INVESTIGATING THE EFFECTS OF COVID-19 FACTORS ON ELECTIONS

216proj.ipynb

Jessica Tang, Victoria Midkiff, Vicky Jin, Ellen Zeng

Part 1: Introduction and Research Questions

COVID-19 has had a large impact on every country: advancing digitization, changing the way education operates, and leading to the loss of millions of lives. The effects of COVID-19 impacted us in particular, as we were navigating the transition to college during the peak of the pandemic, making it an interesting topic for us to investigate. The pandemic also created political controversy, with a wide range of opinions on how to properly handle disease spread and implement relevant legislation. We plan to research how COVID-19 affected political outcomes in the United States. More specifically, we will evaluate how the impact of COVID, and its mask use and deaths in particular, altered states' political beliefs and outcomes, and how this played out in the 2020 election.

1. How did the total number of deaths throughout 2019-2020 due to COVID-19 affect the 2020 elections on a state level?

This research question examines how the number of deaths in US states correlates with political outcomes. We will be combining large datasets together and analyzing the correlation between different trends. This will require us to dig deeper into how several factors related to excess death are related to election data, such as the election of more progressive candidates. The relationship between COVID-19 deaths and political outcomes could provide us with more information to prepare for the next election. This is an important topic that still applies today, as society still grapples with the devastating impact of COVID-19.

2. How do COVID-19 trends relating to mask use frequency impact election results by state?

For this question, we will investigate how levels of mask use are related to the 2020 election results. It is relevant to society today as the level of people's usage of masks could indicate whether people are more left-leaning or right-leaning in elections. Mask use was a controversial and widely debated topic during the time of the COVID-19 pandemic, so it will be interesting to see how political outcomes are correlated with mask use frequency.

Part 2: Data Sources

Our project used four data sets. For data on the number of deaths, we used Covid-19 State data from Kaggle, which contained data about COVID-19 impacts. We are concerned with the deaths and population information in the dataset. The deaths information comes from The COVID Tracking Project, and the population information comes from the World Population Review. The election data we used comes from the US Elections Dataset, from Kaggle, which took the data, which is in the public domain, from opendatasoft. This includes information on the number of votes each candidate received from each State in every presidential election from 1976 to 2020.

For data about mask use, we used Mask Use by County from Kaggle, which contains data about the frequency of mask use by county. For each county, the percentage of people identifying as "Never," "Rarely," "Sometimes," "Frequently," and "Always" with regards to mask use are described. This data is sourced from interviews conducted online by the survey firm Dynata at the request of The New York Times. The responses date from July 2 to July 14, 2020. Lastly, we used the US counties Covid 19

dataset, also from the New York Times via Kaggle, which contains county information including fips identification number and state to identify which state each county is in using their fips identification numbers since the Mask Use data only contains the fips identification number.

We analyzed different trends between which states had correlations between COVID-19 excess deaths and mask use and political outcomes. The data we found provides us with necessary details and statistics regarding our topic, and are all helpful in aiding us to ultimately uncover how COVID-19 and related policies and consequences affected our elections in the United States.

Part 3: What Modules are You Using?

Module 4: Data Wrangling is exemplified through our importation of CSV files onto Google Colab to create dataframes to use to answer our questions. This was the first step of our project when examining our datasets. Following loading in the data, we cleaned the data, which is also covered under this module. For example, the COVID19_state.csv used proper capitalization for the state names, but election_data used all capital letters, so we had to adjust the COVID-19 data to be all capital letters in the state name. Our justification is that we needed to clean our data, read the CSV files into the Jupyter Notebook, and check the data for any mistakes.

Module 6: Combining Data was used in the second stage of our project where we prepared our data. We used several CSV files, so our justification is that we had to combine the data sets by state, in order to find how election results were affected by COVID-19 deaths and mask use trends. We used group by in the election data and merged the election data with the COVID-19 data in our data preparation, which are concepts in Module 6.

Module 5: Statistical Inference was used in the data analysis stage of our project. We used different methods to analyze our data on how different COVID-19 factors are related to election data. Our justification is that we found confidence intervals for our predictions using hypothesis testing. We predicted if deaths in Republican states were significantly different than Democrat states in the 2020 election using hypothesis testing. We used hypothesis testing (using ttest_ind()) to find if our prediction is significant, as outlined in Module 5.

Module 8: Visualization is present in the data visualization stage of our project. Our justification is that we plotted specific trends, so our results can be easily and visually summarized. We used bar graphs to visualize our data and find trends between mask usage, deaths, the 2020 election outcome, and voter turnout.

Part 4: Preliminary Results and Methods

Repository link:
https://colab.research.google.com/drive/1as9im5JqkEjQ5fx1Dw8LVPKGj9EuyBCr?usp=sharing

We started our analysis by cleaning our first dataset, “election_data” from the 1976-2020-president.csv, which contains rows for each presidential candidate per state race and data such as the candidate’s party, how many votes each candidate received, and how many total votes there were in that state. Then, we filtered for rows from the 2016 and 2020 election. We needed to find the winning candidate from each state, so we sorted by the “candidate_votes” variable (which stores the number of votes each candidate received), grouped by state, and then used head(1) to extract the winning candidate from each state, creating a new dataset called “election_states.” Each row had each state’s winning party

in 2016 and 2020, the number of votes that candidate received in 2016 and 2020, as well as the total number of votes in 2016 and 2020.

Next, we worked to add data about mask use from the "mask_use" dataset, with the help of the "us-countries-recent" dataset from us-counties-recent.csv. The "mask_use" dataset from mask-use-by-county.csv had 5 columns we were interested in: COUNTYFP, NEVER, RARELY, FREQUENTLY, ALWAYS. The 4 latter columns measured the percentage of people in each county who answered "never," "rarely," "frequently," or "always" to the New York Times survey question "How often do you wear a mask in public when you expect to be within six feet of another person?" Each row of the dataset represented a different county, only identified by a county FP code, that conveniently, the "us-countries-recent" dataset shared. Subsequently, we merged these two datasets by the FP codes, and then grouped-by state, using the "mean" aggregate function for NEVER, RARELY, FREQUENTLY, and ALWAYS (to get the average percent of people in the state who mask correspondingly). Then, we added these four columns to our "election_states" dataset by merging by state. Now, we had a dataset we could easily work with that had data on both mask use and 2016 and 2020 election results that could help us answer our second research question.

1. How did the total number of deaths throughout 2019-2020 due to COVID-19 affect the 2020 elections on a state level?

We started off our analysis by observing the relationship between how a state voted and how much of its population died from COVID-19. From the combined 2016 and 2020 election data frame named combined_elections, we labeled each state "DEMOCRAT," "REPUBLICAN," or "FLIPPED." "DEMOCRAT" for states that voted Democrat in both 2016 and 2020, "REPUBLICAN," for states that voted Republican in 2016 and 2020, and "FLIPPED," for states that flipped to Democrat in the 2020 election; there were no states that flipped to Republican. Then, we selected relevant columns from the COVID-19_state.csv file, or the COVID dataframe, including state, number of deaths, and population, before merging this with the election data by state. We created a new column: percentage of population that died, which contained the percentage of deaths in the population of each state by dividing the number of deaths by the total population.

Figure 1, created with Seaborn catplot, showcases each state's 2020 election party outcome versus the percentage of their population that died due to COVID-19. Based on the plot, states that voted Republican both years had a lower death rate than states that voted Democrat in 2020. The death percentage for flipped states could be higher than Republican states because people who witnessed more COVID-19 deaths could then care more about voting for a candidate who would protect them against the disease and lessen their community's mortality rate. A potential reason why Democrats had a higher percentage of population that died than Republicans could simply be because generally more urban, highly populated places where people are more likely to transmit and catch the disease are demographically and historically lean Democratic.

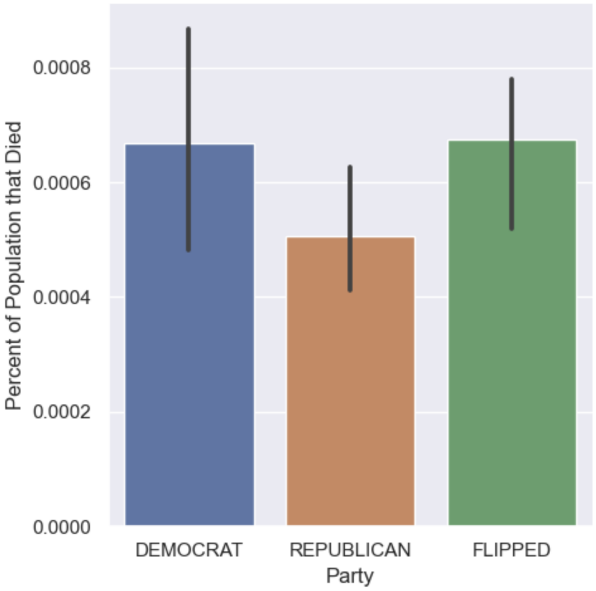

Figure 1. States' 2020 election party outcome VS percentage of population deceased from COVID-19

Figure 2, a Seaborn catplot, visualizes the number of COVID-19 deaths per population for each state and what party won the state in 2020. Based on the image, there seems to be a trend where if there were more deaths per population, they voted more Democrat, but this may not be a very strong correlation as the parties are distributed throughout the graph. A possible explanation for this result could be that after losing more people to COVID, people felt more inclined to vote in ways that would protect citizens from the deadly virus.

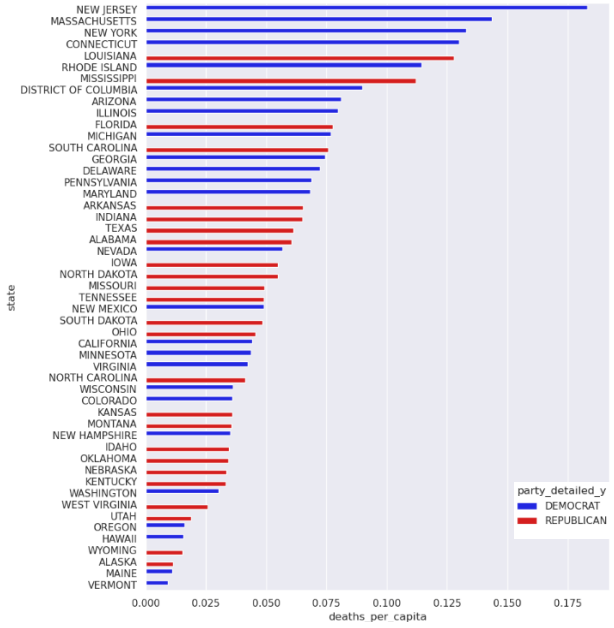

Figure 2. COVID-19 deaths per capita in each state and how they voted in the 2020 election

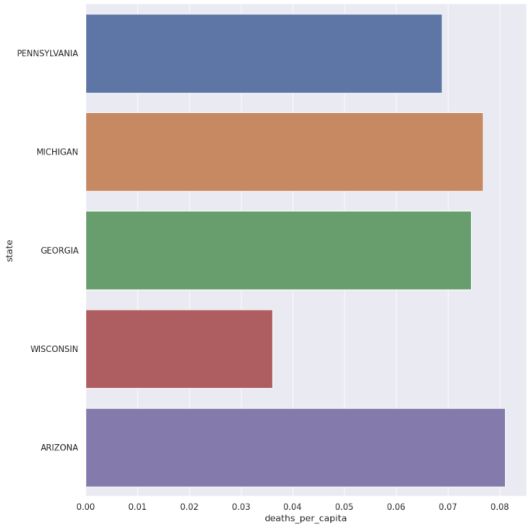

Figure 3. Flipped states and their COVID-19 deaths per capita

Figure 3, another Seaborn catplot, portrays the five swing states that flipped their vote from the 2016 election to the 2020 election: Pennsylvania, Michigan, Georgia, Wisconsin, and Arizona, and how many COVID-19 deaths they had per population. Besides Wisconsin, the other four states had an above-average percentage of deaths per population, which could imply that voters within these states were encouraged to change their vote from Republican to Democratic because of the Democratic party's stricter COVID policies.

We then did hypothesis testing to find whether the winning party of each state was related to the number of deaths per population. We used two-sided t-tests for two independent samples of scores. In this hypothesis test, we used deaths_per_pop from our deaths_elec dataset. Our null hypothesis was: "In the 2020 election, the deaths per population for states who voted Democrat is equal to the deaths per population for states who voted Republican." Our alternative hypothesis was: "In the 2020 election, the deaths per population for states who voted Democrat is not equal to the deaths per population for states who voted Republican."

To see whether or not election results were affected by the deaths per population by state, we first aggregated the death_per_pop values for all states who voted Democratic in the 2020 election and put it into a variable called demo_death. We also aggregated the death_per_pop values for all states who voted Republican in the 2020 election and put it into a variable called rep_death. We then used a two-sided t-test because we are comparing two complete datasets/samples . We obtained a p-value of 0.11962, which indicated that we cannot reject the null hypothesis that the deaths per population for states who voted Democrat are equal to the deaths per population for states who voted Republican in the 2020 election.

We wanted to see if we could find other evidence to support our conclusions from the two-sided t-test, so we created two confidence intervals. The first confidence interval finds the bounds of deaths_per_pop for Democratic states. Using the mean of demo_death, we used stats.norm.interval to find the 95% confidence interval for this group of states. We did the same thing for rep_death. We are 95% confident that the true deaths per population for Democratic states in the 2020 election is between

0.00050258 and 0.00083648. We are 95% confident that the true deaths per population for Republican states in the 2020 election is between 0.000401809 and 0.00061130. These confidence intervals support our two sided t-tests because they overlap, indicating that we fail to reject the null hypothesis. However, the p-value still indicates that there is only a 11.96% chance of getting the results we had. Most likely, the percentage of deaths in states that voted Democrat is not equal to the percentage of deaths in states that voted Republican, even though there is not enough evidence to say so with 95% confidence.

2. How do COVID-19 trends relating to mask use impact election results by state?

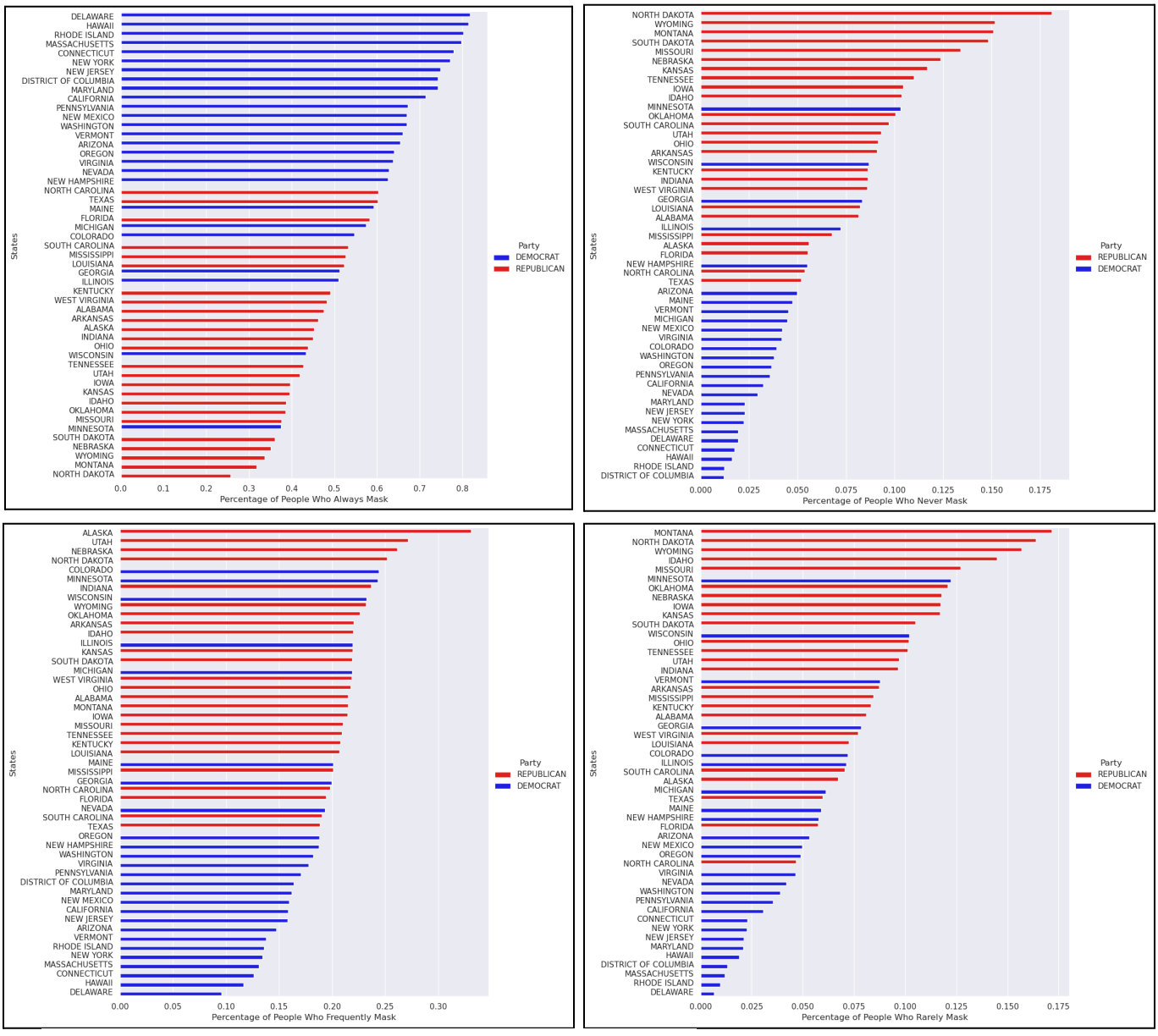

Figure 4. States' 2020 election party outcome VS how frequently people mask

These graphs in Figure 4, created by Seaborn catplot, all depict different ways that states mask and how each state voted in the 2020 election. There are four graphs depicting the people in each state who always mask, frequently mask, rarely mask, and never mask. They are color-coded by their parties.We can observe in the "always masking" and "never masking" graph that Democrats report that

they always mask much more than Republicans and Republicans report that they never mask much more than Democrats. However, the states with high percentages of people who frequently mask and the states with high percentages of people who rarely mask seem to also trend Republican, which could potentially be because this data is collected from a 1-question survey, and Republicans may be more likely to answer "frequently masking" or "rarely masking" as opposed to "always masking.". Additionally, we can see that the states with the highest percentage of people who always mask – Delaware, Hawaii, and Rhode Island –are all highly Democratic states. The same can be said for states that have the highest percentage of people who never mask, North Dakota, Wyoming, and Montana, which are also historically Republican states.

Figure 5, a seaborn cat plot, portrays the five swing states that changed their vote from the 2016 election to the 2020 election and the percentage of people that always wore a mask within these states. Pennsylvania, Michigan, and Arizona had an above-average percentage of people who always masked, while Georgia and Wisconsin had a below-average percentage of people who always masked. This data makes sense because these states are swing states, meaning that they have a close to equal amount of Democratic and Republican voters and the states can swing to either party in an election. Therefore, the mask use within these states would not be as extreme as in states that are consistently Democratic.

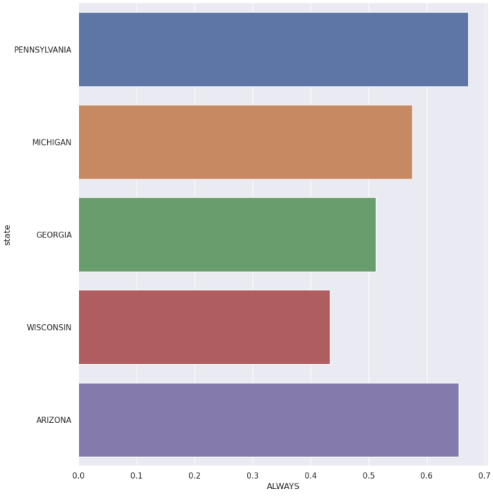

Figure 5. Flipped states and their percentage of population that always masks

We then conducted hypothesis testing to determine if there were significant differences in "always masking" and "never masking" between Democratic and Republican states. We chose to only evaluate these two frequencies because as mentioned earlier, we found that frequently and rarely masking could just be a result of not always or never masking. For our first t-test, our null hypothesis was that in the 2020 election, the always masking proportion for states who voted Democrat is equal to the always masking proportion for states who voted Republican. We received a p-value of 0.00000160816. Since our p-value is significant, we reject the null hypothesis and determine that Democratic states "always mask" at a different proportion than Republican states. We then created two 95% confidence intervals for "always masking," one for Democratic states and one for Republican states. We found, with 95% confidence, that the true "always masking" proportion for Democratic states was between (0.61497, 0.70333), and the "always masking" proportion for Republican states was between (0.40710, 0.47572). Since the confidence interval shows that the true mean "always masking" proportion of Democratic states is higher than the true mean of the other states, we can conclude that Democratic states from the 2020 election on average "always mask" more.

For our second t-test, our null hypothesis was that in the 2020 election, the never masking proportion for states who voted Democrat is equal to the never masking proportion for states who voted Republican. We received a p-value of 4.4750 * 10^-9. Since our p-value is significant at alpha = 0.05, we reject the null hypothesis and determine that Democratic states "never mask" at a different proportion than Republican states. We then created two 95% confidence intervals for "never masking," one for Democratic states and one for Republican states. We found, with 95% confidence, that the true "never masking" proportion for Democratic states was between (0.03135, 0.04927), and the "always masking" proportion for Republican states was between (0.08727, 0.11311). Since the confidence interval shows that the mean "never masking" proportion for Republican states is higher than the mean of all the other states, we can conclude that Republican states from the 2020 election on average "never mask" more.

Part 5: Limitations and Future Work

Upon merging our data frames and understanding the applications of doing this, we noticed that the COVID19_state dataset stops at March 7, 2021. We merged this dataset with "election_states" to understand how the total number of deaths due to COVID-19 affected the 2020 election on a state level. However, to make more sense of the data, it would have been optimal for the COVID19_state data to stop at the end of 2020. There was no way for us to drop the values of the dataset that were collected after 2020 since there was no information about the dates when each observation was collected. We believe that this could have an impact on the findings from the hypothesis test as well as the 3 visualizations that help investigate this question. It would be particularly impactful for the swing state graph, as this limitation could change the states shown in this graph.

Another limitation we noticed was that the data from the mask-use-by-county data frame was collected from July 2 to July 14, 2020. However, the 2020 election took place in November. The data was collected in a short timeframe, and the time gap between July and November could affect the values of the columns in the dataframe (NEVER, RARELY, SOMETIMES, FREQUENTLY, ALWAYS). These columns are categorized into 5 groups that answer the question: "How often do you wear a mask in public when you expect to be within six feet of another person?" (For the purposes of our analysis, we did not look at the "SOMETIMES" column). Conducting the survey for a longer time to collect data would better show the estimated share of people in each county who identify with one of the 5 categories. Additionally, many states implemented mask requirements after the time that the data was collected, which would alter the figures in the dataframe. This directly impacts our visualizations that incorporate the mask-use-by-county data frame. If more mask requirements were put in place after July 14, we could hypothesize that the "FREQUENTLY" and "ALWAYS" categories would have higher values and shift the data to become more left-skewed, which could show more conclusions towards a positive, linear relationship between mask use and voting Democratic.

In future work, we would like to examine the relationship between mask use and election results in greater depth through extending our analysis with more hypothesis testing and logistic regression. We would also like to use logistic regression to see whether excess death relates to the number of voters in an election. Logistic regression may allow us to discover other confounding variables in the data frames and, as a result, pivot from those relationships to investigate other factors that affect the election results.

Part 6: Conclusion

For our first research question, we evaluated how COVID-19 deaths affected the 2020 election on a state-level. From our visualizations, we observed that, for example, eight out of ten of the states with the most deaths per capita voted Democrat in 2020. This could be because more people in these states had their vote affected by COVID-19 after having people in their life pass away from the virus. We performed a two-sided t-test and created confidence intervals, and concluded that the deaths per population did not significantly differ between Democratic and Republican states, as we received a p-value of 0.11.

For our second question, we looked at how mask use affected 2020 election results on a state-level. Our visualizations clearly show a stark partisan divide in our graphs depicting the number of people who report that they "always mask" and "never mask." We performed two two-sided t tests and created confidence intervals, and with p-values of 0.00000160816 (for "always masking") and 4.4750 * 10^-9 (for "never masking), we concluded that Democratic states from the 2020 election on average "always mask" more and Republican states from the 2020 election on average "never mask" more.