Jessie Lee

Ambikha Maharaj

Charan Musunuru

Karthick Ramasubramanian

COVID Vaccination: What's limiting us?

Project Goals

Description

In this project, our mission is to find the most important factors that may contribute to variation in vaccination rates across states in the U.S. By identifying these factors, we can figure out why some states may have higher vaccination rates than others and how states with lower vaccination rates can improve their vaccination rates so that the vaccine reaches as many people as possible. We can predict the vaccination rate of certain states if the current vaccination rate is not known for the respective state. However, we know the vaccination rate for all states and territories for the U.S., so in this project, we focus our attention on finding the most important predictors that may have a significant impact on the vaccination rates in each state. Using the following datasets, we want to gain insights into the vaccination rate distribution among states and counties, and the demographic distribution among vaccinated people.

- CDC Vaccination Data: CDC maintains the total vaccination rate of each county and state and is updated daily
- CDC Vaccination Hesitancy Index: CDC conducted a survey in 2020 to ask if people have hesitation towards the COVID vaccination. The options for them include: no hesitation, unsure(low hesitation), hesitate(median hesitation) and very hesitate (high hesitation)
- CDC Vaccine Coverage Index: CDC published research in February 2020 to predict the vaccination outcome based on some medical concerns of each area (state and counties). These themes include historic undervaccination, sociodemographic barriers, resource constrained Healthcare System, healthcare accessibility barriers and irregular care seeking behavior. Then they also calculated a weighted average index of medical concern (CVAC)
- Religious Landscape Study. (Pew Research): This survey measured the number of people who believe in god and other activities like going to church and praying

- Population and demographic data from the U.S. Census Bureau
- Religious Landscape Study and Nationwide Religion and Index form Pew Research Foundation
- 2020 U.S. Presidential Election, which indicated either Democratic or Republican

<u>Importance of the Problem</u>

It is important to get as many people vaccinated as soon as possible due to the idea of herd immunity. When a population becomes immune to a disease, through vaccination, there is indirect protection provided to those in the population who are not immune to the disease. The higher level of immunity for a population, the more people will benefit from it (D'Souza and Dowdy, 2021). It is therefore imperative for people to get vaccinated to not only protect themselves, but others as well. The work done in this project would be beneficial for states and also local governments to assess how to further combat issues that states face in terms of the factors that are influencing vaccination rates to help boost the number of people receiving vaccines in states with low vaccine rates.

Exploratory Analysis

Our exploratory data analysis starts with Figure 1 which identifies the variation in the number of states that are vaccinated while Figure 2 shows the top 10 vaccinated states and the bottom 10 vaccinated states. By looking at the top 10 states, it is known that these states tend to lean toward the Democratic party while the bottom 10 states are all favored towards the Republican party. The low amount of vaccine rates in some states can be attributed to the lag in the distribution of the administered vaccines. The plot in Figure 3 indicates that vaccines are in fact not being used in these undervaccinated states and 4 out of the bottom 10 states by vaccination have the highest amount of lag. More EDA was conducted on the CVAC index along with the political affiliation index, belief in God index, and hesitancy index as well (shown in Figures 4,5,6,7). It should be noted that in both the belief in God index and hesitancy index graphs that as the vaccination rate decreases the index for these two themes increase. This idea stays consistent with our final result as to the most important predictors for vaccination rates which will be discussed later in this report. Density plots were made to compare and contrast the variation in peaks over intervals in the different themes. Figure 8 illustrates the different distributions among states with both high and low concerns about medical coverage. This factor is derived from 5 sub themes as given by the CDC. It is centered around 70% for the state level with low medical coverage concerns and also shows a different distribution among states that are more religious and less religious. Figure 9 illustrates the data peaks of the Republican and Democratic data are significantly distinct with little to no overlap. This bimodal distribution indicates that there are two different groups, and in this case they are those who are Democratic or Republican. Likewise, the density plot in Figure 10 indicates a large discrepancy between religious and non-religious people as the values of the peaks of the two curves are concentrated over different intervals. This bimodal distribution indicates that there are two different groups, those who are religious and those who are non-religious. Finally, like the religious index and the political affiliation density plots, Figure 11 indicates that there are two distinct groups which are those who are hesitant about receiving the vaccine and those who are not hesitant and which accounts for the 2 separate peaks on the plot itself. Accessibility to healthcare and irregular care seeking behavior are more of a concern as compared to other themes such as historic under vaccination, sociodemographic barriers, and resource constraints (as seen in Figure 12). The peaks for the accessibility to healthcare density plot shown in Figure prove to have greater differentiation in its peaks which is significant in determining the two varied groups present. Next, a correlation matrix was produced, as shown in Figure 13, in which 2 predictor groups were found: medical themes and behavior themes. Medical themes include the vaccination rate, historic under vaccination levels, the sociodemographic barriers, resource constrained healthcare systems, healthcare accessibility, and irregular care seeking behavior. Behavior themes include low, medium, and high hesitancy levels along with party affiliation and religious index. These two predictor groups are the basis for data modeling along to help find the most important predictors that determine vaccination rate.

Modeling

A linear regression model was put in place for datasets in both the state and county levels. We first ran a linear regression on a state level to identify the themes present in each state of the United States to analyze and predict the most important variables and we also ran a linear regression on a county level to see if the same important predictors that we found on the state level matched those on a county level (Figure 15). The images in Figure 14 depict the results of the linear regression using only the medical themes where we found that adding each of the themes, starting with hesitation with Figure 14.2 to adding the political views theme in Figure 14.3 increases the R2 value. The final R2 value including all the themes was found to be 84%. This tells us that adding the additional predictors helps us to improve our regression model and is adding value to the model itself.

The random forest model shown in Figure 16, for the state level, was built in which the parameters were optimized for the model. Each tree was built by bagging around 50 levels rows for the state level and sampled the 50 row dataset many times and found that 850 trees is the optimal level. We then set the depth level from 0-100 and found that the maximum depth is 6.

For the maximum number of leaf nodes for each tree, we found the maximum nodes to be 17 and used to decide on the complexity of our decision trees. The same method was used for the Random Forest for the county level which can be seen in Figure 17.

Based on the results of the Linear Regression and Random Forest Model, our sample was consistent with our data exploration as well. In terms of finding the most important features that determine the vaccination rate across the U.S, they include religion, political affiliation, and the level of hesitancy to receive the vaccine as seen in bar plots shown in Figure 18. The predictors on the county level were also found and similar results were produced as shown in the feature importance in Figure 19.

Solution and Insights

As mentioned, religion plays a large factor in terms of what areas of the U.S. have lower vaccine rates compared to others. Based on our datasets, we expected that religion would play a significant role in the distribution of vaccination rates in a state. This is because it is known that those who refuse vaccines can be due to their religious beliefs. Next, political views in terms of states that are predominantly Democratic or Republican have shown to be a driving factor in the differences in vaccination rates. Democratic states are more likely to receive the vaccine while Republican dominated states are more reluctant to receive the vaccine. Staying consistent with our findings, The Kaiser Family Foundation released a survey to track the red/blue divide in vaccination rates and found that vaccination rates were lower in counties that voted for Trump in the 2020 election compared to those who voted for Biden. By looking at the top 10 states and bottom 10 states, it is known that these states tend to lean toward the Democratic party while the bottom 10 states are all favored towards the Republican party. In terms of the hesitancy index, the CDC put out an HPS survey in which they asked the question of if given the opportunity to receive the vaccine when available, would people willingly choose to get it. This factor makes sense when analyzing vaccination rates in certain states and counties because areas that are more hesitant to receive the vaccine are also usually attributed to the level of Democratic and Republican affiliation in that area. Overall, combating problems such as discrepancies with religion, issues with political affiliation influencing vaccine rates, and hesitancy within U.S. communities will help increase vaccination rates among the U.S. Only by removing these barriers can the U.S. achieve a higher vaccine coverage and develop targeted solutions to achieve herd immunity throughout the nation.

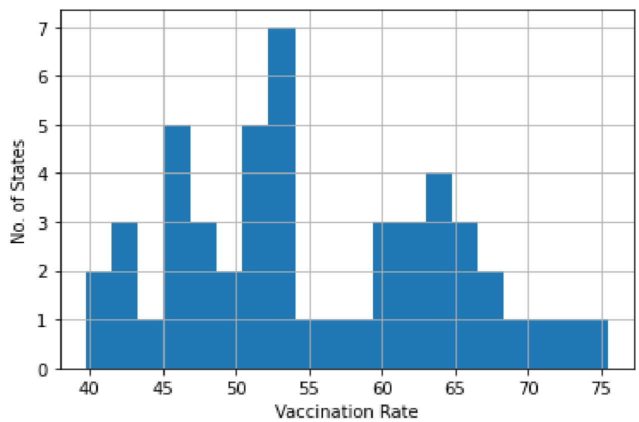

Figure 1

state	vac_rate
Vermont	75.5
Massachusetts	72.6
Hawaii	71.3
Connecticut	69.8
Maine	68.3
Rhode Island	67.3
New Jersey	65.8
Pennsylvania	65.4
New Mexico	65.3
California	64.7

state	vac_rate
Mississippi	39.7
Idaho	41.1
Wyoming	41.6
Louisiana	42.1
Alabama	43.2
Tennessee	44.6
North Dakota	45.4
West Virginia	46.0
Georgia	46.2
South Carolina	46.6

Top 10 Vaccinated States   Bottom 10
Vaccinated States

Figure 2

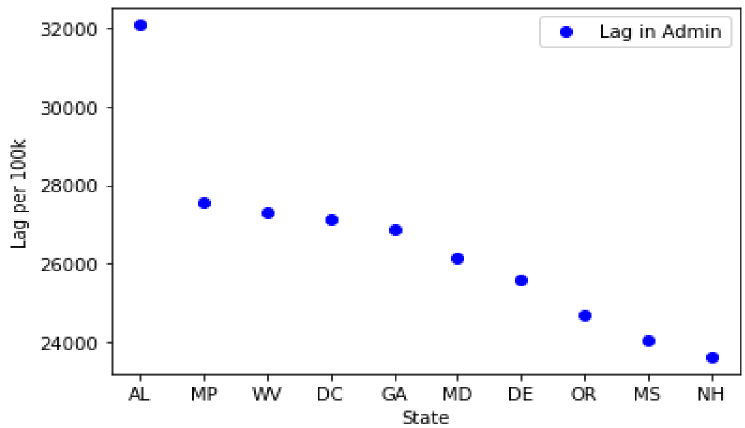

Figure 3

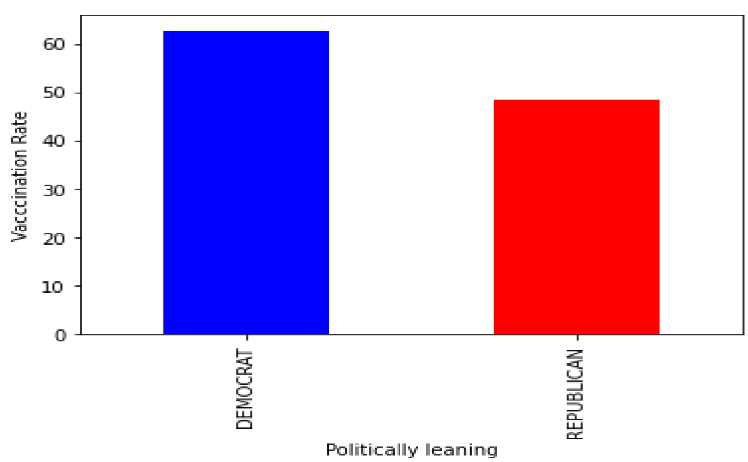

Figure 4

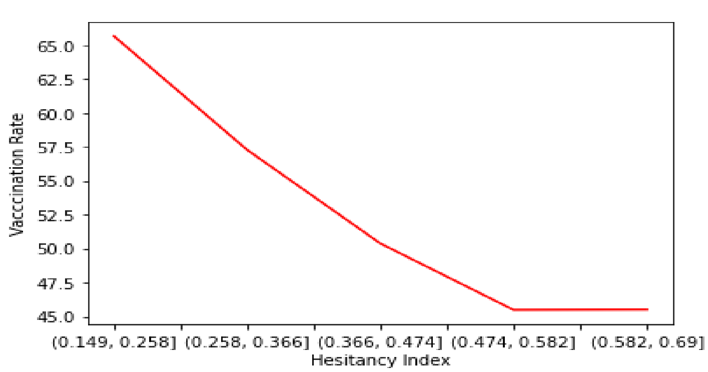

Figure 5

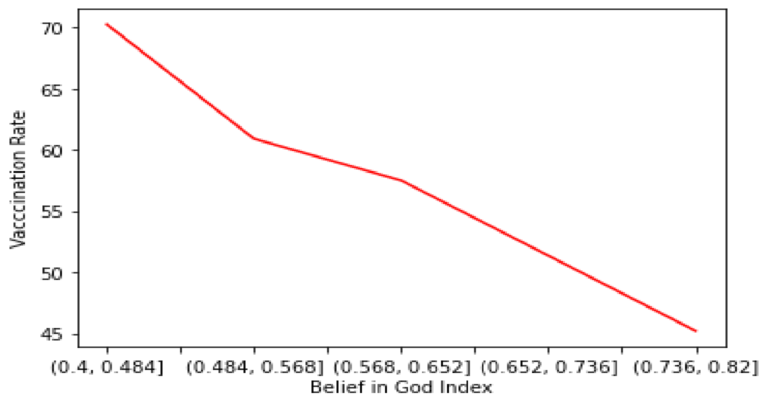

Figure 6

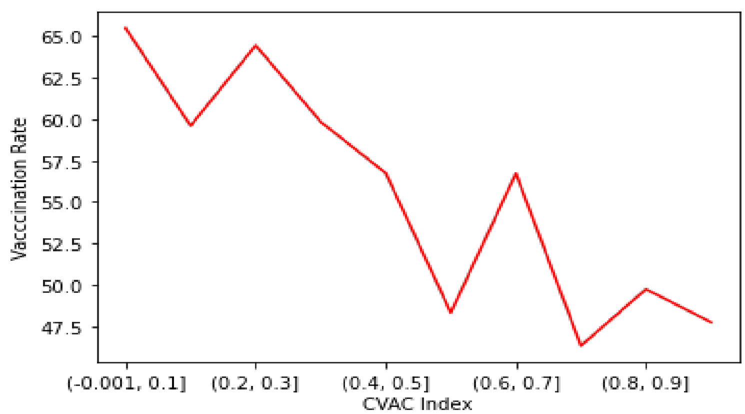

Figure 7

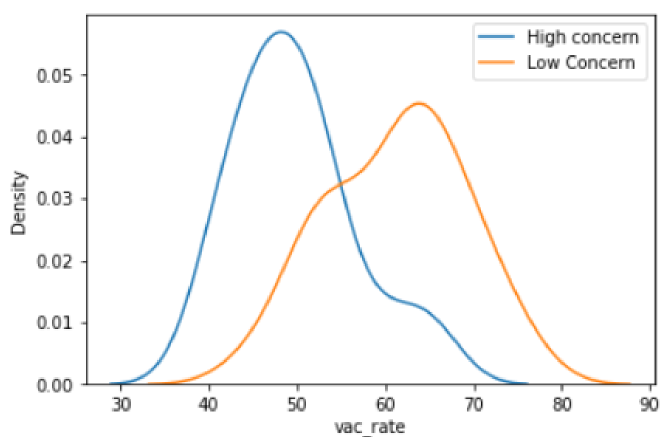

Figure 8

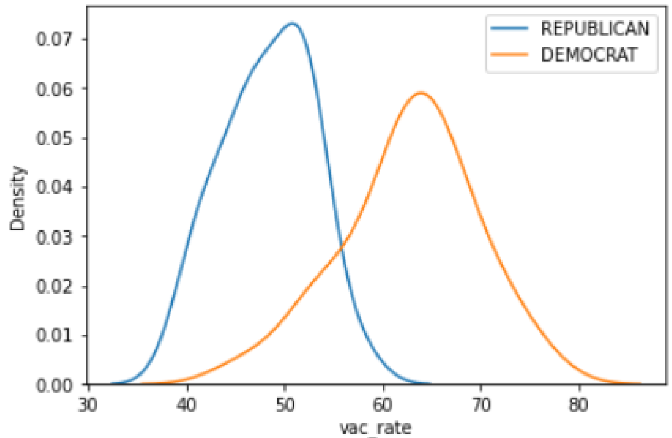

Figure 9

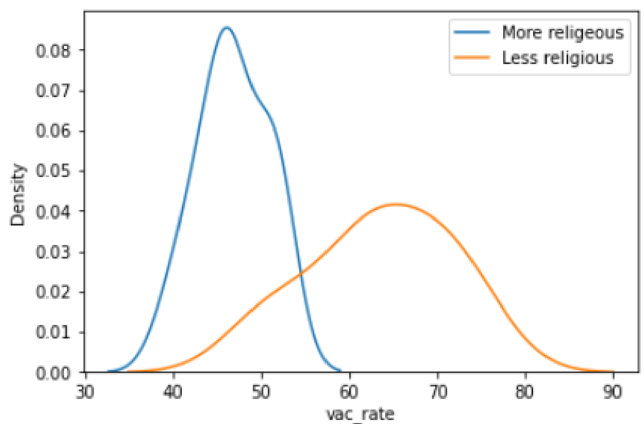

Figure 10

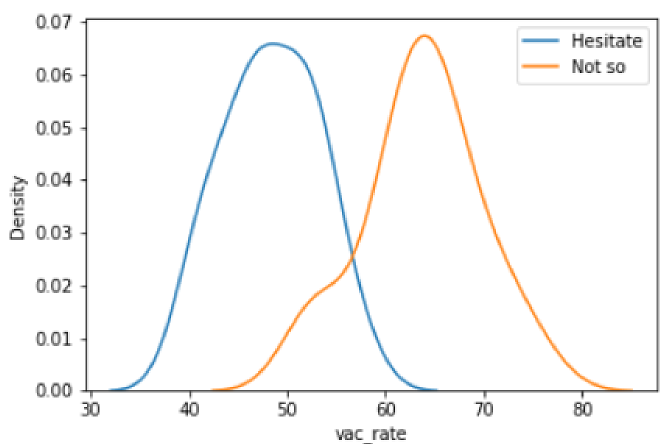

Figure 11

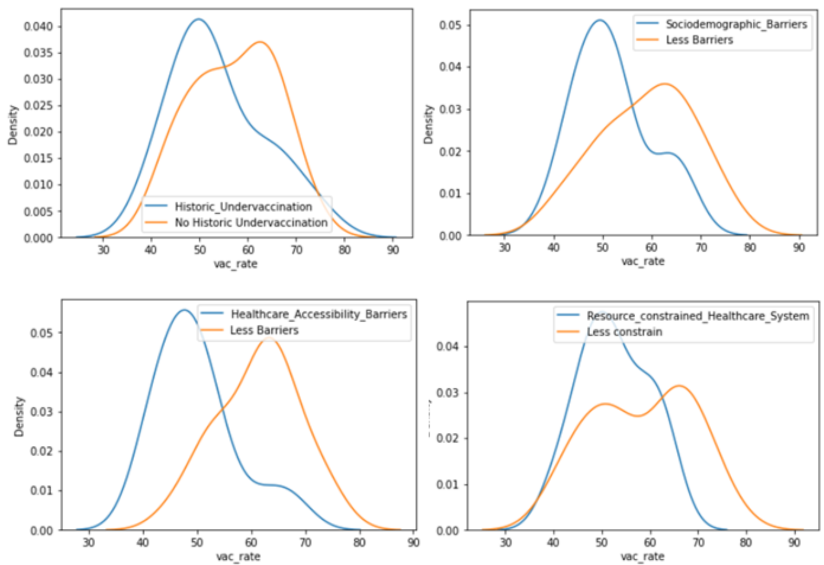

Figure 12

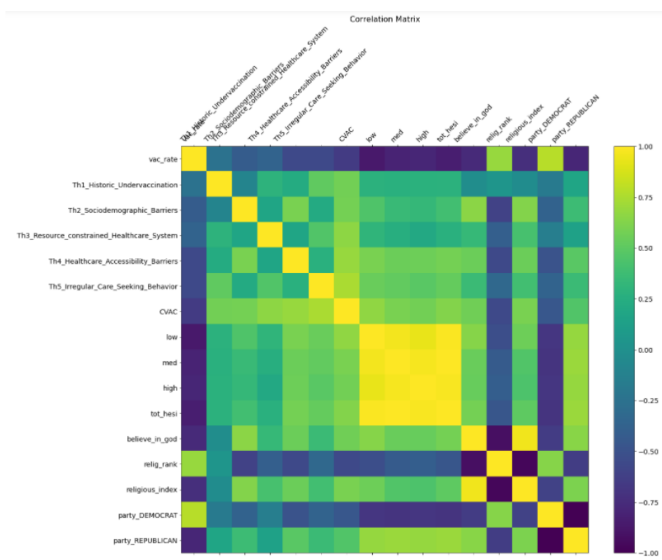

Figure 13

1.

OLS Regression Results

Dep. Variable:	vac_rate	R-squared:	0.424
Model:	OLS	Adj. R-squared:	0.412
Method:	Least Squares	F-statistic:	35.39
Date:	Thu, 05 Aug 2021	Prob (F-statistic):	3.00e-07
Time:	10:03:23	Log-Likelihood:	-167.89
No. Observations:	50	AIC:	339.8
Df Residuals:	48	BIC:	343.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	65.8749	2.012	32.736	0.000	61.829	69.921
CVAC	-20.4754	3.442	-5.949	0.000	-27.396	-13.555

Omnibus:	0.010	Durbin-Watson:	1.527
Prob(Omnibus):	0.995	Jarque-Bera (JB):	0.157
Skew:	-0.019	Prob(JB):	0.925
Kurtosis:	2.729	Cond. No.	4.37

4.

OLS Regression Results

Dep. Variable:	vac_rate	R-squared:	0.894
Model:	OLS	Adj. R-squared:	0.863
Method:	Least Squares	F-statistic:	29.10
Date:	Fri, 06 Aug 2021	Prob (F-statistic):	3.80e-15
Time:	21:18:56	Log-Likelihood:	-125.62
No. Observations:	50	AIC:	275.2
Df Residuals:	38	BIC:	298.2
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	82.1374	6.140	13.377	0.000	69.707	94.568
Q("Th1_Historic_Undervaccination")	0.9836	2.412	0.408	0.686	-3.900	5.867
Q("Th2_Sociodemographic_Barriers")	5.5020	2.514	2.188	0.035	0.412	10.592
Q("Th3_Resource_constrained_Healthcare_System")	-0.8880	2.194	-0.405	0.688	-5.330	3.554
Q("Th4_Healthcare_Accessibility_Barriers")	-0.3354	2.669	-0.126	0.901	-5.739	5.068
Q("Th5_Irregular_Care_Seeking_Behavior")	-3.2998	2.523	-1.308	0.199	-8.407	1.807
Q("low")	-150.9610	41.412	-3.645	0.001	-234.794	-67.128
Q("med")	119.3730	70.823	1.686	0.100	-24.001	262.747
Q("high")	-49.8301	68.129	-0.731	0.469	-187.749	88.089
Q("religious_index")	-14.9315	20.129	-0.742	0.463	-55.681	25.818
Q("believe_in_god")	-11.1835	22.100	-0.506	0.616	-55.922	33.555
Q("party_DEMOCRAT")	5.5283	1.667	3.317	0.002	2.155	8.902

Omnibus:	1.459	Durbin-Watson:	2.117
Prob(Omnibus):	0.482	Jarque-Bera (JB):	1.316
Skew:	-0.251	Prob(JB):	0.518
Kurtosis:	2.384	Cond. No.	357.

2.

OLS Regression Results

Dep. Variable:	vac_rate	R-squared:	0.723
Model:	OLS	Adj. R-squared:	0.712
Method:	Least Squares	F-statistic:	61.45
Date:	Thu, 05 Aug 2021	Prob (F-statistic):	7.67e-14
Time:	10:03:44	Log-Likelihood:	-149.57
No. Observations:	50	AIC:	305.1
Df Residuals:	47	BIC:	310.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	75.4866	1.951	38.693	0.000	71.562	79.411
CVAC	-6.5703	3.102	-2.118	0.039	-12.810	-0.330
tot_hesi	-46.6893	6.551	-7.127	0.000	-59.868	-33.511

Omnibus:	1.643	Durbin-Watson:	1.975
Prob(Omnibus):	0.440	Jarque-Bera (JB):	1.144
Skew:	-0.035	Prob(JB):	0.564
Kurtosis:	2.262	Cond. No.	11.8

3.

OLS Regression Results

Dep. Variable:	vac_rate	R-squared:	0.799
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	61.04
Date:	Thu, 05 Aug 2021	Prob (F-statistic):	4.53e-16
Time:	15:31:25	Log-Likelihood:	-141.56
No. Observations:	50	AIC:	291.1
Df Residuals:	46	BIC:	298.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	65.5740	2.911	22.524	0.000	59.714	71.434
CVAC	-6.1621	2.673	-2.305	0.026	-11.542	-0.782
tot_hesi	-29.3764	7.005	-4.194	0.000	-43.476	-15.276
party_DEMOCRAT	7.0568	1.693	4.169	0.000	3.650	10.464

Omnibus:	0.849	Durbin-Watson:	2.543
Prob(Omnibus):	0.654	Jarque-Bera (JB):	0.891
Skew:	0.174	Prob(JB):	0.640
Kurtosis:	2.446	Cond. No.	16.1

Figure 14 (4 images labeled 1-4)

OLS Regression Results

Dep. Variable:	vac_rate	R-squared:	0.304
Model:	OLS	Adj. R-squared:	0.302
Method:	Least Squares	F-statistic:	149.1
Date:	Sun, 08 Aug 2021	Prob (F-statistic):	1.24e-208
Time:	22:20:07	Log-Likelihood:	-10709.
No. Observations:	2744	AIC:	2.144e+04
Df Residuals:	2735	BIC:	2.149e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.9380	2.160	33.312	0.000	67.704	76.173
Q("Th1_Historic_Undervaccination")	3.1128	0.897	3.469	0.001	1.353	4.872
Q("Th2_Sociodemographic_Barriers")	-6.4606	1.164	-5.553	0.000	-8.742	-4.179
Q("Th3_Resource_constrained_Healthcare_System")	-8.8037	0.897	-9.812	0.000	-10.563	-7.044
Q("Th4_Healthcare_Accessibility_Barriers")	-2.9567	1.285	-2.300	0.022	-5.477	-0.436
Q("Th5_Irregular_Care_Seeking_Behavior")	-3.6186	0.982	-3.683	0.000	-5.545	-1.692
Q("tot_hesi")	14.9817	2.310	6.487	0.000	10.453	19.510
Q("party")	12.0066	0.648	18.524	0.000	10.736	13.278
Q("believe_in_god")	-47.2805	3.851	-12.277	0.000	-54.832	-39.729

Omnibus:	234.423	Durbin-Watson:	1.119
Prob(Omnibus):	0.000	Jarque-Bera (JB):	419.064
Skew:	-0.599	Prob(JB):	1.00e-91
Kurtosis:	4.493	Cond. No.	33.1

Figure 15

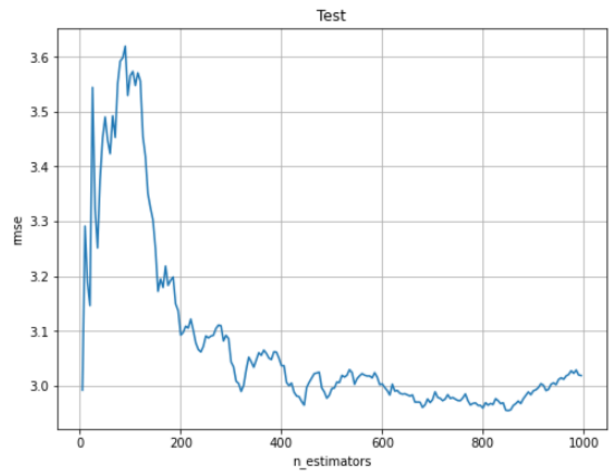

Figure 16

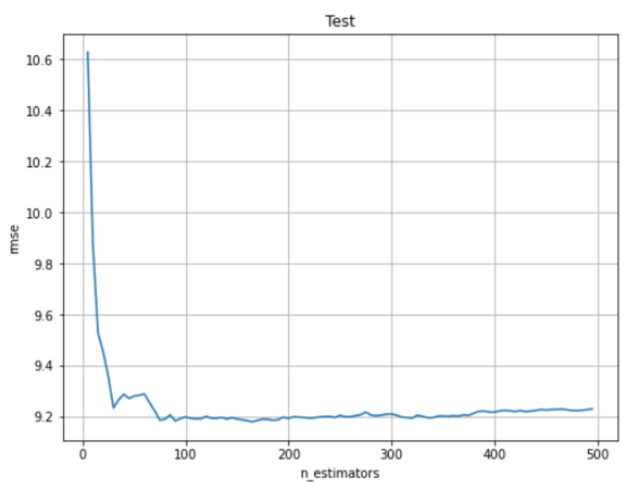

Figure 17

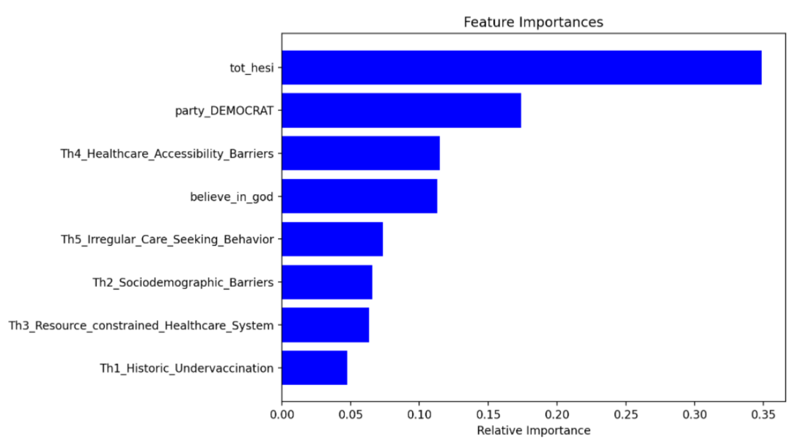

Figure 18

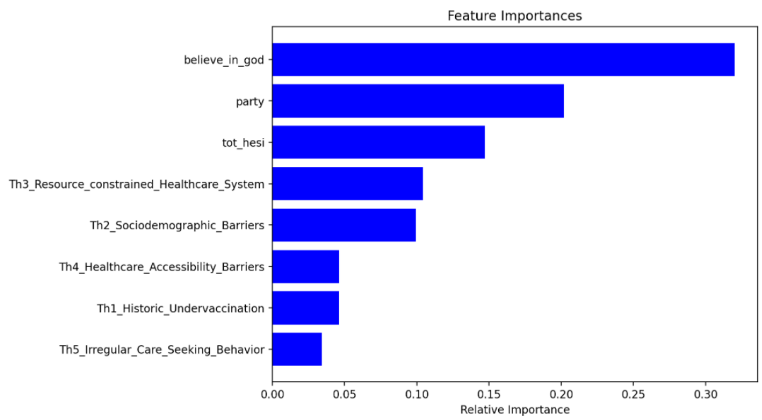

Figure 19

Herd Immunity Information:
https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html