Phase 3
Covid vaccines analysis

(Pre processing, analysis and visualisation)

ABSTRACT

The 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 show the current state of COVID-19 vaccine administration effectively and to anticipate vaccination rates in the United States, India, Brazil, France, and Turkey.

INTRODUCTION

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 2019. The World Health Organization classified COVID-19 as a Public Health Emergency of International Concern on January 30, 2020, and a pandemic on March 11, 2020. More than 167 million cases had occurred as of May 24, 2021, with more than 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). Several previous instances of COVID-19 were connected to persons who may have traveled to the Huanan Seafood Wholesale Market in Wuhan (Sun et al. 2020), though living person dissemination may have occurred prior to this (Hu et al. 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 that sparked the pandemic is identified as SARS-CoV-2, a recently found virus that is directly connected to bat coronaviruses (Perlman 2020), pangolin coronaviruses (Zhang et al. 2020), and SARS-CoV (European Centre for Disease Prevention and Control: Risk assessment 2020).

MATERIALS and METHODS

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 coronavirus 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 Pandemic (COVID-19) dataset, which covers the period from 20 December 2020 to 19 May 2021 (Hannah et al. 2020).

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 everyday 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.

The second part of this study forecasts vaccination rates for those nations over the upcoming fifty days from May 20, 2021, using the autoregressive integrated moving average (ARIMA) model. The ARIMA model would be a type of predictive model that is used to evaluate and forecast data over the period. ARIMA seems to be a short form for AutoRegressive Integrated Moving Average (Swain et al. 2018; Robert 2020). It's really a type of approach that incorporates a variety of conventional temporal information in time series. This expressly responds to a number of fundamental patterns in time series analysis and thus offers a simple though effective way for creating skilled time series predictions. 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 utilized. The ARIMA model's factors would be as shown in l: the number of lag inferences considered with the model, also known here as lag order, d: the number of incidents that baseline assumptions 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 renders it stationary, i.e. to eliminate trend and seasonal features that weaken the regression analysis.

Data Pre-Processing and Normalization

Data cleaning and pre-processing is a crucial step before starting any analysis as it makes our data reliable for further calculations. Data taken from online databases offer structured data on a day-to-day basis. All variables were converted into monthly data, namely, COVID cases, deaths, and people vaccinated for dosage 1 and dosage 2. The data were normalized to fit it within a range and perform statistical computations using Standard scaler and min–max scaler. After data standardization and normalization, logarithmic transformations were also performed so data could become reliable for analysis. Following the data cleaning, statistical calculations were performed using linear, polynomial, OLS regression models and a support vector machine to assess the effect of vaccination over COVID cases and deaths and check the accuracy of the model. All analysis was performed on the Python platform.

Statistical Analysis

Statistical analysis is essential in verifying assumptions and demonstrating them to create a concrete conclusion about a study. This study focuses on the efficacy of vaccination over COVID cases and COVID deaths. The linear regression analysis will investigate the relevance of vaccinations, followed by polynomial and OLS regression models and SVM models. This will provide information about the effectiveness of

being immunized.

Linear Regression

In linear regression, two variables are employed: one is the dependent variable (plotted on the y-axis) on which the prediction is based, and the other is an independent variable (plotted on the x-axis) utilized to make the prediction. Variable-based prediction might be univariate (based on one variable) or multivariate (based on several variables) (Moore et al., 2013). A regression line is a straight line that explains how the dependent variable changes with the change in the independent variable. To contrast the model predictions against several sets of field data, we use vaccine dose data to calculate the number of COVID-19 cases and the number of people dying considering the above factors. We fit the model with

$$\text{Regression line}, y = mx + c,$$

where c is an intercept and m is the slope and y is the dependent variable and x is the independent (explanatory) variable r2 is the coefficient of determination which is calculated by Karl Pearson's coefficient (r) (Calkins, 2005). This coefficient indicates how many variations are explained by the variable being predicted. The greater the slope, the greater the correlation between the variables, and the greater the ability to explain fluctuations in other variables. Linear regression

improves prediction since it focuses on situations with one or more predictor variables (in our study, vaccine data for First Dose and Second Dose) and one outcome variable.

$$\text{Multiple linear regression}, y = ax + bz + c,$$

where a and b are coefficients of regression, c is the intercept while having x and z as multiple explanatory variables.

The outcome variable, y, is a linear function of each predictor variable, x, and z, forcing the regression model to be a straight line (Marill, 2004). For a good model, the r-value should lie in 0.5 to 1.0 so this score gives a good correlation and a good predictor. Regression analysis is also used to predict the p-value for significance testing. The statistical inference approach is based on a complex network that includes assumptions about how data was gathered and analyzed and how the research results were presented.

RMSE and r^2-Value

The RMSE is the square root mean error. This error value gives an idea about the fitness of the model, i.e., how the values deviate from the true value. RMSE is an absolute measure of fit, while R-squared is a relative measure of fit. RMSE can be interpreted as the standard deviation of the

unexplained variance since it is the square root of the variance. It has the advantage of being in the same units as the answer variable. The lower the RMSE value the better will be the prediction. If the main goal of the model is prediction, the RMSE is the essential criterion for fit because it is a valid standard of how well the model predicts the response.

$$\text{Mean Square Error} = \text{True value} - \text{Predicted value}$$

Root mean square error=(True value−Predicted value). We square the error because the estimate can be above or below the true value, resulting in a negative or positive difference. If we didn't square the errors, the sum might fall due to negative differences rather than a strong model fit. Lower values of RMSE indicate a better fit.

Polynomial Regression

For a good predictor model, polynomial regression is the best approach. Some points which do not fit in linear regression fit best in polynomial regression. If the linear regression is underfitting, then we plot polynomial regression to increase complexity in the model by increasing the power of the features and making them new features. All models for degree quadratic, cubic and quartic were constructed using python, and based upon the RMSE value, the best was selected. Equation (5) shows a polynomial regression curve of degree 4.

Polynomial regression of degree4,y=x+ a

where, a, b, and c are coefficients of regression and d is the intercept. The polynomial curve can be studied for the complexities of COVID cases and deaths based upon vaccination Dosage 1 and Dosage 2.

OLS Regression Model and P-Value Interpretation

The p-value is the “probabilities” of hypotheses. When we perform statistical significance analysis based upon the hypothesis we have designed; if there is a condition in which we have a p-value that is very low like 0.0 (although it is not exactly 0), that means that there is a strong correlation between the coefficients and the target (Princeton University). Statistical “significance tests” based on this concept have been a central part of statistical analyses for centuries (Stigler Stephen, 2003). Traditional p-value and statistical significance notions have emphasized null hypotheses, treating all other assumptions used to calculate the p-value as if they were true. Recognizing that the other assumptions are always suspect, if not outright false, we'll look at the p-value in a broader sense as a statistical overview of the compatibility between observed results and what we would expect to see if the entire statistical model (all the assumptions used to compute the p-value) were correct. And then, we have an alternate hypothesis in which it is opposite to the null hypothesis. Based on p-value, i.e.,

If p-value >0.5; so, we accept null hypothesis

If p-value <0.5; so we reject null hypothesis

And we can make a concrete statement about our study do have relevance or significance or not. In logical terms, we can say that the p-value tests all assumptions related to the model developed, not only focusing on the target hypothesis, i.e., the null hypothesis. OLS regression uses Student's t distribution for calculating class intervals from which p-value can be interpreted.

To perform the statistical analysis, we can use the OLS regression model, which stands for ordinary least square regression used to compute unknown parameters in the Regression model (linear or polynomial). The method of OLS provides minimum-variance mean-unbiased estimation when the errors have finite variances and are normally distributed. OLS is the maximum likelihood estimator. The (squared) vertical distance from each data point to the line is reduced overall data points by using OLS regression to match a line to bivariate data. The equation:

$bOLS=Cov(x, y)/var$ describes the slope of this axis (x). As a result, OLS slopes change whether either the way x and y covary or the variance of the x-axis variable changes (Sokal and Rohlf, 2012). This OLS regression calculates a p-value which is easy to interpret based on all variables. OLS model was chosen as it gives a reasonable interpretation of models generated and a better way to access the model's relevancy. Using the stats model package, the OLS regression was calculated and

the hypothesis generated as stated below:

Null hypothesis = There is no significant effect of people vaccinated over COVID cases and COVID deaths

Alternate hypothesis = There is a significant effect of people vaccinated over COVID cases and COVID deaths.

Using this model, we can be clearer about the significance of vaccinations.

Support Vector Machine Algorithm

Calculating the accuracy of a model-designed support vector machine (SVM) algorithm is a good measure. It is a supervised learning algorithm that classifies data into 2 classes based upon which training is done and then, using that, future learning classifications are made. These algorithms are more efficient as their performance is high. Using SVM, a hyperplane can be plotted between datasets which are called a decision boundary. Based upon that classifier, classification can be performed. This is an advanced version of the linear and polynomial regression model analyzed above. By using SVM we can make some predictions and these predictions can be compared with the actual values and in the last, the accuracy of the model can also be obtained (Bruno, 2017). SVM regression analysis will complete our analysis and tell us about the

efficacy of the model and decrease the error value of the model by making it more precise.

Statistical analysis

Common statistical methods like logistic regression, two-sample t-test, Fisher's exact test, K-means clustering, and principal component analysis (PCA) were applied to compare the vaccine-developed and non-vaccine-developed groups of countries using the above-mentioned R&D indicator variables. Continuous variables were standardized to have a mean zero and a standard deviation of one before logistic regression analysis.

For the vaccine policies and the age-group-specific data, the generalized estimation equation (GEE) approach (26, 27) was applied to analyze these two datasets. The model is defined as below:

$$\log(\mu i(t))=\beta 0+\beta 1t+2\log(t)+3Zi(t-lag) \quad (1)$$

where μi = E(Yi), Yi(t) are the daily COVID-19 confirmed cases or specific age-group rate of new cases, t is the number of days or weeks since the first case, β0, β1 and β3 are the regression coefficients, Zi(t)s are the different vaccine policies or specific age-group vaccination, for each country i. The Poisson distribution and the log link function were used. An independent working correlation was assumed. 0, 2 weeks, 1, 2, and 3 months lagging (lag) were assumed when analyzing the vaccine policies' data but not in the age-group-specific analysis because it is bi-weekly and not daily data.

A lag is a fixed time displacement in time series data. This assumes that the effects of policies implemented on a given day may affect the number of confirmed cases several days after implementation. Statistical significance level was taken for p-values (P) < 0.05. All countries and cluster analyses using groups of countries produced from the K-means clustering analysis were performed for both vaccine policy analysis and specific age-group analysis. All analyses were carried out in the R (ver. 4.1.0) software tool.

Results

Vaccine development analysis

In total, 20 variables were analyzed for their association with fast COVID-19 vaccine development. Exploratory analysis using unsupervised clustering methods of PCA and K-means clustering did not reveal noticeable differences between the two groups of countries. Figure 1 shows the results of PCA and K-means clustering. K-means (k = 2) classified the countries into two groups with one cluster having the USA, China, and Germany and the other cluster having the remaining 32 countries. K-means misclassified 8 countries belonging to the vaccine-developed group as belonging to the non-developed vaccine group. The same could be observed with PCA as it could not differentiate between the two groups of countries. Both methods captured 33.5 % and 16% variance of the data in the 1st and 2nd dimensions, respectively.

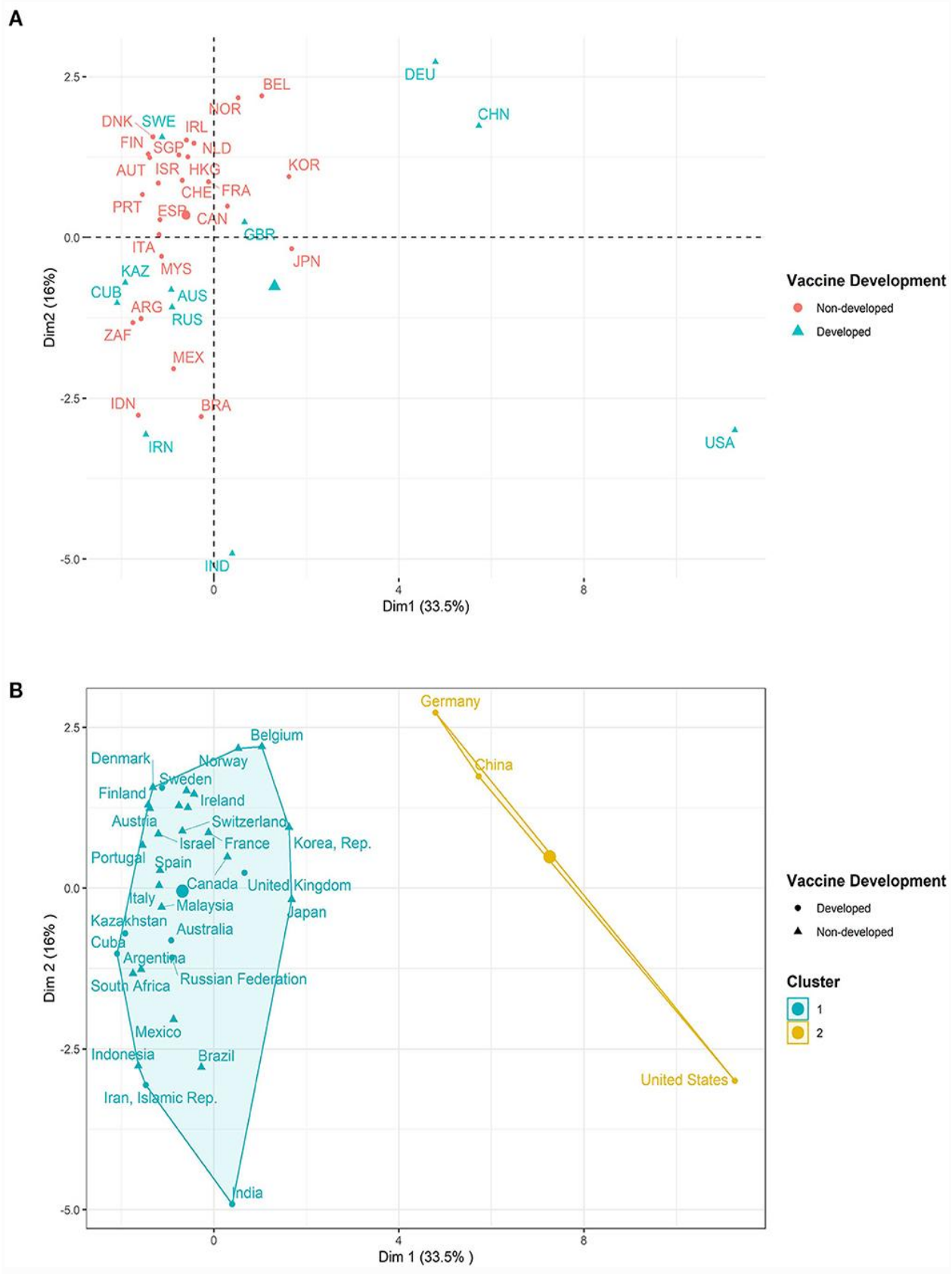

A
BEL
DEU
NOR
CHN
DNK
SWE
IRL
FIN
SGP
NLD
AUT
ISR
HKG
KOR
CHE
FRA
PRT
ESP
CAN
GBR
ITA
JPN
KAZ
MYS
CUB
AUS
ARG
RUS
ZAF
MEX
IDN
BRA
IRN
USA
IND
2.5
0.0
-2.5
-5.0
0
4
8
Dim2 (16%)
Dim1 (33.5%)
Vaccine Development
Non-developed
Developed
B
Germany
Belgium
China
Denmark
Norway
Sweden
Finland
Ireland
Austria
Switzerland
Israel
France
Korea, Rep.
Portugal
Spain
Canada
United Kingdom
Italy
Malaysia
Japan
Kazakhstan
Australia
Cuba
Argentina
South Africa
Russian Federation
Mexico
Indonesia
Brazil
United States
Iran, Islamic Rep.
India
2.5
0.0
-2.5
-5.0
0
4
8
Dim 2 (16%)
Dim 1 (33.5%)
Vaccine Development
Developed
Non-developed
Cluster
1
2

Figure 1. Unsupervised clustering methods. (A) Principal component analysis. (B) K-means clustering.

For numeric variables, a two-sample t-test and the non-parametric Wilcoxon rank-sum test were applied (Figure 2A). T-test found all R&D indicator variables except scientific and technical journal articles to be non-significantly associated with fast vaccine development. However, Wilcoxon rank-sum test found all R&D indicator variables to be not significant except patent applications (residents), COVID-19 Vaccine R&D Funding (investment in the pharmaceutical industry US$), and scientific and technical journal articles. Fisher's exact test found liability to be significantly (P = 0.0088, odd ratio = 0.05982) associated with fast vaccine development, and the income group was not significant. Furthermore, logistic regression found scientific and technical journal articles, liability, and COVID-19 Vaccine R&D Funding (investment in the pharmaceutical industry US$) to be associated with vaccine development (Figure 2B).

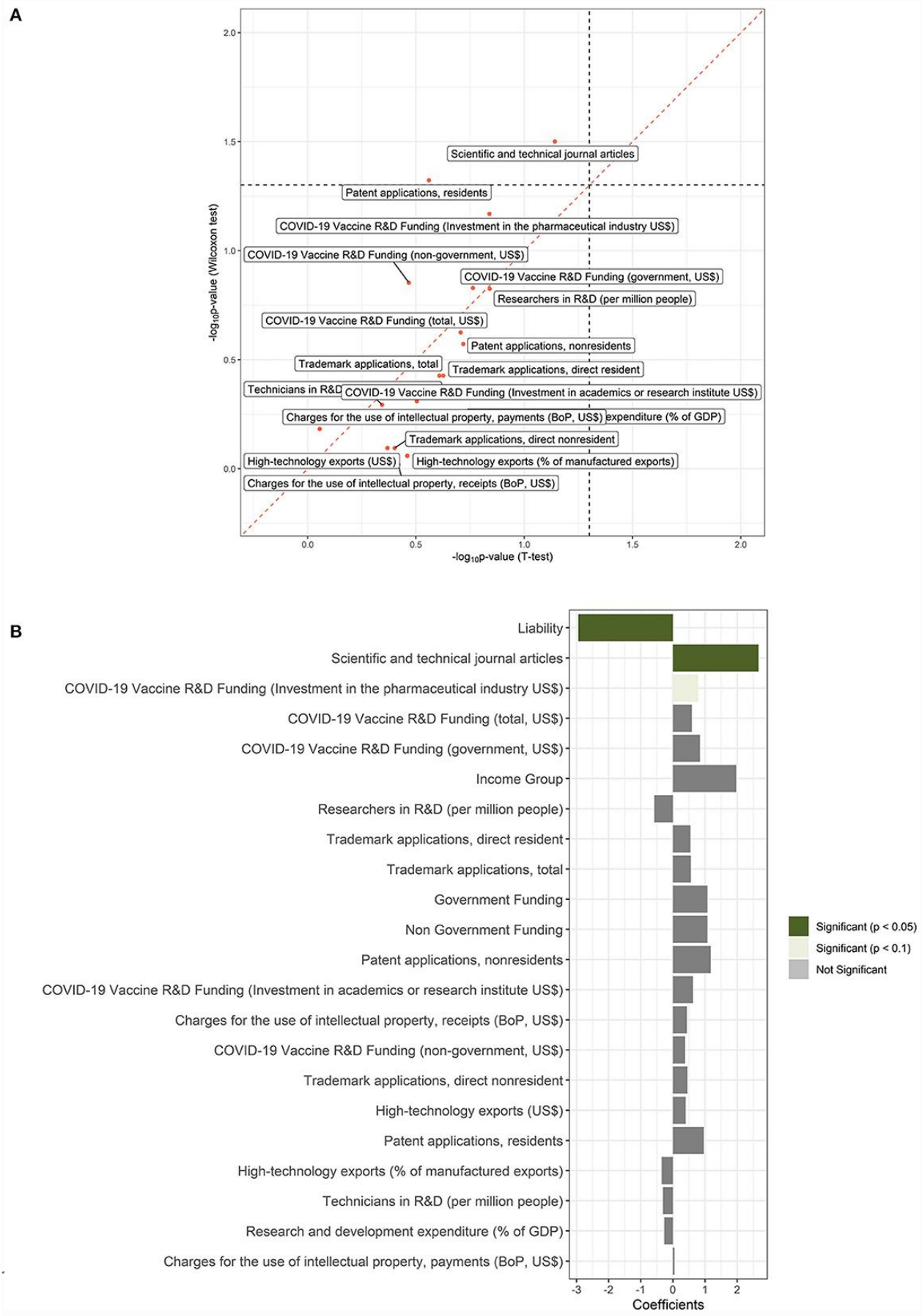

A
Scientific and technical journal articles
Patent applications, residents
COVID-19 Vaccine R&D Funding (Investment in the pharmaceutical industry US$)
COVID-19 Vaccine R&D Funding (non-government, US$)
COVID-19 Vaccine R&D Funding (government, US$)
Researchers in R&D (per million people)
COVID-19 Vaccine R&D Funding (total, US$)
Patent applications, nonresidents
Trademark applications, total
Trademark applications, direct resident
Technicians in R&D
COVID-19 Vaccine R&D Funding (Investment in academics or research institute US$)
Charges for the use of intellectual property, payments (BoP, US$)
expenditure (% of GDP)
Trademark applications, direct nonresident
High-technology exports (US$)
High-technology exports (% of manufactured exports)
Charges for the use of intellectual property, receipts (BoP, US$)
-log10p-value (Wilcoxon test)
-log10p-value (T-test)
B
Liability
Scientific and technical journal articles
COVID-19 Vaccine R&D Funding (Investment in the pharmaceutical industry US$)
COVID-19 Vaccine R&D Funding (total, US$)
COVID-19 Vaccine R&D Funding (government, US$)
Income Group
Researchers in R&D (per million people)
Trademark applications, direct resident
Trademark applications, total
Government Funding
Non Government Funding
Patent applications, nonresidents
COVID-19 Vaccine R&D Funding (Investment in academics or research institute US$)
Charges for the use of intellectual property, receipts (BoP, US$)
COVID-19 Vaccine R&D Funding (non-government, US$)
Trademark applications, direct nonresident
High-technology exports (US$)
Patent applications, residents
High-technology exports (% of manufactured exports)
Technicians in R&D (per million people)
Research and development expenditure (% of GDP)
Charges for the use of intellectual property, payments (BoP, US$)
Coefficients
Significant (p < 0.05)
Significant (p < 0.1)
Not Significant

Figure 2. Results of association analysis. (A) -log10 (P-value) values of the continuous variables of the t-test vs. the Wilcoxon test. (B) Results from the simple logistic regression. The bars show the effect sizes and the colors show the level of significance by P-values. Significance means $P < 0.05$, Evidence means $P < 0.1$ and not-significance means $P > 0.1$.

Vaccine policy and age-group analysis

Analysis of the three vaccine policies using the GEE approach found only vaccine prioritization to be significant with no lagging and at 12 weeks lag, albeit not a negative relationship (Figure 3A). However, the grouping of the population vaccination rate and the confirmed cases into specific age-groups revealed the impact of vaccination in lowering the rate of new confirmed cases for all age-groups except the >79 years age-group, especially among the population aged 25–49 and 50–64 years (Figure 4A). For the age-group >79 years, though the relationship between vaccination and the rate of new cases is significant, vaccination does not have a lowering effect on the rate of COVID-19 cases. The effect size is small (0.025) and the relationship with vaccination is positive. For the combined age-groups, <25 years does not have a significant relationship between vaccination and the rate of new cases. But, we observe the > 65 years age-group shows a lowering impact of vaccination on the rate of new COVID-19 confirmed cases (Figure 4A).

Figure 3

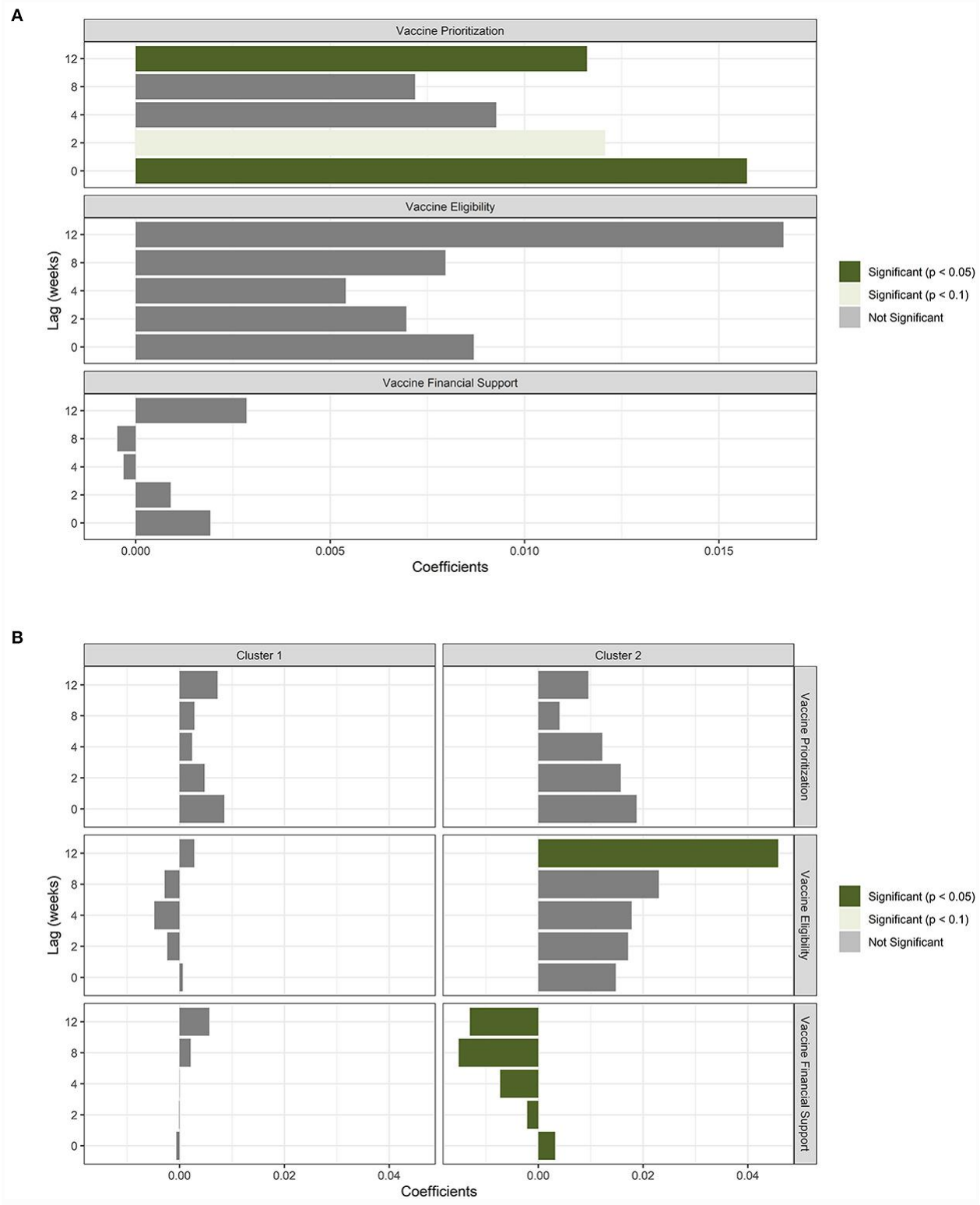


Figure 3. Vaccine policy results in lagging.

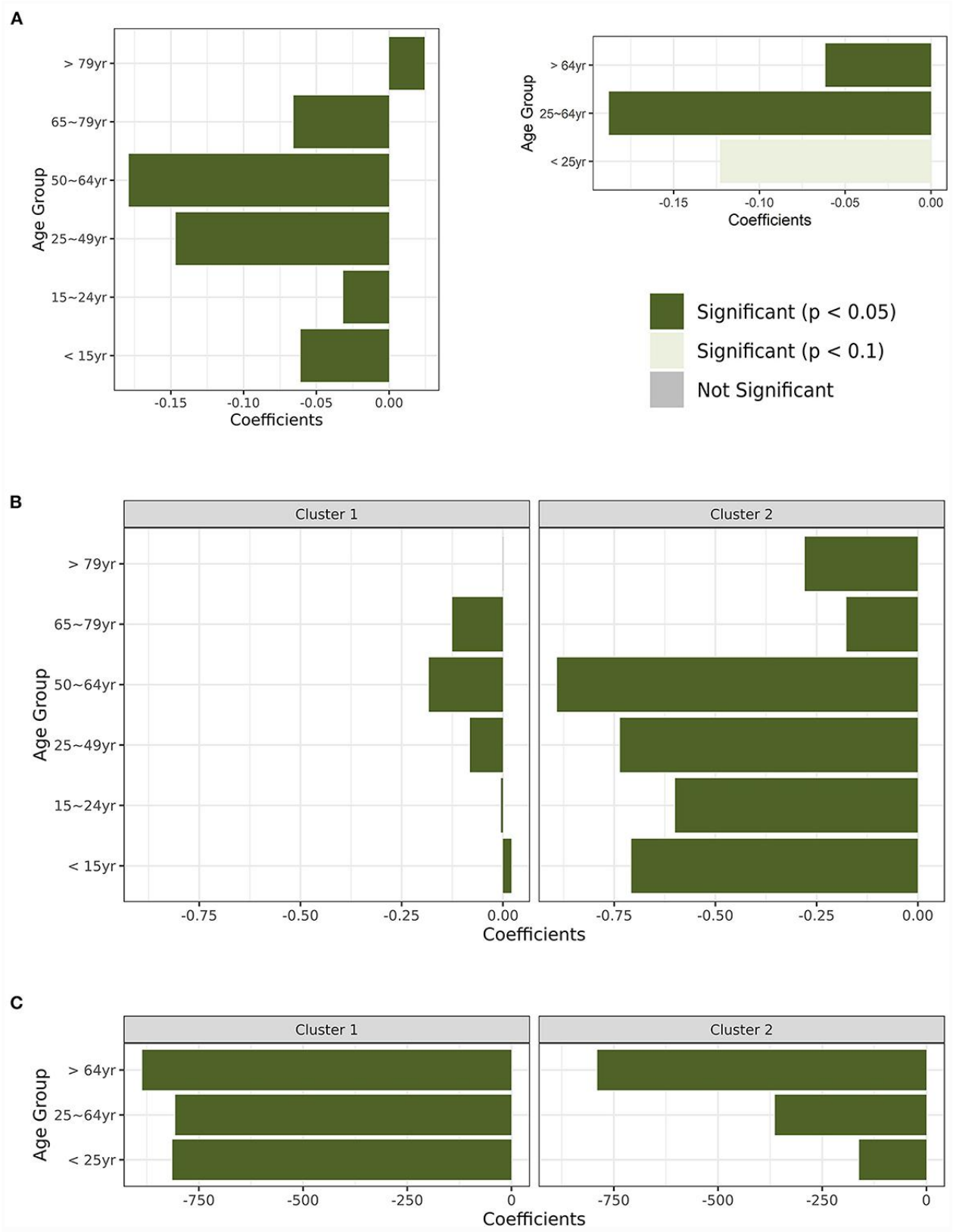

A
> 79yr
65~79yr
50~64yr
25~49yr
15~24yr
< 15yr
Age Group
-0.15
-0.10
-0.05
0.00
Coefficients
> 64yr
25~64yr
< 25yr
Age Group
-0.15
-0.10
-0.05
0.00
Coefficients
Significant (p < 0.05)
Significant (p < 0.1)
Not Significant
B
Cluster 1
Cluster 2
> 79yr
65~79yr
50~64yr
25~49yr
15~24yr
< 15yr
Age Group
-0.75
-0.50
-0.25
0.00
Coefficients
C
Cluster 1
Cluster 2
> 64yr
25~64yr
< 25yr
Age Group
-750
-500
-250
0
Coefficients

Figure 4. Results for specific age-group analysis of vaccination on the rate of new COVID-19 cases. (A) All countries with six age-groups and three age-groups. (B) Cluster groups with six age-groups. (C) Cluster groups with three age-groups.

K-means clustering analysis clustered the countries into two clusters with cluster 1 having 32 countries and cluster 2 having three countries (China, USA, and Germany). Contrasting the three vaccine policies between the two clusters found no policy significant in the Cluster 1 countries. But, cluster 2 countries' analysis found vaccine eligibility at the 12 weeks lag and vaccine financial support at all lag points to be significantly and negatively associated with COVID-19 daily confirmed cases (Figure 3B). For specific age-group analysis, age-group data was provided for only European countries with 12 out of 30 countries belonging to cluster 1 and 1 country (Germany) to cluster 2. The two clusters yielded similar results of the significantly lowering impact of vaccination on the rate of new COVID-19 cases (Figures 4B,C).

PREDICTIVE ANALYSIS

SIR (susceptible–infected–recovered) epidemic model is a very popular and effective model for understanding the dynamics of the spread of any epidemic disease (Chowell et al., 2016). The SIR model can be easily modified to include the vaccination term, if V_S is the vaccination shots per day then the coupled set of equations of SIR model can be modified (Chowell et al., 2016) as follows-

here, the first equation is the susceptibility rate equation with $S(t)$ is the susceptible population at time t, the second and the third equations describe the infection and recovery rate respectively, with $I(t)$ and $R_P(t)$ are the infected and recovered population

at time t. In these equations, β is the transmission rate, γ_0 is the recovery rate, and N_T is the total population in any region. Here, we are assuming that the new births and deaths due to ageing, accidents, non-epidemic diseases, etc. are negligible. Then the total population N_T is always constant, so we have

Eqs. (1), (2), (3) can be reduced to the normalized form by using $s(t) = S(t)/N_T$, $i(t) = I(t)/N_T$, $r_p(t) = R_p(t)/N_T$, and $v = V_s/N_T$, and it can be written as

The transmission parameter, β is a dynamic parameter and dependent over the infected and recovered fraction of people at any time t. Thus, we have taken two different forms of β as discussed in paper (Chakravarty et al., n.d.), i.e.

here, the 9(a) form of β is suitably applied when the epidemic prevention is strict and inhibitive, and the 9(b) is applicable when the preventive measures are less restrictive that allows selective movements depending on the active infection reported in the region (Chakravarty et al., n.d.). The first term β_0 is the *effective infection transmission* (*EIT*) parameter of the epidemic, this term may change depending on the restriction imposed by the government and implementation of social distancing and the second term of β is the additional response term (*RT*) exhibited by the people on increase of cumulative ($i(t) + r_p(t)$) and active cases $i(t)$ following in response of the government restricted measures. The response term has two parameters, c and m in 9(a) and 9(b). c indicates the *strength of infection inhibition or control*, and the larger value implies better disease control. m represents the *sluggishness or promptness of response*, and a smaller value of m implies a broader peak around maximum infection and vice versa.

From Eq. (7), we obtain that the infection increases exponentially at the early epidemic growth phase if we assume that the susceptible population is equal to the total population of any city, region, or country ($S(0) \approx N_T$) at time $t = t_0$, then we have

Thus, from the above expression, we found that the number of secondary cases generated per primary case depends on the ratio of β_0 and γ_0, which is called a

reproduction number R_0, and is given by (Chowell et al., 2016).

Thus, the secondary cases increase if R_0 is greater than 1 and it decays when it is less than 1. Hence the R_0 is the special parameter to study the transmission of infection in any area. But the number of susceptible people decreases with time due to the increase in infection, and the effective reproduction number over time R_t, is given by the product of R_0 and the proportion of susceptible individuals in the population (Chowell et al., 2016).

The above expression shows that the effective reproduction number R_t is a function of susceptible population, recovered population, infected population, effective transmission parameter, recovery rate, strength of infection inhibition or control and promptness of response parameter. When the strength of infection control measure is large then R_t reduces. Thus, R_t describes the dynamics of the transmission of infection in the population. If $R_t > 1$, then the number of infections increases and when $R_t < 1$, then it decays (Volpert et al., 2020). Also, the disease will not spread if R_0 is less than 1 (Kwang et al., 2013).

The variation in R_t for different values of c is shown in Fig. 1(a) when m is 1, for the case when the transmission parameter is $\beta(t) = \beta_0 - c(i(t) + r_p(t))_m$. The black solid straight line at 3 shows the R_0 curve which remains constant due to its dependency over the effective infection transmission and recovery parameter. We noticed the vertical stretch in the R_t curves for different c values (infection inhibition strength parameter), and it maintains the slope as m is same for all curves. R_t varies from 2.991 to 2.844, 2.715, 2.04, and 1.614 for the respective c values 0, 5, 10, 50, and 100 at t = 50 days. R_t decreases from 0.3323 to 0.1896 as c changes from 0 to 100 when t is 100. Thus, as the value of c i.e., the infection inhibition strength parameter increases, the value of response term increases, and hence R_t decreases (see expression (13)). R_t reaches first to 0.2 for the higher inhibition strength parameter. It means the epidemic eradicates faster for higher value of inhibition strength parameter. Fig. 1(b) has been plotted to show the variation in R_t with days for different sets of response term, c and m. Initially, the steepness of the curve decreases with m (promptness of response parameter) and then increases as m reaches near to 0 value. Thus, the promptness of response decides the

steepness or rate of decrease of R_t curve.

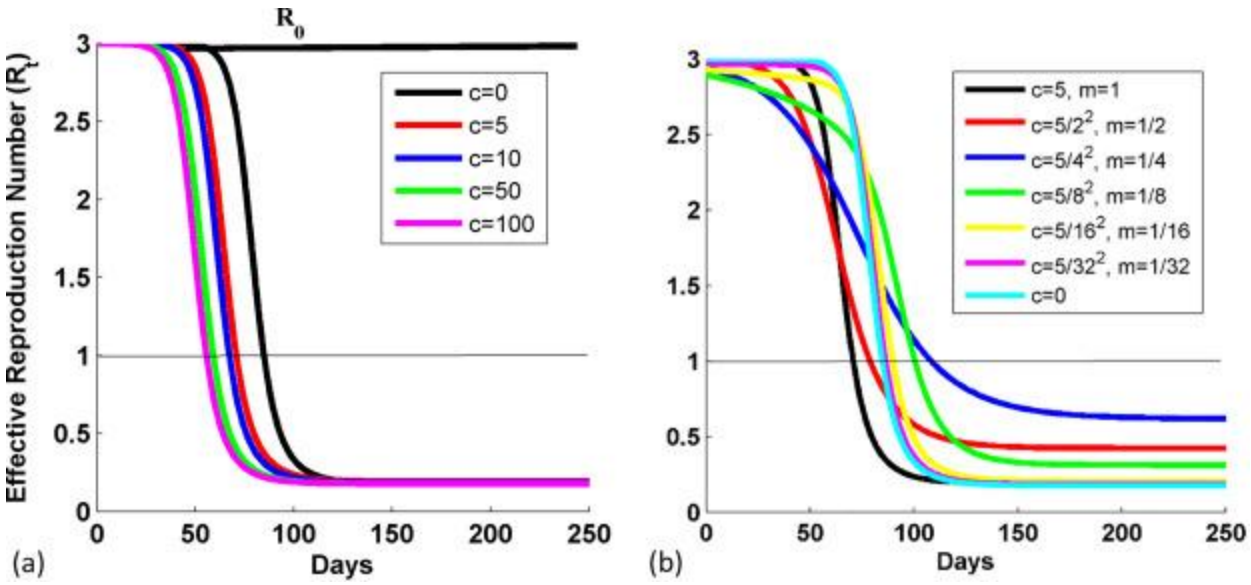


Fig. 1

We also plotted the variation in R_t for different values of c when m is 1, and for different sets of response term in Fig. 2(a) and (b), when the transmission parameter is $\beta(t) = \beta_0 - ci(t)_m$. We see that the lower part of the curve shifts upward for nonzero values of c, when m is 1 (see Fig. 2(a)), and as m decreases, the lower part of the curve shifts downward (see Fig. 2(b)). It means that when infection inhibition strength parameter is zero, then R_t reduces to 0.2, whereas it lies near to 1, when infection inhibition strength parameter is non zero. As the promptness of response decreases, the R_t tends to shift downward. The epidemic eradicates faster for the case either when c is 0 or when m or c tend towards 0.

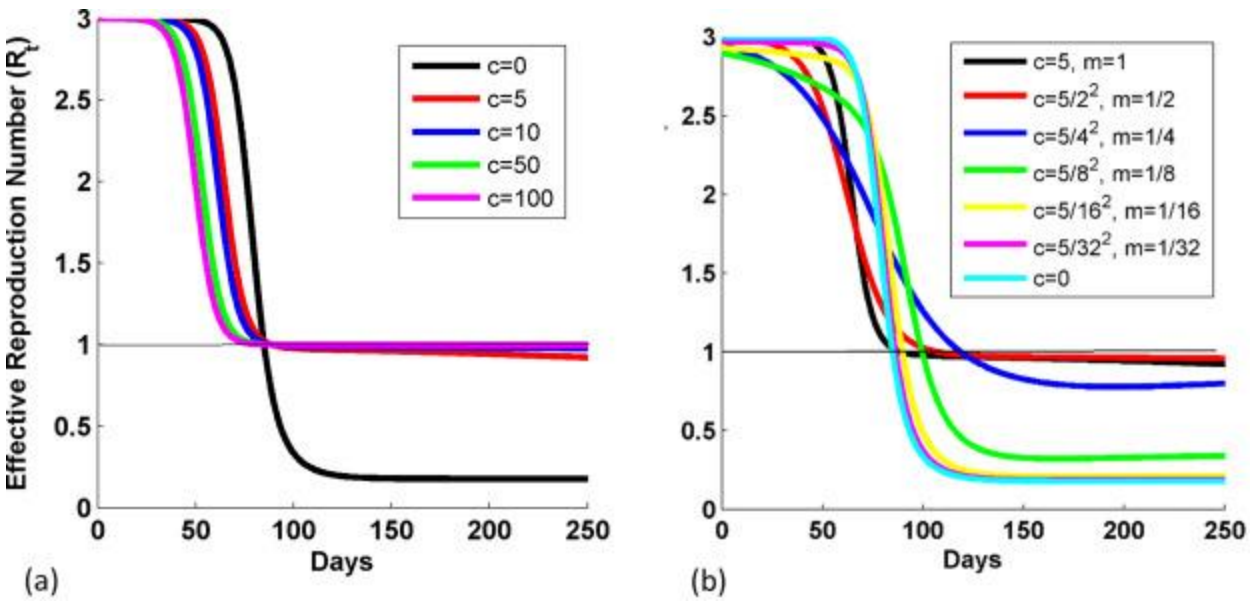


Fig. 2

Code:

Importing Libraries:

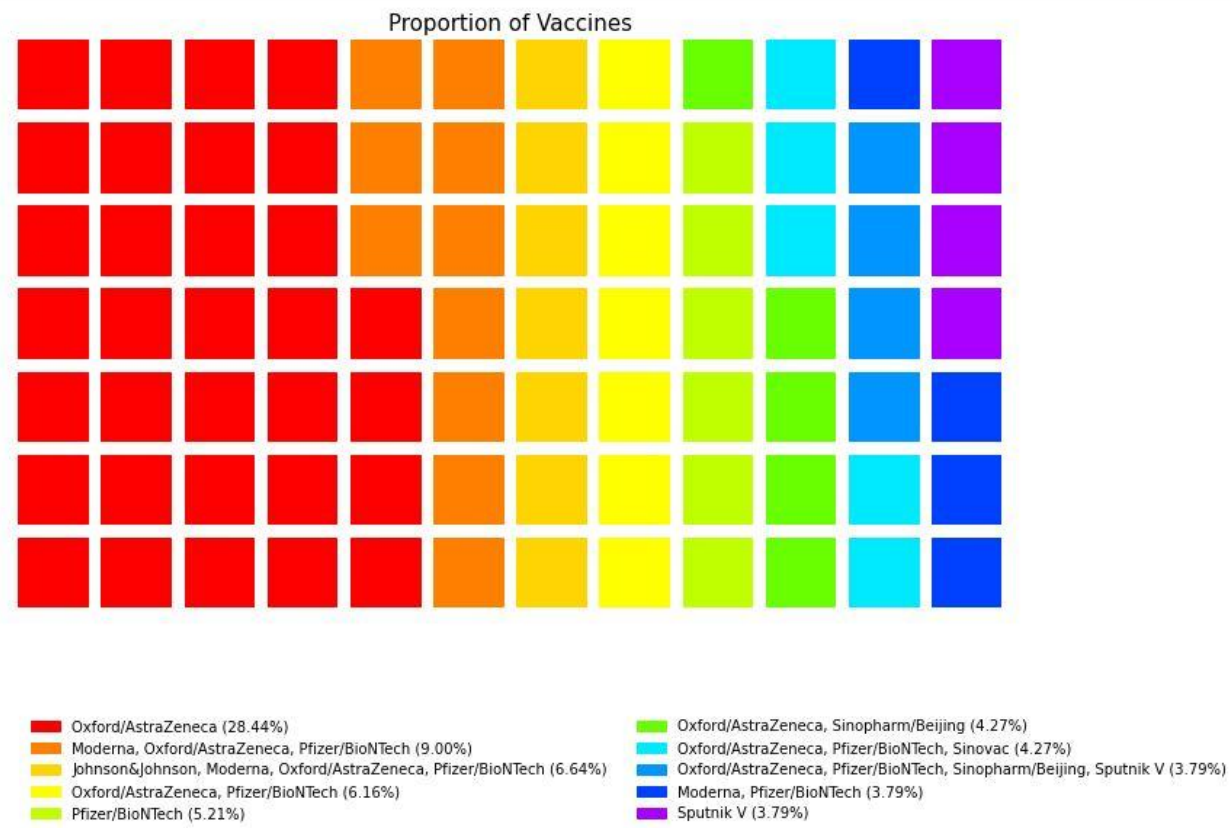
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g.
pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from plotly.offline import
download_plotlyjs,init_notebook_mode,plot,iplot
import plotly.graph_objects as go
import plotly.figure_factory as ff
from plotly.colors import n_colors
from wordcloud import WordCloud,ImageColorGenerator
init_notebook_mode(connected=True)
```

```
from plotly.subplots import make_subplots
from pywaffle import Waffle
import warnings
warnings.filterwarnings("ignore")
top10 = new_df['vaccines'].value_counts().nlargest(10)
top10
```

```
Oxford/AstraZeneca                                                57
Moderna, Oxford/AstraZeneca, Pfizer/BioNTech                      20
Oxford/AstraZeneca, Pfizer/BioNTech                               13
Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech     12
Pfizer/BioNTech                                                   12
Oxford/AstraZeneca, Sinopharm/Beijing                              8
Sinopharm/Beijing                                                  8
Sputnik V                                                          8
Moderna, Pfizer/BioNTech                                           6
Oxford/AstraZeneca, Pfizer/BioNTech, Sinovac                       6
Name: vaccines, dtype: int64
```

```
data = dict(new_df['vaccines'].value_counts(normalize =
True).nlargest(10)*100)
#dict(new_df['vaccines'].value_counts(normalize = True) * 100)
vaccine = ['Oxford/AstraZeneca', 'Moderna, Oxford/AstraZeneca,
Pfizer/BioNTech',
       'Oxford/AstraZeneca, Pfizer/BioNTech',
       'Johnson&Johnson, Moderna, Oxford/AstraZeneca,
Pfizer/BioNTech',
       'Pfizer/BioNTech', 'Sputnik V', 'Oxford/AstraZeneca,
Sinopharm/Beijing',
       'Sinopharm/Beijing', 'Moderna, Pfizer/BioNTech',
       'Oxford/AstraZeneca, Pfizer/BioNTech, Sinovac']
fig = plt.figure(
    rows=7,
    columns=12,
    FigureClass = Waffle,
```

```
    values = data,
    title={'label': 'Proportion of Vaccines', 'loc': 'center',
           'fontsize':15},
    colors=("#FF7F0E", "#00B5F7",
"#AB63FA","#00CC96","#E9967A","#F08080","#40E0D0","#DFFF00","#DE
3163","#6AFF00"),
    labels=[f"{k} ({v:.2f}%)" for k, v in data.items()],
    legend={'loc': 'lower left', 'bbox_to_anchor': (0, -0.4),
'ncol': 2, 'framealpha': 0},
    figsize=(12, 9)
)
fig.show()
```

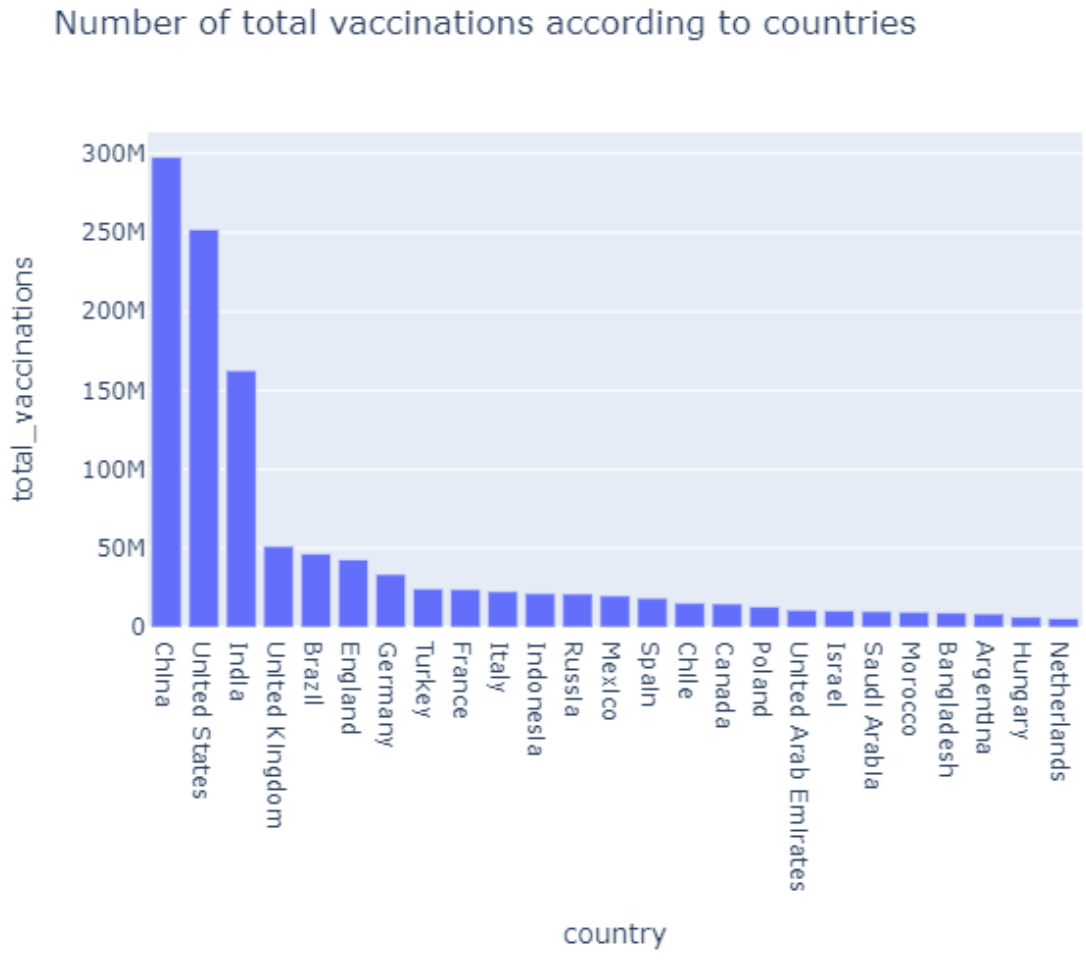


Observation:

- In a range of percentage of vaccines 28.44% used Oxford/AstraZeneca
- Oxford/AstraZeneca is the most used Vaccine

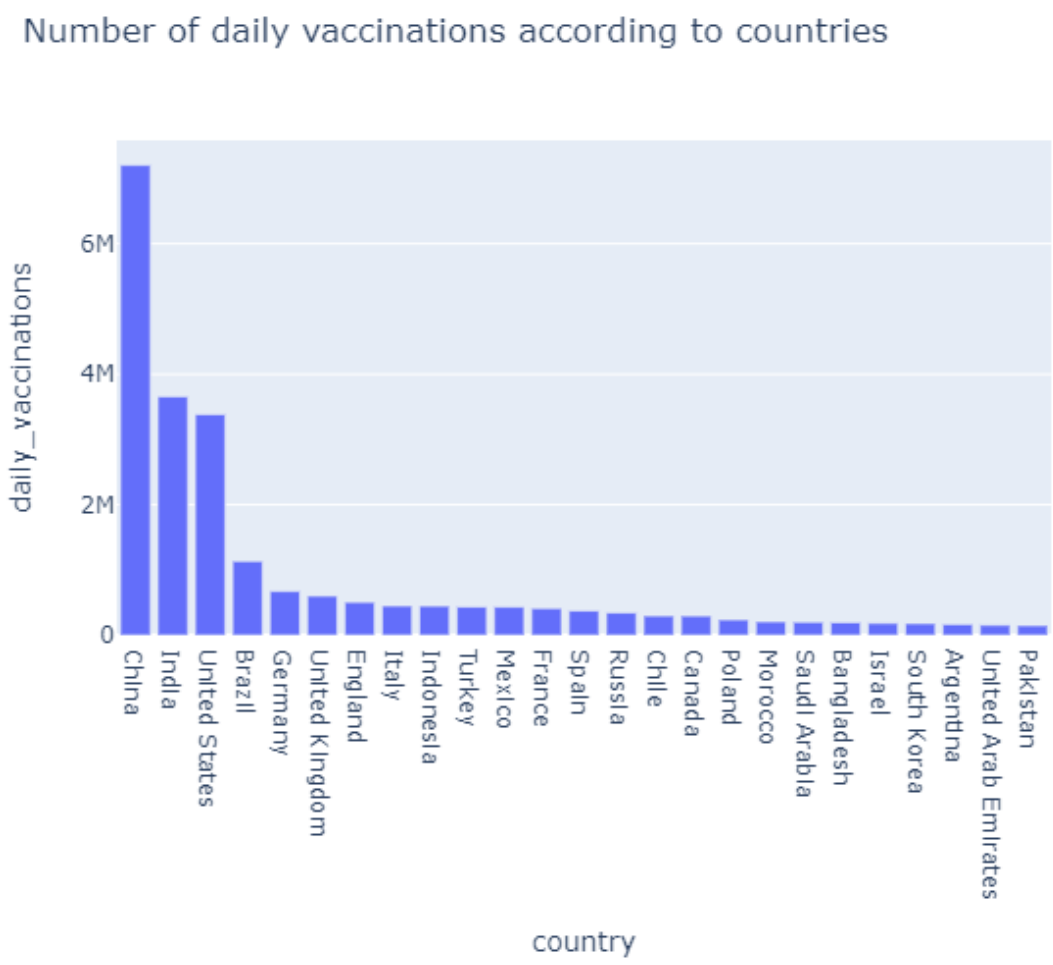
- Later Pfizer/BioNTech was the most used Vaccine and now it's in 5th place also Oxford/AstraZeneca was not in the top 3 & now it's in 1st place. Looks like Oxford/AstraZeneca works best among the vaccines

```
data =
new_df[['country','total_vaccinations']].nlargest(25,'total_vacc
inations')
fig = px.bar(data, x = 'country',y =
'total_vaccinations',title="Number of total vaccinations
according to countries",)
fig.show()
```




```
data =
```

```
new_df[['country','daily_vaccinations']].nlargest(25,'daily_vacc
inations')
fig = px.bar(data, x = 'country',y =
'daily_vaccinations',title="Number of daily vaccinations
according to countries",)
fig.show()
```



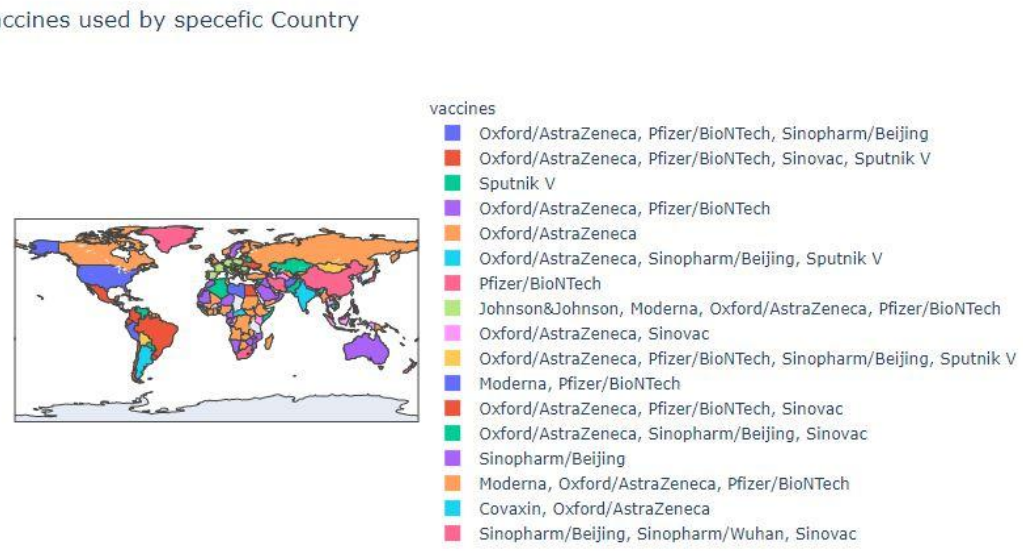
Which vaccine is used by which Country?

```
vacc = new_df["vaccines"].unique()
for i in vacc:
    c = list(new_df[new_df["vaccines"] == i]['country'])
    print(f"Vaccine: {i}nUsed countries: {c}")
```

```
print('-'*70)
```

```
Vaccine: Sputnik V
Used countries: ['Argentina', 'Russia']
----------------------------------------------------------------------
Vaccine: Pfizer/BioNTech
Used countries: ['Austria', 'Belgium', 'Chile', 'Costa Rica', 'Croatia', 'Cyprus', 'Estonia', 'Finland', 'France', 'Gibraltar',
'Greece', 'Hungary', 'Ireland', 'Israel', 'Italy', 'Kuwait', 'Latvia', 'Luxembourg', 'Malta', 'Mexico', 'Netherlands', 'Norwa
y', 'Oman', 'Poland', 'Portugal', 'Romania', 'Saudi Arabia', 'Serbia', 'Singapore', 'Slovakia', 'Slovenia', 'Sweden', 'Switzerl
and']
----------------------------------------------------------------------
Vaccine: Pfizer/BioNTech, Sinopharm
Used countries: ['Bahrain', 'United Arab Emirates']
----------------------------------------------------------------------
Vaccine: Sinovac
Used countries: ['Brazil', 'Turkey']
----------------------------------------------------------------------
Vaccine: Moderna, Pfizer/BioNTech
Used countries: ['Bulgaria', 'Canada', 'Czechia', 'Denmark', 'Germany', 'Iceland', 'Lithuania', 'Spain', 'United States']
----------------------------------------------------------------------
Vaccine: CNBG, Sinovac
Used countries: ['China']
----------------------------------------------------------------------
Vaccine: Covaxin, Covishield
Used countries: ['India']
----------------------------------------------------------------------
Vaccine: Sinopharm
Used countries: ['Seychelles']
----------------------------------------------------------------------
Vaccine: Oxford/AstraZeneca, Pfizer/BioNTech
Used countries: ['United Kingdom']
----------------------------------------------------------------------
```

```
fig = px.choropleth(new_df,locations = 'country',locationmode =
'country names',color = 'vaccines',
                    title = 'Vaccines used by specefic
Country',hover_data= ['total_vaccinations'])
fig.show()
```



Which Vaccine is Used the most?

```
vaccine = new_df["vaccines"].value_counts().reset_index()
vaccine.columns = ['Vaccines','Number of Country']
vaccine.nlargest(5,"Number of Country")
```

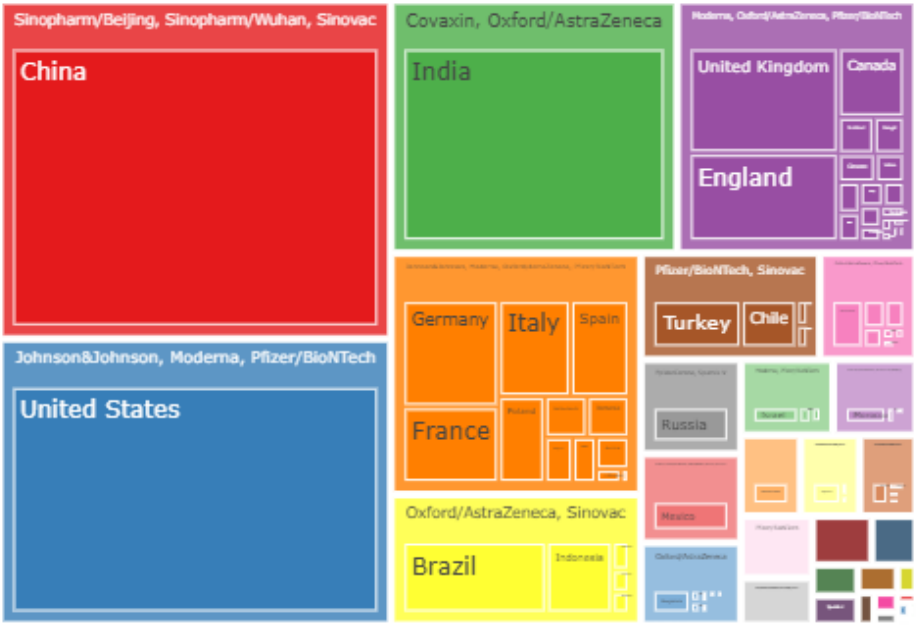
	Vaccines	Number of Country
0	Oxford/AstraZeneca	60
1	Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	19
2	Johnson&Johnson, Moderna, Oxford/AstraZeneca, ...	14
3	Oxford/AstraZeneca, Pfizer/BioNTech	13
4	Pfizer/BioNTech	11

Oxford/AstraZeneca is being used by 60 Countries.

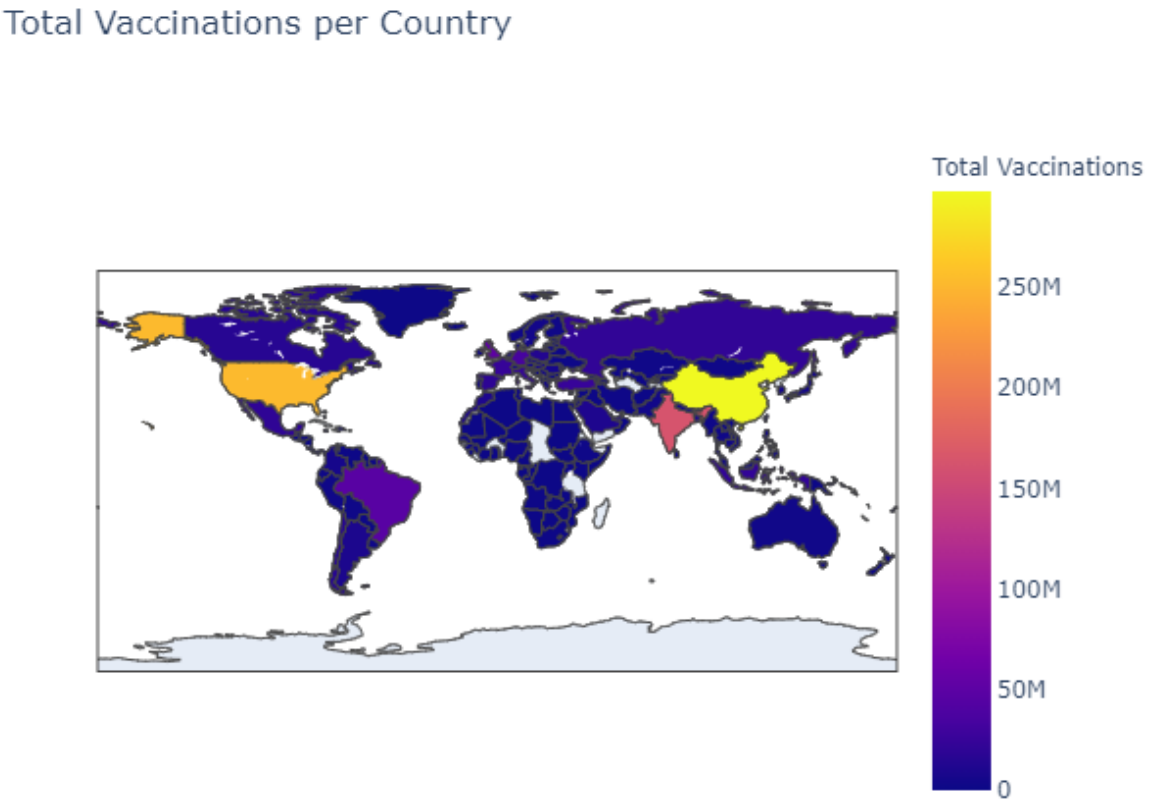
Total Vaccinations per country grouped by Vaccines:

```
fig = px.treemap(new_df,names = 'country',values =
'total_vaccinations',
                 path = ['vaccines','country'],
                 title="Total Vaccinations per country grouped
by Vaccines",
                 color_discrete_sequence
=px.colors.qualitative.Set1)
fig.show()
```

Total Vaccinations per country grouped by Vaccines

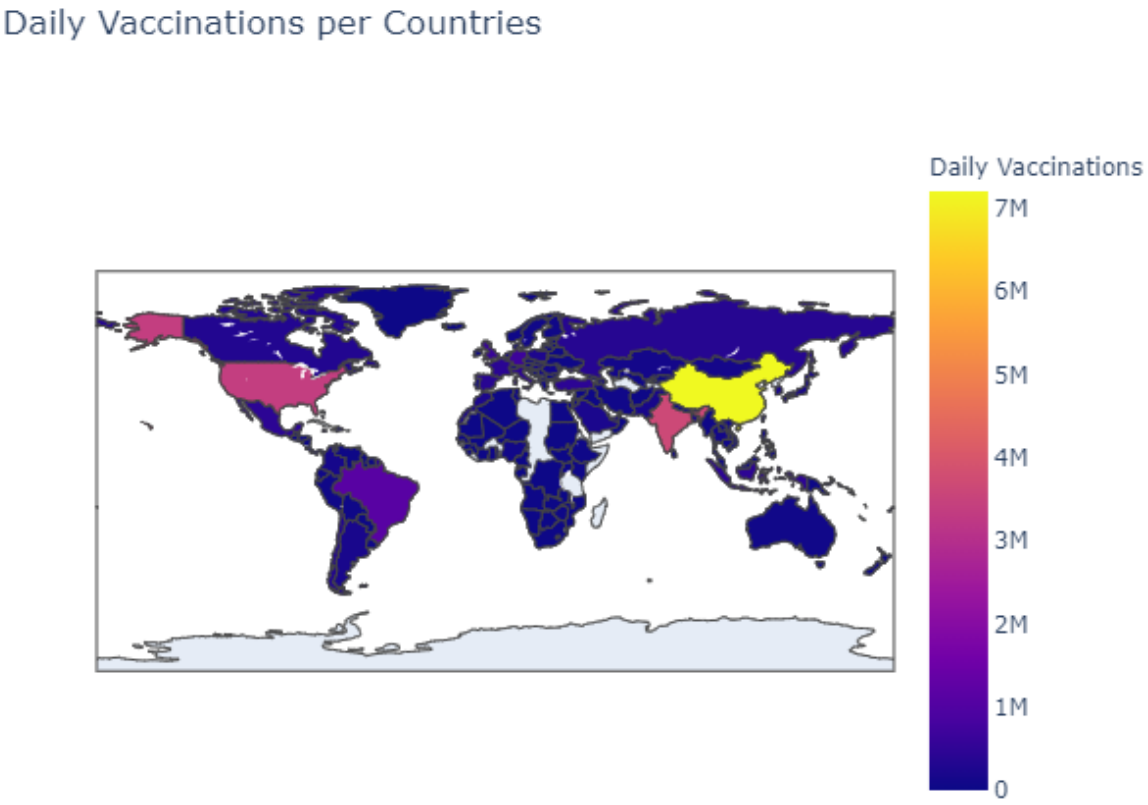



```
fig = go.Choropleth(locations = new_df["country"],locationmode =
'country names',
                    z = new_df['total_vaccinations'],
                    text= new_df['country'],colorbar =
dict(title= "Total Vaccinations"))
data = [fig]
layout = go.Layout(title = 'Total Vaccinations per Country')
fig = dict(data = data,layout = layout)
iplot(fig)
```



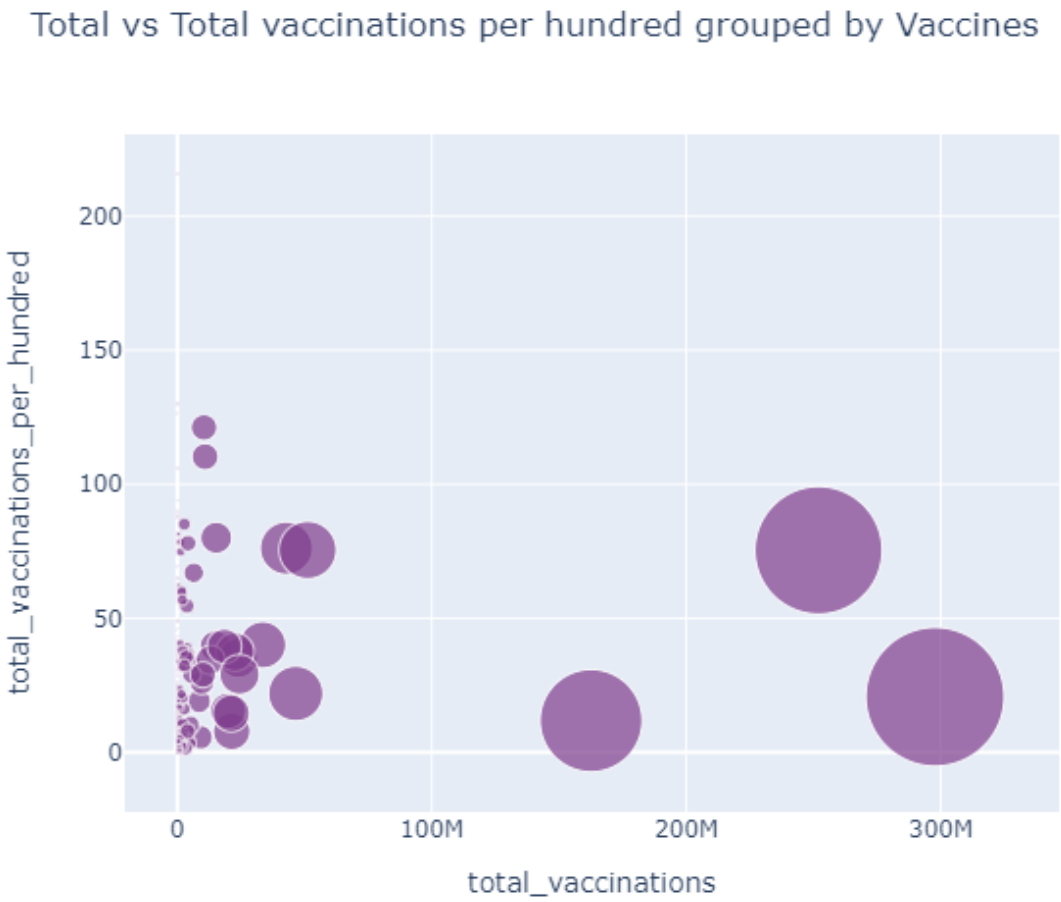
Daily Vaccinations per Countries:

```
fig = go.Choropleth(locations = new_df["country"],locationmode =
'country names',
                    z = new_df['daily_vaccinations'],
                    text= new_df['country'],colorbar =
dict(title= "Daily Vaccinations"))
data = [fig]
layout = go.Layout(title = 'Daily Vaccinations per Countries')
fig = dict(data = data,layout = layout)
iplot(fig)
```



Relation between Total Vaccinations and Total Vaccinations per Hundred:

```
fig = px.scatter(new_df,x =
'total_vaccinations',y='total_vaccinations_per_hundred',
                 size='total_vaccinations',
                 hover_name = 'country',size_max = 50,
                 title="Total vs Total vaccinations per hundred
grouped by Vaccines",
                color_discrete_sequence =
px.colors.qualitative.Bold)
fig.show()
```



If you hover your cursor to the scatters you will also see the country names,number of total vaccinations and number of total vaccinations per hundred. By this we observe that:

- Although USA & China produce the highest number of vaccinations to their citizens, according to their population this is not much.

What is the trend of total vaccinations according to countries?

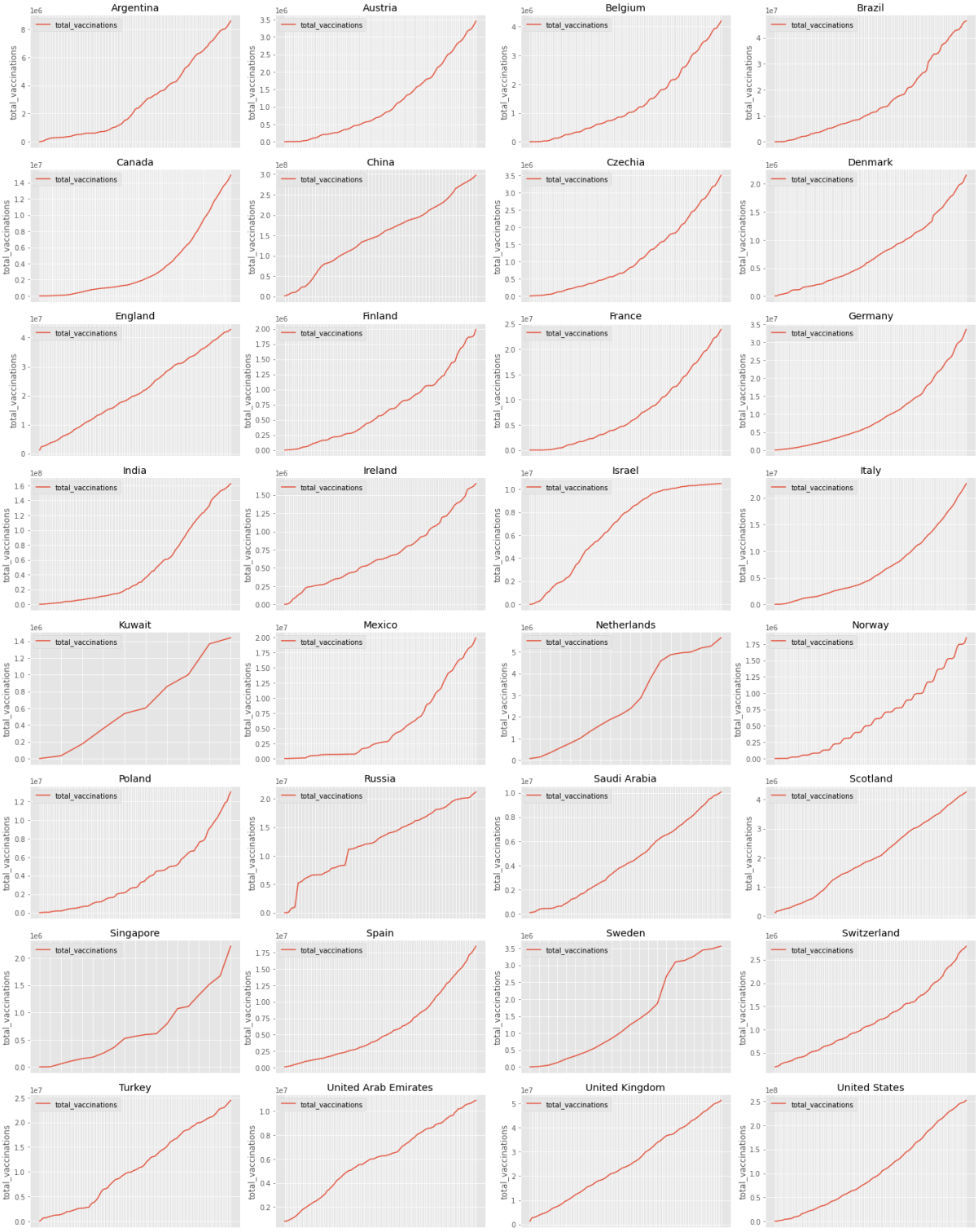
```
def plot_trend(dataframe,feature,title,country):
    plt.style.use('ggplot')
```

```
    plt.figure(figsize=(20,25))

    for i,country in enumerate(country):
        plt.subplot(8,4,<a
onclick="parent.postMessage({'referent':'.kaggle.usercode.144406
04.62732853.plot_trend..i'}, '*')">i+1)
        data = dataframe[dataframe['country'] == country]
        sns.lineplot(x=data['date'] ,y=data[feature],label =
feature)
        plt.xlabel('')

plt.tick_params(axis='x',which='both',top=False,bottom=False,lab
elbottom=False)
        plt.title(country)

    plt.suptitle(title,y=1.05)
    plt.tight_layout()
    plt.show()
country = ['Argentina', 'Austria', 'Belgium',
'Brazil','Canada','China','Czechia',
           'Denmark', 'England','Finland',
'France','Germany','India','Ireland',
           'Israel', 'Italy', 'Kuwait','Mexico',
'Netherlands','Norway', 'Poland',
           'Russia','Saudi Arabia',
'Scotland','Singapore','Spain', 'Sweden',
           'Switzerland', 'Turkey','United Arab Emirates',
'United Kingdom', 'United States']
plot_trend(df,'total_vaccinations','Trend of total
vaccination',country)
```


Trend of total vaccination
Argentina
Austria
Belgium
Brazil
Canada
China
Czechia
Denmark
England
Finland
France
Germany
India
Ireland
Israel
Italy
Kuwait
Mexico
Netherlands
Norway
Poland
Russia
Saudi Arabia
Scotland
Singapore
Spain
Sweden
Switzerland
Turkey
United Arab Emirates
United Kingdom
United States
total_vaccinations

CONCLUSION

In this study, visualize and debate the current state of COVID-19 vaccination in terms of the proportion of top 10 vaccines in the race to combat COVID-19, the number of cumulative vaccinations and every day vaccinations as per the countries, cumulative vaccinations per country grouped by vaccines, daily vaccinations per countries, and the relationship among cumulative vaccinations and cumulative vaccinations per hundred of the top five countries seriously affected by the COVID-19 globally as of May 24, 2021, including the United States, India, Brazil, France, and Turkey. The statistics reveal 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 anticipated 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. However, this will not be enough to save the public, and policymakers in India, Brazil, and Turkey should take the necessary steps.