Failing to Follow the Herd: Factors Explaining Variations in Covid Vaccination Rates in the U.S.

Leo Yi

2021-10-10

Table of Contents

# Introduction

Over the past two years, humanity has been faced with a common enemy whose name is notoriously known across the world. COVID-19, which is caused by the coronavirus SARS-CoV-2, was first officially identified in January 2020 (Allam, 2020). The response by national leaders across the world has varied from apathy and denial to strict lockdowns. Within the United States, national and state leaders continue to push their own idea of how to face this enemy. Additionally, the public’s polarized opinions about how to handle the situation are also hindering the effectiveness of a unified strategy. These political issues are the reality of a world made up of individuals who work and live together.

The scientific community confronting COVID-19 in the US, led by the CDC, has been offering updated guidance for our government, businesses, and individuals (*Coronavirus Disease 2019 (COVID-19)*, n.d.). The guidance was based on the information available at the time and included suggestions of lock downs, face masks, social distancing, quarantines, and vaccines. In December of 2020, the FDA issued Emergency Use Authorization for the Pfizer-BioNTech and Moderna vaccines, followed by the Janssen vaccine in February 2021 (*COVID-19 Vaccines*, n.d.).

Vaccines are proven to be an effective strategy to fight viruses. Each of the vaccines currently available have been proven safe and effective at preventing severe disease (Bajema et al., 2021) and have been suggested and encouraged by the CDC. The federal government set out to vaccinate the public, with the initial strategy focused on logistics – how to produce, supply, deliver, and give the vaccine to every person in the US. As of September 11, 2021 vaccinations are now currently available, for free, for all adults with 53.9% of Americans fully vaccinated (“U.s. COVID-19 Vaccine Tracker,” n.d.).

The situation we face today is the lower than ideal vaccination rates across the country. For various personal reasons, people have chosen to refuse the vaccine, with some actively campaigning against it. The issue now relates to social and political differences rather than logistics and the goal of herd immunity seems to be out of reach (Aschwanden, 2021).

The purpose of this research is to seek to understand the connections that may exist between people who ignore the science of vaccines and the collective knowledge of humanity to believe only what they choose to believe. If vaccine rates have plateaued and we can assume those who have not yet vaccinated are doubtful of science, this may be the first time we have collected data that can tell us who believes in science and humanity. This research will aim to shed some more light on who we are.

# Literature Review

As a nation with a goal to reach vaccine levels sufficient for herd immunity, a major obstacle is that some people refuse vaccines. The anti-vaccine movement has existed well before COVID, with unfounded claims that vaccines cause autism. These claims have been spread far enough that the CDC offers their official perspective on it’s website in order to clarify the now common misinformation (“Autism and Vaccines,” 2021).

In the modern world, advances in all areas of study make it impossible for any one person to know about everything. An injection of some unknown substance may lead to fear without the previous knowledge and understanding of what a vaccine is and how it works (Cerda & García, 2021). Additionally, mistrust of the government, scientists, healthcare workers, or the system altogether can lead to wild speculation about what substances are actually entering our bodies.

The exposure to competing views of information and misinformation in social and mass media adds to the confusion and hinders people’s ability to make a well informed decision (Hussain et al., 2018). Additionally, people who are led into the anti-vaccination movement may be influenced by the misinformation spread by Donald Trump and the surrounding conservative movement, which provides a social network that is engaging, supportive, and based on feelings of community (Germani & Biller-Andorno, n.d.).

# Research Questions

What factors can explain variations in vaccination rates in the US?

If we create a projection for vaccination rates, when might we hit herd immunity of 70% vaccination rates (“Herd Immunity and Covid-19 (Coronavirus),” 2021), if ever?

The idea behind these questions comes from the phenomena that people are choosing to go against the best known strategy humanity has formed so far, with no real effective alternative strategy. Before understanding why people might be doing this, the goal in this research is to see if we can find commonalities in people who choose to ignore the herd and the facts and believe their own version of reality. That might lead us to why vaccination rates are not currently high enough to hit herd immunity.

Based on current research, political party affiliation seems to be a major factor affecting vaccination rates. Additionally, it’s possible that education would affect people’s willingness to accept and trust modern scientific knowledge and also be less likely to be influenced by misinformation campaigns in social media (Germani & Biller-Andorno, n.d.). It’s not clear whether other county level factors will show correlations, but the idea is to check whether we can find any statistically signficant factors that affect vaccination rates, in order to further consider why these connections may exist.

# Data

The available data for vaccine rates are aggregated at the country and day level by the Centers for Disease Control and Prevention (“Covid-19 Vaccinations in the United States,county,” n.d.). The Federal Information Processing Standard (FIPS) code will be the county level code which all variables will be collected and aggregated. The CDC provides vaccination rates starting from the end of 2020, excluding Texas.

The US Census Bureau provides population data by age, sex, and race (Bureau, 2021a), which will be used both as a way to standardize all other variables as a percent of the population, as well as to check if certain age groups or ethnicity might be relevant factors in overall vaccination rates.

The Association of Religion Data Archives provides rates of religious followers by denomination (“Download,” n.d.), however rather than using each minor sect, we’ll only be looking at rates of religious people in general and densities of any religious congregations to see if religious beliefs may affect vaccination rates.

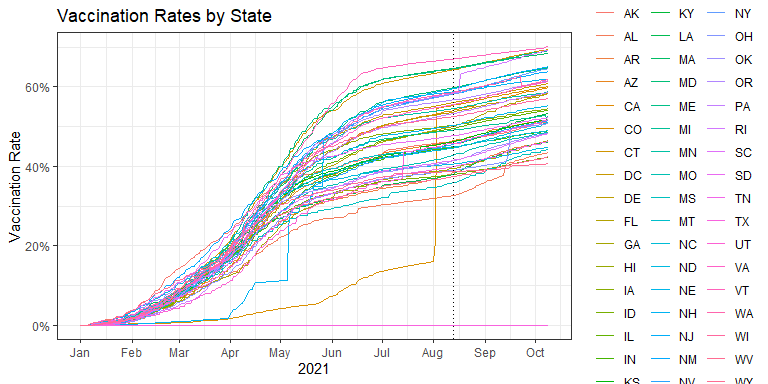
The Economic Research Service of the US Department of Agriculture was used to gather socioeconomic data related to unemployment, median household income, education attainment, poverty estimates, a rural-urban scale, and an urban influence factor (n.d.). Additionally, the ERS provides a data on economic topology, which provides additional variables related to the level of different industries by county, as well as child poverty levels (“County Typology Codes,” n.d.).

For political factors, the 2020 Election Results by County were used in order to assess the policy perspectives of its constituents (Tonmcg, n.d.). It should be noted that Alaska is the only state that has different FIPS codes for the counties and voting districts, which makes vaccine rates and election results incompatible.

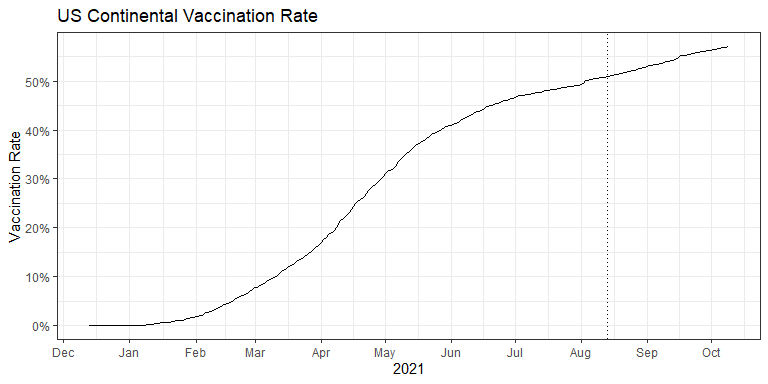
An obvious factor that could influence a counties vaccination rates would be the number of COVID related cases and deaths (Nytimes, n.d.) experienced. A less obvious set of variables would be related to birth data, which includes information about babies and their mothers (“Natality, 2016-2019 Expanded Request,” n.d.).

Finally, some external factors were included, like temperature ranges and average precipitation (NCEI.Monitoring.info@noaa.gov, n.d.), environmental quality (“County-Level Cumulative Environmental Quality Associated with Cancer Incidence.” 2020), and crime statistics (Bureau, 2021b).

There are 2,888 observations or counties after excluding Texas and island territories. After all the variables have been joined, there are 233 counties with missing data which have been omitted from the final dataset, leaving 2,665 complete observations. This includes 1 independent variable, vaccination rate, and 125 dependent variables.



Looking at the plot above showing rates for all states since the beginning of 2021, the vertical dotted line represents August 13th, which is a Friday. This date was chosen as the snapshot date because it is between corrections, before mandates were issued, and also lagged from the present in order to include data that comes in late.



# Statistical Methods

Using vaccination rates as the dependent variable, we can create supervised machine learning models to determine which variables have the highest impact. The independent variables will be county level data sources which are made available by the government and the dependent variables have been sourced from the other government and other organizations.

With 2,665 observations, the data will be split into 75% as a training group with 2,001 observations and 25% as a test group with 664 observations. Each of the machine learning models will use all variables, with all models except decision trees using 10 fold cross validation. We will determine the final model which has the lowest validation error in order to assess the most important variables.

Finally, we can check the distributions of vaccination rates based on the most important variables and assign classes (low / mid / high). We will create two forecasts, one using the overall US vaccination rates as a time series, and another using each of the classes as independent time series. The latter group will be combined, weighted by population, and compared to the overall forecast. This will allow us to estimate when the US might reach herd immunity of 70% vaccinated.

The rates provided are based on the population considered, so the true overall US rates including all counties will vary.

The final time series projection will be selected based on the measure of error against test data using the most recent vaccination rates, held out from the data used to train the models.

After training the models, the predictions on the test data set are compared to determine the best predictor.

### Model Validation Results

## X RMSE Rsquared MAE  
## 1 Cubist 0.07610659 0.6748728 0.04821745  
## 2 ranger 0.08253407 0.6211878 0.05238892  
## 3 Random Forest 0.08271117 0.6214280 0.05256042  
## 4 SVM 0.08463286 0.5990194 0.05653898  
## 5 Neural Network 0.08487343 0.6001813 0.05602094  
## 6 Boosting 0.08865653 0.5609635 0.05727506  
## 7 Bagging 0.09107649 0.5382184 0.06114104  
## 8 LASSO 0.09235012 0.5225623 0.06177234  
## 9 Ridge 0.09356212 0.5168556 0.06365635  
## 10 MARS 0.09612205 0.4881134 0.06330440  
## 11 KNN 0.09813398 0.4670728 0.06726223

### Variable Importance

Here’s a look at the most important variables of the cubist model:

## X Overall  
## 1 per\_gop 100.00000  
## 2 college\_pct 52.94118  
## 3 avg\_age\_of\_mother 52.10084  
## 4 sub\_hs\_pct 50.42017  
## 5 avg\_birth\_weight\_grams 47.89916  
## 6 pct\_black 47.89916  
## 7 per\_dem 47.05882  
## 8 pct\_age\_70\_74 46.21849  
## 9 temp\_max 45.37815  
## 10 temp\_max\_avg 45.37815

### Political Classes

## grp pop\_in\_M fips\_count avg\_pop\_per\_fip\_K  
## 1 center 119.5 709 168.6  
## 2 left 111.3 268 415.2  
## 3 right 67.7 1881 36.0

Since per\_gop was by far the most important variable, we’ll use it to classify each of the counties. The variable per\_gop represents the percent of the county who voted for the GOP in the 2020 election. We can then classify each county as left leaning with less than 40% voting Republican, right leaning with greater than 60% of the county voting Republican, or in the center with 40 - 60% of the county voting Republican.

The vaccine rate data at the time of publishing has 301 days of data and we’ll be using 80% or 240 days of data to train the forecasting models and the remaining 20% or 61 days of data to validate the relative predictive errors.

### Forecasts

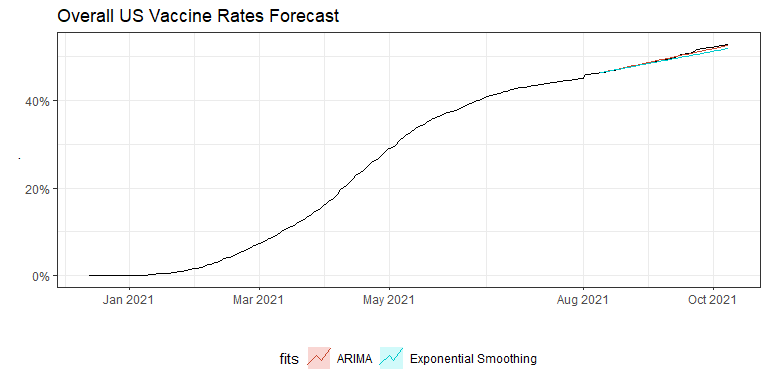
#### Overall Forecast Results

Overall vaccine rates auto.arima() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00002008745 0.000781087 0.0004868231 2.07689435 2.578264 NaN  
## Test set 0.00032016722 0.002606152 0.0022442424 0.04780755 0.446708 NaN  
## ACF1 Theil's U  
## Training set -0.005785283 NA  
## Test set 0.959832252 2.016234

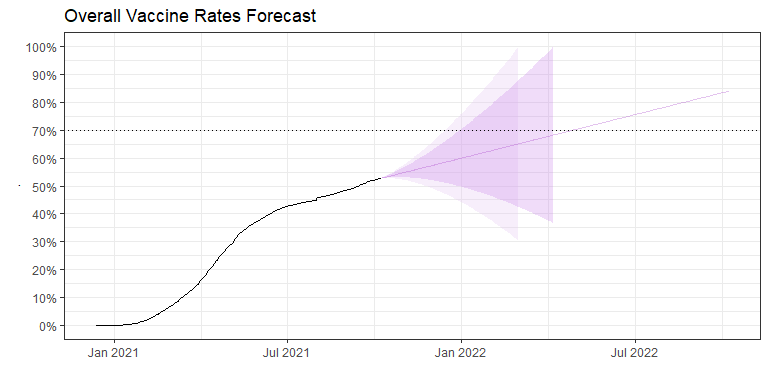
Overall vaccine rates ets() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00001152718 0.0008224402 0.000530931 -Inf Inf NaN  
## Test set 0.00366658154 0.0056788192 0.003978345 0.7084952 0.7749042 NaN  
## ACF1 Theil's U  
## Training set 0.2218056 NA  
## Test set 0.9658865 4.31746



Looking at forecasting overall US vaccine rates using ARIMA and exponential smoothing, the ARIMA model produces a slightly more accurate forecast, which we’ll use for the overall forecast.

##### Retrained Final Forecast, Overall US



After retraining the model using the full time series, we can predict that herd immunity would be reached after 2Q 2022. The 80% prediction interval suggests a possibility of reaching 70% vaccination rates before the end of 2021 and the 95% prediction interval indicates its possible at the end of the year.

Note that the intervals on the chart do not reflect the reality that vaccination rates will not drop and can only go up.

#### Political Class Forecast Results

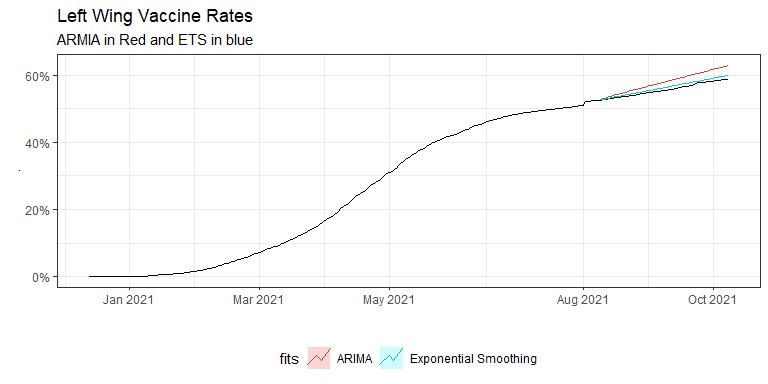
##### Left Leaning

Left wing vaccine rates auto.arima() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00002112706 0.0009236769 0.0005319993 2.117644 2.839497 NaN  
## Test set -0.02277338376 0.0249657374 0.0227733838 -4.035834 4.035834 NaN  
## ACF1 Theil's U  
## Training set 0.01306561 NA  
## Test set 0.93433060 20.1806

Left wing vaccine rates ets() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00001912351 0.0009812393 0.0005901451 -Inf Inf NaN  
## Test set -0.00652918265 0.0070738113 0.0065291826 -1.161448 1.161448 NaN  
## ACF1 Theil's U  
## Training set 0.1648555 NA  
## Test set 0.9064327 5.755536



For the counties who leaned left politically, the exponential smoothing model produced a more accurate forecast.

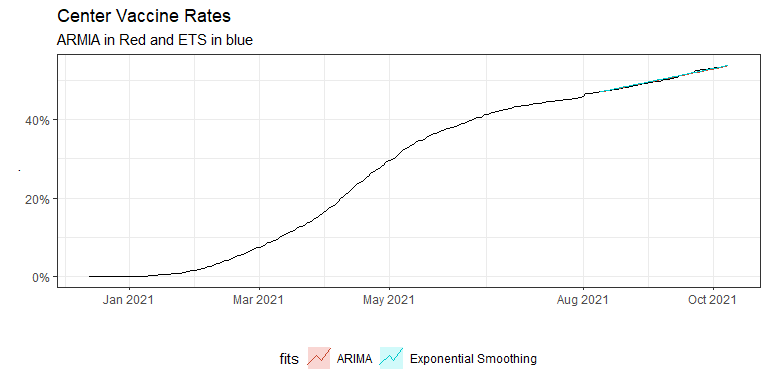
##### Moderate Leaning

Center vaccine rates auto.arima() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00002201552 0.0008352299 0.0005186638 2.1291895 2.9363581 NaN  
## Test set -0.00044794525 0.0021338263 0.0019023698 -0.1015846 0.3779673 NaN  
## ACF1 Theil's U  
## Training set -0.004570358 NA  
## Test set 0.951229291 1.674934

Center vaccine rates ets() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00002572256 0.000875336 0.0005630098 NaN Inf NaN  
## Test set -0.00123349138 0.002255843 0.0019525232 -0.2555926 0.392319 NaN  
## ACF1 Theil's U  
## Training set 0.2749466 NA  
## Test set 0.9418885 1.800418



For politically moderate counties, both forecasts had low error, with ARIMA having a slight edge.

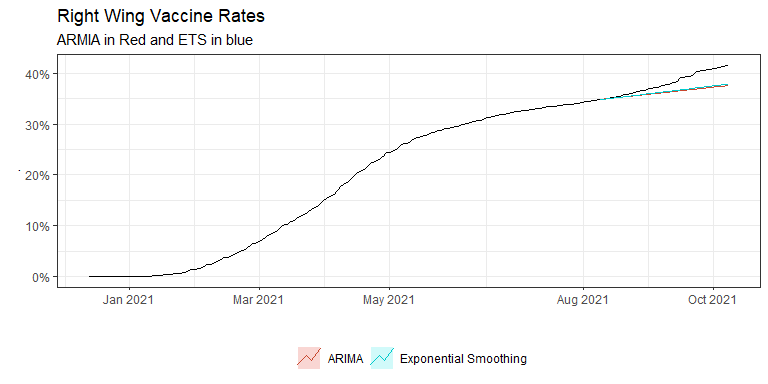
##### Right Leaning

Right wing vaccine rates auto.arima() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.000004015738 0.0005454404 0.000351378 1.423301 3.148563 NaN  
## Test set 0.017895767860 0.0224024682 0.017906635 4.527721 4.530838 NaN  
## ACF1 Theil's U  
## Training set -0.2340338 NA  
## Test set 0.9614654 15.51741

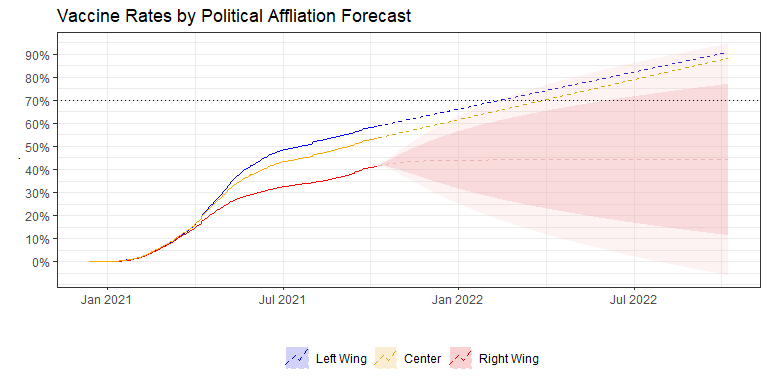
Right wing vaccine rates ets() accuracy:

## ME RMSE MAE MPE MAPE MASE  
## Training set 0.000003087053 0.0006923589 0.0004776573 -Inf Inf NaN  
## Test set 0.016612356158 0.0207964629 0.0166123562 4.203271 4.203271 NaN  
## ACF1 Theil's U  
## Training set 0.2820044 NA  
## Test set 0.9617446 14.40181



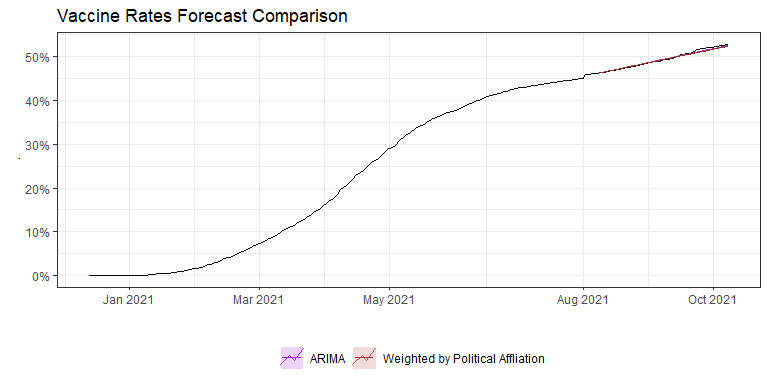
Although exponential smoothing produces a lower test error using RMSE, both forecast methods are underestimating, possibly due to changes in the landscape related to vaccination mandates, surges of cases, deaths, as well as the weakening influence of Donald Trump.

#### Forecasts by Political Affliation Class



The chart above shows each of the classes with models retrained using the methods with the lowest test error. We can see that left leaning counties have the highest rate of vaccinations, followed by moderates, with conservatives trailing. Both left and moderate leaning counties are projected to hit 70% vaccination rates before 2Q 2022, which suggests that the right wing vaccination rates are pulling down the overall average rates.

#### Comparing Final Forecasts



Comparing the forecasts on the test set show both models with very low test error.

Overall US vaccine rates ARIMA error:

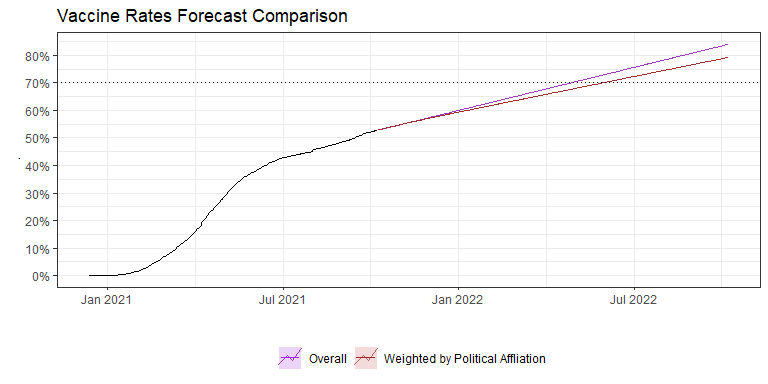
## ME RMSE MAE MPE MAPE MASE  
## Training set 0.00002008745 0.000781087 0.0004868231 2.07689435 2.578264 NaN  
## Test set 0.00032016722 0.002606152 0.0022442424 0.04780755 0.446708 NaN  
## ACF1 Theil's U  
## Training set -0.005785283 NA  
## Test set 0.959832252 2.016234

Weighted Average Time Series Test error:

## ME RMSE MAE MPE MAPE ACF1  
## Test set 0.001090778 0.003211247 0.002532556 0.1993123 0.4996213 0.9639573  
## Theil's U  
## Test set 2.457126

The overall forecast seems to have a slightly lower error based on RMSE. Its possible that the change in trend seen in the right leaning class is being averaged in total and leading to a smoother trend overall.

#### Final Forecast



Above, we see the comparison of the overall US forecast compared to the combined forecasts of political classes, weighted by population. Both models above have been retrained using the entire time series.

The politically weighted forecast slows down the overall rate and suggests herd immunity can be reached between 2Q and 3Q 2022, about a month later than the overall US model.

# Discussion of Results

The attempt to gather as many variables as possible did not lead to any surprising results and resulted in validating earlier research relating to the online behavior of conservatives (Germani & Biller-Andorno, n.d.). The percent of the population who graduated from college was the 2nd most important variable, but still about half as important as the percent of the voters in a county who voted for Donald Trump. The percent of people who did not graduate high school was the 4th most important variable, further validating the assumption that education affects an individual’s ability to understand vaccination and their susceptibility to misinformation online.

The average age of mother at birth and the average weight of a baby were the 3rd and 5th most important variables in the leading cubist model, which may suggest that the lifestyle and health of mothers in a given county could predict people’s affinity to vaccines. Again, both of these factors were about half as important as the percent of voters who voted for Donald Trump. It’s likely that these factors are not directly related to vaccination rates and instead are related to or collinear to other factors not included in this research.

Finally, using the data we have available, we can project that the US could hit herd immunity before the start of summer 2022. This would be possible if we continued vaccinating the population at the current pace, but it’s entirely possible the rate of vaccinations slows and reaches a plateau as we face the sector of the population who refuse to vaccinate. This could keep us from reaching herd immunity as a nation, even if certain left leaning and moderate counties may have already reached over 70% vaccinated.

Validating earlier research suggests that improving education could have long term benefits for receptiveness to the understanding of vaccines as well as reducing the impact of misinformation, while the short term strategy includes actively fighting against misinformation and educating the adult public about vaccines through marketing, or other forms that reach beyond schools.

# References

(n.d.). In *USDA ERS - Download Data*. US Deptartment of Agriculture Economic Research Service. <https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>

Allam, Z. (2020). The first 50 days of covid-19: A detailed chronological timeline and extensive review of literature documenting the pandemic. In *Surveying the Covid-19 Pandemic and its Implications*. U.S. National Library of Medicine. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7378494/>

Aschwanden, C. (2021). Five reasons why COVID herd immunity is probably impossible. In *Nature News*. Nature Publishing Group. <https://www.nature.com/articles/d41586-021-00728-2>

Autism and vaccines. (2021). In *Centers for Disease Control and Prevention*. Centers for Disease Control; Prevention. <https://www.cdc.gov/vaccinesafety/concerns/autism.html>

Bajema, K. L., Dahl, R. M., Prill, M. M., Meites, E., Rodriguez-Barradas, M. C., Marconi, V. C., Beenhouwer, D. O., McMorrow, M., Schrag, S. J., Hall, A. J., & al., et. (2021). Effectiveness of covid-19 MRNA vaccines against COVID-19–associated hospitalization - five veterans affairs medical centers, united states, february 1–august 6, 2021. In *Morbidity and Mortality Weekly Report*. Centers for Disease Control; Prevention. <https://www.cdc.gov/mmwr/volumes/70/wr/mm7037e3.htm>

Bureau, U. C. (2021a). County population by characteristics: 2010-2019. In *Census.gov*. <https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-detail.html>

Bureau, U. C. (2021b). Crime. In *Census.gov*. United States Census Bureau. <https://www.census.gov/library/publications/2010/compendia/databooks/crime.html>

Cerda, A. A., & García, L. Y. (2021). Hesitation and refusal factors in individuals’ decision-making processes regarding a coronavirus disease 2019 vaccination. In *Frontiers*. Frontiers. <https://www.frontiersin.org/articles/10.3389/fpubh.2021.626852/full>

*Coronavirus disease 2019 (COVID-19)*. (n.d.). Centers for Disease Control; Prevention. <https://www.cdc.gov/coronavirus/2019-ncov/index.html>

County typology codes. (n.d.). In *USDA ERS - County Typology Codes*. US Department of Agriculture Economic Research Service. <https://www.ers.usda.gov/data-products/county-typology-codes/>

County-level cumulative environmental quality associated with cancer incidence. (2020). In *County-level cumulative environmental quality associated with cancer incidence. - CKAN*. Publisher U.S. EPA Office of Research; Development (ORD). <https://catalog.data.gov/dataset/county-level-cumulative-environmental-quality-associated-with-cancer-incidence>

Covid-19 vaccinations in the united states,county. (n.d.). In *Centers for Disease Control and Prevention*. Centers for Disease Control; Prevention. <https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh>

*COVID-19 vaccines*. (n.d.). U.S. Food; Drug Administration. <https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines>

Download: RCMSCY10: Data archive: The association of religion data archives. (n.d.). In *Download | RCMSCY10 | Data Archive | The Association of Religion Data Archives*. <https://www.thearda.com/Archive/Files/Downloads/RCMSCY10_DL2.asp>

Germani, F., & Biller-Andorno, N. (n.d.). The anti-vaccination infodemic on social media: A behavioral analysis. In *PLOS ONE*. Public Library of Science. <https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0247642>

Herd immunity and covid-19 (coronavirus): What you need to know. (2021). In *Mayo Clinic*. Mayo Foundation for Medical Education; Research. <https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/herd-immunity-and-coronavirus/art-20486808>

Hussain, A., Ali, S., Ahmed, M., & Hussain, S. (2018). The anti-vaccination movement: A regression in modern medicine. In *Cureus*. US National Library of Medicine National Institutes of Health. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6122668/>

Natality, 2016-2019 expanded request. (n.d.). In *Centers for Disease Control and Prevention*. Centers for Disease Control; Prevention. <https://wonder.cdc.gov/natality-expanded-current.html>

NCEI.Monitoring.info@noaa.gov. (n.d.). Climate at a glance. In *National Climatic Data Center*. <https://www.ncdc.noaa.gov/cag/county/mapping/110/tavg/202108/12/value>

Nytimes. (n.d.). Nytimes/COVID-19-data: An ongoing repository of data on coronavirus cases and deaths in the u.s. In *GitHub*. <https://github.com/nytimes/covid-19-data>

Tonmcg. (n.d.). In *GitHub*. <https://github.com/tonmcg/US_County_Level_Election_Results_08-20>

U.s. COVID-19 vaccine tracker. (n.d.). In *Mayo Clinic*. Mayo Foundation for Medical Education; Research. <https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker>