

COVID-19 VACCINES ANALYSIS

Performing exploratory data analysis





Table of contents



01

ABSTRACT

02

INTRODUCTION

03

**MATERAILS AND
METHOD**

04

**STATISTICAL
ANALYSIS**

05

VISUALIZATION

06

RESULT





ABSTRACT

Goal of this research is to analyze data on vaccinations, vaccination administration, and forecasting vaccination rates on a country-by-country basis for the general public, policymakers, vaccine manufacturers, national governments, and international governments to better understand the current state of COVID-19 vaccination. In this study, two public datasets were used: the Johns Hopkins University coronavirus 2019 dataset and Our World in Data - Coronavirus Pandemic dataset. With datasets, two approaches have been used: visual data analysis for COVID-19 vaccine administration and the autoregressive integrated moving average (ARIMA) model for forecasting vaccination rates.

The findings confirm that Oxford/AstraZeneca is the top vaccine used across the globe with 26.54%, the United States is the top in vaccination, with 277,290,173, India is the top in number of daily vaccinations with 3.659357M, and in total vaccinations per hundred people, the United States has the highest count with 82.91, among the top five countries. It is also estimated that the vaccination rate in the United States will reach almost 60%, while India, Brazil, France, and Turkey will reach about 15%, 28%, 60%, and 23%, respectively, in the following 50 days beginning 20 May 2021. This exploratory study of COVID-19 vaccination data was carried out to effectively show the current state of COVID-19 vaccine administration and to anticipate vaccination rates in the United States, India, Brazil, France, and Turkey.

INTRODUCTION

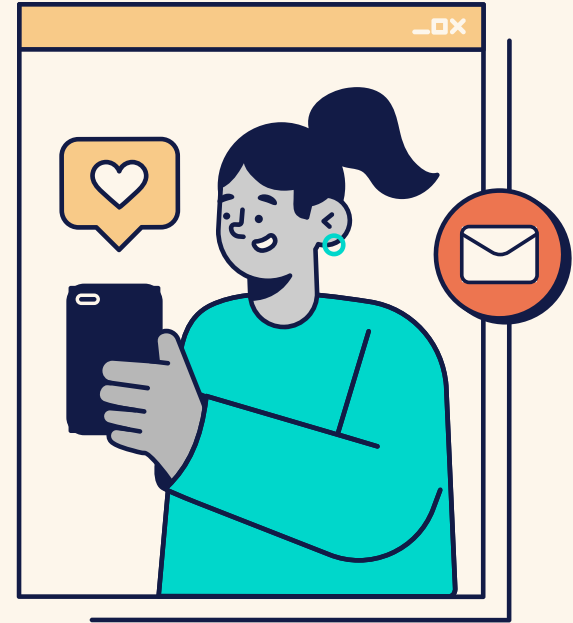
The World Health Organization classified COVID-19 a Public Health Emergency of International Concern on January 30, 2020 and a pandemic on March 11, 2020.

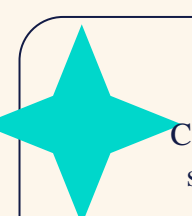


The COVID-19 outbreak, officially identified as the coronavirus disease outbreak, would be a continuing major worldwide public health problem of coronavirus disease 2019 (COVID-19) impacted by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The epidemic happened in Wuhan, China, in December

- 3.46 million authenticated fatalities attributed to COVID-19, making it one of the worst pandemics in history.
- Since this virus's specific source is uncertain, the very first epidemic occurred in late 2019 in Wuhan, Hubei, China (To et al. 2021).
- More than 167 million cases had occurred as of May 24, 2021, with more than

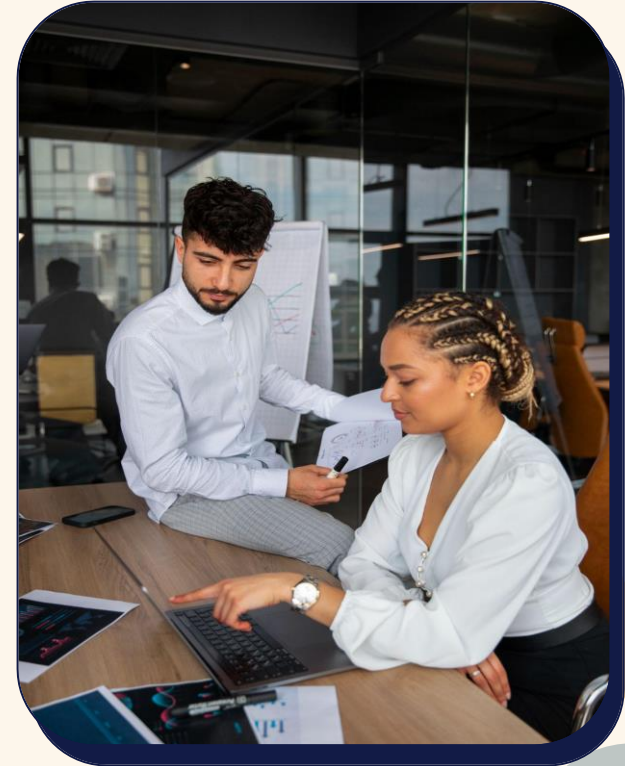

Since this virus's specific source is uncertain, the very first epidemic occurred in late 2019 in Wuhan, Hubei, China (To et al. 2021). Several previous instances of COVID-19 were connected to persons who may have travelled the Hua-nan Seafood Wholesale Market in Wuhan (Suneal. 2020), though living person dissemination may have occurred prior to this (Huetal. 2021). On February 11, 2020, the World Health Organization (WHO) called the sickness "COVID-19," short for coronavirus disease 2019 (World Health Organization 2020). The infection which sparked the pandemic is identified as SARS-CoV-2, a recently found virus which is directly connected to bat coronaviruses





CoViD-19 Data Repository [2020](#); Muthusami and Saritha [2020](#)). Our explicit study used the 2019 corona- virus dataset of Johns Hopkins University (January – May 2021). We provide a contribution to display and evaluate the findings between 22 January 2020 and 24 May 2021. COVID-19 has still spread almost 192 Countries/Regions, 87 Cities/Provinces, and 275 diverse geographic places. Applying time-series data, it approximated the set of possible cases, such as reported infections, fatalities, and cured cases, throughout the world, as well as the top five nations. The United States and India are among top leading countries as of May 24, 2021. In addition to the variety of diverse instances, including such proven illnesses, fatalities, and recoveries in those nations,

The cumulative number of reported infectious cases worldwide is 167,316,360, with a global mean rate of 0.36 and a standard deviation of 1.7. The Top 5 nations' global mean rate is 10.4, with a standard deviation of 7.5. In this case, the United States ranks first on a total of 33,143,662, the worldwide percent is 19.81, from a total of 26,948,874, the global percentage is 16.11, India ranks second, Brazil ranks third at 16,120,756, the worldwide percent is 9.63, with a total of 5,550,143, the worldwide percent is 3.32, France ranks fourth, and Turkey ranks five with 5,194,010, the global mean is 3.10. The total amount of fatalities world- wide is expected to be 3,473,036, with a global mean of 1.81. With this circumstance.



03

Materials and methods





MATERIALS

In this research, two approaches have been used: visual data analysis for COVID-19 vaccine administration and the autoregressive integrated moving average (ARIMA) model for forecasting vaccination rates. In this study, two public datasets were used; (i). the Johns Hopkins University corona virus 2019 dataset, which covers the period from 22 January 2020 to 24 May 2021 (Johns Hopkins University Center for Systems Science and Engineering: CoViD-19 Data Repository [2020](#)), (ii). Our World in Data- Coronavirus Pan-demic (COVID-19) dataset, which covers the period from 20 December 2020 to 19 May 2021 (Hannah et al. [2020](#)).



METHOD

The first part of this study presents a visual data analysis of COVID-19 vaccine administration in terms of various aspects such as the proportion of top 10 vaccines in the race to combat COVID-19, the number of cumulative vaccinations and every day vaccinations as per country, cumulative vaccinations per country grouped by vaccines, daily vaccinations per countries, and the link between cumulative vaccines and cumulative vaccines per hundred of the top five countries heavily affected by the COVID-19 internationally as of May 24, 2021, including the United States, India, Brazil, France, and Turkey. These results are shown using Python language libraries and data from the Our World in Data - Coronavirus Pandemic (COVID-19) dataset.

STATISTICAL ANALYSIS

Comparison to the population of these nations. The daily vaccine administration among these five nations was discovered to be as follows: India ranks first with 3.659357 million, the United States placed second with 3.384387 million, Brazil comes third with 1.135847 million, France ranks fourth with 498.762 thousand, and Turkey comes last with 435.596 thousand. Following that, how much each of these five countries used and which vaccinations they used revealed that the United States used 'Moderna, Johnson & Johnson, Pfizer/BioNTech,' with a total of 277,290,173, while India used 'Covaxin, Oxford/AstraZeneca,' with a total of 186,410,600, 'Oxford/AstraZeneca, Pfizer/BioNTech, Sino-vac' vaccines were used in Brazil with a total of 55,098,913,

'Johnson & Johnson, Moderna, Oxford/AstraZeneca, Pfizer/ BioNTech' vaccines were used in France with a total of 30,264,699, and 'Pfizer/BioNTech, Sinovac' vaccines were used in Turkey with a total of 26,821,460.

Figure 3 depicts the relationship between cumulative vaccinations and cumulative vaccinations per hundred for each country. Among these five countries, the United States ranks first in cumulative vaccinations and cumulative vaccinations per hundred, with 277,290,173 and 82.91, respectively. India, Brazil, France, and Turkey have 186,410,600/13.51, 55,098,913/25.92, 30,264,699/44.79, and 26,821,460/31.8, respectively

<https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>

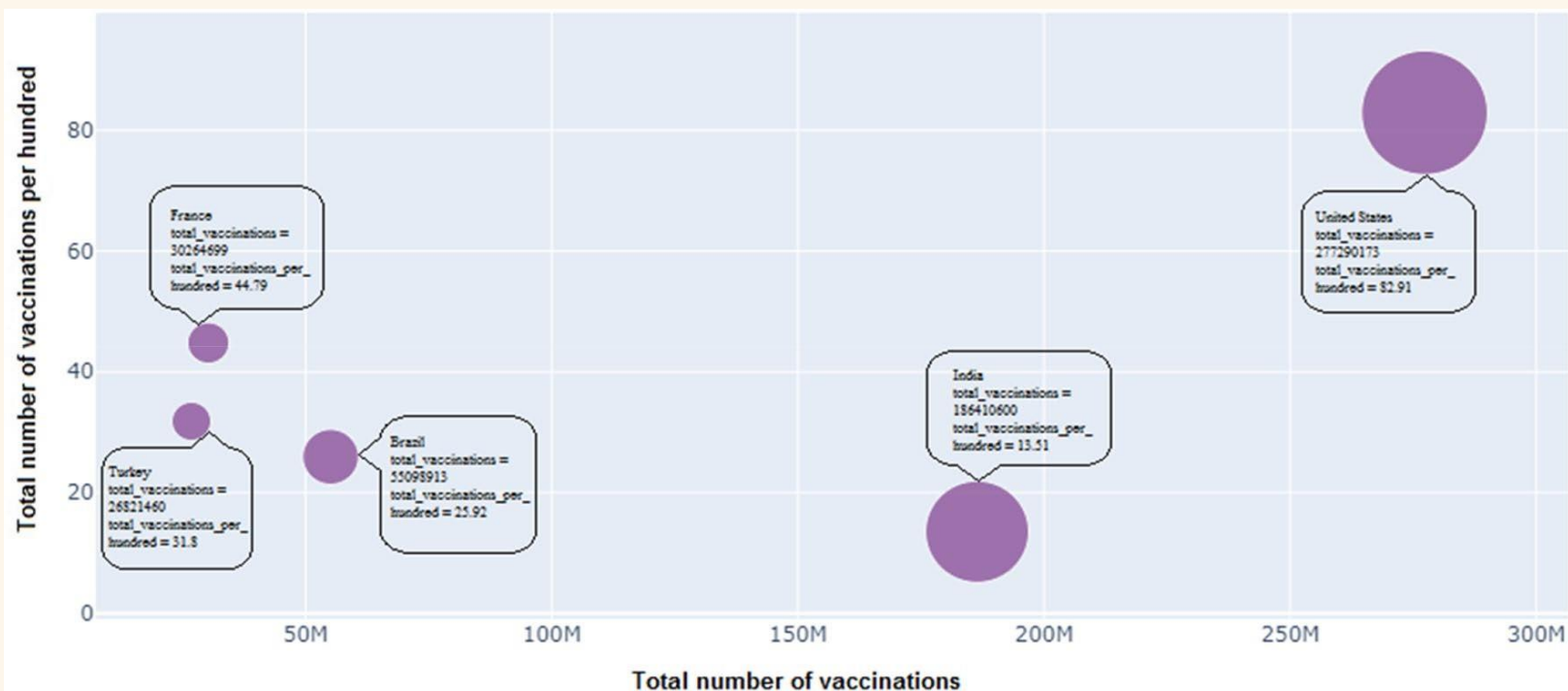
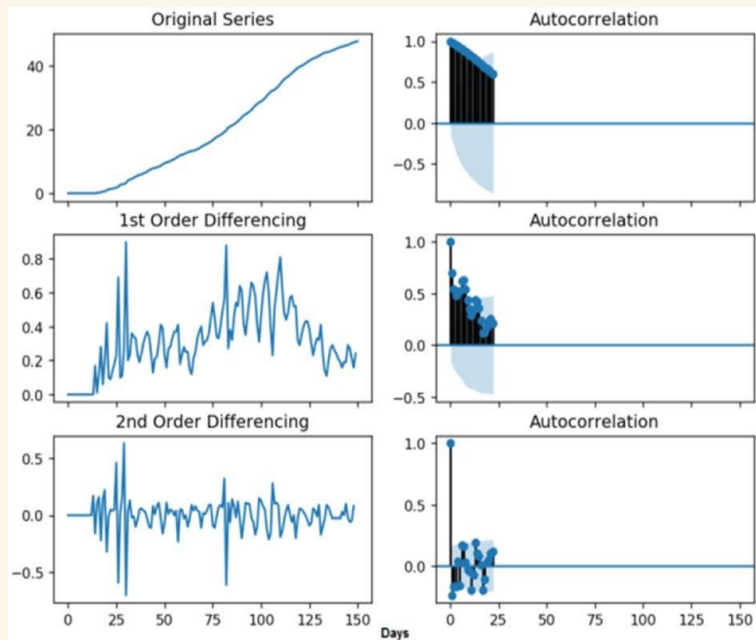
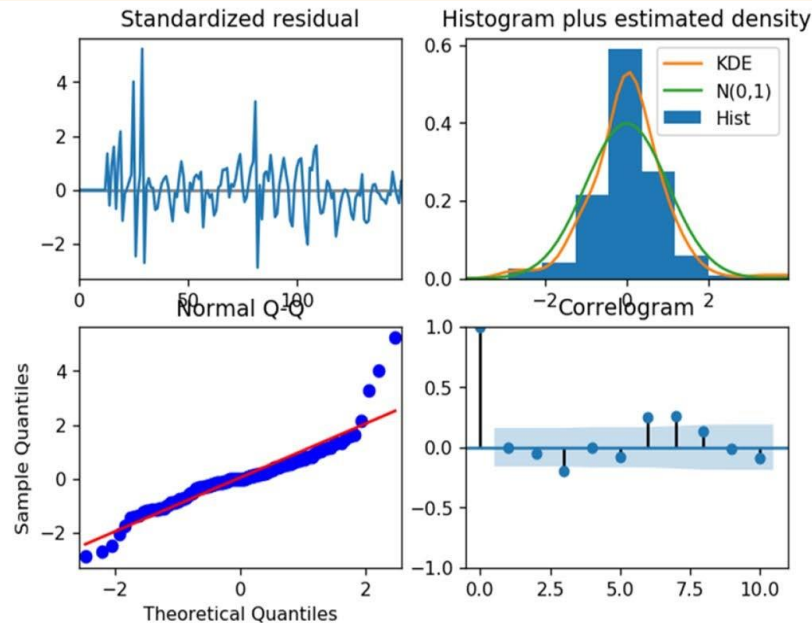


FIG 3. Display the link between cumulative vaccinations and cumulative vaccinations per hundred of the countries, the United States, India, Brazil, France, and Turkey, as of May 19, 2021, using dataset



(A)



(B)

FIG 4.A Displays the differencing and autocorrelation of COVID-19 vaccination time series data of the United States from dataset 2. **B** Displays the ARIMA fit residual and density of the United States time-series data from dataset 2 with $AR(1) = 0$, $I(d) = 2$, and $MA(n) = 2$

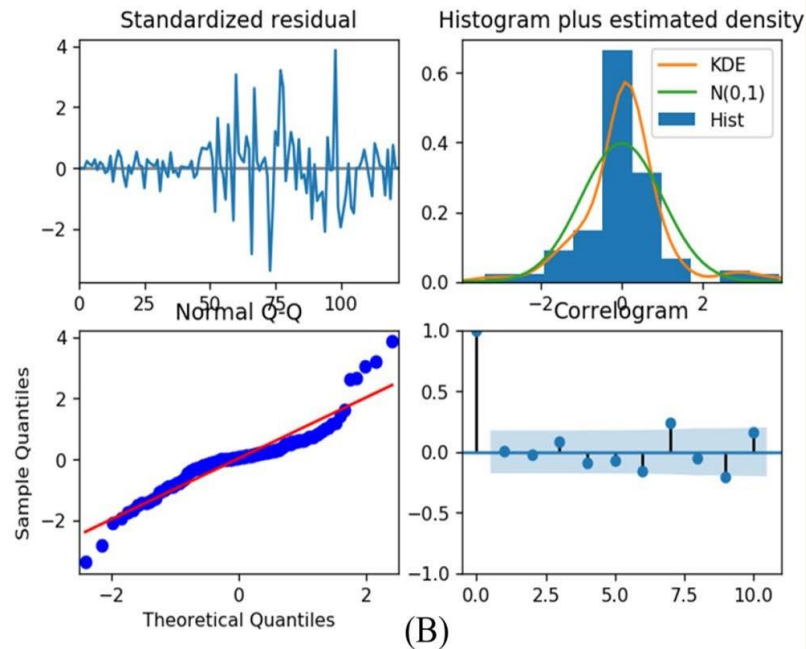
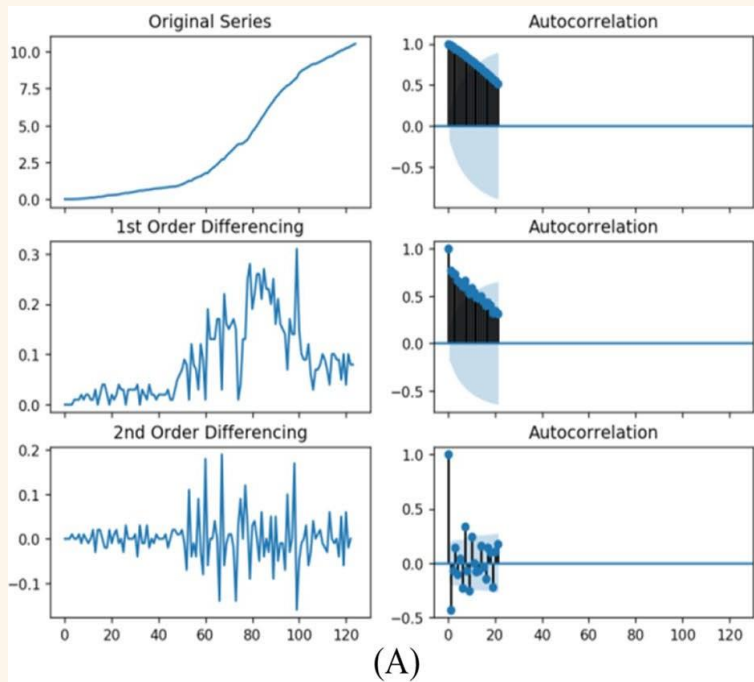


FIG 5.A Displays the differencing and autocorrelation of COVID-19 vaccination time series data of India from dataset 2. **B** Displays the ARIMA fit residual and density of the United States time-series data from dataset 2 with $AR(l) = 0$, $I(d) = 2$, and $MA(n) = 1$

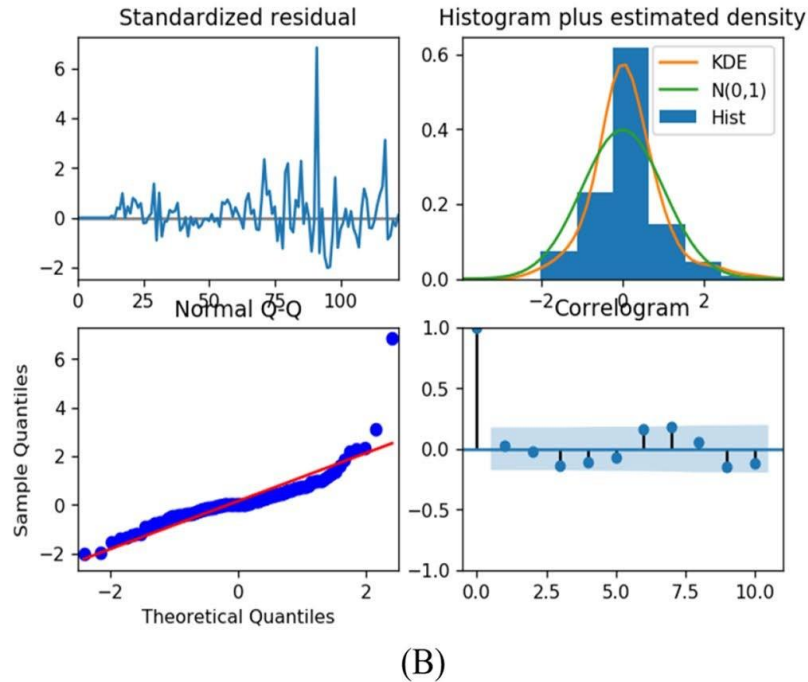
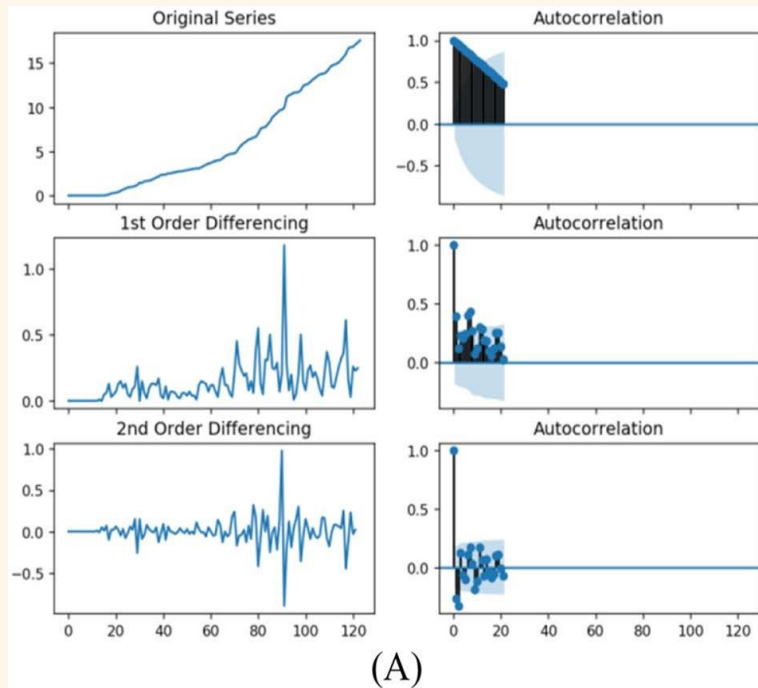


FIG 6. A Displays the differencing and autocorrelation of COVID-19 vaccination time series data of Brazil from dataset 2. B Displays the ARIMA fit residual and density of the United States time-series data from dataset 2 with $AR(l) = 1$, $I(d) = 1$, and $MA(n) = 3$

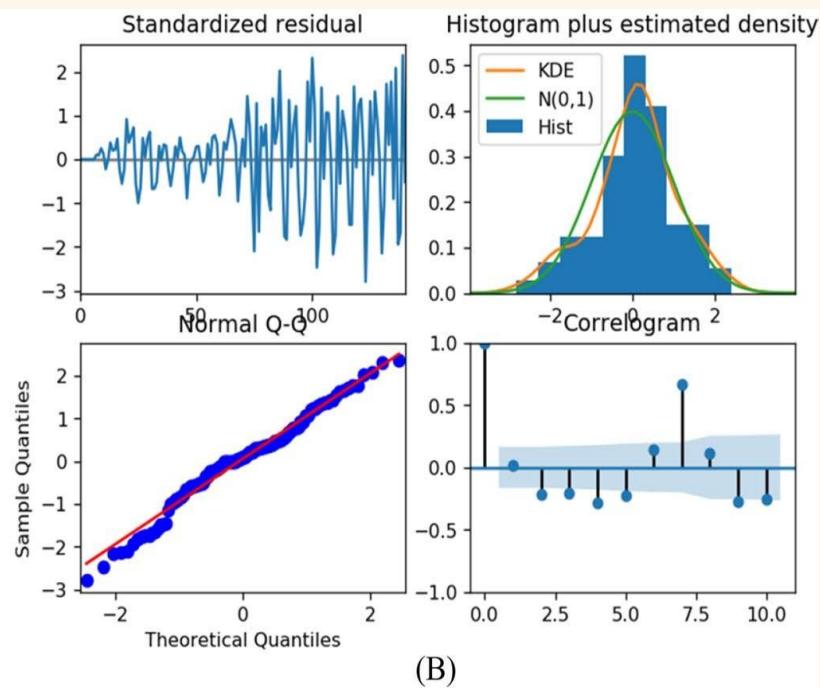
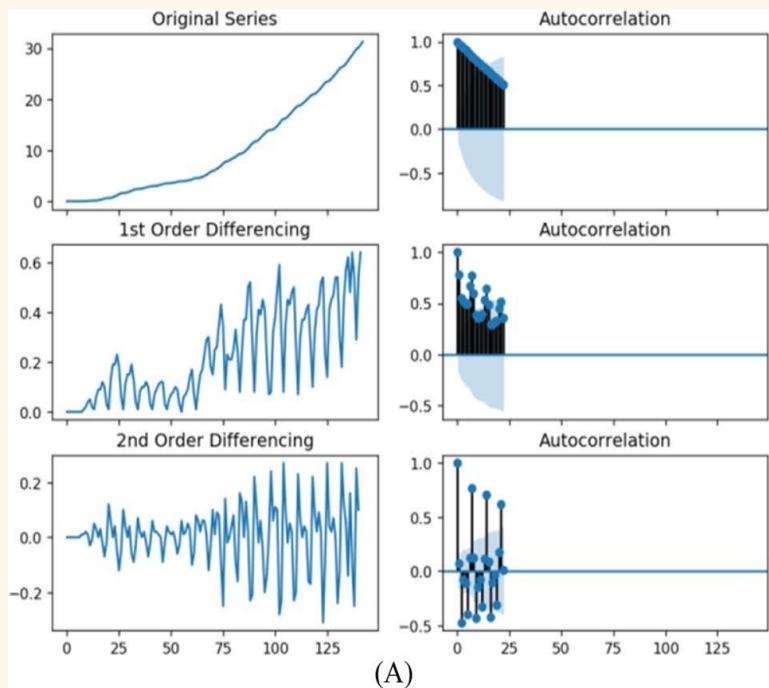
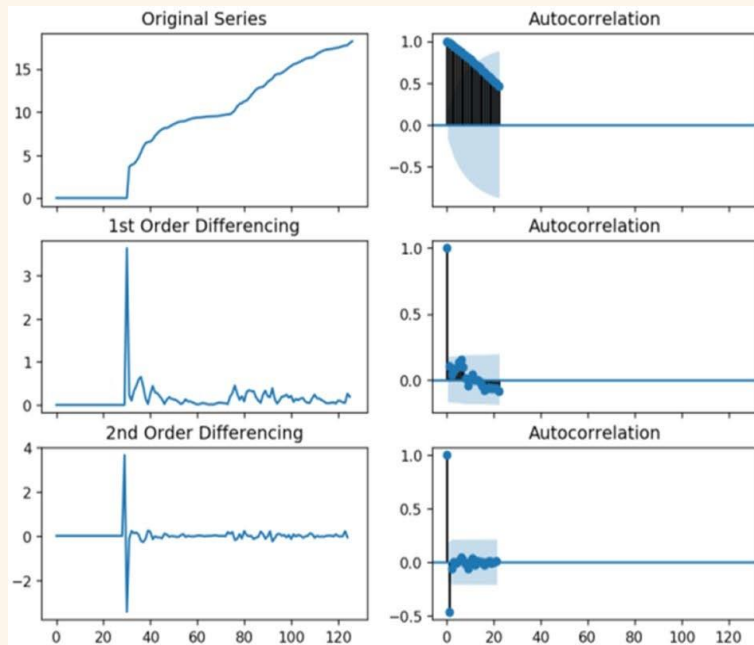
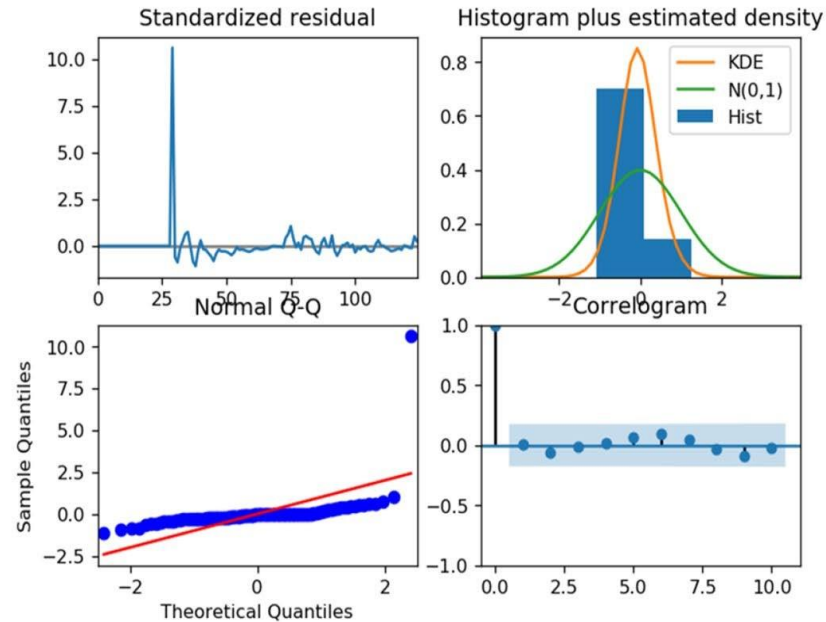


FIG 7 .A Displays the differencing and autocorrelation of COVID-19 vaccination time series data of France from dataset 2. **B** Displays the ARIMA fit residual and density of the United States time-series data from dataset 2 with $AR(1) = 2$, $I(d) = 2$, and $MA(n) = 0$

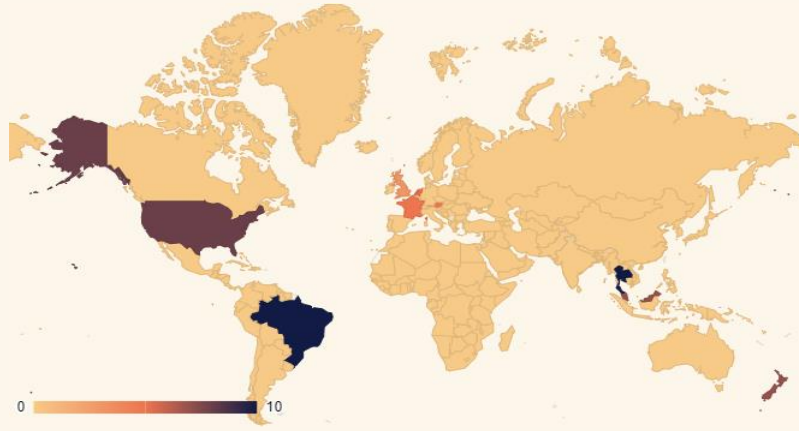


(A)



(B)

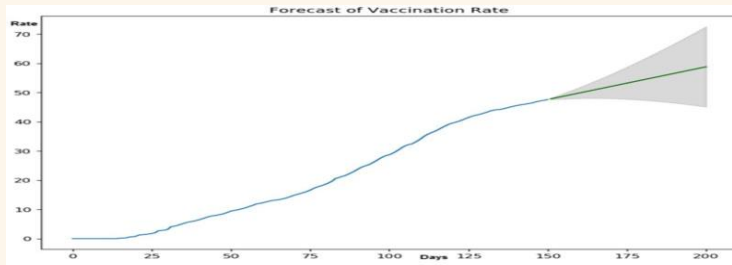
Displays the differencing and autocorrelation of COVID-19 vaccination time series data of Turkey from dataset 2. B Displays the ARIMA fit residual and density of the United States time-series data from dataset 2 with $AR(l) = 0$, $I(d) = 2$, and $MA(n) = 1$



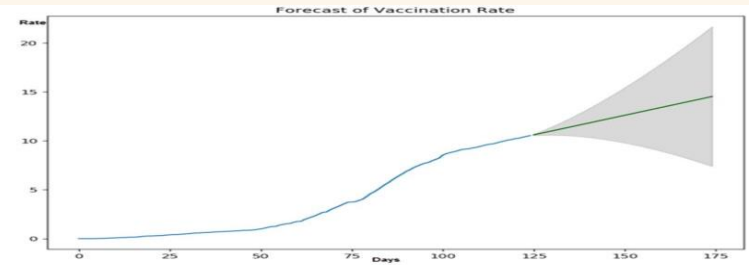
Then, as shown in Figs. 4A, 5A, 6A, 7A, evaluate an autocorrelation plot of a time which illustrates the autocorrelation for possibly high lags in the time series. As it seems that there is a positive correlation with its first 3-to-5 lags, which may be noteworthy for the first 3 lags, indicating that 3 may be a decent baseline for the AR parameter of a model. The ARIMA model is then created using the factors l , d , and n . T

Using the ARIMA approach to investigate a time-series data presupposes that the conceptual model used to analysis the data would be an ARIMA model. It may look simple, which stresses the importance of evaluating the model's assumptions in raw data and residual errors of model forecasts. To discover the trend, we first imported the vaccination time series dataset 2; Fig. 4A disclosed the United States, Fig. 5A showed India, Fig. 6A discovered Brazil, Fig. 7A revealed France and Fig. 8A exposed Turkey.

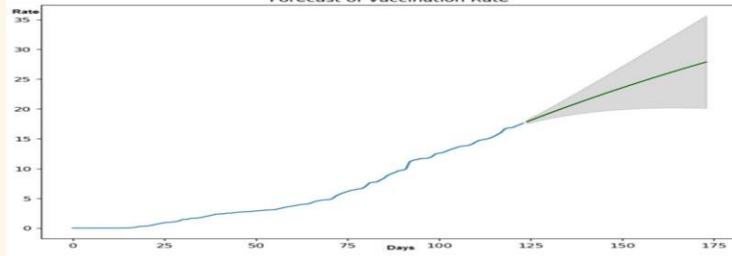
With this circumstance, first-order differencing of time series did not eliminate complete trend and seasonality, therefore second-order differencing of time series was performed to obtain a stationary time series.



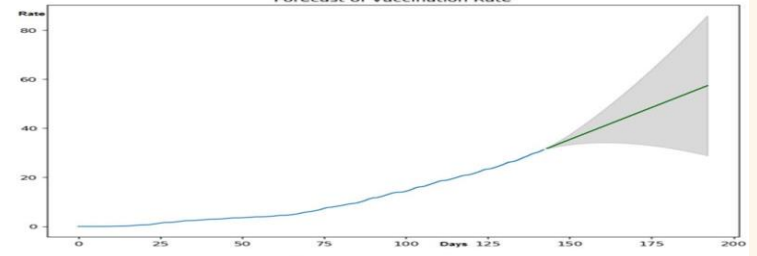
(A) United States



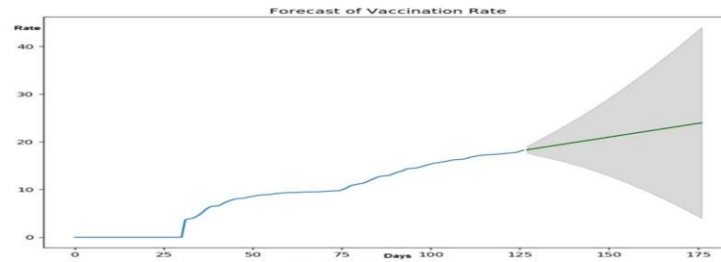
(B) India



(C) Brazil



(D) France

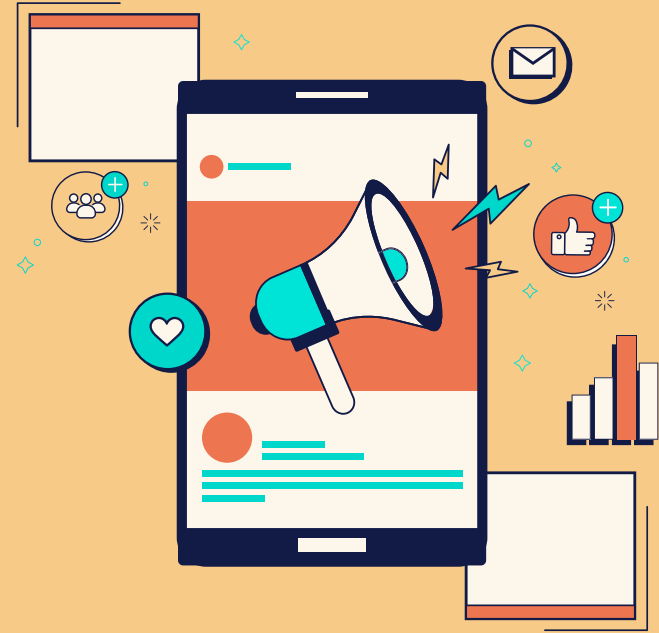


(E) Turkey

A – E Shows the vaccination rates in the United States, India, Brazil, France, and Turkey respectively over the next 50 days beginning on may 19,2021

To determine the order of the fittest ARIMA model and the augmented Dickey–Fuller (ADF) test to determine the appropriate value of differencing. As a result, the best model was chosen as ARIMA (0, 2, 2) in the United States, ARIMA (0, 2, 1) in India, ARIMA (1, 1, 3) in Brazil, ARIMA (2, 2, 0) in France and ARIMA (0, 2, 1) in Turkey. The residual was then subjected to a diagnostic examination, as shown in Figs. 4B, 5B, 6B, 7B, and then it was determined that the residuals are uncorrelated and have a zero mean.

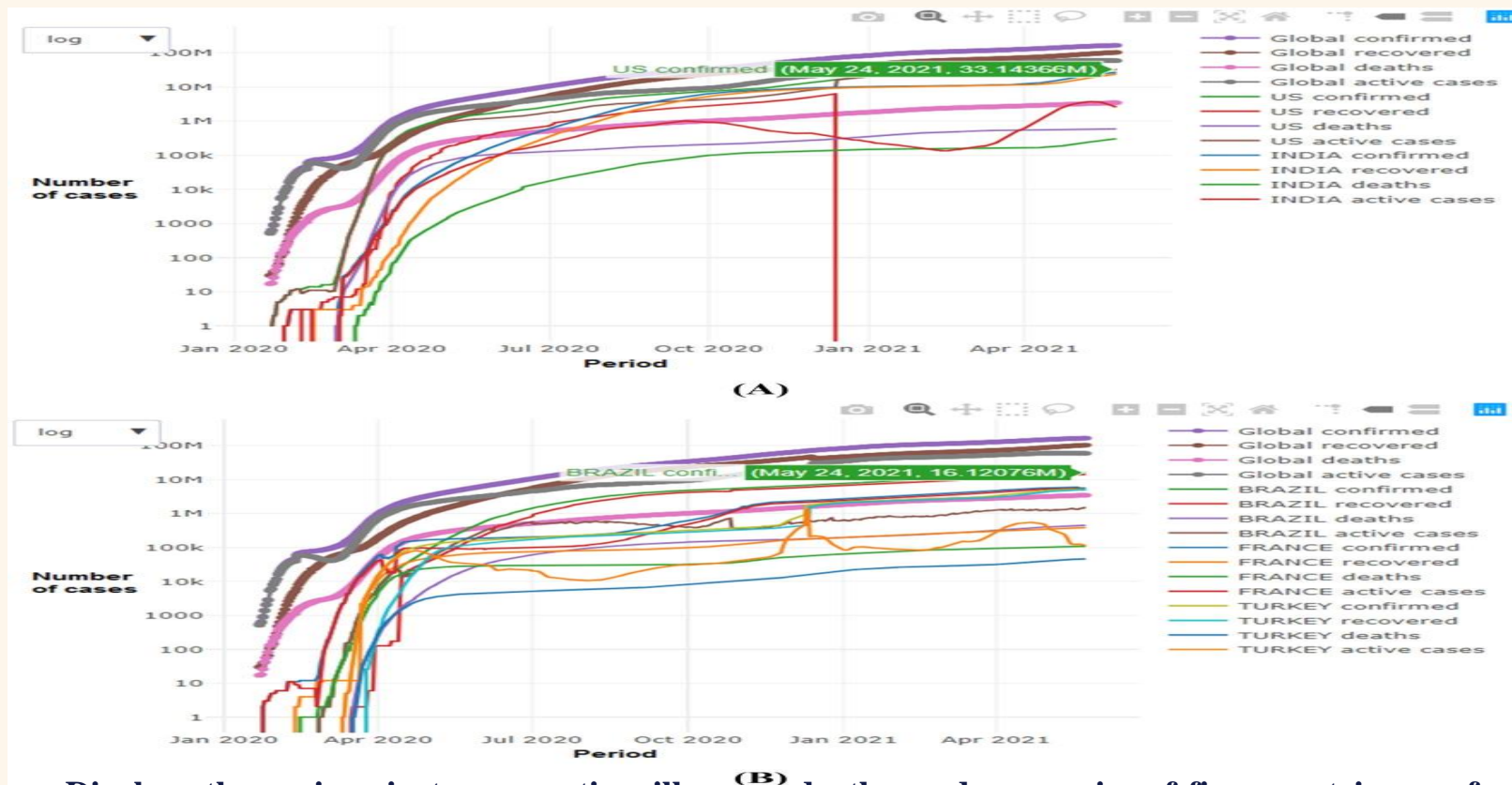
At this moment, the fitted ARIMA model forecasts the vaccination rate in the following 50 days beginning 20 May 2021 for the nations of the United States, India, Brazil, France, and Turkey with dataset 2. According to Fig. 9 (green line), the vaccination rate in the United States will reach almost 60%, while India, Brazil, France, and Turkey will reach about 15%, 28%, 60%, and 23%, respectively, in the following 50 days beginning 20 May 2021.



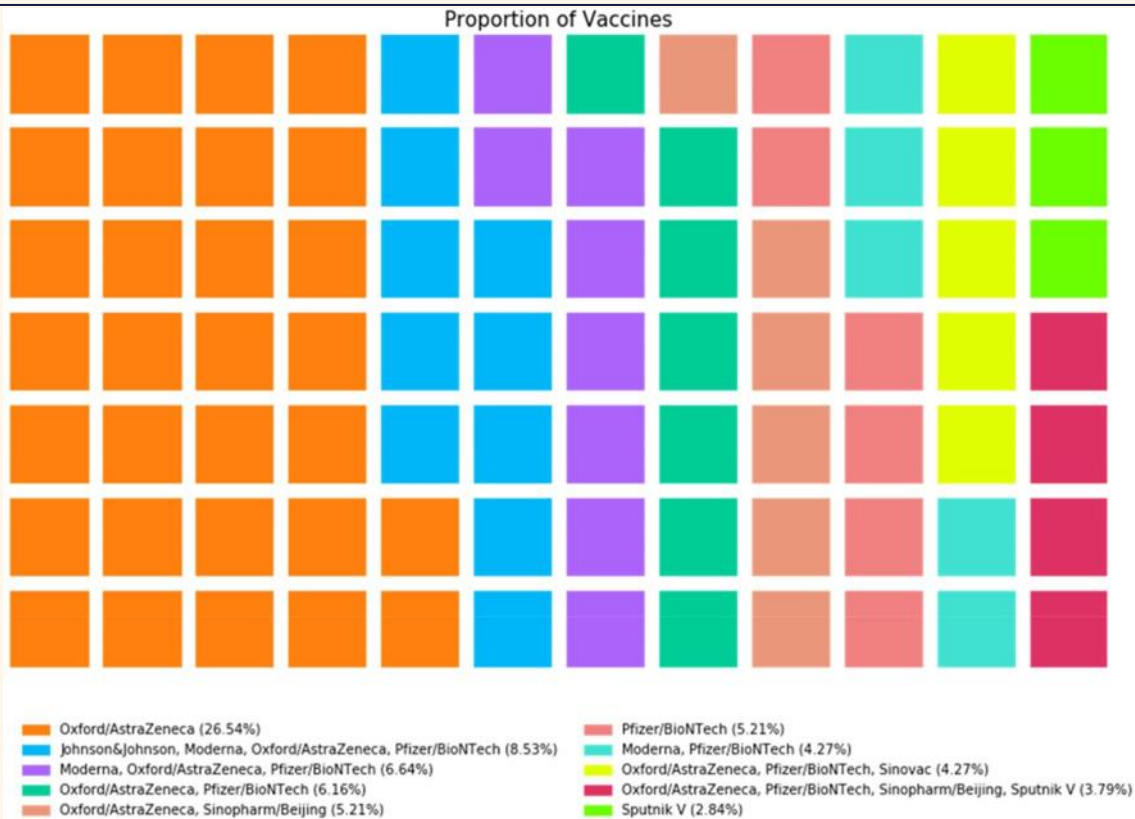
VISUALIZATION

To determine the order of the fittest ARIMA model and the augmented Dickey–Fuller (ADF) test to determine the appropriate value of differencing. As a result, the best model was chosen as ARIMA (0, 2, 2) in the United States, ARIMA (0, 2, 1) in India, ARIMA (1, 1, 3) in Brazil, ARIMA (2, 2, 0) in France and ARIMA (0, 2, 1) in Turkey.

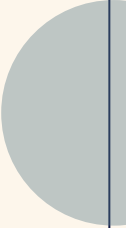







Displays the various instances, active, illnesses, deaths, and recoveries of five countries as of May 24, 2021, using dataset 1. A United States and India, B Brazil, France, and Turkey



Display the proportion of the top ten vaccines in the globe as of May 19, 2021, with dataset 2

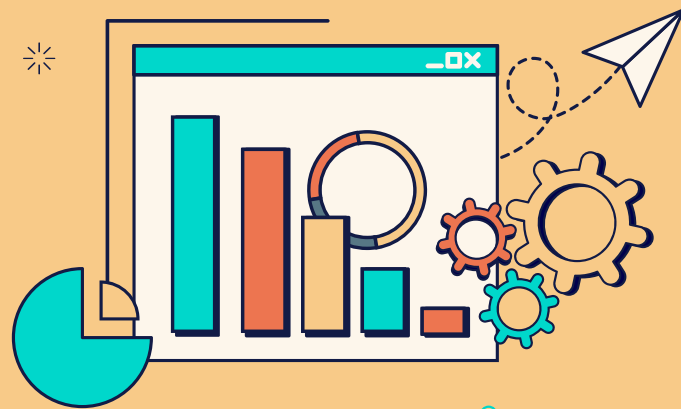



This is a refinement of the conventional AutoRegressive Moving Average that incorporates the concept of convergence (Hyndman and Athanasopoulos 2018). This abbreviation, ARIMA is comprehensive, summarizing the model's major features. In a nutshell, these seem to be: AR - autoregression: the model typically employs the reliant connection among a given feature and a set of deferred data, I - integrated: use of the raw observational quantization to render a time series stationary, MA - moving average: a model which applies the dependence besides an inference and a residual error from a moving average model to deferred data. These kinds of components are clearly specified as being a factor in the model. ARIMA (l, d, n) is a general procedure in which the factors are supplemented by integers to rapidly specify the precise ARIMA model is utilized. The ARIMA model's factors would be as shown in: l: the amount of lag inferences considered with the model, also known here as lag order, d: the amount of incidents that baseline assumptions being differed, generally known as level of residuals; n: a weighted average window density also known as the weighted average order. A simple regression classifier is developed with the necessary quantity and kind of features, as well as the data is processed by a quantity of differencing that render it stationary, i.e. to eliminate trend and seasonal features that weaken the regression analysis.



06

RESULT





An exploratory analysis of COVID-19 vaccination data was performed using Python packages, yielding the results of the global proportion of vaccines as well as the top 10 vaccines in terms of availability on May 19, 2021 around the world, total amount of vaccines, and daily vaccine administration in the United States, India, Brazil, France, and Turkey as well as how much each of these five countries has used, the vaccines they have used, and the relationship between cumulative vaccinations and cumulative vaccinations per hundred for each country. The vaccination rate is then forecasted over the next 50 days commencing 20 May 2021 for the countries of the United States, India, Brazil, France, and Turkey using time series data from dataset 2 and the autoregressive integrated moving average (ARIMA) model.

