Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

COVID-19 Dataset Modeling Final Report

Abstract

For this final project, we will be continuing our design and implementation of a typical data science workflow. We will begin our final report by summarizing our Open-Ended EDA findings from the previous section which provides insight into what motivated our hypothesis.

Our hypothesis: *One question we propose for modeling is how do political leanings across counties affect mask usage and thus COVID-19 cases per capita? We hypothesize that mask usage alone can identify a county as either Republican or Democratic correctly at least ⅔ of the time. Specifically, low mask usage counties will tend to be Republican while high mask usage counties will tend to be Democratic. As a result, we hope to model the effect that these predicted political leanings have on COVID-19 cases per capita through auto-regressive models.*

To answer this question, we implemented logistic regression on an imported dataset to perform a binary classification of each county's political leaning (democratic vs republican) using it's mask usage frequencies. After deriving each county's political leaning, we implemented a multiple linear regression model which took in the county's predicted political leaning and past COVID case data as features that would be used to predict future COVID cases for each respective county.

Finally, we conclude our report by recommending several strategies to improve our binary classification model. We can also perform additional research to gain a qualitative understanding of how each county's political preferences affect their COVID-19 rates. This analysis can help us reshape society and lower the number of COVID-19 cases.

Open-Ended EDA

Through guided, supervised, unsupervised, and open EDA, we generated a number of visualizations that allowed us to understand relationships about COVID-19. One key takeaway from guided supervised EDA was that using COVID-19 cases per capita provides a better comparison between counties/states than the absolute number. We were also able to find a relationship between low mask usage and high COVID-19 cases per capita.

While conducting the guided unsupervised exploration, we were able to determine that the majority of the variance is captured by the first two principal components that encode COVID-19 cases per capita, the proportion of individuals vaccinated, or mask usage. Additionally, we found that states with similar (pc1, pc2) values shared similar locations and political preferences. We discovered the importance of political preferences which led to our hypothesis later on.

Throughout our open EDA, we aimed to build upon questions raised in the previous EDA we had conducted. For example, we displayed the relationship between COVID-19 cases per capita and vaccination rates for four states over time (Figure 1). We selected California and Texas because we have domain knowledge about these two states. We selected Alabama and New York to further explore the similarities we noticed between them and their five closest PCA neighbors.

Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

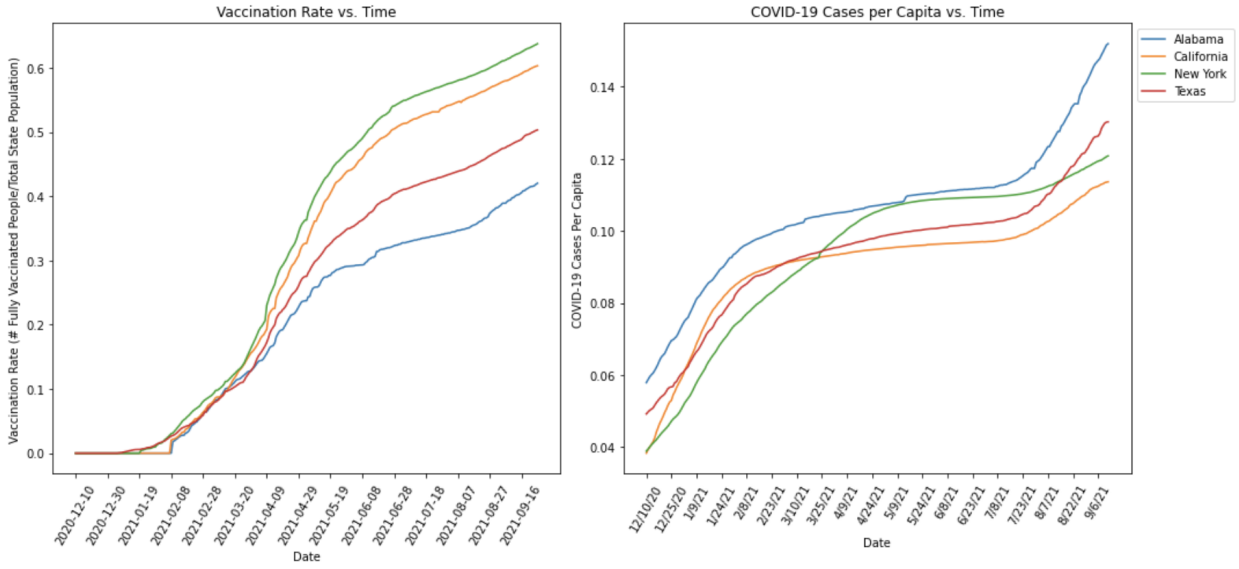
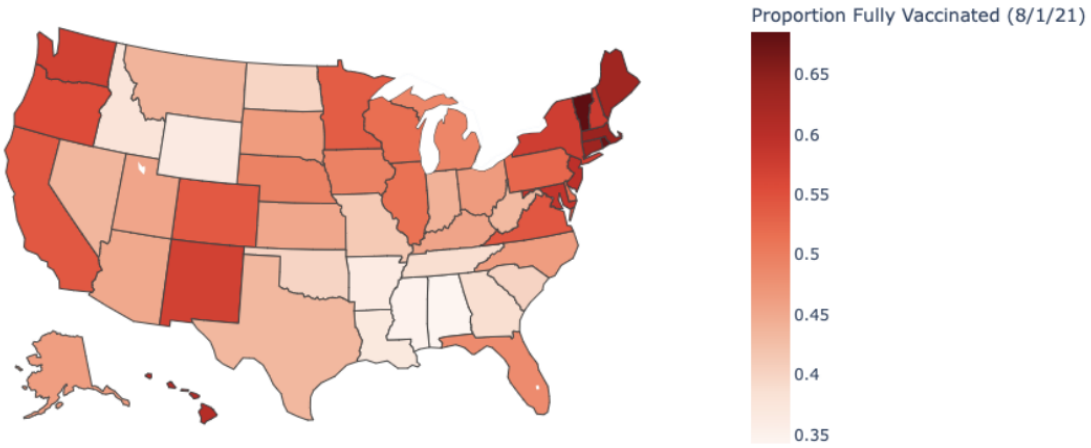


Figure 1

From this visualization, we can see that COVID-19 cases per capita remain relatively constant between March and July for California, Alabama, and Texas. The vaccination rate also begins to steeply increase at the beginning of February. Therefore, there seems to be a buffer between the vaccination rate increases and the corresponding flattening of the COVID-19 per capita curve. This buffer makes it difficult to use vaccination as a feature to directly predict COVID-19 cases or for another supervised learning task. Furthermore, approximately 30% of people were vaccinated in each state by the end of April. However, after reaching this point, we see a large divergence in vaccination rates between these states. In particular, the vaccination rate in Alabama only rises to 40% while the vaccination rate in New York rises to 60%. This may explain why Alabama has more COVID cases per capita than the other states.

To further explore the relationship between vaccination rates and mask usage, we created heat maps of these two quantities across a visualization of the United States (Figure 2).



Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

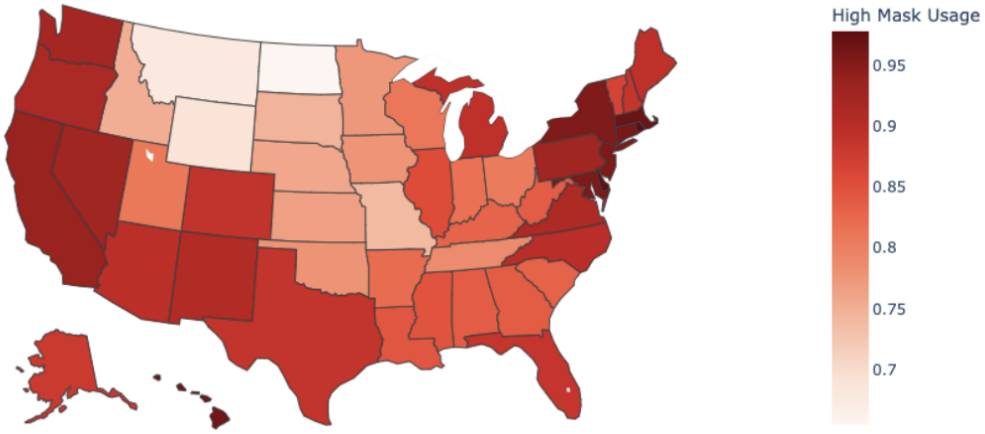


Figure 2

From this visualization, we can observe the relationship between states with high vaccination rates and frequent mask usage. On the East and West coasts, there is a strong correlation between high mask usage and high vaccination rates. In the Central US, there is a bit more ambiguity as states like Texas have high mask usage but a low vaccination rate. Our group expected this relationship because individuals who received the vaccine are more likely to feel that COVID-19 is a serious threat and thus also frequently wear masks. Overall, this analysis was very interesting and led us to form our hypothesis written below.

Problem & Hypothesis

Throughout our EDA, we developed several questions and trends that could be further answered with modeling. *One question we propose for modeling is how do political leanings across counties affect mask usage and thus COVID-19 cases per capita? We hypothesize that mask usage alone can identify a county as either Republican or Democratic correctly at least ⅔ of the time. Specifically, low mask usage counties will tend to be Republican while high mask usage counties will tend to be Democratic. As a result, we hope to model the effect that these predicted political leanings have on COVID-19 cases per capita through auto-regressive models.* Essentially we plan to use political leanings as a proxy for predicting COVID-19 cases from mask usage. So, we plan to use a predicted quantity from one model as a feature in another model. To do this we will need to import county-wide election data from an external dataset as a "creative" data source and compute political leanings from this data to use as a response variable for the first part of the hypothesis and a feature for the second part of the hypothesis. This problem is relevant and intriguing because there seem to be quite a few assumptions nowadays regarding the political nature of mask-wearing, and more broadly, the political nature of COVID-19.

Our open EDA showed that it would be somewhat difficult to focus on vaccination rates due to the time lag needed to see the corresponding effect in COVID-19 rates. We also found a correlation between high vaccination rates and high mask usage. Thus, we thought it would be appropriate to use mask usage as a portion of our modeling problem.

Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

Furthermore, through PCA on vaccination rates, cases per capita, and mask usage, we found that states with similar values for the first two principal components shared similarities geographically and politically. This implies that geography and politics are somehow encoded in the data and could be further explored through modeling.

From our guided EDA, we observed that counties with lower mask usage also had higher COVID cases per capita. Based on this relationship, our group wanted to inquire more about what factors led to certain counties having higher mask usage over others. We will use an external dataset to test part of our hypothesis on whether political preference plays a large role in determining whether people choose to wear masks or not. We can then combine our analysis with our insights from Figure 3 to model future COVID-19 cases per capita through an auto-regressive model.

Modeling

To test our hypothesis, we trained a few models to help us answer our questions. First we trained a logistic regression model to classify each county as either Democrat or Republican. We extracted labels for each county using an external dataset that contained election data. We used the party of the presidential candidate with the most votes in the 2020 election to label each county. We then trained our model using mask usage data as the input and the county-wide political leaning labels as the binary variable to predict. We chose logistic regression for this model because we wanted to predict a binary variable that labels each county as either Democrat (1) or Republican (0). Since these are the only two parties that get the majority of the vote in presidential elections, we thought binary classification would be a good fit.

The second model, an autoregressive model, uses the output from the first model (political preferences) and past COVID-19 case data to predict future COVID-19 case data. In this model, the inputs are the political preferences for each county and several timestamps of COVID-19 cases per capita for each county. The output is the COVID-19 cases per capita on a given day in the future. We chose multiple linear regression because we noticed an inverse correlation between mask usage and COVID-19 cases per capita in our EDA. Additionally, since the majority of the data in the provided dataset was time-based, we chose an autoregressive model to hopefully decrease prediction error. See Figure 4 for a summary of the multi-modeling process.

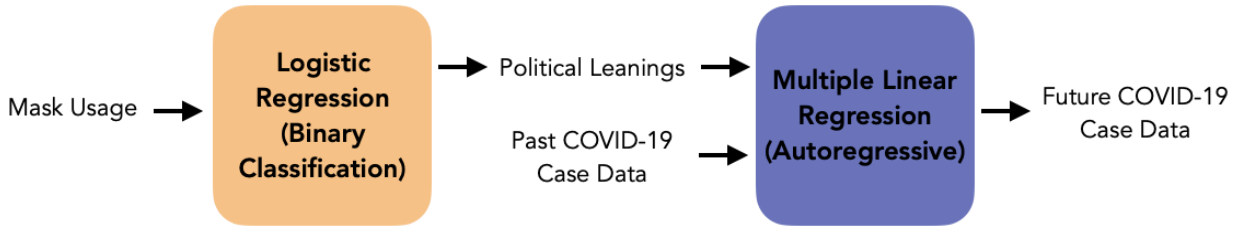


Figure 4

For all models generated, we used train/test splits of 33% of the data for the test set and 67% of the data for the training set. Additionally, in order to accurately make comparisons between models and ensure consistency we set the random state parameter to be the same value in all splits executed.

Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

Model Evaluation and Analysis

Logistic Regression (Binary Classification)

Our logistic regression model trained on the proportion of people in a county that never, rarely, sometimes, frequently, and always wear masks had a training accuracy of 84.32% and a testing accuracy of 85.24%. A baseline random guesser would attain a 50% accuracy, and thus our accuracy for this prediction task can be considered "good."

From the confusion matrix for the test set shown in Figure 5, we can see that our model produced 40 true positives, 13 false positives, 102 false negatives, and 624 true negatives. This resulted in a precision of 75.47% and a recall of 28.17%. The reason that our recall is so low is because there are far more Republican counties in the dataset than Democratic counties, and false negatives are penalized when calculating recall. Nevertheless, we do not believe this low recall will negatively affect the use of this output as a feature moving forward since the accuracy is somewhat more important.

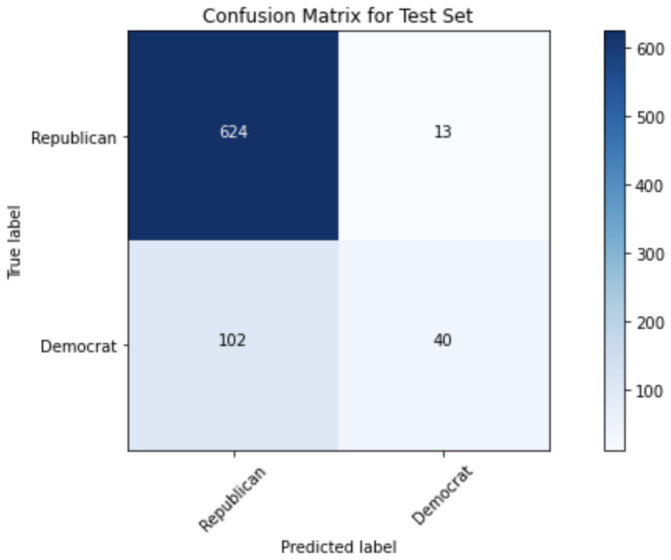


Figure 5

Another way to evaluate the performance of this model is to visualize the real party preferences of a sample of counties against two extreme mask usage variables next to the predicted party preferences of these counties as seen in Figure 6. As we can see, the original data is not linearly separable using these two features. Nevertheless, this model predicts Democrat for counties that have a high proportion of people who always wear masks and Republican for counties that have a low proportion of people who always wear masks and a high proportion of people who never wear masks.

Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

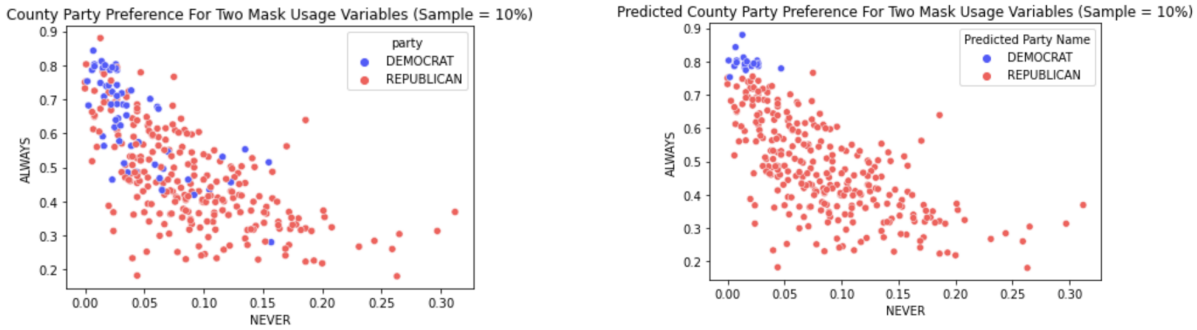


Figure 6

Finally, Figure 7 provides the coefficients of the model for each of the features we trained it on. The intuition from these coefficients leads us to conclude that our intuition in our hypothesis was correct. Never, rarely, or sometimes wearing a mask weighs more heavily towards Republican while frequently or always wearing a mask weighs more heavily toward Democrat as observed by the sign of the values.

	NEVER	RARELY	SOMETIMES	FREQUENTLY	ALWAYS
Coefficients	-2.000009	-2.030086	-2.337286	0.296215	6.05778

Figure 7

Multiple Linear Regression (Autoregressive Modeling)

To perform autoregressive modeling, we used past COVID-19 cases per capita data for each county along with the political party predictions from our binary classifier to predict future COVID-19 cases per capita. Our baseline model which predicts COVID-19 cases per capita on 9/1/2021 using political parties and case data for every previous day available in the dataset had a train RMSE of 0.0002573 cases per capita and a test RMSE of 6.380512 cases per capita. Considering most values for COVID-19 cases per capita for each county in the dataset are between 0 and 0.2, we can see that the test RMSE of this baseline model is quite bad. Particularly, the train RMSE is much lower than the test RMSE which indicates that this baseline model overfits due to the excessive number of features used and does not generalize well. The results and analysis from this model informed our improvement upon this model in the next section.

Model Improvement and Evaluation of Improvement

Due to the extremely high test RMSE, we decided to improve our multiple linear regression model through feature engineering by using only the past two weeks of COVID-19 case data as opposed to using all of the past COVID-19 case data available in the dataset. Since the previous model was overfit, by reducing the number of features, we hoped that our new model would generalize better. By only including the COVID-19 cases per capita from 8/18/21 to 8/30/21 to predict cases per capita on 9/1/21 our model produced a training RMSE of 0.000825 cases per capita and a test RMSE of 0.000859 cases per capita. This is a significant improvement from the previous model which included too many features and

had a training RMSE that was much higher (6.380512 cases per capita). Additionally, now the RMSEs for both the training and test sets are approximately the same.

While this model does quite well with predicting COVID-19 cases per capita for one day in the future, we wanted to see how well it did on predicting COVID-19 cases per capita for multiple days in the future. As seen in Figure 8, the test MSE grows as *k* grows larger where *k* is the number of days in the future for which we are predicting COVID-19 cases per capita. While the test RMSEs still may seem somewhat small, if we put this in context with COVID-19 case data we can better understand what this means. On average there are approximately 100,000 people per county and with a RMSE of 0.007 cases per capita as seen in the prediction for 12 days in the future, that means we have an RMSE of 700 COVID-19 cases. This RMSE can definitely be improved by making further improvements to the model.

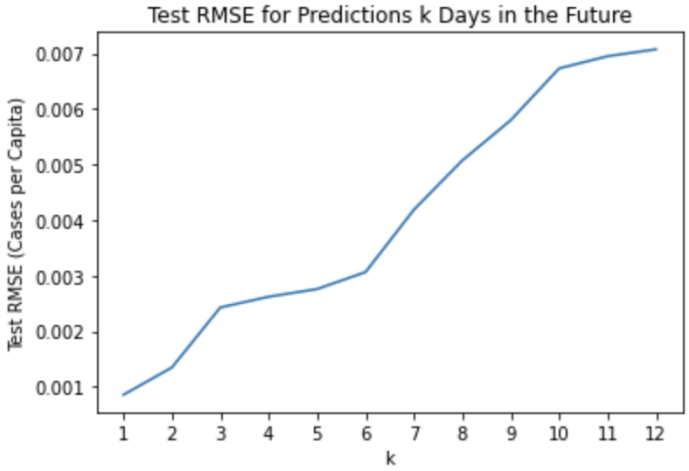


Figure 8

In order to improve this model further, we made an adjustment so that instead of predicting COVID-19 cases per capita for only a single day in the future, the model predicts COVID-19 cases per capita for several days in the future in one shot. In other words, instead of the response being a single-dimensional array with predicted COVID-19 cases per capita for one day in the future, the response is a multi-dimensional array with predicted COVID-19 cases per capita for twelve days in the future. More specifically, this model uses the same input variables as the previously improved model but outputs predictions for twelve different days in a row in one prediction. As seen in Figure 9, this second improved model that executes predictions for multiple future days in one shot has a much lower test RMSE than the first improved model. By making predictions using this model and strategy, we have not only reduced the overall RMSE across all days in the future but also reduced the growth of the RMSE as *k* grows larger.

Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

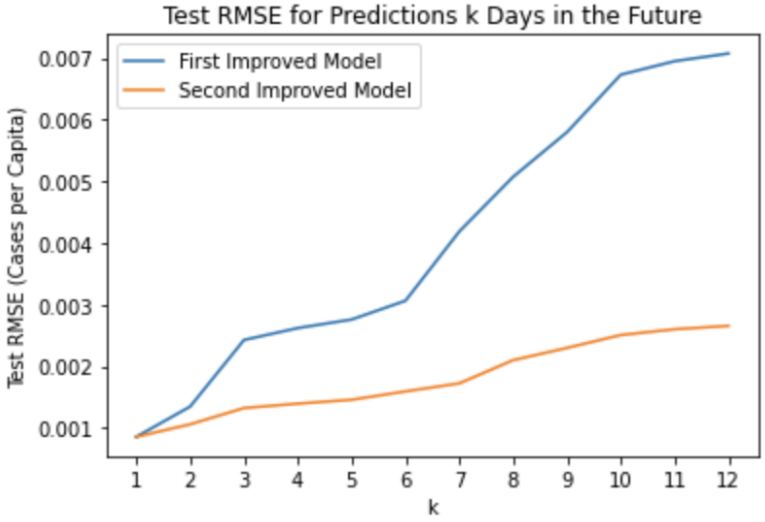


Figure 9

Answer

Now that we have created models, evaluated them, improved upon them, and evaluated this improvement, we can arrive at a conclusion regarding our initial hypothesis. Based on all of the analysis provided above, we confirm the hypothesis. First, we hypothesized that mask usage data alone can predict a county's political preference correctly at least ⅔ of the time. The accuracy for the logistic regression model we built was ~85%, so our hypothesis was correct. Additionally, we hypothesized that low mask usage counties tend to be Republican while high mask usage counties tend to be Democratic. Again, the coefficients from our model in Figure 7 and the corresponding analysis show how this is true. Finally, through all of our autoregressive modeling using political preferences and past case data as features, we arrived at a relatively low test RMSE after improvements, concluding that these political preferences, and more specifically, mask usage, do indeed play a role in predicting COVID-19 cases per capita.

Future Work

For our future work, we believe that there can be additional improvements on a quantitative and qualitative standpoint. From a quantitative standpoint, there are additional ways to improve the binary classification in addition to mask usage. From a qualitative perspective, it would be interesting to learn about how each counties' political preferences shape their COVID-19 cases data.

Improving Binary Classification. There are a couple of additional data sets that may be useful for better predicting a county's political preference. For example, we can utilize a county's past political elections to see if there has been a trend for that county. Moreover, in the previous homeworks, we have utilized demographic information from the United States Census. We can incorporate that data set again to see if adding in ethnicity, income and zip code may improve the accuracy of our function. By adding in these additional features, we do have to be cautious about the risk of overfitting. To mitigate this risk, we can utilize either L1 or L2 regression to reduce model complexity and prevent overfitting.

Shaan Sheth
Curtis Lee
Brandon Cheng
DATA 100

Political Preferences on COVID-19. Recall that we have added political leanings to create a linear regression to predict future COVID-19 cases. Whether a county is Democratic or Republican has large implications on government policies and society. Further qualitative analysis on how politics affects COVID-19 would be a great area of exploration and help us not only better understand society but also prevent further spread of COVID-19. From John Hopkins' University, "Republican governors in 2020 were broadly less strict than their Democrat counterparts in setting policies on … social distancing, and other pandemic-related measures." Therefore, a logical area of exploration would be to look into each governor's policies surrounding social distancing and how much that affects COVID-19. Once we have a better understanding of COVID-19 on a qualitative level, we can use our analysis to create better-informed decisions based on which counties have been more successful in dealing with COVID-19. This will help limit the spread of COVID-19 and potentially save countless lives.