Covid-19 Data 100 Final Project Final Report

Group Members: Harshaan Sall, Curtis Wong, Shan Virani

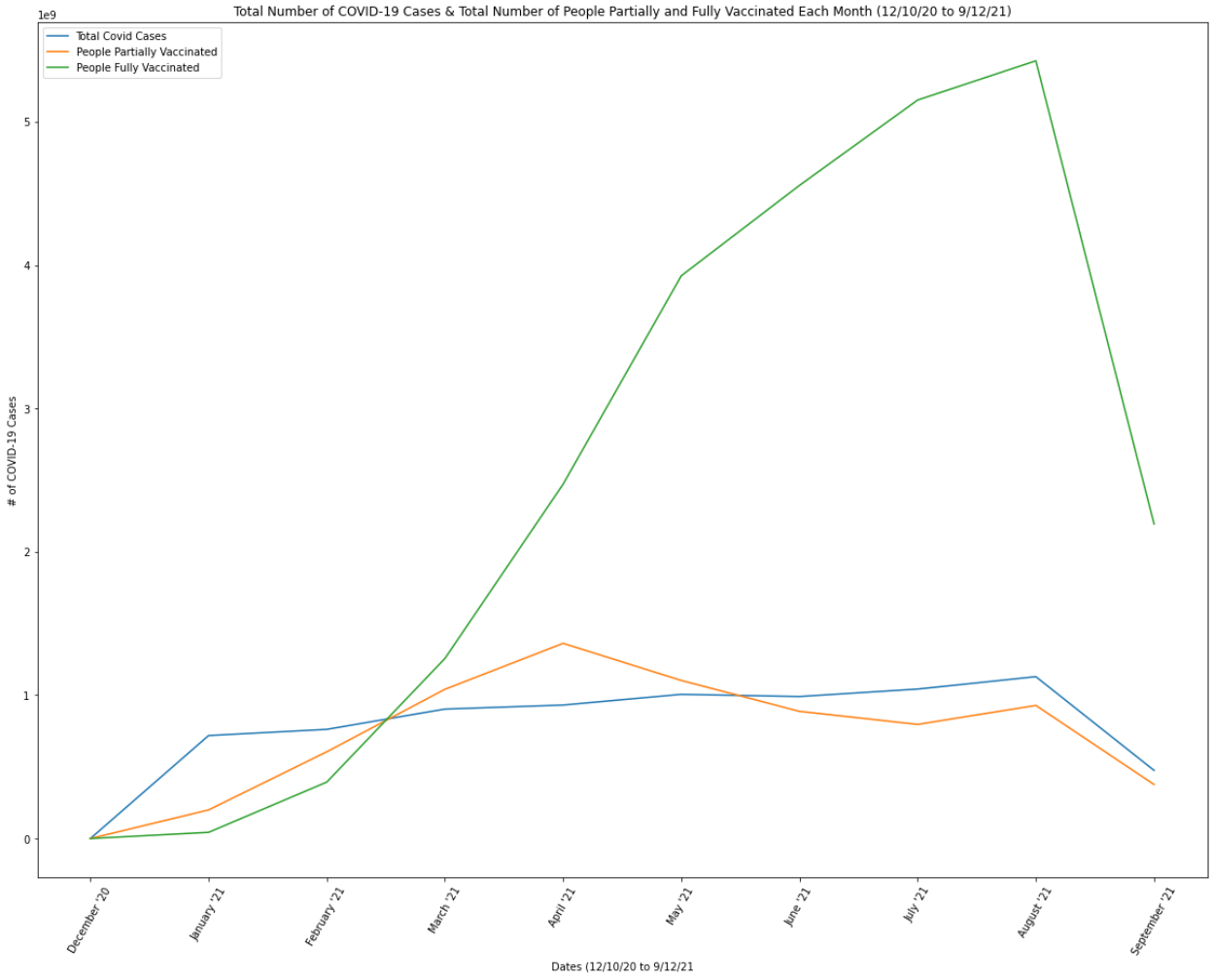
Overview

As mentioned in the Design Document for Part One of our final Project, one of the most important contemporary applications of Data Science is the Covid-19 pandemic. Given the importance of the pandemic and its impact on the planet, we can demonstrate the significance of Data Science as a discipline by analyzing the pandemic through the numerous data that has come as a result of the Covid-19 virus. There are many quantifiable metrics relating to the covid-19 pandemic that lend themselves to analysis using the tools and techniques of Data Science. In this project, we were given 4 datasets: cases, counties, vaccinations, and mask use. We believe another very important (arguably the most important) data relating to the pandemic was the number of deaths caused by the Covid-19 virus. Given the importance of this data, we decided to incorporate it into our hypothesis. Therefore, we imported Covid-19 deaths data to perform our analysis. Using these datasets, we have gone through the Data science lifecycle of asking a question, understanding the data, understanding the world, and reporting. Throughout this report, we will be discussing our hypothesis and corresponding modeling related to the data we decided to use!

Open-Ended EDA

One of the biggest questions surrounding the Covid-19 pandemic throughout 2020 was the development of a vaccine against the Covid-19 virus. Starting in Fall 2020, vaccines to protect against the Covid-19 virus were starting to be available to the public and once people started getting vaccinated, the biggest question was whether they were effective in preventing Covid-19 deaths. Given the importance of this question, our group was interested in vaccinations

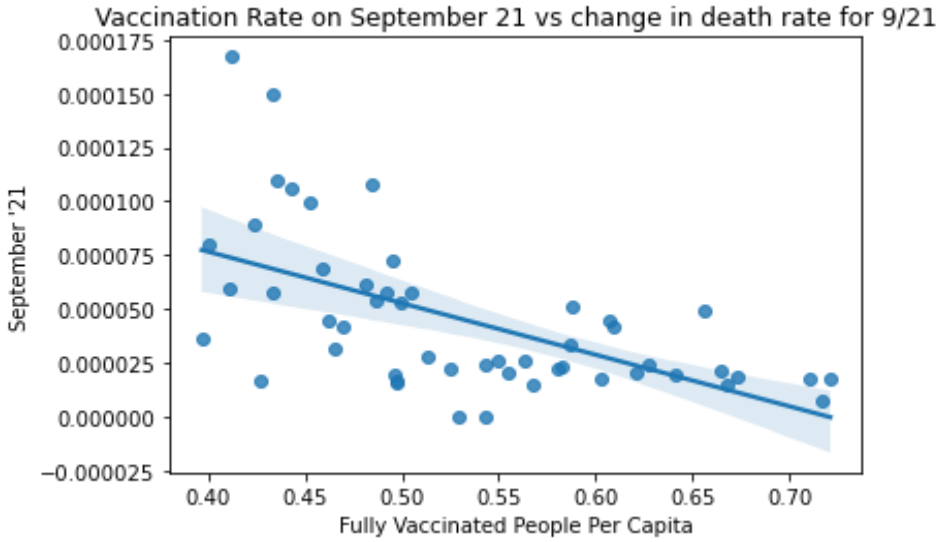
for our exploratory data analysis and we decided to create a visualization of the total number of vaccinations, total number of partial vaccinations, and the total number of new Covid-19 cases all as a function of time from December of 2020 to September of 2020.



Through this visualization, we do see an association between the increase of the number of vaccinated individuals and the flattening of the number of new Covid-19 cases. The green line shows the total number of vaccinated individuals. The orange line represents the total number of partially vaccinated individuals, and the blue line represents the number of new covid cases. As we can see, the blue line tends to flatten out and dip when the number of vaccinated individuals increases. This

visualization serves as a springboard for further analysis as we can create different visualizations and models to analyze the relationship between vaccines and covid-19 cases/deaths.

Another big question that lends itself to analysis is the relationship between the proportion of a state's population that is vaccinated compared to the rate of change in the number of deaths related to the Covid-19 Virus. One of the biggest sources of optimism and "the light at the end of the tunnel" during the pandemic was the development of vaccines against the virus. According to the health officials, vaccines can help fight the virus and can prevent death due to the virus. One way we decided to visualize this relationship was to plot, at the state level, the proportion of a state's population that is vaccinated against the change in covid-19 related death rates from the previous day.



This visualization shows us something interesting about the relationship between the rate of deaths due to the Covid-19 virus and the proportion of the population that is vaccinated. According to the visualization, there is a negative relationship between the proportion of a state's population that is vaccinated and the change in Covid-19 related death rates from the previous day. This plot would suggest that if a higher proportion of the population was vaccinated, there

could potentially be a decrease in the number of Covid-19 related deaths compared to a situation in which there was no vaccine. This visualization serves as a springboard for further analysis as we can create different visualizations and models to analyze the relationship between vaccines and covid-19 deaths. From these visualizations, we thought of some open-ended questions to help guide our analysis: “Do vaccines help prevent deaths?”, “Can vaccination rates help predict the number of Covid-19 deaths for a given day?”, and “Is the vaccine effective?”

Problem

One problem that we can address with modeling is seeing the relationship between vaccinations and covid-19 deaths. As we visualized the relationship between covid-19 vaccinations and covid-19 cases, we did see a sort of relationship between the two variables. However, an important metric during this pandemic is the number of people who have passed away due to the covid-19 virus. This is an important question because if there is a relationship between the number of covid-19 deaths and the percent of the population that’s vaccinated, then we have evidence to encourage more people to get vaccinated. Ultimately, these questions are formalized into our hypothesis, which is: the higher proportion of a population that is vaccinated will have a significant decrease in the number of Covid-19 related deaths as vaccination rates are a really good predictor of death rates. To confirm/test this hypothesis, we will evaluate the performance of our model on the dataset and use the RMSE as the metric to test our hypothesis. Although we can come to a conclusion, this hypothesis cannot ultimately be confirmed or rejected until the end of the pandemic and with more, better data. To test this hypothesis, we will include an external dataset about the Covid-19 related deaths. Using this data, we will compute

the per capita vaccination rate and change in per capita death rates to test our hypothesis and conduct our analysis.

Modeling

As it pertains to modeling, we took a similar approach to the guided modeling portion in question 5 and tailored it more towards our hypothesis! Ultimately, we trained a model to predict the rate change of Covid-19 deaths based on the proportion of a county's population that is vaccinated. The model that we used to make predictions was the linear regression model with multiple features. We tested the model performance by using the root-mean-squared-error metric to get a sense of, roughly, how much error our model has on average for all predictions. Since our hypothesis was that the higher proportion of a population that is vaccinated will have a significant decrease in the number of Covid-19 related deaths as vaccination rates are a good predictor of Covid-19 deaths, we decided to use the linear regression model where we selected features of the dataset as model parameters to predict the change of rate of the number of deaths related to the Covid-19 virus, which is the output to our model. Since, in our open-ended EDA, we noticed there was a relationship between the proportion of a given population that was vaccinated and the change in death rates, we decided to use vaccination rates as model parameters for our baseline model. In addition, we improved this model by including more features to create an autoregressive modeling framework.

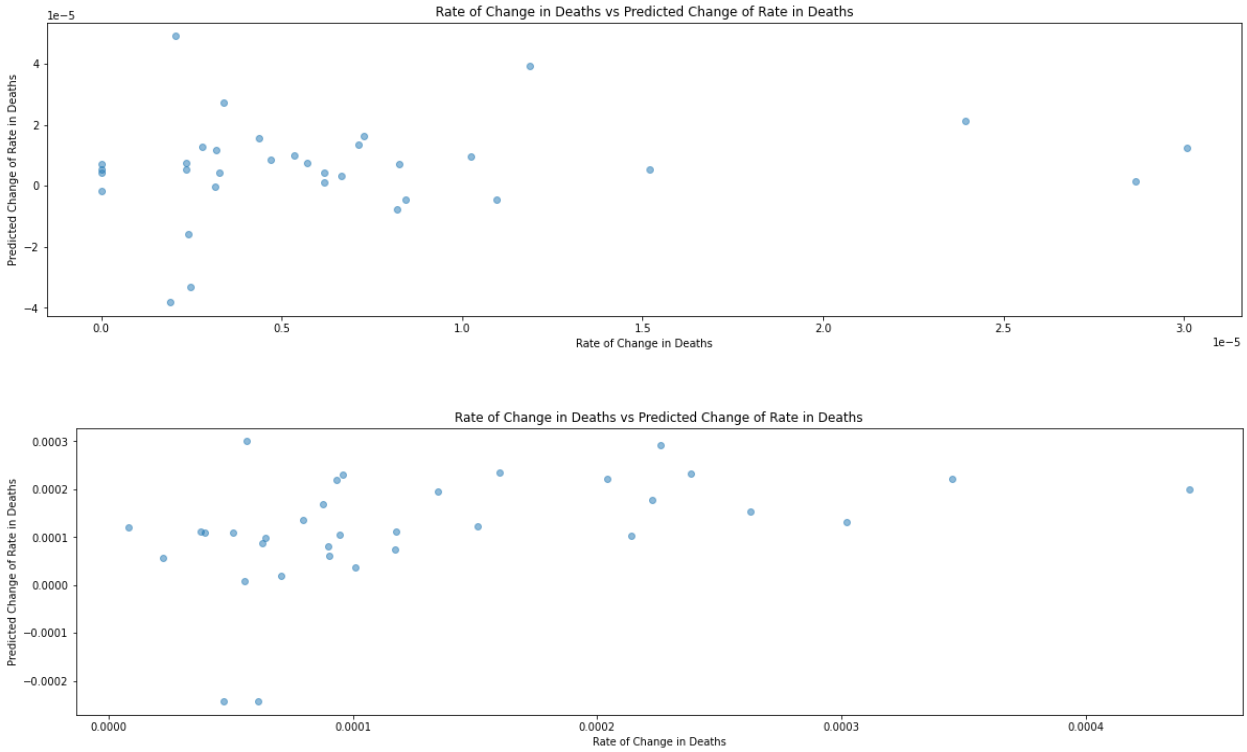
Model Evaluation and Analysis

Given that our hypothesis was that the higher proportion of a population that is vaccinated will have a significant decrease in the number of Covid-19 related deaths as

vaccination rates are a good predictor of Covid-19 deaths, we decided to use the linear regression model to see if vaccination rates were a good predictor of Covid-19 death rates. In order to conduct this analysis, we first had to find an external dataset on Covid-19 deaths. When searching for this Dataset, we found, on Github, the exact dataset we were looking for and it was from the CSSE at John Hopkins University, the same source as the cases dataset! In order to make this dataset ready for analysis, we kept the relevant columns and then grouped the data by months. Next, we merged the Covid-19 death data with the vaccinations data to get a dataset with data for Covid-19 deaths and vaccinations. Finally, we divided the whole dataset by the population for each respective state so we could standardize the data to a per-capita basis. With this new transformed dataset, we began to conduct our analysis. For question 6a, we trained a baseline model using linear regression. For our purposes and hypothesis, we decided to train the model to predict September ‘21 deaths using vaccination data and 2020 monthly Covid-19 death data as features. With this model, we were able to get a low train RMSE of 2.4989378741685852e-05, and a low test RMSE of 3.711825128863507e-05.

These results support our hypothesis as these low values for the RMSE demonstrate that vaccination rates are a good predictor of Covid-19 related death rates. In order to improve this model, however, we decided to follow the guidance in Question 6b and add previous months as features into our model. This created an autoregressive modeling framework in which we predicted the rate change of Covid-19 deaths using previous quantities. In question 6d, we did further analysis on our models from 6a and 6b as we evaluated the performance of our model on both short term and long term scales. The metric, again, that we decided to use to test the accuracy of our model and correspondingly, our hypothesis, was the RMSE of our model’s predictions compared to the actual values. Below, we have two plots describing the relationship

between our baseline model's predicted change in the death rate and the actual change in death rate as long as a visualization of our improved model's predicted change in the death rate and the actual change in death rate.

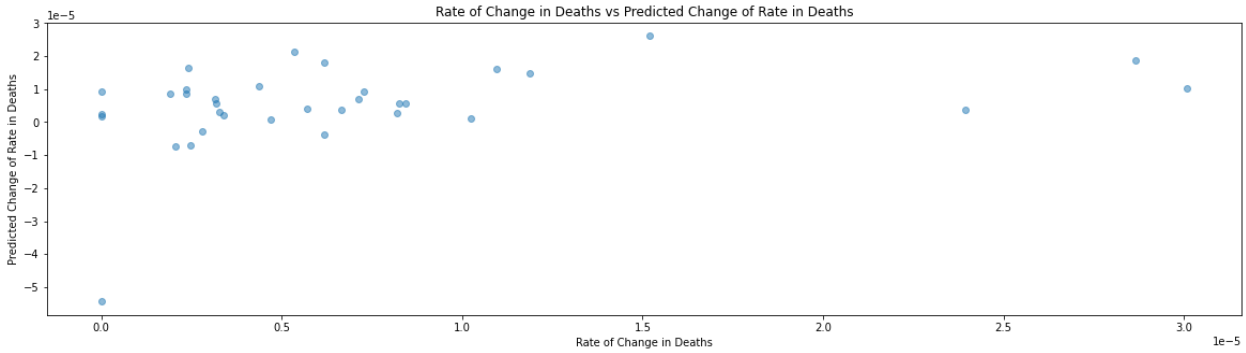


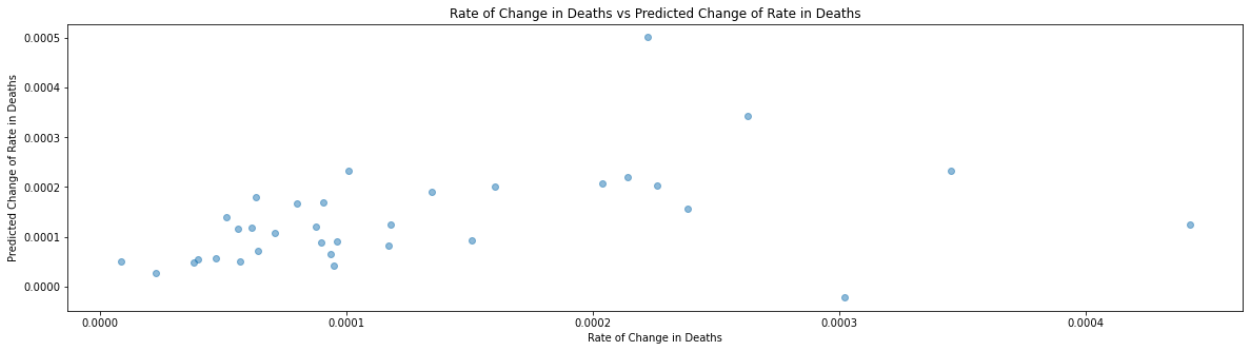
In these plots, we are visualizing the relationship between the rate of change in deaths and the predicted change of rate in deaths for the different time scales of our baseline model. The first is trying to predict the change of rate in deaths per capita for September 13th and September 14th. The second plot is predicting the next two monthly changes of rate in deaths per capita -- October 2021 and November 2021. Overall, there does seem to be a relationship between the values predicted by our model and the actual values, which is also a vote towards the validity of our hypothesis. However, one problem that we came across in this visualization that extends to the whole process of modeling is the presence of negative values in our model. This can be attributed to inconsistencies within the dataset as each sequential day is the previous day's total death plus that day's death count. We noticed that there was a bit of inconsistency within the

original data as some days had less death values than the previous day, which would suggest that someone came back to life, which does not make sense in our case. Even when we tried to correct these inconsistencies through data cleaning, our model still seems to output negative predictions. Overall, however, there was very little of these inconsistencies and these model's still performed well in terms of RMSE!

Model Improvement

In order to produce a refurbished and more improved baseline model, we have decided to keep the fully and partially vaccinated people as features. However, in question 6b, we only included the monthly rate of deaths per capita in the year 2021 as features, starting from January 1, 2021 to September 12, 2021. Unfortunately, the issue that we faced with the baseline model is that having access to features contained only in the year 2020, for the most part, does not involve the role of vaccinations until the very end of the year. With the baseline and improved model, we are able to make a comparison between the two models and how they perform between the short term scale and long term scale. Our results indicated that for the baseline model, in the short term scale, there was about a 63 percent accuracy, while on the flip side, the accuracy decreased to about 48 percent in the long term scale; And while the improved model had a lower training and testing RMSE in question 6b, the model underperformed on both the short term and long term scale -- 40 percent and 38 percent, respectfully. Here below are the visualization results of the improved model performance.

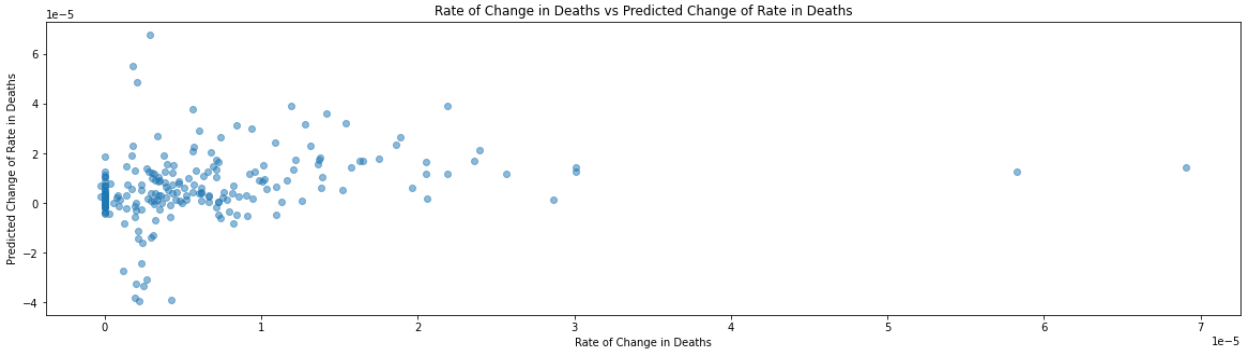


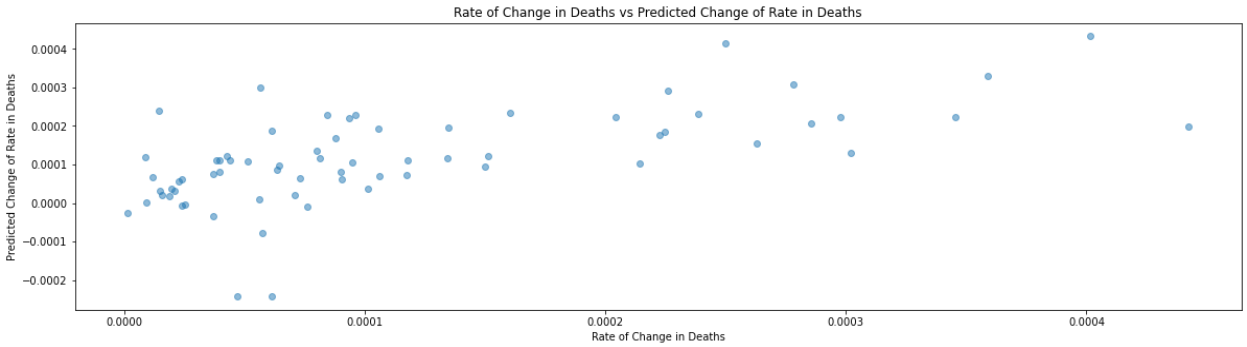


With the improved model, the two visualizations above show a more grouped scatter plot with the data points, while having the short term and long term scales stay the same with the baseline visualizations. In comparison, the improved model seems to have a more positive correlation between the rate of change in deaths and the predicted change of rate in deaths. However, the baseline model performance has a higher accuracy for both short term and long term scale compared to the improved model. Even when the improved model has more recent data on the monthly death rates per capita as features, the baseline model is likely to be biased. Aside from the model being biased with a few outliers, the model performed accurately, and confirmed our hypothesis!

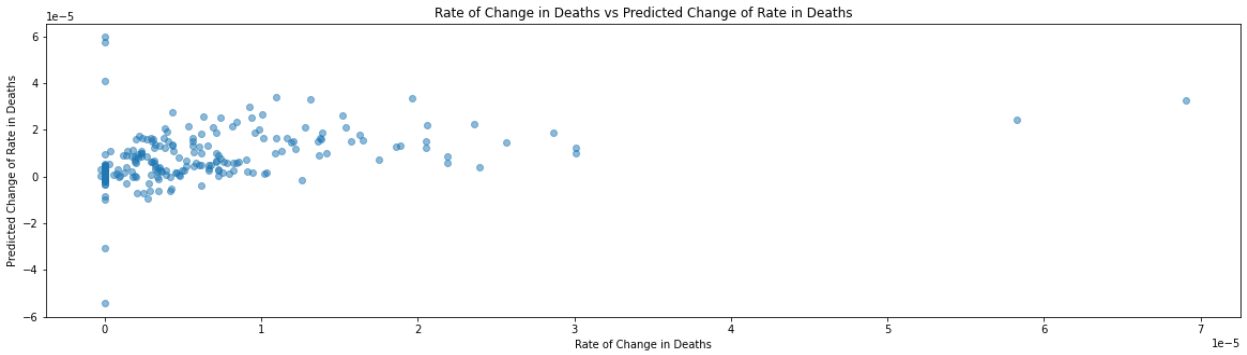
Through further analysis in question 6e, we test the performance of our baseline and improved model predictions by changing the short term scale to two weeks, and four months for the long term scale. Here below are the visualizations between our baseline and improved model:

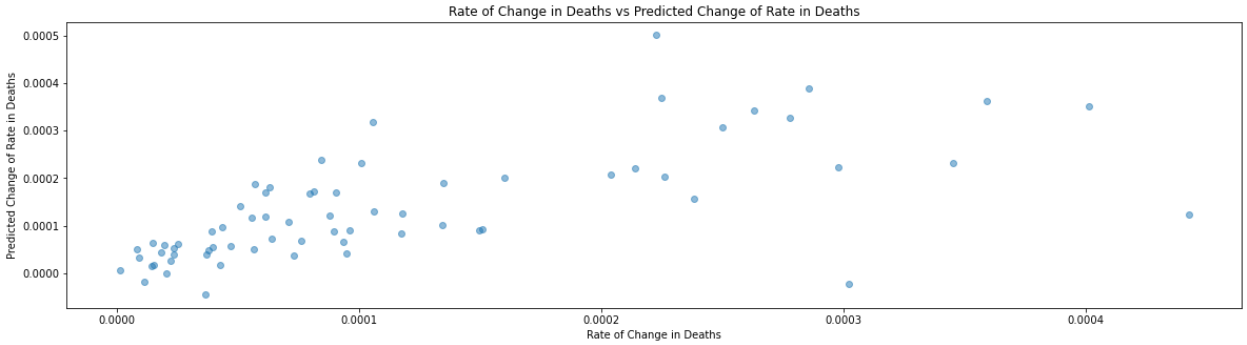
Baseline:





Improved:



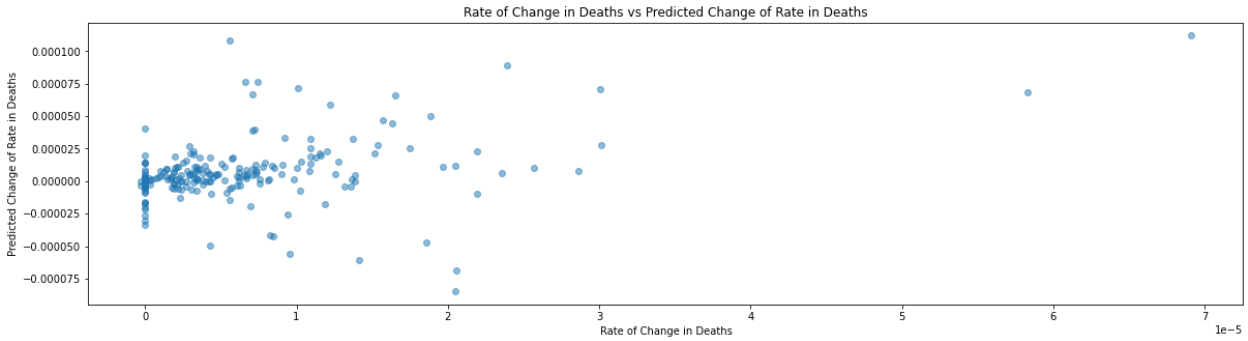


Comparing the visualizations, the improved model plot has more grouped data points for both the short term and long term scale. Again, we notice a trend where the short term and long term scale for the baseline model performs better than the improved model. However, the accuracies for both of the models were higher in the long term scale than the short term scale. Even though we have reduced the train and test RMSE in our improved model, the results of our accuracy in our model may indicate a bias, even in the improved model. Overall, the performance of our model proves that there is a relationship between the rate of deaths and vaccinations, hence the model performed accurately, and confirmed our hypothesis!

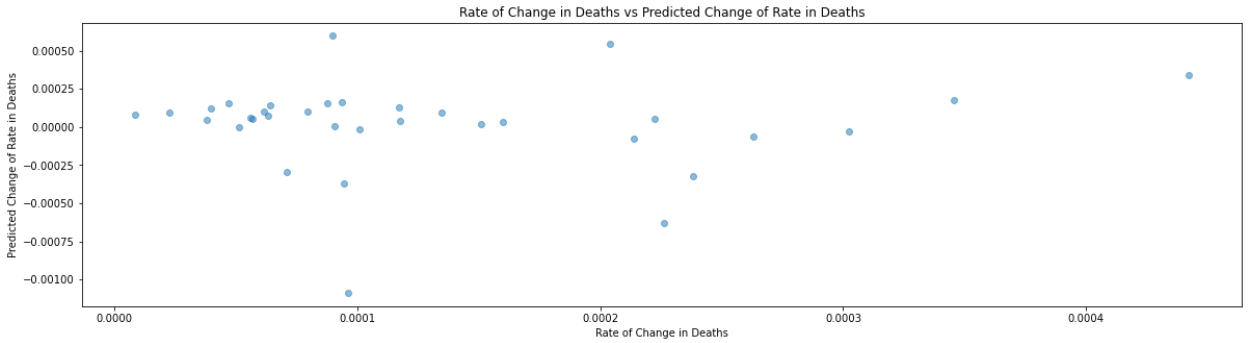
Unsatisfied with our improved model's performance, we decided in question 6f to add additional improvements to the model due to the unexpected results found in 6d and 6e; And this time, instead of relying on a monthly death rate per capita as a feature for the model, we switched to a daily death rate per capita to see if the model can perform better on the short term and long term scale. Likewise, we only used thirty days worth of data, from October 12, 2021 to September 10, 2021, so that the model only utilizes the most recent data to predict future death rates per capita. With a month's worth of data along with the fully and partially vaccinated features, we were able to reduce the train RMSE to 3.796195910327839e-07 and the test RMSE to 7.5208658094055e-06. For the short term scale, we predicted two weeks instead of two days to see how much the model has improved. The results of the short term scale was that the model

had an accuracy of about 99 percent, which is a significant increase from our modeling results in 6d and 6e! We then had our improved model in question 6f to predict the monthly death rates for October 2021 and November 2021, which decreased the performance of the model slightly with an accuracy of about 95 percent. Finally, we tried to predict the daily death rates per capita starting the next day from our improved model features, all the way to December 8, 2021, which is the most recent data from the exported death's dataset. The results improved slightly with about a 96 percent accuracy, a one percent difference from the long-term scale. Here below are the visualizations of question 6f:

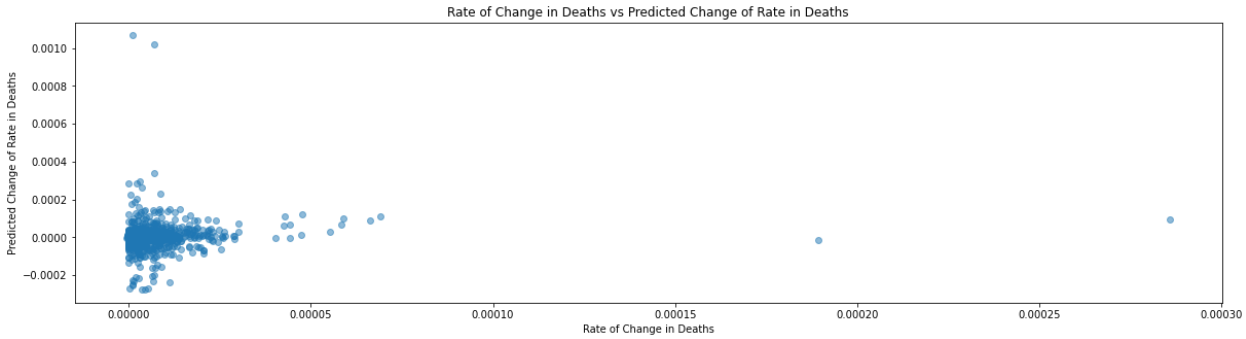
Short-term scale:



Long-term scale:



Long-term scale - All k values:

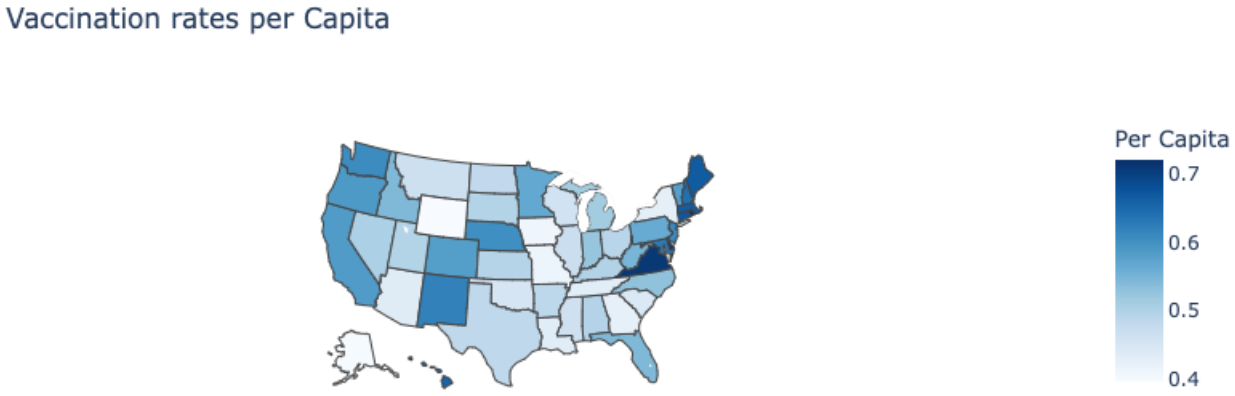


The first visualization shows the data points for two weeks, the second visualization shows data points for monthly death rates per capita for October and November, and the third visualization shows all the data points starting September 13, 2021 to December 8, 2021. Overall, aside from a few outliers, the model performed accurately, and confirmed our hypothesis!

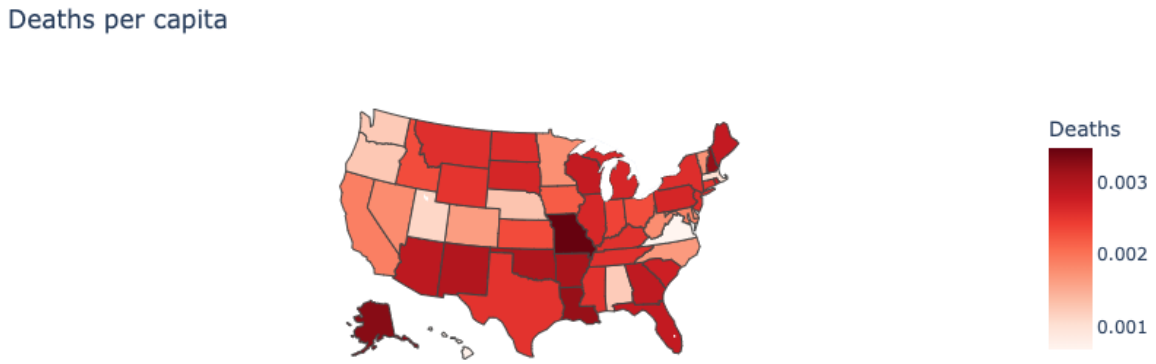
Future Work

Given that the pandemic has more of a profound impact on human life than any other event in recent memory, there is a lot of potential for further work and analysis to help overcome some of the difficulties caused by the pandemic. The Covid-19 Virus has affected life in a variety of different ways and one of the biggest sources of optimism throughout this time has been the development of the vaccine. As we saw in our hypothesis and modeling, vaccination rates are a good predictor of deaths caused by the Covid-19 Virus. Given this, we can do further analysis on the vaccine and the virus and their relationship. First, we created a visual below to show the distribution of the Covid-19 vaccine rate, per capita, across the United States. This visualization shows the states in the USA that have a higher and lower vaccination rate, per capita. With this visualization which serves as EDA for another potential analysis, we can see which states have higher and lower rates of vaccinations and encourage those states with lower vaccination rates to

increase their vaccination rates. With the use of Data and Analysis, we can have solid evidence for convincing state and local governments to encourage more vaccinations.



Another important aspect of the pandemic that can be explored is the deaths per capita for all the states. Given this, and the relationship between deaths and vaccination rates, we can have further evidence for encouraging vaccinations. Below, we created a plot to show the distribution of deaths per capita across the United States.



Given the importance of vaccinations and deaths, another important analysis that can be done is to study the relationship between vaccination rates and political affiliation. This is a controversial topic in the United States as it has been observed that people of certain political

affiliations feel differently about vaccines and vaccine mandates. A common narrative is that those who oppose vaccine mandates are more likely to be on the right side of the political aisle while those who are more supportive of vaccine mandates are likely to be on the left side of the political aisle. Given that vaccines are a very important course of action when it comes to limiting the number of deaths relating to the Covid-19 virus. Ultimately, there is a lot of analysis that can be done, and we understand the importance of Data Science as it pertains to the end of the pandemic!