COVID-19 Data Analysis and Inference

Authors: Ronak Laddha, Rithik Lingineni, and Melissa Wong

I. Background

We chose to analyze the COVID-19 datasets for our final project because we wanted to explore the relationship between US counties' political leanings and how their responses to the novel coronavirus differed. We wanted to analyze the political aspects of the novel coronavirus and how counties across the US have responded to the outbreak.

II. Question

How may we use COVID-19 statistics on counties across the US to predict a region's political leaning?

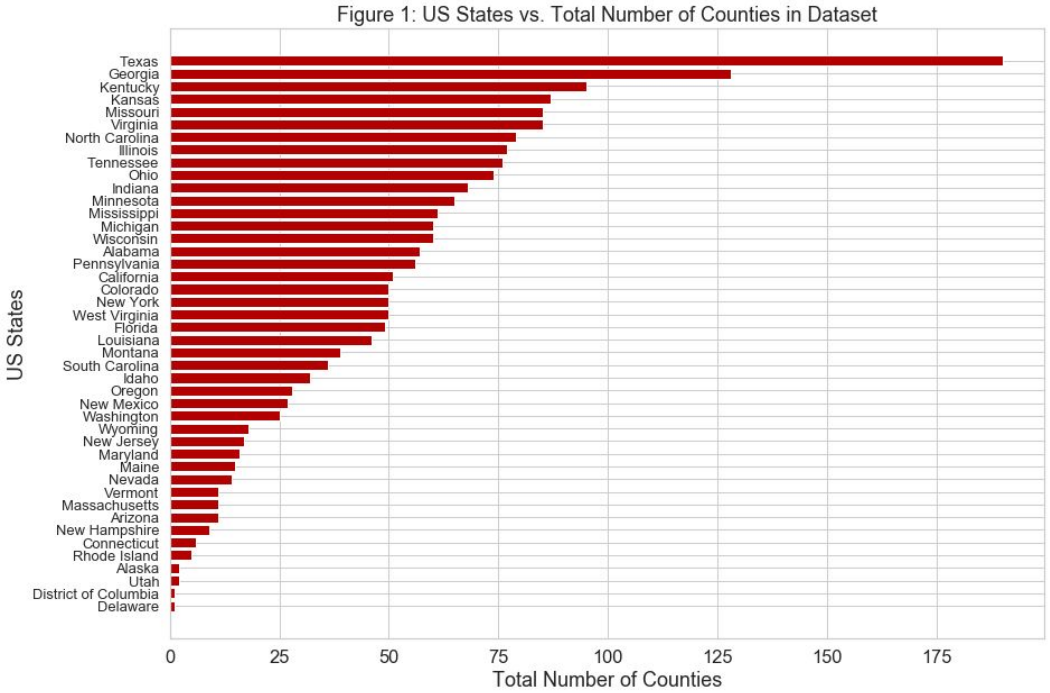
III. Exploratory Data Analysis & Visualizations

We performed EDA on our training set. In doing so, we aimed to answer several questions to help inform the direction we wanted to follow in our analysis.

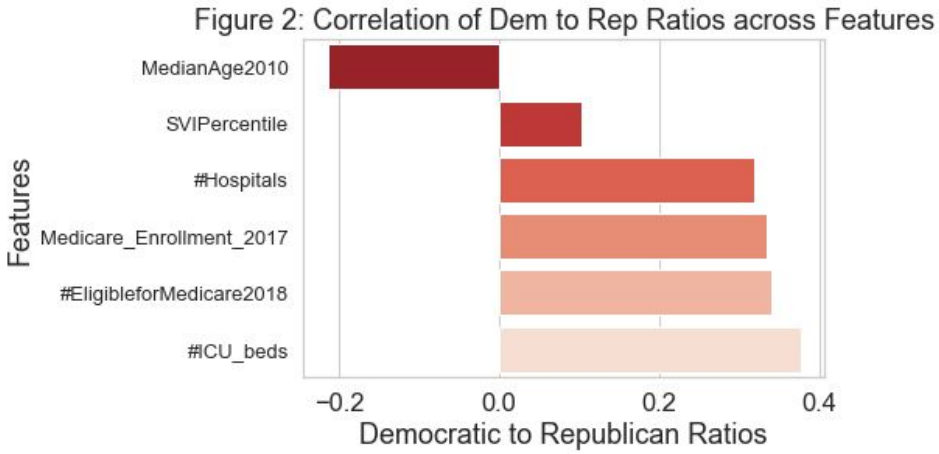
<u>Some questions we posed:</u>

- How is our data distributed throughout our dataset? In terms of states? Counties? Political leanings?
- Are all counties unique?
- Are there any anomalies in our data?
- What is the earliest/latest date a county enacted the stay at home order? Which state were they located in? Their political leaning?
- What is the most democratic leaning county? State? Republican leaning?
- Is there a correlation between dem to rep ratio and the other attributes?

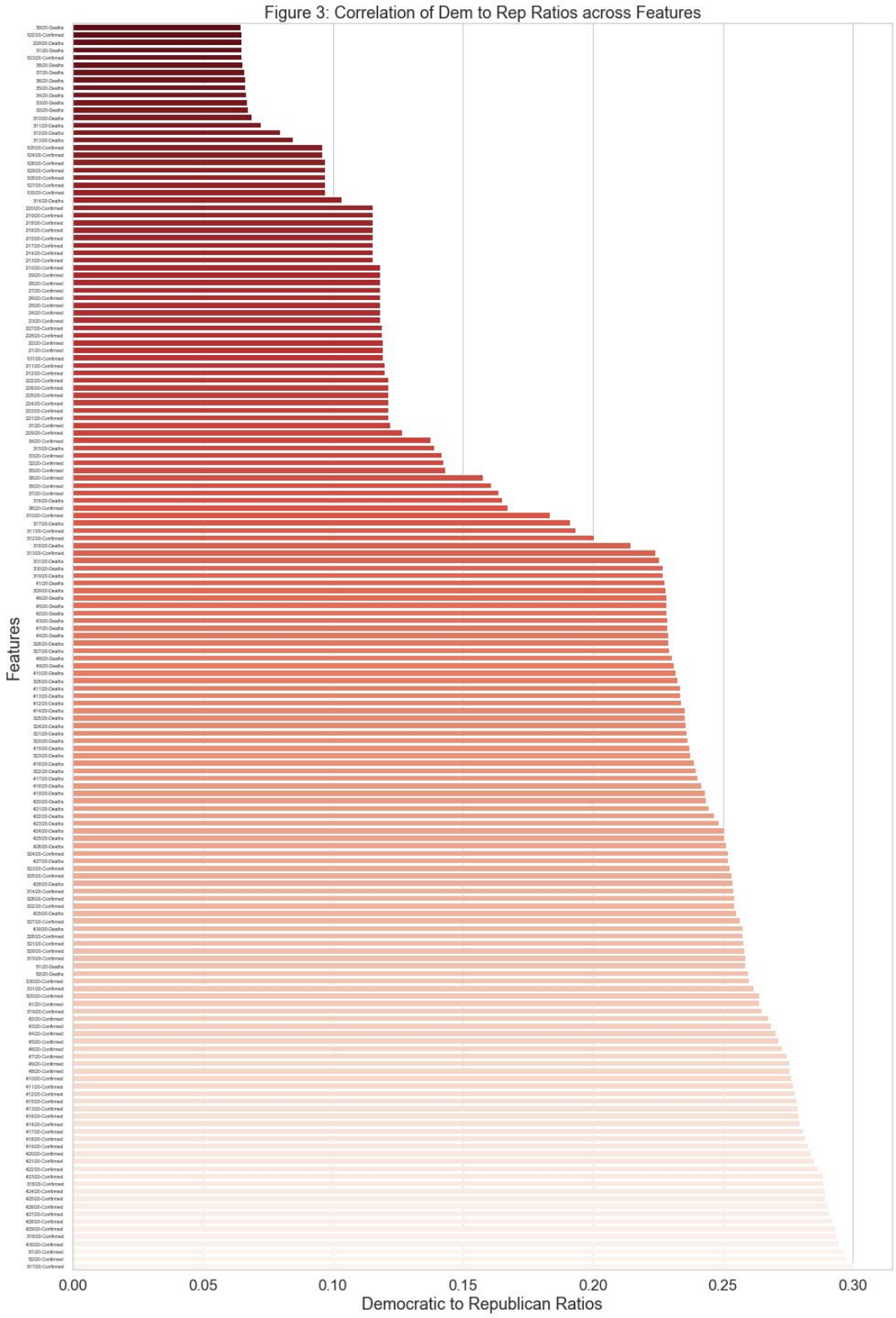
*Below are some visualizations we created to answer some of the questions we had used to decide what our overarching question was going to be.



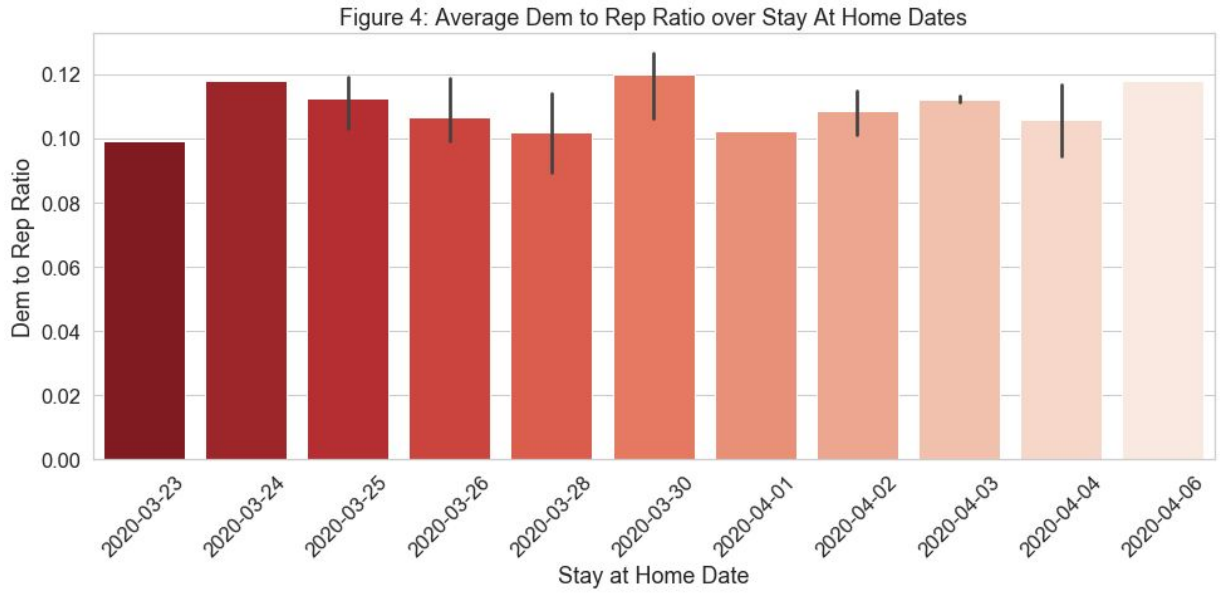
***CAPTION:** Figure 1 is a horizontal bar chart that illustrates the total number of counties we have data on for each state. For example, we have a little under 250 counties for Texas, and around 58 counties for California. This plot shows us the distribution of data of our merged dataset so we have a visual representation of what states are overrepresented in our dataset, and which are underrepresented.*

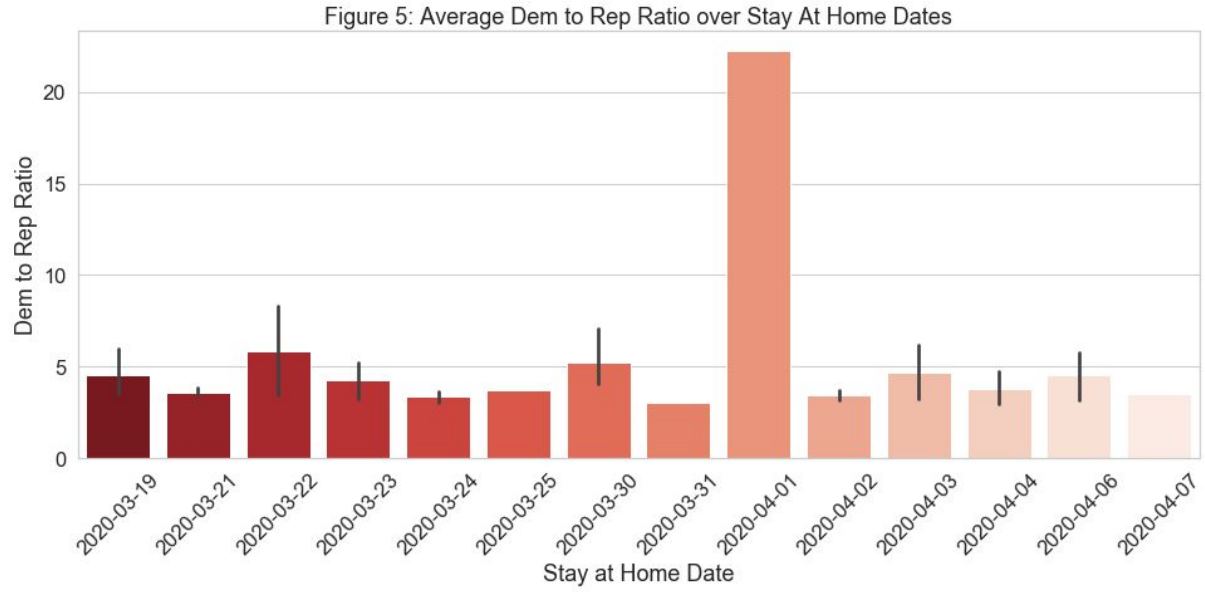


***CAPTION:** Figure 2 illustrates the correlation between the dem to rep ratios of counties within the US states and the following features: median age, SVI Percentile (measures social vulnerability), the number of hospitals, the total number of citizens enrolled in Medicare, the number of people eligible for medicare, and the number of ICU beds.*

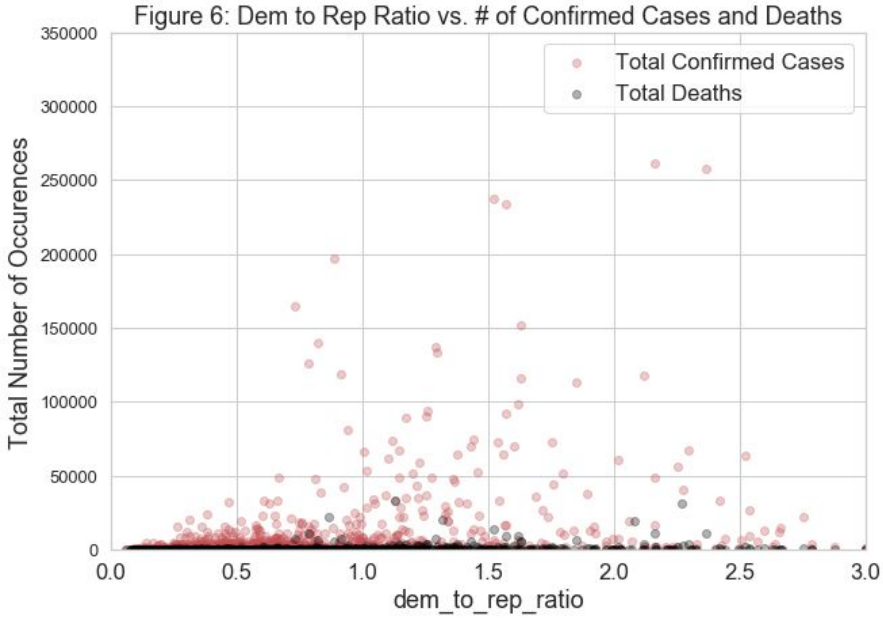


***CAPTION:** Figure 3 illustrates the correlation between the dem to rep ratios of counties within the US states and the following the number of confirmed cases and deaths due to COVID-19 from January to May.*

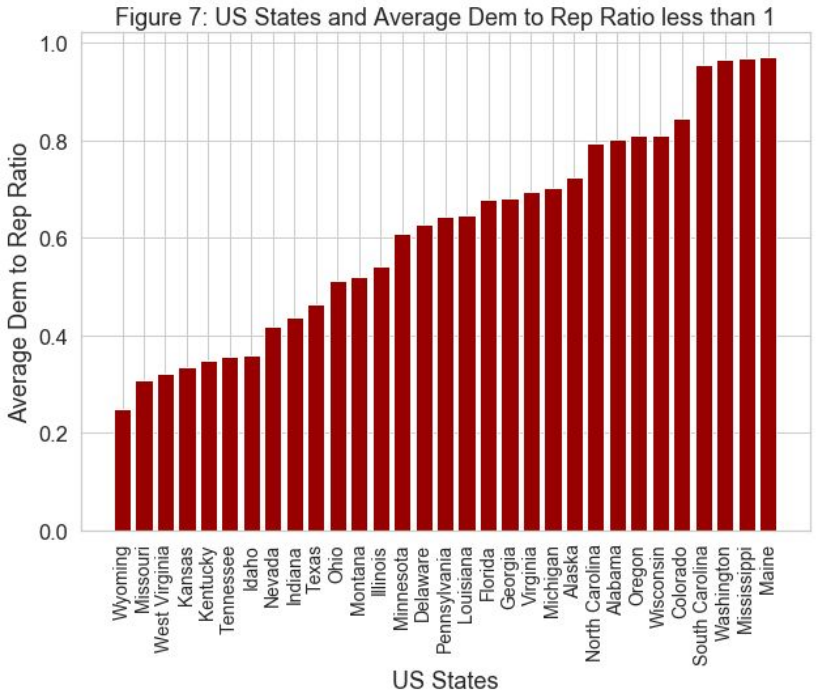




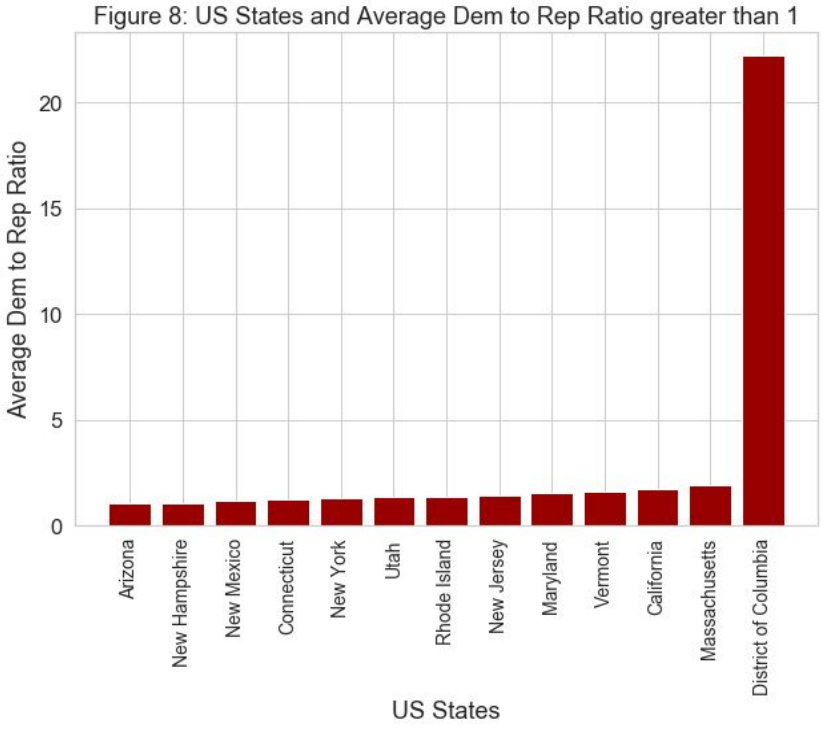
***CAPTION:** Figure 4 and 5 illustrate the average dem to rep ratios of US counties who shared the same stay at home date. The black lines located above some of the bars are the error bars, and they represent the variability of our dataset and the error or uncertainty in the average dem to rep ratio calculation. There appears to be variability because there are multiple states with the same stay at home date and different dem to rep ratios. Figure 4 displays the stay at home dates with the 50 lowest dem to rep ratios and Figure 5 displays the stay at home dates with the 50 highest dem to rep ratios.*



***CAPTION:** Figure 6 represents the all of the US counties' dem to rep ratios as compared to their corresponding total number of confirmed cases, as represented by the red dots, and the total number of deaths, as represented by the black dots. The x-axis only goes up to 3 because we chose to exclude the outliers (any counties with dem to rep ratios greater than 3), because we wanted to focus on the range where a majority of our data points lie in our plot. We wished to graph this regplot so we could see how the number of confirmed cases and deaths were related to US counties' dem to rep ratios.*



***CAPTION:** Figure 7 illustrates the US states with average dem to rep ratios less than 1.*



***CAPTION:** Figure 8 illustrates the US states with average dem to rep ratios greater than 1. We compared the number of states represented within Figure 7 and 8 in order to understand how the difference in distribution of data and the variability of dem to rep ratios. For example, we see in Figure 8 that the District of Columbia has a much higher average dem to rep ratio than all other states represented in the figure. This is because it contains the maximum value in our dataset and is only represented by 1 county.*

Figure 9: Spread of Dem to Rep Ratio Across 50 Counties

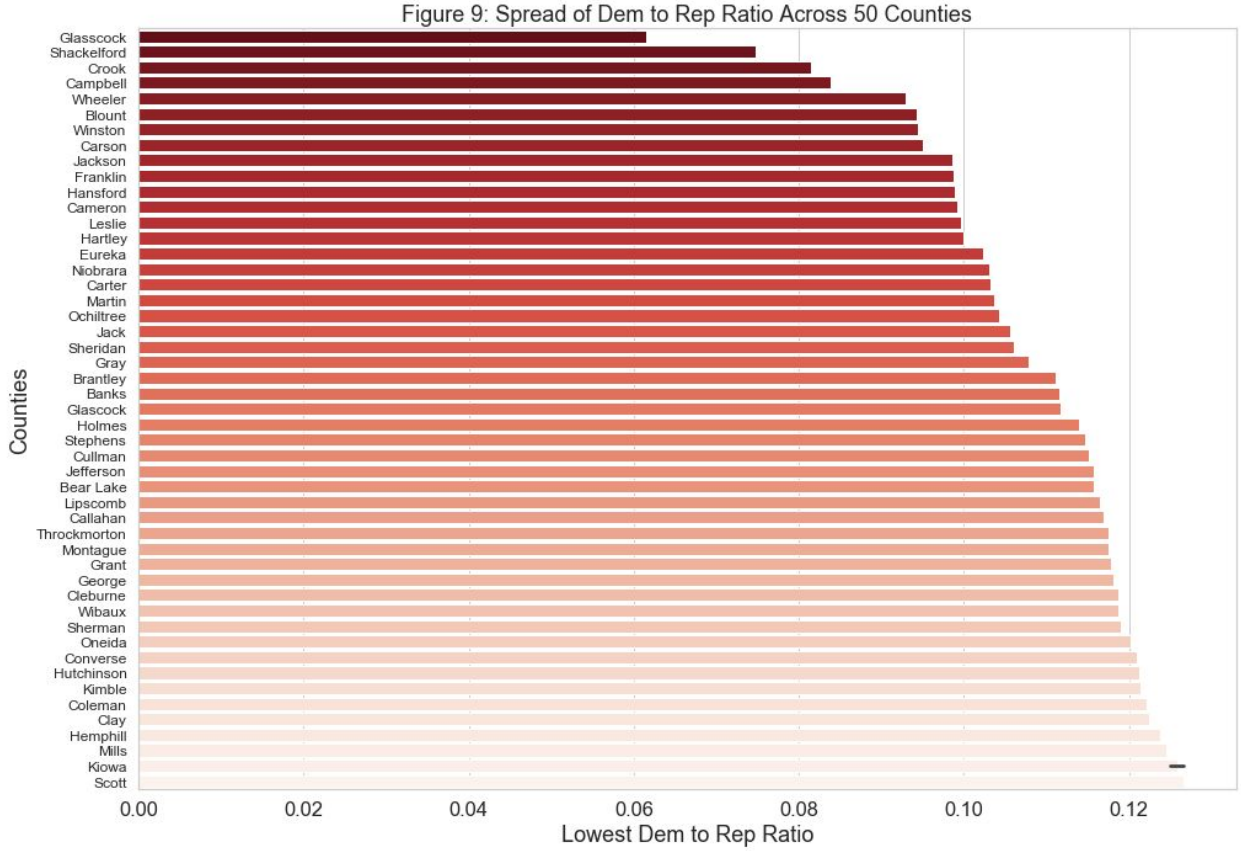
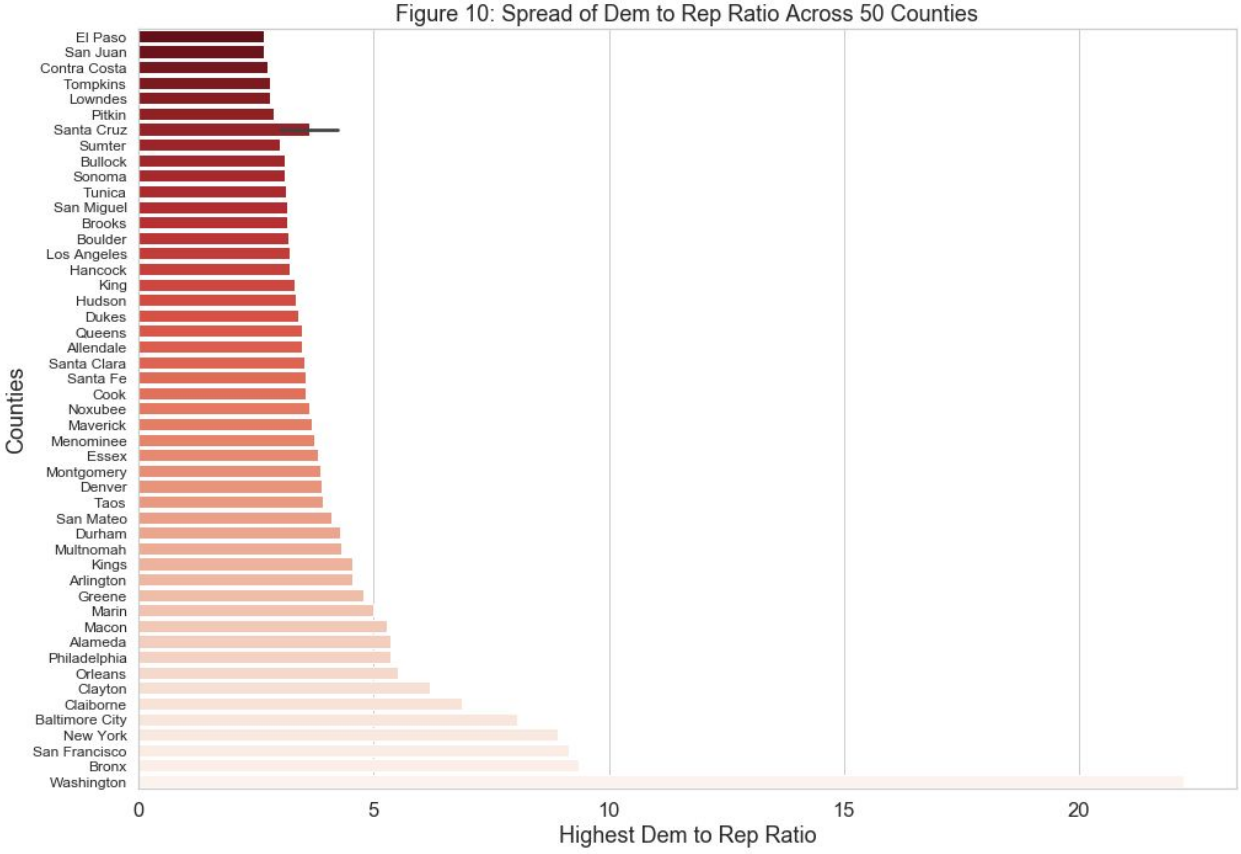
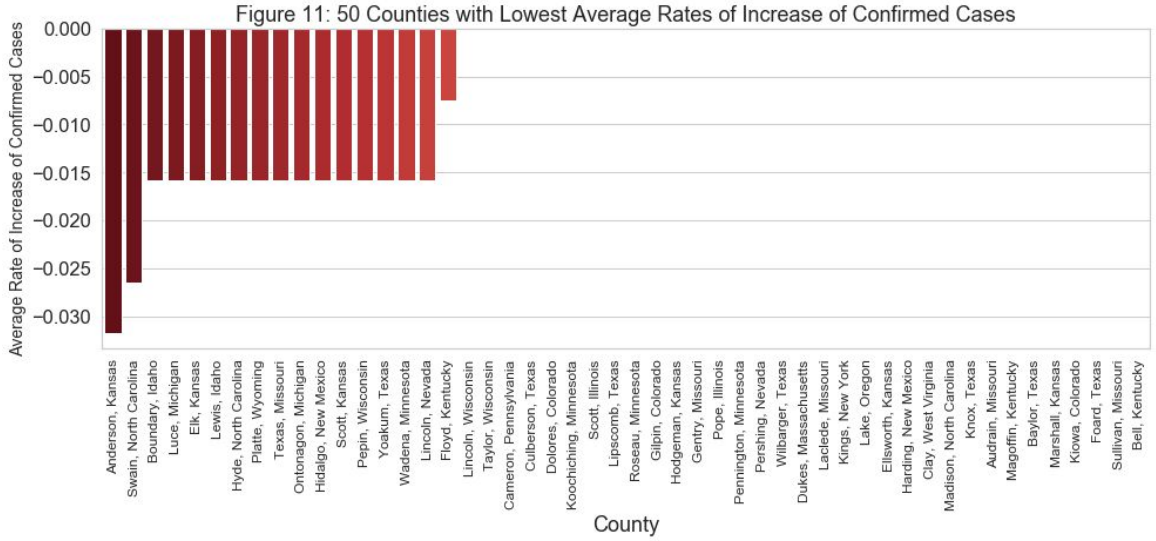
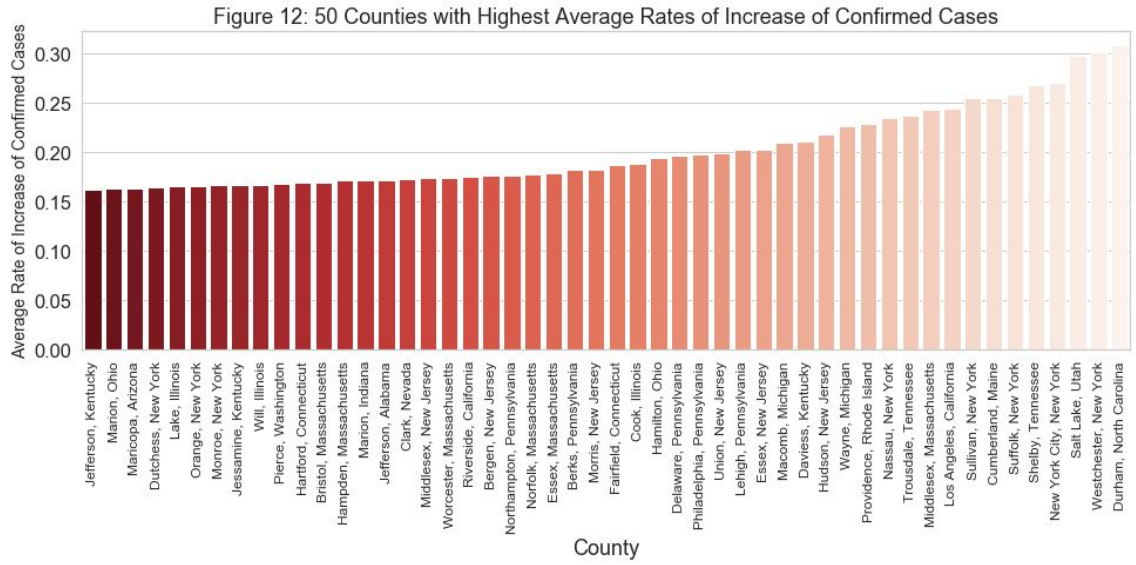


Figure 10: Spread of Dem to Rep Ratio Across 50 Counties

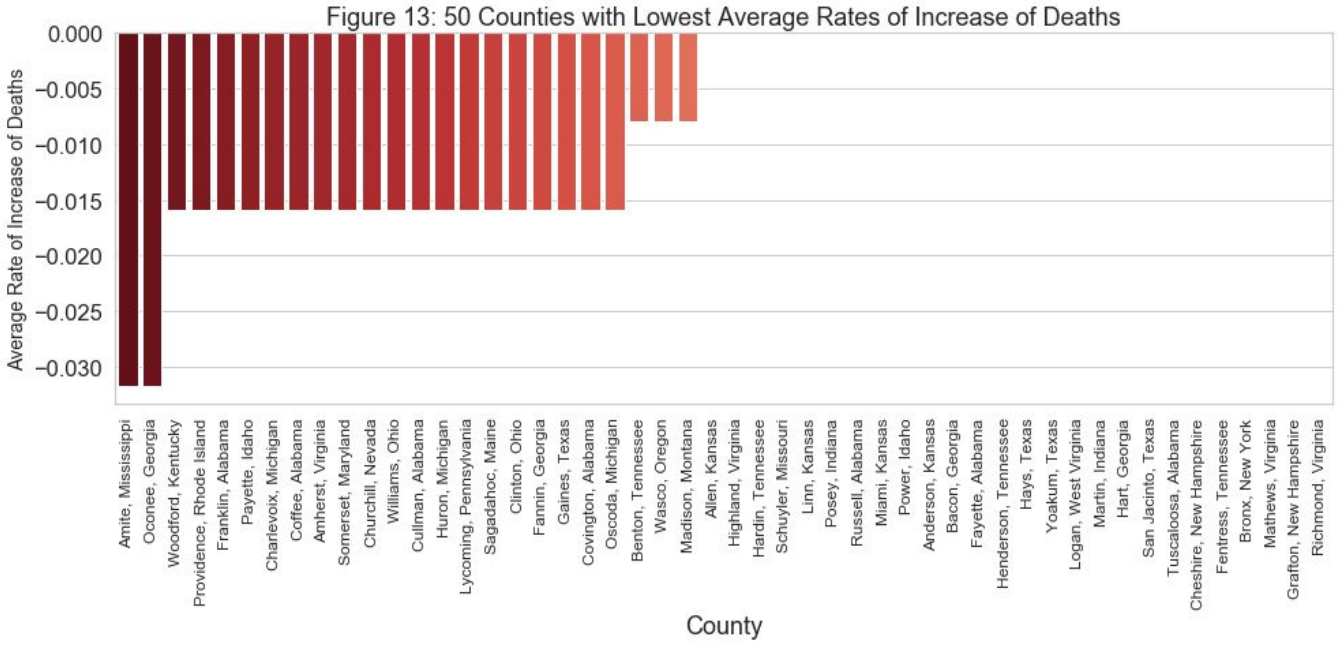


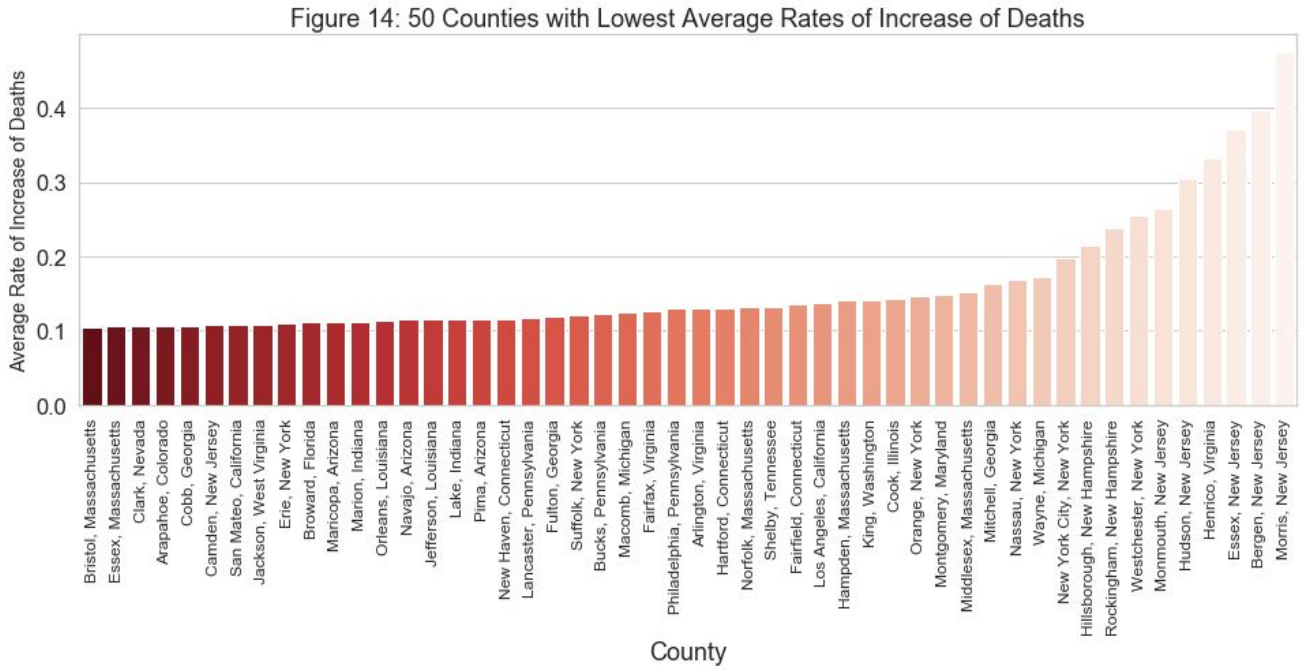
***CAPTION:** Figure 9 illustrates the US counties with the 50 lowest dem to rep ratios. Figure 10 illustrates the US counties with the 50 highest dem to rep ratios. We visualized these metrics because we wanted to see the spread of dem to rep ratios across all of the counties in our dataset. We could not plot all counties on these graphs, so we graphed the lowest and highest 50.*



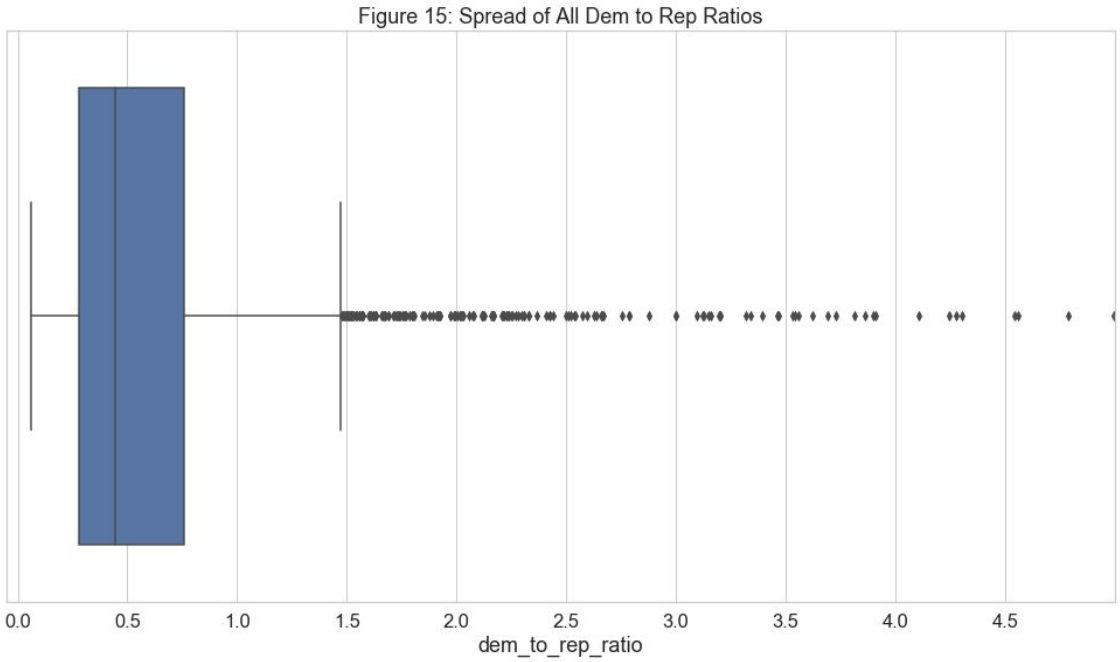


***CAPTION:** Figure 11 illustrates the 50 US counties with lowest average rates of increase of confirmed cases over the months of January to May. Figure 12 illustrates the 50 US counties with highest average rates of increase of confirmed cases over the months of January to May. We could not plot the average ROI for all counties in our dataset because that is over 2000 data points. In addition, graphing this plot made us realize that there are counties with the same name in multiple states, thus the reason why each county is listed with its corresponding state. In addition, we can see that New York and North Carolina notably had the highest average ROI of confirmed COVID-19 cases.*

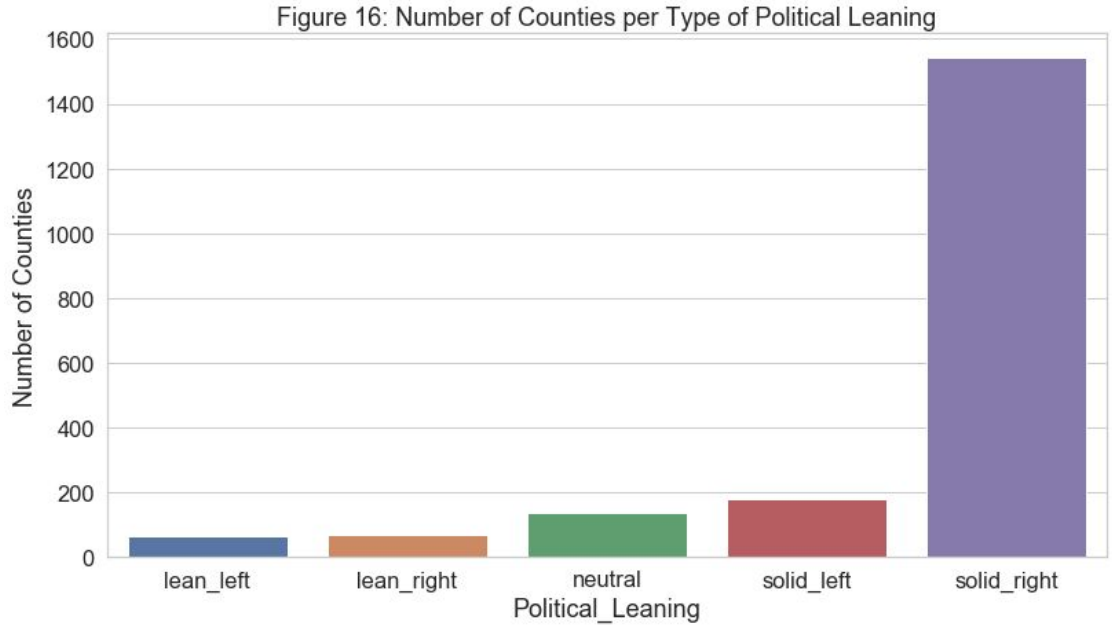




CAPTION: ***Figure 13 illustrates the 50 US counties with lowest average rates of increase of deaths over the months of January to May. Figure 14 illustrates the 50 US counties with highest average rates of increase of deaths over the months of January to May. From Figure 14, we can see that US states located on the east coast, in particular, New Jersey, New Hampshire, and New York notably had the highest average ROI of deaths.***



***CAPTION:** Figure 15 is a boxplot of the spread of all dem to rep ratios across all US counties in our dataset. Most notably, we can see where our data are distributed, as the media is a little less than 0.5 and we have many outliers beyond 1.5.*



***CAPTION:** Figure 16 is a bar plot of the total number of counties that fall into each of the 5 classifications according to their respective dem to rep ratios. We can see that our data is clearly not distributed evenly, as there appears to be a majority of counties falling into the "solid_right" category.*

IV. Data Cleaning & Transformations

We worked with 3 datasets: `counties_df`, `confirmed_df`, and `deaths_df`.

1. How we cleaned `counties_df`:
    - Selected relevant columns that would work as features for our classifier.
    - There was missing data in the `State` column for the rows representing the state of Alaska, so we filled in the missing values.
    - Identified the NaN values in `counties_df` and selected to drop the rows with the null values. Did not make sense to replace the nulls with the mean values of our dataset.
    - The `stay at home` and `restaurant dine-in` column values were in the Gregorian ordinal form of the date, so we converted them all into datetime values so that we could easily interpret, sort, and graph them.
    - Renamed the `State` column value of DC to match the format of the other datasets.
    - Renamed the columns in order to match those of `confirmed_df` and `deaths_df`, so we could merge.

2. How we cleaned `confirmed_df` and `deaths_df`:
    ** Used the same data cleaning techniques because they are identical in types of columns. **
    - Dropped all columns that contained repetitive or unnecessary data.
    - Removed all rows that represent the US territories outside of the United States of America, such as the American Samoa.
    - Appended indicator tags to `confirmed_df` and `deaths_df` column names, because we needed to differentiate them when merging datasets.
    - Renamed the `Admin2` column value of DC to match the format of the `counties_df`.
    - Renamed the columns in order to match those of `counties_df`, so we could merge.

3. Transformations:
    - Once we cleaned the three data frames, we merged them into one in order to combine all of the relevant data into one dataset. To do so, we first merged the `confirmed_clean` and `deaths_clean` datasets through an inner join, as they contained the same columns (except for the number of cases per day). Then, we merged `counties_clean` with the `confirmed_and_deaths` dataframe using an outer join to ensure that we had complete information on the counties, number of confirmed cases, and deaths.
    - We then took this merged dataset and split it into a training and test set using the 80/20 split.

V. Methods

We first used a Logistic Regression model in order to predict the political leanings of a US county based off of a particular set of features; however, we realized that a Random Forest

classifier would be a better model, as our data had a large class imbalance (Figure 16). Thus, the final method that we used for our prediction analysis was Random Forest, as we identified a clear class imbalance in regards to republican leaning counties and democratic leaning counties. We see that in some states, there are many counties represented (some with very small populations), and in some states, there are much less counties represented in our data. This imbalance ultimately skews the data and is represented in the distribution we noticed while performing EDA, as we did not weigh population in our calculations. Furthermore, through performing analysis on the percentiles, we discovered that there were many more dem_to_rep_ratios that were under 1 compared to those that were over 1, in part due to this imbalance in the number of counties. Thus, these reasons motivated our decision to select Random Forest as our model to negate this issue and help us answer our question.

VI. Model & Assumptions

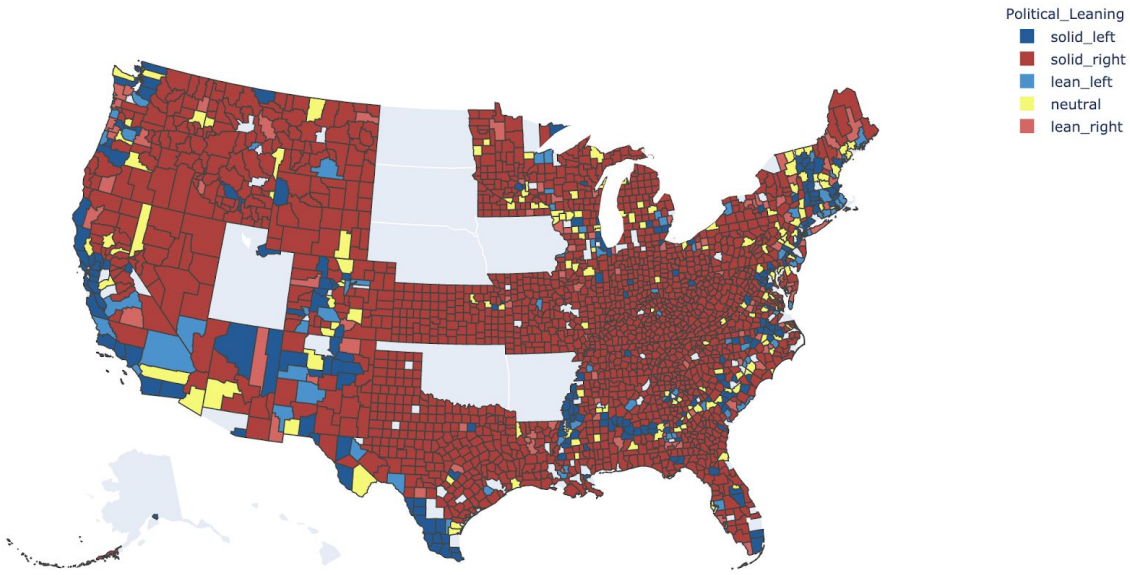
- **Our Model:** We built a Random Forest classifier in order to predict a county's political leaning as either `solid_left`, `solid_right`, `lean_left`, `neutral`, or `lean_right`. Our model uses features such as median age, number of citizens enrolled in Medicare, SVI Percentile (measures social vulnerability), number of ICU beds, and the average rate of increase of confirmed cases and deaths in order to construct the model.

- **Assumptions:** In creating our model, one of the main assumptions we made was in building the classifications for state's political leanings based on their dem_to_rep ratios as represented in our data. Rather than try to normalize the data using our personal/anecdotal opinions, we utilized the "Party Identification and Leaning, by State, 2016" chart on Gallup's website to help us classify the ranges of `solid_left`, `solid_right`, `lean_left`, `neutral`, and `lean_right` counties through the data presented. We calculated the bounds as being the points in which the "State Type" column changed from one political inclination to the next and we calculated the numbers by dividing the Democratic Voting % by the Republican Voting % to get the exact `dem_to_rep_ratio` bounds to help classify our data. Another assumption that we made was that `dem_to_rep_ratio` was a good enough of a statistic to base an entire county's political leaning classification on.

VII. Summary of Results

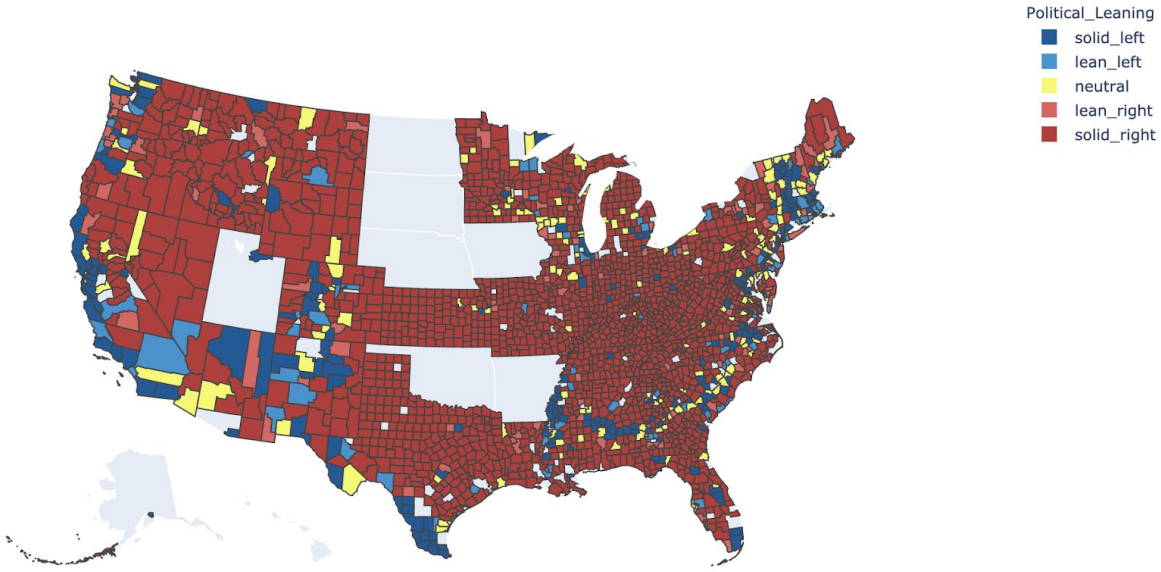
We ran our Random Forest classifier and got a **training accuracy of 0.996 and a test accuracy of 0.784**. Based on these results, we can come to the conclusion that there does appear to be a correlation between our chosen features and US counties' political leanings. In addition, we found that we could classify counties and their political leanings based on COVID-19 related data, such as SVI Percentile, the number of ICU beds, and the average rate of increase of confirmed cases and deaths. Our conclusion is further evident in the following visualizations.

Predicted Political Leanings of US Counties



***CAPTION:** The above map displays the predicted political leanings of all US counties contained within our dataset. The key on the right illustrates the colors that indicate the 5 possible political leanings the US counties could be classified as. The blank areas represent US counties that we did not analyze because they were missing vital data.*

Actual Political Leanings of US Counties



***CAPTION:** The above map displays the actual political leanings of all US counties contained within our dataset. These classifications are based solely on the dem to rep ratio of each of the counties and the bounds that we constructed using Gallup's chart.*

VIII. Methods

A. Most interesting features?

- Two features that we found particularly interesting and were related to our question were `SVIPercentile` and `FIPS`. `SVIPercentile` was an interesting feature because we did not know there was a statistic that measured a county's social vulnerability. In addition, before completing this project, we did not know that the `FIPS` value of states and counties were used to plot their location on maps.
- In regards to our particular question, we found that most of the counties we had complete data on tended to be `solid_right`. This result was interesting; however, not surprising considering the state of our current political climate.

B. One feature you thought would be useful, but turned out to be ineffective?

- One feature we thought might be useful was `MedianAge2010`; however, it turned out to be ineffective because we calculated the correlation between that feature and `dem_to_rep_ratio`, and it appeared to have a negative correlation. We initially thought that it might be useful because we thought there was some relationship between an individual's age and how left/right leaning they are.

C. Any challenges with your data? Where did you get stuck?

- **Null Values:** One challenge we faced was deciding to do with the abundance of null values in each of the three datasets that we worked with. We were debating about whether or not to fill in the null values in some way, or if the best option was to drop the rows that contained nulls. We were worried about not having enough data to work with if we did drop, but ultimately decided that we had enough data.
- **Classification Bounds:** We struggled with the decision of how we would classify counties' political leanings. We knew that we wanted to classify each county as one of the five political leanings; however, we did not know how to create the bounds that would classify each. We tried constructing bounds based on how the data were distributed; however, that approach was quite arbitrary, so we calculated the bounds using real-time data from the chart on Gallup's website.
- **Reformatting FIP Values:** After we constructed our model, we wanted to visualize our predictions on a map of the US. We used the FIP values and Plotly's Choropleth Maps to do so. However, once we graphed the data, we noticed that many counties were not represented in the graph even though we had complete data on them. We realized that this issue was due to the formatting of the FIP values for counties with 4 digit FIP rather than 5, and thus those counties were plotted incorrectly. However, through many trials and discovering that these FIP values could be read as String types, we successfully plotted all of the counties we had complete data on.

D. Any limitations of the analysis? Any assumptions that could be incorrect?

- **Data:** There were limitations in the data, for example the numerous missing rows of data, which prevented us from being able to truly analyze the scope of political

leanings across all of the counties in the United States. This lack of complete data may have led us to create unintended biases as there may have potentially been more counties with certain political leanings that could impact our model.

- **Classification:** The Gallup poll that we used to help us classify our data utilized a selection of 177,788 random telephone surveys in 2016 in order to classify states by their political leanings. We made an assumption that the states classifications could be paralleled to the classifications of counties (regarding the type of political lean) through utilizing the democratic voting % over the republican voting % they listed. Furthermore, we assumed that the classifications of whether a state was Solid Dem -> Solid Rep on Gallup were accurate. These assumptions could be proved to be incorrect, so the data is only reflective of 2016 (as 2016 was the most readily available data), not 2020, so the political landscape could have shifted by now.

E. What ethical dilemmas did you face with this data?

- An ethical dilemma we faced with this data was in regards to how it was collected. Many of the features we used to train our classifier were medical data, which could have been collected from the individuals without their consent or knowledge. We are unaware of how this data was obtained or processed before we worked with it, thus we are questioning whether or not we were working with ethically questionable data.

F. What additional data would strengthen your analysis?

- More recent data on `dem_to_rep_ratio` and political leanings by state in 2020 to help us better classify counties in our model to reflect this year's political landscape. The three datasets we decided to work with have last been updated on May 2, 2020; however, the feature that we were trying to classify on was based off of a 2016 election. It is reasonable to think that perhaps the political leanings of a county might have changed in the four year time span from then up till now. Thus, having additional data on the `dem_to_rep_ratio` of a county in regards to a more recent political event would strengthen our conclusions.

G. Any ethical concerns while studying this problem? How might you address them?

- Ethical concerns we might run into through studying this problem are how politics and political leanings of counties have had a direct impact on people's lives. Through looking at this data we might find how because of political beliefs responses to COVID-19 may have been delayed or not as sufficient, and consequently innocent people might have been negatively impacted. We may find out more about the impact of politics on the health of nations and whether there is an ethical concern regarding letting politics influence medical response during a pandemic.

IX. Approach & Limitations

We used a Random Forest model to solve our classification problem. Originally, we had naively used a Logistic Regression Model, but we noticed that we have a fairly severe class

imbalance, as described earlier in this report (Section V: Methods, Figure 16). To address this class imbalance we decided to use a learning algorithm that would be more robust to class imbalance. We considered using decision trees, but they are extremely susceptible to overfitting. Therefore, we decided to use a Random Forest model because it uses an ensemble of decision trees to model and represent the data. There was no way for us to completely eliminate the class imbalance because we noticed that the states with many counties usually voted Republican (Figure 1). We considered looking at the political leanings of states rather than their counties, but realized that we were working with data respective to the counties' populations, and therefore would likely not scale well when considering the entire state. Thus, our model was limited by the data we were given, in that we were not given the population of each county. If we had been given this, we could have calculated the number of voters for each party. Another limitation in the data was the high number of null values, which we discussed in greater detail above (Section VIII: Methods). Overall, our model uses features such as median age, number of citizens enrolled in Medicare, and the number of ICU beds because we believe them to be robust indicators of how a counties' citizens vote. These features are indicative of how much funds a county designates to its medical sector. Traditionally, blue states have designated more funds towards its medical sector than red states. In addition, states with a higher median age tend to be more red-leaning. Cumulatively, our model leverages features that suggest key insights into a county's voting habits to classify counties' political leanings.

X. Discoveries

Discoveries: One discovery that stood out to us was the highest `dem_to_rep_ratio` in our entire dataset--the District of Columbia with a ratio of 22.23, meaning that for every 22 votes casted for the Democratic candidate, 1 was casted for the Republican candidate in the 2016 presidential election. It was shocking to see how `solid_left` this particular region of the country was. Another discovery was to visually see (Figure 16) how skewed our dataset was in representing republican leaning states as opposed to those who are more democratic leaning. In addition, as shown in Figure 1, Texas appeared to be very overrepresented, not only due to the sheer amount of counties contained within the large state, but also due to the fact that these same counties had complete data. We realize now how important it is to weight states by their population sizes, not just by the number of county representations.

Future Work: One future project we believe could be interesting to do would be to build a model to predict the response time of a county based on their political leaning. We could analyze how political leaning influences a county's preparedness to tackle a medical crisis like COVID-19 based on their current medicare/hospital-inventory data. Given that politics has a huge impact on healthcare in our country, we are very passionate in understanding how counties could better prepare themselves through making better choices, whether they be political, social, or economical.