

Toward Data Visualization and Data Forecasting with COVID-19 Vaccination Statistics



Vaishnavi Kulkarni, Jay Kulkarni, and Anurag Kolhe

Abstract The world runs on data. Various organizations, businesses, and institutions utilize and generate data. This information is a valuable commodity if availed of in the right way. Big data can be large and incomprehensible on its own, but when analyzed computationally, it can be a powerful tool for revealing patterns and trends, forecasting future values of certain data parameters as well as providing clarity about the metrics in the data. Data visualization and forecasting using such data are fields that have applications in every sector—from information technology, to education, to healthcare. Since the world was hit by the debilitating COVID-19 pandemic in 2019, life has become a blur of statistics—daily new case counts, daily deaths and recoveries, number of people vaccinated, etc. Such data are of paramount importance to everyone affected by the pandemic, and presenting it in a way that is easily understandable to a layperson and using it to glean insights into the spread and curb of the disease as well as the efficacy of the vaccines is necessary. This paper takes COVID vaccination statistics as a use case for the fields of data visualization and data forecasting. It elucidates the methodology and benefits of both interactive visualizations of vaccination data and forecasting future trends in vaccine and case metrics based on data over time.

Keywords Data visualization · Data forecasting · COVID vaccinations · Polynomial regression · Machine learning · Data science

1 Introduction

With the exponential growth of the Internet, tremendous volumes of data are getting generated each day. According to a Forbes article [1], 2.5 quintillion bytes of data

V. Kulkarni (✉) · J. Kulkarni
Pune, India
e-mail: vaishnavikulkarni@hotmail.com

A. Kolhe
Akola, India

are created each day and this pace is increasing every day. This data are related to professional fields like medicine, technology, science, communication, engineering, etc., along with other fields like social media, news, sports, etc. It is generally difficult to extract useful information from these heaps of large data. Hence, to make sense of these large volumes of data present, data visualization comes into the picture. Data visualization is the graphical representation of any given information. It gives us a clear picture of what the data are representing by using visual elements like graphs, maps, charts, etc. Properly organizing all the data based on some parameters into graphs or charts helps us to comprehend the meaning of it. Further, it proves beneficial for deriving trends and detecting anomalies in the available information.

Today, when we talk about the tremendous volumes of data, big data are the word that comes first into the picture. Big data refer to the flood of structured and unstructured digital data that are gathered by organizations day in and day out. It can be characterized by the 5 versus of the big data—volume, velocity, variety, veracity, and value. Volume is the volume of produced data. Velocity is the pace of generating and moving data. Variety denotes the diverse type of data collected from varied sources. Veracity is the credibility of the collected data. Finally, value refers to the importance to the data to company's business.

The traditional storage system cannot be used to store big data. Large storage space is required to store such gigantic data. Similarly, powerful processing systems, analytical tools, databases for quick access of information, etc., are essential when dealing with big data. This is where cloud computing comes into the picture. Renowned cloud services such as Azure, AWS, and GCP provide the user with servers, databases, storage, networking, analytics, etc. The cloud infrastructure allows for real-time processing of big data. Data could be efficiently and quickly stored in the storage facilities of cloud services. Similarly, data could be retrieved and looked upon quickly and further could also be interpreted in real-time. Data are fragmented and stored securely on these cloud platforms. This fragmentation process is, however, dispersed and scattered, thus lacking proper order. This increases the time in information collection. In [2], hybridized historical aware algorithm (HHAR) is used to minimize the dispersed and scattered packages.

In the world of big data and data science, data visualization is highly significant. Machine learning and deep learning models require a huge amount of data for training and testing purposes. Without initially learning from the immense data, it will not perform as expected and give the required output. Initially, before building any model, we need to analyze the available data over which we will be using different algorithms. This analysis of data is complex and is not possible manually. Data visualization tools can help solve this quandary and help us with the computational analysis of data to present it in a way that is easily understood by the human eye. Real-world data, though complex, may not be outputted in 2- or 3-dimensional spaces. [3] Hence, dimensionality reduction techniques like PCA, TSNE, LDA, etc., could be used to reduce higher dimensional data into 2D or 3D. Before building the models, visualizing this data into graphs may help us to reveal some hidden patterns, clusters that could be formed within the data, etc. We can also figure out whether the data are

linearly separable, or if it overlaps too much, etc. This initial analysis is thus crucial for selecting a particular ML model for a given set of data.

Data forecasting is another area that can exploit ML techniques with the availability of tremendous data. In data forecasting, ML models are trained using past data to analyze the future trends of upcoming data. This is helpful in areas like sales marketing, climate changes, etc. In sales for example, with the help of historical data, predictions like whether a stock price will fall or rise can be done. The more accurately an ML model is trained, the more accurate its forecasting. [4] Forecasting can be classified as time series forecasting and time series classification. In time series forecasting, techniques are used for predicting future values using an ML algorithm. Whereas in time series classification, by looking at past data, techniques are used to classify an item into a particular class.

In 2019, the entire world was hit by the coronavirus pandemic which affected many people, resulting in high mortality rates. It also crippled the economies of many countries. Vaccines have now been administered globally. However, the efficacy of the vaccines can only be decided by analyzing the past data. In this paper, we will be analyzing the data globally for the COVID-19 pandemic. Visualizing this data in the form of charts and graphs helps reveal how inoculation against the disease is proceeding worldwide. Furthermore, we have used the polynomial regression technique to forecast the trend of the number of new COVID cases based on the number of vaccinations.

2 Related Work

For United States, the authors in [5] have proposed a model to accurately forecast the COVID-19 cases and deaths.

The authors did not use a traditional single model to predict the required trend by keeping the parameters fixed. They instead used a “last-fold partitioning” which is a general learner. This gave them the best parameters for their model, the best combination of features, and the forecasting model with best history-length.

Supervised learning ML models are used in [6] to predict an unknown input instance. Learning methods use regression techniques and classification algorithms for predictive models’ development. Four regression models are used to study the proponents of COVID-19. This helps in the forecasting of factors like recently contracted cases, the mortality rate as well as the count of cured cases over a time frame of almost a fortnight.

In [7], a new prediction model is proposed by the combination support vector regression (SVR) model with conventional random vector functional link (RVFL) model to improve the prediction capabilities of COVID-19 cases. RVFL network is hybridized with 1D discrete wavelet transform, and a wavelet coupled RVFL (WCRVFL) network is proposed.

The study in [8] aims to do a time series forecasting to predict the COVID-19 patients’ rise, recovery, and death in India based on the daily data obtained from

the Indian Government. It uses a statistical model called autoregressive integrated moving average (ARIMA) which has proven effective in short-term forecasting in many other diseases. ARIMA models aim to describe the autocorrelations in the data. ARIMA is also combined with an exponential smoothing model which is based on a description of the trend and seasonality in the data.

In [9], the authors rank different ML classification algorithms with the help of the COVID-19 World Vaccination Progress dataset. Four classification algorithms are considered—decision tree, K-nearest neighbor, random tree, and Naïve Bayes. For this comparison, an open-source Java platform WEKA is used. WEKA contains a series of ML algorithms that enable researchers to analyze their data for patterns, trends, etc. After running the algorithms over a dataset consisting of 6745 instances and more than 14 attributes, findings are recorded. The results show that the decision tree classifier's percentage of correctly classified instances was highest among the four. And that for Naïve Bayes was lowest. Accordingly, the root mean square error (RMSE) is lowest for the decision tree classifier, thus making it the best classification algorithm among above 4. Similarly, for Naïve Bayes, the RMSE value was highest.

In [10], the authors predict the death rates as well as survival rates of COVID-19 patients. Supervised machine learning is used to achieve this. These rates are forecasted by considering the effects of chronic diseases, features unique to particular groups of people belonging to similar age-groups, ethnicities, etc., as well as the initial data from clinical trials. COVID-19 samples from the King Fahad University Hospital, Saudi Arabia were used as the key dataset for this study. Patient records were classified into 2 classes—survived and deceased. Dataset is classified on various features like body temperature, shortness of breath, chronic disease like diabetes, etc. As the mortality rate was low, the deceased class contains fewer values and thus causes class imbalance. To resolve this, the synthetic minority oversampling technique (SMOTE) was used. The authors employed three classification algorithms—logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB). For parameter optimization, the grid technique was used. From the results, the authors concluded that the random forest algorithm was the most effective. It performed much more efficiently compared to the other classifiers using the top 20 features with SMOTE data. On the other hand, logistic regression gave the least accurate results.

In [11], authors have used logistic regression for forecasting the probability of recovery for patients with debilitating COVID symptoms. They have used data of 183 patients from Tongji Hospital, Wuhan. Four variables—lymphocyte count, age, d-dimer, and high-sensitivity C-reactive protein were selected and used to fit the logistic regression model. The areas under the receiver operating characteristic curves (AUROCs) in the logistic regression model were 0.895. Other models have also been considered in this paper but the logistic model ultimately decided upon due to it being minimalistic yet effective. For prediction of death of patients, the AUROC of the external validation set, its sensitivity, and its specificity all had values close to 0.8.

To forecast COVID-19 reproduction rate, research in [12] aims at the performance evaluation of various non-linear regression techniques, such as gradient

boosting, KNN, XGBOOST, SVR, and random forest regressor. It also highlights the importance of hyperparameters tuning and feature selection. For the feature selection, methods such as gradient boosting, GBOOST, and random forest are applied. Depending upon the score of feature importance, the top 7 features that affect the rate of reproduction are recognized. Further, four different experiments are conducted with the presence and absence of feature selection. Across all the experiments, RMSE, mean absolute error (MAE), and relative absolute error (RAE) were used to measure the reproduction rate. In all the experiments performed, KNN approach obtained altogether the best values for RAE, RMSE, and MAE. It was succeeded by the XGBOOST and random forest. Further, tuning of hyperparameter with all the features demonstrated a low prediction error rate and the best performance for these parameters was shown by random forest.

In [13], to evaluate the metrics of ML models in analyzing disease infection, an epidemiology Mexico dataset of COVID-19 cases is used. Different supervised ML algorithms are considered for this purpose. Eight clinical features and two demographic features are considered here. 1 is encoded for positive and 0 for negative. The correlation coefficient analysis is carried out to reveal each independent and dependent feature relationship. Here, also the performance of the models is again evaluated based on accuracy, sensitivity, and specificity. The findings show that among all the models, the highest 94.99% accuracy is achieved by the decision tree. In terms of sensitivity, SVM is better than the rest with a value of 93.34%, and Naïve Bayes achieves the highest specificity of 94.3%. Also, “age” is extracted as the most significant dependent feature by the decision tree.

Authors in [14] predict the number of daily confirmed cases of coronavirus after vaccination using 2 statistical models and a deep learning model—the autoregressive integrated moving average (ARIMA), the generalized autoregressive conditional heteroscedasticity (GARCH), and the stacked long short-term memory deep neural network (LSTM DNN). Dataset of WHO which is obtained from GitHub is used here. Every 3 models are applied to the dataset. The optimal hyperparameters for LSTM—that is the count of LSTM cells and blocks of cell—are obtained by a comprehensive search. When considered along the parameters of RMSE and MAE, the results of experiments based on these models show that LSTM DNN functions with the highest accuracy. Performances of ARIMA and GARCH depend on datasets.

Medical images could also be leveraged for data forecasting. Efficient image retrieval is crucial in such cases. In [15], a framework using adaptive state transition Kalman filtering technique is used to improve retrieval rate. It achieved a success rate of 96.2% of retrieval rate in image retrieval. In [16], classification of COVID-19 is done with the help of chest X-ray images of patients. Authors have used CNN with histogram-oriented gradients (HOG) as a methodology for feature extraction. The proposed CNN architecture consists of 5 layers. Max pooling layer has been used to reduce the space size with rectified linear unit ReLU as an activation function. The final flatten array gives categories to determine whether the result is COVID-19 positive or negative or pneumonia. For testing purposes, Cohen’s dataset with 400 positive COVID-19 X-ray images is used, and an accuracy of 93% was observed.

3 Data Visualization—COVID-19 Vaccinations

3.1 Data Visualization Tools and Techniques

Businesses and organizations produce huge amounts of data every day. This data can be a powerful tool if leveraged correctly. For a person who is unfamiliar with data analytics, large datasets tend to be quite abstruse. Visualization techniques are employed to make datasets meaningful—some commonly preferred ones are bar graphs, pie charts, scatter plots, line graphs, histograms, Gantt charts, map visualizations, candlestick diagrams, etc.

Data visualization tools make it possible to create such charts and make big data easy to interpret and extract useful information from. Tools ranging from basic to advanced such as MS Excel, Google Charts, Tableau, Zoho analytics, and Datawrapper are available for generating relevant data visualizations [17]. Another important tool available is Dash. Dash is a Python framework that is useful for creating a more personalized, interactive dashboard application. Flask, Plotly.js, and React.js form the basis of this framework. Dash applications are comprised of two components, such as the layout and the callbacks. As Dash application development is done using Python, Dash is an ideal framework for tools that integrate visualization with forecasting and analytics.

3.2 Use Case: COVID-19 Vaccination Dashboard

Application Architecture: We have built an interactive dashboard application that combines visualization as well as forecasting of COVID-19 vaccination statistics. The application is built using the Python Dash framework as shown in Fig. 1. With

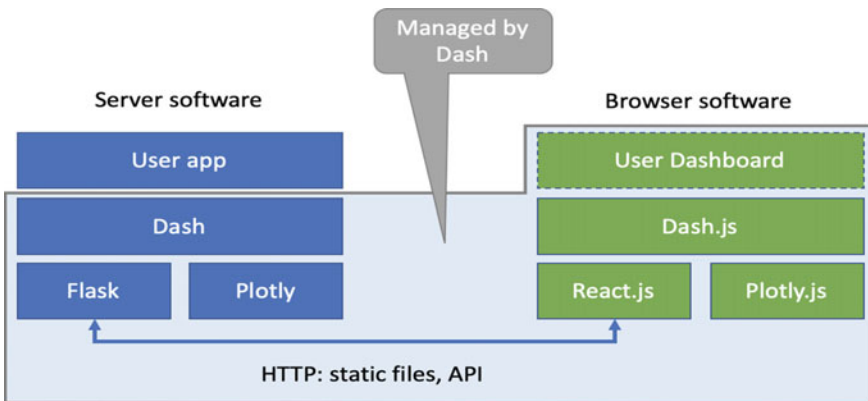


Fig. 1 Dash application architecture [9]

Plotly Dash, all of the user code for the dashboard is in Python. Dash is built on top of Flask (a micro-Web framework in Python) and serves the code over it using HTTP which is then deployed over a Web server like Nginx. Dash generates the necessary JavaScript code. The browser Web application is created and updated with the Web API generated by Dash [18]. Plotly is the library employed by the Dash framework to create interactive data visualizations.

Datasets: The application uses the vaccination datasets provided by our world in data. [19] This dataset uses official vaccination statistics sourced globally from governments and health ministries. The datasets are updated daily based on the latest information received from official sources, which enables our application to be updated daily with the latest statistics. Our dashboard employs two different datasets—one of global vaccination metrics and one of state-wise vaccinations in the United States of America. The latter relies on daily updates sourced from the United States Centers for Disease Control and Prevention.

The global dataset comprises of data with the following headers: location, iso_code, date, total_vaccinations, people_vaccinated, people_fully_vaccinated, total_boosters, daily_vaccinations_raw, daily_vaccinations, total_vaccinations_per_hundred, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, total_boosters_per_hundred, and daily_vaccinations_per_million. The US state dataset has the following data columns—date, location, total_vaccinations, total_distributed, people_vaccinated, people_fully_vaccinated_per_hundred, total_vaccinations_per_hundred, people_fully_vaccinated, people_vaccinated_per_hundred, distributed_per_hundred, daily_vaccinations_raw, daily_vaccinations, daily_vaccinations_per_million, and share_doses_used.

Population estimates for per-capita statistics in the datasets are based on the United Nations World Population Prospects. Income groups are based on the World Bank classification.

Data preprocessing: The data used for the visualizations are read from the source datasets and stored as DataFrames. A DataFrame is a data structure of pandas, a powerful library for Python programming that facilitates data analysis and manipulation. DataFrames are two-dimensional, size-mutable structures containing tabular data (possibly of different types). [20] They contain labeled rows and columns. From the two DataFrames (of global and US state data), two dictionaries are created. A dictionary (dict) is a Python data structure with a key-value pair. The DataFrames are filtered by country or state, respectively, and a dict of DataFrames is formed with each DataFrame in the dict having its corresponding country or state as its key. These dicts are sorted based on the people_fully_vaccinated metric for the latest date in each DataFrame in descending order—i.e., the country or state with the highest number of fully-vaccinated people is first. Further, sorting and preprocessing are done based on the data required for each type of visualization—based on date, daily doses administered, etc.

Visualizations: The first section of the dashboard displays the global vaccination metrics using different graphics. Figure 2 shows the top of the dashboard with the different tabs as well as the total number of vaccinated people worldwide. This lets the user get an idea of the state of vaccinations on a global scale at a glance.



Fig. 2 Dashboard vaccination statistics tab

The top 10 countries with the highest number of fully-vaccinated people are displayed next, with the latest updated count as well as the change in the count of doses administered since the last update (Fig. 3). This graphic enables the user to track the progress of the most highly-vaccinated countries on a daily basis and provides information about total vaccinations as well as the number of doses administered per day.

A line graph showing the daily dose count in these countries since December (based on the “date” and the “daily_vaccinations” columns in the dataset) is the next graphic (Fig. 4). This shows the trend of vaccinations in these top ten countries, and the user can observe how daily vaccination rates have evolved over time—whether they are increasing, whether they peaked and then dropped, or whether they have been constant. This visualization can be interacted with by selecting a single country to display (Fig. 5), which enables the user to observe the vaccination trend for a particular country more closely.

The second section of the dashboard is comprised of state-wise information for the United States of America (USA). As of November 2021, the USA has both the highest number of cases as well as the highest count of active cases globally [21].

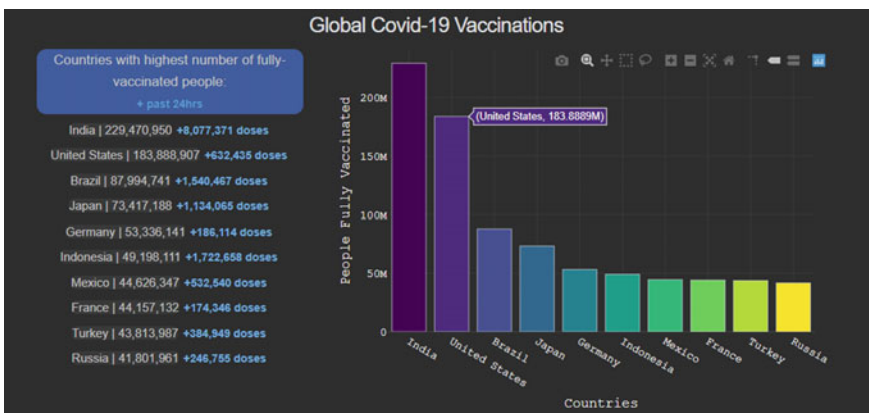


Fig. 3 Top 10 countries based on number of fully-vaccinated people

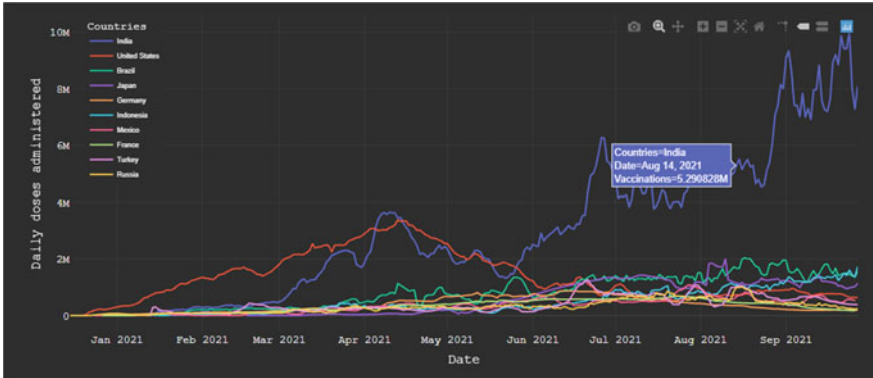


Fig. 4 Daily vaccinations in top 10 countries

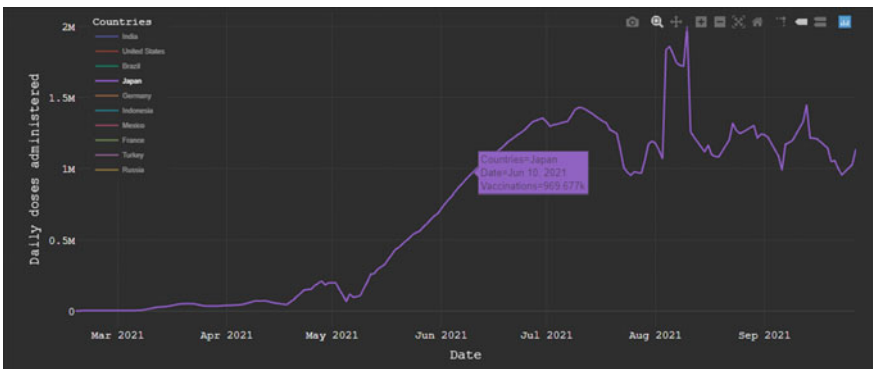


Fig. 5 Single country selected

Thus, visualizations of the vaccinations administered in the states of the USA are useful for tracking which states are implementing vaccination measures well and which states are lacking in their efforts. This section of the dashboard thus consists of an infographic of the 20 states with the highest count of fully-vaccinated people (Fig. 6). An interactive scatter plot of the daily doses administered in a state over time is also displayed, based on the state selected from the table (Fig. 7 and 8). This scatter plot allows users to view and understand the progress of vaccination drives in the selected state over time, and to glean whether vaccination efforts are still being undertaken rigorously or whether they have stagnated.

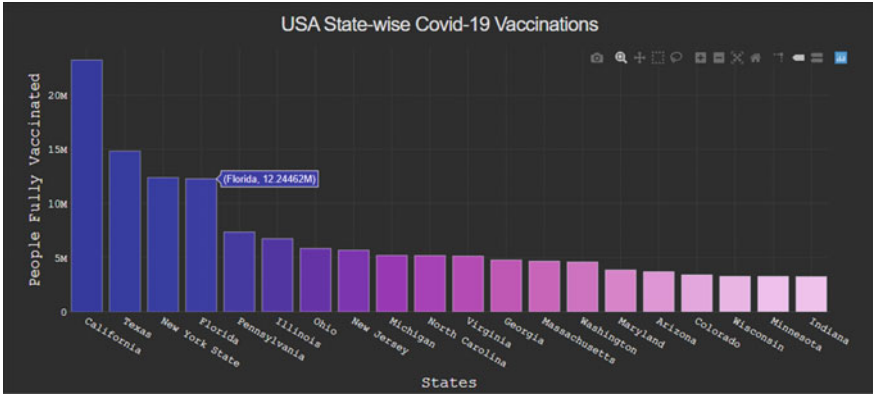


Fig. 6 Top 20 states based on number of fully-vaccinated people

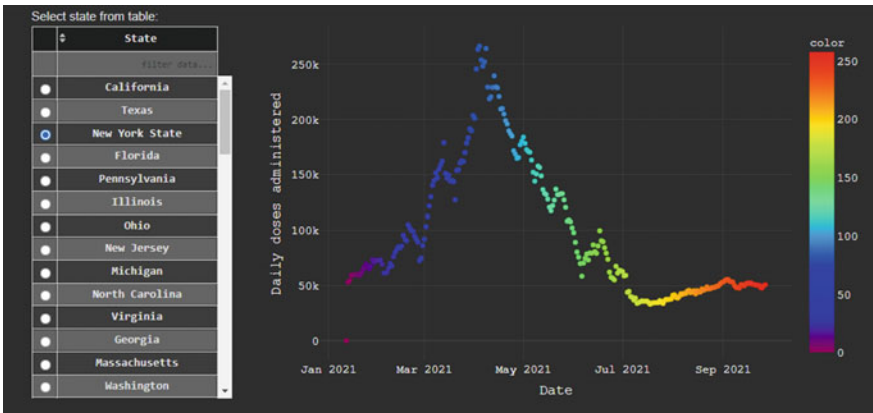


Fig. 7 The state of New York selected

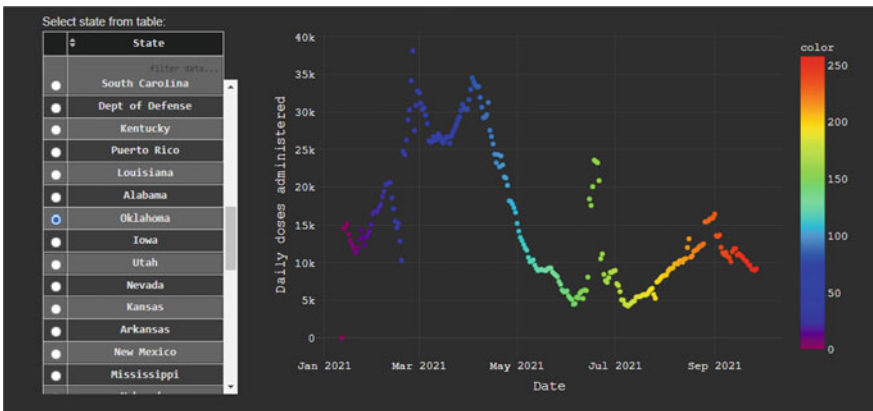


Fig. 8 The state of Oklahoma selected

3.3 Advantages and Future Scope

Data visualization is a powerful tool for data analysis and for building statistical models. Taking the use case of COVID vaccinations which we have detailed here, the spread of misinformation about the recently developed vaccines is rampant. People are hesitant to get vaccinated due to the unknown nature of the vaccines. In such a situation, clear and easily comprehensible visualizations of complex and opaque vaccination datasets can help give people an accurate idea of the situation of vaccination drives all over the world. Additionally, such a tool that is kept up-to-date with the latest information can be advantageous to healthcare providers for streamlining their logistics and operations. They can ensure that they are not short-staffed or falling short in supply of the vaccine by making an informed estimate with the help of easy-to-read, user-friendly visualizations.

The dashboard also integrates data forecasting of new COVID-19 cases using the vaccination data, which will be explained in the next section.

4 Data Forecasting—COVID Vaccinations Per Hundred Versus New Cases

4.1 Motivation

We attempted to model the relationship between the “people vaccinated per hundred” and “number of new cases.” We chose to analyze the relationship between these 2 factors as determining their relationship can prove to be extremely useful to forecast the trend in the number of new cases upon increasing the number of inoculations.

4.2 Algorithm

Various techniques such as regression and correlation are used to analyze the relationship between 2 and more variables. We chose to use the polynomial regression technique as it modeled the best relationship among all the techniques surveyed. According to [22], polynomial regression is a regression technique in which the relationship between the independent variable x and the dependent variable y is modeled as an n th degree polynomial in x . The equation can be modeled as follows wherein $a_0, a_1, a_2, \dots, a_n$ are the coefficients of the regression equation.

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \dots + a_nx^n \quad (1)$$

In our case, the independent variable x was “people_vaccinated_per_hundred,” and the dependent variable y was “number_of_new_cases.”

4.3 Implementation Details

Dataset: After analyzing numerous datasets, we chose to use the dataset made available for public use provided by [23]. Ritchie et al. [23] is one of the most comprehensive COVID datasets and contains extensive data which can be used for various kinds of research. The columns “people_vaccinated_per_hundred” and “new_cases” were used for the algorithm.

Data Preprocessing: The dataset was filtered, and data for India were collated. The data used were after the month of May as the number of vaccinations before May were not significant enough to model the relationship. The dataset contained incomplete data at numerous instances for both the columns mentioned above. The incomplete data were estimated using pandas interpolate method [24]. The interpolation technique used was “polynomial” of order 5.

Polynomial Regression: We used scikit-learn [25] for performing polynomial regression in Python. Scikit-learn has a linear regression module that can be used to train a polynomial regression function. PolynomialFeatures are a tool in scikit-learn which computes the desired degrees or powers of feature variables. It transformed the given feature variable into a matrix containing the degrees of the variables.

The model was trained using a random split of 80% training to 20% test data. Using the fit function, the polynomial features were fit into the polynomial regression model, and the equation of the regression was obtained. The degree of the polynomial regression was chosen to be 8 after extensive experimentations. The following is the equation of the regression where the variables are as defined in Sect. 4.2:

$$y = -487053.151x - 4912.572x^2 + 5008.596x^3 - 414.69x^4 + 16.45x^5 - 0.355x^6 + 0.00403x^7 - 0.0000188x^8 \quad (2)$$

Future values for the number of new cases can be predicted by entering the value of the people vaccinated per hundred in the above equation. A scatterplot of the regression function was plotted using matplotlib [26] as shown in Fig. 9.

4.4 Results

Using the predict method, the values for the test input data were predicted. There are various methods to evaluate regression models such as relative absolute error, range of prediction, mean absolute error, relative squared error, and coefficient of determination (R^2). We chose to use the coefficient of determination denoted popularly as R^2 , which is the amount of variation in the dependent variable [27] which can be predicted from the independent variable. The R^2 score provides a measure of how well the variances of two variables are dependent on each other. E.g., if the R^2 score of the regression is 0.50, then roughly 50 percent of the variation can be explained by the inputs. R^2 was calculated between the predicted and the actual values for the

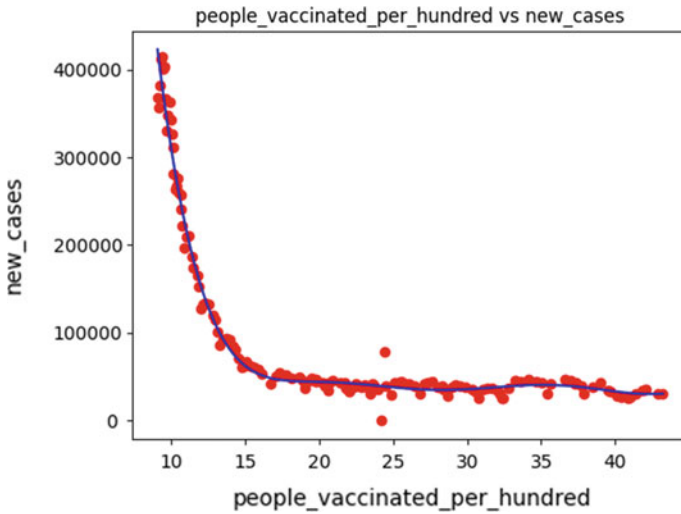


Fig. 9 People_vaccinated_per_hundred versus new_cases

input data. It usually ranges from 0 to 1 with 1 being the highest possible score. The R^2 score is calculated as follows:

$$R^2 = 1 - (SS_{res}/SS_{tot})$$

where SS_{res} is the sum of squares of residuals, also called the residual sum of squares:

$$SS_{res} = \sum (y_i - f_i)^2$$

SS_{tot} is the total sum of squares (proportional to the variance of the data):

$$SS_{tot} = \sum (y_i - \bar{y})^2$$

\bar{y} is mean of the observed data:

$$\bar{y} = 1/n \left(\sum y_i \right)$$

The R^2 score of our algorithm was 0.9878. This shows how closely the variance of the independent variable—people vaccinated per 100 explains the variance of the dependent variable—no of cases.

Training	Test	R^2 score
80% of input data	20% of input data	0.9878

4.5 *Vantages and Future Scope*

The above-mentioned forecasting model using polynomial regression can certainly prove to be beneficial in a lot of ways. The model can be used by the authorities to forecast the cases at a future point in time. This can assist them to manage the existing resources more efficiently and allow for better planning and allocation of resources. The forecasting model will also help the vaccine manufacturers to predict the efficacy of their vaccines.

The model can be further enhanced by expanding the dataset size and entering actual values in place of the missing values in the dataset. Other regression models such as Lasso and Ridge could also be employed to forecast the relation between the number of vaccinations and the number of cases. Various alternative outlier detection methods could also be used to remove the outliers present during the second wave.

5 **Conclusions**

Our data forecasting model predicts the future trend of new cases by taking the past data into consideration and is integrated with a visualization dashboard. This visualization of the data along with forecasting is highly useful to medical professionals. They can easily see the current and past trends as displayed by the dashboard by filtering through various parameters. Further, data visualization combined with data forecasting ensures that the efforts being taken behind vaccination drives are being channeled in the right direction. An optimistic trend resulting in fewer deaths after vaccinations makes a strong point for the government to encourage their citizens to get vaccinated. Vice-a-versa, if the trend is not as positive as expected, then the efficacy of the vaccine could be questioned, thus alerting a particular country to look for an alternative measure. Hence, relevant measures, steps, and decisions could be taken by the medical professionals as well as the government officials by utilizing this data visualization and forecasting.

References

1. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read, <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=1503adeb60ba>. Last accessed 15 Sep 2021
2. Pandian, A.P., Smys, S.: Effective fragmentation minimization by cloud enabled back up storage. *J. Ubiquitous Comput. Commun. Technol. (UCCT)* **2**(01), 1–9 (2020)
3. Data Visualization using Python for Machine Learning and Data science, <https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7>, last accessed 2021/09/16
4. ML time-series forecasting the right way, <https://towardsdatascience.com/ml-time-series-for-ecasting-the-right-way-cbf3678845ff>. Last accessed 19 Sep 2021
5. Ramazi, P., Haratian, A., Meghdadi, M., et al.: Accurate long-range forecasting of COVID-19 mortality in the USA. *Sci Rep* **11**, 13822 (2021). <https://doi.org/10.1038/s41598-021-91365-2>
6. Rustam, F., et al.: COVID-19 future forecasting using supervised machine learning models. *IEEE Access* **8**, 101489–101499 (2020). <https://doi.org/10.1109/ACCESS.2020.2997311>
7. Hazarika, B.B., Gupta, D.: Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks. *Appl. Soft Comput.* **96**, 106626–106626 (2020)
8. Darapaneni, N., Reddy, D., Paduri, A. R., Acharya, P., Nithin, H. S.: “Forecasting of COVID-19 in India using ARIMA model.” In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp. 0894–0899 (2020). <https://doi.org/10.1109/UEMCON51285.2020.9298045>
9. Abdulkareem, N. M., Abdulazeez, A. M., Zeebaree, D. Q., Hasan, D. A.: COVID-19 world vaccination progress using machine learning classification algorithms. *Qubahan Acad. J.*, **1**(2), 100–105 (2021). <https://doi.org/10.48161/qaj.v1n2a53>
10. Aljameel, S. S., Khan, I.U., Aslam, N., Aljabri, M., Alsulmi, E. S.: “Machine learning-based model to predict the disease severity and outcome in COVID-19 patients.” *Sci. Program.*, **2021**, 10 (2021) Article ID 5587188. <https://doi.org/10.1155/2021/5587188>
11. Hu, C., Liu, Z., Jiang, Y., Shi, O., Zhang, X., Xu, K., Suo, C., Wang, Q., Song, Y., Yu, K., Mao, X., Wu, X., Wu, M., Shi, T., Jiang, W., Mu, L., Tully, D.C., Xu, L., Jin, L., Li, S., Tao, X., Zhang, T., Chen, X.: Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *Int. J. Epidemiol.* **49**(6), 1918–1929 (2020). <https://doi.org/10.1093/ije/dyaa171>
12. Jayakumar, K., et al.: Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. *Frontiers Public Health* **9**, 729795 (2021). <https://doi.org/10.3389/fpubh.2021.729795>
13. Muhammad, L.J., Algehyne, E.A., Usman, S.S., et al.: Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Comput. Sci.* **2**, 11 (2021). <https://doi.org/10.1007/s42979-020-00394-7>
14. Kim, M.: Prediction of COVID-19 confirmed cases after vaccination: based on statistical and deep learning models. *Sci. Med. J.* **3**(2), 153–165 (2021)
15. Dhaya, R.: Analysis of adaptive image retrieval by transition kalman filter approach based on intensity parameter. *J. Innovative Image Process. (JIIP)* **3**(01), 7–20 (2021)
16. Chen, J.-Z.: Design of accurate classification of COVID-19 disease in X-ray images using deep learning approach. *J. ISMAC* **3**(02), 132–148 (2021)
17. Stobierski, T.: “Top data visualization tools for business professionals.” 12 January 2021. [Online]. Accessed 6 Sep 2021
18. Plotly Dash or React.js + Plotly.js? A side-by-side comparison, <https://towardsdatascience.com/plotly-dash-or-react-js-plotly-js-b491b3615512>. Last accessed 10 Sep 2021
19. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., et al.: A global database of COVID-19 vaccinations. *Nat. Hum. Behav.* (2021)

20. pandas.DataFrame, <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html>. Last accessed 11 Sep 2021
21. WHO Coronavirus (COVID-19) Dashboard, <https://covid19.who.int/>. Last accessed 01 Nov 2021
22. Polynomial Regression, https://en.wikipedia.org/wiki/Polynomial_regression. Last accessed 15 Sep 2021
23. Ritchie, H., Mathieu, E., Rod s-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., Roser, M.: “Coronavirus pandemic (COVID-19)”. In: Published online at OurWorldInData.org. Retrieved from: ‘<https://ourworldindata.org/coronavirus>’ [Online Resource]
24. Pandas, <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html>. Last accessed 16 Sep 2021
25. Scikit-learn, <https://scikit-learn.org/stable/>. Last accessed 16 Sep 2021
26. Matplotlib, <https://matplotlib.org/>. Last accessed 17 Sep 2021
27. Coefficient of determination, https://en.wikipedia.org/wiki/Coefficient_of_determination. Last accessed 19 Sep 2021