U.S. Covid-19 Vaccinations Analysis at County Level

Department of Computer Science, Stony Brook University

Abstract

The major goal of our project is to investigate the factors influencing vaccination rates across counties in the United States of America and also to forecast future vaccination rates based on our data analysis. Here we are going to handle some hurdles like economic, cultural, and scientific factors. One of the main priorities is to analyze whether the age of a particular person affects the vaccination rate. We will be building a model that predicts the date when the vaccination rate increases over 80%. We will be providing some insights and some suggestions for a county to increase the vaccination program speed to reduce daily infectious numbers based on different aspects of Covid-19 related data.

Keywords

COVID-19, Data Analysis, Time series analysis, ARIMA, LSTM

1 Introduction

Since 2020, the novel Coronavirus or Covid-19, an infectious disease caused by the SARS-CoV-2 virus, has spread to every corner of the globe, resulting in a global pandemic and the deaths of millions. Nonetheless, we were able to quickly develop the first Covid-19 vaccines for general use, and many countries began administering them by the end of 2020. Even with government approval and mass production, vaccination rates have not been consistent or satisfactory, even in higher-income countries such as the United States. Considering these parameters, we aim to find out what factors influenced the vaccination program coverage. We are also using state policies to find out the reason for discrepancy. There has been some ongoing analysis being done on this to predict the period when it reaches the whole vaccination coverage state.

2 Dataset

We are dealing with multiple datasets, in which the primary dataset is taken from CDC [1] and remaining datasets were taken mainly from the U.S Department of Agriculture. Each of these datasets are discussed below.

2.1 Primary Dataset:

The Centers for Disease Control and Prevention (CDC) is a federal health agency in the United States that strives to preserve public health and safety by controlling and preventing disease, injury, and disabilities in the United States and around the world. CDC is using both new and existing information technology (IT) systems to rapidly collect reliable data about how many doses of COVID-19 vaccines have been delivered (distribution) and how many people have been vaccinated with those doses(administration). For this project we have used the dataset - COVID-19 Vaccinations in the United States, County as the primary dataset. Since the beginning of the vaccination process in the United States, different vaccination rates and relevant information have been collected for each county. There are 1.16 million rows and 32 columns in this dataset. The primary dataset gives details on Date, FIPS code unique for each county, Recipient County, Recipient State, social vulnerability index (SVI) for each county, total number of people who got single and double dose of vaccination in each county, total percentage of people who got single and double dose of vaccination in each county.

The dataset considered has information from 2020-12-13 to 2021-10-20.

2.2 Secondary Datasets:

2.2.1 Education:

From the Understanding America Study survey [2], it reveals that when it comes to attitudes and beliefs about COVID-19 vaccine - from willingness to get the vaccine to knowing someone who has been vaccinated to the perceived risks of side effects — there is a substantial gap between more- and less-educated U.S. residents. We have considered the education data from the United States Department of Agriculture [3]. The United States Department of Agriculture maintains different streams of data in the United states. This dataset gives details of the education rate for each county.

2.2.2 Economy:

It is considered that economic status is the biggest factor in COVID-19 vaccination rate. Economic status of an individual may be a common reason for not taking Covid-19 vaccine. Economic status may depend on income and unemployment status also. If this is the reason, then we may suggest the federal government provide funds for the people below the poverty line to take Covid-19 vaccines. We have considered the data provided by the Local Area Unemployment Statistics (LAUS) program [4]. LAUS produces monthly and annual employment, unemployment, and labor force data for Census regions and divisions, States, counties, metropolitan areas, and many cities, by place of residence. From the economy dataset provided by LAUS, we have considered columns specifying for each county. There are other factors which we have considered to analyze the vaccination rate, such as Satte policies, mandate policies, demographics, and race.

3 Data Preprocessing:

Using all the data we obtained from the different sites and APIs we generated the entire dataset. Now, while merging these different sets of data, we encountered several issues and resolved them. We tried to obtain the dataset using web scraping. But as the webpage is static, it's showing the same web address for all the states. So, we generated the dataset manually. As mentioned in the proposal, using all the data from different websites such as CDC and USDA, we structured a dataset manually by choosing the features that are only required for our current data analysis part. For example, we took the fips, county and literature data for a county after 2015 to see how the education rate affects the vaccination progress. In the same way we took economic data, county population data to determine some of the affecting features. We generated these features manually by creating separate csv files for each state. We then merged all the csv files into a master csv file to perform the data analysis part.

4 Data Analysis

4.1 Primary Analysis:

Based on primary data analysis we got to know that the vaccination progress is null in 9 county's of the total 3224 counties. Vaccination progress will be predicted based on 319 dates where vaccination took place. We found out that some countries haven't taken COVID-19 vaccine which is a peculiar observation. We noticed that the intake of vaccines has gradually increased till the month of September and there upon we noticed a sudden decrease in the month of October. We noticed that the count of people taking the first dose was the highest in September of all the months that we considered in our dataset.We concluded that there are many counties whose vaccination rate is zero. There may be two major reasons for this. One is inadequate data and for the second we can assume that the vaccination program hasn't started in that county.

4.2 Vaccination Based on Age :

In the Primary dataset, we have columns which specify, number of people who got single and double vaccinations above age 12, above age 18, above age 65. The column specifying the number of people vaccinated above 12, also has a count of the number of people vaccinated above 18 and 65. As these columns have cumulative data, we divided these columns into 3 columns I.e... number of people got vaccination aged between 12 – 18, number of people got vaccination aged between 18 - 65, number of people got vaccination aged above 65. Same was done for both double doses columns and single dose columns. From figure (1) and (4) we can say that for people between 12–18 most of the vaccinations were done around March and the vaccination rate decreased gradually till June and then started increasing. The main reason for the decrease of vaccination rates of people aged between 12 – 18 could be that, most of the students went for vacations and didn't bother to get vaccinated. We can see that the vaccination rate for this age group increased from August as this is when academics started. Same could be the reason for the decrease in vaccination rate for people aged between 18 – 65, as most of them of this age are employed. From figure (3) and (6), which portrays the vaccination rates of people aged 65 and above, we can observe that most of the people got their first vaccination in the month of January and in February their second vaccination in the month of March. For people aged above 65, vaccination drastically decreased.

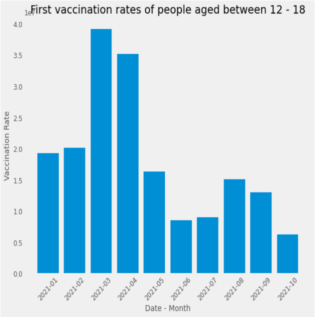

Figure 1: First dose vaccination rate of people between 12 - 18

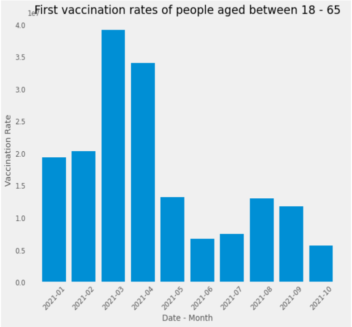

Figure 2 : First dose vaccination rate of people between 18-65

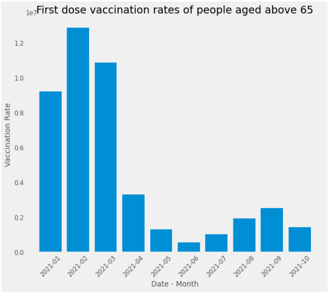

Figure 3 : First dose vaccination rate of people above 65

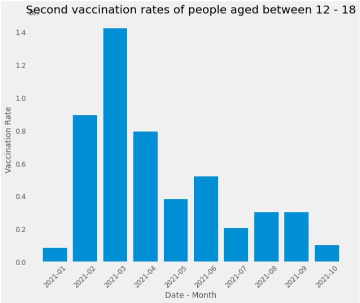

Figure 4: Second dose vaccination rate of people between 18 - 65

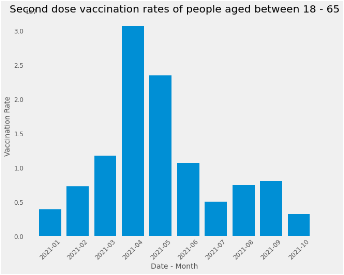

Figure 5: Second dose vaccination rate of people between 12 - 18

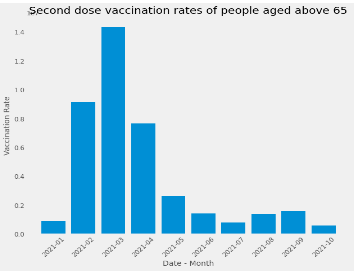

Figure 6: First dose vaccination rate of people above 65

4.3 County level vaccination distribution:

We noticed that the mean of the vaccination rate per all the counties is left skewed. This means that the distribution is positively skewed. There are outliers in the case of percent of total vaccinated people. However, there is another finding where the median is towards the left which says the data is right-skewed. So, here the data is positively skewed. Here Mean Series_Complete_Pop_Pct is 21.429570149230816 and Median Series_Complete_Pop_Pct is 20.4.

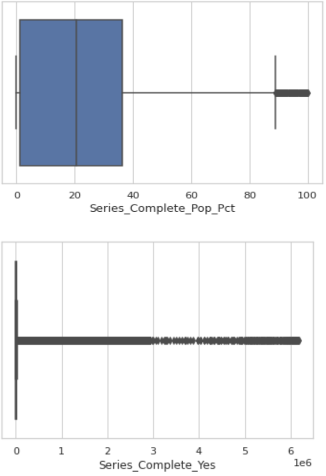

Figure 7: (a) Mean of vaccination percent for population in a county.(b) Mean of total vaccinations done on a particular day.

Even Though, California has most of the vaccinated people. The distribution among the California counties is less. Many other states contributed to the list of top 10 fully vaccinated. We concluded that there are many counties whose vaccination rate is zero. There may be two major reasons for this. One is inadequate data and for the second we can assume that the vaccination program hasn't started in that county.Honolulu county reported the avg least vaccination rate of all of the counties excluding the ones where vaccination progress rate is zero. Chattahoochee County reported the average highest vaccination rate of all of the counties with a rate of 72.9%.

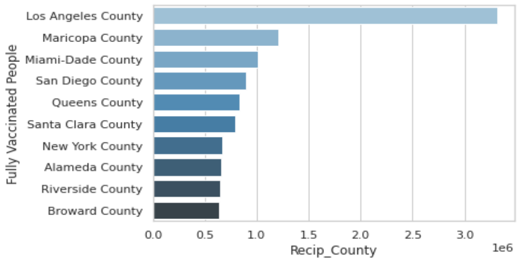

Figure 8: Top 10 fully vaccinated counties

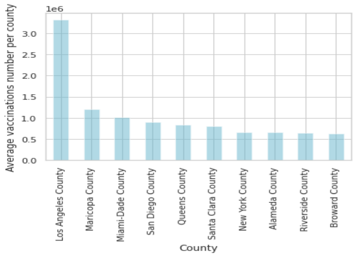

Figure 9: Average vaccinations number per county

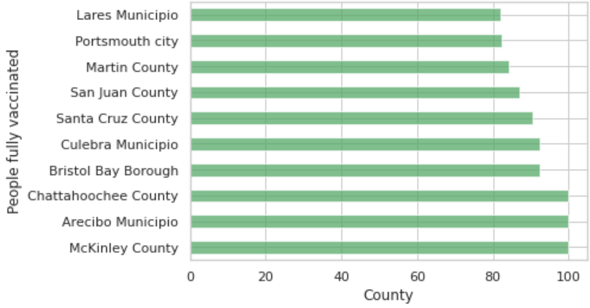

Figure 10: Fully vaccinated people with highest vaccination rate

4.4 Correlation between features of the dataset

Surprisingly, there's no relationship between the education degree of a person and vaccination intake, although this is false in the case of North Carolina State counties. As expected, there is a high correlation between features like education rate, employment rate, poverty, and the vaccination rate. Figure explains the correlation between their features. The below correlation matrix shows the high correlation between population and the covid vaccination rate. This fact is expected.

4.5 Education rate vs Cases and Deaths

People with higher education rates have already taken vaccination considering their awareness about the situation. In Montgomery County we noticed that the education rate (that is who has degree) affected the vaccination rate in a positive way. The people in that county took vaccination seriously and were in the top 10 highest vaccinated counties. This implies that only Montgomery County's education rate and vaccination rate were related when compared with other counties.McKinley County achieved a 100% vaccination rate for a particular date. Vaccination rates in the counties Arecibo Municipio and Chattahoochee County went up to 100% and then decreased to 99.9%. One of the reasons for this may be due to the increase in the death rate.Another finding about Montgomery County is that it's recorded in the top 10 position in regard to the daily number of cases in that county. This is surprising because we thought that as the education rate is high, it brings more awareness among the county population and it in turn reduces the number of daily cases. So, our prior analysis about the education rate and daily cases rate was proved to be wrong. Another finding is that Montgomery County was in the list of top 10 deaths county wise. This is obvious from the above point, as there is relation between number of cases and number of deaths per county But, again the education rate and death rate were not coordinated

4.6 Demographics vs vaccination progress

Concerning with the gender of the county population, we noticed that the vaccination rate among females is more than males.

4.7 Unemployment vs vaccination rate We found out that the unemployment rate and education level are not related. So, we cannot combine these factors to determine the vaccination rate accurately.

4.8 Metro and non-metro vs vaccination rate

People coming under metro population tend to have more vaccination rate when compared to non-metro population.Metro counties have lower vaccination rates for the total population than non-metro counties.

4.9 Vaccination policy vs vaccination rate

Considering the vaccination policy, out of 3024 counties only 607 counties are providing free vaccines to the population of that county. Santa Cruz county has the highest vaccination rate in the pool of counties which are providing free vaccination. For those counties who are not provided free vaccination, the vaccination progress rate is 0. So, we can suggest the government bring a free vaccination policy in those counties. And obviously the counites with 0 level mandatory vaccination rules are having 0 percent vaccination progress. Based on the mandatory level of vaccinations, the progress rate is getting improved in some counties which is a good rule that's being implemented by the government in some counties. We found that McKinley County has 100 percent vaccination rate even though there's no mandatory policy in this county. We tried to find relation with economic data, education data, ethnicity data, but there's no relation of those features which can be glued to this county.

4.10 Race vs vaccination rate

Kalawao, Hawaii, Kauai, Mauli, countys has the largest number of non-Hispanic Native Hawaiian/Pacific Islander, where the vaccination progress is 0 percent. Though Honolulu County has the largest number of non-Hispanic population it's vaccination is not zero, but it has the least vaccination progress rate.

5 Models Training And Predicting

The vaccination dataset for the period from 5 Jan 2021 to 20 Oct 2021 in the United States for 3224 counties is used in this study. A time series, such as the vaccination rate data series in the figure below, exhibits properties such as time-varying mean and variance, which are typical of a non-stationary time series. Based on the plot, we may conclude that the vaccination rate data is non - stationary. Typical patterns cannot be deduced immediately from the signal, and prediction of this type of data requires special care. The main goal of this section is to predict when vaccination rate reaches 80% herd immunity.

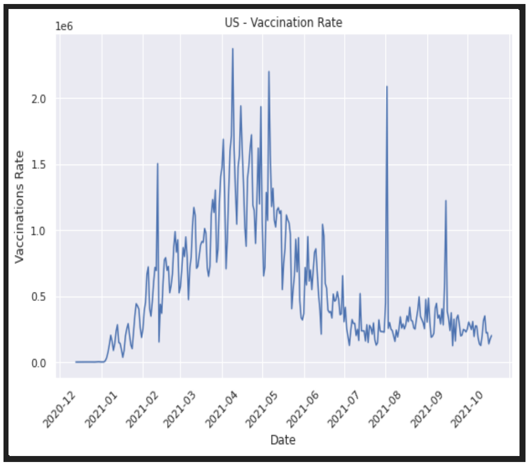

Figure 11 : US Vaccination Rate

5.1 Time Series

Time series analysis is a forecasting and analysis method that is frequently used in business, economics, finance, computing, and science. It is the process of generating scientific forecasts based on time stamped historical information. It entails developing models based on historical data and using them to make observations and drive future strategic decisions. An essential contrast in forecasting is that the future outcome is completely unknown at the time of work and can only be anticipated by meticulous analysis and evidence-based priors. It's not always an exact prediction, and the likelihood of forecasts can vary wildly—especially when dealing with the commonly fluctuating variables in time series data as well as factors outside our control.

5.2 Data for Modelling Model

For modelling we are using only two columns i.e.. 'Date' and 'series_complete_yes'. In feature 'series_complete_yes' cumulative data of vaccination rate until considered date, it is inappropriate to take Cumulative data. The general approach is to model the increment process (the initial difference in the cumulative sum process) and then compute the cumulative sum. Cumulative sums (by definition) have unit roots, and such processes do not lend themselves easily to traditional statistical modeling because they (the processes) are non - stationary. Estimation procedures for statistical models often require the processes to be stationary; otherwise, the estimates of model parameters do not have their nice qualities (consistency, asymptotic normality), which makes obtaining excellent quality forecasts difficult. As cumulative data can't be used for modelling the time series we have created a new column named 'Daily_new_vaccinations' which has a daily count of people vaccinated.

5.3 Baseline Model

Establishing a baseline is essential for time series forecasting modelling. A performance baseline gives us a notion of how well all other models will perform in our scenario. In general, the technique used to construct a prediction and compute baseline performance must be simple to apply and devoid of problem-specific details. For this problem we have used the Moving average smoothing model as the baseline model.

5.4 Moving Average Smoothing

Smoothing is a technique used on time series to remove fine-grained variance between time steps. Moving averages are a simple and common type of smoothing used in time series research and forecasting. As a baseline model, we employed Moving average smoothing. Moving average smoothing is based on the rolling window concept. The underlying concept behind rolling windows is to set a window of a specific size within which certain calculations are performed. The rolling() function on the Series Pandas object will automatically combine observations into a window. You can specify the window size, and by default, a trailing window is generated. Once the window is created, we can take the mean value. Using the rolling() function from pandas we got the rolling average values which we have used to make predictions. Figure ___ shows the moving average prediction. From walk forward evaluations, we got MSE as 83476567447.5 and RMSE as 298923.11684503817, now we want the models which will make better predictions then the baseline model.

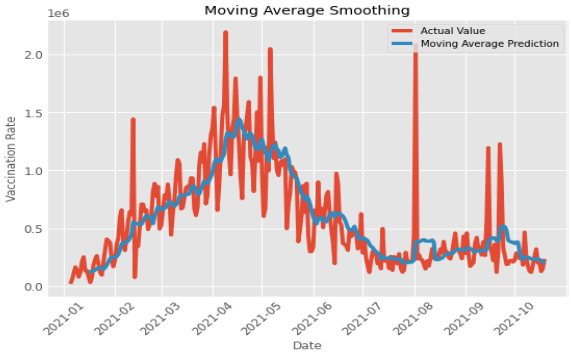

Figure 12: Moving Average Smoothing

5.5 Autoregressive Integrated Moving Average (ARIMA)

Box - Jenkins Analysis refers to a systematic method of identifying, fitting, checking, and using integrated autoregressive, moving average (ARIMA) time series models, These estimation methods are autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), and the autoregressive integrated moving average (ARIMA) processes. The specified time series method is known as the Box–Jenkins method. These time series assume that the series is stationary.

5.6 Checking if time series stationary

A TS is said to be stationary if its statistical properties such as mean, variance remain constant over time. To verify if the time series is stationary , we have used Augmented Dicker Fuller (ADF). The ADF test expands the Dickey-Fuller test equation to include a high order regressive process in the model. To do the ADF test we have used the statsmodel package that provides a reliable implementation of the ADF test via the adfuller() function in statsmodels.tsa.stattools.It returns the following outputs: The p-value, The value of the test statistic, Number of lags considered for the test, The critical value cutoffs when the test statistic is lower than the critical value shown, you reject the null hypothesis and infer that the time series is stationary. After verification if the time series is not stationary it should be made stationary. On running ADF test we got output as ADF Statistic: -1.406605 p-value: 0.579106

Critical Values: 1%: -3.454 ,5%: -2.872,10%: -2.572.The p-value obtained is greater than the significance level of 0.05 and the ADF statistic is higher than any of the critical values. Clearly, there is no reason to reject the null hypothesis. So, the time series is in fact non-stationary. Now our next task was to convert the time series to non-stationery. Two major reasons behind non-stationary data are trends and seasonality. We have terminated the trend and seasonality from the current time series. To convert to non stationary data, we have used decomposition and differencing techniques and converted data to a modified version. On running the ADF test on this modified version we got following results.

ADF Statistic = -3.395297,

p-value=0.011120

CriticalValue(1%) = -3.455754,

CriticalValue(5%)=-2.872721,

CriticalValue(10%) = -2.572728.

This looks like a much better series. Also, the test statistic is smaller than the 5% critical values so we can say with 95% confidence that this is a stationary series. Using ARIMA model from stats package we got below predictions

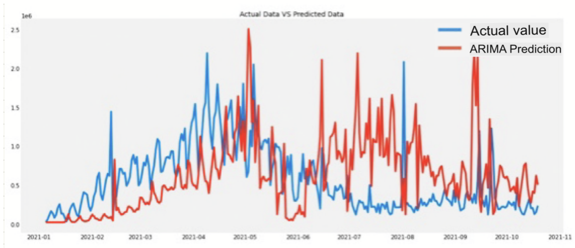

Figure 13: Prediction with help of ARIMA

For ARIMA we got RMSE as 302313.38237 which is not good when compared to RMSE of the baseline model.

5.7 Long short-term memory

Long short-term memory networks (LSTMs) are very powerful when used in time series prediction problems. LSTMs are explicitly designed to reduce the vanishing and exploding gradient problem during backpropagation in recurrent neural networks. We have implemented LSTM with memory between batches. The LSTM network has memory that allows it to recall extended sequences.model. As part of normalizing data for model training, we have used “MinMaxScaler”, package of SKLearn, to normalize vaccination count between 0 to 1.By making the LSTM layer "stateful," we can have more control over when the internal state of the LSTM network is cleared in Keras. This means that it can build state across the whole training phase and even preserve it if necessary to make predictions. When fitting the network, the training data must not be jumbled. It also requires periodically resetting of the network state after each exposure to training data (epoch) using model.reset_states(). This means that we must establish our own outer loop of epochs and call model.fit() and model.reset states() within each epoch.

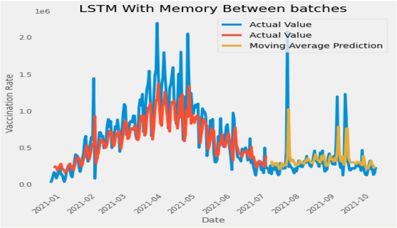

Figure 14: LSTM With Memory Between Batches

In the evaluation phase, we have calculated RMSE for both training and testing data. We got the following results.
Train Score: 261314.03 RMSE
Test Score: 257818.48 RMSE
Based on above scores, we can say that LSTM modeled the data more efficiently when compared to baseline model. As LSTM modeled data efficiently, we used this model to make future predictions to predict when total vaccination rate will reach 80%. Using the predict function on LSTM, we have observed that on 20 Feb 2022 65% of vaccination will be reached. On 20 July 2022 72% of vaccination was reached. On 20 Aug 2022, the US reached 81.476 vaccination rate. Therefore using LSTM model we can predict that US reaches herd immunity around August 2022

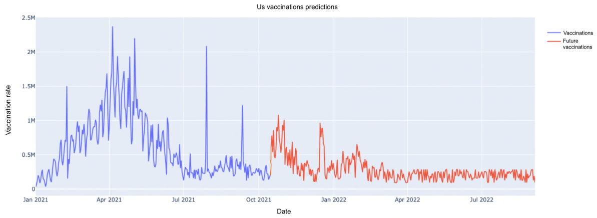

Figure 15: US Vaccination Prediction

6 Conclusion

In the report we have successfully completed the data analysis and data modeling part at county level for counties in the United States.We provided some insights and some suggestions in the data analysis part to increase the vaccination rate at county level.

7 References

1. COVID-19 Vaccinations in the United States,County
2. UNDERSTANDINGAMERICA STUDY
3. USDA - Education
4. https://www.bls.gov/LAU/
5. Household Pulse Survey COVID-19 Vaccination Tracker
6. Path to Normality - COVID-19 Vaccine Projections

7. Vaccination in America Might Have Only One Tragic Path Forward
8. How to Develop LSTM Models for Time Series Forecasting
9. keras-attention-mechanism