Covid-19 Vaccine Compliance Based on Political Partisanship and Forcasted Machine Models

Raine Brookshire - raine.brookshire.25@dartmouth.edu

November 2023

Supervised by Dr. Mia Costa, introduced by Keren Luo and accompanied by Leila Ambrus, Cade Haskins, Michael Hanrahan, Aniket Dey, and Alexander Wojcik

Abstract

The escalating tide of partisanship that has swept through our nation in recent decades has not been an accident. According to the Inlander news site, these are the facts; partisanship breaks down our sense of community by partitioning us into smaller, warring tribes which ameliorates partisanship's capacity to build power. Government, at its core, is best when individuals come together to solve problems. Some aspects of our collective problem solving came about from Covid-19, a major exogenous shock that caused great human suffering and daunting challenges for the entire globe. The onset of Covid brought a set of medical and governance issues that political parties allied to quickly react to extinguish the chaos. This study is motivated by these stylized facts and hopes to illuminate aspects of partisanship based on Covid mandates, vaccines and policies. Additionally, there will be room to expand on the future implications of partisanship on covid policies and individual willingness to take the vaccine.

Keywords – Covid 19, Statistical Analysis, Forecasting, Predictive Model, Machine Learning, Ideology, Political Partisanship, Regression Model

Significance

After programming the full survey questionnaire successfully fielded on a national sample, we examined the role of political partisanship on compliance to take the covid vaccine. We took into accout differing treatment conditions where we looked to decipher how the independent variables of political affiliation and provided health measure influence the outcome variable of an individual's compliance/ willingness to receive the COVID-19 vaccine. In this study there were 6 main treatment variables. These variables contained vignettes of politically involved CDC directors comparing Republican vs Democrat vs No Party endorsement methods and Vaccine vs Hygiene health measures which where meant to be compared to a control group (No vignette being shown) in order to make predictions and access patterns in regard to individual willingness to receive the vaccine in our post-COVID era. This study highlights the nuanced role of how political partisanship influences engagement in a compliance to take the vaccine and adhere to covid health measures.

Contents

1 Introduction and Methodology **4**

1.1 Variables and Conditions . . . 4

1.2 Hypotheses and Mechanisms . . . 5

2 Results **6**

2.1 Interesting Figures and Statistical Models . . . 7

3 3D models and Machine Learning **9**

4 Conclusions, Implications and Limitations **14**

List of Figures

1 Cleveland dot plot for polarization of compliance based on treatment and control group. It is clear that the prediction was correct. In this visualization, after I ran a paired t-test between values, it was clear that significant polarization and partisanship was present. . . . 6

2 3D plotted graph in python revealing polarization between democrats and republicans. Based off of figure 1 but adding another dimension. . . . 7

3 Dot plot assessing the influence of the treatment variables based on a person's likelihood to take COVID-19 vaccine (based on their compliance level). Hypotheses predicted more significant changes for Democrat and vaccine conditions. In this case for the treatment variables, the alternate hypothesis expected that (compliance of treatments) $mu1 > mu2$ (compliance of control). . . . 8

4 2D heatmap plot showing relationships between people who tend not to be willing to take the COVID-19 vaccine because they don't value the importance of health measurea vs people who have high compliance for taking the covid vaccine likely due to the fact that they place importance on serious/ strict covid health measures. . . . 10

5 2D Kernel density plot which shows relationships between individuals who engage less in CDC health measures and whether they are less likely to receive a vaccine based on their personal interest. Based the ranking method, the value 6 (x-axis) represents the idea that the **most** effective health measure is to take the New vaccination (2023-2024 formula) while 1 (x-axis) represents the method of social distancing. . . . 10

6 Linear regression model with confidence intervals representing the correlation of the independent age variable and dependent compliance variable. . . . 11

7 Complexity graph to show how an optimization of degrees produces the best R-squared value (or line of best fit). . . . 11

8 Finalized machine learning model after testing around 20-30 percent of the data. Model was improved and shows more of an exponential function based on 2 variables. . . . 11

9 Graph based on Desmos spatial graphing calculator to construct a surface of best fit for the data (figure 3.5 graph rotated from the top counter clockwise). This plot was eyeballed and is an estimation, of a potential model in Python. The equation to the line is $z = -0.2x \cdot 2.3 - 0.05y \cdot 2.3 + 3.9$ and shows the clear decline in z axis from Republicans to Democrats based on $x = y$. In this case, x is the ranking variable, and y is the level of compliance to get the vaccine. . . . 12

List of Tables

1 The Conditions and predicted mechanisms for the predictors (X axis: mean compliance to take vaccine) and response variables(Y axis:conditions/ mechanisms of interest) . . . 5
2 Statistical test values for cleveland plot based on the difference in means for each Group . . . 7

1 Introduction and Methodology

How do national identifications and party support affect compliance with COVID-19 CDC public policies? Before researching questions, the team pondered on impactful historic effects in recent years. With the entire world sheltered and crippled by human isolation, the spread of COVID-19 pervaded through the streets. The virus has overtaken other serious outbreaks like Ebola, tuberculosis and measles. The death toll is around 30 million and while many accredited sources have documented data in regard to the CDC and the political implications of the COVID-19 virus and the vaccine, most of the studies have been performed during the onset of this virus and where completed as a means of seeking mitigation methods. As one of our avenues of research, we will expound upon how treatment effects based on party affiliation influence the Democrat and Republic mind on their willingness to comply with COVID-19 vaccine mandates. Based on these goals, we first researched design methods to investigate a causal mechanis which will further be revealed at 1.1.

1.1 Variables and Conditions

1. First we formulated a Hypothesis question
2. Planned methods to fully comprehend our variables and conditions of interest
3. Decided on our experiment structure based on either a field experiment, quasi experiment, survey experiment, natural experiment and lab experiment
   - We chose the survey experiment
   - We chose to survey 1500 individuals
   - Created vignettes based on political affiliations and a baseline vignette that represented no political affiliation
4. Ran a survey on a smaller sample size to perform a power analysis.
5. Submitted documents to the IRB, submitted a PAP and then fielded the survey
6. Went through pre-analysis, manulation checks, and debreifing
7. Ran data analysis on treatment variables and I created stellar visualizations based on variables of interest and other interesting mechanism we'd like to study. (Note: Individuals who failed the attenion check were discarded from the survey and csv dataset altogether.)

Variables
1 = DemVaccine
2 = RepVaccine
3 = NoPartyVaccine
4 = DemHygiene
5 = RepHygiene
6 = NoPartyHygiene
7 = control group (no vignette shown)

Understanding these variables is key but my table will later properly highlight the importance of how these values were interpreted based on my predictions. I used methods based on data analysis techniques in python, D3 and MATLAB and assumed that ranking1 was based on the first choice of order out of the 6 choices in the results csv. Using the ranking variable (ranking1 in the csv) was helpful to other hypotheses I formulated and assisted in assessing personal judgments and biases. Interestingly, I didn't use credibility of the health advisor

Conditions	Predicted Mechanism for Republicans	Predicted Mechanisms for Democrats
Democrat + Vaccine	Lower than control variable	Slightly greater than control variable
Republican + Vaccine	Greater than control variable	Slightly greater than control variable
No Party + Vaccine	Greater than control variable	Slightly greater than control Variable
Democrat + Hygiene	Slightly lower than control variable	Around control variable
Republican + Hygiene	Slightly greater than control variable	Around control variable
No Party + Hygiene	Slightly greater than control variable	Slightly Greater than Average
Control Group	Baseline	Baseline

Table 1: The Conditions and predicted mechanisms for the predictors (X axis: mean compliance to take vaccine) and response variables(Y axis:conditions/ mechanisms of interest)

(column in dataset) due to there being no (neutral vignette) control (however I may have likely used a a paired t-test to access the difference in means of Democrats and Republicans based on the variables). Additionally, I flipped the ranking variables with their opposite number (1:6 to 6:1 for ——New vaccination (2023-2024 formula) (1)) because having a higher number should mean a person believes stricter covid mandates are more important/effective.

1.2 Hypotheses and Mechanisms

- For the **major** null hypothesis on partisanship. H0 indicates that there is no difference in means (for compliance to take vaccine) between Democrats and Republicans
  - In this case, we assume that for H0, mu1 (mean value for treatment variables) == mu2 (mean value for control groups)
  - For the alternate Hypothesis: Ha we have that mu1=/= m1 indicating that the difference is significant or that there is a significant difference in perceptions on the issue of covid where Republicans are less likely to comply and Democrats are more likely to adhere to COVID-19 mandates. The mechanism at play could be due to the fact that Republicans usually value freedom of body while Democrats value health mandates (free health for example) and safety of the people.
- For the minor null hypothesis. I have H0 indicates that there is no difference in means (for compliance to take vaccine) for **total averaged** values across the population.
- In the alternate hypothesis I expected values for individuals in the Democrat and vaccine categories to usually be higher than the control while individuals who where in the Republican and hygiene category would be be lower than the mean.
- An additional hypotheses of interest is based on the idea that as the age of individuals increase, there will be more of a willingness to comply. This is based on the causal mechanism that elders are more likely to get covid.
- Later I documented that I would also expect individuals who have high compliance to likely be democrats (vice versa with republicans) but as an additional variable I assumed that these individuals would likely have a high ranking variable because they believe in the effectiveness of strict covid mandates.

Based on these mechanisms I was interested in the partisanship and its effect for both Democrats and Republicans. Diving deeper into the mechanisms at play and the expected

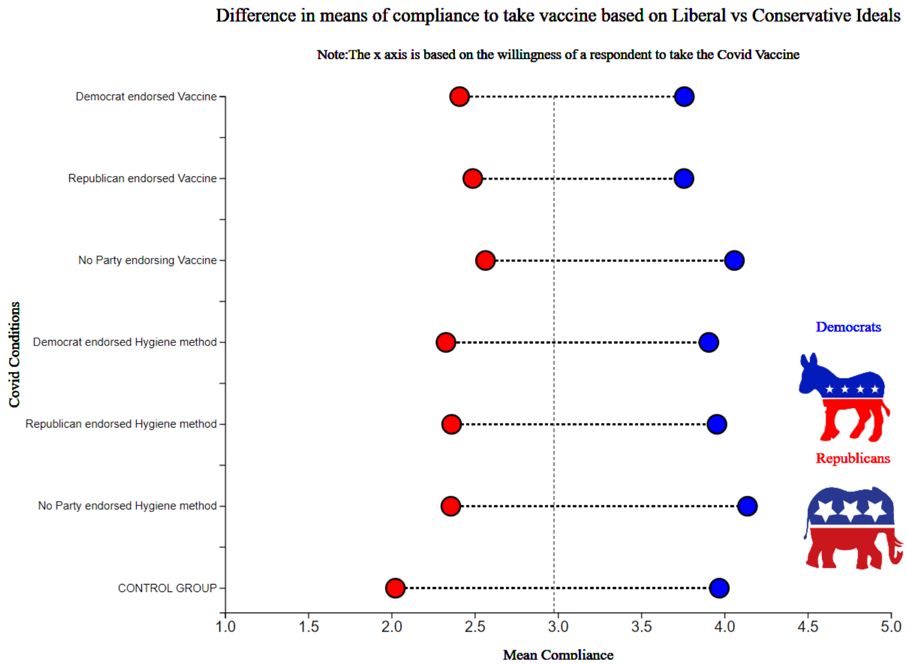

Figure 1: Cleveland dot plot for polarization of compliance based on treatment and control group. It is clear that the prediction was correct. In this visualization, after I ran a paired t-test between values, it was clear that significant polarization and partisanship was present.

results, I first hypothesized that Democrats would mostly be consistent in their tendency to comply regardless of the treatment variables. I also imagined that Republicans would only see an increase in their compliance when the treatment variable is endorsed by another Republican affiliate. Finally, I hypothesized that both groups would be more likely to comply when there is no party affiliation in the vignette because there would be no bias or no political ideal to support.

2 Results

After producing significance t-tests and comparing the mean values based on a difference of means where $\tau(i) = Y_i(1) - Y_i(0)$, I was looking to first see whether the polarization between Democrats and Republicans was significant. To do this, I wanted to calculate mean values across all individuals and filter **out** all respondents who were **not** Republican or Democrat in the political ideology column (pid5). After running one-tailed t-tests on both Republicans and Democrats, the results were interesting.

I later ran a 2 sample paired t -test to determine the validity of my hypothesis based on if the difference between the two parties was statistically significant or based on chance. In python I received a t-statistic of 17.748 giving a p-value of $2.055^{-}6$ indicating an extremely low chance

Condition	Group	T-statistic	P-value
RepVaccine-2	Republican	2.606	0.009
NoPartyVaccine-3	Republican	2.838	0.004
DemPartyVaccine-5	Republican	1.877	0.06
NopartyHygiene-6	Democrat	1.710	0.089
DemPartyHygien-4	Democrat	-0.404	0.260

Table 2: Statistical test values for cleveland plot based on the difference in means for each Group

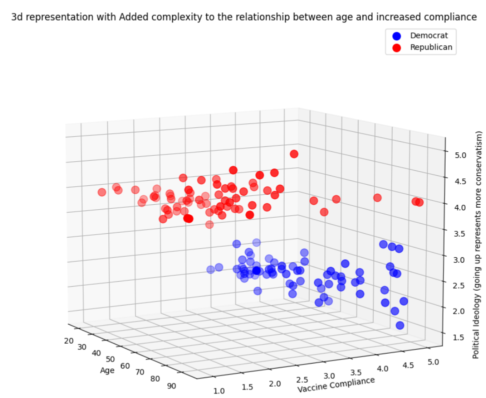

Figure 2: 3D plotted graph in python revealing polarization between democrats and republicans. Based off of figure 1 but adding another dimension.

that the difference in the data is based on chance alone. This supports my alternative hypothesis and suggests that a significant difference between the Democrat and Republican mind on compliance to get the covid vaccine is apparent.

After looking further at some of the data in python, the data immediately captivated my attention due to its deviation from a Gaussian distribution, which is the expected pattern based on the Central Limit Theorem. However, a closer examination revealed a compelling narrative: the data is sharply divided along partisan lines, with Democrats and Republicans exhibiting distinct patterns. This partisan divide is so pronounced that calculating an average across both groups would yield a value close to the control group's mean, masking the underlying partisan disparities. Based on this understanding, we will see how the data highlights the partisan divide (Figure 1), and explains how the partisan split could lead to misleading conclusions if not accounted for.

2.1 Interesting Figures and Statistical Models

Interestingly, when the split is not accounted for and we are solely looking at the averages of all respondents, although I predicted there would be in increase when the vignettes where shown (more so for democrats than republicans) calculating the p-value based on the mean

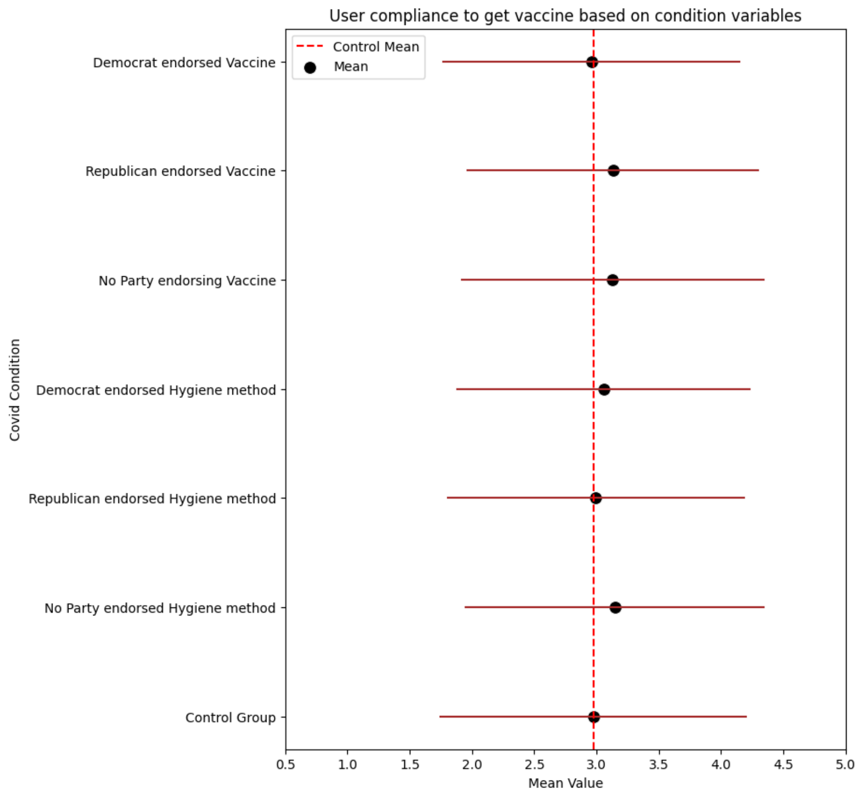

Figure 3: Dot plot assessing the influence of the treatment variables based on a person's likelihood to take COVID-19 vaccine (based on their compliance level). Hypotheses predicted more significant changes for Democrat and vaccine conditions. In this case for the treatment variables, the alternate hypothesis expected that (compliance of treatments) $mu1 > mu2$ (compliance of control).

difference in values gave me p-values ranging from 0.116 -0.8446. The difference in means ranged from -1.5726 - 0.156 which resulted in none of the treatment variables resulting in a significant difference with the control mean. Based on this result, with a p-value set with a significance level of 0.05, I failed to reject the null hypothesis that a person seeing a vignette would **not** result in an increased mean. Due to this fact, does it mean that seeing a vignette doesn't increase a persons willingness to comply but rather reaffirms their prior political inclination to either comply or not ?

There will always be pressing questions but in the case of this experiment, it is clear that the partisanship is present. From this it would be interesting to see whether this divide has worsened or improved over the years. In the end, the differences in the mean showed there was no significant increase of the values based on a one-tailed- t-test. P-values were not within the thresh-hold of $\alpha < 0.05$ for the average compliance score for all individuals but for mean democrat and republican compliance scores. This information signifies that there is a significant difference between the mean of the two groups indicating that Democrats are more likely to comply with vaccine regulations while Republicans are less likely to comply to receive a vaccine.

3 3D models and Machine Learning

On the topic of compliance variables, I was also interested to see if individuals are more likely to comply as they age.
With a 95 percent confidence level, I assessed that there was statistical evidence supporting an increase in the dependent variable (y: compliance) based on both the linear and quadratic terms (Figure 6 and 7), The p-values for x1 and x2 are below the 0.05 significance level, suggesting that these terms are statistically significant in predicting the dependent variable.

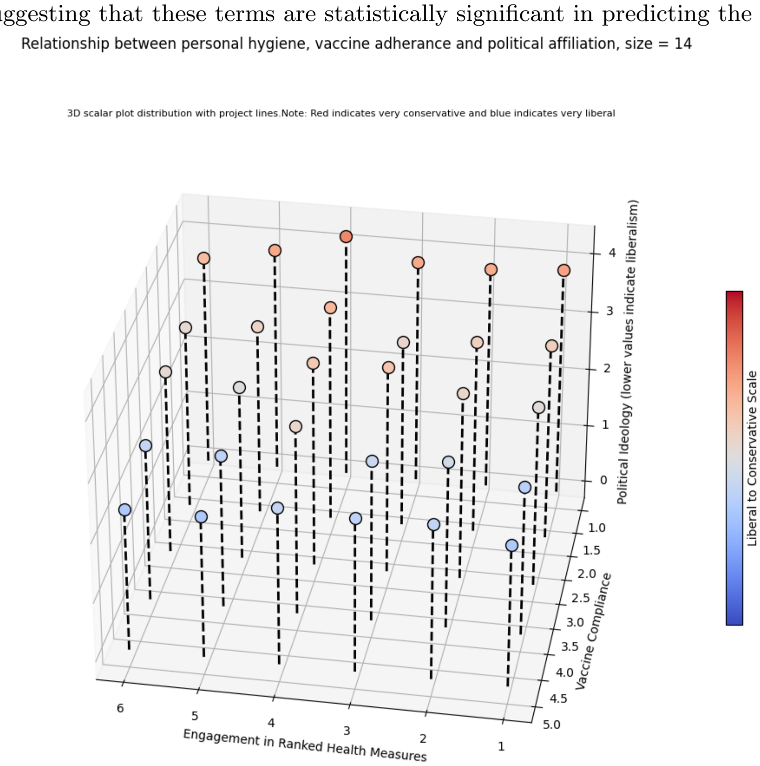

Figure 3.5: Graph based in python structurally modeling the relationship between political ideology, ranked importance of health measures and compliance. We see an interestingly increasing relationship on the z axis indicating that whenindividuals are more democrat leaning, they are more likely to comply with mandates compared to when individuals are republican which makes them less likely to comply. This graph is based on the application of figures 4 and 5 as I suspected that individuals on the top right were republicans and individuals on the bottom left were democrats (further expounding on the polarization)

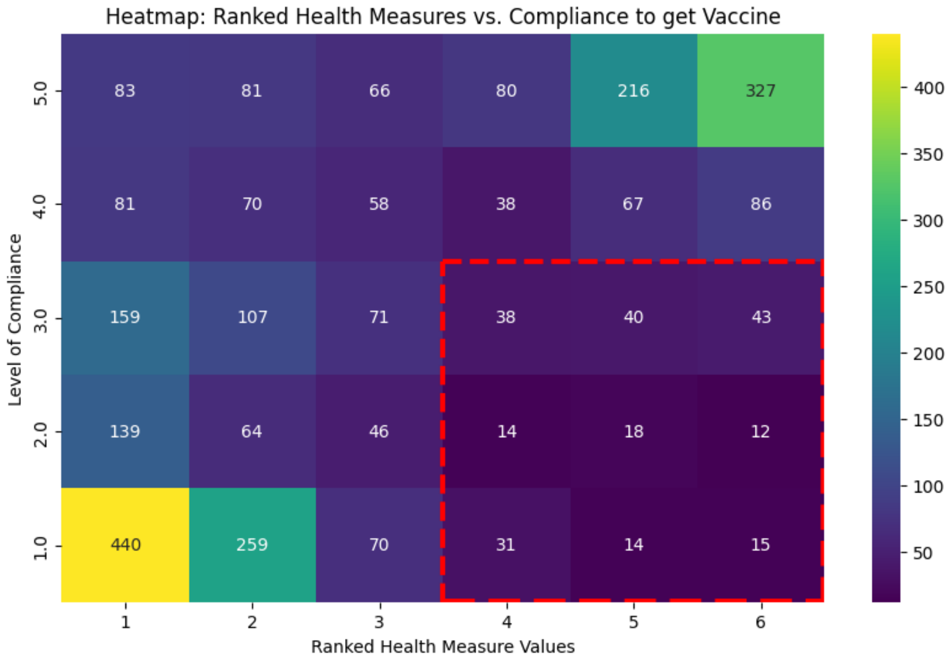

Figure 4: 2D heatmap plot showing relationships between people who tend not to be willing to take the COVID-19 vaccine because they don't value the importance of health measurea vs people who have high compliance for taking the covid vaccine likely due to the fact that they place importance on serious/ strict covid health measures.

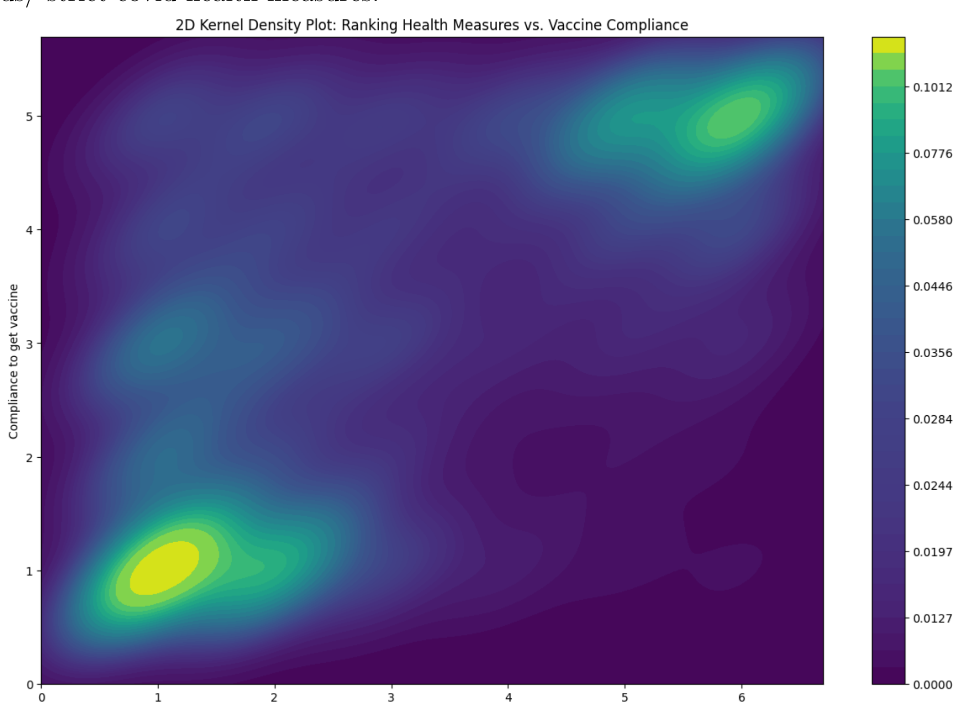

Figure 5: 2D Kernel density plot which shows relationships between individuals who engage less in CDC health measures and whether they are less likely to receive a vaccine based on their personal interest. Based the ranking method, the value 6 (x-axis) represents the idea that the **most** effective health measure is to take the New vaccination (2023-2024 formula) while 1 (x-axis) represents the method of social distancing.

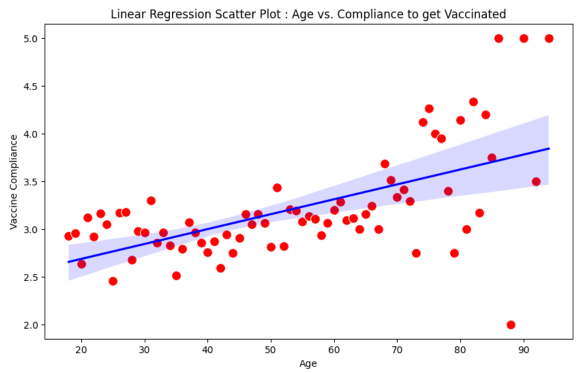

Figure 6: Linear regression model with confidence intervals representing the correlation of the independent age variable and dependent compliance variable.

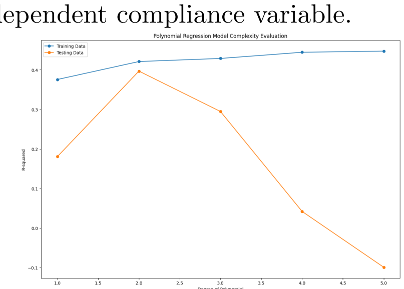

Figure 7: Complexity graph to show how an optimization of degrees produces the best R-squared value (or line of best fit).

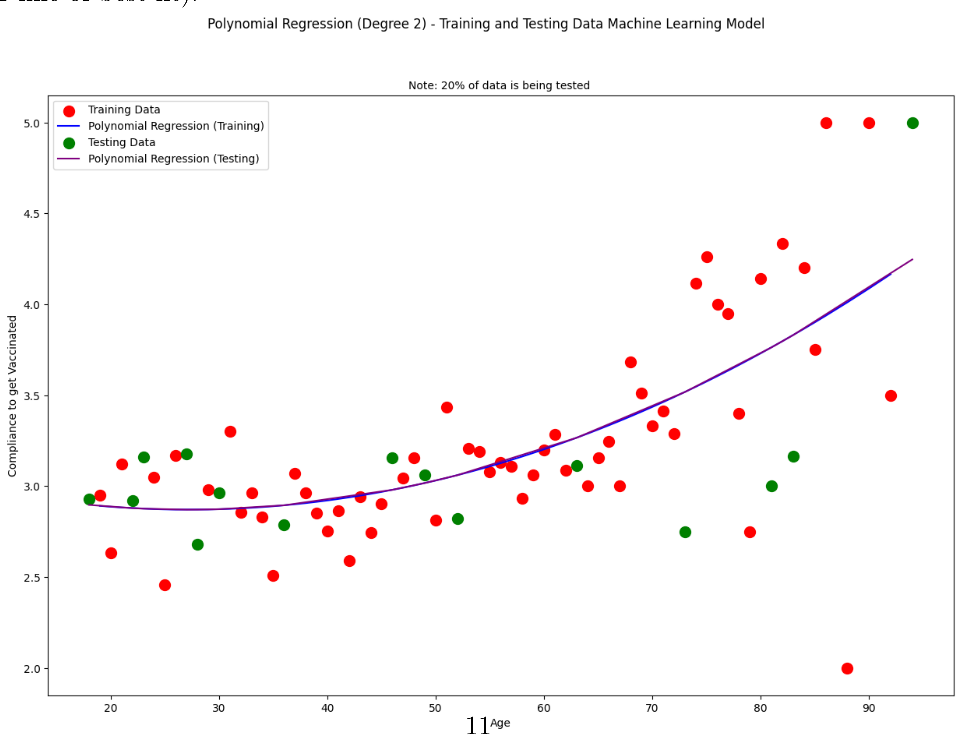

Figure 8: Finalized machine learning model after testing around 20-30 percent of the data. Model was improved and shows more of an exponential function based on 2 variables.

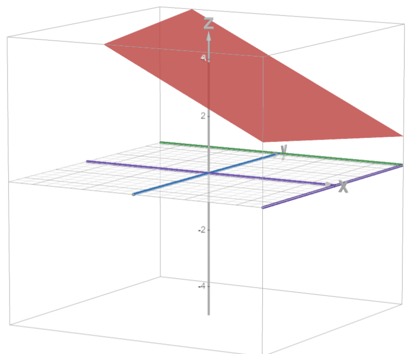

Figure 9: Graph based on Desmos spatial graphing calculator to construct a surface of best fit for the data (figure 3.5 graph rotated from the top counter clockwise). This plot was eyeballed and is an estimation, of a potential model in Python. The equation to the line is $z = -0.2x \cdot 2.3 - 0.05y \cdot 2.3 + 3.9$ and shows the clear decline in z axis from Republicans to Democrats based on $x = y$. In this case, x is the ranking variable, and y is the level of compliance to get the vaccine.

First, after the 3d model, based on figure 3 and 4, it inspired me to find a surface describing the change in data. No statistical analysis was involved but it was interesting to see how the variables z: political ideology, x: ranked effectiveness of health measures and y: level of compliance are related. Going back to my linear regression model (Figure 8), initially I had

an R-squared value of 0.12 for the linear model but after splitting the data to allocate the data into training and testing categories I was able to improve my model to a an R-squared value of 0.42. Later on I over fit my data and almost had a quadratic equation when I had too many predictor variables. Eventually, I had found a sweet spot to provide the most accurate model in figure 8 and received an R-squared value of 0.465. Based on this situation of too many predictors where we have the regression model $\beta_i = E[Y|X_1, \ldots, X_{i-1}, X_i = x_i + 1, X_{i+1}, \ldots, X_K] - E[Y|X_1, \ldots, X_{i-1}, X_i = x_i, X_{i+1}, \ldots, X_K]$, we may be obtaining more related values but I ended up overestimating my data when applying more variables. Therefore, I had to either reduce the degree or change the percentage of tested and trained data.These findings are riveting and further show relationships between variables of interest and the entire idea of political ideologies. After these changes I had a chance to calculate the correlation coefficient: 0.59, p-values: 0.003 and R-squared: 0.45. I had an RMSE of around 0.4 but based on the p-values, there seems to be much support that the age of an individual is positively correlated to the level of compliance for covid vaccines.

To discuss accurate models and enhance the study's methodology, on the survey, we could have considered incorporating a slider scale for participants to indicate their level of agreement or disagreement with statements about vaccines rather than explicitly asking if they agree with a statement in regard to vaccine effectiveness. Additionally, including both a neutral control group and a no-vignette shown group would provide a baseline comparison for the vaccine group. A scatter plot would be a more suitable graphical representation of the data given if the slider was implemented since there would be a non-discrete nature of the variables. While race based on Asian discrimination (due to the virus previously being labeled as the Chinese virus) could be a potential factor influencing vaccine perceptions, analyzing aspects of racial patterns may be more complex yet exciting.

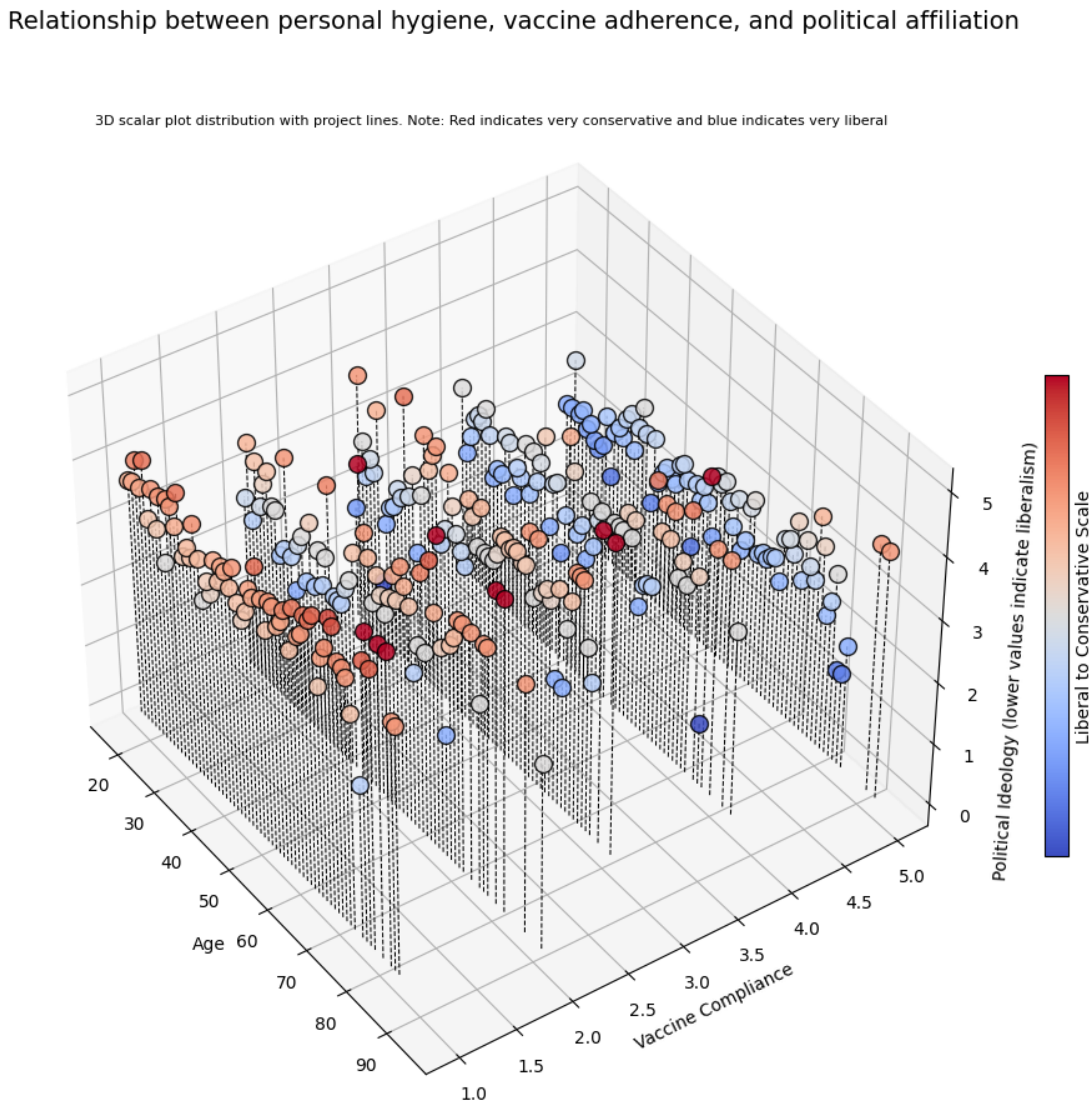

Figure 10: 3D scatter plot with projected lines to emphasize previous relationship between age an an individuals political ideology. z axis represents the level of political ideology. The 2d kernel plot can both be seen in some regard showing an intruiging relationship with the 3 variables (x,y,z).(Note:No statistical test was applied).

4 Conclusions, Implications and Limitations

After modelling the mother of all visualizations (figure 9), I marveled at the elegant layering of the data, revealing its evolutionary patterns in relation to another variable. Essentially, I had constructed a foundational summary of the data, which, if refined and calibrated using predictive models and machine learning techniques, or perhaps an accurate three-dimensional surface curve, is capable of predicting the data based on the three provided variables. This model could be further developed and utilized to create generalized models for other data or experiments.

To delve deeper into the intricate patterns within the data, several promising approaches can be employed. One intriguing technique involves constructing a unique neural network of points by connecting all z data points using breadth-first search algorithms (in python for example). This network can then be further enriched by applying Riemann sums to generate additional data points. Machine learning models, trained and tested on the data, can uncover the underlying relationships and produce a succinct equation for a best-fit surface that effectively represents the data. With these training models, we can evaluate the model's accuracy and assess its ability to explain the data. Additionally, incorporating all axes and variables could lead to the creation of a summarized approximation of the data, enabling the prediction of individual perceptions (based on age, political affiliation etc), basically allowing us to become a skilled geo-guessing expert.

Future research endeavors could explore the impact of contracting vs not contracting COVID-19 on an individual's willingness to get vaccinated by conducting logistic regression analysis. In the end, the conclusions based on the results are as follows. The partisanship and polarization between Democrats and Republicans is significant, the image vignettes had no significant impact on compliance scores for the averaged population. Some conditions such as RepVaccine and NoPartyVaccine were in support of my alternate hypothesis but for the most part, values saw no significant change which was expected for most democrats(as they are likely to be consistent with CDC guidelines regardless).

Finally as a summation of all my **points**, it is clear that increasing the number of parameters, or model complexity, can enhance a model's ability to interpolate any type of data perhaps based on future AI or machine learning algorithms. To end swiftly, I will give a famous quote that highlights how the number of predictors are essential to the accuracy of a model. It is said that "a four-parameter model can effectively fit an elephant, while a five-parameter model can even make the elephant wiggle its trunk" - Von Neumann.

References

1. LightOn (2020) Beyond overfitting and beyond silicon: The double descent curve, Medium. Available at: `https://medium.com/@LightOnIO/beyond-overfitting-and-beyond-silicon-the-double-descent-curve-18b6d9810e1b`
2. Drapala, J. (2023) Kernel density estimator for multidimensional data, Medium. Available at: `https://towardsdatascience.com/kernel-density-estimator-for-multidimensional-data-3e78c9779ed8`
3. Reuter, J.T. (2023) The rise of partisanship, Inlander. Available at: `https://www.inlander.com/comment/the-rise-of-partisanship-4131666`