred'].rolling(10).mean()
```

```
[ ]  df_covid['50 days rolling']=df_covid['people_fully_vaccinated_per_hundred'].rolling(50).mean()
```

```
[ ]  df_covid['75 days rolling']=df_covid['people_fully_vaccinated_per_hundred'].rolling(75).mean()
```

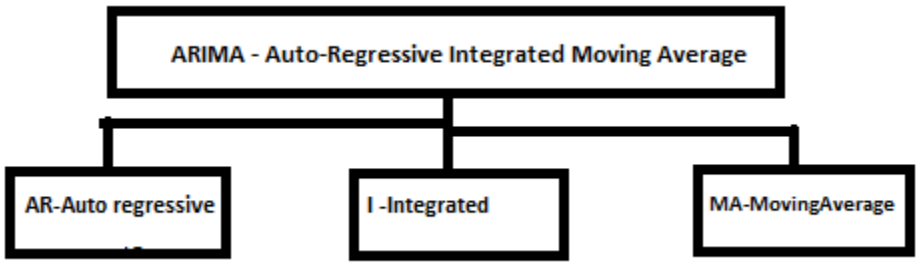


- These graphs indicate that there are no peaks and valleys in the data and there is no seasonality trend available in the data.
- We have completed all the necessary analysis. Next we have to start with Machine learning models

Model Building

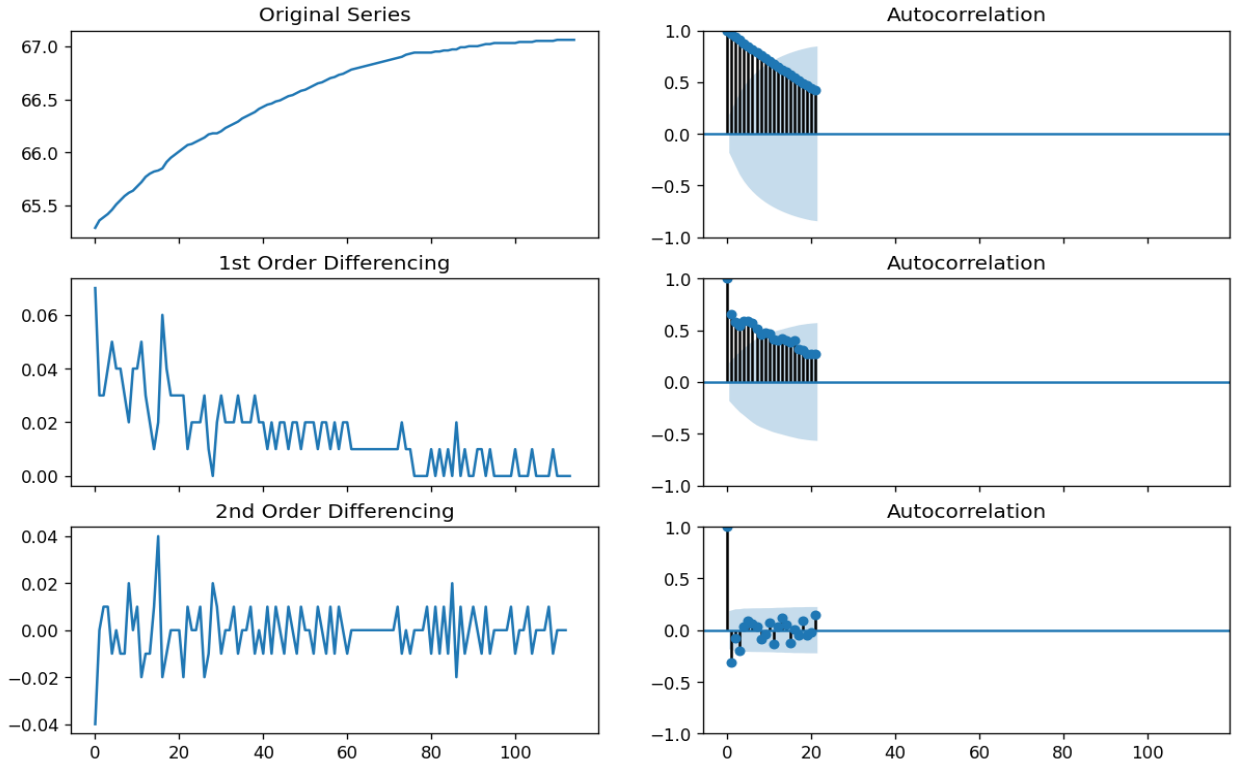
1. ARIMA Model

- Our Objective is to forecast the next 50 days vaccination trend in India.
- To forecast the trend we can use the ARIMA model

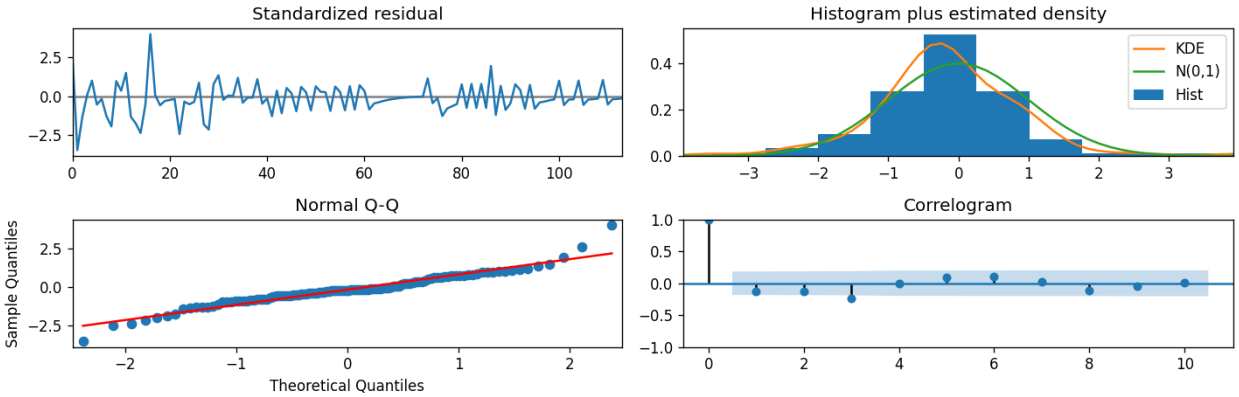


- AR Terms : Past important values in the time series . These values are easily obtained from Autocorrelation plots. Generally it is denoted the variable p . It means Plot having the boundary lines(Upper and lower) , we have to count how many the data crossing the boundary lines is captured in this variable
- I Term : Generally it is denoted the variable d. To accurately forecast the future values we need to convert the Data as stationary. It already confirmed that this data is not seasonal data. Our next task is to make the data as stationary means we have to remove the dependency of time , zero mean and zero standard deviation. This is done by a differentiation operation.Differenctionation means successive subtraction from past values with current values
- Sometimes single differentiation is not sufficient. We need multiple times to differentiate the results until the series becomes stationary.

- But in general practice we have to use the Augmented Dickey–Fuller statistical test to check if the time series is stationary or not.
- Dickey–Fuller statistical test :
  H0 : Data is Stationary
  H1 : Data is Not Stationary.
- Using 5% Confidence interval ADF test confirm the time series is stationary or not
- As a conclusion d value is counted how many times of differentiation makes time series stationary.
- MA Terms : Generally it is denoted the variable q. It indicates the forecasting errors. It also gathered from the Auto regressive plots.
- ARIMA(p,d,q) means ARIMA model without seasonality.

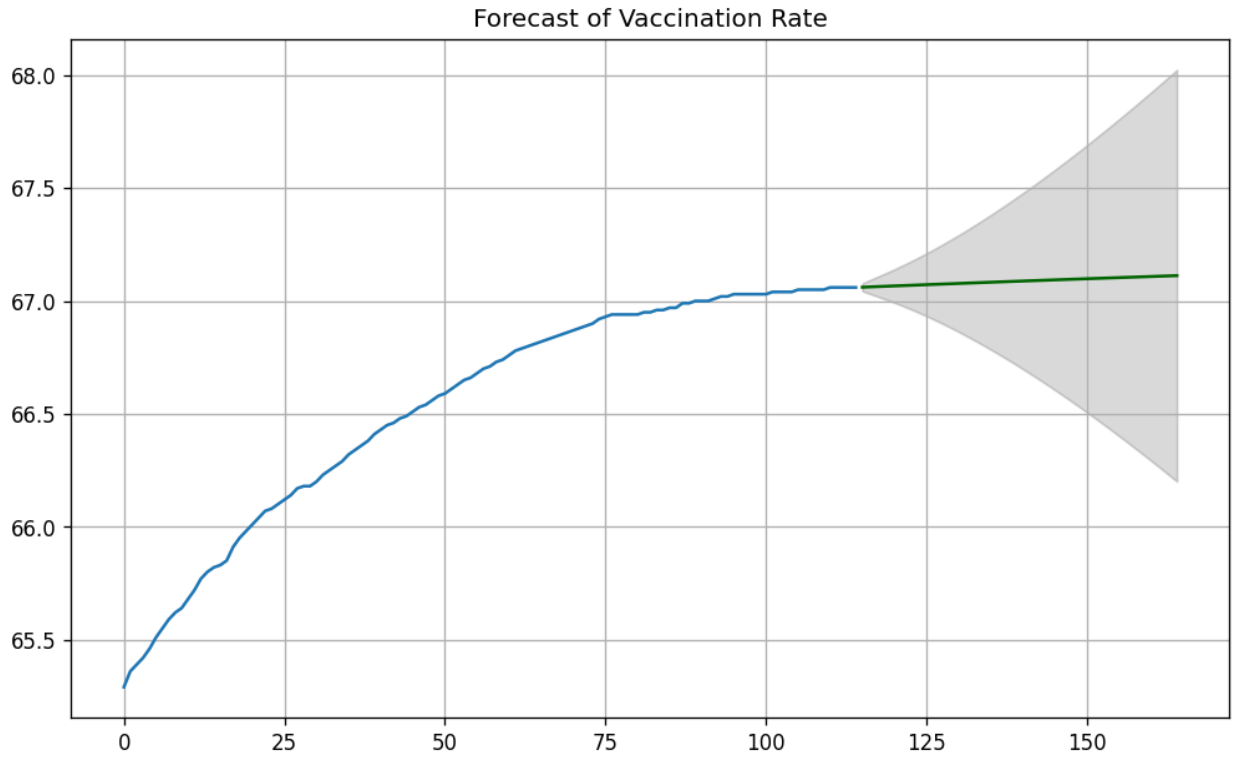


- Best suited model for our data is ARIMA(1,1,3)
- Model diagnosis test plots indicated that real and predicted values matching



2. Forecasting

- We have completed all the necessary steps. Next steps we can forecast the results for the next 50 days.
- Below are the covid vaccination results for the next 50 days. Currently the vaccination rate is 67.06%



Results

- Our forecast results indicate that we never expect much increase in the fully vaccinated rate.
- Also the plots show that vaccination rates .not increased much in last few months

Challenges

- There is no such reliable dataset available to combine vaccinated people and the Indian population.(fully vaccinated people %, 1 dose vaccinated people %, booster dose vaccinated people %)

Future Work

- We have to continue the same kind of analysis / forecast the result for the 1 dose vaccination % and booster dose vaccination %

References

1. https://www.youtube.com/watch?v=vhI0Nr1hHCY&list=PLZoTAELRMXVNty3jyJkYXuyQY3IMSpr3b
2. https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/