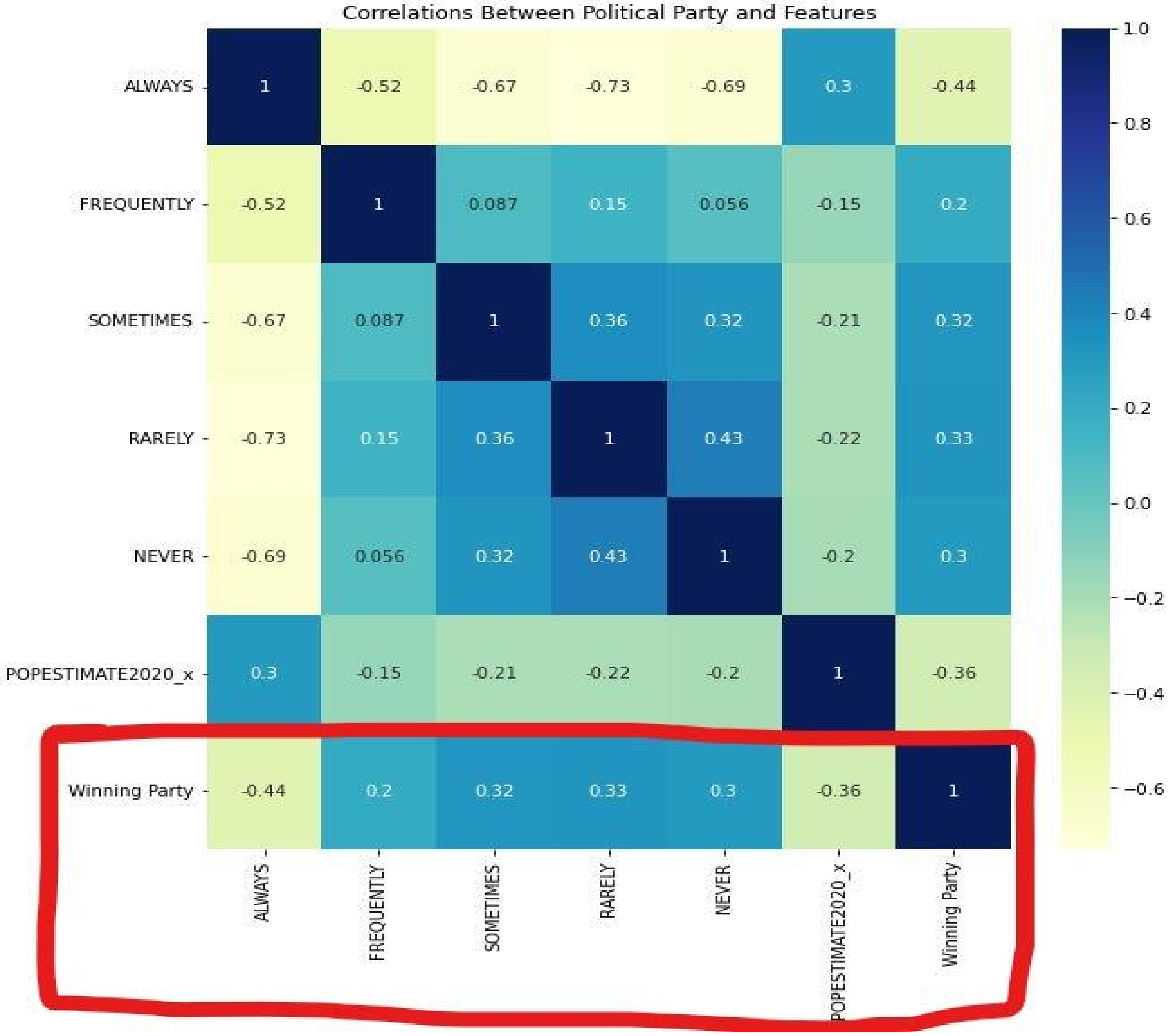
# Open-Ended Modeling Report

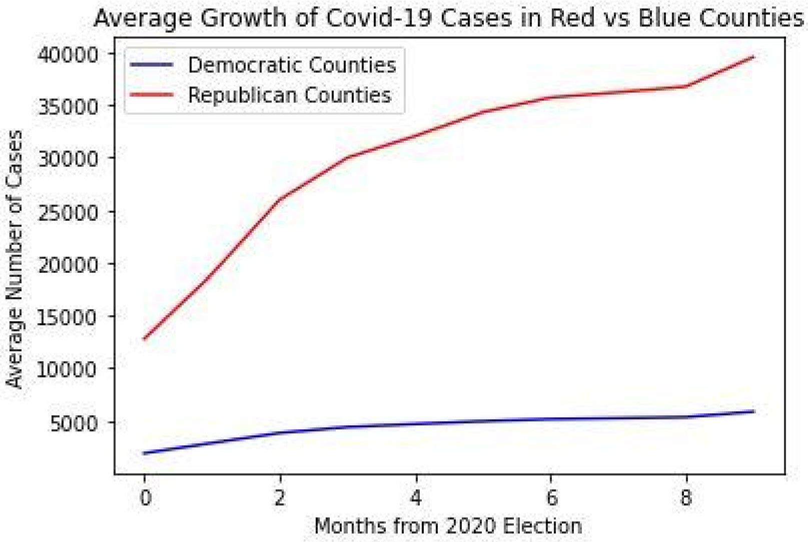
## Plot 1: Heatmap

Motivation: From the start, our team was interested in understanding the role of politics in the Covid-19 situation. Thus, our EDA considers several noteworthy features of the Covid-19 situation across counties in America and their majority political parties (Democrat or Republican). We determined the political party of a county based on their 2020 presidential election, which had the highest voter turnout of all time. This data was provided by MIT Election Lab (Source 1).



We first wanted to verify if there was any merit to our hunch of there being different Covid-19 activity in Republican versus Democratic states. Hence, we created a Heatmap. Focusing on the boxed portion, the correlation between winning parties (Democratic or Republican) and a couple of our features is described. Specifically, we considered mask usage (ALWAYS, FREQUENTLY, SOMETIMES, RARELY, and NEVER) and the 2020 population size. The lowest correlation across the row is -0.44, while the highest is 0.33. This lowest correlation indicates a moderate correlation between 'ALWAYS' mask usage and Democrats since we represent Democrats with a 0. Meanwhile, the highest correlation indicates a near moderate correlation between 'RARELY' mask usage and Republicans since we represent Republicans with a 1.

## Plot 2: Time Series Plot



Next, our group wanted to view the growth of Covid-19 over time, specifically from the onset of the election. We chose 11/20/20 as the start date. Through stratifying the covid-19 county counts by political party, we could compare the average growth in cases. This plot shows a significant difference in the growth rates of covid. Republican counties show a large, positive, and concave trend in case growth, with a sharper upward trend in the 8th month. Meanwhile, Democratic counties show a mild, positive, linear trend in average cases growth. The Democrat plot looks nearly horizontal. Furthermore, the Republican plot starts about 10,000 average cases higher than the Democrat plot. These notable differences in case growth and starting point are worth researching further if we want to understand the relationship between the political majority of counties and their Covid-19 situation.

## EDA Hypothesis:

Based on our visualizations above, our team would like to test the strength between the relationship of certain factors like mask usage and Covid-19 growth rate and political majority in counties. The growth rate of Covid-19 in a county is the difference in Covid-19 counts from November 2020 to September 2021 divided by the county population. Hence, if a county has poor mask usage and Covid-19 growth from the onset of the election, there is a higher likelihood that the county is majority red. Meanwhile, if a county has relatively higher mask usage and shows lower Covid-19 growth from the onset of the election, there is a higher likelihood that the county is majority blue. Our hypothesis extends to individual counties and we plan to test our hypothesis using predictive models (logistic regression to predict red or blue from mask usage and Covid-19 growth rate). Furthermore, the exact definitions of “higher mask usage”, “lower Covid-19 growth”, and “likelihood” will be defined later in the problem and modeling portions of our experiment.

## Further Questions:

Here are three questions our group might further consider for our report:

1. Do red and blue states have significantly different Covid-19 growth rates from the onset of the pandemic and which features are most strongly associated with them? We would consider a logistic regression model to answer this question.
2. What is the relationship between geographic location and the spread of Covid-19? To what extent can we use the latitude and longitudinal data (location) to predict the Covid-19 growth rate? We would consider a linear regression model to answer this question.
3. How do the size and political majority of a county affect the Covid-19 growth rate? To what extent can we use linear regression to predict the Covid-19 growth rate based on these factors? We would consider a linear regression model to answer this question.

---

# Problem

What role does a county’s political affiliation play in determining its COVID-19 response? Is there a trend to begin with indicating that certain political demographics are more mindful about COVID-19? We are concerned with the empirical Covid-19 data of counties and their political affiliations to see if such a trend/connection exists. More specifically, our Covid-19 data will focus on the arguably two most influential factors for counties in Covid-19 spread: vaccinations and masking rates.

Thus, our problem begins with our desire to quantify the inherent difference that exists in counties of differing political affiliations based on their COVID-19 response. This makes no claims of causation, such that mask usage/vaccinations make a county red or blue, and vice-versa, that a county being red or blue determines its response, but rather aims to quantify the difference in response that already exists.

## Hypothesis (Right-Tailed Test)

Given a county’s mask usage and vaccination rates (cumulative vaccinations per capita from 12/13/20 to 9/11/21), we will predict a county's political majority with statistically significant accuracy. More specifically:
**Null Hypothesis:** Our county political majority classification model with accuracy that is not statistically significant from random guess (~50%) using the aforementioned features/data.
**Alternative Hypothesis:** Our county political majority classification model with accuracy that is significantly higher from random guess(~50%) using the aforementioned features/data.

We will define significance (statistical) as having p-value < 0.05 for a normal distribution centered around 0.5 with standard deviation = 0.1. We chose 0.1 according to the NumPy documentation for their percentile() function. More specifically, the accuracy at p-value = 0.05 is **0.6586**. Hence, our model must have accuracy higher than that for the Null Hypothesis to be rejected.

Our hypothesis can be confirmed with our existing datasets since we can create a classification model, like logistic regression, and compare its mean accuracy to the accuracy mentioned in the hypothesis.

Our “creative” datasets are the election data and the vaccination data. The Harvard Dataverse contains a dataset giving the number of votes by county for the candidates, which we used to see the constituent winner and assign a county its color. The dataset provided to us contained the vaccination percentage by state, however, we needed a more robust measure of vaccinations specifically for each county which would take into account the per capita increase in vaccinations. We acquired this data from the CDC database (Source 2) and narrowed it down to the days between Dec 13th, 2020, and Sept 11th, 2021. We took the total number of vaccines administered by Sept 11th and subtracted the total number at Dec 13th to get the total increase.

---

# Modeling

**Model to train:** Logistic regression

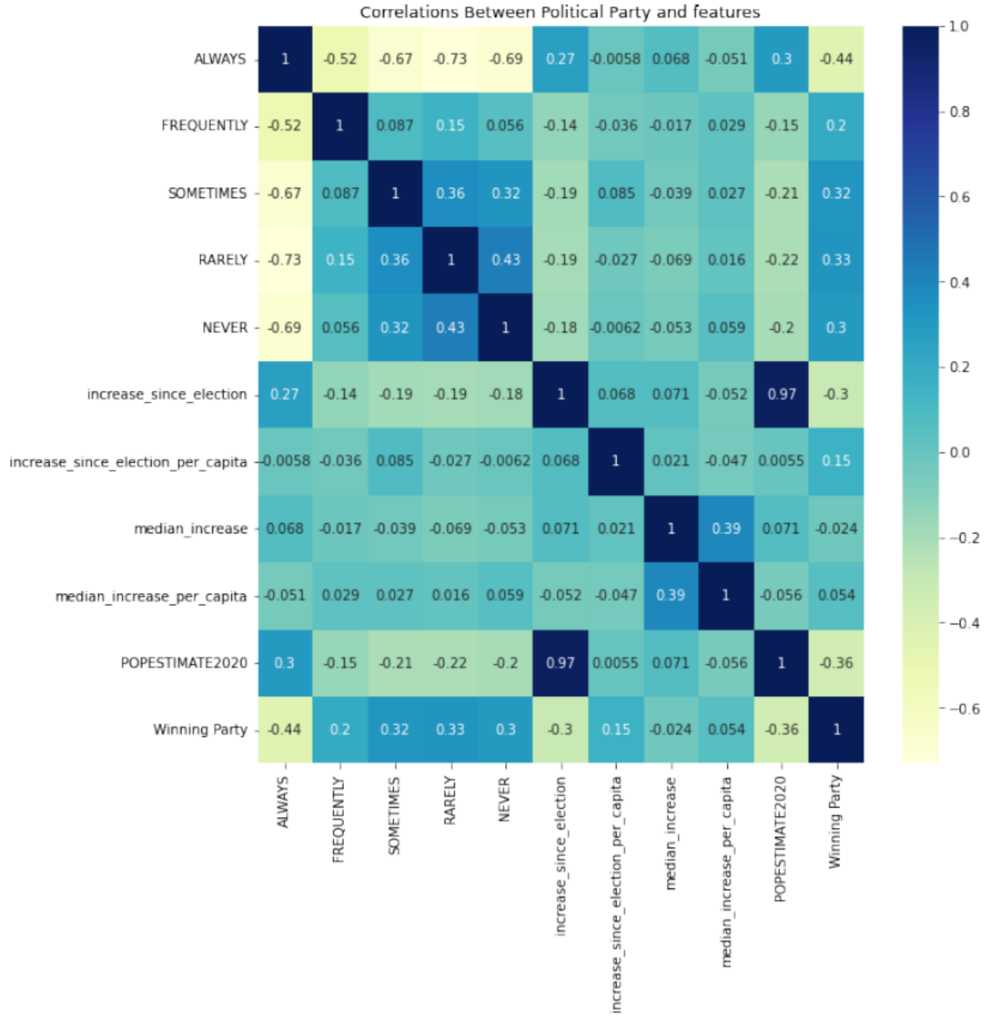
**Inputs:**

1. The “Always” category of mask use.
2. The number of COVID-19 vaccines administered from December 13th, 2020 (The US election day) to September 11th, 2021 (last day on the dataset) found on the dataset from the data.CDC.gov website (Source 2)

**Output:** The political party that the county voted for in the 2020 presidential elections.

**Model Choice:**

Since we are interested in predicting a value that is categorical in nature, and has a binary (republican vs democrat) output, we decided to use logistic regression for our model. Since we were interested in the potential association between COVID-19 spread and the political association of the county in question, we decided to select our features based on their correlation with the political party. In order to select our features, we used a revised version of the features heatmap, found below, using a wider range of features that we thought would be a good fit for our model:



From this, we noticed that the **"Always"** mask use category had the highest absolute correlation with the Winning Party category out of the entire mask use dataset. We decided that it was best to use only one of the categories since they all have relatively high correlations with each other and to reduce redundancy in our feature selections. This would hopefully reduce the complexity of our model, and prevent overfitting.

We also decided to use the total vaccinations per capita (vaccination rates) as a feature since we wanted to isolate this data and not have it depend on the county population. To justify, county population (POPESTIMATE2020) has a 0.071 correlation with median_increase so we deemed it irrelevant to covid spread.

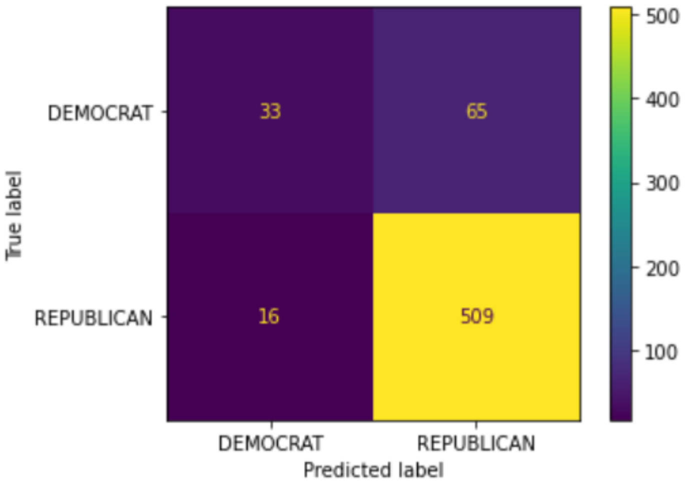
From this, if our model can predict the political party with above 0.6586 accuracy (which is addressed in the Evaluation portion of the report), we can conclude that there is a correlation between better response rate (masking and vaccination) to COVID-19 and the political affiliation of a county, and therefore **reject the null hypothesis**. Since we already established a high correlation between high case rates and political affiliation during the EDA portion of the report, we will not be looking at it as a feature in our model.

# Model Evaluation & Analysis

In order to evaluate our model, we used two different methods. First, we split our data into test data and train data, with 20% of the data being reserved for testing. Once we had the test predictions, we used the actual Political Party values of the test data, and compared them to the predictions in order to calculate the:

1. True positive - number of values that were predicted to be Republican, and were actually republican.
2. True negative - number of values that were predicted to be Democrat, and were actually democrat.
3. False positive - number of values that were predicted to be Republican, but were actually Democrat.
4. False negative - number of values that were predicted to be Democrat, but were actually Republican.

In the above list, “Republican” counties are represented as positive, whereas “Democrat” counties are represented as negative. The heatmap of the confusion matrix can be seen below:

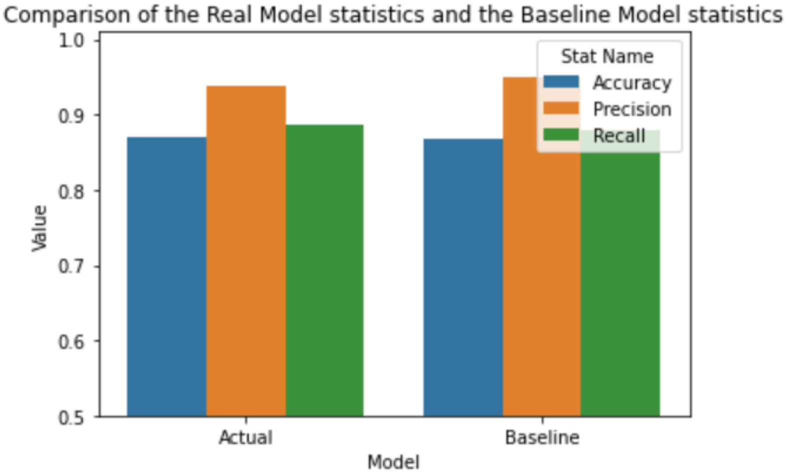


As seen from the confusion matrix, our True Positive value was the largest of all, while the False Positive value was comparatively small. This might suggest that our model is good at predicting if a county is Republican correctly. However our True Negative and False Negative values have the opposite relationship, which suggests that our model might not be as accurate at predicting Democratic counties.

In order to fully understand our model based on the confusion matrix, we also need to compute the **accuracy**, **recall**, and **precision** of our model. We can use the previously computed values for true positive, true negative, false positive, and false negative for our calculations. By doing so we get:

**Model Accuracy: ~**87% on the test set
**Model Precision: ~**94% on the test set
**Model Recall: ~**89% on the test set

From this we can say that our model does a better job than a random guesser, which would be correct ~50% of the time, hence have 50% accuracy. However, in order to test the validity of our model, and the feature selection we used, we decided to use **bootstrapping** in order to test our model to a version with a wider selection of features that have high correlation with political party, in order to compare it to a baseline model, to make sure our feature selection was a useful process. To do this, we created a set of features for our actual mode (discussed in the previous section on inputs), and a model that takes the highest correlation features from our dataset (ie all of the mask usage statistics, the number of case increases per capita since election day, and the median number of daily case increase per capita). Below is the graph representing the accuracy, recall, and precision of the two models:



The above is the result of creating 1000 models with each feature set (the baseline feature set of random features, and the actual feature set) and then calculating the mean of each of the statistics for the 1000 models. This was done with bootstrapped samples of the dataset for each of the models, to make sure that the iterations were random. As we can see, our models perform nearly identically to the randomized models, however, the randomized model takes 7 features from the correlations heatmap that was shown in the model choice section of the report, while the actual model only takes 2 features into account: “Always” mask use category and “Vaccinations per Capita Since the Presidential Election”.

Our current model uses fewer features and achieves the same accuracy, recall, and precision stats as the baseline model with the features taken from the correlations heatmap; however, because it uses fewer features, there is less risk in overfitting our model, and the model has reduced complexity as a result.

# Model Improvement

**Problem:** When we first created the model, we noticed that stats like “daily case increase since election day” and “median increase of daily cases” performed better in predicting political affiliation than their “per capita” counterparts. After constructing the correlation heatmap that is seen in the Model section of the report, it was clear that the “POPESTIMATE2020” which is a stat describing population estimates for the year 2020 has a high correlation with the political affiliation (-0.36), which suggests that the model was using the population in order to predict the political party of a county. In fact, the correlation suggests that there is a correlation between democratic counties and large populations.

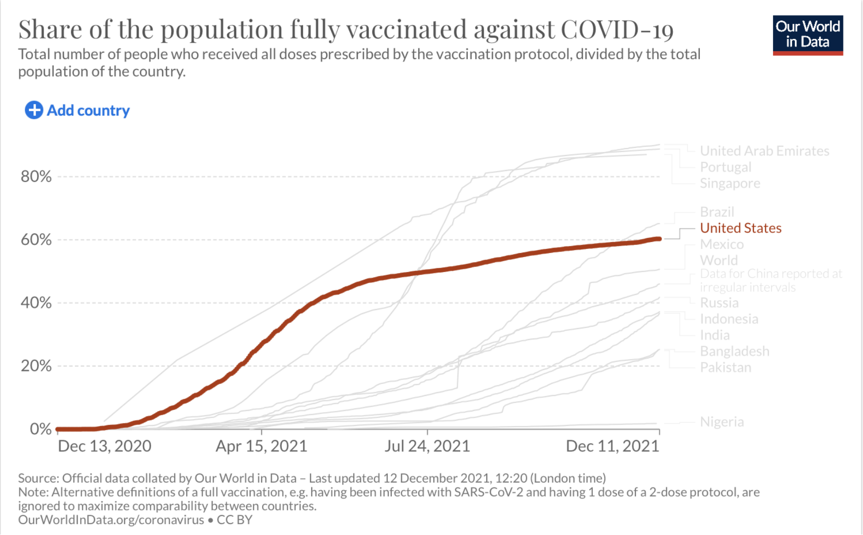
**Solution:** We first fixed the issue by normalizing the stats to be per capita, but found that they had a weak correlation with the political party, and did not provide a high enough accuracy. This is when we decided to look at vaccination statistics for each county. This was also due to the fact that we already observed a strong trend for COVID-19 case growth during our initial EDA and did not want to use stats that have already been seen to have an undeniable correlation with political parties.

**Result:** As seen from the model analysis section, the introduction of vaccinations per capita in our model helped bring the accuracy, recall, and precision of the model to be on par with the baseline model. It made sure that the model did not bias toward population statistics as it previously did, and helped to observe the relationship between specific factors of mask usage and vaccinations and the political party of a county, which directly ties into our hypothesis.

With regards to our A/B Testing, we can safely **reject the null hypothesis in favor of the alternate hypothesis**; our logistic regression model has an overall accuracy of ~0.87 which is much higher than 0.6586. This result confirms our hunch that there is a trend between a county’s political majority and their Covid-19 situation/response. Referring back to the line plot from our Open-Ended EDA, the Democratic counties clearly had a better response to Covid-19 as shown by their lower average cases growth per county and lower average number of cases overall. Through this report, our team solidified that this difference must be related to mask usage and vaccination rates of the counties. Hence, it's important that we get vaccinated and wear masks regularly when outside.

# Future Work

The future of COVID-19 research needs to take into consideration the current vaccination rates, the requirement of booster shots and their regularity, and more importantly, different strains that may arise and how the response to them may appear. The ability to predict a county's political affiliation given factors relating to COVID-19 response is useful for policy measures and knowing where to focus the county's recovery efforts.



Share of the population fully vaccinated against COVID-19

Total number of people who received all doses prescribed by the vaccination protocol, divided by the total population of the country.

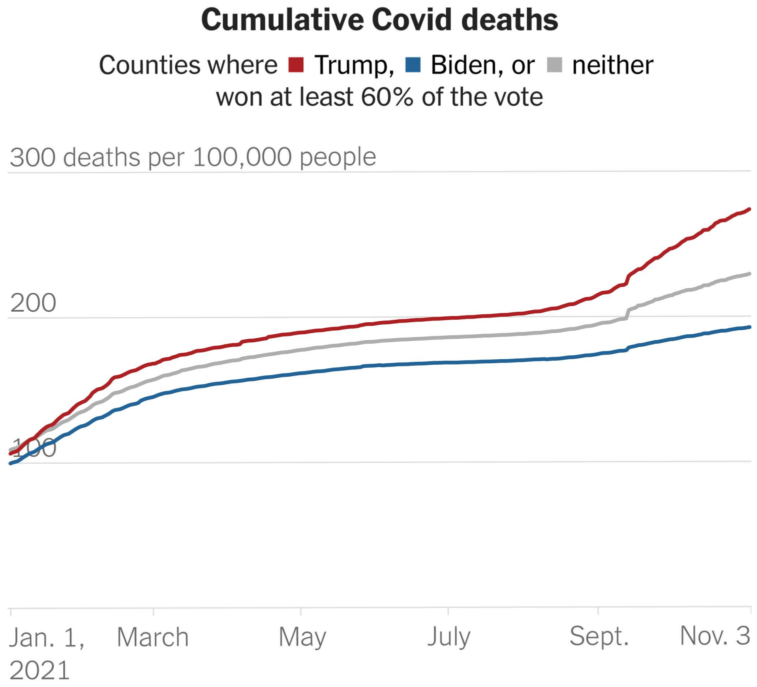
Source: Official data collated by Our World in Data – Last updated 12 December 2021, 12:20 (London time)
Note: Alternative definitions of a full vaccination, e.g. having been infected with SARS-CoV-2 and having 1 dose of a 2-dose protocol, are ignored to maximize comparability between countries.
OurWorldInData.org/coronavirus • CC BY

(Courtesy of Our World in Data)

The percentage of fully vaccinated people in the United States has been stagnating. As initial doses begin wearing off, the percentage even has the ability to fall, and 60% is not nearly an ideal amount to be at for that change. Simultaneously, with the sudden arrival of Omicron, it may be inevitable that more strains pop up.

The study of political affiliations not only helps with focusing the government's effort but also predicting the change in vaccination rates and masking. The vice-versa applies as well. If we were to study the change in vaccinations, masking, and deaths over time, especially considering pivotal moments in the fight against COVID such as the first phase of vaccinations, the first phase of booster doses, and initial impacts of new strains, could this data combine with data of political affiliation to give a better prediction at how changes in the COVID and vaccination landscape can impact different people? Our model would benefit from data regarding vaccination and booster patterns, and a model which takes into account the same features over time would have much higher accuracy.

One aspect we did not explore was the relationship between the location (the longitude and latitude) of a county and the rise in COVID cases and deaths. This could be useful for knowing exactly where deaths may rise or by how much if there was suspicion of a resurgence or a new strain, and thus medical supplies could be redirected there. The accuracy of this prediction could only be strengthened by taking into account political affiliation.



(Courtesy of New York Times Database, Edison Research) (Source 4)

The persistence of the Delta variant combined with slowing vaccinations led to a spike in cases in deaths in September 2021. This spike was unevenly greater in red counties and further increased the gap. Any study of COVID or any effort to collect data exists for the purpose of saving lives. We believe our decision to quantify the involvement of politics is essential because the effects of COVID are spread unevenly across different counties. Perhaps another political representative makes dubious claims about the vaccine. The CDC could potentially predict the quantifiable consequence of these claims and the locations at risk. The inequitable distribution of the rise in deaths and cases is attributed to politics amongst other factors, and any policymaker must then equitably distribute aid and efforts.

# Sources

1. Harvard Dataverse - County Presidential Elections Returns 2000-2020 Dataset : County Presidential Election Returns 2000-2020 - US Presidential Elections
2. Vaccinations per county dataset: COVID-19 Vaccinations in the United States,County | Data | Centers for Disease Control and Prevention (cdc.gov)
3. The Hill: The death rates from Covid in red America and blue America are growing further apart
4. NY Times: US Covid Deaths Get Even Redder